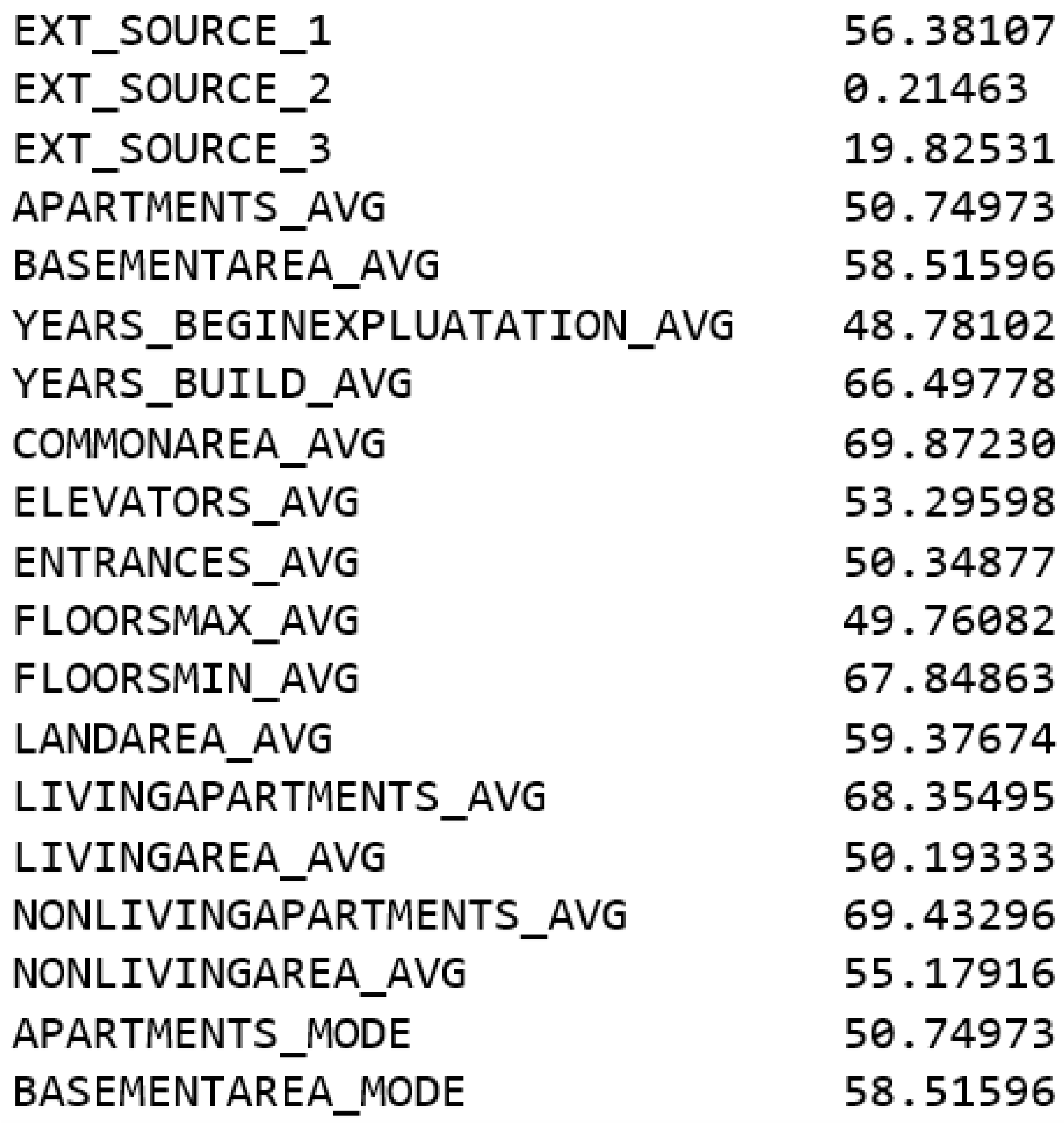
Credit EDA Case Study

Team Members:
Chandrashekhar Vilas Purandare
Vishnuvardhan Reddy Bekkem

Data Quality checks and cleanup

- Identification of the % of missing values for each column
- Drop columns with more than 50% of missing values
- Remove irrelevant columns referring to Data Dictionary
- Impute missing values using Mean, Median & Mode
- Change datatype as appropriate
- Binning of continuous data into categorial bins

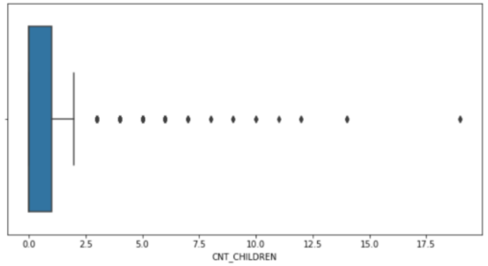


```
EXT_SOURCE_1                    56.38107
EXT_SOURCE_2                    0.21463
EXT_SOURCE_3                    19.82531
APARTMENTS_AVG                  50.74973
BASEMENTAREA_AVG                58.51596
YEARS_BEGINEXPLUATATION_AVG     48.78102
YEARS_BUILD_AVG                 66.49778
COMMONAREA_AVG                  69.87230
ELEVATORS_AVG                   53.29598
ENTRANCES_AVG                   50.34877
FLOORSMAX_AVG                   49.76082
FLOORSMIN_AVG                   67.84863
LANDAREA_AVG                    59.37674
LIVINGAPARTMENTS_AVG            68.35495
LIVINGAREA_AVG                  50.19333
NONLIVINGAPARTMENTS_AVG         69.43296
NONLIVINGAREA_AVG               55.17916
APARTMENTS_MODE                 50.74973
BASEMENTAREA_MODE               58.51596
```

Outlier treatment

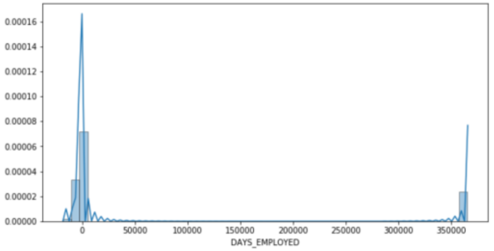
- Plot 1 highlights the outlier values for the client who have children more than 2.
- Plot 2 we can observe that high number of clients have been working more 300K days which we can consider as outlier values and this can be removed for further analysis.
- The bottom table make it very clear using mean and percentile distribution for columns having outliers such as 19 children or 1000 years of employment duration.

