Data Science and Visualization (DSV, F23)

3. Data Visualization

Hua Lu
https://luhua.ruc.dk; luhua@ruc.dk
PLIS, IMT, RUC

Agenda

- **Basic Visualization**
  - Histograms
  - Bar charts
  - Box plots
  - Scatter plots
  - Line charts
- Advanced Visualization

Learning goals

for each plot type:

- What
- Why
- How

Visualization Module and Functions

- import **matplotlib.pyplot** as plt
  - Histogram: plt.hist()
  - Bar chart: plt.bar()
  - Boxplot: plt.boxplot()
  - Scatter plot: plt.scatter()
  - Line chart: plt.plot()
- Parameters
  - Data: *in general*, a Series object
    - Index for the X axis
    - Values for the Y axis
  - Others for labels, ticks, legend, title…

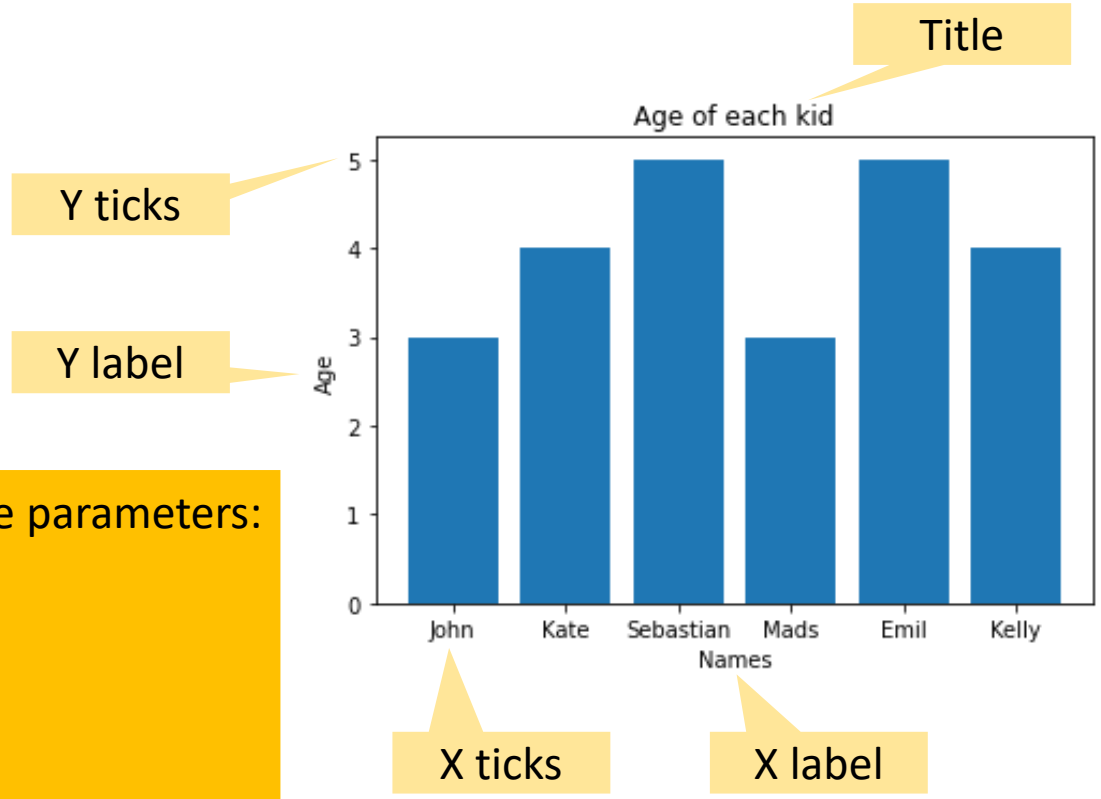
Useful Functions in Plotting

- X and Y labels
  - plt.xlabel(), plt.ylabel()
- X and Y ticks
  - plt.xticks(), plt.yticks()
- Title of a figure
  - plt.title()
- Legend
  - plt.legend()
- Save to file
  - plt.savefig()

These functions have flexible parameters:
- Location
- Font type
- Font size

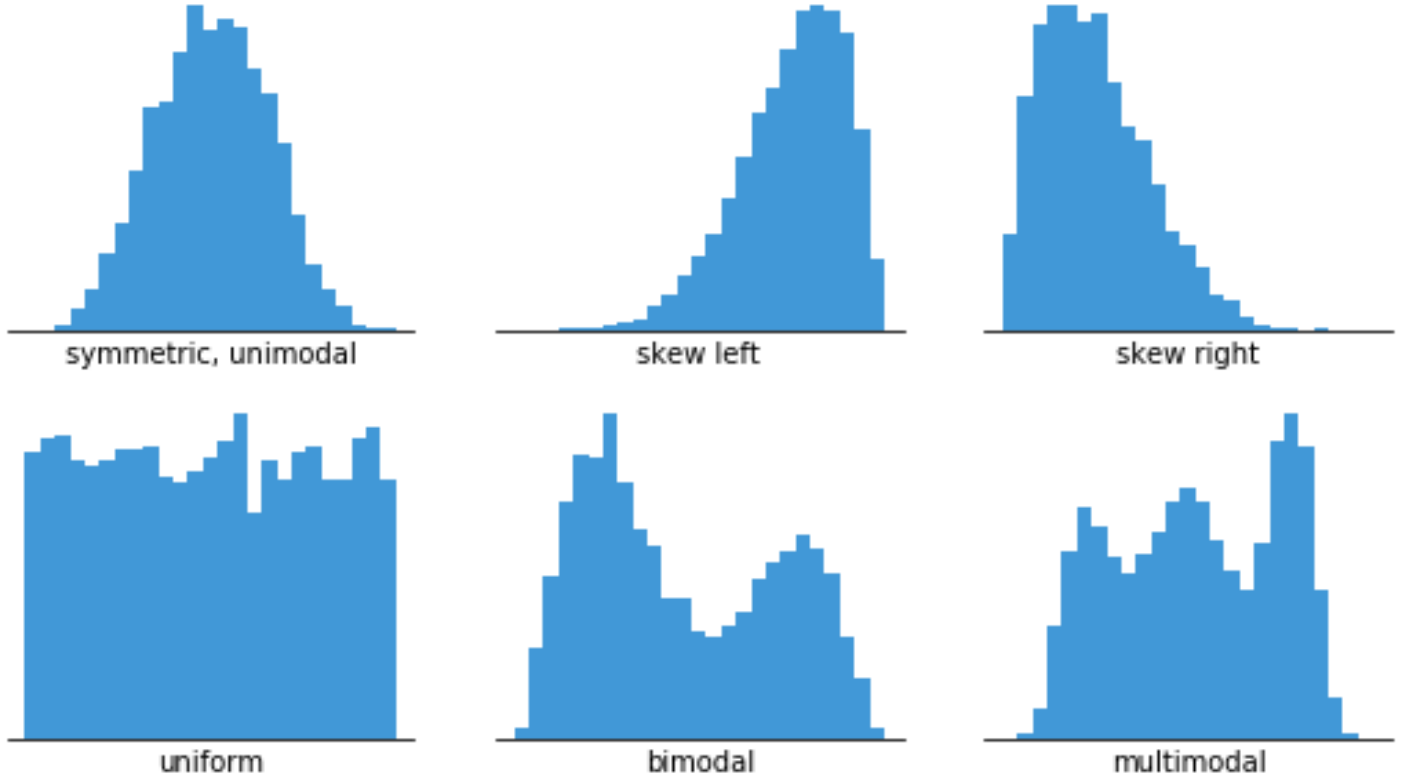
Ticks
- Range and step
- Rotation

Save image
- dpi
- Image type (.jpg, .png, .pdf…)



