Credit EDA Case Study

By-
Anurag Sharma &
Shubham Deshpande

Problem Statement

- To find the possible patterns in the data so that a probability can be set up to find if a particular applicant will default or not and also a genuine applications is not rejected.

Hypothesis

- H0 : The applicant will not default and will repay the loan on time.
- H1 : The applicant will default and will not repay the loan on time.

Types Of Analysis Done

- Handling of null values.
- Standardization of columns
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis

Handling Null Values for categorical and non-continuous numerical variables.

The below mentioned columns are categorical/non-continuous numerical columns, we can use the mode value (most frequent value) for each ach column to impute the null values.

1) OCCUPATION_TYPE

2) AMT_REQ_CREDIT_BUREAU_YEAR

3) AMT_REQ_CREDIT_BUREAU_MON

4) AMT_REQ_CREDIT_BUREAU_WEEK

5) AMT_REQ_CREDIT_BUREAU_DAY

6) AMT_REQ_CREDIT_BUREAU_HOUR

7) AMT_REQ_CREDIT_BUREAU_QRT

8) NAME_TYPE_SUITE

Below is a screenshot from Jupyter notebook denoting the values to be used for imputing null values.

```
--> For column OCCUPATION_TYPE, the mode value - Laborers can used to impute null values
--> For column AMT_REQ_CREDIT_BUREAU_YEAR, the mode value - 0.0 can used to impute null values
--> For column AMT_REQ_CREDIT_BUREAU_MON, the mode value - 0.0 can used to impute null values
--> For column AMT_REQ_CREDIT_BUREAU_WEEK, the mode value - 0.0 can used to impute null values
--> For column AMT_REQ_CREDIT_BUREAU_DAY, the mode value - 0.0 can used to impute null values
--> For column AMT_REQ_CREDIT_BUREAU_HOUR, the mode value - 0.0 can used to impute null values
--> For column AMT_REQ_CREDIT_BUREAU_QRT, the mode value - 0.0 can used to impute null values
--> For column NAME_TYPE_SUITE, the mode value - Unaccompanied can used to impute null values
```

Handling Null Values for Numerical variables.

Below is the logic used for imputing null values:

1) If there are outliers, we'll use median value to impute the missing values.

2) if there are no outliers, we'll use the mean value to impute the missing values.

The below columns had outliers, 'median' values used to replace null columns:

1) AMT_ANNUITY

2) AMT_GOODS_PRICE

3) CNT_FAM_MEMBERS

The below columns don't have any outliers, 'mean' values used to impute the missing values:

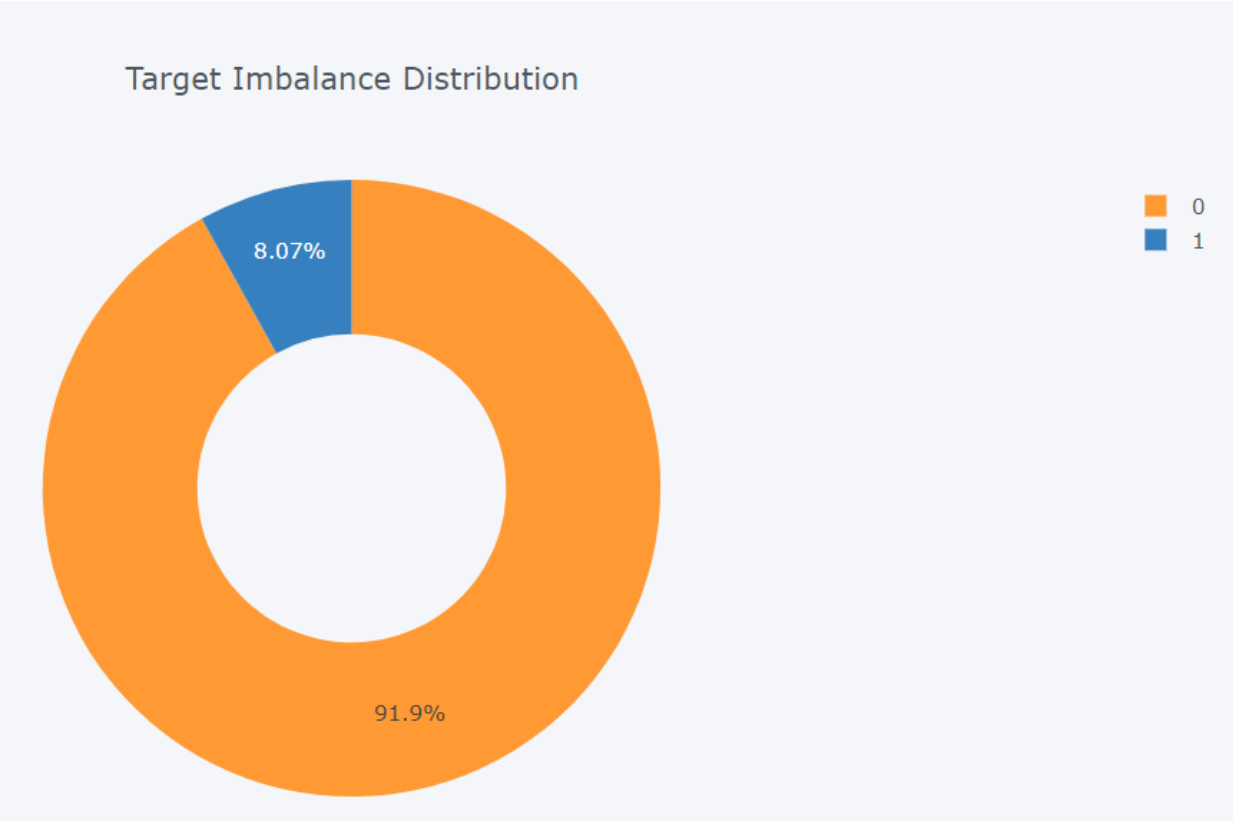
1) EXT_SOURCE_2

2) EXT_SOURCE_3

Below is the screenshot depicting the mean/median values that should be used for imputing null values.

```
--> For column AMT_ANNUITY, the median value - 24903.0 can used to impute null values
--> For column AMT_GOODS_PRICE, the median value - 450000.0 can used to impute null values
--> For column CNT_FAM_MEMBERS, the median value - 2.0 can used to impute null values
--> For column EXT_SOURCE_2, the mean value - 0.5143926741308463 can used to impute null values
--> For column EXT_SOURCE_3, the mean value - 0.5108529061800121 can used to impute null values
```

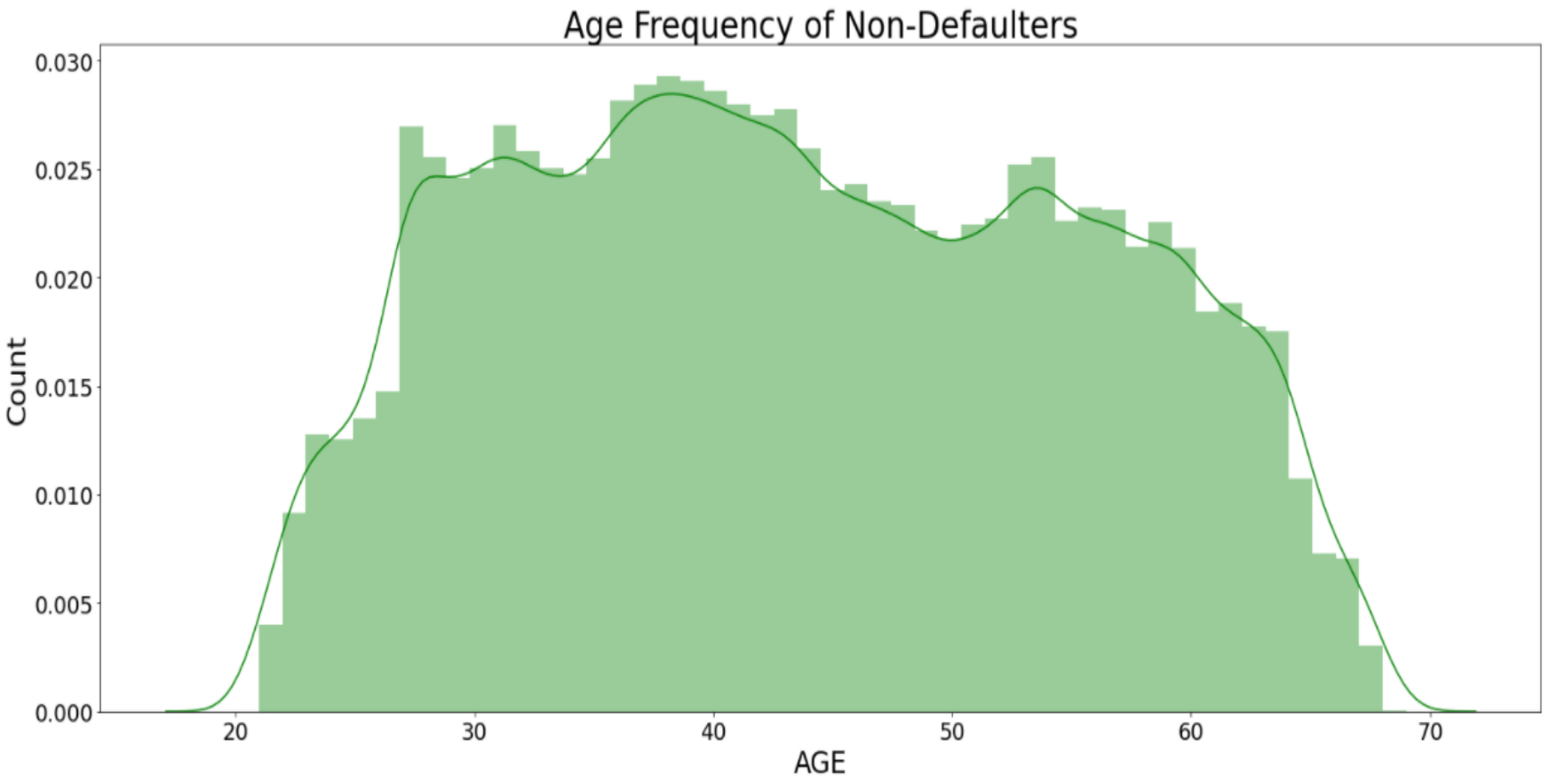
Target Imbalance



- From the pie chart above, we can clearly observe that imbalance is very high between the target variables i.e. applicants with no payment difficulties(\~8%) are way more than applicants with payment difficulties(~92%).

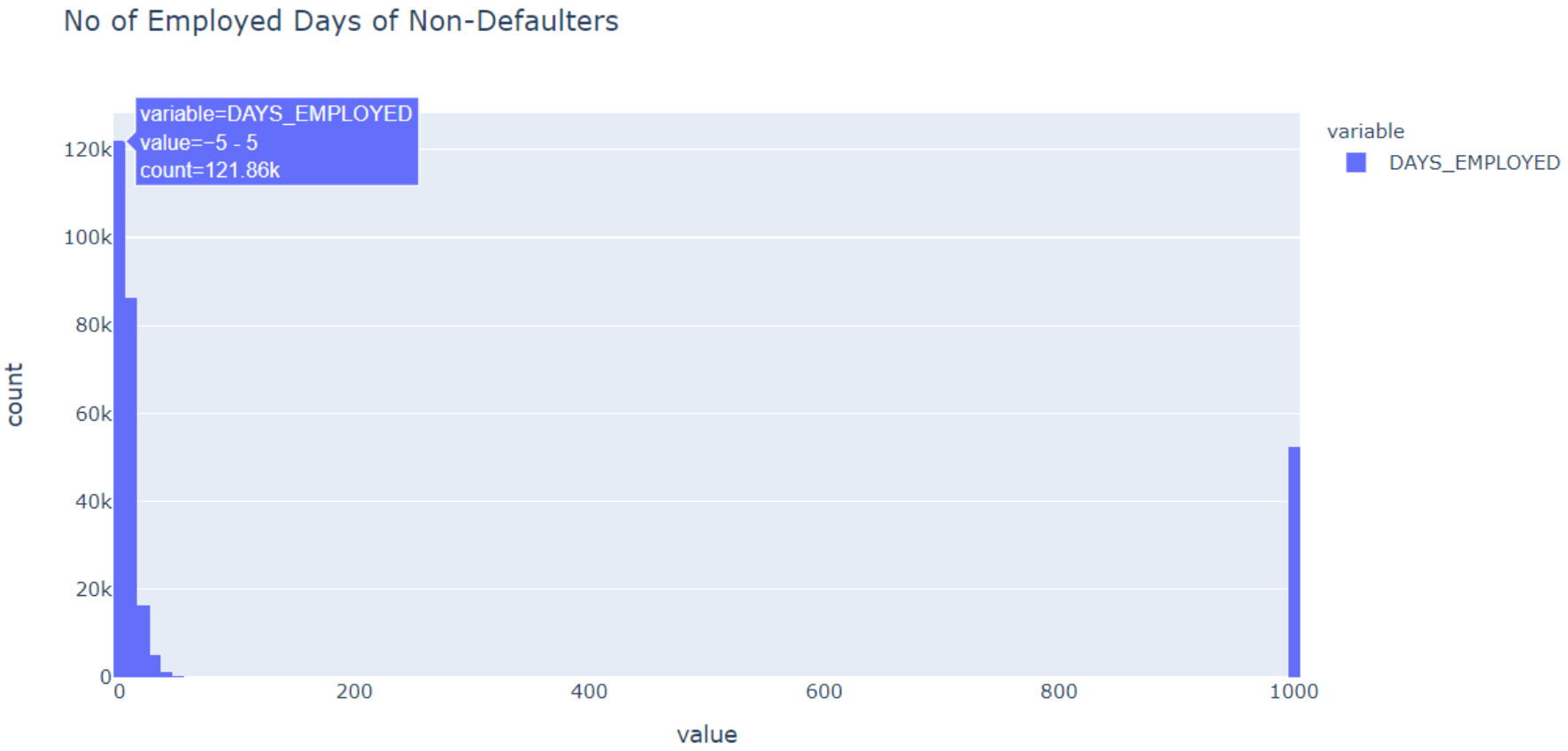
Univariate Analysis – Numeric Data on Target 0

AGE



- Most of the applicants are middle aged. To be more specific, in middle aged they lie in the range of 35-45.

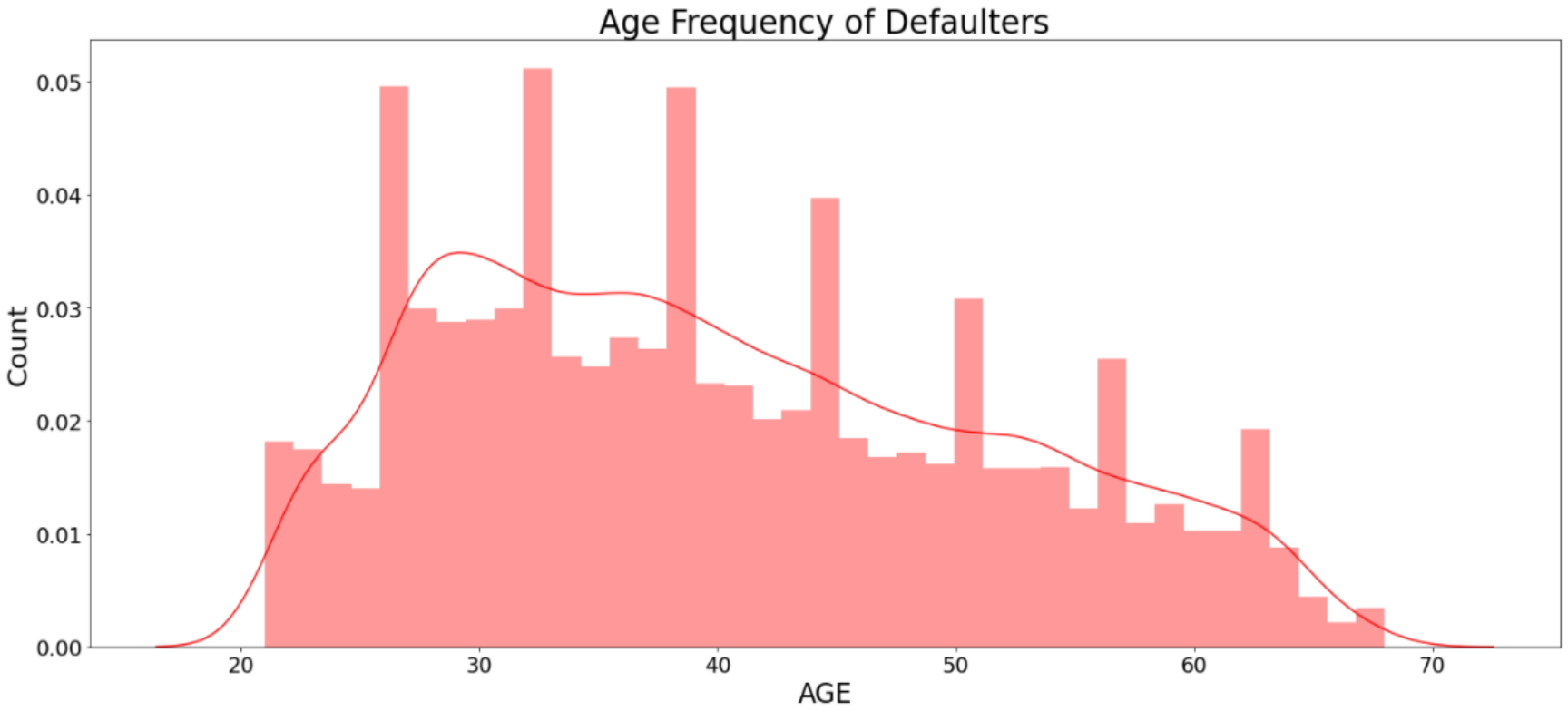
No of Days Employed



- People who are employed for a period of 5 years have highest probability that they will not default.

