Credit EDA Case Study

SUBMITTED BY:

PRASAD SANA

INDUSHREE. L

Problem statement

In this case study we have to make analysis about the driving factors for the company to decide for loan approval based on the applicant’s profile.

- 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.

Importing the Datasets and Routine Data checks

1. Imported Application dataset, it consists of around 3L rows and 122 columns.
2. For the ease of analysis we have taken the random sample of 1000 rows.
3. We used pd.sample() to take the sample and used "random_state" parameter for reproducibility.
4. Used pd.info() to check the column and respective Dtypes.

Missing Value checks and Data type conversions

1. Calculated missing value percentage for each columns
2. We dropped the columns having missing value percentage more than 50 % (as those column wont give us much info)
3. For the columns who has missing values of 15% or less, we have suggested the best imputation method based on the data in that column.
4. NAME_TYPE_SUITE, AMT_REQ_CREDIT_BUREAU_HOUR etc. are few of the examples we have considered for missing value analysis.

Data Type Conversions:

1. Converted all the flag variables to object types as they have only limited values (ex: 0 and 1)
2. Corrected the datatypes for other columns as well
3. Eg: CNT_FAM_MEMBERS – From float to int

Outlier Analysis

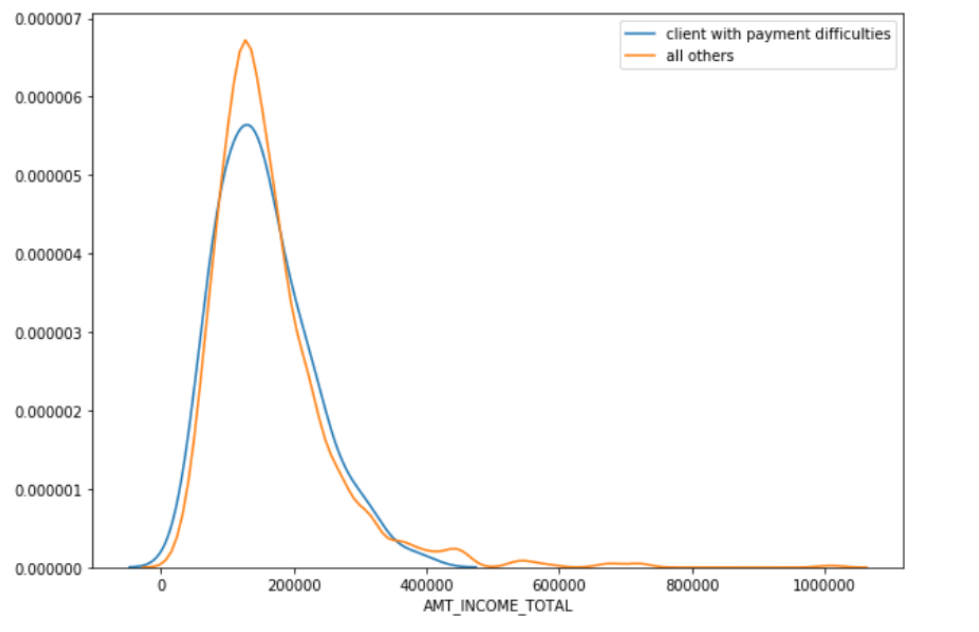
1. We did outlier analysis for few of the columns and suggested best methods for the outlier treatment.
2. **DAYS_EMPLOYED** provides the data about How many days before the application the person started current employment
   From the outlier analysis we found that the data has outlier values after 80th percentile.
   "DAYS_EMPLOYED" data is converted into the values in year for the ease of anlysis and found that the data above 30 years are all same and not valid.
   Hence it is suggested to replace this invalid data with np.NAN values
3. For AMT_INCOME_TOTAL column we binned the income in to 6 categories as <1L, 1L to 2L, ..., >5L

