Análise de Dados Aplicada à Computação

ANÁLISE EXPLORATÓRIA DE DADOS

Prof. M.Sc. Howard Roatti

Análise de Dados 1/5

- **O que é Análise de Dados?**
  - É uma prática que envolve examinar informações para encontrar **padrões, relações e insights** que auxiliem a tomar decisões baseadas em dados.
  - Tem o **Big Data** como uma fonte de dados valiosa, porém **desestruturada**.
  - Para os negócios, as vantagens incluem: identificação de **padrões e tendências**, **tomada de decisões baseadas em dados** e melhoria de processos e desempenho.
  - A análise de dados **usa o raciocínio crítico sobre os dados**, sem viés, com objetivo de **extrair conhecimento**

Análise de Dados 2/5

- **Por que Analisar Dados?**
  - A análise de dados é essencial para a transformação digital e para o sucesso do negócio, pois permite:
  - Rastrear as atividades e resultados do negócio em áreas como marketing, vendas, produtividade e finanças.
  - Compreender o negócio a partir de diferentes perspectivas, possibilitando uma gestão mais informada e eficaz.
  - Extrair conhecimento profundo de cada setor do negócio por meio da análise crítica de dados.
  - Tomar decisões mais precisas e informadas, baseadas em insights obtidos por meio do processo de data mining.

Análise de Dados 3/5

- **Tipos de Análises de Dados:**
  - *Descritiva*: descreve os dados observados, de forma que eles "digam" o que aconteceu. **Estatística descritiva.**
    - Exemplo: descrever os acessos de um site de vendas, no intervalo de uma hora, pela média de usuários que visitaram o site.
  - *Diagnóstica*: Conhecidas as características dos dados, imediatamente vem a pergunta: **qual a causa desses dados?**
    - Exemplo: oferecer o programa de fidelidade para alguns usuários (cenário A), mas não para outros (cenário B), e comparar a média de acessos.

Para mais detalhes, acessem o link:
https://www.digitalhouse.com/br/blog/tipos-de-analise-de-dados/

Análise de Dados 4/5

- **Tipos de Análises de Dados:**
  - *Preditiva*: **fazer previsões**.
    - Exemplo: os bancos identificam riscos de investimento, quando o empréstimo efetivamente não foi feito, mas o modelo desenvolvido pelo banco é capaz de prever o resultado.

Para mais detalhes, acessem o link:
https://www.digitalhouse.com/br/blog/tipos-de-analise-de-dados/

Análise de Dados 5/5

- **Tipos de Análises de Dados:**
  - *Prescritiva*: Sabendo descrever os dados, entender o porquê de eles apresentarem tais características, criar uma representação abstrata deles, capaz de prever um cenário hipotético, é possível então pensar no processo de **introdução de objetivos a serem alcançados**.
    - Exemplo: o dos sistemas de controle da SpaceX, que analisam continuamente o movimento do foguete, para prescrever ações de seu motor. Este trabalho tem o objetivo de trazê-lo em segurança para o solo.

Essa é a etapa mais complexa, pois demanda conhecimento profundo sobre o sistema.

Para mais detalhes, acessem o link:
https://www.digitalhouse.com/br/blog/tipos-de-analise-de-dados/

Atividades Analíticas 1/4

Tarefa	Descrição Geral	Exemplo
Recuperar Valor	Dado um conjunto de casos específicos, encontrar atributos desses casos.	- Qual é a quilometragem por litro do Fiat Toro? - Quanto tempo dura o filme Liga da Justiça Snyder's Cut?
Filtrar	Dadas algumas condições concretas sobre os valores dos atributos, encontrar casos de dados que satisfaçam essas condições.	- Quais cereais Kellogg's têm alto teor de fibras? - Quais comédias ganharam prêmios? - Quais fundos tiveram desempenho inferior ao SP-500?
Calcular Valor Derivado	Dado um conjunto de casos de dados, calcular uma representação numérica agregada desses casos de dados.	- Qual é o conteúdo calórico médio dos cereais Post? - Qual é a receita bruta de todas as lojas combinadas? -Quantos fabricantes de automóveis existem?

https://pt.wikipedia.org/wiki/An%C3%A1lise_de_dados#Atividades_anal%C3%ADticas_de_usu%C3%A1rios_de_dados

