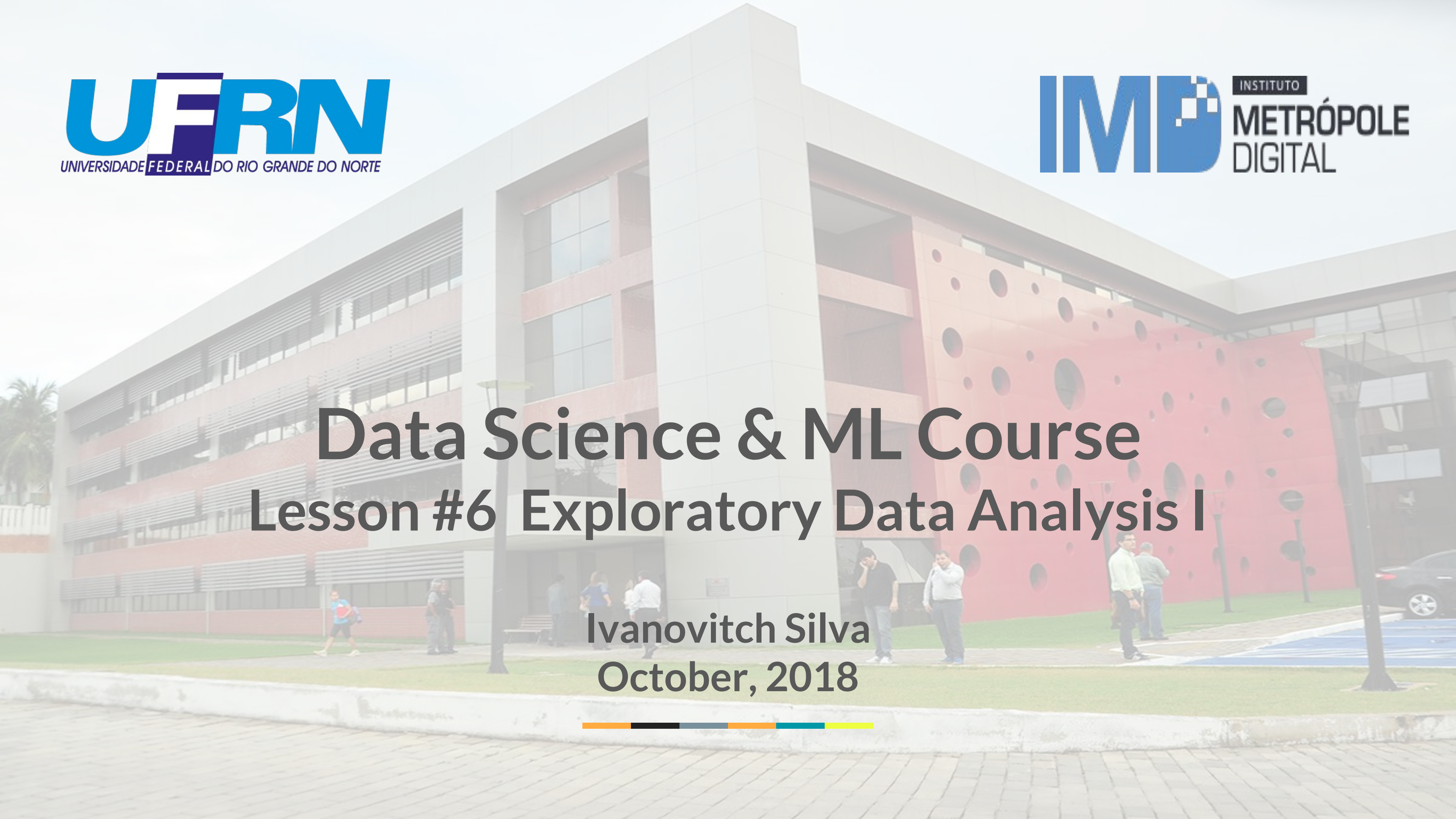


UFRN
UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

IMD INSTITUTO METRÓPOLE DIGITAL

Data Science & ML Course

Lesson #6 Exploratory Data Analysis I

Ivanovitch Silva
October, 2018

Agenda

- Case study: unemployment rate, movie ratings
- Tabular vs Visual representation
- Matplotlib
- Line, Bar and Scatter Plots

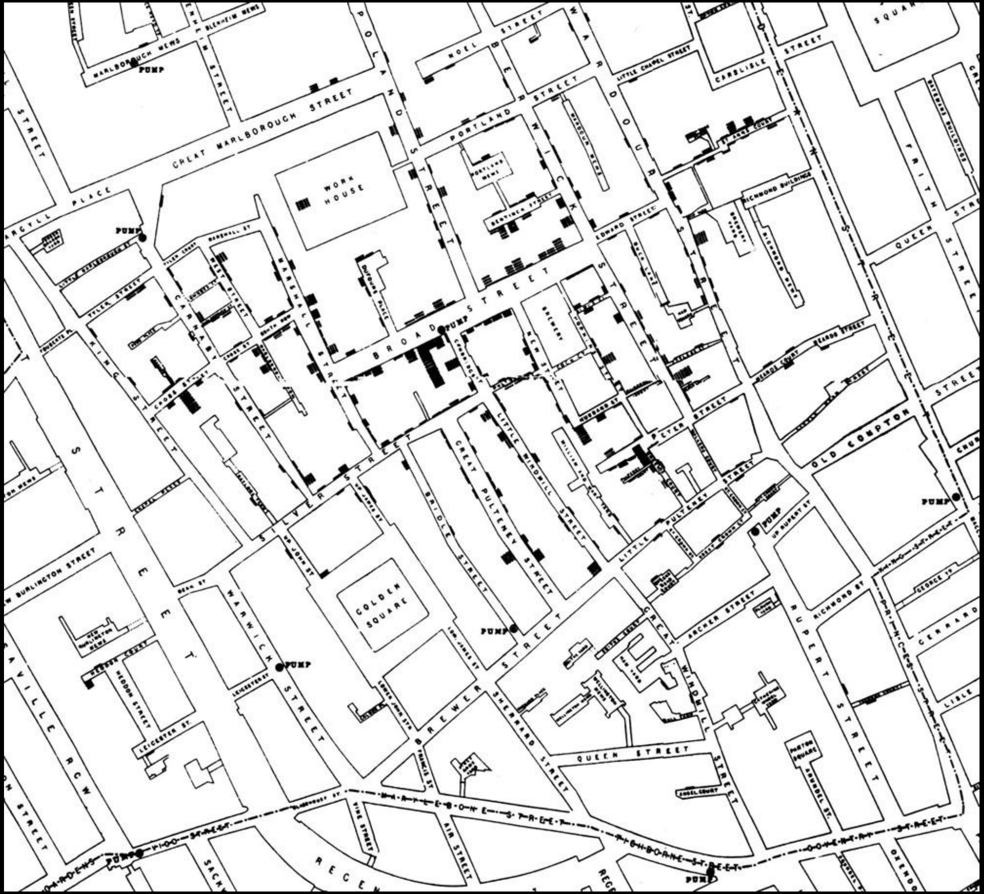
Update from repository

```
git clone https://github.com/ivanovitchm/datascience2machinelearning.git
```

