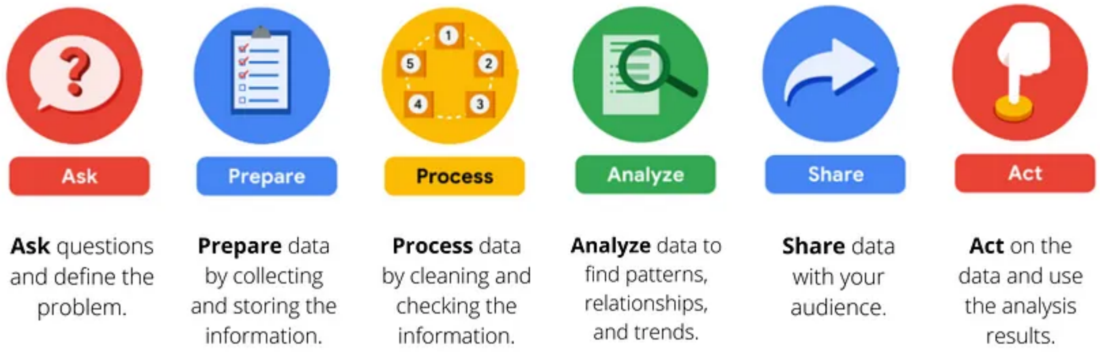
Data Analysis Process

- **Prudhvi Vardhan Notes**


Ask
Prepare
Process
Analyze
Share
Act
Ask questions and define the problem.
Prepare data by collecting and storing the information.
Process data by cleaning and checking the information.
Analyze data to find patterns, relationships, and trends.
Share data with your audience.
Act on the data and use the analysis results.

1. Asking questions

- What features will Contribute to my analysis?
- What Features are not important for my analysis?
- Which of the features have a strong Correaltion?
- Do i need Data Preprocessing?
- What Kind of Feature engineering / Maupulation is required?

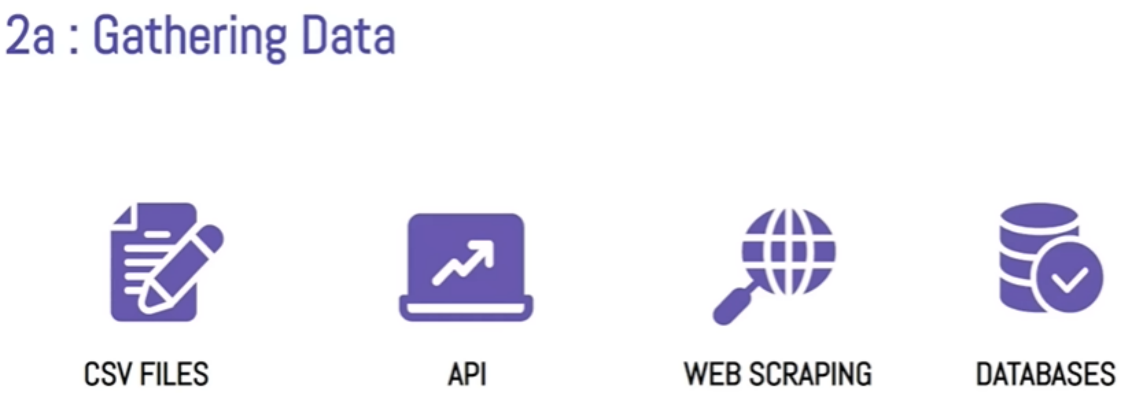
how to Ask Better Questions?

- subject Matter Expertise
- Experience

2. Data Wrangling / Munging

Data Wrangling , sometimes referred to as data munging , it is the process of **Transforming and Mapping data** from on "raw" data form into another format with intnet of making it more appropriate and valuable for a variety of downstream purposes such as analytics

