INSTITUTO UNIVERSITÁRIO DE LISBOA

LICENCIATURA EM CIÊNCIA DE DADOS

Análise Exploratória de Dados

Exame de 2ª época

22 de junho 2020 Duração: 2h

Nome:

Número de aluno: **Turma:**

Exercício 1 (3,5 valores)

Na tabela abaixo está representada a evolução dos três principais grupos de população estrangeira residente em Portugal entre 2010 e 2018:

Anos	Total	Nacionalidade		
		Roménia	Cabo-Verde	Brasil
2010	199.535	36.830	43.510	119.195
2011	194.082	39.312	43.475	111.295
2012	183.122	35.216	42.388	105.518
2013	167.453	34.204	42.011	91.238
2014	157.356	31.505	40.563	85.288
2015	149.384	30.523	38.346	80.515
2016	146.191	30.429	36.193	79.569
2017	148.517	30.750	34.706	83.061
2018	169.856	30.908	34.444	104.504

População estrangeira com estatuto legal de residente: total e por algumas nacionalidades
Fontes de Dados: INE | SEF/MAI - População Estrangeira com Estatuto Legal de Residente
Fonte: PORDATA

(0,75) a) Entre 2010 e 2018 diminuiu o número de residentes em Portugal de qualquer uma destas nacionalidades; todavia, foi na comunidade caboverdiana que se registou uma diminuição mais acentuada. Indique (apresentando os cálculos) qual o decréscimo registado nesse período.

(0,75) b) Verifique (apresentando os cálculos) se a afirmação seguinte é verdadeira ou falsa:
"Por cada 1000 brasileiros residentes em Portugal em 2010, em 2018 não chegavam a 900" .

(1,0) c) Admitindo que a evolução da população romena residente em Portugal, entre 2018 e 2028, será idêntica à que se registou entre 2010 e 2018, determine o valor previsto para 2028.

(1,0) d) Calcule (apresentando os cálculos) um índice de base móvel para o total dos residentes em Portugal destas três nacionalidades e interprete o valor obtido para 2018.

Exercício 2 (7,0 valores)

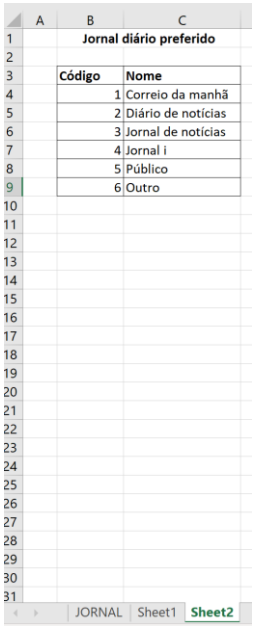
Foi feita uma sondagem a uma amostra de 100 leitores de jornais, com o objetivo de avaliar os seus hábitos de leitura, bem como as características dos jornais mais valorizadas por eles.

As respostas foram obtidas através da aplicação de um questionário, tendo os dados sido introduzidos no Excel por mais do que um indivíduo. A equipa de investigação, antes de iniciar a análise dos dados, pretende assegurar-se de que não existiu duplicação de casos aquando da construção do ficheiro de Excel.

(1,5) a) Descreva sucintamente os procedimentos a seguir para procurar e identificar possíveis casos duplicados.

(1,5) b) Numa das perguntas do questionário pedia-se aos leitores que indicassem qual o seu jornal diário preferido. Ao introduzir as respostas no ficheiro Excel foram introduzidos códigos em vez dos titulos dos jornais conforme pode ver abaixo. Indique como faria (escrevendo a função) para substituir os códigos dos jornais pelos títulos respetivos.

