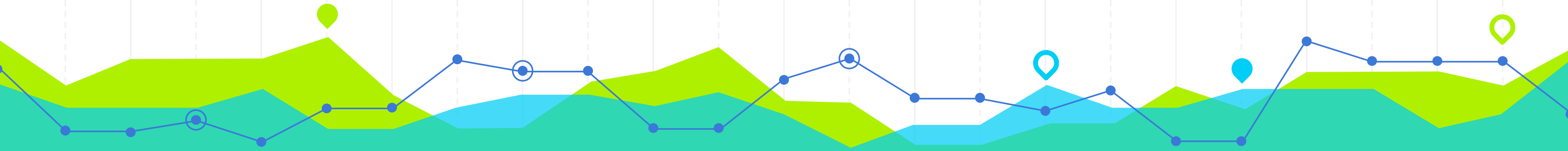
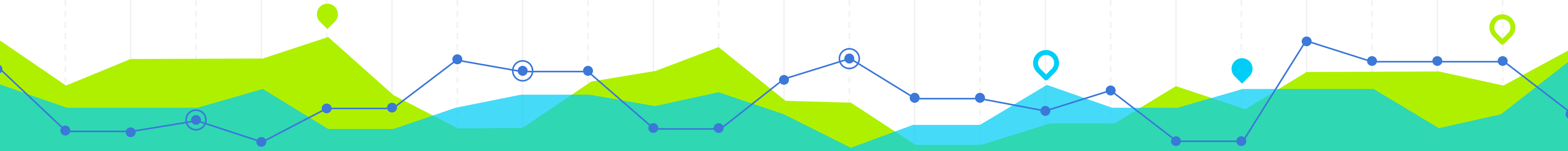


Credit EDA Assignment

Submitted by
Kunal Sadana
DS C40 Batch



Introduction and Problem Statement

Introduction

This case study aims to give us an idea of applying EDA in a real business scenario. In this assignment, we have to apply the techniques of EDA, and 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.

Problem Statement

In this case study, we have to use EDA to understand how consumer attributes and loan attributes influence the tendency of default. We have to use EDA to analyze the patterns present in the data to ensure that the applicants capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

Approach

For Exploratory Data Analysis, the steps followed are:

- Importing necessary Libraries
- Loading data
- Data Inspection: Reading the data
- Data Cleaning: It involves the following:
  - Dealing with Null values
  - Fixing Data Types and removing unnecessary rows or columns
  - Dealing with Outliers
  - Binning the data into groups wherever required
- Univariate Analysis: Analysis of single columns or variable using various Data Visualization Techniques
- Bivariate and Multivariate Analysis: Analysis of multiple columns or variables using various Data Visualization Techniques to find relation between them if present
- Conclusion: Detailing the Insights inferred.

“

Analysis for application_data

Data Cleaning

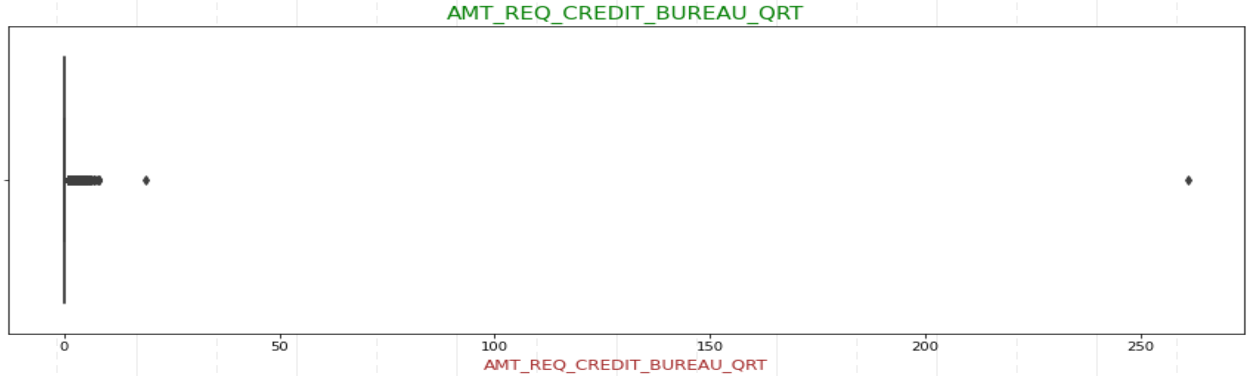
Dealing with Null Values

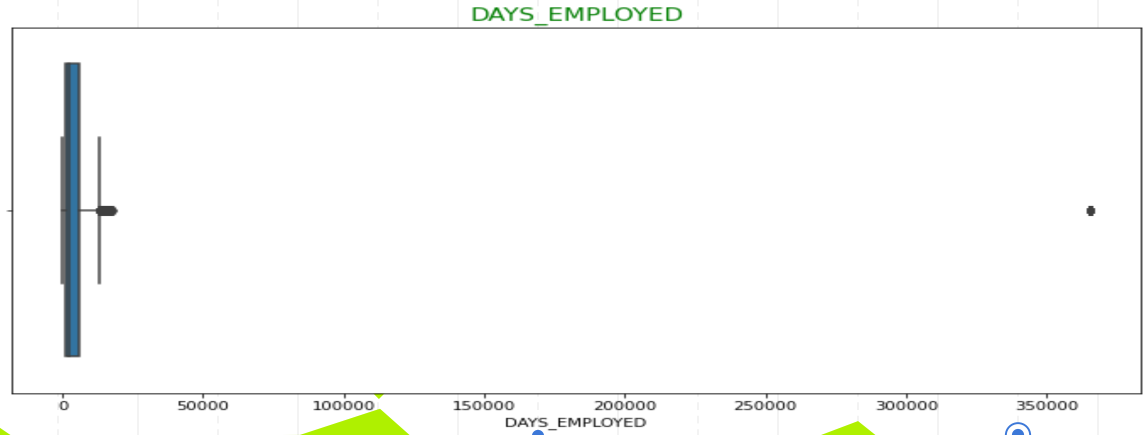
- All the columns having null value percentage more than 45% can be removed.
- For the columns 'AMT_REQ_CREDIT_BUREAU_HOUR', 'AMT_REQ_CREDIT_BUREAU_DAY', 'AMT_REQ_CREDIT_BUREAU_WEEK', and ‘FLAG_DOCUMENT_#’ since 0 occurs more than 95% of time, we can drop the columns as it won't have any major impact on our analysis. (except FLAG_DOCUMENT_3)
- For OCCUPATION_TYPE we can replace null values with 'Unknown' as it is a categorical data type.
- For CODE_GENDER there is 'XNA' value present in the column. Only 4 values are there, so these can be replaced with 'F', which is the mode, as it won't make a big impact.
- For AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT we can replace null values with median

Fixing Data Types

- For few columns, values given were negative, have to convert negative values to positive values.
- FLAG_DOCUMENT_3 shows whether the document submitted or not where 0: not submitted ,and 1: submitted. We can convert the following column into bool type.

Dealing with Outliers



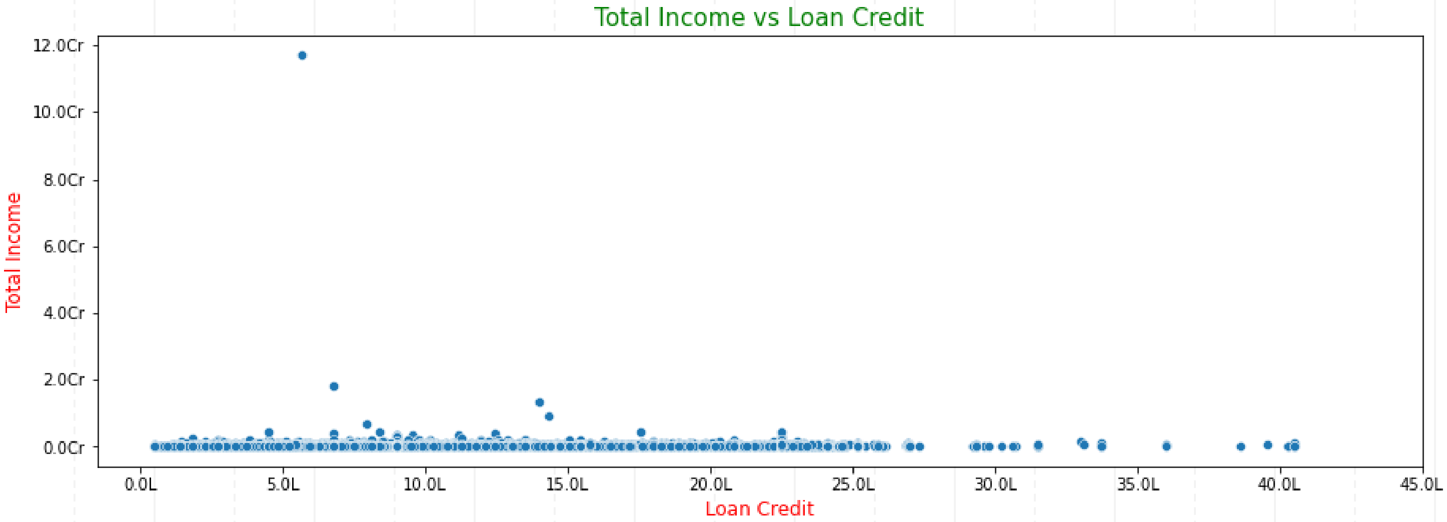


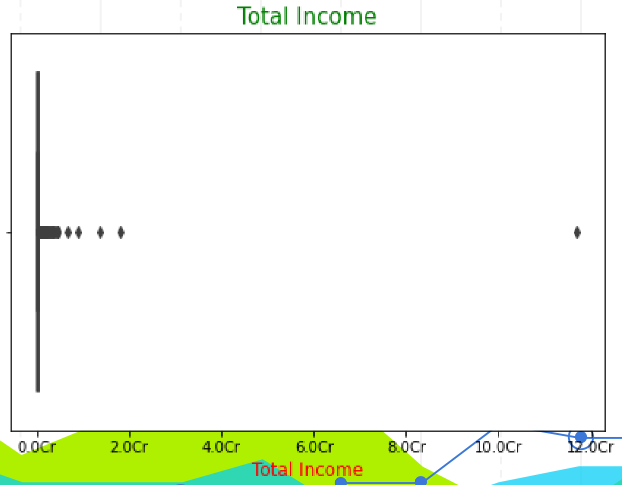
AMT_REQ_CREDIT_BUREAU_QRT

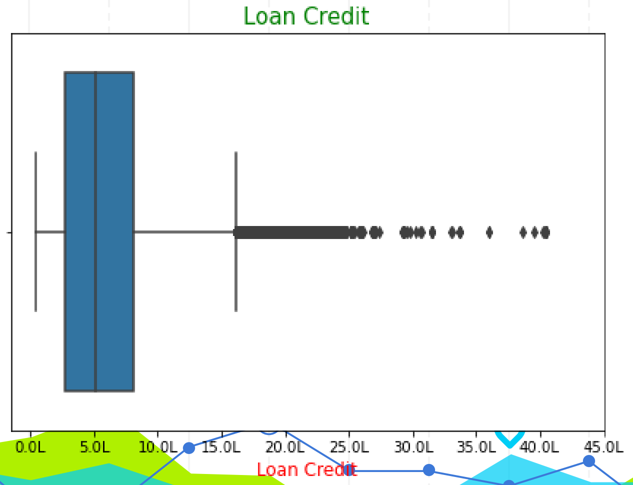
- As can be seen from the box plot, we have on outlier which is much greater than the rest of the values.
- Looking at the description of the variable, this data seems to be a 'wrong data'.
- We can replace this outlier with the median

DAYS_EMPLOYED

- It can be seen here that OUTLIER value is too high and it is clearly an incorrect data as when converted to years, it amounts to 1000 years. Also, 18% of data have this value. We can replace this value with NaN.


Total Income vs Loan Credit


Total Income


Loan Credit

- As can be seen from the boxplot, there is an outlier in *'AMT_INCOME_TOTAL'* whose value is much more than compared to others. As evident from scatterplot (huge difference between Income and Loan value) and person's occupation, this data may be an incorrect data. We can replace this value with the median. Oher outliers can be ignored while dealing with this variable.
- There are few outliers in *'AMT_CREDIT'*. Use median instead of mean while dealing with this variable.

