



Instituto Politécnico de Setúbal

Escola Superior de Tecnologia do Barreiro

Projeto de Análise e Tratamento de Dados Multivariados

Licenciatura em Bioinformática

Título do projeto

Dezembro/2021

Grupo 02

Gilherme Sá (202000201)



João Gomes (202000202)



Tiago Timóteo (202000029)

Índice

1	Introdução	1
2	Caracterização da amostra	1
3	Análise estatística univariada	2
3.1	Idade	2
3.2	Sexo	4
3.3	Curso	4
3.4	Ano Curricular	5
3.5	Qual a primeira Opção	6
3.6	Escolhi este curso	6
3.7	Tempo Deslocação	7
3.8	Horas de Estudo	8
3.9	Horas de Redes	9
3.10	Horas TV	10
3.11	Horas de Sono	10
3.12	Pmentor	11
4	Análise estatística Bivariada	12
4.1	Tabela de contingencia	12
4.2	Coefficientes de associação/correlação	12
4.2.1	Horas de Sono VS Horas de Redes (Pearson)	12
4.2.2	Varivel Sexo VS Escolhi este curso	13
5	Estudo Inferencial	14
5.1	Primeira Pergunta	14
5.2	Segunda Pergunta	15
5.3	Terceira Pergunta	16
6	Métodos de Análise de Dados Multivariados	17
6.1	Regressão Linear Múltipla	17
6.2	Verificação dos pressuposto através de testes	20
6.3	Modelo da Regressão Linear	22
7	Análise Fatorial	23
7.1	Critério de Kaiser	27
7.2	Rotação de Fatores	29
8	Conclusão	32
9	Webgrafia	33

1 Introdução

O trabalho realizado tem como objetivo principal explorar as funções do software , tendo por base a análise estatística univariada e bivariada do Questionário Burnout fornecido para o estudo. Tem-se como objetivo realizar uma análise descritiva univariada, bivariada e também realizar o estudo inferencial da amostra aplicando os vários métodos abordados nas aulas e disponíveis pelo software .

O  é um software gratuito e open-source que é utilizado para auxiliar nos mais diversos estudos de Data Science, Matemática, Estatística entre outros, possui várias ferramentas úteis para manipulação de dados e tratamento dos mesmos, como por exemplo: (modelagem linear e não-linear, testes estatísticos clássicos, análise de séries temporais, classificações, entre outras opções) e suporte gráfico e altamente versátil e de fácil utilização. Foi criado originalmente por Ross Ihaka e por Robert Gentleman no departamento de Estatística da universidade de Auckland, Nova Zelândia. Para este projecto de investigação serão utilizadas principalmente as seguintes ferramentas disponibilizadas não só pelo  mas também por outros programadores como por exemplo, testes para variáveis unilaterais e bilaterais, suporte gráfico para construção de gráficos como ("boxplot", "gráficos circulares", "gráficos de barras", entre outros), utilização de regressões lineares etc...

2 Caracterização da amostra

O questionário "burnout" que nos foi fornecido para a realizar o trabalho de grupo da unidade curricular de Análise e Tratamento de Dados Multivariados - ATDM do curso de Licenciatura em Bioinformática tem por alvo os estudantes da ESTBarreiro/IPS, foi adaptado um questionário real, os dados coletados sobre os mesmos foram utilizados principalmente para verificar a opinião sobre as mais diversas situações escolares, tempo dos estudantes relativamente à escola e lazeres dos mesmos. Para iniciar o trabalho, houve um acerto/correção das variáveis uma vez que nem todos os estudantes responderam da maneira prevista ou responderam com respostas em branco ou inválidas. Houve uma atenção especial para as questões da bivariada uma vez que permite colocar em prova se de facto questões como falta de horas de sono, tempo nas redes sociais entre outros "problemas" na vida de um jovem de facto o afetam a nível escolar e pessoal.

3 Análise estatística univariada

A análise descritiva univariada é um ramo da estatística que aplica várias técnicas para descrever e sumarizar um conjunto de dados. Diferencia-se da estatística inferencial, pelo seu objetivo: organização e sumarização de dados. Algumas medidas que são normalmente usadas para descrever um conjunto de dados são medidas de tendência central ou um conjunto generico de elementos – média, mediana e moda, e medidas de variabilidade ou dispersão – desvio padrão, variâncias, valor máximo e mínimo.

3.1 Idade

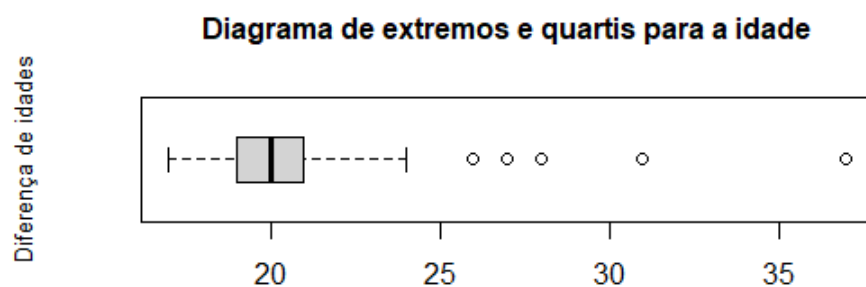


Figura 1: Diagrama de extremos e quartis da variavel idade.

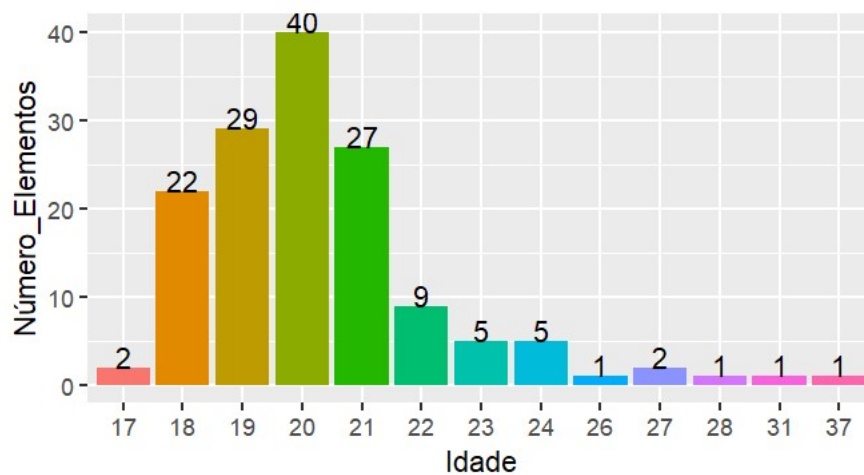


Figura 2: Histograma da variável idade.

Idade Mínima	1ºQuartil	Mediana	Media	3ºQuadrante	Idade Máxima
17.00	19.00	20.00	20.39	21.00	37.00

Figura 3: Tabela que resume a variável idade

Através da análise do histograma [figura 2], é possível verificar que das pessoas que responderam ao questionário têm idades compreendidas entre os **17 e os 37 anos**. As idades dos estudantes encontram-se principalmente entre a faixa etária dos **18 aos 21 anos**. O diagrama de Extremos e Quartis mostra que a distribuição dos dados referentes as idades é muito próxima de uma **distribuição simétrica**, possui cerca de **5 outliers superiores**, sendo que um deles é outlier superior severo (referente a Idade 37).

3.2 Sexo

Relativamente a amostra da variável sexo dos estudantes da ESTBarreiro que responderam ao inquerito pode-se retirar que dos **145** estudados, maioria pertencem ao **sexo feminino com 55.2%**.

