

Instituto Politécnico de Setúbal

Escola Superior de Tecnologia do Barreiro

Análise de Tratamento de Dados Multivariados

David Cabrita (202100320)

João Almeida (202100068)

João Carreira (201901365)

Trabalho apresentado no âmbito da unidade
curricular de Análise e Tratamento de Da-
dos Multivariados, 2^o ano da Licenciatura
em Bioinformática

Docente: Ana Meireles


Barreiro
Dezembro de 2022

Índice

1	Introdução	1
1.1	Caracterização da Amostra	1
1.2	Fazer outra subsecção (se fizer sentido)	1
2	Análise das variáveis individualmente	2
2.1	Idade	2
2.2	Sexo	3
2.3	Curso que frequenta	4
2.4	Ano Curricular que frequenta	4
2.5	Este curso foi a 1ªOpção?	5
2.6	Fui eu que escolhi este curso?	6
2.7	Deslocação Casa-Escola	6
2.8	Nº de horas de estudo por semana	8
2.9	Nº de horas diárias dedicadas às redes sociais	9
2.10	Número de horas diárias dedicadas a ver TV/Netflix/etc	9
2.11	Número médio de horas que dorme por dia de 2ªf a 6ªf	10
2.12	Tem conhecimento que a ESTBarreiro/IPS oferece um Programa de Mentoria aos seus estudantes	11
3	Análise da relação entre duas variáveis	12
3.1	Qual dos sexos sente maior desgaste emocional com os estudos?	12
3.2	Como estão distribuídos os cursos tendo em conta o sexo dos estudantes?	13
3.3	Como as Horas de Estudo estão relacionadas com as Horas dedicadas TV,Netflix,etc?	14
4	Estudo Inferencial para a validação de Questões	15
4.1	O Tempo de Deslocação Casa-Escola é influenciado pelo Ano Curricular?	15
4.2	O Número Horas de Estudos varia de acordo com os alunos que tiveram o seu curso como 1ªOpção?	18
4.3	O Número de Horas de Sono varia entre os Cursos em Estudo?	21
5	Métodos de Análise de Dados Multivariados	23
5.1	As Horas de sono dos estudantes pode ser afetada pelas Horas de estudo e o curso que frequentam?	24
5.2	As Horas de Sono são afetadas pelo conhecimento do programa de Mentoria e o Tempo de Deslocação	29
6	Análise Fatorial	33

7 Conclusão	38
Referências	40
Apêndice A Anexo Utilizado no Estudo	41


Resumo


Neste estudo, foi-nos proposto a exploração do software  para fins estatísticos com o qual devemos fazer a caracterização da amostra, estudo inferencial, regressão linear múltipla e redução da informação.

A nossa amostra foram as respostas dos estudantes da ESTBarreiro a um questionário onde abrangimos perguntas como Sexo, Idade, Curso que frequenta, se sente que é um bom aluno, se tem vindo a perder interesse no seu curso específico, entre outras.

Ao longo deste projeto vamos procurar sempre encontrar as melhores variáveis para relacionar com a vista a obter resultados que consideremos interessantes e que tragam informação relevante para o estudo.

1 Introdução

Neste projeto foi-nos proposto a exploração das funções software  de modo a realizar a análise estatística de um questionário realizado aos alunos da ESTBarreiro.

Criado originalmente por Ross Ihaka e por Robert Gentleman no departamento de Estatística da Universidade de Auckland, na Nova Zelândia, o  é um software gratuito utilizado em numerosos estudos na área de Data-Science, Estatística, entre outras áreas. Dispõe de uma vasta gama de ferramentas úteis para a manipulação de dados e tratamento dos mesmos, aliado a um suporte gráfico altamente versátil e de fácil compreensão e utilização.

O questionário em estudo é constituído por 12 perguntas de resposta simples e outras 15 feitas em escala de **Burnout de Maslach** de modo a entender como os alunos se sentem face às dificuldades dos seus cursos e avaliar também os sentimentos/emoções dos estudantes em contexto escolar.

1.1 Caracterização da Amostra

O questionário que nos foi fornecido para realizar o trabalho de grupo da unidade curricular de ATDM (Análise e Tratamento de Dados Multivariados) do curso de Licenciatura em Bioinformática tem como objeto de estudo os estudantes da ESTBarreiro onde os estudantes começam por identificar a sua idade, o seu sexo, o curso que frequentam (Neste questionário os cursos em estudo foram Bioinformática, Biotecnologia e CTeps Tecnologias de Laboratório Químico e Biológico) e o seu ano curricular.

O total de estudantes que responderam a este questionário é de **142 estudantes**, onde **45,8%** são do sexo feminino enquanto que os restantes **54,2%** são do sexo masculino.

De seguida são realizadas questões relativamente a temas sobre o curso, o tempo de deslocação dos estudantes, Nº de horas de estudo, sono e de horas dedicadas às redes sociais, e pergunta-se se os estudantes têm conhecimento do programa de Mentoria que a ESTBarreiro oferece aos estudantes.

Por fim os estudantes têm 15 afirmações em **Burnout de Maslach** onde se pretende avaliar os sentimentos/emoções dos estudantes em contexto escolar.

1.2 Fazer outra subsecção (se fizer sentido)

Escrever aqui o texto da segunda subsecção.

Se quiser fazer referência a alguma seção ou subseção basta indicar a *label* associada a essa seção. Por exemplo, esta subseção é a segunda subseção da seção 1 e vem depois da subseção 1.1.

2 Análise das variáveis individualmente

A **análise estatística univariada** é aquela que permite a análise de cada variável separadamente juntamente com os métodos de **estatística inferencial** para uma determinada variável, podendo ser medida por uma ou mais amostras independentes.

Iremos agora fazer a análise estatística univariada referente às 12 primeiras perguntas do questionário em estudo, onde, através de gráficos, tabelas e histogramas iremos analisar todos os dados fornecidos e comentá-los consoante os resultados obtidos.

2.1 Idade

Iremos começar pela primeira questão, referente à idade de cada estudante.

Por se tratar de uma **variável quantitativa contínua**, para podermos realizar uma análise completa iremos proceder à criação de um **histograma**, de um **diagrama de extremos e quartis** e a sua **mediana e média**.

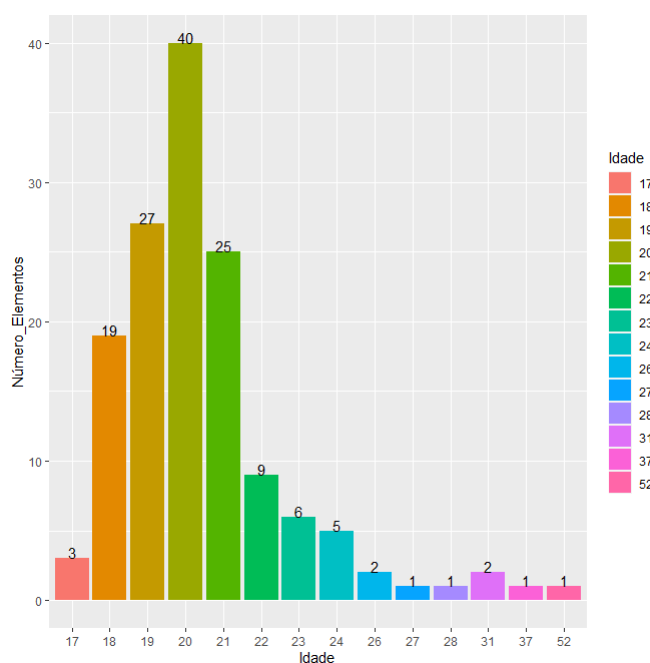


Figura 2.1: Histograma relativamente à variável Idade

Através deste histograma 2.1 podemos observar que as idades variam entre os 17 e os 52 anos.

Podemos também observar que as idades mais comuns entre os estudantes estão compreendidas entre os 18 e os 21 anos.

Vamos então avançar para o diagrama de extremos e quartis:

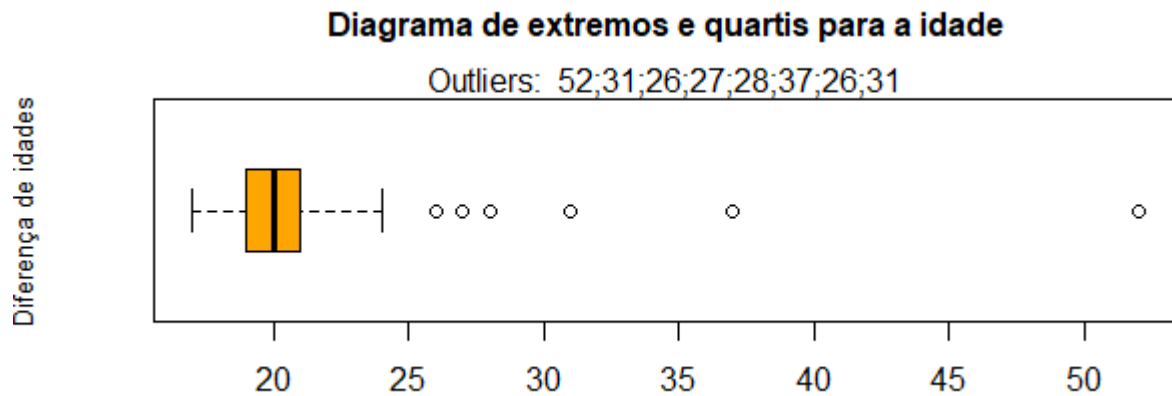


Figura 2.2: Diagrama de extremos e quartis relativamente á variável Idade

Como podemos observar através do diagrama de extremos e quartis 2.2, a distribuição de idades aproxima-se muito de uma **distribuição simétrica**, podemos identificar a presença de 8 **outliers**, sendo 2 deles repetidos (26 e 31) e dentro dos 8 outliers 5 deles são **severos**, sendo eles 31,31,52,37 e 28.

Agora vamos obter a tabela com a **média e mediana**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
17.00	19.00	20.00	20.74	21.00	52.00

Tabela 2.1: Resumo da estatística de teste para a variável Idade

Podemos então verificar pela tabela 2.1 que a **média** é de 20.74 e a **mediana** é 20, podemos também retirar a informação sobre os **máximos e os mínimos**(17 e 52) juntamente com os valores do **1ºQuartil** e do **3ºQuartil**(19 e 21).

2.2 Sexo

Agora vamos para a variável Sexo, trata-se de uma variável **qualitativa nominal**, por isso começamos pela construção de um **gráfico circular** de modo a analisar a percentagem de estudantes do sexo masculino e feminino.

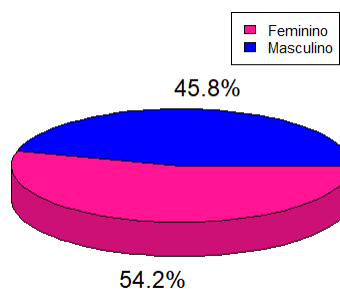


Figura 2.3: Gráfico circular para a variável Sexo

O que podemos retirar do gráfico 2.3 é que, dentro dos estudantes que responderam ao questionário, **54.2%** é do **sexo feminino** enquanto que **45.8%** é do **sexo masculino**.

2.3 Curso que frequenta

Vamos agora para a variável **curso que frequenta**, estamos novamente a falar de uma **variável qualitativa nominal** por isso iremos novamente proceder á criação de um **gráfico circular** para a analisarmos os resultados obtidos:

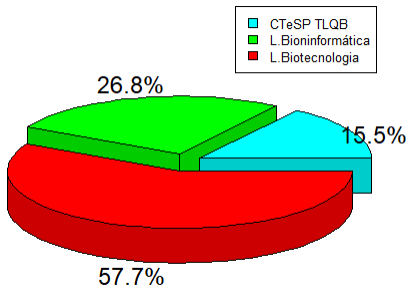


Figura 2.4: Gráfico circular para a variável Curso que frequenta

Podemos então observar pela figura 2.4 que, dentro dos alunos que responderam ao questionário, **57.7%** são estudantes que frequentam o curso de **Biotecnologia**, **26.8%** frequenta o curso de **Bioninformática** enquanto que os restantes **15.5%** frequentam o curso **CTeSP TLQB**.

Concluimos então que o curso com mais estudantes a responder a este questionário é o curso de **Biotecnologia**.

2.4 Ano Curricular que frequenta

Estamos perante uma **variável quantitativa discreta** de modo que, iremos proceder á criação de um **gráfico de barras**:

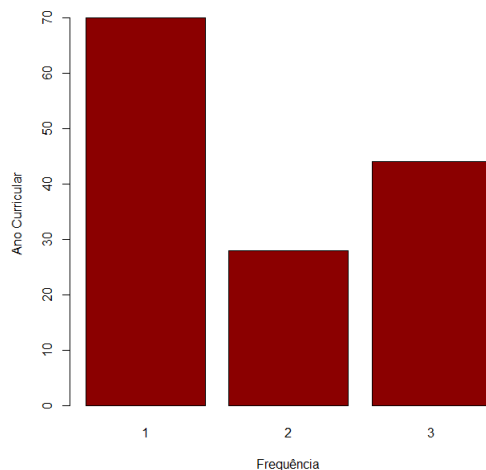


Figura 2.5: Gráfico de Barras para a variável Ano Curricular

Após analisarmos o gráfico 2.5, podemos observar que a maioria dos estudantes que responderam ao questionário está no **1ºano**, enquanto os alunos do **2ºano** estão em menor quantidade.

2.5 Este curso foi a 1ªOpção?

Estando agora perante uma **variável qualitativa nominal**, para esta análise iremos recorrer á criação de um **gráfico circular**:

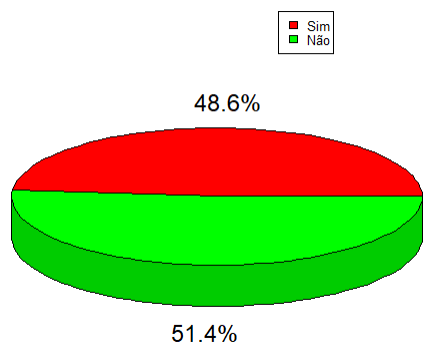


Figura 2.6: Gráfico Circular para a variável 'Este curso foi a 1ªOpção?'

