



Credit EDA Case Study

Presented by:

- **Akash Verma**
- **Sachin Verma**



Business Objective



- This case study aims to identify patterns which indicate if a client has difficulty paying their installments 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.
- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.



Overall Solution Approach

- Step 1: Data Sourcing
 - a) Importing and Reading of Data.
 - b) Data Inspection.
 - c) Inspecting Null Values.
- Step 2: Data Cleaning & Manipulation
 - a) Checking for Data Quality Issues.
 - b) Binning of Continuous Variables & Categorical Variables.
- Step 3: Data Analysis
 - a) Check for Data Imbalance.
 - b) Univariate and Bivariate Analysis on both Application Data & Previous Application Data.
 - c) Drawing Insights from the Analysis.
- Step 4: Conclusion
 - a) Drawing conclusions from Analysis done on the given data.

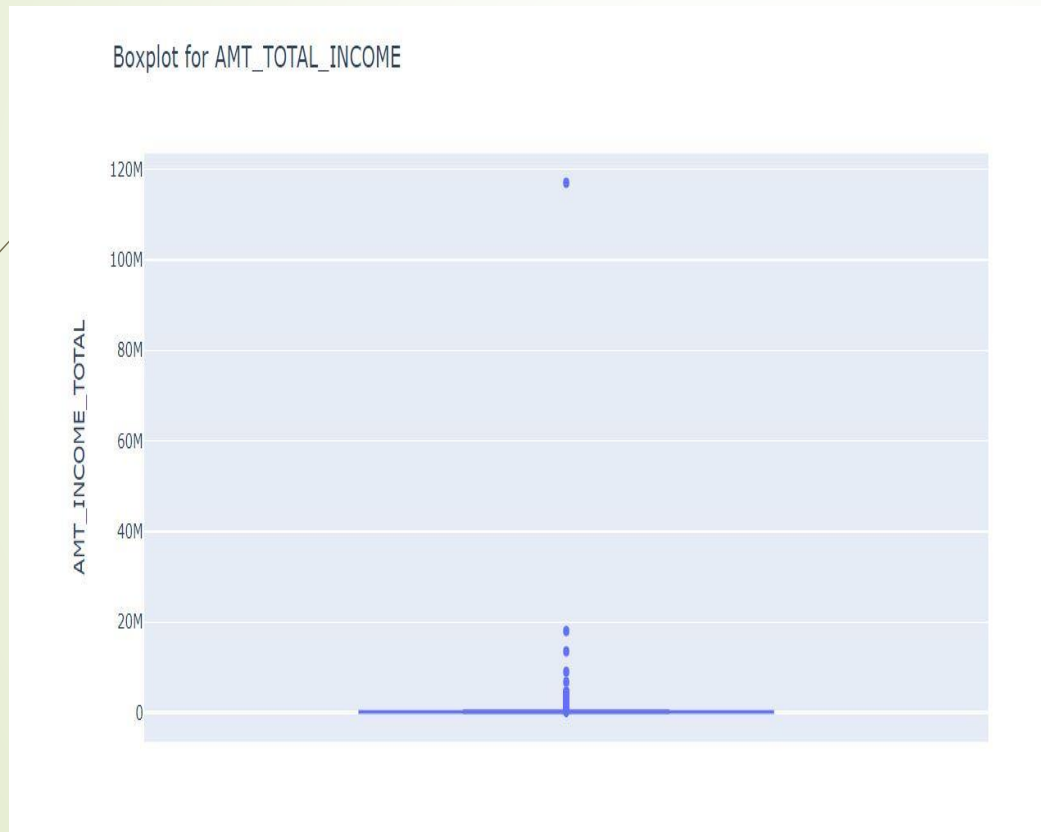


I. APPLICATION DATA



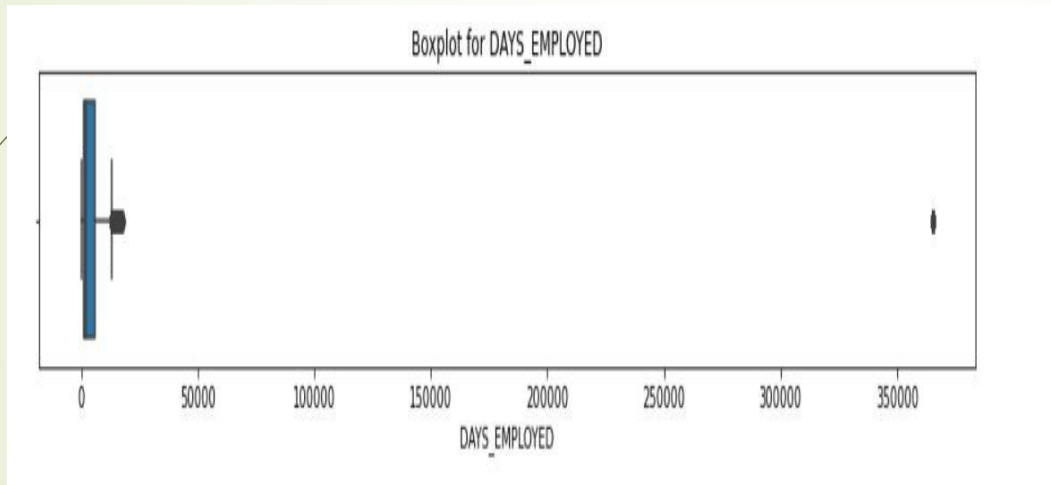
Check for Outliers in Data

Boxplot for AMT_TOTAL_INCOME



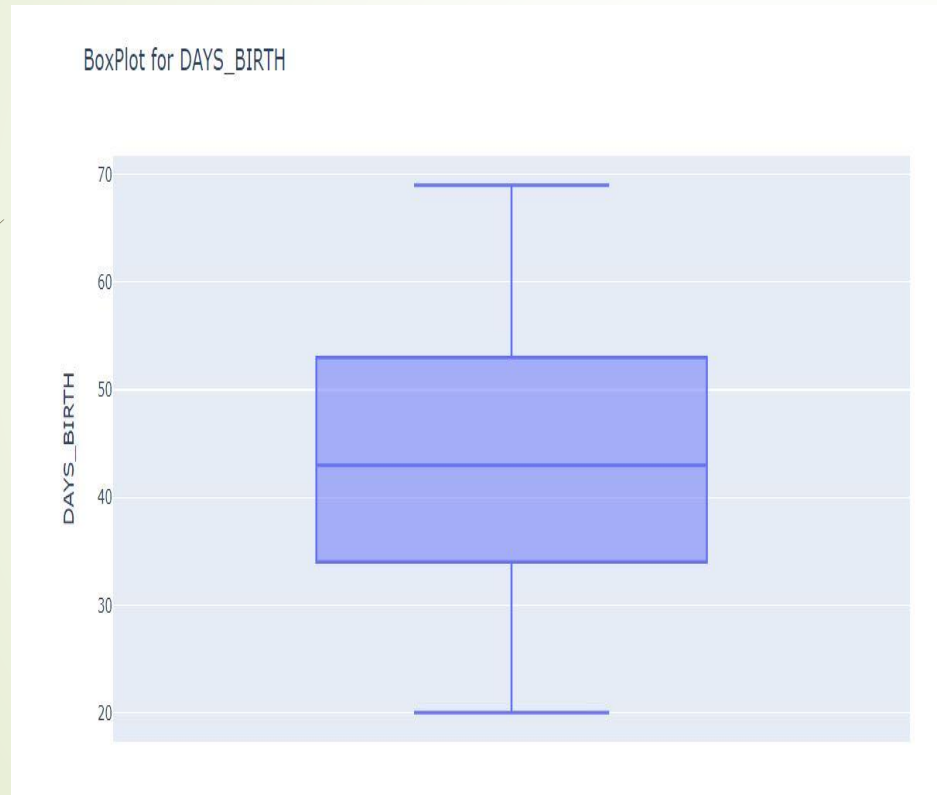
- Here AMT_INCOME_TOTAL is the income of the client. From the above boxplot we can observe that the point at 117M is an outlier.

Boxplot for DAYS_EMPLOYED



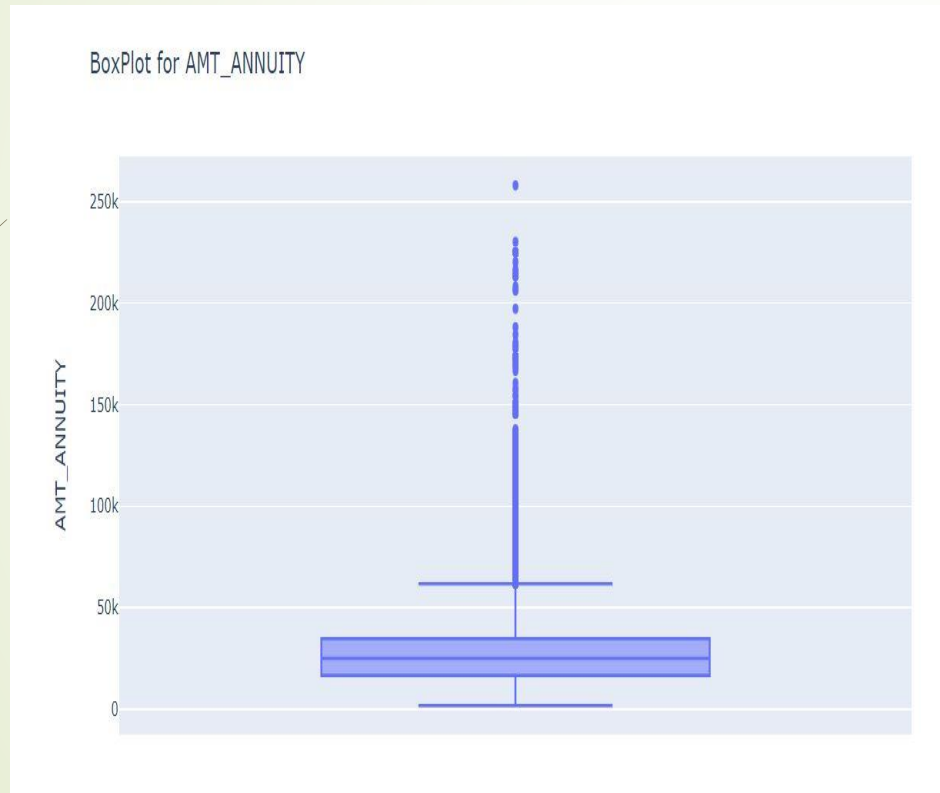
- Here DAYS_EMPLOYED column shows how many days before the application the person started current employment. We can clearly see a point above 350k which is clearly an outlier.