Correlation Matrix

1. We have divided the data into two dataframes based on target variable (0 and 1)
2. Plotted the correlation matrix and identified the top correlated variables in both the data frames
3. Choosed a few columns for Univariate numerical variables out of those top correlated variables
4. Ex :AMT_CREDIT, DEF_30_CNT_SOCIAL_CIRCLE etc

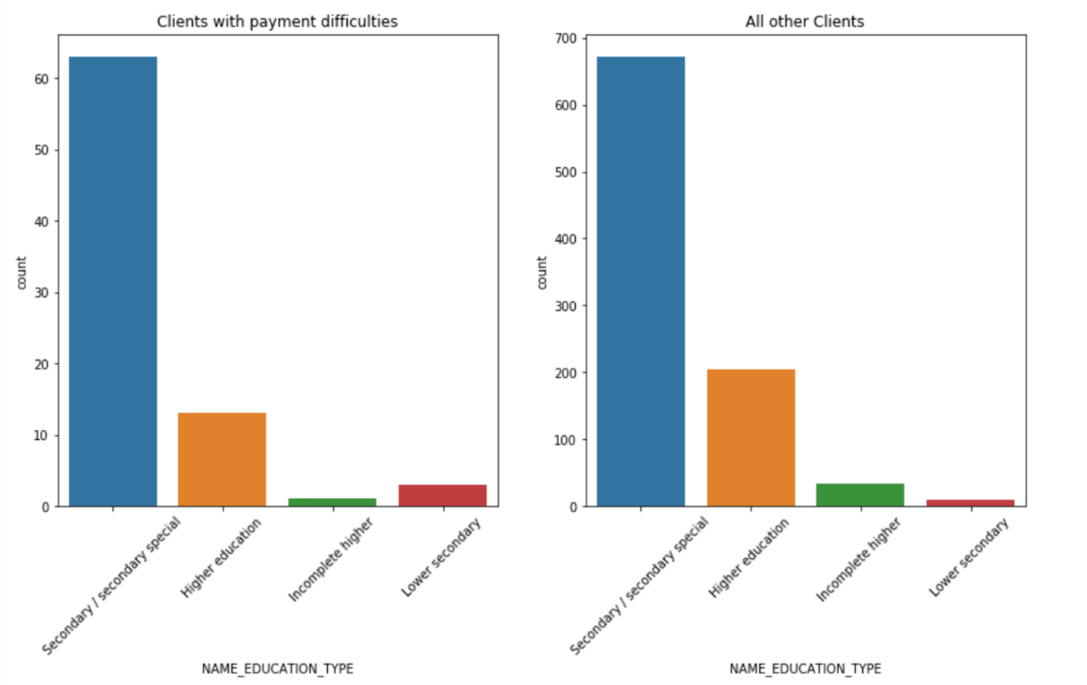
Univariate Analysis

