Bank Loan Case Study

Project Description

In this project, we will be doing data analysis of a bank loan dataset. The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. We will be doing EDA to find out the applicants who can repay the loans and who cannot and will help the bank in making the right decision of giving out loans.

Approach

First, we clean the data by using the appropriate methods to identify the missing values and remove/replace them. Then, look for any imbalance in the data and spot the outliers if any using various graphs and plots. Then, using the right methods we find out the top 10 correlations for the target variables.

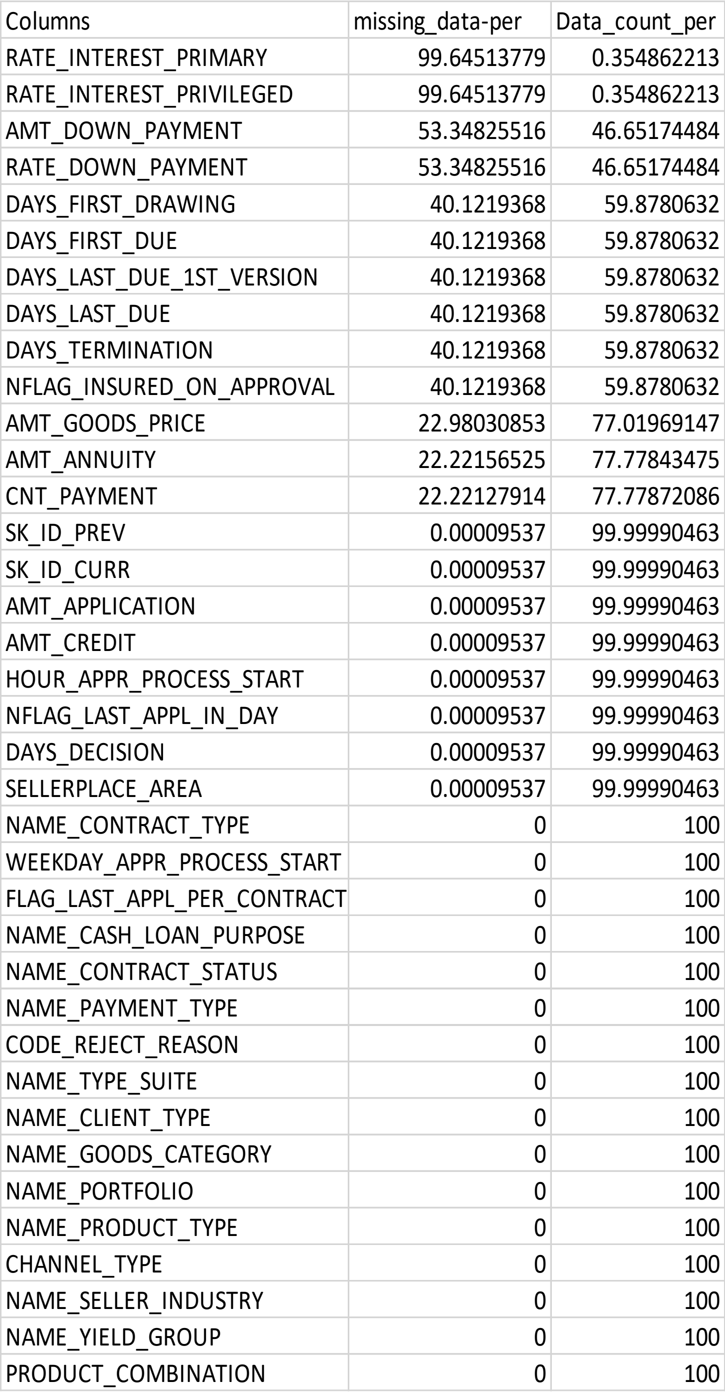
Tech-Stack Used

For this project, we used Microsoft Excel 2021, the flagship software used across the world for spreadsheets.

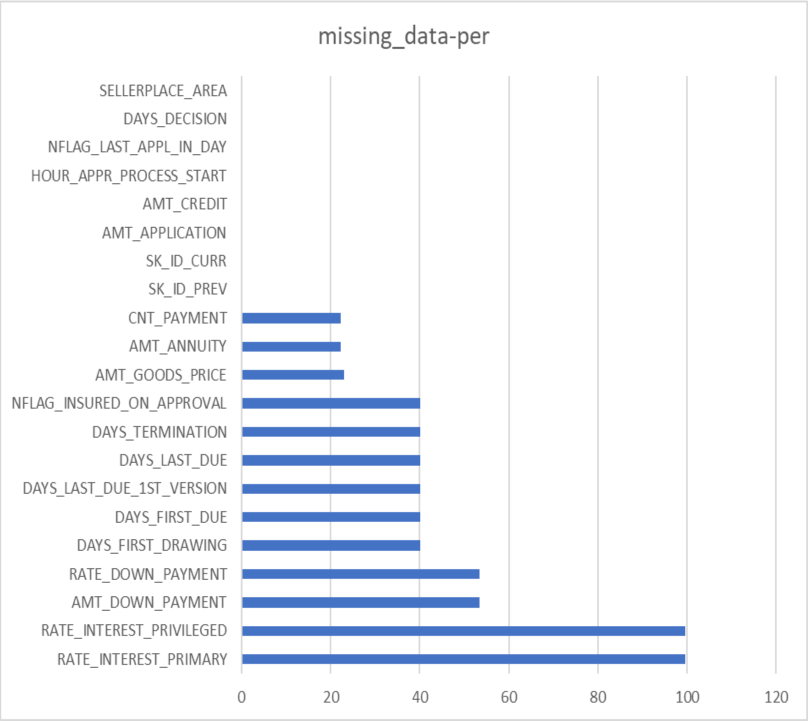
Insights

The analysis of the dataset gave us various insights which are as follows –

1. **Cleaning the data.**
   - For this, we first calculated the missing value percentages for each feature in the previous application data. As we can see in the table that there are 21 columns/features that have missing values in the dataset. Since the missing data percentage in four columns is higher than 50%, we permanently drop these columns.
   - As in columns SK_ID_PREV to SELLERPLACEAREA has only 0.000095% of data missing, we can delete/impute these rows as they won't affect the analysis.
   - In the application_data dataset, we can replace the missing values in the columns AMT_GOODS_PRICE and EXT_SOURCE_2 with the mean values of both the columns i.e., 538396 and 0.514392 respectively.
   - Similarly, in the NAME_TYPE_SUITE column we can replace the missing values with "Unaccompanied".

