

Estatística Descritiva com Python





Oi, sou a Kay!


kayleighmeneghini@gmail.com

- Sou **astrônoma** pelo IAG-USP e **divulgadora científica**;
- **PyLady** desde 2016;
- Há três anos trabalho no mercado de **Marketing Digital**
- Atualmente do time de **Data Science** da **DP6**.

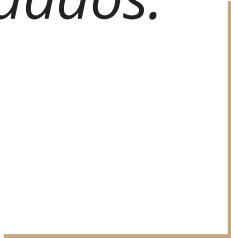
Quem são vocês?

Acesse: sli.do
Event code: pyladies

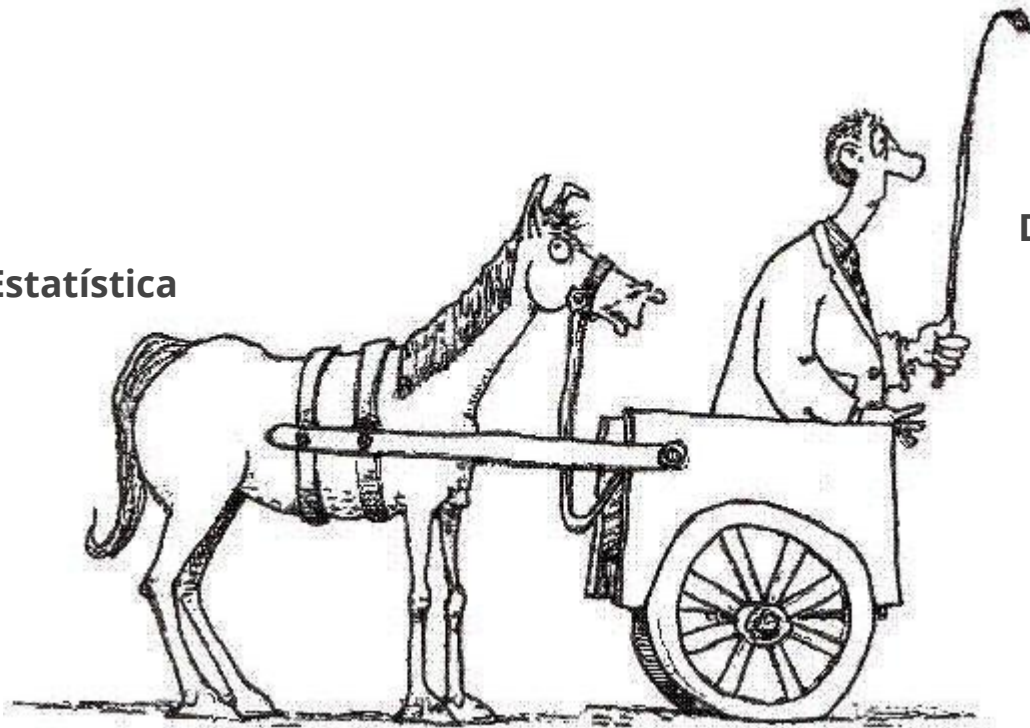
Por que estatística?



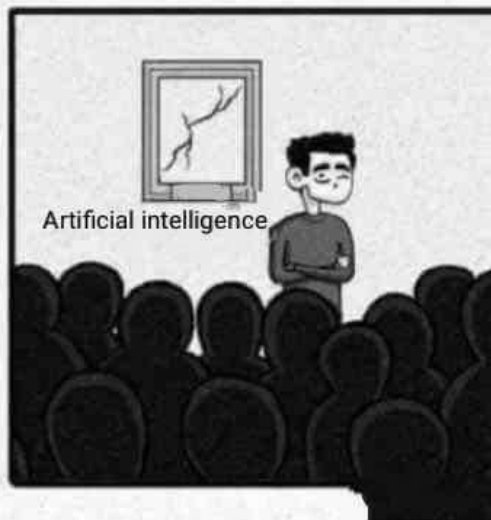
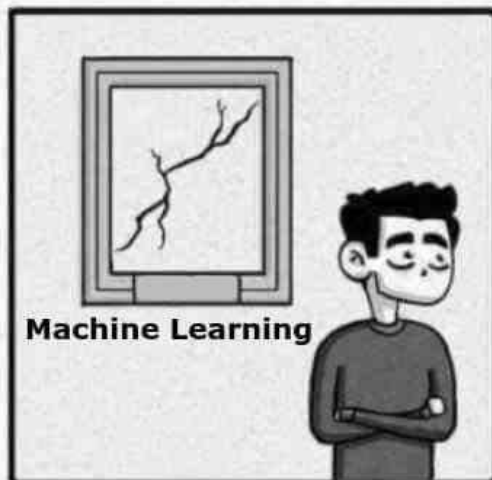
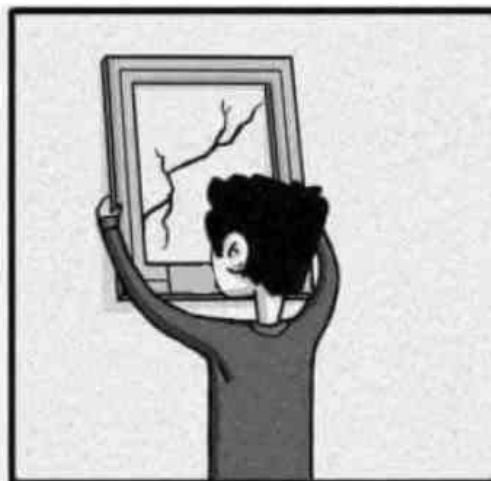
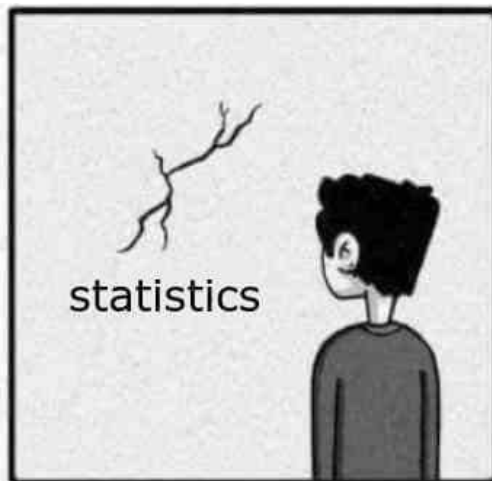
*“Estatística é o ramo da matemática que trata da **coleta**, da **análise**, da **interpretação** e da **apresentação** de dados.”*



Estatística



Data Science



Por que Python?

Exemplo do mesmo programa em diferentes linguagens

Java

```
1 public class Hello
2 {
3     public static void main(String args[]) {
4         java.util.Scanner s = new java.util.Scanner(System.in);
5         System.out.print("Digite seu nome:");
6         String nome = s.nextLine();
7         System.out.println("Olá, " + nome);
8     }
9 }
```

C

```
1 #include <stdio.h>
2 int main()
3 {
4     char nome[200];
5     printf("Digite seu nome:");
6     scanf("%s", nome);
7     printf("Olá, %s\n", nome);
8     return 0;
9 }
```

Python

```
1 nome = input('Digite seu nome:')
2 print ('Olá,', nome)
```

Código Aberto

Comunidade Python ❤️

Documentação (inclusive em pt-br)

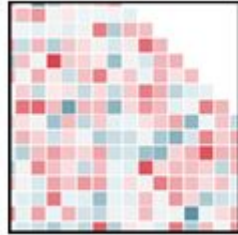
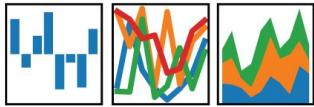
Bibliotecas fáceis

~~Stack Overflow~~

Ferramentas para análise de dados

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Seaborn



NumPy



SciPy




StatsModels

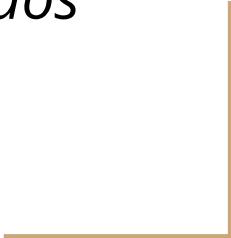
Statistics in Python



Estatística Descritiva



“Estatística descritiva é um ramo da estatística que aplica várias técnicas para resumir e descrever um conjunto de dados”



Resumo dos dados

Estatística Descritiva

Tipos de variáveis

Categóricas (Dimensões)

Dispositivo

Cidade

Dia da semana

Data

Numéricas (Métricas)

Visitas no site

Taxas

Receita

Novos Usuários

Nominal

Dispositivo

Cidade

Ordinal

Dia da semana

Data

Discreta (int)

Novos Usuários

Visitas no site

Contínua (float)

Receita

Bounce Rate



```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
df = sns.load_dataset("tips")
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 7 columns):
total_bill    244 non-null float64
tip           244 non-null float64
sex           244 non-null category
smoker        244 non-null category
day           244 non-null category
time          244 non-null category
size          244 non-null int64
dtypes: category(4), float64(2), int64(1)
memory usage: 7.3 KB
```

```
df.head()
```

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3

Primeiro passo:
importando as bibliotecas

Importando os dados e salvando em
uma variável df

Visualizando o **tipo dos dados**.