Plot 1

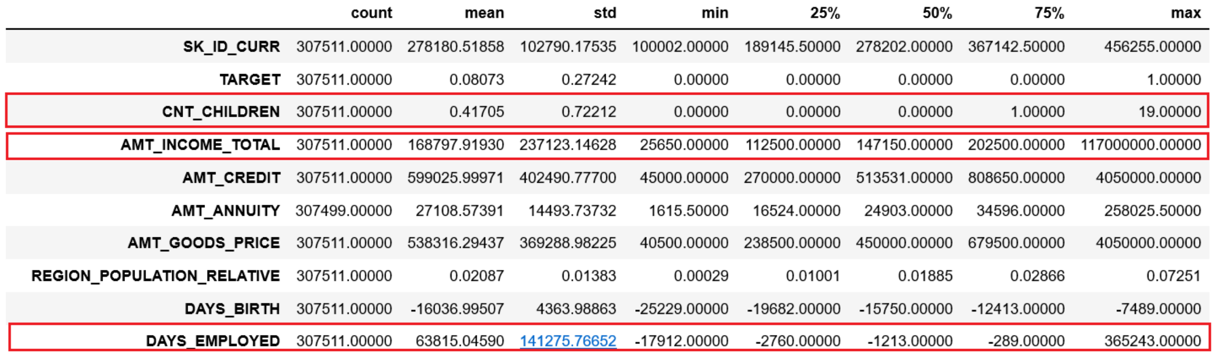


Plot for Children outlier

Plot 2

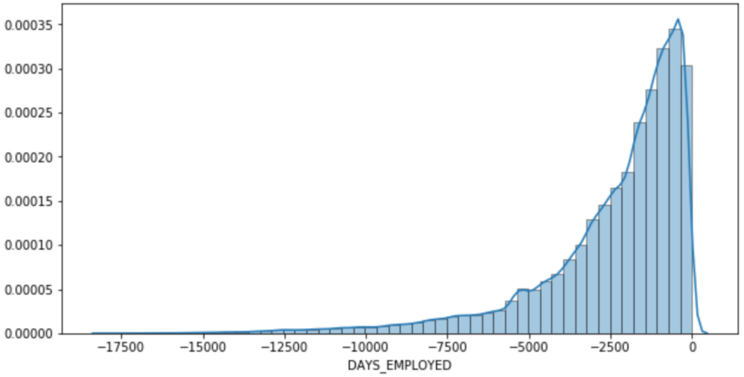


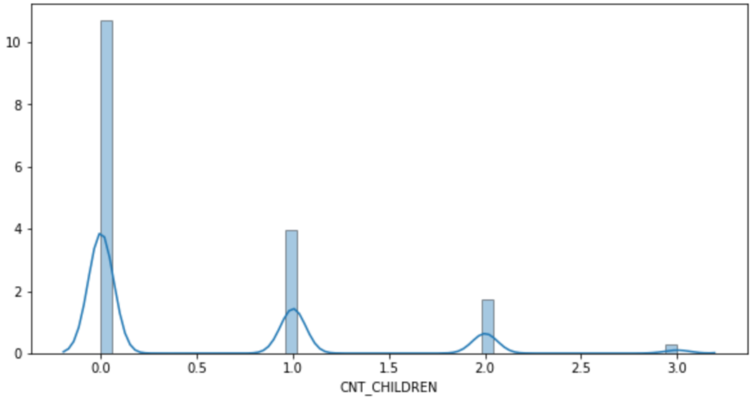
Plot for Days employed

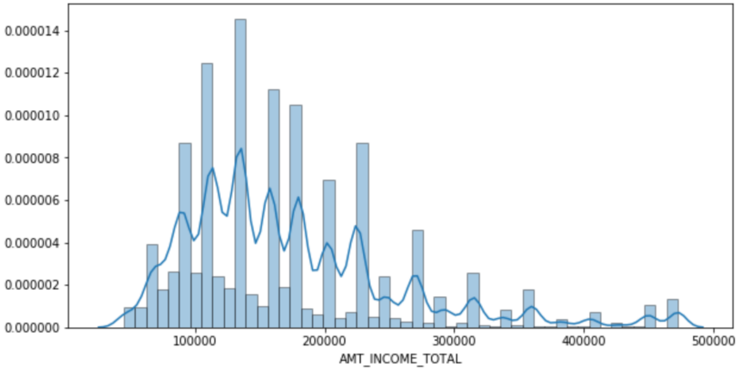


	count	mean	std	min	25%	50%	75%	max
SK_ID_CURR	307511.00000	278180.51858	102790.17535	100002.00000	189145.50000	278202.00000	367142.50000	456255.00000
TARGET	307511.00000	0.08073	0.27242	0.00000	0.00000	0.00000	0.00000	1.00000
CNT_CHILDREN	307511.00000	0.41705	0.72212	0.00000	0.00000	0.00000	1.00000	19.00000
AMT_INCOME_TOTAL	307511.00000	168797.91930	237123.14628	25650.00000	112500.00000	147150.00000	202500.00000	117000000.00000
AMT_CREDIT	307511.00000	599025.99971	402490.77700	45000.00000	270000.00000	513531.00000	808650.00000	4050000.00000
AMT_ANNUITY	307499.00000	27108.57391	14493.73732	1615.50000	16524.00000	24903.00000	34596.00000	258025.50000
AMT_GOODS_PRICE	307511.00000	538316.29437	369288.98225	40500.00000	238500.00000	450000.00000	679500.00000	4050000.00000
REGION_POPULATION_RELATIVE	307511.00000	0.02087	0.01383	0.00029	0.01001	0.01885	0.02866	0.07251
DAYS_BIRTH	307511.00000	-16036.99507	4363.98863	-25229.00000	-19682.00000	-15750.00000	-12413.00000	-7489.00000
DAYS_EMPLOYED	307511.00000	63815.04590	141275.76652	-17912.00000	-2760.00000	-1213.00000	-289.00000	365243.00000

