Credit EDA Assignment

By Gopalakrishanan Narayanan – EDS22020105

Data Cleaning

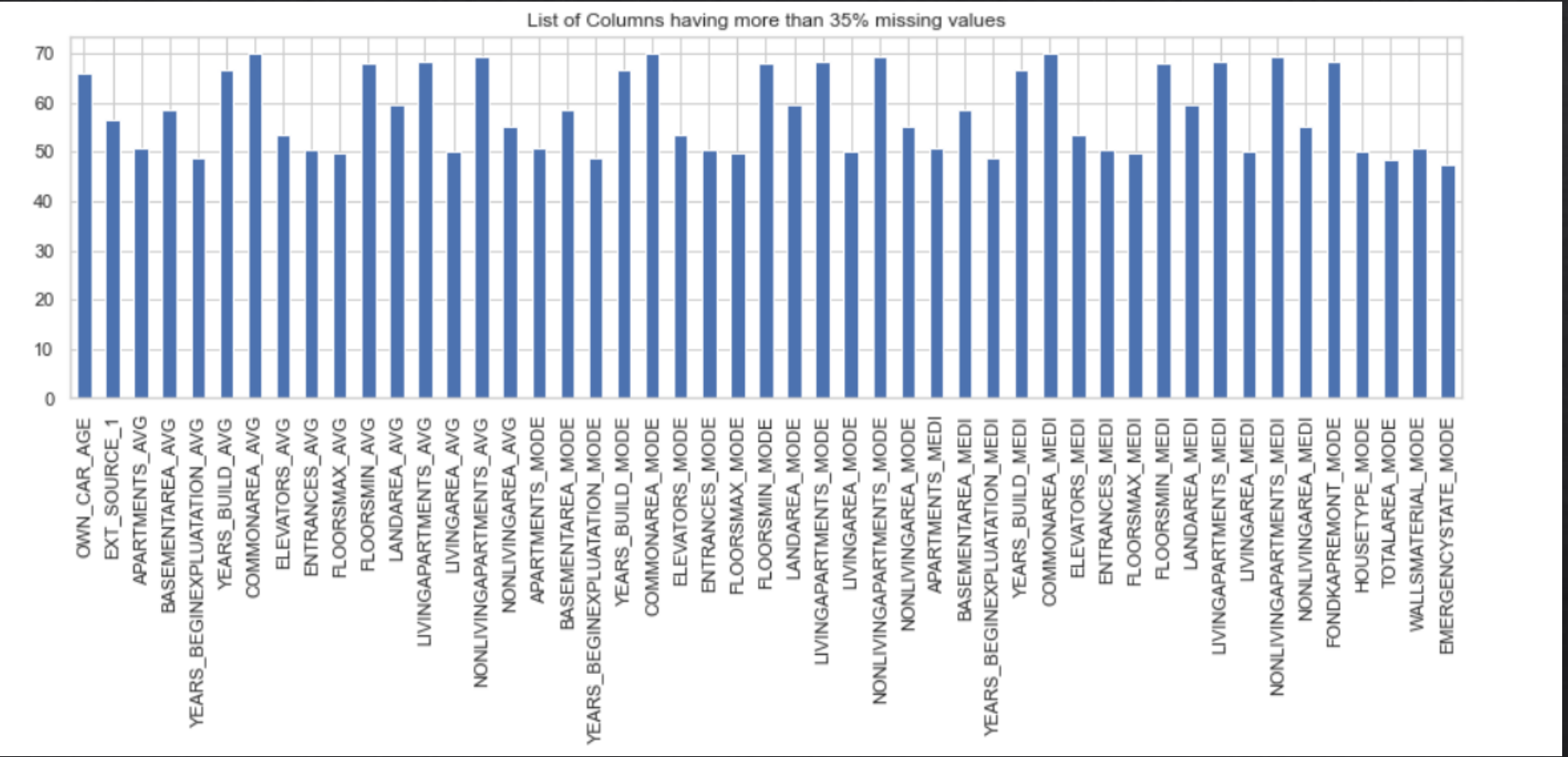
- After reading the dataset we can see that there are 122 columns in the dataframe.
- We can see that many of the columns have missing values.
- I have set the cut-off for missing values at 35%.
- Selecting and dropping all the columns having missing values above 35%.
- 49 columns where there in the dataframe with more than 35% missing values were dropped.

Introduction and objectives

- This case study aims to give us an idea of applying EDA in a real business scenario. In this we will do risk analysis of loan process.

- This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

Graphical presentation of columns with more than 35% missing values



Data Imputing

- After dropping the column the data frame was separated with columns having categorical data and numerical data.
- Few columns in both category have missing values.
- Depending on the data we will impute them with mean, median or mode values.

Categorical Columns

- There are 12 categorical columns if that only 2 have missing value NAME_TYPE_SUITE and OCCUPATION_TYPE column.
- NAME_TYPE_SUITE for this column , Imputing the missing with mode value as the percentage of mode value in the column will not change much.
- OCCUPATION_TYPE column has huge no of missing values(96391) than compared the mode value(55186) in column.
  - If we impute them with mode value it will skew the data set.
  - So the missing values are assigned as a separate category as “Unknown”

Numerical Columns

- When we check the list of numerical columns we see that there are 68 numerical columns. Of these not all 68 have missing columns, so we procced to handle those that have missing values.
- For the columns : AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_YEAR. For these columns, we can use mode to replace the missing values as median and mode are the same. Hence we replace the missing values

Numerical Columns

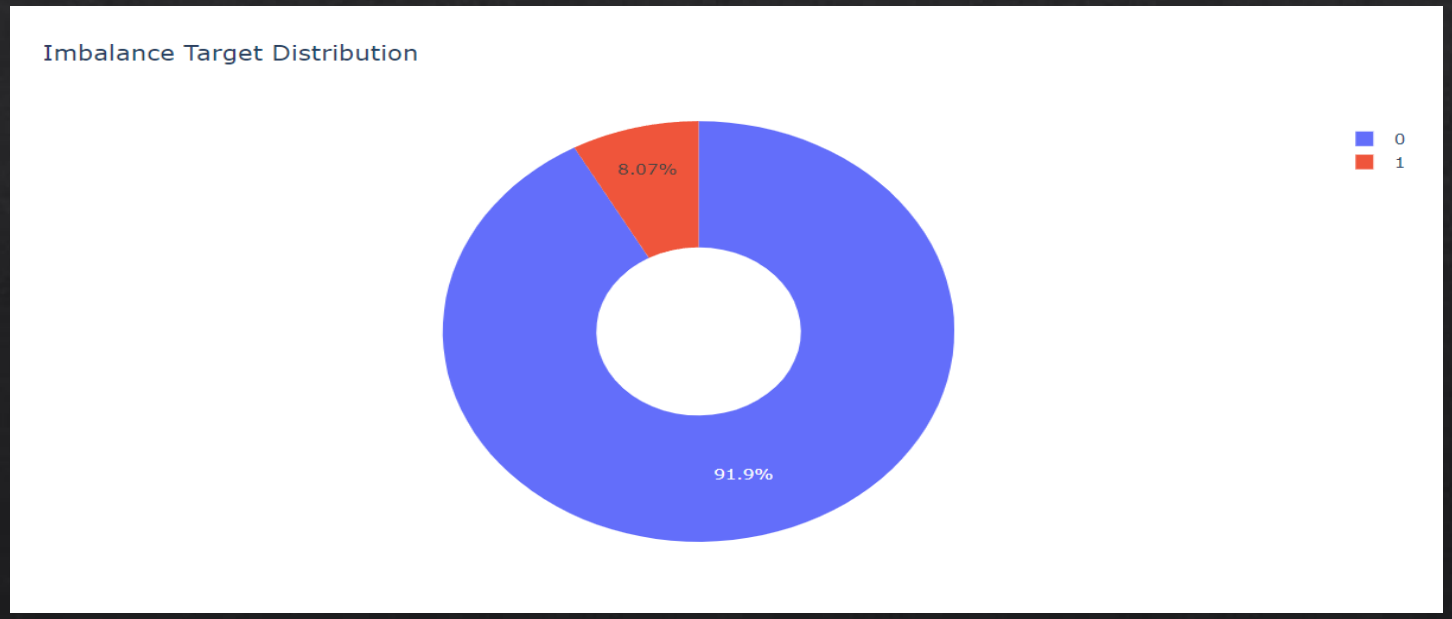
- For other missing data in numerical columns So we will be replacing missing data with the help of median.
- Once the missing data has been imputed , we recheck columns values again to ensure that no missing data remains.
- After this only Binning of columns is remaining in data cleaning

Data Binning

- If we look at the columns that denote date of some form, we see that the dates in most of them are in the form of days and are negative.
- So we shall we converting all dates to positve using the absolute method. Also we will need the dates in 365 days format so we also use floor divison to convert them.
- Further we bin few columns to categorical so that it is easier during analysis
- DAYS_BIRTH_BINS, AMT_CREDIT_BIN columns are binned into categories using pd.qcut this function
- With this the dataset is clean.

