CREDIT EXPLORATORY DATA ANALYSIS

CASE STUDY

BY:
Nitanshu Joshi
Anshika Dua

Problem Statement

This case study aims to identify patterns which indicate if a client has difficulty paying their installments 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.

AVAILABLE DATASETS:

- **application_data:** The data is about whether a client has payment difficulties.
- **previous_application:** It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

The Approach

STEPS UNDERTAKEN FOR EDA:

- Data Understanding
- Data Cleaning & Manipulation
- Data Analysis- Univariate,Bivariate & Multivariate
- Combining the two available datasets for further analysis
- Drawing Inferences and conclusions.

Handling Missing Values.

The missing values were majorly of two types:

- Visible
- Invisible

Visible missing values were identified as the columns having any more that 13% missing values.
Invisible missing values were hidden under the names XNA, XPA etc.
These values were either dropped, imputed with mean, median or mode; whichever was appropriate or renamed for better understanding.

Some columns which were unwanted for the analysis were also dropped

The final data frame had these columns with 307511 non-nulls.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Data columns (total 32 columns):
 #   Column                       Non-Null Count   Dtype
---  ------                       --------------   -----
 0   SK_ID_CURR                   307511 non-null  int64
 1   TARGET                       307511 non-null  int64
 2   NAME_CONTRACT_TYPE           307511 non-null  object
 3   CODE_GENDER                  307511 non-null  object
 4   FLAG_OWN_CAR                 307511 non-null  object
 5   FLAG_OWN_REALTY              307511 non-null  object
 6   CNT_CHILDREN                 307511 non-null  int64
 7   AMT_INCOME_TOTAL             307511 non-null  float64
 8   AMT_CREDIT                   307511 non-null  float64
 9   AMT_ANNUITY                  307511 non-null  float64
 10  AMT_GOODS_PRICE              307511 non-null  float64
 11  NAME_TYPE_SUITE              307511 non-null  object
 12  NAME_INCOME_TYPE             307511 non-null  object
 13  NAME_EDUCATION_TYPE          307511 non-null  object
 14  NAME_FAMILY_STATUS           307511 non-null  object
 15  NAME_HOUSING_TYPE            307511 non-null  object
 16  REGION_POPULATION_RELATIVE   307511 non-null  float64
 17  DAYS_BIRTH                   307511 non-null  int64
 18  DAYS_EMPLOYED                307511 non-null  int64
 19  DAYS_REGISTRATION            307511 non-null  float64
 20  DAYS_ID_PUBLISH              307511 non-null  int64
 21  CNT_FAM_MEMBERS              307511 non-null  float64
 22  REGION_RATING_CLIENT         307511 non-null  int64
 23  REGION_RATING_CLIENT_W_CITY  307511 non-null  int64
 24  WEEKDAY_APPR_PROCESS_START   307511 non-null  object
 25  REG_REGION_NOT_LIVE_REGION   307511 non-null  int64
 26  REG_REGION_NOT_WORK_REGION   307511 non-null  int64
 27  LIVE_REGION_NOT_WORK_REGION  307511 non-null  int64
 28  REG_CITY_NOT_LIVE_CITY       307511 non-null  int64
 29  REG_CITY_NOT_WORK_CITY       307511 non-null  int64
 30  LIVE_CITY_NOT_WORK_CITY      307511 non-null  int64
 31  ORGANIZATION_TYPE            307511 non-null  object
dtypes: float64(7), int64(14), object(11)
memory usage: 75.1+ MB
```

Manipulating Data

- Certain columns having negative values were identified and we took their absolute value to use them better:
- DAYS_BIRTH
- DAYS_REGISTRATION
- DAYS_ID_PUBLISH

2. We used the column DAYS_BIRTH to get the approximate age of the applicants and divided them in the range:

20-25	45-50
25-30	50-55
30-35	55-60
35-40	60-65
40-45	65-70

35-40	43680
40-45	39997
30-35	39437
25-30	36488
50-55	35097
45-50	34404
55-60	32722
60-65	24359
20-25	16317
65-70	5009

DAYS_BIRTH	DAYS_REGISTRATION	DAYS_ID_PUBLISH
9461	3648.0	2120
16765	1186.0	291
19046	4260.0	2531
19005	9833.0	2437
19932	4311.0	3458

3. Bins were created for income and credit amount to use the columns better.

INCOME

