**UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL**
**ESCOLA DE ENGENHARIA**
**PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO**

**Carla Fiorioli**

# QUANTITATIVE ANALYSIS OF STUDENTS' PERFORMANCE IN THE BRAZILIAN HIGH SCHOOL EXAM

**Porto Alegre**
**2019**

Carla Fioriolli

# QUANTITATIVE ANALYSIS OF STUDENTS' PERFORMANCE IN THE BRAZILIAN HIGH SCHOOL EXAM

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do título de Mestre em Engenharia de Produção, modalidade Acadêmica.

Orientador: Flávio Sanson Fogliatto, *Ph.D.*

Porto Alegre
2019

Carla Fiorolli

# QUANTITATIVE ANALYSIS OF STUDENTS' PERFORMANCE IN THE BRAZILIAN HIGH SCHOOL EXAM

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Engenharia de Produção na modalidade Acadêmica e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul.

**Flávio Sanson Fogliatto, *Ph.D.***
Orientador PPGEP/UFRGS

**Flávio Sanson Fogliatto, *Ph.D.***
Coordenador PPGEP/UFRGS

**Banca Examinadora:**

Professor José Luís Duarte Ribeiro, Dr. (PPGEP/UFRGS)

Professor Michel José Anzanello, *Ph.D.* (PPGEP/UFRGS)

Priscila Goergen Brust Renck, *Ph.D.* (PPGEP/UFRGS)

*Dedico essa dissertação à minha família e em*
*especial aos meus pais José Carlos e Rosana,*
*à minha irmã Thais e à minha sobrinha e*
*afilhada Kiara.*

# AGRADECIMENTOS

Em primeiro lugar, agradeço aos meus pais, José Carlos e Rosana, pelos ensinamentos, pelo carinho, pelo apoio incondicional, por sempre acreditarem na minha capacidade e pelo papel decisivo que tiveram na conclusão desta dissertação.

Agradeço também ao meu professor e orientador, Flávio Sanson Fogliatto, por sua dedicação e por sua valiosa orientação, também imprescindíveis à conclusão deste trabalho.

À minha irmã, Thais, por sempre acreditar em mim, pelas palavras de apoio nos momentos mais difíceis e pelo carinho. À minha sobrinha e afilhada Kiara, pelos sorrisos, pela leveza e pelos momentos de descontração proporcionados. À minha tia e madrinha Neiva e à minha tia Graça pelo carinho, pelo apoio e pela compreensão. À toda minha família pelo carinho e pela compreensão mesmo na minha ausência.

A todos os meus amigos, em especial às minhas amigas Camila, Joana e Marina, e ao meu amigo Pedro pela compreensão, pelos momentos de descontração e pelas palavras de incentivo. Aos amigos Mateus Hexsel e Gabriela Cardoso pelo apoio e pelas intervenções em momentos fundamentais.

Aos meus colegas de curso pelo apoio e pelo companheirismo, em especial à amiga Raísa e ao amigo Bernardo pelas conversas e pelas palavras de incentivo e à colega Gabrielli pelo ágil apoio em um momento fundamental para a conclusão deste trabalho.

Aos professores, coordenação e equipe do PPGEP, por todo suporte e dedicação. Ao CNPq pelo auxílio financeiro.

Agradeço também a todos que passaram pela minha vida e que me proporcionaram aprendizados e me fizeram crescer.

Por fim, agradeço pela energia universal e sagrada que me proporcionou a força necessária para trilhar esse caminho.

*"Querer ser livre é também querer livres os outros."*

*Simone de Beauvoir*

# ABSTRACT

This dissertation presents an analysis of the relationship between sociodemographic attributes and students' performance in the Brazilian High School Exam (ENEM). The main objective is to analyze the impact of sociodemographic variables on students' performance by modeling their score in the five knowledge areas evaluated in ENEM as a function of variables that characterize these students. The method adopted to perform the analysis was an adaptation of the CRISP-DM (Cross-Industry Standard Process for Data Mining) method, which is implemented in five phases: (i) data/environment understanding, (ii) data preparation, (iii) modeling, (iv) evaluation/analysis, and (v) discussion. In all models obtained, the variable that most explained the variance in students' performance was a dummy variable associated with the type of school attended by the student: those who attended only private schools without a scholarship had an advantage in test scores. A dummy variable related to race was also retained in all models: auto declared white students had an advantage in scores. The sex-related effect varied depending on the area of knowledge analyzed. The most positive effect for males occurred in the mathematics knowledge area, while the most negative effect occurred in the essay. Other variables such as students' fathers' and mothers' level of education, fathers' occupation, and the ownership of a computer were included in all models. Models obtained yielded an average variance explained of 17.90%, which is consistent to what is observed in other studies of the same nature, and suggests that the method employed is suitable for this type of analysis.

Keywords: Assessment exams, ENEM, sociodemographic attributes, data mining, partial least squares regression.

# RESUMO

Esta dissertação apresenta uma análise quantitativa do desempenho dos estudantes no Exame Nacional do Ensino Médio (ENEM) em relação a variáveis sociodemográficas. O objetivo deste trabalho é a análise do impacto das variáveis sociodemográficas no desempenho dos estudantes através da modelagem deste desempenho nas 5 áreas de conhecimento avaliadas no ENEM como função das variáveis que caracterizam esses estudantes. Para a realização da análise, utilizou-se uma versão adaptada do método CRISP-DM (*Cross-Industry Standard Process for Data Mining*) implementado em cinco fases: (i) compreensão dos dados/ambiente, (ii) preparação dos dados, (iii) modelagem, (iv) avaliação/análise e (v) discussão. Em todos os modelos gerados, a variável que melhor explica a variação no desempenho dos estudantes é a variável *dummy* associada ao tipo de escola frequentada pelo aluno; aqueles que frequentavam apenas escolas privadas sem bolsa de estudos tiveram vantagem nos escores. A variável *dummy* relacionada à raça também foi mantida em todos os modelos: os estudantes auto declarados brancos tiveram vantagem nos escores. O efeito relacionado ao gênero variou dependendo da área de conhecimento analisada. O efeito mais positivo para o sexo masculino ocorreu na área de conhecimento de matemática, enquanto o efeito mais negativo ocorreu na redação. Outras variáveis, como o nível de escolaridade dos pais e mães dos estudantes, a ocupação dos pais e a posse de um computador foram incluídas em todos os modelos. A modelagem desenvolvida nesta dissertação explica, em média, 17,90% da variância do desempenho dos estudantes no ENEM, o que é consistente com os resultados obtidos em estudos de mesma natureza. Esta condição sugere que o método utilizado é adequado à realização deste tipo de análise.

Palavras-chave: Sistemas de avaliação, ENEM, atributos sociodemográficos, mineração de dados, regressão por mínimos quadrados parciais.

## CONTENTS

# 1. INTRODUÇÃO

Sistemas de avaliação têm como um dos propósitos a promoção da igualdade na educação (Airasian, 1988). Desde o início do século 20, o *British School Certificate* passou a ser o documento oficial que atesta a conclusão do ensino médio e permite que os estudantes concorram por vagas em universidades Britânicas (Broadfoot, 2012). Ao longo do tempo, países como os Estados Unidos da América, a Austrália, a Nova Zelândia, e o próprio Reino Unido passaram a associar o seu desenvolvimento econômico com o desempenho educacional dos seus estudantes, e certificados como esses passaram a ser utilizados para monitorar e melhorar seus sistemas educacionais (Brown e Lauder, 1996).

Existem, no entanto, pontos positivos e negativos na utilização de sistemas de avaliação, tanto do ponto de vista de certificação e seleção, como do ponto de vista de utilização para desenvolvimento de políticas sociais. A natureza objetiva das avaliações fornece não apenas um meio de efetivamente comparar indivíduos, mas também induz um senso de justiça e representa a ordem e o controle (Airasian, 1988). Todavia, existem argumentos a favor da subjetividade do conhecimento que corroboram a teoria de que se somos seres sociais que desenvolvem percepções de valores e experiências; assim, as avaliações projetadas por seres sociais carregam o viés de dois grupos: aqueles que as desenvolveram e aqueles que as avaliam. Os vieses transferidos para esses sistemas tendem a sistematizar e reforçar a estratificação social (Gipps, 1999).

Independentemente das desvantagens dos sistemas de avaliação educacionais, há uma consciência de que o desenvolvimento econômico das nações depende do desenvolvimento educacional de sua população (Eckstein, 1996). A Alemanha, a França e o Japão, considerando prós e contras, decidiram sistematizar os procedimentos de avaliação e centralizar os programas nacionais de educação; esses países estão geralmente posicionados no topo dos rankings de educação (Green, 1997).

No Brasil, o ENEM (Exame Nacional do Ensino Médio) é o Exame Nacional do Ensino Médio anual realizado por estudantes. Foi criado em 1998 e é administrado pelo Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) com o objetivo principal de avaliar o desempenho dos alunos na conclusão do Ensino Médio. Desde 2009, o exame tem sido amplamente adotado como uma nota principal ou complementar para a entrada em universidades públicas. A participação no exame aumentou de cerca de 160.000 estudantes em 1998 para mais de 4 milhões em 2008. Após a adoção pelas universidades, a participação aumentou ainda mais, alcançando mais de 8,5 milhões de estudantes em 2016. O ENEM não é e nunca foi um exame obrigatório (INEP, 2017).

Independentemente dos pontos de vista divergentes em relação aos exames de avaliação, o ENEM é realizado por milhões de estudantes, sendo uma fonte prolífica de informações para entender o que está impactando o desempenho dos estudantes brasileiros (Vahdat et al., 2015). Seguindo experiências internacionais bem-sucedidas, políticas públicas e currículos centralizados/padronizados poderiam ser desenvolvidos com base no conhecimento extraído desse tipo de dado (Runci et al., 2017), permitindo o desenvolvimento de melhores estratégias educacionais (Vahdat et al., 2015; Connelly et al., 2016; De Rosa, 2017).

## 1.1. TEMA E OBJETIVOS

De acordo com a contextualização anteriormente exposta, propõe-se o seguinte tema: a análise quantitativa do desempenho dos alunos no Exame Nacional do Ensino Médio (ENEM) em relação a variáveis sociodemográficas.

O objetivo geral deste trabalho é a análise do impacto das variáveis sociodemográficas no desempenho dos estudantes através da modelagem deste desempenho nas 5 áreas de conhecimento avaliadas no ENEM, como função das variáveis que caracterizam esses estudantes.

Os objetivos específicos são: (i) a análise dos coeficientes gerados pelos modelos, (ii) as discussões acerca da análise dos coeficientes, e (iii) a análise comparativa dos resultados obtidos com resultados da literatura.

## 1.2. JUSTIFICATIVA

Como mencionado anteriormente, o ENEM tem sido amplamente adotado pelas universidades públicas como um escore complementar e/ou principal e os microdados fornecidos pelo INEP (2017) têm uma riqueza de detalhes e podem ser explorados com eficiência usando ferramentas de mineração de dados. Eckstein (1996) argumenta que as nações estão conscientes de que seu desenvolvimento econômico depende do desenvolvimento educacional de sua população. Brown e Lauder (1996) acrescentam ainda que diversos países têm utilizado os resultados de exames como o ENEM para monitorar e melhorar seus sistemas educacionais.

Experiências internacionais mostram que o estudo dos microdados pode trazer subsídios para o desenvolvimento do plano educacional e de políticas públicas relacionadas à educação (Runci et al., 2017), permitindo a elaboração de melhores estratégias educacionais (Vahdat et al., 2015; Connelly et al., 2016; De Rosa, 2017). Desta forma, é de interesse do INEP (2017) e socialmente relevante que os dados do ENEM sejam objeto de estudos que contemplem diversas abordagens.

A julgar pela pesquisa bibliográfica conduzida neste trabalho, este é o primeiro estudo a usar técnicas de regressão de segunda geração para mapear e analisar as relações entre os atributos sociodemográficos e o desempenho dos estudantes usando o conjunto de dados do ENEM.

## 1.3. DELIMITAÇÕES DO TRABALHO

Este trabalho foi realizado tendo como referência a base de dados do ENEM do ano de 2017. O estudo contempla os três estados da região Sul (Rio Grande do Sul, Santa Catarina e Paraná). O Brasil é um país amplo, com diferentes culturas e características. Os três estados foram selecionados devido às semelhanças dessas populações e suas diferenças com relação a outros estados e regiões. Esta foi uma abordagem necessária, uma vez que a análise de todo o conjunto de dados poderia resultar em conclusões pouco generalizáveis. Menezes-Filho (2007) e Viggiano e Mattos (2013) apresentam evidências do efeito dessas diferenças. Em função disso, há a oportunidade de aplicação do mesmo método também para as demais regiões brasileiras.

Em relação aos dados utilizados nesse estudo, não foram explorados os aspectos relacionados à localização dos alunos dentro da região e de cada estado, à qualidade individual das escolas que cada aluno frequentou, bem como características individuais de saúde, deficiências ou necessidades especiais. Essas variáveis poderiam servir de subsídio para outros estudos que avaliem outras dimensões do desempenho dos alunos, não comtempladas no presente estudo.

O ENEM pode ser realizado por pessoas de qualquer idade. Nesse estudo, foi tomada a decisão de se manter todos os estudantes na análise, independentemente de idade. Estudos similares poderiam ser desenvolvidos utilizando-se somente as observações cujos alunos se enquadram em alguma faixa etária específica.

Em relação aos métodos e ferramentas utilizadas durante a preparação dos dados e da geração e avaliação dos modelos, estes foram escolhidos conforme a sua adaptabilidade para o caso em questão, mas não necessariamente são os únicos existentes. Esse estudo não tem como pretensão cobrir todos os métodos e ferramentas possíveis para cada etapa do processo. No entanto, há a responsabilidade de utilizar métodos e ferramentas adequadas para cada situação.

## 1.4. ESTRUTURA DO TRABALHO

Este trabalho está estruturado em 5 capítulos. O capítulo 1 apresenta uma visão geral do trabalho, apresentando seu tema e objetivos, justificativa, delimitações e a sua estrutura. O capítulo 2 traz uma revisão da literatura acerca do tema proposto e dos métodos utilizados no desenvolvimento do trabalho. O capítulo 3 apresenta a metodologia e o detalhamento das etapas utilizadas no desenvolvimento do modelo enquanto que o capítulo 4 traz uma discussão acerca dos resultados obtidos. Por fim, o capítulo 5 faz o fechamento do trabalho, apresentando as principais conclusões obtidas, bem como sugestões para trabalhos futuros.

# 2. BACKGROUND

## 2.1. ASSESSMENT EXAMS

Examination systems can play the role of selection, certification, and monitoring for either educational or professional purposes (Keeves, 1994). From the educational perspective, British universities were the first entities to institute entrance exams (Gipps, 1999); at the start of the 20th century, the British School Certificate became the official document assuring school termination and eligibility to enter universities in England (Broadfoot, 2012). By the end of the 20th century, some countries (e.g. United Kingdom, United States of America, Australia and New Zealand) began to associate economic development with educational performance, using assessment results to monitor and improve their educational systems (Brown and Lauder, 1996).

One of the purposes for the creation of assessment systems (namely, decreasing the privileges of the wealthy and patronized) is to promote equality in education. The objective nature of assessments provides not only a means to effectively compare individuals, but also induces a sense of justice and represents order and control (Airasian, 1988). They play social and economic functions of equally distributing roles, adding value to those who undertake them (Broadfoot, 1996; Gipps, 1999). However, there are arguments in favor of the subjectivity of knowledge that corroborates the theory that if we are social beings that develop perceptions from values and experiences, then assessments – designed by social beings – will similarly carry the bias of two groups: those who developed them, and those who evaluate them. Bias carried over to these systems tend to systematize and reinforce social stratification (Gipps, 1999).

Regardless of the drawbacks of educational examination systems, Eckstein (1996) argues that nations are conscious that their economic development relies on the educational development of their population. Germany, Japan, and France, considering pros and cons, decided to systematize assessment procedures and centralize national education programs; these countries are usually positioned at the top of education rankings (Green, 1997).

Abitur is an exam taken after secondary education in Germany, being used for certification purposes. The exam follows the national curricula, but it is not mandatory; however, it is a requirement for students willing to apply for tertiary education (Shavit, 2007). The exam consists of both a written and an oral phase, and it is administered by teachers within schools. The evaluation is a responsibility of teachers from the students' school and it is not conducted by any independent entity (Reichelt, 1997).

Reichelt (1997) investigated the quality of form and content of texts written in English by German students in relation to their perceptions of the writing process. The author used Abitur information as well as interviews to accomplish this goal and concluded that the quality of the content is better than the form if compared to native English speakers and that it may be related to the method employed to teach English.

Randler and Frech (2006) showed that the circadian rhythm influences the scores of the Abitur exam. Their study categorized students as evening or morning people in relation to their circadian rhythm, however all of them study in the morning. Evening students tended to display worse scores than morning students. Therefore, they argue that early school start times may have a negative impact on evening students, mainly because their learning curve might be harmed, but also because the Abitur scores are used for university entrance, which could lead to an unfair selection of the best students.

Pilz (2009) analyzed the difference between students that took Abitur and decided not to go to university but rather to enter the financial services sector. He combined a written survey and Abitur scores to understand students' motivations to make such decision. The conclusion is that the decision to take the University path could not be related to poor performance in school (as reflected by Abitur scores); instead, he found the decision to be based in the perception of value in taking this step for the students' careers.

A few studies used Abitur scores in combination with other information (Reichelt, 1997; Randler and Frech, 2006; Pilz, 2009); however, no studies using statistical approaches to analyze Abitur data were found in the consulted literature.

Japan adopts the National Center Test for University Admissions, which is the first stage of a two-step process to enter national and local public (and some private) universities. It is electively taken after the last year of secondary education, being considered a certification. The exam is aligned with guidelines set by the Japanese Ministry of Education and is composed of multiple choice questions, except for the English portion that comprises a writing and a listening phase. The exam is applied to all students at the same time in two days of the same weekend in January following a strict protocol, such that every student has the same conditions (Guest, 2008; Watanabe, 2013).

Arai and Matsuzaki (2014) presented the development status of a robot built to answer a National Center Test. The study used questions from a past exam edition and compared the robot's performance with that of students. Although the robot performed far worse than students who were accepted at the University of Tokyo, it performed well enough to reach the entrance scores of 404 (out of 744) private universities.

In France, the equivalent exam for university admission is the *baccalauréat*, which entitles students to enter higher education. The French exam has three streams: hard sciences, economics, and social sciences and humanities. As well as in the other countries mentioned, the exam is not mandatory (Duru-Bellat and Kieffer, 2008). In spite of the importance of this exam in the country, no quantitative studies analyzing its results were found in the consulted literature.

### 2.1.1. PISA

At a multinational level, the Programme for International Student Assessment (PISA) is an international large-scale assessment launched by the Organisation for Economic Co-operation and Development (OECD) in 1999. It does not follow any particular curriculum; instead, it focuses on a literacy approach (OECD, 2000). The exam takes place once every three years, and each edition focuses on a given area that can be either reading, scientific, or mathematical literacy. PISA is a reliable source to understand each country's position in educational development and has been largely used to discuss educational policies worldwide (Hopfenbeck et al., 2018). Several studies presented in the sequence use different PISA versions to analyze several aspects of students' performance.

Nonoyama-Tarumi (2008) examined the sensitivity of the relationship between family background and educational performance through Ordinary Least Squares. Using the PISA 2000 dataset it was found that the relationship is stronger when multidimensional measures such as parental occupation, parental education, PISA index of home educational resources, PISA index of home possessions related to classical culture, and the number of books at home are used, rather than unidimensional measures (e.g. only parental education and occupation). Lafontaine and Monseur (2009) explored impacts of the assessment format on students' performance according to their sex using a multinomial logit model. Their results point to a larger sex gap associated with how long it takes to answer the questions. Females tend to perform better at questions demanding longer answers. It was also found that open-ended questions generate a significant higher sex gap than multiple-choice questions, which give to females an advantage when answering open-ended questions and to males when answering multiple-choice questions. According to them, the reasons for that could be related to the type of written material that males and females customarily read.

Martins and Veiga (2010) used hierarchical linear models to measure inequalities in the performance of students from 15 European countries as a function of socioeconomic factors. Results show that Sweden and Finland presented the lowest performance impact associated with socioeconomic factors, and that Belgium, Germany, Great Britain, Greece, and Portugal presented the highest impact. Perry and McConney (2010) examined the characteristics of students' performance and socioeconomic attributes of schools in Australia using descriptive statistics. Their findings show that students who attend schools in which the average students' socioeconomic status are higher perform better than those who attend schools with a lower average, regardless of their own socioeconomic status.

De Lange et al. (2014) used multilevel analysis to explore the relationship between student's performance and the presence of two biological parents at home. They show that student's performance is affected by the number of single-parent children in a school, especially when children are from single-mother families. The less present the parents are, the lower the performance of the students. Other authors use different methods to analyze educational exams datasets, such as meta-analysis (Sirin, 2005; Lietz, 2006) and quantile regression (Jerrim, 2012). These studies show a variety of possibilities to explore these datasets.

### 2.1.2. ENEM

ENEM is the annual National High School Exam taken by students in Brazil. It was created in 1998 and run by the Brazilian National Institute of Studies and Educational Research (INEP) with the primary objective of evaluating students' performance upon High School completion. Since 2009 the exam has been largely adopted by public universities as complementary or as main score for granting entrance to students. Participation in the exam increased from around 160,000 in 1998 to more than 4 million in 2008. After adoption by universities, participation increased even further, reaching more than 8.5 million in 2016. ENEM is not and has never been a mandatory exam (INEP, 2017).

ENEM is comprised of 180 objective questions covering four areas of knowledge (45 questions per area): languages and codes (LC), human sciences (HS), natural sciences (NS) and mathematics (MA). An essay (ES) on a chosen subject is also part of the exam. ENEM is taken in two weekend days: LC and HS questions and the essay are presented on the first day; NS and MA questions close the exam on the second day. Students are also required to answer a questionnaire covering socioeconomic attributes upon subscription. Answers to the questionnaire, students' anonymized data and their respective scores in each area of knowledge are made available annually at the INEP website (INEP, 2017).

Considering the ENEM dataset analyzed in this study there was one article found, due to Stearns et al. (2017), that uses a complex analytical approach: the Boosted Regression Trees techniques. They compare two such techniques, AdaBoost and Gradient Boosting, to predict student's score and conclude that Gradient Boosting gives the best results.

## 2.2. PARTIAL LEAST SQUARES (PLS)

ENEM microdata provided by INEP has a wealth of detail that can be efficiently explored using data mining tools. The multivariate nature of the information contained in the dataset cannot be appropriately described by univariate descriptive statistics (Vahdat et al., 2015) or first-generation regression techniques (Haenlein and Kaplan, 2004). First-generation regression techniques such as multiple regression analysis, discriminant analysis, logistic regression and analysis of variance have known limitations. Haenlein and Kaplan (2004) enumerate three of them: (i) the assumption of no correlation between independent variables, (ii) the assumption that variables are observable – i.e., measurable in practice – and (iii) the assumption that observations are measured without systematic error.

Partial Least Squares (PLS) regression provides an alternative for the limitations above. It is a principal component (PC)-based regression, in which independent variables are rewritten as orthogonal PCs that capture most of the variance useful for the model. PLS is suitable to model multicollinear datasets and unobservable attributes that are representable by a combination of observable variables, being also robust to measurement errors (Haenlein and Kaplan, 2004; Cassel et al., 1999). PLS is not based on any distributional assumption for the variables (Fornell and Bookstein, 1982) broadening its applicability in comparison to first-generation regression techniques. Due to its characteristics and advantages, PLS has been used in social research (Esposito Vinzi et al., 2010) as well as in the food industry (Granato and Ares, 2014), genomics (Zhou et al., 2014) and more remarkably in chemometrics (Wold et al., 2001).

As mentioned before, PLS regression is based on PC regression. Furthermore, PC regression is based on PCA (Principal Component Analysis). For PC regression, a PCA is conducted and then the first $\boldsymbol{k}$ principal components are used to perform the regression. The explanation that follows was elaborated in line with the algorithm description from Geladi and Kowalski (1986) and Wold et al. (2001).

Assuming **X** is an $\boldsymbol{n} \times \boldsymbol{m}$ matrix with $\boldsymbol{n}$ data points and $\boldsymbol{m}$ variables, the objective of PCA is to find an orthogonal $\boldsymbol{m} \times \boldsymbol{m}$ matrix **W** comprised of new variables $\boldsymbol{t_1}, \ldots, \boldsymbol{t_m}$ that are uncorrelated and organized in order of decreasing variance, such that:

$$\mathbf{T} = \mathbf{XW} \tag{1}$$

where matrix **W** is an $\boldsymbol{m} \times \boldsymbol{m}$ matrix of weights for variables in **X**, and **T** is an $\boldsymbol{n} \times \boldsymbol{m}$ matrix of scores.