Data imbalance

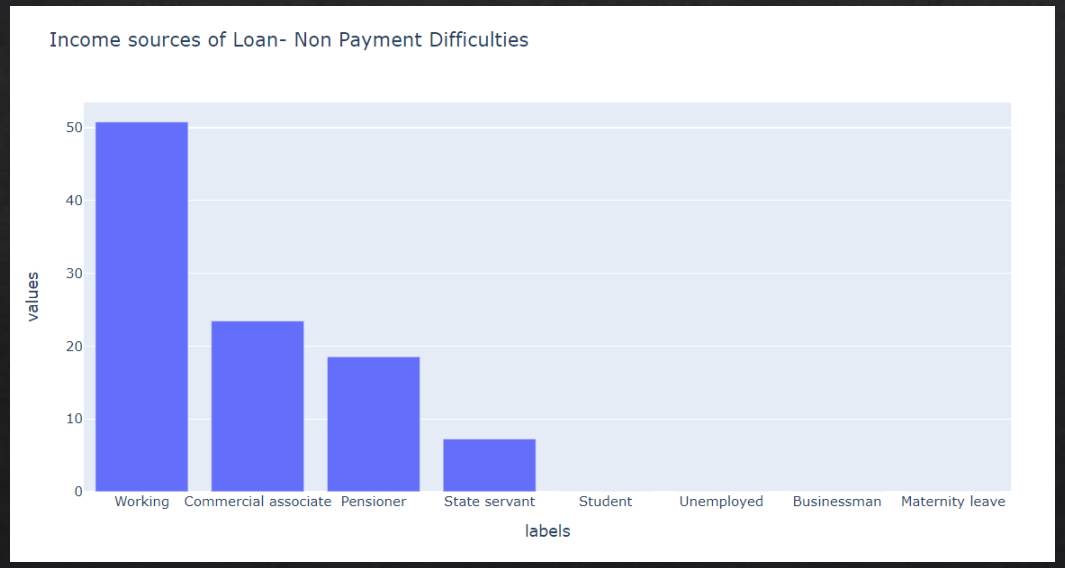
- We can see that TARGET column has a huge Data imbalance favouring 1 side.
- Hence we wil be splitting them into Target_0(Loan Non-Payement) and Target_1(Loan payment) data frames.

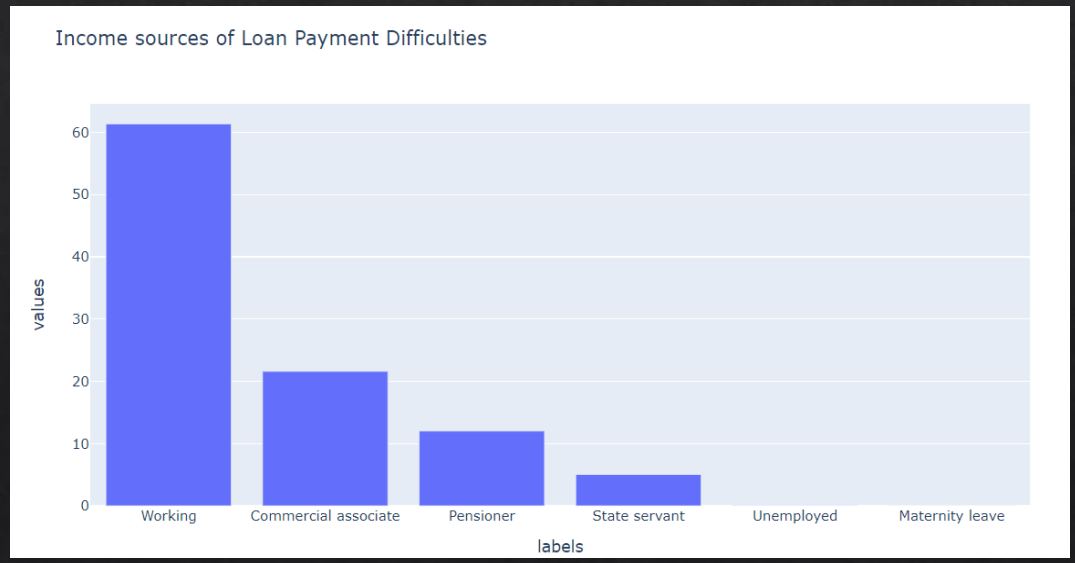


Univariate Analysis

Income Source

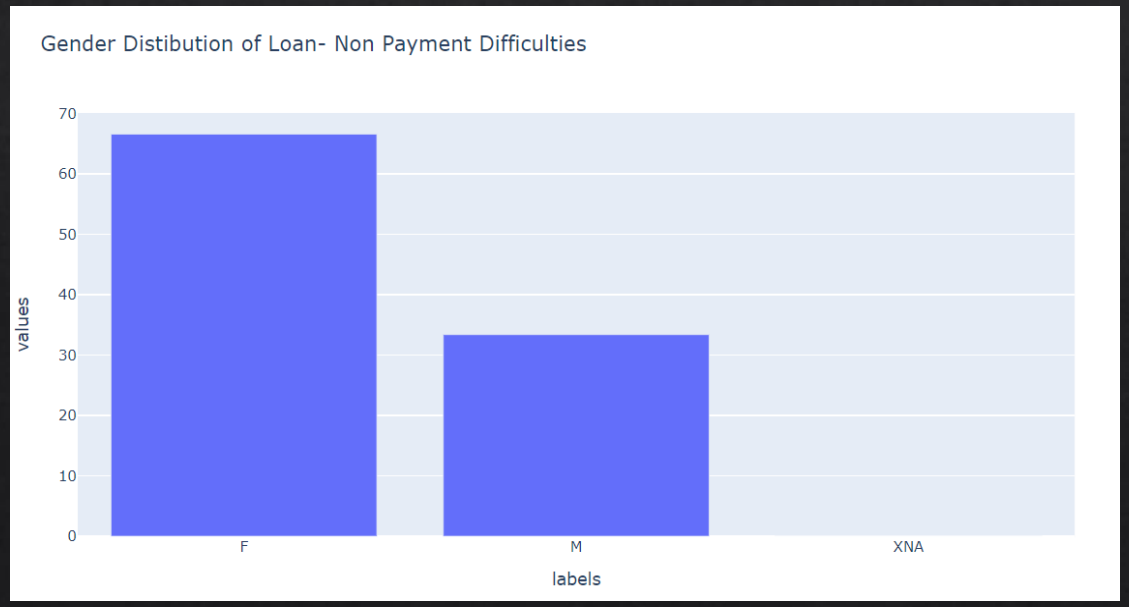
- We can see that there is a decrease in the percentage of Payment Difficulties who are pensioners, State servant and Commercial associate and an increase in the percentage of Payment Difficulties who are Working class when compared the percentages of both Payment Difficulties and non-Payment Difficulties.

