Bank Loan Case Study

Arcot Navya Sai (navyasaipatnaik@gmail.com) +91-8074851394

Project Description: This Project is about handling the unclean data in the industry where we will be focusing on Exploratory Data Analysis on the given Dataset. As the case study we are dealing with is the Bank Loan Case Study we will be developing 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.

In the current Dataset we are having three datasets namely

1. **`application_data.csv`** contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
2. **`previous_application.csv`** contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.
3. **`columns_descrption.csv`** is data dictionary which describes the meaning of the variables.

Approach: As we are handling the unclean data our first step is to understand the data and the columns we are having three datasets where one dataset columns_descrption.csv is totally explaining about columns of the other two datasets namely application_data.csv, previous_application.csv.

After understanding the data, we will explore the columns and description of the data along with info and later we will be handling the missing data in the columns and rows, we will also drop the irrelevant rows and columns.

Later we will do the analysis of outliers and remove them followed by Univariate and Bivariate analysis on the discrete and the continuous variables Finally we summarize the data and the analysis by visualizing the relations in excel.

Tech-Stack Used: In this Case Study We have used

1. Microsoft Excel Professional Plus 2016 version
2. Jupyter Notebook from Anaconda 3
3. Python 3
4. Pandas
5. Numpy etc. and all the necessary libraries in python

Insights: Initially we have started with exploring the columns, info, description etc. Of the dataset given in the Jupyter Notebook.

Missing values in application Dataset: https://drive.google.com/file/d/1jI5YVqZ-AiRK7LhOvOjzX-JnKhKmJcsT/view?usp=sharing

We have removed the columns which were irrelevant for the data Analysis

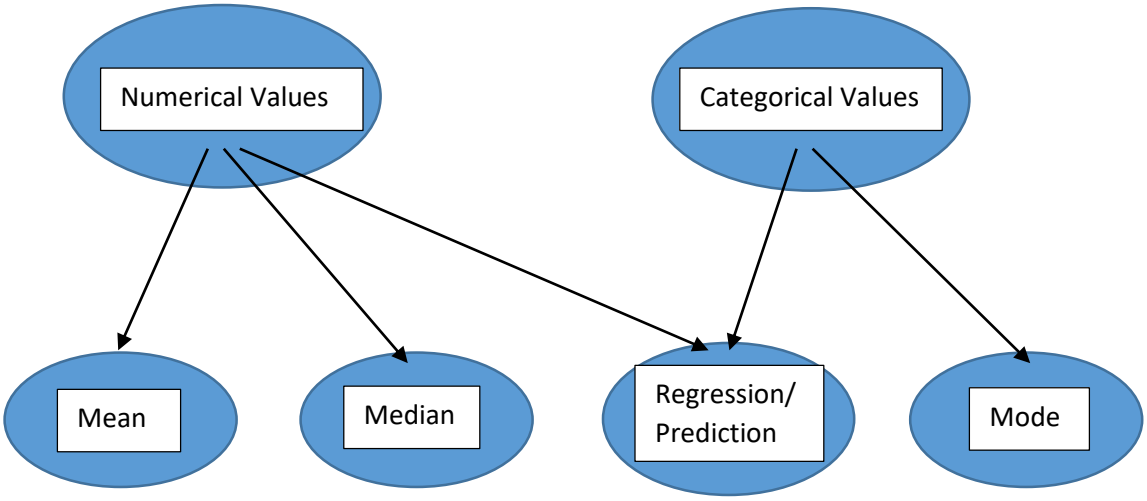
'DAYS_REGISTRATION','FLAG_MOBIL','FLAG_EMP_PHONE','FLAG_WORK_PHONE','FLAG_CONT_MOBILE',FLAG_PHONE','FLAG_EMAIL','WEEKDAY_APPR_PROCESS_START','HOUR_APPR_PROCESS_START','LIVE_REGION_NOT_WORK_REGION','REG_CITY_NOT_LIVE_CITY','REG_CITY_NOT_WORK_CITY','LIVE_CITY_NO

T_WORK_CITY','DAYS_LAST_PHONE_CHANGE','OBS_30_CNT_SOCIAL_CIRCLE','DEF_30_CNT_SOCIAL_CIRCLE','OBS_60_CNT_SOCIAL_CIRCLE','DEF_60_CNT_SOCIAL_CIRCLE','NAME_TYPE_SUITE'

After the Whole Cleaning Process, the size of the application data set is (307511, 51) nearly 58% of columns were removed many were irrelevant and others had large missing values

Then we checked for the rows having more than 50% of missing values.

How did we handle the missing values?

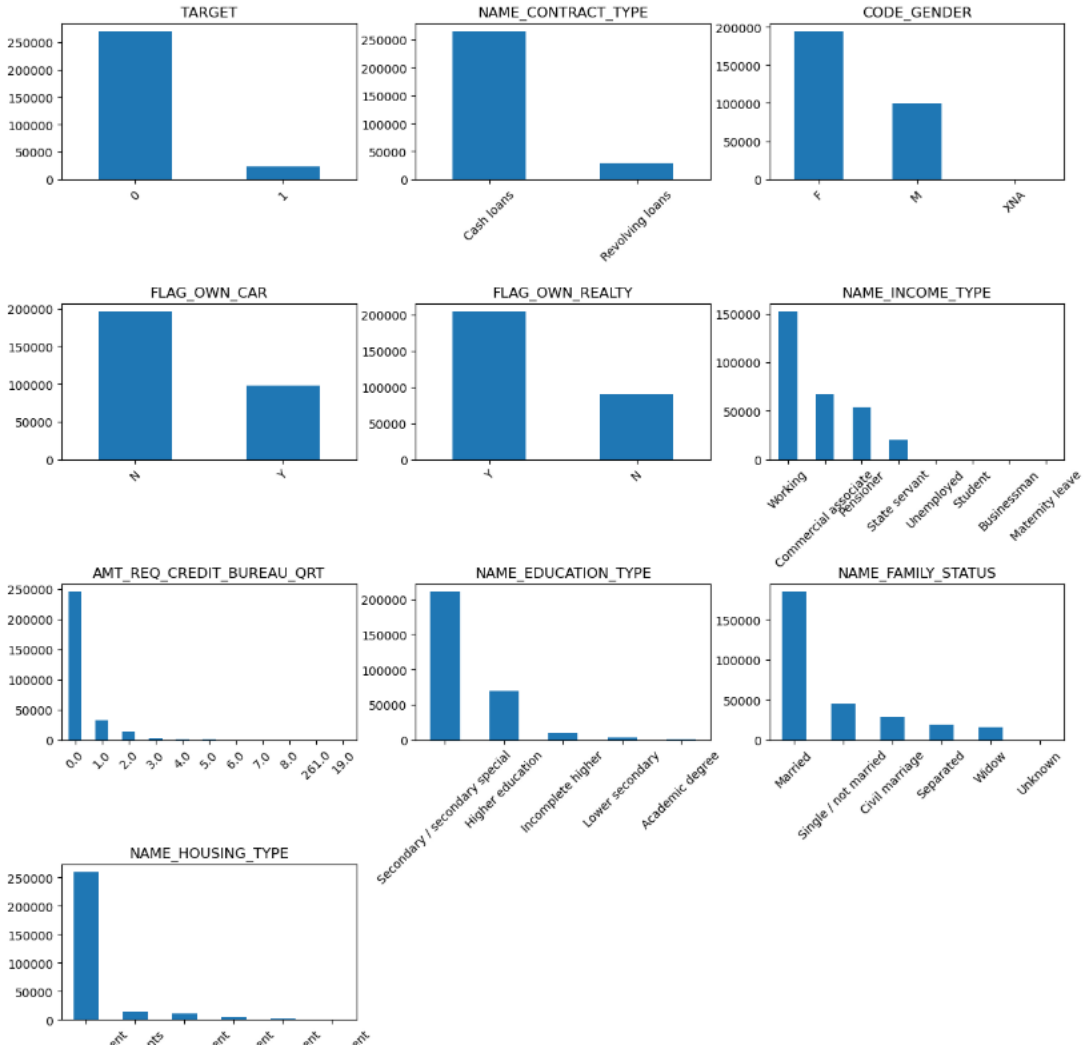


Outlier Analysis: For the Outliers we have used the Z score to Identify and remove the outliers in the numerical columns. (294063, 53) is the shape of the Dataset. After the outlier analysis 4.37% of rows got deleted now we have the data set where there are no missing values and the outliers.

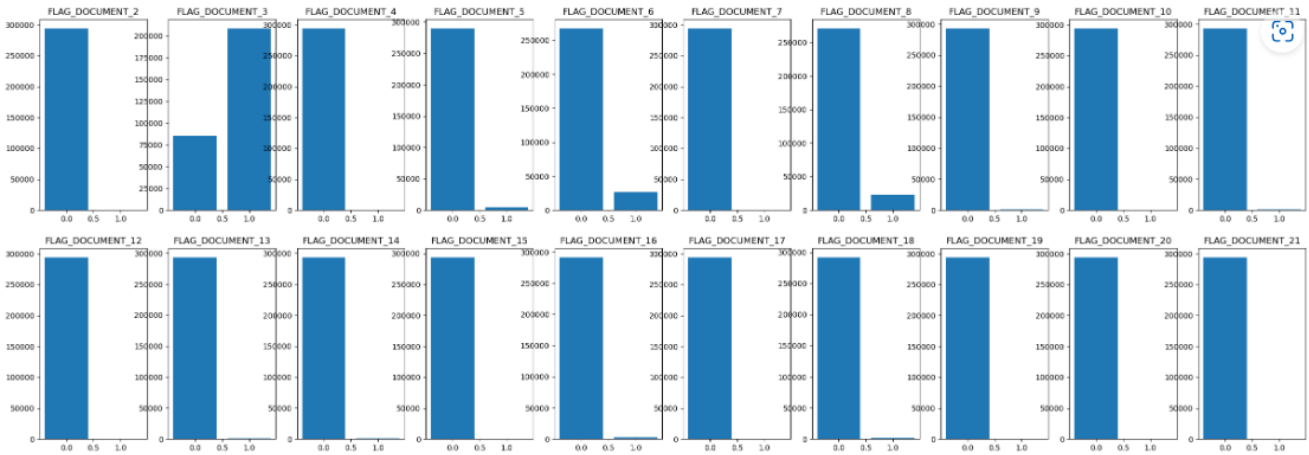
Data Imbalance:

Then we did the Data Imbalance Analysis to check the Numerical Data and Categorical Data
We can see that there is data imbalance in below columns: -

TARGET - There are very few defaulters (1) compare to non-defaulters (0)
NAME_CONTRACT_TYPE - There are very few Revolving loans than Cash loans
NAME_EDUCATION_TYPE - Most of the loans applied by Secondary/Secondary special educated people
NAME_FAMILY_STATUS - Most of the loans applied by Married people.
NAME_HOUSING_TYPE - Most of the application came from Home/apartment owner