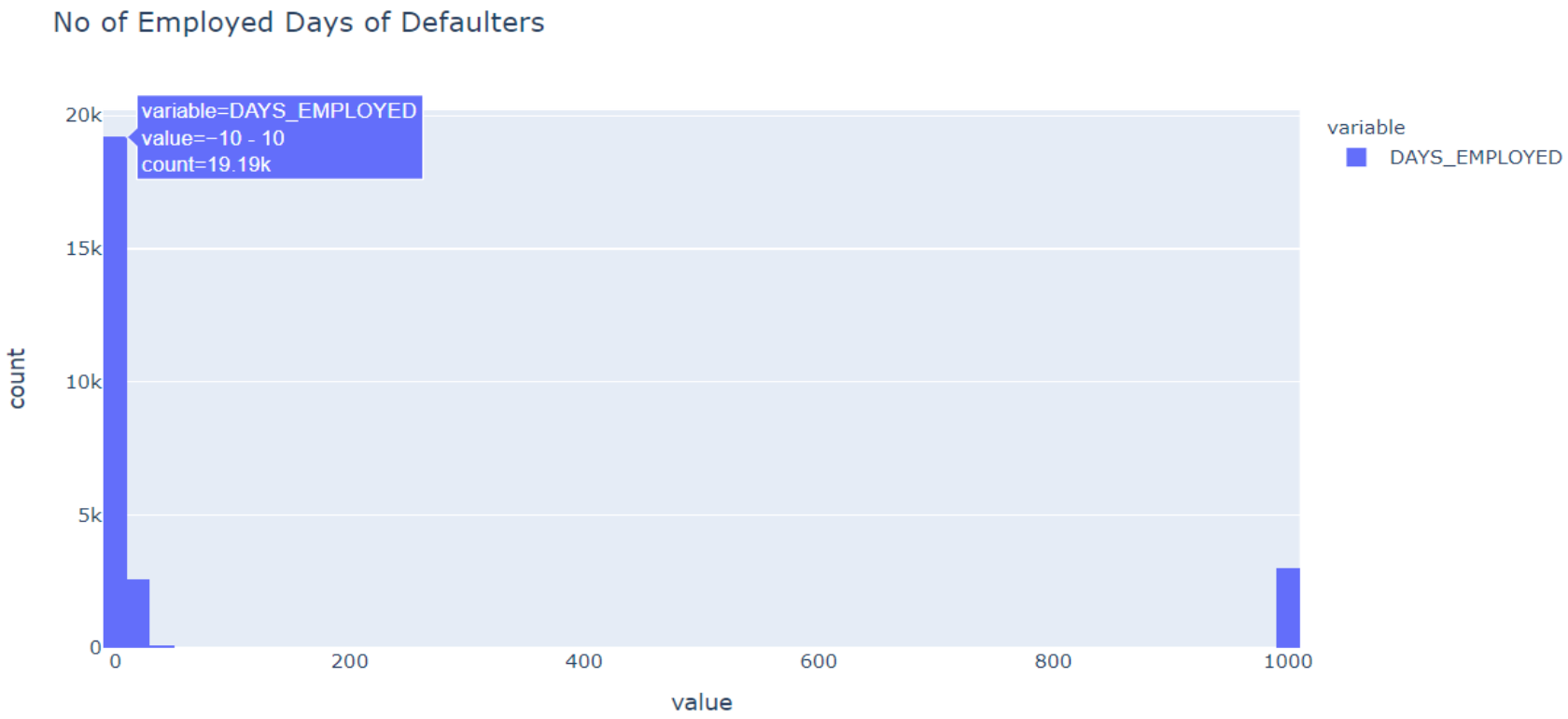
Univariate Analysis – Numeric Data on Target 1

AGE



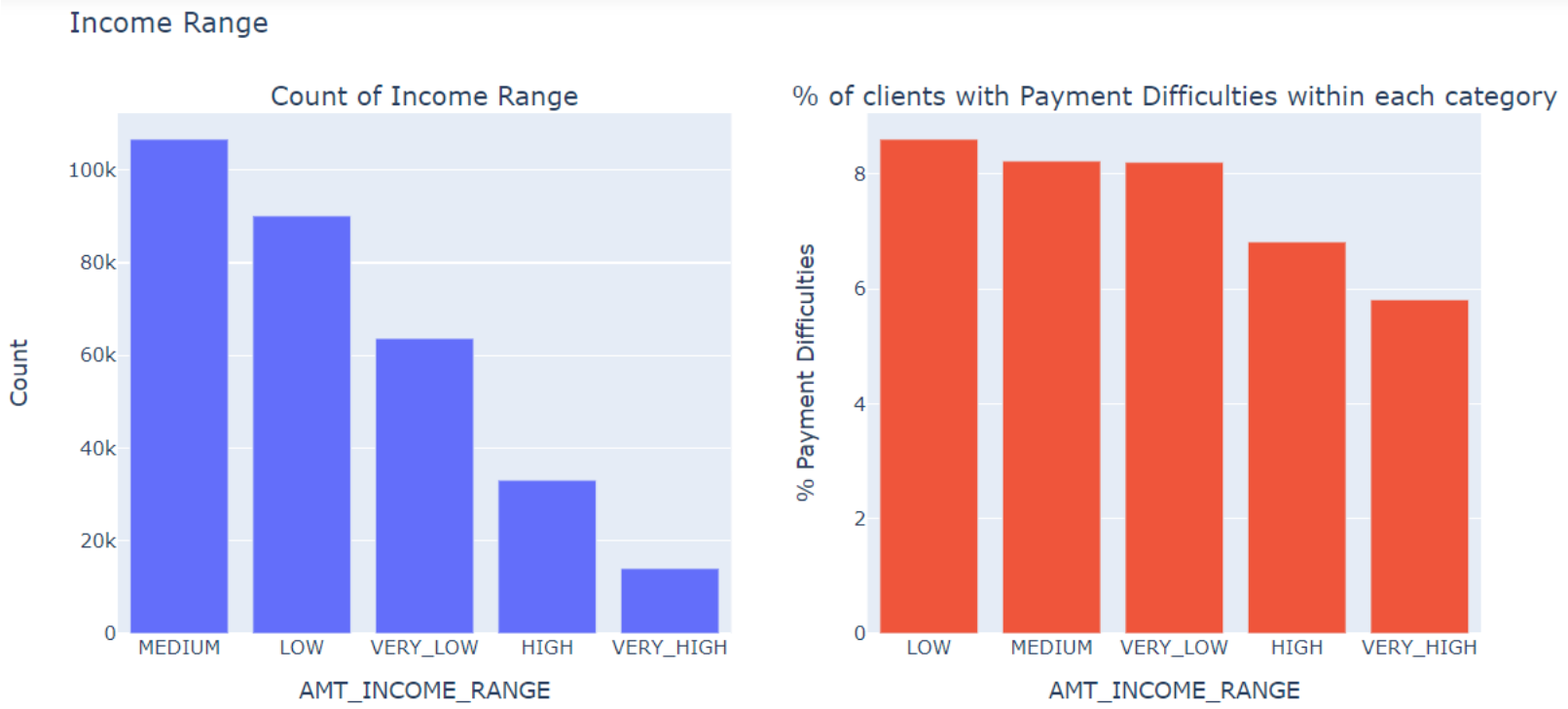
- It can be seen here that the highest number of defaulters lie in the age group of around 30.
- As the age increases, the possibility of the person defaulting decreases.

No of Days Employed



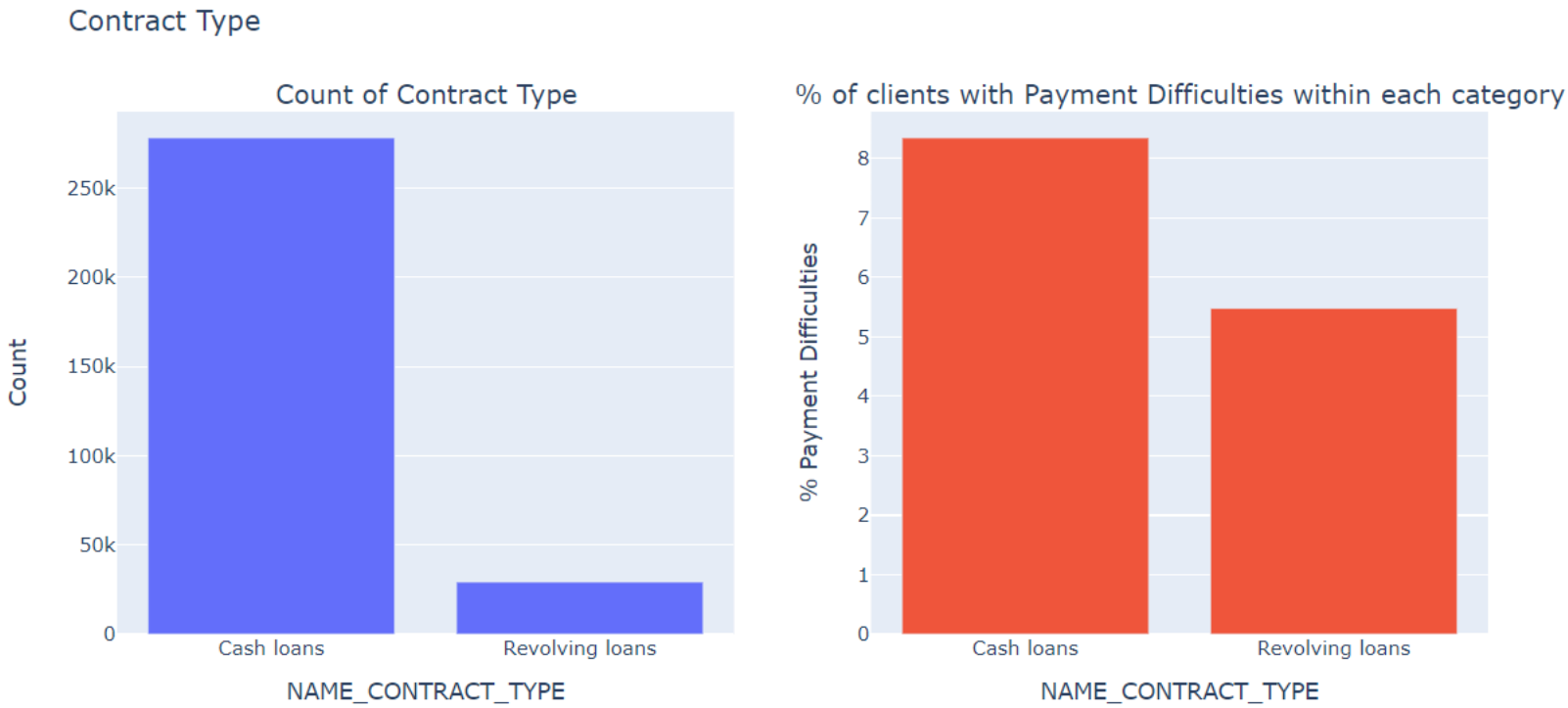
- People who are employed for a period of 10 years have highest probability that they will default.

Defaulters % : Income Range



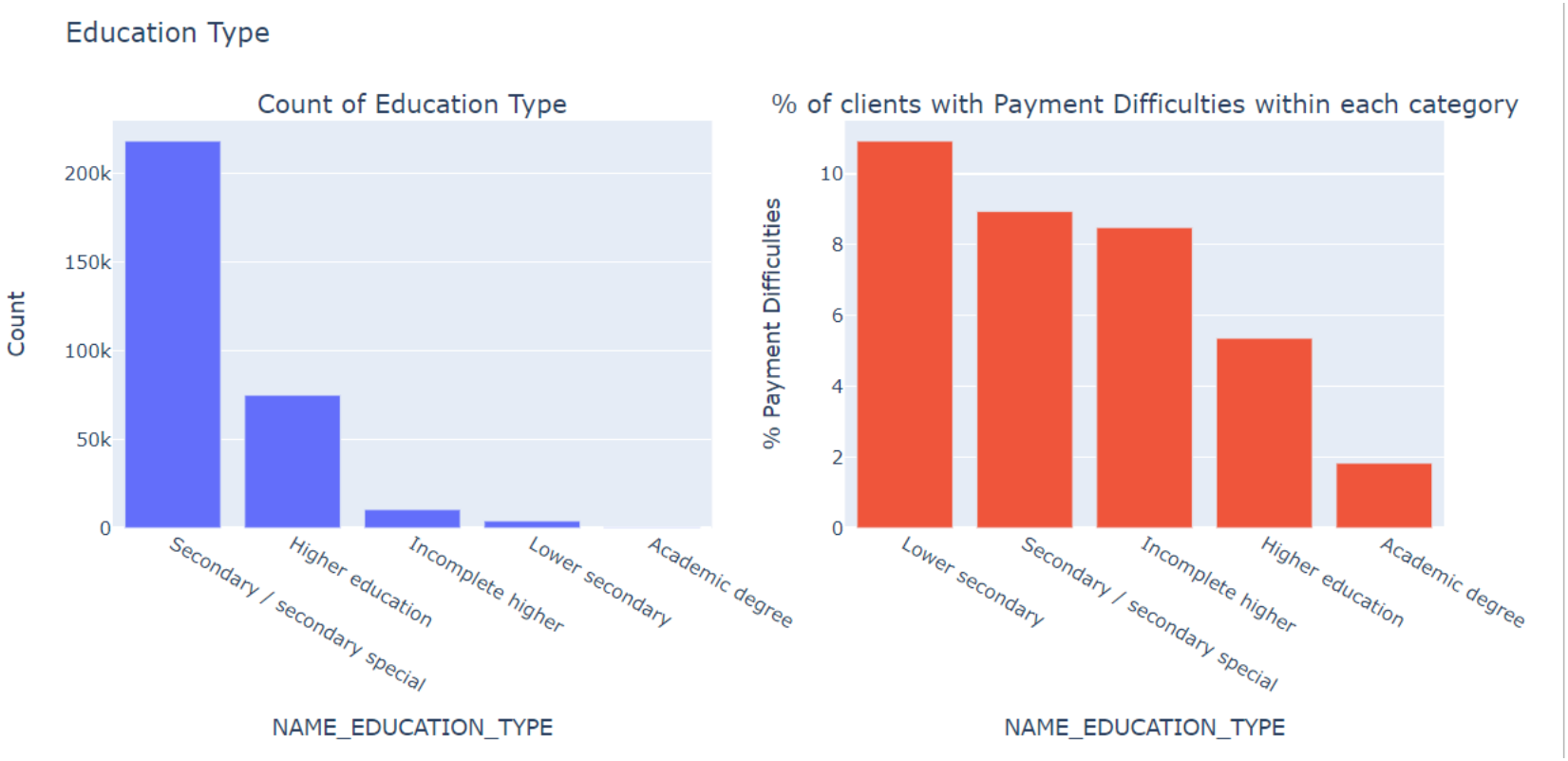
- Highest defaulters lie in the low income range followed by medium income range.

