

LICENCIATURA EM CIÊNCIA DE DADOS

Análise Exploratória de Dados

Teste/Exame de 1ª época

05 de junho 2021

Duração: 2h

Nome:

Número de aluno:

Turma:

1. Num estudo acerca do alojamento local (AL) efetuado em 3 regiões de Portugal (Alentejo, Centro e Norte), foi recolhida informação acerca da nacionalidade dos hóspedes. Essa informação pode ser vista na tabela 1.

Tabela 1: Nacionalidade dos hóspedes por região

Nacionalidade dos hóspedes	Região										
	Alentejo			Centro			Norte			Total	
	n	% coluna	%linha	n	% coluna	%linha	n	% coluna	%linha	n	% coluna
Alemã	76	22,1	15,0	197	10,1	39,0	232	11,6	45,9	505	11,8
Inglesa	89	25,9	15,0	195	10,0	32,8	311	15,5	52,3	595	13,9
Francesa	30	8,7	6,1	271	13,9	55,1	191	9,5	38,8	492	11,5
Espanhola	32	9,3	3,9	301	15,5	36,4	494	24,7	59,7	827	19,3
Brasileira	31	9,0	8,0	150	7,7	38,8	206	10,3	53,2	387	9,0
Portuguesa	63	18,3	5,6	615	31,6	54,4	453	22,6	40,1	1131	26,3
Outras	23	6,7	6,4	219	11,2	61,3	115	5,7	32,2	357	8,3
Total	344	100,0	8,0	1948	100,0	45,4	2002	100,0	46,6	4294	100,0

(3,0) Complete as afirmações seguintes:

- 1.1. No total do país a nacionalidade de hóspedes de AL mais frequente é a _____ (____%).
- 1.2. A percentagem de hóspedes de nacionalidade francesa que escolhem a região Centro (____%) é _____ (igual/inferior/superior) à percentagem de hóspedes dessa mesma nacionalidade que escolhem o Norte (____%).

1.3. A nacionalidade dos hóspedes de AL mais frequente no Alentejo é_____ (igual/diferente) à nacionalidade predominante no Centro.

2. Na tabela 2 pode observar a distribuição dos hóspedes segundo a principal razão para a opção pelo alojamento local e as habilitações.

Tabela 2: Principal razão para a opção pelo AL segundo as habilitações

Principal razão para a opção pelo AL	Habilitações						Total	
	Básico		Secundário		Superior			
	n	%	n	%	n	%	n	%
Preço	83	30,9	513	25,7	432	19,6	1028	23,0
Recomendação de amigos	72	26,8	291	14,6	261	11,8	624	14,0
Pontuação do AL na plataforma de reservas	31	11,5	438	22,0	623	28,3	1092	24,4
Reviews de outros hóspedes na plataforma de reserv	26	9,7	421	21,1	519	23,5	966	21,6
Localização	57	21,2	332	16,6	370	16,8	759	17,0
Total	269	100,0	1995	100,0	2205	100,0	4469	100,0

(1,0) **2.1.** Qual a natureza das variáveis em análise? Justifique

(1,5) **2.2.** Poderemos afirmar que, nesta amostra, a principal razão para a opção pelo AL é diferente consoante as habilitações dos hóspedes? Justifique.

3. Os valores do rendimento obtido com dormidas por parte dos estabelecimentos de Alojamento Local, entre 2016 e 2019, encontram-se na tabela 3, onde se adicionou também um deflator, o Produto Interno Bruto (PIB), com base em 2016.

Tabela 3: Proveitos com dormidas nos AL e deflator (PIB) (2016-2019)

Anos	Alojamento Local	PIB
2016	170040	1
2017	227461	1,015
2018	277424	1,034
2019	340580	1,052
2020	123643	1,078

Nota: Valores preliminares para 2020

(1,5) **3.1.** Calcule os valores dos proveitos a preços constantes de 2016, considerando o deflator PIB (apresente os cálculos realizados). Interprete o valor encontrado para 2019.

(1,5) **3.2.** De acordo com os dados preliminares de 2020 é possível antever uma queda nos proveitos face a 2019. Calcule a taxa de variação prevista para 2020, relativamente ao ano anterior (apresente os cálculos realizados). Interprete o valor encontrado.

