Análise Descritiva dos Dados

Luciano Barbosa

Contexto

- Proposta pelo estatístico John Tukey
- Etapa que deve preceder a criação de modelos

Motivação

- Entender os dados
- Encontrar problemas

Tipos de Dados

- Categórico
  - Binário: 2 categorias (ex: manhã ou noite)
  - Nominal: várias categorias (ex.: cores)
  - Ordinal: ordem importa (ex.: dia do mês)
- Contínuo
  - Ex.: peso, tempo para realizar uma tarefa etc.

Dimensionalidade dos Dados

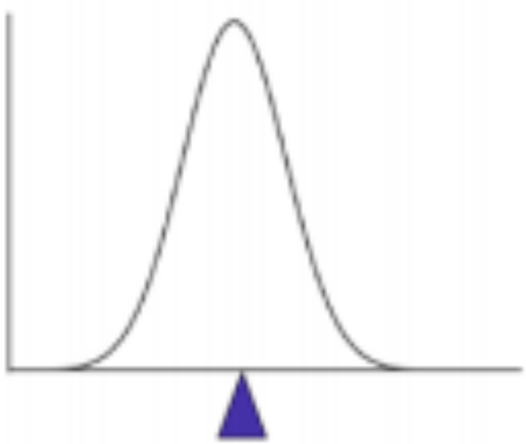
- Univariado
- Bivariado
- Multi-variado

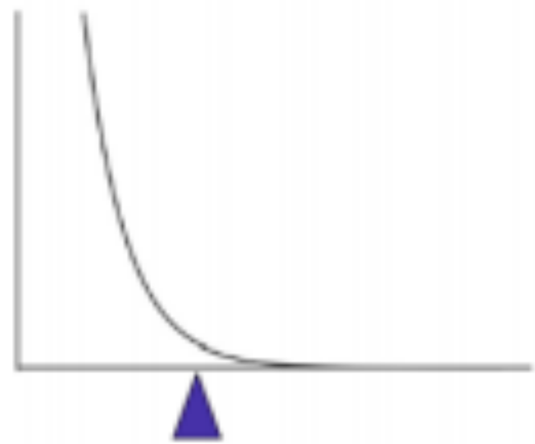
Resumos Numéricos dos Dados

- Medidas de valor central: ponto central ao redor do qual os dados estão distribuídos
  - Ex: média, mediana
- Medidas de variabilidade: descreve como os dados estão distribuídos ou quão distante estão do centro
  - Variância e desvio padrão
- Medidas relativas: descreve posições relativas de pontos nos dados
  - Quartil e percentil

Medidas de valor central: Média

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



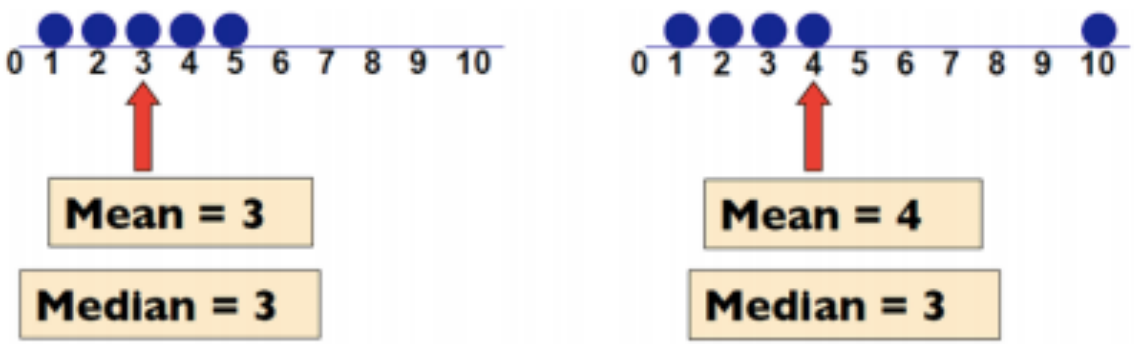


Medidas de valor central: Mediana

- Valor na metade dos valores ordenados
- Se o número de valores for ímpar, usa-se o valor médio
- Se for par, usam-se os dois valores na metade e calcule-se a média
- Ex: 17, 19, 21, 22, 23, 23, 23, 38
  - Mediana = (22+23)/2 = 22.5

Onde Usar Média ou Mediana?

