TABLE OF CONTENTS

Project Description

This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop 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.

Approach

I first analysed the data and cleaned it by deleting irrelevant columns, empty cells, and outliers. I then ran different analyses, including univariate and bivariate analysis, on the dataset.

Data Cleaning

First cleaning the data

At first, I deleted some irrelevant columns which in my opinion are not relevant for our analysis.

Columns I deleted:
['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_NOT_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']

Counting percentage of blank cells in each column using below formulas:

For column having only numbers:

```
fx  =100*COUNTBLANK(B1:B307512)/COUNT(B1:B307512)
```

For column having only text:

```
fx  =100*COUNTBLANK(H1:H307512)/COUNTA(H1:H307512)
```

This will give us the percentage of blank cells for each column & after sorting in descending order we can see the column which has highest percentage of blank cells.

	A	B	C	D	E	F	G	H	I	J
1		COMMON	COMMON	COMMON	NONLIVIN	NONLIVIN	NONLIVIN	FONDKAPI	LIVINGAP/	LIVINGAP/
307509		0.0022	0.0022	0.0022	0	0	0	reg oper a	0.0202	0.022
307510		0.0123	0.0124	0.0124	0	0	0	reg oper a	0.0841	0.0918
307511										
307512		0.0176	0.0178	0.0177						
307513	**Percentage of empty cells in column**	**231.9204**	**231.9204**	**231.9204**	**227.1498**	**227.1498**	**227.1498**	**216.315**	**216.0052**	**216.0052**

Now I deleted the columns which have more than 30% of blank cells in them which are:

['OWN_CAR_AGE', 'OCCUPATION_TYPE', 'EXT_SOURCE_1', 'APARTMENTS_AVG', 'BASEMENTAREA_AVG', 'YEARS_BEGINEXPLUATATION_AVG', 'YEARS_BUILD_AVG', 'COMMONAREA_AVG', 'ELEVATORS_AVG', 'ENTRANCES_AVG', 'FLOORSMAX_AVG', 'FLOORSMIN_AVG', 'LANDAREA_AVG', 'LIVINGAPARTMENTS_AVG', 'LIVINGAREA_AVG', 'NONLIVINGAPARTMENTS_AVG', 'NONLIVINGAREA_AVG', 'APARTMENTS_MODE', 'BASEMENTAREA_MODE', 'YEARS_BEGINEXPLUATATION_MODE', 'YEARS_BUILD_MODE', 'COMMONAREA_MODE', 'ELEVATORS_MODE', 'ENTRANCES_MODE', 'FLOORSMAX_MODE', 'FLOORSMIN_MODE', 'LANDAREA_MODE', 'LIVINGAPARTMENTS_MODE', 'LIVINGAREA_MODE', 'NONLIVINGAPARTMENTS_MODE', 'NONLIVINGAREA_MODE', 'APARTMENTS_MEDI', 'BASEMENTAREA_MEDI', 'YEARS_BEGINEXPLUATATION_MEDI', 'YEARS_BUILD_MEDI', 'COMMONAREA_MEDI', 'ELEVATORS_MEDI', 'ENTRANCES_MEDI', 'FLOORSMAX_MEDI', 'FLOORSMIN_MEDI', 'LANDAREA_MEDI', 'LIVINGAPARTMENTS_MEDI', 'LIVINGAREA_MEDI', 'NONLIVINGAPARTMENTS_MEDI', 'NONLIVINGAREA_MEDI', 'FONDKAPREMONT_MODE', 'HOUSETYPE_MODE', 'TOTALAREA_MODE', 'WALLSMATERIAL_MODE', 'EMERGENCYSTATE_MODE']

Treating of missing values:

As all the values in both columns are non-zero, I started by counting the rows for which EXT SOURCE 2 and EXT SOURCE 3 are both empty. To do this, I added the values in these two columns and counted any cells where the sum was zero.

AA	AB	AC
EXT_SOURCE_2	**EXT_SOURCE_3**	**AB+AC**
0.616890562		0.616891
0.200589492	0.176652579	0.377242
nt		0
0.357648626		0.357649
0.743712069	0.6041126	1.347825
0.312901058	0.832785025	1.145686

Now, using below formula, I found the rows where these two columns are empty by counting the zeros in the sum column:

fx =COUNTIF(AC2:AC307512,0)

AC
AB+AC
1.175186
0.822491
230

Thus, there are 230 such cells with empty rows in both columns.

After determining the means for these columns, I discovered that there wasn't any much difference in the means for these columns.

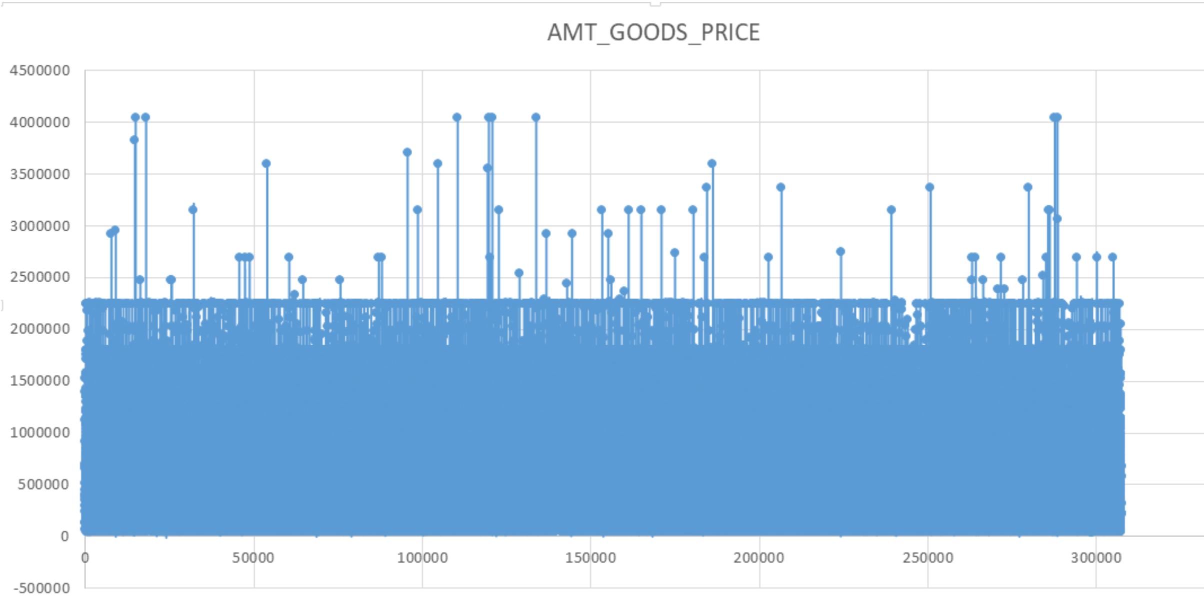
Z	AA	AB
ORGANIZA	**EXT_SOURCE_2**	**EXT_SOURCE_3**
Business E	0.51416282	0.661023539
Business E	0.708568896	0.113922396
0	**0.215088105**	**24.72763703**
Mean	**0.514392674**	**0.510852906**

Therefore, we can use mean to replace the empty cells in these columns.

There are no vacant cells in these columns after replacing with mean values.

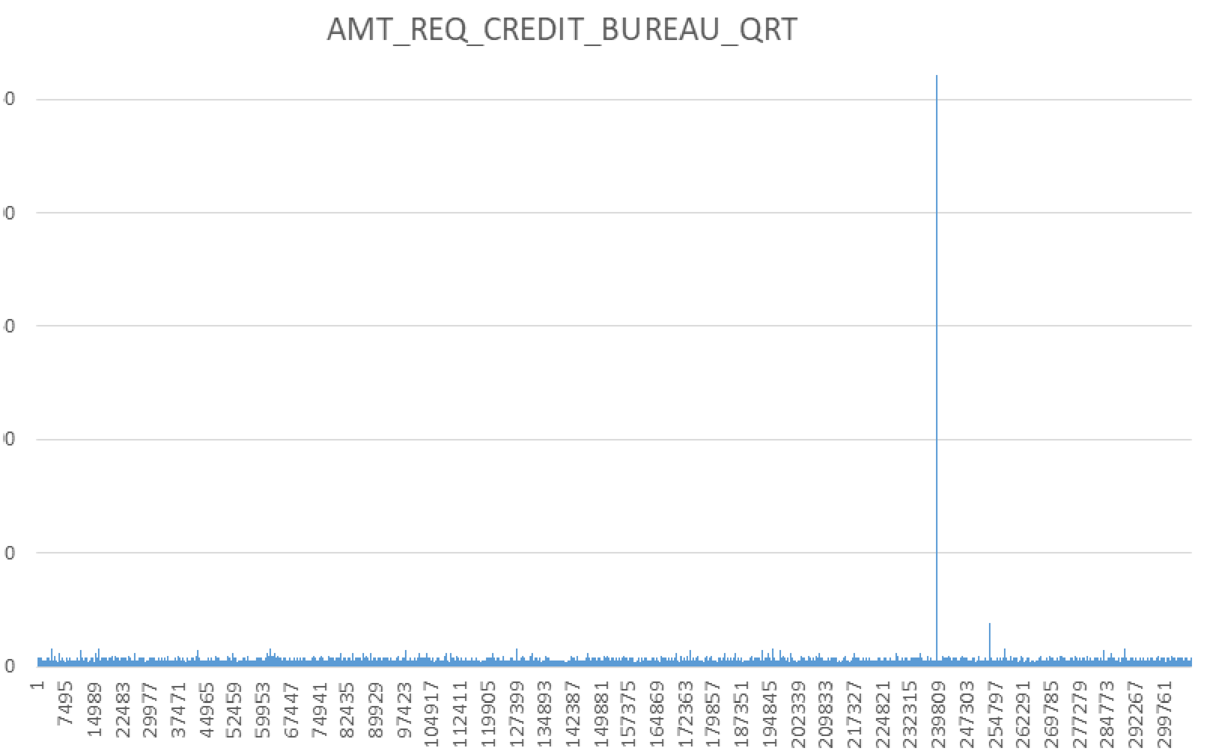
A	B	AA	AB
	SK_ID_CU	**EXT_SOURCE_2**	**EXT_SOURCE_3**
	100002	0.262948593	0.13937578
	100003	0.622245775	0.510852906
	100004	0.555912083	0.729566691
	100006	0.65044169	0.510852906
	456254	0.51416282	0.661023539
	456255	0.708568896	0.113922396
Percentage of empty cells in column	**0**	**0**	**0**

For AMT_GOODS_PRICE column, we can see there are many outliers in below plot:



Therefore, we are unable to substitute the mean value for the empty cells in this column. I consequently deleted the empty cells from this column. Additionally, the database won't be significantly impacted because only 0.09% of the cells are empty.