Considering most of the variance in **X** is described by the first principal components, an approximation of **X** can be constructed using the first $\boldsymbol{k} \leq \boldsymbol{m}$ PCs:

$$\mathbf{T}_{|\boldsymbol{k}} := \mathbf{XW}_{|\boldsymbol{k}} \tag{2}$$

where $\mathbf{T}_{|\boldsymbol{k}}$ is an $\boldsymbol{n} \times \boldsymbol{k}$ condensation of the original $\boldsymbol{n} \times \boldsymbol{m}$ matrix that captures most of the variance in matrix **X**. The main goal in PC regression is to use the $\mathbf{T}_{|\boldsymbol{k}}$ matrix instead of the complete **X** matrix, considering a convenient value for $\boldsymbol{k}$. By definition, each variable in $\mathbf{T}_{|\boldsymbol{k}}$ is uncorrelated.

PLS uses the basic concept of PC regression. PLS accepts multiple variables both for the independent **X** and dependent **Y** variables matrices, decomposing them as:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \tag{3}$$

and

$$\mathbf{Y} = \mathbf{UC}' + \mathbf{F} \tag{4}$$

where **U** is a score matrix equivalent to **T** for **Y**, **C** is a weights matrix equivalent to **W** from Equation 1, and **P** is a matrix of loadings that represents the relationship between **T** and **U**. Matrices **E** and **F** give the residual errors for **X** and **Y**, respectively from Equations 3 and 4. With **T** being used as a good estimator for **Y**, then

$$\mathbf{Y} = \mathbf{TC}' + \mathbf{F}^* \tag{5}$$

where $\mathbf{F}^*$ is the matrix of residual errors considering the relationship between $\mathbf{T}$ and $\mathbf{U}$. Figure 1 shows a graphical representation of the scheme.

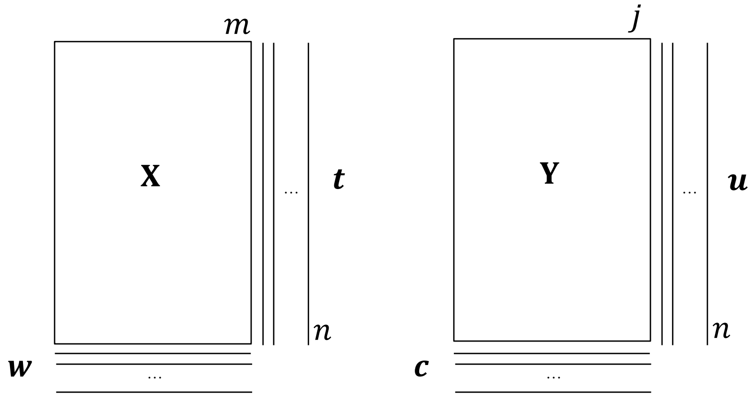

**Figure 1:** Graphical representation of PLS matrices and vectors

PLS does not consist of performing PC regressions on **X** and **Y**, individually. Instead, it searches for a model that maximizes the covariance between **X** and **Y.** PLS allows scores in **U** to be predicted by scores in **T** since the covariance between them is maximum; i.e.,

$$\mathbf{u} = \mathbf{pt} \tag{6}$$

with $\mathbf{p}$ being the vector that maximizes the covariance between $\mathbf{u}$ and $\mathbf{t}$.

In addition, scores in **T** can be estimated as a linear combination of the original matrix **X** and the matrix of loadings **P**; namely:

$$\mathbf{T} := \mathbf{XP} \tag{7}$$

which, from equations (5) and (7), leads to the following:

$$\mathbf{Y} = \mathbf{XPC}' + \mathbf{F}^* \tag{8}$$

or

$$\mathbf{Y} = \mathbf{XB} + \mathbf{F}^* \tag{9}$$

where $\mathbf{B}$ is the matrix of regression coefficients.

The PLS regression algorithm is an iterative process. Consider two multivariate matrices $\mathbf{X}$ and $\mathbf{Y}$, centered and scaled. The process starts with (i) a random $\mathbf{u}$ vector that can be a single $\mathbf{y}$ column, (ii) $\mathbf{w}$ is then computed as $\mathbf{X}'\mathbf{u}/\mathbf{u}'\mathbf{u}$, (iii) $\mathbf{t} = \mathbf{Xw}$, (iv) $\mathbf{c} = \mathbf{Y}'\mathbf{t}/\mathbf{t}'\mathbf{t}$, and finally (v) $\mathbf{u} = \mathbf{Yc}/\mathbf{c}'\mathbf{c}$. The iterative process stops when $\mathbf{t}$ converges. However, the process needs to take place for each principal component, that in PLS are named latent variables (LV), from 1 to $\boldsymbol{k}$. After each round of iterations, $\mathbf{X}$ and $\mathbf{Y}$ are recalculated as $\mathbf{X} = \mathbf{X} - \mathbf{tp}'$, with $\mathbf{p} = \mathbf{X}'\mathbf{t}/\mathbf{t}'\mathbf{t}$ and $\mathbf{Y} = \mathbf{Y} - \mathbf{tc}'$ and the iteration starts again for the next LV. Once all LVs have been calculated and there is no more information in $\mathbf{X}$ about $\mathbf{Y}$, then all $\mathbf{p}$ vectors form the $\mathbf{P}$ matrix and all $\mathbf{c}$ vectors form the $\mathbf{C}$ matrix and the regression coefficients can be extracted from $\mathbf{B}$ matrix which is formed by the $\mathbf{b}$ vectors, calculated by $\mathbf{b} = \mathbf{pc}'$.

The aim of this section is to give a brief overview of the PLS algorithm. More details are available in Wold et al. (1984), Geladi and Kowalski (1986), Wold et al. (2001), and Haenlein and Kaplan (2004).

## 3. DATA AND METHODS

The Cross Industry Standard Process for Data Mining (CRISP-DM) method was created to be adaptable to any industry, tool or application. It is the suggested method to be followed in any data science project by Provost and Fawcett (2013), being comprised of six phases in its original proposition (Shearer, 2000), and adapted to five in the present study; they are: (i) Data/Environment Understanding (originally denoted as Business Understanding and Data Understanding), (ii) Data Preparation, (iii) Modeling, (iv) Evaluation/Analysis (originally denoted as Evaluation), and (v) Discussions (originally denoted as Deployment). The adapted method phases are depicted in Figure 2.

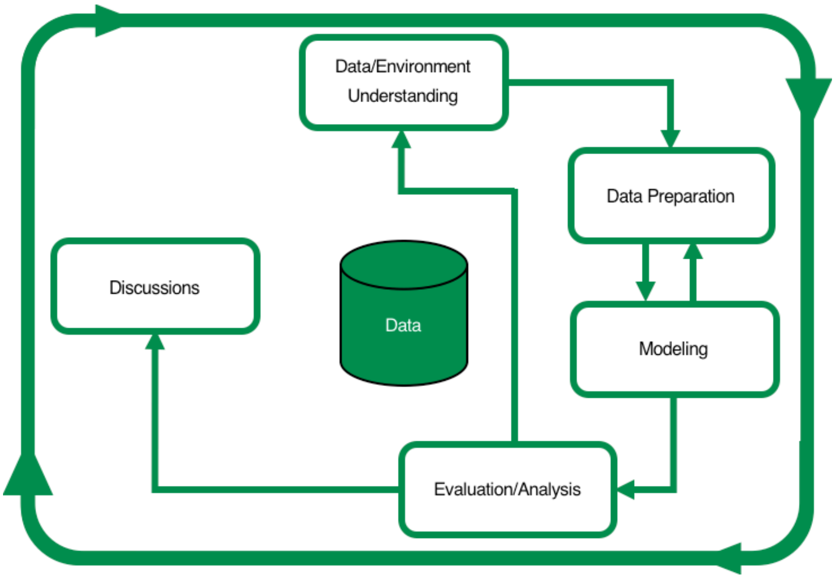

**Figure 2:** Adapted CRISP-DM method. Source: Adapted from Shearer (2000)

Although ENEM is applied in all Brazilian regions, the analysis was restricted to students with residence in the three southernmost states of Brazil (Paraná, Santa Catarina, and Rio Grande do Sul). This was a necessary approach due to the similarities of these populations and their differences to other states and regions. Brazil is a broad country with different cultures and characteristics; an analysis of the entire dataset could result in misleading conclusions. Menezes-Filho (2007) and Viggiano and Mattos (2013) studies present evidence regarding these differences.

## 3.1. DATA DESCRIPTION/EXPLORATION

The portion of the dataset analyzed here has information on 448,949 students that live and took the exam in the South region of Brazil. Only students who provided answers to all four areas of knowledge (AK) and wrote the essay obtaining valid scores were considered in this study. In order to explore and understand the data, a descriptive analysis was run for this group of students.

Shapiro-Wilk and D'Agostino and Pearson's $K^2$ tests were used to check the normality of the distributions of scores. The Mann-Whitney Test was used to check the null hypothesis of equality between groups' average scores in each area of knowledge. The Shapiro-Wilk test verifies the null hypothesis that a sample belongs to a normally distributed population. Hence, in case the *p-value* is smaller than a predefined alpha level, the null hypothesis cannot be rejected and there is evidence that the sample data is not normally distributed. The same procedure is applied in D'Agostino and Pearson's $K^2$ Test for the same purpose. However, tests calculate their statistic differently. Shapiro-Wilk uses the sample values and mean, the expected values if it was a normal distribution and the covariance between those values to obtain the test statistic; D'Agostino and Pearson's uses the sample kurtosis and skewness and the sample mean to calculate it (Shapiro and Wilk, 1965; D'Agostino et al., 1990). In the context of this study, these tests were used to verify if the distribution of scores in each group from selected independents variables followed a normal distribution. Such information was necessary to decide on which test to use to compare scores and conclude if their differences could be considered statistically significant or not.

The Mann-Whitney Test is a non-parametric test that does that. It verifies the null hypothesis that the mean from one sample is higher or lower than the mean from a second sample. The test may be used on independent samples, which is suitable for the comparison of groups within independent variables. Being a non-parametric test, it does not require samples to follow any particular distribution (Mann and Whitney, 1947).

## 3.2. DATA PREPARATION

The dataset was normalized such that scores obtained for each area of knowledge ranged from 0 to 100. Only students who provided answers to all four AKs and wrote the essay obtaining valid scores larger than zero were considered in this study, yielding the final sample of 448,949 observations. Except for Age, all independent variables were rewritten as dummy variables, allowing the information in each category to be captured in raw format and avoiding the assumption of unrealistic relationships between categories. A set of $j - 1$ dummy variables were created for each categorical variable with $j$ categories, following recommendations in Hardy (1993). Thirty-one of the independent variables available were selected for this study; once transformed to dummy variables a total of 91 independent variables became available, in addition to the 5 dependent variables. Appendix A presents the selected variables and their descriptions, as well as reference groups.

### 3.3. MODELING

Data modeling was carried out using both 'pls' Package (Mevik et al., 2016) and 'plsVarSel' Package (Liland et al., 2017) in R. PLS regression was the technique chosen to model the data. The dataset was split into training and testing portions, in the proportion of 3 to 1. The testing portion was used to generate the final performance metrics for the models obtained using the training portion (Provost and Fawcett, 2013).

The first set of models to be generated were the complete models, each comprising of all variables. Each dependent variable had its own model generated. Such approach had the purpose of selecting the proper number of latent variables to each dependent variable of the dataset and helps eliminating noise from the data (information that does not explain the dependent variables). Since models were generated using the 10-fold cross-validation technique, error estimates could be computed, and the number of latent variables could be selected by the one-sigma rule. According to that rule, the result of the Root Mean Square Error (RMSE) estimate for each latent variable is compared to the minimum error obtained; the minimum number of latent variables that yield an RMSE within one sigma of the minimum error estimate is selected (Friedman et al., 2001).

New models were generated taking into account the number of LVs selected in the previous step. The Akaike's Information Criterion (AIC) of each model was calculated and set as baseline for model comparison (Gujarati, 2009). Additionally, the significance Multivariate Correlation (sMC) for each independent variable was obtained and compared to the critical $F$ value; variables with sMC values lower than the critical $F$ were removed from the list of candidate variables in the final model. The Variable Importance in the Projection (VIP) was obtained for the remaining variables and those with value smaller than 1 were also removed (Tran et al., 2014). More on the AIC, sMC, and VIP can be found in section 3.4.

Since the goal was to understand the relationship between sociodemographic attributes and students' performance, and given the overall number of variables, the forward variable selection approach was implemented at this stage. In each step, the independent variables were added to the model one at a time. The order in which they entered the model was determined by the variable's sMC; those with higher sMC entered the model first. The criterion to keep the variable in the model was its contribution to the AIC. In case of a positive contribution, the variable was kept in the model; otherwise, the variable was discarded. Once all remaining variables were tested in the model and only those with a positive contribution were kept, the *p-values* of the regression coefficients were verified. Variables with non-significant coefficients (we used a 0.05 significance level) were also removed from the model to obtain the final model for each of the five dependent variables (Gujarati, 2009; Tran et al., 2014).

### 3.4. EVALUATION

The metrics used to build and evaluate the models were the Root Mean Square Error (RMSE), Akaike's Information Criterion (AIC), significance Multivariate Correlation (sMC), Variable Importance in the Projection (VIP) and the adjusted coefficient of determination $R^2$. RMSE gives the error generated by the model when it is used to predict new scores. In the context of this study, it was used in two moments: in the selection of the number of LVs after the complete model was generated as part of the one-sigma rule (Friedman et al., 2001) and in the evaluation of the final model. The RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{10}$$

where $y_i$ is an observed value, $\hat{y}_i$ is a corresponding predicted value and $n$ is the number of observations (Evans and Olson, 2003).

The Akaike Information Criterion (AIC) is used to compare models and its objective is to minimize the RMSE adjusting the result by the number of variables inserted in the model. AIC is a metric that allows maximizing the tradeoff between model complexity (i.e., number of variables) and its goodness-of-fit. The criterion has been used in this study to select the final models instead of the $R^2$ since parsimonious models are more desirable. When two models are compared by the AIC metric, the best model is the one with the lowest AIC value. The AIC score is calculated as follows:

$$AIC = e^{2k/n}\,\frac{SSE}{n} \tag{11}$$

where $k$ is the number of regressors, $n$ is the number of observations and SSE is the Sum of Squared Errors (Gujarati, 2009).

sMC and VIP are filtering methods for variable selection. VIP is one of the most popular; however, as mathematically proved by Tran et al. (2014), VIP might yield false positive or negative values in situations where there is excessive noise (irrelevant variation) in the data or when a large number of LVs are demanded. To overcome that, the sMC method was created. sMC uses predicted values as a new LV and the regression coefficients generated by the model. Differently from classic methods, sMC does not use the orthogonal variance decomposition, which is the portion that may contain the irrelevant information that wants to be avoided (Tran et al., 2014).

The cutoff value in the VIP method to retain a variable in the model is 1.0 (or higher). For the sMC, an $F$-test is used to select the variables. The degrees of freedom for the numerator is 1 and for the denominator is $(n-2)$, and $F(1-\alpha, n-2)$ is used in the statistical test. sMC and VIP are calculated as follows:

$$sMC_j = \frac{\left\| \dfrac{\left(\hat{\boldsymbol{y}}\,\hat{\boldsymbol{b}}'_{PLS_j}\right)}{\left\|\hat{\boldsymbol{b}}_{PLS_j}\right\|^2} \right\|^2}{\left\| \boldsymbol{x}_i - \dfrac{\left(\hat{\boldsymbol{y}}\,\hat{\boldsymbol{b}}'_{PLS_j}\right)}{\left\|\hat{\boldsymbol{b}}_{PLS_j}\right\|^2} \right\|^2 \Big/ (n-2)} \tag{12}$$

where $\hat{\boldsymbol{b}}_{PLS}$ is the PLS regression coefficient vector.

$$VIP_j = \sqrt{d \sum\nolimits_{k=1}^{h} \boldsymbol{s}_k^2 \boldsymbol{t}_k{}' \boldsymbol{t}_k \left(\boldsymbol{q}_{kj}\right)^2 \Big/ \sum\nolimits_{k=1}^{h} \boldsymbol{s}_k^2 \boldsymbol{t}_k{}' \boldsymbol{t}_k} \tag{13}$$

where $d$ is the number of variables, $h$ is the number of latent variables, $\boldsymbol{q}_{kj}$ is the covariance between $\boldsymbol{X}$ and $\boldsymbol{y}$ for each variable $\boldsymbol{j}$ and

$$s_k = \frac{\boldsymbol{t_k}'\boldsymbol{y}(\boldsymbol{k})}{\boldsymbol{t_k}'\boldsymbol{t_k}} \tag{14}$$

In this study, $\alpha$ was set at 0.01. Considering that all models were built with the same number of observations, the critical $F$ value used to evaluate variables in each model was 6.63 (Tran et al., 2014). Given the number of independent variables and LVs demanded by each model, we used a combination of these two methods (VIP and sMC), as exposed in section 3.3, which was shown to be efficient in the selection of variables.

Finally, the adjusted $R^2$ gives the proportion of the variance in the dataset explained by the model, penalizing for the number of regressors. Although not used here as a criterion in the development of the models, the adjusted $R^2$ is the standard metric in regression studies. The Adjusted $R^2$ is calculated as follows:

$$R^2Adj = 1 - \frac{SSE/(n-k)}{TSS/(n-1)} \tag{15}$$

where SSE is the Sum of Squared Errors, TSS is the Total Sum of Squares, $n$ is the number of observations and $k$ is the number of regressors (Gujarati, 2009).

## 4. RESULTS AND DISCUSSIONS

This section is organized in three subsections: in section 4.1, a brief descriptive analysis of the dataset is presented; in section 4.2, the partial results during the modeling process are presented; in section 4.3, final results are presented and discussed.

### 4.1. DESCRIPTIVE ANALYSIS

The portion of the dataset analyzed in this study has information on 448,949 students that live and took the exam in the South region of Brazil. Only students who provided answers to all four areas of knowledge (AK) and wrote the essay obtaining valid scores were considered. The present descriptive analysis was conducted based on graphs, normality tests, and the Mann-Whitney test. Results for all normality tests are reported in Appendix B; none of the variables tested could be described by a normal distribution. Hence, only the Mann-Whitney test for equality between groups within variables was used. Normality and hypothesis tests were only conducted on variables reported in the literature as potential predictors of students' performance: parent's years of education, type of school, sex, race, and age. Complete results are presented in Appendix B and Appendix C, respectively.

Ages in the complete sample ranged from 10 to 81 years, with an average of 20.94 years and a standard deviation (SD) of 6.57 years. The most frequent age in the dataset was 17 years old. The ratio between male and female students was 42:58. Figure 3 presents the distribution of students stratified by age and sex.

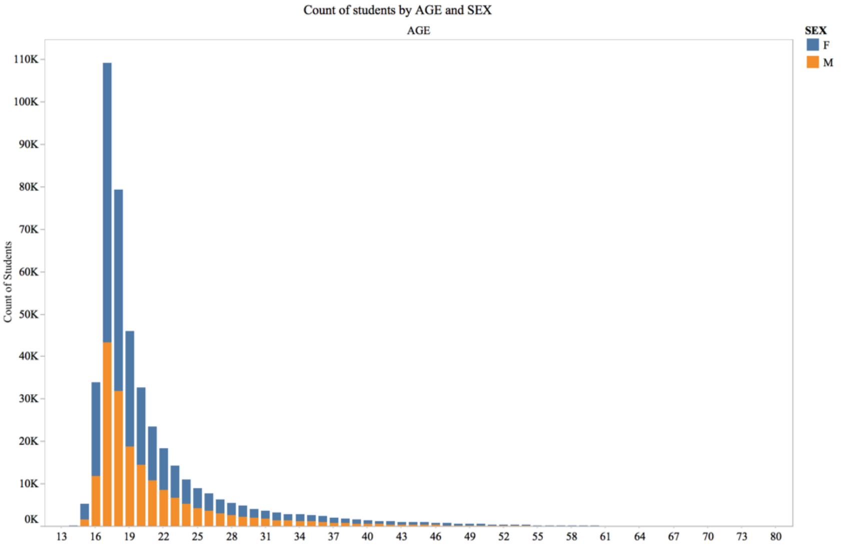

**Figure 3:** Count of students stratified by Age and Sex

From the average scores by age displayed in Figure 4, it can be seen that the effect of age on LC, HS and NS scores is much smaller than the effect observed on ES and MA scores. It is noteworthy that score behaviors for ES and MA are different between the ages of 16 and 25 (where 84.35% of the students are concentrated): ES average scores display a peak at the age of 17 while MA average scores are almost constant from ages 19 to 25.

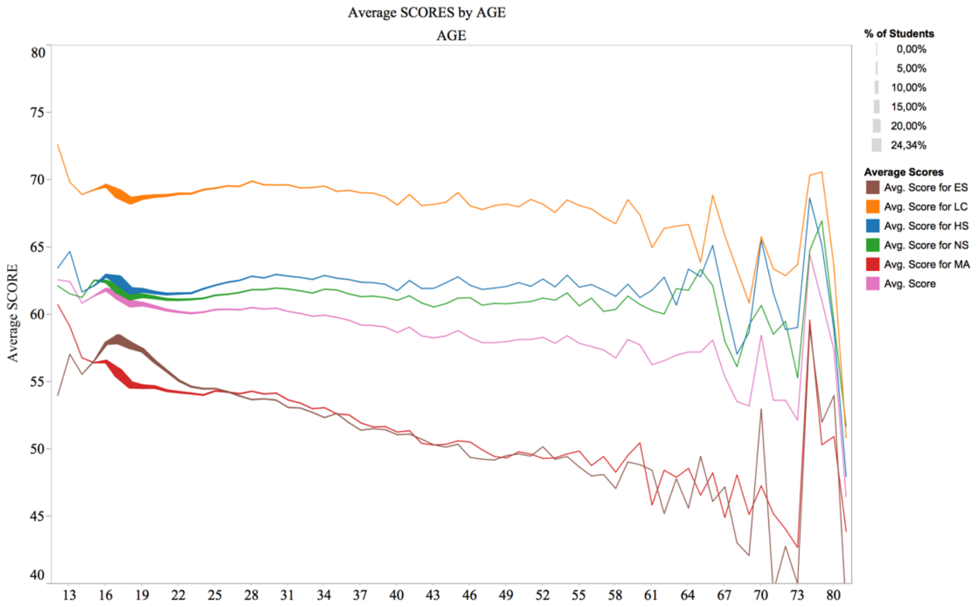

**Figure 4:** Average Scores by Age

The Mann-Whitney test results, as presented in Table 1, confirm at a 95% confidence level (which is the confidence level in all analyses to follow) that in the ES AK there are significant differences in score averages from ages 16 to 23. In the LC and MA AKs, the same occurs from ages 16 to 19; in the HS and NS areas, it occurs from ages 16 to 18. After these age intervals, the behavior of scores varies. However, when plotted against age intervals (see Figures 4 and 5) no noticeable trend appears in LC, HS, NS and MA average scores, which are rather stable. On the other hand, ES average scores tend to stabilize only from age 23 on.

Average scores for students of ages 16, 17 and 18 resulted significantly different for all areas of knowledge, indicating that age may be associated with performance in the range from 16 to 18 years old. However, it might not be that much relevant from the age of 19 years on.

