CREDIT EXPLORARTORY DATA ANALYSIS CASE STUDY

SUBMITTED BY:

1.KAVITHA BALAMURUGAN

2.AKSHADA JOSHI

PROBLEM STATEMENT

- This case study aims to give you an idea of applying EDA in a real business scenario.
- A basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

BUSINESS UNDERSTANDING

Two types of risks are associated with the bank's decision:

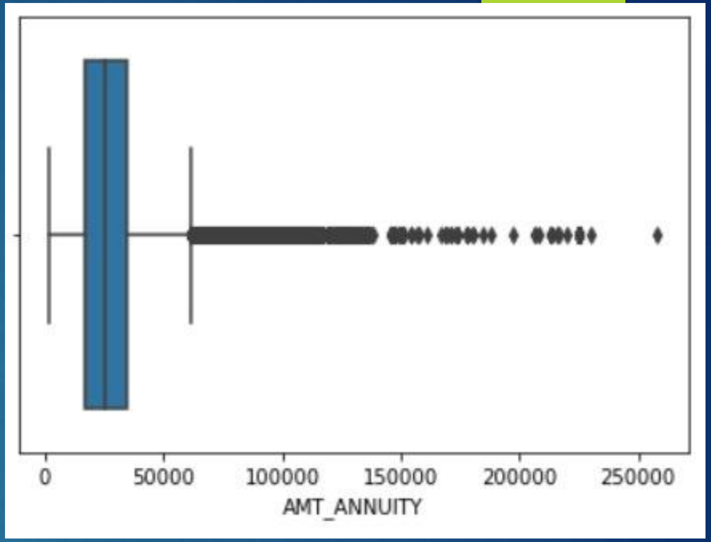
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

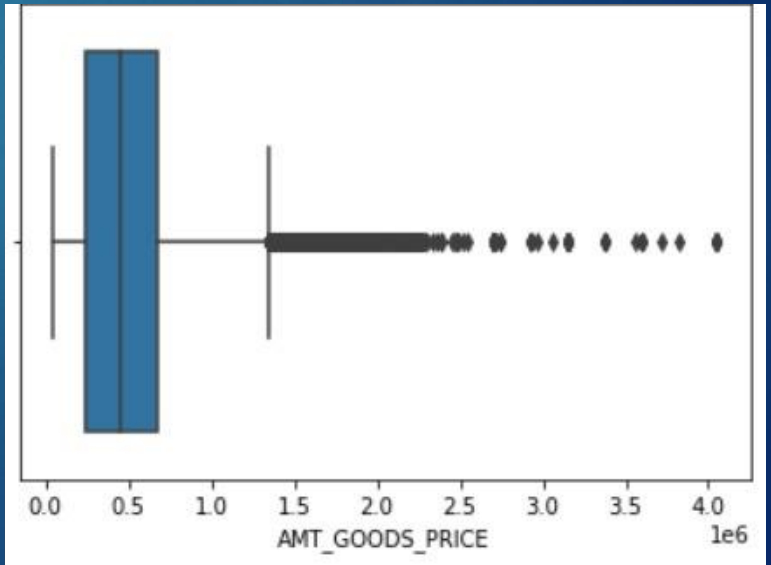
DATA CLEANING

- APPLICATION.CSV FILE:
- The total number of Rows and columns- (307511, 122)
- The after deleting of columns with missing values more than 50%- (307511,81)
- The total number of missing values less than 13%- 16 columns
- Dropped unwanted columns not needed for EDA analysis-'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', etc.
- Data imputation:
- For Numerical variables we replaced the missing data with the mean.
- For Categorical variables we replaced the missing values with the most occurred value/highest frequency value.

HANDLING OUTLIERS:

- Outlier Analysis was done for the following columns:
- 1. AMT_ANNUITY –
- LOWER BOUND- -9465.75, UPPER BOUND-59060.25
-
- 2. AMT_GOODS_PRICE-
- LOWER BOUND—438750.0 , UPPER BOUND- 1343250.0





