

EDIT.

Módulo 3 – Sessão 2

EXPLORATORY
DATA ANALYSIS

TUTORA
Carla Cardoso
Freelancer AI Manager

30 de Janeiro 2025





1

ORGANIZAÇÃO DA SESSÃO


EDIT.

Trabalho 2 Sessão 1





Pausa para
jantar das
20:30 às 21:00



Quizz Final



Alternar entre
conteúdos e
experimentação



2



3

EXERCICIO EM EQUIPA

EDIT.

Ser-vos-á disponibilizado um conjunto de dados referentes a **características de carros**.

O objetivo deste desafio realizar um conjunto de análises ao conteúdo desse conjunto de dados

Para tal vocês devem:

- **Importar** os dados para excel
- **Descrever** as variáveis no que toca ao seu tipo
- Criar **tabelas de frequência** das variáveis **body-style** e **wheel-base** de acordo com o seu **tipo**
- Criar **representação gráfica** para as variáveis **body-style** e **wheel-base** de acordo com o seu **tipo**
- Criar **representação gráfica** para a relação entre as variáveis **length** e **height**

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 2

4

4


EXERCICIO EM EQUIPA

E

D

I

T.



O resultado do trabalho é um relatório, a ser entregue no final dos **30 minutos** que têm disponíveis para a realização deste exercício.

O relatório deve ter **EXATAMENTE 4 slides**, organizado da seguinte forma:

1. Apresentação dos **dados**
2. Tabelas de frequência : **body-style** e **wheel-base**
3. Representação gráfica : **body-style** e **wheel-base**
4. Representação gráfica : **length** vs. **height**

No final, será escolhido **1 grupo para apresentar** o seu relatório (10 minutos) e **1 grupo para comentar** os resultados apresentados (10 minutos).

A escolha dos destes 2 grupos será **aleatória**.

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 2

5

5


EXERCICIO EM EQUIPA #2

E


D

I


T.




Equipa 1




Equipa 2



Equipa 3



Equipa 4



Equipa 5

Slide 1	João	Andreia	Alexandre	Carolina L	Filipa
Slide 2	João	José M	Ana	José F	Filipa
Slide 3	José P M	Rui	Carolina M	Tamara	Joana
Slide 4	Sara Gomes	Nuno	Stefane	Susana	Yohanna

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 1

6

6

AGENDA

EDIT.



REVISÃO DE PYTHON



TRATAMENTO E PREPARAÇÃO



REPRESENTAÇÃO DE DADOS

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 2

7

7

AGENDA

EDIT.



REVISÃO DE PYTHON



TRATAMENTO E PREPARAÇÃO



REPRESENTAÇÃO DE DADOS

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 2

8

8

REVISÃO DE PYTHON: TIPOS DE DADOS

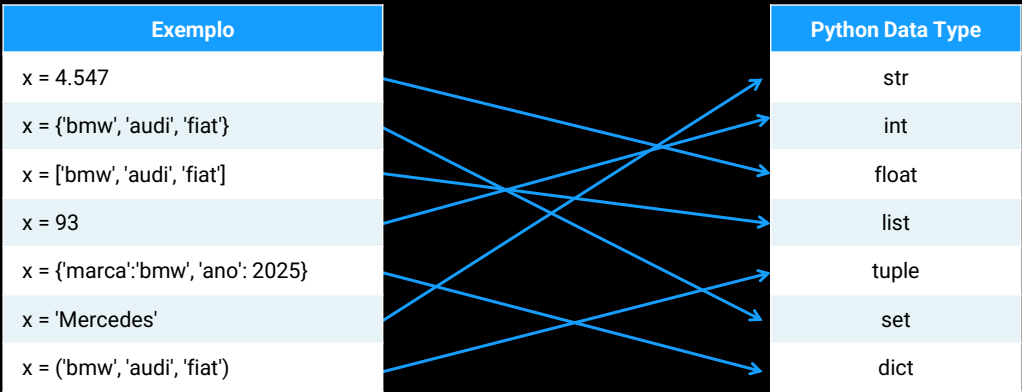
E D I T.