**Table 1:** Mann-Whitney test results for Age

| | ES | | LC | | HS | | NS | | MA | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | *p-value* | Statistic | *p-value* | Statistic | *p-value* | Statistic | *p-value* | Statistic | *p-value* |
| 16 - 17 | -2.47 | 0.007 | -13.87 | 0.000 | -6.86 | 0.000 | -12.78 | 0.000 | -14.93 | 0.000 |
| 17 - 18 | -11.06 | 0.000 | -14.34 | 0.000 | -20.60 | 0.000 | -16.58 | 0.000 | -18.89 | 0.000 |
| 18 - 19 | -6.44 | 0.000 | -5.83 | 0.000 | **-0.21** | **0.416** | **-0.04** | **0.485** | -3.10 | 0.001 |
| 19 - 20 | -9.66 | 0.000 | **-1.61** | **0.054** | -2.85 | 0.002 | -2.31 | 0.011 | **-0.03** | **0.489** |
| 20 - 21 | -5.86 | 0.000 | **-0.52** | **0.303** | **-1.16** | **0.124** | **-1.19** | **0.117** | -2.33 | 0.010 |
| 21 - 22 | -5.30 | 0.000 | -1.84 | 0.033 | **-0.79** | **0.214** | **-0.55** | **0.290** | **-0.30** | **0.382** |
| 22 - 23 | -3.05 | 0.001 | **-0.16** | **0.436** | **-0.52** | **0.303** | **-0.88** | **0.188** | **-0.50** | **0.309** |
| 23 - 24 | **-1.17** | **0.120** | -2.96 | 0.002 | -3.03 | 0.001 | **-0.84** | **0.201** | **-0.95** | **0.171** |
| 24 - 25 | **-0.39** | **0.347** | **-0.70** | **0.243** | -1.70 | 0.045 | -2.16 | 0.015 | -2.07 | 0.019 |

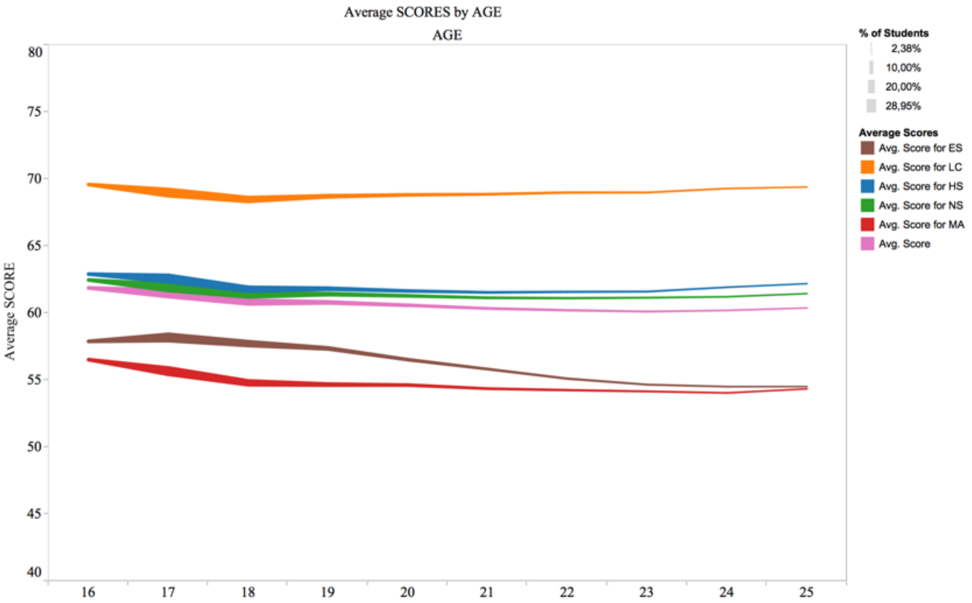

**Figure 5:** Average Scores by Age from ages 16 to 25

When average scores are stratified by race (Figure 6), it is noticeable that among the students who declared their race, Group D (Yellow) had the best average score, while Group E (Indigenous) had the worst. However, these groups represent only 1.24% and 0.19% of the sample, respectively. Group C (White) represents 74.06% of the students and had the second highest average scores. Seventeen percent of the students belong to Group B (Mixed) which had an intermediate performance compared to other groups. Group A (Black) represents 5.35% of the students and had an average score worse than every other group, except for Group E.

Table 2 presents the Mann-Whitney test results for race. It can be observed that among the students who declared their race, the score averages are not significantly different only between Groups A (Black) and E (Indigenous) in the NS and MA AKs, and between Groups C (White) and D (Yellow) in the ES AK. However, as mentioned, these two groups (D and E) together represent only 1.43% of the students. It is noteworthy that the majority of the test results yielded significant differences between the average scores.

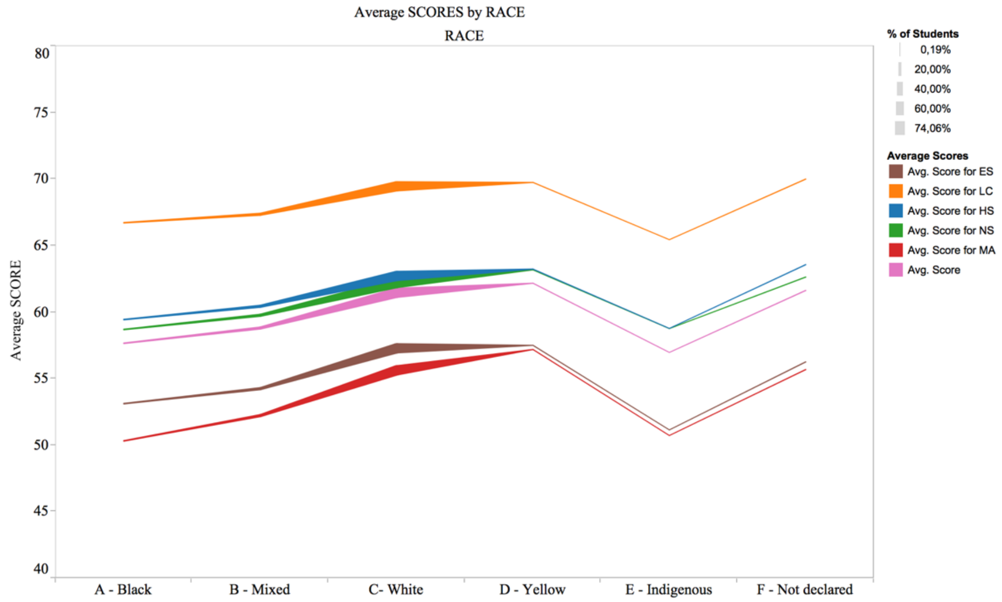

**Figure 6:** Average Scores stratified by Race

**Table 2:** Mann-Whitney test results for Race

| | ES | | LC | | HS | | NS | | MA | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | *p-value* | Statistic | *p-value* | Statistic | *p-value* | Statistic | *p-value* | Statistic | *p-value* |
| A - B | -12.41 | 0.000 | -10.77 | 0.000 | -15.35 | 0.000 | -17.55 | 0.000 | -25.98 | 0.000 |
| A - C | -50.57 | 0.000 | -53.06 | 0.000 | -55.75 | 0.000 | -61.75 | 0.000 | -73.92 | 0.000 |
| A - D | -22.76 | 0.000 | -25.48 | 0.000 | -28.10 | 0.000 | -32.41 | 0.000 | -39.00 | 0.000 |
| A - E | -5.15 | 0.000 | -4.69 | 0.000 | -2.16 | 0.015 | **-0.33** | **0.370** | **-1.31** | **0.094** |
| A - F | -20.34 | 0.000 | -32.94 | 0.000 | -35.83 | 0.000 | -35.87 | 0.000 | -37.80 | 0.000 |
| B - C | -61.99 | 0.000 | -69.16 | 0.000 | -65.69 | 0.000 | -71.42 | 0.000 | -78.15 | 0.000 |
| B - D | -18.05 | 0.000 | -21.80 | 0.000 | -22.48 | 0.000 | -26.15 | 0.000 | -29.49 | 0.000 |
| B - E | -7.71 | 0.000 | -6.95 | 0.000 | -5.47 | 0.000 | -3.45 | 0.000 | -4.24 | 0.000 |
| B - F | -14.53 | 0.000 | -29.81 | 0.000 | -30.36 | 0.000 | -28.80 | 0.000 | -26.08 | 0.000 |
| C - D | **-0.77** | **0.219** | -2.62 | 0.004 | -4.49 | 0.000 | -7.08 | 0.000 | -8.43 | 0.000 |
| C - E | -14.41 | 0.000 | -14.54 | 0.000 | -12.98 | 0.000 | -11.73 | 0.000 | -13.13 | 0.000 |
| C - F | -7.25 | 0.000 | -7.18 | 0.000 | -8.80 | 0.000 | -4.79 | 0.000 | **-0.49** | **0.309** |
| D - E | -13.38 | 0.000 | -14.11 | 0.000 | -13.26 | 0.000 | -12.81 | 0.000 | -14.65 | 0.000 |
| D - F | -5.11 | 0.000 | -2.44 | 0.007 | -2.01 | 0.022 | -2.62 | 0.004 | -6.76 | 0.000 |
| E - F | -11.66 | 0.000 | -15.32 | 0.000 | -14.51 | 0.000 | -12.34 | 0.000 | -12.09 | 0.000 |

**Codes:** A = Black; B = Mixed; C = White; D = Yellow; E = Indigenous; F = Not declared

Regarding the type of school attended, 77.65% of the students attended only public schools and had the worst average scores in every AK. Sixteen percent of the students attended only private schools either with scholarships (Group E) or without it (Group D) and were the ones with highest average scores. Groups B (attended both public and private – with scholarship) and C (attended both public and private – without scholarship) represent 6.00% of the students and have an intermediate performance compared to the other groups. Again, those without scholarships performed better, as shown in Figure 7. These visual conclusions are confirmed by Mann-Whitney test results: there are significant differences in average scores from all AK and all groups.

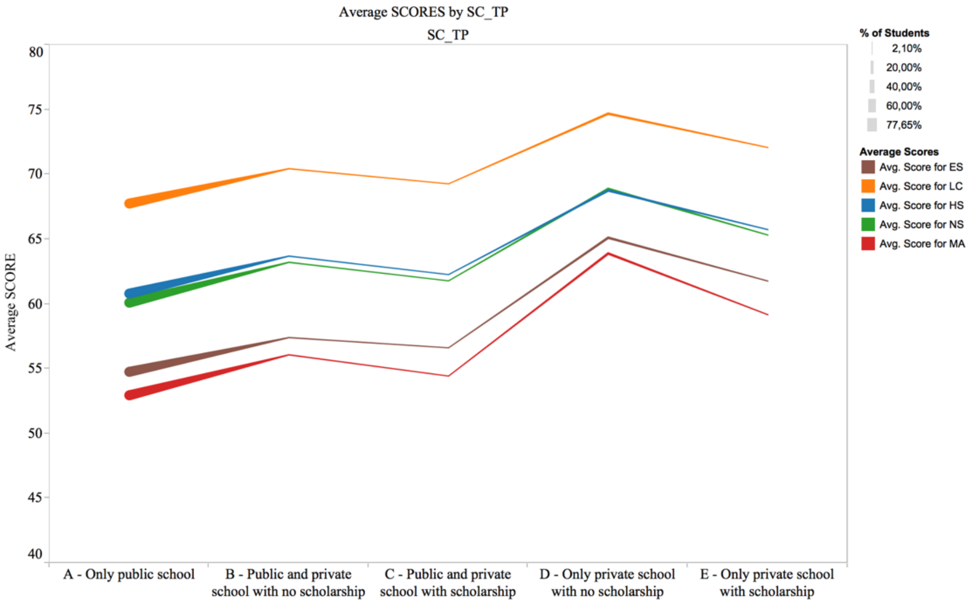

**Figure 7:** Average Scores stratified by Type of School

The percentage of students' parents who completed primary education was 17.69% (father) and 15.57% (mother). In addition, 15.31% of the fathers and 15.59% of the mothers completed only the lower secondary education, which corresponds to 9 years of formal mandatory education, while 30.62% of fathers and 32.37% of mothers completed the upper secondary education. On the other hand, 15.91% (20.87%) of the students' fathers (mothers) graduated from a university or had post-graduate degrees. Lastly, 20.47% of the fathers did not complete primary education or never studied, in contrast to 15.59% of the mothers.

As depicted in Figures 8 and 9, the larger the parents' number of schooling years, the better the average students' performances. For those whose parents did not complete primary education or never studied, the most affected areas of knowledge were ES and MA. There are no major differences between graphs for fathers and mothers. Visual conclusions are again corroborated by Mann-Whitney test results: there are significant differences in average scores from all AK and all groups, for both fathers and mothers.

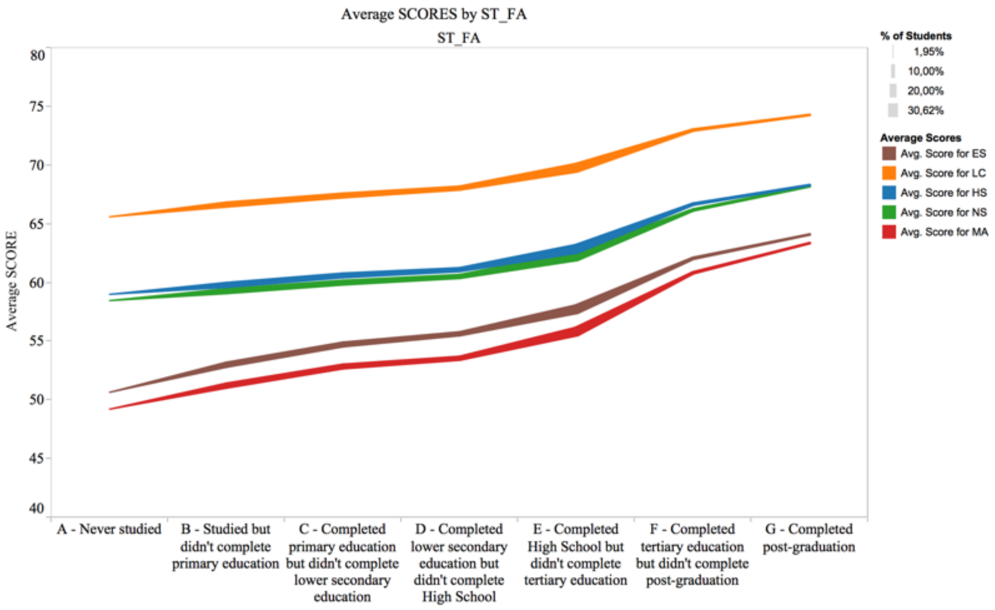

**Figure 8:** Average Scores stratified by years of schooling of student's father

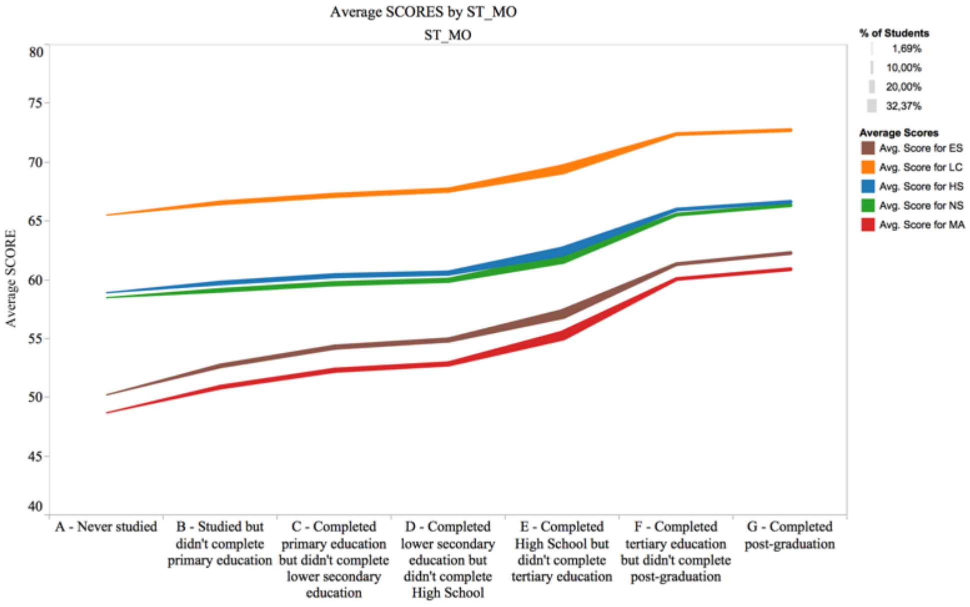

**Figure 9:** Average Scores stratified by years of schooling of student's mother

In addition to the main variables presented above, some other variables are also worth an overview. More than half of the students in the sample (53.71%) already finished High School (HS) by the time the exam was taken, and 34.88% of the students were in senior HS year. The remaining 11.41% of students either still had more than one year left to complete HS or were not attending school at all and never completed HS. The exam also requires students to choose between English and Spanish as foreign language; 49.58% of them chose Spanish and 50.42% chose English.

The most common numbers of persons per households are 4 (32.99%), 3 (29.26%), 2 (14.64%) or 5 (13.20%); the most common household income is from R$ 0,00 to R$ 1,405.50 (32.12%) followed by an income from R$ 1,405.51 to R$ 1,874.00 (13.00%). More than half of the households own one or more cars (69.25%), while 77.95% have one or more computers. Internet connection is available in 83.83% of the households.

Figure 10 presents score frequencies in each area of knowledge stratified by sex. Essay scores display the largest variability, which is similar for both sexes. The distribution of LC scores is the one closest to a normal distribution, in opposition to the MA scores, with a flat distribution. It is interesting to observe that males tend to score better in the HS, NS and MA areas of knowledge, with averages scores of 63.23, 62.77 and 57.24 versus 61.37, 60.73 and 52.96 of females, respectively. Females, on the other hand, perform better on the essay, with an average score of 57.73 versus 54.79 from males. LC is the area of knowledge with smallest differences across sexes, with females averaging 68.96 and males 68.93. According to the Mann-Whitney test results, average scores are significantly different between males and females for all AKs.

Finally, considering the complete dataset of students who live and took the exam in the South region and obtained valid scores, LC and HS questions yielded an average score of 68.95 (SD = 7.79) and 62.15 (SD = 8.80), respectively; the average score obtained in the essay was 56.50 (SD = 12.06). NS and MA questions were presented on the second day of exams and yielded average scores of 61.59 (SD = 8.39) and 54.75 (SD = 10.69), respectively.

## 4.2. MODELING

Following the method presented in section 3.3, the first set of models were generated and the appropriate number of LVs for each dependent variable were selected. Figure 11 presents the selection according to the one-sigma rule already described. The number of LVs selected for each dependent variable were 10 for ES, LC, HS, NS and 11 for MA.

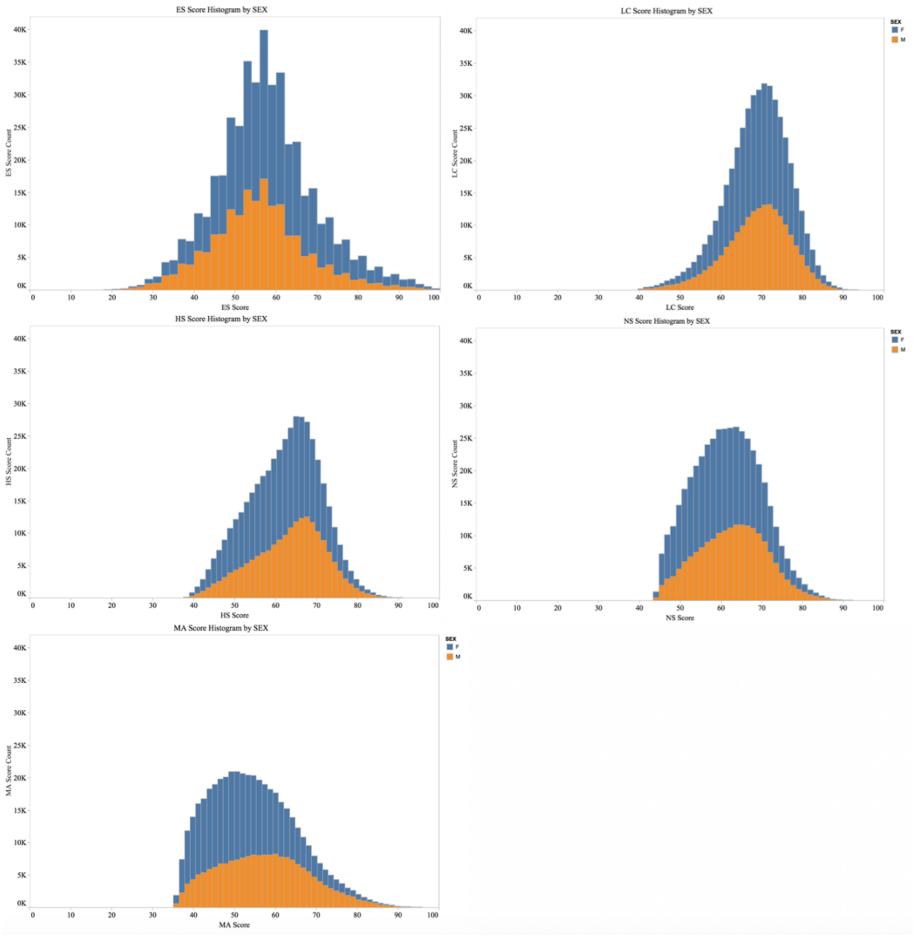

**Figure 10:** Histograms of Scores stratified by Sex

After selecting the proper number of LVs, new models were generated, and their corresponding AICs calculated to be used as baseline. AIC results for the complete model were 140.75 for ES, 53.04 for LC, 65.77 for HS, 67.41 for NS and 111.72 for MA. From the new models, sMC values were also calculated for each independent variable. A critical ***F*** value of 6.63 was also calculated. sMC values for each independent variable were compared to the critical ***F***; variables that did not meet the test criterion were discarded. For the remaining variables, VIP values were obtained; variables with VIP values less than 1 were also removed. The sMC and VIP values for all variables can be found in Appendix D.

To guarantee that only variables that improve the model were retained the forward variable selection approach was adopted at this stage. At each iteration step, one independent variable was added to the model (from largest to smallest sMC value) and the AIC of the model generated was calculated. Variables with positive contribution to AIC were retained in the model. Table 3 presents information on the progression of forward variable selection iterations for each model obtained.

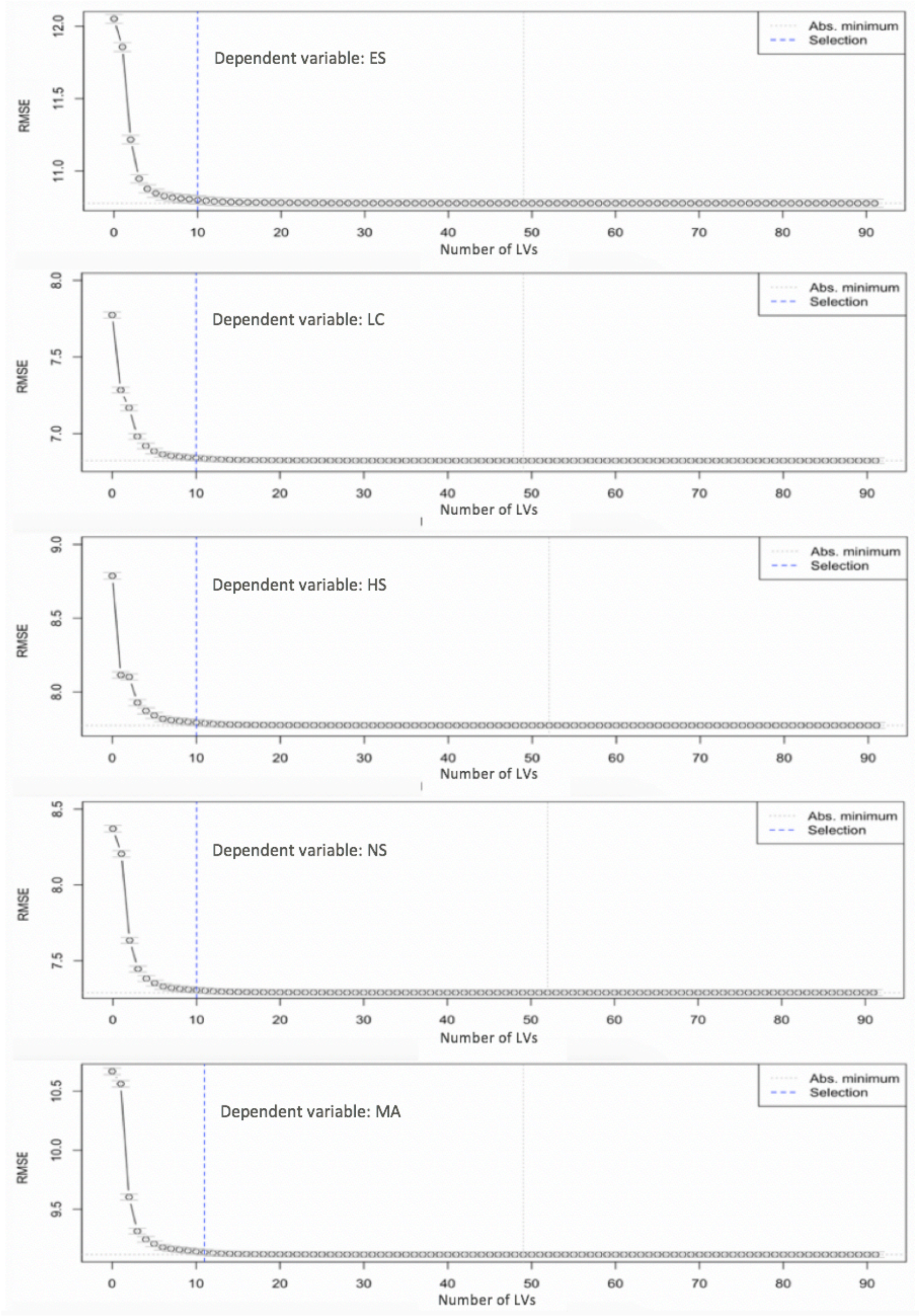

**Figure 11:** Selection of LVs for each dependent variable

### 4.3. FINAL MODELS AND COEFFICIENT ANALYSES

From Table 3 it was possible to identify independent variables to be used in the final model for each dependent variable. Performance statistics for the final models were obtained using the testing portion of the dataset. Table 4 presents results obtained modeling the training and testing portions of the dataset.