Histogram

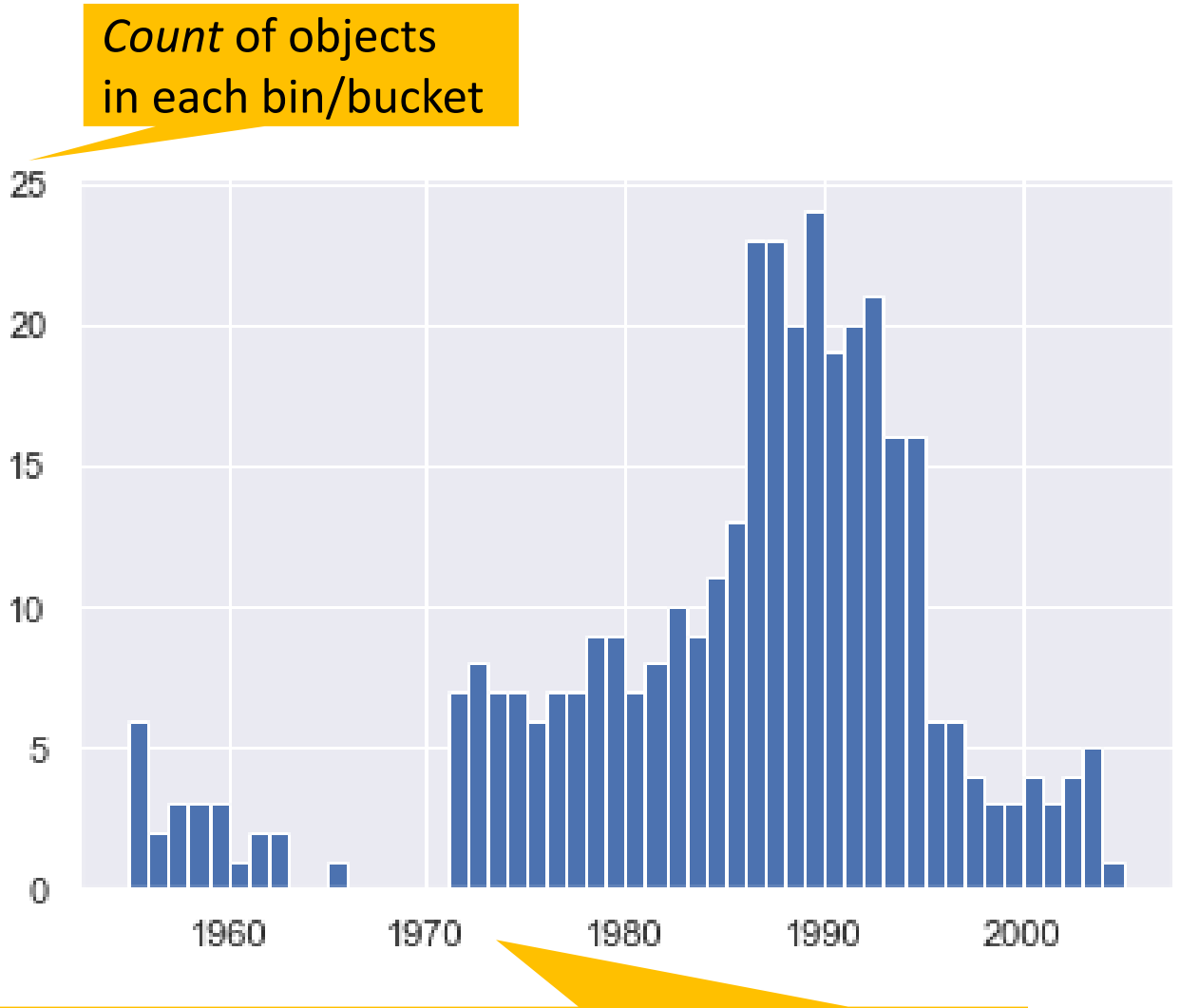
- An approximate representation of the distribution of numerical data.
- It plots the frequency distribution of numerical data (continuous or discrete).
  - X axis: values
  - Y axis: frequencies



https://chartio.com/learn/charts/histogram-complete-guide/

Histogram in Python

- plt.hist(series_1)
  - series_1 = df.col1: To illustrate the frequency of each particular value of column df.col1
  - series_1 = a column with filtering
- Examples
  - Lecture3_Visualization.ipynb

