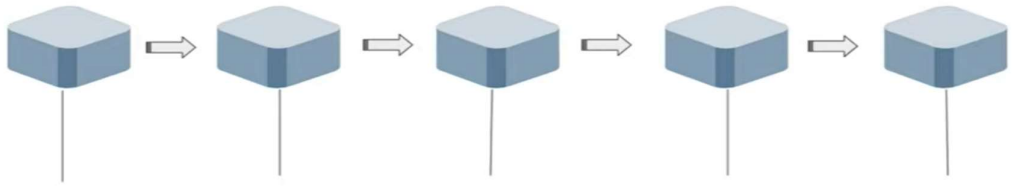
What is Data Analysis

Data analysis is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusions and supporting decision-making.

—Wikipedia

DATA ANALYSIS PROCESS



Asking Questions → Data Wrangling → Exploratory Data Analysis → Drawing Conclusions → Communicating Results

Step 1 : Asking Questions

Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S

1. What features will contribute to my analysis?
2. What features are not important for my analysis?
3. Which of the features have a strong correlation?
4. Do I need data preprocessing?
5. What kind of feature manipulation/engineering is required?

How can I ask better questions?

Subject Matter Expertise

Experience

Step 2 : Data Wrangling/Munging

Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.

1. Gathering Data
2. Assessing Data
3. 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)

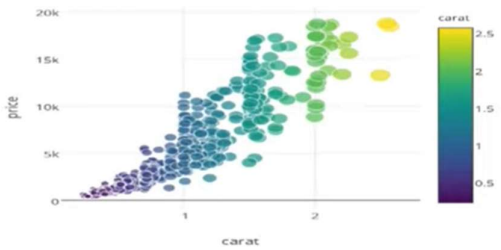
2c : Cleaning Data

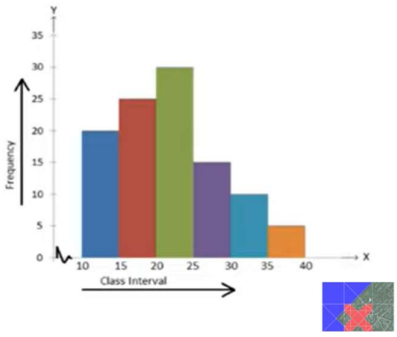
1. Missing Data (e.g mean)
2. Remove duplicate data (drop_duplicates)
3. Incorrect data type (astype)

Step 3 : Exploratory Data Analysis

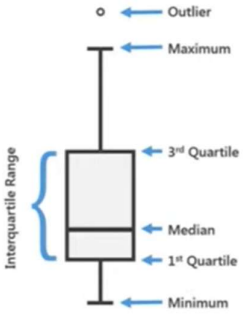
3a : Exploring Data

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

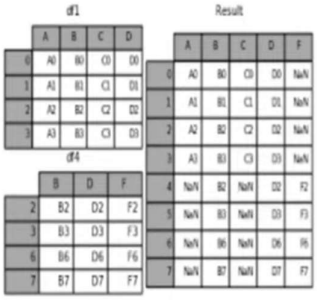




3b : Augmenting Data



Removing Outliers



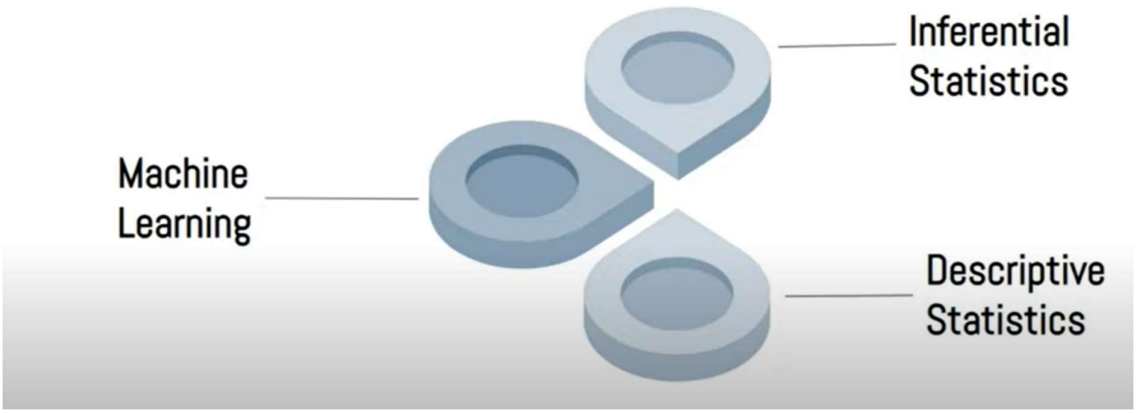
Merging Dataframes

	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**

Step 4 : Drawing Conclusions



Some example conclusions based on Descriptive Statistics

1. Is Rohit Sharma a better batsman in 2nd innings (IPL Dataset)?
2. Does being a female increases your chances of Survival (Titanic Dataset)?
3. Is Delhi the most costly place for eating out(Zomato Dataset)?

Step 5 : Communicating Results/ Data Storytelling

The fun part...