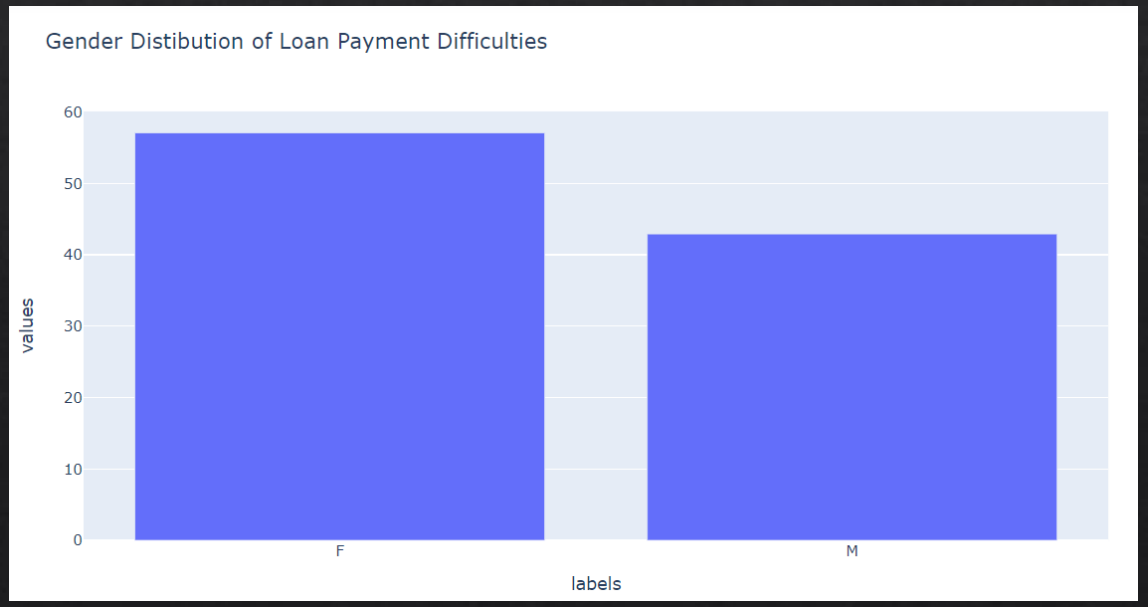




Gender Distribution

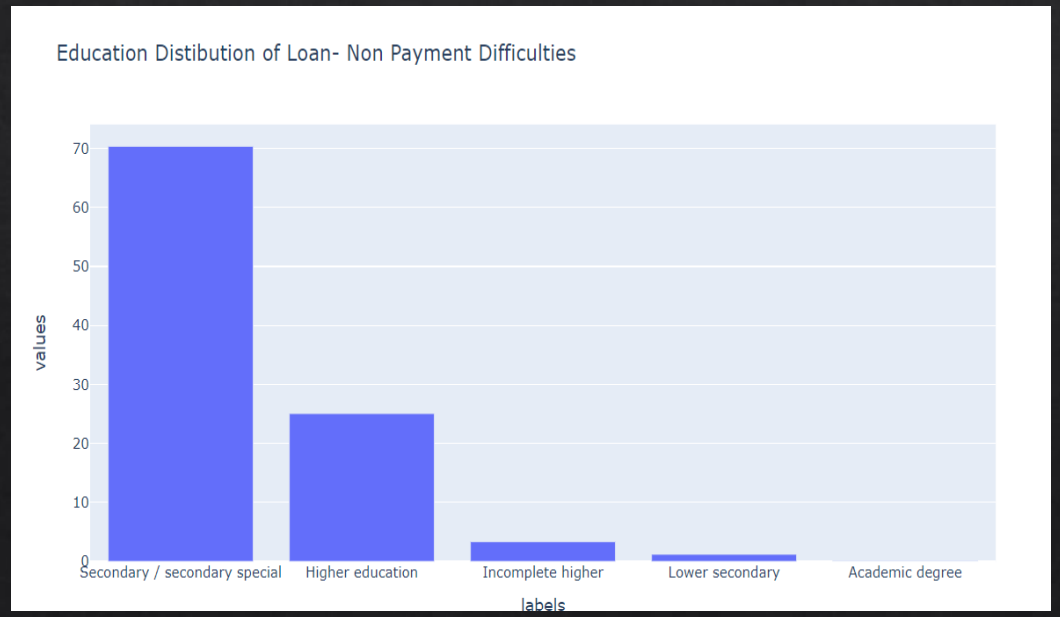
- We can see that Females are the majority in both cases. The % slightly increase in female loan non payment and % decreases for males in Loan payment as compared to loan non-payment.
- XNA – are people who have not disclosed their gender in application.

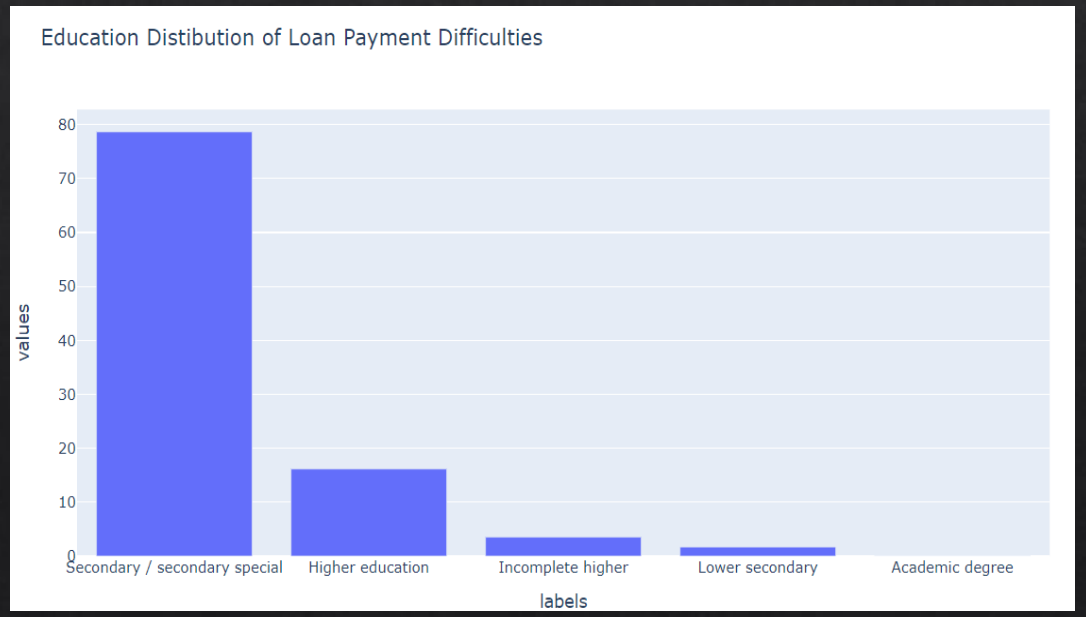




Education Distribution

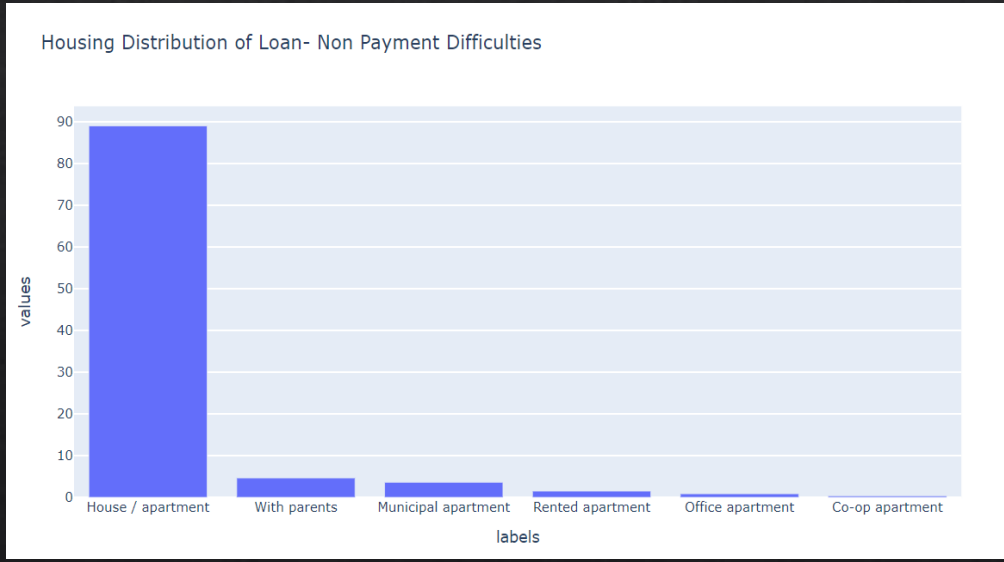
- In both cases people with Academic Degree do not face difficulties in repaying loan. Whereas people who have done Secondary schooling have faced more difficulties while repaying the loan.
- Higher education category has increased in loan non payment difficulties comparatively.

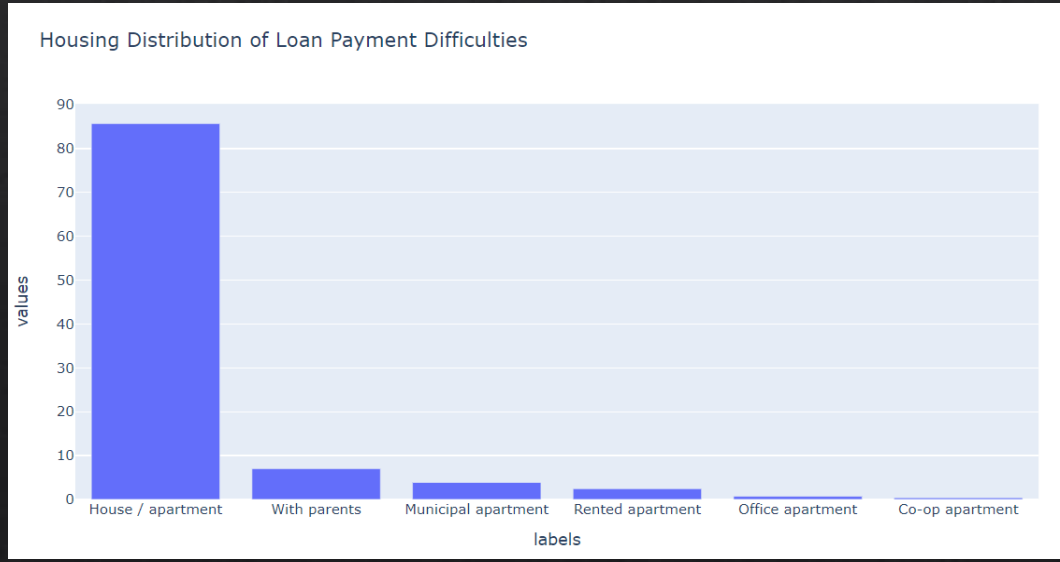




Housing Distribution

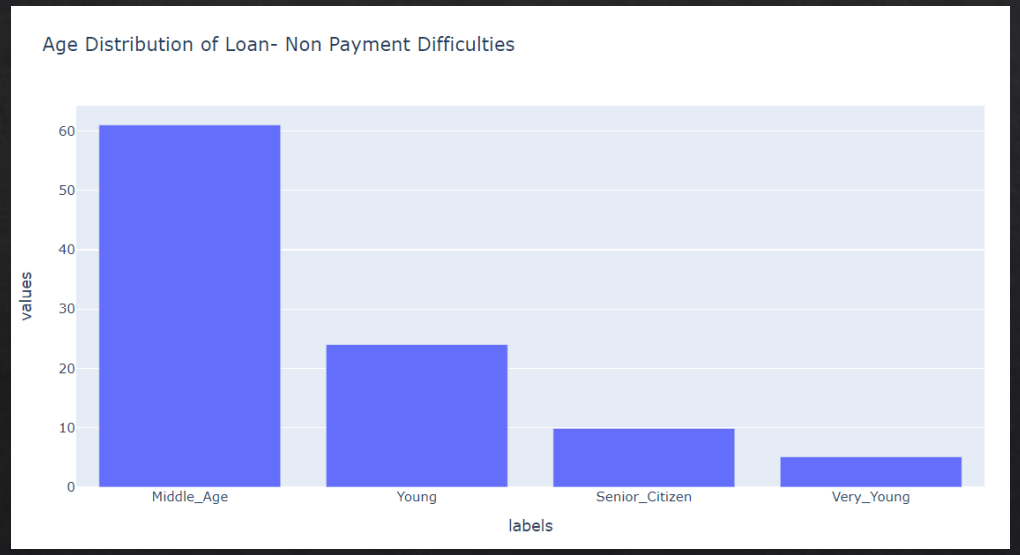
- In both cases we can see that people who are accommodated House/Apartment face more difficulties in loan re payment.
- In case of loan payment difficulties there is a slight increase in people living with parents.