Or ….

```
git pull
```

One Picture Worth Ten Thousand Words



London, 1854

NumPy

pandas

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

Keras

TensorFlow

scikit learn

pgmpy

NetworkX

PyGSP

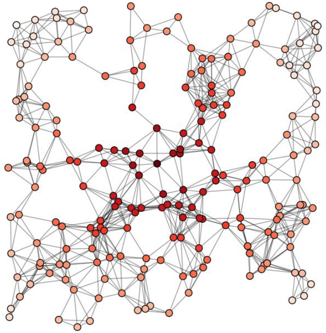
PySpark

DASK

matplotlib

Folium

GeoJSON



Seaborn

plotly

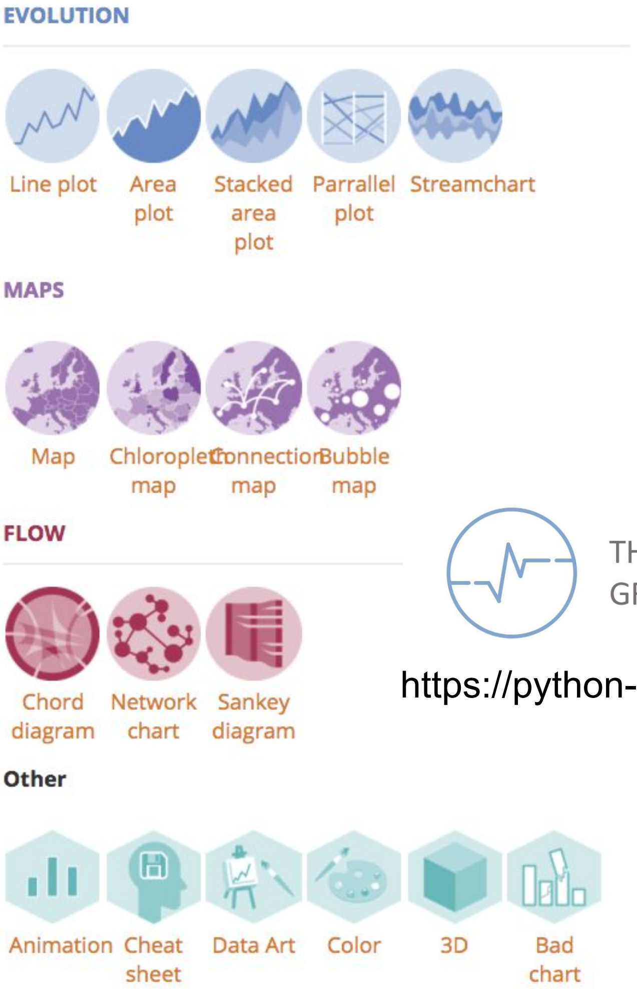
Requests

bokeh

Leaflet

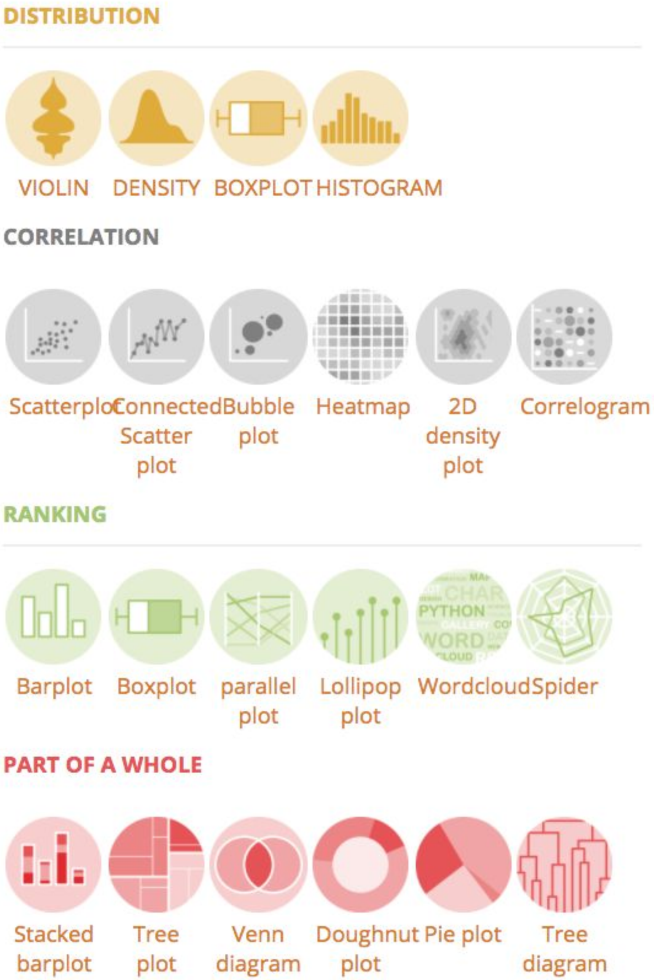
Beautiful Soap

Scrapy

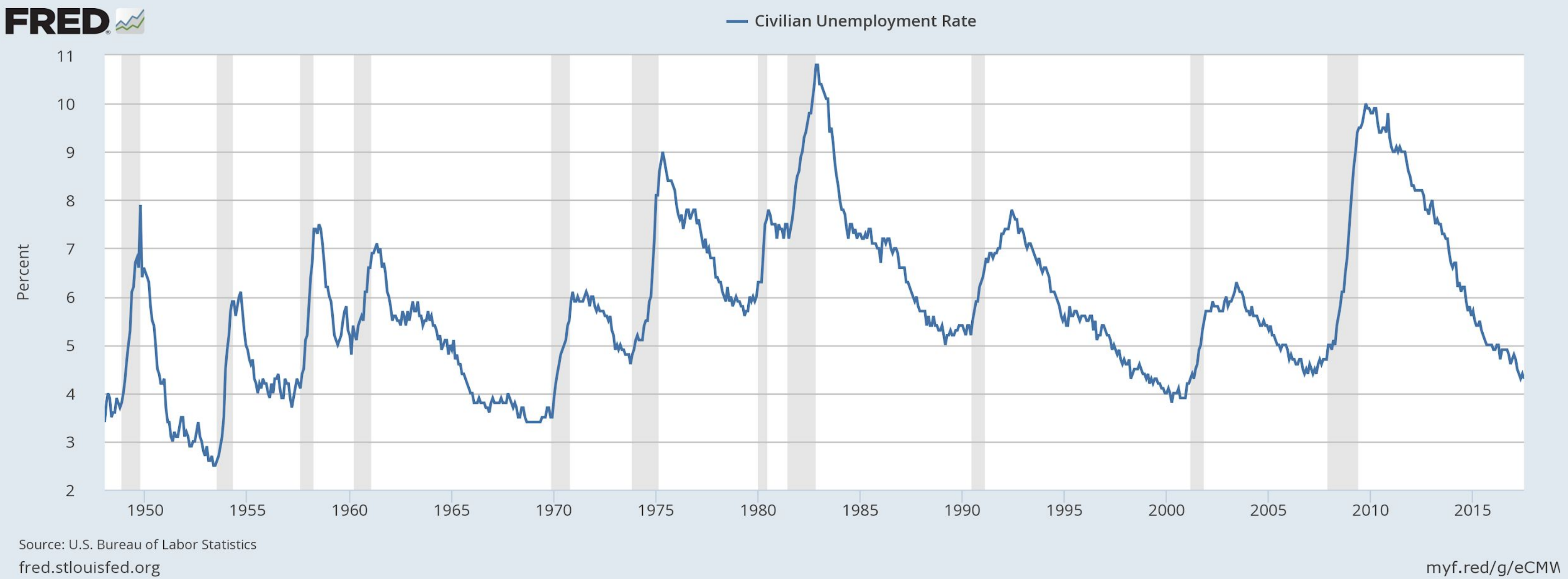


THE PYTHON
GRAPH GALLERY

https://python-graph-gallery.com/



Case study: unemployment rate (US)

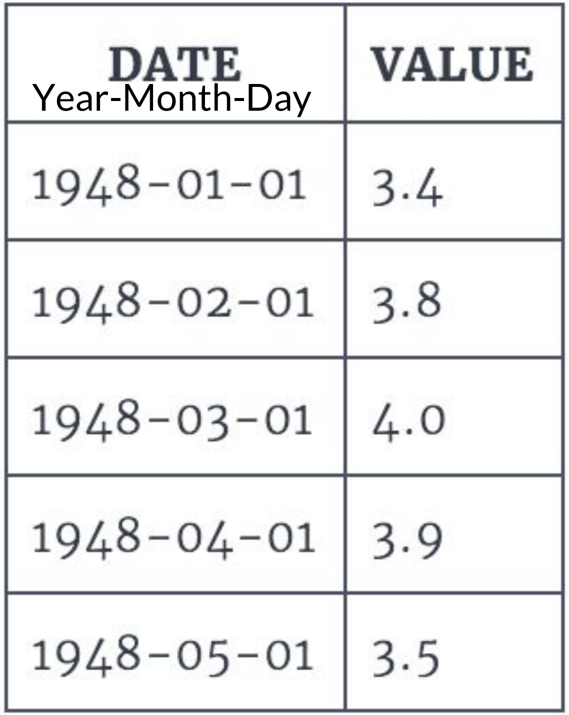


Source: U.S. Bureau of Labor Statistics

fred.stlouisfed.org

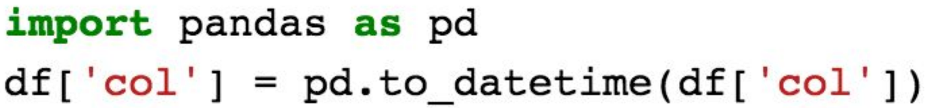
myf.red/g/eCMW

Investigating the dataset

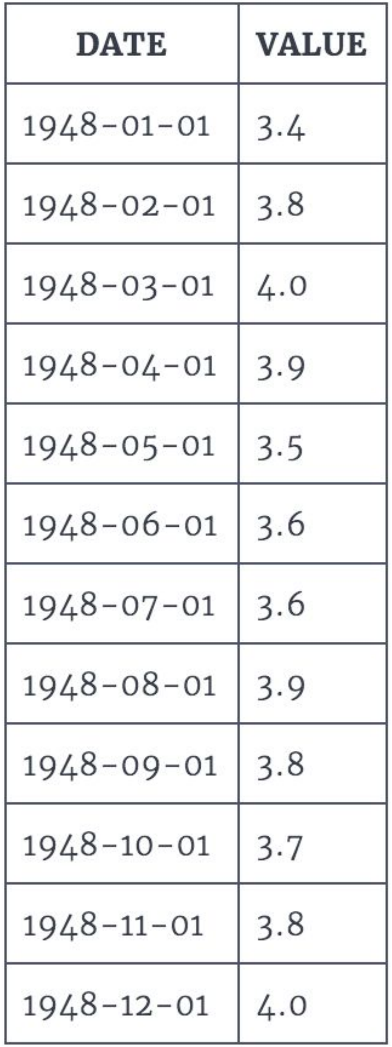


DATE Year-Month-Day	VALUE
1948-01-01	3.4
1948-02-01	3.8
1948-03-01	4.0
1948-04-01	3.9
1948-05-01	3.5