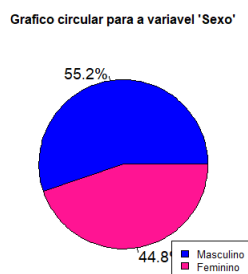


Figura 4: Gráfico de circular da variavel sexo.

3.3 Curso

Relativamente, a distribuição entre cursos, **Biotechnologia é o curso onde se obteve uma maior número de escolhas** entre todos os cursos disponíveis, sendo este o curso mais desejado entre os alunos da EstBarreiro com um **total de 67.4%**, onde essa mesma conclusão se pode retirar do gráfico circular (figura 5) e do gráfico de barras (figura 6) onde a dispersão entre cursos se torna mais evidente.

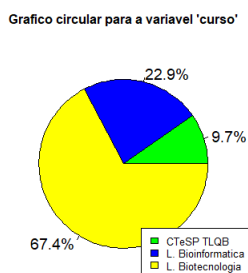


Figura 5: Grafico circular da variável curso

3.4 Ano Curricular

Os alunos da EstBarreiro encontram-se distribuídos entre 3 anos curriculares, podemos verificar que o 1º ano representa o maior numero de estudantes(48%) e o 2º ano representa o menor numero de estudantes(23%), como podemos observar no gráfico de barras[figura 6]. Através do diagrama de Extremos e Quartis verificamos uma distribuição simétrica.

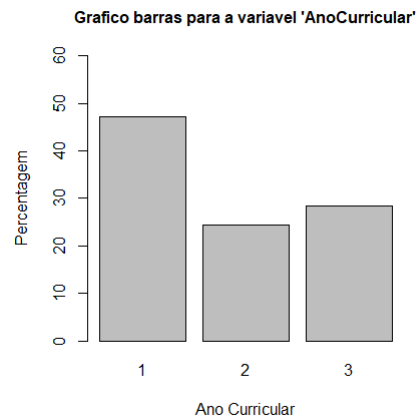


Figura 6: Gráfico de barras para ano curricular

3.5 Qual a primeira Opção

Na escolha das opções, existe uma pouca divergencia entre respostas sim e não, contudo 51,4% dos alunos colocou como primeira opção o curso.

Gráfico circular para a variável 'Opção'

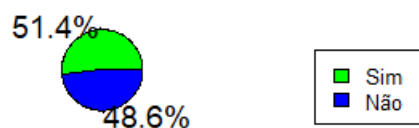


Figura 7: Grafico circular da variavel primeira opção

3.6 Escolhi este curso

Sobre a decisão da escolha do curso, uma larga maioria escolheu o curso, apenas **0.7%** dos alunos não foram os próprios a escolher qual curso pretendiam frequentar, fica claramente evidente através do gráfico circular e de barras a grande discrepância entre as respostas sim e não.

Grafico circular para a variavel 'Escolhi_Este_Curso'

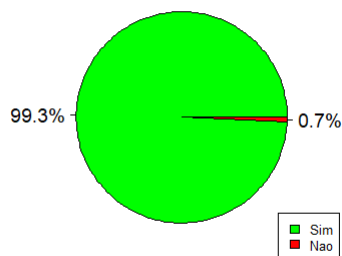


Figura 8: Grafico Circular da variavel escolhi este grupo

3.7 Tempo Deslocação

rama de extremos e quartis para a Variável "Tempo de De:

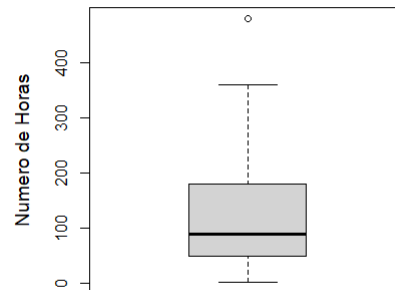


Figura 9: diagrama de extremos e quartis da variavel tempo deslocação

Através do diagrama de extremos e quartis podemos verificar que os dados têm uma distribuição assimétrica positiva. O tempo mínimo de ida e volta é de cerca de **2 minutos** e o máximo são **480 minutos**. A variável tempo de Deslocação **apresenta outliers**, ou seja, valores que apresentam um afastamento da sequência do tempo de deslocação um exemplo disso é o **tempo máximo de deslocação com 1 outlier, 480 minutos**.

Tempo médio de deslocação é de 111 minutos.

3.8 Horas de Estudo

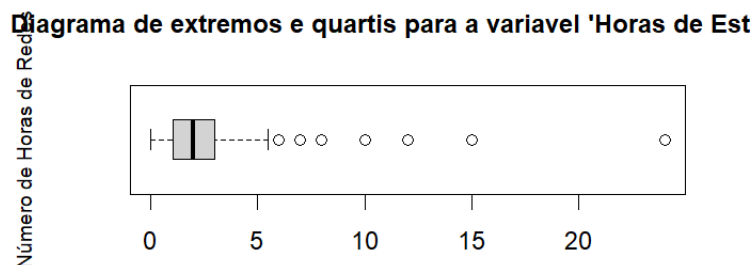


Figura 10: diagrama de extremos e quartis da variável horas de estudo.

Com a análise do diagrama de extremos e quartis é possível verificar que os dados têm uma distribuição assimétrica positiva. O tempo mínimo de estudo é de cerca de **0 horas** e o máximo são **70 horas**. A variável Tempo de Estudo apresenta outliers, ou seja, valores que apresentam um afastamento da sequência do tempo de estudo um exemplo disso é o tempo máximo de estudo com 4 outliers, o mais afastado é o que corresponde a 70 horas de Estudo. **Tempo médio de estudo é de cerca de 9 horas.**

3.9 Horas de Redes

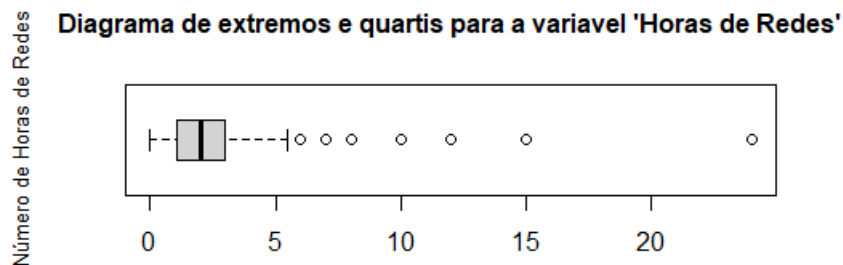


Figura 11: Diagrama de extremos e quartis da variável Horas de Redes

Através da análise dos resultados do diagrama de extremos e quartis podemos verificar que os dados têm uma ligeira distribuição assimétrica positiva. O tempo mínimo de horas de redes é de cerca de **0 horas** e o **máximo são 24 horas**. A variável Tempo de Redes apresenta outliers, ou seja, valores que apresentam um afastamento da sequência do tempo de deslocação um exemplo disso é o tempo máximo de Redes com **7 outliers**, o mais afastado é o que corresponde a **24 horas** de Redes. **Tempo médio de deslocação é de cerca de 3 horas.**

3.10 Horas TV

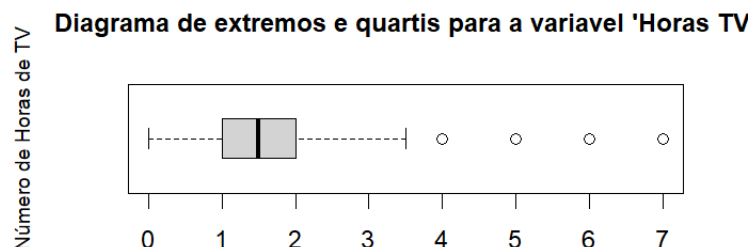


Figura 12: diagrama de extremos e quartis da variavel horas de TV.

