**Exploratory Data Analysis**

**Introduction:**

First step in any analytical project is to understand data using exploratory data analysis. Once you understand data then you can drive data according to the problem statement. Exploratory data analysis is an approach for summarizing and visualizing the important characteristics of a data set. Exploratory data analysis focuses on exploring data to understand the data’s underlying structure and variables, to develop intuition about the data set, to consider how that data set came into existence, and to decide how it can be investigated with more formal statistical methods. To accomplish this task (Understand the data), we have graphical and non-graphical tools. Graphical tools includes all visualizations like scatter plot for correlation, boxplot for outlier analysis, histogram for distribution, bar plot for missing value analysis, heat map for locations. Non-Graphical tools include all central statistics and statistical techniques/test.

**Objective**:

1. Maximize insight into a data set
2. Uncover underlying structure
3. Extract important variables
4. Detect outliers and anomalies
5. Test underlying assumptions
6. Develop parsimonious models and
7. Determine optimal factor settings.

**Problems we generally face in the data**

* Column with wrong name
* Rows with missing values
* Runtime column has units
* Revenue in multiple scales
* Wrong file format

Below are the steps involved to understand, clean and prepare your data for building your predictive model:

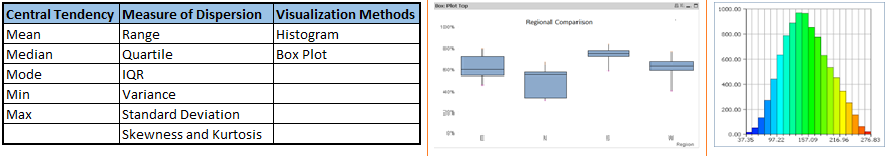
1. Variable Identification
2. Univariate Analysis
3. Bi-variate Analysis
4. Missing values treatment
5. Outlier treatment
6. Variable transformation
7. Variable creation

Few steps we will discuss here

### Univariate Analysis:

At this stage, we explore variables one by one. Method to perform uni-variate analysis will depend on whether the variable type is categorical or continuous. Let’s look at these methods and statistical measures for categorical and continuous variables individually:

Continuous Variables: In case of continuous variables, we need to understand the central tendency and spread of the variable. These are measured using various statistical metrics visualization methods as shown below:



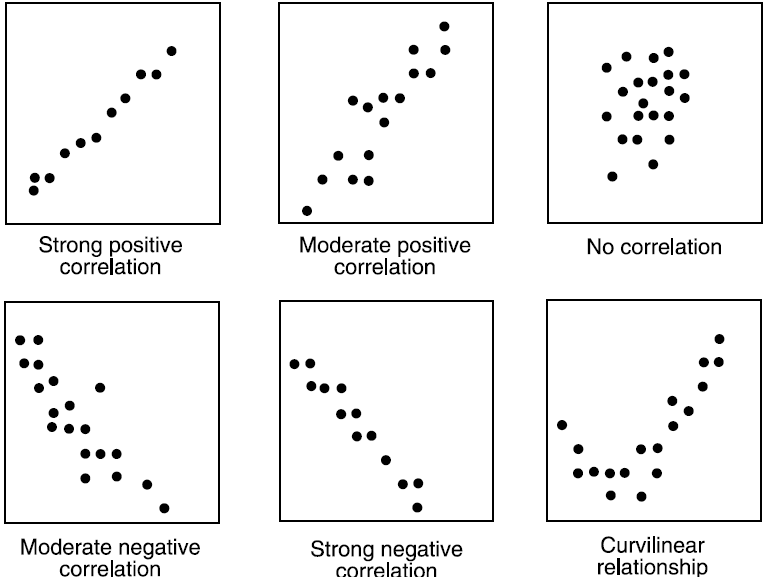
Categorical Variables:For categorical variables, we’ll use frequency table to understand distribution of each category. We can also read as percentage of values under each category. It can be be measured using two metrics, Count and Count% against each category. Bar chart can be used as visualization.

**Bi-variate Analysis:**

Bi-variate Analysis finds out the relationship between two variables. Here, we look for association and disassociation between variables at a pre-defined significance level. We can perform bi-variate analysis for any combination of categorical and continuous variables. The combination can be: Categorical & Categorical, Categorical & Continuous and Continuous & Continuous. Different methods are used to tackle these combinations during analysis process.

Let’s understand the possible combinations in detail:

1. Continuous & Continuous: While doing bi-variate analysis between two continuous variables, we should look at scatter plot. It is a nifty way to find out the relationship between two variables. The pattern of scatter plot indicates the relationship between variables. The relationship can be linear or non-linear.



Scatter plot shows the relationship between two variable but does not indicates the strength of relationship amongst them. To find the strength of the relationship, we use Correlation. Correlation varies between -1 and +1.

* -1: perfect negative linear correlation
* +1: Perfect positive linear correlation and
* 0: No correlation

1. Categorical & Categorical:To find the relationship between two categorical variables, we can use following methods:

* **Two-way table:** We can start analyzing the relationship by creating a two-way table of count and count%. The rows represent the category of one variable and the columns represent the categories of the other variable. We show count or count% of observations available in each combination of row and column categories.
* **Chi-Square Test:** This test is used to derive the statistical significance of relationship between the variables. Also, it tests whether the evidence in the sample is strong enough to generalize that the relationship for a larger population as well. Chi-square is based on the difference between the expected and observed frequencies in one or more categories in the two-way table. It returns probability for the computed chi-square distribution with the degree of freedom.

Probability of 0: It indicates that both categorical variable are dependent

Probability of 1: It shows that both variables are independent.

Probability less than 0.05: It indicates that the relationship between the variables is significant at 95% confidence.

The chi-square test statistic for a test of independence of two categorical variables is found by:



Where O represents the observed frequency. E is the expected frequency   
 **Interview Questions:**

1. **Explain how data is aggregated in R?**
2. Explain difference between inner join and left outer join?
3. **What is the function used for adding datasets in R?**
4. **What is the use of subset() function and sample() function in R ?**

This will be more like programming questions.