EDA ON LOAN APPLICATION DATA

CASE STUDY FROM: SANJAY NAGARAJ

AGENDA

- Introduction & Problem Statement
- Approach
- Data Understanding
- Data Cleansing and Enrichment
- Data Analysis
- Summary and Recommendations

THE PROBLEM STATEMENT

Business

The Loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history

Some consumers use it as their advantage by becoming a defaulter.

Context

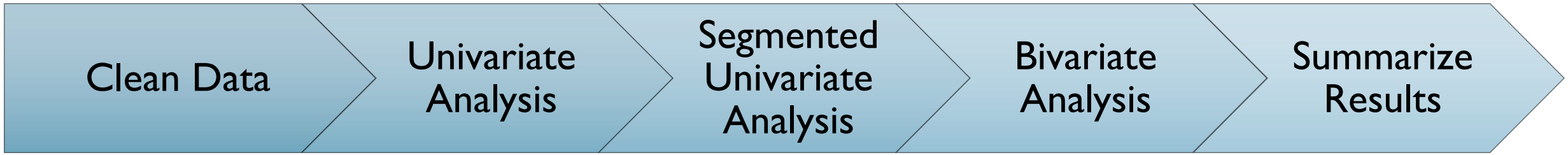
Finance company wants to understand driving factors for loan default through analysis

The company wants to use the analysis for risk assessment on loan application

Problem

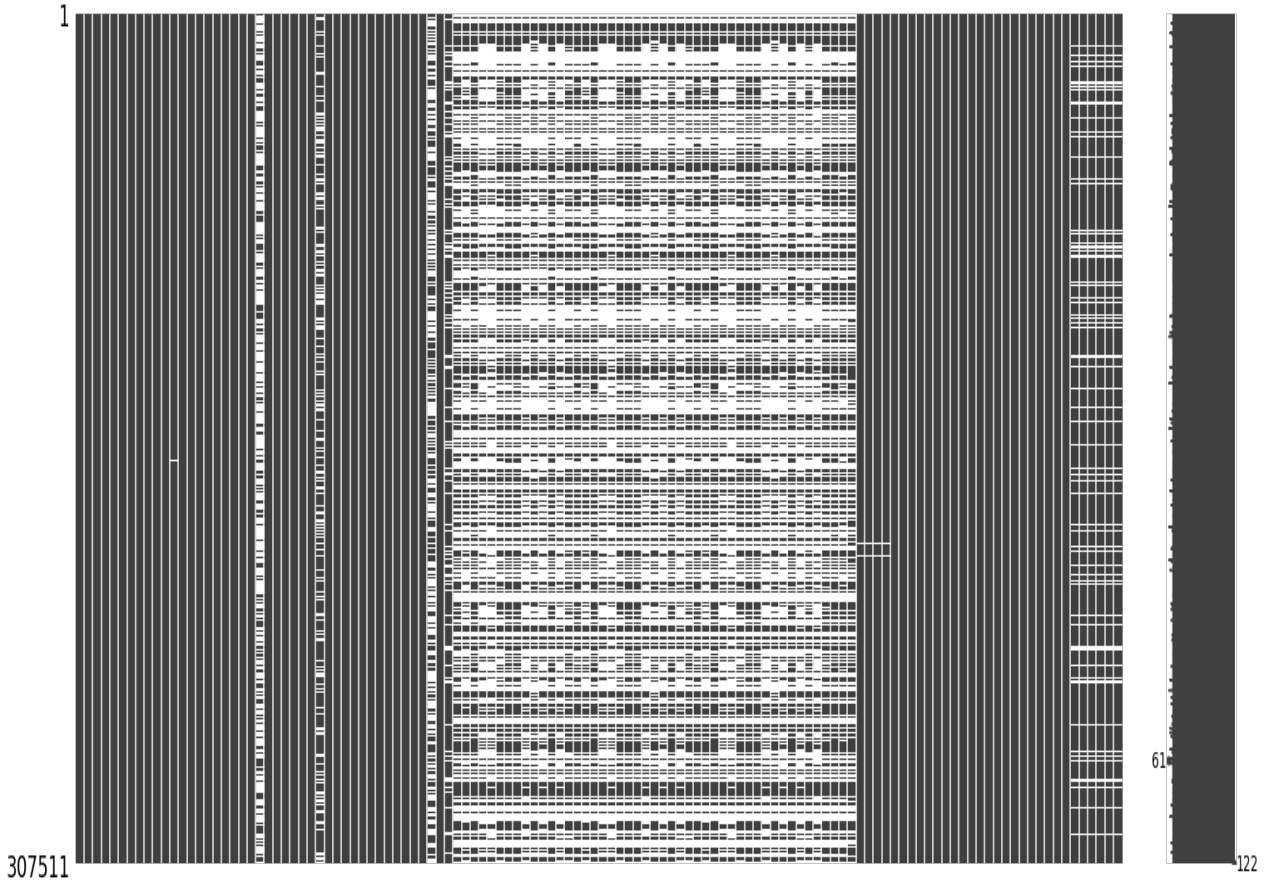
As a Data Science Engineer at this company, the problem is to analyse dataset containing loan application & previous application using EDA to understand how the attribute may affect the tendency of loan Default

ANALYSIS APPROACH

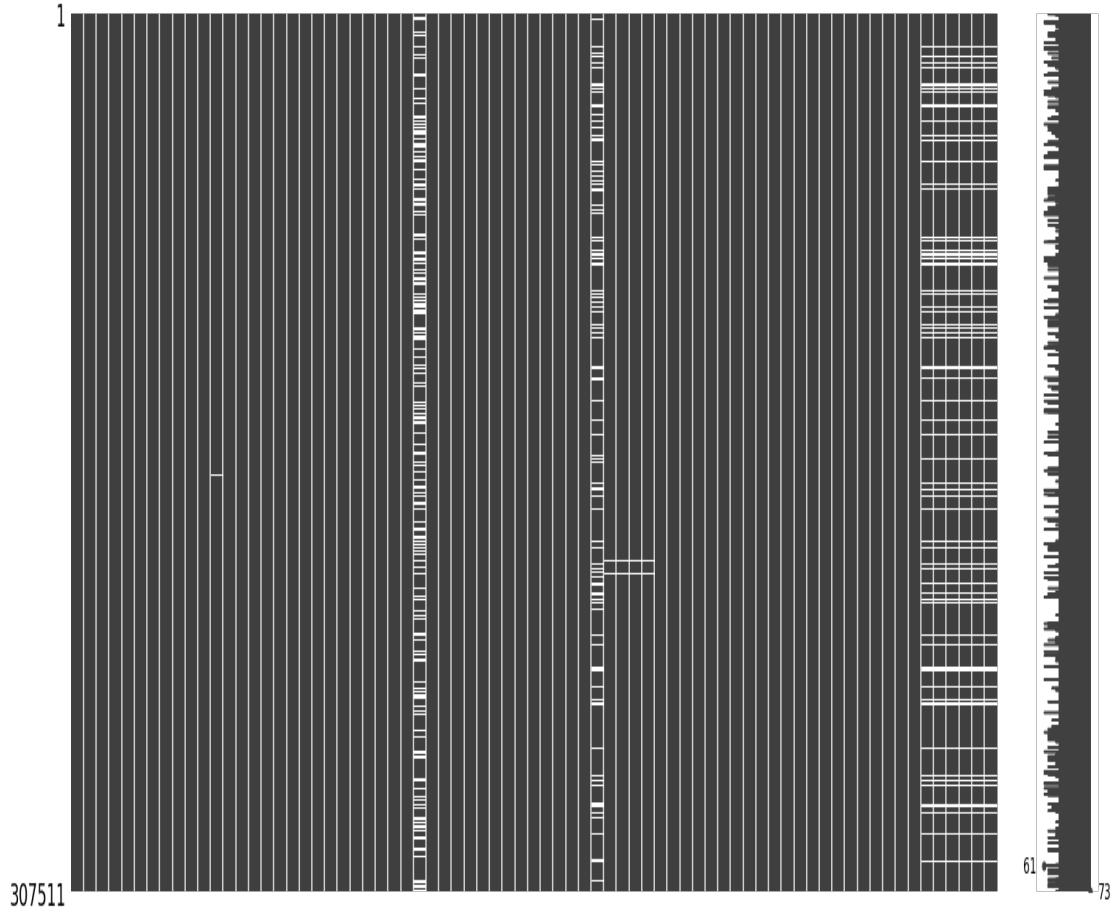


Let's see each block of pipeline individually below

DATA UNDERSTANDING - DISTRIBUTION



Before Column delete



After column delete

DATA UNDERSTANDING, CONTD

- The shape is pretty high having 3 Lakh+ rows and 122 columns
- The data consumes 286+ in memory resource.
- From above data distribution Visual, we can see that we have around 40+ columns are having more than 35% null values in overall data set.
- We can observe the TARGET columns which is a critical attribute with which we can corelate other attributes w.r.t this Attribute

DATA CLEANSING

we see columns with > 35% missing values

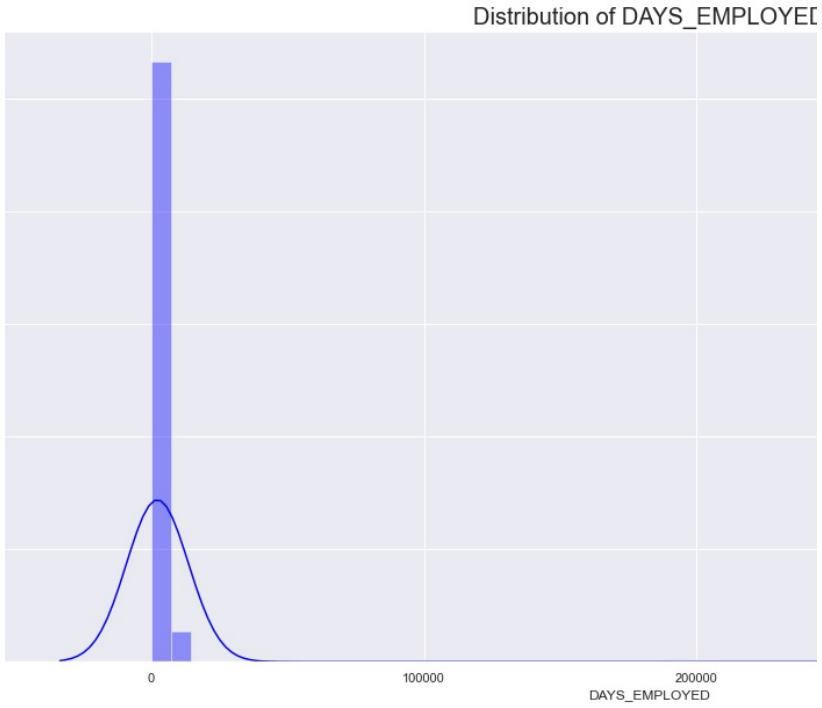
Hence, delete those column straightaway So the number of columns are reduced to 73

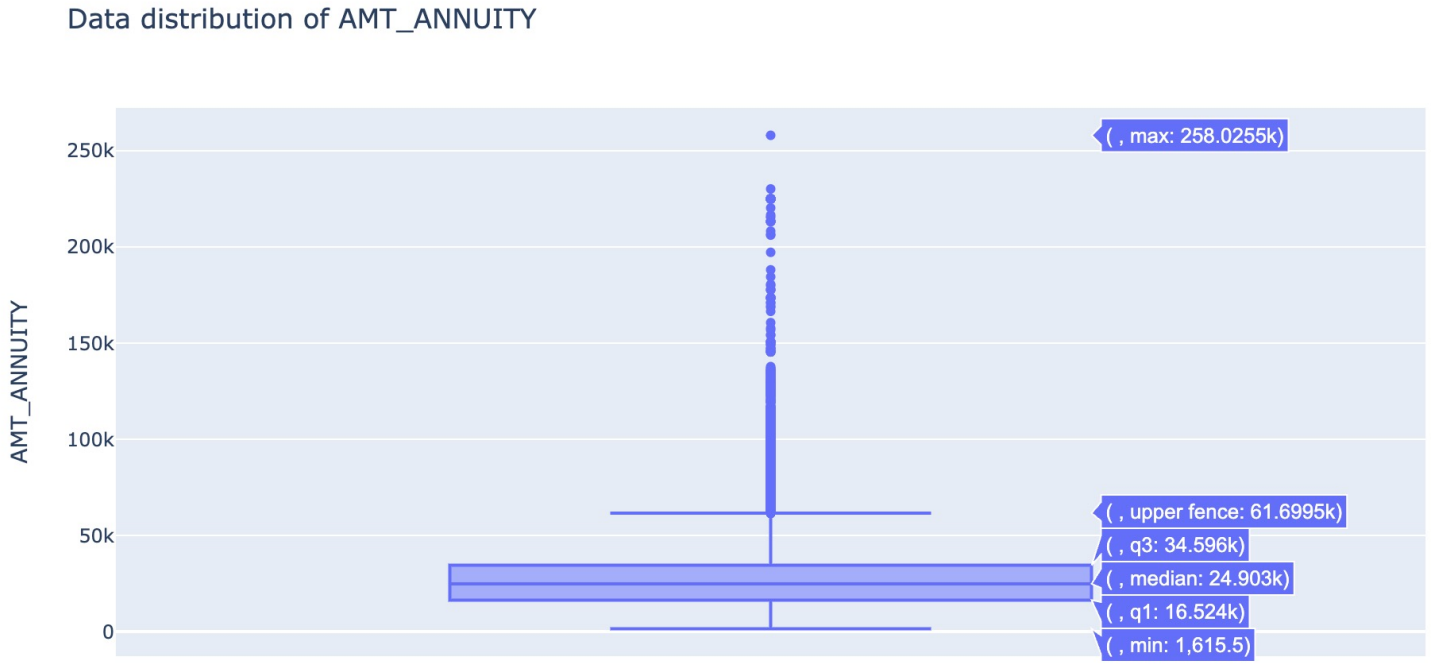
After deleting columns, we need to impute the rest of missing values with Mode/Median/Mean as per use case

Identify and correct/remove wrong values as necessary

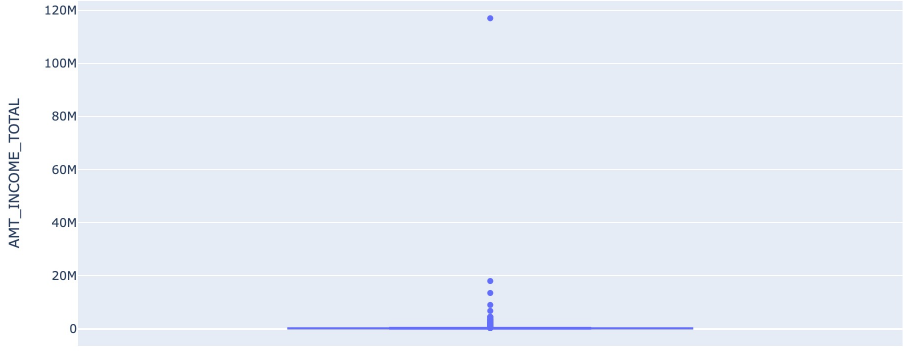
Enrich he data by adding additional column for continuous variable to make it as categorical range