The we understood the Flag data columns Data Imbalance



There is the data only in Flag Document 3

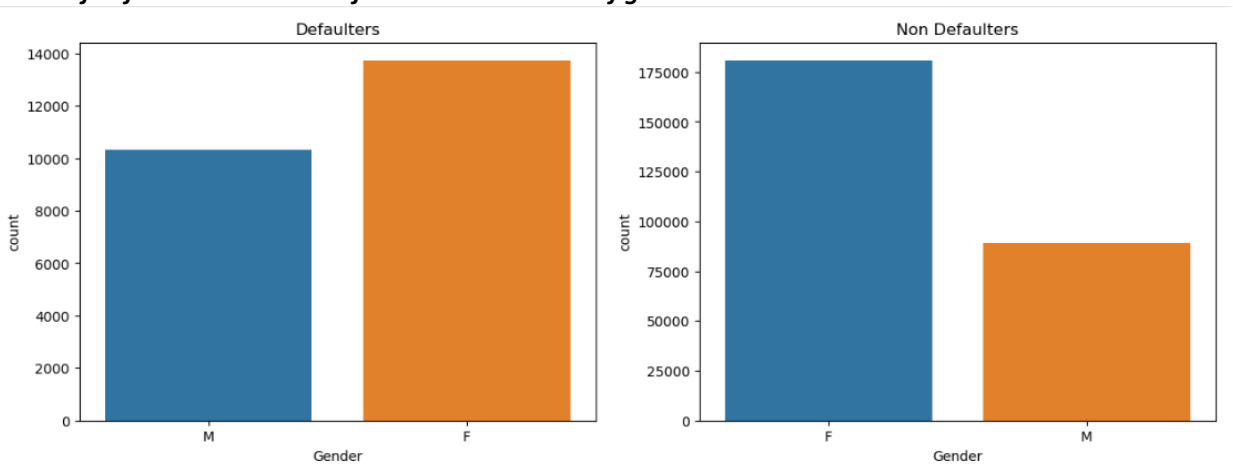
Binning the columns: Binning is the process of grouping a set of numerical values into a smaller number of "bins" or "buckets.

Binning of AGE, AMT_INCOME_TOTAL, AMT_CREDIT and EXT_SOURCE_SCORE columns

AGE	Years_Employed	Age_Group	Income_Group	Credit_Group	EXT_SOURCE_SCORE	Ext_source_group
26	2	Young	High	Low	0.20	Low
46	3	Mid Age	High	High	0.52	Medium
52	1	Mid Age	Low	Low	0.54	Medium
52	8	Mid Age	Medium	Low	0.60	High
55	8	Mid Age	Medium	Medium	0.38	Low

Univariate analysis:

Count of defaulters and non-defaulters on the basis of gender

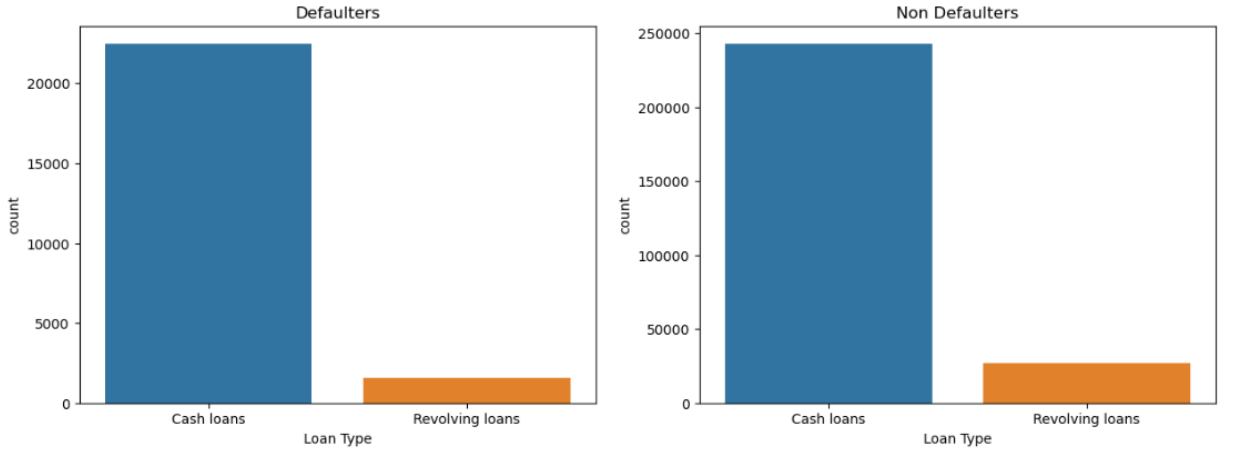


Analysis

1.Defaluters - We can see that females are slightly more in number of defaulters than male.

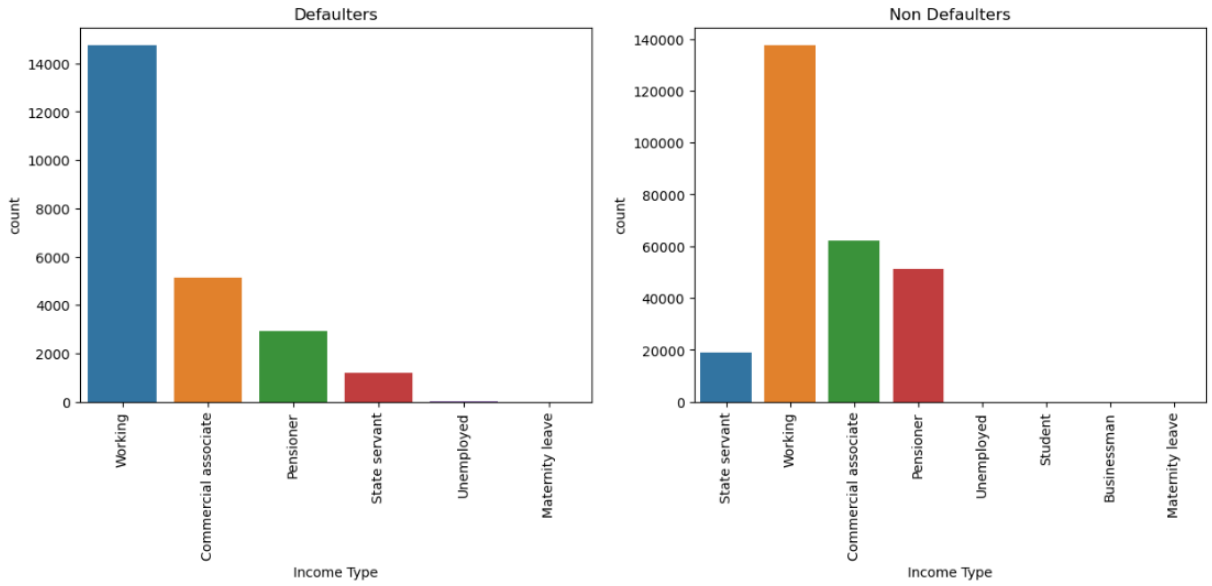
2.Non-defaluters - The same pattern continues for non-defaulters as well. The females are more in number here than male.

Count of defaulters and non-defaulters on the basis Loan Type



As we can see the Revolving loans are very less in number both in Defaulters and non-Defaulters

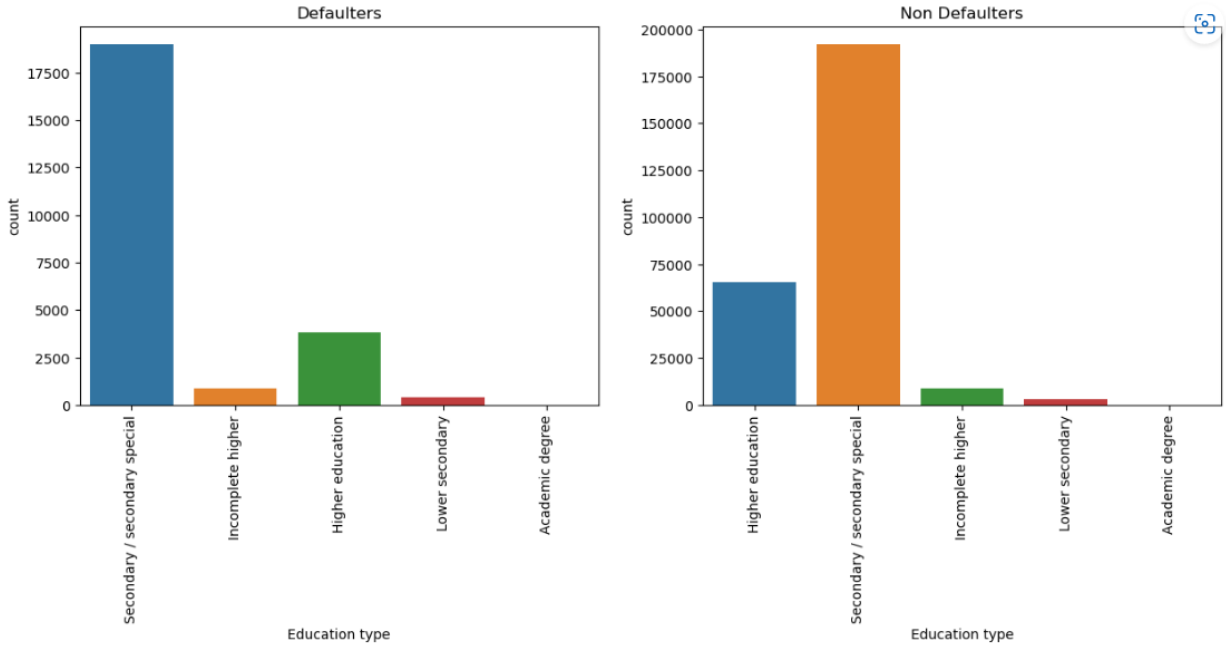
Count of defaulters and non-defaulters on the basis Income Type



1. As we can see from the above plot Working type of income is highest in number both in Defaulter and non-defaulters

2. And also State Servant is comparatively low among all the types

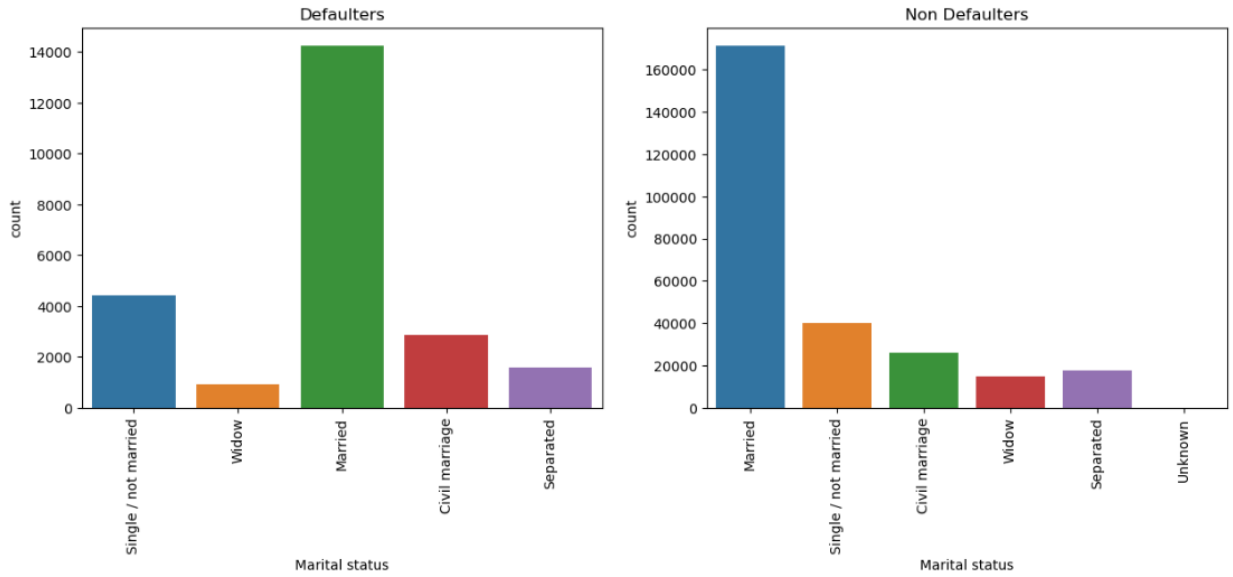
Count of defaulters and non-defaulters on the basis Education type