We are now left with 6 columns of information about credit bureaus for different time periods, including hour, day, week, month, quarter, and year. But I just consider the quarter-specific information because the rest is irrelevant. I thus eliminated other columns.



There is only one outlier for this column, and by filtering, we can see all of its values:

	AW	AX	AY	AZ
1	AMT_R▾_CREDIT_BUREAU_QRT			

Sort Smallest to Largest
Sort Largest to Smallest
Sort by Color
Clear Filter From "AMT_REQ_CREDIT_BU..."
Filter by Color
Number Filters
Search
☑ (Select All)
☑ 0
☑ 1
☑ 2
☑ 3
☑ 4
☑ 5
☑ 6
☑ 7
☑ 8
☑ 19
☑ 261
☑ (Blanks)

	AV	AW	AX	AY
OC	FLAG_DOC	AMT_REQ_CREDIT_BUREAU_QRT		
0	0	0		
0	0	0		
0	0	15.60543		
	Mean	0.265525		
	Mode	0		

Since they are all integer values and the mean is 0.265525, the blanks cannot be filled in with the mean value. So I identified the column's mode, which is zero (the value that occurs the most often), and I changed the blanks to that value.

1		SK_ID_CU	TARGI	NAME_CONTR	CODE_GEN	FLAG_OWI	FLAG_OWN_	CNT_CHI	AMT_INCON	AMT_CRE	AMT_ANN	AMT_GOC	NAME_IN	NAME_ED	NAME_FA	NAME_HC	REGION_F	DAYS_BIR	DAYS_EM	DA
307217		456249	0	Cash loans	F	N	Y	0	112500	225000	22050	225000	Pensioner	Secondary	Single / no	House / ap	0.0228	-24384	365243	
307218		456251	0	Cash loans	M	N	N	0	157500	254700	27558	225000	Working	Secondary	Separated	With pare	0.03256	-9327	-236	
307219		456252	0	Cash loans	F	N	Y	0	72000	269550	12001.5	225000	Pensioner	Secondary	Widow	House / ap	0.02516	-20775	365243	
307220		456253	0	Cash loans	F	N	Y	0	153000	677664	29979	585000	Working	Higher edu	Separated	House / ap	0.005	-14966	-7921	
307221		456254	1	Cash loans	F	N	Y	0	171000	370107	20205	319500	Commerci	Secondary	Married	House / ap	0.00531	-11961	-4786	
307222		456255	0	Cash loans	F	N	N	0	157500	675000	49117.5	675000	Commerci	Higher edu	Married	House / ap	0.04622	-16856	-1262	
307223	**Percentage of empty cells in column**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	

In our database, there are no longer any empty cells. The total number of rows was 307512 at the start, and it is now 307222. Only 0.09% of the database is lost after cleaning.

100*(307512-307222)/307512	=	**0.09431**

The age of the applicants was then calculated by dividing the "DAYS BIRTH" column by 365 using the absolute function, and the old column was eliminated. The "YEARS EMPLOYED" column will be constructed similarly to the "DAYS EMPLOYED" column.

fx =ROUND(ABS(R2/365),0)

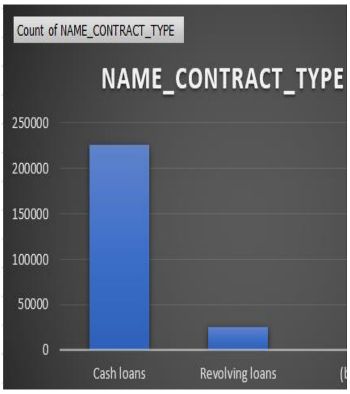
	R	S	T	
_I	DAYS_BIRTH	AGE	DAYS_EM	D
8	-9461	26	-637	
4	-16765	46	-1188	
3	-19046	52	-225	
2	-19005	52	-3039	
6	-19932	55	-3038	
9	-16941	46	-1588	

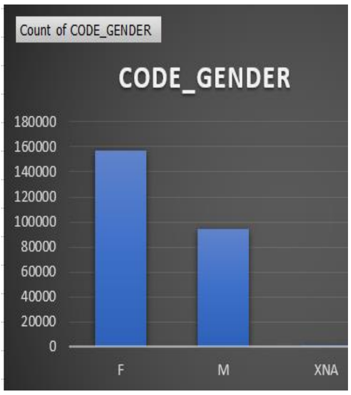
We lost about 18% of the data after manually deleting the outliers from the "AMT INCOME TOTAL," "AMT CREDIT," "AMT ANNUITY," "AMT GOODS PRICE," and "AGE" columns, but the data is now free of outliers.

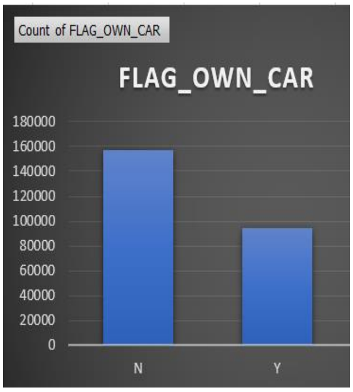
100*(307222-251788)/307222	=	**18.044**

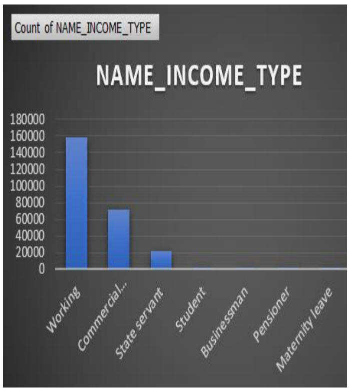
Checking Data Imbalance

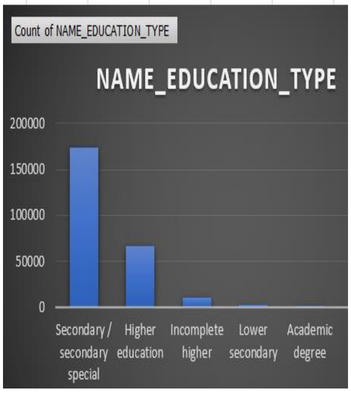
After that, I created pivot tables and plotted their graph as shown below to examine the data imbalance:

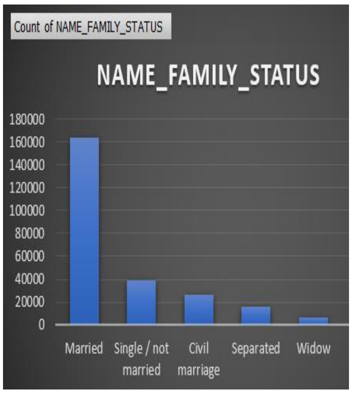


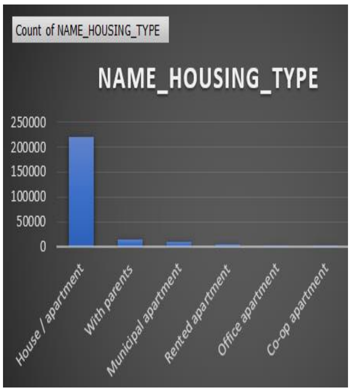










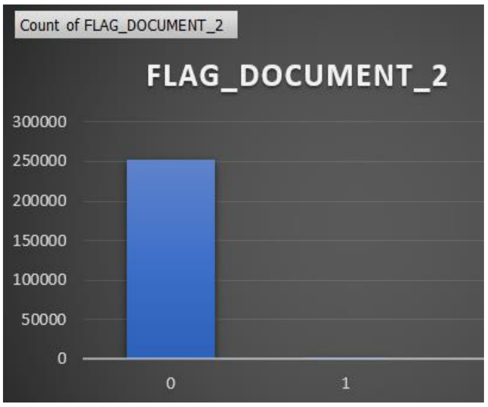


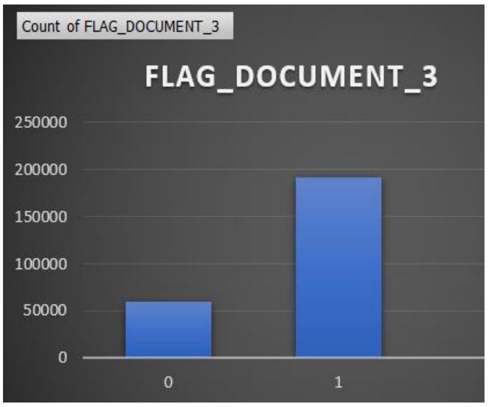
These graphs allow me to deduce the following:

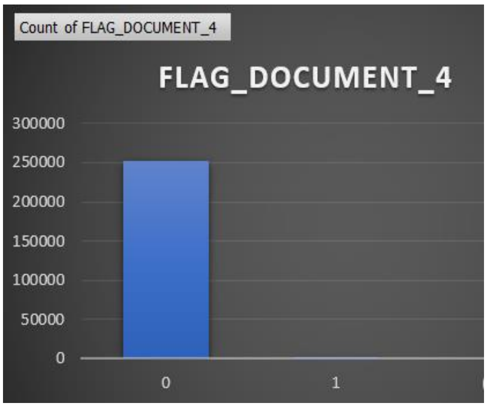
1. Target - There are minimal data for those who experienced payment difficulties (defaulters).
2. NAME_CONTRACT_TYPE - Cash loans are more prevalent than revolving loans in terms of quantity.
3. CODE_GENDER - Female applicants default less frequently than male applications.
4. FLAG_OWN_CAR – The majority of applicants don't own a vehicle.
5. FLAG_OWN_REALTY - The majority of candidates own a home or apartment.
6. NAME_INCOME_TYPE - The majority of candidates are employed professionals (9 to 5 job).

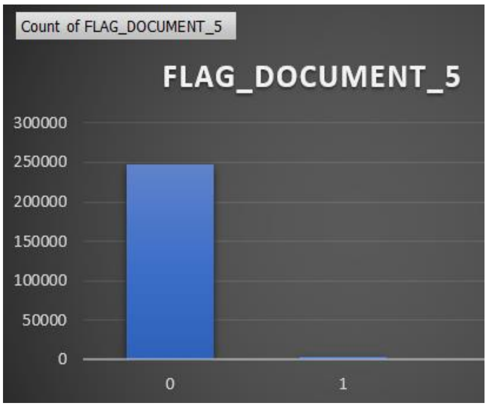
7. NAME_EDUCATION_TYPE - The majority of candidates have completed secondary or secondary special education.
8. NAME_FAMILY_STATUS - The majority of applicants are married families.
9. NAME_HOUSING_TYPE - The majority of candidates own their home or apartment.

