Credit EDA Assignment

BY PANKAJ KUMAR

Problem Statement I

- 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. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You 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 has to 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.

Problem Statement - II

- Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.
- Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)
- **Hint:** Note that in EDA, since it is not necessary to replace the missing value, but if you have to replace the missing value, what should be the approach. Clearly mention the approach.
- Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.
- Identify if there is data imbalance in the data. Find the ratio of data imbalance.
- Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.

Assumptions and steps taken

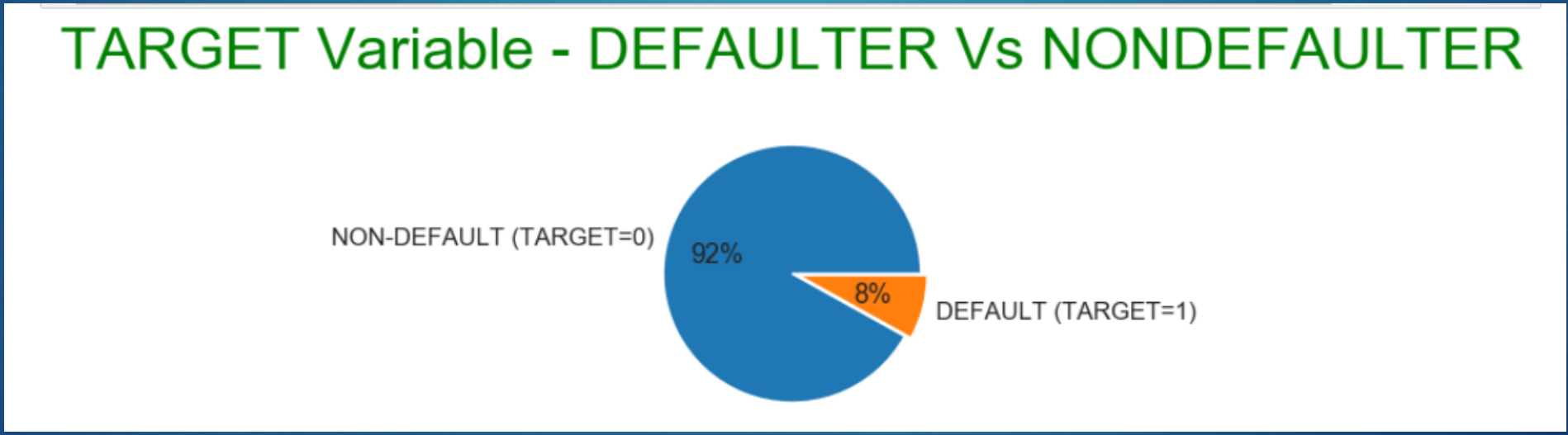
- From both the data set i.e. application_data.csv and previous_application_data.csv columns with more than 40% of missing values are dropped
- In application_data.csv below steps taken for categorical variable
  - NAME_TYPE_SUITE imputed with mode() value, for NaN
  - CODE_GENDER imputed with mode() where it was "XNA“
  - ORGANIZATION_TYPE imputed with mode() where it was "XNA“
  - OCCUPATION_TYPE have many missing values. It is left as it is. No records dropped.

Assumptions and steps taken continued

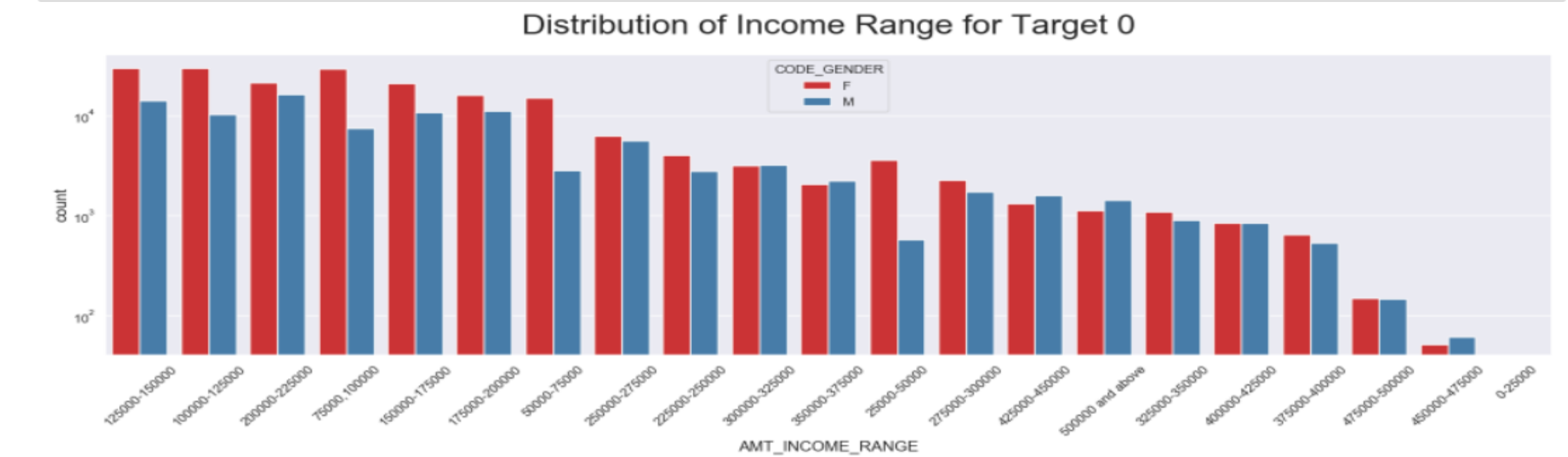
- In previous_application_data.csv below steps taken for categorical variable
  - NAME_CONTRACT_TYPE is imputed with mode where it was "XNA"
  - NAME_CLIENT_TYPE is imputed with mode where it was "XNA"
  - NAME_CASH_LOAN_PURPOSE majority of records have value "XNA" and "XNP". No steps taken for these.
  - NAME_PAYMENT_TYPE majority of records have value "XNA". No steps taken for these.
  - CODE_REJECT_REASON majority of records have value "XNA" and "XNP". No steps taken for these.
  - NAME_CLIENT_TYPE is imputed with mode where it was "XNA"
  - NAME_GOODS_CATEGORY majority of records have value "XNA". No steps taken for these.
  - NAME_PORTFOLIO majority of records have value "XNA". No steps taken for these.
  - NAME_PRODUCT_TYPE majority of records have value "XNA". No steps taken for these.
  - NAME_SELLER_INDUSTRY majority of records have value "XNA". No steps taken for these.
  - NAME_YIELD_GROUP majority of records have value "XNA". No steps taken for these.

Data imbalance in the data

- Its clear that there is an imbalance between people who defaulted and who didn't default. More than 92% of people didn't default as opposed to 8% who defaulted.



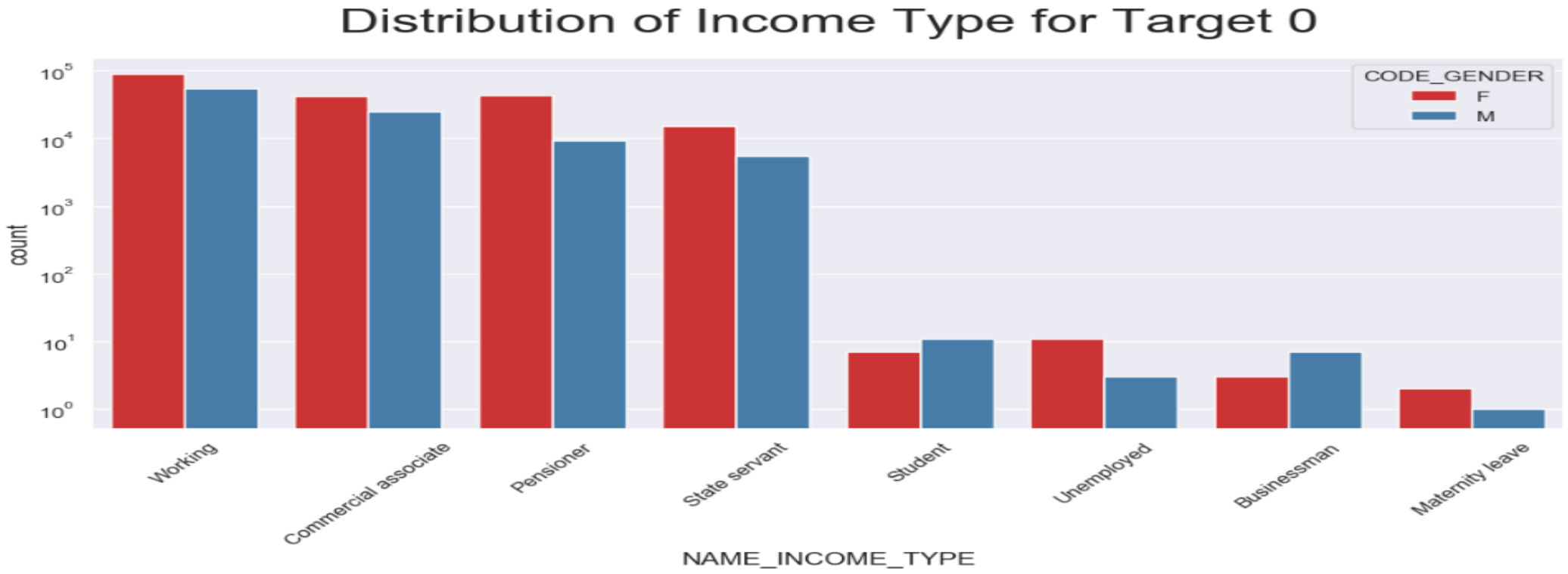
Income Range Distribution for Male and Female



Points to be concluded from the above graph.

Female counts are higher than male.
Income range from 100000 to 200000 is having more number of credits.
This graph show that females are more than male in having credits for that range.
Very less count for income range 400000 and above.

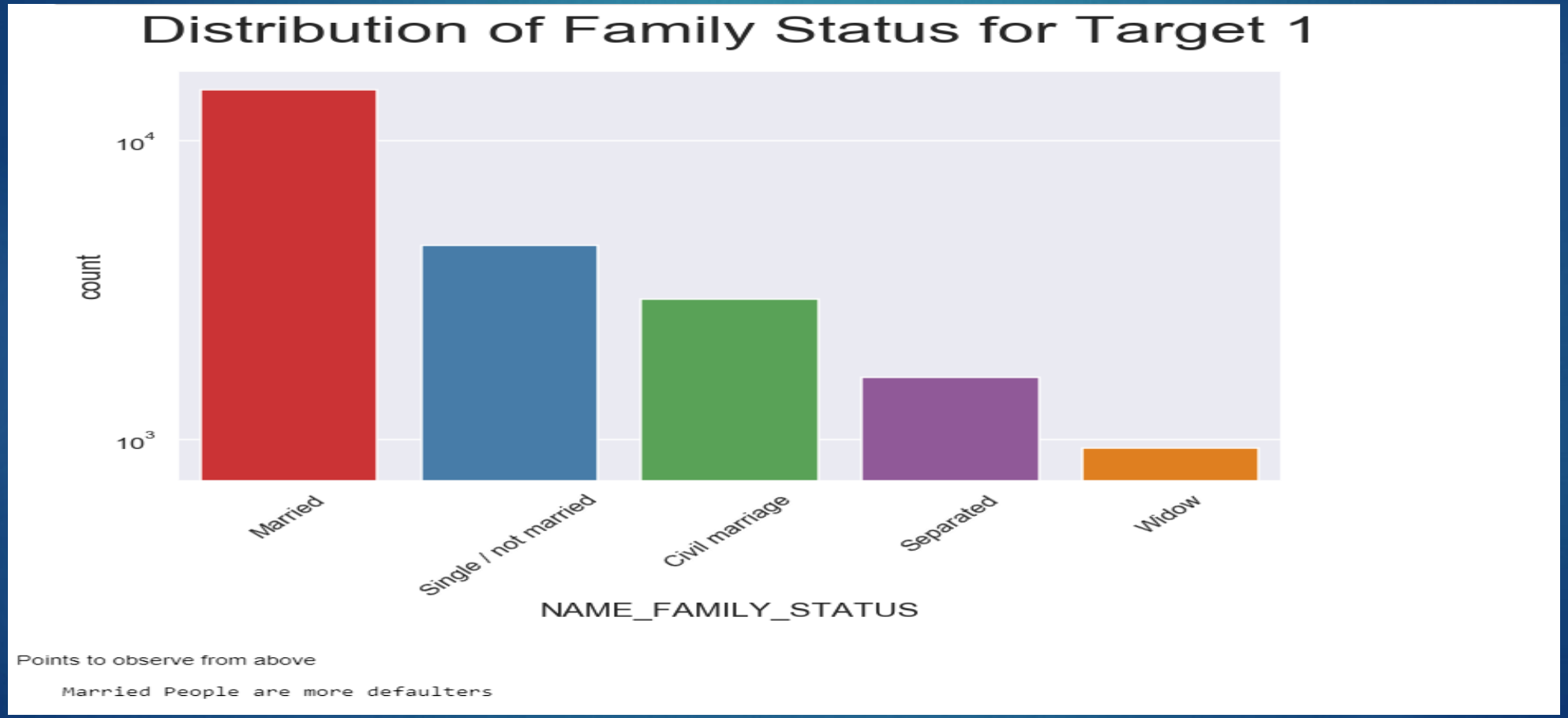
Income Type distribution for Male and Female



Points to be concluded from the above graph.