**Table 3:** Progression of forward variable selection iterations for dependent variables ES, LC, and HS

| | Dependent variable: ES | | Dependent variable: LC | | Dependent variable: HS | |
|---|---|---|---|---|---|---|
| Order | Variable | AIC | Variable | AIC | Variable | AIC |
| Complete model | All | 140.75 | All | 53.04 | All | 65.77 |
| 1 | SC_TP_D | 133.56 | SC_TP_D | 55.27 | SC_TP_D | 70.41 |
| 2 | ST_FA_G | 132.83 | ST_FA_G | 54.88 | ST_FA_G | 69.87 |
| 3 | ST_MO_F | 132.17 | OC_FA_E | 54.73 | ST_MO_F | 69.43 |
| 4 | OC_FA_E | 131.95 | ST_FA_F | 54.29 | ST_MO_G | 68.80 |
| 5 | ST_MO_G | 131.01 | ST_MO_F | 54.06 | ST_FA_F | 68.36 |
| 6 | ST_FA_F | 130.55 | OC_FA_D | 53.31 | LG_E | 67.44 |
| 7 | SEX_M | 128.22 | ST_MO_G | 53.04 | OC_FA_E | 67.23 |
| 8 | OC_FA_D | 126.85 | HS_ST_C | 52.38 | RACE_C | 66.97 |
| 9 | LG_E | 126.57 | LG_E | 51.42 | SEX_M | 66.55 |
| 10 | PC_Y | 125.04 | RACE_C | 51.26 | PC_Y | 65.52 |
| 11 | RACE_C | 124.78 | PC_Y | 50.72 | **OC_FA_D** | **65.07** |
| 12 | FZR_Y | 124.65 | CELL_B | 50.65 | OC_MO_B† | 65.26 |
| 13 | INT_Y | 124.61 | FZR_Y† | 50.88 | FZR_Y† | 65.21 |
| 14 | BATH_B† | 125.04 | OC_MO_D† | 50.79 | VA_CL_Y† | 65.34 |
| 15 | CELL_B† | 124.65 | CAR_Y† | 51.08 | OC_MO_D† | 65.20 |
| 16 | HK_Y | 124.55 | HK_Y† | 50.68 | BATH_B† | 65.28 |
| 17 | OC_MO_D† | 124.64 | TP_Y† | 50.87 | CAR_Y† | 65.47 |
| 18 | CAR_Y† | 125.35 | VA_CL_Y† | 50.99 | MW_Y† | 65.30 |
| 19 | OC_MO_B† | 125.10 | BATH_B† | 51.04 | CB_TV_Y† | 65.32 |
| 20 | **ST_MO_B** | **124.36** | ST_FA_B† | 50.67 | INT_Y† | 65.14 |
| 21 | VA_CL_Y† | 125.27 | TV_B† | 50.99 | ST_FA_B† | 65.15 |
| 22 | - | - | **ST_MO_B** | **50.62** | TP_Y† | 65.40 |
| 23 | - | - | INT_Y† | 50.74 | TV_B† | 65.50 |
| Final model | | 124.36 | | 50.62 | | 65.07 |

† Variables that were tested but were not retained in the model

**cont. Table 3:** Progression of forward variable selection iterations for dependent variables NS and MA

| | Dependent variable: NS | | Dependent variable: MA | |
|---|---|---|---|---|
| Order | Variable | AIC | Variable | AIC |
| Complete model | All | 67.41 | All | 111.72 |
| 1 | SC_TP_D | 61.80 | SC_TP_D | 100.77 |
| 2 | ST_FA_G | 61.21 | ST_FA_G | 99.64 |
| 3 | ST_MO_F | 60.80 | ST_MO_F | 98.78 |
| 4 | ST_FA_F | 60.36 | ST_MO_G | 97.58 |
| 5 | ST_MO_G | 59.78 | ST_FA_F | 96.88 |
| 6 | LG_E† | 59.81 | SEX_M | 93.64 |
| 7 | OC_FA_E† | 59.83 | RACE_C | 92.82 |
| 8 | SEX_M | 59.19 | LG_E | 92.21 |
| 9 | RACE_C | 59.01 | OC_FA_E | 91.73 |
| 10 | OC_MO_B† | 59.76 | HK_Y | 91.51 |
| 11 | PC_Y | 58.31 | FZR_Y | 91.06 |
| 12 | OC_FA_D | 58.14 | OC_FA_D | 90.32 |
| 13 | FZR_Y† | 58.49 | **PC_Y** | **89.65** |
| 14 | VA_CL_Y† | 58.55 | OC_MO_B† | 90.05 |
| 15 | BATH_B† | 58.37 | CELL_B† | 89.71 |
| 16 | OC_MO_D† | 58.63 | VA_CL_Y† | 90.30 |
| 17 | ST_FA_B† | 58.26 | BATH_B† | 90.04 |
| 18 | CB_TV_Y† | 58.58 | OC_MO_D† | 89.88 |
| 19 | **HK_Y** | **58.02** | CB_TV_Y† | 90.24 |
| 20 | MW_Y† | 58.34 | - | - |
| 21 | OC_FA_C† | 58.42 | - | - |
| 22 | TP_Y† | 58.50 | - | - |
| 23 | DR_MA_Y† | 58.31 | - | - |
| 24 | TV_B† | 58.71 | - | - |
| Final model | | 58.02 | | 89.65 |

† Variables that were tested but were not retained in the model

**Table 4:** $R^2$ Adjusted and RMSE results

| | | ES | LC | HS | NS | MA |
|---|---|---|---|---|---|---|
| Training | $R^2$ Adj | 0.1562 | 0.1710 | 0.1630 | 0.1806 | 0.2246 |
| | RMSE | 11.070 | 7.079 | 8.039 | 7.577 | 9.395 |
| Testing | RMSE | 11.168 | 7.121 | 8.064 | 7.627 | 9.490 |

The variance explained by the models as given by their adjusted $R^2$ values vary between 14% and 23%. Given the variables considered in this study, these values are aligned with the literature. White (1982) and Sirin (2005) presented a comprehensive meta-analysis on studies associating sociodemographic variables and students' performance and reported an average variance explained of 10%. The only other study using the ENEM dataset available in the literature reported $R^2$ values of 0.35 and 0.18 for the two methods applied (Gradient Boosting and AdaBoost, respectively; Stearns et al., 2017). Our $R^2$ values also indicate that students' performance in ENEM may be explained by variables not available in the test's dataset. Other studies on different test results investigate variables such as family environment, school structure, students' health, effort and psychological factors, among others, which are not available in the ENEM questionnaire (McLoyd, 1998; Menezes-Filho, 2007; Stewart, 2008; de Oliveira Barbosa, 2009; Sampaio and Guimarães, 2009; Perry and McConney, 2010; De Lange et al., 2014).

However, results from this study support several relevant conclusions; variables retained provide information related to the type of school attended by the student, fathers' and mothers' years of schooling, fathers' occupation, foreign language chosen by the student, sex, and race of the students and families' possessions.

### 4.3.1. ESSAY

The model generated for the dependent variable *Essay Score* comprises 14 independent variables, with an adjusted $R^2$ of 0.1562 and an RMSE of 11.070. The AIC of the final model is 124.36; the complete model with 91 independent variables yielded an AIC value of 140.75. Table 5 presents the coefficients for the retained variables as well as their Standard Errors (SEs) and significance.

**Table 5:** Coefficients for the ES final model

| Variable | Estimate (SE) signif |
|---|---|
| SEX_M | -3.545 (0.037) *** |
| RACE_C | 1.076 (0.076) *** |
| LG_E | 1.716 (0.035) *** |
| SC_TP_D | 5.262 (0.101) *** |
| ST_FA_F | 1.217 (0.107) *** |
| ST_FA_G | 1.322 (0.147) *** |
| ST_MO_B | -1.420 (0.078) *** |
| ST_MO_F | 1.963 (0.140) *** |
| ST_MO_G | 2.149 (0.086) *** |
| OC_FA_D | 2.230 (0.085) *** |
| OC_FA_E | 2.679 (0.156) *** |
| PC_Y | 2.183 (0.053) *** |
| INT_Y | 1.499 (0.062) *** |
| FZR_Y | 1.455 (0.042) *** |

Significance codes: '***' 0.001 '**' 0.01

Variables related to sex, race, foreign language, type of school, father's and mother's years of schooling, father's occupation, and family's possessions were retained in the model. These variables are explored in the following sections.

#### *4.3.1.1. TYPE OF SCHOOL*

The variable Type of School (SC_TP), represented by one of its five categories, the dummy variable SC_TP_D (attended only private school with no scholarship), is the one with the highest coefficient in the Essay model. It is also the variable that yielded the largest VIP value. All remaining categories of SC_TP were not included in the model. Furthermore, as shown in Table 3, it was the first variable to enter the model. These aspects indicate the variable's importance in explaining the students' performance in the Essay. Attending only private schools with no scholarship (13.55% of the sample) is a condition that increases the score in the Essay by 5.262 points. All remaining school attendance conditions have no impact on the score.

Sampaio and Guimarães (2009) analyzed students who took private admission exams to enter tertiary education and concluded that there is a significant difference between the performance of students from public and private schools in Brazil, which corroborate our findings. They present two main reasons for that. First, there is what they denote by school efficiency, which is the school's capacity to develop each student to her maximum potential. Second, there is the students' entry level, which represents all knowledge acquired before entering the school. They conclude that public schools usually present lower efficiency rates and receive students with lower entry levels.

A study conducted by Silva and Araújo (2009) reveals a difference of perception in students from private and public schools regarding what is the Essay's most important aspect. To private school students, grammar is the most important aspect, while to students from public schools the essay's theme is more important. These findings are not necessarily linked to the difference in their performance; however, being aware of these differences in perception could subsidize new studies on the subject.

In the ENEM dataset there are no variables related to high school efficiency or to aspects of the Essay deemed as most important by students. In case these effects exist, they are aggregated in the type of school variable. Regarding the entry level of each student, other variables, such as those related to students' socioeconomic status, could be related to that.

#### *4.3.1.2. SOCIOECONOMIC STATUS*

The Essay model comprises 6 categorical variables that may be viewed as socioeconomic status dimensions. They are: student's father years of schooling (ST_FA), student's mother years of schooling (ST_MO), student's father occupation (OC_FA), if the student owns a computer (PC), a freezer (FZR) and has internet access (INT) at home. These variables are represented by 11 dummy variables: student's father completed tertiary education but did not complete graduate studies (ST_FA_F), student's father completed graduate studies (ST_FA_G), student's mother went to school but did not complete primary education (ST_MO_B), student's mother completed tertiary education but did not complete graduate studies (ST_MO_F), student's mother completed graduate studies (ST_MO_G), student's father occupation is of high intellectual level (OC_FA_D), student's father occupation is of highest intellectual level (OC_FA_E), student owns a computer (PC_Y) and a freezer (FZR_Y) at home, as well as access to internet service (INT_Y) and the English foreign language (LG_E).

From these 11 variables, OC_FA_E is the one with the highest coefficient, followed by OC_FA_D. In 2017, 7.26% of the students belonged to the category described by OC_FA_E and 21.64% belong to the category described by OC_FA_D. Students with these characteristics had a score advantage of more than 2.2 points over the remaining 71.10% whose fathers have occupations of lower intellectual levels.

Groups represented by variables ST_MO_B (students whose mothers did not complete elementary school), ST_MO_F (students whose mothers completed tertiary studies) and ST_MO_G (students whose mothers completed graduate studies) comprise 13.58%, 10.07% and 10.33% of the sample, respectively. Considering as baseline the remaining categories which did not enter the model, groups ST_MO_F and ST_MO_G yielded an advantage of approximately 2 points in the final score. In opposition, students from group ST_MO_B lost around 1.4 points in score.

Regarding years of education of students' fathers, 8.85% of the students are from group ST_FA_F (students whose fathers completed tertiary studies) and 5.93% are from group ST_FA_G (students whose fathers completed graduate studies). These students' scores are slightly over 1.2 points higher than the rest of the sample, whose fathers' years of education did not impact on Essay scores.

Regarding variables associated with possessions, the highest coefficient is shown by the one describing the ownership of a computer at home (77.95% of the sample) which adds around 2 points to the final student's score. Additionally, owning a freezer at home (54.69% of the sample) and having access to internet service (83.84% of the sample) add over 1.4 points each to the final score.

Lastly, variable LG_E reports a better performance of students who chose English (50.42% of the sample) as foreign language instead of Spanish. This choice represented an increase of around 1.7 points in the final Essay score.

Many studies have corroborated the fact that socioeconomic status has an effect on students' performance; e.g. Stewart (2008), Lafontaine and Monseur (2009), Tucker-Drob (2013) and Hair et al. (2015). Sirin (2005) reported that variables related to parents' occupation and family's possessions are the most correlated with performance, although parent's years of schooling and family's income are also listed as important variables. Apart from the fact that income was not included in our model, remaining variables are aligned with Sirin's (2005) findings.

Coleman (1988) proposes that student's performance in school is related to an aggregate variable named "family background", comprised of three components: financial capital, human capital and social capital. Financial capital comprises aspects such as infrastructure available for studying, materials and resources that ease the family's life. Human capital refers to parents' education, which contributes to the environment where the student is raised and her potential for cognitive development. Social capital is a more abstract concept that is related to the effort and time spent by parents to transmit to students aspects such as knowledge, confidence, and empowerment.

The inclusion of the foreign language chosen by the student as a variable in the model could also be related to "family background" as defined by Coleman (1988) since the Essay does not evaluate the knowledge on the foreign language. Such conclusion is corroborated by the fact that students who chose English tend to come from higher socioeconomic levels. Regarding parents' occupation, 38.44% and 37.19% of fathers and mothers of students who chose English have high intellectual demanding occupations, while in the complete sample these values are 28.90% and 28.65%, respectively. The same happens to parents' years of schooling: 53.57% of fathers and 61.75% of mothers of those who chose English completed at least High School, in contrast to the complete sample, in which only 43.27% of fathers and 52.03% of mothers displayed that condition. Regarding the income, 53.45% of families from students who chose English earned more than R$ 2,342.50 monthly, while in the complete sample only 43.20% receive the same amount.

Comparing Coleman's (1988) proposition with our results, variables included in the Essay model could be classified as related to financial and human capital, and their inclusion is coherent with the real importance of these aspects. There are no variables in the ENEM dataset related to social capital aspects.

#### *4.3.1.3. SEX AND RACE*

In 2017, 41.90% of the students who took the ENEM test in the South region were males. The Essay model indicates that male students present an average score 3.545 points lower than that of female students.

Steinmayr and Spinath (2008) claim that differences in performance between male and female individuals exist but are due to behavioral factors. To Lafontaine and Monseur (2009), the more cognitively demanding the task is, the larger the female advantage in performance. They found the larger gap between male and female performances in open-ended questions (equivalent to an essay), which is consistent with findings in this study, considering the Essay is the only portion of the exam that does not follow the multiple-choice standard. The inclusion of the male sex dummy variable in the model and its negative coefficient is also aligned with results from the Mann-Whitney test, in which the average score for males and females appeared as significantly different, with female students scoring higher than male students.

Regarding race, in 2017 74.06% of the students declared themselves as white. The variable RACE_C is the dummy for RACE = White, and was included in the Essay model; all remaining dummies derived from Race were not included. According to that variable's coefficient, white students scored over 1 point higher in the essay compared to students of other races. Several studies address the issue of race, especially in Brazil, and there is a consensus regarding the variable's effect on students' performance (Fernandes, 2004; Sirin, 2005; Menezes-Filho, 2007; Stewart, 2008; de Oliveira Barbosa, 2009). Components of this effect could be related to socioeconomic factors as non-white students are more prone to attend worse schools and come from families of low income, with parents that are less educated (Sirin, 2005). That is also related to the financial and human capital dimensions described by Coleman (1988).

### 4.3.2. LANGUAGES AND CODES

The model generated for the dependent variable *Languages and Codes Score* comprises 13 independent variables, with an adjusted $R^2$ of 0.1710 and an RMSE of 7.079. The AIC of the final model is 50.62. The complete model comprised of 91 independent variables yielded an AIC value of 53.04. Table 6 presents the coefficients for the retained variables as well as their SEs and significance.

**Table 6:** Coefficients for the LC final model

| Variable | Estimate (SE) signif |
|---|---|
| RACE_C | 0.735 (0.043) *** |
| LG_E | 2.178 (0.047) *** |
| SC_TP_D | 3.464 (0.055) *** |
| HS_ST_C | -1.921 (0.053) *** |
| ST_FA_F | 1.021 (0.087) *** |
| ST_FA_G | 1.076 (0.092) *** |
| ST_MO_B | -0.740 (0.049) *** |
| ST_MO_F | 1.396 (0.059) *** |
| ST_MO_G | 1.231 (0.065) *** |
| OC_FA_D | 1.702 (0.045) *** |
| OC_FA_E | 1.846 (0.088) *** |
| CELL_B | 0.856 (0.039) *** |
| PC_Y | 1.647 (0.057) *** |

Significance codes: '***' 0.001 '**' 0.01

Variables related to race, foreign language, type of school attended, year the student will complete high school, father's and mother's years of schooling, father's occupation, and family's possessions were retained in the model. These variables are explored in the following sections.

#### *4.3.2.1. TYPE OF SCHOOL*

The variable Type of School (SC_TP), represented by one of its five categories (the dummy variable SC_TP_D = attended only private schools with no scholarship), is the one with the highest coefficient in the model. All remaining categories of SC_TP were not included in the model.

Attending only private schools with no scholarship is a condition that increases the score in the Languages and Codes AK by 3.464 points. All remaining school attendance conditions have no impact on the score. The portion of students who attended only private schools with no scholarship in the sample is 13.55%. As mentioned in section 4.3.1.1, the higher performance of private school students may be related either to school efficiency and/or to students' entry level (Sampaio and Guimarães, 2009). However, the ENEM dataset does not measure directly those factors and we cannot test any hypotheses related to them.

#### *4.3.2.2. SOCIOECONOMIC STATUS*

The Language and Codes model comprises 6 variables that can be considered socioeconomic status dimensions. They are: student's father years of schooling (ST_FA), student's mother years of schooling (ST_MO), student's father occupation (OC_FA), if the student owns a computer (PC) and two or more mobile phones (CELL) at home, and foreign language (LG). These variables are represented in the model by 10 dummy variables: student's father completed tertiary education but did not complete graduate studies (ST_FA_F), student's father completed graduate studies (ST_FA_G), student's mother went to school but did not complete primary education (ST_MO_B), student's mother completed tertiary education but did not complete graduate studies (ST_MO_F), student's mother completed graduate studies (ST_MO_G), student's father occupation is of high intellectual level (OC_FA_D), student's father occupation is of highest intellectual level (OC_FA_E), student owns a computer (PC_Y) and two or more mobile phones (CELL_B) at home, and English chosen as foreign language (LG_E).

Foreign language is the socioeconomic-related variable yielding the highest coefficient; the score in the LC AK increases by 2.178 points for students who chose English as foreign language (50.42% of the sample). It is important to remark that this AK directly includes foreign language questions (unlike the other AKs, which does not test the knowledge of English or Spanish).

For the remaining socioeconomic variables, OC_FA_E is the one with the highest coefficient, followed by OC_FA_D; 7.26% and 21.64% of students in the sample were represented by these conditions, which are related to fathers' occupations. Students whose fathers work in jobs that are intellectually more demanding perform better in the LC AK by around 1.8 points.

Variables related to father and mother years of schooling were also included in the model. Students whose mothers did not complete elementary school (ST_MO_B) perform worse in this AK by 0.74 points; they represent 13.58% of the sample. In opposition, students whose mothers completed tertiary studies (ST_MO_F), representing 10.07% of the sample, or completed graduate studies (ST_MO_G), representing 10.33% of the sample, score higher in this AK by around 1.3 points. The same analysis applies to students whose fathers completed tertiary studies (ST_FA_F), representing 8.85% of the sample, or completed graduate studies (ST_FA_G), representing 5.93% of the sample: these students score over 1 point higher than the remaining 85.19% of students whose fathers are in other schooling conditions.

Regarding variables associated with possessions, the highest coefficient is associated with the variable PC_Y, which indicates students who own a computer at home; this condition increases the score in the LC AK by over 1.6 points. The second variable related to possessions included in the model is CELL_B, denoting the condition of students with two or more cell phones in the household; these students score 0.856 points higher in this AK.

These results are also consistent with those available in the literature, as exposed in section 4.3.1.2, in which socioeconomic status is found to be related to student's performance, especially parent's years of schooling and parent's occupation, as well as possessions. They may be related to the concept of "family background" described by Coleman (1988).

#### *4.3.2.3. RACE*

As mentioned in section 4.3.1.3, there is a consensus with respect to the influence of race on students' performance. For the LC AK, Race appears in the model represented by one of its six categories (the dummy variable RACE_C = White), which represents 74.06% of the students in the sample. White students score almost 0.8 points higher than students of other races. Although statistically significant, the coefficient associated with RACE_C is the smallest in the model. As previously exposed (section 4.3.1.3), the superior performance of students may be related to the quality of school they attended to and the financial, human and social capital of their families (Coleman, 1988; Sirin, 2005).

The inclusion of the white race variable in the model is also aligned with results from the Mann-Whitney test, in which the average score of students who declared black and mixed races appeared as significantly different from those who declared white race, with the latter scoring higher. These three groups represent 96.69% of the students.

### 4.3.3. HUMAN SCIENCES

The model generated for the dependent variable *Human Sciences Score* comprises 11 independent variables, with an adjusted $R^2$ of 0.1630 and an RMSE of 8.039. The AIC of the final model is 65.07. The complete model (comprised of 91 variables) yielded an AIC value of 65.77. Table 7 presents the coefficients for the retained variables as well as their SEs and significances.

**Table 7:** Coefficients for the HS final model

| Variable | Estimate (SE) signif |
|---|---:|
| SEX_M | 1.295 (0.047) *** |
| RACE_C | 0.880 (0.041) *** |
| LG_E | 2.205 (0.036) *** |
| SC_TP_D | 4.051 (0.080) *** |
| ST_FA_F | 1.240 (0.077) *** |
| ST_FA_G | 1.434 (0.090) *** |
| ST_MO_F | 1.568 (0.067) *** |
| ST_MO_G | 1.626 (0.069) *** |
| OC_FA_D | 1.658 (0.055) *** |
| OC_FA_E | 1.862 (0.106) *** |
| PC_Y | 2.311 (0.049) *** |

Significance codes: ‘***’ 0.001 ‘**’ 0.01

Variables related to sex, race, foreign language, type of school, father’s and mother’s years of schooling, father’s occupation, and family’s possessions were retained in the model. These variables are explored in the following sections.

#### *4.3.3.1. TYPE OF SCHOOL*

The dummy variable SC_TP_D represents students who attended only private schools with no scholarship, which is one category of variable Type of School (SC_TP), and is the one with highest coefficient in the model. As shown in Table 3, that was also the first variable to enter the model. All remaining categories of SC_TP were not included.

Attending only private schools with no scholarship (13.55%) is an aspect that increases the score in the Human Sciences AK by 4.051 points. All remaining school attendance conditions have no impact on the score. These results may be related either to school efficiency and/or to students' entry level in school (Sampaio and Guimarães, 2009). However, those factors are not available in the ENEM dataset and we cannot test any hypotheses related to them.

#### *4.3.3.2. SOCIOECONOMIC STATUS*

The Human Sciences model comprises 8 socioeconomic-related dummy variables: student's father completed tertiary education but did not complete graduate studies (ST_FA_F), student's father completed graduate studies (ST_FA_G), student's mother completed tertiary education but did not complete graduate studies (ST_MO_F), student's mother completed graduate studies (ST_MO_G), student's father occupation is of high intellectual level (OC_FA_D), student's father occupation is of highest intellectual level (OC_FA_E), student owns a computer (PC_Y) at home, and English chosen as foreign language (LG_E). These are dummy variables derived from 5 categorical variables, which may be considered socioeconomic status dimensions; they are: student's father years of schooling (ST_FA), student's mother years of schooling (ST_MO), student's father occupation (OC_FA), if the student owns a computer (PC) at home, and foreign language (LG). All remaining categories for which the corresponding dummy variables did not enter the model have no impact on the score.

Regarding variables associated with possessions, the ownership of a computer at home (PC_Y) was the only dummy variable retained in the model. This variable is the one with the second highest coefficient in the model resulting in an additional 2.3 points in the student's final score in this AK.

