CREDIT EDA CASE STUDY

PRANAB DEY

Problem Statement

To understand how the bank approves and refuses loan. Find out different patterns and represent the outcomes to help the bank reduce the credit risk and interest risk.

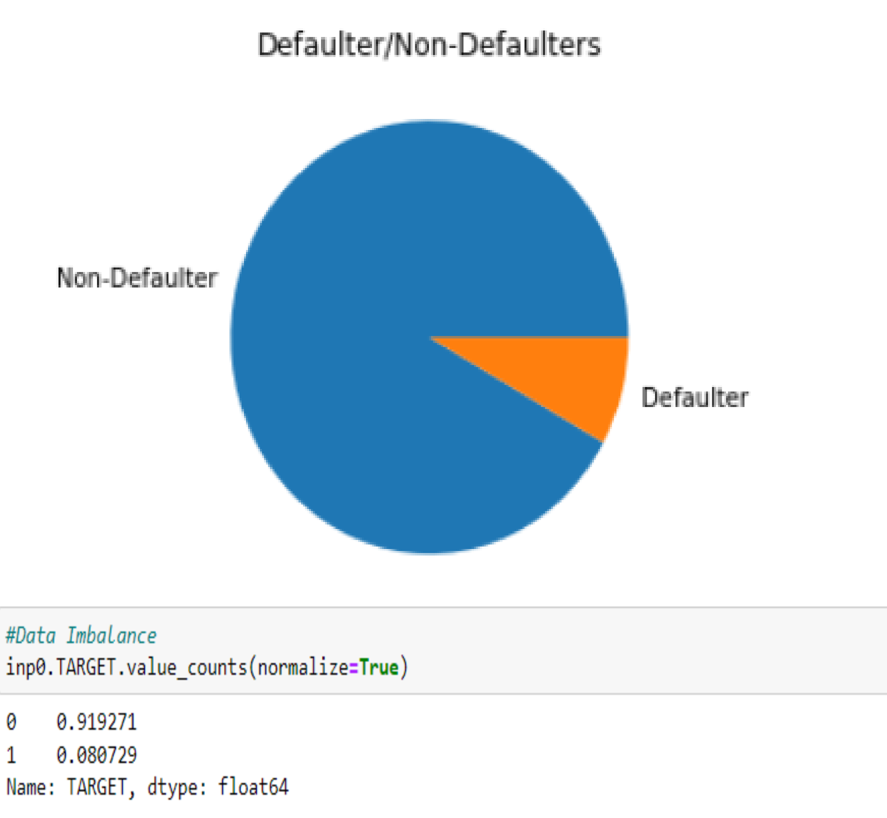
The two input files are loaded, cleaned and few columns are analyzed via different charts generated using different Python libraries. Then some observations are made based on the outcomes.

APPROACH

- Data Understanding
- Cleaning the data
- Identifying Missing value (with minimal treatment)
- Finding Outliers
- Analysis
  - Univariate
  - Segmented univariate
  - Bivariate/ Multivariate
  - Correlation
- Merging data for further analysis

Application Data

DATA IMBALANCE

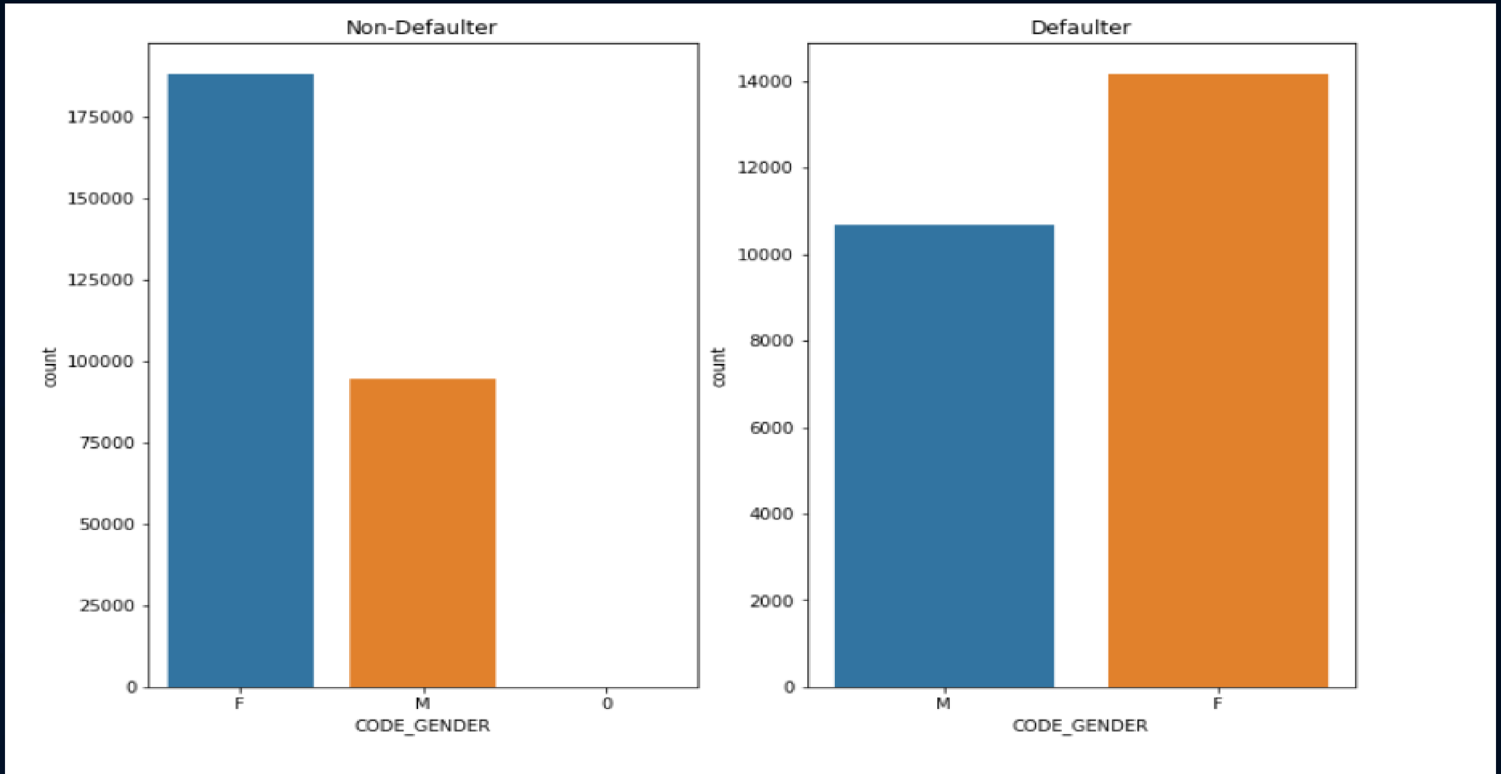


```
#Data Imbalance
inp0.TARGET.value_counts(normalize=True)
```

```
0    0.919271
1    0.080729
Name: TARGET, dtype: float64
```

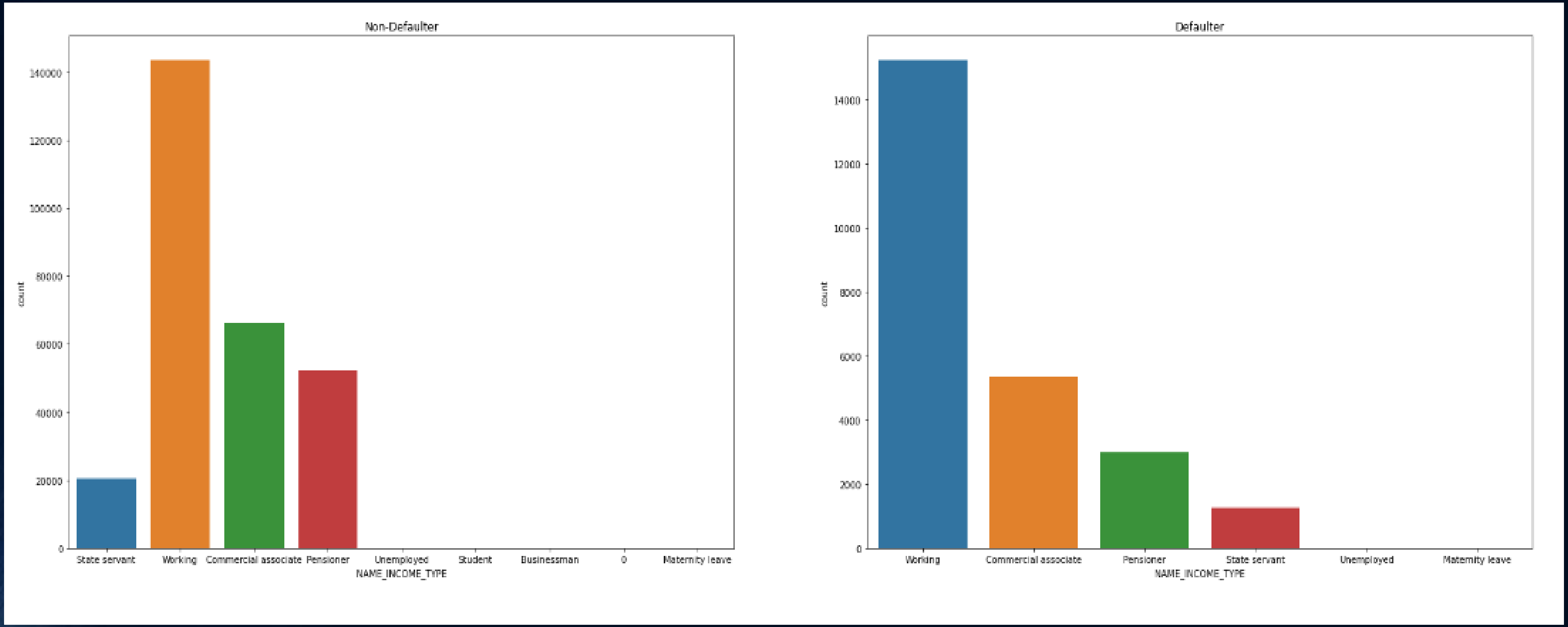
- Here we can observe that majority of people are Non-Defaulters.(91%).

UNIVARIATE ANALYSIS ON APPLICATION DATA



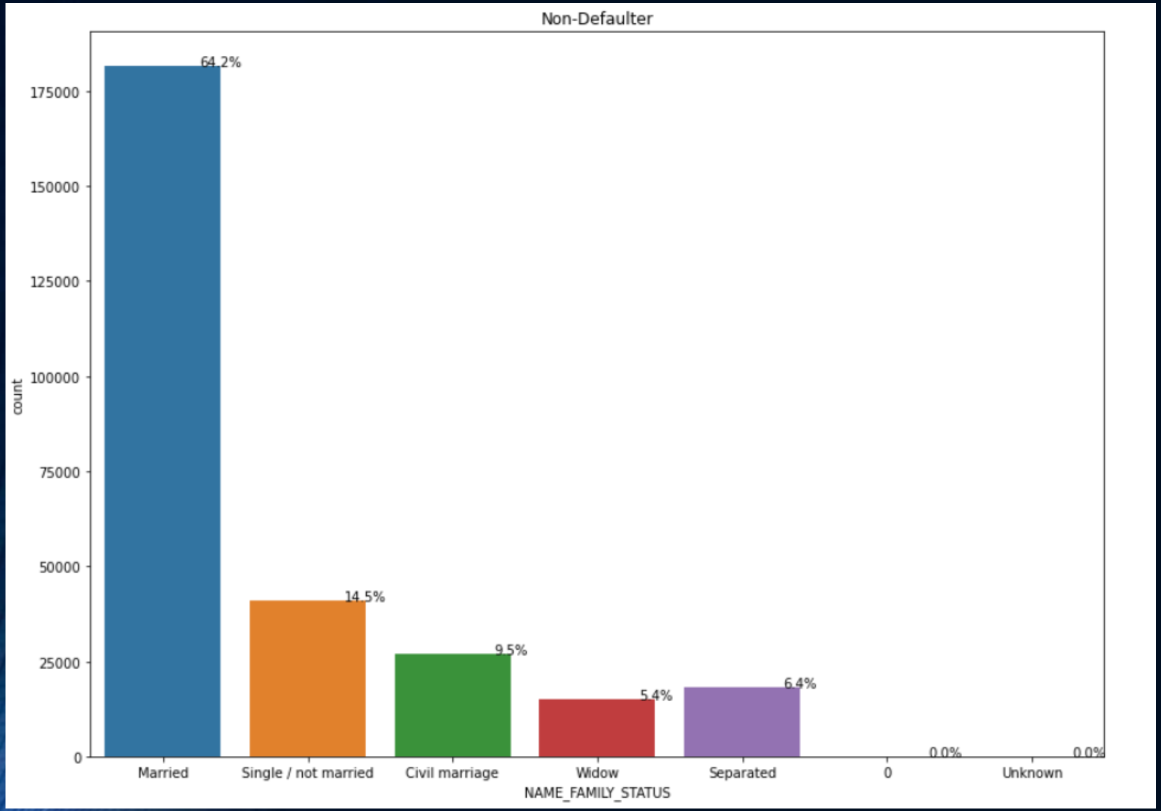
- Females are in majority in both Defaulter and non- defaulter case list
- There is an increase of Males in defaulter list

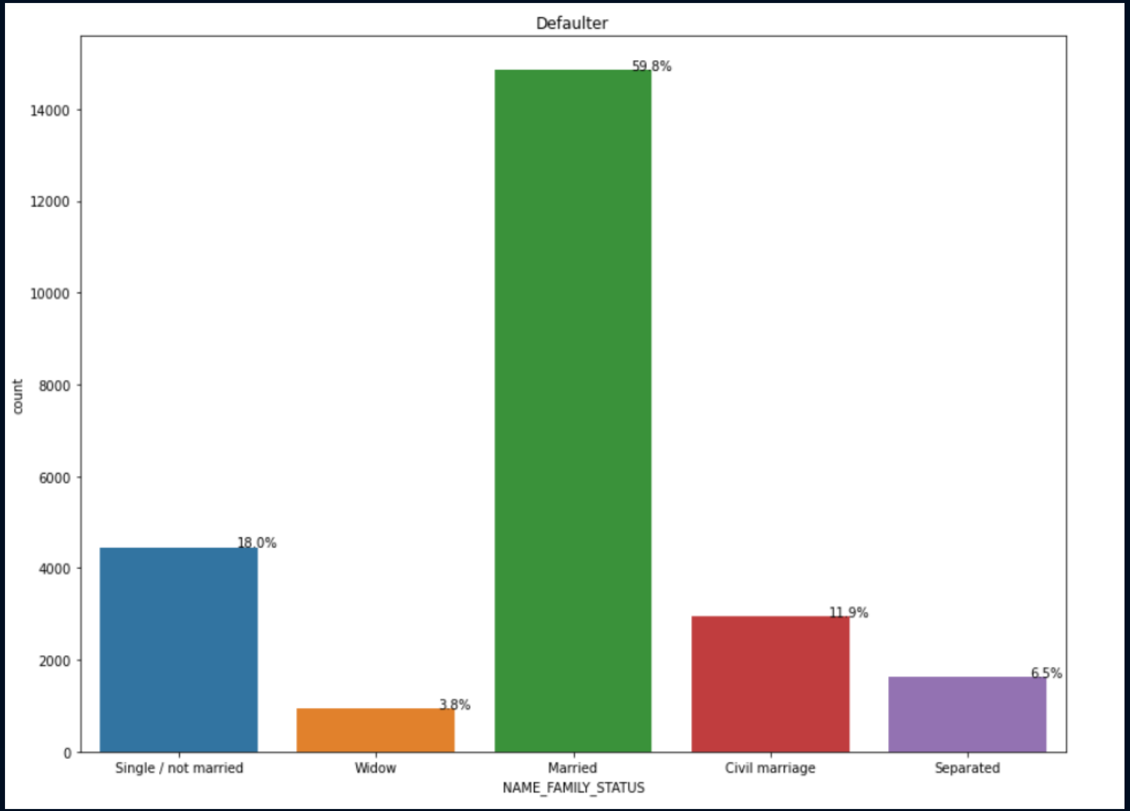
UNIVARIATE ANALYSIS ON APPLICATION DATA



- Here we are comparing the income of defaulters and non-defaulters and we observe that majority of the distribution is among Working in both defaulter and non defaulter, however chances of defaulting is more.
- Business man and student seems not to default
- There is a decrease in the percentage of being defaulters among Pensioners.

