CREDIT EDA - ASSIGNMENT

By:- Rashmi Singh

Introduction

This assignment aims to give an idea of applying EDA in a real business scenario. In this assignment, apart from applying the EDA techniques we also develop 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.

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected

Business Objective

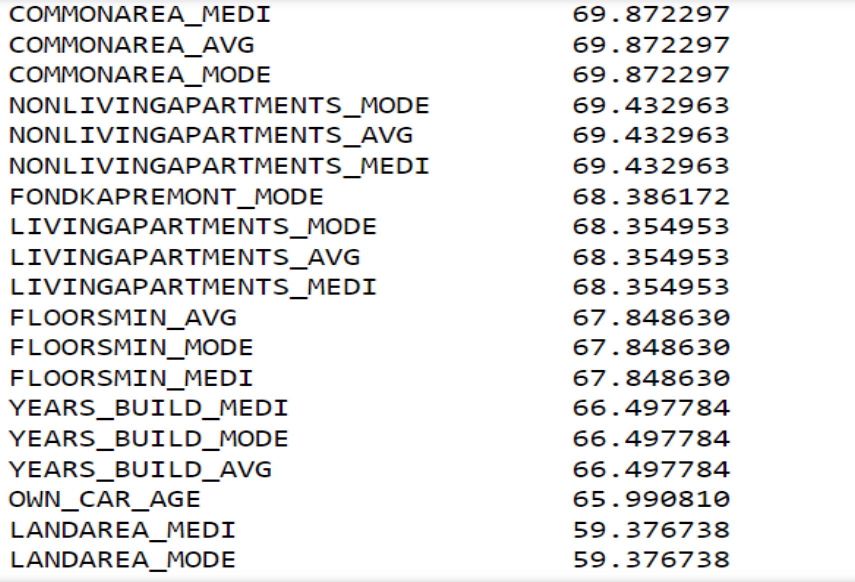
This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study. In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment

Data Cleaning Approach

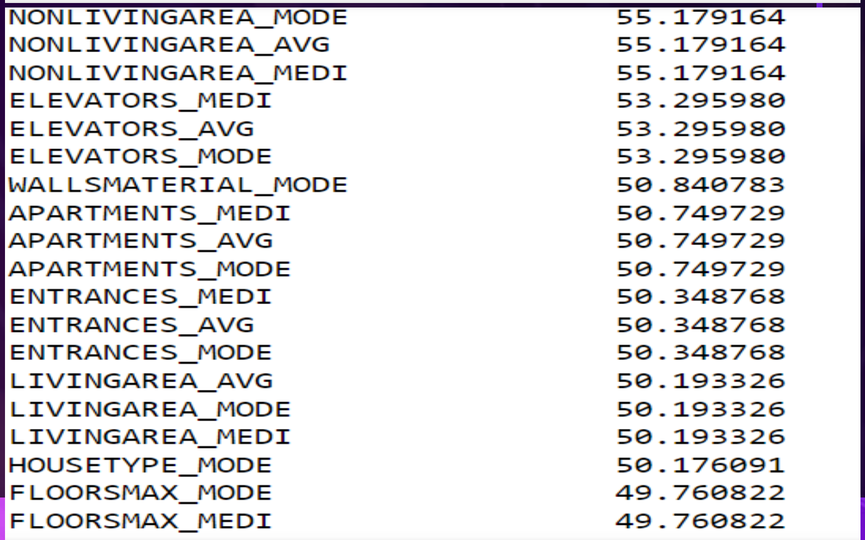
- Due to Threshold theory for any dataset if we have more than 40% missing value we can drop the column because that column don't give accurate result.
- I have useless columns like flag and in this values are 0 or 1 and which is not giving any useful information so I have removed all the unnecessary columns
- **checked days column where negative values are there so we need to convert it to positive by using absolute**
- **Also by analysis the data which I get is Female applicants are more as compare to male applicants**

Missing Value

Application Data

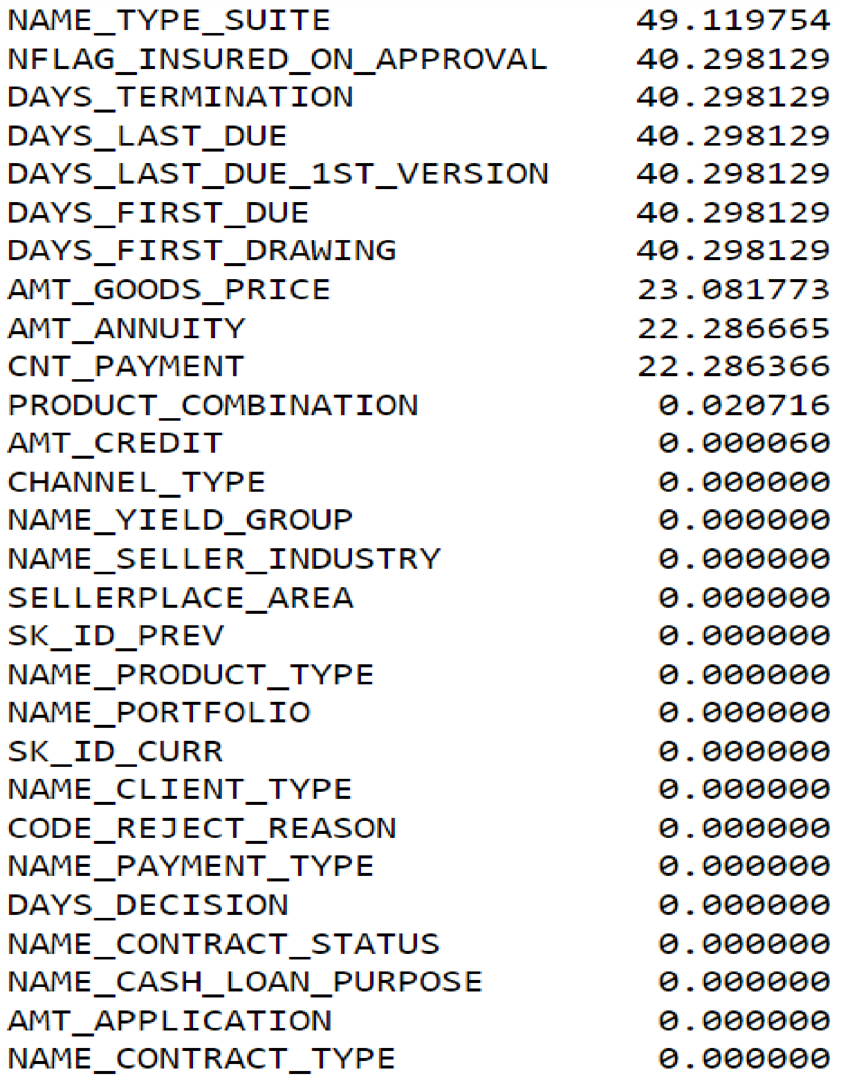


```
COMMONAREA_MEDI                 69.872297
COMMONAREA_AVG                  69.872297
COMMONAREA_MODE                 69.872297
NONLIVINGAPARTMENTS_MODE        69.432963
NONLIVINGAPARTMENTS_AVG         69.432963
NONLIVINGAPARTMENTS_MEDI        69.432963
FONDKAPREMONT_MODE              68.386172
LIVINGAPARTMENTS_MODE           68.354953
LIVINGAPARTMENTS_AVG            68.354953
LIVINGAPARTMENTS_MEDI           68.354953
FLOORSMIN_AVG                   67.848630
FLOORSMIN_MODE                  67.848630
FLOORSMIN_MEDI                  67.848630
YEARS_BUILD_MEDI                66.497784
YEARS_BUILD_MODE                66.497784
YEARS_BUILD_AVG                 66.497784
OWN_CAR_AGE                     65.990810
LANDAREA_MEDI                   59.376738
LANDAREA_MODE                   59.376738
```



```
NONLIVINGAREA_MODE                 55.179164
NONLIVINGAREA_AVG                  55.179164
NONLIVINGAREA_MEDI                 55.179164
ELEVATORS_MEDI                     53.295980
ELEVATORS_AVG                      53.295980
ELEVATORS_MODE                     53.295980
WALLSMATERIAL_MODE                 50.840783
APARTMENTS_MEDI                    50.749729
APARTMENTS_AVG                     50.749729
APARTMENTS_MODE                    50.749729
ENTRANCES_MEDI                     50.348768
ENTRANCES_AVG                      50.348768
ENTRANCES_MODE                     50.348768
LIVINGAREA_AVG                     50.193326
LIVINGAREA_MODE                    50.193326
LIVINGAREA_MEDI                    50.193326
HOUSETYPE_MODE                     50.176091
FLOORSMAX_MODE                     49.760822
FLOORSMAX_MEDI                     49.760822
```