Defaulters % : Contract Type



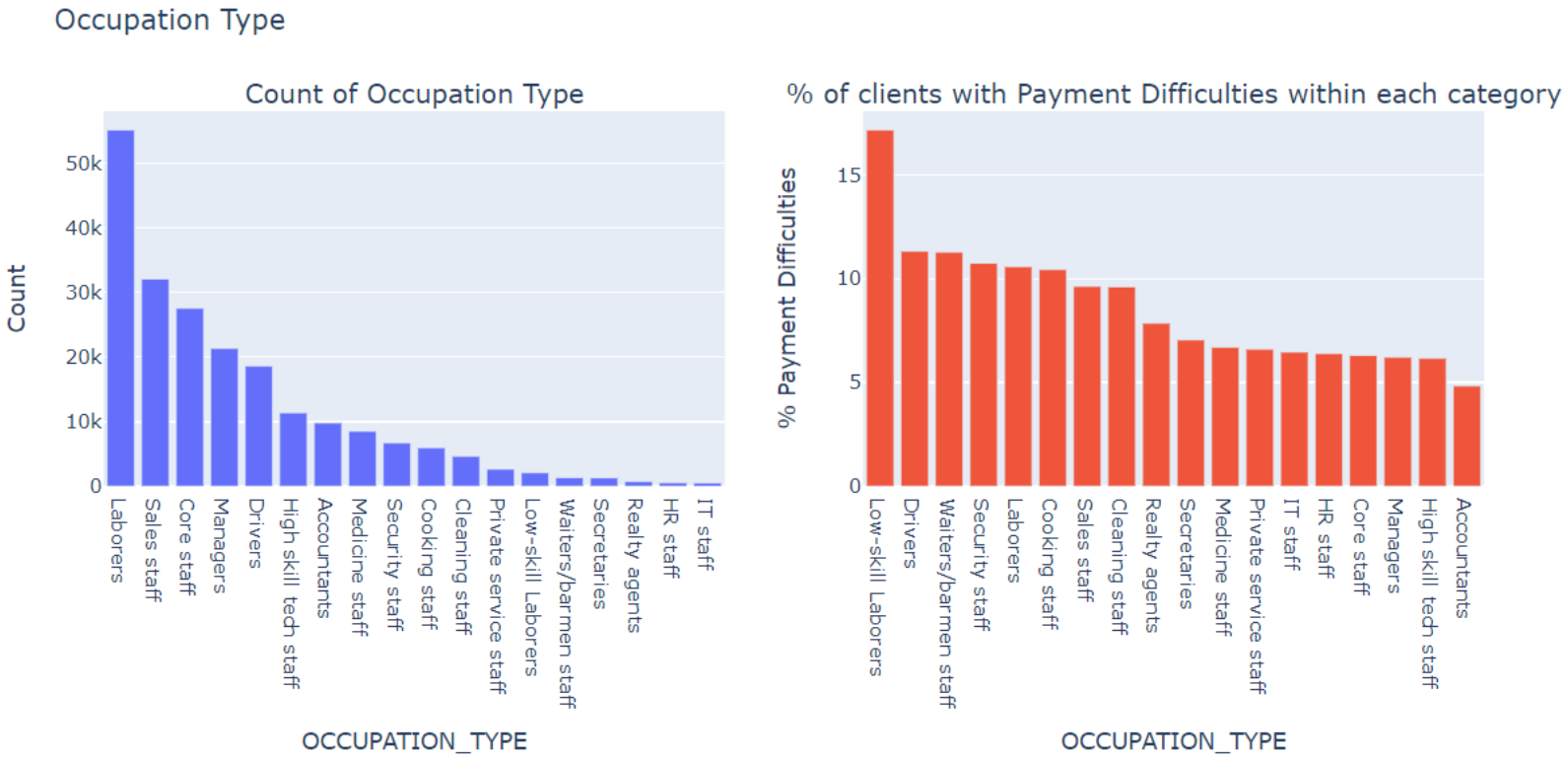
- Even though there are less applications of revolving loans yet the number of defaulters is higher in revolving loans when proportion is compared.

Defaulters % : Education Type



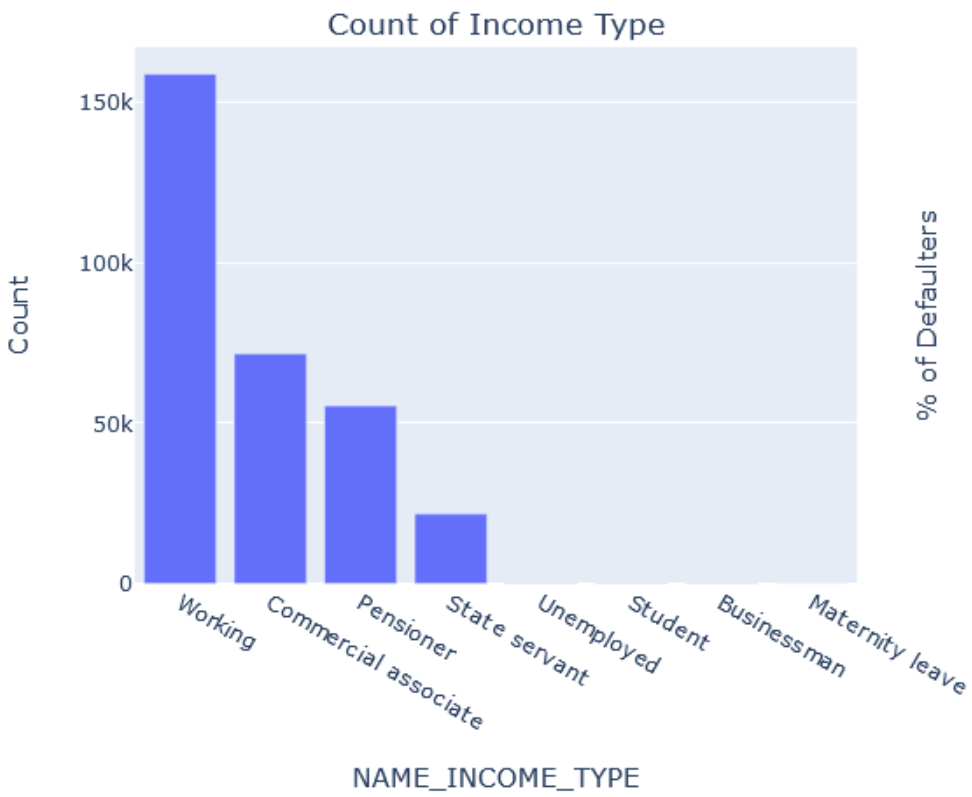
- Lower secondary, secondary and Incomplete higher have most of the defaulters in proportion to number of applications.

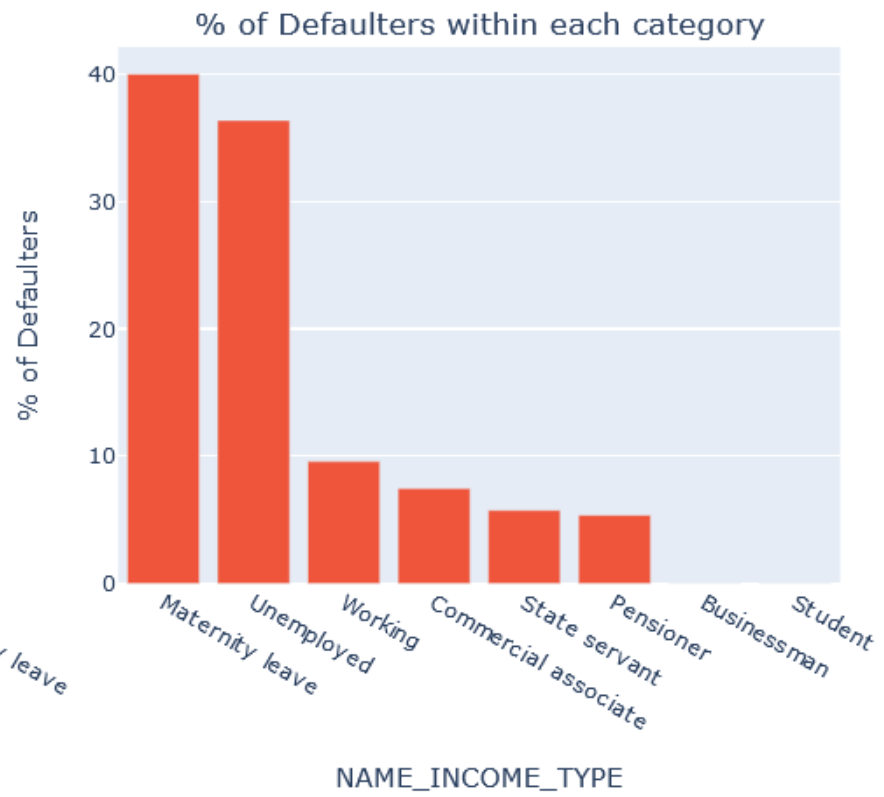
Defaulters % : Occupation Type



- Low skill laborers are less when it comes to number of loan applications but they are the top most in terms of defaulters.

Defaulters % : Income Type

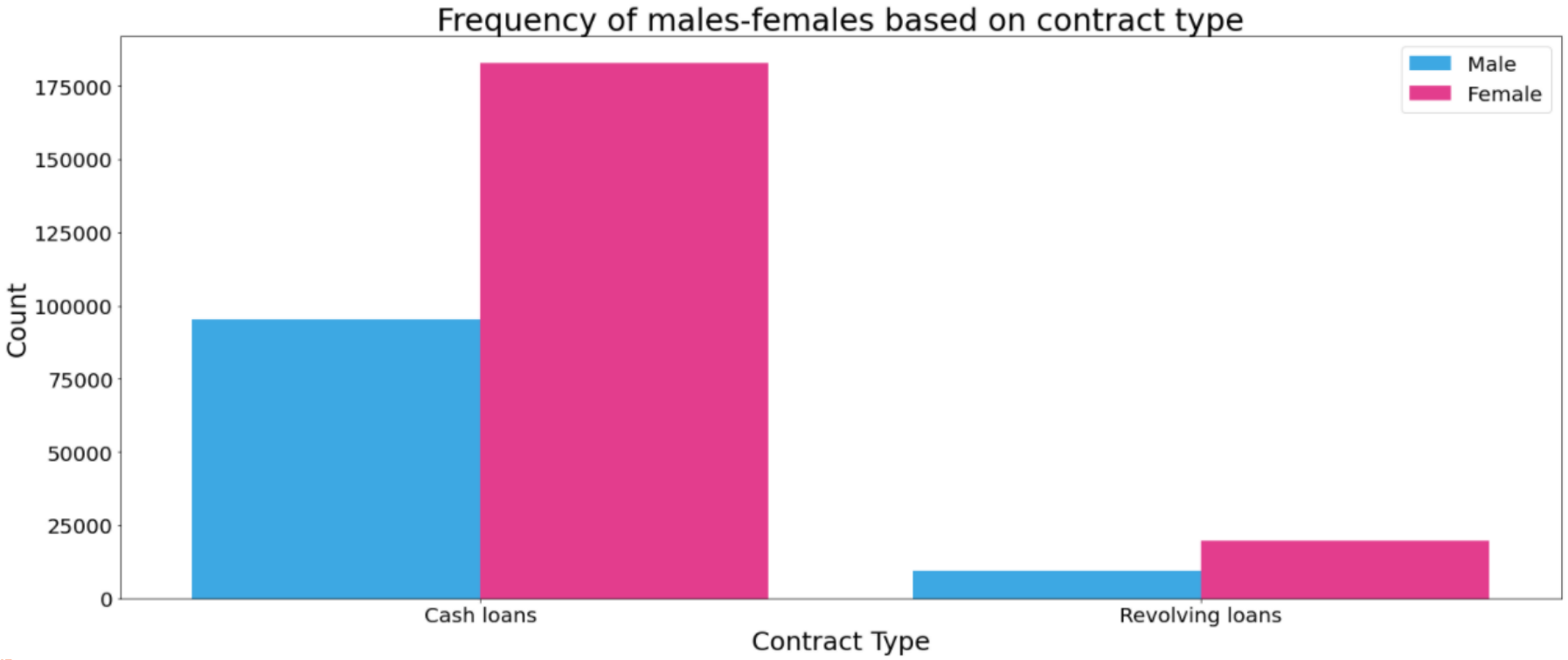




- From the plots above we can say that clients with 'Maternity leave' Income type have maximum % of payment difficulties followed by 'UNEMPLOYED' although the count of both these categories is far less than others.

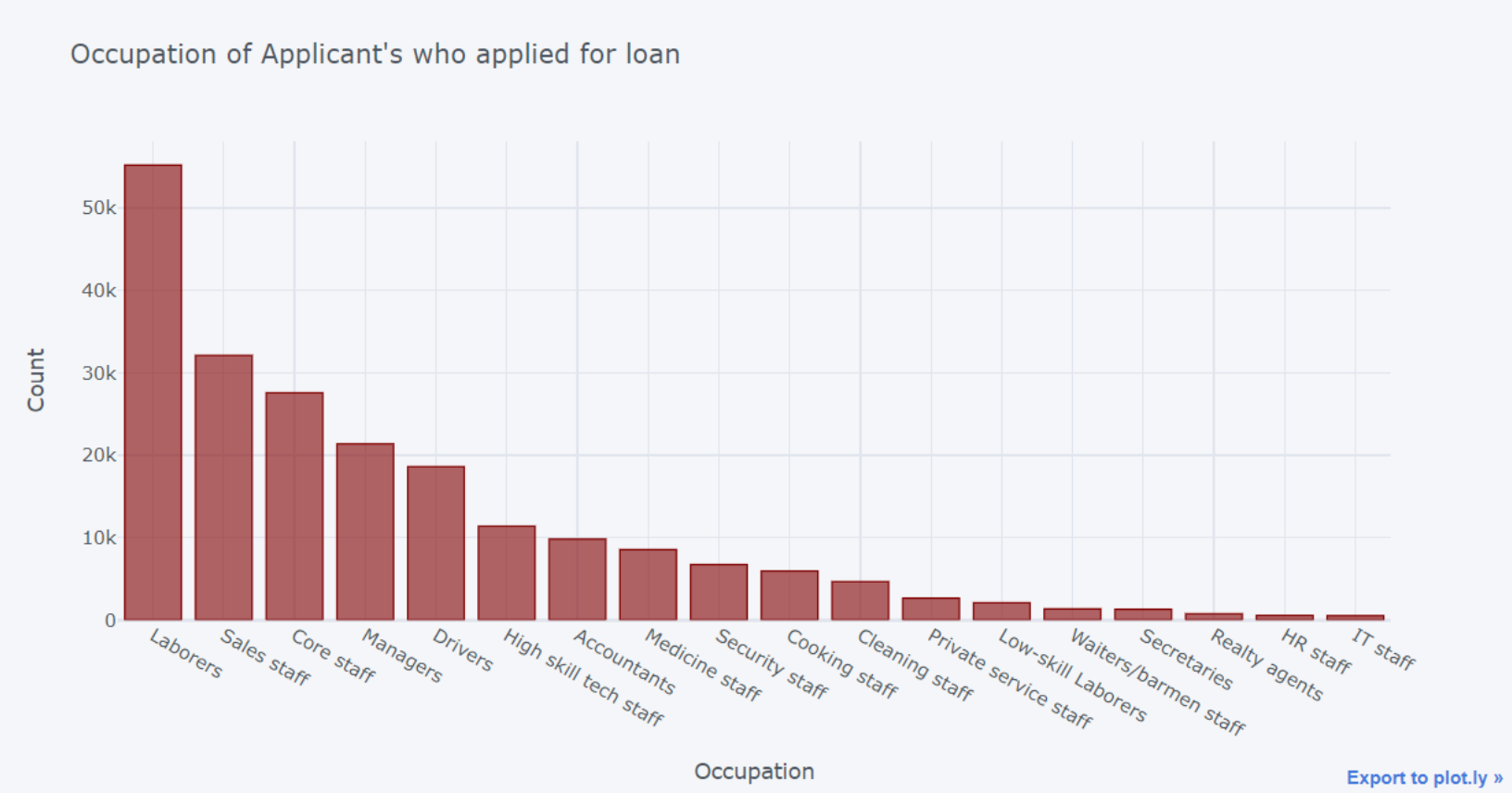
Univariate Analysis – Categorical Data

Frequency of Contract Type



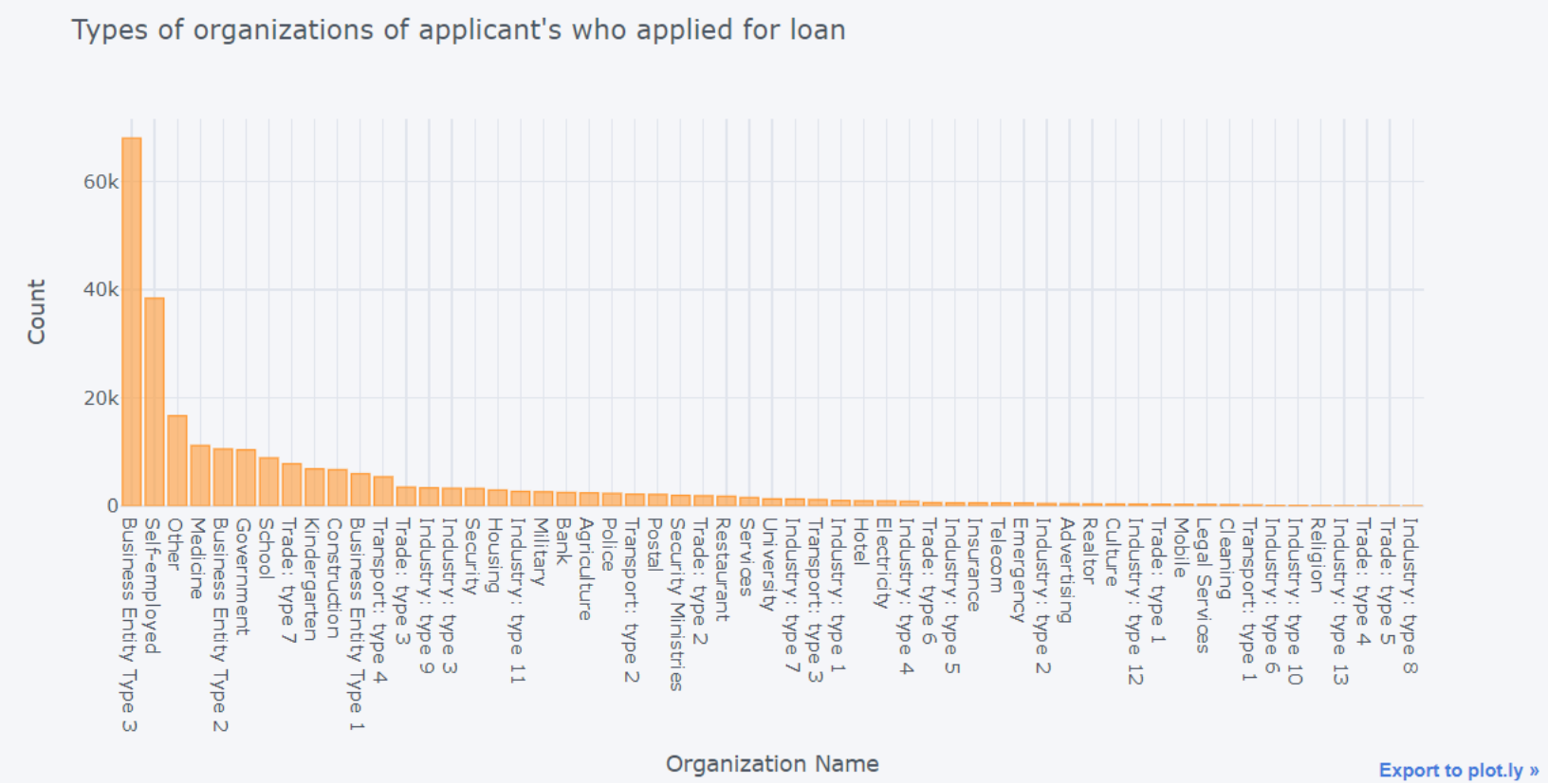
- Cash loans is the highest amongst applications.
- Females are highest in both the contracts.