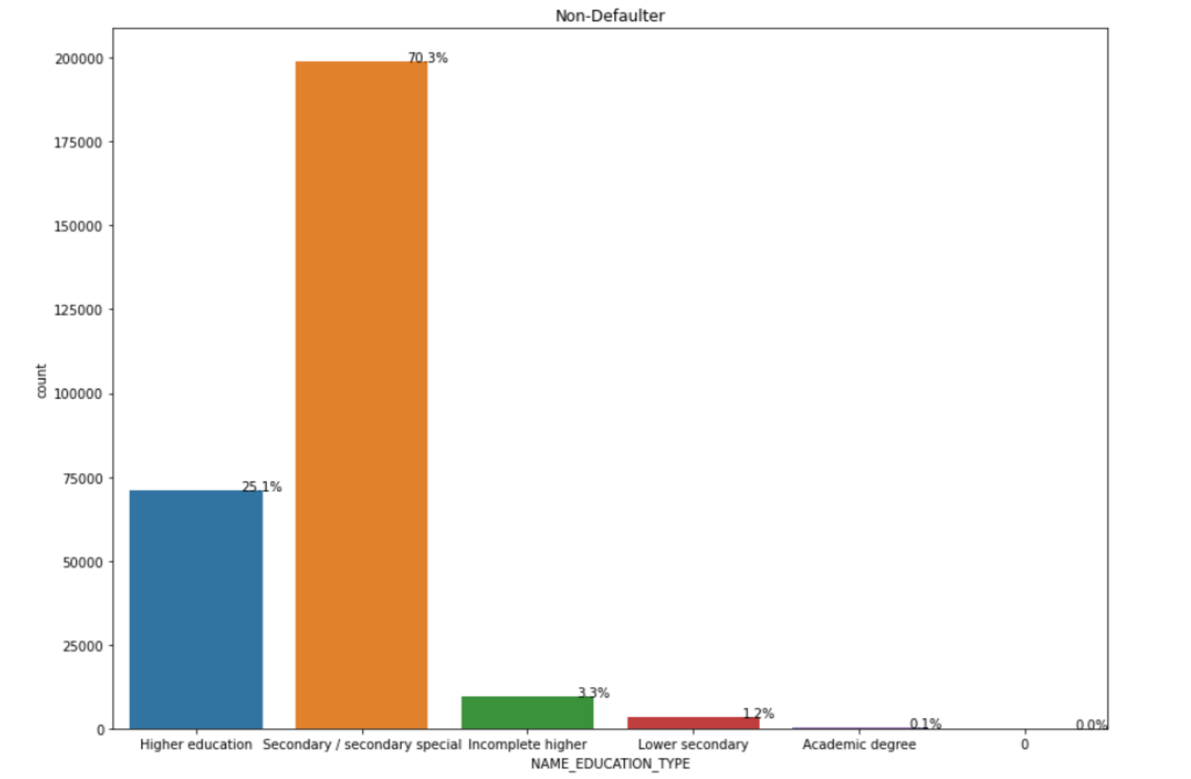
UNIVARIATE ANALYSIS ON APPLICATION DATA

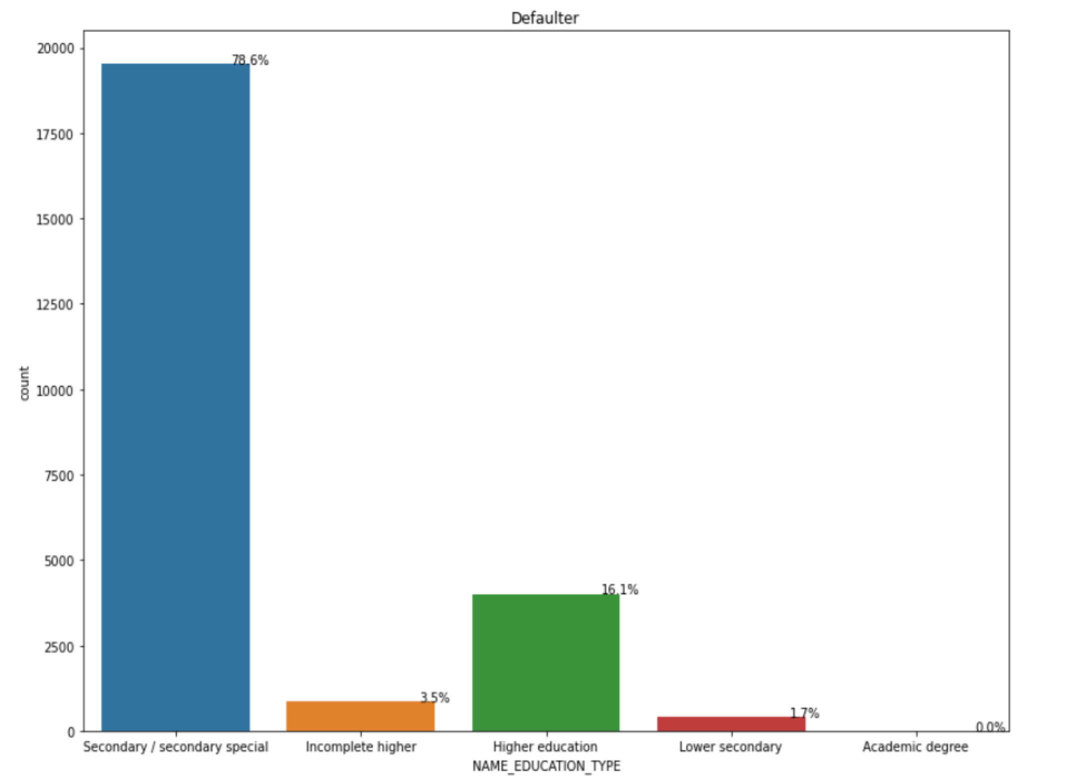




- We observe that people who are married tend to be non- defaulters more than being defaulters.
- There is an increase in the percentage of people who are single to be a defaulter than to be a non-defaulter so there is more risk associated with it

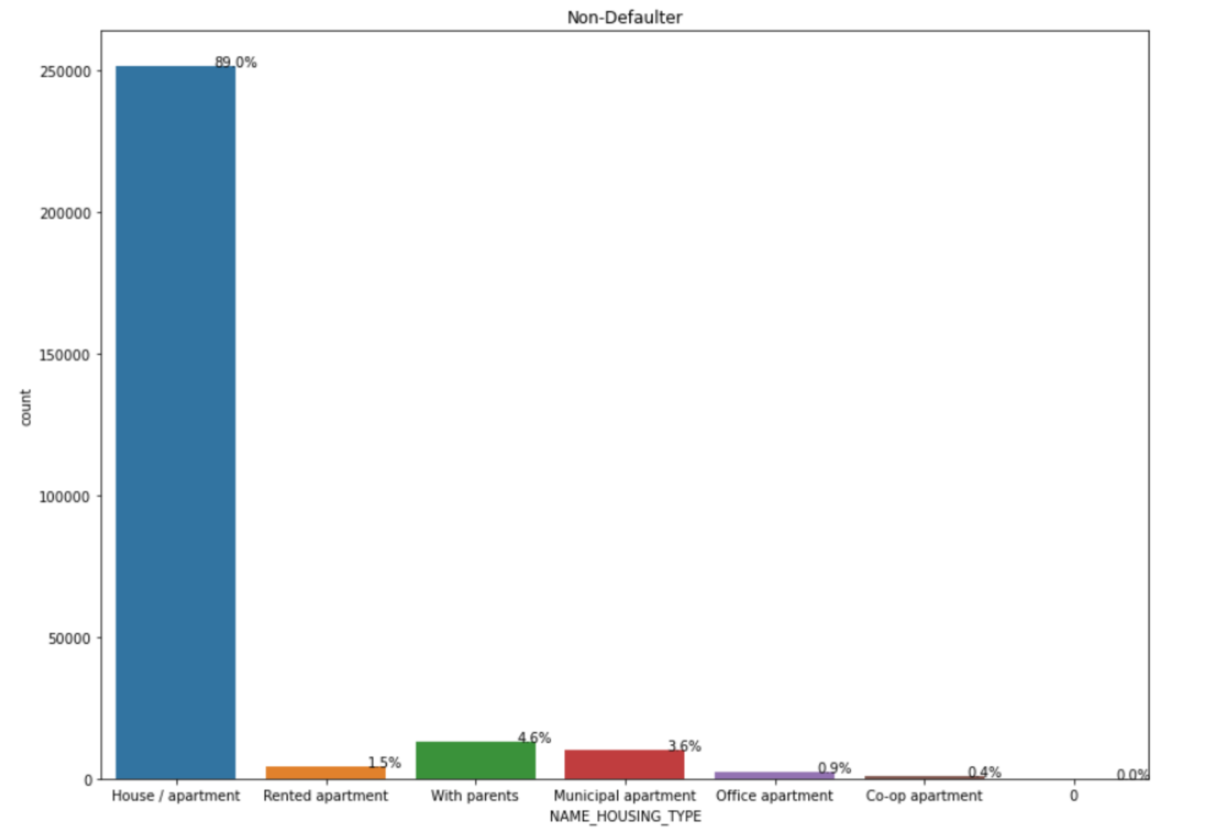
UNIVARIATE ANALYSIS ON APPLICATION DATA

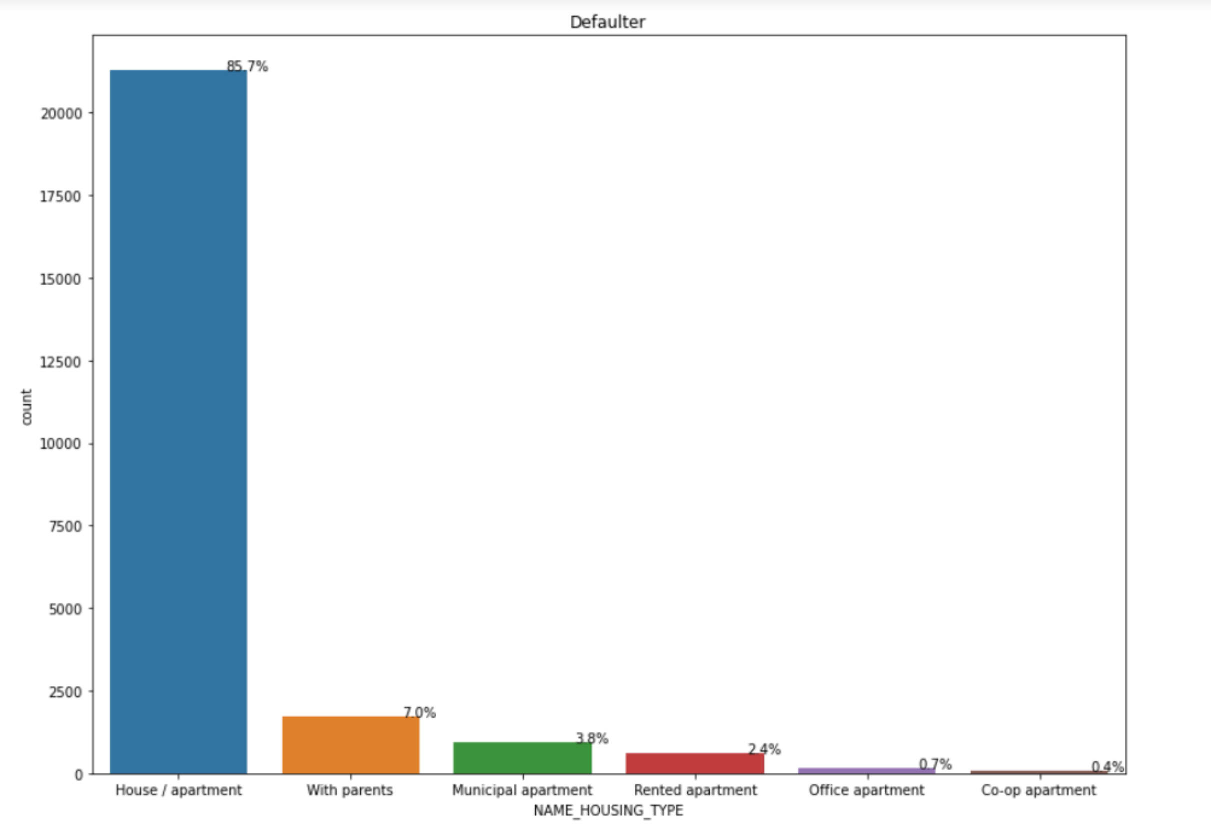




- We observe here that majority of the non-defaulters are having education qualification as Secondary/secondary special, so more risk is associated with it.
- There is more chances of people being non defaulter if they have higher education

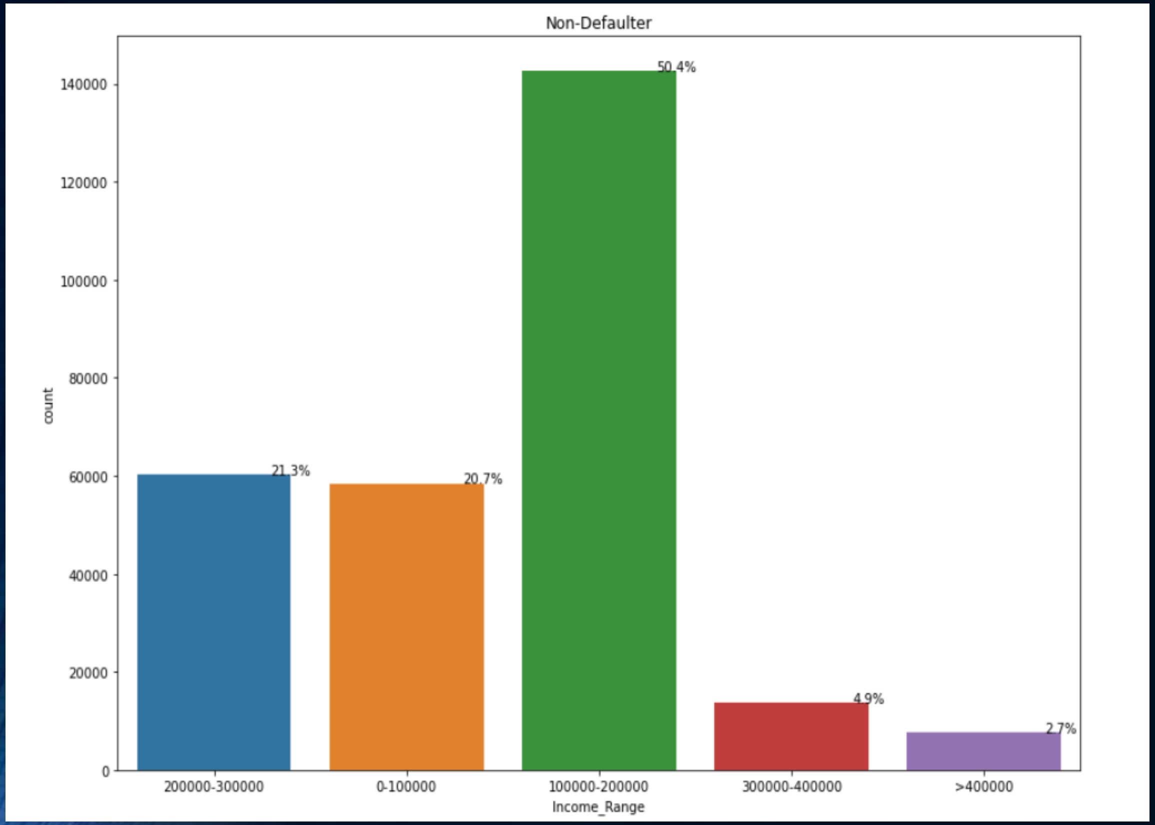
UNIVARIATE ANALYSIS ON APPLICATION DATA

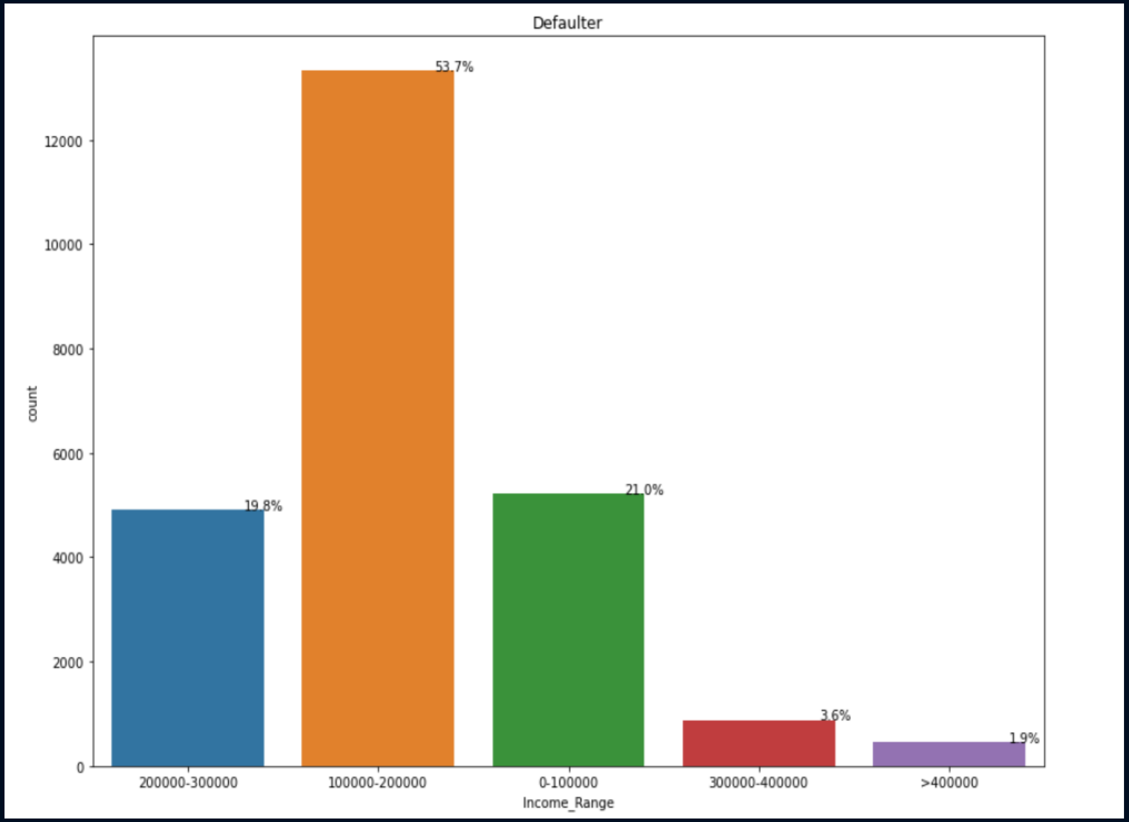




- We observe that majority of people in both defaulter and non-defaulters are with house/apartment
- There is an increase in the chances of people being defaulter if they live with parents, maybe due to their overhead expenses.

