Bank Loan Case Study

Final Project 2

Project By: Sumit Gope

Project Description:

The aim of this project is to apply exploratory data analysis (EDA) techniques to understand the patterns present in loan application data and identify the driving factors behind loan default. The dataset used in this project contains information about loan applications at the time of applying for the loan, as well as previous loan data for clients.

The problem statement for this project is to identify patterns that indicate if a client has difficulty paying their installments, which can be used to take actions such as denying the loan, reducing the amount of loan, or lending to risky applicants at a higher interest rate. The company wants to understand the driving factors behind loan default, i.e. the variables that are strong indicators of default, which can be used for portfolio and risk assessment.

Approach

The analysis approach for this project involves the following steps:

1.Data Understanding: This involves understanding the structure and contents of the dataset and data dictionary provided, and identifying any missing data and outliers.

2.Data Cleaning: This involves dealing with missing data and outliers using appropriate methods, such as removing columns or replacing missing data with an appropriate value.

3.Univariate Analysis: This involves analyzing individual variables in the dataset and identifying any trends or patterns.

4.Bivariate Analysis: This involves analyzing the relationship between two variables in the dataset and identifying any trends or patterns.

5.Top Correlation Analysis: This involves finding the top 10 correlations for the client with payment difficulties and all other cases, segmented by the target variable, to identify the driving factors behind loan default.

Tech-Stack Used

Excel 2010 and Microsoft 365 web

Power BI

PowerPoint

Overall Approach of analysis:

The bank's problem statement is to identify the major cause of bank loan default. So below are the data sets provided to analysis.

1. *application_data.csv* contains all of the client's information at the time of application. The information related to whether or not a client is having financial issues.

2. *previous_application.csv* provides data from the client's previous loans. It indicates if the prior application was Accepted , cancelled Refused or Unused.

Both sets of data contained many undesired columns that will not be used for risk analytics as well as many blanks. So I cleaned up the data.

Data sets contains Categorical variables as well as Numerical variables.

Insights (Application_data.csv)

Data Cleaning

The dataset contained 3 lac+ rows and 161 columns of data.

I deleted numbers of columns containing more than 45% of blank data.

For this I used =*countblank()* function.

After got total blank rows of each column I divided total number of column and multiply by 100 so we got percentage of blank column.

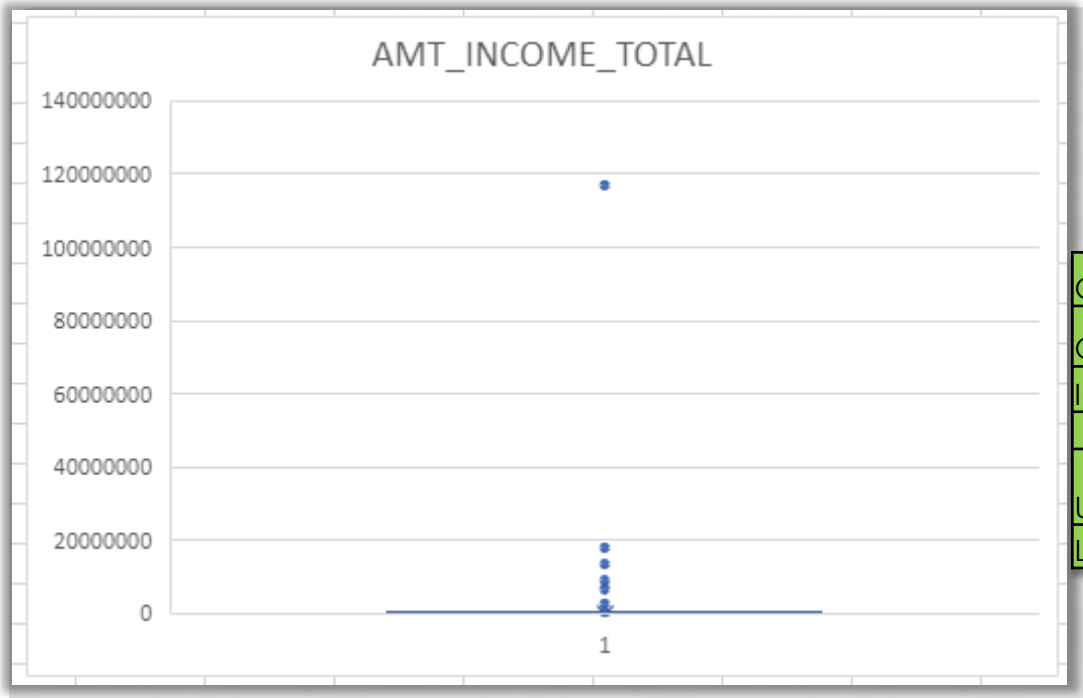
After deleting unwanted columns we got 72 remaining columns.

Outliers

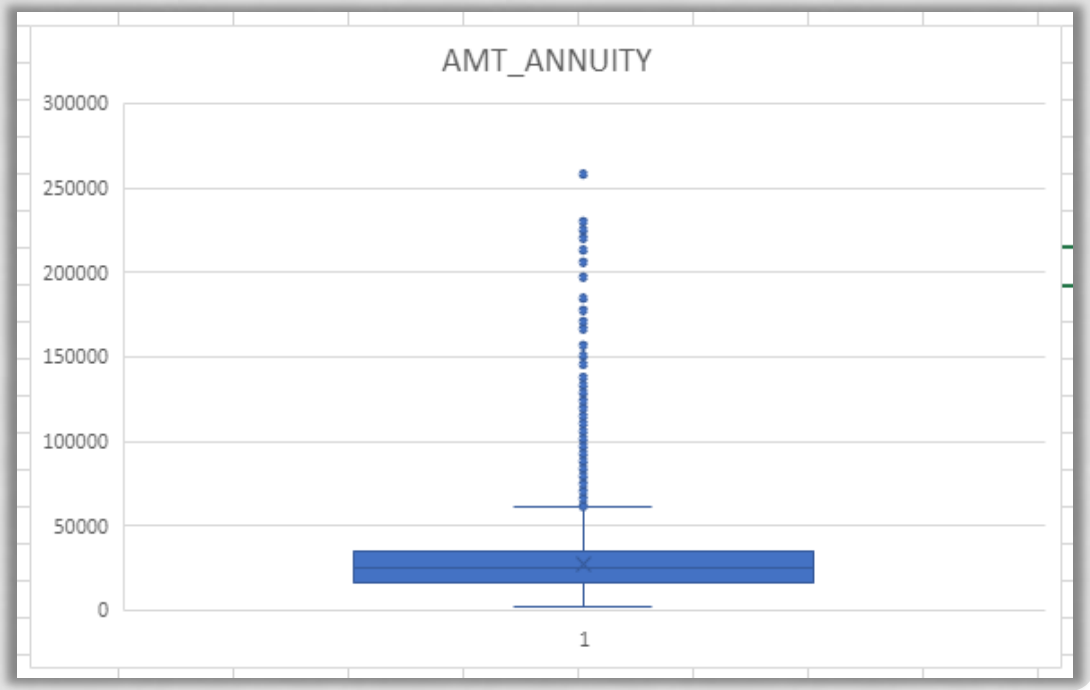
Outliers can only be identified on Numerical variable.
To find outlier basically I used Box and Whisker chart.

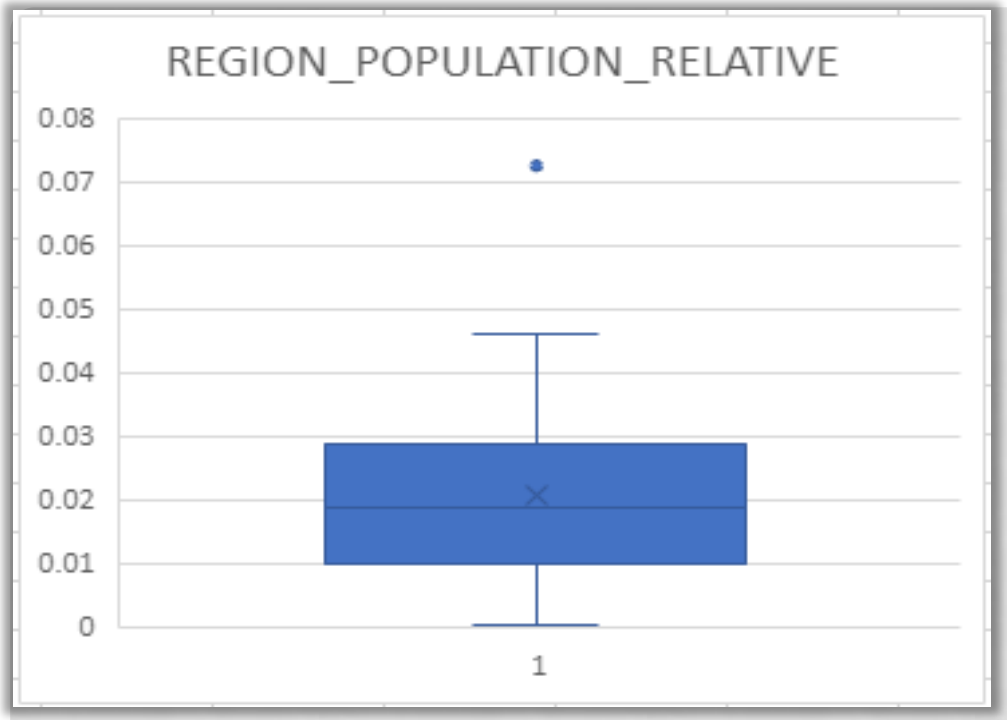
Statistical formula on excel that, why it is outlier.

Q1	270000
Q3	808650
IQR (Q3-Q1)	538650
UPPER BOUND [Q3+(1.5*IQR)]	1616625
LOWER BOUND [Q1-(1.5*IQR)]	-942975



Q1	112500
Q3	202500
IQR (Q3-Q1)	90000
UPPER BOUND [Q3+(1.5*IQR)]	337500
LOWER BOUND [Q1+(1.5*IQR)]	-22500



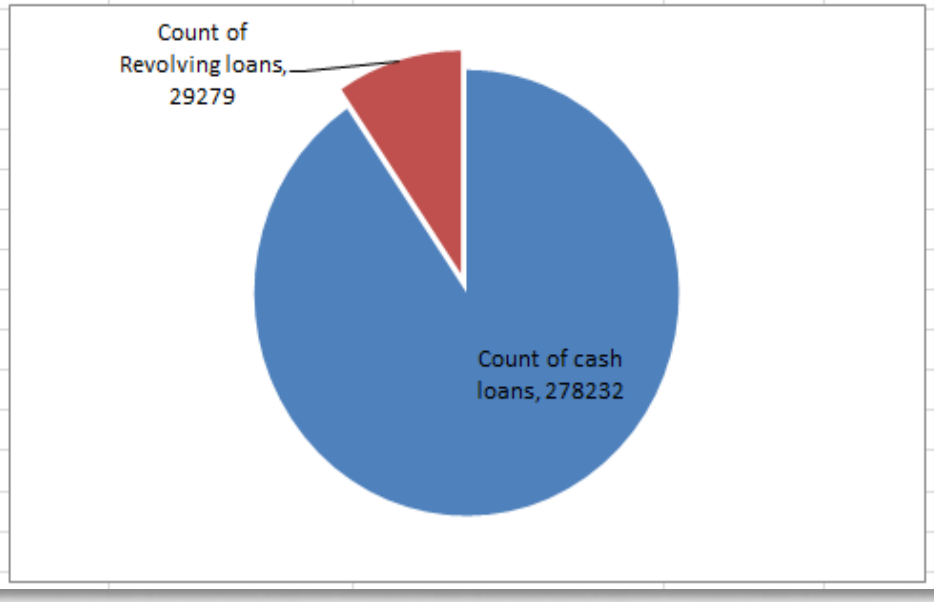

REGION_POPULATION_RELATIVE
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1


AMT_GOODS_PRICE
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1

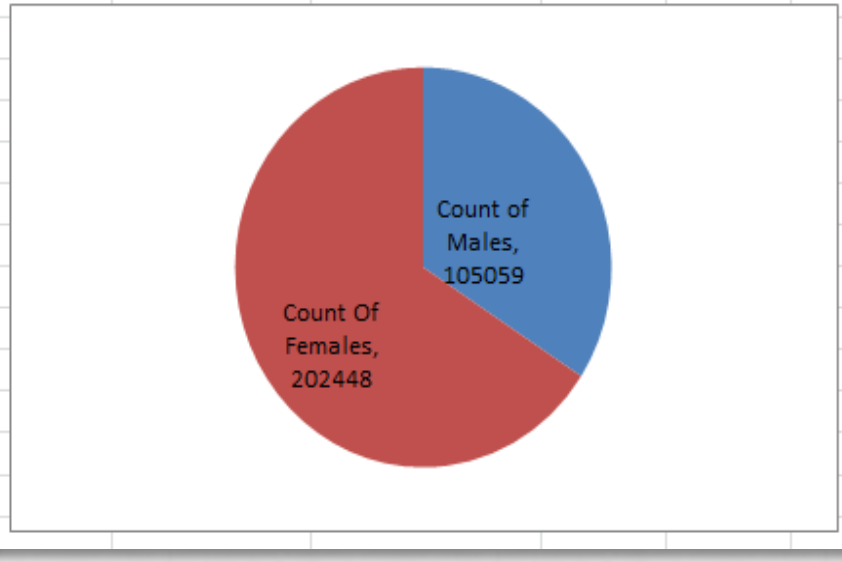
Data Imbalance and Ratio

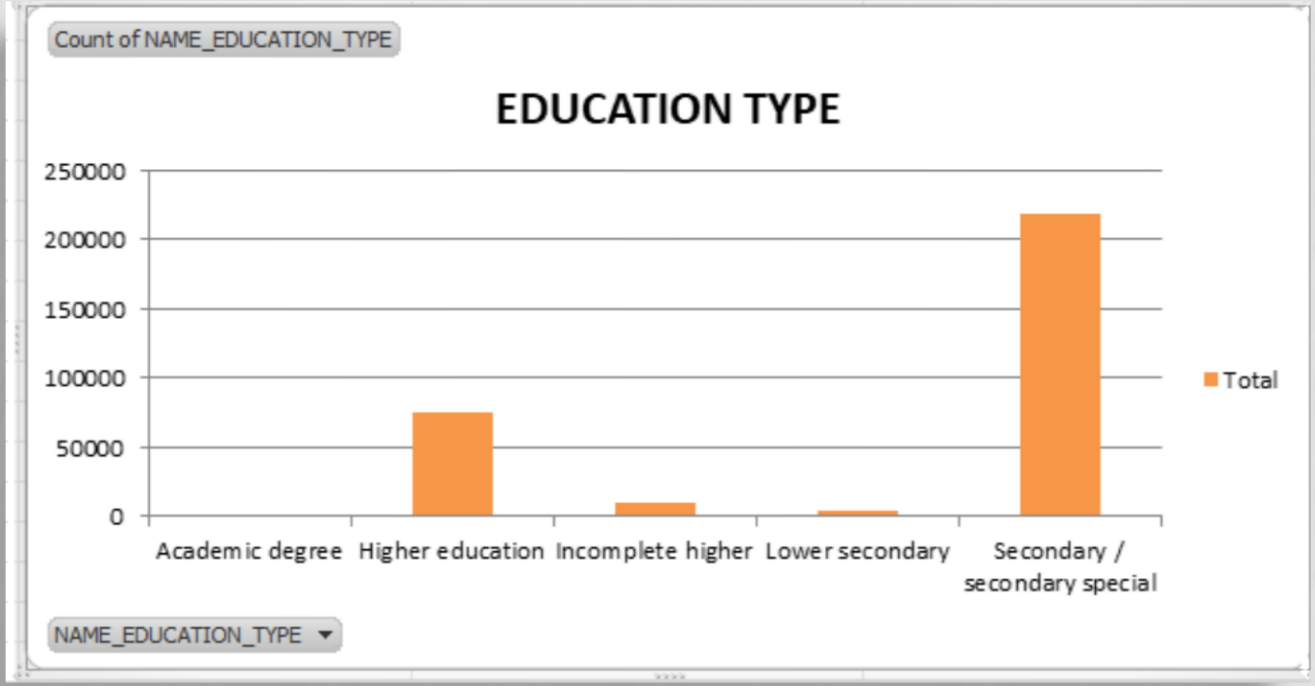
Data Imbalance occurs when data is disseminated is an unequal manner. I plotted data imbalance using pivot table.