Nem sempre (~~quase nunca~~)
todos vão estar no formato certo.

É aqui que passamos 80% do nosso
tempo.
Mas uma hora acaba!

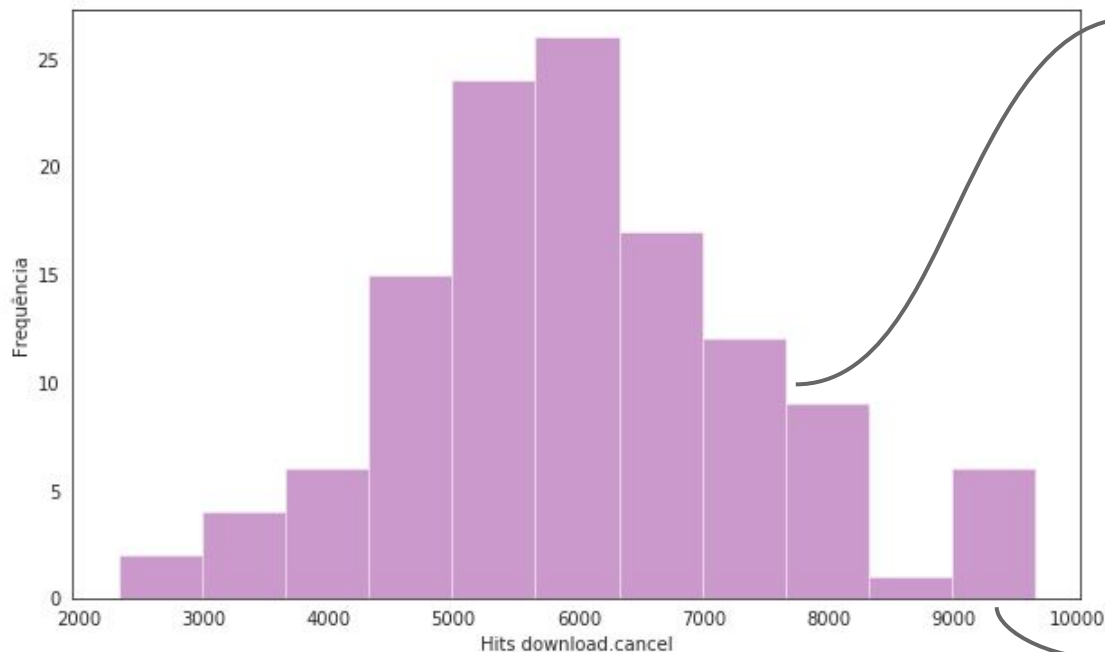
Visualizando a base de
dados

Tudo certo? Sim!
#Partiu explorar as
variáveis.



Distribuição de frequências: Histograma

```
ax, fig = plt.subplots(figsize=(10,6))
g = sns.distplot(df_android['download.cancel'], kde=False, color='purple')
g.set_xlabel('Hits download.cancel')
g.set_ylabel('Frequência')
```

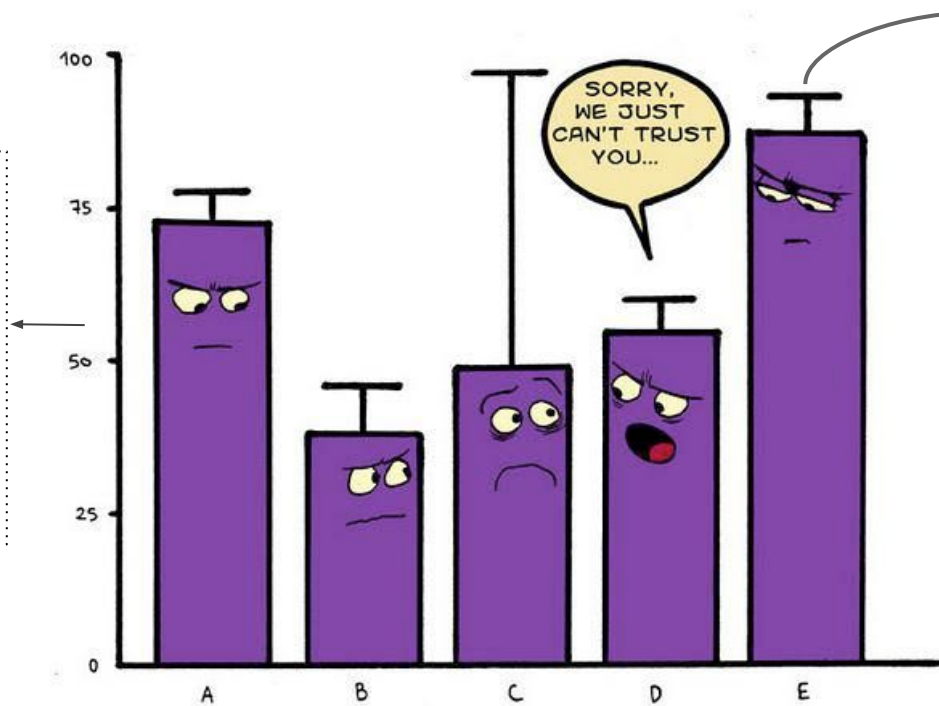


A altura é proporcional à **frequência absoluta** dos valores das variáveis numéricas, isto é, quantas vezes aquele valor (ou valores do intervalo) aparece no conjunto de dados

A largura das barras representa intervalos de valores de **variáveis numéricas**

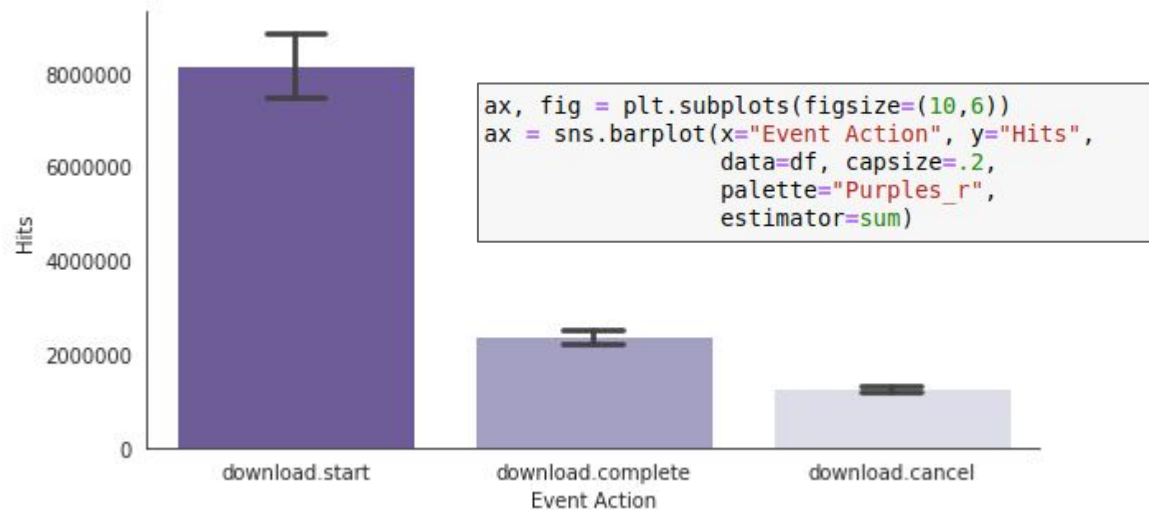
Comparação entre variáveis categóricas: gráfico de barras

A altura de cada barra é proporcional a uma **agregação específica** (por ex., a soma dos valores na variável categórica que representa)