```
100K-150K          91591
150K-200K          64307
200K-250K          48137
75K-100K           39806
50K-75K            19375
250K-300K          17039
300K-350K           8874
350K-400K           5802
400K-450K           4924
25K-50K             4517
500K and above      2702
450K-500K            437
0-25K                  0
Name: AMT_INCOME_RANGE, dtype: int64
```

CREDIT

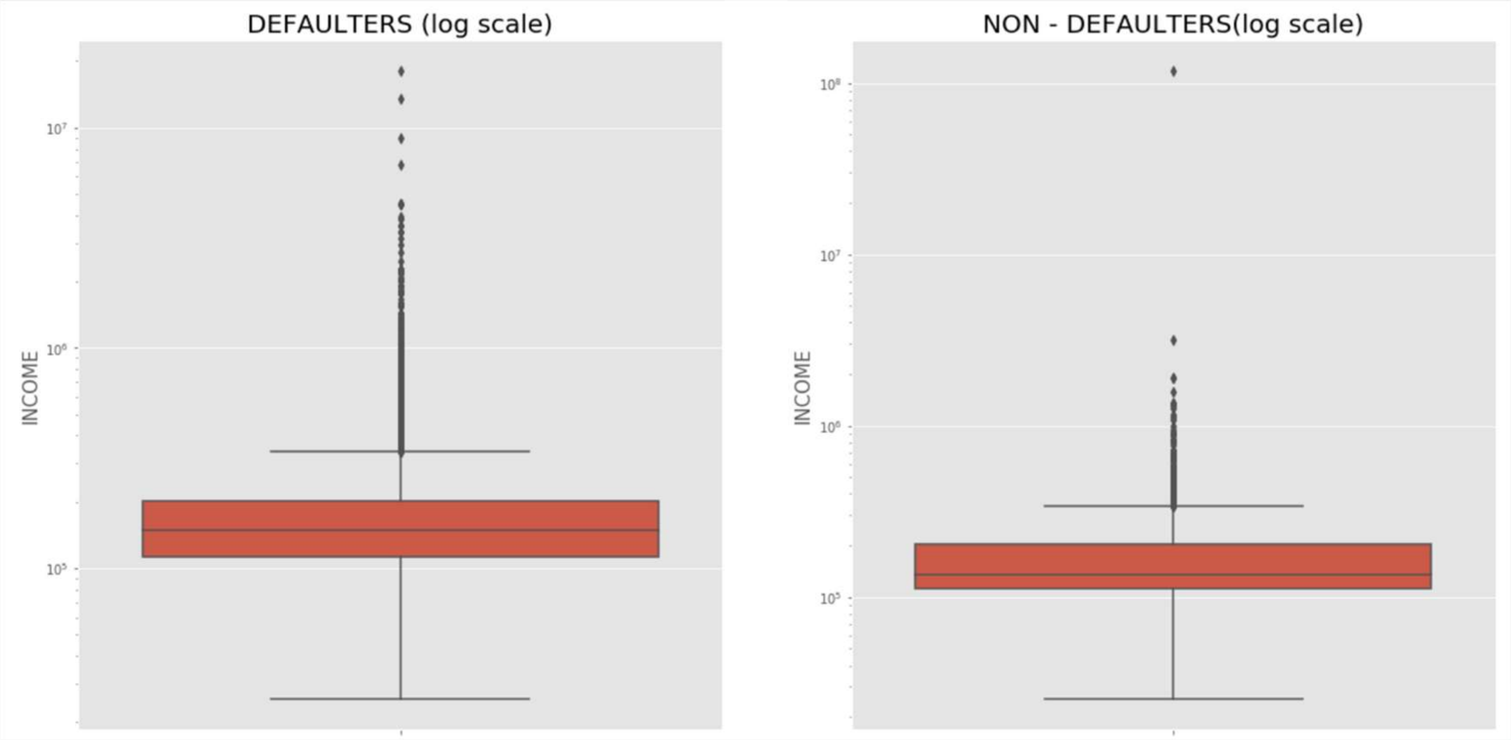
```
900K & above    58912
200K-300K       54813
500K-600K       34232
400K-500K       32038
100K-200K       30140
300K-400K       26338
600K-700K       24049
800K-900K       21792
700K-800K       19193
40K-100K         6004
Name: AMT_CREDIT_RANGE, dtype: int64
```

Befor proceeding to analysis, we divide the dataset on the basis of TARGET.

target_1
Client with payment difficulties.

target_0
All other cases

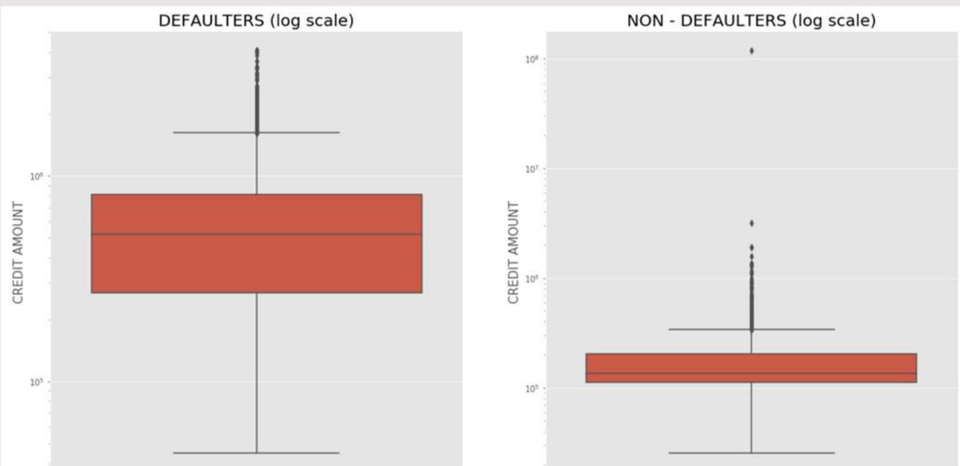
OUTLIER ANALYSIS FOR INCOME.



- We observe that the inter-quartile range of clients WHO HAVE DEFAULTED is slightly higher than clients WHO HAVE NOT DEFAULTED.
- We also observe that number of outliers for DEFAULTER clients is more and high as compared to clients who are NOT Defaulters.

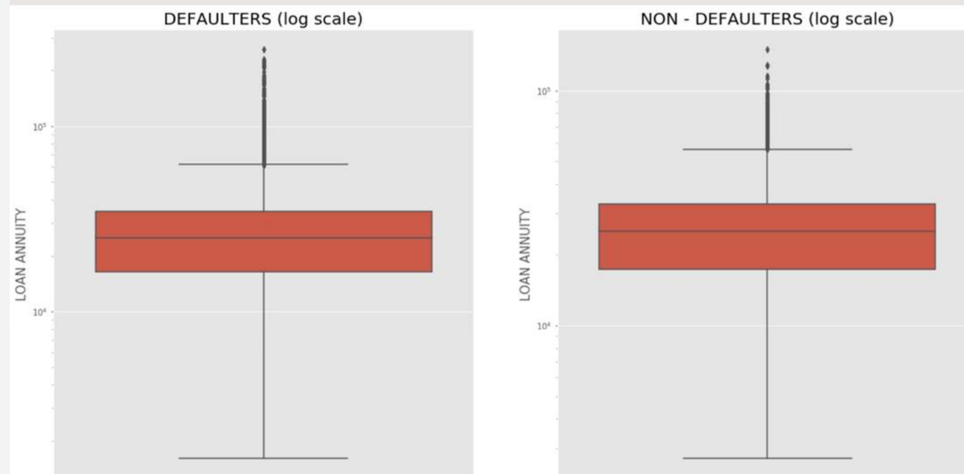
OUTLIER ANALYSIS FOR CREDIT AND ANNUITY.

CREDIT



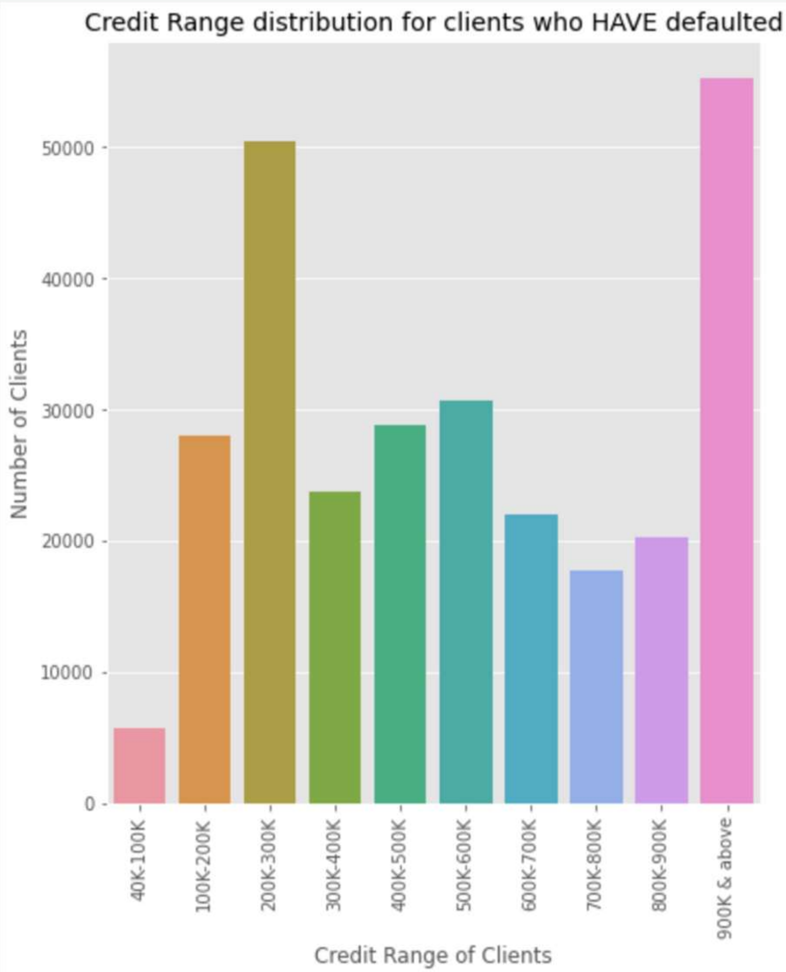
- We see that the interquartile range for the credit amount data is higher for clients WHO ARE defaulters.
- We see that outliers are present in both cases, but for Targer = 0, outliers are more on the higher side.
- We can infer that clients who are seeking a higher credit loan amount tend to default more as compared to clients who take a lower credit loan amount.

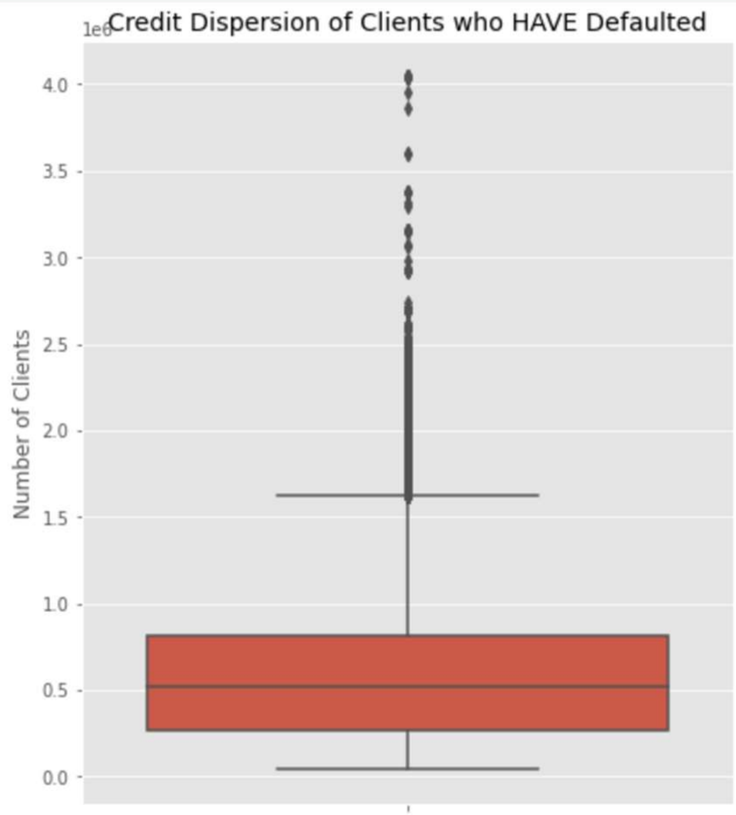
ANNUITY



We see that for Loan Annuity, both types (i.e. Defaulters and Non-Defaulters) follow a similar trend of values, with Non-Defaulters having a slighlty higher Inter-Quartile range.