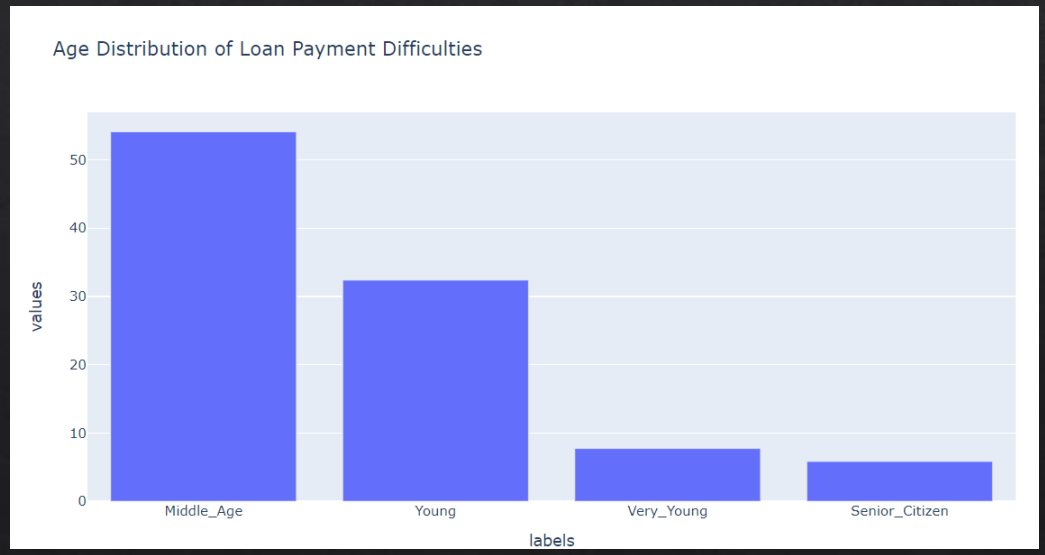




Age Distribution

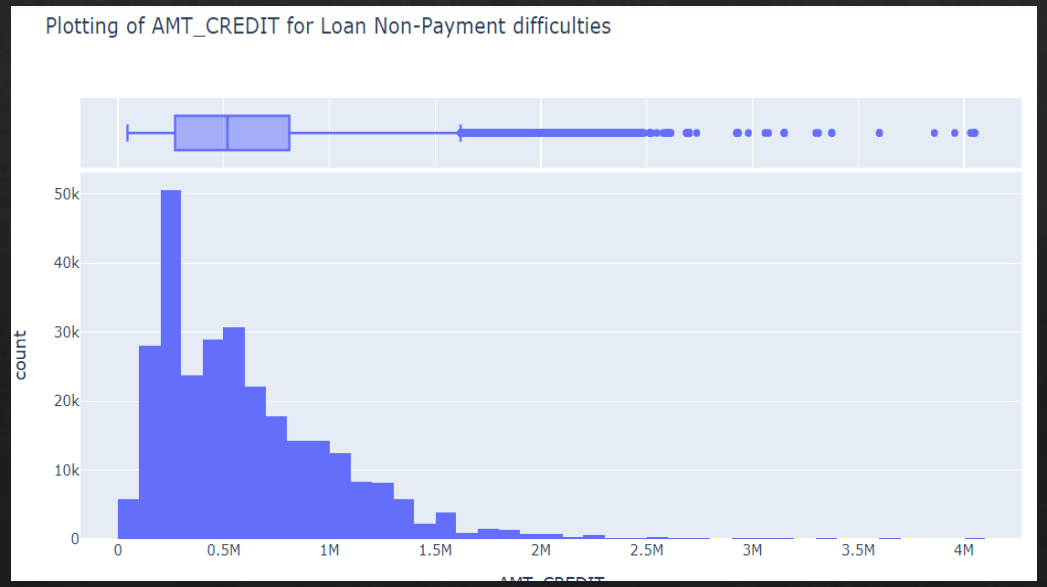
- In both categories we can observe that Middle aged people are having more difficulties.
- In Loan payment segment Youngsters percentage is comparatively more than in Loan Non payment. Senior citizens are also not able to re pay the loan amount monthly.

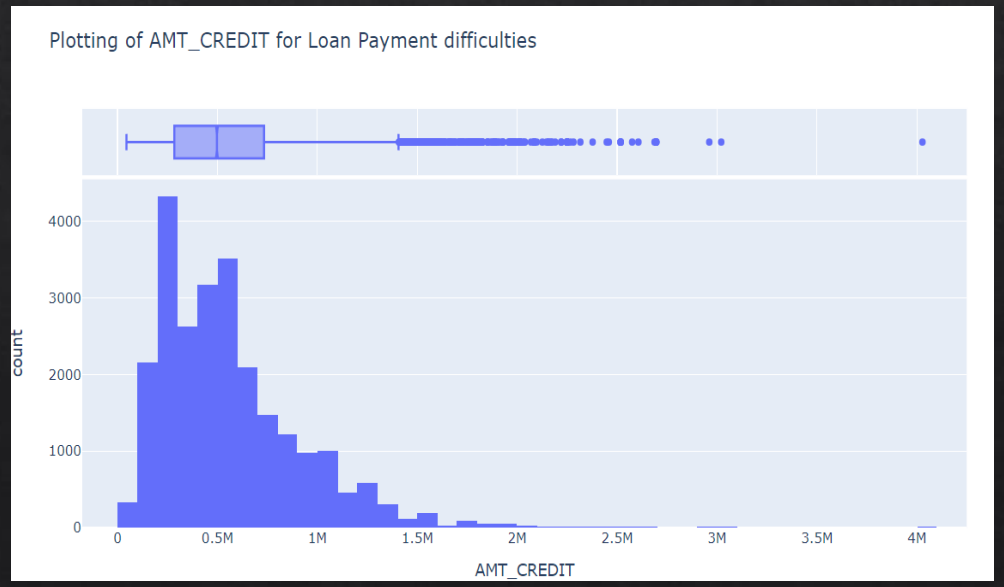




Credit Amount Distribution

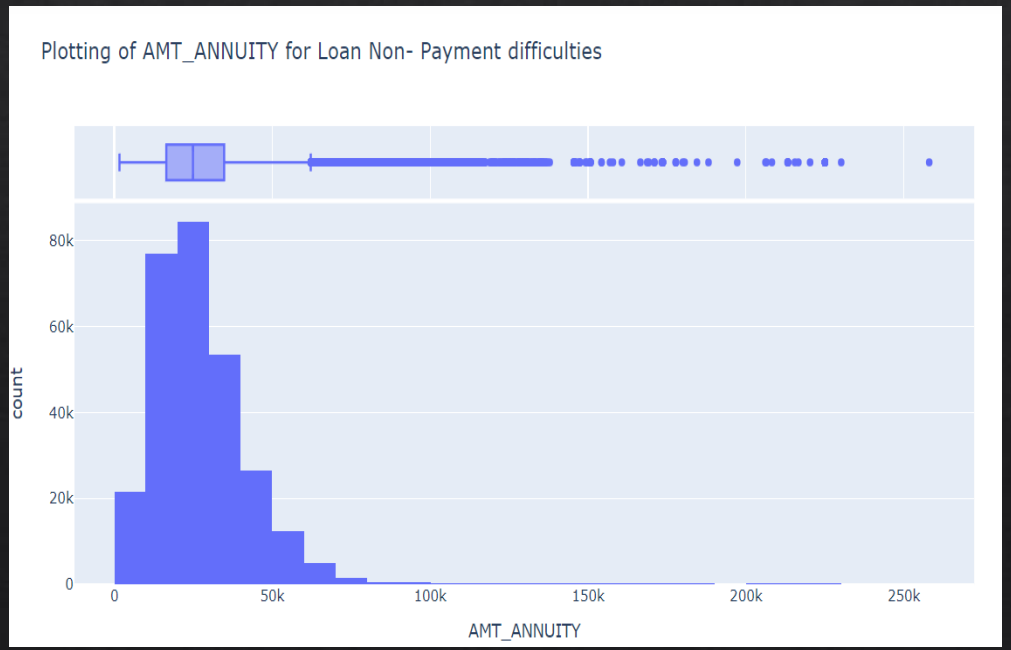
- We can see that in both target 0 and target 1 majority of the difficulties have amt_credit in range 200k and 600k.
- In target 0 we can see few outliers above the range of 3 million whereas in target 1 there are very few outliers above the 3 million region.