Com a análise do diagrama de extremos e quartis é possível verificar que os dados têm uma distribuição simétrica. O tempo mínimo de estudo é de cerca de 0 horas e o máximo são 7 horas. A variavel Tempo de TV apresenta outliers, ou seja, valores que apresentam um afastamento da sequência do tempo de estudo um exemplo disso é o tempo máximo de TV com 4 outliers, o mais afastado é o que corresponde a 7 horas de Estudo. Tempo médio de estudo é de cerca de 2 horas.

3.11 Horas de Sono

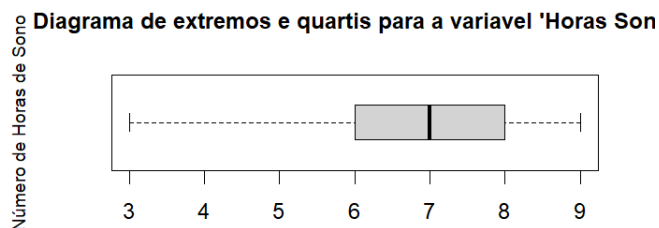


Figura 13: diagrama de extremos e quartis da variavel horas de sono.

Com a análise do diagrama de extremos e quartis é possível verificar que os dados têm uma distribuição simétrica. O tempo mínimo de estudo é de cerca de 3 horas e o máximo são 9 horas. Tempo médio de estudo é de cerca de 7 horas.

3.12 Pmentor

Gráfico circular para a variável 'Mentoria'

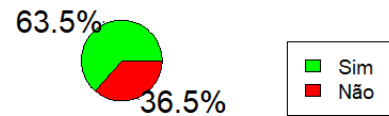


Figura 14: grafico circular da variavel Pmentor.

Relativamente sobre a existencia de um mentor, os alunos apresentam uma maioria de respostas afirmativas, cerca de 63.5%.

4 Análise estatística Bivariada

Com a análise da estatística bivariada (duas variáveis) é possível observar como duas variáveis se comportam na presença uma da outra. Esta análise tanto pode ser feita em termos de distribuição (para duas variáveis ordinais) como em termos de frequências para variáveis nominais.

A tabela seguinte (tabela 1) sistematiza os dados referentes a pergunta "BurnoutP3" e a variável sexo.

Pretende-se com análise estatística Bivariada estudar a relação entre duas variáveis de diferentes tipos.

4.1 Tabela de contingência

	Feminino	Masculino
Nunca	13	21
Quase nunca	19	23
Algumas vezes	15	19
Regularmente	1	5
Muitas vezes	5	4
Quase sempre	8	5
Sempre	3	2

Tabela 1: Tabela de contingência da questão BurnoutP3 segundo o sexo dos estudantes

Sendo a tabela de contingência a maneira mais simples de resumir as informações de duas variáveis num estudo bivariado é possível tirar uma rápida conclusão. Que numa primeira análise, verifica-se que não existe grande diferença entre o número de elementos do sexo Feminino e Masculino.

4.2 Coeficientes de associação/correlação

4.2.1 Horas de Sono VS Horas de Redes (Pearson)

Para verificar se duas variáveis tinham algum tipo de relação foi utilizado alguns testes como pearson e spearman respetivamente, as variáveis avaliadas foram "Horas de Sono" (HS) e "Horas de TV" (HTV) respetivamente para o coeficiente de pearson.

⇒ **Variáveis Quantitativas**

Através do coeficiente de pearson, verificamos que existe uma correlação positiva, embora muito fraca ou até mesmo sem significância. Concluindo, podemos afirmar que não existe praticamente nenhuma ligação entre o tempo de sono dos estudantes e as horas que passam a ver televisão.

```

Pearson's product-moment correlation

data: Grupo2$HorasTV and Grupo2$HorasSono
t = 1.761, df = 141, p-value = 0.0804
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.01787925  0.30353904
sample estimates:
      cor
0.1466999

```

Figura 15: Summary do teste de Pearson entre as Var. HS e TV.

4.2.2 Varivel Sexo VS Escolhi este curso

Para o estudo da correlação entre as variáveis Sexo e Escolhi este curso, será utilizado o teste de phi (Uma excessão a Cramer, uma vez que se trata de uma variável dicotômica).

- Variaveis Qualitativas;
- Amostra total é ≥ 20 elementos;
- Nenhuma das células possui freq. esperadas < 5 ;
- Todas as células possuem valores de freq. esperadas > 5 ;

Realizou-se um **teste de Qui-Quadrado** para verificar se as variáveis são independentes. Foi possível obter um **p-value(0.436)** que é > 0.05 .

Logo pode se concluir que de facto as variáveis Sexo e Escolhi este curso são independentes logo pode-se seguir para o teste de Phi e tirar-se as ultimas conclusões.

Freq Esperada	Feminino	Masculino
Sim	36	38
Não	29	42

Tabela 2: Tabela com as frequencias esperadas

Com análise do valor do coeficiente de Phi, **0.0784** verificamos que a correlação entre a escolha do curso e o sexo tem pouca ou nenhuma relação entre elas, o que nos permite concluir que não existe qualquer relação.

5 Estudo Inferencial

A Inferência estatística é um ramo da Estatística cujo objetivo é fazer afirmações a partir de um conjunto de valores representativo (amostra) sobre um universo (população), assume-se que a população é muito maior do que o conjunto de dados observados, uma vez que é praticamente impossível contabilizar todos os casos então será necessário generalizar a amostra. Tal tipo de afirmação deve sempre vir acompanhada de uma medida de precisão sobre sua veracidade. Para realizar este trabalho, será necessário recolher informações/dados de dois tipos, experimentais (as amostras).

Pretende-se verificar se é possível responder as questões que serão levantadas. Neste estudo serão levantadas 3 possíveis hipóteses de investigação sendo elas:

5.1 Primeira Pergunta

O curso de biotecnologia é igualmente repartido entre alunos do sexo masculino e do sexo feminino.

Para responder à primeira hipótese, visto que a probabilidade de um aluno do curso de biotecnologia ser rapaz ou rapariga é de 50 %, será utilizado um teste binomial para a resolução desta hipótese.

Como possíveis respostas ao nosso problema serão;

⇒ O curso de biotecnologia **é igualmente** repartido entre os sexos.

⇒ O curso de biotecnologia **não é igualmente** repartido entre os sexos.

```
Exact binomial test

data: 52 and 96
number of successes = 52, number of trials = 96, p-value = 0.4752
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4368634 0.6438297
sample estimates:
probability of success
 0.5416667
```

Figura 16: Resumo do teste Binomial.

Para um nível de significância de 5 %, pode-se admitir que a distribuição dos sexos no curso de biotecnologia é igualmente repartido. Para um p-value de 0.475 como podemos observar na figura acima [Figura 16]. Logo não se rejeita a hipótese inicial (hipótese nula).

5.2 Segunda Pergunta

Nesta segunda pergunta verificou-se se existe alguma relação de dependência entre primeira opção que o aluno escolheu e o sexo do mesmo.

Sendo assim duas respostas possíveis para o nosso estudo serão:

⇒ A opção do aluno escolher o curso **é independente** do sexo do aluno.

⇒ A opção do aluno escolher o curso **não é independente** do sexo do aluno.

