BANK LOAN CASE STUDY

BY - SUKHMANI ARORA

Final Application data -
https://docs.google.com/spreadsheets/d/12OYY691bFqlcRV4kv8s9MZc56UkE67yI/edit?usp=sharing&ouid=108310215373032149971&rtpof=true&sd=true

Final Previous application data -
https://docs.google.com/spreadsheets/d/1krui64sjOLOK3LO0blYoAH_oZu29kawi/edit?usp=sharing&ouid=108310215373032149971&rtpof=true&sd=true

Tableau Workbook -
https://public.tableau.com/views/BankLoanCaseStudy_16887984938020/NAMECONTRACTTYPEvsNameContractStatusvsAmtApplication?:language=en-US&:display_count=n&:origin=viz_share_link

PROBLEM STATEMENT

For this case study, Exploratory Data Analysis(EDA) is to be performed to analyze the loan application data of a finance company which gives various types of loans to urban customers. The company find it hard to provide loans due to the insufficient credit history of applicants. There are two risks involved for the company :

- The company will suffer business loss if they reject loans of customers who can repay.
- The company will suffer financial loss if they approve the loans of customers who cannot repay.

By performing analysis, the trends and insights is to be provided regarding the factors which should help in identifying the defaulters and non-defaulters from the given data so that the company can take appropriate action for every loan application in order to increase overall profit of the company.

APPROACH

Understanding data and data cleaning are the most important steps before doing the analysis.

- Application dataset contains 307511 rows and 122 columns.
- Previous application dataset contains 1048575 rows and 37 columns.

For data cleaning of both datasets, the irrelevant columns are removed and also, the columns containing more than 10% of missing values are removed. There were a lot of columns whose most of the values were "XNA" which means unknown, that are also removed.

After that, analysis is performed on both the datasets which include outlier analysis, data imbalance, univariate analysis, bivariate analysis, and correlation, and useful insights are obtained.

APPROACH

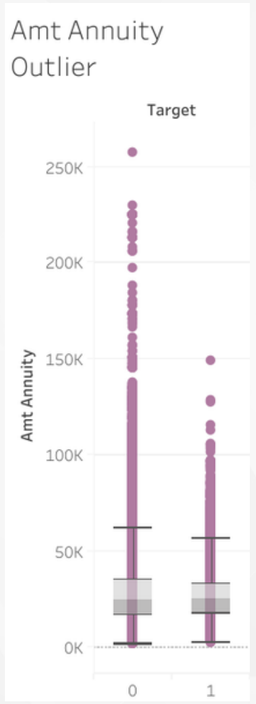
1	No of Missing column values	0	0	0	0	0	0	0	0	12	278
2	% of Missing column values	0	0	0	0	0	0	0	0	0.003902299	0.09040327
3											
4	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NA
5	100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Un
6	100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Fa
7	100004	0	Revolving loans	M	Y	Y	0	67500	135000	6750	135000	Un
8	100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Un
9	100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Un
10	100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Sp
11	100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Un
12	100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Un
13	100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Ch
14	100012	0	Revolving loans	M	N	Y	0	135000	405000	20250	405000	Un
15	100014	0	Cash loans	F	N	Y	1	112500	652500	21177	652500	Un
16	100015	0	Cash loans	F	N	Y	0	38419.155	148365	10678.5	135000	Ch
17	100016	0	Cash loans	F	N	Y	0	67500	80865	5881.5	67500	Un
18	100017	0	Cash loans	M	Y	N	1	225000	918468	28966.5	697500	Un
19	100018	0	Cash loans	F	N	Y	0	189000	773680.5	32778	679500	Un
20	100019	0	Cash loans	M	Y	Y	0	157500	299772	20160	247500	Fa
21	100020	0	Cash loans	M	N	N	0	108000	509602.5	26149.5	387000	Un
22	100021	0	Revolving loans	F	N	Y	1	81000	270000	13500	270000	Un
23	100022	0	Revolving loans	F	N	Y	0	112500	157500	7875	157500	Ot
24	100023	0	Cash loans	F	N	Y	1	90000	544491	17563.5	454500	Un
25	100024	0	Revolving loans	M	Y	Y	0	135000	427500	21375	427500	Un

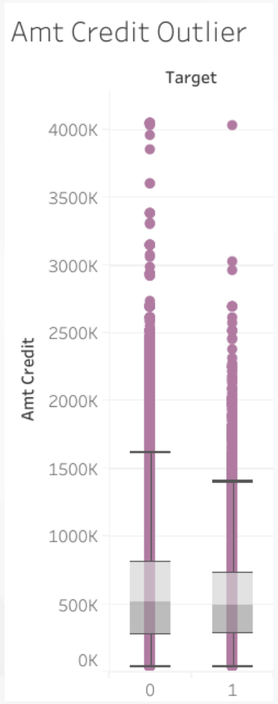
application_data | final data

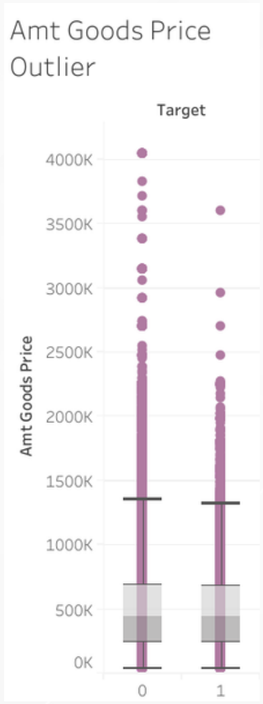
ANALYSIS

Analysis of Application dataset

OUTLIER ANALYSIS :







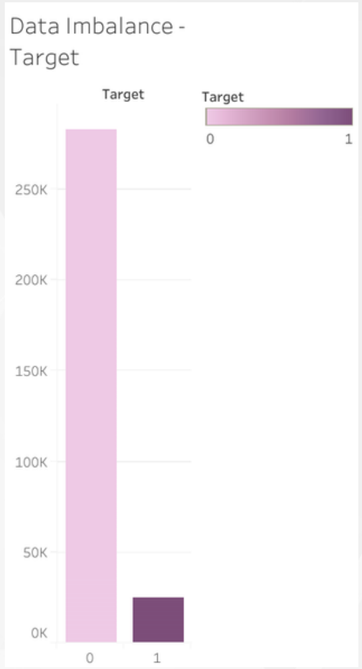
Firstly, outliers of all the numerical columns are analyzed by box plotting the data of these columns and the results are :

- All these columns have a very high number of outliers.
- The data in AMT_ANNUITY, AMT_CREDIT, and AMT_GOODS_PRICE columns is positively skewed.

