Instituto Politécnico de Setúbal

Escola Superior de Tecnologia do Barreiro

Projeto de Análise e Tratamento de Dados Multivariados

Licenciatura em Bioinformática

Estudo estatístico sobre os estudantes da ESTBarreiro

Janeiro de 2023

Grupo 04
Pedro Pacheco (202100957)
Hélder Marques (202100959)

Índice

1 Introdução

Neste projeto vamos realizar uma análise do comportamento dos estudantes da Escola Superior de Tecnologia do Barreiro. Os temas que vamos abordar para a seguinte análise destes dados são: análise descritiva univariada, análise descritiva bivariada, estudo inferencial, regressão linear múltipla e por fim análise fatorial e de fiabilidade. Com este estudo prentendemos saber se os hábitos dos estudantes são divergentes ou semelhantes visto que frequentam a mesma universidade.

2 Metodologia

2.1 Conceitos utilizados

Desenvolvemos este projeto com recurso à análise estatística. Análise estatística permite recolher, explorar, organizar e apresentar grandes quantidades de dados. Desenvolvemos estas atividades através de testes de hipóteses, análise exploratória e modelação de dados. Utilizámos uma população alvo que é um conjunto de individuos sobre os quais é efetuada uma avaliação e/ou uma análise estatística.

2.2 População alvo e variáveis

A nosssa população de estudo é constituida por 136 (cento e trinta e seis) alunos. Para avaliar a nossa população alvo utilizámos as seguintes variáveis: idade, sexo, curso, ano curricular, se o curso foi primeira escolha, se foi o aluno que escolheu esse curso, tempo total de deslocação, número de horas semanais de estudo, número de horas diárias nas redes sociais, número médio de horas diárias de horas de sono nos dias úteis, número de horas dedicadas à TV, se o aluno tem conhecimento do programa de mentoria e uma escalha de Burnout de Maslach com 15 (quinze) perguntas.

2.3 Análise estatística

Para esta população vamos utlilizar análise univariada e bivariada. A análise univariada analisa a distribuição e dispersão dos dados, inclui a análise dos valores da média, moda, mediana e quartis. A análise bivariada permite observar como duas variáveis se comportam na persença uma da outra, ou seja a sua associação, correlação e dependência.

No nosso projeto utilizaremos todas as medições estudadas da estatistica univariada, ou seja, média, mediana, moda, quartis e máximos/mínimos. Em relação à análise bivariada vamos utilizar a correlação de Spearman e de Pearson.

Começaremos por fazer uma tabela de contigência com a variável Sexo e a variável p5 (pergunta 5 da escala de Burnout: "Os meus estudos deixam-me completamente esgotado(a).")

Iremos utilizar a correlação de Spearman pois a primeira questão que avaliámos continha duas variáveis, uma delas quantitativa e outra qualitativa ordinal. Na segunda questão utilizaremos a correlação de Pearson pois continha duas variáveis quantitativas.

Vamos realizar um estudo inferencial, com o nível de significância de $\boldsymbol{\alpha}$=0.05, para três questões:

1 - "Será que os alunos que estudam mais semanalmente estão em anos curriculares mais elevados?".

2 - "Será que existe diferenças entre o número médio diário de horas de sono nos dias úteis em cada um dos diferentes cursos".

3 - "Será que o número de horas de estudo é influênciado pelo facto de os alunos terem escolhido o curso em regime de primeira opção?".

Iremos realizar uma regressão linear múltipla para duas questões com o nível de significância de $\boldsymbol{\alpha}$=0.05:

1 - "Será que o número de horas de sono diário é influenciado pela idade do estudante e pelo tempo de deslocação?".

2 - "Será que o número de horas de estudo é influenciado pelo Sexo e pelas horas de lazer?".

Por último, vamos realizar uma Análise Fatorial e uma Análise de Fiabilidade sobre as questões da escalha de Burnout de Maslach.

3 Caraterização da amostra

3.1 Análise Descritiva Univariada e Bivariada

3.2 Análise descritiva Univariada

3.2.1 Análise da variável Idade

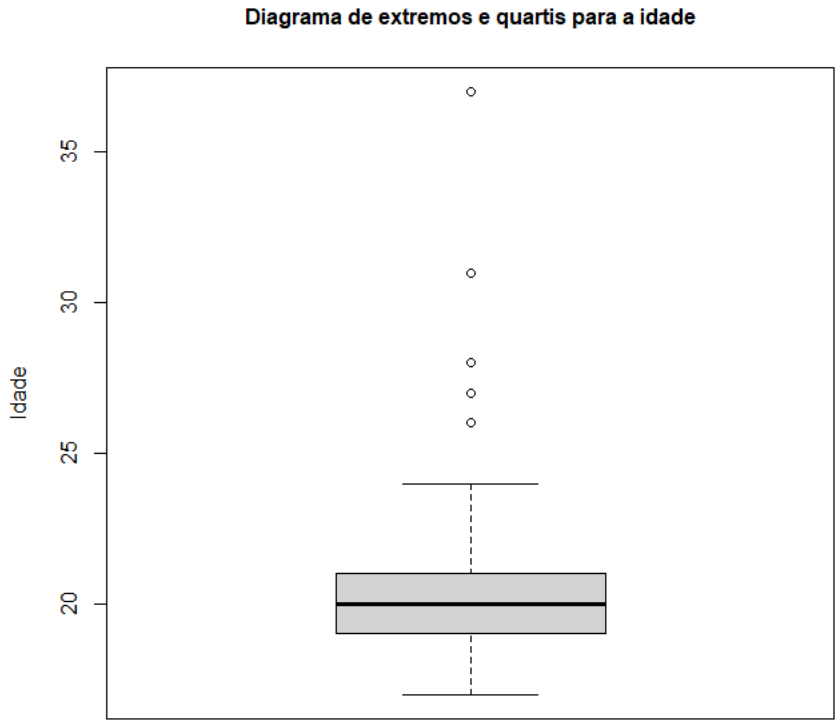

Figura 1: Diagrama de extremos e quartis para a variável idade

Como podemos observar neste diagrama de extremos e quartis, a variavél idade tem um minimo de 17 (dezassete) e um máximo de 37 (trinta e sete), o primeiro quartil de 19 (dezanove), a mediana é 20 (vinte), o terceiro quartil de 21 (vinte e um), uma média de 20,5 (vinte vírgula cinco) e por fim um desvio padrão de 2.79 (dois vírgula setenta e nove).