Univariate Analysis Of Credit Amounts Taken By The Clients Who Have Defaulted





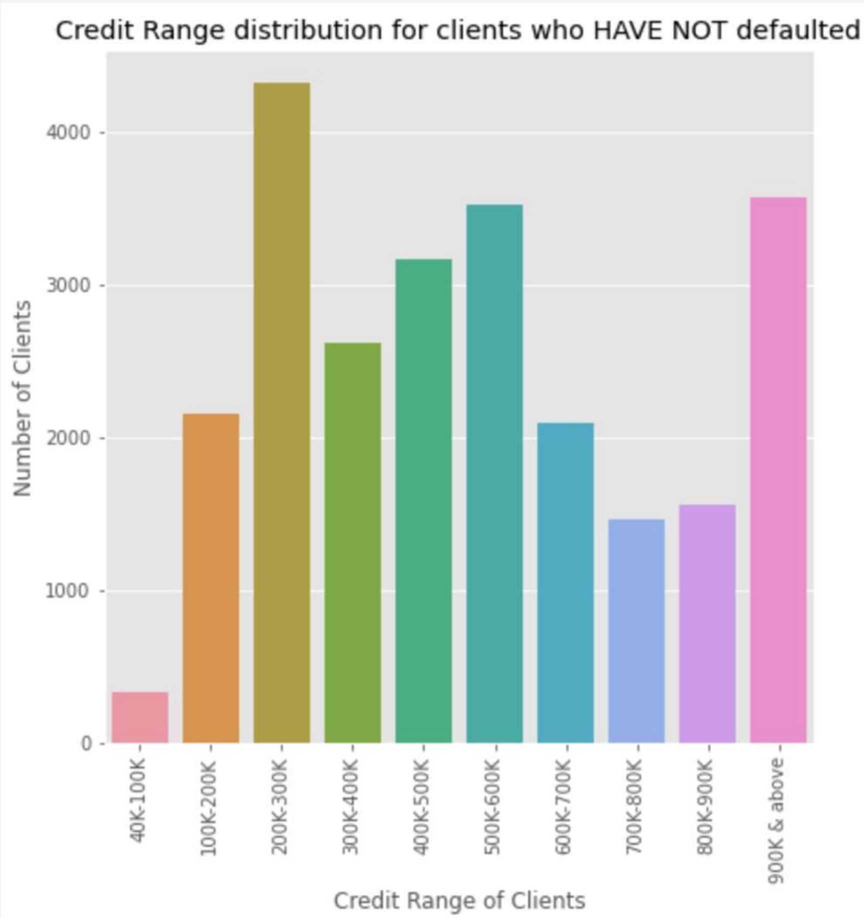
From Bar plot:

We see that people who Deafult the most have either taken a very high loan, i.e. in the range of 900 thousand and above, OR have taken a loan in the range of 200 thousand to 300 thousand.

From Box plot:

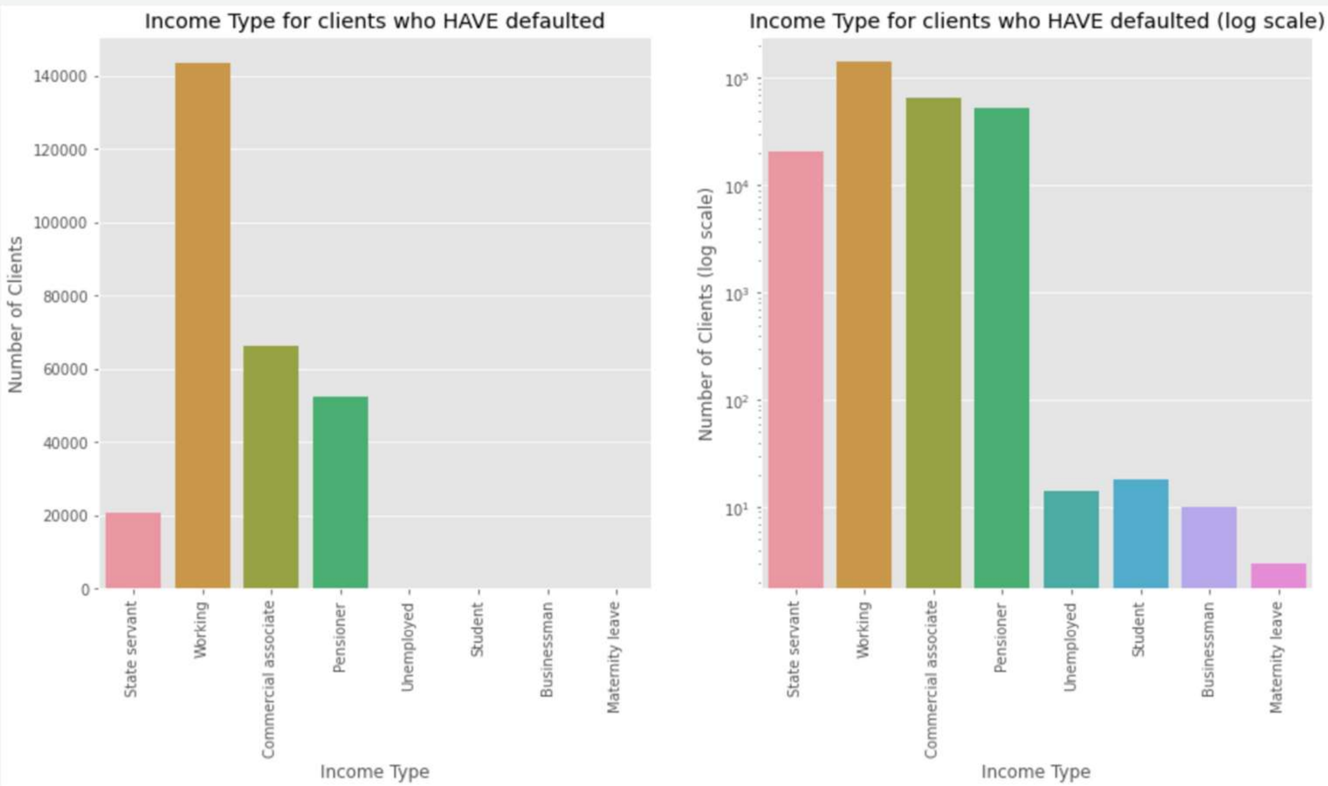
- From the box plot we can observe that 75% of clients who have defauler have taken a loan amount in the range of 45 thousand dollars and 81 thousand dollars.
- Apart from that quite a few outliers are present in the distribution.From this it can be inferred that, there might have been some clients, who tool a huge loan to start a business, but have defaulted due to their business faulure or similar problems.

Univariate analysis of Credit Amounts taken by the clients who HAVE NOT defaulted.



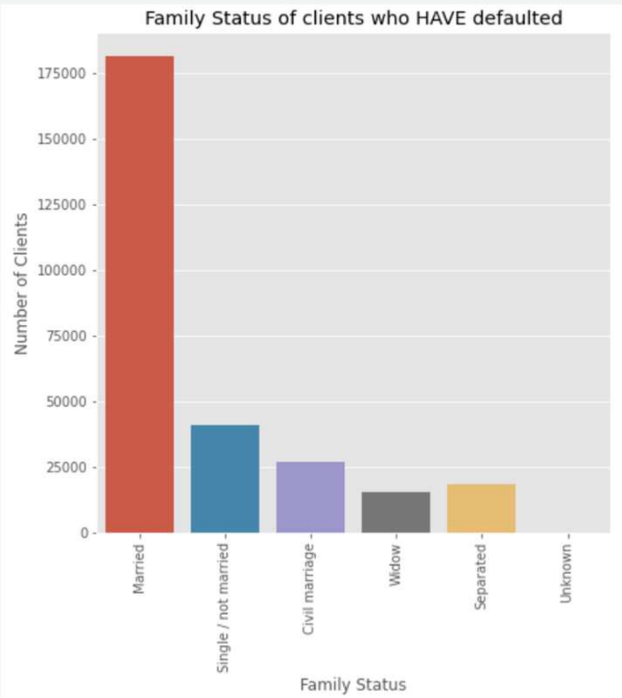
The number of clients who HAVE NOT defaulted follow a simliar pattern for their credit amount range as compared to the clients who HAVE defaulted.

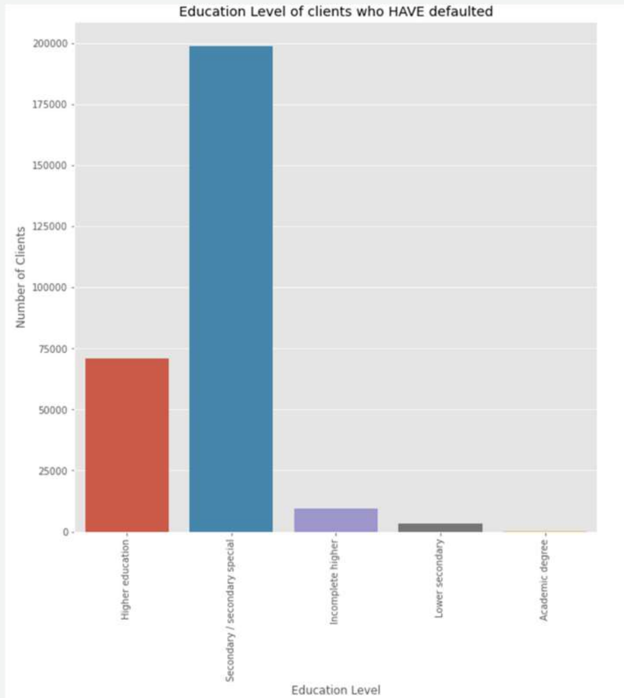
Univariate Analysis Income Type Of The Clients Those Who Have Defaulted

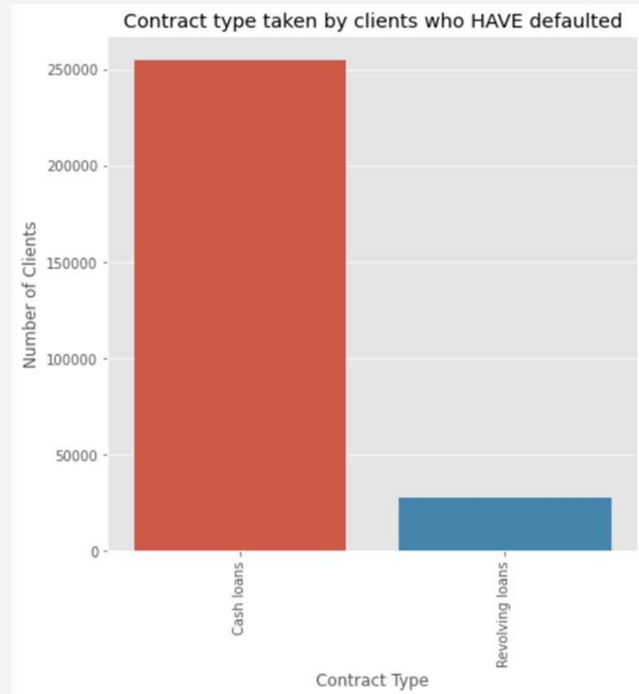


- We observe that the working class clients have the highest frequency of defaulters.
- We also observe that clients who are unemployed, students, businessman and clients who are on maternity leave have a low chance of defaulting.
- We can infer that these clients have less probability of taking loans.
- We can also infer State Servants are comparatively less likely to Default, since they have a safe and regular salary.

Univariate Analysis For Family Status, Education Level And Contract Type For Clients





