EDA Case Study

By: Sainatth Wagh

Nidhi G.

Problem Statement

There are two types of risks associated with any loan request:

H0: If applicant likely to repay the loan, then not approving results in loss of business.

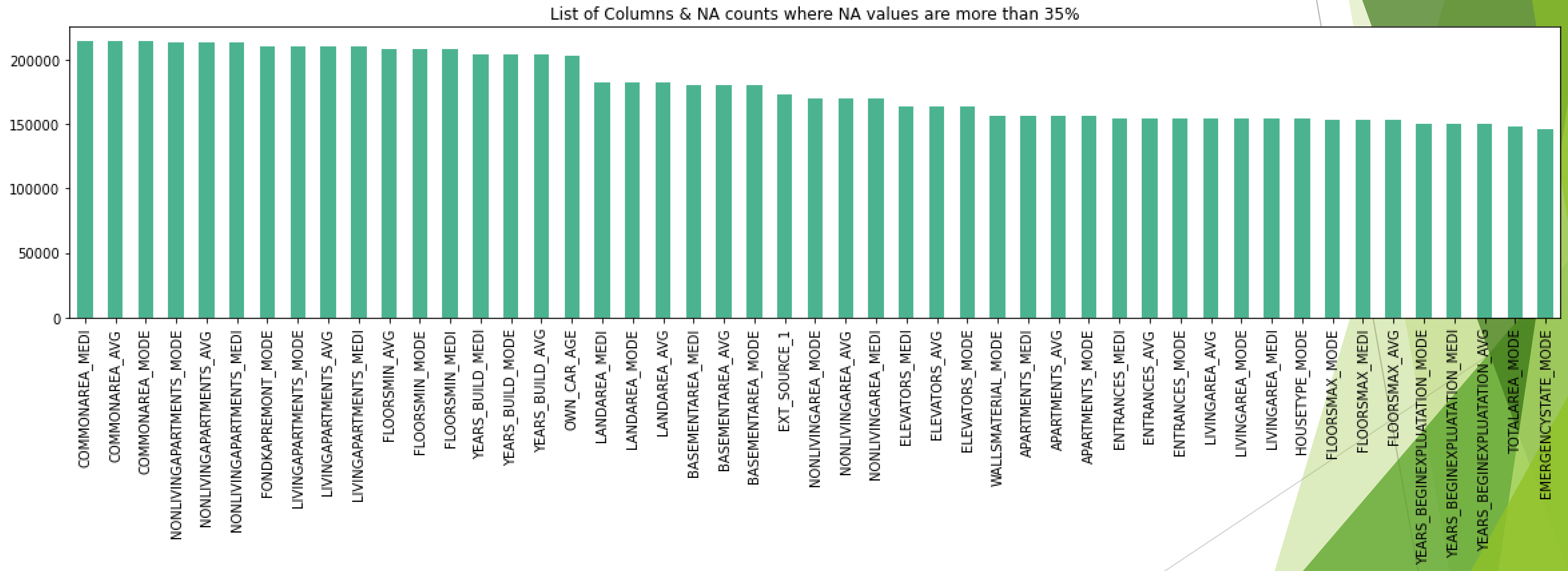
H1: If applicant not likely to repay the loan likely to default then approving that loan would result in financial loss to the company.

Analysis of data set has been done in Python Jupyter notebook.

Steps for Analysis Done:

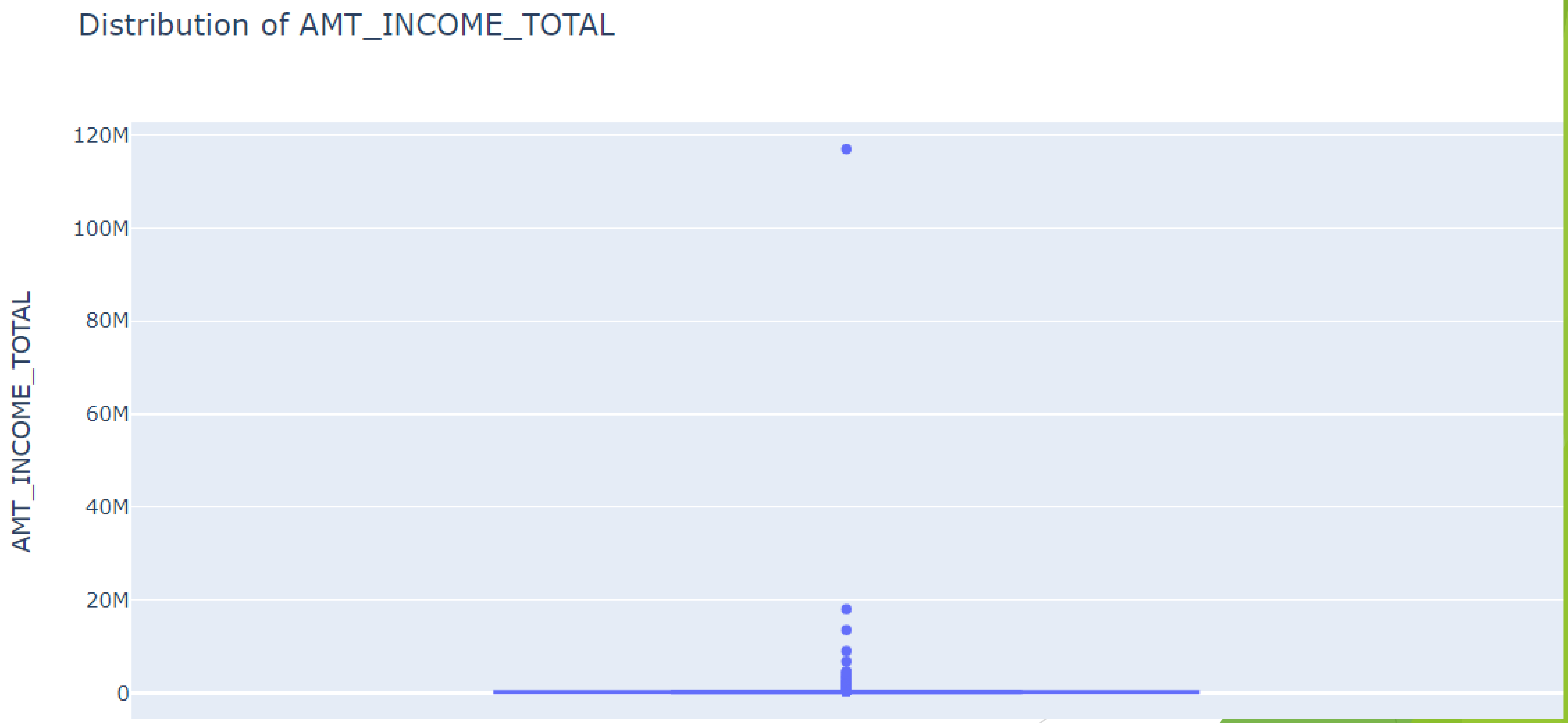
- Checking the missing values and handling the same.
- Checking outliers and imbalance.
- Top 10 correlation factors.
- Which correlation is most important.

