**Problem definition:**

## Problem definition depicts an explanation a project seeks to address but in a very concise manner. The problem definition gives us information regarding the current state, the future state the project desires and gaps if there is any. A good problem definition should be conclusive & specific.  A problem definition is a very important conveying process which provides an understanding regarding “what the problem is needed to be addressed” and “why is it important”. Problem definition helps identifying the goals of a project and draft the span of the project. It also helpful for assisting the activities and conclusions of those are working on the project. It also helps majorly a business or organization grow by pointing out the problems it is going through and also indicating the result it is hoping to address in a straight forward and to the point manner.

## Key elements of a problem definition: We need to keep few things in mind whenever we are formulating a problem definition. Hence those are:

## Apt Situation: this explains what is the ideal situation without the problem which is needed to be addressed.

## Reality: This displays the current scenario and talks about the problem, explain why it’s a problem and identify whom the problem is having an impact.

## End result: This part expresses the consequences and issues caused due to the problem.

## Proposal: The particular section of the problem definition proposes several possible solutions to the issue.

## ****Data Analysis****

Data analysis can be explained as the process of giving “order, structure, and meaning” to the data which has been collected. Data analysis aims at unearthing patterns or regularities by exploring, organizing, transforming, observing and modelling the data collected.

Describing, exhibiting, and evaluating the data, are some practical ways to apply statistical techniques. These are important tools in making meaningful insights, reach at conclusions, and support the decision making process. This is also a process of ordering, summarizing data which in turn helps, also to get answers to hypothetical questions. Another side of data analysis is Exploratory data. It helps in understanding the relationships between the variables found in between the data.

**There are five types of data analysis:**

* Descriptive Analysis
* Diagnostic Analysis
* Predictive Analysis
* Prescriptive Analysis
* Cognitive Analysis

**Descriptive analysis:** Descriptive Analysis, as the name suggests, mainly focusses on the data**. The foundation step** simply talks about the past data and everything that has happened in the past.

### **Diagnostic Analysis: Why did it happen?**

After the identification, of what has happened, the next step in the process would be to find the answer to the basic question – why this. Diagnostic analysis helps in digging the truth further by creating a descriptive, informative and interactive dashboards help in finding answers. It identifies the root cause of the problem and separates the source of the patterns.

### **Predictive Analysis: What is most likely to happen in the future?**

After the RCA of the problem is identified and the causal relationship between the variables is understood, the next we would want to know if there is any possibility that event is about to happen again? Predictive analysis is all about getting these facts. It predicts the occurrence probability of any event and forecasting the risk and segmenting customers into groups.

### **Prescriptive Analysis: How to make it happen?**

Prescriptive analysis is result-oriented. It analyses the learnings from the basic questions - what, why with what might happen to help in understanding what is needed to better the primary business metrics. It helps in understanding the best course of action and strategies. Prescriptive analysis is not about one particular, singular event but a collection of events that might happen in the future, using techniques like - simulation and optimization

### **Cognitive Analysis: Mimicking the human brain to carry out tasks**

This is an advanced type of analysis which basically works at mimicking the human brain to perform tasks just as a human does. It works on basics from technologies such as artificial intelligence, semantics, machine learning, and deep learning algorithms.

**EDA Concluding Remark**

EDA which is known as exploratory data analysis is a process of describing [data sets](https://en.wikipedia.org/wiki/Data_set) to point out the important characteristics, by using various [statistical tools and graphics](https://en.wikipedia.org/wiki/Statistical_graphics) and other [data visualization](https://en.wikipedia.org/wiki/Data_visualization) methods.

Primarily EDA is a visualizing method which is used for observing what can be deduced out of the data apart from the normal modelling or hypothesis testing task.

Patterns and insights are drawn by using data visualization for EDA. EDA also involves cleansing of datasets by handling irregularities in the data

EDA is a very important process. If it is not done properly then it will jeopardize the further steps in the machine learning model building process. If the EDA is carried out properly then it will increase efficacy of the dataset.

## Parts of EDA :

## Data Sourcing: Data Sourcing is defined as the process of finding and loading the data into the system for analysis. this can be done in 2 ways.

* + Private Data
  + Public Data
    - 1. **Private Data:** As the name suggests, private data is provided by private organizations. There are some security and privacy aspects to it. This kind of data is used majorly for internal analysis in organizations.
      2. **Public Data:** This is the kind of Data which is available to everyone. Such data can be found in government websites, public organizations etc. Anyone can access public data, and no special permissions or approvals are required for the same.

## Data Cleaning: After Data Sourcing, the next step in EDA is **Data Cleaning**. It is essential to clean the data and get rid of the irregularities, after sourcing it into system. Irregularities can be of different types within data -

* Missing Values
* Incorrect Format
* Incorrect Headers
* Anomalies/Outlier

## Univariate Analysis: If we observe data over a single variable/column from a dataset, it is termed as Univariate Analysis.