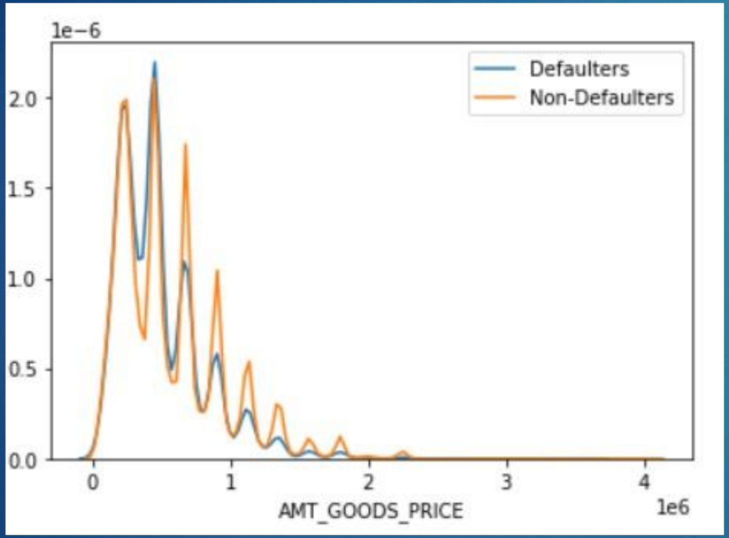
BINNING

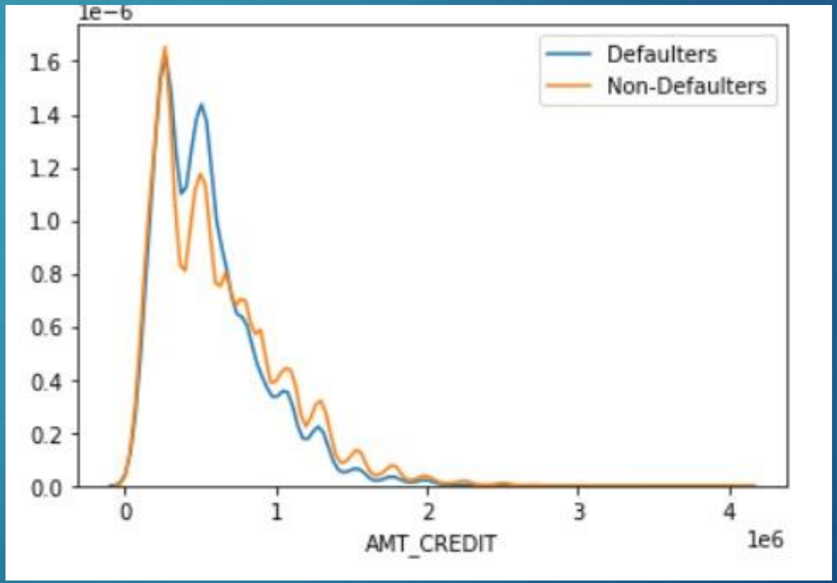
- For AMT_INCOME_TOTAL and AMT_CREDIT:

FOR AMT_INCOME_TOTAL and AMT_CREDIT we have binned the numerical values into categorical values.

CREATING TWO DATAFRAMES FOR TARGET VARIABLE

- We created two data frames namely df_0 and df_1, df_0= Non-defaulters and df_1= defaulters.
- 1.UNIVARIATE ANALYSIS FOR NUMERICAL DATASET:
- Plotted graph between defaulters and non defaulters for the columns AMT_GOODS_PRICE and AMT_CREDIT.

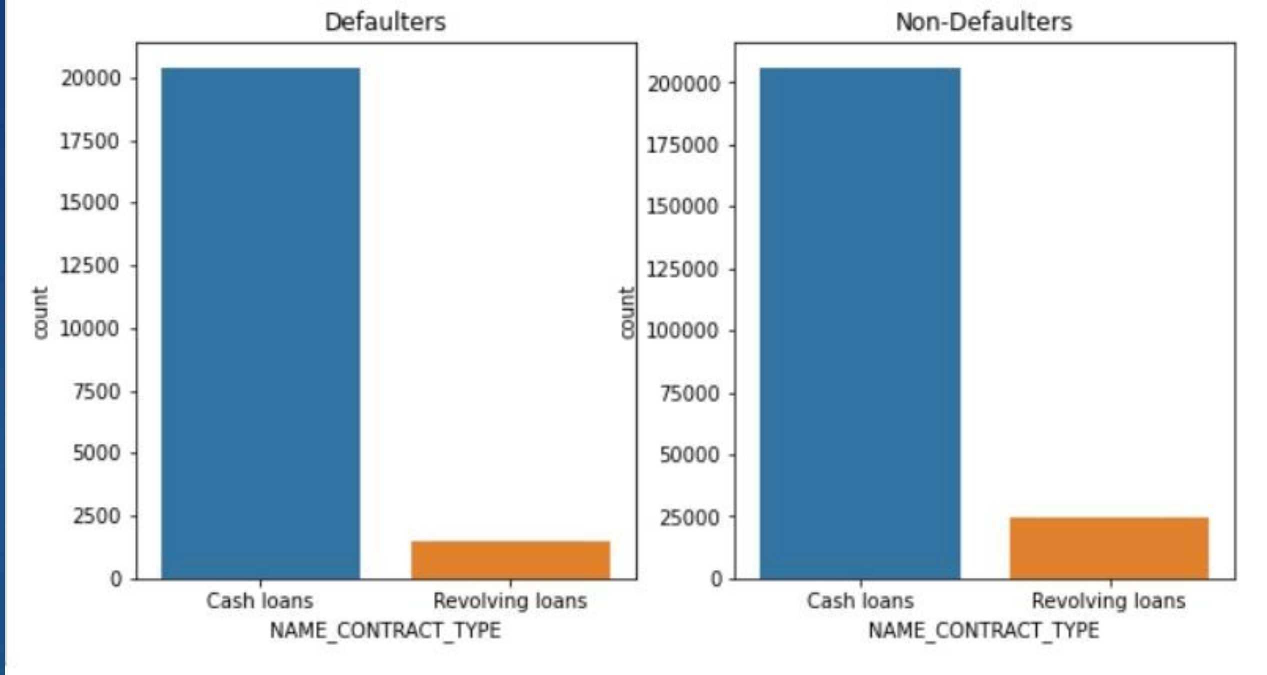




INFERENCES- For AMT_CREDIT the number of defaulters and non-defaulters overlap for the range 0-1 and 1.6(y-axis) then the defaulter count increases, and AMT_GOODS_PRICE the defaulters count is more comparatively for the range (0,1), >2.0 (y-axis).