Data Cleaning checks where missing values greater than 35%

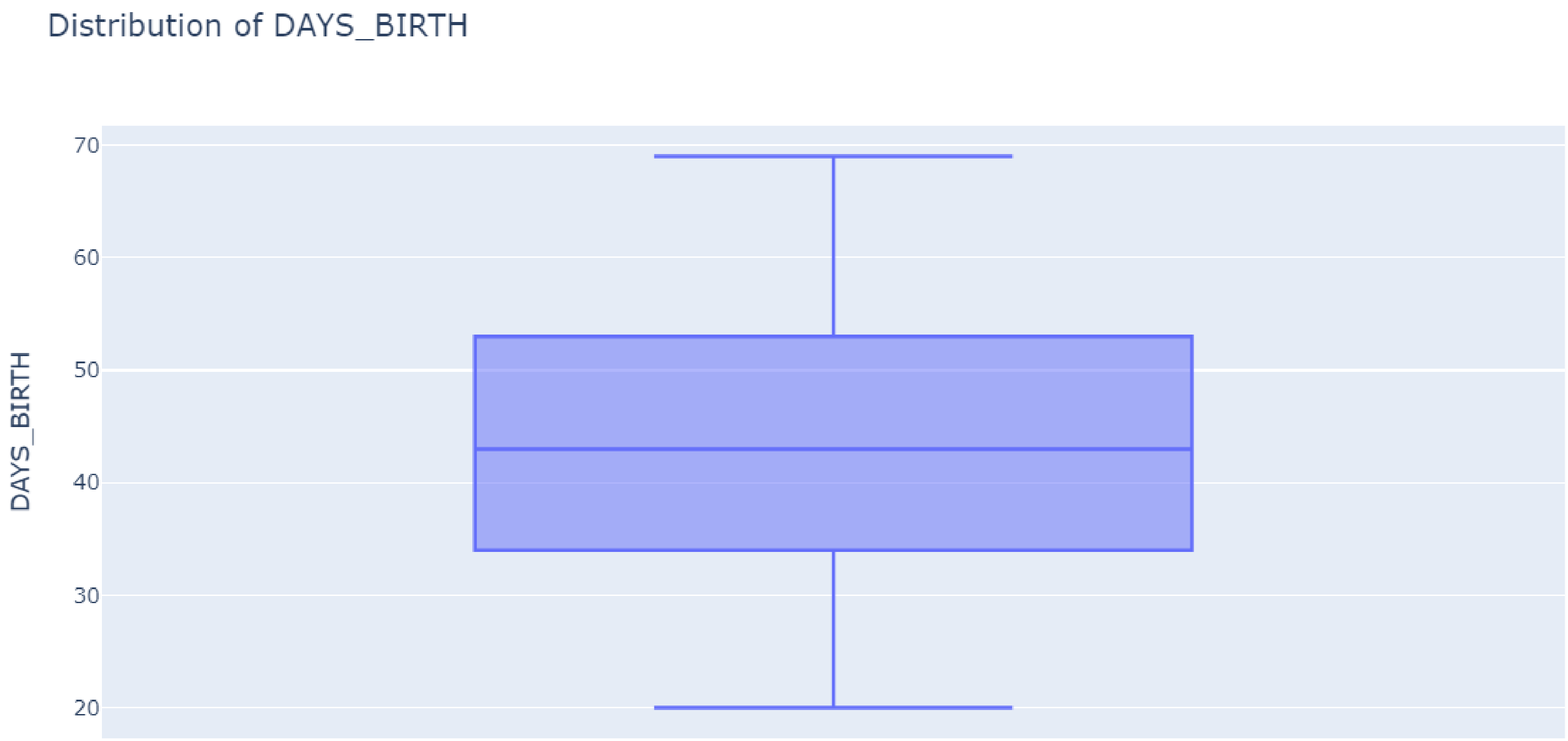


List of Columns & NA counts where NA values are more than 35%

Outliers for Amt_Income_Total



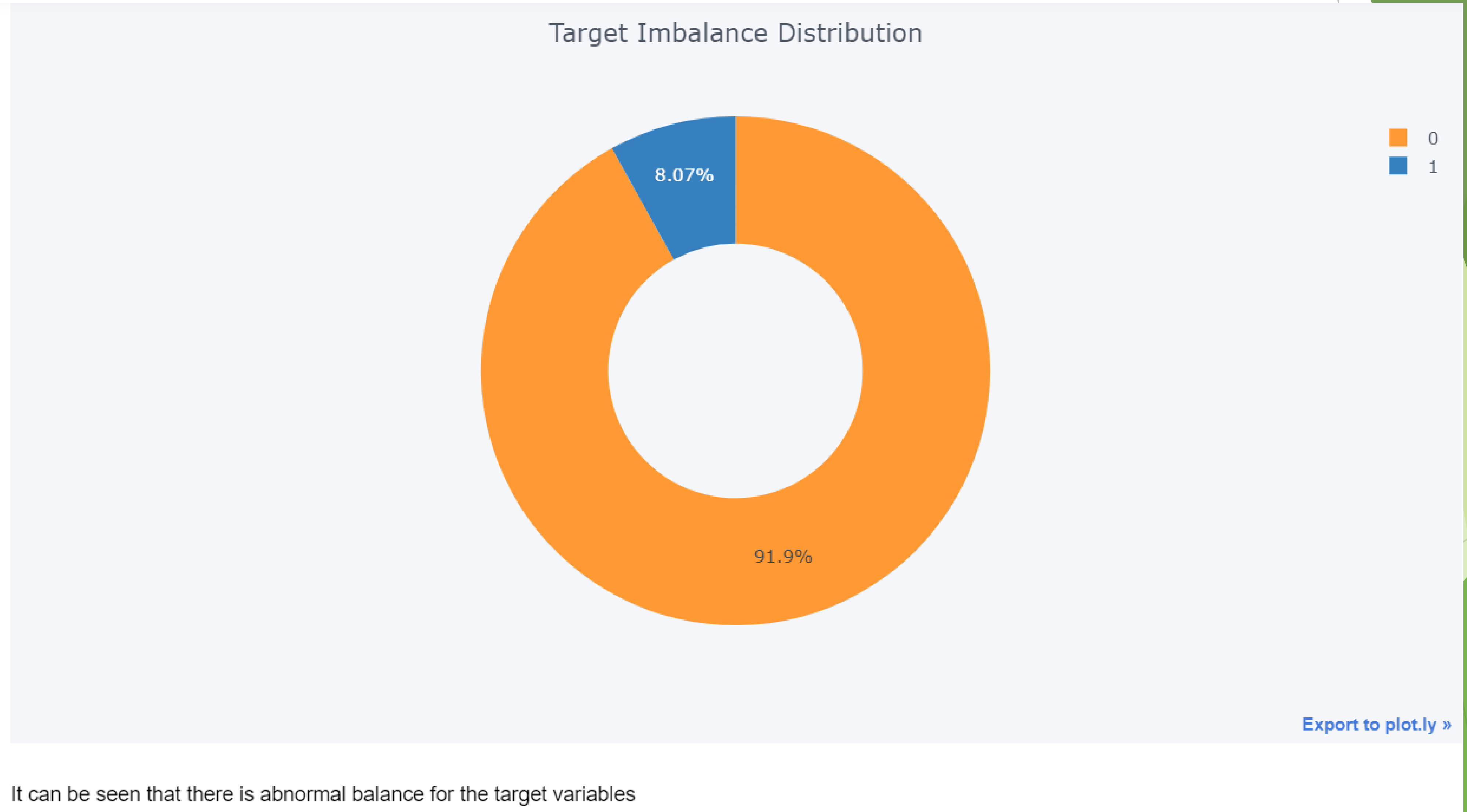
Outliers wrt Days_by_birth



Distribution of DAYS_BIRTH

It is very evident there are no outliers for the DAYS_BIRTH

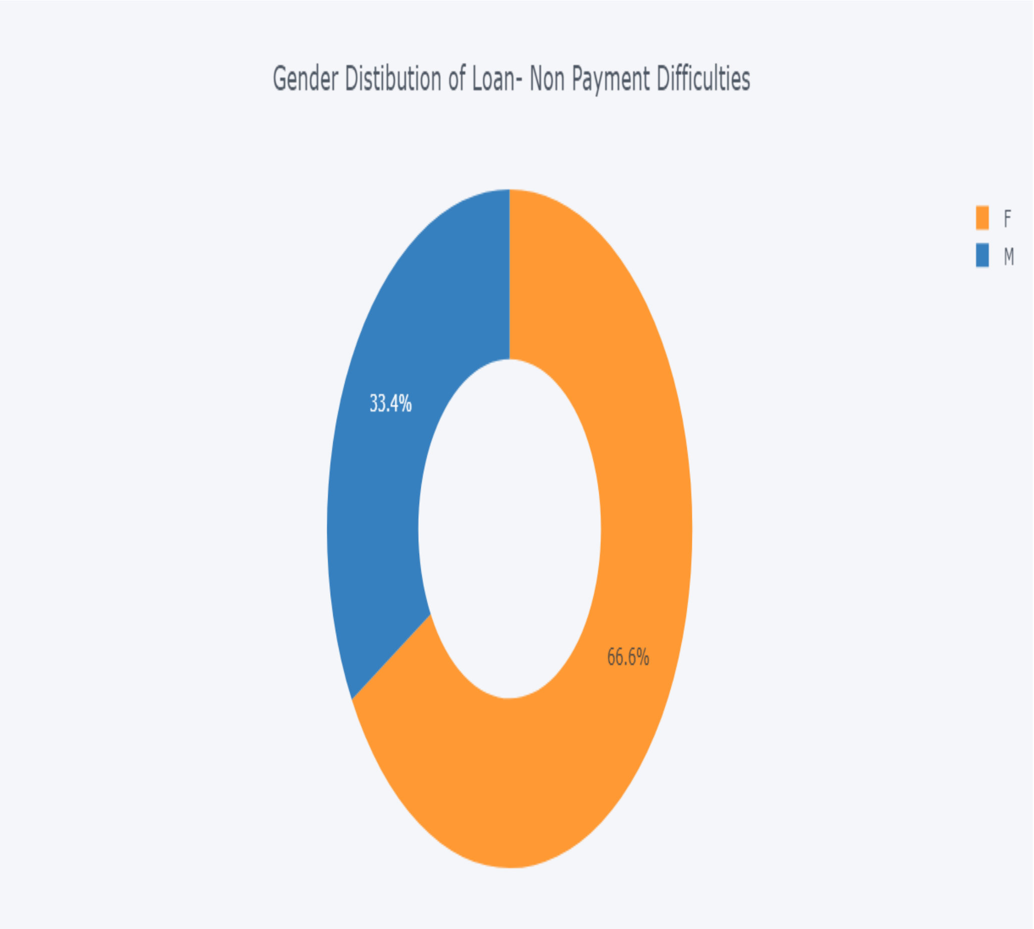
Checking Imbalance

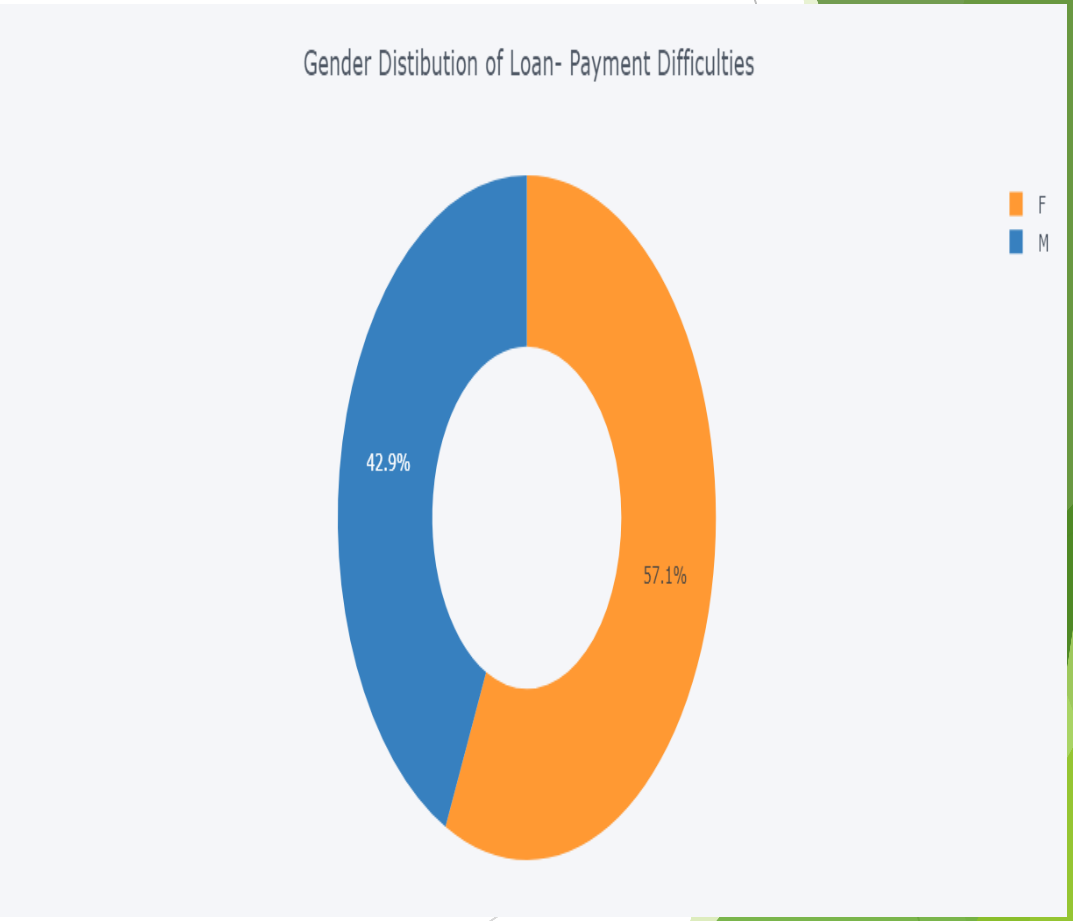


It can be seen that there is abnormal balance for the target variables

Univariate Analysis : Gender

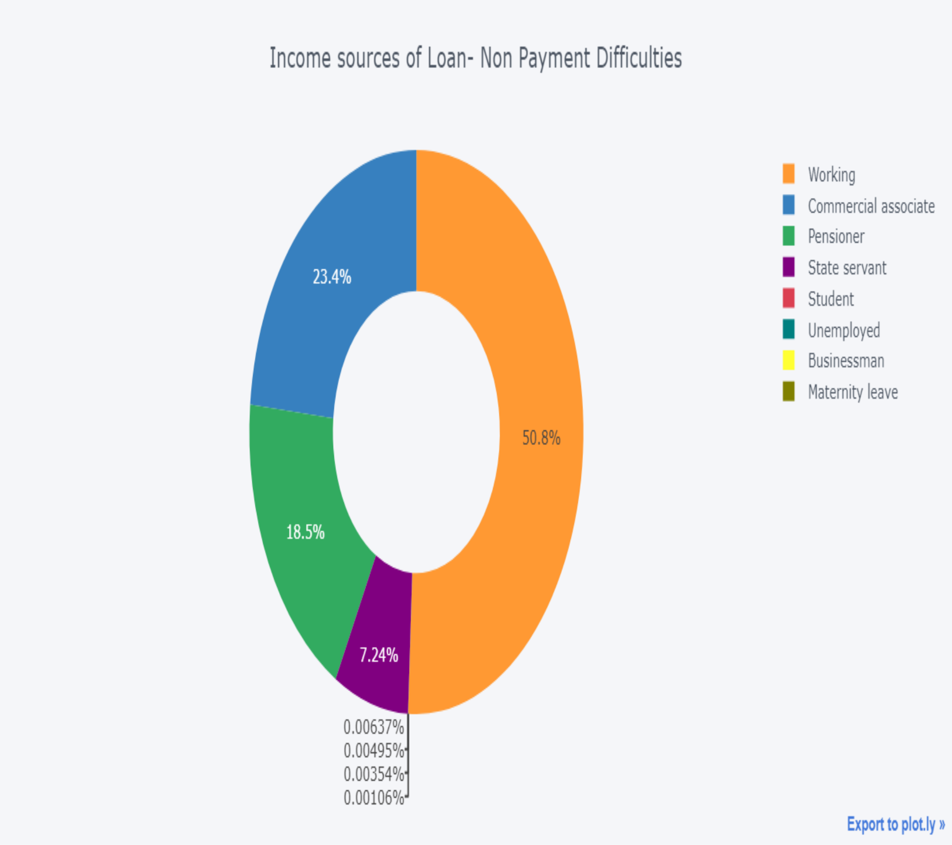
Inference: Comparing the Payment Difficulties and Non Payment Difficulties on the basis of Gender, we observe that Females are in majority in both the cases although there is an increase in the percentage in Male Payment Difficulties when compared with Non-Payment Difficulties

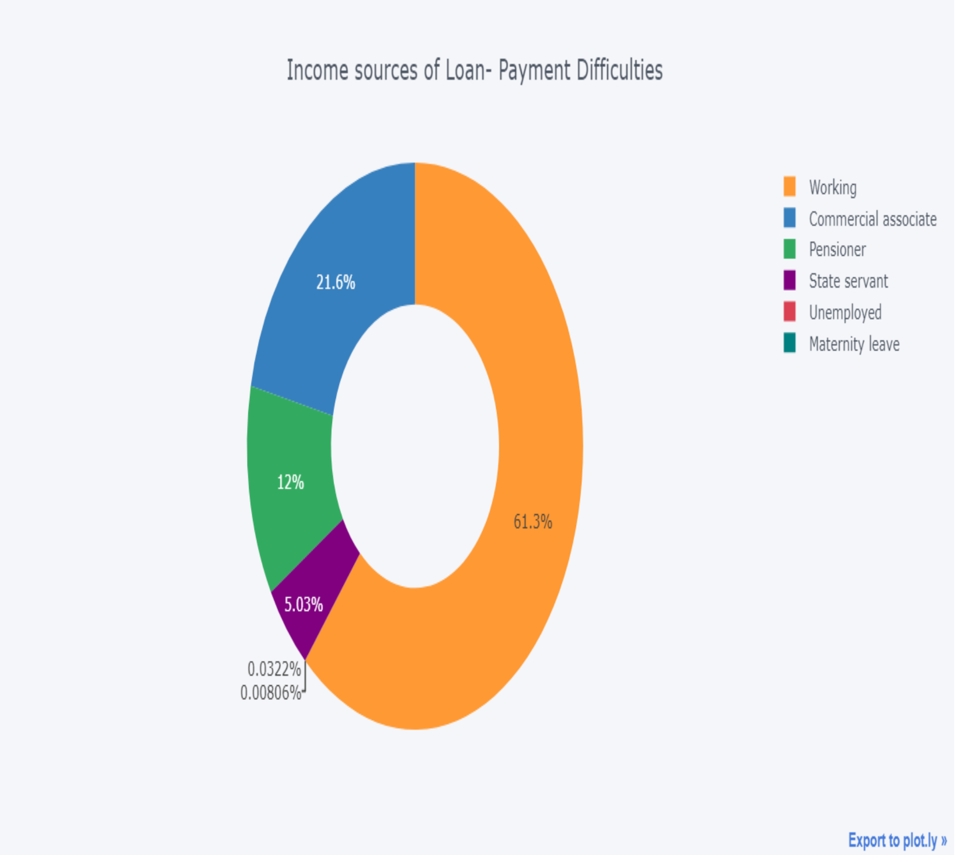




Income

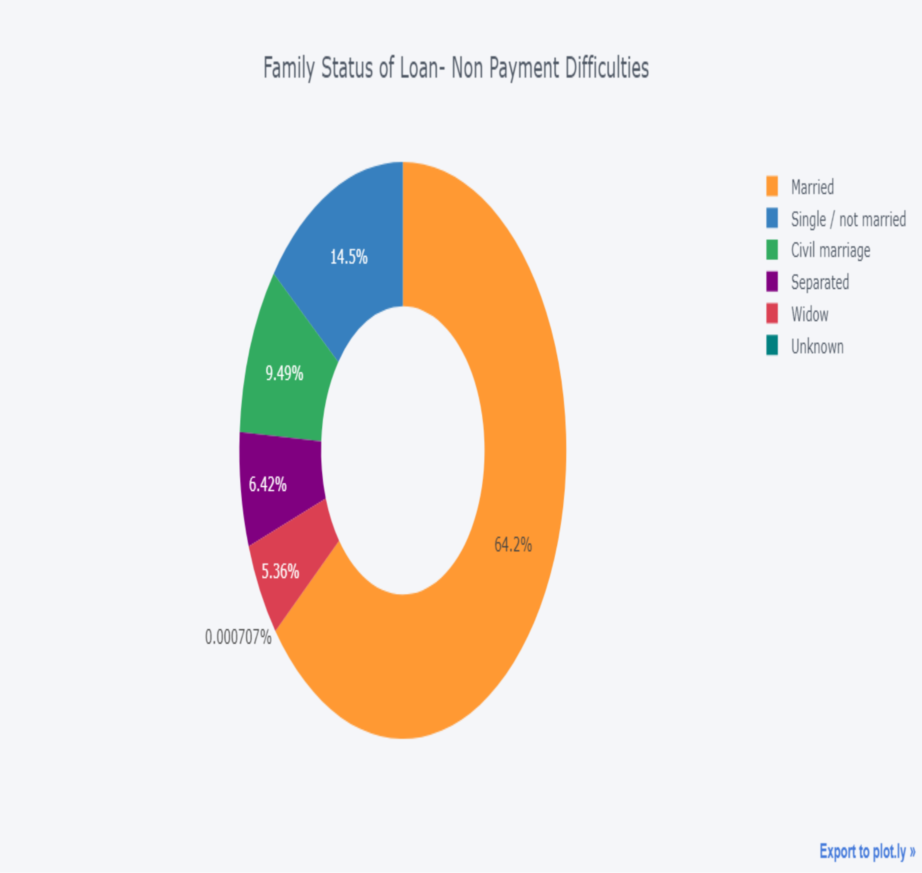
Inference: We can observe there is steep decrease for the category Pensioner when compared the percentages of both Loan Payment and Non Payment Difficulties