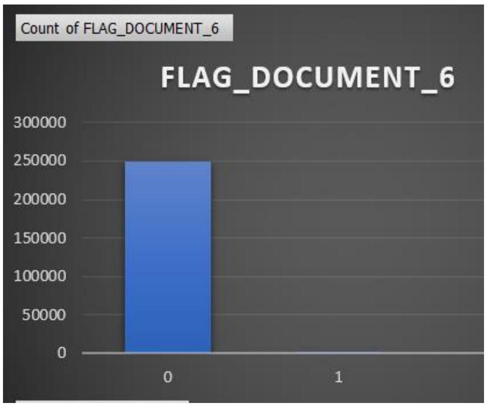
Following this, I created graphs in several "FLAG_DOCUMENT" columns to determine whether or not each one was pertinent.

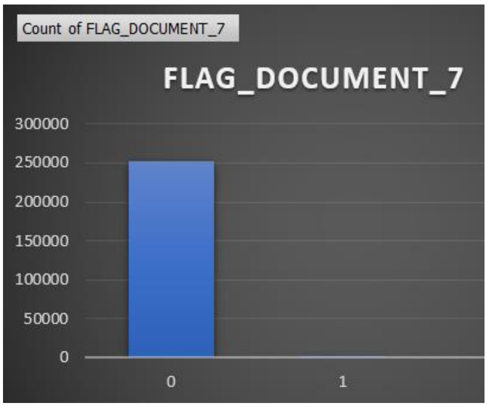


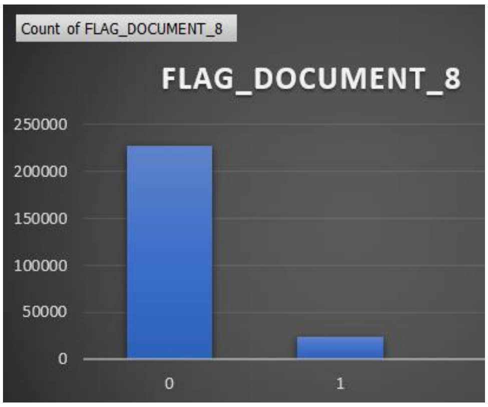


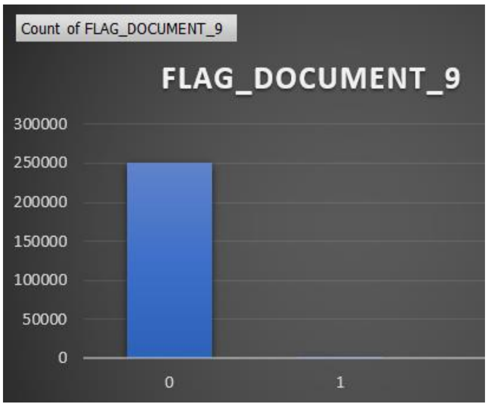


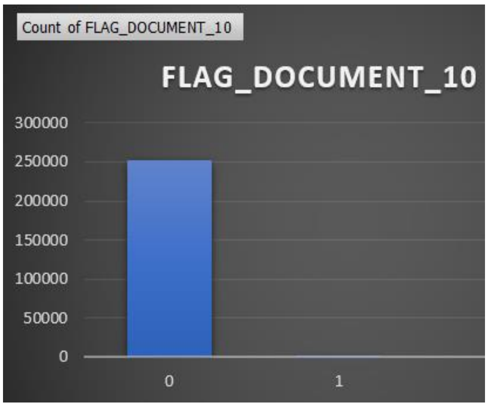


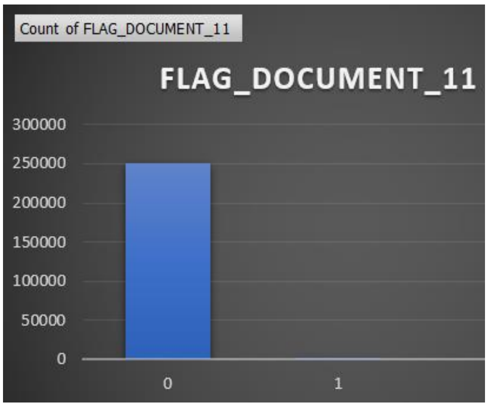


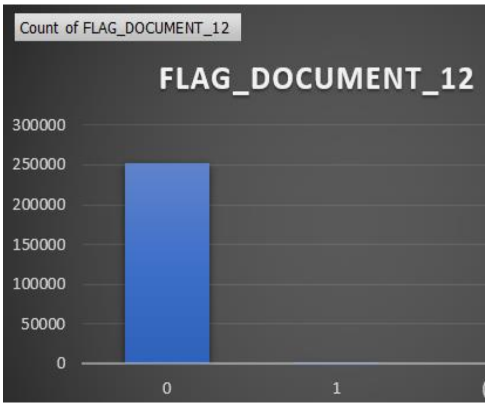


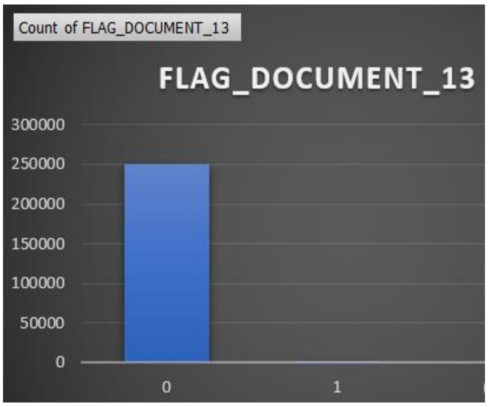

Count of FLAG_DOCUMENT_8
FLAG_DOCUMENT_8
250000
200000
150000
100000
50000
0
0
1

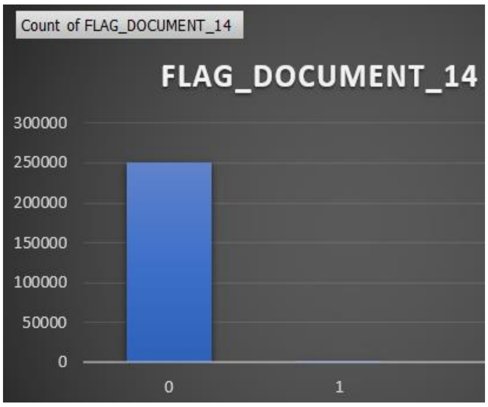

Count of FLAG_DOCUMENT_9
FLAG_DOCUMENT_9
300000
250000
200000
150000
100000
50000
0
0
1

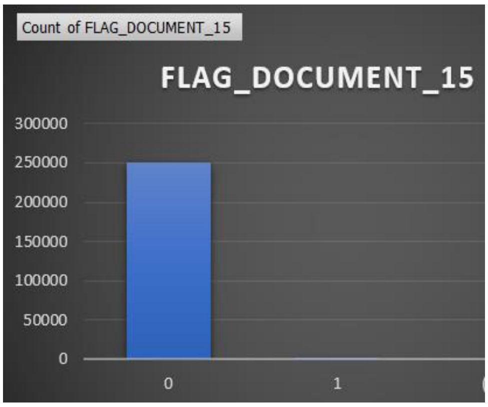

Count of FLAG_DOCUMENT_10
FLAG_DOCUMENT_10
300000
250000
200000
150000
100000
50000
0
0
1

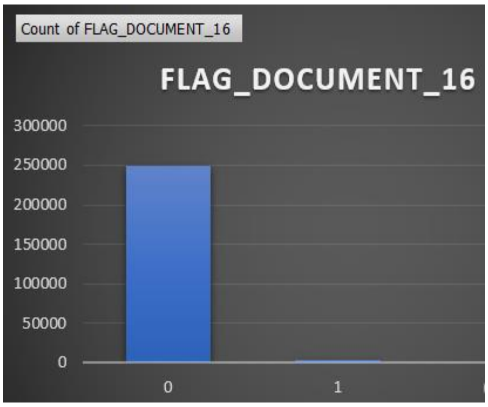

Count of FLAG_DOCUMENT_11
FLAG_DOCUMENT_11
300000
250000
200000
150000
100000
50000
0
0
1

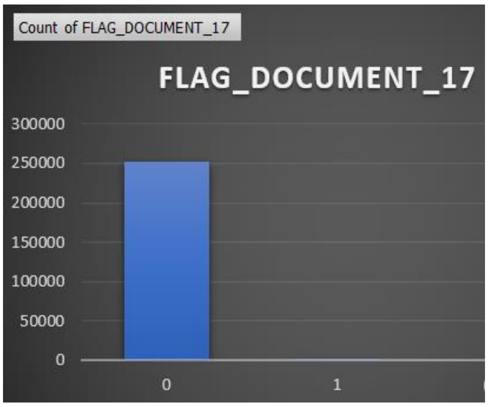

Count of FLAG_DOCUMENT_12
FLAG_DOCUMENT_12
300000
250000
200000
150000
100000
50000
0
0
1

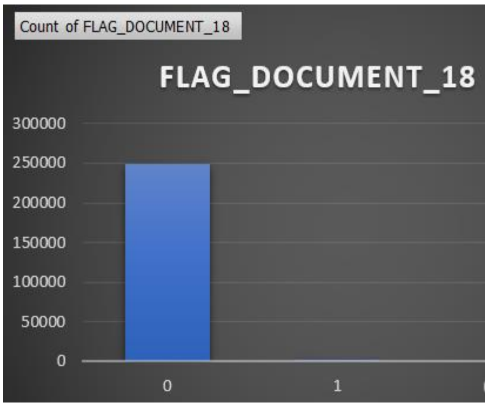

Count of FLAG_DOCUMENT_13
FLAG_DOCUMENT_13
300000
250000
200000
150000
100000
50000
0
0
1

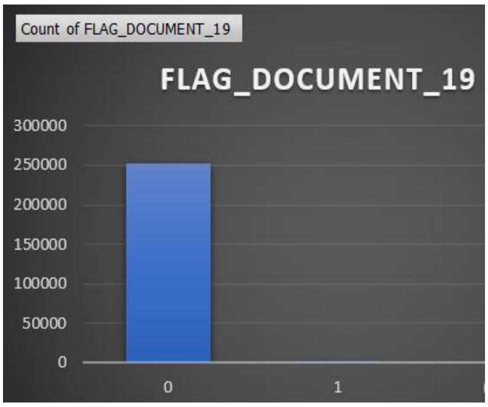

Count of FLAG_DOCUMENT_14
FLAG_DOCUMENT_14
300000
250000
200000
150000
100000
50000
0
0
1

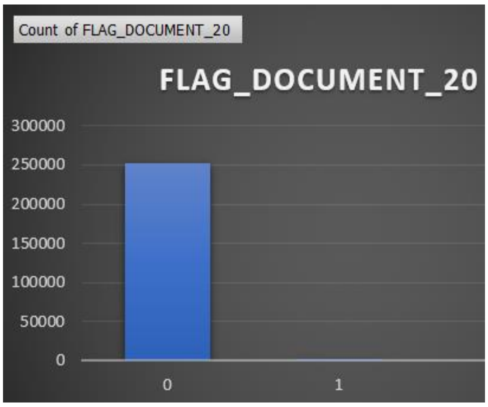

Count of FLAG_DOCUMENT_15
FLAG_DOCUMENT_15
300000
250000
200000
150000
100000
50000
0
0
1

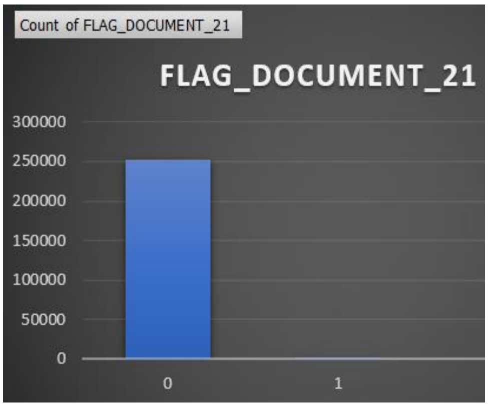












The majority of these graphs, with the exception of FLAG_DOCUMENT_3, have a minimal number of 1s, as can be seen. I therefore eliminated all columns aside from this one.

