CREDIT – EDA ASSIGNMENT

Introduction:

- The dataset which has the information regarding loans for various types of customers.
- Loans can be approved related to various categories like salaried, employed, business and so on .
- The companies which are providing loans are facing many problems to find the right customer who can repay within timeline.
- Now we will start with EDA analysis to filter out the data which can give us good knowledge about customers background

DATA CLEANING:

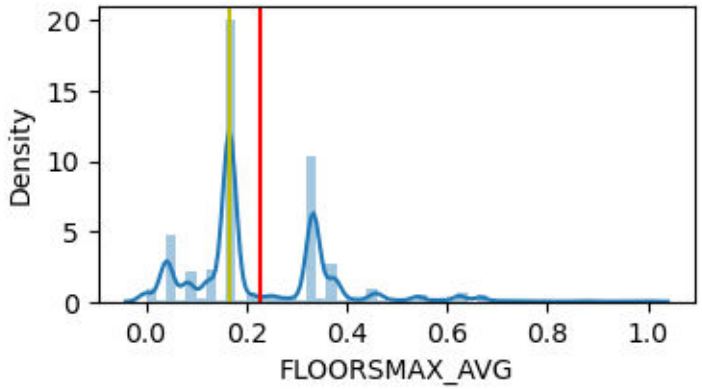
- Predict the proportion of missing data and apply proper methods like imputation,dropping ,assigning the values accordingly.
- Check the data type of each column and assign proper data type for each column .
- Removal of Irrelevant Data and assigning new columns using feature engineering concepts.
- Replace NaN with a Scalar Value and drop the Duplicates

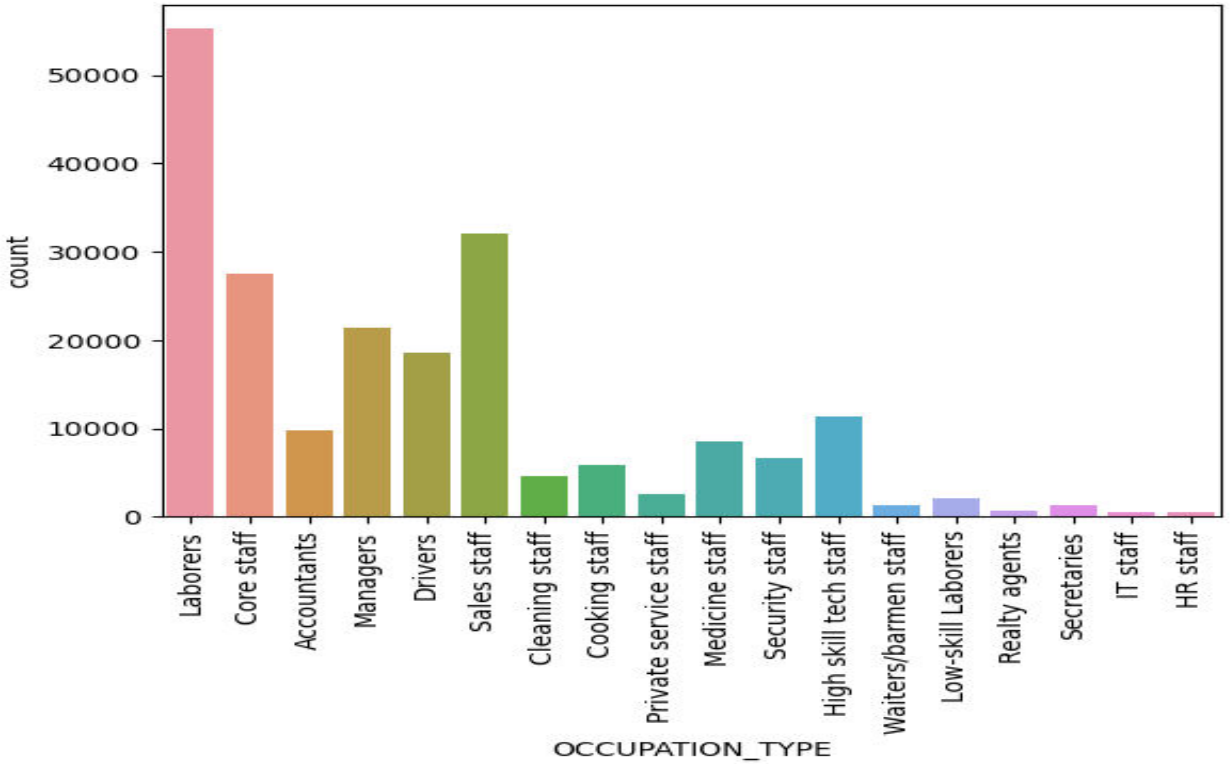
- We are plotting with mean and median so that we can have clear picture of filling the missing valuese either with mean or median.

