Método de regressão linear aplicado ao dataset *Students Performance in Exams*

 João Cunha (202103227)

Roberto Pinto (202104006)

1. **Introdução**

O sucesso escolar no ensino secundário é um tema com bastante relevância nos dias de hoje. Sendo um requisito praticamente obrigatória nos países da Europa Ocidental, este ensino situa-se entre o ensino universitário e a escolaridade básica. Sendo que, os sistemas educativos atuais são muito importantes para os jovens de todo o mundo, tendo um impacto crucial na vida dos estudantes, através da qual são feitas escolhas importantíssimas para o seu futuro pessoal e profissional.

Embora o estudo das variáveis associadas ao desempenho escolar tenha sido, historicamente, uma preocupação mundial, como é referido na publicação de Coleman (1966) (1), o papel central das variáveis socioeconômicas e a relevância das práticas e políticas escolares, deu início a uma pesquisa linha cuja relevância se estende por mais de cinco décadas, sendo ainda muito relevante nos dias de hoje. Apesar de, existirem muitas fontes diferentes de dados para a realização deste tipo de estudos com variáveis relacionadas ao desempenho dos alunos, as avaliações em larga escala demonstraram ser uma fonte valiosa, devido ao grande volume de variáveis e observações que oferecem aos investigadores. (2)

Para a avaliação destes fatores condicionam os resultados de grande maioria dos estudantes é possível recorrer a vários tipos de análises. Neste trabalho recorremos à utilização da Regressão Linear, avaliando 8 variáveis características de todo o tipo de alunos, mais concretamente: Género, Etnicidade, Educação Parental, Almoço, Curso de preparação para o teste, Notas de matemática, Pontuação de leitura e Pontuação de escrita. Esta mesma análise permitiu-nos retirar conclusões muito relevantes em relação à dependência e efeito destas variáveis no valor das “Notas de Matemática”, tomando-a como o *target* deste modelo estatístico.

1. **Descrição do dataset *Students Performance in Exams***

Este conjunto de dados consiste numa amostra de 1000 alunos e pretende perceber o desempenho dos estudantes em várias disciplinas tendo em conta vários aspetos, como a preparação de cada aluno para um teste e a situação académica familiar.

As variáveis descritas neste dataset são:

* Género: Masculino ou feminino
* Etnicidade: 5 grupos (A, B, C, D, E)
* Educação Parental: 6 grupos (Algum Secundário, Ensino Secundário, Diploma Associado, Alguma Faculdade, Licenciatura, Mestrado)
* Almoço: 2 grupos (Grátis/Reduzido, Padrão)
* Curso de preparação para o teste: 2 grupos (Completo, Nenhum)
* Notas de matemática: de 0 a 100
* Pontuação de leitura: de 0 a 100
* Pontuação de escrita: de 0 a 100

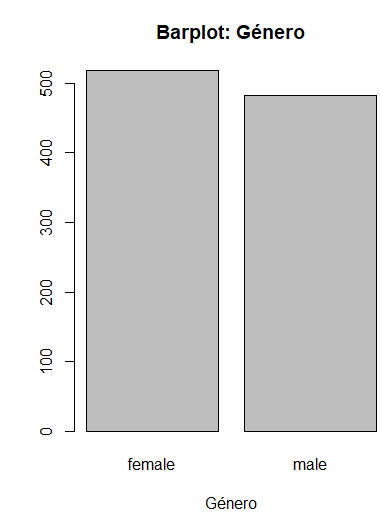
Assim sendo, este trabalho tem como finalidade perceber o impacto que todas as variáveis descritas têm nas notas de matemática e conseguir um modelo de regressão linear que demonstre a relação entre as várias variáveis.

Uma imagem com texto, computador, captura de ecrã, portátil

Descrição gerada automaticamente

Figura 1:Dataset Students Performance in Exams

1. **Descrição das variáveis explicativas** 
   1. **Género**

****

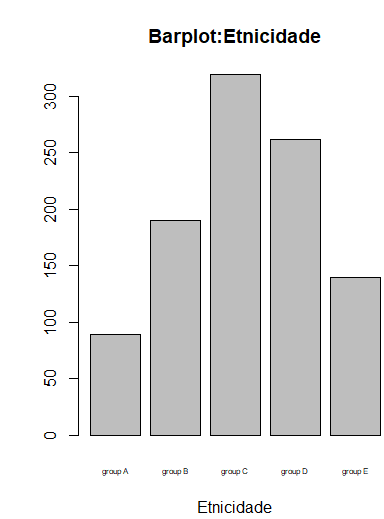
|  |  |
| --- | --- |
| **Frequências absolutas e relativas** | |
| **Feminino** | **Masculino** |
| 518 (52%) | 482 (48 %) |