Continuous



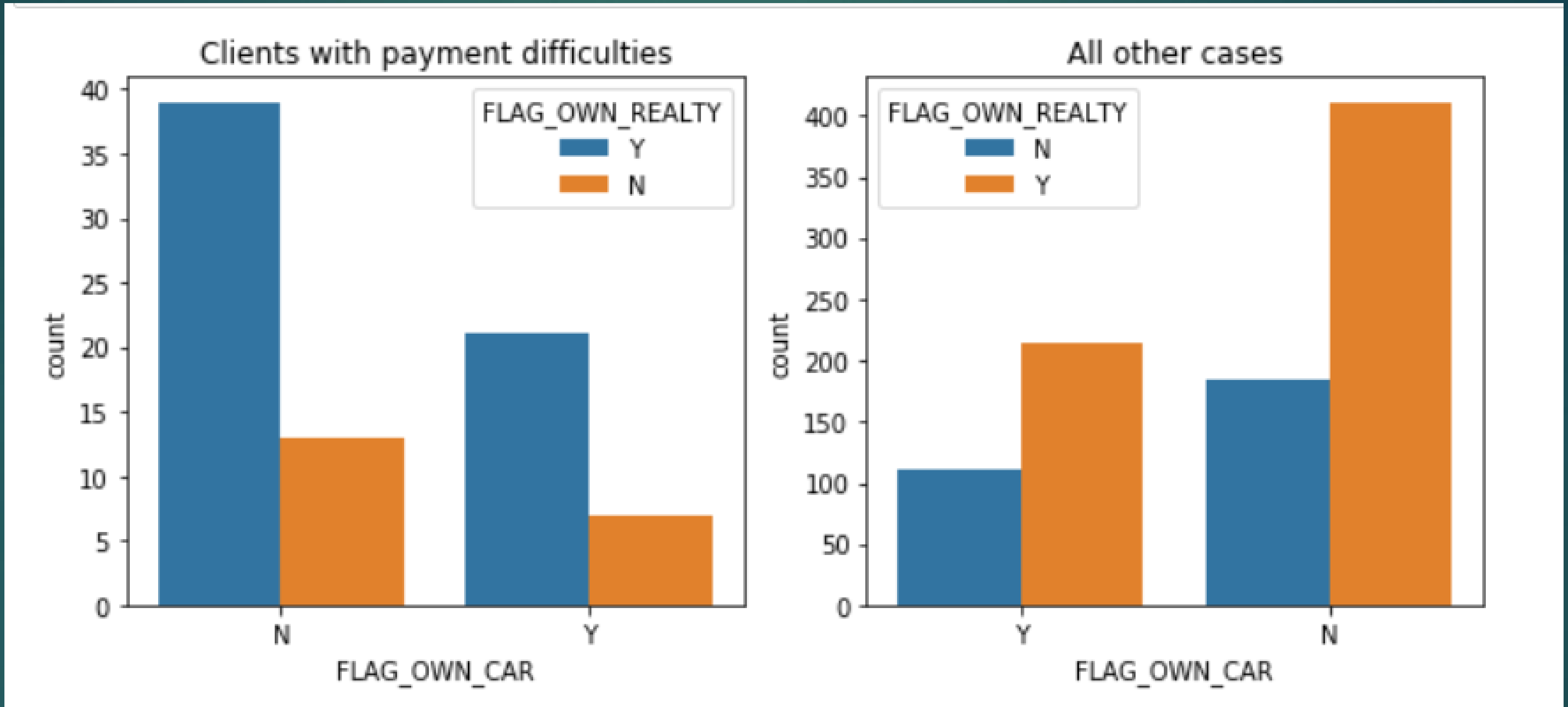
All the applicants with higher income dont have any difficulty in repaying the loan, applicants with income below 4L are facing difficulty in repaying the loan amount

Categorical



Among the applicants with/without payment difficulties, the applicants whose education level is "Secondary/secondary special" are more in number compared to other education level people