Conversion of types (Object to Datetime)



```
import pandas as pd
df['col'] = pd.to_datetime(df['col'])
```

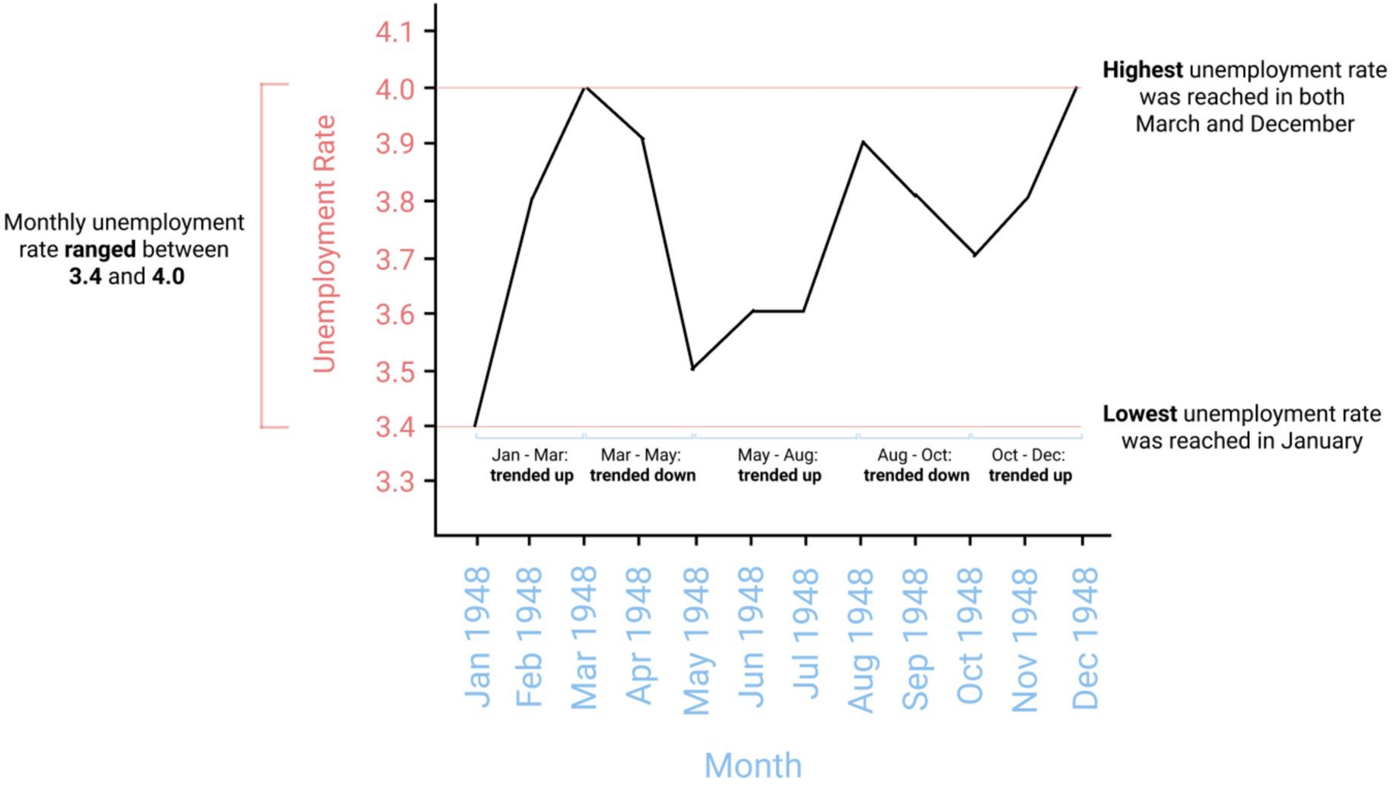


DATE	VALUE
1948-01-01	3.4
1948-02-01	3.8
1948-03-01	4.0
1948-04-01	3.9
1948-05-01	3.5
1948-06-01	3.6
1948-07-01	3.6
1948-08-01	3.9
1948-09-01	3.8
1948-10-01	3.7
1948-11-01	3.8
1948-12-01	4.0

Observation from the table representation

- What is the minimum value?
- What is the maximum value?
- Is there seasonality?
- What are the trend up periods?
- What are the trend down periods?
- Is the table representation really useful?