Analysis

The Secondary Special is comparatively high both in defaulters and non-defaulters

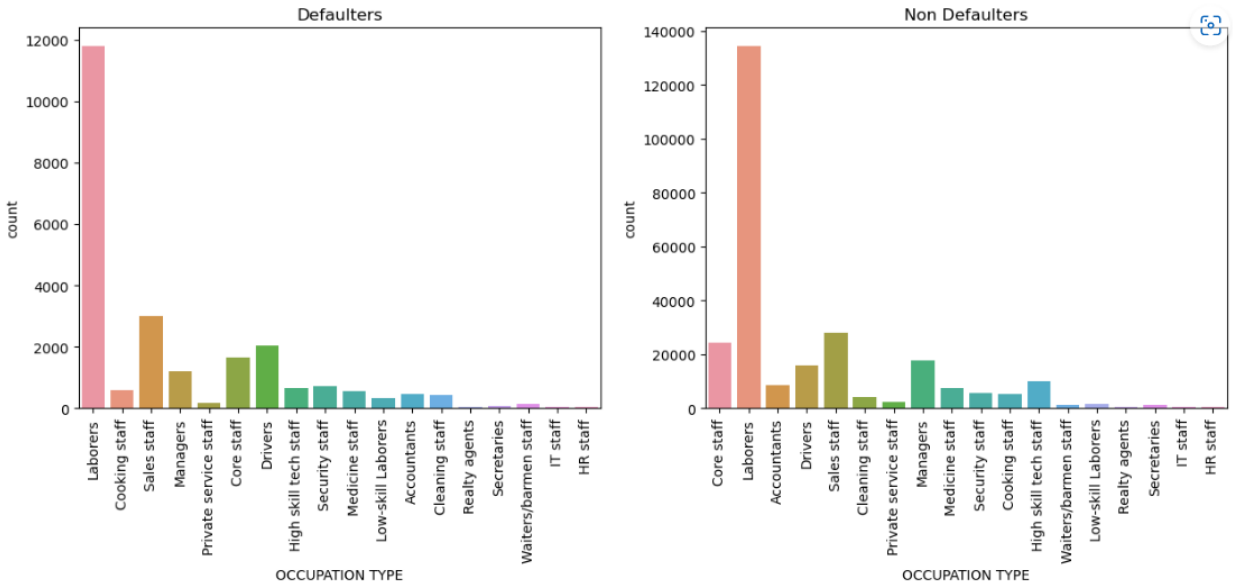
Count of defaulters and non-defaulters on the basis Marital status



Analysis

We can clearly understand that the marital status as Married are in more number both in defaulters and non-defaulters.

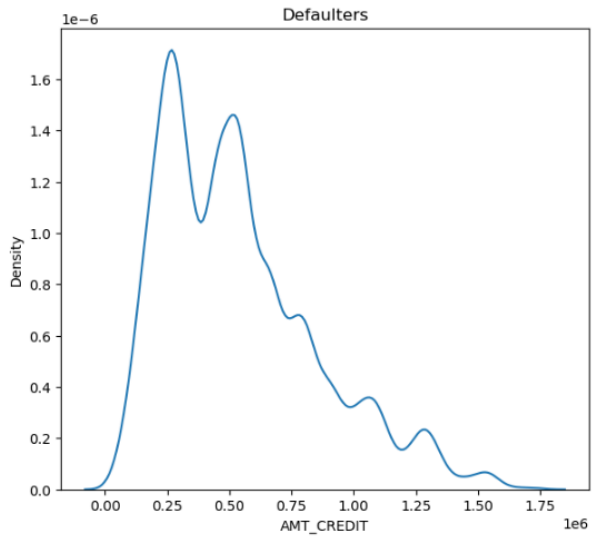
Count of defaulters and non-defaulters on the basis OCCUPATION_TYPE

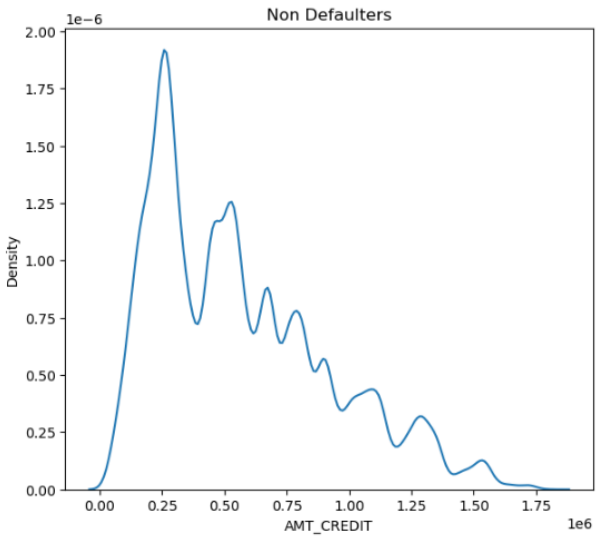


Analysis

Laborers is having more number comparing with other category both in defaulters and non-defaulters

Defaulters and non-defaulters on the basis of credit amount of the loan



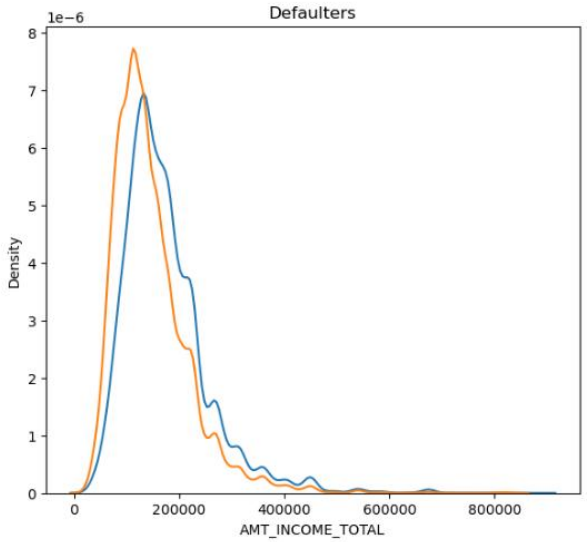


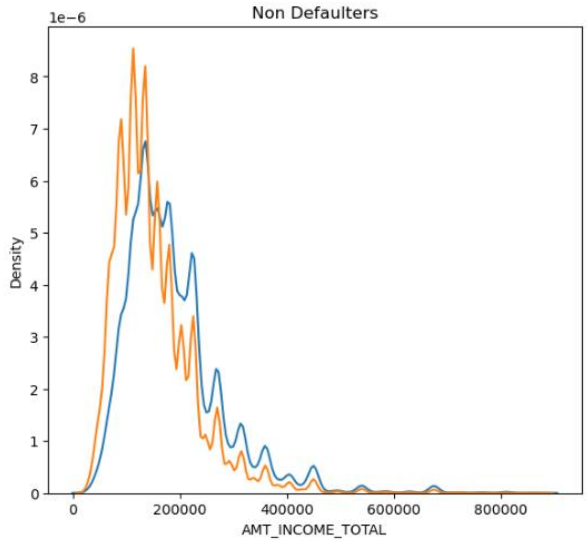
Analysis

Defaulters - We can notice that the lesser the credit amount of the loan, the more chances of being defaulter. The spike is till 500000.

Non defaulters - If the credit amount is less, there is lesser chance of being defaulted. And gradually the chance is being decreased with the loan credit amount.

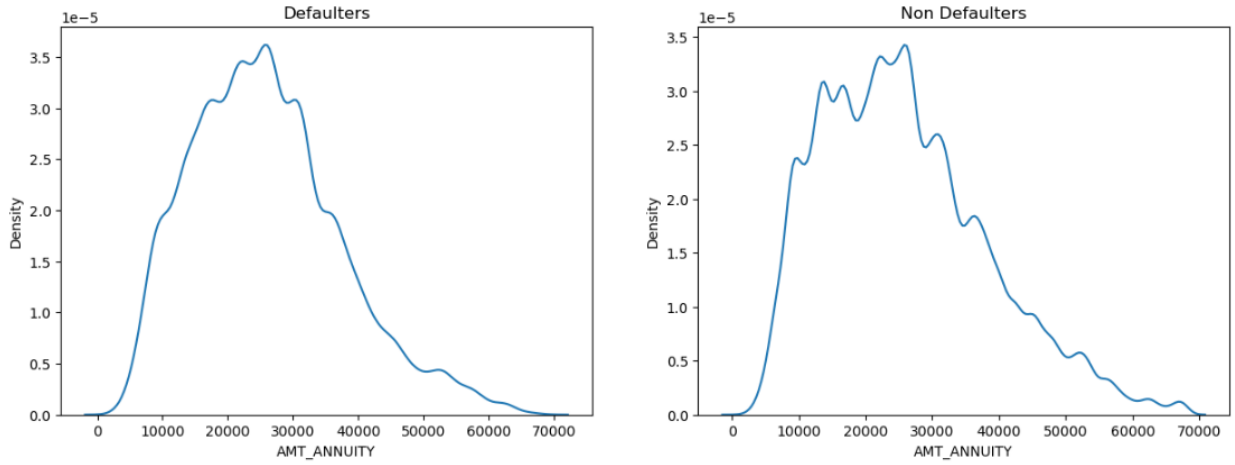
Defaulters and non-defaulters on the basis of gender and their total income





1. **Defaulters** - We can notice by looking at the pattern that for being a defaulter both the genders (male and female) are almost equal in all income levels. The spike of being defaulters is from 50000 to 200000.
2. **Non defaulters** - Here we see an interesting pattern. Females are more non defaulter on the lower income level but lesser non defaluter in higher income level. The spike is more for both the genders from 75000 to 150000.