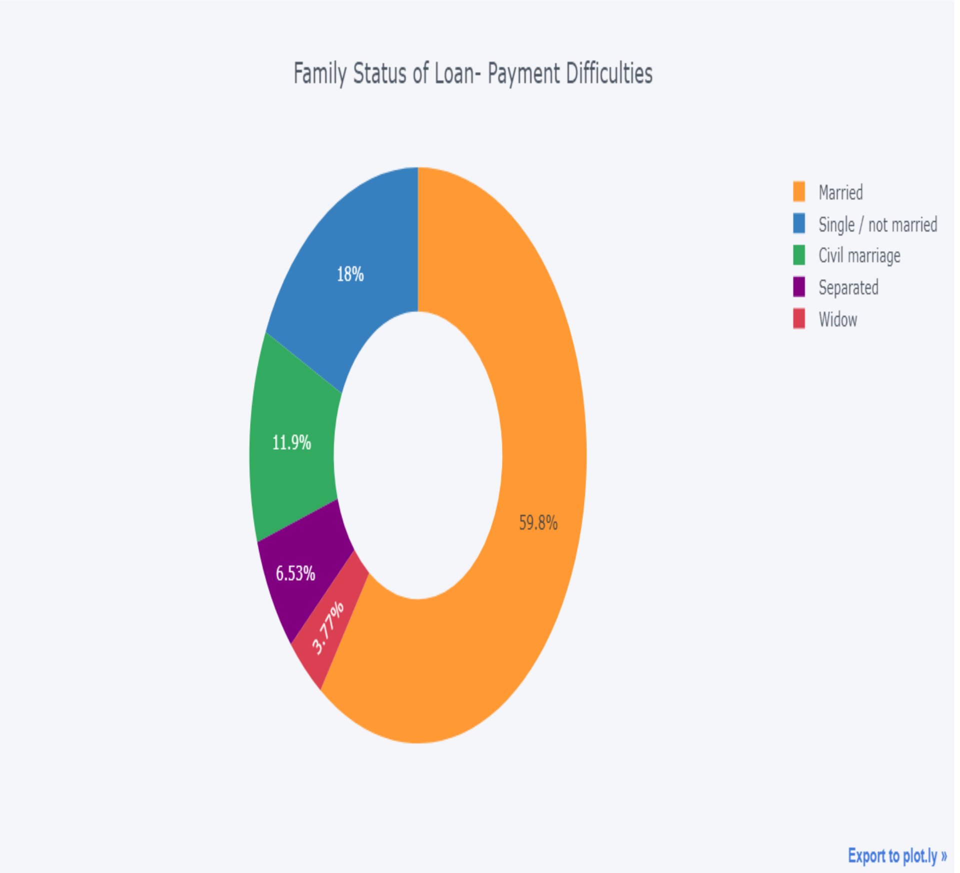




Family Status

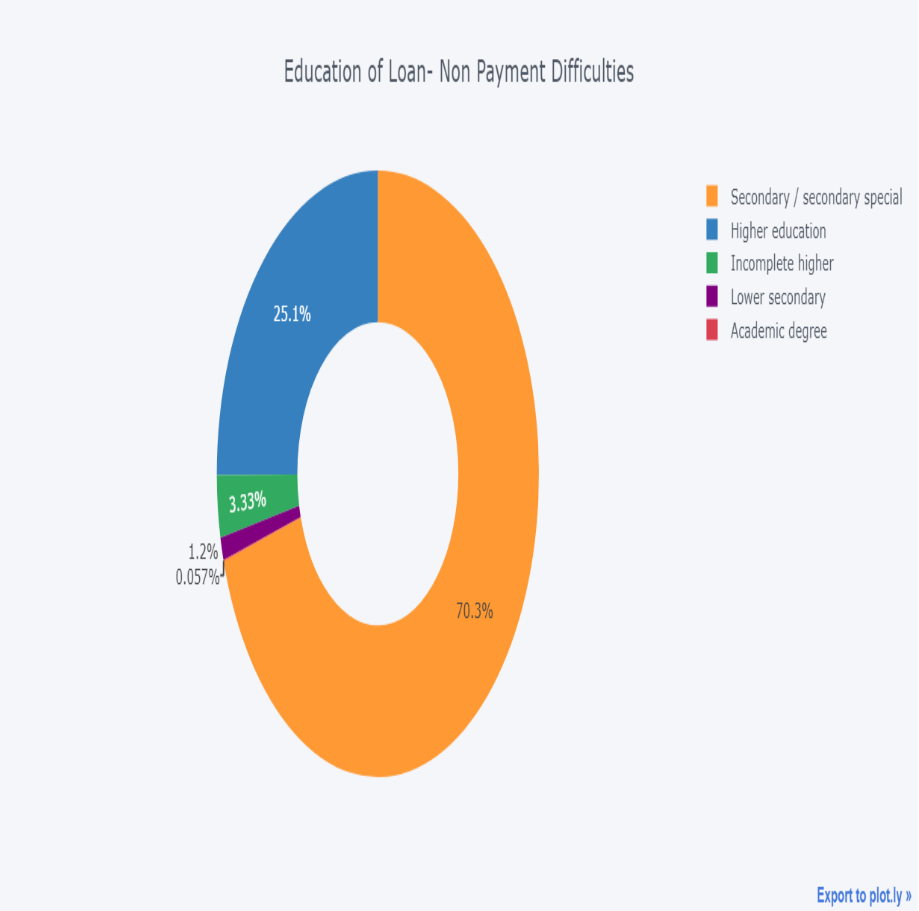
Inference: We observe a decrease in the percentage of married and widowed with Loan Payment Difficulties and an increase in the percentage of single and civil married with Loan Payment Difficulties when comapred with the percentages of both Loan Payment Difficulties and Loan Non-Payment Difficulties

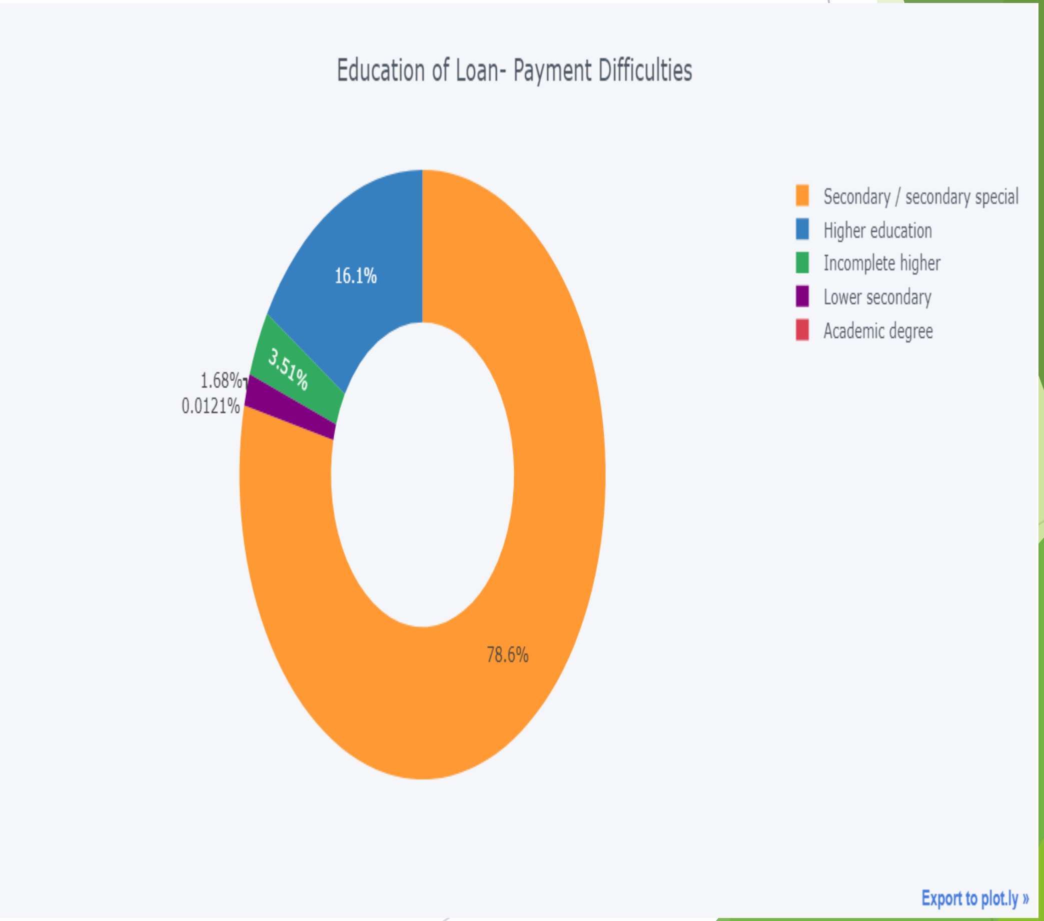




Education

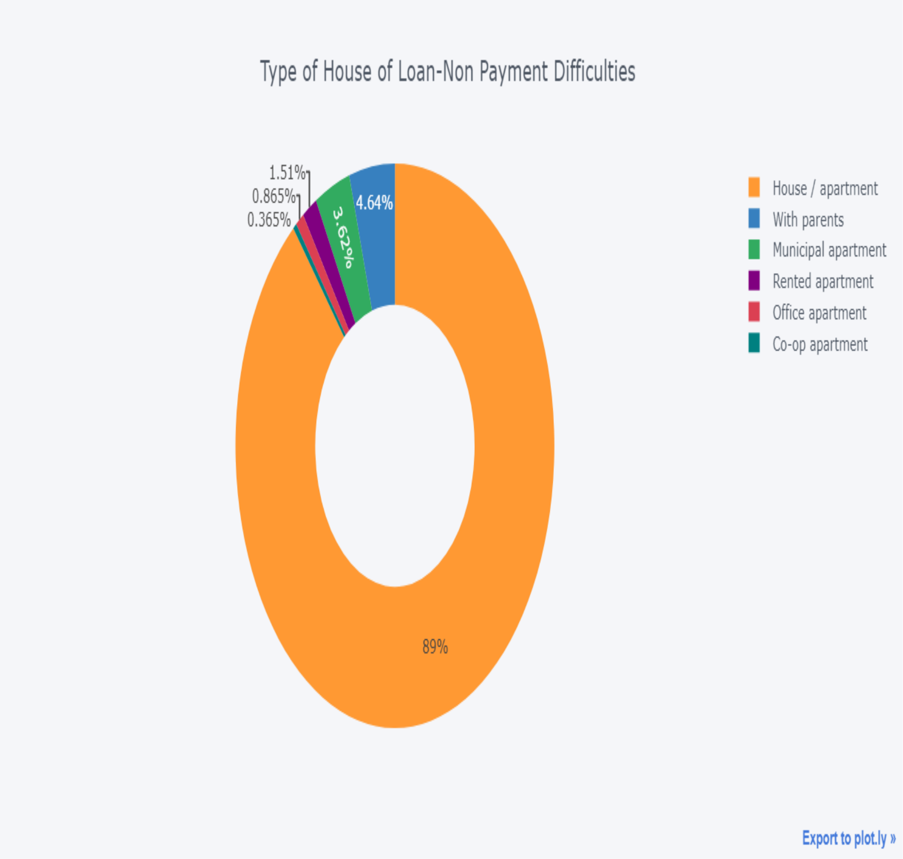
Inference: We observe an increase in percentage of Loan Payment Difficulties whose educational qualifications are secondary/secondary special and a decrease in the percentage of Loan Payment Difficulties who have completed higher education when compared with the percentages of Loan Payment Difficulties and Loan Non-Payment Difficulties

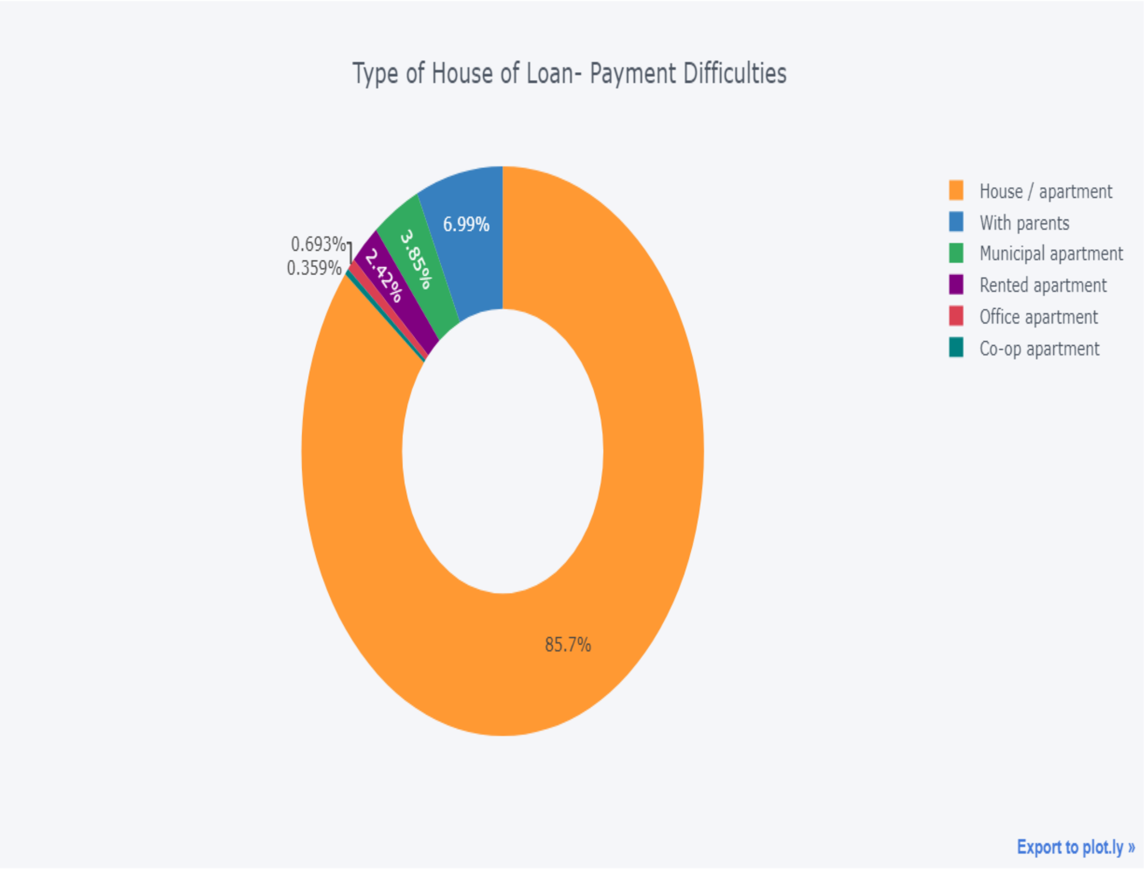




Housing

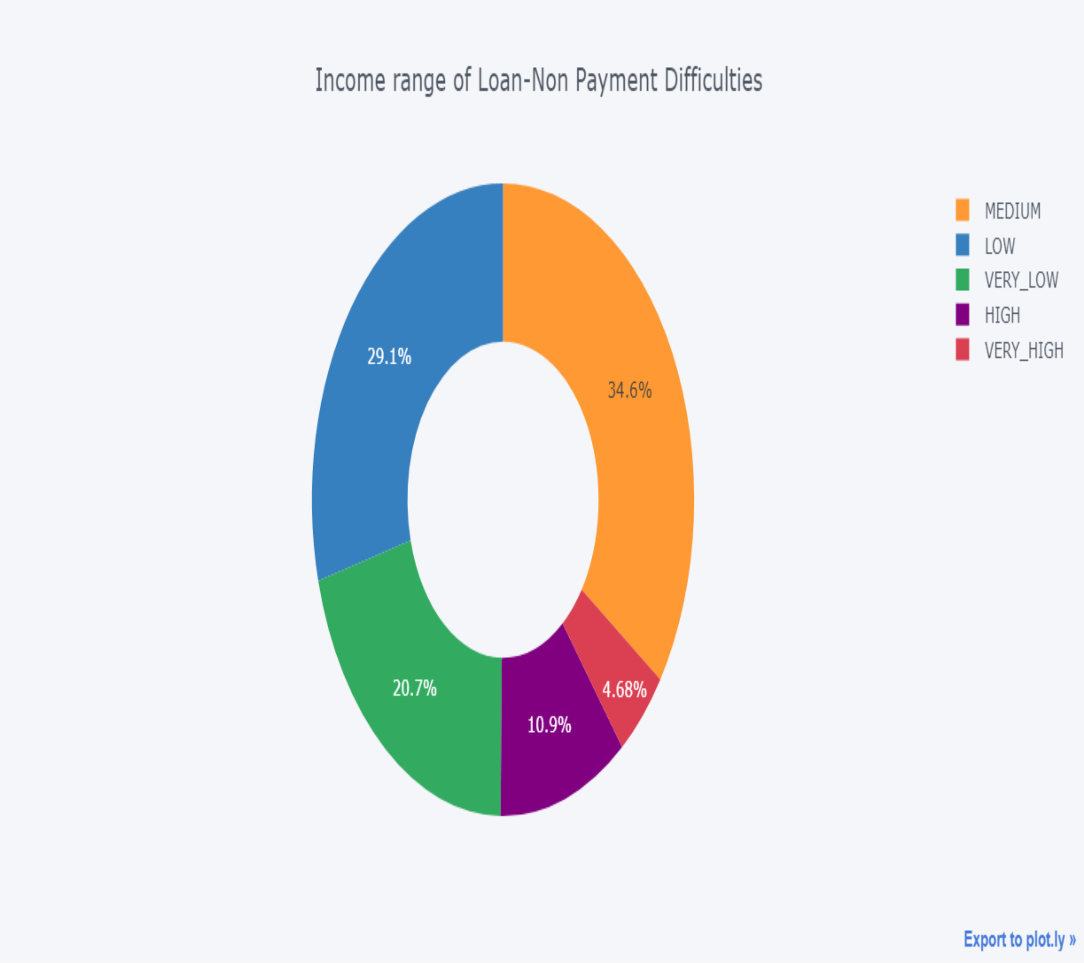
Inference: We observe an increase in the percentage of Payment Difficulties who live with their parents when compared to the
percentages of Payment Difficulties and non-Payment Difficulties