Occupation of the Applicants



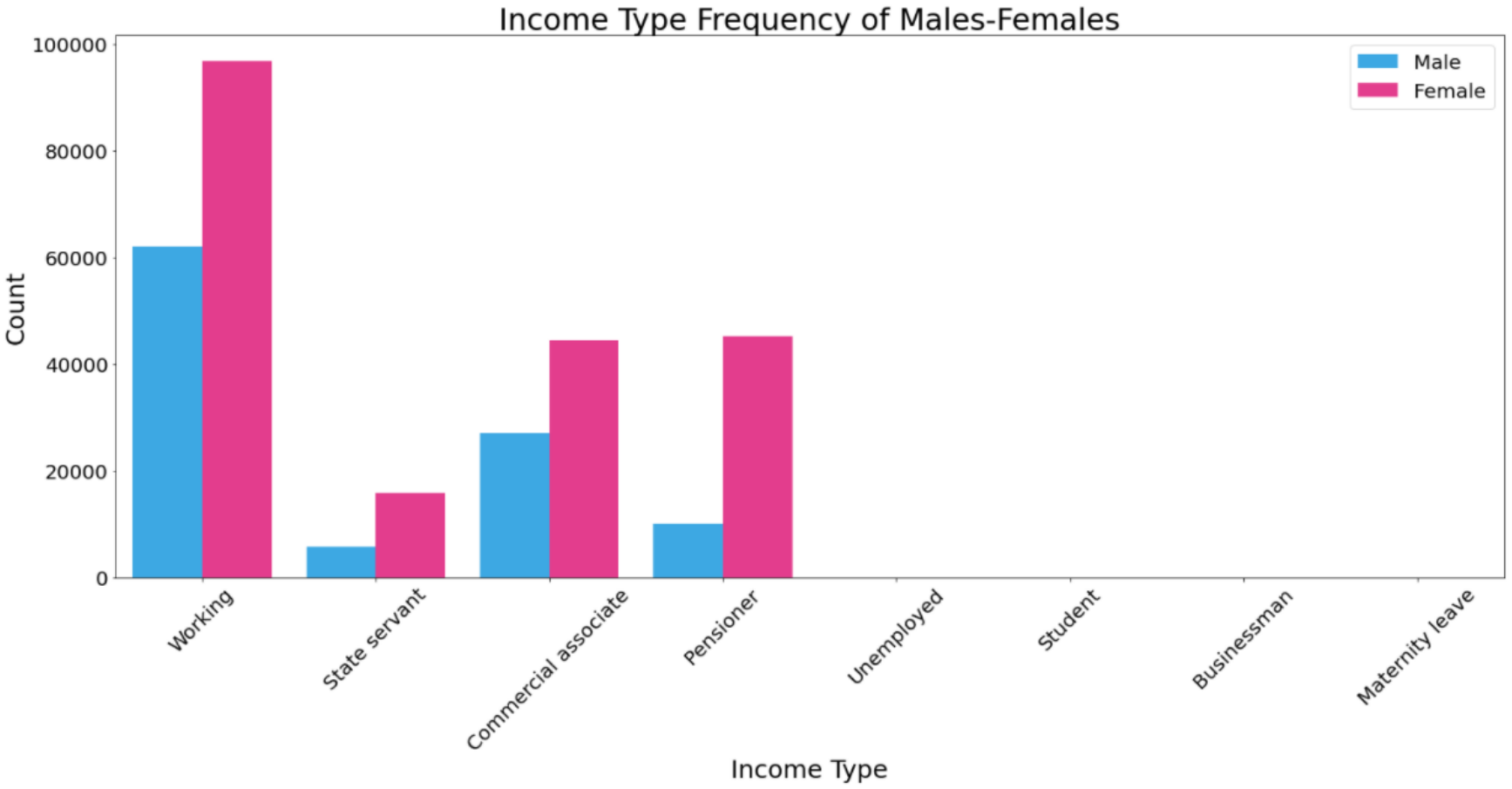
- Laborers is the highest occupation type among applicants while IT staff is the lowest

Organizations of Applicants



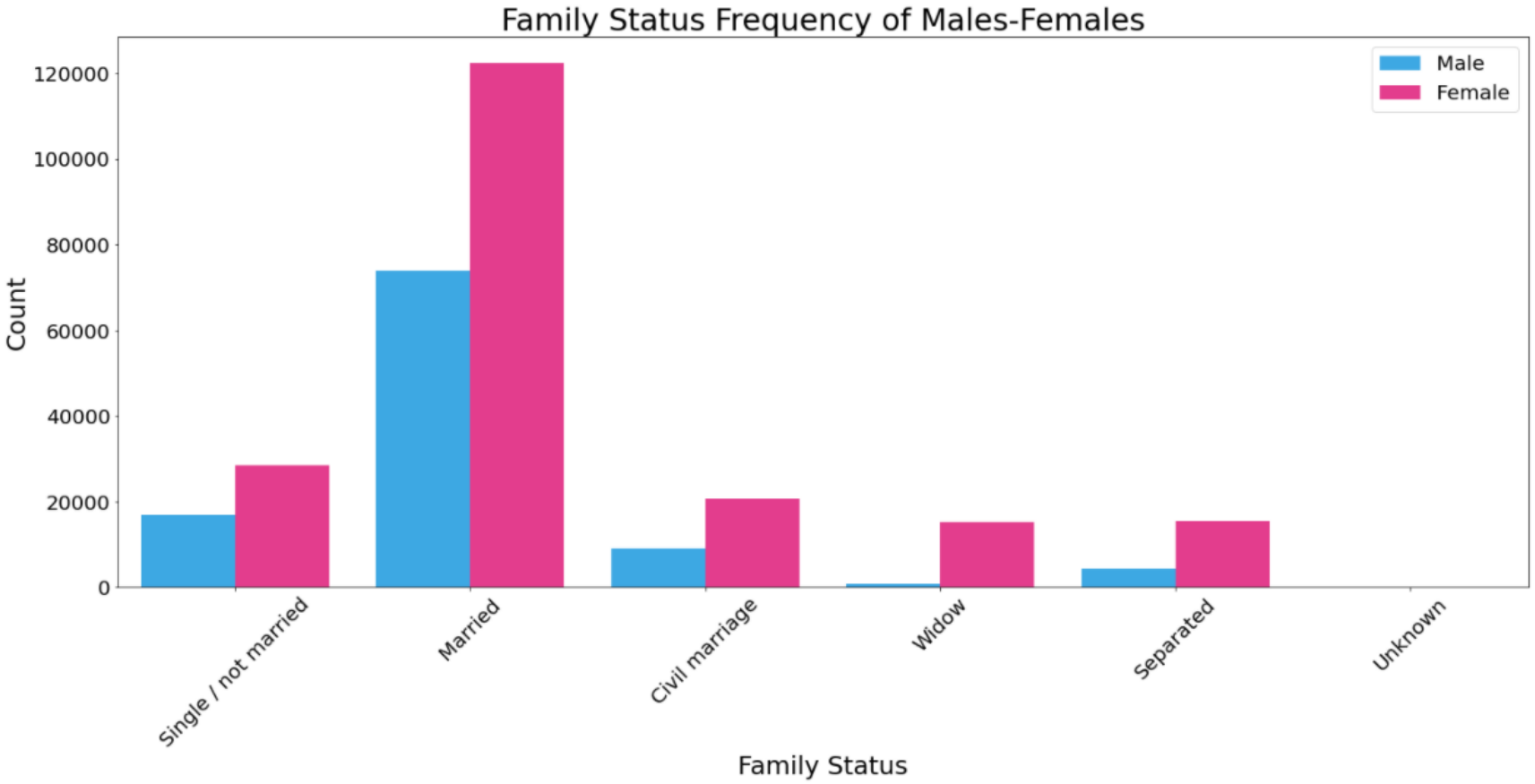
- Business Entity Type 3 and self employed are the highest for applying loans
- There are less clients from ‘Industry: Type 8’, ‘Trade : Type 5’

Income Source



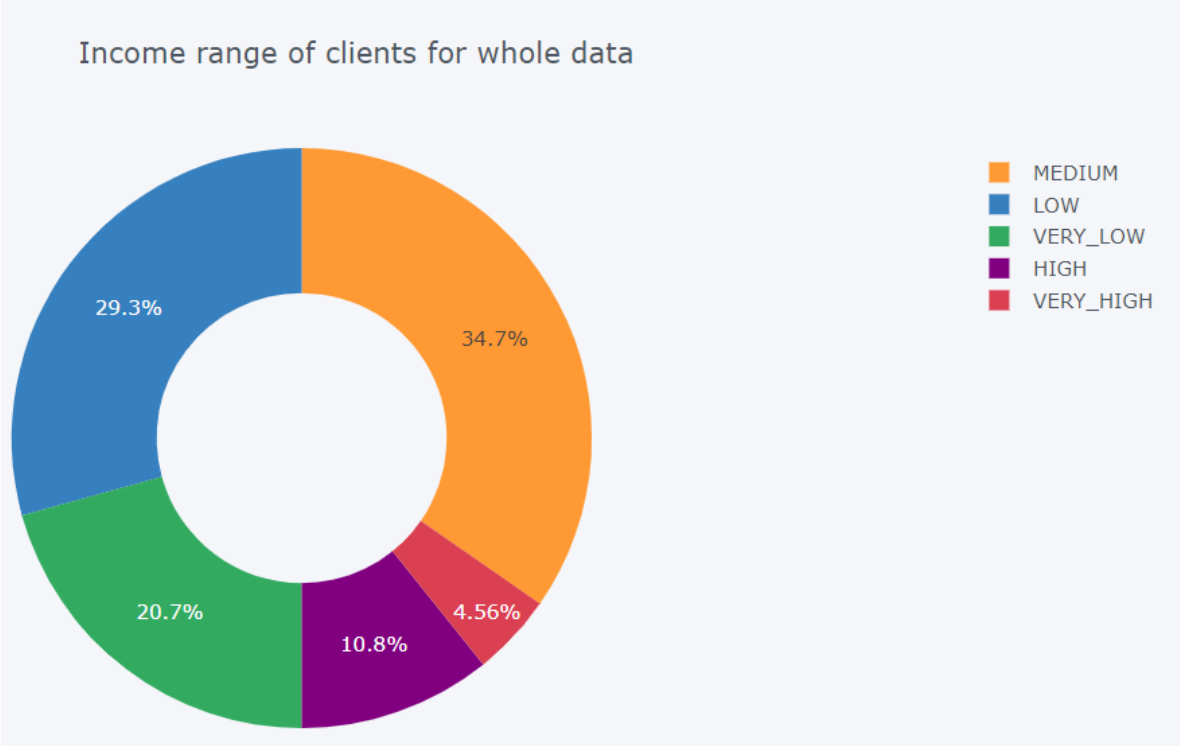
- Working people have the highest number of applications.

Family Status



- Married people have the highest number of applications
- Widow males apply for loans very less than that of females.

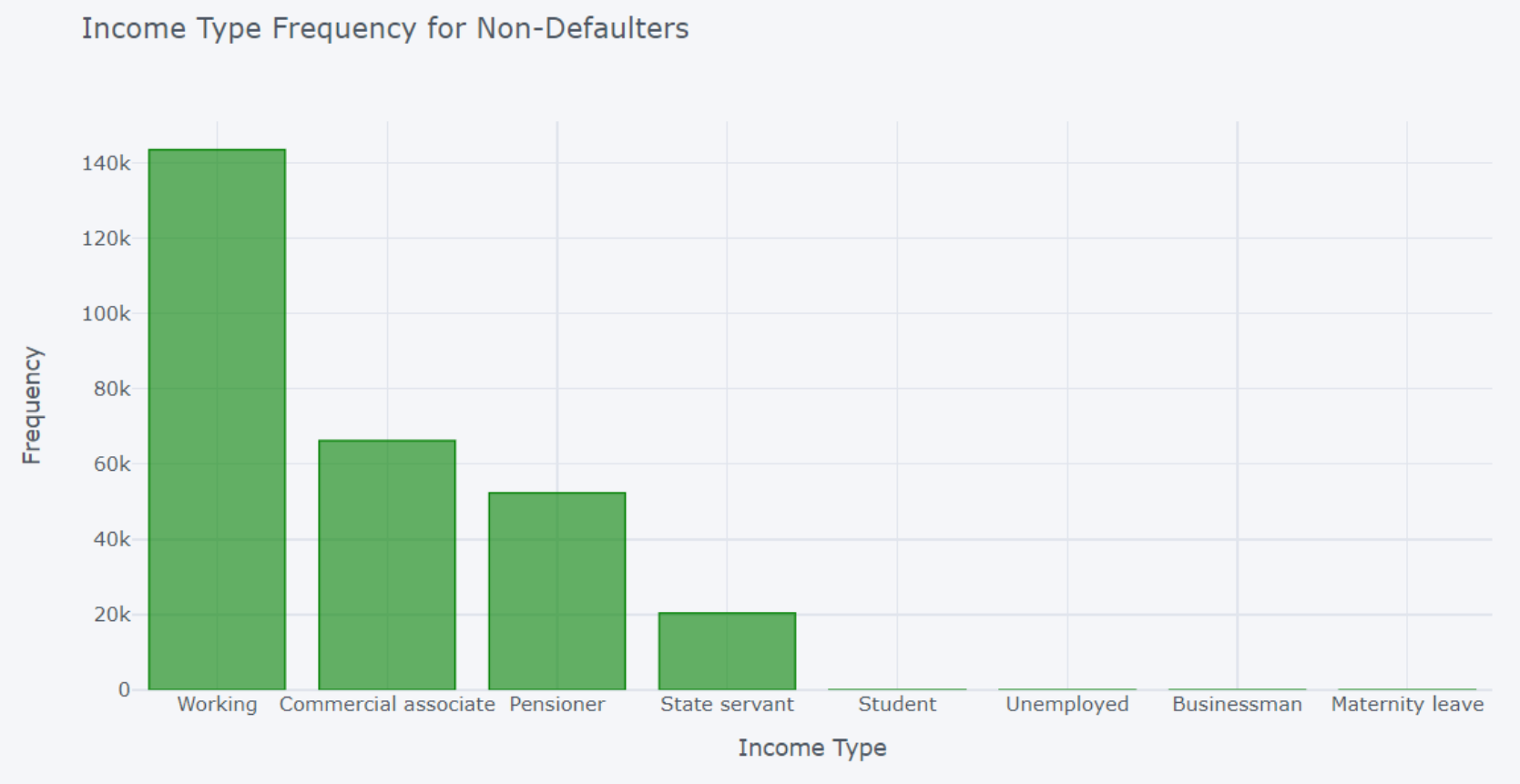
Income Range



- As the income of people increases, they do not take much loans
- Medium income range people have the highest number of applications.