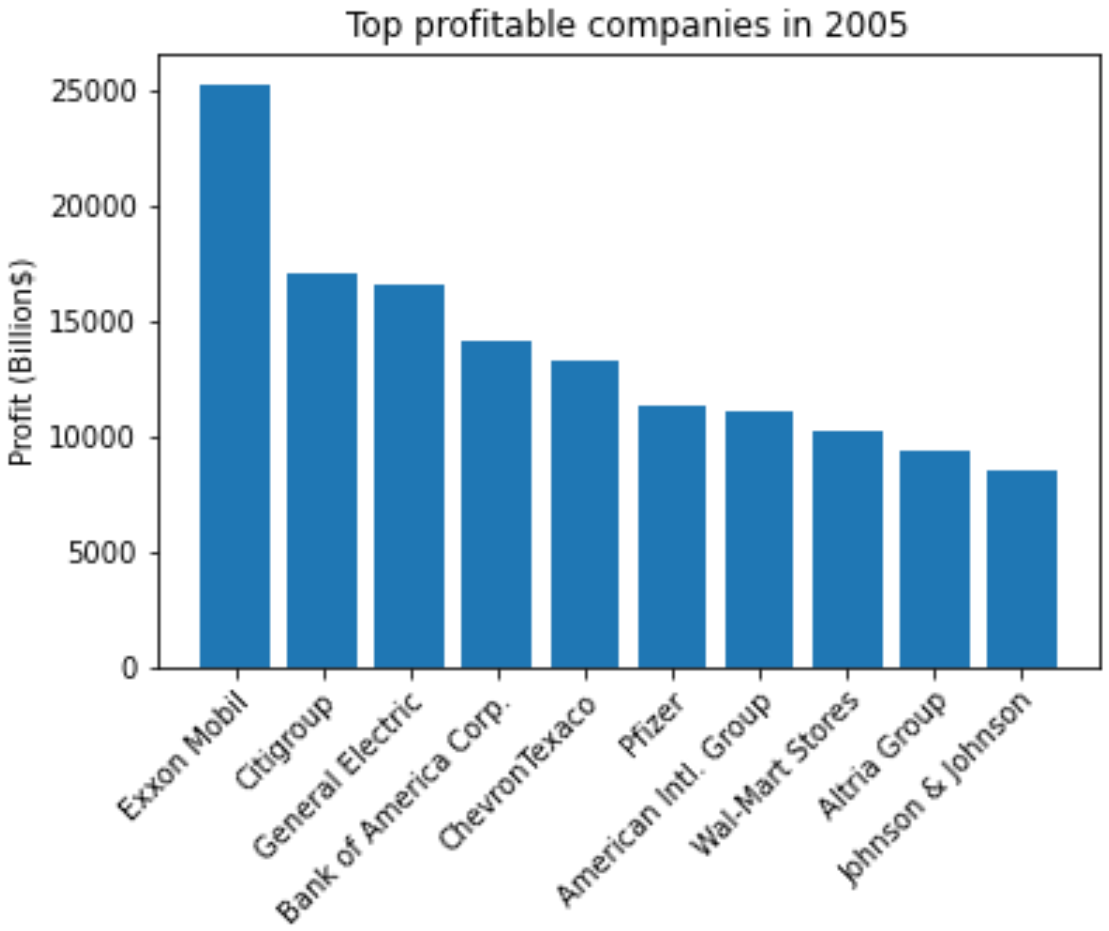


jupyter

DEMO

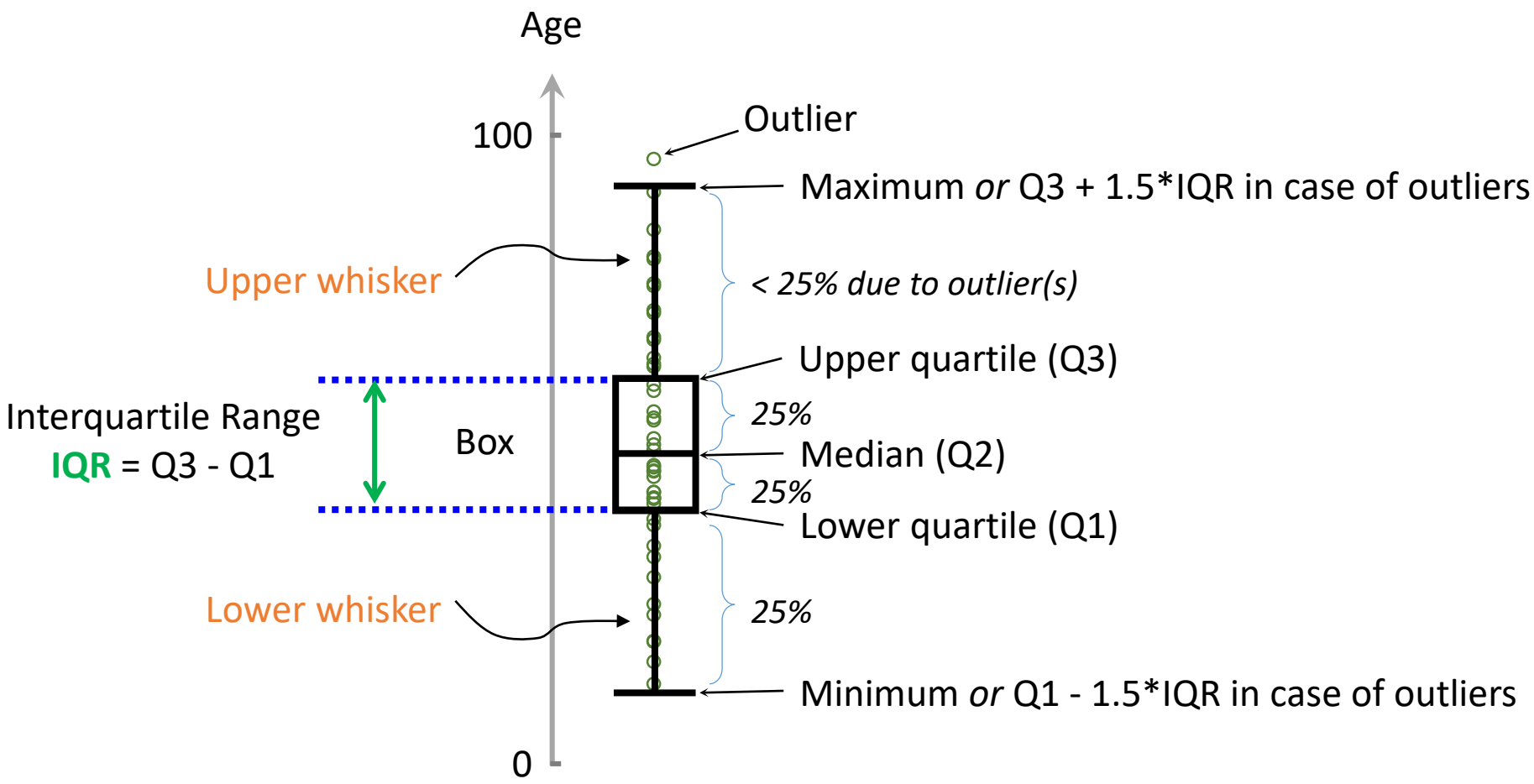
Bar Chart

- A generalization of histogram
  - For categorical data (X axis)
- plt.bar(x_data, y_data)
- Examples
  - Lecture3_Visualization.ipynb



