Análise de Dados Aplicada à Computação

ANÁLISE EXPLORATÓRIA DE DADOS

Prof. M.Sc. Howard Roatti

Explorando a Distribuição de Dados

- **Boxplot**: um modo rápido de visualizar a distribuição de dados
- **Tabela de Frequências**: Contagem de valores dentro de um intervalo
- **Histograma**: visualização da tabela de frequências
- **Gráfico de Densidade**: versão simplificada do histograma

Percentil e Boxplot

- Em “Estimativas Baseadas em Percentis”, explora-se os percentis para medir a dispersão dos dados
- Os percentis são valiosos para resumir toda a distribuição
- É comum registrar os quartis (25º, 50º e 75º percentis) e os decis (10º, 20º, ..., 90º percentis).

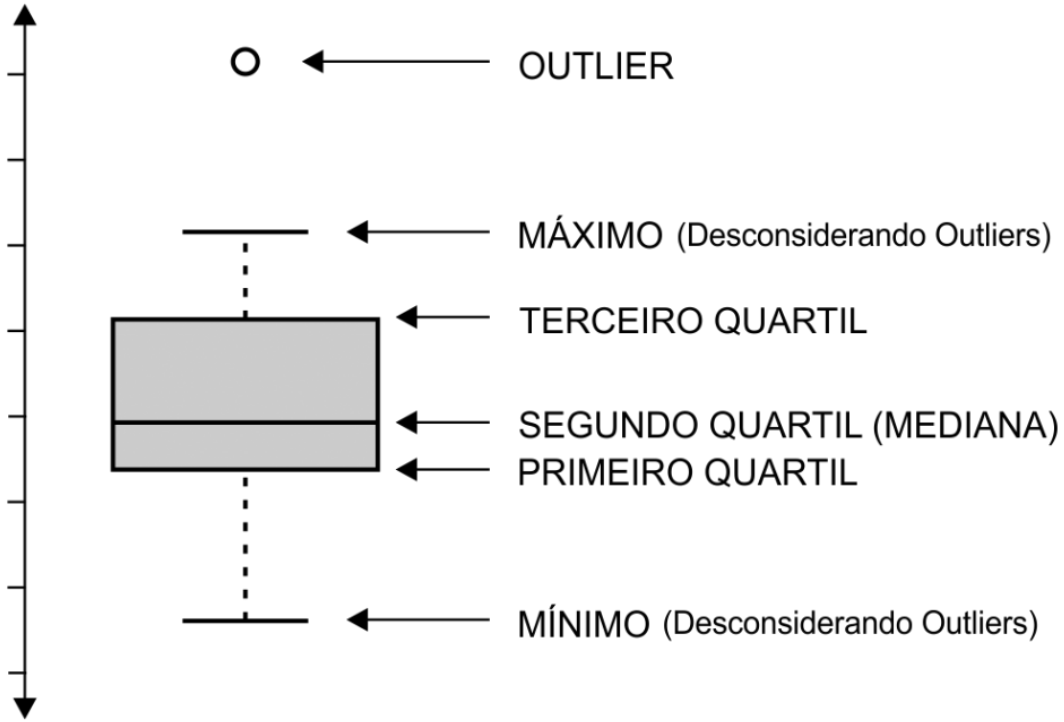
5%	25%	50%	75%	95%
1.60	2.42	4.00	5.55	6.51

- Percentis das Taxas de Homicídios por Estado

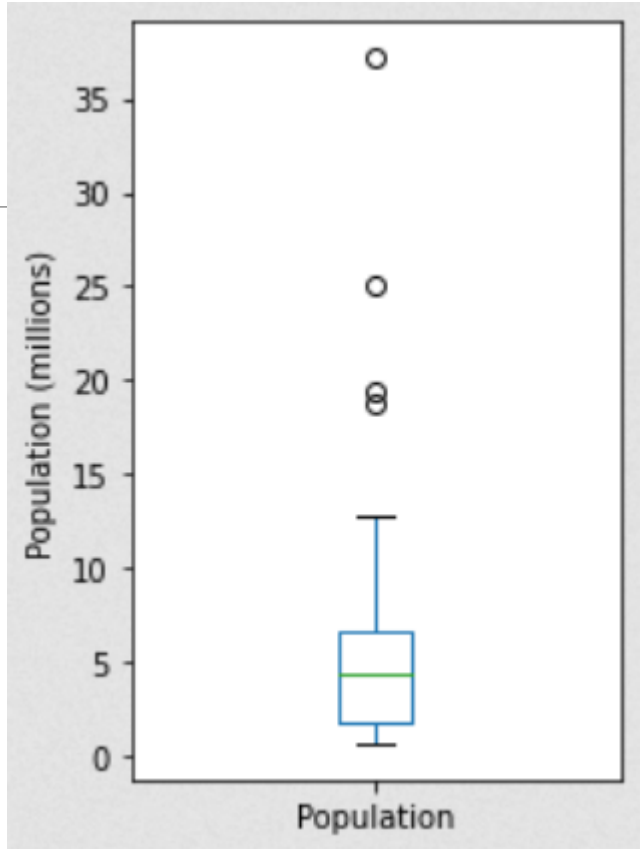
```
print(state['Murder.Rate'].quantile([0.05, 0.25, 0.5, 0.75, 0.95]))
```