Visual representation



Monthly unemployment rate **ranged** between **3.4** and **4.0**

Highest unemployment rate was reached in both March and December

Lowest unemployment rate was reached in January

Jan - Mar: **trended up**

Mar - May: **trended down**

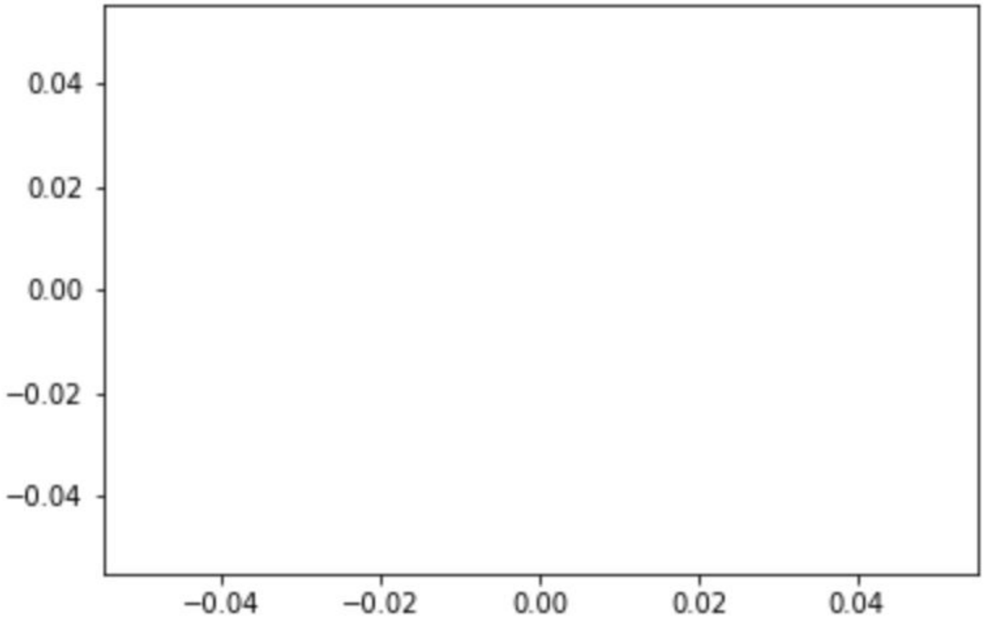
May - Aug: **trended up**

Aug - Oct: **trended down**

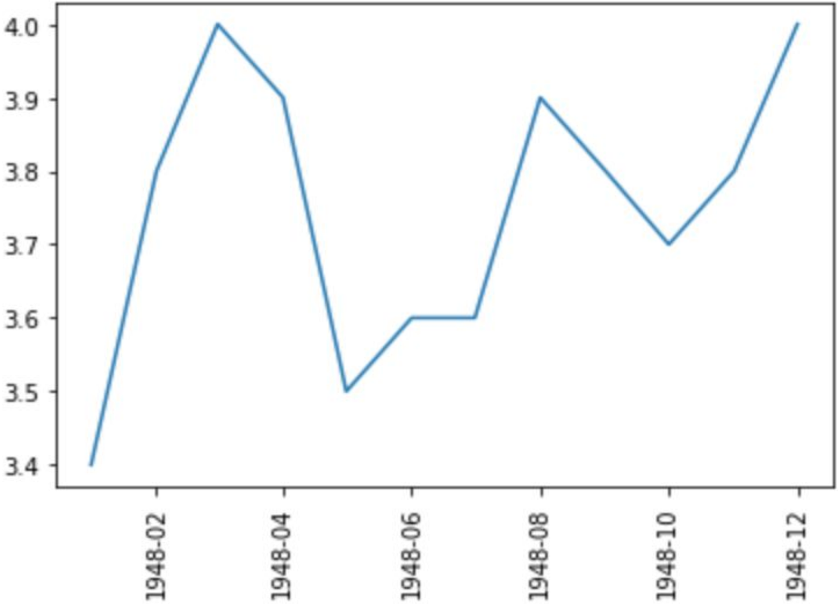
Oct - Dec: **trended up**

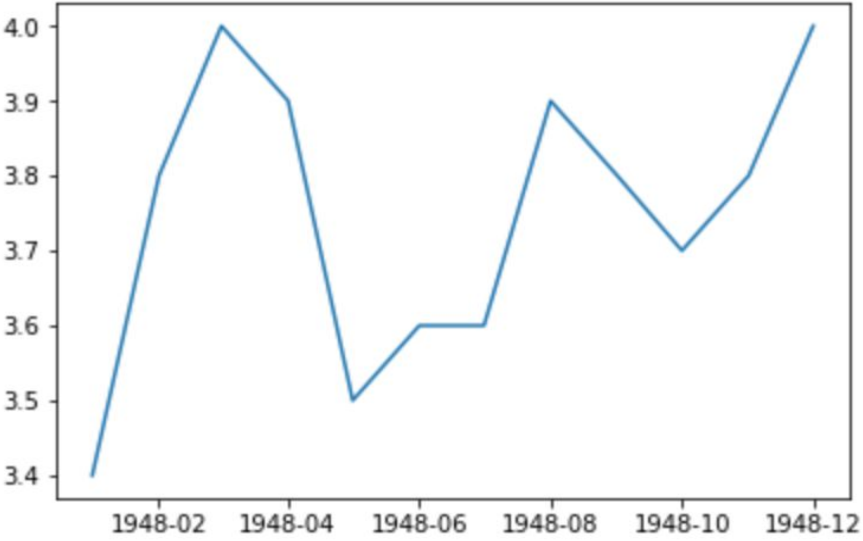
matplotlib

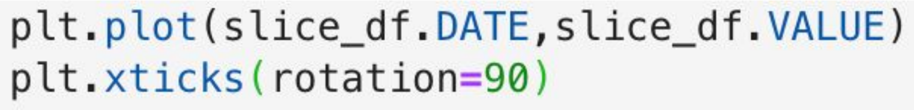
```
import matplotlib.pyplot as plt
plt.plot()
plt.show()
```



Adding and Fixing Axis Ticks

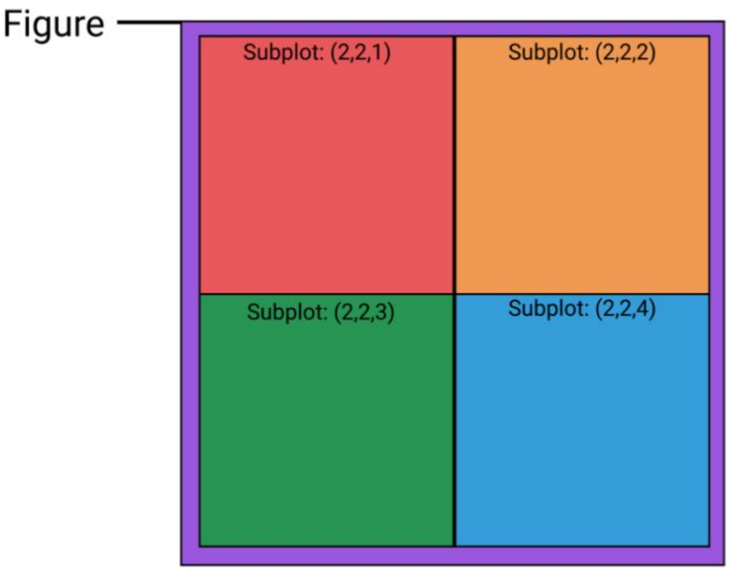


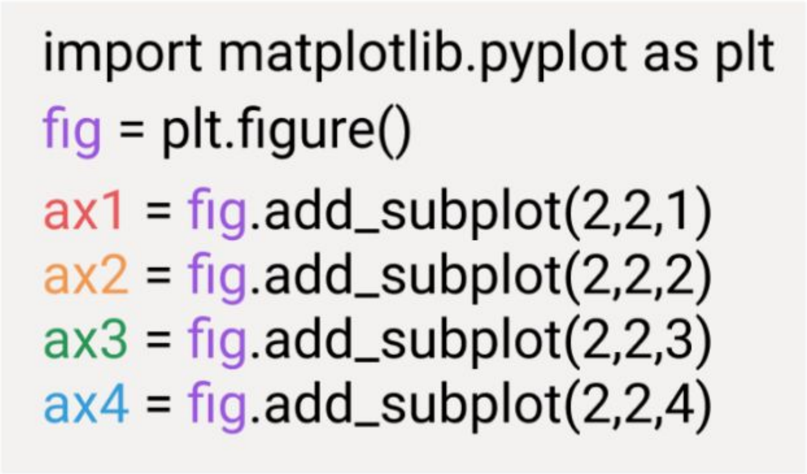




```
plt.plot(slice_df.DATE,slice_df.VALUE)
plt.xticks(rotation=90)
```

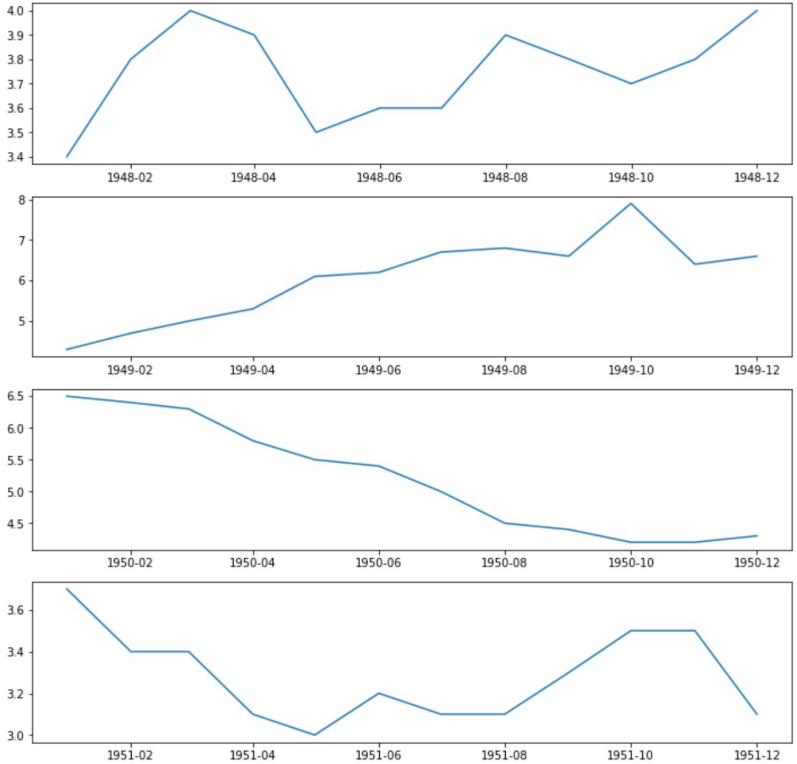
Multiples Charts

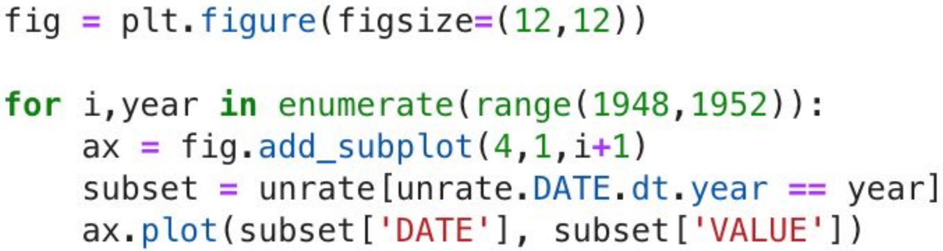




```
import matplotlib.pyplot as plt
fig = plt.figure()
ax1 = fig.add_subplot(2,2,1)
ax2 = fig.add_subplot(2,2,2)
ax3 = fig.add_subplot(2,2,3)
ax4 = fig.add_subplot(2,2,4)
```