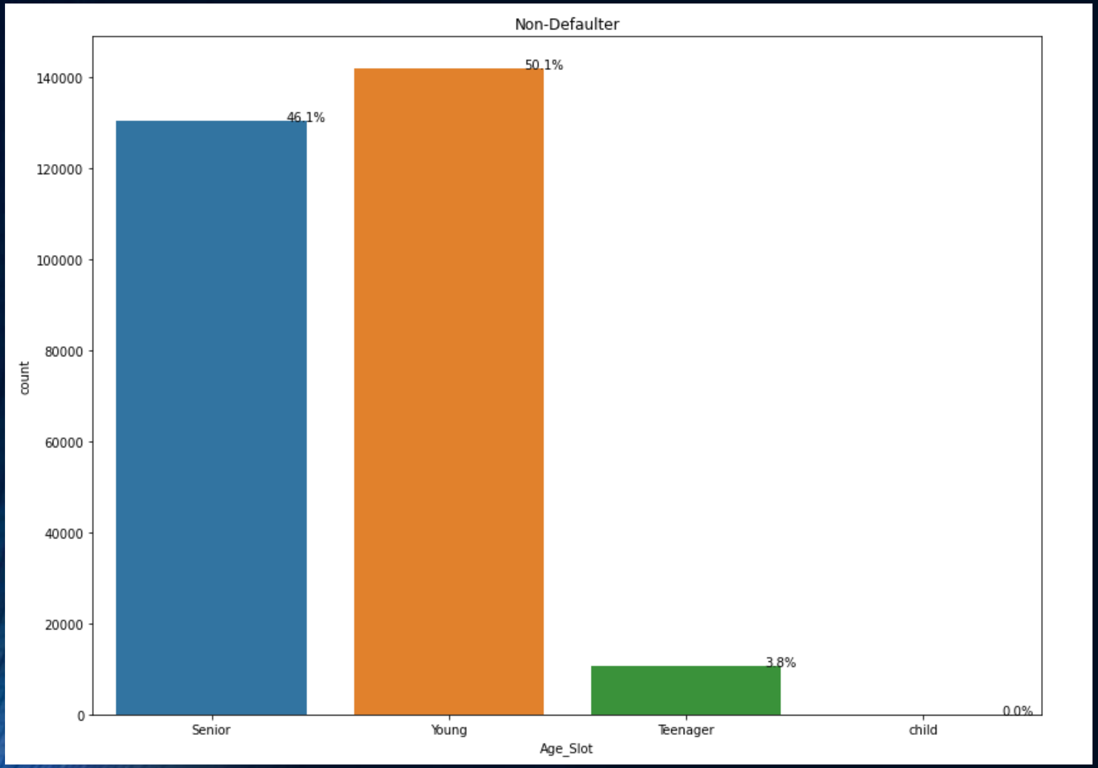
UNIVARIATE ANALYSIS ON APPLICATION DATA

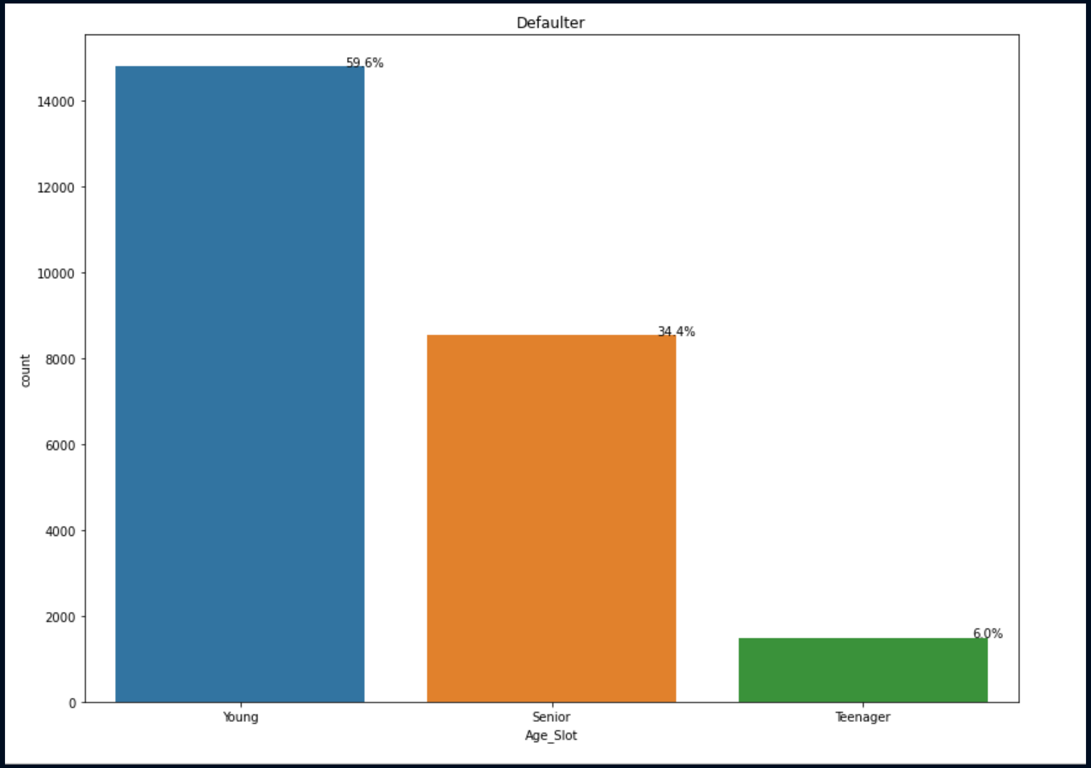




We observe that people who are earning between 100000 to 200000 are more likely to be at risk

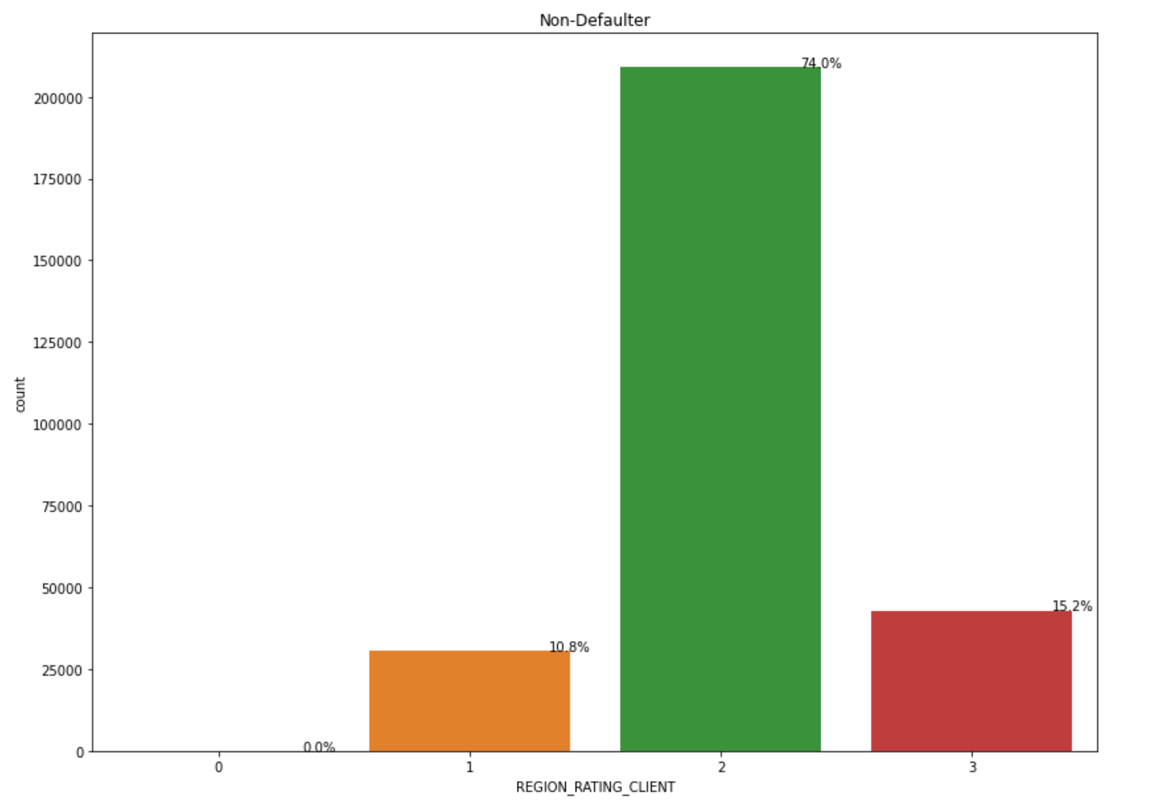
UNIVARIATE ANALYSIS ON APPLICATION DATA

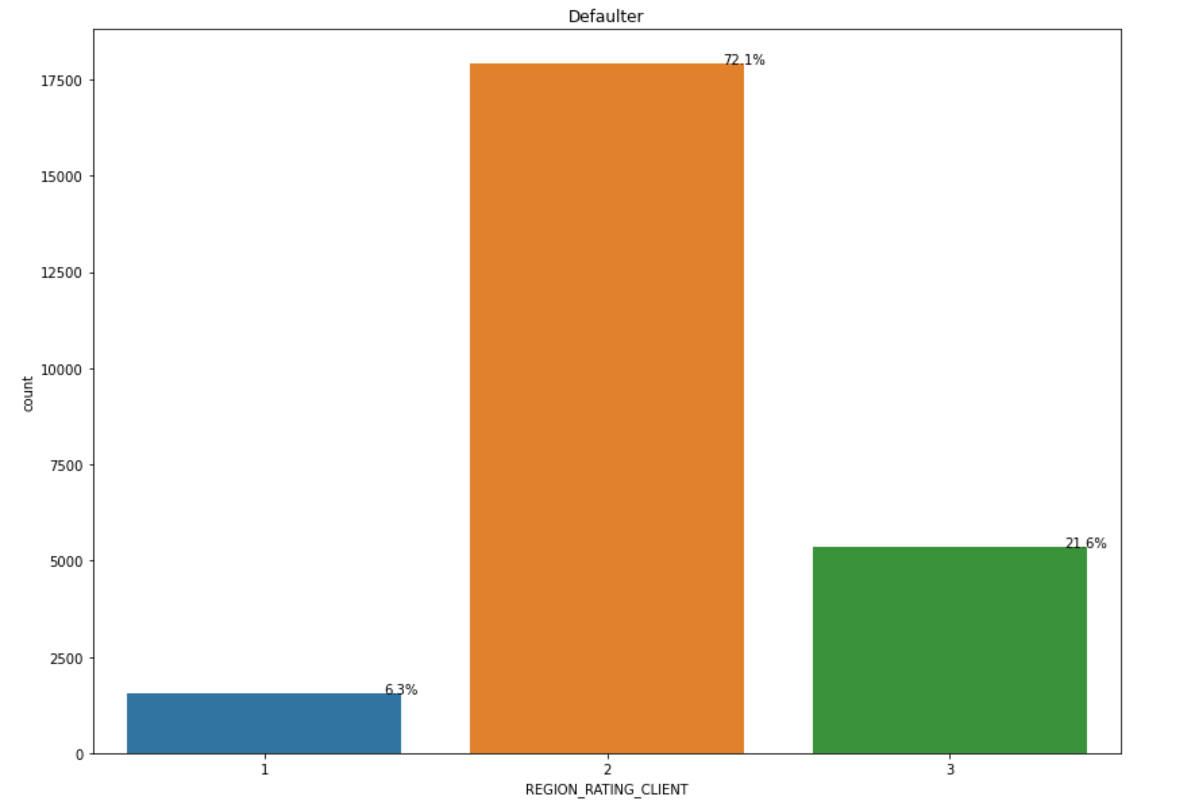




- We observe that young (25-45) and teenagers (18-25) are more prone to be defaulters and high at risk

