

EDIT.

Módulo 3 – Sessão 4

EXPLORATORY  
DATA ANALYSIS

TUTORA

Carla Cardoso

Freelancer AI Manager


5 de Fevereiro 2025




1

AGENDA


EDIT.



REVISÃO DE  
PYTHON



TRATAMENTO E  
PREPARAÇÃO



REPRESENTAÇÃO DE  
DADOS

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 4

2

2

PREPARAÇÃO DE DADOS

EDIT.

OUTLIERS

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 4

3

3

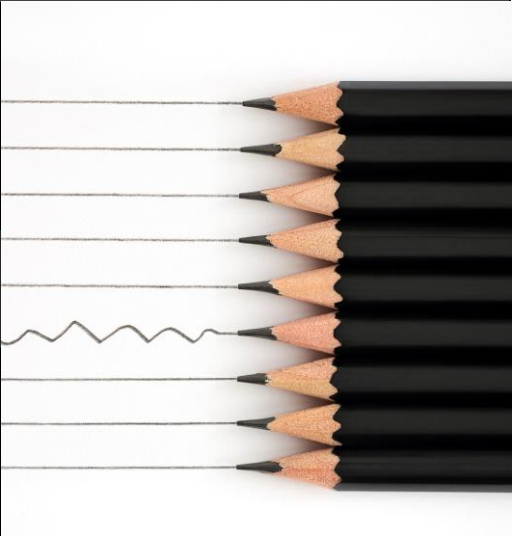
OUTLIERS: PROBLEMA

EDIT.

Os **outliers**, ou valores anómalos, são **valores atípicos** de uma variável, ou seja, são valores que estão fora do padrão comportamental da variável.

❖ Um estudo, após analisar o salário de **todos os seus trabalhadores**, revela que o salário médio na nossa empresa é de 5,000€ mensais.

Na sessão 2 referimos que o seguinte exemplo poderia ser impactado pela presença de **outliers**, pois inclui salários de elementos da administração, que tipicamente são atípicos.



DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 4

4

4

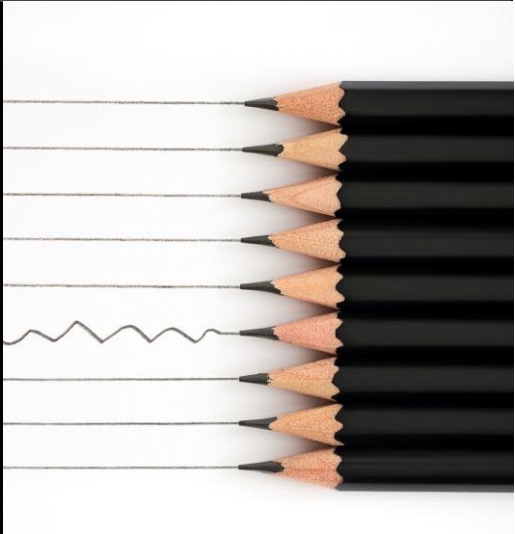
OUTLIERS: PROBLEMA

E D I T.

Estas observações podem **impactar os resultados** dos estudos feitos sobre eles, mas **não devem ser excluídos** nem **"manipulados"** sem conhecer a sua causa e o seu papel no estudo, visto que podem ser de elevada relevância.

Alguns exemplos:

- ❖ No desenvolvimento de um modelo de **deteção de Fraude**, verificamos que variável "Total transações por minuto" tem alguns valores muito elevados
- ❖ Num estudo do **tempo de espera** nos centro de saúde do país, verificamos que há um centro de saúde onde, num certo dia, existem tempos de espera **superiores a 24h**
- ❖ Na análise do **consumo médio** por família de determinado produto, encontramos alguns **valores muito elevados** na variável "Valor total das compras"



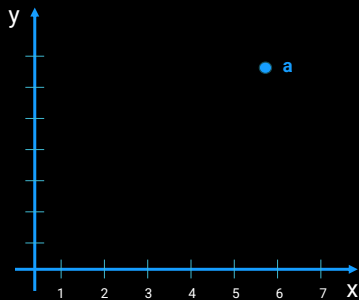
5

OUTLIERS: COMO IDENTIFICAR?

E D I T.

Devemos começar por **compreender as variáveis** que estamos a analisar para perceber o seu **comportamento esperado**, qual o seu **papel** no estudo e verificar se têm **outliers**.