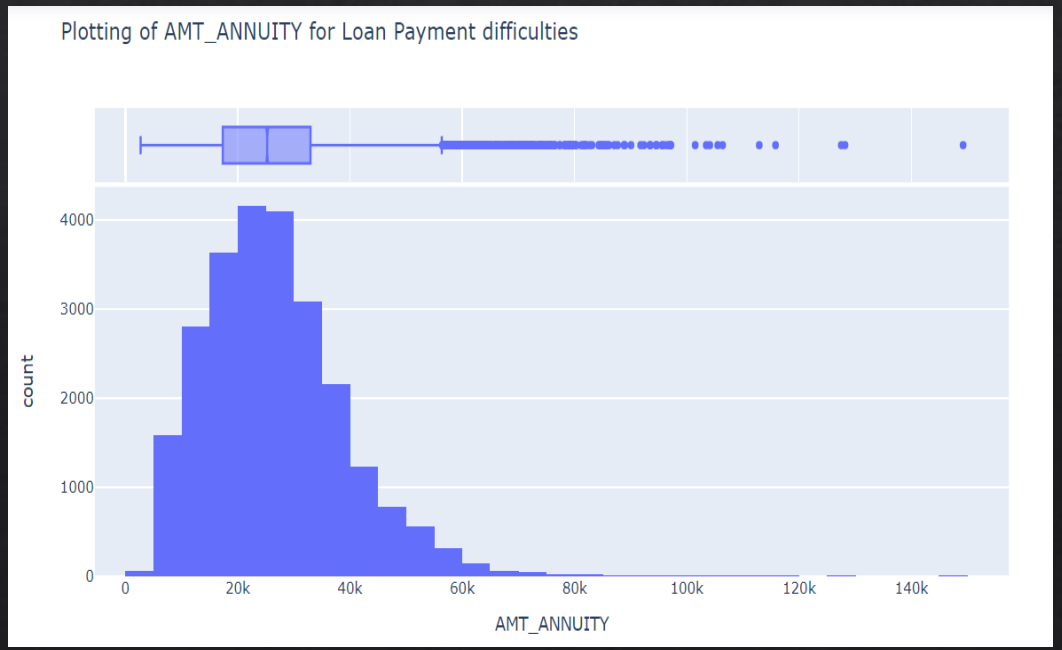




Annuity Amount

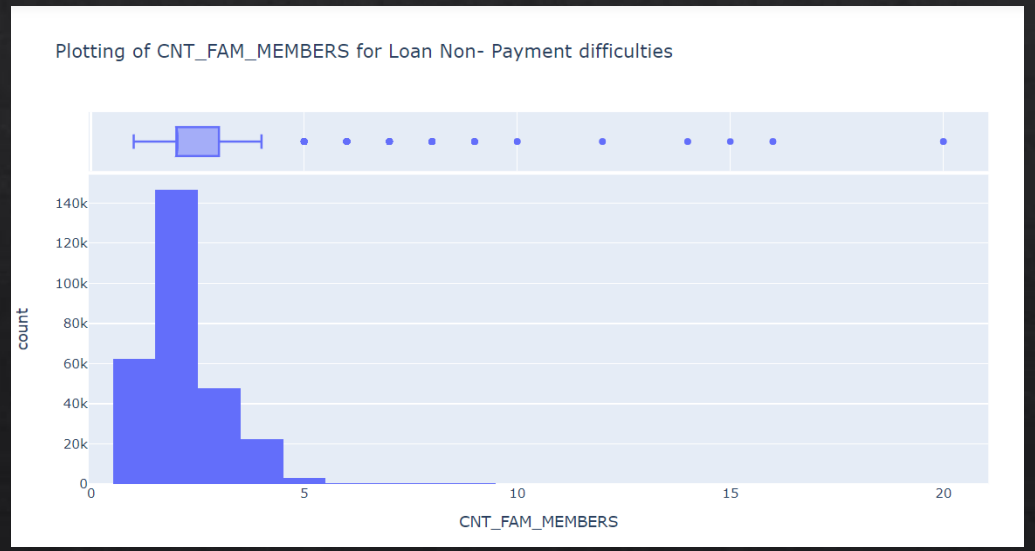
- In non payment difficulties the majority of annuity amt is in the 15k to 35k range with median around 25k and in payment difficulty the annuity amt is in the range of 17k to 32k with 25k median.
- Using outliers we can see a lot of people with non payment(target_0) are having high annual annuity.

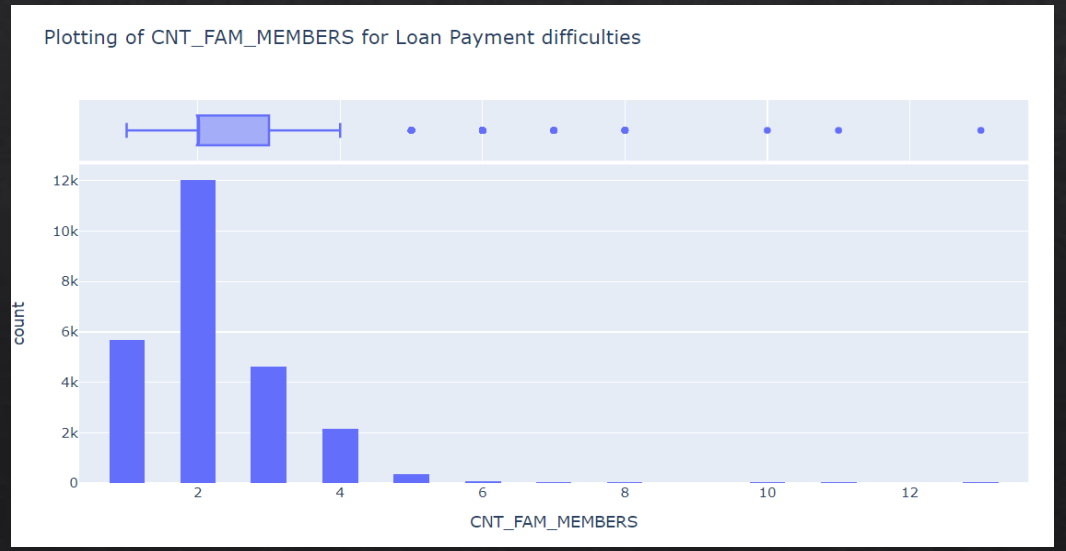




Family Members

- We can see that in both scenario people with less family members have more difficulty in repaying loans.in target 0 case majority of the family are less than 2.
- In target 1 case with median at 2 members in a family with majority in the range of 1 to 3 members.