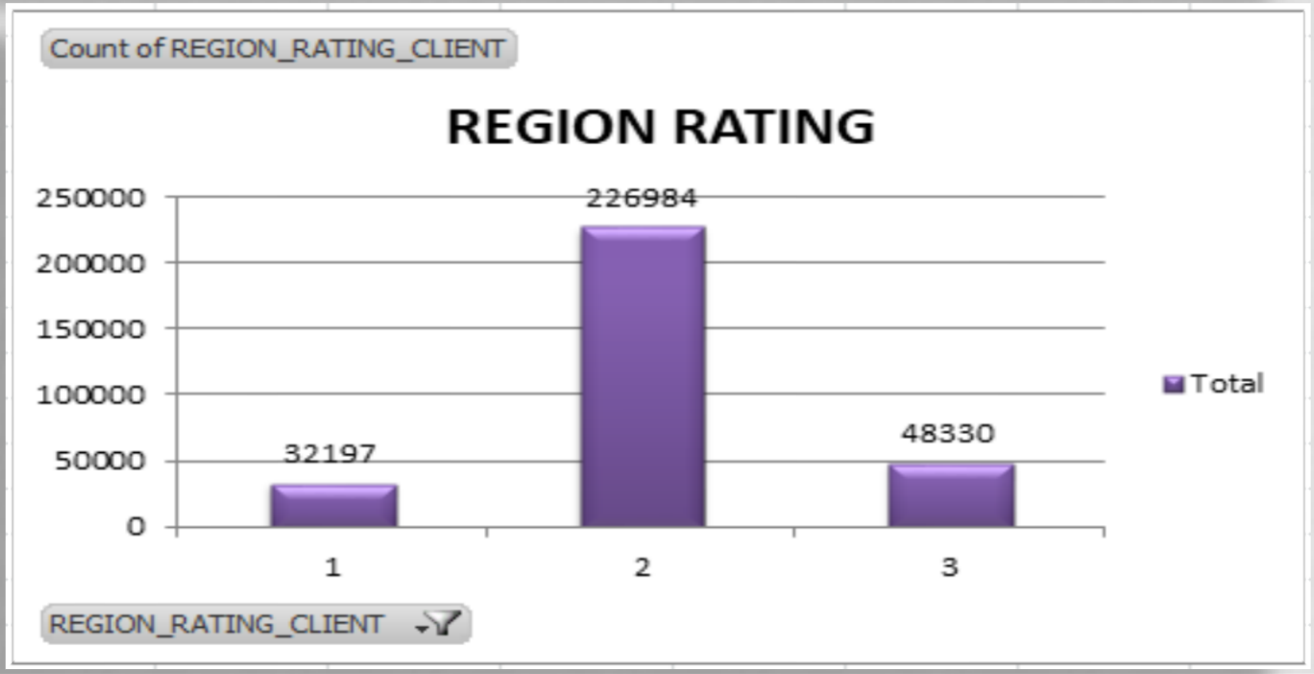
Count of cash loans	Count of Revolving loans	Ratio
278232	29279	9.502783565

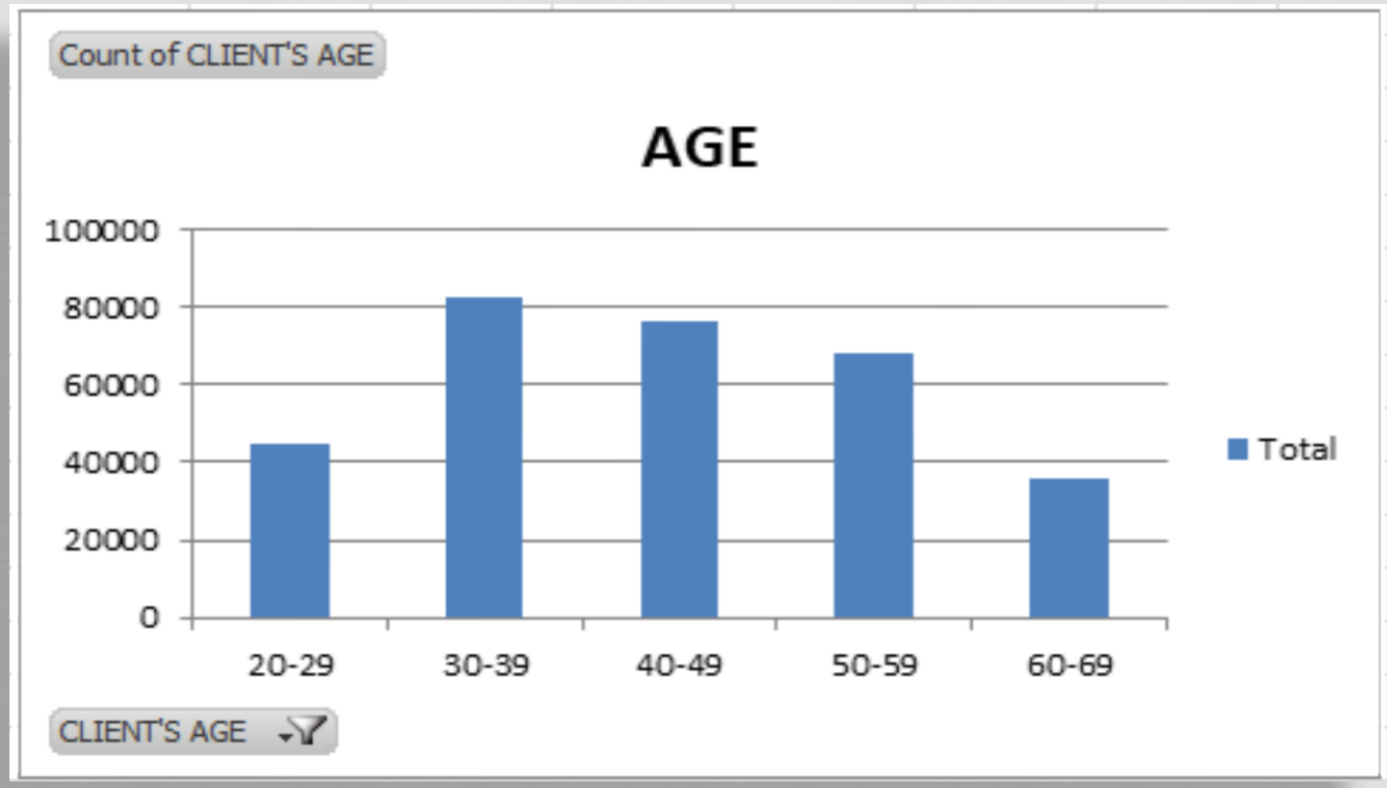


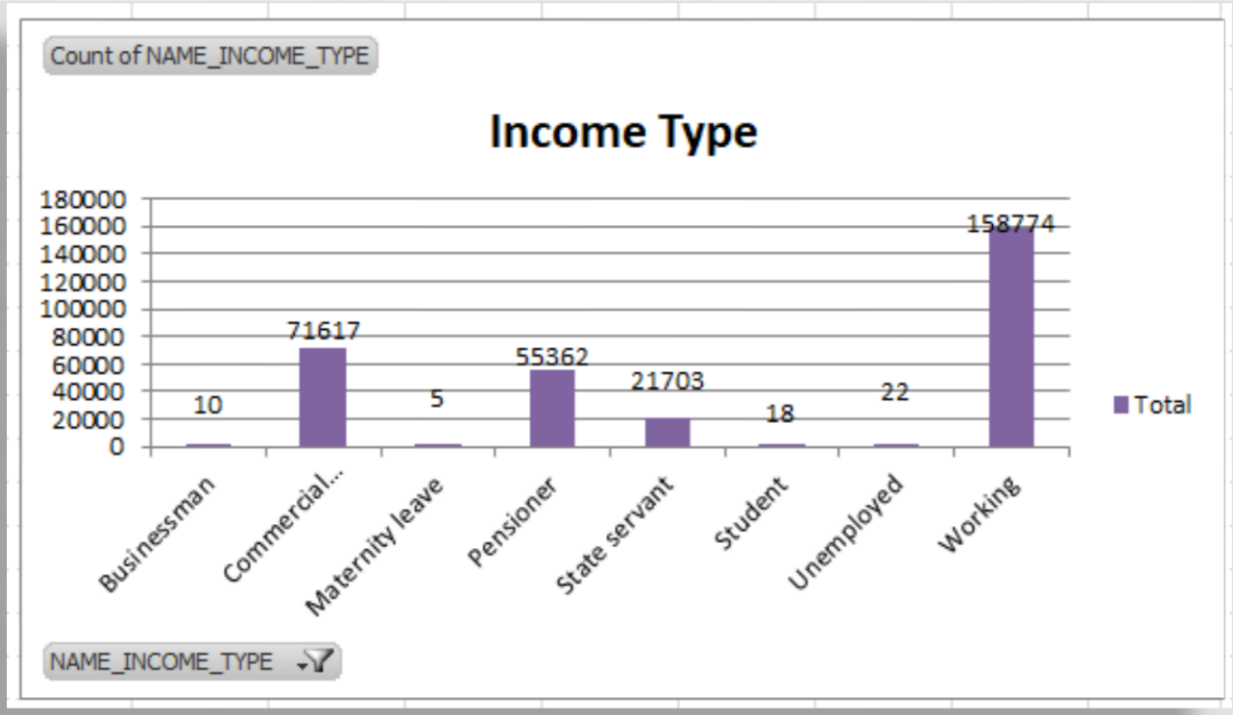
Count of Males	Count Of Females	Ratio
105059	202448	0.518943

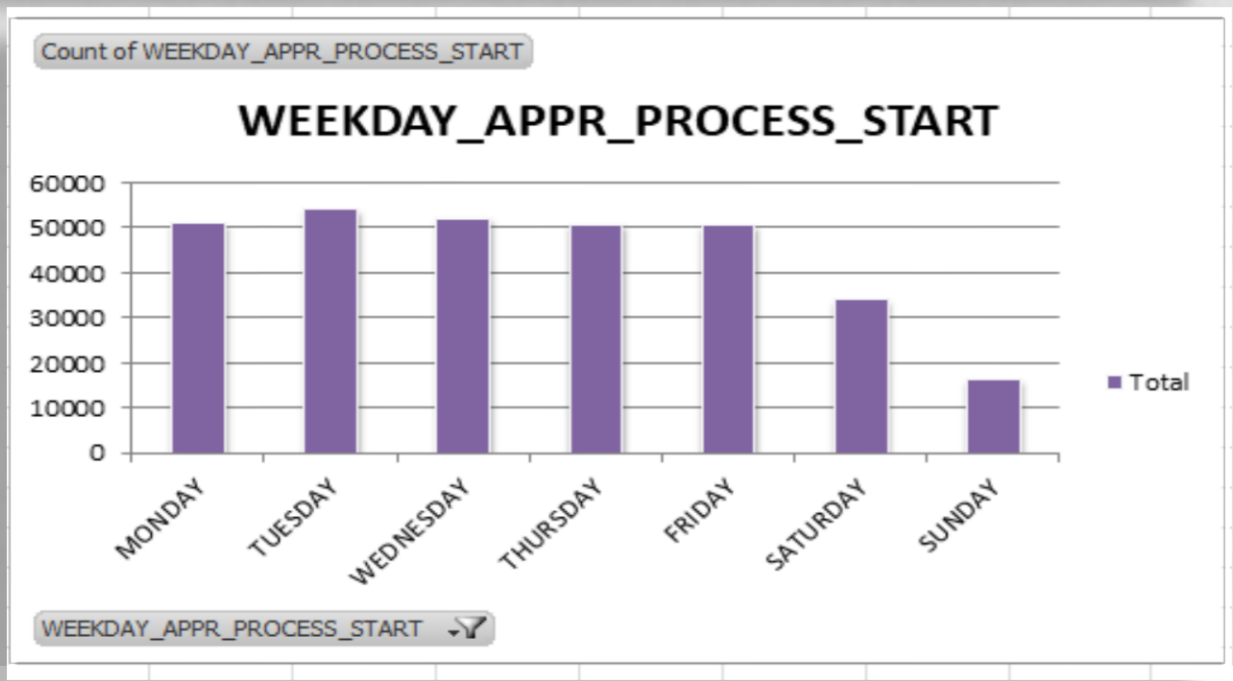


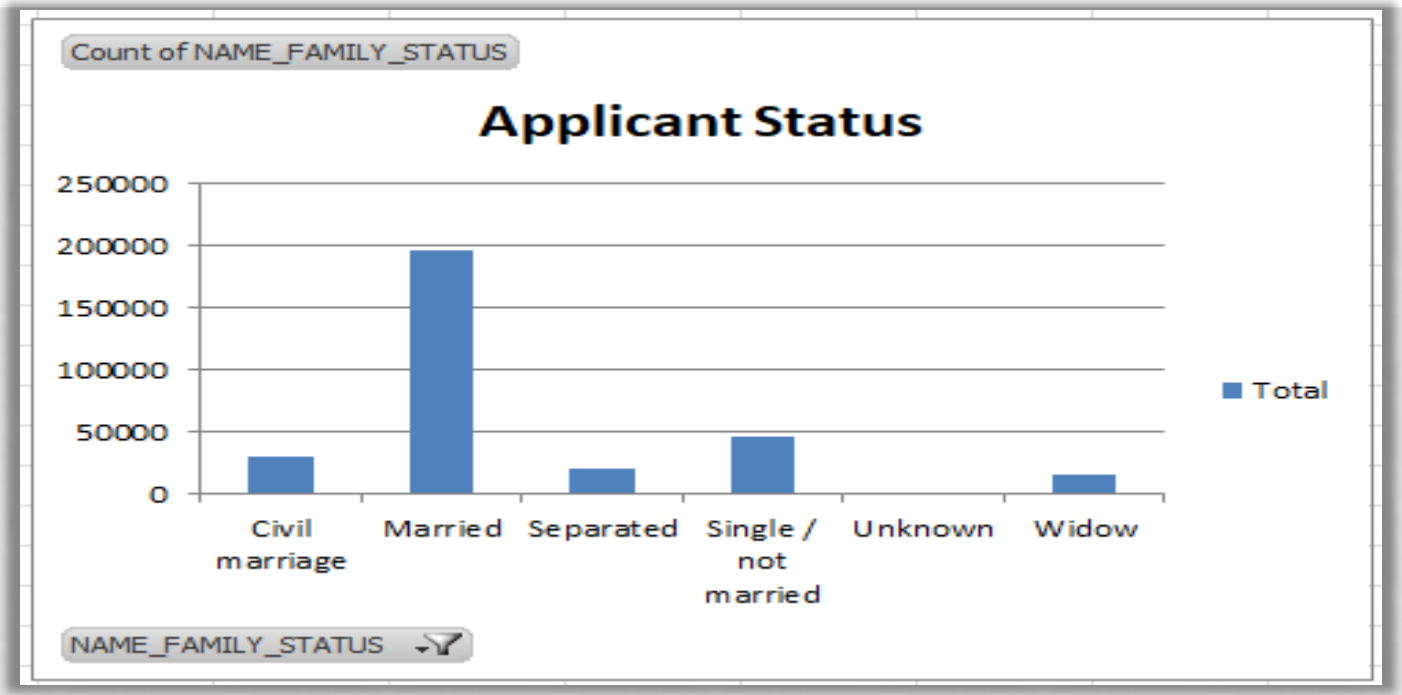

Count of NAME_EDUCATION_TYPE
EDUCATION TYPE
250000
200000
150000
100000
50000
0
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
Total
NAME_EDUCATION_TYPE