Tabela :Descrição do Género

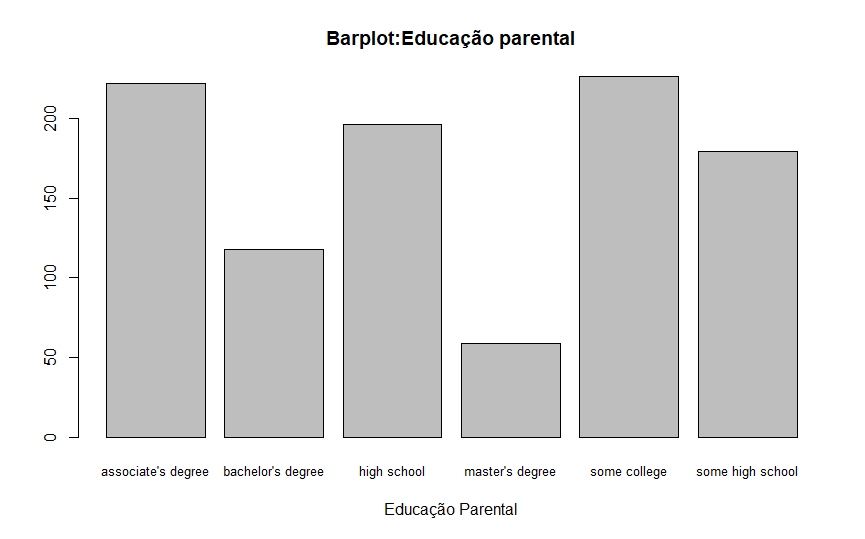
* 1. **Etnicidade**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Frequências absolutas e relativas** | | | | |
| **A** | **B** | **C** | **D** | **E** |
| 89 (9%) | 190 (19%) | 319 (32%) | 262 (26%) | 140 (14%) |

Tabela :Descrição da Etnicidade

****

* 1. **Educação Parental**

****

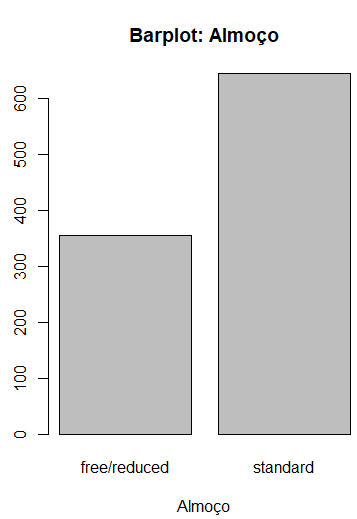
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Frequências absolutas e relativas** | | | | | |
| **Diploma Associado** | **Licenciatura** | **Ensino Secundário** | **Mestrado** | **Alguma Faculdade** | **Algum Secundário** |
| 222 (22%) | 118 (12%) | 196 (20%) | 59 (6%) | 226 (22%) | 179 (18%) |

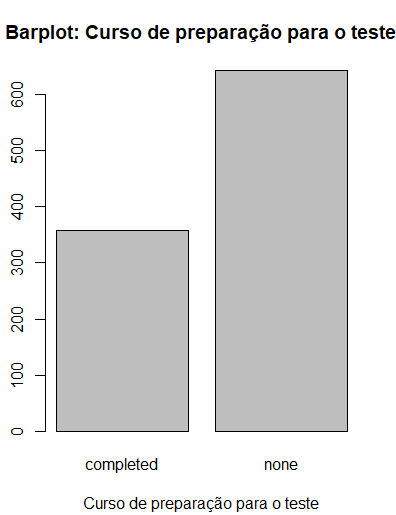
Tabela :Descrição da Educação Parental

* 1. **Almoço**

|  |  |
| --- | --- |
| **Frequências absolutas e relativas** | |
| **Grátis/Reduzido** | **Padrão** |
| 355 (36%) | 645 (64 %) |

Tabela :Descrição do Almoço

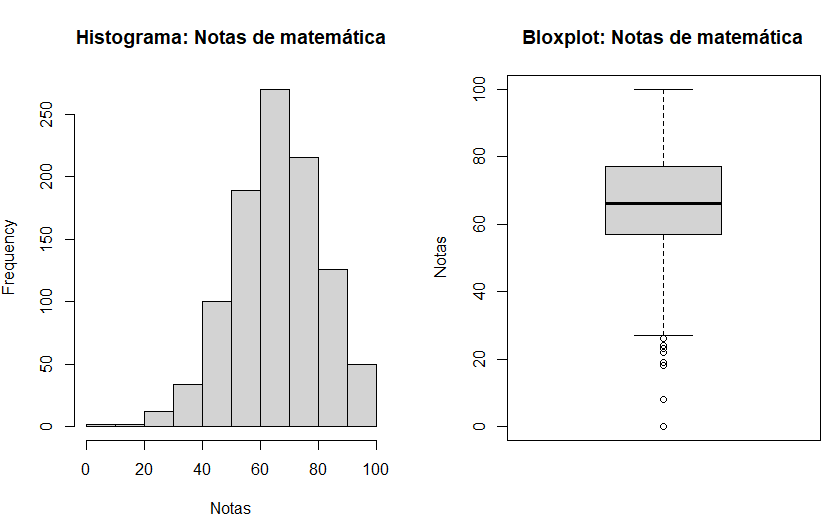
****

* 1. **Curso de Preparação para o teste**

|  |  |
| --- | --- |
| **Frequências absolutas e relativas** | |
| **Completo** | **Nenhum** |
| 358 (36%) | 642 (64%) |