Defaulters and non-defaulters on the basis of Loan annuity

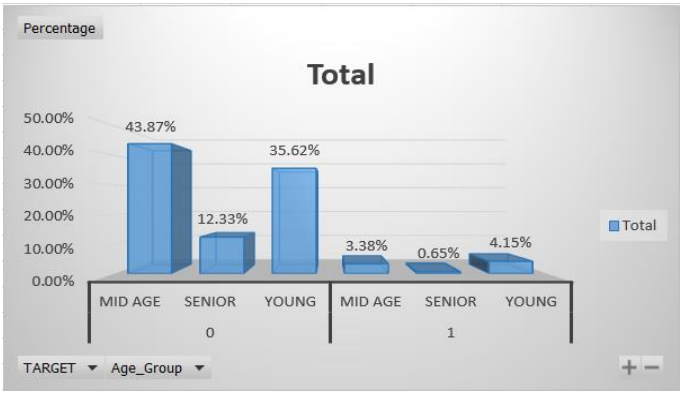


We can notice from the above distribution plot that in both the cases the loan annuity is concentrated more from 10000 to 40000.

Bivariate Analysis:

Age group & Target

Age_Group	Percentage
0	**91.82%**
Mid Age	43.87%
Senior	12.33%
Young	35.62%
1	**8.18%**
Mid Age	3.38%
Senior	0.65%
Young	4.15%
Grand Total	**100.00%**

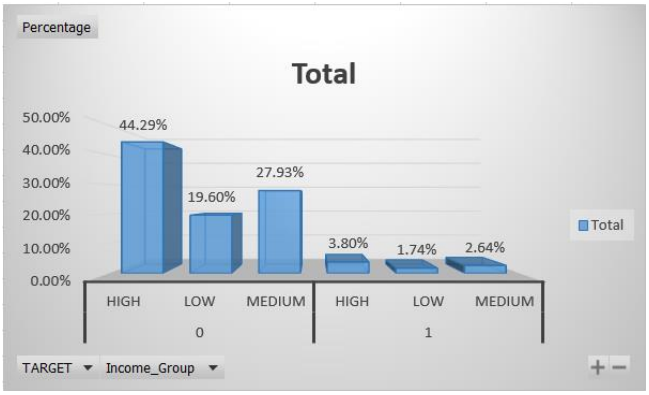


For defaulters, young age is given more preference compared to other groups.

For Non-defaulters, all age groups are given similar importance.

Income group & Target:

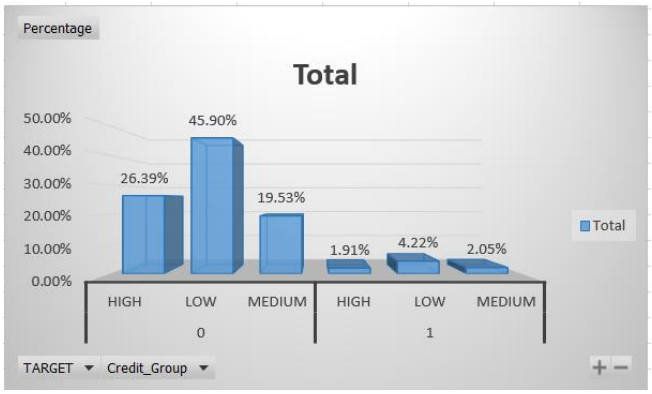
Income_Group	Percentage
0	91.82%
High	44.29%
Low	19.60%
Medium	27.93%
1	8.18%
High	3.80%
Low	1.74%
Medium	2.64%
Grand Total	100.00%



Low Income Group is given the least importance in both defaulters and non-defaulters.

Credit group & Target:

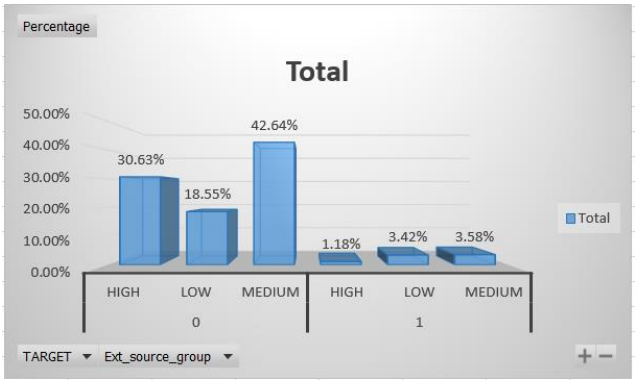
Credit_Groups	Percentage
0	91.82%
High	26.39%
Low	45.90%
Medium	19.53%
1	8.18%
High	1.91%
Low	4.22%
Medium	2.05%
Grand Total	100.00%



Credit groups having low value are more in number in both the defaulters and non-defaulters.

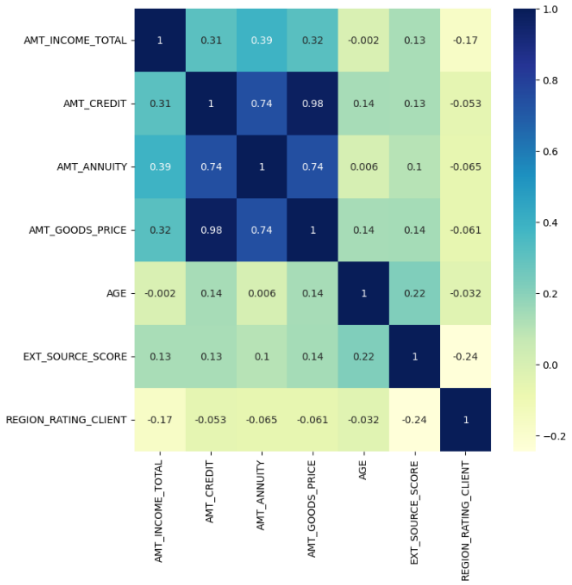
External Source Score group & Target:

Ext_score	Percentage
0	91.82%
High	30.63%
Low	18.55%
Medium	42.64%
1	8.18%
High	1.18%
Low	3.42%
Medium	3.58%
Grand Total	100.00%



Medium value dominates the range in both the categories.

Co-relation Analysis between top numerical columns:

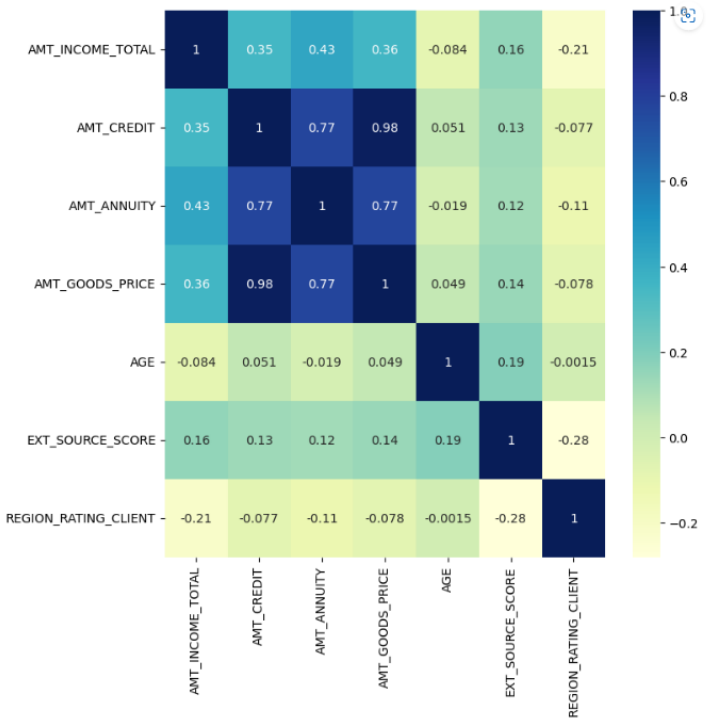


Highly co-relate columns for defaulters

AMT_CREDIT and AMT_ANNUITY (0.74)

AMT_CREDIT and AMT_GOODS_PRICE (0.98)

AMT_ANNUITY and AMT_GOODS_PRICE (0.74)



Highly correlate columns for non-defaulters

AMT_CREDIT and AMT_ANNUITY (0.76)

AMT_CREDIT and AMT_GOODS_PRICE (0.98)

AMT_ANNUITY and AMT_GOODS_PRICE (0.76)

Conclusion - We can see that for both defaulters and non-defaulters the same pairs of columns are highly correlated.

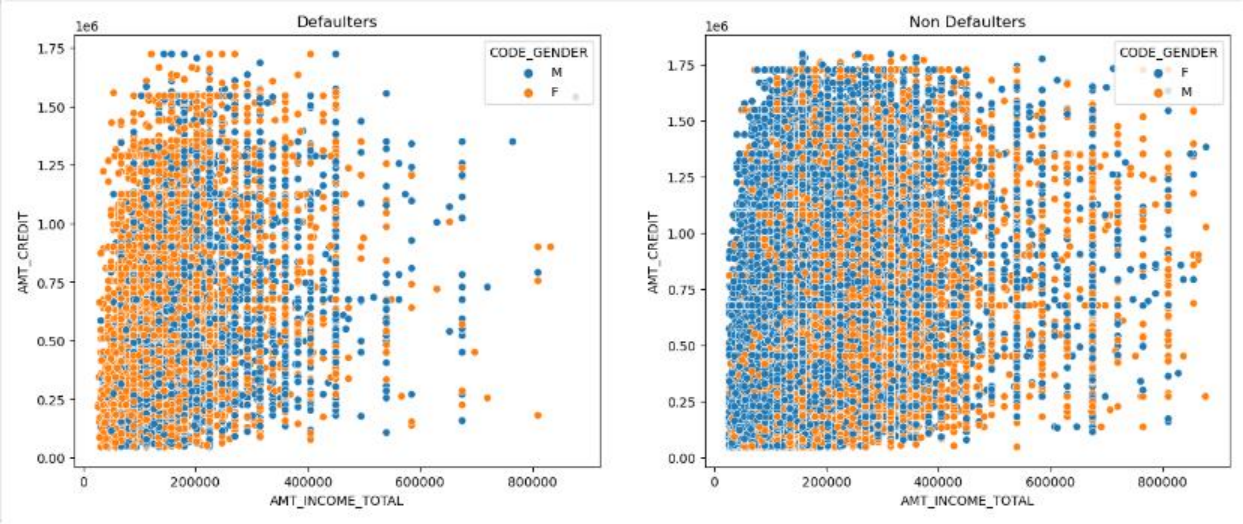
Bivariate analysis on continuous variable