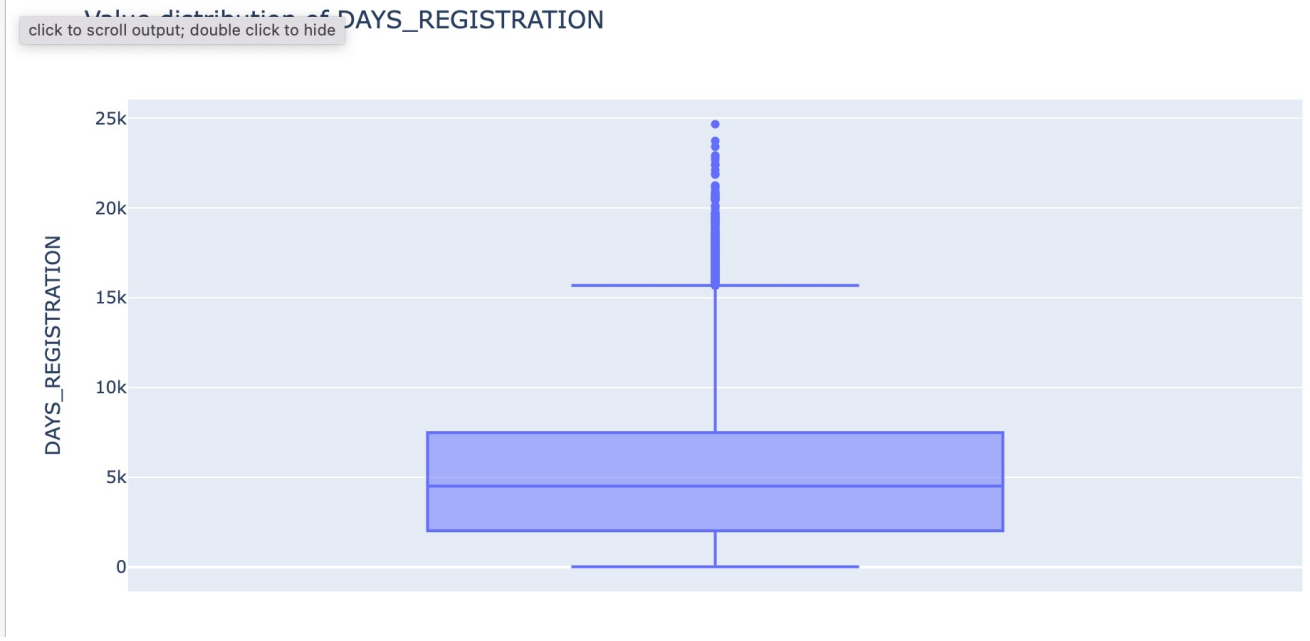
DATA ANALYSIS - OUTLIERS



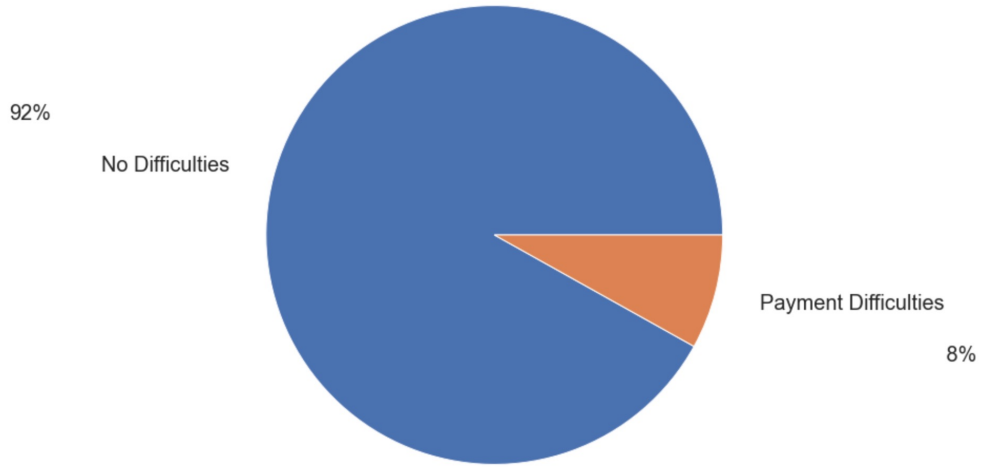


Boxplot visual of AMT_INCOME_TOTAL



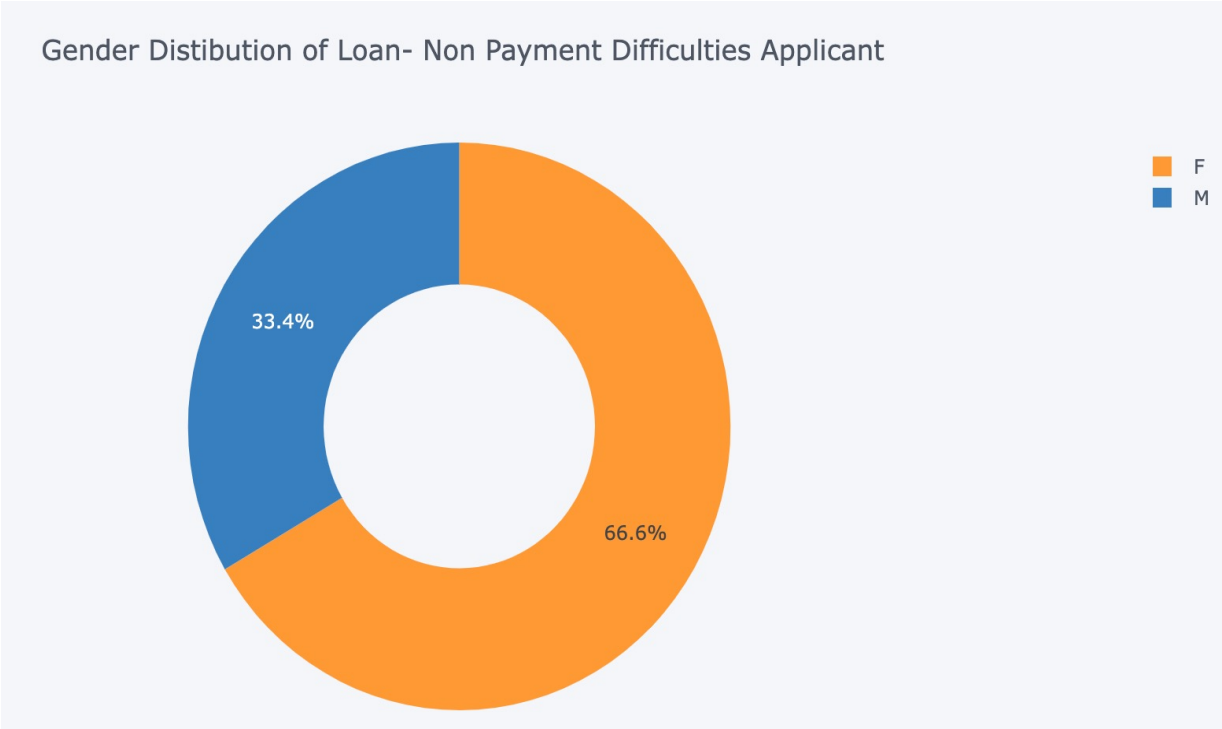


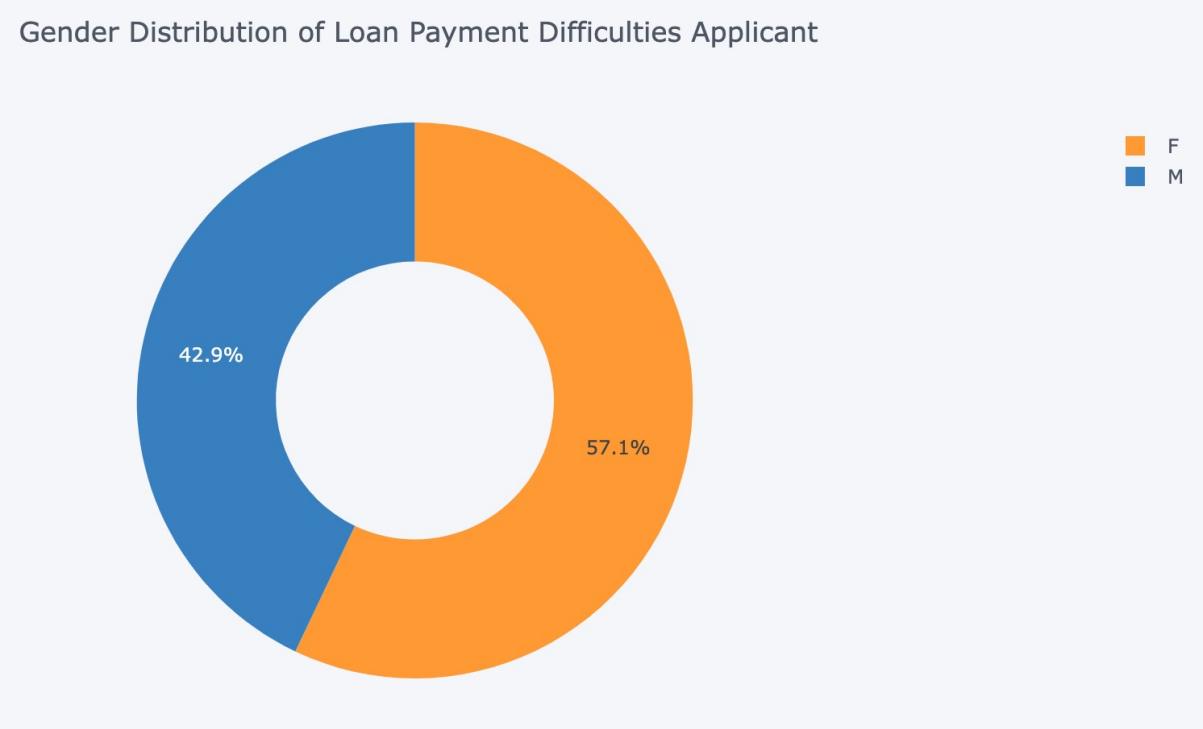
DATA ANALYSIS - IMBALANCE ANALYSIS



- Observed data imbalance in TARGET attribute
- We have 92% of applicant without payment difficulties and rest 8% with difficulties.
- We will divide the data into two subset based on Target attribute for further analysis

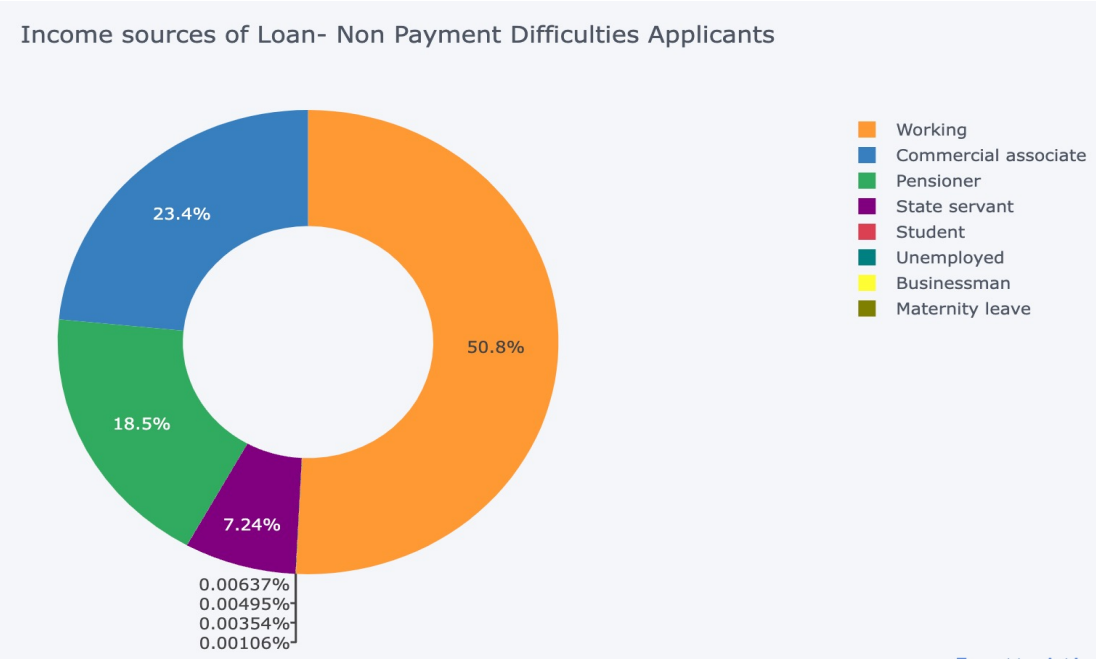
UNIVARIATE ANALYSIS

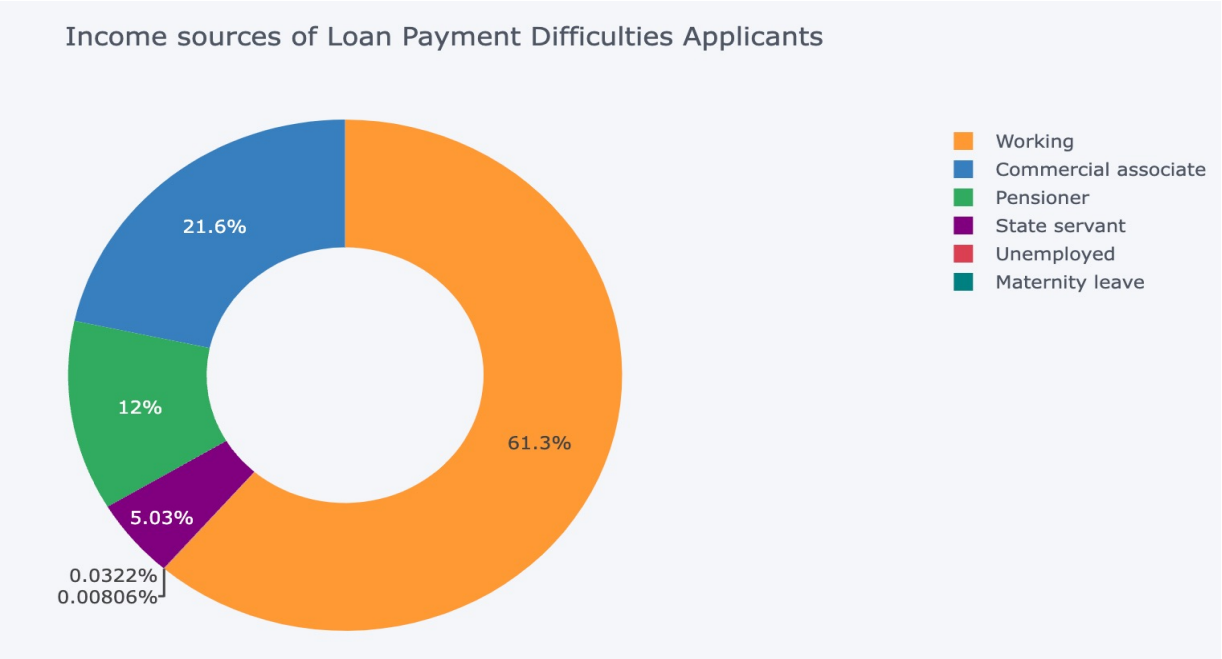




- Females are major applicant overall
- Increment in Female percentage seen from nonpayment difficulties to payment difficulties
- Essentially, female applicants are at higher rick