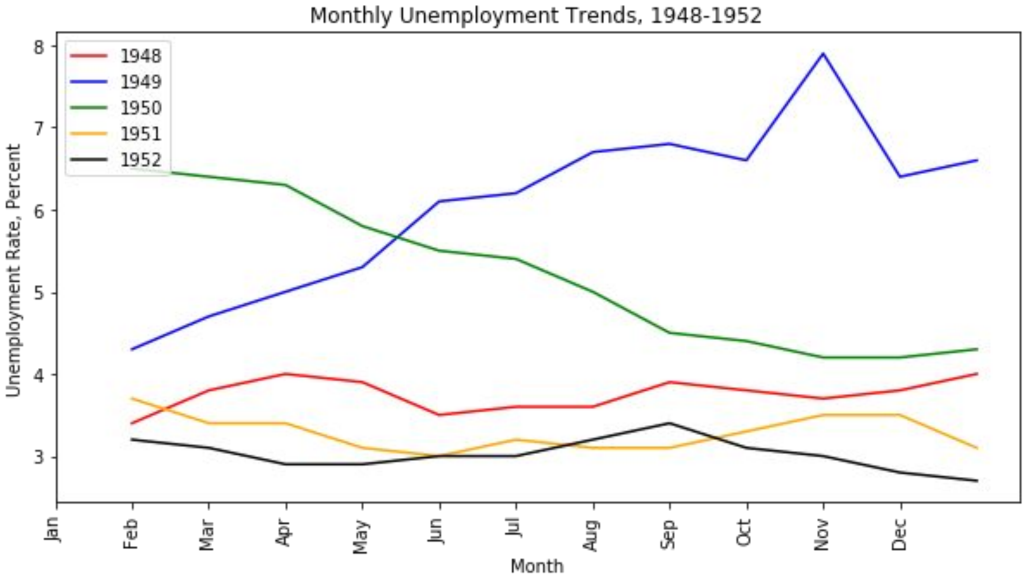
Comparing across more years

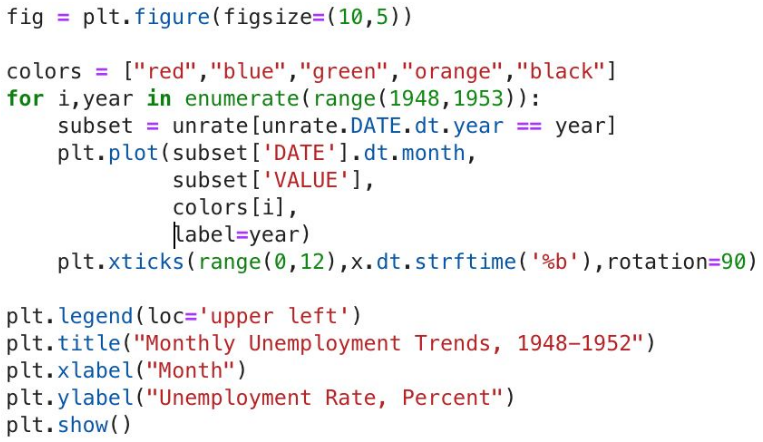




```
fig = plt.figure(figsize=(12,12))

for i,year in enumerate(range(1948,1952)):
    ax = fig.add_subplot(4,1,i+1)
    subset = unrate[unrate.DATE.dt.year == year]
    ax.plot(subset['DATE'], subset['VALUE'])
```