Atividades Analíticas 2/4

Tarefa	Descrição Geral	Exemplo
Encontrar Extremo	Encontrar casos de dados que possuem um valor extremo de um atributo em seu intervalo no conjunto de dados.	- Qual é o carro com maior MPG? - Qual diretor/filme ganhou mais prêmios? - Qual filme da Marvel Studios tem a data de lançamento mais recente?
Ordenar	Dado um conjunto de casos de dados, ordená-los de acordo com alguma métrica ordinal.	- Ordenar os carros por peso. - Classificar os cereais por calorias.
Determinar Intervalo	Dado um conjunto de casos de dados e um atributo de interesse, encontrar o intervalo de valores dentro do conjunto.	- Qual é a gama de comprimentos de filme? - Qual é a faixa de potência do carro? - Quais atrizes estão no conjunto de dados?

https://pt.wikipedia.org/wiki/An%C3%A1lise_de_dados#Atividades_anal%C3%ADticas_de_usu%C3%A1rios_de_dados

Atividades Analíticas 3/4

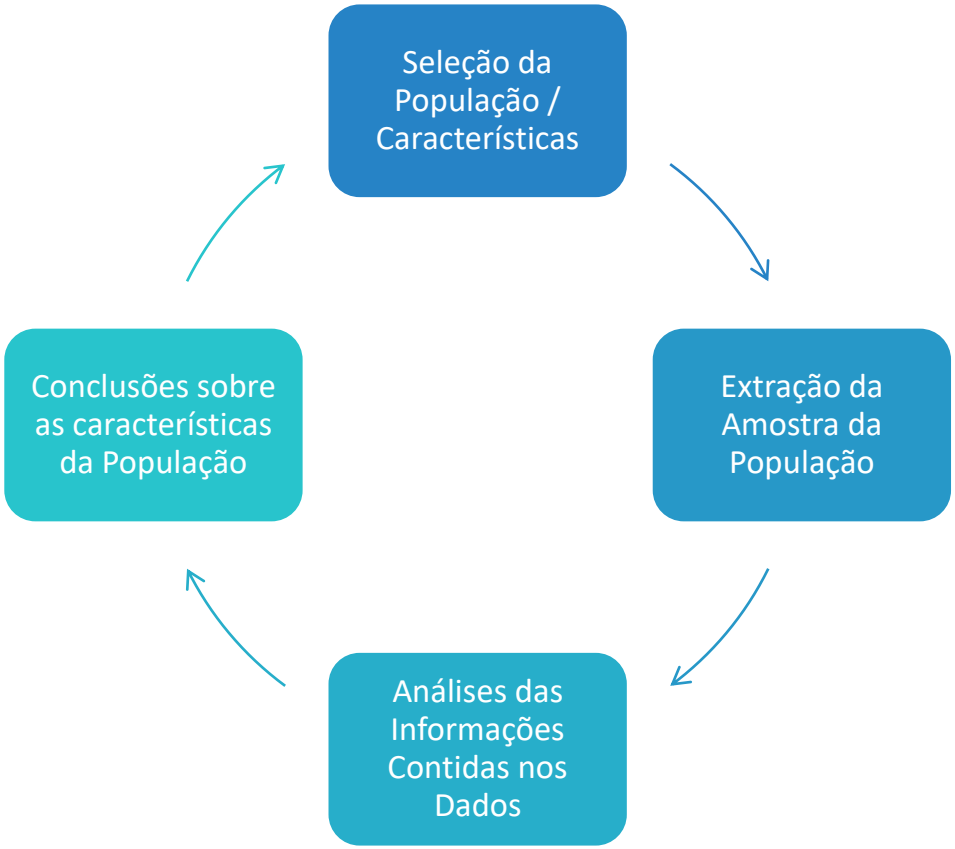
Tarefa	Descrição Geral	Exemplo
Caracterizar Distribuição	Dado um conjunto de casos de dados e um atributo quantitativo de interesse, caracterizar a distribuição dos valores desse atributo ao longo do conjunto.	- Qual é a distribuição dos carboidratos nos cereais? - Qual é a distribuição da idade dos compradores?
Encontrar Anomalias	Identificar quaisquer anomalias dentro de um determinado conjunto de casos de dados com respeito a um determinado relacionamento ou expectativa, por exemplo, *outliers* estatísticos.	- Existem exceções para a relação entre potência e aceleração? - Existem outliers na proteína?

https://pt.wikipedia.org/wiki/An%C3%A1lise_de_dados#Atividades_anal%C3%ADticas_de_usu%C3%A1rios_de_dados

Atividades Analíticas 4/4

Tarefa	Descrição Geral	Exemplo
Agrupar	Dado um conjunto de casos de dados, encontrar grupos de valores de atributos semelhantes.	- Existem grupos de cereais com gordura/calorias/açúcar semelhantes? - Existe um grupo de durações de filme típicas?
Correlacionar	Dado um conjunto de casos de dados e dois atributos, determinar relações úteis entre os valores desses atributos.	- Existe correlação entre carboidratos e gordura? - Existe correlação entre país de origem e MPG? - Os diferentes gêneros têm um método de pagamento preferido? - Existe uma tendência de aumento da duração dos filmes ao longo dos anos?
Contextualizar	Dado um conjunto de casos de dados, encontrar a relevância contextual dos dados para os usuários.	- Existem grupos de restaurantes que oferecem alimentos com base na minha ingestão calórica atual?

