EDIT.

WORKSHOP
Digital Data Analytics
Analytics descritivo

Jorge da Costa Ferreira

WORKSHOP

Digital Data Analytics

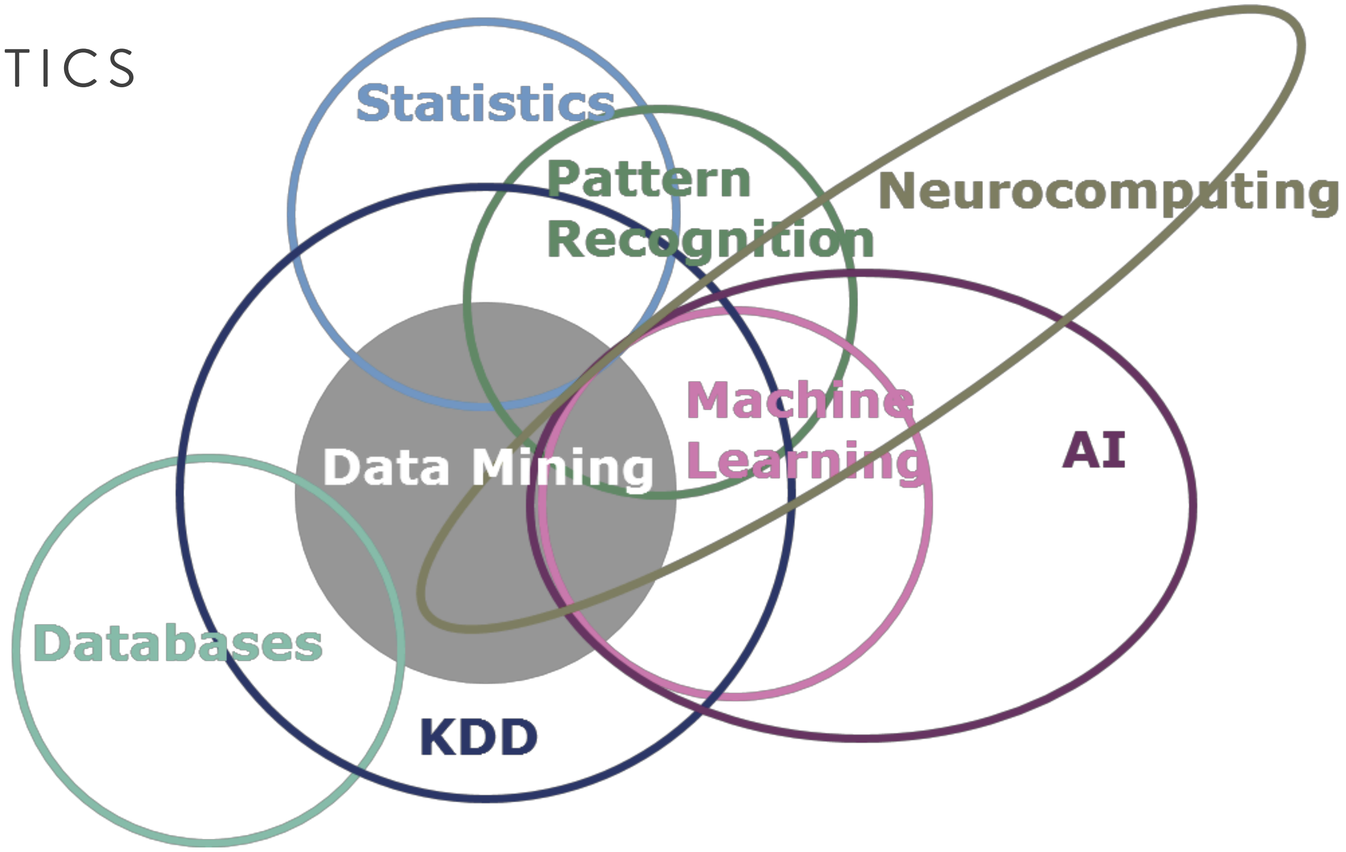
QUAIS SÃO AS ABORDAGENS DE ANALYTICS DESCRITIVO? QUE ANÁLISES SE FAZEM DEPENDENDO DO TIPO DE DADOS?

Jorge da Costa Ferreira

FERRAMENTAS DO DATA ANALYTICS

O data analytics é composto por diversas áreas de aplicação que vão desde a **estatística**, as **bases de dados**, o **data mining**, o **machine learning**, a **inteligência artificial**, etc.