ANALYSIS

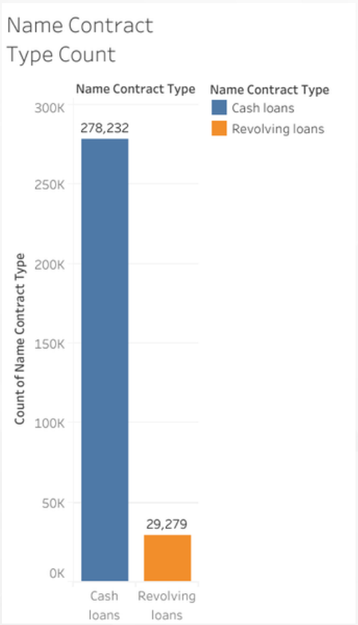
DATA IMBALANCE :

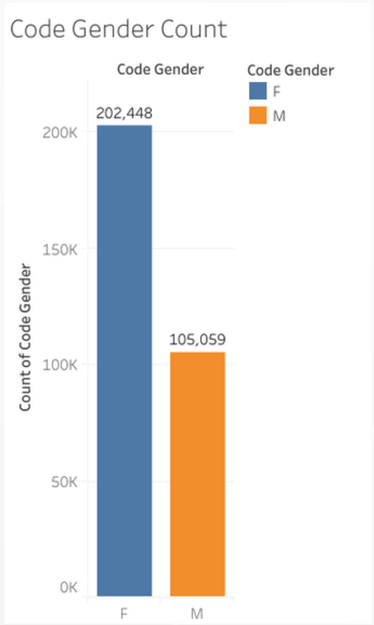
Data Imbalance shows the unequal distribution of different classes or categories in a variable of the dataset. Here, most of the applicants are non-defaulters in the data.

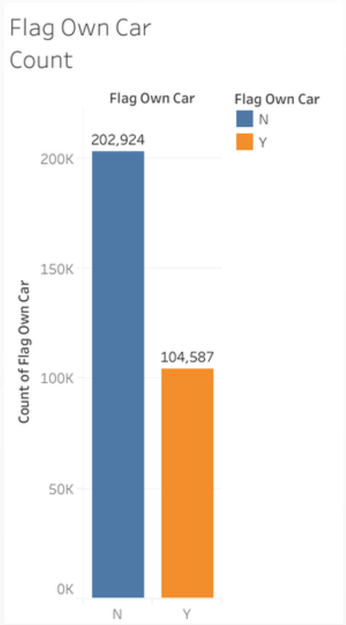


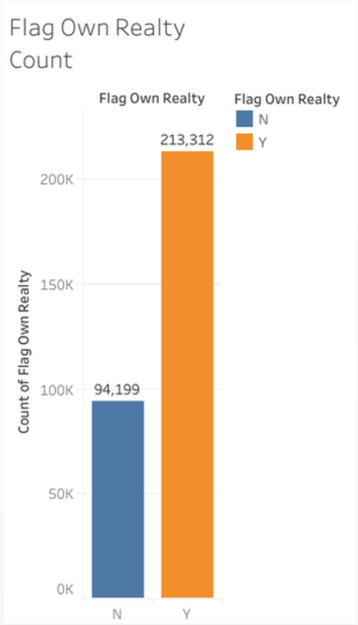
UNIVARIATE ANALYSIS :

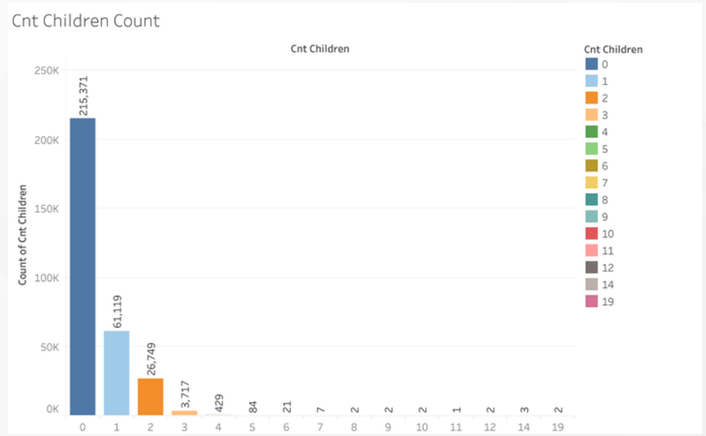
Univariate analysis is when only one variable is analyzed. We can easily see the data distribution pattern in these variables from these visualizations.





