CREDIT EDA CASE STUDY

It basically gives an idea on applying EDA in a real business scenario. It develops a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

PREPARED BY:
ADITYA RANJAN BEHERA & ABHIJEET BEBARTA

➝ INTRODUCTION

- The aim of this report is to analyse the patterns present in the data using EDA which helps to indicate whether the client has difficulty in paying the loan.

- This analysis will help the loan paying organizations to take actions such as denying the loan, reducing the amount of loan, lending to risky applicants at a higher interest rate, etc.

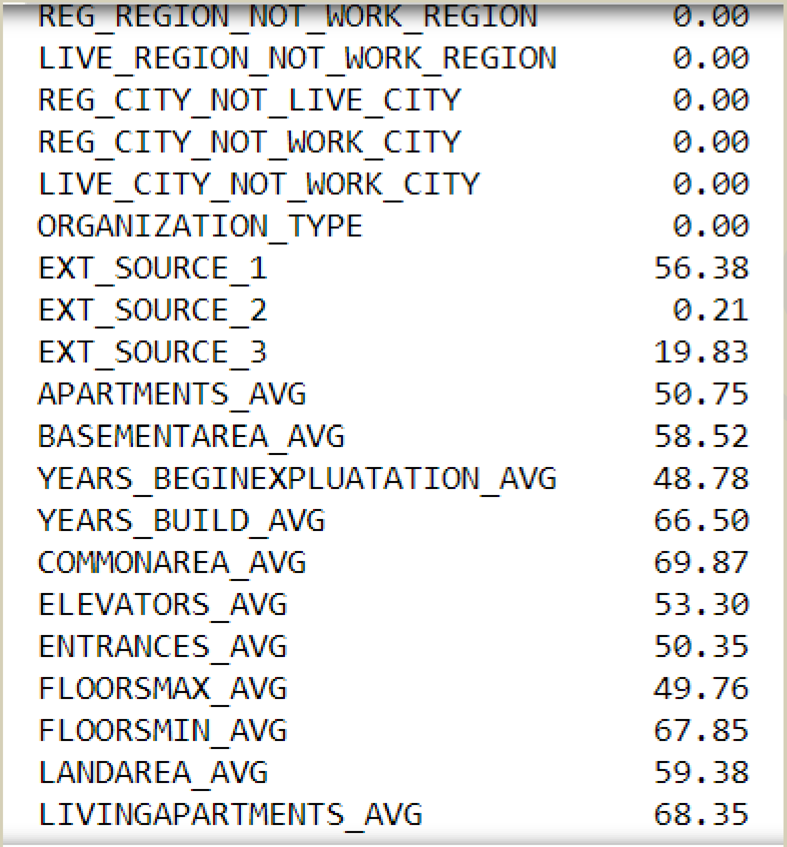
- It will give an insight to the company to understand the driving factors behind loan default.

- This analysis can be utilized by the company to assess its portfolio and risk management.

➞ Identifying the missing data and using the appropriate method to deal with it

We observed a lot of columns were having missing values. So, if we remove those missing columns, then we will be left with very few data to start with our analysis.

We used indexing and count of missing values in each column to identify the percentage of missing values in each column.



```
REG_REGION_NOT_WORK_REGION        0.00
LIVE_REGION_NOT_WORK_REGION       0.00
REG_CITY_NOT_LIVE_CITY            0.00
REG_CITY_NOT_WORK_CITY            0.00
LIVE_CITY_NOT_WORK_CITY           0.00
ORGANIZATION_TYPE                 0.00
EXT_SOURCE_1                     56.38
EXT_SOURCE_2                      0.21
EXT_SOURCE_3                     19.83
APARTMENTS_AVG                   50.75
BASEMENTAREA_AVG                 58.52
YEARS_BEGINEXPLUATATION_AVG      48.78
YEARS_BUILD_AVG                  66.50
COMMONAREA_AVG                   69.87
ELEVATORS_AVG                    53.30
ENTRANCES_AVG                    50.35
FLOORSMAX_AVG                    49.76
FLOORSMIN_AVG                    67.85
LANDAREA_AVG                     59.38
LIVINGAPARTMENTS_AVG             68.35
```

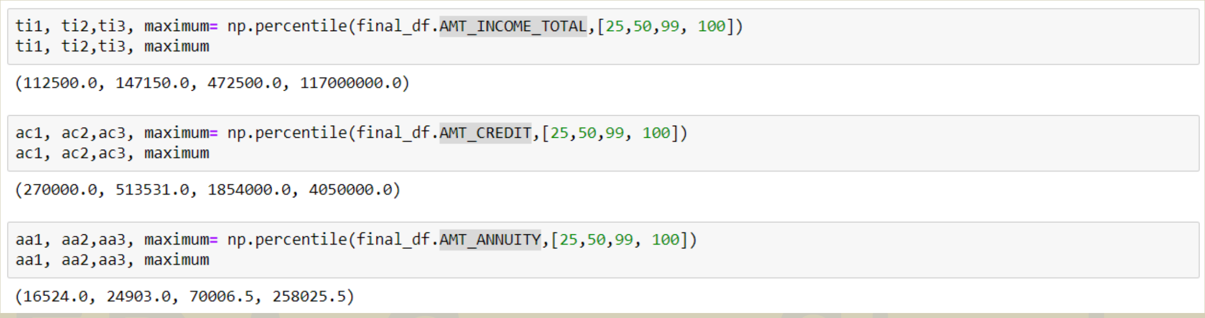
Hence, we decided to impute the missing data.

- The Numerical values were imputed with Median
- The categorical values were imputed with Mode.

➙ Identifying the outliers and the reason for considering them as an Outlier.

To find outliers, we considered the below three columns and analysed the statistics of them.

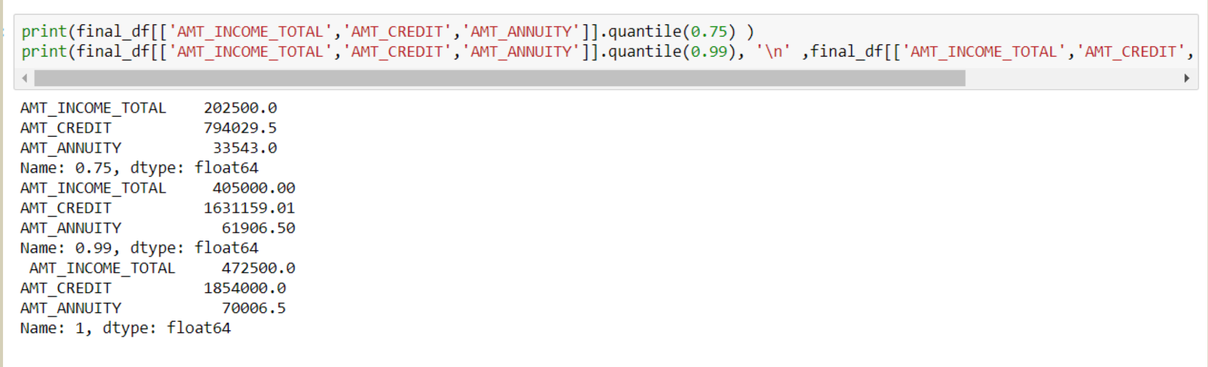
- AMT_INCOME_TOTAL
- AMT_CREDIT
- AMT_ANNUITY



```
ti1, ti2,ti3, maximum= np.percentile(final_df.AMT_INCOME_TOTAL,[25,50,99, 100])
ti1, ti2,ti3, maximum

(112500.0, 147150.0, 472500.0, 117000000.0)

ac1, ac2,ac3, maximum= np.percentile(final_df.AMT_CREDIT,[25,50,99, 100])
ac1, ac2,ac3, maximum

(270000.0, 513531.0, 1854000.0, 4050000.0)

aa1, aa2,aa3, maximum= np.percentile(final_df.AMT_ANNUITY,[25,50,99, 100])
aa1, aa2,aa3, maximum

(16524.0, 24903.0, 70006.5, 258025.5)
```

From the above data, we can clearly observe that there is a huge difference between the 99% interval and maximum value. So, we consider them outliers and removed them for our analysis as it may impact our results.

After removal of records, we checked our results.

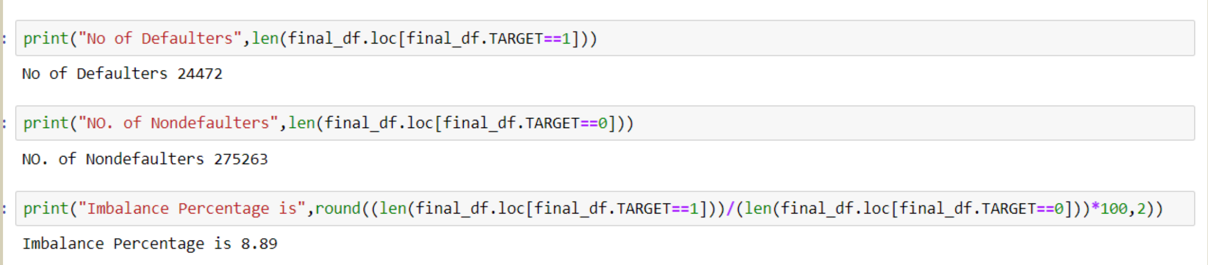


```
print(final_df[['AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY']].quantile(0.75) )
print(final_df[['AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY']].quantile(0.99), '\n' ,final_df[['AMT_INCOME_TOTAL','AMT_CREDIT',

AMT_INCOME_TOTAL     202500.0
AMT_CREDIT           794029.5
AMT_ANNUITY           33543.0
Name: 0.75, dtype: float64
AMT_INCOME_TOTAL      405000.00
AMT_CREDIT           1631159.01
AMT_ANNUITY            61906.50
Name: 0.99, dtype: float64
 AMT_INCOME_TOTAL      472500.0
AMT_CREDIT           1854000.0
AMT_ANNUITY            70006.5
Name: 1, dtype: float64
```

We observe there is no significant difference between Q99 and Max. Hence this proves that our outliers are removed.

➙ Identifying the data imbalance and their ratio.

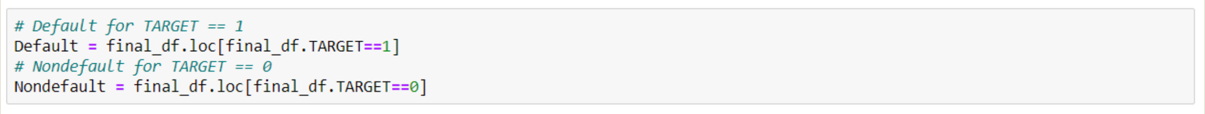
While working with the data frame based on target variable, we can see the imbalance between target type 1 and 0.



```
print("No of Defaulters",len(final_df.loc[final_df.TARGET==1]))

No of Defaulters 24472

print("NO. of Nondefaulters",len(final_df.loc[final_df.TARGET==0]))

NO. of Nondefaulters 275263

print("Imbalance Percentage is",round((len(final_df.loc[final_df.TARGET==1]))/(len(final_df.loc[final_df.TARGET==0]))*100,2))

Imbalance Percentage is 8.89
```

The data imbalance ratio is 8.89

Dividing the datasets into two parts according to target columns



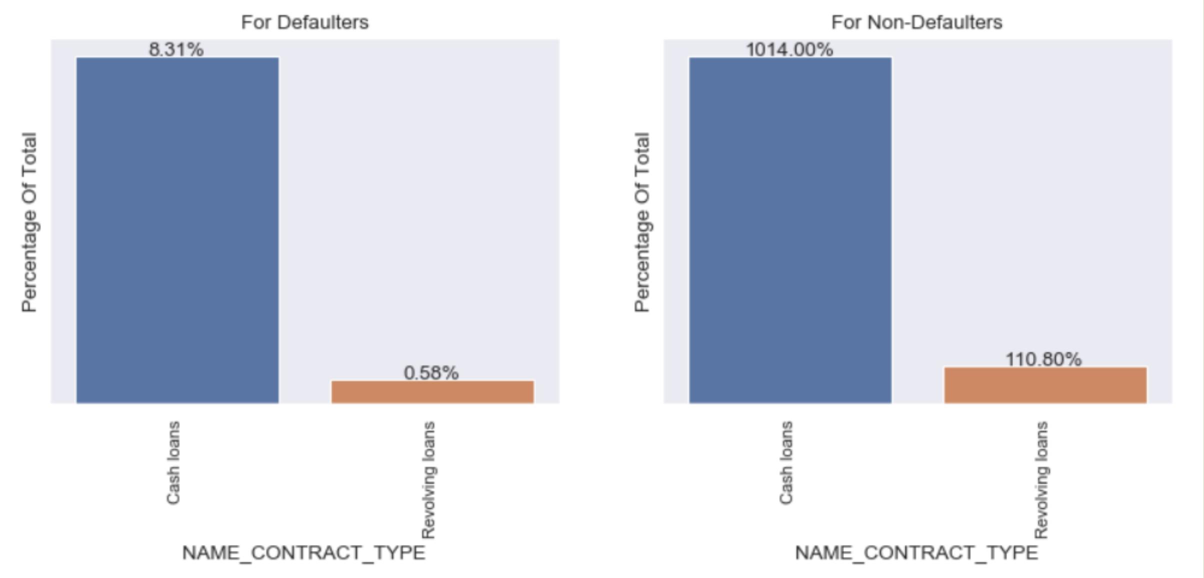
```
# Default for TARGET == 1
Default = final_df.loc[final_df.TARGET==1]
# Nondefault for TARGET == 0
Nondefault = final_df.loc[final_df.TARGET==0]
```

➙ Plot more than one type of plot to analyse the different aspects due to data imbalance. Do this analysis for the '**Target variable**' in the dataset (**clients with payment difficulties and all other cases**). Use a mix of univariate and bivariate analysis etc.

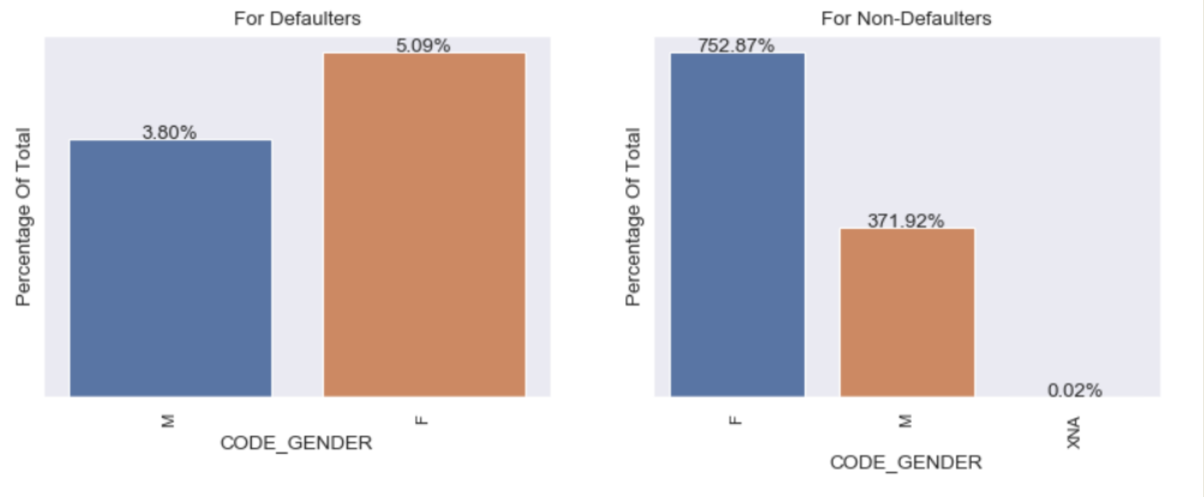
UNIVARIATE ANALYSIS

In order to perform Univariate analyis on categorical variables available in the data set for both 0 and 1, we have choosen few columns in terms of variable and compared against Percentage of Total in data frame and based on Target variable.

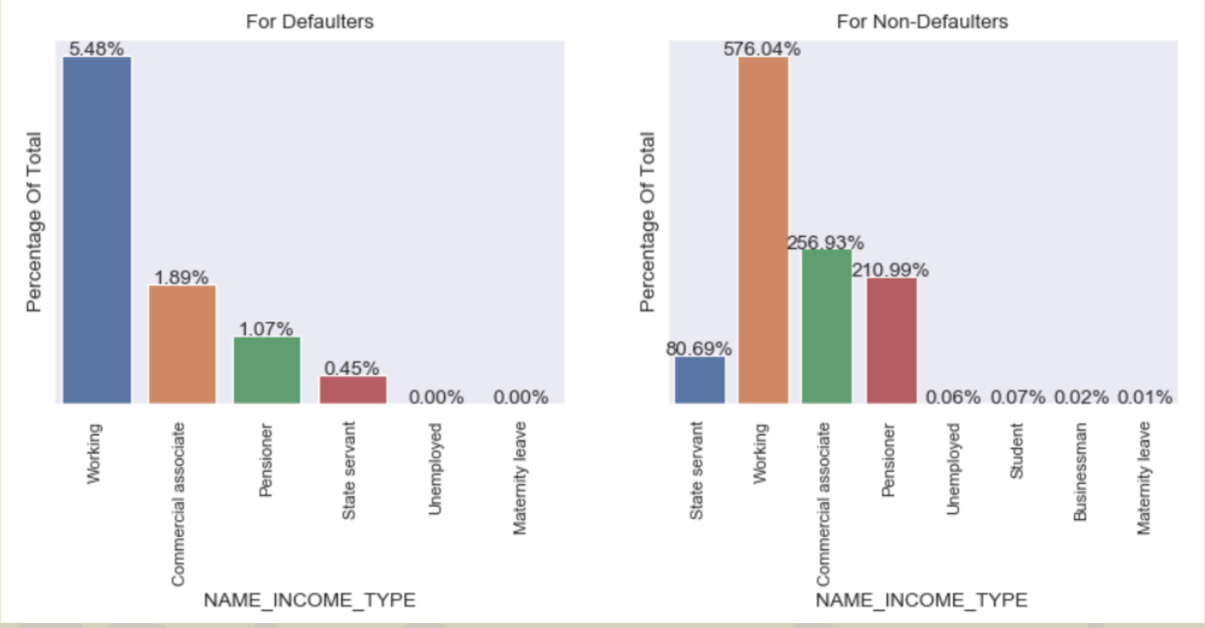
NAME_CONTRACT_TYPE: Larger number of applicants are interested in cash loan and found that the percentage of defaulters are high in Cash loan as compared to % of non-defaulters.