Muitas destas áreas se sobrepõem e têm propósitos semelhantes e/ou complementares.



ANALYTICS DESCRITIVO

Analytics descritivo é uma fase preliminar do processamento de dados para no data mining. Esta área do analytics cria um sumário dos dados históricos para retirar informação útil e possivelmente preparar os dados para análises futuras.

ANALYTICS DESCRITIVO

O analytics descritivo divide-se em duas principais áreas:

- Análise exploratória univariada: avalia a distribuição de cada variável de forma unitária

- Análise exploratória bivariada: é realizada com o intuito de perceber o impacto da variação de uma variável no comportamento de outra

Estas análises tem abordagens diferentes dependendo do tipo de dados em questão, tendo uma abordagem para os dados qualitativos e para os dados quantitativos.

CARACTERIZAÇÃO DOS DADOS 

1. TIPO

Define se o atributo representa quantidades, sendo denominado como **quantitativo** (ou numérico), ou qualidades, sendo designado de **qualitativo** (ou simbólico ou categórico). Os valores quantitativos podem ainda ser classificados como **contínuos** ou **discretos**.

1. Qualitativos (e.g região do país)

2. Quantitativos
- Discretos (e.g. número de visitas de clientes a uma farmácia)
- Contínuos (e.g. Altura)

CARACTERIZAÇÃO DOS DADOS 

2. ESCALA

Qualitativos:

Nominais - Valores consistem apenas em nomes diferentes, carregam a menor quantidade de informação possível (e.g. Cidade)

Ordinais - Os valores têm uma ordenação, sendo possível aplicar operadores lógicos >, < (e.g faixa etária)

Quantitativos:

Intervalar - Atributos que variam dentro de um determinado intervalo. Não existe um zero absoluto (e.g. temperatura em graus celsius)

Racional - Existe um zero absoluto (e.g. número de visitas de clientes a uma farmácia)

EXPLORAÇÃO DOS DADOS UNIVARIADOS

1. MEDIDAS DE CENTRALIDADE

Variam se os dados são numéricos ou simbólicos. Para simbólicos a métrica mais utilizada é a **moda** (valor com maior frequência). Já para dados numéricos as métricas mais utilizadas são a **média** e a **mediana**. Esta última é mais robusta à presença de outliers.

nota	frequência (fi)
3	4
4,5	5
5	2
6,5	3
7	6
8	5
9	4
10	1
Total	30

(Moda)

$$\bar{x} = \frac{\sum x_i}{n}$$

(Média)

2, 2, 3, **7**, 8, 9, 9
Mediana = 7

1, 4, 4, **5**, **6**, 7, 7, 7
Mediana = (5+6) ÷ 2
= 5.5

(Mediana)

EXPLORAÇÃO DOS DADOS UNIVARIADOS

1. MEDIDAS DE CENTRALIDADE

Variam se os dados são numéricos ou simbólicos. Para simbólicos a métrica mais utilizada é a **moda** (valor com maior frequência). Já para dados numéricos as métricas mais utilizadas são a **média** e a **mediana**. Esta última é mais robusta à presença de outliers.

Outras medidas de centralidade são:

Trimean: 1/4 Q1 + 1/2 Q2 + 1/4 Q3

x% trimmed mean: Média aritmética computacionada sem os x% valores mais altos e os x% valores mais baixos da amostra

EXPLORAÇÃO DOS DADOS UNIVARIADOS

2. MEDIDAS DE LOCALIZAÇÃO

Para além das medidas apresentadas anteriormente, existem outras métricas de localização que visam explicar não apenas onde se centra a distribuição da variável, mas também onde se localizam as observações ao longo desta. Exemplo disso são os **Quartis** (Q1, Q2 e Q3) bem como o valor **máximo** e **mínimo** da distribuição.

Os quartis dividem a amostra em 4 secções com igual frequência (25% dos dados). O intervalo interquartil [Q1, Q3[contem 50% dos valores observados.

O segundo quartil (Q2) é a mediana.

Percentís: 5%, 10%, 90%, 95%

EXPLORAÇÃO DOS DADOS UNIVARIADOS

3. MEDIDAS DE DISPERSÃO

Medem a variabilidade dos valores do atributo. Avaliam se os valores estão amplamente dispersos ou concentrados. As medidas mais comuns são o **intervalo interquartil**, a **amplitude**, **variância** e **desvio padrão**.