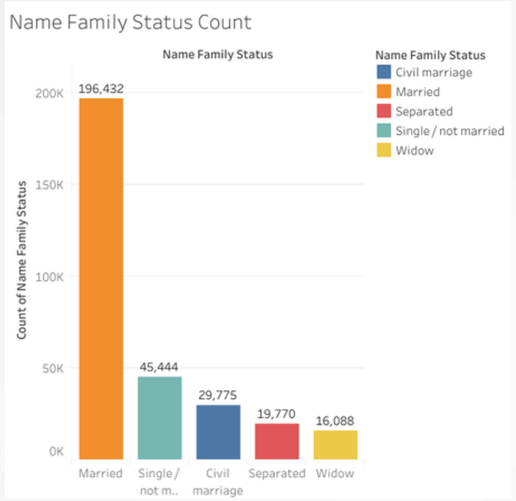


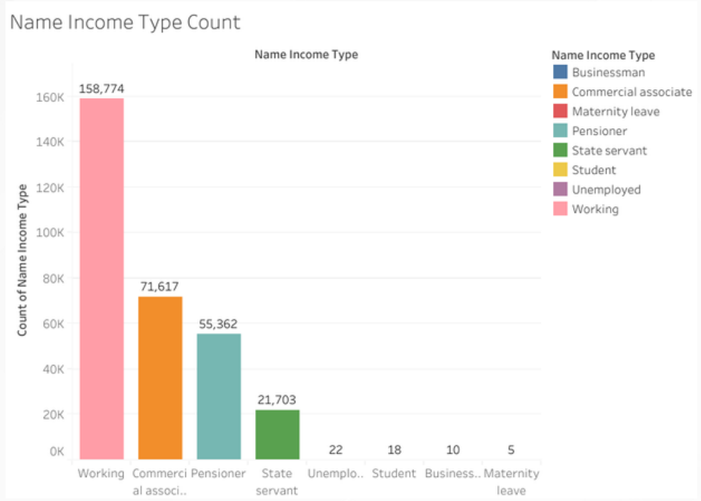


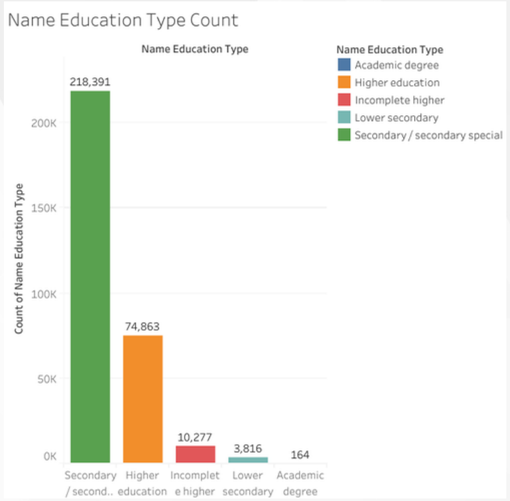


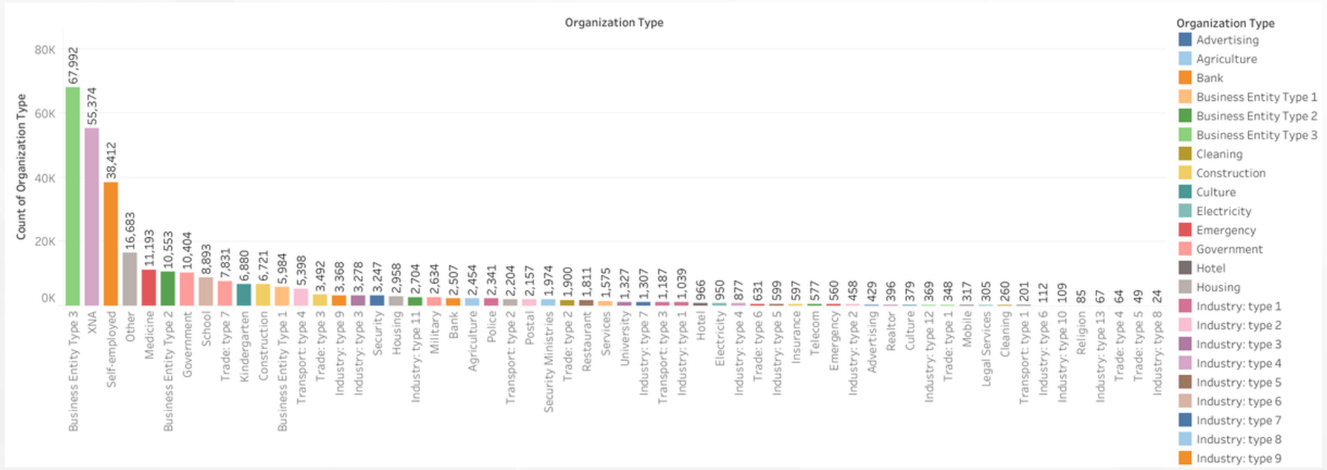
ANALYSIS

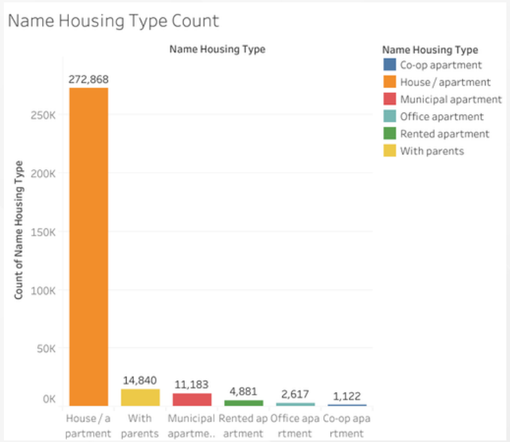
UNIVARIATE ANALYSIS :





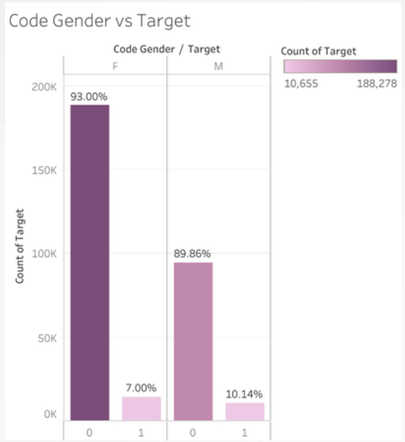


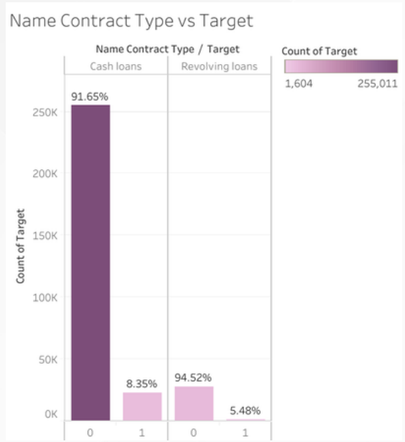


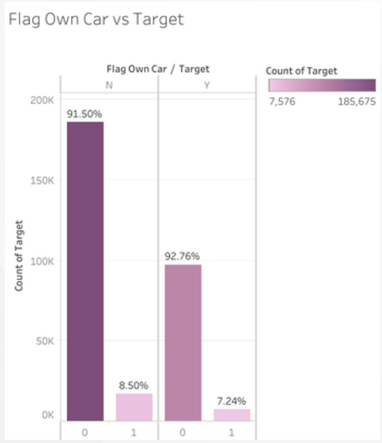


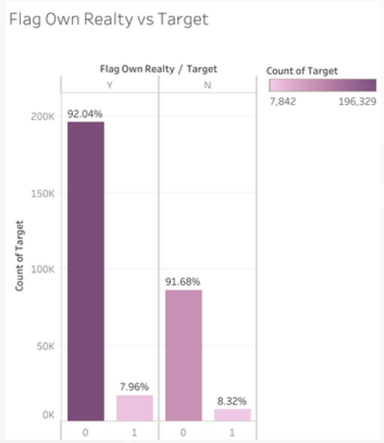
ANALYSIS

BIVARIATE ANALYSIS :