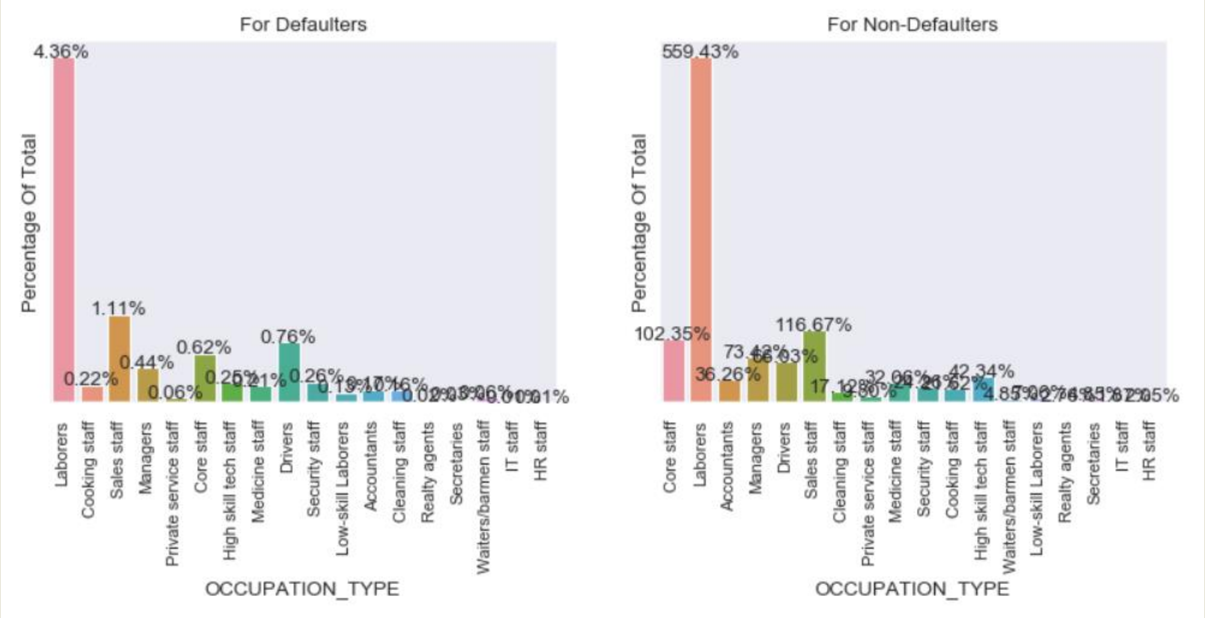
CODE_GENDER: Females are defaulters as compared to Males. Also, as a whole there are more female applicants than male



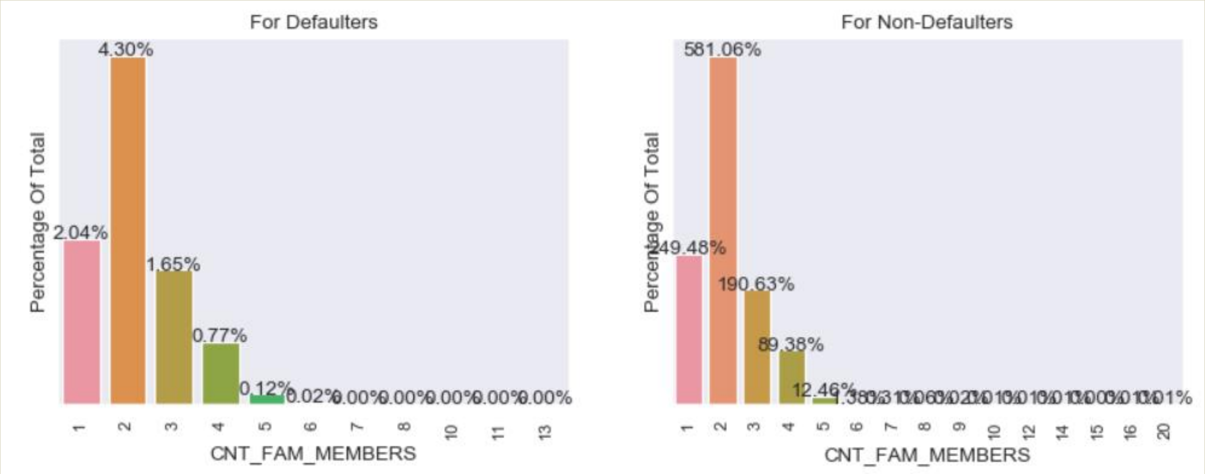
NAME_INCOME_TYPE: Working Professionals have greater ratio/chance of missing loan payment where as other income type has lesser trend of defaulting the loan



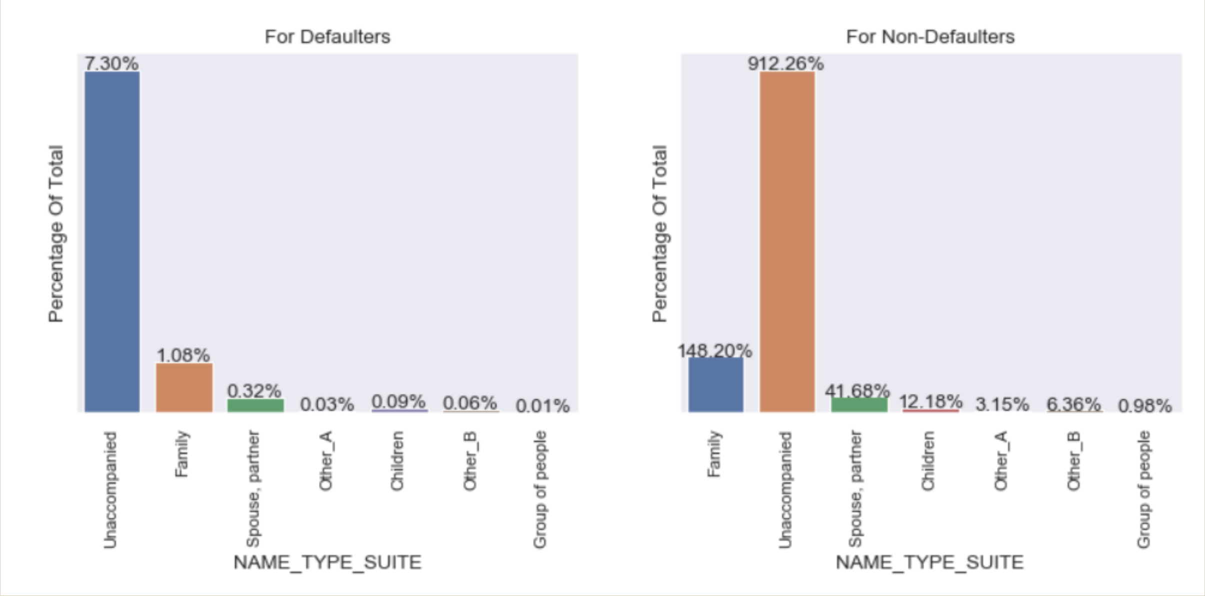
OCCUPATION_TYPE: The percentage of paying or not paying the loan amount are probably same for those working as labour. And the core staff people are more in no. of Non-Defaulter than Defaulters.



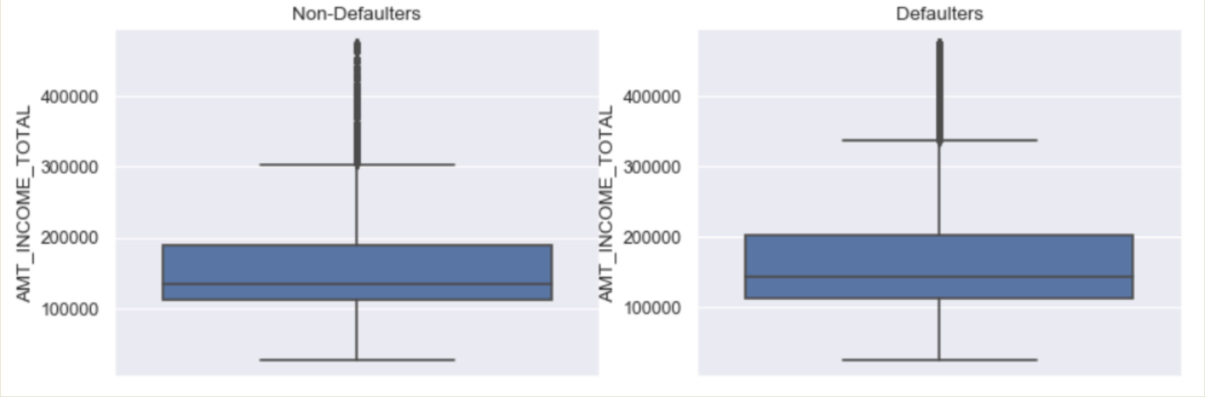
CNT_FAM_MEMBERS: Similar to Children Count most of the applicants are duos and there is a slight chance the loan default in case of 3 family members as per the sample data



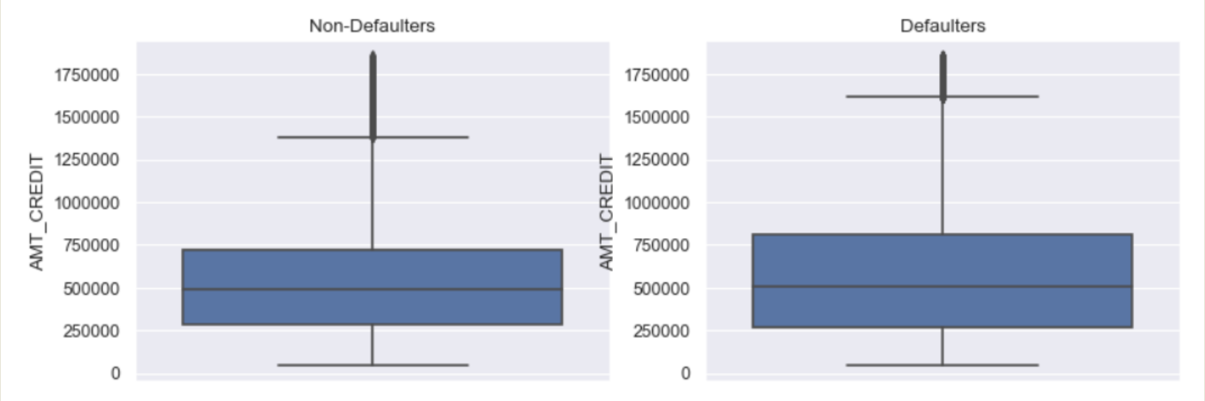
NAME_TYPE_SUITE: The percentage of unaccompanied is same in both the cases i.e. defaulters and non-defaulters.



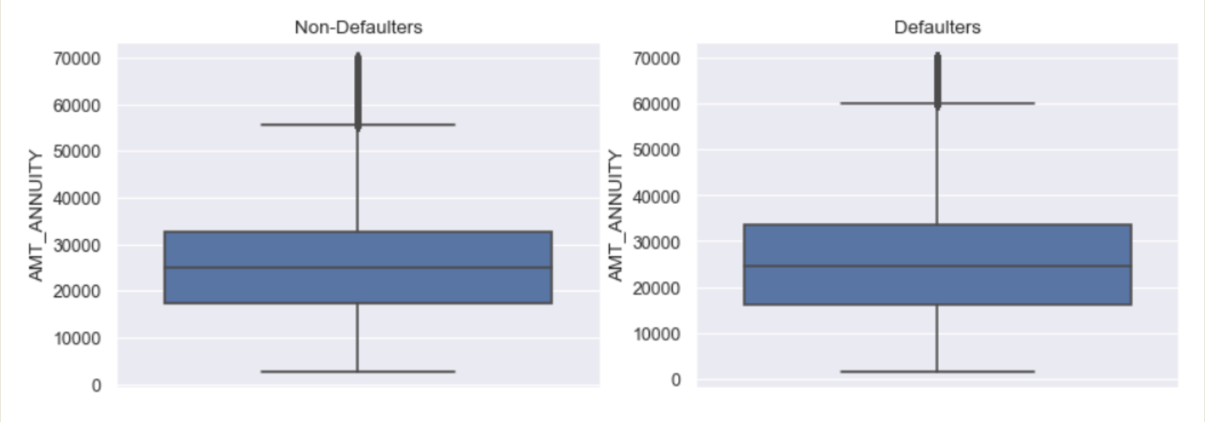
AMT_INCOME_TOTAL: The income total for Non-defaulters is more than Defaulters.



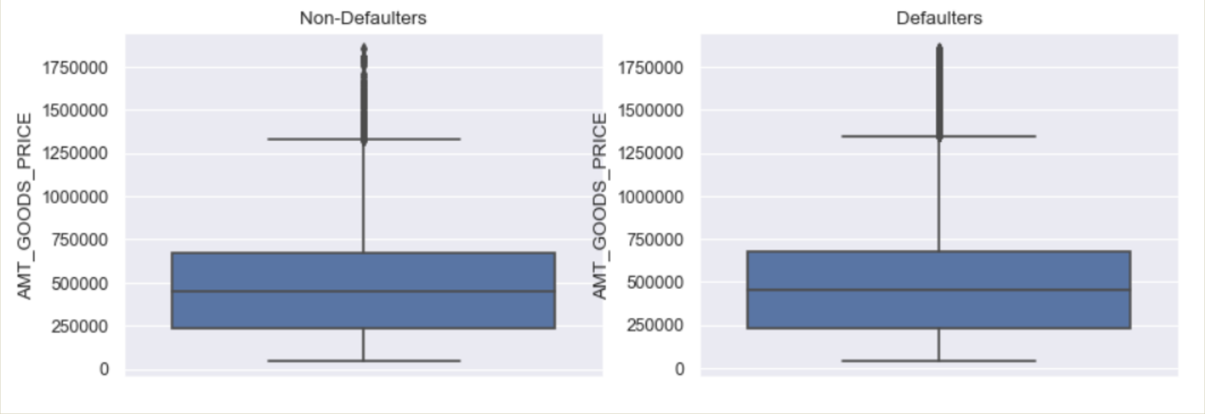
AMT_CREDIT: The total amount of credit for Non-defaulters is more than Defaulters.



AMT_ANNUITY: The amount of annuity Non-defaulters is probably same as Defaulters because it depends on total amount credit.



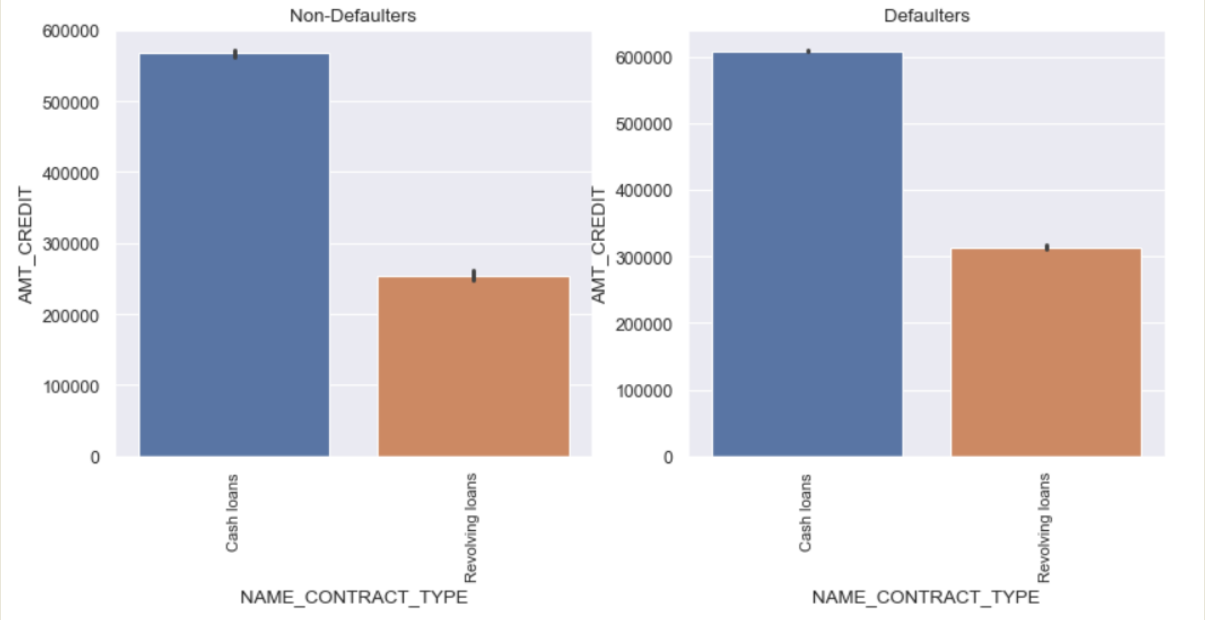
AMT_GOODS_PRICE: The total amount of goods price for non-defaulters is more than Defaulters.