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

(Variância)

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

(Desvio padrão)

EXPLORAÇÃO DOS DADOS UNIVARIADOS

4. MEDIDAS DE DISTRIBUIÇÃO

Duas métricas importantes de distribuição são a **skewness** e a **kurtosis**. A primeira indica a simetria da distribuição e a segunda o seu achatamento. Ambas são. Instanciações de uma métrica denominada momento.

$$\textbf{Skew} = \frac{n}{(n-1)(n-2)}\sum\left(\frac{x_j - \bar{x}}{s}\right)^3$$

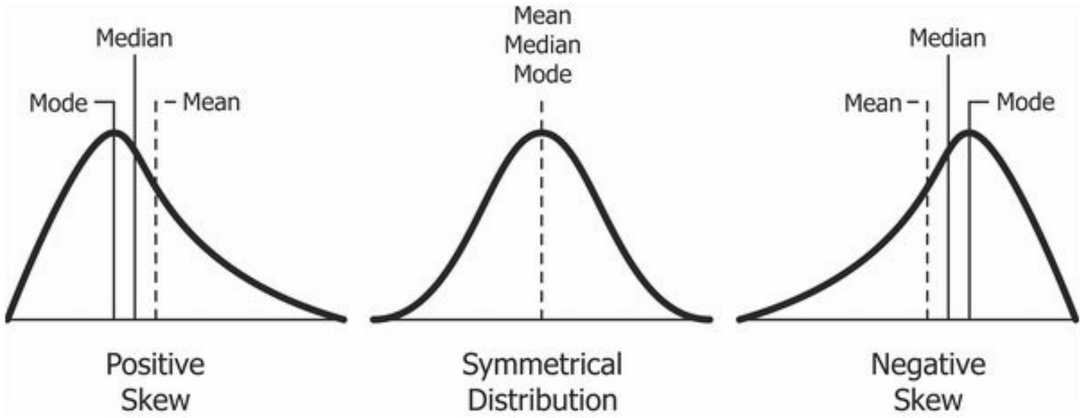
$$\textbf{Kurtosis} = \left\{\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{x_j - \bar{x}}{s}\right)^4\right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

EXPLORAÇÃO DOS DADOS UNIVARIADOS

4. MEDIDAS DE DISTRIBUIÇÃO

Duas métricas importantes de distribuição são a **skewness** e a **kurtosis**. A primeira indica a simetria da distribuição e a segunda o seu achatamento. Ambas são. Instanciações de uma métrica denominada momento.

Skewness = 0 (simétrica)
Skewness > 0 (a distribuição encontra-se mais do lado esquerdo, diz-se assimétrica à direita)
Skewness < 0 (a distribuição encontra-se mais do lado direito, diz-se assimétrica à esquerda)

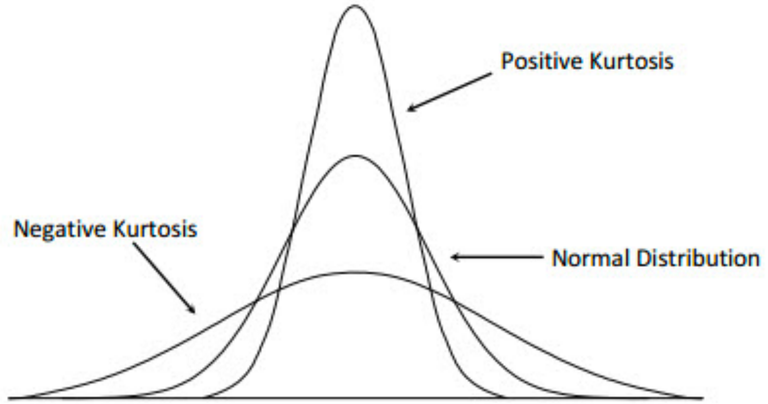


EXPLORAÇÃO DOS DADOS UNIVARIADOS

4. MEDIDAS DE DISTRIBUIÇÃO

Duas métricas importantes de distribuição são a **skewness** e a **kurtosis**. A primeira indica a simetria da distribuição e a segunda o seu achatamento. Ambas são. Instanciações de uma métrica denominada momento.