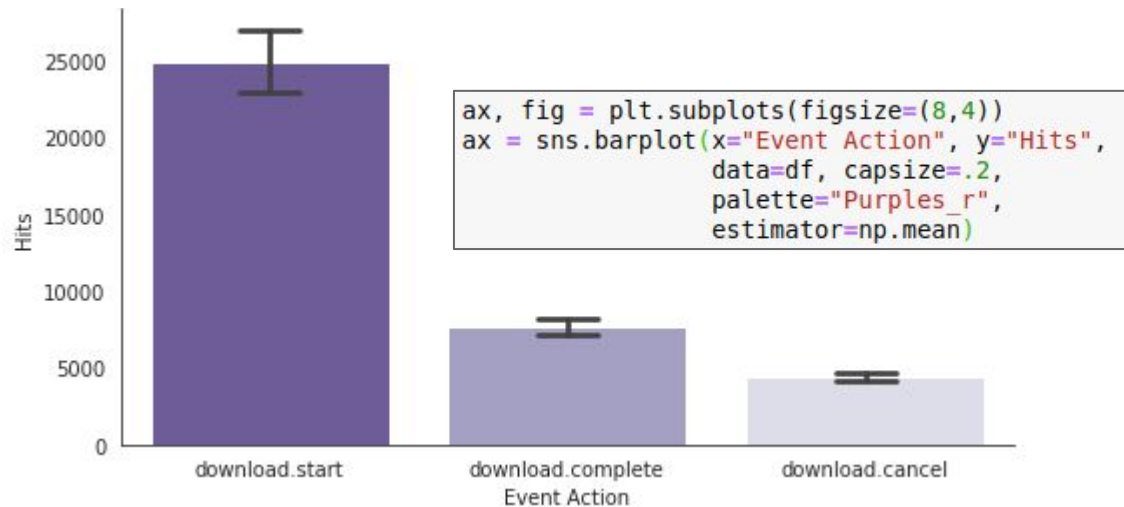
Barras de erro: indica a incerteza na medida ou intervalo de confiança.

Agora no eixo x temos **variáveis categóricas** ou **variáveis discretas**, representadas por barras.



O parâmetro **estimator** indica a **agregação** da variável categórica.

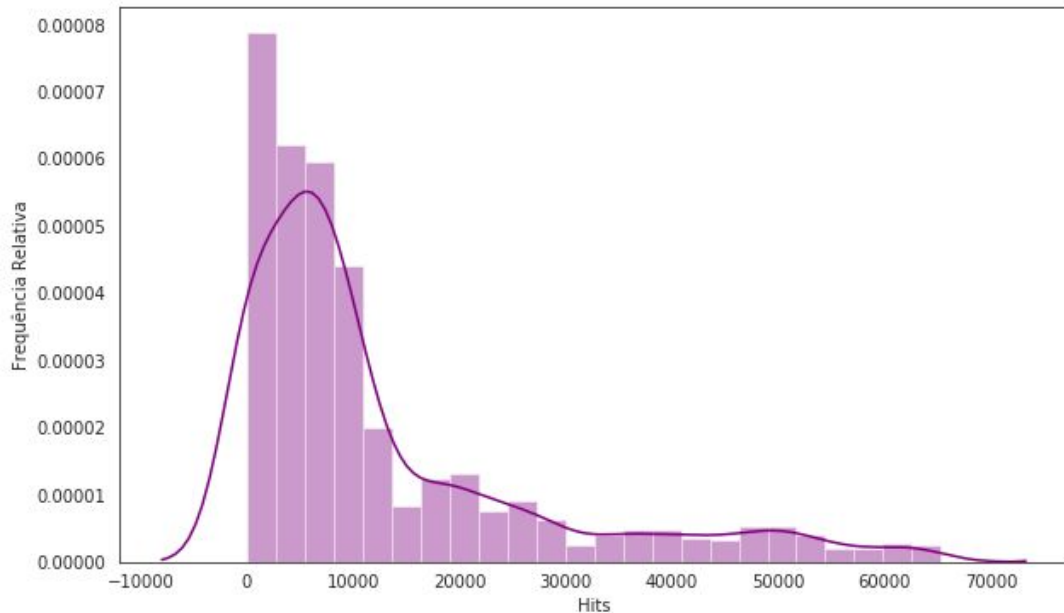
Ao lado, a **soma** dos *hits* de cada evento representado no eixo x.



Aqui temos a **média** dos *hits* dos eventos no eixo x.

Distribuição de frequências

```
ax, fig = plt.subplots(figsize=(10,6))
g = sns.distplot(df['Hits'], color='purple')
g.set_xlabel('Hits')
g.set_ylabel('Frequência Relativa')
```



$$\text{Frequência Relativa} = \frac{\text{Frequência Absoluta}}{\text{Total de observações}}$$

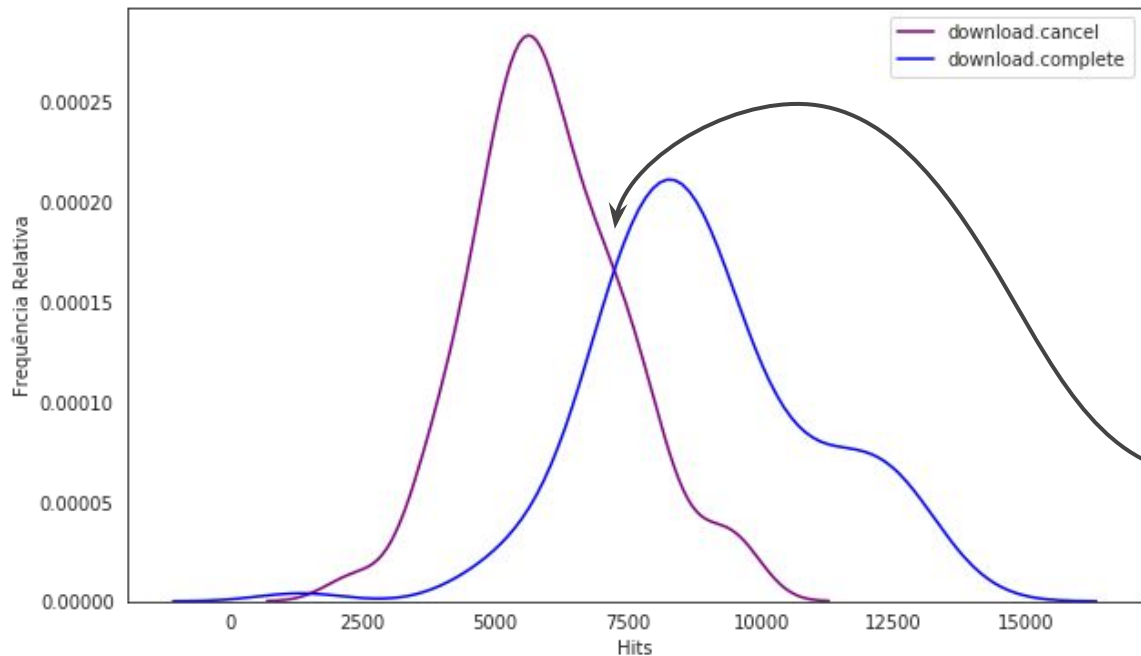
A frequência relativa representa uma **porcentagem**.

Além de mostrar a distribuição dos dados, traz uma visão de **composição**.

A curva desenhada é uma estimativa para a **função densidade de probabilidade**, vulgo *fdp*.