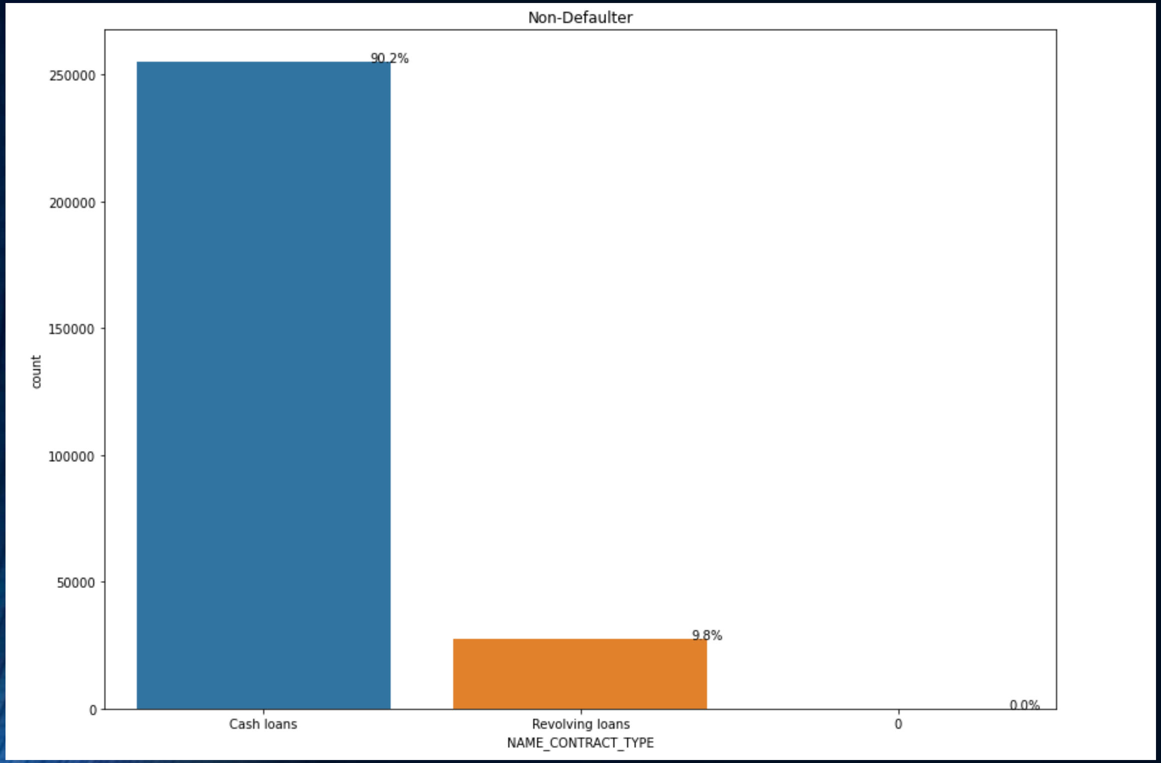
UNIVARIATE ANALYSIS ON APPLICATION DATA

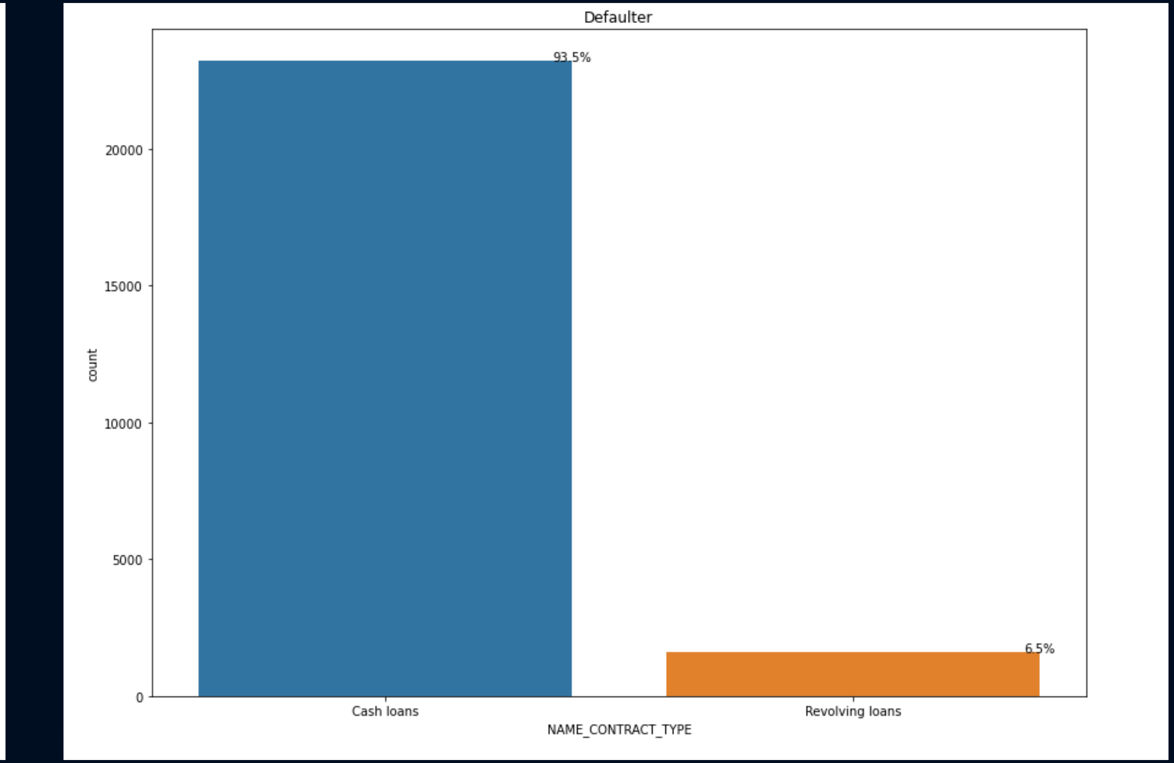




- Majority of the people belong to a region which has the rating of 2
- People belong to region with rating 3 have more chances of being a defaulter.

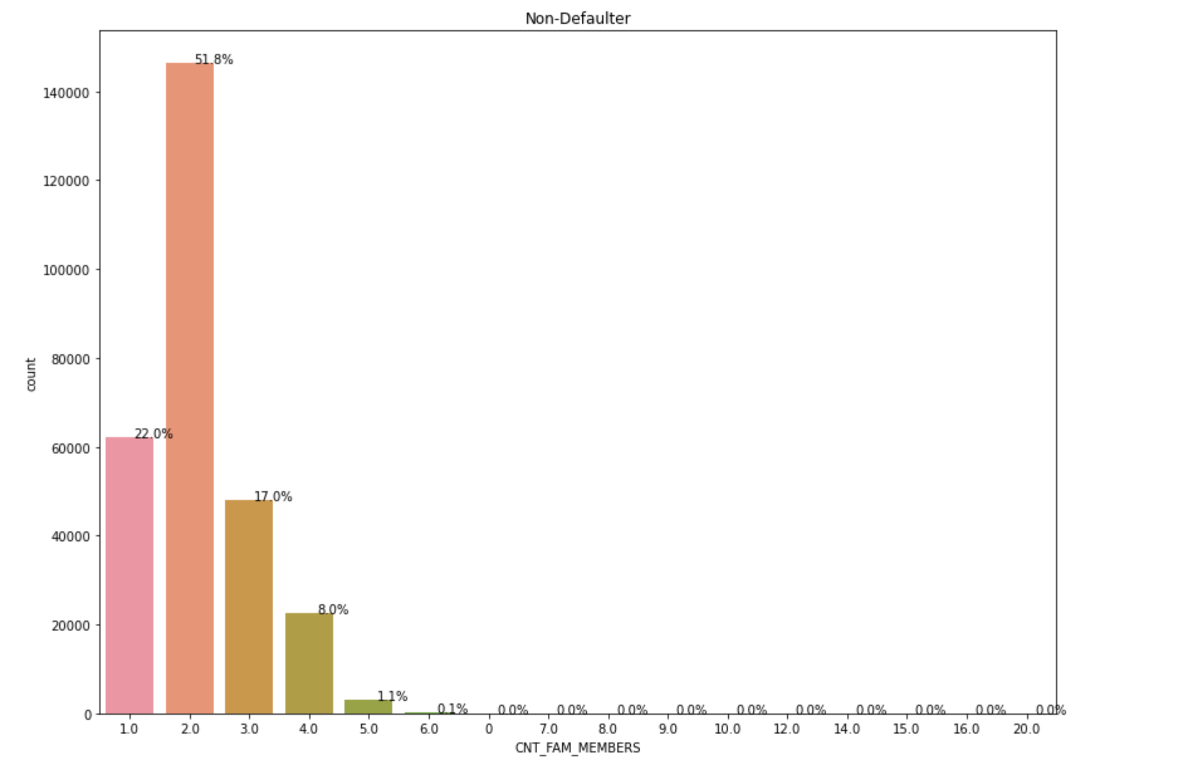
UNIVARIATE ANALYSIS ON APPLICATION DATA

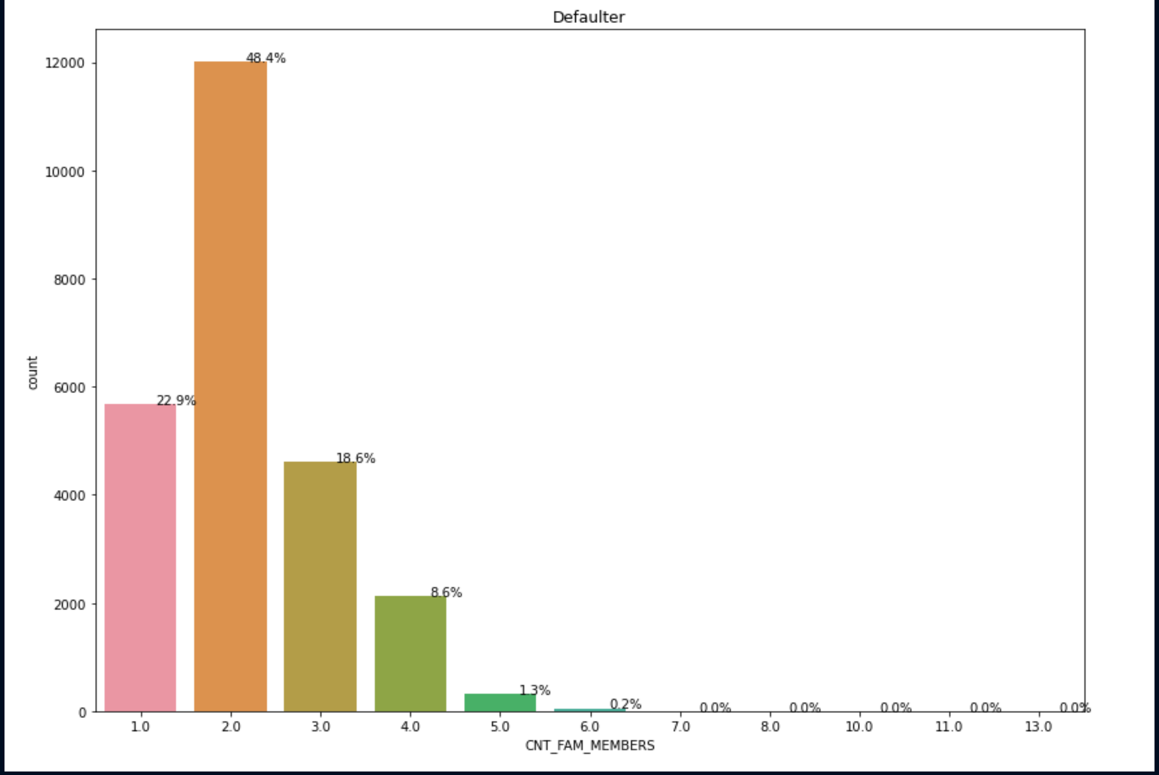




- We can observe that the majority of people opted for cash loans in both defaulter and non defaulter list and there is a slight increase in percentage of being a defaulter.
- People who opted for Resolving loan tend to have less chances of being a defaulter

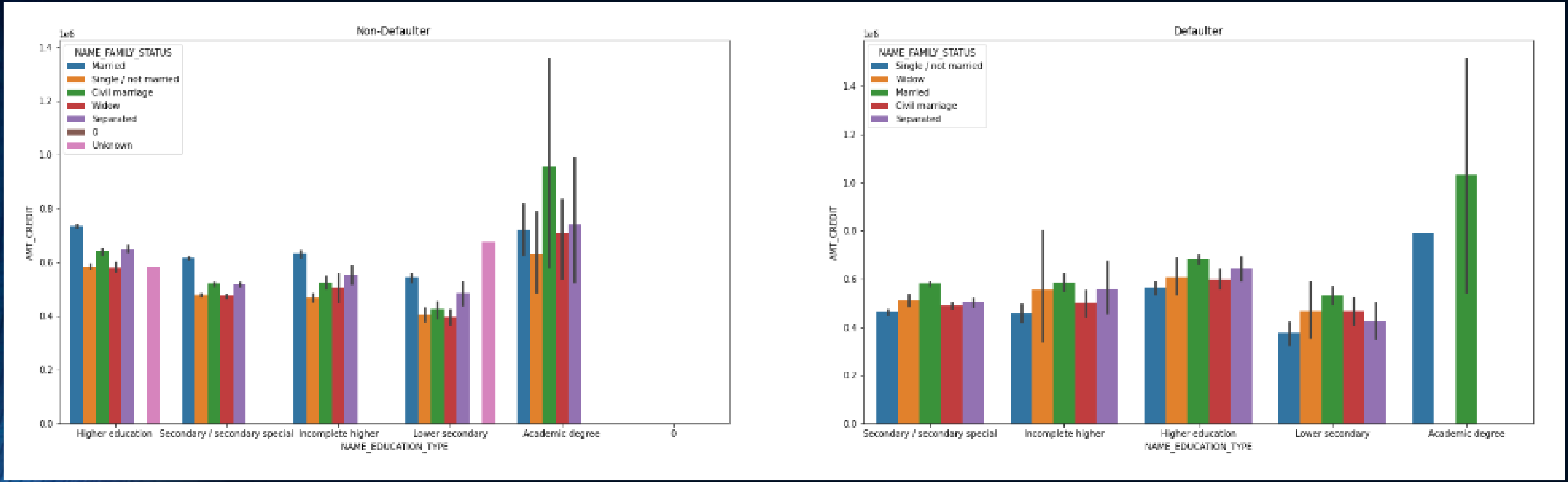
UNIVARIATE ANALYSIS ON APPLICATION DATA





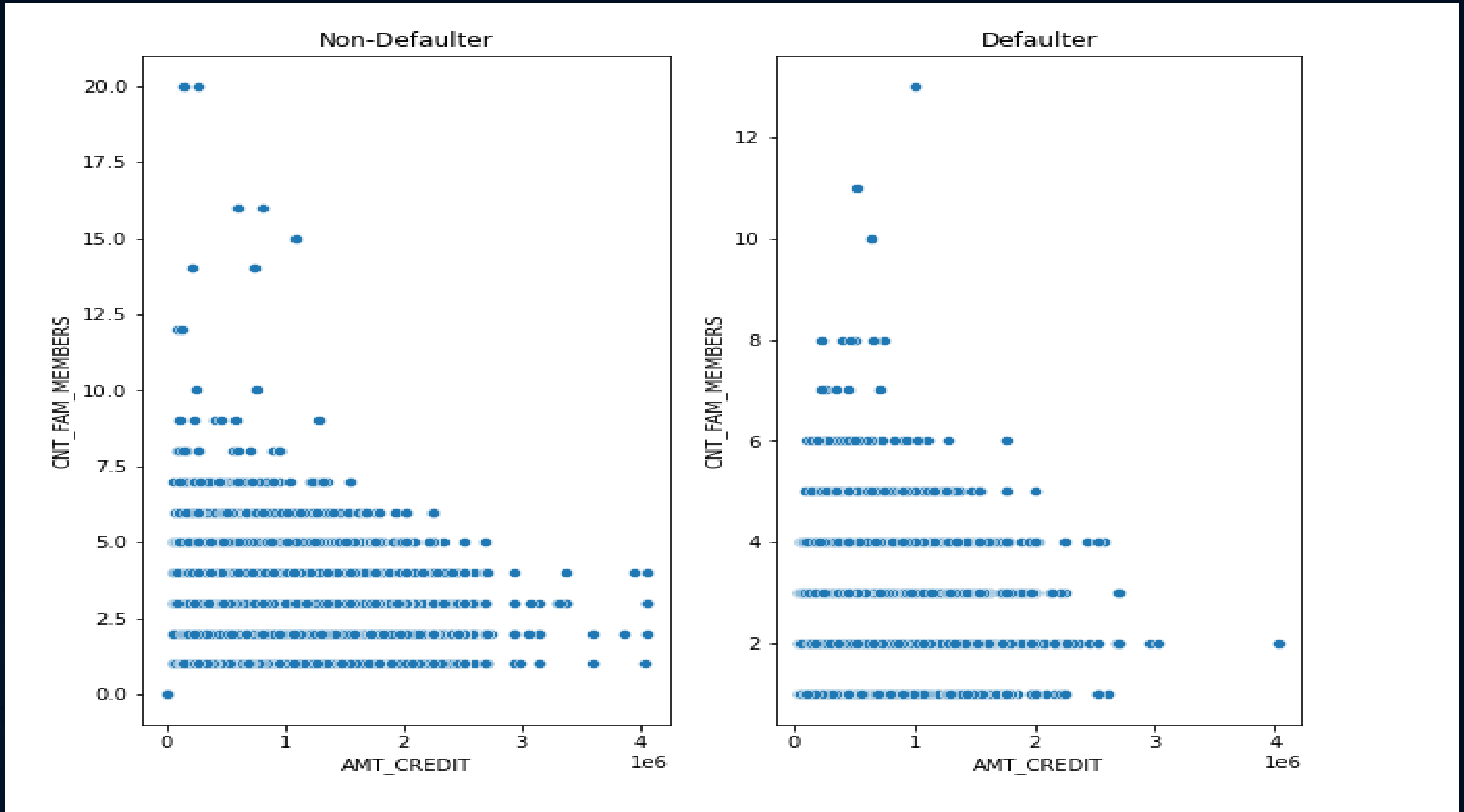
- Here we observe that families with 2 members are in mjority in both defaulters and non-defaulters, however there is a slightly more chance of being a non-defaulter.
- However, we can see a pattern, although its not significant but very minimal pattern of family with more number of members being a defaulter.