Regarding the foreign language, students who chose English (50.42% of the sample) instead of Spanish had a better performance. This choice represented an increase of around 2.2

points in the HS AK final score. That would not be an expected result considering that the Human Sciences exam does not test the student's knowledge of a foreign language. However, the variable may be considered as part of the "family background" aspect (Coleman, 1988), since students who chose English are from a higher socioeconomic level in the sample analyzed.

From the remaining socioeconomic variables, OC_FA_E (7.26%) is the one with highest coefficient, followed by OC_FA_D (21.64%). Variable OC_FA is related to fathers' occupation. Students' whose fathers work in more intellectually demanding jobs perform around 1.7 points better in the HS AK than those whose fathers work in less intellectually demanding jobs (71.10%).

Variables related to father and mother years of schooling were also included in the model. Groups represented by variables ST_MO_F = students whose mothers completed tertiary studies (10.07%) and ST_MO_G = students whose mothers completed graduate studies (10.33%) score higher in this AK by around 1.6 points comparing to the remaining categories which did not enter the model. The same analysis applies to ST_FA_F = students whose fathers completed tertiary studies (8.85%) and to ST_FA_G = students whose fathers completed graduate studies (5.93%): these students score around 1.3 points higher than the remaining students whose fathers are in other schooling conditions (85.19%).

#### *4.3.3.3. SEX AND RACE*

Variables related to sex and race were also included in the model. The Human Sciences model indicates that male students (41,90%) present an average score 1.295 points higher than that of female students in this AK. This difference could be related to Lafontaine and Monseur's (2009) finding that males tend to perform better than females at multiple-choice questions.

For the HS AK, the dummy variable RACE_C = White (74.06%) appears in the model and represents one of the six race categories. White students score almost 0.9 points higher than students of other races. Although statistically significant, the coefficient associated with RACE_C is the smallest in the model. The superior performance of white students may be related to the quality of school they attended to and the financial, human and social capital of their families (Coleman, 1988; Sirin, 2005).

### 4.3.4. NATURAL SCIENCES

The model generated for the dependent variable *Natural Sciences Score* comprises 10 independent variables, with an adjusted $R^2$ of 0.1806 and an RMSE of 7.577. The AIC of the final model is 58.02. The complete model comprises 91 independent variables and yielded an AIC value of 67.41. Table 8 presents the coefficients for the retained variables as well as their SEs and significance.

**Table 8:** Coefficients for the NS final model

| Variable | Estimate (SE) signif |
|---|---|
| SEX_M | 1.777 (0.048) *** |
| RACE_C | 1.038 (0.040) *** |
| SC_TP_D | 5.515 (0.081) *** |
| ST_FA_F | 1.755 (0.086) *** |
| ST_FA_G | 2.561 (0.100) *** |
| ST_MO_F | 1.832 (0.107) *** |
| ST_MO_G | 1.879 (0.063) *** |
| OC_FA_D | 1.367 (0.036) *** |
| PC_Y | 2.173 (0.040) *** |
| HK_Y | 0.579 (0.049) *** |

Significance codes: '***' 0.001 '**' 0.01

Variables related to sex, race, type of school attended, father's and mother's years of schooling, father's occupation, and family's possessions were retained in the model. These variables are explored in the following sections.

#### *4.3.4.1. TYPE OF SCHOOL*

The variable Type of School (SC_TP), represented by one of its five categories, the dummy variable SC_TP_D (attended only private school with no scholarship), is the one with highest coefficient in the Natural Sciences model. It is also the variable that yielded the largest VIP value. All remaining categories of SC_TP were not included in the model. As shown in Table 3, it was also the first variable to enter the model. These aspects indicate the variable's importance in explaining students' performance in the NS AK. Attending only private schools with no scholarship (13.55%) is a condition that increases the score in this AK by 5.515 points. All remaining school attendance conditions have no impact on the score.

#### *4.3.4.2. SOCIOECONOMIC STATUS*

The Natural Sciences model comprises 5 variables that may be considered socioeconomic status dimensions; they are: student's father years of schooling (ST_FA), student's mother years of schooling (ST_MO), student's father occupation (OC_FA), if the student owns a computer (PC) and if a housekeeper works in the family's house (HK). These variables are represented by 7 dummy variables: student's father completed tertiary education but did not complete graduate studies (ST_FA_F), student's father completed graduate studies (ST_FA_G), student's mother completed tertiary education but did not complete graduate studies (ST_MO_F), student's mother completed graduate studies (ST_MO_G), student's father occupation is of a high intellectual level (OC_FA_D), student owns a computer (PC_Y) and a housekeeper works in the family's house (HK_Y).

ST_FA_G (5.93%) is the socioeconomic-related variable yielding the highest coefficient. The score in the NS AK increases by over 2.5 points for students in this group. Variable PC_Y (student owns a computer) is the one with second highest coefficient, resulting in additional 2.1 points in the student's final score. The portion of students who own a computer in the sample is 77.95%.

Considering as baseline the remaining categories which did not enter the model, groups ST_MO_F and ST_MO_G yielded an advantage of over 1.8 points in the final score. Groups represented by variables ST_MO_F and ST_MO_G comprise 10.07% and 10.33% of the sample. In terms of fathers' occupation, scores of students in group OC_FA_D is around 1.3 points higher than the score of remaining students (78.36%).

Another retained socioeconomic-related variable is the one reflecting the situation in which the students' family has a housekeeper (HK_Y), which represents 8.14% of the sample. Although statistically significant, the coefficient associated with this variable is the smallest in the model. Students from this group score around 0.5 points higher than the remaining.

#### *4.3.4.3. SEX AND RACE*

Variables related to sex and race were also included in the model. The NS model indicates that male students (41.90%) present an average score almost 1.8 points higher than that of female students. This difference could be related to the conclusion by Lafontaine and Monseur (2009) that males tend to perform better at multiple-choice questions. The inclusion of the male sex variable in the model and its positive coefficient is also aligned with results from the Mann-Whitney test, in which the average score for males and females appeared as significantly different, with male students scoring higher than female students.

It is a consensus that race is a variable that has an effect on students' performance. For the NS AK, Race appears in the model represented by one of its six categories (the dummy variable RACE_C = White), which represents 74.06% of the students in the sample. White students score over 1 point higher than students of other races.

The inclusion of the white race variable in the model is also aligned with results from the Mann-Whitney test, in which the average score for students who declared black and mixed as their race appeared as significantly different to those who declared being white, with the latter scoring higher. These three groups represent 96.69% of the students in the sample.

### 4.3.5. MATHEMATICS

The model generated for the dependent variable *Mathematics Score* comprises 13 independent variables, with an adjusted $R^2$ of 0.2246 and an RMSE of 9.395. The AIC of the final model is 89.65. The complete model (comprised of 91 variables) yielded an AIC value of 111.72. Table 9 presents the coefficients for the retained variables as well as their SEs and significance.

**Table 9:** Coefficients for the MA final model

| Variable | Estimate (SE) signif |
|---|---|
| SEX_M | 3.738 (0.050) *** |
| RACE_C | 1.608 (0.045) *** |
| LG_E | 2.135 (0.059) *** |
| SC_TP_D | 5.567 (0.085) *** |
| ST_FA_F | 1.509 (0.088) *** |
| ST_FA_G | 2.072 (0.157) *** |
| ST_MO_F | 2.247 (0.076) *** |
| ST_MO_G | 2.255 (0.067) *** |
| OC_FA_D | 1.989 (0.038) *** |
| OC_FA_E | 2.571 (0.128) *** |
| PC_Y | 2.169 (0.045) *** |
| HK_Y | 0.636 (0.143) ** |
| FZR_Y | 1.377 (0.058) *** |

Significance codes: '***' 0.001 '**' 0.01

Variables related to sex, race, foreign language, type of school, father's and mother's years of schooling, father's occupation, and family's possessions were retained in the model. These variables are explored in the following sections.

#### *4.3.5.1. TYPE OF SCHOOL*

The variable Type of School (SC_TP), represented by one of its five categories (the dummy variable SC_TP_D = attended only private schools with no scholarship), is the one with highest coefficient in the model. All remaining categories of SC_TP were not included in the model.

Attending only private schools with no scholarship (13.55%) is a condition that increases the score in the Mathematics AK by 5.567 points. All remaining school attendance conditions have no impact on the score. The higher performance of private school students may be related either to school efficiency and/or to students' entry level (Sampaio and Guimarães, 2009).

#### *4.3.5.2. SOCIOECONOMIC STATUS*

The Mathematics model comprises 10 socioeconomic-related dummy variables: student's father completed tertiary education but did not complete graduate studies (ST_FA_F), student's father completed graduate studies (ST_FA_G), student's mother completed tertiary education but did not complete graduate studies (ST_MO_F), student's mother completed graduate studies (ST_MO_G), student's father occupation is of a high intellectual level (OC_FA_D), student's father occupation is of the highest intellectual level (OC_FA_E), student owns a computer (PC_Y) and a freezer (FZR_Y) at home, a housekeeper works in the family's house (HK_Y), and English as chosen foreign language (LG_E). These variables represent 7 categorical variables; they are: student's father years of schooling (ST_FA), student's mother years of schooling (ST_MO), student's father occupation (OC_FA), if the student owns a computer (PC), if the student owns a freezer (FZR), if a housekeeper works in the family's house (HK), and foreign language (LG). All remaining categories for which the corresponding dummy variables did not enter the model have no impact on the score.

From these 10 dummy variables, OC_FA_E (7.26%) is the one with highest coefficient. Students whose fathers' occupation are of highest intellectual level (OC_FA_E) had a score advantage of 2.571 points. Students whose fathers' occupation are of high intellectual level

(OC_FA_D), which represent 21.64% of the sample, had a score advantage of 1.989 points. All remaining fathers' occupation categories have no impact on the score.

Variables related to father and mother years of schooling were also included in the model. Groups represented by variables ST_FA_F = students whose fathers completed tertiary studies (8.85%) and ST_FA_G = students whose fathers completed graduate studies (5.93%) score higher in this AK by around 1.5 and 2 points comparing to the remaining categories which did not enter the model (85.19%). The same analysis applies to ST_MO_F = students whose mothers completed tertiary studies (10.07%) and to ST_MO_G = students whose mothers completed graduate studies (10.33%): these students score around 2.2 points higher than the remaining students whose mother are in other schooling conditions (79.60%).

Regarding variables associated with possessions, the highest coefficient is associated with owning a computer at home (77.95%), which adds around 2 points to the final student's score. Additionally, owning a freezer at home (54.69%) and having the help of a housekeeper (8.14%) add over 1.3 and 0.6 points, respectively, in the final score.

#### *4.3.5.3. SEX AND RACE*

Variables related to sex and race were also included in the model. The Mathematics model indicates that male students (41,90%) present an average score 3.738 points higher than that of female students in this AK. This difference could be related to what Lafontaine and Monseur (2009) claim: that males tend to perform better than females at multiple-choice questions. The inclusion of the male sex variable in the model and its positive coefficient is also aligned with results from the Mann-Whitney test, in which the average score for males and females appeared as significantly different, with male students scoring higher than female students.

For the MA AK, the dummy variable RACE_C = White (74.06%) appears in the model and represents one of the six race categories. White students score over 1.6 points higher than students of other races. The inclusion of the white race variable in the model is also aligned with results from the Mann-Whitney test, in which the average score for students who declared to belong to black and mixed races appeared as significantly different from those who declared white race, the latter scoring higher. These three groups represent 96.69% of the students.

### 4.3.6. COMPARISON OF MODELS

The five models detailed in previous sections were generated through the same method, however independently. It is interesting to observe that from the 91 independent variables available in the dataset only 17 were retained in one or more models, while 8 were included in all models. That is consistent with the nature of the dependent variables, that measure different aspects of an aggregate output (students' performance). Results are also aligned to what is found in the literature identifying most common variables included in student performance models, which are related to socioeconomic status, sex and race.

In all models, the variable that most explained the variance in student's performance was the type of school attended by the student. It was the first variable selected to enter each model, the one yielding the highest coefficients and sMC values. That is a meaningful result, especially considering the dataset is from a Brazilian exam, where private schools perform consistently better according to the index of basic education development (IDEB – *Índice de Desenvolvimento da Educação Básica*) since 2007, when it started to be reported by INEP (2018). It points to the importance of quality in basic education on the students' development, which also impacts on the development of society as a whole. As mentioned before, these results are aligned with those from other authors (Menezes-Filho, 2007; Viggiano and Mattos, 2013).

Variable SC_TP_D, a dummy representing level D (private school, no scholarship) of type of school was included in all models with coefficient ranging from 3.464 (LC AK) to 5.567 (MA AK). It is clear that students from private schools have an advantage over those from public schools, particularly in mathematics. Carraher et al. (2013) argue that mathematics, as taught in schools, do not always take advantage of more instinctive methods and considering that the background of students' who attended public schools tend to be of families with less financial, human and social capital (Sampaio and Guimarães, 2009), they might lack in logical and mathematics thinking. SC_TP_D's coefficient in the NS model is 5.515, the second largest among models. That is an expected result since NS also measures quantitative and logical proficiency.

Following the type of school variable, the ones retained in all models were mother's and father's years of schooling. Those are categorical variables with levels deployed as binary dummy variables. All models retained the dummies associated with levels F (completed tertiary studies but not graduate studies) and G (completed graduate studies). For father's education, coefficients ranged from 1.021 (LC model) to 1.755 (NS model), and from 1.076 (LC model) to 2.561 (MA model), respectively. Mother's education coefficients ranged from 1.396 (LC model) to 2.247 (MA model) and from 1.231 (LC model) to 2.255 (MA model), respectively.

Note the impact of parent's higher level of education on MA students' scores. Based on Coleman's (1988) and Carraher et al.'s (2013) theories, it may be considered that a high level of parents' education provides better conditions for students, especially in areas of knowledge in which schools may lack more. That is backed by the fact that the highest the parents' education the highest the human and social capital they are able to transmit. The effect is more noticeable in analytical areas of knowledge.

Another variable related to level of education and that appeared in all models is the dummy OC_FA_D, corresponding to level D (father works on intellectually demanding jobs) the categorical variable OC_FA (occupation of student's father). Coefficients for this variable ranged from 1.367 (NS model) to 2.230 (ES model). Although expected to have a similar behavior to ST_FA, with coefficients weighing higher is scores from the same AKs, that is not the behavior of OC_FA_D. We could not find studies in the literature relating fathers' occupation and student's performance in tests. However, its higher effect on Essay scores may be related to the level of literacy transmitted to the student.

Another important variable related to socioeconomic status that was included in all models is the ownership of one or more computers at home (PC_Y), with coefficients ranging from 1.647 (LC model) to 2.311 (HS model), with other coefficients closer to the latter. That indicates LC to be the area of knowledge in which students least benefit from owning a computer, with all remaining areas displaying similar importance. It is clear that the ownership of at least one computer at home adds value to the development of the student, as pointed out by authors such as Menezes-Filho (2007) and de Oliveira et al. (2010). Our results reinforce that point.

The variable related to the race also was retained in all the models, indicating a homogeneous effect across all areas of knowledge. Variable coefficients in the models ranged from 0.735 (LC model) to 1.608 (MA model). That seems to be following the same pattern of parents' education variables, which makes sense if we take into account the fact that, in Brazil, the majority of people with tertiary and/or graduate degrees are auto declared white (IBGE, 2017). That is also an expected result as race is found to be related to student's performance worldwide (Fernandes, 2004; Sirin, 2005; Menezes-Filho, 2007; Stewart, 2008; de Oliveira Barbosa, 2009).

A final interesting variable to analyze is sex, which was represented by a binary dummy variables. The variable was not included in one of the models (LC), indicating that the performance between males and females in this area of knowledge are not affected by sex. In all other models the sex dummy variable SEX_M (M = male) was included, but not always with the same behavior. In the ES and MA areas of knowledge the variable's coefficients are almost the same in magnitude but have opposite signals, which was also detected in the data exploration and confirmed by the Mann-Whitney test. SEX_M's coefficient for ES is -3.545 while for MA it is 3.738. This result is aligned with the theory by Lafontaine and Monseur (2009) and Steinmayr and Spinath (2008) who advocate the gap in performance between males and females to be attributed to behavior factors as well as to a female advantage in more cognitively demanding tasks, such as essays about open-ended questions.

Table 10 summarizes the direction and magnitude of effects of independent variables in each area of knowledge. It is clear that the MA model is the one with highest coefficients in all variables discussed in this section, followed by the ES model. Finally, the LC model is the one consistently yielding the lowest coefficients.

**Table 10**: Summary of direction and magnitude of variables' effects in models

| Variable | ES | LC | HS | NS | MA |
|---|---|---|---|---|---|
| SEX_M | ↓↓ | - | ↑ | ↑ | ↑↑ |
| RACE_C | ↑↑ | ↑ | ↑ | ↑ | ↑↑ |
| SC_TP_D | ↑↑ | ↑ | ↑ | ↑↑ | ↑↑ |
| ST_FA_F | ↑ | ↑ | ↑ | ↑↑ | ↑↑ |
| ST_FA_G | ↑ | ↑ | ↑ | ↑↑ | ↑↑ |
| ST_MO_F | ↑↑ | ↑ | ↑ | ↑↑ | ↑↑ |
| ST_MO_G | ↑↑ | ↑ | ↑ | ↑↑ | ↑↑ |
| OC_FA_D | ↑↑ | ↑ | ↑ | ↑ | ↑↑ |
| PC_Y | ↑↑ | ↑ | ↑↑ | ↑↑ | ↑↑ |

**Codes:** ↑ *moderate positive effect*
↑↑ *strong positive effect*
↓↓ *strong negative effect*

### 4.3.7. DISCARDED INDEPENDENT VARIABLES

Age is the first discarded variable to be analyzed here, which is frequently mentioned in the literature, although controversial (White, 1982; Sirin, 2005). Data exploration provides some insights on why the variable is not included in the models. Regardless of the dependent variable, the impact of age on scores is not linear as age increases.

In all models, there are age ranges at which the score is not influenced by this variable. In HS and NS AKs, age only influences the score of students older than 18 years of age. In LC and MA AKs, the same phenomenon happens only with students older than 19 years of age. In the ES AK age starts to influence the score only for students older than 23 years of age.

The removal of this variable during modeling is aligned with its observed behavior in data exploration. Unless the variable had entered non-linearly in the model, it is expected to be discarded. In addition, despite being frequently mentioned in the literature, White (1982) and Sirin (2005) conclude in their meta-analyzes that age is not a unanimous variable. They show that in some studies the effect of age decreases as students become older.

Another variable that has been discarded from every model is Income. The variable may be considered as part of the aggregate variable "family background", under the financial capital concept (Coleman, 1988). However, there are three problems with this variable in the ENEM dataset. First, this kind of information, when provided by students and not by their parents, tends to be inaccurate. Second, the family income might not be the same every month, which makes the information less reliable. Finally, the student may not want to disclose this kind of information, leading to missing or false data (White, 1982; Sirin, 2005; Alves and Soares, 2009).

Another relevant aspect is that parents' years of schooling and occupation are highly correlated to family income (Duncan, 1961; White, 1982). Since the method used to generate the models tends to exclude variables that bring redundant information (already captured by other variables), that could justify the absence of Income in all five models. The same explanation applies to variables related to family possessions, which were not included in any of the models. They all cover the same financial dimension as other variables included in the models, being excluded based on redundancy.

In summary, the reasons for a variable not to be included in a model are related to the precision of the information they carry, their relevance in explaining the dependent variable, the violation of the linearity assumption underlying the regression model and the redundancy of information they represent.

# 5. FINAL CONSIDERATIONS

This section presents the conclusions of this dissertation, as well as suggestions for further studies.

## 5.1. CONCLUSION

The objective of this dissertation was to analyze the relationship between sociodemographic attributes and students' performance through the modeling of students' performance in the 5 knowledge areas evaluated on the ENEM and the variables that characterize these students. The specific objectives were: (i) the analysis of the coefficients generated by the models, (ii) discussions on the analysis of the coefficients, and (iii) a comparative analysis of the results obtained with other results found in the related literature.

The objectives of this study were achieved through the development of each of the five phases proposed by the adapted CRISP-DM method. First, a descriptive analysis of the ENEM data was carried out in order to generate a **data/environment understanding**. This analysis was performed through visual tools such as histograms and comparative graphs, as well as through normality tests and statistical comparisons between average scores from different groups of students. This first step allowed for a better understanding of the data. This analysis showed groups' average scores do not follow a normal distribution, and that most groups present significant differences between their average scores. This step also supported the second phase, **data preparation**, since the data/environment understanding made data cleaning and organization possible.

The next phase used the PLS as the regression tool for **modeling**, which consisted of three stages: (i) selection of the number of latent variables for each model; (ii) variable selection and their order of entry in the models; (iii) models' construction through an iterative process. At first, the one-sigma rule based on the RMSE was used for the selection of LVs. Next, the VIP and sMC values were used to select and order the entry of variables in each of the five models. Finally, an iterative process was carried out in which the variables were added one by one in the models and retained only when the AIC generated was better than the best AIC obtained. In this step, metrics such as RMSE and $\boldsymbol{R^2}$were also generated and reported.

The fourth phase of the CRISP-DM method corresponds to the individual **analysis** of both the variables retained in the models and the respective coefficients generated (in relation to their magnitude and the direction of their effect: positive or negative). This analysis took into account the order of dummy variables entered into the models as well as the percentages that each group represents in the sample. In this study, variables that entered the models were able to explain on average a 17.90% of variance in students' performance.

The type of school was the main variable in all the models since it was the first variable entered in all of them and showed the highest coefficients. This variable entered the model represented by the dummy variable of students who attended only private schools with no scholarship. It was the aspect that most differentiate the students' scores. Race is another variable that was retained in all five models. It was represented by the dummy variable of students who auto declared white. Since, in Brazil, the majority of black and mixed students frequent public schools, while private schools have a majority of white students, these two dummy variables are related and complement each other in the analysis and in the composition of the final scores. This effect could be related to what is called school efficiency, which is the capacity of the school to develop each student to their maximum potential as well as to what is called the students' entry level to the school, which represents all the knowledge and baggage previously acquired.

Sex was also identified as an important variable and generated different effects in each of the models: either positive, negative, or neutral. Females have better scores only in the ES AK. Males have better scores in HS, NS and MA AKs. For the LC AK, this variable was not considered relevant. The differences in results between the models could be related to the fact, found in previous studies, that females tend to have an advantage when answering open-ended questions while males have this advantage when answering multiple-choice questions.

The **discussion** was done through a comparison between the results obtained in the individual model analyses and previous results from similar studies. The results obtained in this study were consistent with literature results. The discussion also compared the 5 models in order to identify similarities and differences between them as well as to identify patterns among the different dimensions of performance. From the total 91 variables available, seventeen of them entered at least one model. Variables related to parents' years of schooling, fathers' occupation and family's possessions also were retained in all models. Finally, the variables discarded were analyzed and the results were compared to the previous literature consulted. Age was the discarded variable that most stood out in this analysis since it is frequently mentioned in the literature as a potential variable to enter in this kind of analysis. Although, the impact of age on scores in the ENEM dataset is not linear as age increases. However, this result is consistent with the descriptive analysis of the data and with the results reported in the previous literature, which noted this variable as controversial in this type of analysis.

There are several other factors that can influence students' performance, such as school efficiency and students' entry level in school, which are not captured by the ENEM questionnaire. However, results obtained are satisfactory and consistent with what was found in the consulted literature. This suggests that the research method used in this study is suitable for this kind of analysis. The analysis of the relationship between sociodemographic attributes and students' performance in ENEM could be used to understand the Brazilian educational system, and potentially help lead to improvements.

## 5.2. SUGGESTIONS FOR FURTHER STUDIES

Future research may be developed as extensions of the method proposed in this study:

(i) Conduct studies of the other four Brazilian regions: Southeast, Midwest, North, and Northeast;

(ii) Perform a clustering process in order to generate models for more homogeneous groups;

(iii) Carry out a study taking the non-linear age behavior into account;

(iv) Evaluate the use of other variable selection methods to enter the model.

## REFERENCES

Airasian, P. W. (1988). Measurement driven instruction: A closer look. Educational Measurement: Issues and Practice, 7(4), 6-11.

Alves, M. T. G., & Soares, J. F. (2009). Medidas de nível socioeconômico em pesquisas sociais: uma aplicação aos dados de uma pesquisa educacional. Opinião Pública, 15(1), 1-30.

Arai, N. H., & Matsuzaki, T. (2014). The impact of ai on education–can a robot get into the university of Tokyo. In Proc. ICCE (pp. 1034-1042).

Broadfoot, P. (1996). Education, assessment and society: A sociological analysis. Open University Pres.

Broadfoot, P. (2012). Assessment, schools and society. Routledge.

Brown, P., & Lauder, H. (1996). Education, globalization and economic development. Journal of Education Policy, 11(1), 1-25.