Como é possível verificar pelo gráfico 2.6 que, os resultados são bastante equilibrados, dentro dos estudantes que responderam ao questionário, **51,4%** não teve o seu curso como 1ªopção, enquanto que os restantes **48,6%** tiveram o seu curso como 1ªopção.

2.6 Fui eu que escolhi este curso?

Estando perante outra **variável qualitativa nominal** vamos novamente recorrer á elaboração de um **gráfico circular**:

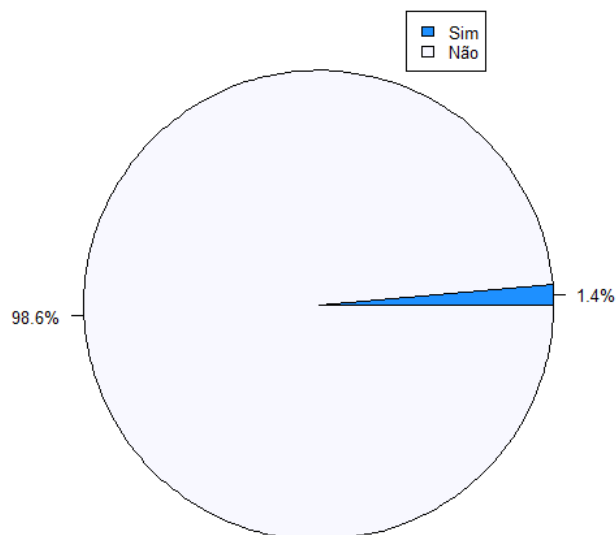


Figura 2.7: Gráfico Circular para a variável Fui eu que escolhi este curso?

De acordo com o gráfico 2.7, vemos que **98.6%** dos estudantes foram eles que escolheram o curso em que estão e apenas **1.4%** não escolheram o curso em que estão.

2.7 Deslocação Casa-Escola

Vamos agora avaliar o tempo de deslocação dos estudantes desde a sua residência até a sua escola, desta vez estamos perante uma **variável quantitativa contínua**, de modo que, para podermos analisar corretamente todos os dados iremos construir um **histograma**, um **diagrama de extremos e quartis** e iremos calcular a sua **média e mediana**.

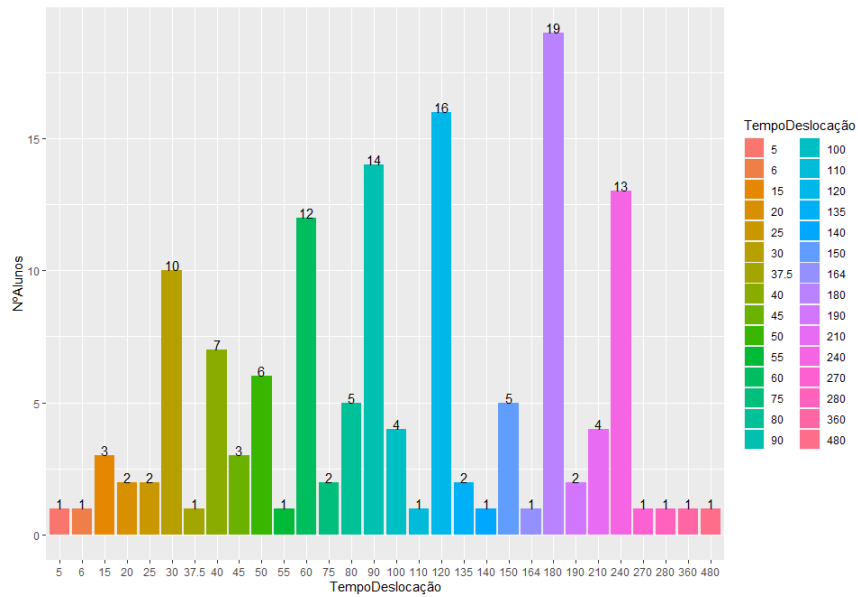


Figura 2.8: Histograma para a variável Tempo de Deslocação

Através do histograma 2.8, vemos que o intervalo de números é bastante elevado, sendo que o tempo de deslocação mais comum entre os estudantes é **180 minutos(19)**, **120 minutos(16)** e **90 minutos(14)**.

Vamos agora avançar para o **diagrama de extremos e quartis**:

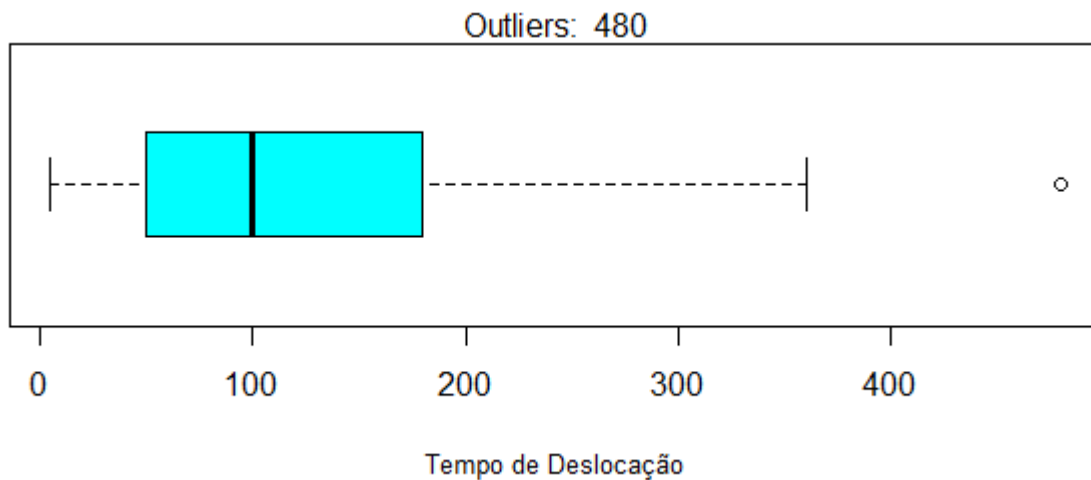


Figura 2.9: Diagrama de extremos e quartis para a variável Tempo de Deslocação

Podemos então verificar pelo diagrama 2.9 que os dados têm uma distribuição **assimétrica positiva**, temos 1 **outlier**, que é **480 minutos**, porém **não é severo**. Vamos agora calcular a sua **média e mediana**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
5.00	51.25	100.00	117.52	180.00	480.00

Tabela 2.2: Resumo da estatística de teste para a variável Tempo de Deslocação

Podemos então retirar desta tabela 2.2 o **mínimo** que é **5 minutos** e o valor **máximo** que é **480 minutos**, o valor do **1º e 3º Quartil** é de **51 e 180 minutos** respetivamente, a **mediana** é **100 minutos** e o tempo de deslocação **médio** é de **117 minutos**.

2.8 Nº de horas de estudo por semana

Vamos agora para o número de horas por semana que os alunos dedicam para estudar. Estamos perante uma **variável quantitativa discreta**, de modo que vamos proceder á criação de um **diagrama de extremos e quartis**:

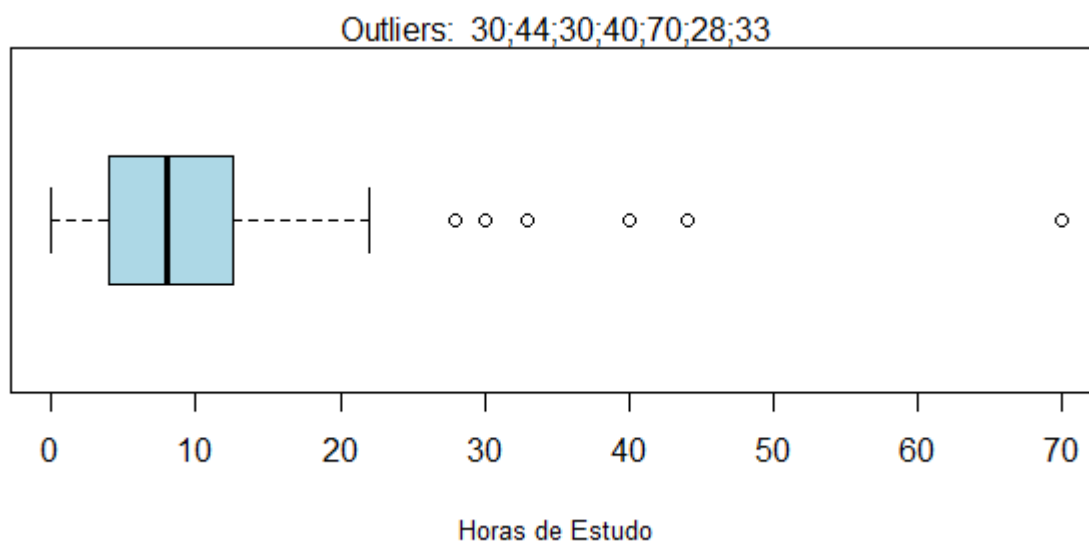


Figura 2.10: Diagrama de extremos e quartis para a variável Horas de Estudo

Podemos então verificar pelo diagrama 2.10 que existe uma **distribuição assimétrica positiva**, e verificamos a existência de **7 outliers**, sendo eles 30,30,28,33,40,44,70. Dentro dos outliers, temos **3 severos**, sendo eles **40,44 e 70**. Vamos agora prosseguir para a tabela onde temos os valores da **média** e da **mediana**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	4.000	8.000	9.701	12.375	70.000

Tabela 2.3: Resumo da estatística de teste para a variável Horas de Estudo

Como podemos observar pela tabela 2.3, o valor mínimo é de **0 horas** e o valor máximo é de **70 horas**, o valor do **1º e 3º Quartil** é de **4 e 12 horas** respetivamente. A **mediana** tem o valor de **8 horas**, e o valor médio de horas de estudo por semana é de **9.7 horas**.

2.9 N^o de horas diárias dedicadas às redes sociais

Agora vamos avaliar o número de horas que os estudantes gastam por dia em redes sociais, tratando-se de uma **variável quantitativa discreta**, criaremos um **diagrama de extremos e quartis** e iremos calcular o valor da **média** e da **mediana**:

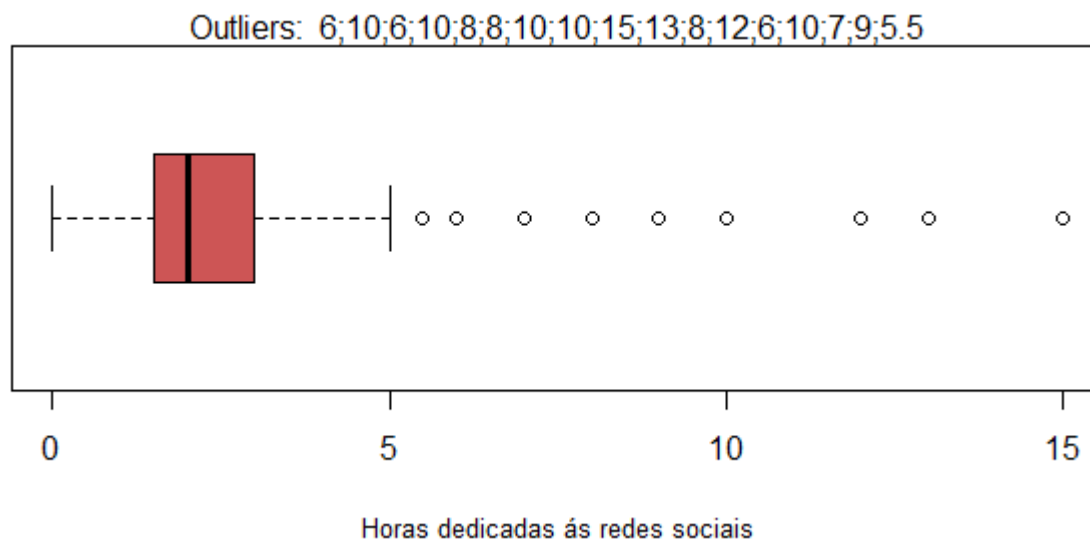


Figura 2.11: Diagrama de extremos e quartis para a variável horas diárias dedicadas às redes sociais

Como podemos observar pelo diagrama 2.11 estamos perante uma **distribuição assimétrica positiva** e estamos perante 10 **outliers**, sendo eles 5,5,6,7,8,9,10,12,13 e 15. De entre estes outliers, temos que 6 são **outliers severos**, sendo eles 8,9,10,12,13 e 15. Vamos agora para a tabela com o valor da **média** e da **mediana**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	1.500	2.000	3.092	3.000	15.000

Tabela 2.4: Resumo da estatística de teste para a variável número de horas diárias dedicadas às redes sociais

Como podemos observar pela tabela 2.4, o valor mínimo e máximo são **0 e 15 horas** respetivamente, os valores do **1^o e 3^oQuartil** é de **1.5 e 3 horas** respetivamente, a **mediana** tem um valor de **2 horas** e o valor médio que os estudantes gastam por dia em redes sociais é de **3 horas**.