Estruturando uma Análise 1/2



Etapas da Análise Estatística

Estruturando uma Análise 2/2

- Coletar dados;
- Agrupar dados por tópicos;
- Utilizar uma ferramenta para procurar por correlações nos dados;
- Montar gráficos com informações encontradas pela correlação;
- Interpretar os gráficos de acordo com o sistema analisado;
- Montar um relatório de análise.

Análise Efetiva 1/3

- Comece planejando: Essa etapa é fundamental para que as informações colhidas sejam úteis e, principalmente, para que o procedimento possa ser finalizado com sucesso. Lembre-se de que, sem planejamento, você pode se perder durante o processo.
- Defina o foco de sua análise: Se você pretende ter uma previsão de vendas, não adianta analisar dados de RH, não é mesmo?

Análise Efetiva 2/3

- Escolha as hipóteses e perguntas que serão trabalhadas:
  - Determinado produto está tendo uma boa aceitação no mercado?
  - A renda obtida com as vendas é suficiente para cobrir os custos?
  - Qual é o estoque necessário para uma data sazonal?
  - Quais produtos precisam de um trabalho de marketing mais efetivo?
  - Colete os dados e faça uma boa análise.

Análise Efetiva 3/3

- Tire as suas conclusões e predições:
  - Nessa etapa, é interessante trabalhar os dados em diversos cenários e, se necessário, coletar novas informações. O importante é chegar à melhor resposta para as suas dúvidas, ajudando-o a tomar a decisão mais vantajosa.

Elementos de Dados Estruturados

- A origem dos dados: Texto, IoT, Web, Sistemas Transacionais
- Geralmente sem estrutura formal(Tabular): Imagens (RGB), Textos, Clickstreams
- Um dos maiores desafios da Ciência de Dados: transformar esses dados em **Informação**
- Para aplicar conceitos estatísticos, **dados não estruturados** precisam ser manipulados e processados para se tornarem **dados estruturados**

Tipos de Dados

- **Contínuo:** Pode assumir qualquer valor em um intervalo (velocidade, duração)
- **Discreto:** Só pode assumir números inteiros (contagens)
- **Categórico:** Apenas um conjunto de valores possíveis (marca de celulares, estados de um país)
- **Binário:** Apenas dois valores (Verdadeiro ou Falso. 0 ou 1. Sim ou Não)
- **Ordinal:** Categórico mas com ordenação explícita (Classificação 1-5 estrelas)

Estimativas de Localização

Variáveis numéricas ou de contagem podem ter centenas de valores distintos. Um passo básico na exploração de dados é encontrar o valor típico para cada característica

Estimativas de Localização

- **Média**: A soma dos valores, dividida pelo número de valores.
- **Média Ponderada**: A soma dos valores, multiplicada por um peso e dividida pela soma dos pesos.
- **Mediana**: O valor que ocupa a posição central dos dados.
- **Mediana Ponderada**: Valor cuja posição está no centro da soma dos pesos, estando metade da soma antes e metade depois desse dado.
- **Média Aparada**: A média dos valores depois da exclusão de um número fixo de valores extremos.
- **Robustez**: Não sensível a valores extremos ou anômalos.
- ***Outlier***: valores anômalos.

Estimativas de Localização

- Apesar de fácil de computar e entender, a média não é sempre a melhor medida para encontrar o valor central.
- Valores calculados são chamados pelos **estatísticos** de **estimativas**
- Entretanto os **cientistas de dados** chamam de **métricas**

Média

$$\text{média} = \bar{x} = \frac{\sum_i^n x_i}{n}$$

Em Estatística **n** e **N** possuem significados diferentes. **N** se refere a **população**. **n** uma **amostra** da população. Em *Data Science* tanto faz

Média Aparada

$$\text{médiaAparada} = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n - 2p}$$

- A média aparada elimina a influência de valores extremos.
- Exemplo: nos desfiles das escolas de sambas, as maiores e menores notas são descartadas
- Isso dificulta a manipulação do placar por um único jurado

Média Ponderada

$$\text{médiaPonderada} = \overline{x_w} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i}^{n} w_i}$$

- Alguns valores podem ser menos importantes que outros, por exemplo, em se tratando de sensores em que um deles esteja descalibrado, então devemos dar uma menor importância para ele
- Dados coletados podem não representar igualmente diferentes grupos, por exemplo, um grupo de usuários que não foi representado adequadamente poderia receber um peso maior para que sua representação seja coerente.