Tabela :Descrição do Curso de Preparação para o Teste

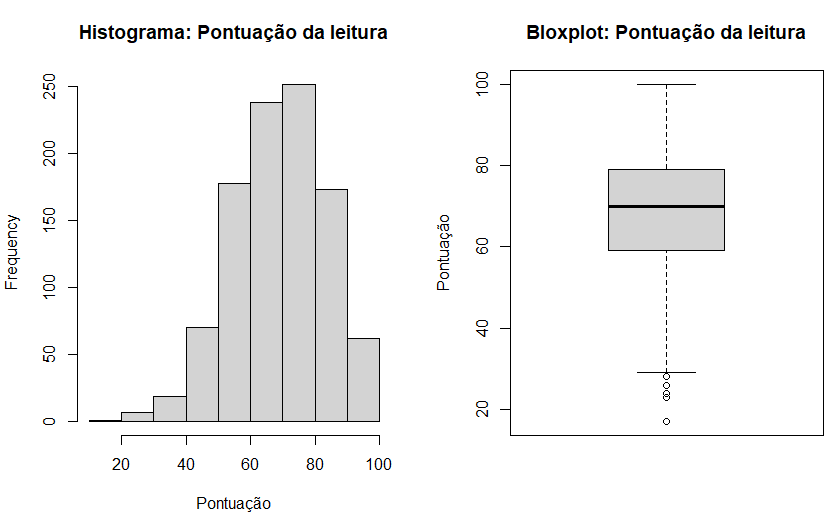
* 1. **Notas de Matemática**

****

|  |  |  |
| --- | --- | --- |
| **Mínimo:** | 0.00 | |
| **Máximo:** | 100.00 | |
| **Mediana:** | 66.00 | |
| **Média:** | 66.089 | |
| **Desvio padrão:** | | 15.16308 |

Tabela :Descrição das Notas de Matemática

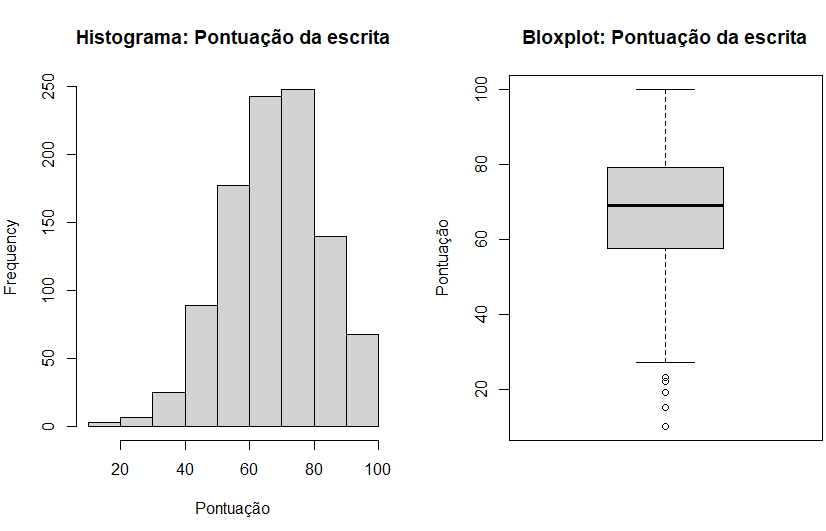
* 1. **Pontuação de Leitura**

****

|  |  |  |
| --- | --- | --- |
| **Mínimo:** | 0.00 | |
| **Máximo:** | 100.00 | |
| **Mediana:** | 70.00 | |
| **Média:** | 69.169 | |
| **Desvio padrão:** | | 14.60019 |

Tabela :Descrição da Pontuação de Leitura

* 1. **Pontuação de Escrita**

****

|  |  |  |
| --- | --- | --- |
| **Mínimo:** | 0.00 | |
| **Máximo:** | 100.00 | |
| **Mediana:** | 69.00 | |
| **Média:** | 68.054 | |
| **Desvio padrão:** | | 15.19566 |

Tabela :Descrição da Pontuação de Escrita

1. **Correlação entre variáveis**

Para realizar a correlação entre todas as variáveis desta amostra, ou seja 8 variáveis (Género, Etnicidade, Educação Parental, Almoço, Curso de preparação para o teste, Notas de matemática, Pontuação de leitura, Pontuação de escrita), foi necessário transformar as 5 variáveis categóricas em variáveis numéricas, com o recurso à função *as.numeric().*

Uma imagem com texto

Descrição gerada automaticamente

Apenas após este passo foi possível realizar a correlação entre as variáveis.