Binning

- *DAYS_BIRTH*

Convert the column into years and divide it

- 20-30
- 30-40
- 40-50
- 50-60
- 60+'

- *AMT_INCOME_TOTAL*

According to quantiles

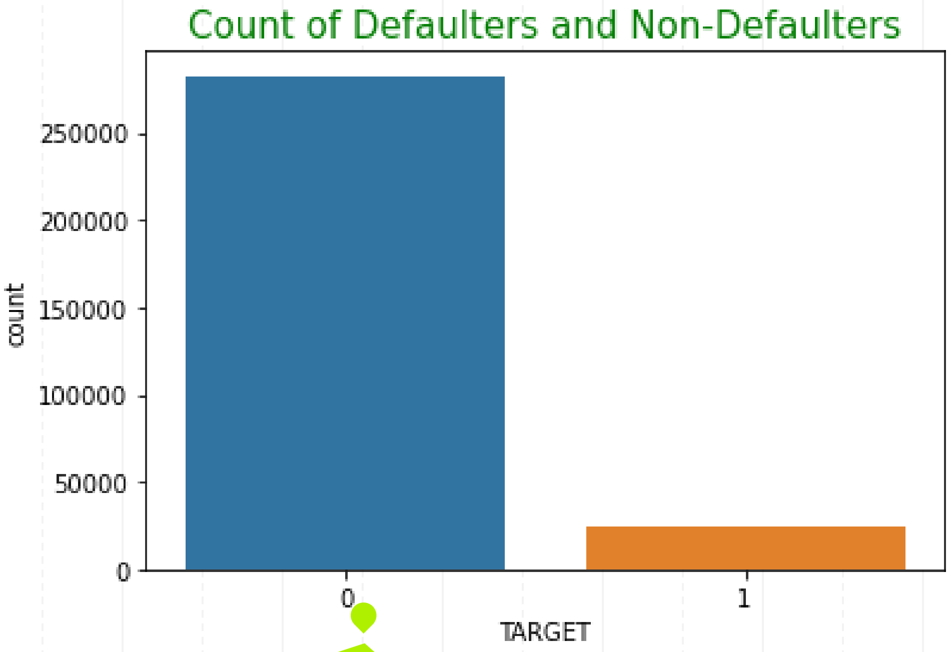
- Very Low (0.0 - 0.2)
- Low (0.2 – 0.4)
- Medium (0.4 – 0.6)
- High (0.6 – 0.8)
- Very High (0.8 – 1.0)

- *CREDIT_GROUP*

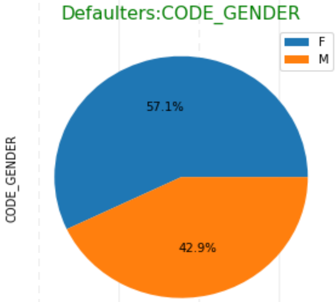
According to quantiles

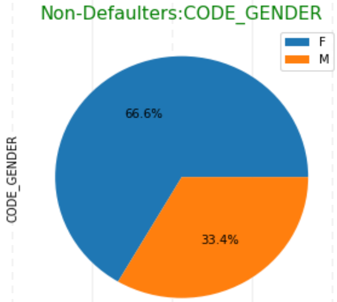
- Very Low (0.0 - 0.2)
- Low (0.2 – 0.4)
- Medium (0.4 – 0.6)
- High (0.6 – 0.8)
- Very High (0.8 – 1.0)

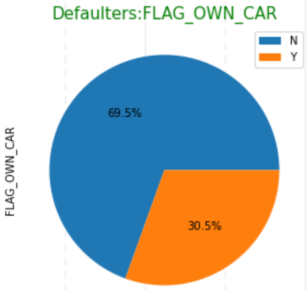
Univariate Analysis

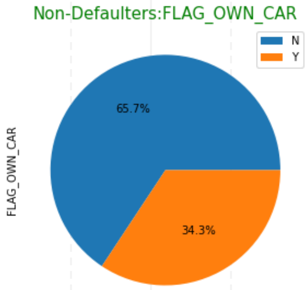


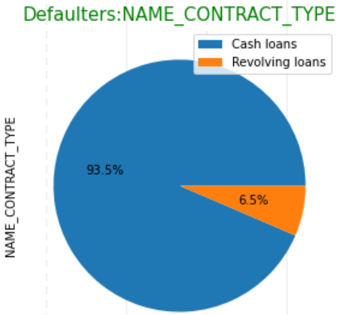
- Majority (91%) of people lie in Non-Defaulters category.
- Shows the imbalance in data
- We thus divide the data set into two parts: Defaulters and Non-defaulters

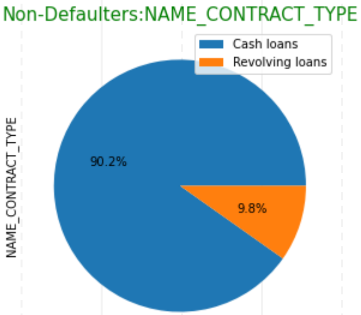


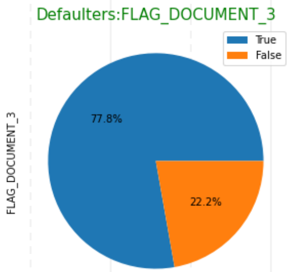


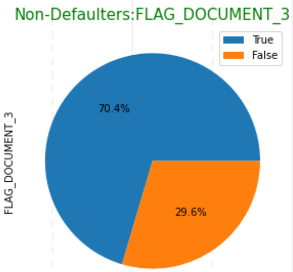


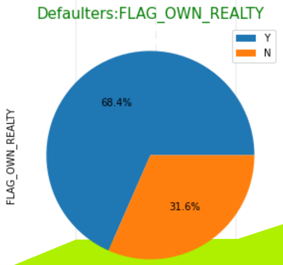


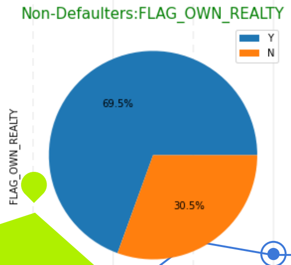






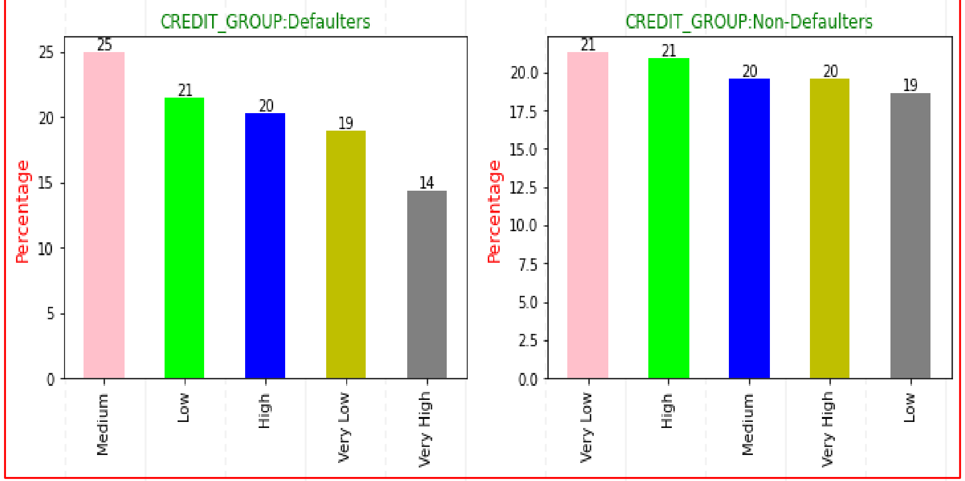


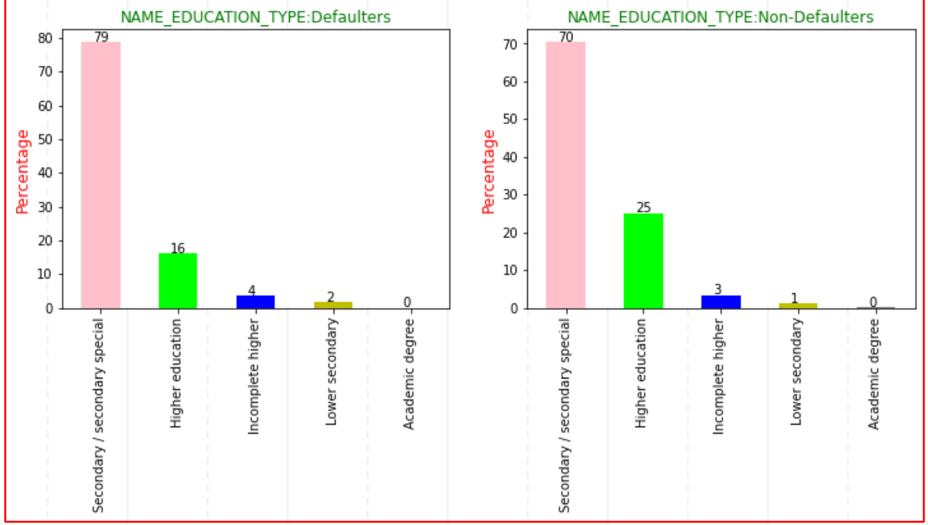


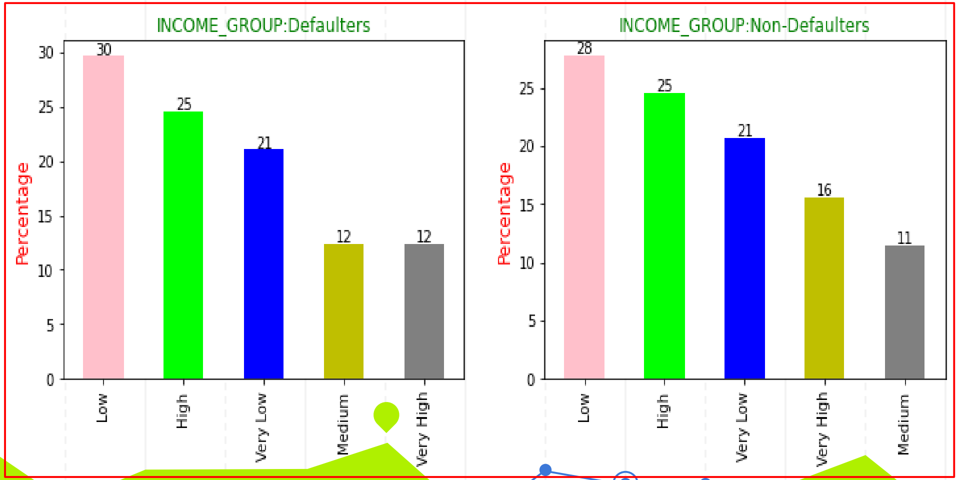


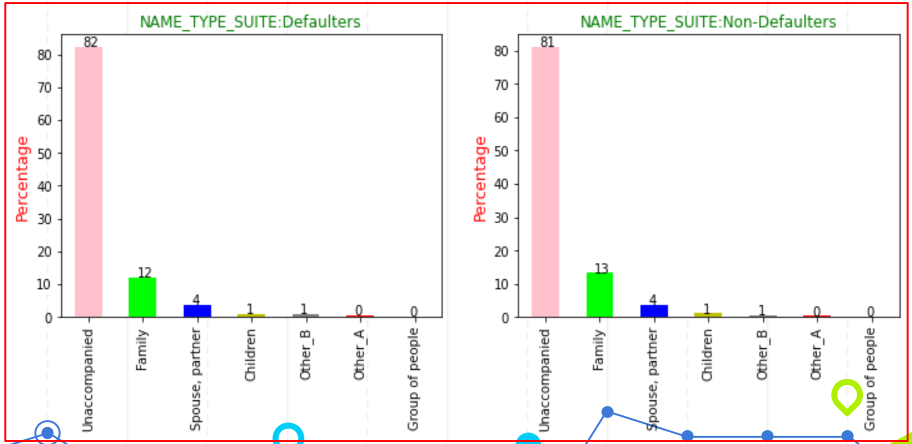
Categorical Variables

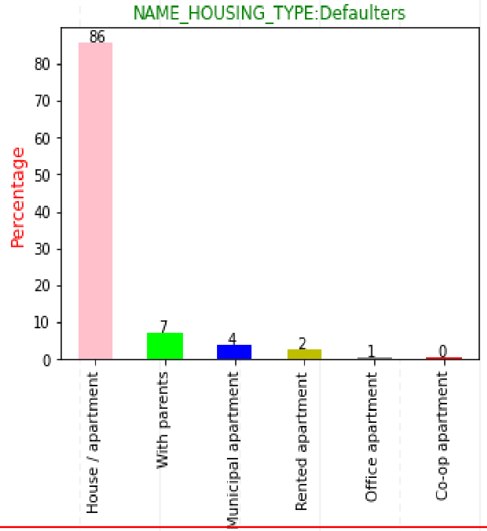
- More females have applied for loans. Female percentage in Non-Defaulters is more than that in Defaulters. Thus we can make a conclusion here that Females are less likely to default than Males
- Most people have applied for Cash Loans
- People who don't own a Car have applied for more loans
- People who own a flat or a house have applied for more loans
- People who have submitted Document 3 are shown to be more likely to Default

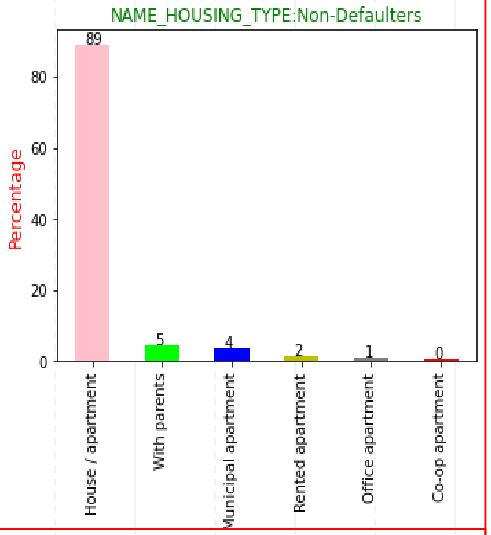

CREDIT_GROUP:Defaulters
Percentage
0
5
10
15
20
25
25
21
20
19
14
Medium
Low
High
Very Low
Very High
CREDIT_GROUP:Non-Defaulters
Percentage
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
21
21
20
20
19
Very Low
High
Medium
Very High
Low