Creating Loan Credit Amount Group

Now for "AMT_CREDIT" column, I created a "LOAN_LEVEL" column as low, medium and high categorization of loan.

H	I	J	K	L
AMT_INCOMI	AMT_CREDIT		AMT_ANN	AMT_GOC
103500	481495.5		36130.5	454500
90000	165024		8154	108000
Median	521280	**MEDIAN(I2:I251788)**		
3rd Quartile	829224	**QUARTILE.EXC(I2:I251788,3)**		

I found the median and 3rd quartile of the AMT_CREDIT column to get an idea about the level.

```
fx  =IF(I2>900000,"High",IF(I2>500000,"Medium","Low"))
```

Using this formula, levels are created as shown below.

I	J
AMT_CREDIT	**LOAN_LEVEL**
157500	Low
622413	Medium
733315.5	Medium
648000	Medium
1054773	High
450000	Low
1772352	High
1105632	High
272520	Low
450000	Low

Creating Income Group

The same I done with "AMT_INCOME_TOTAL" column and created a column "INCOME_GROUP".

```
fx  =IF(H2>250000,"High",IF(H2>100000,"Medium","Low"))
```

H	I
AMT_INCOME_TOTAL	INCOME_GROUP
112500	Medium
180000	Medium
166500	Medium
90000	Low
216000	Medium
292500	High
360000	High
540000	High
112500	Medium
234000	Medium

Now for better analysis, I created a new column “AVG_EXT_SOURCE” as average score of “EXT_SOURCE_2” & “EXT_SOURCE_3” and deleted these columns.

AC	AD	AE	AF
EXT_SOURCE_2	EXT_SOURCE_3	AVG_EXT_SOURCE	FLAG_DOC A
0.247163892	0.050631513	=ROUND(AVERAGE(AC2,AD2),2)	
0.487305014	0.321735282	0.4	1
0.524157672	0.262248971	0.39	1
0.46067787	0.743559314	0.6	1

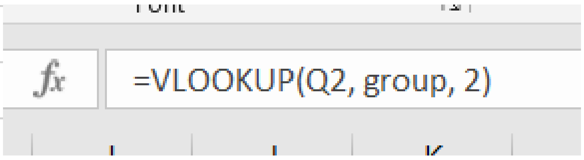
And categorized this column in three levels: low, medium and high.

AC	AD	AE	AF	AG
AVG_EXT_SOURCE	EXT_SOURCE_CATE	FLAG_DOC	AMT_REQ_CREDIT_	
0.15	=IF(AC2>0.6,"High",IF(AC2>0.4,"Medium","Low"))			
0.4	Low	1	1	
0.39	Low	1	0	
0.6	Medium	1	0	
0.64	High	1	0	
0.59	Medium	1	0	
0.57	Medium	1	0	

Creating Age Groups

After that, I used VLOOKUP on a range of ages in the age column to generate the age group.

Age Group	Name
Age<31	Young
30<Age<56	Mid Age
Age>55	Old

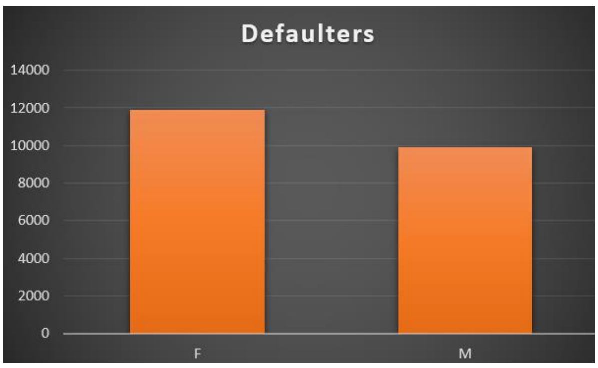


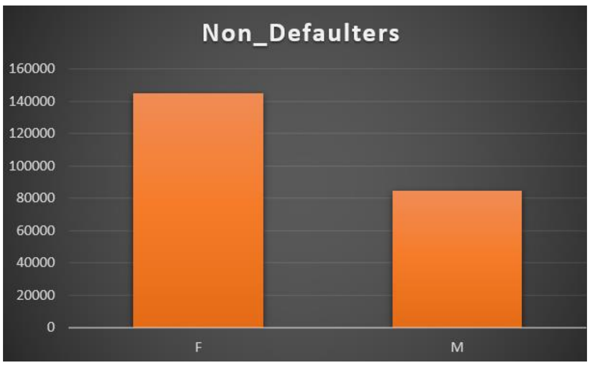
AGE	AGE_GROUP
32	Mid Age
52	Mid Age
34	Mid Age
39	Mid Age
45	Mid Age
31	Mid Age
63	Old
44	Mid Age
36	Mid Age
26	Young
31	Mid Age
32	Mid Age
27	Young

Univariate Analysis

Count of defaulters (Target = 1) & non-defaulters (Target = 0)

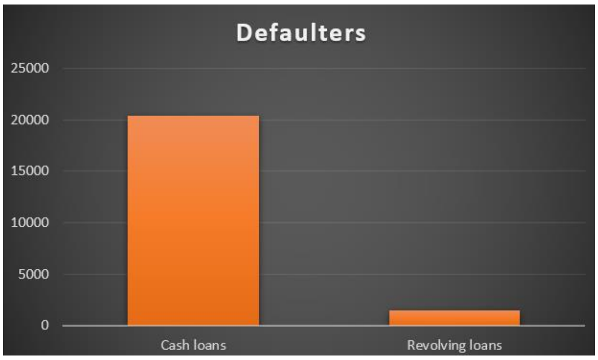
1. On the basis of gender

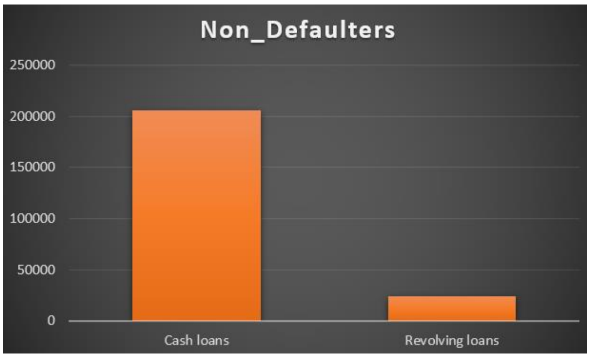




From these graphs, we can see Females are in higher number than males for both defaulter and non-defaulter.

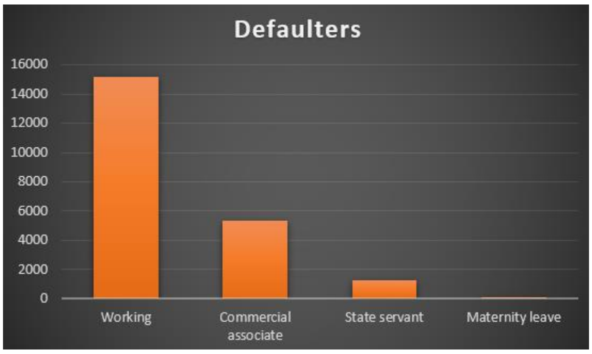
2. On the basis of Loan type

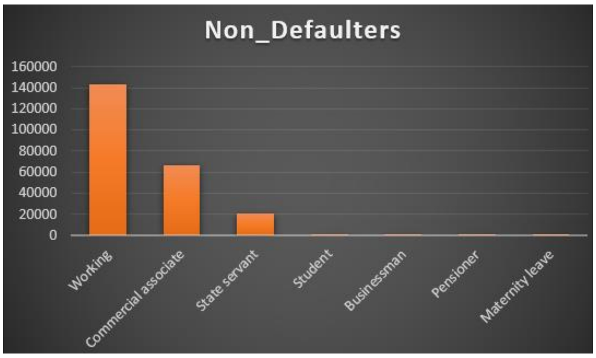




From these graphs, we can see Cash loans are in higher number than revolving loans for both defaulter and non-defaulter.

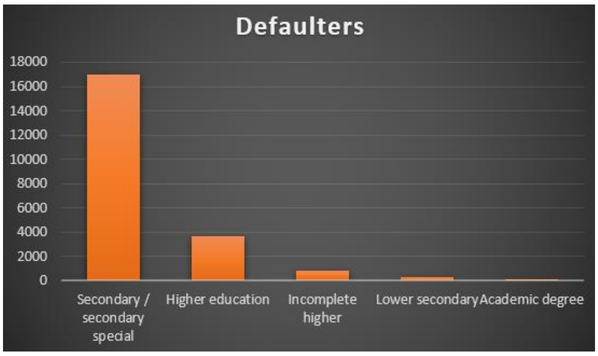
3. On the basis of income type

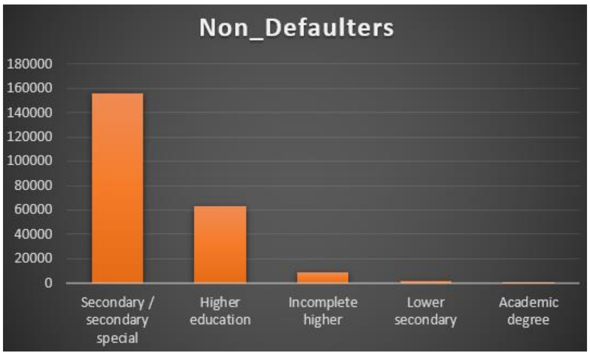




From these graphs, we can see working professionals are in higher number than other professions for both defaulter and non-defaulter.

