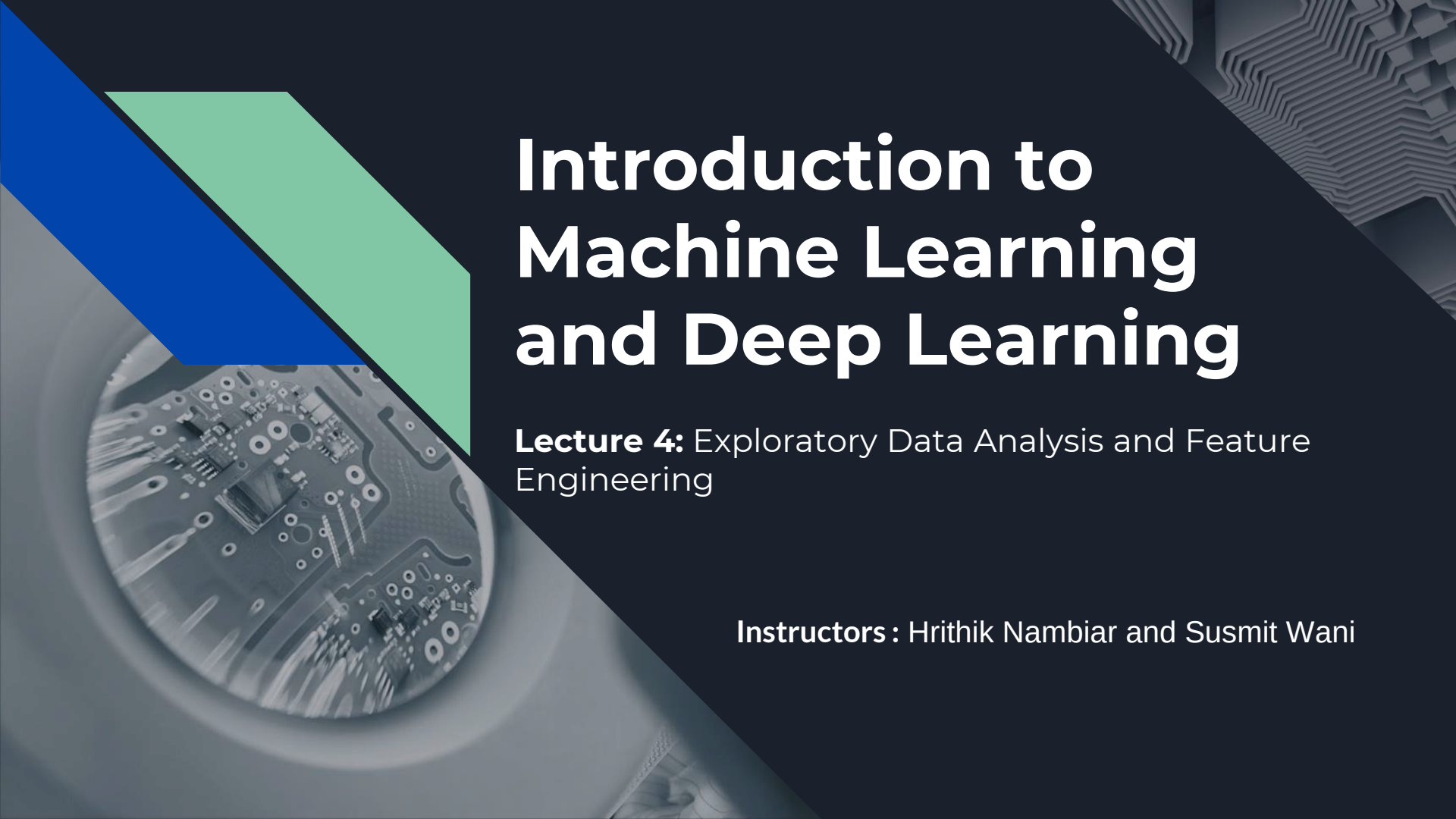


Introduction to Machine Learning and Deep Learning

Lecture 4: Exploratory Data Analysis and Feature Engineering

Instructors : Hrithik Nambiar and Susmit Wani

Recap

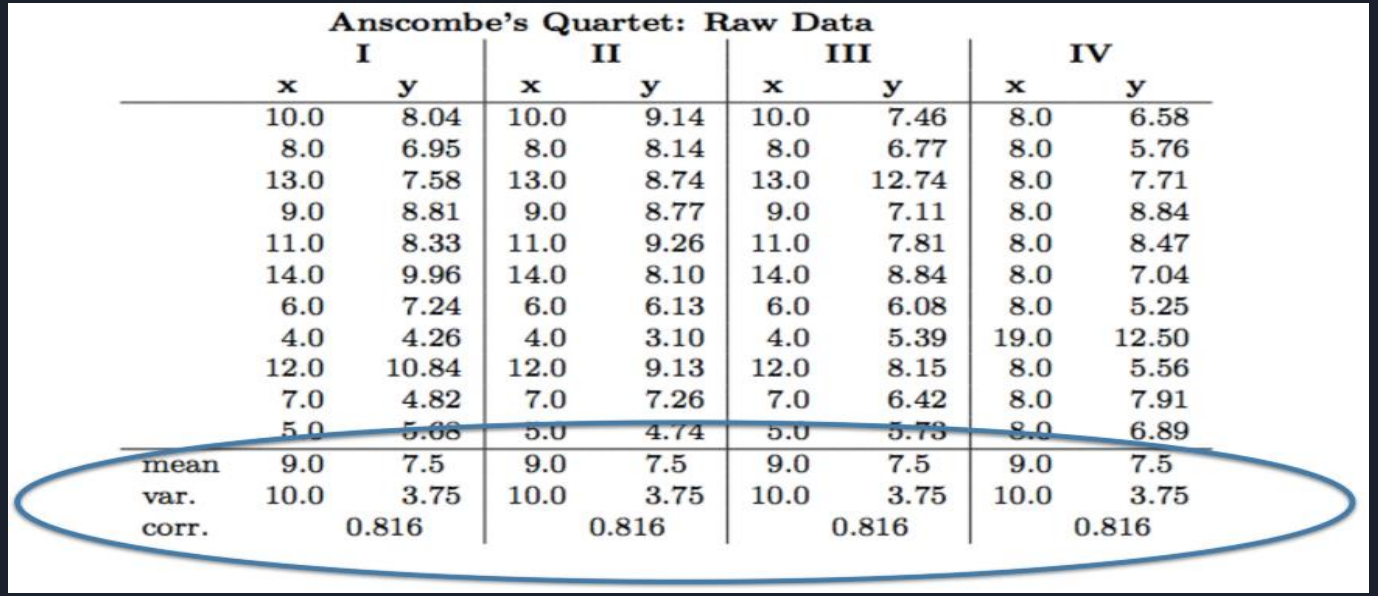
- What is ML and its history
- Introduction to Regression based problems.
- Intro to classification based questions.
- Intro to python (conditional loops, tuples, function call etc)
- Intro to Pandas library (working with CSV files and dataframes)
- Intro to Numpy library- mathematical functions of Numpy, Why to use Numpy.
- Probabilistic approach to ML, Explanation of Bayes theorem.