Freq Esperada	Sexo Feminino	Sexo Masculino
Escolhi este curso	65	78
Não escolhi este curso	0	1

Tabela 3: Tabela com as frequências esperadas

O teste do Qui-Quadrado de independência só pode ser aplicado com rigor se:

- Variáveis Qualitativas;
- $N > 20$;
- Todas as classes possuam frequências esperadas > 1
- Pelo menos 80% das classes possuam frequências esperadas ≥ 5

Uma vez que um dos pressupostos para aplicação do teste do Qui-Quadrado falhou, será necessário utilizar outro teste, tratando-se de uma tabela de 2×2 , o teste de Fisher será o mais indicado para a resolução da questão anteriormente colocada.

Fisher's Exact Test for Count Data

```
data: tab
p-value = 0.4045
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.6757774 2.7905991
sample estimates:
odds ratio
 1.369028
```

Figura 17: Resumo do teste Fisher.

Para um intervalo de confiança de 95 %, pode-se admitir que a escolha do curso é independente do sexo do aluno, pois para um p-value de 0.404 como podemos observar na figura acima [Figura 17]. Logo não se rejeita a primeira hipótese colocada (hipótese nula).

5.3 Terceira Pergunta

Foi estudado o tempo que os alunos da ESTBarreiro passam nas redes sociais, o resultado do teste indica uma media de cerca 3 horas entre os dois sexos, sexo masculino e feminino. Pretende-se verificar se o tempo que os alunos passam nas redes sociais difere entre os sexos.

⇒ O tempo que os alunos passam é **diferente** do sexo do aluno.

⇒ A opção do aluno escolher o curso **não é independente** do sexo do aluno.

Para verificação da normalidade de uma amostra que neste caso é "Horas de Redes" será utilizado um teste de lilliefors este que é um teste adaptado de Kolmogorov-Smirnov.

```
Lilliefors (Kolmogorov-Smirnov) normality test  
  
data: Grupo2$HorasRedes  
D = 0.2771, p-value < 2.2e-16
```

Figura 18: Resumo do teste Lille.

Através do valor do p-value que é bastante inferior a 0.001, podemos neste tipo de situações será necessario utilizar um teste não paramétrico para chegar as conclusões. O teste não paramétrico utilizado é de Wilcoxon, este teste que só é utilizado quando não se verificam os pressupostos anteriores.

```
Wilcoxon signed rank test  
  
data: Grupo2$HorasRedes  
V = 151.5, p-value < 2.2e-16  
alternative hypothesis: true location is not equal to 10
```

Figura 19: Resumo do teste Wilcoxon.

6 Métodos de Análise de Dados Multivariados

6.1 Regressão Linear Múltipla

Análise dos dados

Na regressão linear simples são utilizados os conceitos e técnicas para analisar e utilizar a relação linear entre duas variáveis. Da análise das regressões seguintes resulta 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 que podem ser várias se necessário).

Para a análise dos dados do questionário Burnout, será escolhida uma (**variável, NHE) (Número de Horas de Estudo.)** Esta será a variável dependente neste próximo estudo, sendo que as restantes variáveis serão variáveis independentes. As outras variáveis são: Número de Horas de Redes (**NHR**), Número de Tempo de Deslocação(**NTD**). Com este próximo estudo será utilizado uma série de teste e modelos gráficos para tentar verificar se existe algum tipo de relação entre as variáveis referidas anteriormente.

Para os estudos em diante será utilizado exclusivamente intervalos de confiança de **95%**.

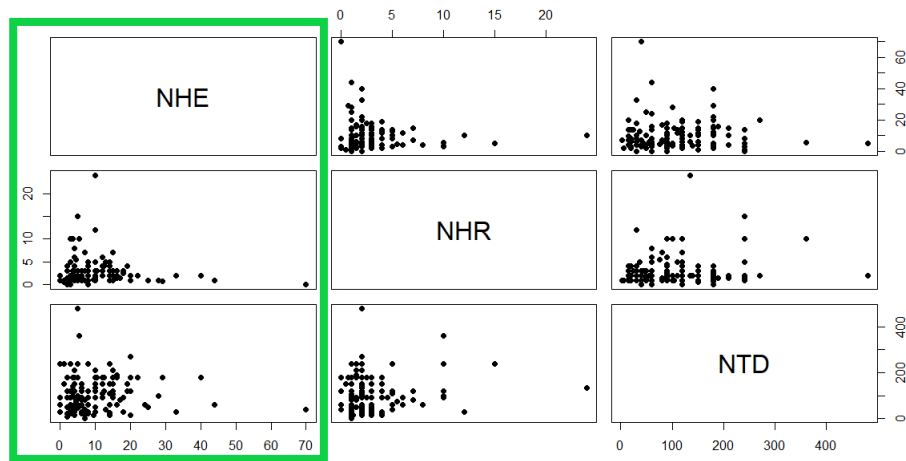


Figura 20: Gráficos da relação entre variáveis.

Sendo que a nossa **variável NHE** será a **dependente** e todas as outras independentes, apenas será analisado as relações entre os gráficos que estão em torno do retângulo verde.

Através de uma primeira análise dos graficos verificamos que a partida não são cumpridos os pressupostos para a regressão linear. Apesar dessa análise gráfica serão verificados todos os pressupostos nas próximas alíneas com auxílio de testes estatísticos.

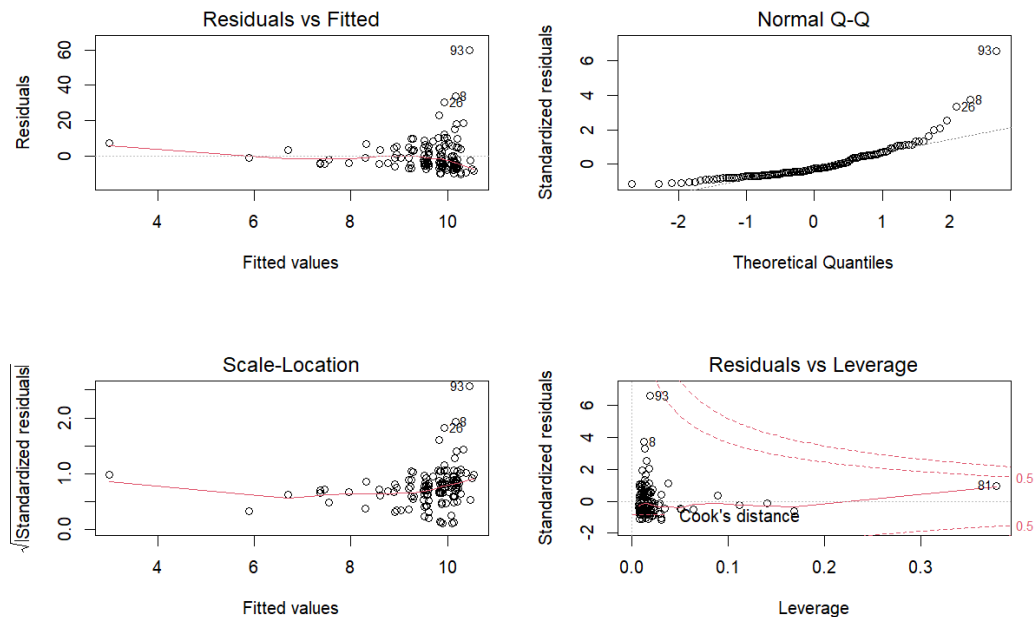


Figura 21: Gráficos da relação entre variáveis.

Dos 4 gráficos, o primeiro **Gráfico dos resíduos pelos valores previstos**, este gráfico permite analisar os pressupostos de linearidade e de homogeneidade das variâncias, com análise deste gráfico verifica-se que de facto não existe uma reta linear onde provavelmente estariam os resíduos (pontos) o que indica que não existe linearidade.