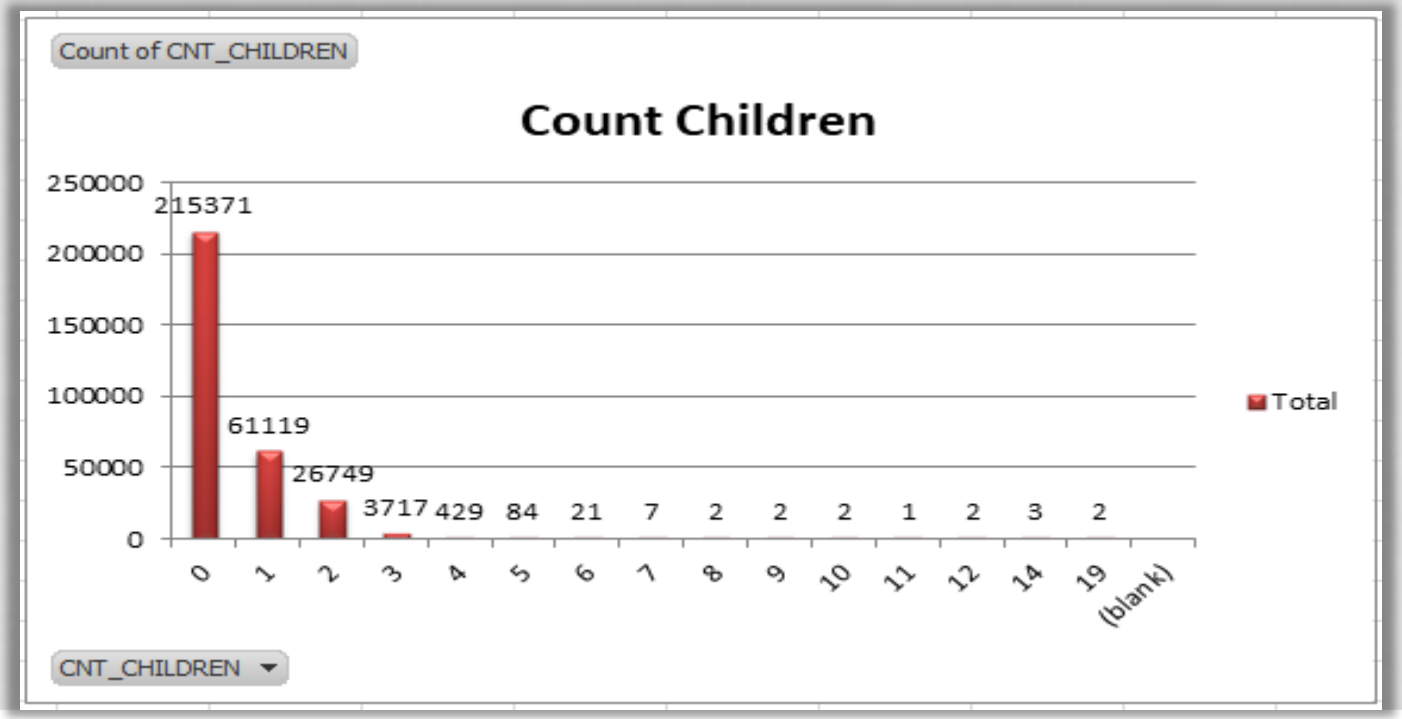

Count of REGION_RATING_CLIENT
REGION RATING
250000
200000
150000
100000
50000
0
226984
32197
48330
1
2
3
Total
REGION_RATING_CLIENT


Count of CLIENT'S AGE
AGE
100000
80000
60000
40000
20000
0
20-29
30-39
40-49
50-59
60-69
Total
CLIENT'S AGE


Count of NAME_INCOME_TYPE
Income Type
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
10
71617
5
55362
21703
18
22
158774
Businessman
Commercial...
Maternity leave
Pensioner
State servant
Student
Unemployed
Working
Total
NAME_INCOME_TYPE

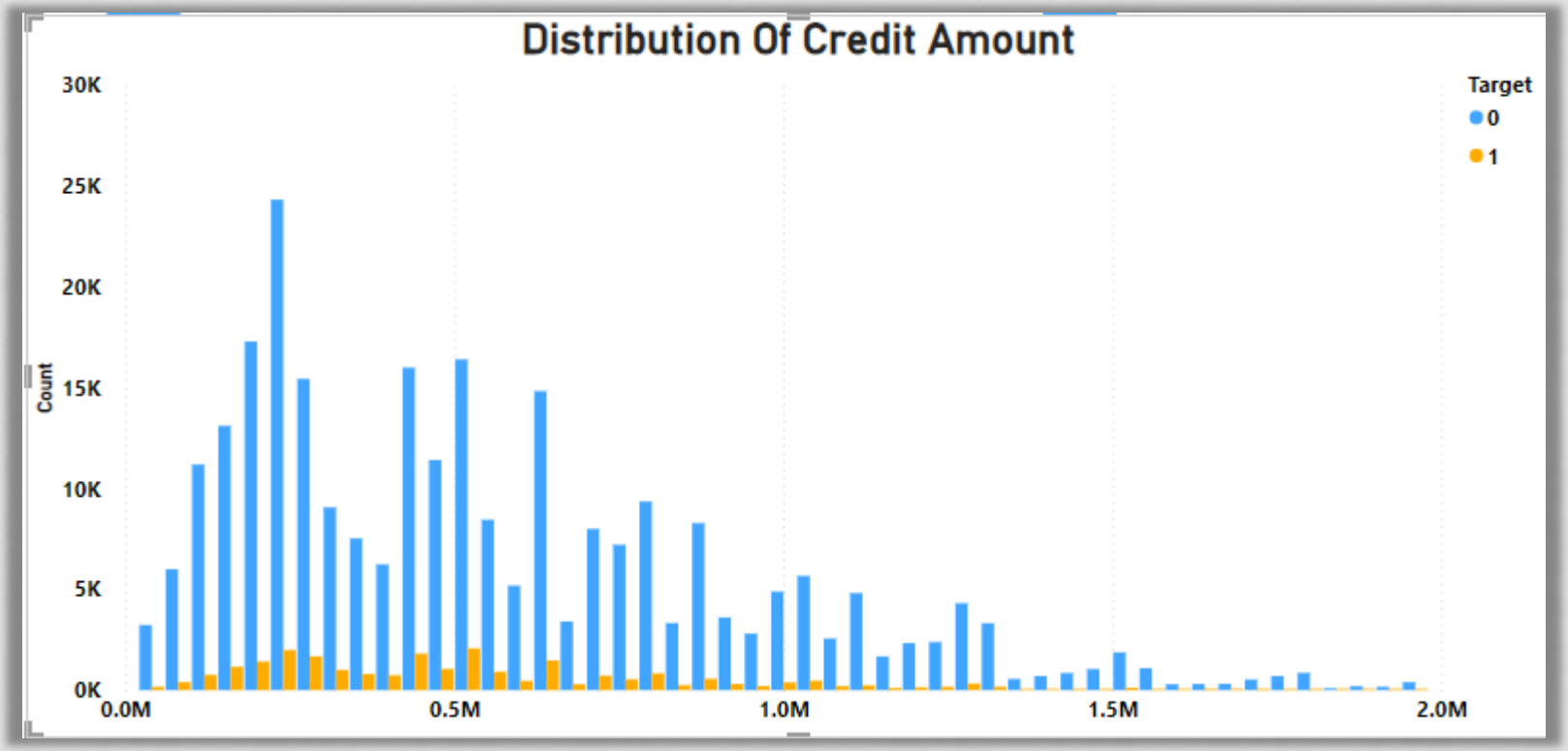

Count of WEEKDAY_APPR_PROCESS_START
WEEKDAY_APPR_PROCESS_START
60000
50000
40000
30000
20000
10000
0
MONDAY
TUESDAY
WEDNESDAY
THURSDAY
FRIDAY
SATURDAY
SUNDAY
Total
WEEKDAY_APPR_PROCESS_START


Count of NAME_FAMILY_STATUS
Applicant Status
250000
200000
150000
100000
50000
0
Civil marriage
Married
Separated
Single / not married
Unknown
Widow
Total
NAME_FAMILY_STATUS


Count of CNT_CHILDREN
Count Children
250000
200000
150000
100000
50000
0
215371
61119
26749
3717
429
84
21
7
2
2
2
1
2
3
2
0
1
2
3
4
5
6
7
8
9
10
11
12
14
19
(blank)
Total
CNT_CHILDREN

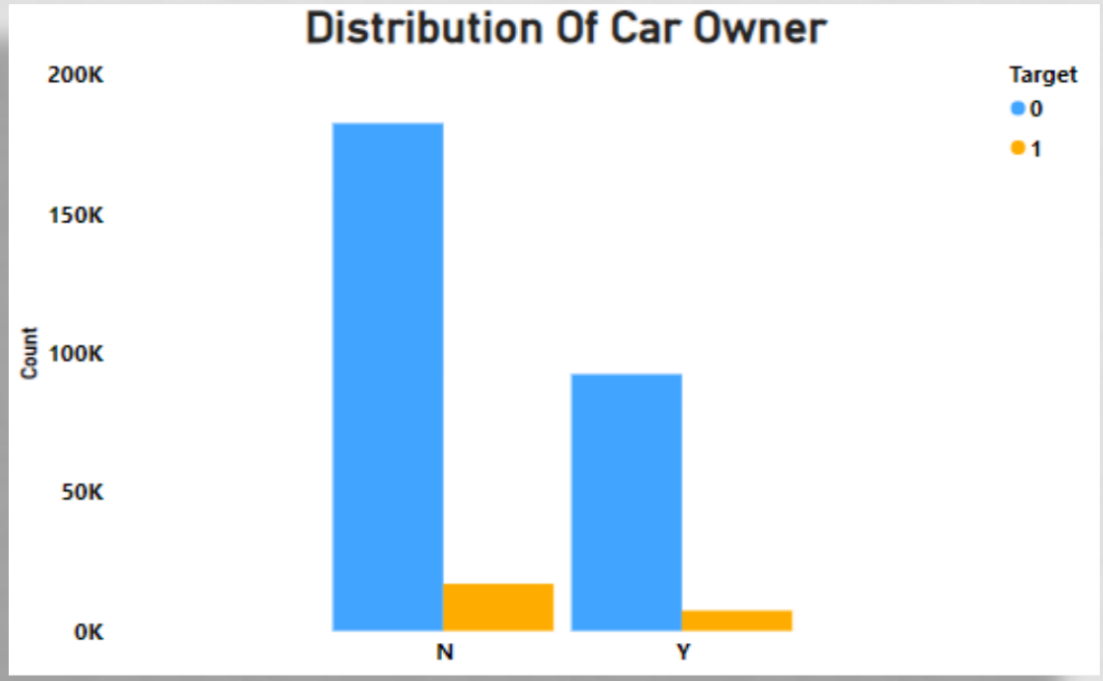
Univariate Analysis:

So for univariate analysis I used power BI tool I drag the credit amount into x-axis and drag target variable into y-axis as well as into legend.



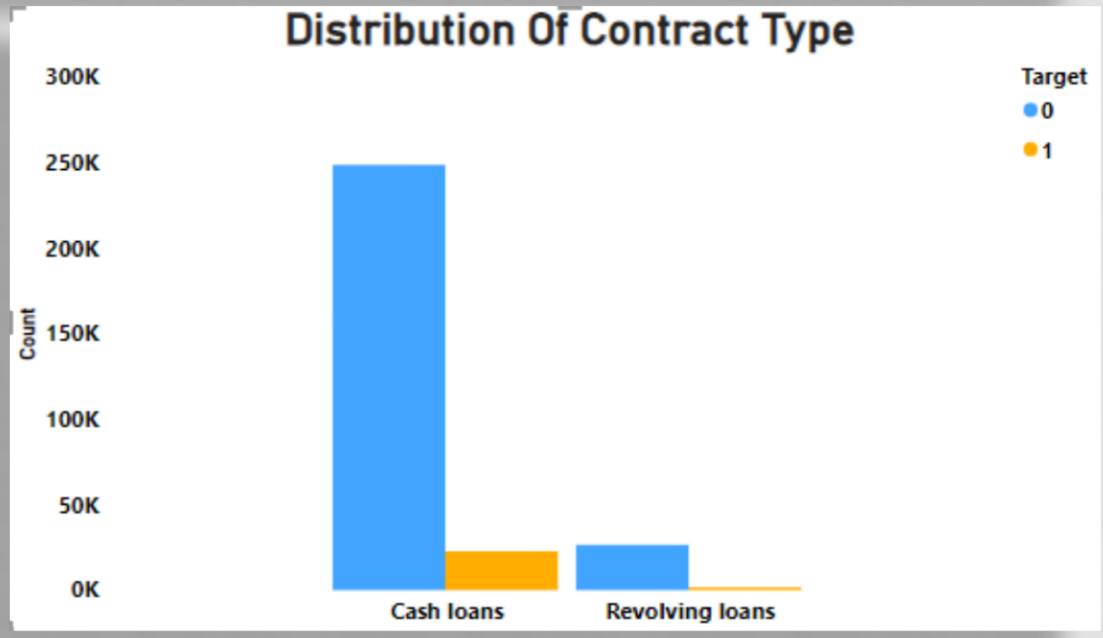
Clients with credit range lying between 0.2M-0.6M are capable of paying loan bank. Least number of clients lying in income range 1.3M-above are not capable of paying.

1. Its shows that clients without having car ownership have no payment difficulties than clients having car.

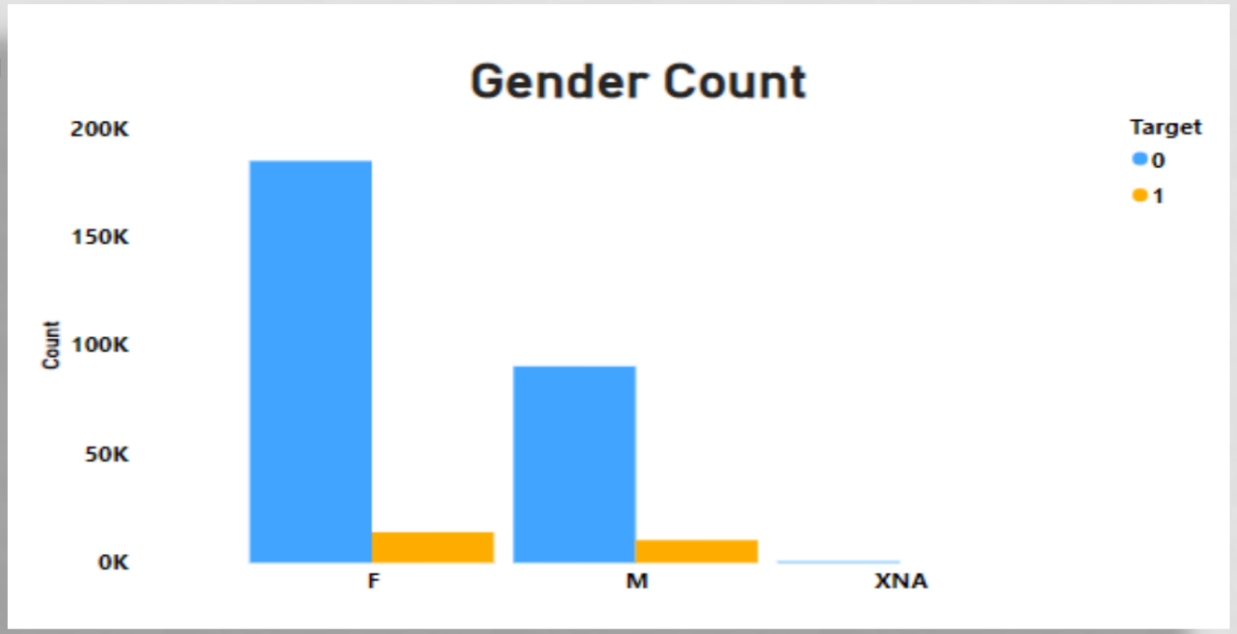


1.Most cash loans applicants have payment difficulties.

