**Exploratory Data Analysis -**

1. Overall Data Distribution, shape
2. Target Feature - **Expected CTC** 
   1. Histogram plot - **Yes/No**
   2. Check the distribution - skewed or not? - \*\***Insights\*\* (we will tackle in the modelling Notes-2)**
   3. Pair plot with all the numerical features
   4. Separate the categorical and numerical features
   5. **Categorical -** 
      1. Univariate Analysis -
         1. Hygiene checks in the data
            1. a)  Check the categories of the columns - more than expected categories like – Grad/undergrad (convert it into two categories)
            2. b)  Suppose city - 60% - Delhi, 1% Ahmedabad - try to merge for later purpose
         2. **Missing values -** Treat them with a method - Mode/Max freq/KNN imputer from sklearn/Unknown
         3. Check that - "?"/special characters - value counts on each of the categorical - if you can run a loop
         4. Create a few count plots to show freq - run a loop to get all the plots in 1 go - \*\***Insights\*\***
      2. **Bi-variate Analysis -** 
         1. Categorical to categorical (X1 v/s X2) - **stack bar plot**
         2. Categorical to numerical (X1\_cat v/s X2\_num) - bar plot/swarm/violin/box - \*\***Insights\*\***
         3. Categorical to Target Feature (X1\_cat v/s Target conversion) – bar plot/swarm/violin/box Plot - \*\***Insights\*\***
   6. **Numerical -** 
      1. Univariate Analysis -

1)  Hygiene checks on the data

2)  Missing values - Mean/Median/**KNN imputer**/simple imputer

3)  Distribution and box plots with a loop - \*\***Insights\*\***

4)  Outliers - boxplot - IQR method/**percentile method (99%,95%)**

5)  Distribution and box plots with a loop - verify the outliers are removed - \*\***Insights\*\***

6)  Skewness in the data - right skewed - take a log else take a squareroot

* + 1. Bi-variate Analysis -
       1. **Correlation -** 
          1. a)  Correlation between (X1\_num v/s X2\_num) - heatmap - \*\***Insights\*\***
          2. b)  Scatter plots (X1\_num v/s X2\_num) - regplot - \*\***Insights\*\***
       2. Relation with target feature (X1\_num v/s Target) - BOX/Swarm/violin - \*\***Insights\*\***
       3. Relation with Categorical feature (X1\_num v/s X1\_cat) - BOX/Swarm/violin - \*\***Insights\*\***
  1. **Overall Pairplot -** Try to see the separation between the - creation the distribution plot with a hue of target - **Pair plot**