- Gathering data
- Assesing data
- Cleaning data



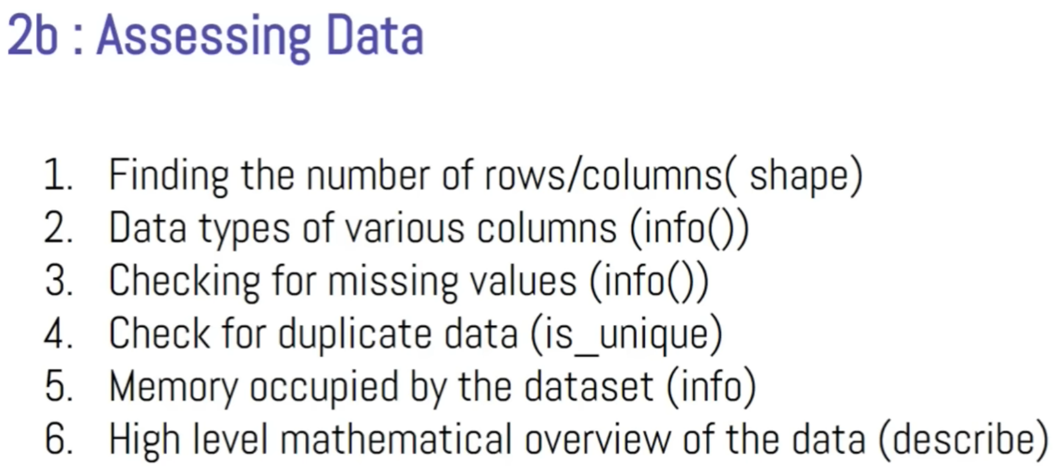
2a : Gathering Data

CSV FILES

API

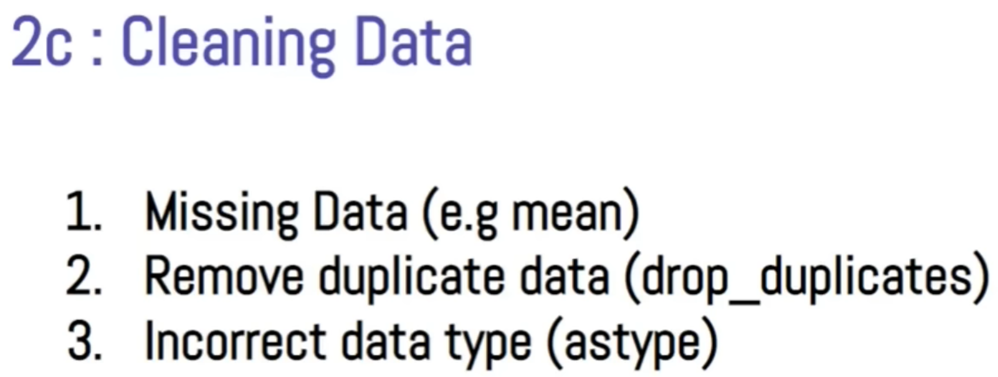
WEB SCRAPING

DATABASES



2b : Assessing Data

1. Finding the number of rows/columns(shape)
2. Data types of various columns (info())
3. Checking for missing values (info())
4. Check for duplicate data (is_unique)
5. Memory occupied by the dataset (info)
6. High level mathematical overview of the data (describe)