UNIVARIATE ANALYSIS FOR CATEGORICAL DATASET

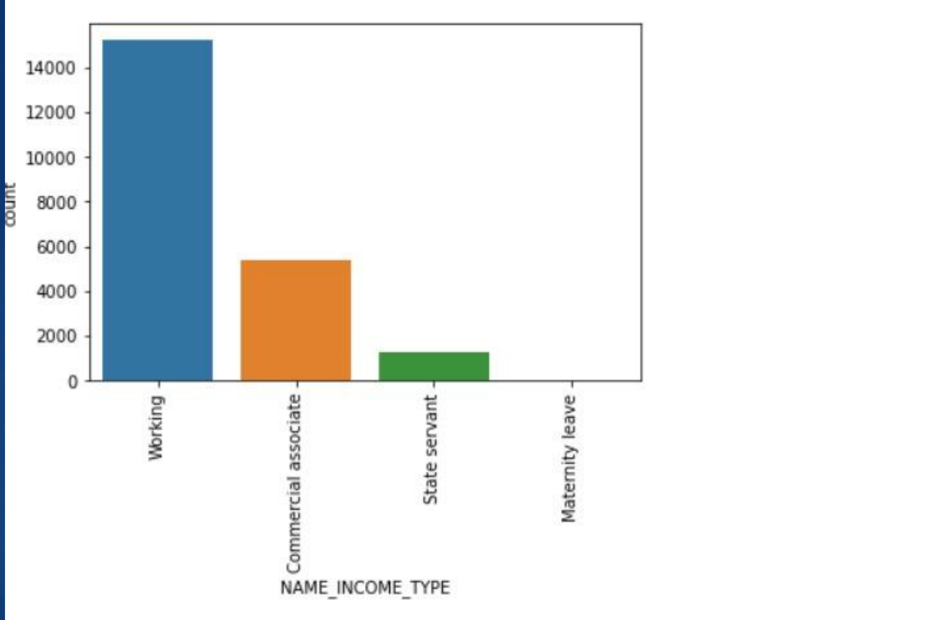
- Plotted graph for NAME_CONTRACT_TYPE for defaulters and non-defaulters:

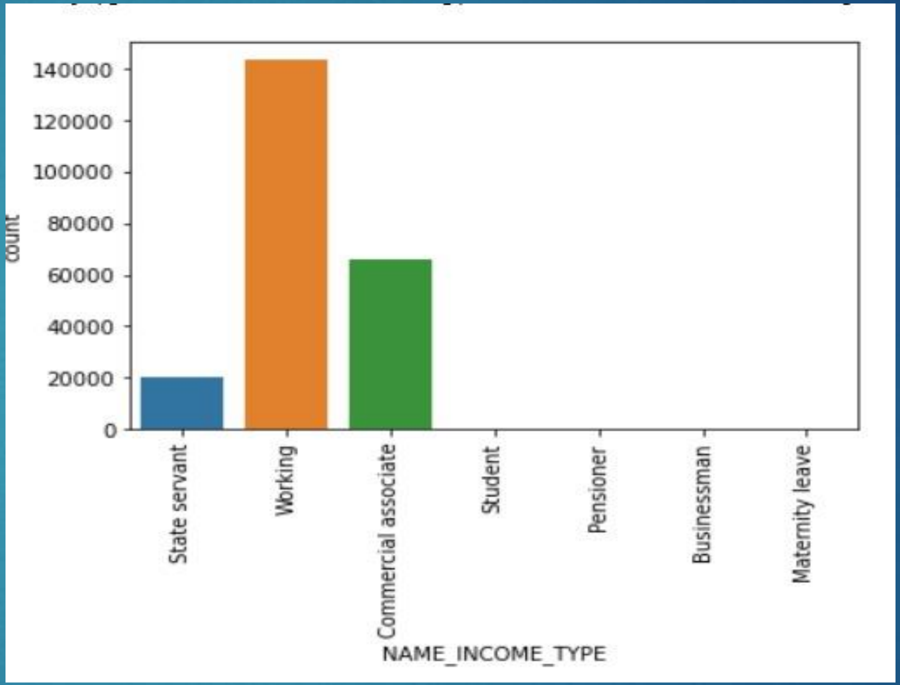


INFERENECES- The proportion of non-defaulters for cash loan is high as compared to the proportion of non-defaulters for cash loans.

SEGMENTED UNIVARIATE ANALYSIS:

- Plotting graph between defaulters and non-defaulters for column name NAME_INCOME_TYPE:

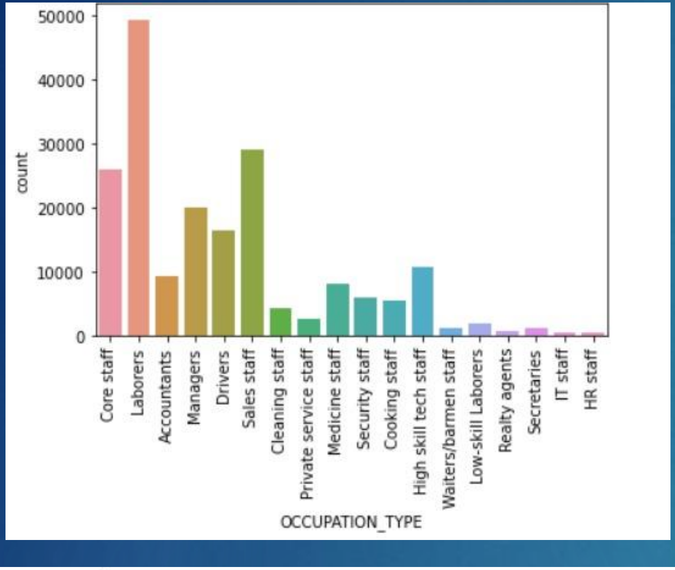


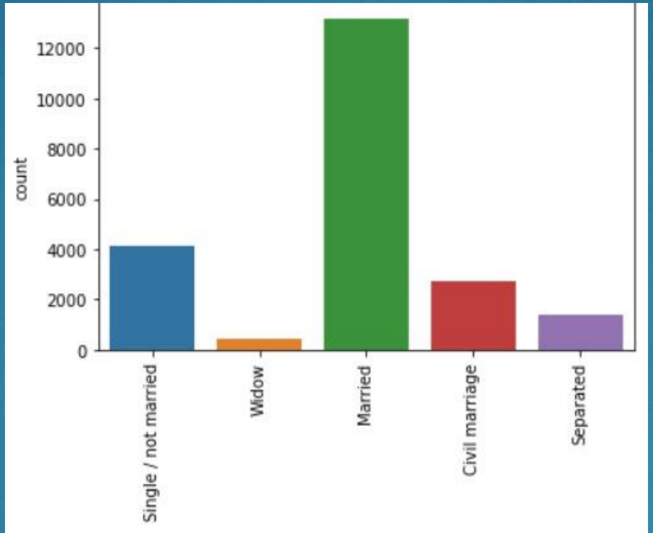


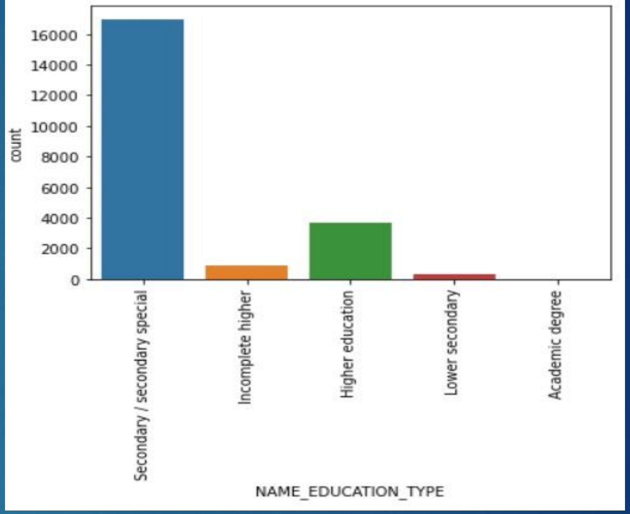
- INFERENECES- The count of non-defaulters is 10 times higher than the count of defaulters for the income type as working. In similar way count of non-defaulters for commercial associates and state servant is 10 times more than the defaulters.

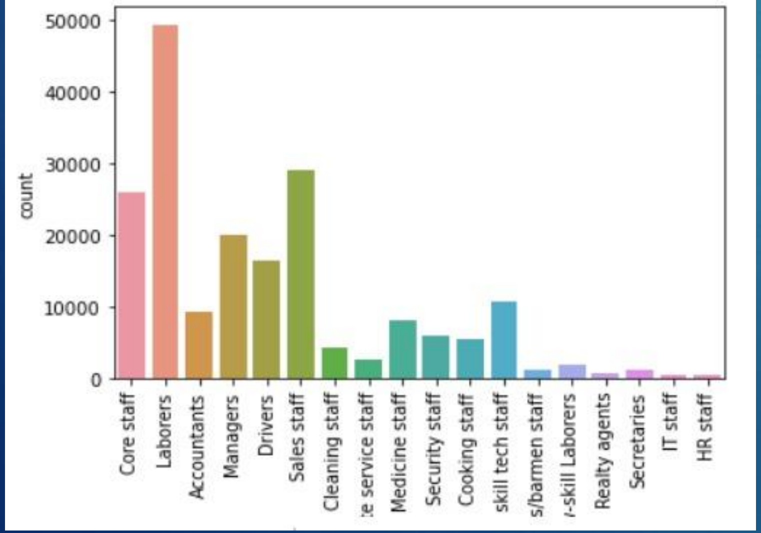
SEGEMENTED ANALYSIS:

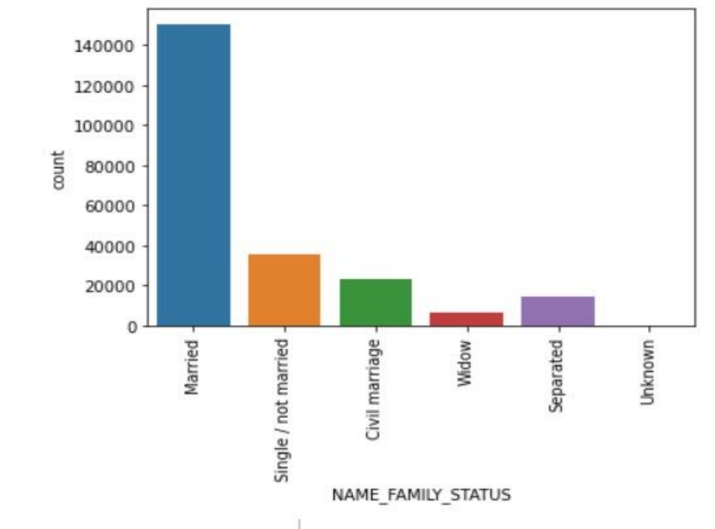
- Plotting graph for defaulters and non-defaulters for the column name OCCUPATION_TYPE, NAME_EDUCATION_TYPE and NAME_FAMILY_STATUS:

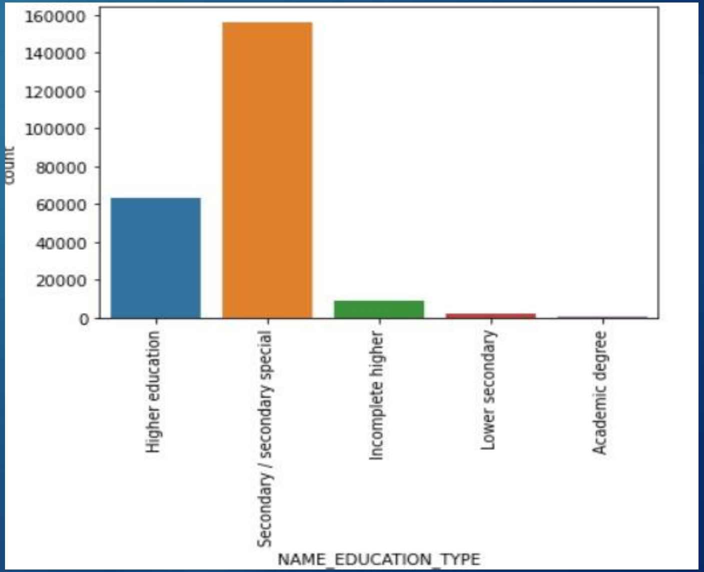










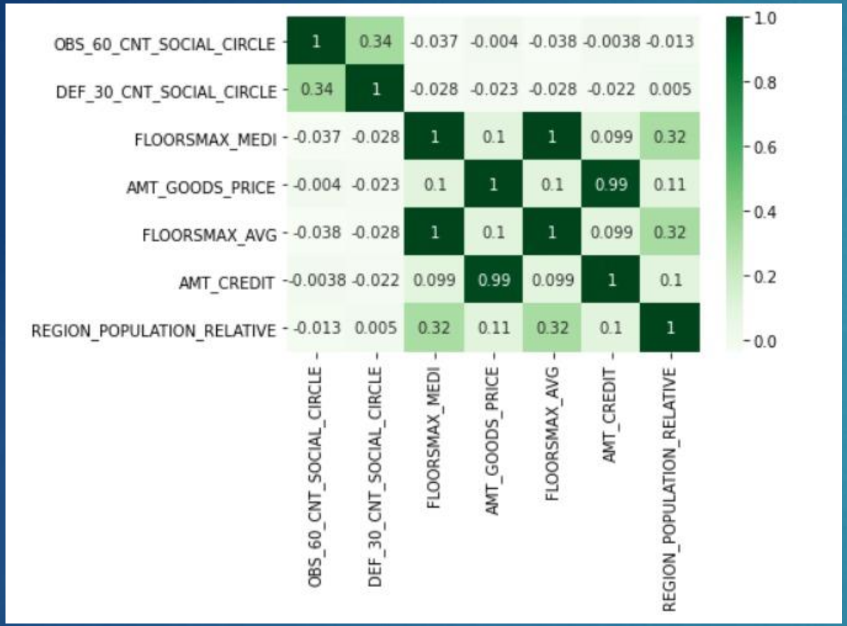


INFERENCES FROM SEGEMENTED UNIVARIATE ANALYSIS:

- INFERENCES- 1. Customers with profession as Laborer have higher proportion of defaulters
- Another observation is as IT/HR have lower proportion of defaulting
- Customers with Secondary education have high proportion of defaulting if compared to non-defaulters
- Customers with higher education tend to default less as their proportion is reduced