Kurtosis = 0 (normal)
Kurtosis > 0 (a distribuição é mais alta e concentrada que a distribuição normal)
Kurtosis < 0 (a distribuição é mais achatada do que a distribuição normal)



EXPLORAÇÃO DOS DADOS UNIVARIADOS

5. OUTLIERS

Outliers são valores atípicos que se encontram fora do espectro de valores da restante série. Existem varias formas de calcular outliers sendo que a mais simples e uma das mais utilizadas se baseia no intervalo interquartil (método de Tukey).

IQR = Q3 - Q1

<u>Outliers moderados:</u>

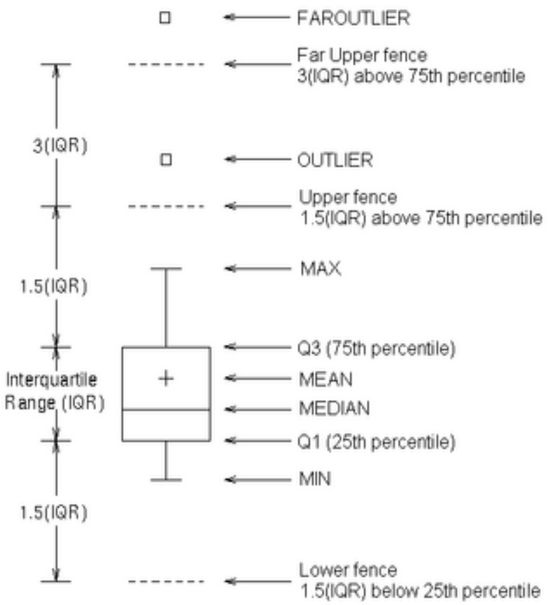
Limite inferior = Q1- 1.5 * IQR

Limite superior = Q3 + 1.5 * IQR

<u>Outliers severos:</u>

Limite inferior = Q1- 1.5 * IQR

Limite superior = Q3 + 1.5 * IQR



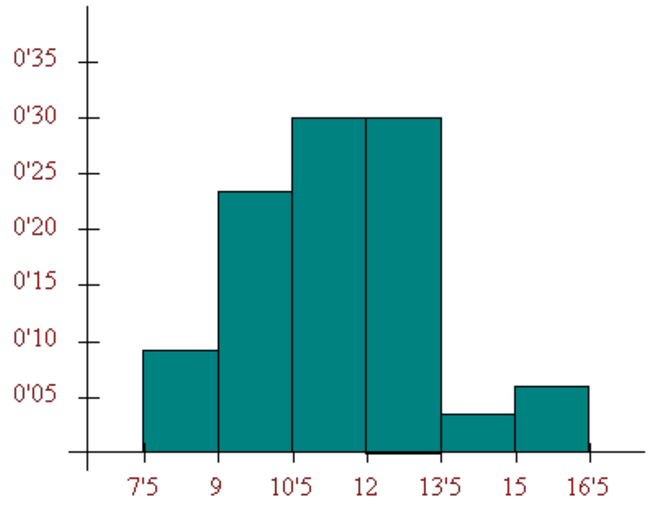
Outros métodos de identificação de outliers:

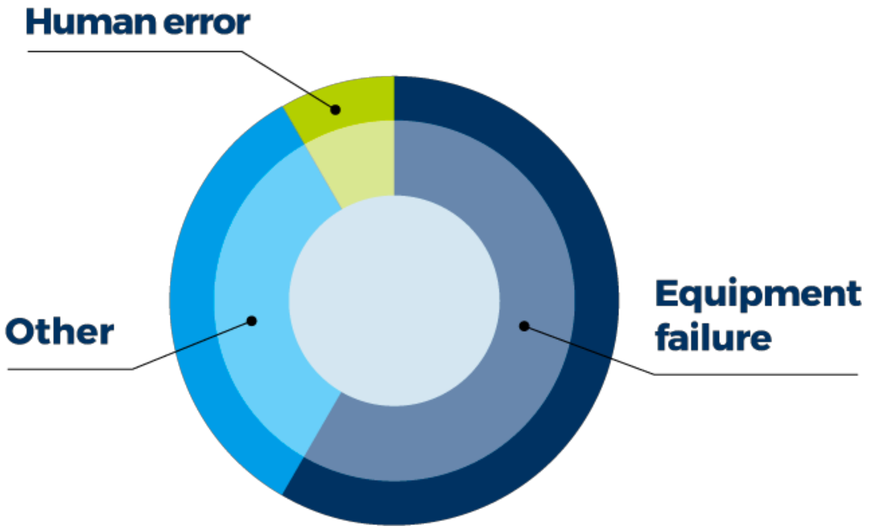
1. Generalized ESD
2. Grubbs' test.
3. Dixon's Q Test.
4. Modified Thompson Tau Test
5. Pierce's Criterion