Boxplot for DAYS_BIRTH



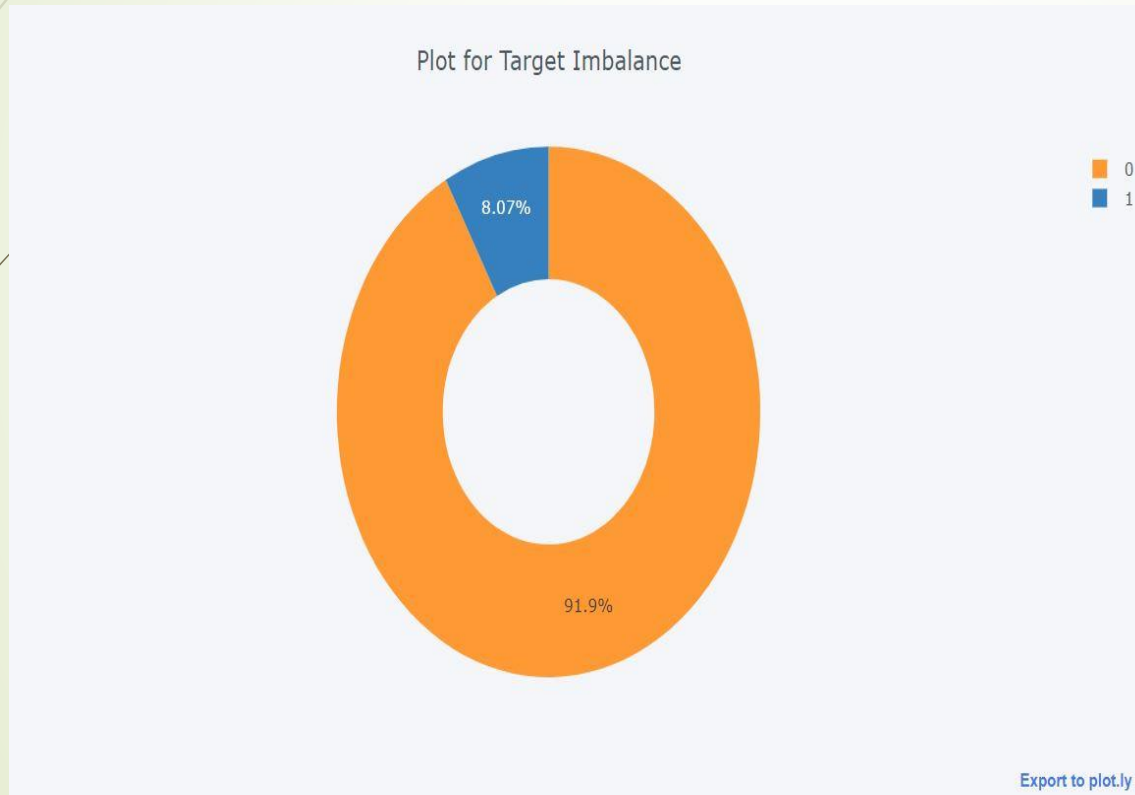
- As we can observe here in this column DAYS_BIRTH there's no outlier present.

Boxplot for AMT_ANNUIITY



- AMT_ANNUIITY colum shows loan annuity. Here we can observe a point at 258k which is an outlier.

Data Imbalance



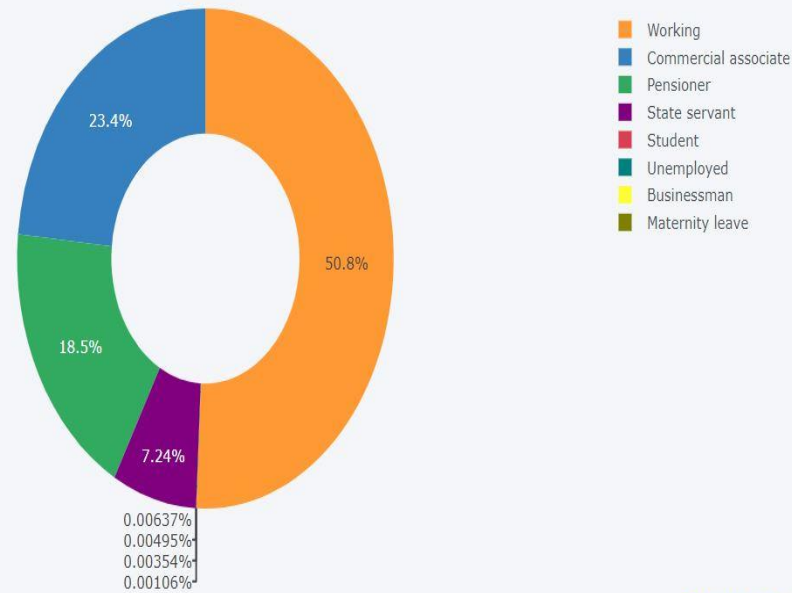
- We can observe there's high data imbalance between the target variables



Univariate Analysis on Application Data

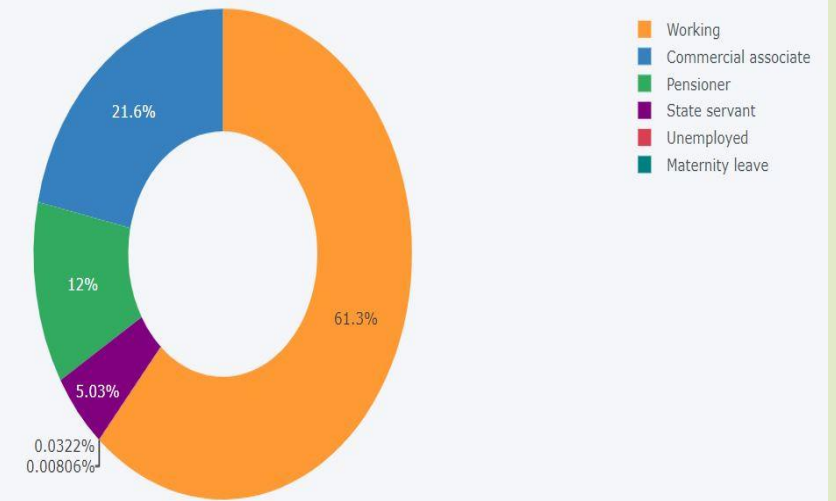
Income Source

Income Source Plot for Loan Non-Payment Difficulties



[Export to plot.ly »](#)

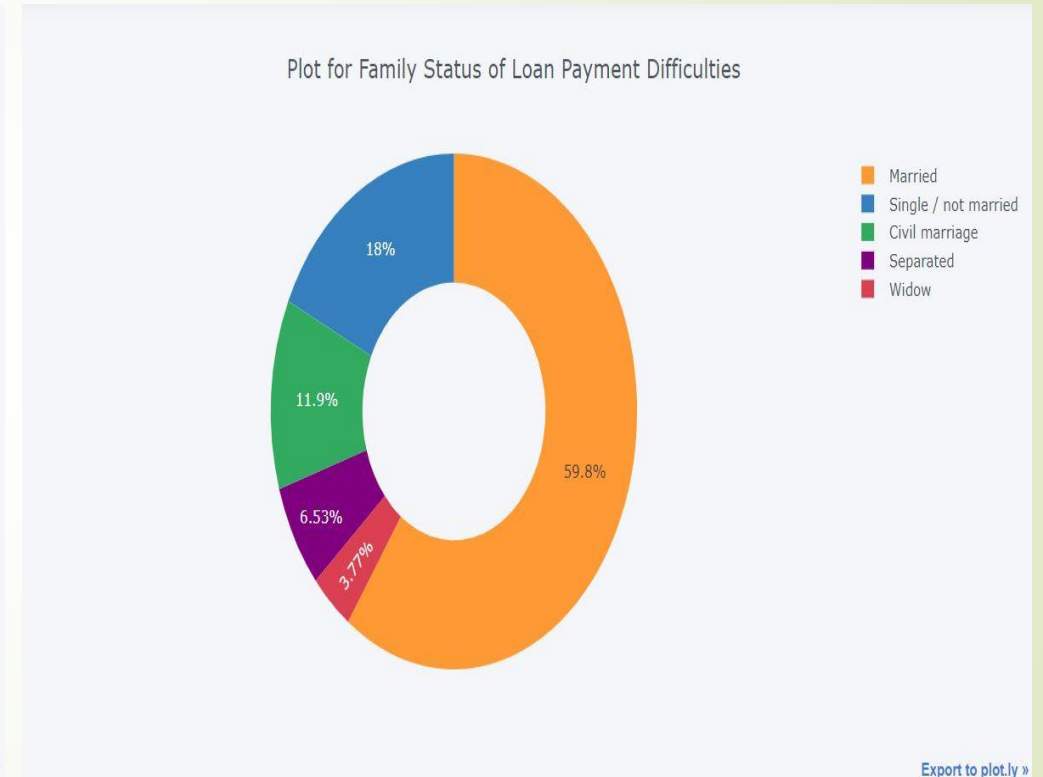
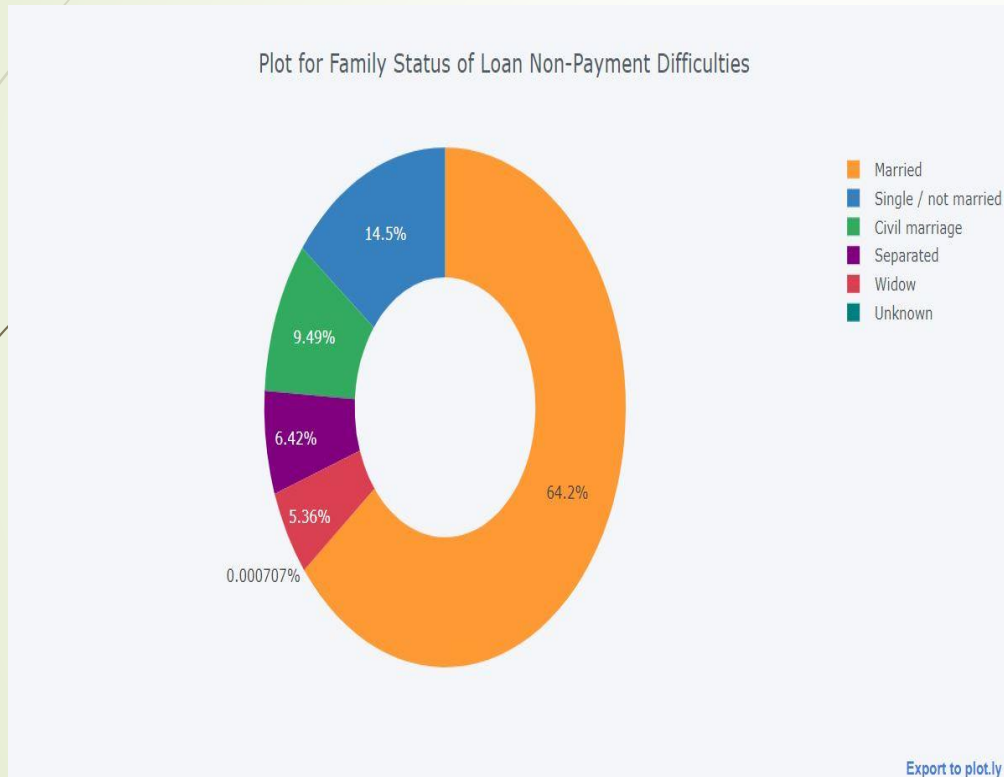
Income Source Plot for Loan Payment Difficulties