3.2.2 Análise da variável Sexo

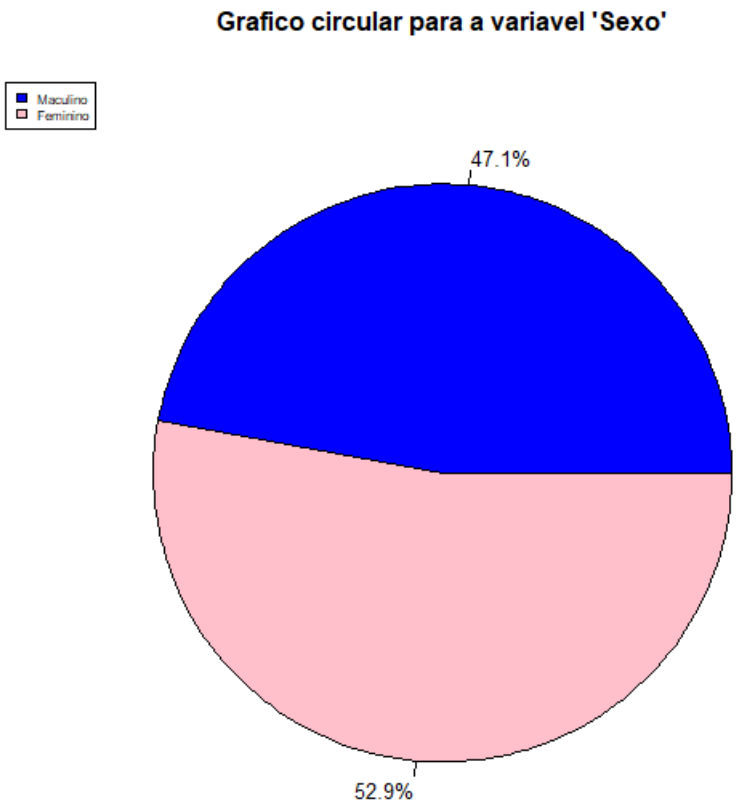

Figura 2: Gráfico circular para a variável sexo

Como podemos observar neste gráfico circular 47,1%(quarenta e sete virgula um por cento) serão alunos do sexo masculino e 52,9%(cinquenta e dois virula nove por cento) serão alunos do sexo feminino.

3.2.3 Análise da variável Curso

Figura 3: Gráfico circular para a variável curso

Como podemos observar neste gráfico circular 27,2%(vinte e sete virgula dois por cento) serão alunos do curso de Bioinformática , 56,6%(cinquenta e seis virula seis por cento) serão alunos do curso de Biotecnologia, e por fim 16,2%(dezasseis virgula dois por cento) dos alunos pertencem ao CTeSP TLQB.

3.2.4 Análise da variável ano curricular

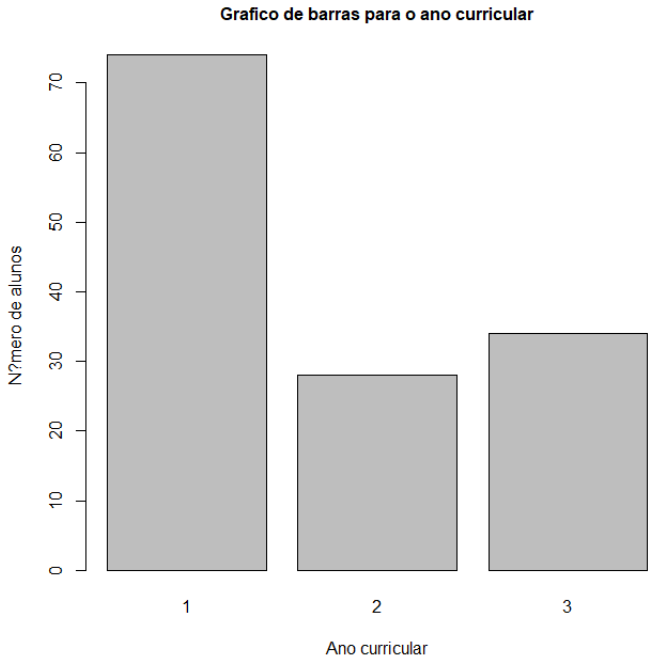

Figura 4: Gráfico de barras para a variavel ano curricular

Olhando para o grafico podemos observar que dos 136 (cento e trinta e seis) alunos que fizeram o questionário cerca de 70 (setenta) são alunos do primeiro ano, cerca de 30 (trinta) alunos pertencem ao segundo ano e a volta de 36 (trinta e seis) pertencem ao terceiro ano. Conseguimos concluir que a maioria dos alunos que respondeu ao questionário pertence ao primeiro ano e a minoria pertence ao segundo ano.

3.2.5 Análise da variável primeira opção

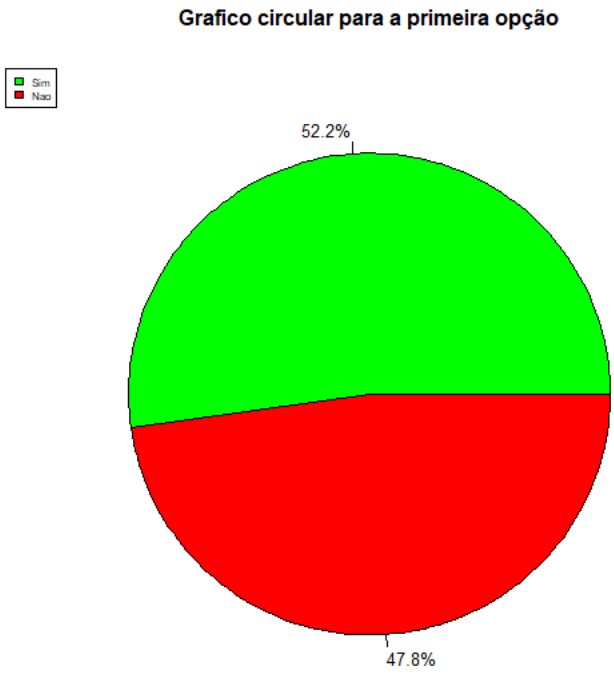

Figura 5: Gráfico circular para a variável primeira opção

Ao olhar paa o grafico circular conseguimos rapidamente analisar que 52,2% (cinquenta e dois virgula dois por cento) dos alunos escolheram o seu curso como primeira opção e 47,8%(quarenta e sete virgula oito por cento) dos alunos não escolheu o seu curso como primeira opção.