Previous Data



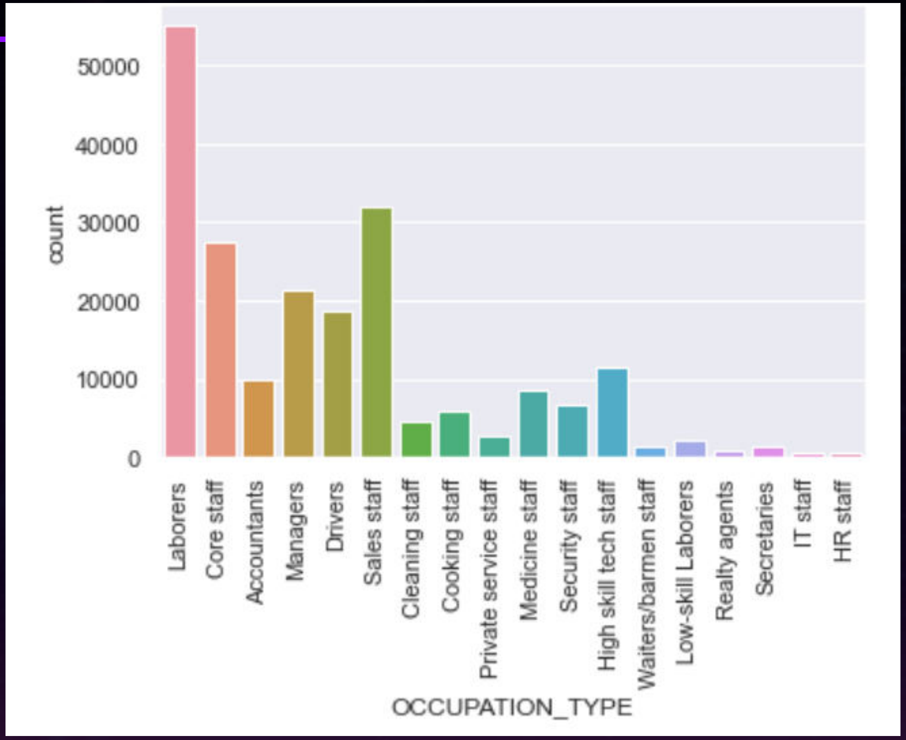
```
NAME_TYPE_SUITE              49.119754
NFLAG_INSURED_ON_APPROVAL    40.298129
DAYS_TERMINATION             40.298129
DAYS_LAST_DUE                40.298129
DAYS_LAST_DUE_1ST_VERSION    40.298129
DAYS_FIRST_DUE               40.298129
DAYS_FIRST_DRAWING           40.298129
AMT_GOODS_PRICE              23.081773
AMT_ANNUITY                  22.286665
CNT_PAYMENT                  22.286366
PRODUCT_COMBINATION           0.020716
AMT_CREDIT                    0.000060
CHANNEL_TYPE                  0.000000
NAME_YIELD_GROUP              0.000000
NAME_SELLER_INDUSTRY          0.000000
SELLERPLACE_AREA              0.000000
SK_ID_PREV                    0.000000
NAME_PRODUCT_TYPE             0.000000
NAME_PORTFOLIO                0.000000
SK_ID_CURR                    0.000000
NAME_CLIENT_TYPE              0.000000
CODE_REJECT_REASON            0.000000
NAME_PAYMENT_TYPE             0.000000
DAYS_DECISION                 0.000000
NAME_CONTRACT_STATUS          0.000000
NAME_CASH_LOAN_PURPOSE        0.000000
AMT_APPLICATION               0.000000
NAME_CONTRACT_TYPE            0.000000
```