Segundo gráfico, trata-se dos **Distribuição dos resíduos standarizados**, a análise deste gráfico permite verificar que pressuposto referente a distribuição dos resíduos necessita de possuir uma distribuição Normal, o que com análise do gráfico verifica-se que numa primeira análise poderia-se aceitar a possibilidade de existir uma distribuição Normal.

Terceiro gráfico, trata-se dos **resíduos standarizados**, mas neste caso com uma vertente para a **Homogeneidade** entre variáveis, novamente graficamente seria possível aceitar a homogeneidade dos resíduos.

O ultimo gráfico, permite um estudo dos resíduos que podem ser considerados **outliers**, ou seja, todos os pontos que se encontrem além das linhas picotadas e para além da reta (**acima de 2 e -2**) são considerados outliers, neste caso ainda existem uns quantos.

6.2 Verificação dos pressuposto através de testes

Numa primeira análise gráfica, já é possível tirar algumas conclusões, embora seja sempre preferível retirar conclusões através de análises com os respectivos testes estatísticos.

•Teste de Shapiro-Wilk Teste a Normalidade em amostras > 30

Será avaliada a normalidade dos resíduos.

Shapiro-wilk normality test

```
data: mod$residuals
W = 0.75463, p-value = 5.124e-14
```

Figura 22: Teste a Normalidade(Shapiro-Wilk).

Verifica-se que com a utilização do **teste de Shapiro-Wilk** o **p-value** ($5.12e^{-14}$) que é um valor extremamente pequeno pode ser arredondado para (0.001) e continua a ser inferior ao nosso **alpha** que foi definido a partida como **5%**, (**0.05**) isto significa que o teste a normalidade falhou, ou seja, não existe normalidade nestes resíduos.

Neste pressuposto será analisado a existencia de outliers

•Verificação de outliers

Min	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.145755	-0.623966	-0.240044	0.000821	0.424735	6.595820

Tabela 4: Sumario dos residuos da variavel mod

Verifica-se que teremos outliers tanto inferiores como superiores uma vez que , sendo os maiores os que correspondem simultaneamente aos mínimos e aos máximos.

Procura-se avaliar a homogeneidade através deste teste

•**Teste de Breusch-Pagan a homogeneidade**

```
studentized Breusch-Pagan test

data:  mod
BP = 2.854, df = 2, p-value = 0.24
```

Figura 23: Teste de Breusch-Pagan (homogeneidade).

Com análise dos resultados do teste de Breush-Pagan [Figura 23], pode-se retirar um **p-value (0.24)** que é muito superior ao nosso alpha (0.05) para um nível de significância de 5%, sendo que é possível concluir que as variáveis são homogenias.

Neste pressuposto será utilizado um teste de independência

•**Teste de Durbin-Watson a independência**

```
lag Autocorrelation D-W Statistic p-value
1      0.1392877      1.716964    0.088
Alternative hypothesis: rho != 0
```

Figura 24: Teste de Durbin-Watson (independência).

Após análise dos resultados do teste de Durbin-Watson relativamente a independência [Figura 24], pode-se concluir que existe independência, uma vez que **p-value(0.088)** é superior ao **alpha (0.05)** anteriormente definido.

No último pressuposto será avaliada a multicolinearidade

NHE	NHTV
1.00846	1.00846

Tabela 5: Analise das multicolinearidades

Neste último caso verifica-se que os resultados são ambos inferiores a 10, o que indica a ausência de multicolinearidade.

6.3 Modelo da Regressão Linear

É possível extrair uma fórmula matemática que representa um gráfico a 3 dimensões a partir do sumário da equação. Será construída uma reta do tipo $Y=B_0+B_1.X_1+B_2.X_2$. A variável dependente corresponde ao nosso B_0 (NHE), e as restantes variáveis são as independentes, logo B_1 e B_2 , são os valores da nossa estimativa. O X_1 e X_2 correspondem as variáveis NHR e NTD.

$$NHE = -0.3134.NHR + 0.0007.NTD + \varepsilon$$

```
Call:
lm(formula = NHE ~ NHR + NTD, data = Grupo2)

Residuals:
    Min       1Q   Median       3Q      Max
-10.280  -5.623  -2.165   3.818   59.542

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.4305445   1.4927206   6.988 1.11e-10 ***
NHR          -0.3134148   0.2640364  -1.187   0.237
NTD           0.0006797   0.0098473   0.069   0.945
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.114 on 137 degrees of freedom
(6 observations deleted due to missingness)
Multiple R-squared:  0.01019,    Adjusted R-squared:  -0.004258
F-statistic: 0.7053 on 2 and 137 DF,  p-value: 0.4957
```

Figura 25: Análise do modelo.

Com análise do sumário pode-se tirar algumas conclusões:

- Em média, por cada hora utilizada em redes a nota diminui 0.3134 valores.
- Em média, por cada hora utilizada em transportes a nota aumenta 0.0007 valores.
- Significa que indica que 1,02 % da variância é explicada pela variável independente.
- O p-value > alpha logo o modelo não é estatisticamente significativo
- O modelo apresenta uma significância prática negativa, ou seja, consideramos o valor como 0 e que não possui qualquer significância prática.

Multiple R-squared(R^2)

Permite avaliar quão bom será o nosso gráfico, se multiplicarmos por 100, vamos obter uma percentagem.

Adjusted R-squared

Permite avaliar a utilidade prática do nosso gráfico para um estudo realista pode ser também expressa em percentagens.

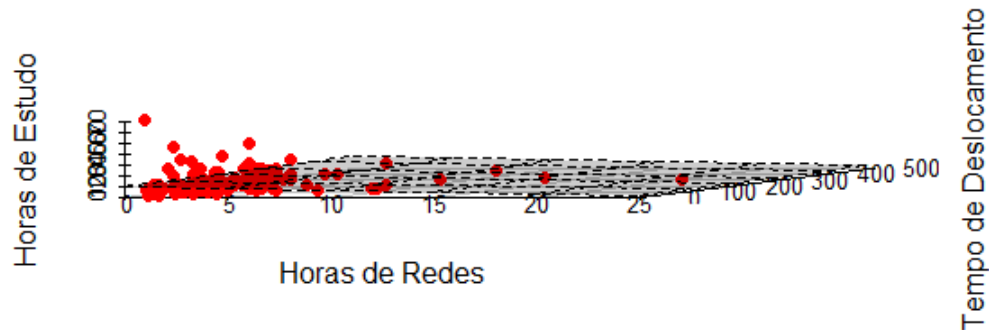


Figura 26: Análise do modelo.

7 Análise Fatorial

Análise Fatorial, divide-se em pelo menos dois modelos estatísticos, Análise Fatorial Exploratória (**AFE**), 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 em o menos numero fatores possíveis.

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

O Principal objetivo deste teste é

- encontrar tantos fatores quantos os grupos de correlação semelhantes (entre si e diferentes de outros grupos).
- se as variáveis estiverem todas correlacionadas entre si de modo idêntico (vamos ter apenas um fator.)

Para se efetuar uma análise fatorial é **necessario primeiro verificar** alguns promenores primeiro:

- Dimensão da amostra é suficiente (pelo menos 50 observações e o rácio entre observações e variáveis não deve ser inferior a 5).
- As correlações entre as variáveis originais são elevadas:
- Análise da matriz de correlações:
(Apenas possível se o número de variáveis é reduzido);
- Medida de adequabilidade de Kaiser-Mayer-Olkin (KMO);
- Teste de esfericidade de Barlett.