For income type 'working', 'commercial associate', and 'State Servant' the number of credits are higher than others.
For this Females are having more number of credits than male.
Less number of credits for income type 'student' ,'pensioner', 'Businessman' and 'Maternity leave'.

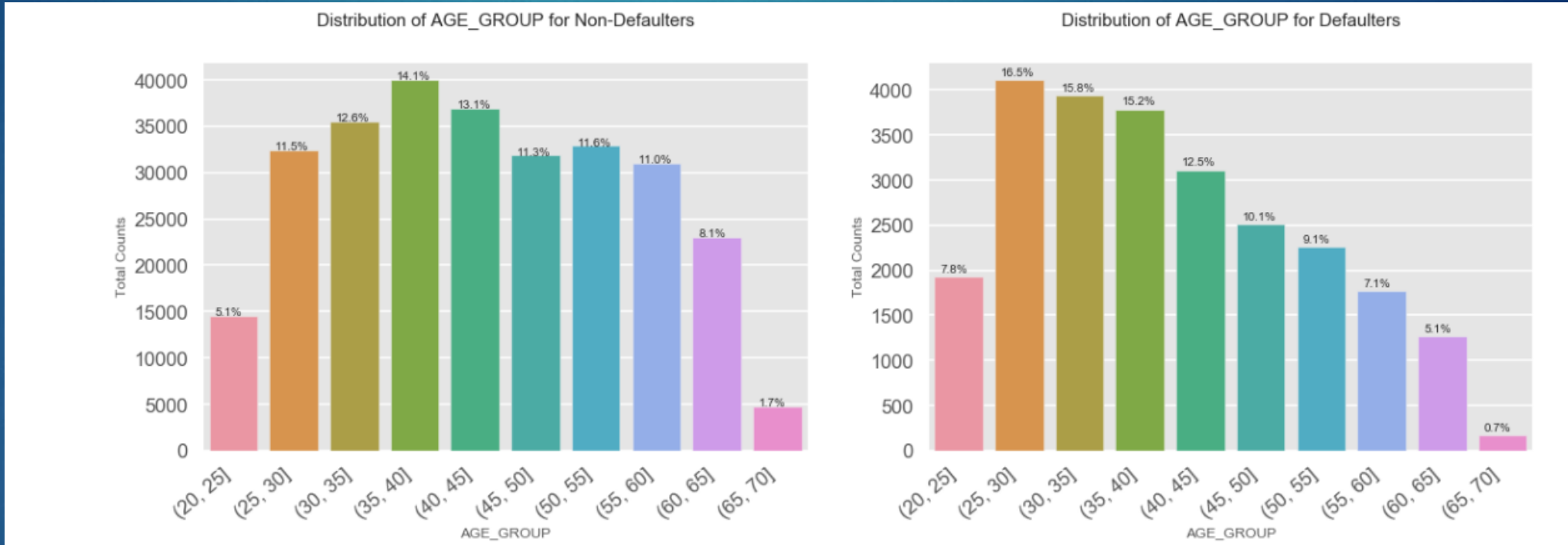
Distribution of family Status of Defaulters



Points to observe from above

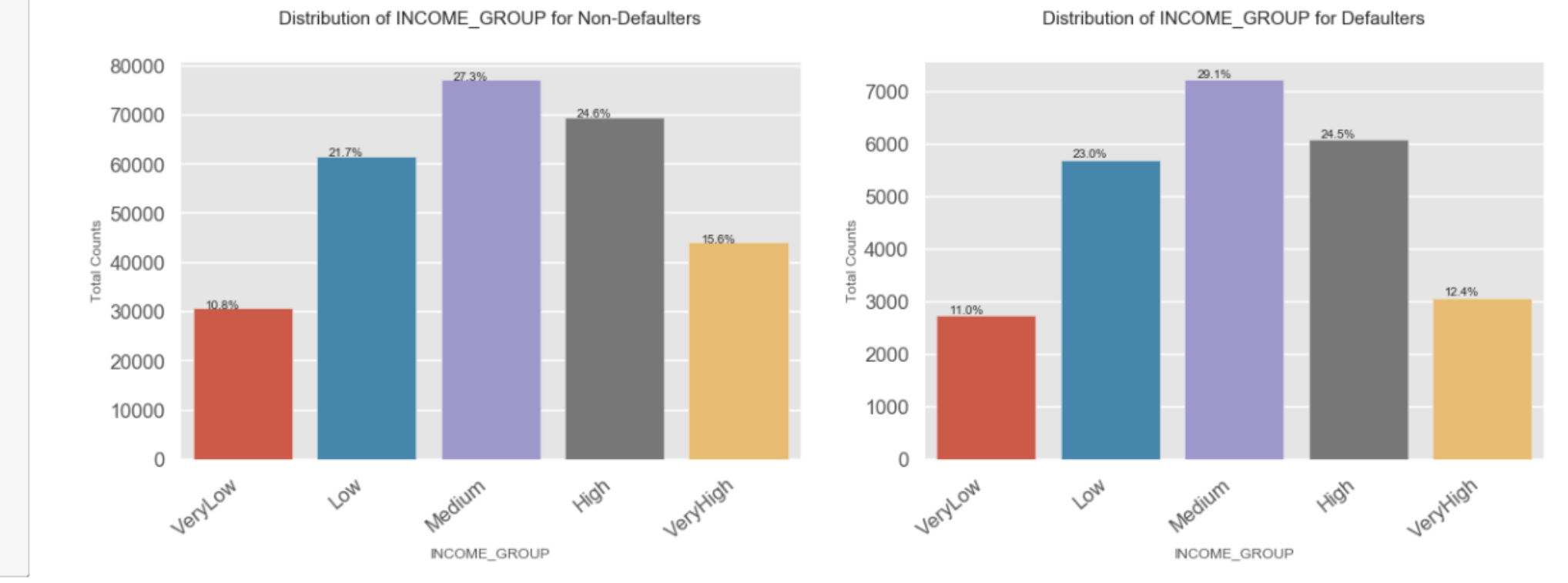
Married People are more defaulters

Defaulters and Non-Defaulters based on age group



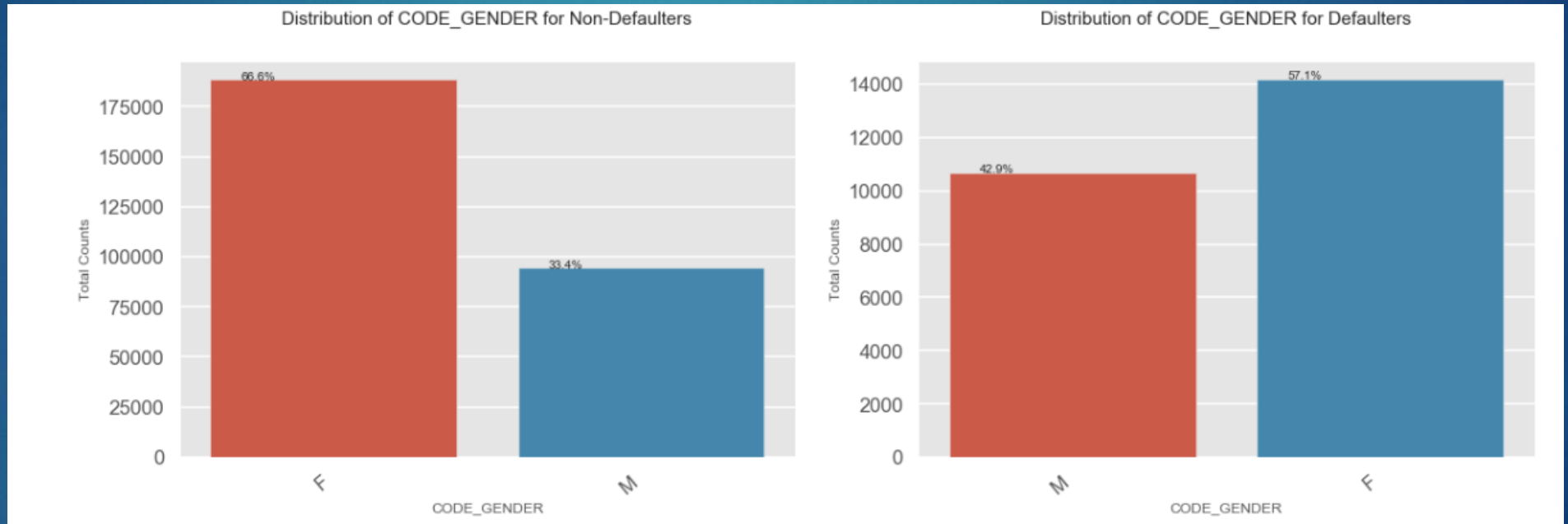
We see that (25,30] age group tend to default more often. So they are the riskiest people to loan to. With increasing age group, people tend to default less starting from the age 25. One of the reasons could be they get employed around that age and with increasing age, their salary also increases.

Defaulters and Non-Defaulters based on Income



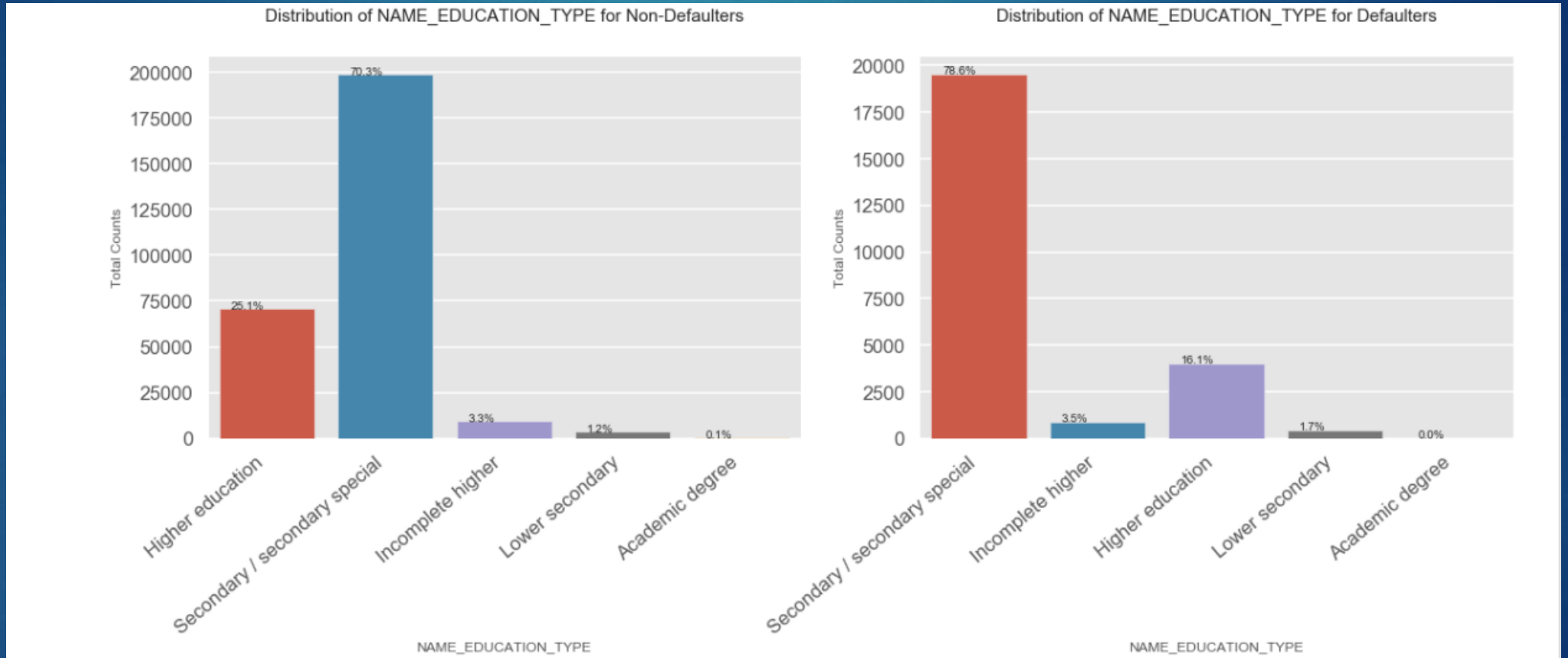
The Very High income group tend to default less often. They contribute 12.4% to the total number of defaulters, while they contribute 15.6% to the Non-Defaulters.

