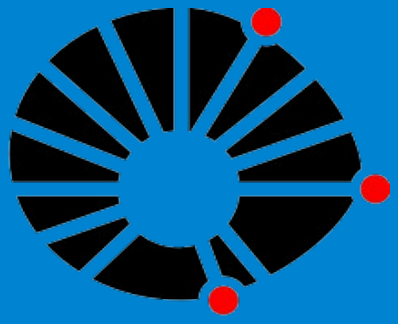
Especialização em Mineração de Dados

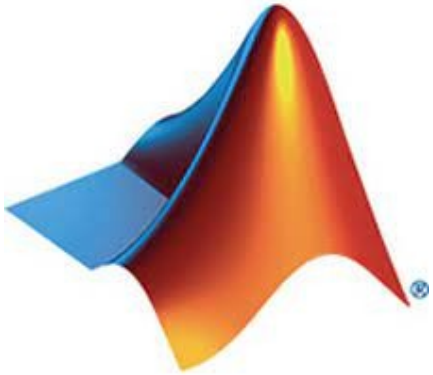
Análise de Dados

Thiago H Silva



UNICAMP

Linguagens apropriadas ajudam



Matlab

Neste curso:

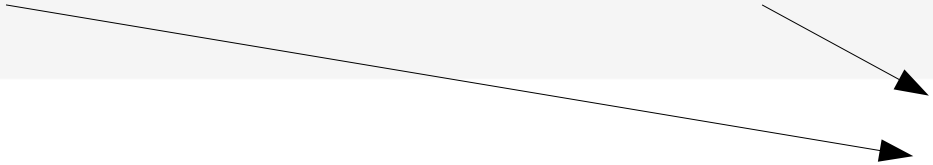
python + pandas NumPy seaborn scikit learn

Dataframe

Suportam variáveis de vários tipos

Exemplo baseado nos dados “businessToronto.csv”

city	state	postal_code	latitude	longitude	stars	review_count	is_open
Toronto	ON	M5V 1K4	43.645041	-79.395799	4.0	23	1
Toronto	ON	M6J 1J5	43.642889	-79.425429	3.0	57	1
Toronto	ON	M5R 2C7	43.670744	-79.391385	5.0	12	1



Podem ter tipos diferentes em cada coluna

Dataframe

UNICAMP

Fácil manipulação e sumarização de dados

```
dfBusiToronto['stars']
```

```
0        4.0
1        3.0
2        5.0
3        3.5
4        4.5
        ...
18901    3.5
18902    4.0
18903    3.5
18904    4.0
18905    4.5
Name: stars, Length: 18906, dtype: float64
```

Tendências centrais

A média é a medida mais comum de localização de um conjunto de pontos

- No entanto, é muito sensível a outliers
- Assim, a mediana é também muito usada

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

```
> dfBusiToronto['stars'].mean()
3.444

> dfBusiToronto['stars'].median()
3.5
```

Mediana = {(m + 1) ÷ 2}ésimo valor.

Onde m é o número de valores.

{25.0, 25.2, 25.6, 25.7, 26.1.}

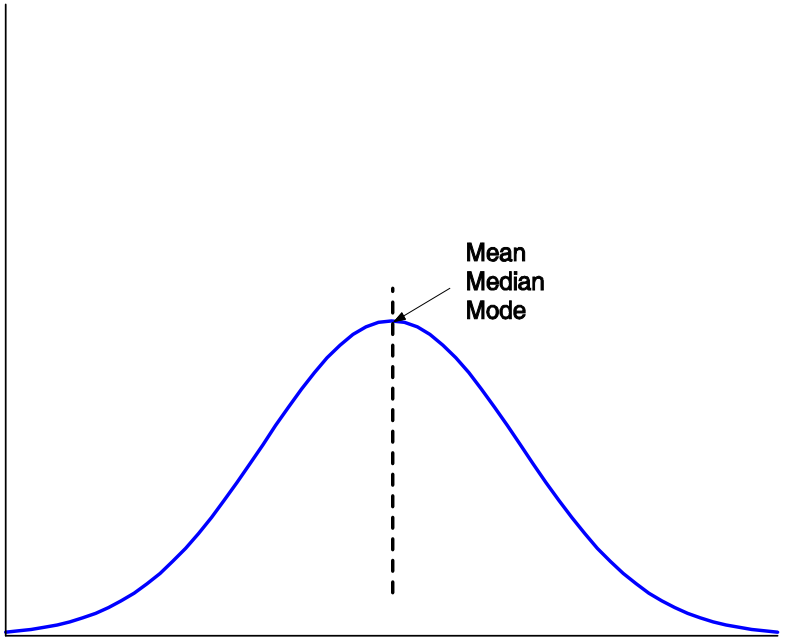
= (5 + 1) ÷ 2
= 3

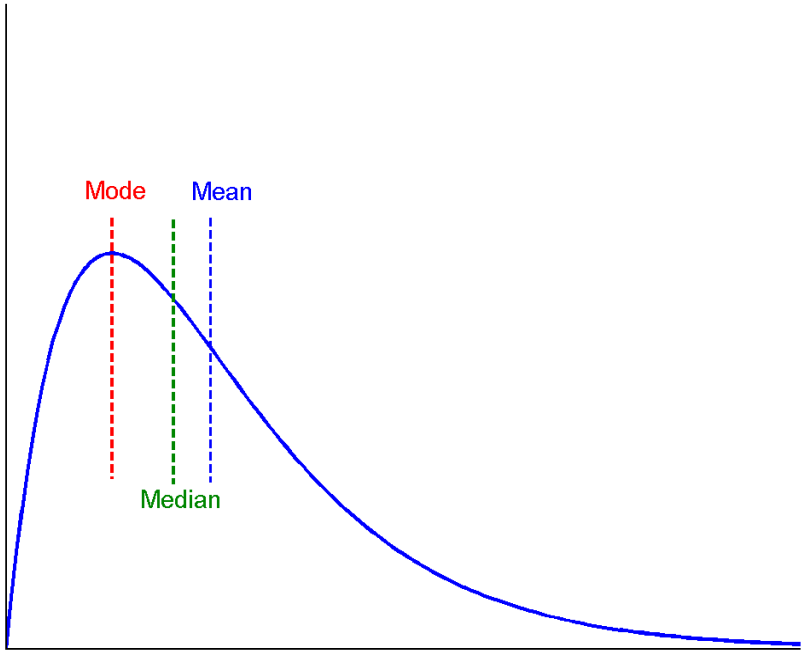
Se o valor não é inteiro calcular a média dos valores do intervalo em que se encontra