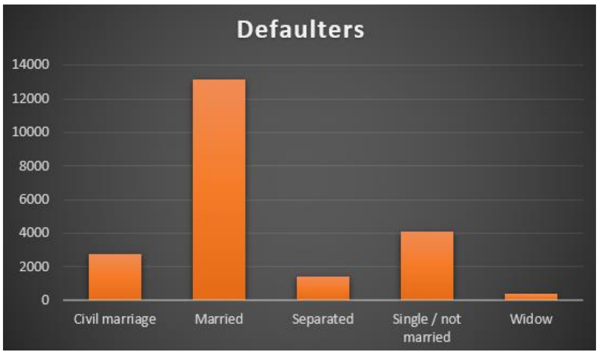
4. On the basis of education type

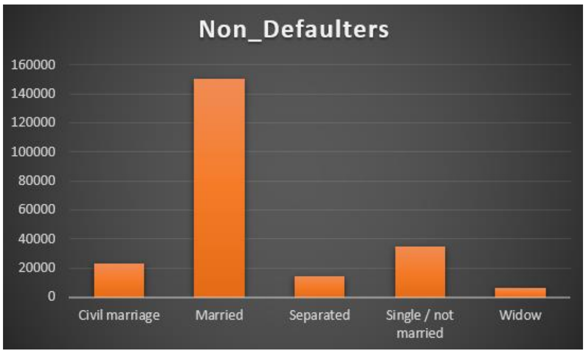




From these graphs, we can see people with secondary/ secondary special education are in higher number than other education background for both defaulter and non-defaulter.

5. On the basis of Family status



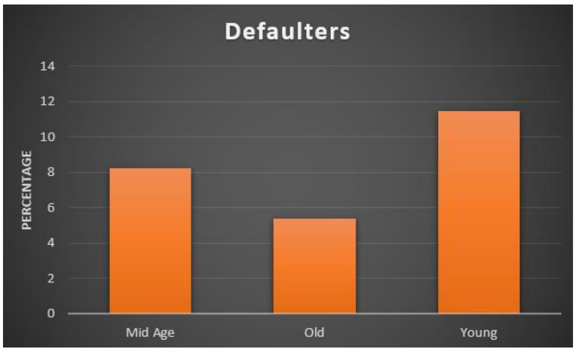


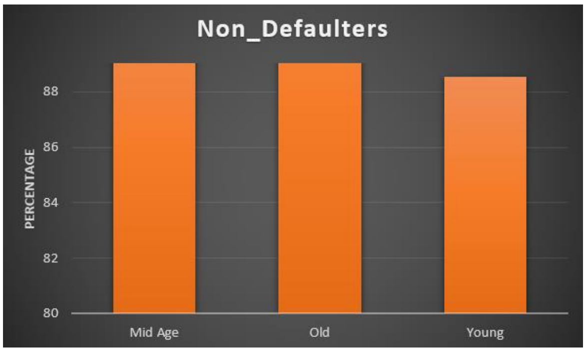
From these graphs, we can see married people are in higher number than others for both defaulter and non-defaulter.

Segmented Univariate Analysis

1. According to Age Group

Now I calculated the percentage of each age group for defaulters and non-defaulters.

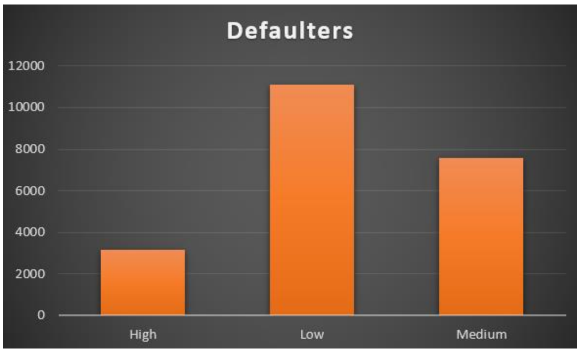


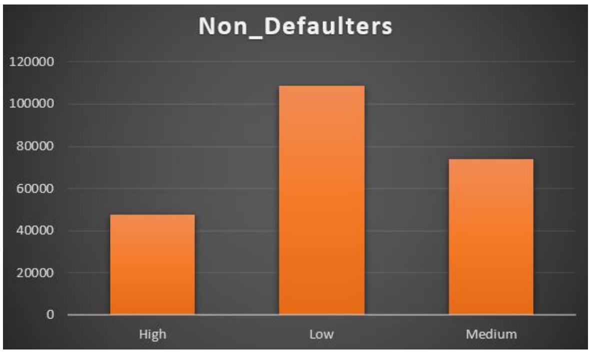


From these graphs, we can see that young people are more likely to default than other age groups. But in case of non – defaulters there is not much significant difference among all the age groups.

2. According to Loan Level

Now I calculated the count of people for each loan amount level for default and non-default.

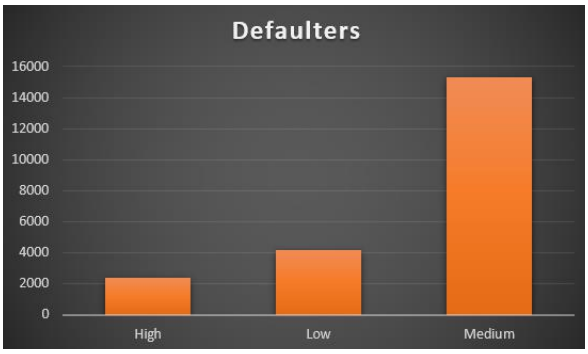


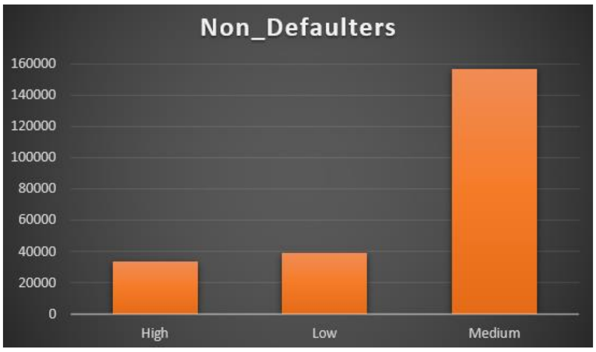


From these graphs, we can see that most defaulters are from low and medium loan credit group. And low amount loan credit groups are more in non-defaulters.

3. According to Income Group

Now I calculated the count of defaulters and non-defaulters for each income group.

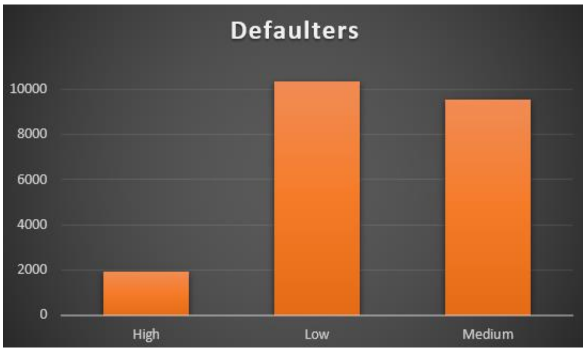


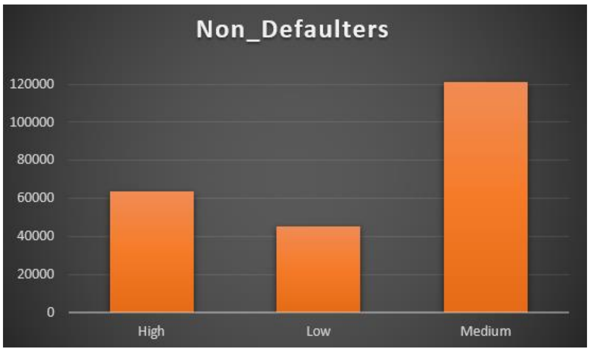


From these graphs, we can see that most defaulters are from medium income group. And same in non-defaulters, there are high number of medium income group than others.

4. According to External Source Category

Now I calculated the count of defaulters and non-defaulters for each external source category.





From these graphs, we can see that most defaulters are from low and medium external source category. And in non-defaulters, there are high number of medium external source category than others.

Bivariate analysis

Correlation

Correlation of numerical data for following columns for both defaulters and non-defaulters: ['AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY','AMT_GOODS_PRICE','AGE','EXT_SOURCE_SCORE','REGION_RATING_CLIENT']

➢ Correlation of defaulters:

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	AGE	REGION_RATING_CLIENT	AVG_EXT_SOURCE
AMT_INCOME_TOTAL	1						
AMT_CREDIT	0.362210842	1					
AMT_ANNUITY	0.433585961	0.76203507	1				
AMT_GOODS_PRICE	0.369828962	0.986352451	0.766359904	1			
AGE	0.056057041	0.157835287	0.092754031	0.152990103	1		
REGION_RATING_CLIENT	-0.207858142	-0.105133839	-0.129051223	-0.107231888	-0.042166115	1	
AVG_EXT_SOURCE	0.079255319	0.133883493	0.11685045	0.141866396	0.207387088	-0.214776389	1

Highly correlated columns for defaulters:

1. 0.76 - AMT_ANNUITY & AMT_CREDIT
2. 0.99 – AMT_GOODS_PRICE & AMT_CREDIT

3. 0.77 – AMT_GOODS_PRICE & AMT_ANNUITY

➢ Correlation of non - defaulters:

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	AGE	REGION_RATING_CLIENT	AVG_EXT_SOURCE
AMT_INCOME_TOTAL	1						
AMT_CREDIT	0.362244196	1					
AMT_ANNUITY	0.433627822	0.762049893	1				
AMT_GOODS_PRICE	0.369863652	0.98635344	0.766377555	1			
AGE	0.056077489	0.157860166	0.092769348	0.153017639	1		
REGION_RATING_CLIENT	-0.207855689	-0.105141048	-0.12904239	-0.107236929	-0.042180576	1	
AVG_EXT_SOURCE	0.079273122	0.13391621	0.116879593	0.141895288	0.207391986	-0.214774547	1