Percentil e Boxplot

- Boxplot oferece uma forma de visualizar a distribuição de dados utilizando percentil



Percentil e Boxplot




```
ax = (state['Population']/1_000_000).plot.box(figsize=(3, 4))
ax.set_ylabel('Population (millions)')

plt.tight_layout()
plt.show()
```

Tabelas de Frequências e Histogramas

- O maior interesse é conhecer o comportamento da variável, observando as ocorrências para se ter uma ideia global de sua distribuição.
- Tomando como base a tabela a seguir:

Tabela 2.1 Informações sobre estado civil, grau de instrução, número de filhos, salário (expresso como fração do salário mínimo), idade (medida em anos e meses) e procedência de 36 empregados da seção de orçamentos da Companhia MB.

Nº	Estado civil	Grau de instrução	Nº de filhos	Salário (x sal. mín.)	Idade		Região de procedência
					anos	meses	
1	solteiro	ensino fundamental	–	4,00	26	03	interior
2	casado	ensino fundamental	1	4,56	32	10	capital
3	casado	ensino fundamental	2	5,25	36	05	capital
4	solteiro	ensino médio	–	5,73	20	10	outra
5	solteiro	ensino fundamental	–	6,26	40	07	outra
6	casado	ensino fundamental	0	6,66	28	00	interior
7	solteiro	ensino fundamental	–	6,86	41	00	interior
8	solteiro	ensino fundamental	–	7,39	43	04	capital
9	casado	ensino médio	1	7,59	34	10	capital
10	solteiro	ensino médio	–	7,44	23	06	outra
11	casado	ensino médio	2	8,12	33	06	interior
12	solteiro	ensino fundamental	–	8,46	27	11	capital
13	solteiro	ensino médio	–	8,74	37	05	outra

Tabelas de Frequências e Histogramas

14	casado	ensino fundamental	3	8,95	44	02	outra
15	casado	ensino médio	0	9,13	30	05	interior
16	solteiro	ensino médio	—	9,35	38	08	outra
17	casado	ensino médio	1	9,77	31	07	capital
18	casado	ensino fundamental	2	9,80	39	07	outra
19	solteiro	superior	—	10,53	25	08	interior
20	solteiro	ensino médio	—	10,76	37	04	interior
21	casado	ensino médio	1	11,06	30	09	outra
22	solteiro	ensino médio	—	11,59	34	02	capital
23	solteiro	ensino fundamental	—	12,00	41	00	outra
24	casado	superior	0	12,79	26	01	outra
25	casado	ensino médio	2	13,23	32	05	interior
26	casado	ensino médio	2	13,60	35	00	outra
27	solteiro	ensino fundamental	—	13,85	46	07	outra
28	casado	ensino médio	0	14,69	29	08	interior
29	casado	ensino médio	5	14,71	40	06	interior
30	casado	ensino médio	2	15,99	35	10	capital
31	solteiro	superior	—	16,22	31	05	outra
32	casado	ensino médio	1	16,61	36	04	interior
33	casado	superior	3	17,26	43	07	capital
34	solteiro	superior	—	18,75	33	07	capital
35	casado	ensino médio	2	19,40	48	11	capital
36	casado	superior	3	23,30	42	02	interior

Fonte: Dados hipotéticos.

Tabelas de Frequências e Histogramas

- Foi desenvolvida a tabela de frequências a seguir:

Exemplo 2.2 A Tabela 2.2 apresenta a *distribuição de frequências* da variável grau de instrução, usando os dados da Tabela 2.1.

Tabela 2.2 Frequências e porcentagens dos 36 empregados da seção de orçamentos da Companhia MB segundo o grau de instrução.

Grau de instrução	Frequência n_i	Proporção f_i	Porcentagem $100\,f_i$
Fundamental	12	0,3333	33,33
Médio	18	0,5000	50,00
Superior	6	0,1667	16,67
Total	36	1,0000	100,00

Fonte: Tabela 2.1.