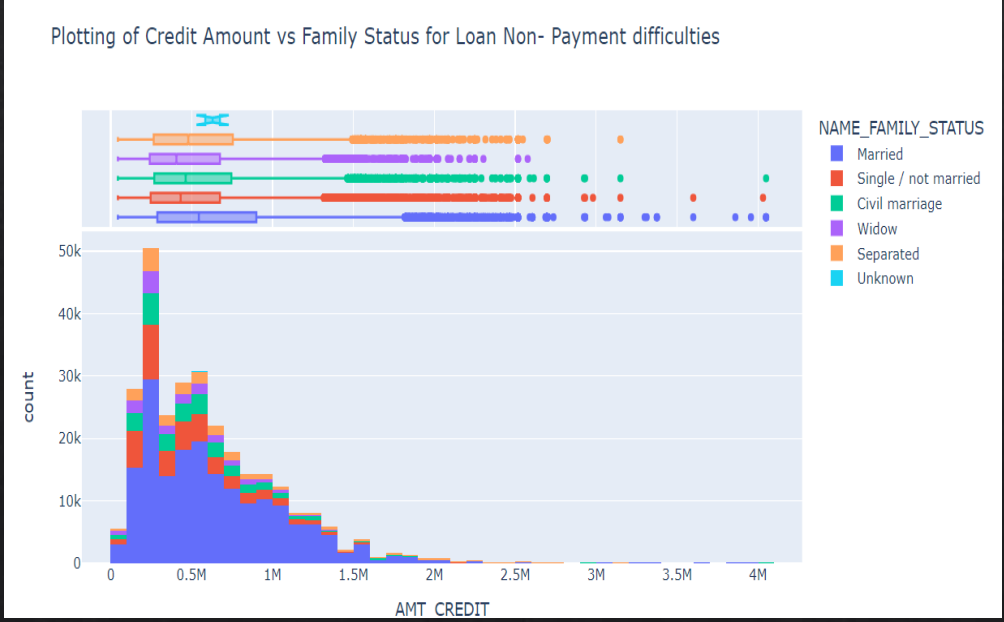


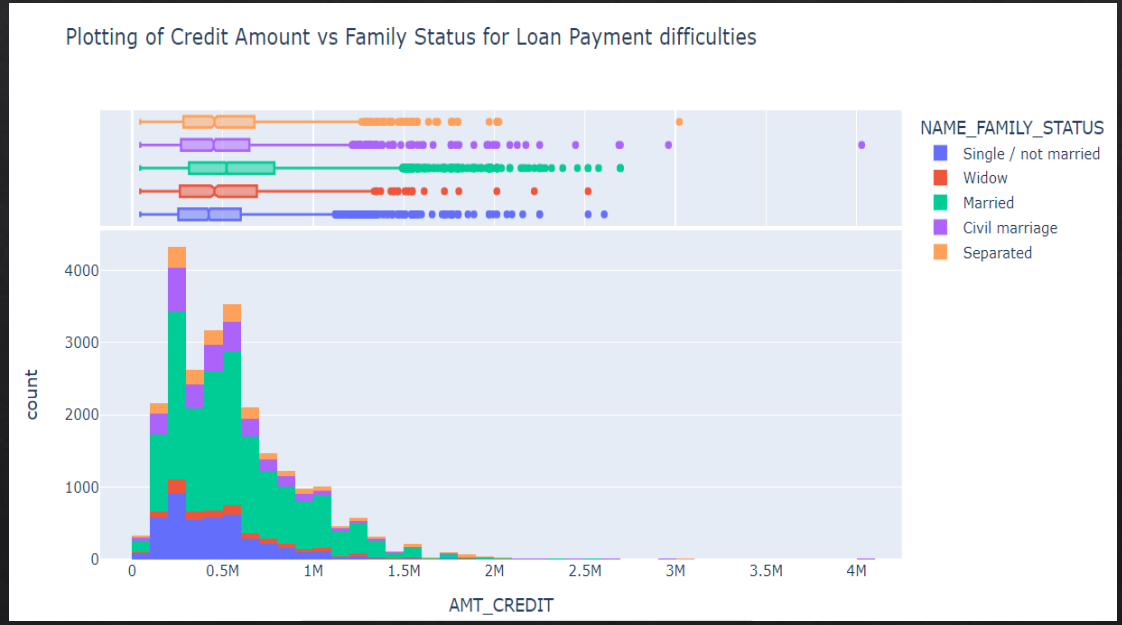


Bivariate Analysis

Income range vs Family Status

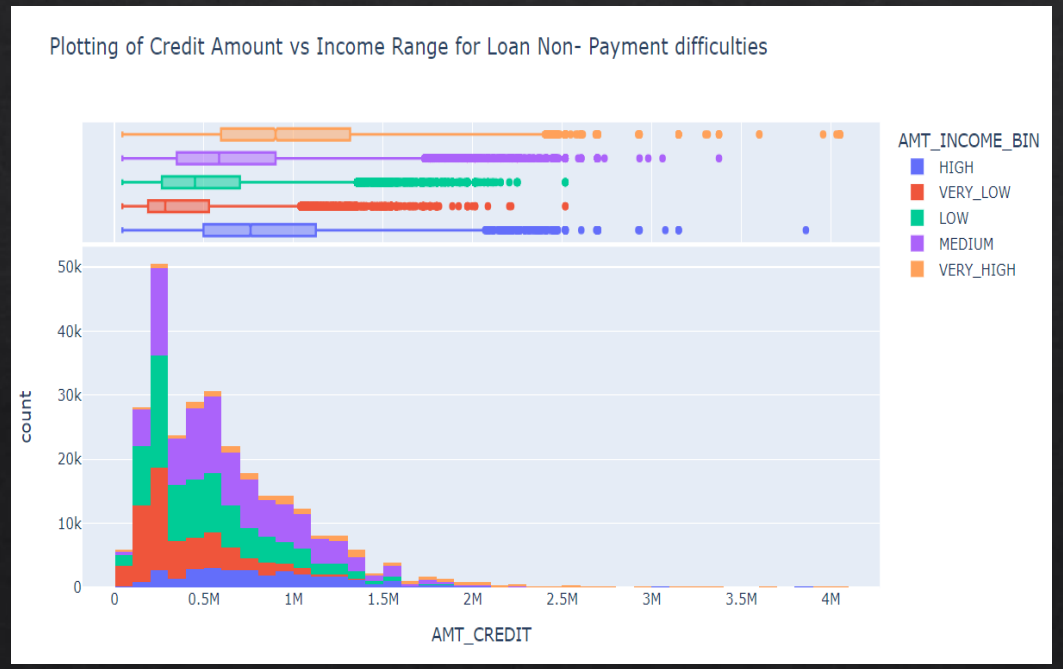
- In both cases we can observe that married people have the highest credit amount.
- In Loan non payment difficulties case single have an increased count as compared to Loan payment difficulties.

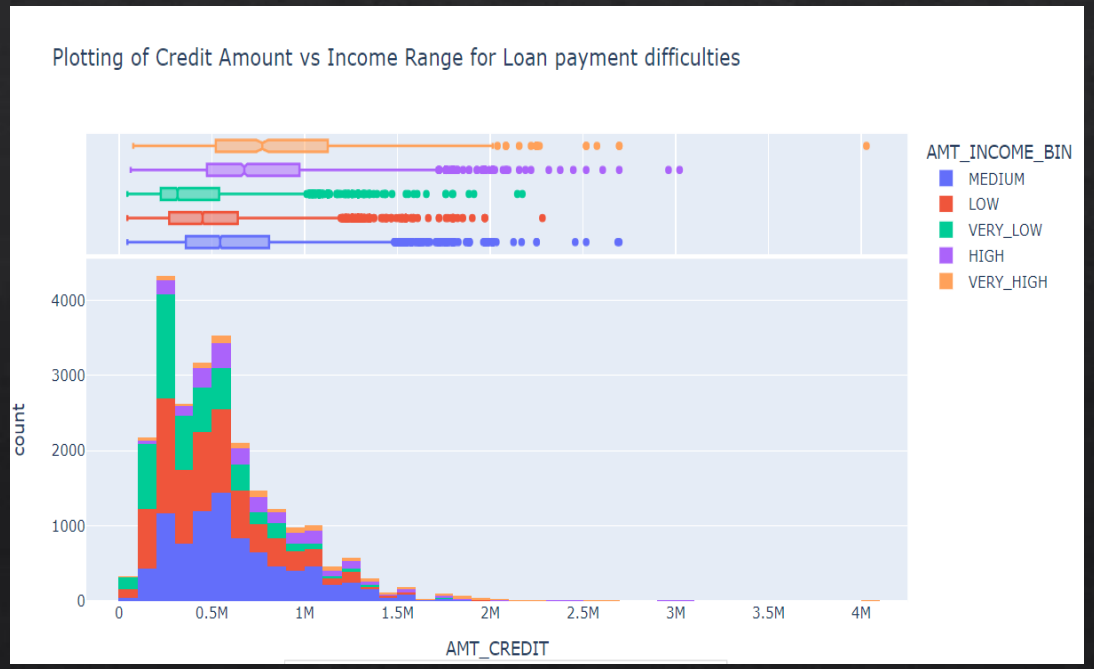




Income Range vs Credit Amount

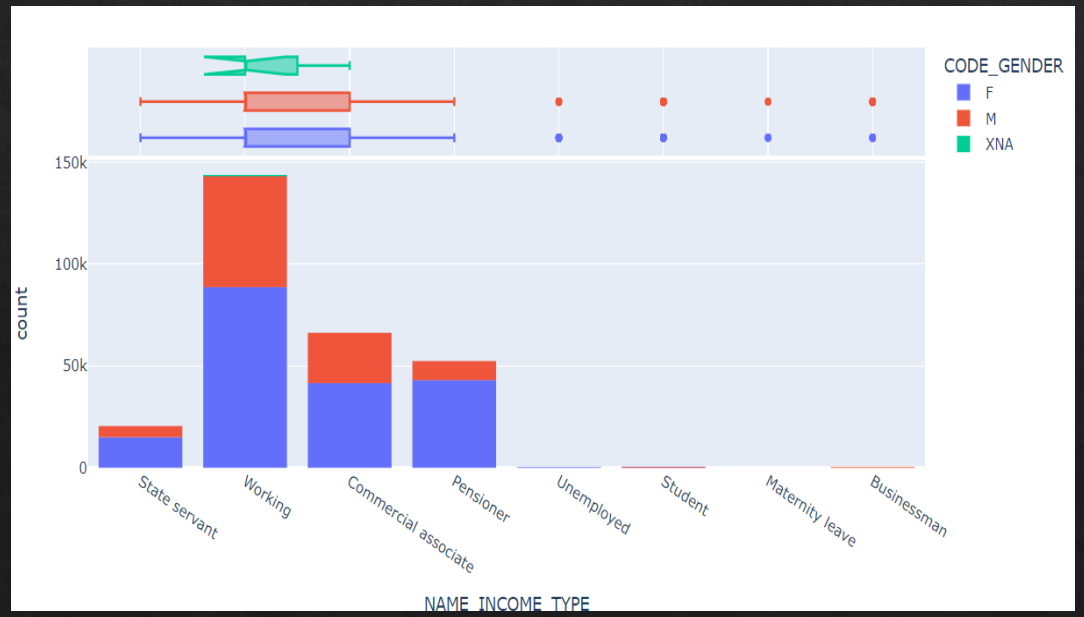
- In both scenario we can see that Medium , low , and very low income range have high credits.
- In case of loan non repayment medium income range has higher credits comparing the median and q3 quartile value in both scenarios.