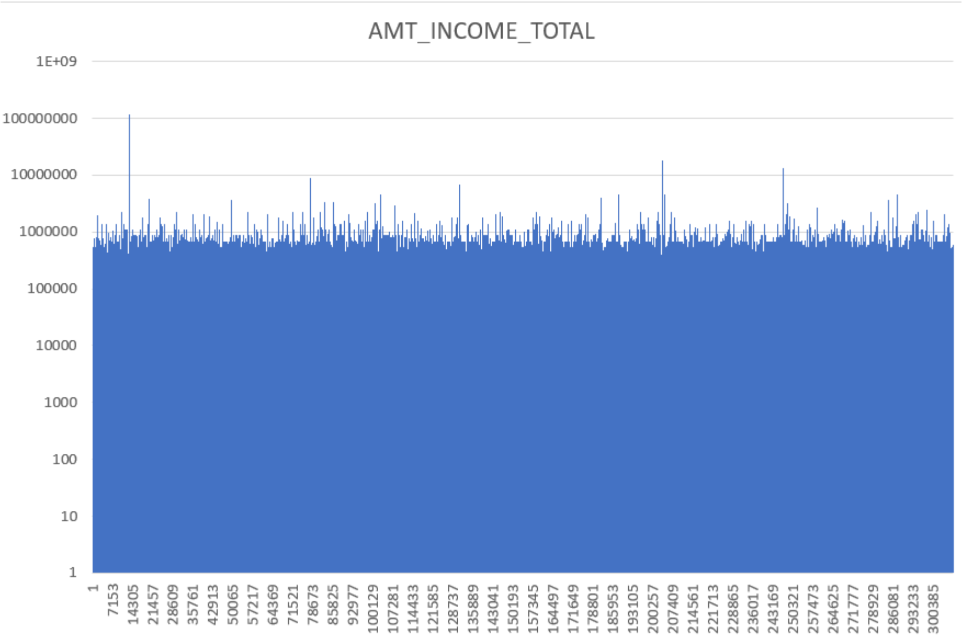


Columns	missing_data-per	Data_count_per
RATE_INTEREST_PRIMARY	99.64513779	0.354862213
RATE_INTEREST_PRIVILEGED	99.64513779	0.354862213
AMT_DOWN_PAYMENT	53.34825516	46.65174484
RATE_DOWN_PAYMENT	53.34825516	46.65174484
DAYS_FIRST_DRAWING	40.1219368	59.8780632
DAYS_FIRST_DUE	40.1219368	59.8780632
DAYS_LAST_DUE_1ST_VERSION	40.1219368	59.8780632
DAYS_LAST_DUE	40.1219368	59.8780632
DAYS_TERMINATION	40.1219368	59.8780632
NFLAG_INSURED_ON_APPROVAL	40.1219368	59.8780632
AMT_GOODS_PRICE	22.98030853	77.01969147
AMT_ANNUITY	22.22156525	77.77843475
CNT_PAYMENT	22.22127914	77.77872086
SK_ID_PREV	0.00009537	99.99990463
SK_ID_CURR	0.00009537	99.99990463
AMT_APPLICATION	0.00009537	99.99990463
AMT_CREDIT	0.00009537	99.99990463
HOUR_APPR_PROCESS_START	0.00009537	99.99990463
NFLAG_LAST_APPL_IN_DAY	0.00009537	99.99990463
DAYS_DECISION	0.00009537	99.99990463
SELLERPLACE_AREA	0.00009537	99.99990463
NAME_CONTRACT_TYPE	0	100
WEEKDAY_APPR_PROCESS_START	0	100
FLAG_LAST_APPL_PER_CONTRACT	0	100
NAME_CASH_LOAN_PURPOSE	0	100
NAME_CONTRACT_STATUS	0	100
NAME_PAYMENT_TYPE	0	100
CODE_REJECT_REASON	0	100
NAME_TYPE_SUITE	0	100
NAME_CLIENT_TYPE	0	100
NAME_GOODS_CATEGORY	0	100
NAME_PORTFOLIO	0	100
NAME_PRODUCT_TYPE	0	100
CHANNEL_TYPE	0	100
NAME_SELLER_INDUSTRY	0	100
NAME_YIELD_GROUP	0	100
PRODUCT_COMBINATION	0	100

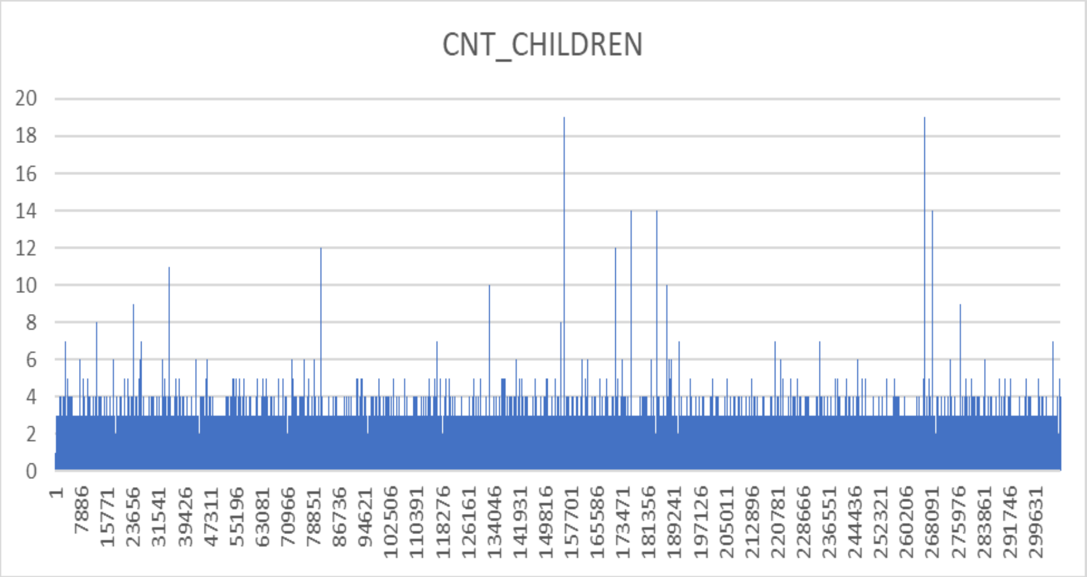


2. **Identifying outliers.**
   - Total Income:

     We can see in the following graph which shows the total annual income of all the applicants. We used exponential values on the y-axis to show the income amount so that we could identify the outliers as the values here are very big.
     Here, we can clearly see 117000000 and 18000090 are very high and above the mean of the values hence we consider them as an outlier.

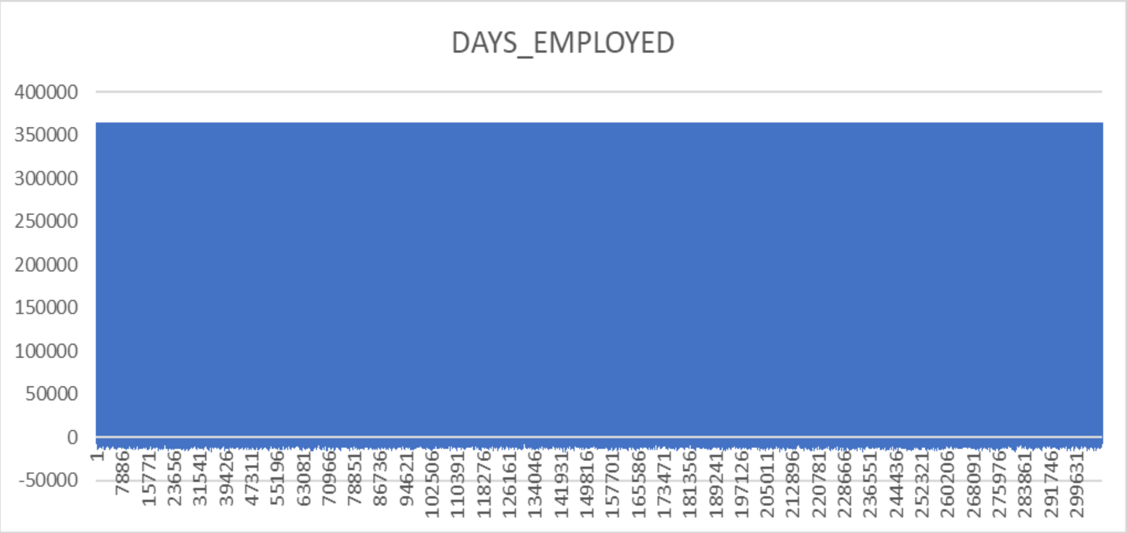


- Count of Children:
  Here, as shown in the plot below, we can see that some values are very well above the mean count of number of children. Since humans generally cannot/do not have 19 children, we will consider these as outliers in the data.



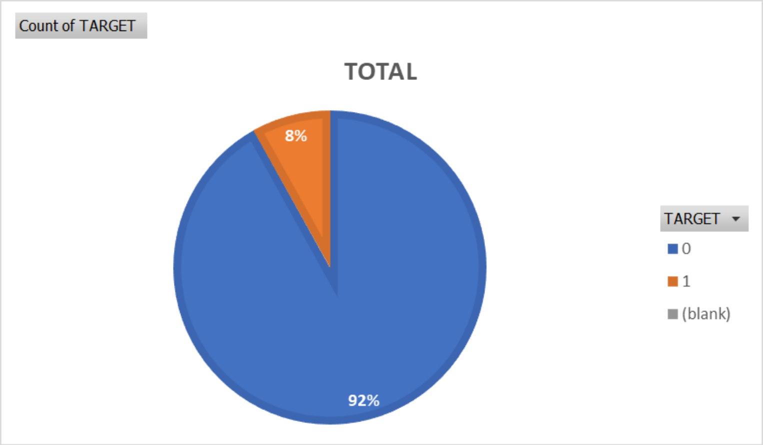
- Days Employed:

  There is a value 365243 that is present and reoccurs very often in the data making the data look as shown in the graph below. This might be due to an error hence; we will treat it as an outlier.