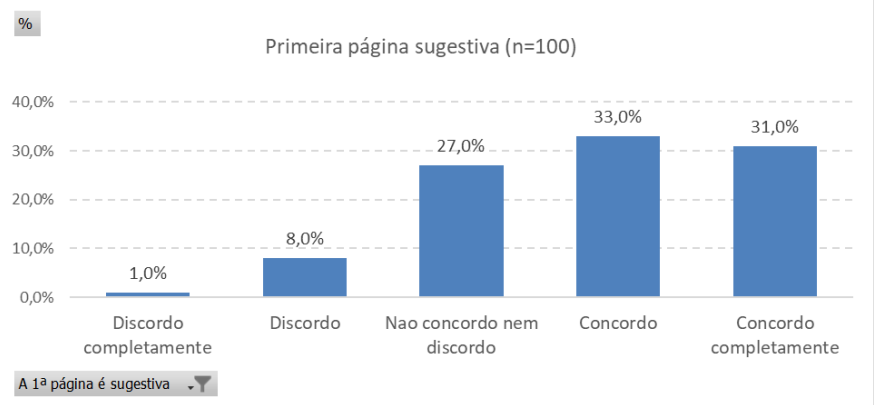
	A	B	C	D	E	F
1	N_questior	Sexo	Idade	Grupo etário	Habilitaçõe	Diário preferido
2	1	Feminino	33	de 30 a 34 anos	até ensino	2
3	2	Feminino	20	< 25 anos	até ensino	1
4	3	Feminino	99		até ensino	1
5	4	Feminino	32	de 30 a 34 anos	até ensino	3
6	5	Feminino	33	de 30 a 34 anos	até ensino	2
7	6	Feminino	37	de 35 a 39 anos	até ensino	4
8	7	Feminino	30	de 30 a 34 anos	até ensino	6
9	8	Feminino	30	de 30 a 34 anos	até ensino	5
10	9	Feminino	28	de 25 a 29 anos	até ensino	1
11	10	Feminino	25	de 25 a 29 anos	até ensino	1



Apresentam-se em baixo alguns resultados obtidos em Excel relativamente ao número de jornais lidos por mês, qual o jornal preferido, bem como o grau de concordância quanto à importância da 1ª página ser sugestiva.

Número de jornais lidos por mês	
Mean	7,0
Standard Error	0
Median	6,0
Mode	6,0
Standard Deviation	3,21
Sample Variance	10,28
Kurtosis	0,9
Skewness	1,0
Range	16,0
Minimum	0,0
Maximum	16,0
Sum	669,0
Count	96,0

Jornal preferido	**n**	**%**
Correio da manhã	43	44,8%
Público	16	16,7%
Diário de notícias	15	15,6%
Jornal de notícias	14	14,6%
Jornal i	8	8,3%
Total	**96**	**100,0%**



(2,5) c) Escreva um pequeno texto com a descrição e interpretação dos resultados apresentados.

(1,5) d) Com base na leitura das percentagens que lhe parecerem mais adequadas (apresentadas na tabela de cruzamentos abaixo) verifique se a hipótese, apontada pela equipa de investigação, de que homens e mulheres têm diferentes formas de ler o jornal, é verdadeira.

Como costuma ler o jornal?	**Sexo**						**Total**
	Feminino			**Masculino**			
	n	**% Coluna**	**% Linha**	**n**	**% Coluna**	**% Linha**	
Dar uma vista de olhos	3	8,8%	60,0%	2	3,2%	40,0%	5
Ler algumas secções	19	55,9%	38,0%	31	49,2%	62,0%	50
Ler com atenção	12	35,3%	28,6%	30	47,6%	71,4%	42
Total	**34**	**100,0%**	**35,1%**	**63**	**100,0%**	**64,9%**	**97**

Exercício 3 (7,0 valores)

Para além do Excel, a equipa de investigação utilizou igualmente o R.

(2,5) a) Indique a finalidade de cada linha de código apresentado abaixo:

a1 `bd<-read.xlsx("Jornais.xlsx")`

a2 `colnames(bd)`

a3 `names(bd)[9]<-c("Nº.jornais.lidos")`

a4 `is.na(bd$Nº.jornais.lidos)`

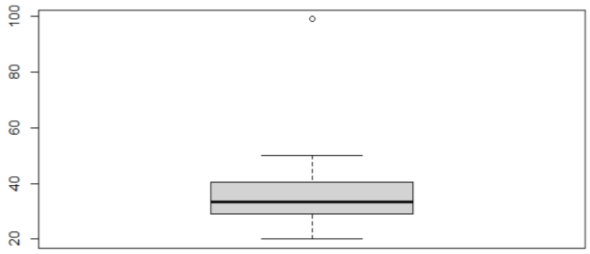
a5 `bd$Nº.jornais.lidos[which(is.na(bd$Nº.jornais.lidos))]<-median(bd$Nº.jornais.lidos,na.rm=T)`

a6 `summary(bd$Nº.jornais.lidos)`

a7 `tab1<-table(bd$Grupo.etário)`

a8 `tab2<-round(prop.table(tab1)*100, digits=1)`

(2,0) b) Aquando da análise da distribuição das idades foi construído o gráfico que pode ver em seguida:



Na sequência da análise deste gráfico o investigador escreveu as linhas de código seguintes:

```
Idade.nova <-bd$Idade[which(bd$Idade<60)]

boxplot(bd$Idade.nova)
```

Indique o que foi feito e qual a razão pela qual terá sido feito.