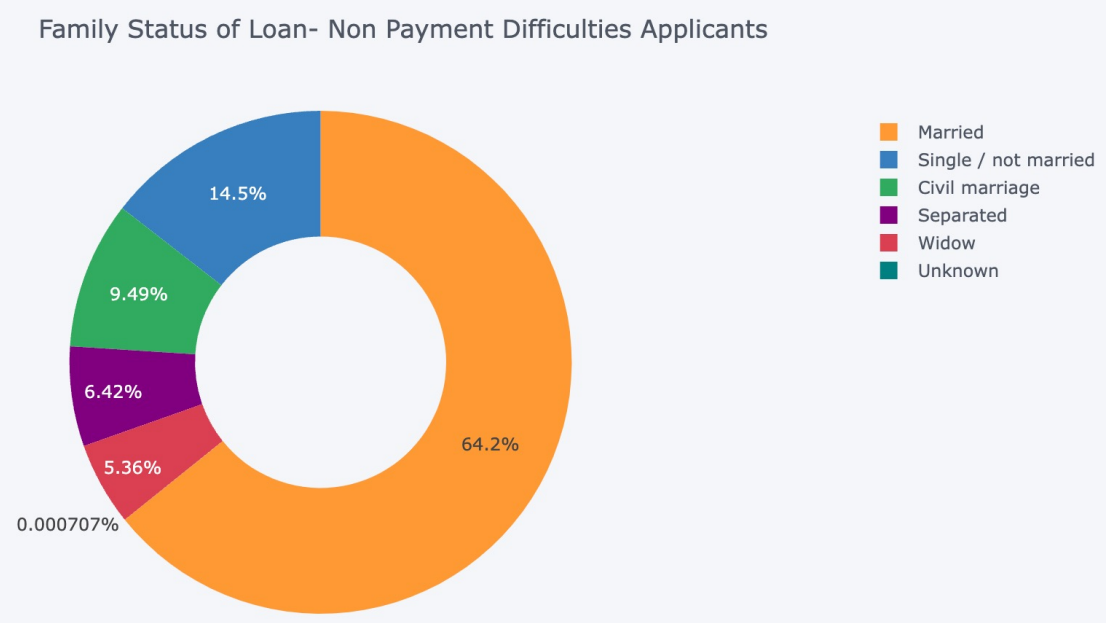
UNIVARIATE ANALYSIS CONTINUED

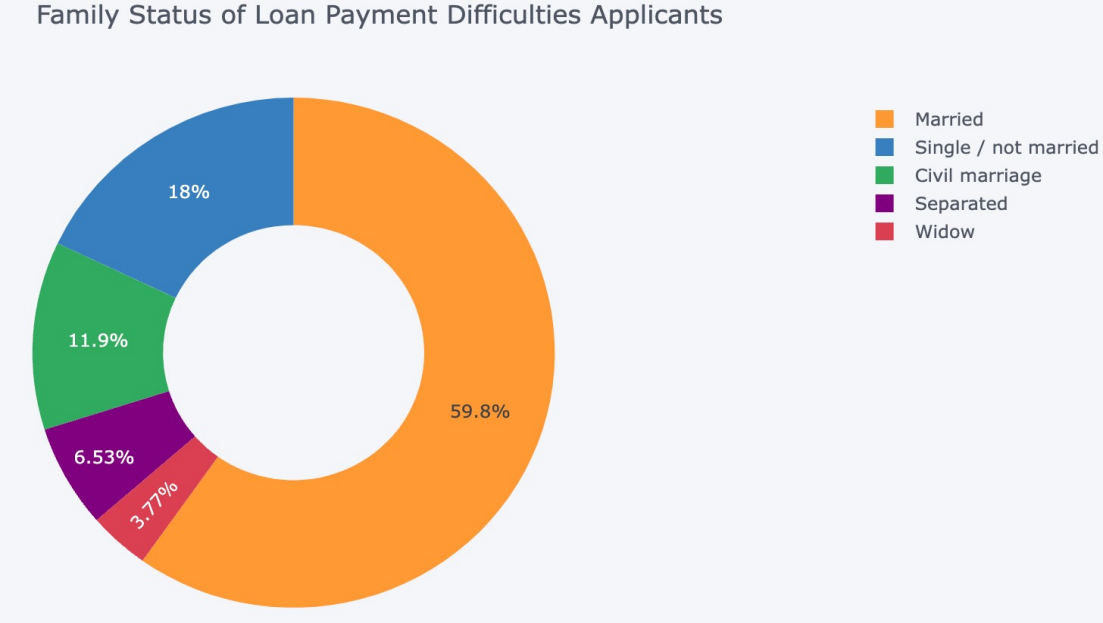




- Overall, working category are major applicants
- Increase in percentage in Difficulties in payment category w.r.t Non difficulty category for Working group
- Decrease in percentage in Commercial associate, Pensioner and State Servant from non-payment difficulties to Payment difficulties group
- Essentially, 'Working' professionals are at border of risk. Hence, we need to consider more factors for this category of applicants

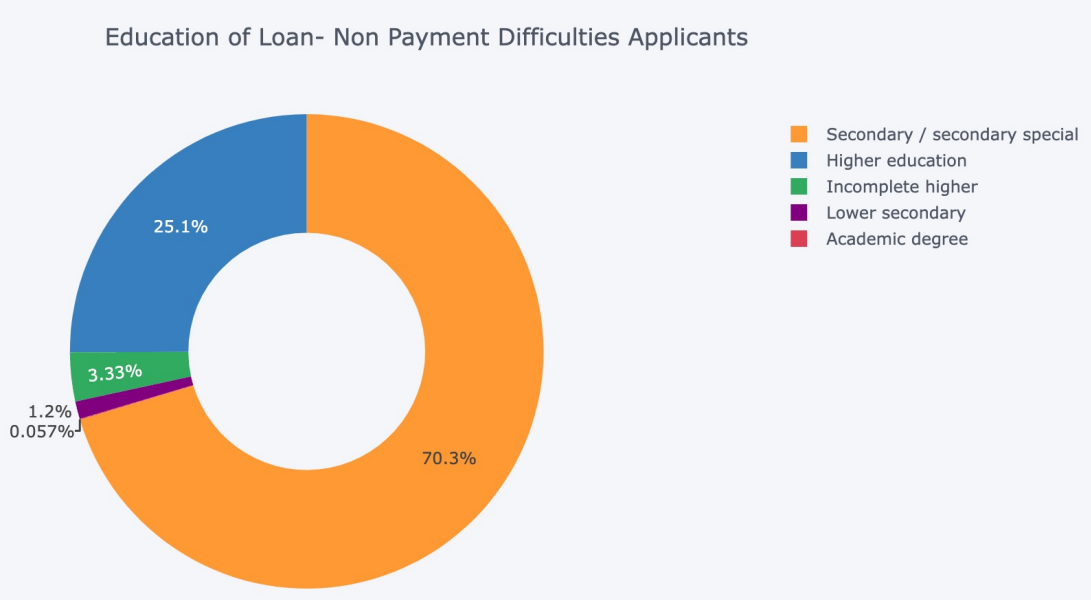
UNIVARIATE ANALYSIS CONTD.

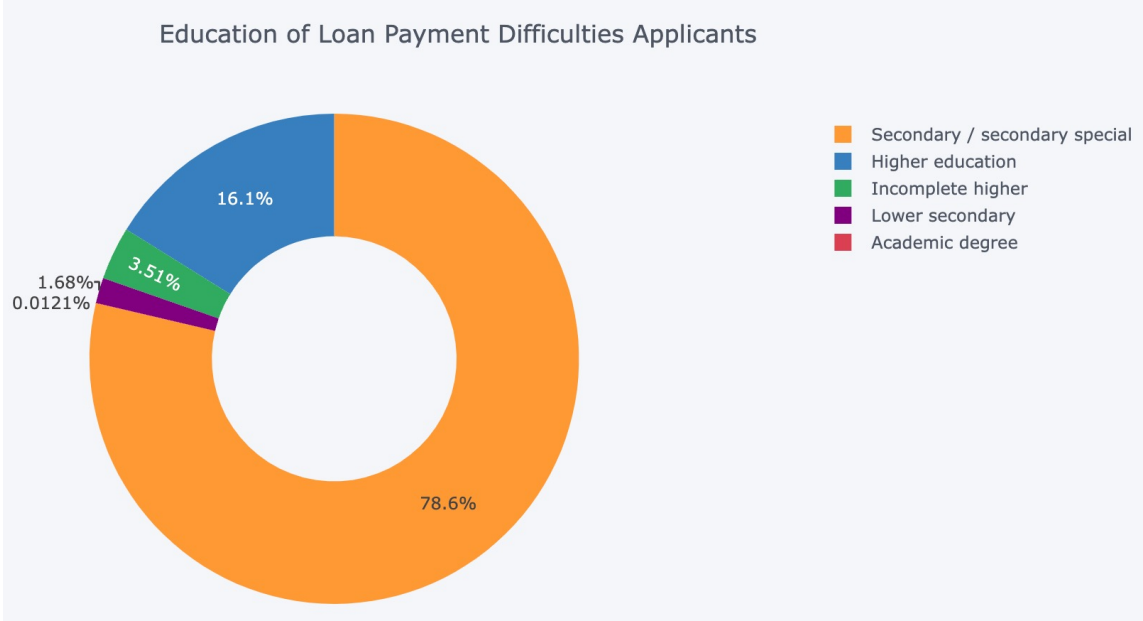




- Overall, Married applicants are majority in number
- Observed a decrease in the percentage of Married and Widowed with Loan Payment Difficulties and an increase in the the percentage of single and civil married with Loan Payment Difficulties when compared with the percentages of both Loan Payment Difficulties and Loan Non-Payment Difficulties
- Essentially, single and civil married are at high risk

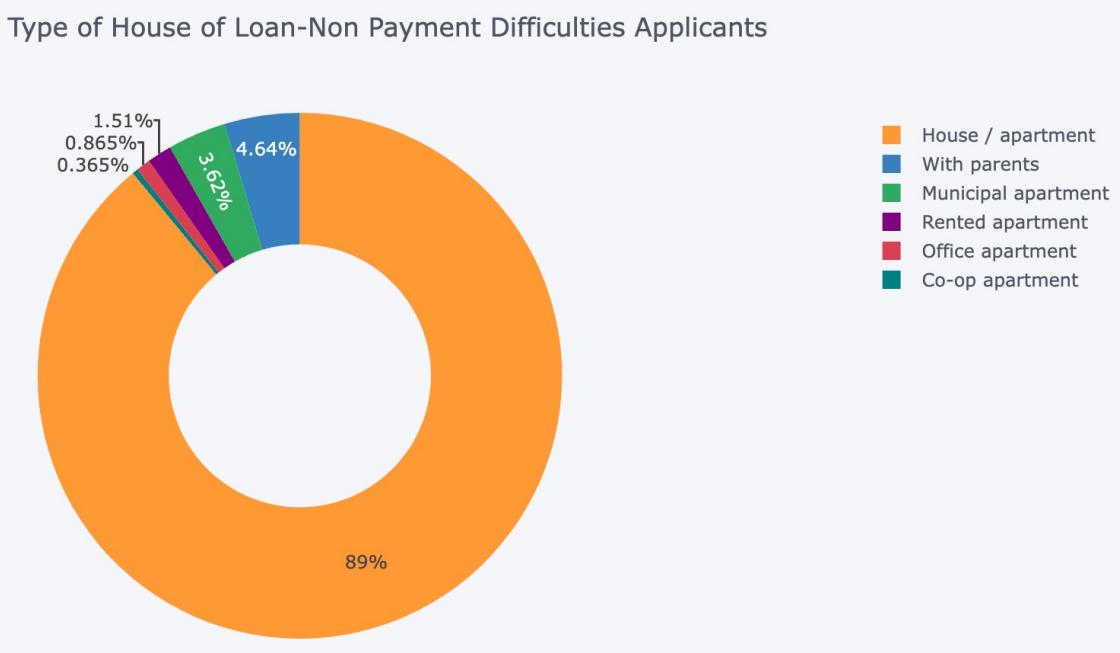
UNIVARIATE ANALYSIS CONTD.

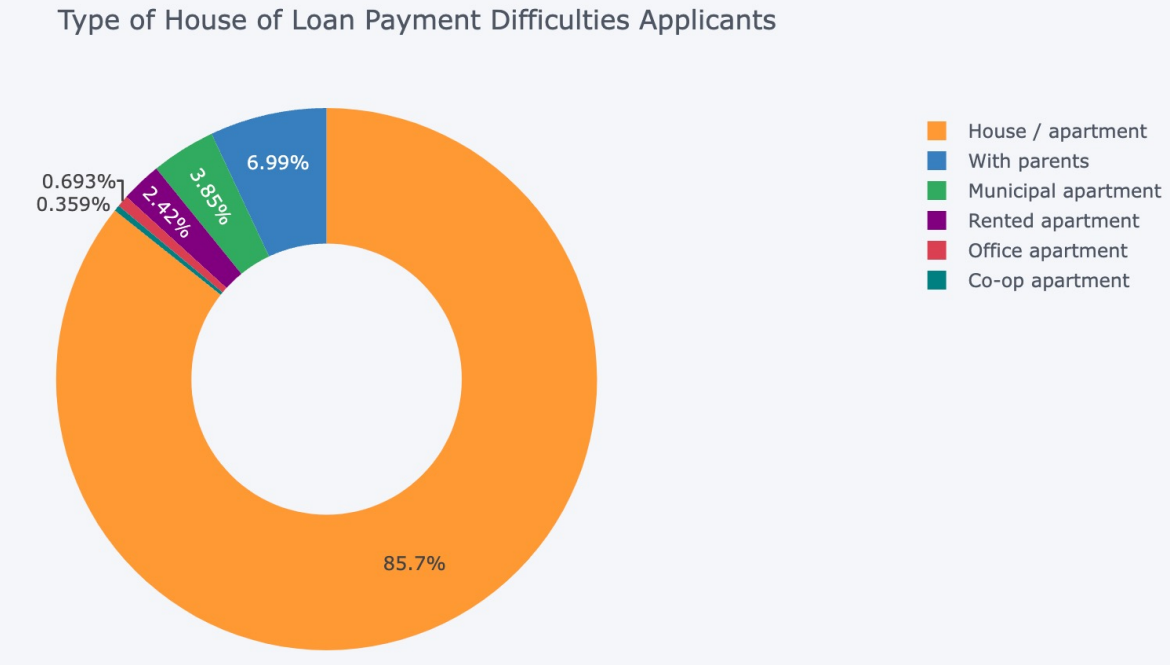




- Overall, secondary/secondary special category are major applicants
- observe an increase in percentage of Loan Payment Difficulties whose educational qualifications are secondary/secondary special and a decrease in the percentage of Loan Payment Difficulties who have completed higher education when compared with the percentages of Loan Payment Difficulties and Loan Non-Payment Difficulties
- Essentially, secondary/secondary special category is at high risk

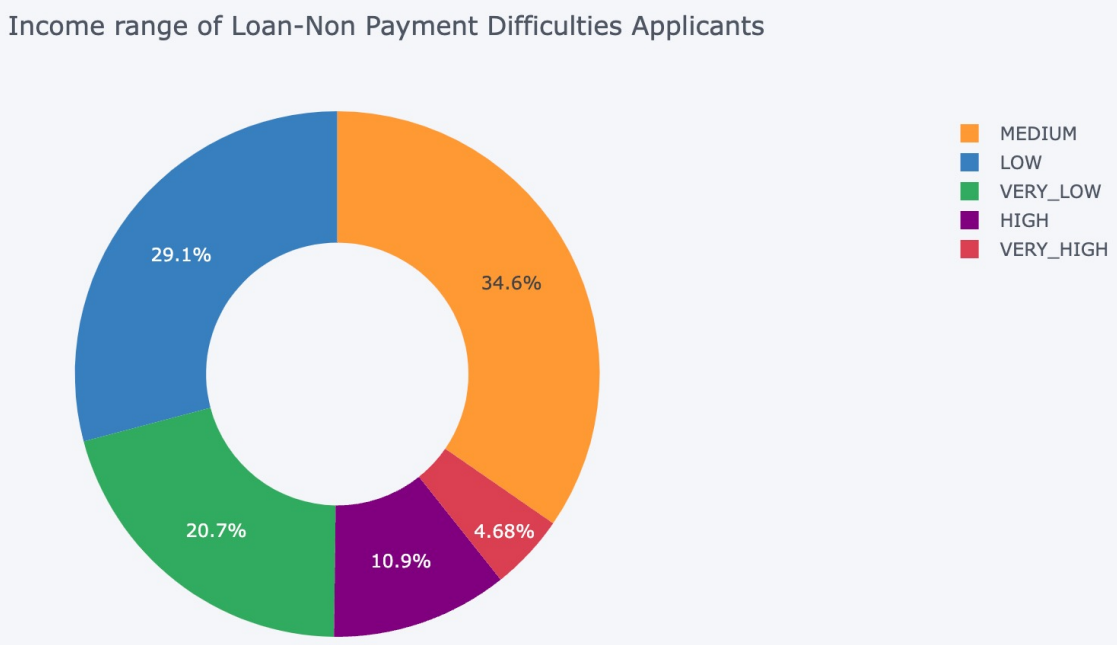
UNIVARIATE ANALYSIS CONTD.