Distribuição de probabilidades

```
g = sns.distplot(df_android['download.cancel'], hist=False, color='purple', label='download.cancel')
g = sns.distplot(df_android['download.complete'], hist=False, color='blue', label='download.complete')
g.set_xlabel('Hits')
g.set_ylabel('Frequência Relativa')
```

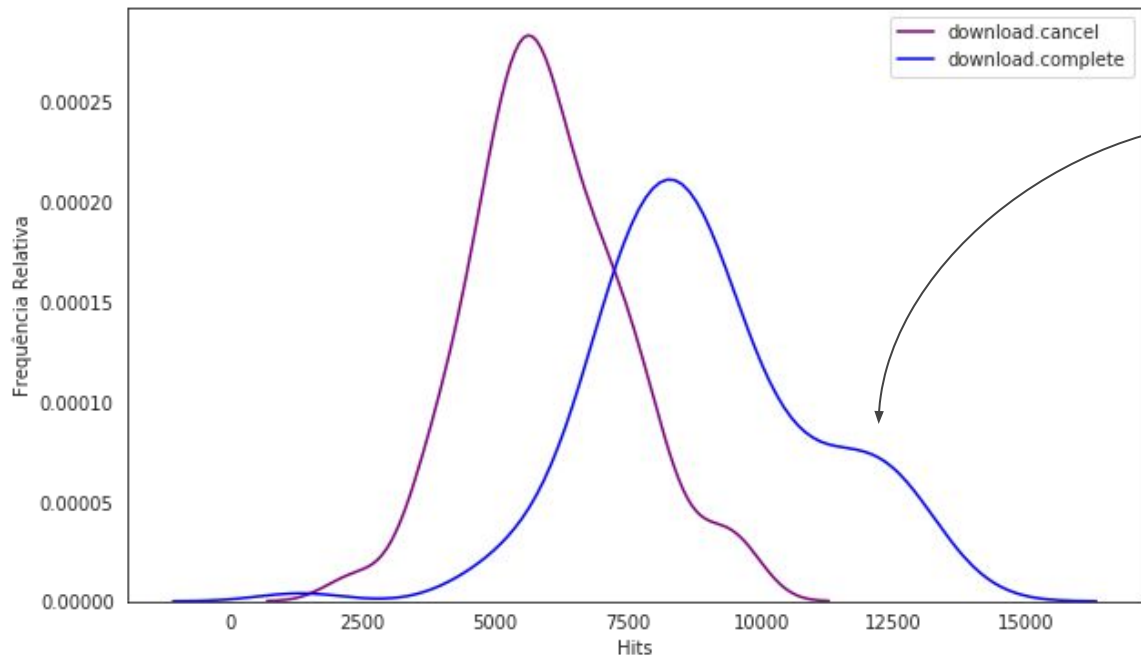


Podemos comparar as probabilidades de variáveis distintas assumirem determinados valores.

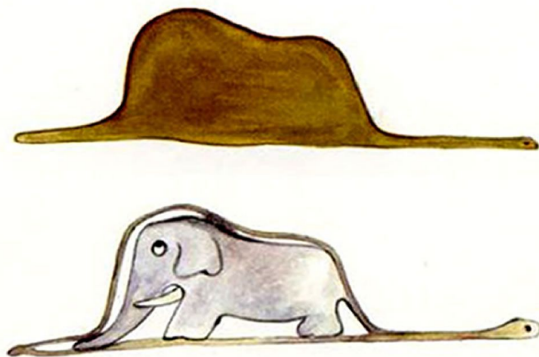
Onde as curvas se cruzam as variáveis têm a mesma probabilidade de ocorrência.
(*Mesma frequência relativa*)

Distribuição de probabilidades

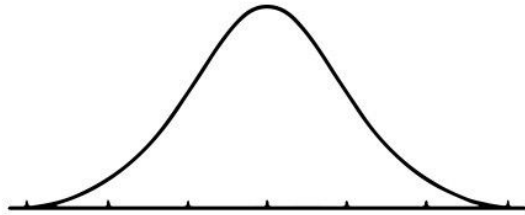
```
g = sns.distplot(df_android['download.cancel'], hist=False, color='purple', label='download.cancel')  
g = sns.distplot(df_android['download.complete'], hist=False, color='blue', label='download.complete')  
g.set_xlabel('Hits')  
g.set_ylabel('Frequência Relativa')
```



Isso não é uma distribuição



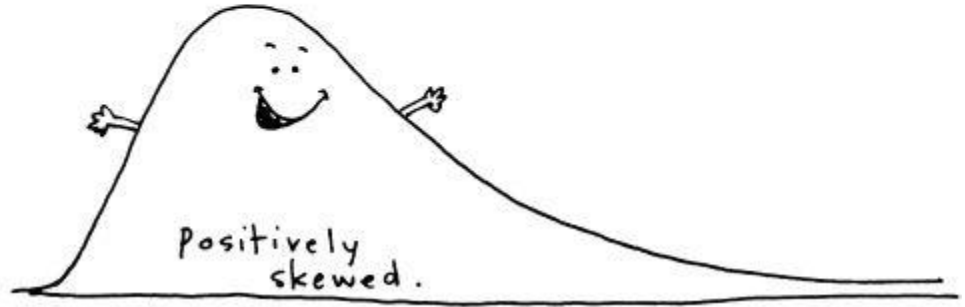
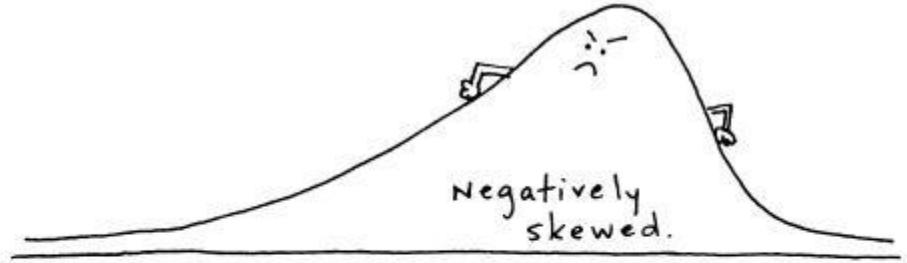
Distribuição de probabilidades



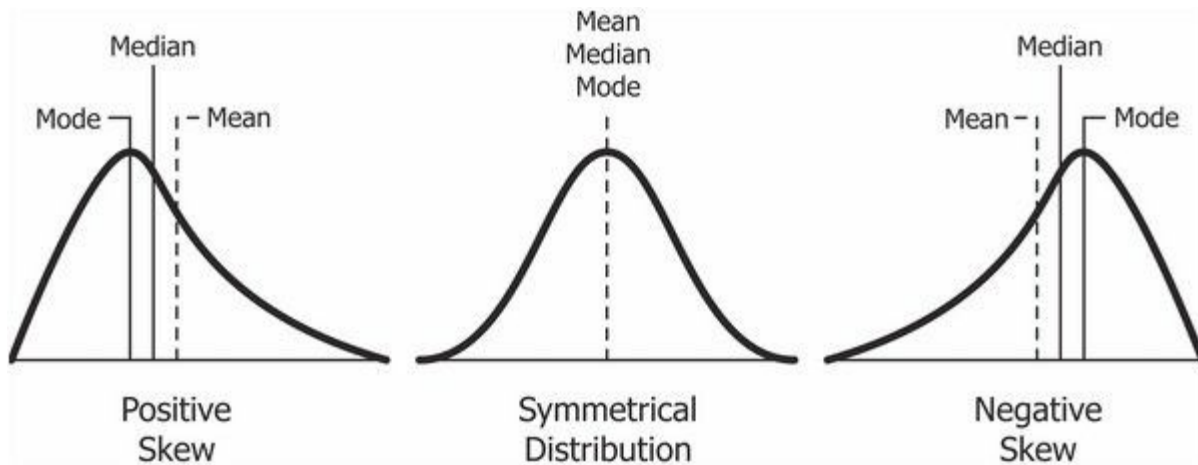
NORMAL DISTRIBUTION



PARANORMAL DISTRIBUTION



Skewness (viés)



Medidas Resumo

Estatística Descritiva

Medidas de Posição

Medidas de Tendência Central

Média Simples

$$M_s = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Média Ponderada

$$M_p = \frac{p_1 * x_1 + p_2 * x_2 + \dots + p_n * x_n}{p_1 + p_2 + \dots + p_n}$$

Média

é soma dos elementos de uma amostra dividida pela quantidade de elementos

Mediana

É definida como o valor tal que 50% das observações são menores e 50% são maiores que ela.

$$Q_2 = x_{(\frac{n+1}{2})}$$

n ímpar

n par

$$Q_2 = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2} + 1)}}{2}$$

Ex.: [2, 3, 4, 2, 5, 6, 2, 8]

Moda = 2

Moda

é o valor (ou atributo) que ocorre com maior frequência

Ex.: [2, 3, 3, 2, 5, 3, 2, 8]

Moda = 2



Medidas de Dispersão

Variância (σ^2)

é uma medida de dispersão que mostra o quão distante cada valor desse conjunto está do valor central (média).

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N}$$

Pode ser difícil de interpretar o valor, por ser uma soma quadrática dos termos.
Ex.: a variância do ticket médio é 59599.

Desvio Padrão (σ)

É o resultado da raiz quadrada da variância.

$$\sigma = \sqrt{\sigma^2}$$

Indica qual seria o “erro” se substituíssemos um dos valores coletados pelo valor da média.