[Export to plot.ly »](#)

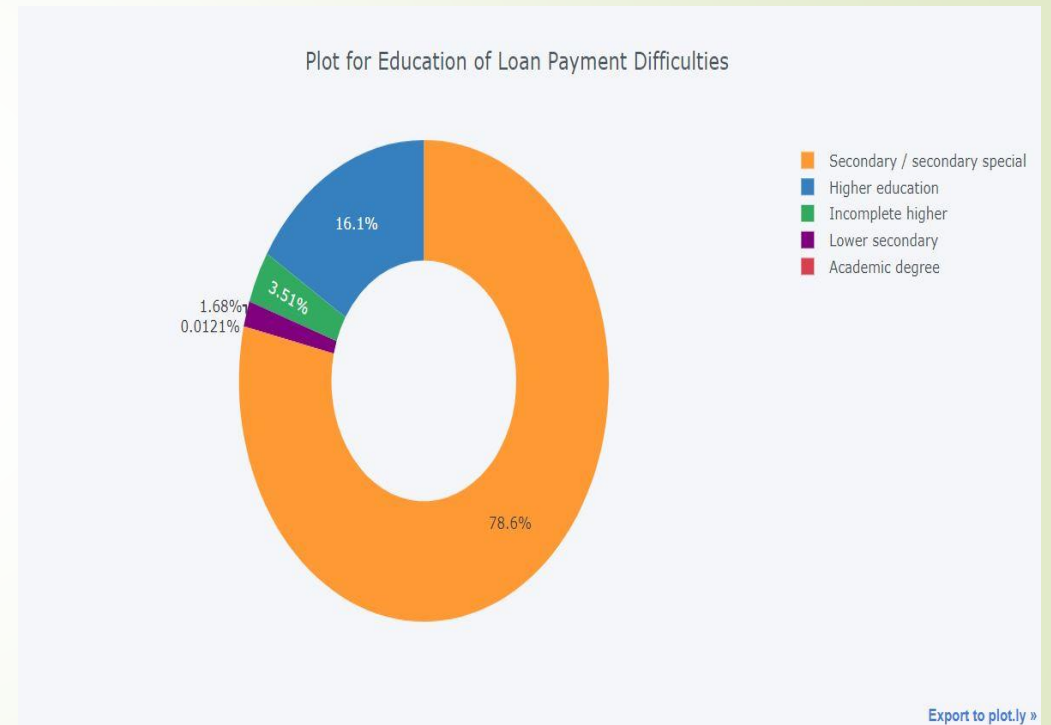
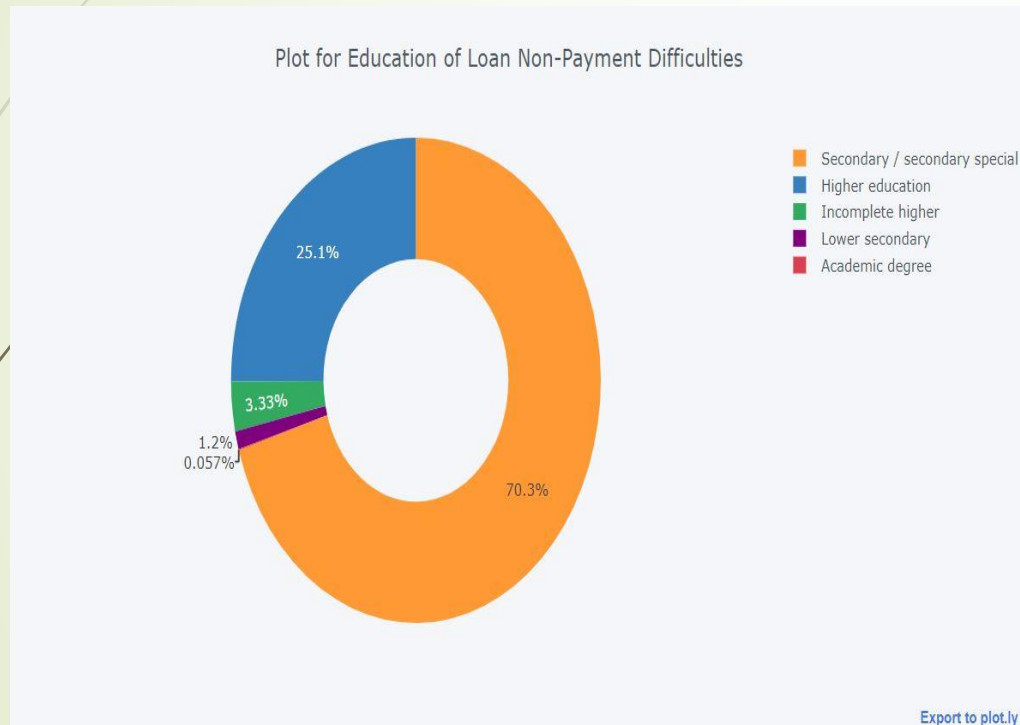
- ❖ From above graphs comparing both loan non-payment difficulties and loan payment difficulties we can observe that there's decrease in percentage of loan payment difficulties for pensioner, commercial associates and state servants whereas we can observe increase in percentage of loan payment difficulties for working clients.

Family Status



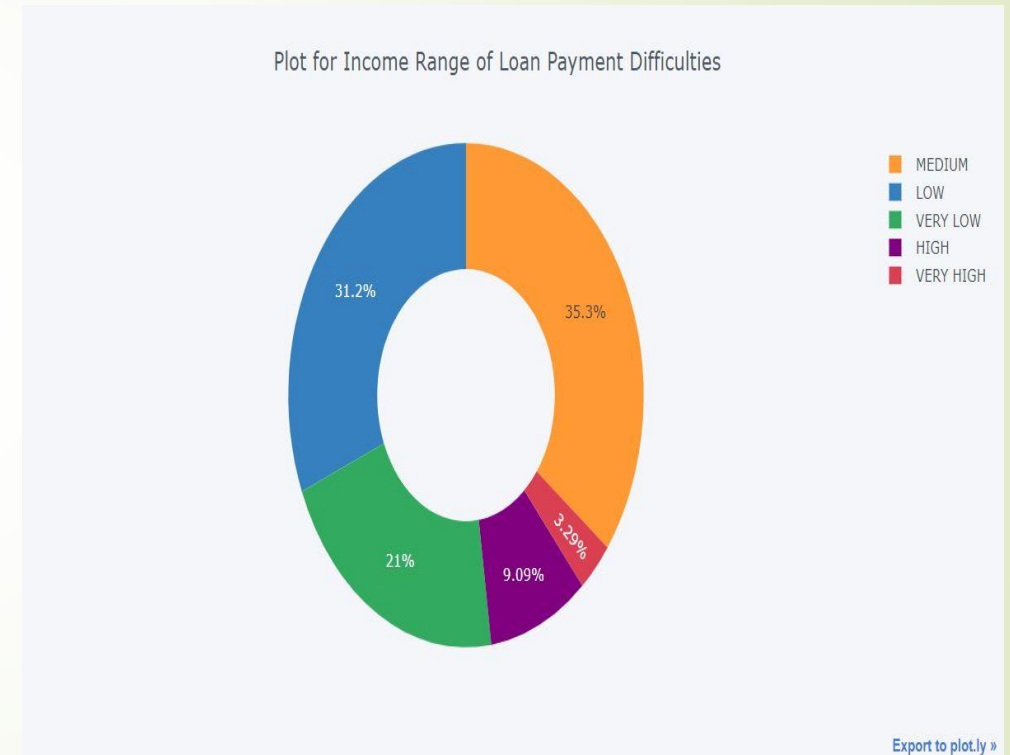
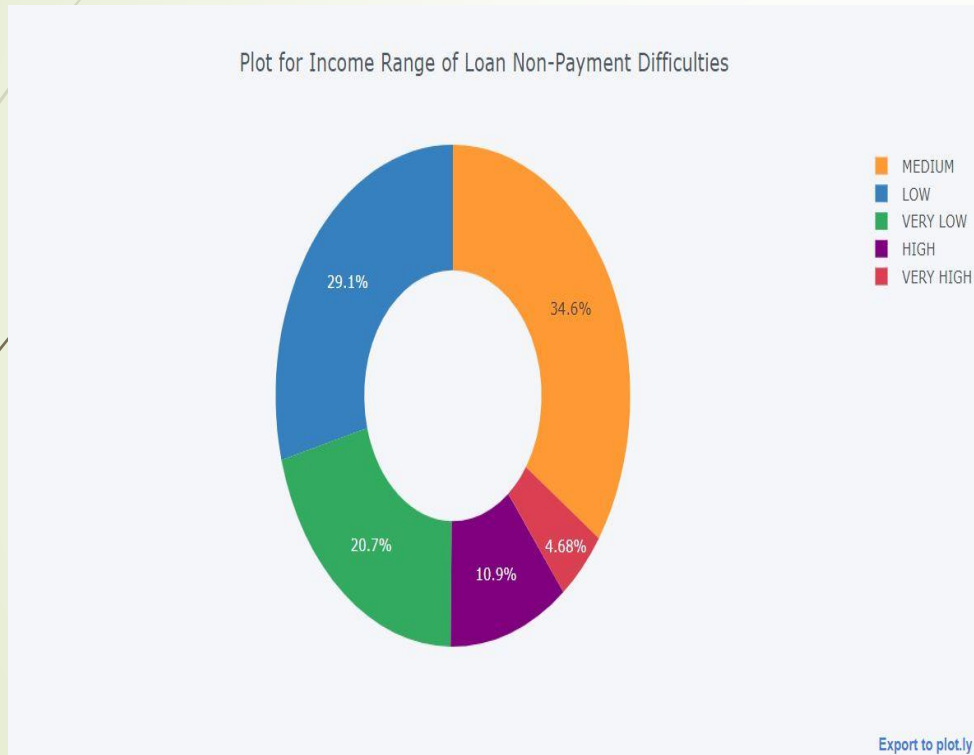
- ❖ By comparing both loan non-payment difficulties and loan payment difficulties graph we can observe that there's a decrease in percentage of married and widowed people with loan difficulties whereas there's increase in percentage of single/not married, civil married and separated loan payment difficulties.