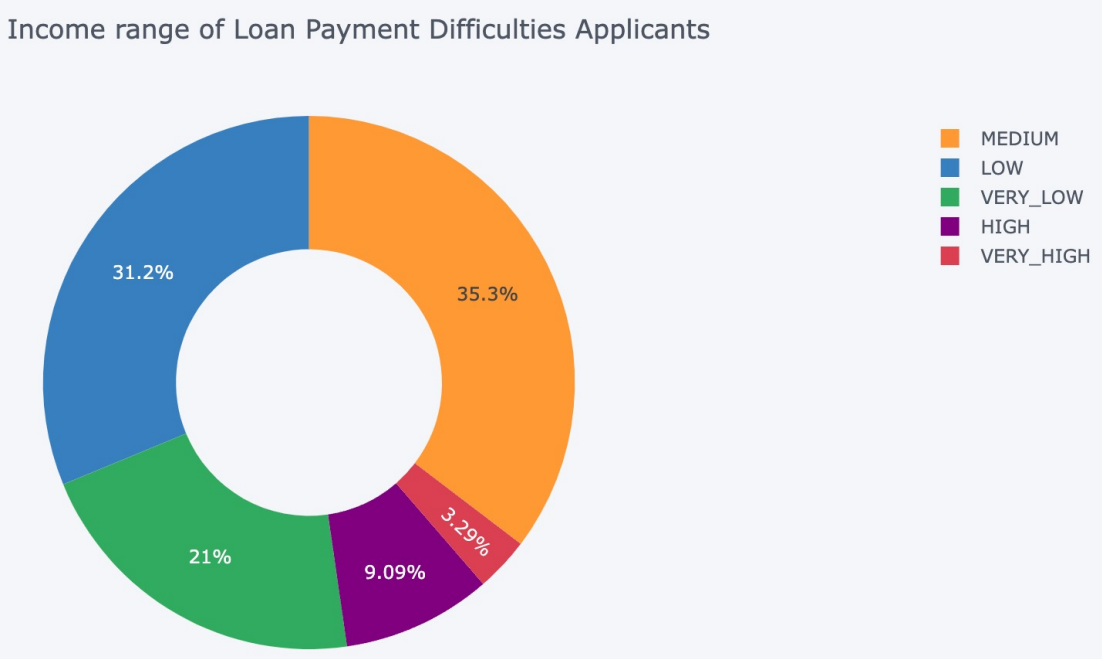




- Overall, applicants live in House/apartment are major in number
- Observed an increase in the percentage of Payment Difficulties who live with their parents when compared to the percentages of Payment Difficulties and non-Payment Difficulties
- Decrease in percentage for Payment Difficulties who live in House / Apartments when compared to the percentages of Payment Difficulties and non-Payment Difficulties.
- Essentially, applicants living with parents are at high risk

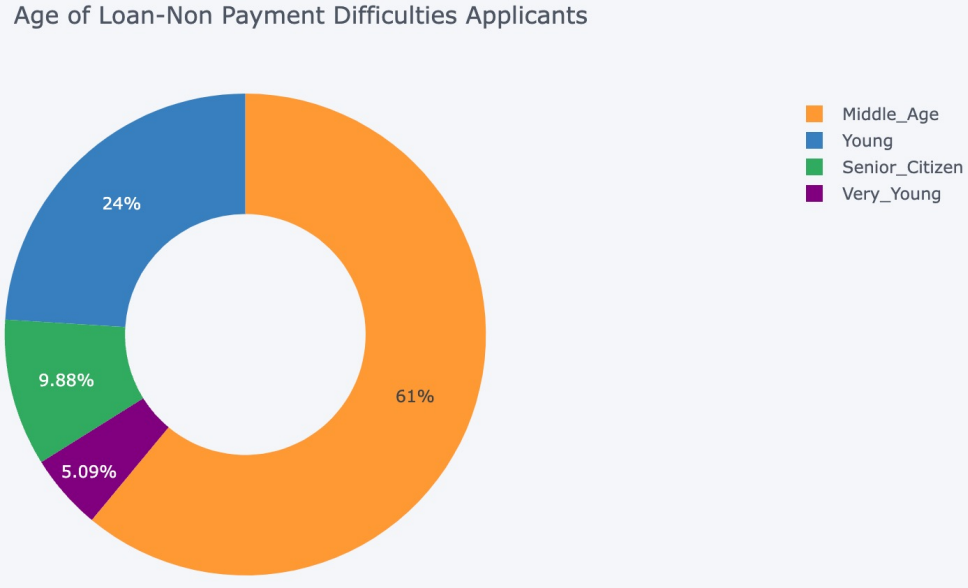
UNIVARIATE ANALYSIS CONTD.

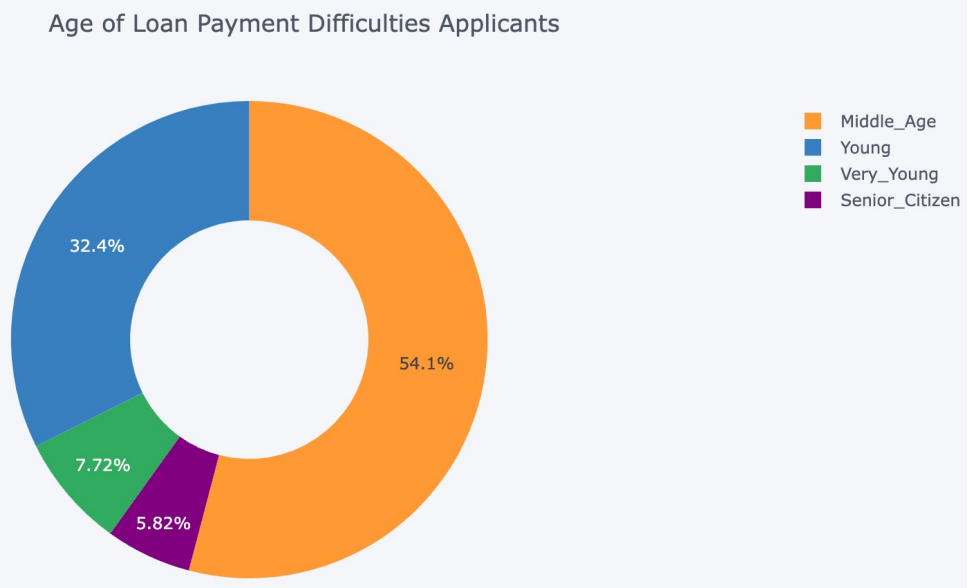




- Overall, Medium income range applicants are major in number
- Observed an increase in the percentage of Loan Payment Difficulties whose income is low when compared with the percentages of Payment Difficulties and Loan non-Payment Difficulties
- Low-income category are at higher risk

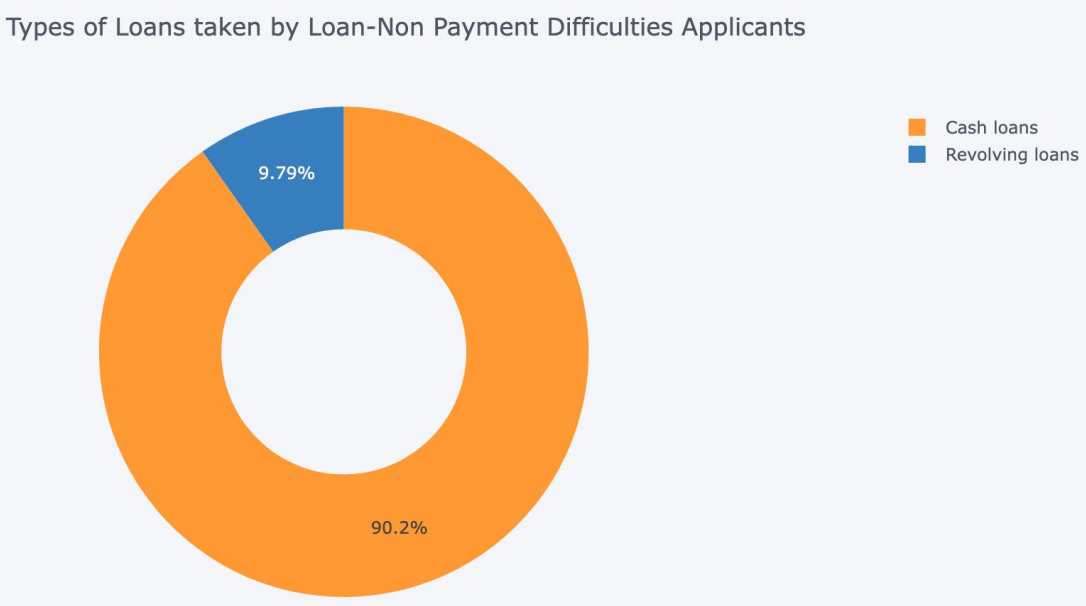
UNIVARIATE ANALYSIS CONTD.

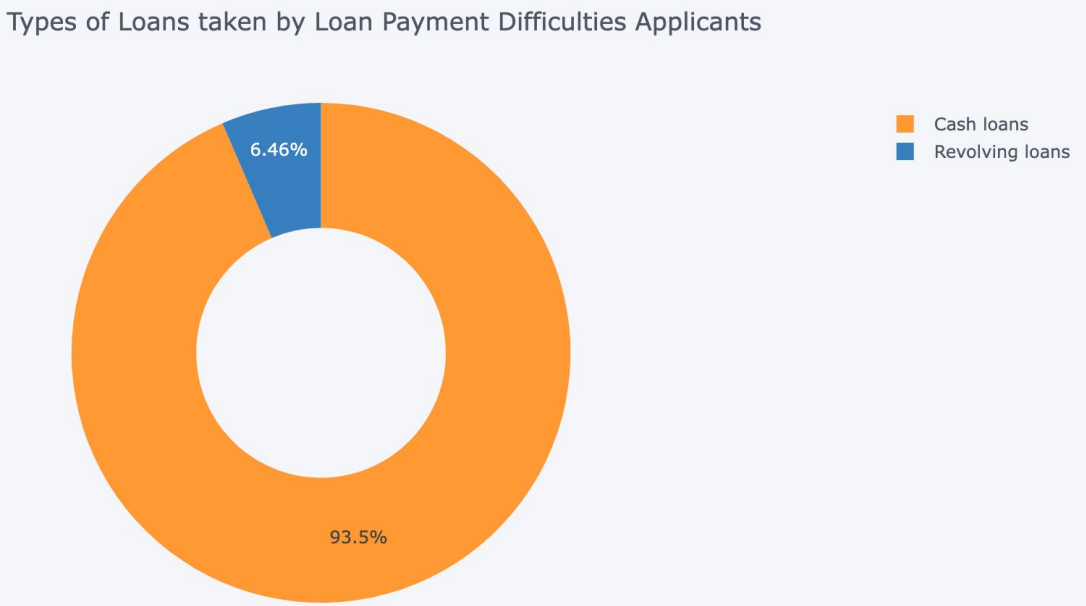




- Overall, Mid age applicants are major in number
- Observed that there is an increase in the percentage of Loan Payment Difficulties who are young in age when compared to the percentages of Payment Difficulties and Loan-Non Payment Difficulties
- Young age applicants are at high risk

UNIVARIATE ANALYSIS CONTD.

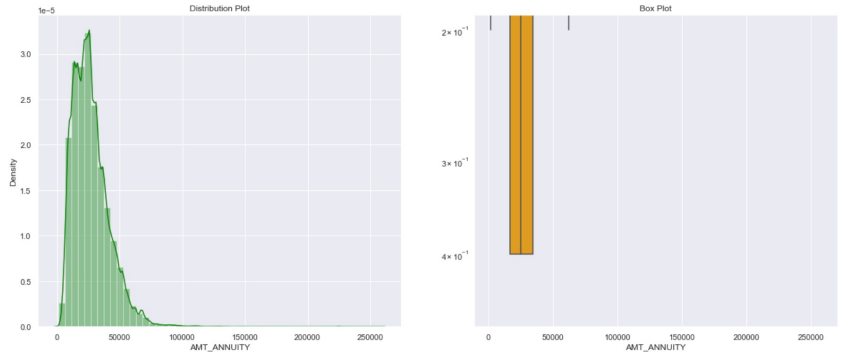




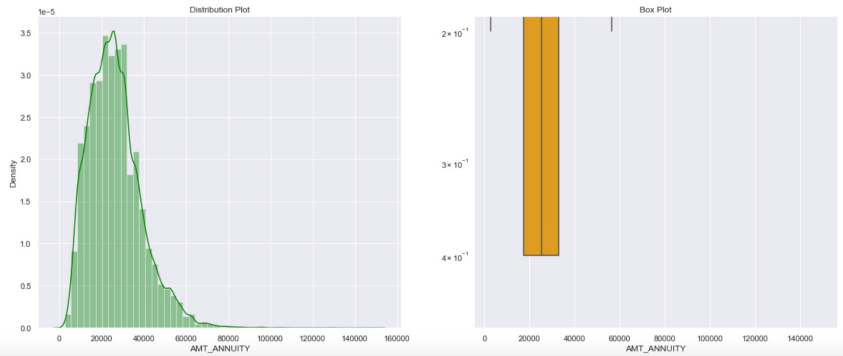
- Overall, Cash loan applicants are higher in number
- Observed increment in percentage of cash loan from applicant not having payment difficulties to ones having payment difficulties
- Cash loans are at high risk

UNIVARIATE ANALYSIS CONTD.

