Bank Loan Case Study

Description:

This case study aims to gives us an idea of applying EDA in a real business scenario.The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. As an analyst in a consumer finance company which specialises in lending various types of loans to urban customers, I have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company must decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample
- All other cases: All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company:

1. **Approved:** The company has approved loan application
2. **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
3. **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
4. **Unused Offer:** Loan has been cancelled by the client but on different stages of the process.

In this case study, I will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Approach:

Here I followed below steps to get the key insights:

1. First I looked into the dataset given to me i.e
    a. application_data.csv which contains all the current application details
    b. previous_application.csv which contains all the information from previous loan application
    c. columns_description.csv where we get to know the meaning of the column names
2. Then I loaded the application_data.csv in the pandas dataFrame
3. Then I looked for the basic information about the dataset
4. Then I looked for the missing values percentage in the columns
5. Then I dropped the columns where >50% missing values were there.
6. Then I did some analysis what needs to be done for the missing values in the remaining dataset. I came to conclusion that in some columns missing values should be replaced with mean value in some cases with mode value i.e the value which is most occurring one and some columns will just be replaced by 0.
7. Then I dropped a few columns which were irrelevant for us.
8. Done some outlier analysis using box plot.
9. Then I did univariate, bivariate analysis on the numerical and categorical variables
10. Checked Data imbalance in the dataset
11. Generated insights & visuals from the data based on defaulter & non-defaulter.
12. Then loaded the previous_application.csv in the pandas DataFrame
13. Then again did the missing value analysis
14. Then I dropped the columns where >50% missing values were there
15. Then merged both the datasets based on common key SK_ID_CURR
16. Then generated insights based on contract status i.e Approved,Cancled,Refused,Unused Offer

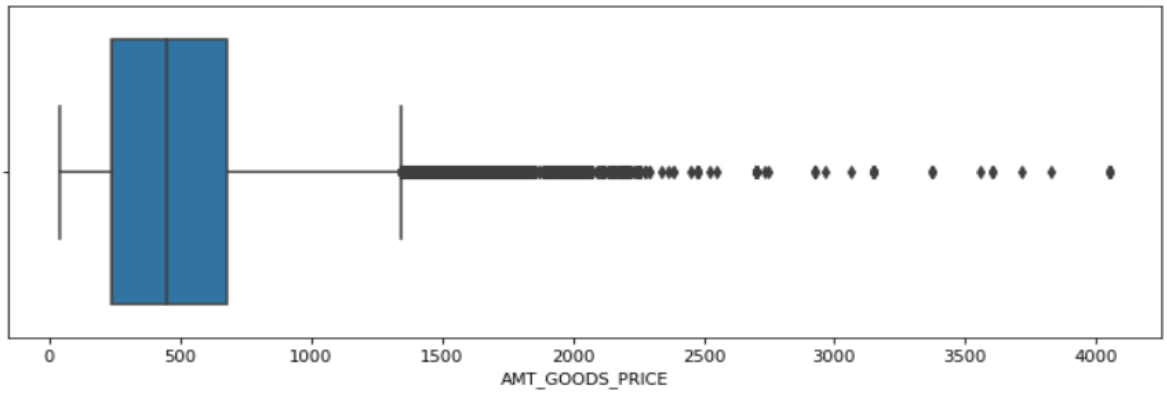
Tech-Stack Used:

- Python as the programming language
- Pandas,Numpy library for analysis
- Matplotlib.pyplot and seaborn for visualization
- Jupyter Notebook as the IDE

Insights

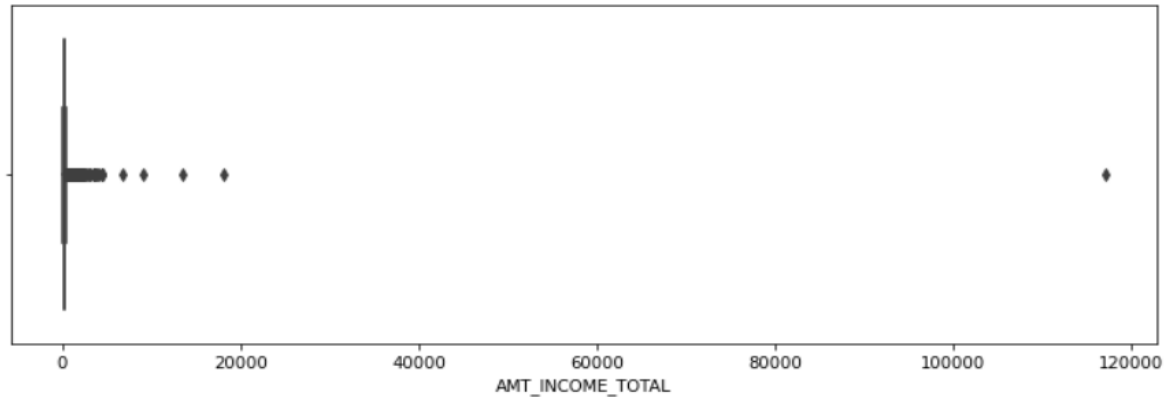
From the Current Application Data:

1. **Outlier analysis for AMT_GOODS_PRICE**



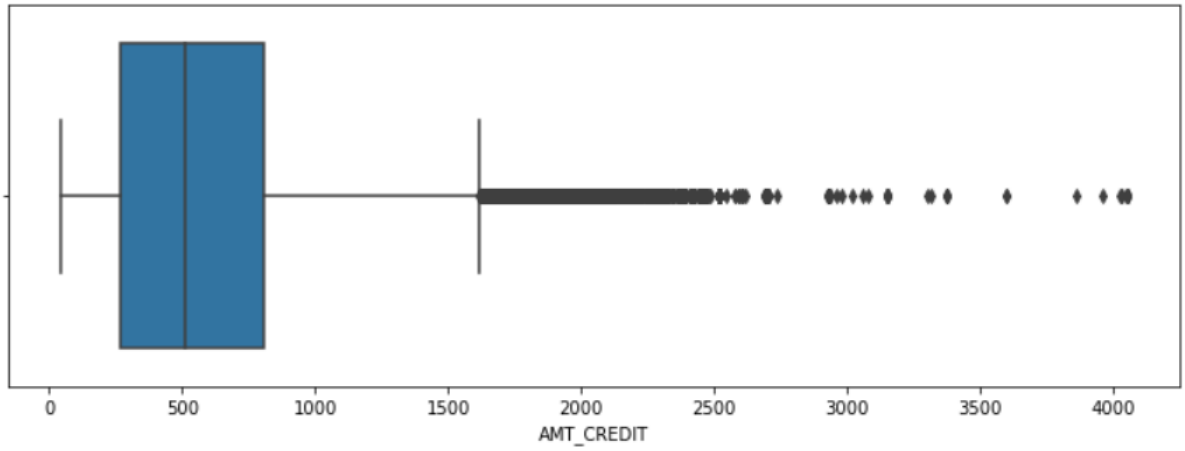

```
count    307233.000000
mean        538.396207
std         369.446461
min          40.500000
25%         238.500000
50%         450.000000
75%         679.500000
max        4050.000000
Name: AMT_GOODS_PRICE, dtype: float64
Upper Limit 1341.0
Outlier % 4.79
```

2. **Outlier analysis for AMT_INCOME_TOTAL**



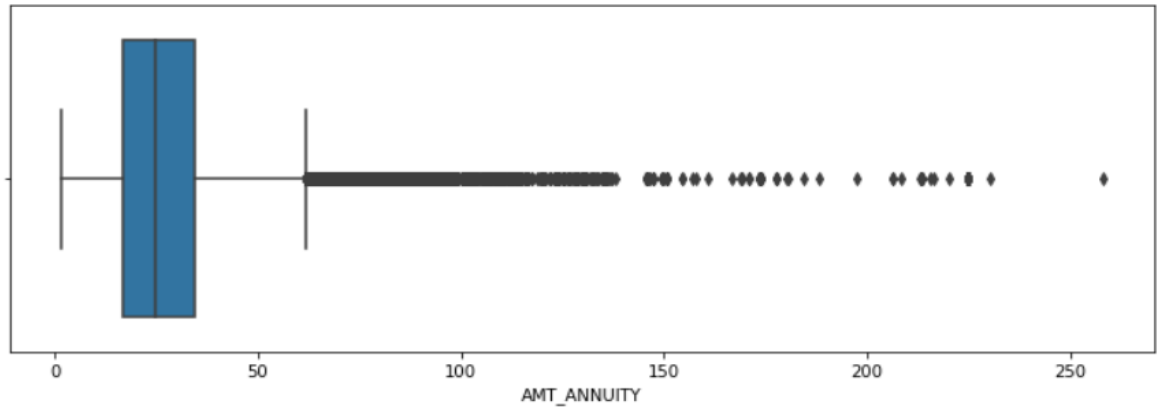

```
count    307511.000000
mean        168.797919
std         237.123146
min          25.650000
25%         112.500000
50%         147.150000
75%         202.500000
max      117000.000000
Name: AMT_INCOME_TOTAL, dtype: float64
Upper Limit 337.5
Outlier % 4.56
```

