*Exploratory Data Analysis (EDA)*

**O que é uma EDA?**

A Análise Exploratória de Dados (AED), também conhecida como Exploratory Data Analysis (EDA em suas siglas em inglês), tem como objetivo examinar os dados antes de aplicar qualquer técnica estatística. Dessa forma, o(a) Cientista de Dados consegue um entendimento básico de seus dados e das relações existentes entre as variáveis analisadas.

A EDA fornece métodos simples para organizar e preparar os dados, detectar falhas no design e coleta de dados, tratar e avaliar dados ausentes, identificar casos atípicos e muito mais.

É importante destacar que o teste prévio dos dados é uma etapa necessária, que leva tempo, e que costuma ser negligenciada pelos analistas de dados. As tarefas contidas nesses testes podem parecer insignificantes e inconsequentes à primeira vista, mas são uma parte essencial de qualquer análise estatística.



**UTILIDADES da EDA**

Algumas das perguntas que podemos responder por meio de uma EDA são:

* Existe algum viés nos dados coletados?
* Existem erros na codificação dos dados?
* Como a informação contida em um conjunto de dados é sintetizada e apresentada?
* Existem dados atípicos (outliers)? Quais são? Como os tratamos?
* Existem dados ausentes (missing)? Tem algum padrão sistemático? Como os tratamos?

***Etapas da EDA – Parte 1***

Para realizar uma EDA, é praxe seguir as seguintes etapas ou fases:

1) Preparar os dados para torná-los acessíveis a qualquer técnica estatística.

2) Realizar um exame gráfico da natureza das variáveis individuais a serem analisadas e uma

análise estatística que permita quantificar alguns aspectos gráficos dos dados.

3) Analisar correlações entre as variáveis e dependências.

4) Avaliar, se for necessário, algumas hipóteses sobre a distribuição das variáveis, assimetrias, formas, etc.

5) Identificar os possíveis casos atípicos (outliers) e determinar o impacto potencial que

podem exercer em análises estatísticas posteriores.

6) Estabelecer, se for necessário, o impacto potencial que os dados ausentes (missing) podem ter sobre a representatividade dos dados analisados.

Vamos entender um pouco mais sobre o que acontece em cada etapa ou fase. 🤓

**Etapas da EDA – 1. Prep. dos Dados**

Como bem comentamos, o primeiro passo de uma EDA é deixar os dados acessíveis para qualquer técnica estatística. Para isso, teremos que realizar um input de dados, os quais, lembremos, podem vir de diferentes origens, como, por exemplo: Excel, csv, Bancos de Dados, etc. Em seguida, teremos que escolher o software de análise de dados que utilizaremos para a manipulação e o processamento do dataset. Em nosso caso, utilizaremos o Python.



A grande maioria dos softwares orientados para a análise de dados permitem realizar manipulações de dados antes de analisá-los. Algumas operações úteis são:

- Combinar conjuntos de dados de dois ou mais arquivos distintos.

- Selecionar subconjuntos dos dados.

- Dividir o arquivo dos dados em várias partes.

- Transformar variáveis.

- Filtrar e ordenar o dataset.

- Acrescentar novos dados e/ou variáveis.

- Eliminar dados e/ou variáveis.

- Armazenar dados e/ou resultados.

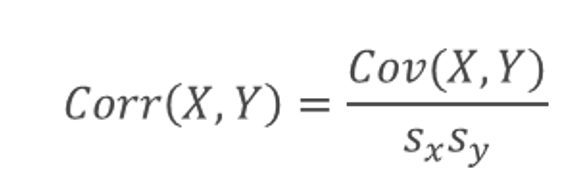
**Etapas da EDA – 2. Análise Estatística**

Uma vez organizados os dados, o segundo passo de uma EDA consiste em realizar uma análise estatística gráfica e numérica das variáveis do dataset, com o objetivo de obter uma ideia inicial sobre a informação contida no conjunto de dados, assim como detectar, caso existam, erros de codificação.

Dentro desse contexto, é importante entender que o tipo de análise que deveremos fazer dependerá da escala de medida da variável analisada.

***Etapas da EDA –* 3. Correlações e Dependências**

A correlação é a covariância, mas dividida pelos desvios padrão das duas variáveis. Apresenta a seguinte fórmula matemática:



A correlação sempre dará um número entre **-1 e 1**

* Quanto mais próximo do valor 1, mais forte é a **relação linear direta** entre as variáveis.
* Quanto mais próximo do valor -1, mais forte é a **relação linear inversa** entre as variáveis.
* Se nos dá 0, então **não há relação linear** entre as variáveis.