Defaulters and Non-Defaulters ratio of Male and Female



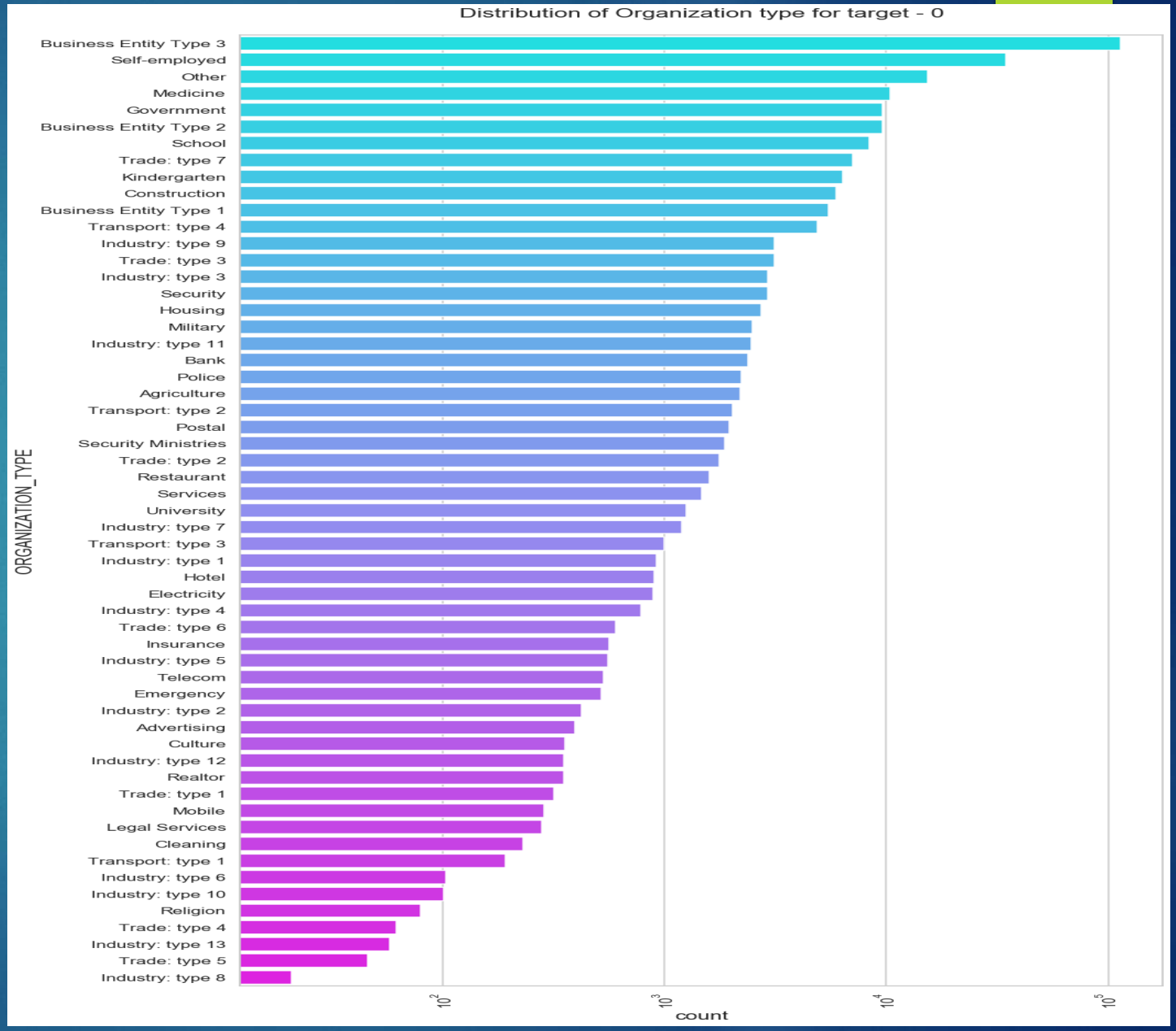
We can see that Female contribute 67% to the non-defaulters while 57% to the defaulters. We can conclude that We see more female applying for loans than males and hence the more number of female defaulters as well. But the rate of defaulting of FEMALE is much lower compared to their MALE counterparts.

Defaulters and Non-Defaulters Based on Education

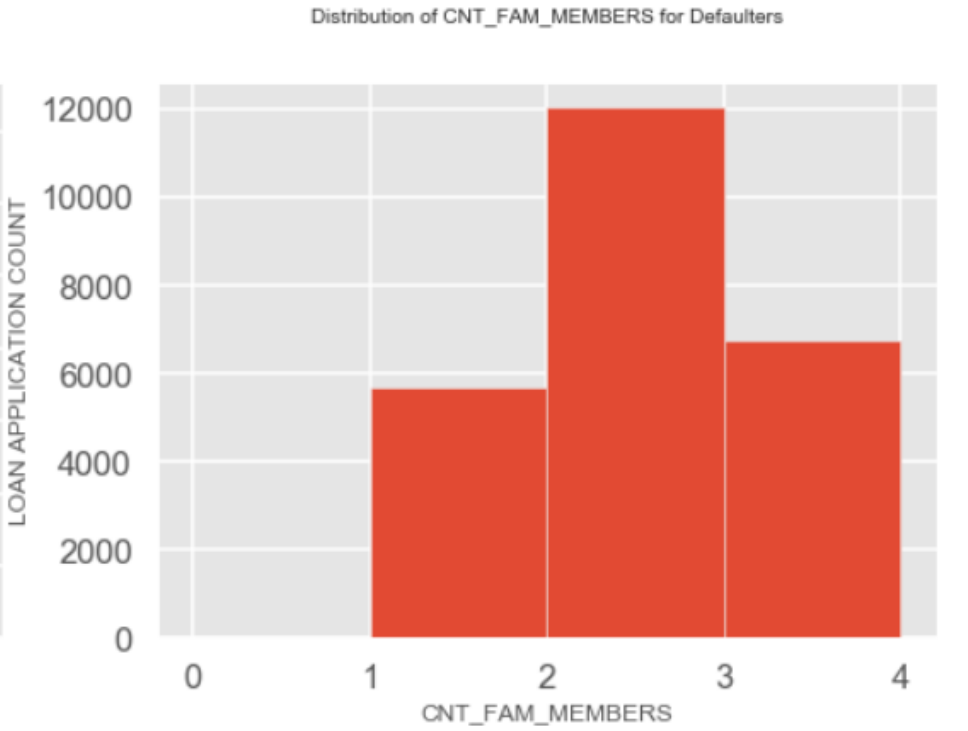


Almost all of the Education categories are equally likely to default except for the higher educated ones who are less likely to default and secondary educated people are more likely to default