BIVARIATE / MULTIVARIATE ANALYSIS ON APPLICATION DATA



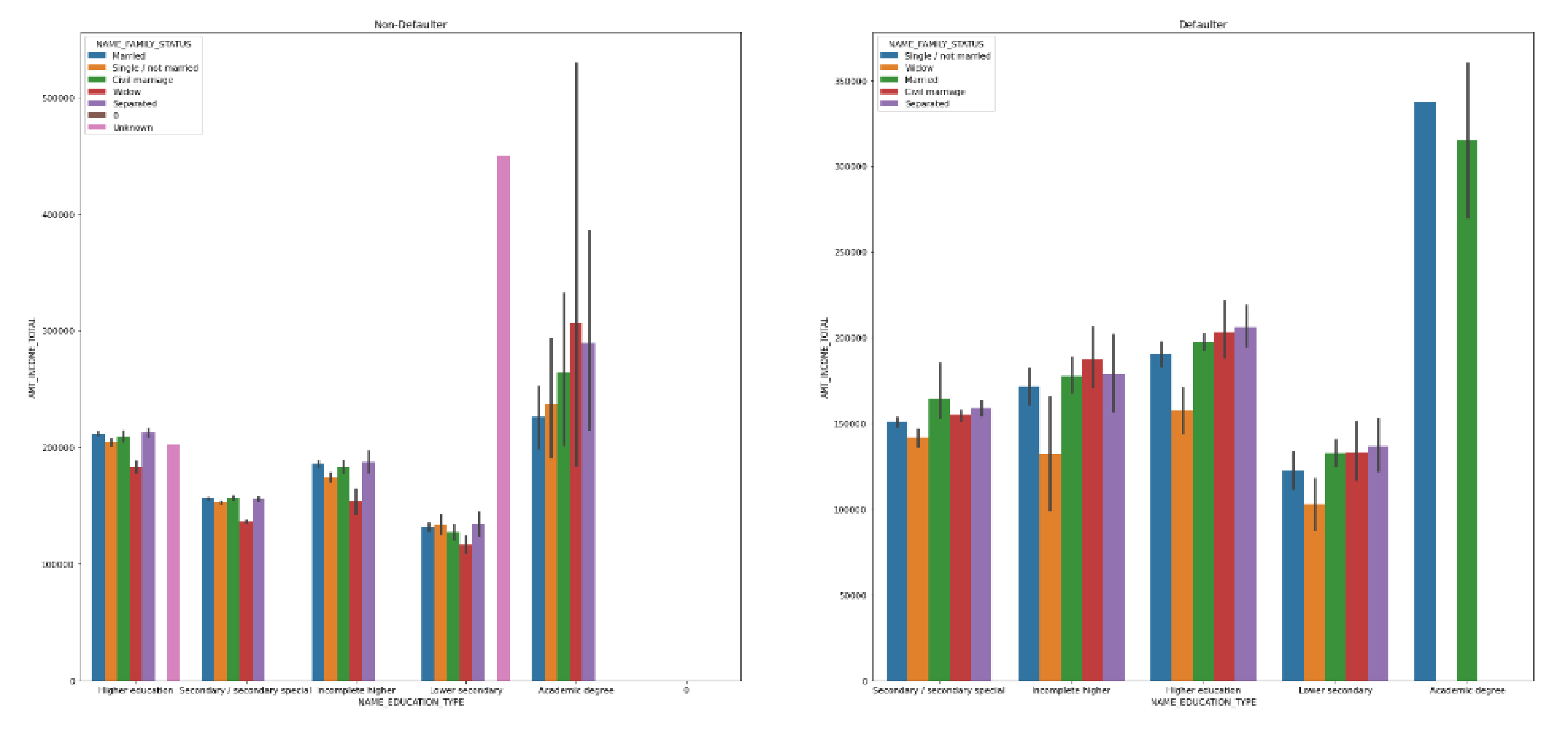
- We observe that people belong to Family of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having more credits than others but they are non-defaulters.

BIVARIATE / MULTIVARIATE ANALYSIS ON APPLICATION DATA



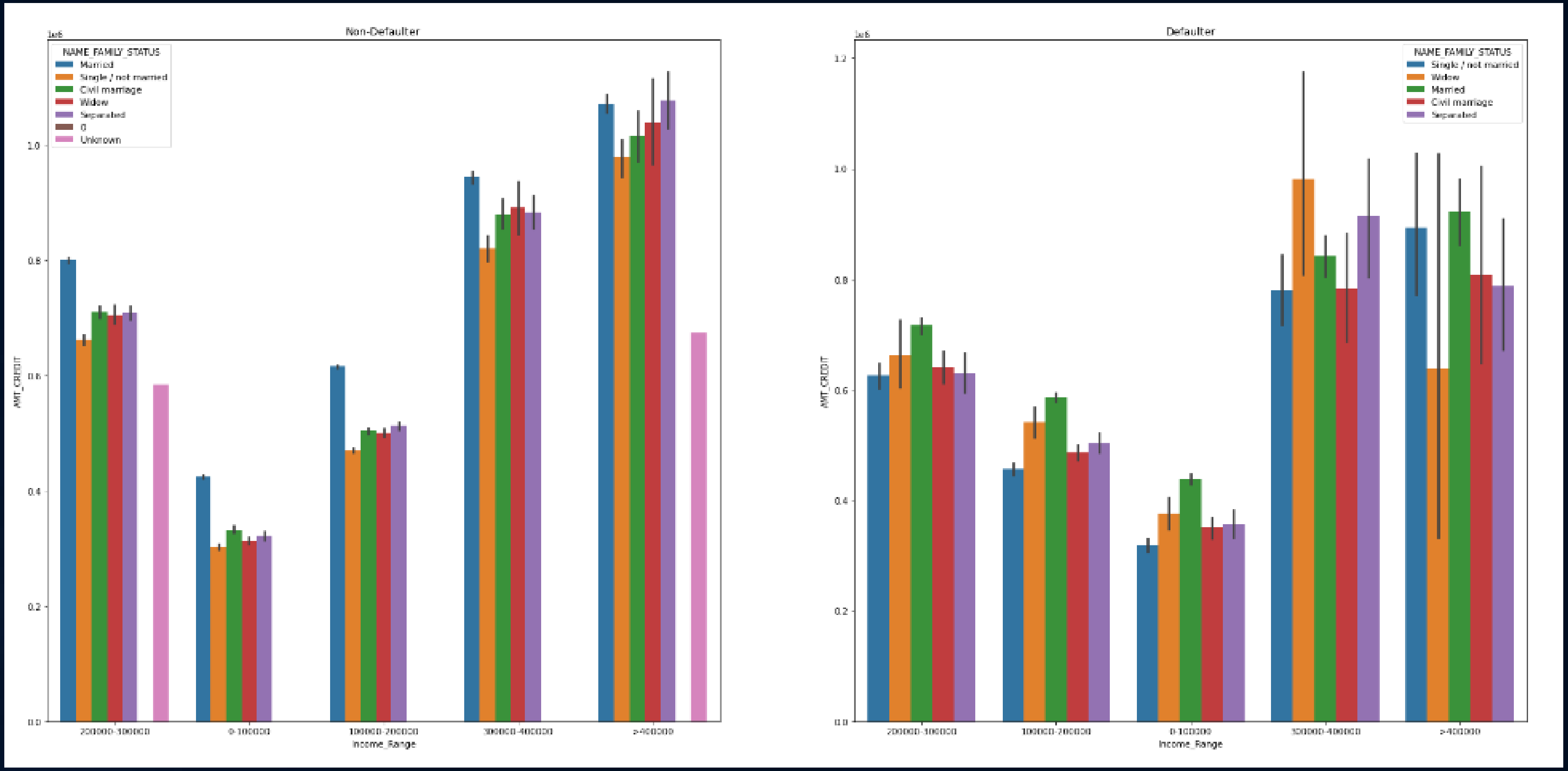
- As we can observe here, if we see the down left corner of the graph in both defaulter and non-defaulter, it has more density so as when income is less people tend to default even though family members are less.
- Also we see here that families have higher income tend to default very less even though they have more no. of family members.

BIVARIATE / MULTIVARIATE ANALYSIS ON APPLICATION DATA



- As we can observe from above plot, Education type 'Higher education' the income amount is mostly equal with family status.
- People earn less who are having Lower secondary with civil marriage family status

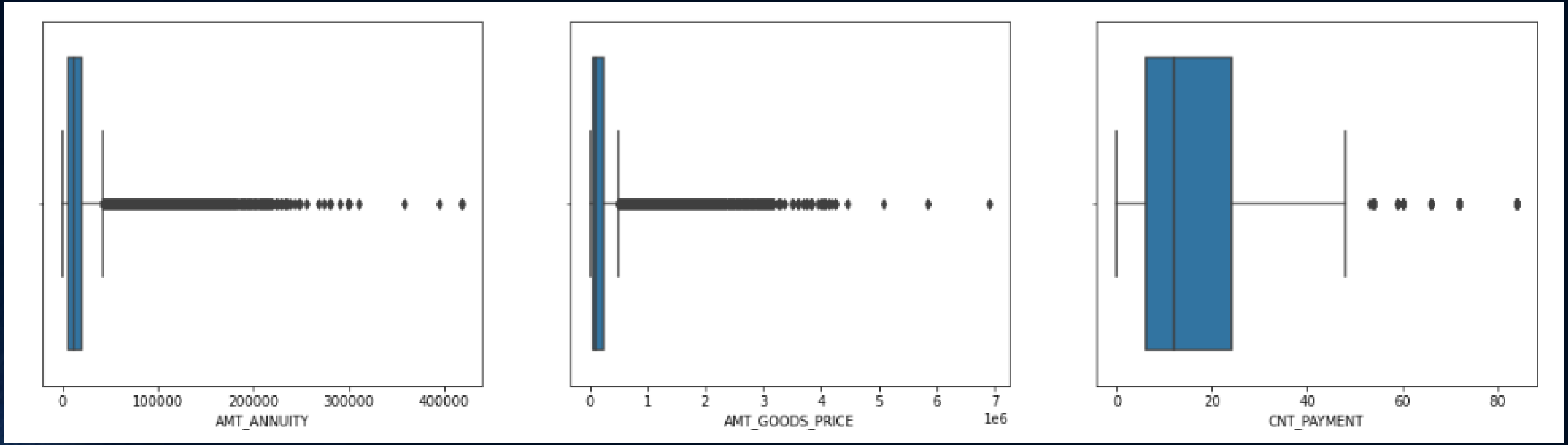
BIVARIATE / MULTIVARIATE ANALYSIS ON APPLICATION DATA



- As we can observe, the graph of both Defaulters and Non-Defaulters are almost similar in nature We observe that people who are 'single’ or ’separated' or 'married’ and earn more than 300000 are having higher number of credits than others.

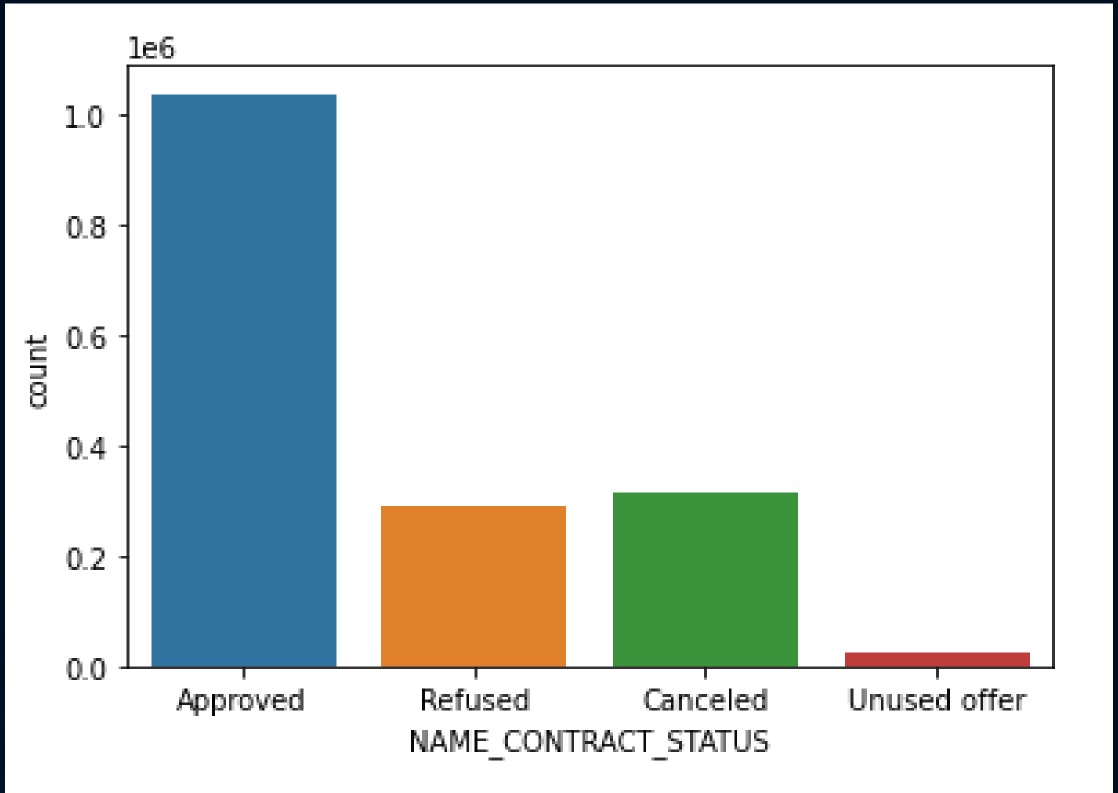
PREVIOUS APPLICATION DATA

Identifying Outliers



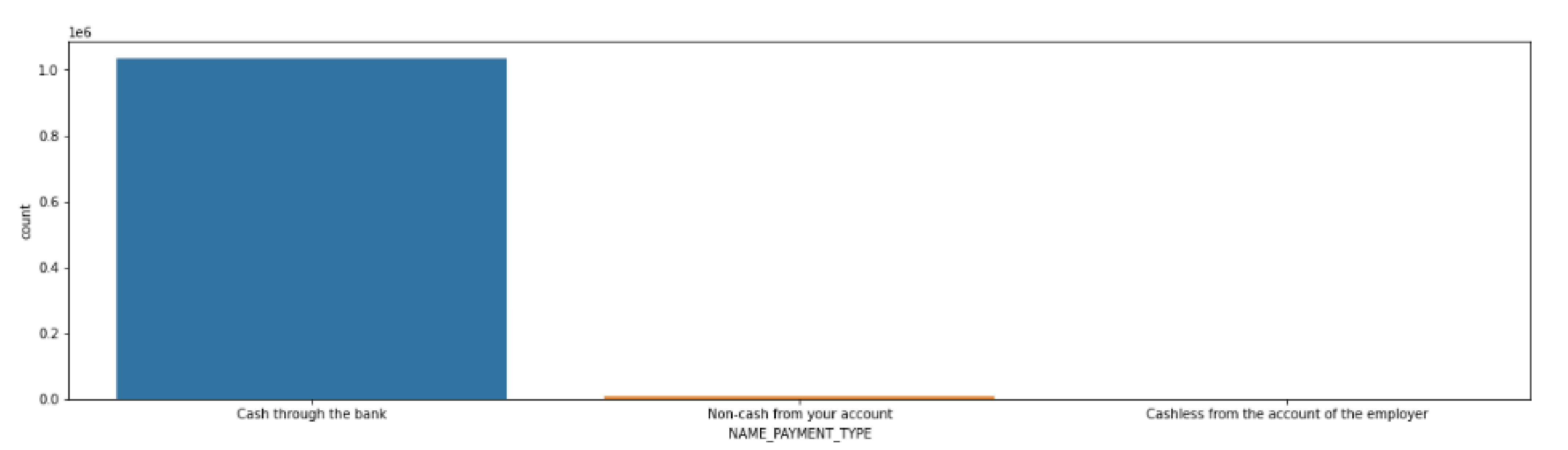
- We can observe here that AMT_ANNUITY, AMT_GOODS_PRICE, CNT_PAYMENT,AMT_DOWN_PAYMENT have quite a few outliers.

UNIVARIATE ANALYSIS ON PREVIOUS DATA



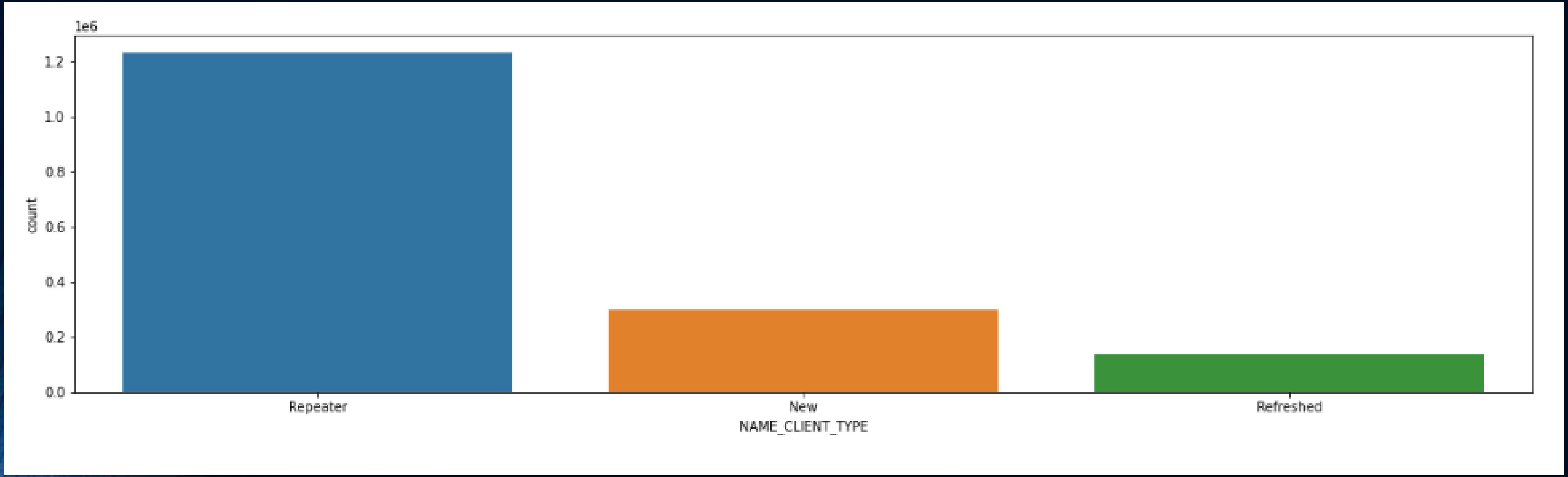
- Here we can observe that majority of the applications were approved.