Uma imagem com texto

Descrição gerada automaticamente

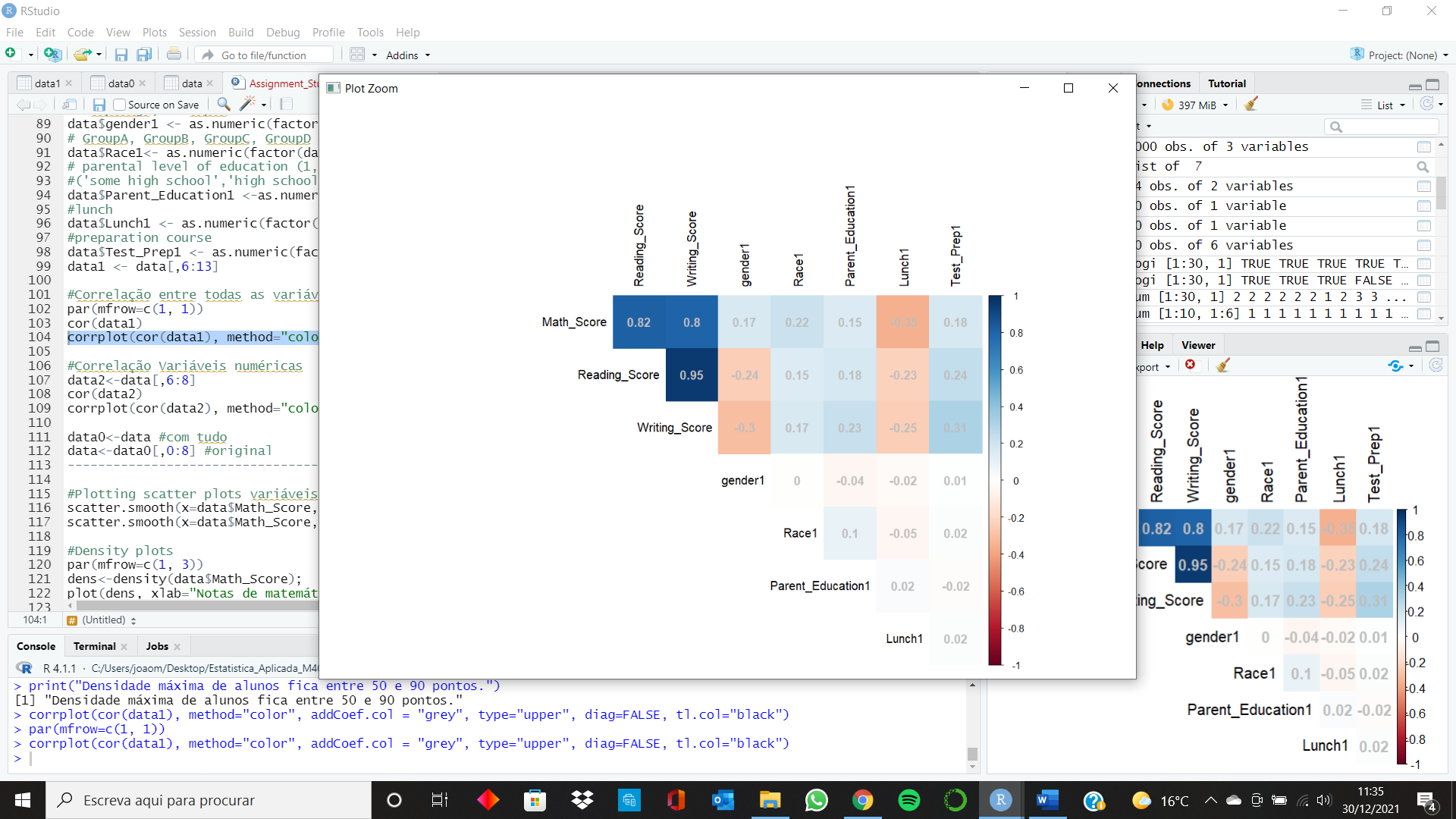


Figura : Matriz de Correlação entre as variáveis

Através da observação da matriz de correlação é possível perceber que as variáveis com maior correlação entre si são: Notas de matemática, Pontuação da leitura e Pontuação da escrita.