Tipo de Dado	Python Data Type	Exemplo
Texto	str	x = 'Hello World'
Números Inteiros	int	x = 20
Números Decimais	float	x = 20.5
Listas Mutável, ordenada	list	x = ['apple', 'banana', 'cherry']
Tuple Imutável, ordenada	tuple	x = ('apple', 'banana', 'cherry')
Conjunto Imutável, não ordenada	set	x = {'apple', 'banana', 'cherry'}
Dicionários Imutável, ordenada, sem duplicados	dict	x = {'name': 'John', 'age': 36}

Para saber o Data Type do objeto x usamos o comando: `print(type(x))`

REVISÃO DE PYTHON: TIPOS DE DADOS

E D I T.





O Pandas é uma biblioteca dedicada à **manipulação e análise de dados**.

Possui funções para analisar, limpar, explorar e manipular dados.

O nome é derivado do termo “panel data”, um termo econométrico.



<https://pandas.pydata.org/>

Comando	Ação
<code>import pandas as pd</code>	Importa a biblioteca Pandas e atribui-lhe a designação pd .
<code>pd.__version__</code>	Informa sobre a versão do Pandas que estamos a usar.

Atenção: O nome das variáveis, funções, etc. é case sensitive

Series



- Matriz (array) com 1 só dimensão
- Pode conter qualquer tipo e formato de dados (int, float, str...)
- É declarado com `pd.Series()`

DataFrame



- Matriz (array) com 2 dimensões (linhas e colunas)
- Cada coluna pode conter diferentes formatos de dados (int, float, str...)
- É declarado com `pd.DataFrame()`

Fonte: Modulo 1 - Data Science Fundamentals

Comando	Ação
<code>series = pd.Series(lista)</code>	Transforma uma lista numa Series da biblioteca Pandas
<code>dataframe = pd.DataFrame(dicionario)</code>	Transforma um dicionario numa DataFrame da biblioteca Pandas

Exemplo:

```
exercicio = {'calorias': [420,380,390],
             'tempo': [50,40,45]}

DF_exercicio = pd.DataFrame(exercicio)
```

Comando	Ação
<code>print(dataframe.loc[n])</code>	Devolve a linha n-1 da DataFrame . Para devolver mais que 1 linha, colocar n:m
<code>dataframe = pd.DataFrame(dicionario, index = ['a','b','c'])</code>	Transforma um dicionario numa DataFrame da biblioteca Pandas , mas agora adiciona também índices
<code>print(dataframe.loc['c'])</code>	Devolve a linha onde o index é igual a c da DataFrame

Nota: Também podemos usar o `iloc[]` no caso de não termos indexes, ou estes serem números inteiros.

Exemplo:

```
exercicio = {'calorias': [420,380,390],
             'tempo': [50,40,45]}

DF_exercicio = pd.DataFrame(exercicio, index = ['dia1','dia2','dia3'])

print(DF_exercicio.loc['dia3'])
```

REVISÃO DE PYTHON

EDIT.

BORA LÁ POR A MÃO NA MASSA

Inteligência Artificial: Gen AI & LLM

Módulo 3 – Exploratory Data Analysis – Sessão 2

2323

23

EXERCÍCIO I

EDIT.

Linha 1	Importar a biblioteca Pandas para o Python com o alias pd	João B.
Linha 2	Criar uma lista chamada a com os valores: Renault, Fiat e BMW	José P.M.
Linha 3	Imprimir a	Sara
Linha 4	Imprimir tipologia de a	Andreia
Linha 5	Criar uma lista chamada b com os valores: Twingo, 600 e X5	José M.
Linha 6	Imprimir b	Nuno
Linha 7	Imprimir tipologia de b	Rui
Linha 8	Criar um dicionário chamado dicionario com: marca → lista de valores presentes em a modelo → lista de valores presentes em b	Alexandre
Linha 9	Imprimir dicionario	Ana
Linha 10	Imprimir tipologia de dicionario	Carolina M.

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 2

2424

24

EXERCÍCIO I

EDIT.