- From the second graph ,the highest occupation is in laborers but its highly impossible to fill with highest value like mode.Either we have to leave the column or replace the NaNs with proper values.

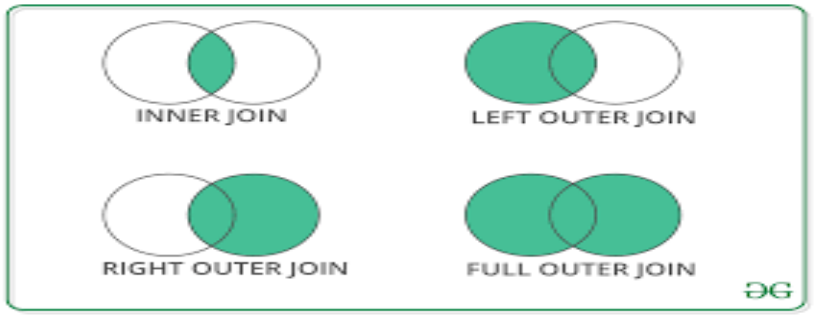
- Skipping unnecessary rows in the dataset and make sure that keep the columns which are required for further analysis

- Extracting more information from your dataset to get more variables and check the unique values of columns





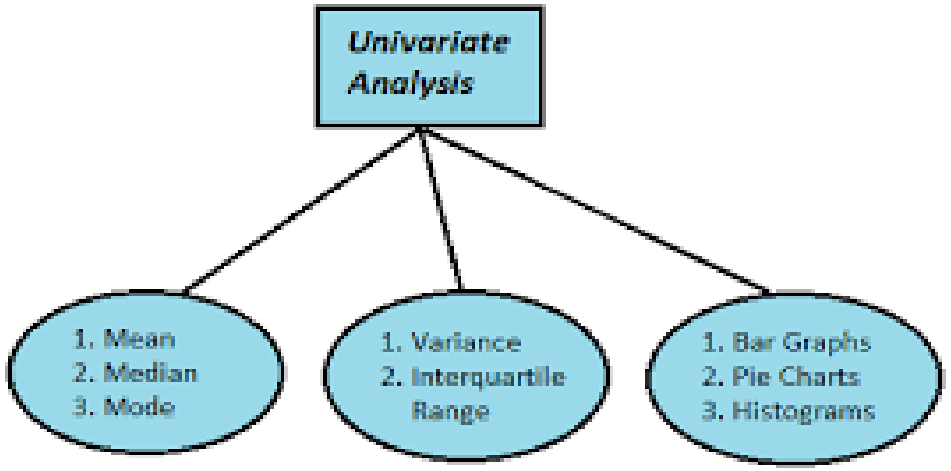
MERGING METHODOLOGY



- Merging data sets is to combine two or more data sets horizontally or vertically by matching the keys from both data sets.
- We have so many methods to merge data like vstack, hstack, concatenation,join methods and so on.
- Here in this dataset ,merge method is used with inner join concept .