Tabelas de Frequências e Histogramas

- Para construção de tabela de frequência para variáveis contínuas, como salário, a solução é agrupar por faixas de salário.

Exemplo 2.3 A Tabela 2.4 dá a distribuição de frequências dos salários dos 36 empregados da seção de orçamentos da Companhia MB por faixa de salários.

Tabela 2.4 Frequências e porcentagens dos 36 empregados da seção de orçamentos da Companhia MB por faixa de salário.

Classe de salários	Frequência n_i	Porcentagem $100\,f_i$
4,00 ⊢ 8,00	10	27,78
8,00 ⊢ 12,00	12	33,33
12,00 ⊢ 16,00	8	22,22
16,00 ⊢ 20,00	5	13,89
20,00 ⊢ 24,00	1	2,78
Total	36	100,00

Fonte: Tabela 2.1.

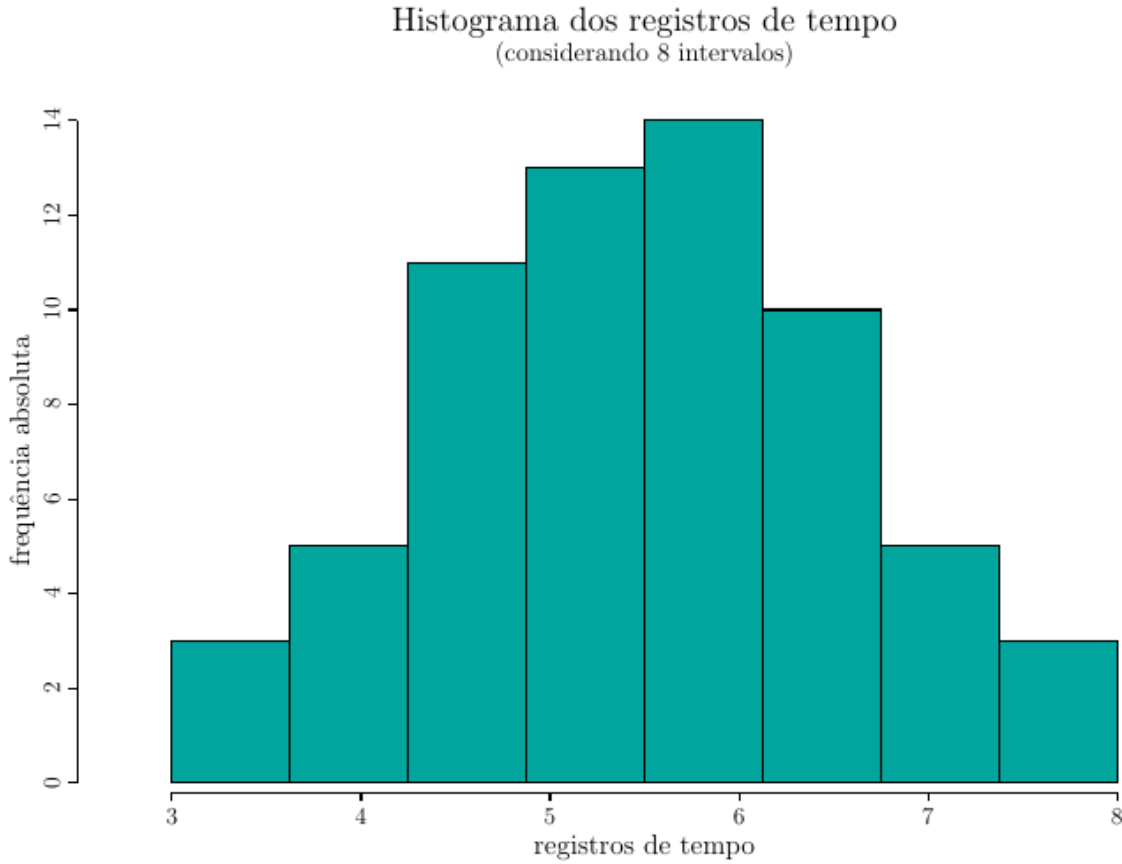
Tabelas de Frequências e Histogramas

- Seguindo o *Data Set state*, podemos desenvolver uma tabela de frequências no R seguindo os passos a seguir:

```
binnedPopulation = pd.cut(state['Population'], 10)
print(binnedPopulation.value_counts())
```

Tabelas de Frequências e Histogramas

- Um Histograma é uma maneira de visualizar uma tabela de frequências
- O usuário define o número de barras
- As barras possuem a mesma largura
- Barras vazias são incluídas no gráfico
- As barras são contínuas, ou seja, sem espaços entre elas, a menos que seja uma coluna vazia