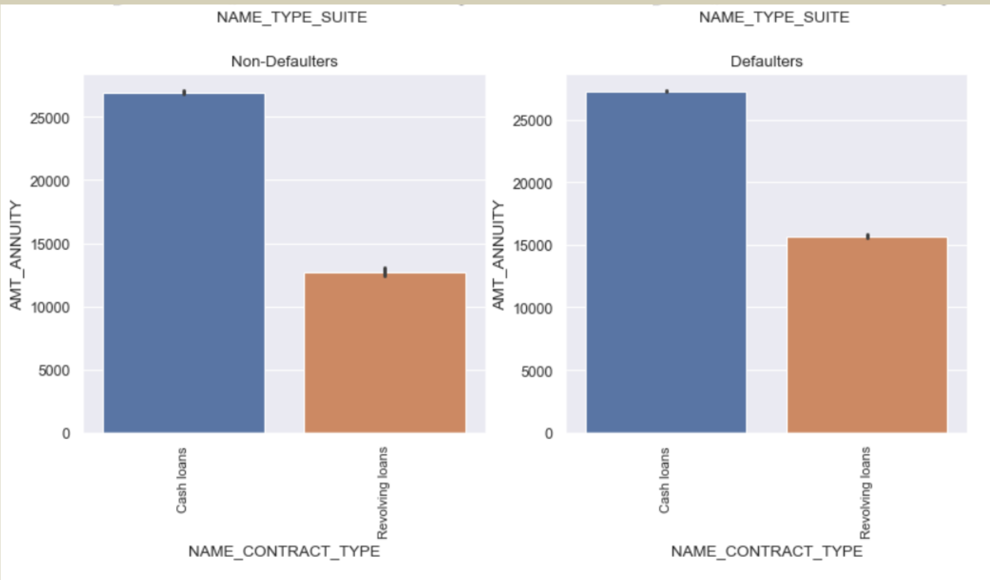


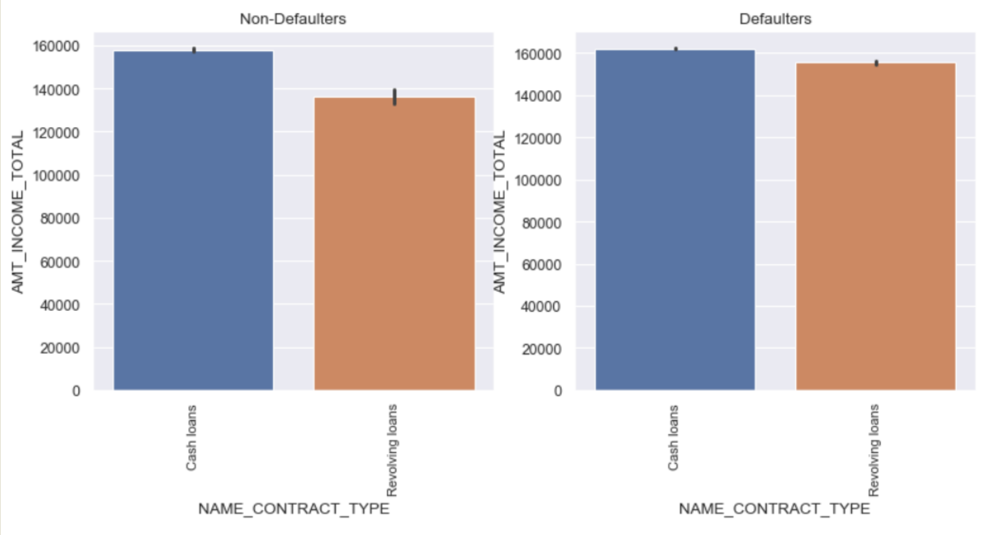
BIVARIATE ANALYSIS

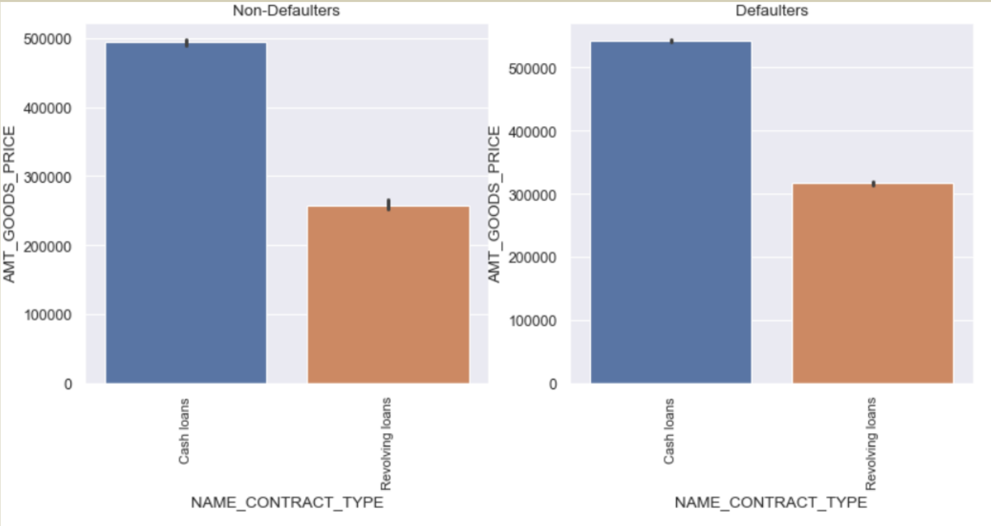
In order to perform Bivariate analysis on categorical variables available in the data set for both 0 and 1, we have chosen few columns in terms of variable and compared against AMT_CREDIT, AMT_ANNUITY, AMT_INCOME_TOTAL & AMT_GOODS_PRICE in data frame and based on Target variable.

NAME_CONTRACT_TYPE:

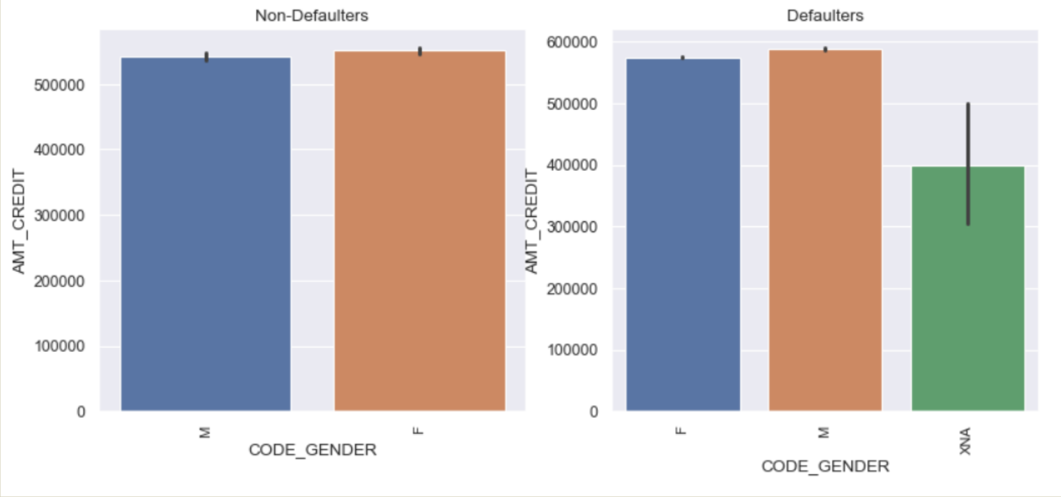


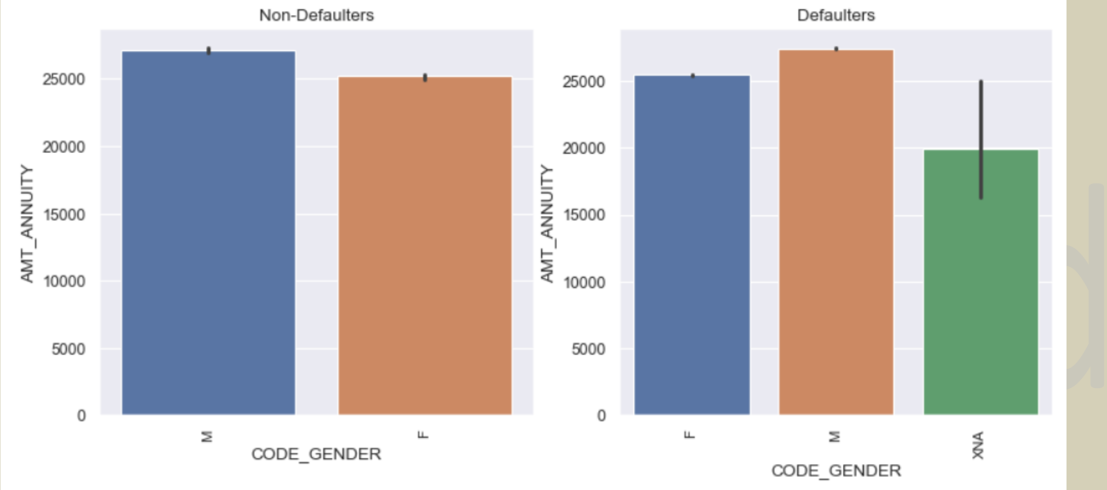

NAME_TYPE_SUITE
Non-Defaulters
AMT_ANNUITY
25000
20000
15000
10000
5000
0
Cash loans
Revolving loans
NAME_CONTRACT_TYPE
NAME_TYPE_SUITE
Defaulters
AMT_ANNUITY
25000
20000
15000
10000
5000
0
Cash loans
Revolving loans
NAME_CONTRACT_TYPE

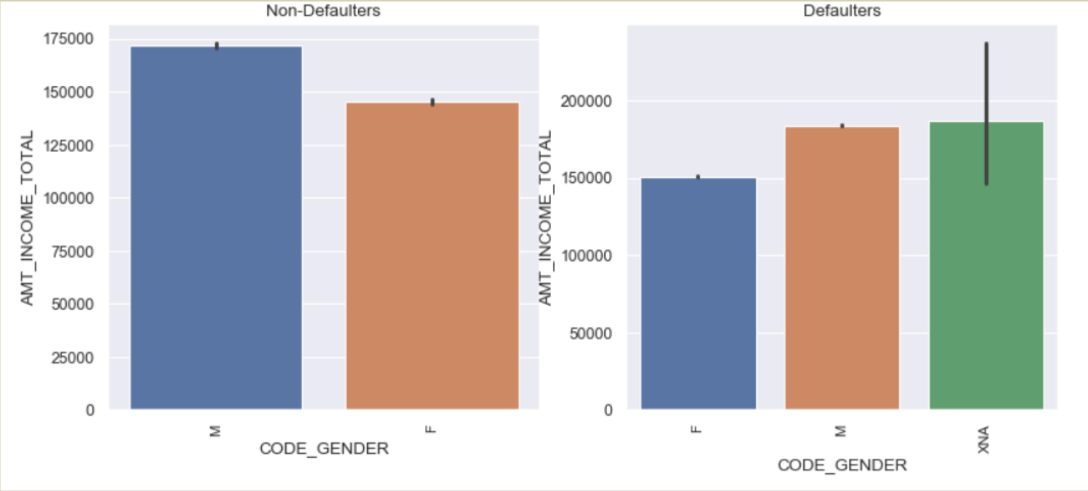

Non-Defaulters
AMT_INCOME_TOTAL
160000
140000
120000
100000
80000
60000
40000
20000
0
Cash loans
Revolving loans
NAME_CONTRACT_TYPE
Defaulters
AMT_INCOME_TOTAL
160000
140000
120000
100000
80000
60000
40000
20000
0
Cash loans
Revolving loans
NAME_CONTRACT_TYPE

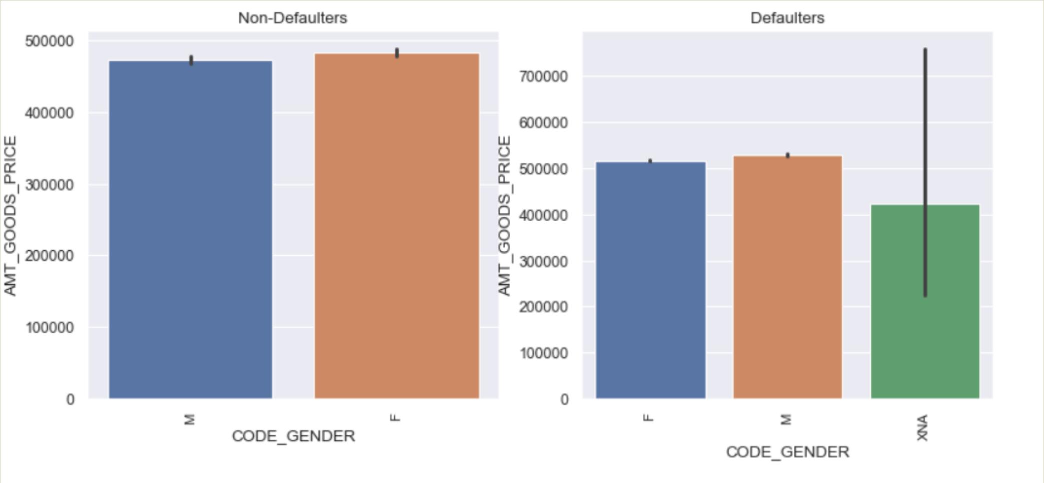

Non-Defaulters
AMT_GOODS_PRICE
500000
400000
300000
200000
100000
0
Cash loans
Revolving loans
NAME_CONTRACT_TYPE
Defaulters
AMT_GOODS_PRICE
500000
400000
300000
200000
100000
0
Cash loans
Revolving loans
NAME_CONTRACT_TYPE

CODE_GENDER:

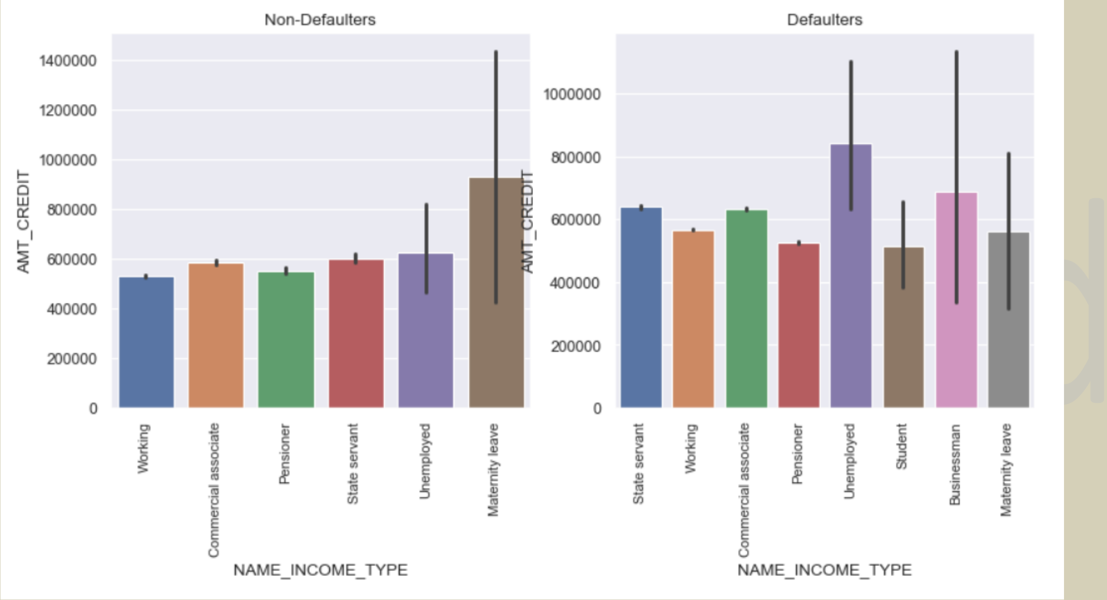

Non-Defaulters
Defaulters
AMT_CREDIT
0
100000
200000
300000
400000
500000
600000
M
F
XNA
CODE_GENDER

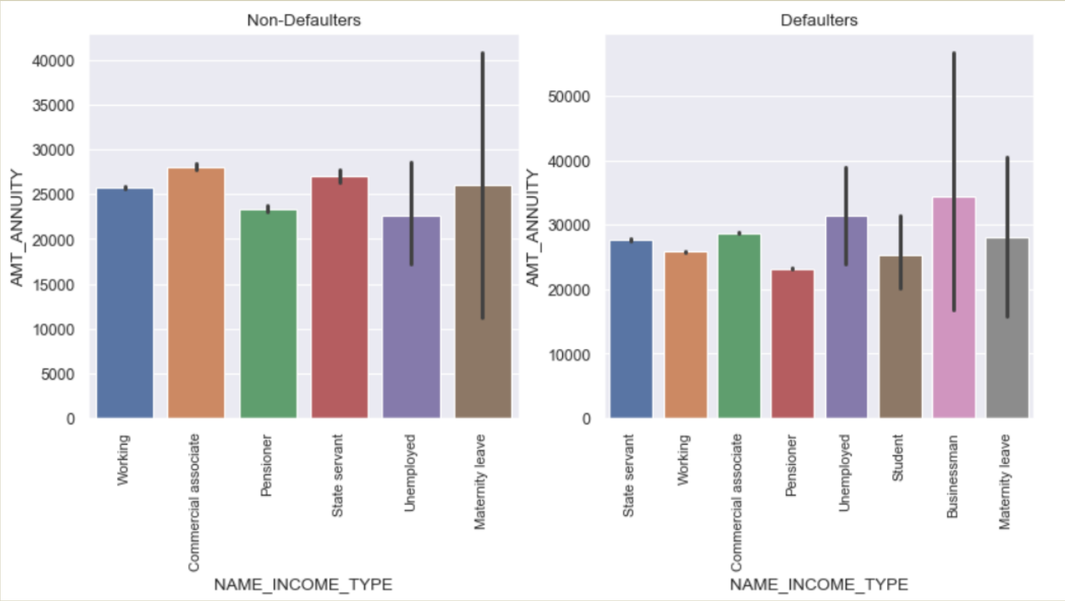

Non-Defaulters
Defaulters
AMT_ANNUITY
0
5000
10000
15000
20000
25000
M
F
XNA
CODE_GENDER