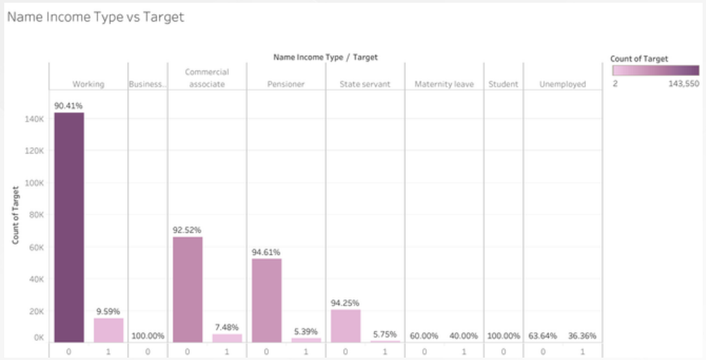


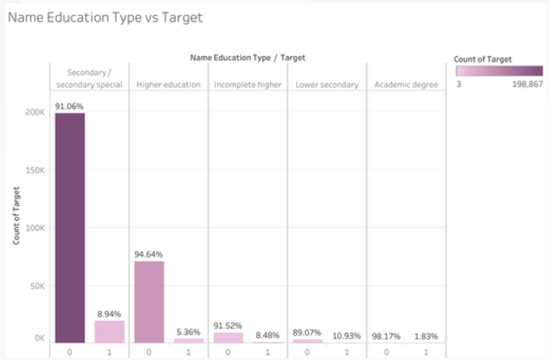
Bivariate data analysis is when exactly two variables are analyzed. Here, all the variables are analyzed with target variable.

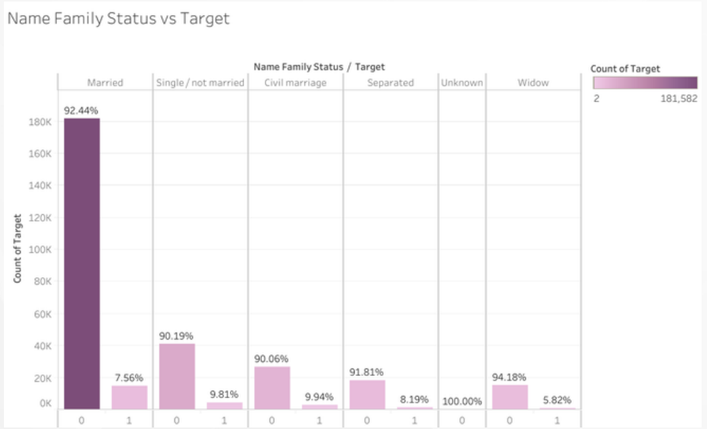
- CODE_GENDER : Females have applied for loans more than males but still defaulters have a high percentage of males.
- NAME_CONTRACT_TYPE : More people applied for cash loans rather than revolving loans.
- FLAG_OWN_CAR : Most of the applicants do not own a car.
- FLAG_OWN_REALTY : A lot of applicants own real estate property.

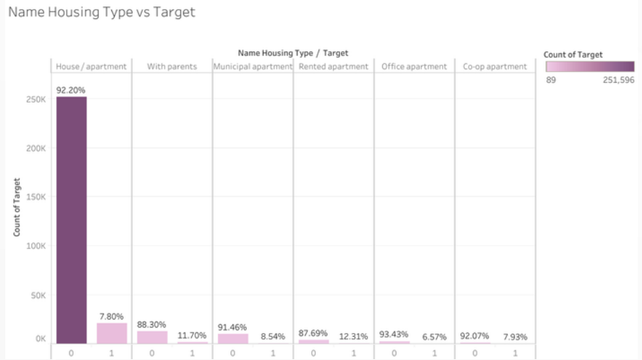
ANALYSIS

BIVARIATE ANALYSIS :





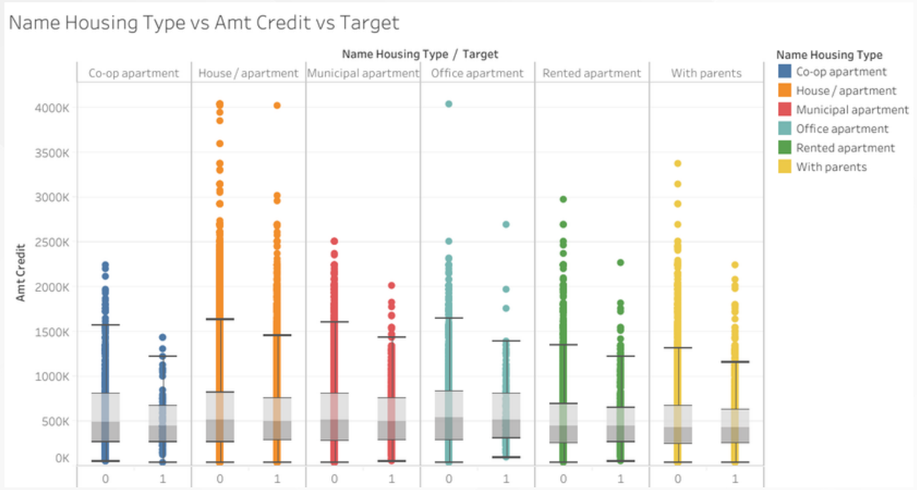




- NAME_INCOME_TYPE : People in the working category are high in both non-defaulters and non-defaulters.
- NAME_EDUCATION_TYPE : The people who are doing secondary/secondary special education are high and also has high percentage of defaulters.
- NAME_FAMILY_STATUS : Most of the applicants are married but single people have high percentage of defaulters.
- NAME_HOUSING_TYPE : People who own a house/apartment are most likely to apply for loans and are high in both non-defaulters and non-defaulters.

