Credit EDA Case Study

Presentation

By SAPTARSHI MANDAL

Introduction

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.

BUSINESS UNDERSTANDING

When the company receives a loan application, the company has to decide for loan approval based on the applicant’s profile. Two types of risks are associated with the bank’s decision:

- 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.

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
- **All other cases:** All other cases when the payment is paid on time.

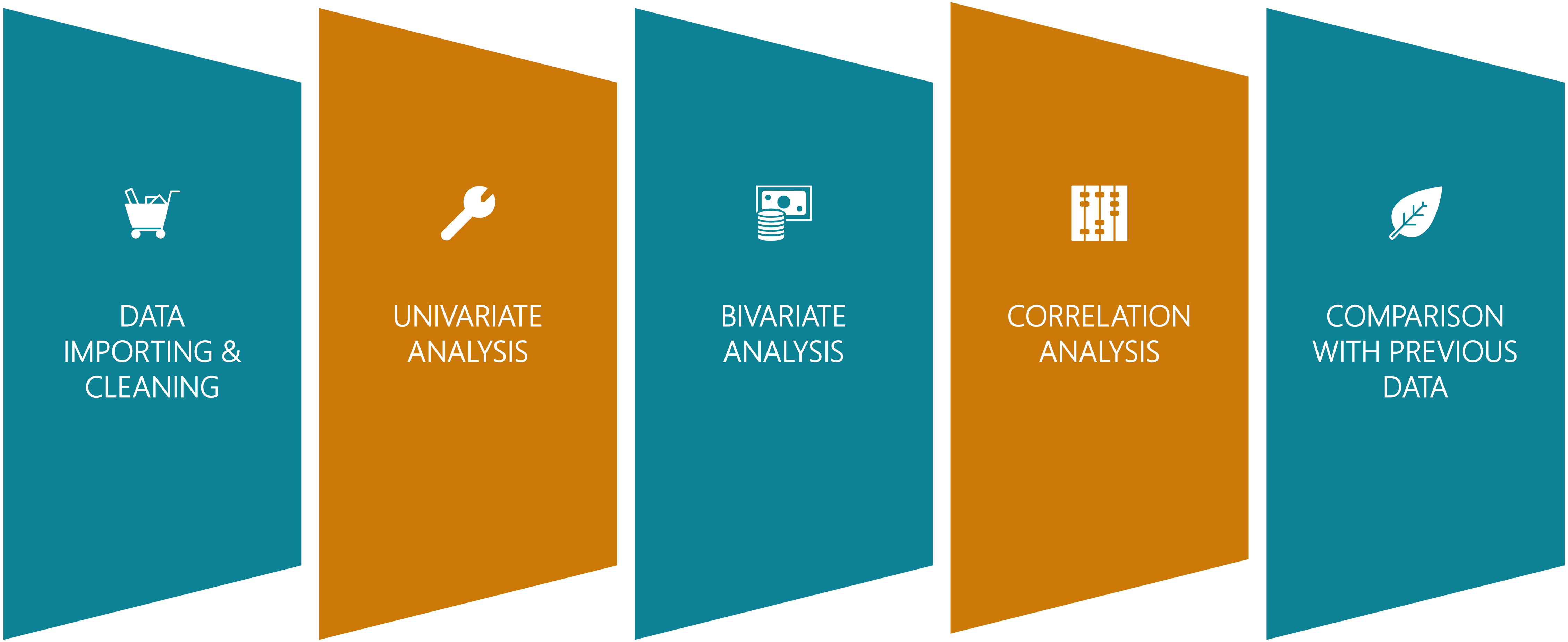
When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

- **Approved:** The Company has approved loan Application
- **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.
- **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
- **Unused offer:** Loan has been cancelled by the client but at different stages of the process.

Data Understanding

- This dataset has 3 files as explained below:
- 'application_data.csv' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
- 'previous_application.csv' contains information about the client’s previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.
- 'columns_description.csv' is data dictionary which describes the meaning of the variables.