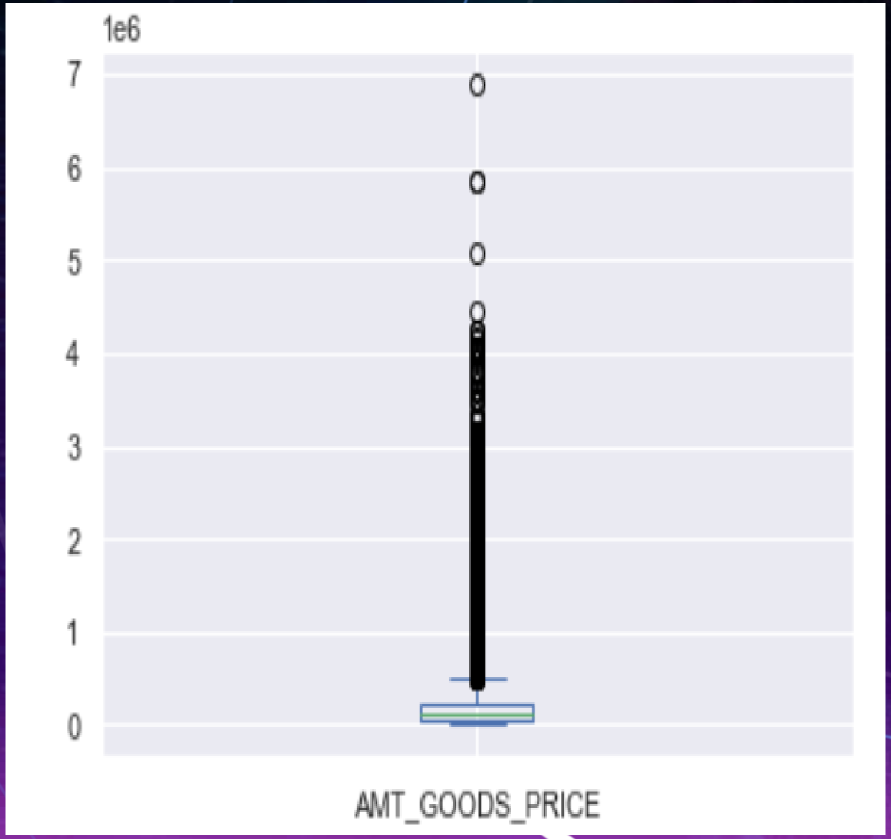
MISSING VALUE TREATMENT

1.if variable is **object** means categorical so will fill with **Mode**

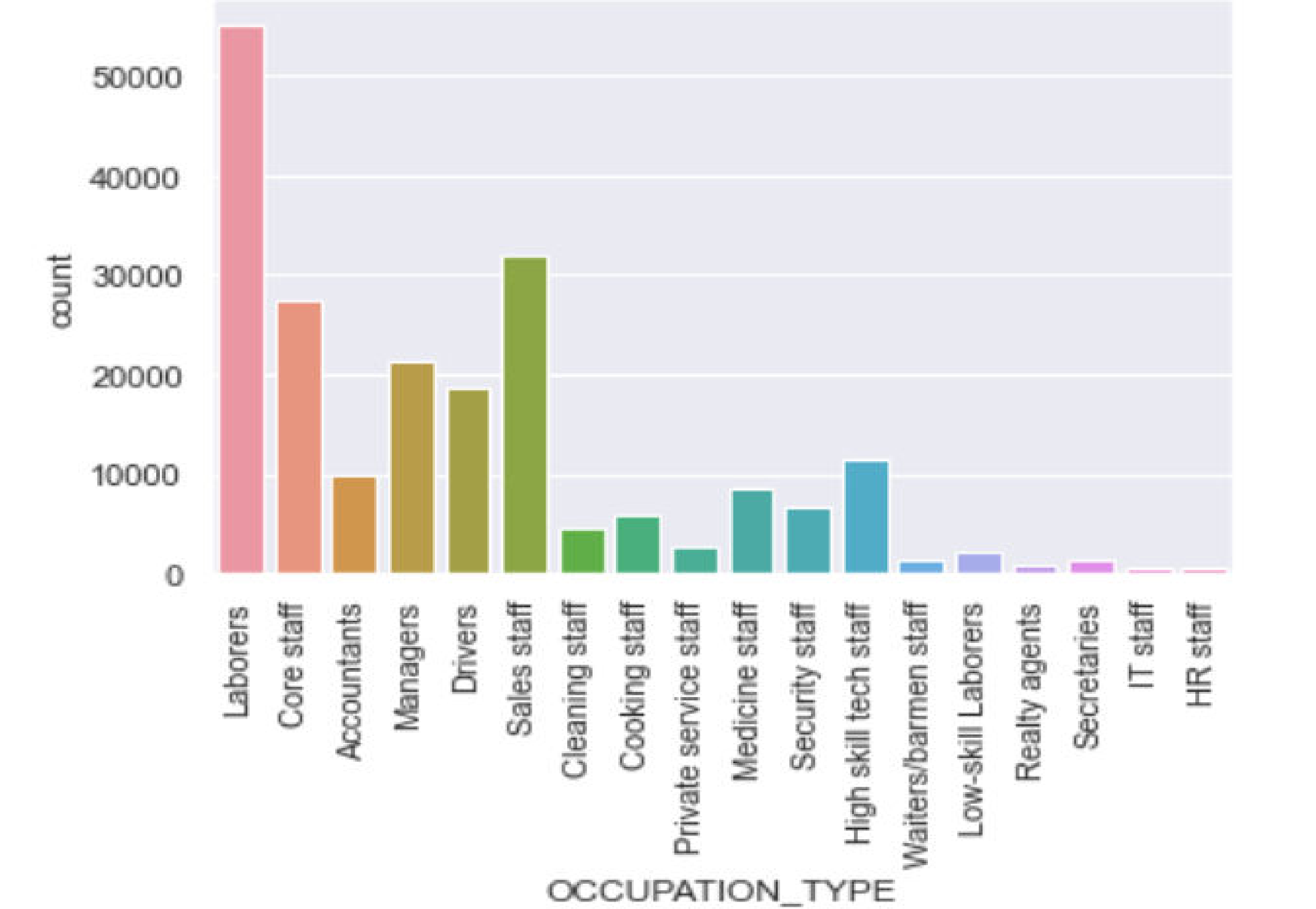
2.if variable is **int** or **float** means numerical so will fill with **Median** or **Mean**

If we have numerical variable so we preferred fill with median because due to outlier in data set they impact mean but not median