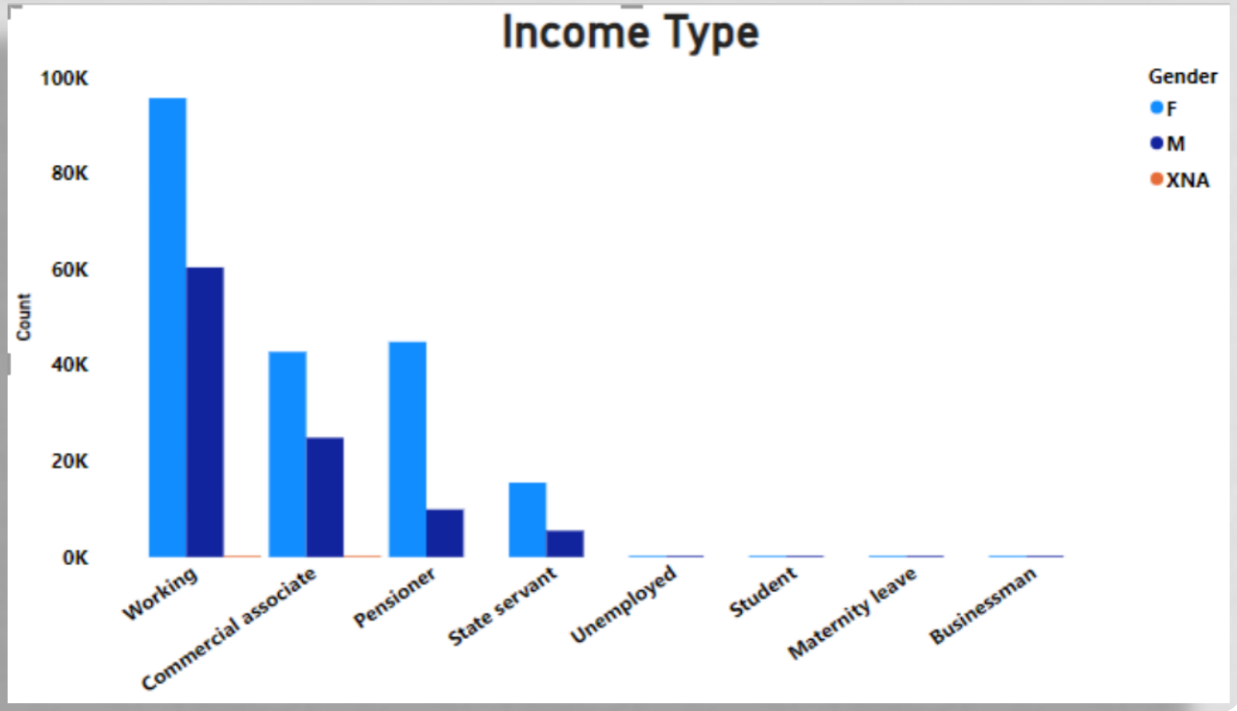
2.The same type of loans also have the most applicants with payment difficulties.

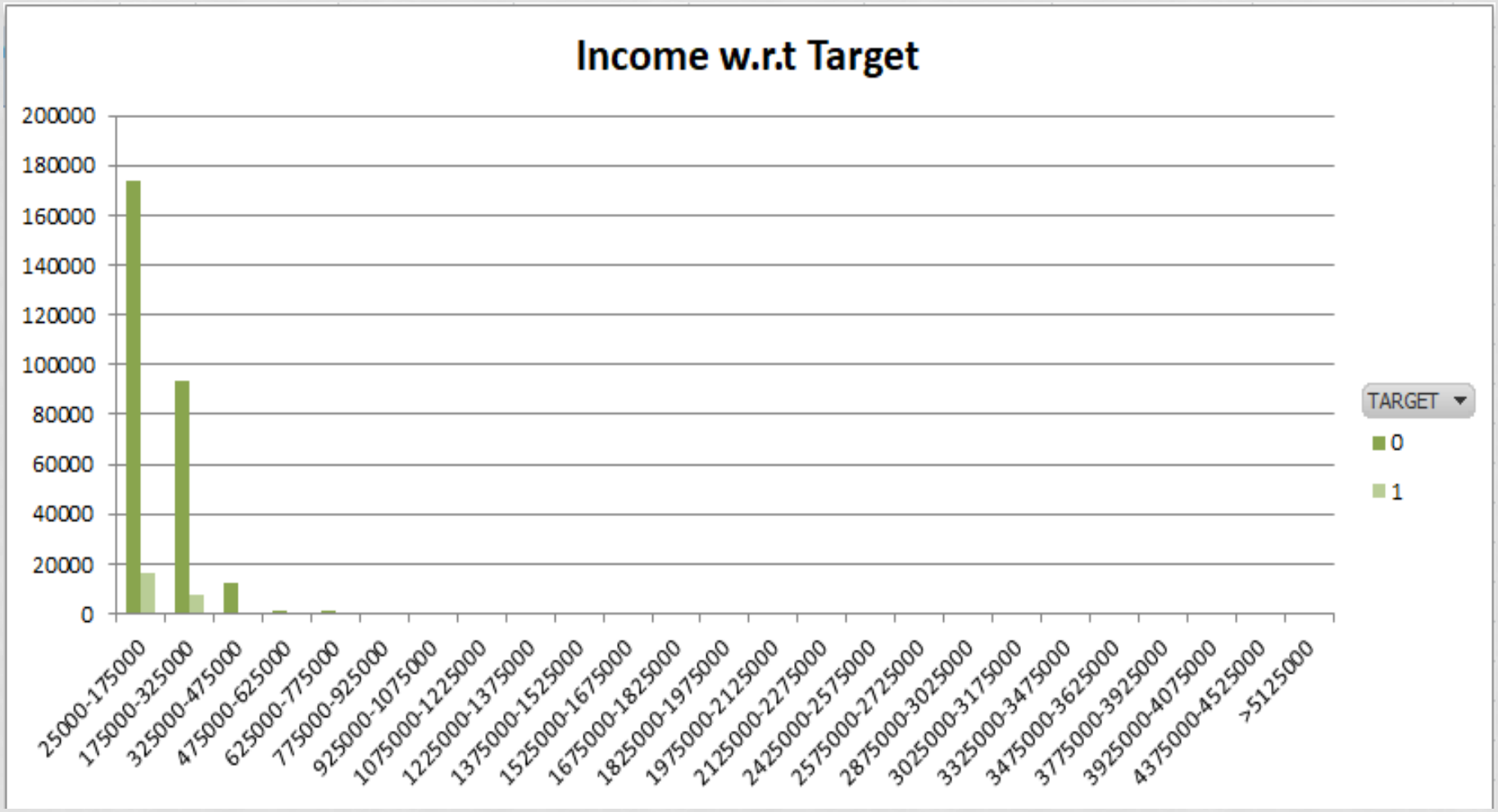


1. It can be seen that both number of males and females are same for having payment difficulties
2. It can also be seen that number of females is more than males for not having payment difficulties.

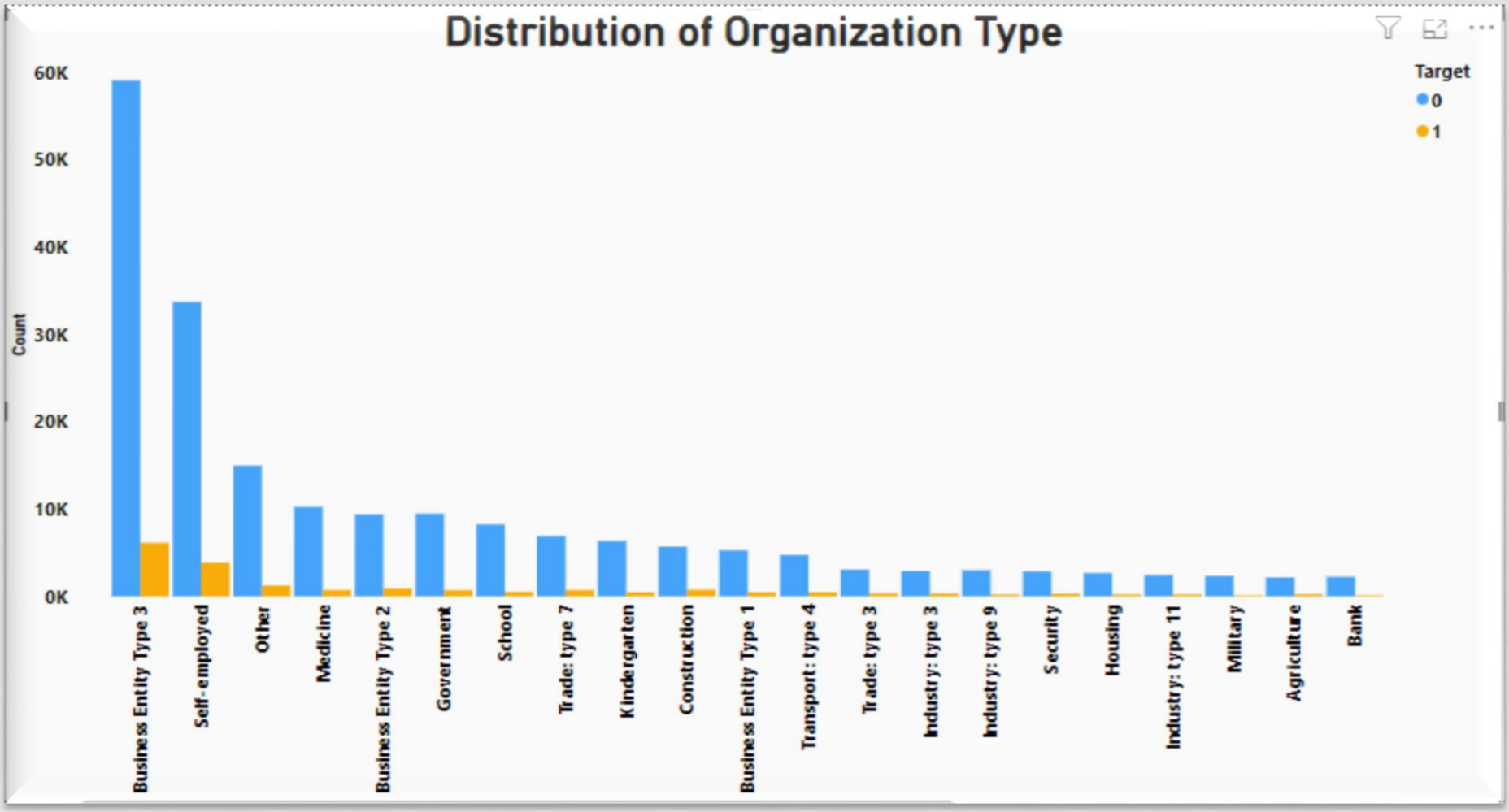


1. We can conclude that most clients who fall in working type income category are applying for loans.





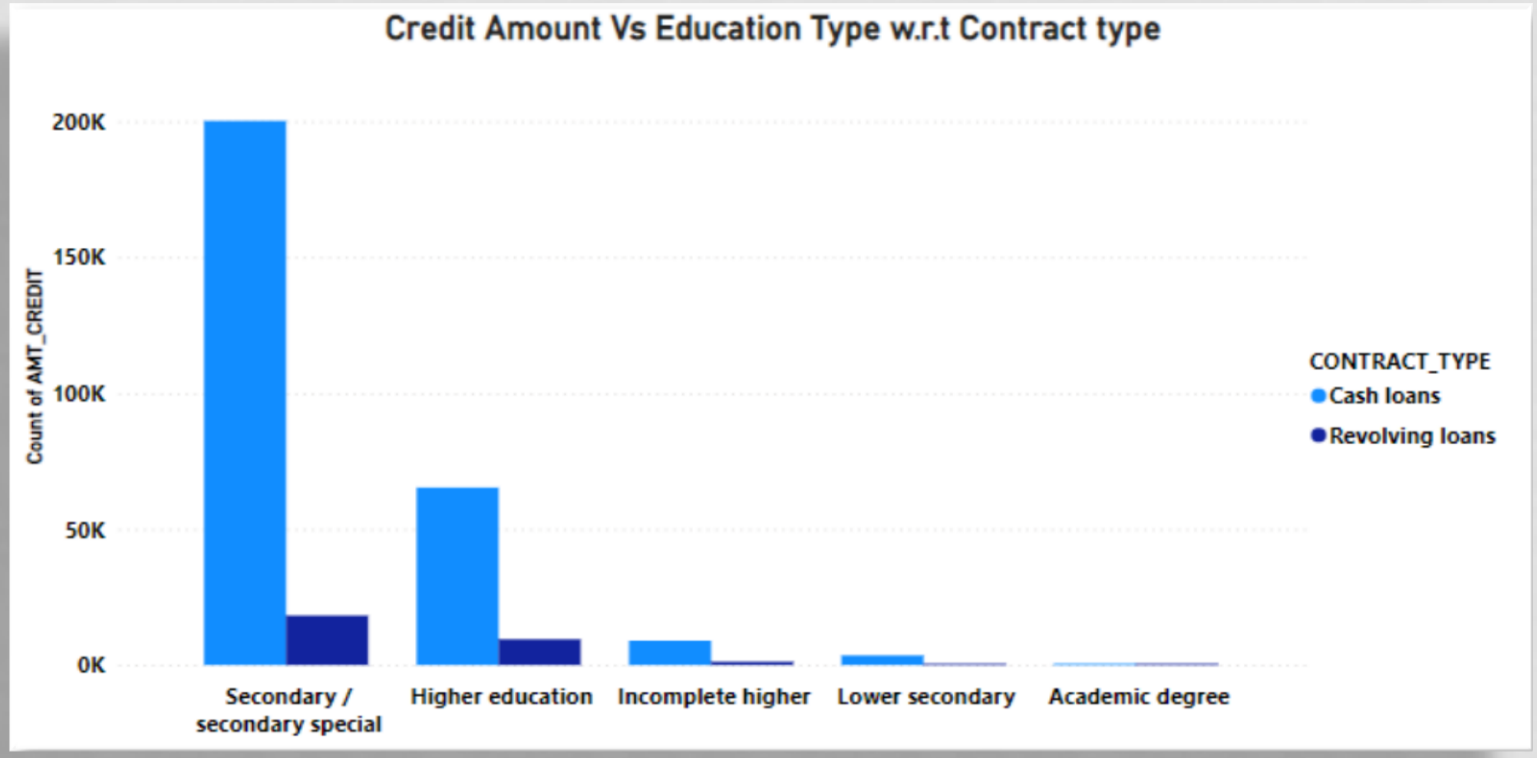
From the above plot we can infer that the maximum number of clients without payment difficulties lie in the income range 25K-475k.
but maximum number of client with payment difficulties lie in range 25k-175k.



Most clients with no payment difficulties lie in organization type named Business Entity Type 3.
And as you can see that after Medkine organization type all are having least number of payment difficulties clients than no payment difficulties .