* 1. **Gráficos de dispersão entre as variáveis com maior correlação**

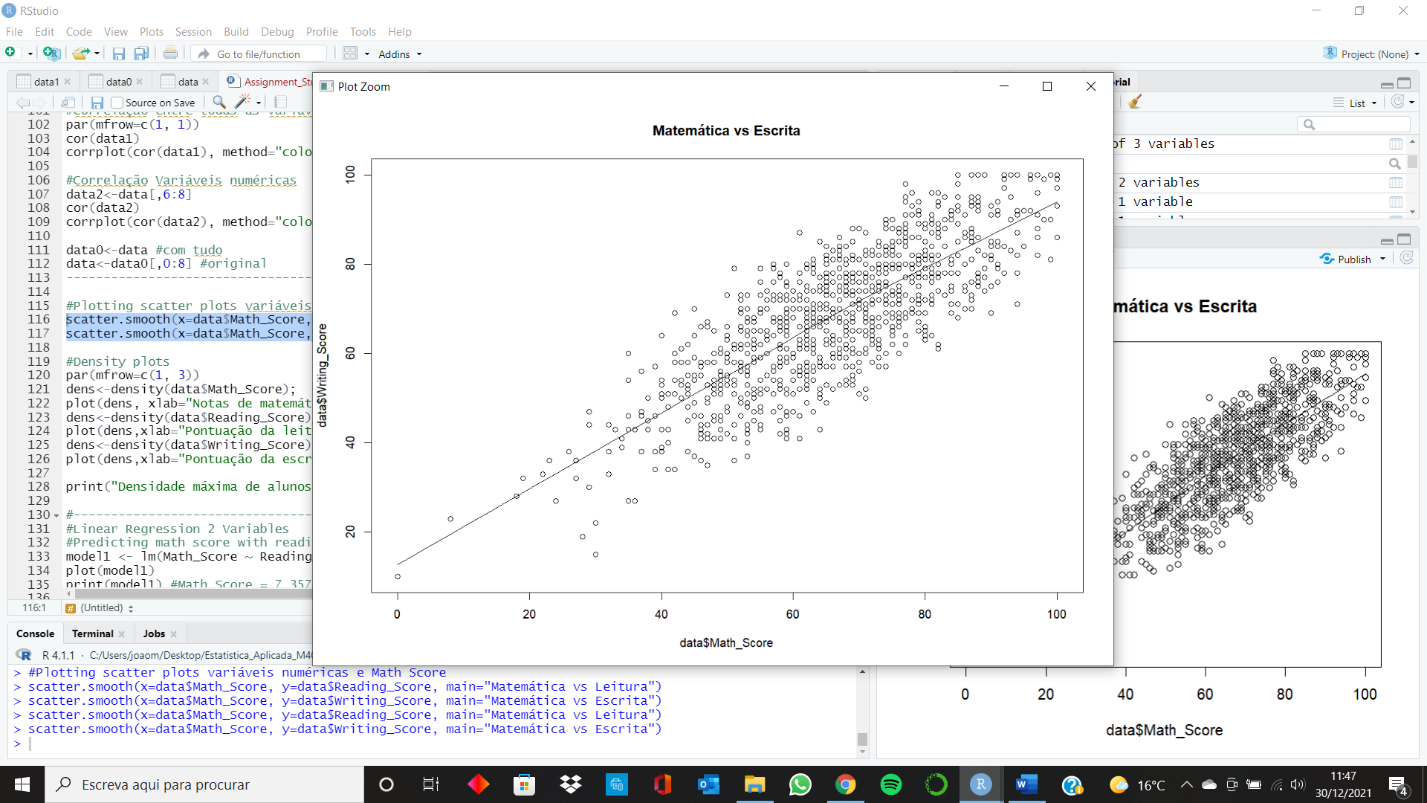
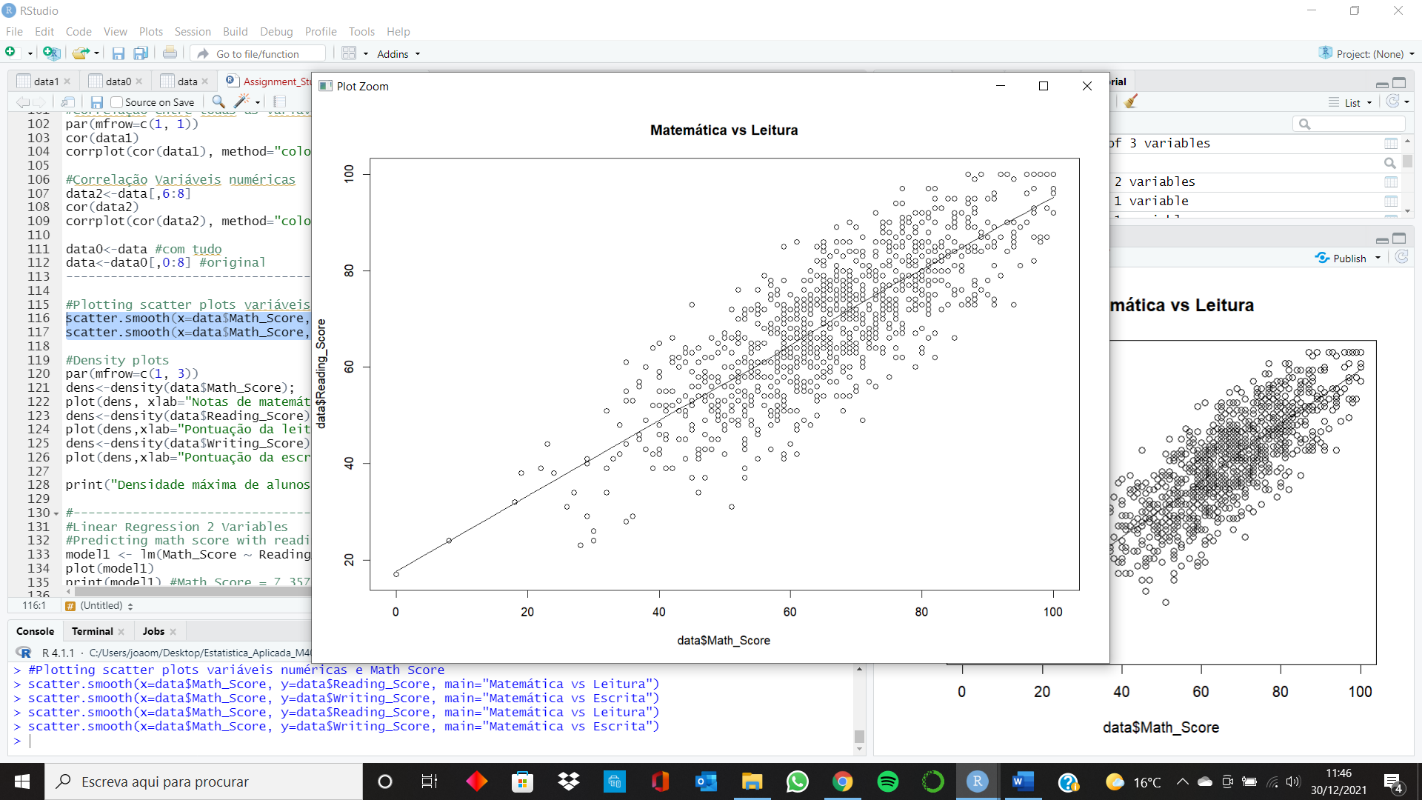
Um gráfico de dispersão mostra a relação entre duas variáveis quantitativas medidas para os mesmos indivíduos. Os valores de uma variável aparecem no eixo horizontal e os valores da outra variável aparecem no eixo vertical. Cada indivíduo nos dados aparece como um ponto no gráfico. (3)