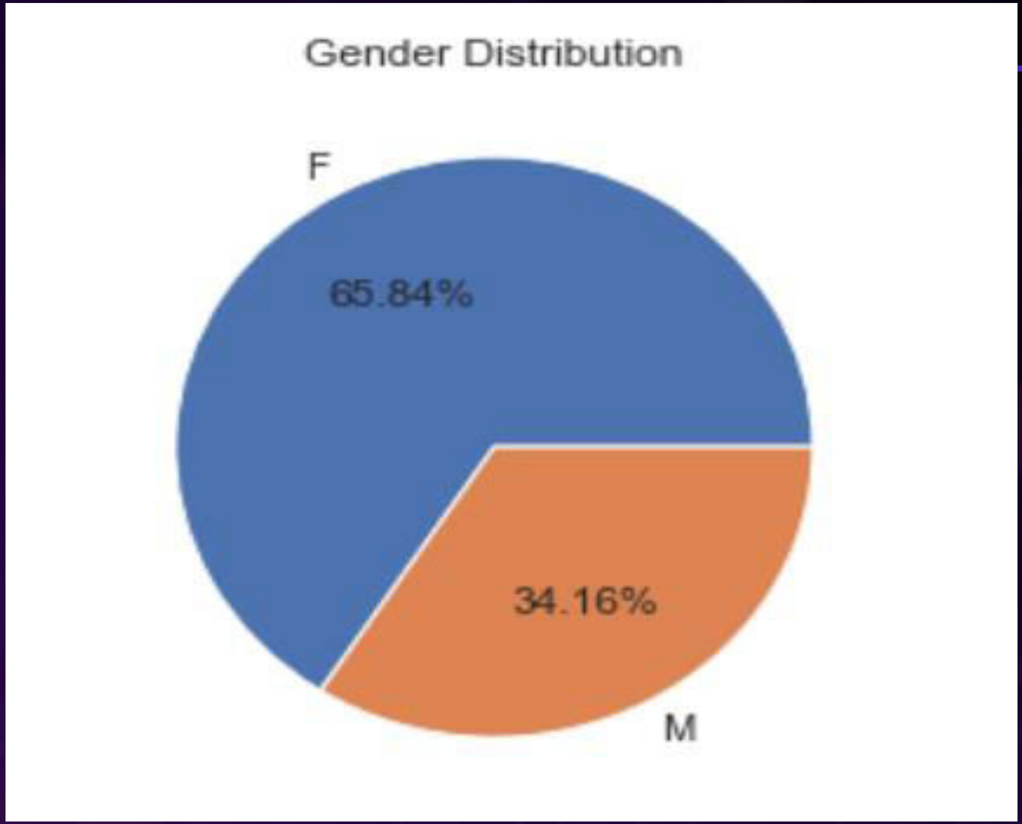


HANDLING OUTLIER



The occupation Type is categorical so we need to fill it with mode

Gender Distribution

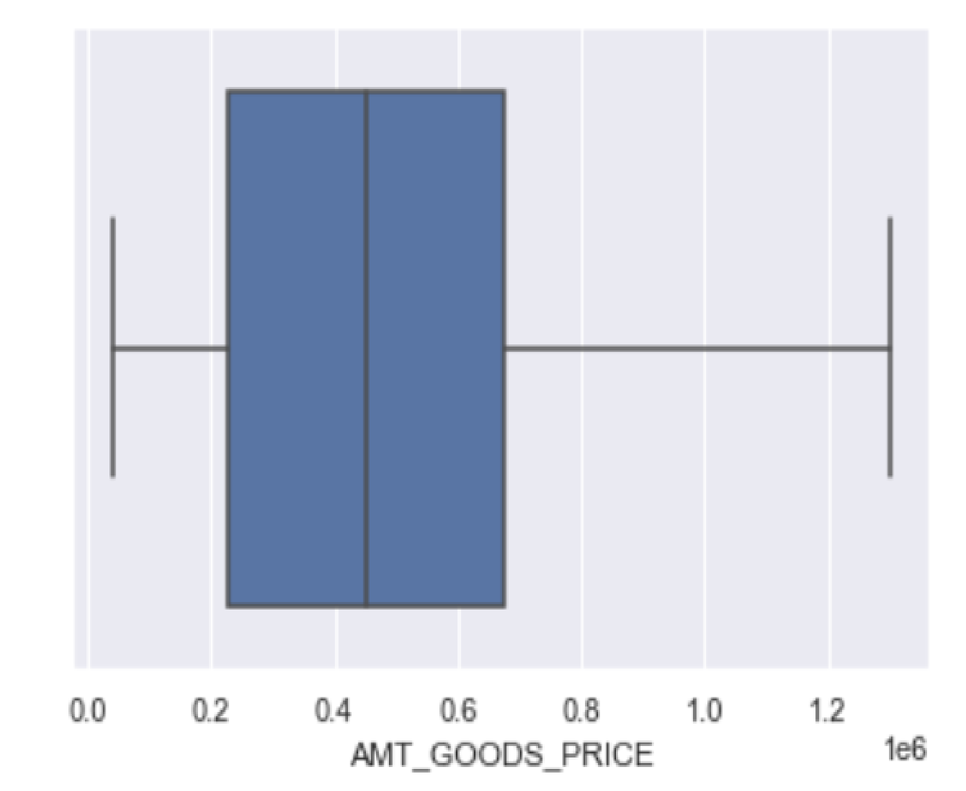


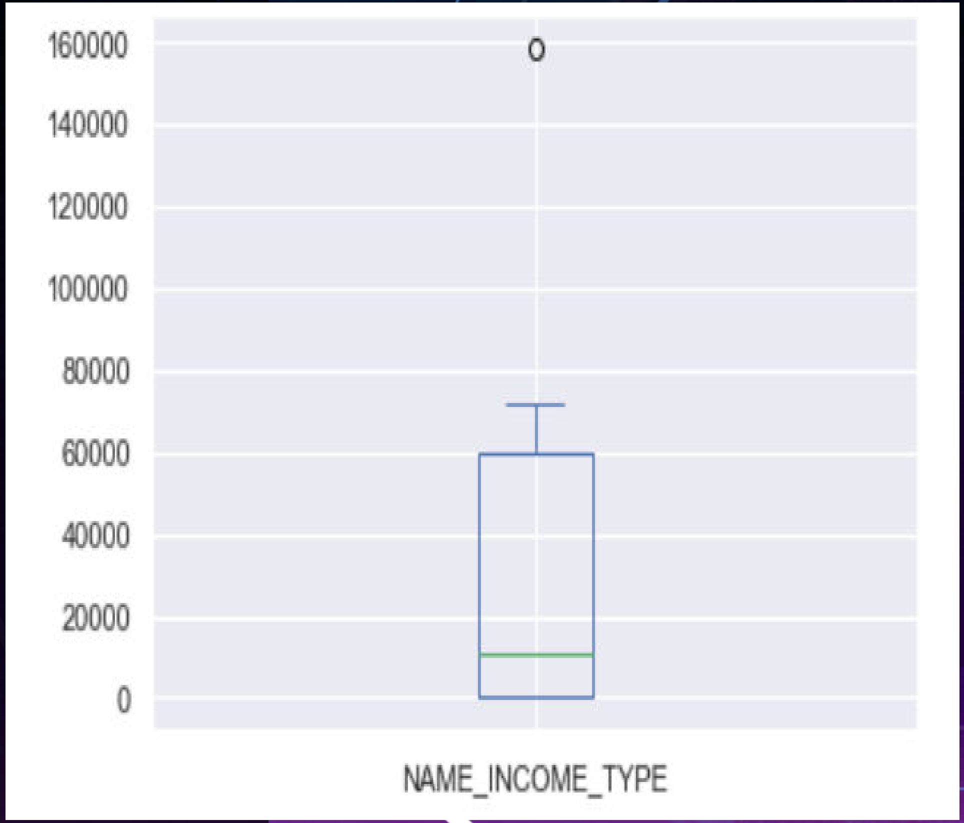
Female applicants are more as compare to male applicants

OUTERLIERS

Outliers can be removed after setting the value upto 0.95 percentile

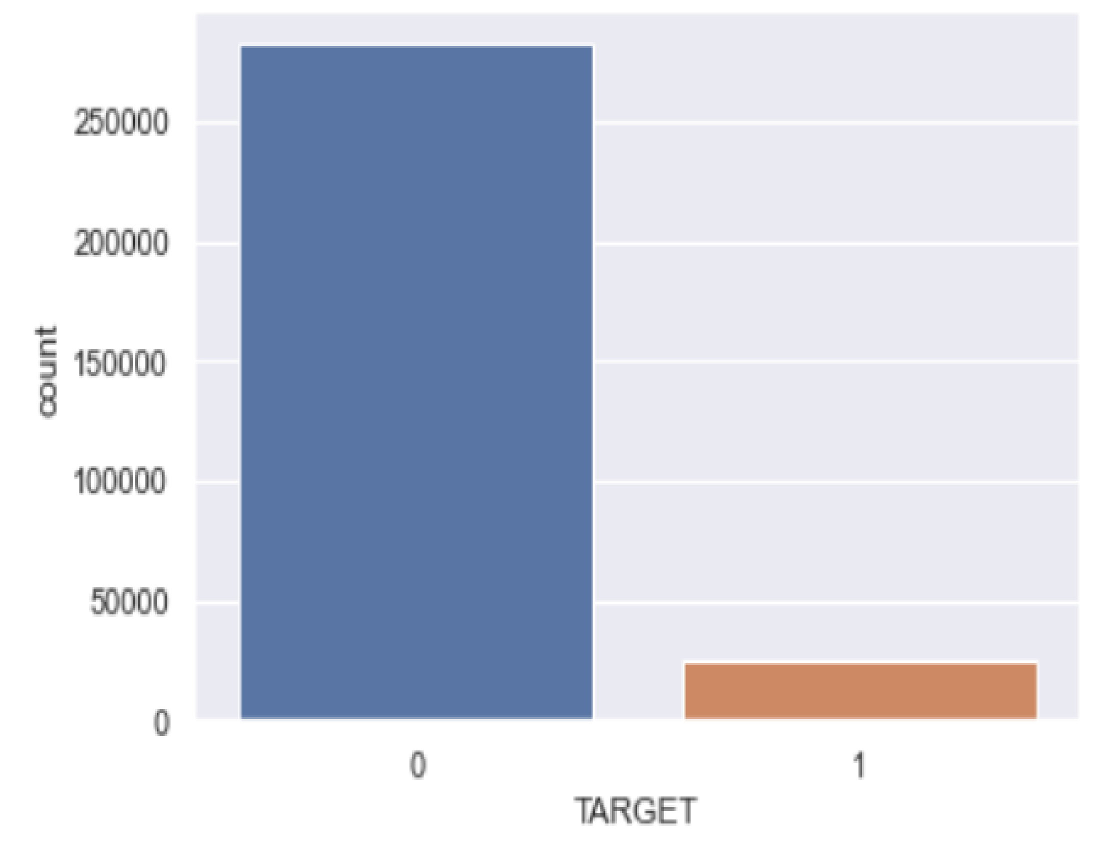
- Outliers identified which is at the max point