Outlier Treatment Results





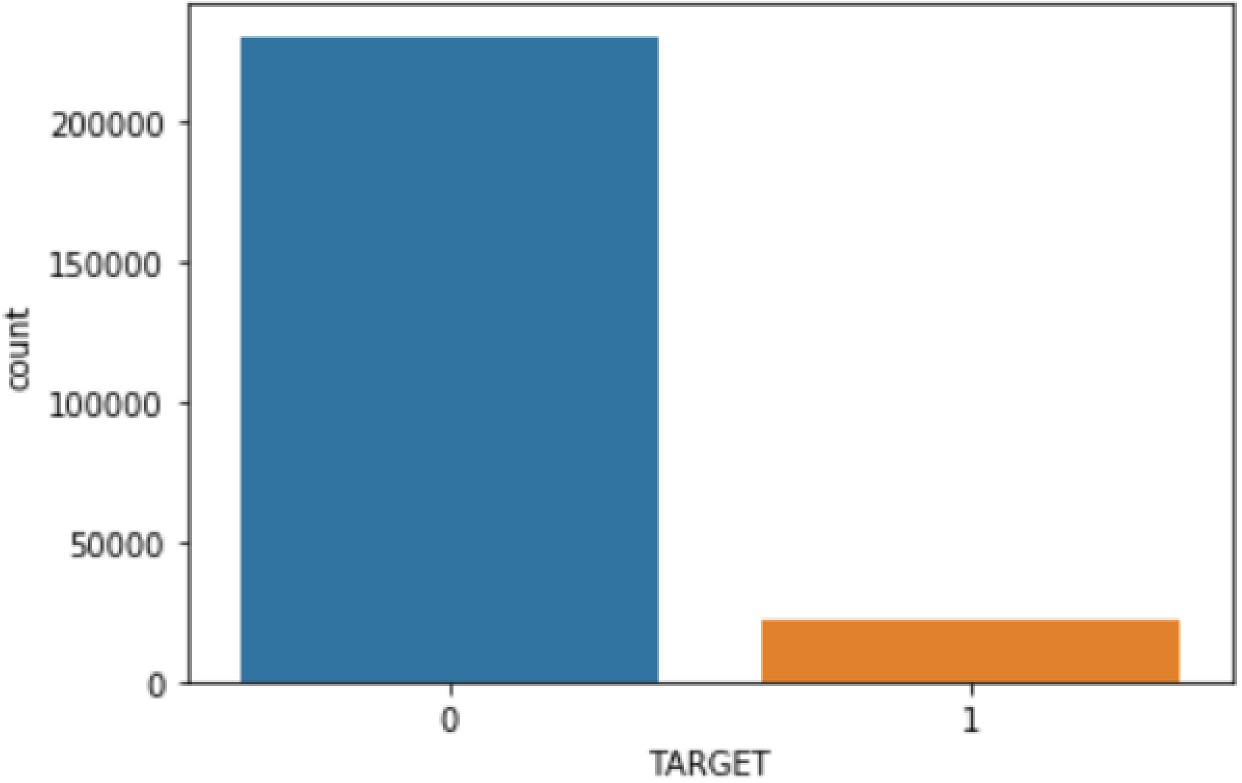


Conclusion: Above 3 plots are related to clients working experience, number of children's and total income of client.

Maximum loan applicants are in early years of their employment ,The majority of loan applicants don't have any child and Majority of the applicants income is between 45K to 300K.

Data Imbalance Analysis

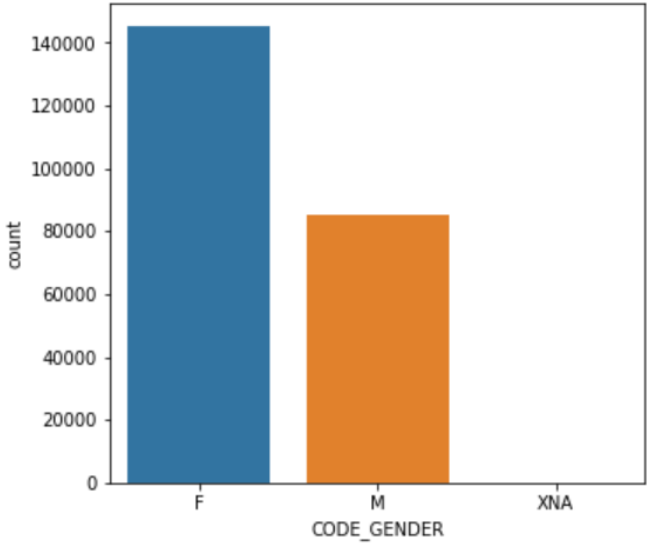
- It is clear from the shown graph and percent distribution for Target column is a highly imbalanced.
- The ratio of dataset imbalance is 91.34 : 8.66
- Hence, we have decided to split this dataset for further analysis.



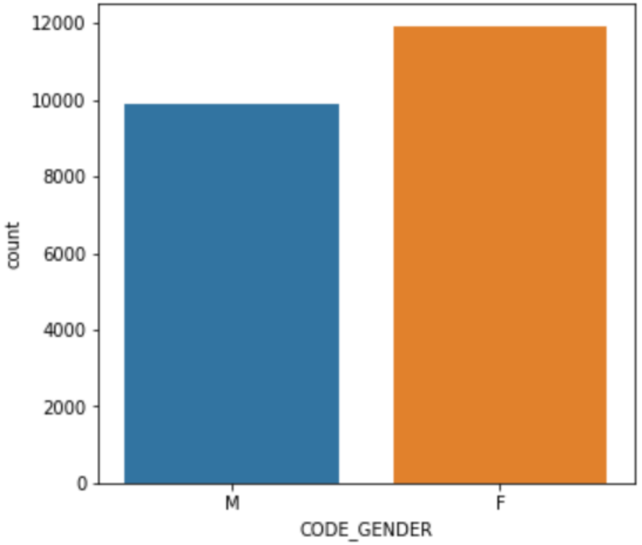
Univariate Analysis - Gender

- From the plot it can be clearly observed that the number of female clients in non-defaulted dataset is almost double of male clients.
- Ratio of defaulted loans between male and female is 11.66 : 8.21
- That indicates male clients have a higher chance of not returning their loans (12%), comparing with female clients (8%).

Univariate Analysis for Categorical variables for both datasets (target 0 and 1)



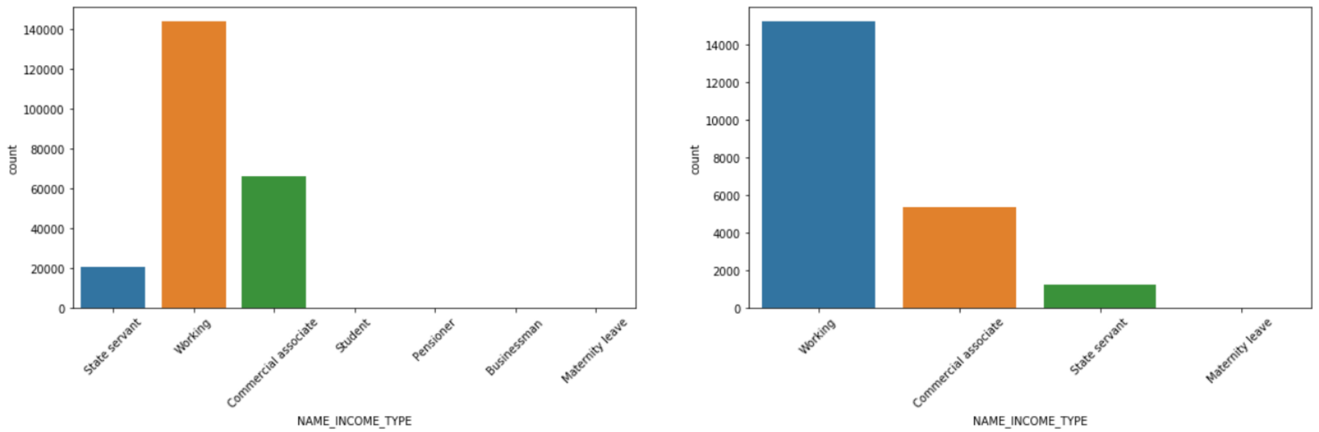
Plot 1(Target 0)