Bivariate – Categorical Categorical

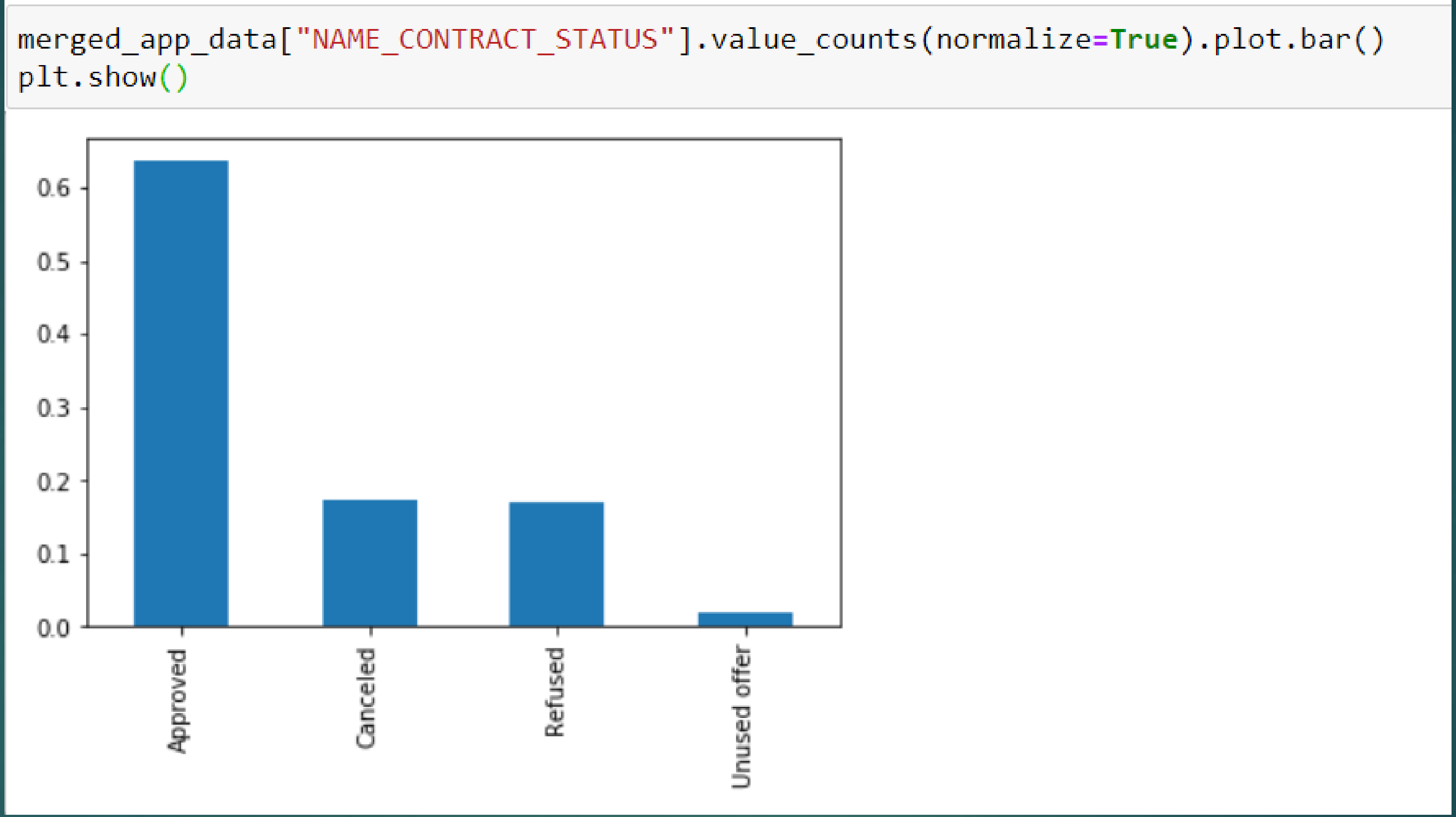


Clients who don’t own a house and a own car are more likely to be defaulters i.e, they are clients with payment difficulties.
In clients with payment difficulties category, people with own house and car are very less.

Merged Data

1. Loaded the “previous_application.csv” file
2. Merged both the application and previous application data based on the “SK_ID_CURR” column(Loan ID column)
3. Doing merging this way, gives the previous loan applications of the same customer