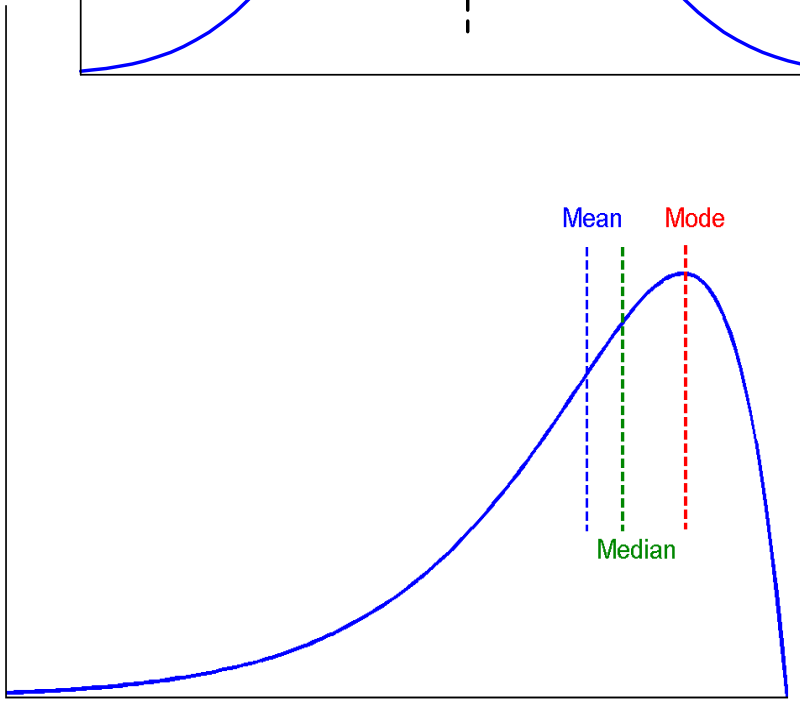
Tendências centrais

Visualização em diferentes distribuições

Medidas que podem ser úteis, por exemplo, para a criação de atributos







Visualização

Visualização de dados é uma das formas mais poderosas e atraentes de exploração de dados

- Humanos tem boa habilidade para analisar grandes quantidades de dados apresentadas visualmente
- Podemos detectar padrões e tendências
- Ajuda a detectar outliers e padrões incomuns

Passo essencial em análises de dados

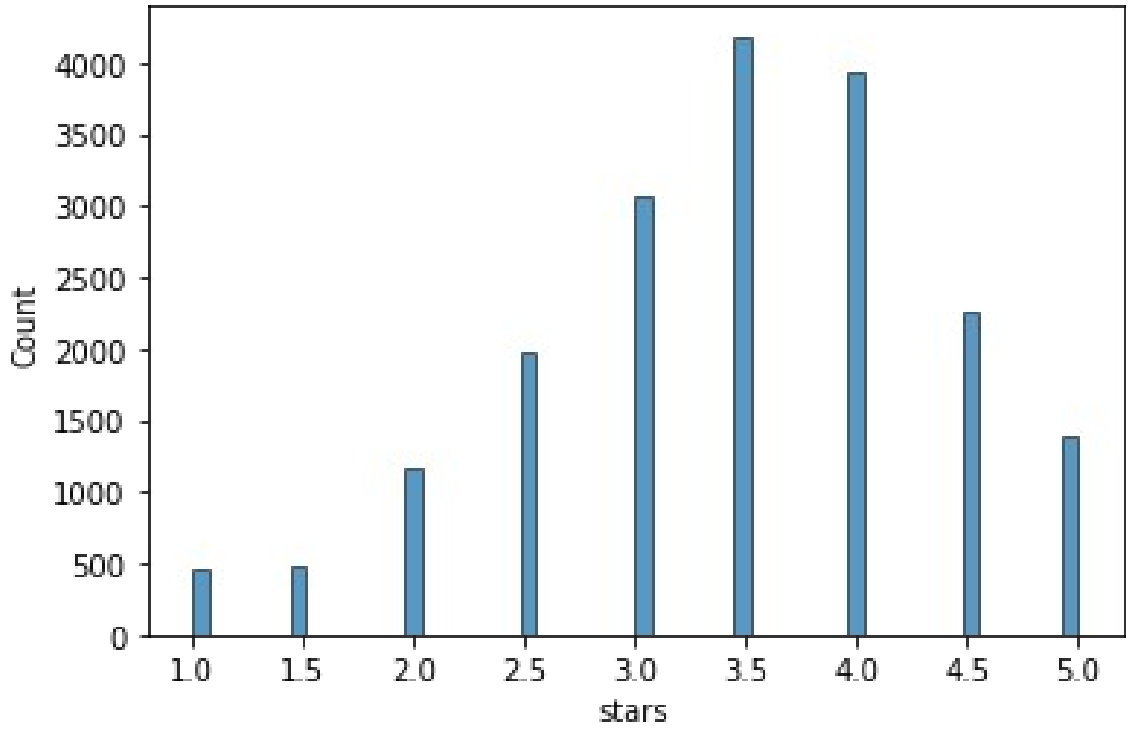
Histograma

Divide os valores em intervalos (bins) e mostra um gráfico de barras do número de objetos em cada bin