Exemplo:



Acham que o ponto **a** é um **outlier**?

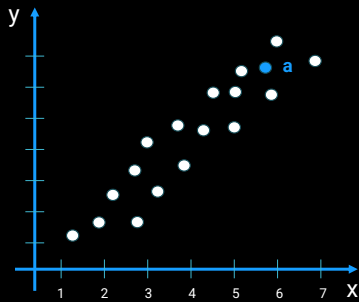
6

OUTLIERS: COMO IDENTIFICAR?

E D I T.

Devemos começar por **compreender as variáveis** que estamos a analisar para perceber o seu **comportamento esperado**, qual o seu **papel** no estudo e verificar se têm **outliers**.

Exemplo:



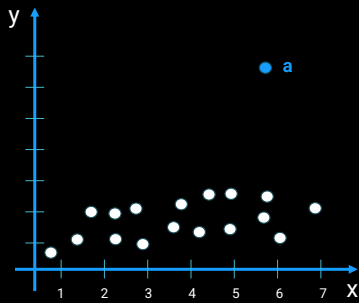
E agora? Será **a** um **outlier**?

OUTLIERS: COMO IDENTIFICAR?

E D I T.

Devemos começar por **compreender as variáveis** que estamos a analisar para perceber o seu **comportamento esperado**, qual o seu **papel** no estudo e verificar se têm **outliers**.

Exemplo:



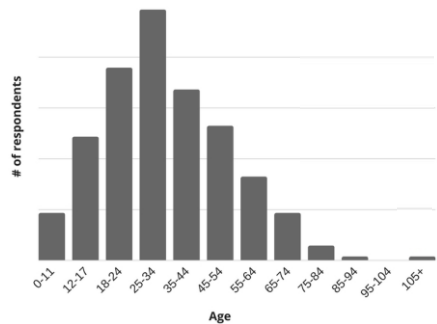
E agora? Será **a** um **outlier**?

OUTLIERS: COMO IDENTIFICAR?

EDIT.

Por forma a “estimar” o que é o **comportamento esperado** de uma variável é útil recorrer a **indicadores estatísticos** e/ou **representações gráficas** da sua distribuição.

Vamos imaginar a variável **idade** com a seguinte **distribuição**:

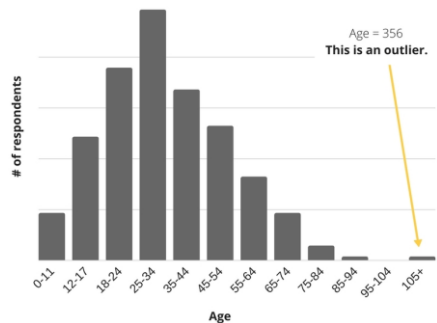


OUTLIERS: COMO IDENTIFICAR?

EDIT.

Por forma a “estimar” o que é o **comportamento esperado** de uma variável é útil recorrer a **indicadores estatísticos** e/ou **representações gráficas** da sua distribuição.

Vamos imaginar a variável **idade** com a seguinte **distribuição**:

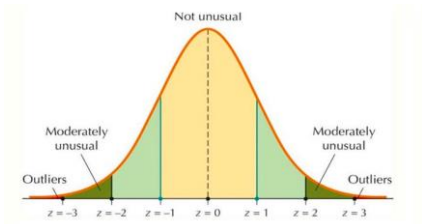


OUTLIERS: COMO IDENTIFICAR?

EDIT.

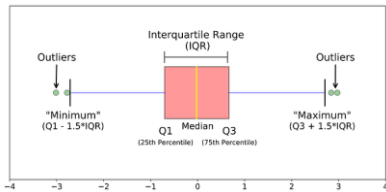
Dado que a análise gráfica pode **consumir muito tempo**, podemos recorrer ao uso de determinados **indicadores estatísticos**. Os métodos mais comuns para a classificação de um valor como **outliers**:

❖ Valores afastados da **média mais do que 3 desvios-padrão**



Indicado para variáveis com distribuições Gaussianas ("Normais")

❖ Valores **afastados mais que 1,5x IQR** (InterQuartile Range) do 1º e 3º Quartil



Não pressupõe nenhuma distribuições da variável

OUTLIERS: COMO RESOLVER?