- Média: distribuições simétricas sem outliers
- Mediana: distribuições não simétricas ou dados com outliers



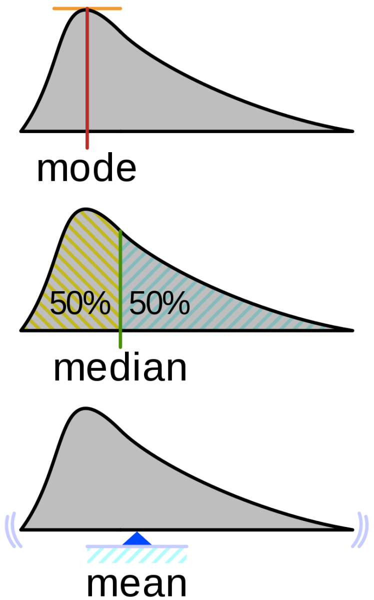
Medidas de valor central: Moda

- Valor mais frequente de um atributo
- Usada para dados categóricos ou numéricos
- Bimodal: dois atributos mais frequentes
- Multimodal: muitos atributos mais frequentes

Comparação: Média, Mediana e Moda

Comparison of common averages of values { 1, 2, 2, 3, 4, 7, 9 }

Type	Description	Example	Result
Arithmetic mean	Sum of values of a data set divided by number of values: $\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i$	(1+2+2+3+4+7+9) / 7	**4**
Median	Middle value separating the greater and lesser halves of a data set	1, 2, 2, **3**, 4, 7, 9	**3**
Mode	Most frequent value in a data set	1, **2**, **2**, 3, 4, 7, 9	**2**



Medidas de Variabilidade: Variância

- Média da diferença dos valores com relação à média

$$\sigma^2 = \frac{\sum (x_r - \mu)^2}{n}$$

- Elevado a 2 elimina números negativos
- Valores absolutos não possui boas propriedades matemáticas

Medidas de Variabilidade: Desvio Padrão

- Variância é difícil de interpretar
- O que significa uma variância de 10.8 ou 2.2
- Padronização da variância: desvio padrão
- Mesma unidade dos dados originais
- Raiz quadrada da variância

$$\sigma = \sqrt{\frac{\sum (x_r - \mu)^2}{n}}$$

Exemplo:
Peso de Ovos

Weight (x)	$(x - \bar{x})$	$(x - \bar{x})^2$
60	1	1
56	-3	9
61	2	4
68	9	81
51	-8	64
53	-6	36
69	10	100
54	-5	25
472		**320**

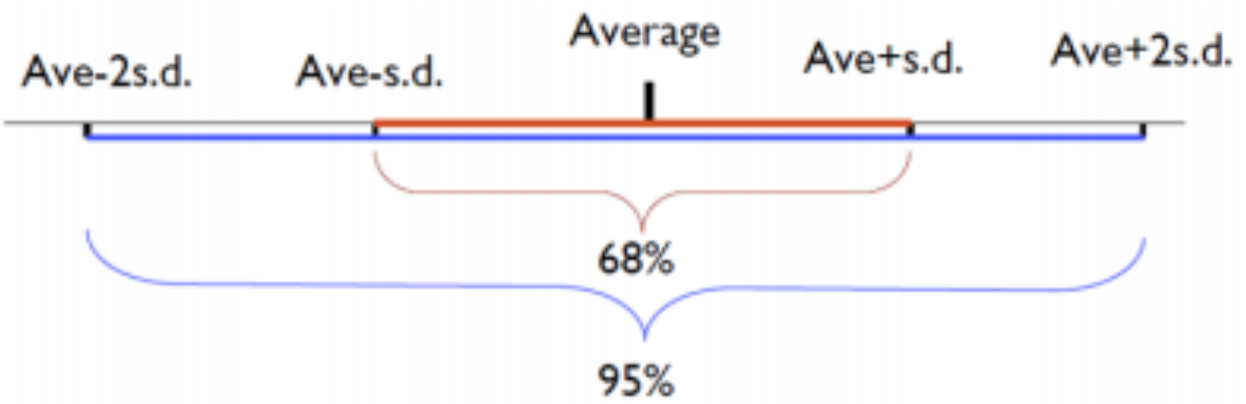
$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

