Datasets

PhD. Msc. David C. Baldears S.

TC3007C

Datasets

- Different types of problems require different data and techniques to solve them.
- Datasets”, fundamental to the data processing revolution
- A Dataset is a set of tabulated data in any structured data storage system.
- The term refers to a single source database.
- Normally tablewise:
  - Each column of the Dataset represents a variable.
  - Each row corresponds to any data that we are dealing with.

https://keepcoding.io/blog/que-son-datasets/

Activity 01

Load data set into Notebook Download one of the data set below and

Classification and regression data sets (features and labels)

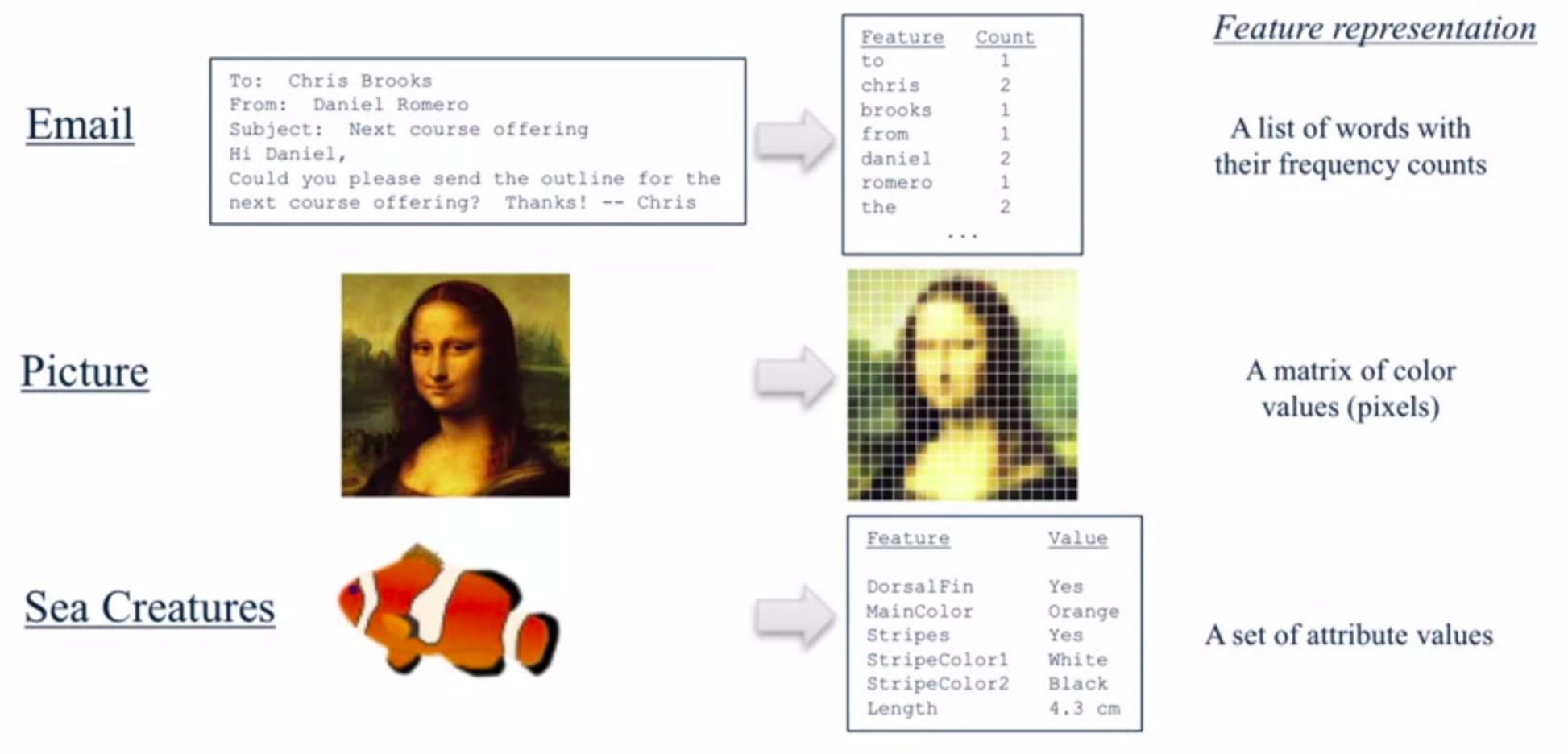
- data set for regression iris data description
- data set for classification wine data description