A altura da barra indica o número de objetos

O formato do histograma depende do número de bins

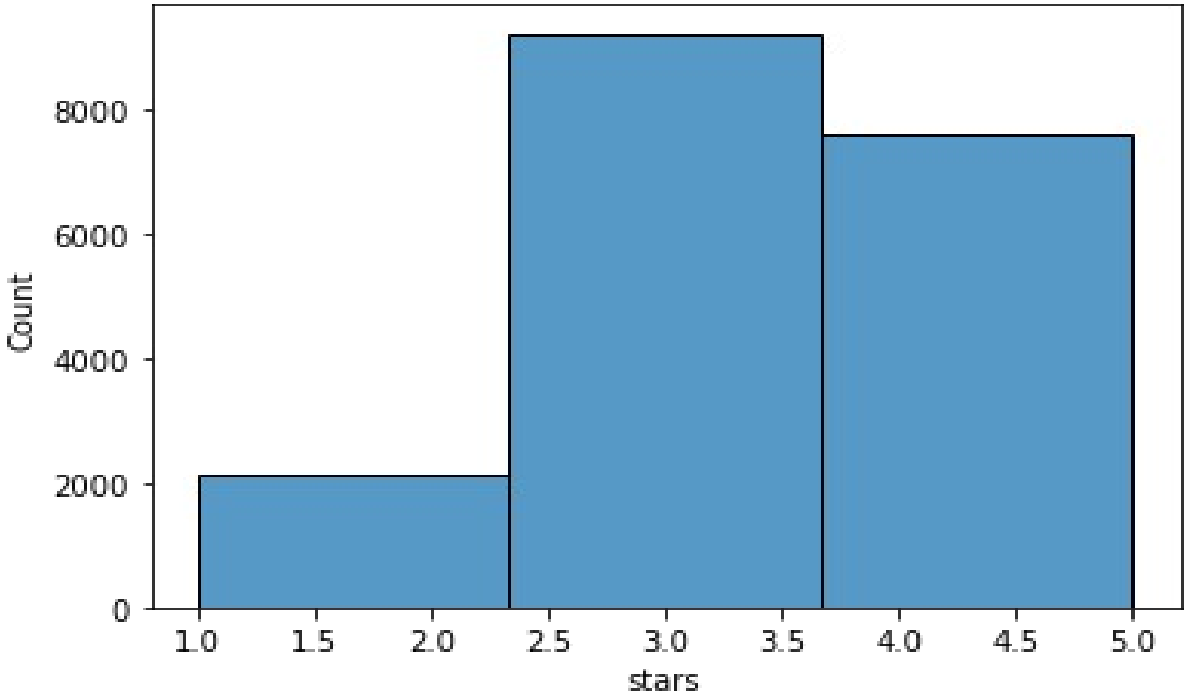
```
sns.histplot(dfBusiToronto['stars'])
```



Histograma

Definindo os bins

```
sns.histplot(dfBusiToronto['stars'],  bins=3)
```



CDF

Função de Distribuição Cumulativa
(CDF cumulative distribution function)

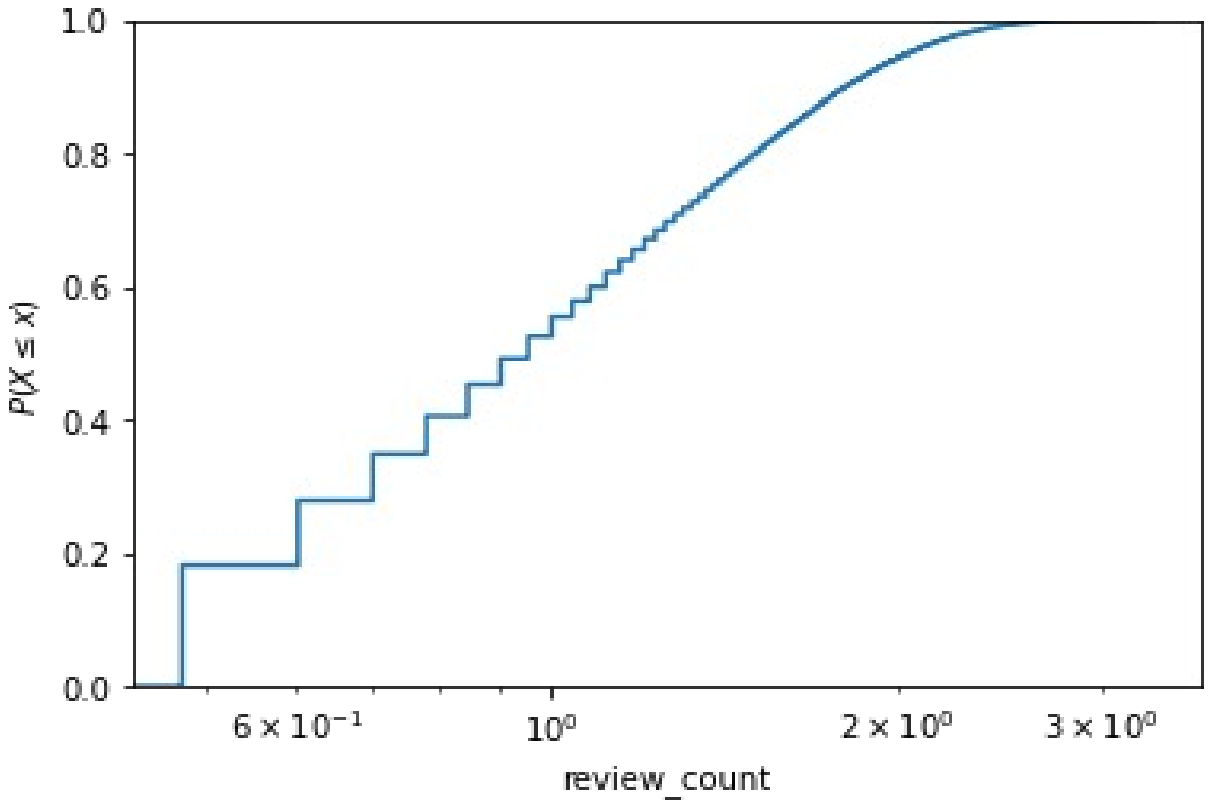
Mapeia um valor para uma probabilidade cujo resultado é menor ou igual a x:

$$F_X(x) = \mathrm{P}(X \leq x),$$

Probabilidade de uma variável aleatória X assumir um valor menor ou igual a x

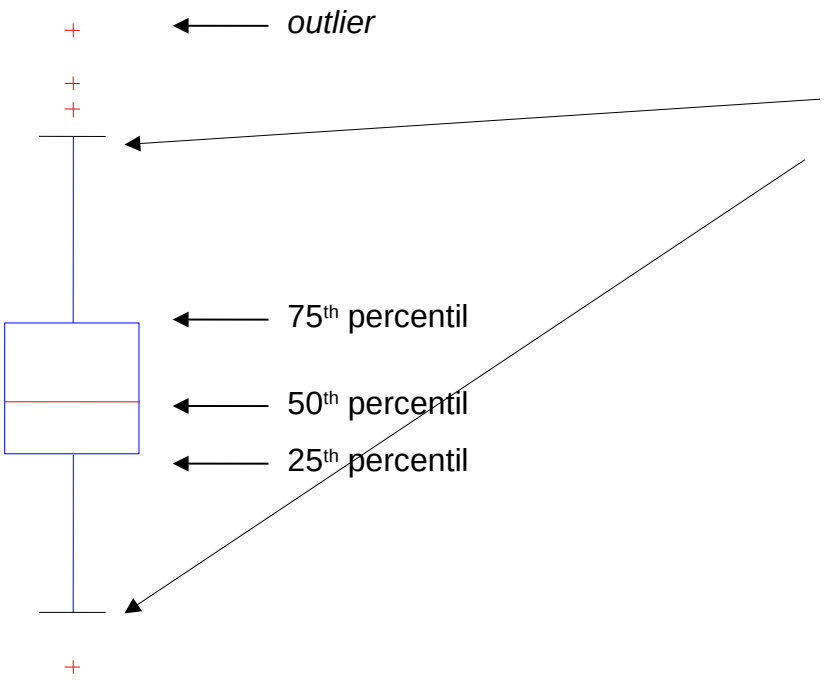
CDF

```
ax = sns.ecdfplot(data = dfBusiToronto, x='review_count', log_scale=True )
ax.set(ylabel=r'$P(X \leq x)$')
plt.show()
```



Box plot

Outra forma de mostrar uma distribuição de dados

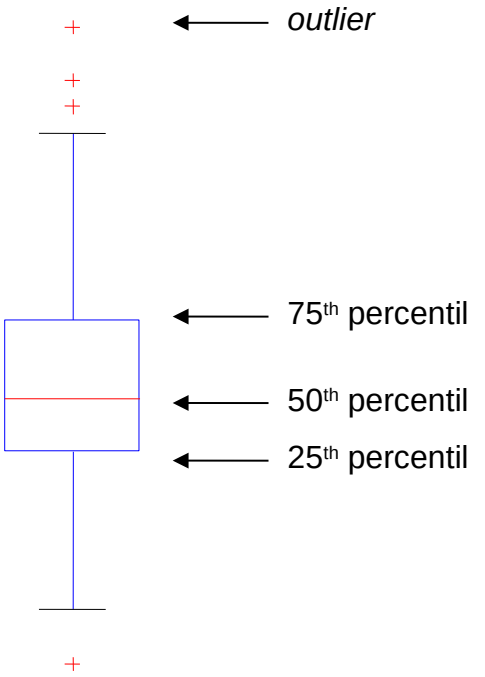


upper whisker = min(max(x), Q_3 + 1.5 * IQR)

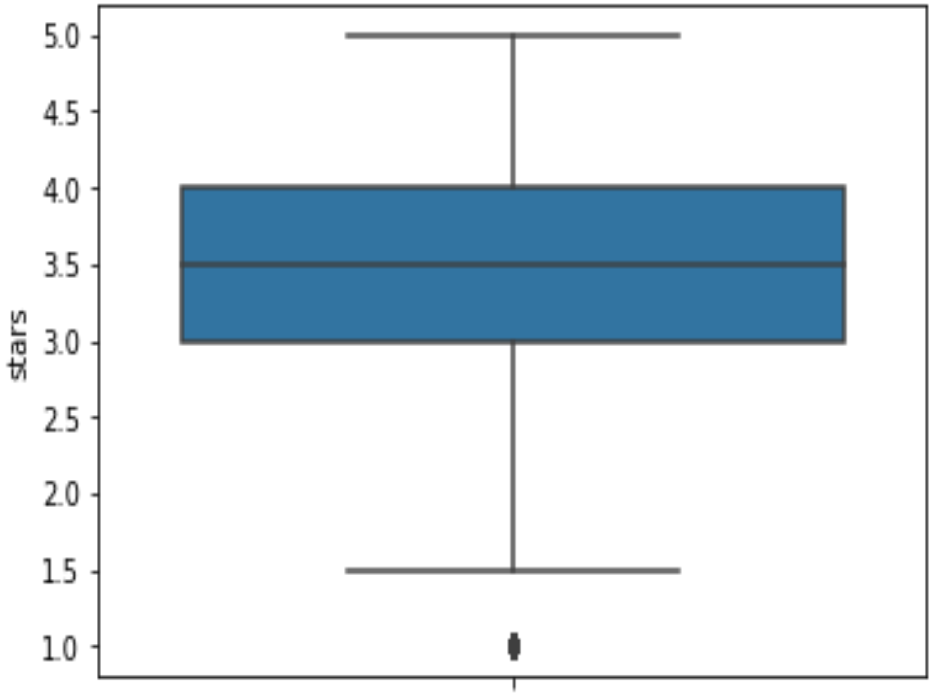
lower whisker = max(min(x), Q_1 – 1.5 * IQR)

