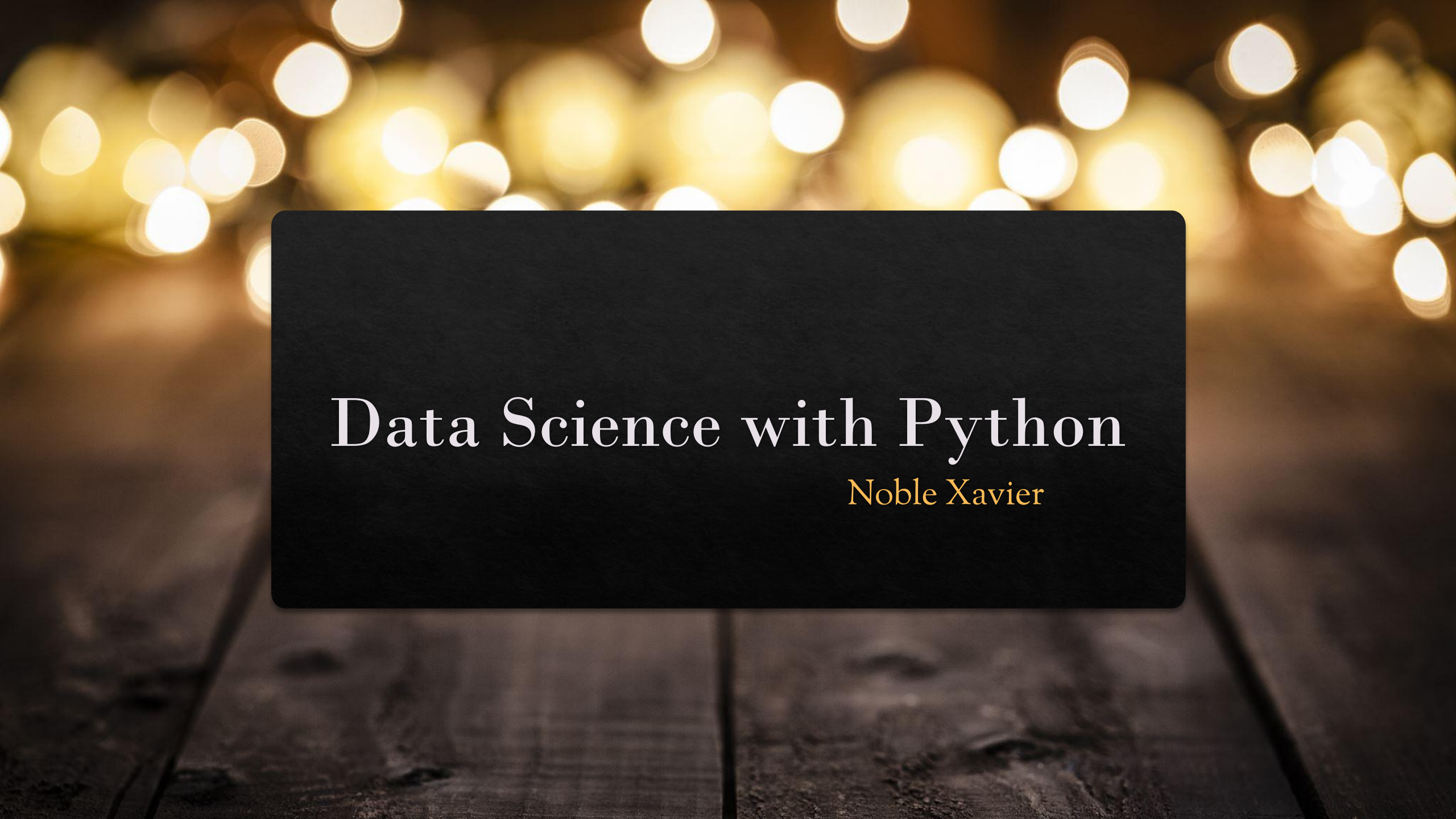

Data Science with Python
Noble Xavier

Introduction to Data manipulation

- Using NumPy, Pandas and Matplotlib we will analyse, visualize and manipulate large datasets in order to extract valuable information and insights

Pre-Processing Steps

- Import relevant libraries
- Import the Data Set to analyse and clean for prediction
- Missing value treatment(Impute the dataset)
  - In numerical data
    - Populate with mean –When there is no outlier
    - Populate with median – When there is outlier
  - Categorical data– Mode imputation – Maximum data category

