DATA EXPLORATION WITH PANDAS

CHEATSHEET FOR LESSON 2

1- PANDAS DATAFRAME EXPLORATION GUIDE

- Reading a CSV file as a data frame : df = pd.read_csv('path_to_the_file.csv')
- Explore the first lines : df.head()
- Get shape informations :
  - len(df) #will give us the number of rows
  - len(df.columns) #will give us the number of columns
  - df.shape # or we can use the shape property of our data frame directly.
- Get some general information about each variable of the dataset: df.info()

2- IMPORTANT STATISTICAL MEASURMENTS

- **The mean:** The average value of the variable, also called the "expected value" if the variable is continuous. Noted as μ or E(X) or $\bar{X}$.

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- **The Median:** The median of a variable is the middle value, to calculate it we sort the values in ascending order and pick the value that is at the middle index.
- **The mode:** The mode is the value that is repeated the most.
- **The variance:** The variance is a way to measure the deviation of a variable from its mean. it has many notations: $\sigma^2 = V = \mathbb{V}(X) = \mathbf{Var(X)}$

$$V = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- **The standard deviation:** Another way to measure the deviation of a variable from its mean. The smaller it is, the closest are the values to the mean. The standard deviation is the square root of the variance.

$$\sigma = \sqrt{V} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2}.$$

- The quartiles:
  - Q1 is the same as the 25th percentile
  - Q2 is the same as the 75th percentile.
  - Q2 is the median and it's the 50th percentile.