3.2.6 Análise da variável eu escolhi este curso

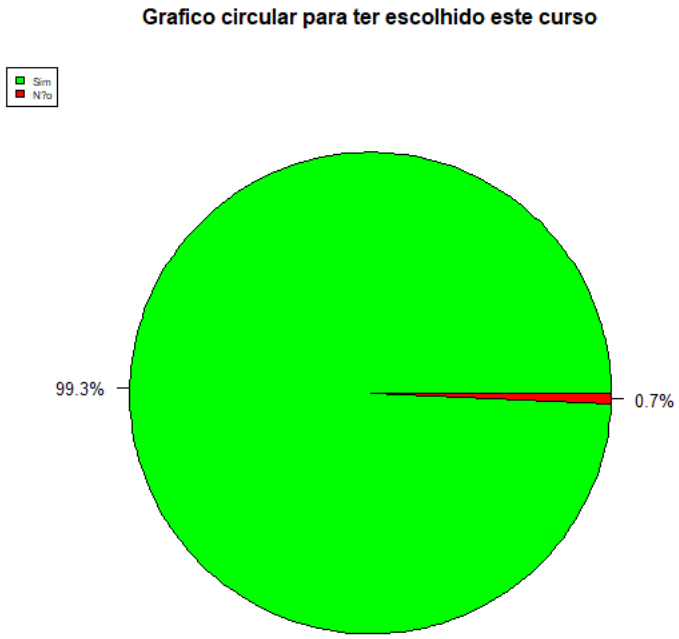

Figura 6: Gráfico circular para a variável curso escolhido

Com a análise deste gráfico circular observamos que 99,3%(noventa e nove virgula três por cento) escolheram o seu curso e 0,7%(zero virgula sete por cento) não escolheram o seu curso.

3.2.7 Análise da variável tempo de deslocação

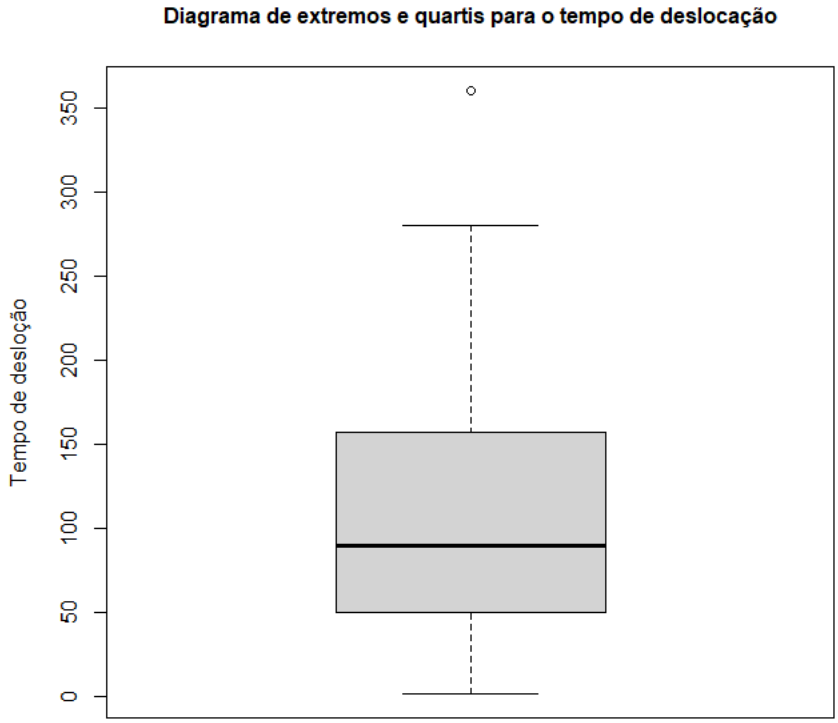

Figura 7: Diagrama de extremos e quartis para a variável tempo de deslocação

Como podemos observar neste diagrama de extremos e quartis, a variavél tempo de deslocação tem um minimo de 2 (dois) e um máximo de 360 (trezentos e sessenta), o primeiro quartil de 50 (cinquenta), a mediana é 90 (noventa), o terceiro quartil de 154 (cento e cinquenta e quatro), uma média de 108 (cento e oito) e por fim um desvio padrão de 71,1 (setenta e um vírgula um).

3.2.8 Análise da variável horas de estudo por semana

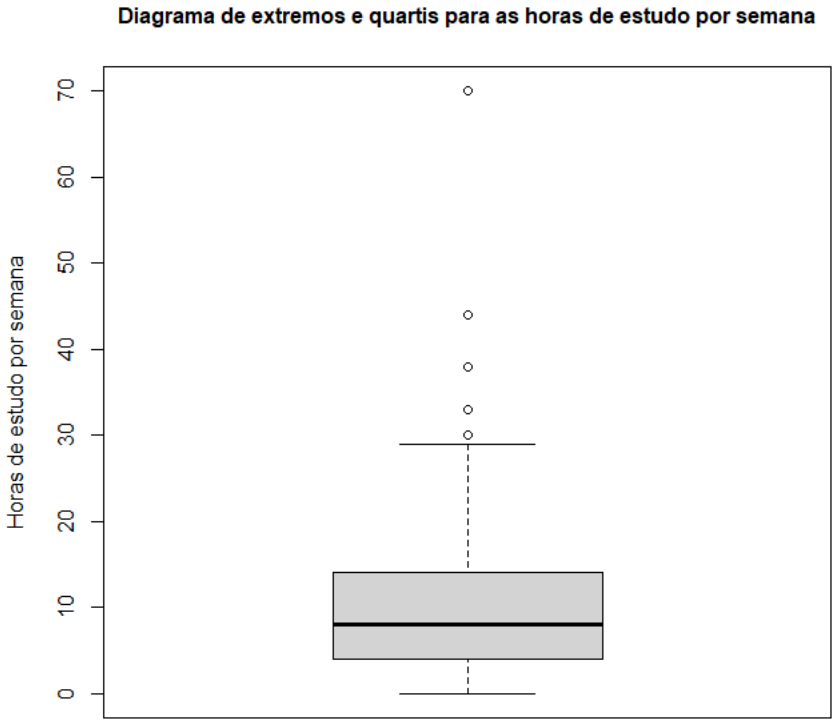

Figura 8: Diagrama de extremos e quartis para a variável horas de estudo por semana

Como podemos observar neste diagrama de extremos e quartis, a variavél horas de estudo tem um minimo de 0 (zero) e um máximo de 70 (trezentos e sessenta), o primeiro quartil de 4 (quatro), a mediana é 8 (oito), o terceiro quartil de 14 (quatorze), uma média de 10,1 (dez vírgula um) e por fim um desvio padrão de 9,30 (nove vírgula trinta).