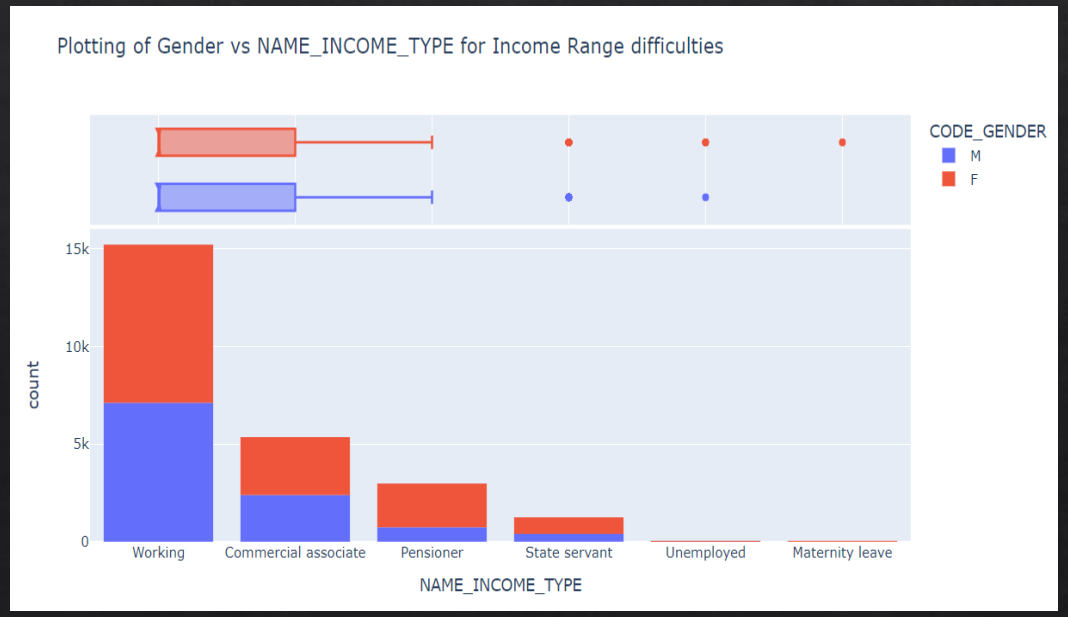




Income type vs Gender

- We can see that in both scenario the applicants are of the working income type for both genders.
- In Loan non payment_0 segment we can observe higher no of Females of Commercial Associate and Pensioner income type as compared to loan payment.

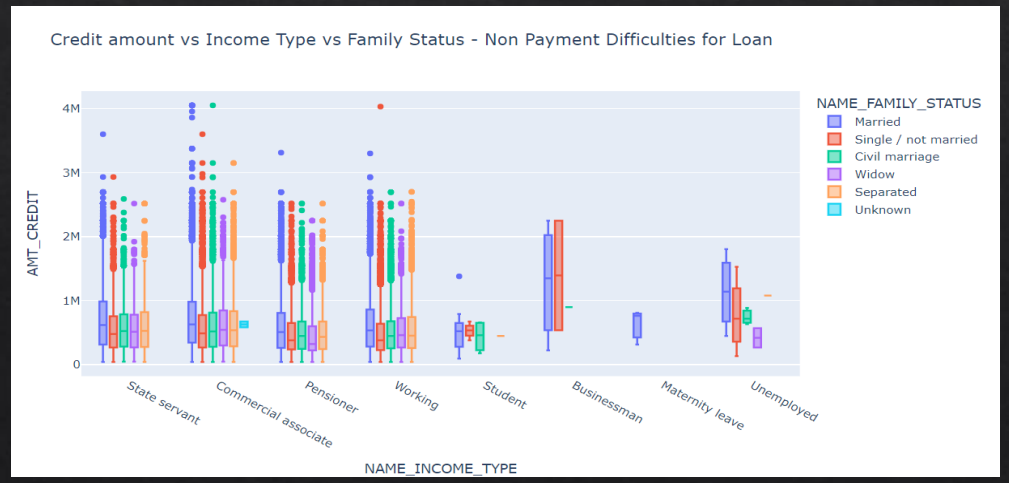


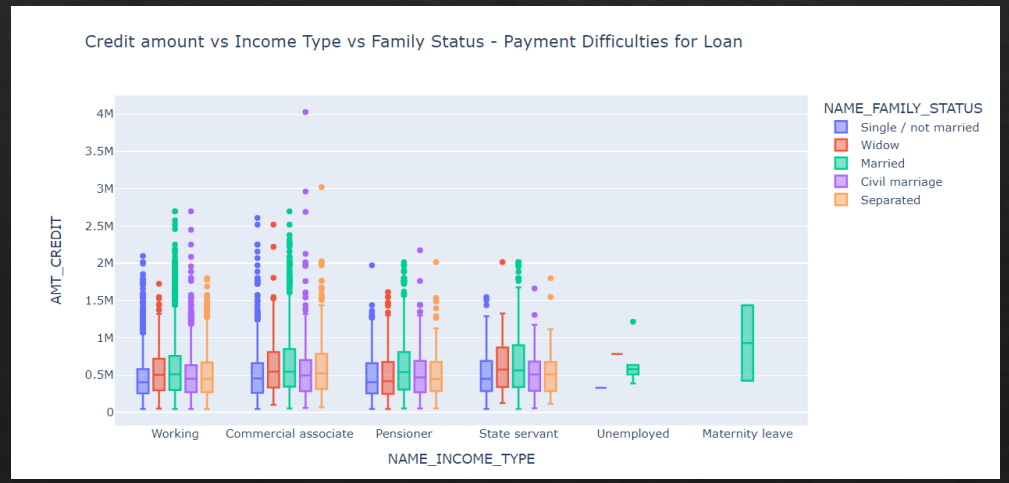


Multivariate Analysis

Credit amount vs Income Type vs Family Status

- We can observe that the median credit amt granted is the highest in the married family status in both the scenario of loan non payment and loan payment.
- Single family status has lowest median credit amt in loan payment difficulties and similarly in loan non payment single family status has the lowest median in most income type.





Data Correlation

- We take the abs value and apply quicksort to find columns with high correlation for both target segment.

```
corr_0.sort_values(ascending = False).head(10)
```

```
6]: FLAG_EMP_PHONE                DAYS_EMPLOYED                0.999756
    DAYS_EMPLOYED                 FLAG_EMP_PHONE               0.999756
    OBS_60_CNT_SOCIAL_CIRCLE      OBS_30_CNT_SOCIAL_CIRCLE     0.998510
    OBS_30_CNT_SOCIAL_CIRCLE      OBS_60_CNT_SOCIAL_CIRCLE     0.998510
    AMT_GOODS_PRICE               AMT_CREDIT                   0.987022
    AMT_CREDIT                    AMT_GOODS_PRICE              0.987022
    REGION_RATING_CLIENT_W_CITY   REGION_RATING_CLIENT         0.950149
    REGION_RATING_CLIENT          REGION_RATING_CLIENT_W_CITY  0.950149
    CNT_CHILDREN                  CNT_FAM_MEMBERS              0.878571
    CNT_FAM_MEMBERS               CNT_CHILDREN                 0.878571
    dtype: float64
```

```
7]: corr_1.sort_values(ascending = False).head(10)
```

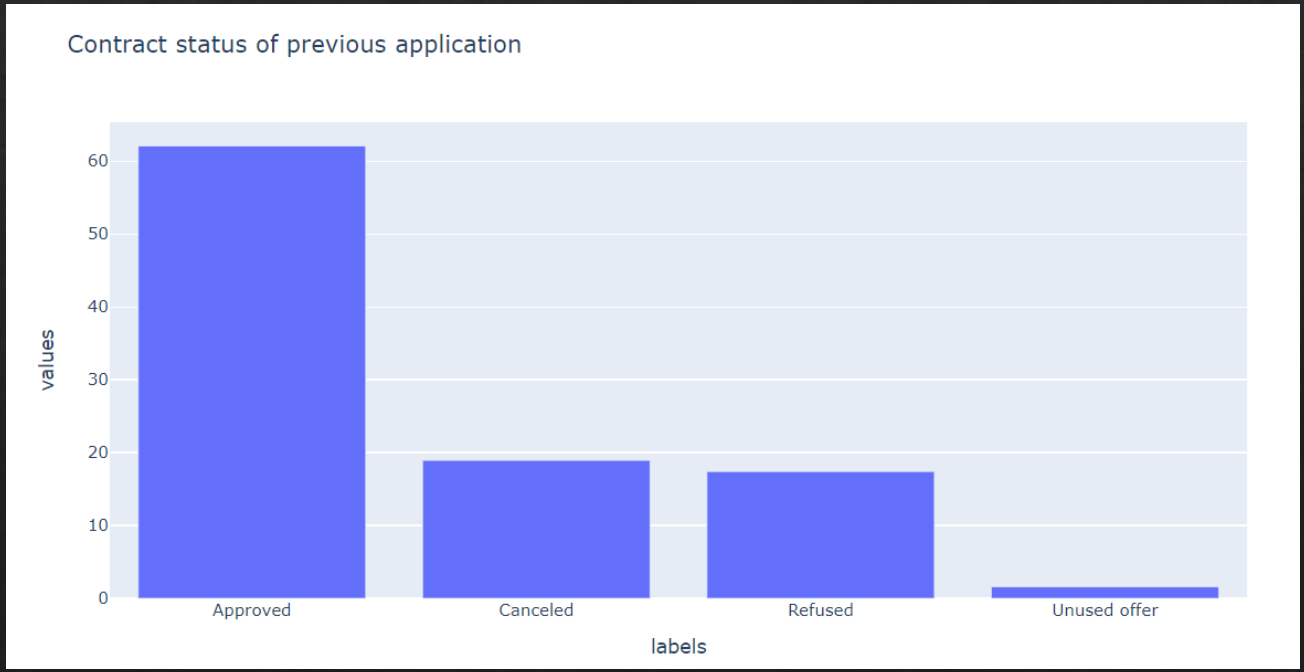
```
7]: DAYS_EMPLOYED                 FLAG_EMP_PHONE               0.999705
    FLAG_EMP_PHONE                DAYS_EMPLOYED                0.999705
    OBS_60_CNT_SOCIAL_CIRCLE      OBS_30_CNT_SOCIAL_CIRCLE     0.998270
    OBS_30_CNT_SOCIAL_CIRCLE      OBS_60_CNT_SOCIAL_CIRCLE     0.998270
    AMT_CREDIT                    AMT_GOODS_PRICE              0.982783
    AMT_GOODS_PRICE               AMT_CREDIT                   0.982783
    REGION_RATING_CLIENT_W_CITY   REGION_RATING_CLIENT         0.956637
    REGION_RATING_CLIENT          REGION_RATING_CLIENT_W_CITY  0.956637
    CNT_CHILDREN                  CNT_FAM_MEMBERS              0.885484
    CNT_FAM_MEMBERS               CNT_CHILDREN                 0.885484
    dtype: float64
```

Previous Application Data set

- In this there is no need to clean the data set apart from removing NaN values for few columns.
- And we have applied few plotting technique in few columns and merge the data set with application data set using the column SK_ID_CURR.
- We have used inner join to merge both the data set.