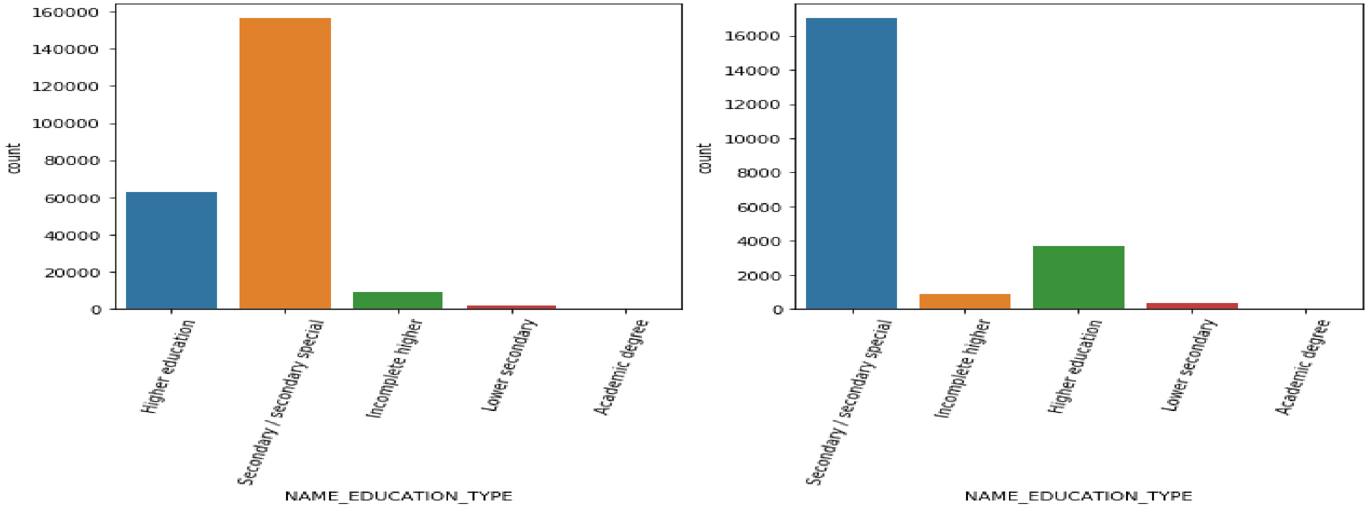
Plot 2(Target 1)

Univariate Analysis – Income Type

- From Income type graph it can be observed that Applicants having Income Types as Working and Commercial associate has the highest percentage of clients.
- From Education type graph we can observe that a large number of applications (approx. 150K) are filed by people having secondary education followed by people with Higher Education.