Box plot

Outra forma de mostrar uma distribuição de dados

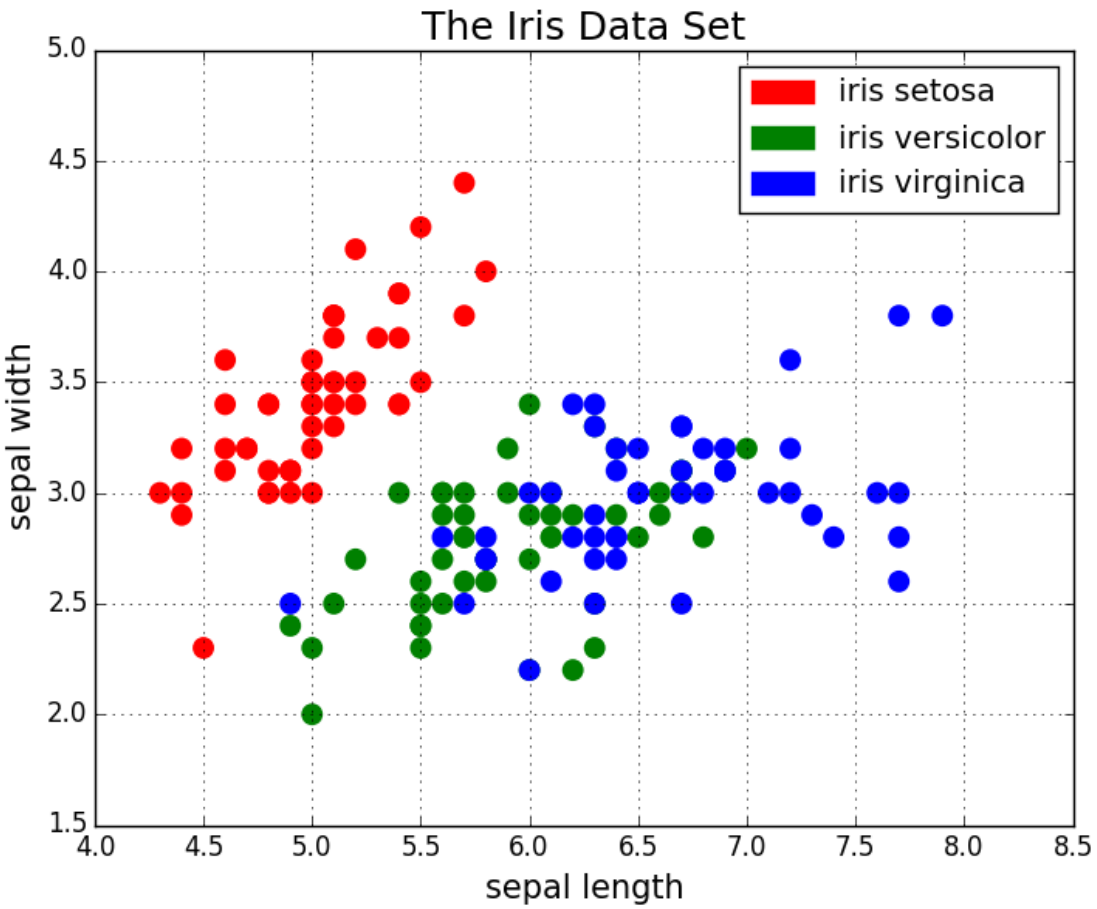



```
> sns.boxplot(data=dfBusiToronto, y='stars')
```