VISUALIZATION

“A picture is worth thousand words “

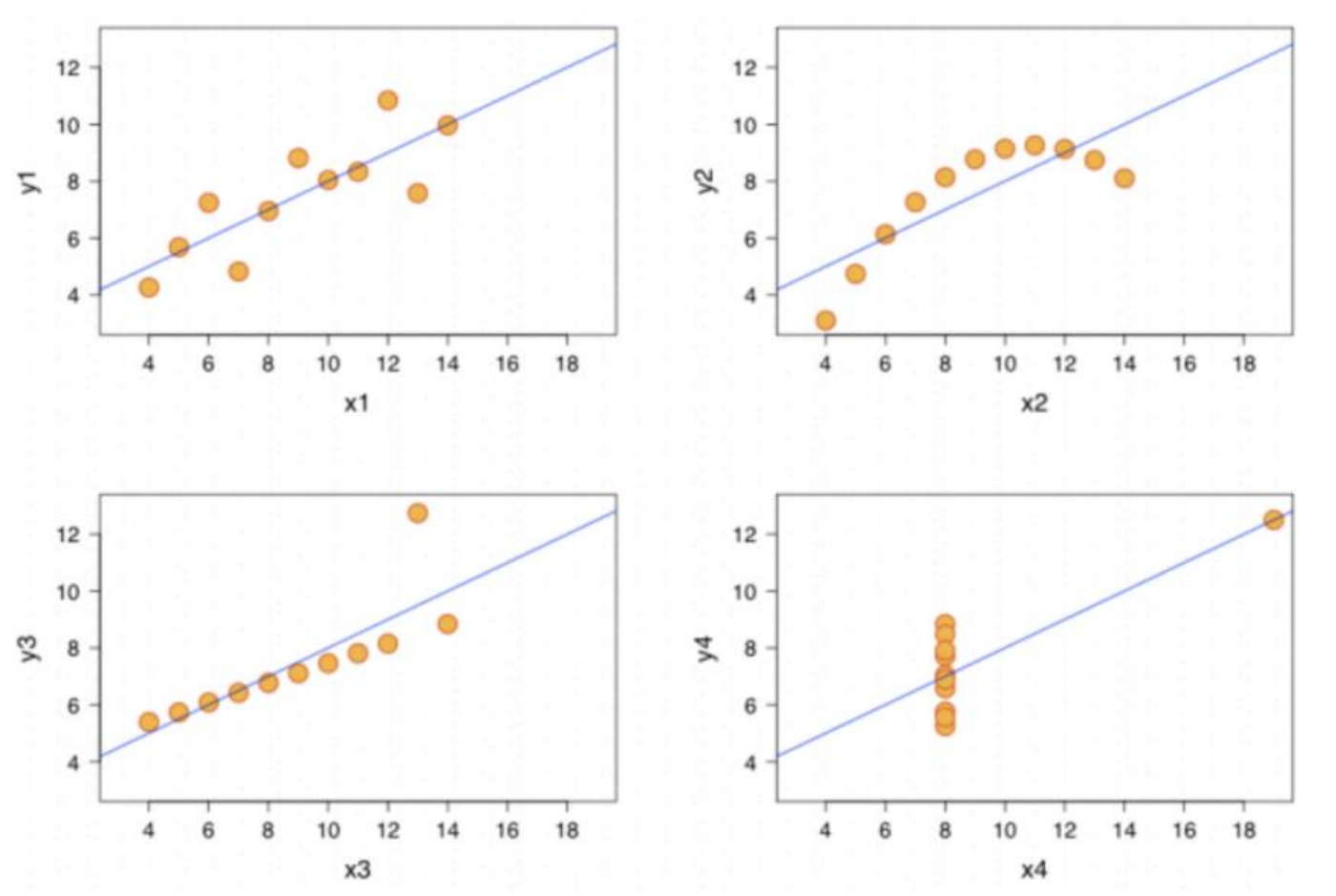
Anscombe’s Quartet

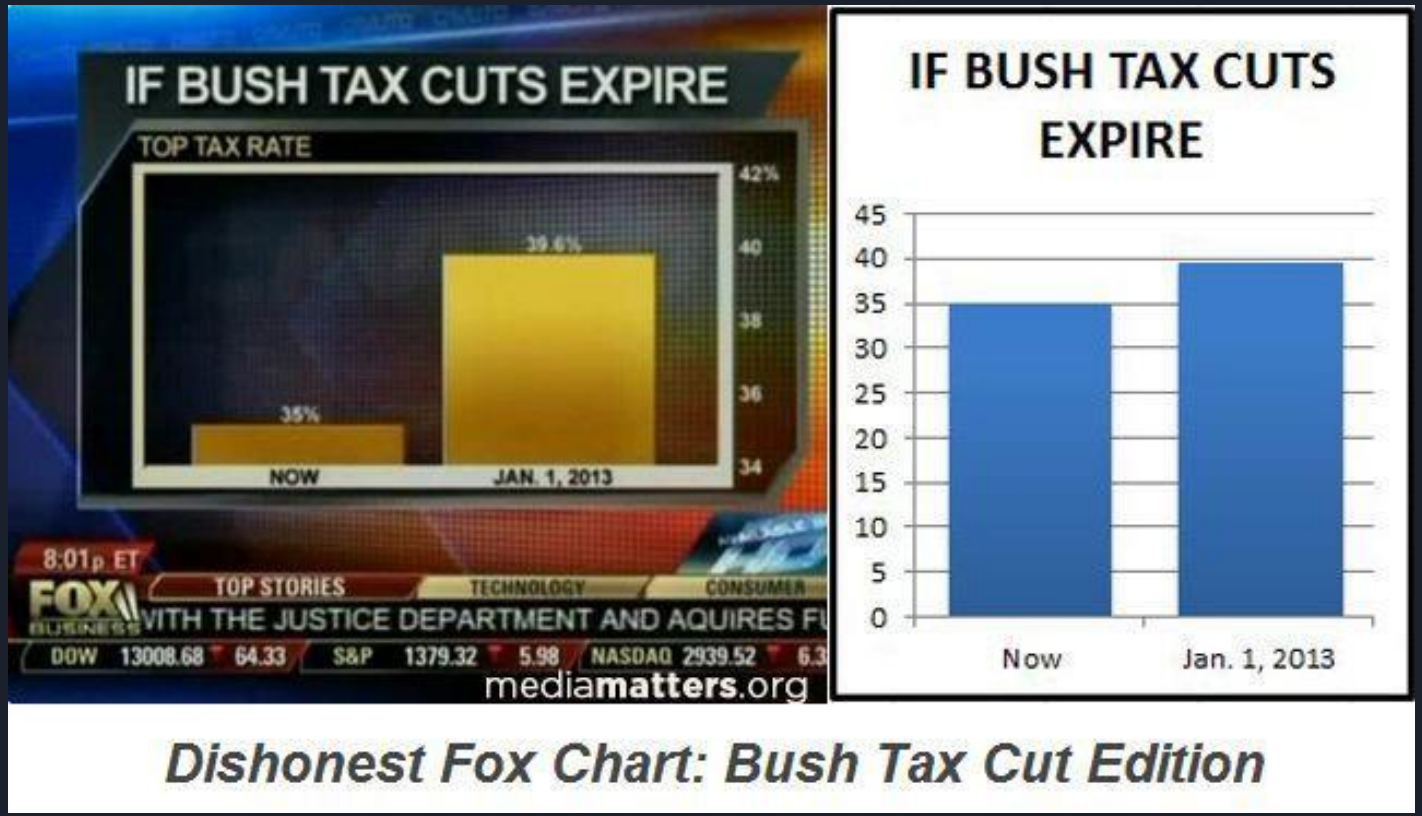
Number’s can be misleading.



Anscombe’s Quartet: Raw Data

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
corr.		0.816		0.816		0.816		0.816