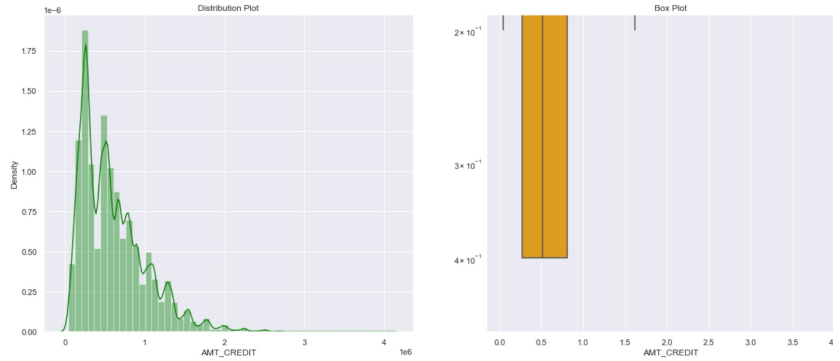
Non-Payment Difficulties Applicant



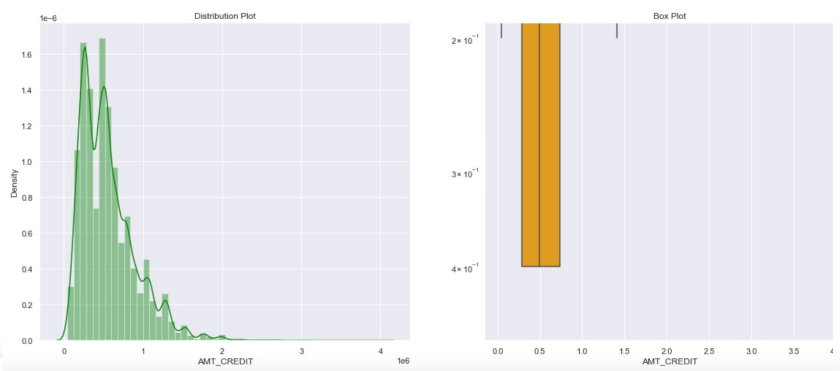
Payment Difficulties Applicant



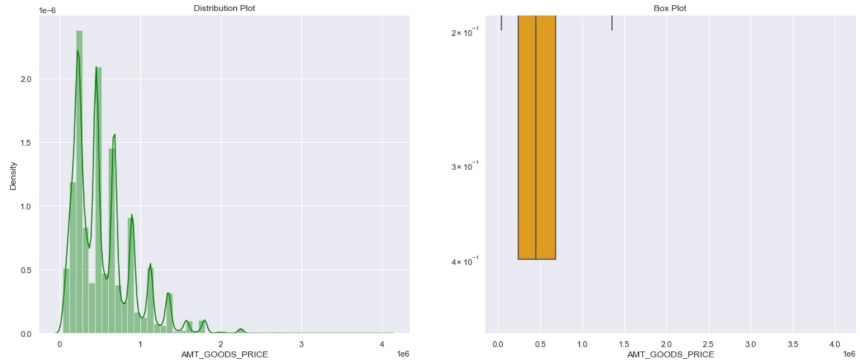
Distribution and Box plot for 'AMT_CREDIT' for Loan Non-Payment Difficulties



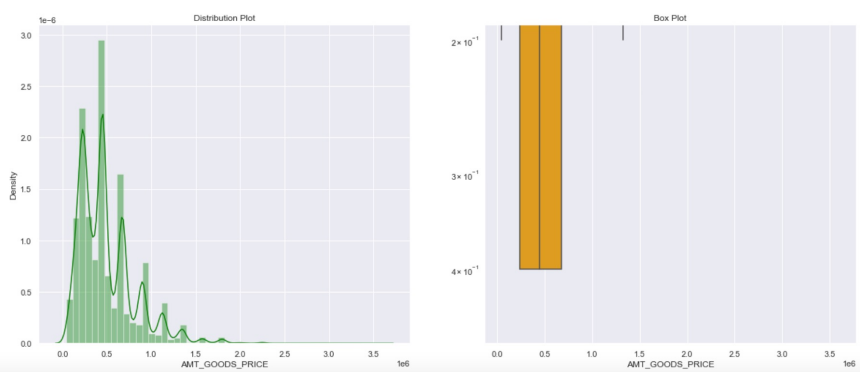
Distribution and Box plot for 'AMT_CREDIT' for Loan Payment Difficulties



Distribution and Box plot for 'AMT_GOODS_PRICE' for Loan Non-Payment Difficulties



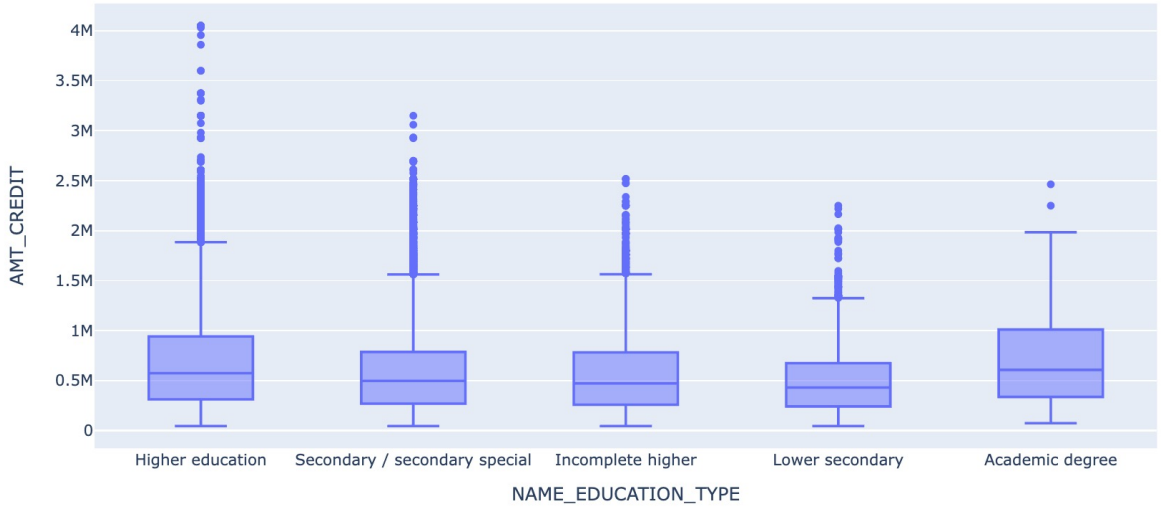
Distribution and Box plot for 'AMT_GOODS_PRICE' for Loan Payment Difficulties



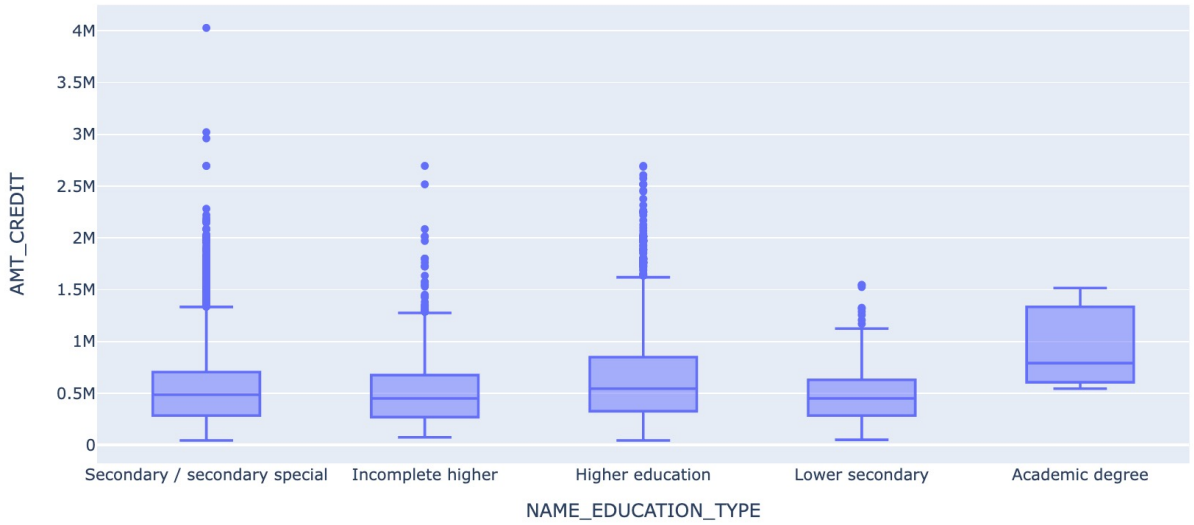
- Overall, Loan Annuity, Credit Amount¶, Goods Price falls in first quartile and is been aligned towards left in histogram and hence more applicants are having low Loan Annuity, Credit Amount¶, Goods Price

BIVARIATE ANALYSIS

Credit amount vs Education of Loan- Non Payment Difficulties Applicants



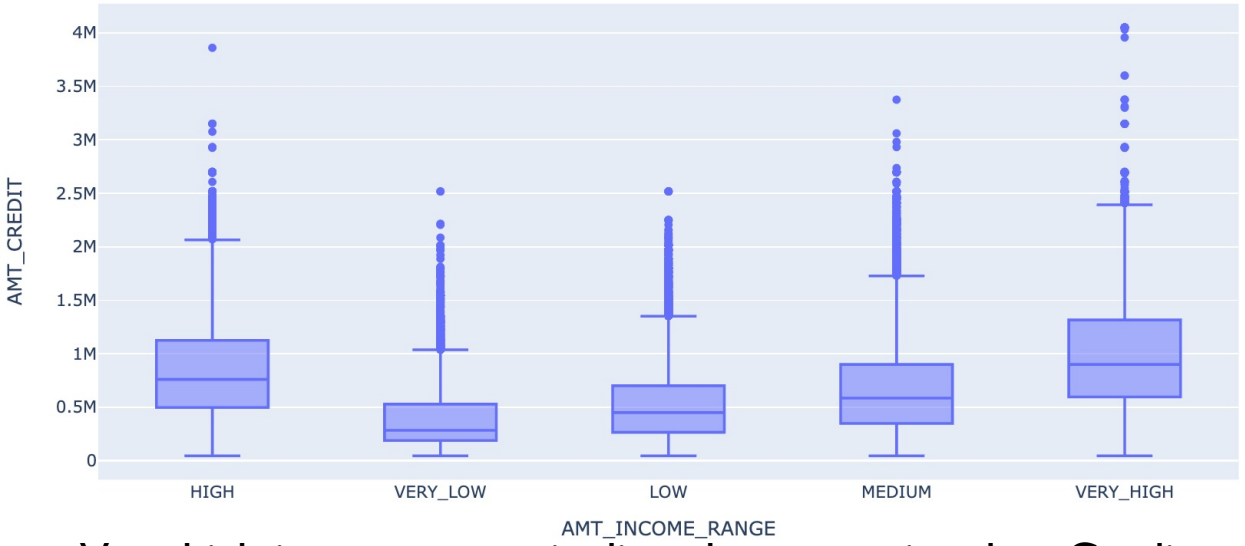
Credit amount vs Education of Loan Payment Difficulties Applicants



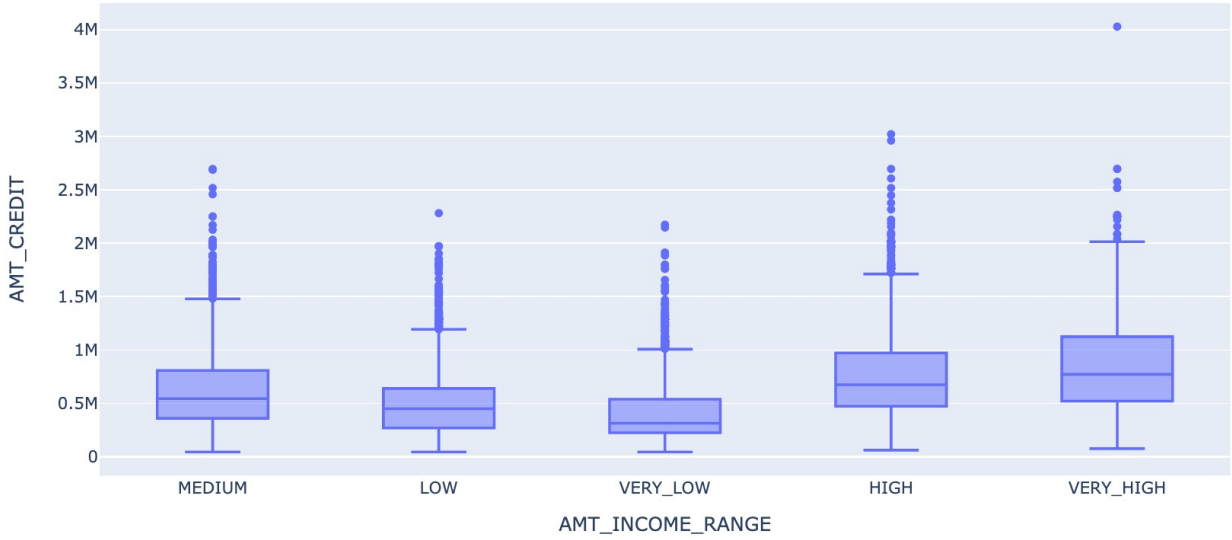
- Higher education applicant having highest values of Credit amount
- decent ranges are seen from Academic degree category
- Payment difficulties candidates are mostly in Secondary/ Secondary special
- More Outliers are seen both in Secondary as well as in Higher Education

BIVARIATE ANALYSIS

Income range vs Credit amount of Loan Non- Payment Difficulties



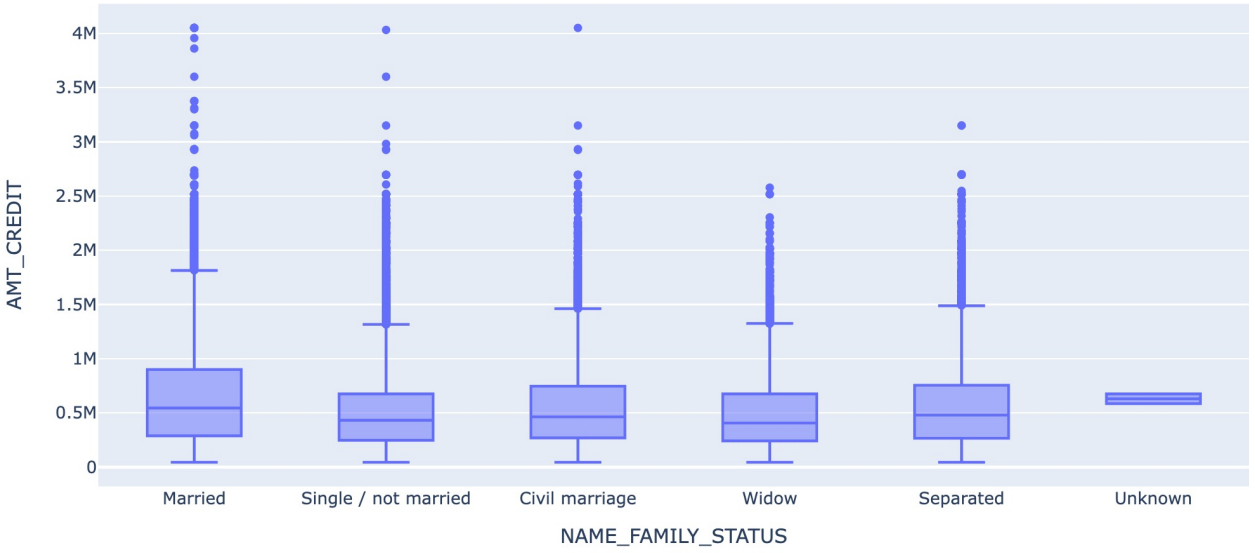
Income range vs Credit amount of Loan Payment Difficulties



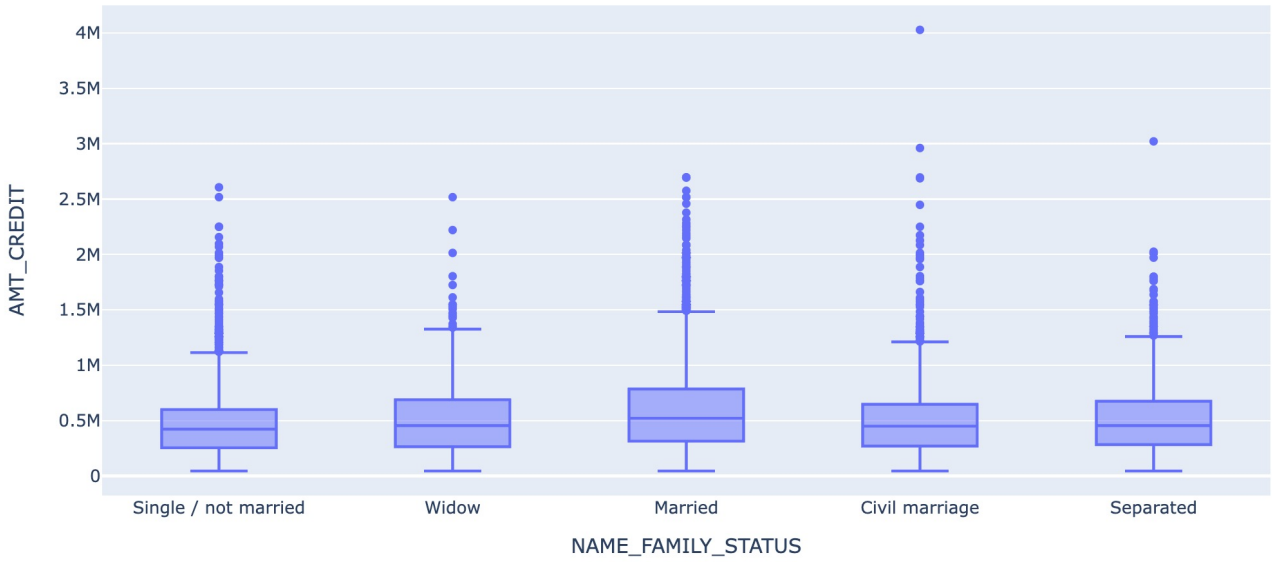
- Very high income range is directly proportional to Credit amount
- Very low income applicant max credit amount is less than all other category
- Very low income category has more number of applicant in high credit amount comparatively in Loan payment difficulties applicant
- Very high income category has one outlier near 4M in payment difficult categories with which it follows the below equation for rick
- RISK ~ Very LOW > Low, Medium, HIGH, VERY_HIGH