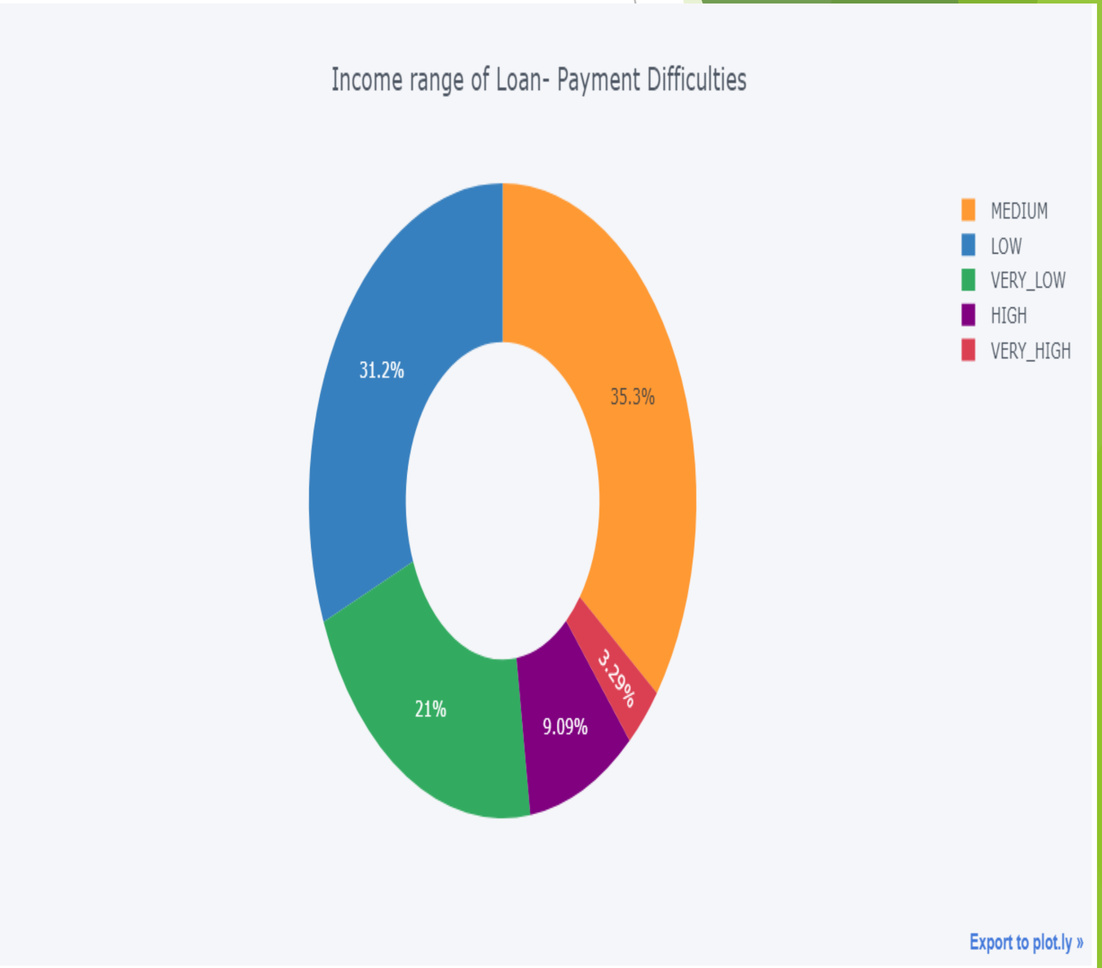




Income Range

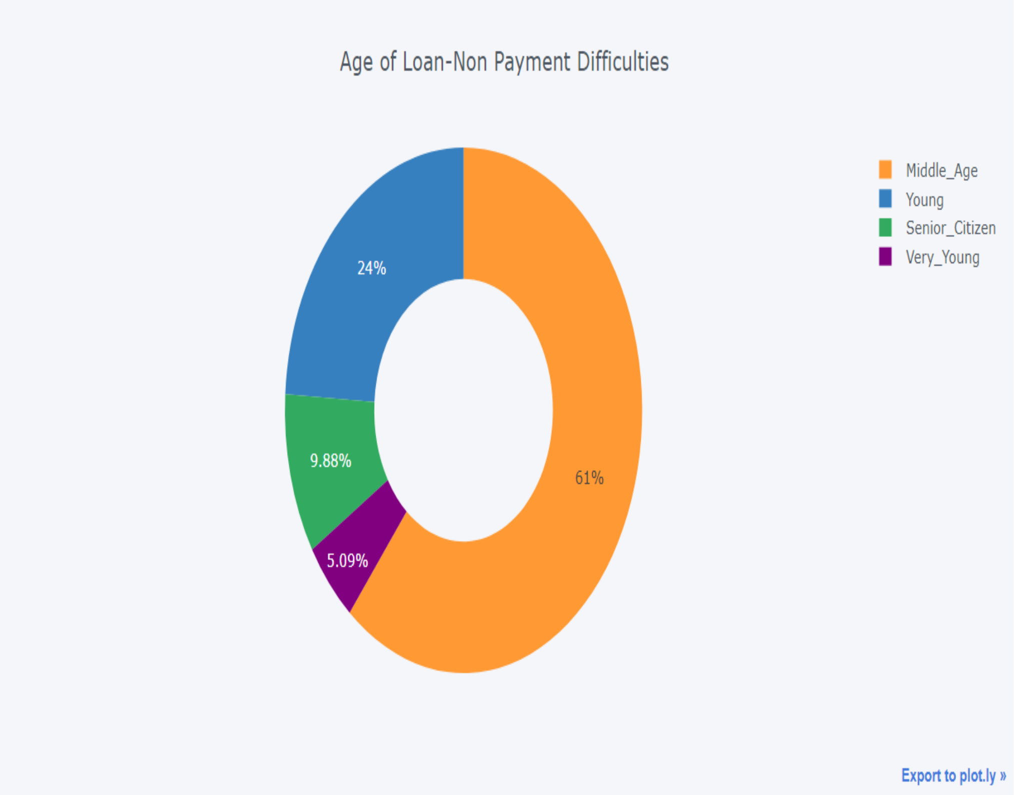
Inference: We can see that an increase in the percentage of Loan Payment Difficulties whose income is low when compared with the percentages of Payment Difficulties and Loan-Non Payment Difficulties

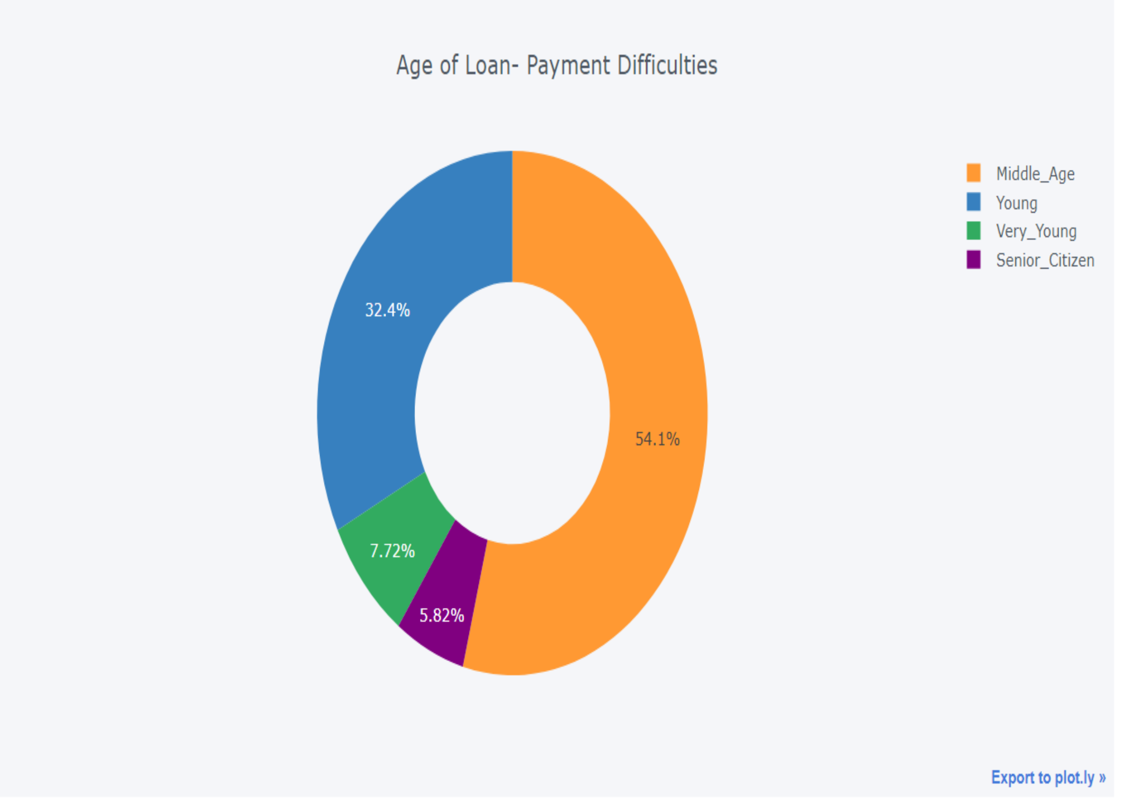




Age

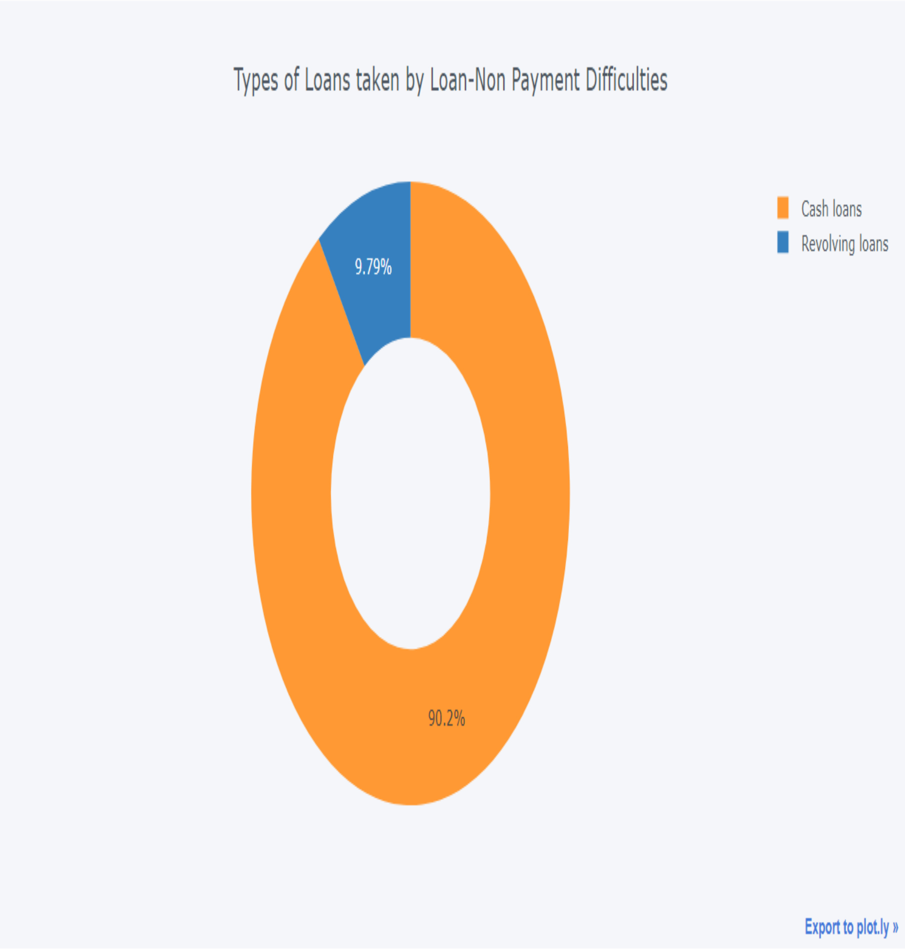
Inference: We observe that there is an increase in the percentage of Loan Payment Difficulties who are young in age when
compared to the percentages of Payment Difficulties and Loan-Non Payment Difficulties.



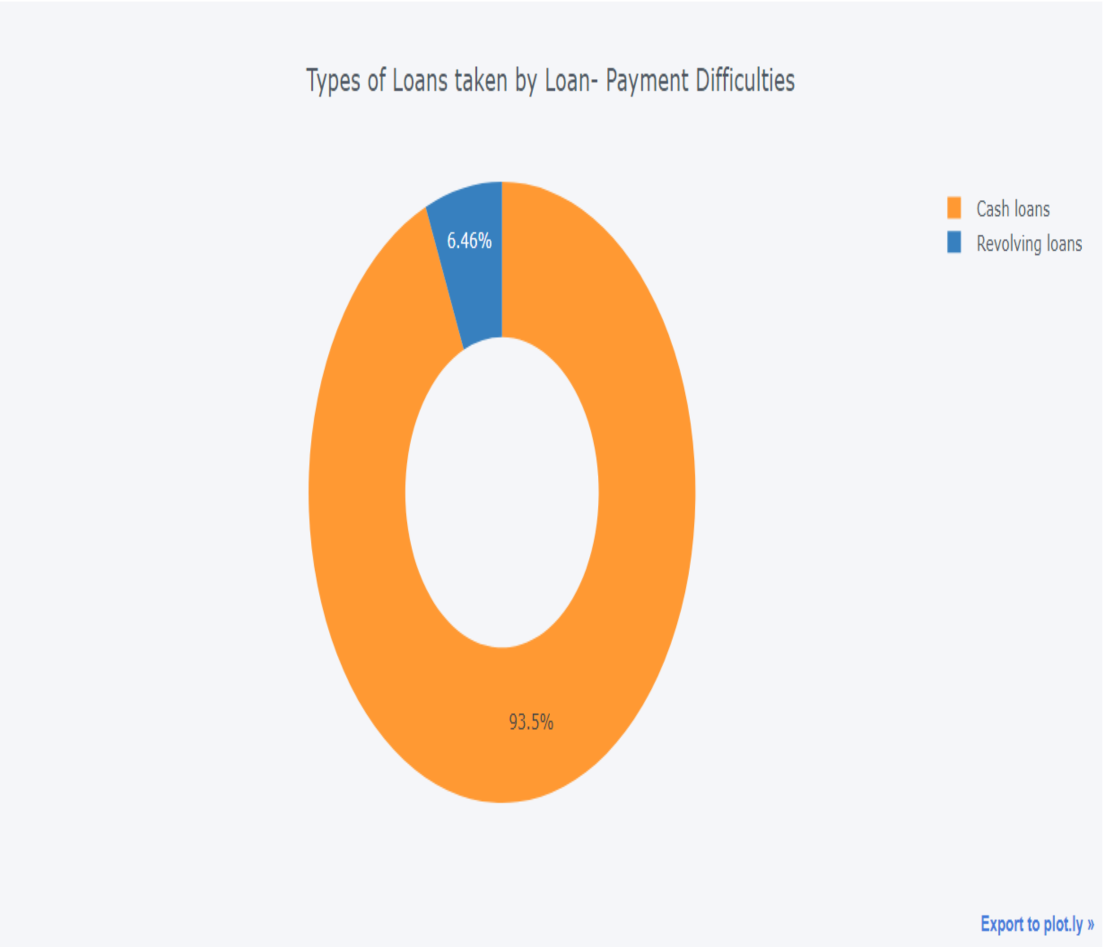


Loan Types

Inference: It can be seen cash loans are preferred by both Loan Payment and Non payment difficulties and also there is decrease in the percentage of payment difficulties who opt for Revolving Loans