These are repositories of datasets

- UC Irvin https//archive.ics.uci.edu/ml/index.php
- Data set from https//index.okfn.org/place/
- UN http//data.un.org/
- World bank https//data.worldbank.org/

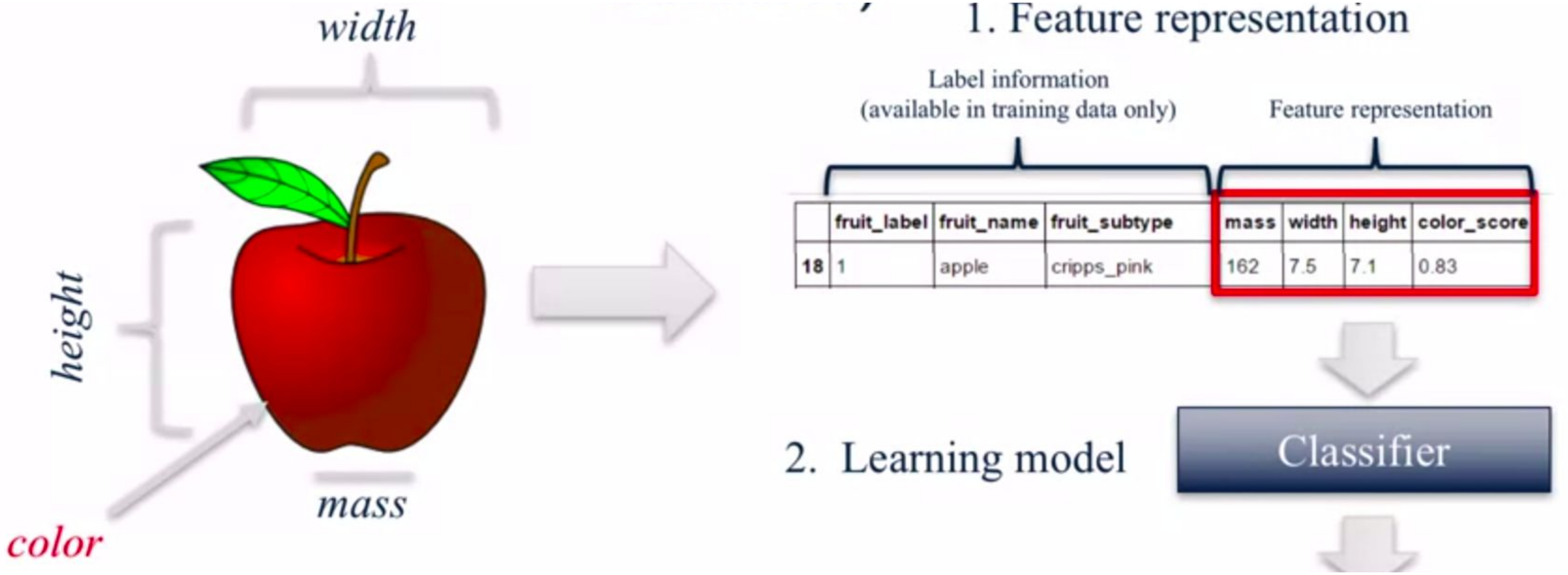
Go into one of the above links, download a dataset, open it in a plain text editor (notepad, nano, pico, etc...) and locate the instances, the features/attributes, the values of the attributes, the labels/classes. Now load it into pandas in your own notebook