Coeficiente de Variação (CV)

comparar a variação de conjuntos de observações que diferem na média ou são medidos em grandezas diferentes.

$$CV = \frac{\sigma}{\bar{X}} * 100$$

O **CV** é o desvio padrão expresso como uma porcentagem média.

Quantis

Tanto a média como o desvio padrão podem não ser medidas adequadas para representar um conjunto de dados, pois:

- são afetados por valores extremos;
- apenas com esses dois valores não temos ideia de simetria ou assimetria da distribuição dos dados.

A mediana no entanto não é afetada por valores discrepantes, por isso dizemos que é um parâmetro **robusto**.

Definimos **$q(p)$** o quantil de ordem p ou **p-quantil** ($0 < p < 1$) como o valor da variável tal que **100p%** das observações sejam menor que ele na amostra ordenada.

Alguns exemplos de quantis:

$q(0.25)$ = 1º Quartil = 25º Percentil
 $q(0.50)$ = Mediana = 2º Quartil
 $q(0.75)$ = 3º Quartil
 $q(0.40)$ = 4º Decil
 $q(0.95)$ = 95º Percentil



Quantis e simetria

Os valores $X_{(1)}$, q_1 , q_2 , q_3 e $X_{(n)}$ são importantes para se ter uma boa ideia da assimetria dos dados.

Para uma distribuição simétrica (ou aproximadamente simétrica):

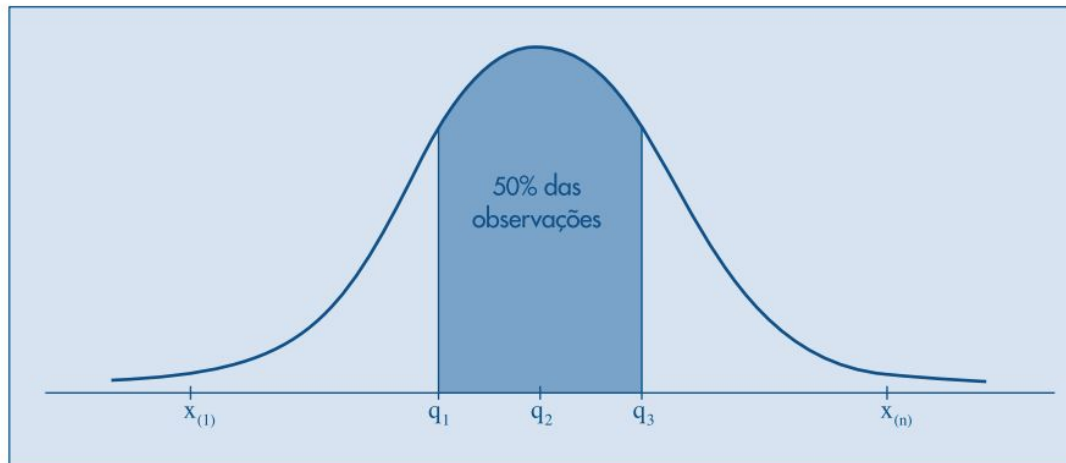
(a) $q_2 - x_{(1)} \approx x_{(n)} - q_2$

(b) $q_2 - q_1 \approx q_3 - q_2$

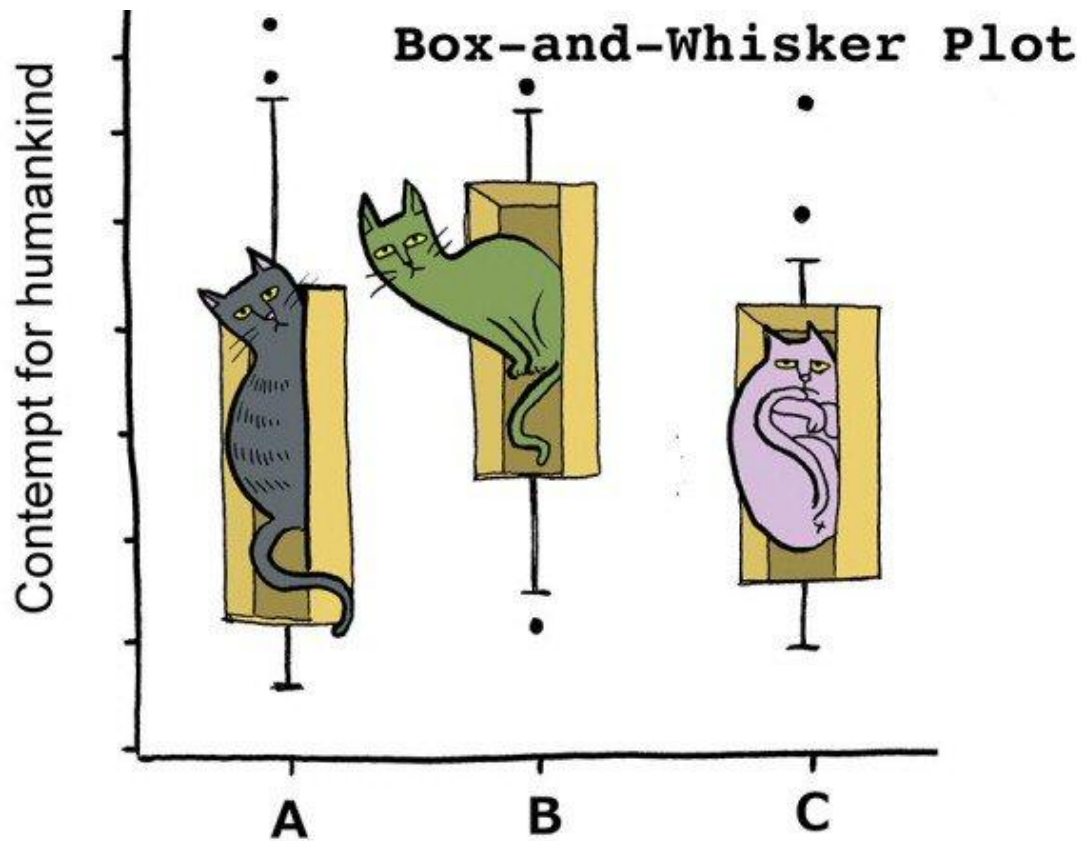
(c) $q_1 - x_{(1)} \approx x_{(n)} - q_3$

(d) distâncias entre mediana e q_1 , q_3 menores do que distâncias entre os extremos e q_1 , q_3

Distribuição simétrica: normal ou *gaussiana*



Boxplot



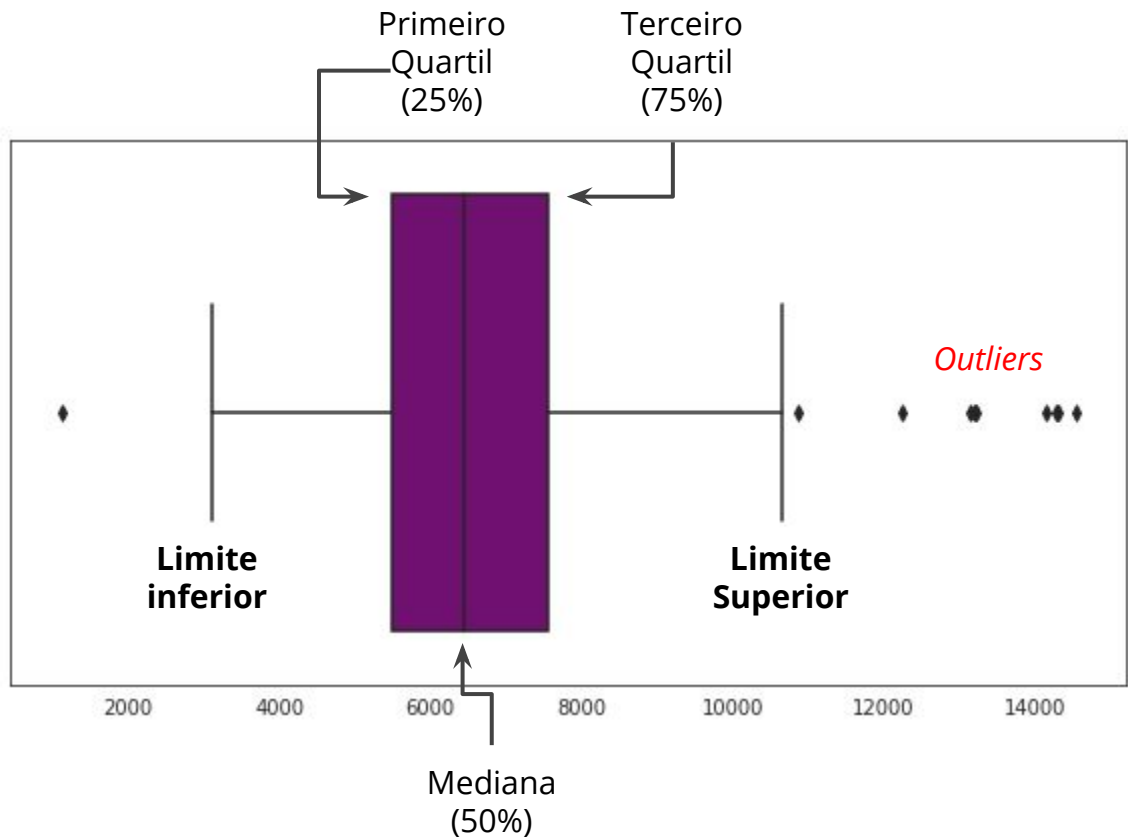
Mostra a **distribuição** dos dados com base em parâmetros descritivos (*Mediana e Quartis*)