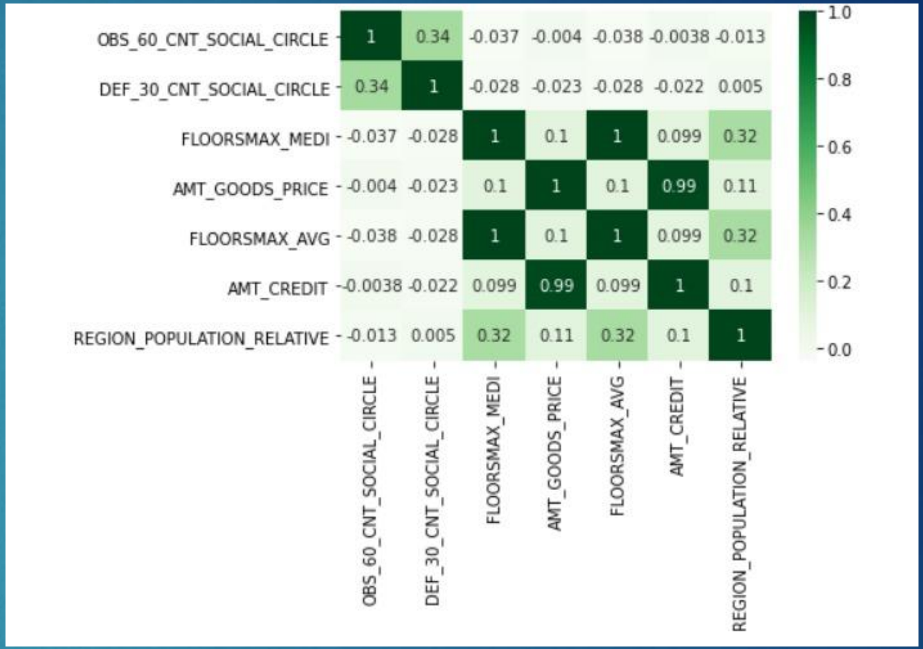
CORRELATION

- The list of correlations between variables are as follows:



	OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	FLOORSMAX_MEDI	AMT_GOODS_PRICE	FLOORSMAX_AVG	AMT_CREDIT	REGION_POPULATION_RELATIVE
OBS_60_CNT_SOCIAL_CIRCLE	1	0.34	-0.037	-0.004	-0.038	-0.0038	-0.013
DEF_30_CNT_SOCIAL_CIRCLE	0.34	1	-0.028	-0.023	-0.028	-0.022	0.005
FLOORSMAX_MEDI	-0.037	-0.028	1	0.1	1	0.099	0.32
AMT_GOODS_PRICE	-0.004	-0.023	0.1	1	0.1	0.99	0.11
FLOORSMAX_AVG	-0.038	-0.028	1	0.1	1	0.099	0.32
AMT_CREDIT	-0.0038	-0.022	0.099	0.99	0.099	1	0.1
REGION_POPULATION_RELATIVE	-0.013	0.005	0.32	0.11	0.32	0.1	1

Non-defaulters



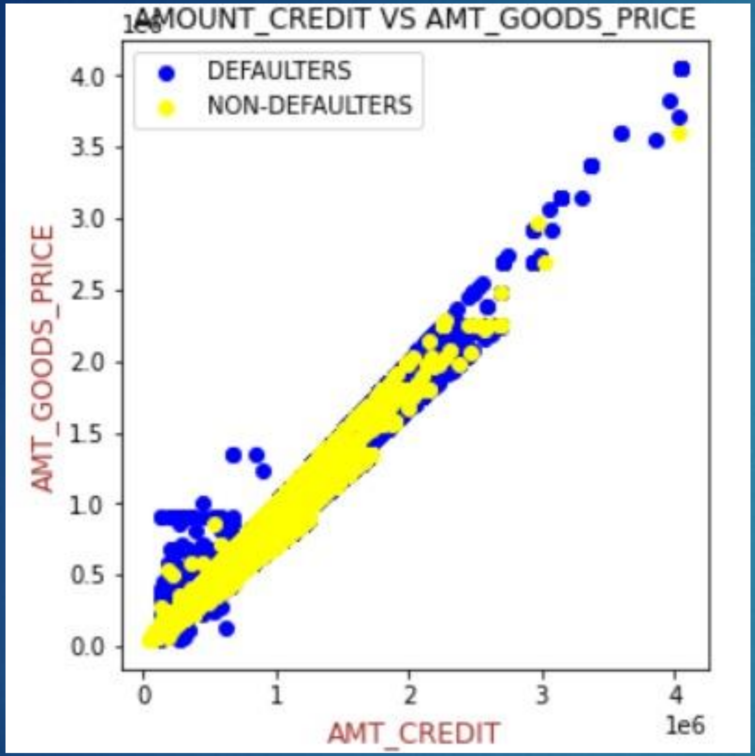
	OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	FLOORSMAX_MEDI	AMT_GOODS_PRICE	FLOORSMAX_AVG	AMT_CREDIT	REGION_POPULATION_RELATIVE
OBS_60_CNT_SOCIAL_CIRCLE	1	0.34	-0.037	-0.004	-0.038	-0.0038	-0.013
DEF_30_CNT_SOCIAL_CIRCLE	0.34	1	-0.028	-0.023	-0.028	-0.022	0.005
FLOORSMAX_MEDI	-0.037	-0.028	1	0.1	1	0.099	0.32
AMT_GOODS_PRICE	-0.004	-0.023	0.1	1	0.1	0.99	0.11
FLOORSMAX_AVG	-0.038	-0.028	1	0.1	1	0.099	0.32
AMT_CREDIT	-0.0038	-0.022	0.099	0.99	0.099	1	0.1
REGION_POPULATION_RELATIVE	-0.013	0.005	0.32	0.11	0.32	0.1	1