3.2.9 Análise da variável horas de redes

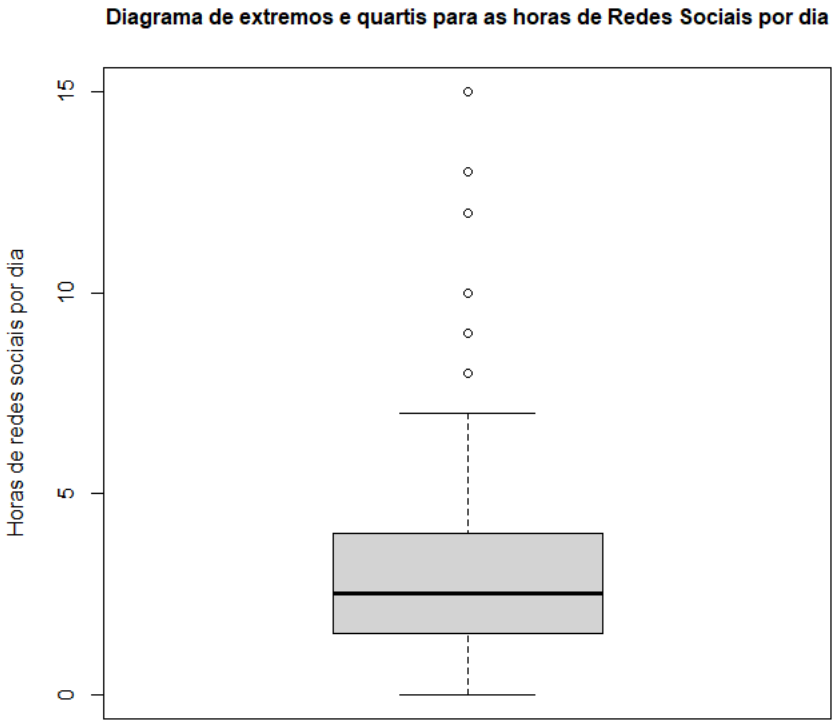

Figura 9: Diagrama de extremos e quartis para a variável horas de redes

Como podemos observar neste diagrama de extremos e quartis, a variavél horas de redes tem um minimo de 0 (zero) e um máximo de 15 (quinze), o primeiro quartil de 1,5 (um vírgula cinco), a mediana é 2,5 (dois vírgula cinco), o terceiro quartil de 4 (quatro), uma média de 3,24 (tres vírgula vinte e quatro) e por fim um desvio padrão de 2,76 (dois vírgula setenta e seis).

3.2.10 Análise da variável horas de TV

Como podemos observar neste diagrama de extremos e quartis, a variavél horas de redes tem um minimo de 0 (zero) e um máximo de 8 (oito), o primeiro quartil de 1 (um), a mediana é 2 (dois), o terceiro quartil de 3 (tres), uma média de 1,99 (um vírgula noventa e nove) e por fim um desvio padrão de 1,51 (um vírgula cinquenta e um).

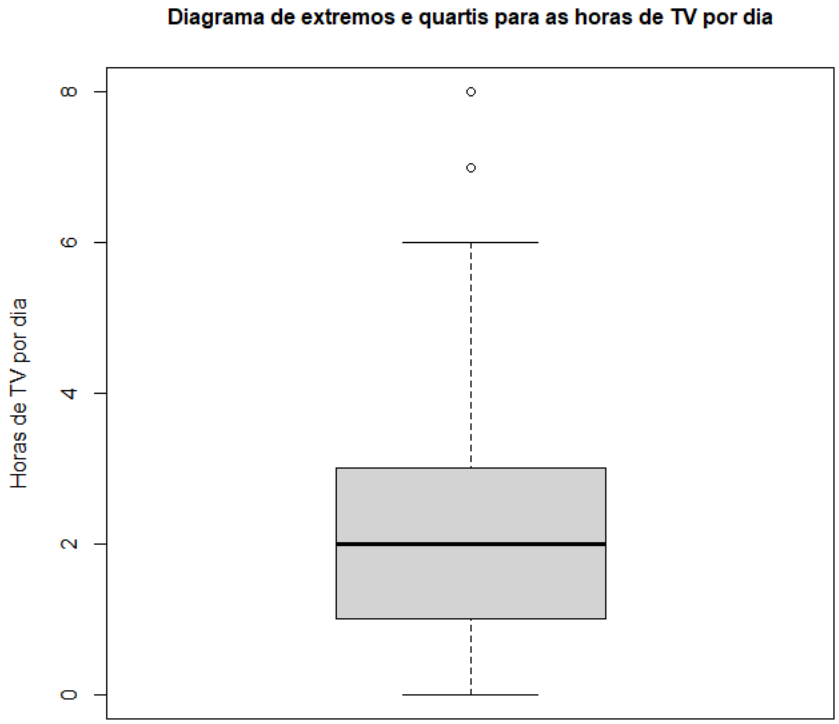

Figura 10: Diagrama de extremos e quartis para a variável horas de TV

3.2.11 Análise da variável horas de Sono

Como podemos observar neste diagrama de extremos e quartis, a variavél horas de redes tem um minimo de 3 (tres) e um máximo de 9 (nove), o primeiro quartil de 6 (seis), a mediana é 7 (sete), o terceiro quartil de 8 (oito), uma média de 6,95 (seis vírgula noventa e cinco) e por fim um desvio padrão de 1 (um).

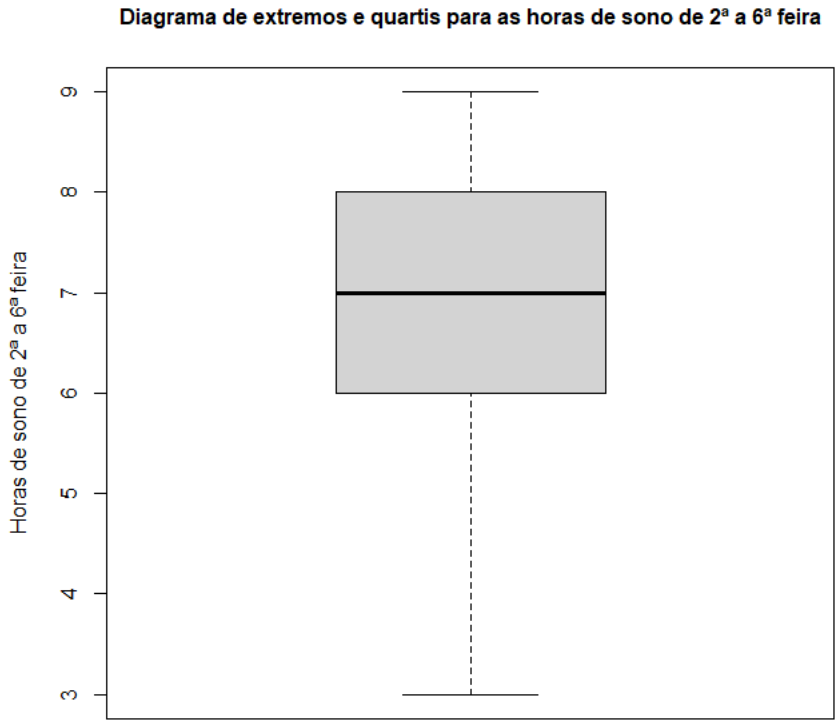

Figura 11: Diagrama de extremos e quartis para a variável horas de Sono

3.2.12 Análise da variável programa de mentoria

Ao olhar para o gráfico circular conseguimos rápidamente analisar que 56,6% (cinquenta e seis vírgula seis) dos alunos têm conhecimento do programa de mentoria e 43,4%(quarenta e três vírgula quatro) dos alunos não tinha conhecimento do programa de mentoria.