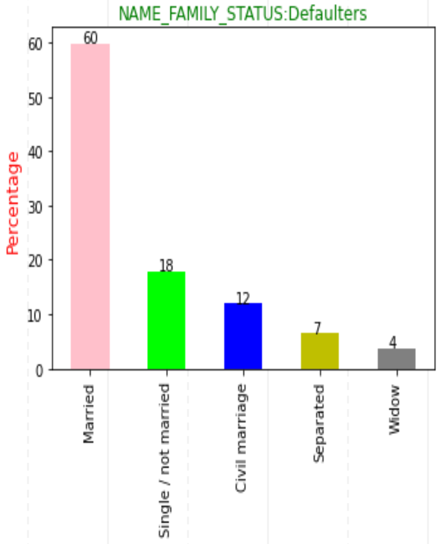

NAME_EDUCATION_TYPE:Defaulters
Percentage
0
10
20
30
40
50
60
70
80
79
16
4
2
0
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE:Non-Defaulters
Percentage
0
10
20
30
40
50
60
70
70
25
3
1
0
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree

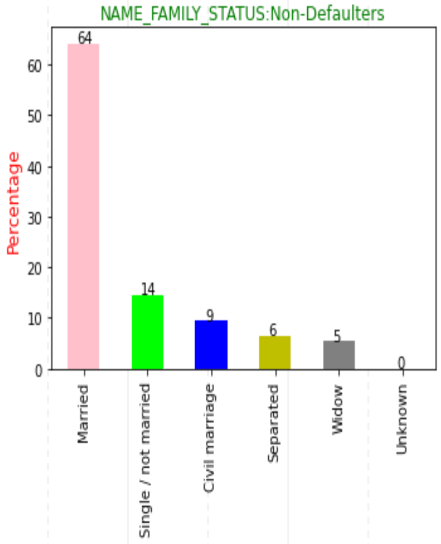

INCOME_GROUP:Defaulters
Percentage
0
5
10
15
20
25
30
30
25
21
12
12
Low
High
Very Low
Medium
Very High
INCOME_GROUP:Non-Defaulters
Percentage
0
5
10
15
20
25
28
25
21
16
11
Low
High
Very Low
Very High
Medium

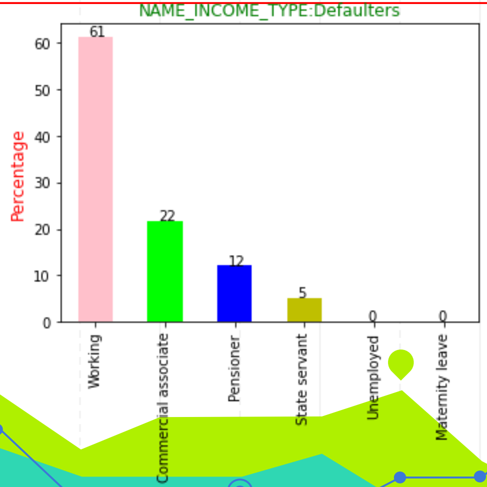

NAME_TYPE_SUITE:Defaulters
Percentage
0
10
20
30
40
50
60
70
80
82
12
4
1
1
0
0
Unaccompanied
Family
Spouse, partner
Children
Other_B
Other_A
Group of people
NAME_TYPE_SUITE:Non-Defaulters
Percentage
0
10
20
30
40
50
60
70
80
81
13
4
1
1
0
0
Unaccompanied
Family
Spouse, partner
Children
Other_B
Other_A
Group of people

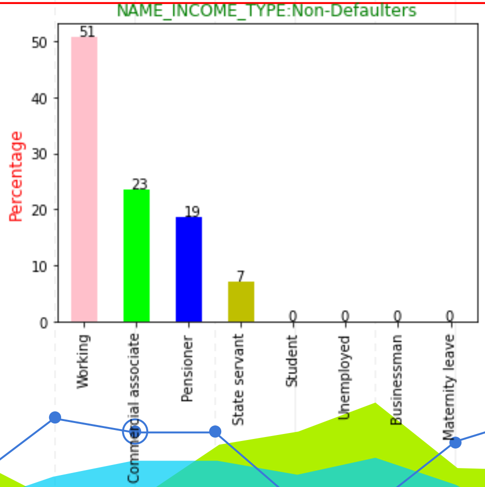

NAME_HOUSING_TYPE:Defaulters
Percentage
0
10
20
30
40
50
60
70
80
86
7
4
2
1
0
House / apartment
With parents
Municipal apartment
Rented apartment
Office apartment
Co-op apartment

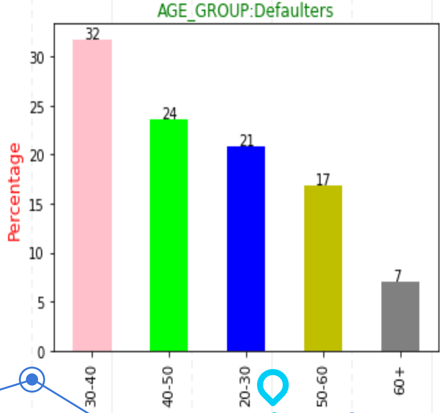

NAME_HOUSING_TYPE:Non-Defaulters
Percentage
0
20
40
60
80
89
5
4
2
1
0
House / apartment
With parents
Municipal apartment
Rented apartment
Office apartment
Co-op apartment

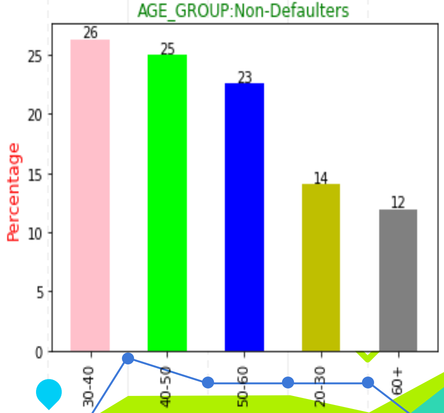

NAME_FAMILY_STATUS:Defaulters
Percentage
0
10
20
30
40
50
60
60
18
12
7
4
Married
Single / not married
Civil marriage
Separated
Widow


NAME_FAMILY_STATUS:Non-Defaulters
Percentage
0
10
20
30
40
50
60
64
14
9
6
5
0
Married
Single / not married
Civil marriage
Separated
Widow
Unknown

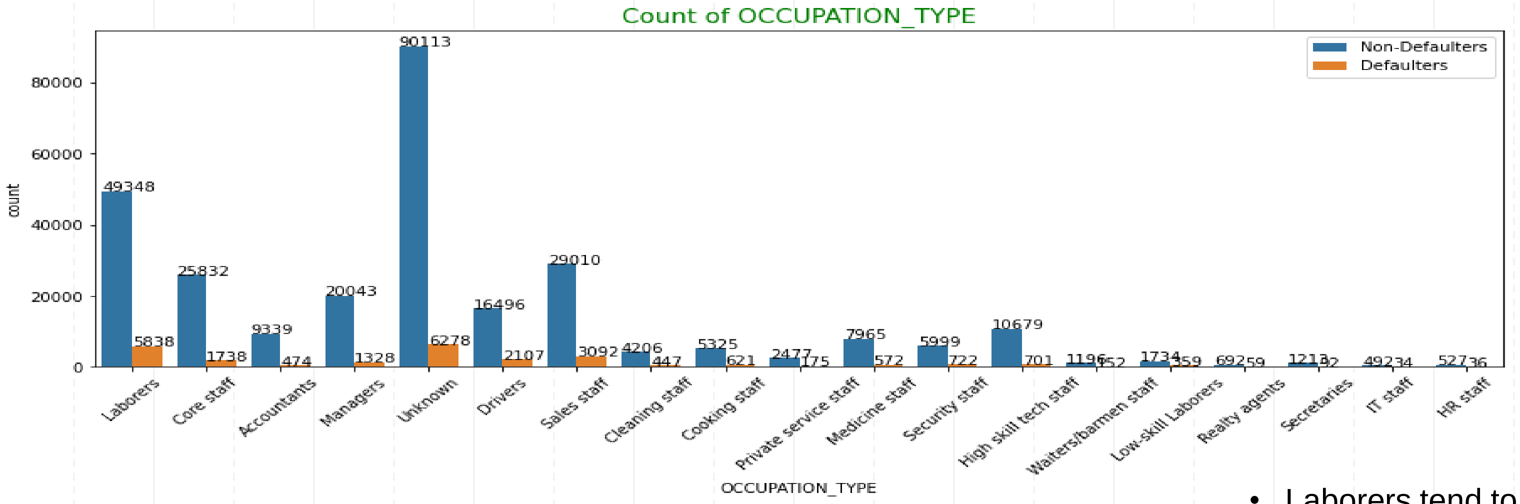

NAME_INCOME_TYPE:Defaulters
Percentage
0
10
20
30
40
50
60
61
22
12
5
0
0
Working
Commercial associate
Pensioner
State servant
Unemployed
Maternity leave

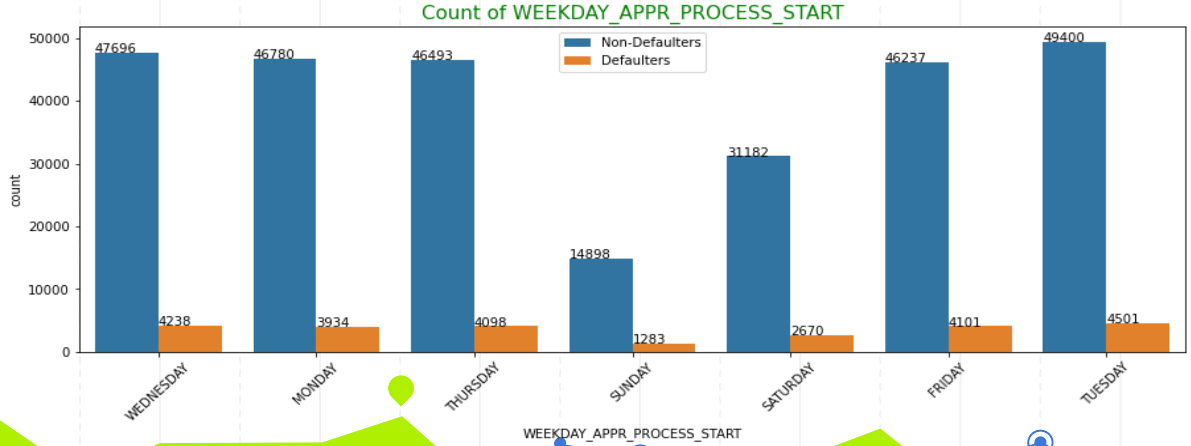

NAME_INCOME_TYPE:Non-Defaulters
Percentage
0
10
20
30
40
50
51
23
19
7
0
0
0
0
Working
Commercial associate
Pensioner
State servant
Student
Unemployed
Businessman
Maternity leave


AGE_GROUP:Defaulters
Percentage
0
5
10
15
20
25
30
32
24
21
17
7
30-40
40-50
20-30
50-60
60+

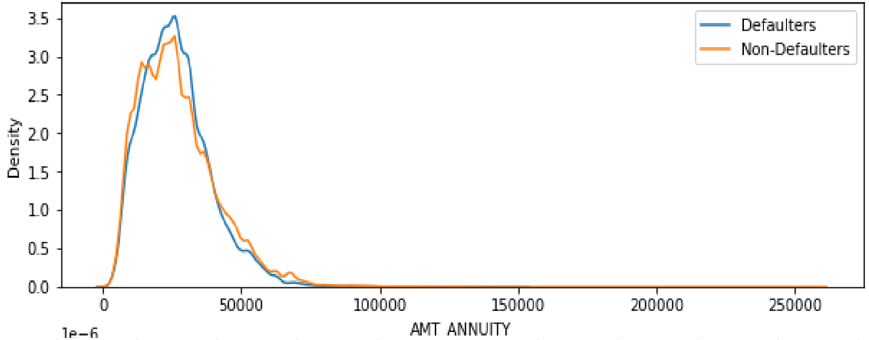

AGE_GROUP:Non-Defaulters
Percentage
0
5
10
15
20
25
26
25
23
14
12
30-40
40-50
50-60
20-30
60+

