

EDIT.

Módulo 3 – Sessão 3

EXPLORATORY
DATA ANALYSIS

TUTORA
Carla Cardoso
Freelancer AI Manager


4 de Fevereiro 2025



1


AVALIAÇÃO DO MÓDULO

EDIT.




Assiduidade e Pontualidade
15%

Presença nas aulas




Motivação e Participação
5%

Participação nas aulas




Domínio dos Conteúdos das Aulas
25%

Questionários



Aplicação Técnica
25%

Exercícios Python



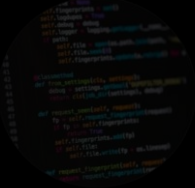
Trabalho Final
30%

Trabalhos


2

AGENDA


EDIT.



REVISÃO DE
PYTHON



TRATAMENTO E
PREPARAÇÃO



REPRESENTAÇÃO DE
DADOS

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 3

3

3

TRATAMENTO E PREPARAÇÃO

EDIT.

MISSING VALUES

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 3

4

4

MISSING VALUES: PROBLEMA

E D I T.

Os *missing values* são **valores em falta nos nossos dados** e por isso podem gerar problemas nas análises e modelos que os utilizam, mas devemos ter cuidado a lidar com eles, pois não devem ser tratados todos da mesma maneira.

A primeira coisa a perceber é se estamos perante um **valor em falta** ou um **valor que não existe**. Vejamos o seguinte exemplo:

transaction_id	cust_id	tran_date	prod_subcat_code	prod_cat_code	Qty	Rate	Tax	total_amt	Store_type	return_date
25963520987	274829	20/02/2014	4	4	3	502	158.13	1664.13	Flagship store	
17183929085	266863	20/02/2014	1	2	1	1359	142.695	1501.695		
44783317894	269452	20/02/2014	3	1	3	825	259.875	2734.875	TeleShop	21/02/2014

5

MISSING VALUES: PROBLEMA

E D I T.

Os *missing values* são **valores em falta nos nossos dados** e por isso podem gerar problemas nas análises e modelos que os utilizam, mas devemos ter cuidado a lidar com eles, pois não devem ser tratados todos da mesma maneira.

A primeira coisa a perceber é se estamos perante um **valor em falta** ou um **valor que não existe**. Vejamos o seguinte exemplo:

transaction_id	cust_id	tran_date	prod_subcat_code	prod_cat_code	Qty	Rate	Tax	total_amt	Store_type	return_date
25963520987	274829	20/02/2014	4	4	3	502	158.13	1664.13	Flagship store	
17183929085	266863	20/02/2014	1	2	1	1359	142.695	1501.695		
44783317894	269452	20/02/2014	3	1	3	825	259.875	2734.875	TeleShop	21/02/2014

Claramente o campo em vazio na **Store_type** é um **valor em falta**, dado que todas as transações são realizadas numa loja (mesmo que virtual), e todas as lojas têm um tipo. Mas no caso da **return_date** já **não é um valor em falta**, simplesmente algumas (a maioria) transações não foram devolvidas e por isso não existe data de devolução.

6

MISSING VALUES: PROBLEMA

EDIT.

Os valores em falta podem ser de diferentes tipos:

- Missing Completely At Random (**MCAR**)
- Missing At Random (**MAR**)
- Missing Not At Random (**MNAR**)

Estas tipologias estão relacionadas com as **causas** da ausência de informação, e por consequência, com o **impacto** que podem gerar nas conclusões a que pretendemos chegar.

Diferentes tipos de missing values devem ser **tratados de forma diferente**.

MISSING VALUES: MCAR

EDIT.

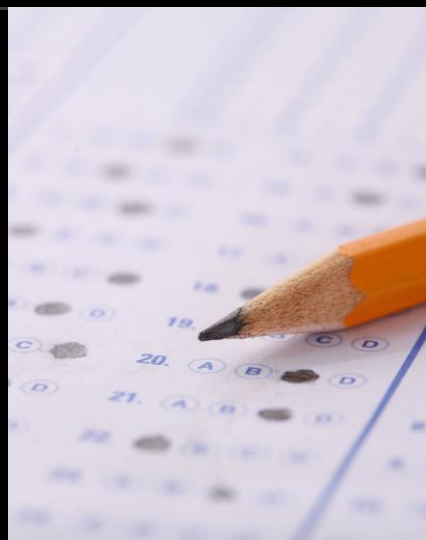
Nos MCAR, Missing Completely At Random, a **probabilidade de falta** de dados é a **mesma para todas as observações, não havendo relação** entre os dados em falta e quaisquer outros valores observados ou não observados no nosso conjunto de dados.