Steps of Analysis



DATA IMPORTING & CLEANING

UNIVARIATE ANALYSIS

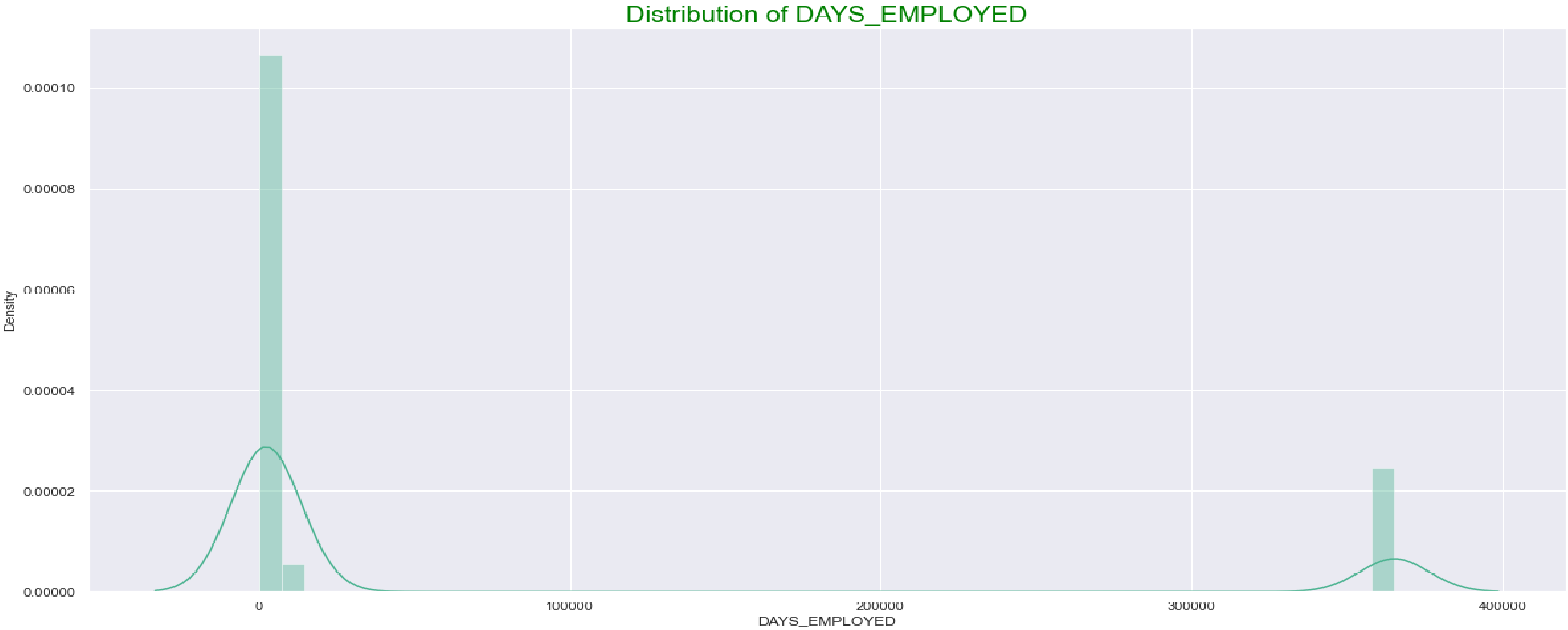
BIVARIATE ANALYSIS

CORRELATION ANALYSIS

COMPARISON WITH PREVIOUS DATA

DATA CLEANING

Outlier Checking

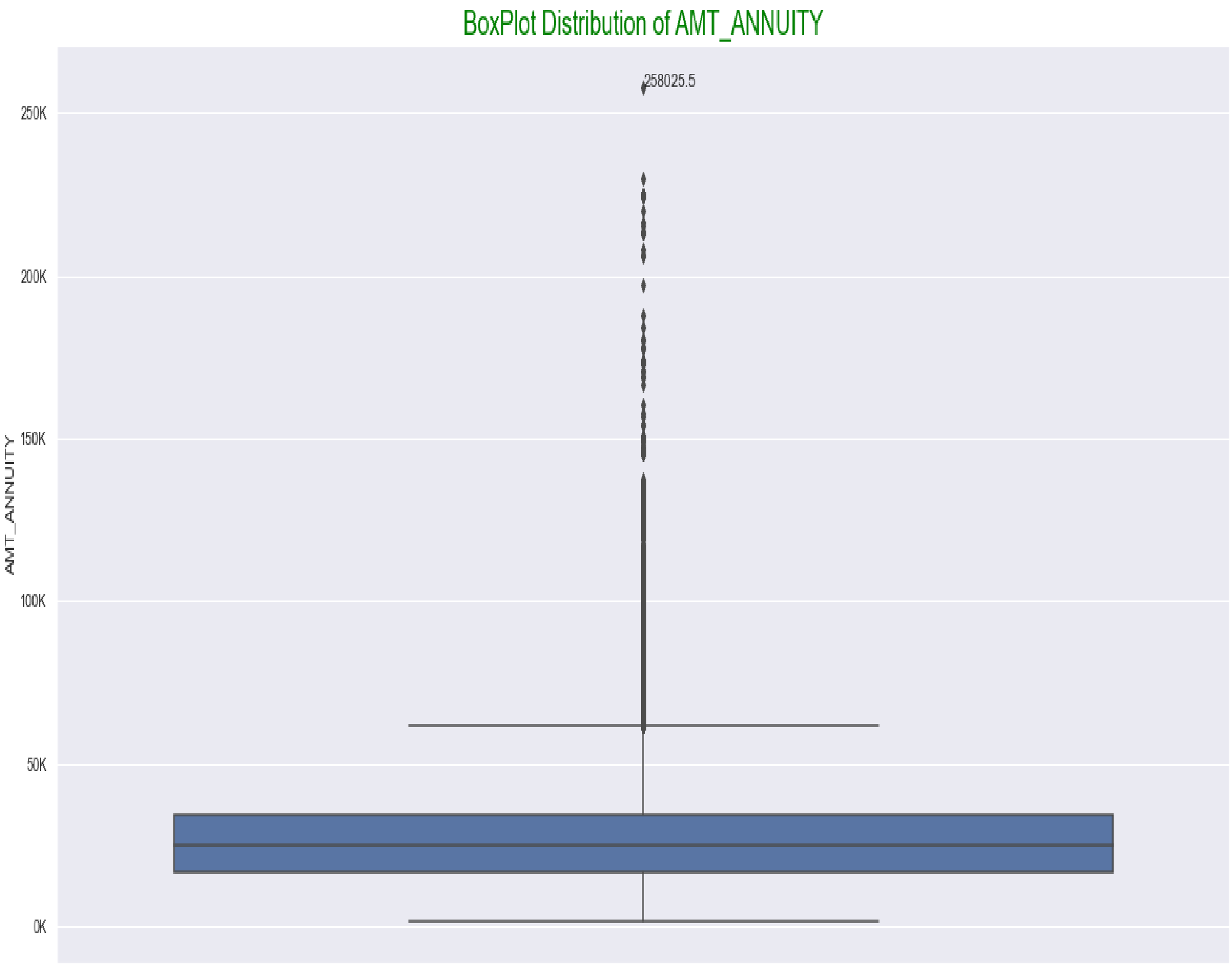


Inference:

- Tells us about the outliers present in DAYS_EMPLOYED column.
- Here we can take anything greater than 20,000 as an outlier since its more than 50 years. Considering a person will be over 70 years of age if that person starts working at 20 years old.