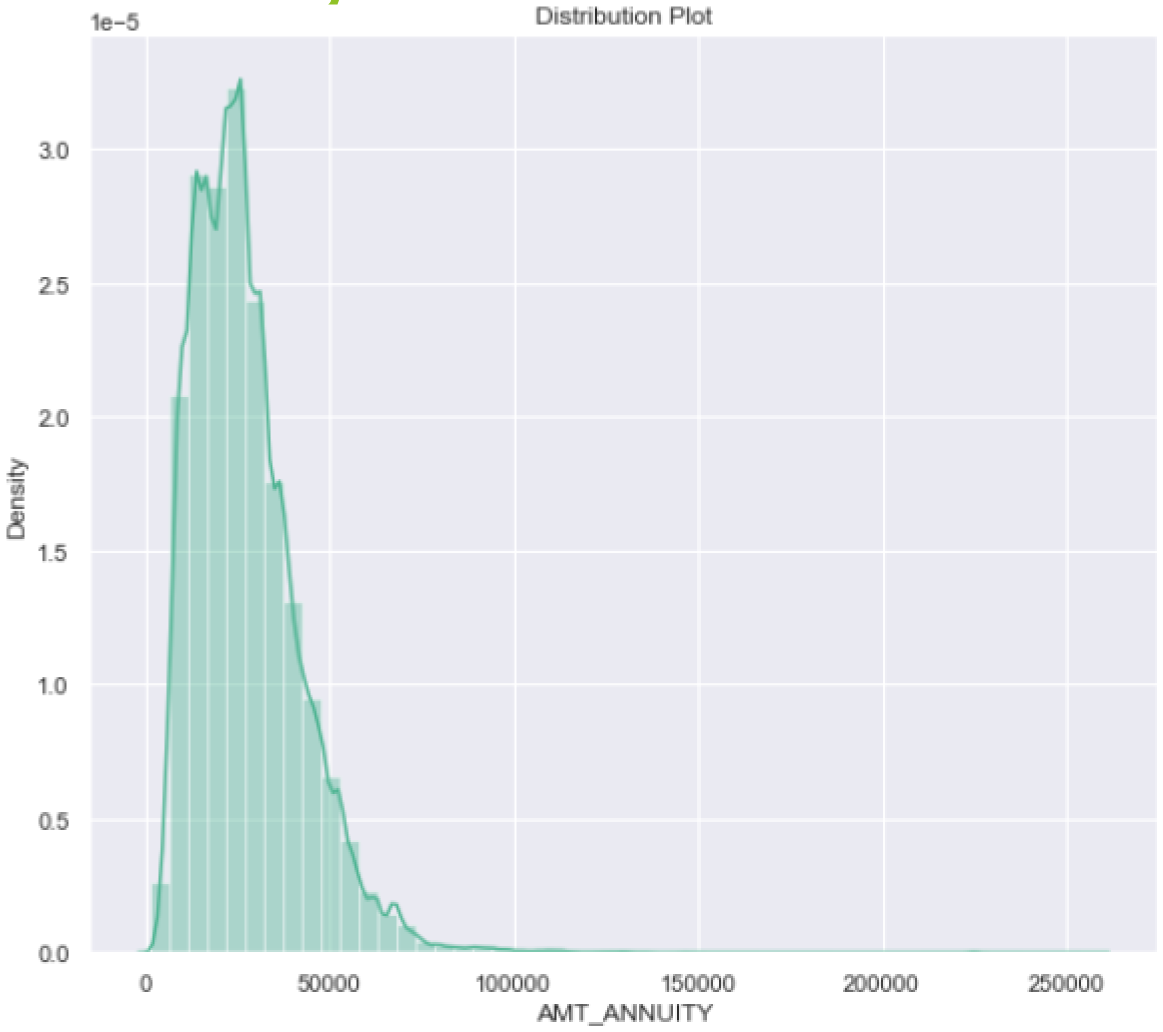


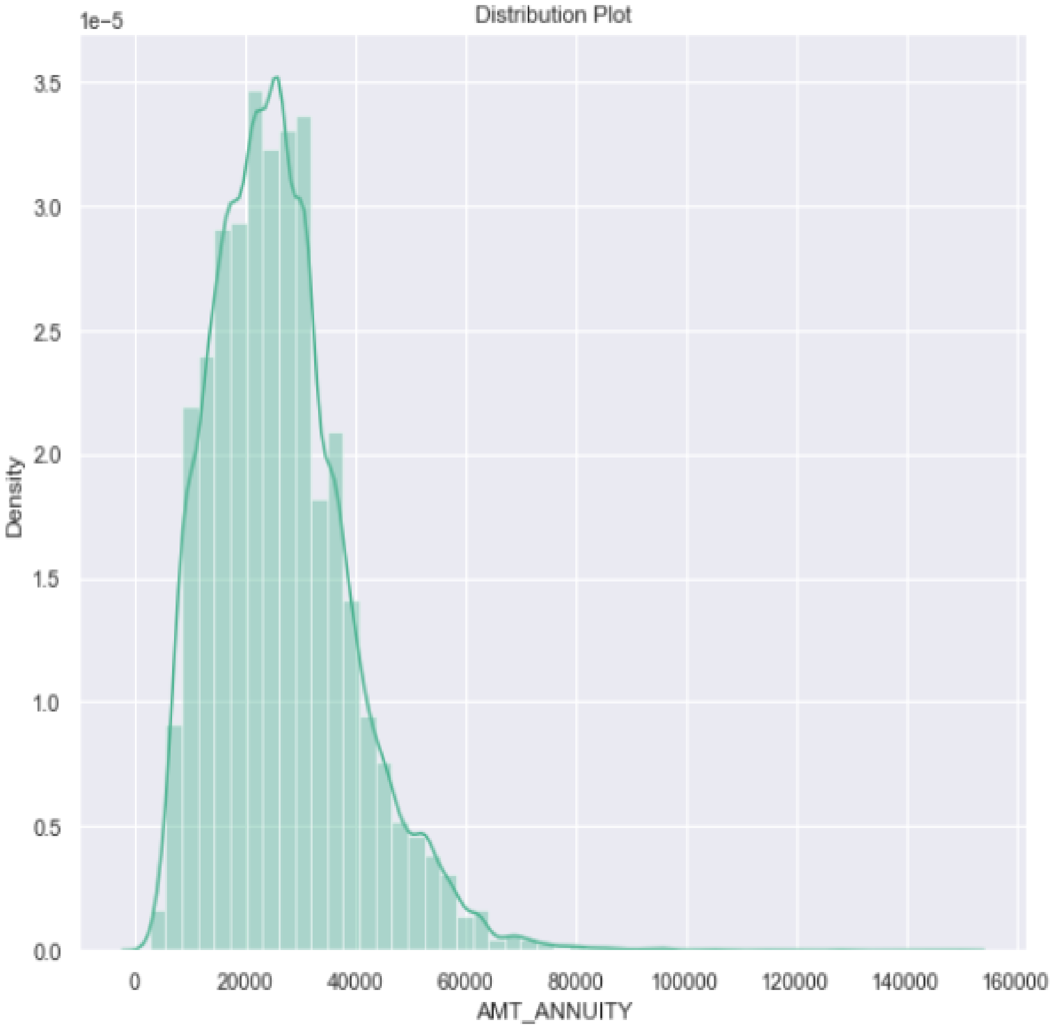
Types of Loans taken by Loan-Non Payment Difficulties



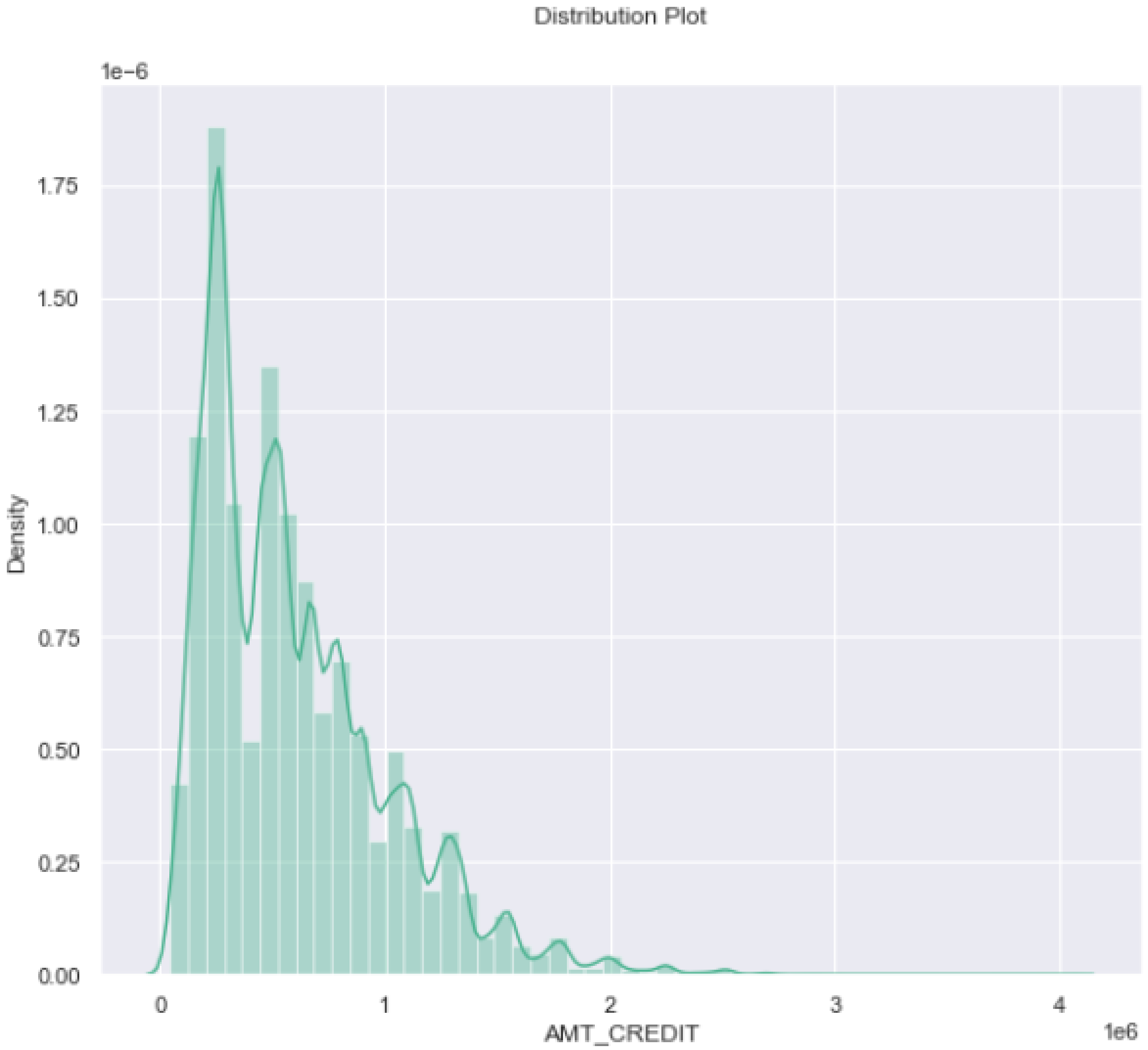
Types of Loans taken by Loan- Payment Difficulties

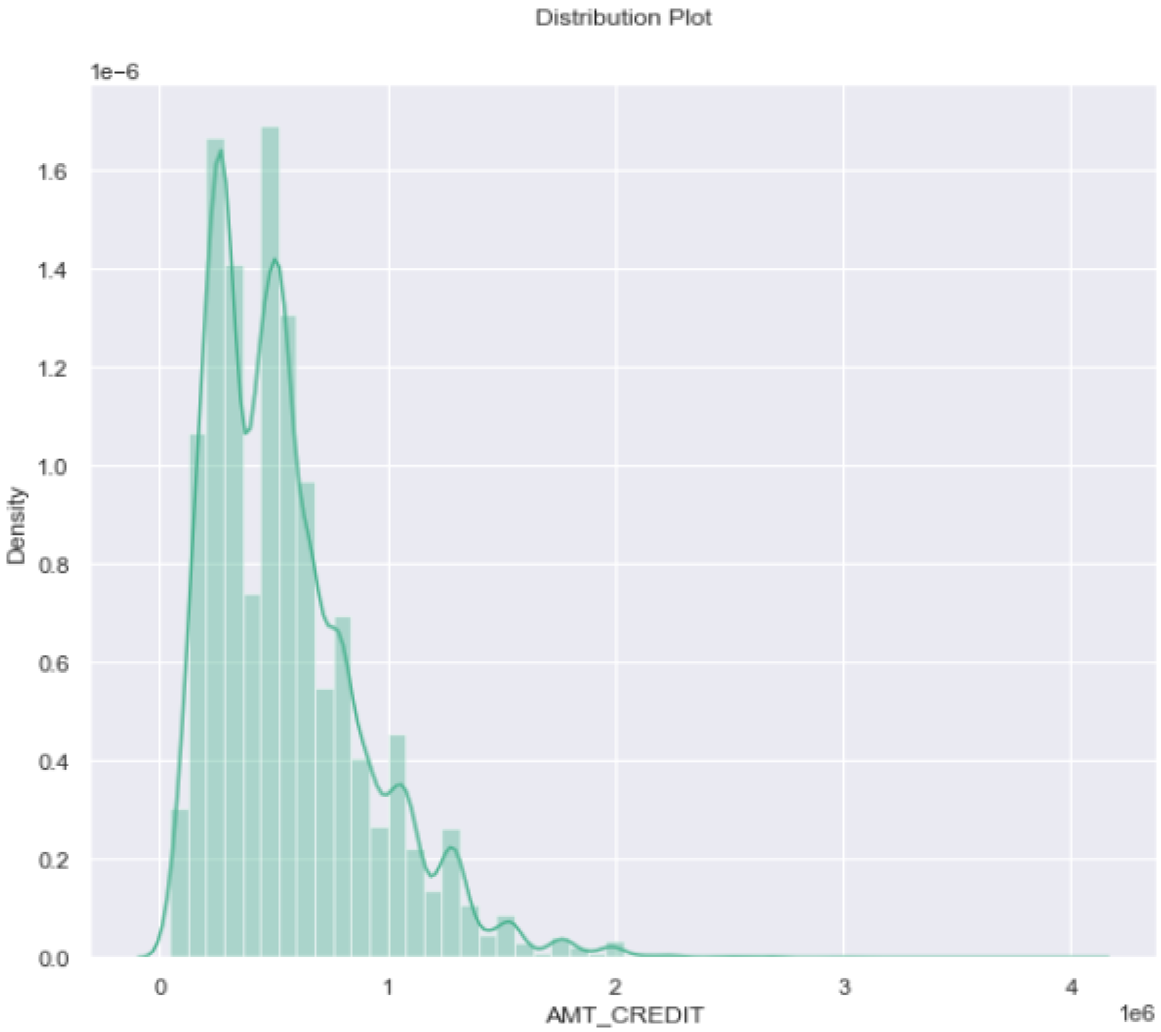
Univariate Numerical Variables: Loan Annuity