BIVARIATE ANALYSIS

Family Status vs Credit amount of Loan Non- Payment Difficulties



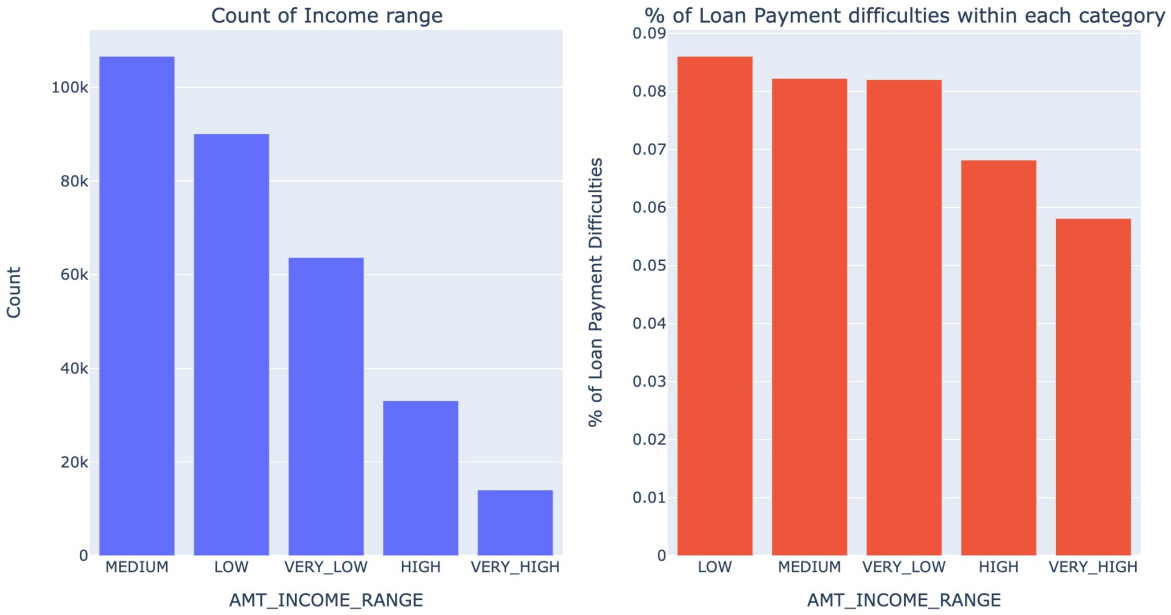
Income range vs Credit amount of Loan Payment Difficulties



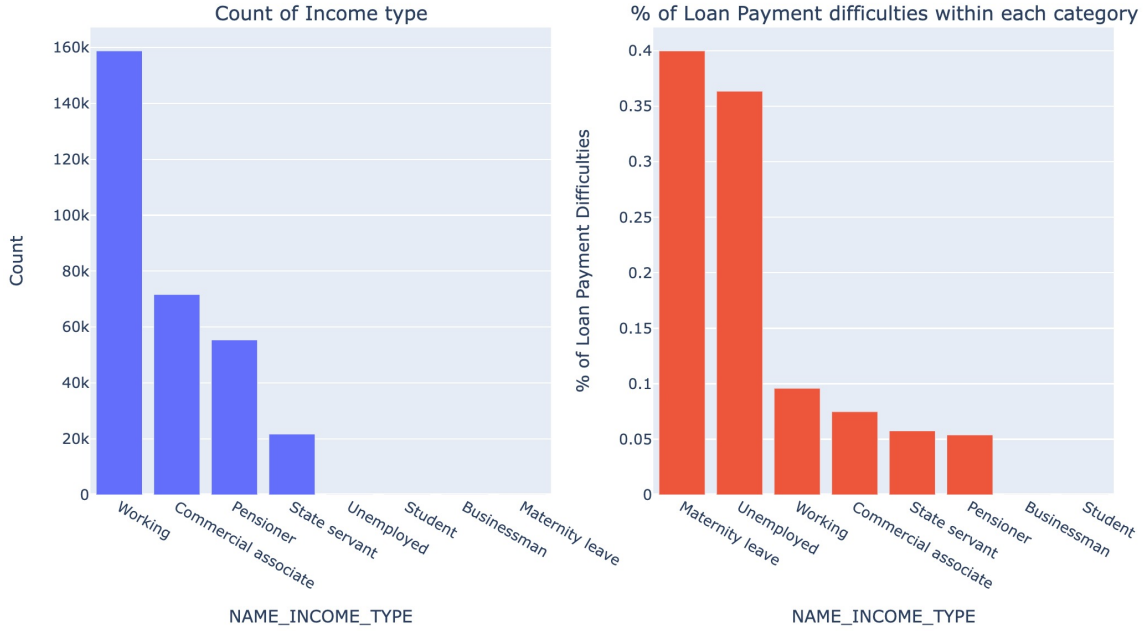
- Among all, Married cateegory occupies major number of applicants with higher credit amount
- The unknown category has very low range of credit amount
- We can observe Civil married and Single applicants are most in Loan payment difficulties than any other category

BIVARIATE ANALYSIS

Income range

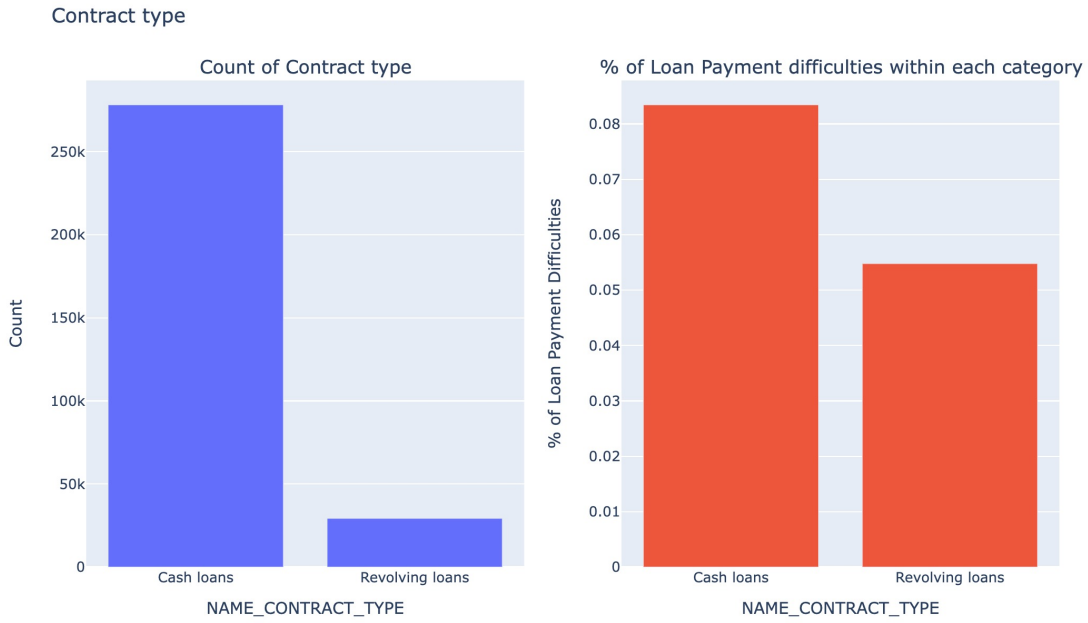


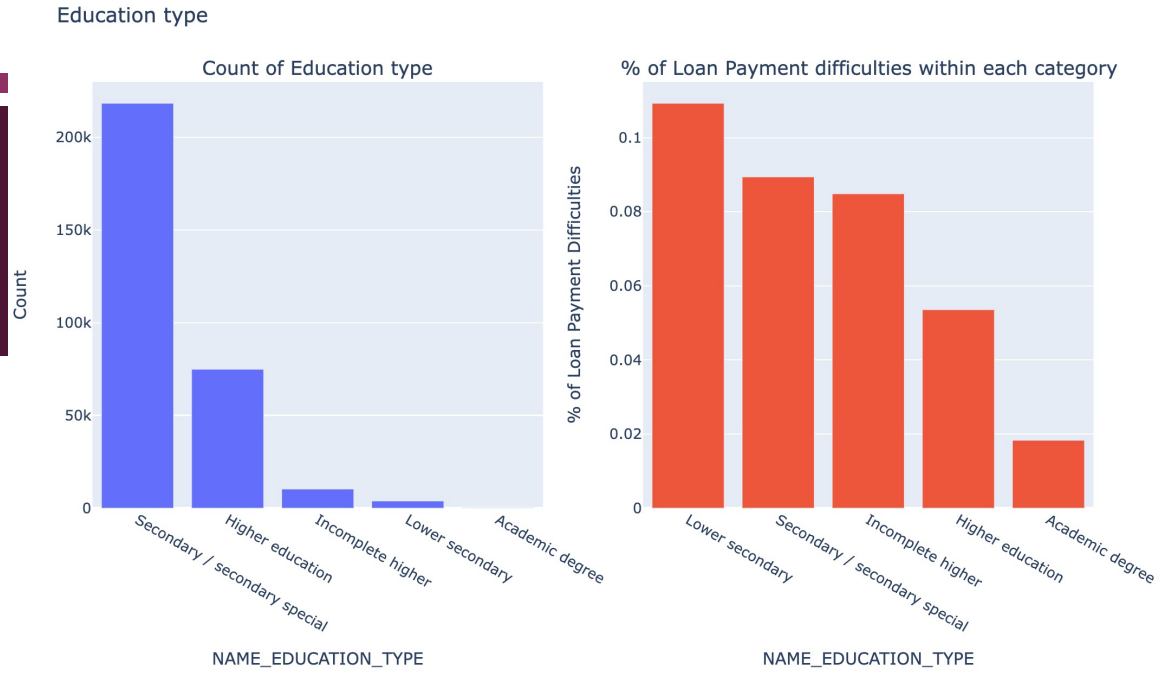
Income type

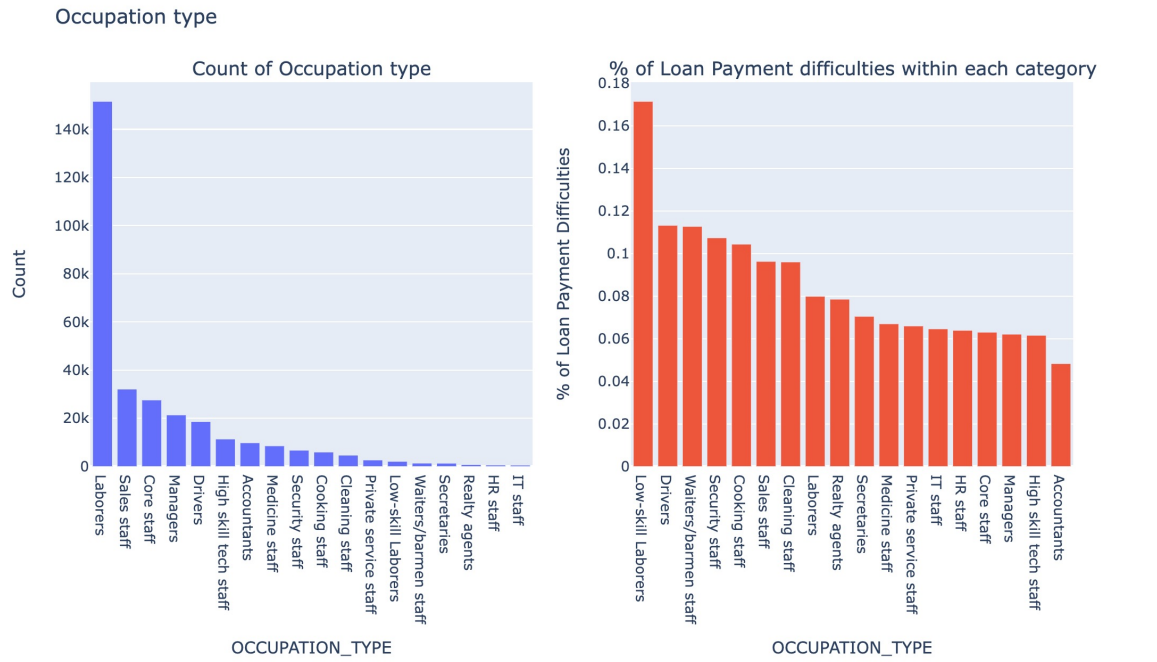


- From the plot above we can say that clients with 'LOW' Income range have maximum percent of Loan-Payment Difficulties
- From the plot above we can say that clients with 'Maternity leave' Income type have maximum percent of Loan-Payment Difficulties

BIVARIATE ANALYSIS



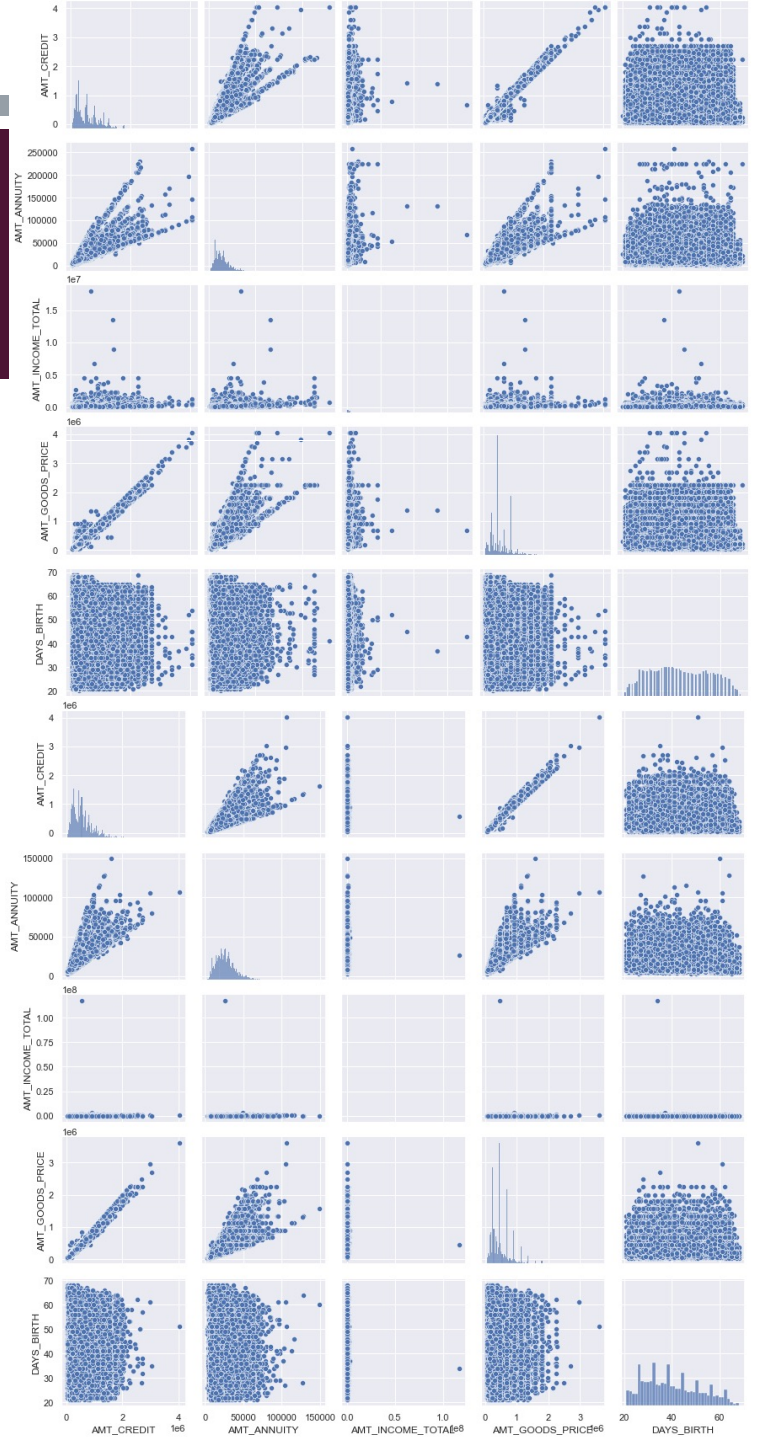




- From the plot above we can say that clients with 'Cash loans' contract type have maximum % of Loan-Payemnt Difficulties
- Clients with 'Lower secondary' education type have maximum % of Loan-Payment Difficulties.
- Applicants with 'Lower skill Laborers' occupation type have maximum percent of Loan-Payment Difficulties.