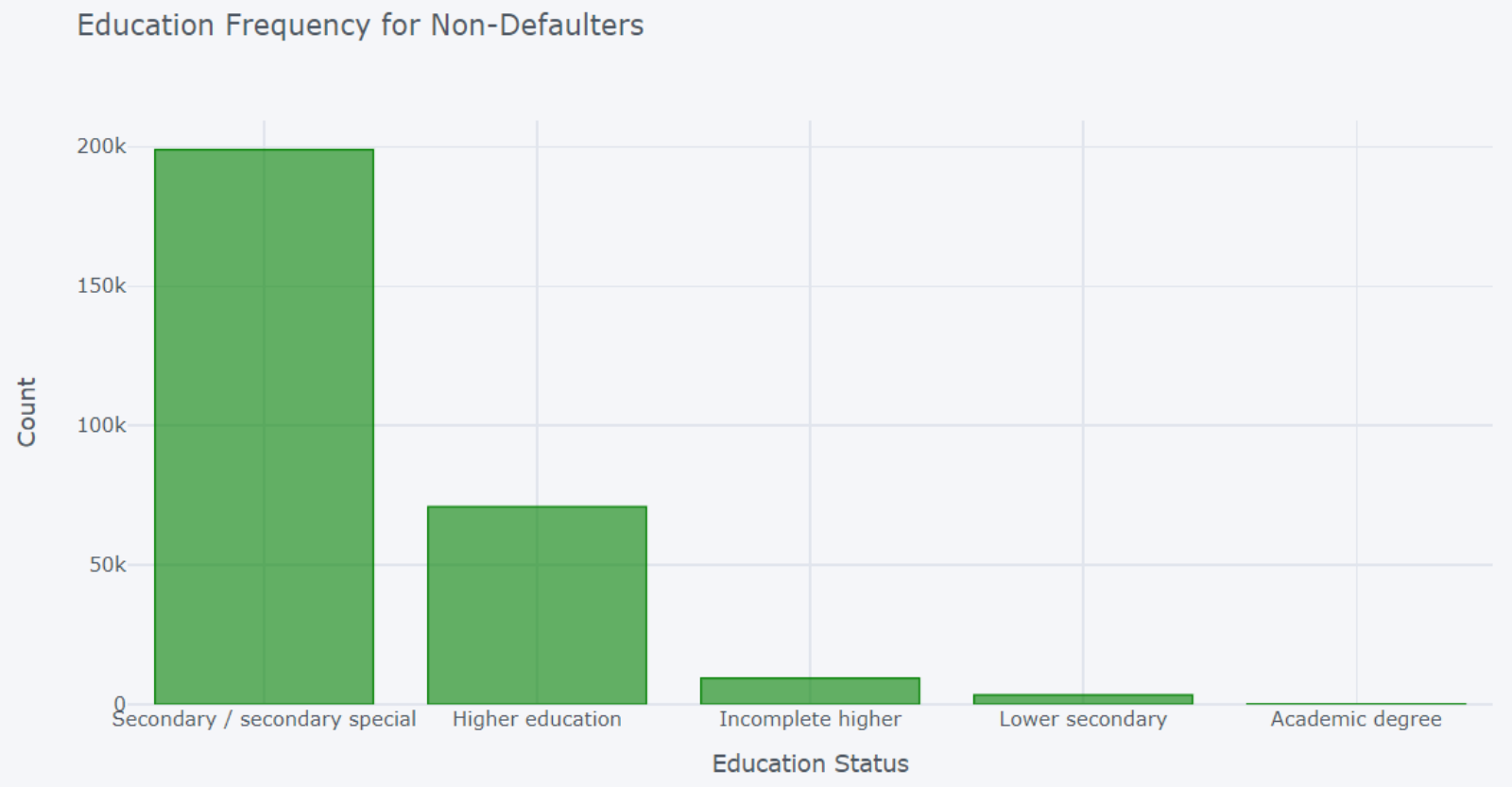
Univariate Analysis – Categorical Data On Target 0

Income Source



- Working people have the highest application.

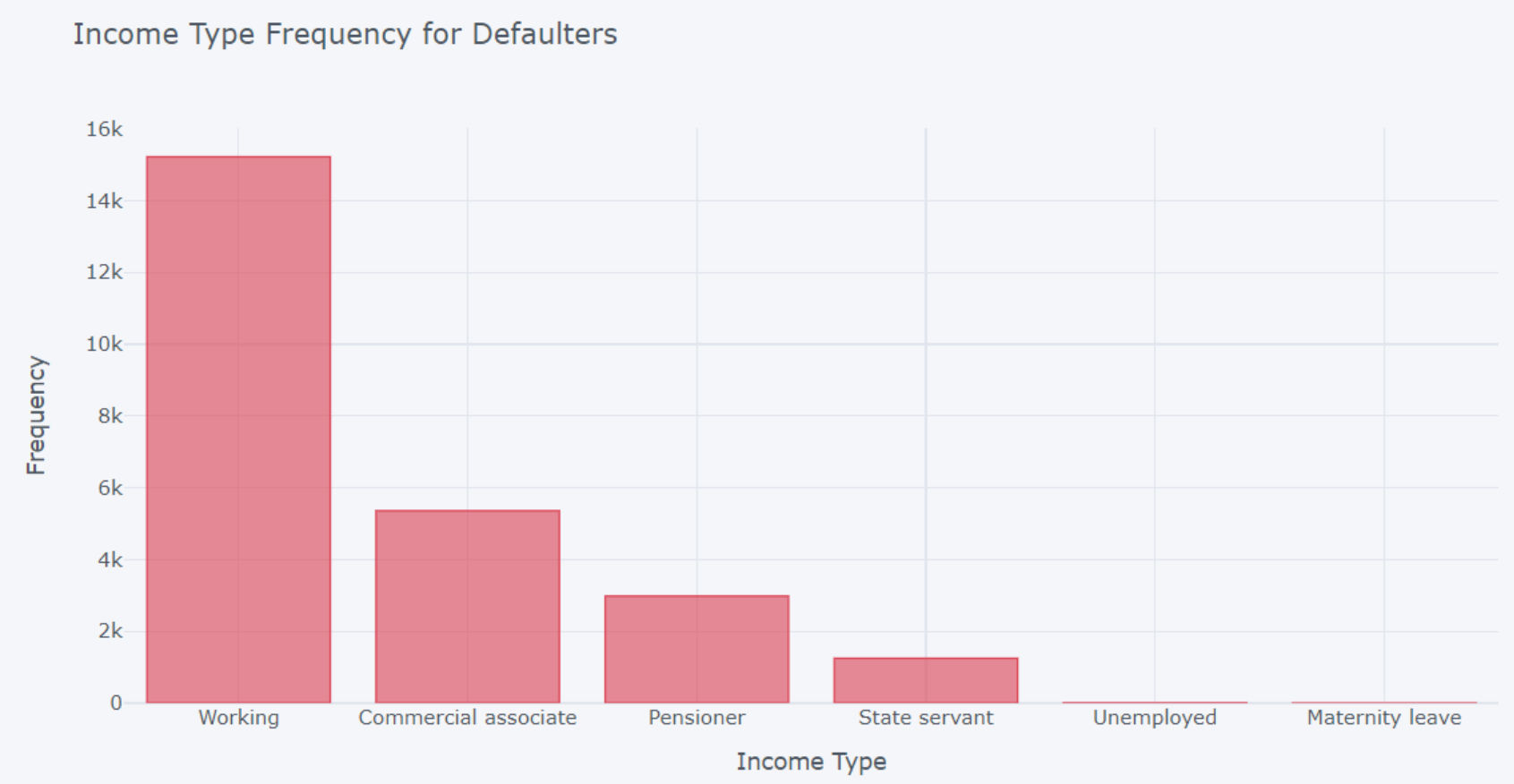
Education



- People with Secondary Education have the highest applications of loans.

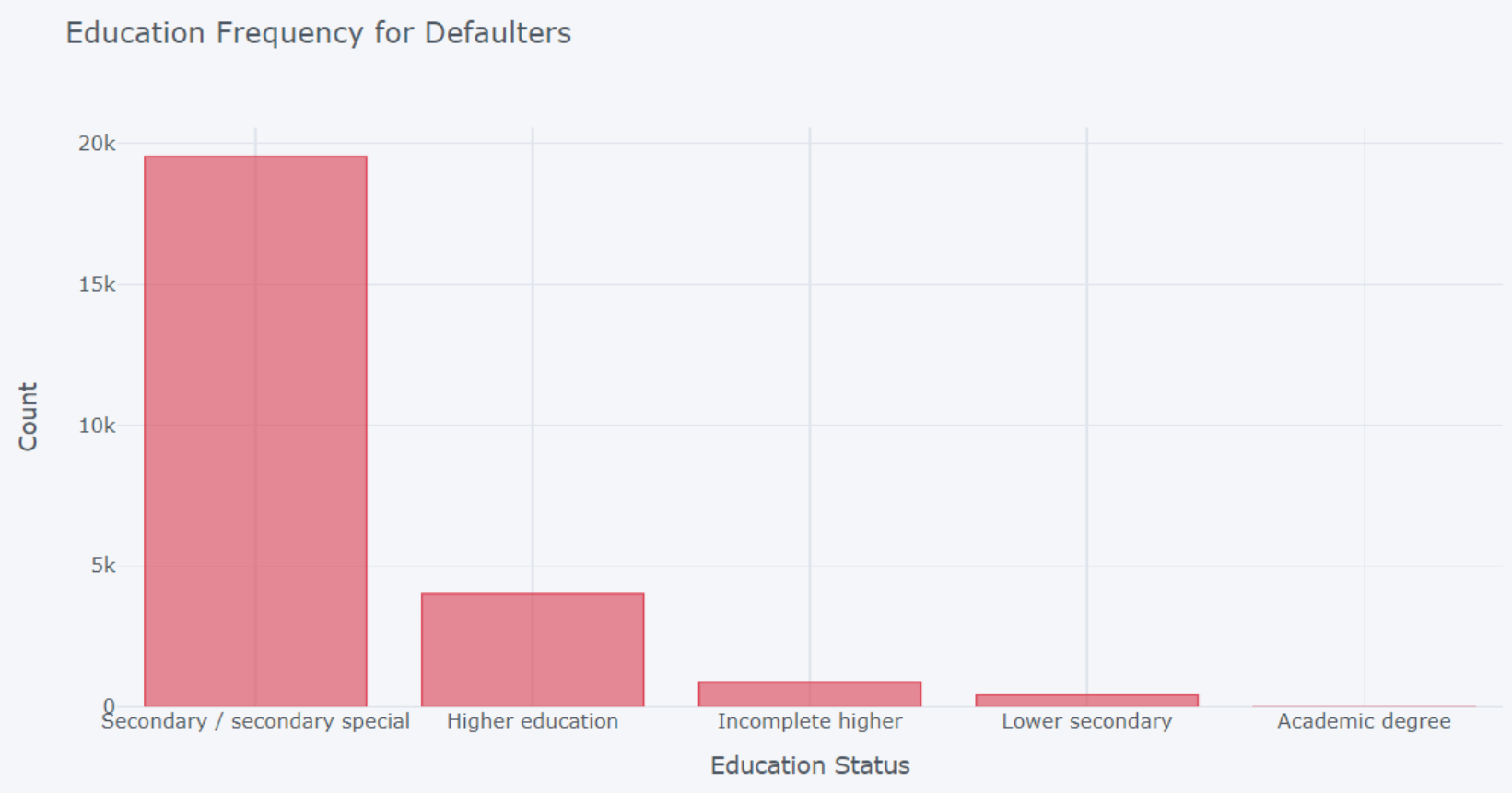
Univariate Analysis – Categorical Data On Target 1

Income Source



- Students do not default according to the data we have

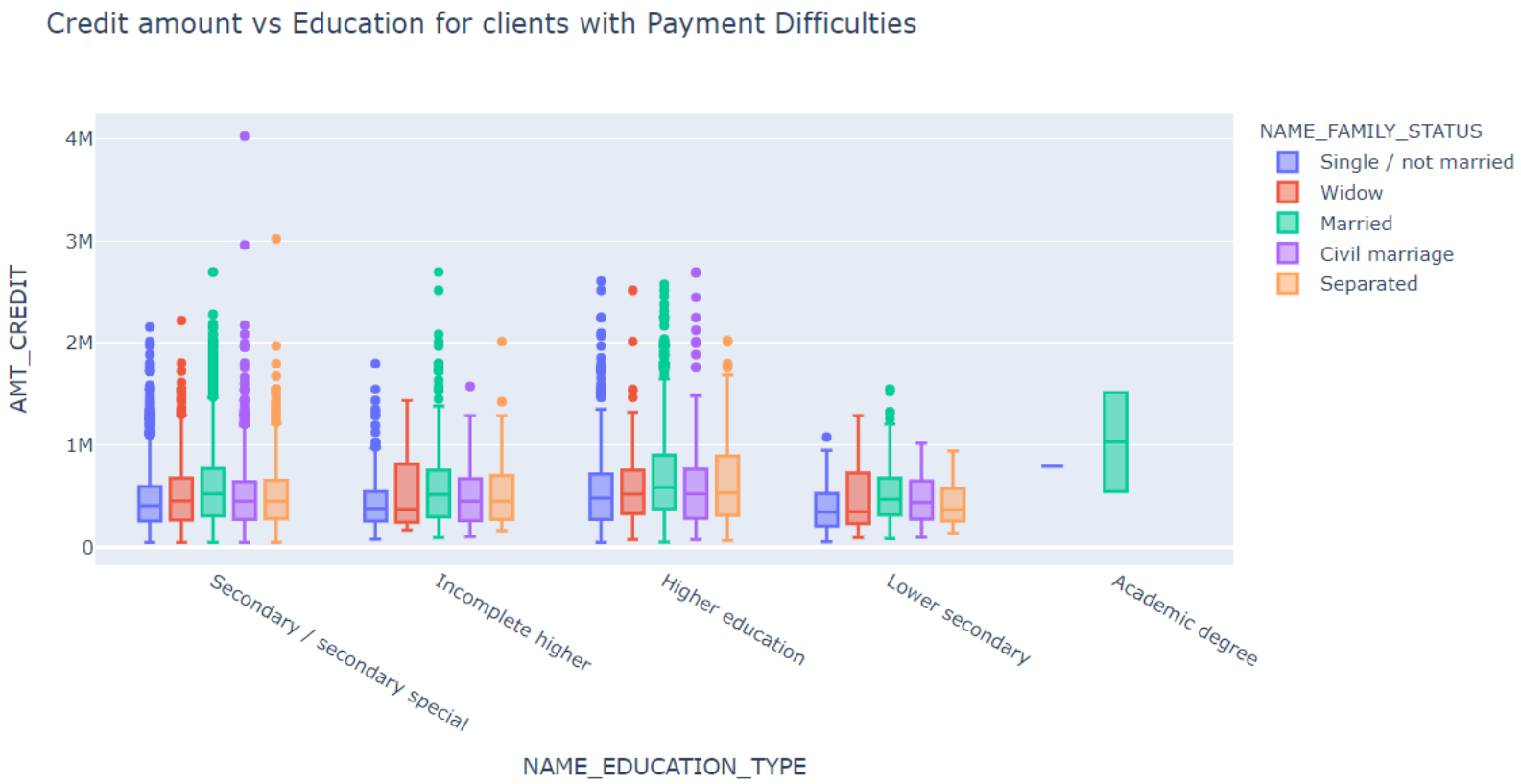
Education



- People with Higher education defaults less when compared to the non-default rate.

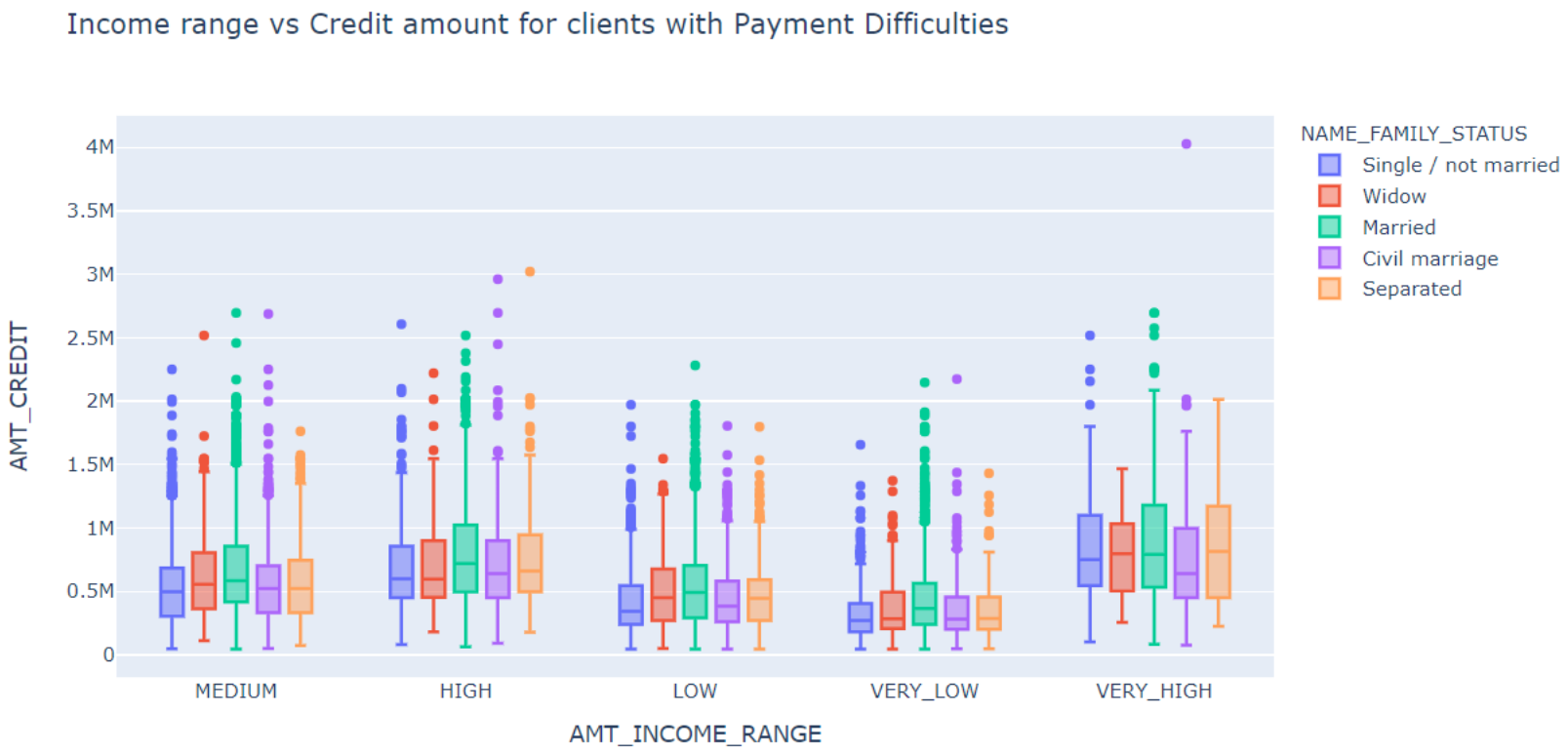
Bivariate Analysis – Categorical VS Numerical