3. **Outlier analysis for AMT_CREDIT**



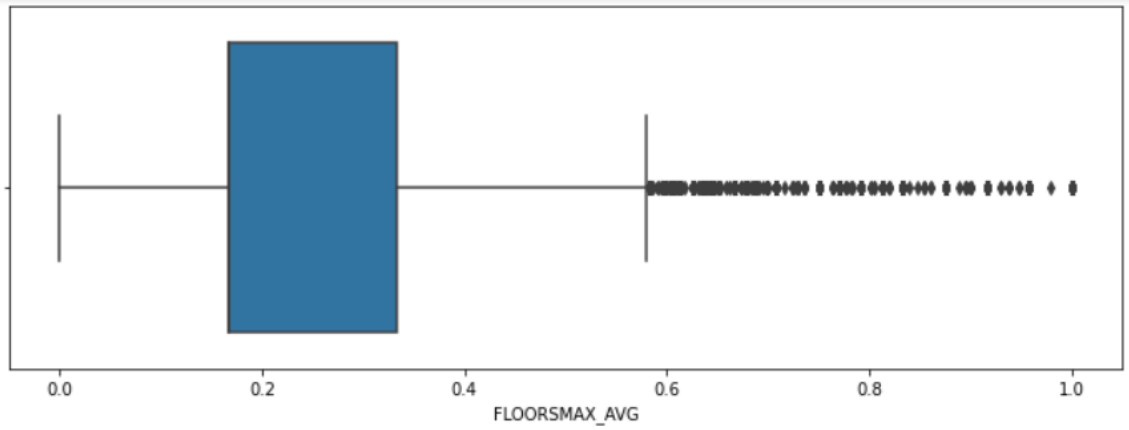

```
count    307511.000000
mean        599.026000
std         402.490777
min          45.000000
25%         270.000000
50%         513.531000
75%         808.650000
max        4050.000000
Name: AMT_CREDIT, dtype: float64
Upper Limit 1616.625
Outlier % 2.13
```

4. **Outlier analysis for AMT_ANNUITY**




```
count    307499.000000
mean         27.108574
std          14.493737
min           1.615500
25%          16.524000
50%          24.903000
75%          34.596000
max         258.025500
Name: AMT_ANNUITY, dtype: float64
Upper Limit 61.70399999999999
Outlier % 2.44
```

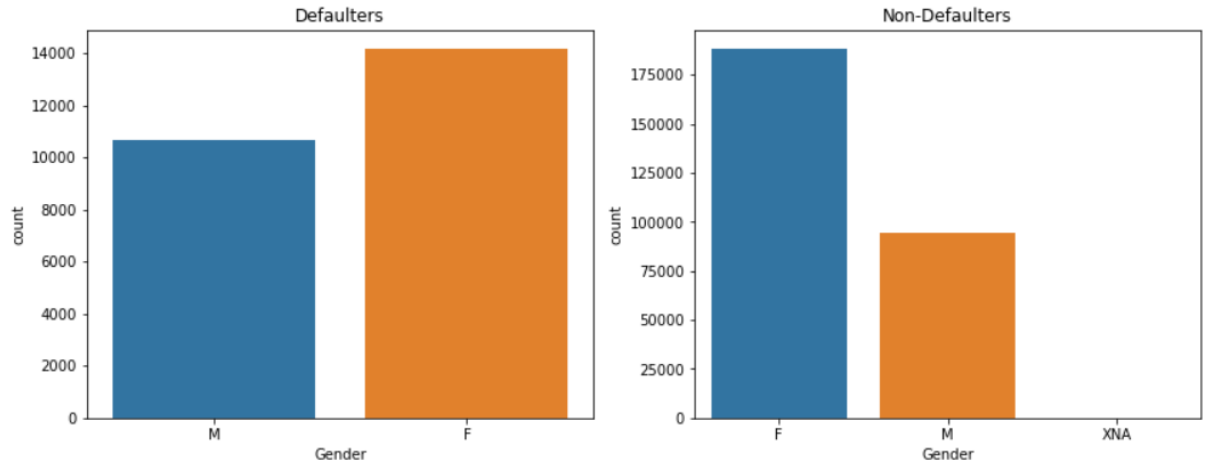
5. Outlier analysis for FLOORSMAX_AVG




```
count    154491.000000
mean          0.226282
std           0.144641
min           0.000000
25%           0.166700
50%           0.166700
75%           0.333300
max           1.000000
Name: FLOORSMAX_AVG, dtype: float64
Upper Limit 0.5831999999999999
Outlier % 1.7
```

6. Comparison of defaulters and non-defaulters based on gender

```
[Text(0.5, 0, 'Gender')]
```



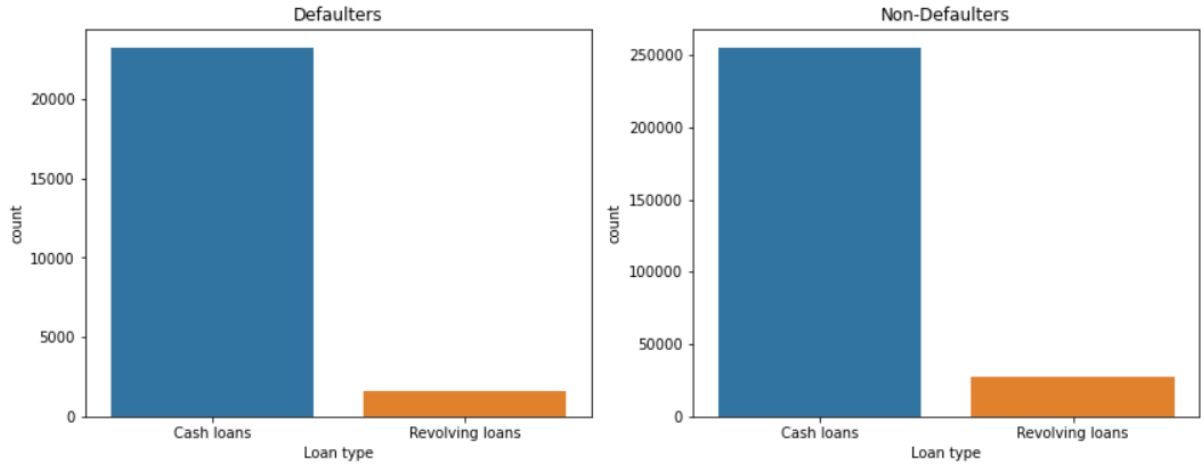
Insights

Defaluters - We can see that females are more in number of defaulters than male.

Non-defaluters - The same pattern continues for non-defaluters as well. The females are more in number here than male.

7. Comparison of defaulters and non-defaulters based on Loan Type

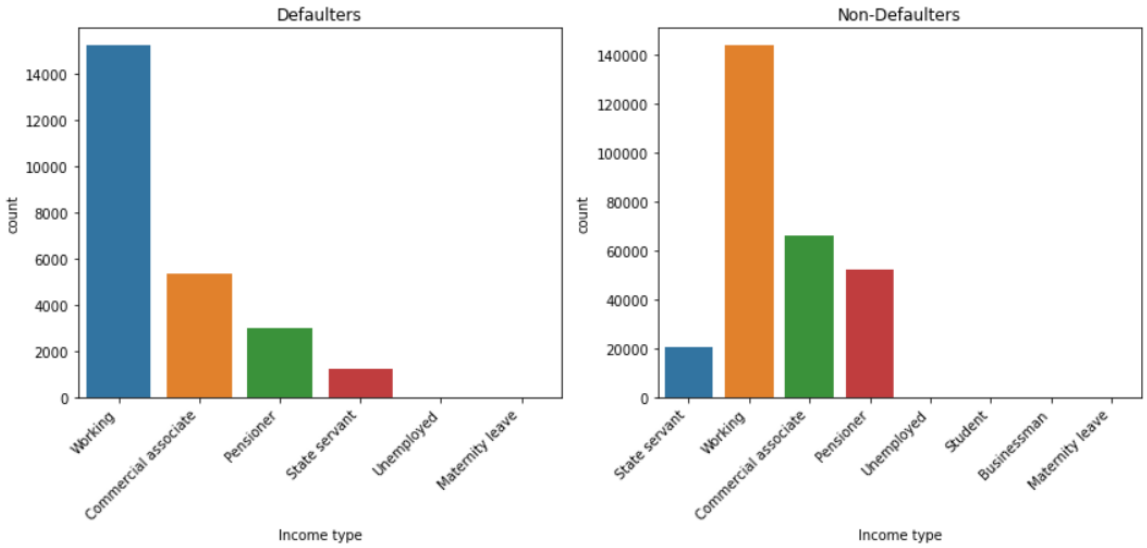
```
[Text(0.5, 0, 'Loan type')]
```



Insights

We see in both the cases that Revolving loans are very less in number compared to Cash loans.

8. Comparison of Defaulters and non-defaulters based on Income type

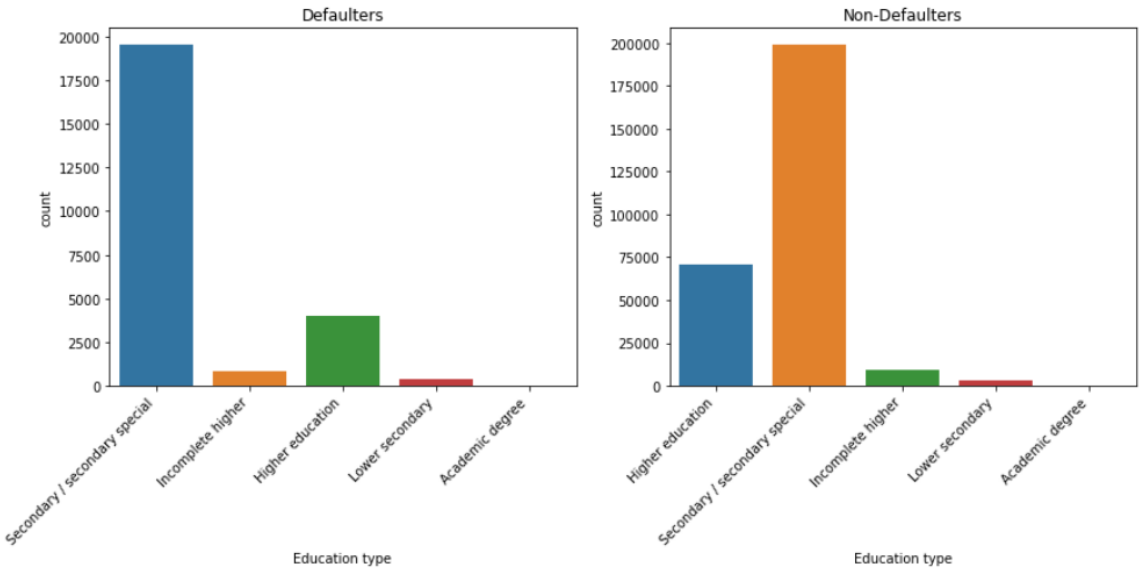


Insights

Defaulters - Working people are mostly defaulted as their numbers are high with compare to other professions.