Distribution of Organization Type for Non-Defaulters

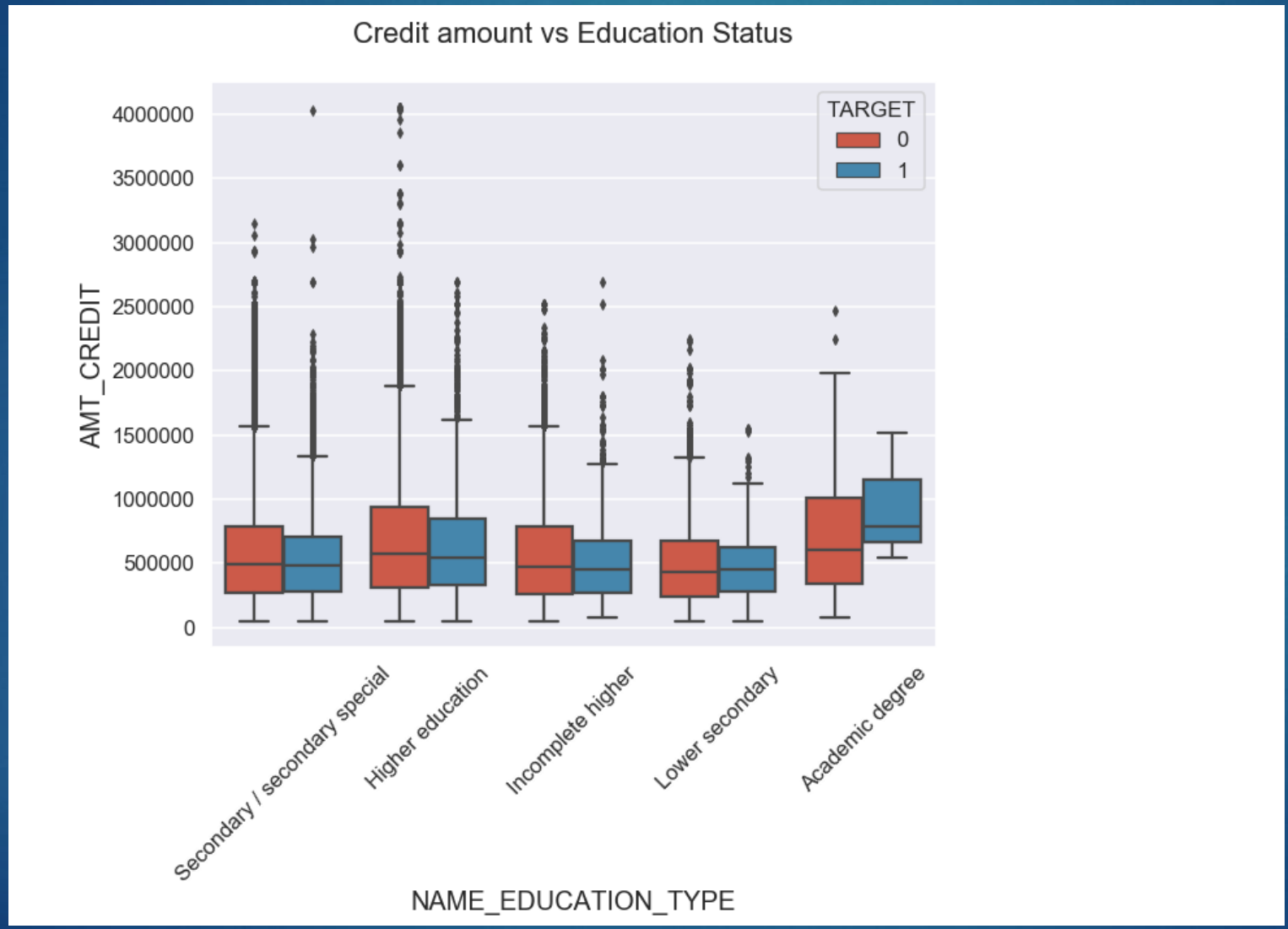


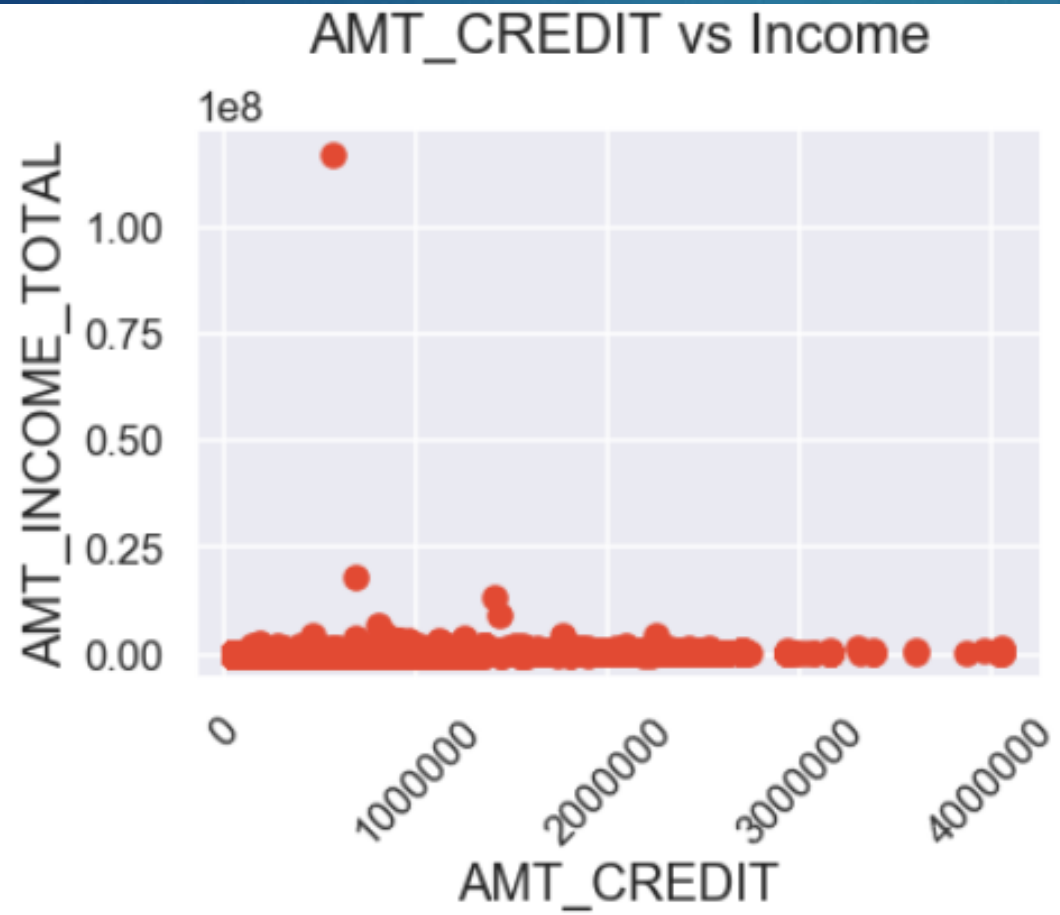
Number of applicants respect to family member count



We can see that a family of 3 applies loan more often than the other families

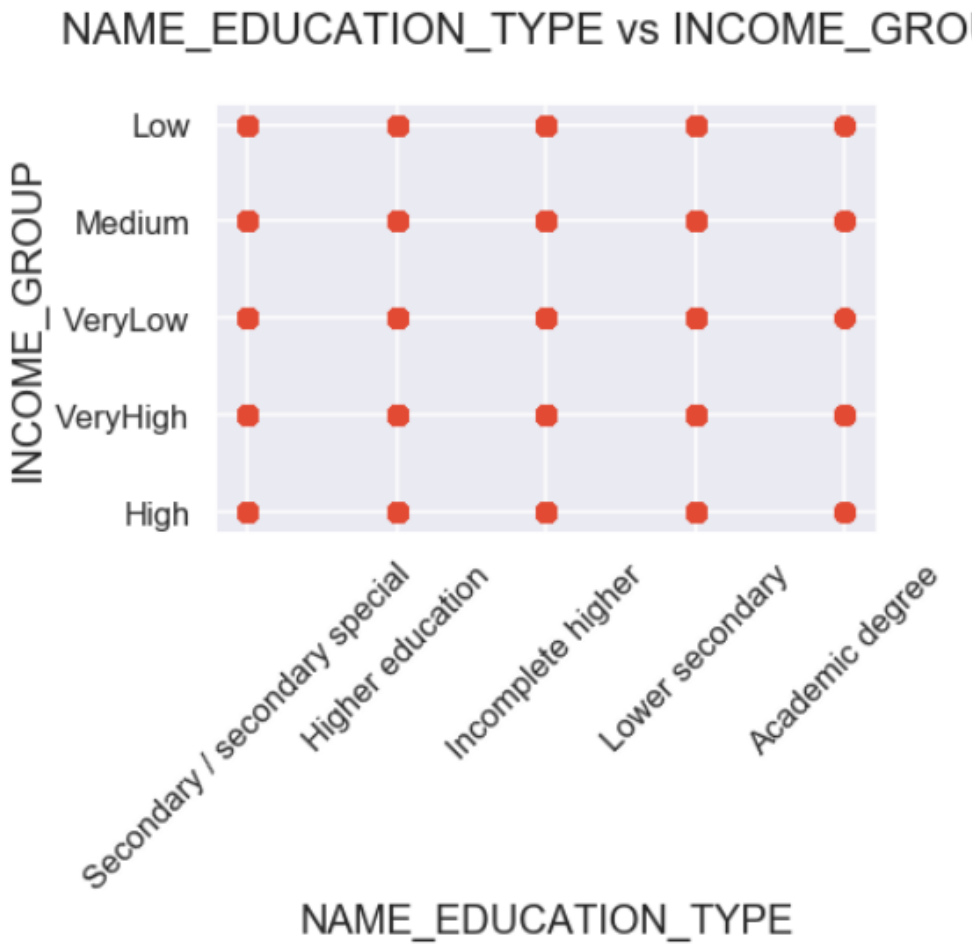
Bivariate Analysis



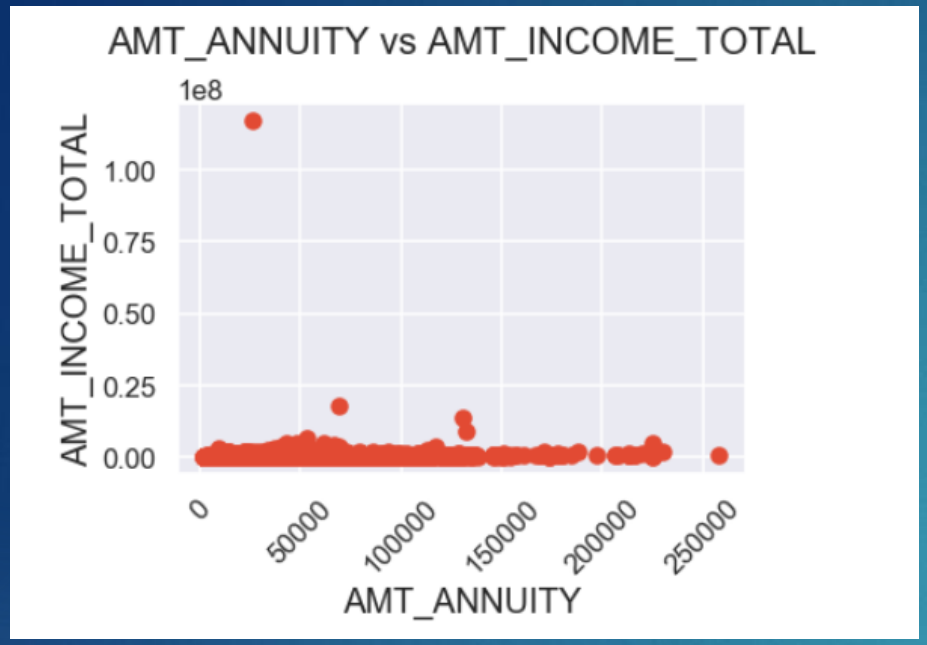


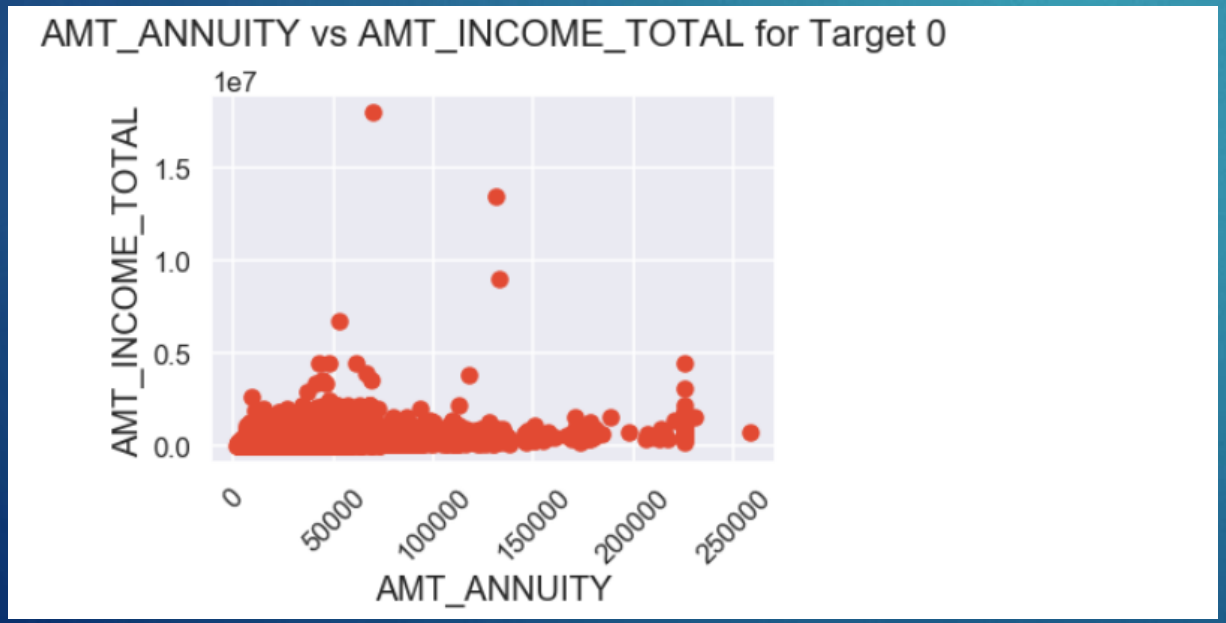
No Strong relationship found between AMT_INCOME_TOTAL and AMT_CREDIT