Feature Representation



source by Kevyn Collins-Thompson, University

Feature Representation



source by Kevyn Collins-Thompson, University

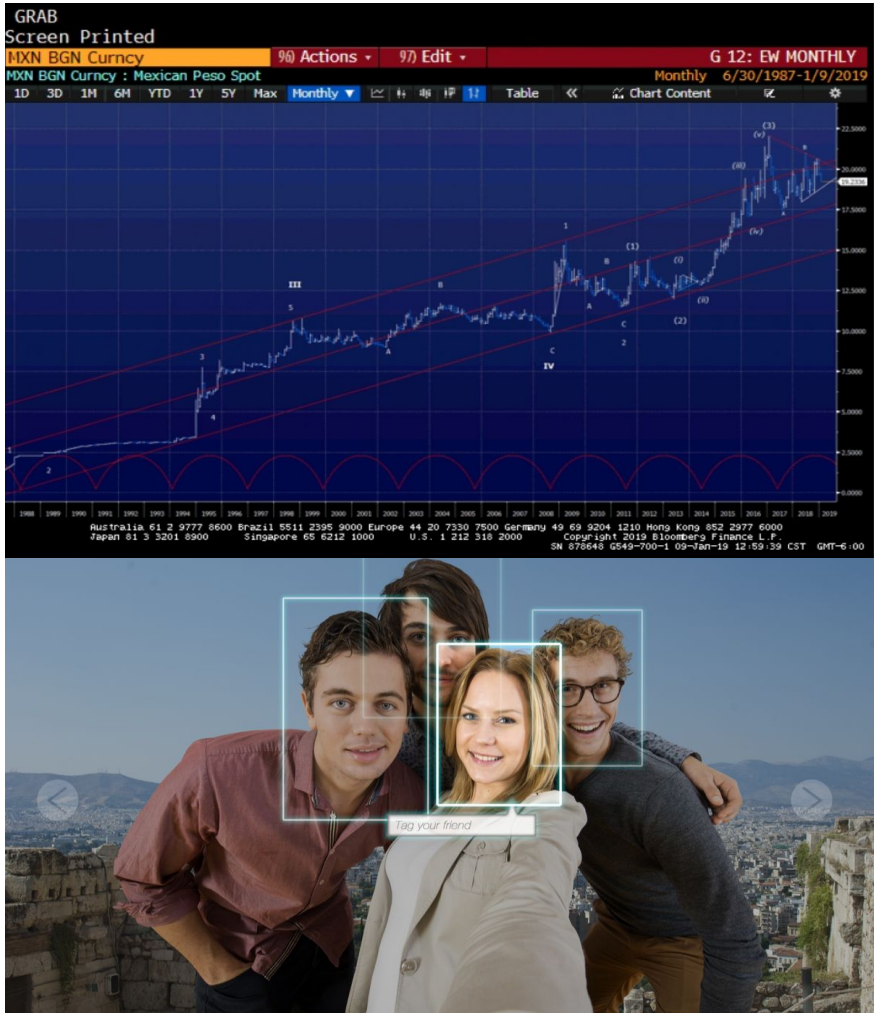
How are they used?

Regression to predict a numeric value given a set of inputs.

Finance market prediction.

Classification to predict a class or label given a set of inputs.

Facebook face recognition tagging.



Big Data Principles

More data versus Better Algorithms

The worst algorithm can beat the best algorithm when the size of the dataset is dramatically increased.

More data beats clever algorithms, but better data beats more data.

Parametric versus Non-parametric Models

Parametric model: Regard data as random samples from a distribution, and try to estimate its parameters.

Nonparametric model: Ignore any distribution and treat the data on its own.