O valor em falta pode estar ausente devido a:

- erro humano
- falha do sistema/equipamento
- perda de amostra
- detalhes técnicos insatisfatórios durante o registro dos valores

Por exemplo, falhas de preenchimento num questionário.

Tratando-se de um **“erro” raro e não sistemático**, e o **impacto na análise estatística é reduzido**.



MISSING VALUES: MAR

EDIT.



MAR, Missing At Random, significa que apesar do motivo dos valores em falta ser **aleatório**, o valor em falta **pode ser estimado por variáveis sobre as quais existe informações**, existindo uma relação entre os dados em falta e os dados existentes.

Por exemplo, numa base de dados de livros numa biblioteca podem **faltar alguns dados** como a editora ou ano de publicação devido a páginas em falta ou ilegíveis, situações **mais comuns em livros muito antigos**. Ou seja, o ano de edição influencia a probabilidade de outros dados estarem em falta.

Portanto, e ainda que a probabilidade de falta de dados na variável editora estar correlacionada com o ano de publicação, **não é possível saber o valor em falta** referente à editora.

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 3

9

9

MISSING VALUES: MNAR

EDIT.

Significa que o facto da informação estar em falta, está **correlacionado com o valor real** dessa informação.

Estes casos são os **mais críticos**, pois **impactam significativamente** as conclusões de estudos ou modelos realizadas sobre esses dados.

Alguns exemplos:

- Falta da dados de pacientes que ficaram **demasiado doentes** para virem fazer o check-up
- Falha na **recolha de feedback** de alunos que não terminaram o curso por desagrado, visto não estarem presentes no momento da recolha de informação
- Falha na recolha de informação de precipitação dadas **condições extremas** de precipitação



DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 3

10

10

E agora....?



11



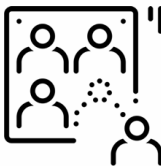
Ignorar
(manter como
está)



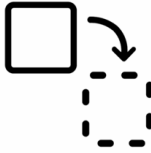
Recuperar a
informação



Descartar
variáveis



Descartar
observações



Substituir
valores

Qualquer que seja a escolha, haverá **impactos** na sua implementação, e por isso é importante fazê-la tendo em conta **vantagens** e **desvantagens**.

12

MISSING VALUES: IGNORAR

EDIT.

Indicado quando os casos são **raros**, são do tipo **MCAR** e as técnicas que vamos usar **conseguem lidar com missing values**



Ignorar
(manter como
está)



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



- **Limitação nas técnicas** possíveis de utilizar, pois algumas irão ignorar os dados que tenham *missing values*, impactando consequentemente os **resultados**

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 3

13

13

MISSING VALUES: RECUPERAR INFORMAÇÃO

EDIT.

Indicado quando os casos são do tipo **MNAR** ou em grande quantidade **inviabilizando** o estudo / análise.



Recuperar a
informação



- Os dados são corrigidos com a informação real, **não havendo distorção**
- **Não haverá limitação** nas técnicas possíveis de utilizar



- Pode ter impactos elevados em termos de **tempo e custo**

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 3

14

14

MISSING VALUES: DESCARTAR VARIÁVEIS

EDIT.

Descartar
variáveis

Indicado quando são variáveis com uma **percentagem significativa de missing values** ou quando as variáveis **não são muito relevantes** para o estudo.



- Processo **simples e rápido**
- Os dados originais são mantidos, **sem manipulação** da nossa parte
- **Não haverá limitação** nas técnicas possíveis de utilizar



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

15

MISSING VALUES: DESCARTAR OBSERVAÇÕES

EDIT.

Descartar
observações

Indicado quando os valores em falta são do tipo **MCAR** e o **número de observações** com **missing values não é significativo** para o estudo.