2.10 Número de horas diárias dedicadas a ver TV/Netflix/etc

Iremos agora avaliar o número de horas diárias dedicadas a ações como ver TV,Netflix,etc, estando novamente perante uma **variável quantitativa discreta**, porém desta vez iremos criar um **gráfico de barras** e proceder ao cálculo da sua **mediana** e da sua **média**:

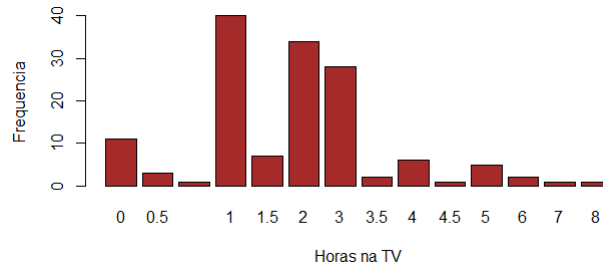


Figura 2.12: Gráfico de barras para a variável número de horas diárias dedicadas a TV,Netflix,etc

Como podemos observar pelo gráfico 2.12, as horas dedicadas a ver TV,Netflix,etc está entre as 0 e as 8 horas, e que a maioria dos estudantes dedica entre **1 a 3 horas** por dia a ver TV,Netflix,etc. Possuímos ainda **2 outliers**, sendo eles **7 e 8**, porém nenhum deles é severo. Vamos agora para o cálculo da **média** e da **mediana**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	1.000	2.000	2.058	3.000	8.000

Tabela 2.5: Resumo da estatística de teste para a variável número de horas diárias dedicadas a TV,Netflix,etc

Vemos então pela tabela 2.5 que o valor mínimo e máximo é **0 e 8 horas** respetivamente, o valor do **1º e 3ºQuartil** é **1 e 3 horas** respetivamente, o valor da **mediana** é **2 horas**, e o valor médio de horas que os estudantes gastam por dia a ver TV,Netflix,etc é de **2 horas**.

2.11 Número médio de horas que dorme por dia de 2^a a 6^a

Vamos agora avaliar o número de horas que os estudantes dormem de 2^a a 6^afeira, tratando-se novamente de uma **variável quantitativa discreta**, vamos proceder á criação de um **diagrama de extremos e quartis** e proceder ao cálculo da sua **mediana** e da sua **média**:

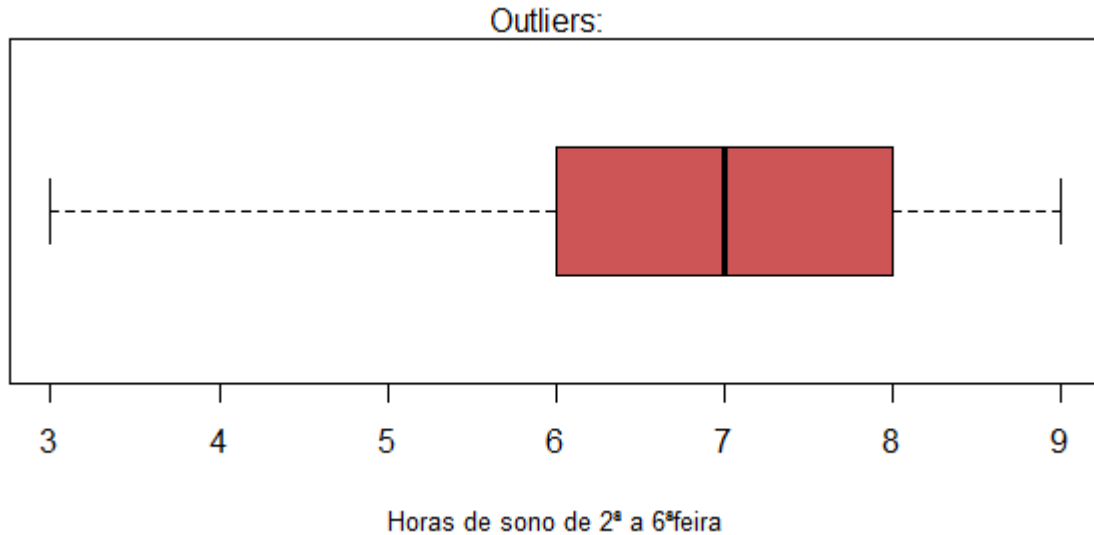


Figura 2.13: Diagrama de extremos e quartis para a variável número de horas de sono de 2ª a 6ªfeira

Podemos então verificar que, pelo diagrama 2.13, estamos perante uma **distribuição simétrica** e não se verifica a existência de outliers. Vamos agora para o cálculo da **média** e da **mediana**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.000	6.000	7.000	6.956	8.000	9.000

Figura 2.14: Resumo da estatística de teste para a variável número de horas de sono de 2ª a 6ªfeira

Podemos então verificar pela tabela 2.14 que, os valores mínimos e máximos são de **3 e 9 horas**, o valor do **1º e 3ºQuartil**, a **mediana** tem o valor de **7 horas** e, o número médio de horas que os estudantes dormem de 2ª a 6ªfeira é de **6.9 horas**.

2.12 Tem conhecimento que a ESTBarreiro/IPS oferece um Programa de Mentoria aos seus estudantes

Vamos agora avaliar se os estudantes que responderam ao questionário têm conhecimento da existência do programa de mentoria que a ESTBarreiro/IPS oferece aos seus estudantes. Estamos perante uma **variável quantitativa nominal** e para podermos analisar os dados da melhor maneira possível, vamos criar um **gráfico circular**:

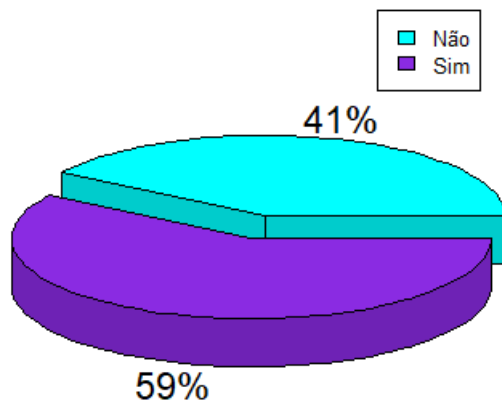


Figura 2.15: Gráfico circular para a variável Programa de Mentoria

Podemos então verificar pelo gráfico 2.15 que cerca de **59%** dos estudantes que respondeu ao questionário têm conhecimento do Programa de Mentoria que a ESTBarreiro/IPS dispõe, enquanto que **41%** não tem tal conhecimento. Logo, podemos concluir que a maioria dos estudantes tem conhecimento que a ESTBarreiro/IPS oferece um Programa de Mentoria.

3 Análise da relação entre duas variáveis

Vamos agora avaliar a maneira como duas variáveis se comportam na presença uma da outra, para tal iremos utilizar a **Análise Descritiva Bivariada**.

Iremos proceder á criação de uma tabela de contingência de modo a resumir a informação de uma das questões da Escala de Burnout de Maslach de acordo com o sexo do Estudante.

Iremos também abordar dois coeficientes de correlação que nos permitam estudar 2 pares distintos de variáveis.

3.1 Qual dos sexos sente maior desgaste emocional com os estudos?

Vamos então começar com a relação entre a variável **Sexo** e a **Burnout P1**, que é "Os meus estudos deixam-me emocionalmente exausto(a)", com este estudo pretendemos averiguar se algum dos sexos tem vindo a sentir maior desgaste emocional que o outro.

Vamos então proceder para a criação da **tabela de contingência**:

Sexo	BurnoutP1							Total
	0 - Nunca	1 - Quase Nunca	2 - Algumas Vezes	3 - Regularmente	4 - Muitas Vezes	5 - Quase Sempre	6 - Sempre	
Feminino	0 0 %	6 9.2 %	18 27.7 %	6 9.2 %	14 21.5 %	12 18.5 %	9 13.8 %	65 100 %
Masculino	3 3.9 %	14 18.2 %	26 33.8 %	12 15.6 %	15 19.5 %	4 5.2 %	3 3.9 %	77 100 %
Total	3 2.1 %	20 14.1 %	44 31 %	18 12.7 %	29 20.4 %	16 11.3 %	12 8.5 %	142 100 %

$$\chi^2=15.788 \cdot df=6 \cdot \text{Cramer's } V=0.333 \cdot \text{Fisher's } p=0.015$$

Tabela 3.1: Tabela de Contingência para a variável Sexo e a Burnout P1

Nesta análise da tabela 3.1, foram avaliados **142 estudantes**, sendo **65** do sexo **feminino** e **77** do sexo **masculino**. Podemos observar que, as respostas mais frequentes foram **"Algumas vezes"** e **"Muitas vezes"**, com **31%** e **20.4%** das respostas respectivamente. Existiu uma grande discrepância entre os sexos para a resposta **"Quase Sempre"**, onde o sexo feminino registou **18.5%** dos estudantes a colocar esta alternativa, enquanto que o sexo masculino, apenas **5.2%** colocaram esta alternativa, outra das discrepâncias foi para a resposta **"Sempre"** onde o sexo feminino registou **13.8%** dos estudantes a colocar esta alternativa, já o sexo masculino apenas **3.9%** colocou esta alternativa. Por fim, a última discrepância foi para a resposta **"Quase Nunca"**, onde o sexo feminino registou **9.2%** dos estudantes a colocar esta alternativa, enquanto que, para o sexo masculino, registou-se **18.2%** dos estudantes a colocarem esta alternativa.

Por fim, outras das discrepâncias acontece na resposta **"Quase Sempre"**, onde **9.2%** dos estudantes do sexo feminino diz que, desde que ingressou no curso, tem sentido quase sempre desinteresse pelo mesmo, enquanto que para os estudantes do sexo masculino, registamos apenas **5.2%** dos estudantes com a mesma resposta.

Podemos concluir que, os estudantes do sexo **feminino** têm tendência a que os estudos as esgotem emocionalmente quando comparado aos estudantes do sexo **masculino**.

O **Coefficiente de associação de Cramer** assume o valor de **0.333**, revelando uma **associação baixa**.

3.2 Como estão distribuídos os cursos tendo em conta o sexo dos estudantes?

Vamos agora realizar a relação entre as variáveis Sexo e Curso que frequenta, com este estudo pretendemos como estão organizados os cursos de acordo com o sexo dos estudantes, visando entender qual curso possui uma maior taxa de indivíduos de um determinado gênero, para isso, iremos proceder á criação de uma **tabela de contingência** e de seguida vamos calcular o **coeficiente de associação de Cramer**:

<i>Sexo</i>	<i>Curso</i>			<i>Total</i>
	CTeSP TLQB	L. Bioinformática	L. Biotecnologia	
Feminino	17 26.2 %	7 10.8 %	41 63.1 %	65 100 %
Masculino	5 6.5 %	31 40.3 %	41 53.2 %	77 100 %
Total	22 15.5 %	38 26.8 %	82 57.7 %	142 100 %

$$\chi^2=20.838 \cdot df=2 \cdot \text{Cramer's } V=0.383 \cdot p=0.000$$

Tabela 3.2: Tabela de Contingência para a variável Sexo e Curso que Frequenta

Após a análise da tabela 3.2 em questão, é possível verificar que o curso onde o sexo feminino está mais presente seria o curso de **L.Biotecnologia** com **63.1%** e onde está menos presente seria no curso de **L.Bioinformática**, com apenas **10.8%**.

Para o sexo masculino, podemos verificar que o curso de **L.Biotecnologia** continua a ser aquele com mais indivíduos do sexo masculino, com o total de **53.2%** e estando menos presente no curso **CTeSP TLQB**, com apenas **6.5%**.

Podemos verificar que o curso de **L.Bioinformática** e o curso **CTeSP TLQB** foram aqueles onde se apresentou uma maior discrepância, uma vez que, no curso de **L.Bioinformática**, dentro dos 65 estudantes femininos que realizaram o questionários, somente 7 estavam presentes neste curso, porém para os estudantes do sexo masculino, dentro dos 77 envolvidos no questionário, 31 estão inscritos no mesmo curso, e no curso **CTeSP TLQB**, dos 65 estudantes do sexo feminino, 17 estavam inscritas neste curso, e dos 77 estudantes do sexo masculino, somente 5 estavam inscritos no curso.

O **Coefficiente de associação de Cramer** assume o valor de **0.383**, revelando uma **associação baixa**.

3.3 Como as Horas de Estudo estão relacionadas com as Horas dedicadas TV,Netflix,etc?

Vamos agora realizar a análise bivariada entre as variáveis **Horas dedicadas a TV,Netflix,etc** e **Número de horas de estudo por semana**, para isso iremos proceder ao cálculo do **Coefficiente de Correlação de Pearson**.

```
> cor(dados$HorasEstudo,dados$HorasTV)
[1] -0.1408496
```

Figura 3.1: Cálculo do Coeficiente de Pearson

Podemos então observar pela figura 3.1 que o valor do coeficiente de pearson foi, aproximadamente, **-0.141**, neste caso, como o coeficiente de correlação se mostrou negativo, indica-nos que existe uma **correlação linear negativa**. Isto significa que, as variáveis estão **inversamente relacionadas**, isto é, quando o número de horas de estudo aumenta, o número de horas de TV diminui, e vice-versa.

4 Estudo Inferencial para a validação de Questões

Vamos agora para o **Estudo Inferencial**, este estudo tem como objetivo estimar parâmetros populacionais a partir do estudo de uma determinada amostra. Neste projetos iremos levantar **3 questões de investigação** tendo em conta os dados em estudo, onde iremos dar uso a métodos de inferência estatística, considerando como nível de significância $\alpha = 0.05$.

4.1 O Tempo de Deslocação Casa-Escola é influenciado pelo Ano Curricular?

A análise das diferenças entre os tempos de deslocação por parte dos estudantes dos diferentes cursos será primeiramente apresentada em termos descritivos.

Neste estudo foi considerado um total de **142 estudantes**, sendo eles **49%** do **primeiro ano**, **20%** do **segundo ano** e **31%** do **terceiro ano**. O ano curricular que registou um maior tempo médio de deslocação Casa-Escola foi o segundo ano com uma **média** de **176.4 ± 99.35** minutos, seguido do primeiro ano com uma média de **106.1 ± 62.2** minutos, e por fim o terceiro ano com o tempo mais baixo, com **98.18 ± 73.68**, como se apresenta na tabela 4.1. Foi verificado a presença de **2 outliers**, ambos outliers moderados superiores, sendo um no segundo ano e outro no terceiro ano (480min e 360min respetivamente), como podemos observar na tabela 4.2.

Atrvés do diagrama de extremos e quartis apresentado na figura 4.1, pode-se verificar que o tempo de deslocação para os estudantes do primeiro e terceiro ano se distribuem de uma forma **assimétrica positiva**, isto é, há uma maior tendência para que os estudantes demorem menos que o tempo médio acima definido. Situação contrária é verificada nos estudantes do segundo ano, uma vez que se regista uma distribuição **assimétrica negativa**, isto é, os estudantes tendem a demorar mais tempo que o tempo médio anteriormente calculado.