Credit Amount VS Education



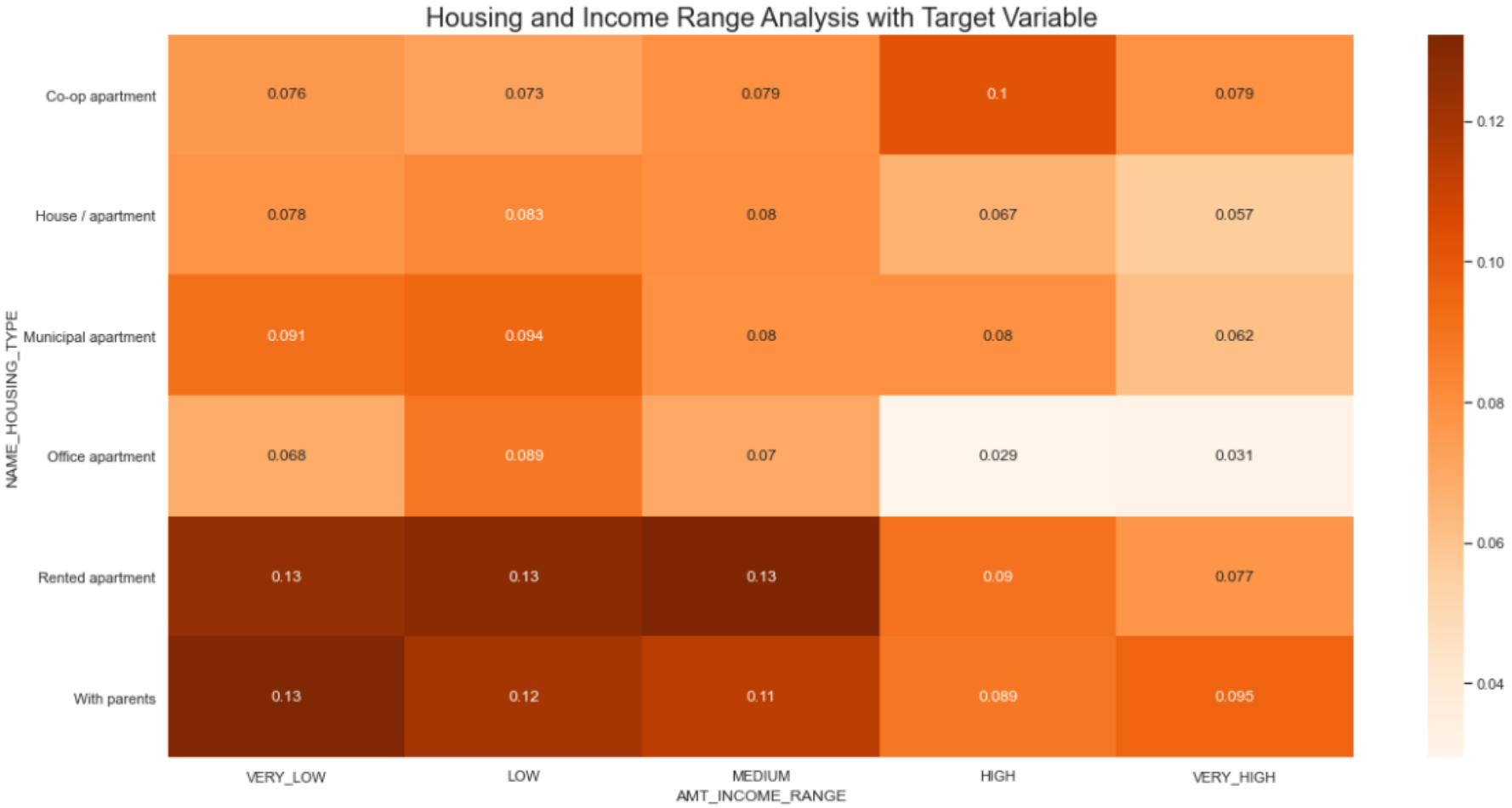
- It can be said that people who have academic degree and have status other than married are safe to provide loans.

Income Range VS Credit Amount



- It can be seen here that the people who have taken credit amount for more than 3M have less probability of defaulting for Medium, High and very high income range.

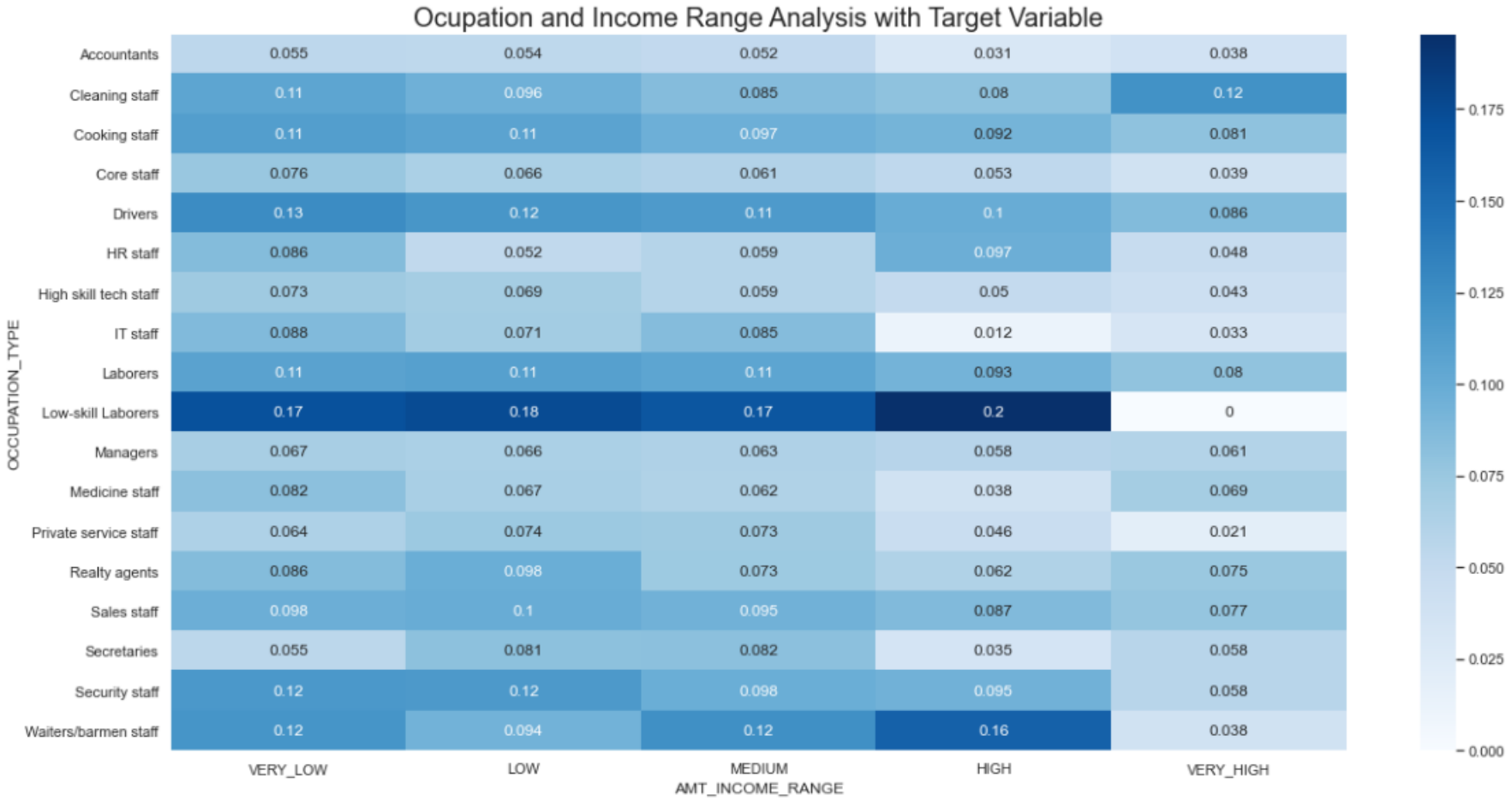
Housing – Income Range with Payment Difficulties



Housing – Income Range with Payment Difficulties : Insights

- People staying with parents or staying in rented apartments with income of medium of less have high chances of defaulting.
- People staying in office apartments and have very high income have very less chances of defaulting.

Occupation – Income Range with Payment Difficulties

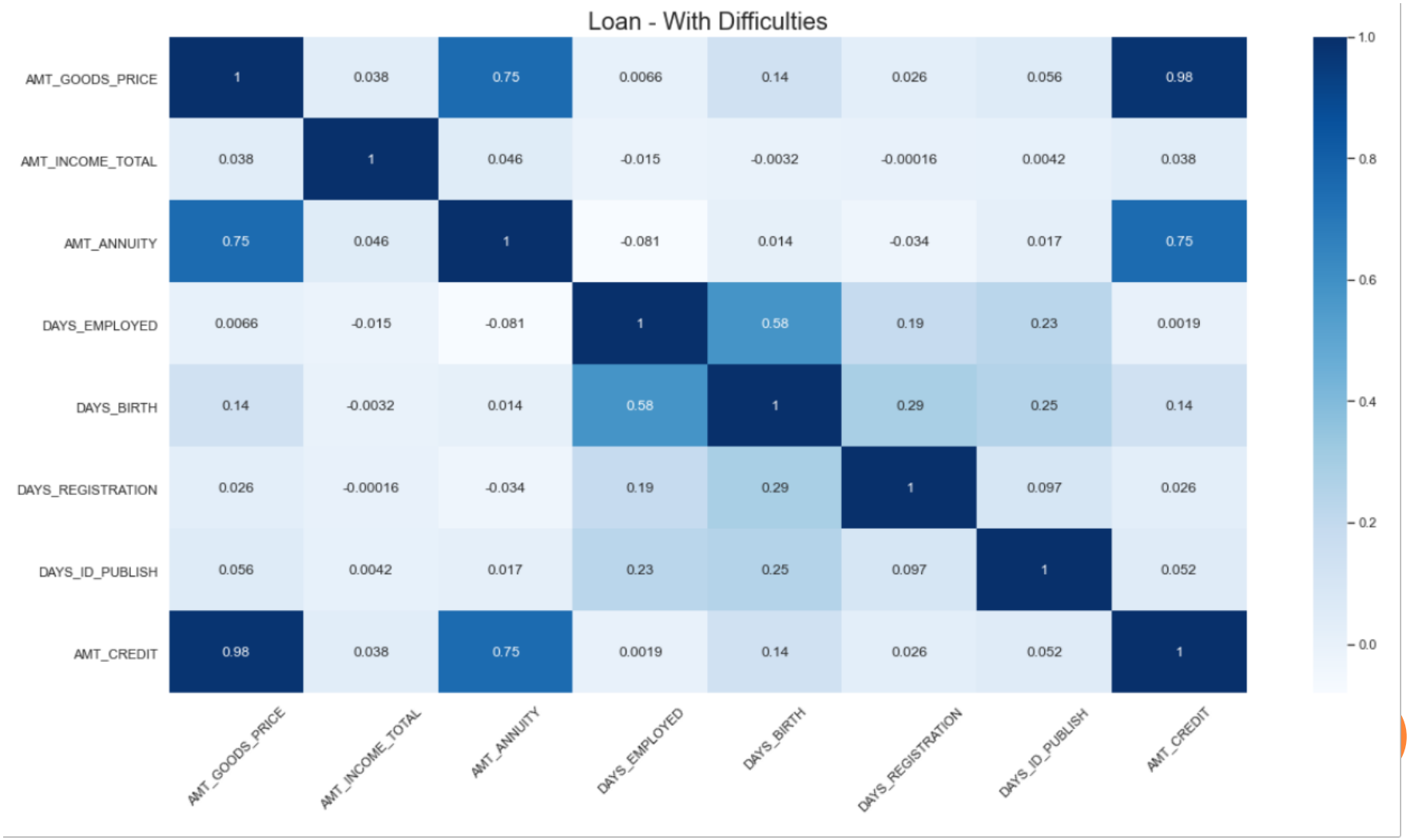


Occupation – Income Range with Payment Difficulties : Insights

- Low skill laborers have high chance of defaulting.
- IT staff with High salary income has the lowest chances of defaulting.
- Accountants overall have less chance of defaulting.

Bivariate Analysis –Numerical VS Numerical

Loan – With Payment Difficulties



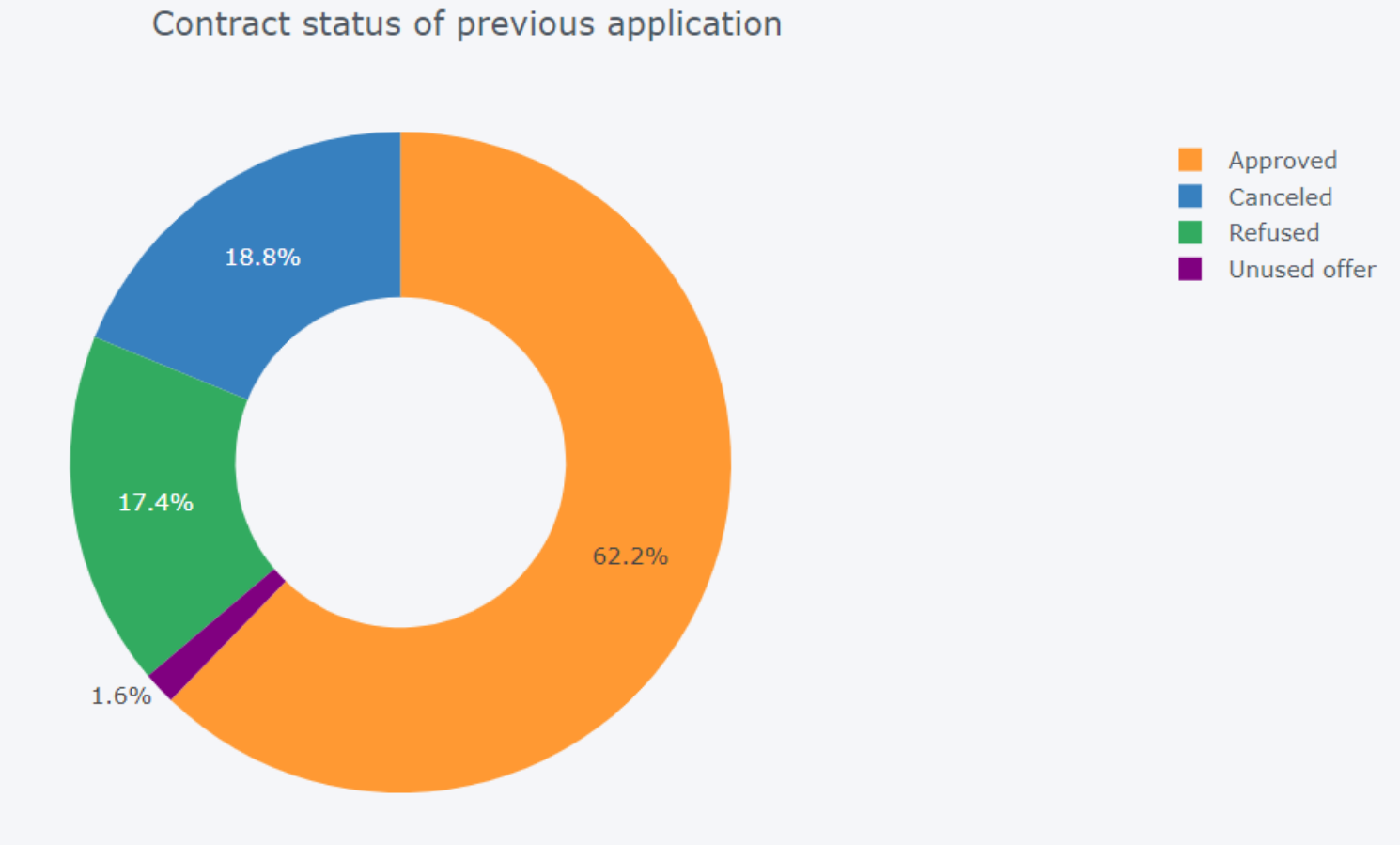
Loan – With Payment Difficulties Insights

- We have high correlation between credit price and goods price, as it can be because the loan is credited only as much as the price of the goods.

- We also have high correlation between credit price and annuity as this can be because higher the credit is given, higher will be the installments.