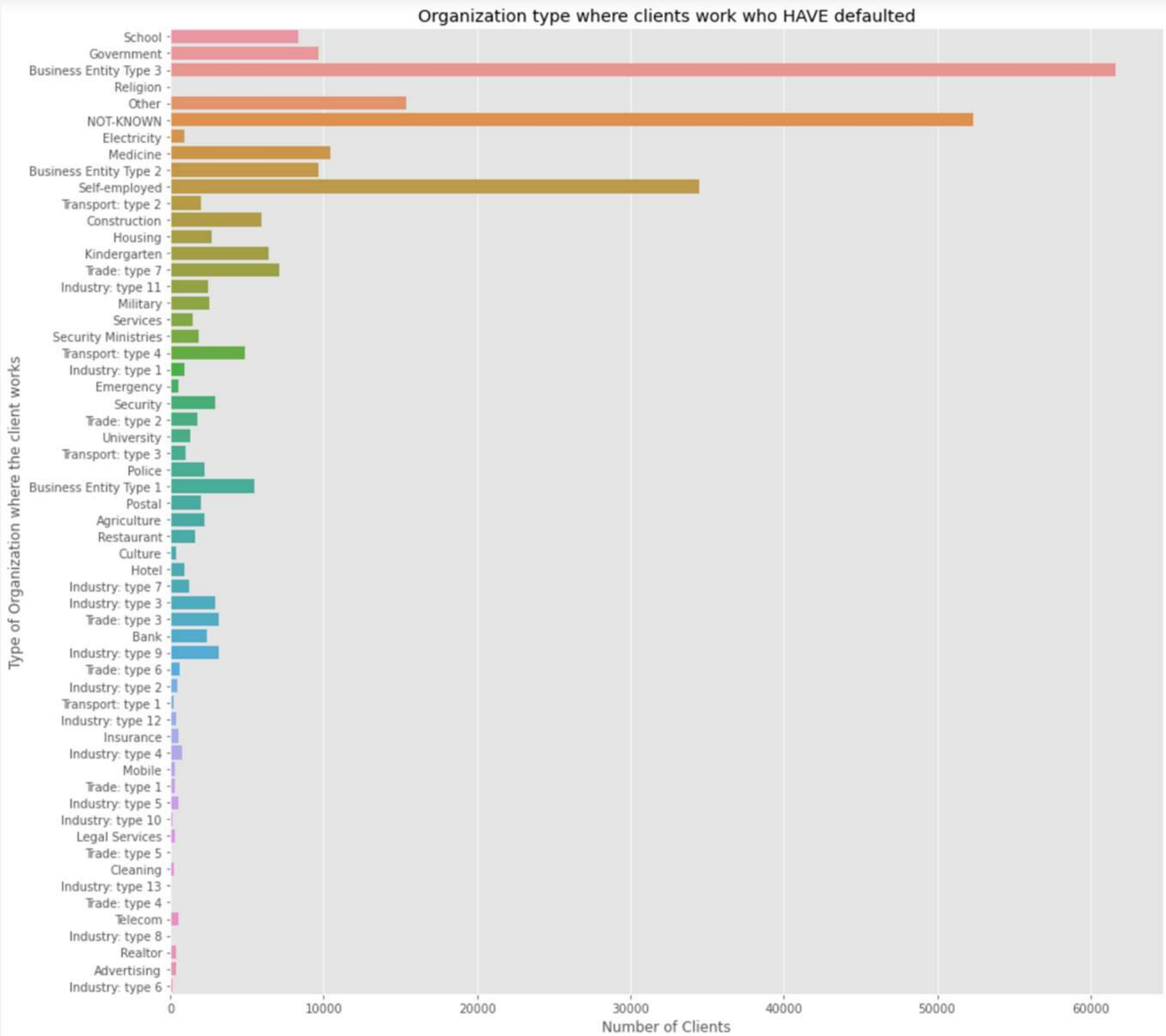


- We observe that married people are more likely to default.
- We also observe that people who are widow or widower are less likely to default.

We can clearly see that number of clients with only a secondary education have a very high probability to default as compared to a client who has completed his/her higher education. From the above observation, we can infer the clients with higher with higher level of education might be earning weill and may be less susceptible to be a defaulter.

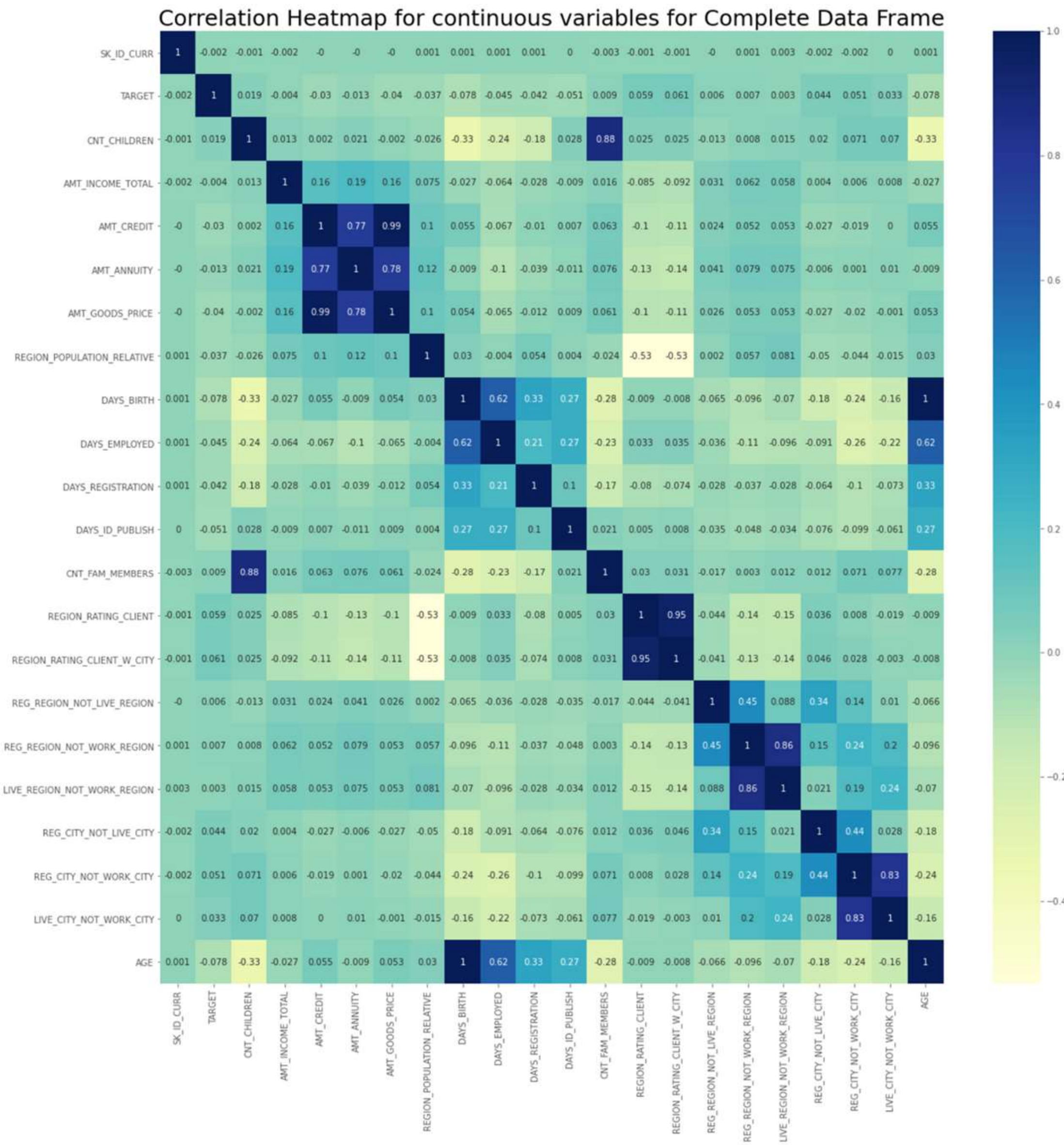
We observe that people who took cash loans are more likely to default as compared to revolving loans.

Univariate analysis of organization type of clients who HAVE defaulted



- We observe that clients who work in a business entity are more likely to take loans and default.
- Another category of clients which are highly susceptible to Default are self-employed people.

Bivariate/Multivariate Analysis



We make a correlation matrix of the numeric/continuous columns in the dataset, make its heatmap and derive the top 10 correlations.

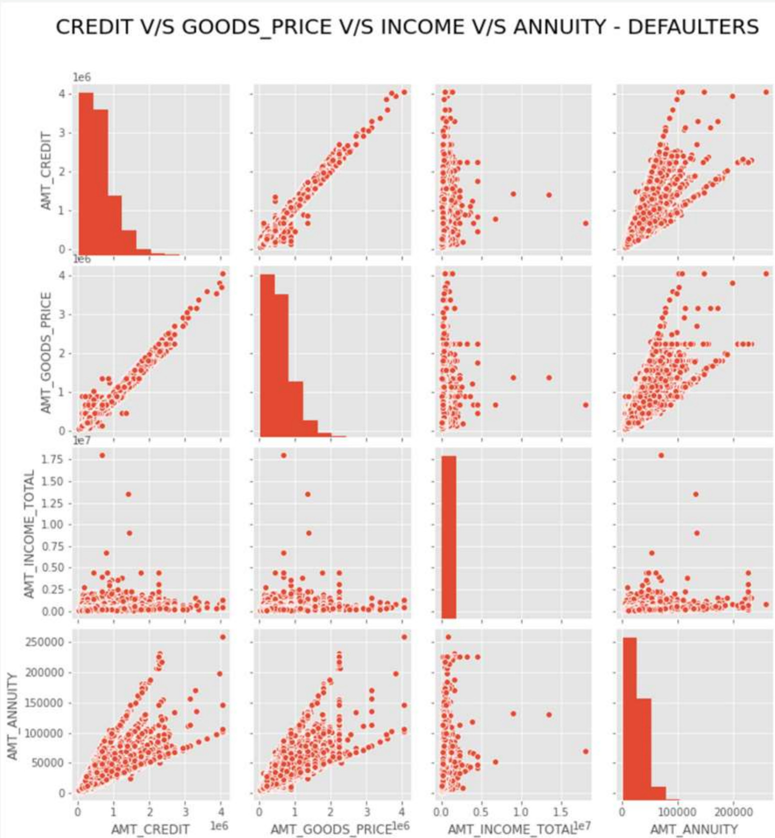
Top 10 Correlations

Variable 1	Variable 2	Correlation
AMT_CREDIT	AMT_GOODS_PRICE	0.999885
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.998678
CNT_CHILDREN	CNT_FAM_MEMBERS	0.989672
AMT_ANNUITY	AMT_GOODS_PRICE	0.963801
AMT_CREDIT	AMT_ANNUITY	0.962958
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.948332
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.930532
DAYS_BIRTH	DAYS_EMPLOYED	0.903728
DAYS_EMPLOYED	AGE	0.903727
REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT_W_CITY	0.866640

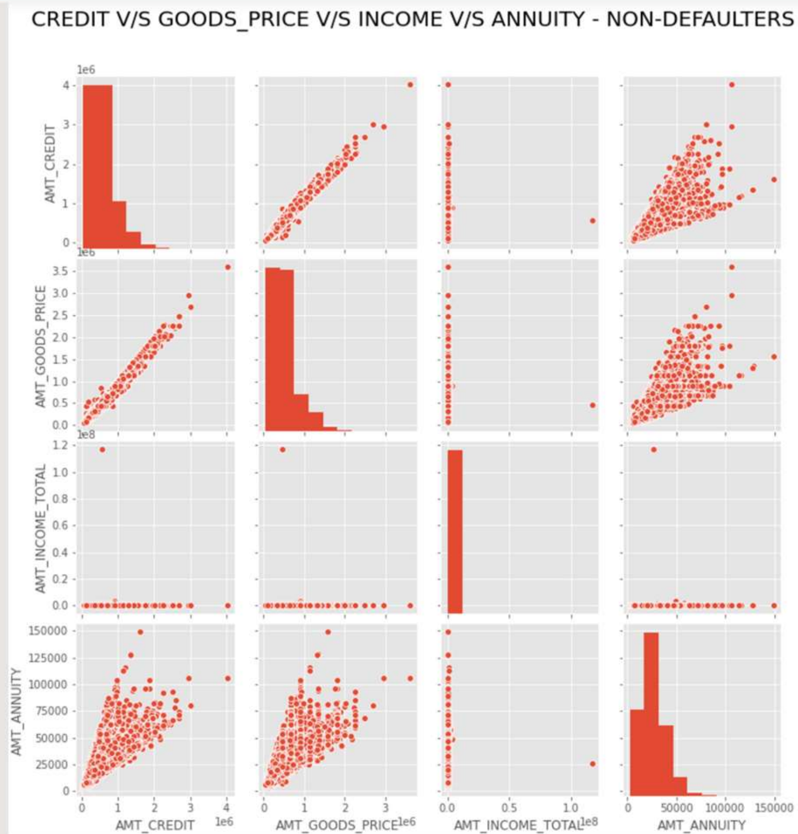
Observations and Inferences from Heatmap