Credit amount of the loan on the basis of client income for both male and female

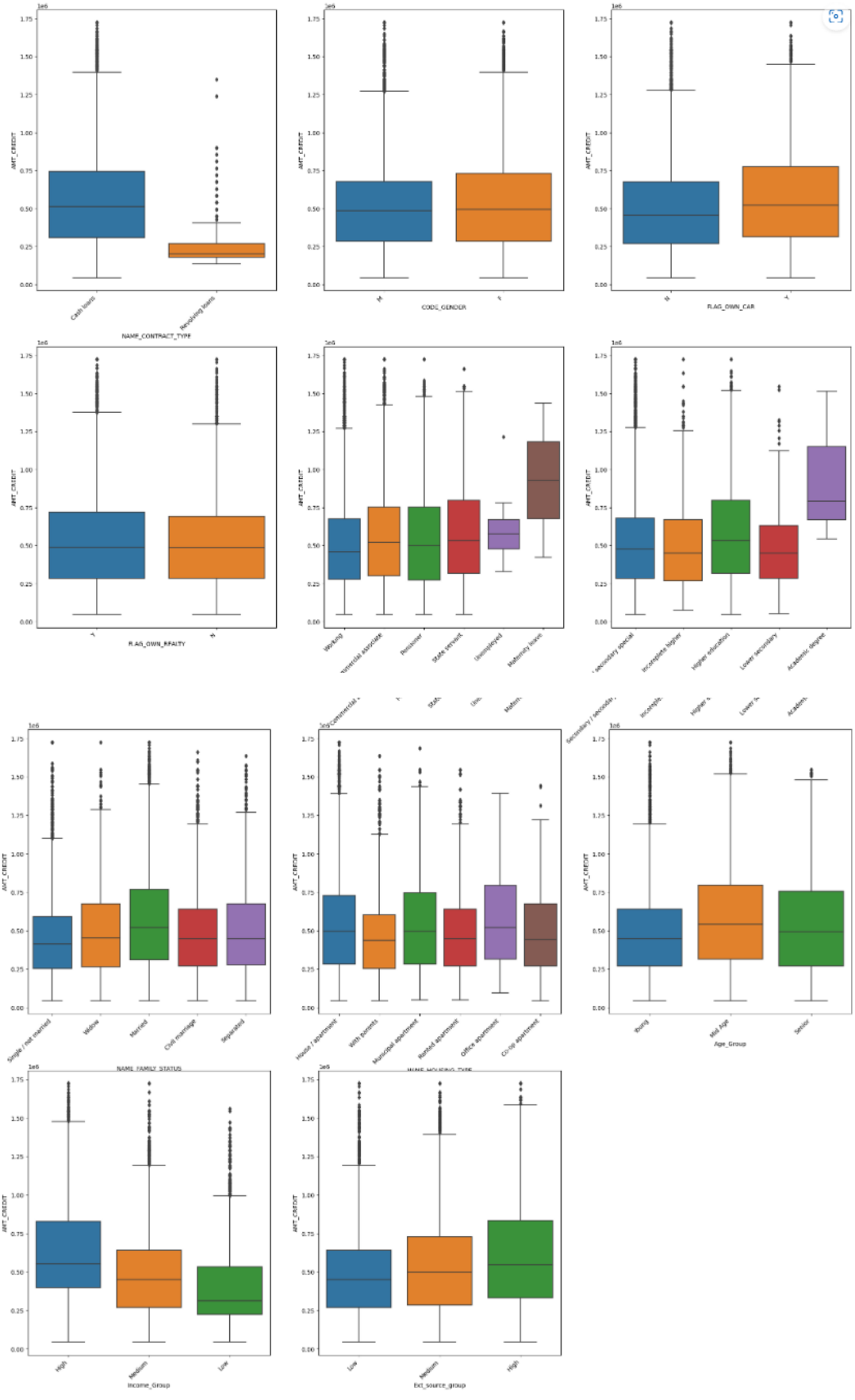
Analysis

Defaulters - We can slightly figure out that the values are more concentrated on the lower income and lower credit of the loan. That means as the income is increased, the amount of loan is also increased. This is true for both the genders.

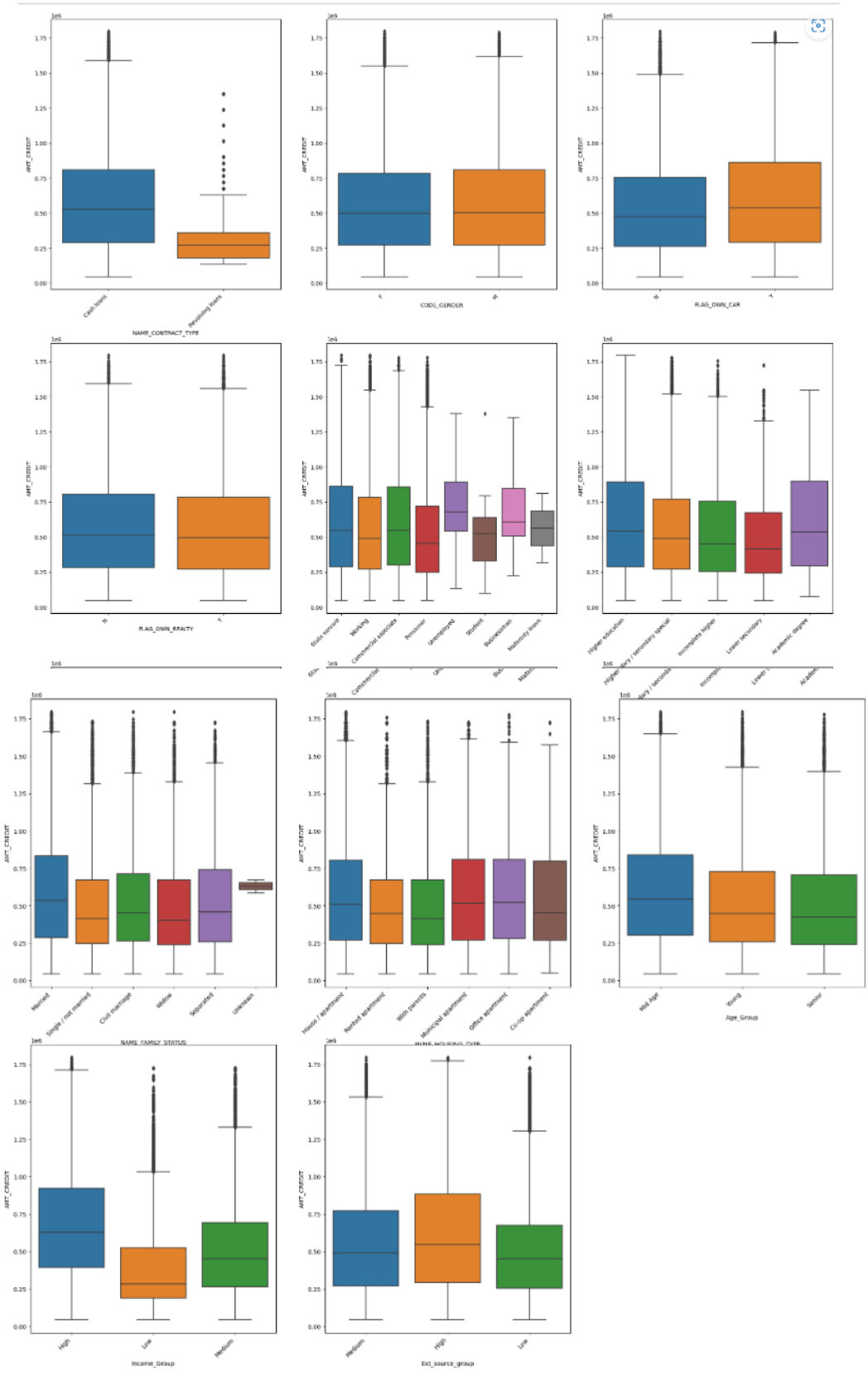
Non defaulters - We can hardly figure out any pattern out of this.



Credit amount of the loan of various categories:

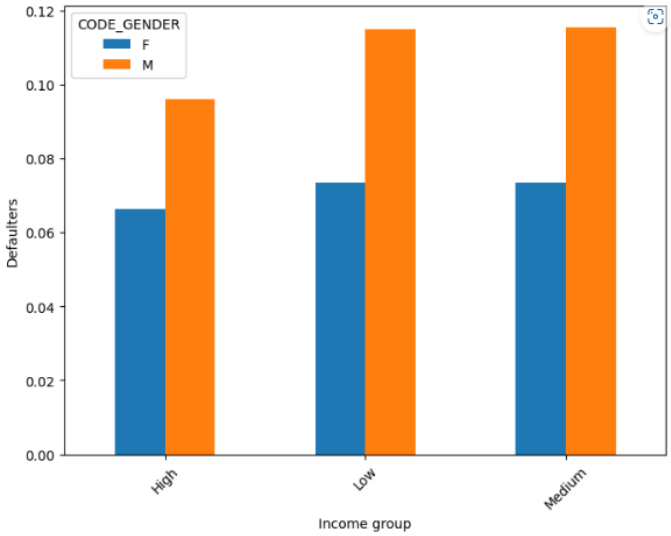

1e6
AMT_CREDIT
Cash loans
Revolving loans
NAME_CONTRACT_TYPE
CODE_GENDER
FLAG_OWN_CAR
FLAG_OWN_REALTY
Working
Pensioner
State servant
Unemployed
Maternity leave
Single / not married
Widow
Married
Civil marriage
Separated
House / apartment
With parents
Municipal apartment
Rented apartment
Office apartment
Co-op apartment
Young
Mid Age
Senior
Age_Group
High
Medium
Low
Income_Group
Ext_source_group

Non-defaulters



Analysis of two segmented variables

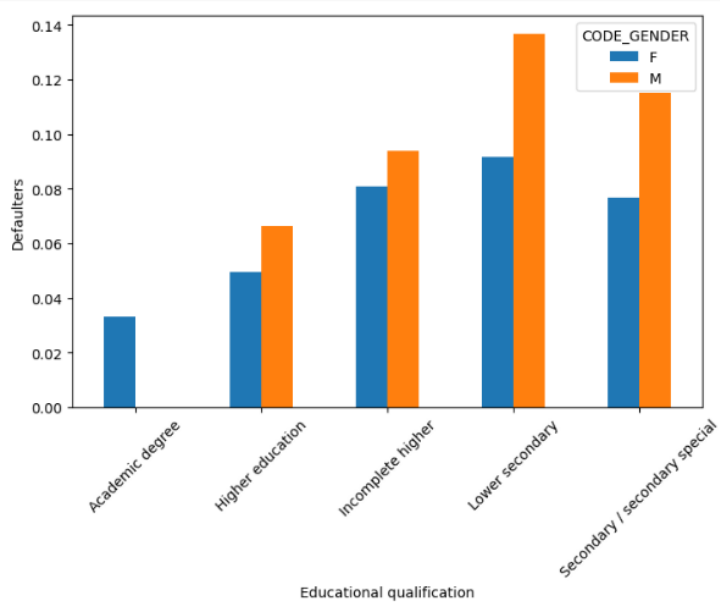
Income group and gender



Analysis

We can see that Males are more likely defaulted than Females across all income groups

Code Gender and Education Qualification



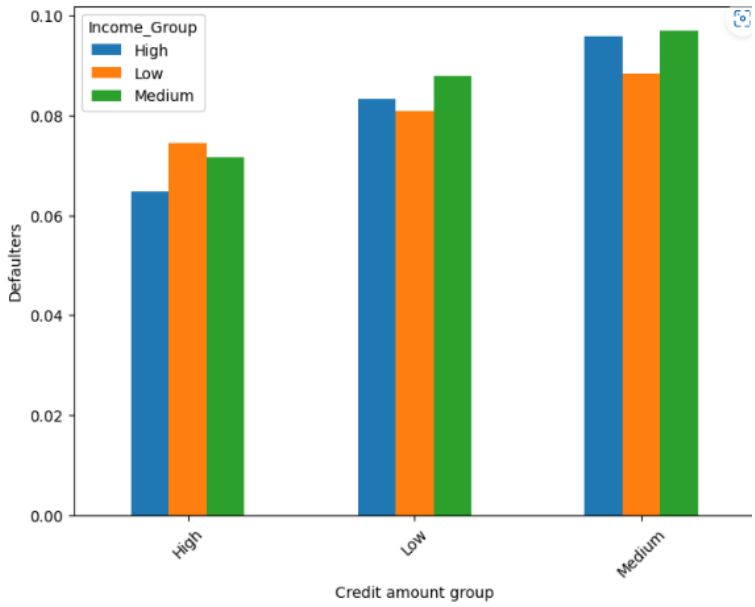
Analysis

Lower secondary educated clients are more defaulted followed by Secondary and Incomplete higher educated clients.