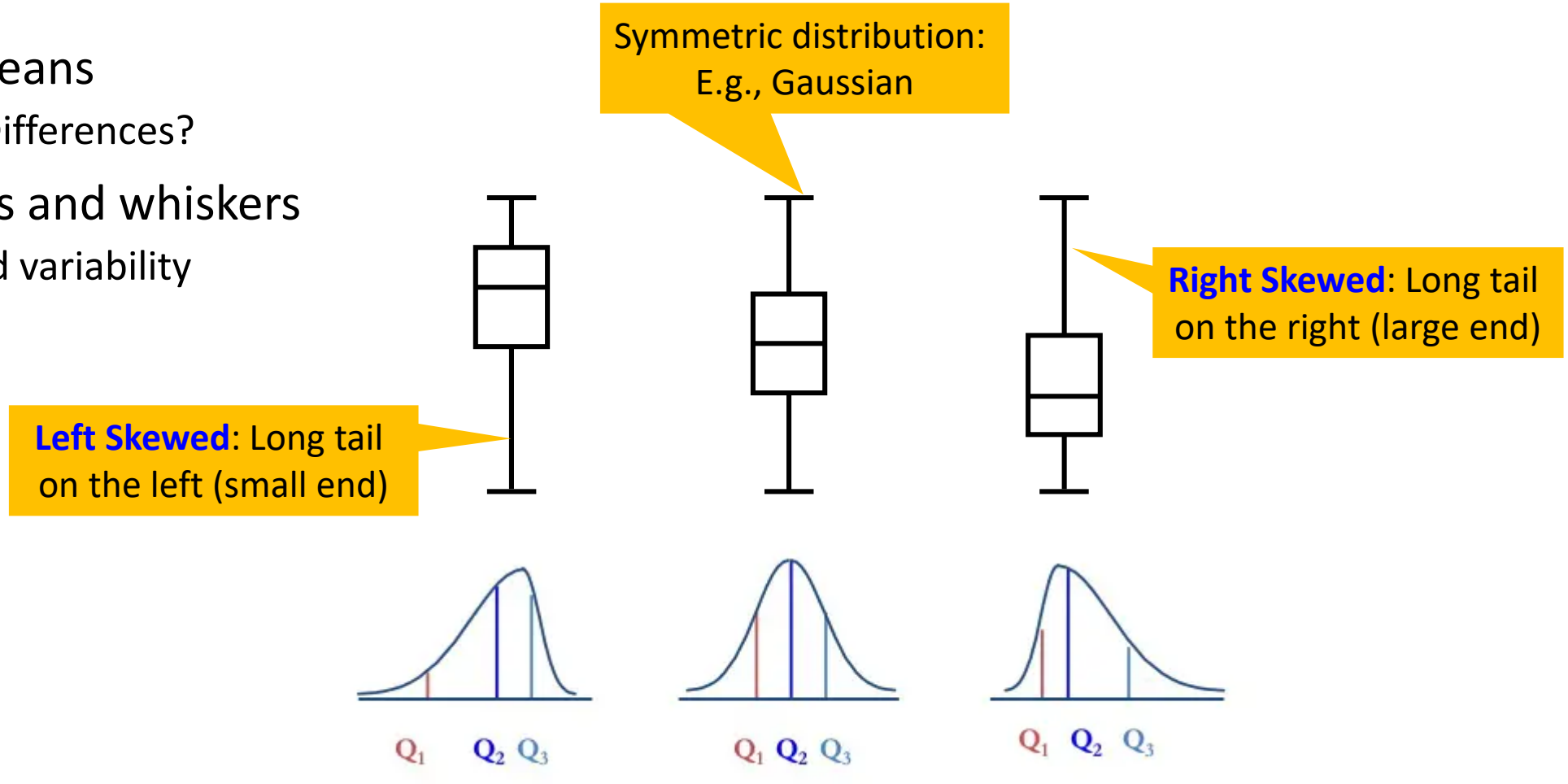
jupyter DEMO

Boxplot Anatomy



Compare Boxplots

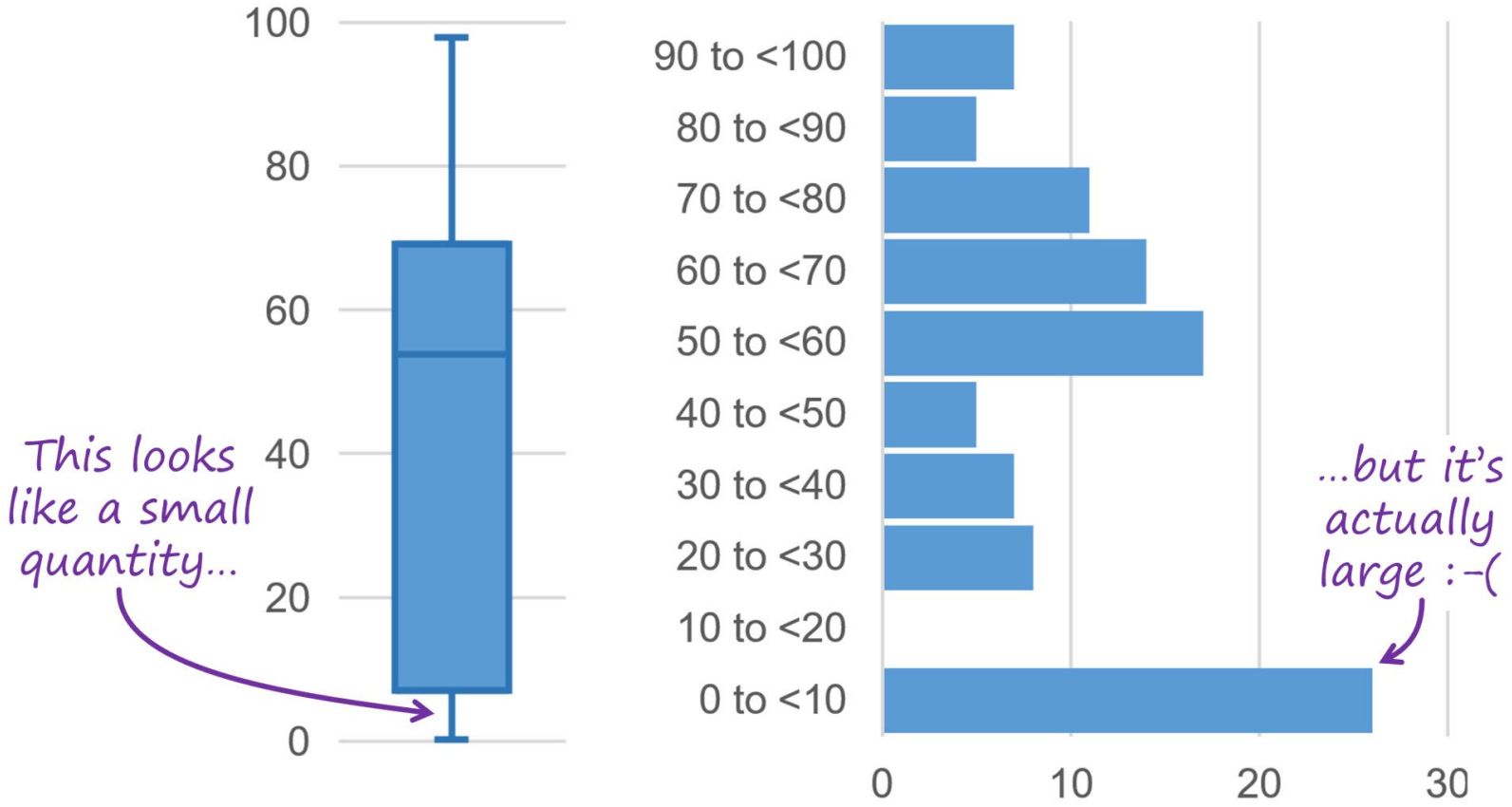
- Boxes and means
  - Overlap? Differences?
- sizes of boxes and whiskers
  - Ranges and variability
- Outliers?