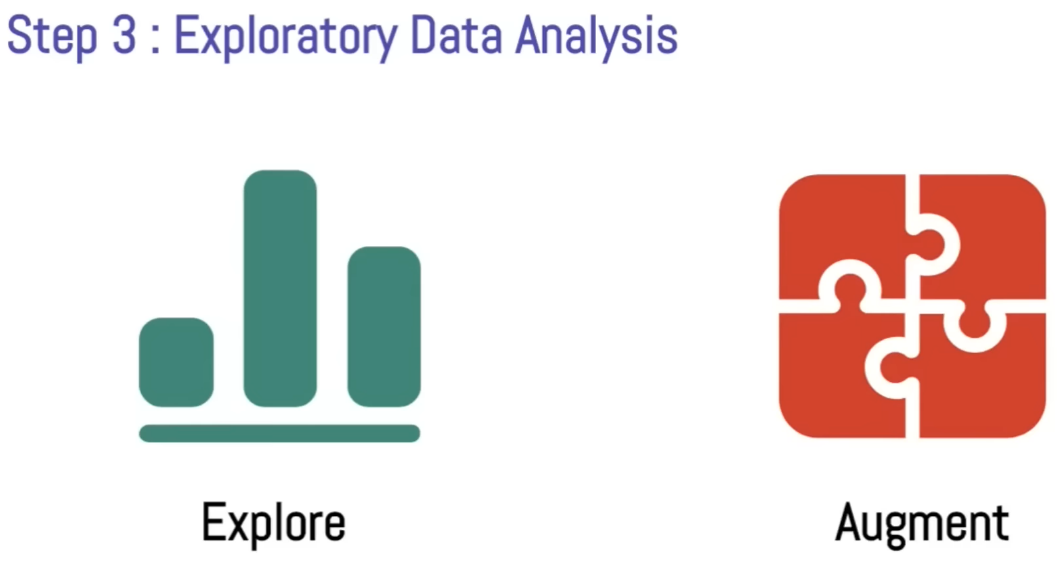
2c : Cleaning Data

1. Missing Data (e.g mean)
2. Remove duplicate data (drop_duplicates)
3. Incorrect data type (astype)

3. Exploratory Data Analysis

To analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

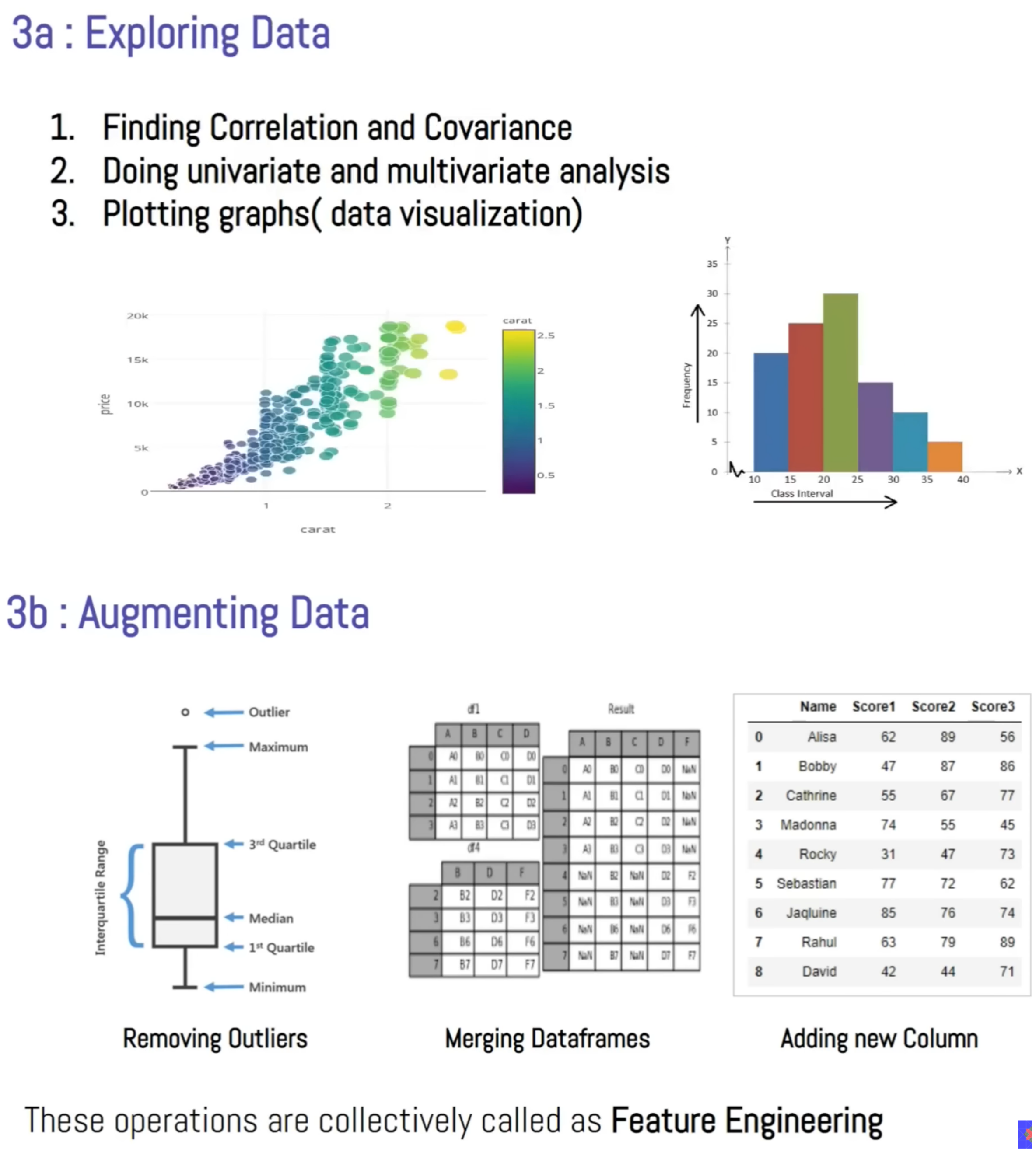
- Exploring Data
- Augmenting Data



Step 3 : Exploratory Data Analysis

Explore

Augment



3a : Exploring Data

1. Finding Correlation and Covariance
2. Doing univariate and multivariate analysis
3. Plotting graphs(data visualization)

3b : Augmenting Data

Removing Outliers

df1

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3

df4

	B	D	F
2	B2	D2	F2
3	B3	D3	F3
6	B6	D6	F6
7	B7	D7	F7

Result

	A	B	C	D	F
0	A0	B0	C0	D0	NaN
1	A1	B1	C1	D1	NaN