Outlier Checking

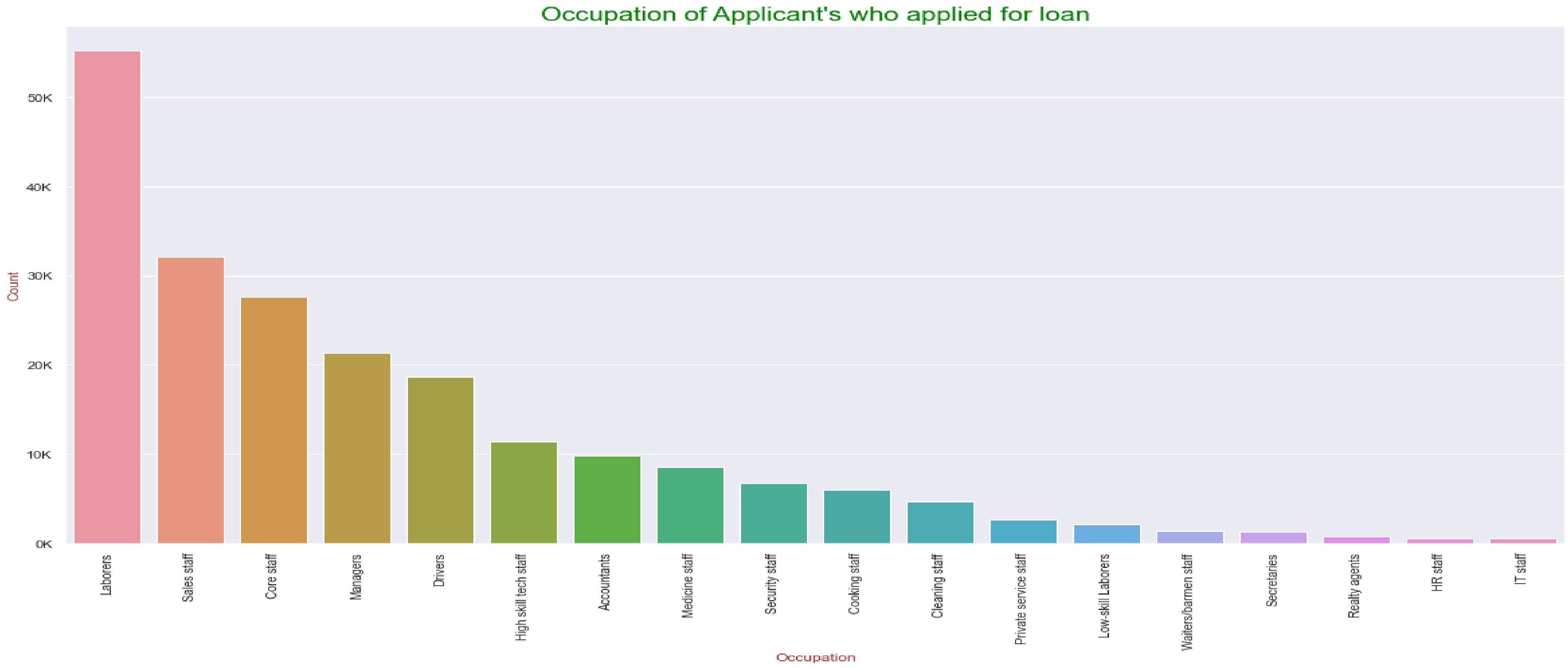
BoxPlot Distribution of AMT_ANNUITY



Inference:

- Tells us about the outliers present in AMT_ANNUITY column.
- Here we can take 258025.5 as an outlier.

Distribution Checking

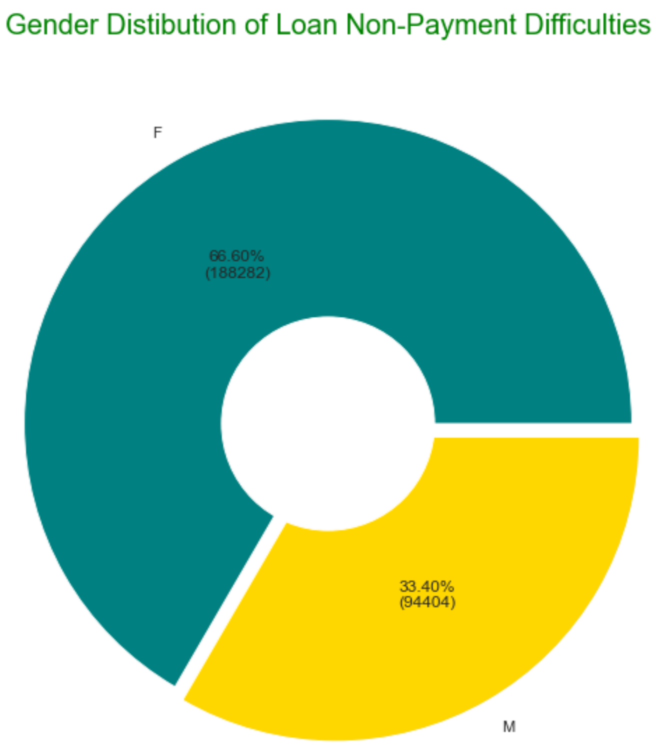