UNIVARIATE ANALYSIS ON PREVIOUS DATA



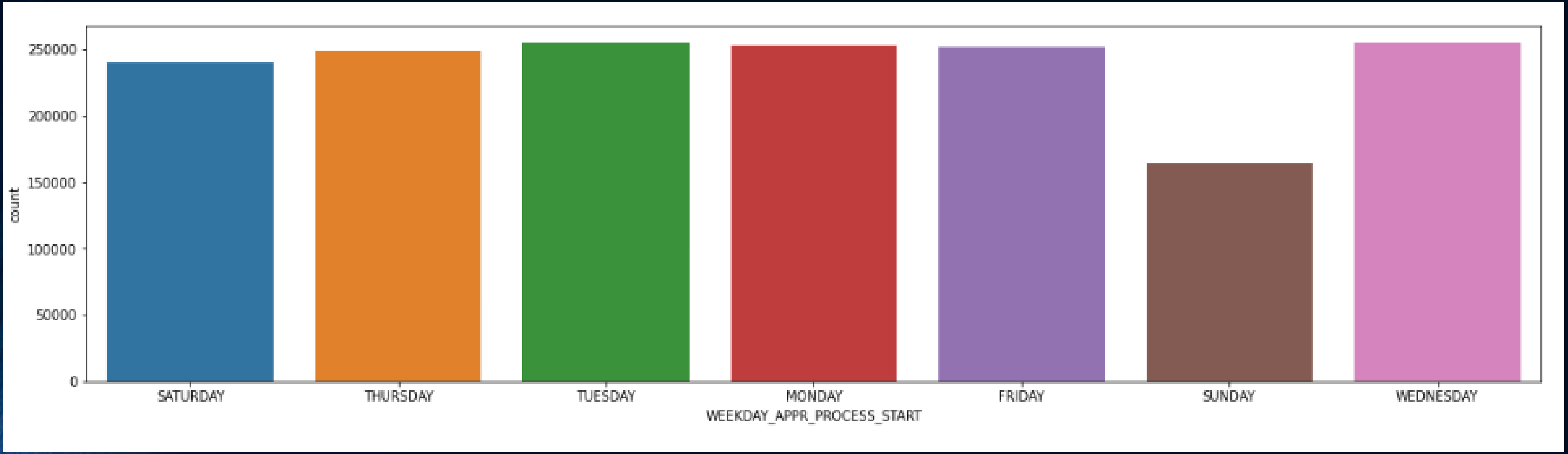
- As we can observe here, majority of people choose cash through bank

UNIVARIATE ANALYSIS ON PREVIOUS DATA



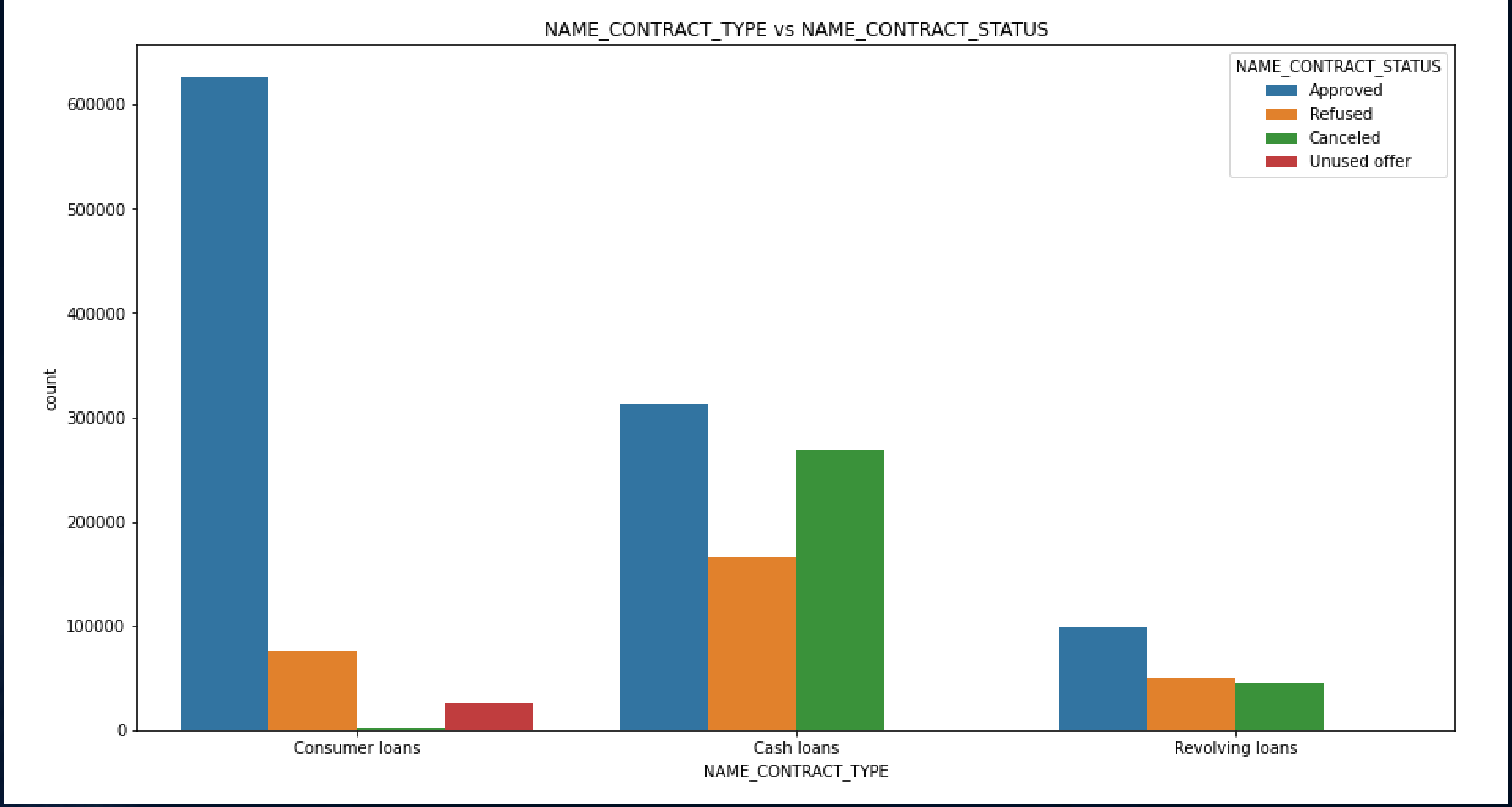
- As we can observe here, majority of people are Repeaters

UNIVARIATE ANALYSIS ON PREVIOUS DATA



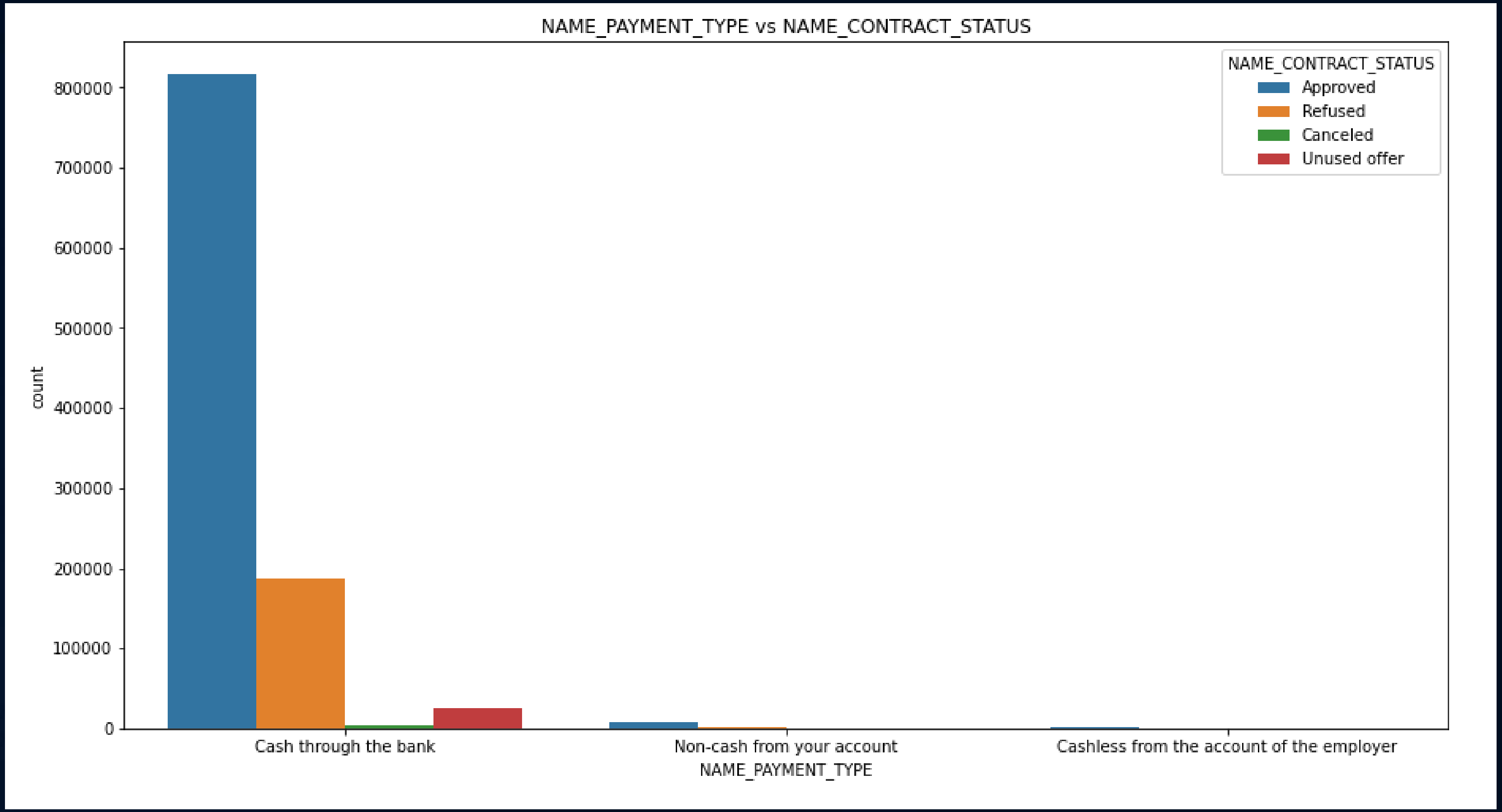
- As we can observe here, less people come in weekends

BIVARIATE ANALYSIS ON PREVIOUS DATA



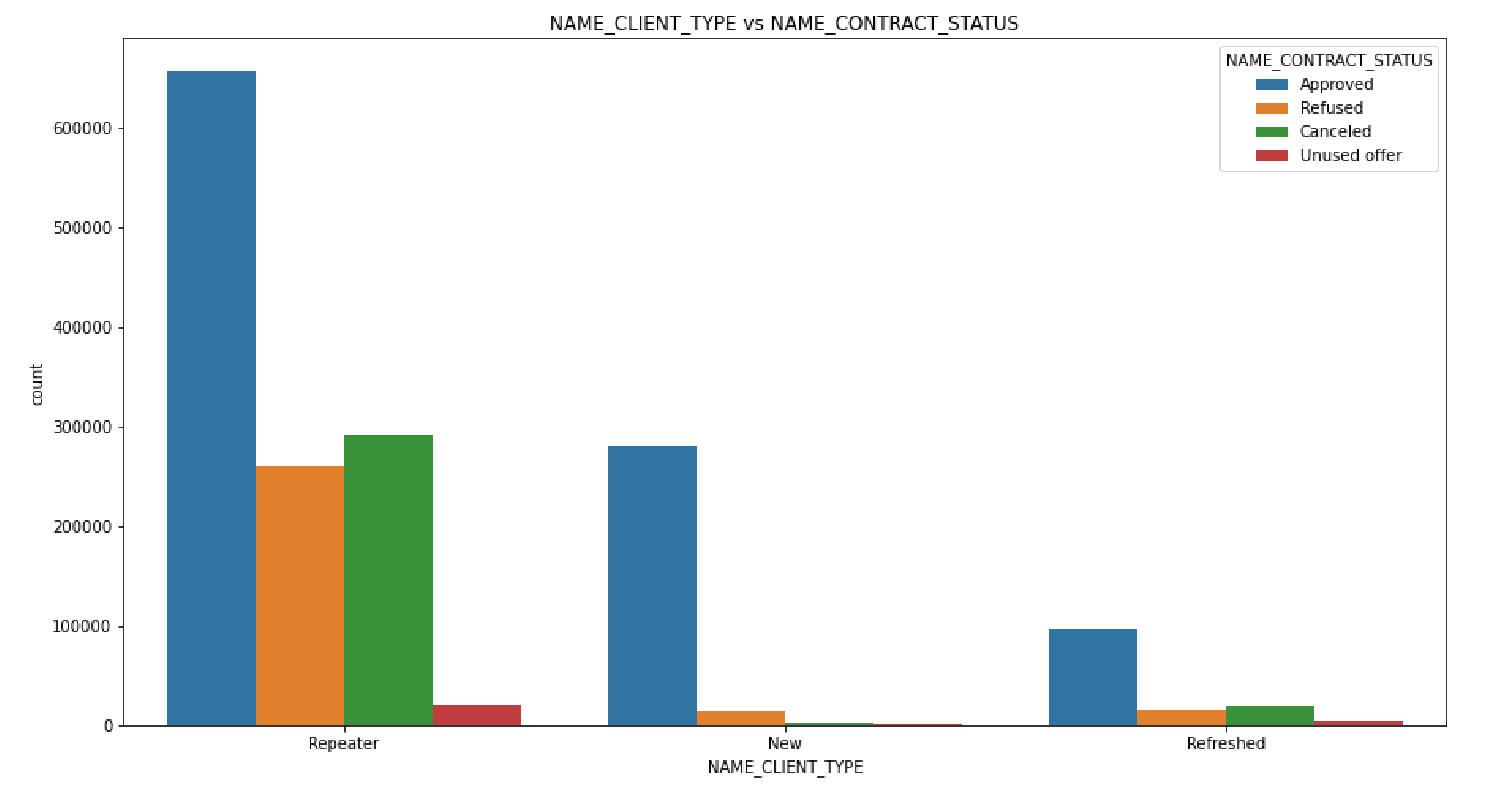
- Here we can observe that consumer loans are mostly approved
- Most of the loans are Consumer loans and cash loans
- Cash loans tend to get refused more often than others

BIVARIATE ANALYSIS ON PREVIOUS DATA



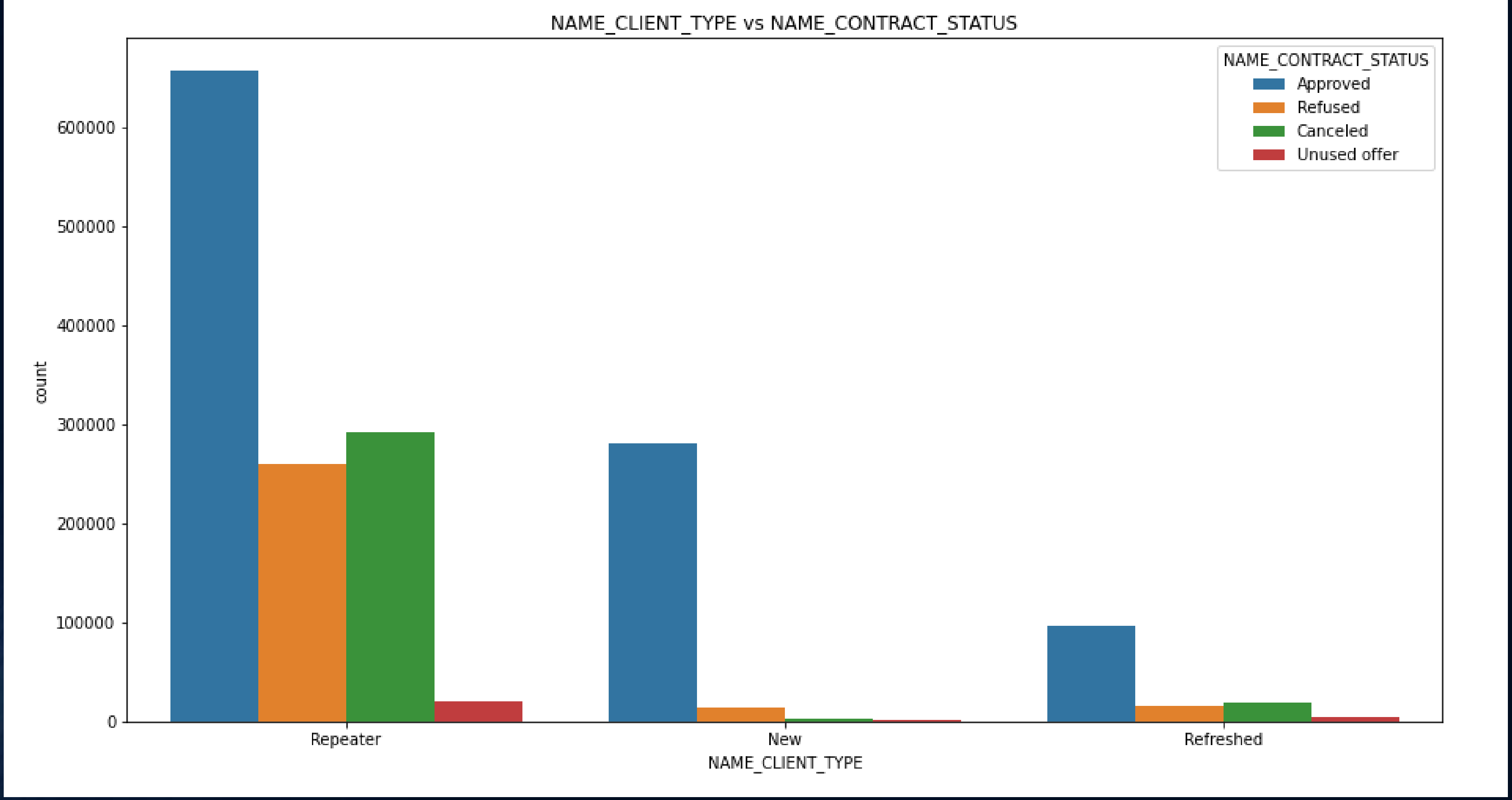
- As we can observe here people choose Cash through bank option more often than others
- Cash through bank option has the highest approval rate
- Non-cash from your account and cashless are quite unpopular options