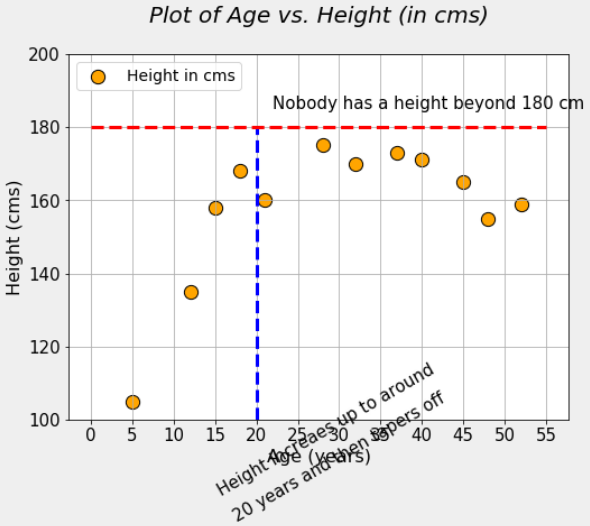

y1
x1
y2
x2
y3
x3
y4
x4
4
6
8
10
12
14
16
18

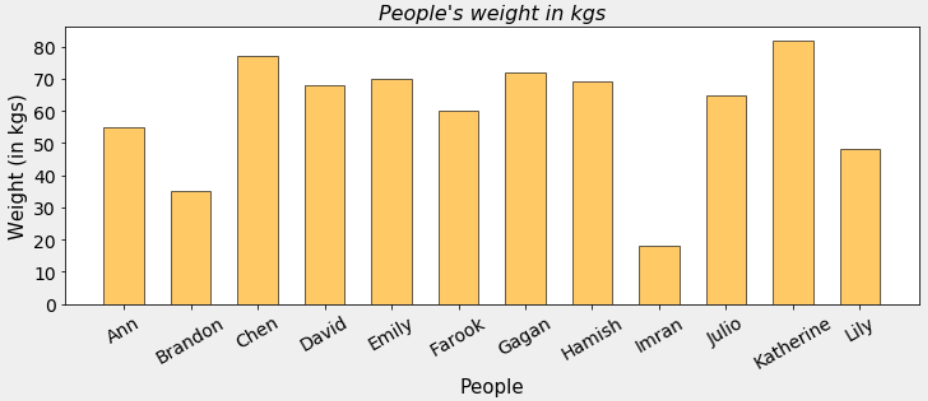


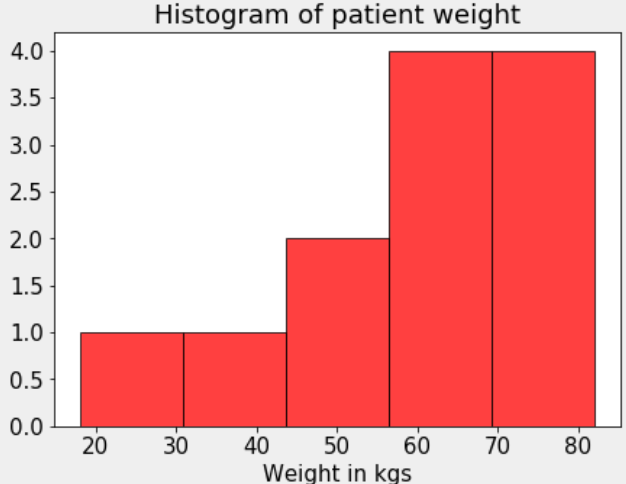
Dishonest Fox Chart: Bush Tax Cut Edition

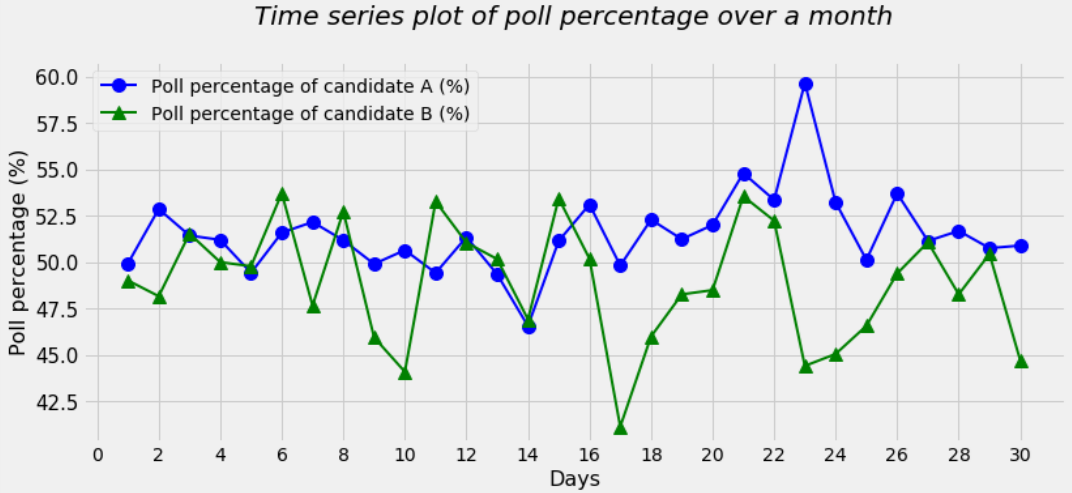
Types of plot

- **scatterplot**
- **boxplot**
- **histogram**
- **bar charts**
- **heatmap**
- **time series line plot**