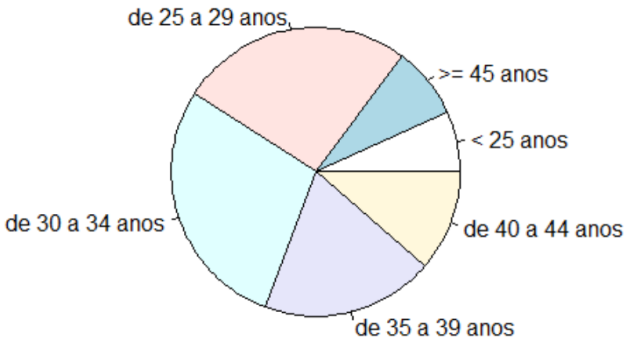
(1,0) c) Durante a análise dos dados, ao mandar executar o comando seguinte, com o objetivo de construir um gráfico circular:

```
pie (bd$Grupo.etário)
```

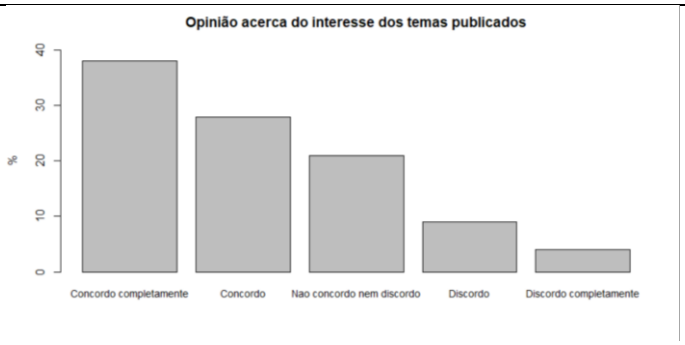
o investigador obteve a seguinte mensagem de erro:

```
Error in pie(bd$Grupo.etário)
```

Corrija o código de forma a obter o gráfico seguinte:



(1,5) d) Apresente o comando necessário para construir o gráfico apresentado abaixo:




```
t1<-prop.table(bd$Temas.de.interesse))*100
t1
Concordo    Concordo completamente
       28                        38
    Discordo    Discordo completamente
        9                         4
Nao concordo nem discordo
                       21
```

Exercício 4 (2,5)

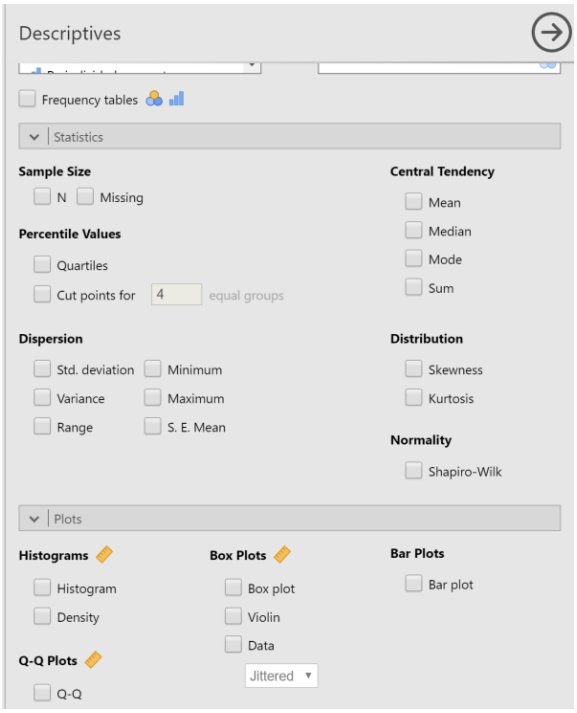
A equipa de investigação utilizou também o Jamovi para fazer algumas das análises. Abaixo pode ver a janela onde foram pedidas as medidas de estatística descritiva e gráficos para analisar as respostas às variáveis "Tempo de leitura de jornais diários por semana (minutos)" e "Grupo etário".

Indique, justificando, **qual a natureza** destas variáveis e assinale, nas janelas abaixo, **as opções necessárias para fazer a análise** das respostas dos leitores a estas duas questões.

(1,25) a) Tempo de leitura de jornais diários por semana (minutos).

(1,25) b) Grupo etário.

a) Tempo de leitura de jornais diários por semana (minutos)



b) Grupo etário