4. Foi também perguntado aos donos dos alojamentos como eram feitas as reservas. Na tabela 3 pode ser vista a distribuição percentual pelos vários tipos de operadores bem como a % de reservas feita diretamente com o alojamento.

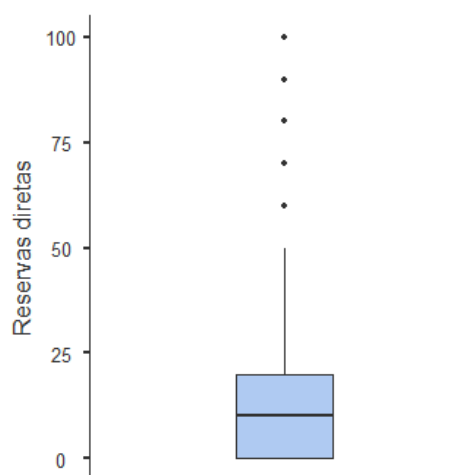
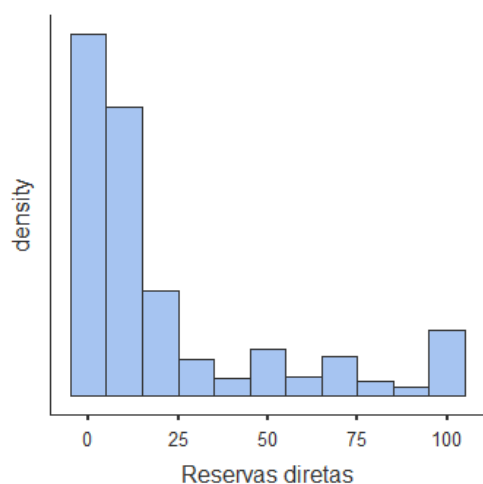
Tabela 3: Percentagem de reservas por operador

	Reservas diretas	HomeAway	Booking	Expedia	Airbnb	Outros
N	1761	1761	1761	1761	1761	1761
Mean	21.6	8.36	42.2	2.17	17.9	7.29
Median	10	0	50	0	10	0
Standard deviation	29.5	21.7	34.1	6.08	26.8	17.7
Minimum	0	0	0	0	0	0
Maximum	100	100	100	70	100	100
Skewness	1.57	3.05	0.028	3.94	1.81	3.64

(1,0) Indique se a afirmação seguinte é verdadeira ou falsa. Justifique a sua resposta.

“A dispersão é maior na percentagem de reservas feitas através do *Booking* do que nas reservas feitas através do *Airbnb*”

(1,0) **5.** Abaixo apresentamos dois gráficos onde pode ver a distribuição das reservas feita diretamente no AL. Enquanto investigador/analista de dados, qual dos gráficos escolheria para efetuar a sua análise?



(1,5) **6.** Ainda no mesmo estudo, pediu-se aos proprietários dos alojamentos que indicassem como perspetivavam o futuro (a 5 anos) relativamente à evolução do preço (médio) dos alojamentos e da taxa de ocupação. As respostas foram dadas numa escala de 1 a 5, em que 1= vai piorar muito e 5= vai melhorar muito.

Abaixo pode ver o *output* produzido com o objetivo de analisar a relação entre estes dois aspetos. Indique, justificando, qual dos coeficientes de correlação escolheria para fazer a análise e interprete o resultado obtido.

```
jmv::corrMatrix(
  data = data,
  vars = vars(Taxa_Ocupação, Preço_médio),
  spearman = TRUE,
  sig = FALSE)
```


Correlation Matrix
#

		Taxa_Ocupação	Preço_médio
Taxa_Ocupação	Pearson's r	–	
	Spearman's rho	–	
Preço_médio	Pearson's r	0.523	–
	Spearman's rho	0.484	–

#

(1,5) **7.** Uma equipa de analistas de dados precisa de apresentar e analisar o rendimento dos agregados familiares, por município, em Portugal. Após algumas discussões sobre o melhor indicador a utilizar, encontram-se indecisos entre dois indicadores:

1. rendimento médio dos agregados familiares por município;
2. rendimento mediano dos agregados familiares por município

Contribua para este debate, apresentando e fundamentando a sua opinião sobre qual o indicador a utilizar.