Tabelas de Frequências e Histogramas

- Podemos criar um Histograma no Python utilizando a instrução a seguir:

```
ax = (state['Population'] / 1_000_000).plot.hist(figsize=(4, 4))
ax.set_xlabel('Population (millions)')

plt.tight_layout()
plt.show()
```

Estimativas de Densidade

- Relacionado ao histograma é o diagrama de densidade
- Mostra a distribuição dos valores como uma linha contínua
- Pode ser pensado como uma versão atenuada do histograma
- Uma diferença é a escala de Y, que é uma proporção ao invés de contagem
- Para Data Science a estimativa padrão é suficiente

Estimativas de Densidade

- Em Python, podemos reproduzir assim:

```
ax = state['Murder.Rate'].plot.hist(density=True, xlim=[0, 12],
                                    bins=range(1,12), figsize=(4, 4))
state['Murder.Rate'].plot.density(ax=ax)
ax.set_xlabel('Murder Rate (per 100,000)')

plt.tight_layout()
plt.show()
```

Estimativas de Densidade

- **Histogramas** mostram a frequência de contagem dos valores em Y e os valores em X
- **Tabela de frequência** é uma versão em tabela do histograma
- **Boxplot ou diagrama em caixa** tem como topo e fundo da caixa os percentis 25% e 75%. Normalmente é usado para comparar distribuições
- **Gráfico de Densidade** é uma versão alisada do histograma e requer uma função de densidade

Explorando Dados Binários e Categóricos

- **Moda:** Categoria ou valor de maior ocorrência
- **Gráfico de Barra:** a frequência ou proporção mostrados como barras
- **Gráfico de Pizza:** a frequência ou proporção mostrados como fatias de pizza

Gráfico de Barra

- Muito usado para visualizar variáveis categóricas
- Categorias ficam no eixo X e a frequência ou proporção no eixo Y
- Histograma possui valores de uma variável na escala numérica no eixo X
- Enquanto o gráfico de barra apresenta diferentes categorias
- No histograma as barras 'encostam‘
- Gráficos de Pizza também são usados como alternativa da barra, mas pouco usado por experts

Gráfico de Barra

- Podemos plotar um gráfico de barras utilizando o comando em Python a seguir:

```
ax = dfw.transpose().plot.bar(figsize=(4, 4), legend=False)
ax.set_xlabel('Cause of delay')
ax.set_ylabel('Count')

plt.tight_layout()
plt.show()
```

Correlação

- A correlação trata da interdependência de duas ou mais variáveis
- O objetivo do estudo da **correlação** é determinar (mensurar) o grau de relacionamento entre duas variáveis
- Quanto maior for o valor absoluto do coeficiente, mais forte é a relação entre as variáveis.

Correlação

- Na análise exploratória de dados muitas vezes temos que examinar a correlação entre preditores e entre preditores e a variável alvo
- Duas variáveis X e Y possuem correlação positiva se valores altos de X vão com altos de Y e baixos de X com baixos de Y
- Se valores altos de X vão com valores baixos de Y e vice-versa, as variáveis são negativamente correlacionadas

Correlação

- **Coeficiente de Correlação**, métrica que mede a associação entre variáveis, vai de -1 a 1
- **Matriz de Correlação**, uma tabela onde as variáveis são mostradas nas colunas e nas linhas, a célula contém a correlação entre as variáveis
- **Diagrama de Dispersão**, um gráfico onde o eixo X mostra os valores de uma variável e o eixo Y os valores de outra

Coeficiente de Correlação de Pearson

$$r = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

- Variáveis podem ter associações não lineares, nesse caso o coeficiente pode não ser uma métrica útil
- Exemplo: com o aumento de imposto crescendo de 0, o valor arrecadado também cresce
- Mas quando o imposto fica alto e se aproxima de 100% sobe a sonegação de impostos

Matriz de Correlação

	T	CTL	FTR	VZ	LVLT
T	1.000	0.475	0.328	0.678	0.279
CTL	0.475	1.000	0.420	0.417	0.287
FTR	0.328	.0420	1.000	0.287	0.260
VZ	0.678	0.417	0.287	1.000	0.242
LVLT	0.279	0.287	0.260	0.242	1.000

- Uma tabela de correlação expõe visualmente o relacionamento entre múltiplas variáveis
- Aqui vemos a correlação entre os maiores fundos diários (*Exchange Traded Funds – ETFS*) da SP500.