Carraher, T. N., Carraher, D. W., & Schliemann, A. D. (2013). Na vida dez; na escola zero: os contextos culturais da aprendizagem da matemática. Cadernos de pesquisa, (42), 79-86.

Cassel, C., Hackl, P., & Westlund, A. H. (1999). Robustness of partial least-squares method for estimating latent variable quality structures. Journal of Applied Statistics, 26(4), 435-446.

Coleman, J. S. (1988). Social capital in the creation of human capital. American journal of sociology, 94, S95-S120.

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. Social Science Research, 59, 1-12.

D'Agostino, R. B., Belanger, A., & D'Agostino Jr, R. B. (1990). A suggestion for using powerful and informative tests of normality. The American Statistician, 44(4), 316-321.

De Lange, M., Dronkers, J., & Wolbers, M. H. (2014). Single-parent family forms and children’s educational performance in a comparative perspective: Effects of school’s share of single-parent families. School Effectiveness and School Improvement, 25(3), 329-350.

de Oliveira Barbosa, M. L. (2009). Desigualdade e desempenho: uma introdução à sociologia da escola brasileira. Argumentum.

de Oliveira, I. S. V., da Silva, M. V. B., & de Siqueira, L. B. O. (2010). Determinantes do desempenho dos estudantes no vestibular da Universidade Federal da Paraíba. Revista Economia e Desenvolvimento, 7(2).

De Rosa, R. (2017). Governing by Data: Some Considerations on the Role of Learning Analytics in Education. In Data Science and Social Research (pp. 67-77). Springer, Cham.

Duncan, O. D. (1961). A socioeconomic index for all occupations. Class: Critical Concepts, 1, 388-426.

Duru-Bellat, M. & Kieffer, A. (2008). From the baccalauréat to higher education in France: Shifting inequalities. Population, 63(1), 119-154.

Eckstein, M. A. (1996). A comparative assessment of assessment. 233-240.

Esposito Vinzi, V., Chin, W. W., Henseler, J., & Wang, H. (2010). Handbook of partial least squares: Concepts, methods and applications. Heidelberg, Dordrecht, London, New York: Springer.

Evans, J. R., & Olson, D. L. (2003). Statistics, data analysis, and decision modeling. Prentice Hall.

Fernandes, D. C. (2004). Race, socioeconomic development and the educational stratification process in Brazil. Research in social stratification and mobility, 22, 365-422.

Fornell, C., & Bookstein, F. L. (1982). Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. Journal of Marketing Research, 440-452.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York, NY, USA:: Springer series in statistics.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. Analytica chimica acta, 185, 1-17.

Gipps, C. (1999). Chapter 10: Socio-cultural aspects of assessment. Review of research in education, 24(1), 355-392.

Granato, D., & Ares, G. (Eds.). (2014). Mathematical and statistical methods in food science and technology. John Wiley & Sons. 137-149.

Green, A. (1997). Education, globalization and the nation state. In Education, Globalization and the Nation State (pp. 130-186). Palgrave Macmillan, London.

Guest, M. (2008). Japanese university entrance examinations: What teachers should know. LANGUAGE TEACHER-KYOTO-JALT-, 32(2), 15.

Gujarati, D. N. (2009). Basic econometrics. Tata McGraw-Hill Education.

Haenlein, M., & Kaplan, A. M. (2004). A beginner's guide to partial least squares analysis. Understanding statistics, 3(4), 283-297.

Hair, N. L., Hanson, J. L., Wolfe, B. L., & Pollak, S. D. (2015). Association of child poverty, brain development, and academic achievement. JAMA pediatrics, 169(9), 822-829.

Hardy, M. A. (1993). Regression with dummy variables (Vol. 93). Sage.

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J. A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the programme for international student assessment. Scandinavian Journal of Educational Research, 62(3), 333-353.

IBGE: Instituto Brasileiro de Geografia e Estatística. Síntese de indicadores sociais: uma análise das condições de vida da população brasileira, 2017.

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2017). Microdata from Enem 2017. http://portal.inep.gov.br/microdados. Accessed 31 July 2018.

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2018). http://ideb.inep.gov.br/resultado/. Accessed 27 February 2019.

Jerrim, J. (2012). The Socio-Economic Gradient in Teenagers' Reading Skills: How Does England Compare with Other Countries? Fiscal Studies, 33(2), 159-184.

Keeves, J. P. (1994). National examinations: design, procedures and reporting (Vol. 50). Unesco.

Lafontaine, D., & Monseur, C. (2009). Gender Gap in Comparative Studies of Reading Comprehension: to what extent do the test characteristics make a difference? European Educational Research Journal, 8(1), 69-79.

Lietz, P. (2006). A meta-analysis of gender differences in reading achievement at the secondary school level. Studies in Educational Evaluation, 32(4), 317-344.

Liland, K. H., Mehmood, T., Sæbø, S. (2017). Package ‘plsVarSel’. 23p.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics, 50-60.

Martins, L., & Veiga, P. (2010). Do inequalities in parents' education play an important role in PISA students' mathematics achievement test score disparities? Economics of Education Review, 29(6), 1016-1033.

McLoyd, V. C. (1998). Socioeconomic disadvantage and child development. American psychologist, 53(2), 185.

Menezes-Filho, N. A. (2007). Os determinantes do desempenho escolar do Brasil (pp. 1-31). IFB.

Mevik, B. H., Wehrens, R. & Liland, K. H. (2016). Package 'pls'. 59p.

Nonoyama-Tarumi, Y. (2008). Cross-national estimates of the effects of family background on student achievement: A sensitivity analysis. International Review of Education, 54(1), 57-82.

OECD. (2000). Measuring Student Knowledge and Skills: The PISA 2000 Assessment of Reading, Mathematical and Scientific Literacy. Paris: OECD.

Perry, L. B., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. Teachers College Record, 112(4), 1137-1162.

Pilz, M. (2009). After Abitur, First an Apprenticeship and then University? Why German Abitur Holders Are Taking Vocational Training in the Financial Services Sector. European Journal of vocational training, 46(1), 41-65.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.".

Randler, C., & Frech, D. (2006). Correlation between morningness–eveningness and final school leaving exams. Biological Rhythm Research, 37(3), 233-239.

Reichelt, M. (1997). Writing instruction at the German Gymnasium: A 13th-grade English class writes the Abitur. Journal of Second Language Writing, 6(3), 265-291.

Runci, M. C., Di Bella, G., & Cuppone, F. (2017). Integrated Education Microdata to Support Statistics Production. In Data Science and Social Research (pp. 283-290). Springer, Cham.

Sampaio, B., & Guimarães, J. (2009). Diferenças de eficiência entre ensino público e privado no Brasil. Economia Aplicada, 13(1), 45-68.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3/4), 591-611.

Shavit, Y. (Ed.). (2007). Stratification in higher education: A comparative study. Stanford University Press.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4), 13-22.

Silva, E. M. D., & Araújo, D. L. D. (2009). University entrance exam: the washback effect of textual genre knowledge. Trabalhos em Linguística Aplicada, 48(1), 133-152.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. Review of Educational Research, 75(3), 417-453.

Stearns, B., Rangel, F., Rangel, F., de Faria, F. F., Oliveira, J., & Ramos, A. A. D. S. (2017). Scholar Performance Prediction using Boosted Regression Trees Techniques. In European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN). Citeseer.

Steinmayr, R., & Spinath, B. (2008). Sex differences in school achievement: What are the roles of personality and achievement motivation? European Journal of Personality: Published for the European Association of Personality Psychology, 22(3), 185-209.

Stewart, E. B. (2008). School structural characteristics, student effort, peer associations, and parental involvement: The influence of school-and individual-level factors on academic achievement. Education and urban society, 40(2), 179-204.

Tran, T. N., Afanador, N. L., Buydens, L. M., & Blanchet, L. (2014). Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC). Chemometrics and Intelligent Laboratory Systems, 138, 153-160.

Tucker-Drob, E. M. (2013). How many pathways underlie socioeconomic differences in the development of cognition and achievement? Learning and Individual Differences, 25, 12-20.

Vahdat, M., Ghio, A., Oneto, L., Anguita, D., Funk, M., & Rauterberg, M. (2015). Advances in learning analytics and educational data mining. Proc. of ESANN2015, 297-306.

Viggiano, E., & Mattos, C. (2013). O desempenho de estudantes no Enem 2010 em diferentes regiões brasileiras. Revista Brasileira de Estudos Pedagógicos, 94(237).

Watanabe, Y. (2013). The National Center Test for University Admissions. Language Testing, 30(4), 565-573.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. Psychological bulletin, 91(3), 461.

Wold, S., Ruhe, A., Wold, H., & Dunn, III, W. J. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM Journal on Scientific and Statistical Computing, 5(3), 735-743.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. Chemometrics and intelligent laboratory systems, 58(2), 109-130.

Zhou, Y., Zhu, Y., & Leung, S. W. (2014). PLS-Frailty Model for Cancer Survival Analysis Based on Gene Expression Profiles. In International Conference on Partial Least Squares and Related Methods (pp. 189-199). Springer, Cham.

# APPENDIX A – VARIABLE DESCRIPTION

**INDEPENDENT VARIABLES**

| VARIABLE | DESCRIPTION | CATEGORY | CATEGORY DESCRIPTION |
|---|---|---|---|
| AGE | Student's age | - | - |
| SEX | Student's sex | F* | Female |
| | | M | Male |
| RACE | Student's auto declared race | A* | Black |
| | | B | Mixed |
| | | C | White |
| | | D | Yellow |
| | | E | Indigenous |
| | | F | Not declared |
| LG | Student’s foreign language option | S* | Spanish |
| | | E | English |
| SC_TP | Student's type of school attended | A* | Only public school |
| | | B | Part in public school and part in private school with no scholarship |
| | | C | Part in public school and part in private school with scholarship |
| | | D | Only private school with no scholarship |
| | | E | Only private school with scholarship |
| HS_ST | Situation of student's High School diploma | A* | Never completed High School and it is not attending classes to complete |
| | | B | Still have more than one year to finish High School |
| | | C | Is in the last year of High School |
| | | D | Completed High School by 2016 |
| | | E | Completed High School by 2015 |
| | | F | Completed High School by 2014 |
| | | G | Completed High School by 2013 |
| | | H | Completed High School by 2012 |
| | | I | Completed High School by 2011 |
| | | J | Completed High School by 2010 |
| | | K | Completed High School by 2009 |
| | | L | Completed High School by 2008 |
| | | M | Completed High School by 2007 |
| | | N | Completed High School before 2007 |

**INDEPENDENT VARIABLES**

| VARIABLE | DESCRIPTION | CATEGORY | CATEGORY DESCRIPTION |
|---|---|---|---|
| IN | Family's monthly income | A* | Up to R$ 1,405.50 (1.5 minimum wages) |
| | | B | From R$ 1,405.51 up to R$ 1,874.00 (from 1.5 to 2 minimum wages) |
| | | C | From R$ 1,874.01 up to R$ 2,342.50 (from 2 to 2.5 minimum wages) |
| | | D | From R$ 2,342.51 up to R$ 2,811.00 (from 2.5 to 3 minimum wages) |
| | | E | From R$ 2,811.01 up to R$ 3,748.00 (from 3 to 4 minimum wages) |
| | | F | From R$ 3,748.01 up to R$ 4,685.00 (from 4 to 5 minimum wages) |
| | | G | From R$ 4,685.01 up to R$ 5,622.00 (from 5 to 6 minimum wages) |
| | | H | From R$ 5,622.01 up to R$ 6,559.00 (from 6 to 7 minimum wages) |
| | | I | From R$ 6,559.01 up to R$ 7,496.00 (from 7 to 8 minimum wages) |
| | | J | From R$ 7,496.01 up to R$ 8,433.00 (from 8 to 9 minimum wages) |
| | | K | From R$ 8,433.01 up to R$ 9,370.00 (from 9 to 10 minimum wages) |
| | | L | From R$ 9,370.01 up to R$ 11,244.00 (from 10 to 12 minimum wages) |
| | | M | From R$ 11,244.01 up to R$ 14,055.00 (from 12 to 15 minimum wages) |
| | | N | From R$ 14,055.01 up to R$ 18,740.00 (from 15 to 20 minimum wages) |
| | | O | More than R$ 18,740.00 (20 minimum wages) |
| ST_FA | Years of schooling of student's father | A* | Never studied |
| | | B | Studied but did not complete primary education |
| | | C | Completed primary education but did not complete lower secondary education |
| | | D | Completed lower secondary education but did not complete High School |
| | | E | Completed High School but did not complete tertiary education |
| | | F | Completed tertiary education but did not complete graduate studies |
| | | G | Completed graduate studies |
| | | H | Not informed |
| ST_MO | Years of schooling of student's mother | A* | Never studied |
| | | B | Studied but did not complete primary education |
| | | C | Completed primary education but did not complete lower secondary education |
| | | D | Completed lower secondary education but did not complete High School |
| | | E | Completed High School but did not complete tertiary education |
| | | F | Completed tertiary education but did not complete graduate studies |
| | | G | Completed graduate studies |
| | | H | Not informed |

| INDEPENDENT VARIABLES | | | |
|---|---|---|---|
| VARIABLE | DESCRIPTION | CATEGORY | CATEGORY DESCRIPTION |
| OC_FA | Student's father occupation | A* | Farmer, livestock farmer (cattle, pigs, chickens, sheep, horses, etc.), beekeeper, fisherman, lumberjack, rubber tapper, extractivist |
| | | B | Housekeeper, elderly caregiver, nanny, cook (in private homes), private driver, gardener, janitor, guard, porter, postman, administrative assistant, receptionist, bricklayer, repositories of merchandise |
| | | C | Baker, industrial or restaurant cook, cobbler, dressmaker, jeweler, mechanic, machine operator, welder, factory worker, mining worker, painter, electrician, plumber, driver, truck driver, taxi driver |
| | | D | Teacher (primary or secondary education, language, music, arts etc.), technician (nursing, accounting, electronics etc.), police officer, low military officer (soldier, corporal, sergeant), supervisor, manager, master builder, pastor, micro entrepreneur (owner of a company with less than 10 employees), small trader, small landowner, self-employed or self-employed |
| | | E | Doctor, engineer, dentist, psychologist, economist, lawyer, judge, promoter, defender, delegate, lieutenant, captain, colonel, university professor, director in public or private companies, politician, owner of companies with more than 10 employees |
| | | F | Not informed |
| OC_MO | Student's mother occupation | A* | Farmer, livestock farmer (cattle, pigs, chickens, sheep, horses, etc.), beekeeper, fisherwoman, lumberjack, rubber tapper, extractivist |
| | | B | Housekeeper, elderly caregiver, nanny, cook (in private homes), private driver, gardener, janitor, guard, porter, postman, administrative assistant, receptionist, bricklayer, repositories of merchandise |
| | | C | Baker, industrial or restaurant cook, cobbler, dressmaker, jeweler, mechanic, machine operator, welder, factory worker, mining worker, painter, electrician, plumber, driver, truck driver, taxi driver |
| | | D | Teacher (primary or secondary education, language, music, arts etc.), technician (nursing, accounting, electronics etc.), police officer, low military officer (soldier, corporal, sergeant), supervisor, manager, master builder, pastor, micro entrepreneur (owner of a company with less than 10 employees), small trader, small landowner, self-employed or self-employed |
| | | E | Doctor, engineer, dentist, psychologist, economist, lawyer, judge, promoter, defender, delegate, lieutenant, captain, colonel, university professor, director in public or private companies, politician, owner of companies with more than 10 employees |
| | | F | Not informed |

**INDEPENDENT VARIABLES**

| VARIABLE | DESCRIPTION | CATEGORY | CATEGORY DESCRIPTION |
|---|---|---|---|
| RES | Quantity of residents in the student's house (including the student) | A* | 1 |
| | | B | 2 |
| | | C | 3 |
| | | D | 4 |
| | | E | 5 |
| | | F | 6 |
| | | G | 7 |
| | | H | 8 |
| | | I | 9 |
| | | J | 10 or more |
| CAR | If there are one or more cars in the student's house | N* | No |
| | | Y | Yes |
| MOTO | If there are one or more motorcycles in the student's house | N* | No |
| | | Y | Yes |
| CELL | Quantity of cell phones in the student's residence | A* | None or one |
| | | B | Two or more |
| PC | If there are one or more computers in the student's house | N* | No |
| | | Y | Yes |
| INT | If there is access to the internet in the student's house | N* | No |
| | | Y | Yes |
| CB_TV | If there is cable TV installed in the student's house | N* | No |
| | | Y | Yes |

**INDEPENDENT VARIABLES**

| VARIABLE | DESCRIPTION | CATEGORY | CATEGORY DESCRIPTION |
|---|---|---|---|
| TV | Quantity of televisions in the student's house | A*<br>B | None or one<br>Two or more |
| DVD | If there is a DVD player in the student's house | N*<br>Y | No<br>Yes |
| TP | If there is a fixed telephone installed in the student's house | N*<br>Y | No<br>Yes |
| HK | If a housekeeper works in the student's house | N*<br>Y | No<br>Yes |
| BATH | Quantity of bathrooms in the student's house | A*<br>B | None or one<br>Two or more |
| BED | Quantity of bedrooms in the student's house | A*<br>B | None or one<br>Two or more |
| MW | If there are one or more microwaves in the student's house | N*<br>Y | No<br>Yes |
| FGE | Quantity of fridges in the student's house | A*<br>B | None or one<br>Two or more |
| FZR | If there are one or more freezers in the student's house | N*<br>Y | No<br>Yes |

### INDEPENDENT VARIABLES

| VARIABLE | DESCRIPTION | CATEGORY | CATEGORY DESCRIPTION |
|---|---|---|---|
| WA_MA | If there are one or more washing machines in the student's house | N*<br>Y | No<br>Yes |
| DR_MA | If there are one or more drying machines in the student's house | N*<br>Y | No<br>Yes |
| DI_WA | If there are one or more dish washers in the student's house | N*<br>Y | No<br>Yes |
| VA_CL | If there is a vacuum cleaner in the student's house | N*<br>Y | No<br>Yes |

### DEPENDENT VARIABLES

| VARIABLE | DESCRIPTION | CATEGORY | CATEGORY DESCRIPTION |
|---|---|---|---|
| ES_SC | Student's score for the Essay | - | - |
| LC_SC | Student's score for Languages and Codes | - | - |
| HS_SC | Student's score for Human Sciences | - | - |
| NS_SC | Student's score for Natural Sciences | - | - |
| MA_SC | Student's score for Mathematics | - | - |

* Variable categories that were used as the control group. No dummy variables were created from these categories.

## APPENDIX B – SHAPIRO-WILK AND D’AGOSTINO-PEARSON’S NORMALITY TESTS RESULTS

| | ES | | | | LC | | | | HS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SHAPIRO-WILK | | D’AGOSTINO-PEARSON’S | | SHAPIRO-WILK | | D’AGOSTINO-PEARSON’S | | SHAPIRO-WILK | | D’AGOSTINO-PEARSON’S | |
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| ST_FA | | | | | | | | | | | | |
| A | 0.992 | 0.000 | 80.069 | 0.000 | 0.986 | 0.000 | 302.227 | 0.000 | 0.993 | 0.000 | 189.628 | 0.000 |
| B | 0.989 | 0.000 | 1587.714 | 0.000 | 0.987 | 0.000 | 2850.620 | 0.000 | 0.992 | 0.000 | 1882.769 | 0.000 |
| C | 0.989 | 0.000 | 1522.840 | 0.000 | 0.987 | 0.000 | 2918.597 | 0.000 | 0.992 | 0.000 | 1464.435 | 0.000 |
| D | 0.987 | 0.000 | 1656.136 | 0.000 | 0.987 | 0.000 | 2547.816 | 0.000 | 0.991 | 0.000 | 1245.872 | 0.000 |
| E | 0.988 | 0.000 | 2781.047 | 0.000 | 0.983 | 0.000 | 6619.970 | 0.000 | 0.989 | 0.000 | 2479.229 | 0.000 |
| F | 0.992 | 0.000 | 271.168 | 0.000 | 0.973 | 0.000 | 3250.026 | 0.000 | 0.980 | 0.000 | 1536.390 | 0.000 |
| G | 0.993 | 0.000 | 113.528 | 0.000 | 0.963 | 0.000 | 3009.463 | 0.000 | 0.972 | 0.000 | 1673.757 | 0.000 |
| H | 0.990 | 0.000 | 597.712 | 0.000 | 0.987 | 0.000 | 1188.277 | 0.000 | 0.991 | 0.000 | 730.272 | 0.000 |
| ST_MO | | | | | | | | | | | | |
| A | 0.991 | 0.000 | 80.372 | 0.000 | 0.985 | 0.000 | 283.431 | 0.000 | 0.992 | 0.000 | 163.693 | 0.000 |
| B | 0.990 | 0.000 | 974.942 | 0.000 | 0.985 | 0.000 | 2519.229 | 0.000 | 0.991 | 0.000 | 1496.603 | 0.000 |
| C | 0.989 | 0.000 | 1356.555 | 0.000 | 0.986 | 0.000 | 2871.110 | 0.000 | 0.991 | 0.000 | 1577.737 | 0.000 |
| D | 0.989 | 0.000 | 1532.405 | 0.000 | 0.986 | 0.000 | 2879.694 | 0.000 | 0.992 | 0.000 | 1342.734 | 0.000 |
| E | 0.989 | 0.000 | 3235.965 | 0.000 | 0.985 | 0.000 | 6437.257 | 0.000 | 0.990 | 0.000 | 2734.623 | 0.000 |
| F | 0.991 | 0.000 | 391.603 | 0.000 | 0.976 | 0.000 | 3324.687 | 0.000 | 0.982 | 0.000 | 1488.920 | 0.000 |
| G | 0.992 | 0.000 | 269.375 | 0.000 | 0.973 | 0.000 | 3667.772 | 0.000 | 0.979 | 0.000 | 1752.509 | 0.000 |
| H | 0.989 | 0.000 | 254.399 | 0.000 | 0.991 | 0.000 | 236.175 | 0.000 | 0.993 | 0.000 | 207.465 | 0.000 |
| SEX | | | | | | | | | | | | |
| M | 0.987 | 0.000 | 5.122.707 | 0.000 | 0.986 | 0.000 | 7.368.572 | 0.000 | 0.989 | 0.000 | 3.378.040 | 0.000 |
| F | 0.987 | 0.000 | 6.026.447 | 0.000 | 0.990 | 0.000 | 7.677.271 | 0.000 | 0.994 | 0.000 | 4.523.133 | 0.000 |

| | ES | | | | LC | | | | HS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SHAPIRO-WILK | | D’AGOSTINO-PEARSON’S | | SHAPIRO-WILK | | D’AGOSTINO-PEARSON’S | | SHAPIRO-WILK | | D’AGOSTINO-PEARSON’S | |
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| SC_TP | | | | | | | | | | | | |
| A | 0.988 | 0.000 | 8019.444 | 0.000 | 0.986 | 0.000 | 14203.329 | 0.000 | 0.991 | 0.000 | 7069.660 | 0.000 |
| B | 0.988 | 0.000 | 344.713 | 0.000 | 0.980 | 0.000 | 1009.799 | 0.000 | 0.987 | 0.000 | 419.200 | 0.000 |
| C | 0.988 | 0.000 | 201.528 | 0.000 | 0.984 | 0.000 | 489.678 | 0.000 | 0.987 | 0.000 | 216.877 | 0.000 |
| D | 0.992 | 0.000 | 285.611 | 0.000 | 0.966 | 0.000 | 6819.560 | 0.000 | 0.969 | 0.000 | 4645.034 | 0.000 |
| E | 0.990 | 0.000 | 138.684 | 0.000 | 0.981 | 0.000 | 818.439 | 0.000 | 0.979 | 0.000 | 540.138 | 0.000 |
| RACE | | | | | | | | | | | | |
| A | 0.991 | 0.000 | 352.002 | 0.000 | 0.985 | 0.000 | 968.664 | 0.000 | 0.991 | 0.000 | 757.144 | 0.000 |
| B | 0.989 | 0.000 | 1751.176 | 0.000 | 0.988 | 0.000 | 2608.077 | 0.000 | 0.993 | 0.000 | 1577.074 | 0.000 |
| C | 0.988 | 0.000 | 7786.179 | 0.000 | 0.988 | 0.000 | 12065.974 | 0.000 | 0.992 | 0.000 | 5521.982 | 0.000 |
| D | 0.988 | 0.000 | 113.730 | 0.000 | 0.991 | 0.000 | 124.531 | 0.000 | 0.991 | 0.000 | 128.569 | 0.000 |
| E | 0.992 | 0.000 | 9.031 | 0.011 | 0.987 | 0.000 | 24.979 | 0.000 | 0.990 | 0.000 | 18.067 | 0.000 |
| F | 0.988 | 0.000 | 180.864 | 0.000 | 0.986 | 0.000 | 298.013 | 0.000 | 0.991 | 0.000 | 142.489 | 0.000 |
| AGE | | | | | | | | | | | | |
| 16 | 0.992 | 0.000 | 386.881 | 0.000 | 0.984 | 0.000 | 1.602.990 | 0.000 | 0.984 | 0.000 | 949.981 | 0.000 |
| 17 | 0.991 | 0.000 | 1.636.296 | 0.000 | 0.989 | 0.000 | 3.488.876 | 0.000 | 0.990 | 0.000 | 2.381.776 | 0.000 |
| 18 | 0.987 | 0.000 | 1.781.937 | 0.000 | 0.991 | 0.000 | 2.074.967 | 0.000 | 0.993 | 0.000 | 1.452.235 | 0.000 |
| 19 | 0.984 | 0.000 | 1.299.520 | 0.000 | 0.990 | 0.000 | 1.333.308 | 0.000 | 0.994 | 0.000 | 823.759 | 0.000 |
| 20 | 0.982 | 0.000 | 1.218.373 | 0.000 | 0.990 | 0.000 | 892.774 | 0.000 | 0.994 | 0.000 | 520.399 | 0.000 |
| 21 | 0.983 | 0.000 | 890.581 | 0.000 | 0.990 | 0.000 | 712.993 | 0.000 | 0.994 | 0.000 | 335.733 | 0.000 |
| 22 | 0.983 | 0.000 | 688.225 | 0.000 | 0.989 | 0.000 | 597.870 | 0.000 | 0.993 | 0.000 | 278.814 | 0.000 |
| 23 | 0.983 | 0.000 | 535.734 | 0.000 | 0.988 | 0.000 | 483.427 | 0.000 | 0.992 | 0.000 | 239.909 | 0.000 |
| 24 | 0.983 | 0.000 | 435.156 | 0.000 | 0.987 | 0.000 | 423.949 | 0.000 | 0.992 | 0.000 | 207.531 | 0.000 |
| 25 | 0.985 | 0.000 | 309.684 | 0.000 | 0.988 | 0.000 | 338.846 | 0.000 | 0.993 | 0.000 | 101.155 | 0.000 |