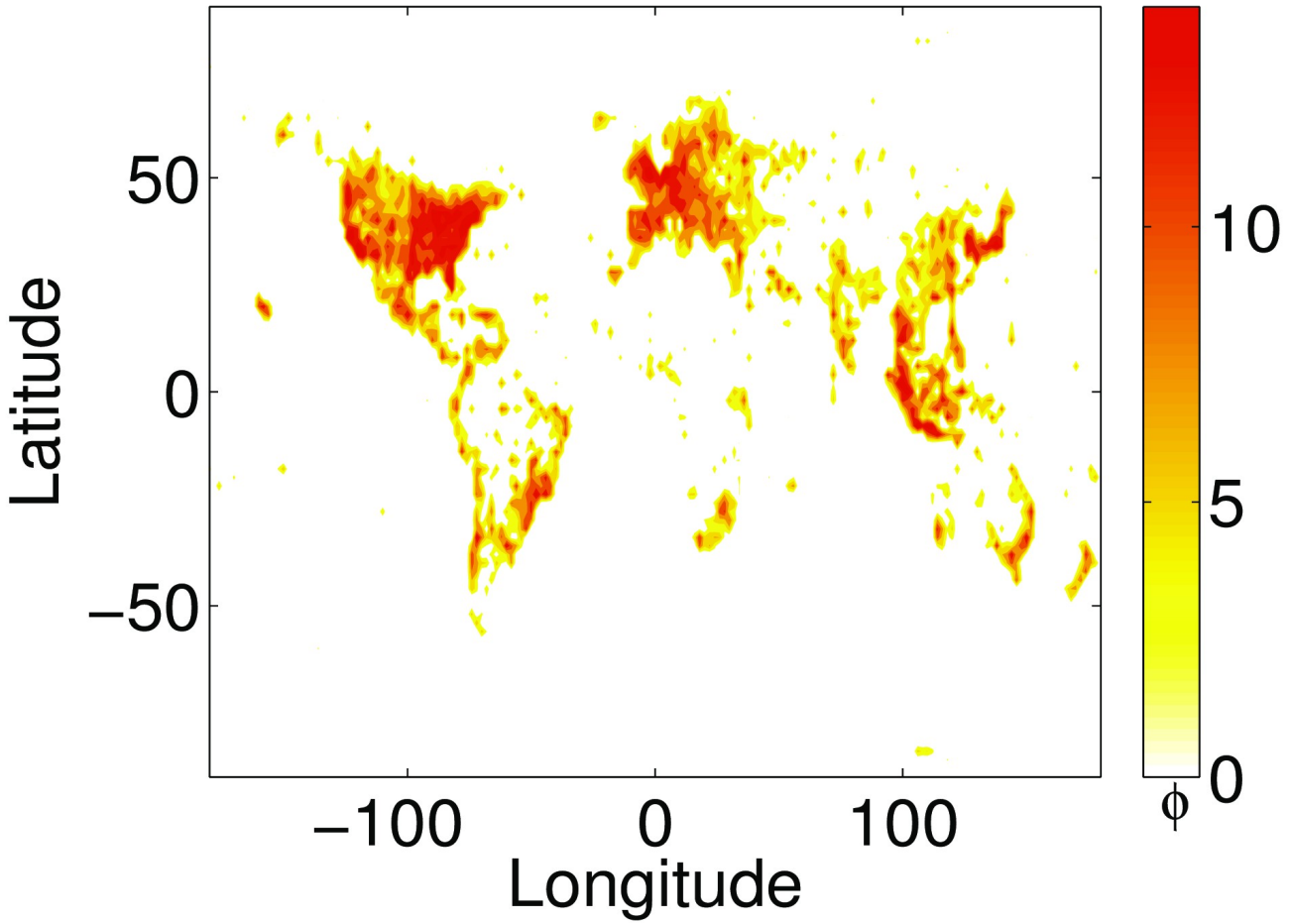
Scatter plot

Os valores dos atributos determinam a posição do objeto

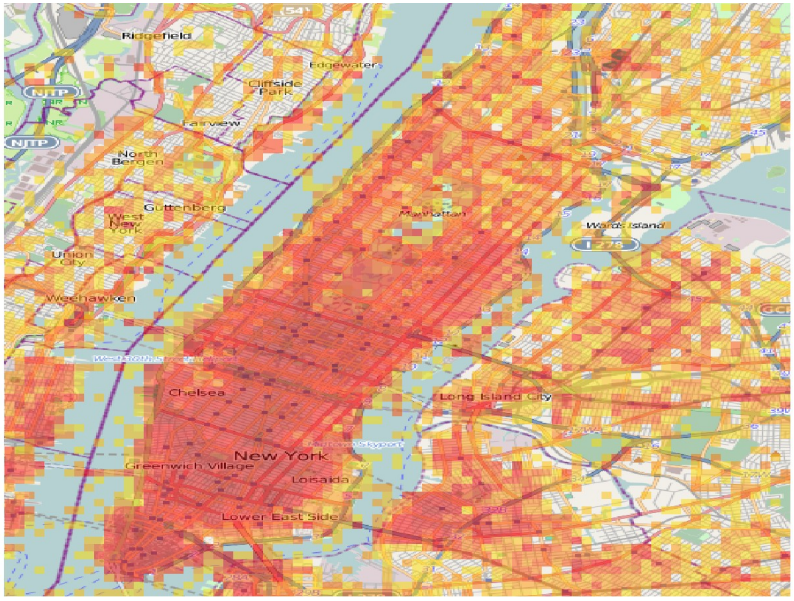


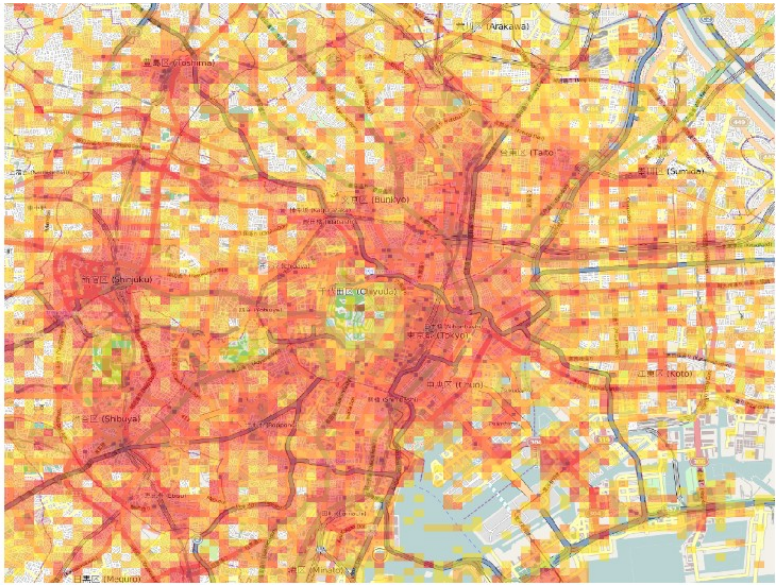
Qual é a tendência observada?