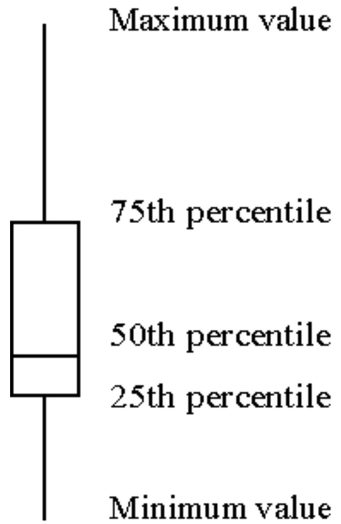
EXPLORAÇÃO DOS DADOS UNIVARIADOS

6. ANÁLISE GRÁFICA

Análises univariadas gráficas recorrem a visualizações como histogramas, pie charts e box plots







EXPLORAÇÃO DOS DADOS BIVARIADA

Tal como a análise univariada depende do tipo das variáveis a analisar, as técnicas utilizadas para a análise bivariada, depende do par de variáveis em análise.

Jorge da Costa Ferreira

EXPLORAÇÃO DOS DADOS BIVARIADA

PAR DE VARIÁVEIS CATEGÓRICAS - TABELAS DE CONTIGÊNCIA

Para pares de variáveis categóricas utilizam-se tabelas de contingência.

Tendo a variável A e B, cada categoria da variável A define uma classe e compara-se a sua distribuição ao longo das diferentes classes de B face à sua distribuição univariada.

EXPLORAÇÃO DOS DADOS BIVARIADA

PAR DE VARIÁVEIS CATEGÓRICAS - TABELAS DE CONTIGÊNCIA

Para pares de variáveis categóricas utilizam-se tabelas de contingência.

Tendo a variável A e B, cada categoria da variável A define uma classe e compara-se a sua distribuição ao longo das diferentes classes de B face à sua distribuição univariada.

		Regiao				Total
		Europa central	Europa de leste	Europa de norte	Europa de sul	
Moeda	Euro	7	2	4	6	19
	Nao-euro	1	6	2	0	9
Total		8	8	6	6	28

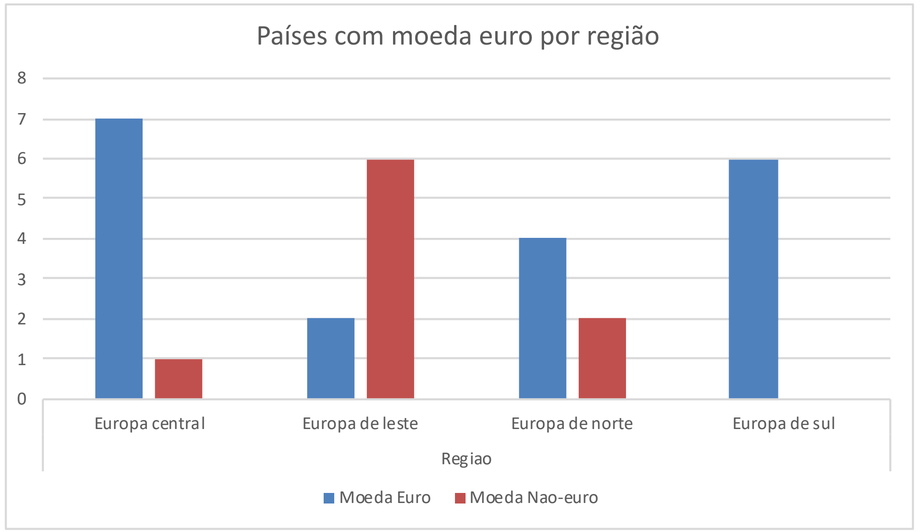
EXPLORAÇÃO DOS DADOS BIVARIADA

PAR DE VARIÁVEIS CATEGÓRICAS - TABELAS DE CONTIGÊNCIA

Para pares de variáveis categóricas utilizam-se tabelas de contingência.

Tendo a variável A e B, cada categoria da variável A define uma classe e compara-se a sua distribuição ao longo das diferentes classes de B face à sua distribuição univariada.