Assim, em termos descritivos, pode-se concluir que os tempos de deslocação dos estudantes do segundo ano diferem dos registados para os estudantes dos restantes anos curriculares.

Tabela 4.1: Resumo das principais estatísticas descritivas entre as variáveis em questão.

Ano Curricular	variable	n	min	max	q1	median	q3	mean	sd
Primeiro	TempoDesloca	70	5	270	60	90	150	106.1	62.2
Segundo	TempoDesloca	28	15	480	90	180	240	176.4	99.35
Terceiro	TempoDesloca	44	15	360	40	60	146.2	98.18	73.68

Antes de concluir os resultados desta tabela, vamos verificar a **presença de outliers**:

Tabela 4.2: Identificação dos outliers

Ano Curricular	Idade	TempoDesloca	is.outlier	is.extreme
2	20	480	TRUE	FALSE
3	23	360	TRUE	FALSE

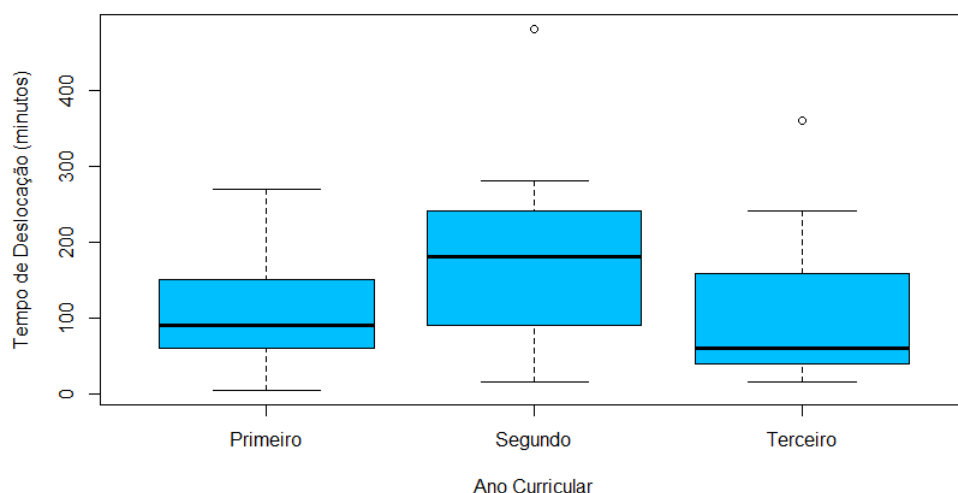


Figura 4.1: Diagrama de extremos e quartis

Pretendemos então verificar se, para $\alpha = 0.05$, existem diferenças significativas entre os valores médio de uma variável quantitativa (Tempo de deslocação Casa-Escola) em 3 grupos independentes (3 anos curriculares envolvidos no estudo).

O teste mais adequado será o teste **ANOVA one-way** para amostras independentes, porém teremos de aplicar um teste á normalidade de modo a poder saber se podemos avançar com este teste.

De modo a verificar a normalidade de cada um dos grupos, iremos aplicar o teste de normalidade de **Kolmogorov-Smirnov**, uma vez que $n > 30$.

```
Two-sample Kolmogorov-Smirnov test

data:  EscalaBurnoutGrupo2$TempoDesloca and EscalaBurnoutGrupo2$`Ano Curricular`
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figura 4.2: Teste á normalidade Kolmogorov-Smirnov

Como podemos observar, através da figura 4.2, para um nível de confiança 0.05, nenhum dos tempos de deslocação dos diferentes anos curriculares provém de uma **distribuição normal** ($\alpha > 2.2e-16$). Não verificado o pressuposto de normalidade, não será possível aplicar o teste paramétrico ANOVA onw-way, em vez disso ,vamos avançar para a sua alternativa não paramétrica, que será o **Teste de Kruskal-Wallis** .

```
kruskal-wallis rank sum test

data:  TempoDesloca by dados$`Ano Curricular`
kruskal-wallis chi-squared = 15.339, df = 2, p-value = 0.0004667
```

Figura 4.3: Teste de Kruskal-Wallis

.y.	n	effsize	conf.low	conf.high	method	magnitude
TempoDesloca	142	0.09597	0.01	0.23	eta2 [H]	moderate

Figura 4.4: Dimensão do efeito para o teste de Kruskal-Wallis

Conclusão:

Para um nível de significância 0.05, pode-se considerar que existem diferenças significativas entre o tempo de deslocação e o ano curricular dos estudantes, uma vez que $p > 0.0004$, e o seu efeito é **moderado**.

Uma vez que se verificaram diferenças significativas no tempo de deslocação de acordo com os anos curriculares, é necessário analisar essas diferenças. Para tal, vamos recorrer ao **teste de Dunn**:

```
Comparison      Z      P.unadj      P.adj
1 Primeiro - Segundo -3.2800595 0.0010378519 0.0031135558
2 Primeiro - Terceiro 0.9295647 0.3525965001 1.0000000000
3 Segundo - Terceiro 3.7737004 0.0001608439 0.0004825318
```

Figura 4.5: Teste de Dunn

A avaliação dos tempos de deslocação nos 3 anos curriculares foi avaliado através do teste não paramétrico **Kruskal-Wallis**. A utilização deste teste é justificada pelo facto de, após ser aplicado o

teste de normalidade de **Kolmogorov-Smirnov** ($n > 30$), conclui-se que não foi verificado o **pressuposto de normalidade** ($p = 0.007$). Após realização do teste de Kruskal-Wallis, verificamos que, para um nível de significância de 5%, existem diferenças significativas entre o tempo de deslocação dos diferentes anos curriculares ($p = 0.0004$). O maior tempo médio de deslocação casa-escola foi do segundo ano (176,4 minutos), seguido do primeiro ano (106,1 minutos) e do terceiro ano (98,18 minutos). A comparação entre pares foi realizada com recurso ao **teste de Dunn** que realiza comparações múltiplas de média de ordens. Através deste teste foi possível observar que existem diferenças estatisticamente significativas no tempo de deslocação entre o primeiro e o segundo ano ($p = 0.003$), e entre o segundo e o terceiro ano ($p = 0.0004$), sendo que o tempo médio de deslocação revelou-se superior nos estudantes do **segundo ano**.

Podemos concluir que, os alunos do 2º ano têm um tempo de Deslocação Casa-Escola superior aos alunos dos restantes anos curriculares.

4.2 O Número Horas de Estudos varia de acordo com os alunos que tiveram o seu curso como 1ª Opção?

Vamos agora realizar o estudo inferencial entre a variável **1ª Opção** e o **número de horas de estudo por semana** por parte dos estudantes envolvidos no estudo.

Neste estudo, foram considerados **142 estudantes**, sendo que **48%** não teve o seu curso como 1ª Opção, e os restantes **52%** tiveram o curso que frequentam como 1ª Opção. Os estudantes que registaram um Número de horas de Estudo superior foram os alunos em que o curso foi 1ª Opção, com uma **média de 10.05 ± 11.08 horas de estudo por semana**, enquanto que os alunos que não tiveram o curso que frequentam como 1ª Opção registaram uma média de **9.32 ± 6.17 horas de estudo por semana** como se pode observar na tabela 4.3. Foram registados **8 outliers**, sendo 3 deles outliers moderados superiores, que pertencem a 2 estudantes que tiveram o seu curso como 1ª Opção (22 e 28 horas de estudo por semana) e 1 por parte de um estudante que não teve o seu curso como 1ª Opção (30 horas de estudo por semana), os outros 5 outliers encontrados pertencem todos aos estudantes que tiveram o seu curso como 1ª Opção e tratam-se de outliers superiores severos (30,33,40,44 e 70 horas de estudo por semana), como se pode observar na tabela 4.4. Através do diagrama de extremos e quartis representado na figura 4.6 podemos observar que, assim como para os estudantes que tiveram o curso como 1ª Opção como aqueles que não tiveram, em ambos os casos vemos que as horas de estudo por semana provém de uma distribuição **assimétrica positiva**, ou seja, os estudantes tendem a ter um número de horas de estudo por semana inferior ao valor médio acima definido, porém essa distribuição é mais acentuada nos estudantes que tiveram o seu curso como 1ª Opção.

Assim, em termos descritivos, pode-se concluir que o número de horas de estudo por semana não difere com o facto dos estudantes terem ou não o seu curso como 1ª Opção.

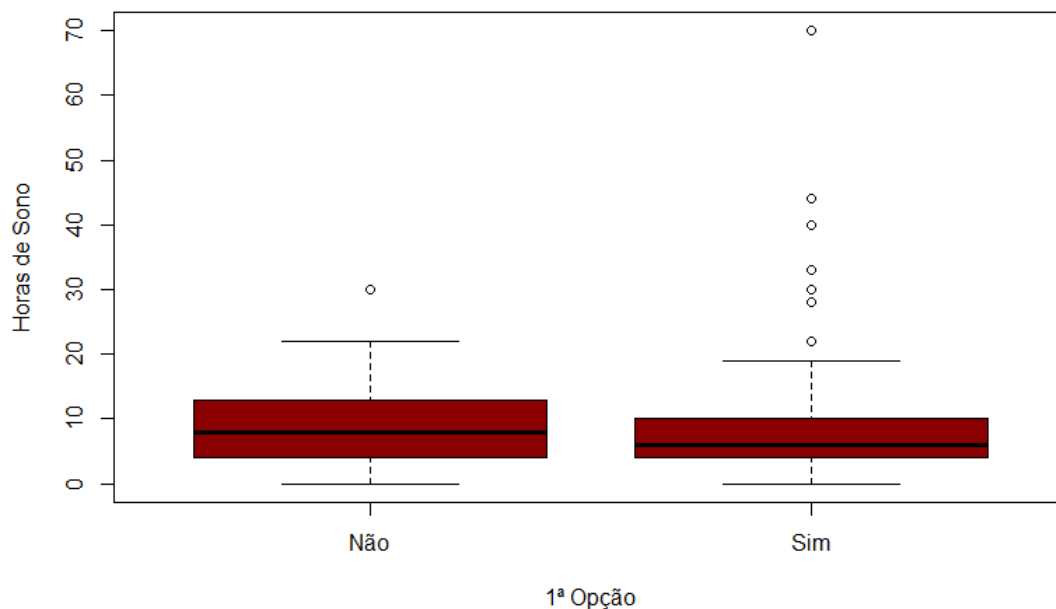
Tabela 4.3: Resumo da estatística de teste para as variáveis número de horas de estudo e 1ª opção

1 Opção	variable	n	min	max	q1	median	q3	mean	sd
Não	HorasEstudo	69	0	30	4	8	13	9.326	6.172
Sim	HorasEstudo	73	0	70	4	6	10	10.05	11.08

Tabela 4.4: Identificação de outliers

1 Opção	HorasEstudo	is.outlier	is.extreme
Não	30	TRUE	FALSE
Sim	44	TRUE	TRUE
Sim	30	TRUE	TRUE
Sim	40	TRUE	TRUE
Sim	22	TRUE	FALSE
Sim	70	TRUE	TRUE
Sim	28	TRUE	FALSE
Sim	33	TRUE	TRUE

Figura 4.6: Diagrama de extremos e quartis



Pretendemos verificar se, para $\alpha=0.05$, existem diferenças significativas entre os valores médios de uma variável quantitativa (Horas de Estudo por semana) em 2 grupos independentes (1ª Opção).

O teste mais adequado será o teste paramétrico **t-Student** para amostras independentes, porém, te-

remos de aplicar um teste á normalidade de modo a saber se podemos avançar com este teste.

Para que possamos verificar o pressuposto de normalidade, iremos aplicar o teste de **Kolmogorov-Smirnov**, uma vez que $n > 30$. Para um nível de significância $\alpha = 0.05$, pode-se considerar que o números de de horas de estudo dos alunos não provém de uma **distribuição normal**, como podemos observar na figura 4.7.

Não se verificando o pressuposto de normalidade, não será possível aplicar o teste t-Student de forma correta, sendo necessário recorrer á sua alternativa não paramétrica, o teste de **Wilcoxon-Mann-Whitney** (figura 4.8) e de modo a saber qual o valor do teste de Z de modo a sabermos a dimensão do efeito (figura 4.9).

Figura 4.7: Teste á normalidade Kolmogorov-Smirnov

```
data: EscalaBurnoutGrupo2$HorasEstudo and EscalaBurnoutGrupo2$`1 opção`
D = 0.92958, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figura 4.8: Teste de Wilcoxon-Mann-Whitney

```
wilcoxon rank sum test with continuity correction

data: HorasEstudo by 1 opção
W = 2734, p-value = 0.3786
alternative hypothesis: true location shift is not equal to 0
```

Figura 4.9: Valor do teste Z do teste de Wilcoxon-Mann-Whitney

	U	Z	p
	<dbl>	<dbl>	<dbl>
1	2734	-0.880	0.379

A avaliação do número de horas de estudo (variável dependente, quantitativa) dos estudantes que tiveram o curso como 1ª Opção com os estudantes que não tiveram o seu curso como 1ª Opção foi avaliado através do teste de Wilcoxon-Mann-Whitney. A utilização deste teste é justificado pelo facto de, pela aplicação do teste de normalidade de Kolmogorov Smirnov não ter sido verificado o pressuposto de normalidade ($p = 2.2e - 16$).

Este estudo permitiu verificar que, os estudantes que tiveram o seu curso como 1ª Opção possuem um número médio de horas de estudo por semana de 10.05 horas, enquanto que os estudantes que não tiveram o seu curso como 1ª Opção possuem apenas um tempo médio de 9.32 horas de estudo por semana.