ANALYSIS

MULTIVARIATE ANALYSIS :



NAME_HOUSING_TYPE vs AMT_CREDIT vs TARGET

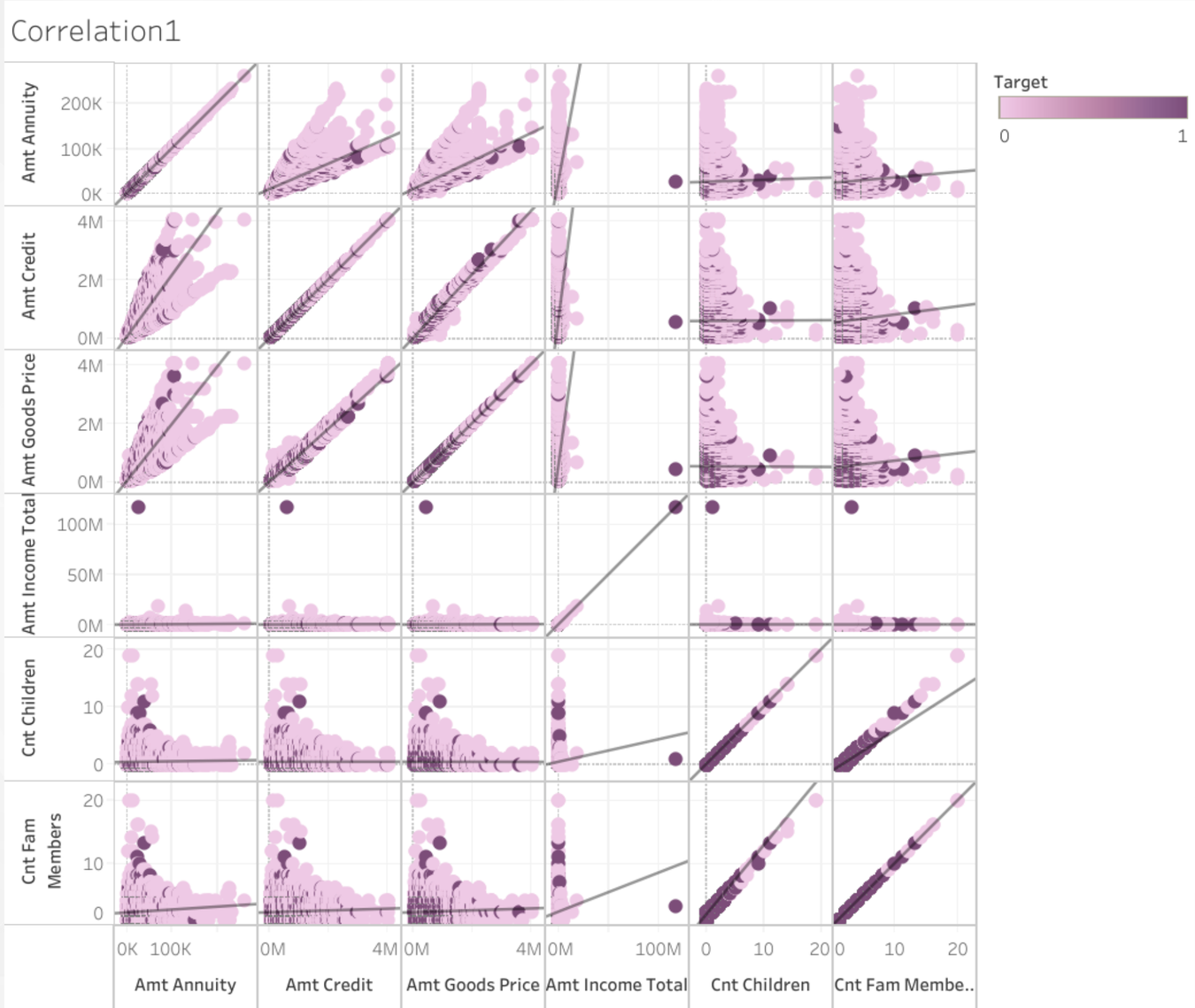
People who own a house/apartment have applied for very high credit loans and have a high number of defaulters as well.

NAME_EDUCATION_TYPE vs AMT_CREDIT vs TARGET

People who are doing higher education have applied for very high credit loans. People with academic degree are very less in defaulters.

ANALYSIS

CORRELATION ANALYSIS :



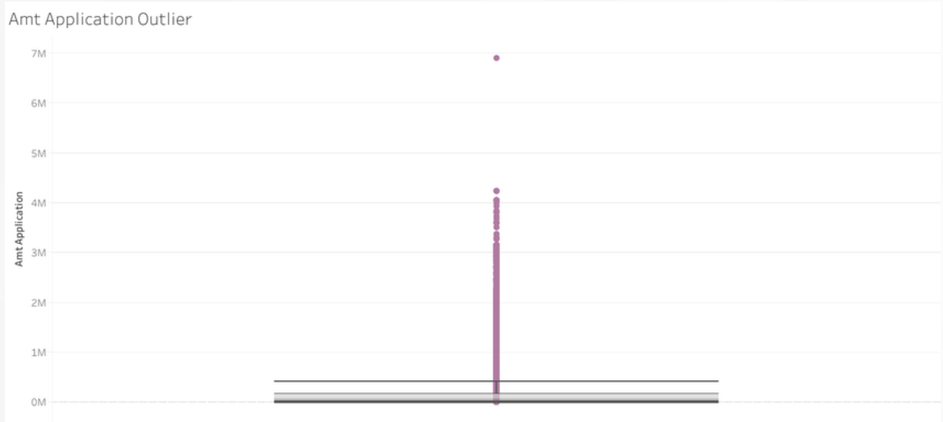
Here, the correlation analysis is performed on different numerical columns and trend lines are created to observe the relationship. The trend line which shows the R-squared value closer to 1 has a higher correlation. The variables with the highest correlations are :