Útil para **comparação** da distribuição dos dados em diferentes dimensões.

Descreve a **simetria** dos dados.

Aponta possíveis *outliers*.

Boxplot



Limite Inferior

$$Q1 - 1.5 \cdot (IQR)$$

Limite Superior

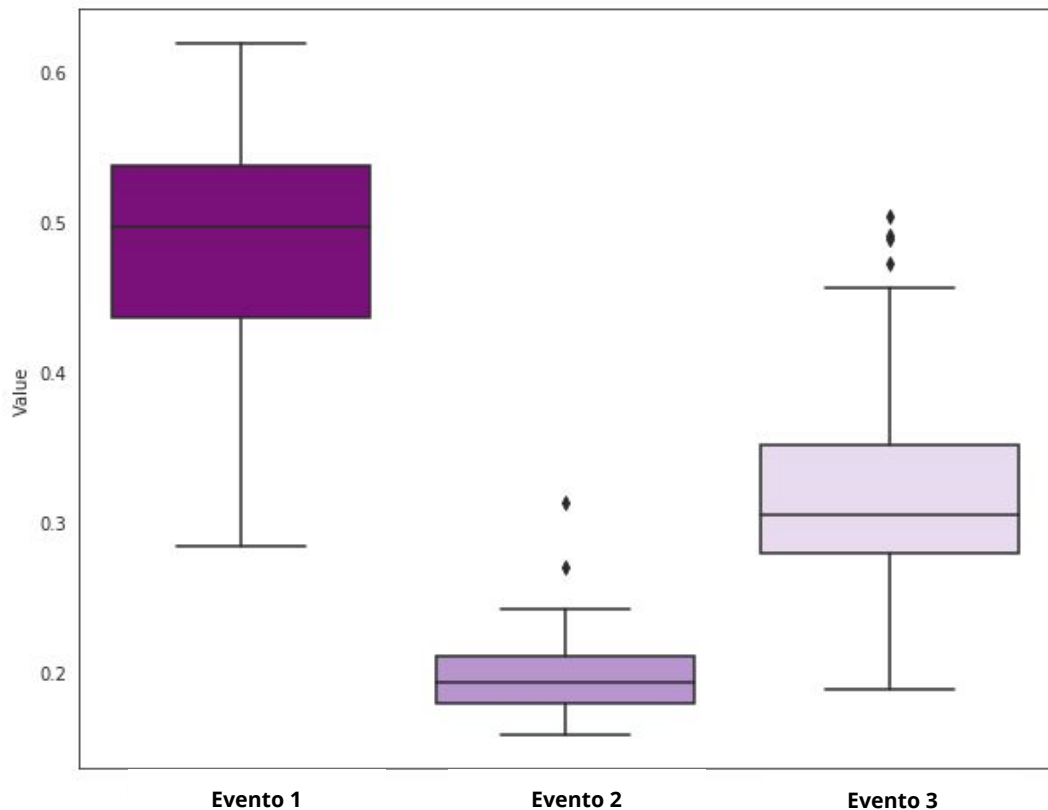
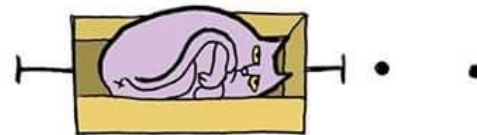
$$Q3 + 1.5 \cdot (IQR)$$

Intervalo Interquartil

$$IQR = Q3 - Q1$$

Serão considerados *outliers* valores acima do limite superior ou abaixo do limite inferior

Boxplot



Qual tem maior variabilidade?

Maior intervalo interquartil e limites.

Qual está mais propenso a outliers?

O que tiver distribuição mais simétrica ou menor intervalo interquartil.

Qual possui distribuição mais simétrica?

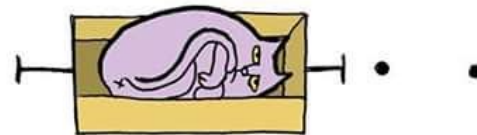
Mediana aproximadamente equidistante ao primeiro e terceiro quartil.

Qual mais assimétrica e qual o viés?

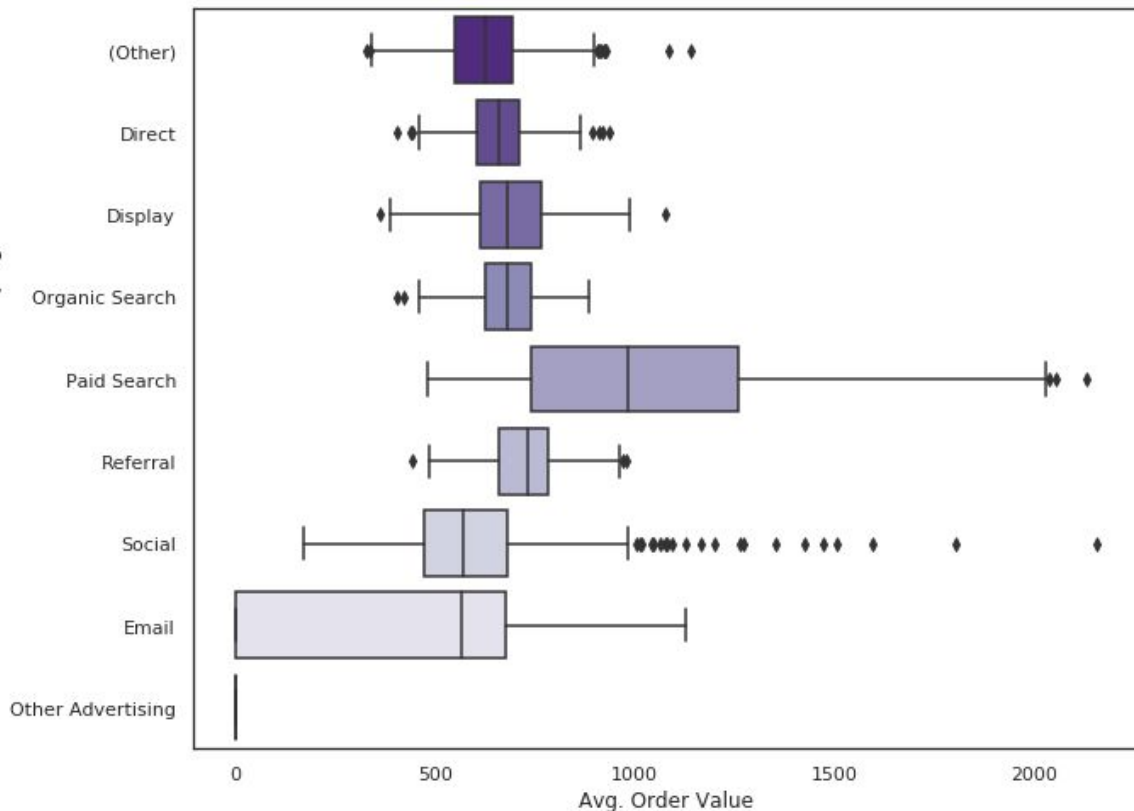
Maior distância entre Q1 e Mediana:
Viés Negativo