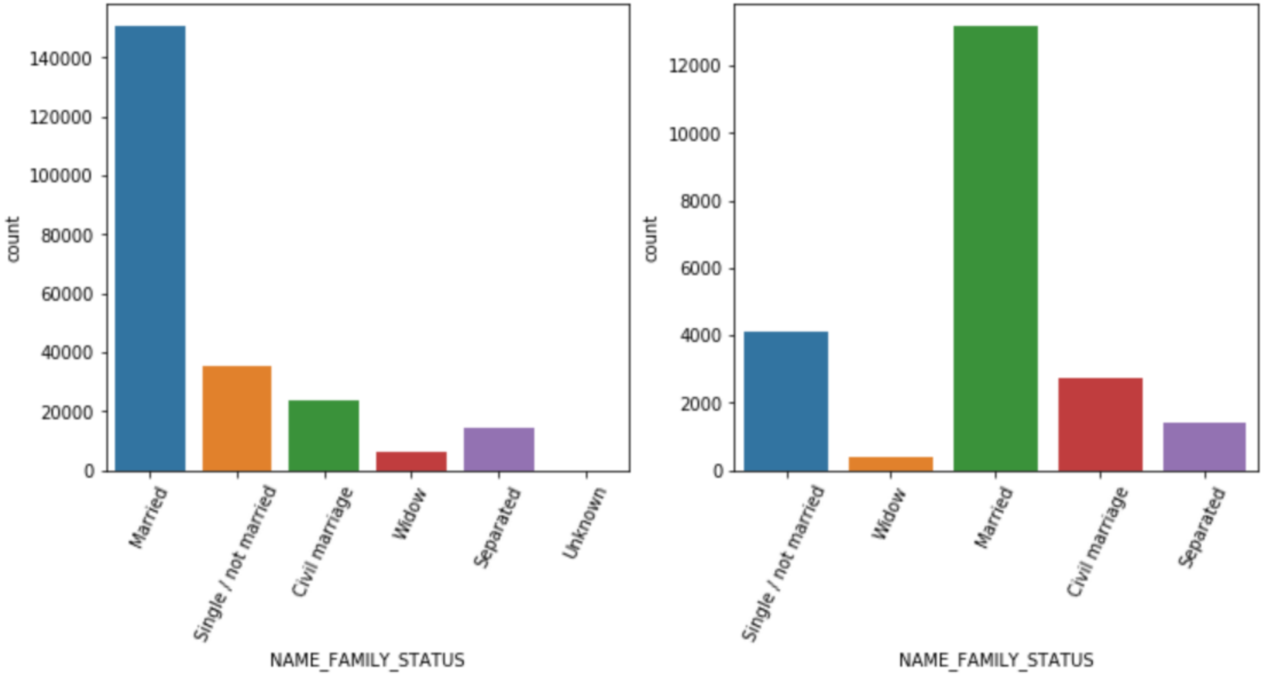
Graph for Income Type



Graph for Education type

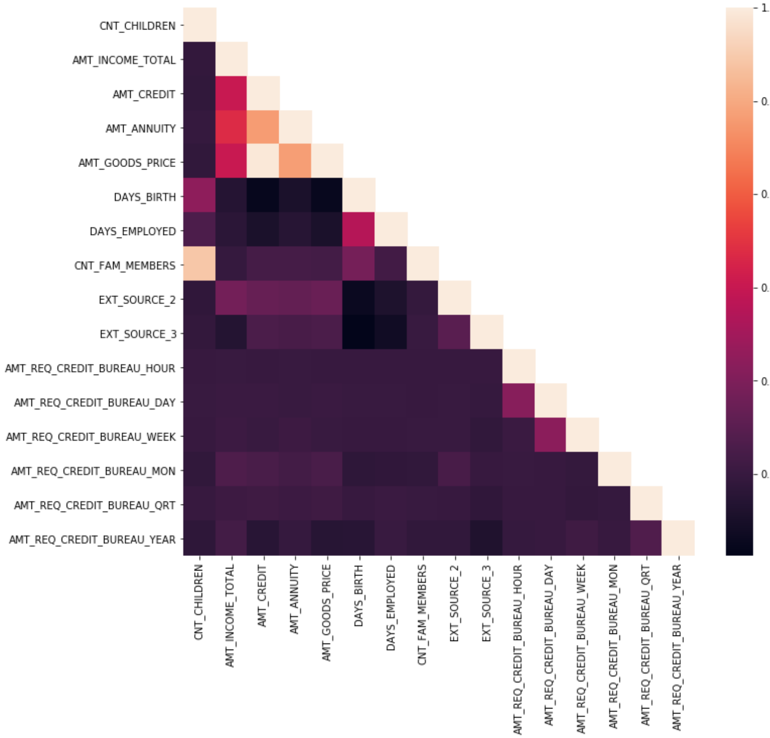
Univariate Analysis – Categorial Data

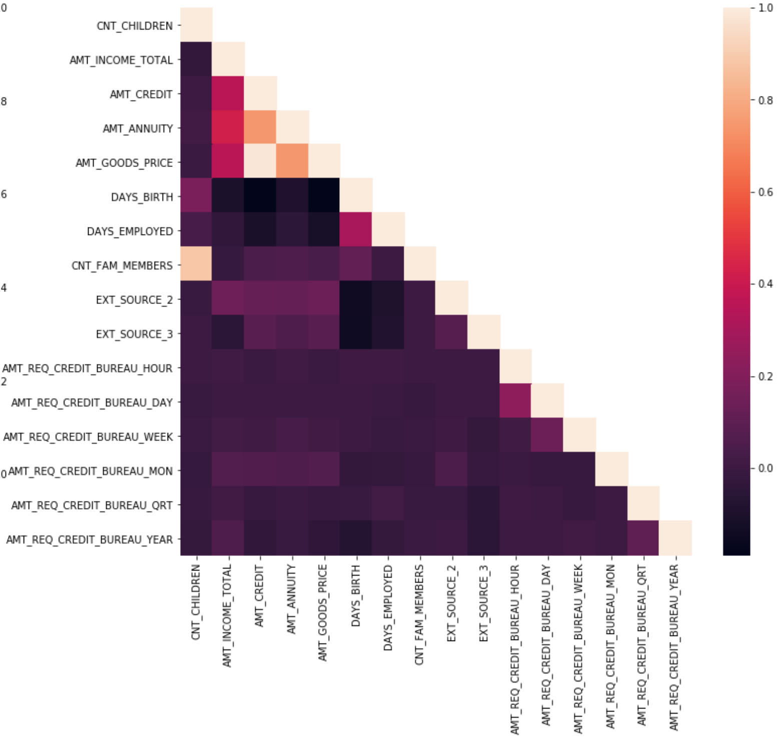
- As seen in the graphs, majority of clients are married followed by Single (not married) and Civil marriage.
- Client with Civil marriage category has the highest percent of repayment difficulties and clients in Widow category has the lowest.



Data Quality checks and cleanup

- Based on analysis of side by side heat maps showing correlation for both targets 0 and 1, we can conclude that the highest correlation features are similar in both the datasets.



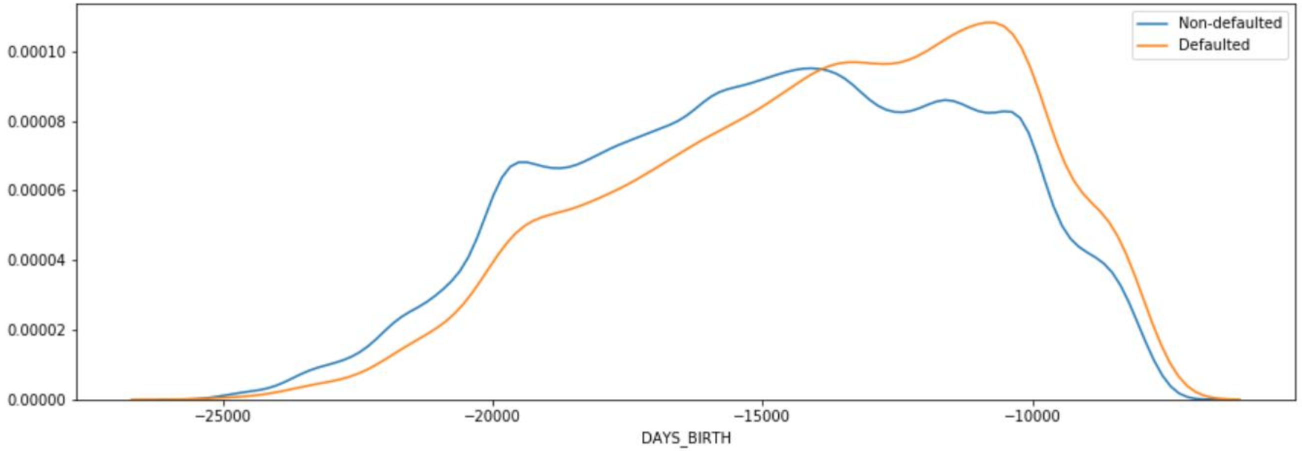


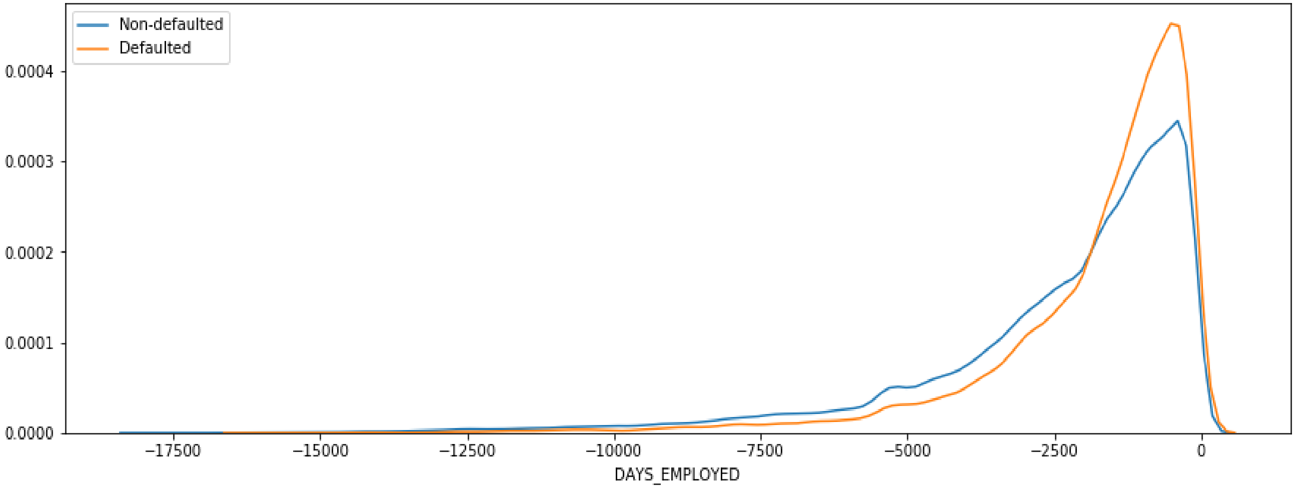
Top 10 Highly Corelated Features

- REGION_RATING_CLIENT
- REGION_RATING_CLIENT_W_CITY
- REGION_POPULATION_RELATIVE
- DAYS_EMPLOYED
- DAYS_BIRTH
- EXT_SOURCE_2
- EXT_SOURCE_3
- AMT_CREDIT
- AMT_INCOME_TOTAL
- AMT_GOODS_PRICE