Linha 11	Criar um DataFrame chamado <code>df_dicionario</code> com base no dicionário <code>dicionario</code>	Stéfane
Linha 12	Imprimir <code>df_dicionario</code>	Filipa
Linha 13	Imprimir tipologia de <code>df_dicionario</code>	Joana
Linha 14	Imprimir apenas 2ª linha de <code>df_dicionario</code>	Yhoanna

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 2

25

25

REVISÃO DE PYTHON: IMPORTAÇÃO / EXPORTAÇÃO

EDIT.

Comando	Ação
<code>path = 'C:/Users/...../Ficheiro.csv'</code>	Cria uma variável chamada path , do tipo <i>string</i> e é-lhe atribuído o valor 'C:/Users/...../Ficheiro.csv'
<code>Tabela = pd.read_csv(path)</code>	Importa os dados localizados no caminho guardado na variável path para uma dataframe chamada Tabela através do comando read_csv do pandas (pd).
<code>Tabela = pd.read_csv(path, header = None)</code>	Igual ao anterior, mas dá a indicação de que o ficheiro não tem cabeçalhos (header = None)
<code>Tabela.to_csv(path)</code>	Exporta os dados da Tabela para o ficheiro formato csv indicado na variável path . Caso já exista um ficheiro na pasta indicada no path, ele será sobreposto .
<code>Tabela</code>	Imprime no ecrã as primeiras e últimas 5 linhas da tabela

Atenção: Em vez de **csv** podemos ter outros formatos, como **json**, **excel** e **sql**


DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 2

26

26

Comando	Ação
<code>pd.set_option('display.max_columns', n)</code>	Limita a visualização a n colunas (sem limite None)
<code>pd.set_option('display.max_rows', n)</code>	Limita a visualização a n linhas (sem limite None)
<code>Tabela.head(n)</code>	Imprime no ecrã as primeiras n linhas da tabela
<code>Tabela.tail(n)</code>	Imprime no ecrã as últimas n linhas da tabela
<code>print(Data.to_string())</code>	Imprime no ecrã toda a tabela
<code>Colunas = ['coluna1','coluna2',....]</code>	Cria uma variável chamada Colunas do tipo <i>list</i> e atribui-lhe os valores ['coluna1','coluna2',....]
<code>Tabela.columns = Colunas</code>	Muda o nome das colunas da Tabela para os valores guardados na lista Colunas
<code>print(Tabela.info())</code>	Imprime informação sobre os conteúdo da Tabela

BORA LÁ POR A MÃO NA MASSA 

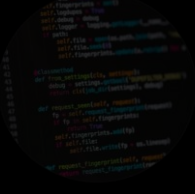


EXERCÍCIO II		EDIT.
Linha 1	Criar uma variável chamada <code>path</code> com o caminho onde guardaram ficheiro <code>data</code>	Carolina L.
Linha 2	Importar os dados do ficheiro <code>data</code> para uma <code>DataFrame</code> chamada <code>Dados</code>	José F.
Linha 3	Imprimir conteúdo <u>total</u> de <code>Dados</code>	Tamara
Linha 4	Imprimir tipologia da <code>Dados</code>	Susana
Linha 5	Imprimir apenas os 3 primeiros elementos da <code>Dados</code>	João B.
Linha 6	Imprimir apenas os 4 últimos elementos da <code>Dados</code>	José P.M
Linha 7	Imprimir informação sobre o conteúdo de <code>Dados</code>	Sara

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 2

29

29

AGENDA		EDIT.
 REVISÃO DE PYTHON	 TRATAMENTO E PREPARAÇÃO	 REPRESENTAÇÃO DE DADOS

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 2

30

30

PREPARAÇÃO DE DADOS: MISSÃO

EDIT.

O pré-processamento ou limpeza de dados, é um dos processos mais importantes e **um dos mais morosos** dentro de um processo de **Data Science**, e talvez também os dos menos agradáveis 😞

É este processo que garante que temos ingredientes de qualidade e no formato adequado para treinar os nossos modelos!