Bivariate Analysis

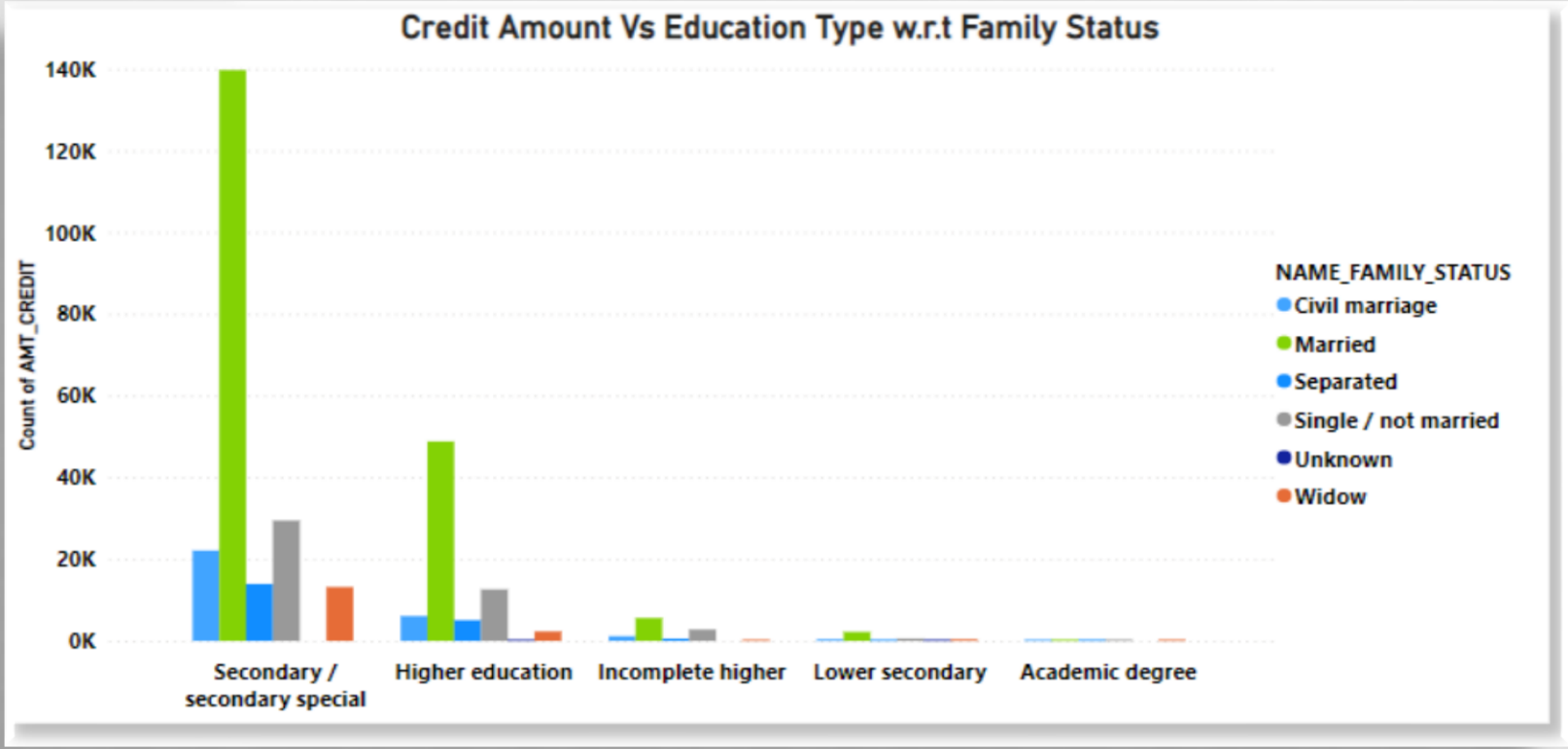
So to bivariate analysis I drag the Credit amount into value, contract type into legend and education type into x-axis.



1.Education status of higher education and secondary/secondary special have most clients for contract type cash loans

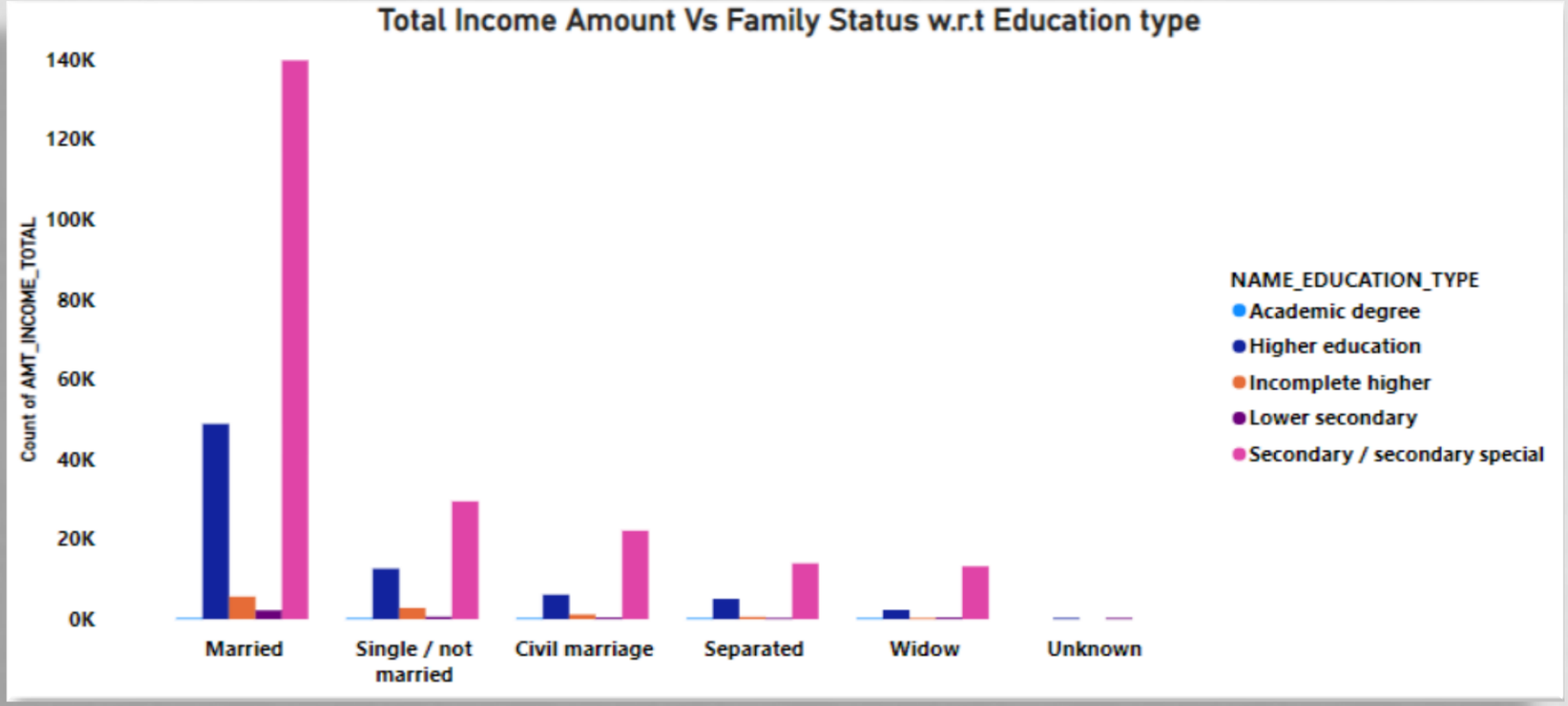
2.Most number of clients applying for revolving loans are in education status also are secondary/secondary special and higher education than other education type.

I drag the Credit amount into value, Education type into legend and family status into x-axis.



1. Family status of married of Secondary/secondary special education are having higher number of credits than others, then family status of civil marriage and separated of secondary/ secondary special education type having higher number of credits than others.

I drag the Credit amount into value, family status into legend and education type into x-axis.



1.Family status of 'civil marriage', 'marriage' and ‘Single/not married' of secondary/ secondary special are having higher number of income than others.

2.All family status of academic degree education type having minority of income rate. It means bank having less loan customers in education type academic degree. As we see earlier in contract type.

Correlations (target 0)

The below chart provides correlation of different variables for clients with no payment difficulties.
For Correlation I filter the target variable to 0 and selected the all variables then clicked the data bar then click on Data analysis.

As we can see top 10 correlations.

1. Family member & children count
2. Goods price& credit
3. Annuity & credit
4. Goods price& annuity
5. Annuity &total income
6. Credit &total income
7. Goods price& total income
8. Day birth & children count
9. Day registration & days birth
10. Days publish & days birth

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS
CNT_CHILDREN	1									
AMT_INCOME_TOTAL	0.0274	1								
AMT_CREDIT	0.00308	0.3428	1							
AMT_ANNUITY	0.02091	0.41895	0.77131	1						
AMT_GOODS_PRICE	-0.00052	0.34946	0.98725	0.77669	1					
DAYS_BIRTH	0.33697	0.06261	-0.04738	0.01226	-0.04456	1				
DAYS_EMPLOYED	-0.24336	-0.14125	-0.07251	-0.10642	-0.07105	-0.61805	1			
DAYS_REGISTRATION	0.18579	0.06494	0.01348	0.03944	0.01592	0.33315	-0.21019	1		
DAYS_ID_PUBLISH	-0.02875	0.0229	-0.00146	0.01411	-0.00365	0.27131	-0.27429	0.10024	1	
CNT_FAM_MEMBERS	0.87857	0.03426	0.06454	0.07579	0.06281	0.28582	-0.23741	0.17563	-0.02046	1

Correlation (target 1)

Same as before I filter the target variable to 1 and selected the variables then clicked the data bar then click on Data analysis.

So these are the top 10 correlations

1. Family mem &children count
2. Goods price & credit
3. Annuity & credit
4. Goods price & annuity
5. Days birth & family member
6. Days registration & days birth
7. Days publish & days birth
8. Days registration & family member
9. Days registration & children count
10. Days id publish & days registration

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	CNT_FAM_MEMBERS	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH
CNT_CHILDREN	1									
AMT_INCOME_TOTAL	0.004796	1								
AMT_CREDIT	-0.00167	0.038131	1							
AMT_ANNUITY	0.031257	0.046421	0.752195	1						
AMT_GOODS_PRICE	-0.00811	0.037583	0.983103	0.752699	1					
CNT_FAM_MEMBERS	0.885484	0.006654	0.051224	0.075711	0.047388	1				
DAYS_BIRTH	0.259109	0.003096	-0.13532	-0.0143	-0.13581	0.203267	1			
DAYS_EMPLOYED	-0.19194	-0.01498	-0.00097	-0.08255	0.003587	-0.18656	-0.5751	1		
DAYS_REGISTRATION	0.149154	0.000158	-0.02585	0.034279	-0.02568	0.145828	0.289114	-0.18893	1	
DAYS_ID_PUBLISH	-0.0323	-0.00421	-0.05233	-0.01677	-0.05609	-0.03178	0.252863	-0.22647	0.096833	1

The above chart provides correlation of different variables for clients with payment difficulties

Insights (previous_application.csv)

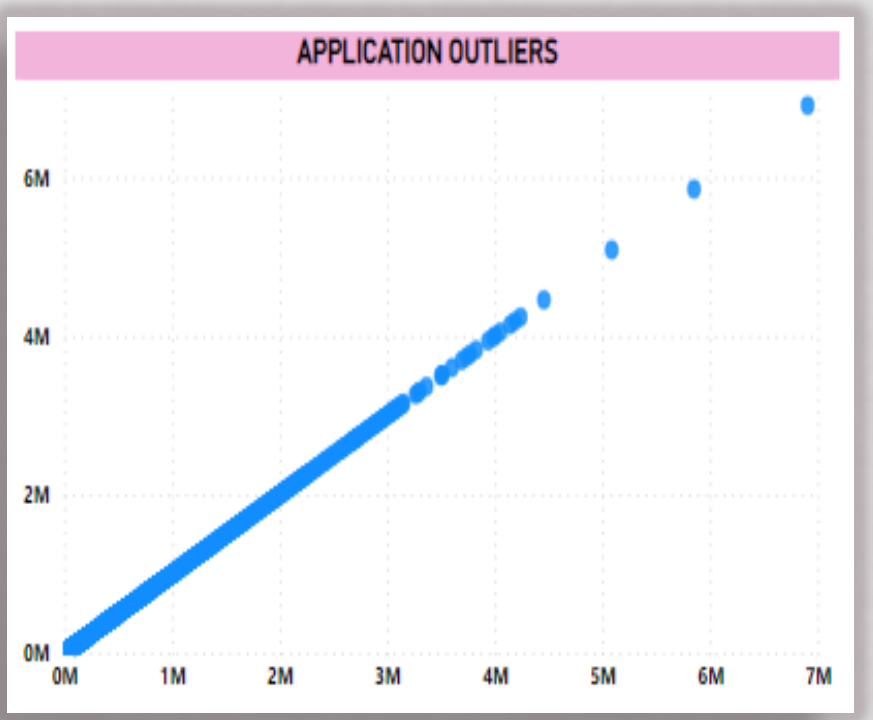
Data Cleaning: The dataset contained more than 1.6 million rows and 37 columns of data.

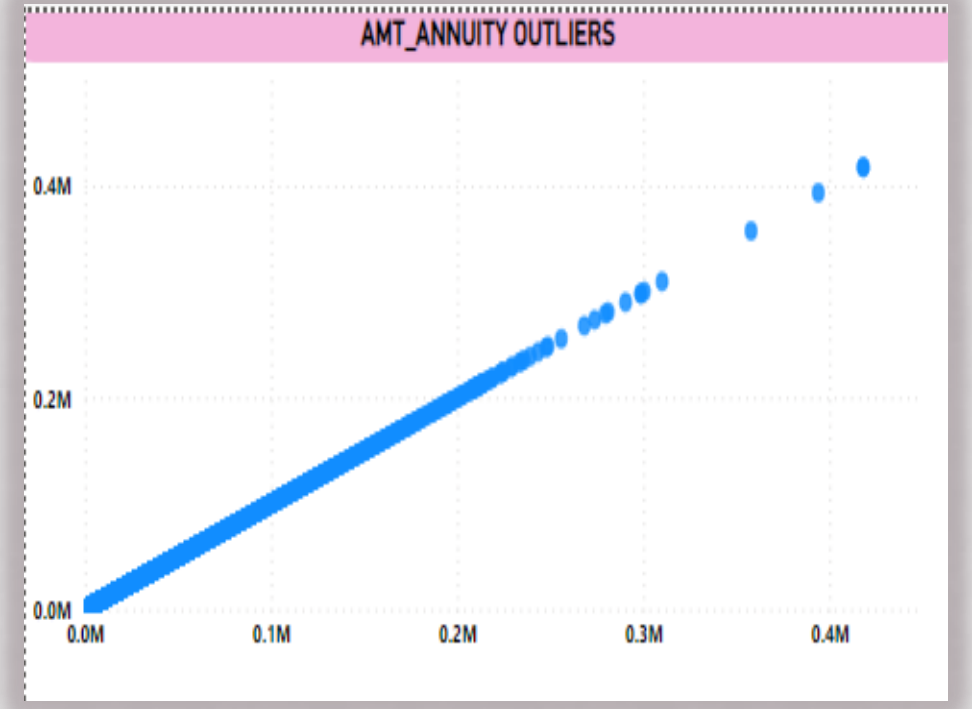
So I used Power BI software because there are more than million rows of data which cannot be import in Excel, after importing the dataset into power BI basically I transform the data by removing unwanted column contained 45% of blank data.

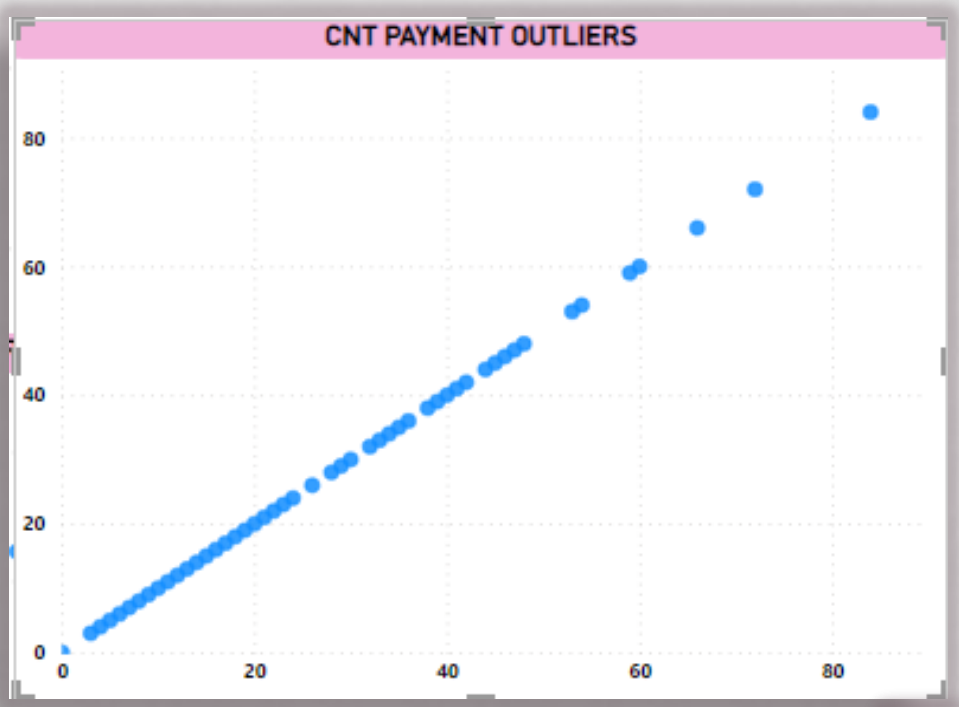
After deleting unwanted columns we got 34 remaining columns.