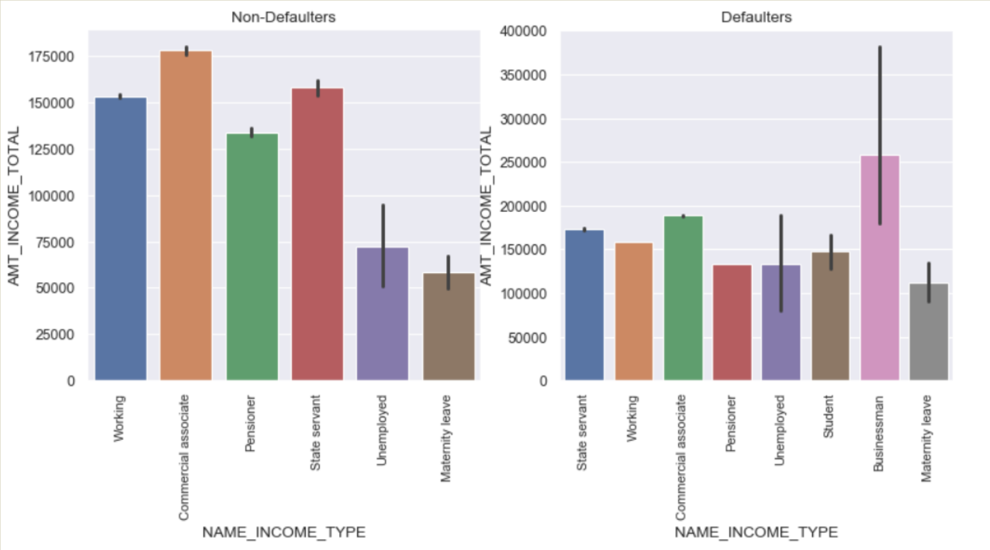

Non-Defaulters
Defaulters
AMT_INCOME_TOTAL
0
25000
50000
75000
100000
125000
150000
175000
200000
M
F
XNA
CODE_GENDER

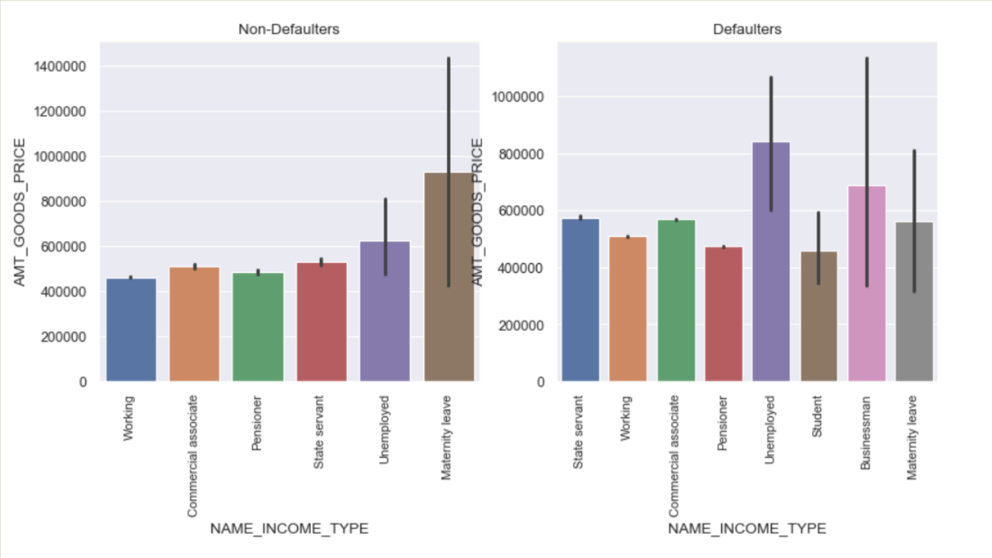


NAME_INCOME_TYPE:

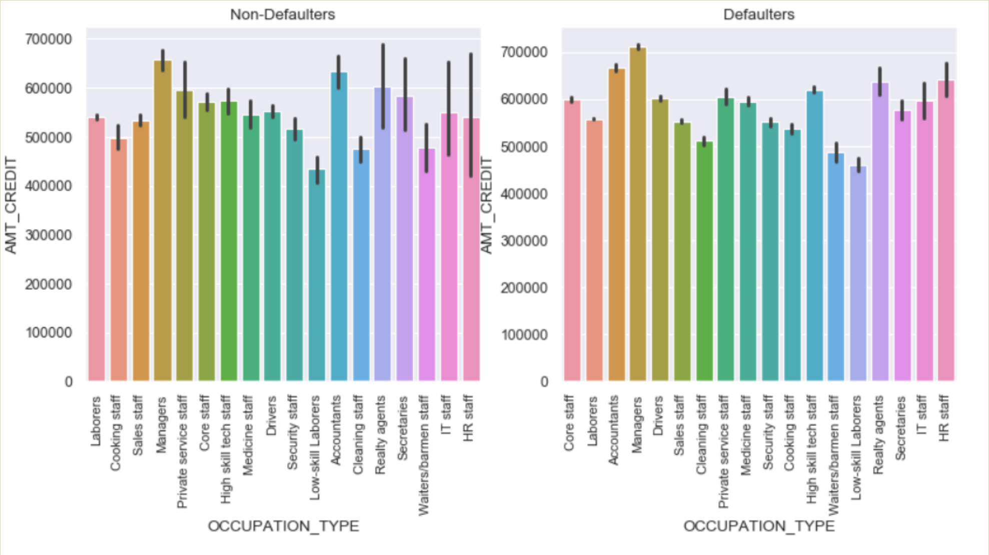


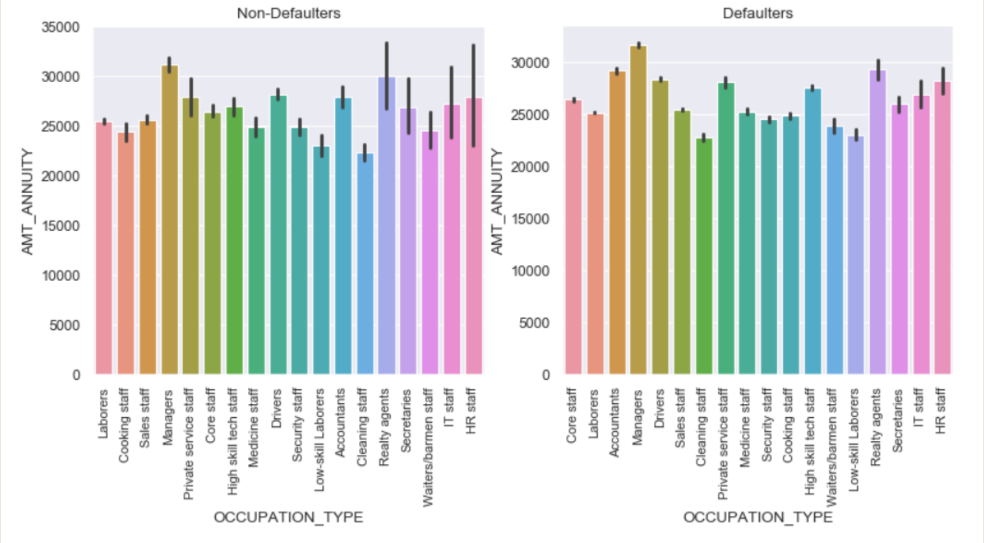


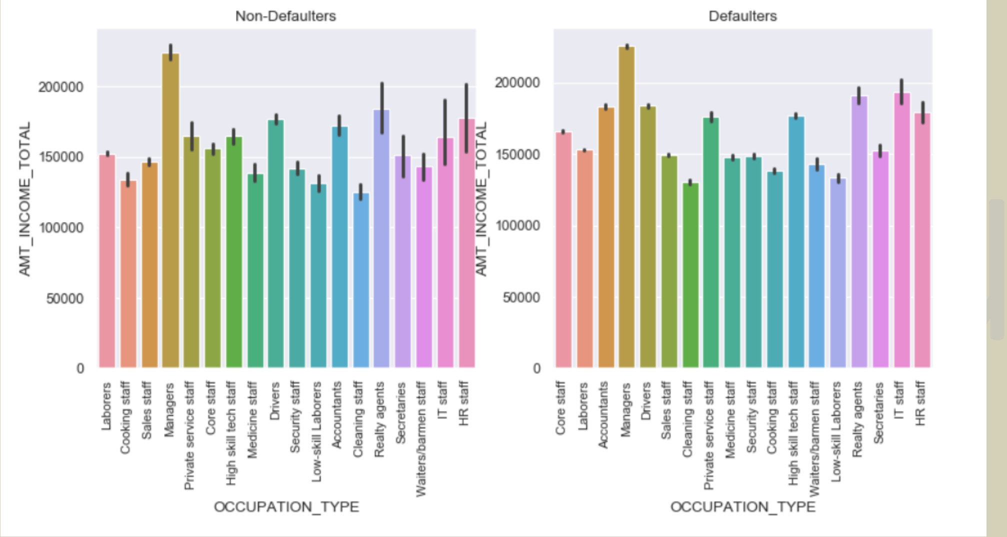


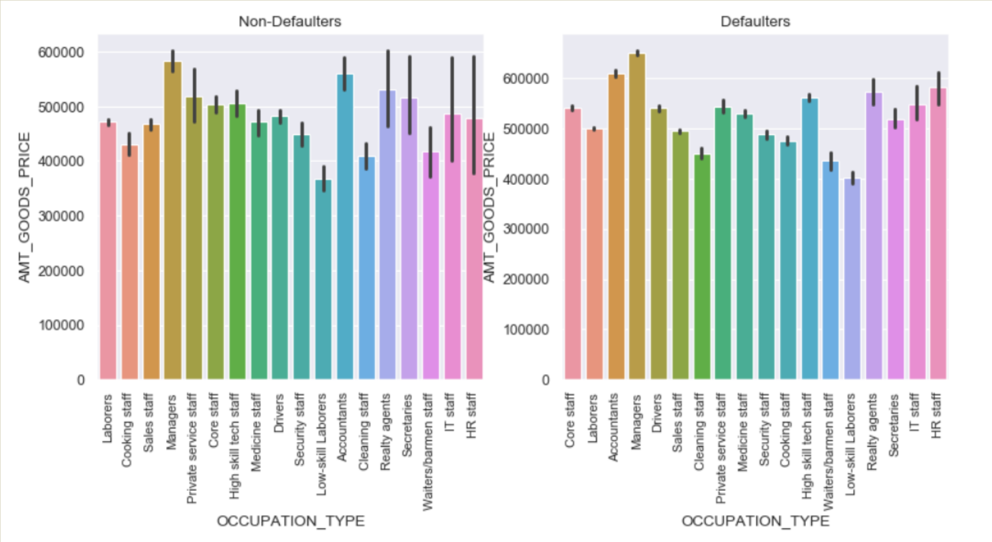


OCCUPATION_TYPE:

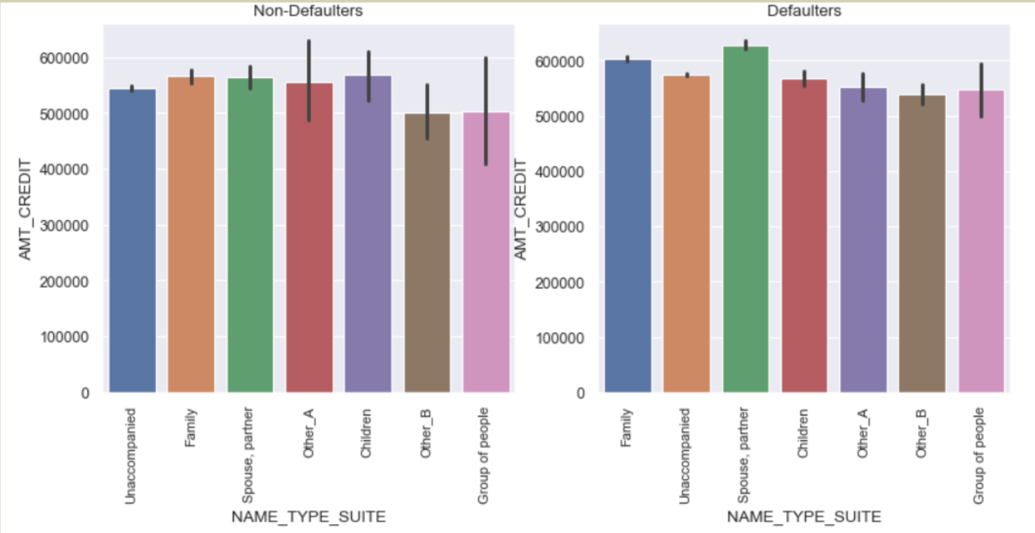


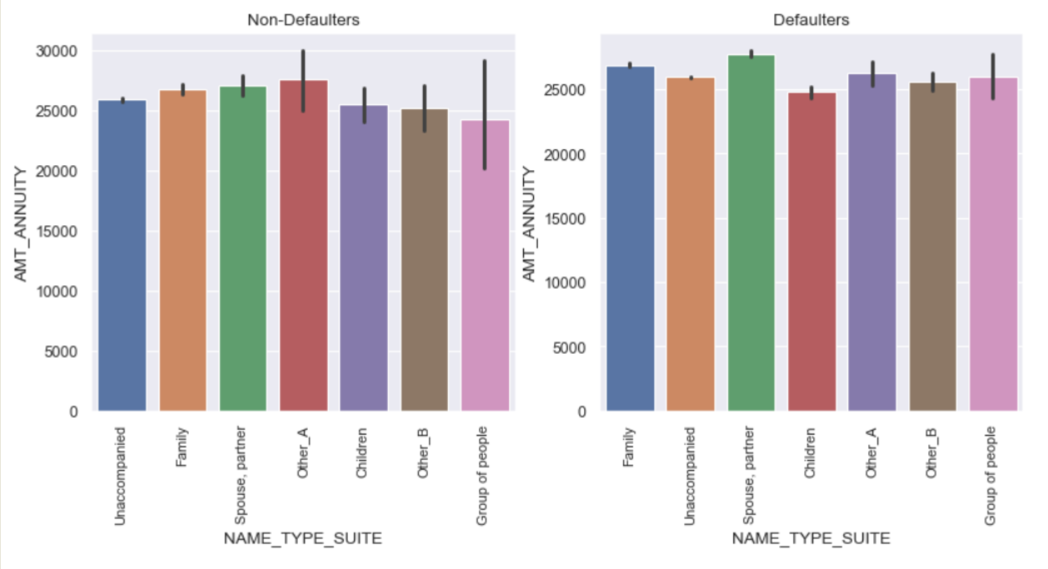

Non-Defaulters
AMT_ANNUITY
35000
30000
25000
20000
15000
10000
5000
0
Laborers
Cooking staff
Sales staff
Managers
Private service staff
Core staff
High skill tech staff
Medicine staff
Drivers
Security staff
Low-skill Laborers
Accountants
Cleaning staff
Realty agents
Secretaries
Waiters/barmen staff
IT staff
HR staff
OCCUPATION_TYPE
Defaulters
AMT_ANNUITY
30000
25000
20000
15000
10000
5000
0
Core staff
Laborers
Accountants
Managers
Drivers
Sales staff
Cleaning staff
Private service staff
Medicine staff
Security staff
Cooking staff
High skill tech staff
Waiters/barmen staff
Low-skill Laborers
Realty agents
Secretaries
IT staff
HR staff
OCCUPATION_TYPE

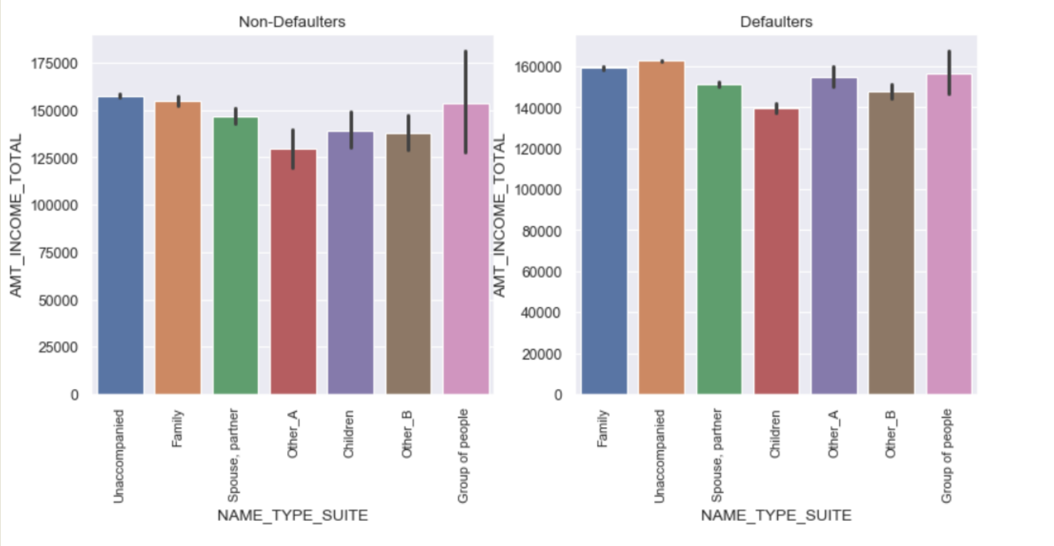

Non-Defaulters
AMT_INCOME_TOTAL
200000
150000
100000
50000
0
Laborers
Cooking staff
Sales staff
Managers
Private service staff
Core staff
High skill tech staff
Medicine staff
Drivers
Security staff
Low-skill Laborers
Accountants
Cleaning staff
Realty agents
Secretaries
Waiters/barmen staff
IT staff
HR staff
OCCUPATION_TYPE
Defaulters
AMT_INCOME_TOTAL
200000
150000
100000
50000
0
Core staff
Laborers
Accountants
Managers
Drivers
Sales staff
Cleaning staff
Private service staff
Medicine staff
Security staff
Cooking staff
High skill tech staff
Waiters/barmen staff
Low-skill Laborers
Realty agents
Secretaries
IT staff
HR staff
OCCUPATION_TYPE