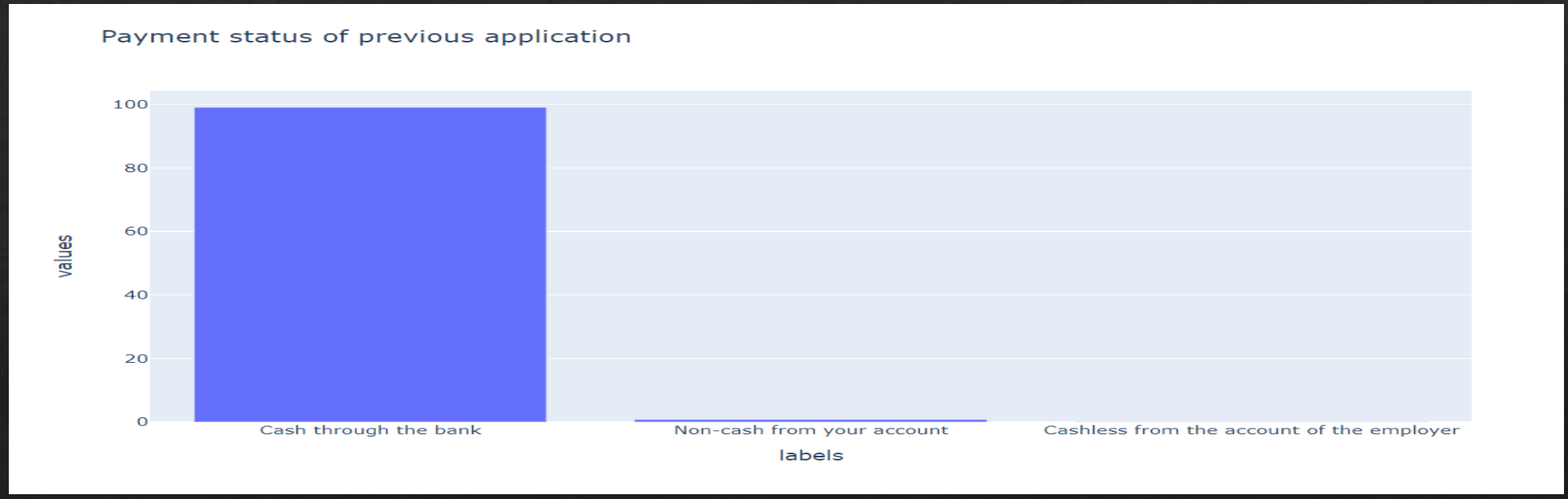
Univariate Analysis for Contract Status

- We can see that more that 50% of previous application has been approved and less than 20% have been cancelled. Very less no of loan application have been unused.



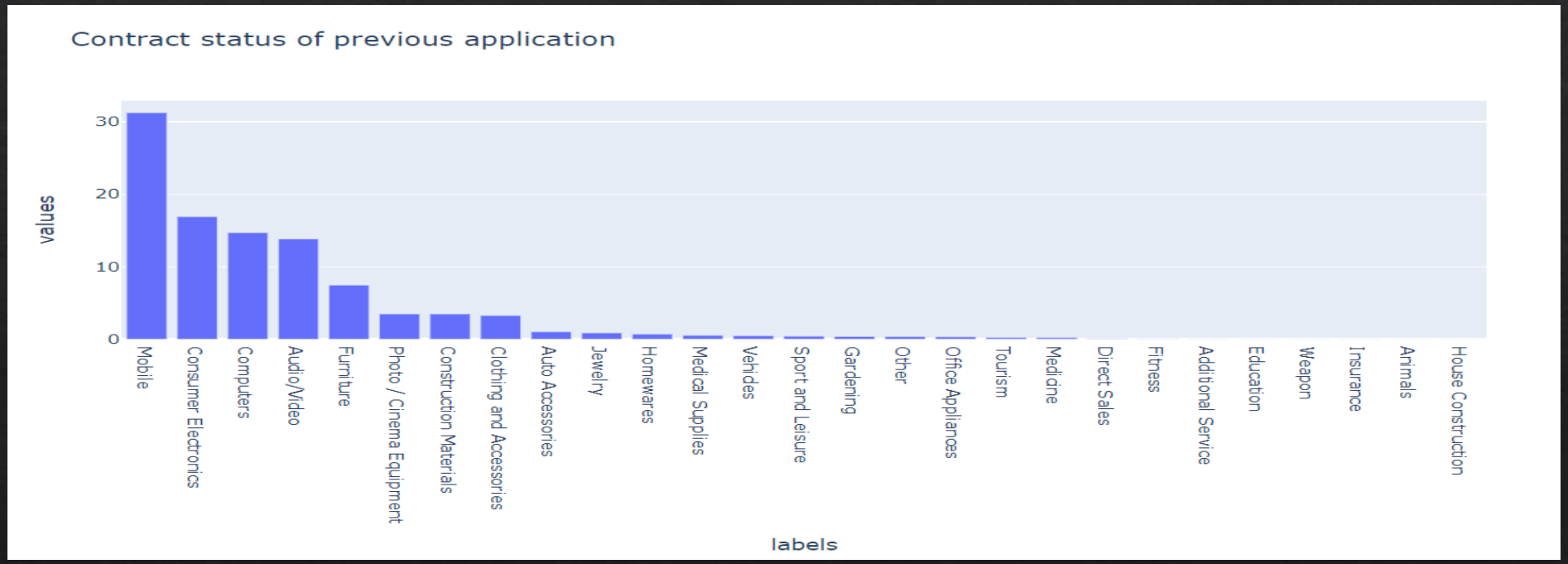
Univariate Analysis for Contract Status

- We can observe that majority of the application have been Cash through bank.



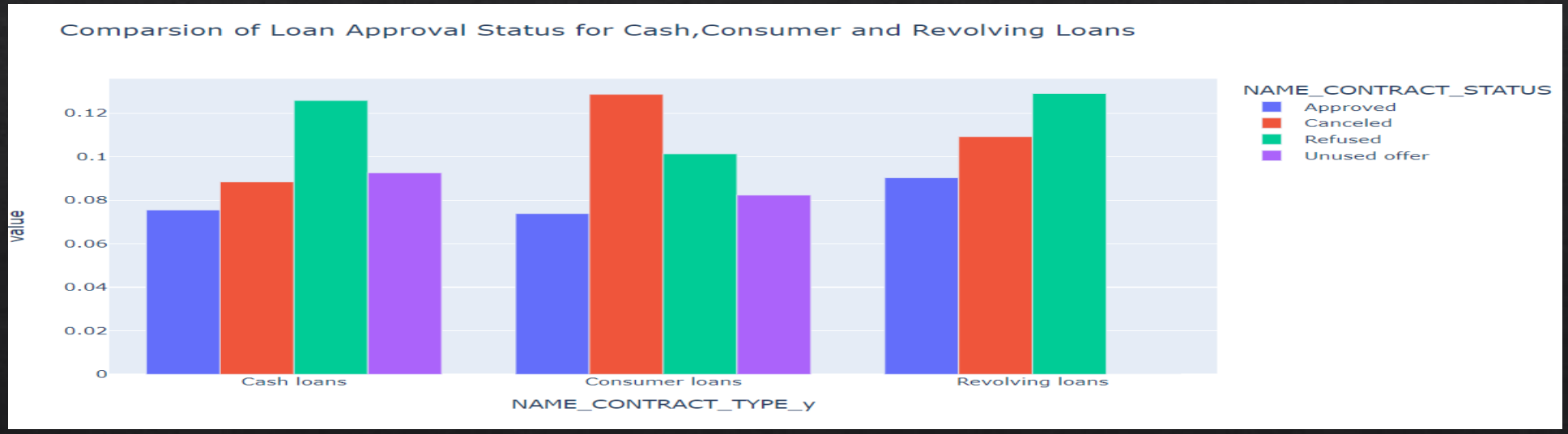
Univariate Analysis for Goods types

- We can observe that loan for Mobile is the highest where as for loan for House Construction is the lowest.



Comparison of Loan Approval Status for Cash, Consumer and Revolving Loans

- We can see that Refused loan application is the highest for Cash loan.
- For consumer loan type has the highest application refused.
- For Revolving loans there are no unused offers and majority loan application have been refused.



Thank You