8. Num ficheiro Excel, numa folha chamada *Dados* existem 4 colunas/variáveis: 2 variáveis com os nomes Q1 e Q1.1 (variáveis numéricas que diferem apenas porque a Q1.1 tem valores em branco) e 2 variáveis com os nomes Q1.texto e Q1.1.texto, variáveis convertidas com base em funções de transformação das variáveis Q1 e Q1.1. Numa outra folha do ficheiro Excel, com o nome *Códigos*, existe o código de codificação correspondente, como se pode observar na imagem seguinte.

	A	B	C	D		A	B
1	Q1	Q1.1	Q1.texto	Q1.1.texto	1	Q1 - Interesse pela política	
2		1	Muito	99	2		1 Muito
3		2	2 Algum	Algum	3		2 Algum
4		3	3 Pouco	Pouco	4		3 Pouco
5		4	4 Nenhum	Nenhum	5		4 Nenhum

Dados

Códigos

Escreva o comando completo o comando executado em Excel que permitiu de uma forma automatizada:

(1,0) **8.1.** Construir a variável Q1.texto (coluna C):

(1,0) **8.2.** Construir a variável Q1.1.texto (coluna D):

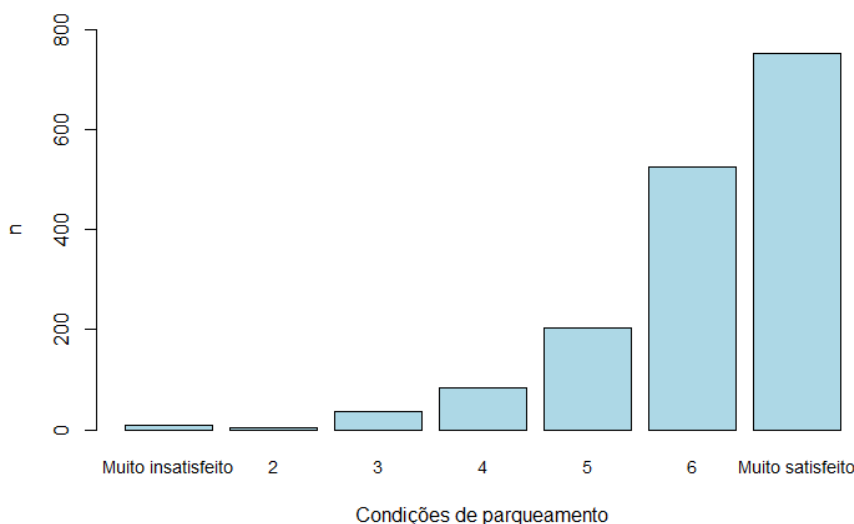
(1,5) **9.** Uma base de dados carregada em R (com o nome **bd**), contém uma variável chamada **Condições.de.parqueamento**. Esta variável foi construída com base em respostas a um questionário e contém as categorias listadas no output 1. Foi igualmente construindo o gráfico de barras correspondente (gráfico 1).

Tendo em conta a apresentação de categorias do output 1, escreva os comandos em linguagem R, que, permitiram construir o gráfico apresentado. (assuma que o objeto bd já se encontra carregado em memória).

Output 1 – Categorias únicas da variável *Condições.de.Pagamento*

<code>unique(bd\$Condições.de.parqueamento)</code>			
## [1]	"Muito satisfeito"	"6"	"4"
## [4]	"5"	"3"	"Muito insatisfeito"
## [7]	NA	"2"	

Gráfico 1 – Satisfação com as *Condições de Estacionamento*



10. Na **base de dados bd** apresentada na questão anterior, encontra-se também a variável **Voto**, constituída com respostas a um questionário e que reflete se os inquiridos votaram nas últimas eleições. Com base na seguinte tabela (t1):

Não	Não era elegível	Sim
359	46	1055

(1,0) **10.1** Apresente o comando necessário para construir também uma tabela com os valores percentuais

(1,5) **10.2** Apresente o comando necessário para construir uma *flextable* para a variável **Voto**, de forma a obter o resultado que se apresenta a seguir. Assuma que a biblioteca está importada, e que a cor de fundo utilizada tem o seguinte valor Hexadecimal: #3895D3.

Voto	n	Percentagem
Não	359	24.6
Não era elegível	46	3.2
Sim	1,055	72.3