**Project brief**

*Problem statement:*

When looking at large amounts of data that we want to use in machine learning problems, these data need to be cleaned and checked thoroughly before they are utilised in the machine learning process. The data must be analysed, checked for missing values, anomalies and so on. In most machine learning problem, data is always front and centre and is the most important.

Exploratory data analysis (EDA) is a method used to quickly build a relationship with new data. It consists of checking for missing values, anomalies, using visual and quantitative methods to get an idea of the story the data is trying to tell.

The business wants this process to be automatic whenever the business is dealing with new data, the process of EDA should be automatic, instead of manually going through each step.

We are required to build a web application for this purpose alone. The user should be able to enter a dataset and then select any steps of EDA that he/she wishes to run on the dataset. Some of the steps should be mandatory. The application should be able to take any dataset and run the necessary EDA steps, providing a refined dataset at the end of it.

Exploratory Data Analysis: steps

A brief description of all steps used in EDA.

1. Variable identification (Categorical/Numerical)
2. Identifying all data types in dataset
3. Analysing the basic metrics - size, shape, and a description
4. Non-graphical univariate analysis - checking for unique values, checking for null values
5. Graphical Univariate analysis –

* Histograms - check distribution of data
* Boxplot - checking for anomalies
* count plots - checking the distribution of categorical variables

1. Bi-variate analysis – To find out relationships between two variables.

Combinations:

* Categorical and Continuous – Z-test/T-test
* Categorical and categorical – Chi-square
* Continuous and continuous – Correlation matrix, Variance Inflation factor

1. Variable transformation– standardization
2. Variable creation – dummy variables, derived variables

The web application will be created using the python language. We will be using streamlit to deploy the web app.