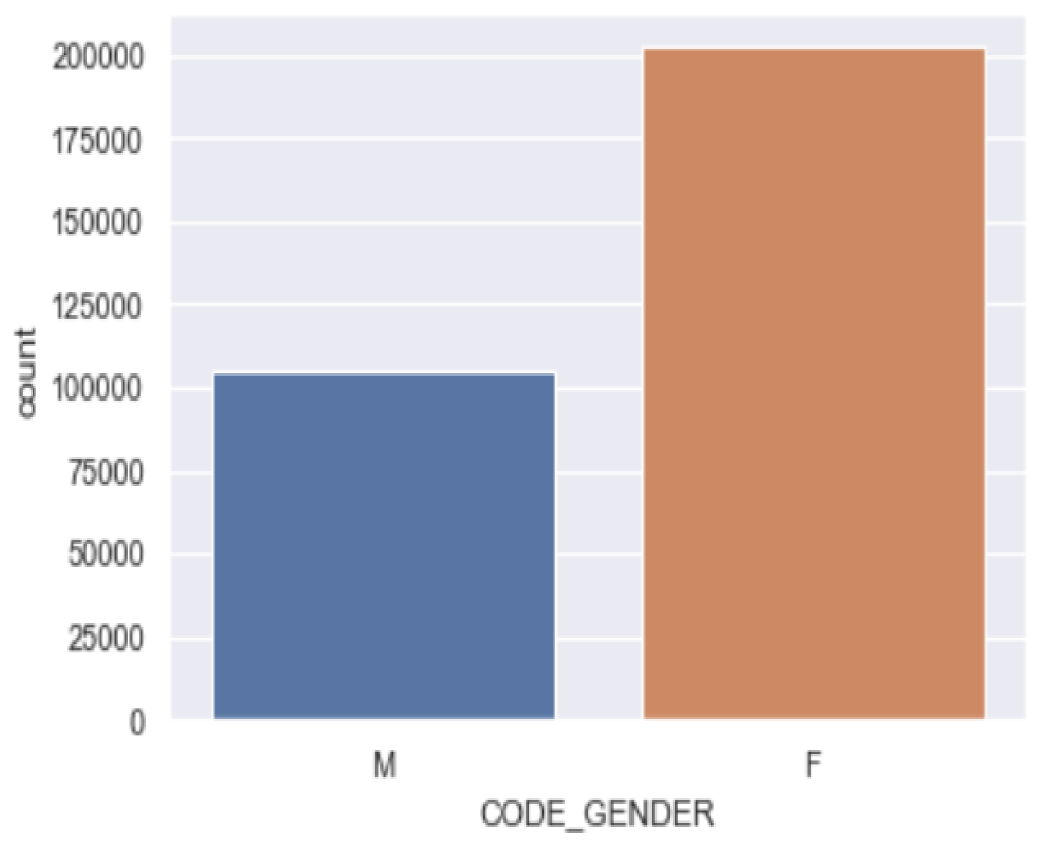


APPLICATION DATA

Cash Loans Are More In Comparison Of Revolving Loans



The Number Of Female Are More In Comparison to Male



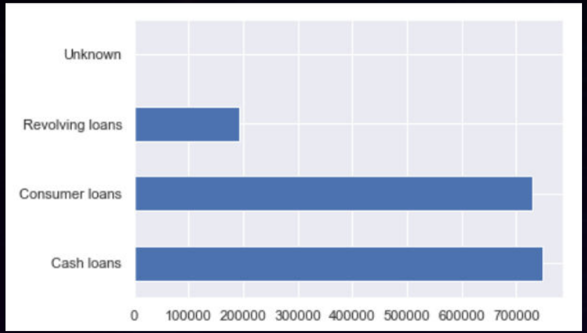
Previous Application Data

Univariate Analysis

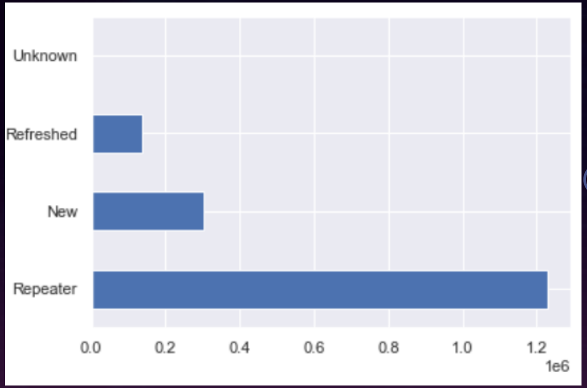
Bivariate Analysis

The CASH LOAN ARE MORE AS COMPARE TO OTHER TYPE

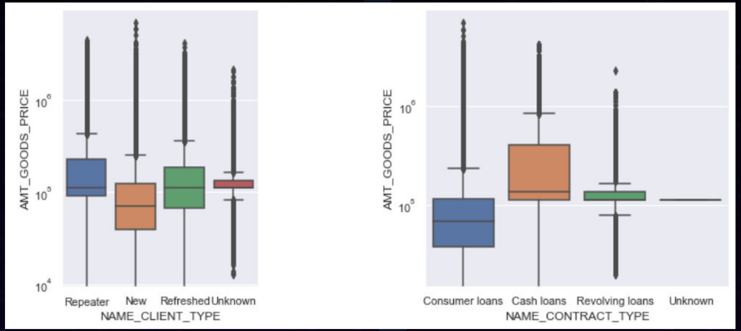
OF LOANS



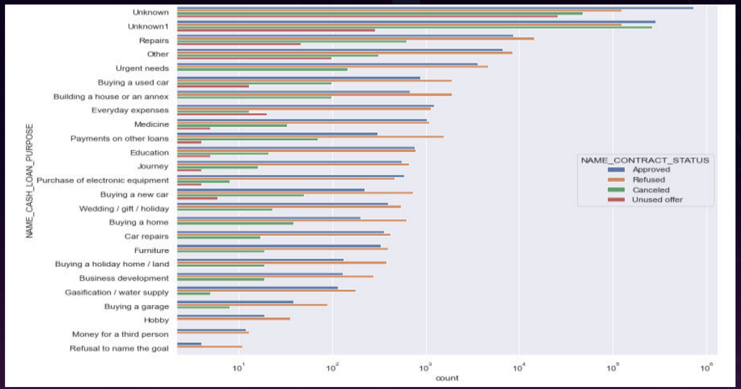
Repeater clients are more as compare to others



Repeaters have the highest AMT_GOODS_PRICE Cash Loans are also more
as compare to other loans¶