Non-defaulters - Similarly here also working people are more in number who are not defaulted.

9. Comparison of Defaulters and non-defaulters based on Education type

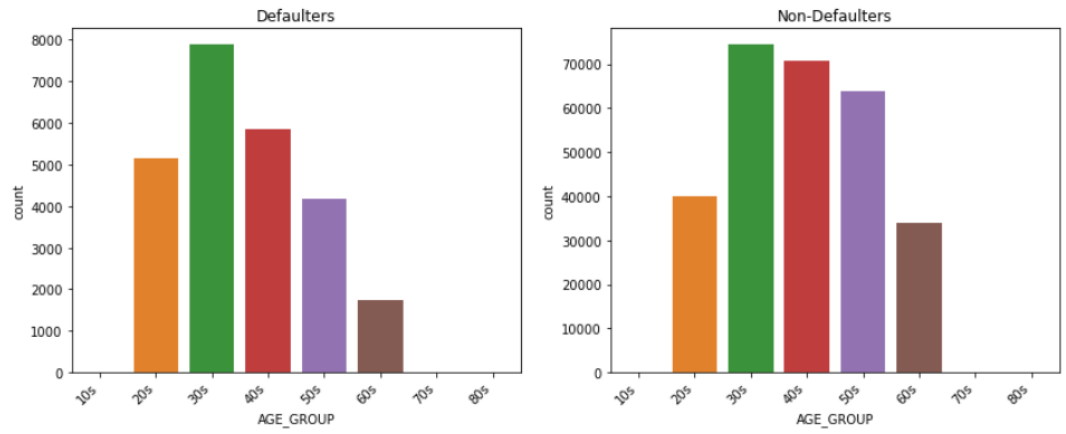


Insights

Defaulters - Education with Secondary/Secondary sepcial customers are more number in defaulters comapre with other level of eduacted poeple.
Non defaulters - Here also Secondary/Secondary sepcial are more in numbers.

10. Comparison of Defaulters and non-defaulters based on AGE_GROUP

I have divided the DAYS_BIRTH column by 365 and taken absolute value of that to get the age and then grouped them according to range 20-29,30-39 like this

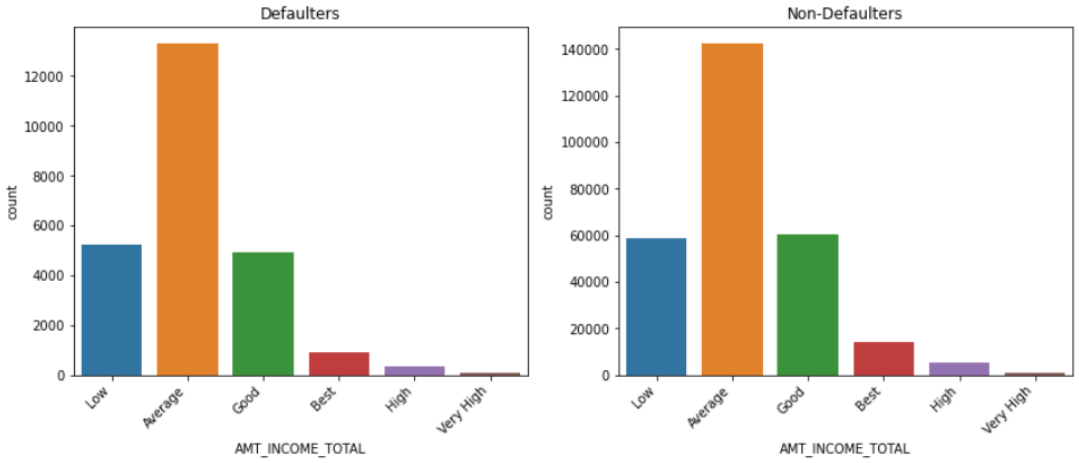


Insights

Defaulters - Customers who are in their early days in career(specially in 30's) are more number in defaulters comapre with other Age group. Older people are less likely to be defaulter
Non defaulters - Here also the same trend is followed.

11. Comparison of Defaulters and non-defaulters based on AMT_INCOME_TOTAL

Like age I have also created a bin for AMT_CATEGORY as AMT_INCOME_TOTAL

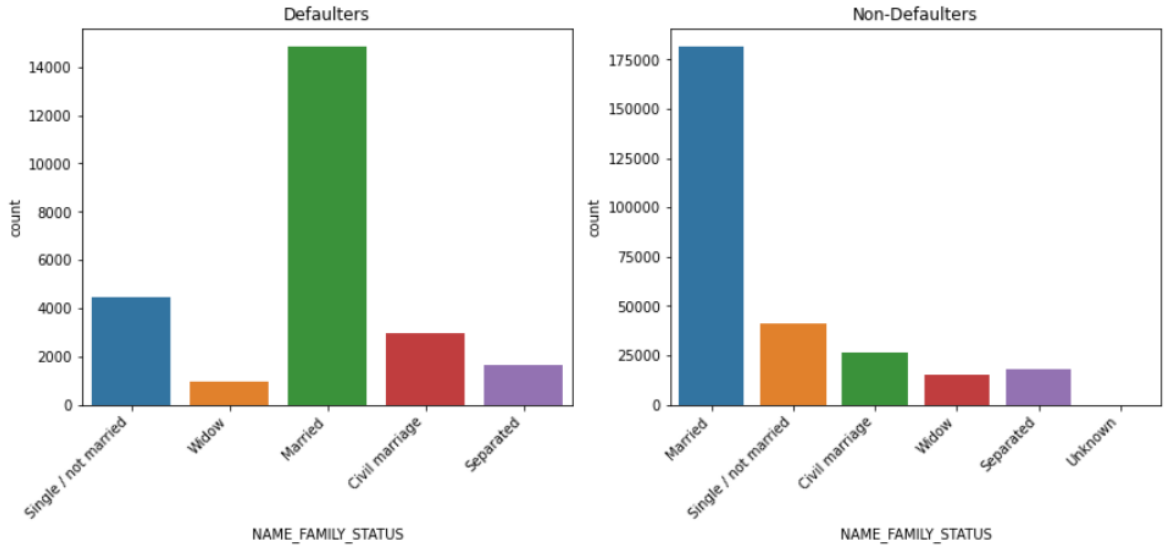


Insights

Defaulters - Customers who are average income group are most likely to be defaulter and who are in best,high and very high income group are less likely defaulter

Non defaulters - Here also the same trend is followed.

12. Comparison of Defaulters and non-defaulters based on NAME_FAMILY_STATUS

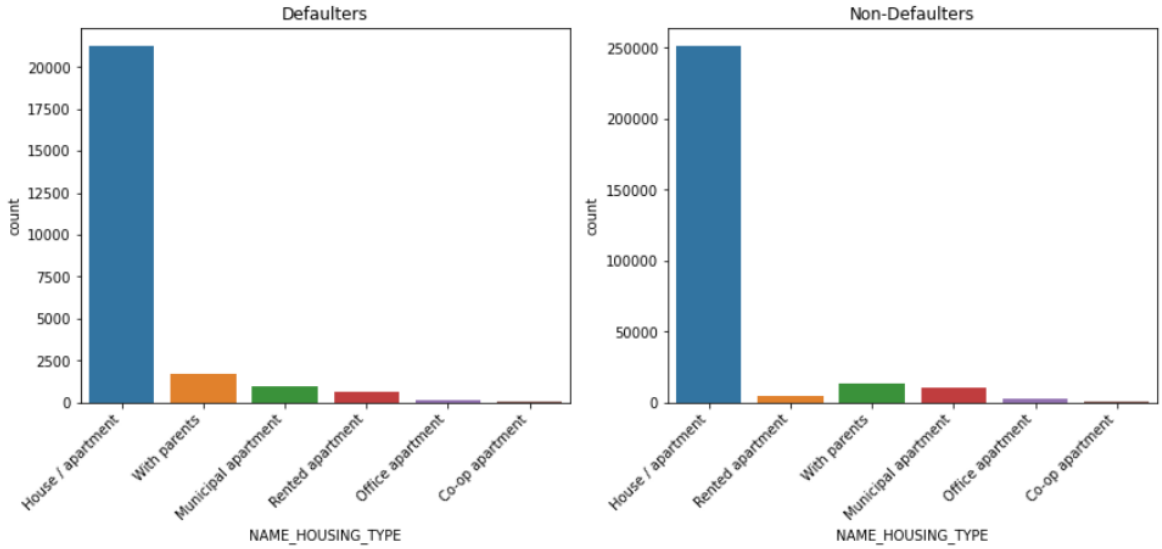


Insights

Defaulters - Customers who are married are most likely to be defaulter

Non defaulters - Here also the same trend is followed.