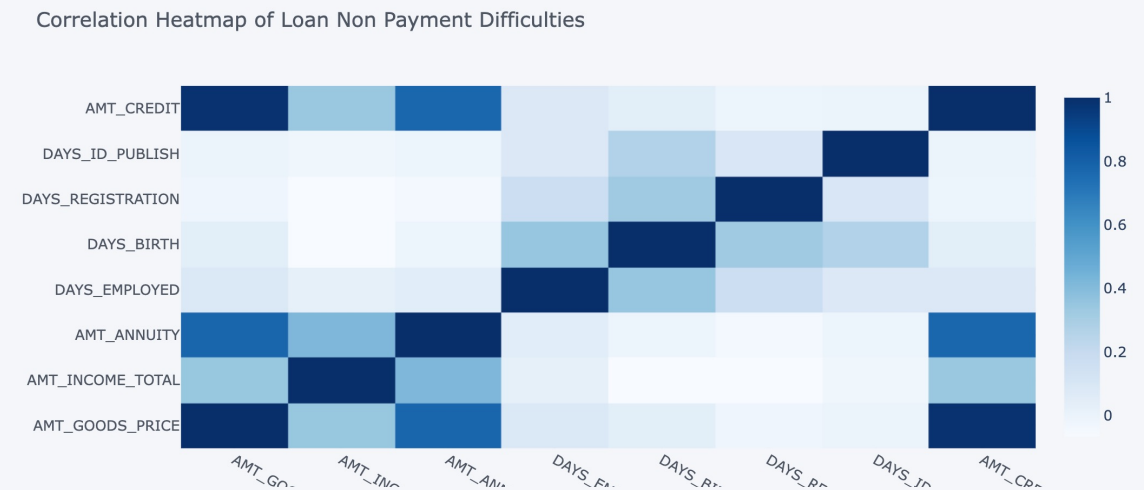
BIVARIATE ANALYSIS

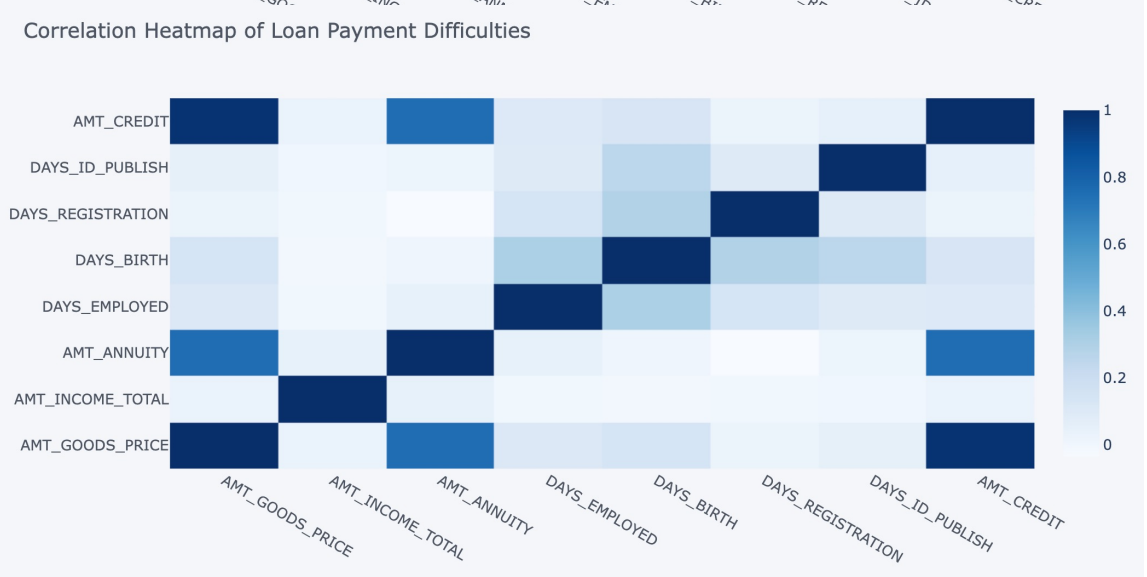
CODE_GENDER	NAME_EDUCATION_TYPE / AMT_INCOME_RANGE	Academic degree	Higher education	Incomplete higher	Lower secondary	Secondary / secondary special
F	VERY_LOW	0.000000	0.056068	0.086399	0.080193	0.076778
	LOW	0.000000	0.049022	0.080075	0.113889	0.079523
	MEDIUM	0.000000	0.050254	0.078431	0.096983	0.075692
	HIGH	0.105263	0.041516	0.074313	0.038961	0.070736
	VERY_HIGH	0.076923	0.037289	0.082251	0.066667	0.065930
M	VERY_LOW	0.000000	0.080411	0.123967	0.125000	0.118066
	LOW	0.000000	0.073305	0.097778	0.142857	0.123693
	MEDIUM	0.000000	0.070086	0.095130	0.150515	0.113466
	HIGH	0.000000	0.055911	0.074627	0.081633	0.093484
	VERY_HIGH	0.000000	0.044080	0.077586	0.064516	0.089939



- From Female category Clients who have LOW income and ACADEMIC DEGREE education have maximum percent of Loan-Payment Difficulties
- From Male category Clients wo have MEDIUM income and LOWER SECONDARY education have maximum percent of Loan-Payment Difficulties

MULTIVARIATE ANALYSIS - CORRELATION HEATMAP

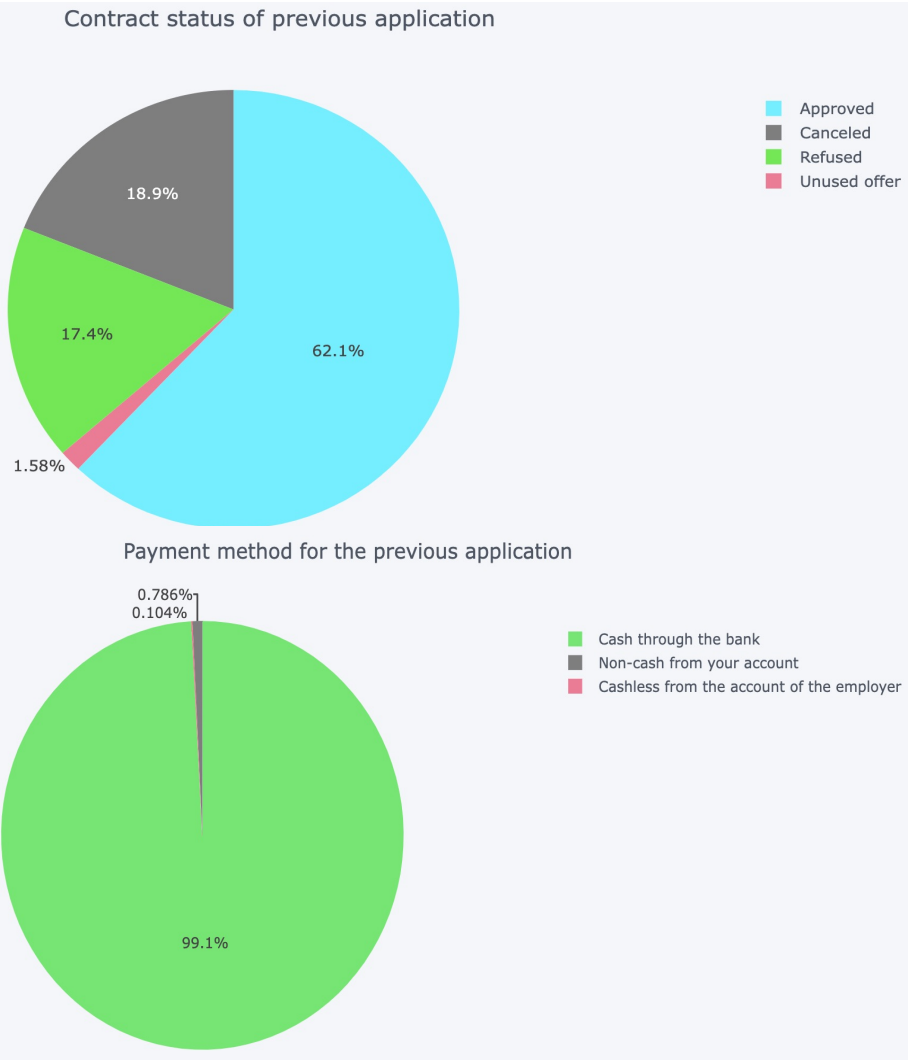




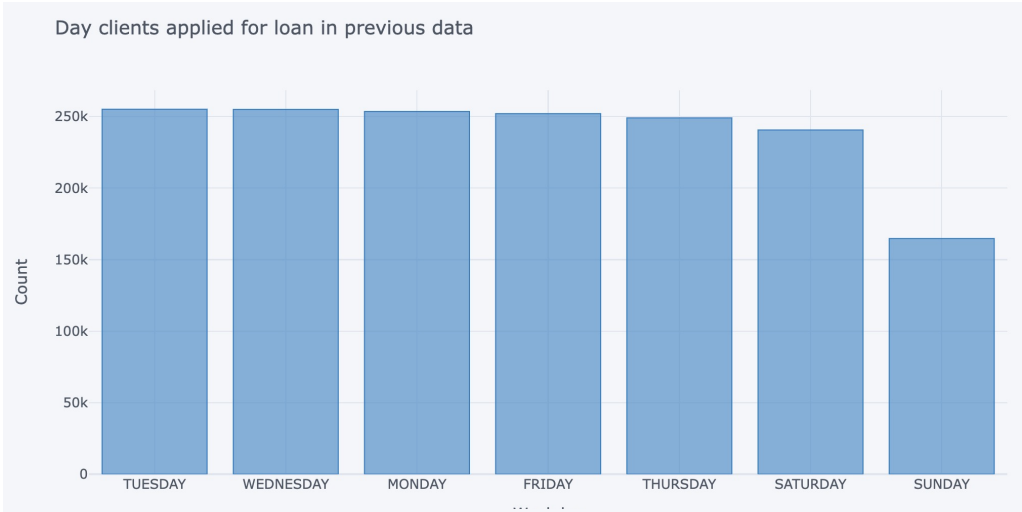
	VAR1	VAR2	CORRELATION	CORR_ABS
56	AMT_CREDIT	AMT_GOODS_PRICE	0.982783	0.982783
16	AMT_ANNUITY	AMT_GOODS_PRICE	0.752295	0.752295
58	AMT_CREDIT	AMT_ANNUITY	0.752195	0.752195
35	DAYS_BIRTH	DAYS_EMPLOYED	0.306726	0.306726
44	DAYS_REGISTRATION	DAYS_BIRTH	0.288872	0.288872
52	DAYS_ID_PUBLISH	DAYS_BIRTH	0.252256	0.252256
32	DAYS_BIRTH	AMT_GOODS_PRICE	0.135532	0.135532
43	DAYS_REGISTRATION	DAYS_EMPLOYED	0.135465	0.135465
60	AMT_CREDIT	DAYS_BIRTH	0.135070	0.135070
24	DAYS_EMPLOYED	AMT_GOODS_PRICE	0.111886	0.111886

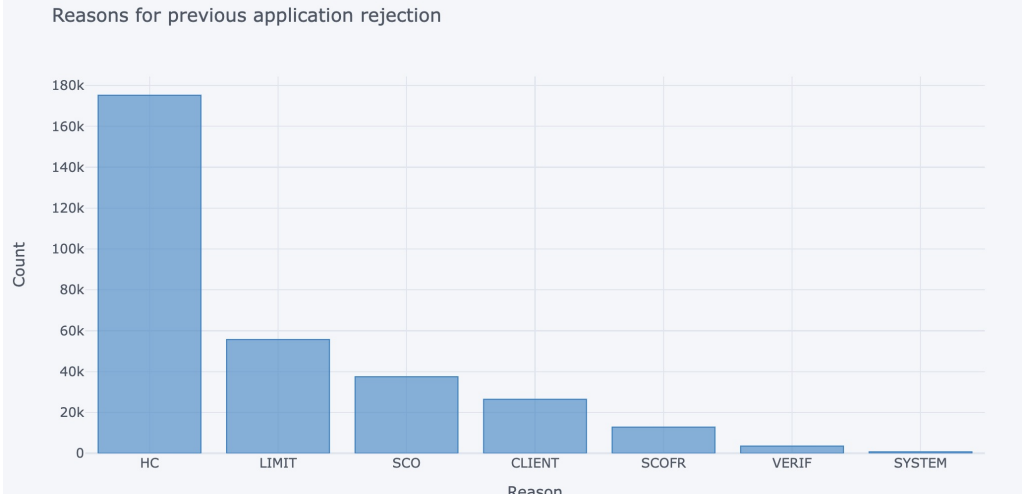
- We observe that there is a high correlation between credit amount and goods price. There appears to be some deviancies in the correlation of Loan-Payment Difficulties and Loan- non-payment Difficulties such as credit amount v/s income.
- Above is top 10 correlation for clients with payment difficulties