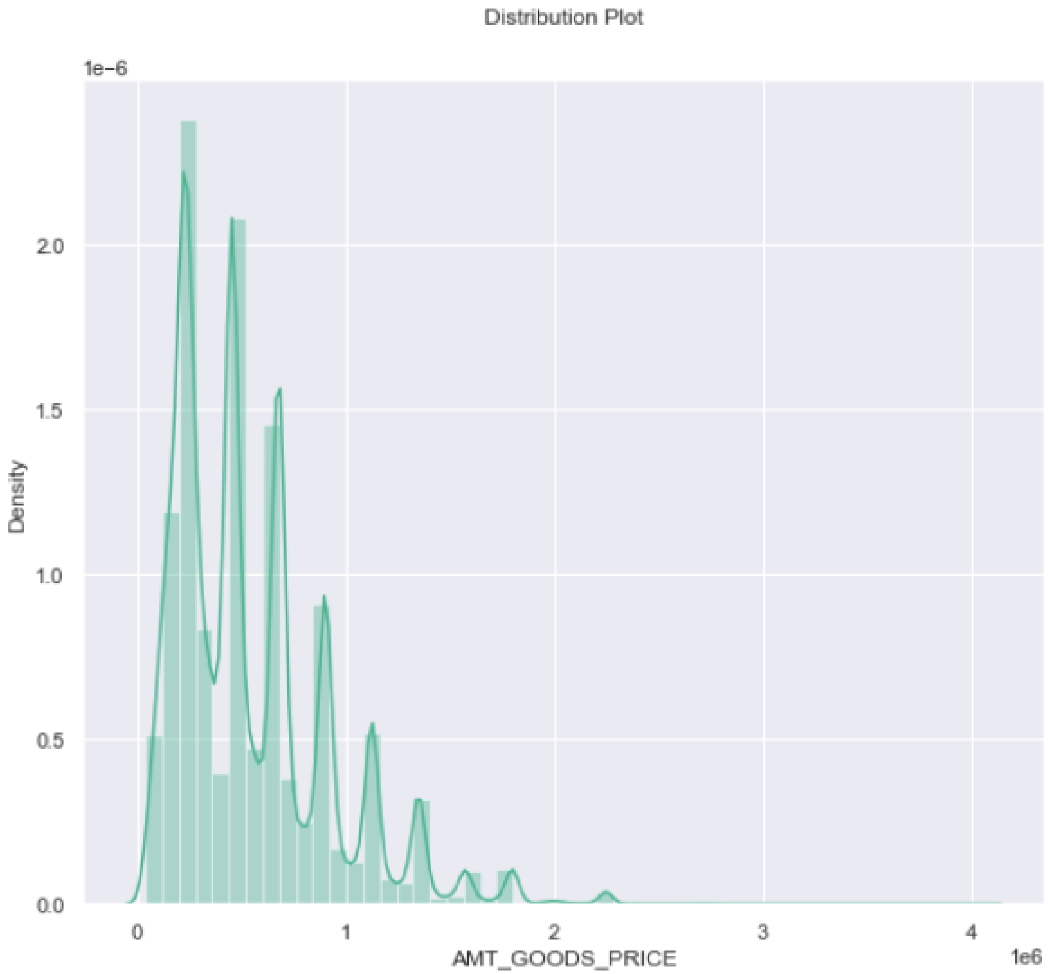


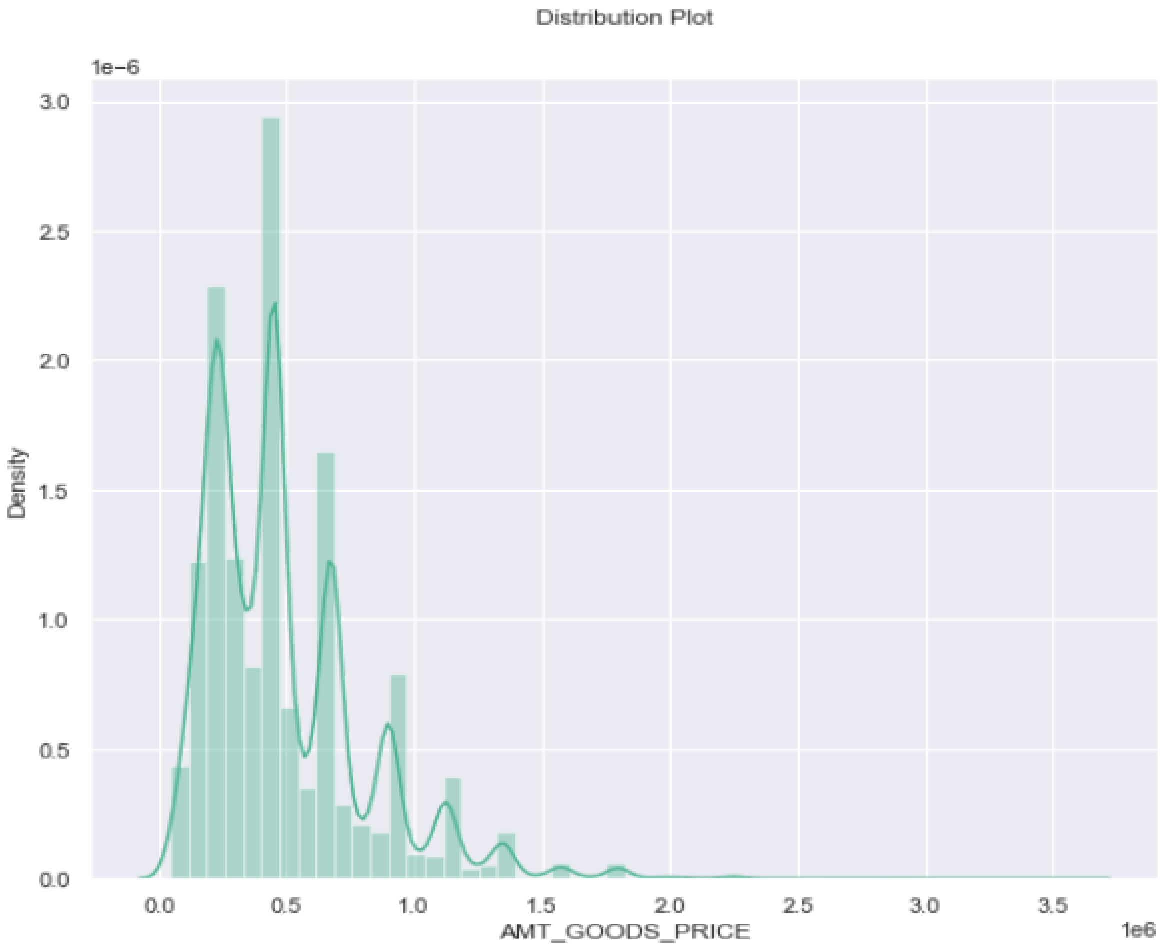
CREDIT AMOUNT





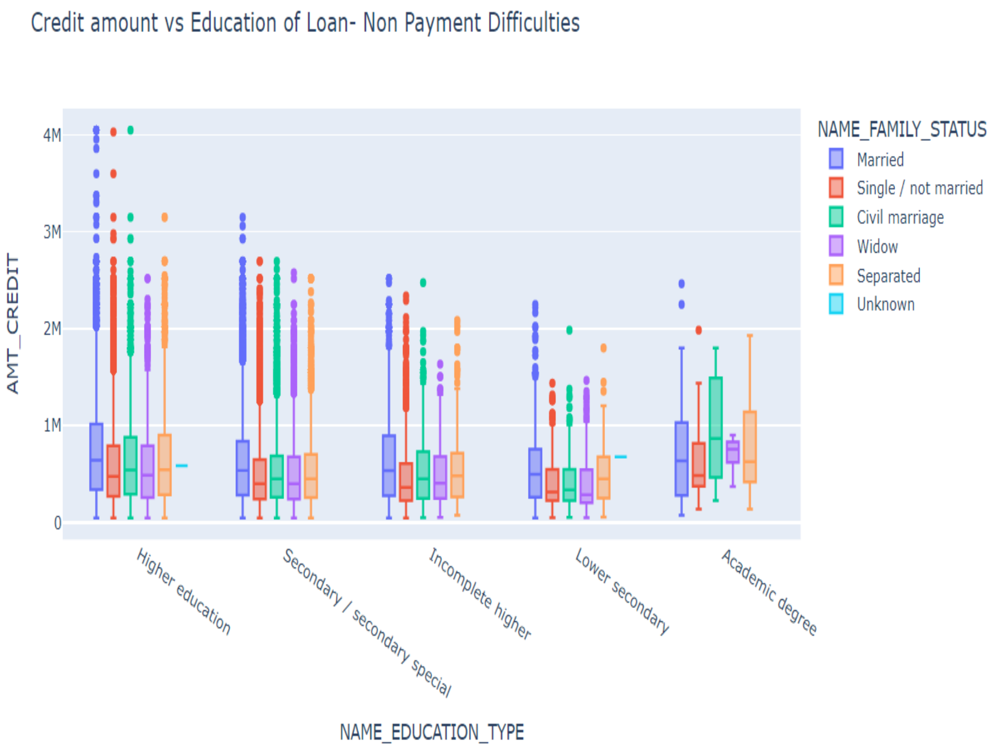
Goods Price

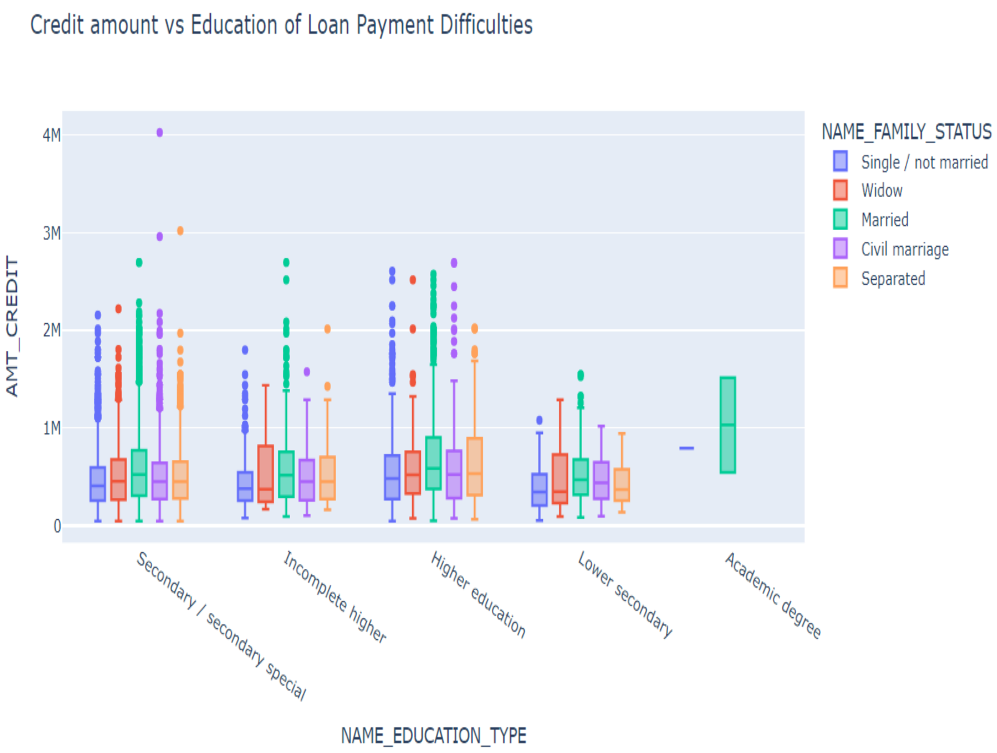

Distribution Plot
1e−6
2.0
1.5
1.0
0.5
0.0
Density
0
1
2
3
4
AMT_GOODS_PRICE
1e6


Distribution Plot
1e−6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
AMT_GOODS_PRICE
1e6

Bivariate Analysis: Analysis - 1

INFERENCE: The graphs for Loan Payment Difficulties and Loan Non-Payment Difficulties apperas to be similar. We observe that Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others. Most of the outliers are from Education type 'Higher education' and 'Secondary'. Civil marriage for Academic degree is having most of the credits in the third quartile.

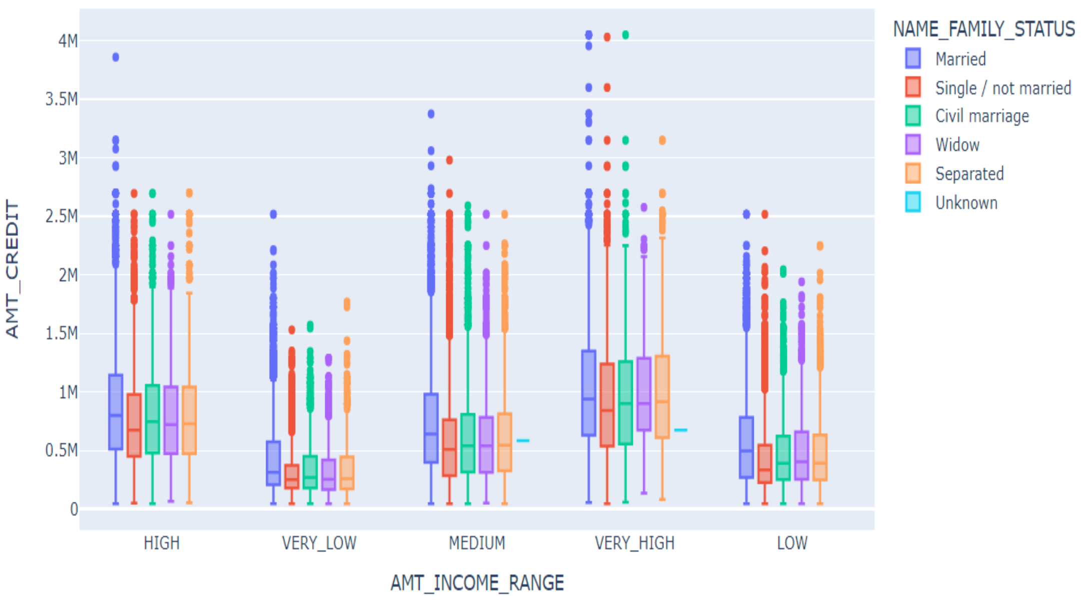




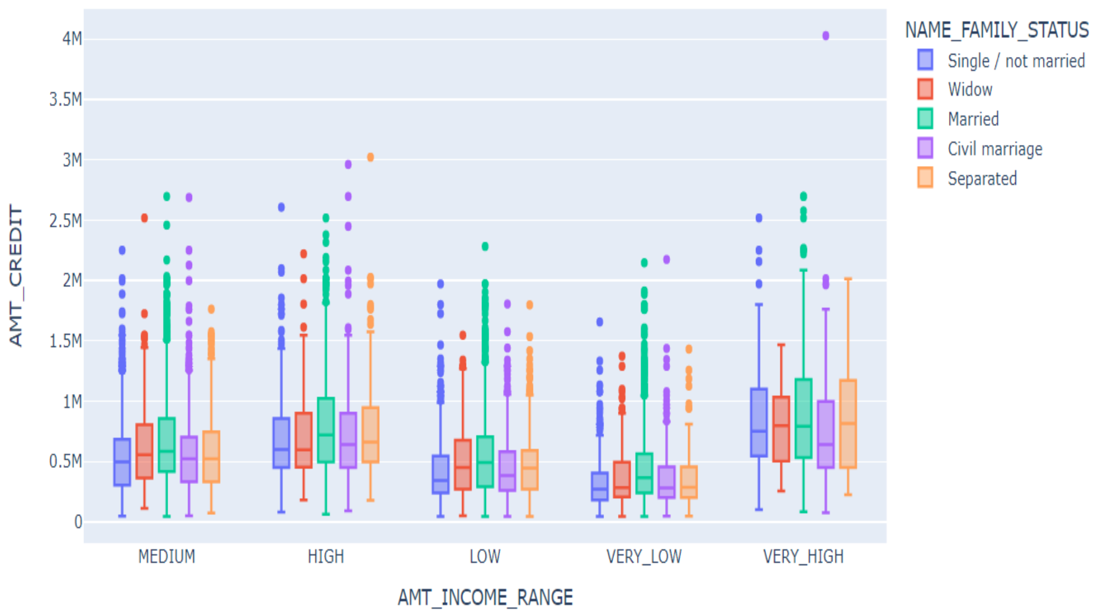
Bivariate Analysis – Analysis 2

INFERENCE: The graphs for Loan Payment Difficulties and Loan- Non Payment Difficulties apperas to be similar. We observe that Family status of 'single', 'seperated' and 'married' of income range veryhigh are having higher number of credits than others.