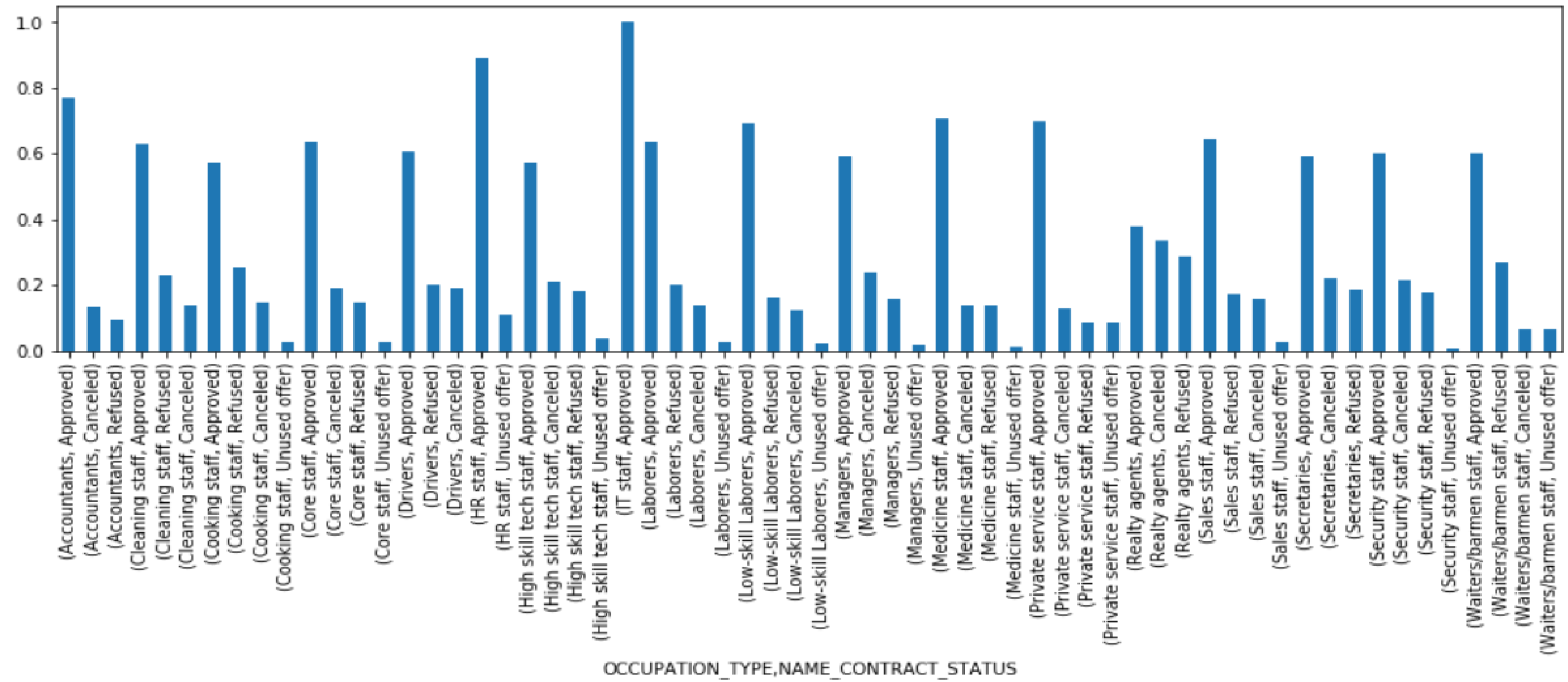
Univariate Analysis



```
merged_app_data["NAME_CONTRACT_STATUS"].value_counts(normalize=True).plot.bar()
plt.show()
```

