Capítulo 1 - Exploratory Data Analysis

- A Estatística foi desenvolvida, em sua maioria, no século passado.
- Teoria da probabilidade (função matemática para estatísitica)
  - Thomas Bayes
  - Pierre-Simon Laplace
  - Carl Gauss
- Ao contrário da natureza teórica pura da probabilidade, estatística é a ciência aplicada, que se preocupa e analisar e modelar os dados.

- Este capítulo foca nos primeiros passos de qualquer projeto de ciência de dados: Análise Exploratória de dados (EDA). É uma nova área da estatística.
- **Estatística clássica** é focada quase que exclusivamente em **inferência.**
- **Inferência**: um complexo conjunto de procedimentos

- Em 1962, John W. Tukey propôs uma nova disciplina científica chamada **análise de dados** que inclui inferência estatística como apenas uma componente.
- Ele forjou links com engenharia e comunidade de ciência da computação.

- Análise de dados evolui bem além do seu escopo original.
  - Rápido desenvolvimento de novas tecnologias.
  - Acesso a cada vez mais e mais dados.

    - Acesso a cada vez mais e mais dados.
    - O grande uso de análises quantitativas.

Elementos de dados estruturados

- **Fontes de dados**: Medidas de sensores, eventos, textos, imagens, vídeos.
- Podemos citar alguns exemplos de **dados não estruturados**:
    - imagens: coleção de pixels
    - textos: sequências de palavras e caracteres especiais
    - cliques: sequência de ações geradas por um usuário, etc.
- Para que os conceitos estatísticos sejam aplicados, dados não estruturados precisam ser manipulados de modo que fiquem estruturados.
- **Dados estruturados** são comumente representados em uma tabela com linhas e colunas.
- Tipos básicos de dados estruturados:
    - Numérico
        - Contínuo: velocidade do vento, duração de um evento
        - Discreto: quantidade de ocorrências de um evento
    - Categórico
        - Binário: dois valores apenas (0/1, sim/não, verdadeiro/falso, etc)
        - Ordinal: ruim, regular, bom e ótimo (escala de avaliação)
- Saber os tipos de dados com os quais estamos lidando, nos ajuda a determinar qual a melhor visualização, tipo de análise e modelo estatístico a ser aplicado.

Dados retangulares

- Dados retangulares possuem linhas (registros de casos), e colunas (indicando as features ou variáveis).
- Sinônimos de *features*: atributos, input, preditor, variável.

- *Saída*: output, target, variável resposta, variável dependente.

Estruturas de dados não-retangulares

- **Séries Temporais**: tratam-se de medidas repetidas da mesma variável num dado intervalo de tempo (por hora, dia, mês, ano, etc).
- **Dados Espaciais:** mapeamento de localizações. São dados mais complexos e variados do que estruturas retangulares.
- **Grafos ou network (redes):** usados para representar relacionamentos físico, social e abstratos (Facebook, LinkedIn são exemplos de network). São usados para certos tipos de problemas como otimização e sistemas de recomendação.

- Leituras adicionais sobre dataframes:
  - **R:** https://stat.ethz.ch/R-manual/R-devel/library/base/html/data.frame.html
  - **Python:** https://pandas.pydata.org/pandas-docs/stable/user_guide/dsintro.html

Estimativas de localização

- **Tendência central**: uma estimativa de onde a maioria dos dados está localizada.
- **Média:** Soma de todos os valores dividida pela quantidade de valores.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Média truncada:** É uma variação da média, é calculada removendo-se um número fixo de valores ordenados no final e, então, calcula-se a média dos valores restantes:

$$\overline{x} = \frac{1}{n-2p}\sum_{i=p+1}^{n-p} x_i$$

em que x_1 é o menor valor e x_n é o maior. Este tipo de média elimina valores extremos.

- **Média Ponderada:** A soma de todos os valores multiplicada por um peso, dividida pela soma dos pesos.

$$\overline{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  → Dois motivos para usar a média ponderada:

  1. Alguns valores são intrinsecamente mais variáveis que outros
  2. Os dados coletados não representam igualmente as diferenças entre os grupos que estamos interessados em medir.