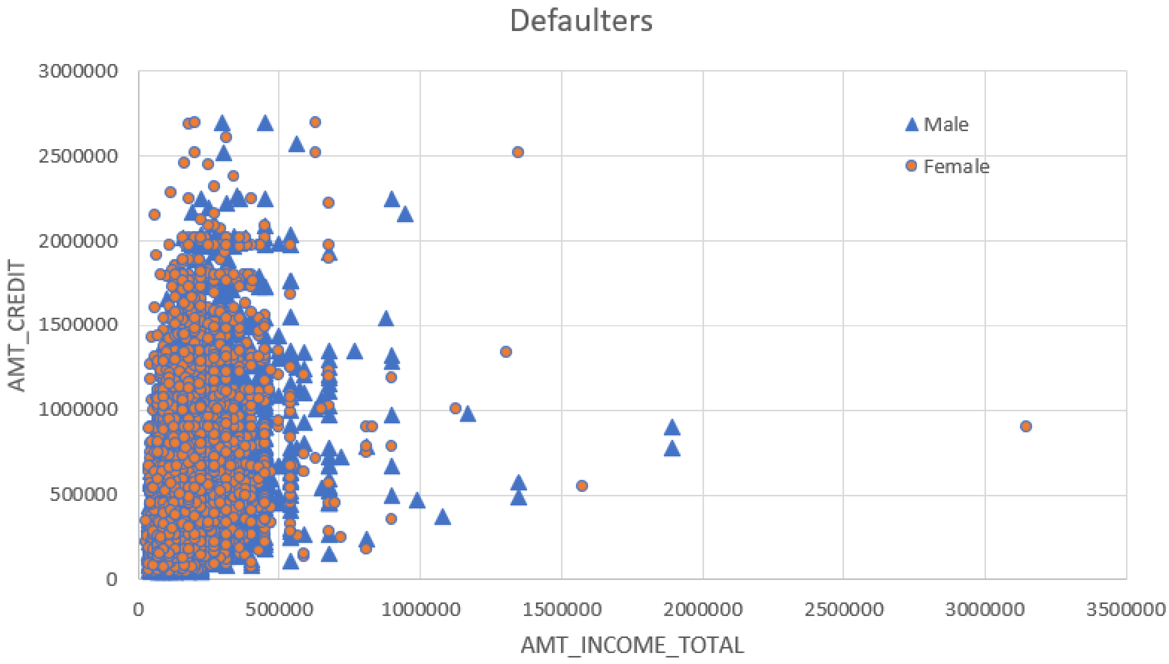
Highly correlated columns for non - defaulters:

1. 0.76 - AMT_ANNUITY & AMT_CREDIT
2. 0.99 – AMT_GOODS_PRICE & AMT_CREDIT
3. 0.77 – AMT_GOODS_PRICE & AMT_ANNUITY

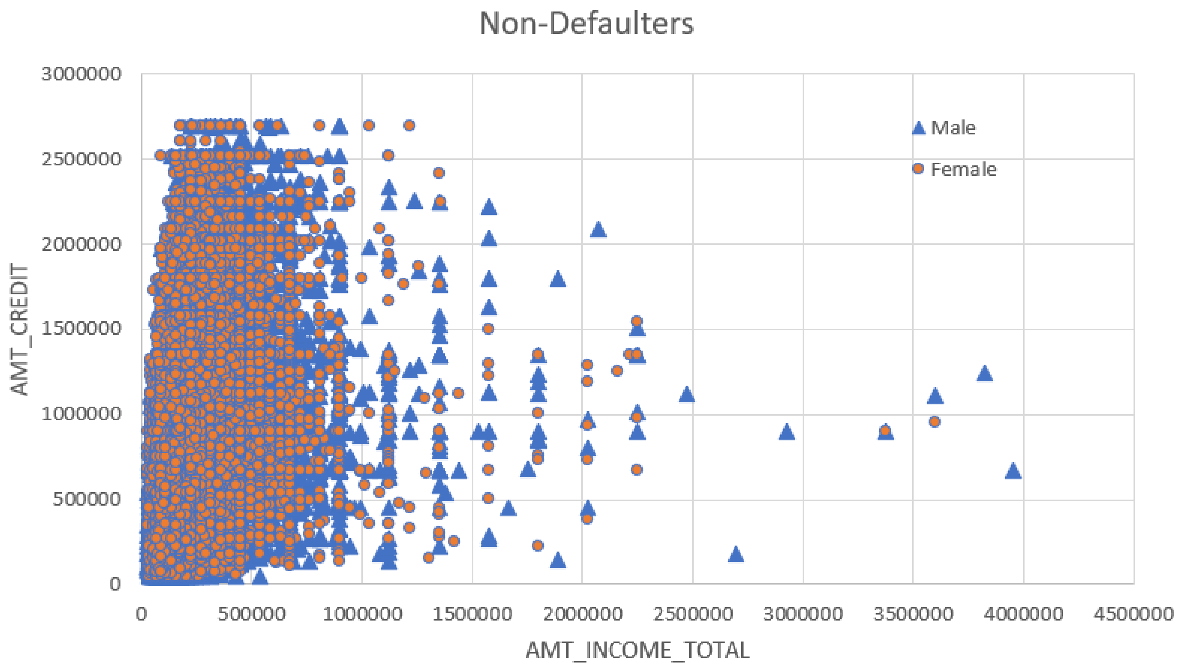
So, there is correlation between same pairs in both defaulter and non – defaulters.

Continuous Variables

Bivariate analysis of credit amount of loan & total income based on gender for both defaulters and non-defaulters:



We can observe that the majority of the data points are clustered on the lower income and lower loan amount side of the scatter plot between Income & Loan Amount for Defaulters. Additionally, the loan amount is rising for both men and women as income rises.



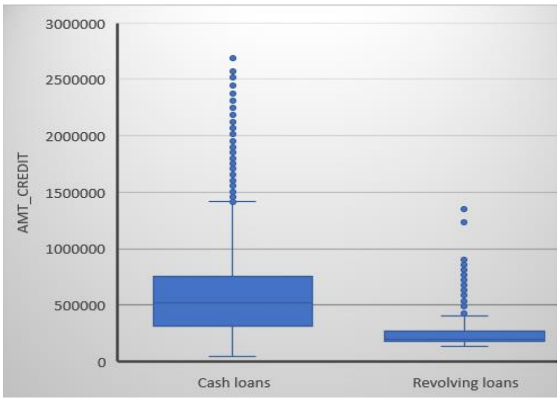
We can observe that the majority of the data points are centred on the side of lower income and lower loan amount from this scatter plot between Income & Loan Amount for Non-Defaulters. Additionally, the loan amount is rising for both men and women as income rises.

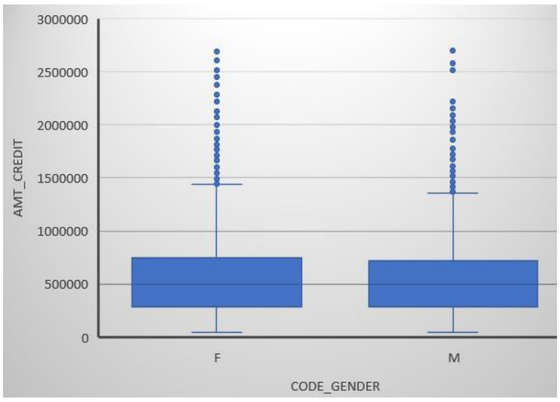
Categorical Variables

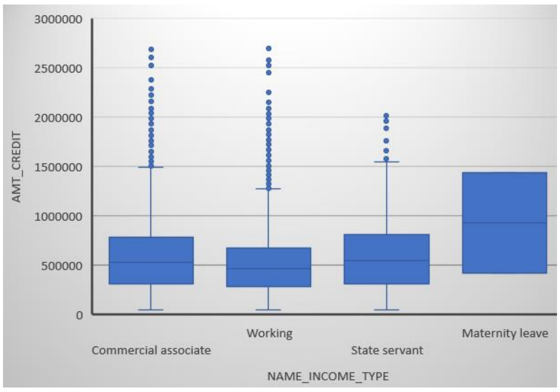
Bivariate analysis of credit amount of loan with categorical variables for both defaulters and non – defaulters:

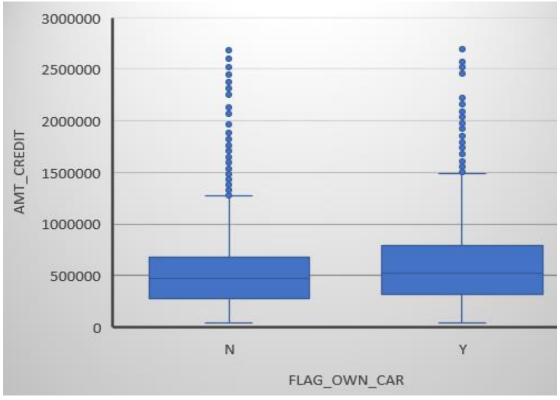
Categorical variables that I took for this analysis are: ['NAME_CONTRACT_TYPE','CODE_GENDER','FLAG_OWN_CAR','FLAG_OWN_REALTY','NAME_INCOME_TYPE','NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS','NAME_HOUSING_TYPE','AGE_GROUP','INCOME_GROUP','EXT_SOURCE_CATEGORY']

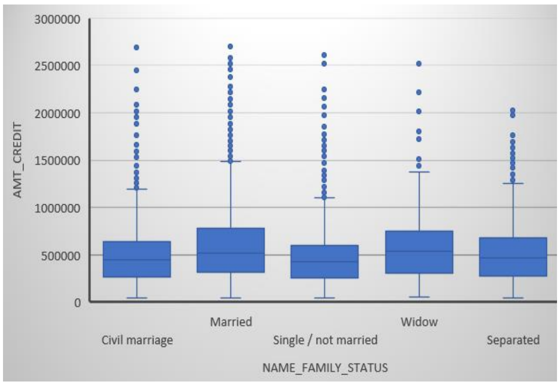
1. Defaulters

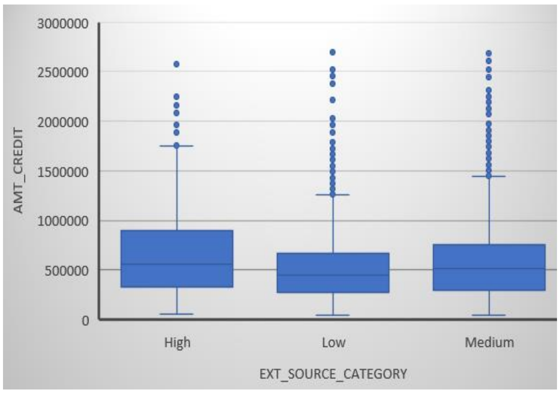


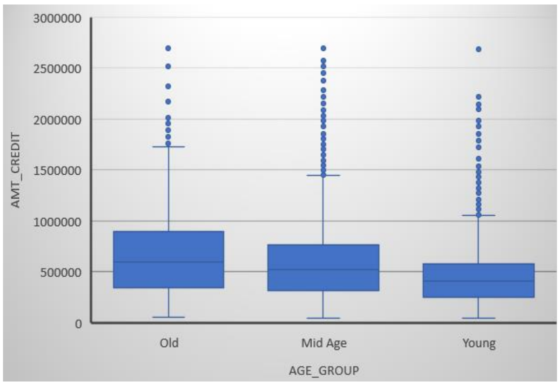


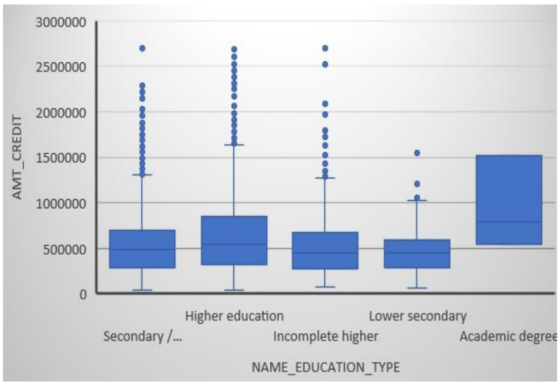

AMT_CREDIT
Commercial associate
Working
State servant
Maternity leave
NAME_INCOME_TYPE