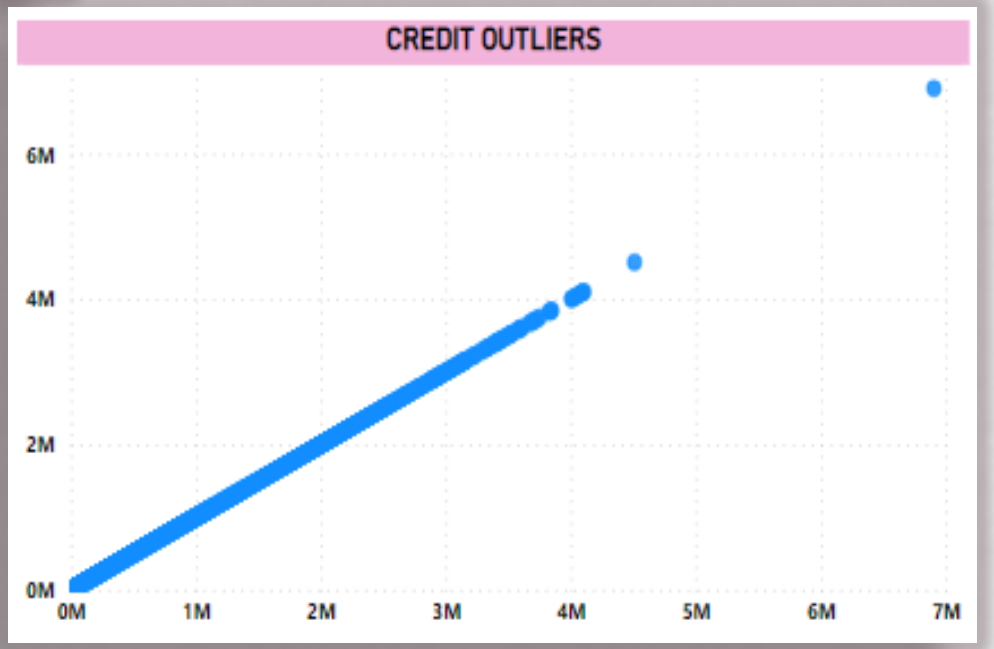
Outliers

Outliers can only be identified on Numerical variable.
To find outlier basically I used Scatter plot chart.





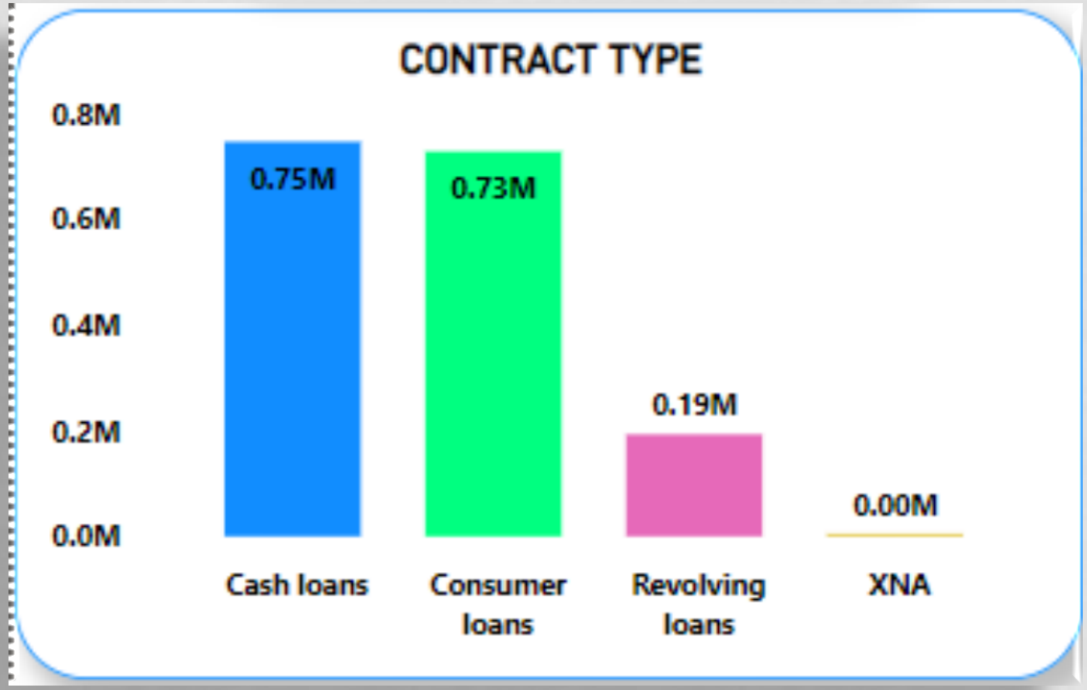

CNT PAYMENT OUTLIERS
80
60
40
20
0
0
20
40
60
80

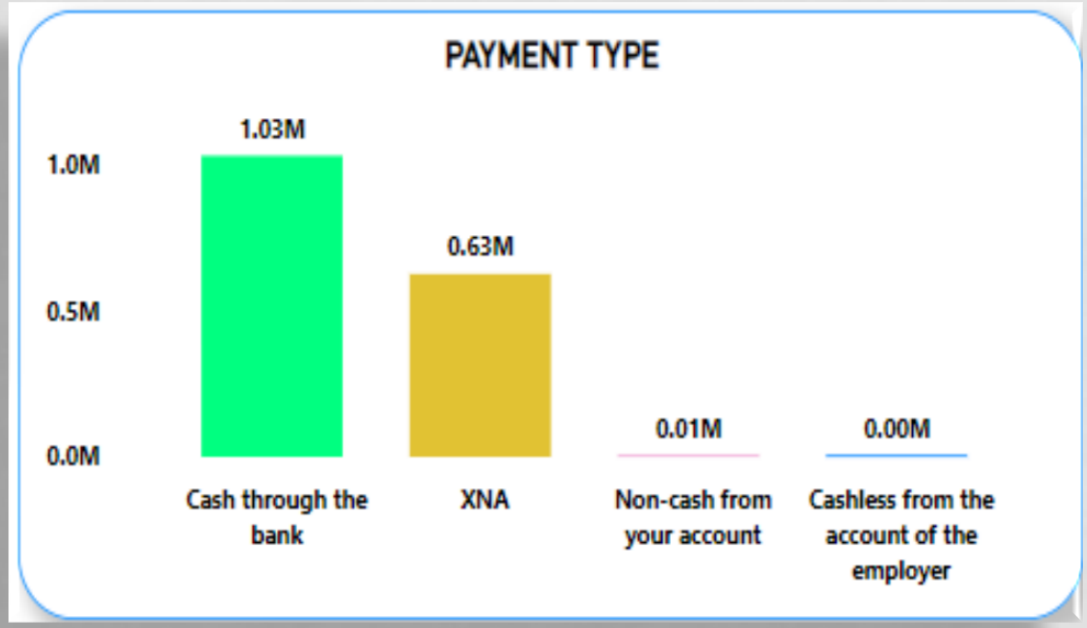

CREDIT OUTLIERS
6M
4M
2M
0M
0M
1M
2M
3M
4M
5M
6M
7M

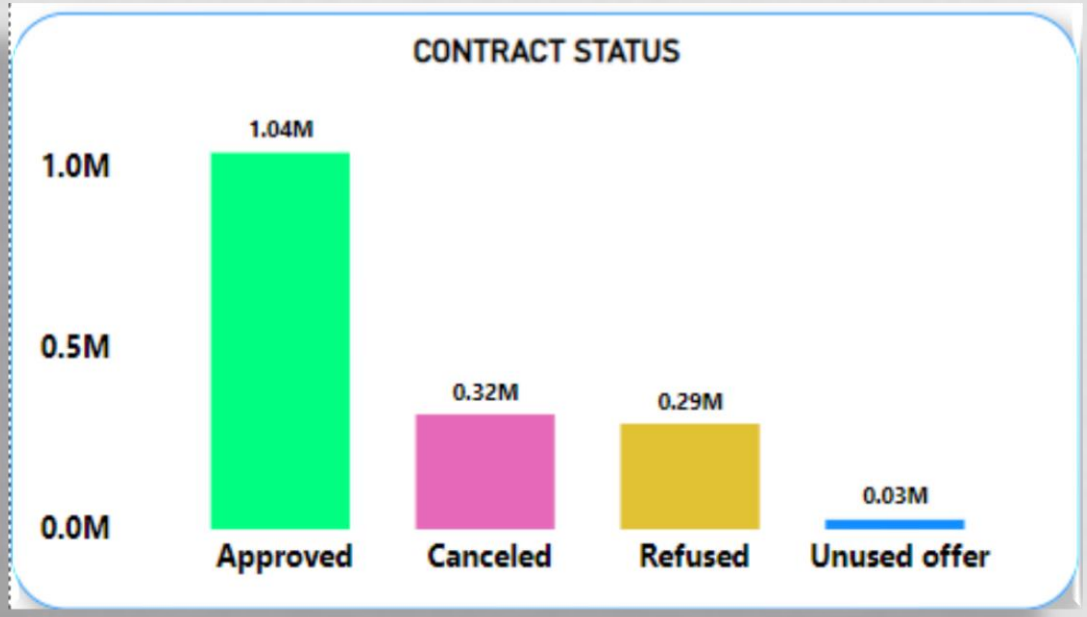

GOODS PRICE OUTLIERS
6M
4M
2M
0M
0M
1M
2M
3M
4M
5M
6M
7M

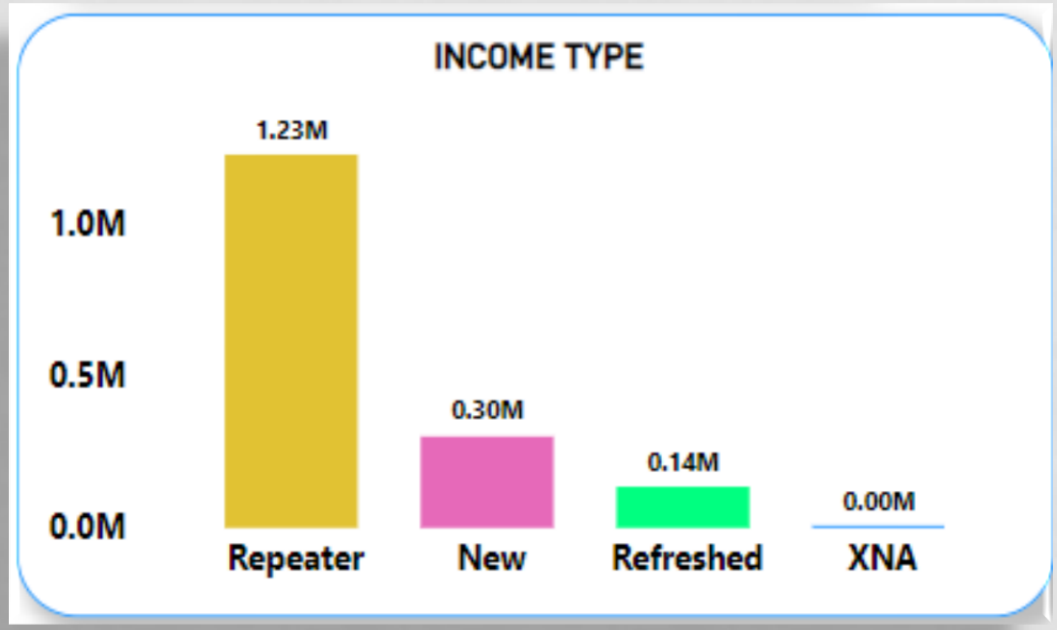
Data Imbalance:

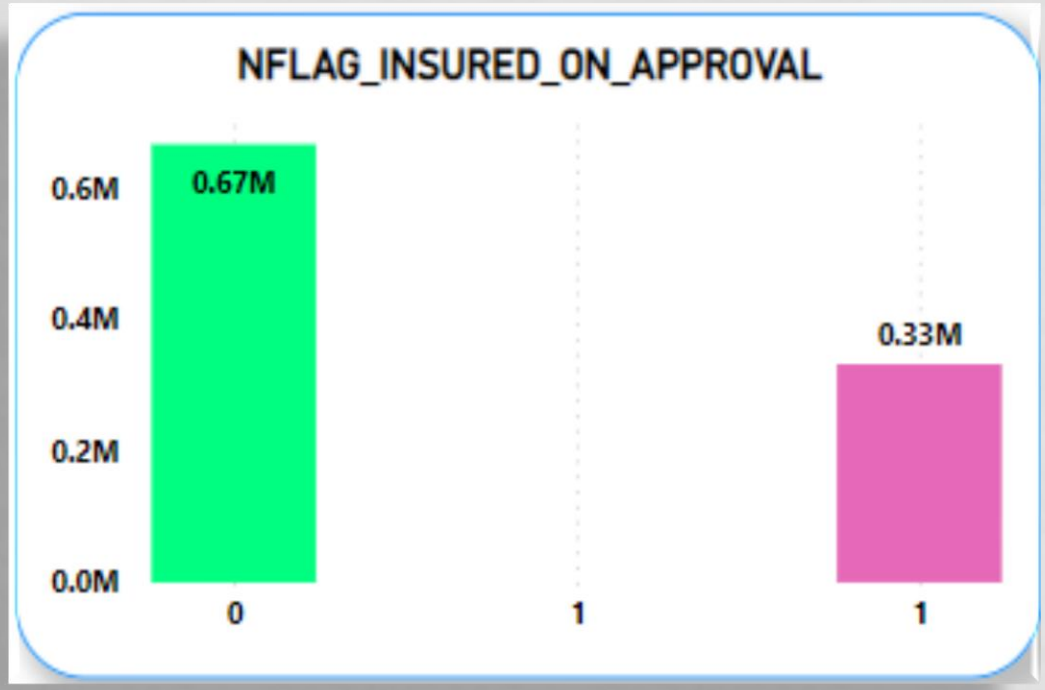
Below are the columns where data is unevenly distributed

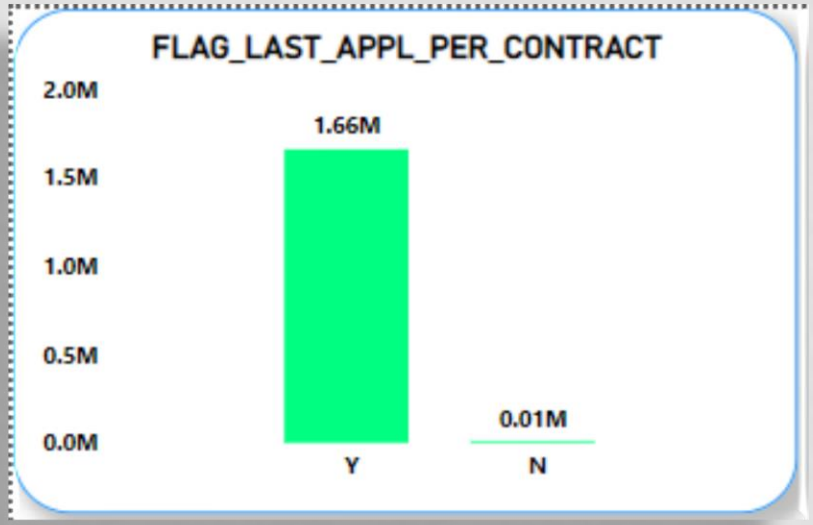


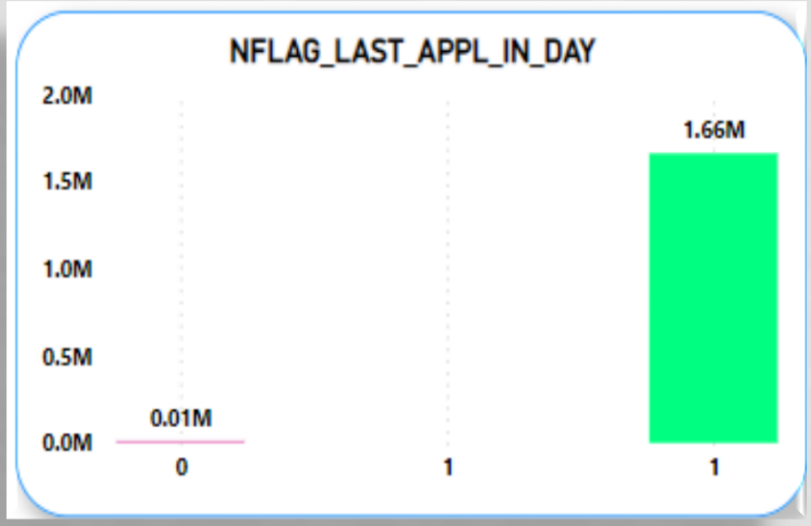

PAYMENT TYPE
1.03M
1.0M
0.5M
0.0M
0.63M
0.01M
0.00M
Cash through the bank
XNA
Non-cash from your account
Cashless from the account of the employer