Univariate Analysis – Client Age

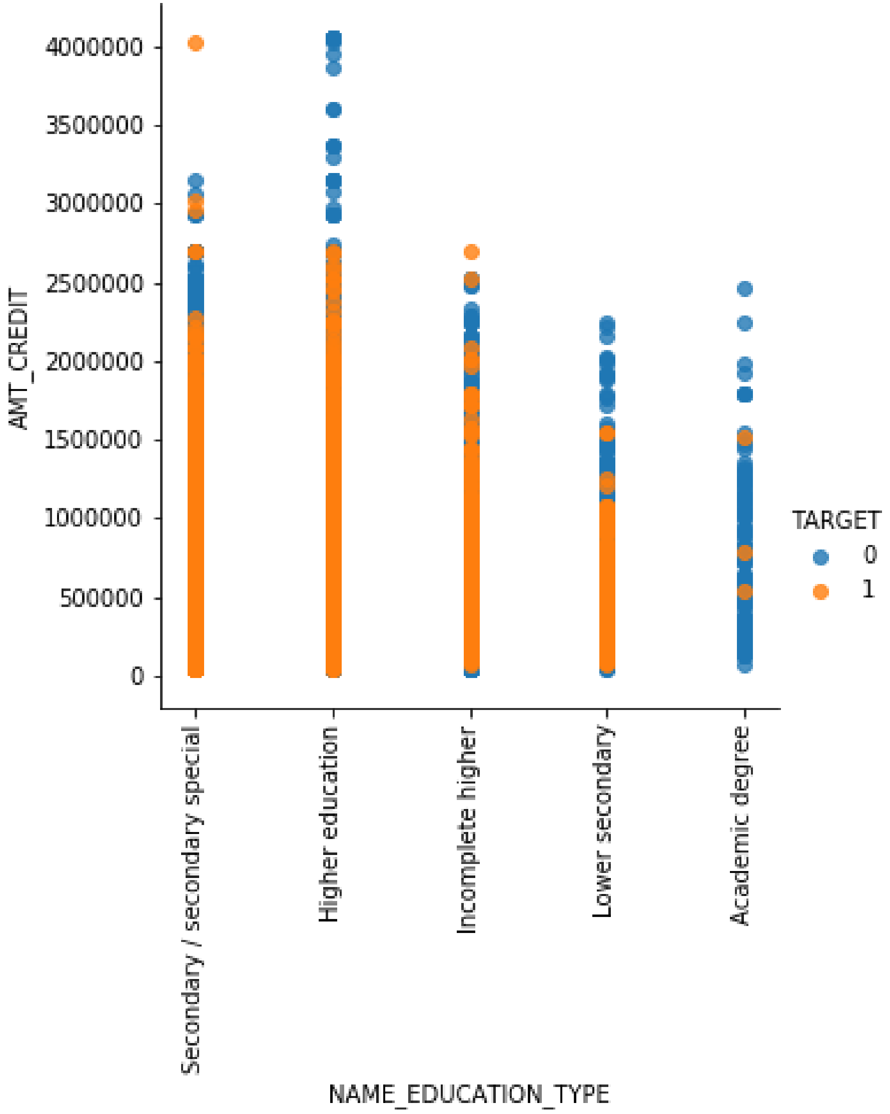
- As seen in the given graph, majority of clients are married followed by Single (not married) and Civil marriage. Client with Civil marriage category has the highest percent of repayment difficulties and clients in Widow category has the lowest.
- The employment duration for applicants is approximately ranges between 0 to 30 years. Looking at the spike in the graph we can conclude that majority of applicants between 0-6 year are having difficulties in making repayment of load. The number of defaulter get reduced as the employment duration increased.





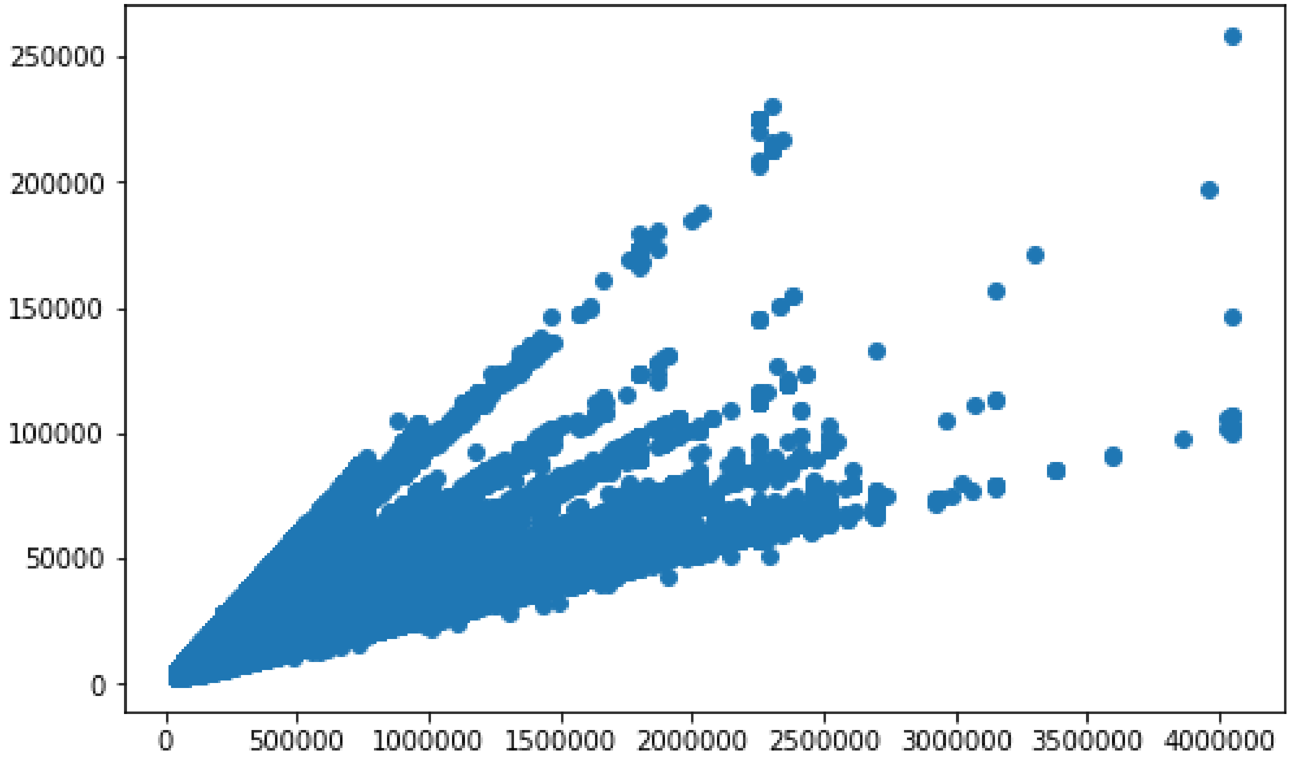
Bivariate Analysis – Loan amount & Education type