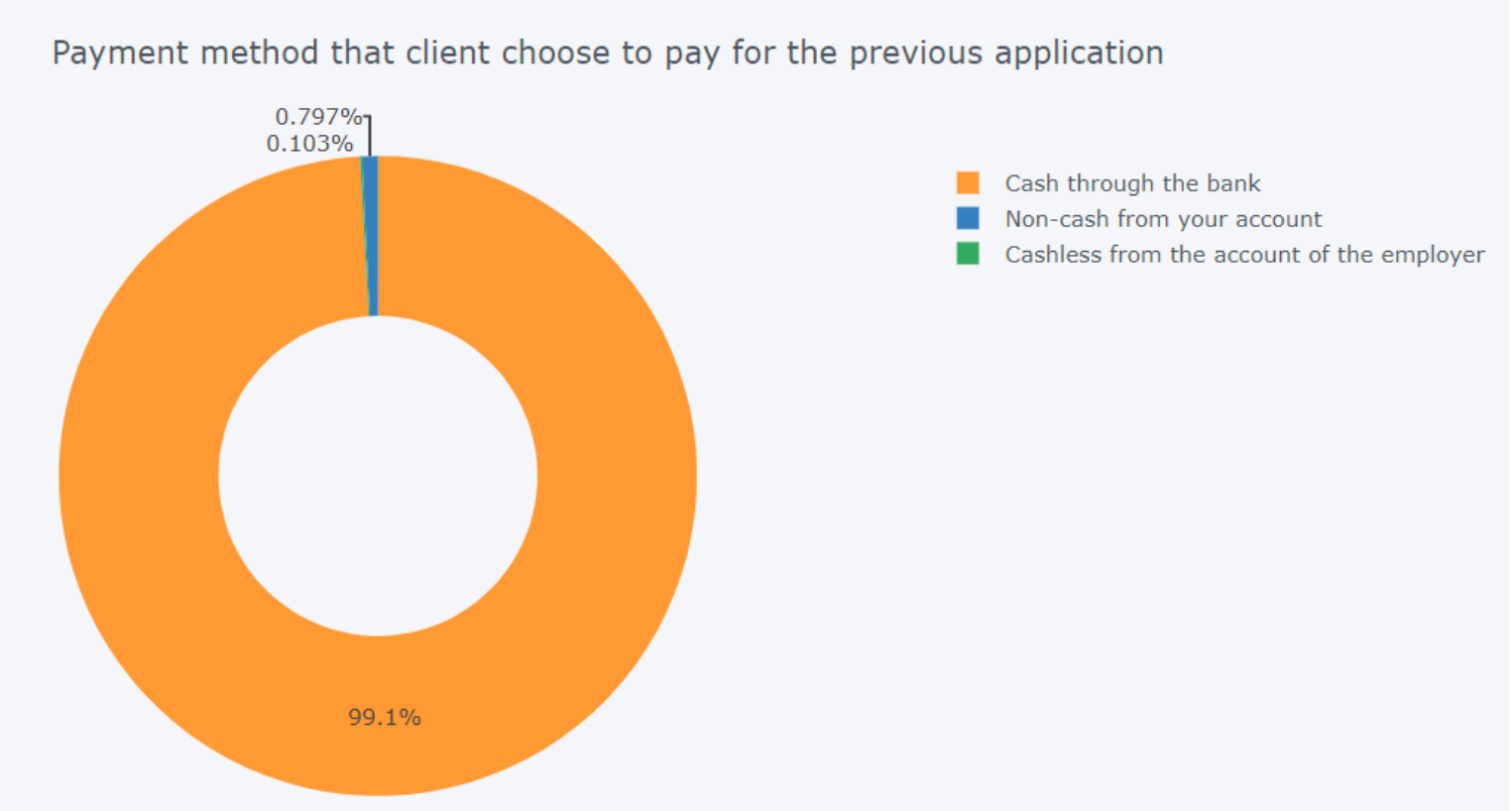
Previous Application Dataset

Contract Status



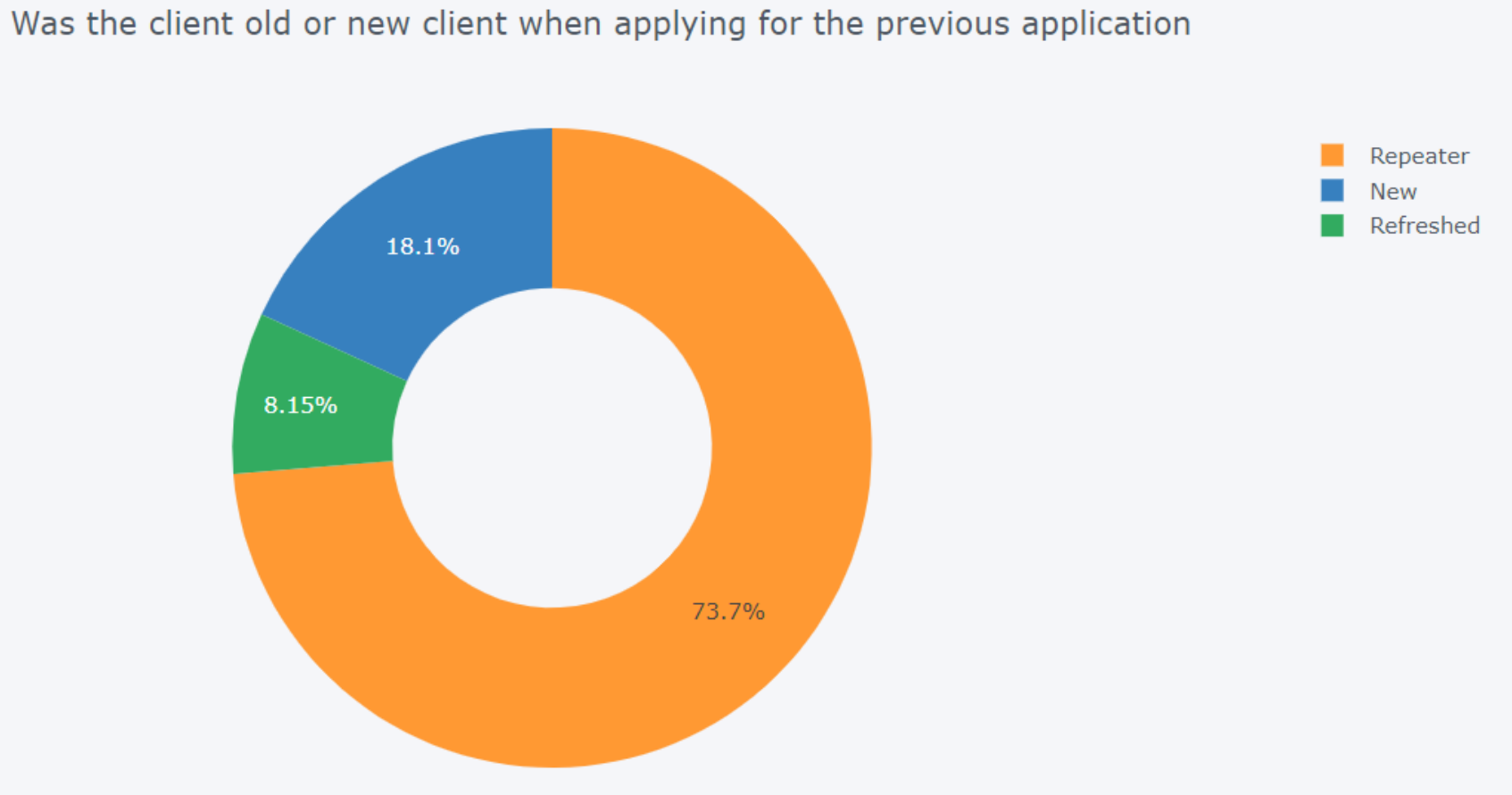
- Most of the loans do get approved.

Payment Method



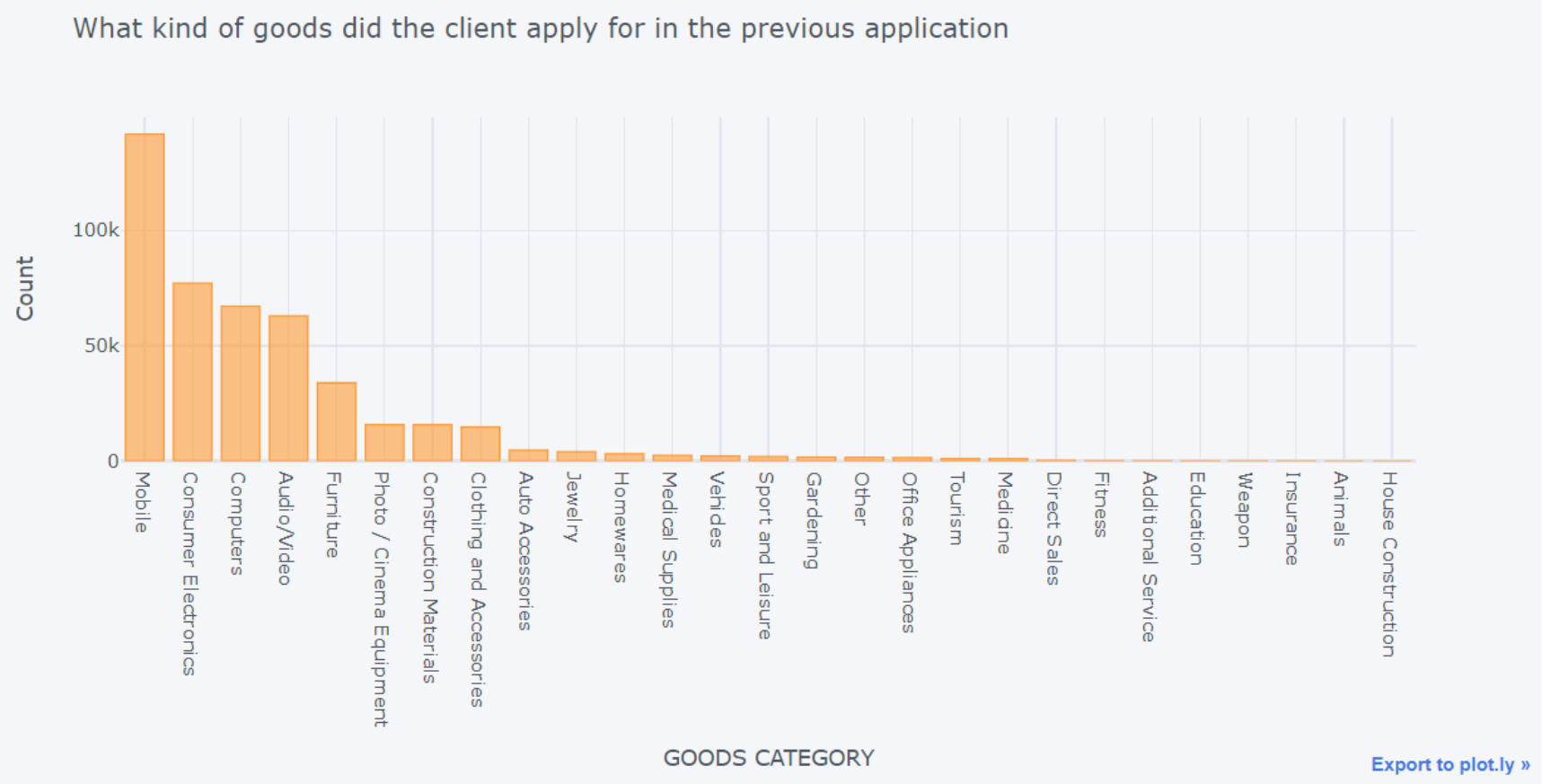
- 99% of the people uses cash through the bank for payments.

Client Type



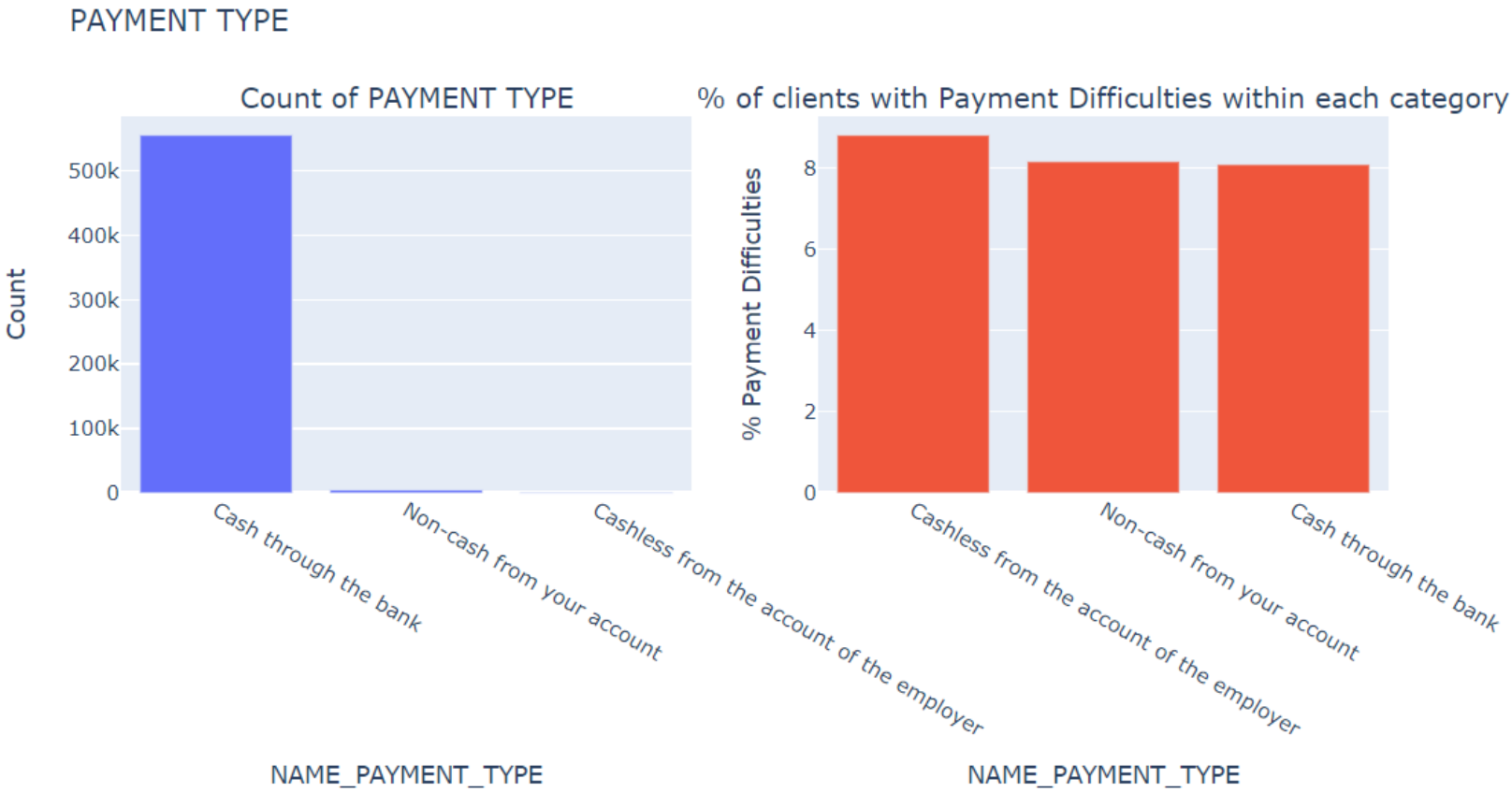
- Maximum people are the ones who have taken loan previously.

Type of Goods



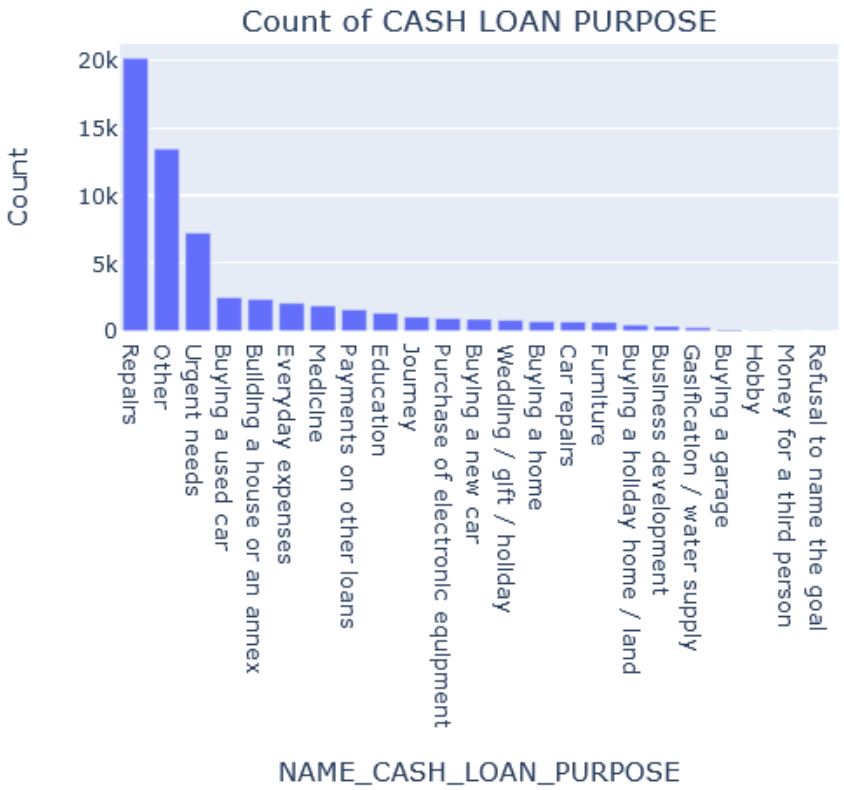
- Maximum loans are taken to buy mobile phones while least loans are taken for house construction.