CONTRACT STATUS
1.04M
1.0M
0.5M
0.0M
0.32M
0.29M
0.03M
Approved
Canceled
Refused
Unused offer


INCOME TYPE
1.23M
1.0M
0.5M
0.30M
0.14M
0.00M
0.0M
Repeater
New
Refreshed
XNA

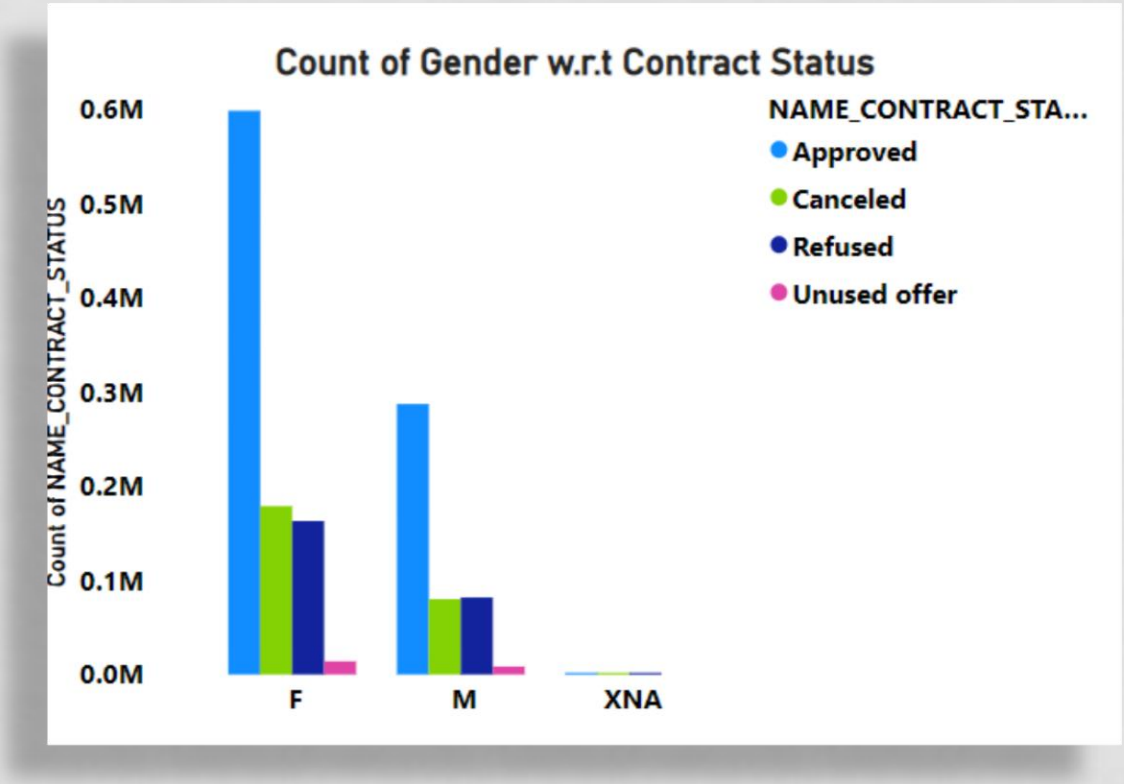

NFLAG_INSURED_ON_APPROVAL
0.6M
0.67M
0.4M
0.33M
0.2M
0.0M
0
1
1


FLAG_LAST_APPL_PER_CONTRACT
2.0M
1.5M
1.0M
0.5M
0.0M
1.66M
0.01M
Y
N

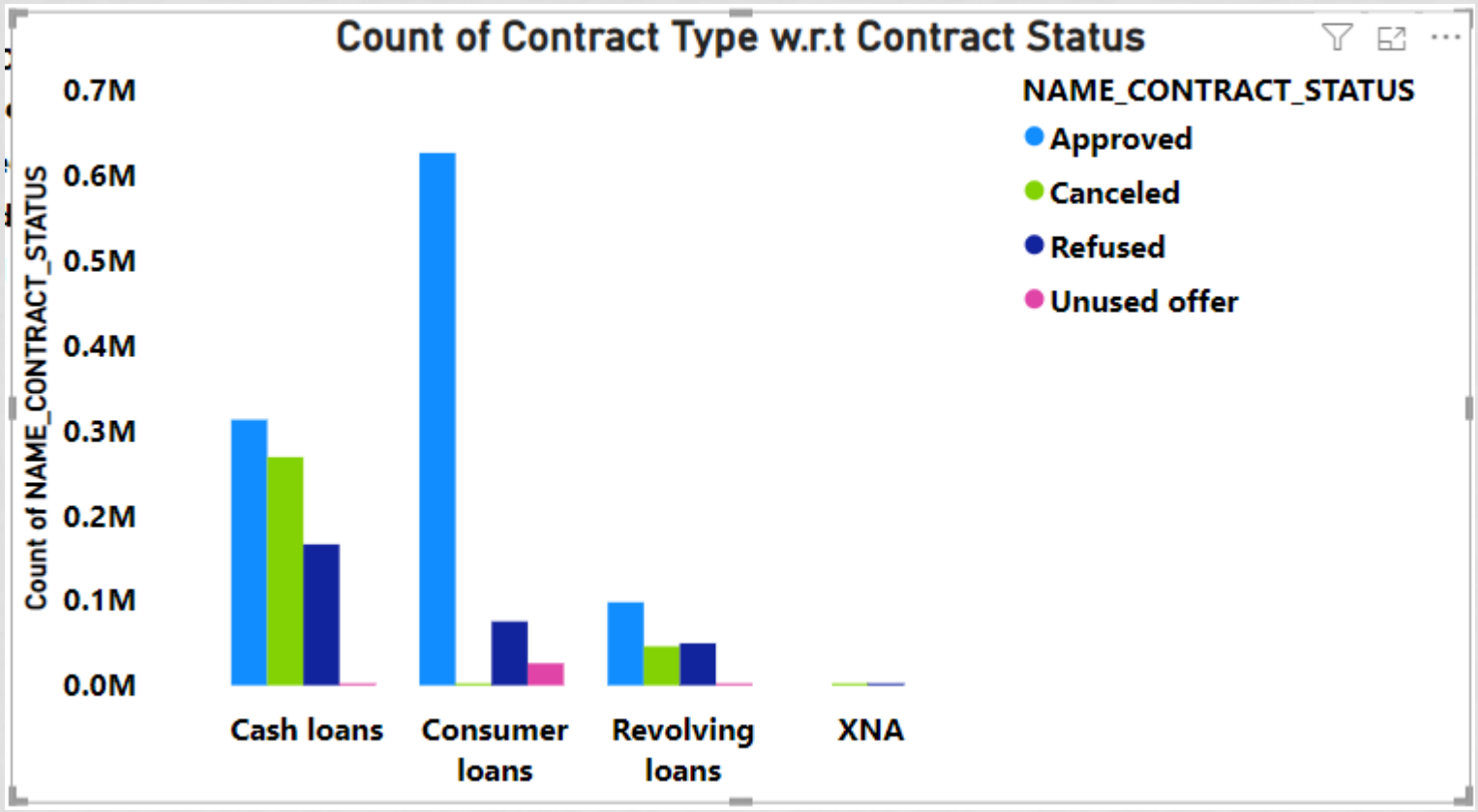

NFLAG_LAST_APPL_IN_DAY
2.0M
1.5M
1.0M
0.5M
0.0M
0.01M
1.66M
0
1
1

Univariate Analysis

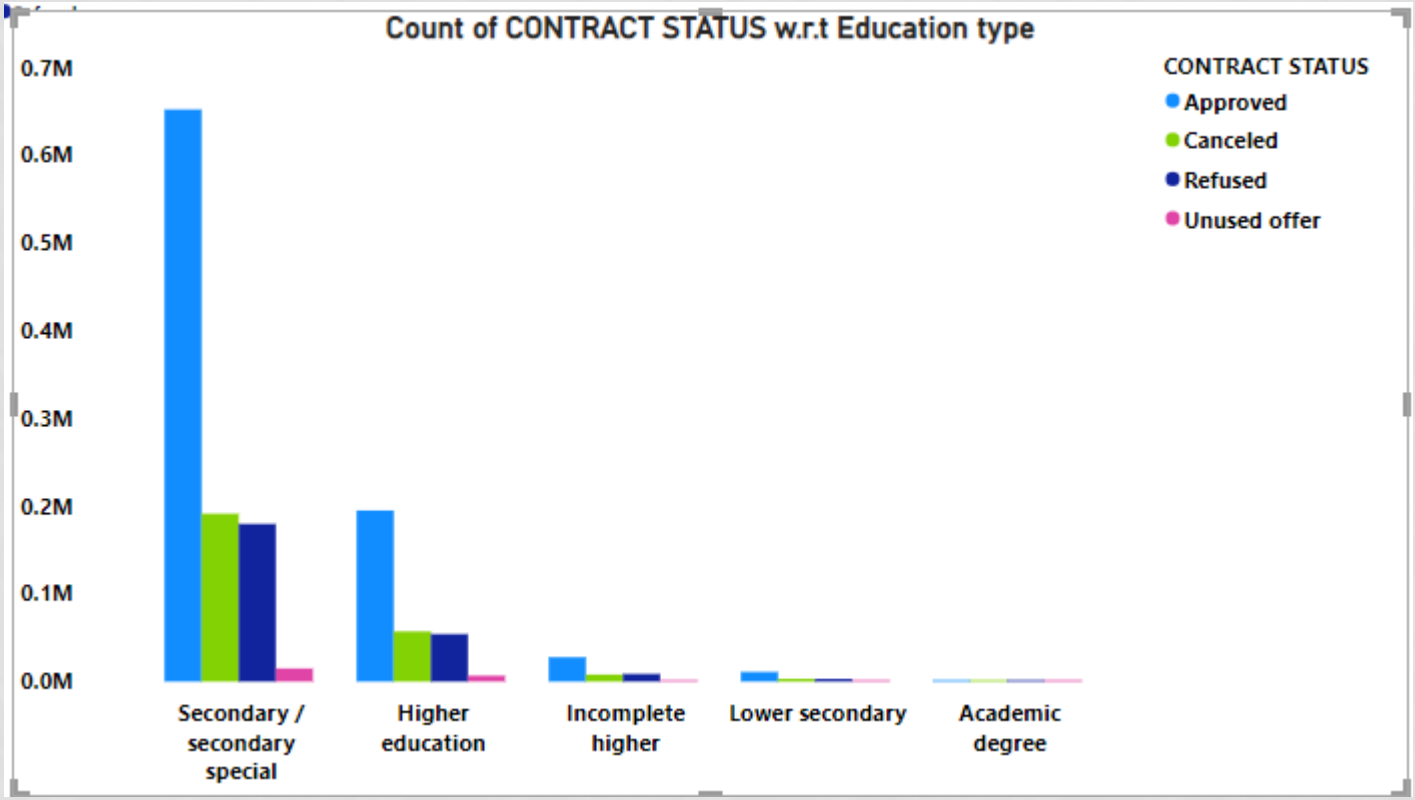
So for univariate analysis I did same as application_data.csv dataset.
And after this I create relationship between both dataset table.



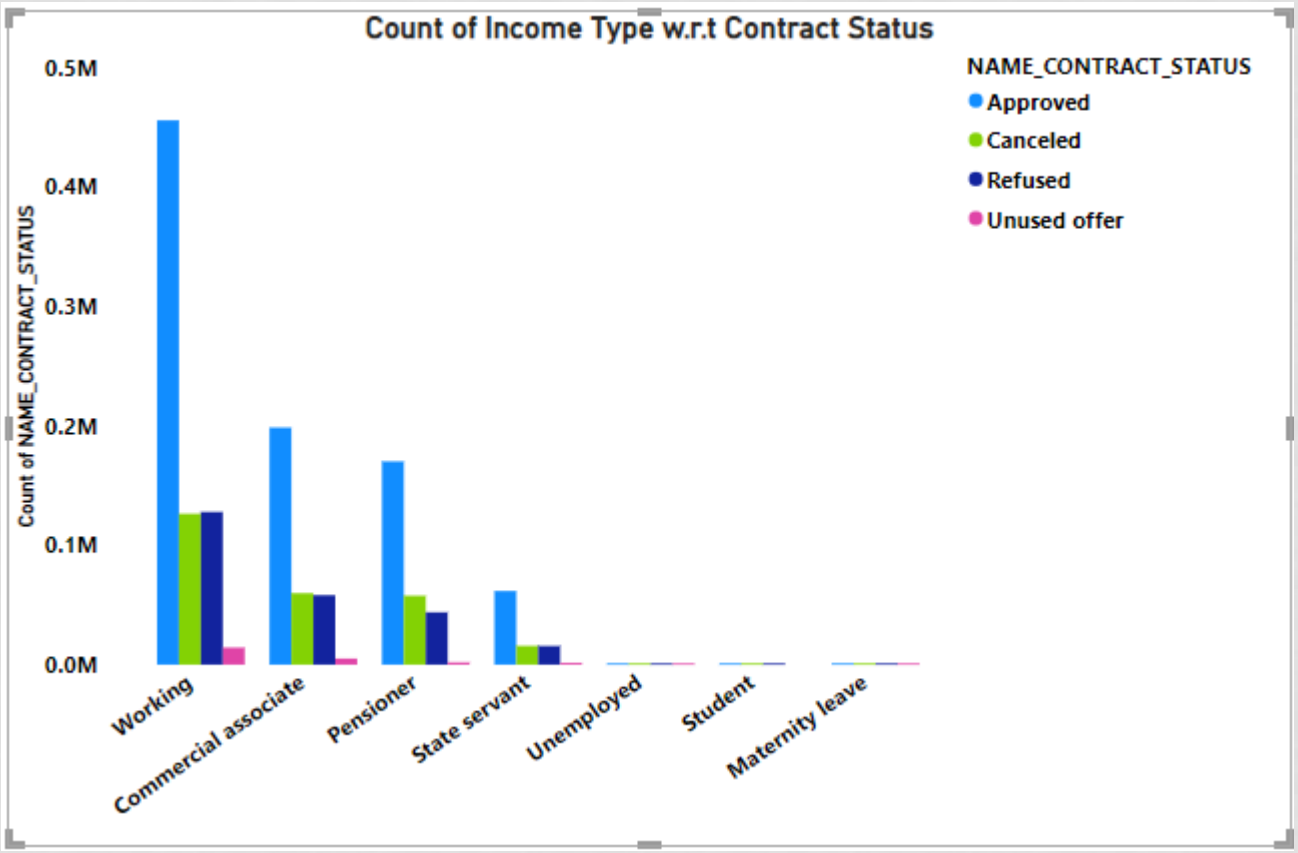
- **Maximum approved loans are for female clients**
- **It can also be observed that male clients use most of the offers of loans as unused offers are very less for male clients than that of female clients.**