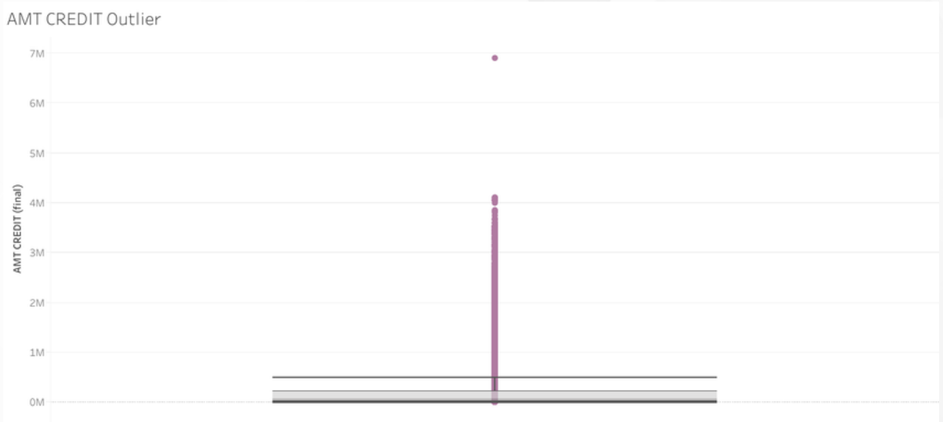
- AMT_GOODS_PRICE and AMT_CREDIT
- AMT_GOODS_PRICE and AMT_ANNUITY
- AMT_ANNUITY and AMT_CREDIT

ANALYSIS

Analysis of Previous Application dataset

OUTLIER ANALYSIS :



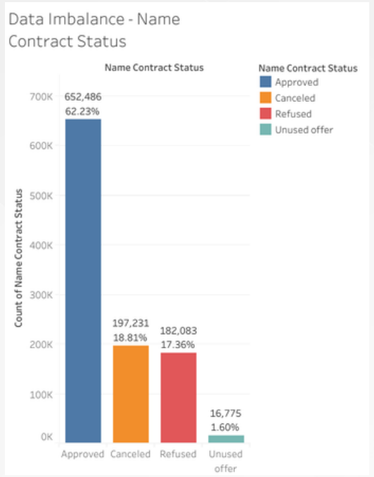


Firstly, outliers of all the numerical columns are analyzed by box plotting the data of these columns and the results are :

- Both columns, AMT_APPLICATION and AMT_CREDIT have high number of outliers.

ANALYSIS

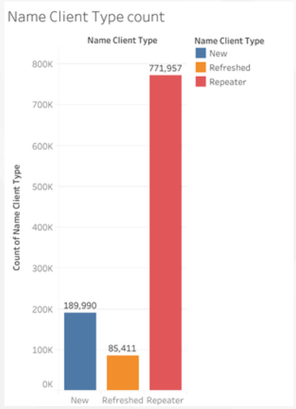
DATA IMBALANCE :

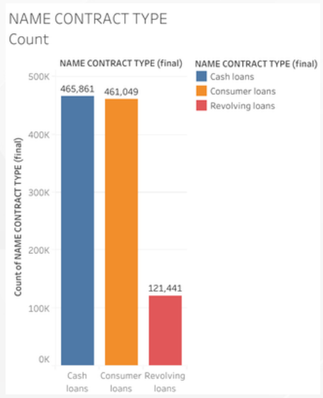


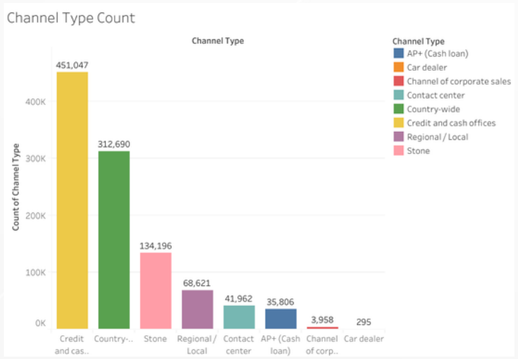
Data Imbalance shows the unequal distribution of different classes or categories in a variable of the dataset.

Here, we can see that most of the loans are approved which is about 62.23%, and refused and canceled categories 17.36% and 18.81% respectively and only 1.60% of total applicants are in the unused offer category.

UNIVARIATE ANALYSIS :