Education Qualification



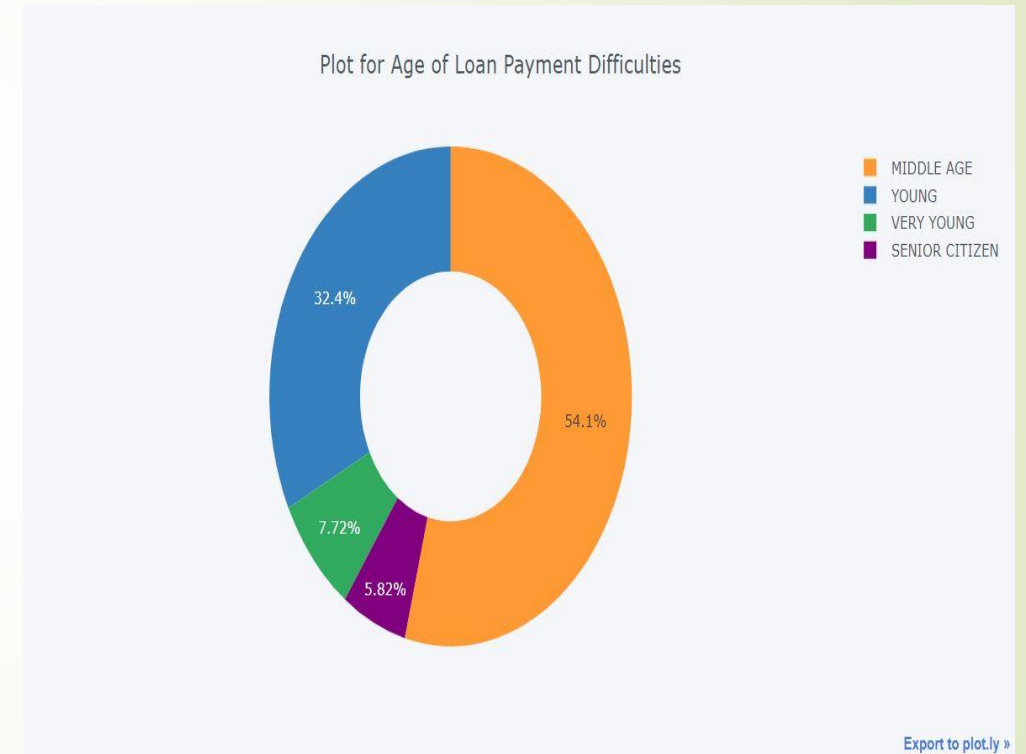
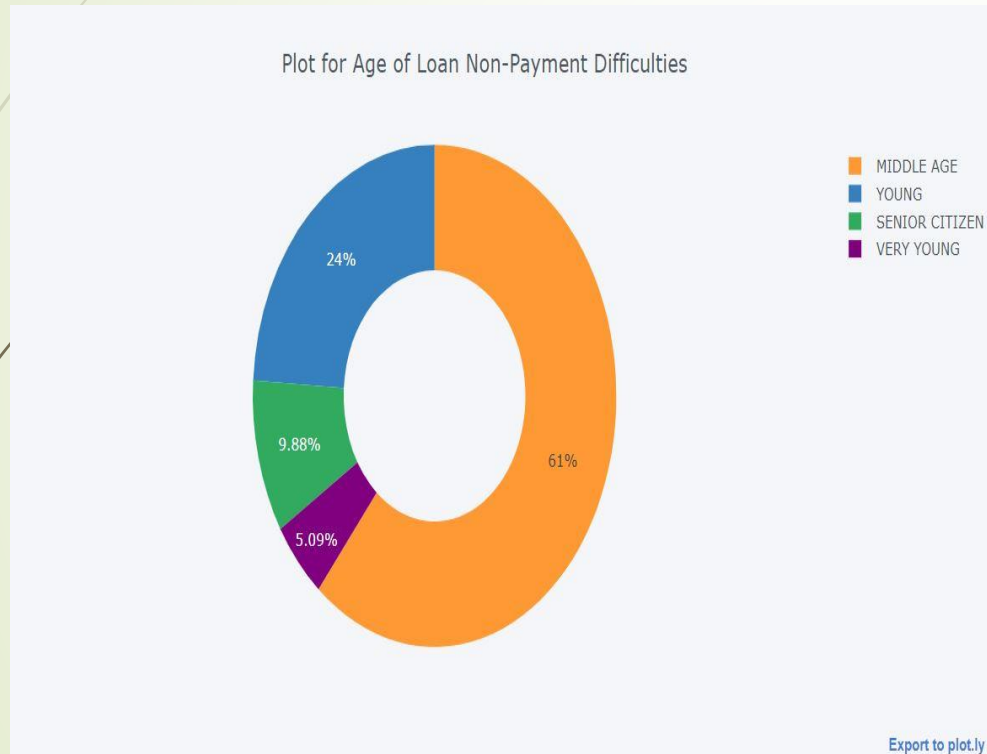
- ❖ By comparing both loan non-payment difficulties and loan payment difficulties graph we can observe that percentage of loan payment difficulties increased for clients with educational qualification of secondary/secondary special, incomplete higher and lower secondary. Whereas percentage of loan payment difficulties decreased for clients with educational qualifications of higher education and academic degree.

Income Range



- ❖ By comparing both loan non-payment difficulties and loan payment difficulties graph we can observe that percentage of loan payment difficulties increased for clients whose income is medium, low and very low. Whereas there's a decrease in loan payment difficulty for clients whose income is very high and high.

Age

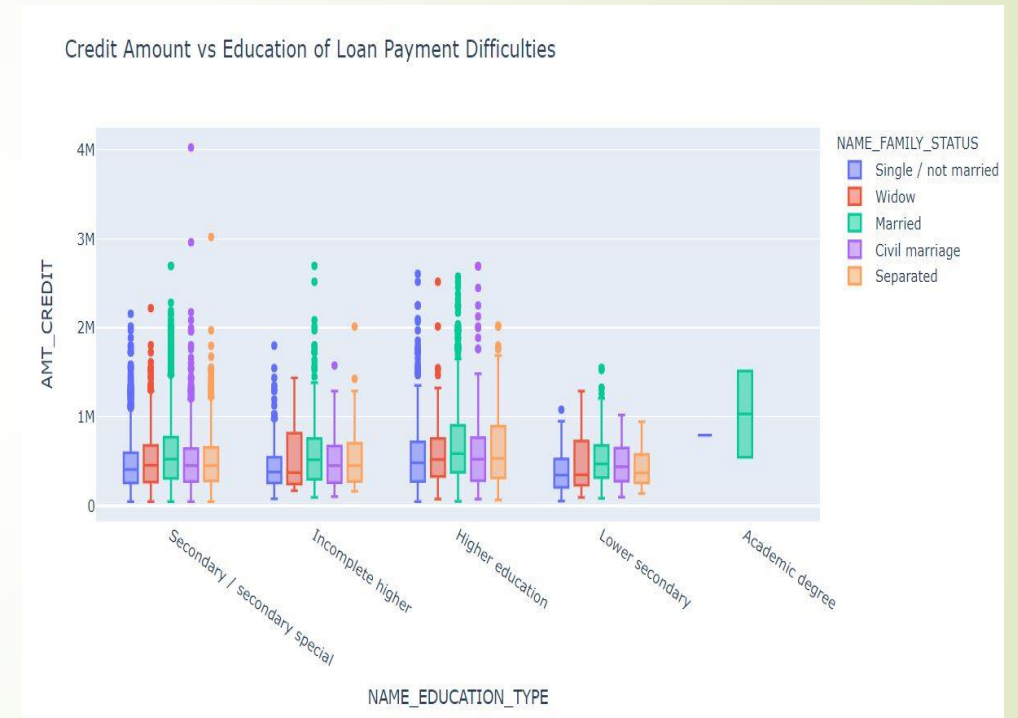
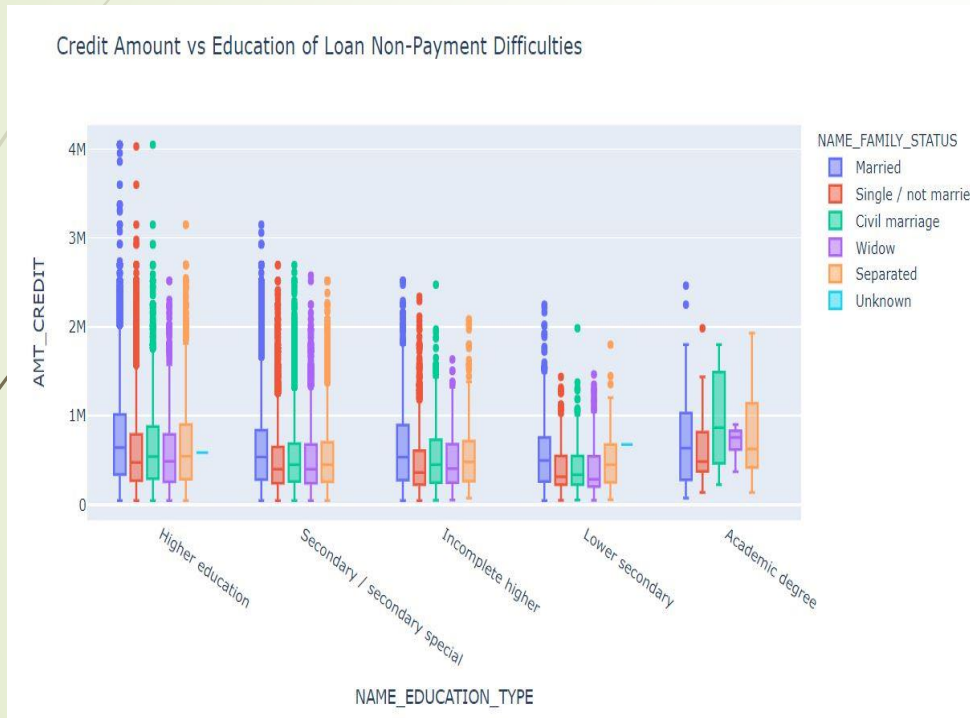