		Regiao				Total
		Europa central	Europa de leste	Europa de norte	Europa de sul	
Moeda	Euro	7	2	4	6	19
	Nao-euro	1	6	2	0	9
Total		8	8	6	6	28



EXPLORAÇÃO DOS DADOS BIVARIADA

PARES DE VARIAVEIS NUMÉRICAS - CORRELAÇÃO E COVARIÂNCIA

A covariância mede o grau com que dois atributos variam juntos (para atributos numéricos). O mesmo acontece com a correlação. Enquanto que a covariância é afetada pela magnitude dos atributos, este fenômeno não acontece com a correlação, pelo que esta métrica permite comparar melhor a relação entre dois atributos.

Correlação = 1 (correlação positiva máxima)
Correlação = -1 (correlação negativa máxima)

EXPLORAÇÃO DOS DADOS BIVARIADA

PARES DE VARIAVEIS NUMÉRICAS - CORRELAÇÃO E COVARIÂNCIA

A covariância mede o grau com que dois atributos variam juntos (para atributos numéricos). O mesmo acontece com a correlação. Enquanto que a covariância é afetada pela magnitude dos atributos, este fenômeno não acontece com a correlação, pelo que esta métrica permite comparar melhor a relação entre dois atributos.

Correlação = 1 (correlação positiva máxima)
Correlação = -1 (correlação negativa máxima)

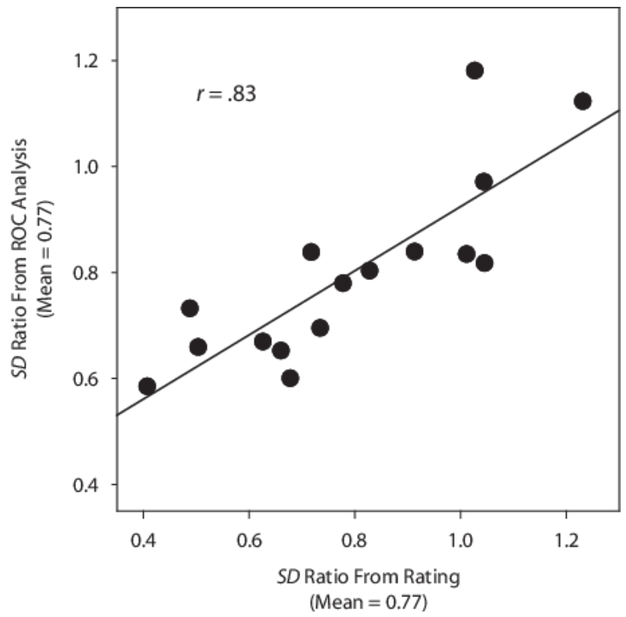
Análises de correlação

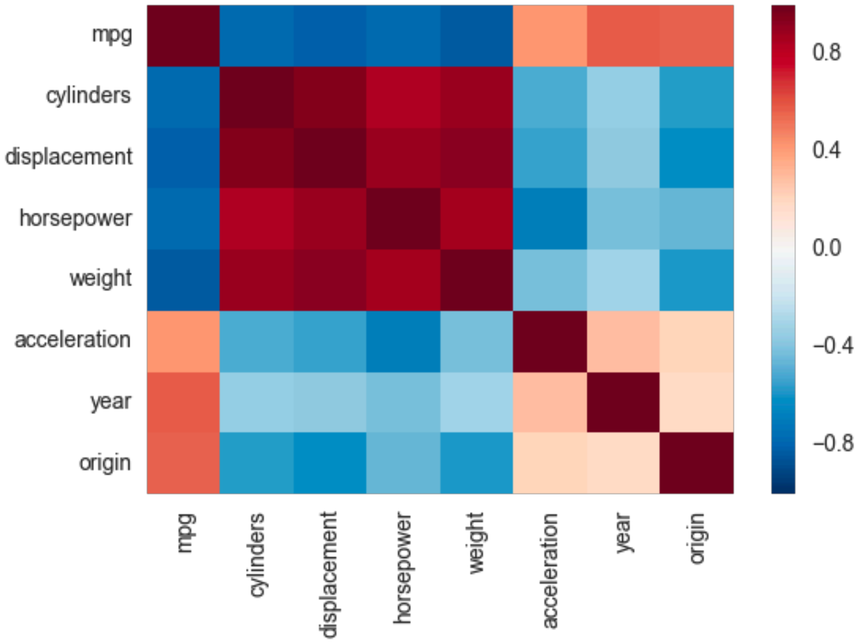
- Correlação linear:
  - Coeficiente de Pearson
- Correlação ordinal:
  - Coeficiente de Spearman
  - Kendall's Tau

EXPLORAÇÃO DOS DADOS BIVARIADA

2. ANÁLISE GRÁFICA

A análise gráfica multivariate pode ser feita através de scatter plots ou heat maps de correlação.