Figura 12: Gráfico circular para a análise da variável PMentoria

3.3 Análise Descritiva Bivariada

3.3.1 Tabela de contingencia

```
          p5
sexo       Nunca Quase nunca Algumas vezes Regularmente Muitas vezes Quase sempre Sempre
  Masculino    11          19            11           7            6            7      0
  Feminino     10          27            10           5            6            3      0
```

Figura 13: Tabéla de contingencia da variavel sexo pela pergunta 5(cinco) do teste de burnout

Na tabéla de contingencia a cima observada conseguimos analisar que tanto os alunos do sexo feminino com do sexo masculino dizem que "Nunca"ou "Quase nunca"ou "algumas vezes"ficam esgotados completamente esgotados com os seus estudos sendo que a maioria dos alunos de ambos os sexos respondeu "Quase Nunca"e a minoria respondeu "Quase Sempre".

3.3.2 Coéficiente de Spearman - Horas de Lazer por P5

```
> cor(EscalaBurnoutGrupo4$horaslazer, EscalaBurnoutGrupo4$BurnoutP5) #Nao ha correlacao significativa -0.04
[1] -0.04174405
> 
```

Figura 14: Coéficiente de Spearman

Como o valor é negativo as duas variáveis tem sentidos opostos afastando se uma da outra sendo inversamente proporcionais.

3.3.3 Análise: Será que os estudantes ficam menos esgotados com mais horas de lazer?

A nossa variável "horaslazer"é constituida pela soma das variáveis "horasTV"e "horasRedes", a sua amostra total são os 136 (cento e trinta e seis) alunos, apresenta um mínimo de 0 e um máximo de 18, a sua mediana é de 4,5 e a sua média de 5,24, apresenta como primeiro quartil 3,0 e como terceiro quartil 7,0.

Utilizámos a correlação de Spearman porque vamos avaliar uma variável quantitativa e uma qualitativa ordinal, verificámos então que a sua correlação (-0.0417) é negativa, ou seja, as duas variáveis têm sentidos opostos e muito fraca ou praticamente não existente.

3.3.4 Análise: Será que os estudantes mais velhos tendem a estudar mais?

Utilizámos a correlação de Pearson porque a nossa questao apresenta duas variáveis quantitativas, então verificámos que a sua correlação (-0.1466) é negativa, ou seja, as duas variáveis têm sentidos opostos e muito fraca pelo que as duas variáveis apresentam uma correlação muito fraca.

3.4 Estudo Inferencial

3.4.1 Será que existe diferenças entre o numero medio de horas de sono por semana em cada um dos três diferentes cursos?

A avaliação do número médio de horas de sono por semana (variável dependente definida numa escala quantitativa) para cada um dos cursos (L. Bioinformática, L.Biotecnologia e CTeSP TLQB) foi avaliada através do teste não paramétrico de Kruskal-Wallis.

A utilização deste teste é justificada pelo facto de, pela aplicação do teste de normalidade de Kolmogorov-Smirnov, se ter concluído que o número médio de horas semanais pelo curso L. Bioinformática não verificar o pressuposto de normalidade (p=3.177e-05), e pelo curso L.Biotecnologia (p=1.57e-06) não verificarem o pressuposto de normalidade. As diferenças entre o número medio de horas de sono por semana em cada um dos cursos, detetadas por aplicação do teste de Kruskal-Wallis, foram analisadas por recurso ao teste de Dunn que realiza comparações múltiplas de médias de ordens. Todas as análises estatísticas foram realizadas com o auxílio do software Rstudio (R version 4.2.2 (2023 01 20)). Os resultados são apresentados como média +- desvio-padrão da média e considera-se que as diferenças são estatisticamente significativas para valores de p-value do teste igual a 0.05.

A aplicação do teste não paramétrico de Kruskal-Wallis, mostrou que, para um nível de significância de 5%, a origem dos indivíduos tem um efeito estatisticamente significativo e de pequena dimensão ($\chi^2$2=6.06, p = 0.048, (ηH2= 0.415)) sobre as horas de estudo por curso. Os maiores níveis são registados para os indivíduos das populações L. Bioinformática (6,8 +/- 0,85) e L.Biotecnologia (7,12 +/- 0,99), seguido da população CTeSP TLQB (6,64 +/- 1,20). A comparação entre pares realizada pelo teste post-hoc de Dunn mostrou que não existe diferenças significativas entre as populações.

3.4.2 Será que o número de horas de estudo é influenciado pelo facto de o curso ter sido escolhido em regime de primeira opção?

```
# A tibble: 3 x 10
  Curso             variable      n   min   max    q1 median    q3  mean    sd
  <chr>             <fct>     <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
1 CTeSP TLQB        HorasSono    22     4     9   6      6.5  7.75  6.64 1.20
2 L. Bioinformática HorasSono    37     5     8   6      7    7.5   6.80 0.854
3 L. Biotecnologia  HorasSono    77     3     9   6.5    7    8     7.12 0.992
> |
```

Figura 15: Estatisticas descritivas por "Grupo"@

A avaliação do numero de horas de estudo (variável dependente, quantitativa) ser influenciado pelo facto do curso ter sido escolhido na primeira opção (Sim, Não) foi avaliado através do teste não paramétrico de Wilcoxon-Mann-Whitney. A utilização deste teste é justificada pelo facto de, pela aplicação do teste de normalidade de Kolmogorov-Smirnov, se ter concluído que o número de horas de estudo influenciado pelo facto do curso ter sido escolhido na primeira opção(pS=3,11e-10) e não ter sido escolhido na primeira opção (pN=0,016) não verificarem o pressuposto de normalidade. Todas as análises estatísticas foram realizadas com o auxílio do software Rstudio (Rversion 4.2.2 (2023 01 22)). Os resultados são apresentados como média desvio-padrão da média e considera-se que as diferenças são estatisticamente significativas para valores de p-value do teste iguais a 0.05.