Inference:

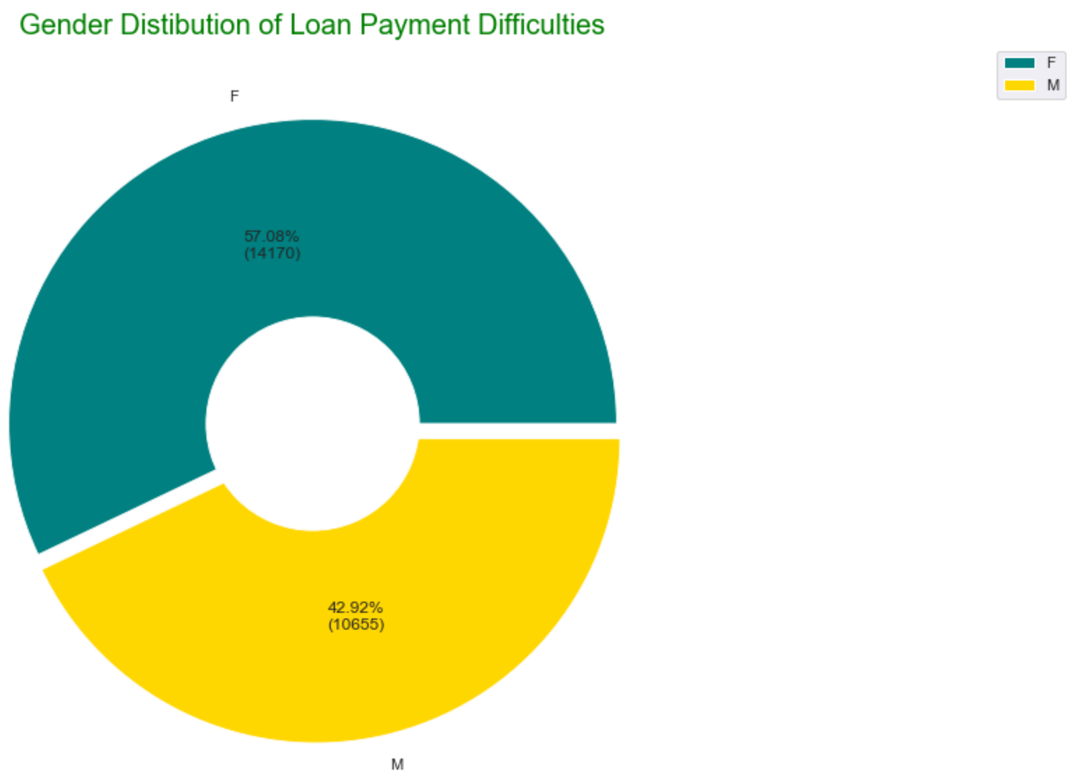
- Tells us about the distribution of occupation among the customers present in OCCUPATION_TYPE column.
- Laborers, Sales Staff and Core Staff constitutes the major occupation whereas IT Staff is on the minor side.