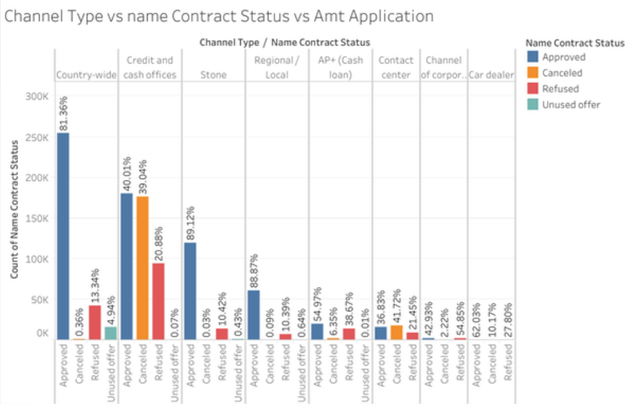


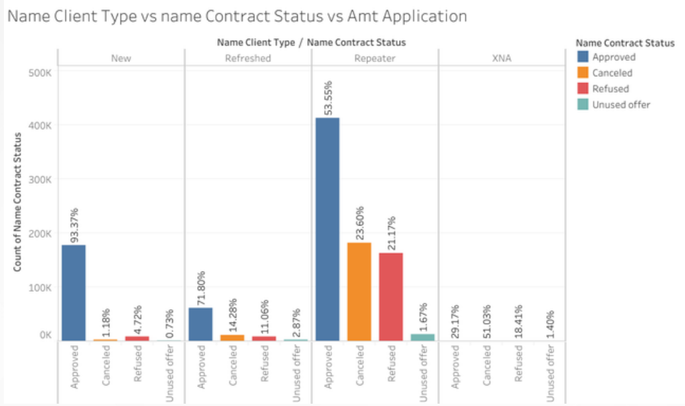


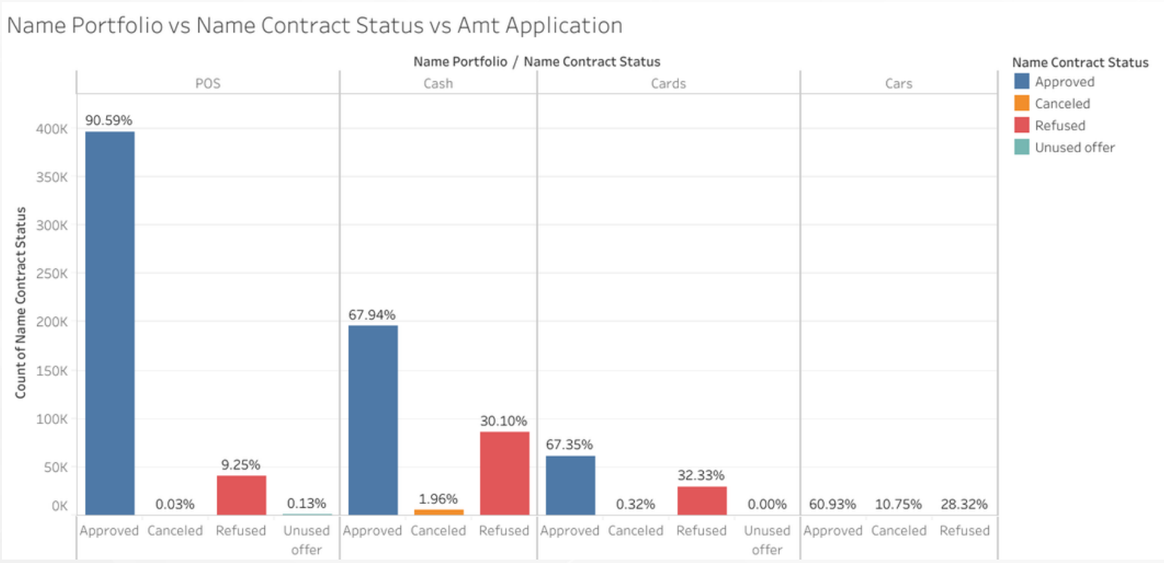
- Repeater clients are very high in number.
- Cash loans and consumer loans are higher than revolving loans.
- Most of the applications are with channel type cash and credit offices.

ANALYSIS

BIVARIATE ANALYSIS :







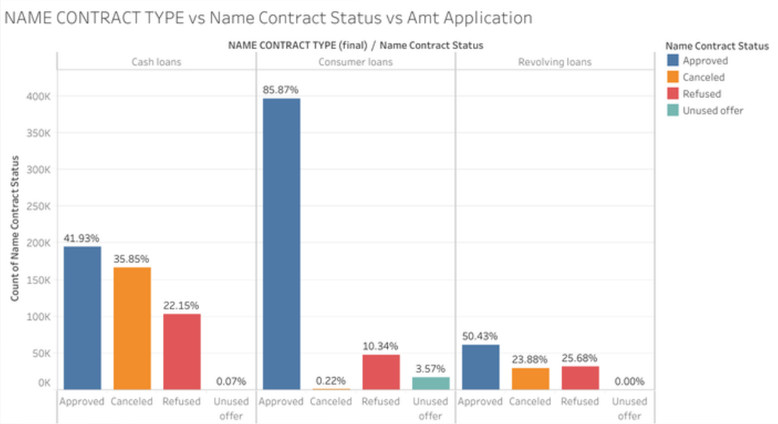
- CHANNEL_TYPE vs NAME_CONTRACT_STATUS: Country-wide has both the highest approval and refusal rate and credit and cash offices have a high canceled rate.

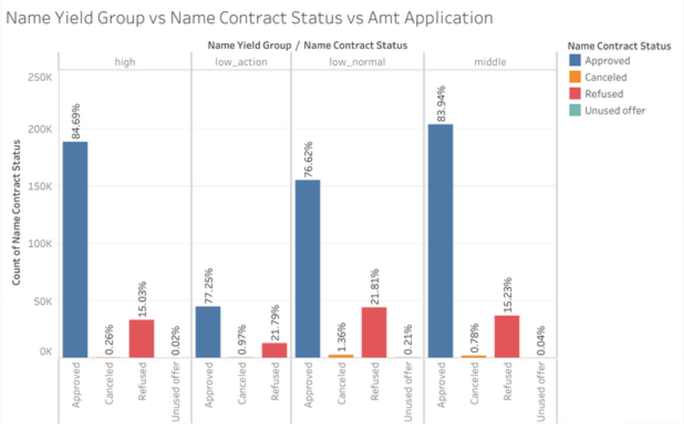
- NAME_CLIENT_TYPE vs NAME_CONTRACT_STATUS : Most of the repeaters' applications are either refused or canceled and the applications of new applicants are mostly approved.

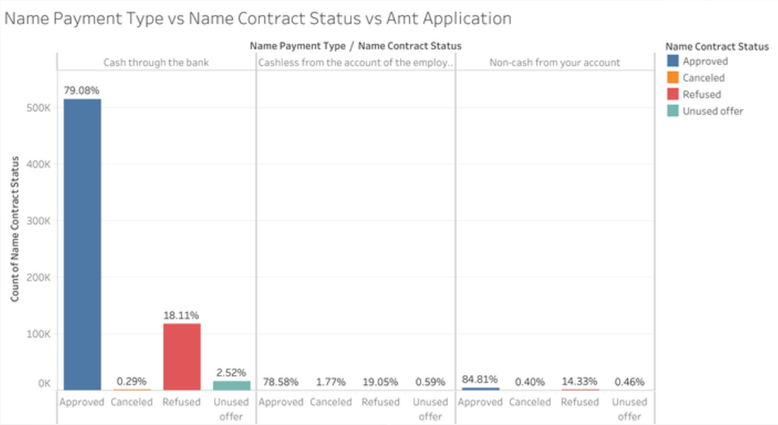
- NAME_PORTFOLIO vs NAME_CONTRACT_STATUS : Loans of people in the POS category are mostly approved, it has a very low cancellation rate.