Este estudo permitiu verificar que o número de horas de estudo pela escolha "Sim"da primeira opção é em média 9,99 +/- 11,2 horas, com mediana de 6 horas, enquanto que o número de horas de estudo pela escolha "Não"é em média 10,3 +/- 6,75 horas, com mediana de 10 horas. O estudo inferencial, realizado por recurso ao teste não paramétrico de Wilcoxon Mann-Whitney permitiu concluir que o número de horas de estudo é influenciado pelo facto do curso ter sido escolhido em regime primeira opção. (W = 1946,5 ; Z = -1,59, p = 0,116, r = 0,135), sendo a dimensão do efeito fraca.

3.4.3 Será que o conhecimento do plano de mentoria é independente do sexo?

Para avaliar se o conhecimento do plano de mentoria é independente do sexo recorreu-se ao teste de independência do Qui-Quadrado, seguido de umaanálise de resíduos. As análises estatísticas foram efetuadas com o auxílio do software Rstudio (R version 4.2.2 (2023 01 25)). Considera-se um resultado estatisticamente significativo para valores de p-value do teste inferiores ou iguais a 0.05.

O estudo realizado demonstrou que o conhecimento do plano de mentoria foi mais elevado entre os sujeitos do sexo feminino (n=40), do que no sexo masculino (n=37) e o não conhecimento foi maior entre os alunos do sexo feminino (n=32), do que no sexo masculino (n=27), sendo então o total da dimensão da amostra 136(cento e trinta e seis). Todos os pressupostos da aplicação do teste do

qui-quadrado são verificados tendo então todas as classes frequências esperadas superior a 1 e 80(%) a cima de 5. A análise inferencial, realizada através do teste de independência do QuiQuadrado, mostrou que o o conhecimento do plano de mentoria é independente do sexo (($\chi^2(1)2 = 0.0084$;p=0.927 > 0.05).

3.5 Regressão Linear Multipla

3.5.1 Será que o tempo de deslocação é afetado pela idade e pelas horas de sono?

3.5.2 Verificação da utilização da Regressão Linear

Como estamos a considerar apenas duas variáveis independentes (p = 2) é possível visualizar a nuvem de n pontos em R3, através de um gráfico a 3 dimensões.

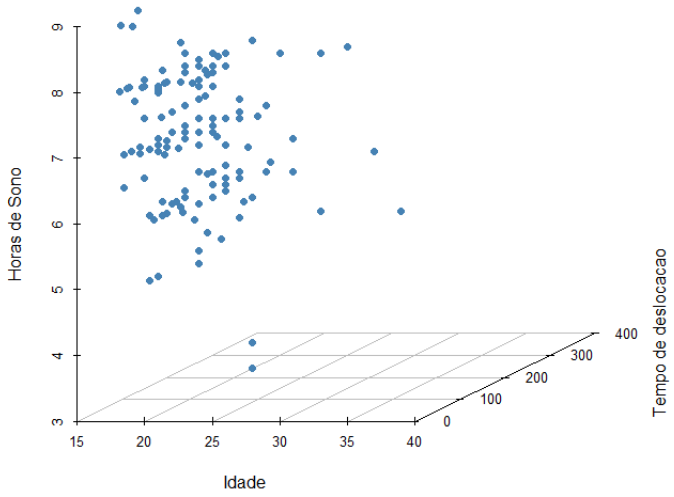

Figura 16: nuvem de n pontos em R3

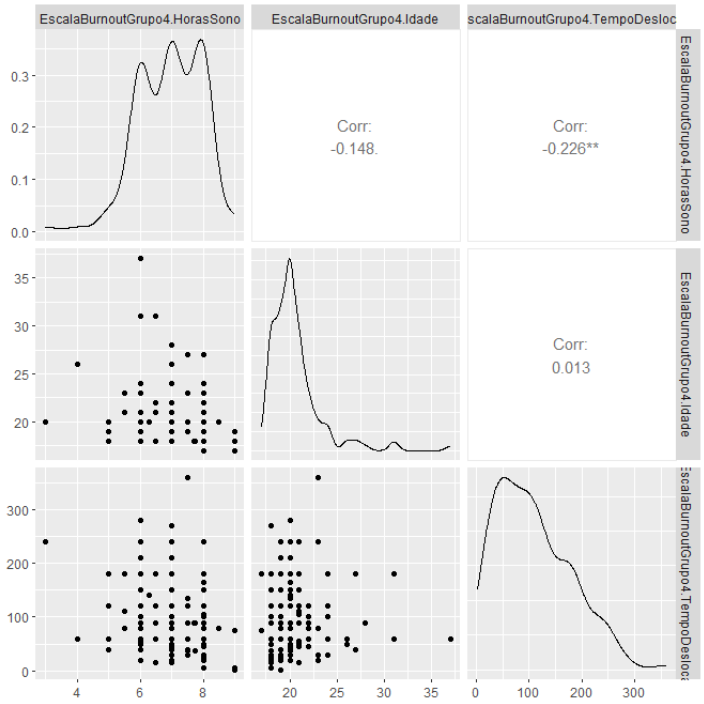

Figura 17: nuvem de n pontos em R3

A equação Y= $\beta 0$ + $\beta 1$x1 + $\beta 2$x2

Em que o nosso Y corresponde a variável tempo de deslocação, o nosso parametro $\beta 0$ = 8.366211, o parametro relacionado com a idade $\beta 1$ = -0.052175, o x1 representa o valor da variável idade, o parametro relacionado com as horas de sono $\beta 2$ = -0.003167 e o x2 representa o valor da variável horas de sono.

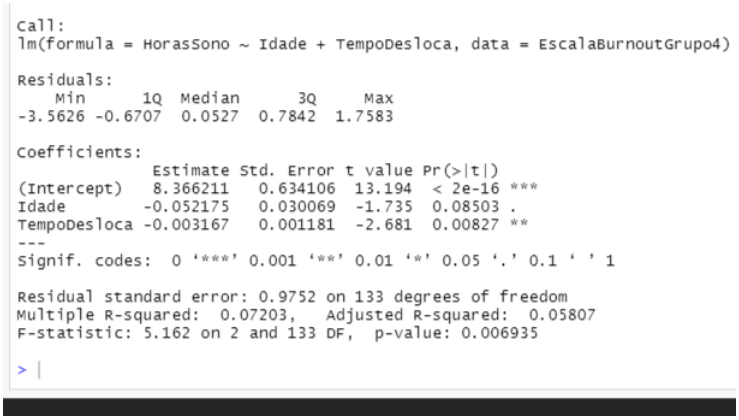

Figura 18: Construção do modelo de regressão linear múltiplo

3.5.3 Verificação de pressupostos do modelo

Ao tentar verificar a Linearidade no gráfico "Residuals vs Fitted"da figura 19(dezanove) conseguimos observar que a linha horizontal encarnada está a cima da linha a tracejado em determinadas partes logo podemos concluir que existe linearidade.