Education Vs Income

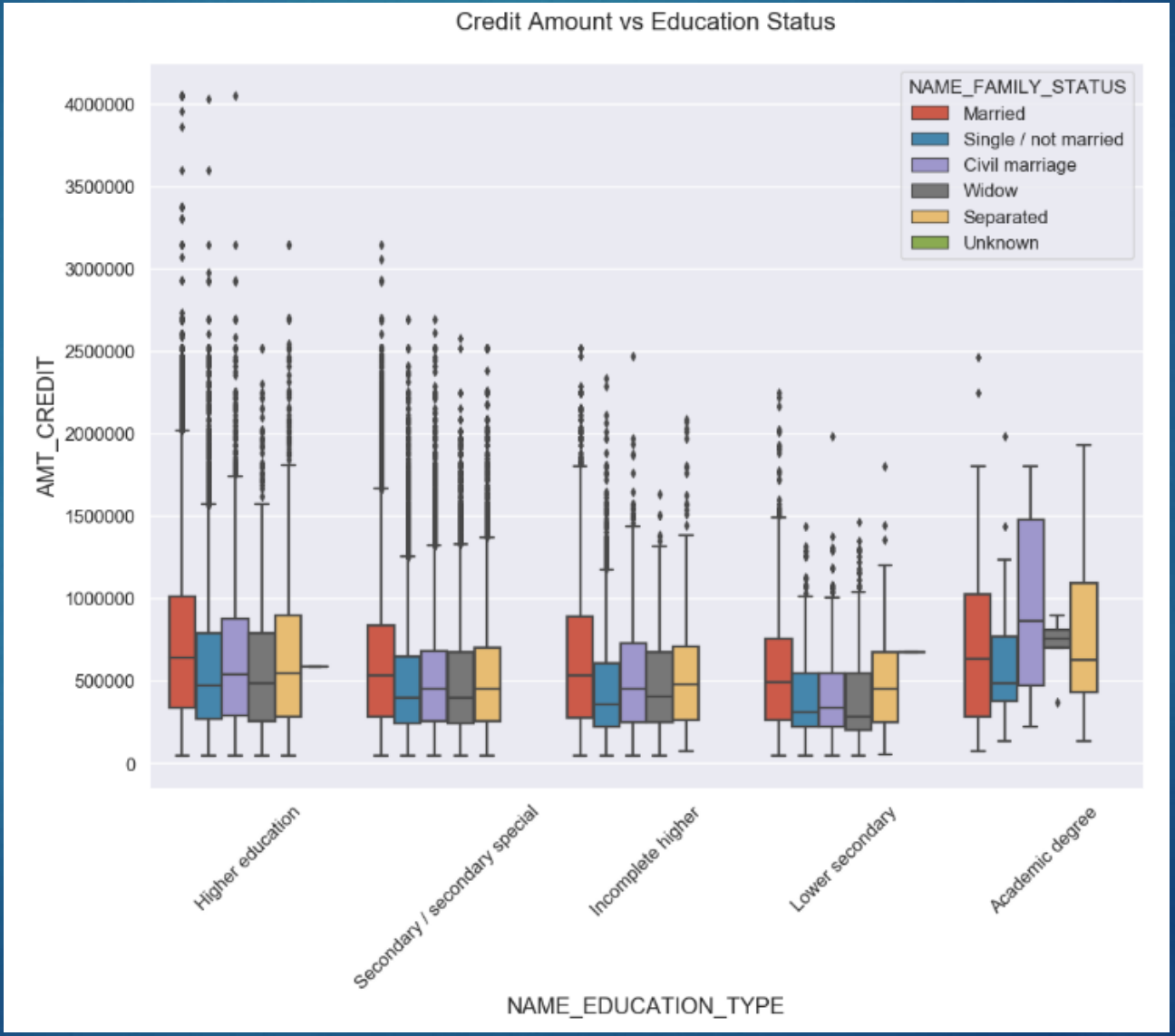


Income is equally distributes across all the Education


AMT_ANNUITY vs AMT_INCOME_TOTAL
1e8
AMT_INCOME_TOTAL
1.00
0.75
0.50
0.25
0.00
0
50000
100000
150000
200000
250000
AMT_ANNUITY


AMT_ANNUITY vs AMT_INCOME_TOTAL for Target 0
1e7
AMT_INCOME_TOTAL
1.5
1.0
0.5
0.0
0
50000
100000
150000
200000
250000
AMT_ANNUITY

Credit Amount Vs Education for Target 0



Top 10 Correlation for Target Value 0

FLAG_EMP_PHONE	DAYS_BIRTH	0.622073
AMT_ANNUITY	AMT_CREDIT	0.771309
AMT_GOODS_PRICE	AMT_ANNUITY	0.776686
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830381
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859332
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861
CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149
AMT_GOODS_PRICE	AMT_CREDIT	0.987250
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508

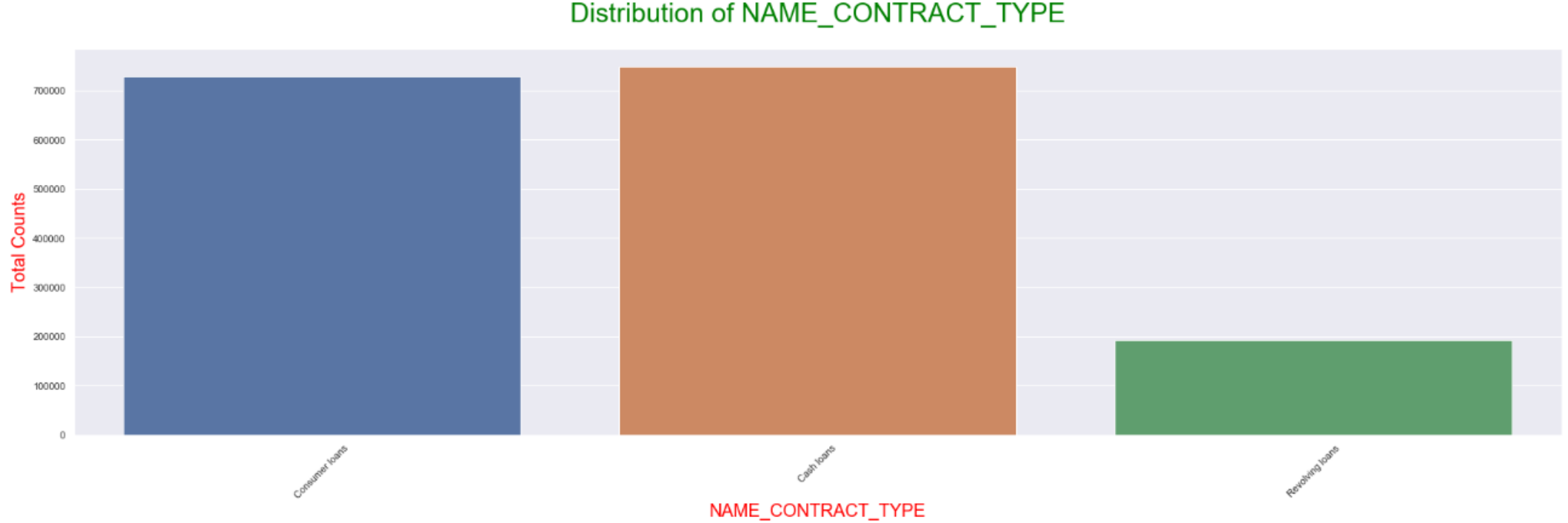
dtype: float64

Top 10 Correlation for Target Value 1

FLAG_DOCUMENT_6	DAYS_EMPLOYED	0.617307
AMT_ANNUITY	AMT_CREDIT	0.752195
AMT_GOODS_PRICE	AMT_ANNUITY	0.752699
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778540
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.868994
CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637
AMT_GOODS_PRICE	AMT_CREDIT	0.983103
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998269

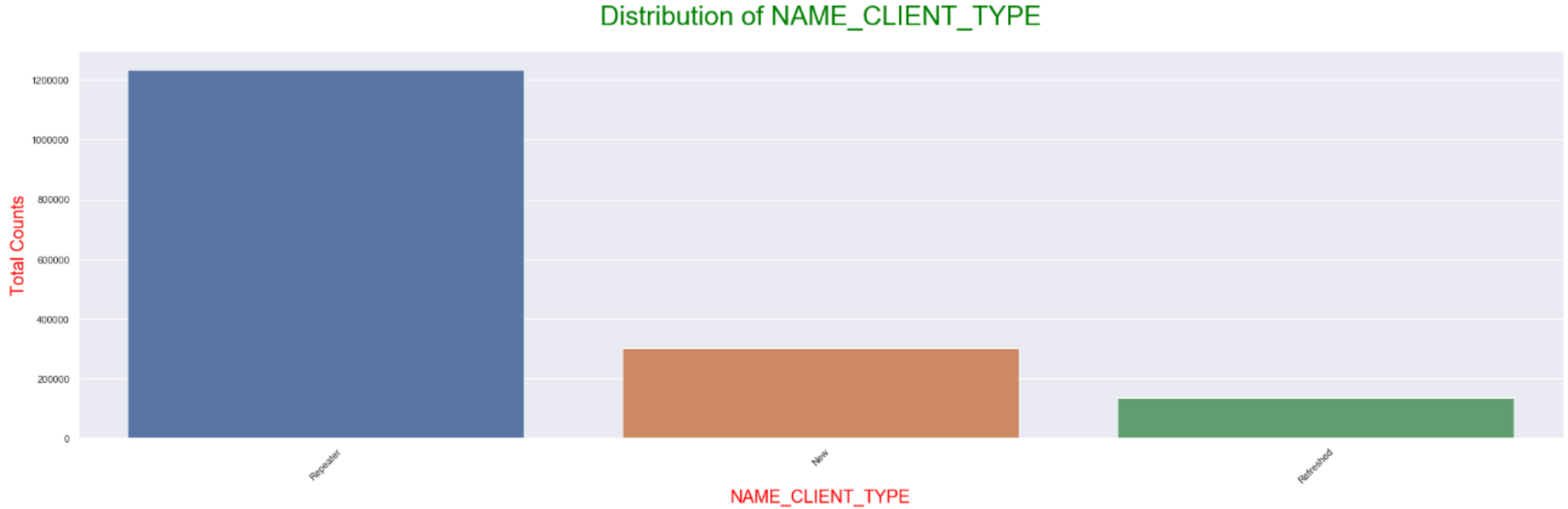
dtype: float64

Previous Application Data Univariate Analysis



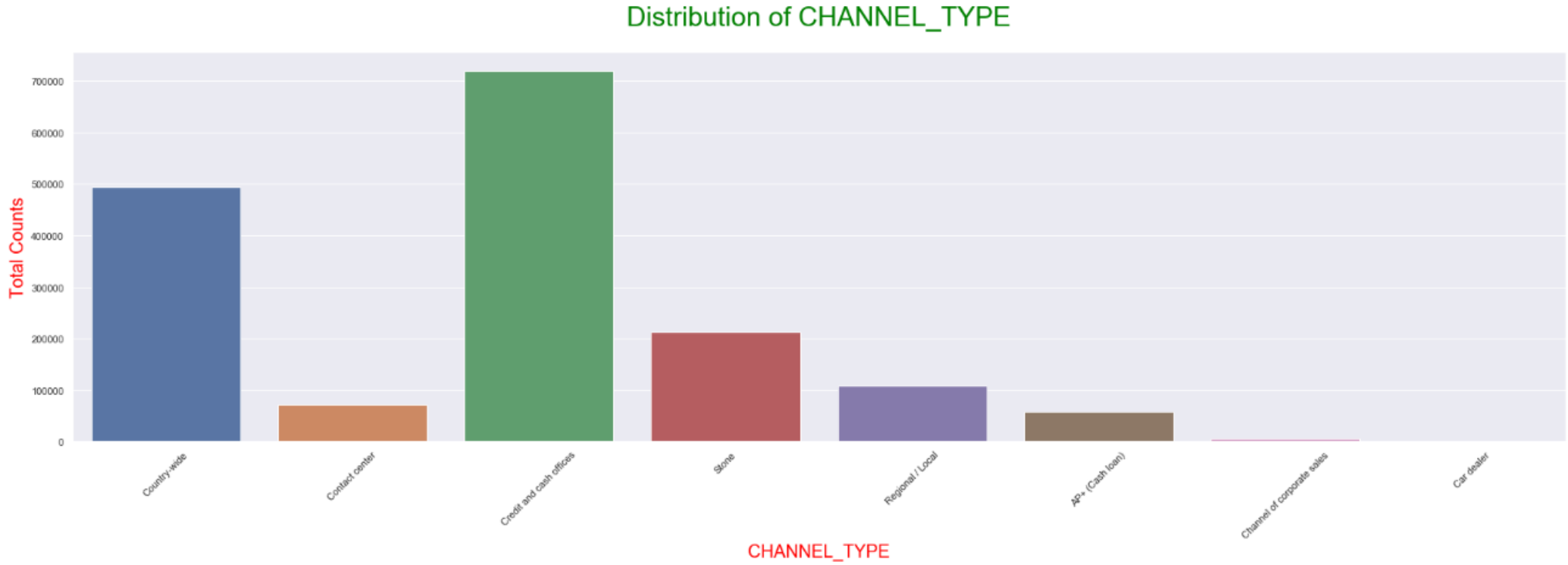
From above graph we can conclude that that Cash Loans are highest.