	Br_P1	Br_P2	Br_P3	Br_P4	Br_P5	Br_P6	Br_P7	Br_P8	Br_P9	B_P10	B_P11	B_P12	B_P13	B_P14	B_P15
Burnout_P1	1.00														
Burnout_P2	0.68	1.00													
Burnout_P3	0.53	0.71	1.00												
Burnout_P4	0.58	0.68	0.65	1.00											
Burnout_P5	0.64	0.70	0.56	0.70	1.00										
Burnout_P6	0.36	0.43	0.60	0.51	0.48	1.00									
Burnout_P7	0.34	0.40	0.57	0.41	0.33	0.70	1.00								
Burnout_P8	0.33	0.41	0.49	0.52	0.46	0.61	0.54	1.00							
Burnout_P9	0.32	0.31	0.50	0.52	0.45	0.62	0.53	0.75	1.00						
Burnout_P10	0.03	0.04	0.05	-0.03	-0.02	-0.03	-0.04	-0.02	-0.06	1.00					
Burnout_P11	0.21	0.15	0.08	0.13	0.24	0.02	-0.12	0.14	0.12	0.31	1.00				
Burnout_P12	0.09	0.12	0.01	0.02	0.07	-0.07	-0.16	-0.08	-0.13	0.35	0.39	1.00			
Burnout_P13	0.01	0.08	-0.06	-0.08	-0.01	-0.19	-0.27	-0.04	-0.15	0.09	0.27	0.33	1.00		
Burnout_P14	-0.04	-0.03	-0.20	-0.11	0.01	-0.36	-0.43	-0.37	-0.37	0.12	0.28	0.16	0.41	1.00	
Burnout_P15	-0.11	-0.13	-0.18	-0.26	-0.20	-0.27	-0.24	-0.11	-0.17	0.30	0.27	0.39	0.27	0.26	1.00

Figura 27: Análise da Matriz de Correlação de Pearson:

O 1º pressuposto já foi verificado uma vez que temos uma amostra suficiente, sendo elas um total de 133 [Figura 31].

Possuimos também uma amostra com correlações de variáveis elevadas logo o 2º pressuposto também está verificado.

O 3º refere-se a matriz que está representada na figura a cima [Figura 27].

Outra maneira de interpretar os resultados, os mesmos da [Figura 27], pode ser por exemplo utilizar a interface gráfica do  e produzir-mos uma outra matriz.

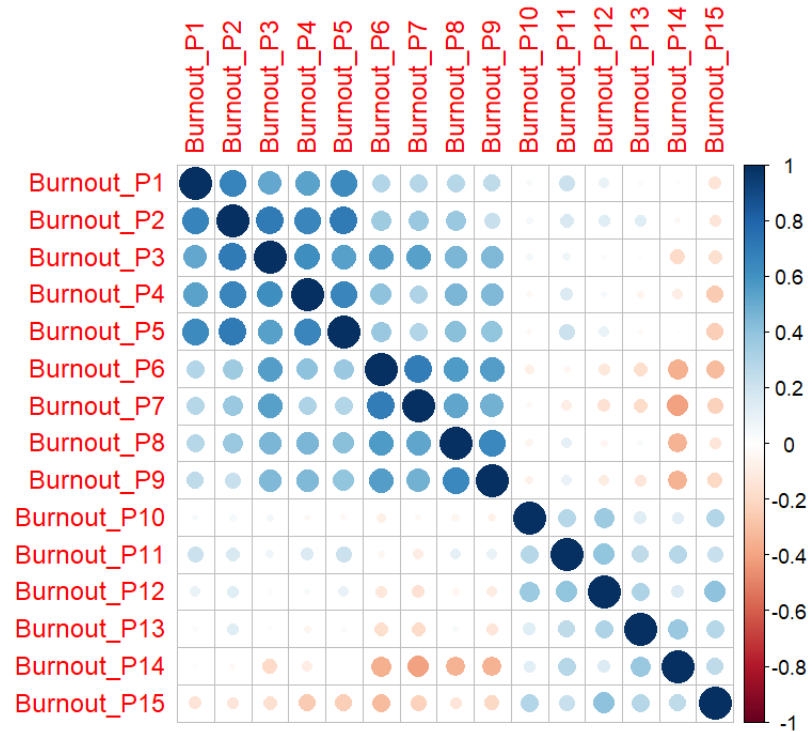


Figura 28: Análise da Matriz de Correlação de Pearson:

Através desta representação é possível retirar algumas conclusões:

Quanto maior o tom das bolas azuis, maior é a correlação.

Quanto maior o tom das bolas vermelhas, maior é a correlação, embora neste caso seja inversamente proporcional.

O valor de KMO trata-se de um valor empírico e permite avaliar a homogeneidade entre as variáveis, como **KMO (0.85)**, possui-mos uma boa AFE, sendo que este valor varia entre 0 e 1, teremos bons coeficientes de correlação parciais podemos prosseguir com a nossa análise.

Relativamente as variáveis, os valores também variam entre **0 e 1** a **correlação mais baixa é de 0.67** o que indica que todas as variáveis são válidas para o estudo e também apresentam bons valores.

O **4º ponto foi verificado com a conclusão que a adequabilidade apresenta um valor superior a 0.6**, este que é considerado o mínimo para aplicação do teste.

```

Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = Burnout)
Overall MSA = 0.85
MSA for each item =
  Burnout_P1 Burnout_P2 Burnout_P3 Burnout_P4 Burnout_P5
      0.91      0.83      0.90      0.92      0.88
  Burnout_P6 Burnout_P7 Burnout_P8 Burnout_P9 Burnout_P10
      0.90      0.88      0.83      0.84      0.72
  Burnout_P11 Burnout_P12 Burnout_P13 Burnout_P14 Burnout_P15
      0.74      0.67      0.67      0.74      0.78

```

Figura 29: Medida de adequabilidade de Kaiser-Mayer-Olkin

Teste de esfericidade de Bartlett:

Aplicação do teste de esfericidade testa a hipótese de a matriz de correlações ser a matriz identidade, ou seja, testa a hipótese de as correlações entre as diversas variáveis serem nulas.

```

$chisq
[1] 898.6965

$ p.value
[1] 2.90855e-126

$df
[1] 105

```

Figura 30: Teste de Esfericidade(Bartlett)

O valor de p-value é extremamente pequeno o que pode ser arredondado para **0.001**, o **alpha (0.05)** continua a ser maior, o que nos leva a rejeitar a hipótese de a matriz de correlações ser a matriz identidade.

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

```

Call:
princomp(x = st_dados, cor = TRUE)

Standard deviations:
  Comp.1   Comp.2   Comp.3   Comp.4   Comp.5   Comp.6   Comp.7   Comp.8   Comp.
2.3305759 1.6155856 1.1739544 0.9775075 0.8778055 0.8117252 0.7848564 0.7171908 0.682797
  Comp.11  Comp.12  Comp.13  Comp.14  Comp.15
0.5524355 0.5241767 0.4955930 0.4684986 0.4073007

15 variables and 133 observations.

```

Figura 31: Desvio padrão das variáveis

Com uma rapida analise pode-se observar os devios padrões das variaveis que serão utilizadas para o AFE. E como referido anteriormente o **número de observações (133)**

7.1 Critério de Kaiser

Existem várias regras empíricas que são utilizadas para ajudar na tomada de decisão relativamente ao número correto de fatores a considerar neste caso será utilizado o Critério de Kaiser, teriamos ainda o Critério do Scree plot este que é uma representação grafica dos fatores.