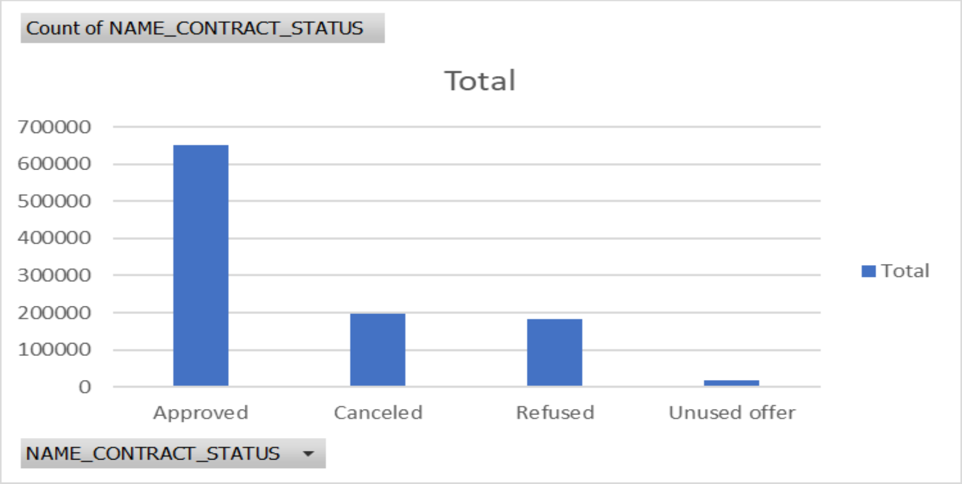
3. **Data Imbalance.**

As we can observe here, the data is highly imbalanced. The Defaulted population is 8% and the Non-Defaulted or Repayors population is at 92%. The ratio here comes out to be 23:2 or 11.5.

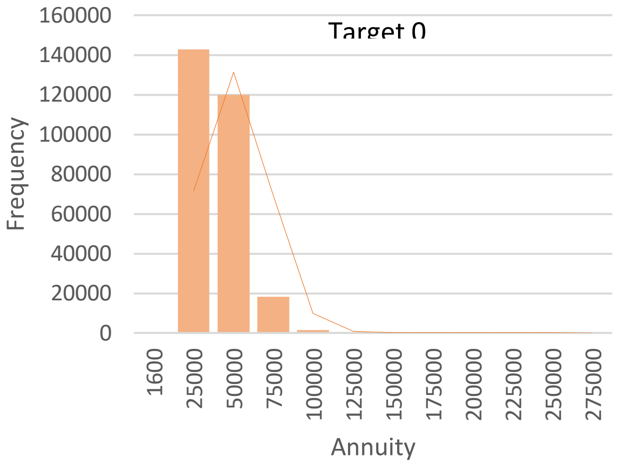


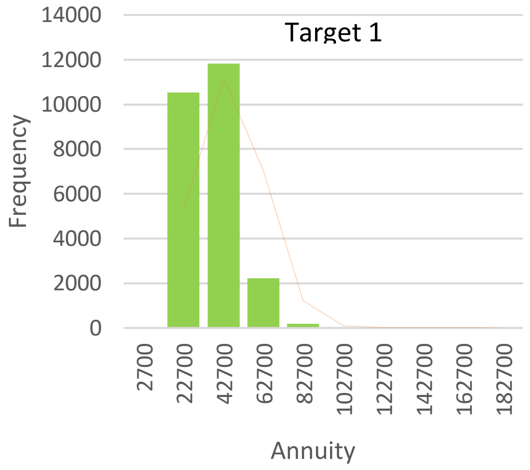
- In the previous_application, when we look at the contract status of the applications received, we can see that the approval rate is very high compared to refusal rate. This shows that the data is imbalanced.

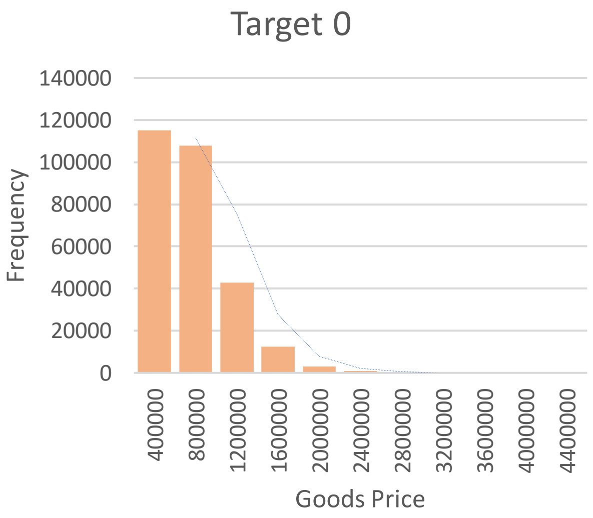
Row Labels	Count of NAME_CONTRACT_STATUS	Percentage
Approved	652486	62.22597334
Canceled	197231	18.80943185
Refused	182083	17.36480462
Unused offer	16775	1.599790191
Grand Total	**1048575**	**100**

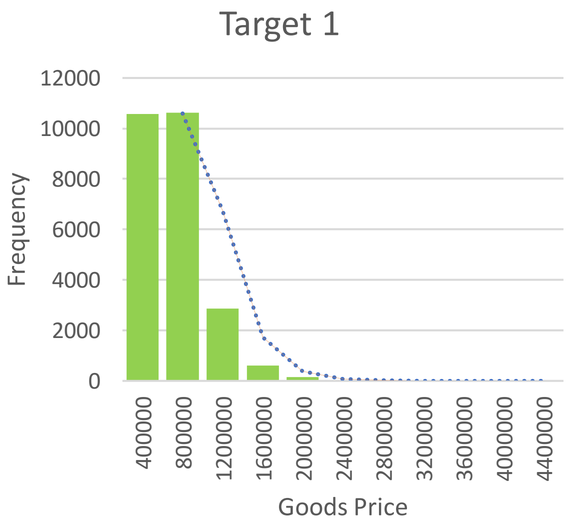


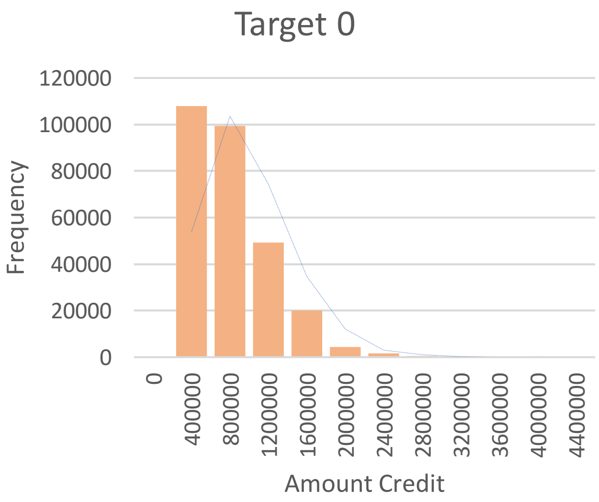
4. Univariate Analysis W.R.T Target Variable.