Univariate Analysis

Univariate Analysis



Gender Distibution of Loan Non-Payment Difficulties



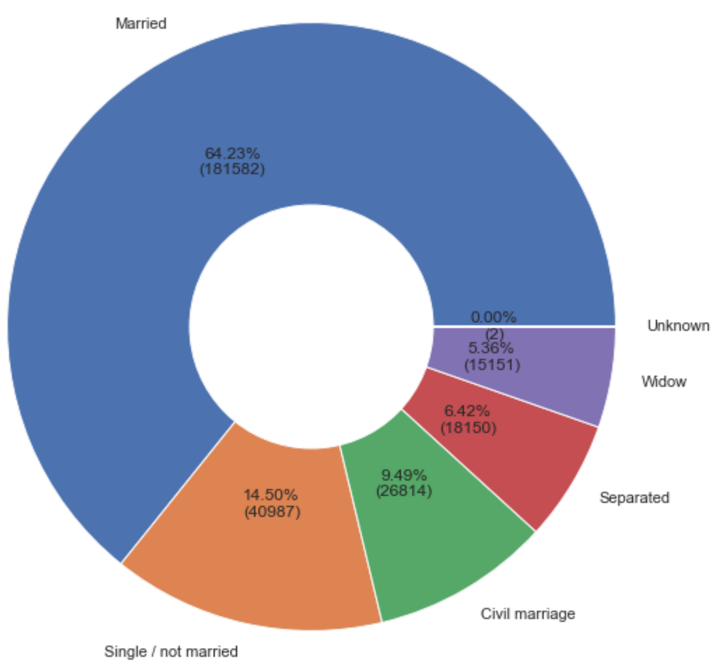
Gender Distibution of Loan Payment Difficulties

Inference:

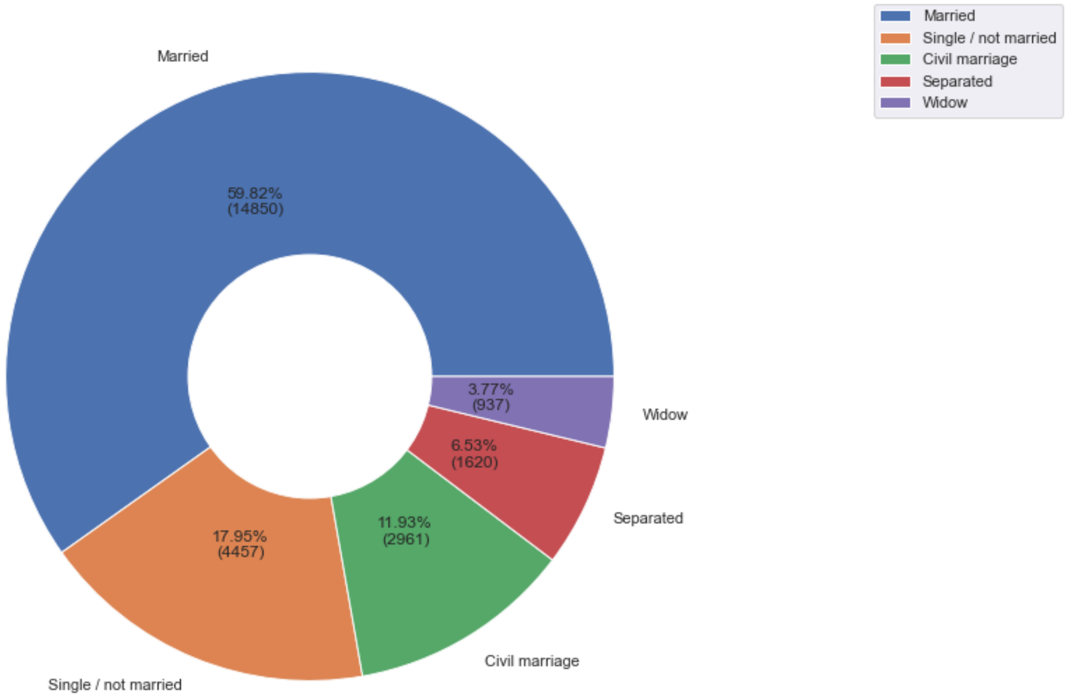
- On the basis of gender, Female group is majority in both Loan Payment Difficulties and Loan Non-Payment Difficulties.
- There is an increase in the percentage of Male group in Loan Payment Difficulties from Loan Non-Payment Difficulties.