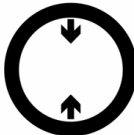
EDIT.



Ignorar  
(manter os  
valores)



Eliminar as  
observações



Restringir os  
valores



Aplicar  
transformações

Tal como no caso dos **missing values**, qualquer que seja a escolha de tratamento de **outliers**, haverá **impactos** na sua implementação, e por isso é importante fazê-la tendo em conta **vantagens** e **desvantagens**.

OUTLIERS: IGNORAR

E D I T.

Indicado para situações quando o **valores atípicos têm um papel relevante** no trabalho a realizar, como por exemplo no caso de modelos de deteção de Fraude.



Ignorar  
(manter como está)



- **Sem esforço adicional**
- Os dados originais são mantidos, **sem manipulação** da nossa parte



- **Impactos em métricas** usadas no desenvolvimento do nosso trabalho, podendo impactando consequentemente nos **resultados**

OUTLIERS: DESCARTAR OBSERVAÇÕES

E D I T.

Indicado quando os valores atípicos são **raros** e são derivados de **erros**, não havendo uma perda de informação significativa.



Eliminar as  
observações



- Processo **simples e rápido**
- Os dados originais são mantidos, **sem manipulação** da nossa parte
- **Não haverá impactos** nas métricas e técnicas que iremos usar



- Há uma **perda de informação**
- **Conclusões** alcançadas podem **não ser representativas** dos dados como um todo

## OUTLIERS: SUBSTITUIR VALORES

EDIT.

Indicado quando existem **várias variáveis** com valores atípicos e esses valores **não são críticos** para o estudo em questão.



Restringir os valores



- **Não há perda de observações**
- Os dados são truncados, procurando o **menor impacto possível**,
- **Não haverá impactos** nas métricas e técnicas que iremos usar



- **Complexidade** do processo de restrição
- Há manipulação dos dados, podendo haver **distorção** dos resultados
- Pode ter impactos em termos de **tempo**

15

## OUTLIERS: APLICAÇÃO DE TRANSFORMAÇÕES

EDIT.

Indicado quando têm **distribuições demasiado concentradas** ou **demasiado dispersas** ou com **pouca relevância** para o trabalho em questão, podendo a transformação levar a corrigir *outliers* e aumentar a sua relevância.



Aplicar transformações



- **Não há perda de observações**
- Os dados são transformados de forma não-linear, por forma a eliminar *outliers*, o pode **potenciar a relevância** da informação para o estudo
- **Não haverá impactos** nas métricas e técnicas que iremos usar



- **Complexidade** do processo de transformação
- Há manipulação dos dados, podendo haver **distorção** dos resultados e **piorar os resultados**
- Pode ter impactos elevados em termos de **tempo**

16



OUTLIERS: COMO ESCOLHER A TRANSFORMAÇÃO ADEQUADA?



Aplicar transformações

Nem sempre é fácil decidir que transformação de dados é a melhor, a melhor solução muitas vezes passa por **testar o impacto de diferentes** transformações e escolher a que permite atingir **melhores resultados**.

Em termos gerais, temos seguintes impactos:

- ❖ **Logaritmos e raízes quadradas** – retirar “caudas” da distribuição à direita
- ❖ **Raízes cúbicas** – retirar “caudas” da distribuição à direita e esquerda
- ❖ **Decis / Percentis**: Útil para categorizar dados e reduzir a influência de valores extremos
- ❖ **Ranking**: Útil para modelos não-paramétricos (árvores de decisão) ou quando apenas a ordem relativa importa

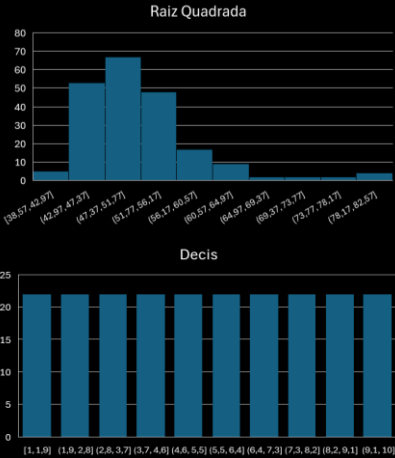
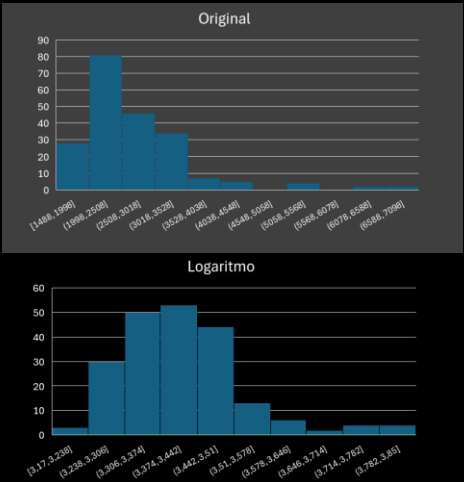
OUTLIERS: COMO ESCOLHER A TRANSFORMAÇÃO ADEQUADA?