Importance of components:								
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	2.3305759	1.6155856	1.17395442	0.9775075	0.8778055	0.81172521	0.78485638	0.71719084
Proportion of Variance	0.3621056	0.1740078	0.09187793	0.0637014	0.0513695	0.04392652	0.04106664	0.03429085
Cumulative Proportion	0.3621056	0.5361134	0.62799133	0.6916927	0.7430622	0.78698875	0.82805539	0.86234623
	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15	
Standard deviation	0.68279777	0.62261561	0.55243549	0.52417669	0.49559305	0.46849864	0.40730075	
Proportion of Variance	0.03108085	0.02584335	0.02034566	0.01831741	0.01637416	0.01463273	0.01105959	
Cumulative Proportion	0.89342709	0.91927043	0.93961610	0.95793351	0.97430768	0.98894041	1.00000000	

Figura 32: Critério de Kaiser

Podemos concluir que serão selecionados um total de 5 fatores, com base na proporção de variancia, o nosso **alpha (0.05)** deve ser inferior a proporção de variancia e o 5º fator é o ultimo a possuir um valor superior ao alpha.

Proporção acumulativa diz nos a quantidade de fatores que explicam uma determinada percentagem de variabilidade total dos dados, por exemplo o **4 fator explica um total de 69.16 % da variabilidade total.**

Reduzir ou aumentar o numero de fatores não terá nenhum grande impacto benefico, aumentar o número de fatores não aumenta praticamente nada a informação(h2) a mais adequirida e diminuir o numero de fatores de um modo geral apresenta grandes perdas de informação nas variaveis, logo é preferivel manter os 5 fatores.

É possivel observar também que 74% da variabilidade total(Cumulative Var) é explicada.

```

Principal Components Analysis
Call: principal(r = st_dados, nfactors = 5, rotate = "none",
  n.obs = nrow(Burnout), scores = TRUE)
Standardized loadings (pattern matrix) based upon correlation matrix
      PC1  PC2  PC3  PC4  PC5  h2  u2 com
Burnout_P1  0.67  0.33 -0.30 -0.18 -0.02 0.68 0.32 2.1
Burnout_P2  0.75  0.35 -0.31 -0.18  0.18 0.84 0.16 2.1
Burnout_P3  0.81  0.12 -0.05 -0.16  0.20 0.74 0.26 1.3
Burnout_P4  0.81  0.17 -0.23 -0.01 -0.10 0.74 0.26 1.3
Burnout_P5  0.76  0.30 -0.29  0.02 -0.18 0.78 0.22 1.8
Burnout_P6  0.79 -0.16  0.24  0.03  0.07 0.71 0.29 1.3
Burnout_P7  0.72 -0.28  0.25 -0.13  0.23 0.73 0.27 1.9
Burnout_P8  0.75 -0.04  0.35  0.37 -0.01 0.83 0.17 1.9
Burnout_P9  0.74 -0.13  0.33  0.36 -0.15 0.83 0.17 2.1
Burnout_P10 -0.05  0.48  0.47 -0.49 -0.19 0.73 0.27 3.3
Burnout_P11  0.10  0.69  0.22  0.21 -0.48 0.81 0.19 2.3
Burnout_P12 -0.07  0.68  0.28 -0.20  0.21 0.63 0.37 1.8
Burnout_P13 -0.18  0.60 -0.09  0.48  0.44 0.82 0.18 3.1
Burnout_P14 -0.36  0.57 -0.41  0.17 -0.11 0.66 0.34 2.9
Burnout_P15 -0.32  0.52  0.42  0.02  0.24 0.61 0.39 3.1

      PC1  PC2  PC3  PC4  PC5
SS loadings      5.43 2.61 1.38 0.96 0.77
Proportion Var    0.36 0.17 0.09 0.06 0.05
Cumulative Var    0.36 0.54 0.63 0.69 0.74
Proportion Explained 0.49 0.23 0.12 0.09 0.07
Cumulative Proportion 0.49 0.72 0.85 0.93 1.00

Mean item complexity = 2.1
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 90.62 with prob < 8.5e-06

```

Figura 33: Sumario dos fatores

A matriz das componentes apresenta, para cada variável, o seu peso fatorial (loading) nos diferentes fatores.

Permite também retirar uma equação de cada variável, por exemplo:

$$Z_{Burn1} = 0.670.F1 + 0.326.F2 - 0.299.F3 - 0.178.F4 + \psi$$

$$Z_{Burn4} = 0.809.F1 + 0.165.F2 - 0.225.F3 + \psi$$

$$Z_{Burn8} = 0.749.F1 + 0.350.F3 - 0.373.F4 + \psi$$

Loadings:					
	PC1	PC2	PC3	PC4	PC5
Burnout_P1	0.670	0.326	-0.299	-0.178	
Burnout_P2	0.750	0.346	-0.311	-0.181	0.183
Burnout_P3	0.813	0.122		-0.165	0.196
Burnout_P4	0.809	0.165	-0.225		
Burnout_P5	0.756	0.300	-0.295		-0.180
Burnout_P6	0.789	-0.157	0.236		
Burnout_P7	0.721	-0.285	0.246	-0.129	0.228
Burnout_P8	0.749		0.350	0.373	
Burnout_P9	0.740	-0.129	0.333	0.359	-0.152
Burnout_P10		0.483	0.473	-0.487	-0.192
Burnout_P11	0.103	0.689	0.221	0.208	-0.479
Burnout_P12		0.682	0.275	-0.196	0.214
Burnout_P13	-0.177	0.599		0.481	0.441
Burnout_P14	-0.358	0.569	-0.408	0.166	-0.114
Burnout_P15	-0.323	0.522	0.423		0.237
	PC1	PC2	PC3	PC4	PC5
SS loadings	5.432	2.610	1.378	0.956	0.771
Proportion Var	0.362	0.174	0.092	0.064	0.051
Cumulative Var	0.362	0.536	0.628	0.692	0.743

Figura 34: Matriz de componentes

7.2 Rotação de Fatores

A rotação dos fatores tem como objetivo melhorar a interpretação dos fatores. Ou seja, fazer com que cada variável seja explicada pelo menor nº possível de fatores e, se possível, só por um, apesar de que maior parte das situações não é possível.

Será utilizado o método Varimax:

Este método baseia-se na determinação de novos loadings que são os valores que maximizam a variância dos quadrados dos loadings originais correspondentes a cada um dos fatores (este processo não garante que cada variável tenha um só loading elevado - situação ideal). A interpretação dos fatores faz-se através da matriz dos loadings obtida após rotação, onde se procuram em cada linha, os loadings significativamente elevados.

```

Principal Components Analysis
Call: principal(r = st_dados, nfactors = 5, rotate = "varimax",
  n.obs = nrow(Burnout), scores = TRUE)
Standardized loadings (pattern matrix) based upon correlation matrix
      RC1  RC3  RC2  RC4  RC5  h2  u2 com
Burnout_P1 0.81 0.10 0.05 -0.02 0.07 0.68 0.32 1.1
Burnout_P2 0.89 0.15 0.07 0.10 -0.08 0.84 0.16 1.1
Burnout_P3 0.73 0.42 0.08 -0.02 -0.17 0.74 0.26 1.8
Burnout_P4 0.77 0.33 -0.11 -0.06 0.14 0.74 0.26 1.5
Burnout_P5 0.81 0.22 -0.09 -0.01 0.27 0.78 0.22 1.4
Burnout_P6 0.41 0.71 -0.03 -0.13 -0.10 0.71 0.29 1.7
Burnout_P7 0.36 0.67 0.00 -0.19 -0.35 0.73 0.27 2.3
Burnout_P8 0.28 0.84 -0.07 0.08 0.16 0.83 0.17 1.3
Burnout_P9 0.25 0.83 -0.14 -0.05 0.23 0.83 0.17 1.4
Burnout_P10 0.03 -0.04 0.80 -0.23 0.18 0.73 0.27 1.3
Burnout_P11 0.15 0.06 0.39 0.14 0.78 0.81 0.19 1.7
Burnout_P12 0.12 -0.10 0.71 0.31 0.06 0.63 0.37 1.5
Burnout_P13 0.03 -0.12 0.14 0.88 0.11 0.82 0.18 1.1
Burnout_P14 0.09 -0.58 0.04 0.39 0.39 0.66 0.34 2.7
Burnout_P15 -0.25 -0.05 0.62 0.40 0.07 0.61 0.39 2.1

      RC1  RC3  RC2  RC4  RC5
SS loadings      3.79 3.10 1.76 1.34 1.16
Proportion Var    0.25 0.21 0.12 0.09 0.08
Cumulative Var    0.25 0.46 0.58 0.67 0.74
Proportion Explained 0.34 0.28 0.16 0.12 0.10
Cumulative Proportion 0.34 0.62 0.78 0.90 1.00

Mean item complexity = 1.6
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 90.62 with prob < 8.5e-06