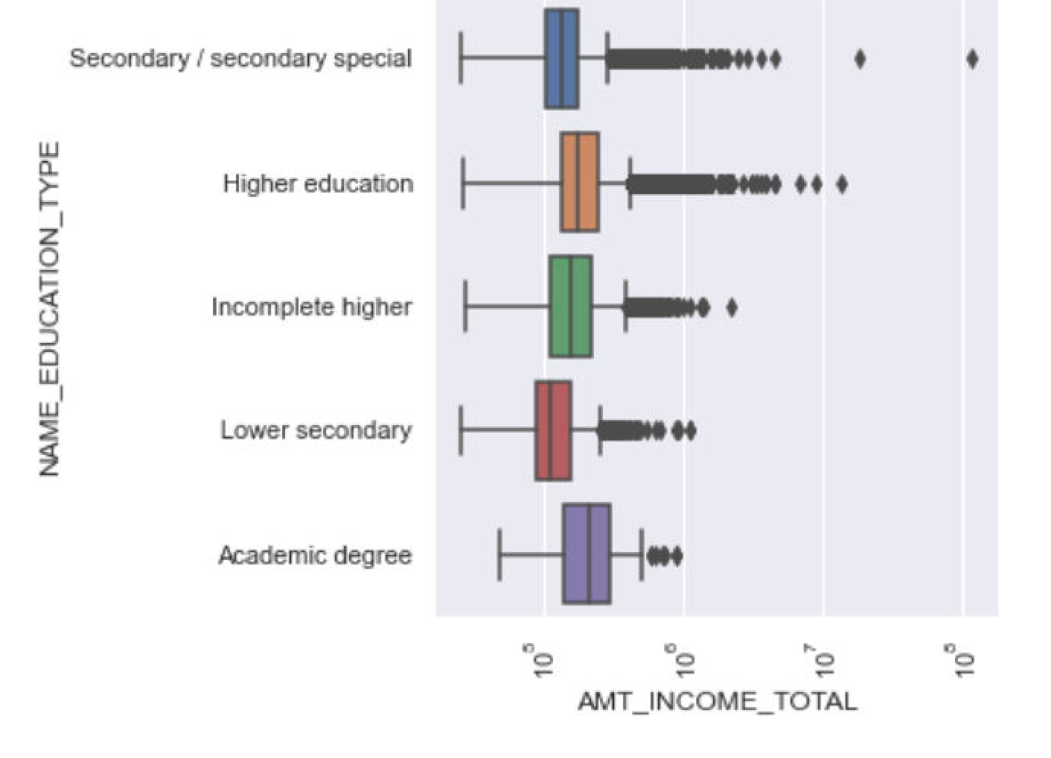


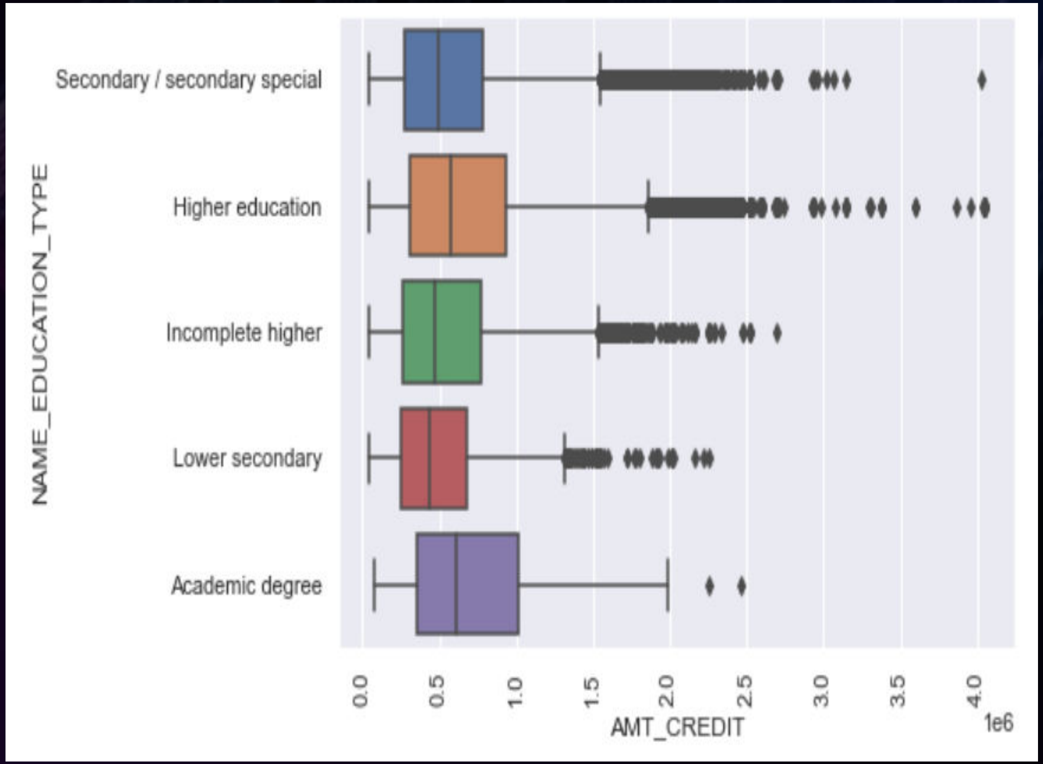
THE NUMBER OF REFUSED LOANS ARE for LOANS AND REPAIR PURPOSE
UNKNOWN AND UNKNOWN 1 ARE ALSO VERY POPULAR



Amount Income Total/Amount Credit

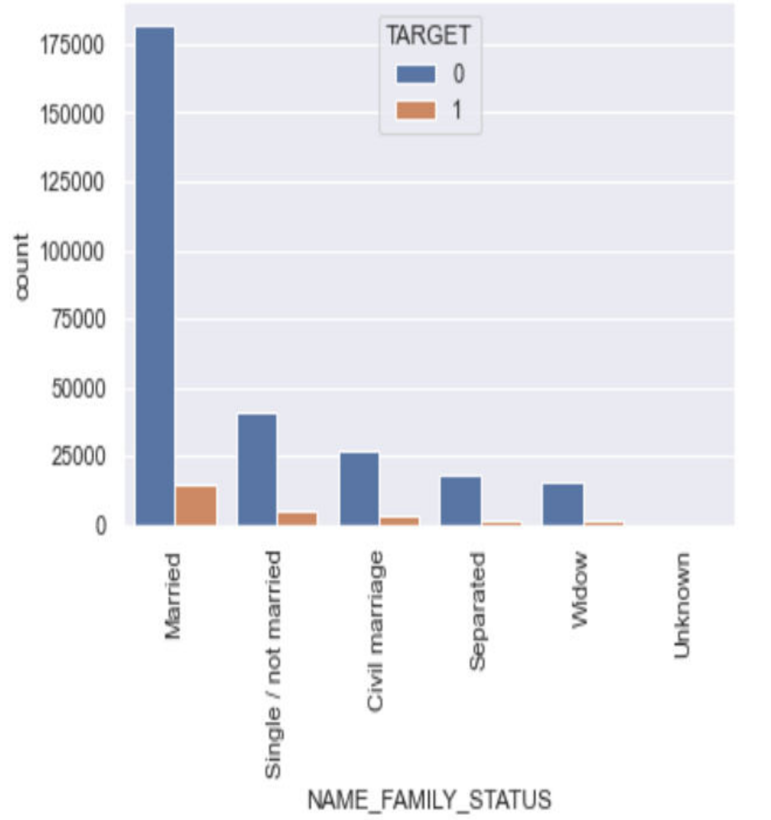
Clearly the business Income type is domaining the other business types with working which as a verity of range