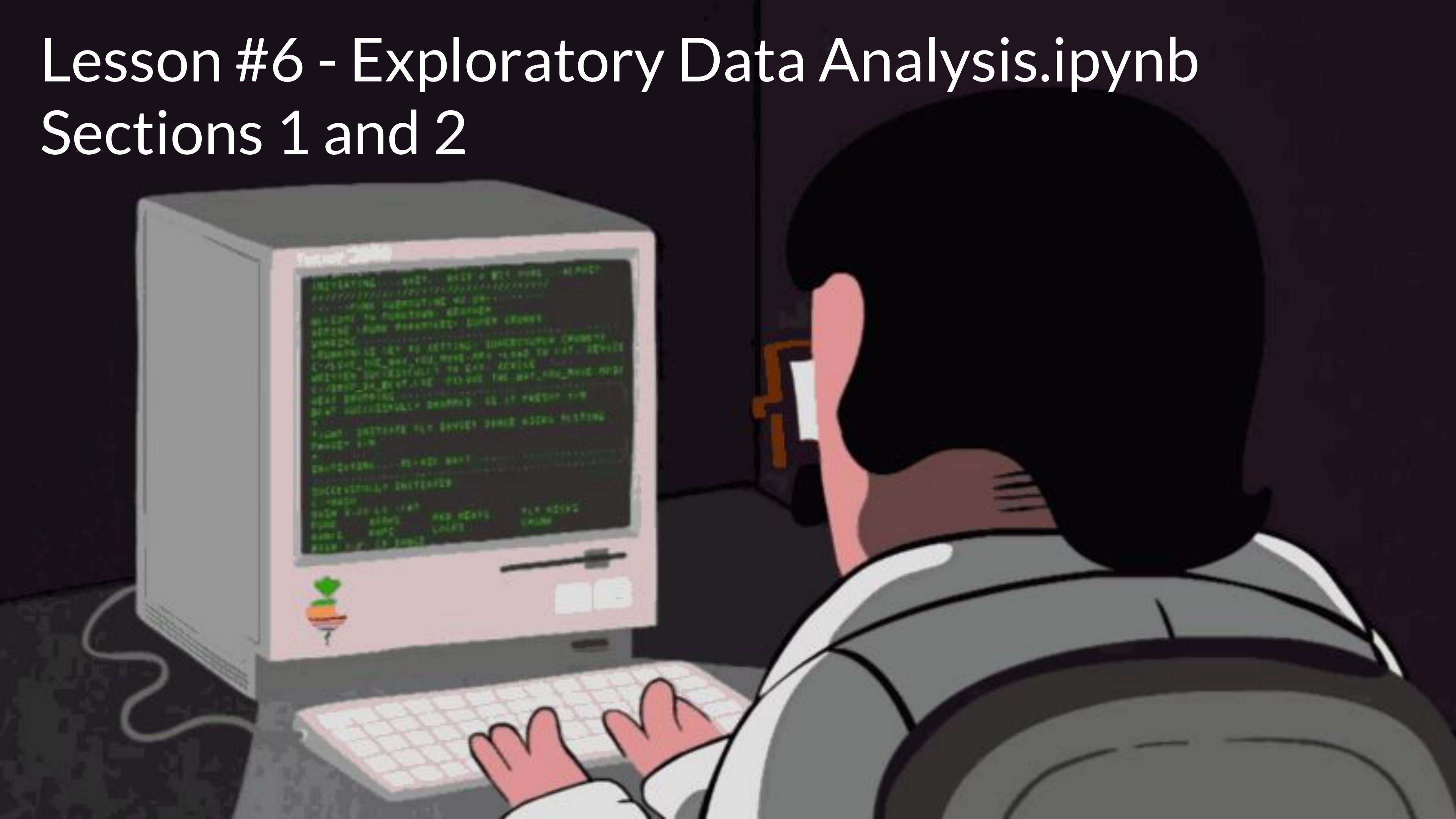
Overlaying line charts

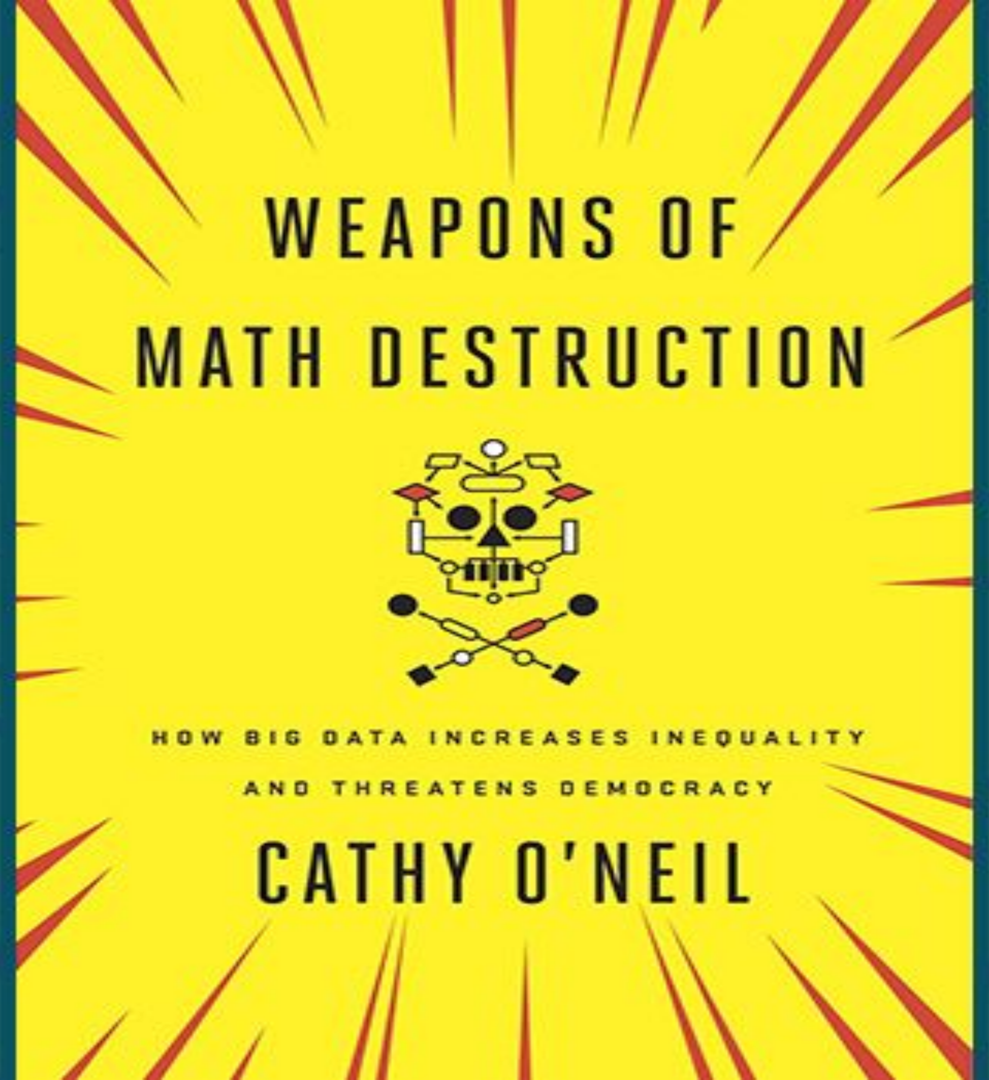


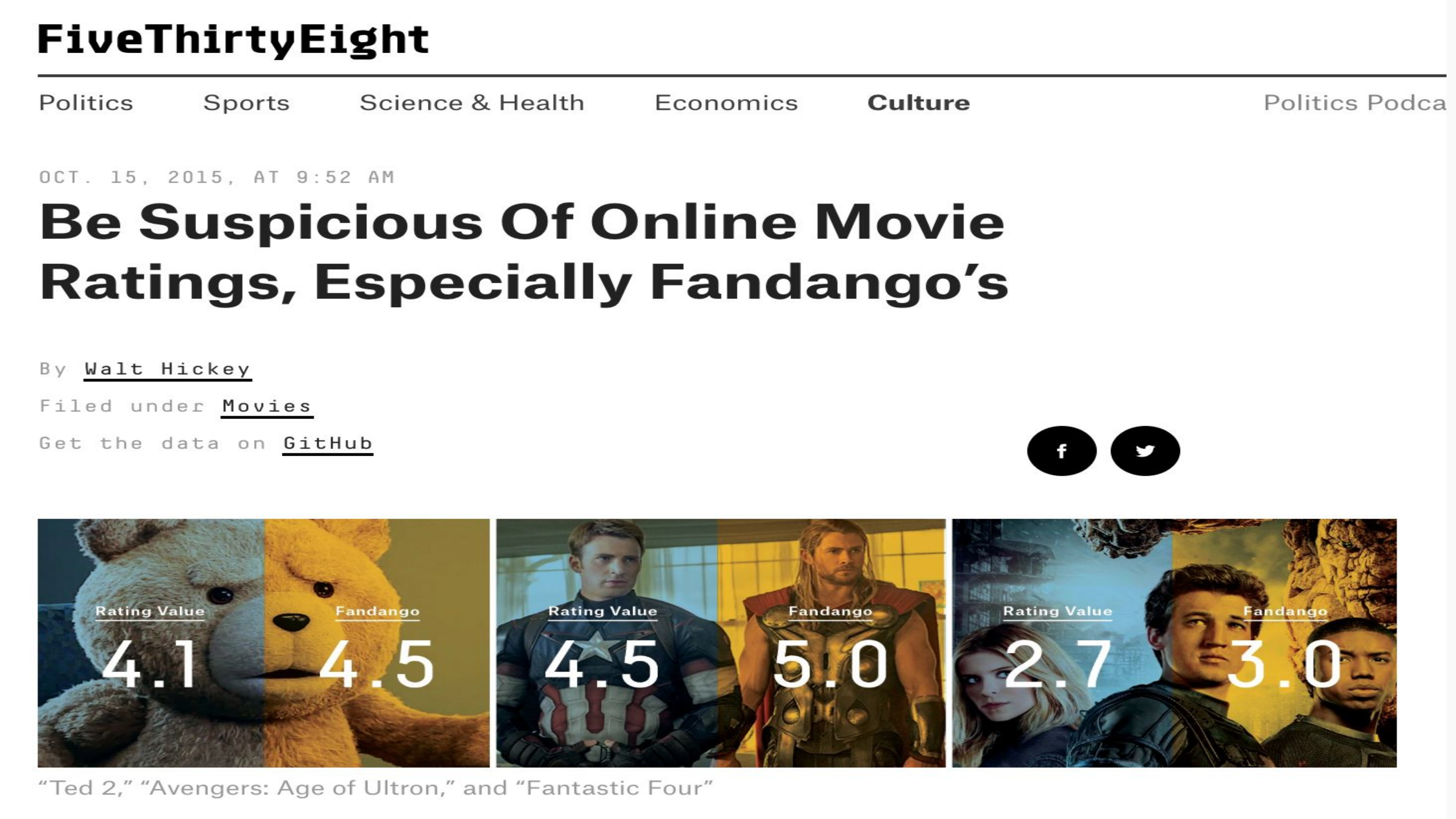


```
fig = plt.figure(figsize=(10,5))

colors = ["red","blue","green","orange","black"]
for i,year in enumerate(range(1948,1953)):
    subset = unrate[unrate.DATE.dt.year == year]
    plt.plot(subset['DATE'].dt.month,
             subset['VALUE'],
             colors[i],
             label=year)
    plt.xticks(range(0,12),x.dt.strftime('%b'),rotation=90)

plt.legend(loc='upper left')
plt.title("Monthly Unemployment Trends, 1948-1952")
plt.xlabel("Month")
plt.ylabel("Unemployment Rate, Percent")
plt.show()
```

Lesson #6 - Exploratory Data Analysis.ipynb
Sections 1 and 2




WEAPONS OF MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY
CATHY O'NEIL



OCT. 15, 2015, AT 9:52 AM

Be Suspicious Of Online Movie Ratings, Especially Fandango’s

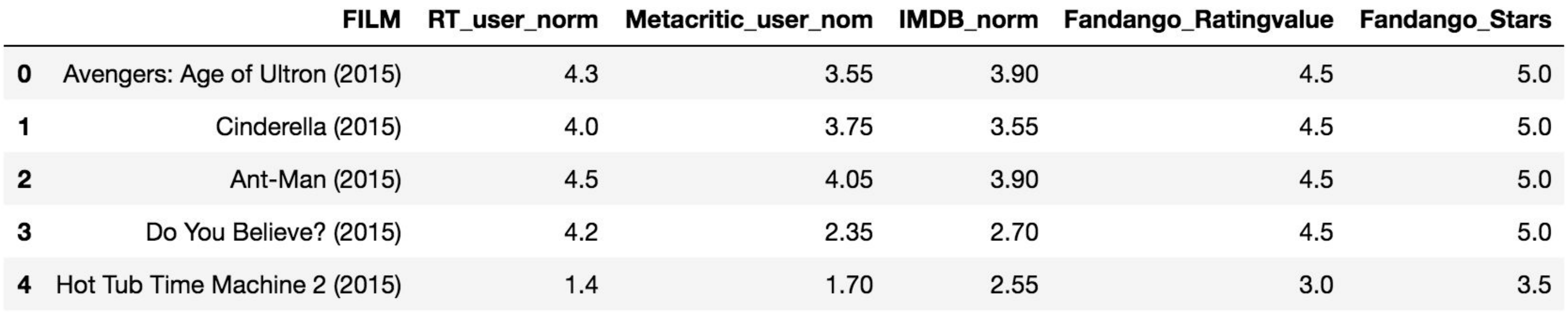
By Walt Hickey

Filed under Movies

Get the data on GitHub

“Ted 2,” “Avengers: Age of Ultron,” and “Fantastic Four”