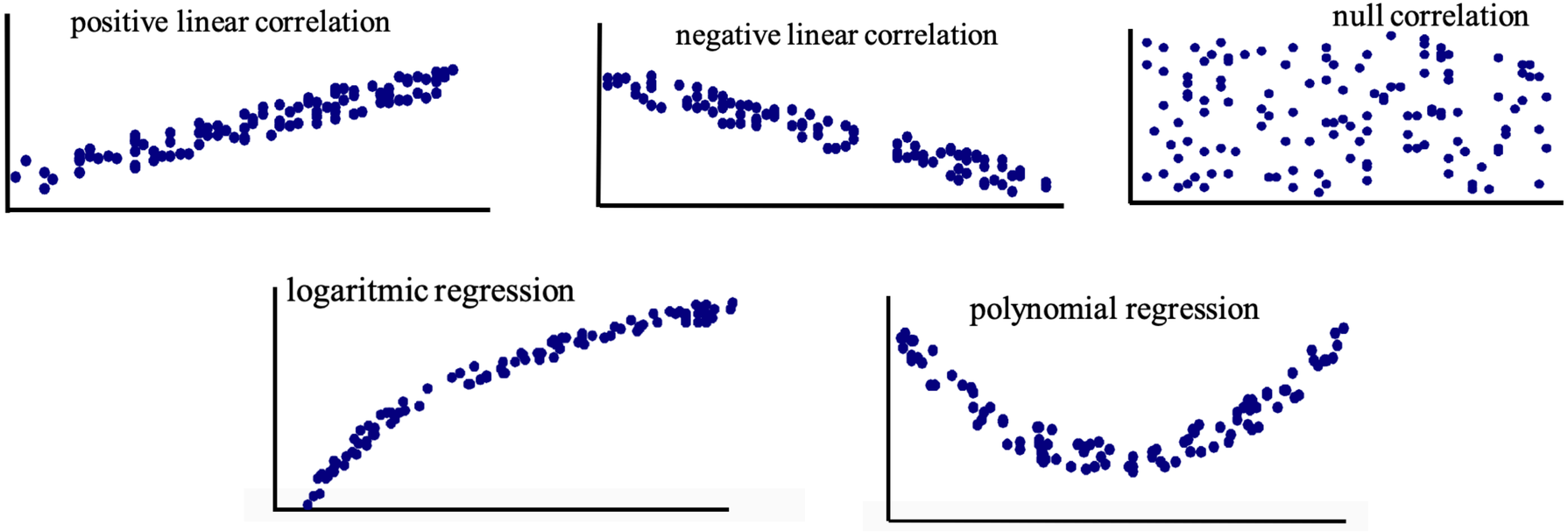




EXPLORAÇÃO DOS DADOS BIVARIADA

2. ANÁLISE GRÁFICA

Scatter-plots:



EXPLORAÇÃO DOS DADOS BIVARIADA

PARES DE VARIAVEIS NUMÉRICAS-CATEGÓRICAS - DIVISÃO EM GRUPOS

Divisão das amostra em K grupos pelas categorias da variável categórica.

Se K=2:

- Comparação dos valores médios para populações Normais (t test ou teorema do limite central para grandes amostras)
- Comparação de medianas, teste de Mann-Whitney para populações que não normalmente distribuídas

Se K>2:

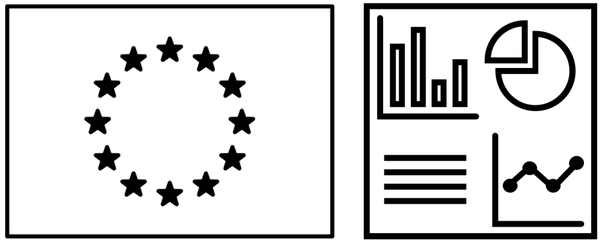
- Análise da variância (ANOVA) para populações Normais
- Teste não paramétrico de Kruskal-Wallis' para populações não Normais

WORKSHOP

Digital Data Analytics

EXERCÍCIO PRÁTICO

Data set - Índices demográficos e econômicos de rendimento e consumo dos países da União Europeia com foco na moeda e região



Jorge da Costa Ferreira

DATA VISUALISATION

Data visualisation é a representação gráfica de informação e dados. Utiliza elementos como gráficos, tabelas, mapas, entre outros. Ferramentas de representação gráfica fornecem uma compreensão rápida de tendências, outliers, padrões e aproximação aos objetivos.