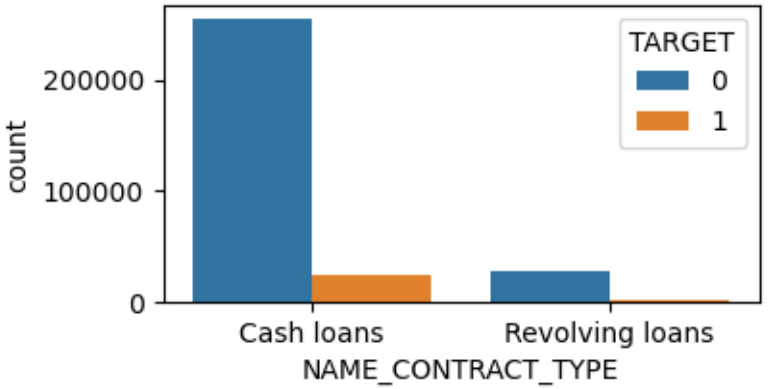
UNIVARIATE ANALYSIS

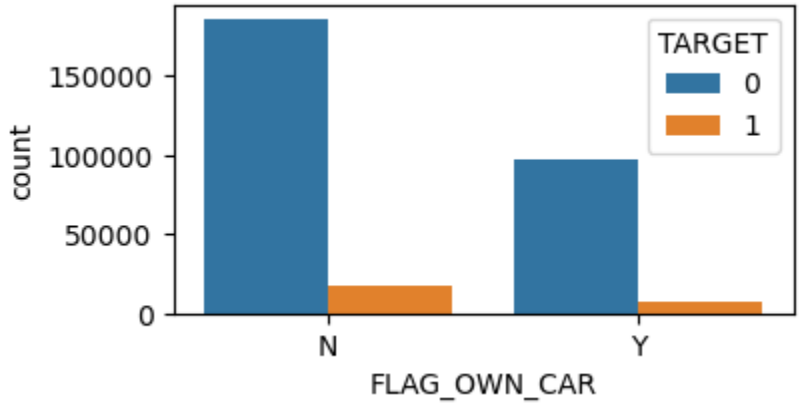
- Univariate Analysis is a type of data visualization where we visualize only a single variable at a time.
- Univariate Analysis helps us to analyze the distribution of the variable present in the data so that we can perform further analysis.



UNIVARIATE ANALYSIS USING PLOTS

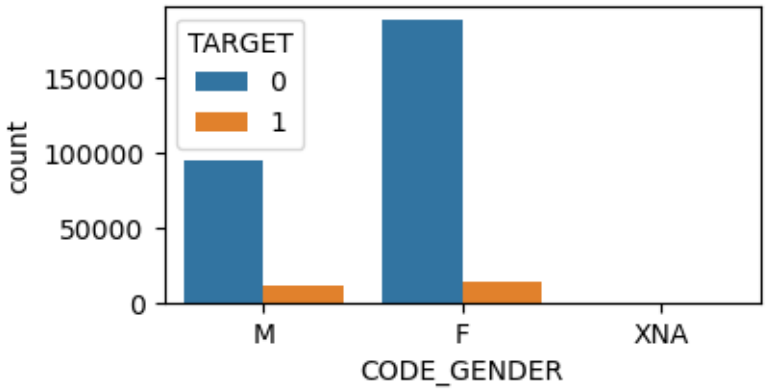
- For univariate analysis plotting can be done using single value and the data can be in the form of categorical,numerical and continuous.
- Typed of plots can be used frequently for the univariate analysis **are Histograms, Distplot and Boxplot.**
- Here we have plotted the customers having own car and types of loans
- By the graph cash loans are mostly accessed by customers,and customer who have own car does not go for loan