```

Figura 35: Rotação de fatores

Com análise da Rotação de fatores através do método Varimax, é possível separar em 5 grandes fatores, sendo estes respetivamente :

- RC1:Desgaste Emocional
- RC2:Desinteresse Escolar
- RC3:Bom Rendimento Escolar
- RC4:Proatividade Escolar
- RC5:Interesse Escolar

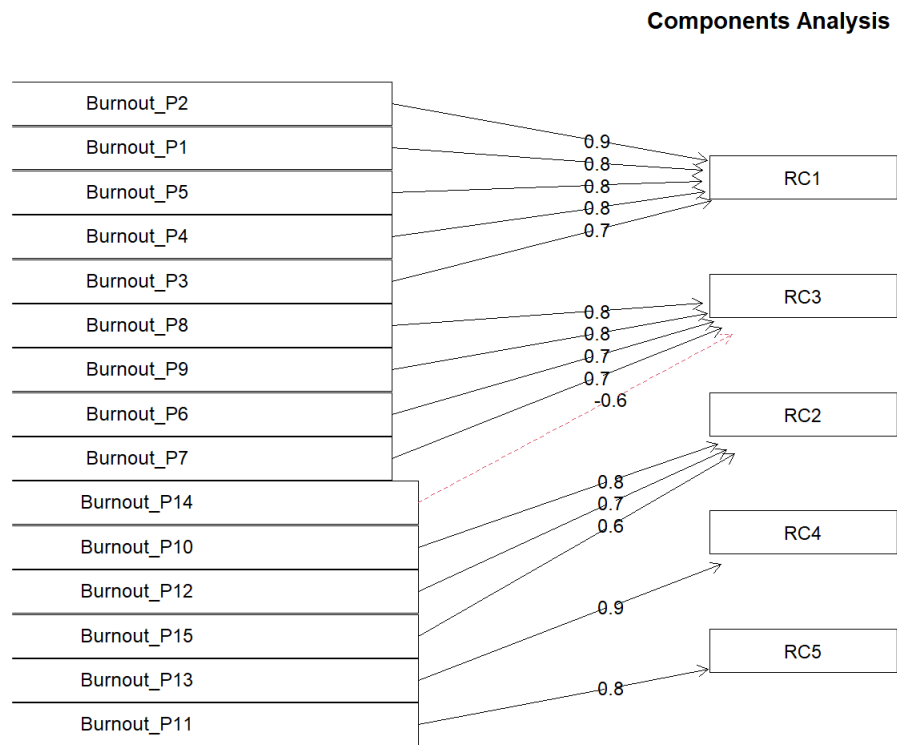


Figura 36: Modelo grafico das cargas fatoriais

Com análise deste diagrama é possível reforçar aquilo já mencionado, os 5 grandes fatores.

8 Conclusão

Análise estatística univariada, foi aplicado principalmente funções e interpretações gráficas em quase todas as questões, utilização e construção de determinados gráficos (Deste histogramas, gráficos de barras, Diagramas de extremos e quartis, gráficos circulares) para determinadas variáveis (Deste de quantitativas e qualitativas), com auxílio do Software R. Verificou-se um problema principalmente na questão dos outliers, alguns gráficos ficaram distorcidos e afetaram um pouco as interpretações dos mesmos. O estudo da Análise estatística Bivariada, teve um foco especial para as perguntas Burnout e a comparação entre elas, uso de tabelas de contingência para armazenar os dados, comparações de variáveis de diferentes tipos com os mais diversos coeficientes. Estudo Inferencial, permitiu levantar 3 hipóteses estatísticas e provar cada uma delas através de testes aprendidos no âmbito desta disciplina, tendo em conta também os pressupostos de cada teste, foi utilizado um teste binominal, teste de qui-quadrado e um teste t-student para auxiliar na resposta as hipóteses levantadas. Entrando na regressão linear, houve um foco também para Regressão Linear Múltipla, onde foi estudado as correlações entre variáveis quantitativas e construção de modelos gráficos para suporte do estudo, onde se chegou a conclusão que as variáveis não apresentavam as condições necessárias para aplicação de uma regressão linear, pode-se verificar isso através dos gráficos ou mesmo com os pressupostos que apresentavam grandes discrepâncias entre eles, foram utilizados testes para verificar também os pressupostos. A parte da regressão linear foi concluída com a equação genérica do modelo e a construção de um modelo em 3 dimensões que representa como os pontos se encontram interligados. O estudo foi finalizado com a Análise Fatorial, onde se procurava fatores comuns entre as variáveis Burnout, utilizou-se matrizes de correlação e a verificação dos mesmos pressupostos, suporte gráfico e testes estatísticos. Com a AFE foi possível destacar 5 grandes fatores anteriormente mencionados e a construção de um diagrama do mesmo.

9 Webgrafia

<https://math.uoregon.edu/wp-content/uploads/2014/12/compsymb-1qyb3zd.pdf>
(Consultado a 30-11-2020)

<https://pt.overleaf.com/learn/latex/Lists> (Consultado a 30-11-2020)

https://www.overleaf.com/learn/how-to/Using_the_symbol_palette_in_overleaf (Consultado a 19-12-2020)

[https://pt.overleaf.com/learn/latex/LaTeX_video_tutorial_for_beginners\(video3\)](https://pt.overleaf.com/learn/latex/LaTeX_video_tutorial_for_beginners(video3)) (Consultado a 19-12-2020)

<https://www.caam.rice.edu/heinken/latex/symbols.pdf> (Consultado a 19-12-2020)

https://pt.overleaf.com/learn/latex/List_of_Greek_letters_and_math_symbols (Consultado a 19-12-2020)

<https://www.stat.berkeley.edu/s133/Lr-a.html> (Consultado a 02-01-2021)

<https://study.sagepub.com/stinerock/student-resources/exercises/chapter-13-multiple-regression> (Consultado a 02-01-2021)

<http://www.sthda.com/english/wiki/scatterplot3d-3d-graphics-r-software-and-data-visualization> (Consultado a 02-01-2021)

https://www.researchgate.net/post/Interpretation_of_negativeAdjusted_R_squared (Consultado a 06-01-2021)

<https://www.statology.org/bartlettstest/> (Consultado a 10-01-2021)

<https://stattrek.com/anova/homogeneity/bartlettstest.aspx> (Consultado a 14-01-2021)

<https://medium.com/psicodata/fundamentos-da-analise-fatorial-confirmatoria-d56f49dac236> (Consultado a 19-01-2021)