DATA VISUALISATION

Data visualisation é a representação gráfica de informação e dados. Utiliza elementos como gráficos, tabelas, mapas, entre outros. Ferramentas de representação gráfica fornecem uma compreensão rápida de tendências, outliers, padrões e aproximação aos objetivos.

Power BI

tableau

ClickView

EDIT.

WORKSHOP

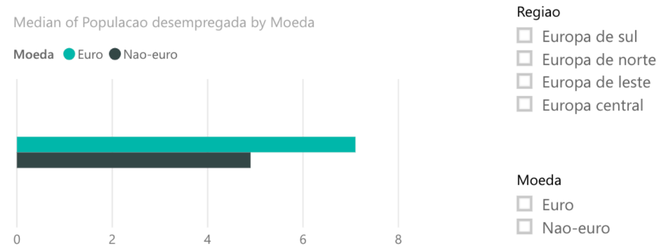
Digital Data Analytics

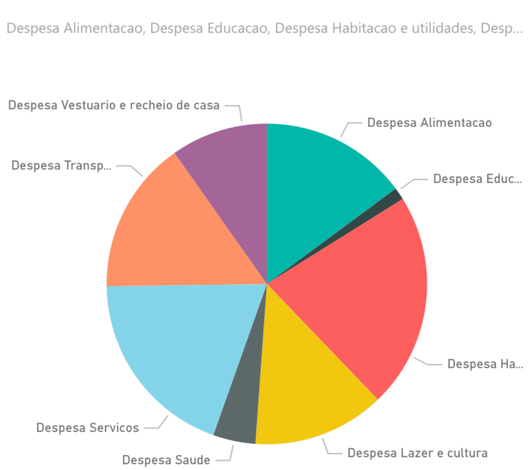
DATA VISUALISATION

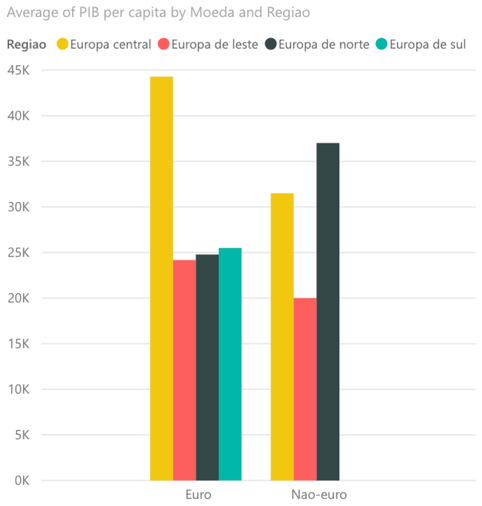
13.17K
Average of Rendimento Ensino basico

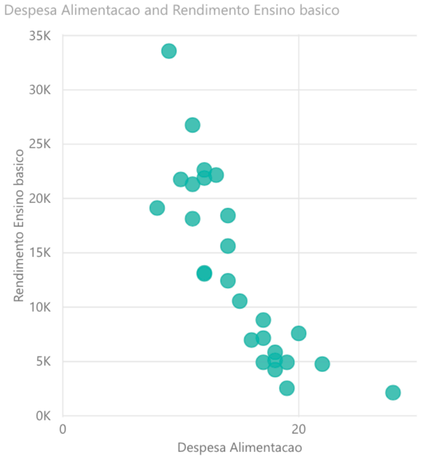
21.99K
Average of Rendimento Ensino superior

16.35K
Average of Rendimento Ensino pos-secundario









Jorge da Costa Ferreira

EDIT.

WORKSHOP
Digital Data Analytics
Áreas de estudo

Jorge da Costa Ferreira

SUPORTE:

- Cálculo da correlação de Pearson e Spearman excel
- Escolher número de bins para um histograma
- Testes de correlação