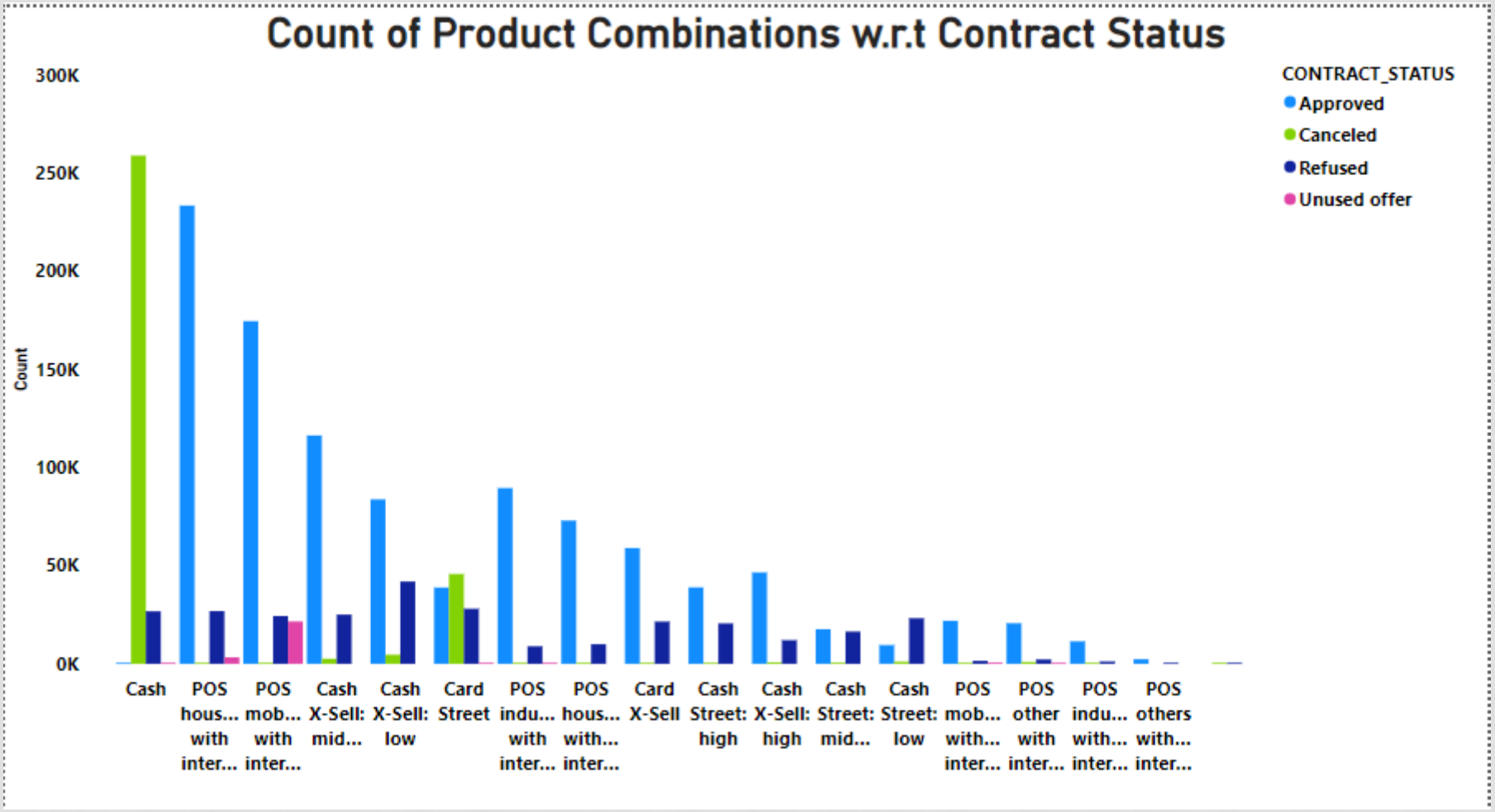
- Contract type cash loans are maximum in number where all kinds of contract statuses are more than contract type revolving loans
- Consumer loans are most approved loans.



Most clients with all kinds of education types have approved rate is better than canceled, refused rate.

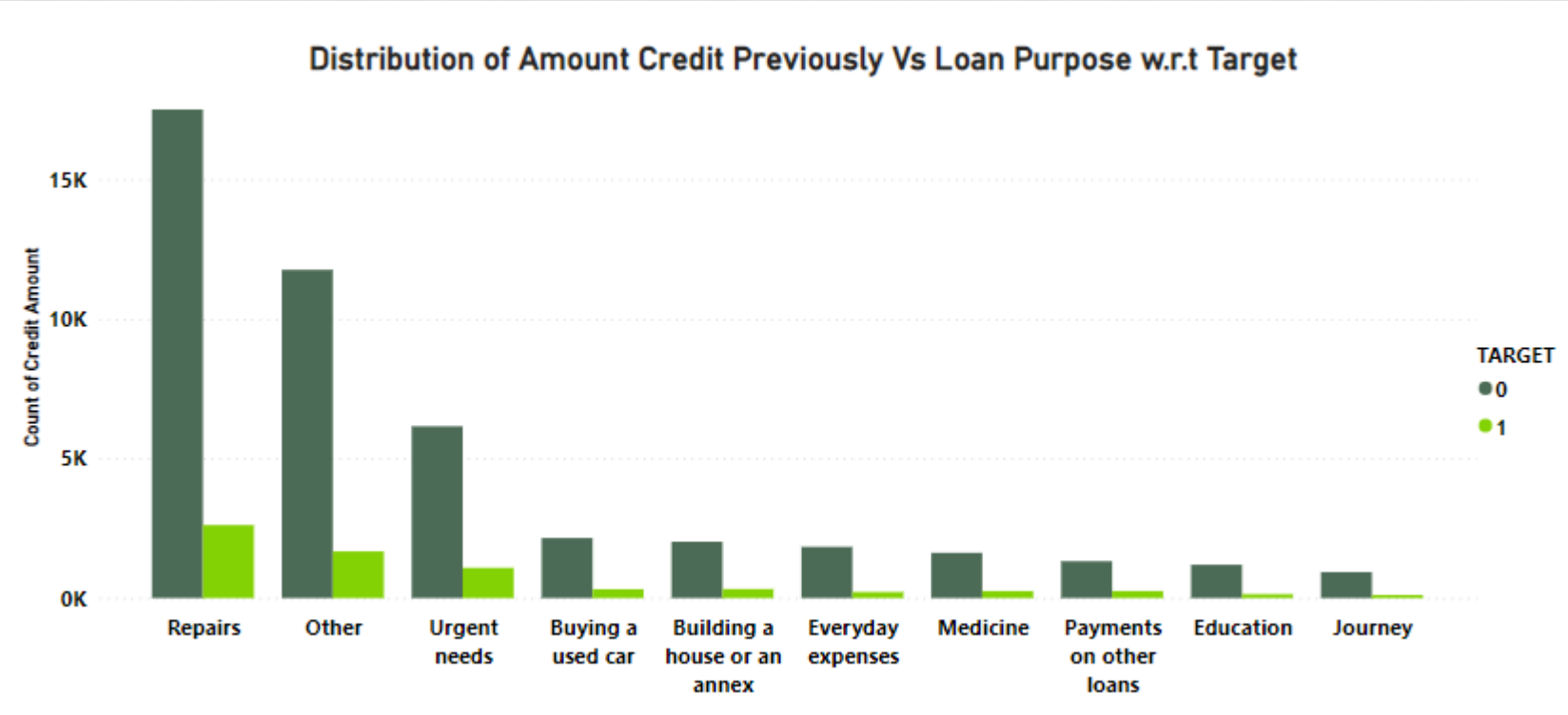


- There are very less rate of getting approved and unused offer for students and for maternity leave income type.
- The number of approved loans for state servants is almost equal to the refusal or canceled loans for Commercial associates.
- Maximum unused offers is by working clients and more approved is by working clients.

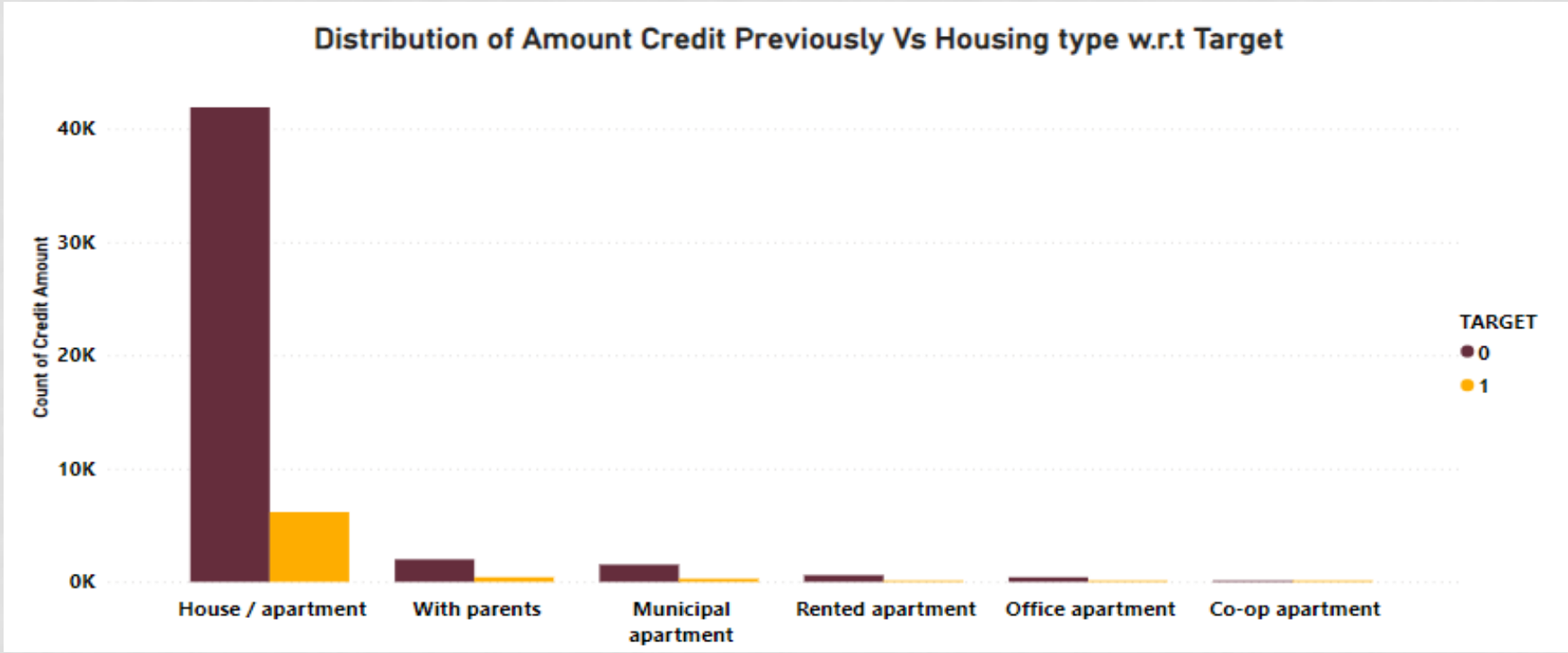


- Most canceled loans are of the product combination, Cash.
- Most refused loans are of product combination Cash X-Sell: low
- Some product combinations are null incase of unused offers as well as cancelled loans
  - POS industry without interest
  - POS others without interest

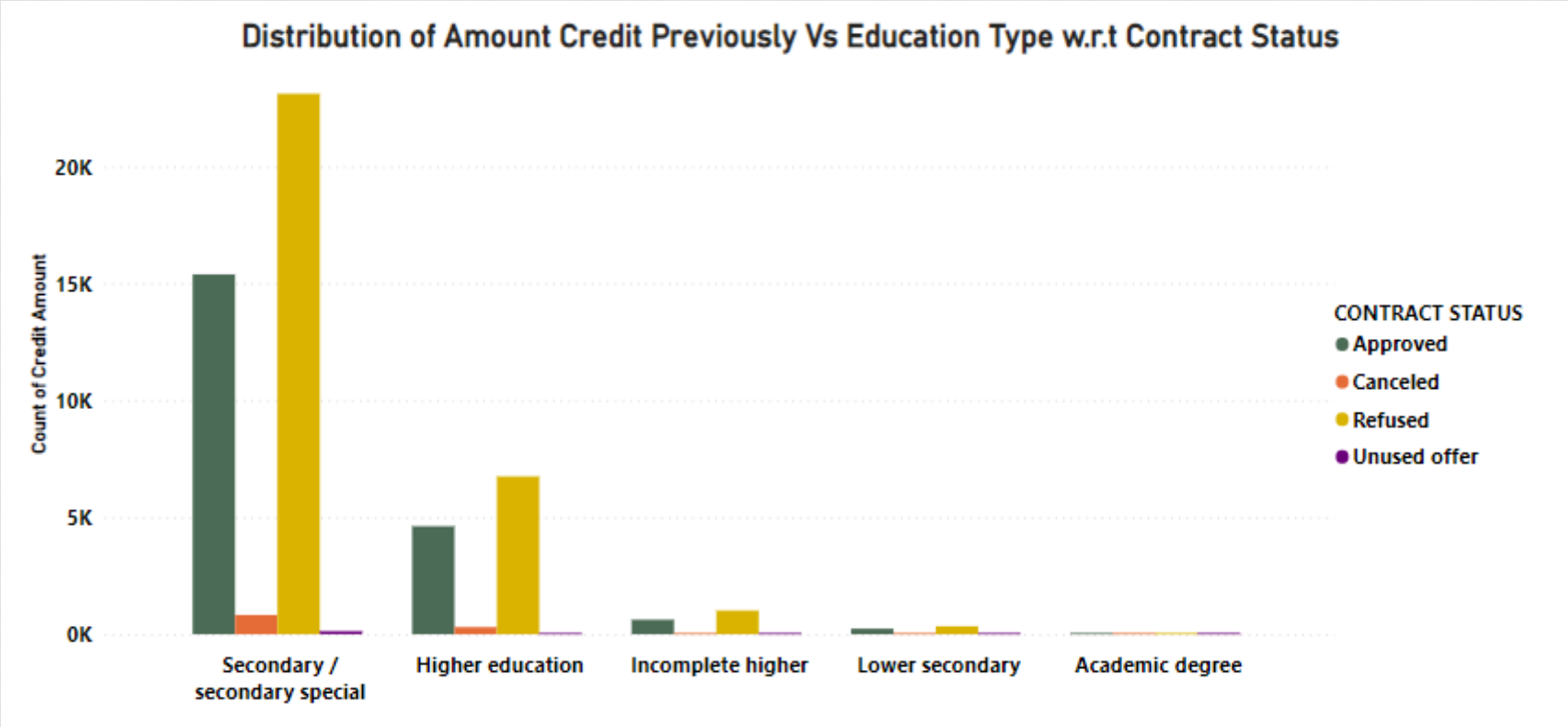
Bivariate Analysis for previous_application.csv



- The credit amount of Loan purposes like ‘urgent needs’, ‘buying a used car’ and ‘building a house’ are higher and also they are having no payment difficulties.
- Repairs loan purpose have maximum number of payment difficulties clients and maximum number of no payment difficulties clients.



- Here for Housing type, house/apartment is having higher credit of target 1.
- So, we can conclude that bank should avoid giving loans to the housing type of office apartment , rented apartment and co-op apartment as they are having difficulties in payment.
- Bank can focus mostly on housing type with parents or House or apartment or municipal apartment for successful payments.



It can be seen that number of refusals are for clients has nothing to do with their education levels

Almost all education level clients have equal unused offers.

Results

- Bank should focus on education type higher education and secondary/secondary special who took cash loan or revolving loan are capable of successful payment on time.
- All type of organization don't have any payment difficulties. Except business entity type 3 and self employed have some payment difficulties but it is negligible.
- Contract type revolving loans have leas client with payment difficulties.
- Bank should focus on income type working, commercial associate and pensioner they are capable of successful payment back to bank.
- Housing type house/apartment, with parents and munsipal apartment have more clients with no payment difficulties than payment difficulties.


The end