$$= \sqrt{\frac{320}{8}}$$

= 6.32 grams

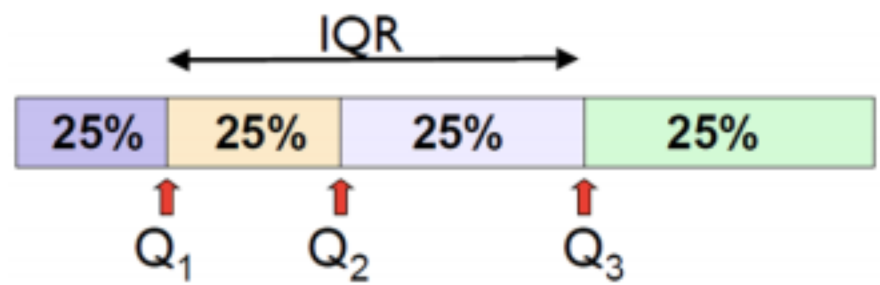
Desvio Padrão

- Se a distribuição é próxima à gaussiana

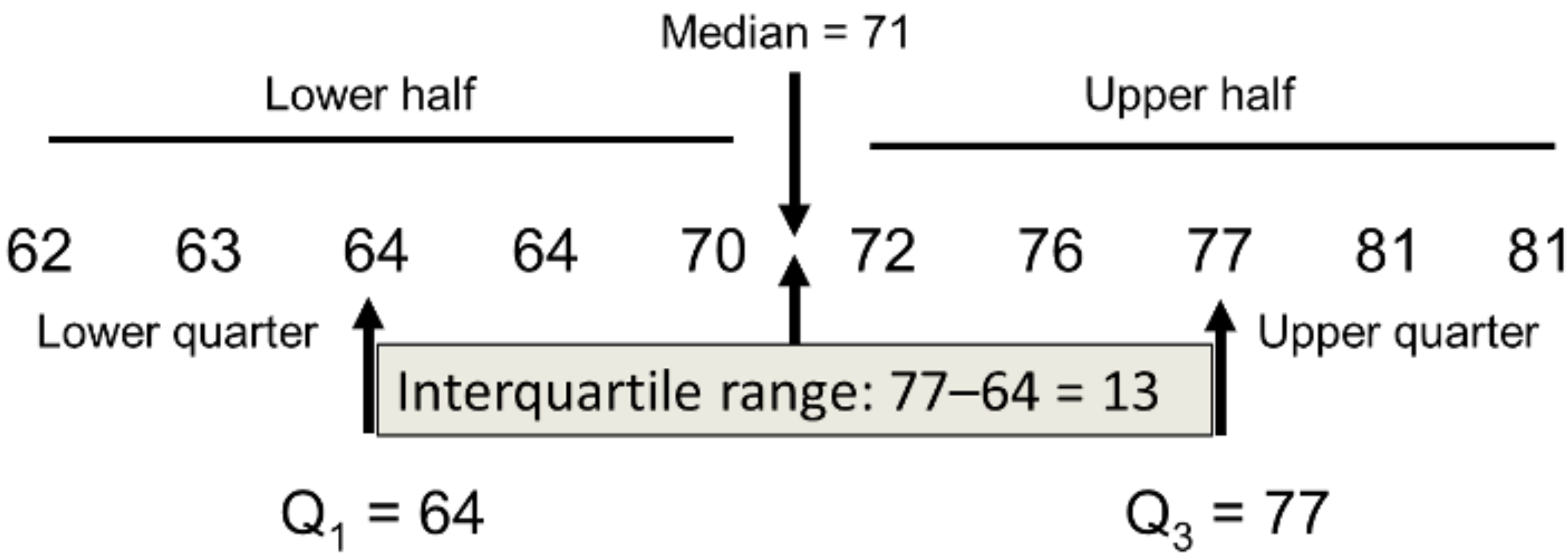


Medidas Relativas: Quantil

- Quantis: dividem os valores em intervalos com a mesma frequência (mesma quantidade de elementos)
- Mediana: 2-quantil
- Quartil: 4-quantil
- Percentil: 100-quantil
- IQR (amplitude interquartile): $Q_3 - Q_1$