The Higher educated people are less defaulted.

Across all educated level Females are less defaulted than male.

Credit amount group and Income group

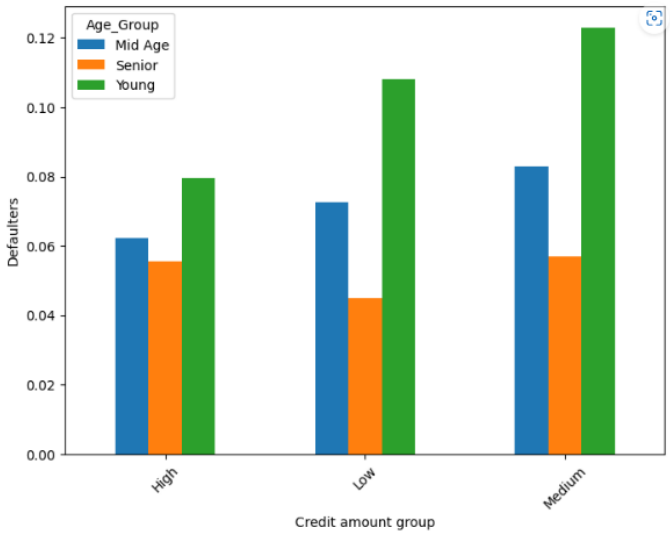


Analysis

Medium credit amount group are highly defaulted in all income groups.

High credit amount groups are less likely to default in all income groups.

Credit amount group and Age group

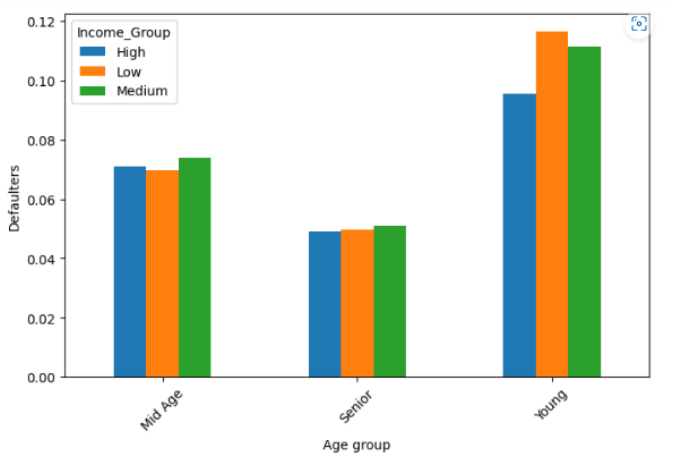


Analysis

Young clients with medium and low credit amount group are highly defaulted.

Senior citizens across all credit amount groups are less likely defaulted.

Age group and Income group



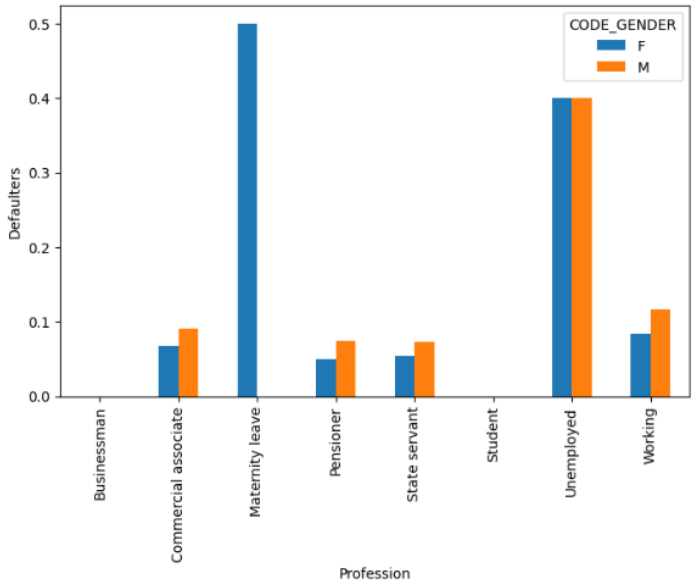
Analysis

Young clients are more defaulted than Mid age and senior.

Young low income people are more defaulted.

For Mid age and senior people, the default rate is almost same in all income group.

Profession and Gender



Analysis

No surprise the unemployed clients are more defaulted.

Clients with maternity leave are expected to be defaulted more.

The default rate is lesser in all other professions.

Males are more defaulted with their respective professions compared to females.

Analysis on Previous application:

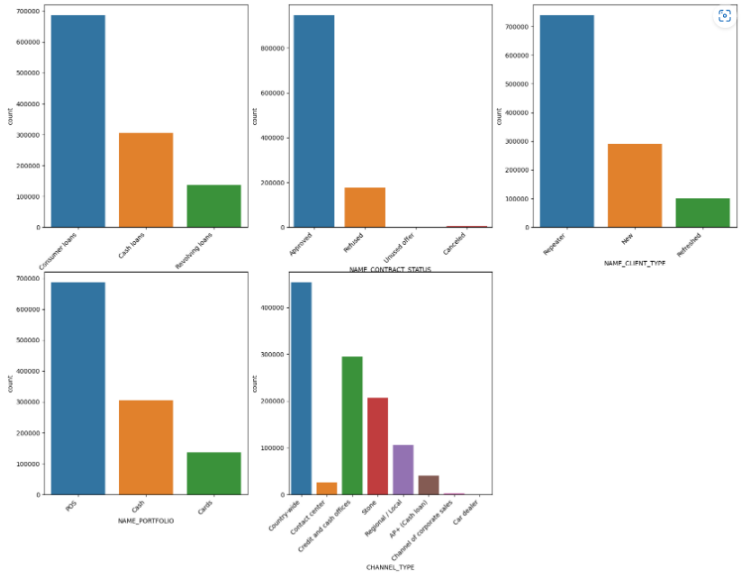
Previous application dataset size (1670214, 37)

Similar to the application_data we have also analyzed the previous application data

The outlier Analysis were also done: In our data cleaning exercise we have lost around 32% rows.

As this is huge dataset, we still have 1129387 records for analysis.

Checking data imbalance



We can see that there is data imbalance in below columns: -

NAME_CONTRACT_TYPE - There are very few Revolving Loans
NAME_CONTRACT_STATUS - There are very few Refused loans. Almost negligible Canceled loans.
NAME_CLIENT_TYPE - There are very few New applicants. Even fewer Refreshed applicants.
NAME_PORTFOLIO - Very few applications for Cards and Cars
CHANNEL_TYPE - Except Country-Wide, Credit and Cash offices and Stone all other channels are very few in number

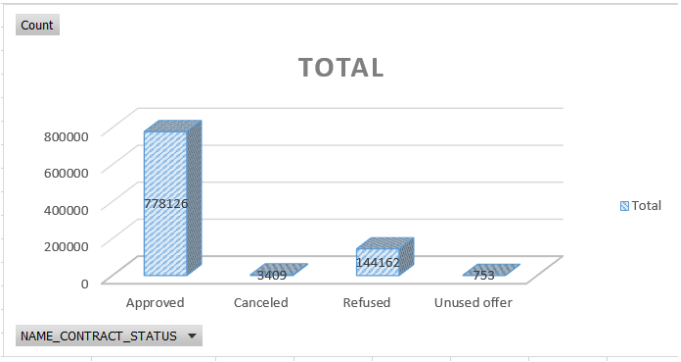
Merging Current application and Previous application dataset

All the initial data exploration is done in the merged dataset

https://drive.google.com/file/d/1YtliyTrEZNzTvbBTLiKxQWyFq6mQqUBz/view?usp=sharing

Univariate analysis on unordered categorical variable:
Previous Loan status

NAME_CONTRACT_STATUS	Count
Approved	778126
Canceled	3409
Refused	144162
Unused offer	753
Grand Total	**926450**

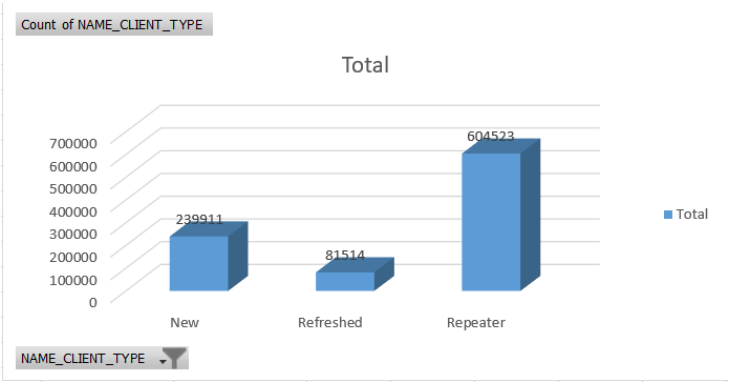


Analysis

There are huge number of Approved loan than Refused. Hardly, there are any Canceled or Unused offer loan.

Client type

NAME_CLIENT_TYPE	Count of NAME_CLIENT_TYPE
New	239911
Refreshed	81514
Repeater	604523
Grand Total	**925948**

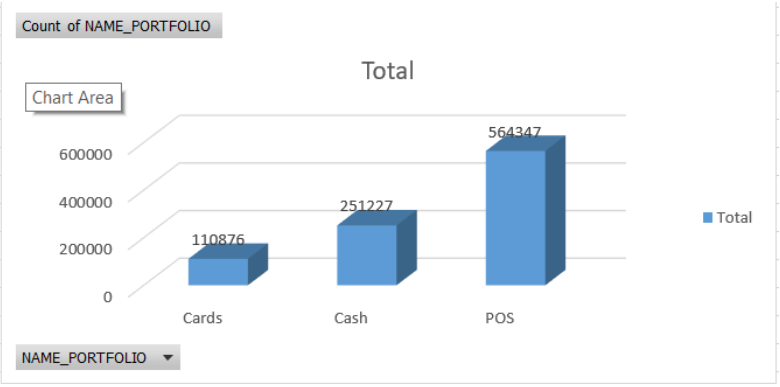


Analysis

Mostly the applicants were Repeater

Portfolio of the previous applications

NAME_PORTFOLIO	Count of NAME_PORTFOLIO
Cards	110876
Cash	251227
POS	564347
Grand Total	**926450**

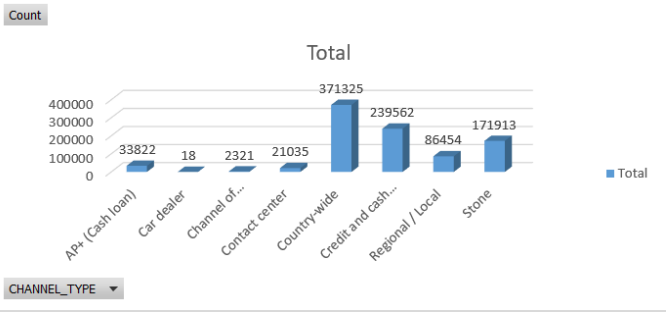


Analysis

The highest number of the previous applications was for POS. Applications for Cash also has good number. Applications for Cards were very few.