Dataset Types

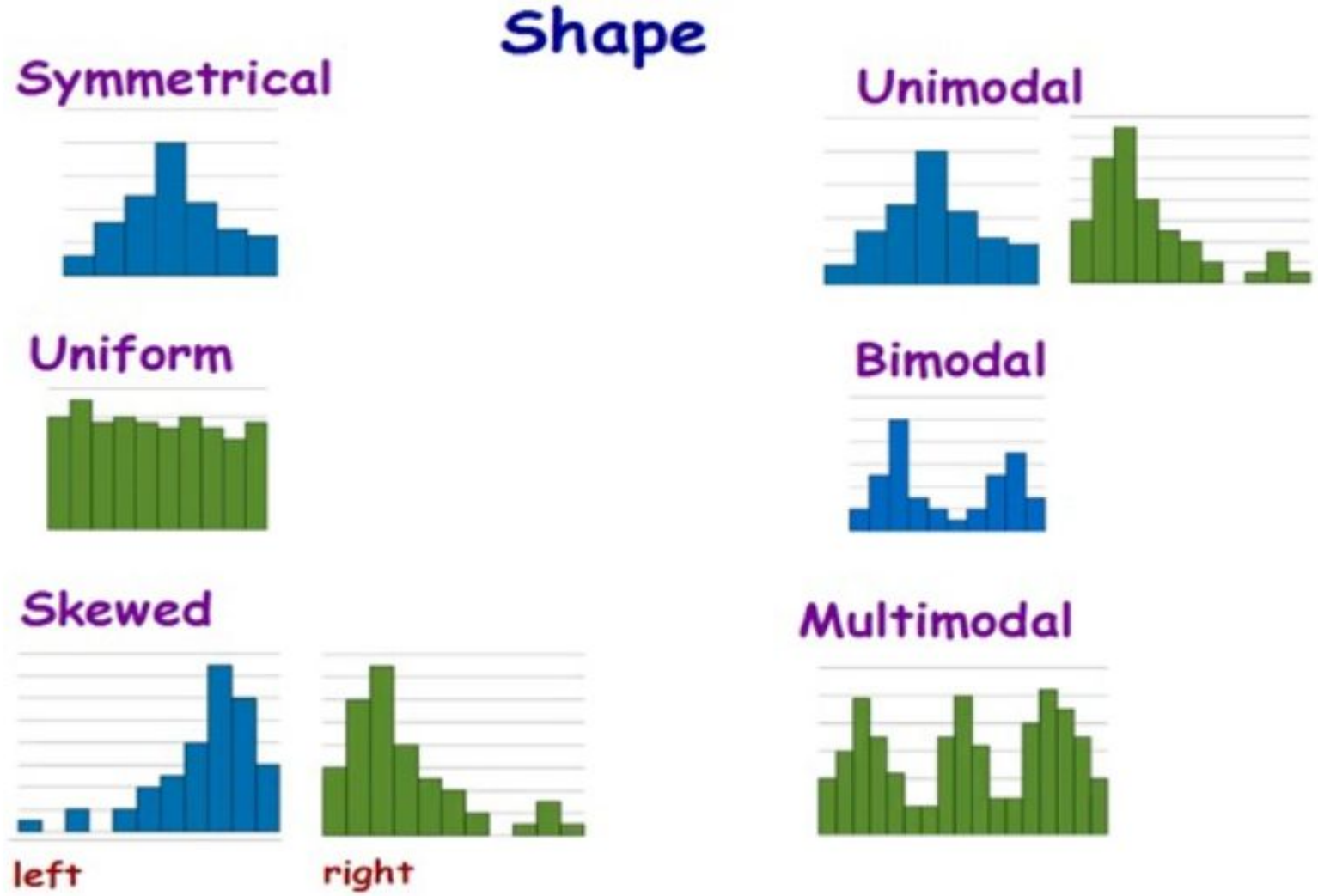
There are four types of Datasets:

1. File: it is an independent file in which all the information is stored.
2. Folder: is the sum of different Datasets stored in the same folder, which are connected to each other. These files must share the same format such as .csv, .mif or dxf.
3. Databases: they are specific formats designed for specific programs.
4. Web: is the compilation of data that is stored within a website.