By comparing both loan non-payment difficulties and loan payment difficulties graph we can observe that percentage of loan payment difficulties increased for clients who's young and senior citizen. Whereas percentage of loan payment difficulties decreased for clients who's middle age and very young.



Bivariate Analysis of Categorical vs Numerical Variables

Credit Amount vs Education

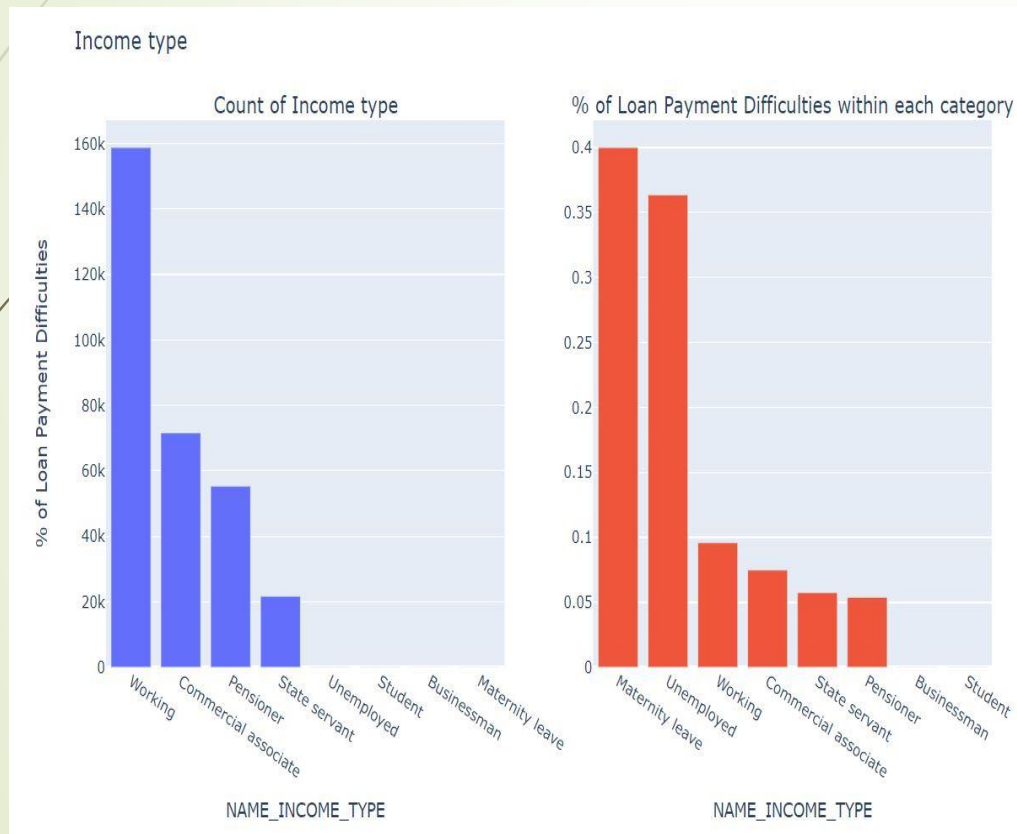


- ❖ The plot for Loan Payment Difficulties and Loan Non-Payment Difficulties is almost similar. From it we can observe and draw insight that Family status of Civil marriage and Separated of Academic degree Education are having higher number of credits than others. We can also observe that most of the outliers are from Education Type - Higher Education and Secondary. Civil marriage for Academic degree is having most of the credits in the third quartile.



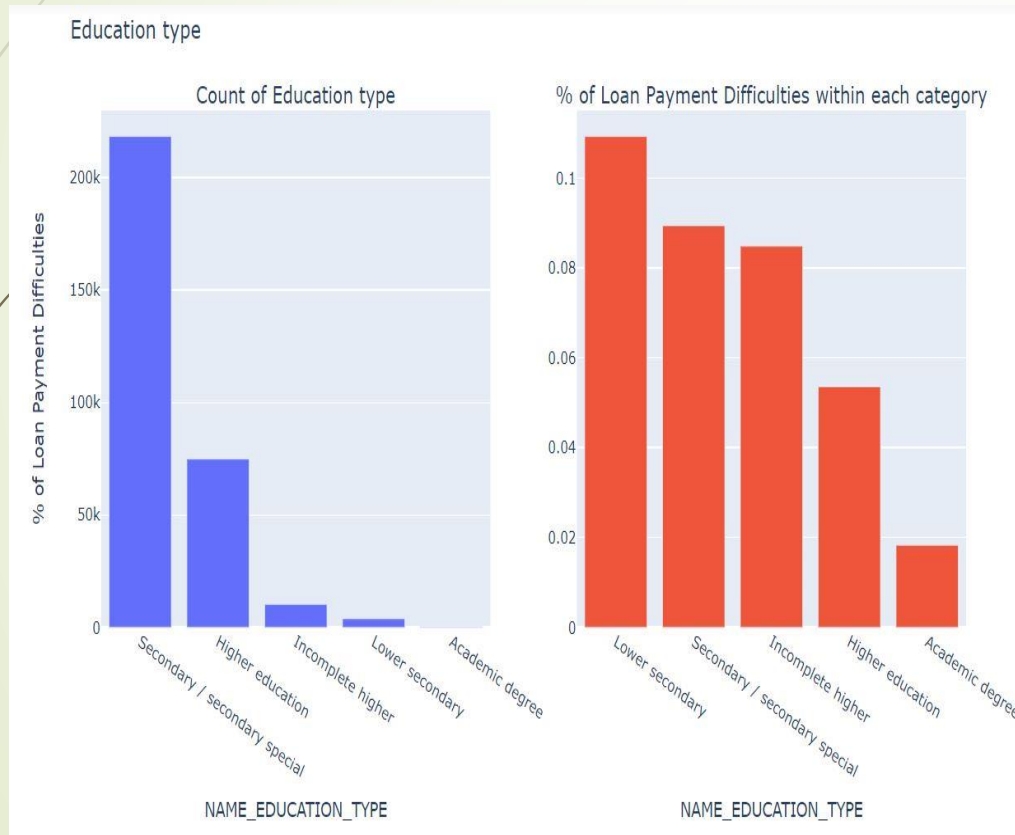
Bivariate Analysis of Categorical vs Categorical Variable

Distribution of income range and the category type with maximum loan payment difficulties



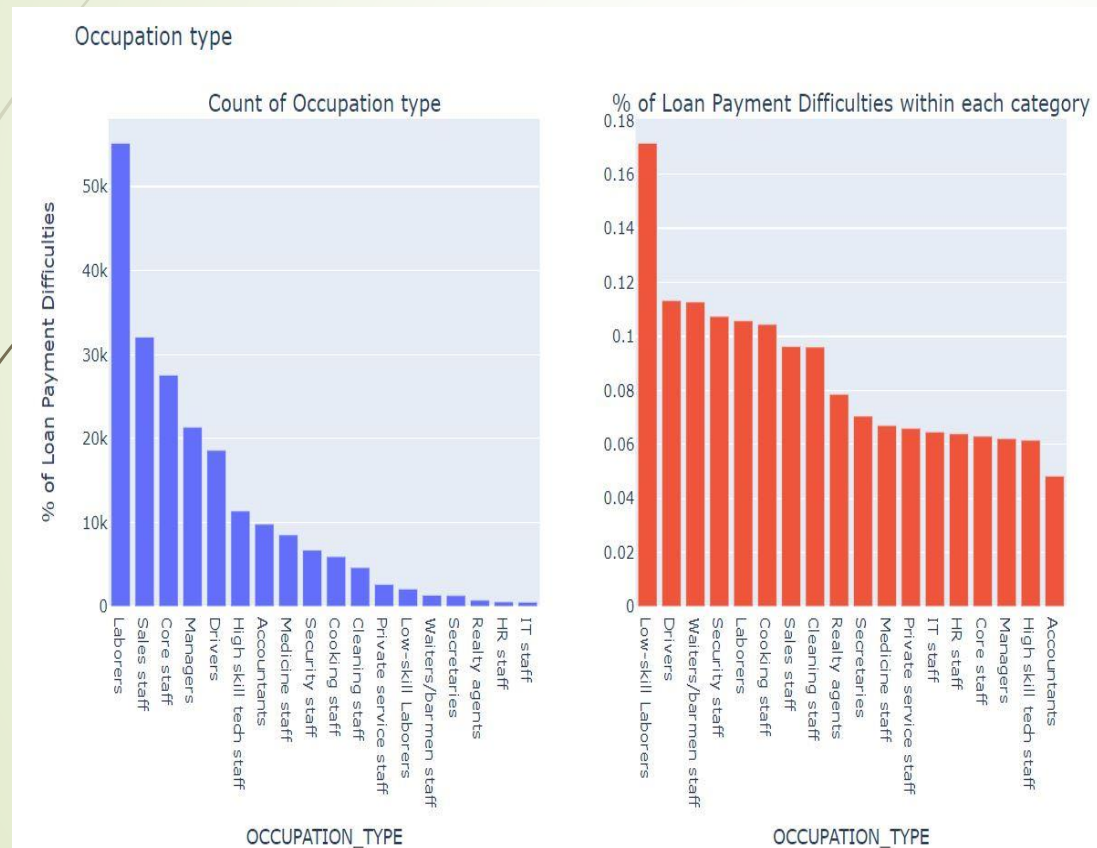
- From the graph we can observe that Maternity leave has the maximum percentage of loan payment difficulties.