- Highest credit amount (400K+) is given to client with higher education followed by Secondary education level.
- Lowest loan amount is provided to clients in Lower Secondary education type
- Clients in Academic degree has lowest loan defaulted percentage



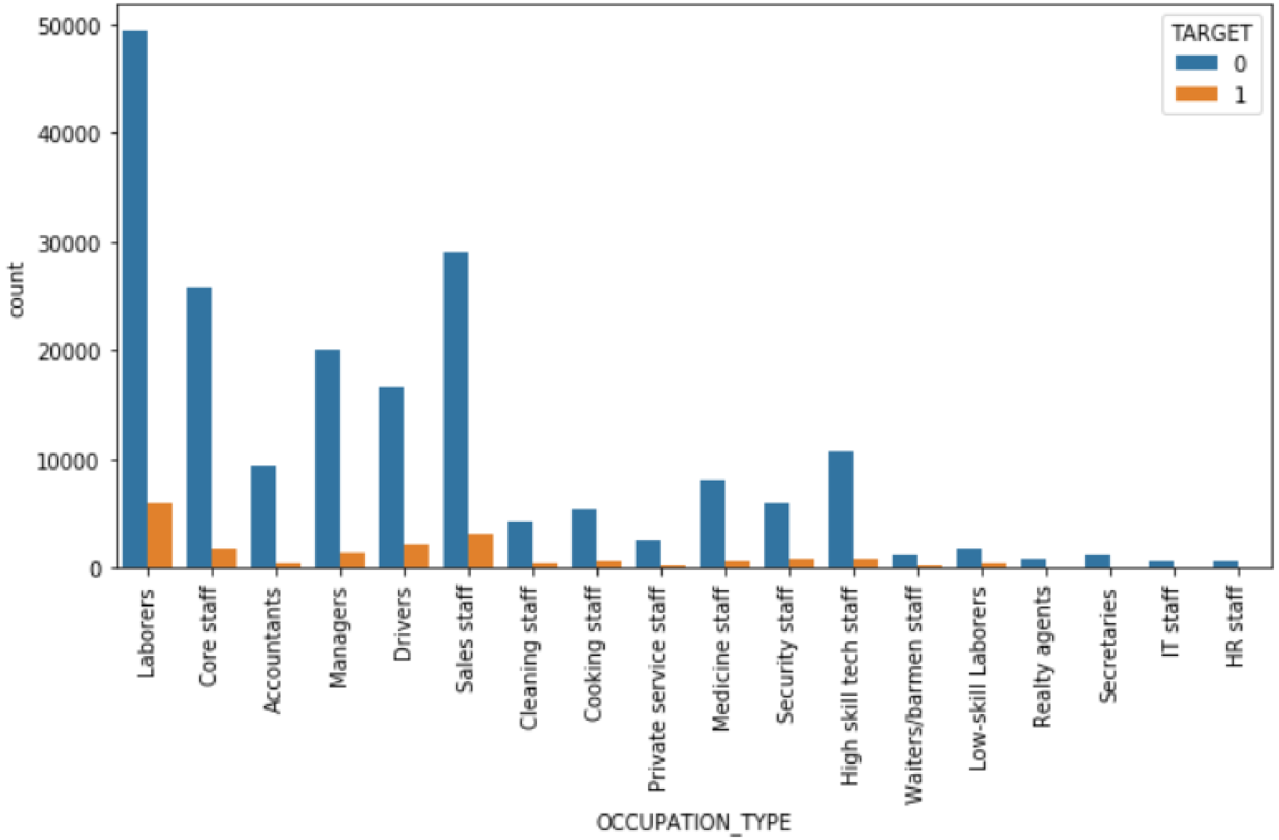
Bivariate Analysis – Loan amount & Loan annuity

- There is positive correlation between loan amount and annuity amount.
- This indicates that larger the amount of annuity, client will get that much of bigger loan sanctioned
- Clients in Academic degree has lowest loan defaulted percentage



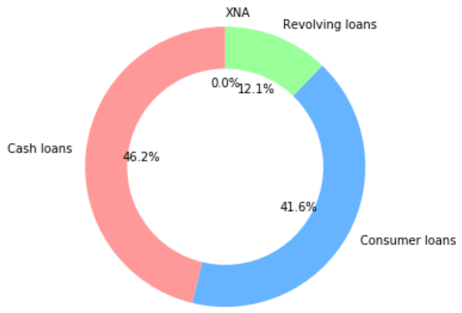
Bivariate Analysis – Loan defaulted & Occupation Type

- There is positive correlation between loan amount and annuity amount.
- This indicates that larger the amount of annuity, client will get that much of bigger loan sanctioned
- Clients in Academic degree has lowest loan defaulted percentage

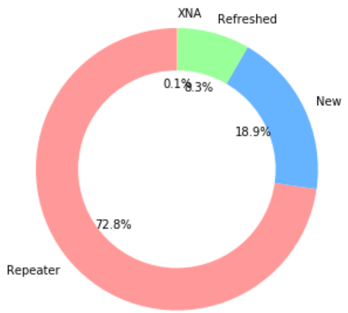


Univariate Analysis on Merged Data – Contract Status, Contract Type, Client Type & Insurance

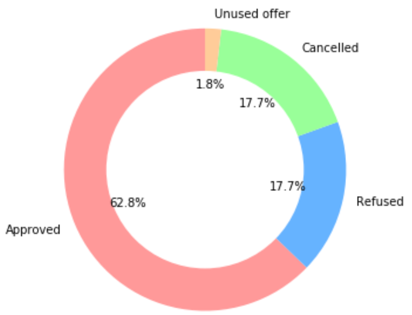
- There is positive correlation between loan amount and annuity amount.
- This indicates that larger the amount of annuity, client will get that much of bigger loan sanctioned
- Clients in Academic degree has lowest loan defaulted percentage



