**Approach For The Model**

**Data Cleaning and Feature Engineering:**

1. We see that there are multiple rows for the same employee with different Total Business Values hence we will group the data on the basis of Emp\_ID and add up the Total Business Value to get single record for Each employee.
2. For the ones who have already left where separated from the training dataframe.
3. A new feature for tenurity was made for all the employees, taking present day as 2017-12-31 as the present date .
4. One more new dataframe is made with all the details in it for every unique employee and adding the total business value for each of the employee.
5. Both the dataframe is then merged together.
6. A new feature is made from the both the dataframe named Attrition with values – 1( those who have left the organisation) and 0(for the ones who have not left it) for analysing the results of clusters prediction.
7. Columns **MMM-YY,Dateofjoining\_y , LastWorkingDate** removed as new features where already made with these and MMM-YY was not taking into consideration as important for building the model.
8. Converted the data type to datetime for **Dateofjoining\_x** to fill the na values in the column **Tenurity** of the remaining employees who are still in the organisation taking 2017-31-12 as the present working date.
9. Created new feature Promoted based on existing feature of Designation and joining Designation.
10. New Feature for **Joining day, Joining month, Joining year** is created from the Dateofjoinin feature in the dataset.
11. Changed the datatype Tenurity and removed unwanted string values from each of the value.

**Data-Visualization:**

1 . To check for **Outliers** in the dataset boxplot was plotted for every continuous feature and found out it is present in the features **Salary** and **Total Business Value** and **Age.** Outliers were then treated with z-score method.

2. Displot was plotted for continuous feature to check if there is any Skewness present in the dataset or not, again it was found on the features, **Salary, Total Business Value and Age which is practically not possible as well as in the Tenurity.** Skewness was then removed using quantile method from these feature.

3.Plotted heatmap inorder to see the co-linearity amongst the features. There was hardly any Co-relation found amoungs the features.

**Data- Preprocessing:**

1 - Dropped off the columns Emp\_ID and attrition as after checking everyting are not found to be important.

2 - Removed the skewness from the features **Age, Salary, Total Business Value and Tenurity** with the help of quantile method.

3 – Removed outliers from **Total Business Value** feature with the help of z-score method.

4 – With the help of Label Encoder we performed label encoding on the feature Education\_Level as it Ordinal.

5 - With the help of get dummies method we performed encoding for the fetures Gender and City as these are nominal in nature.

7 - Scaled the features for building our model with the help of Normalize method.

**Model Building:**

We used Kmeans Clustering unsupervised algorithm for the predictions and chose n\_clusters 2 as we needed two clusters one for who will leave and one for we will remain.

**Predictions:**

For predictions test dataset with **Emp\_ID** was imported and all the features where extracted from the df2 that we had created from train dataset.

Values where then predicted using the model that we prepared after following the pre-processing steps which we performed on the train set.

The predicted results were added to the test dataset and Emp\_ID and the Target where extracted and then saved to local system for Submission.