Exploratory Data Analysis (EDA)

Session-1

In the world of data & data science, EDA plays a prominent role. It is the first step in any data science use-case after data gathering & cleaning.

Let's take a look at the life cycle of Data Science Project:

Life Cycle of a Data Science (DS) Use-Case begins with:

1) Requirement Gathering: At this stage, the problem statement is understood based on which team is formed. Roles are defined & responsibilities are distributed. After this, the team begins to think what data do they require & how would they get that.

2) EDA: At this stage, the data is gathered through various sources like, third party APIs, Web Scrapers & other data vendors. Sometimes, the organization for which we're working gives us the data.
Data is stored in various SQL & NoSQL Databases.
Statistical Analysis done over the data.

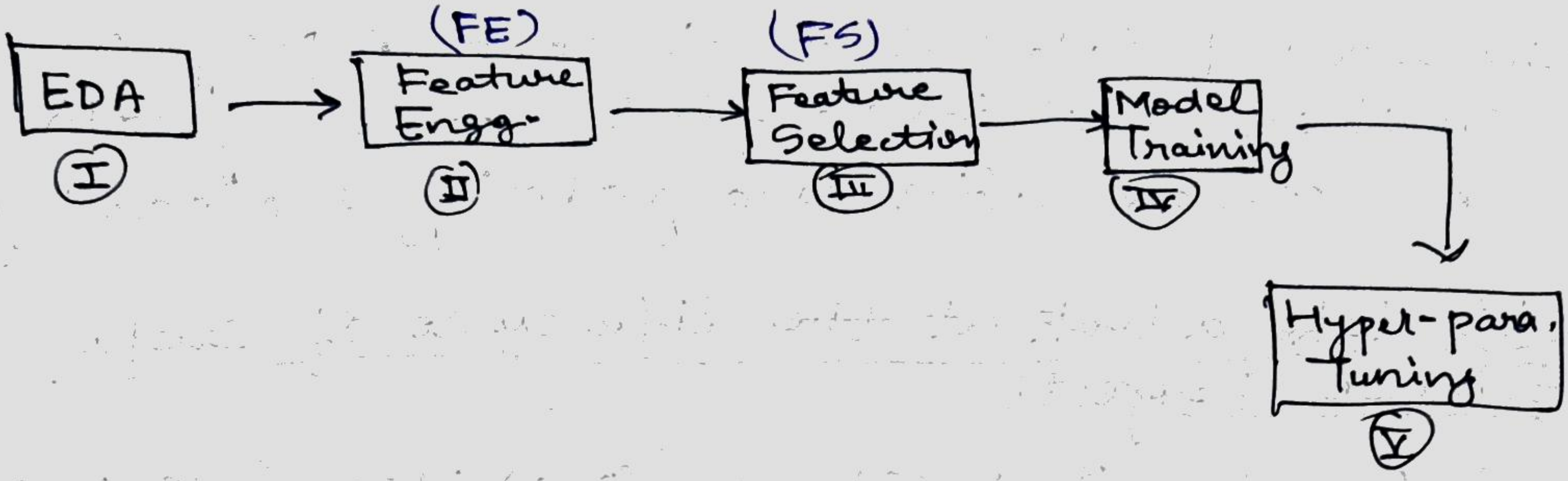
3) Feature Enggineering: It is the next step in which using domain knowledge we extract features from raw data.

4) Feature Selection: Based on knowledge & experience, we select the features & discard others that ~~will~~ are not going to play a prominent role in our ML Algo.

5) Model Creation: A suitable ML model is created based on the requirement & data.

6) Hyper-parameter Tuning: This helps in increasing the over-all efficiency of the model by tuning the parameters of Mathematical equations of ML Model.
It's a never ending process.

Diagrammatic Representation of Life Cycle of DS Project:



Since, these sessions will primarily focus on EDA.
∴ We're not going to worry about the ongoing steps but we'll cover them gradually!!

EDA or Exploratory Data Analysis or Statistical Data Analysis requires Maths & Statistics, as the name suggests.

∴ Statistics: It is the science of collecting, organizing, analyzing, presenting/visualizing & drawing ~~dat~~ conclusions from data.

Data: " Facts or pieces of Information"

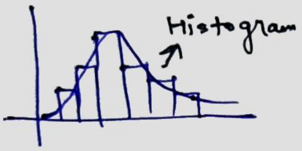
E.g. Ages of Students, weights of people.