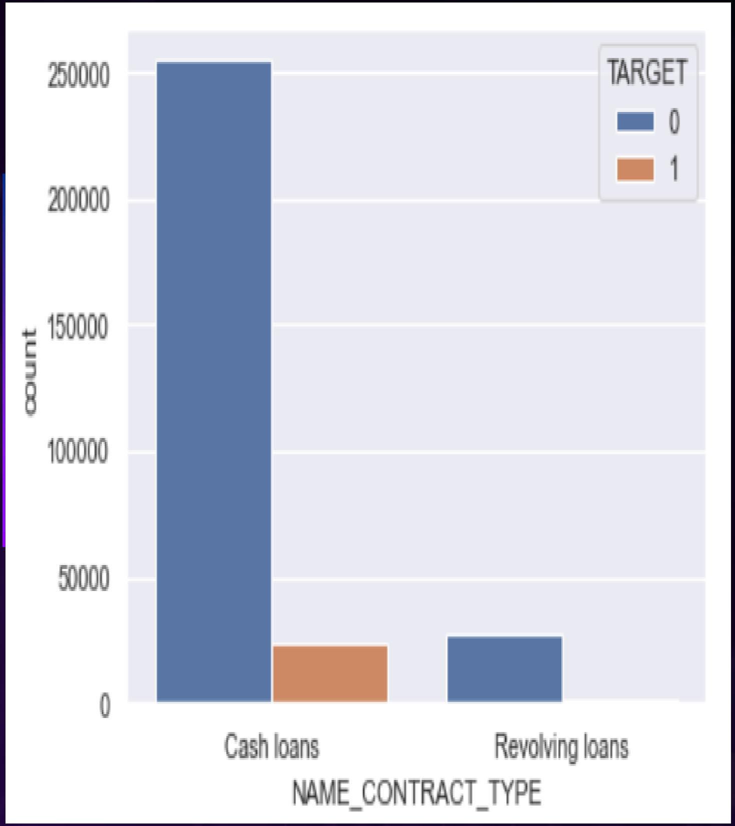


BIVARIATE ANALYSIS

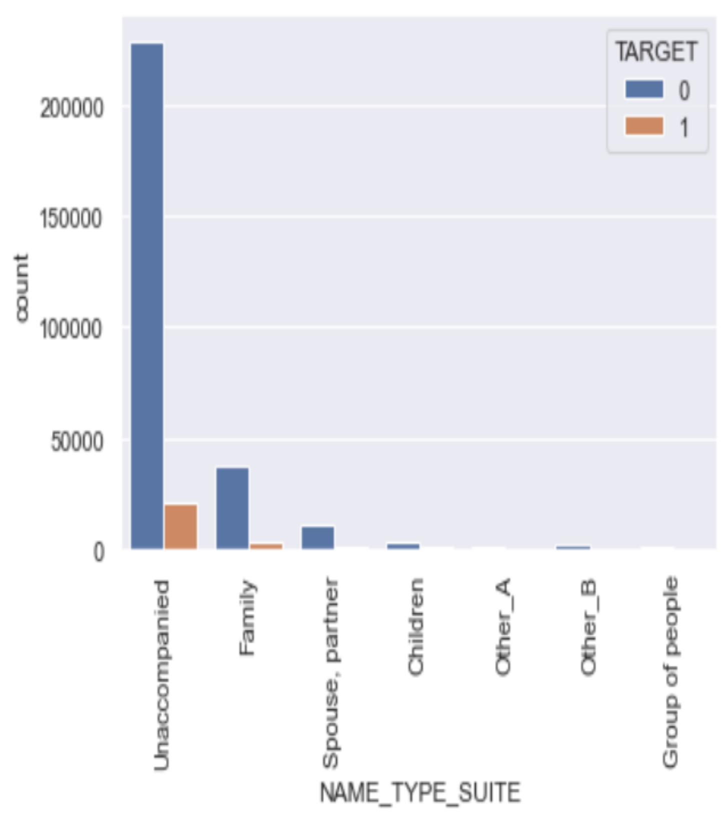
Married and single are top 2 category to target which has highest no. of non defaulters



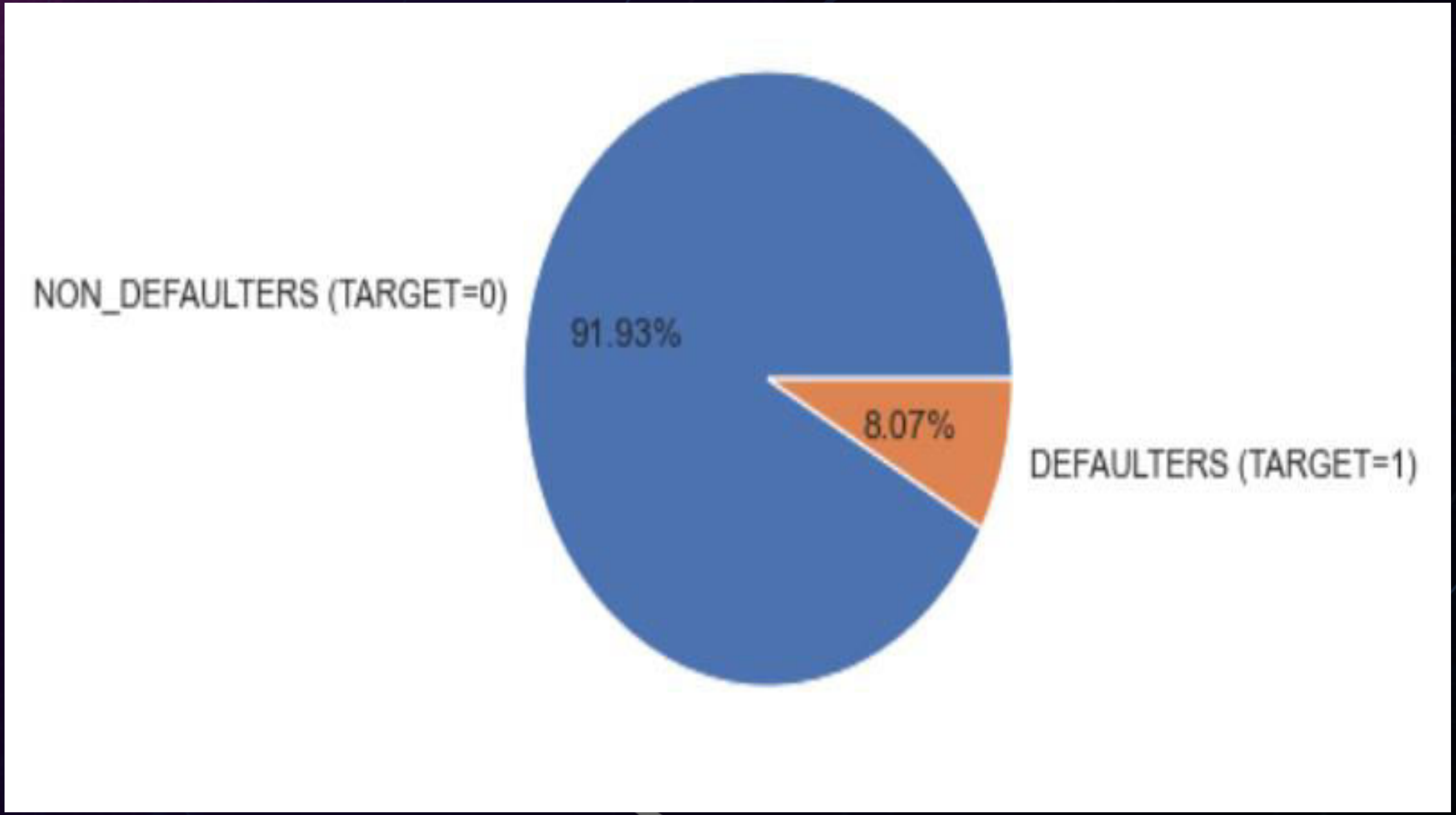
Cash loans have more non defaulters than revolving loans



Unaccompanied and family are the people who have less defaulters than other category