Ways to describe datasets

Symmetrical
Mean Mode and Median roughly equal

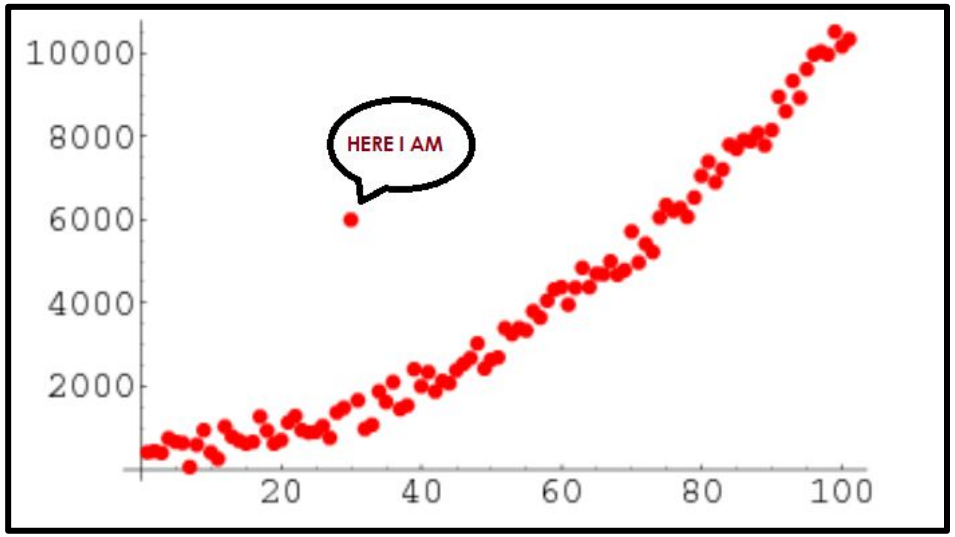
Non-Symmetrical



Mean < Median < Mode Mean > Median > Mode

Outliers

- Definitions:
  - An Outlier is an extreme value of the data (extremely high or extremely low).
  - Its an observation value that is significantly different from the rest of the data.
  - An outlier is a data point that lies outside the overall pattern in a distribution.
- There might be more than one outlier in a set of data
- Possible reasons for an outlier:
  - An error was made while taking the measurement or entering into the computer.
  - The individual belongs to a different group than the bulk of individual measured.
  - The outlier is legitimate, though extreme data value.

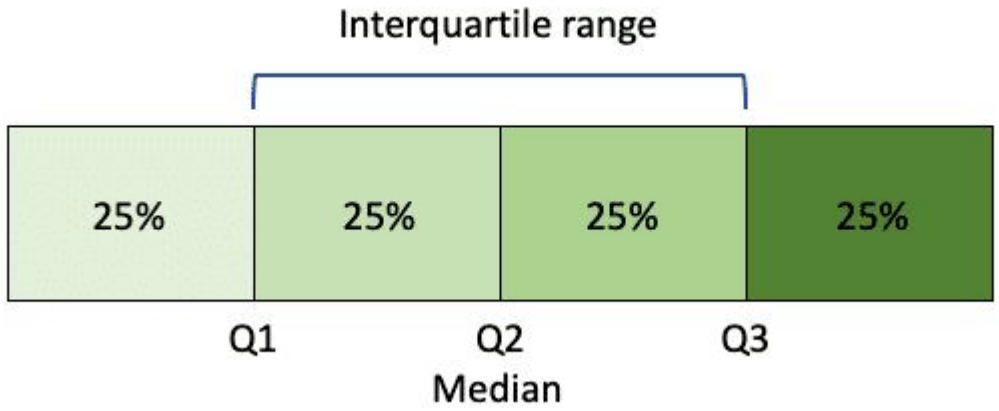


https://miro.medium.com/max/697/1*O3lOgPwuHP7Vfc1T6NDRrQ.png

Identifying outliers

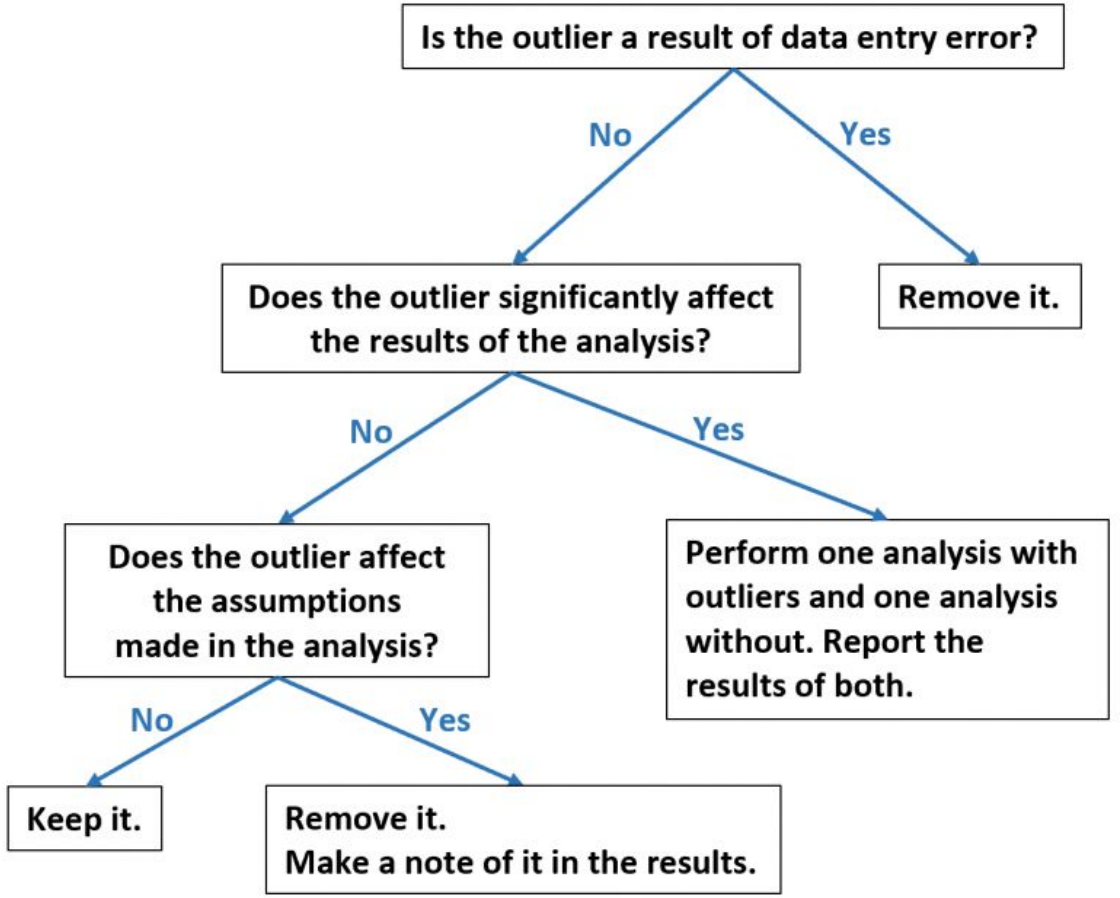
- An outlier if it is:
  - Less than Q1 - 1.5IQR
  - Greater than Q3 + 1.5IQR