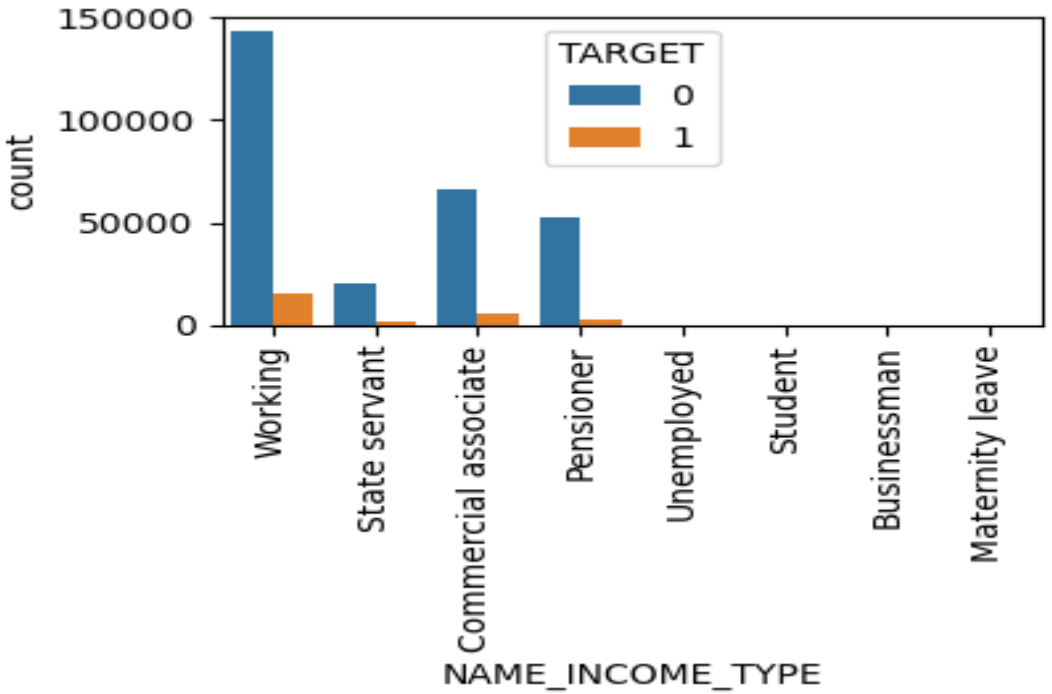




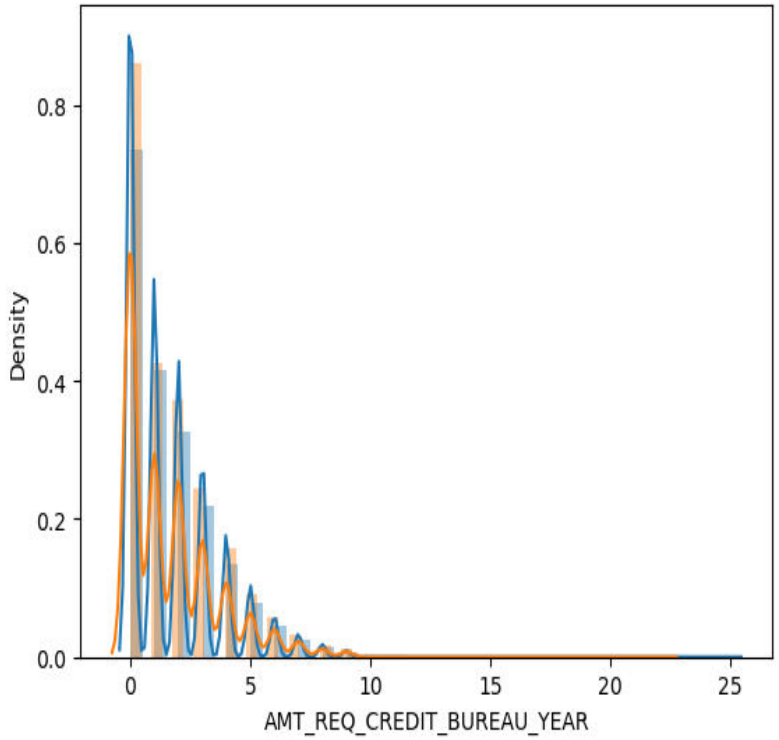
Gender and Income Type Plots

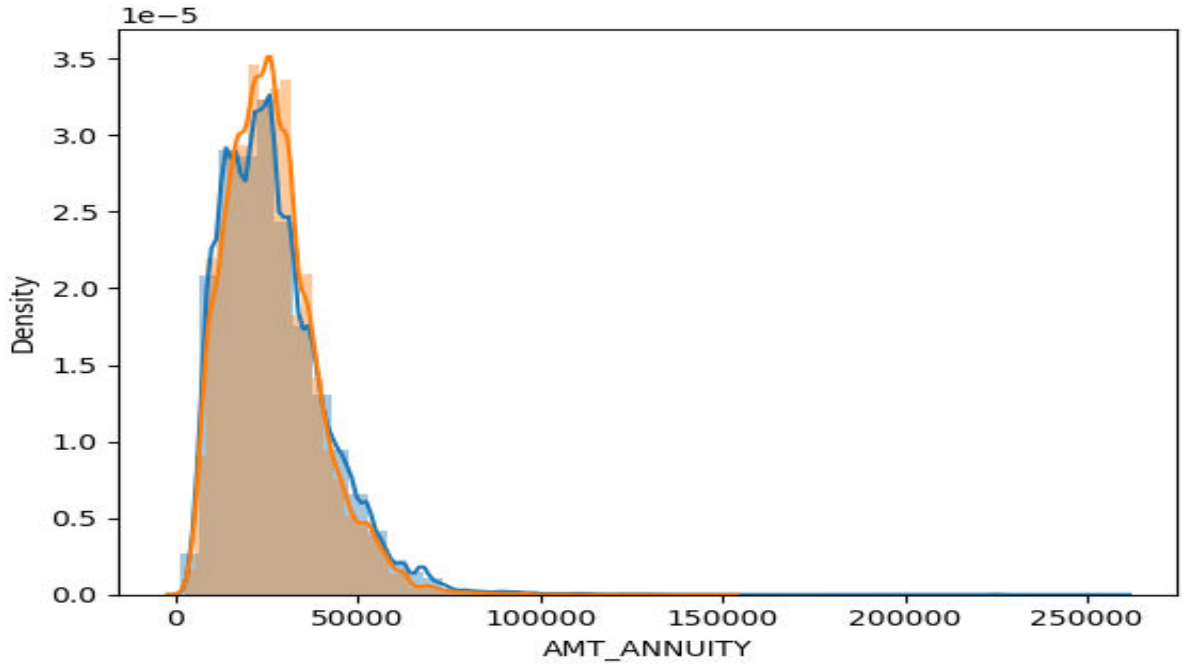
- From the graph we can predict that both male and female are in thr targets for approval of loans.
- Presently by the income type , mostly working professionals are mostly approved depending on their backgrounds.
- Govt servants,commercial ,pensioner are given preference in the next level.
- All the other customers they have not targeted as they may don't have regular income