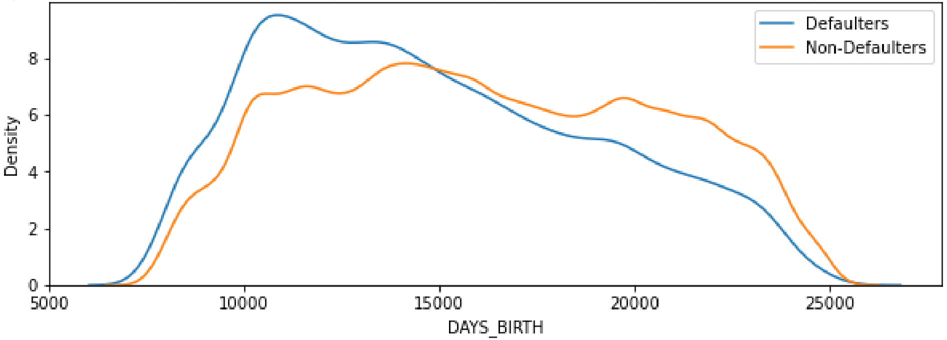
- As seen, people with *'Higher Education'* have a higher percentage in Non-Defaulters' case as compared to their percentage in Defaulters' case. Thus we infer that such people are less likely to Default on loan. Also, people with 'Academic Degree' are less likely to default but their numbers is very less, hence not much profitable. People with 'Secondary' and 'Incomplete Higher' are more likely to Default
- Married people tend to take more loans. Single and Civil marriage people are more likely to Default.
- People who have their own House/apartment tend to take more loans. People living with parents are more likely to Default.
- Working applied for most number of loans and more likely to Default, but Pensioners have least chance of Defaulting.
- People in 20-30 and 30-40 age group are more likely to default than those in 60+
- People with very high income are least likely to default
- People who have applied for loans in medium range are more likely to default

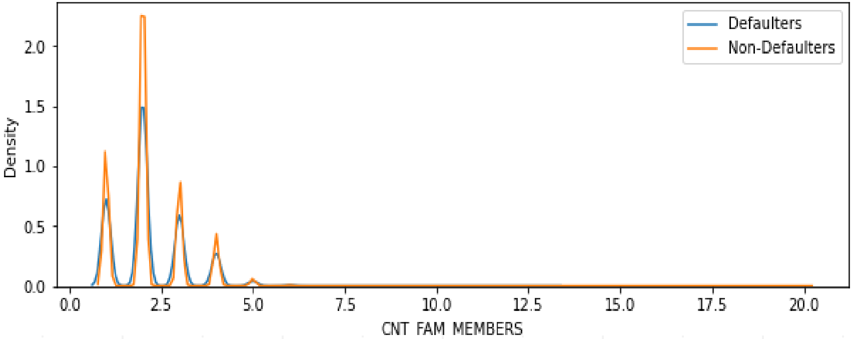


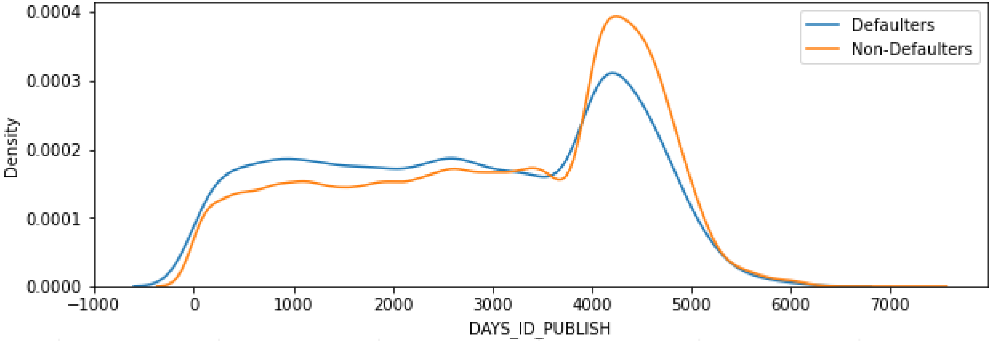


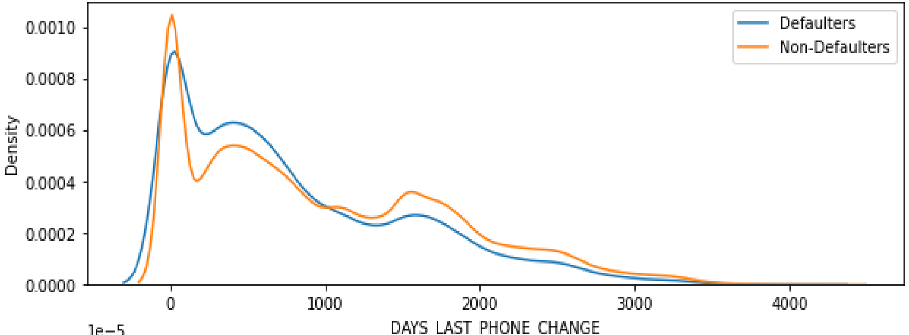
- Laborers tend to take more loans and are more likely to Default compared to other Occupation Types. IT Staff have taken least number of loans.
- People prefer mostly to apply for loans on Tuesday and least on Sunday. There can be some underlying reason for it that Sunday is mostly holiday for the Financial Institutions

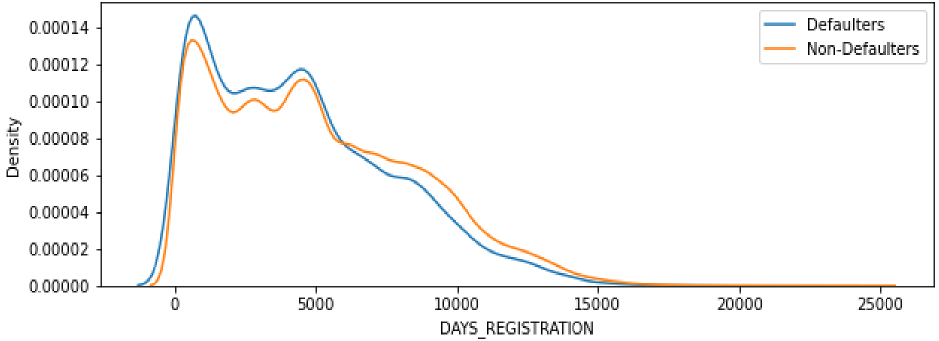

Defaulters
Non-Defaulters
Density
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
50000
100000
150000
200000
250000
AMT_ANNUITY
1e−6

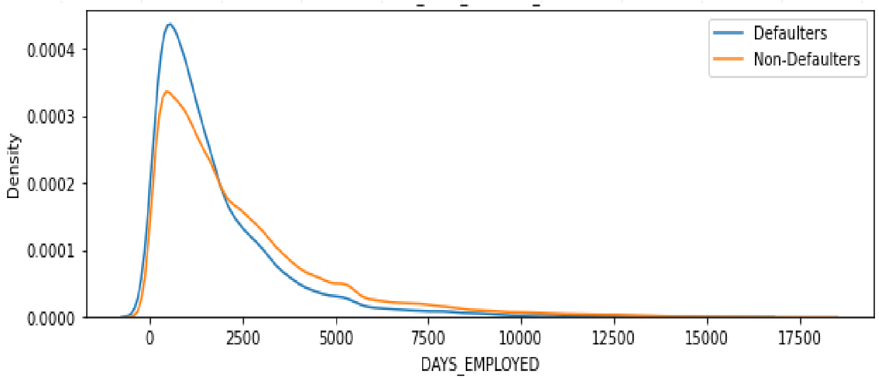

Defaulters
Non-Defaulters
Density
8
6
4
2
0
5000
10000
15000
20000
25000
DAYS_BIRTH

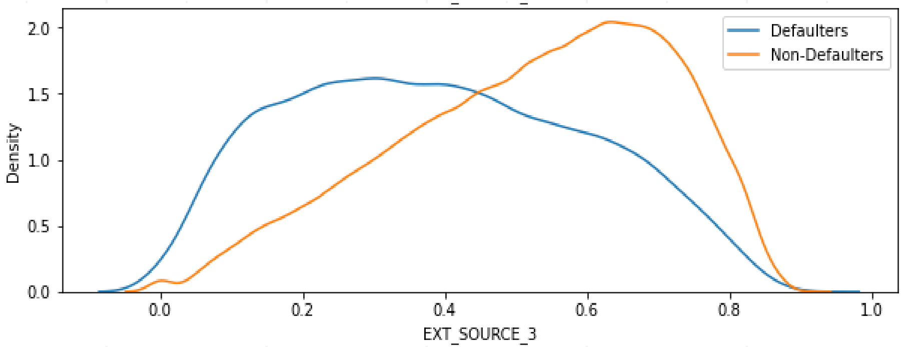

Defaulters
Non-Defaulters
Density
2.0
1.5
1.0
0.5
0.0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
CNT_FAM_MEMBERS

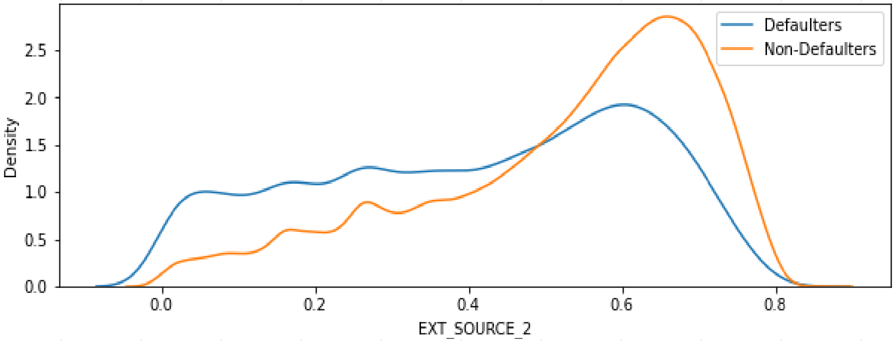

Defaulters
Non-Defaulters
Density
0.0004
0.0003
0.0002
0.0001
0.0000
−1000
0
1000
2000
3000
4000
5000
6000
7000
DAYS_ID_PUBLISH


Defaulters
Non-Defaulters
Density
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
1000
2000
3000
4000
DAYS_LAST_PHONE_CHANGE
1e−5


Defaulters
Non-Defaulters
Density
0.00014
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
5000
10000
15000
20000
25000
DAYS_REGISTRATION





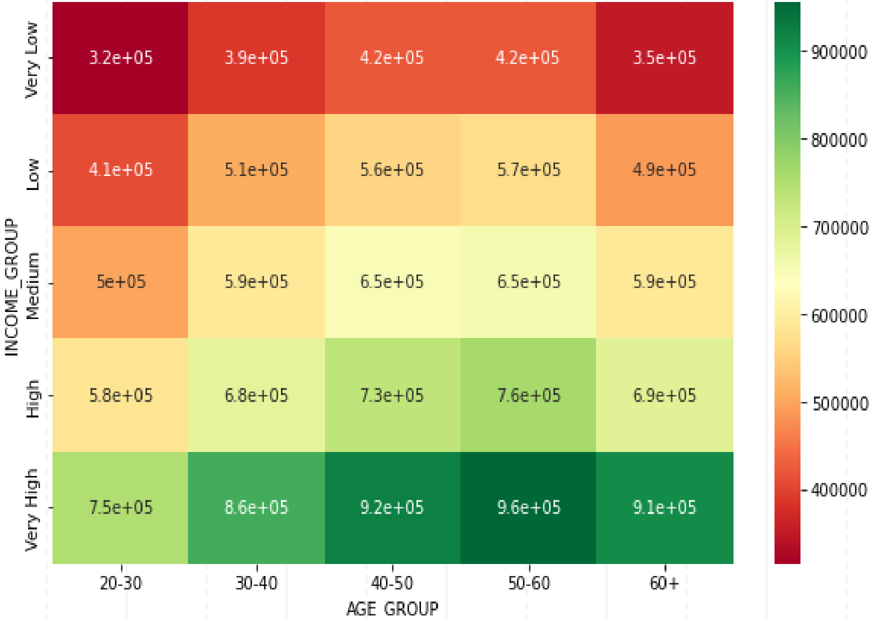


Numerical Variables

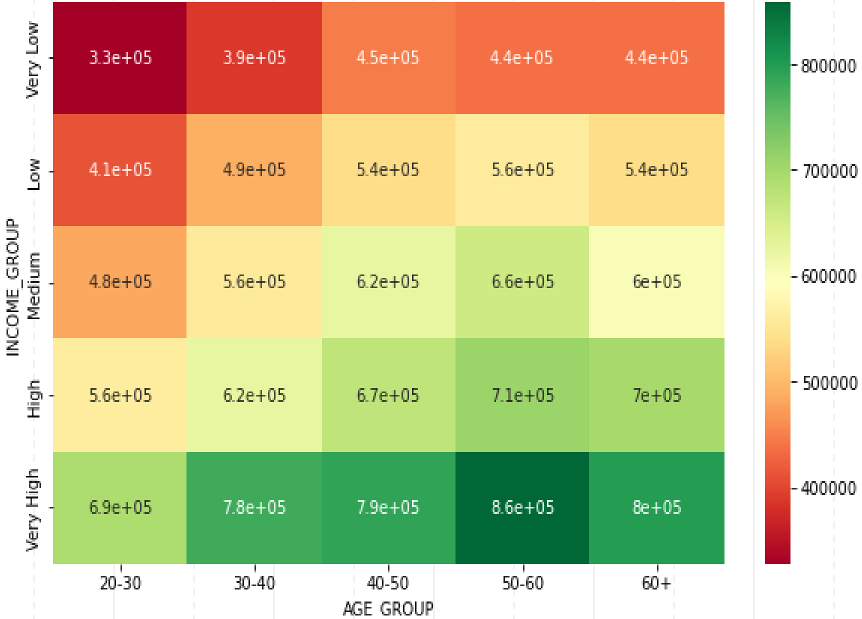
- People who changed their phone numbers recently are more likely to default
- People who changed their ID and registration recently are more likely to default
- People have applied more for the loans in the hours of 9 to 15
- People with low amount of Annuity have taken more loans
- People who are new in their jobs are more likely to apply for loans and also more likely to default
- Higher Defaulter rate for lower values of EXT_SOURCE_2 and EXT_SOURCE_3