IQR = Interquartile Range



Eliminating outliers

- Some outliers represent natural variations in the population, and they should be left as is in your dataset.
  - These are called true outliers.
- Other outliers are problematic
  - measurement errors
  - data entry or processing errors
  - or poor sampling.



Is the Outlier a Result of Data Entry Error?

Example

- Suppose a biologist is collecting data on the height of a certain species of plants and records the following data:

6.83 cm	7.83 cm
7.51 cm	755 cm
5.21 cm	6.53 cm
5.84 cm	6.31 cm
5.83	5.91 cm

Calculate the mean with and without the outlier

Does the Outlier Significantly Affect the Results of the Analysis?

Fit a simple linear regression model using fertilizer as the predictor variable and plant height as the response variable. Does the outlier affects the significantly the regression model?

Fertilizer	Plant Height
2	4
3	5
4	8
5	10
5	12
6	12
7	14
7	13
8	15
9	20
10	22
35	70

Second test: change the last pair for 35 and 20. Does it still follow the data?

```
import numpy as np
import matplotlib.pyplot as plt

# dataset
x = np.array([2, 3, 4, 5, 5, 6, 7, 7, 8, 9, 10, 35])
y = np.array([4, 5, 8, 10, 12, 12, 14, 13, 15, 20, 22, 70])

# prediction model
model = np.polyfit(x, y, 1)
predict = np.poly1d(model)

# new data prediction
x_lin_reg = np.arange(min(x)-1, max(x)+1, 0.1)
y_lin_reg = predict(x_lin_reg)

# plot
plt.scatter(x, y)
plt.plot(x_lin_reg, y_lin_reg, c = 'r')
```

Example

Make a Whiskers plot

With data = 'Weight':[45, 88, 56, 15, 71],
'Name':['Sam', 'Andrea', 'Alex', 'Robin', 'Kia'],
'Age':[14, 25, 55, 8, 21

Code Outlier

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
plt.figure()

plt.suptitle("Boxplot")
ax = plt.gca()

df2 = pd.DataFrame({'Weight':[45, 88, 56, 15, 71],
                    'Name':['Sam', 'Andrea', 'Alex', 'Robin', 'Kia'],
                    'Age':[14, 25, 55, 8, 21]})
df2.boxplot(showmeans=True)
# Rotate x axis text values
for tick in ax.get_xticklabels():
        tick.set_rotation(30)

print("\nIn the boxplot below, the box extends from the lower to upper quartile values of the data, with a line at the median.\n \
The whiskers extend from the box to show the range of the data. The triangle indicates the mean value.\n")
```