Previous Application Data Univariate Analysis continued……



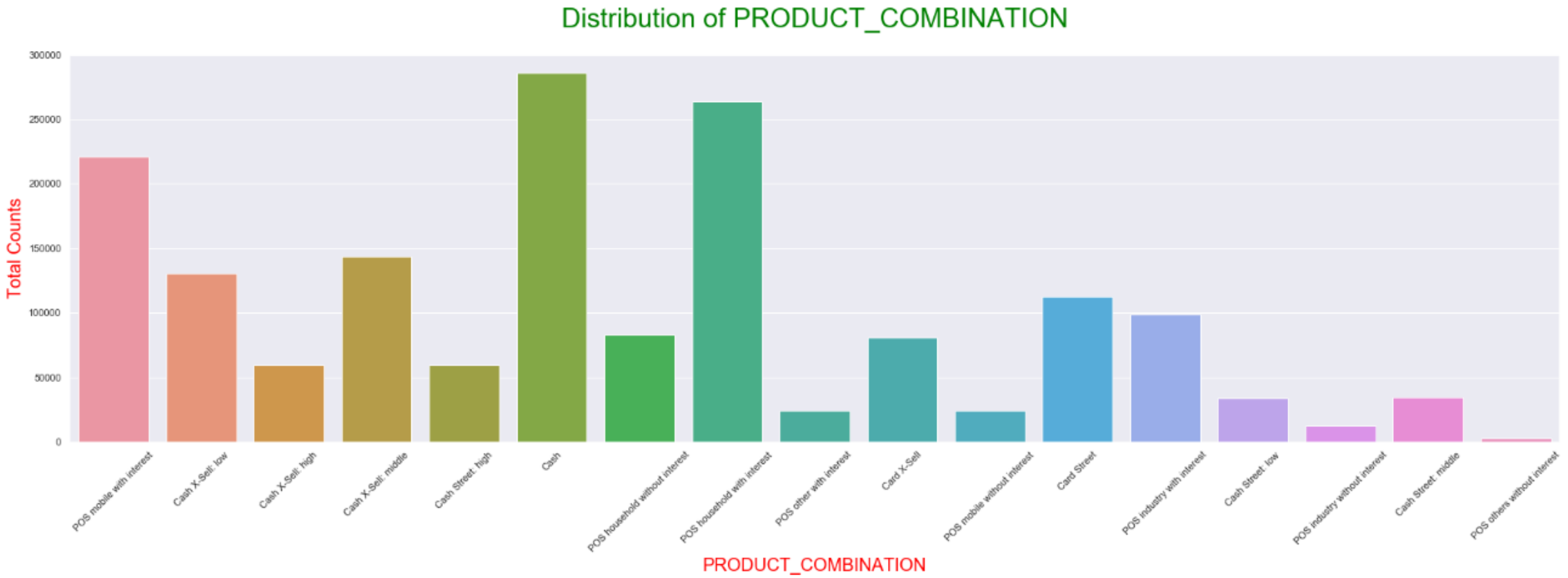
From the Above graph we can conclude that Repeater type clients are marginally very high.

Previous Application Data Univariate Analysis continued……

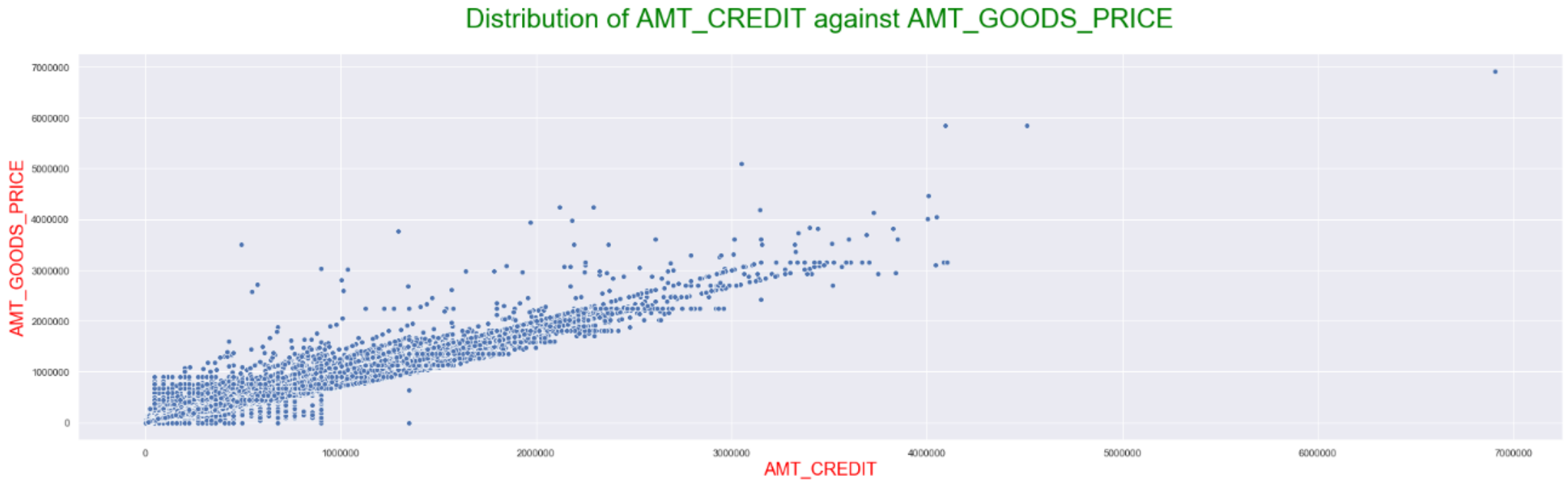


From CHANNEL_TYPE distribution we can observe that "Credit And Cash Officers" has highest volume of clients acquired followed by "Country-Wide"

Previous Application Data Univariate Analysis continued……

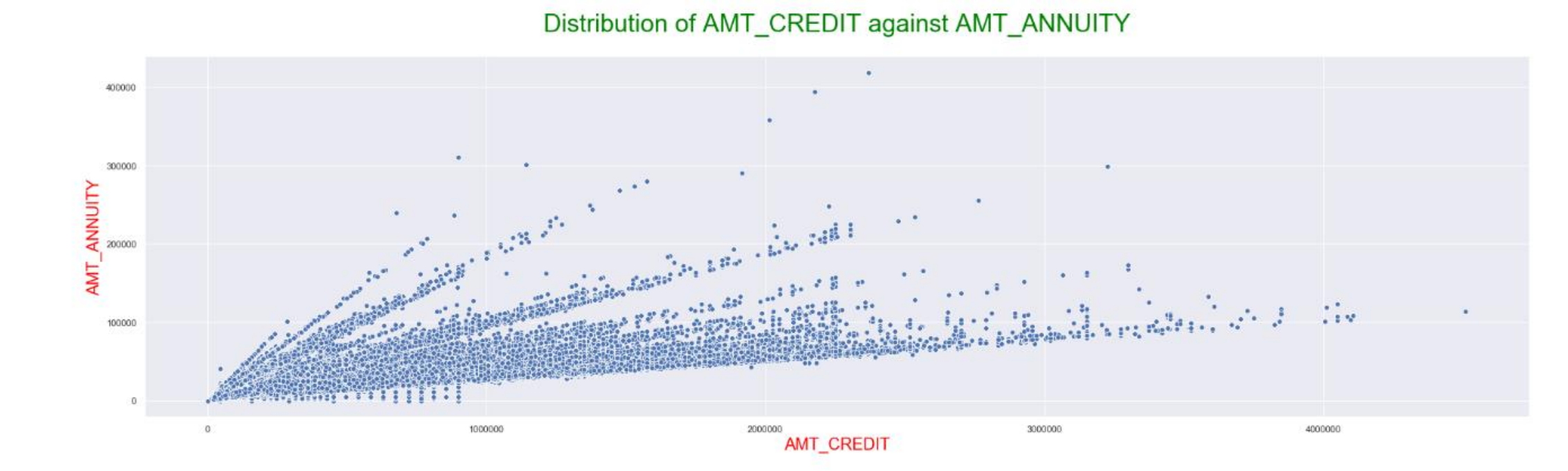


Previous Application data Bivariate Analysis



From the above plot, we can coclude that there is positive linear relationship between "AMT_CREDIT" and "AMT_GOODS_PRICE"

Previous Application data Bivariate Analysis continued ....



Top 10 correlation in Previous Application Data

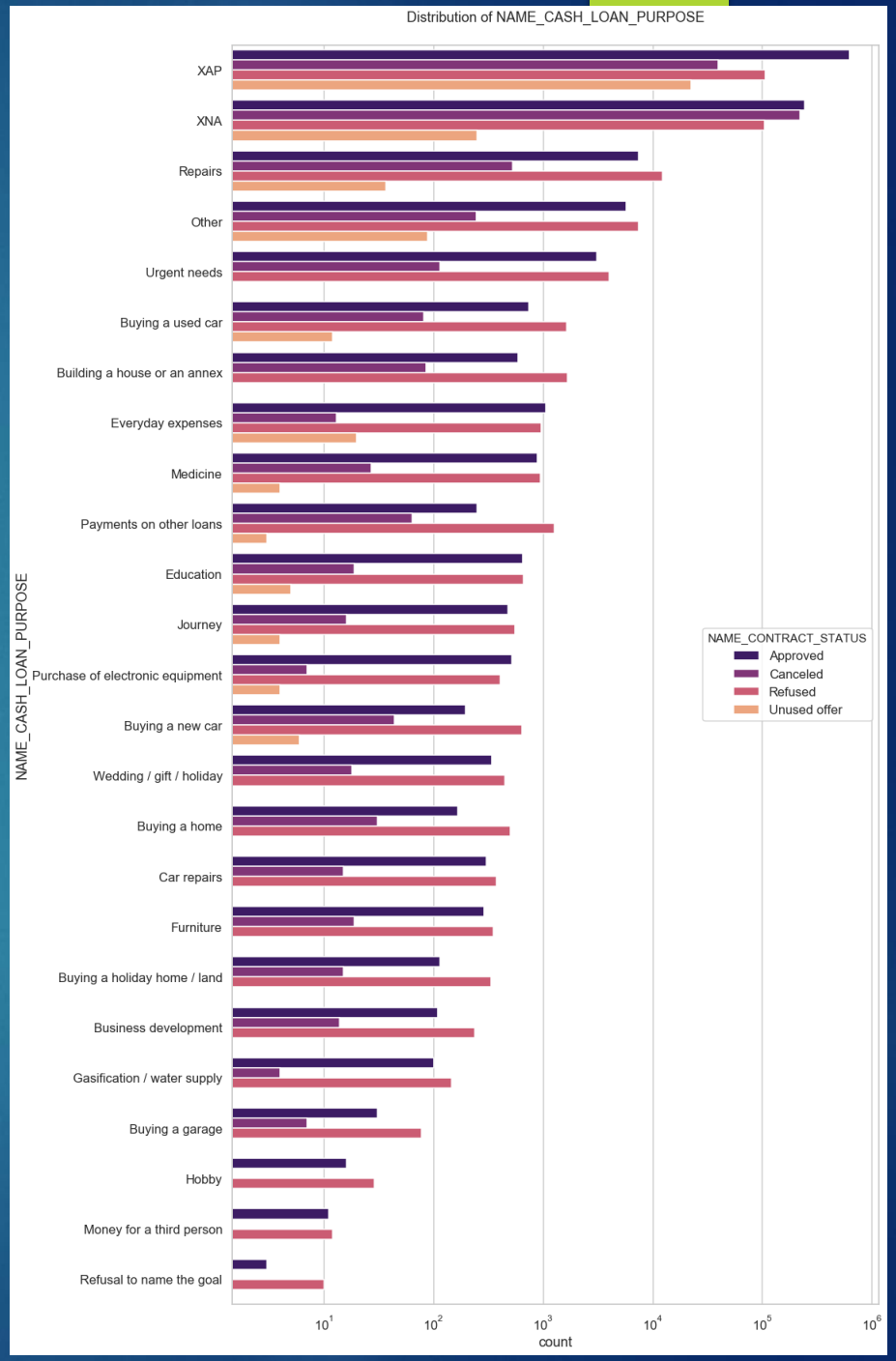
```
CNT_PAYMENT       AMT_ANNUITY         0.394535
                  AMT_GOODS_PRICE     0.672129
                  AMT_CREDIT          0.674278
                  AMT_APPLICATION     0.680630
AMT_APPLICATION   AMT_ANNUITY         0.808872
AMT_CREDIT        AMT_ANNUITY         0.816429
AMT_GOODS_PRICE   AMT_ANNUITY         0.820895
AMT_CREDIT        AMT_APPLICATION     0.975824
AMT_GOODS_PRICE   AMT_CREDIT          0.993087
                  AMT_APPLICATION     0.999884
dtype: float64
```

