Exploratory Data Analysis

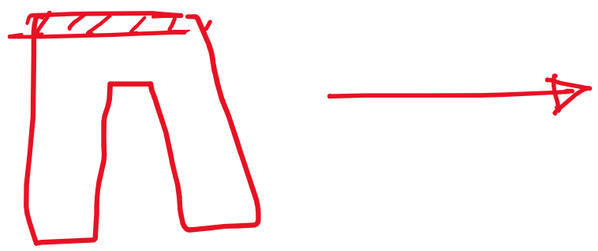
Exploratory data analysis (EDA) is an integral & very important part of (machine learning) → Learning the patterns from data.

For building a good machine learning model we need to have good data.

In real life we often come across data which are noisy, sometimes redundant & may contain outliers & missing values.

Example:- Suppose we want to build a machine learning based model to identify whether a customer is likely buy a recommended product or not.

In order to create that model we need data.



User

pant, brand.

The ML model will try to find out how likely the user is going to purchase that pant.

	u_1	u_2	u_3	.....	u_m
pant 1	0.02				
pant 2	0.05				
pant 3	0.1				
⋮					
pant k	0.01				

u_1 →

For building this kind of ML model what type of data do we need?

The data that we need to build the model

(1) The product data.

product type, price, previous purchase, discount, brand, ------

(2) The user data

age, gender, purchase history, search history, married/unmarried, pet's name.

<1> To identify which data are required? (We need domain knowledge)

<2> De-noising / Data cleaning.

<3> Find out if there is any outlier? → Anomaly / Mistake

<4> Check for the missing values.

<5> Identify the important features.

All these part of Exploratory Data Analysis. (EDA)

Exploratory Data Analysis (EDA) is a collection of techniques that will help us to understand the data better.

Objectives of- EDA :—

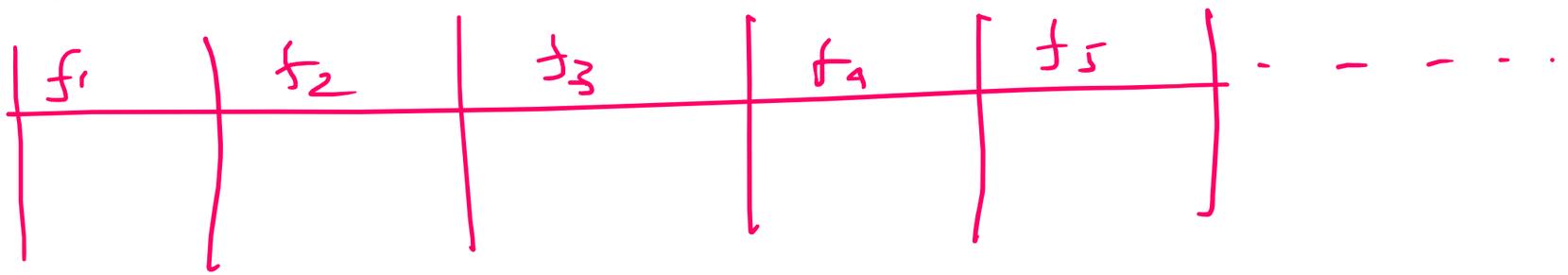
<1> Discover patterns (both analytically & visually)

<2> Spot any anomalies / outliers.

<3> Frame hypothesis.

<4> Check assumptions / hypothesis.

What type of- Analysis we do in EDA ?



→ Univariate Analysis (taking one variable at a time)

- → Categorical variables → count & plot the frequencies of different categories.
- → numerical (quantitative) variables → distribution of the variable.

We can do → spotting anomalies / outliers.

outlier treatment

missing value treatment.

→ Bivariate Analysis (two variables at a time).

- → trends or patterns over time.
- → Correlation
- → Hypothesis formation & testing
- → Feature importance analysis.