Plot for Education type and category with max Loan Payment Difficulties



- From the graph we can observe that Lower Secondary education type has the highest percentage of loan payment difficulties

Plot for Occupation Type and category with max Loan Payment Difficulties

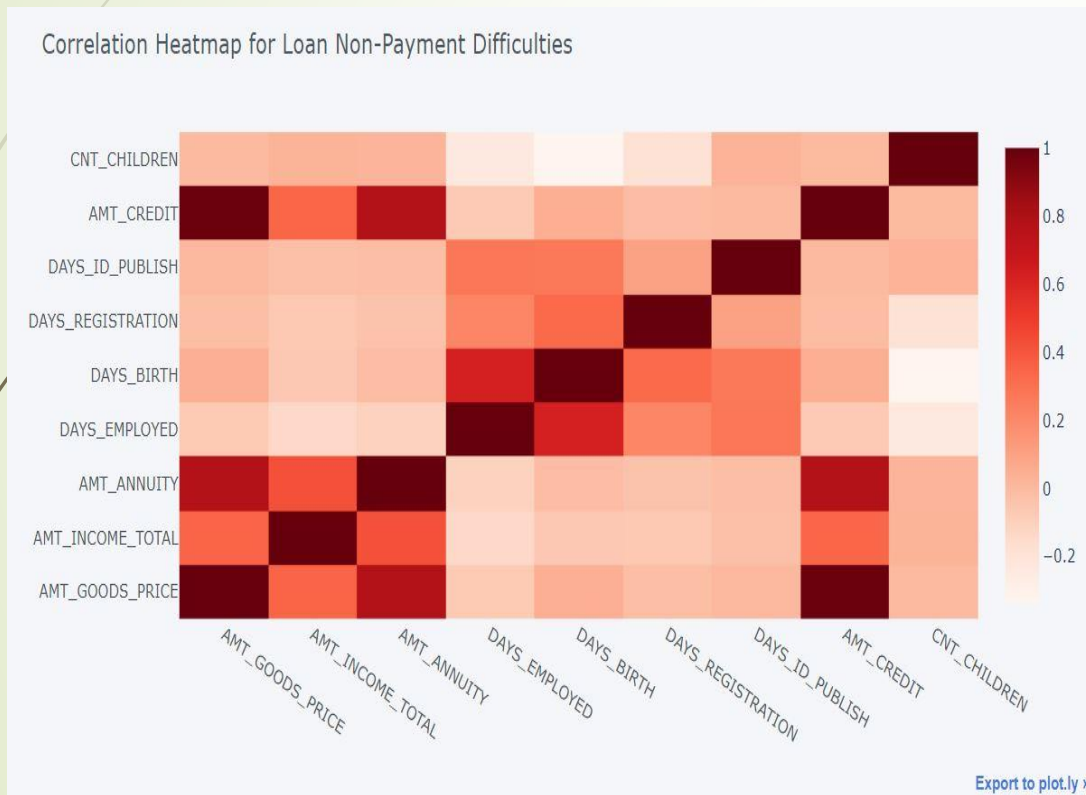


- From the graph we can observe that Low-skill Laborers occupation type has the highest percentage of loan payment difficulties



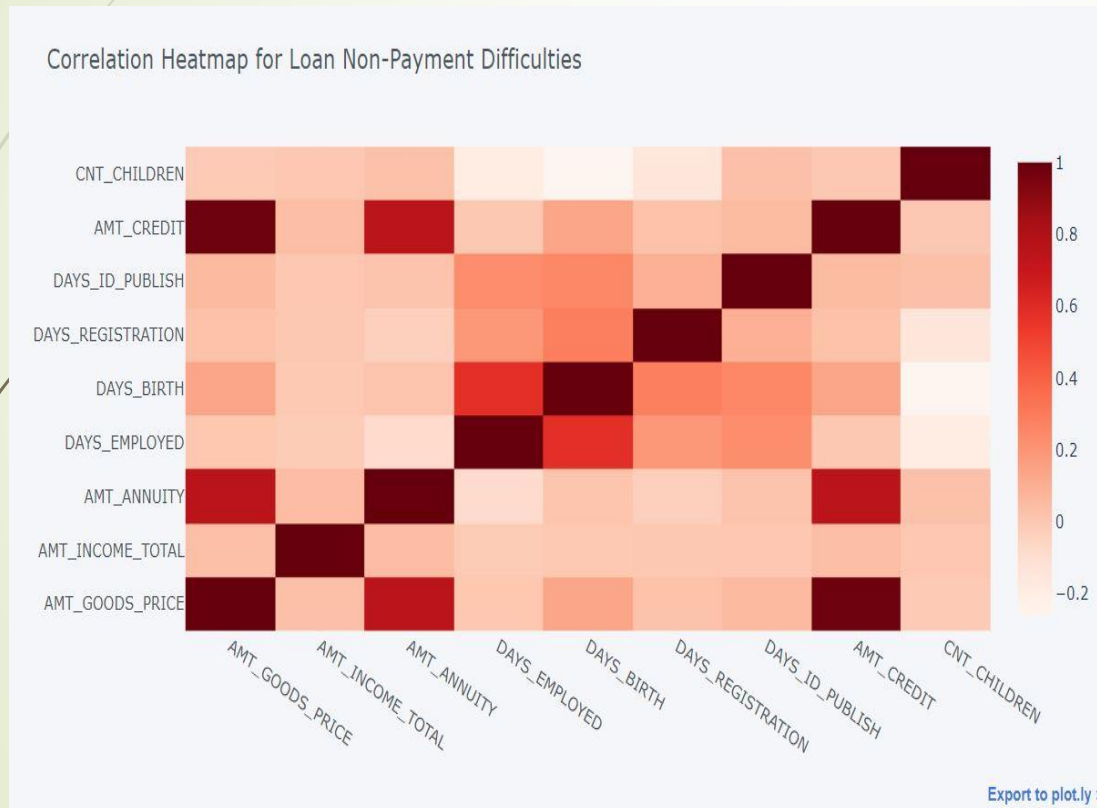
Correlation for Numerical columns for both Target Cases

Heatmap for Loan Non-Payment Difficulties



- ❖ From this correlation heatmap we can draw the following insights:
- ❖ Credit amount is inversely proportional to date of birth i.e. credit amount is higher for low age and vice versa.
- ❖ Credit amount is inversely proportional to number of children a client has i.e. credit amount is higher for less children count a client has and vice-versa.
- ❖ Income amount is inversely proportional to the number of children a client has i.e. more income for less children a client has and vice versa.

Heatmap for Loan Payment Difficulties



- ❖ From this correlation heatmap we can draw the following insights:
- ❖ We can observe there's high correlation between credit amount and goods price.
- ❖ We can also observe some deviancies in the correlation of loan payment difficulties and loan non-payment difficulties such as credit amount vs income.