BIVARIATE ANALYSIS ON PREVIOUS DATA



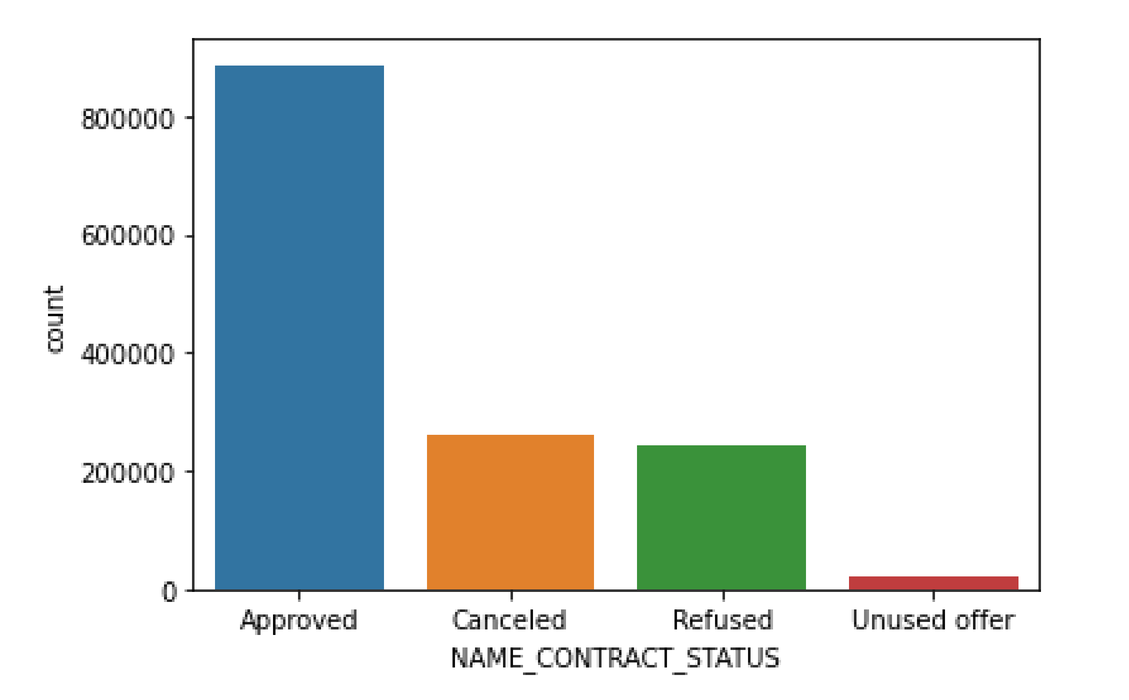
- Here we can observe, majority of customers are Repeaters
- Repeaters and New Customers have high approval rate

BIVARIATE ANALYSIS ON PREVIOUS DATA



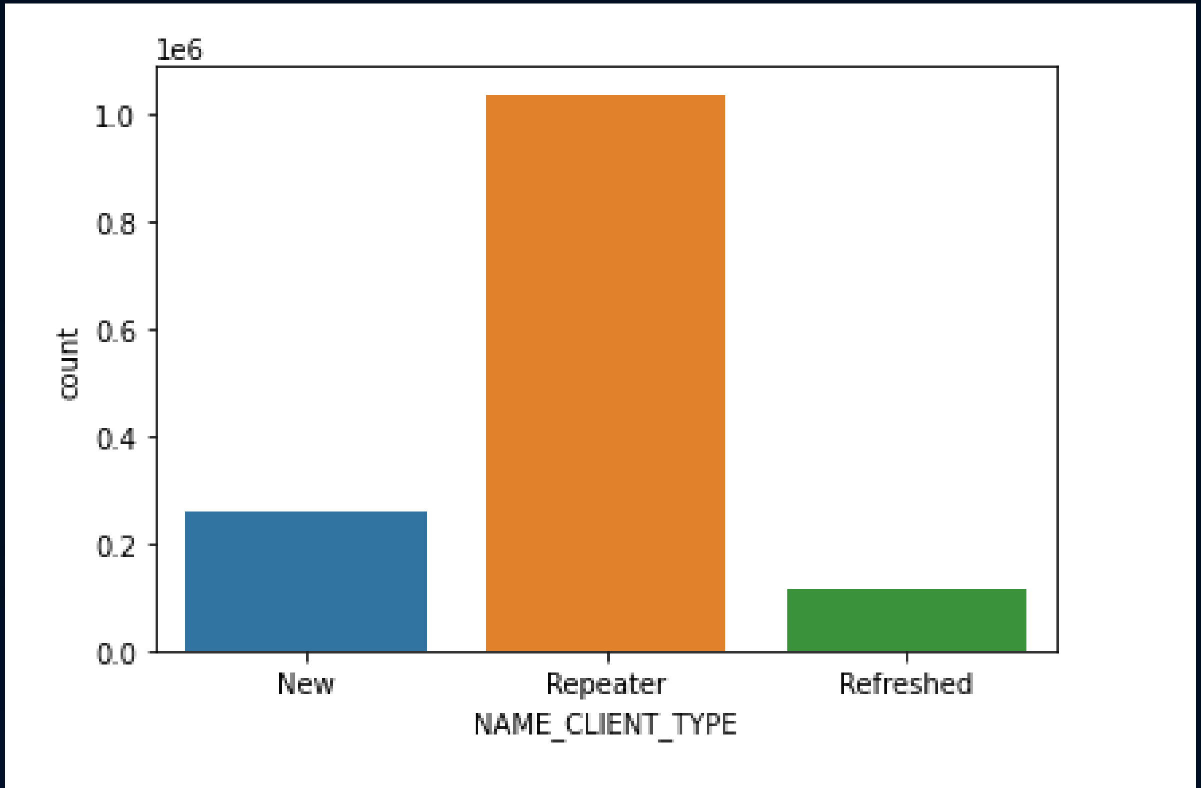
- Here we can observe, majority of customers are Repeaters
- Repeaters and New Customers have high approval rate

UNIVARIATE ANALYSIS ON MERGED DATA



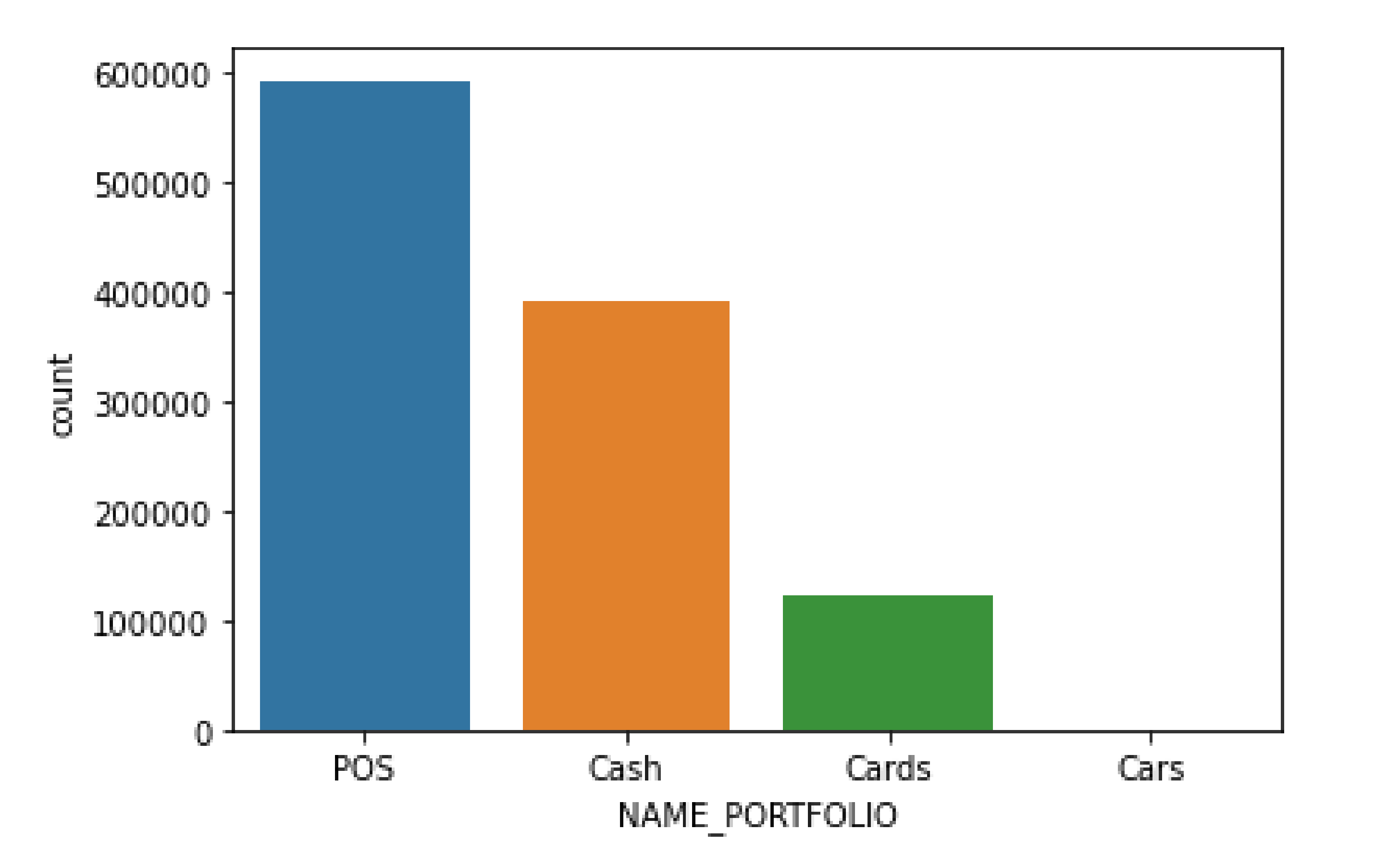
- We can observe here that Approved loan status is in majority compare to rejected or canceled.

UNIVARIATE ANALYSIS ON MERGED DATA



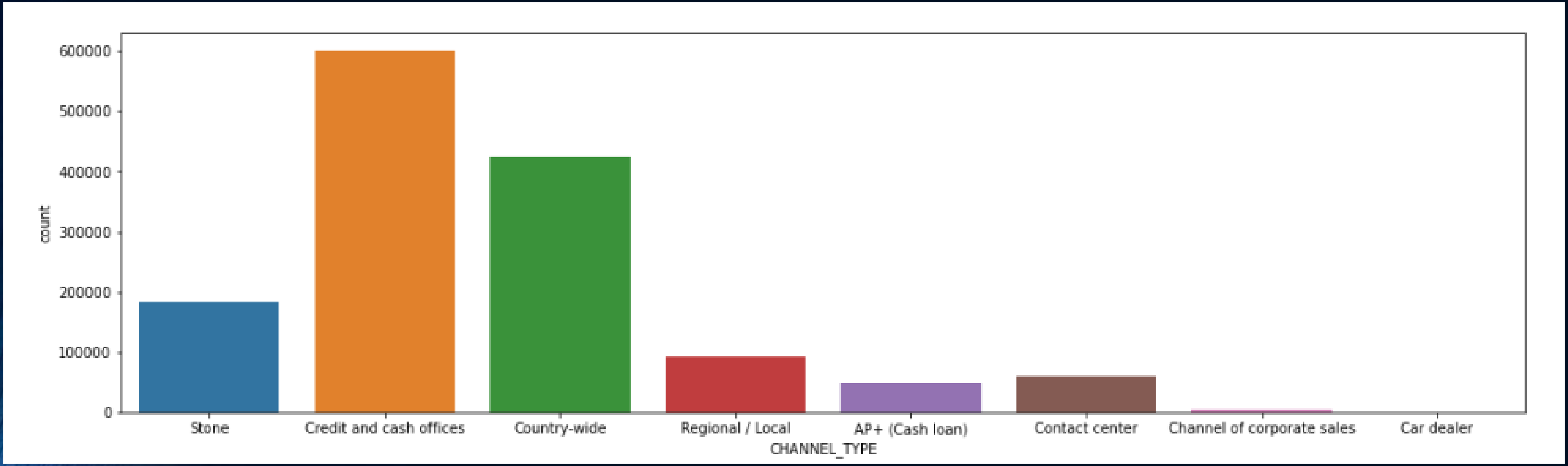
- Here we can observe that customers who are Repeaters are in majority compare to New and Refreshed customers

UNIVARIATE ANALYSIS ON MERGED DATA



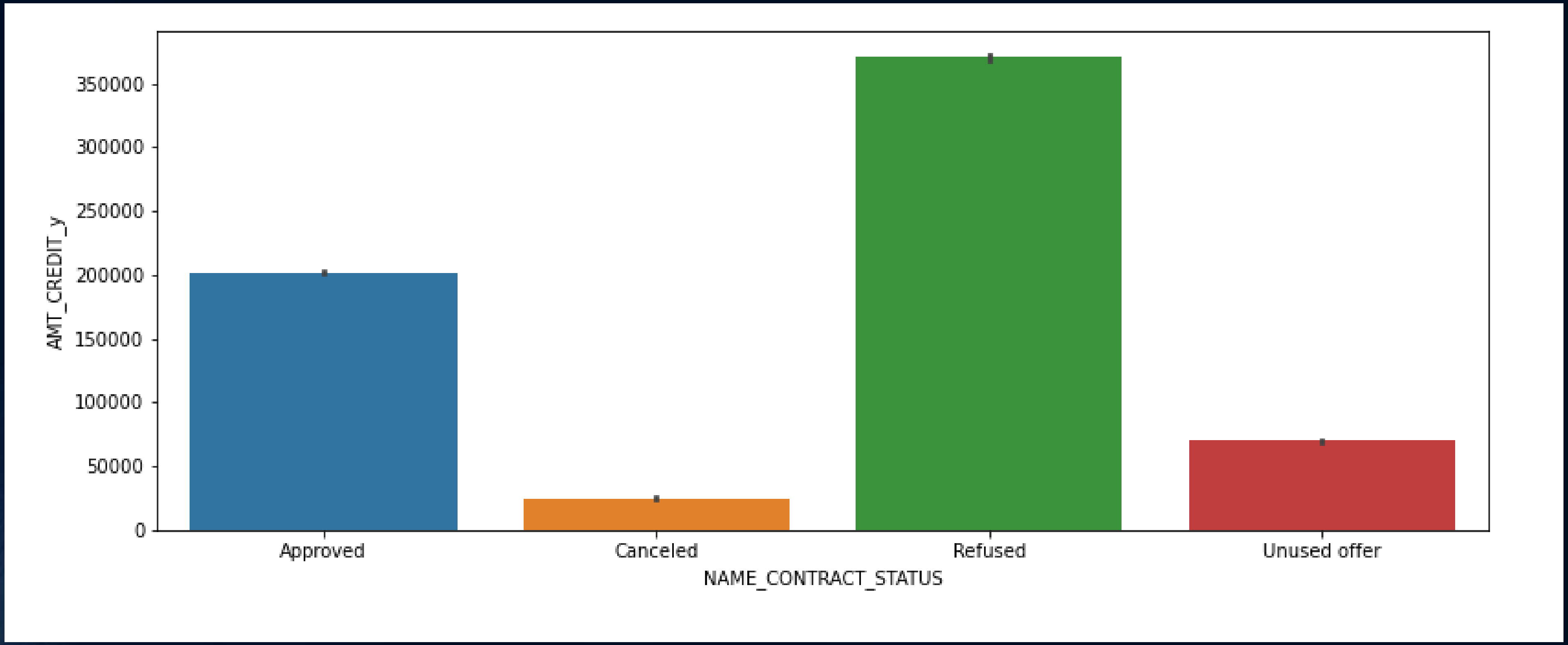
- Here we can observe POS Loans are mostly opted compare to cash, card or any other loans