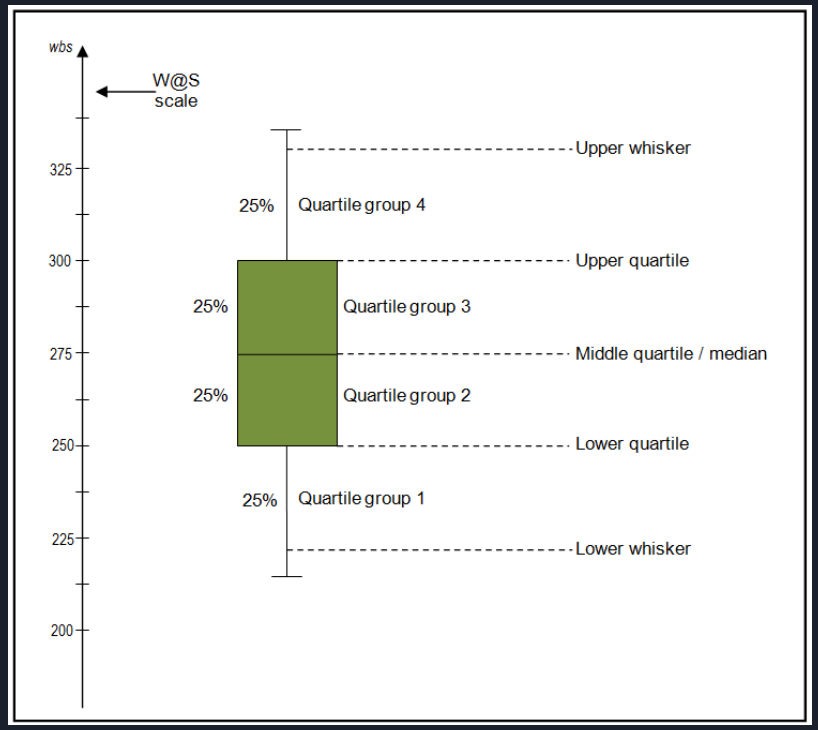

Plot of Age vs. Height (in cms)
Height in cms
Nobody has a height beyond 180 cm
Height (cms)
Height age (keeps increases up to around 20 years and stops off)