https://www.simplypsychology.org/boxplots.html

Don’t be fooled by Boxplot

- Boxplot indicates
  - Range
  - Spread
  - Quartiles
- But not
  - Absolute quantity



https://nightingaledvs.com/ive-stopped-using-box-plots-should-you/

Boxplot Creation

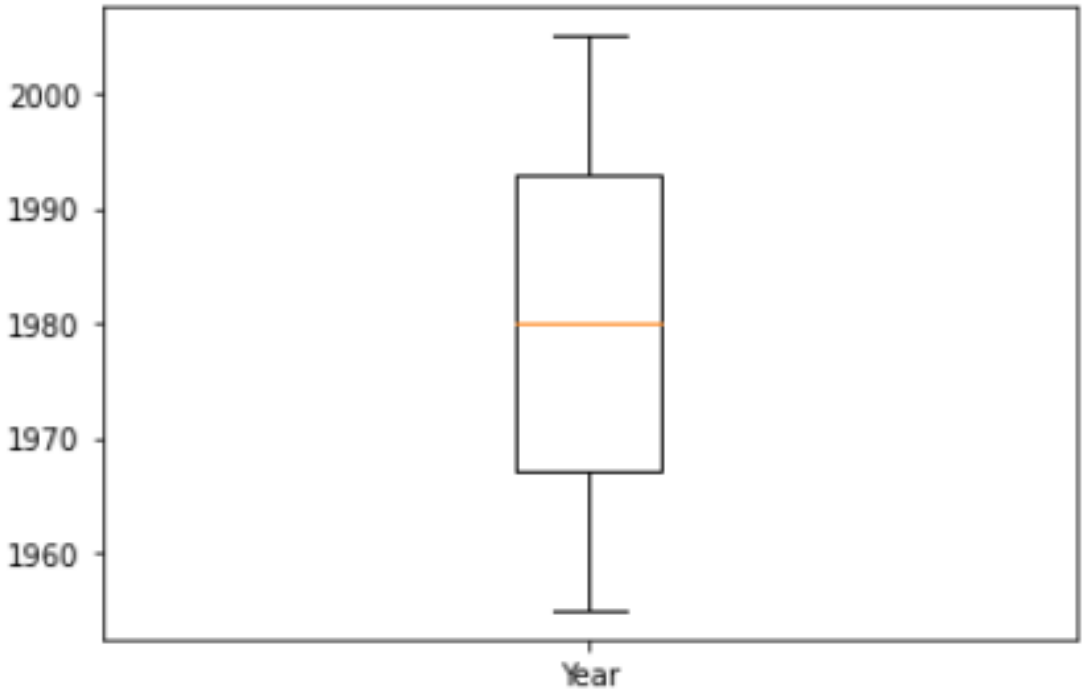
- plt.boxplot()
  - Lecture3_Visualization.ipynb

```
In [6]: data['year'].describe()

Out[6]: count    25500.00000
        mean      1980.00000
        std         14.71989
        min       1955.00000
        25%       1967.00000
        50%       1980.00000
        75%       1993.00000
        max       2005.00000
        Name: year, dtype: float64
```

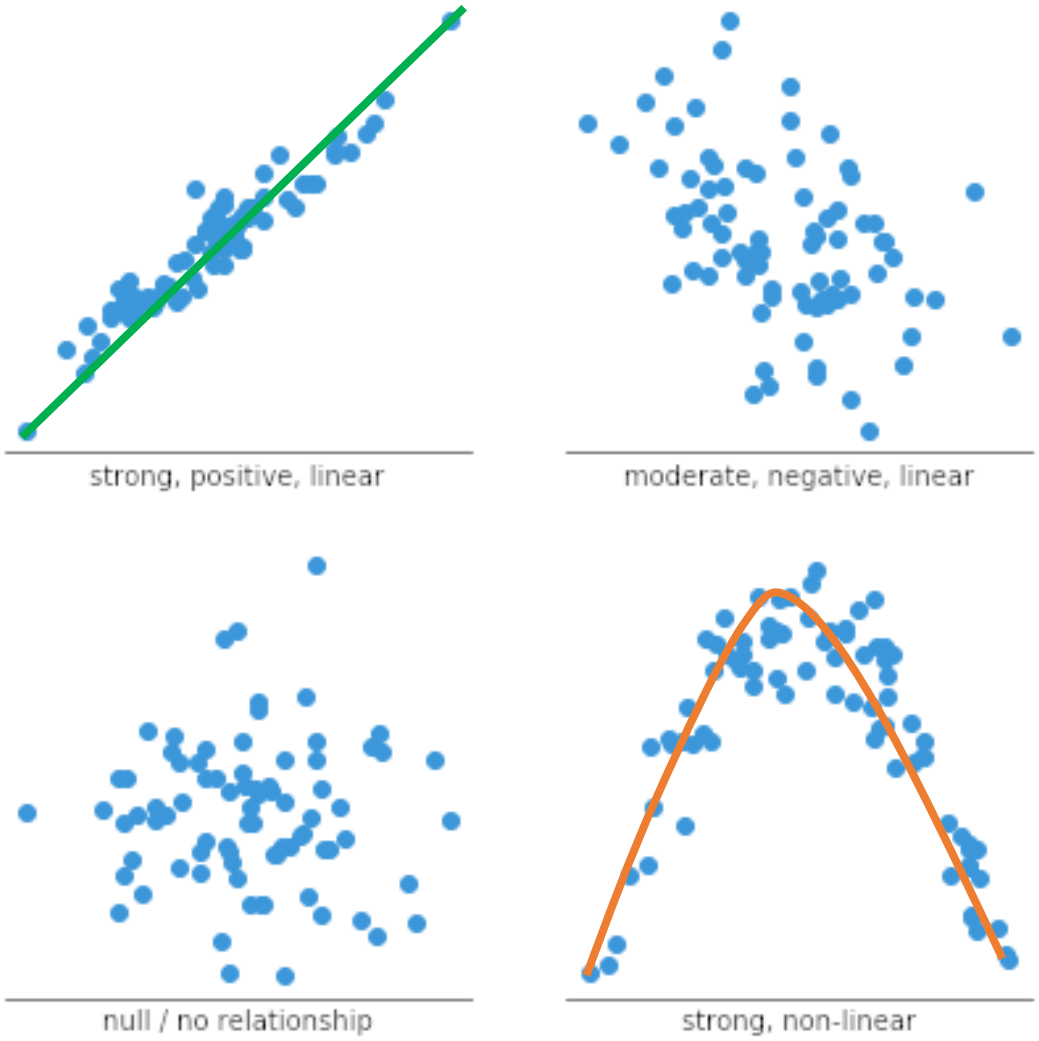
DEMO

```
plt.boxplot(data['year'])
plt.xticks([1], ['Year'])
```