* **Categorical Unordered Univariate Analysis:** A categorical variable that has no defined order, is an unordered variable.
* **Categorical Ordered Univariate Analysis:** The variables that have a natural rank of order are Ordered variables.

## Bivariate Analysis: When we study data by focussing on two variables/columns into consideration from a dataset, it is called Bivariate Analysis.

## Numeric-Numeric Analysis: The study of two numeric variables from any dataset is called numeric-numeric analysis. We can understand it in three different ways.

## ****Numeric - Categorical Analysis:**** A numeric-categorical analysis includes Analysing a numerical variable and a categorical variable from a dataset. It is analysed by them using mean, median, and box plots.

## ****Categorical — Categorical Analysis****: In this case the target column is categorical and the ultimate answer will be binary as in yes/no or 0/1.

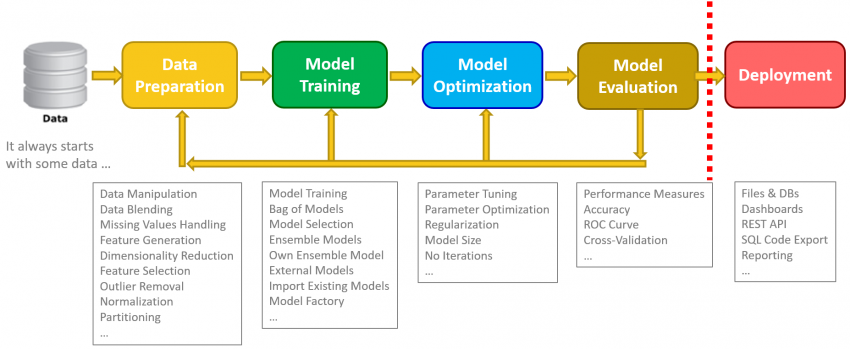
* **Multivariate Analysis:** Multivariate analysis is analysing data by taking into account more than two variables/columns into consideration to analyse from a dataset.

**Pre-Processing Pipeline**

pre-processing pipeline is a principal step in machine learning to submit highly precise, exact and perceptive results. if the quality of data is greater, the reliability of the produced results also increases. the inherent nature of real-world datasets are the **Incomplete, noisy, and inconsistent data**. it helps in improving the quality of data by handling missing values, smoothing noise, and resolving inconsistencies.

**Different stages of pre-processing pipeline are:**

* **Data Cleaning**
* **Data Integration**
* **Data Transformation**
* **Data Reduction**
* **Data cleaning: it** imputes missing values, smooths out noises, resolves inconsistencies, removes outliers in the dataset.it clears out the impurities of the dataset.
* **Data integration: Data integration** contributes by integrating the data from a multitude of sources into a single data warehouse. It is similar to collating multiple data which belong to same issue into a single one.
* **Data transformations: Data transformations** is referred as data normalization. Data transformations are carried out to improve the accuracies and efficiency of machine learning algorithms.
* **Data reduction: Data reduction** is referred to as reducing the dataset size by dropping out redundant features, features not required, or features creating high variance etc. Feature selection and feature extraction techniques can be used for solving the purpose.



**Building machine learning models**

To build a proper and effective machine learning model we have to proceed in a stepwise manner after performing apt EDA and pre processing pipeline. This is categorized under data preparation.

Model building can be done in three stepwise processes effectively.

* **Model Training:** the phase where data scientists try to fit the best combination of weights and bias to a machine learning algorithm to minimize a loss function over the prediction range in the data science development lifecycle is termed as model training. The main objective of model training is to create the best fitting mathematical representation to establish the relationship between data features and a target label (in supervised learning) or among the features themselves (unsupervised learning).
* **Model optimization:** The process of adjusting hyperparameters in order to minimize the cost function by using one of the optimization techniques is termed as model optimization. One of the important tasks is to minimize the cost function because it elaborates the discrepancy between the true value of the estimated parameter and the predicted model.
* **Model evaluation**: Model Evaluation is an important part of the model development process. It contributes in finding the best model which represents our data and also tells how well the chosen model will work in the future. Evaluating a model’s performance by using the data which is used for training purposes and is not acceptable in data science as it can facilitate generating overoptimistic and overfitted models. So, there are two ways for solving the purpose of evaluating models in data science. Those are Hold-Out and Cross-Validation. To avoid overfitting, both processes use a test set (not seen by the model) for evaluation of model performance.

**Conclusion:** Data, data science and analytics are an integral part of our lives and are used on a daily basis in our everyday lives to help businesses grow, drive efficiencies, tackle inefficiencies, glean deeper operational sights and finally generating more and more revenue. Though the impact data science makes is far beyond only helping businesses and is helping the mankind by solving many pressing issues. Data science is the new age technology which is minimizing the time, reducing manpower requirement, solving pressing issues and enhancing the output of the results.