- Processo **simples e rápido**
- Os dados originais são mantidos, **sem manipulação** da nossa parte
- **Não haverá limitação** nas técnicas possíveis de utilizar



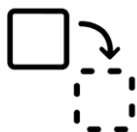
- Há uma **perda de informação**
- **Conclusões** alcançadas podem **não ser representativas** dos dados como um todo

16

MISSING VALUES: SUBSTITUIR VALORES

EDIT.

Indicado quando os casos são do tipo **MAR**.



Substituir
valores



- **Não há perda de informação**
- Os dados são corrigidos com o **menor impacto possível**, com informação estimada, tendo em conta a informação disponível
- **Não haverá limitação** nas técnicas possíveis de utilizar

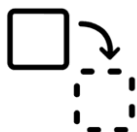


- **Complexidade** do processo de substituição
- Há manipulação dos dados, podendo haver **distorção** dos resultados
- Pode ter impactos elevados em termos de **tempo**

MISSING VALUES: SUBSTITUIR VALORES – COMO?

EDIT.

Dado que não sabemos o valor real em falta, **que valor devemos usar** para preencher os valores em falta?



Substituir
valores

Dependendo do **tipo de dado**, temos diferentes opções:

- **Imputação**: Preenchemos **todos** os valores em falta por um valor único que pode ser um **valor arbitrário**, como 0 ou 99999, ou pode ser o resultado de uma função dos dados, como a **média, mediana, moda**, etc.
- **Estimação**: Preenchemos **cada** valor em falta por um **valor estimado** em **função** da restante informação que temos ao nosso dispor sobre cada indivíduo. Ou seja, **criamos um modelo** para preencher a informação em falta.
- **Vizinho mais próximo**: Usamos a informação existente no **indivíduo mais próximo** para preencher o valor em falta.

MISSING VALUES: SUBSTITUIR VALORES – EXEMPLOS

EDIT.

Substituir valores

Data	Hora	Temperatura	Humidade	Precipitação	Vento
01/12/2024	9:00	10	60%	0.2	24 km/h
01/12/2024	10:00	12	65%	0.4	22 km/h
01/12/2024	11:00	NaN	67%	0.5	23 km/h
01/12/2024	12:00	14	85%	0.8	24 km/h
01/12/2024
01/12/2024	20:00	8	65%	0.5	25 km/h
01/12/2024	21:00	7	67%	0.2	30 km/h
01/12/2024	22:00	NaN	60%	0.0	40 km/h
01/12/2024	23:00	7	60%	0.0	42 km/h

Média

Mediana

Moda

Estimação

Vizinho

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 3

19

19

MISSING VALUES: SUBSTITUIR VALORES – EXEMPLOS

EDIT.

Substituir valores

Data	Hora	Temperatura	Humidade	Precipitação	Vento
01/12/2024	9:00	10	60%	0.2	24 km/h
01/12/2024	10:00	12	65%	0.4	22 km/h
01/12/2024	11:00	NaN	67%	0.5	23 km/h
01/12/2024	12:00	14	85%	0.8	24 km/h
01/12/2024
01/12/2024	20:00	8	65%	0.5	25 km/h
01/12/2024	21:00	7	67%	0.2	30 km/h
01/12/2024	22:00	NaN	60%	0.0	40 km/h
01/12/2024	23:00	7	60%	0.0	42 km/h

Vizinho

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 3

20

20

MISSING VALUES: SUBSTITUIR VALORES – EXEMPLOS

E

D

I

T.

Substituir valores

Aluno	Disciplina	Assiduidade	Nota
122775	Francês	90%	12
232571	Francês	85%	15
224166	Francês	NaN	17
288412	Francês	90%	16
...
258874	Português	100%	12
247887	Português	100%	15
122271	Português	NaN	14
233637	Português	95%	11

Média

Mediana

Moda

Estimação

Vizinho

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 3

21

MISSING VALUES: SUBSTITUIR VALORES – EXEMPLOS

E

D

I

T.

Substituir valores

Aluno	Disciplina	Assiduidade	Nota
122775	Francês	90%	12
232571	Francês	85%	15
224166	Francês	NaN	17
288412	Francês	90%	16
...
258874	Português	100%	12
247887	Português	100%	15
122271	Português	NaN	14
233637	Português	95%	11

Média

Mediana

Estimação

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 3

22


MISSING VALUES: EXEMPLOS

EDIT.