13. Comparison of Defaulters and non-defaulters based on NAME_HOUSING_TYPE

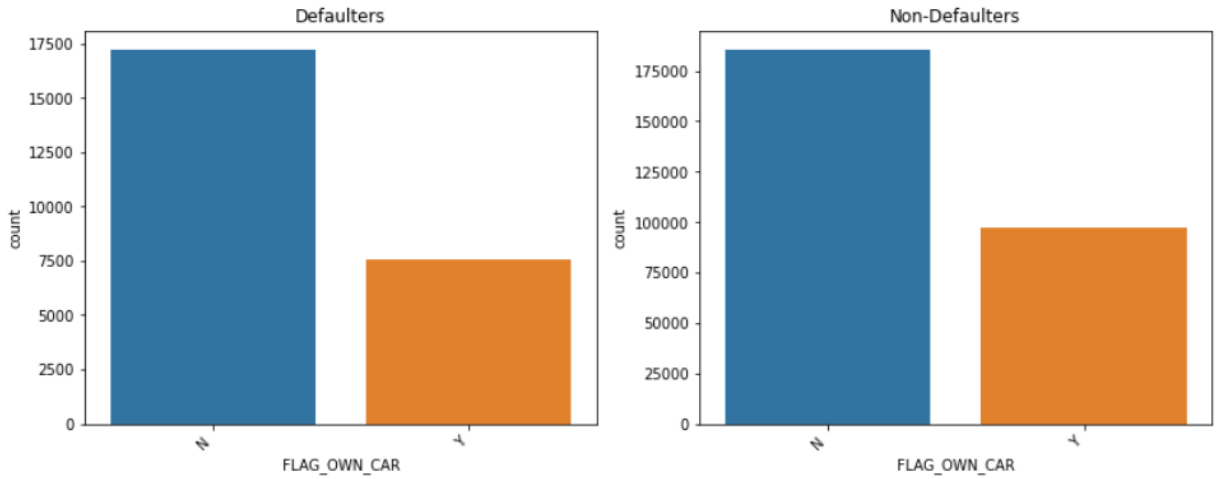


Insights

Defaulters - Customers who are owining a house/apartment are most likely to be defaulter

Non defaulters - Here also the same trend is followed.

14. Comparison of Defaulters and non-defaulters based on FLAG_OWN_CAR

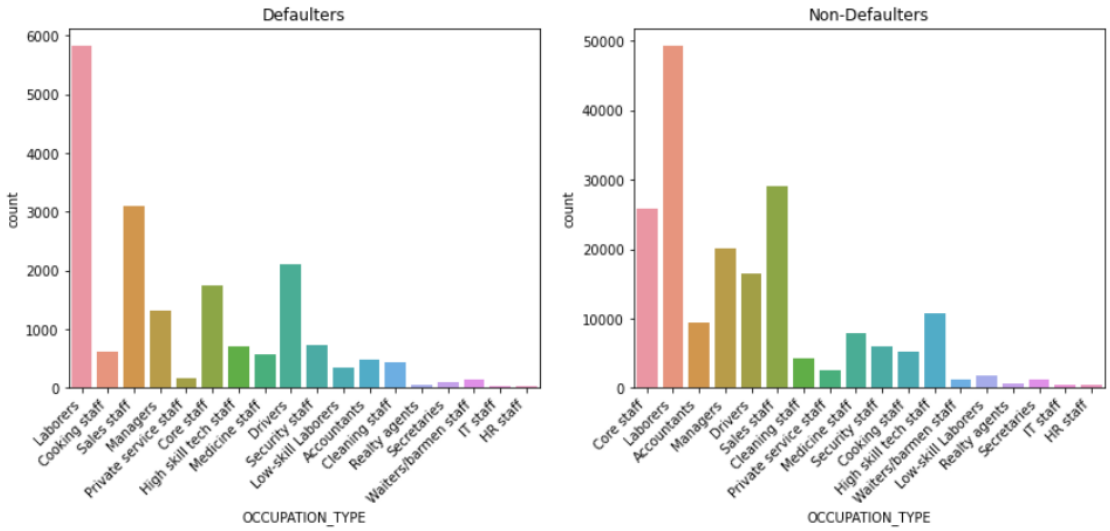


Insights

Defaulters - Customers who are not owining a car are most likely to be defaulter

Non defaulters - Here also the same trend is followed.

15. Comparison of Defaulters and non-defaulters based on OCCUPATION_TYPE



Insights

Defaulters - Customers who are having OCCUPATION_TYPE as laborers,Sales Staff,Drivers,Core Staff are most likely to be defaulter

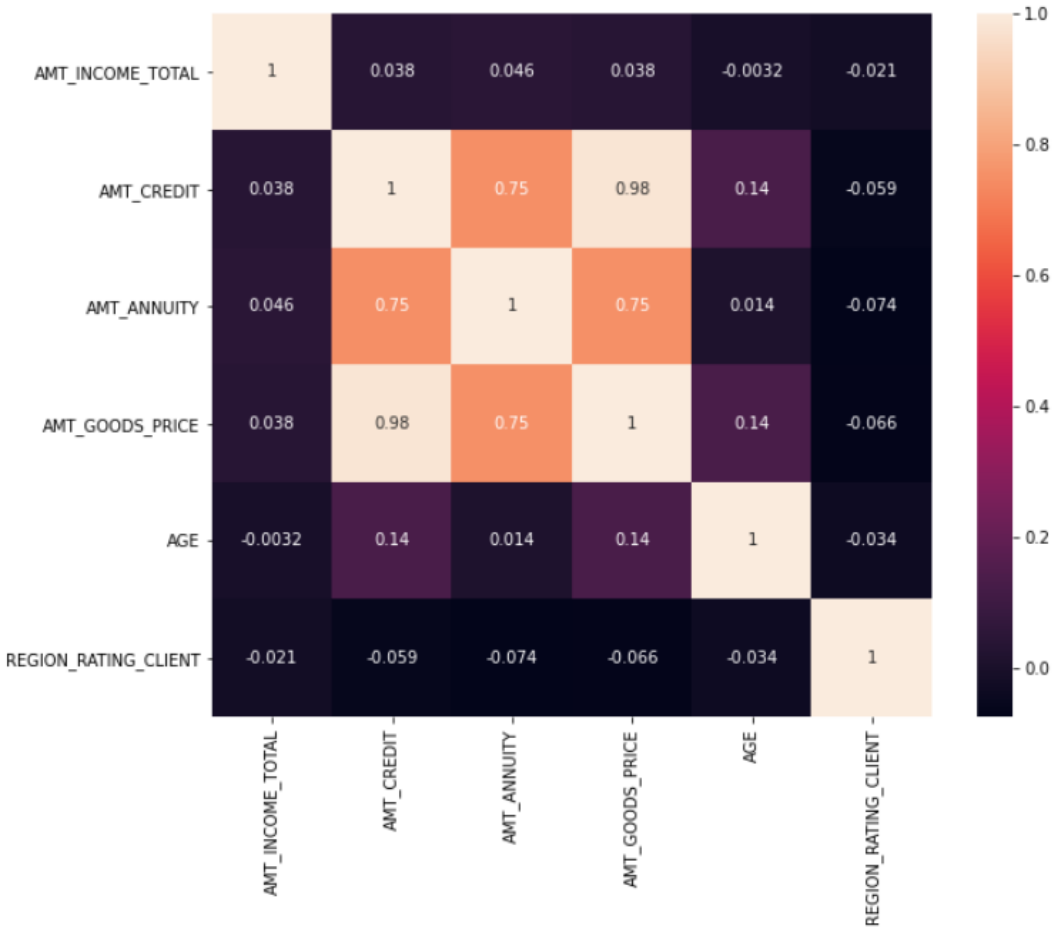
Non defaulters - Customers who are having OCCUPATION_TYPE as laborers,Sales Staff,Managers,Core Staff are most likely to be non-defaulter

16. Correlation matrix for target 1 i.e Defaulters

Correlation matrix for target 1

```
df_corr_target_1.corr()
```

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	AGE	REGION_RATING_CLIENT
AMT_INCOME_TOTAL	1.000000	0.038131	0.046421	0.037583	-0.003154	-0.021486
AMT_CREDIT	0.038131	1.000000	0.752195	0.983103	0.135070	-0.059193
AMT_ANNUITY	0.046421	0.752195	1.000000	0.752699	0.014028	-0.073784
AMT_GOODS_PRICE	0.037583	0.983103	0.752699	1.000000	0.135603	-0.066390
AGE	-0.003154	0.135070	0.014028	0.135603	1.000000	-0.033648
REGION_RATING_CLIENT	-0.021486	-0.059193	-0.073784	-0.066390	-0.033648	1.000000



Highly correlate columns for defaulters

AMT_CREDIT and AMT_ANNUITY (0.75)

AMT_CREDIT and AMT_GOODS_PRICE (0.98)

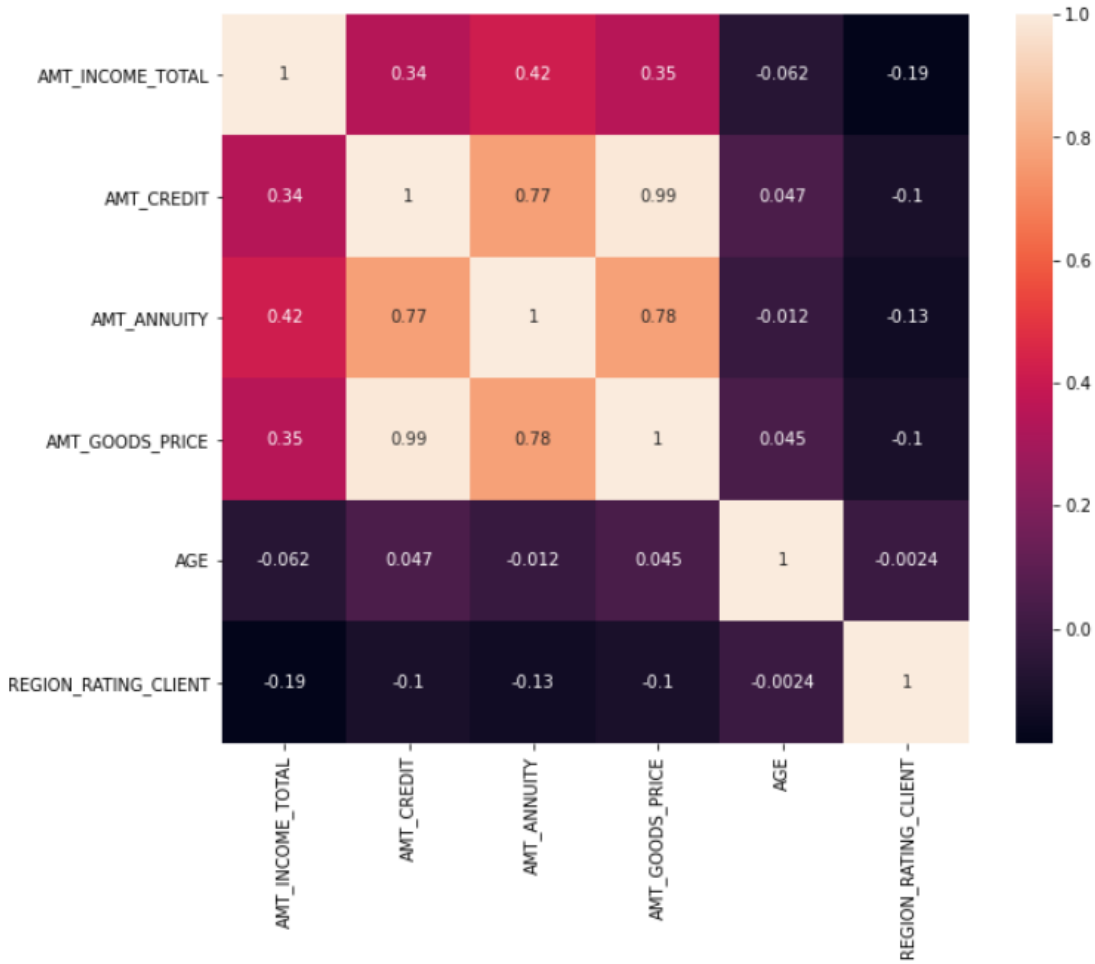
AMT_ANNUITY and AMT_GOODS_PRICE (0.75)