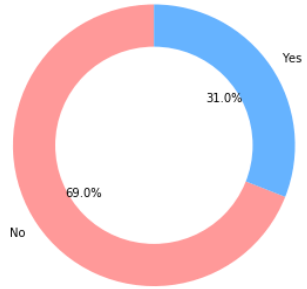
Contract Type



Client Type



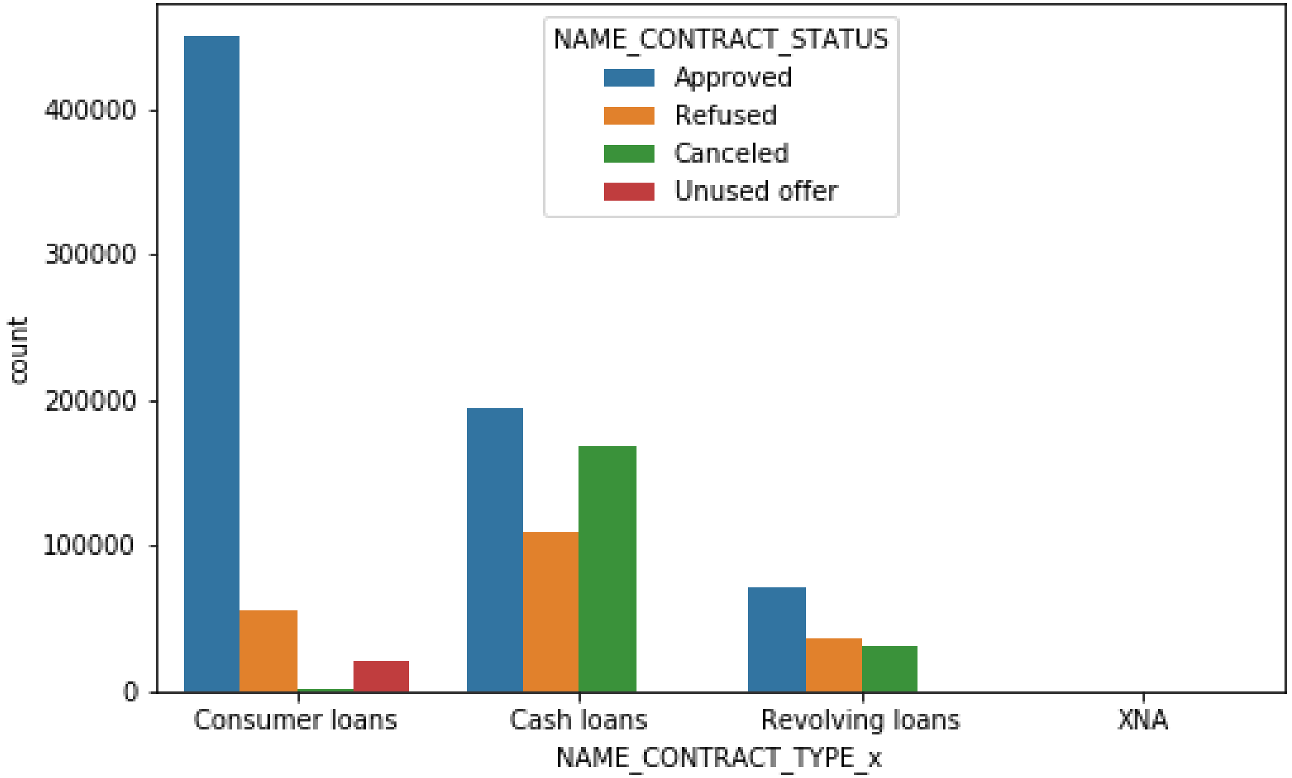
Contract Status



Insured Flag

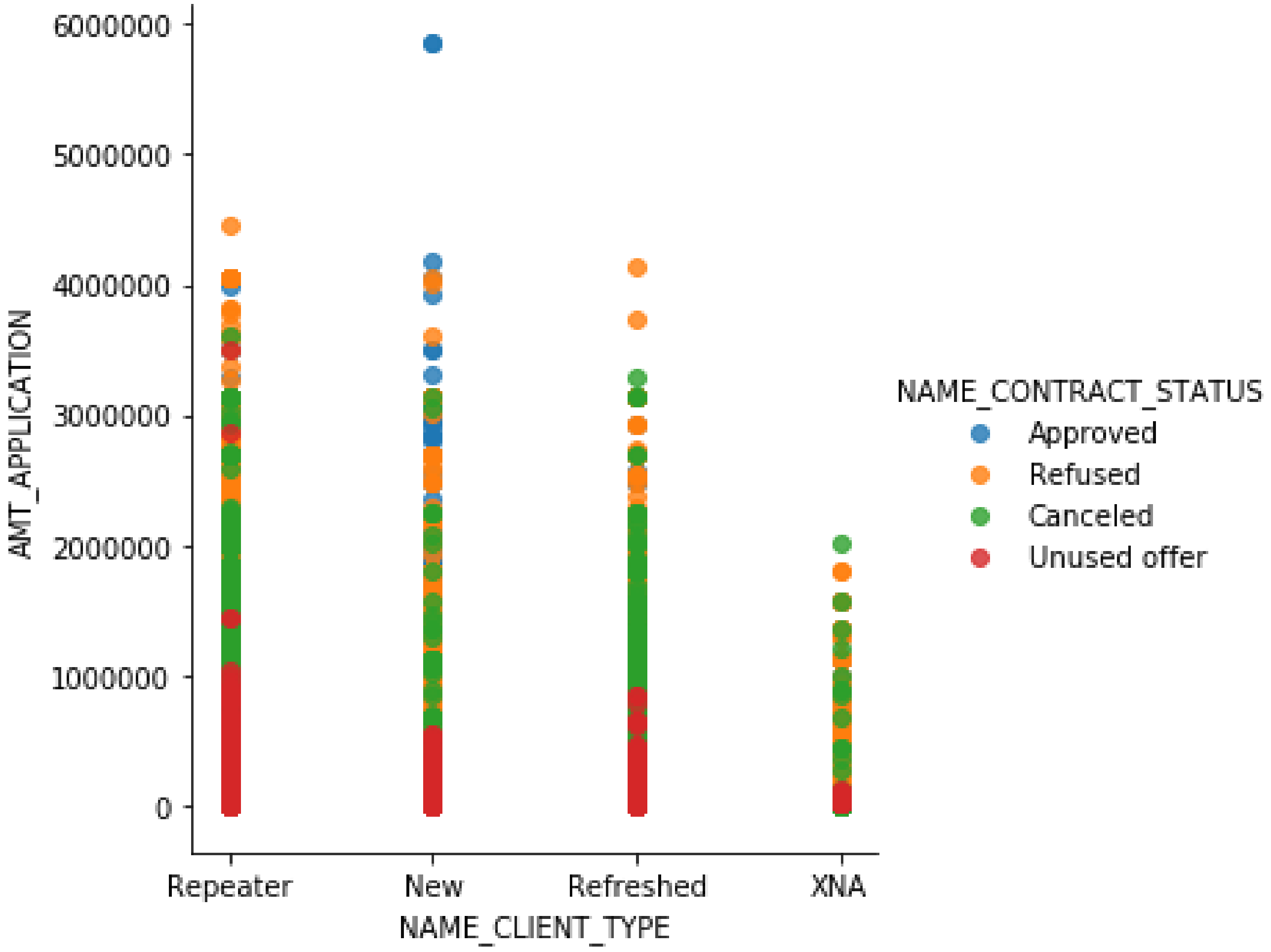
Bivariate Analysis on merged data – Application Amount & Client Type

- The largest percent of approved loans belongs to Consumer loan category followed by Cash loans.
- The majority of loans are contributed by personal loan and consumer durable loans such as loan for buying electronics appliances or gadgets.



Bivariate Analysis on merged data – Application Amount & Client Type

- A large portion of application amount is contributed by Repeater client followed by new client type.



Thank you