Figura :Gráficos de Dispersão entre as 3 variáveis

Em ambos os casos as duas variáveis têm uma associação positiva visto que tanto os valores acima da média como os valores abaixo da média tendem a ocorrer “juntos”, seguem uma tendência semelhante.

Além disso demonstram uma relação linear, devido à forma do gráfico. Isso significa que os pontos no gráfico de dispersão são semelhantes a uma linha reta. Uma relação é linear se uma variável aumenta aproximadamente na mesma taxa que as outras variáveis mudam gradualmente.

* 1. **Gráficos de Densidade**

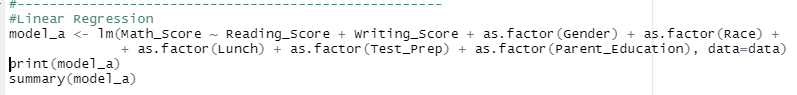
Uma imagem com texto, computador, captura de ecrã

Descrição gerada automaticamenteRealizamos também os gráficos de densidade para estas três variáveis, evidenciando ainda mais a sua relação e linearidade.

Figura :Gráficos de Densidade entre as 3 variáveis

É possível perceber, através destes gráficos, que a densidade máxima de alunos fica entre 50 e 90 pontos.

1. **Discussão do modelo final selecionado**

O conjunto de dados selecionado para este trabalho contém 8 variáveis, sendo que o y são as notas de matemática (*Math\_Score*). Assim, para iniciar a regressão linear múltipla construiu-se um primeiro modelo onde constavam todas as variáveis. Esse modelo foi denominado de *model\_a*.

Através das funções *print()* e *summary()* conseguiu-se obter algumas informações sobre o modelo inicial (*model\_a*).

Uma imagem com texto

Descrição gerada automaticamenteUma imagem com texto

Descrição gerada automaticamente

Através dos resultados apresentados anteriormente, consegue-se afirmar que o R quadrado (R2) para este modelo é de 0,8767. Ainda, é possível verificar o nível de significância de cada variável, através do *p-value* dado pela linha *Signif. Codes*, em que se esse valor for inferior a 0.05, significa que essa variável não é significativa para o modelo e assim pode-se remover essa variável. É importante realçar que nas variáveis categóricas que tenham mais de que uma categoria, basta uma categoria satisfazer a condição *p-value* < 0.05, que essa variável passa a ser significativa para o modelo (4).

Observando então as figuras anteriores é possível verificar que a variável *Parent\_Education*, em todas das categorias, tem um *p-value* superior a 0.05, podendo se então remover essa variável.

Assim sendo, procedeu-se á elaboração de um novo modelo, o model\_b, onde se removeu a variável *Parent\_Education* e que consiste em 6 variáveis (*Reading\_Score, Writing\_Score, Gender, Race, Lunch, Test\_Prep*).