Mediana e Robustez

- A mediana é o valor do meio de uma lista ordenada
  - Se houver uma quantidade par de números, a mediana é a média dos dois valores que dividem os dados
- A mediana é considerada robusta a *outliers*
- Ser um *outlier* não torna o dado inválido, mas em muitos casos eles aparecem por erro de conversão ou medição
  - *Outliers* devem ser identificados e valem investigações

Média Aparada

- Também considerada como uma métrica robusta
- Normalmente são eliminados 10% dos valores extremos, mas pode variar de acordo com a percepção do analista
- Existem outras métricas para localização que podem ser mais eficientes que a média e úteis para pequenos dados, mas não trazem muitos benefícios para bases médias ou maiores

Exemplos de Estimativa de Localização

- Considere o dataset state.csv compartilhado anteriormente...

```
import pandas as pd
import numpy as np
from scipy.stats import trim_mean
import wquantiles

state = pd.read_csv(STATE_CSV)
print(state['Population'].mean())
print(trim_mean(state['Population'], 0.1))
print(state['Population'].median())
print(np.average(state['Murder.Rate'],
                 weights=state['Population']))
print(wquantiles.median(state['Murder.Rate'],
                        weights=state['Population']))
```

Estimativas de Variabilidade

Uma das maneiras de se resumir os dados é utilizando a localização para uma característica. Outra maneira é a variabilidade ou dispersão. Ela mede o quanto os dados estão agrupados ou espalhados.

Estimativas de Variabilidade

- **Desvio:** A diferença entre valores observados e a estimativa de localização
- **Variância:** A soma dos quadrados dos desvios da média, divididos por n-1
- **Desvio-Padrão:** A raiz quadrada da variância
- **Desvio Absoluto Médio:** A média do valor absoluto dos desvios da média
- **Desvio absoluto mediano da mediana:** a mediana do valor absoluto dos desvios da mediana

Estimativas de Variabilidade

- **Amplitude:** A diferença do maior e o menor valor
- **Estatísticas Ordinais:** Métricas baseadas nos valores ordenados do menor para o maior
- **Percentil:** Valor tal que P por cento dos valores assuma esse valor ou menos, e (100-p) são maiores ou iguais
- **Amplitude Iterquartil (IQR):** A diferença entre o 75º percentil e o 25º percentil.

Desvio Padrão e Estimativas Relacionadas

- As estimativas de variação mais usadas são baseadas nas diferenças, ou desvios, entre estimativas de localização e os valores observados
- Os dados **[1,4,4]** possuem **média 3** e **mediana 4**
- O desvio da média é **1 - 3 = -2**, **4 – 3 = 1**, **4 – 3 = 1**
- Indicam o quão dispersos do centro os valores são
- Se calcularmos a média, ao somar os negativos anulariam os positivos, por isso devemos utilizar os valores absolutos: **(2 + 1 + 1) / 3 = 1.33**

Desvio Padrão e Estimativas Relacionadas

$$\text{Desvio Absoluto Médio} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

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

$$Desvio\ Padrão = s = \sqrt{variância}$$

- As melhores estimativas de variabilidade são a variância e o desvio padrão
- O desvio padrão é mais fácil de interpretar, uma vez que está na mesma escala que os dados originais

MAD
Desvio Absoluto Mediano

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

- **Onde m é a mediana**
- **Nem a variância nem o desvio padrão são robustos**

Graus de Liberdade

- Uma forma de entender os graus de liberdade na estatística é pensar em um quebra-cabeça. Imagine que você tem um quebra-cabeça com 10 peças e cada peça tem um número de 1 a 10. Você sabe que a soma dos números das peças deve ser 55. Agora, se você colocar as peças em qualquer ordem, você pode escolher livremente 9 delas, mas a última peça vai depender das outras 9 para que a soma seja 55. Por exemplo, se as primeiras 9 peças somam 50, a última peça tem que ser o número 5. Então, você só tem 9 graus de liberdade para escolher as peças do quebra-cabeça.