Scatter Plot

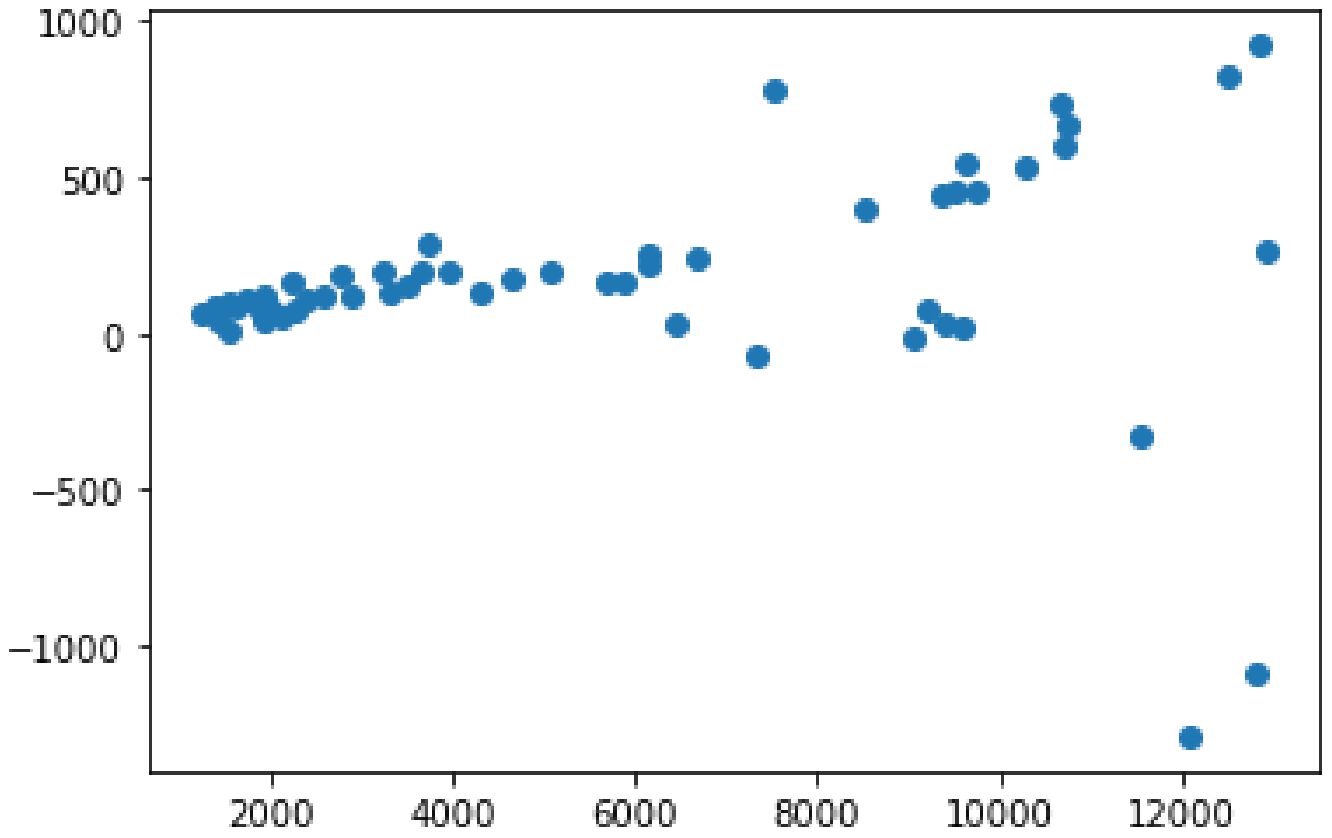
- A.k.a. scatter chart.
- A chart that shows the *relationship* between two variables (x *vs.* y).
- Regression
  - Linear
  - Polynomial



https://chartio.com/learn/charts/what-is-a-scatter-plot/

Scatter Plot in Python

- plt.scatter(x_data, y_data)
- Examples
  - Lecture3_Visualization.ipynb



DEMO

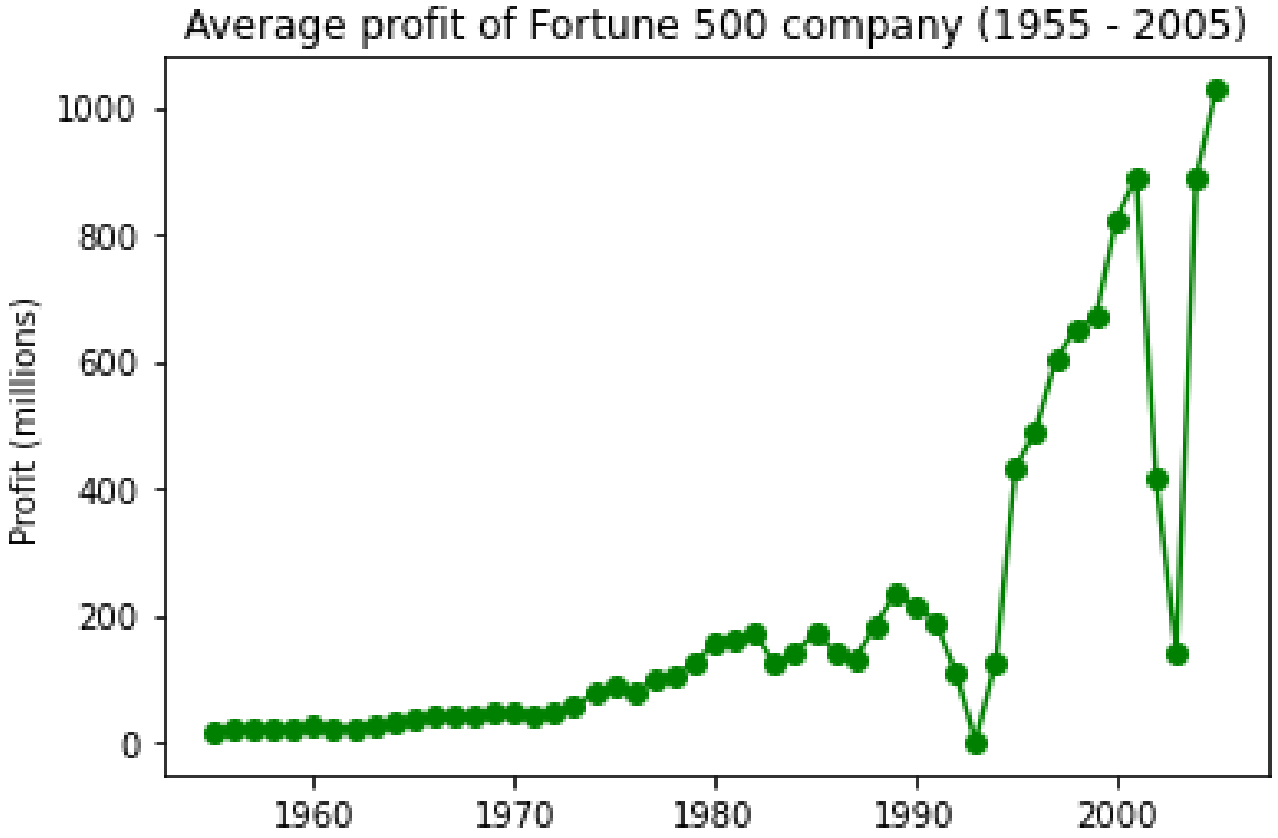
Line Chart

- A line chart uses points connected by line segments from left to right to demonstrate changes in value.
  - X axis: a continuous progression, e.g., time.
  - Y axis: values corresponding to the progression.
- When to use it?
  - If you want to see changes of a variable
  - If you want to see the trend of a variable
  - If you want to compare the changes/trends of two or more variables
    - Two or more lines

https://www.xe.com/

Line Chart in Python

- plt.plot(x_data, y_data)
- Examples
  - Lecture3_Visualization.ipynb



jupyter DEMO

Agenda

- Basic Visualization
- **Advanced Visualization**
  - Advanced bar chart
  - Pairplot
  - Correlation heatmap