17. Correlation matrix for target 0 i.e Non-defaulters

Correlation matrix for target 0

```
df_corr_target_0.corr()
```

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	AGE	REGION_RATING_CLIENT
AMT_INCOME_TOTAL	1.000000	0.342799	0.418953	0.349462	-0.062494	-0.186573
AMT_CREDIT	0.342799	1.000000	0.771309	0.987250	0.047366	-0.103337
AMT_ANNUITY	0.418953	0.771309	1.000000	0.776686	-0.012254	-0.132128
AMT_GOODS_PRICE	0.349462	0.987250	0.776686	1.000000	0.044552	-0.104382
AGE	-0.062494	0.047366	-0.012254	0.044552	1.000000	-0.002415
REGION_RATING_CLIENT	-0.186573	-0.103337	-0.132128	-0.104382	-0.002415	1.000000



Highly corelate columns for non defaulters

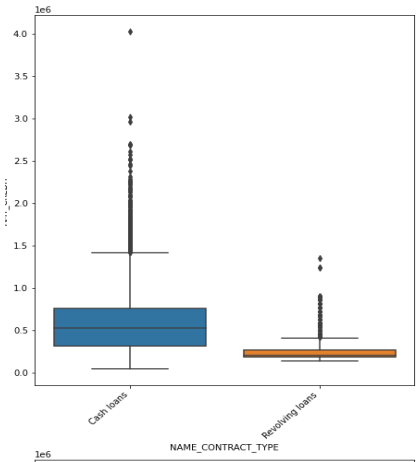
AMT_CREDIT and AMT_ANNUITY (0.77)

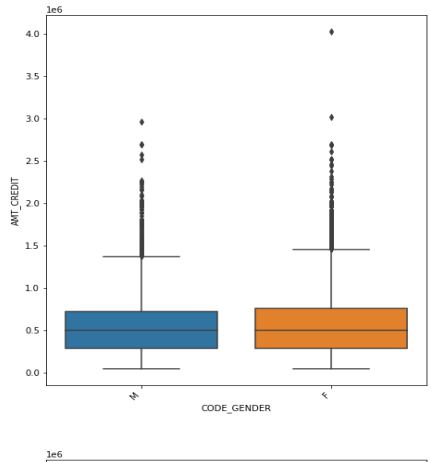
AMT_CREDIT and AMT_GOODS_PRICE (0.99)

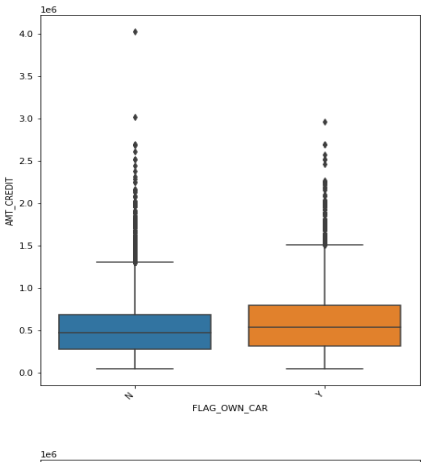
AMT_ANNUITY and AMT_GOODS_PRICE (0.78)

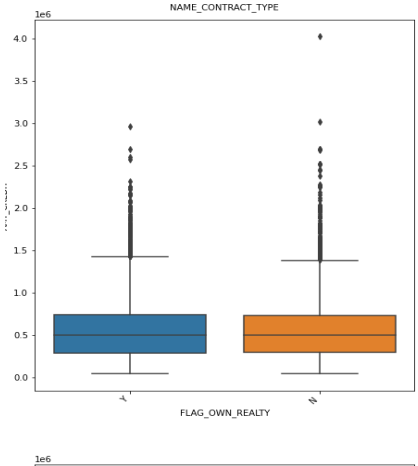
18. Bivariate analysis on categorical variable for defaulters

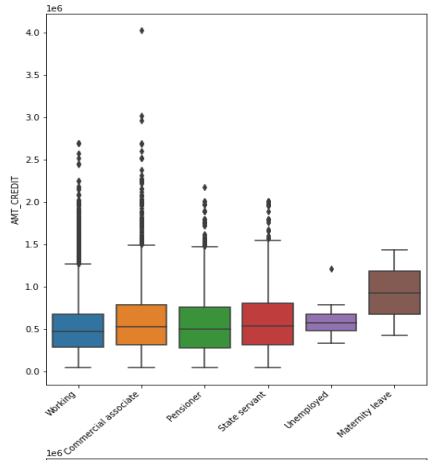
categories =
['NAME_CONTRACT_TYPE','CODE_GENDER','FLAG_OWN_CAR','FLAG_OWN_REALTY','NAME_INCOME_TYPE','NAME_EDUCATION_TYPE',
'NAME_FAMILY_STATUS','NAME_HOUSING_TYPE','AGE_GROUP','AMT_CATEGORY']

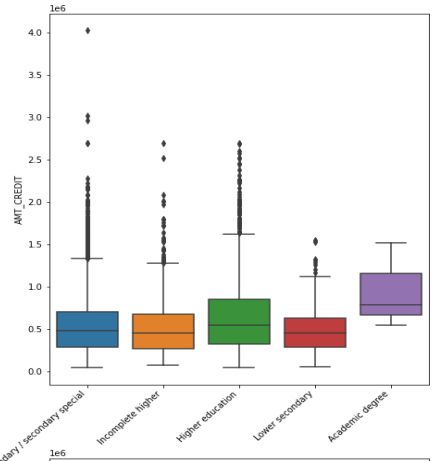

1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Cash loans
Revolving loans
NAME_CONTRACT_TYPE

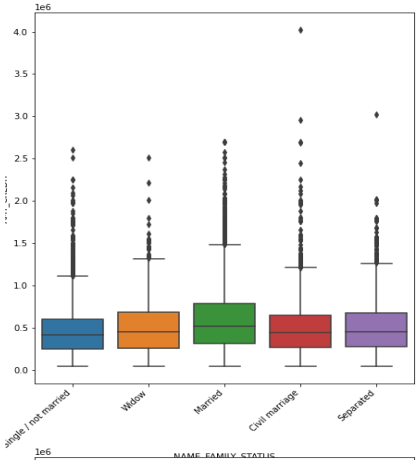

1e6
AMT_CREDIT
M
F
CODE_GENDER

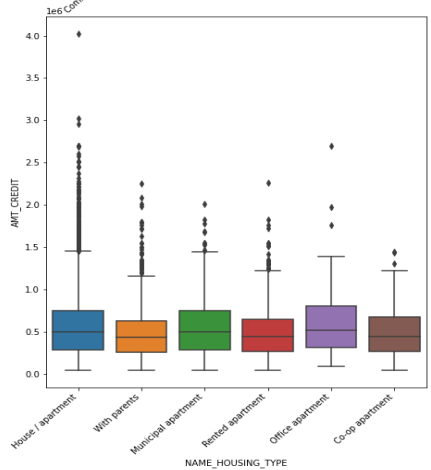

1e6
AMT_CREDIT
N
Y
FLAG_OWN_CAR

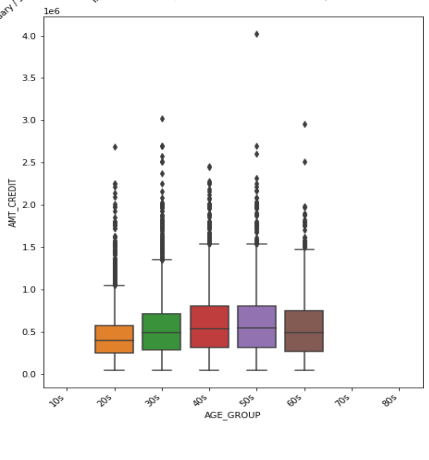

1e6
Y
N
FLAG_OWN_REALTY

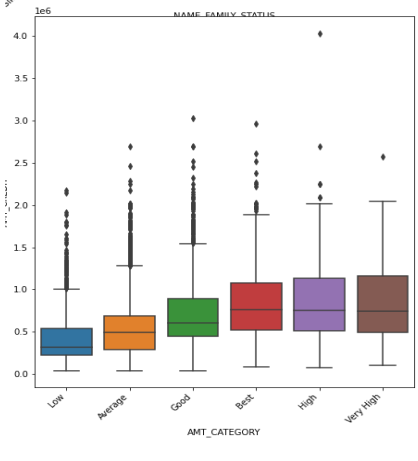

1e6
AMT_CREDIT
Working
Commercial associate
Pensioner
State servant
Unemployed
Maternity leave