O estudo inferencial, realizado através do teste de Wilcoxon-Mann-Whitney permitiu concluir que não existem diferenças significativas entre o número de horas de estudo por semana dos estudantes que tiveram e não tiveram o seu curso como 1ª Opção ($W = 2734, Z = -0.880, p = 0.3786$), tratando-se de uma dimensão do efeito fraca ($p = 0.3786$).

4.3 O Número de Horas de Sono varia entre os Cursos em Estudo?

Queremos agora determinar se, o número de horas que os estudantes dormem de 2^a a 6^a feira varia de acordo com o curso que frequentam. Para isso iremos recorrer a uma análise Inferencial onde iremos apresentar primeiramente em termos descritivos.

Neste estudo foram considerados **142 estudantes**, sendo **15%** do curso CTeSP TLQB, **27%** do curso de Bioinformática, e **58%** do curso de Biotecnologia.

Através da tabela 4.5, verificamos que o curso que registou um maior numero de horas de sono diárias de 2^a a 6^afeira foram os estudantes do curso de Biotecnologia, com uma **média** de **7.12±1.01** horas de sono por dia, de seguida o curso de Bioinformática com **6.79±0.84** e o curso com um menor número de horas de sono por dia é o curso CTeSP TLQB com **6.64±1.2** horas. Foram registados a presença de **5 outliers**, sendo todos eles pertencentes aos estudantes do curso de Biotecnologia, os 5 outliers mostraram ser outliers negativos sendo um deles um outlier negativo severo(3 horas) e os restantes sendo outliers moderados negativos (4,5,5,5 horas), como podemos observar na tabela 4.6. Através do diagrama de extremos e quartis da figura 4.10 verificamos que as horas de sono dos estudantes do curso de Bioinformática possui uma distribuição **assimétrica negativa**, enquanto que os estudantes dos cursos de Biotecnologia e de CTeSP TLQB possuem uma distribuição **assimétrica positiva**, ou seja, os estudantes de Bioinformática tendem a dormir mais horas do que o valor da média anteriormente calculada, enquanto que os estudantes dos restantes cursos tendem a dormir menos horas que a média acima calculada.

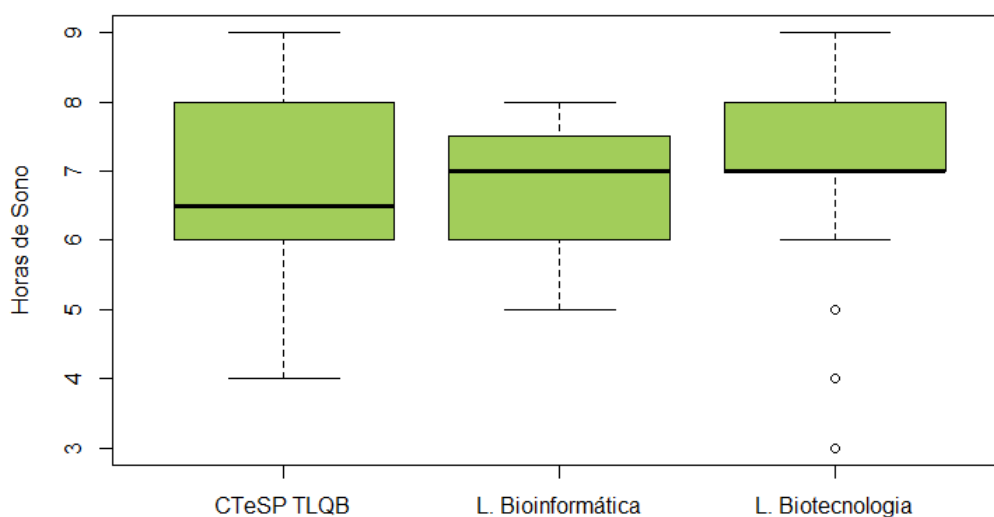
Tabela 4.5: Resumo da estatística de teste para as variáveis "Curso" e "Horas de Sono".

Curso	variable	n	min	max	median	q1	q3	iqr	mad	mean	sd	se	ci
CTeSP TLQB	HorasSono	22	4	9	6.5	6	7.75	1.75	0.741	6.636	1.197	0.255	0.531
L. Bioinformática	HorasSono	38	5	8	7	6	7.5	1.5	1.483	6.789	0.835	0.136	0.275
L. Biotecnologia	HorasSono	82	3	9	7	7	8	1	1.483	7.118	1.011	0.112	0.222

Tabela 4.6: Identificação de outliers.

Curso	HorasSono	is.outlier	is.extreme
L. Biotecnologia	4	TRUE	FALSE
L. Biotecnologia	5	TRUE	FALSE
L. Biotecnologia	3	TRUE	TRUE
L. Biotecnologia	5	TRUE	FALSE
L. Biotecnologia	5	TRUE	FALSE

Figura 4.10: Diagrama de extremos e quartis.



Pretendemos então verificar se, para $\alpha=0.05$, existem diferenças significativas entre os valores médios de uma variável quantitativa (Horas de Sono) em 3 grupos independentes (3 Cursos existentes).

O teste mais adequado será o teste paramétrico **ANOVA one-way** para amostras independentes, porém, antes de podermos avançar é necessário verificar alguns pressupostos, nomeadamente a normalidade. Para que possamos verificar se a população dos estudantes de cada curso provém de uma distribuição normal é necessário selecionar o teste mais adequado, uma vez que, as dimensões da nossa amostra são 22, 38 e 82, iremos aplicar 2 testes de normalidade de Kolmogorov-Smirnov e um teste de normalidade de Shapiro-Wilk para os estudantes do curso CTeSP TLQB, como mostram as figuras 4.11, 4.12 e 4.13. Uma vez que não foi verificado o pressuposto de normalidade, iremos prosseguir para a alternativa não paramétrica ao teste ANOVA one-way, o teste de **Kruskal-Wallis**, como mostra a figura 4.14.

shapiro-wilk normality test

```
data: Dados3$HorasSono[Dados3$Curso == "CTeSP TLQB"]
w = 0.95444, p-value = 0.3857
```

Figura 4.11: Teste de normalidade de Shapiro-Wilk para os estudantes de CTeSP TLQB

Lilliefors (kolmogorov-smirnov) normality test

```
data: Dados3$HorasSono[Dados3$Curso == "L. Bioinformática"]
D = 0.22244, p-value = 5.466e-05
```

Figura 4.12: Teste de Kolmogorov-Smirnov aos estudantes de Bioinformática

```

Lilliefors (Kolmogorov-Smirnov) normality test

data:  Dados3$HorasSono[Dados3$Curso == "L. Biotecnologia"]
D = 0.23393, p-value = 4.738e-12

```

Figura 4.13: Teste de Kolmogorov-Smirnov aos estudantes de Biotecnologia.

```

kruskal-wallis rank sum test

data:  HorasSono by Dados3$Curso
kruskal-wallis chi-squared = 7.5023, df = 2, p-value = 0.02349

```

Figura 4.14: Teste de Kruskal-Wallis.

A avaliação entre o número de horas de sono de 2^a a 6^afeira (variável dependente, quantitativa) dos estudantes e os 3 diferentes cursos pelos quais estão distribuídos foi avaliado através do teste de Kruskal-Wallis. A utilização deste teste é justificada pelo facto de, na tentativa de aplicação do teste paramétrico ANOVA one-way para amostras independentes, ao aplicarmos o teste de normalidade de Kolmogorov-Smirnov e de Shapiro-Wilk, podemos observar que não era possível verificar o pressuposto de normalidade ($\rho=4.738e-12$).

Este estudo permitiu verificar que os estudantes do curso de Biotecnologia possuem um número de horas de sono médio de 7.12 horas, os estudantes de Bioinformática possuem um número médio de 6.79 horas e os estudantes de CTeSP TLQB possuem um número médio de horas de sono de 6.64 horas. O estudo inferencial, realizado através do teste de Kruskal-Wallis (figura 4.14) permitiu concluir que existem diferenças significativas entre o número de horas de estudo dos estudantes de Bioinformática e os estudantes dos restantes cursos ($r^2=7.5$, $df=2$ e $\rho=0.02$).

5 Métodos de Análise de Dados Multivariados

Vamos agora avançar para a regressão linear simples, nesta regressão são utilizados conceitos e técnicas para analisar e utilizar a relação linear entre duas variáveis. Na análise das regressões a que iremos proceder, estas resultam de uma equação que pode ser utilizada para "prever" possíveis valores de uma variável dependente e de uma variável independente. Na regressão linear múltipla assume-se que existe uma relação linear entre uma variável Y (a variável dependente) e k variáveis independentes (estas são aquelas que podem ser várias se for necessário).

Para a análise do questionário que nos foi fornecido, iremos escolher como variável dependente o **Número de horas de sono**. Esta será a variável dependente para ambos os estudos que vamos realizar, iremos apenas alterar as variáveis independentes consoante o estudo que estamos a fazer.

5.1 As Horas de sono dos estudantes pode ser afetada pelas Horas de estudo e o curso que frequentam?

Com este estudo pretendemos verificar se existe algum tipo de relação entre as variáveis horas de sono(v.a dependente), horas de estudo e curso que frequenta(v.a independentes).

Antes que possamos avançar para os métodos e as análises, é necessário recorrer á criação de uma **variável dummy**, uma vez que a variável Curso se trata de uma variável qualitativa, esta não poderá ser incluída no modelo em estudo, para tal iremos utilizar as variáveis dummy, para possamos incluir esta variável no modelo, tornando-as variáveis quantificadas. Logo a variável curso é representada da seguinte forma:

Variável Qualitativa:

- Biotecnologia
- Bioinformática
- CTeSP TLQB

Variável Dummy:

$$\text{Biotec} = \begin{cases} 1 & \text{se Biotec} \\ 0 & \text{se caso contrario} \end{cases}$$
$$\text{Bioinf} = \begin{cases} 1 & \text{se Bioinf} \\ 0 & \text{se caso contrario} \end{cases}$$

Após a identificação das variáveis, prosseguimos á construção do modelo:

$$Y_{HSono} = \beta_0 + \beta_{HEstudo} * x_{HEstudo} + \beta_{Biotec} + \beta_{Bioinf} * \text{Bioinf}$$

Procedida a fase de construção do modelo, realizamos uma tabela do modelo de regressão. Ao analisar a tabela(tabela 5.1) podemos conferir as estimativas dos parâmetros do modelo de regressão, os valores obtidos nessas estimativas foram atribuídos ao nosso modelo que foi feito anteriormente:

$$HSono = 6.712 - 0.009 * X_{HEstudo} + 0.501 * \text{Biotec} + 0.182 * \text{Bioinf}$$

De acordo com este modelo podemos interpretá-lo da seguinte forma.

As horas de Sono dos estudantes de Biotecnologia é inferior ás Horas de Sono dos estudantes de CTeSP TLQB. A diferença média estimada das Horas de Sono entre esses estudantes é **0.009**.

As Horas de Sono dos estudantes de Bioinformática é inferior ás Horas de Sono dos estudantes de CTeSP TLQB. A diferença média das Horas de Sono entre esses estudantes é **0.182**.

O acréscimo médio nas Horas de Sono corresponde ao decréscimo de uma unidade nas Horas de Estudo($\beta_{HEstudo}$) é o mesmo para os diferentes casos.

Após interpretar-mos o modelo podemos partir para a análise da **significância prática e estatística**

do modelo de regressão(5.1), isto é, vamos verificar se o modelo possui ou não **significância estatística**. Para fazer esta análise iremos verificar na tabela os valores que foram obtidos pelo teste de significância da anova.

Após observar os valores(tabela 5.1) podemos concluir que, para um nível de significância de 5%, o modelo feito anteriormente **não é estatisticamente significativo**($F_{(3,138)}=2.187$; $\rho=0.092$; $R_a^2=0.025$). De seguida vamos realizar a análise das variáveis independentes em estudo usadas no ajustamento do modelo para verificar se estas variáveis possuem efeito significativo sobre as Horas de sono dos Estudantes(5.1).

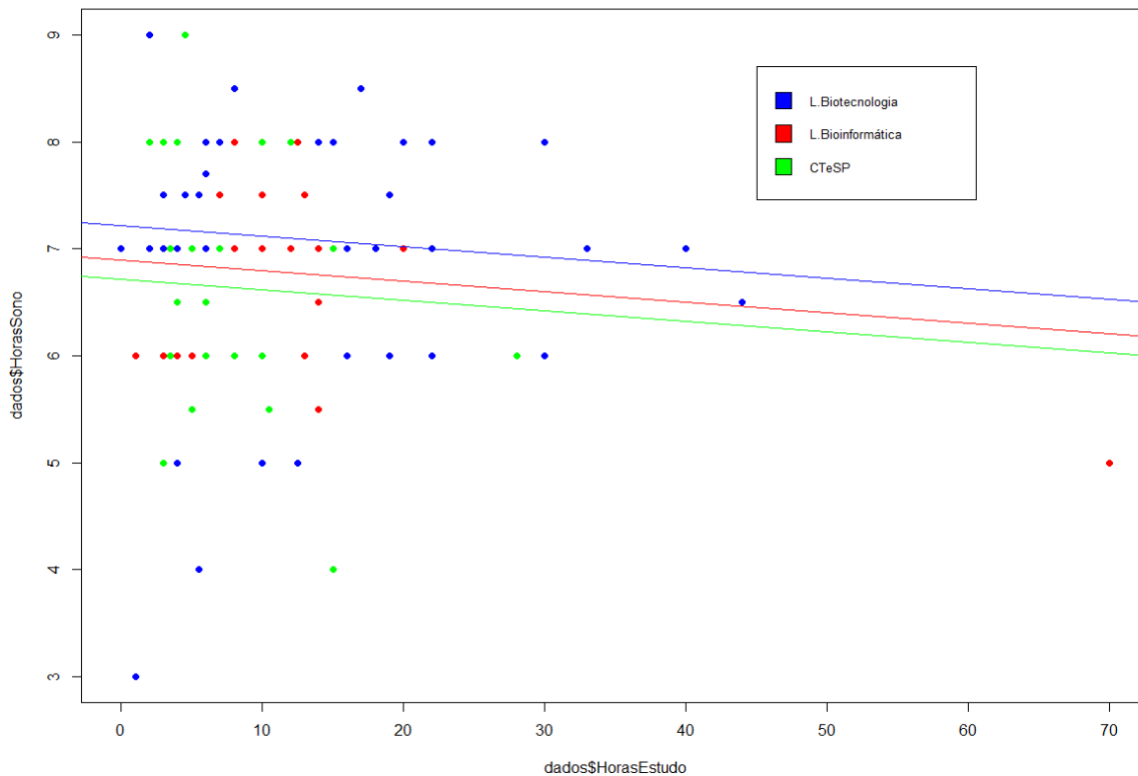
Tabela 5.1: Tabela dos Coeficientes

```
Residuals:
    Min       1Q   Median       3Q      Max
-4.2037 -0.6744 -0.0472  0.8403  2.3316

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.712400    0.225094   29.820  <2e-16 ***
HorasEstudo   -0.009783    0.009386   -1.042    0.2991
Biotecnologia  0.501093    0.240516    2.083    0.0391 *
Bioinformatica 0.182235    0.269031    0.677    0.4993
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9988 on 138 degrees of freedom
Multiple R-squared:  0.04538,    Adjusted R-squared:  0.02463
F-statistic: 2.187 on 3 and 138 DF,  p-value: 0.09236
```