Maior distância entre Q3 e Mediana:
Viés Positivo

Boxplot



Default Channel Grouping



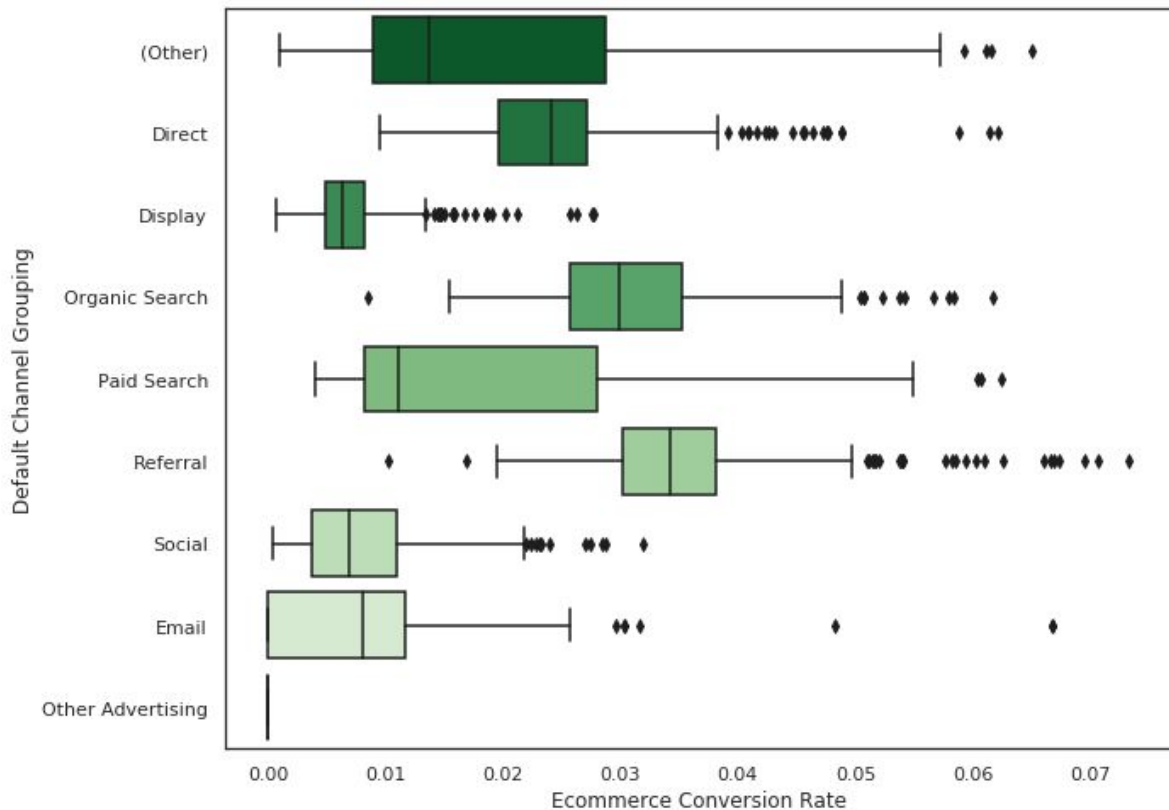
Análise Estatística

- Qual tem maior variabilidade?
- Qual está mais propenso a *outliers*?
- Qual possui distribuição mais simétrica?
- Qual mais assimétrica e qual o viés?

Análise de Negócio

- Melhor ticket médio?
- Canal mais estável/consistente?
- Qual provável motivo de *outliers*?
- Como as **particularidades** dos canais se refletem na distribuição de valores?

Boxplot



Perguntas

- Qual tem maior variabilidade?
- Qual está mais propenso a *outliers*?
- Qual possui distribuição mais simétrica?
- Qual mais assimétrica e qual o viés?

Pandas Describe

```
df_ios[['download.start',  
        'download.complete',  
        'download.cancel']].describe()
```

O método `.describe()` traz os valores descritos no *boxplot* além da média, desvio padrão e número de dados das variáveis.

	download.start	download.complete	download.cancel
count	122.000000	122.000000	122.000000
mean	21684.721311	10460.795082	4291.885246
std	4896.082065	2976.841543	1180.326248
min	3632.000000	1947.000000	577.000000
25%	18233.500000	8390.000000	3603.000000
50%	21168.500000	10184.500000	4277.000000
75%	24912.000000	11912.000000	4966.250000
max	35593.000000	17605.000000	11158.000000

Sugestão de abordagem

Comparar média e mediana.
Pois a média é afetada por outliers.

Comparar média e desvio padrão.

Estimar a distância entre a média e a mediana em termos do desvio padrão.

Verificar valores mínimos e máximos;

E agora, José?

	Users	New Users	Sessions	Bounce Rate	Pages / Session	Avg. Session Duration	Ecommerce Conversion Rate	Transactions	Avg. Order Value
count	3697.000000	3697.000000	3697.000000	3697.000000	3697.000000	3696.000000	3697.000000	3697.000000	3697.000000
mean	45727.556397	18582.534217	56115.506357	0.202332	5.133541	5.074223	0.019523	1406.295375	693.789775
std	51495.911380	22976.269330	62071.045055	0.124176	1.429403	3.492824	0.013508	1756.958170	236.642036
min	1.000000	0.000000	1.000000	0.000000	1.110000	1.000000	0.000000	0.000000	0.000000
25%	10841.000000	3171.000000	13786.000000	0.128400	4.270000	3.500000	0.007800	104.000000	592.460000
50%	22369.000000	7779.000000	28444.000000	0.168200	4.980000	5.040000	0.016800	441.000000	674.740000
75%	71843.000000	29833.000000	89619.000000	0.249000	5.930000	6.060000	0.029400	2469.000000	760.150000
max	580360.000000	344101.000000	724749.000000	0.888900	19.000000	58.000000	0.073300	13901.000000	2158.950000

Agrupamento de variáveis em dimensões com comportamento distintos pode causar assimetria e grande variância.

Por exemplo

Será que o padrão de navegação é o mesmo em *Web* e *Mobile*?

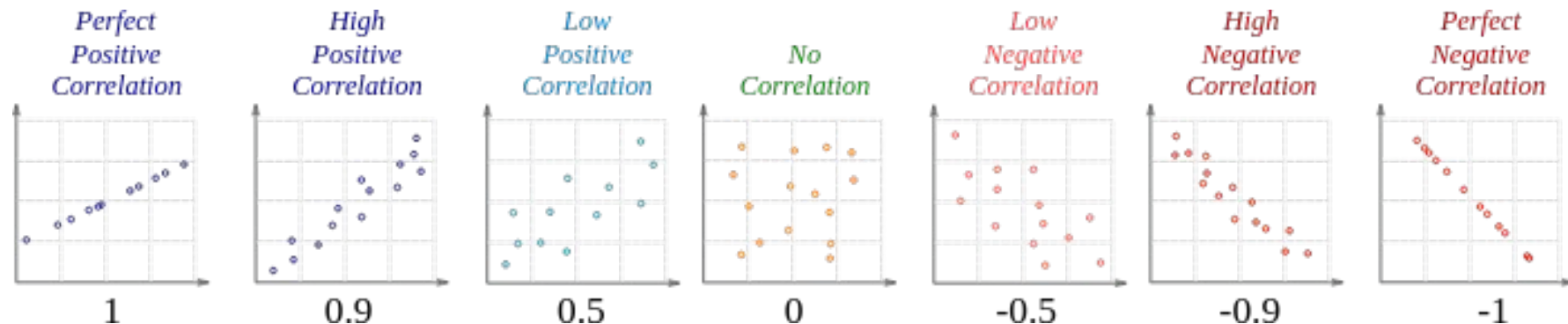
Será que tratar as métricas sem essa quebra é uma boa generalização?