1e6
AMT_CREDIT
Secondary / secondary special
Incomplete higher
Higher education
Lower secondary
Academic degree


1e6
Widow
Married
Civil marriage
Separated
NAME_FAMILY_STATUS


1e6
AMT_CREDIT
House / apartment
With parents
Municipal apartment
Rented apartment
Office apartment
Co-op apartment
NAME_HOUSING_TYPE

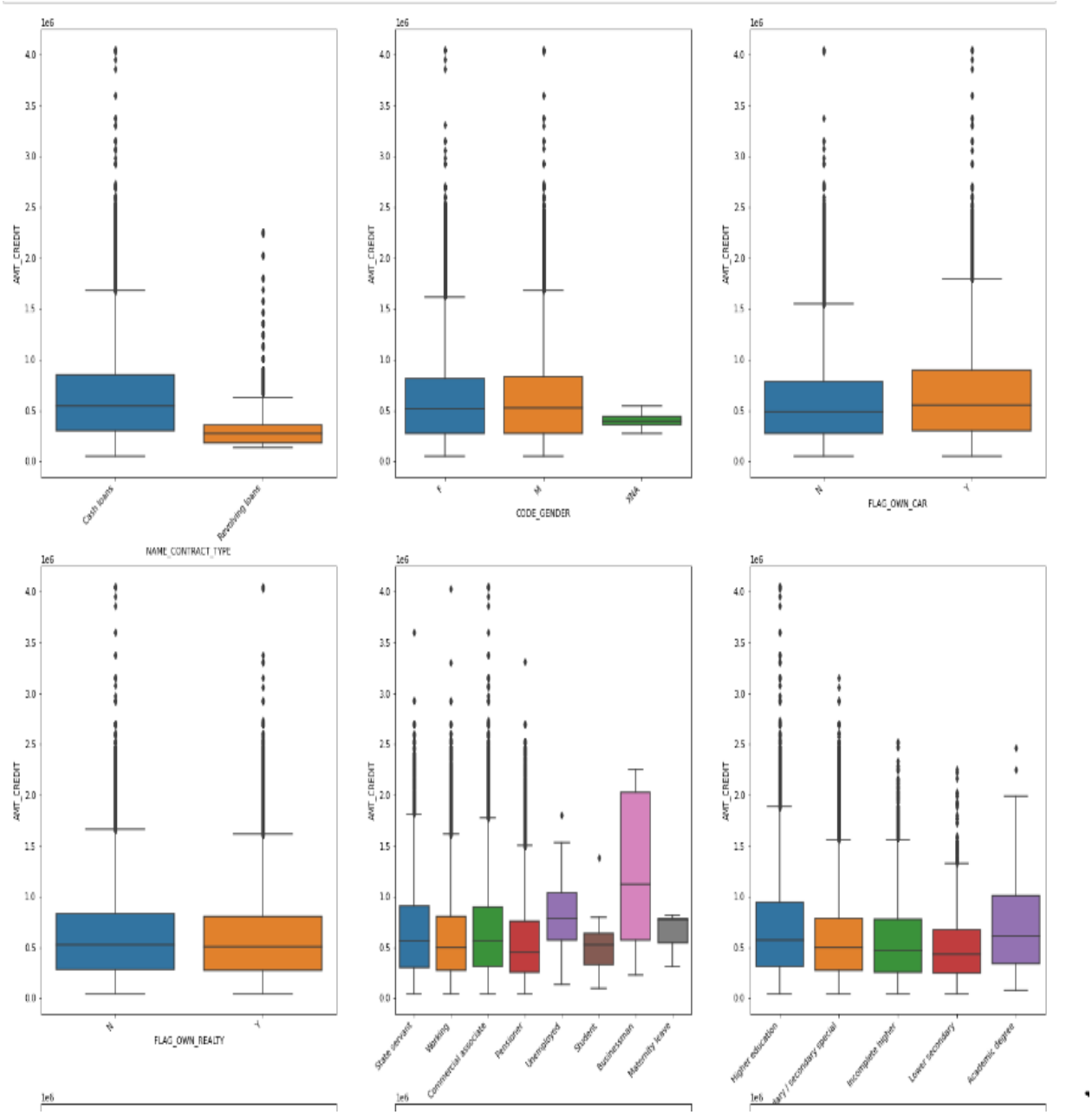

1e6
AMT_CREDIT
10s
20s
30s
40s
50s
60s
70s
80s
AGE_GROUP

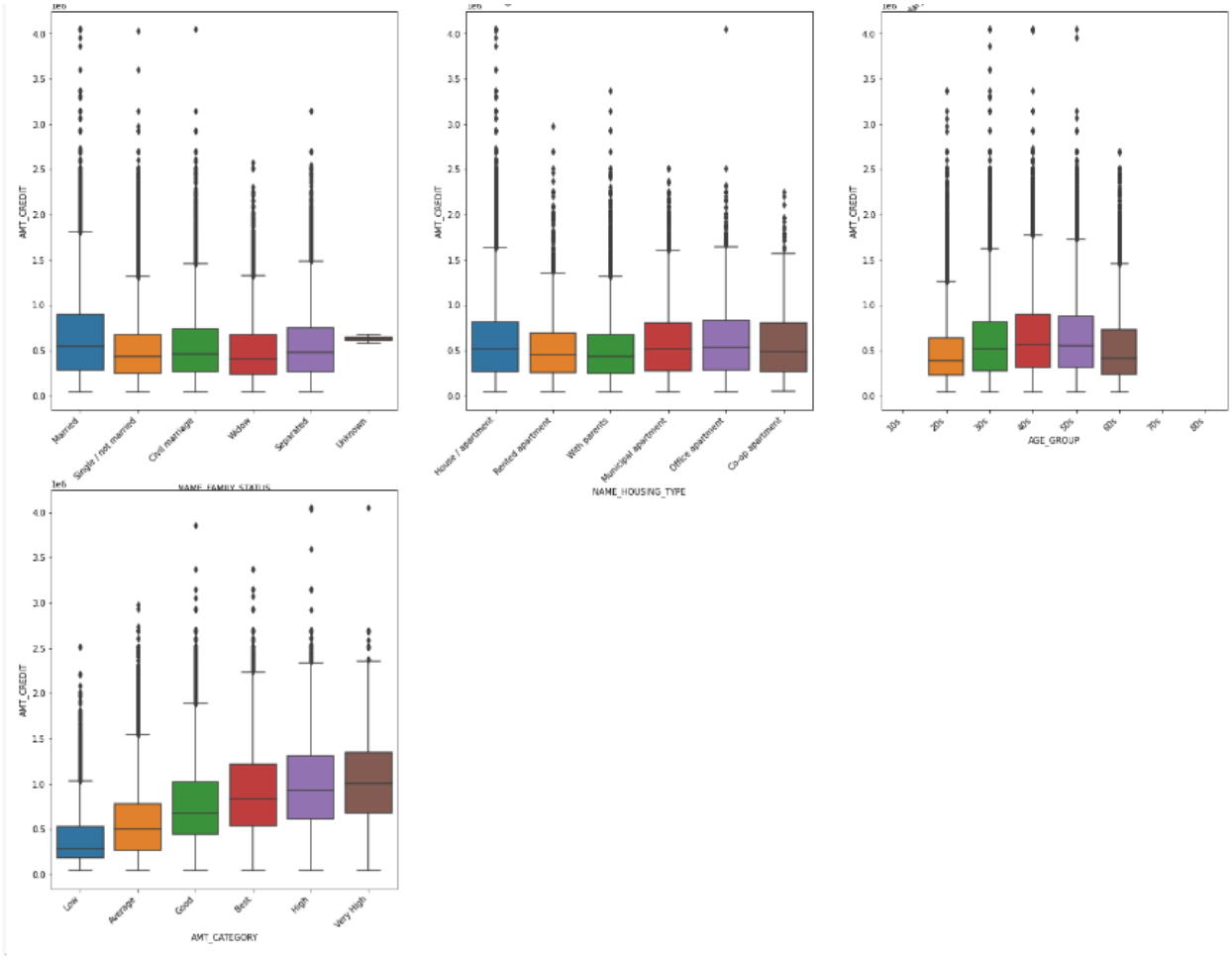

1e6
Low
Average
Good
Best
High
Very High
AMT_CATEGORY

Analysis

- Credit amount of the loans are very low for Revolving loans
- There is no credit amount difference between genders, client owning cars or realty.
- The Young age group got less amount of loan credited compared to mid age and senior citizen.
- Higher income group have more loan amount credited.
- Clients having higher external score have more loan amount.

19. Bivariate analysis on categorical variable for non-defaulters



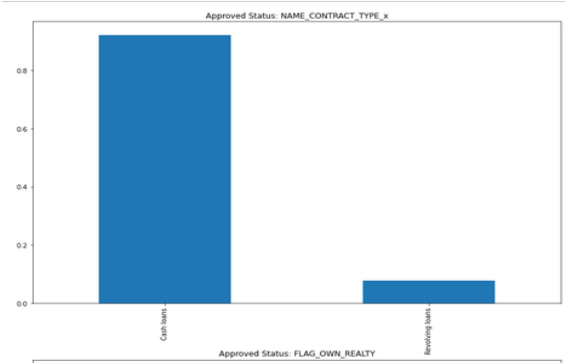


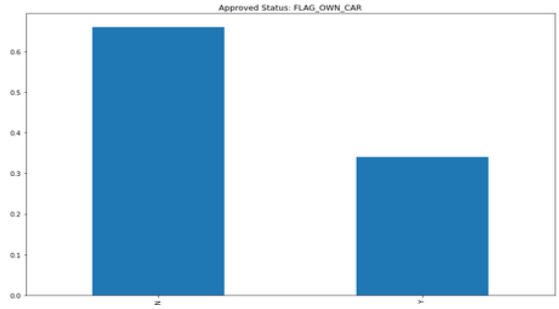
Analysis

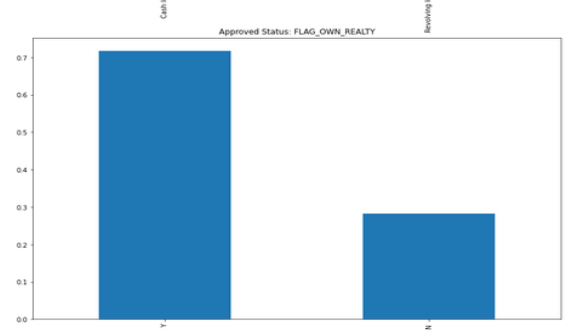
- Credit amount of the loans are very low for Revolving loans
- There is no credit amount difference between genders, client owning cars or realty.
- The mid age group got more amount of loan credited compared to young and senior citizen.
- Higher income group have more loan amount credited and lower the lowest.
- Clients having higher external score have more loan amount.
- Surprisingly the unemployed people have spike in credit amount of loan
  The Married people have more loan amount credited.