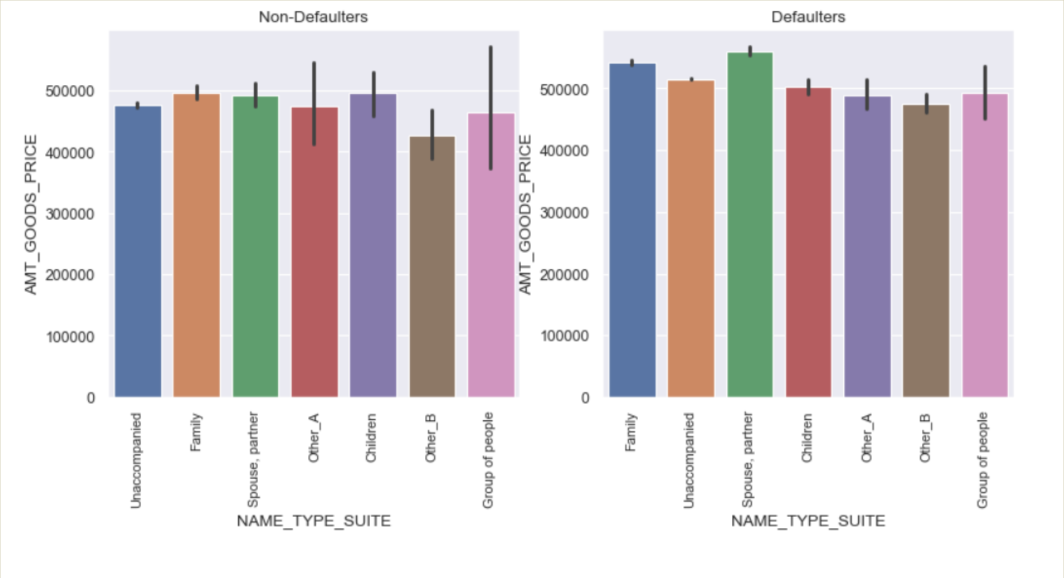

Non-Defaulters
AMT_GOODS_PRICE
600000
500000
400000
300000
200000
100000
0
Laborers
Cooking staff
Sales staff
Managers
Private service staff
Core staff
High skill tech staff
Medicine staff
Drivers
Security staff
Low-skill Laborers
Accountants
Cleaning staff
Realty agents
Secretaries
Waiters/barmen staff
IT staff
HR staff
OCCUPATION_TYPE
Defaulters
AMT_GOODS_PRICE
600000
500000
400000
300000
200000
100000
0
Core staff
Laborers
Accountants
Managers
Drivers
Sales staff
Cleaning staff
Private service staff
Medicine staff
Security staff
Cooking staff
High skill tech staff
Waiters/barmen staff
Low-skill Laborers
Realty agents
Secretaries
IT staff
HR staff
OCCUPATION_TYPE

NAME_TYPE_SUITE:







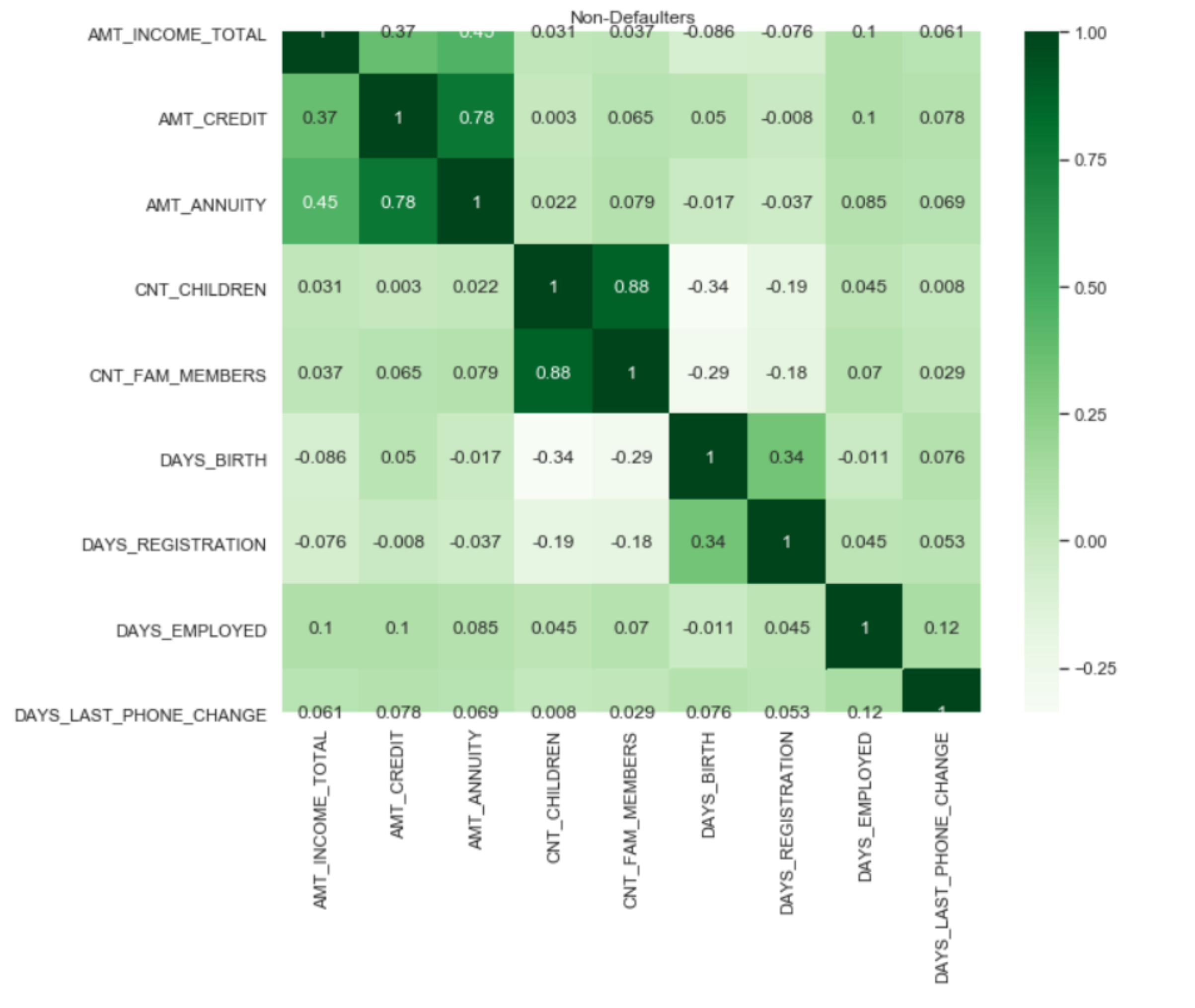


We have plotted graphs for various categorical variables and measures and some of the analysis are mentioned below

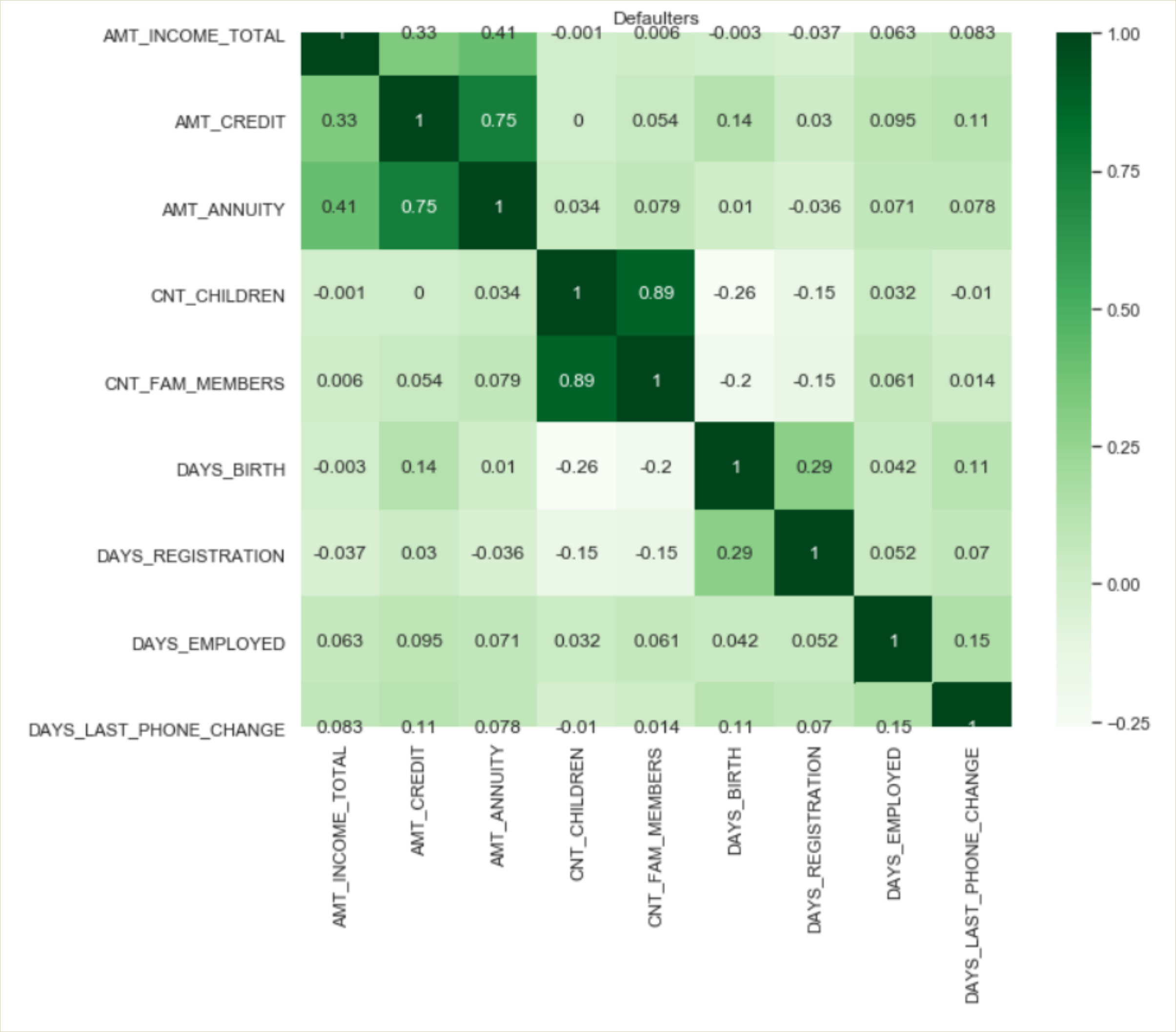
- **Credit Amount and Family Type: As per the Data above married couples have with higher credit amount mean are less likely to default and their credit amount is higher compared to other non-defaulters so it is safe to handover loan to that category.**
- **Annual income and Family Type: Overall people with higher income are less likely to default but if we compare the family type also, we found that defaulters with higher income values re group of people**
- **Work type and Credit Amount: Unemployed persons are having higher average of non-defaulter. as per above graph the are very less likely to default and have highest average of credit as non-default.**
- **Occupation type and Credit Amount: People of manager occupation having higher no. of non-defaulter than Defaulter, and People of Accountants occupation having higher no. in Defaulter than non-defaulter**

➙ Finding the top 10 correlation by segmenting the data-frame w.r.t target variable

Correlation Plot for non-defaulters: -



Correlation Plot for defaulters: -



As per the correlation matrix and heatmap for both non-defaulters and Defaulters

Correlation between Family Member and Count of Children have highest correlation Value as usual.

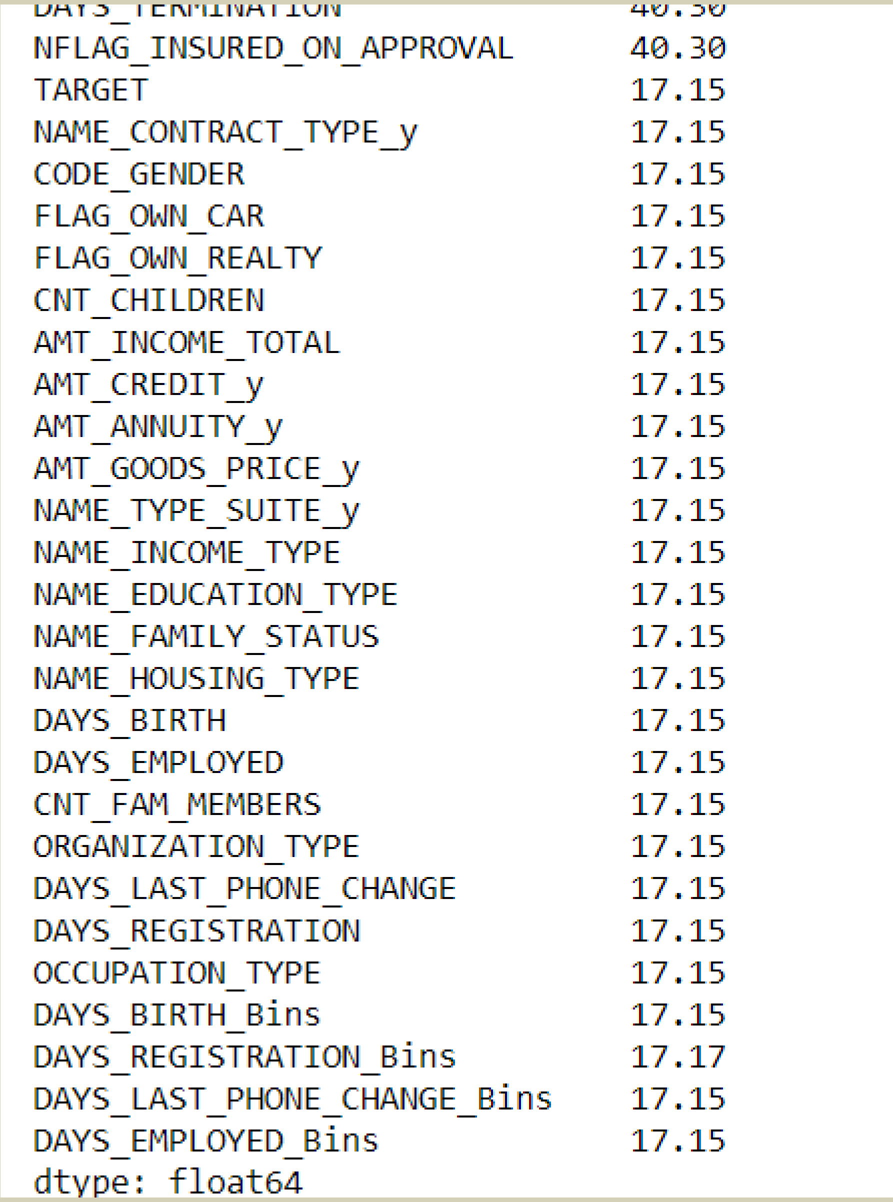
Secondly the Annuity amount and Credit amount are also highly correlated for both default and non-default cases.

After that Days after registration and Days birth (Age) are have highly correlated values.

→Previous Application Dataset

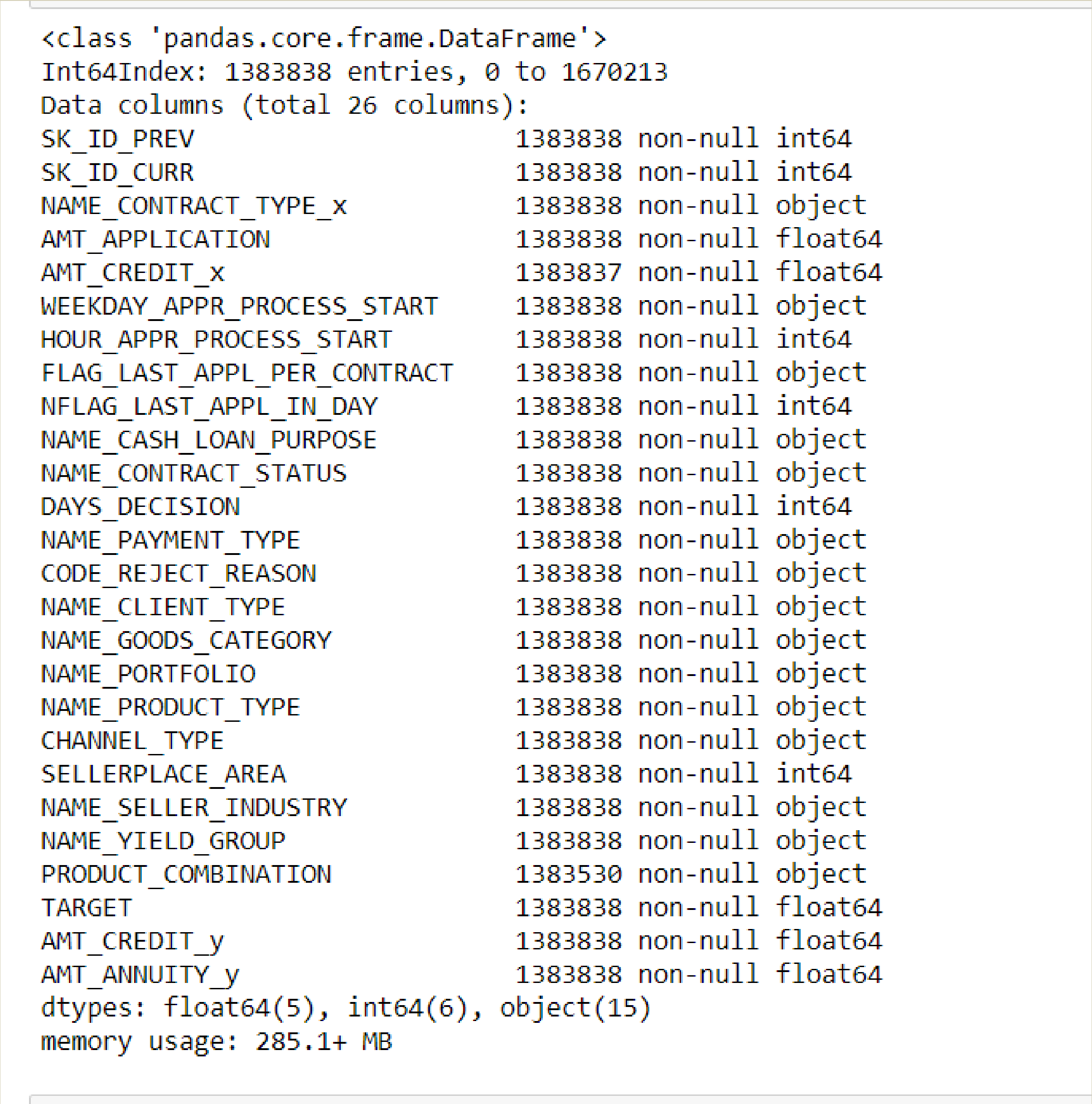
- Merged the datasets Application and previous on SK_ID_CURR (left join)
- Took sample out of the output data
- Started visualisation