Coeficiente de Correlação de Pearson

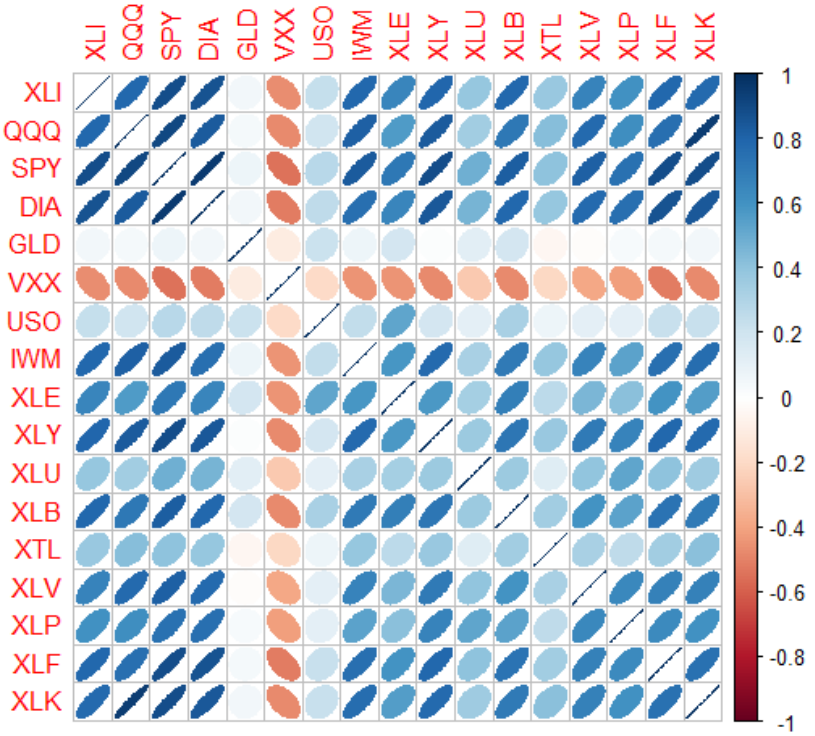
```
etfs = sp500_px.loc[sp500_px.index > '2012-07-01',
                    sp500_sym[sp500_sym['sector'] == 'etf']['symbol']]

from matplotlib.collections import EllipseCollection
from matplotlib.colors import Normalize

m = plot_corr_ellipses(etfs.corr(), figsize=(5, 4), cmap='bwr_r')
cb = fig.colorbar(m)
cb.set_label('Correlation coefficient')

plt.tight_layout()
plt.show()
```

Coeficiente de Correlação de Pearson



- Os ETFs da S&P500 (SPY) e o índice Down Jones(DIA) têm alta correlação
- QQQ e XLK são empresas tecnológicas relacionadas
- GLD(Ouro), USO(Petróleo) e VXX(Volatilidade do Mercado) tendem a ter relacionamentos negativos com outras.

Gráficos de Dispersão

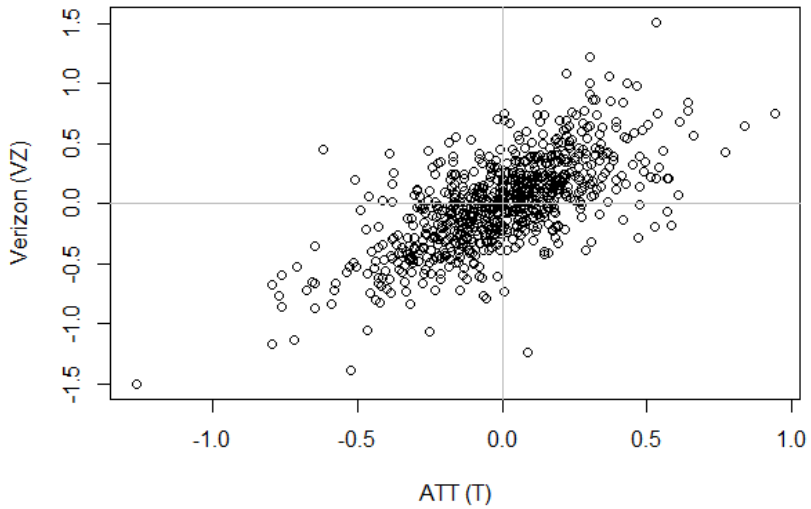
- O modo de visualizar a relação entre duas variáveis é através dos gráficos de dispersão
- O eixo X representa a variável, o eixo y outra
- Cada ponto no gráfico é um registro
- O gráfico a seguir mostra os retornos diários da Verizon e ATT, ambas de telecomunicação

Gráficos de Dispersão

```
ax = telecom.plot.scatter(x='T', y='VZ', figsize=(4, 4),
marker='$\u25EF$')
ax.set_xlabel('ATT (T)')
ax.set_ylabel('Verizon (VZ)')
ax.axhline(0, color='grey', lw=1)
ax.axvline(0, color='grey', lw=1)

plt.tight_layout()
plt.show()
```

