# Life Insurance dataset

Sunday, December 19, 2021 2:13 PM

EDA -

1) **Numerical - 10 - create a separate data frame for all the features**
    a. Sanity check on the data - Anomalies in the data
    b. Univariate Analysis - pick up one variable at a time
        i. Overall Summary
        ii. Age - histogram, boxplot - distribution, outliers
    c. Outliers Treatment -
        i. IQR method - interquartile range
        ii. z-score - +-3 cutoff
    d. Missing Values -
        i. Treat missing values -
            1) Imputation - Median/Mean/**sklearn - KNN imputation**/mean/median
    e. Bi-variate Analysis
        i. X to X - relationship between input features
            1) Patterns - scatter plots, heatmaps - correlation, VIF (multicollinearity), pair plot (numerical to numerical) - Findings
            2) Numerical to Categorical - Bar plots (X- category, Y-numerical), Boxplots (X-categorical, x-numerical) - Findings
        ii. X to Y -relationship between X & Y
            1) Scatter (X-numerical, Y -numerical) - feature importance , correlation - Findings

2) Categorical - 8
    a. Sanity check on the data - Anomalies in the data
    b. Missing Values
        i. Treat missing values - Mode, Unknown category, Frequency based
    c. Univariate Analysis -
        i. Count plots, pie chart - merging the category
    d. Bi-variate Analysis
        i. X&X - stacked bar
        ii. Box plot plots - Y & X relationship - freelancer
        iii. X&X relationship as well

3) Transformation -
    a. Normalization/Standardization (x-min/max-min), (x-u/std.)
    b. Yes

4) Feature engineering