Graus de Liberdade

- Os graus de liberdade na estatística são parecidos com isso. Eles representam o número de valores nos dados que podem variar livremente sem violar alguma restrição ou condição. Por exemplo, se você quer calcular a média de uma amostra de dados, você precisa somar todos os valores e dividir pelo número de valores. Mas se você já sabe a média e o número de valores, então você só pode escolher livremente n-1 valores da amostra, onde n é o número de valores. O último valor vai depender dos outros n-1 para que a média seja igual àquela que você já sabe. Então, nesse caso, os graus de liberdade são n-1.

Graus de Liberdade

- Os graus de liberdade são importantes na estatística porque eles afetam como nós calculamos e interpretamos os testes estatísticos e as medidas de variabilidade dos dados. Eles também nos ajudam a avaliar o quanto os nossos resultados são confiáveis e precisos.

Estimativas com Percentis

- A medida mais simples é a amplitude (Max - Min), mas é extremamente sensível a outliers
- Para dar robustez podemos olhar o alcance após remover valores dos extremos
- Essas estimativas são baseadas em percentis diferentes
- Pegamos a diferença entre os percentis
- Uma medida comum é o IQR onde pegamos a distancia entre o 25º e o 75º
- Para uma massa muito grande de dados percentis exatos podem ser computacionalmente caros

Exemplos de Estimativa de Variabilidade

- Considere o dataset state.csv compartilhado anteriormente...

```
print(state['Population'].std()) #Desvio Padrão

print(state['Population'].quantile(0.75) –
      state['Population'].quantile(0.25)) #IQR

#Desvio Absoluto da Mediana
print(robust.scale.mad(state['Population']))
```

- A variância e desvio padrão são as métricas de variabilidade mais utilizadas
- As duas são sensíveis a outliers!
- Métricas mais robustas incluem média e mediana dos desvios absolutos da média e percentil

Exemplos de Conclusão

- Um exemplo de conclusão que pode ser tirada a partir das estimativas de localização e variabilidade é comparar a distribuição de alturas de pessoas em diferentes países. Para isso, podemos usar o Python e um dataset conhecido chamado Gapminder, que contém dados sobre vários indicadores socioeconômicos de países do mundo.
- Uma forma de fazer isso é importar o dataset usando a biblioteca pandas e calcular a média e o desvio padrão das alturas por país usando a biblioteca numpy. A média é uma medida de localização que indica o valor central dos dados, e o desvio padrão é uma medida de variabilidade que indica o quanto os dados se afastam da média. Por exemplo:

Exemplos de Conclusão

```
import pandas as pd
import numpy as np

df = pd.read_csv("gapminder.csv")
df = df[["country", "height"]]

# Agrupando os dados por país
df = df.groupby("country").agg({"height": [np.mean, np.std]})

# Renomeando as colunas
df.columns = ["mean_height", "std_height"]

# Ordenando os dados pela média das alturas em ordem decrescente
df = df.sort_values(by="mean_height", ascending=False)

# Mostrando os primeiros 10 países
print(df.head(10))
```

Exemplos de Conclusão

A partir desses resultados, podemos concluir que:

- O país com a maior média de altura é a Holanda (Netherlands), com cerca de **182 cm**.
- O país com o menor desvio padrão de altura é a Eslovênia (Slovenia), com cerca de **5 cm**.
- Os países com as maiores alturas estão concentrados na Europa, especialmente na região norte e central.
- Há uma grande variabilidade nas alturas dos países, indicando que há outros fatores que influenciam esse indicador além da localização geográfica.

Exemplos de Conclusão

```
import matplotlib.pyplot as plt

# Criando uma figura e um eixo
fig, ax = plt.subplots()

# Definindo o tamanho da figura
fig.set_size_inches(10, 6)

# Plotando o gráfico de barras
ax.bar(df.index, df["mean_height"])

# Rotacionando os nomes dos países
plt.xticks(rotation=90)

# Adicionando um título ao gráfico
plt.title("Média das alturas por país“)
# Mostrando o gráfico na tela
plt.show()
```

Referências

- Bruce, P.; Bruce, A.; **Estatística Prática para Cientista de Dados**: 50 Conceitos Essenciais; Rio de Janeiro; Alta Books; 2019.
- https://pt.wikipedia.org/wiki/An%C3%A1lise_de_dados
- https://www.voitto.com.br/blog/artigo/analise-de-dados
- https://rockcontent.com/br/blog/tipos-de-analise-de-dados/
- https://www.digitalhouse.com/br/blog/tipos-de-analise-de-dados/

FAESA
CENTRO UNIVERSITÁRIO

Análise de Dados Aplicada à Computação

PROF. M.SC HOWARD ROATTI