Bivariate/Multivariate Analysis

Income vs Age vs Loan Credit for Non-Defaulters



Income vs Age vs Loan Credit for Defaulters



- Income group Very High in the age group 50-60 have a high amount of Credit for both Defaulters and Non-Defaulters, they can bring high profit to the company.
- For very low and low income group, mean of Credit amount is higher for Defaulters

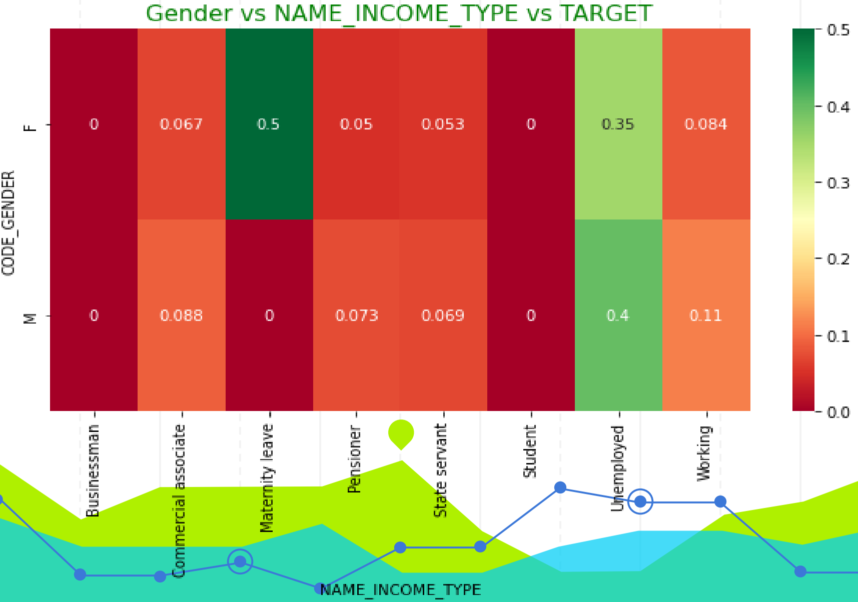
<u>Following have a higher chance of defaulting</u>

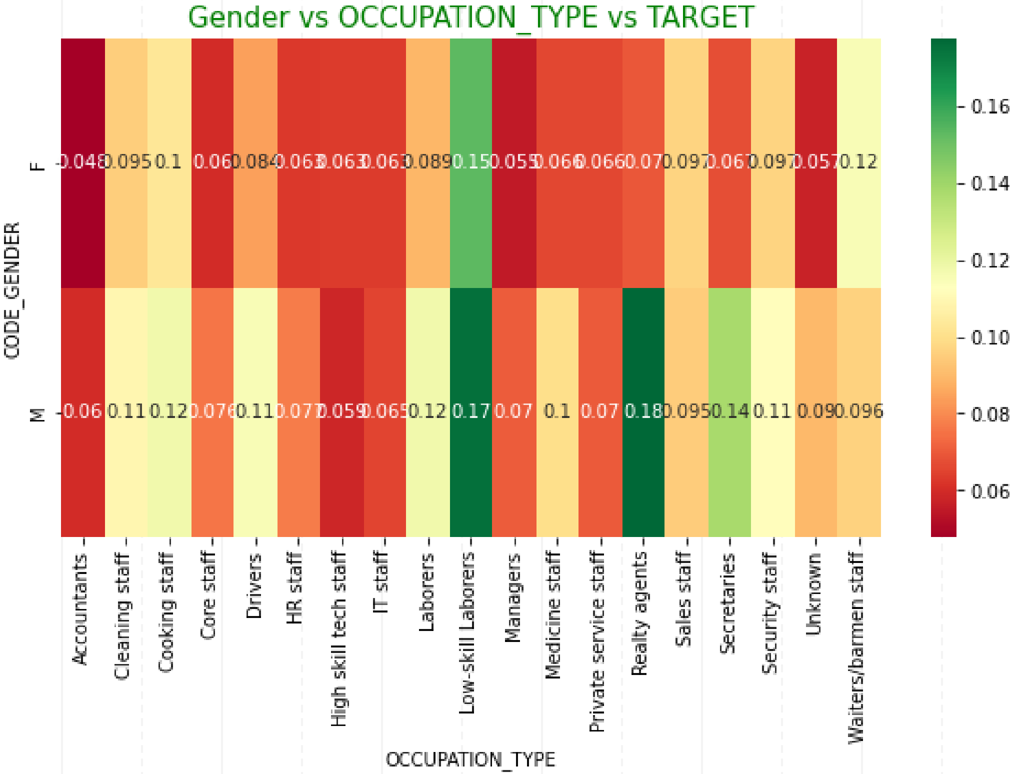
Males with

- Lower Secondary and Secondary education
- Civil Married, Separated, Single or Widowed
- Living with Parents or Rented Apartments
- Low-skilled Laborers or Realty Agents
- 20-30 followed by 30-40 age group
- Low and Medium Credit
- Very Low, Low and Medium Income

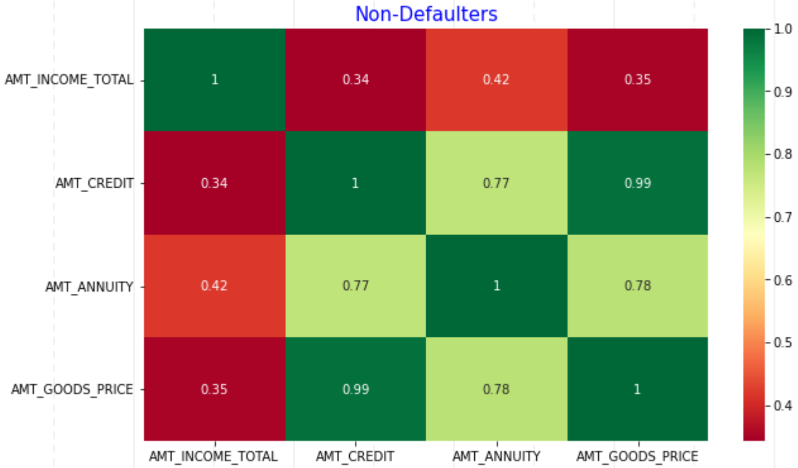
Females with

- On Maternity Leave

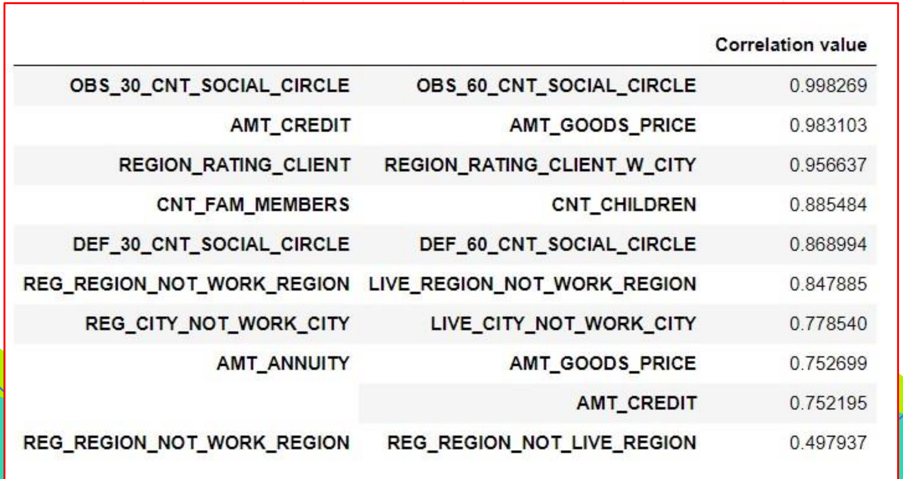




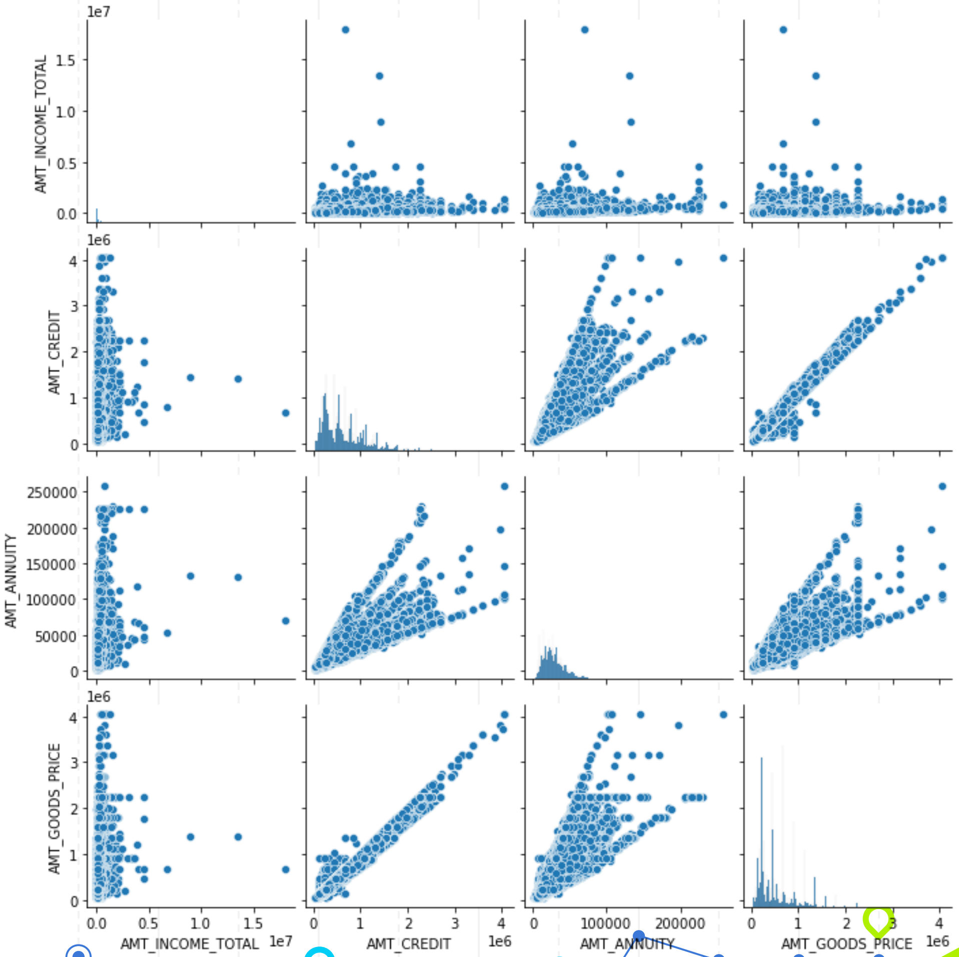
<u>Gender vs Categorical columns with value as Target mean</u>