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

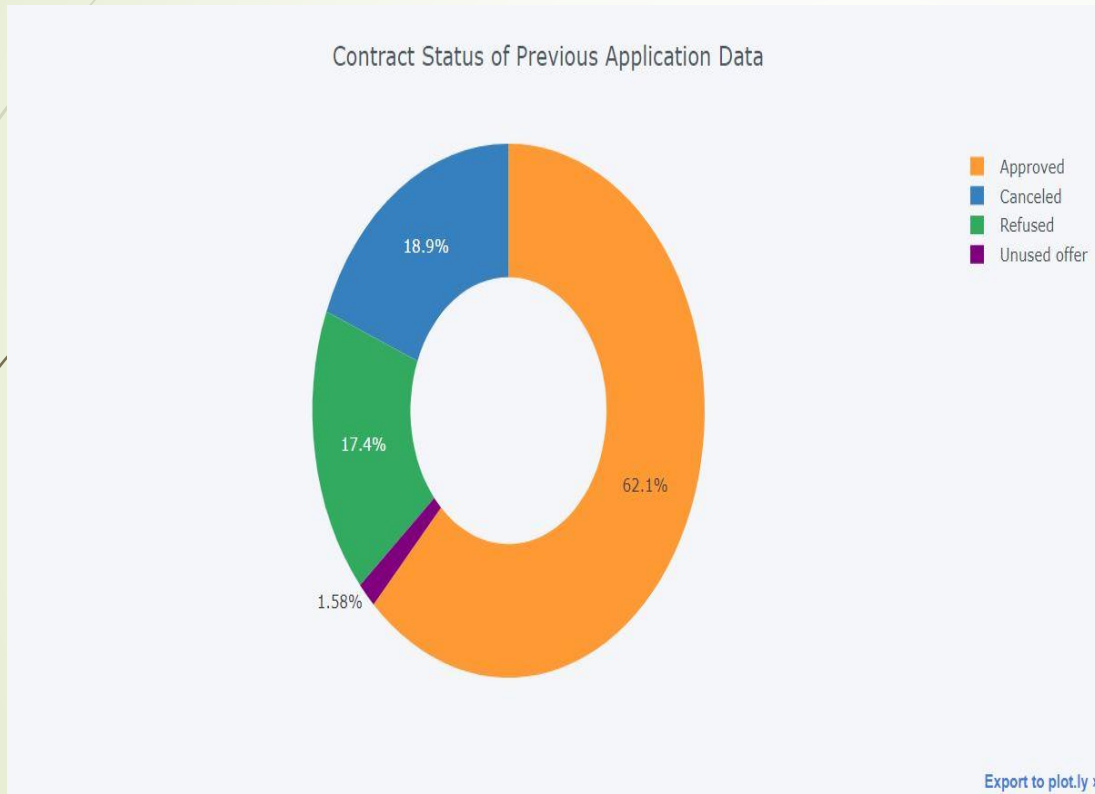


II. PREVIOUS APPLICATION DATA



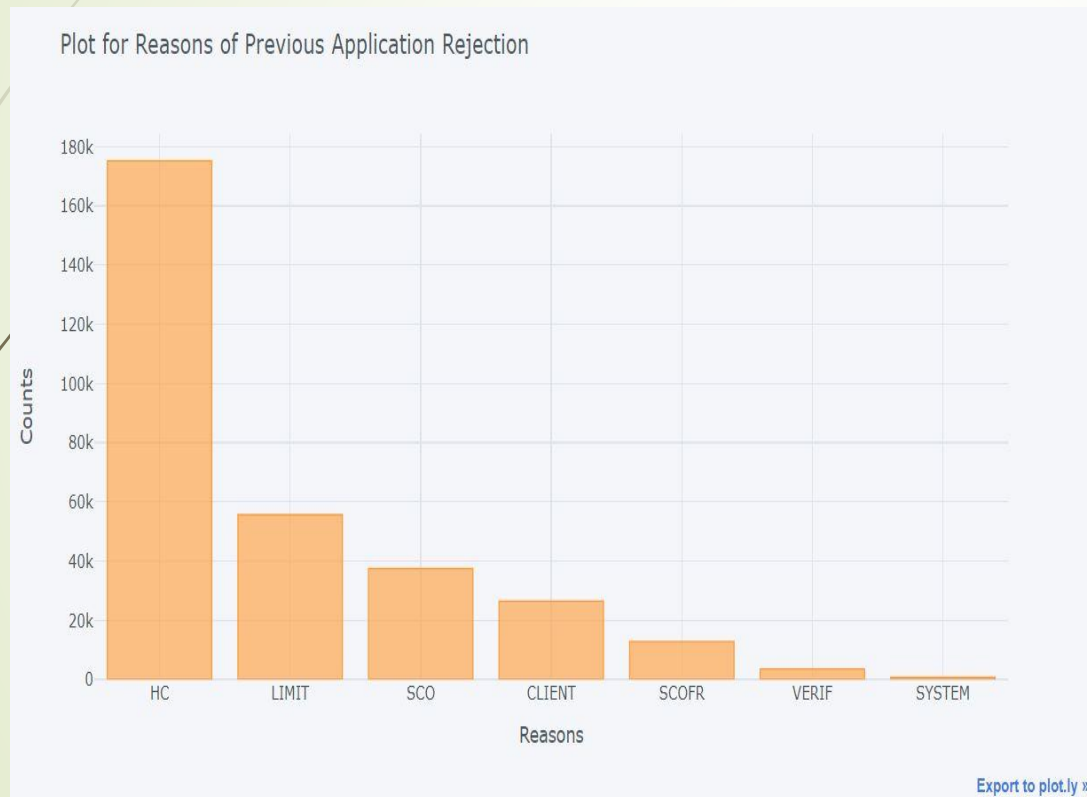
Univariate analysis on Previous Application Data

Previous Application Data's Contract Status



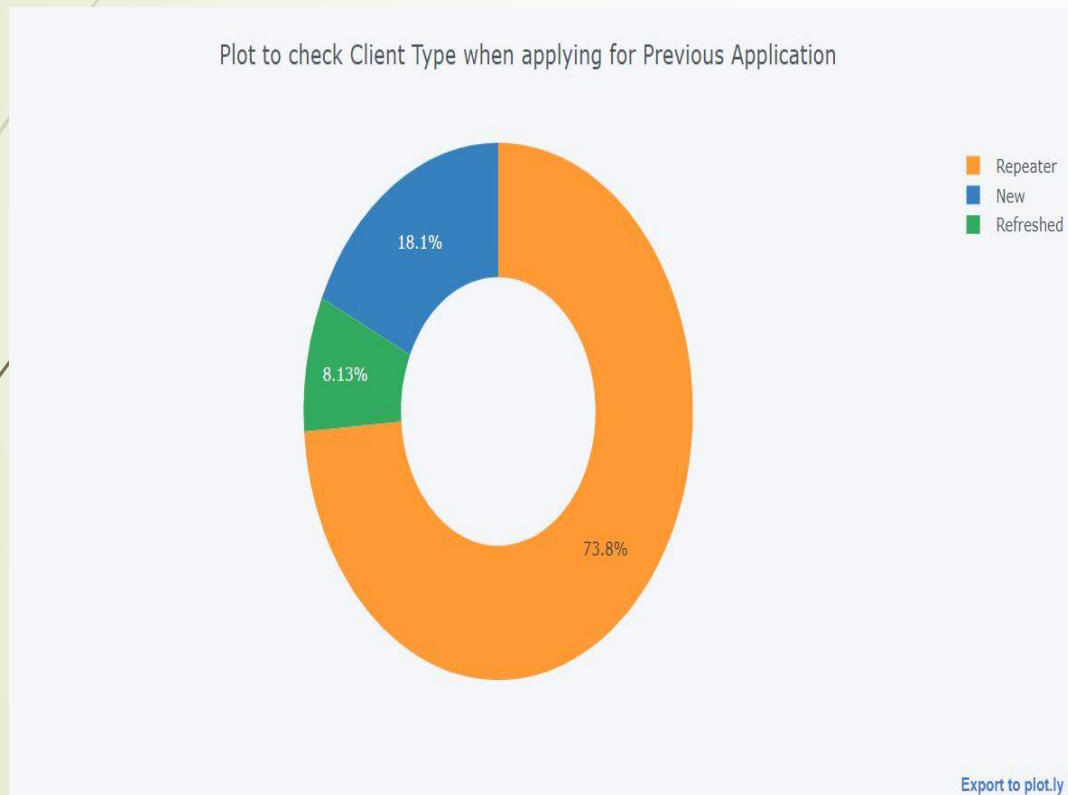
- From the graph we can observe that majority of loans are Approved and very less percentage of loans are Unused offer.

Reasons for Previous Application Rejection



- As we can observe from the graph that majority of applications got rejected due to HC reason i.e. precisely above 175k applications got rejected due to this reason.

Client Type when applying for Previous Application

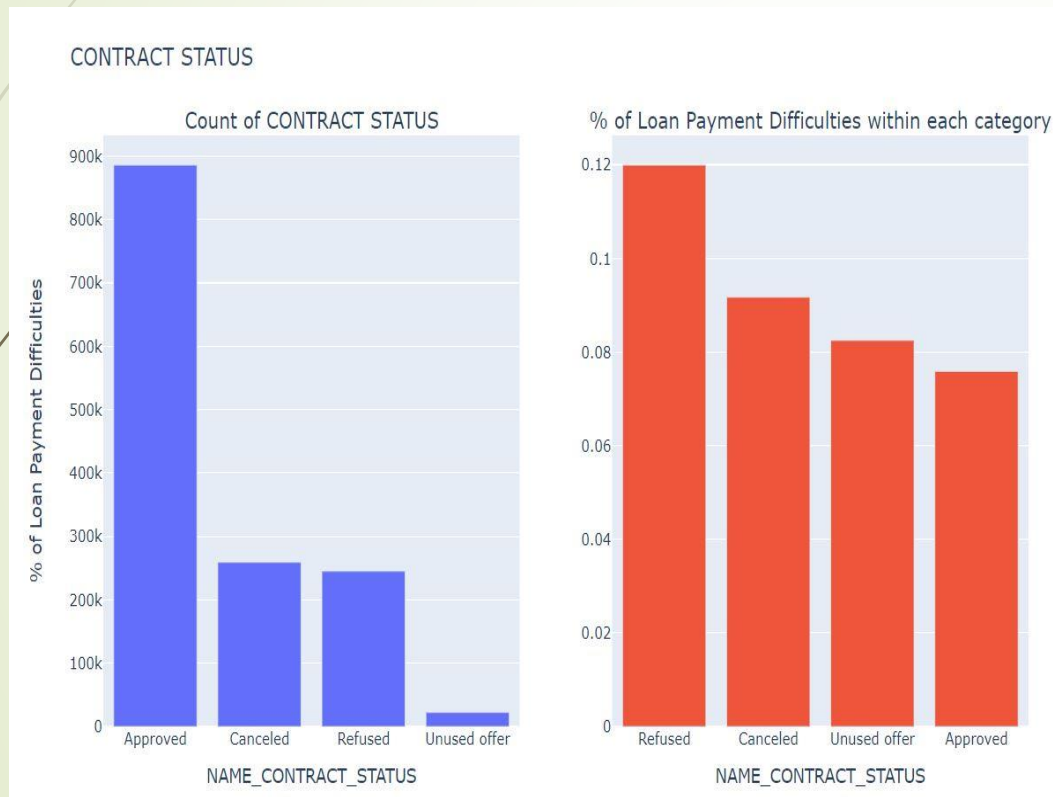


- We can observe in the above graph that majority of clients are Repeater i.e. around 73.8% .