Exemplo:  
**curb-weight**



Aplicar transformações



OUTLIERS: COMO ESCOLHER A TRANSFORMAÇÃO ADEQUADA?

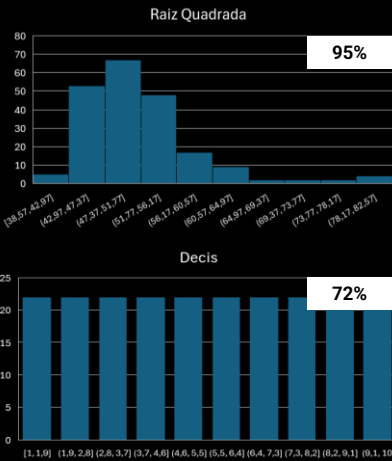
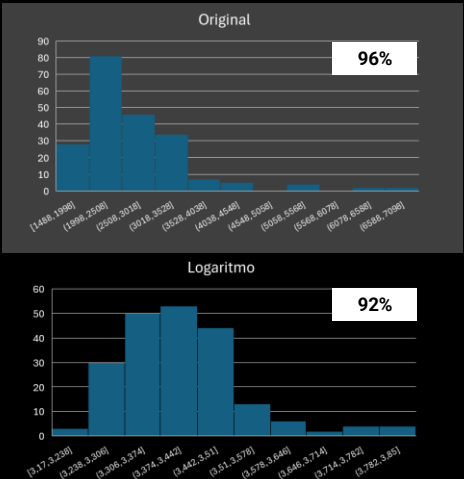
EDIT.

Exemplo: **curb-weight**



Aplicar transformações

☐ Correlação com variável Price (target)



OUTLIERS: COMO ESCOLHER A TRANSFORMAÇÃO ADEQUADA?

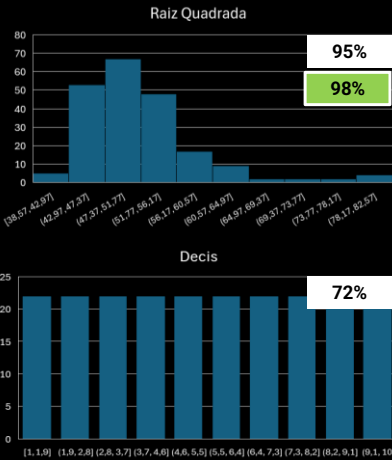
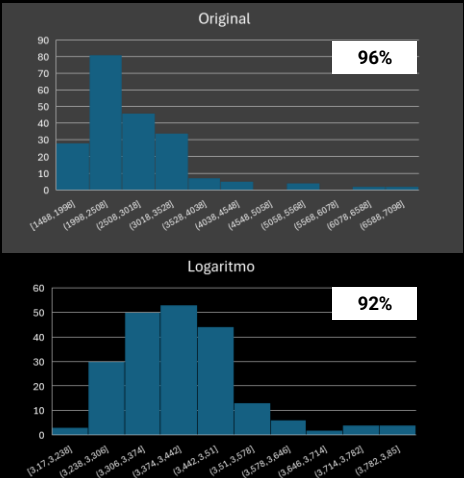
EDIT.

Exemplo: **curb-weight**



Aplicar transformações

☐ Correlação com variável Price (target)  
☒ Correlação com variável sqrt Price



OUTLIERS: EXEMPLOS

E D I T.

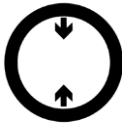


Ignorar  
(manter como  
está)

No estudo sobre as **alterações climáticas** são observados **valores fora do padrão** para a amplitude térmica no último ano. Esses valores representam **10%** das observações.