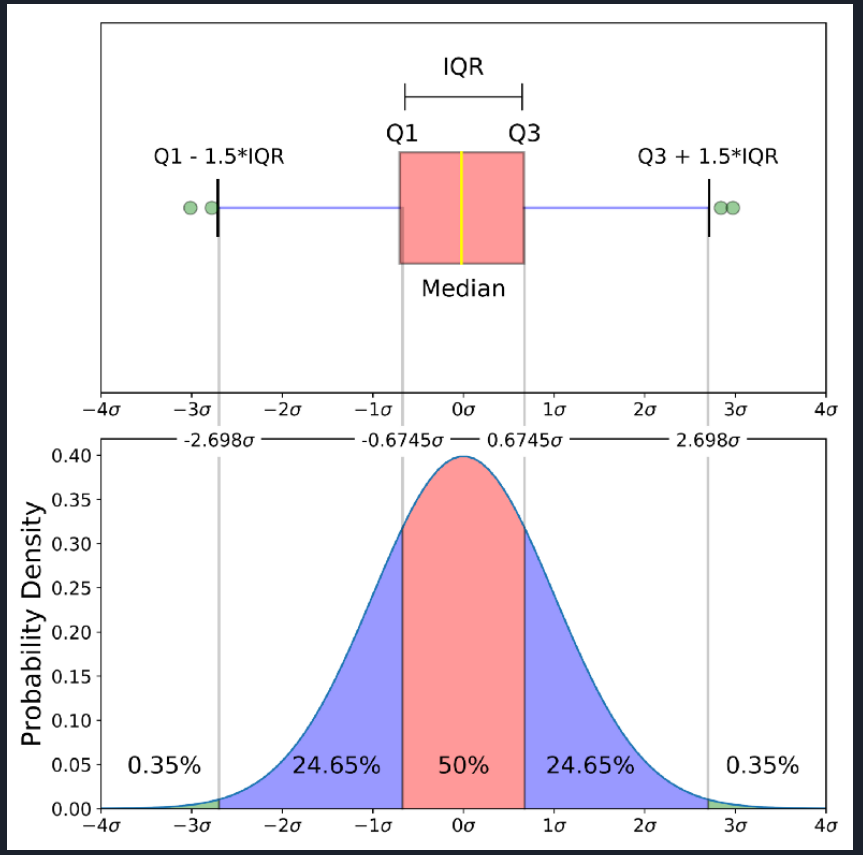

People's weight in kgs
Weight (in kgs)
Ann
Brandon
Chen
David
Emily
Farook
Gagan
Hamish
Imran
Julio
Katherine
Lily
People