Application channel type

CHANNEL_TYPE	Count
AP+ (Cash loan)	33822
Car dealer	18
Channel of corporate sa	2321
Contact center	21035
Country-wide	371325
Credit and cash offices	239562
Regional / Local	86454
Stone	171913
Grand Total	**926450**



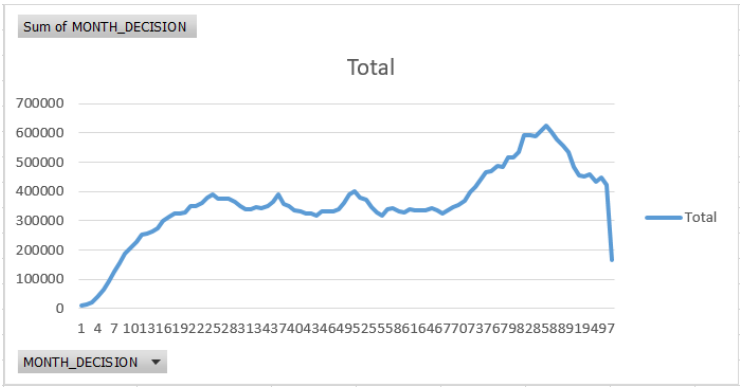
Analysis

We see that Country-wide was heavily used for previous applications followed by Credit and Cash offices, Stone and Regional. Rest other channels are hardly used.

Univariate analysis for continious variables

Months took for the pervious application decision relative to the current application

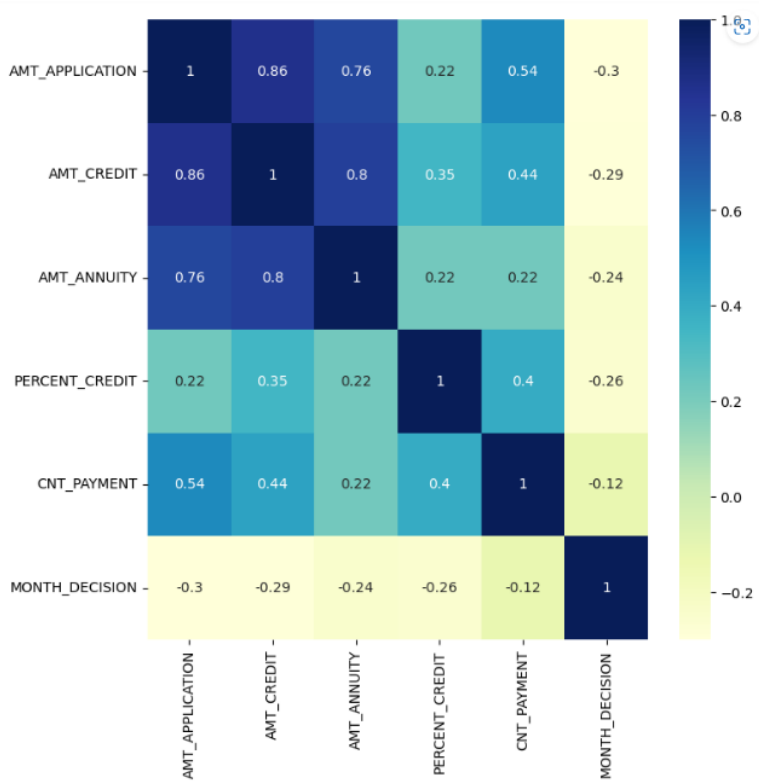
Row Labels	Sum of MONTH_DECISION
1	9001
2	15200
3	20949
4	40884
5	64680
6	94512
7	126490
8	154072
9	185760
10	205490



Analysis

We can see that most of the applications decision took approximately 30 months. The time taken spread up to 100 months

Correlation of relevant numerical columns

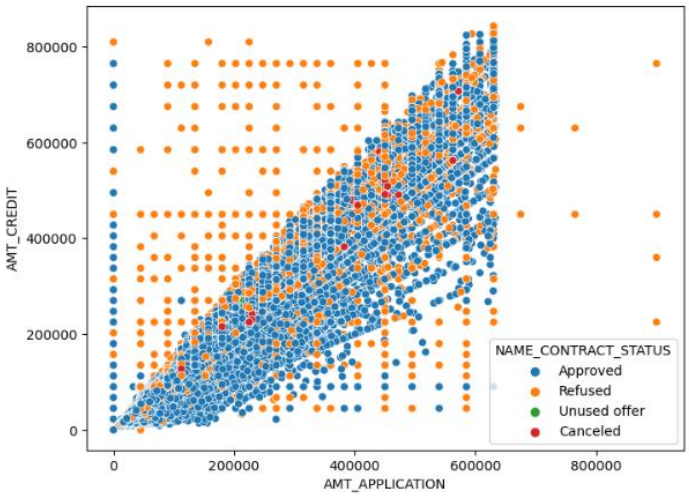


Highly correlate columns

AMT_APPLICATION and AMT_CREDIT
AMT_APPLICATION and AMT_ANNUITY
AMT_CREDIT and AMT_ANNUITY
Moderately correlated columns
AMT_APPLICATION and CNT_PAYMENT
AMT_CREDIT and CNT_PAYMENT

Bivariate analysis on continuous variable

Application amount and credited amount

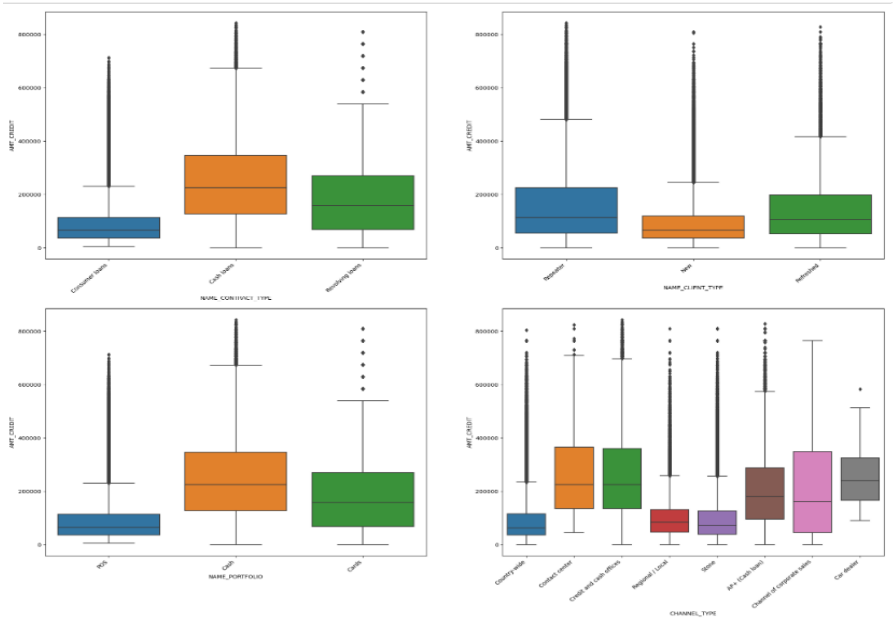


Analysis

We can see a pattern here that the more the application amount of the loan, the lesser the months taken prior to current application. That means, most of the higher amount of the loan application decision made in the recent time comapred to the lower loan amount application.

Bivariate analysis on categorical variable

Credit amount of the loan of various categories



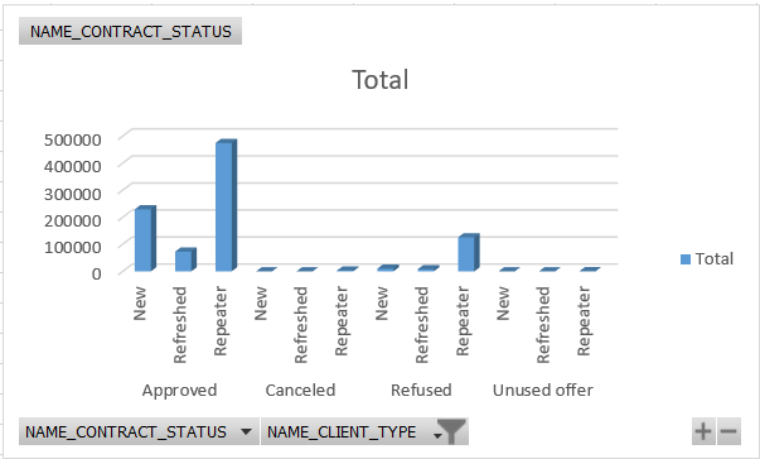
Analysis

1. Cash loans are more credited in amount than Revolving and Consumer loans.
2. Repeater clients get more amount loan than new and refreshed clients.
3. The loan with portfolio Cars are more amount credited followed by Cash.
4. The credit amount of the loan is more from the application channel type as car dealer followed by Channel of corporate sales, Credit and cash offices and Contact center. The amount is very less for Regional, Stone and Country-wide channels.

Analysis of two segmented variables

Status and Client type:

Row Labels	NAME_CONTRACT_STATUS
Approved	**777763**
New	229541
Refreshed	73282
Repeater	474940
Canceled	**3404**
New	114
Refreshed	308
Repeater	2982
Refused	**144029**
New	10234
Refreshed	7842
Repeater	125953
Unused offer	**752**
New	22
Refreshed	82
Repeater	648
Grand Total	**925948**

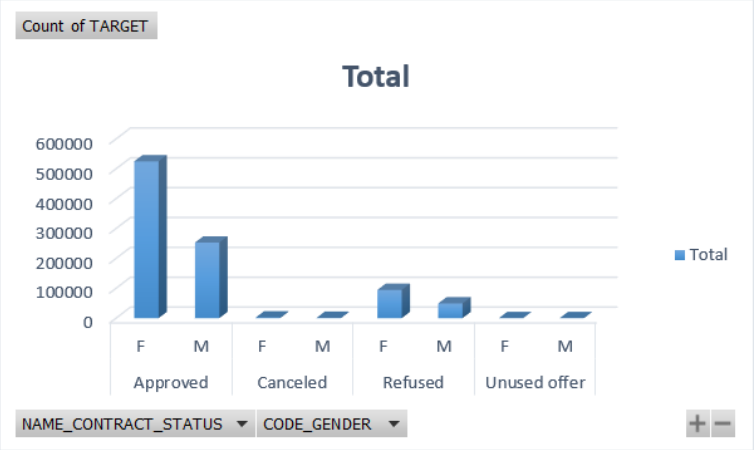


Analysis

We see that the Repeater clients have more approved loans than New and Refreshed clients.