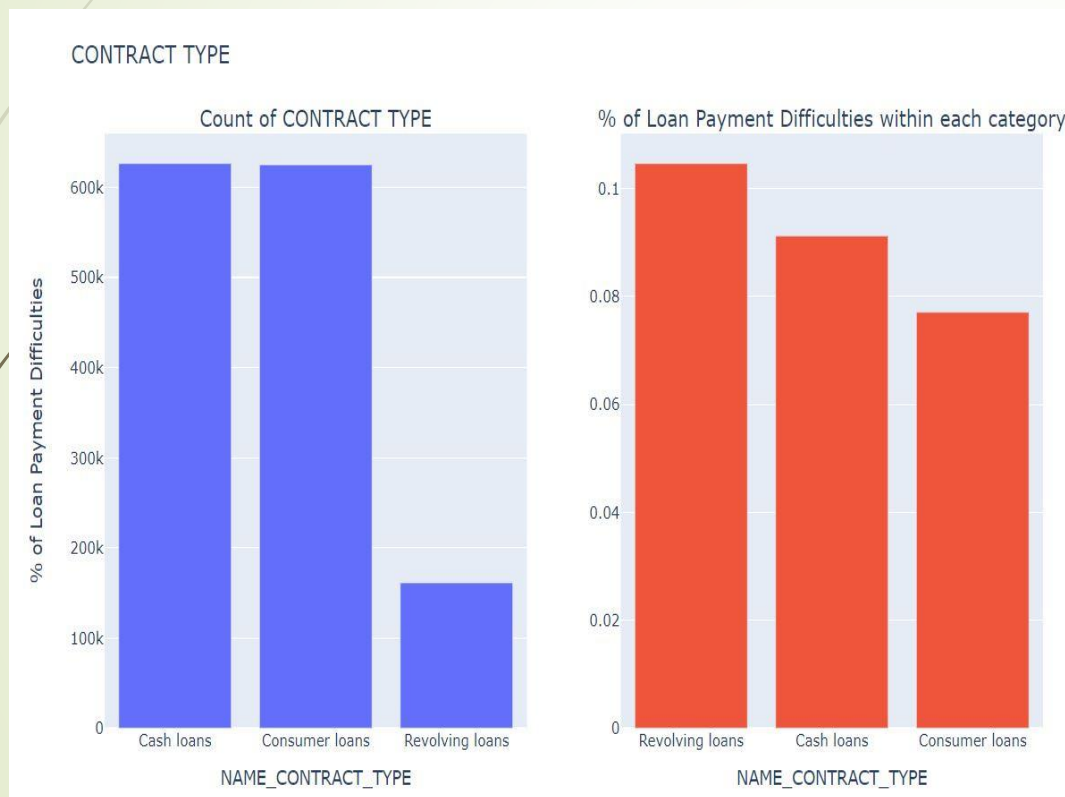
Bivariate Analysis by merging Application Data & Previous Application Data

Contract Status and it's category with max percentage of Loan Payment Difficulties



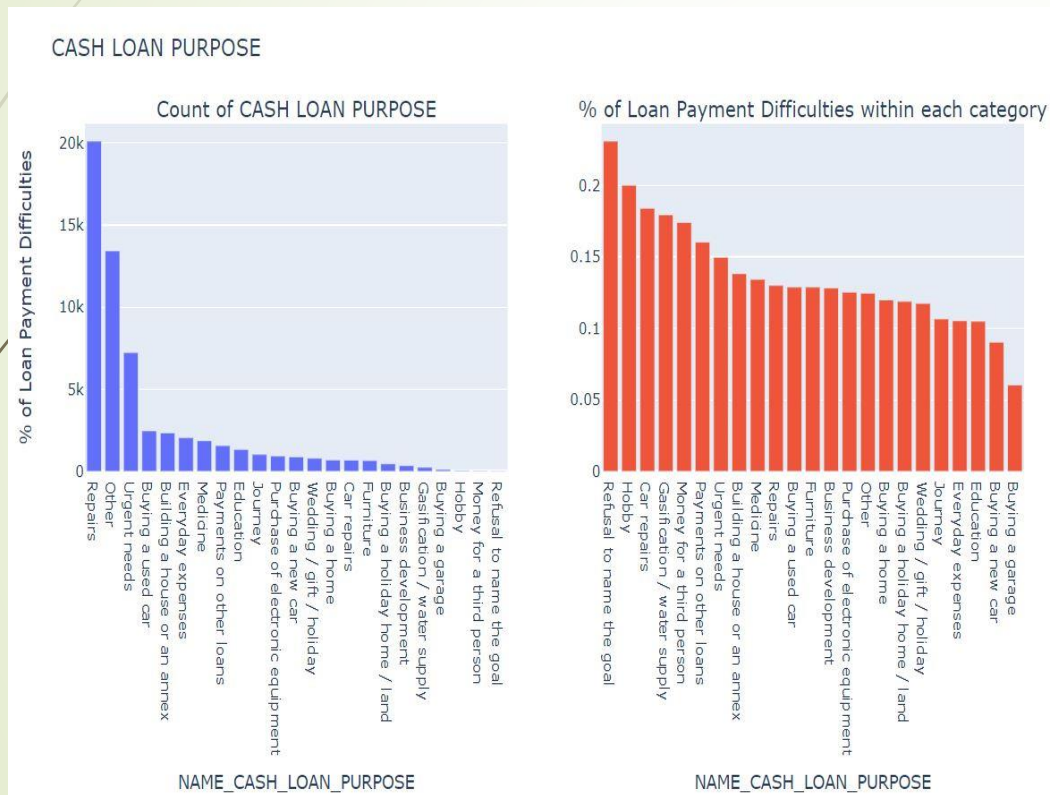
- ❖ We can observe from the first graph that most of the application from previous application has been approved by the bank.
- ❖ From the second graph we can draw the following observations:
 - ❖ From previous application Refused contracts are the ones who have maximum percentage of loan payment difficulties from current application.
 - ❖ From previous application Approved contracts are the ones who have minimum percentage of loan payment difficulties from current application.

Contract Type and it's category with max percentage of Loan Payment Difficulties



- ❖ From the first graph we can observe that most of the contract type from previous application was 'Car loans'
- ❖ From the second graph we can draw the following observations:
 - ❖ From the previous application, Revolving loans contract are the ones who have maximum percentage of loan payment difficulties from current application.
 - ❖ From the previous application, Consumer loans contract are the ones who have minimum percentage of loan payment difficulties from current application

Cash Loan Purpose and it's category with max percentage of Loan Payment Difficulties



- ❖ From 1st graph it can be observed that purpose of cash loan in previous application was maximum for Repairs
- ❖ From the 2nd graph we can observe that:
 - ❖ Refusal to name the goal for cash loans from previous application are the ones who have maximum percentage of loan payment difficulties in current application.
 - ❖ Buying a garbage for cash loans from previous application are the ones who have minimum percentage of loan payment difficulties in current application.

% of Loan Payment Difficulties for NAME_CONTRACT_STATUS & NAME_CLIENT_TYPE'



- From this graph we can observe that clients that were new and canceled their previous application tend to have more percentage of loan payment difficulties in current application.

% of Loan Payment Difficulties for NAME_CONTRACT_STATUS & NAME_CONTRACT_TYPE



- From this graph we can observe that clients with Revolving loans and Refused previous application tend to have more percentage of loan payment difficulties in current application.



INSIGHTS



From Application Data

- ❖ We've observed that the count of 'Maternity leave' in 'NAME_INCOME_TYPE' column is very less but it has maximum percentage of loan payment difficulties i.e. around 40%. Hence we can conclude that, clients with income type as 'Maternity leave' are most likely to default or I can say clients with income type as 'Maternity leave' are the driving factors for Loan Defaulters.
- ❖ We've also observed that count of 'Lower Secondary' in 'NAME_EDUCATION_TYPE' is comparatively very less and it also has maximum percentage of loan payment difficulties i.e. around 11%. Hence we can conclude that, client with education type as 'Lower Secondary' are most likely to default or I can say clients with education type as 'Lower Secondary' are the driving factors for Loan Defaulters.
- ❖ We've also observed that count of 'Low-Skill Laborers' in 'OCCUPATION_TYPE' is comparatively very less and it also has maximum percentage of loan payment difficulties i.e. around 17%. Hence we can conclude that, client with occupation type as 'Low-Skill Laborers' are most likely to default or I can say clients with occupation type as 'Low-Skill Laborers' are the driving factors for Loan Defaulters.



From Previous Application Data

- We've observed that count of 'Refused' in 'NAME_CONTRACT_STATUS' is comparatively very less and it also has maximum percentage of payment difficulties i.e. around 12%. Hence we can conclude that, clients with contract status as 'Refused' in previous application are most likely to default or I can say that clients with contract status as 'Refused' in previous application are driving factors for Loan Default.
- We've observed that count of 'Revolving Loans' in 'NAME_CONTRACT_TYPE' is comparatively very less and it also has maximum percentage of payment difficulties i.e. around 10%. Hence we can conclude that, clients with contract type as 'Revolving Loans' in previous application are most likely to default or I can say that clients with contract type as 'Revolving Loans' in previous application are driving factors for Loan Default.
- We've observed that count of 'Refusal to name the goal' in 'NAME_CASH_LOAN_PURPOSE' is comparatively very less and it also has maximum percentage of payment difficulties i.e. around 23%. Hence we can conclude that, clients who have 'Refused to name the goal' for Cash Loan in previous application are most likely to default or I can say that clients who have 'Refused to name the goal' for Cash Loan in previous application are driving factors for Loan Default.
- We've observed that clients with 'Revolving Loans' and with 'Refused' previous application are most likely to have higher percentage of payment difficulties in current application. Since the count of both 'Revolving Loans' and 'Refused' is comparatively less, hence we can conclude that clients with 'Revolving Loans' and 'Refused' previous application are most likely to default or I can say that clients with 'Revolving Loans' and 'Refused' previous application are driving factors of Loan Default.



CONCLUSION




Recommended Group(Less LIKELY to be defaulter)

- I. Clients working as a state servant.
- II. Old people of any income group.
- III. Client with high income category.
- IV. Old female client.
- V. Client with higher education(female).
- VI. Any client who's previous loan was approved.



Risky Group(More likely to be defaulter)

- I.** Lower secondary educated clients are the most in number to be defaulted when their previous loans were cancelled or refused.
 - II.** Male clients with civil marriage.
 - III.** Previously refused loan status group.
- 



THANK YOU