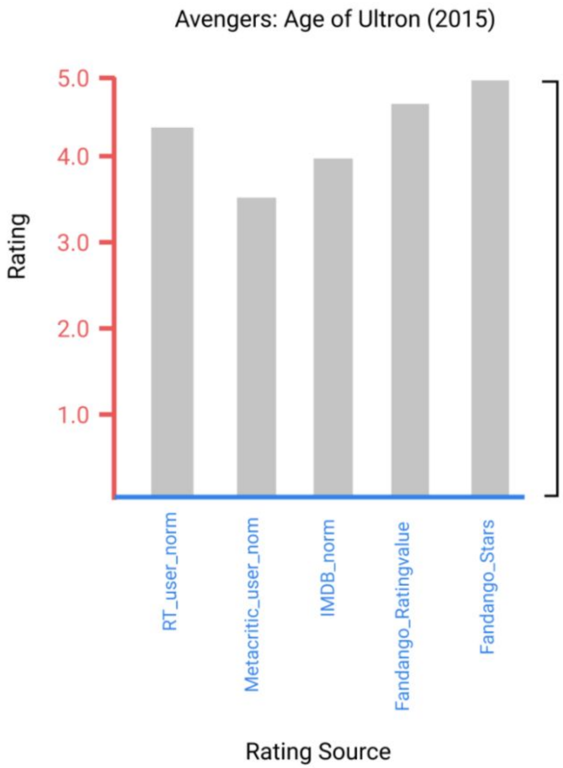
Introduction to the data



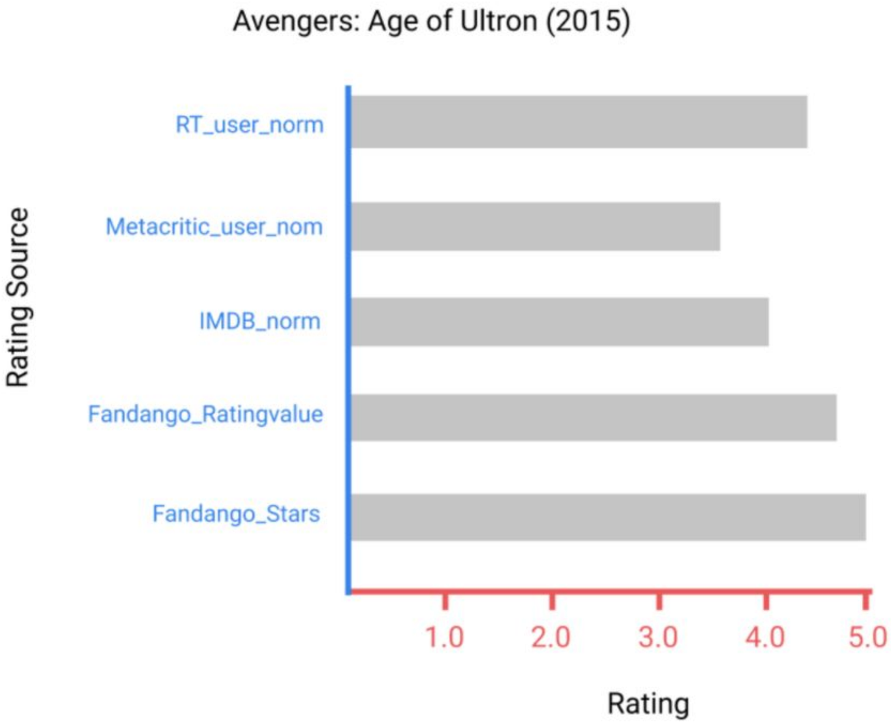
	FILM	RT_user_norm	Metacritic_user_nom	IMDB_norm	Fandango_Ratingvalue	Fandango_Stars
0	Avengers: Age of Ultron (2015)	4.3	3.55	3.90	4.5	5.0
1	Cinderella (2015)	4.0	3.75	3.55	4.5	5.0
2	Ant-Man (2015)	4.5	4.05	3.90	4.5	5.0
3	Do You Believe? (2015)	4.2	2.35	2.70	4.5	5.0
4	Hot Tub Time Machine 2 (2015)	1.4	1.70	2.55	3.0	3.5

https://github.com/fivethirtyeight/data/tree/master/fandango

Bar plot

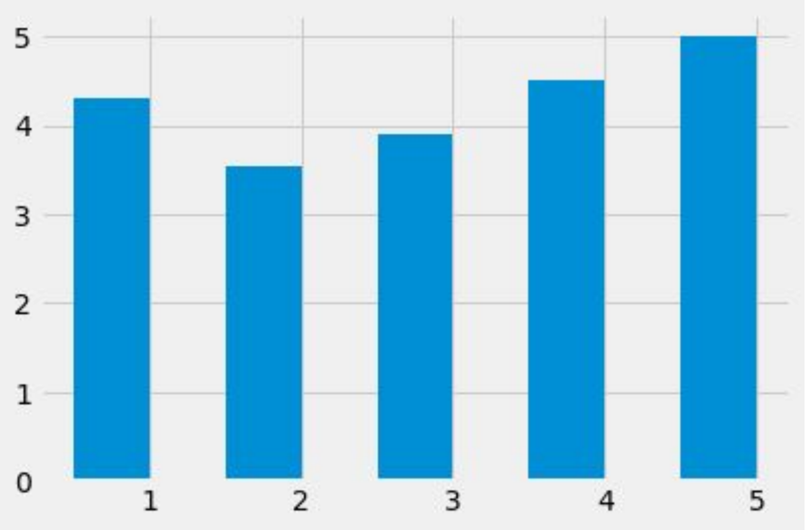


Length of each bar corresponds to numerical value

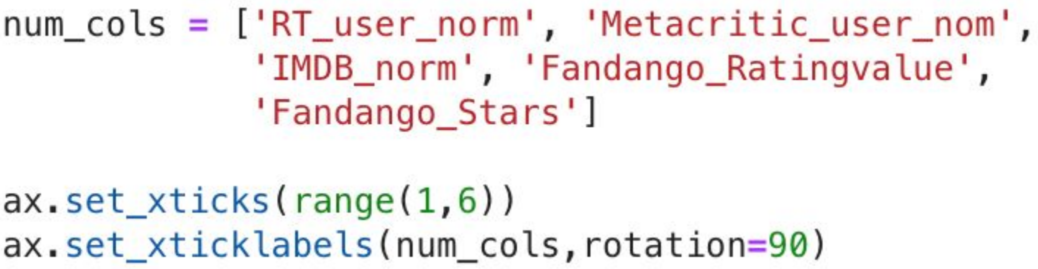


Creating Bars

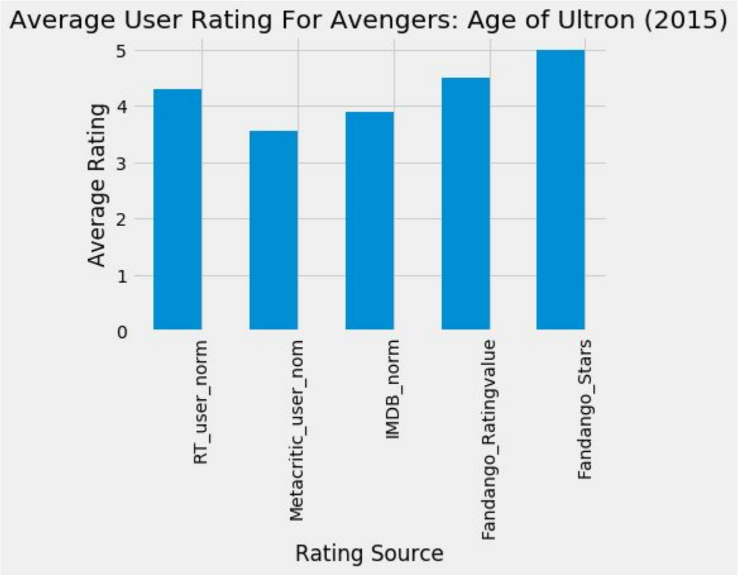
```
import numpy as np

plt.style.use('fivethirtyeight')

# create a subplot
fig, ax = plt.subplots()

# position of bars
bar_positions = np.arange(5) + 0.75

# Average rating for the first movie in the dataset.
num_cols = ['RT_user_norm', 'Metacritic_user_nom',
            'IMDB_norm', 'Fandango_Ratingvalue',
            'Fandango_Stars']
bar_heights = norm_reviews[num_cols].iloc[0]

# create a bar plot
ax.bar(bar_positions,bar_heights,0.5)

plt.show()
```