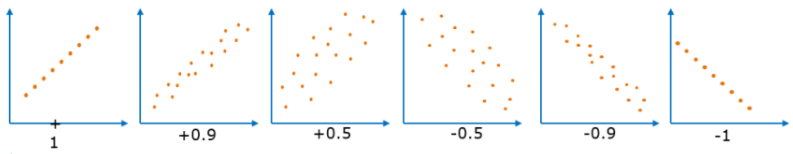
Em relação à força da correlação, falamos sempre de uma correlação:

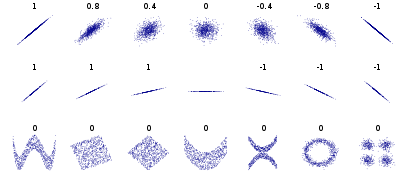
* Nula.
* Fraca.
* Forte.

Também é importante considerar 2 aspectos relevantes:

1. A ausência de correlação significa que não há uma relação linear, mas não que não há relação.
2. A correlação não é, nem implica, uma relação de causalidade.

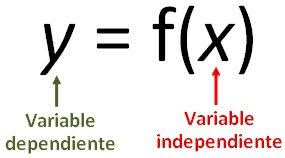
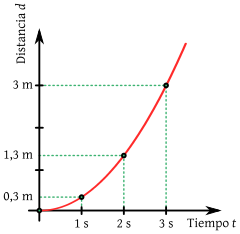
**Correlações:**

****

****

**Dependências:**

A variável dependente é aquela cujo valor depende do valor numérico que a variável independente assume dentro da função matemática.



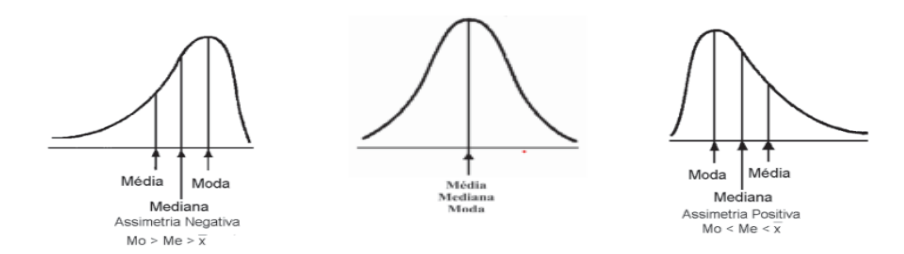
**Etapas da EDA – 4. Distribuição das variáveis**

Chegando nessa etapa, é importante estudar, por exemplo, as “Medidas de Dispersão” dentro do tema da Estatística.

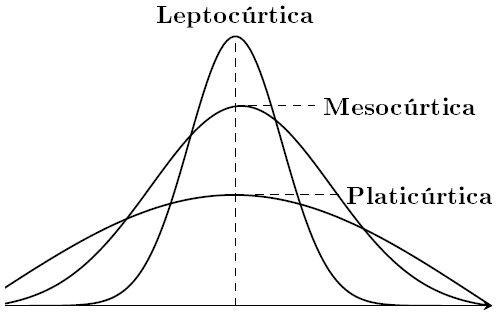
Mas o que são as medidas de dispersão? São aquelas que estudam as características da distribuição de probabilidades observada. Podemos destacar:

* Simetria.
* Curtose.

**Simetria:** uma variável é simétrica se os valores equidistantes da média forem iguais. Para uma maior compreensão, observemos a seguinte imagem:



**Curtose:** acurtose mede o grau de apontamento ou achatamento da curva de distribuição de frequência. Ou seja, nos ajuda a entender “o quão íngreme está a curva”. Adicionalmente, existem diferentes tipos de curtose:

****

***Etapas da EDA –* 5. Identificação de Outliers**

Como já estudamos em outras unidades do curso, temos que prestar atenção redobrada aos outliers, visto que podem ter um potencial negativo dentro de nossa EDA.

Também é muito importante deixar claro que não devemos eliminar os outliers pelo mero fato de serem outliers. A menos que estejamos 100% seguros de que esse valor extremo se deve a um erro de registro, uma falha no instrumento de medição ou algum problema externo que seja verificável, os outliers são observações tão válidas quanto qualquer outra e fazem parte da realidade de nossos dados. 😀

***Etapas da EDA –* 6. Valores Missing**

Uma situação frequentemente encontrada por qualquer cientista de dados é o tratamento dos valores perdidos. Os valores ausentes são aqueles que, para determinada variável, não constam em algumas linhas ou padrões.

Os 3 principais motivos pelos quais normalmente tratamos os valores ausentes são:

* Podem introduzir um viés considerável (uma diferença significativa entre os dados observados e os não observados).
* Deixam a análise e o gerenciamento dos dados mais complicados.
* Geralmente, ocasionam perdas de informação.

Existem vários procedimentos para aplicar quando temos valores perdidos. Mesmo que, basicamente, existam duas aproximações possíveis:

* Eliminar amostras ou variáveis que têm dados ausentes.
* Imputar os valores perdidos, ou seja, substituí-los por estimações.