DEFAULTERS

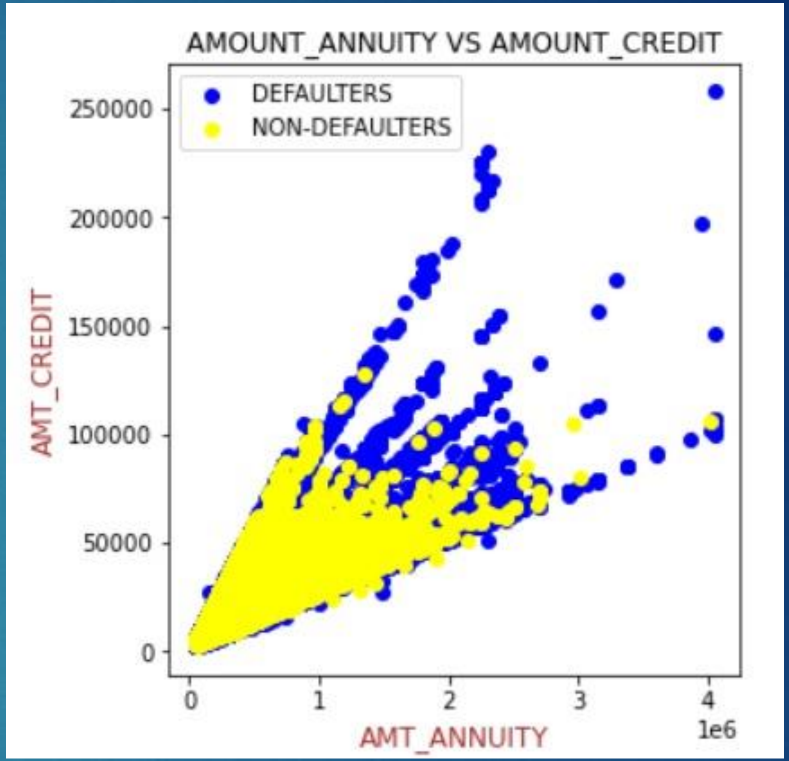
- INFERENCES- The correlation for defaulters is comparatively high than non-defaulters.

BIVARIATE ANALYSIS FOR NUMERICAL DATA:

- AMT_GOODS_PRICE VS AMT_CREDIT:



AMT_ANNUITY VS AMT_CREDIT:



INFERENCES- 1.From the above analysis the AMT_GOODS_PRICE and AMT_CREDIT are highly correlated so AMT_GOODS_PRICE increases then the AMT_CREDIT also increases.

2. From the above analysis the AMT_ANNUITY and AMT_CREDIT are strongly correlated so AMT_CREDIT increases then the AMT_ ANNUITY also increases.

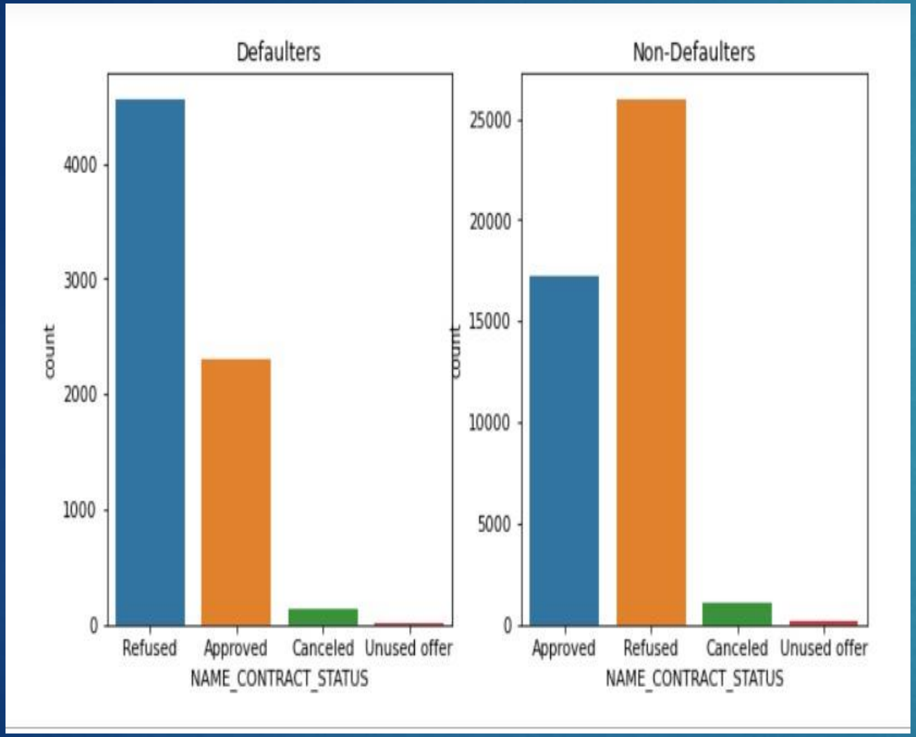
ANALYSIS ON PREVIOUS APPLICATION DATA FILE