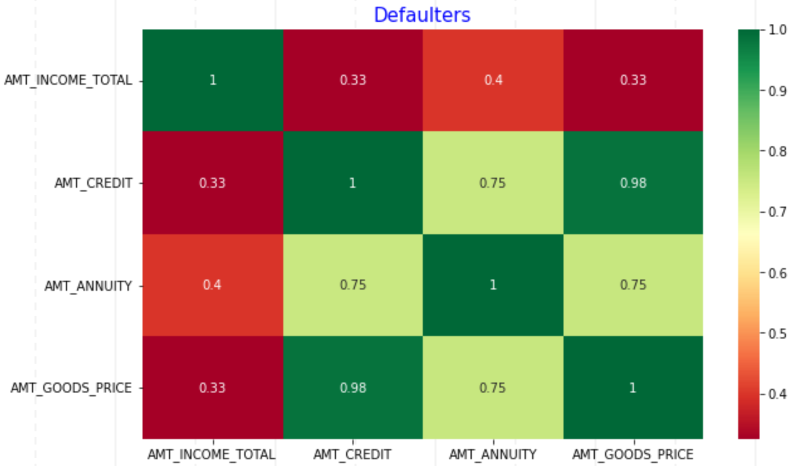


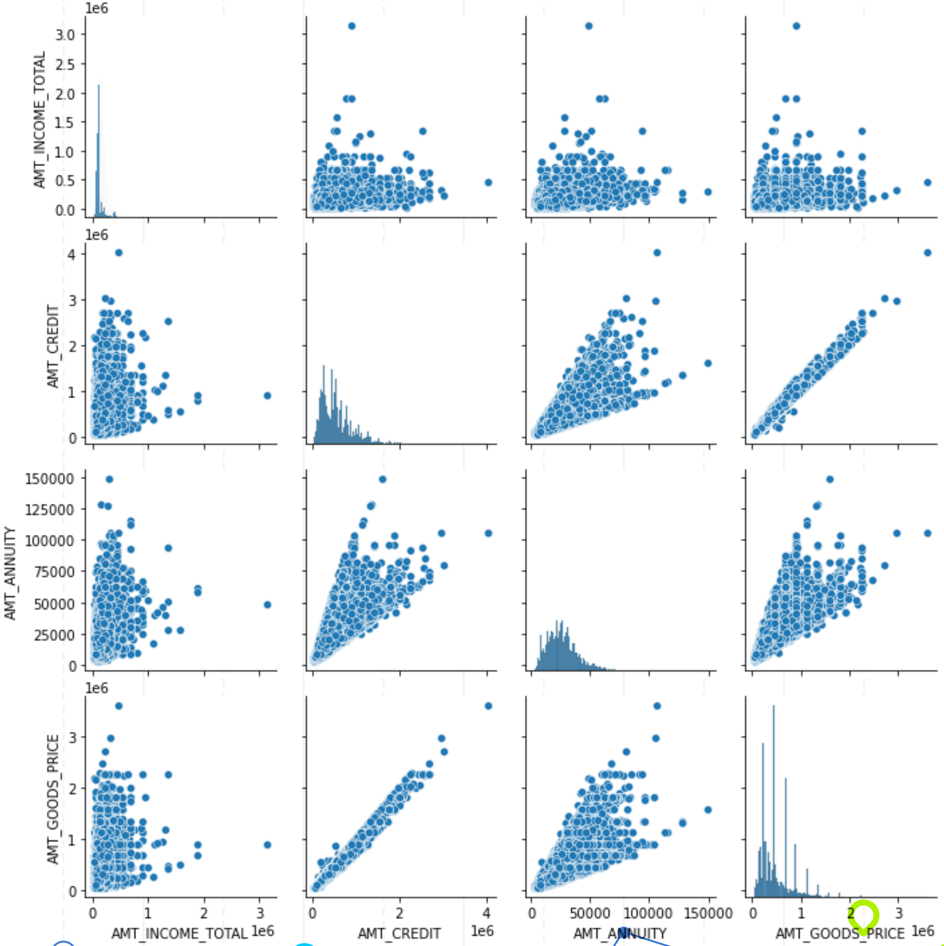
Top correlations for TARGET=0 (Non-Defaulters)



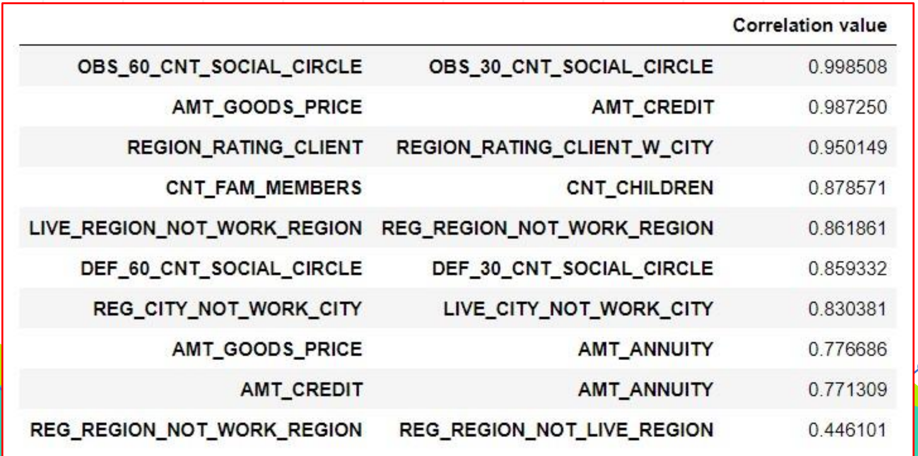
		Correlation value
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998269
AMT_CREDIT	AMT_GOODS_PRICE	0.983103
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.956637
CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.868994
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.847885
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.778540
AMT_ANNUITY	AMT_GOODS_PRICE	0.752699
	AMT_CREDIT	0.752195
REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.497937







Top correlations for TARGET=1 (Defaulters)



		Correlation value
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508
AMT_GOODS_PRICE	AMT_CREDIT	0.987250
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950149
CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859332
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.830381
AMT_GOODS_PRICE	AMT_ANNUITY	0.776686
AMT_CREDIT	AMT_ANNUITY	0.771309
REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.446101

“

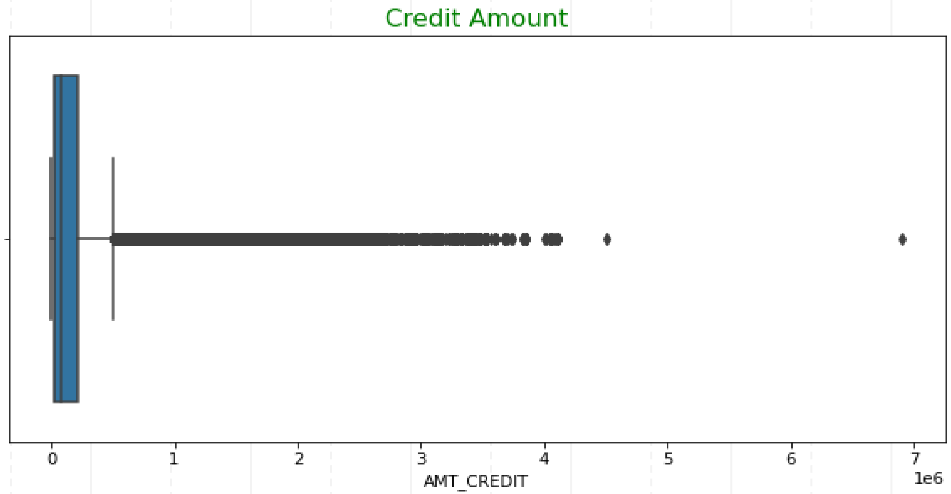
Analysis for previous_application

Data Cleaning

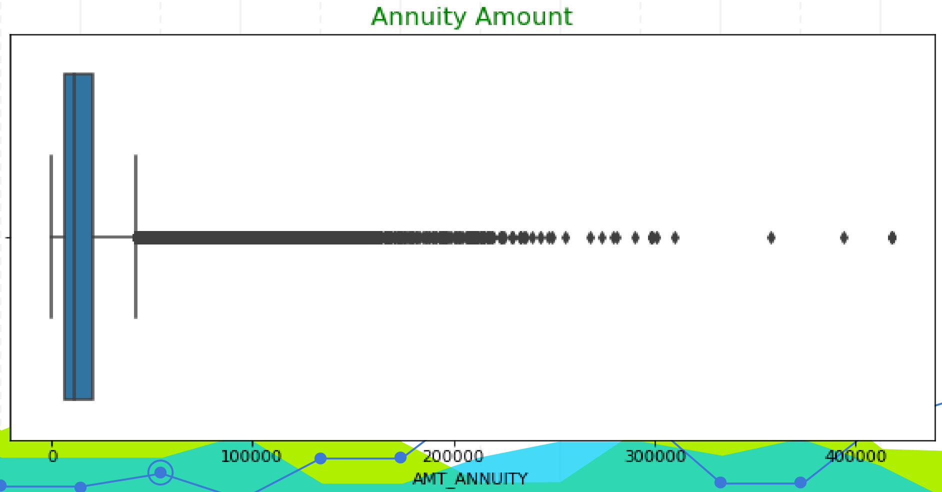
Dealing with Null values and removing unwanted columns

- Columns having Null value percentage > 40% can be removed
- *Dropped columns:*
  - *NAME_CASH_LOAN_PURPOSE*: High XAP and XNA values (95%)
  - *WEEKDAY_APPR_PROCESS_START*: Not relevant
  - *WEEKDAY_APPR_PROCESS_START*: 99% values are same
  - *NAME_GOODS_CATEGORY*: High XNA values (56%)
  - *SELLERPLACE_AREA*: not relevant
  - *FLAG_LAST_APPL_PER_CONTRACT* : more than 99% value of same type (Y)
- Replacing all XNA and XAP values with 'Unknown‘ for categorical columns

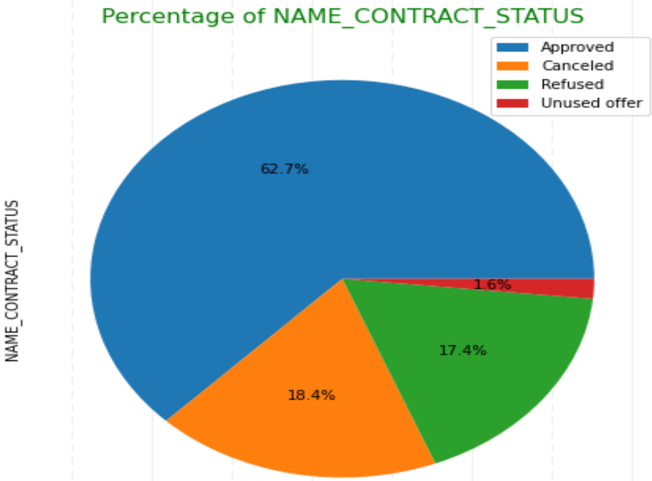
Dealing with Outliers

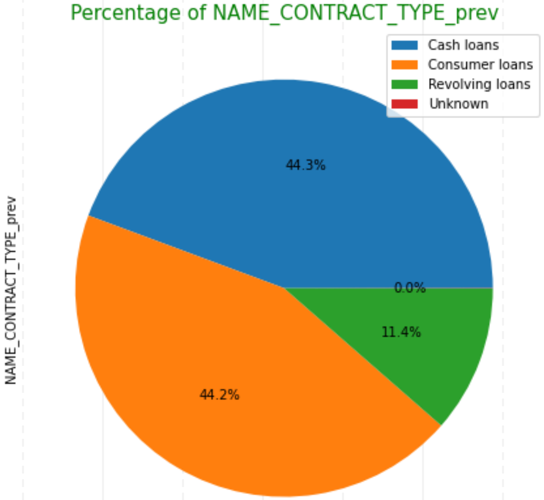


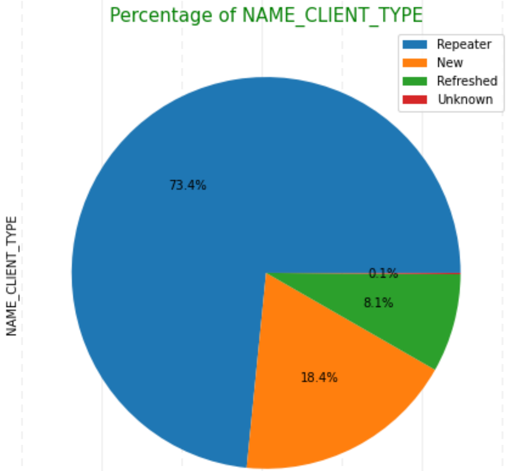
- As seen, there is one high value of outlier
- While dealing with this variable, use median instead of mean

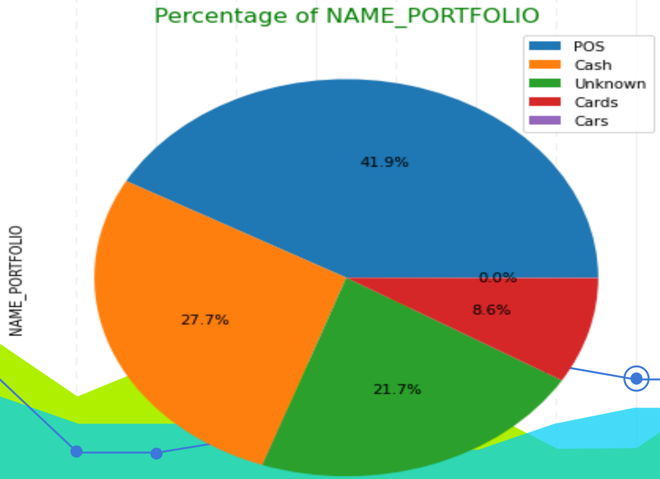


- As seen, there are few high values of outlier
- While dealing with this variable, use median instead of mean