Figura 5.1: Gráfico do modelo



Após interpretar a expressão do modelo de regressão linear múltipla que nos permite analisar a hipótese em estudo, iremos desta forma verificar se os pressupostos do modelo obtido anteriormente são verificados. O gráfico dos diagramas permite-nos avaliar a linearidade e a Homocedasticidade(figura 5.3). Assim sendo de acordo com a análise feita pelo gráfico diz-nos que o modelo possui **linearidade**, isto é, como a linha horizontal encarnada está acima da linha horizontal a tracejada, pode-se considerar que existe linearidade.A análise também mostra que o modelo em estudo analisado anteriormente também possui Homocedasticidade, ou seja, as distâncias (dispersão) dos dados são aproximadamente constantes ao longo dos valores previstos de Y, portanto os dados estão dispostos com uma forma semelhante a um retângulo.

Após a verificação da análise da normalidade dos resíduos diz-nos também que o modelo possui normalidade, isto é, os resíduos distribuem-se segundo uma lei normal, estes **não estão muito afastados da linha tracejada**.

Por fim, podemos também verificar a existência de outliers, que, de acordo com o gráfico existem outliers no nosso modelo, ou seja, existem resíduos com valores **inferiores a -3**.

A análise do gráfico também conseguimos ver que o modelo construído **não possui pontos influentes**, isto é, não são verificados resíduos fora da linha encarnada a tracejada na qual é calculada com base na distância de Cook.

Após a análise dos valores e para um nível de significância de 5%, somente a variável "Biotecnolo-

gia”possuí efeito significativo sobre os valores ”Horas de Sono”dos estudantes, ou seja, a ordenada na origem da reta da reta de regressão populacional dos estudantes de CTeSP($t_{(138)}=2.083$; $\rho=0.0391$).

Após este estudo, foi feito um gráfico sobre este modelo também foi realizada uma análise sobre o respetivo gráfico. Posteriormente da análise do gráfico foi constatado um **paralelismo** entre as retas de regressão, isto deve-se ao facto de não se ter considerado o **fator de interação entre as variáveis**. Por fim, foi realizado uma comparação entre dois modelos, para essa comparação utilizamos os critérios de **AIC** e **BIC** que nos dizem que, quanto menor o valor de AIC e BIC melhor será o ajustamento do modelo aos nossos dados.

Os modelos utilizados para essa compração foi o modelo em estudo e um modelo criado(figura 5.4) apenas com a variável Horas de Estudo.

$$\text{Modelo analisado} = Y_{HSono} = 7.043 - 0.009 * x_{HEstudo}$$

Após a realização da análise dos valores obtidos dos critérios de AIC e BIC(figura 5.2) podemos concluir que o modelo em estudo possui um melhor ajustamento aos nossos dados. Porém para o critério de BIC o modelo analisado possui um melhor ajustamento para os nossos dados.

Entretanto podemos também realizar a comparação de modelos alinhados pelo teste da anova, isto é, saber se os dois modelos são ou não equivalentes.

Neste caso, de acordo com a análise dos dados obtidos podemos concluir que os dois modelos são **equivalentes**.

```
> AIC(modelo, modelo_hs_1)
      df      AIC
modelo    3 410.2487
modelo_hs_1 5 408.5767
> BIC(modelo, modelo_hs_1)
      df      BIC
modelo    3 419.1161
modelo_hs_1 5 423.3558
> |
```

Figura 5.2: Valores de AIC e BIC

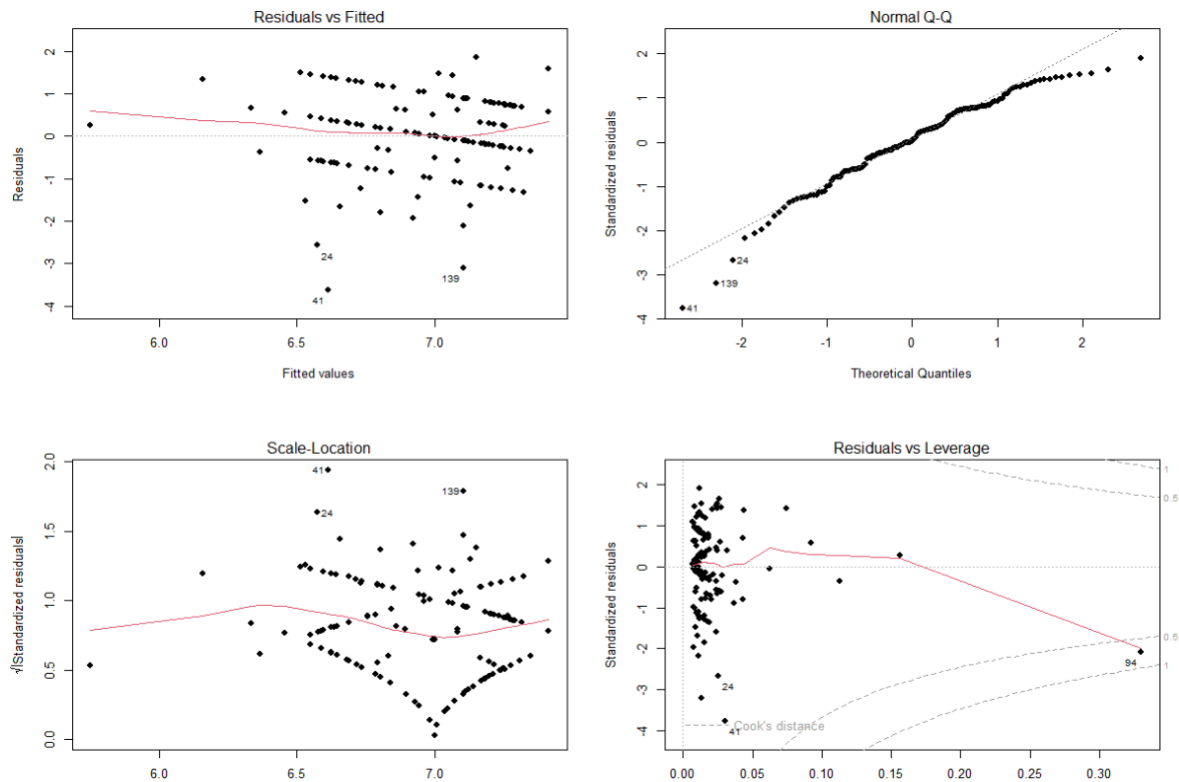


Figura 5.3: Modelo gráfico para a verificação de pressupostos

```
call:
lm(formula = HorasSono ~ HorasEstudo, data = dados)

Residuals:
    Min       1Q   Median       3Q      Max
-4.0343 -0.8919  0.0471  0.9838  1.9974

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.043293   0.124982  56.354  <2e-16 ***
HorasEstudo -0.009036   0.009456  -0.956   0.341
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.012 on 140 degrees of freedom
Multiple R-squared:  0.006481, Adjusted R-squared:  -0.0006152
F-statistic: 0.9133 on 1 and 140 DF, p-value: 0.3409
```

Figura 5.4: Construção do Modelo

5.2 As Horas de Sono são afetadas pelo conhecimento do programa de Mentoria e o Tempo de Deslocação

Com este próximo estudo pretendemos averiguar se existe algum tipo de relação entre as variáveis horas de sono(v.a dependente), horas de estudo e conhecimento do programa de mentoria(v.a independentes). Antes que possamos avançar para os métodos e análises, é necessário recorrer á criação de uma **variável dummy**, uma vez que a variável Programa de Mentoria se trata de uma variável qualitativa, para que esta v.a possa ser incluída no modelo, iremos torna-la em uma variável quantificada. Logo a variável Programa de Mentoria é representada da seguinte forma:

Variável Qualitativa:

- Sim
- Não

Variável Dummy:

$$\text{Biotec} = \begin{cases} 1 & \text{se Sim} \\ 0 & \text{se Não} \end{cases}$$

Após a identificação das variáveis, prosseguimos á construção do modelo:

$$Y_{HSono} = \beta_0 + \beta_{PMentoria} * x_{Sim} + \beta_{TDeslocacao} * TDeslocação$$

Após a construção do modelo criamos uma tabela do modelo de regressão (tabela 5.2).

Após a análise da tabela com as estimativas dos parâmetros do modelo de regressão, pegamos nesses valores obtidos e atribuímos ao nosso modelo feito anteriormente:

$$Y_{HSono} = 7.143 - 0.003 * x_{TDeslocacao} + 0.328 * x_{Sim}$$

O acréscimo médio nas Horas de Sono é o correspondente ao **decréscimo de uma unidade** do Tempo de Deslocação.

Através da interpretação deste modelo podemos realizar a análise da significância prática e estatística do modelo de regressão, agora pretendemos verificar se o modelo possui ou não significância estatística. Para ser feita esta análise, recorreremos aos valores obtidos na tabela 5.2 pelo teste da Anova. Após a aobservação dos valores é possível concluir que, para um nível de significância de 5%, o modelo é estatisticamente significativo, mas a sua significância prática é **baixa**, cerca de **8%**($F_{(2,139)}$; $\rho = 0.00074$; $R_a^2 = 0.085$).

Agora vamos realizar a análise do modelo das variáveis em estudo usadas no ajustamento do modelo para verificar se estas variáveis possuem efeito significativo sobre Horas de Sono dos estudantes.

Para isto, vamos utilizar os valores das estatísticas de teste e p-value do teste t-student para coeficientes de modelos de regressão.

Após a análise dos valores, temos que, para um nível de significância de 5%, podemos averiguar que ambas as variáveis **possuem efeito significativo** sobre as Horas de Sono($T_{(28)}$; $\rho = 0.0937$; $T_{(28)}$; $\rho = 0.002$).

Após este estudo ser feito, foi criado um gráfico sobre esse modelo, e uma análise da mesma(figura

5.5).E foi possível verificar se existe paralelismo entre as retas de regressão, ou seja, a existência desse paralelismo deve-se possivelmente ao fator de interação entre as variáveis.

Por isso também será testado a interação entre as v.a(5.8).

Residuals vs Fitted

Este é o diagrama de dispersão dos valores estimados vs resíduos que permite verificar se a existência de linearidade no caso de a linha vermelha estar por cima da linha tracejada, e homocedisticidade se as distâncias dos dados forem aproximadamente constantes ao longo dos valores de Y.

Com este diagrama é possível determinar que o modelo é **linear**, mas não o apresenta, mas não existe Homocedisticidade, visto que os valores de dispersão dos dados são muito grandes.

QQ-plot Normalidade

Este diagrama permite verificar se os resíduos se distribuem segundo uma lei normal. Os resíduos não se distribuem segundo uma lei normal porque estão afastados da linha tracejada.

Diagrama de localização da escala

Apesar do Diagrama de dispersão dos valores estimados vs resíduos conseguir determinar a Linearidade e Homocedisticidade, o Diagrama localização da escala é mais recomendado para avaliar a homocedisticidade. Como foi confirmado anteriormente o modelo não apresenta Homocedisticidade, visto que a linha vermelha não tem um formato horizontal.