Current loan defaulter status with respect to previous loan application status

Row Labels	Count of TARGET
Approved	**778126**
F	525008
M	253118
Canceled	**3409**
F	2399
M	1010
Refused	**144162**
F	94665
M	49497
Unused offer	**753**
F	449
M	304
Grand Total	**926450**

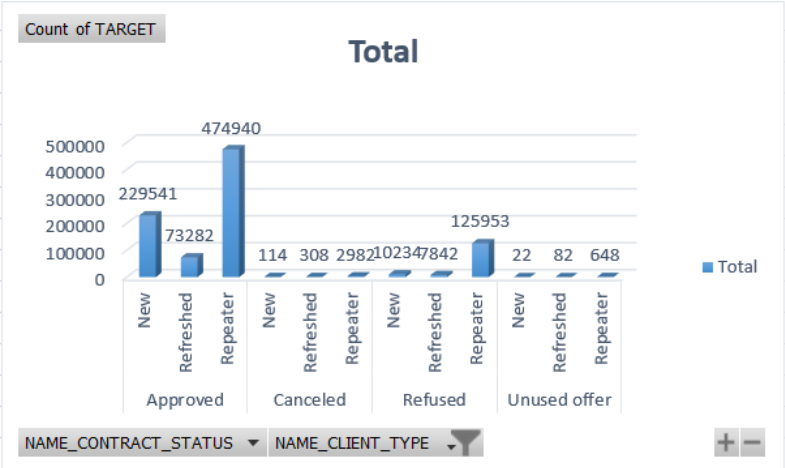


Analysis

We see that previously Refused client is more defaulted than previously Approved clients. Also, in all the cases the Males are more defaulted than Females.

Current loan defaulter status with respect to previous loan application status and client types

Row Labels	Count of TARGET
Approved	**777763**
New	229541
Refreshed	73282
Repeater	474940
Canceled	**3404**
New	114
Refreshed	308
Repeater	2982
Refused	**144029**
New	10234
Refreshed	7842
Repeater	125953
Unused offer	**752**
New	22
Refreshed	82
Repeater	648
Grand Total	**925948**

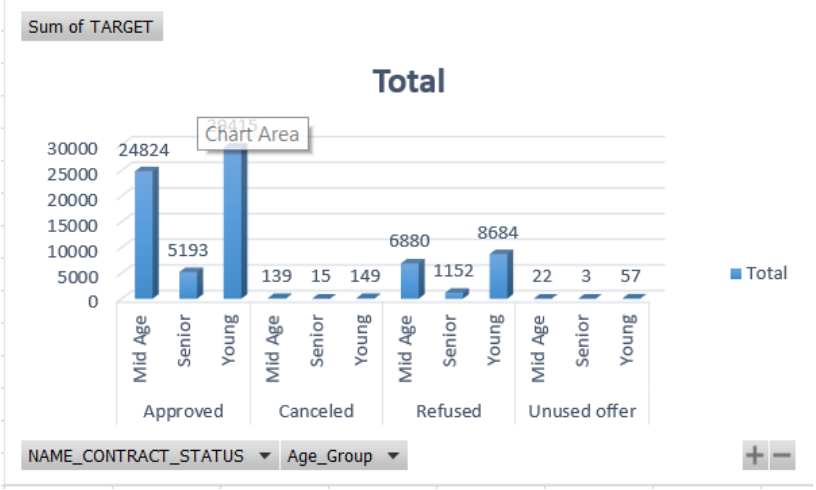


Analysis

1. We can see that the Defaulters are more for previously Unused offers loan status clients, who were New.
2. For previously Approved status the New clients were more defaulted followed by Repeater.
3. For previously Refused applicants the Defaulters are more Refreshed clients.
4. For previously Canceled applicants the Defaulters are more New clients.

Current loan defaulter status with respect to previous loan application status and age group

Row Labels	Sum of TARGET
Approved	**59432**
Mid Age	24824
Senior	5193
Young	29415
Canceled	**303**
Mid Age	139
Senior	15
Young	149
Refused	**16716**
Mid Age	6880
Senior	1152
Young	8684
Unused offer	**82**
Mid Age	22
Senior	3
Young	57
Grand Total	**76533**

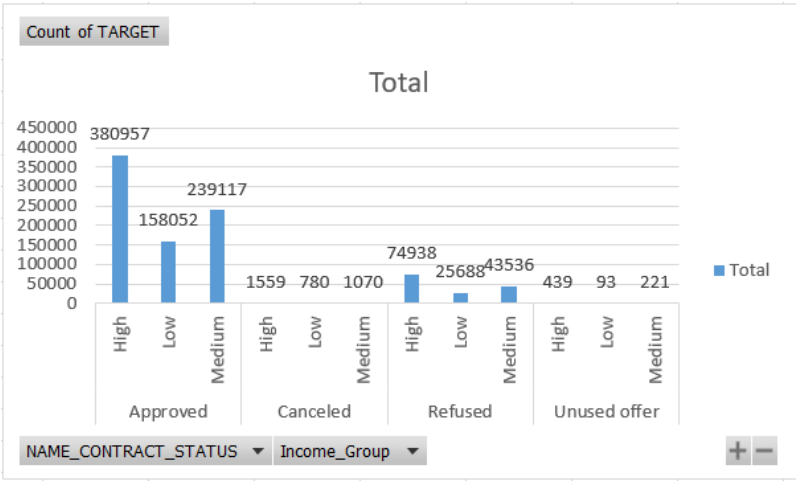


Analysis

1. For all the previous status Young applicants are more defaulted.
2. For all the previous status Senior applicants are less defaulted compared to others.

Current loan defaulter status with respect to previous loan application status and income group

Row Labels	Count of TARGET
Approved	**778126**
High	380957
Low	158052
Medium	239117
Canceled	**3409**
High	1559
Low	780
Medium	1070
Refused	**144162**
High	74938
Low	25688
Medium	43536
Unused offer	**753**
High	439
Low	93
Medium	221
Grand Total	**926450**

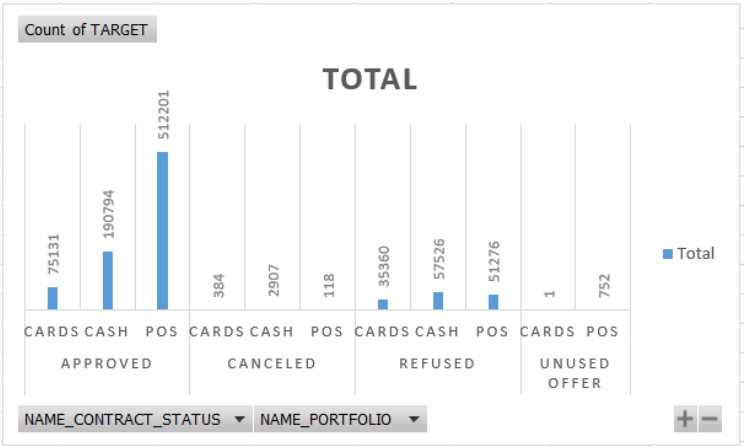


Analysis

1. For previously Unused offer the Medium income group was more defaulted and Low income group is the least.
2. For other application status more or less all the income groups are equally defaulted.

Current loan defaulter status with respect to previous loan application status and portfolio of the loan

Row Labels	Count of TARGET
Approved	**778126**
Cards	75131
Cash	190794
POS	512201
Canceled	**3409**
Cards	384
Cash	2907
POS	118
Refused	**144162**
Cards	35360
Cash	57526
POS	51276
Unused offer	**753**
Cards	1
POS	752
Grand Total	**926450**

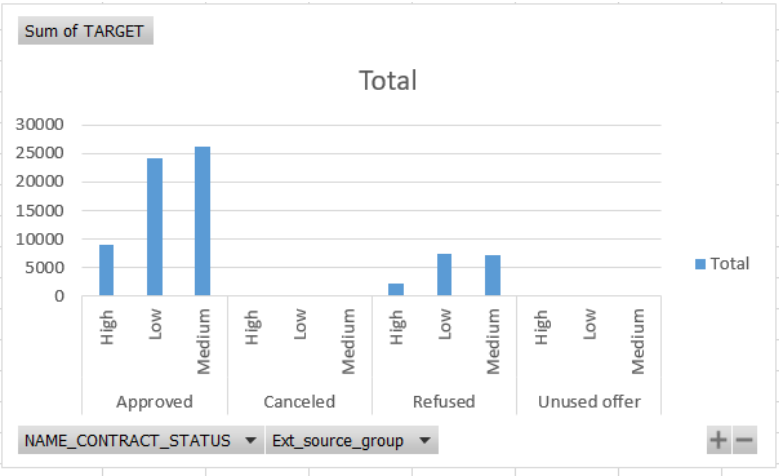


Analysis

1. Most of the clients were defaulted, who previously applied loan for Cards.
2. For approved loan status the clients applied for Cars are less defaulted.
3. For Refused loan status the clients applied for POS are less defaulted.

Current loan defaulter status with respect to previous loan application status and external source score category

Row Labels	Sum of TARGET
Approved	**59432**
High	8929
Low	24225
Medium	26278
Canceled	**303**
High	43
Low	131
Medium	129
Refused	**16716**
High	2213
Low	7328
Medium	7175
Unused offer	**82**
High	5
Low	47
Medium	30
Grand Total	**76533**



1. Applicants with low external source score are highly defaulted.
2. Higher scorer applicants are very unlikely to default irrespective of their previous loan status.

Result: Working with such industry kind data set is very good experience. I have learned to do Exploratory Data Analysis on the real time Data set. Handling such a huge dataset was a task but I gradually learned to handle it better and use the necessary tools.

Drive Link:

Jupyter Notebook file
https://drive.google.com/file/d/1nozOKMVvqjVV3zOCHxGLcX1mEcoCGIcU/view?usp=sharing

Univariate Analysis on Application Dataset

https://docs.google.com/spreadsheets/d/1uv9Tf9Y758DBw6OMpA5L4vcXGLXIURZB/edit?usp=sharing&ouid=102292071000492049204&rtpof=true&sd=true

Univariate and Bivariate Analysis on merged dataset

https://docs.google.com/spreadsheets/d/1OSmqkiOMppKq7oVSK41hc5iw1CP6ON2O/edit?usp=sharing&ouid=102292071000492049204&rtpof=true&sd=true

Whole File

https://docs.google.com/spreadsheets/d/1OSmqkiOMppKq7oVSK41hc5iw1CP6ON2O/edit?usp=sharing&ouid=102292071000492049204&rtpof=true&sd=true