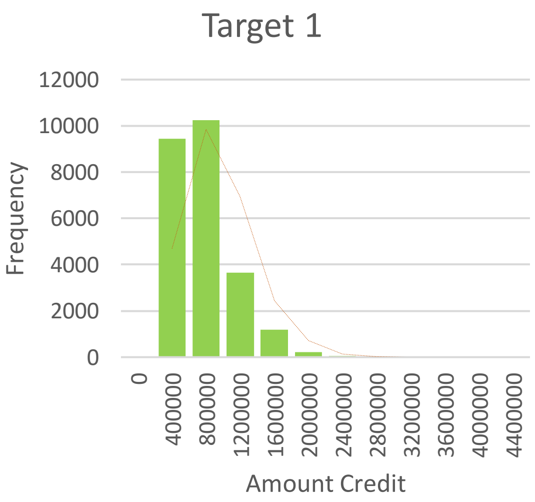


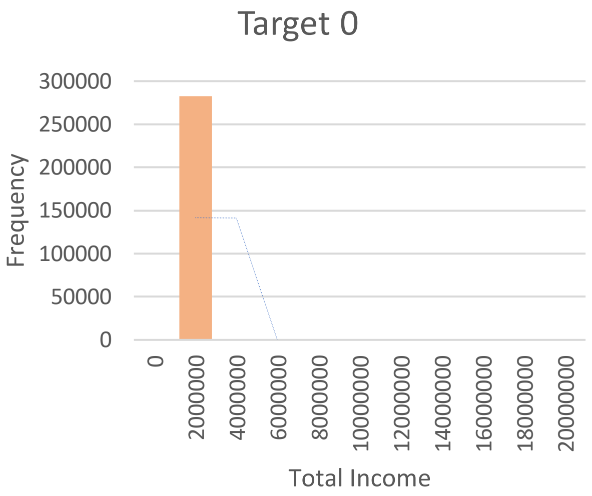


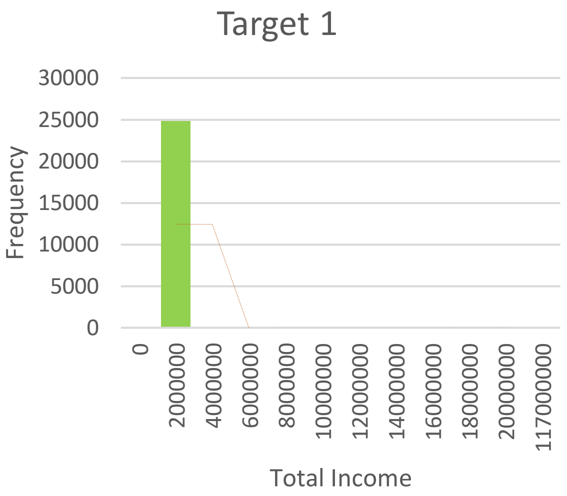










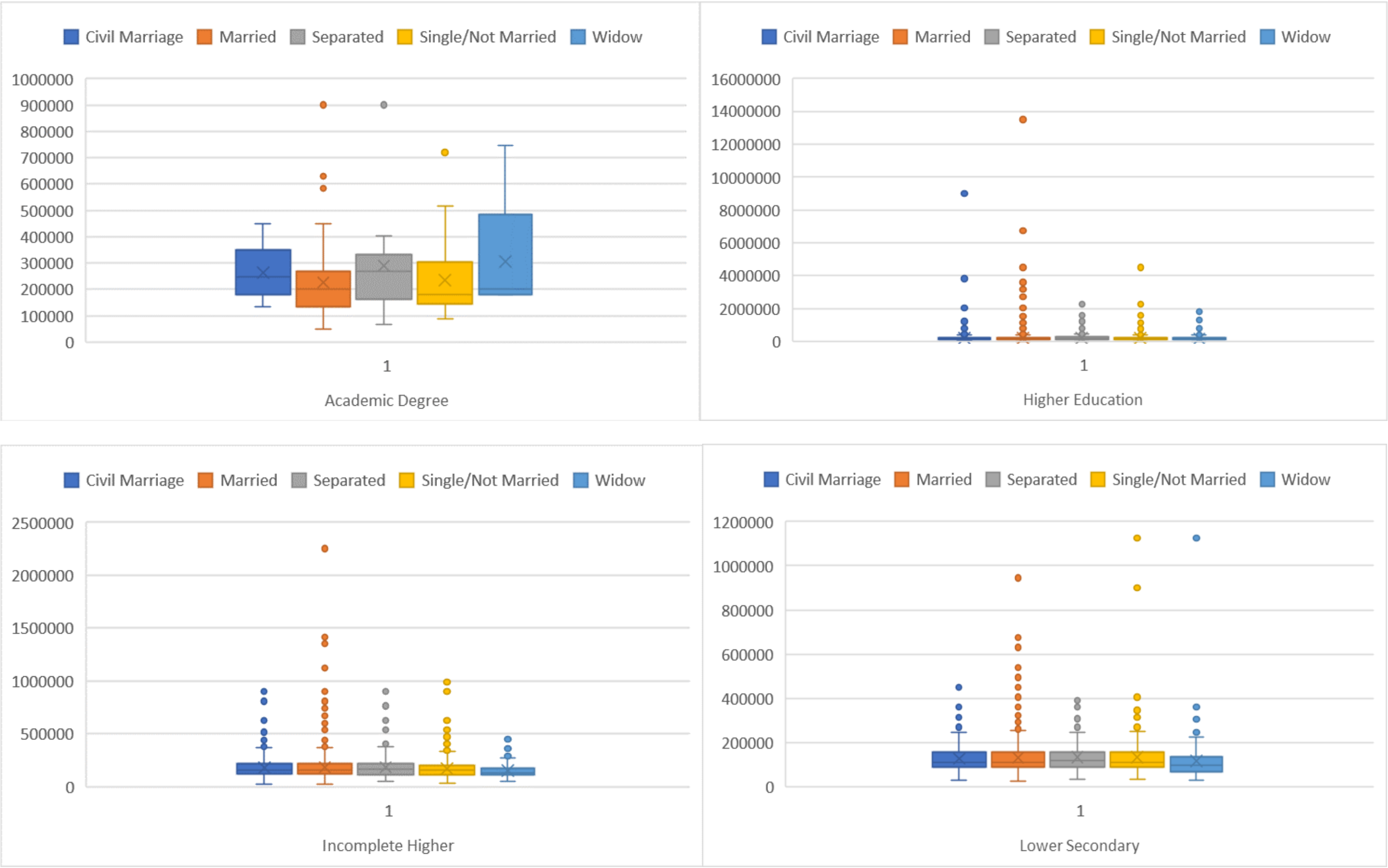


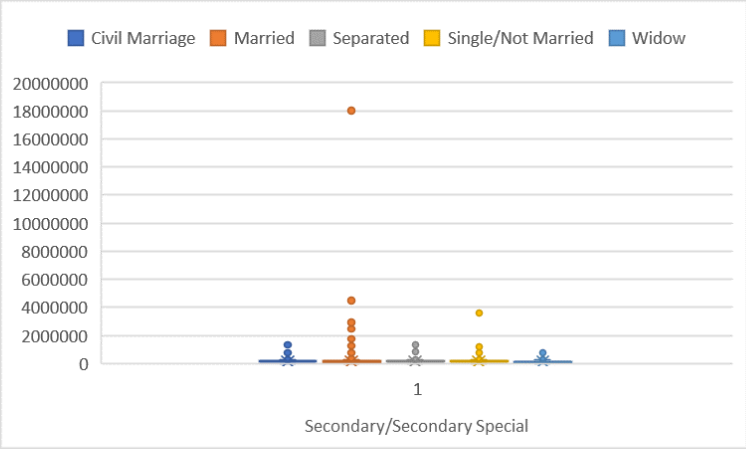
- We can observe that the income of people is highly staggered
- As we can also observe here, people with target 1 have similar plots when compared to Target 0.
- The plots also highlights the people who have difficulty in paying back loans with respect to their income, loan amount, price of goods against which loan is procured.

5. Bivariate Analysis W.R.T Target Variable.

Income Amount Vs Education Status

(For Target 0)