Eliminar as  
observações



Restringir os  
valores



Aplicar  
transformações

OUTLIERS: EXEMPLOS

E D I T.



Ignorar  
(manter como  
está)

No estudo sobre as **alterações climáticas** são observados **valores fora do padrão** para a amplitude térmica no último ano. Esses valores representam **10%** das observações.

OUTLIERS: EXEMPLOS

E D I T.

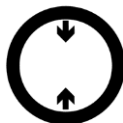


Ignorar  
(manter como  
está)

No desenvolvimento de um **modelo preditivo do sucesso escolar** foi considerada informação sobre **tempo de deslocação à escola** e verificou-se que **0.2%** dos alunos apresentavam valores atípicos.



Eliminar as  
observações



Restringir os  
valores



Aplicar  
transformações

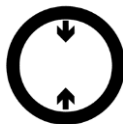
OUTLIERS: EXEMPLOS

E D I T.

No desenvolvimento de um **modelo preditivo do sucesso escolar** foi considerada informação sobre **tempo de deslocação à escola** e verificou-se que **0.2%** dos alunos apresentavam valores atípicos.



Aplicar  
transformações



Restringir os  
valores

OUTLIERS: EXEMPLOS

E D I T.



Ignorar  
(manter como  
está)

Num processo de **controle de qualidade**, verificamos que os **dados recolhidos** de um dos **sensores** a partir de determinado momento tinham comportamentos atípicos, posteriormente associados a uma **avaría**.



Aplicar  
transformações



Eliminar as  
observações



Restringir os  
valores

OUTLIERS: EXEMPLOS

E D I T.

Num processo de **controle de qualidade**, verificamos que os **dados recolhidos** de um dos **sensores** a partir de determinado momento tinham comportamentos atípicos, posteriormente associados a uma **avaría**.



Eliminar as  
observações

OUTLIERS: COMANDOS ÚTEIS EM PYTHON



E D I T.

Comando	Ação
Descartar Observações	
<code>q1 = Tabela['var'].quantile(0.25)</code>	Cria uma variável nova chamada <b>q1</b> que tem o valor do <b>1º quartil</b> (↔ quantil 25%) da variável <b>var</b> .
<code>q3 = Tabela['var'].quantile(0.75)</code>	Cria uma variável nova chamada <b>q3</b> que tem o valor do <b>3º quartil</b> (↔ quantil 75%) da variável <b>var</b> .
<code>IQR = q3 - q1</code>	Cria uma variável nova chamada <b>IQR</b> que representa a <b>distância interquartil</b> da variável <b>var</b> acima referenciada
<code>limite_inf = q1 - 1.5 * IQR</code> <code>limite_sup = q3 + 1.5 * IQR</code>	Definição dos <b>limites</b> inferiores e superiores
<code>Tabela_Nova = Tabela[(Tabela['var'] &gt;= limite_inf) &amp; (Tabela['var'] &lt;= limite_sup)]</code>	Cria uma nova dataframe chamada <b>Tabela_Nova</b> onde apenas as observações que, para a variável <b>var</b> , apresentam valores entre <b>limite_inf</b> e <b>limite_sup</b> . Esta nova dataframe <b>deixa de ter outliers na variável var</b> com base na distância <b>interquartil</b> .


OUTLIERS: COMANDOS ÚTEIS EM PYTHON



E D I T.

Comando	Ação
Restringir Valores	
<code>Conta = ((Tabela['var'] &gt;= limite_inf) &amp; (Tabela['var'] &lt;= limite_sup)).sum()</code>	Calcula <b>quantas</b> observações da coluna <b>var</b> tem valores no <b>intervalo</b> definido pelos limites <b>inferior</b> e <b>superior</b>
<code>Tabela_Nova = Tabela[(Tabela['var']).clip(lower=limite_inf, upper=limite_sup)]</code>	Cria uma nova dataframe chamada <b>Tabela_Nova</b> onde as observações que, para a variável <b>var</b> , apresentavam valores inferiores a <b>limite_inf</b> , <b>passam a ter o valor limite_inf</b> , e as observações com valores superiores a <b>limite_sup</b> <b>passam a ter o valor limite_sup</b> .  Esta nova dataframe <b>deixa de ter outliers na variável var sem perda de observações</b> .