Figured out the null columns



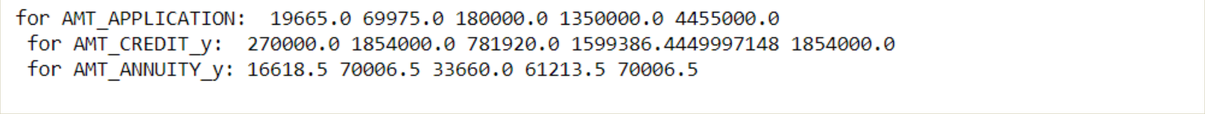
```
DAYS_TERMINATION                  40.30
NFLAG_INSURED_ON_APPROVAL         40.30
TARGET                            17.15
NAME_CONTRACT_TYPE_y              17.15
CODE_GENDER                       17.15
FLAG_OWN_CAR                      17.15
FLAG_OWN_REALTY                   17.15
CNT_CHILDREN                      17.15
AMT_INCOME_TOTAL                  17.15
AMT_CREDIT_y                      17.15
AMT_ANNUITY_y                     17.15
AMT_GOODS_PRICE_y                 17.15
NAME_TYPE_SUITE_y                 17.15
NAME_INCOME_TYPE                  17.15
NAME_EDUCATION_TYPE               17.15
NAME_FAMILY_STATUS                17.15
NAME_HOUSING_TYPE                 17.15
DAYS_BIRTH                        17.15
DAYS_EMPLOYED                     17.15
CNT_FAM_MEMBERS                   17.15
ORGANIZATION_TYPE                 17.15
DAYS_LAST_PHONE_CHANGE            17.15
DAYS_REGISTRATION                 17.15
OCCUPATION_TYPE                   17.15
DAYS_BIRTH_Bins                   17.15
DAYS_REGISTRATION_Bins            17.17
DAYS_LAST_PHONE_CHANGE_Bins       17.15
DAYS_EMPLOYED_Bins                17.15
dtype: float64
```

Target Column had large number of missing values, so it is useful to drop the rows

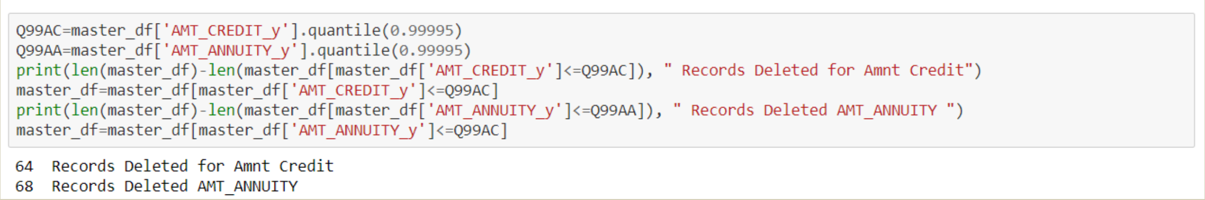


```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1383838 entries, 0 to 1670213
Data columns (total 26 columns):
SK_ID_PREV                     1383838 non-null int64
SK_ID_CURR                     1383838 non-null int64
NAME_CONTRACT_TYPE_x           1383838 non-null object
AMT_APPLICATION                1383838 non-null float64
AMT_CREDIT_x                   1383837 non-null float64
WEEKDAY_APPR_PROCESS_START     1383838 non-null object
HOUR_APPR_PROCESS_START        1383838 non-null int64
FLAG_LAST_APPL_PER_CONTRACT    1383838 non-null object
NFLAG_LAST_APPL_IN_DAY         1383838 non-null int64
NAME_CASH_LOAN_PURPOSE         1383838 non-null object
NAME_CONTRACT_STATUS           1383838 non-null object
DAYS_DECISION                  1383838 non-null int64
NAME_PAYMENT_TYPE              1383838 non-null object
CODE_REJECT_REASON             1383838 non-null object
NAME_CLIENT_TYPE               1383838 non-null object
NAME_GOODS_CATEGORY            1383838 non-null object
NAME_PORTFOLIO                 1383838 non-null object
NAME_PRODUCT_TYPE              1383838 non-null object
CHANNEL_TYPE                   1383838 non-null object
SELLERPLACE_AREA               1383838 non-null int64
NAME_SELLER_INDUSTRY           1383838 non-null object
NAME_YIELD_GROUP               1383838 non-null object
PRODUCT_COMBINATION            1383530 non-null object
TARGET                         1383838 non-null float64
AMT_CREDIT_y                   1383838 non-null float64
AMT_ANNUITY_y                  1383838 non-null float64
dtypes: float64(5), int64(6), object(15)
memory usage: 285.1+ MB
```

Found the outliers and had to drop those values



```
for AMT_APPLICATION:  19665.0 69975.0 180000.0 1350000.0 4455000.0
 for AMT_CREDIT_y:  270000.0 1854000.0 781920.0 1599386.4449997148 1854000.0
 for AMT_ANNUITY_y: 16618.5 70006.5 33660.0 61213.5 70006.5
```

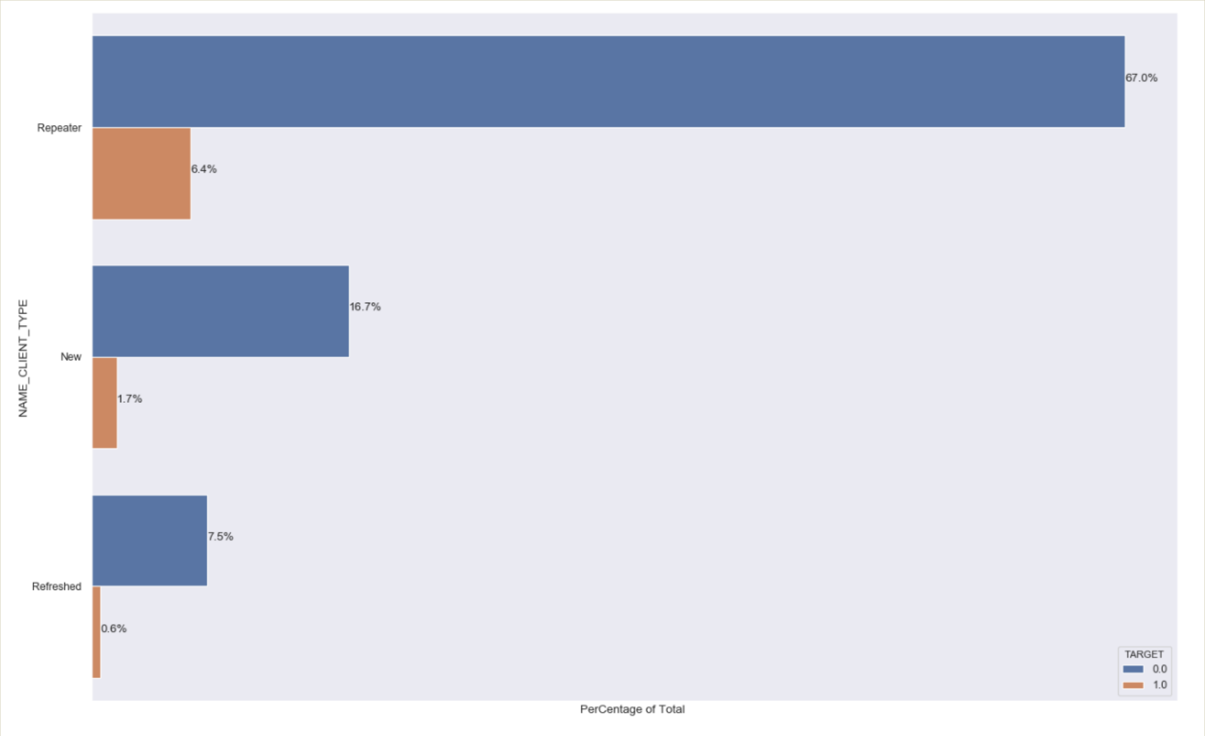


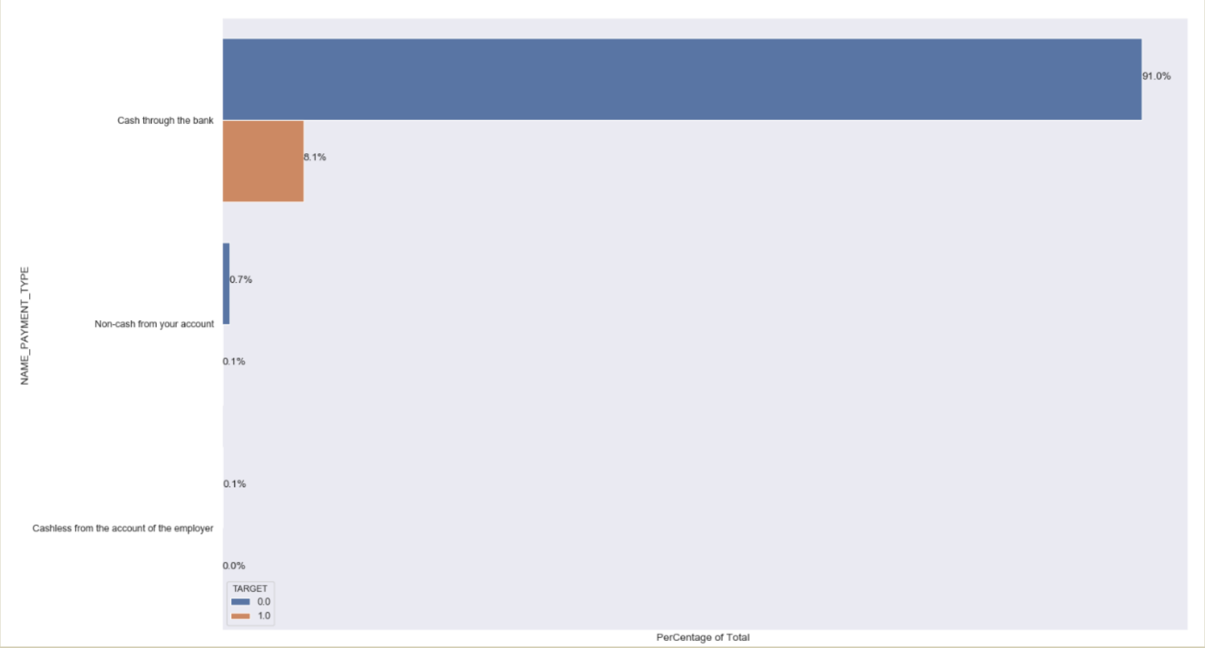
```
Q99AC=master_df['AMT_CREDIT_y'].quantile(0.99995)
Q99AA=master_df['AMT_ANNUITY_y'].quantile(0.99995)
print(len(master_df)-len(master_df[master_df['AMT_CREDIT_y']<=Q99AC]), " Records Deleted for Amnt Credit")
master_df=master_df[master_df['AMT_CREDIT_y']<=Q99AC]
print(len(master_df)-len(master_df[master_df['AMT_ANNUITY_y']<=Q99AA]), " Records Deleted AMT_ANNUITY ")
master_df=master_df[master_df['AMT_ANNUITY_y']<=Q99AC]
```

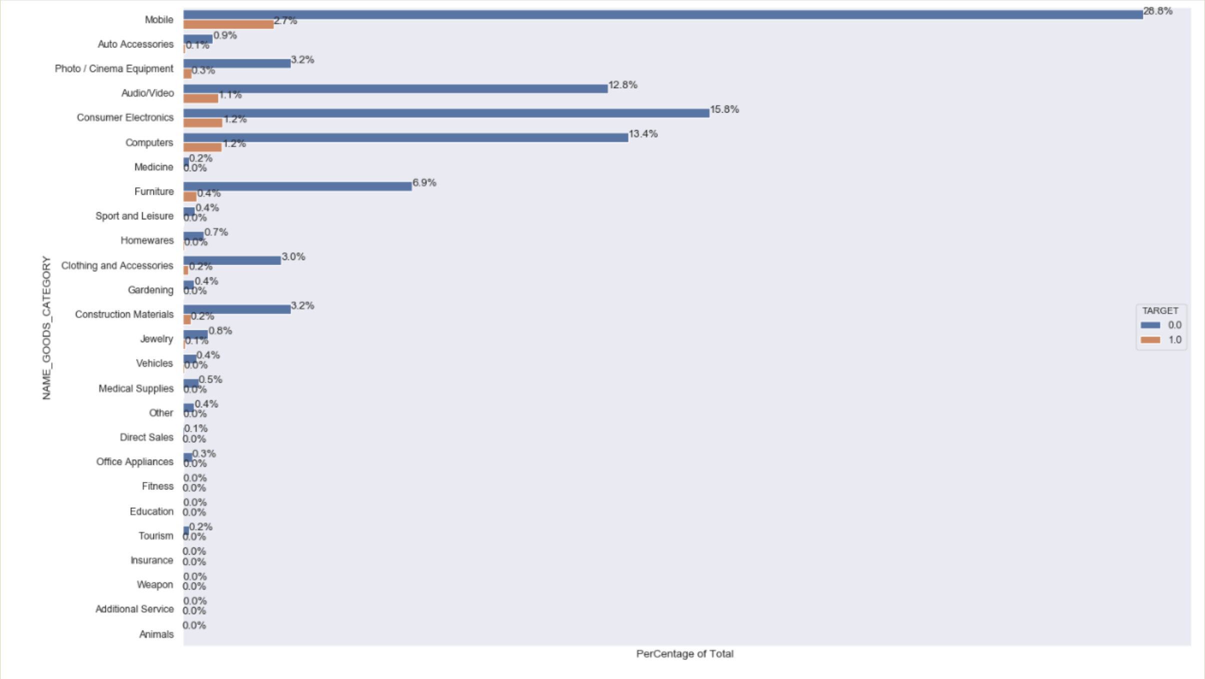
```
64  Records Deleted for Amnt Credit
68  Records Deleted AMT_ANNUITY
```