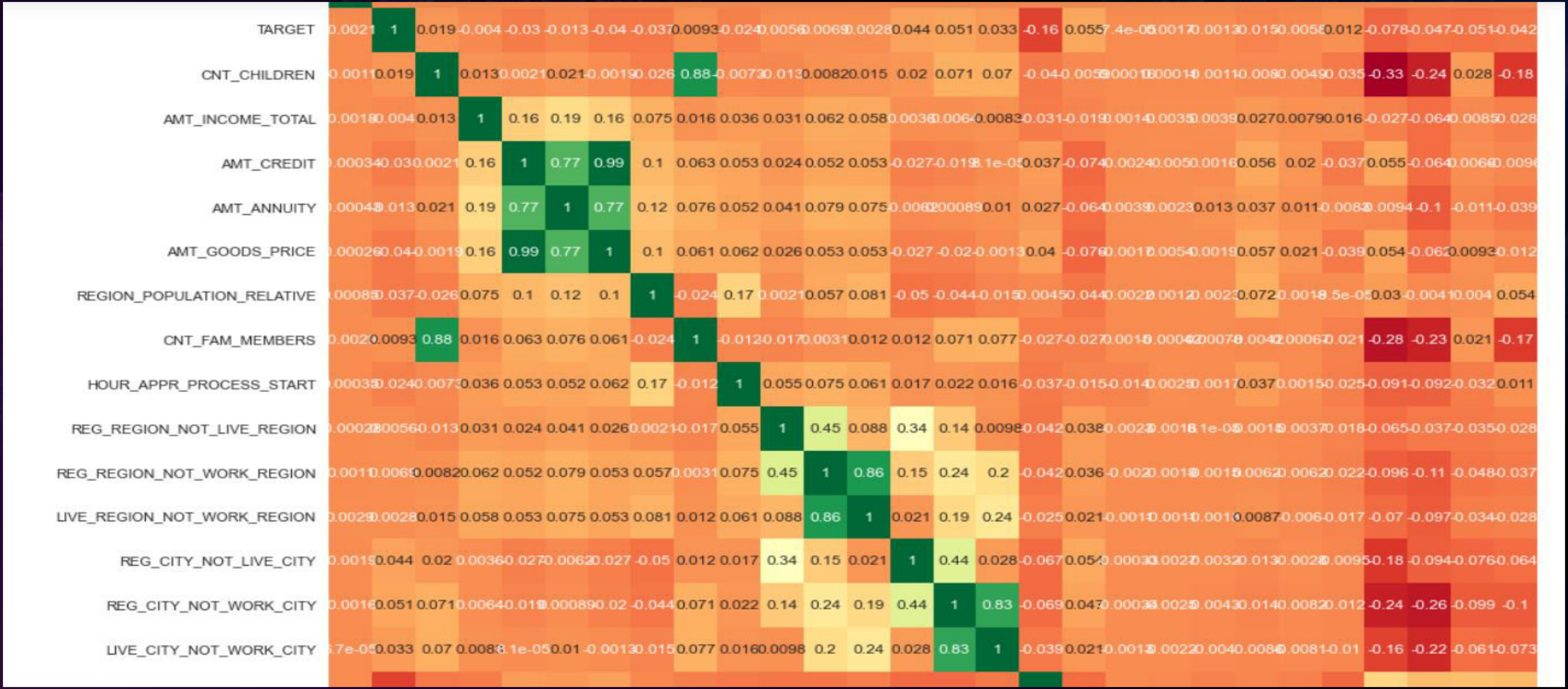
SEGMENTED VARIABLE



We can see there is a huge difference in the data
The Defaulter rate is less 8.07 % and Non-Defaulter rate is high 91.93%

MULTIVARIATE ANALYSIS

CHECKING For Co-relation ALL THE NUMERIC VARIABLE AT ONES



TARGET	0.0021	1	0.019	-0.004	-0.03	-0.013	-0.04	-0.037	0.0093	-0.024	0.0056	0.0069	0.0028	0.044	0.051	0.033	-0.16	0.055	7.4e-05	0.0017	0.0013	0.015	0.0058	0.012	-0.078	-0.047	-0.051	-0.042
CNT_CHILDREN	0.0011	0.019	1	0.013	0.0021	0.021	-0.0019	0.026	0.88	-0.0073	0.013	0.0082	0.015	0.02	0.071	0.07	-0.04	-0.0059	0.00016	0.00014	0.0011	0.008	0.0049	0.035	-0.33	-0.24	0.028	-0.18
AMT_INCOME_TOTAL	0.0018	0.004	0.013	1	0.16	0.19	0.16	0.075	0.016	0.036	0.031	0.062	0.058	0.0036	0.006	-0.0083	-0.031	-0.019	0.0014	0.0035	0.0039	0.027	0.0079	0.016	-0.027	-0.064	0.0085	0.028
AMT_CREDIT	0.00034	0.03	0.0021	0.16	1	0.77	0.99	0.1	0.063	0.053	0.024	0.052	0.053	-0.027	-0.019	8.1e-05	0.037	-0.074	0.0024	0.005	0.0016	0.056	0.02	-0.037	0.055	-0.064	0.0066	0.0096
AMT_ANNUITY	0.00043	0.013	0.021	0.19	0.77	1	0.77	0.12	0.076	0.052	0.041	0.079	0.075	0.0062	0.00089	0.01	0.027	-0.064	0.0039	0.0023	0.013	0.037	0.011	-0.0083	0.0094	-0.1	-0.011	-0.039
AMT_GOODS_PRICE	0.00026	0.04	-0.0019	0.16	0.99	0.77	1	0.1	0.061	0.062	0.026	0.053	0.053	-0.027	-0.02	-0.0013	0.04	-0.076	0.0017	0.0054	0.0019	0.057	0.021	-0.039	0.054	-0.062	0.0093	-0.012
REGION_POPULATION_RELATIVE	0.00085	0.037	-0.026	0.075	0.1	0.12	0.1	1	-0.024	0.17	0.0021	0.057	0.081	-0.05	-0.044	-0.015	0.0045	0.044	0.0022	0.0012	0.0023	0.072	0.0018	8.5e-05	0.03	-0.0041	0.004	0.054
CNT_FAM_MEMBERS	0.002	0.0093	0.88	0.016	0.063	0.076	0.061	-0.024	1	-0.012	-0.017	0.0031	0.012	0.012	0.071	0.077	-0.027	-0.027	0.0015	0.00042	0.00078	0.0042	0.00067	0.021	-0.28	-0.23	0.021	-0.17
HOUR_APPR_PROCESS_START	0.00035	0.024	0.0073	0.036	0.053	0.052	0.062	0.17	-0.012	1	0.055	0.075	0.061	0.017	0.022	0.016	-0.037	-0.015	-0.014	0.0025	0.0017	0.037	0.0015	0.025	-0.091	-0.092	-0.032	0.011
REG_REGION_NOT_LIVE_REGION	0.00028	0.0056	-0.013	0.031	0.024	0.041	0.026	0.0021	-0.017	0.055	1	0.45	0.088	0.34	0.14	0.0098	-0.042	0.038	0.0023	0.0016	.1e-05	0.0015	0.0037	0.018	-0.065	-0.037	-0.035	-0.028
REG_REGION_NOT_WORK_REGION	0.0011	0.0069	0.0082	0.062	0.052	0.079	0.053	0.057	0.0031	0.075	0.45	1	0.86	0.15	0.24	0.2	-0.042	0.036	-0.002	0.0018	0.0015	0.0062	0.0062	0.022	-0.096	-0.11	-0.048	-0.037
LIVE_REGION_NOT_WORK_REGION	0.0029	0.0028	0.015	0.058	0.053	0.075	0.053	0.081	0.012	0.061	0.088	0.86	1	0.021	0.19	0.24	-0.025	0.021	-0.001	0.001	0.001	0.0087	0.006	-0.017	-0.07	-0.097	-0.034	-0.028
REG_CITY_NOT_LIVE_CITY	0.0019	0.044	0.02	0.0036	-0.027	0.0062	0.027	-0.05	0.012	0.017	0.34	0.15	0.021	1	0.44	0.028	-0.067	0.054	0.0003	0.0027	0.0032	0.013	0.0028	0.0095	-0.18	-0.094	-0.076	-0.064
REG_CITY_NOT_WORK_CITY	0.0016	0.051	0.071	0.0064	-0.019	0.00089	0.02	-0.044	0.071	0.022	0.14	0.24	0.19	0.44	1	0.83	-0.069	0.047	0.0003	0.0025	0.0043	0.014	0.0082	0.012	-0.24	-0.26	-0.099	-0.1
LIVE_CITY_NOT_WORK_CITY	.7e-05	0.033	0.07	0.0083	.1e-05	0.01	-0.0013	0.015	0.077	0.016	0.0098	0.2	0.24	0.028	0.83	1	-0.039	0.021	0.0013	0.0022	0.004	0.0086	0.0081	-0.01	-0.16	-0.22	-0.061	-0.073

FINAL OBSERVATION

i. Target variable for Application dataset - "TARGET“

ii. Target variable for Previous dataset - "NAME_CONTRACT_STATUS"

iii. The rate of defaulters are less in the range of 20-40 & 40-60 are good target audience.

iv. Laborers , Core and Sales Staff is the occupation type that has the loan approved and has the highest non defaulter rate.

v. Married people are more likely to get loan approved in comparison to any other Marital Status of the people so this is also a good target audience .

vi. Secondary Education has the Highest Approval rate ,although the Income of Academic degree holder are more as Compare to Secondary education still the approval rate is more than Academic Degree holders.

THANK YOU