Heatmaps

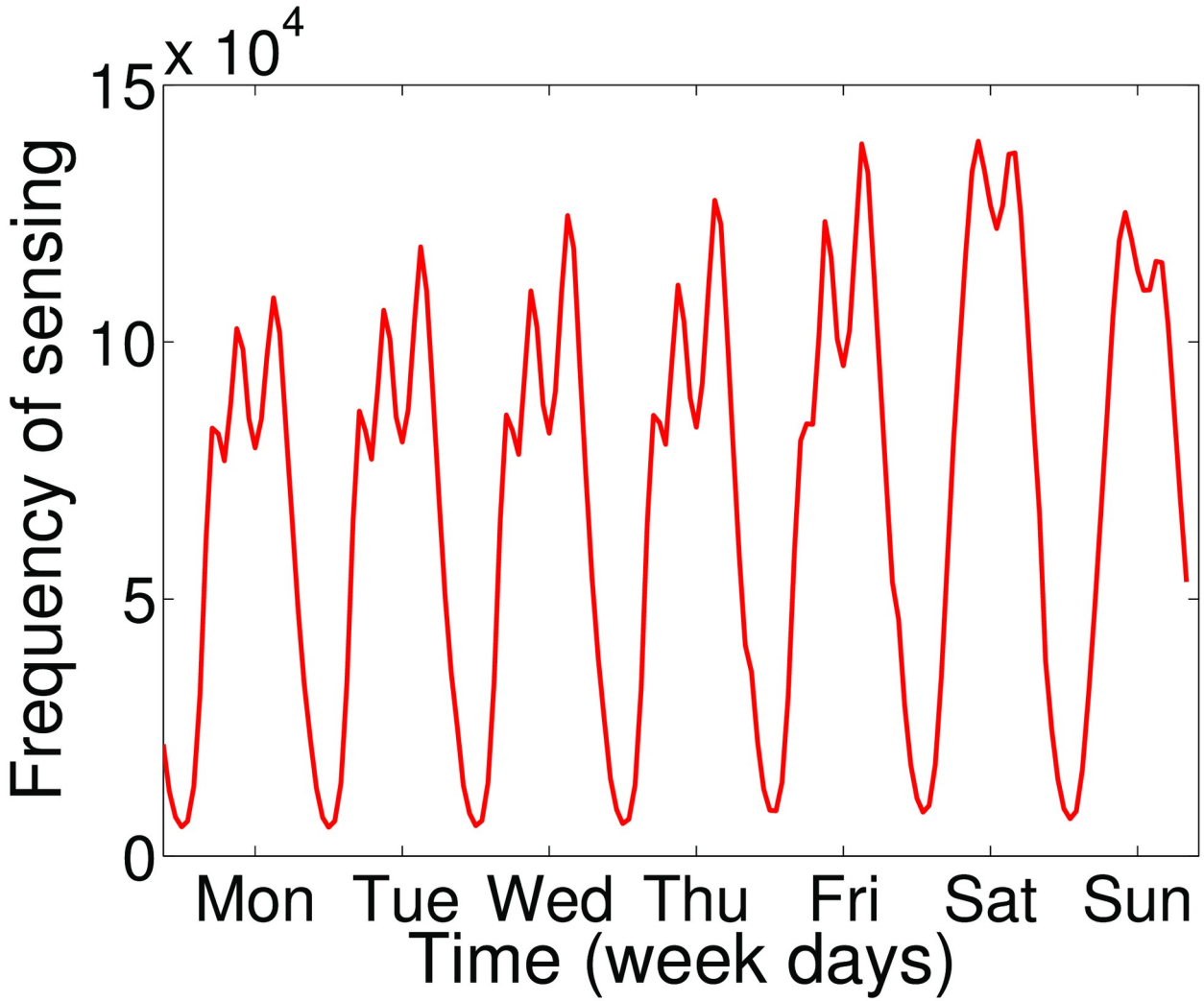


Heatmaps


Ridgefield
Edgewater
Cliffside Park
Fairview
North Bergen
Guttenberg
West New York
Manhattan
Union City
Weehawken
Chelsea
Long Island City
New York
Greenwich Village
Loisaida
Lower East Side