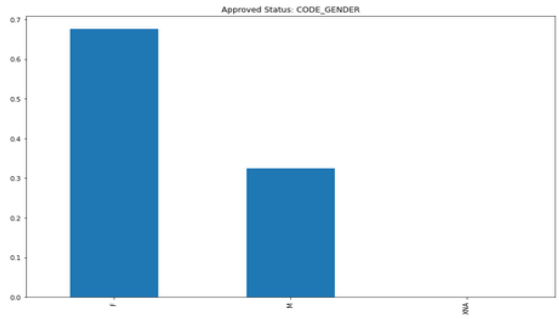
From the Combined Application Data: Analysis on People with Contract Status as Approved

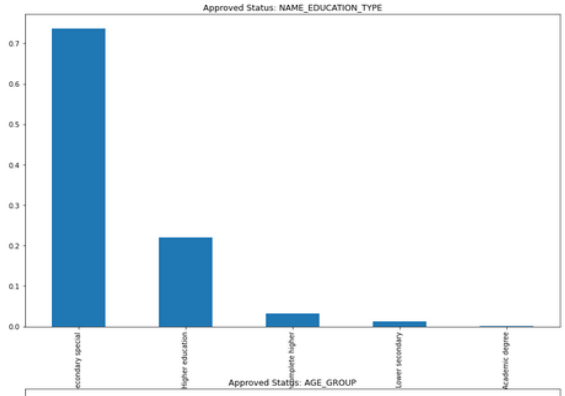
1. Univariate Analysis for few categorical columns in combined dataframe for Approved

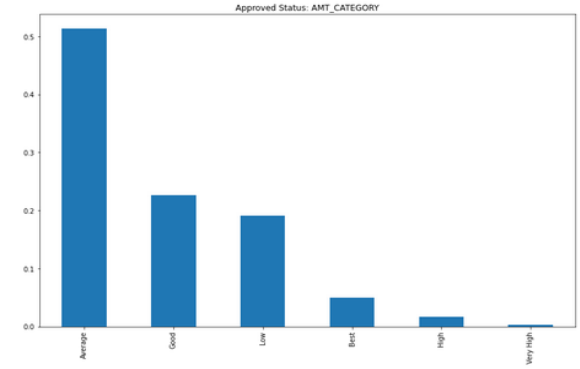


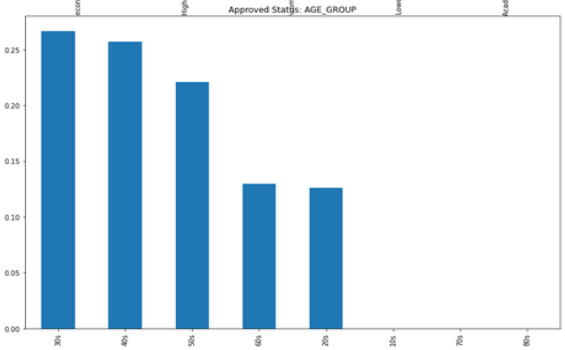








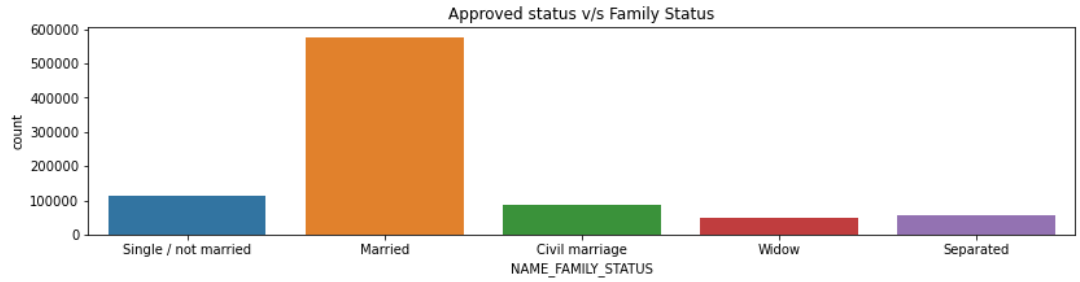


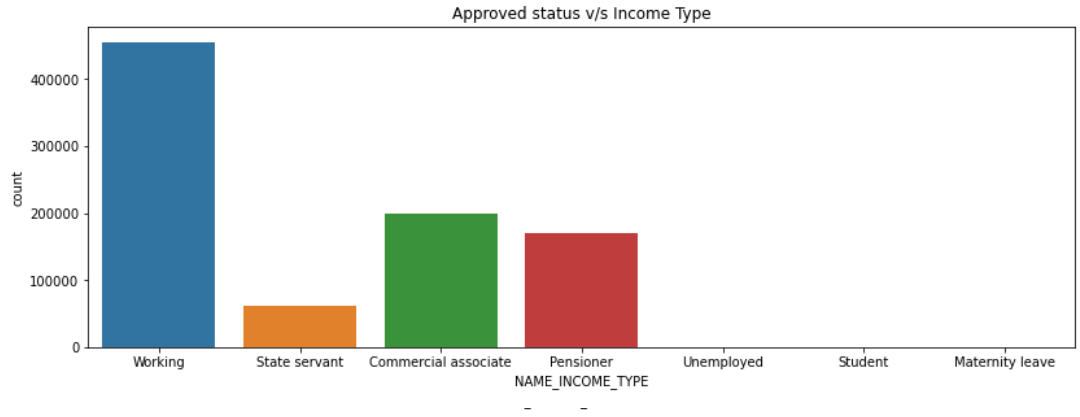


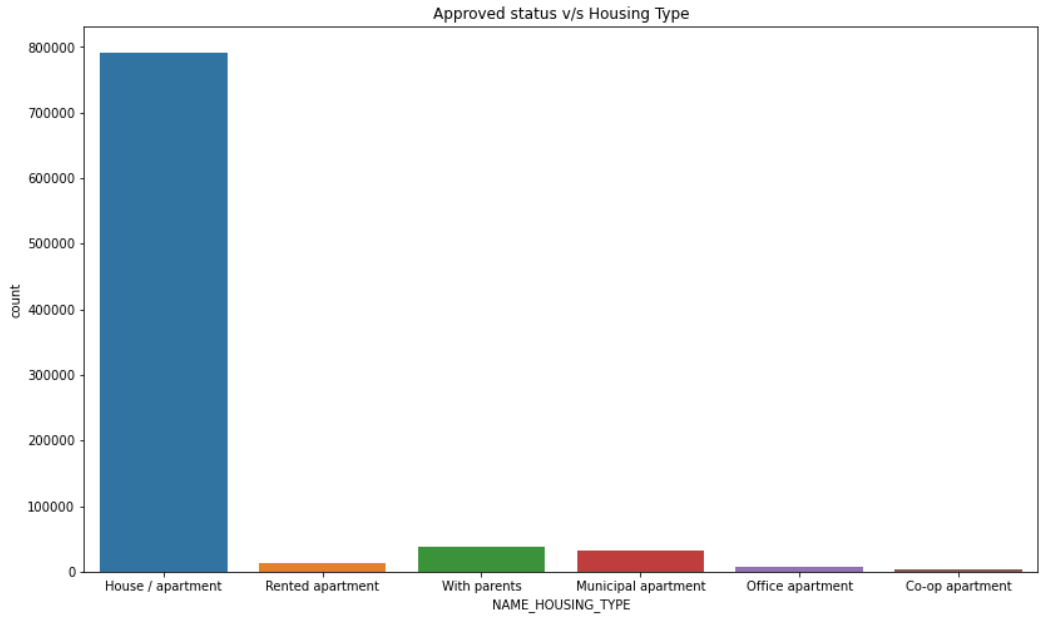
Analysis

- Cash loans are more likely to get approved than revolving loans.
- People who does not own a car are more likely to get approved than who owns a car
- People who own a realty is more likely to get approved
- Female applicant is more likely to get approved.
- People with secondary education in their mid ages i.e 30-50 with AVG income category is likely to get the approval.