OUTLIERS: COMANDOS ÚTEIS EM PYTHON



EDIT.

Comando	Ação
Aplicar Transformações	
Tabela[ <b>var_log</b> ] = np.log1p(Tabela[ <b>var</b> ])	Cria nova variável como <b>Logaritmo</b> da variável <b>var</b>
Tabela[ <b>var_sqrt</b> ] = np.sqrt(Tabela[ <b>var</b> ])	Cria nova variável como <b>Raiz Quadrada</b> da variável <b>var</b>
Tabela[ <b>var_decile</b> ] = pd.qcut(Tabela[ <b>var</b> ], <b>q=10</b> , labels = False) + 1	Criação de nova variável <b>var_decile</b> que corresponde ao valor de <b>var</b> quando agrupada em <b>10 decis</b> .
Visualizar Distribuições	
Tabela[ <b>var</b> ].hist(bins= <b>10</b> , edgecolor='black')	Cria <b>histograma</b> com <b>10 colunas</b> da variável <b>var</b>
Calcular correlações lineares	
Corr = (Tabela[ <b>var2</b> ].corr(Tabela[ <b>var1</b> ]))	Calcula a <b>correlação linear</b> entre a <b>var</b> e a <b>var1</b>

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 4

29

29

OUTLIERS



EDIT.

BORA LÁ POR A MÃO NA MASSA



Inteligência Artificial: Gen AI & LLM

Módulo 3 – Exploratory Data Analysis – Sessão 4

30

30

EXERCÍCIO I		EDIT.
Linha 1	Importar os dados do ficheiro <b>Carros4</b> para uma DataFrame chamada <b>Carros4</b> e <b>apagar</b> colunas criadas via <b>transformações</b> das variáveis originais.	Susana
Linha 2	Analisar as variáveis quantitativas e calcula o <b>número de outliers</b> , segundo o <b>IQR</b> , que cada uma tem.	João B
Linha 3	Considerar a variável quantitativa com <b>menor número de outliers</b> segundo os cálculos do passo anterior e <b>eliminar</b> os registos classificados como <b>outliers</b> .	José P M
Linha 4	Verificar que <b>já não existem outliers</b> , segundo o <b>IQR</b> , na variável em questão.	Tamara
Linha 5	Analisar novamente as variáveis quantitativas e, calcular o <b>número de outliers</b> , agora segundo o método de <b>Estandardização</b> , que cada uma tem.	Nuno
Linha 6	Considerar as <b>duas</b> variáveis quantitativas com <b>menor número de outliers</b> segundo os cálculos do passo anterior, e <b>criar uma nova variável</b> com base nessas, como o sufixo <b>"_lim"</b> , onde se restringe os seus valores com base nos limites inferior e superior calculados no método de <b>Estandardização</b> .	Gonçalo
Linha 7	Validar em que <b>observações</b> as variáveis originais e as <b>"_lim"</b> tomam <b>valores diferentes</b>	Stéfane
DATA SCIENCE & BUSINESS ANALYTICS		Módulo 3 – Exploratory Data Analysis – Sessão 431

31

EXERCÍCIO I		EDIT.
Linha 8	Considerando as variáveis do passo anterior, original e a limitada, <b>comparar</b> os seus <b>histogramas</b> .	Yhoanna
Linha 9	Calcular a <b>correlação</b> entre as variáveis do passo 7, original e a limitada, e a variável <b>price</b> . Que podemos concluir?	Andreia
Linha 10	Considerando a variável quantitativa com <b>maior número de outliers</b> segundo os cálculos do passo 5, criar uma <b>nova variável</b> com base nessa, como o sufixo <b>"_decil"</b> , que corresponde ao valor da variável quando agrupada em 10 decis.	José M
Linha 11	Considerando as variáveis do passo anterior, original e a limitada, <b>compara</b> os seus <b>histogramas</b> .	Rui
Linha 12	Calcula a <b>correlação</b> entre as variáveis do passo 7, original e a limitada, e a variável <b>price</b> . Que podes concluir?	José F
Linha 13	Exportar os dados de <b>Carros4</b> num ficheiro csv com o nome <b>Carros5</b>	Alexandre
DATA SCIENCE & BUSINESS ANALYTICS		Módulo 3 – Exploratory Data Analysis – Sessão 432

32