PREVIOUS APPLICATION DATA ANALYSIS

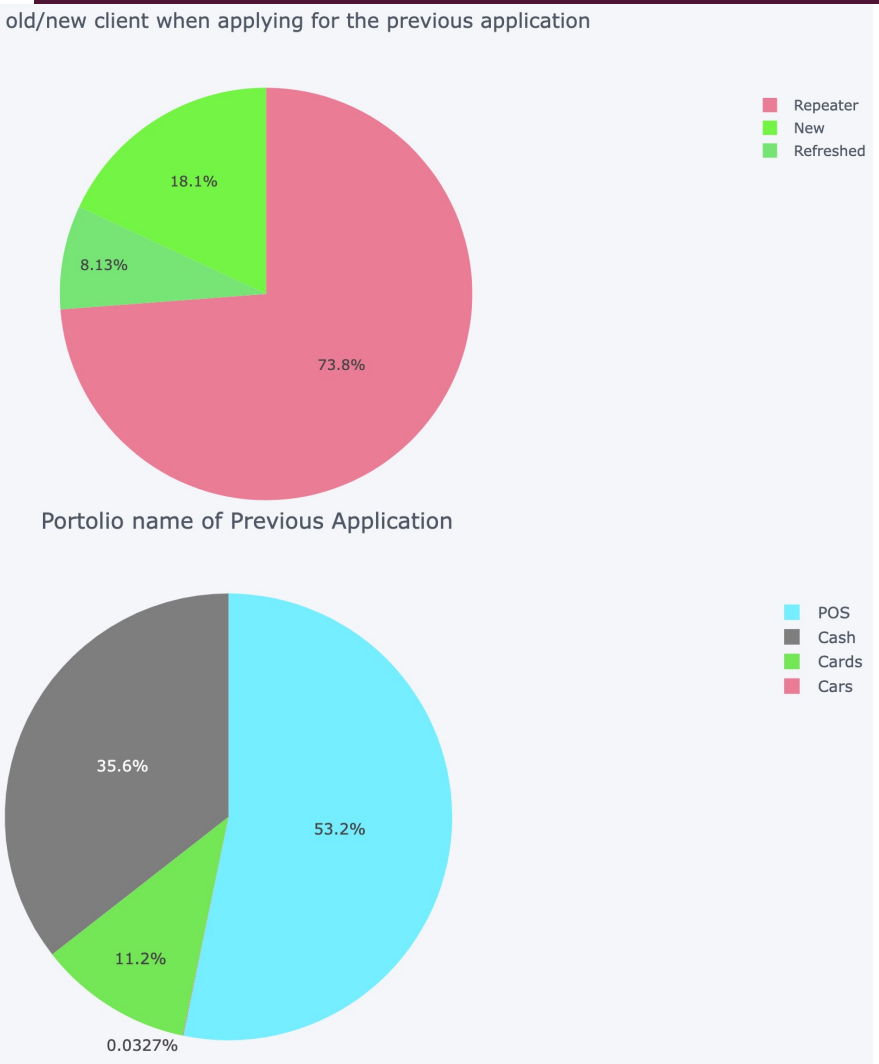


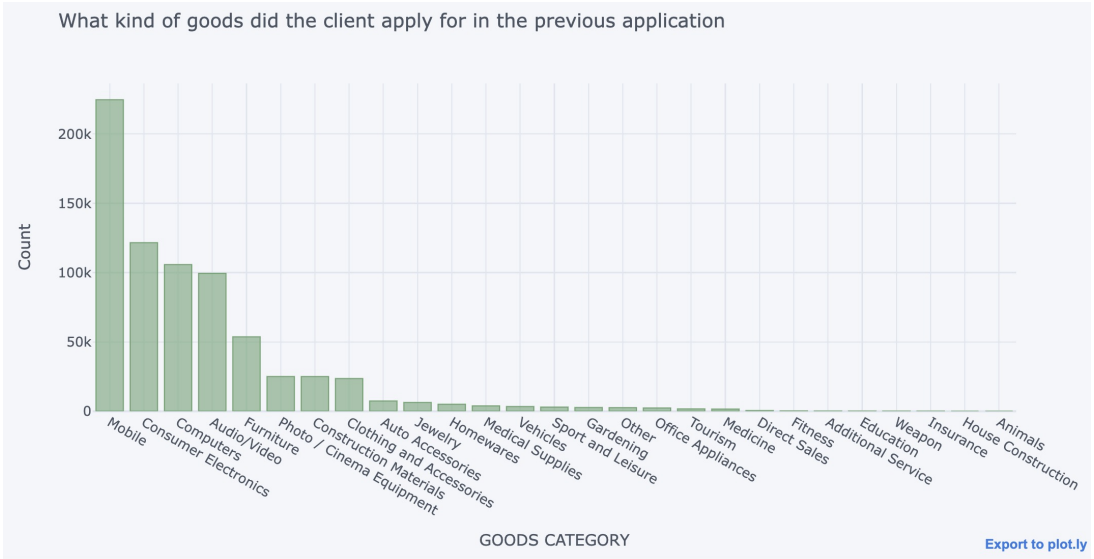
- Majority of loans are approved and very less percentage of loans are unused offer
- Majority of them have applied in start of week and less on weekends
- 99% of the clients chose to pay cash through bank in previous application
- HC is major reason for rejection





PREVIOUS APPLICATION DATA ANALYSIS

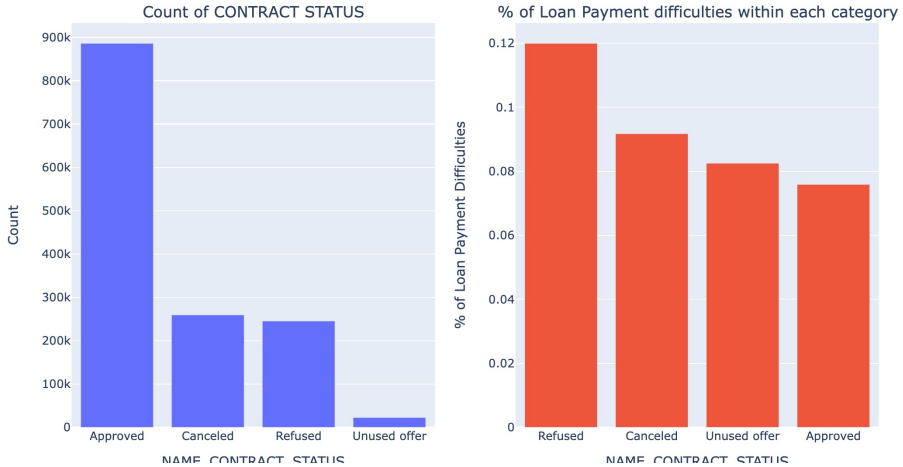




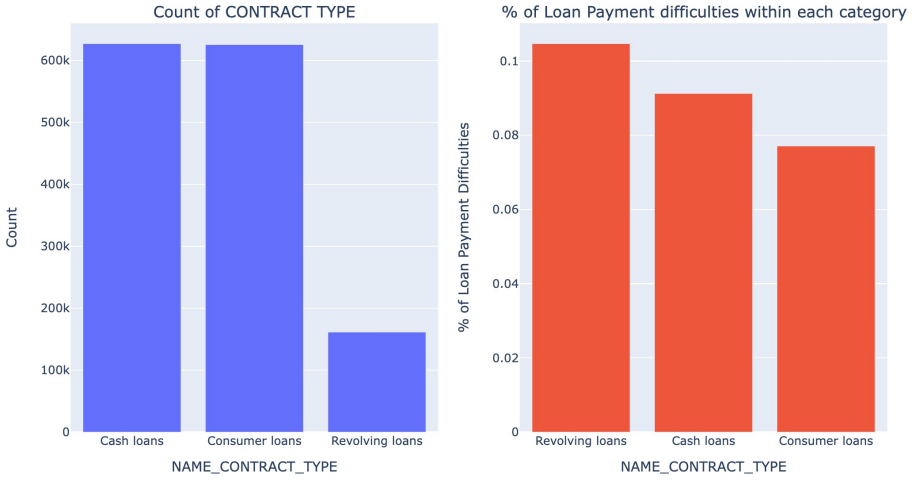
- Major are old client in previous application
- Majority of previous application is for POS and a good amount of it is for cash
- Majority of loans are for mobiles, consumer electronics, computers and furniture's

PREVIOUS APPLICATION DATA ANALYSIS

CONTRACT STATUS

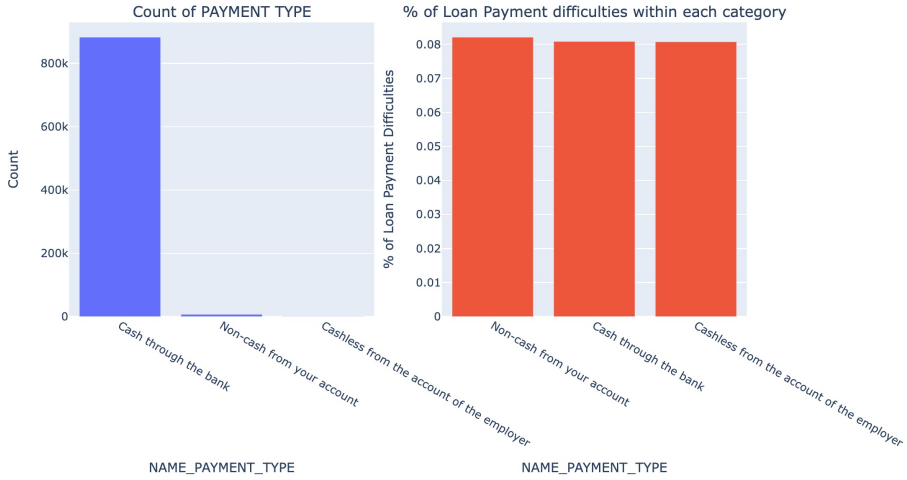


CONTRACT TYPE

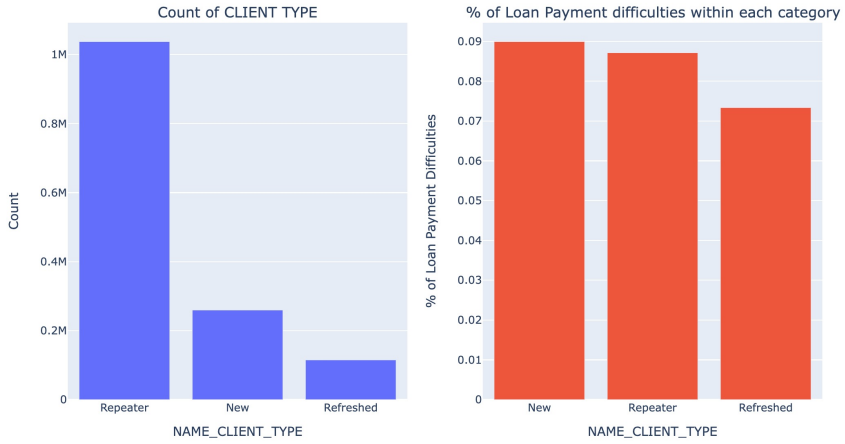


- Refused' contracts from previous application are the ones who have maximum percent of Loan-Payment Difficulties from current application.
- Approved' contracts from previous application are the ones who have minimum percent of Loan-Payment Difficulties from current application
- Revolving Loans contracts 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
- all three types of payments from the previous application have almost same % of Loan-Payment Difficulties from current application
- New clients from previous application are the ones who have maximum % of Loan-Payment Difficulties from current application.
- Refreshed clients from previous application are the ones who have minimum % of Loan-Payment Difficulties from current application.

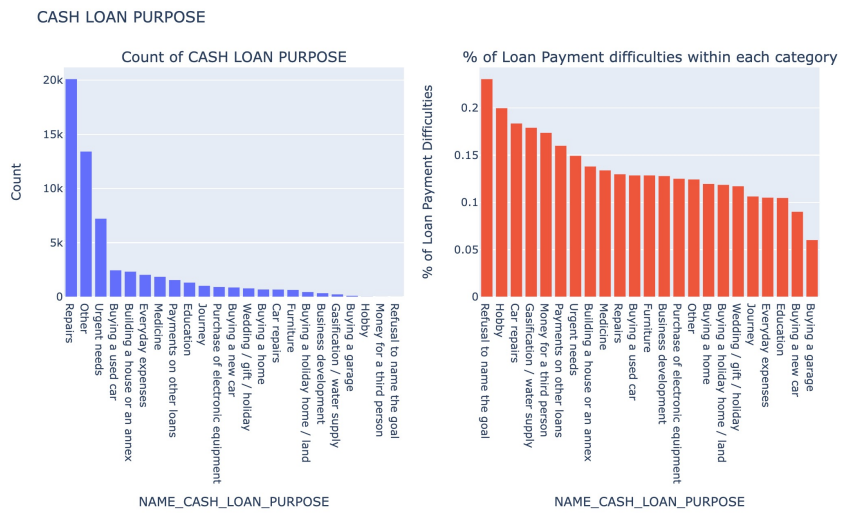
PAYMENT TYPE

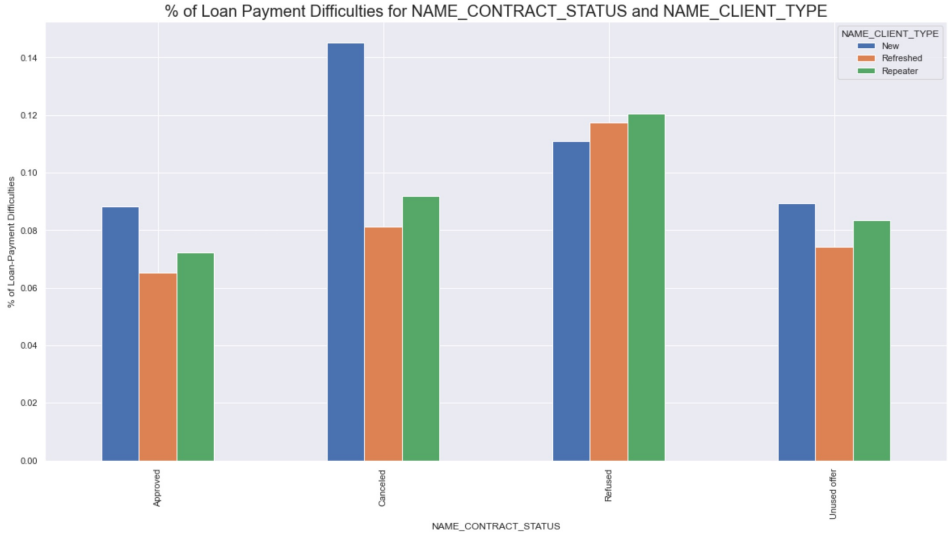


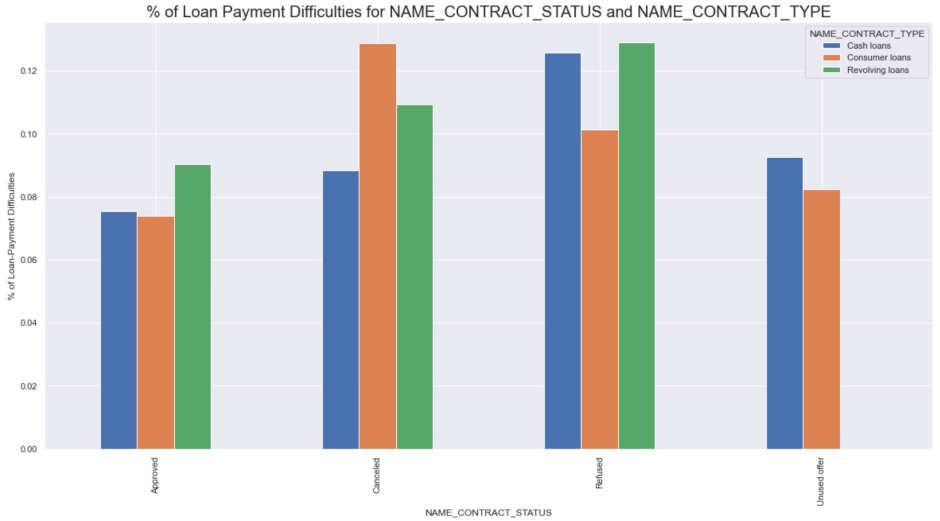
CLIENT TYPE



PREVIOUS APPLICATION DATA ANALYSIS







- purpose of cash loan from previous data was maximum for 'Repairs'
- Refusal to name the goal for cash loan from previous application are the ones who have maximum % of Loan-Payment Difficulties from current application.

- Client who where 'New' and had 'Cancelled' previous application tend to have more % of Loan-Payment Difficulties in current application

- Clients with 'Revolving loans' and with 'Refused' previous application tend to have more % of Loan-Payment Difficulties in current application

SUMMARY AND RECOMMENDATIONS

Application Data

- 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.

Previous Application Data

- 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.
- It can be observed from the graph that 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