- **Mediana**: O valor que divide os dados no meio, 50% acima e 50% abaixo deste valor.
  - Também é possível calcular a *mediana ponderada*.
  - Assim como a mediana, a mediana ponderada é robusta para outliers.
- **Outliers:** A mediana é dita ser uma *estimativa robusta* de localização, uma vez que não é influenciada por *outliers* (casos extremos).
  - Um **outlier** é um valor que está muito distante dos outros valore no dataset.
  - Os outliers devem ser identificados, e vale à pena investigá-los de maneira mais profunda.

- A métrica básica de localização é a **média,** mas ela pode ser sensível à valores extremos (outliers).
- Outras métricas (**mediana, média truncada**) são menos sensíveis aos outliers e distribuições não usuais, e portanto, mais robustas.

Estimativas de Variabilidade

- Variabilidade ou dispersão são medidas de quão próximos ou espalhados estão os dados.
- **Desvios**
  - A diferença entre valores observados e valores estimados de localização.

    - Sinônimos: erros, resíduos
- **Variância**
    - A soma dos desvios ao quadrado a partir da média dividida por $n - 1$ em que n é o número de valores.
    - Sinônimo: erro quadrático médio
- **Desvio-padrão**
    - A raiz quadrada da variância
- **Desvio Absoluto da Média (MAD - Mean Absolut Deviation)**
    - A média dos valores absolutos dos desvios em relação à média
    - Sinônimo: l1-norm, ou Manhatttan norm
- **Desvio Absoluto Mediano da Mediana**
    - A mediana dos valores absolutos dos desvios em relação à mediana
- **Amplitude**
    - A diferença entre o maior e o menor valor numa base de dados
- **Estatísticas de ordem**
    - Métricas baseadas em valores de dados ordenados do menor para o maior
    - Sinônimo: rank (classificação, posição)
- **Percentil**
    - O valor tal que P% dos valores são iguais a ele ou menores, e (100 - P)% dos valores são maiores.
    - Sinônimo: quantil
- **Amplitude Interquartil (Ou Intervalo Interquartil)**
    - A diferença entre or percentil 75º e o 25º.
    - Sinônimo: IQR (Interquartile Range)

Desvio padrão e estimativas relacionadas

- As estimativas de variação estão baseadas em diferenças, ou *desvios*, entre a estimativa de localização (média) e os dados observados.
- Imagine o conjunto de dados: {1, 4, 4}
  - Média: 3
  - Mediana: 4
  - Os desvios da média seriam:
    - 1 - 3 = -2
    - 4 - 3 = 1
    - 4 - 3 = 1
- Os desvios nos dão uma ideia de quão dispersos os dados estão em relação ao valor central, neste caso, a média.
- A soma dos desvios sempre dá zero, então precisamos de outras maneira para quantificar esta dispersão.

Desvio Absoluto da Média

$$\text{Desvio absoluto da média} = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

em que $\overline{x}$ é a média amostral.

Variância e Desvio padrão

- São as estimativas de variabilidade mais conhecidas, e são baseadas no quadrado dos desvios.

Variância	Desvio padrão
$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$	$s = \sqrt{s^2}$

- O desvio padrão, s, está na mesma escala dos dados originais.

- Dos estimadores acima, nenhum é robusto a outliers. A variância é mais sensível, pois usa o quadrado dos desvios.
- Uma estimativa robusta de variabilidade é o **desvio absoluto mediano da mediana** ou **MAD**:

$$\text{MAD} = \text{Mediana}(|x_1 - m|, |x_2 - m|, \ldots, |x_N - m|)$$

em que m é a mediana.

- MAD não é influenciado por valores extremos.
- As medidas de variabilidade apresentadas não são equivalentes:

$$s > \text{desvio absoluto da média} > MAD$$

Estimativas baseadas em percentis

- Estatísticas baseadas em dados ordenados (rankeados) são chamados de *estatística de ordem*.
- A medida mais comum é a *amplitude*: diferença entre o maior e o menor valor.
- A amplitude é extremamente sensível a outliers e não é uma medida de dispersão muito útil.
- Para evitar esta sensibilidade a outliers, podemos olhar para a amplitude depois de “remover” valores de cada extremidade. Isto é feito usando a diferença entre percentis.
- A mediana é o mesmo que o 50º percentil.
- Percentil é essencialmente o mesmo que quartil, com quantis indexados por frações (por exemplo de 10 em 10% ou 5 em 5%)

- Uma medida comum de variabilidade é a diferença entre o 25º percentil e o 75º percentil, chamado de **Intervalo Interquartil ou Amplitude Interquartil (IQR -** do inglês).