Univariate Analysis

Family Status of Loan Non-Payment Difficulties



Family Status of Loan Payment Difficulties

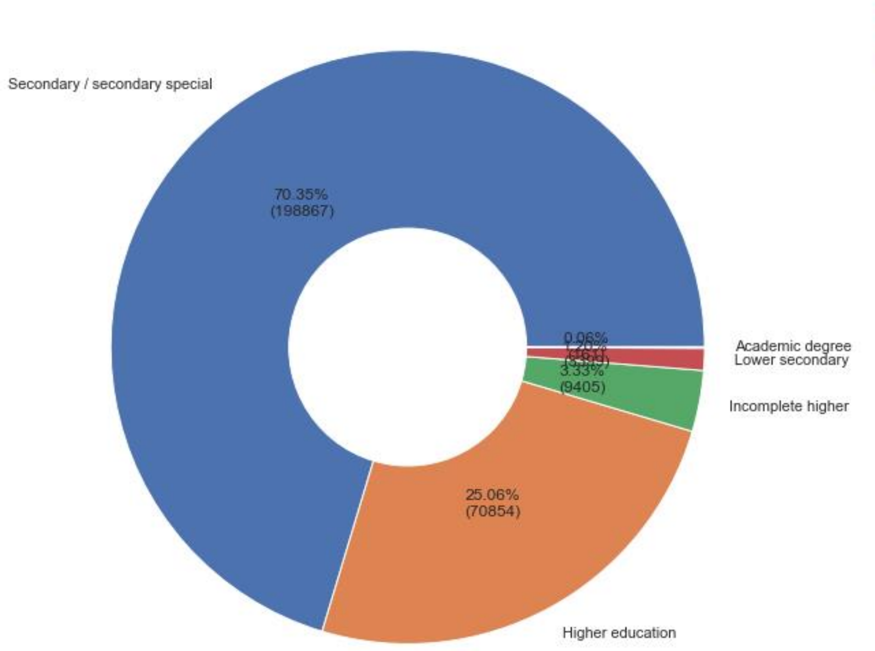


Inference:

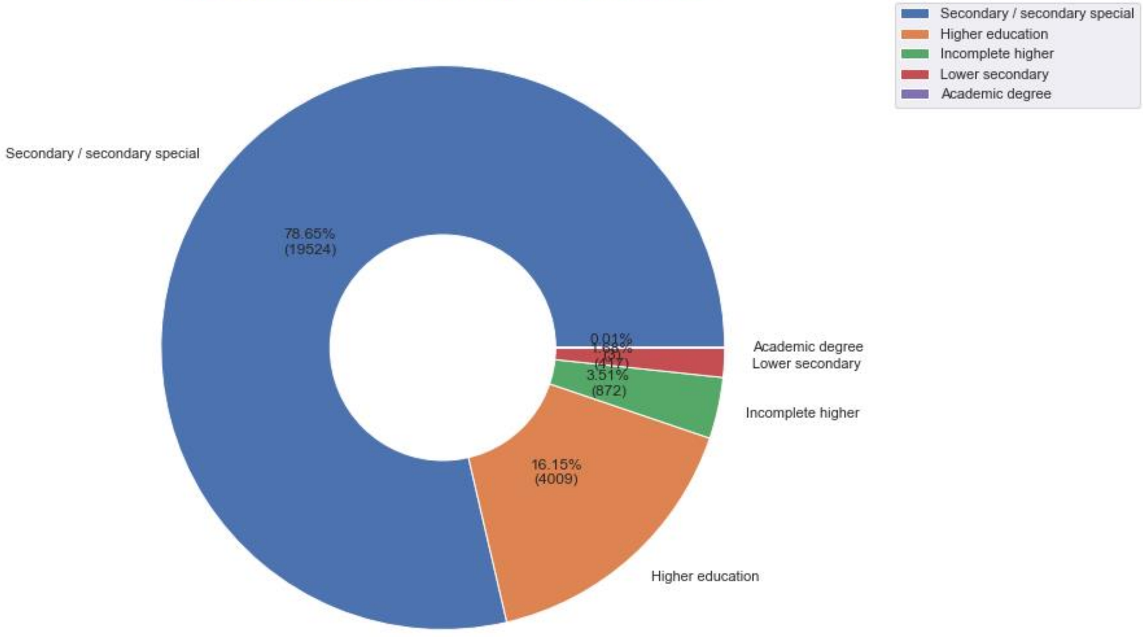
- There is a decrease in the percentage of Loan Payment Difficulties who are Married and Widowed.
- There is an increase in the percentage of Loan Payment Difficulties who are Single and Civil Married when compared the percentages of both Loan Payment Difficulties and Loan Non-Payment Difficulties.