Este processo inclui identificação, transformação e processamento de:

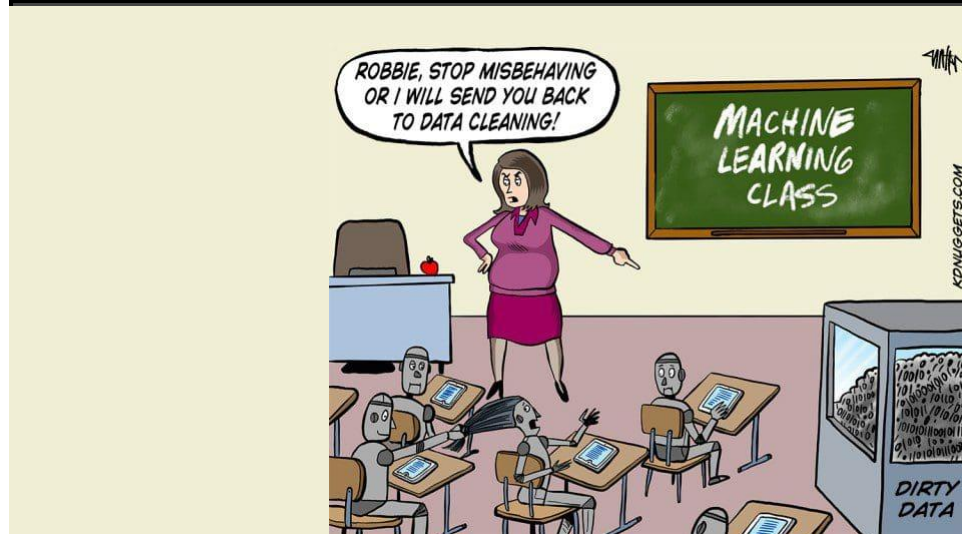
- ✓ Duplicados
- ✓ Formatos e Normalização
- ✓ *Missing Values*
- ✓ Escalas
- ✓ *Outliers*



31

PREPARAÇÃO DE DADOS: A TAREFA INDESEJADA

EDIT.



32

PREPARAÇÃO DE DADOS: MOTIVAÇÃO



DATA SCIENCE & BUSINESS ANALYTICS

Imaginem os seguintes cenários:

- ❖ Um estudo, após analisar o salário de todos os seus trabalhadores, revela que o salário médio na nossa empresa é de 5,000€ mensais
- ❖ O resultado dos inquéritos de satisfação dos alunos da EDIT que terminaram o curso DSBA é de 100%
- ❖ Verificamos que cerca de 20% das nossas encomendas não estão a chegar ao seu destino
- ❖ O nosso modelo baseado na idade, género e salário anual, não está a dar bons resultados a prever preferências musicais

Módulo 3 – Exploratory Data Analysis – Sessão 2

33

33

PREPARAÇÃO DE DADOS: MOTIVAÇÃO



DATA SCIENCE & BUSINESS ANALYTICS

Imaginem os seguintes cenários:

- ❖ Um estudo, após analisar o salário de **todos os seus trabalhadores**, revela que o salário médio na nossa empresa é de 5,000€ mensais
- ❖ O resultado dos inquéritos de satisfação dos alunos da EDIT que terminaram o curso DSBA é de 100%
- ❖ Verificamos que cerca de 20% das nossas encomendas não estão a chegar ao seu destino
- ❖ O nosso modelo baseado na idade, género e salário anual, não está a dar bons resultados a prever preferências musicais

Módulo 3 – Exploratory Data Analysis – Sessão 2

34

34

PREPARAÇÃO DE DADOS: MOTIVAÇÃO



DATA SCIENCE & BUSINESS ANALYTICS

Imaginem os seguintes cenários:

- ❖ Um estudo, após analisar o salário de todos os seus trabalhadores, revela que o salário médio na nossa empresa é de 5,000€ mensais
- ❖ O resultado dos inquéritos de satisfação dos alunos da EDIT que **terminaram** o curso DSBA é de 99%
- ❖ Verificamos que cerca de 20% das nossas encomendas não estão a chegar ao seu destino
- ❖ O nosso modelo baseado na idade, género e salário anual, não está a dar bons resultados a prever preferências musicais

Módulo 3 – Exploratory Data Analysis – Sessão 2

35

35

PREPARAÇÃO DE DADOS: MOTIVAÇÃO



DATA SCIENCE & BUSINESS ANALYTICS

Imaginem os seguintes cenários:

- ❖ Um estudo, após analisar o salário de todos os seus trabalhadores, revela que o salário médio na nossa empresa é de 5,000€ mensais
- ❖ O resultado dos inquéritos de satisfação dos alunos da EDIT que terminaram o curso DSBA é de 99%
- ❖ Verificamos que cerca de 20% das nossas encomendas não estão a chegar ao seu **destino**
- ❖ O nosso modelo baseado na idade, género e salário anual, não está a dar bons resultados a prever preferências musicais

Módulo 3 – Exploratory Data Analysis – Sessão 2

36

36

PREPARAÇÃO DE DADOS: MOTIVAÇÃO

EDIT.



Imaginem os seguintes cenários:

- ❖ Um estudo, após analisar o salário de todos os seus trabalhadores, revela que o salário médio na nossa empresa é de 5,000€ mensais
- ❖ O resultado dos inquéritos de satisfação dos alunos da EDIT que terminaram o curso DSBA é de 99%
- ❖ Verificamos que cerca de 20% das nossas encomendas não estão a chegar ao seu destino
- ❖ O nosso modelo baseado na idade, género e **salário anual**, não está a dar bons resultados a prever preferências musicais

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 2

37

37

PREPARAÇÃO DE DADOS

EDIT.



DUPLICADOS

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 2

38

38

DUPLICADOS: PROBLEMA

E D I T.

Tomando o ficheiro do exercício 1 da aula anterior como referência, imaginem que se deparam com estes dados para 2 das suas colunas: →

Há algum problema que vos salte à vista?

tran_id	tran_date
235448	28/02/2014
235449	27/02/2014
235450	24/02/2014
235450	24/02/2014
235451	23/02/2014
235452	23/02/2014
235453	22/02/2014
235453	22/02/2014
235454	22/02/2014
235455	2014/02/21

DUPLICADOS: PROBLEMA

E D I T.

Tomando o ficheiro do exercício 1 da aula anterior como referência, imaginem que se deparam com estes dados para 2 das suas colunas: →

Há algum problema que vos salte à vista?

Verificam-se **2 linhas repetidas**. Esta situação induz erros nas análises e estudos que fazamos sobre estes dados.

Há que **eliminar** estas situações.

tran_id	tran_date
235448	28/02/2014
235449	27/02/2014
235450	24/02/2014
235450	24/02/2014
235451	23/02/2014
235452	23/02/2014
235453	22/02/2014
235453	22/02/2014
235454	22/02/2014
235455	2014/02/21



Comando	Ação
Tabela.duplicated().any()	Devolve um valor True ou False caso hajam ou não linhas duplicadas na Tabela
print(Tabela.duplicated())	Devolve, para cada uma das linhas, um valor booleano indicando se a linha é (True), o não (False), um duplicado
Tabela.drop_duplicates(inplace = True)	Elimina as linhas duplicadas. O parâmetro inplace garante que as alterações são feitas na DataFrame em questão, e não é criado nova DataFrame.



FORMATOS E NORMALIZAÇÃO

FORMATOS E NORMALIZAÇÃO: PROBLEMA

E D I T.

Tomando o ficheiro do exercício 1 da aula anterior como referência, imaginem que se deparam com estes dados para 2 das suas colunas: →

Há algum problema que vos salte à vista?

tran_date	Store_type
28/02/2014	E-Shop
27/02/2014	e-Shop
24/02/2014	Teleshop
24-02-2014	e-Shop
23/02/2014	teleshop
23.02.2014	TeleShop
22/02/2014	e-Shop
22/02/2014	Mbr
22.02.2014	MBR
2014/02/21	e-Shop

FORMATOS E NORMALIZAÇÃO: PROBLEMA

E D I T.

Tomando o ficheiro do exercício 1 da aula anterior como referência, imaginem que se deparam com estes dados para 2 das suas colunas: →

Há algum problema que vos salte à vista?