Heat Map for Previous Application Correlation

Merged data Analysis

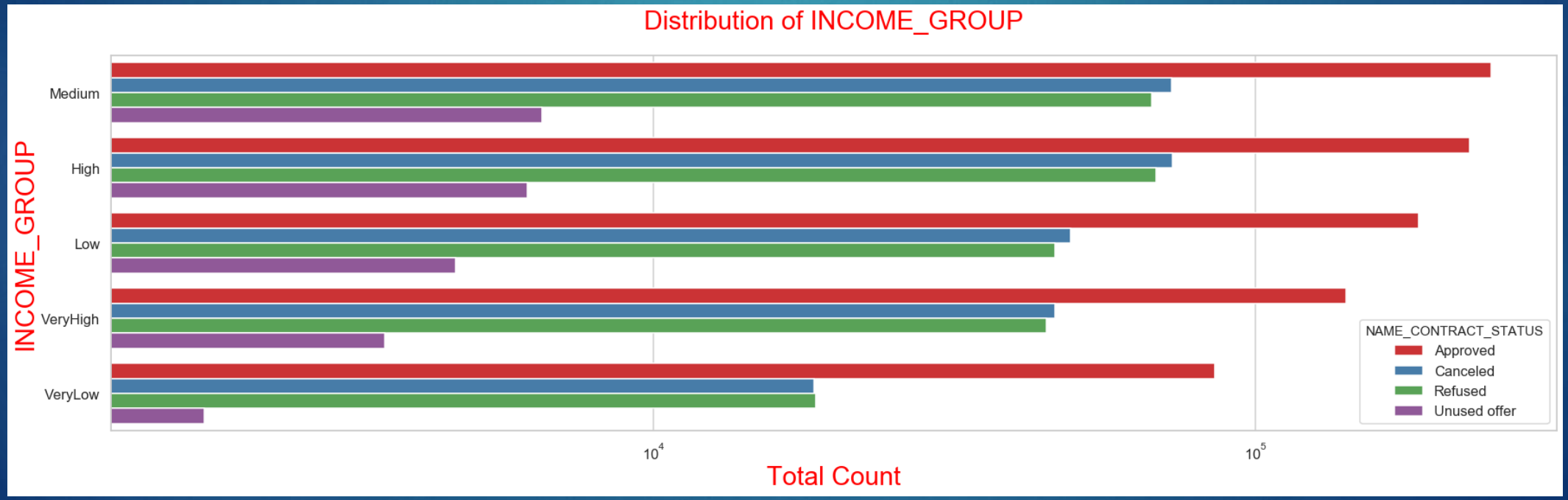
Ignoring XNA and XNP, we can observe below points from above Graph

1. Repairs has highest volume of loan Approved as well as loan Refused
2. Other category have most Unused load status
3. Where purpose is "Refusal to name the goal" bank has refused more than the approved
4. Paying other loans and buying a new car is having significant higher rejection than approves.

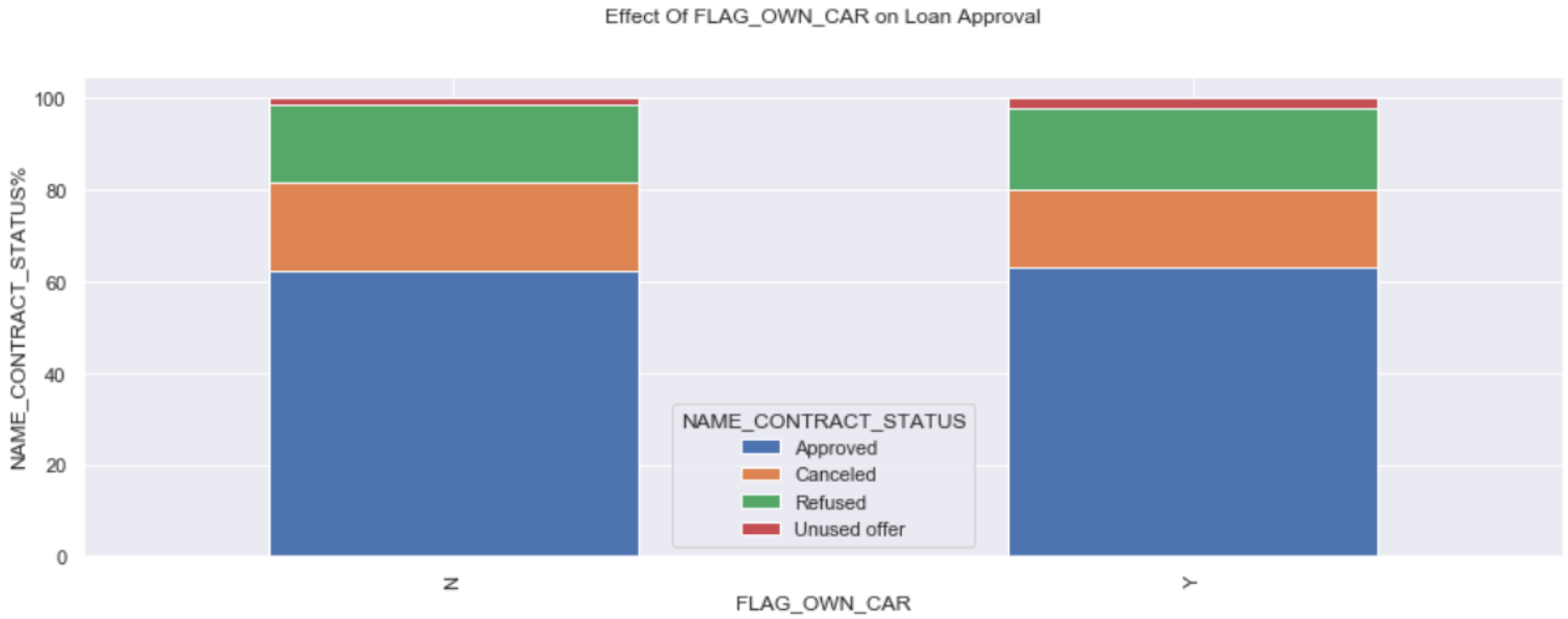


Points to observe from Below graph

1. Income Group Medium and High have almost same ratio of Loan Approve, Canceled, Refused, Unused Offer
2. Income Group Very low have same ratio of Canceled and Refused loan status
3. Income group VeryLow has highest approved to unused offer ratio



Impact on Loan Status if owing a car or not in Merged data

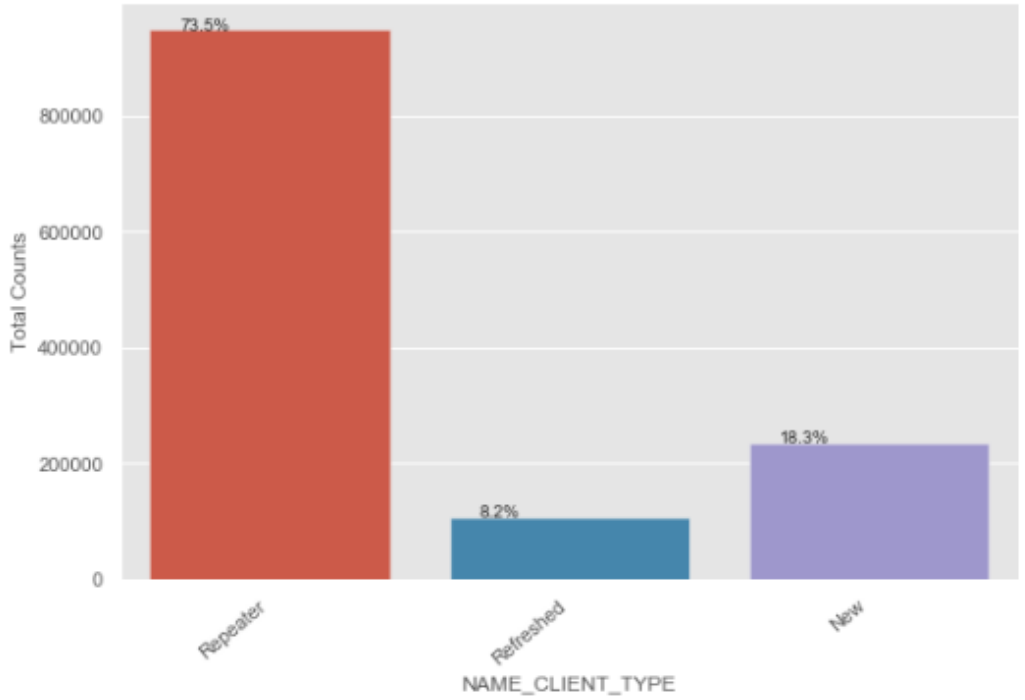


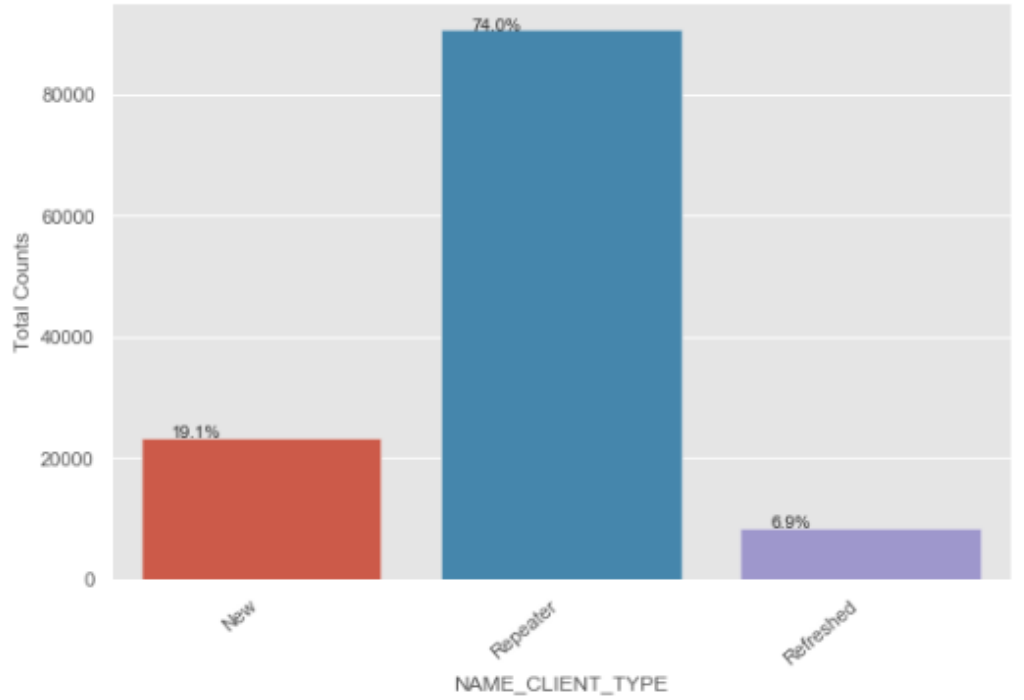
from above graph we can observe that owing a car has not much impact on loan approval or refuse

Loan Status by Age Group in Merged Data

Age_group has no impact on loan status

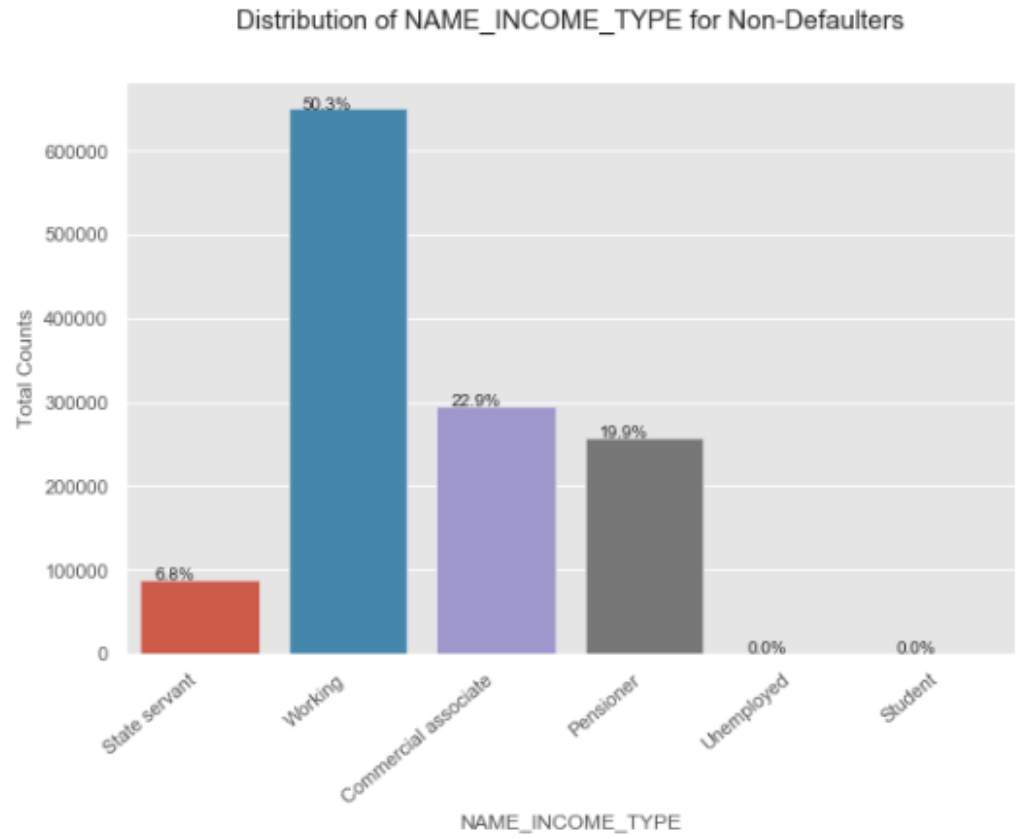
Client Type for Defaulters and Non-defaulters