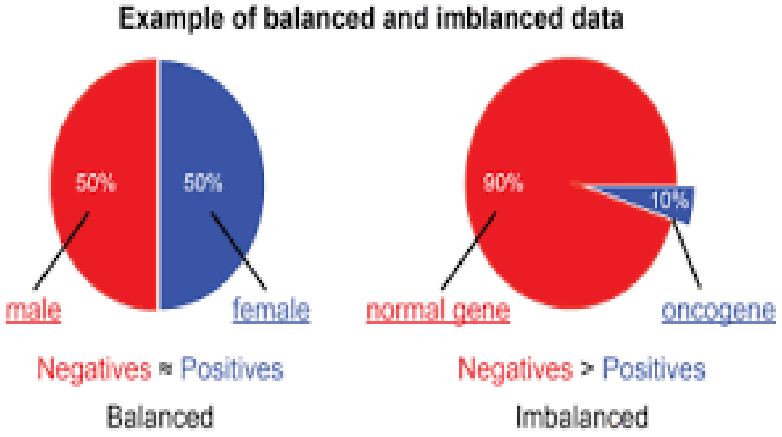


UNIVARIATE ANALYSIS FOR NUMERICAL DATA





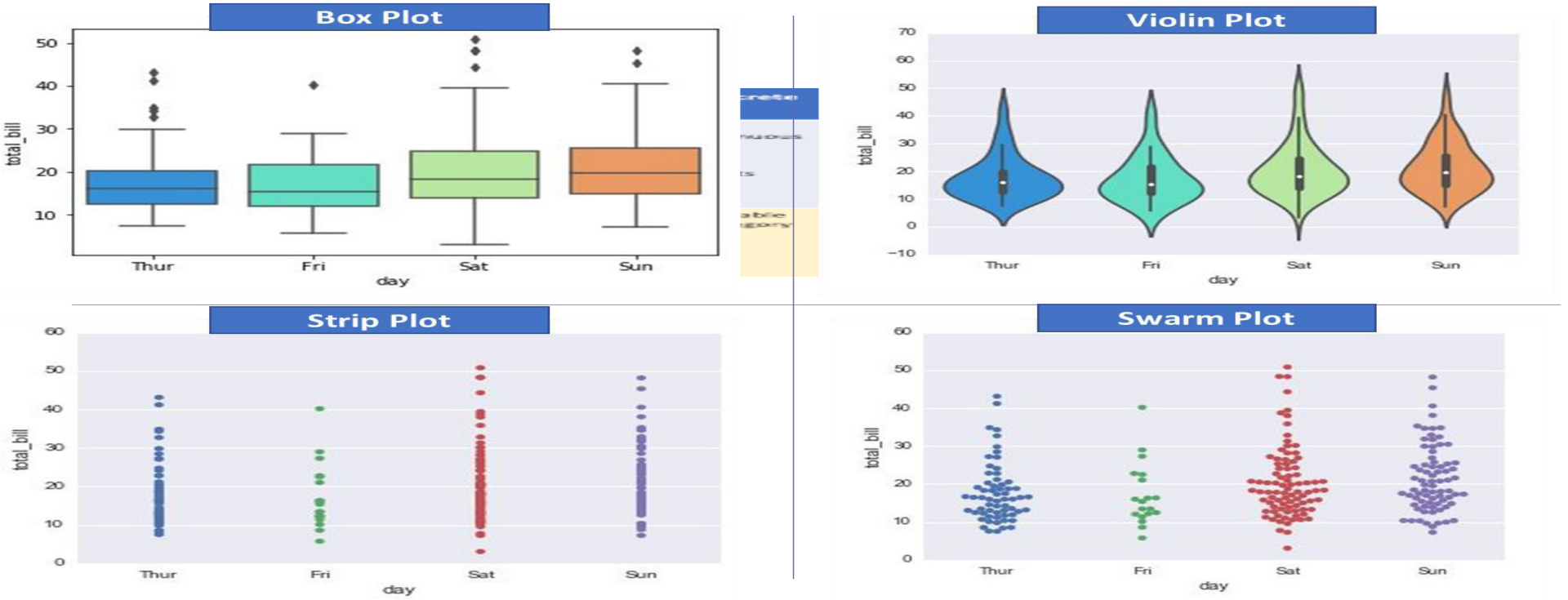
IMBALANCED DATA



- The ratio of the value counts of classes is much higher. Such data set is known as an imbalanced dataset.
- A widely adopted and most straight forward method for dealing the highly imbalanced datasets is called resampling.

BIVARIATE ANALYSIS

Bivariate analysis is the simultaneous analysis of two variables. It explores the concept of the relationship between two variable



Type of Variables (Vs.)	Categorical (incl. discrete numerical)	Continuous

Categorical (incl. discrete numerical)	• Frequency of the two categories/ other continuous variables' range
	  • Crosstab
	  • Heatmaps
	  • Stacked bar charts
		• Range of continuous variable with respect to each category
		  • Boxplots
		  • Violin plots
		  • Swam plots
		  • Count plots
		  • Bar plot

Continuous	• Range of continuous variable with respect to each category
	  • Boxplots
	  • Violin plots
	  • Swam plots
	  • Count plots
	  • Bar plot
		• How the increase or decrease in one variables changes with the other
		  • Scatterplot
		  • Line plots

TYPES OF BIVARIATE ANALYSIS

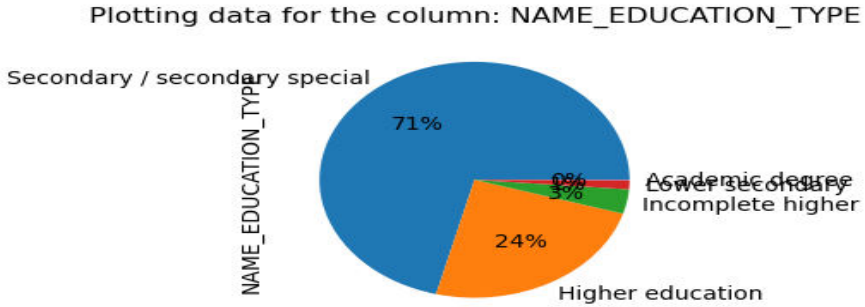
- The main three types we will see here are:

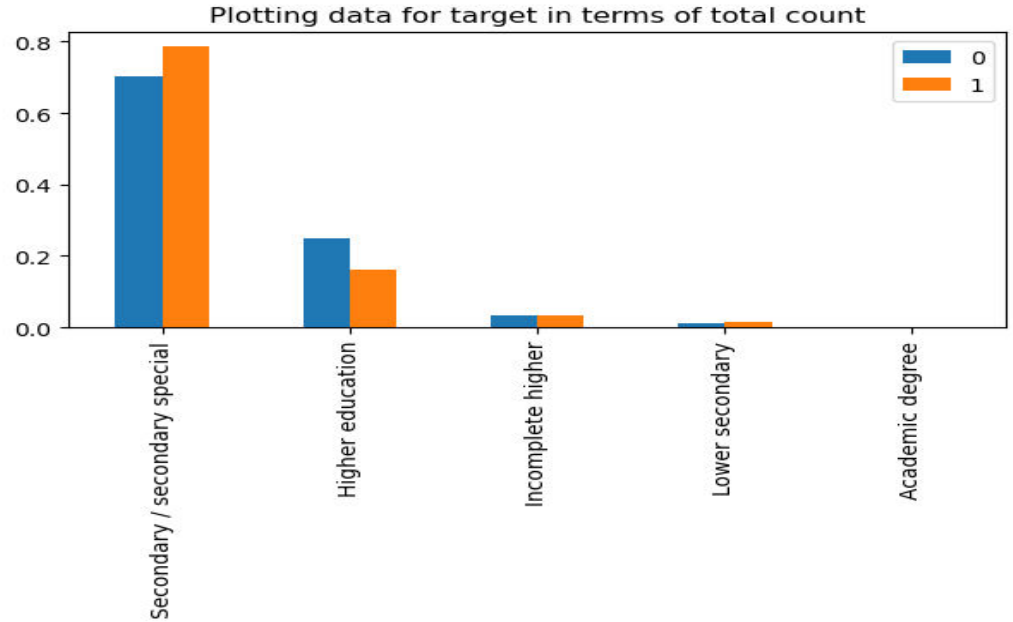
  Categorical v/s Numerical

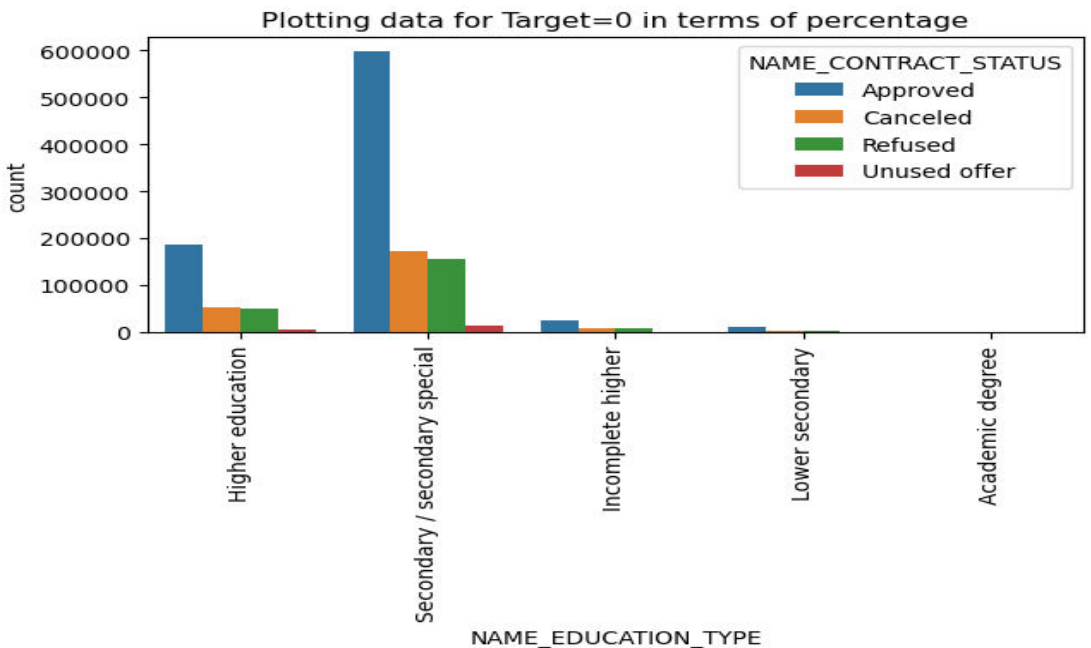
  Numerical V/s Numerical

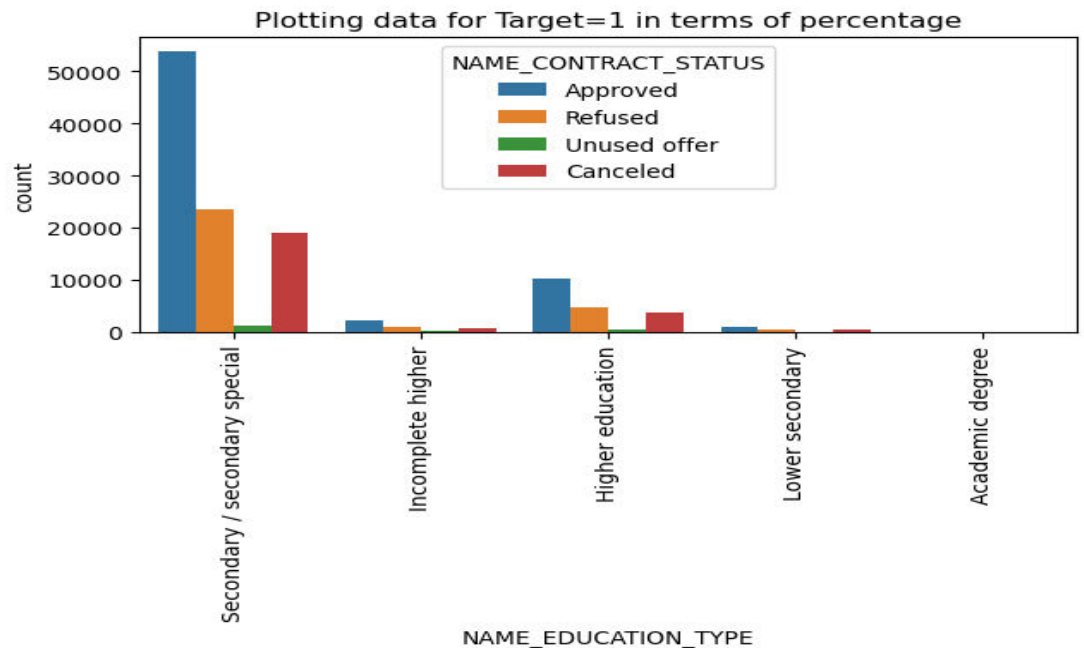
  Categorical V/s Categorical data

- A bivariate statistical test is a test that studies two variables and their relationships with one another.
- For example: Ice cream sales compared to the temperature that day. Traffic accidents along with the weather on a particular day.


Plotting data for the column: NAME_EDUCATION_TYPE
Secondary / secondary special
71%
NAME_EDUCATION_TYPE
24%
Higher education
Academic degree
Lower secondary
Incomplete higher

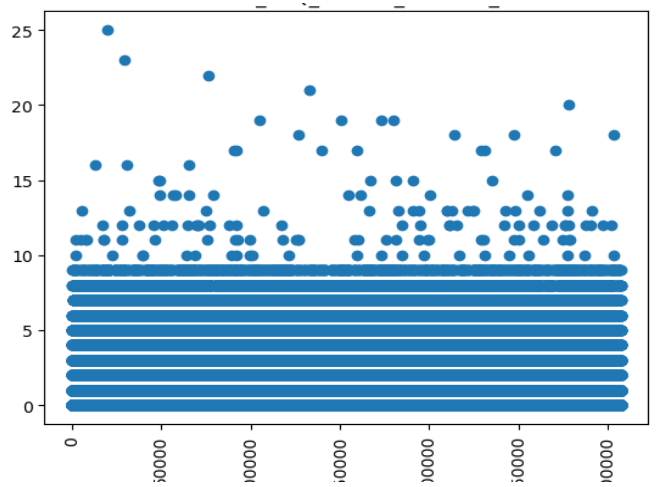

Plotting data for target in terms of total count
0
1
0.8
0.6
0.4
0.2
0.0
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree


Plotting data for Target=0 in terms of percentage
NAME_CONTRACT_STATUS
Approved
Canceled
Refused
Unused offer
600000
500000
400000
300000