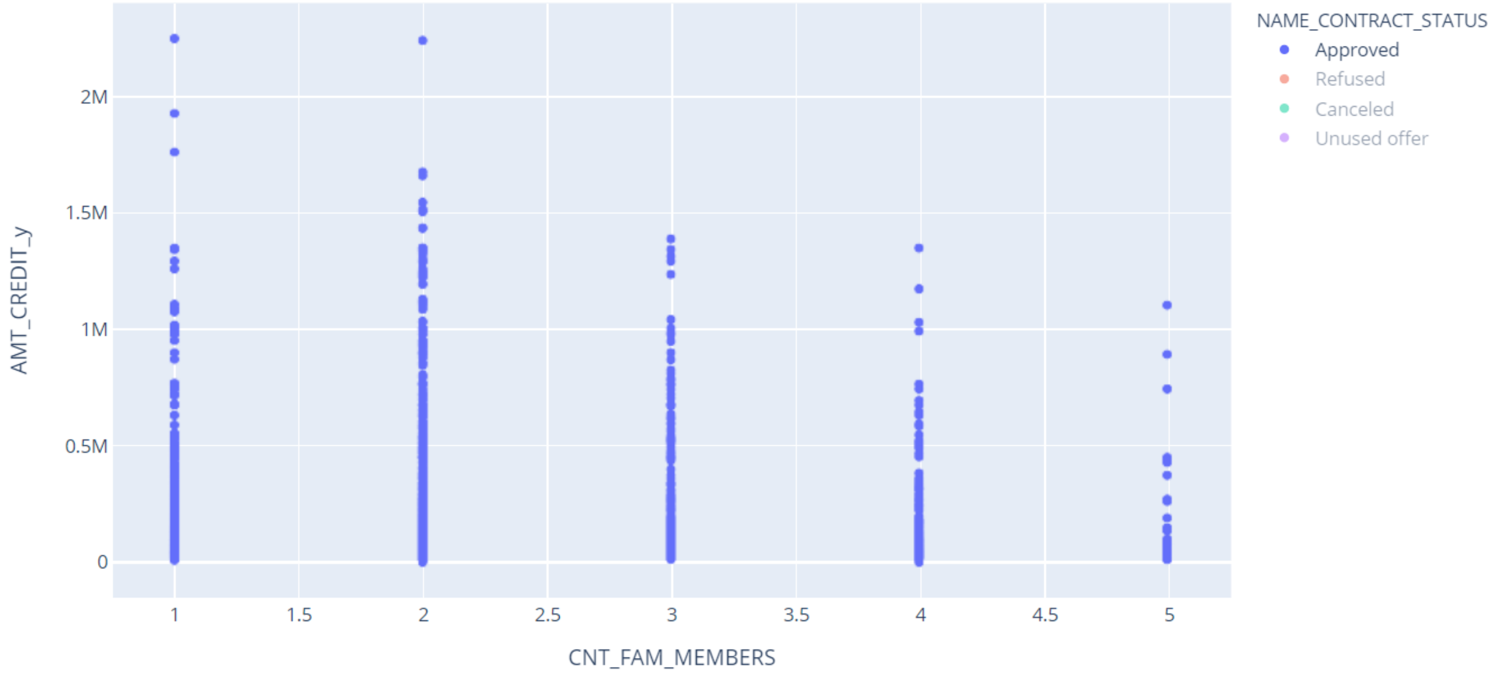
Out of all the applications, it can be seen that only around 60% of the applications are *approved*. Rest of the applications are either *cancelled* or *rejected*