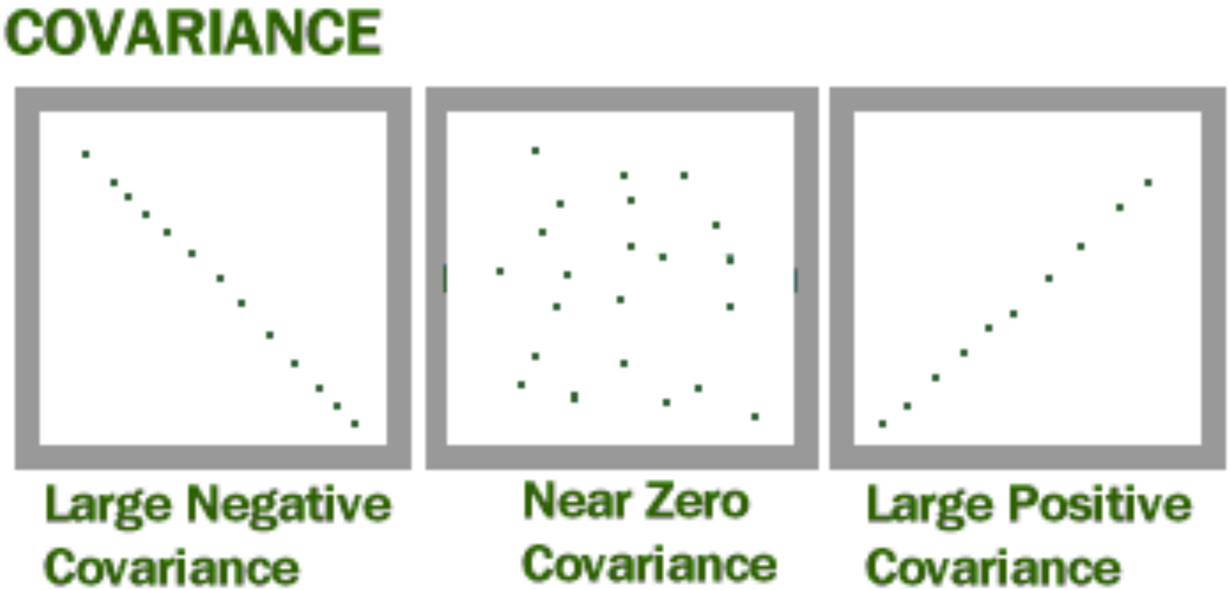
Exemplo: IQR

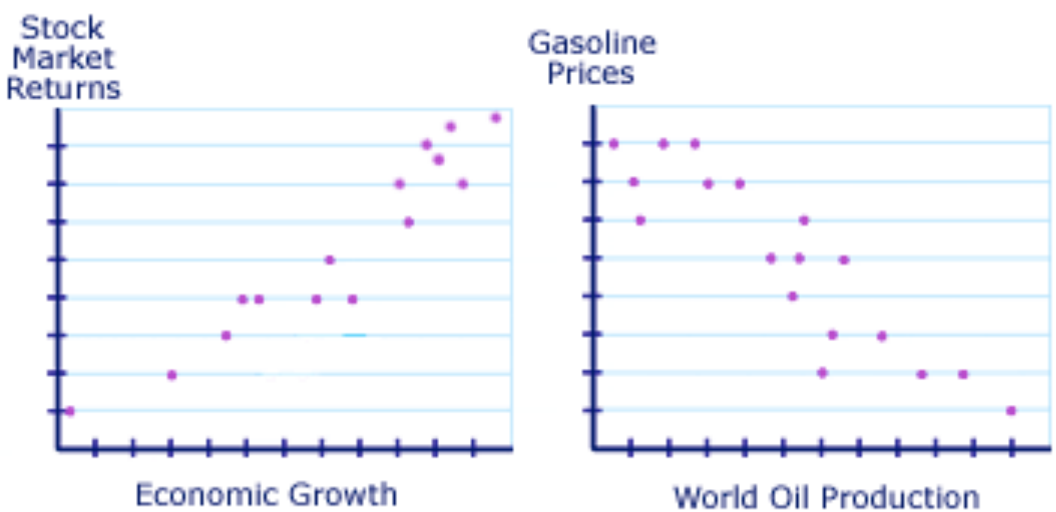

Median = 71
Lower half
Upper half
62 63 64 64 70 72 76 77 81 81
Lower quarter
Upper quarter
Interquartile range: 77–64 = 13
$Q_1 = 64$
$Q_3 = 77$

Medidas de Variabilidade: Covariância

- Avalia a variância conjunta de dois atributos (bivariada)
- Se a variação dos valores de um atributo acompanha a do outro
- Covariância positiva: valores altos para um atributo X estão associados a valores altos para outro atributo Y
- Covariância negativa: X aumenta Y diminui
- Zero: ausência de relação