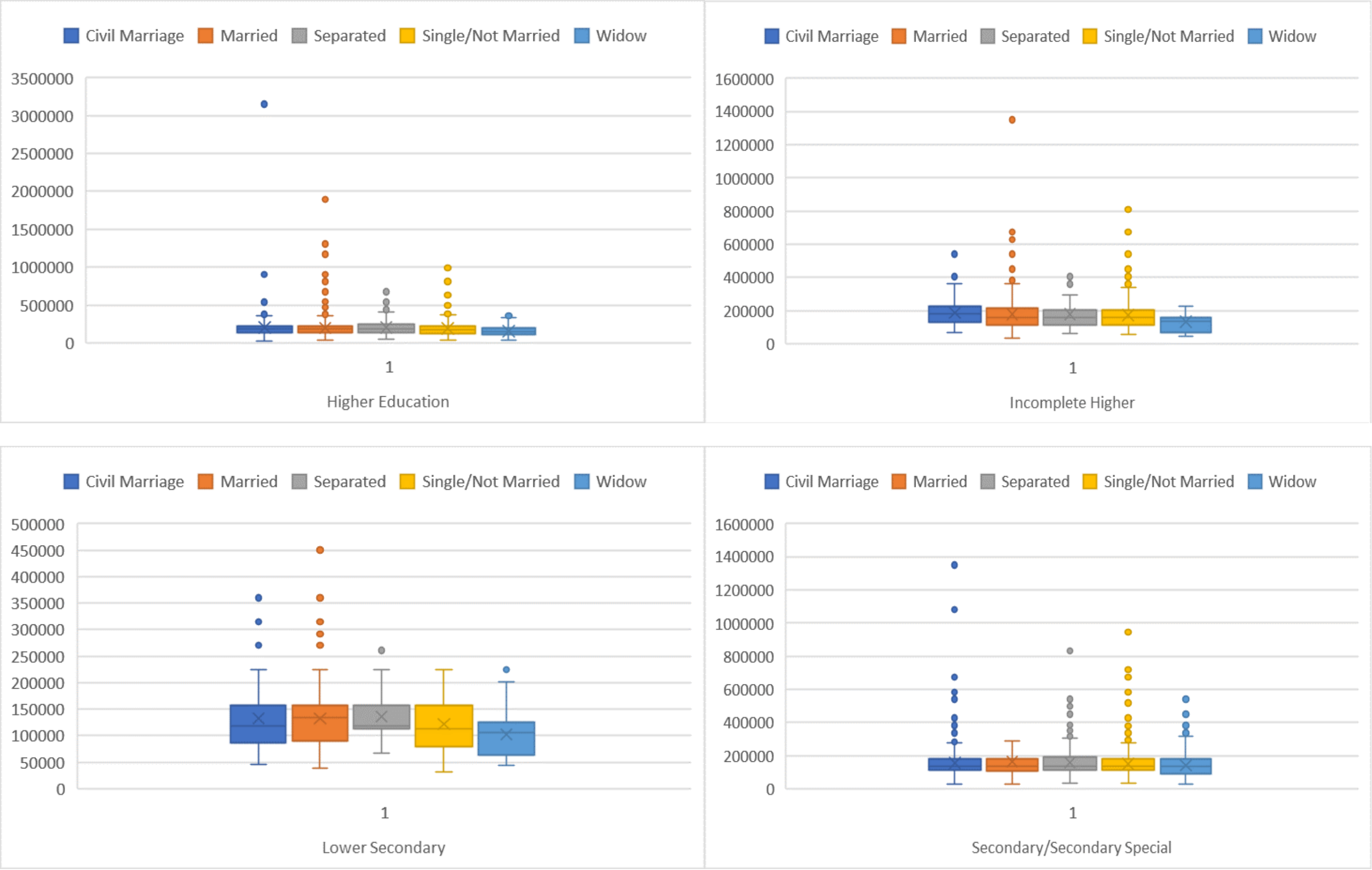




- From the above figures we can say that some of the clients having Higher Education tend to have a higher income compared to others.
- Some of the clients having Secondary/Secondary Special Education tend to have higher incomes.
- Clients having Higher Education, Incomplete Higher Education, Lower Education and Secondary/Secondary Special have a higher number of outliers.
- Clients with all types of family statuses having academic degrees have very less outliers as compared to other types of education.

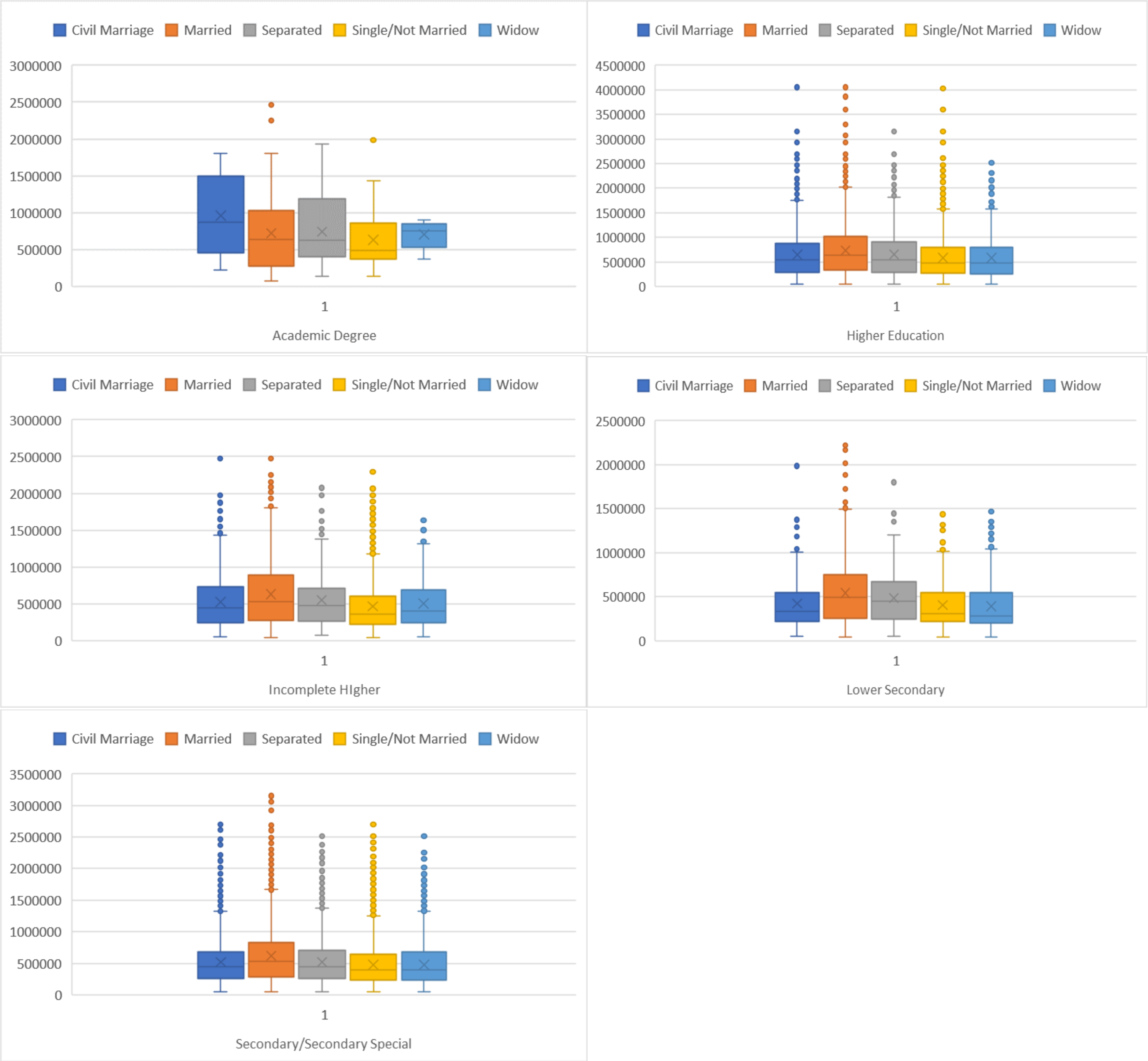
Income Amount Vs Education Status

(For Target 1)



- The income amount for married clients with an academic degree is much lesser as compared to others. The data for academic degree has only three entries, two (270000, 360000) in Married column and one (337500) in Single/Not Married column.
- Clients who defaulted have relatively less income as compared to Non-Defaulters.