Diagrama de leverage vs resíduos

Este gráfico permite verificar se os resíduos são outliers. Se existirem outliers eles corresponderão a resíduos com valor inferior a -3 ou com valores superiores a 3. Com este modelo é possível verificar que existem 3 outliers localizados no -3, e também possível verificar que não apresentam nenhum ponto influente.

```

Call:
lm(formula = Horassono ~ TempoDesloca + Sim, data = dados)

Residuals:
    Min       1Q   Median       3Q      Max
-3.3628 -0.5759  0.1133  0.6824  1.8762

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.143350   0.181738  39.306 < 2e-16 ***
TempoDesloca -0.003252   0.001030  -3.158  0.00195 **
Sim          0.328756   0.165808   1.983  0.04937 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9671 on 139 degrees of freedom
Multiple R-squared:  0.09846,    Adjusted R-squared:  0.08549
F-statistic: 7.591 on 2 and 139 DF,  p-value: 0.0007436

```

Tabela 5.2: Tabela do modelo de regressão

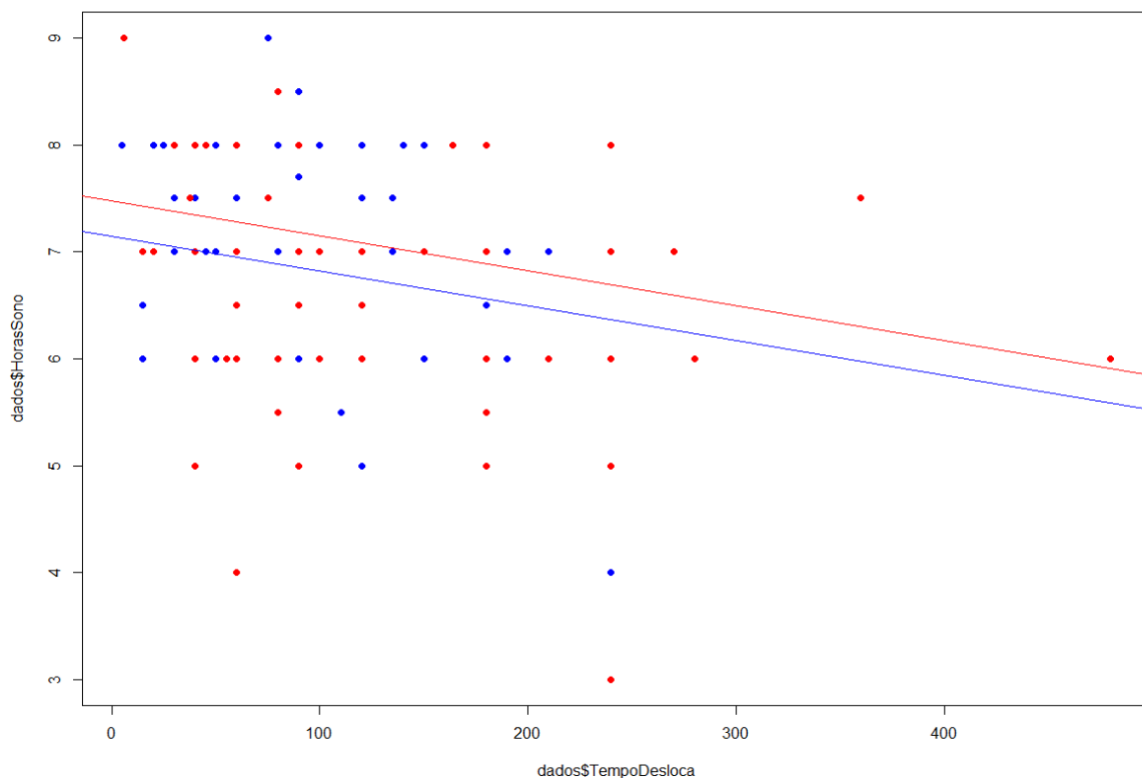


Figura 5.5: Gráfico sobre o modelo em estudo

Após as análises feitas, prosseguimos á criação de outro modelo para verificar o efeito de interação entre as duas variáveis:

$$Y_{H\text{Sono}} = 7.07 + 0.049 * x_{Sim} - 0.003 * x_{TDesloca} - 0.001 * x_{TDesloca} * x_{Sim}$$

Para um nível de significância de 5%, pode-se considerar que a interação "Tempo de Deslocação" e "Programa de Mentoria" dos estudantes **não tem efeito significativo** sobre as "Horas de Sono" dos estudantes ($T_{(26)} = -0.662; p = 0.509$).

Por fim foi realizado uma comparação entre os dois modelos utilizando critérios de AIC e BIC, onde quanto menor for o valor de AIC/BIC melhor o ajustamento do modelo aos nossos dados (figura 5.6). Após a realização da análise dos valores obtidos dos critérios de AIC e BIC podemos concluir que o modelo em estudo tem o melhor ajustamento aos nossos dados tanto para AIC como para BIC.

Também foi realizado a comparação dos modelos aninhados pelo teste da anova, isto é, para verificar se os modelos são ou não equivalentes, neste caso, os modelos não são equivalentes, logo, o "melhor" modelo será aquele com um maior número de parâmetros (figura 5.7).

Figura 5.6: Valores de AIC e BIC

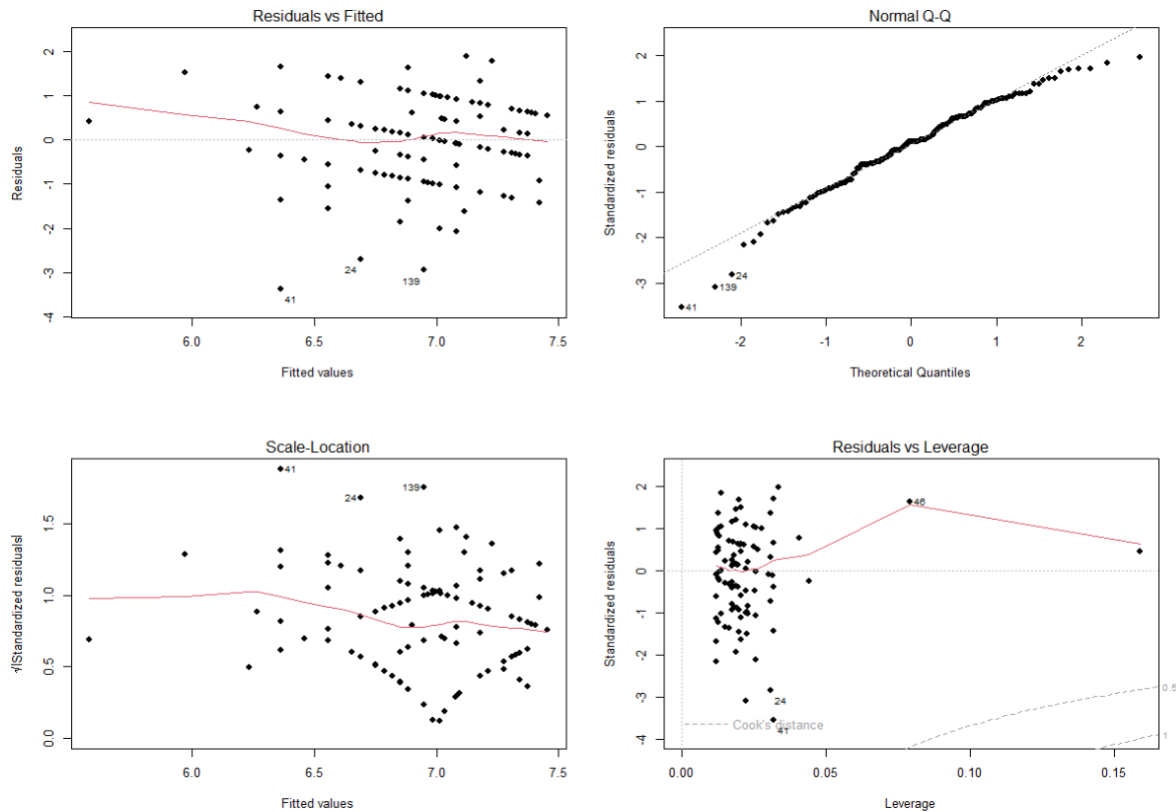
```
> AIC(modelo, modelo_sim)
      df      AIC
modelo    3 400.4134
modelo_sim 4 398.4530
> BIC(modelo, modelo_sim)
      df      BIC
modelo    3 409.2809
modelo_sim 4 410.2763
```

Figura 5.7: Teste anova para comparação dos modelos

Analysis of Variance Table

```
Model 1: Horassono ~ TempoDesloca
Model 2: Horassono ~ TempoDesloca + Sim
      Res.Df    RSS Df Sum of Sq    F Pr(>F)
1         140 133.69
2         139 130.01   1     3.6771 3.9313 0.04937 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 5.8: Teste anova para comparação dos modelos



6 Análise Fatorial

Vamos agora realizar o estudo em relação as 15 perguntas de Burnout que se encontram no fim do questionário em estudo, para isso iremos recorrer à Análise Fatorial.

A análise Fatorial divide-se em pelo menos dois modelos estatísticos, a Análise Fatorial Exploratória (AFE), e a Análise Fatorial Confirmatória (AFC).

A análise fatorial exploratória é uma técnica que tem como objetivo descobrir e analisar a estrutura de um conjunto de dados de variáveis interrelacionadas, de modo a reduzir a dimensionalidade dos dados, ou seja, manter o máximo de informação e qualidade com o menor número de fatores possíveis.

A partir da AFE obtemos um modelo específico, onde cada variável é expressa como função de um conjunto de fatores comuns às variáveis e de um fator que define a especificidade dessa variável, onde se conclui se são boas ou más as variáveis.

A Análise Fatorial Confirmatória é uma estatística multivariada que serve para estimar o quão bons são os dados e se os mesmos se ajustam aos mais diversos modelos.

Para o nosso estudo será utilizado exclusivamente a análise fatorial exploratória (AFE).

Os nossos principais objetivos com este teste é encontrar tantos fatores quanto os grupos de correlação semelhantes entre si e diferentes de outros grupos), queremos também verificar se as variáveis estiverem


todas correlacionadas entre si de modo idêntico (vamos ter apenas um fator).

Porém, antes de avançarmos para a análise fatorial é necessário verificar alguns pormenores primeiro, tais como a **Dimensão da amostra**, precisamos de verificar se a amostra possui uma dimensão suficiente (pelo menos 50 observações e o rácio entre observações não deve ser inferior a 5), de seguida precisamos de verificar se as correlações entre as variáveis originais são elevadas, depois realizar a análise da matriz de correlações (apenas possível se o número de variáveis for reduzido), calcular a medida de adequabilidade de Kaiser-Mayer-Olkin (KMO) e realizar o teste de esfericidade.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
P1	1.00														
P2	0.65	1.00													
P3	0.42	0.65	1.00												
P4	0.48	0.57	0.58	1.00											
P5	0.61	0.65	0.52	0.62	1.00										
P6	0.31	0.39	0.54	0.46	0.46	1.00									
P7	0.17	0.30	0.45	0.25	0.20	0.64	1.00								
P8	0.22	0.41	0.52	0.53	0.46	0.58	0.42	1.00							
P9	0.17	0.23	0.46	0.48	0.38	0.49	0.37	0.68	1.00						
P10	-0.02	0.01	0.06	-0.03	0.01	0.04	0.01	0.08	0.08	1.00					
P11	0.18	0.10	0.01	0.14	0.25	0.01	-0.19	0.10	0.07	0.36	1.00				
P12	0.04	0.11	0.00	-0.01	0.08	-0.04	-0.20	-0.09	-0.13	0.39	0.34	1.00			
P13	0.01	0.10	-0.04	-0.03	0.07	-0.12	-0.31	0.04	-0.09	0.13	0.25	0.37	1.00		
P14	-0.04	-0.01	-0.16	-0.13	0.03	-0.33	-0.42	-0.32	-0.36	0.03	0.23	0.20	0.41	1.00	
P15	-0.14	-0.08	-0.19	-0.23	-0.22	-0.25	-0.28	-0.07	-0.13	0.27	0.23	0.41	0.32	0.17	1.00

Figura 6.1: Análise da Matriz de Correlação de Pearson

Através da figura 6.1 podemos verificar os pressupostos anteriormente referidos, o pressuposto da dimensão da amostra fica verificado uma vez que o tamanho da amostra é suficiente, que no caso a sua dimensão é **142**, é possível verificar que existe uma amostra com correlações de variáveis elevadas.

Outra das maneiras que podíamos usar para interpretar os mesmos resultados obtidos na figura 6.1 podia ser, por exemplo a utilização da interface gráfica do  e produzirmos uma matriz do seguinte tipo:

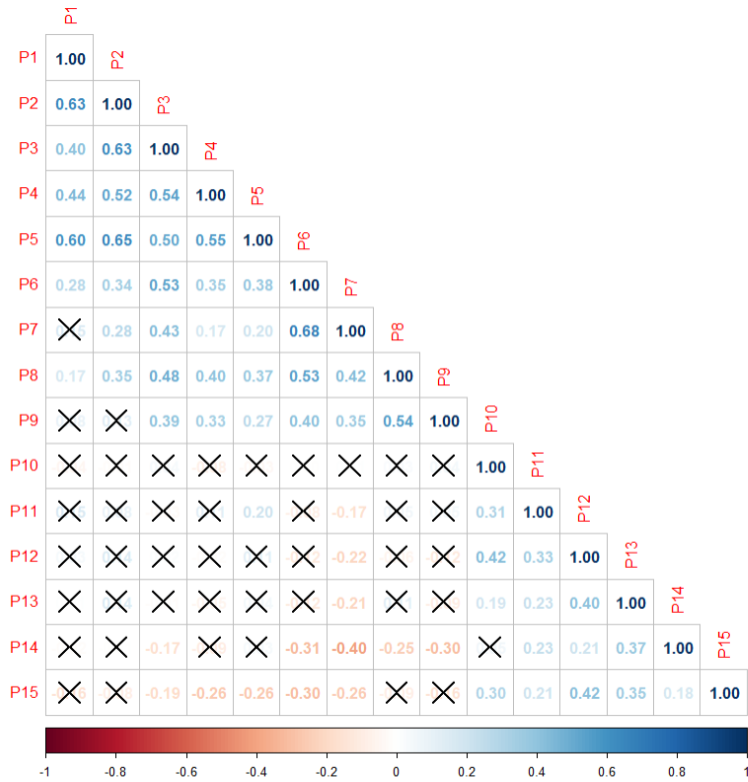


Figura 6.2: Análise da Matriz de Correlação de Pearson

Através desta representação podemos retirar algumas conclusões, tais como, quanto maior os números azuis, maior a correlação, e quanto maior o valor dos números a encarnado, maior a correlação porém neste caso é inversamente proporcional, as cruces que aparecem são relativamente a correlações muito baixas e pouco relevantes para o estudo.

De seguida calculamos o valor do **KMO**, este valor trata-se de um valor empírico e permite avaliar a homogeneidade entre as variáveis, como o valor do KMO é 0.82(figura 6.3), possuímos uma boa AFE, lembrando que este valor varia entre 0 e 1, podemos presumir que teremos bons coeficientes de correlação, podemos então prosseguir com a nossa análise.

Relativamente às variáveis (figura 6.3), estas também variam entre 0 e 1 e a correlação mais baixa encontrada toma o valor de 0.65 o que nos indica que todas as variáveis são validas para o estudo e também apresentam valores bastantes bons.

Com a obtenção do valor mínimo de 0.65, fica também cumprido a adequabilidade, que é necessária possuir um valor superior a 0.6.


