Introduction to EDA

What is Exploratory Data Analysis?

A person recently broke their phone while traveling by train to their hometown. Initially, they felt upset but then decided to purchase a new phone. They decided on a budget of around 35-40K for the new phone. Considering their love for gaming, they prioritized getting a smartphone with the latest processor and sufficient RAM. Once they finalized their budget, their next step involved visiting various websites to gather information about the latest gaming smartphones priced under 40K. The most intriguing part of purchasing their new smartphone was **analyzing** and **comparing** the gathered information to identify the smartphones that best matched their requirements.

Most of us go through these same steps while purchasing any new item. You fix your budget and other requirements, then go through several options, and after comparing those options, you choose the best one that can fulfil your needs.

Data Analysts must also go through these steps while working on any problem.

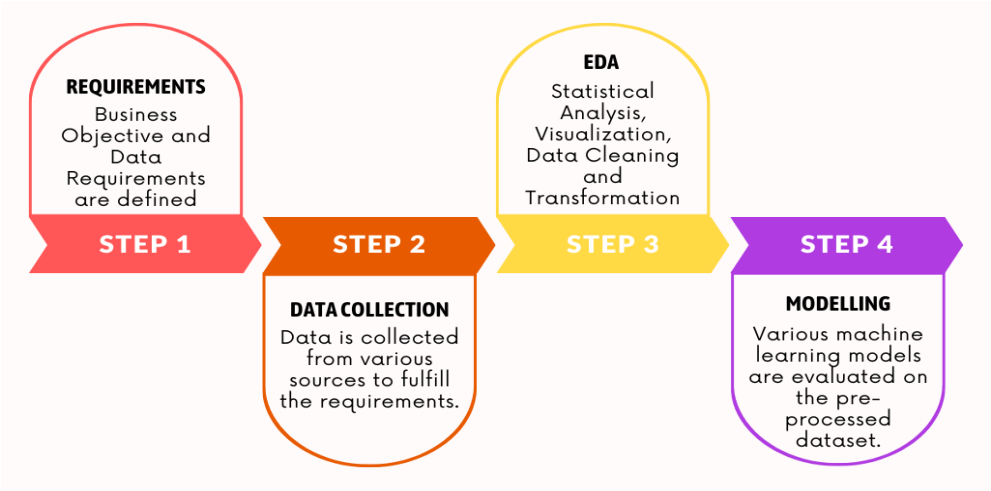
- First, **define the requirements**, like fixing the budget or the kind of processor you need.
- **Collect data** related to those requirements from various sources or websites.
- Then perform **Exploratory Data Analysis (EDA)** on the collected dataset using various visualization and statistical tools. This step is

crucial because while performing EDA, you understand what your data looks like and what kind of anomalies it contains. Here "**Exploratory**" means "**to discover something or learn the truth about something**."

Now let's start learning about EDA in the context of Data Analysis.

Defining EDA

Before defining EDA, let's see where it fits in the Data Science lifecycle.



Let's walk through this lifecycle diagram with the help of an example.

Problem Statement

Imagine you are trying to make a website that helps users to find out which movie they want to watch next.

Step 1: Business Objective and Data Requirements

In this first step, you will define the business requirements:

1. Finding a movie for a particular user based on the watch pattern of that user.
2. Creating a cluster of users with a particular taste in movies requires a machine-learning model that can learn user behaviour.

Then you define your data requirements like this:

To search for movie collections based on users' habits, you need a movie dataset which contains information about the movies, like the movie genre,

the running time, the reviews, the ratings, the cast and the main character, etc. Also, you might need information about the users who have watched a particular movie. The movies and users' information together will help us make clusters of various types of users and the movies they like.

Step 2: Data Collection

After defining the type of data that needs to be collected, in the data collection step, you can go through various movie review sites like IMDb, Rotten Tomatoes, and Metacritic and collect information regarding the users who have reviewed movies and the metadata of the movies from all sorts of genres and regions. You will use various data mining techniques to collect movies from these resources.

Step 3: Exploratory data analysis

Until now, you have decided on the type of data you require and collected the data from various sources. The sample of this collected dataset looks like the image below.

	movieId	title	genres	userId	rating
66	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	185	4.0
68	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	191	4.0
184	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	534	4.0
308	2	Jumanji (1995)	Adventure|Children|Fantasy	534	4.5
463	6	Heat (1995)	Action|Crime|Thriller	191	4.0
727	10	GoldenEye (1995)	Action|Adventure|Thriller	534	4.0
895	16	Casino (1995)	Crime|Drama	191	4.0
976	17	Sense and Sensibility (1995)	Drama|Romance	191	5.0
1115	19	Ace Ventura: When Nature Calls (1995)	Comedy	534	4.0
1174	21	Get Shorty (1995)	Comedy|Crime|Thriller	191	4.0
1239	22	Copycat (1995)	Crime|Drama|Horror|Mystery|Thriller	99	4.0
1259	22	Copycat (1995)	Crime|Drama|Horror|Mystery|Thriller	429	4.0
1272	23	Assassins (1995)	Action|Crime|Thriller	99	4.0
1340	25	Leaving Las Vegas (1995)	Drama|Romance	191	5.0
1552	32	Twelve Monkeys (a.k.a. 12 Monkeys) (1995)	Mystery|Sci-Fi|Thriller	191	5.0

If the data is directly fed into a machine learning model, analyzing the output becomes challenging due to the presence of various parameters such as genre, title, and user ratings. The nature of these parameters and their correlations are unknown. Additionally, the dataset may contain missing

values and outliers, which can impact the training process. Therefore, it is crucial to address these issues before training a model based on the dataset. To gain a better understanding of the data before feeding it to a machine learning model, visualization techniques can be employed to evaluate and enhance the desired model's results. Several steps can be taken to familiarize ourselves with the data. This includes cleaning the dataset, transforming it into the desired format, and generating new features based on the available information. The process of exploring and summarizing the data using various visualization and statistical techniques is known as exploratory data analysis (EDA). EDA allows us to gain insights into the data, identify patterns, and make informed decisions on preprocessing steps and model selection.

Step 4: Modelling

After conducting Exploratory Data Analysis (EDA) on your dataset, you can gain confidence in the quality and suitability of your data for training a machine learning model. Once the model is trained on the given dataset, you can evaluate its performance using various clustering metrics. After obtaining the results, you can easily enhance the efficiency of the model based on your EDA findings. This is possible because during EDA, you have knowledge of which variables are important and which ones can be dropped to increase accuracy.

Process of performing EDA:

The following are the steps of EDA which you will work on in this series of lectures:

1. **Understanding the data**
2. **Studying various features based on descriptive statistics.**
3. **Analyzing and removing outliers.**
4. **Analyzing and treating missing values.**
5. **Transforming the data based on requirements.**
6. **Feature Scaling and Encoding to help the data better fit for machine learning models.**
7. **Selecting the most important features and creating new features to reduce the runtime for training a model and improve the model's accuracy.**