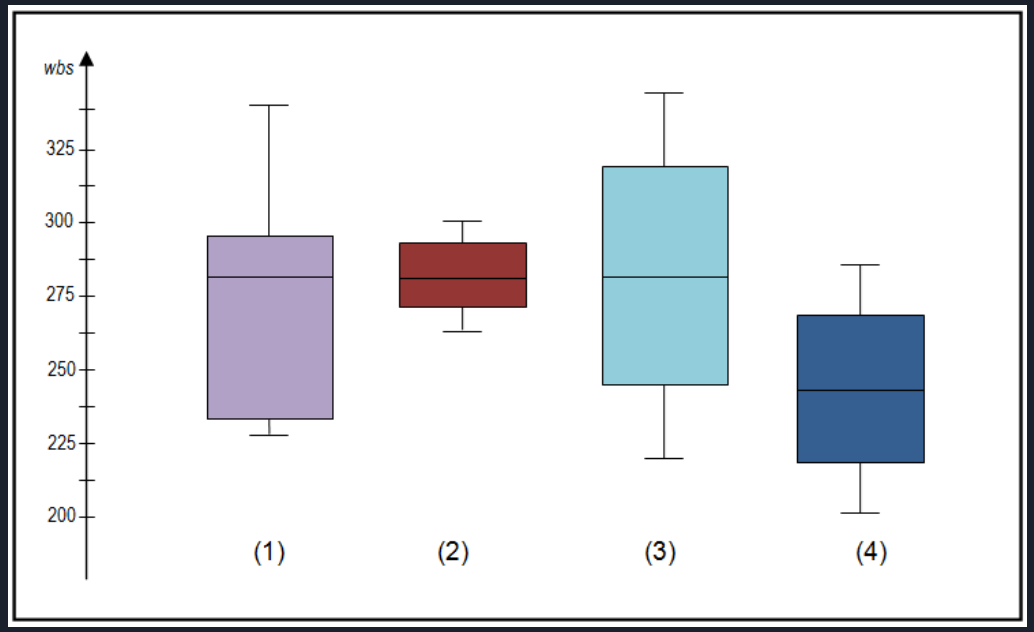

Histogram of patient weight
Weight in kgs

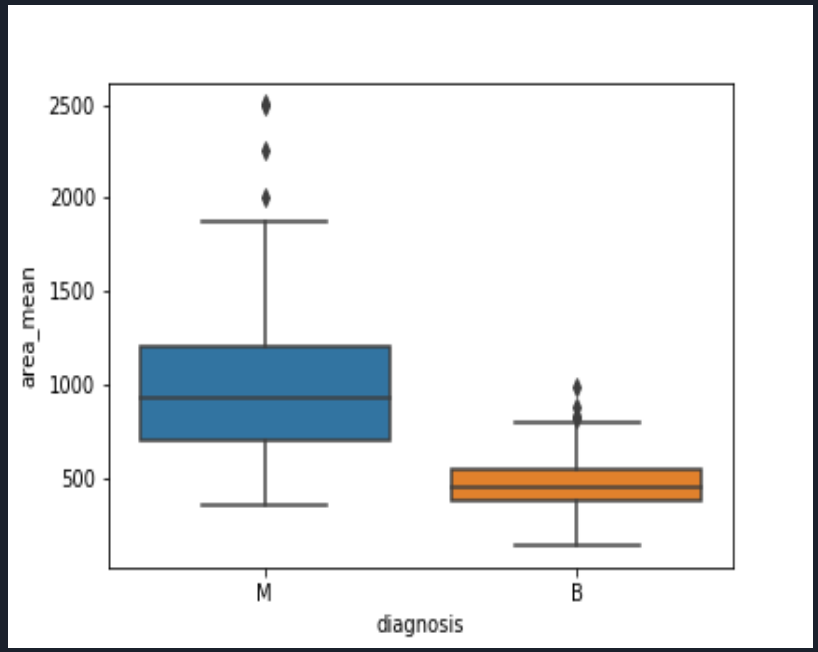

Time series plot of poll percentage over a month
Poll percentage of candidate A (%)
Poll percentage of candidate B (%)
Poll percentage (%)
Days

BOX PLOT


wbs
W@S scale
325
300
275
250
225
200
Upper whisker
25%
Quartile group 4
Upper quartile
25%
Quartile group 3
Middle quartile / median
25%
Quartile group 2
Lower quartile
25%
Quartile group 1
Lower whisker


IQR
Q1
Q3
Q1 - 1.5*IQR
Q3 + 1.5*IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
-2.698σ
-0.6745σ
0.6745σ
2.698σ
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.35%
24.65%
50%
24.65%
0.35%


wbs
325
300
275
250
225
200
(1)
(2)
(3)
(4)

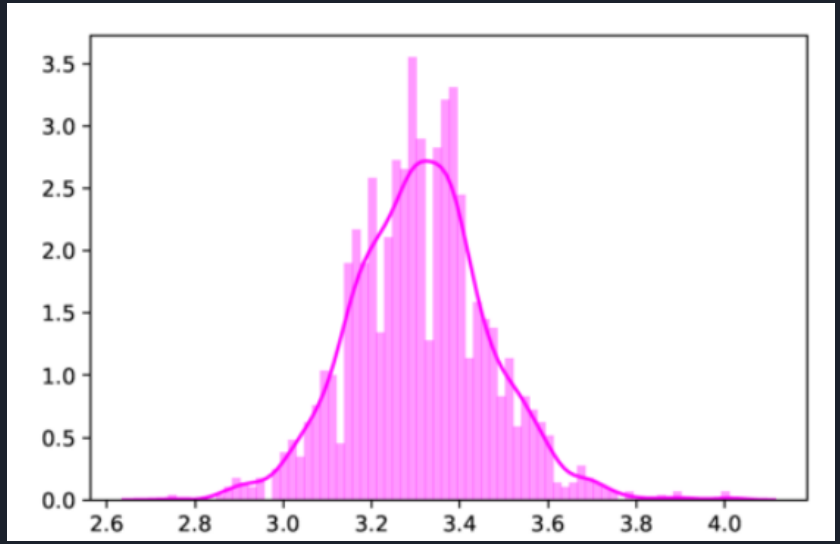

2500
2000
1500
1000
500
area_mean
M
B
diagnosis

MATPLOTLIB

Exploratory Data Analysis

What is EDA?

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.



Why EDA?

The main purpose of EDA is to help look at data before making any assumptions.It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables. As the name suggests, it’s all about exploring the data that we have.

EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning.

The general path

- Import dataset
- Check for missing values and fill/drop
- Check for outliers and drop them
- Check skewness of data (Distribution of data)
- Regularisation of data
- Correlation matrix
- New Feature generation
- Handling Object type data

Thank you!