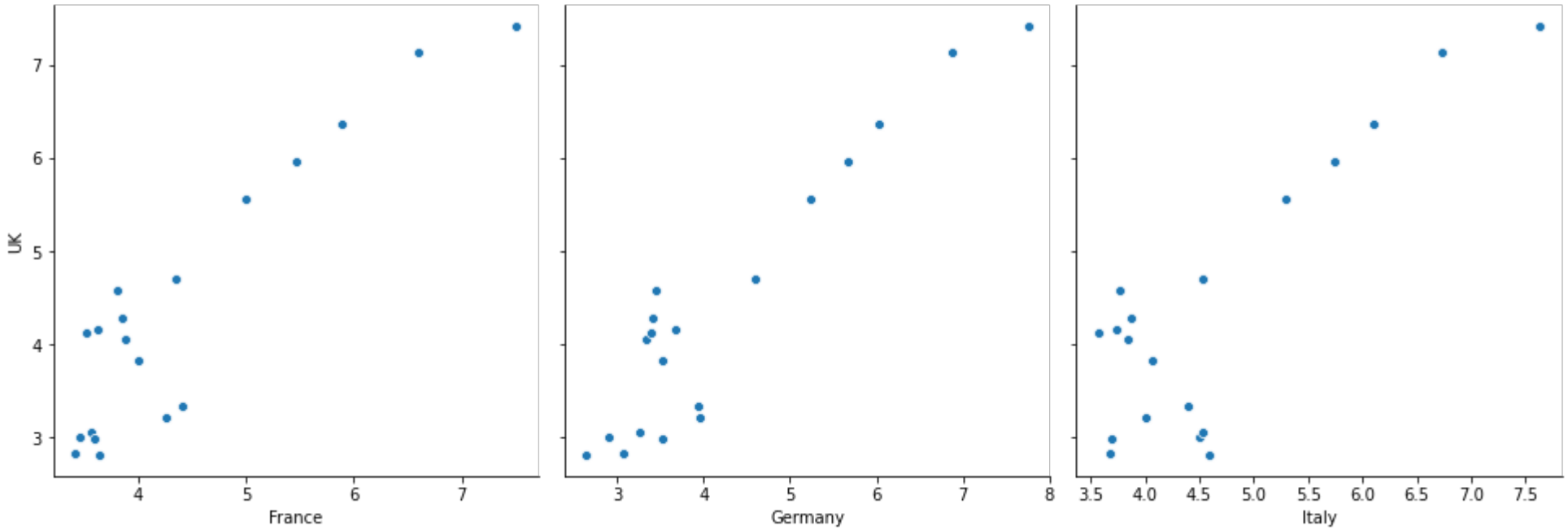
Advanced Bar Chart

DEMO

- We can plot horizontal bars with richer colors
  - import **seaborn** as sns
    - A matplotlib based library for nicer plots
- Lecture3_Advanced.ipynb
  - Data: gas_prices.csv (in Moodle)