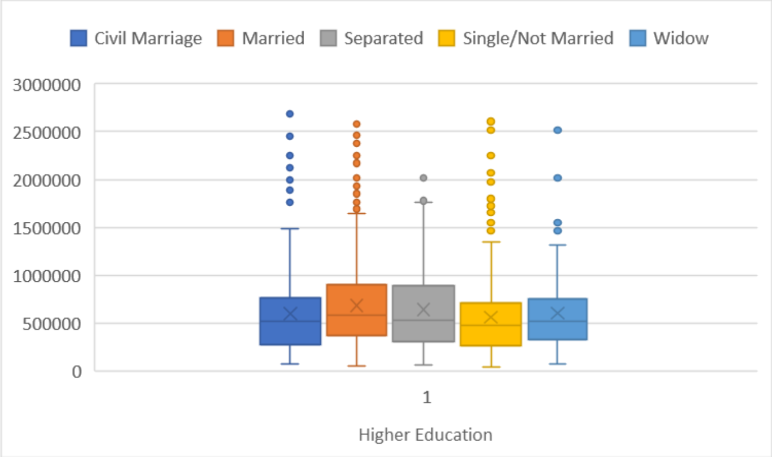
Credit Vs Education
(For Target 0)

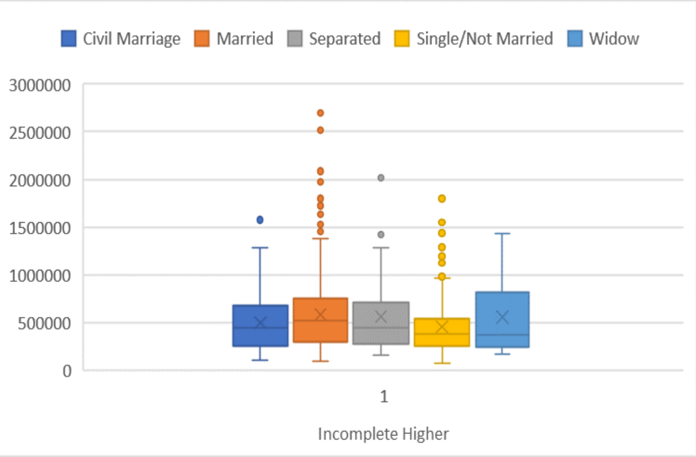


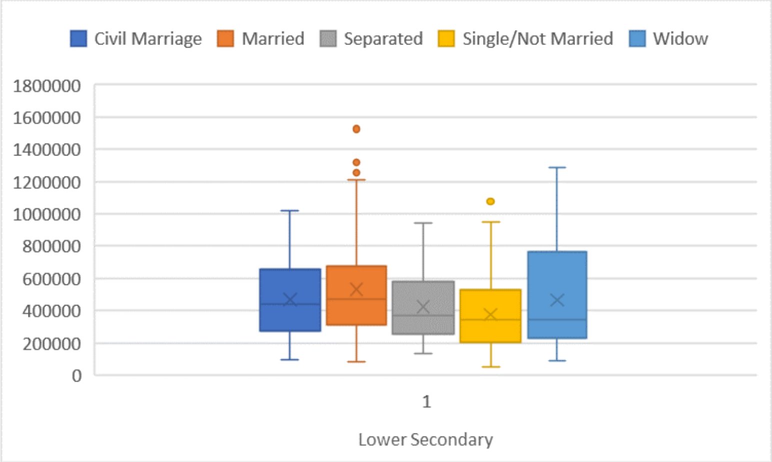
- Clients with different education types except Academic degrees have a large number of outliers.
- Most of the population’s credit amount lie below 25%.
- Clients with a n Academic degree and who is a widow tend to take higher credit loan.
- Some of the clients with Higher Education, Incomplete Higher education, Lower Secondary and Secondary/Secondary Special education are more likely to take a high amount of credit loan.

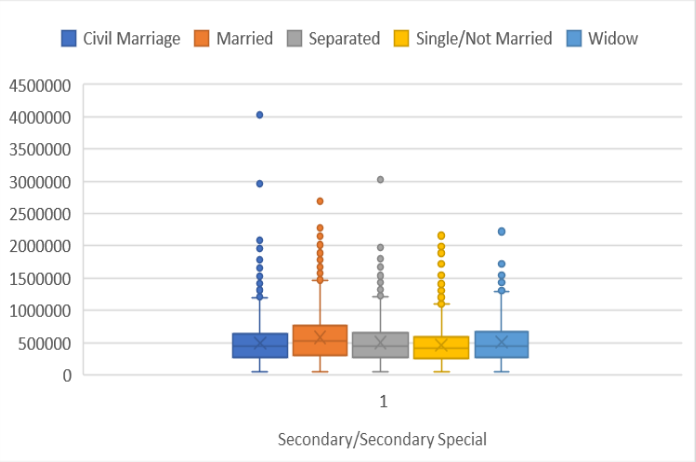
Credit Vs Education

(For Target 1)





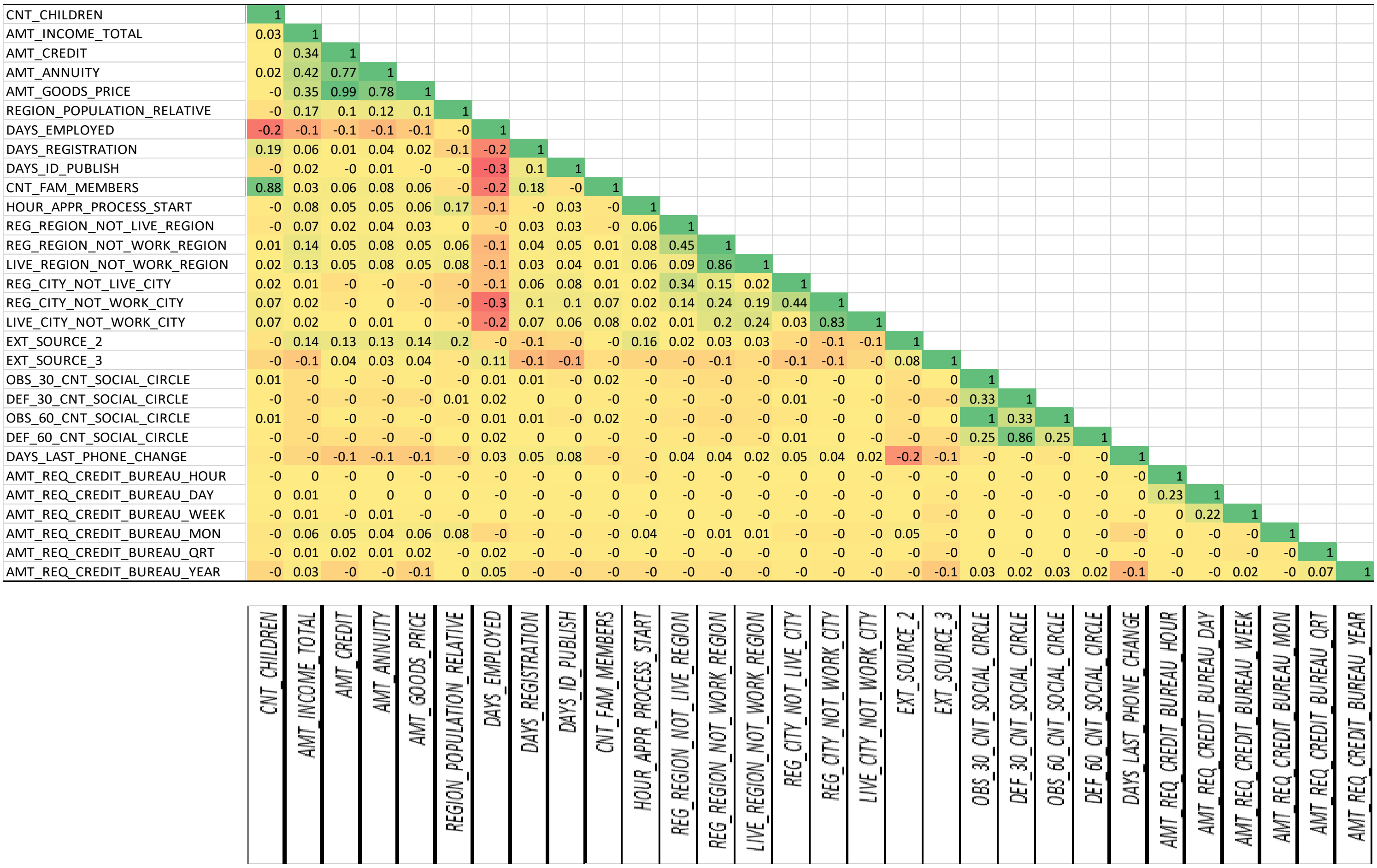




- Married client with academic degrees applied for a higher credit loan and doesn’t have any outliers (as there’s only two clients) and Single client with academic degree has only one entry therefore, no outliers either.
- Some of the clients with Higher Education and Secondary/Secondary Special, Lower Education and Incomplete education are more likely to take a higher amount of credit loans.