ANALYSIS

BIVARIATE ANALYSIS :



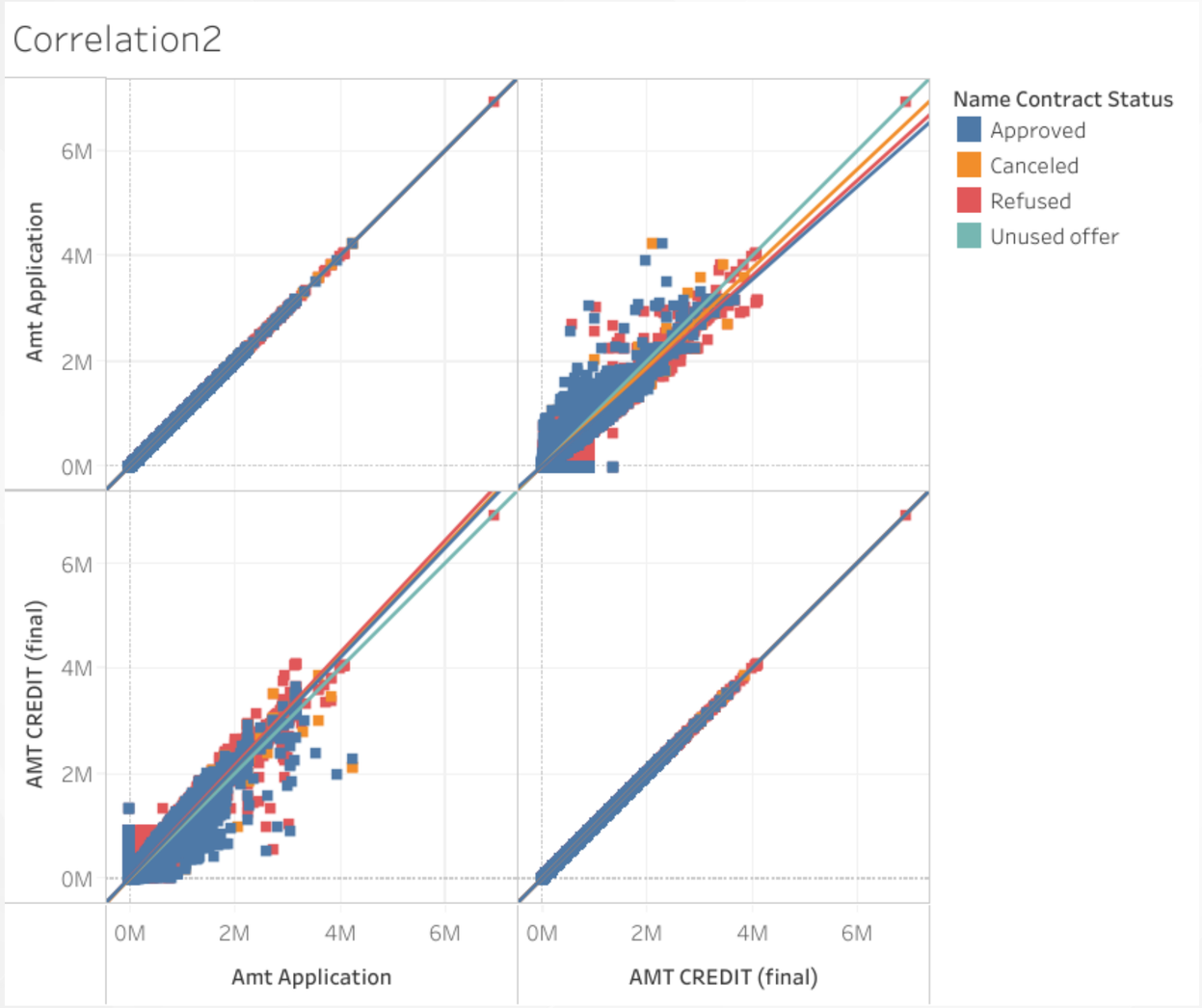




- NAME_CONTRACT_TYPE vs NAME_CONTRACT_STATUS: Consumer loans have a very high approval rate and a very low cancelation rate.
- NAME_YIELD_GROUP vs NAME_CONTRACT_STATUS: People in the high and middle categories have a high chance of approval.
- NAME_PAYMENT_TYPE vs NAME_CONTRACT_STATUS: Most of the applications prefer cash through the bank payment type.

ANALYSIS

CORRELATION ANALYSIS :



Here, the correlation analysis is performed on different numerical columns and trend lines are created to observe the relationship. The trend line which shows the R-squared value closer to 1 has a higher correlation.

AMT_APPLICATION and AMT_CREDIT have a good correlation.

TECH-STACK USED

MICROSOFT EXCEL

Excel is a powerful spreadsheet software that offers a wide range of tools and features for data manipulation and visualization. I have used excel for data cleaning.

TABLEAU PUBLIC

Tableau Public is a free platform to explore, create and publicly share data visualizations online. I have used this as it provides various features to create visually appealing charts.

INSIGHTS

- Female applicants are very less defaulters in comparison to males.
- People who are married apply more for loans but defaults are mostly done by single people.
- People doing secondary/secondary special education type are high in defaulters and very less people doing academic degrees are defaulters.
- Working applicants applied for loans more and defaulted in large numbers.
- The repeaters' loans are rejected in high percentage earlier because they are most likely to do defaults based on the data.
- People who own House/Apartment are more likely to get loan approved in comparison to other housing types but high in number for defaulters too.

RESULTS

- This case study has helped in developing more understanding of Exploratory Data Analysis.
- I have gained more understanding of data cleaning and its importance in the overall analysis.
- I have improved my skills in plotting visualizations and getting valuable insights.
- This case study has demonstrated the power of data visualization in understanding important factors.
- I got to know the significance of Exploratory data analysis for finding helpful insights and thus making informed decisions.