MCAR


MAR

MNAR




Ignorar
(manter como está)


Numa **análise descritiva** sobre a situação familiar dos alunos de um agrupamento de escolas, temos **10% de missing values** na variável **localidade**.




Substituir valores



Recuperar a informação



Descartar variáveis



Descartar observações

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 3


23

23

MISSING VALUES: EXEMPLOS


EDIT.

MCAR



Ignorar
(manter como está)

Numa **análise descritiva** sobre a situação familiar dos alunos de um agrupamento de escolas, temos **10% de missing values** na variável **localidade**.



Substituir valores

Imputação

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 3

24

24

MISSING VALUES: EXEMPLOS

EDIT.

MCAR

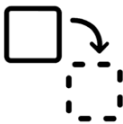
MAR

MNAR



Ignorar
(manter como
está)

Num estudo sobre o **impacto de um medicamento** sobre um grupo de utentes, temos em falta informação mais recente para um conjunto de utentes (~30%) que foram **hospitalizados** e passaram a ser seguidos por outra equipa.



Substituir
valores



Recuperar a
informação



Descartar
variáveis



Descartar
observações

MISSING VALUES: EXEMPLOS

EDIT.

MNAR

Num estudo sobre o **impacto de um medicamento** sobre um grupo de utentes, temos em falta informação mais recente para um conjunto de utentes (~30%) que foram **hospitalizados** e passaram a ser seguidos por outra equipa.



Recuperar a
informação

MISSING VALUES: EXEMPLOS

EDIT.

MCAR

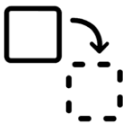
MAR

MNAR



Ignorar
(manter como
está)

No desenvolvimento de um estudo sobre o desemprego em Portugal existe falta de informação do total de inscritos em alguns centro de emprego para alguns meses, sobretudo em zonas do país com menor densidade populacional.



Substituir
valores



Recuperar a
informação



Descartar
variáveis



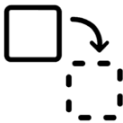
Descartar
observações

MISSING VALUES: EXEMPLOS

EDIT.

MAR

No desenvolvimento de um estudo sobre o desemprego em Portugal existe falta de informação do total de inscritos em alguns centro de emprego para alguns meses, sobretudo em zonas do país com menor densidade populacional.



Substituir
valores

Vizinho

DUPLICADOS: COMANDOS ÚTEIS EM PYTHON



EDIT.

Comando	Ação
Import numpy as np	A biblioteca de numpy é muito útil quando usada em conjunto com o pandas e facilita cálculos matemáticos
Tabela. describe()	Devolve algumas métricas para cada variável , número de observações não vazias, min., max., média...
Identificar <i>missing values</i>	
Tabela = Tabela. sort_values (by='var')	Ordena a Tabela em função da coluna var . Pode ser útil para identificar caracteres "estranhos" em variáveis numéricas
Tabela.replace('?', np.nan , inplace=True)	Nem sempre os valores em falta aparecem a vazio, e nesse caso para um melhor tratamento é melhor substituí-los por nan
print(Tabela. isnull().any().any())	Indica se existem missing values na Tabela

Nota: O pandas também tem uma função NaN equivalente ao numpy, simplesmente a do numpy é mais amplamente utilizada


DUPLICADOS: COMANDOS ÚTEIS EM PYTHON



EDIT.

Comando	Ação
print(Tabela. isnull().sum())	Número de <i>missing values</i> por cada coluna
Tabela[Tabela. isnull().any (axis= 1)]	Devolve as observações com missing values em alguma coluna
Apagar <i>missing values</i>	
Tabela. dropna (inplace = True)	Apaga todas as observações que tenham <i>missing values</i> em alguma coluna. Equivalente a (axis = 0).
Tabela. dropna (axis = 1 , inplace = True)	Apaga todas as colunas que existam <i>missing values</i>
Tabela. dropna (subset = ['var'], axis = 0 , inplace = True)	Apaga todas as observações com <i>missing values</i> na coluna var .

DUPLICADOS: COMANDOS ÚTEIS EM PYTHON



EDIT.