Análise Bidimensional

Estatística Descritiva

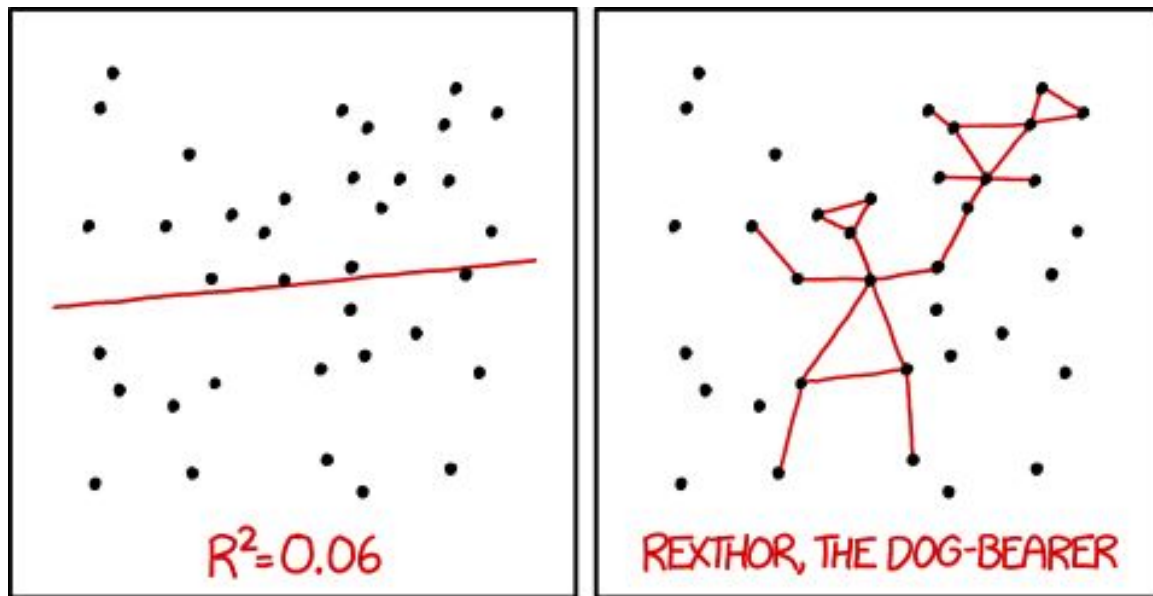
Gráfico de dispersão



Com o gráfico de dispersão podemos ter uma ideia de que tipo de função ajusta os nossos dados.

Ex.: retas, parábolas etc

Como identificar correlação em gráficos de dispersão?



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER
TO GUESS THE DIRECTION OF THE CORRELATION FROM THE
SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Covariância e Correlação

Covariância entre as variáveis X e Y:
a média dos produtos dos valores
centrados das variáveis.

$$Cov(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

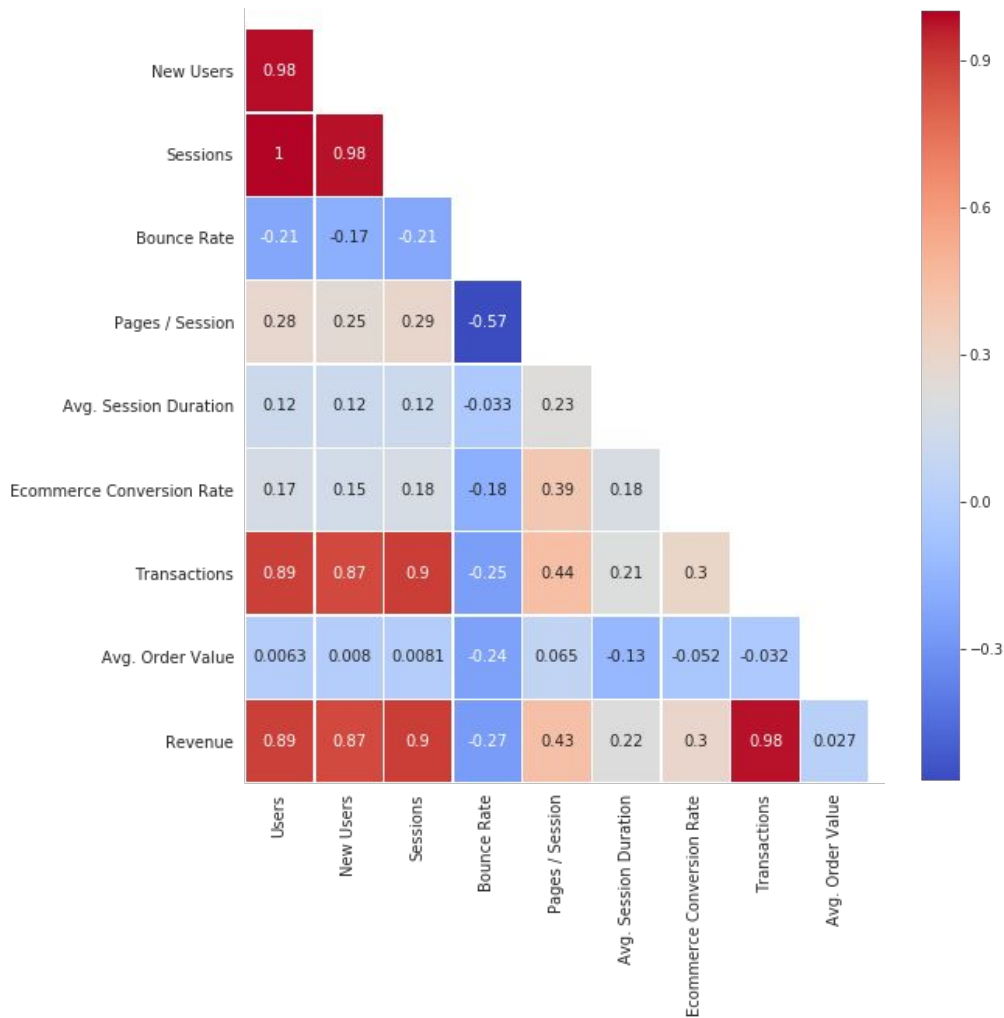
A interpretabilidade pode ser
difícil por conta da escala

Coefficiente de Correlação:

$$-1 \leq Corr(X, Y) \leq 1$$

Resolve o problema da escala.

$$Corr(X, Y) = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{dp(X)} \right) \left(\frac{y_i - \bar{y}}{dp(Y)} \right)$$



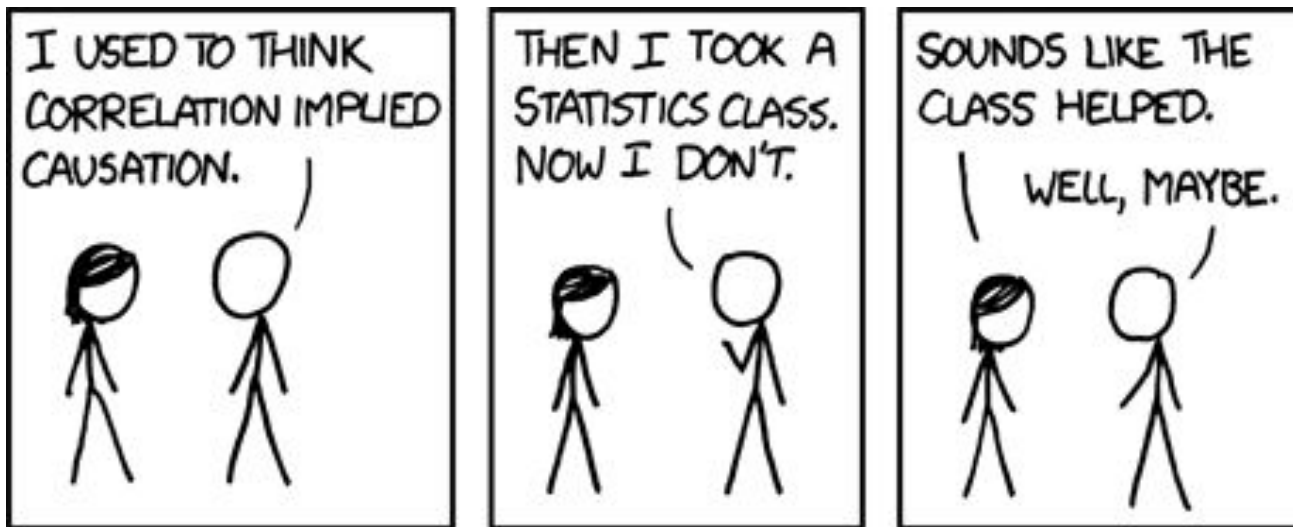
```
ax = sns.heatmap(df.corr(), annot = True)
```

Algumas correlações não são tão úteis para tomada de decisão, por exemplo:

- Users e Sessions;
- Revenue e Transactions;

Além disso, em modelos preditivos, podemos (e devemos) dispensar variáveis que são altamente correlacionadas.

Correlação e causalidade



Gracias!

kayleighmeneghini@gmail.com