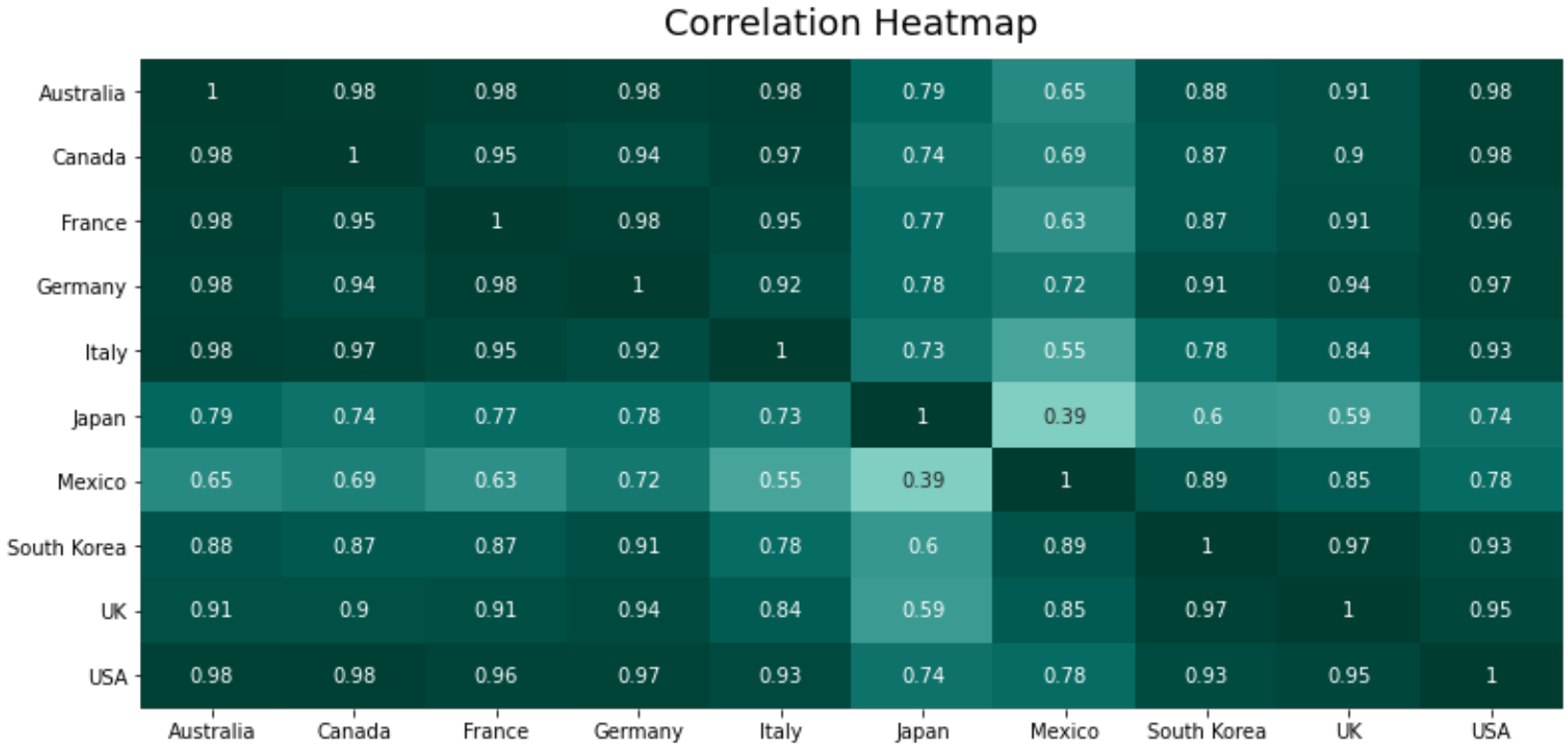
Pairplot

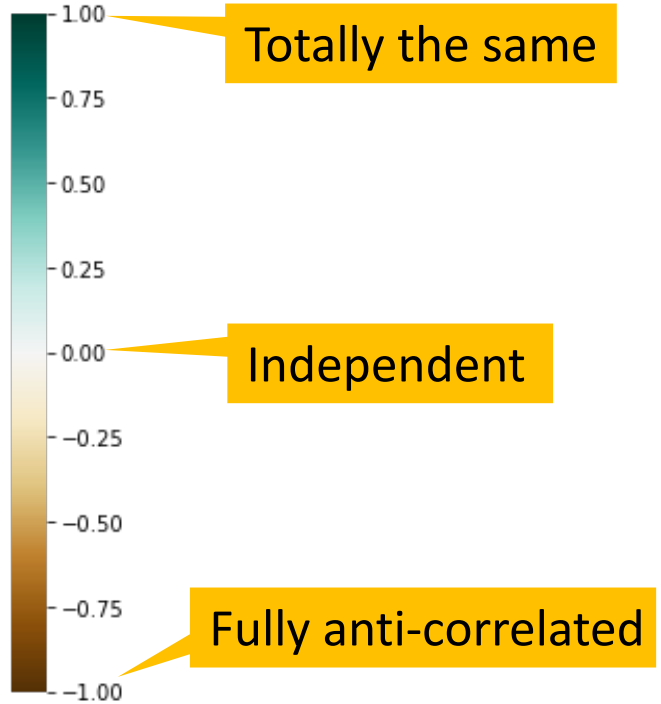
- Compare a number of columns with another column
  - A series of scatter plots, still about correlation
  - sns.pairplot(data, x_vars=['France', 'Germany', 'Italy'], y_vars='UK', …)
  - Lecture3_Advanced.ipynb



Correlation Heatmap

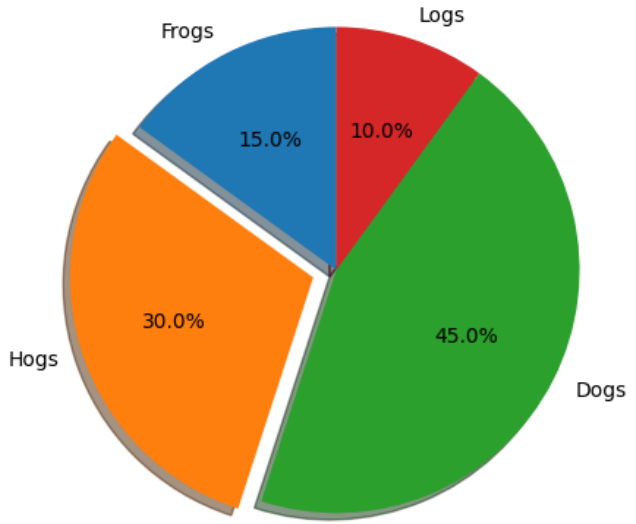
- To check the correlation for each pair of *numeric* columns
  - A color-encoded matrix: sns.heatmap(data.corr(), …)
  - Lecture3_Advanced.ipynb

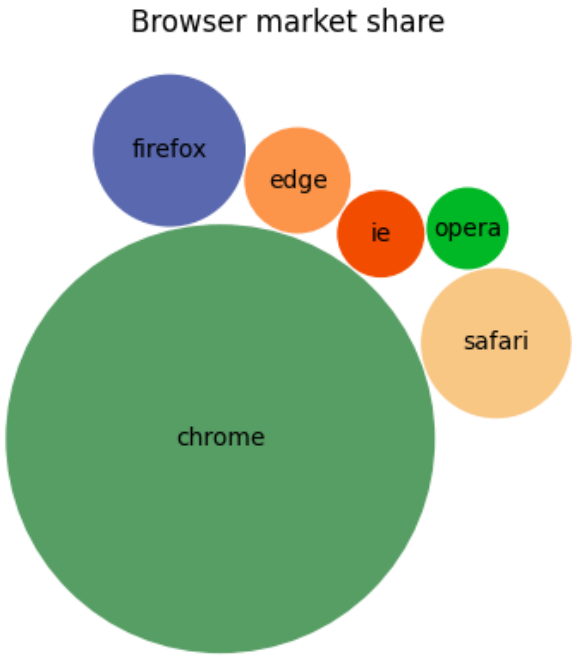


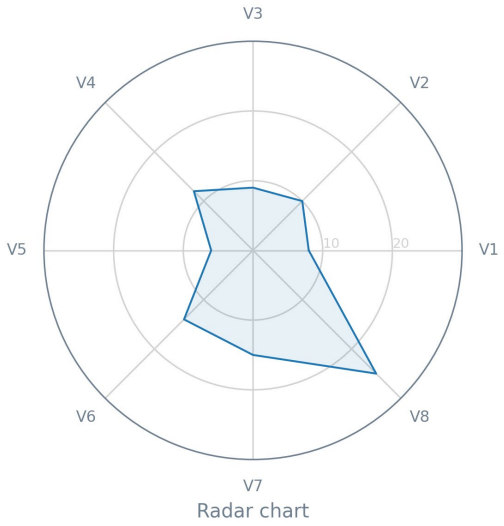


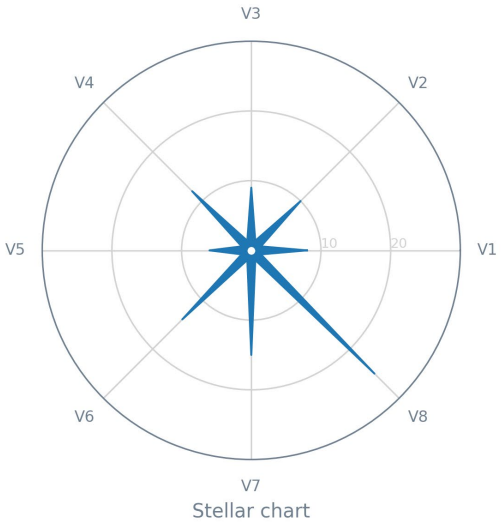
More Plot Types

- bubble
- pie
- Radar chart
- 3D
- …









https://matplotlib.org/

Reminders

- Exercises
  - Read the sample code before you do the exercises
  - At least take a quick look
- Groups for mini-project
  - Please use this Padlet to form your groups
  - Padlet: https://padlet.com/luhua/dsv-f23-mini-project-group-formation-sgw8mxkajzhh0tli

Summary

- Basic Visualization (import **matplotlib.pyplot** as plt)
  - Histograms
  - Bar charts
  - Boxplot
  - Scatterplots
  - Line charts
- Advanced Visualization (import **seaborn** as sns)
  - Advanced bar chart
  - Pairplot (2 dimensions/columns)
  - Correlation heatmap (A full pair-wise analysis)

References

- Matplotlib official website
  - https://matplotlib.org/
- Matplotlib Tutorial
  - https://www.w3schools.com/python/matplotlib_intro.asp
  - https://www.geeksforgeeks.org/matplotlib-tutorial/?ref=lbp
- Seaborn official website
  - https://seaborn.pydata.org/

Exercises

1. Find the exercises in Lecture3_Visualization.ipynb
2. Find the exercises in Lecture3_Advanced.ipynb

- **NB**: For all these exercises, you just fill in the blank cells left in a notebook. Nevertheless, remember to run the cells that import the libraries and load/create the data before running your own code.