**EDA or** Data Exploration and Preparation

**Univariate analysis:**

Analyzing each variable individually.

* Continuous variable: mean, median, min, max, range, variance, sd.
  + Plots: Histogram, Box plot
* Categorical variable: Count, count %
  + Plots: bar charts

**Bivariate analysis:**

Find the relationship between two variables

1. Continuous & Continuous: Correlation, scatter plots.
2. Categorical & Categorical: Two way table, Chi-square test

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.

1. Categorical & Continuous:

**Z-Test/ T-Test:** Either test assess whether mean of two groups are statistically different from each other or not.

**ANOVA:** It assesses whether the average of more than two groups is statistically different.

**Missing value treatment:**

If the data is huge and the proportion of missing values is less then delete the observations.

Treat them using appropriate values: Mean, Median, count, or using KNN imputation.

**Outlier detection:** Outlier is an observation that appears far away and diverges from an overall pattern in a sample.

Box plots are the best to detect outliers.

***Treat outliers:*** delete the observation or treat it (replace it by minimum or maximum)

**Feature Engineering**

There are two steps: Variable transformation and Variable creation.

* Transformation refers to the replacement of a variable by a function. For instance, replacing a variable x by the square / cube root or logarithm x is a transformation.
  + When should we use Variable Transformation?
    - Scale (when all the variables are of different scale)
    - If there is a complex non-linear relationship in the data and you want a linear relationship.
    - If the data is Skewed. Symantec distribution is preferred over Skewed.
    - Common techniques include: Log, square/ cube roots, binning
* Creation refers to creating new variables from the existing once. Eg: Name to first name and second name.

# **What is a hypothesis test?**

A hypothesis test is a statistical test that is used to determine whether there is enough evidence in a sample of data to infer that a certain condition is true for the entire population.

The null hypothesis is the statement being tested. The alternative hypothesis is the statement you want to be able to conclude is true.

You use a p-value, to make the determination. If the p-value is less than level of significance (0.05), which is a cut-off point that you define you accept the null hypothesis. If it’s greater than 0.05 there is enough evidence (data) to support the alternative hypothesis.