- We find that the highest correlation exists between Credit Amount and the Goods Price. We can infer that the loan amount must be taken to buy the goods.
- We also find a very high correlation between Credt Amount and Loan Annuity Amount.
- We can also observe a high correlation between days after birth and the days after starting work.
- There is also a high correlation between LIVE_REGION_NOT_WORK_REGION and REG_REGION_NOT_WORK_REGION. It can be inferred that there is a high probability where a person who did not mention his/her permanent address as work address will not have mentioned his/her contact address as work address.
- There is also a high correlation between LIVE_CITY_NOT_WORK_CITY and REG_CITY_NOT_WORK_CITY. It can be inferred that there is a high probability where a person who did not mention his/her permanent address as work address will not have mentioned his/her contact address as work address.
- Credit Amount and Age show negative correlation. It can be inferred that credit amount is higher for low age and vice-versa.
- There is a negative correlation between Region_population_Relative and CNT_CHILDREN. It can be inferred that clients with less children live in densely populated areas.
- We observe a positive correlation between Credit amount and Region_Population_Relative. We can infer that people living in densely populated region take higher amount loans.
- We also observe a positive correlation between clients income amount and Region_Population_Relative. It can be inferred that clients living in densely populated region have a higher income.

Pair plot for CREDIT V/S GOODS_PRICE V/S INCOME V/S ANNUITY

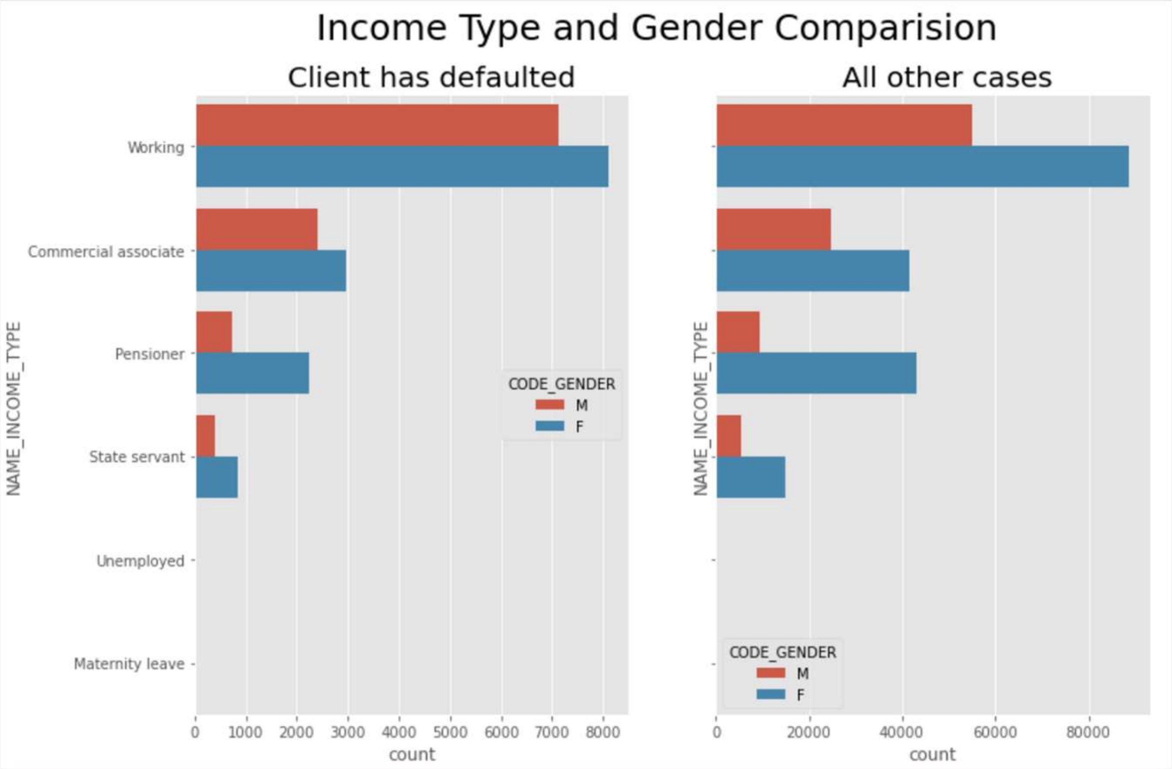


- The clients having low income are more likely to default.
- There seem to be no defaulters having a higher income.
- The more is the goods price, the higher credit is available in both the cases.
