

EDIT.

Módulo 3 – Sessão 5

EXPLORATORY  
DATA ANALYSIS

TUTORA  
Carla Cardoso  
Freelancer AI Manager


13 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 5

2

2

TRATAMENTO E PREPARAÇÃO

EDIT.

TRANSFORMAÇÃO

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 5

3

3

TRANSFORMAÇÃO: CONTEXTO

EDIT.

Muitas vezes pode ser útil em termos de **negócio** ou em termos **analíticos transformar** as variáveis existentes ou mesmo **criar novas variáveis**.

Exemplo:

car-id	release-date	num-of-doors	body-style	wheel-base	length	width	height	price
15471	11/02/2001	four	sedan	93,7	157,3	63,8	50,6	7.609
22441	14/07/2014	?	hatchback	93,7	157,3	63,8	50,6	8.558
32254	14/11/2024	four	wagon	103,3	174,6	64,6	59,8	48.921

Talvez seja interessante saber há quanto tempo foi o lançamento do carro.

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 5

5

5

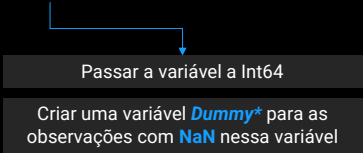
TRANSFORMAÇÃO: CONTEXTO

E D I T.

Muitas vezes pode ser útil em termos de **negócio** ou em termos **analíticos transformar** as variáveis existentes ou mesmo **criar novas variáveis**.

Exemplo:

car-id	release-date	num-of-doors	body-style	wheel-base	length	width	height	price
15471	11/02/2001	four	sedan	93,7	157,3	63,8	50,6	7.609
22441	14/07/2014	?	hatchback	93,7	157,3	63,8	50,6	8.558
32254	14/11/2024	four	wagon	103,3	174,6	64,6	59,8	48.921



\* Será explicado nos seguintes slides que tipo de variáveis são as **Dummy**

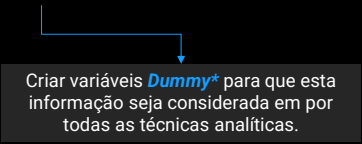
TRANSFORMAÇÃO: CONTEXTO

E D I T.

Muitas vezes pode ser útil em termos de **negócio** ou em termos **analíticos transformar** as variáveis existentes ou mesmo **criar novas variáveis**.

Exemplo:

car-id	release-date	num-of-doors	body-style	wheel-base	length	width	height	price
15471	11/02/2001	four	sedan	93,7	157,3	63,8	50,6	7.609
22441	14/07/2014	?	hatchback	93,7	157,3	63,8	50,6	8.558
32254	14/11/2024	four	wagon	103,3	174,6	64,6	59,8	48.921



\* Será explicado nos seguintes slides que tipo de variáveis são as **Dummy**

TRANSFORMAÇÃO: CONTEXTO



Muitas vezes pode ser útil em termos de **negócio** ou em termos **analíticos transformar** as variáveis existentes ou mesmo **criar novas variáveis**.

Exemplo:

car-id	release-date	num-of-doors	body-style	wheel-base	length	width	height	price
15471	11/02/2001	four	sedan	93,7	157,3	63,8	50,6	7.609
22441	14/07/2014	?	hatchback	93,7	157,3	63,8	50,6	8.558
32254	14/11/2024	four	wagon	103,3	174,6	64,6	59,8	48.921

Variáveis em polegadas. Devemos transformar para metros.

8

TRANSFORMAÇÃO: CONTEXTO



Muitas vezes pode ser útil em termos de **negócio** ou em termos **analíticos transformar** as variáveis existentes ou mesmo **criar novas variáveis**.

Exemplo:

car-id	release-date	num-of-doors	body-style	wheel-base	length	width	height	price
15471	11/02/2001	four	sedan	93,7	157,3	63,8	50,6	7.609
22441	14/07/2014	?	hatchback	93,7	157,3	63,8	50,6	8.558
32254	14/11/2024	four	wagon	103,3	174,6	64,6	59,8	48.921

Variáveis em dólares. Devemos transformar para euros.

Aparenta ser um outliers. De modo a não perder informação, devemos criar uma variável **Dummy\*** a indicar que se tratava de um outlier.

Pode ser interessante **criar classes** como "Alto", "Médio" e "Baixo"

\* Será explicado nos seguintes slides que tipo de variáveis são as **Dummy**

9

TRANSFORMAÇÃO: DUMMY



As variáveis **Dummy** são variáveis booleanas que nos indicam a presença ou ausência de determinada característica.