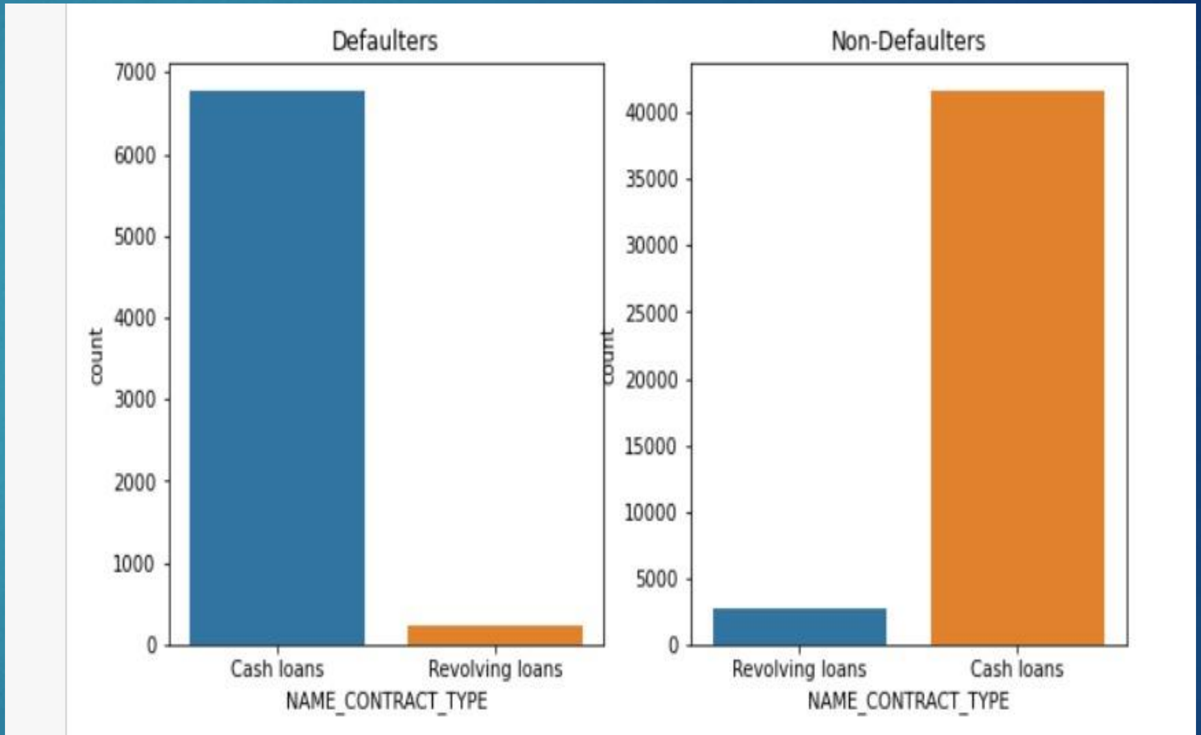
- 1. Read the previous_data.csv file.

- DATA CLEANING:
- 1. Removed 50% of data having null value, and dropped the unwanted columns not needed for the analysis.
- 2.Remove the missing values in the data columns i.e, XNA,XPA.
- 3.Merge the two data frames application_data and previous_data csv files.
- 4.Rename the columns for the ease of analysis.
- 5.Created two target variables final_df1 = Defaulters and final_df0= Non-Defaulters.

UNIVARIATE ANALYSIS:

- Univariate analysis for defaulters and non-defaulters between the column names NAME_CONTRACT_STATUS and NAME_CONTRACT_TYPE:

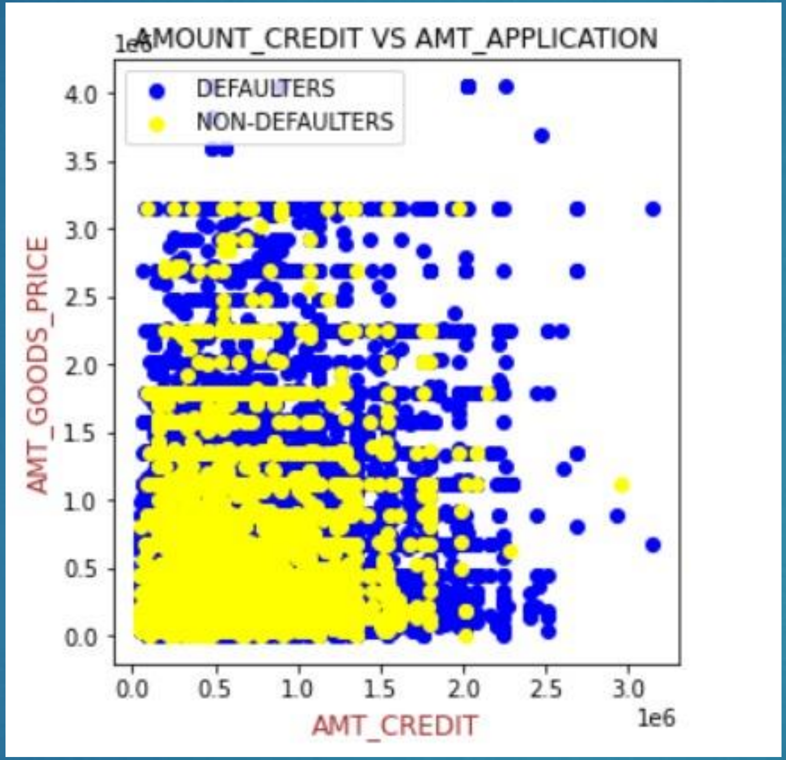




- INFERENCES- Client have larger proportion of REFUSED applications and the defaulters are more of previous application were cash loans.

BIVARIATE ANALYSIS:

- Univariate analysis for defaulters and non-defaulters between the column names AMT_CREDIT and AMT_APPLICATION:



- INFERENCES: The previous application amount and the credit amount are about 0.97.
- The amount in current application and previous application amount for non-defaulters is low as compared to defaulters.

- THANK YOU!!!