- In the case of default, a very high credit is not available even if the price of goods is high while in all other cases a higher credit may be available in case of a high goods price.

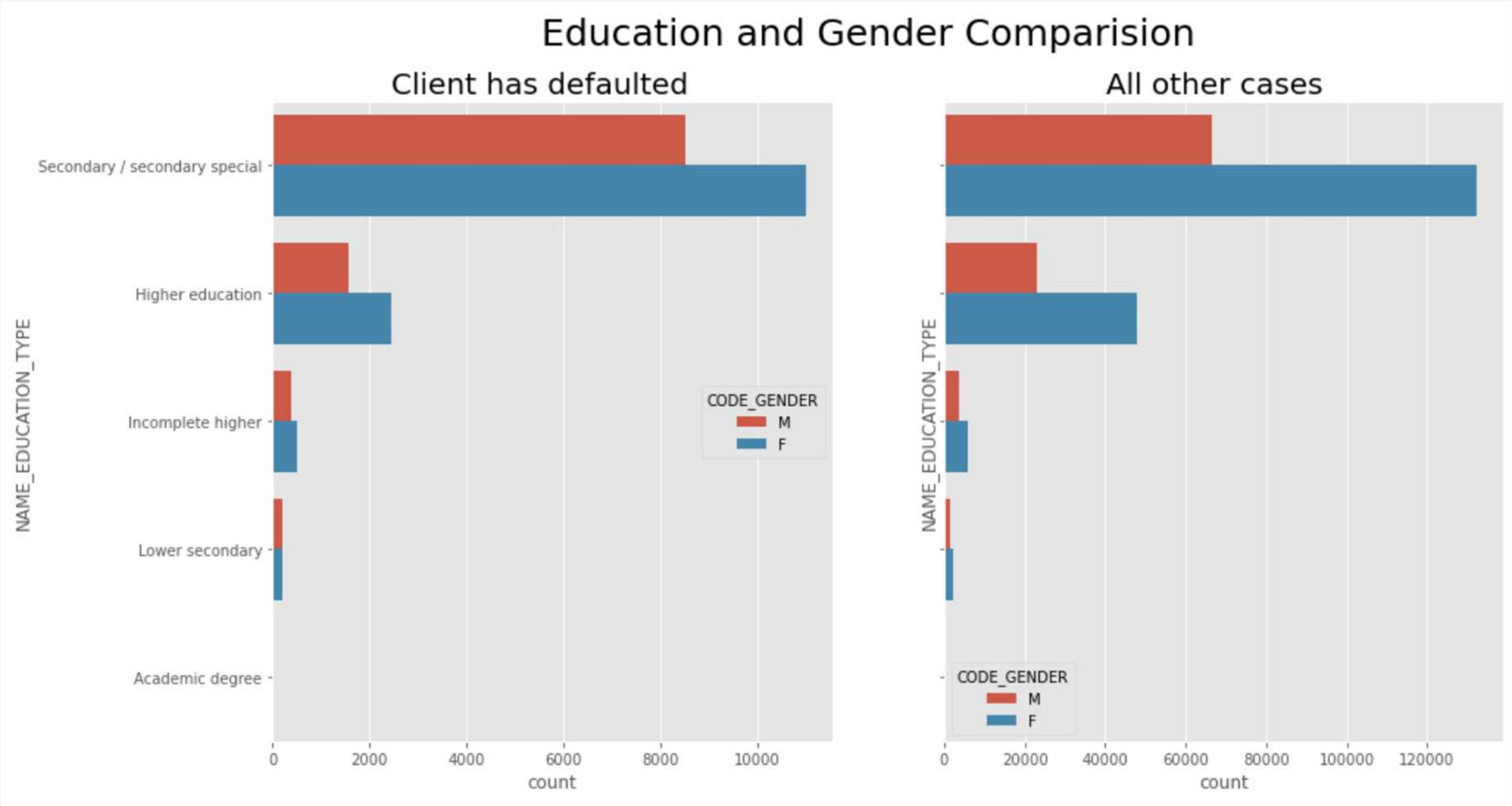


Bivariate Analysis for Income Type vs Gender

- There are no cases of default in the case of students and businessman.
- The proportion of females in case of all income types is higher as compared to males in case of all other cases than in case of default.

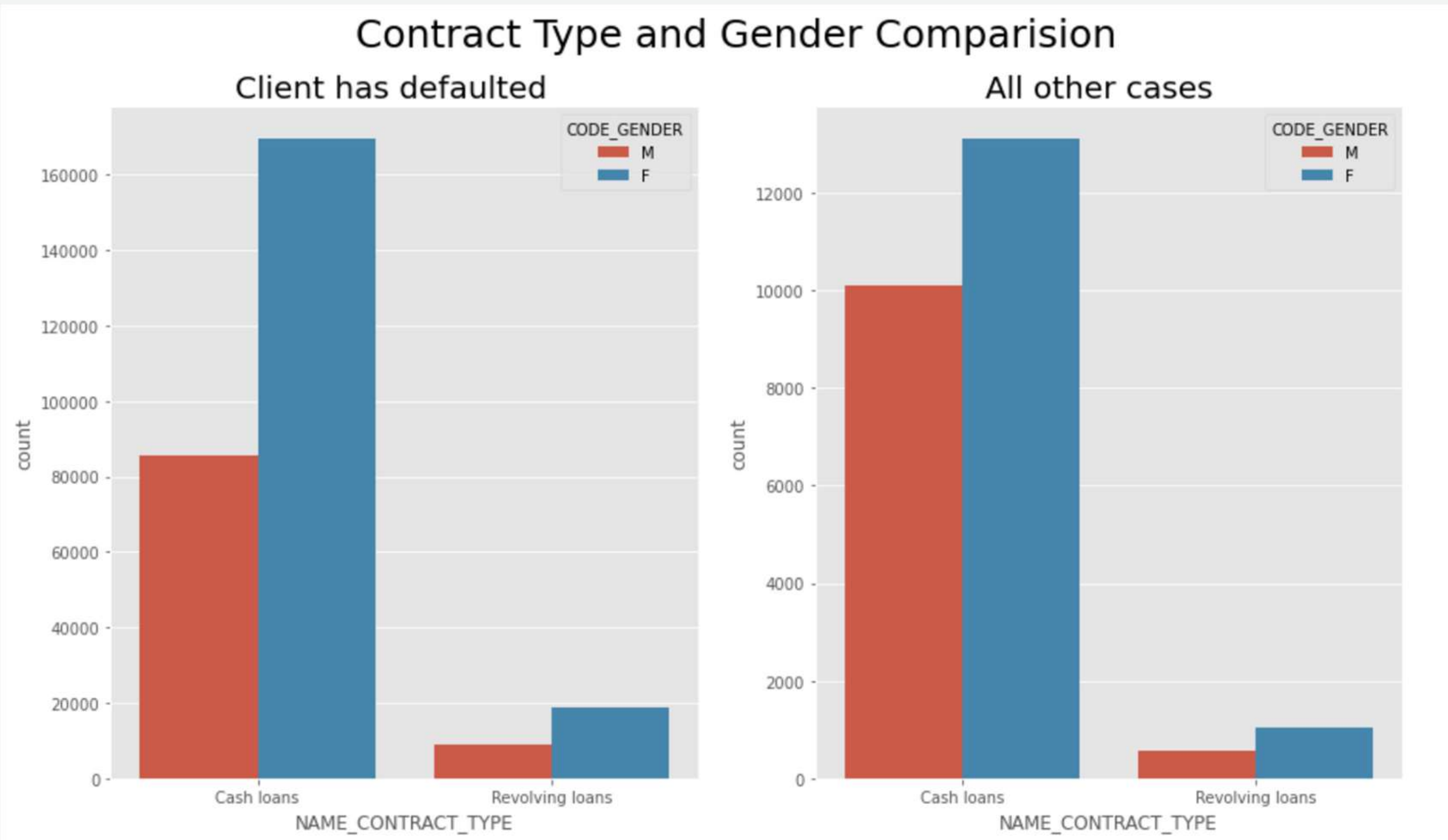


Bivariate Analysis Education Level vs Gender



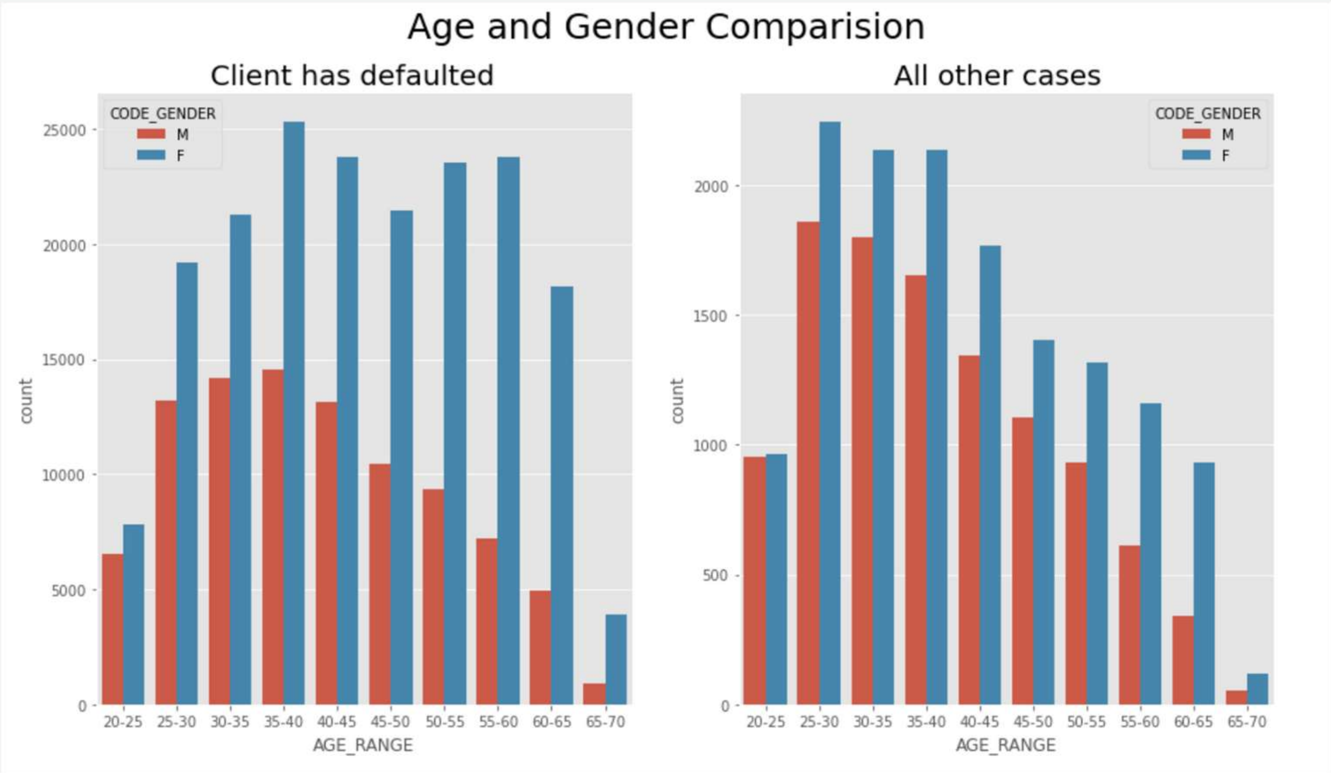
- A large number of applications are by people having secondary education followed by Higher Education both in the case of payment difficulties and all other cases.
- Female applicants are more than male applicants in all cases.

Bivariate Analysis for Contract Type vs Gender



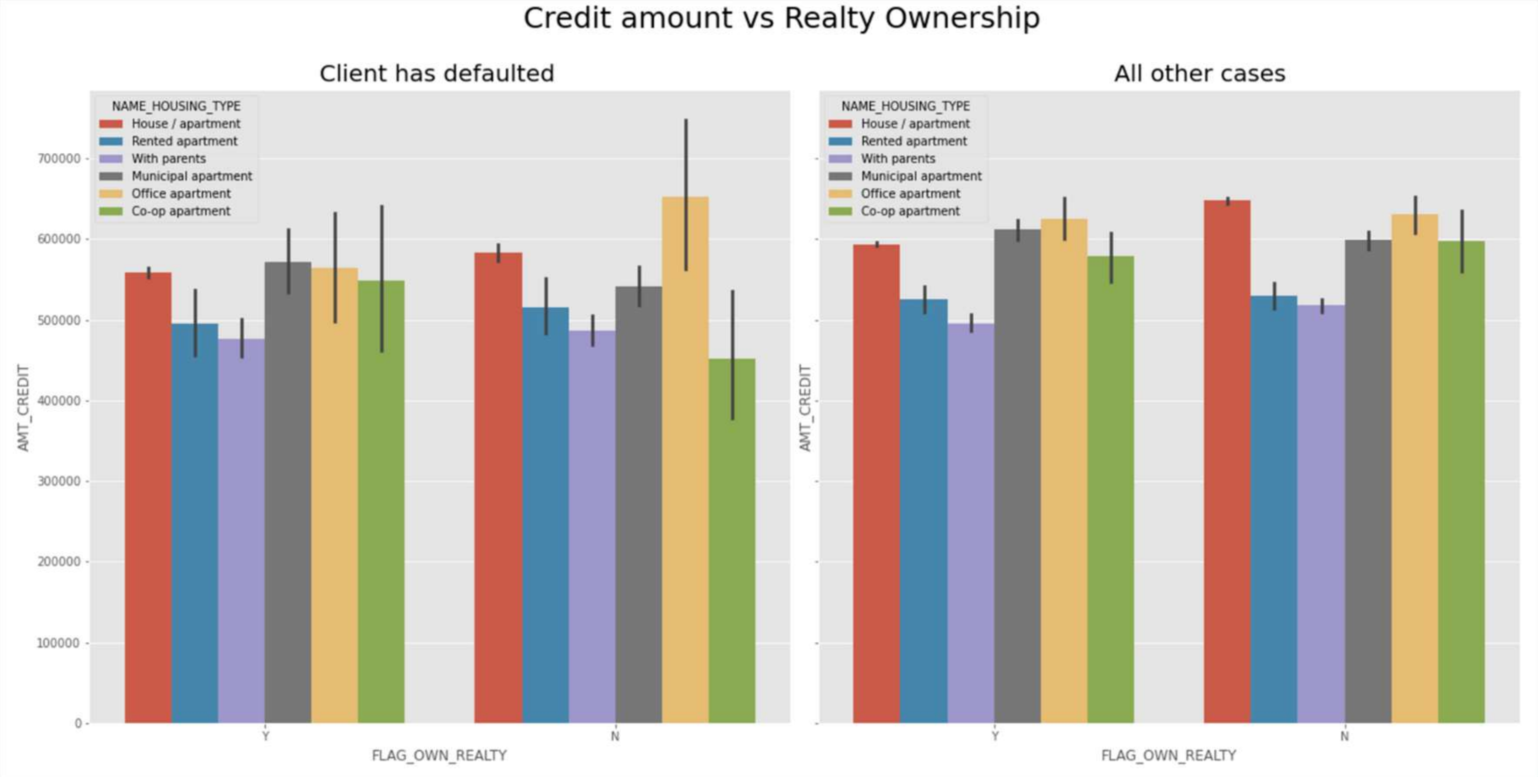
The proportion of no-default by females is higher than males in case of cash loans when compared with default.

Bivariate Analysis for Age vs Gender



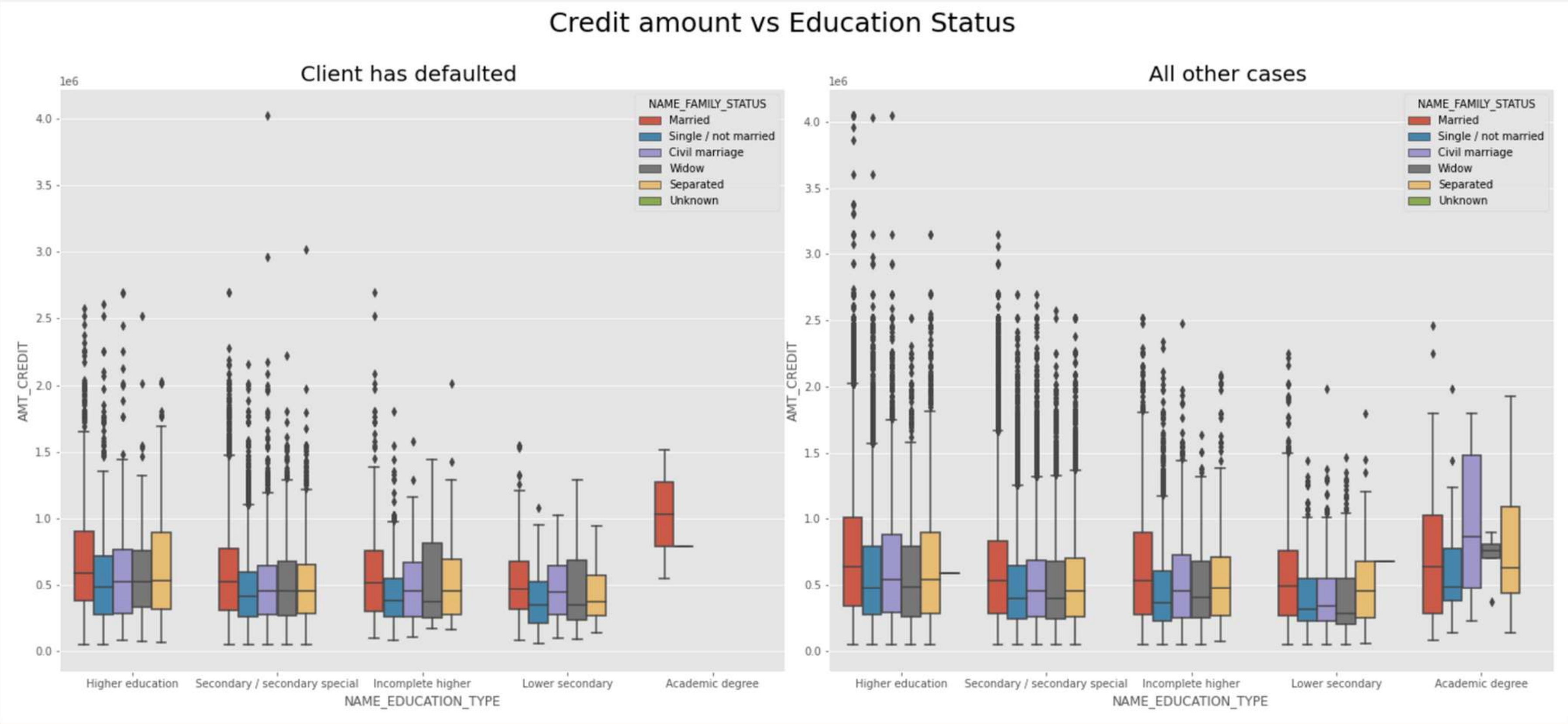
- With the increase in age, there is a decreasing trend of applications in all cases.
- The females are less likely to default from the ages 25 to 65 while the possibility of default is decreasing with increase in age.
- Overall, the age group of 25 to 45 is less likely to default.

Multivariate Analysis for Credit Amount to Reality owning status for various housing type categories for the clients



- The Office apartment which are not owned by the client have a higher default as compared to the office apartment owned by client.
- The default possibility is less in the case the Co-op Apartments not owned by client as compared to all other cases.