Uma imagem com texto

Descrição gerada automaticamenteDe seguida, e como foi feito para o modelo inicial (model\_a), utilizou-se a função *print()* e *summary()* para obter informações relevantes sobre o *model\_b*

No output de ambas as funções é possível verificar que o valor de R quadrado desceu ligeiramente (0.8748). No entanto, olhando novamente para os níveis de significância das variáveis, isto é, o *p-value*, percebe-se que todas as variáveis têm um valor inferior a 0.05, podendo concluir, por isso, que este é o modelo final da regressão linear deste conjunto de dados.

Após escolher o modelo é necessário proceder á avaliação da sua qualidade. Isto é possível através da função *anova()* e *plot()*, obtendo-se, assim, uma tabela de análise da variância das variáveis e também 4 gráficos relativos ao *model\_b*, que podem ser observados de seguida.

|  |  |
| --- | --- |
|  |  |
|  |  |

Figura :Gráficos do model\_b

Observando os gráficos do *model\_b* é possível verificar que a normalidade dos resíduos, visto que no gráfico normal Q-Q, estes coincidem todos com uma linha diagonal, que representa a distribuição normal teórica e os pontos a distribuição teórica dos resíduos. Através dos restantes gráficos consegue-se perceber a ausência de um padrão no comportamento dos resíduos em relação ao modelo. Assim, devido á normalidade dos resíduos e á ausência de um comportamento padronizado pode-se inferir que este modelo (model\_b) é um modelo homocedástico (5).

1. **Apresentação da equação matemática do modelo**

**G: *Gender***

G0: *female* (referência); G1: *male*

Esta variável será representada por 1 *Dummy*:

G1: 1 – male; 0 - otherwise

**L: *Lunch***

L0: *Free/Reduced* (referência); L1: *standard*

Esta variável é representada por 1 *Dummy*:

L1: 1- *standard*; 0 - *otherwise*

**TP: *Test\_Prep***

TP0: *completed* (referência); TP1: *none*

Esta variável é representada por 1 *Dummy*:

TP1: 1 – *none*; 0 - *otherwise*

**R: *Race***

R0: *group* A (referência); R1: *group* B; R2: *group* C; R3: *group* D; R4: *group* E

Esta variável é representada por 4 *Dummys*:

R1: 1 – *group* B; 0 - *otherwise*

R2: 1 – *group* C; 0 - *otherwise*

R3: 1 – *group* D; 0 - *otherwise*

R4: 1 – *group* E; 0 - *otherwise*

1. **Para uma variável continua X1 e uma variável categórica com mais de 2 categorias X2 que constem do modelo final**

Para esta parte do trabalho, vai se considerar uma variável contínua, X1, que será a variável *Writing\_Score* e uma variável categórica X2, que será a variável *Race*.

* 1. **Interpretação do efeito bruto de *Writing\_Score* e o efeito ajustado de *Race*.**

Uma imagem com texto

Descrição gerada automaticamentePara se interpretar o efeito bruto da variável *Writing\_Score* e o efeito ajustado de *Race*, tem-se que fazer um modelo de regressão linear simples apenas com essas varáveis (*model1* e *model2*) de modo a obter os efeitos brutos. De seguida, com o modelo selecionado (*model\_b*) obtêm-se os efeitos ajustados.

Uma imagem com texto

Descrição gerada automaticamente

Através da função *print* foi possível obter os coeficientes dos 3 modelos e assim perceber o efeito bruto e o efeito ajustado de ambas as variáveis, demonstrado na tabela abaixo.

|  |  |  |
| --- | --- | --- |
|  | Efeito Bruto | Efeito ajustado |
| *Writing\_Score* | 0.8009 | 0.6669 |
| *Racegroup* B | 1.823 | 0.8766 |
| *Racegroup* C | 2.835 | 0.1335 |
| *Racegroup* D | 5.733 | 0.1143 |
| *Racegroup* E | 12.192 | 5.0305 |

**Interpretação – efeitos brutos:**

Olhando primeiro para a variável contínua *Writing\_Score*, é possível afirmar que as notas de matemática (*Math\_Score*) aumentam 0.801 pontos.

No que respeita a variável categórica *Race*, caso um aluno seja de uma etnia do grupo B, existe um aumento das notas de matemática em 1,82 pontos. Esse aumento é verificado para todos os grupos étnicos, sendo que no grupo C as notas de matemática aumentam 2,84 pontos, no grupo D 5,73 e no grupo E existe um aumento mais acentuado com 12,19 pontos na nota de matemática. Caso o aluno seja do grupo étnico A, a sua nota de matemática será 61,63 pontos.

**Interpretação – efeitos ajustados:**

No que concerne a variável *Writing\_Score*, mantendo as restantes variáveis *Reading\_Score*, *Gender*, *Test\_Prep*, *Lunch* e *Race* constantes e após ajustamento dessas variáveis, as notas de matemática aumentam 0,67 pontos.