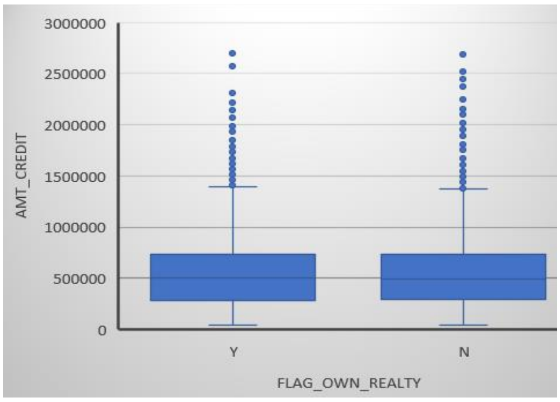

AMT_CREDIT
N
Y
FLAG_OWN_CAR

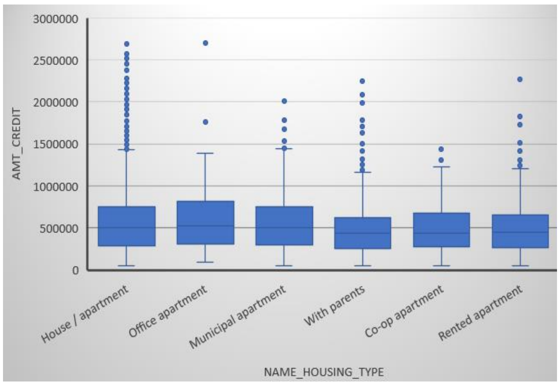

AMT_CREDIT
Civil marriage
Married
Single / not married
Widow
Separated
NAME_FAMILY_STATUS

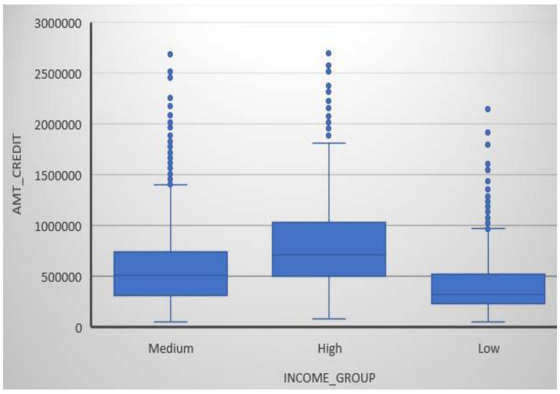

AMT_CREDIT
High
Low
Medium
EXT_SOURCE_CATEGORY


AMT_CREDIT
Old
Mid Age
Young
AGE_GROUP


AMT_CREDIT
Secondary /...
Higher education
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE


AMT_CREDIT
Y
N
FLAG_OWN_REALTY

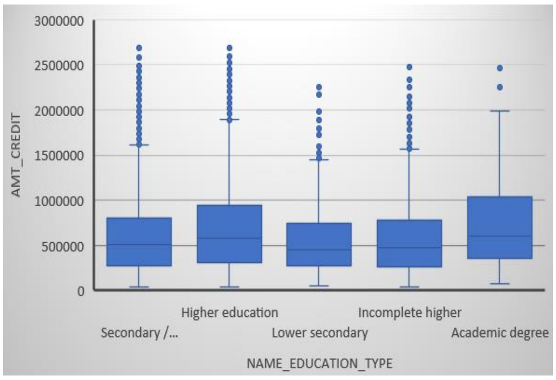

AMT_CREDIT
House / apartment
Office apartment
Municipal apartment
With parents
Co-op apartment
Rented apartment
NAME_HOUSING_TYPE

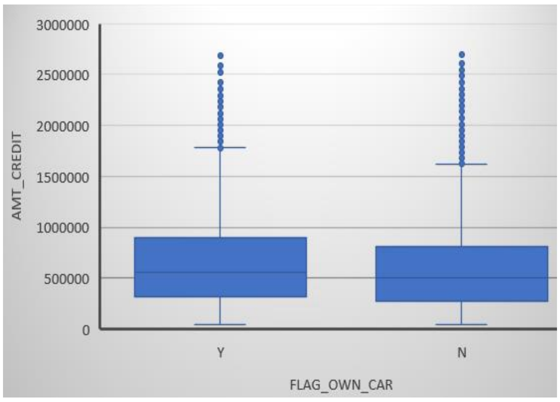


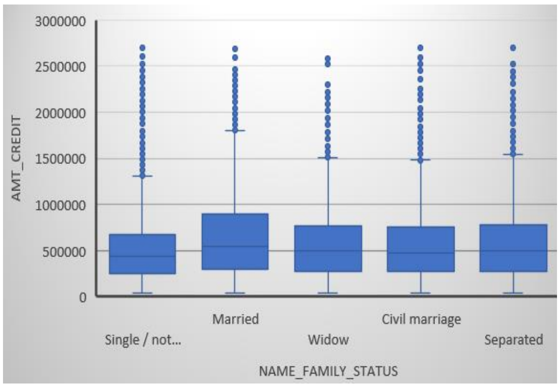
These box plots for defaulters led me to draw the following conclusions:

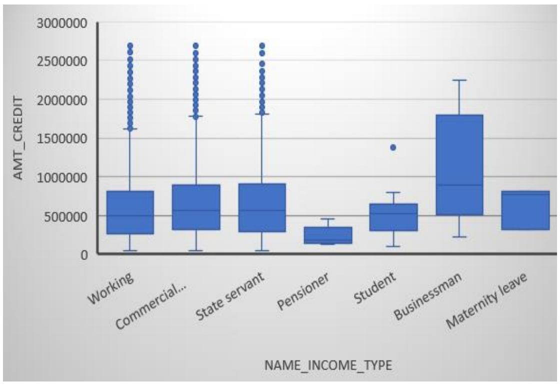
1. Revolving loans have lower credit limits than cash loans.
2. Income from maternity leave received more credit than income from other sources.
3. Credit amounts were larger for people with high EXT_SOURCE_CATEGORY.
4. Compared to other age groups, young people received fewer loans.
5. Those with academic degrees received loans in higher amounts.
6. The credit amounts for the categories of car and real estate owners are not significantly different.
7. People with greater incomes received loans with higher credit limits.
8. Compared to other groups, single people received less loan credit.