Multivariate Analysis for Credit Amount to Client's Education status for various categories of family status.



- The median credit amount is the highest in case of married person having an academic degree.
- The credit amount of people in civil marriage having an academic degree are mostly in the third quartile.
- The education type 'Academic Degree' has the least outliers, which means there is not much variability in the credit amount of this category.

We now move on to the previous_ application dataset.

Handling Missing Value

Columns having missing values greater than 13% were dropped.

Unwanted columns

Columns that were unnecessary for analysis were dropped.

Null Values-*Visible and Invisible*

Null values were either dropped or imputed with mean, median or mode, whichever was appropriate

Merging the two datasets

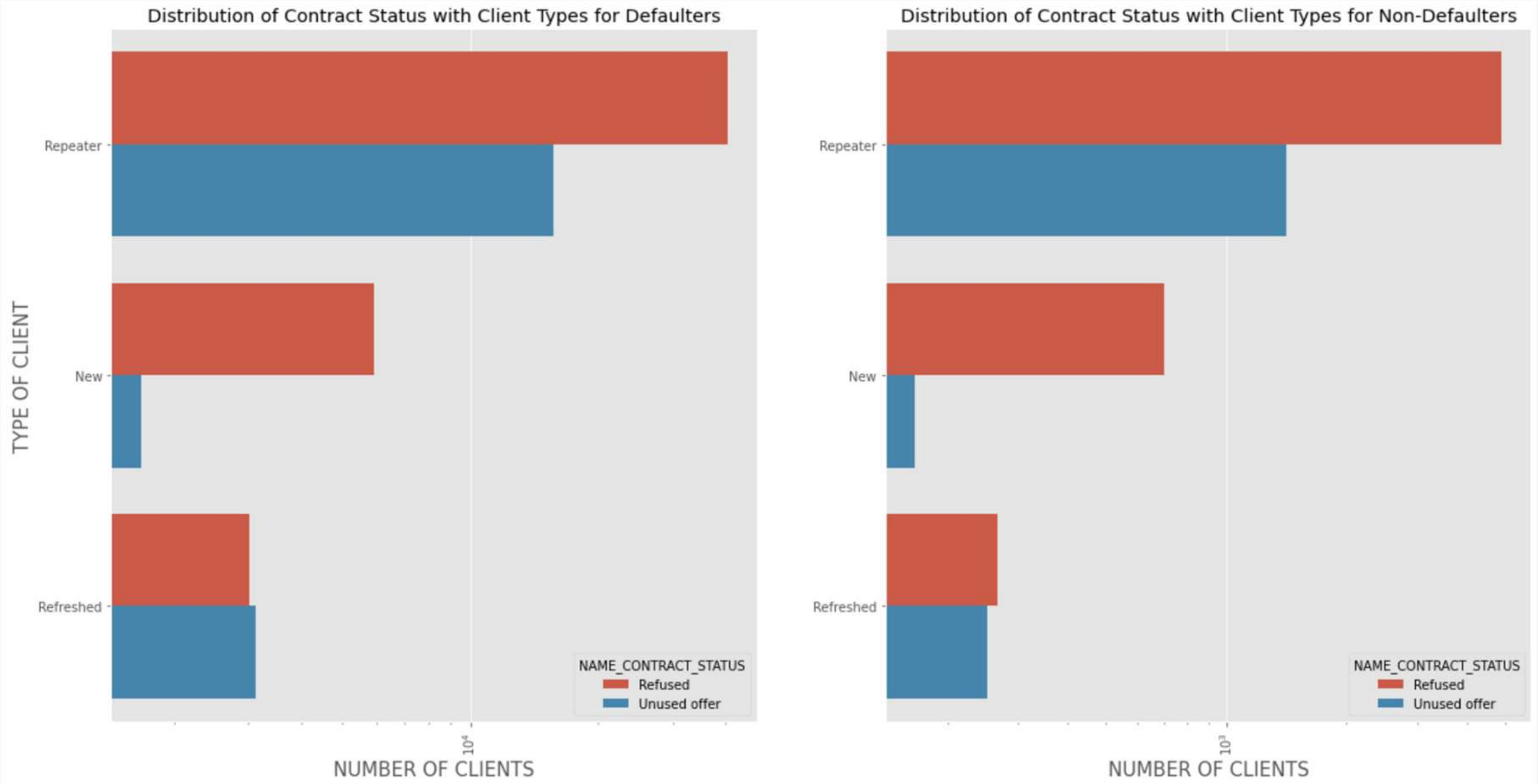
Finally the two datasets: application_data and previous_applictaion were merged on the basis of Current ID of the client.

Dividing the dataset

The final dataframe was split into two on the basis of TARGET: df2_merged_target_0 & df2_merged_target_1.

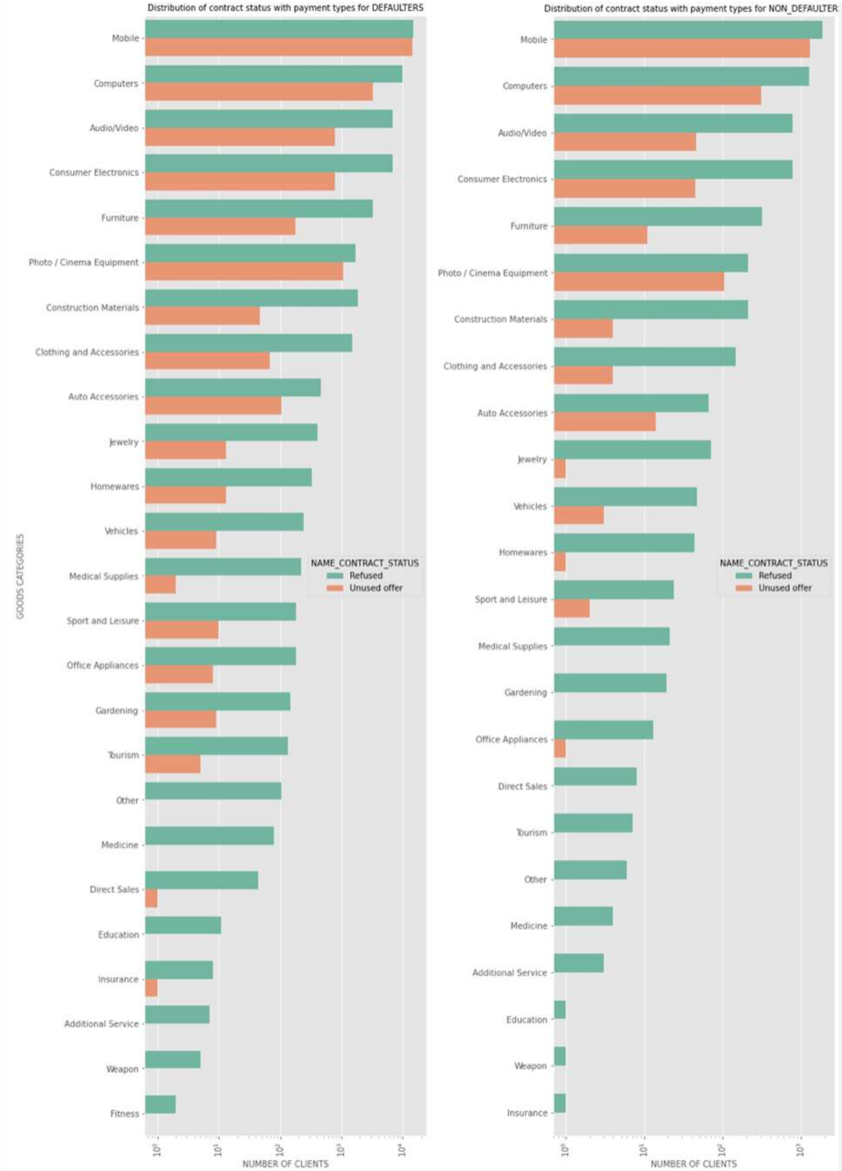
Distribution of Contract status with client types for Defaulters and Non-Defaulters

- We see that the clients who in the past are repeaters are more susceptible to Default on their loans. Majority of them have contract status as Refused.
- We also see that Number of Defaulter clients in the category - 'New' have contract type - 'Refused' are far more than clients in the same category and having contract type - 'Unused Offer'.
- We also see a near similar trend follows for non-defaulters as well.

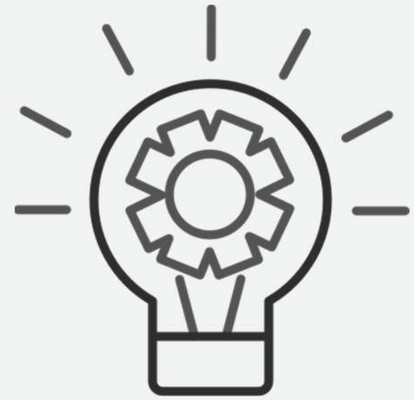


Distribution of Contract status with Goods Categories