Pre-Processing Steps

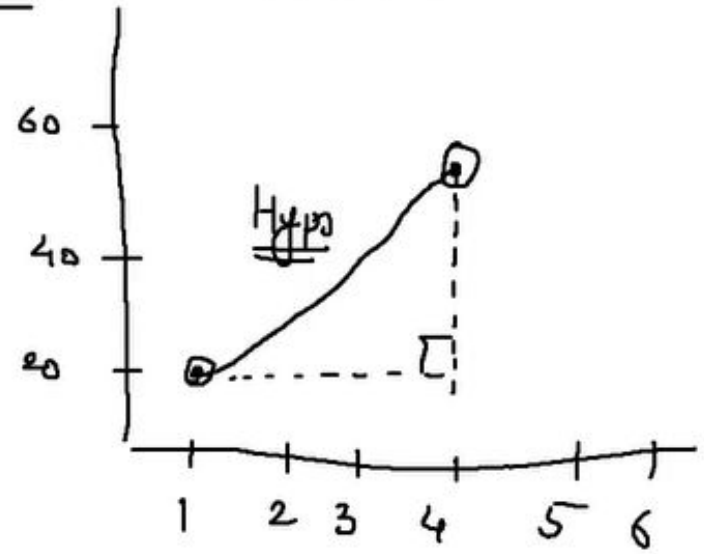
- Encoding
  - Label Encoding – Convert to number
  - One hot Encoding – Create multiple columns based on number of unique record count
- Feature Scaling
  - Standardization – When there no outlier
  - Normalization/ Min Max Scaler : When there is outlier

Measure	Populations	Samples
Size	N	n
Mean	$\mu = \frac{\sum x}{N}$	$\bar{x} = \frac{\sum x}{n}$
Variance	$\sigma^2 = \frac{\sum (x-\mu)^2}{N}$	$s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$
Standard Deviation	$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$	$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$

Reason for Feature Scaling

Salary	Experience
20000	1
40000	2
45000	3
50000	4
y	x

Distance



$$\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

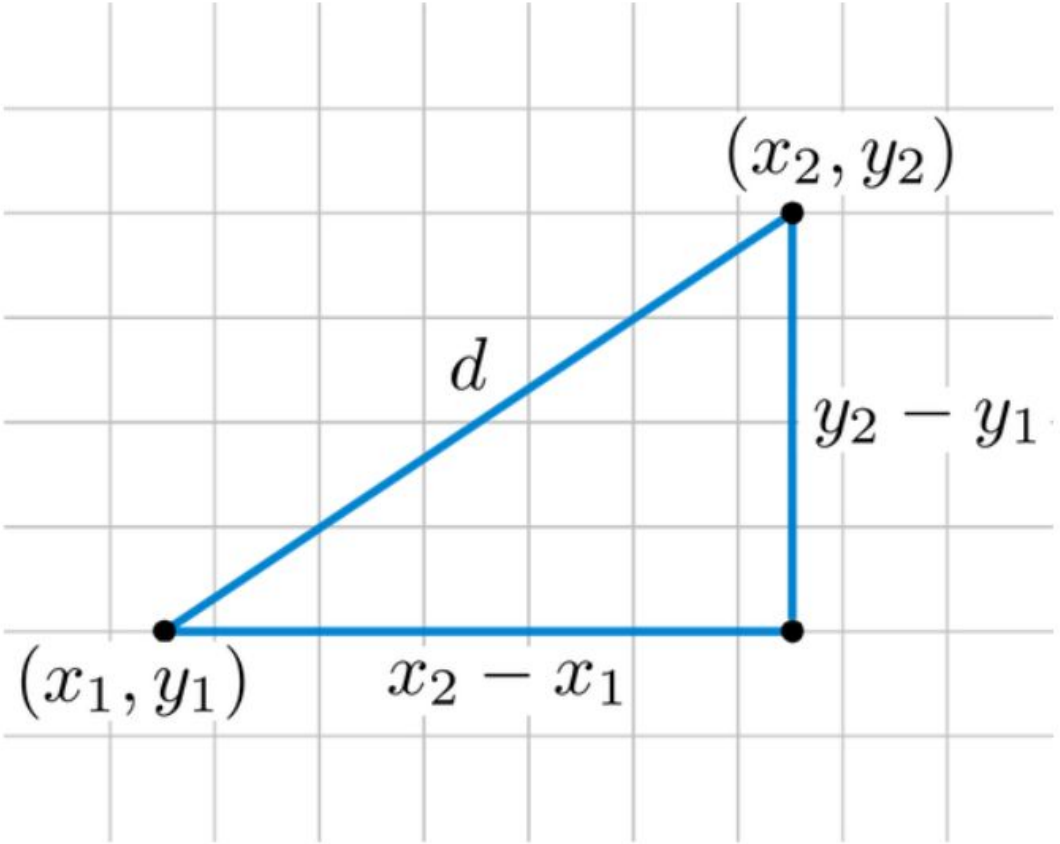
$\sqrt{(4-1)^2+(50000-20000)^2}$ => Salary is dominating Exp.

$\sqrt{9}+$ - Huge = Ignore

Euclidean distance

The Euclidean distance between two points in either the plane or 3-dimensional space measures the length of a segment connecting the two points. It is the most obvious way of representing distance between two points.

The Pythagorean Theorem can be used to calculate the distance between two points, as shown in the figure below. If the points (x_1, y_1) and (x_2, y_2) are in 2-dimensional space, then the Euclidean distance between them is $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$.



For points (x_1, y_1, z_1) and (x_2, y_2, z_2) in 3-dimensional space, the Euclidean distance between them is $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$. For example, the Euclidean distance between $(-1, 2, 3)$ and $(4, 0, -3)$ is $\sqrt{25 + 4 + 36} = \sqrt{65}$.

Standardization

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

Min Max Scaler

$$z = \frac{x - min(x)}{[\max(x) - min(x)]}$$