Data Analysis Framework

Tools and Technology

You have understood *what data analysis is* and *how it is performed*. Let's try to understand which tools are available to perform the activity.

Business Problem Understanding, Framing Problems in Data Analytics terms, and **Validation of Business Problem Assumptions** mostly requires analysts to discuss with **stakeholders, subject matter experts (SMEs)** and **domain experts.**

Data Acquisition: Mostly, ETL (Extract transform and load) tools like **SSIS, Informatica, Alteryx, Tableau** etc., are used. These tools help fetch data from multiple sources like databases and flat files in various formats. There might be a vast amount of data present in these sources, but only the required data must be extracted and merged.

Data Preparation: This is one of the essential steps in the process to ensure the success of analysis, as rightly there are various tools available to perform the same. A few important ones are **ETL, SAS, Python, R, and Spark.**

Data Analysis: This is the core of the process and is mostly performed in an ad hoc way which is helpful to do in **Python, R, SAS, Excel** and using various visualization tools like **Power BI, Tableau, Spotfire** etc., by creating multiple plots and dashboards.

Storytelling/Analysis Presentation: This is the end result of the analysis and needs to be presented in a way easily understood by the stakeholders. Over here, you can use **Google Slides**, **PowerPoint** presentations and there are tools like **tableau**, which aids the storytelling by streamlining the flow of information by arranging various points.

Terms & Definitions

Take a moment and review these essential terms and definitions.

Database: A collection of data stored in a computer system

Data Engineering: Data engineering refers to the building of systems to enable the collection and usage of data.

Feature: A distinctive attribute or independent variable in the dataset. (e.g., If you have a dataset of House prices then the size of the house, locality, and number of bedrooms might be some of the features). Note that the variable you are predicting (the price of the house in this scenario) is not a feature, this will be taught in detail in future machine-learning modules.

Outlier: An outlier is a data point that differs significantly from other observations.

Missing Value: When no data value is stored for a variable in the dataset, then missing data or missing values occur. These are common occurrences and can have a significant effect on the conclusions that can be drawn from the data.

Stakeholders: People who invest time and resources into a project and are interested in its outcome.

Subject Matter Expert: The subject matter expert (SMEs) provides the knowledge and expertise in a specific subject, business area, or technical area for a project or program.