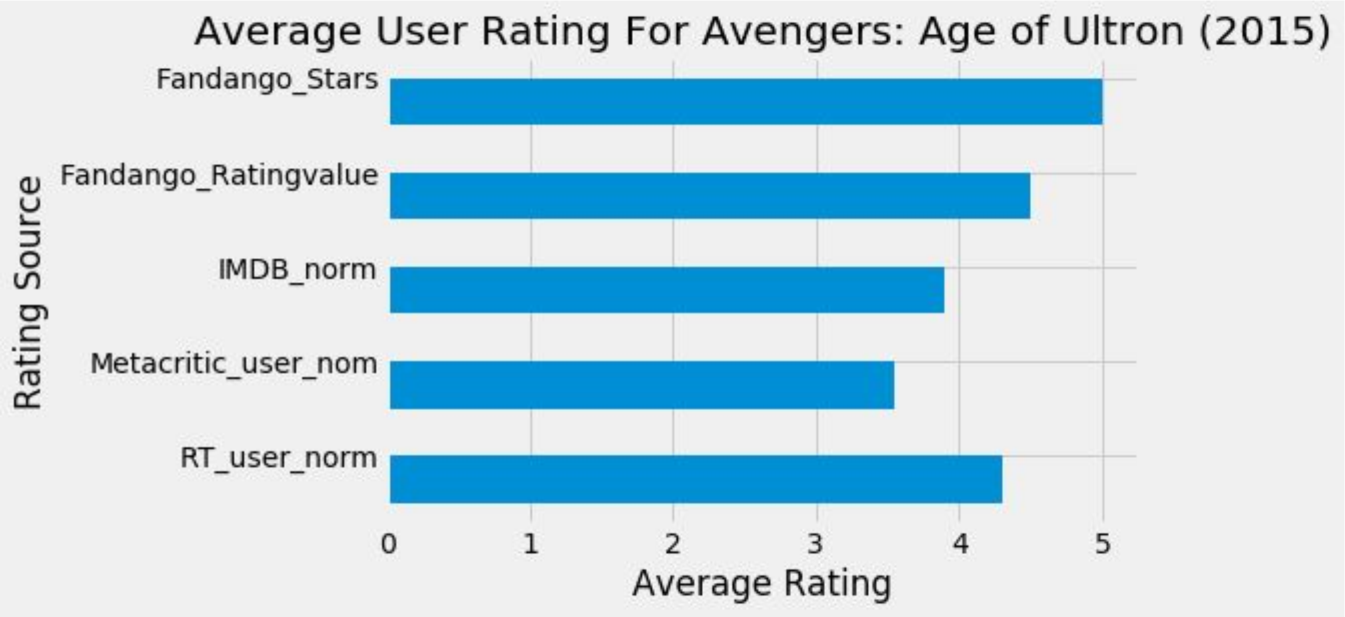
Aligning axis ticks and labels



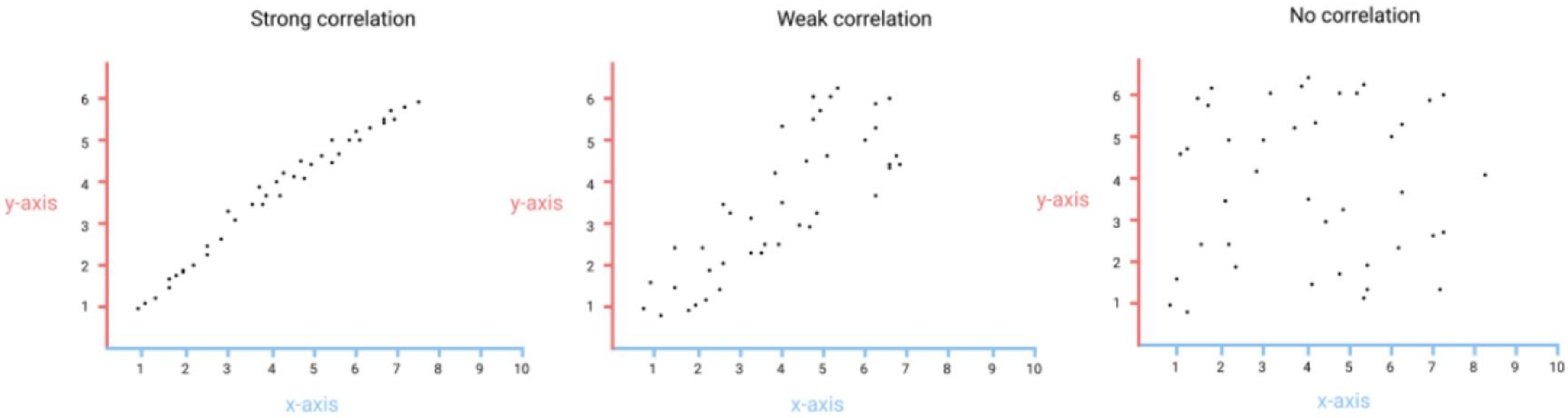
```
num_cols = ['RT_user_norm', 'Metacritic_user_nom',
            'IMDB_norm', 'Fandango_Ratingvalue',
            'Fandango_Stars']

ax.set_xticks(range(1,6))
ax.set_xticklabels(num_cols,rotation=90)
```



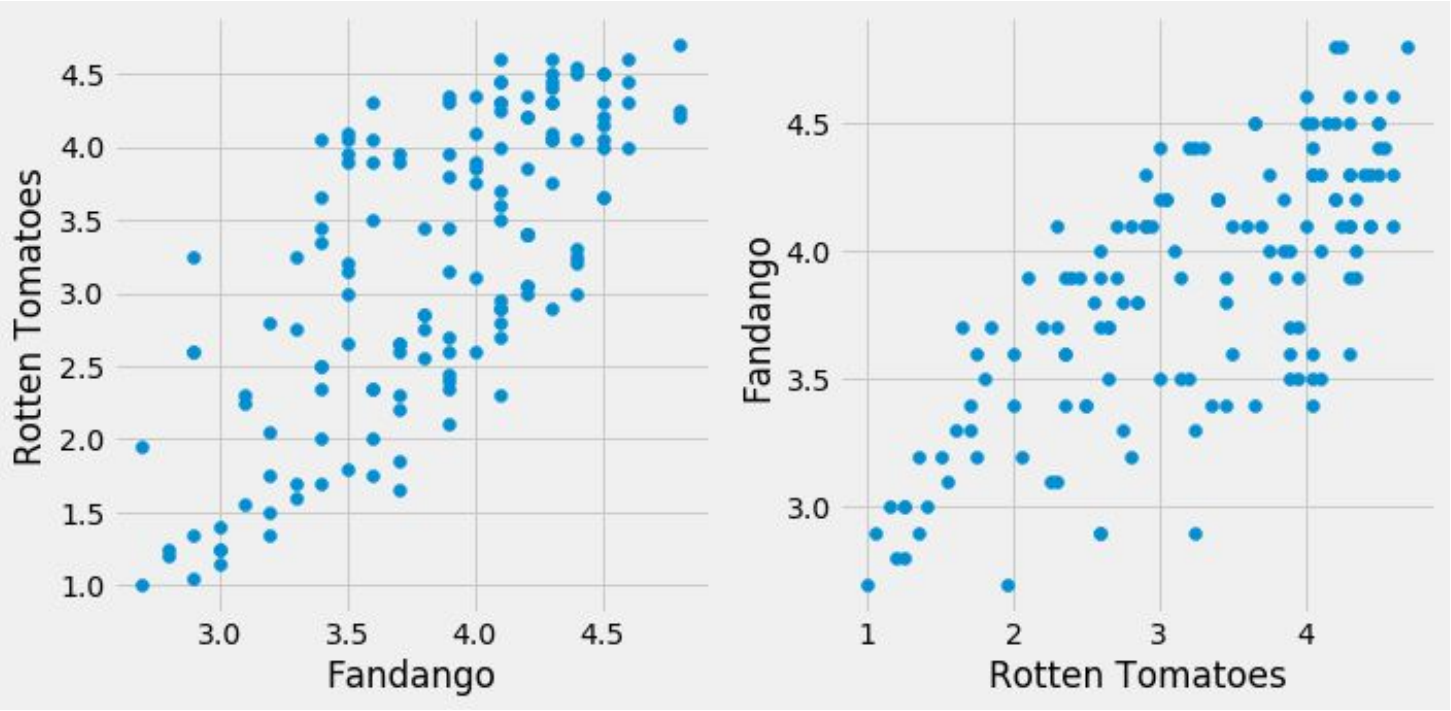
Horizontal bar plots



Scatter plot



Switching axes



Lesson #6 - Exploratory Data Analysis.ipynb
Section 3