Univariate Analysis

Education Status of Loan Non-Payment Difficulties



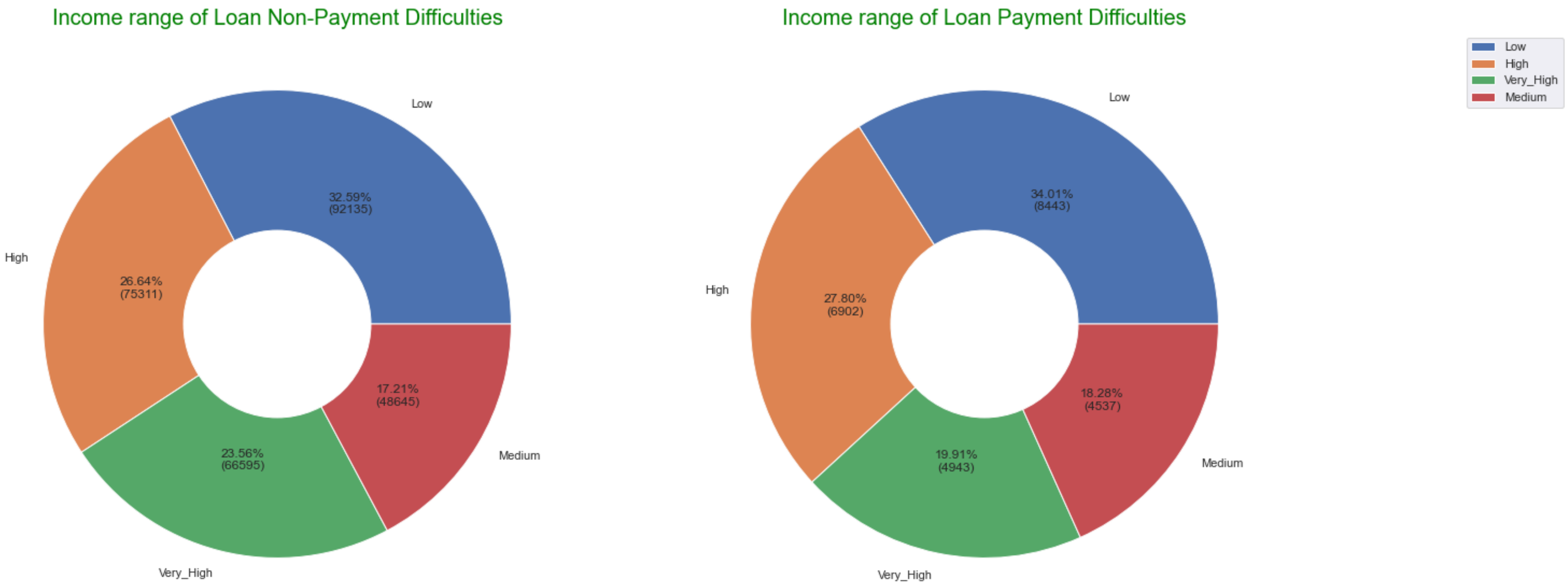
Education Status of Loan Payment Difficulties



Inference:

- There is a decrease in the percentage of Loan Payment Difficulties who have completed Secondary/Secondary Special.
- There is an increase in the percentage of Loan Payment Difficulties who have completed Higher Education when compared the percentages of both Loan Payment Difficulties and Loan Non-Payment Difficulties.