Em relação à Homocedasticidade conseguimos observar no gráfico "Scale-Location"da figura 19(dezanove) que a linha vermelha apresenta uma forma horizontal que é resultante do facto de existir homocedasticidade.

Ainda na figura 19(dezanove) conseguimos observar a existencia de outliers no gráfico "Residuals vs Leverage"pois existem dois pontos inferiores a -3(menos três) que quer dizer que são outliers e esses pontos são o ponto 34(trinta e quatro) e o ponto 101/cento e um).

Podemos também observar no gráfico "Residuals vs Leverage"da figura 19(dezanove) a existencia de três pontos influentes e estes pontos influenciam a estimação do modelo.

3.5.4 Conclusão de resultados

Será que ambas as variáveis independentes usadas no ajustamento do modelo têm efeito significativo sobre a variável tempo de deslocação?

Para existir efeito significativo por parte das variaveis independentes o valor dos paramentros β têm de ser diferentes de 0 (zero) para existir significância estatística.

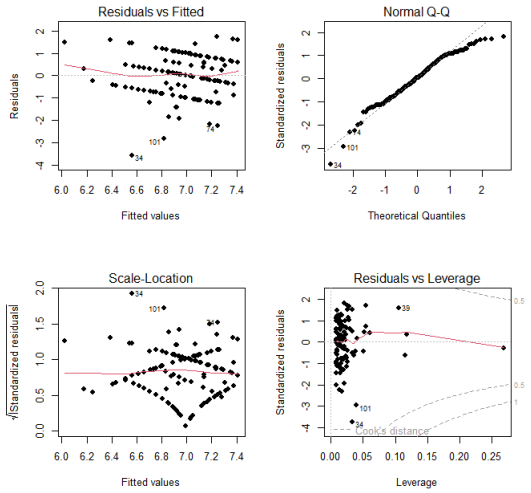

Figura 19: Construção do modelo de regressão linear múltiplo

	Estimate	Standardized	Std. Error	t value	Pr(>|t|)
(Intercept)	8.366	NA	0.6341	13.19	6.23e-26
Idade	-0.05217	-0.1449	0.03007	-1.735	0.08503
TempoDesloca	-0.003167	-0.224	0.001181	-2.681	0.00827

Table: Fitting linear model: HorasSono ~ Idade + TempoDesloca

Figura 20: Coeficientes padronizados

Com a análise da figura podemos observar que existe efeito significativo para a variável para o TempoDesloca e para a Idade não. Podemos concluir com isto que a variável TempoDesloca tem uma maior influência na mariavel dependente do que a variável Idade.

3.5.5 Será que as horas de estudo são influenciadas pelas horas de lazer e pelo sexo?

3.5.6 Verificação da utilização da Regressão Linear

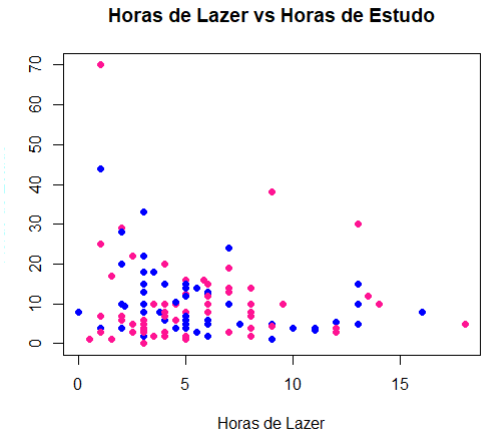

Figura 21: Diagrama de dispersão

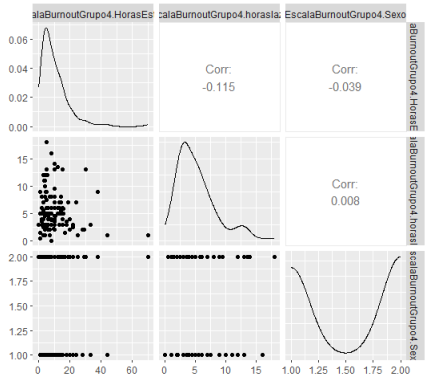

Figura 22: Scatterplot matrix

A equação Y= $\beta 0 + \beta 1x1 + \beta 2x2$ Em que o nosso Y corresponde a variável horas de estudo, o nosso parametro $\beta 0 = 12.8775$, o parametro relacionado com o sexo $\beta 1$ = -0.7094, o x1 representa o valor da variável sexo, o parametro relacionado com as horas de lazer $\beta 2$ = -0.3142 e o x2 representa o valor da variável horas de lazer.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.8775     2.8441   4.528 1.31e-05 ***
Sexo         -0.7094     1.5974  -0.444    0.658
horaslazer   -0.3142     0.2363  -1.330    0.186
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.298 on 133 degrees of freedom
Multiple R-squared:  0.01463,   Adjusted R-squared:  -0.0001889
F-statistic: 0.9873 on 2 and 133 DF,  p-value: 0.3753
```

Figura 23: Construcao do modelo de regressão linear múltiplo

Com a análise da figura 23 podemos concluir que nenhuma das variáveis indenpendentes têm influencia na variável dependente sendo que ambas têm um p-value inferior a 0.05(zero virgula zero cinco).Logo procedemos para a criação de uma variável "dummy"que permite construir um único modelo que comporta as duas regressões de uma vez só.

Variável "dummy": fem = 1 se Feminino; fem = 0 caso contrario Modelo: yi = b0 + b1 * x1 + b2 * fem , onde variavel dependente = yi = "RendEsc"variável independente = xi = "CompLei"= fem = 1 se feminino; 0 se masculino.