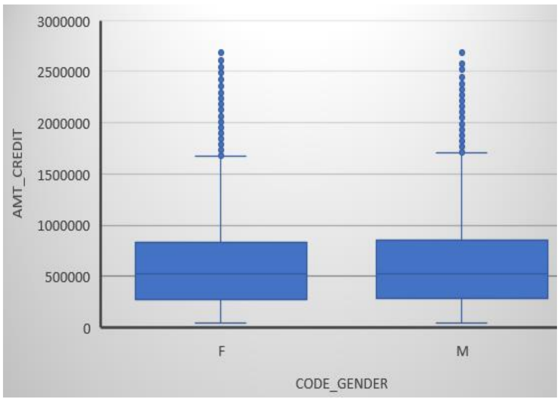
2. Non_Defaulters

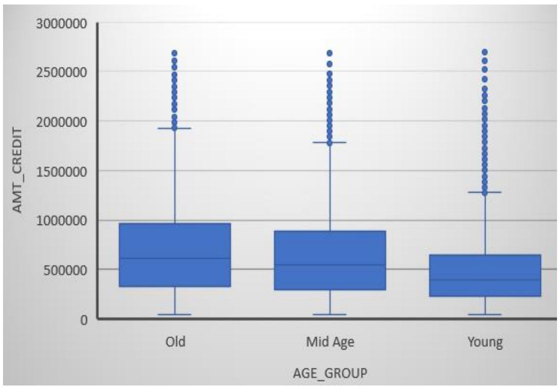


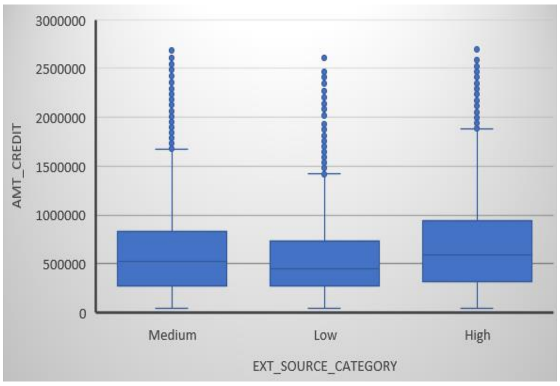


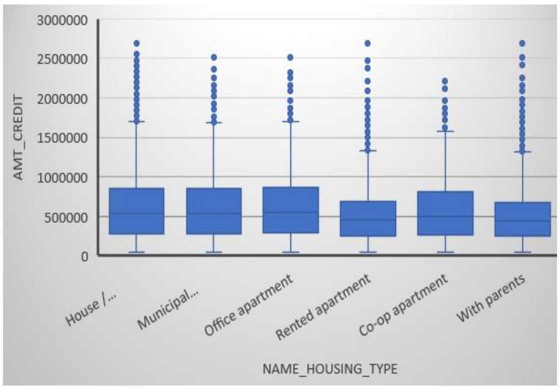


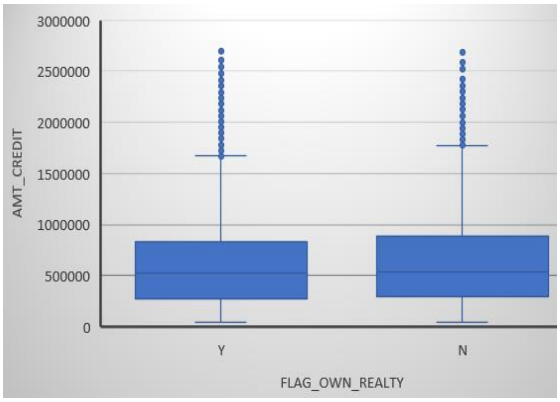


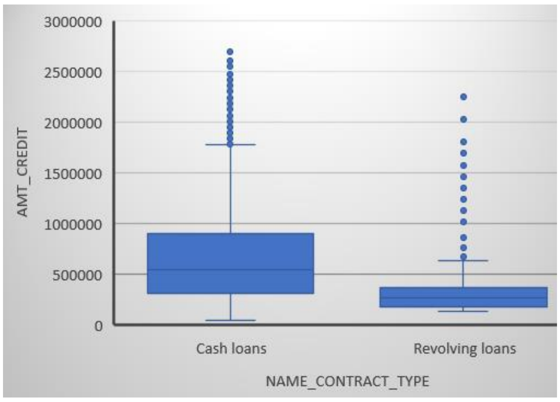


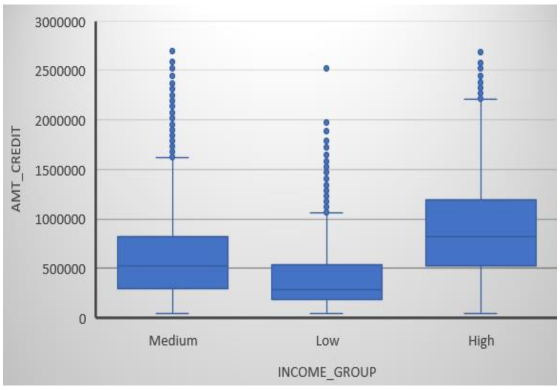










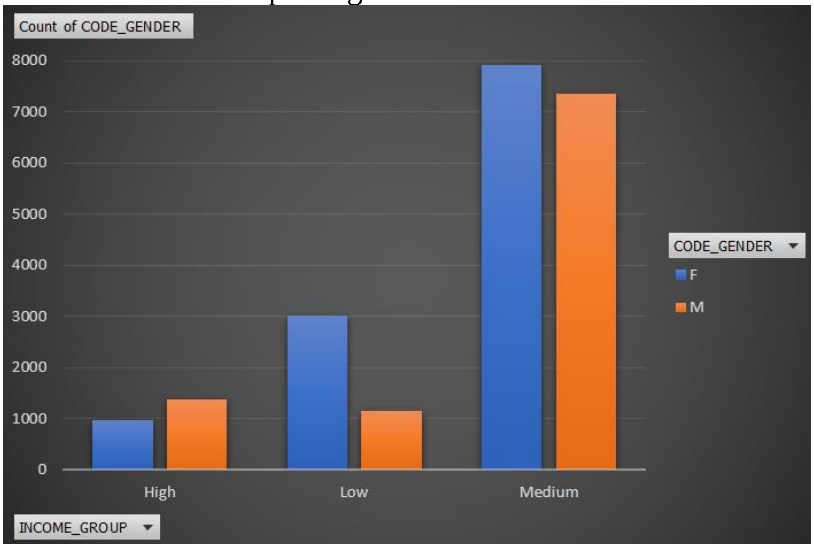


These box plots for non-defaulters led me to the following conclusions:

1. Revolving loans have lower credit limits than cash loans.
2. Businessmen received better credit than other types of income.
3. Credit amounts were larger for people with high EXT_SOURCE_CATEGORY.
4. Compared to other age groups, young people received fewer loans.
5. Those with academic degrees received loans in higher amounts.
6. The credit amounts for the categories of car and real estate owners are not significantly different.
7. People with greater incomes received loans with higher credit limits.
8. Renters and dependent people received less loan credit than other groups of people.

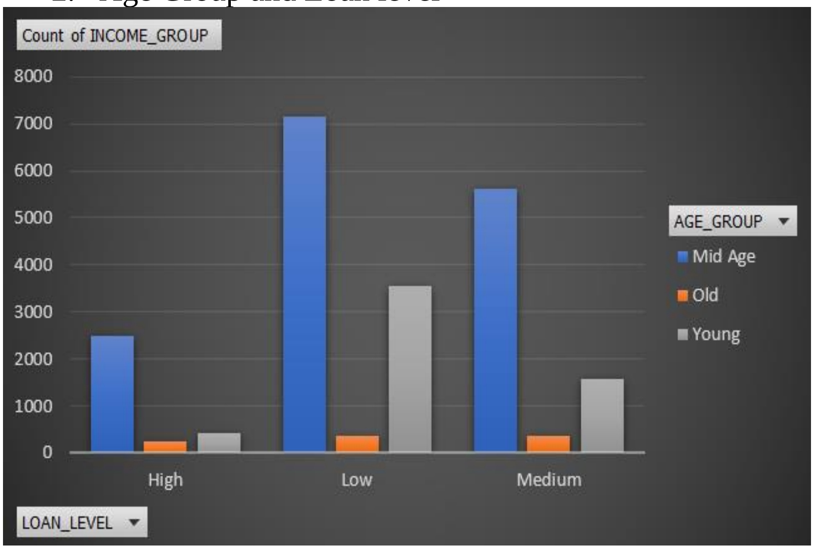
Analysis of defaulters using two segemented variables:

1. Income Group and gender



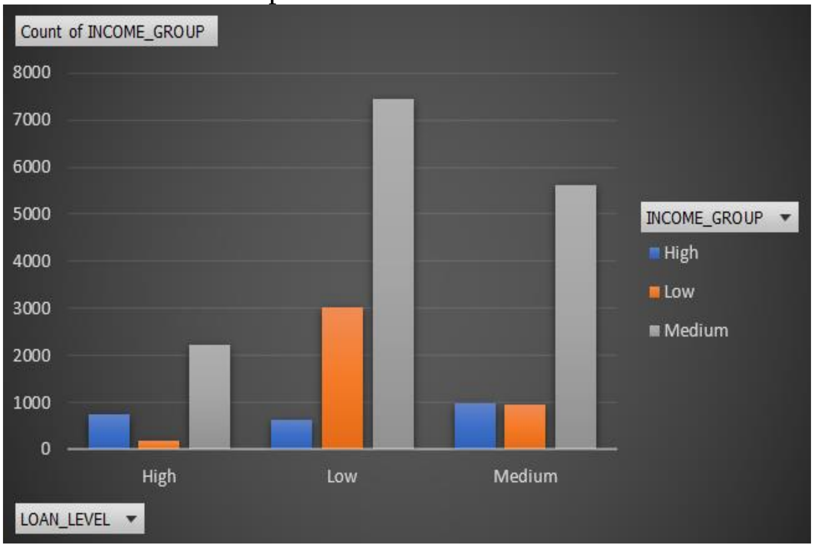
➢ Females are more defaulted than males across low and medium income groups.

2. Age Group and Loan level



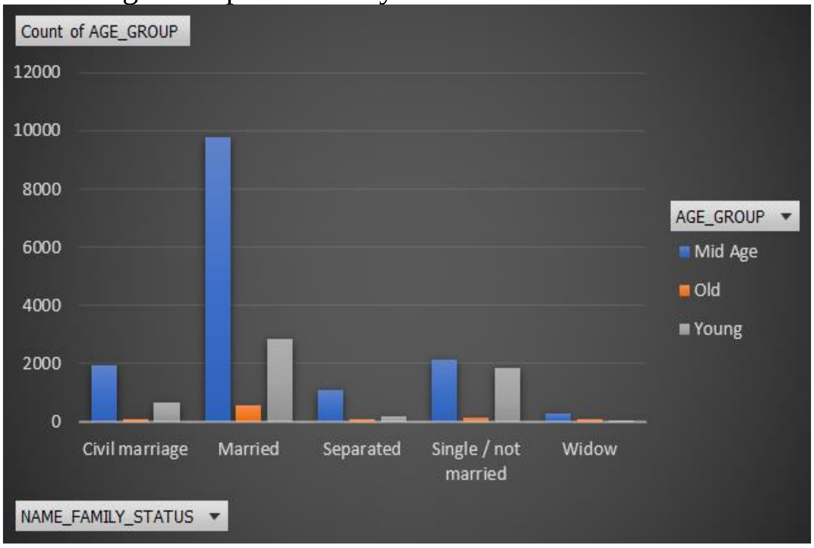
➢ Mid age applicants are more defaulted across all levels of loan credit amount.

3. Income Group and Loan level



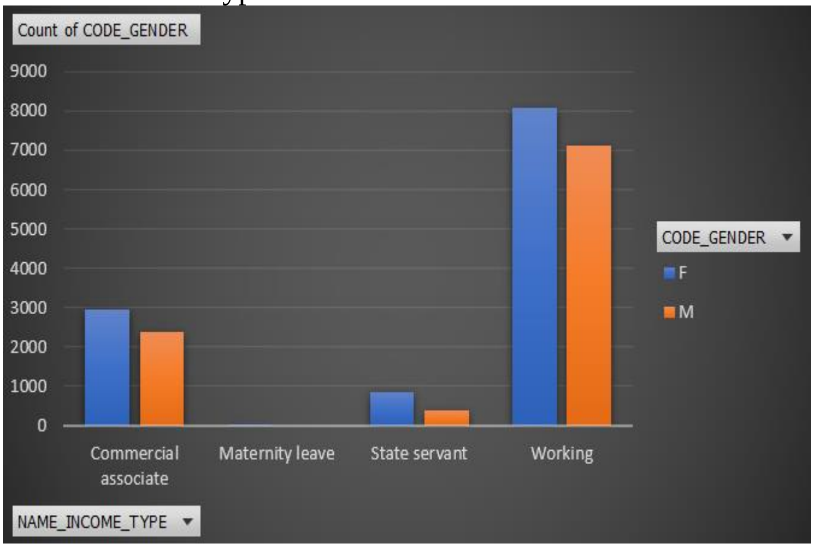
➢ Medium income groups are more defaulted across all loan levels.

4. Age Group and Family status



➢ Mid age people are more defaulted across all family statuses and old people are less.

5. Income Type and Gender

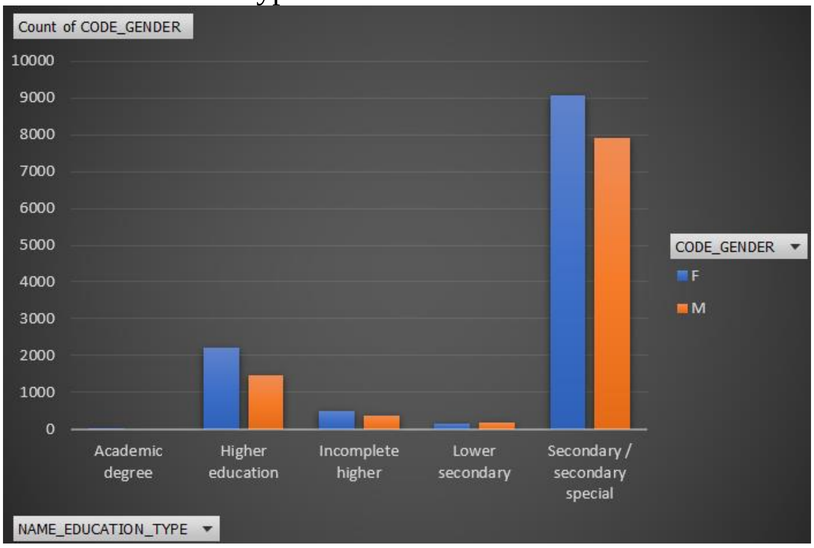


- ➢ Females are more defaulted than males across all income types.

6. Family Status and Gender

- ➢ Across all family statuses, mostly females are more defaulted than males.

7. Education Type and Gender



- ➢ Same for this, females are more defaulted across all education levels.