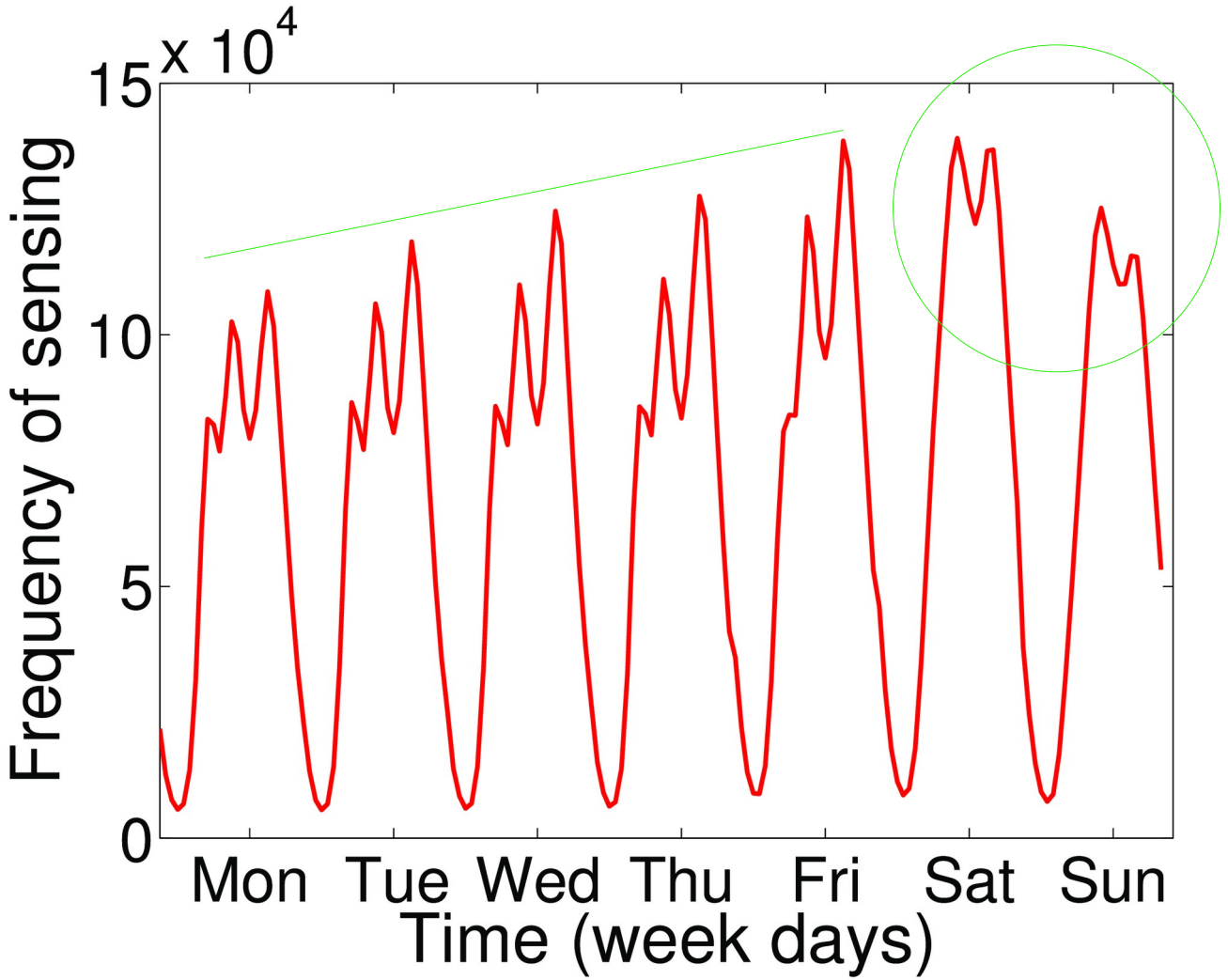

(Arakawa)
(Taito)
(Bunkyo)
(Shinjuku)
(Chiyoda)
(Shibuya)
(Koto)
(Meguro)

Séries temporais



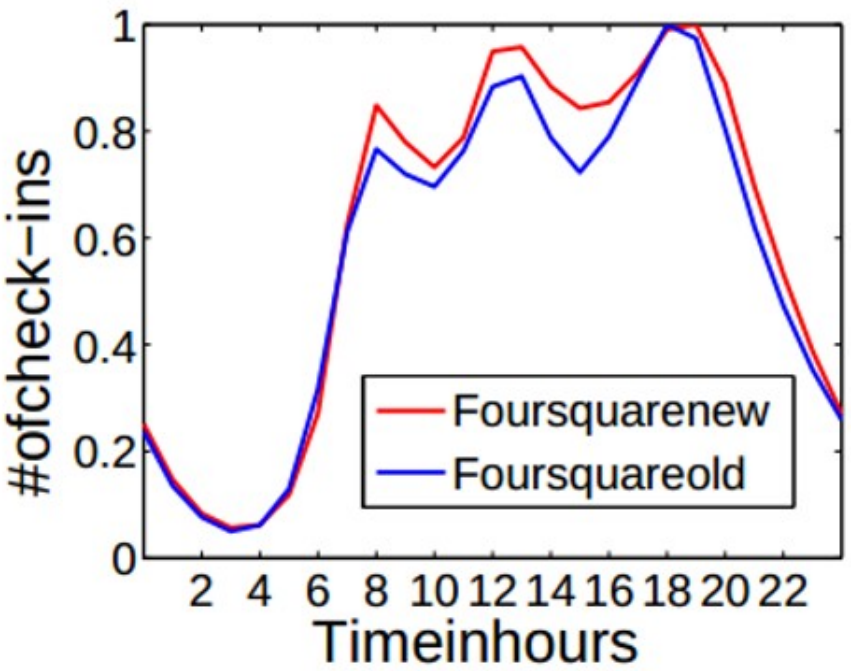
Check-ins do Foursquare - Global

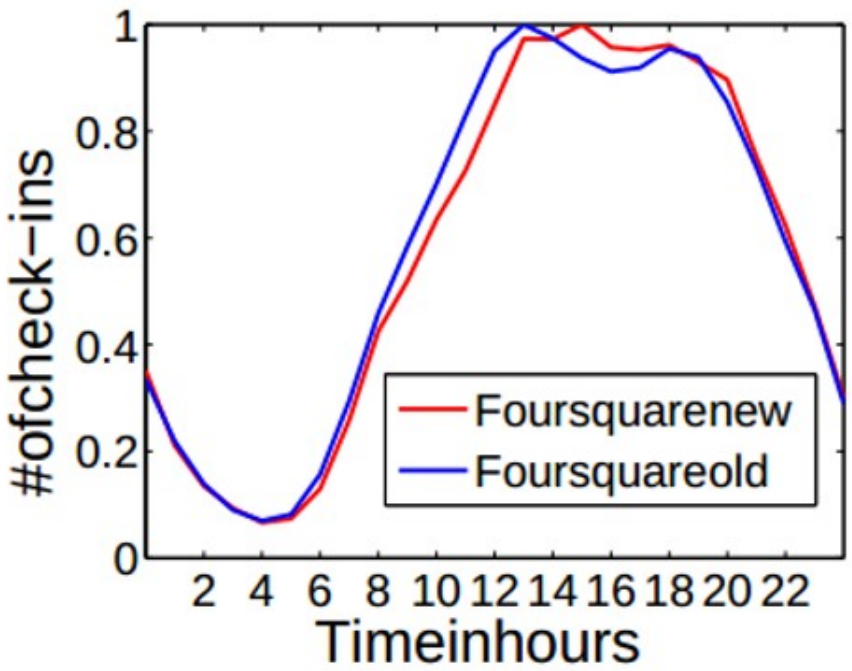
Séries temporais



Check-ins do Foursquare - Global

Séries temporais





Informações complementares

Outros exemplos de visualizações:
http://www.datavizcatalogue.com/

Exemplos de análise de dados:

An Empirical Study of Geographic User Activity Patterns in Foursquare. ICWSM 2011

A picture of Instagram is Worth More than a Thousand Words: Workload Characterization and Application. DCOSS 2013