Gráficos de Dispersão



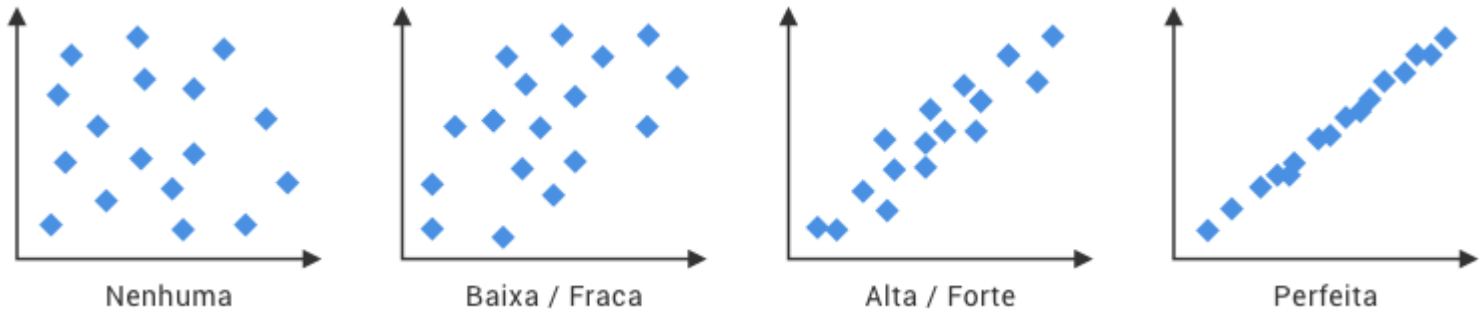
- Na maior parte do dia, os retornos têm uma forte relação positiva: na maioria dos dias as duas ações sobem ou descem paralelamente.
- Poucos são os casos em que uma ação cai significativamente e a outra ação sobe (ou vice-versa)

Correlação

- O Coeficiente de correlação mede a associação entre duas variáveis
- Podem ser positivamente associado o negativamente
- Os valores variam de -1 a 1
- Um coeficiente de zero indica que não há correlação

Correlação

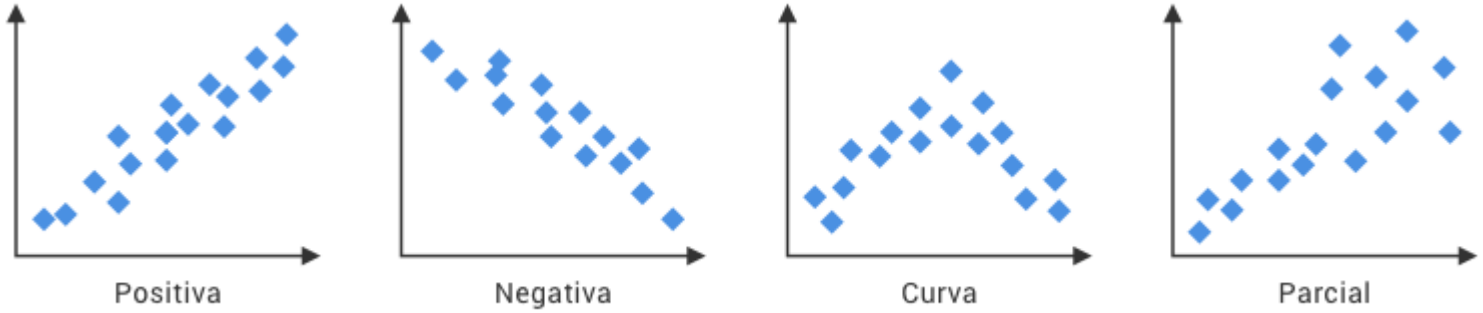
- **Níveis de Correlação:**



• **Nenhuma**: Nessa situacão, as 2 variaveis que estão sendo analisadas não possuem nenhuma correlação.
• **Alta / Forte:** Quanto menor for a dispersão dos pontos (ou seja, mais próximos de uma reta), maior será a correlação entre os dados.
• **Baixa / Fraca:** Então quanto maior for a dispersão dos pontos, menor será o grau de correlação entre os dados, ou seja, eles quase não possuem uma correlação.
• **Perfeita:** A correlação é perfeita quando não há uma grande dispersão entre os pontos, a correlação será total entre os dados, independente da tendência, seja ela positiva ou negativa.