São extremamente úteis para converter **variáveis qualitativas** em **quantitativas**, pois algumas técnicas **não permitem o uso** de variáveis quantitativas, como é o caso dos modelos de segmentação baseados em distâncias.

Exemplo:

car-id	body-style	body-style-sedan	body-style-wagon
15471	sedan	True	False
22441	hatchback	False	False
32254	wagon	False	True

Variáveis **Dummy**.  
Em teoria uma variável com **n** categorias, necessita de **n-1** variáveis **Dummy**



TRANSFORMAÇÃO: COMANDOS ÚTEIS EM PYTHON



Comando	Ação
Criação de variáveis Dummy	
<pre>Tabela2 = pd.get_dummies(Tabela, columns=['var'], prefix='var', drop_first=False)</pre>	<p>Cria uma nova Tabela2 com novas variáveis <b>Dummy</b> criadas com base na variável quantitativa <b>var</b>. A opção <b>drop_first</b> permite eliminar ou não a classe "redundante".</p> <p>Esta função elimina a variável original (qualitativo) <b>var</b>.</p>

TRANSFORMAÇÃO

EDIT.

BORA LÁ POR A MÃO NA MASSA

Inteligência Artificial: Gen AI & LLM

Módulo 3 – Exploratory Data Analysis – Sessão 5

12

12

EXERCÍCIO I

EDIT.

Linha 1	Sobre o Carros5 criar uma nova variável numérica com base na variável num-of-doors	Ana
Linha 2	Criar variáveis Dummy para a variável body-style	Sara
Linha 3	Converter as variáveis wheel-base, length, width e height para centímetros sem casas decimais	Carolina M
Linha 4	Converter a variável price para euros com 2 casas decimais	Carolina L
Linha 5	Com base na análise dos valores da variável price, criar uma nova variável com 3 níveis de preço: alto, médio e baixo.	Joana
Linha 6	Rever o exercício I da sessão 4 e criar variáveis Dummy que indicam, para cada observação / variável se foi ou não alvo de tratamento outliers.	Filipa

DATA SCIENCE & BUSINESS ANALYTICS

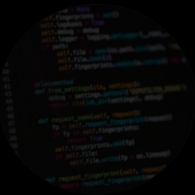
Módulo 3 – Exploratory Data Analysis – Sessão 5

13


13

AGENDA


EDIT.



REVISÃO DE PYTHON



TRATAMENTO E PREPARAÇÃO



VISUALIZAÇÃO DE DADOS

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 5

14

14

VISUALIZAÇÃO DE DADOS: MOTIVAÇÃO

EDIT.

Ainda que as métricas **estatísticas descritivas** das variáveis nos permitam ter informação sobre o comportamento dos dados, muitas vezes elas **pode ser enganadoras**.

Imaginem 4 conjuntos de dados, cada um com 2 variáveis:

Conjunto	Pares	Média X	Desvio Padrão X	Média Y	Desvio Padrão Y	Correlação entre X e Y
1	$(x_1, y_1)$	9	11	7,5	4,125	0,816 $Y = 3 + 0,5 X$
2	$(x_2, y_2)$					
3	$(x_3, y_3)$					
4	$(x_4, y_4)$					

Imaginemos que o nosso objetivo é **prever os valores de Y** em função dos valores de **X**, parece que em qualquer dos conjuntos, a **equação  $Y = 3 + 0,5 X$  é igualmente ajustada**.

Fonte: [Quarteto de Anscombe – Wikipédia, a enciclopédia livre](#)

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 5

19

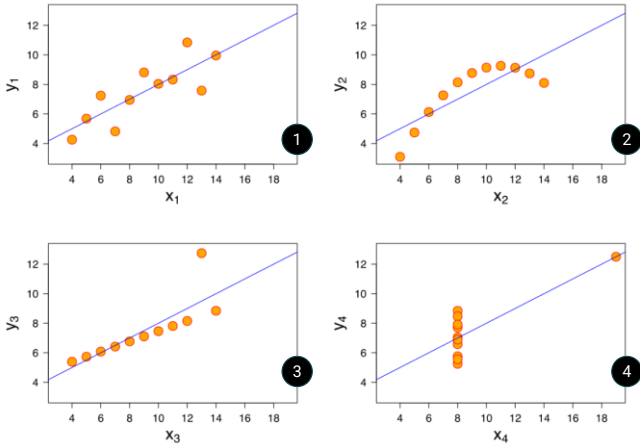
19

VISUALIZAÇÃO DE DADOS: MOTIVAÇÃO

EDIT.