Income range vs Credit amount of Loan Non- Payment Difficulties

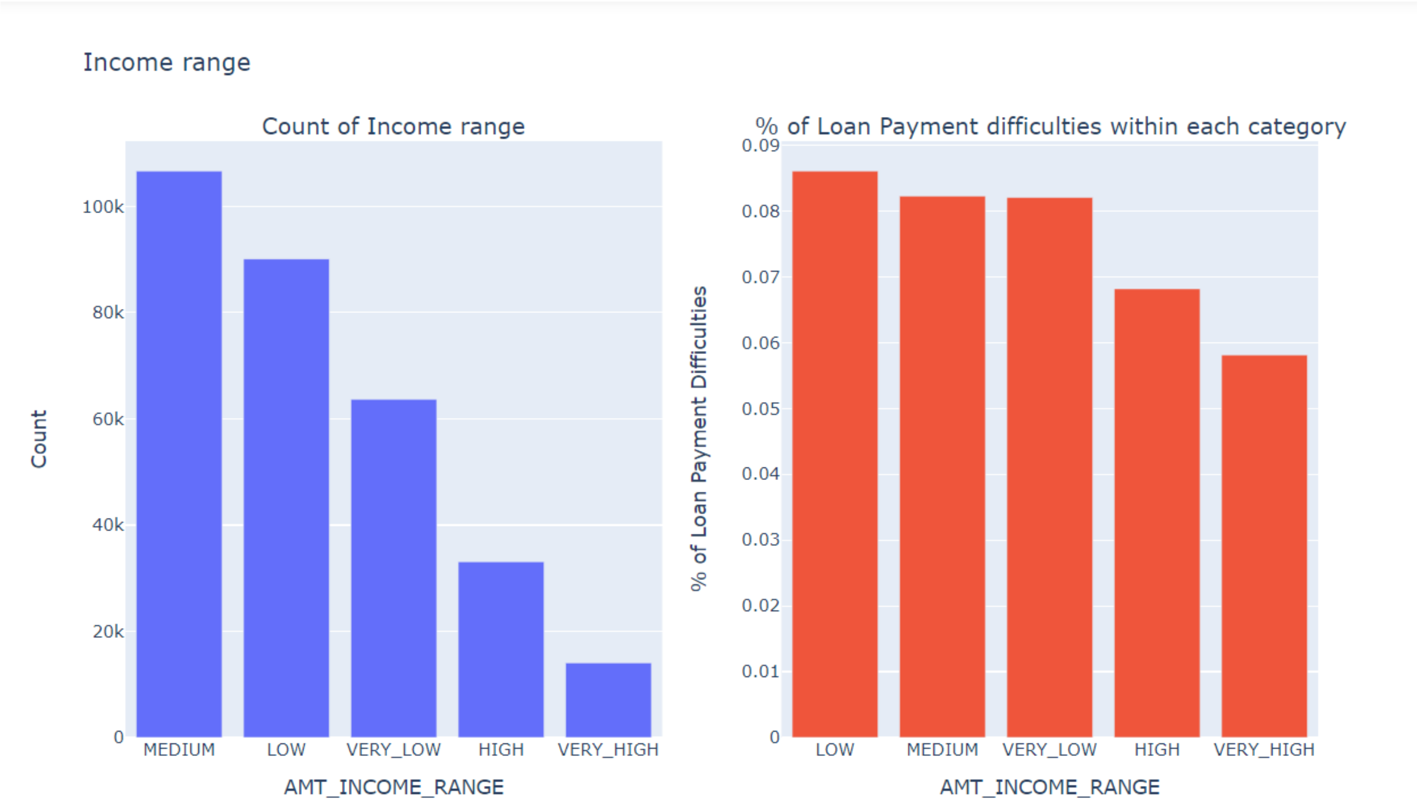


Income range vs Credit amount of Loan Payment Difficulties



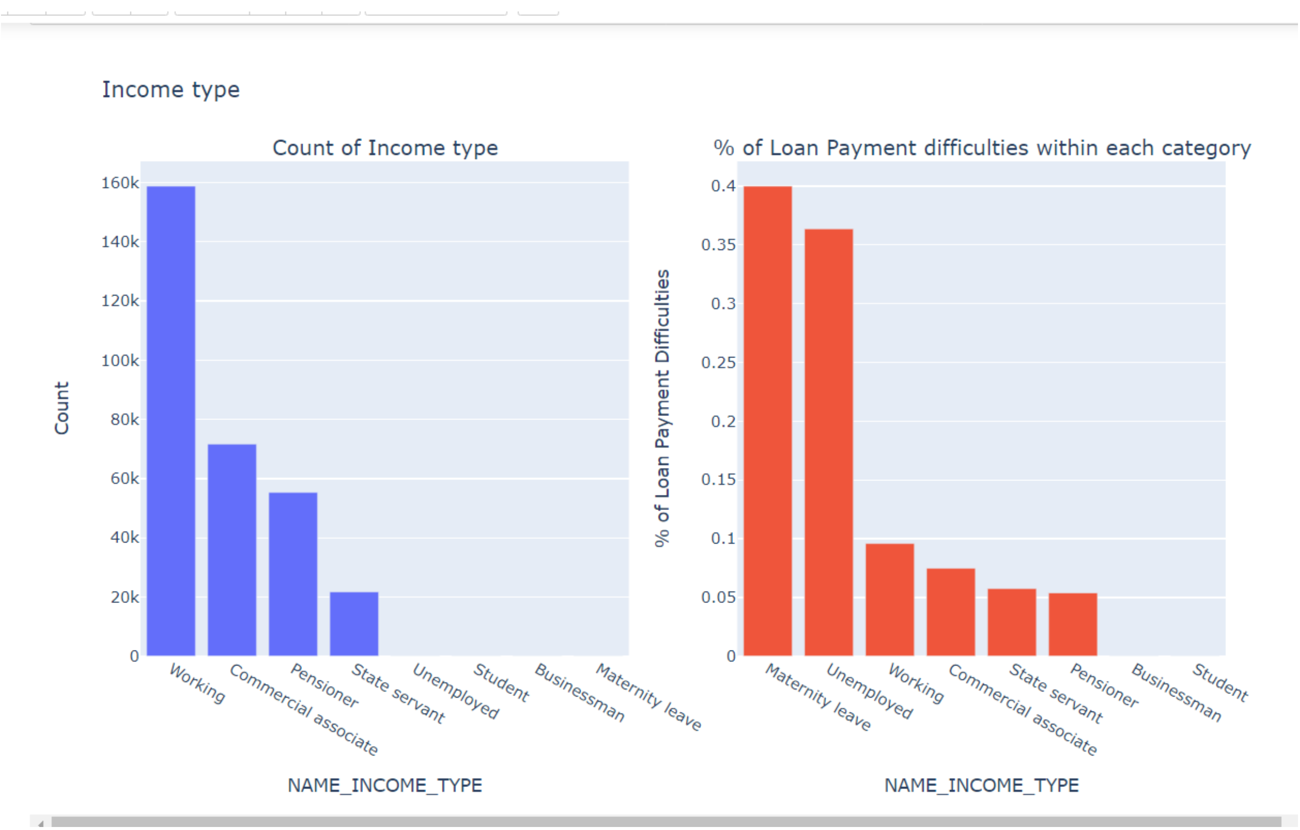
Bivariate Analysis Categorical – Income Range

INFERENCE: From the plot above we can say that clients with 'LOW' Income range have maximum % of Loan-Payment Difficulties.



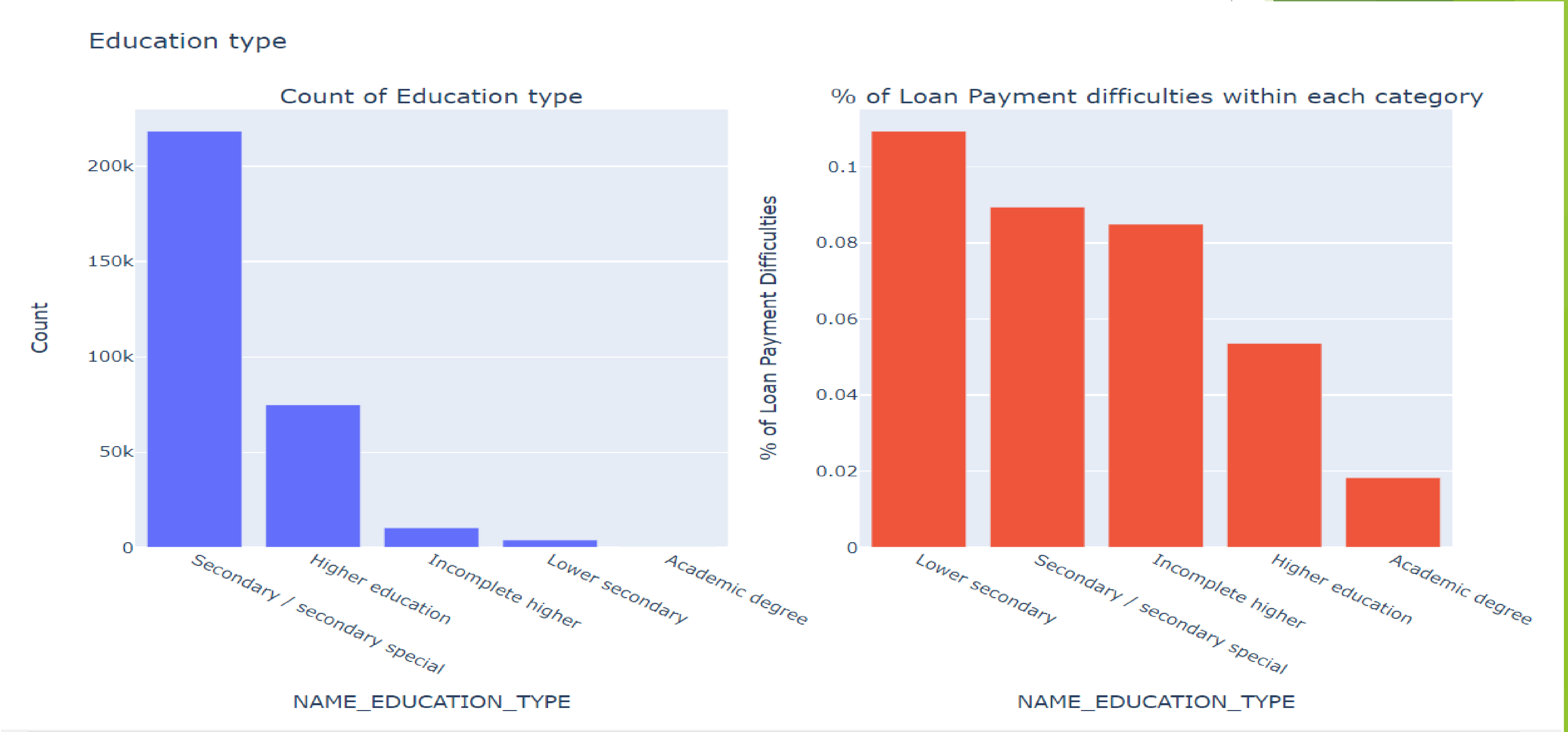
Income Type

From the plot above we can say that clients with 'Maternity leave' Income type have maximum % of Loan-Payment Difficulties.