Correlação

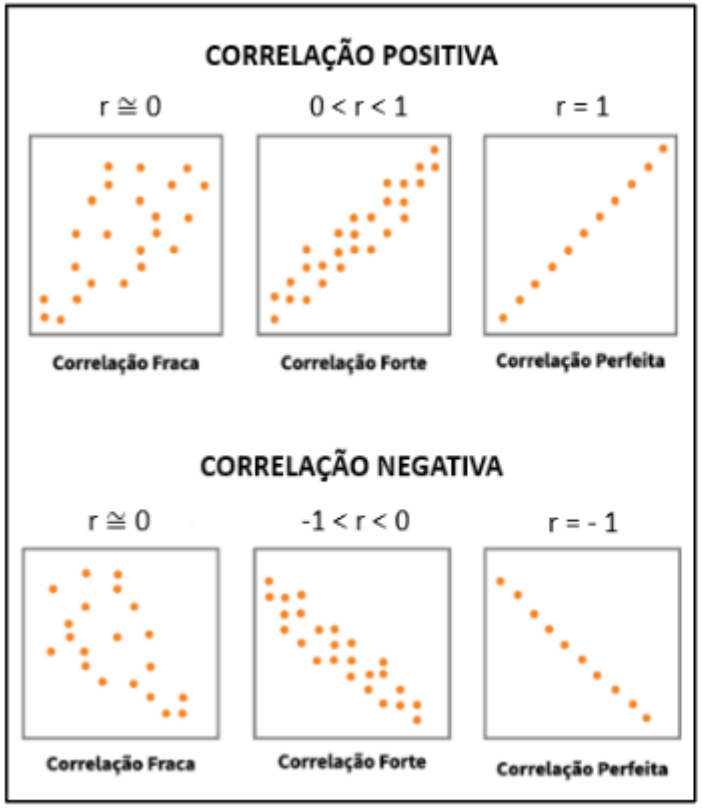
- **Tipos de Correlação:**



• **Positiva:** Este tipo de correlação acontece quando há uma tendência crescente entre os pontos. Conforme uma variável aumenta, a outra variável também aumenta proporcionalmente.
• **Negativa:** Já essa correlação é quando se concentram em uma linha decrescente. Conforme uma variável aumenta, a outra diminui
• **Curva:** É uma combinação de correlação positiva e negativa. Ela se dá quando em algum momento, a correlação entre as variáveis começa a se tornar contraria
• **Parcial:** A correlação parcial indica que até determinado momento entre as variáveis, a correlação é positiva ou negativa, mas que após um ponto máximo a correlação se perde.

Correlação

- **Tipos Vs Níveis:**



O tipo de correlação pode ser visualizado a partir do **Coeficiente de Correlação** (ou Coeficiente de Pearson - r). Os valores obtidos neste coeficiente relacionam-se da seguinte forma:

r = 0: Correlação nula ou inexistente entre variáveis.
r = 1: Correlação positiva entre variáveis.
r = -1: Correlação Negativa entre variáveis.

Correlação Positiva:
No qual as duas variáveis crescem no mesmo sentido. Ou seja, enquanto um aumenta, o outro também aumenta.

Correlação negativa:
No qual as duas variáveis variam em sentidos contrários. Ou seja, enquanto um aumenta, o outro diminui

Correlação Nula:
Não há interação entre variáveis.

Explorando Duas ou Mais Variáveis

- Estimativas como média e variância olham apenas uma variável por vez
- Correlação é um método importante para comparar duas variáveis