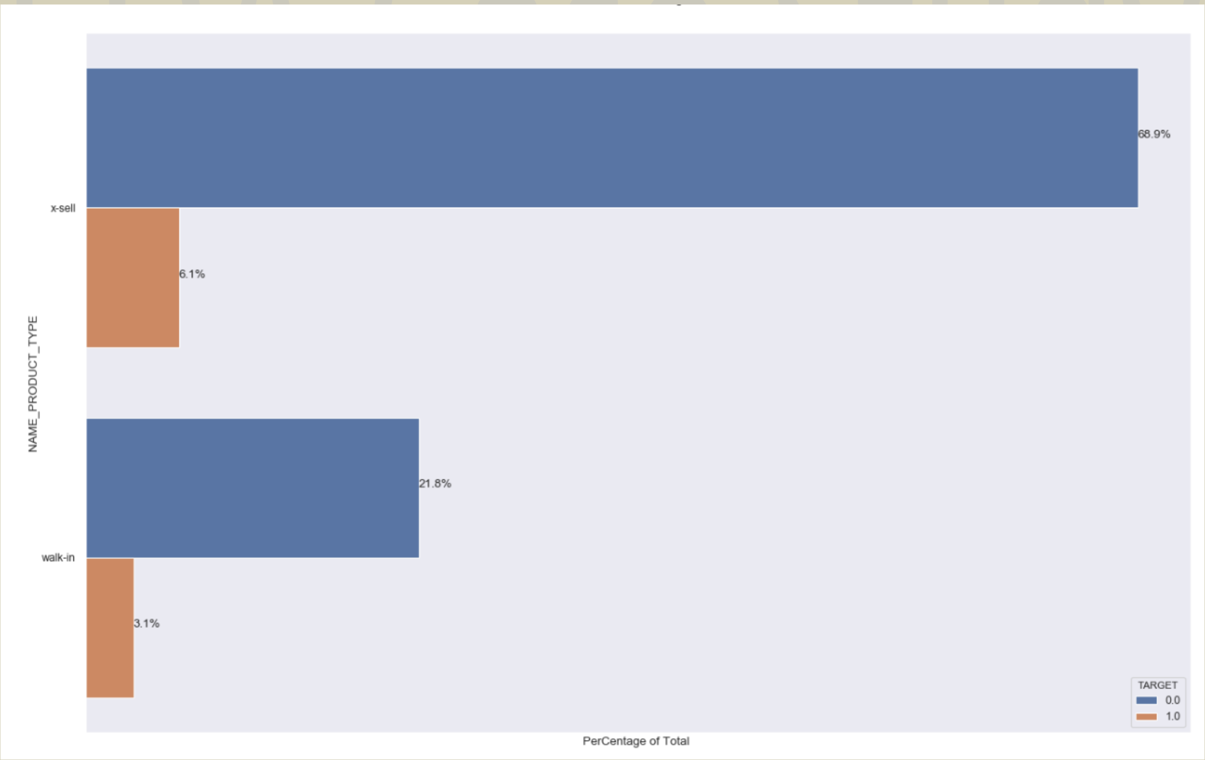
UNIVARIATE ANALYSIS

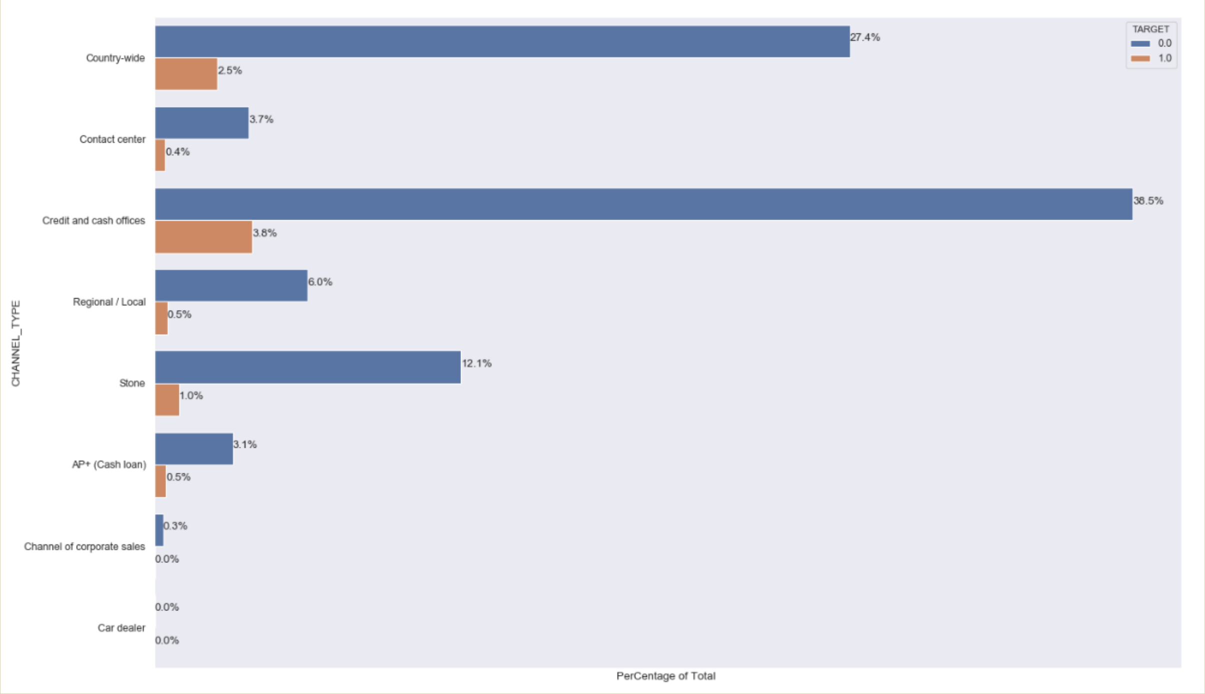
Univariate Analysis for Categorical Columns

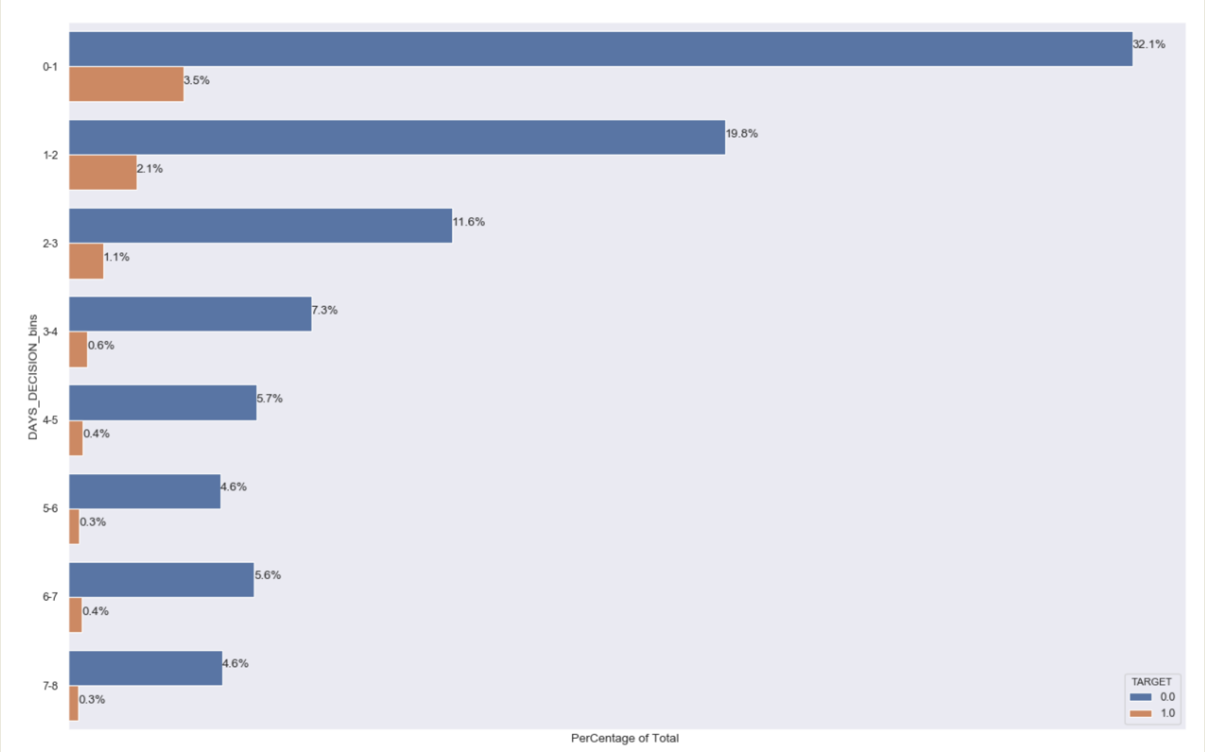




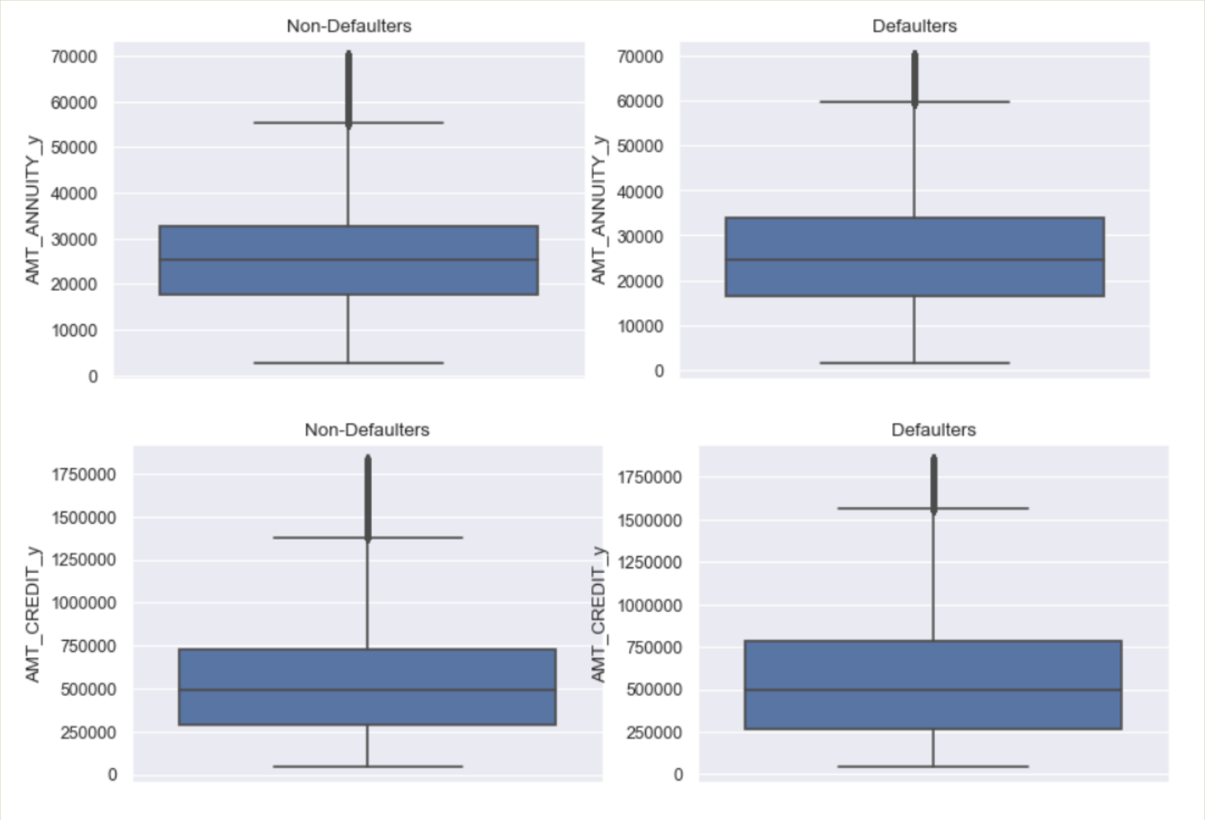

NAME_GOODS_CATEGORY
Mobile
Auto Accessories
Photo / Cinema Equipment
Audio/Video
Consumer Electronics
Computers
Medicine
Furniture
Sport and Leisure
Homewares
Clothing and Accessories
Gardening
Construction Materials
Jewelry
Vehicles
Medical Supplies
Other
Direct Sales
Office Appliances
Fitness
Education
Tourism
Insurance
Weapon
Additional Service
Animals
28.8%
2.7%
0.9%
0.1%
3.2%
0.3%
12.8%
1.1%
15.8%
1.2%
13.4%
1.2%
0.2%
0.0%
6.9%
0.4%
0.4%
0.0%
0.7%
0.0%
3.0%
0.2%
0.4%
0.0%
3.2%
0.2%
0.8%
0.1%
0.4%
0.0%
0.5%
0.0%
0.4%
0.0%
0.1%
0.0%
0.3%
0.0%
0.0%
0.0%
0.0%
0.0%
0.2%
0.0%
0.0%
0.0%
0.0%
0.0%
0.0%
0.0%
0.0%
PerCentage of Total
TARGET
0.0
1.0


NAME_PRODUCT_TYPE
x-sell
walk-in
68.9%
6.1%
21.8%
3.1%
PerCentage of Total
TARGET
0.0
1.0


TARGET
0.0
1.0
CHANNEL_TYPE
Country-wide
27.4%
2.5%
Contact center
3.7%
0.4%
Credit and cash offices
38.5%
3.8%
Regional / Local
6.0%
0.5%
Stone
12.1%
1.0%
AP+ (Cash loan)
3.1%
0.5%
Channel of corporate sales
0.3%
0.0%
Car dealer
0.0%
0.0%
PerCentage of Total


DAYS_DECISION_bins
0-1
32.1%
3.5%
1-2
19.8%
2.1%
2-3
11.6%
1.1%
3-4
7.3%
0.6%
4-5
5.7%
0.4%
5-6
4.6%
0.3%
6-7
5.6%
0.4%
7-8
4.6%
0.3%
TARGET
0.0
1.0
PerCentage of Total

Univariate Analysis of Numerical Columns



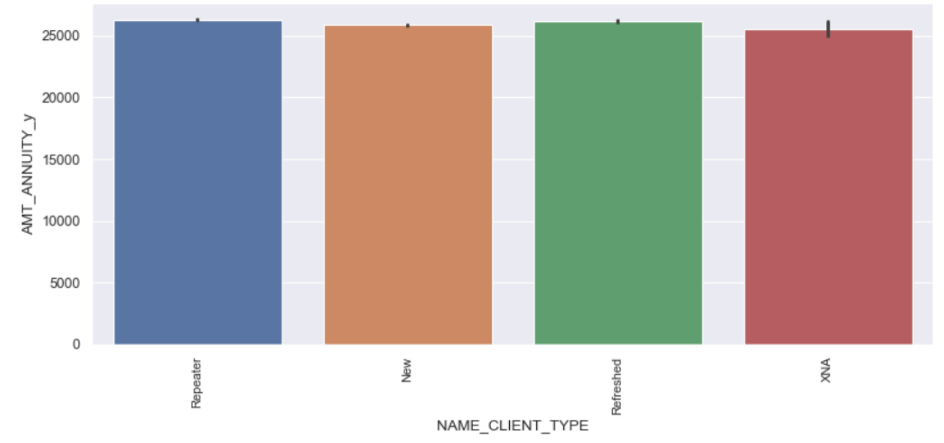
Results of Univariate Analysis of Categorical variables

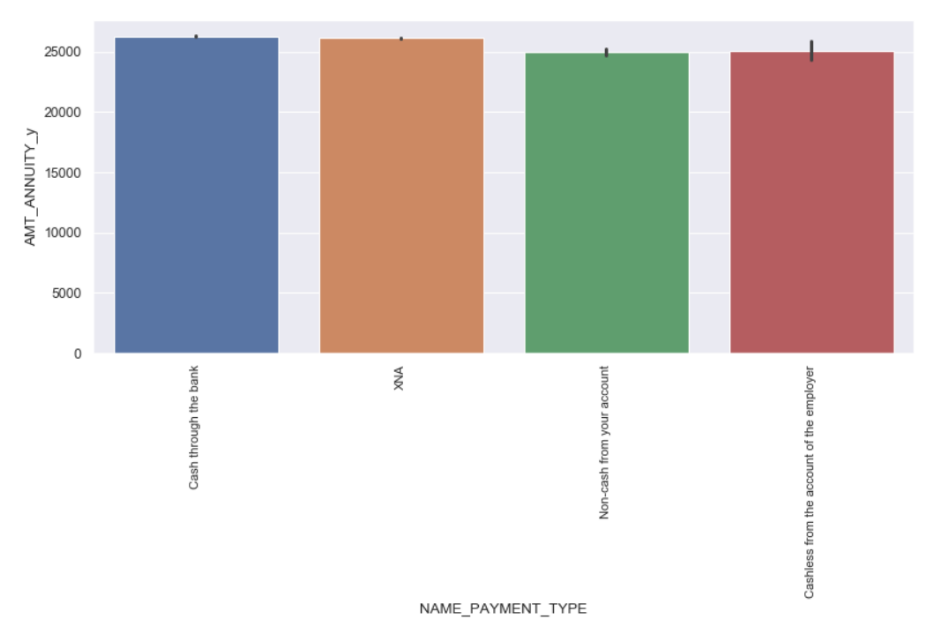
- Client type: Most the repeater clients have applied for loan.
- Payment type: Most of the Peoples are taking credit amount taking form bank by cash only.
- Goods type: People who are applying loan for mobile are more in no. than others.
- Channel type: From the above observation it is observed that most of the selling's are done by Credit and cash officers, because most no. of peoples are coming from that channel only.

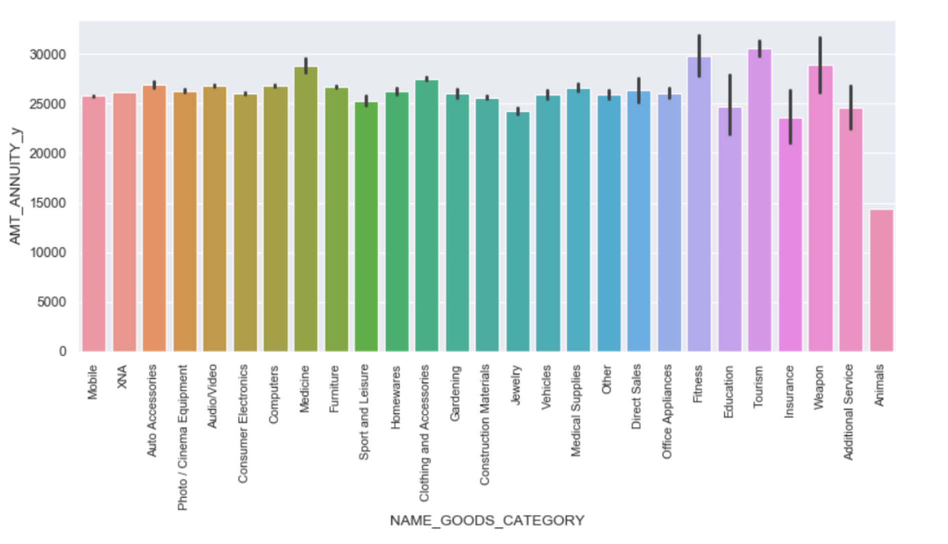
Results of Univariate Analysis of Continuous Variables

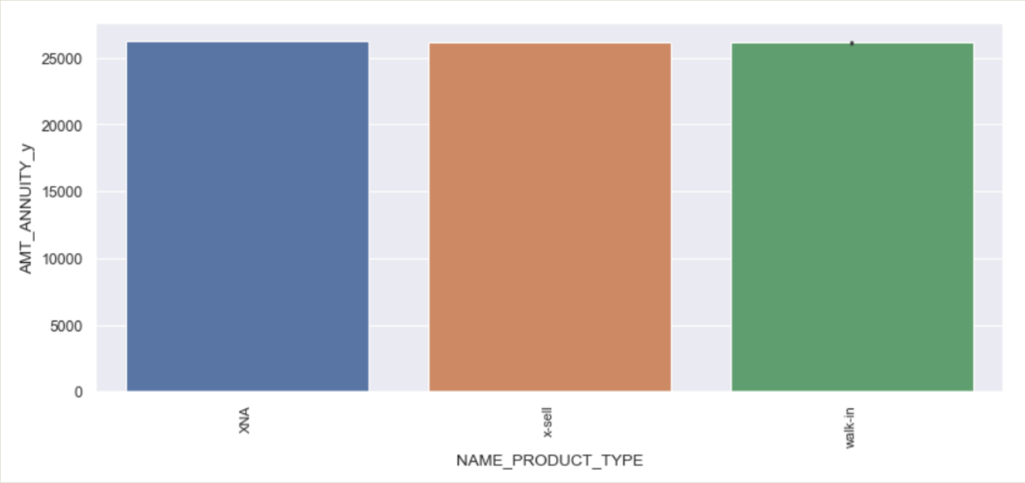
Annuity Amount: There is no significant change between Defaulters and Non-defaulters for annuity amount because of the total credit amounts are likely to be same for both.