Ainda que todas as métricas **estatísticas descritivas** dos 4 conjuntos **sejam iguais**, inclusive a correlação entre as variáveis X e Y, o comportamento de cada conjunto de dados é muito diferente.

No que toca à **previsão de Y com base em X**, o **que podemos concluir** da visualização dos dados?



Fonte: Quarteto de Anscombe – Wikipédia, a enciclopédia livre

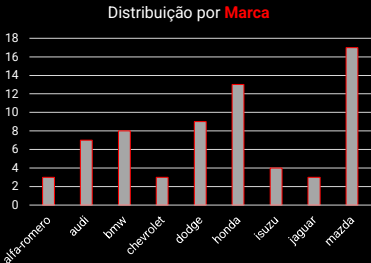
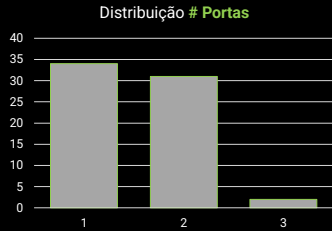
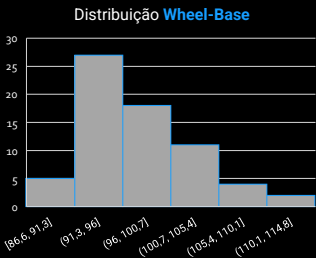
VISUALIZAÇÃO DE DADOS: TIPOS DE VISUALIZAÇÃO

EDIT.

Existem várias formas para **visualizar os dados**, tendo em conta o **tipo de variáveis** e o **objetivo** da visualização:

❖ **Análise Univariada** (apenas 1 variável)

Útil para compreender distribuição das variáveis, escalas, identificação de missing values, outliers, etc.





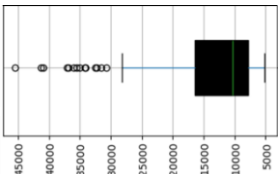
VISUALIZAÇÃO DE DADOS: TIPOS DE VISUALIZAÇÃO



❖ **Análise Univariada** (cont.)

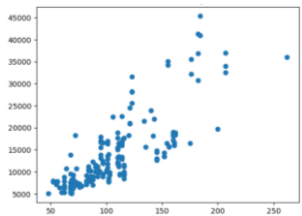
Outra visualização muito útil que já falamos foi a Box-Plot. Permite o nível de concentração ou dispersão dos dados e existências de **outliers**.

BoxPlot Price

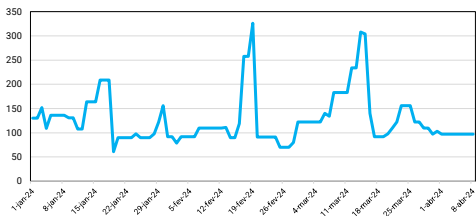


❖ **Análise Bivariada** (2 variáveis)

Scatter Plot: Horsepower vs. Price



Series Temporais: Vendas por Dia



VISUALIZAÇÃO DE DADOS: TIPOS DE VISUALIZAÇÃO



❖ **Análise Multivariada** (2 ou mais variáveis)

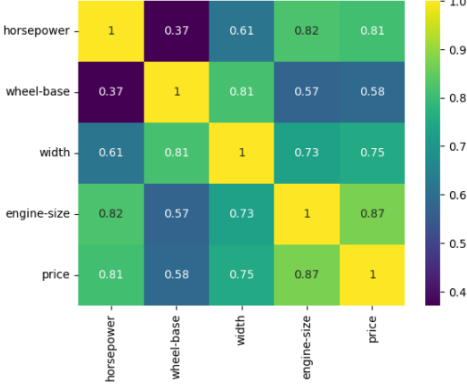
Um exemplo de uma análise multivariada são os **Heatmaps** que permitem avaliar **correlações** entre variáveis.

Se quisermos **prever o valor da variável price**, qual parece ser a **variável mais relevante**?

Porque é que essa conclusão **pode estar incorreta**?



Correlation Heatmap



VISUALIZAÇÃO DE DADOS: COMANDOS ÚTEIS EM PYTHON



EDIT.