Bivariate Analysis



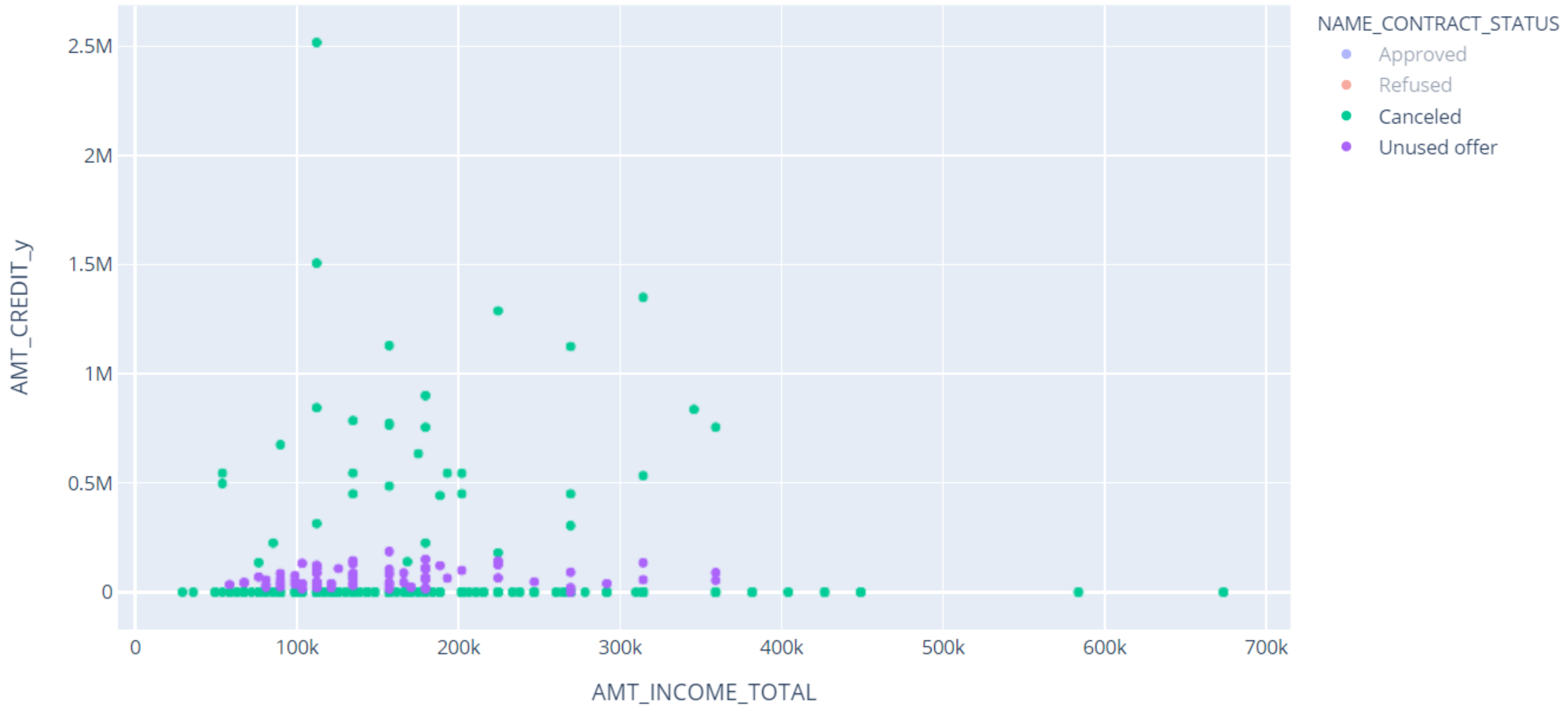
We observe that clients who work as It staff, HR staff, Accountants, Medicine staff have more approval rate. Particularly, IT staff have 100% approval In general, we can say clients in IT industry have more approval rate

Bivariate Analysis



As the count of family members increases, amount credit is less i.e, less amount is approved for the application which has more family members

Bivariate Analysis



- Loan applications with credit amount below 3 Lakhs are more approved irrespective of the annual income.
- Loan applications are more refused for clients with annual income below 4 Lakhs i.e, loans for clients with high income is easily approved
- Clients with less income(less than 3 Lakhs) are cancelling their loans more frequently compared with high income clients

Summary of Observations

- Most of the loan applicants who loans status is “Approved” have an education level **Secondary/Secondary Special** and are **working** with **income range less than 2 Lakhs** and belong to the **age group 30 – 40**
- We observe that clients who work as It staff, HR staff, Accountants, Medicine staff have more approval rate. Particularly, IT staff have 100% approval In general, we can say clients in IT industry have more approval rate
- As the count of family members increases, amount credit is less i.e, less amount is approved for the application which has more family members
- Approval rate is less for clients, If the number of defaulters(60 DPD) in their social circle is more. Even for the loan approved cases, the credit amount is less(below 5 Lakhs observed from the plot)
- Most of the loans rejected have rejection reason as HC(Higher Credit) - around 60%
- Refusal rate is more for clients who already took a loan 1 or 2 years ago. (If previous loan is taken 1 or 2 years ago, it is more likely that the loan is still active and the client is applying for another loan)

Thank You