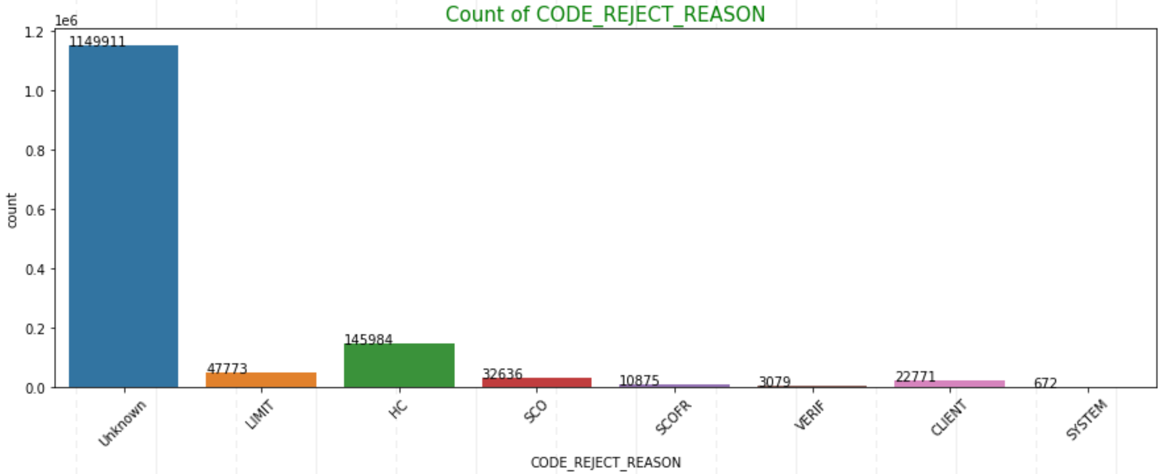


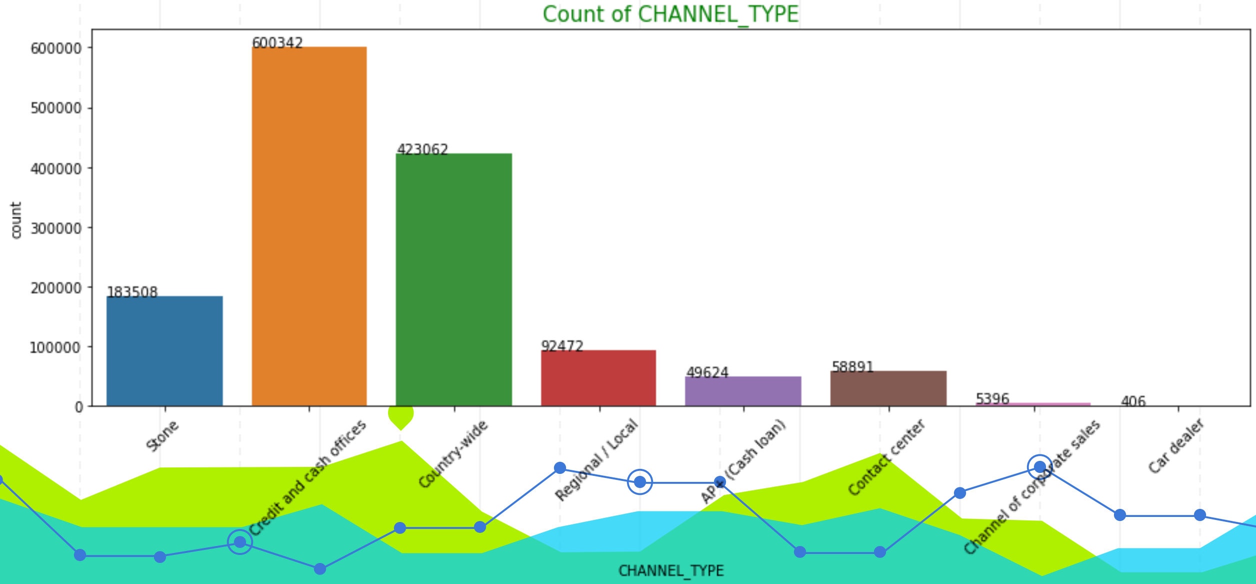






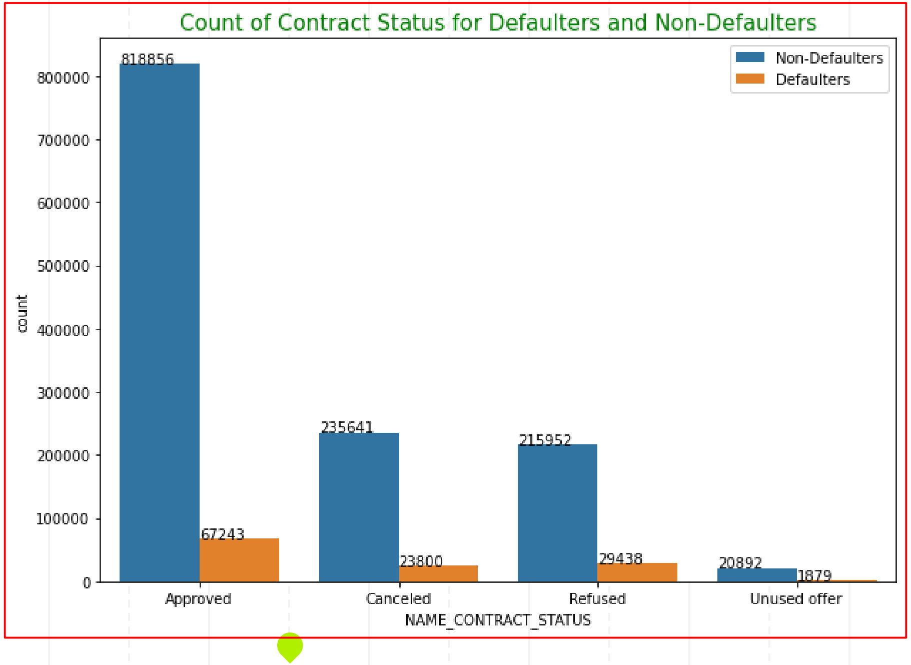
Univariate Analysis





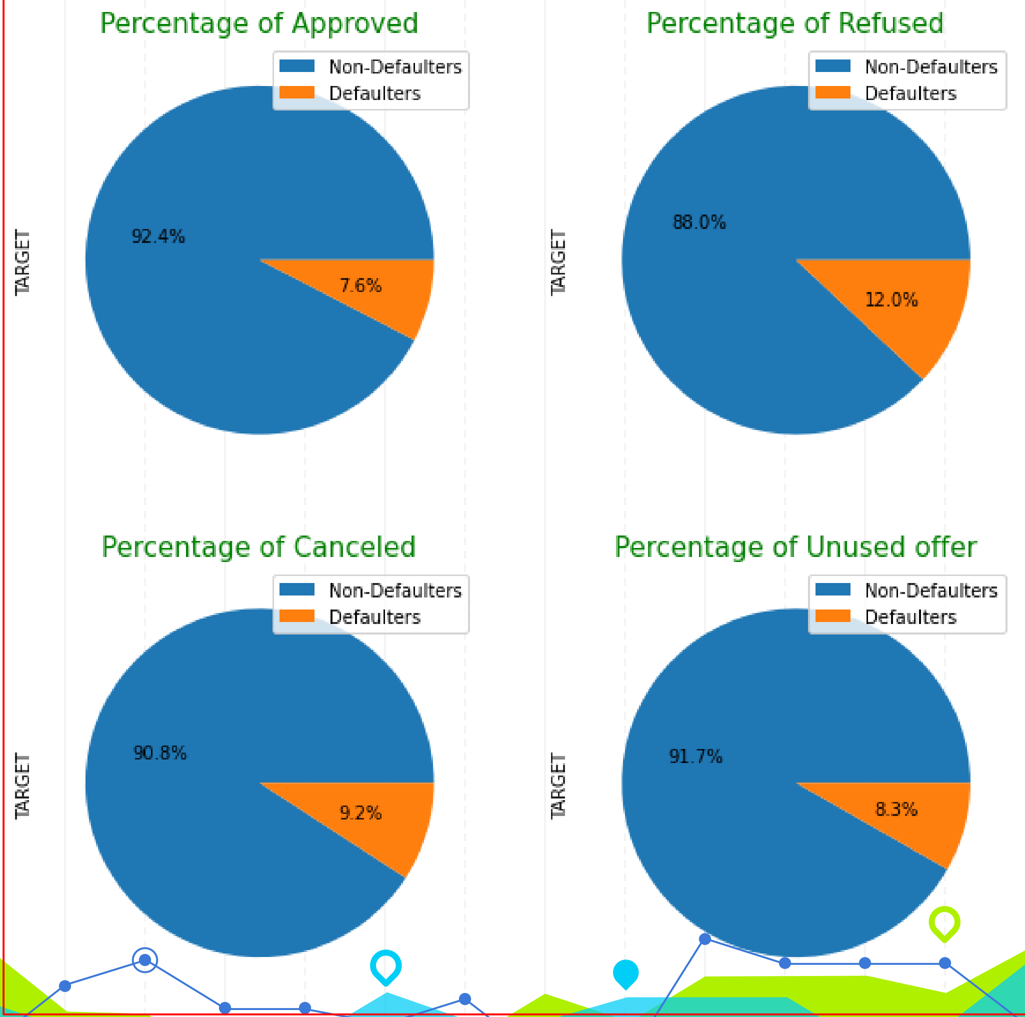
- As seen, most of the loans (63%) which were applied got approved
- Most loan types are Consumer Loans and Cash Loans
- HC is the most given reason for rejection of loans
- Highest number of consumers are of Middle group
- 73% of people are repeaters
- Most people are from Consumer Electronics
- Most clients were acquired by Cash and Credit offices

Bivariate/Multivariate Analysis

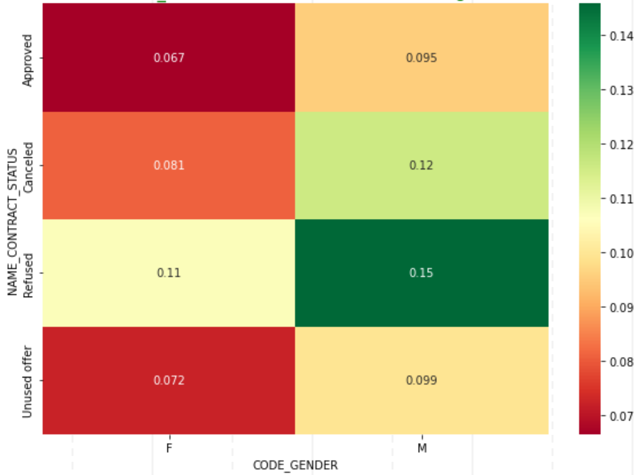


Count for different categories of Contract status for Defaulters and Non-Defaulters

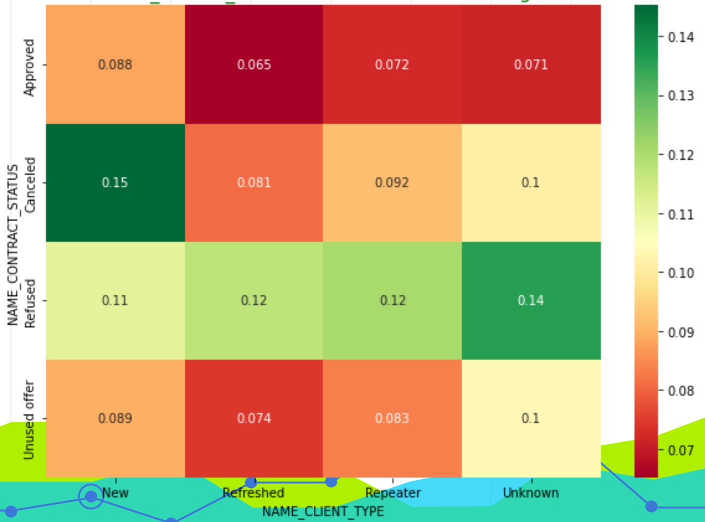
- Around 7.6 % people whose contract was approved previously are in Defaulters category. This percentage has to be brought down even further to enhance profit.
- Around 88.0% people whose contract was refused previously are in Non-Defaulters category. This percentage is very high which can lead to loss to the company.
- Around 90% people whose contract got Canceled previously are in Non-Defaulters category. This percentage is very high which can lead to loss to the company.



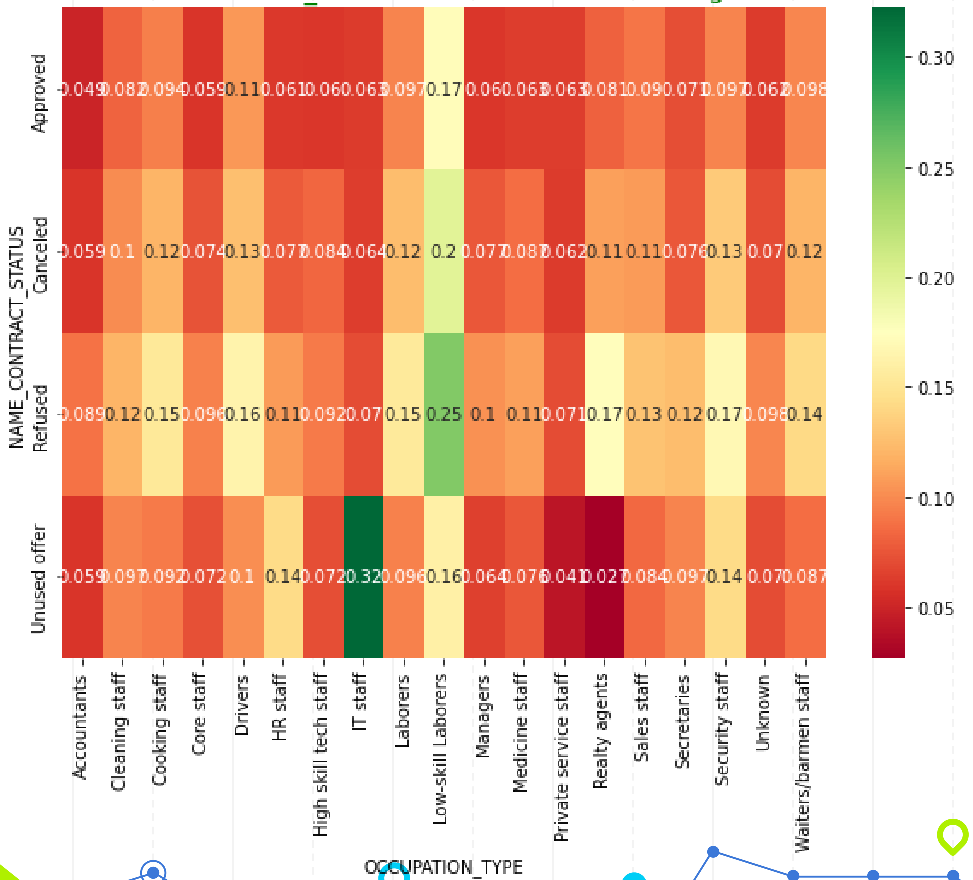
CODE_GENDER vs Contract Status vs Target