No que toca aos valores presentes na coluna **tran_date**, há claramente uma diversidade de **formatos**, o que levará a **incorretas interpretações** de alguns dos valores

Esta multiplicidade de formatos é mais comum em colunas do tipo **data**, valores numéricos com diferentes **marcadores de milhares**, valores **monetários**, etc.

tran_date	Store_type
28/02/2014	E-Shop
27/02/2014	e-Shop
24/02/2014	Teleshop
24-02-2014	e-Shop
23/02/2014	teleshop
23.02.2014	TeleShop
22/02/2014	e-Shop
22/02/2014	Mbr
22/02/2014	MBR
2014/02/21	e-Shop

FORMATOS E NORMALIZAÇÃO: PROBLEMA

E D I T.

Tomando o ficheiro do exercício 1 da aula anterior como referência, imaginem que se deparam com estes dados para 2 das suas colunas:

Há algum problema que vos salte à vista?

No caso do campo Store_type temos o tipo de loja escrito de diferentes formas. Isto pode levar análises e modelos a considerar como tendo **diferentes tipos lojas da mesma tipologia incorretamente**. Estes valores precisam de ser **normalizados**.

Os temas de normalização são muito comuns em **campos de escrita livre**, como nomes, moradas, profissão, etc. Devem ser implementadas **regras de normalização** para não gerar resultados incorretos quando utilizados os dados.

tran_date	Store_type
28/02/2014	E-Shop
27/02/2014	e-Shop
24/02/2014	Teleshop
24-02-2014	e-Shop
23/02/2014	teleshop
23.02.2014	TeleShop
22/02/2014	e-Shop
22/02/2014	Mbr
22/02/2014	MBR
2014/02/21	e-Shop

FORMATOS E NORMALIZAÇÃO: COMANDOS ÚTEIS EM PYTHON

E D I T.

Comando	Ação
Tabela['var'] = Tabela['var'].str.replace('a','b')	Substitui os caracteres a por b na variável var da Tabela
Tabela.loc[n, 'var'] = m	Atribui o valor m campo var presente na linha n
Tabela['var'] = Tabela['var'].replace(Dicionario) ex.: Dicionario = {'Ctt': 'CTT', 'email': 'e-mail'}	Substitui os valores da coluna var da Tabela usando o dicionário
Tabela['var'] = Tabela['var'].str.upper()	Converte os valores da coluna var tudo para maiúsculas
Tabela['var'] = Tabela['var'].str.lower()	Converte os valores da coluna var tudo para minúsculas

FORMATOS E NORMALIZAÇÃO: COMANDOS ÚTEIS EM PYTHON

EDIT.

Comando	Ação
<code>Tabela.dtypes</code>	Indica a tipologia de cada coluna da Tabela
<code>Tabela[['var1', 'var2']] = Tabela[['var1', 'var2']].astype('float')</code>	Converte a tipologia das colunas var1 e var2 para float
<code>Tabela['var'] = pd.to_datetime(Tabela['var'])</code>	Converte a coluna var para tipo datetime

DUPLICADOS, FORMATOS E NORMALIZAÇÃO

EDIT.

BORA LÁ POR A MÃO NA MASSA 

EXERCÍCIO III		EDIT.
Linha 1	Criar uma variável chamada path com o caminho onde guardaram ficheiro data2	Joana
Linha 2	Importar os dados do ficheiro data2 para uma DataFrame chamada Dados2	Yhoanna
Linha 3	Explorar conteúdo da tabela	Stéfane
Linha 4	Verificar tipologia dos campos	Carolina M.
Linha 5	Verificar se existem duplicados	Ana
Linha 6	Remover linhas duplicadas	Alexandre
Linha 7	Normalizar coluna tran_date para o formato DD/MM/AAAA	Rui
Linha 8	Converter coluna tran_date para o tipologia datetime	Nuno
Linha 9	Criar dicionário chamado dicio para usar na normalização da coluna Store_type	Susana
Linha 10	Normalizar coluna Store_type usando o dicio	Andreia
DATA SCIENCE & BUSINESS ANALYTICS		Módulo 3 – Exploratory Data Analysis – Sessão 2
		49

TRATAMENTO E PREPARAÇÃO – PARTE 1

EDIT.

Questionário

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 2

50