2	A2	B2	C2	D2	NaN
3	A3	B3	C3	D3	NaN
4	NaN	B2	NaN	D2	F2
5	NaN	B3	NaN	D3	F3
6	NaN	B6	NaN	D6	F6
7	NaN	B7	NaN	D7	F7

Merging Dataframes

	Name	Score1	Score2	Score3
0	Alisa	62	89	56
1	Bobby	47	87	86
2	Cathrine	55	67	77
3	Madonna	74	55	45
4	Rocky	31	47	73
5	Sebastian	77	72	62
6	Jaqluine	85	76	74
7	Rahul	63	79	89
8	David	42	44	71

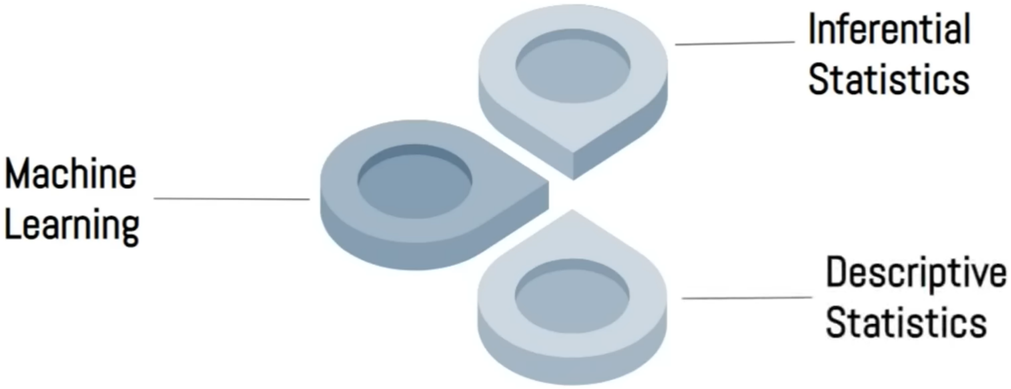
Adding new Column

These operations are collectively called as **Feature Engineering**

4. Drawing Conclusions

- FROM Machine learning Algorithms
- Descriptive statistics
- Inferential Statistics

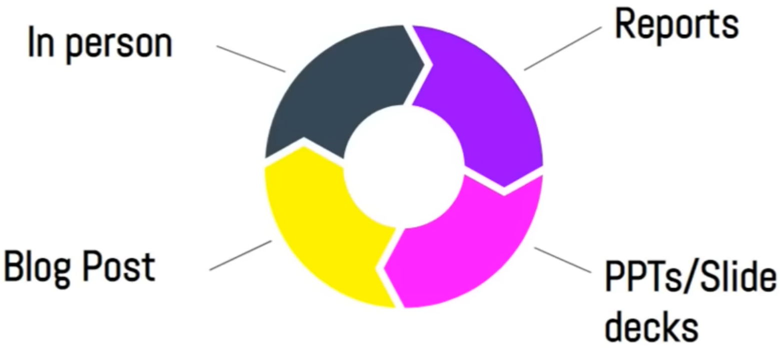
Step 4 : Drawing Conclusions



5. Communicating Result / story Telling

- Using PowerBI
- Tableau

Step 5 : Communicating Results/ Data Storytelling



.

.

.