```

Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = Dados3)
Overall MSA = 0.82
MSA for each item =
  P1  P2  P3  P4  P5  P6  P7  P8  P9  P10 P11 P12 P13 P14 P15
0.83 0.79 0.89 0.91 0.84 0.86 0.80 0.82 0.83 0.65 0.73 0.68 0.69 0.75 0.75

```

Figura 6.3: Medida de adequabilidade de Kaiser-Mayer-Olkin

Vamos agora prosseguir para o **teste de esfericidade de Bartlett**, aplicação deste teste busca testar a hipótese de a matriz de correlações ser a matriz identidade, ou seja, testa a hipótese das correlações entre as variáveis serem nulas.

```

$chisq
[1] 818.3756

$p.value
[1] 6.556378e-111

$df
[1] 105

```

Figura 6.4: Teste de Esfericidade(Bartlett)

Como é possível verificar na figura 6.4, o valor de ρ é extremamente pequeno, sendo menor que α que assumimos inicialmente como 0.05, assim rejeitamos a hipótese da matriz de correlação ser a matriz identidade.

Sendo assim todos os pressupostos estão verificados, podemos então prosseguir com a aplicação do AFE ao conjunto de dados.

```

Standard deviations:
Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9  Comp.10  Comp.11  Comp.12  Comp.13  Comp.14  Comp.15
2.2049559 1.6137954 1.2424282 0.9815485 0.9430115 0.8570989 0.7544714 0.7494833 0.7001157 0.6267114 0.5840568 0.5653287 0.5436466 0.4922669 0.4363494

15 variables and 142 observations.

```

Figura 6.5: Desvio padrão das variáveis

Com uma rápida análise(figura 6.5) podemos observar os valores do desvio padrão das variáveis que serão utilizadas para o AFE. E como referido anteriormente o número de observações (142).

Critério de Kaiser

Ao longo da unidade curricular, aprendemos que existem várias regras empíricas que são utilizadas para ajudar na tomada de decisão relativamente ao número correto de fatores a considerar, neste estudo iremos utilizar o **Critério de Kaiser**, teríamos ainda opções como o Critério de Scree plot que é a representação gráfica de fatores.

Importance of components:												
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12
Standard deviation	2.204956	1.6137954	1.2424282	0.98154847	0.94301145	0.85709893	0.75447139	0.74948333	0.70011573	0.62671144	0.58405679	0.56532875
Proportion of Variance	0.324122	0.1736224	0.1029085	0.06422916	0.05928471	0.04897457	0.03794847	0.03744835	0.03267747	0.02618448	0.02274149	0.02130644
Cumulative Proportion	0.324122	0.4977444	0.6006529	0.66488208	0.72416679	0.77314136	0.81108983	0.84853818	0.88121565	0.90740013	0.93014162	0.95144806
	Comp.13	Comp.14	Comp.15									
Standard deviation	0.54364658	0.49226689	0.43634939									
Proportion of Variance	0.01970344	0.01615511	0.01269339									
Cumulative Proportion	0.97115150	0.98730661	1.00000000									

Figura 6.6: Critério de Kaiser

Tendo em conta a figura 6.6, ao utilizar o critério de Kaiser queremos saber quantos ou quais são os fatores a reter para a nossa solução inicial, isto é, devemos, apenas reter o número mínimo de fatores que nos permita explicar convenientemente o fenómeno e estudo. Assim sendo, a regra de Kaiser nos diz que devemos reter os fatores que expliquem mais informação (variância) do que a informação (standardizada) de uma variável original, ou seja, os valores devem ser superiores a 1. Portanto ao analisar o output do critério de Kaiser foi constatado que devemos reter 3 fatores (Comp. 1, Comp. 2 e Comp. 3) para nos explicar o fenómeno em estudo.

Após a análise do critério de Kaiser, partimos para a construção da solução inicial com o número de fatores sugeridos pela regra de Kaiser, sem rotação (figura 6.7). Após analisar a tabela da solução inicial queremos saber sobre a qualidade da solução, para tal iremos avaliar a solução tendo em conta a percentagem de variabilidade total explicada pelos fatores e pelos valores das comunalidades. A percentagem da variabilidade total explicada pelos fatores nos diz que a solução é admissível quando a percentagem de variabilidade total explicada pelos fatores que constituem a solução deve ser superior ou próximo de 70%, ao analisar os valores do “Cumulative Var” na tabela vemos que os valores de PC3 são de 0.60, isto significa que os 3 primeiros fatores explicam cerca de 60% da variabilidade total dos dados, um valor abaixo do admissível. Os valores das comunalidades nos falam que a solução só é admissível quando os valores das comunalidades são superiores ou próximos de 0.60. Assim sendo, ao observar os valores das comunalidades podemos ver que das 15 variáveis em estudo somente 3 não se aproximam dos valores admissíveis, como são poucas as variáveis que não apresentam os valores admissíveis não há problema em não aceitar os restantes valores. Porém a solução inicial deve ser modificada pois não se verificou um compromisso entre os indicadores anteriormente apresentados. Portanto um método a considerar para ser verificado os compromissos entre os indicadores é aumentar o número de fatores na solução, ou seja, a percentagem de variabilidade total dos dados explicada pelos fatores e os valores das comunalidades também irão aumentar. Sendo assim, aumentamos os números de fatores da solução para 5 (figura 6.8). Ao ter realizado isso verificamos que a percentagem total dos dados explicada pelos fatores aumentou para 72%, um número que está dentro dos parâmetros do admissível. O valor das comunalidades também sofreram alterações no qual todas as variáveis em estudos possuem valores igual ou superior a 0.60, que também está dentro dos parâmetros do aceitável. Em suma, a solução ideal é a solução com 5 fatores.

```
Principal Components Analysis
Call: principal(r = Dados3, nfactors = 3, rotate = "none", n.obs = nrow(Dados3),
  scores = TRUE)
Standardized loadings (pattern matrix) based upon correlation matrix
      PC1  PC2  PC3  h2  u2 com
P1  0.60  0.28 -0.43  0.62  0.38 2.3
P2  0.72  0.33 -0.32  0.72  0.28 1.8
P3  0.79  0.08 -0.02  0.63  0.37 1.0
P4  0.77  0.15 -0.16  0.64  0.36 1.2
P5  0.74  0.33 -0.30  0.75  0.25 1.7
P6  0.77 -0.12  0.22  0.65  0.35 1.2
P7  0.61 -0.40  0.26  0.59  0.41 2.1
P8  0.75  0.03  0.33  0.67  0.33 1.4
P9  0.67 -0.10  0.38  0.61  0.39 1.6
P10 0.01  0.45  0.60  0.56  0.44 1.9
P11 0.07  0.66  0.16  0.47  0.53 1.1
P12 -0.10  0.70  0.25  0.56  0.44 1.3
P13 -0.13  0.66 -0.02  0.45  0.55 1.1
P14 -0.35  0.53 -0.42  0.58  0.42 2.7
P15 -0.33  0.51  0.39  0.51  0.49 2.6

      PC1  PC2  PC3
SS loadings      4.86 2.60 1.54
Proportion var    0.32 0.17 0.10
Cumulative var    0.32 0.50 0.60
Proportion Explained 0.54 0.29 0.17
Cumulative Proportion 0.54 0.83 1.00

Mean item complexity = 1.7
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is 0.08
with the empirical chi square 170.75 with prob < 7e-12

Fit based upon off diagonal values = 0.94
```

Figura 6.7: Solução Fatorial com 3 fatores

```
Principal Components Analysis
Call: principal(r = Dados3, nfactors = 5, rotate = "none", n.obs = nrow(Dados3),
  scores = TRUE)
Standardized loadings (pattern matrix) based upon correlation matrix
      PC1  PC2  PC3  PC4  PC5  h2  u2 com
P1  0.60  0.28 -0.43 -0.29 -0.11 0.72 0.28 3.0
P2  0.72  0.33 -0.32 -0.20  0.23 0.81 0.19 2.3
P3  0.79  0.08 -0.02 -0.07  0.21 0.68 0.32 1.2
P4  0.77  0.15 -0.16  0.16 -0.13 0.68 0.32 1.3
P5  0.74  0.33 -0.30  0.03 -0.16 0.77 0.23 1.9
P6  0.77 -0.12  0.22 -0.09  0.12 0.67 0.33 1.3
P7  0.61 -0.40  0.26 -0.29  0.22 0.73 0.27 3.0
P8  0.75  0.03  0.33  0.38  0.06 0.82 0.18 1.9
P9  0.67 -0.10  0.38  0.41 -0.14 0.80 0.20 2.5
P10 0.01  0.45  0.60 -0.34 -0.23 0.73 0.27 2.9
P11 0.07  0.66  0.16  0.04 -0.56 0.78 0.22 2.1
P12 -0.10  0.70  0.25 -0.30  0.19 0.68 0.32 1.9
P13 -0.13  0.66 -0.02  0.40  0.39 0.76 0.24 2.5
P14 -0.35  0.53 -0.42  0.17  0.04 0.61 0.39 3.0
P15 -0.33  0.51  0.39  0.01  0.31 0.61 0.39 3.4

      PC1  PC2  PC3  PC4  PC5
SS loadings      4.86 2.60 1.54 0.96 0.89
Proportion var    0.32 0.17 0.10 0.06 0.06
Cumulative var    0.32 0.50 0.60 0.66 0.72
Proportion Explained 0.45 0.24 0.14 0.09 0.08
Cumulative Proportion 0.45 0.69 0.83 0.92 1.00

Mean item complexity = 2.3
Test of the hypothesis that 5 components are sufficient.

The root mean square of the residuals (RMSR) is 0.06
with the empirical chi square 109.09 with prob < 2.5e-08

Fit based upon off diagonal values = 0.96
```

Figura 6.8: Solução Fatorial com 5 fatores

7 Conclusão

Com este trabalho podemos evoluir ainda mais os nossos conhecimentos na análise de dados, conseguimos trabalhar muito bem no que toca á comunicação entre os membros do grupo, onde ocasionalmente discutimos várias hipóteses a considerar para o estudo assim como resultados obtidos.

O facto de o nosso objeto de estudo ser algo relacionado conosco estudantes permitiu uma rápida percepção dos dados, porém sentimos algumas ligeiras dificuldades na criação de hipóteses para a utilização de certos métodos, em especial na parte da regressão linear, onde discutimos em grupo várias e várias opções até conseguirmos alcançar aquela que mais se adequava para o estudo em questão.

Em suma podemos dizer que este projeto ajudou-nos muito no que toca a ampliar os nossos conhecimentos, e deu-nos uma nova visão sobre a área de análise de dados e como podemos vir a trabalhar com este tipo de análises num futuro próximo.

Referências

- [1] <https://analise-estatistica.pt/2016/11/analise-estatistica-bivariada.html> <https://analise-estatistica.pt/2018/04/analise-estatistica-univariada.html> Demos uso a vários Powerpoints facultados nas cadeiras de ATDM e IE.

Apêndice A Anexo Utilizado no Estudo

ID: _____



ESCOLA SUPERIOR DE TECNOLOGIA DO BARREIRO

INSTITUTO POLITÉCNICO DE SETÚBAL

QUESTIONÁRIO – Suporte Aulas

Trata-se de um questionário que tem por alvo os estudantes da ESTBarreiro/IPS.

Os dados recolhidos serão tratados em termos estritamente académicos pelos estudantes inscritos nas unidades curriculares: IE, PE, ATED, EA e ATMD, da ESTBarreiro/IPS.

1. Idade: _____
2. Sexo: ☐ Feminino ☐ Masculino
3. Curso que frequenta: _____
4. Ano curricular que frequenta: _____
5. Este curso foi a minha 1ª opção: ☐ Sim ☐ Não
6. Fui eu que escolhi este curso: ☐ Sim ☐ Não
7. Tempo **total (em minutos)** de deslocação entre Escola e Casa (ida e volta): _____
8. Nº horas estudo por semana: _____
9. Nº horas diários que se dedica às redes sociais (Facebook; WhatsApp, Instagram, etc...): _____
10. Nº horas diários que se dedica a ver TV, Netflix, etc...: _____
11. Nº **médio** de horas que dorme por dia de 2ªf a 6ªf : _____
12. Tem conhecimento que a ESTBarreiro/IPS oferece um Programa de Mentoria aos seus estudantes: ☐ Sim ☐ Não

13. Escala de Burnout de Maslach para Estudantes:

As afirmações seguintes são referentes aos sentimentos/emoções de estudantes em contexto escolar. Leia cuidadosamente cada afirmação e decida sobre a frequência com que se sente da forma descrita e de acordo com o quadro seguinte:

	Nunca	Quase nunca	Algumas vezes	Regularmente	Muitas vezes	Quase sempre	Sempre
	0	1	2	3	4	5	6
1. Os meus estudos deixam-me emocionalmente exausto(a)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2. Sinto-me de "rastos" no final de um dia na escola.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3. Sinto-me cansado(a) quando me levanto de manhã e penso que tenho de enfrentar mais um dia na escola.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4. Estudar ou assistir a uma aula deixam-me tenso(a).	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5. Os meus estudos deixam-me completamente esgotado(a).	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6. Tenho vindo a desinteressar-me pelos meus estudos desde que ingressei na escola.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
7. Sinto-me pouco entusiasmado(a) com os meus estudos.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
8. Sinto-me cada vez mais cínico(a) relativamente à utilidade potencial dos meus estudos.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
9. Tenho dúvidas sobre o significado dos meus estudos.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
10. Consigo resolver, de forma eficaz, os problemas que resultam dos meus estudos.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
11. Acredito que participo, de forma positiva, nas aulas a que assisto.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
12. Sinto que sou um(a) bom aluno(a).	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
13. Sinto-me estimulado(a) quando alcanço os meus objetivos escolares.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
14. Tenho aprendido muitas matérias interessantes durante o meu curso.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
15. Durante a aula, sinto que consigo acompanhar as matérias de forma eficaz.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

OBRIGADO PELA SUA COLABORAÇÃO!

Professora Anabela Marques e Professora Ana Meireles