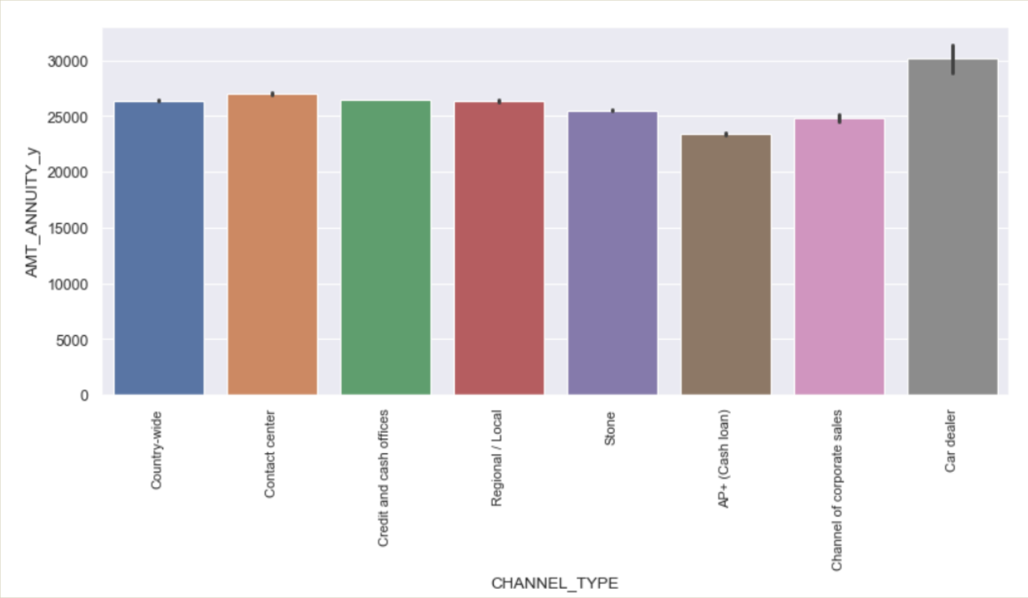
BIVARIATE ANALYSIS

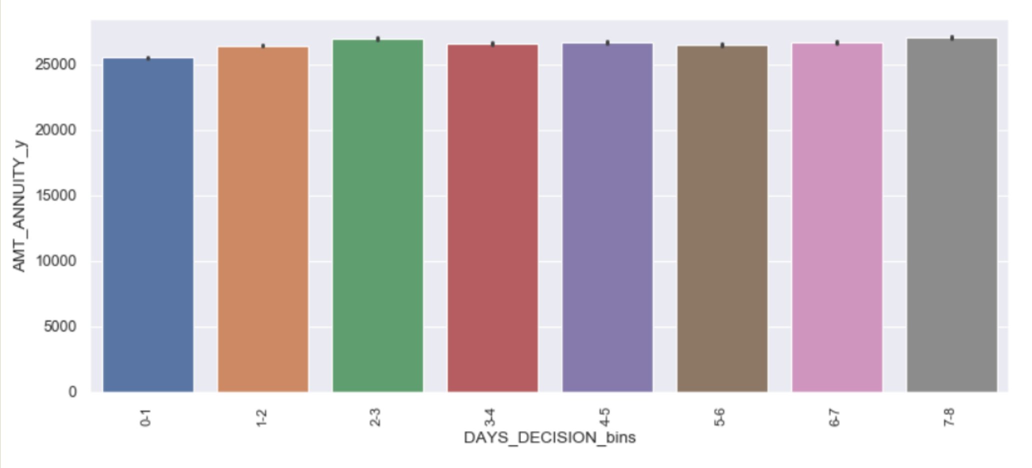


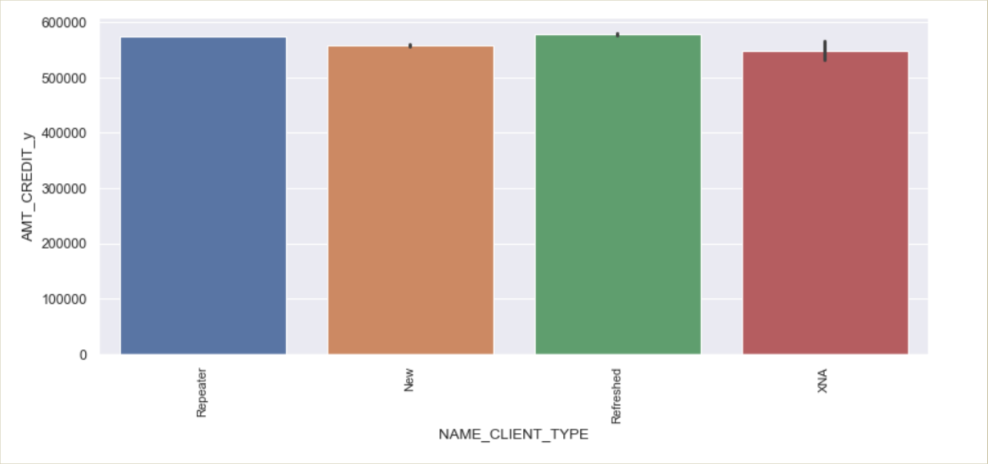


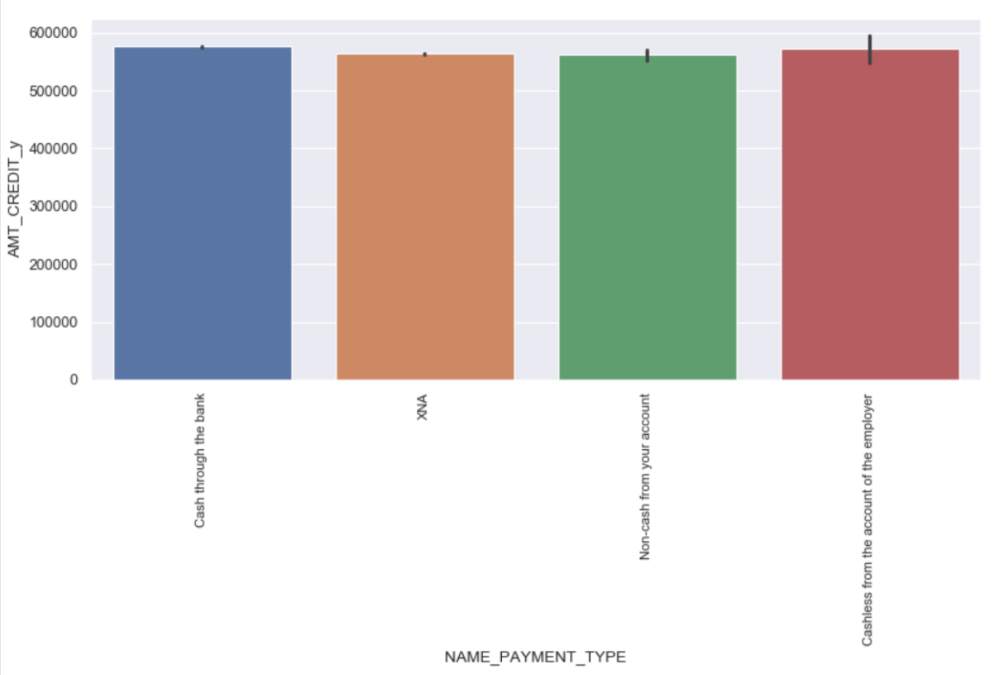


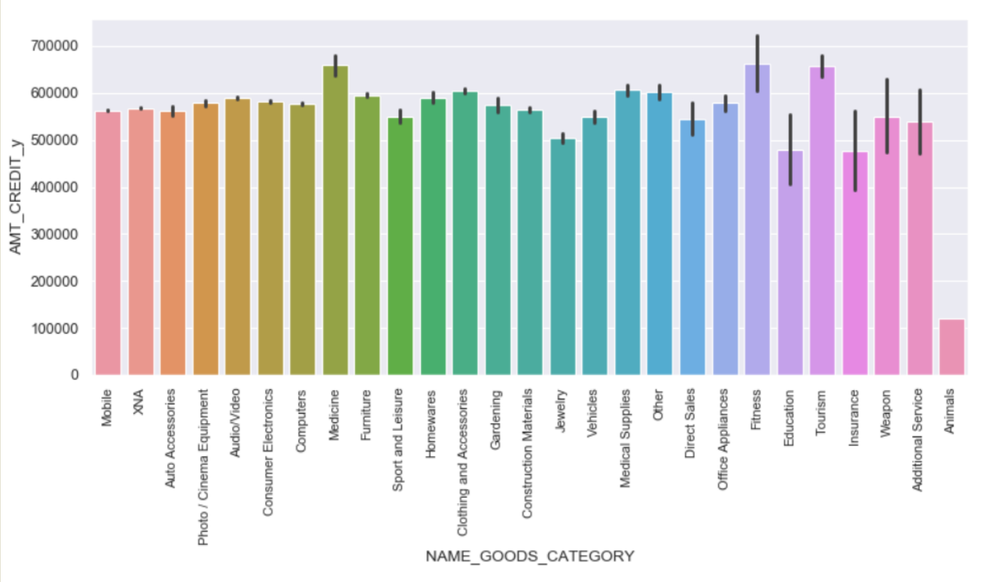

AMT_ANNUITY_y
25000
20000
15000
10000
5000
0
XNA
x-sell
walk-in
NAME_PRODUCT_TYPE

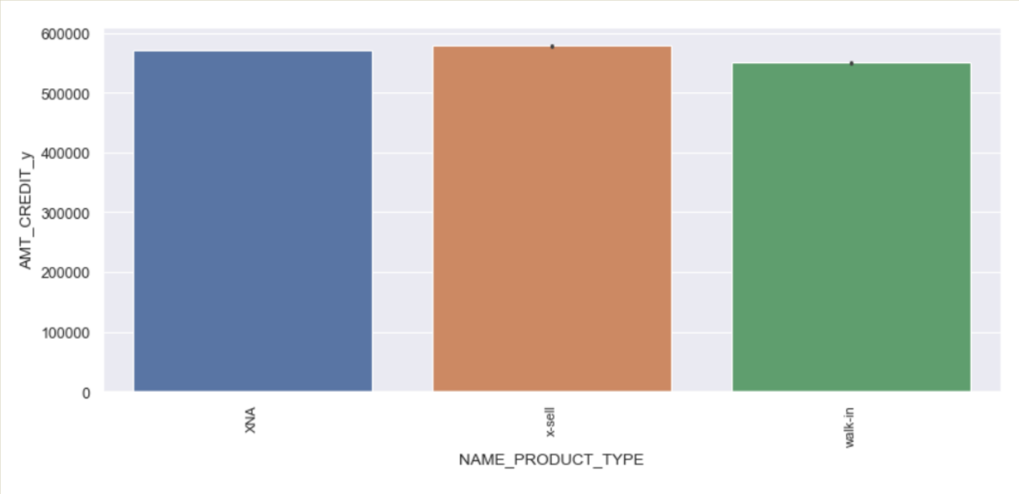

AMT_ANNUITY_y
30000
25000
20000
15000
10000
5000
0
Country-wide
Contact center
Credit and cash offices
Regional / Local
Stone
AP+ (Cash loan)
Channel of corporate sales
Car dealer
CHANNEL_TYPE

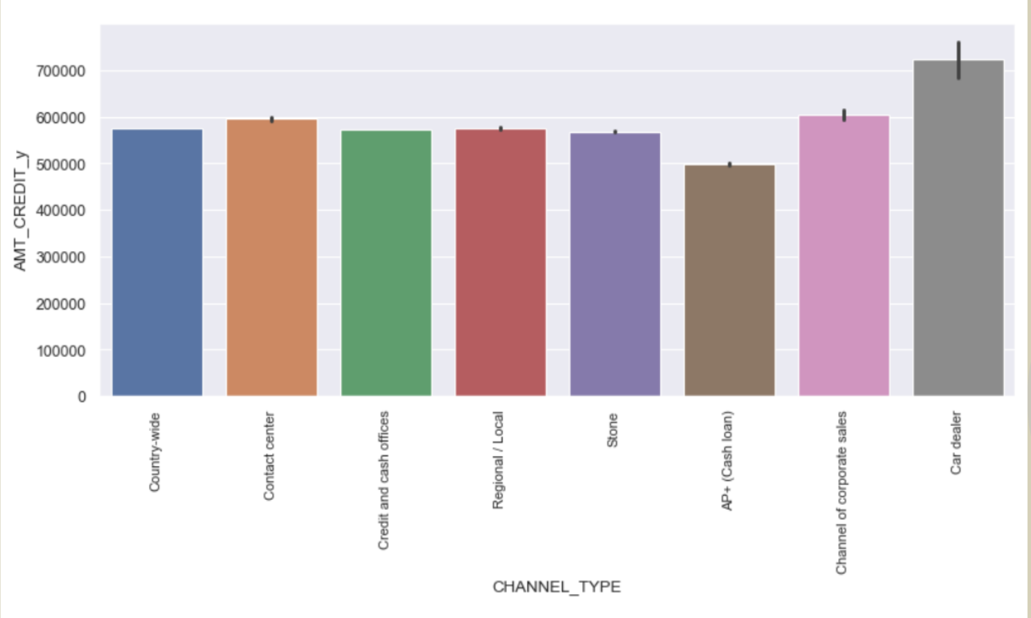

AMT_ANNUITY_y
25000
20000
15000
10000
5000
0
0-1
1-2
2-3
3-4
4-5
5-6
6-7
7-8
DAYS_DECISION_bins

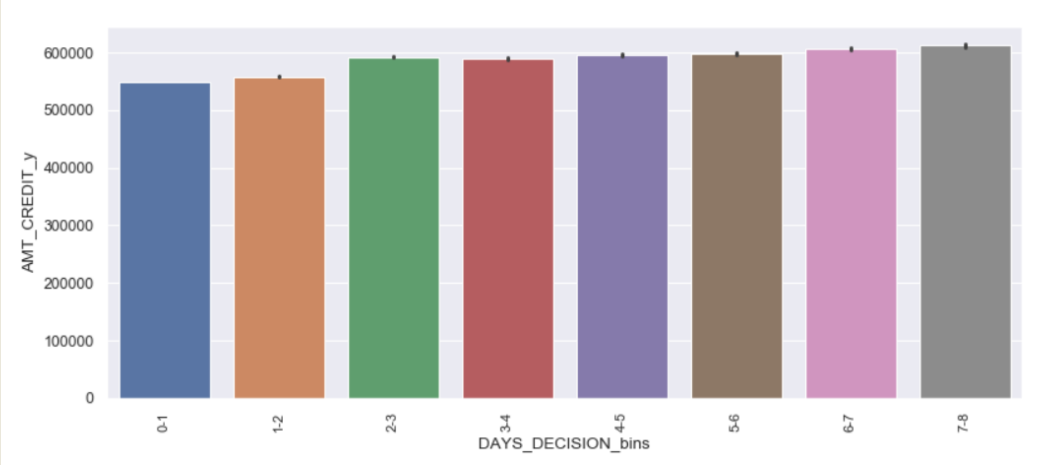

600000
500000
400000
300000
200000
100000
0
AMT_CREDIT_y
Repeater
New
Refreshed
XNA
NAME_CLIENT_TYPE


600000
500000
400000
300000
200000
100000
0
AMT_CREDIT_y
Cash through the bank
XNA
Non-cash from your account
Cashless from the account of the employer
NAME_PAYMENT_TYPE


700000
600000
500000
400000
300000
200000
100000
0
AMT_CREDIT_y
Mobile
XNA
Auto Accessories
Photo / Cinema Equipment
Audio/Video
Consumer Electronics
Computers
Medicine
Furniture
Sport and Leisure
Homewares
Clothing and Accessories
Gardening
Construction Materials
Jewelry
Vehicles
Medical Supplies
Other
Direct Sales
Office Appliances
Fitness
Education
Tourism
Insurance
Weapon
Additional Service
Animals
NAME_GOODS_CATEGORY


600000
500000
400000
300000
200000
100000
0
AMT_CREDIT_y
XNA
x-sell
walk-in
NAME_PRODUCT_TYPE


700000
600000
500000
400000
300000
200000
100000
0
AMT_CREDIT_y
Country-wide
Contact center
Credit and cash offices
Regional / Local
Stone
AP+ (Cash loan)
Channel of corporate sales
Car dealer
CHANNEL_TYPE


600000
500000
400000
300000
200000
100000
0
AMT_CREDIT_y
0-1
1-2
2-3
3-4
4-5
5-6
6-7
7-8
DAYS_DECISION_bins

Analysis of Result of Bivariate Analysis

- Annuity and Product Category: According product category most peoples are ready to pay annuity by taking loan for tourism after that fitness and thirdly for weapon
- Annuity and Channel type: Peoples who are taking loans through car dealers giving more annuity as taking more credit amount.
- Amount credit and client type: The refresher clients are taking more amount of loans.

From the observation of days to taking decisions for giving loans most of decision's are taken at 7-8 days

➙Include visualisations and summarise the most important results in the presentation.

After merging the Previous Application, we get some Necessary Insights for Defaulter People, Some of The Insights Are Mentioned Here

- Previous Payment Type: As per previous record most of the applications are cash through bank (99%) and 1/10th of the applicant is likely to default.
- Client Type: There are lot of repeated application received and only 1 out of 12 applications likely to default so it is safe for the bank to give loans,
- Goods Category: Loans Availed for Cell phone have higher chances of default and it is most common reason for loan application as well.
- Days to decide: The longer it took to decide the lower is the risk of default as the ratio of default for days decision is pretty low for higher values for days taken.
- Previous Credit Amount and Requirement status: Looking at the outliers of defaulter/approved box plot, there are a lot of approved applications with higher than usual amount of money is defaulted.