Comando	Ação
Substituir <i>missing values</i>	
<code>Tabela[var].replace(np.nan, a, inplace = True)</code>	Substitui, na coluna var , os <i>missing values</i> pela valor da variável a
Métricas	
<code>Tabela[var].mean()</code>	Devolve a média da coluna var
<code>Tabela[var].median()</code>	Devolve a mediana da coluna var
<code>Tabela[var].value_counts().idxmax()</code>	Devolve a moda da coluna var
<code>Tabela[Tabela[var1] == 'a'][var2].mean()</code>	Devolve a média (ou outra métrica) da coluna var2 para as observações em que a coluna var1 é igual a a

Nota: Métodos mais avançados que impliquem modelação sessão referidos no **módulo de Machine Learning**

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 3

31

31

MISSING VALUES



EDIT.

BORA LÁ POR A MÃO NA MASSA



Inteligência Artificial: Gen AI & LLM

Módulo 3 – Exploratory Data Analysis – Sessão 3

32

32

EXERCÍCIO I		E D I T.
Linha 1	Importar os dados do ficheiro Carros2 para uma DataFrame chamada Carros2	José P.M.
Linha 2	Explorar conteúdo da tabela e identificar potenciais situações de erro	Nuno
Linha 3	Identificar <i>missing values</i> e converte-los para o formato NaN	Carolina M.
Linha 4	Corrigir tipologia das colunas, caso necessário	Andreia
Linha 5	Indicar o total de observações com <i>missing values</i> e o número de missing values por variável	Susana
Linha 6	No caso da variável price , caso ela tenha <i>missing values</i> , fazer drop	Tamara
Linha 7	Caso existam <i>missing values</i> noutras variáveis quantitativas , substitui-os pela média	José M.
Linha 8	Caso existam <i>missing values</i> noutras variáveis qualitativas , substitui-os pela moda	Stefane
Linha 9	Confirmar que não existem mais valores em falta	José F
Linha 10	Exportar os dados de Carros2 num ficheiro csv com o nome Carros3	Alexandre
DATA SCIENCE & BUSINESS ANALYTICS		Módulo 3 – Exploratory Data Analysis – Sessão 3
		33

PREPARAÇÃO DE DADOS

E D I T.

NORMALIZAÇÃO DE ESCALAS

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 3

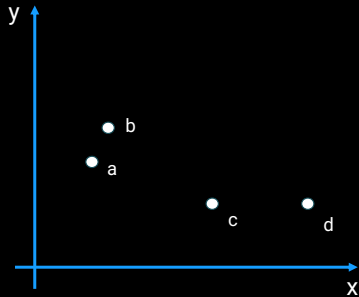
34

NORMALIZAÇÃO DE ESCALAS: PROBLEMA

EDIT.

Os modelos de segmentação, e outros modelos de **Data Science**, são feitas com base em várias **característica** e envolvem **cálculos matemáticos** (ex.: calculo de distâncias) que podem ser influenciados pelo nível de grandeza pelo qual essas características são representados.

Exemplo:



Podemos dizer que o ponto **a** e o **b** estão mais perto um do outro que os pontos **c** e **d**?

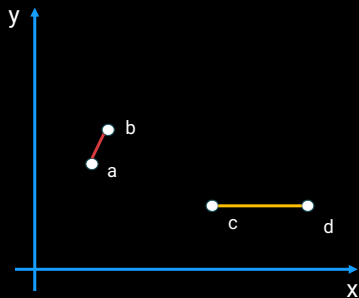
35

NORMALIZAÇÃO DE ESCALAS: PROBLEMA

EDIT.

Os modelos de segmentação, e outros modelos de **Data Science**, são feitas com base em várias **característica** e envolvem **cálculos matemáticos** (ex.: calculo de distâncias) que podem ser influenciados pelo nível de grandeza pelo qual essas características são representados.

Exemplo:



A **linha vermelha** é maior que a **laranja**, logo **a** está mais perto de **b**, do que **c** de **d**

Mas será mesmo assim?

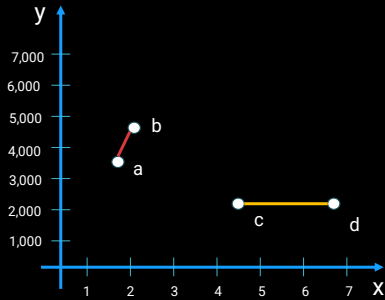
36

NORMALIZAÇÃO DE ESCALAS: PROBLEMA

EDIT.