6. Correlation Between Variables

A. For Target 0



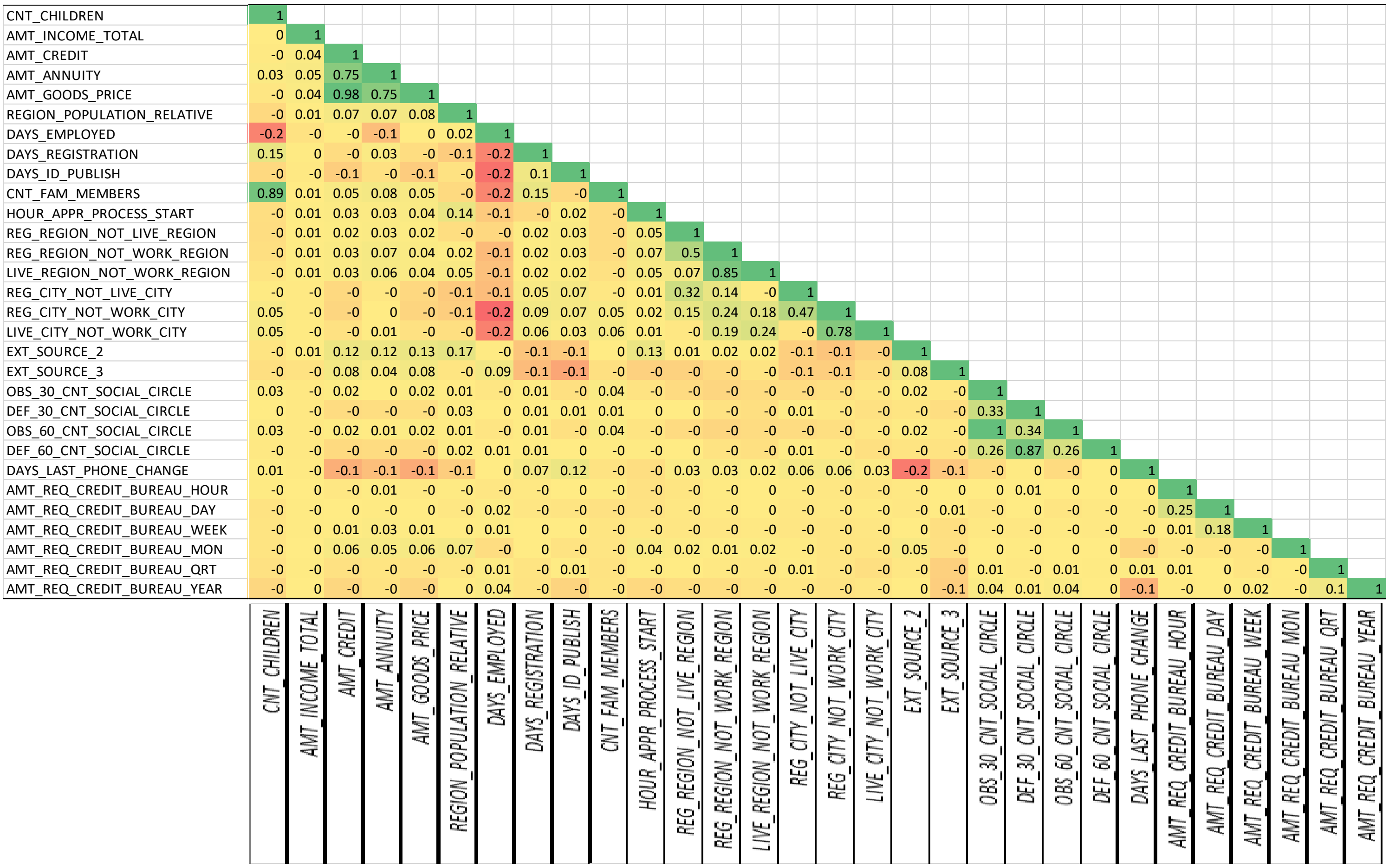
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	HOUR_APPR_PROCESS_START	REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	REG_CITY_NOT_LIVE_CITY	REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	EXT_SOURCE_2	EXT_SOURCE_3	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	DAYS_LAST_PHONE_CHANGE	AMT_REQ_CREDIT_BUREAU_HOUR	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1																													
AMT_INCOME_TOTAL	0.03	1																												
AMT_CREDIT	0	0.34	1																											
AMT_ANNUITY	0.02	0.42	0.77	1																										
AMT_GOODS_PRICE	-0	0.35	0.99	0.78	1																									
REGION_POPULATION_RELATIVE	-0	0.17	0.1	0.12	0.1	1																								
DAYS_EMPLOYED	-0.2	-0.1	-0.1	-0.1	-0.1	-0	1																							
DAYS_REGISTRATION	0.19	0.06	0.01	0.04	0.02	-0.1	-0.2	1																						
DAYS_ID_PUBLISH	-0	0.02	-0	0.01	-0	-0	-0.3	0.1	1																					
CNT_FAM_MEMBERS	0.88	0.03	0.06	0.08	0.06	-0	-0.2	0.18	-0	1																				
HOUR_APPR_PROCESS_START	-0	0.08	0.05	0.05	0.06	0.17	-0.1	-0	0.03	-0	1																			
REG_REGION_NOT_LIVE_REGION	-0	0.07	0.02	0.04	0.03	0	-0	0.03	0.03	-0	0.06	1																		
REG_REGION_NOT_WORK_REGION	0.01	0.14	0.05	0.08	0.05	0.06	-0.1	0.04	0.05	0.01	0.08	0.45	1																	
LIVE_REGION_NOT_WORK_REGION	0.02	0.13	0.05	0.08	0.05	0.08	-0.1	0.03	0.04	0.01	0.06	0.09	0.86	1																
REG_CITY_NOT_LIVE_CITY	0.02	0.01	-0	-0	-0	-0	-0.1	0.06	0.08	0.01	0.02	0.34	0.15	0.02	1															
REG_CITY_NOT_WORK_CITY	0.07	0.02	-0	0	-0	-0	-0.3	0.1	0.1	0.07	0.02	0.14	0.24	0.19	0.44	1														
LIVE_CITY_NOT_WORK_CITY	0.07	0.02	0	0.01	0	-0	-0.2	0.07	0.06	0.08	0.02	0.01	0.2	0.24	0.03	0.83	1													
EXT_SOURCE_2	-0	0.14	0.13	0.13	0.14	0.2	-0	-0.1	-0	-0	0.16	0.02	0.03	0.03	-0	-0.1	-0.1	1												
EXT_SOURCE_3	-0	-0.1	0.04	0.03	0.04	-0	0.11	-0.1	-0.1	-0	-0	-0	-0.1	-0	-0.1	-0.1	-0	0.08	1											
OBS_30_CNT_SOCIAL_CIRCLE	0.01	-0	-0	-0	-0	-0	0.01	0.01	-0	0.02	-0	-0	-0	-0	-0	-0	0	-0	0	1										
DEF_30_CNT_SOCIAL_CIRCLE	-0	-0	-0	-0	-0	0.01	0.02	0	0	-0	-0	-0	-0	-0	0.01	-0	-0	-0	-0	0.33	1									
OBS_60_CNT_SOCIAL_CIRCLE	0.01	-0	-0	-0	-0	-0	0.01	0.01	-0	0.02	-0	-0	-0	-0	-0	-0	0	-0	0	1	0.33	1								
DEF_60_CNT_SOCIAL_CIRCLE	-0	-0	-0	-0	-0	0	0.02	0	0	-0	-0	-0	-0	-0	0.01	0	-0	-0	-0	0.25	0.86	0.25	1							
DAYS_LAST_PHONE_CHANGE	-0	-0	-0.1	-0.1	-0.1	-0	0.03	0.05	0.08	-0	-0	0.04	0.04	0.02	0.05	0.04	0.02	-0.2	-0.1	-0	-0	-0	-0	1						
AMT_REQ_CREDIT_BUREAU_HOUR	-0	0	-0	0	-0	-0	-0	-0	0	0	-0	-0	-0	-0	0	0	-0	-0	-0	0	-0	0	-0	-0	1					
AMT_REQ_CREDIT_BUREAU_DAY	0	0.01	0	0	0	0	-0	-0	-0	0	0	-0	-0	-0	-0	-0	-0	0	-0	-0	-0	-0	-0	0	0.23	1				
AMT_REQ_CREDIT_BUREAU_WEEK	-0	0.01	-0	0.01	-0	-0	0	-0	-0	-0	-0	0	-0	-0	-0	-0	-0	0	-0	0	-0	0	-0	-0	0	0.22	1			
AMT_REQ_CREDIT_BUREAU_MON	-0	0.06	0.05	0.04	0.06	0.08	-0	-0	-0	-0	0.04	-0	0.01	0.01	-0	-0	-0	0.05	-0	0	0	0	-0	-0	0	-0	-0	1		
AMT_REQ_CREDIT_BUREAU_QRT	-0	0.01	0.02	0.01	0.02	-0	0.02	-0	-0	-0	-0	-0	-0	-0	0	-0	-0	-0	-0	0	-0	0	-0	-0	-0	-0	-0	-0	1	
AMT_REQ_CREDIT_BUREAU_YEAR	-0	0.03	-0	-0	-0.1	0	0.05	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0.1	0.03	0.02	0.03	0.02	-0.1	-0	-0	0.02	-0	0.07	1

- We can see here that Credit is inversely proportional to the age column; people belonging to the low-age group take high credit and vice-versa.
- Clients have fewer children in a densely populated area.
- Income is inversely proportional to the number of children, means more income for fewer children clients have and vice-versa.
- Credit is inversely proportional to the number of children, means the credit amount is higher for fewer children count and vice-versa.
- Both credit and incomes are higher in a densely populated area.