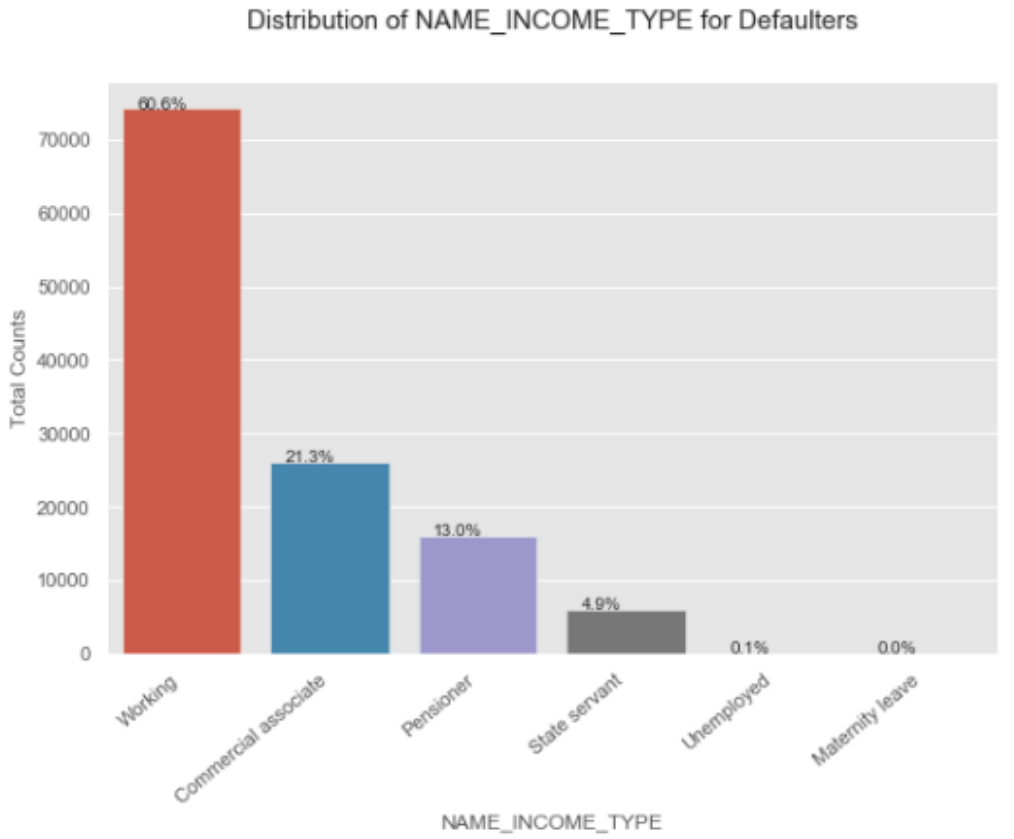




Not much diffrence in Client type of Defaulters and Non-defaulters

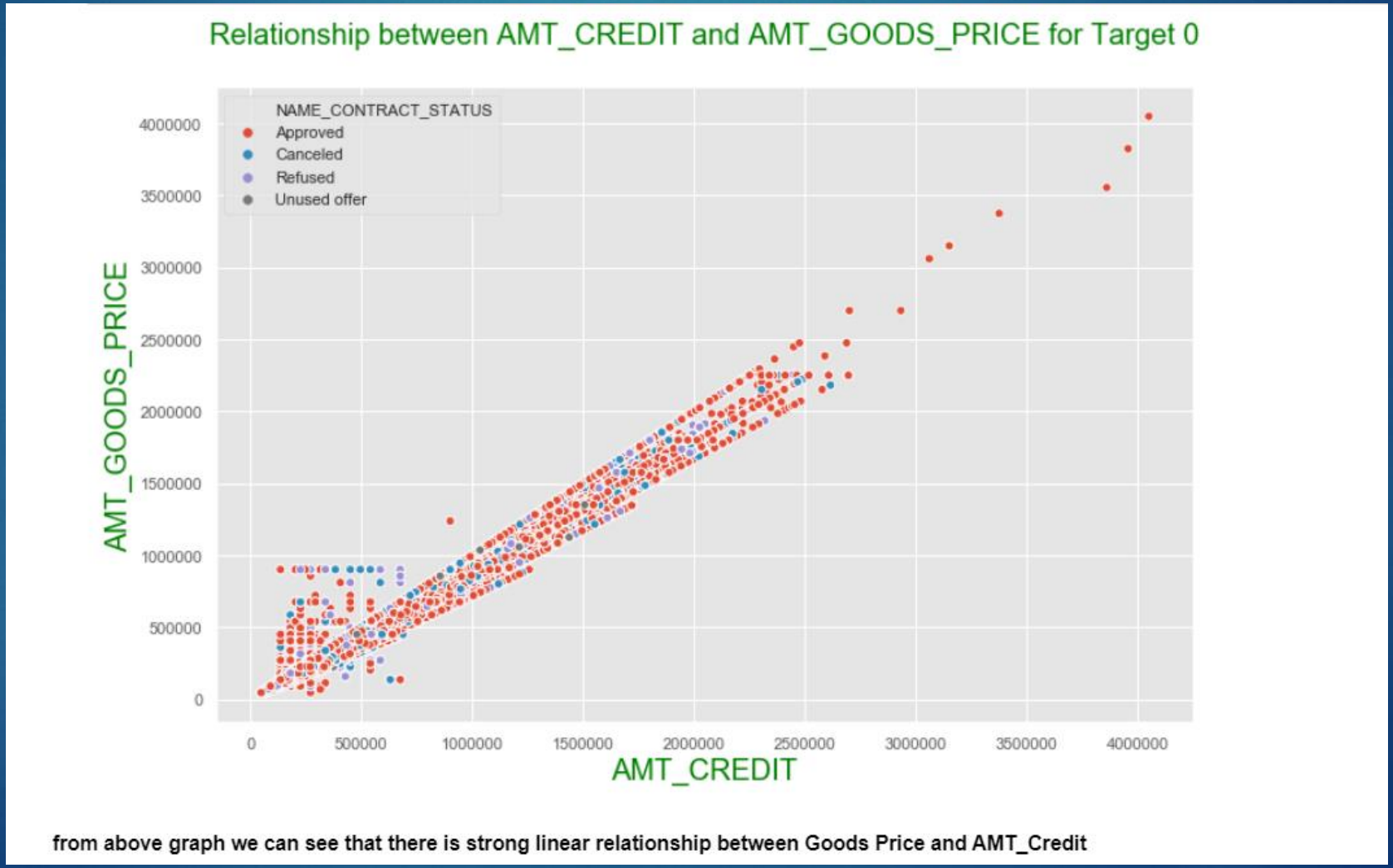
Defaulters VS Non-defaulters for INCOME_TYPE in merged data





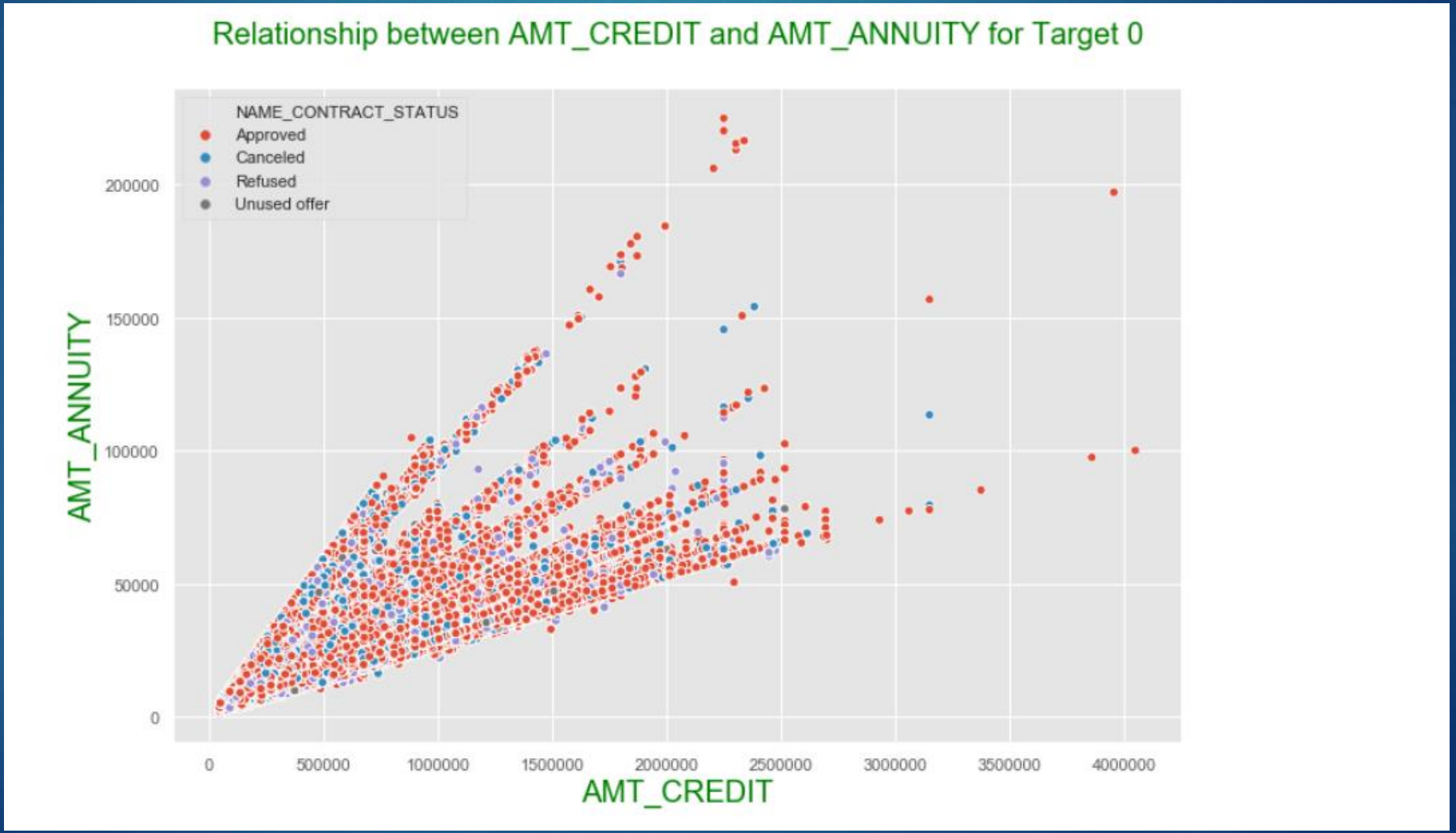
Working Group INCOME_TYPE are more defaulters, but same time they are the most Non-Defaulters as well. But, Defaulters % is more

Relationship between Goods Price and AMT_Credit for Target 0



from above graph we can see that there is strong linear relationship between Goods Price and AMT_Credit

AMT_CREDIT vs AMT_ANNUITY for Merged Data Target 0



Conclusion

- Bank Should Focus on All Age Group
- Bank Should focus on All Education
- Bank is getting Most business from INCOME_TYPE working, but the same time this group contributes most in Defaulters
- Channel "Credit and Cash Officiers" has acquired highest volume of clients.
- From previous application data we observed Repeaters clients are most, so bank should keep targeting its Non-defaulters client.
- Bank has more Female clients than the Male
- Most applicants have 3 Family members
- So bank should focus on clients with any age, education and Having salary range High and Medium with Family member 3

Thank You !!!