|  | NS |  |  |  | MA |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | SHAPIRO-WILK |  | D’AGOSTINO-PEARSON’S |  | SHAPIRO-WILK |  | D’AGOSTINO-PEARSON’S |  |
|  | Statistics | p-value | Statistics | p-value | Statistics | p-value | Statistics | p-value |
| ST_FA |  |  |  |  |  |  |  |  |
| A | 0.989 | 0.000 | 211.464 | 0.000 | 0.954 | 0.000 | 650.446 | 0.000 |
| B | 0.991 | 0.000 | 1829.843 | 0.000 | 0.972 | 0.000 | 3354.775 | 0.000 |
| C | 0.991 | 0.000 | 1610.361 | 0.000 | 0.978 | 0.000 | 2525.070 | 0.000 |
| D | 0.992 | 0.000 | 1173.432 | 0.000 | 0.979 | 0.000 | 2206.588 | 0.000 |
| E | 0.993 | 0.000 | 2668.791 | 0.000 | 0.984 | 0.000 | 3519.617 | 0.000 |
| F | 0.991 | 0.000 | 797.031 | 0.000 | 0.992 | 0.000 | 1213.530 | 0.000 |
| G | 0.989 | 0.000 | 516.013 | 0.000 | 0.992 | 0.000 | 897.848 | 0.000 |
| H | 0.990 | 0.000 | 629.238 | 0.000 | 0.971 | 0.000 | 1496.660 | 0.000 |
| ST_MO |  |  |  |  |  |  |  |  |
| A | 0.989 | 0.000 | 242.305 | 0.000 | 0.954 | 0.000 | 549.018 | 0.000 |
| B | 0.990 | 0.000 | 1693.101 | 0.000 | 0.970 | 0.000 | 2750.805 | 0.000 |
| C | 0.991 | 0.000 | 1637.741 | 0.000 | 0.977 | 0.000 | 2505.696 | 0.000 |
| D | 0.991 | 0.000 | 1350.691 | 0.000 | 0.978 | 0.000 | 2421.978 | 0.000 |
| E | 0.992 | 0.000 | 2839.977 | 0.000 | 0.983 | 0.000 | 4078.049 | 0.000 |
| F | 0.992 | 0.000 | 893.707 | 0.000 | 0.990 | 0.000 | 1433.383 | 0.000 |
| G | 0.991 | 0.000 | 939.956 | 0.000 | 0.991 | 0.000 | 1547.341 | 0.000 |
| H | 0.988 | 0.000 | 208.637 | 0.000 | 0.960 | 0.000 | 706.511 | 0.000 |
| SEX |  |  |  |  |  |  |  |  |
| M | 0.994 | 0.000 | 1.975.530 | 0.000 | 0.986 | 0.000 | 4.717.196 | 0.000 |
| F | 0.989 | 0.000 | 6.036.796 | 0.000 | 0.970 | 0.000 | 13.942.656 | 0.000 |

| | NS | | | | MA | | | |
|---|---|---|---|---|---|---|---|---|
| | SHAPIRO-WILK | | D’AGOSTINO-PEARSON’S | | SHAPIRO-WILK | | D’AGOSTINO-PEARSON’S | |
| | Statistics | p-value | Statistics | p-value | Statistics | p-value | Statistics | p-value |
| SC_TP | | | | | | | | |
| A | 0.991 | 0.000 | 6968.476 | 0.000 | 0.977 | 0.000 | 12782.299 | 0.000 |
| B | 0.993 | 0.000 | 371.909 | 0.000 | 0.982 | 0.000 | 571.375 | 0.000 |
| C | 0.992 | 0.000 | 167.959 | 0.000 | 0.983 | 0.000 | 238.304 | 0.000 |
| D | 0.986 | 0.000 | 1627.098 | 0.000 | 0.994 | 0.000 | 1326.280 | 0.000 |
| E | 0.994 | 0.000 | 88.956 | 0.000 | 0.993 | 0.000 | 205.623 | 0.000 |
| RACE | | | | | | | | |
| A | 0.988 | 0.000 | 745.230 | 0.000 | 0.965 | 0.000 | 1305.368 | 0.000 |
| B | 0.989 | 0.000 | 1631.412 | 0.000 | 0.970 | 0.000 | 3791.862 | 0.000 |
| C | 0.993 | 0.000 | 5704.618 | 0.000 | 0.979 | 0.000 | 11365.635 | 0.000 |
| D | 0.989 | 0.000 | 177.351 | 0.000 | 0.976 | 0.000 | 229.642 | 0.000 |
| E | 0.990 | 0.000 | 9.516 | 0.009 | 0.967 | 0.000 | 50.613 | 0.000 |
| F | 0.990 | 0.000 | 294.492 | 0.000 | 0.974 | 0.000 | 357.758 | 0.000 |
| AGE | | | | | | | | |
| 16 | 0.990 | 0.000 | 1.243.691 | 0.000 | 0.987 | 0.000 | 862.187 | 0.000 |
| 17 | 0.991 | 0.000 | 2.907.707 | 0.000 | 0.979 | 0.000 | 3.754.267 | 0.000 |
| 18 | 0.989 | 0.000 | 1.923.554 | 0.000 | 0.973 | 0.000 | 3.654.484 | 0.000 |
| 19 | 0.987 | 0.000 | 1.226.818 | 0.000 | 0.969 | 0.000 | 2.281.556 | 0.000 |
| 20 | 0.986 | 0.000 | 816.545 | 0.000 | 0.971 | 0.000 | 1.589.416 | 0.000 |
| 21 | 0.988 | 0.000 | 527.946 | 0.000 | 0.972 | 0.000 | 1.027.080 | 0.000 |
| 22 | 0.991 | 0.000 | 309.016 | 0.000 | 0.975 | 0.000 | 715.323 | 0.000 |
| 23 | 0.992 | 0.000 | 227.488 | 0.000 | 0.976 | 0.000 | 544.272 | 0.000 |
| 24 | 0.992 | 0.000 | 151.151 | 0.000 | 0.975 | 0.000 | 439.712 | 0.000 |
| 25 | 0.993 | 0.000 | 132.403 | 0.000 | 0.977 | 0.000 | 319.462 | 0.000 |

## APPENDIX C – MANN-WHITNEY TEST RESULTS

ST_FA

| ES | B | | C | | D | | E | | F | | G | | H | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -18.20 | 0.000 | -30.83 | 0.000 | -36.17 | 0.000 | -51.45 | 0.000 | -72.41 | 0.000 | -79.18 | 0.000 | -21.89 | 0.000 |
| B | | | -30.03 | 0.000 | -42.11 | 0.000 | -87.39 | 0.000 | -116.69 | 0.000 | -120.44 | 0.000 | -9.65 | 0.000 |
| C | | | | | -13.41 | 0.000 | -53.73 | 0.000 | -93.94 | 0.000 | -101.83 | 0.000 | -13.16 | 0.000 |
| D | | | | | | | -36.16 | 0.000 | -80.39 | 0.000 | -90.37 | 0.000 | -23.14 | 0.000 |
| E | | | | | | | | | -60.54 | 0.000 | -74.19 | 0.000 | -52.77 | 0.000 |
| F | | | | | | | | | | | -19.91 | 0.000 | -87.13 | 0.000 |
| G | | | | | | | | | | | | | -95.56 | 0.000 |

| LC | B | | C | | D | | E | | F | | G | | H | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -11.80 | 0.000 | -20.65 | 0.000 | -27.56 | 0.000 | -48.13 | 0.000 | -76.32 | 0.000 | -84.03 | 0.000 | -21.60 | 0.000 |
| B | | | -20.53 | 0.000 | -35.89 | 0.000 | -93.74 | 0.000 | -134.45 | 0.000 | -137.52 | 0.000 | -20.40 | 0.000 |
| C | | | | | -16.17 | 0.000 | -70.59 | 0.000 | -119.44 | 0.000 | -125.69 | 0.000 | -4.90 | 0.000 |
| D | | | | | | | -49.47 | 0.000 | -104.09 | 0.000 | -113.28 | 0.000 | -7.69 | 0.000 |
| E | | | | | | | | | -77.47 | 0.000 | -92.20 | 0.000 | -45.81 | 0.000 |
| F | | | | | | | | | | | -24.02 | 0.000 | -93.18 | 0.000 |
| G | | | | | | | | | | | | | -103.24 | 0.000 |

| HS | B | | C | | D | | E | | F | | G | | H | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -8.53 | 0.000 | -16.72 | 0.000 | -21.46 | 0.000 | -39.94 | 0.000 | -70.52 | 0.000 | -80.02 | 0.000 | -13.77 | 0.000 |
| B | | | -18.81 | 0.000 | -29.07 | 0.000 | -79.93 | 0.000 | -126.56 | 0.000 | -133.05 | 0.000 | -11.01 | 0.000 |
| C | | | | | -10.96 | 0.000 | -58.49 | 0.000 | -112.21 | 0.000 | -121.64 | 0.000 | -3.27 | 0.001 |
| D | | | | | | | -43.61 | 0.000 | -100.72 | 0.000 | -112.23 | 0.000 | -11.69 | 0.000 |
| E | | | | | | | | | -77.78 | 0.000 | -94.20 | 0.000 | -45.87 | 0.000 |
| F | | | | | | | | | | | -25.50 | 0.000 | -94.13 | 0.000 |
| G | | | | | | | | | | | | | -105.59 | 0.000 |

**ST_FA**

| **NS** | B | | C | | D | | E | | F | | G | | H | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -9.23 | 0.000 | -17.03 | 0.000 | -21.74 | 0.000 | -39.79 | 0.000 | -71.37 | 0.000 | -82.83 | 0.000 | -11.96 | 0.000 |
| B | | | -18.08 | 0.000 | -28.49 | 0.000 | -78.29 | 0.000 | -127.52 | 0.000 | -137.63 | 0.000 | -6.65 | 0.000 |
| C | | | | | -11.20 | 0.000 | -57.73 | 0.000 | -113.62 | 0.000 | -126.62 | 0.000 | -6.85 | 0.000 |
| D | | | | | | | -42.30 | 0.000 | -101.09 | 0.000 | -116.11 | 0.000 | -15.20 | 0.000 |
| E | | | | | | | | | -78.92 | 0.000 | -99.04 | 0.000 | -48.51 | 0.000 |
| F | | | | | | | | | | | -29.10 | 0.000 | -96.95 | 0.000 |
| G | | | | | | | | | | | | | -110.90 | 0.000 |

| **MA** | B | | C | | D | | E | | F | | G | | H | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -19.87 | 0.000 | -33.68 | 0.000 | -38.59 | 0.000 | -56.17 | 0.000 | -80.94 | 0.000 | -88.54 | 0.000 | -25.42 | 0.000 |
| B | | | -32.88 | 0.000 | -43.78 | 0.000 | -96.40 | 0.000 | -133.62 | 0.000 | -138.55 | 0.000 | -13.52 | 0.000 |
| C | | | | | -12.22 | 0.000 | -60.12 | 0.000 | -110.38 | 0.000 | -120.53 | 0.000 | -11.43 | 0.000 |
| D | | | | | | | -43.91 | 0.000 | -98.20 | 0.000 | -110.63 | 0.000 | -20.58 | 0.000 |
| E | | | | | | | | | -74.85 | 0.000 | -92.65 | 0.000 | -55.32 | 0.000 |
| F | | | | | | | | | | | -26.62 | 0.000 | -98.35 | 0.000 |
| G | | | | | | | | | | | | | -109.15 | 0.000 |

**ST_MO**

| **ES** | B | | C | | D | | E | | F | | G | | H | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -18.57 | 0.000 | -29.89 | 0.000 | -33.76 | 0.000 | -48.63 | 0.000 | -68.90 | 0.000 | -73.55 | 0.000 | -12.00 | 0.000 |
| B | | | -25.72 | 0.000 | -34.85 | 0.000 | -77.03 | 0.000 | -110.74 | 0.000 | -121.95 | 0.000 | -3.80 | 0.000 |
| C | | | | | -9.56 | 0.000 | -50.27 | 0.000 | -92.56 | 0.000 | -104.52 | 0.000 | -16.92 | 0.000 |
| D | | | | | | | -39.23 | 0.000 | -84.52 | 0.000 | -96.57 | 0.000 | -21.55 | 0.000 |
| E | | | | | | | | | -62.55 | 0.000 | -76.54 | 0.000 | -39.21 | 0.000 |
| F | | | | | | | | | | | -11.18 | 0.000 | -64.28 | 0.000 |
| G | | | | | | | | | | | | | -70.07 | 0.000 |

**ST_MO**

| **LC** | B | | C | | D | | E | | F | | G | | H | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -11.16 | 0.000 | -18.17 | 0.000 | -23.01 | 0.000 | -42.67 | 0.000 | -68.93 | 0.000 | -71.68 | 0.000 | -8.07 | 0.000 |
| B | | | -15.75 | 0.000 | -26.76 | 0.000 | -79.95 | 0.000 | -123.57 | 0.000 | -130.46 | 0.000 | **-0.77** | **0.220** |
| C | | | | | -11.32 | 0.000 | -64.66 | 0.000 | -113.81 | 0.000 | -121.16 | 0.000 | -8.67 | 0.000 |
| D | | | | | | | -51.87 | 0.000 | -105.13 | 0.000 | -112.70 | 0.000 | -14.16 | 0.000 |
| E | | | | | | | | | -76.69 | 0.000 | -86.09 | 0.000 | -36.76 | 0.000 |
| F | | | | | | | | | | | -7.90 | 0.000 | -67.55 | 0.000 |
| G | | | | | | | | | | | | | -70.97 | 0.000 |

| **HS** | B | | C | | D | | E | | F | | G | | H | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -8.09 | 0.000 | -14.06 | 0.000 | -16.16 | 0.000 | -34.46 | 0.000 | -62.97 | 0.000 | -68.09 | 0.000 | -5.68 | 0.000 |
| B | | | -13.28 | 0.000 | -18.01 | 0.000 | -65.88 | 0.000 | -114.86 | 0.000 | -126.43 | 0.000 | **-0.92** | **0.179** |
| C | | | | | -4.89 | 0.000 | -52.81 | 0.000 | -106.64 | 0.000 | -118.79 | 0.000 | -7.66 | 0.000 |
| D | | | | | | | -47.16 | 0.000 | -102.64 | 0.000 | -114.87 | 0.000 | -10.06 | 0.000 |
| E | | | | | | | | | -77.53 | 0.000 | -92.16 | 0.000 | -31.05 | 0.000 |
| F | | | | | | | | | | | -12.02 | 0.000 | -63.92 | 0.000 |
| G | | | | | | | | | | | | | -69.97 | 0.000 |

| **NS** | B | | C | | D | | E | | F | | G | | H | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -6.67 | 0.000 | -12.52 | 0.000 | -15.02 | 0.000 | -33.47 | 0.000 | -63.84 | 0.000 | -70.06 | 0.000 | -6.67 | 0.000 |
| B | | | -12.96 | 0.000 | -18.60 | 0.000 | -66.57 | 0.000 | -118.37 | 0.000 | -132.06 | 0.000 | -2.20 | 0.014 |
| C | | | | | -5.90 | 0.000 | -53.93 | 0.000 | -110.37 | 0.000 | -124.66 | 0.000 | -4.29 | 0.000 |
| D | | | | | | | -46.82 | 0.000 | -104.86 | 0.000 | -119.18 | 0.000 | -7.19 | 0.000 |
| E | | | | | | | | | -79.87 | 0.000 | -96.76 | 0.000 | -27.98 | 0.000 |
| F | | | | | | | | | | | -13.76 | 0.000 | -62.72 | 0.000 |
| G | | | | | | | | | | | | | -69.97 | 0.000 |

| | ST_MO | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MA** | B | | C | | D | | E | | F | | G | | H | |
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -20.50 | 0.000 | -32.63 | 0.000 | -36.71 | 0.000 | -54.42 | 0.000 | -77.55 | 0.000 | -81.51 | 0.000 | -19.61 | 0.000 |
| B | | | -27.32 | 0.000 | -36.87 | 0.000 | -87.55 | 0.000 | -127.66 | 0.000 | -137.53 | 0.000 | -5.31 | 0.000 |
| C | | | | | -9.98 | 0.000 | -60.08 | 0.000 | -110.53 | 0.000 | -121.21 | 0.000 | -8.55 | 0.000 |
| D | | | | | | | -48.80 | 0.000 | -102.85 | 0.000 | -113.68 | 0.000 | -13.43 | 0.000 |
| E | | | | | | | | | -76.55 | 0.000 | -89.39 | 0.000 | -34.81 | 0.000 |
| F | | | | | | | | | | | -10.35 | 0.000 | -66.00 | 0.000 |
| G | | | | | | | | | | | | | -71.25 | 0.000 |

| | SEX | |
|---|---|---|
| **ES** | F | |
| | Statistic | p-value |
| M | -81.02 | 0.000 |
| **LC** | F | |
| | Statistic | p-value |
| M | -4.27 | 0.000 |
| **HS** | F | |
| | Statistic | p-value |
| M | -72.51 | 0.000 |
| **NS** | F | |
| | Statistic | p-value |
| M | -81.38 | 0.000 |
| **MA** | F | |
| | Statistic | p-value |
| M | -129.49 | 0.000 |

| SC_TP | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **ES** | B | | C | | D | | E | |
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -26.79 | 0.000 | -15.37 | 0.000 | -185.09 | 0.000 | -64.49 | 0.000 |
| B | | | -4.19 | 0.000 | -70.49 | 0.000 | -31.33 | 0.000 |
| C | | | | | -61.18 | 0.000 | -31.92 | 0.000 |
| D | | | | | | | -27.53 | 0.000 |
| **LC** | B | | C | | D | | E | |
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -47.43 | 0.000 | -19.90 | 0.000 | -208.39 | 0.000 | -64.61 | 0.000 |
| B | | | -13.82 | 0.000 | -67.55 | 0.000 | -17.90 | 0.000 |
| C | | | | | -69.06 | 0.000 | -29.35 | 0.000 |
| D | | | | | | | -40.94 | 0.000 |
| **HS** | B | | C | | D | | E | |
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -43.91 | 0.000 | -17.31 | 0.000 | -206.74 | 0.000 | -65.05 | 0.000 |
| B | | | -13.48 | 0.000 | -68.60 | 0.000 | -20.13 | 0.000 |
| C | | | | | -69.65 | 0.000 | -31.10 | 0.000 |
| D | | | | | | | -39.91 | 0.000 |
| **NS** | B | | C | | D | | E | |
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -48.02 | 0.000 | -20.73 | 0.000 | -223.81 | 0.000 | -69.23 | 0.000 |
| B | | | -13.08 | 0.000 | -76.36 | 0.000 | -21.58 | 0.000 |
| C | | | | | -74.24 | 0.000 | -31.47 | 0.000 |
| D | | | | | | | -44.81 | 0.000 |

| SC_TP | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **MA** | B | | C | | D | | E | |
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -36.45 | 0.000 | -14.34 | 0.000 | -208.60 | 0.000 | -62.95 | 0.000 |
| B | | | -10.91 | 0.000 | -76.82 | 0.000 | -24.32 | 0.000 |
| C | | | | | -72.33 | 0.000 | -31.77 | 0.000 |
| D | | | | | | | -41.77 | 0.000 |

| RACE | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **ES** | B | | C | | D | | E | | F | |
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -12.41 | 0.000 | -50.57 | 0.000 | -22.76 | 0.000 | -5.16 | 0.000 | -20.34 | 0.000 |
| B | | | -61.99 | 0.000 | -18.05 | 0.000 | -7.71 | 0.000 | -14.54 | 0.000 |
| C | | | | | **-0.78** | **0.219** | -14.41 | 0.000 | -7.25 | 0.000 |
| D | | | | | | | -13.39 | 0.000 | -5.11 | 0.000 |
| E | | | | | | | | | -11.66 | 0.000 |
| **LC** | B | | C | | D | | E | | F | |
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -10.77 | 0.000 | -53.07 | 0.000 | -25.49 | 0.000 | -4.70 | 0.000 | -32.94 | 0.000 |
| B | | | -69.17 | 0.000 | -21.81 | 0.000 | -6.95 | 0.000 | -29.82 | 0.000 |
| C | | | | | -2.62 | 0.004 | -14.55 | 0.000 | -7.18 | 0.000 |
| D | | | | | | | -14.12 | 0.000 | -2.44 | 0.007 |
| E | | | | | | | | | -15.32 | 0.000 |

| **RACE** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **HS** | B | | C | | D | | E | | F | |
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -15.36 | 0.000 | -55.75 | 0.000 | -28.11 | 0.000 | -2.17 | 0.015 | -35.83 | 0.000 |
| B | | | -65.70 | 0.000 | -22.49 | 0.000 | -5.48 | 0.000 | -30.37 | 0.000 |
| C | | | | | -4.49 | 0.000 | -12.99 | 0.000 | -8.81 | 0.000 |
| D | | | | | | | -13.26 | 0.000 | -2.01 | 0.022 |
| E | | | | | | | | | -14.52 | 0.000 |
| **NS** | B | | C | | D | | E | | F | |
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -17.55 | 0.000 | -61.75 | 0.000 | -32.41 | 0.000 | **-0.33** | **0.370** | -35.87 | 0.000 |
| B | | | -71.43 | 0.000 | -26.15 | 0.000 | -3.45 | 0.000 | -28.81 | 0.000 |
| C | | | | | -7.08 | 0.000 | -11.73 | 0.000 | -4.80 | 0.000 |
| D | | | | | | | -12.82 | 0.000 | -2.63 | 0.004 |
| E | | | | | | | | | -12.34 | 0.000 |
| **MA** | B | | C | | D | | E | | F | |
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| A | -25.98 | 0.000 | -73.93 | 0.000 | -39.00 | 0.000 | **-1.32** | **0.094** | -37.80 | 0.000 |
| B | | | -78.16 | 0.000 | -29.50 | 0.000 | -4.24 | 0.000 | -26.08 | 0.000 |
| C | | | | | -8.43 | 0.000 | -13.14 | 0.000 | **-0.50** | **0.309** |
| D | | | | | | | -14.66 | 0.000 | -6.76 | 0.000 |
| E | | | | | | | | | -12.09 | 0.000 |