200000
100000
0
count
Higher education
Secondary / secondary special
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE

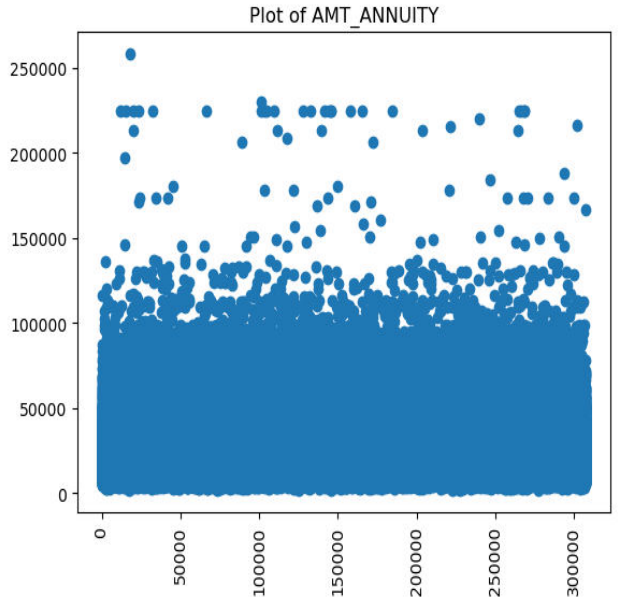

Plotting data for Target=1 in terms of percentage
NAME_CONTRACT_STATUS
Approved
Refused
Unused offer
Canceled
50000
40000
30000
20000
10000
0
count
Secondary / secondary special
Incomplete higher
Higher education
Lower secondary
Academic degree
NAME_EDUCATION_TYPE

OUTLIERS

- An outlier is a value in the data set that is extremely distinct from most of the other values.
- Outliers can be usually identified using scatter plot and boxplot.
- Outliers can be treated by Imputation,Deletion,Binning ,Cap .
- CORRELATION COEFFICIENT is highly sensitive to outliers. Since it measures the strength of a linear relationship between two variables.
- Outliers can be found using univariate,Bivariate and Multivariate analysis



PLOT_AMT_REQ_CREDIT_BUREAU



PLOT OF AMT_CREDIT