```
Call:
lm(formula = HorasEstudo ~ horaslazer + Fem, data = EscalaBurnoutGrupo4)

Residuals:
    Min      1Q  Median      3Q     Max
-10.516  -6.259  -2.242   3.839  58.856

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.1680     1.6929   7.188 4.27e-11 ***
horaslazer   -0.3142     0.2363  -1.330    0.186
Fem          -0.7094     1.5974  -0.444    0.658
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.298 on 133 degrees of freedom
Multiple R-squared:  0.01463,   Adjusted R-squared:  -0.0001889
F-statistic: 0.9873 on 2 and 133 DF,  p-value: 0.3753
```

Figura 24: Construcao do modelo de regressão linear múltiplo

Ao tentar verificar a Linearidade no gráfico "Residuals vs Fitted"da figura 25(vinte e cinco) conseguimos observar que a linha horizontal vermelha está a cima da linha a tracejado em determinadas partes logo podemos concluir que existe linearidade.

Em relação à Homocedasticidade conseguimos observar no gráfico "Scale-Location"da figura 25(vinte e cinco) que a linha vermelha apresenta uma forma

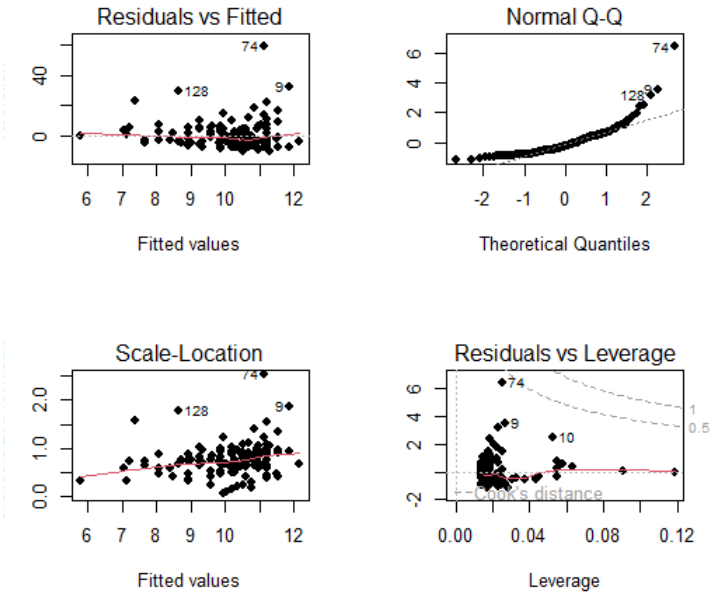

Figura 25: Construcao do modelo de regressão linear múltiplo

aproximadamente horizontal sendo que está ligeiramente incluinada mas mesmo assim concluimos que existe homocedasticidade.

Ainda na figura 25(vinte e cinco) conseguimos observar a existencia de outliers no gráfico "Residuals vs Leverage"pois existem 3 pontos superiores a 3(três), logo podemos concluit a existencia de outliers superiores.

Podemos também observar no gráfico "Residuals vs Leverage"da figura 25(vinte e cinco) a existencia de dois pontos influentes que influenciam a estimação do modelo.

3.5.7 Conclusão de resultados

Será que ambas as variáveis independentes usadas no ajustamento do modelo têm efeito significativo sobre a variável tempo de deslocação?

Para existir efeito significativo por parte das variaveis independentes o valor dos paramentros β têm de ser diferentes de 0 (zero) para existir significância estatística.

Com a análise da figura podemos observar que existe efeito significativo para a variável para o Fem e para a horas de lazer não. Podemos concluir com isto que a variável Fem tem uma maior influência na variável dependente do que a variável horas de lazer.

3.6 Análise Fatorial

Para efetuar uma Análise Fatorial necessitamos de verificar a adequabilidade dos dados e vê-se com 4 (quatro) pressupostos:

- Número de observações e rácio - Teste de esfericidade de Barlett - Análise da Matriz de correlação - Medida de adequabilidade de Kaiser-Mayer-Olkin

A dimensão da amostra é suficiente sendo 136 por 15 variáveis dando um rácio de 9, superior aos 5 exigidos.

O teste de esfericidade de Barlett testa a hipótese de a matriz de correlações ser a matriz identidade, e como o p=3.53e-143<0.05=α podemos confirmar que há correlações, então a nossa matriz de correlações não é a matriz identidade.

Após a análise da matriz de correlação podemos verificar que apenas 189 variáveis possuem correlações entre si. Após verificarmos a medida de adequabilidade de Kaiser-Mayer-Olkin podemos afirmar que os nossos dados apresentam 0.84, sendo então uma boa adequabilidade.

A qualidade da solução fatorial pode ser avaliada pela% da variabilidade total explicada pelos fatores e pelos valores da comunalidade. Como verifificamos a adequabilidade dos nossos dados podemos avançar oara a solução inicial, na solução inicial verificámos que pelo método do Screenplot (menos fiável) apresenta 3 ou 4 fatores dependendo da interpretação, então vamos avaliar pelo método mais fiável, o critério de Kaiser que nos apresenta 4 fatores.

A solução com 4 fatores explica 68% que embora não esteja a cima do valor de referência está próximo (70%), os valores das comunalidades variam entre 0.487 e 0.808, três variáveis representa valores a baixo de 0.6. Das 15 variáveis consideradas apenas 3 se afastaram do valor de referência (0.6), Burnout11 = 0.487, Burnout14 = 0.556, Burnout15 = 0.555, as restantes 12 variáveis apresentam valores entre 0.616 e 0.808. Tendo em conta a informação conjunta, destes dois indicadores pode-se considerar que é adequada. No entanto poderia-se considerar uma análise de 5 fatores uma vez que esta permitia aumentar a variabilidade total explicada pelos fatores e os valores das comunalidades.

Realizámos a rotação dos fatores para facilitar a interpretação, recorrendo à função varimax.

Podemos concluir com recurso à arvore de fatores que as variáveis BurnoutP1, BurnoutP2, BurnoutP3, BurnoutP4, BurnoutP5 e BurnoutP15 estão associados ao fator 1, as variáveis BurnoutP6, BurnoutP7, BurnoutP8, BurnoutP9 estão associados ao fator 2, as variáveis BurnoutP11, BurnoutP12, BurnoutP13, BurnoutP14 estão associados ao fator 3 e a variável BurnoutP10 associada ao fator 4.

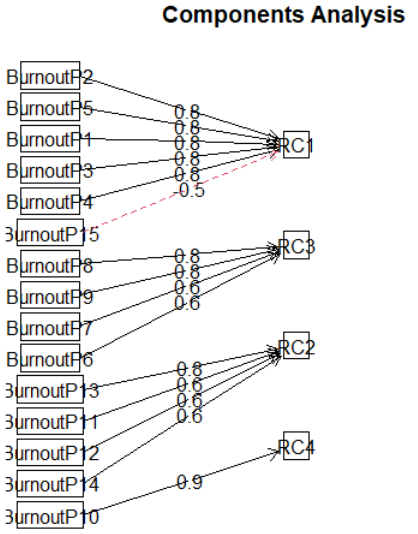

Figura 26: Visualização gráfica do modelo e das suas cargas fatoriais