| | AGE | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **ES** | 17 | | 18 | | 19 | | 20 | | 21 | | 22 | | 23 | | 24 | | 25 | |
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| 16 | -2.47 | 0.007 | -5.77 | 0.000 | -10.61 | 0.000 | -19.26 | 0.000 | -24.15 | 0.000 | -28.36 | 0.000 | -29.59 | 0.000 | -28.68 | 0.000 | -25.98 | 0.000 |
| 17 | | | -11.06 | 0.000 | -16.08 | 0.000 | -25.63 | 0.000 | -30.14 | 0.000 | -33.93 | 0.000 | -34.33 | 0.000 | -32.50 | 0.000 | -29.02 | 0.000 |
| 18 | | | | | -6.44 | 0.000 | -16.53 | 0.000 | -21.76 | 0.000 | -26.17 | 0.000 | -27.32 | 0.000 | -26.20 | 0.000 | -23.39 | 0.000 |
| 19 | | | | | | | -9.66 | 0.000 | -15.13 | 0.000 | -19.85 | 0.000 | -21.59 | 0.000 | -21.09 | 0.000 | -18.90 | 0.000 |
| 20 | | | | | | | | | -5.86 | 0.000 | -11.03 | 0.000 | -13.48 | 0.000 | -13.71 | 0.000 | -12.22 | 0.000 |
| 21 | | | | | | | | | | | -5.30 | 0.000 | -8.11 | 0.000 | -8.80 | 0.000 | -7.75 | 0.000 |
| 22 | | | | | | | | | | | | | -3.05 | 0.001 | -4.08 | 0.000 | -3.39 | 0.000 |
| 23 | | | | | | | | | | | | | | | **-1.17** | **0.120** | **-0.69** | **0.246** |
| 24 | | | | | | | | | | | | | | | | | **-0.39** | **0.347** |
| **LC** | 17 | | 18 | | 19 | | 20 | | 21 | | 22 | | 23 | | 24 | | 25 | |
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| 16 | -13.87 | 0.000 | -23.50 | 0.000 | -16.30 | 0.000 | -13.66 | 0.000 | -12.10 | 0.000 | -9.26 | 0.000 | -8.37 | 0.000 | -4.28 | 0.000 | -3.11 | 0.001 |
| 17 | | | -14.34 | 0.000 | -5.74 | 0.000 | -3.23 | 0.001 | -2.24 | 0.013 | **-0.24** | **0.404** | **-0.43** | **0.334** | -4.12 | 0.000 | -4.67 | 0.000 |
| 18 | | | | | -5.83 | 0.000 | -7.01 | 0.000 | -6.87 | 0.000 | -8.43 | 0.000 | -7.81 | 0.000 | -10.70 | 0.000 | -10.68 | 0.000 |
| 19 | | | | | | | **-1.61** | **0.054** | -2.01 | 0.022 | -3.88 | 0.000 | -3.73 | 0.000 | -6.86 | 0.000 | -7.18 | 0.000 |
| 20 | | | | | | | | | **-0.52** | **0.303** | -2.41 | 0.008 | -2.40 | 0.008 | -5.57 | 0.000 | -5.98 | 0.000 |
| 21 | | | | | | | | | | | -1.84 | 0.033 | -1.88 | 0.030 | -4.98 | 0.000 | -5.44 | 0.000 |
| 22 | | | | | | | | | | | | | **-0.16** | **0.436** | -3.26 | 0.001 | -3.81 | 0.000 |
| 23 | | | | | | | | | | | | | | | -2.96 | 0.002 | -3.51 | 0.000 |
| 24 | | | | | | | | | | | | | | | | | **-0.70** | **0.243** |
| **HS** | 17 | | 18 | | 19 | | 20 | | 21 | | 22 | | 23 | | 24 | | 25 | |
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| 16 | -6.86 | 0.000 | -21.63 | 0.000 | -19.10 | 0.000 | -20.49 | 0.000 | -20.25 | 0.000 | -17.97 | 0.000 | -15.93 | 0.000 | -10.97 | 0.000 | -8.17 | 0.000 |
| 17 | | | -20.60 | 0.000 | -16.78 | 0.000 | -18.19 | 0.000 | -17.63 | 0.000 | -14.98 | 0.000 | -12.80 | 0.000 | -7.60 | 0.000 | -4.76 | 0.000 |
| 18 | | | | | **-0.21** | **0.416** | -3.01 | 0.001 | -4.07 | 0.000 | -2.76 | 0.003 | -1.87 | 0.031 | -2.08 | 0.019 | -4.08 | 0.000 |
| 19 | | | | | | | -2.85 | 0.002 | -3.85 | 0.000 | -2.69 | 0.004 | -1.88 | 0.030 | -1.83 | 0.033 | -3.79 | 0.000 |
| 20 | | | | | | | | | **-1.16** | **0.124** | **-0.25** | **0.402** | **-0.32** | **0.375** | -3.71 | 0.000 | -5.47 | 0.000 |
| 21 | | | | | | | | | | | **-0.79** | **0.214** | **-1.28** | **0.101** | -4.48 | 0.000 | -6.15 | 0.000 |
| 22 | | | | | | | | | | | | | **-0.52** | **0.303** | -3.66 | 0.000 | -5.33 | 0.000 |
| 23 | | | | | | | | | | | | | | | -3.03 | 0.001 | -4.65 | 0.000 |
| 24 | | | | | | | | | | | | | | | | | -1.70 | 0.045 |

|  | AGE |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **NS** | 17 |  | 18 |  | 19 |  | 20 |  | 21 |  | 22 |  | 23 |  | 24 |  | 25 |  |
|  | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| 16 | -12.78 | 0.000 | -24.33 | 0.000 | -21.80 | 0.000 | -22.58 | 0.000 | -22.26 | 0.000 | -20.55 | 0.000 | -18.08 | 0.000 | -15.64 | 0.000 | -11.89 | 0.000 |
| 17 |  |  | -16.58 | 0.000 | -13.77 | 0.000 | -14.96 | 0.000 | -14.81 | 0.000 | -13.01 | 0.000 | -10.72 | 0.000 | -8.61 | 0.000 | -5.13 | 0.000 |
| 18 |  |  |  |  | **-0.04** | **0.485** | -2.63 | 0.004 | -3.77 | 0.000 | -2.87 | 0.002 | **-1.59** | **0.056** | **-0.43** | **0.333** | -2.23 | 0.013 |
| 19 |  |  |  |  |  |  | -2.31 | 0.011 | -3.38 | 0.000 | -2.58 | 0.005 | **-1.41** | **0.079** | **-0.35** | **0.362** | -2.18 | 0.015 |
| 20 |  |  |  |  |  |  |  |  | **-1.19** | **0.117** | **-0.54** | **0.293** | **-0.43** | **0.334** | **-1.30** | **0.096** | -3.64 | 0.000 |
| 21 |  |  |  |  |  |  |  |  |  |  | **-0.55** | **0.290** | **-1.41** | **0.079** | -2.20 | 0.014 | -4.42 | 0.000 |
| 22 |  |  |  |  |  |  |  |  |  |  |  |  | **-0.88** | **0.188** | -1.71 | 0.044 | -3.94 | 0.000 |
| 23 |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **-0.84** | **0.201** | -3.04 | 0.001 |
| 24 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | -2.16 | 0.015 |
| **MA** | 17 |  | 18 |  | 19 |  | 20 |  | 21 |  | 22 |  | 23 |  | 24 |  | 25 |  |
|  | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| 16 | -14.93 | 0.000 | -27.87 | 0.000 | -27.46 | 0.000 | -25.51 | 0.000 | -25.67 | 0.000 | -24.32 | 0.000 | -23.09 | 0.000 | -22.16 | 0.000 | -17.81 | 0.000 |
| 17 |  |  | -18.89 | 0.000 | -18.93 | 0.000 | -16.79 | 0.000 | -17.51 | 0.000 | -16.26 | 0.000 | -15.31 | 0.000 | -14.89 | 0.000 | -10.77 | 0.000 |
| 18 |  |  |  |  | -3.10 | 0.001 | -2.80 | 0.003 | -5.17 | 0.000 | -5.05 | 0.000 | -5.18 | 0.000 | -5.83 | 0.000 | -2.67 | 0.004 |
| 19 |  |  |  |  |  |  | **-0.03** | **0.489** | -2.51 | 0.006 | -2.64 | 0.004 | -2.97 | 0.001 | -3.81 | 0.000 | **-1.01** | **0.157** |
| 20 |  |  |  |  |  |  |  |  | -2.33 | 0.010 | -2.47 | 0.007 | -2.83 | 0.002 | -3.66 | 0.000 | **-0.95** | **0.171** |
| 21 |  |  |  |  |  |  |  |  |  |  | **-0.30** | **0.382** | **-0.79** | **0.214** | -1.76 | 0.039 | **-0.69** | **0.244** |
| 22 |  |  |  |  |  |  |  |  |  |  |  |  | **-0.50** | **0.309** | **-1.46** | **0.072** | **-0.91** | **0.181** |
| 23 |  |  |  |  |  |  |  |  |  |  |  |  |  |  | **-0.95** | **0.171** | **-1.30** | **0.097** |
| 24 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | -2.07 | 0.019 |

# APPENDIX D – sMC AND VIP VALUES

| | ES | | LC | | HS | | NS | | MA | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | sMC | VIP | sMC | VIP | sMC | VIP | sMC | VIP | sMC | VIP |
| AGE | 0.021 | - | 0.002 | - | 0.001 | - | 0.000 | - | 0.006 | - |
| SEX_M | 5707.661 | 1.664 | 503.227 | 0.046 | 925.951 | 1.271 | 1707.030 | 1.434 | 10437.647 | 2.324 |
| RACE_B | 1047.005 | 0.878 | 1344.721 | 0.889 | 1632.630 | 0.803 | 1275.279 | 0.861 | 3030.721 | 0.961 |
| RACE_C | 710.864 | 1.323 | 1058.677 | 1.211 | 1285.917 | 1.137 | 1200.583 | 1.209 | 3105.124 | 1.382 |
| RACE_D | 11593.757 | 0.025 | 16466.935 | 0.035 | 19703.270 | 0.034 | 36561.891 | 0.051 | 98264.323 | 0.064 |
| RACE_E | 21564.101 | 0.024 | 17058.889 | 0.024 | 9457.157 | 0.019 | 2381.204 | 0.016 | 684.578 | 0.016 |
| RACE_F | 25009.480 | 0.018 | 70993.796 | 0.079 | 74163.562 | 0.079 | 57654.616 | 0.060 | 80455.299 | 0.041 |
| LG_E | 1383.472 | 2.235 | 4184.511 | 3.047 | 3729.836 | 2.857 | 1891.869 | 2.490 | 2224.961 | 2.702 |
| SC_TP_B | 1432.920 | 0.065 | 5083.937 | 0.197 | 3781.141 | 0.166 | 11159.206 | 0.174 | 38.052 | 0.098 |
| SC_TP_C | 24589.357 | 0.006 | 18733.479 | 0.026 | 16156.491 | 0.010 | 24847.770 | 0.013 | 4811.679 | 0.024 |
| SC_TP_D | 61531.910 | 2.719 | 42392.906 | 2.645 | 50224.653 | 2.545 | 96333.155 | 2.775 | 68178.287 | 2.733 |
| SC_TP_E | 131000.270 | 0.349 | 94830.983 | 0.292 | 107824.185 | 0.289 | 120678.728 | 0.299 | 109020.004 | 0.263 |
| HS_ST_B | 14069.296 | 0.102 | 26986.878 | 0.142 | 22023.232 | 0.084 | 18324.977 | 0.013 | 14877.479 | 0.138 |
| HS_ST_C | 12.547 | 0.152 | 4527.229 | 1.046 | 2786.064 | 0.646 | 2587.247 | 0.632 | 1608.477 | 0.312 |
| HS_ST_D | 7894.582 | 0.632 | 1603.787 | 0.133 | 988.745 | 0.077 | 423.965 | 0.131 | 88.804 | 0.149 |
| HS_ST_E | 10014.609 | 0.238 | 123.105 | 0.117 | 135.837 | 0.048 | 13.123 | 0.068 | 375.414 | 0.073 |
| HS_ST_F | 7257.196 | 0.094 | 0.553 | 0.121 | 260.977 | 0.027 | 1.107 | 0.059 | 1209.375 | 0.081 |
| HS_ST_G | 7923.490 | 0.003 | 367.876 | 0.107 | 0.047 | 0.039 | 312.924 | 0.046 | 1272.146 | 0.033 |
| HS_ST_H | 4577.567 | 0.063 | 4479.041 | 0.091 | 581.180 | 0.025 | 1267.683 | 0.023 | 2689.300 | 0.007 |
| HS_ST_I | 11542.467 | 0.047 | 25166.059 | 0.094 | 16059.173 | 0.053 | 13124.350 | 0.038 | 17066.190 | 0.018 |
| HS_ST_J | 6644.407 | 0.073 | 25988.196 | 0.052 | 16351.714 | 0.019 | 6949.886 | 0.004 | 14084.757 | 0.015 |
| HS_ST_K | 5206.708 | 0.063 | 44421.116 | 0.071 | 33958.897 | 0.044 | 27003.001 | 0.031 | 24311.970 | 0.002 |
| HS_ST_L | 286.582 | 0.060 | 46484.250 | 0.056 | 44322.665 | 0.040 | 37369.065 | 0.032 | 30377.192 | 0.002 |
| HS_ST_M | 1764.209 | 0.065 | 65084.969 | 0.058 | 61784.685 | 0.045 | 45627.129 | 0.032 | 36519.176 | 0.003 |
| HS_ST_N | 390.878 | 0.510 | 15976.883 | 0.362 | 19218.037 | 0.377 | 8728.301 | 0.221 | 2016.293 | 0.110 |
| IN_B | 177.836 | 0.561 | 0.112 | - | 31.966 | 0.588 | 58.123 | 0.571 | 67.034 | 0.619 |
| IN_C | 482.010 | 0.316 | 490.964 | 0.233 | 542.780 | 0.258 | 108.063 | 0.335 | 725.413 | 0.279 |
| IN_D | 3453.209 | 0.004 | 2206.716 | 0.053 | 2571.879 | 0.027 | 1490.679 | 0.015 | 4206.395 | 0.041 |
| IN_E | 5806.535 | 0.272 | 3546.197 | 0.313 | 4403.106 | 0.293 | 3300.978 | 0.262 | 5657.622 | 0.283 |
| IN_F | 9639.088 | 0.365 | 5303.903 | 0.391 | 7663.117 | 0.390 | 5586.333 | 0.371 | 10830.755 | 0.400 |

| | ES | | LC | | HS | | NS | | MA | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | sMC | VIP | sMC | VIP | sMC | VIP | sMC | VIP | sMC | VIP |
| IN_G | 26517.171 | 0.466 | 15301.390 | 0.468 | 19204.703 | 0.462 | 16440.650 | 0.471 | 29878.199 | 0.499 |
| IN_H | 46621.200 | 0.372 | 27893.967 | 0.367 | 30305.388 | 0.356 | 30042.985 | 0.376 | 50760.857 | 0.384 |
| IN_I | 84211.711 | 0.284 | 53415.224 | 0.276 | 66436.049 | 0.277 | 60112.693 | 0.281 | 93557.880 | 0.290 |
| IN_J | 86719.069 | 0.244 | 51373.931 | 0.230 | 61927.267 | 0.229 | 59239.973 | 0.240 | 98250.563 | 0.250 |
| IN_K | 81037.432 | 0.211 | 51864.272 | 0.205 | 57831.113 | 0.203 | 47766.434 | 0.209 | 98312.456 | 0.224 |
| IN_L | 72142.375 | 0.335 | 55307.635 | 0.338 | 57024.343 | 0.327 | 39170.403 | 0.330 | 80772.648 | 0.344 |
| IN_M | 89471.452 | 0.258 | 60194.232 | 0.251 | 74147.951 | 0.251 | 60496.160 | 0.257 | 107846.629 | 0.272 |
| IN_N | 93864.827 | 0.218 | 87224.119 | 0.225 | 92144.538 | 0.221 | 89390.513 | 0.233 | 123932.348 | 0.243 |
| IN_O | 98471.260 | 0.322 | 84698.719 | 0.325 | 80476.427 | 0.318 | 74485.172 | 0.336 | 128158.089 | 0.370 |
| ST_FA_B | 1.151 | 1.389 | 30.253 | 1.280 | 25.981 | 1.093 | 107.988 | 1.092 | 1.498 | - |
| ST_FA_C | 463.356 | 0.678 | 1.202 | - | 3.291 | - | 3.619 | 0.714 | 716.381 | 0.683 |
| ST_FA_D | 1032.656 | 0.245 | 199.501 | 0.416 | 97.712 | 0.390 | 1.604 | - | 945.613 | 0.350 |
| ST_FA_E | 1197.786 | 0.832 | 778.059 | 0.839 | 343.429 | 0.597 | 196.894 | 0.510 | 1507.711 | 0.700 |
| ST_FA_F | 9011.570 | 1.132 | 10881.417 | 1.186 | 8302.974 | 1.121 | 5956.667 | 1.126 | 16717.112 | 1.155 |
| ST_FA_G | 13199.381 | 1.045 | 20088.187 | 1.062 | 20370.797 | 1.033 | 19044.664 | 1.097 | 42846.770 | 1.132 |
| ST_FA_H | 159.514 | 0.456 | 653.667 | 0.354 | 3.526 | - | 15.857 | 0.398 | 1028.480 | 0.426 |
| ST_MO_B | 26.475 | 1.226 | 9.946 | 1.107 | 0.113 | - | 46.261 | 0.944 | 0.487 | - |
| ST_MO_C | 1268.739 | 0.765 | 480.898 | 0.863 | 577.300 | 0.767 | 87.164 | 0.784 | 1773.949 | 0.801 |
| ST_MO_D | 1273.469 | 0.574 | 534.389 | 0.690 | 641.018 | 0.686 | 130.287 | 0.685 | 2214.518 | 0.632 |
| ST_MO_E | 1786.032 | 0.502 | 1240.794 | 0.545 | 1241.614 | 0.297 | 529.467 | 0.216 | 3515.436 | 0.406 |
| ST_MO_F | 12593.689 | 1.146 | 9861.313 | 1.193 | 12265.066 | 1.109 | 6844.746 | 1.140 | 26149.027 | 1.198 |
| ST_MO_G | 10609.565 | 1.366 | 5320.634 | 1.294 | 10132.207 | 1.285 | 5000.721 | 1.339 | 19670.720 | 1.380 |
| ST_MO_H | 4102.886 | 0.206 | 8322.719 | 0.184 | 495.254 | 0.154 | 3599.269 | 0.136 | 604.665 | 0.157 |
| OC_FA_B | 4.448 | - | 1034.147 | 0.658 | 170.316 | 0.811 | 400.669 | 0.870 | 447.698 | 0.861 |
| OC_FA_C | 4.568 | 0.915 | 1237.831 | 0.761 | 6.618 | - | 32.133 | 1.017 | 1.674 | - |
| OC_FA_D | 1968.136 | 1.879 | 5879.808 | 1.957 | 720.202 | 1.743 | 547.822 | 1.729 | 476.911 | 1.748 |
| OC_FA_E | 11204.561 | 1.295 | 17916.088 | 1.291 | 1390.474 | 1.244 | 1844.547 | 1.314 | 1020.781 | 1.376 |
| OC_FA_F | 543.294 | 0.511 | 480.806 | 0.412 | 930.318 | 0.452 | 868.136 | 0.449 | 1635.239 | 0.477 |
| OC_MO_B | 33.613 | 1.606 | 0.142 | - | 397.502 | 1.769 | 659.581 | 1.920 | 308.440 | 1.888 |
| OC_MO_C | 34.513 | 0.271 | 2.496 | 0.244 | 574.074 | 0.278 | 1374.478 | 0.333 | 119.449 | 0.210 |

|  | ES |  | LC |  | HS |  | NS |  | MA |  |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | sMC | VIP | sMC | VIP | sMC | VIP | sMC | VIP | sMC | VIP |
| OC_MO_B | 33.613 | 1.606 | 0.142 | - | 397.502 | 1.769 | 659.581 | 1.920 | 308.440 | 1.888 |
| OC_MO_C | 34.513 | 0.271 | 2.496 | 0.244 | 574.074 | 0.278 | 1374.478 | 0.333 | 119.449 | 0.210 |
| OC_MO_D | 55.316 | 2.067 | 239.024 | 2.077 | 141.437 | 1.920 | 160.343 | 1.929 | 90.253 | 1.959 |
| OC_MO_E | 4887.345 | 0.876 | 5422.684 | 0.864 | 235.465 | 0.834 | 206.104 | 0.884 | 1034.041 | 0.938 |
| OC_MO_F | 918.173 | 0.268 | 344.520 | 0.172 | 2268.422 | 0.154 | 2025.276 | 0.108 | 1704.792 | 0.150 |
| RES_B | 136.352 | 0.154 | 125.977 | 0.203 | 368.324 | 0.022 | 82.215 | 0.040 | 252.366 | 0.140 |
| RES_C | 243.045 | 0.185 | 379.392 | 0.176 | 409.870 | 0.144 | 199.451 | 0.156 | 391.596 | 0.116 |
| RES_D | 541.576 | 0.604 | 798.395 | 0.210 | 778.683 | 0.310 | 394.415 | 0.382 | 715.099 | 0.475 |
| RES_E | 3043.909 | 0.211 | 2846.378 | 0.330 | 2656.288 | 0.241 | 1916.636 | 0.229 | 1839.216 | 0.152 |
| RES_F | 12362.599 | 0.189 | 10717.611 | 0.218 | 10595.147 | 0.200 | 7986.501 | 0.201 | 7567.294 | 0.168 |
| RES_G | 26102.611 | 0.079 | 32949.605 | 0.101 | 26875.825 | 0.089 | 32463.848 | 0.098 | 31316.037 | 0.087 |
| RES_H | 49570.125 | 0.040 | 16475.540 | 0.033 | 37154.851 | 0.039 | 29484.786 | 0.038 | 49278.447 | 0.038 |
| RES_I | 34538.145 | 0.016 | 22210.365 | 0.016 | 10688.468 | 0.012 | 28809.450 | 0.017 | 25621.635 | 0.013 |
| RES_J | 16469.612 | 0.012 | 27995.284 | 0.018 | 21079.811 | 0.014 | 19525.325 | 0.014 | 15995.851 | 0.012 |
| CAR_Y | 51.440 | 1.534 | 137.435 | 1.167 | 84.641 | 1.372 | 5.759 | - | 1.709 | - |
| MOTO_Y | 782.693 | 0.624 | 897.353 | 0.764 | 483.053 | 0.494 | 118.897 | 0.351 | 170.984 | 0.271 |
| CELL_B | 144.531 | 1.243 | 385.733 | 1.128 | 207.617 | 0.932 | 42.578 | 0.800 | 229.412 | 1.056 |
| PC_Y | 876.295 | 2.050 | 508.545 | 2.092 | 873.232 | 2.103 | 592.293 | 1.962 | 445.563 | 2.006 |
| INT_Y | 277.776 | 1.667 | 8.799 | 1.548 | 29.013 | 1.476 | 5.423 | 1.359 | 0.405 | - |
| CB_TV_Y | 0.093 | - | 3.909 | - | 30.433 | 1.894 | 105.228 | 1.842 | 27.854 | 2.007 |
| TV_B | 0.036 | - | 19.402 | 2.147 | 8.412 | 1.862 | 17.001 | 1.846 | 3.570 | - |
| DVD_Y | 0.980 | - | 0.098 | - | 4.031 | - | 2.143 | - | 0.114 | - |
| TP_Y | 1.436 | 2.071 | 109.416 | 2.385 | 20.586 | 2.119 | 24.231 | 2.099 | 2.541 | - |
| HK_Y | 64.945 | 1.182 | 133.760 | 1.097 | 2.619 | - | 74.879 | 1.204 | 709.148 | 1.292 |
| BATH_B | 153.060 | 2.790 | 38.344 | 2.585 | 98.219 | 2.582 | 179.439 | 2.714 | 179.179 | 2.877 |
| BED_B | 6.046 | - | 41.084 | 0.171 | 0.029 | - | 1.420 | - | 49.652 | 0.373 |
| MW_Y | 6.584 | - | 3.343 | - | 49.486 | 1.321 | 32.603 | 1.298 | 0.017 | - |
| FGE_B | 35.803 | 0.526 | 142.392 | 0.469 | 0.610 | - | 1.901 | - | 63.827 | 0.596 |
| FZR_Y | 614.163 | 2.169 | 346.997 | 1.626 | 372.921 | 1.645 | 238.805 | 1.601 | 649.861 | 1.978 |
| WA_MA_Y | 26.885 | 0.946 | 16.848 | 0.894 | 14.325 | 0.843 | 3.445 | - | 4.656 | - |
| DR_MA_Y | 3.769 | - | 2.560 | - | 0.885 | - | 19.411 | 1.171 | 1.217 | - |
| DI_WA_Y | 899.557 | 0.723 | 455.690 | 0.727 | 139.645 | 0.701 | 1370.592 | 0.769 | 1289.074 | 0.796 |
| VA_CL_Y | 15.740 | 2.350 | 100.568 | 2.492 | 161.668 | 2.395 | 196.229 | 2.436 | 188.340 | 2.597 |