Na variável categórica *Race*, mantendo as restantes variáveis *Reading\_Score*, *Gender*, *Test\_Prep*, *Lunch* e *Writing\_Score* constantes e após ajustamento dessas variáveis, as notas de matemática podem aumentar 0.88, 0.13, 0.11 ou 5.03, se o aluno pertencer ao grupo étnico B, C, D ou E, respetivamente.

* 1. **Determinação gráfica das bandas de confiança e de predição em função dos valores de *Writing\_Score*, fixando os restantes preditores contínuos nos seus valores medianos e os categorias nas respetivas modas.**

Para proceder à determinação gráfica das bandas de confiança e de predição em função dos valores de Writing\_Score, é necessário fixar os preditores contínuos na sua mediana e os preditores categóricos na sua moda criando um novo *dataframe* (*data3*), como é possível verificar na figura a seguir. No entanto, é importante realçar que se procedeu á elaboração de um novo modelo, igual ao modelo selecionado (model\_b), de modo a não alterar nada desse modelo.

Uma imagem com texto

Descrição gerada automaticamente

Criou De seguida, elaborou-se a predição, através da função *predict()* e de seguida fez-se um gráfico com o resultado desta função, utilizando, posteriormente, a função *ggplot().* Uma imagem com texto

Descrição gerada automaticamente

Uma imagem com texto, captura de ecrã, computador, interior

Descrição gerada automaticamente

Figura :Gráfico da Previsão

* 1. **Interpretação o efeito provocado na resposta por uma mudança da terceira categoria de X2 para a segunda, e indique um intervalo de confiança a 95% para esse efeito.**

A mudança da terceira categoria de Race para segunda categoria tem um efeito de na resposta. Para verificar a significância do efeito provocado, procedeu-se a um *T-test* para a diferença de médias.

Uma imagem com texto

Descrição gerada automaticamente

Para um nível de significância de α=0.05, não se pode rejeitar a hipótese nula, visto que o *p-value* é superior a 0,05 (0.4695), podendo afirmar, assim, que a média das notas de matemática do grupo étnico B é igual ás médias das notas de matemática do grupo étnico C.

* 1. **Interpretação do efeito provocado por um aumento em X1 correspondente a dois desvios padrão dos seus valores.**

Uma imagem com texto

Descrição gerada automaticamente Para se conseguir verificar o efeito provocado por um aumento correspondente a dois desvios padrão de *Writing\_Score*, tem-se que usar a função *lm.beta()*. Aplicando esta função ao *model\_b* obtém-se o seguinte resultado.

O coeficiente estandardizado de *Writing\_Score* é 0.6683, significando que, sempre que a Pontuação da Escrita (*Writing\_Score*) aumenta 1 desvio padrão, o modelo prevê um aumento das notas de matemática (*Math\_Score*) em, aproximadamente, 0.67 desvios padrão, ajustando para. Sendo assim, mantendo as restantes variáveis, sempre que a Pontuação da Escrita (*Writing\_Score*) aumenta 2 desvios padrão o modelo prevê um aumento das notas de matemática (*Math\_Score*) em, aproximadamente, 1.34 desvios padrão.

* 1. **Averiguação da existência de uma interação significativa entre *Writing\_Score* e *Race* e interpretação dos efeitos estimados nessa interação.**

Para averiguar a existência de uma interação significativa entre *Writing\_Score* e *Race* foi criado um novo modelo, o *model3*.



Uma imagem com mesa

Descrição gerada automaticamenteApós se criar o modelo, e como se tem feito para os outros modelos, tem se procedido á verificação do modelo através das funções *print()* e *summary()*.

Uma imagem com mesa

Descrição gerada automaticamente

Para estudar a possível interação entre as variáveis *Writing\_Score* e *Race* foi feito um gráfico.

Uma imagem com texto, recibo, captura de ecrã

Descrição gerada automaticamente

O gráfico apresentado acima, sugere que poderá haver uma interação entre as variáveis *Writing\_Score* e Race. No entanto, os termos de interação do modelo não são significativos.

**Interpretação:**

A Pontuação na Escrita (Writing\_Score) faz aumentar 0.601 pontos nas notas de matemática, dependendo do grupo étnico a que o aluno pertence.

1. **Referências Bibliográficas**

1. COLEMAN JS |AN. O. EQUALITY OF EDUCATIONAL OPPORTUNITY. 1966.

2. Sing Chai C, Van Den Noortgate W, Leuven Kulak K, Bo Ning B, Martínez-Abad F, Gamazo A. Citation: Gamazo A and Martínez-Abad F (2020) An Exploration of Factors Linked to Academic Performance in PISA 2018 Through Data Mining Techniques. Front Psychol. 2020;11:575167.

3. Marshall E. Scatterplots and correlation in SPSS SPSS. 2018.

4. Long J, Teetor P. R Cookbook [Internet]. 2019. Available from: https://rc2e.com/index.html

5. Helena M. Tutorial — Ajuste e Interpretação de Regressão Linear com R [Internet]. 2019. Available from: https://medium.com/data-hackers/tutorial-ajuste-e-interpretação-de-regressão-linear-com-r-5b23c4ddb72