Os modelos de segmentação, e outros modelos de **Data Science**, são feitas com base em várias **característica** e envolvem **cálculos matemáticos** (ex.: calculo de distâncias) que podem ser influenciados pelo nível de grandeza pelo qual essas características são representados.

Exemplo:



Agora já parece que **c** e **d** estão mais perto

Mas **como podemos normalizar** as escalas das 2 variáveis, x e y, por forma a que seja possível realmente **comprar** as distâncias entre os pontos?

NORMALIZAÇÃO DE ESCALAS: COMO RESOLVER?

EDIT.

Existem várias transformações que se podem aplicar, com diferentes aplicações, vantagens e desvantagens. Exemplo de algumas transformações **comuns** :

Novo Intervalo de Valores

❖ **Min-Max**

$$X_{normalizado} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

onde X_{min} e X_{max} são o mínimo e o máximo de X respetivamente

[0 : 1]

❖ **Estandarização**

$$X_{normalizado} = \frac{X - \mu}{\sigma}$$

onde μ e σ são o valor esperado e a variância de X respetivamente

[-∞ : +∞]

Mas **concentrado** em:

[0 : 1]

NORMALIZAÇÃO DE ESCALAS: COMO RESOLVER?



Existem várias transformações que se podem aplicar, com diferentes aplicações, vantagens e desvantagens. Exemplo de algumas transformações **comuns** :

Novo Intervalo de Valores

❖ **Logaritmo***

$$X_{normalizado} = \log(X)$$

[-∞ : +∞]

Mas mais concentrado

❖ **Quantis***

Transforma a variável em percentis

[1 : 100]

! * Estas transformações **mudam a distribuição** da variável, e por consequência a relação com as restantes variáveis, podendo **impactar na sua relevância** para o estudo / análise / modelo em desenvolvimento.

NORMALIZAÇÃO DE ESCALAS: COMANDOS ÚTEIS EM PYTHON



Comando	Ação
Min = Tab[‘var’].min() Max = Tab[‘var’].max()	Cálculo do mínimo da variável var Cálculo do máximo da variável var
Tabela[‘var’] = (Tabela[‘var’] - Min) / (Max - Min)	Aplicação da transformação Min-Max à variável var
Mean = Tab[‘var’].mean() Var = Tab[‘var’].std()	Cálculo da média (~valor esperado) da variável var Cálculo do desvio padrão (~variância) da variável var
Tabela[‘var’] = (Tabela[‘var’] - Mean) / (Var)	Estandardização da variável var
Tabela[‘var’] = np.log1p(Tabela[‘var’])	Aplicação da transformação Logaritmo à variável var
Tabela[‘var_decile’] = pd.qcut(Tabela[‘var’], q=10, labels = False) + 1	Criação de nova variável var_decile que corresponde ao valor de var quando agrupada em 10 decis .

NORMALIZAÇÃO DE ESCALAS

EDIT.

BORA LÁ POR A MÃO NA MASSA



Inteligência Artificial: Gen AI & LLM

Módulo 3 – Exploratory Data Analysis – Sessão 3

41

41

EXERCÍCIO II

EDIT.

Linha 1	Importar os dados do ficheiro Carros3 para uma DataFrame chamada Carros3	Carolina L.
Linha 2	Avaliar os principais indicadores das variáveis quantitativas do DataFrame Carros3 e verifica se existe um problema de escalas .	Yhoanna
Linha 3	Criar uma nova variável length_min_max através da transformação Min-Max aplicada a length . Validar se os novos valores da variável são os previstos pela transformação.	Sara
Linha 4	Criar uma nova variável width_std através da standardização da variável width . Validar se os novos valores da variável são os previstos pela transformação.	Filipa
Linha 5	Criar uma nova variável height_log através de uma transformação Logarítmica aplicada a height . Validar se os novos valores da variável são os previstos pela transformação.	Ana
Linha 6	Criar uma nova variável price_decile através da transformação em decis da variável price . Validar se os novos valores da variável são os previstos pela transformação.	Rui
Linha 7	Exportar os dados de Carros3 num ficheiro csv com o nome Carros4	Joana

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 3

42

42