UNIVARIATE ANALYSIS ON MERGED DATA



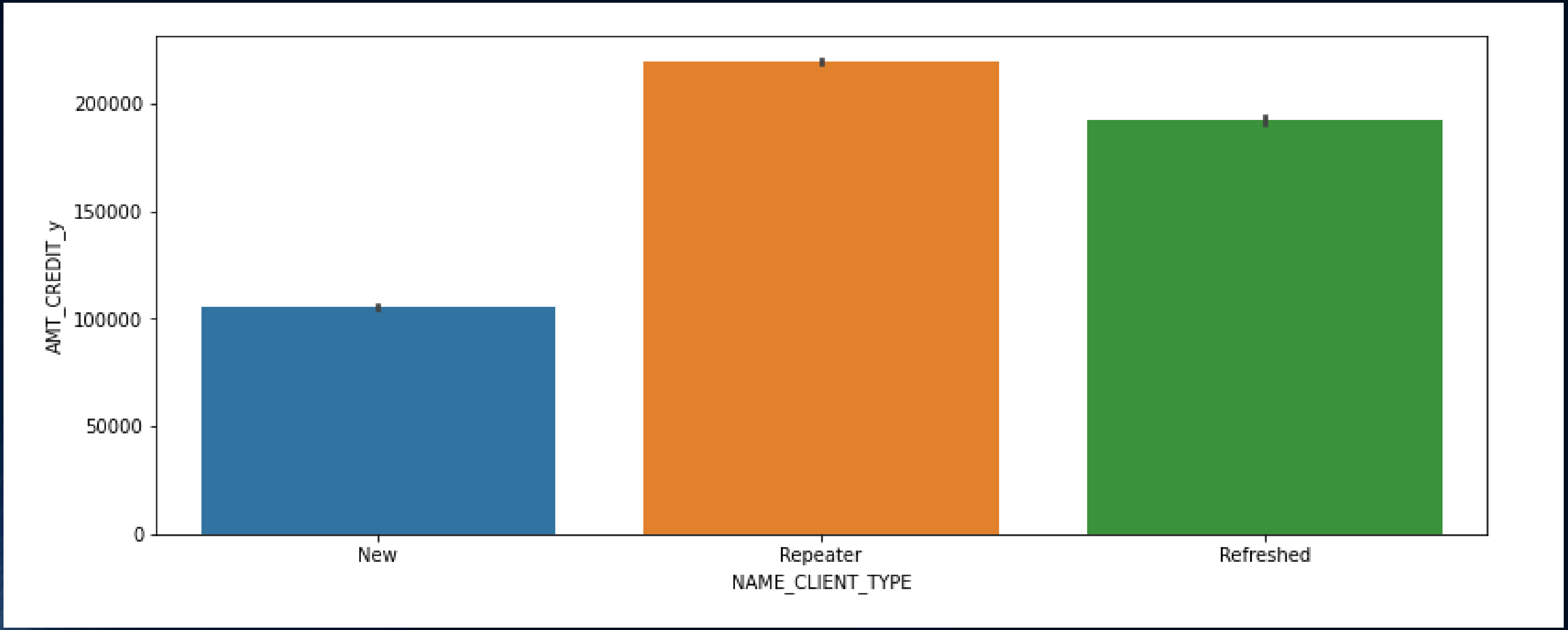
- Here we can observe that Credit and Cash offices is in majority followed by Country-wide.

BIVARIATE ANALYSIS ON MERGED DATA



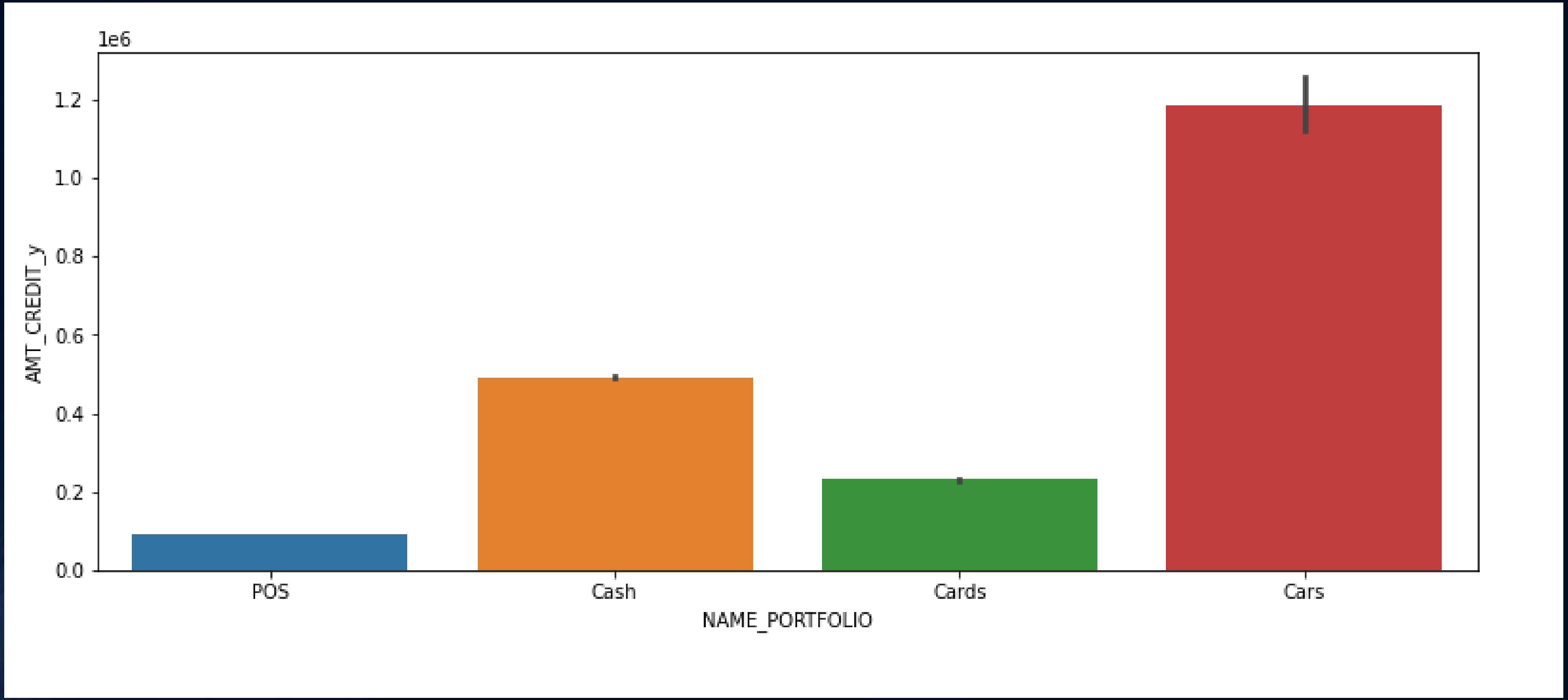
- Here we can observe that Most of the amount credit was refused in status

BIVARIATE ANALYSIS ON MERGED DATA



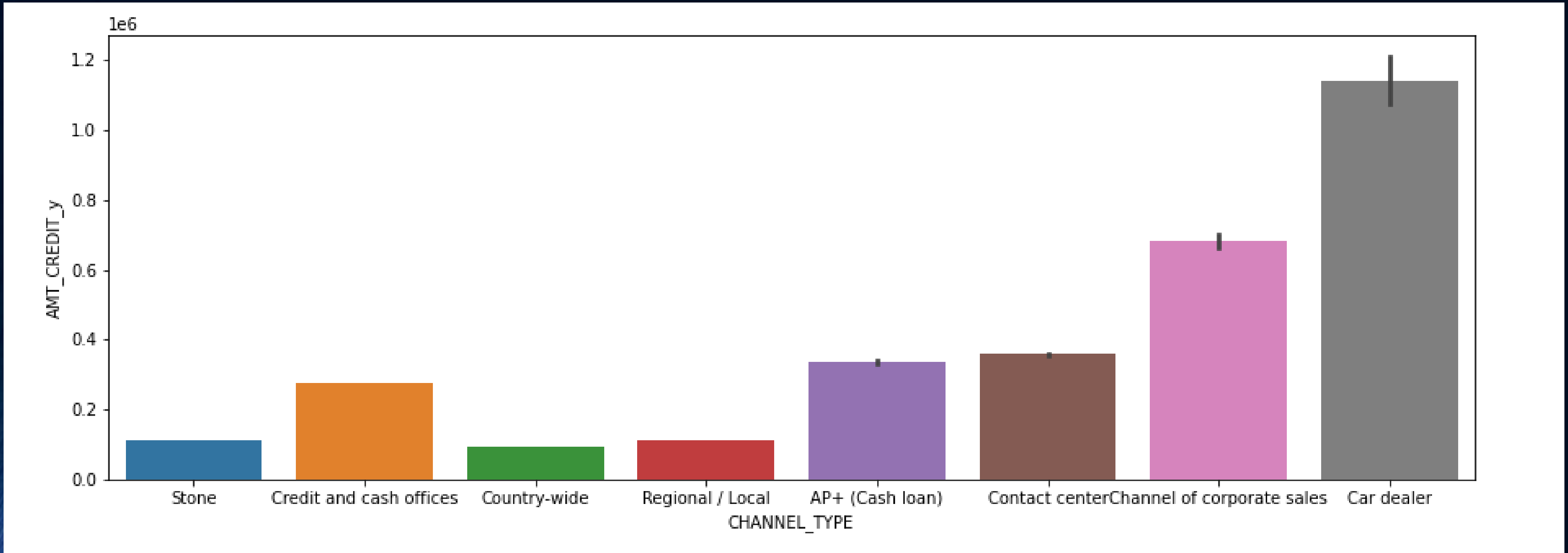
- Here we can observe Repeater client get more loan credit

BIVARIATE ANALYSIS ON MERGED DATA



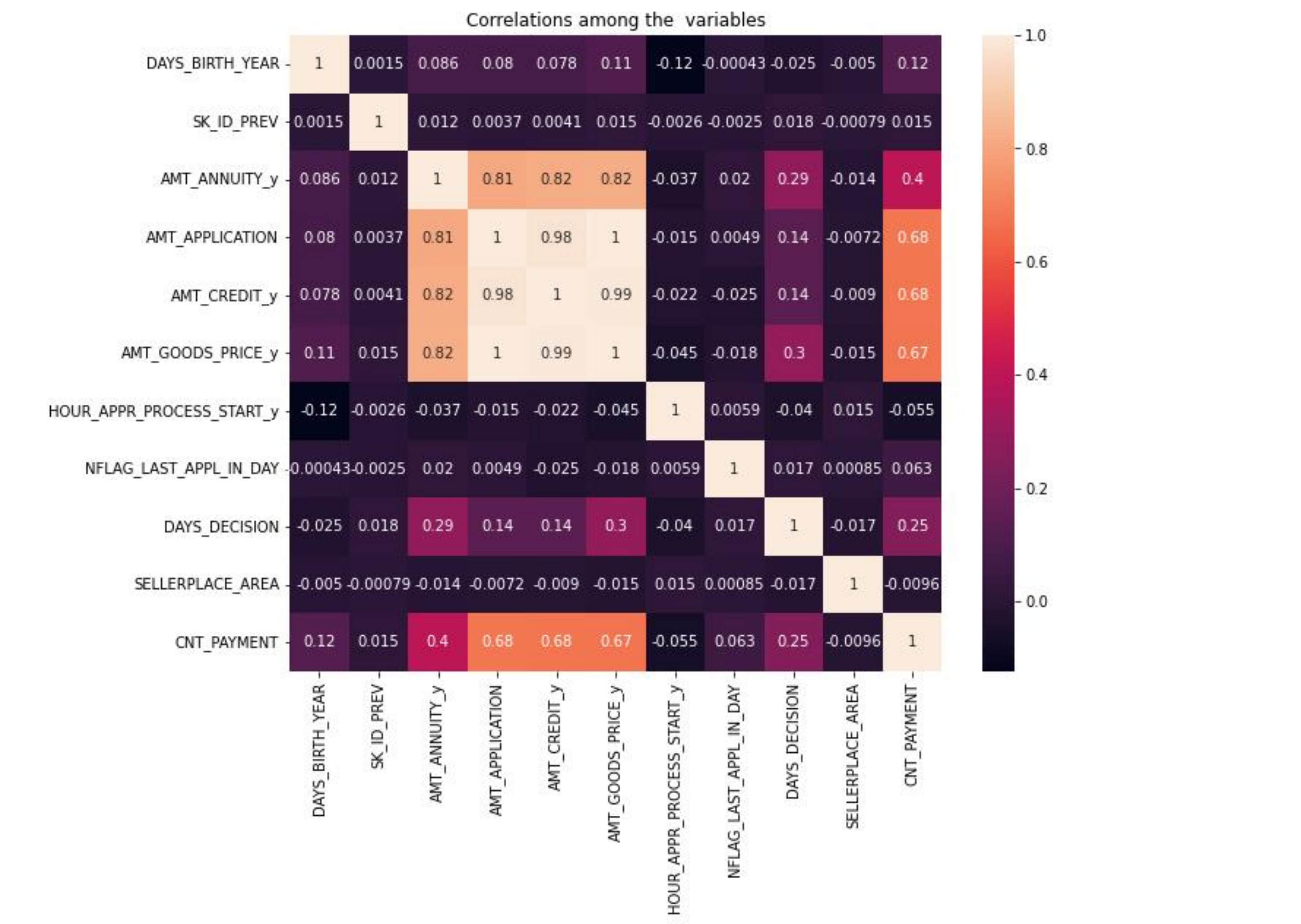
- Here we can observe, Car loans get more loan credit followed by Cash and card loans

BIVARIATE ANALYSIS ON MERGED DATA



- Here we can observe that in Car dealer channel most of the loans got credited followed by Channel of corporate sales, contact center and so on.

CORRELATION ON MERGED DATA



	DAYS_BIRTH_YEAR	SK_ID_PREV	AMT_ANNUITY_y	AMT_APPLICATION	AMT_CREDIT_y	AMT_GOODS_PRICE_y	HOUR_APPR_PROCESS_START_y	NFLAG_LAST_APPL_IN_DAY	DAYS_DECISION	SELLERPLACE_AREA	CNT_PAYMENT
DAYS_BIRTH_YEAR	1	0.0015	0.086	0.08	0.078	0.11	-0.12	-0.00043	-0.025	-0.005	0.12
SK_ID_PREV	0.0015	1	0.012	0.0037	0.0041	0.015	-0.0026	-0.0025	0.018	-0.00079	0.015
AMT_ANNUITY_y	0.086	0.012	1	0.81	0.82	0.82	-0.037	0.02	0.29	-0.014	0.4
AMT_APPLICATION	0.08	0.0037	0.81	1	0.98	1	-0.015	0.0049	0.14	-0.0072	0.68
AMT_CREDIT_y	0.078	0.0041	0.82	0.98	1	0.99	-0.022	-0.025	0.14	-0.009	0.68
AMT_GOODS_PRICE_y	0.11	0.015	0.82	1	0.99	1	-0.045	-0.018	0.3	-0.015	0.67
HOUR_APPR_PROCESS_START_y	-0.12	-0.0026	-0.037	-0.015	-0.022	-0.045	1	0.0059	-0.04	0.015	-0.055
NFLAG_LAST_APPL_IN_DAY	-0.00043	-0.0025	0.02	0.0049	-0.025	-0.018	0.0059	1	0.017	0.00085	0.063
DAYS_DECISION	-0.025	0.018	0.29	0.14	0.14	0.3	-0.04	0.017	1	-0.017	0.25
SELLERPLACE_AREA	-0.005	-0.00079	-0.014	-0.0072	-0.009	-0.015	0.015	0.00085	-0.017	1	-0.0096
CNT_PAYMENT	0.12	0.015	0.4	0.68	0.68	0.67	-0.055	0.063	0.25	-0.0096	1

- Credit amount of the loan is highly correlated or can say directly proportional to credit the client ask on the previous application and Goods price.

THANK YOU