- We can observe that most number of defaulters clients are taking loans for Goods categories - 'Mobile', followed by 'Computers' and 'Audio Video Equipment'.
- We see that most of the Defaulters as well as the Non-Defaulters have majority of their contract status as - 'Refused'.
- We also find that clients associated with 'Fitness' as their Goods Category are very less susceptible to Default on their loans.



CONCLUSION



There are some variables which are greatly impacting the chances of a client being a Loan Defaulter. They include:

- Client's Total Income
- Credit Amount: Total Credit Amount taken by the Client as Loan
- Income Type of Client
- Family Status of the Client
- Education Level of the Client
- Type of client: whether a client is a Repeater or a New client or a Refreshed client
- Goods_Categories: Goods for which the loan is taken.

Precautions that can be taken to avoid defaults

- Bank should check the profile of a client thoroughly, who have income in the brackett of 100 thousand dollars and 250 thousand dollars, before granting them loan.
- Bank should also check the profile of a client thoroughly, who are taking a loan either in the brackett of 200 to 300 thousand dollars or in the range 900 thousand dollars and above.
- Working class clients should be thoroughly processed, since they are the majority loan seekers and have a high probability to default.
- Well educated clients (Education level of Higher Education) should be granted more loans as compared to the less educated clients (Education level of Secondary/Secondary Special Education).
- Clients seeking loan for buying goods like Mobile Phones, Computers and Audio/Video equipments should be processed thoroughly before being granted the loan.
- Clients who are repeating loan seeker pose a high threat for Loan Default, so they should also be checked before being granted loan.