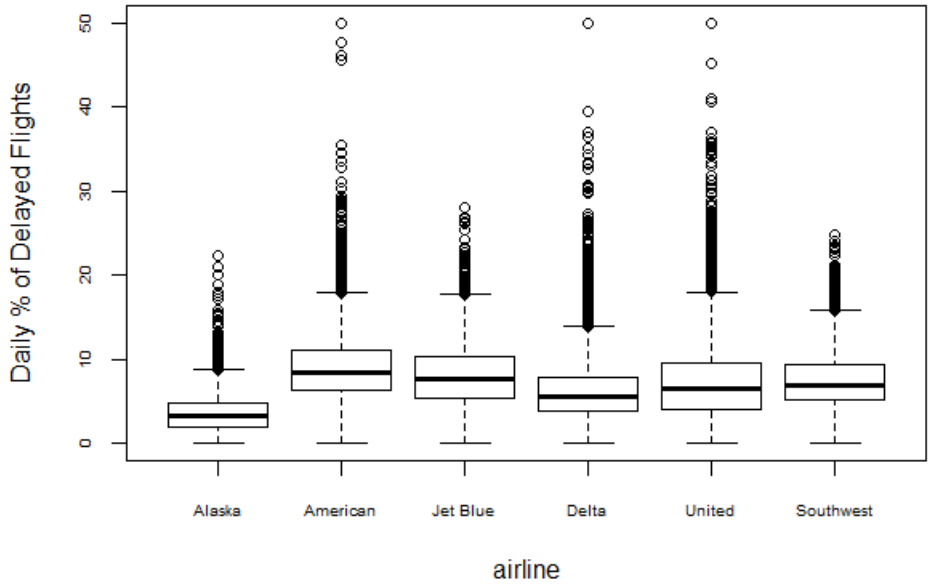
Dados Categóricos e Numéricos

- Boxplot é uma forma simples de visualizar a distribuição de dados
- Podemos querer comparar como os atrasos variam dependendo da empresa aérea

```
airline_stats = pd.read_csv(AIRLINE_STATS_CSV)
airline_stats.head()
ax = airline_stats.boxplot(by='airline', column='pct_carrier_delay',
                           figsize=(5, 5))
ax.set_xlabel('')
ax.set_ylabel('Daily % of Delayed Flights')
plt.suptitle('')

plt.tight_layout()
plt.show()
```

Dados Categóricos e Numéricos



- A Alaska se destaca por ter menos atrasos, enquanto a American tem a maioria deles: o quartil inferior da American é maior que o quartil superior da Alaska.

Visualizando Múltiplas Variáveis

- Os gráficos usados para duas variáveis (dispersão, hexagonal e boxplot) podem ser estendidos a várias variáveis usando a noção de condicionamento
- Lembram do gráfico de hexagonais? Podemos quebrar eles por zip code (cep) e identificar as diferenças

Análise Exploratória de Dados

- O ponto central da Análise Exploratória de Dados é que primeiro devemos olhar para os dados.
- Através da sumarização/visualização podemos ter insights e compreender o valor dos projetos.

Exemplos de Conclusão

Boxplot: Esse gráfico permite ver a localização, a variabilidade e a simetria dos dados, bem como identificar possíveis outliers ou valores extremos. Por exemplo, podemos usar um boxplot para comparar as notas de diferentes disciplinas em uma escola e ver qual disciplina tem a maior ou menor média, qual disciplina tem a maior ou menor variância e qual disciplina tem mais ou menos outliers nas notas.

Exemplos de Conclusão

Histograma: Esse gráfico permite ver a frequência com que cada valor ou intervalo de valores aparece nos dados, bem como identificar possíveis modas ou picos na distribuição. Por exemplo, podemos usar um histograma para analisar a idade dos pacientes com câncer de mama e ver se há uma faixa etária mais prevalente ou se há uma distribuição bimodal ou multimodal na incidência da doença.

Exemplos de Conclusão

Gráfico de densidade: Esse gráfico permite ver a probabilidade de cada valor ou intervalo de valores aparecer nos dados, bem como identificar possíveis modas ou picos na distribuição. Por exemplo, podemos usar um gráfico de densidade para comparar a altura dos homens e das mulheres em um país e ver se há uma diferença significativa entre as médias ou se há uma sobreposição considerável entre as distribuições.

Exemplos de Conclusão

Gráfico de barra: Esse gráfico permite ver a quantidade ou proporção de cada categoria ou grupo nos dados, bem como identificar possíveis padrões ou tendências. Por exemplo, podemos usar um gráfico de barra para mostrar o número de vendas por mês em uma loja e ver se há uma sazonalidade ou variação nas vendas ao longo do ano.

Exemplos de Conclusão

Gráfico de dispersão: Esse gráfico permite ver a relação entre duas variáveis numéricas nos dados, bem como identificar possíveis correlações ou associações. Por exemplo, podemos usar um gráfico de dispersão para mostrar o consumo de energia elétrica em função da temperatura ambiente em uma cidade e ver se há uma relação linear ou não linear entre as variáveis.

Referências

- Bruce, P.; Bruce, A.; **Estatística Prática para Cientista de Dados**: 50 Conceitos Essenciais; Rio de Janeiro; Alta Books; 2019.
- Morettin, P. A.; Bussab, W. O.; **Estatística Básica**. 8 ed. São Paulo: Saraiva, 2013.
- https://vidadeproduto.com.br/diagrama-de-dispersao/ Acessado em 26/04/2022
- https://edtisensei.zendesk.com/hc/pt-br/articles/360033866451-Diagrama-de-Dispers%C3%A3o-Defini%C3%A7%C3%A3o Acessado em 26/04/2022

FAESA
CENTRO UNIVERSITÁRIO

Análise de Dados Aplicada à Computação

PROF. M.SC HOWARD ROATTI