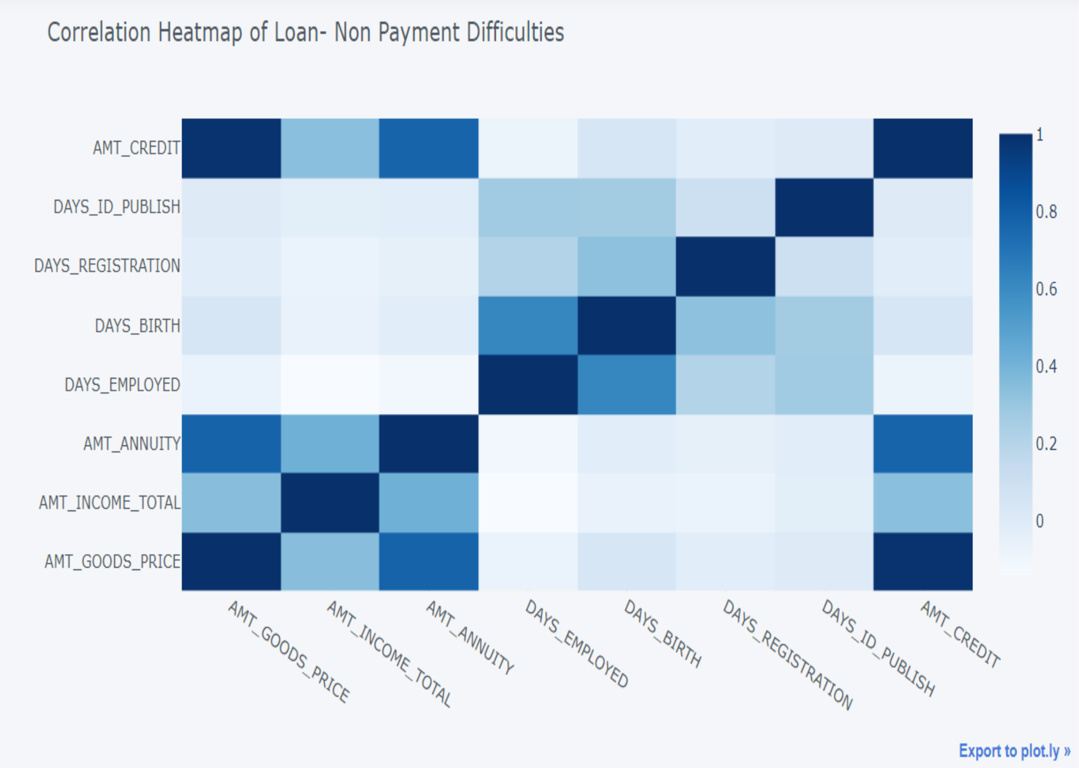
Education Type

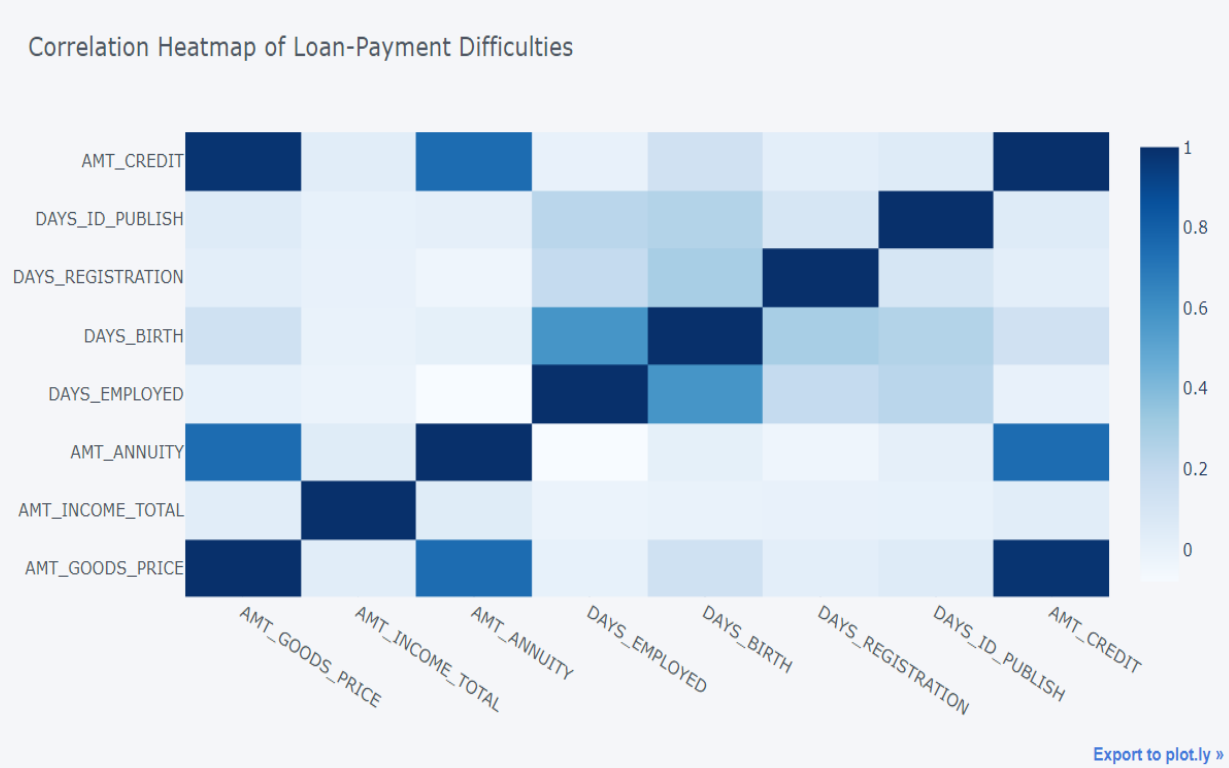
From the plot above we can say that clients with 'Lower secondary' education type have maximum % of Loan-Payment Difficulties.



Correlation

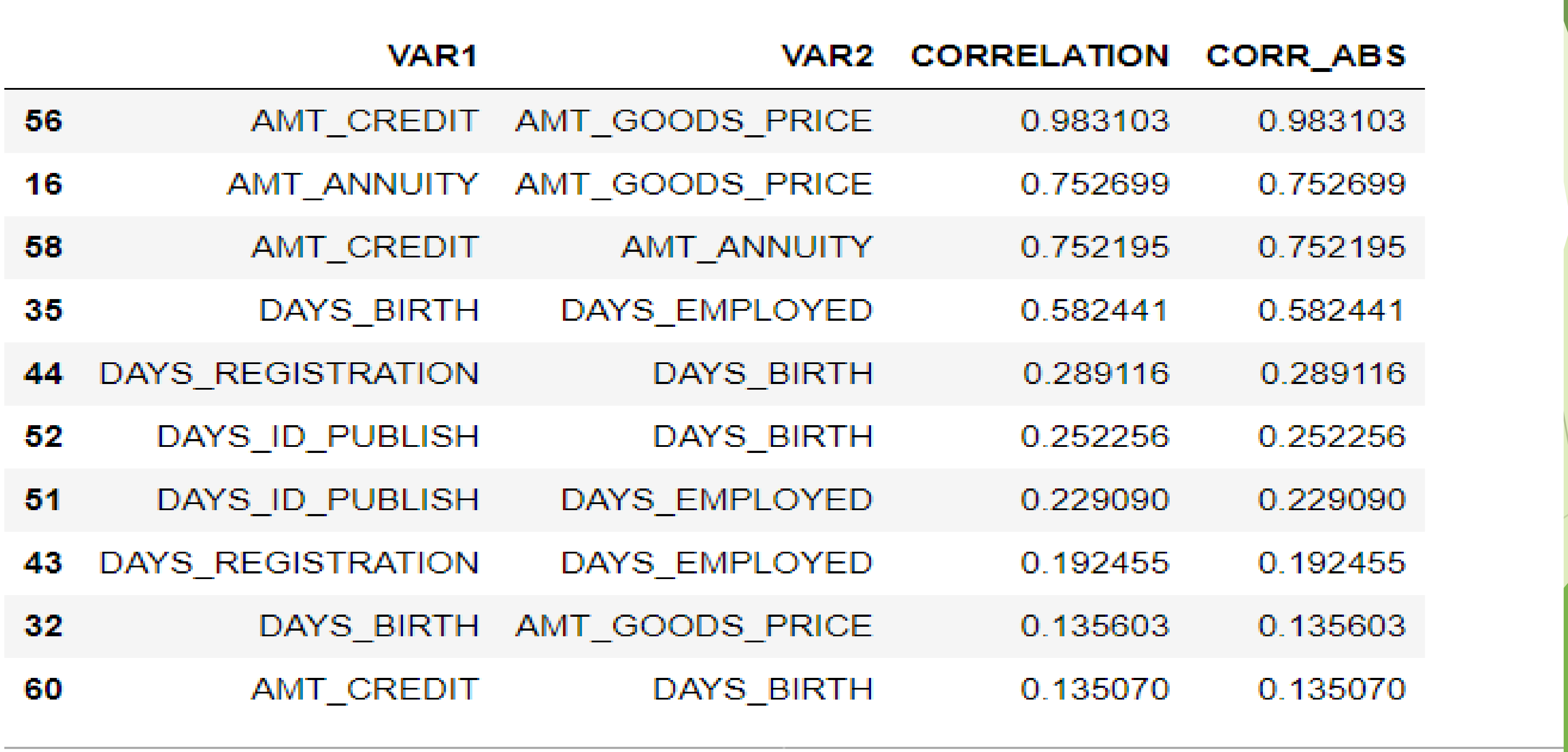
INFERENCE: We observe that there is a high correlation between credit amount and goods price. There appears to be some deviancies in the correlation of Loan-Payment Difficulties and Loan- Non Payment Difficulties such as credit amount v/s income.





Top 10 Correlation

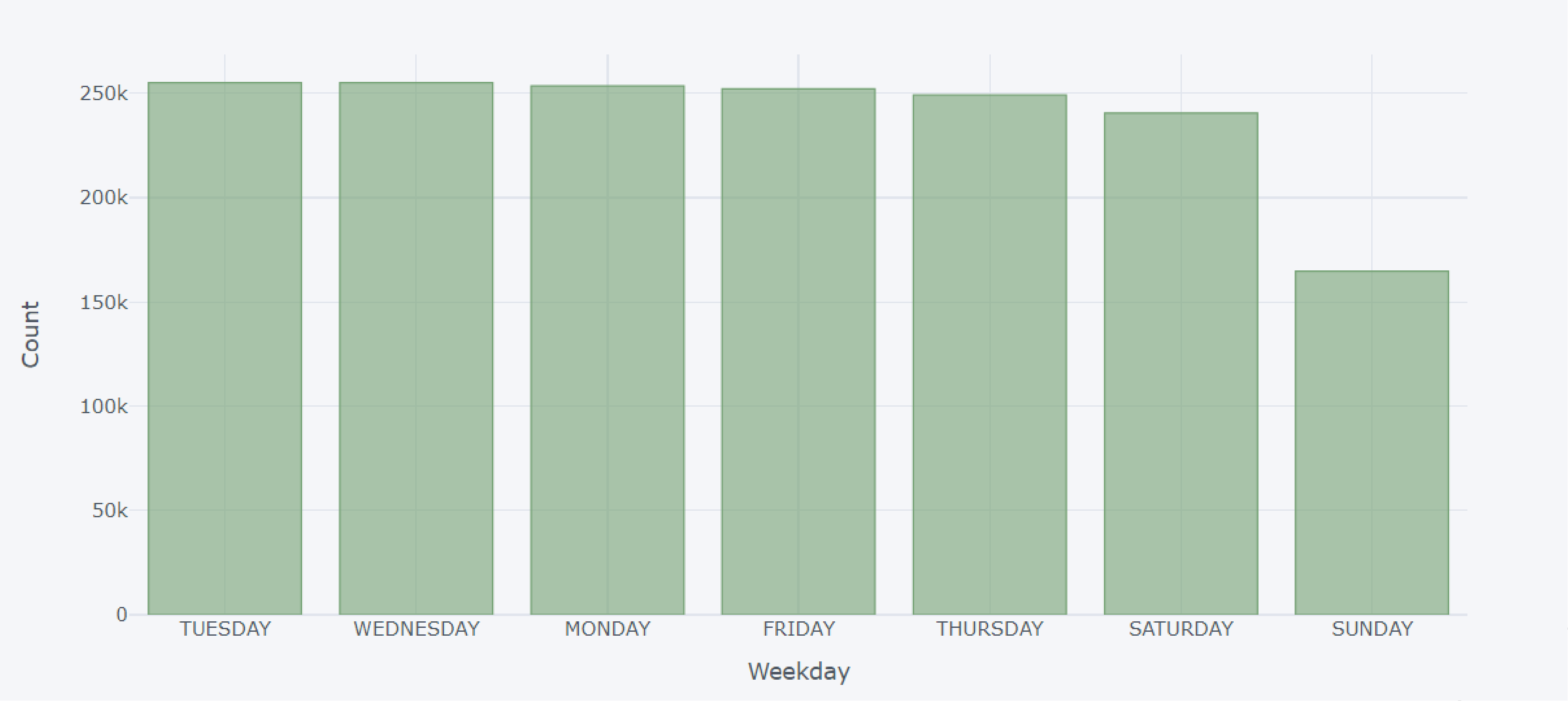
INFERENCE: The below table shows top 10 correlation for clients with payment difficulties



	VAR1	VAR2	CORRELATION	CORR_ABS
56	AMT_CREDIT	AMT_GOODS_PRICE	0.983103	0.983103
16	AMT_ANNUITY	AMT_GOODS_PRICE	0.752699	0.752699
58	AMT_CREDIT	AMT_ANNUITY	0.752195	0.752195
35	DAYS_BIRTH	DAYS_EMPLOYED	0.582441	0.582441
44	DAYS_REGISTRATION	DAYS_BIRTH	0.289116	0.289116
52	DAYS_ID_PUBLISH	DAYS_BIRTH	0.252256	0.252256
51	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.229090	0.229090
43	DAYS_REGISTRATION	DAYS_EMPLOYED	0.192455	0.192455
32	DAYS_BIRTH	AMT_GOODS_PRICE	0.135603	0.135603
60	AMT_CREDIT	DAYS_BIRTH	0.135070	0.135070

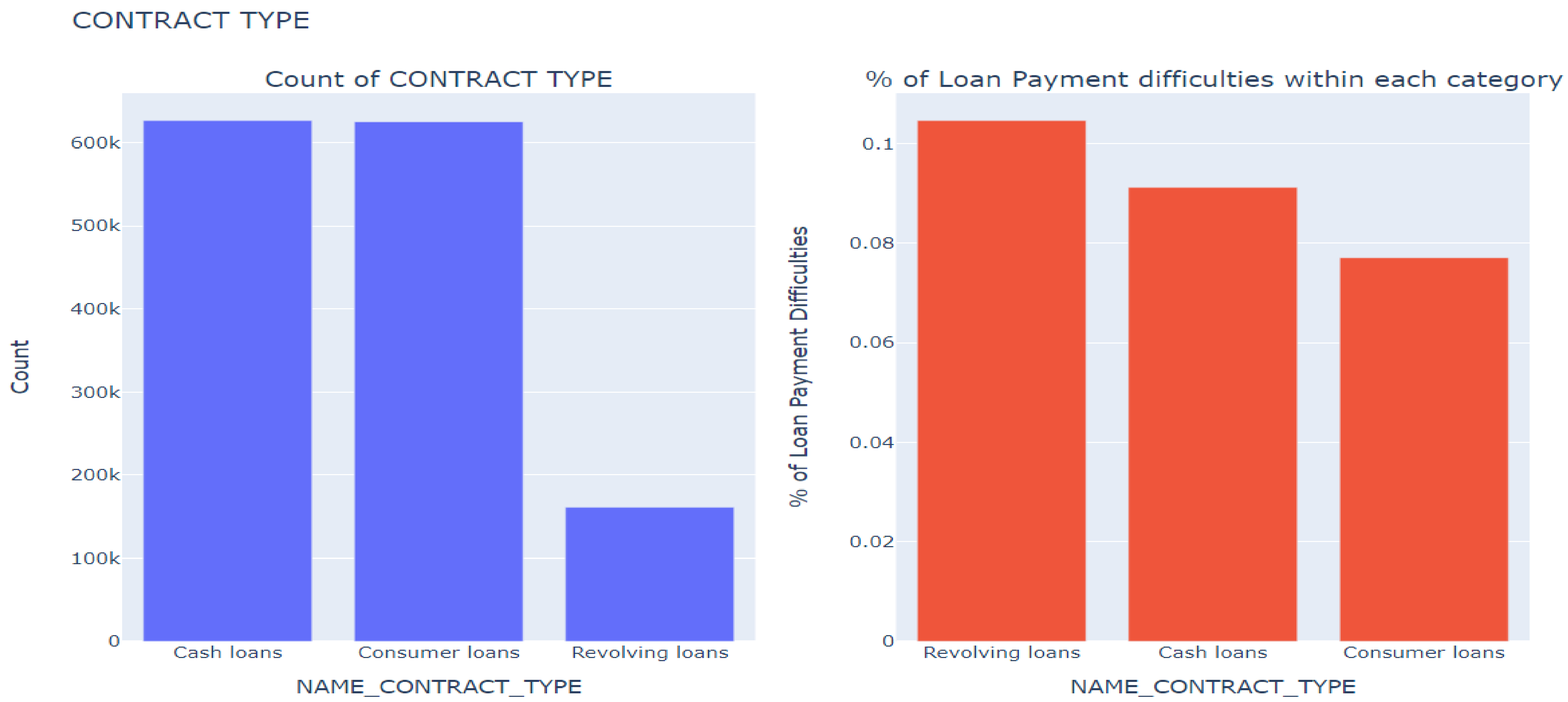
Days of week where people applied for loans

INFERENCE: LESS NUMBER OF PEOPLE APPLIED FOR LOANS ON WEEKENDS



Contract Type of previous application

Inference: From the first graph it can be seen that most of the contract type from previous application was 'Cash loans'



Conclusions

The data set case study reveals following information:

- The proportion of defaulters is 8.7%
- The bank lends more to females.
- More cash loans go into default. Hence more revolving loans should be used.
- Proportion of working defaults more.
- Old people as well as Higher educated people default less.
- Singles default more than married.
- Higher loans higher income less defaults.
- Loans previously cancelled or refused higher probability of default.

Thank You