2. Bivariate Analysis for few Categorical Variable of combined dataframe



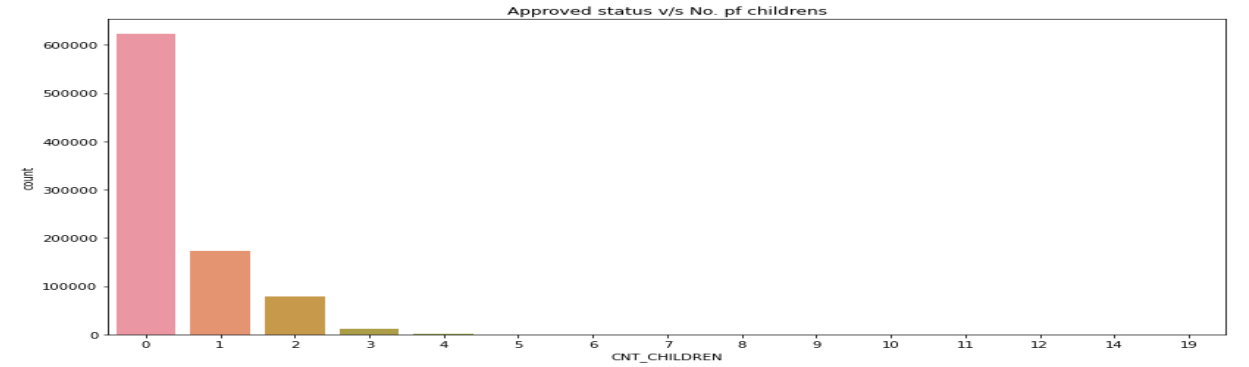


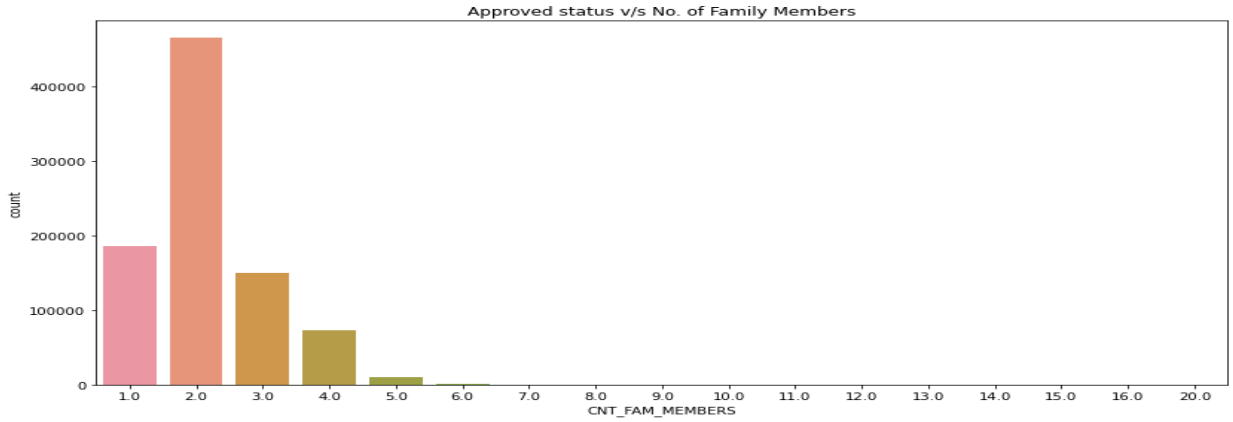


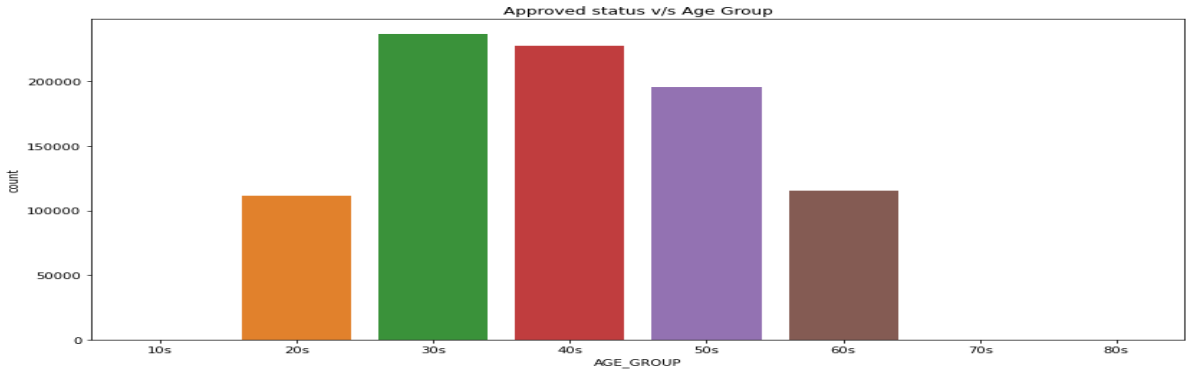
Insights from the above Bivariant analysis for categorical data is as below:

1. Approved status v/s Family Status: People who are married are more likely to get loan approved
2. Approved status v/s Income Type: People who are working are more likely to get loan approved compared to students who are least likely to get loan approved
3. Approved status v/s Housing_type: People who own House/apartment are more likely to get loan approved then compared to rented appartments/ co-op apartment types

3. Bivariant analysis for Numerical Data for the approved data.



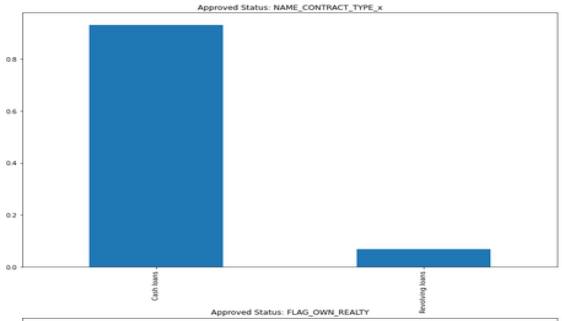


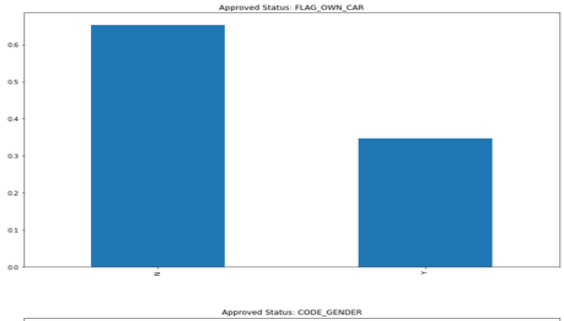


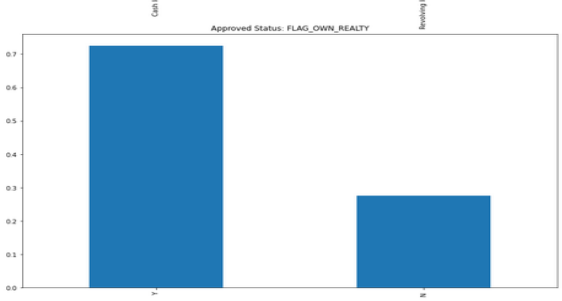
Inference from the above Bivariant analysis for numerical data is as below:

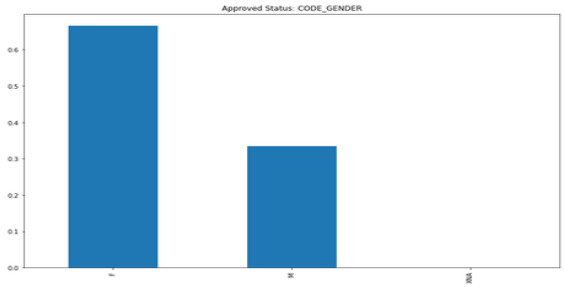
1. Approved status v/s No. of children: People with 0 children are more likely to get loan approved
2. Approved status v/s No. of family members: If the number of people in a family is 2 they are more likely to get loan approved.
3. Approved status v/s Age: People with age in between 30 -50 years are more likely to get loan approved compared to the people in 20s and 60s

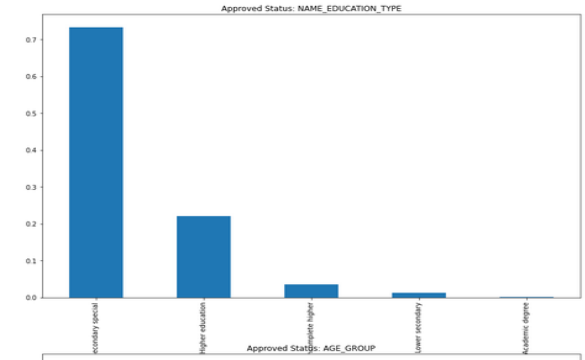
Analysis on People with Contract Status as Refused

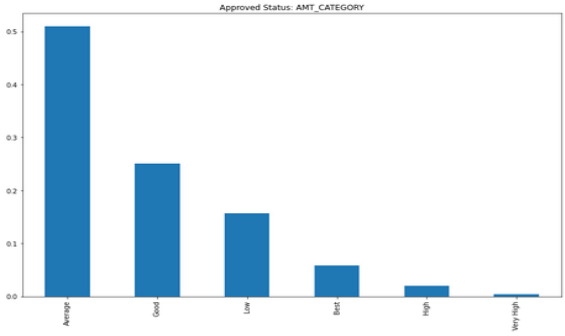


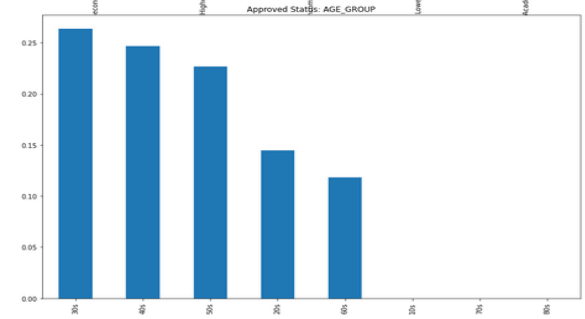












Analysis

- Cash loans are more likely to get refused than revolving loans.
- People who does not own a car are more likely to get refused than who owns a car
- People who own a realty is more likely to get refused
- Female applicant is more likely to get refused.
- People with secondary education in their mid-ages i.e 30-50 with AVG income category is likely to get refused.

Conclusion:

1. People who are in their early days in career are facing difficulties in payment.
2. Retired people are likely to do the payments on time.
3. People earning avg to high income groups are likely to do payments on time.
4. People with proper educational background are likely to do the payments on time.
5. People who are not having children are more likely to return the due amount.
6. People living in small family is more likely to do the payment on time.
7. People who own House/apartment are more likely to get loan approved then compared to rented apartments/ co-op apartment types

Result:

- Learnt to work with huge datasets in python.
- Figured out missing data percentages in the dataset.
- Figured out how to impute the missing values.
- Figured out it is wise to drop irrelevant columns.
- Done univariate bi-variate analysis on numerical & categorical columns.
- Found high correlated columns