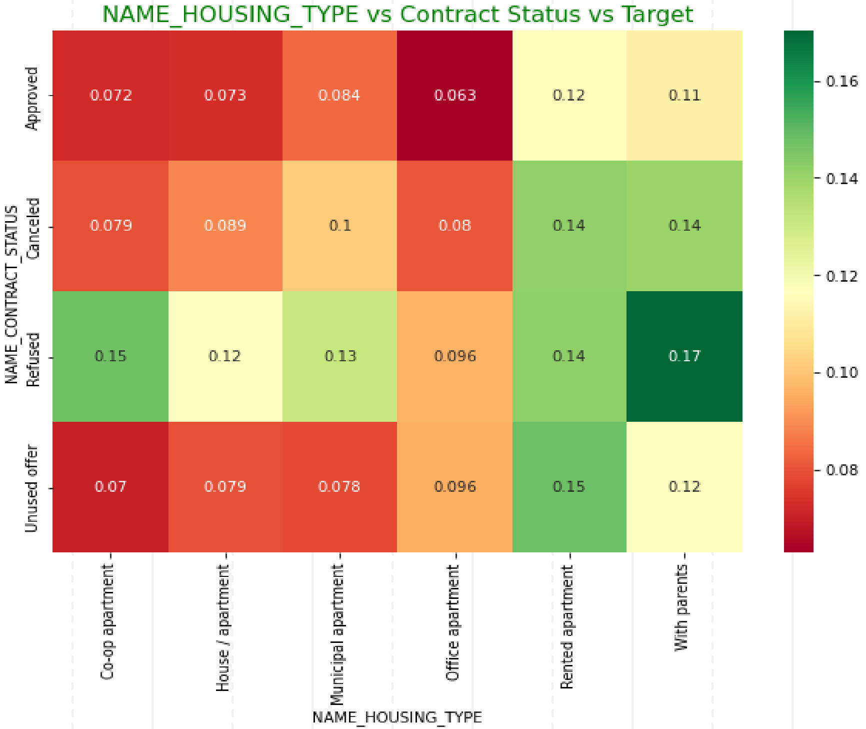


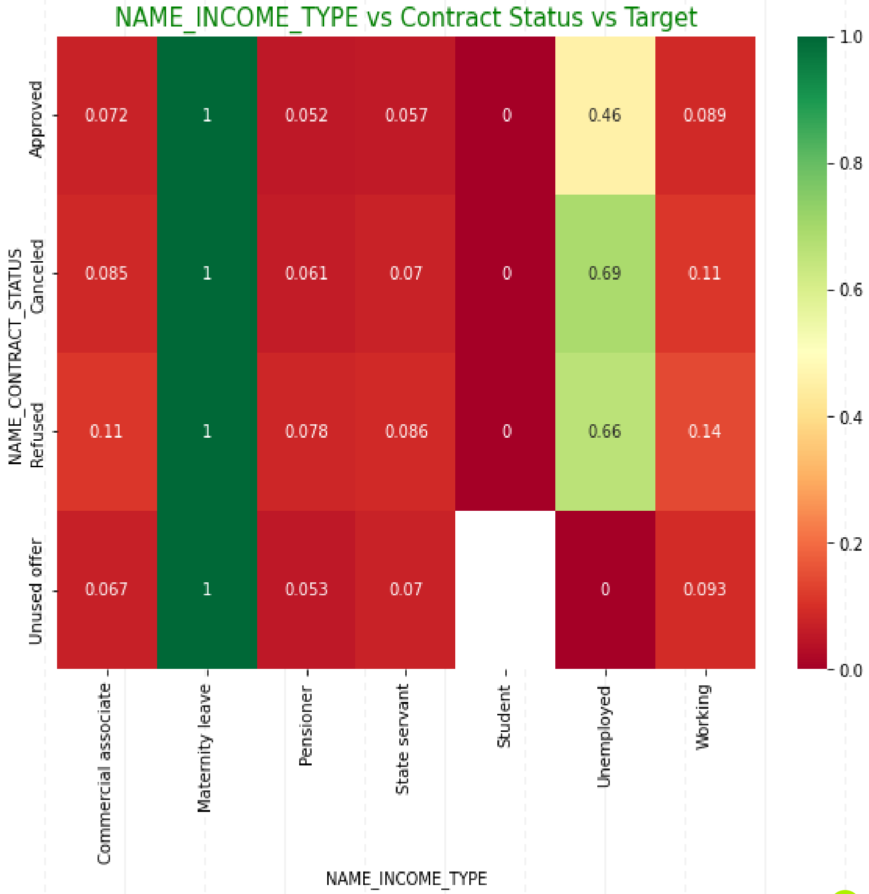
NAME_CLIENT_TYPE vs Contract Status vs Target



OCCUPATION_TYPE vs Contract Status vs Target

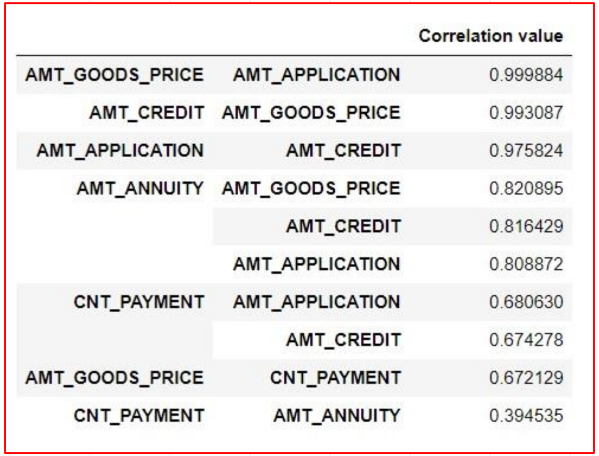






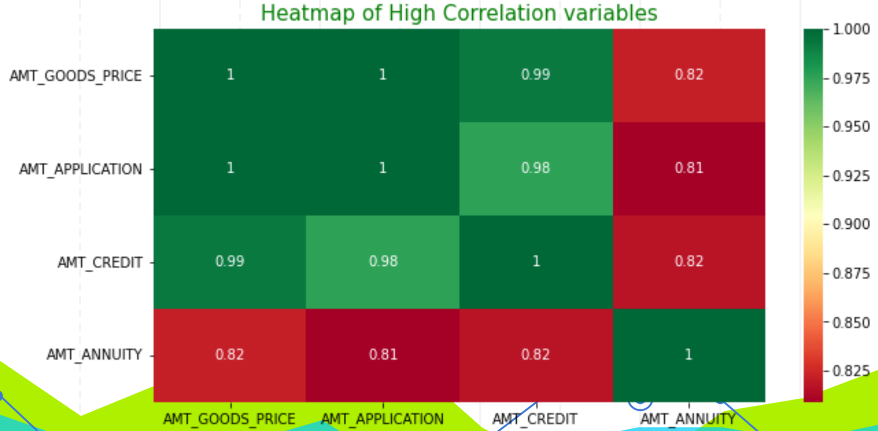
__Following people have a higher chance of Defaulting__

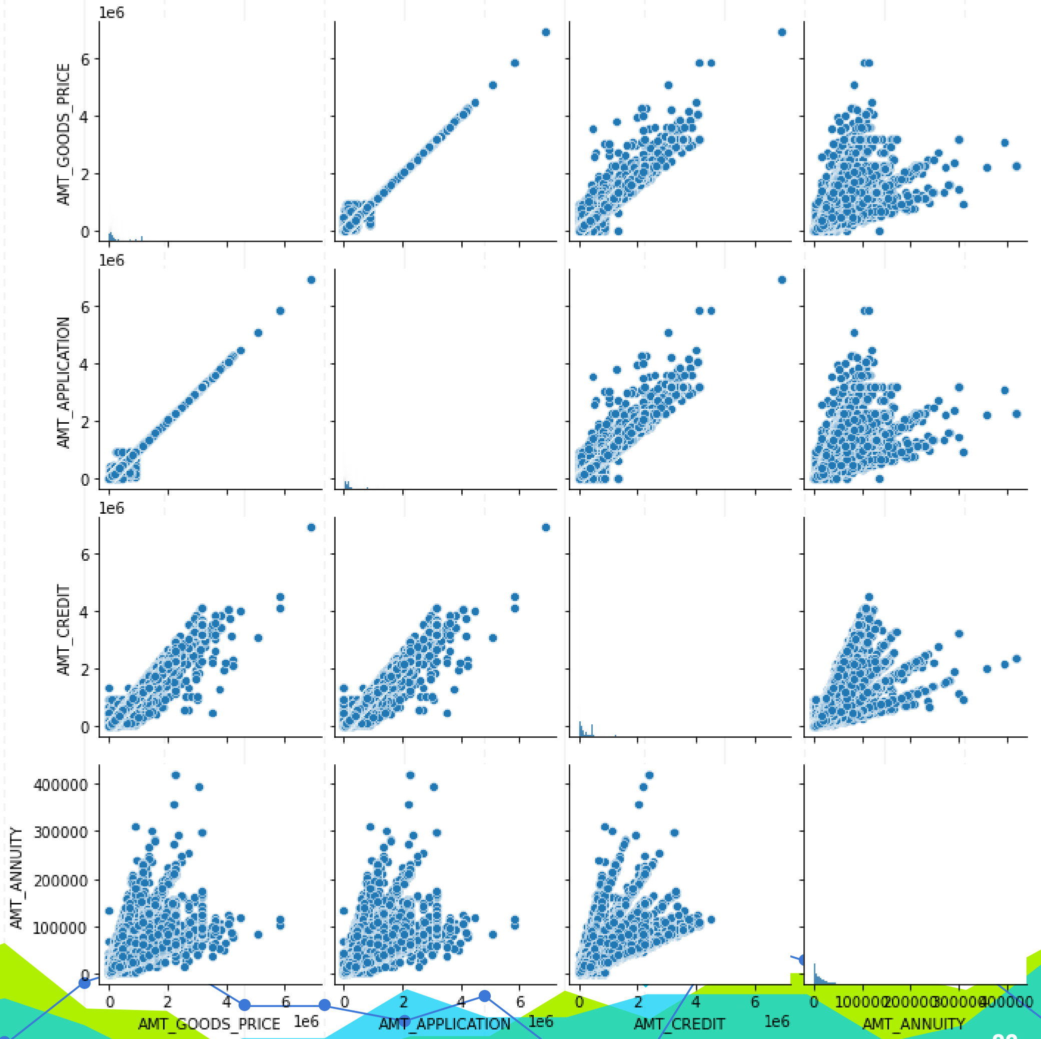
- Males who were previously refused
- Unemployed people who were previously refused or canceled
- Females on maternity leave
- People living with Parents who were previously refused
- IT-Staff who had previously unused offer and Low-skilled laborers who were previously refused or canceled
- 20-30 or 30-40 age group who were previously refused
- New clients who got canceled previously



		Correlation value
AMT_GOODS_PRICE	AMT_APPLICATION	0.999884
AMT_CREDIT	AMT_GOODS_PRICE	0.993087
AMT_APPLICATION	AMT_CREDIT	0.975824
AMT_ANNUITY	AMT_GOODS_PRICE	0.820895
	AMT_CREDIT	0.816429
	AMT_APPLICATION	0.808872
CNT_PAYMENT	AMT_APPLICATION	0.680630
	AMT_CREDIT	0.674278
AMT_GOODS_PRICE	CNT_PAYMENT	0.672129
CNT_PAYMENT	AMT_ANNUITY	0.394535

Top 10 correlations for numeric variables





Conclusions

application_data

- Income group Very High in the age group 50-60 have a high amount of Credit for both Defaulters and Non-Defaulters, they can bring high profit to the company.
- The following have resulted in defaulting more than others:
  - Males
  - Unemployed
  - Females on maternity leave
  - 20-30 and 30-40 age group
  - Laborers ,low-skilled laborers
  - People with less working experience
- The following should also be kept an eye on
  - People who changed their phone numbers recently
  - Lower values of EXT_SOURCE_2 and EXT_SOURCE_3

Conclusions

previous_application

- Around 7.6 % people whose contract was approved previously are in Defaulters category. This percentage has to be brought down even further to enhance profit.
- Around 88.0% people whose contract was refused previously are in Non-Defaulters category. This percentage is very high which can lead to loss to the company.
- Around 90% people whose contract got Canceled previously are in Non-Defaulters category. This percentage is very high which can lead to loss to the company.
- The following people are credible but were refused loan previously:
    - Females
    - Widows
    - Pensioners
    - 60+ age group
    - Car dealers

THANKS!