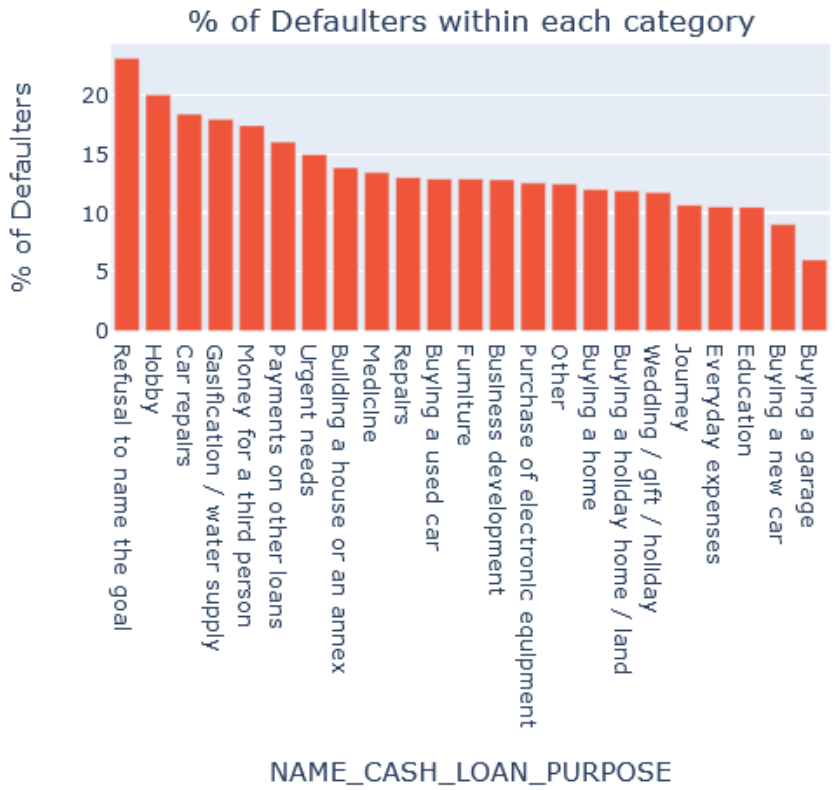
Defaulters in Payment Type



- Even though maximum payments are done by cash through bank, it has less defaults than the other two methods.

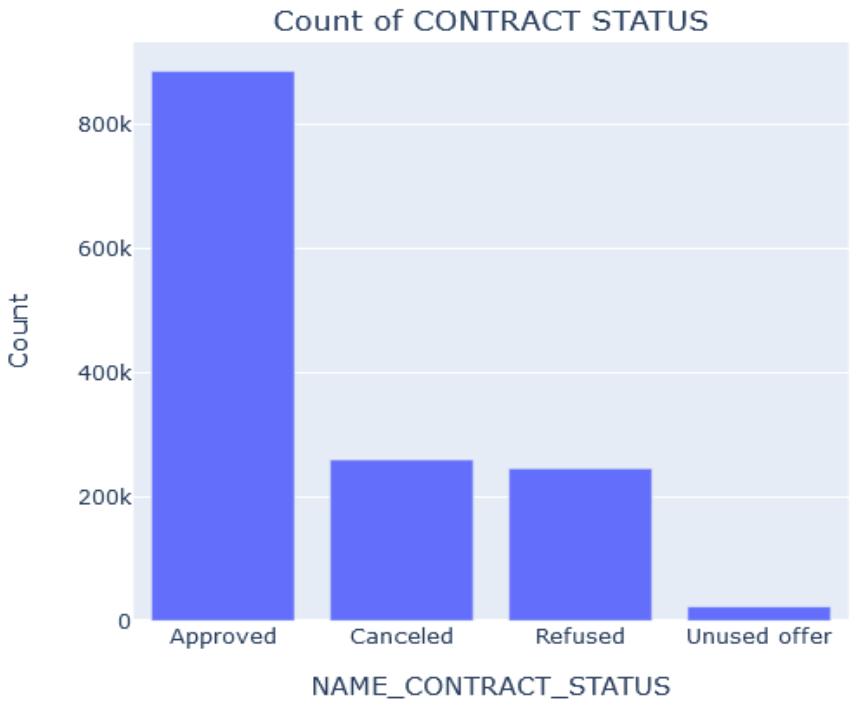
Defaulters Per Cash Loan Purpose

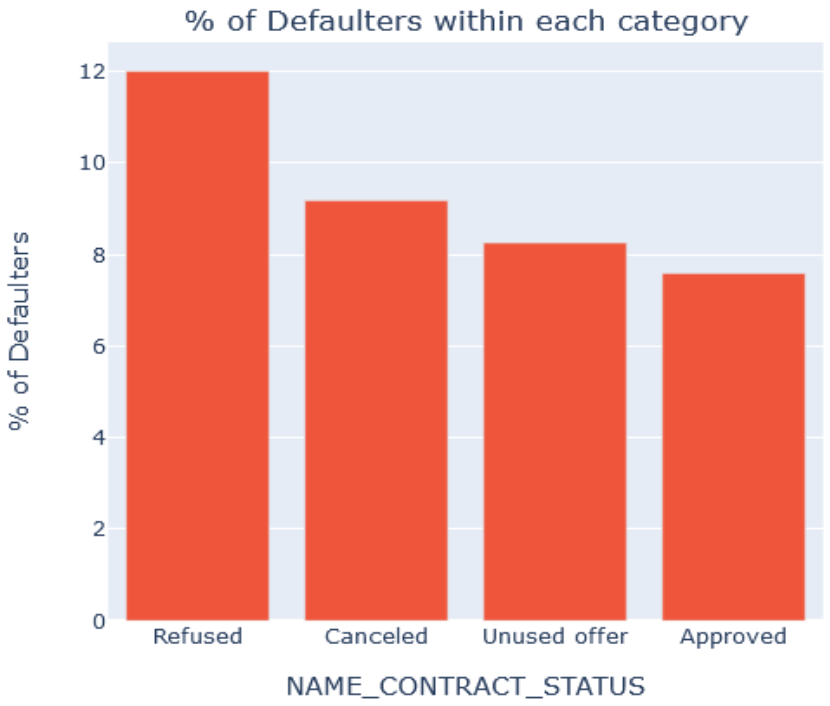




- From the first graph it can be seen that purpose of cash loan from previous data was maximum for 'Repairs'
- It can be clearly seen from the second graph that the 'Refusal to name the goal' for cash loan from previous application are the ones who have maximum % of Loan-Payment Difficulties from current application.

Defaulters Per Contract Status

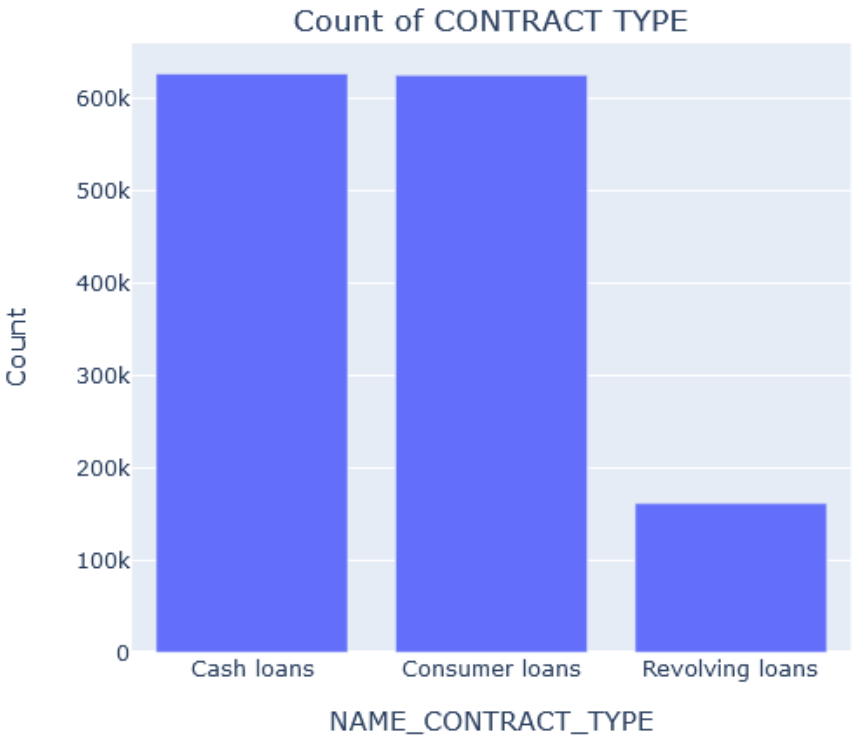


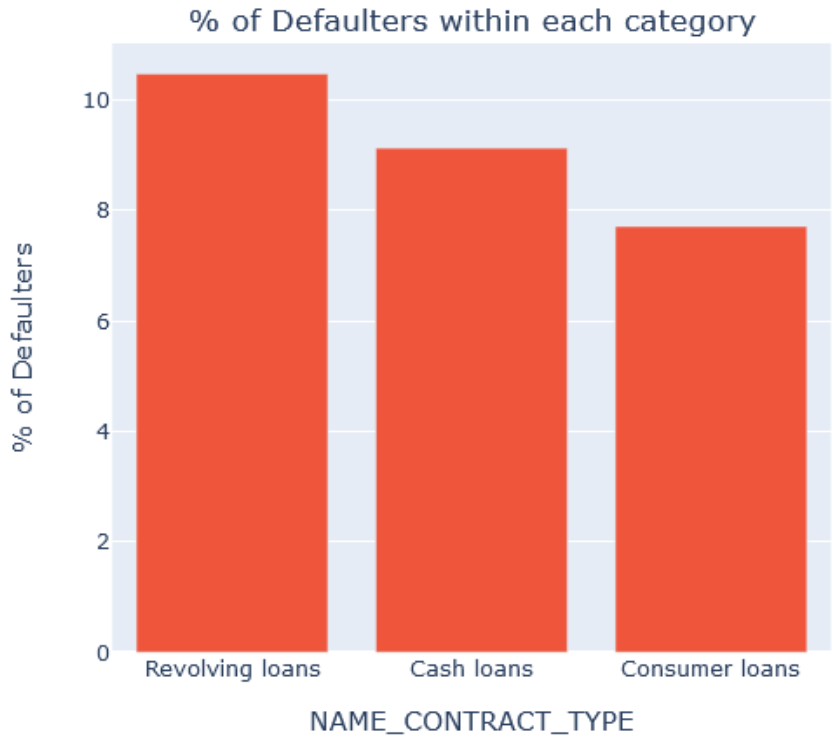


- From the first graph it can be seen that most of the contracts from previous application have been Approved.

- It can be clearly seen from the second graph that:
  - 'Refused' contracts from previous application are the ones who have maximum % of Loan-Payment Difficulties from current application.
  - Approved' contracts from previous application are the ones who have minimum % of Loan-Payment Difficulties from current application.

Defaulters Per Contract Type

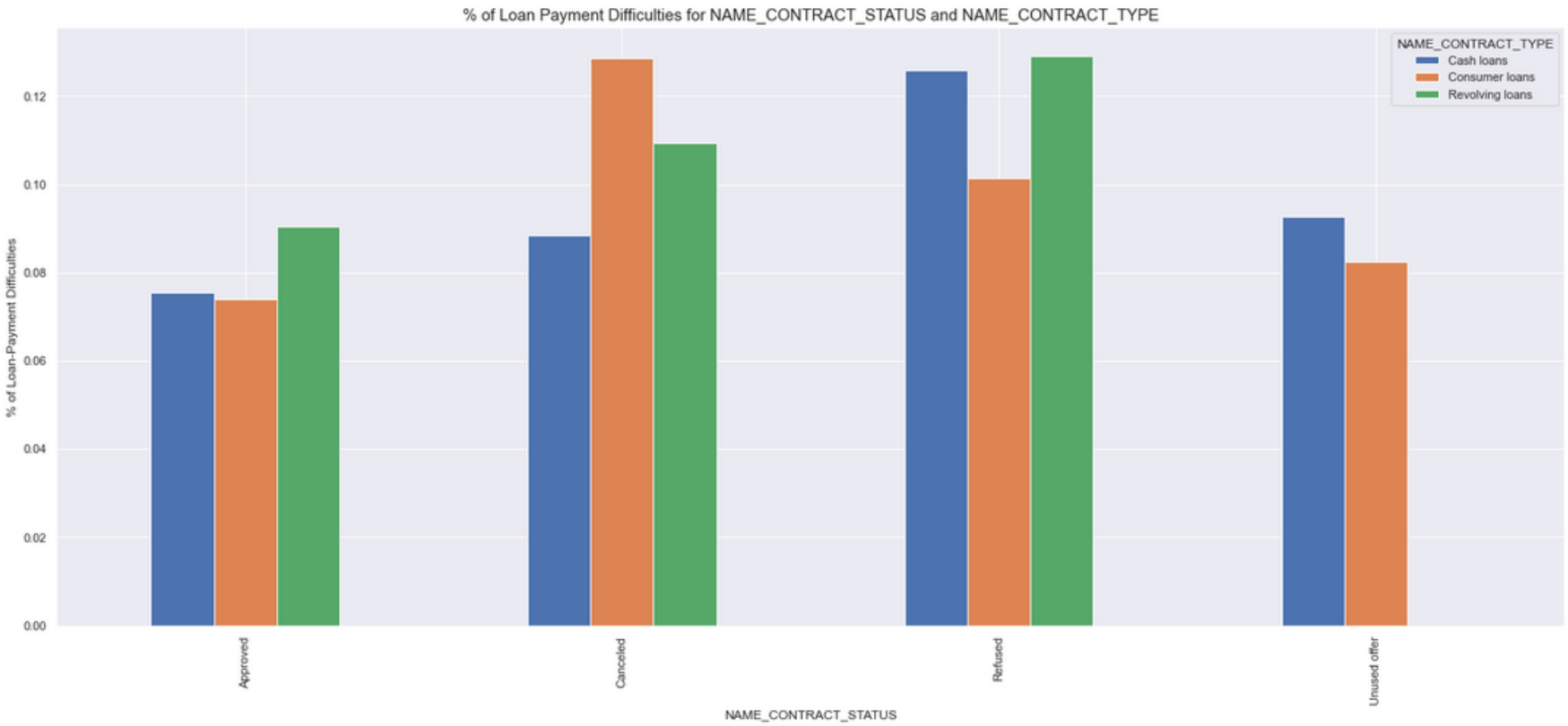




- From the first graph it can be seen that most of the contract types from previous application were 'Cash loans'

- It can be clearly seen from the second graph that:
  - 'Revolving Loans' contracts from previous application are the ones who have maximum % of Loan-Payment Difficulties from current application.
  - 'Consumer loans' contracts from previous application are the ones who have minimum % of Loan-Payment Difficulties from current application.

Defaulters Per Contract Type and Status



- It can be observed from the above graph that Clients with 'Revolving loans' and with 'Refused' previous application tend to have more % of Loan-Payment Difficulties in current application.

Major Insights – New Application Dataset

- The count of ‘Maternity Leave’ in ‘NAME_INCOME_TYPE’ is very less and it also has maximum % of payment difficulties- around 40%. Hence, client with income type as ‘Maternity leave’ are the driving factors for Loan Defaulters.

- The count of ‘Low skilled Laborers’ in ‘OCCUPATION_TYPE’ is comparatively very less and it also has maximum % of payment difficulties- around 17%. Hence, client with occupation type as ‘Low skilled Laborers’ are the driving factors for Loan Defaulters.

- The count of ‘Lower Secondary’ in ‘NAME_EDUCATION_TYPE’ is comparatively very less and it also has maximum % of payment difficulties- around 11%. Hence, client with education type as ‘Lower Secondary’ are the driving factors for Loan Defaulters.

Major Insights –Previous Application Dataset

- The count of ‘Refusal to name the goal’ in ‘NAME_CASH_LOAN_PURPOSE’ is comparatively very less and it also has maximum % of payment difficulties- around 23%. Hence, clients who have ‘Refused to name the goal’ for cash loan in previous application are the driving factors for Loan Defaulters.

- The count of ‘Refused’ in ‘NAME_CONTRACT_STATUS’ is comparatively less and it also has maximum % of payment difficulties- around 12%. Hence, client with contract status as ‘Refused’ in previous application are the driving factors for Loan Defaulters.

- The count of ‘Revolving Loans’ in ‘NAME_CONTRACT_TYPE’ is comparatively very less and it also has maximum % of payment difficulties- around 10%. Hence, client with contract type as ‘Revolving loans’ in previous application are the driving factors for Loan Defaulters.

- Clients with 'Revolving loans' and with 'Refused' previous application tend to have more % of payment difficulties in current application. Since the count of both 'Revolving loans' and ‘Refused’ is comparatively less(from the graphs in previous slide), clients with ‘Revolving Loans’ and ‘Refused’ previous application are driving factors for Loan Defaulters.

Thank You!