B. For Target 1

This heatmap for Target 1 is having very similar observations as Target 0 except for a few points as mentioned below:

- The clients whose permanent address does not match the contact address are having fewer children.
- The clients whose permanent address does not match the work address are also having fewer children.



	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	HOUR_APPR_PROCESS_START	REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	REG_CITY_NOT_LIVE_CITY	REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	EXT_SOURCE_2	EXT_SOURCE_3	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	DAYS_LAST_PHONE_CHANGE	AMT_REQ_CREDIT_BUREAU_HOUR	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1																													
AMT_INCOME_TOTAL	0	1																												
AMT_CREDIT	-0	0.04	1																											
AMT_ANNUITY	0.03	0.05	0.75	1																										
AMT_GOODS_PRICE	-0	0.04	0.98	0.75	1																									
REGION_POPULATION_RELATIVE	-0	0.01	0.07	0.07	0.08	1																								
DAYS_EMPLOYED	-0.2	-0	-0	-0.1	0	0.02	1																							
DAYS_REGISTRATION	0.15	0	-0	0.03	-0	-0.1	-0.2	1																						
DAYS_ID_PUBLISH	-0	-0	-0.1	-0	-0.1	-0	-0.2	0.1	1																					
CNT_FAM_MEMBERS	0.89	0.01	0.05	0.08	0.05	-0	-0.2	0.15	-0	1																				
HOUR_APPR_PROCESS_START	-0	0.01	0.03	0.03	0.04	0.14	-0.1	-0	0.02	-0	1																			
REG_REGION_NOT_LIVE_REGION	-0	0.01	0.02	0.03	0.02	-0	-0	0.02	0.03	-0	0.05	1																		
REG_REGION_NOT_WORK_REGION	-0	0.01	0.03	0.07	0.04	0.02	-0.1	0.02	0.03	-0	0.07	0.5	1																	
LIVE_REGION_NOT_WORK_REGION	-0	0.01	0.03	0.06	0.04	0.05	-0.1	0.02	0.02	-0	0.05	0.07	0.85	1																
REG_CITY_NOT_LIVE_CITY	-0	-0	-0	-0	-0	-0.1	-0.1	0.05	0.07	-0	0.01	0.32	0.14	-0	1															
REG_CITY_NOT_WORK_CITY	0.05	-0	-0	0	-0	-0.1	-0.2	0.09	0.07	0.05	0.02	0.15	0.24	0.18	0.47	1														
LIVE_CITY_NOT_WORK_CITY	0.05	-0	-0	0.01	-0	-0	-0.2	0.06	0.03	0.06	0.01	-0	0.19	0.24	-0	0.78	1													
EXT_SOURCE_2	-0	0.01	0.12	0.12	0.13	0.17	-0	-0.1	-0.1	0	0.13	0.01	0.02	0.02	-0.1	-0.1	-0	1												
EXT_SOURCE_3	-0	-0	0.08	0.04	0.08	-0	0.09	-0.1	-0.1	-0	-0	-0	-0	-0	-0.1	-0.1	-0	0.08	1											
OBS_30_CNT_SOCIAL_CIRCLE	0.03	-0	0.02	0	0.02	0.01	-0	0.01	-0	0.04	-0	-0	-0	-0	-0	-0	-0	0.02	-0	1										
DEF_30_CNT_SOCIAL_CIRCLE	0	-0	-0	-0	-0	0.03	0	0.01	0.01	0.01	0	0	-0	-0	0.01	-0	-0	-0	-0	0.33	1									
OBS_60_CNT_SOCIAL_CIRCLE	0.03	-0	0.02	0.01	0.02	0.01	-0	0.01	-0	0.04	-0	-0	-0	-0	-0	-0	-0	0.02	-0	1	0.34	1								
DEF_60_CNT_SOCIAL_CIRCLE	-0	-0	-0	-0	-0	0.02	0.01	0.01	0	-0	-0	0	-0	-0	0.01	-0	-0	-0	-0	0.26	0.87	0.26	1							
DAYS_LAST_PHONE_CHANGE	0.01	-0	-0.1	-0.1	-0.1	-0.1	0	0.07	0.12	-0	-0	0.03	0.03	0.02	0.06	0.06	0.03	-0.2	-0.1	-0	0	-0	0	1						
AMT_REQ_CREDIT_BUREAU_HOUR	-0	0	-0	0.01	-0	-0	-0	-0	0	-0	-0	-0	-0	0	-0	-0	-0	-0	0	0	0.01	0	0	0	1					
AMT_REQ_CREDIT_BUREAU_DAY	-0	-0	0	-0	0	-0	0.02	-0	-0	-0	-0	-0	-0	0	-0	-0	-0	-0	0.01	-0	0	-0	-0	-0	0.25	1				
AMT_REQ_CREDIT_BUREAU_WEEK	-0	0	0.01	0.03	0.01	0	0.01	0	0	-0	-0	-0	-0	-0	-0	-0	-0	0	-0	-0	-0	-0	-0	-0	0.01	0.18	1			
AMT_REQ_CREDIT_BUREAU_MON	-0	0	0.06	0.05	0.06	0.07	-0	0	-0	-0	0.04	0.02	0.01	0.02	-0	-0	-0	0.05	-0	0	-0	0	0	-0	-0	-0	-0	1		
AMT_REQ_CREDIT_BUREAU_QRT	-0	-0	-0	-0	-0	-0	0.01	-0	0.01	-0	-0	0	-0	-0	0.01	-0	-0	-0	-0	0.01	-0	0.01	0	0.01	0.01	0	-0	-0	1	
AMT_REQ_CREDIT_BUREAU_YEAR	-0	0	-0	-0	-0	0	0.04	-0	-0	-0	-0	-0	-0	-0	-0	-0	-0	0	-0.1	0.04	0.01	0.04	0	-0.1	-0	0	0.02	-0	0.1	1

7. Top 10 Correlation Fields for Target 0

Var1	Var2	Correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.99850846
AMT_GOODS_PRICE	AMT_CREDIT	0.98687958
CNT_FAM_MEMBERS	CNT_CHILDREN	0.87857137
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.86186136
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.85933184
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.83038113
AMT_GOODS_PRICE	AMT_ANNUITY	0.77626269
AMT_ANNUITY	AMT_CREDIT	0.77130895
REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.44610086
REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.43551371

8. Top 10 Correlation Fields for Target 1

Var1	Var2	Correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.99826866
AMT_GOODS_PRICE	AMT_CREDIT	0.98256585
CNT_FAM_MEMBERS	CNT_CHILDREN	0.88548371
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86899444
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.84788518
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.77853974
AMT_ANNUITY	AMT_CREDIT	0.75219474
AMT_GOODS_PRICE	AMT_ANNUITY	0.7520224
REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.49793654
REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.47205229

Result

From this project we were able to get an insight into the bank loan datasets and were able to identify patterns which indicate if a client has difficulty paying their installments or not. We used various univariate and bivariate analysis with the help of graphs and plots to get better understanding of the data and found the top 10 correlations which can help us identify loan defaulters and re-payers.