Covariância





Covariância

- A covariância (amostral) entre dois atributos X e Y:

$$cov(X,Y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{N-1}$$

Economic Growth % (x_i)	S & P 500 Returns % (y_i)
2.1	8
2.5	12
4.0	14
3.6	10

Exemplo

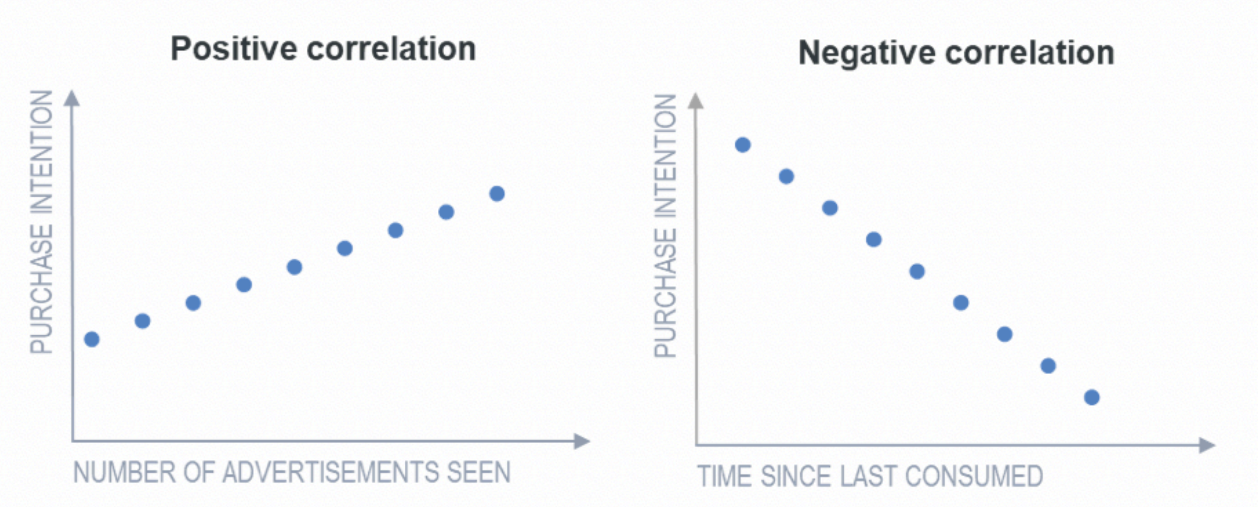
Economic Growth % (x_i)	S & P 500 Returns % (y_i)
2.1	8
2.5	12
4.0	14
3.6	10

$$\bar{x} = 3.1$$
$$\bar{y} = 11$$

$$
\begin{aligned}
COV(x,y) &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \\
&= \frac{(2.1 - 3.1)(8 - 11) + \ldots}{4-1} \\
&= \frac{(-1)(-3) + (-0.6)(1) + (0.9)(3) + \ldots}{3} \\
&= \frac{3 + (-0.6) + 2.7 + (-0.5)}{3} \\
&= \frac{4.6}{3} \\
&= 1.53
\end{aligned}
$$

Correlação

- Covariância mostra se atributos se relacionam positivamente ou negativamente mas não o grau que eles se relacionam
- Correlação padroniza a medida de relacão entre os atributos:
  - Valores entre 1 e -1
  - 0: sem correlação



Correlação de Pearson

- Normaliza a covariância pelo desvio padrão dos atributos
- Suposições:
  - Variáveis seguem uma gaussiana
  - Variáveis contínuas
  - Linearidade
- Quantifica a existência de uma relação linear entre as variáveis.

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)}$$

Exemplo: Correlação de Pearson

Economic Growth % (x_i)	S & P 500 Returns % (y_i)
2.1	8
2.5	12
4.0	14
3.6	10

$$r_{(x,y)} = \frac{COV(x,y)}{s_x s_y}$$

$$r_{(x,y)} = \frac{1.53}{(.90)(2.58)}$$

$$= .66$$

Correlação de Spearman

- Não paramétrico: atributos são relacionados por qualquer função monotônica
- Variáveis podem ser ordinais

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

- d^2: quadrado da diferença entre os ranqueamentos dos atributos
- n: número de instâncias

Exemplo:
Correlação de Spearman

	Marks									
English	56	75	45	71	62	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63

Correlação de Spearman

- Ordeno cada atributo e gero um ranking

	Maths (mark)	Rank (English)	Rank (maths)	d	d^2
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	3	9
62	65	6	5	1	1
64	56	5	9	4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	1	1
61	63	7	6	1	1