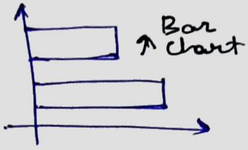
Types of Statistics

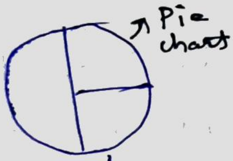
Descriptive Stats → (EDA + FE)

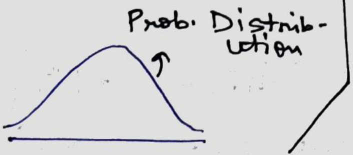
Inferential Stats

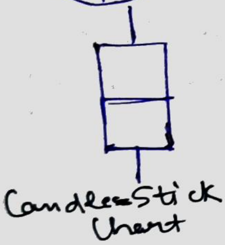
→ Consists of organizing & summarizing the data.

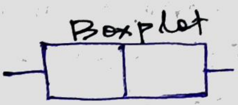


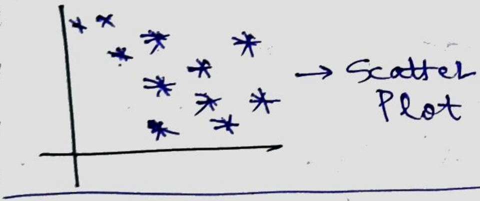












→ consists of collecting sample data & making conclusions about ~~the~~ population data using some Statistical experiments.

→ Hypothesis Testing

→ Z-test | → P-Value

→ t-test

→ χ^2 (Chi-Square)-Test

→ F-test (ANOVA)

e.g. Exit Polls of News channels after Elections, Surveys a handful people & predicts the results for wholestate/country.

Sample Data → Population Data

E.g. Let's say there are 20 classrooms in a University & you've collected the age & weights of students in one class room:

Ages: {21, 20, 18, 34, 17, 22, 24, 25, 26, 23, 22}

Weights: { 60, 65, 56, 70, ..., 67}

Based on ~~the~~ classification that we've done above of statistics, we can ask the following questions:

Descriptive Stats: What is the average age of students in the classroom?

→ Relationship b/w Age & weight?

Inferential Stats: Is the average age of the students in the classroom less than | greater than | equal to the average age of students in the university?

→ Sampling Techniques:

1) Simple Random Sampling.

2) ~~Strat~~ Stratified Sampling.

3) Systematic Sampling

4) Convenience Sampling

1) Simple Random Sampling: This sampling is simplest of all samplings. In this, we choose a fixed number of items from the population. The probability of every item getting selected is equal because of which the sampling becomes unbiased.

E.g. Drawing a lottery winner.

2) Stratified Sampling: The word ~~means~~ 'strata' means to group/layers. We group the data based on some categorical feature & then sample data from them as we're doing previously.

Eg. In Exit poles: Making Groups

Less than 18 (Not useful) ∵ Didn't Vote

More than 18 (Useful) ∵ They vote

3) Systematic Sampling: We draw every nth item from the population.

E.g. Banks choosing that they'll call every 3rd customer for Loan / Credit Card.

→ Door-to-Door Salesman decides to visit every 2nd house that comes in the way.

4) **Convenience Sampling:** It is used when we want correct, true & quality data.

We draw only those people that are ready to comply with us.

E.g. Youtube, these days runs a survey for its improvement. Survey contains some questions regarding our personal experience with youtube ~~but~~. But, it does not forces anyone to fill that survey. We can easily skip that.

Variable: A variable is a property that can take any values.

↳ (Vary - Able) → Able to Vary

Types of Variables

Quantitative Variables

→ can be measured numerically

→ e.g. Age, weight, height, rainfall, temp, distance

→ They can contain any number of values.

Age | Weight

↓

-∞ ← → ∞

Qualitative Variables

→ Categorical Variables

~~(Based on som~~

→ e.g. Gender, Marital Status

→ They can only have a fixed no. of values.

Gender

M F O

---

* Quantitative Variables can be classified further:

e.g. Age: 15, 16, 17, 18, ... {There can't be any other no b/w 15 & 16}

(Whole Numbers)

Weight: 35, 36, 36.5, 37, ... {There can be another value b/w 36 & 36.5 like 36.4}

(Real Numbers)

Quantitative Variables

Discrete

→ The whole number data that we discussed previously, comes under discrete variables

→ E.g. Pincode
(Fixed, WholeNumbers)

Continuous

→ The Real Numbers data comes under Continuous Variables.

→ E.g. Height, Rainfall
(Not Fixed, Real Number)

e.g. Let's classify variables to the data:

→ Marital Status | Gender : Categorical | Qualitative

→ River Length: Continuous Variable

→ Movie Duration : Continuous Variable

→ Pincode : Discrete (100110, 100111, 100112,.... etc)

→ IQ : Discrete (100, 110, 120,... etc.)