Comando	Ação
Biblioteca	
import matplotlib.pyplot as plt	Biblioteca com funções que permitem criar <b>vários tipos de gráficos</b> , histogramas e muito mais.
import seaborn as sns	Biblioteca especializada na <b>opções visuais</b> dos gráficos
Univariadas	
Tabela[ <b>var</b> ].hist(bins=10, edgecolor='black')	Cria <b>histograma</b> com <b>10 colunas</b> da variável <b>var</b>
Tabela.boxplot(column='var', patch_artist=True)	Cria <b>boxplot</b> variável <b>var</b>
freq_tab = Tabela[ <b>var</b> ].value_counts()	Cria uma <b>tabela de frequências</b> com base em <b>var</b>
freq_tab.plot(kind='bar')	Cria um <b>gráfico de barras</b> segundo a <b>tabela de frequências</b>
freq_tab.plot.pie()	Cria um <b>Pie Chart</b> segundo a <b>tabela de frequências</b>

VISUALIZAÇÃO DE DADOS: COMANDOS ÚTEIS EM PYTHON



EDIT.

Comando	Ação
Bivariadas	
plt.scatter(Tabela[ <b>var1</b> ], Carros3[ <b>var2</b> ])	Cria um gráfico <b>Scatter Plot</b> para as variáveis <b>var1</b> e <b>var2</b>
Multivariadas	
correl = Tabela[['var1','var2',...]].corr()	Cria uma matriz de <b>correlações</b> chamada <b>correl</b>
sns.heatmap(correl, annot=True, cmap='viridis')	Desenha um <b>heatmap</b> em função da matriz <b>correl</b> . O parâmetro <b>annot = True</b> mostra o <b>valor</b> das correlações e <b>cmap</b> permite configurar as <b>cores</b> .

VISUALIZAÇÃO DE DADOS

EDIT.

BORA LÁ POR A MÃO NA MASSA

Inteligência Artificial: Gen AI & LLM

Módulo 3 – Exploratory Data Analysis – Sessão 5

29

29

EXERCÍCIO II

EDIT.

Linha 1	Analisar a distribuição da variável <b>price</b> através de uma <b>Boxplot</b> . Que <b>conclusões</b> se podem tirar deste gráfico?	Alexandre
Linha 2	Escolher uma variável que seja adequado ser analisada via <b>gráfico de barras</b> e começar por criar uma <b>tabela de frequências</b> .	Carolina L
Linha 3	Criar um <b>gráfico de barras</b> e um <b>Pie Chart</b> com base na tabela de frequências criada no passo anterior.	Filipa
Linha 4	Explorar via <b>Scatter Plot</b> a relação entre a variável <b>hoursepower</b> e a variável <b>price</b> . Que <b>tipo de relação</b> parece que as variáveis apresentam?	Nuno
Linha 5	Criar um <b>Heatmap</b> para explorar a <b>correlação</b> entre todas as variáveis quantitativas. Que <b>conclusões</b> podemos retirar desta representação gráfica?	José P M

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 5

30

30

FINAL

EDIT.

RESUMO

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 5

31

31

RESUMO

EDIT.

- ❖ Verificar que os dados que temos **são adequados ao objetivos** do nosso trabalho
- ❖ Analisar a informação recebida através da **visualização dos dados** e **análise de indicadores**
- ❖ Realizar os ajustes e correções necessários para garantir que os dados estão preparados para análise
- ❖ Verificar se são **válidos os pressupostos** dos modelos estatísticos que pretendemos aplicar
- ❖ Sugerir **hipóteses** para as causas dos fenómenos observados
- ❖ Ter em atenção aos dados que temos na **seleção de técnicas** adequadas ao problema
- ❖ **Maximizar o conhecimento** da base de dados e da sua estrutura
- ❖ Garantir que as **fontes** de dados são **fiáveis** e **consistentes**
- ❖ Visualizar potenciais **relações entre variáveis** independentes e dependentes
- ❖ Criar **novas variáveis** e **eliminar** variáveis **pouco relevantes** ou com problemas

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 5

32

32

RESUMO

EDIT.

❖ Verificar que os dados que temos **são adequados ao objetivos** do nosso trabalho

❖ Analisar a informação recebida através da **visualização dos dados** e **análise de indicadores**

❖ Garantir que as **fontes** de dados são **fiáveis** e **consistentes**

❖ Visualizar potenciais **relações entre variáveis** independentes e dependentes

❖ Criar **novas variáveis** e **eliminar** variáveis **pouco relevantes** ou com problemas

Compreender

Analisar

Transformar

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 5

33

33

FINAL

EDIT.

Questionário

DATA SCIENCE & BUSINESS ANALYTICS

Módulo 3 – Exploratory Data Analysis – Sessão 5

34

34