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

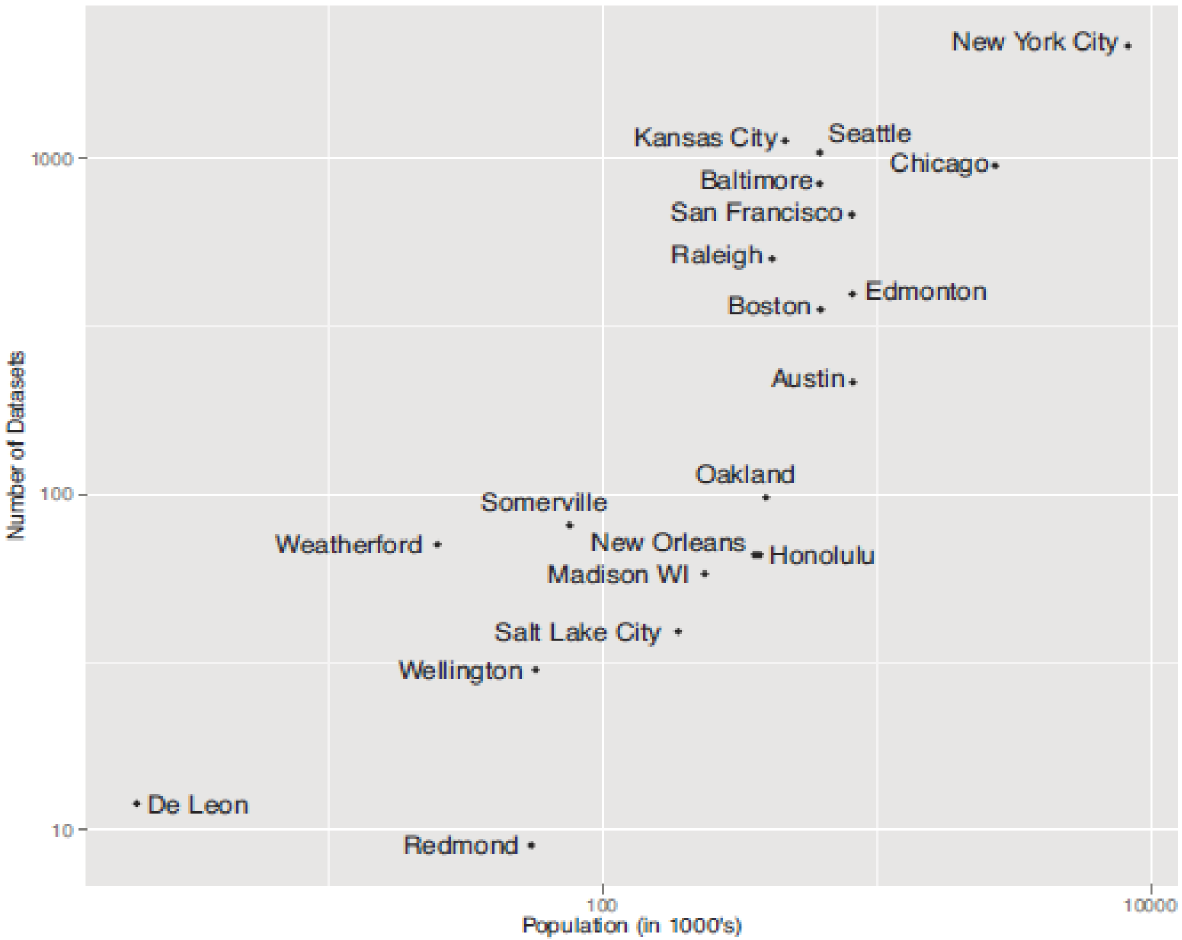
$$\rho = 1 - \frac{6 \times 54}{10(10^2-1)}$$

$$\rho = 1 - \frac{324}{990}$$

$$\rho = 1 - 0.33$$

$$\rho = 0.67$$

Uso em Dados Urbanos



Exercício

- Dataset: https://github.com/if1015-datascience/material/blob/gh-pages/data/recife.csv
- Computar os valores abaixo para os campos numéricos:
  - Média: mean
  - Mediana: median
  - Moda: mode
  - Desvio padrão: std
  - Quantis: describe
- Computar covariância e correlações entre colunas:
  - Covariânca: cov
  - Correlação de pearson: corr(method=pearson)
  - Correlação de spearman: corr(method=spearman)
- Plotar
  - Histograma: df[‘col’].hist()
  - Boxplot: df.boxplot(column=[‘col'])
  - Scatterplot: df.plot.scatter(x=’col1',y=’col2’)