Univariate Analysis



Income range of Loan Non-Payment Difficulties

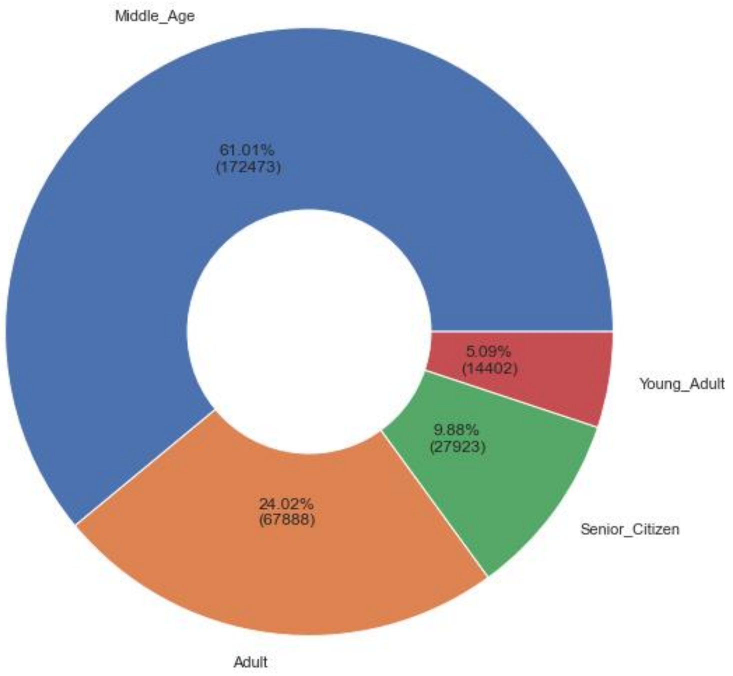
Income range of Loan Payment Difficulties

Inference:

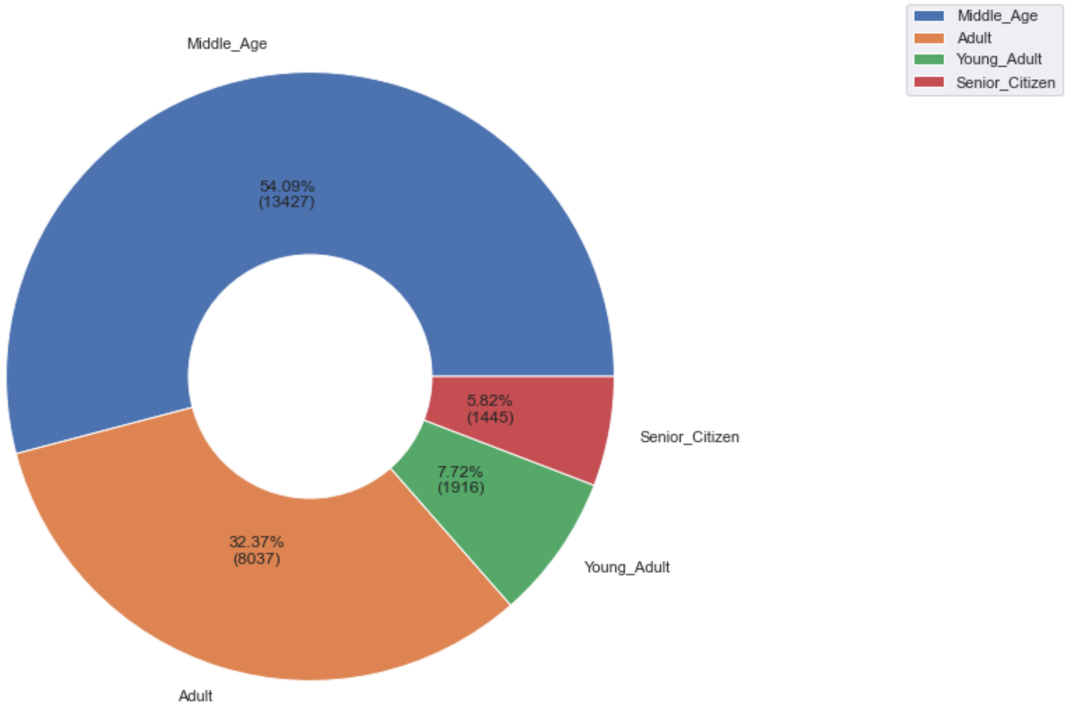
- There is an increase in the percentage of Loan Payment Difficulties whose Income is Low when compared the percentages of both Loan Payment Difficulties and Loan Non-Payment Difficulties.

Univariate Analysis

Age of Loan Non-Payment Difficulties



Age of Loan Payment Difficulties

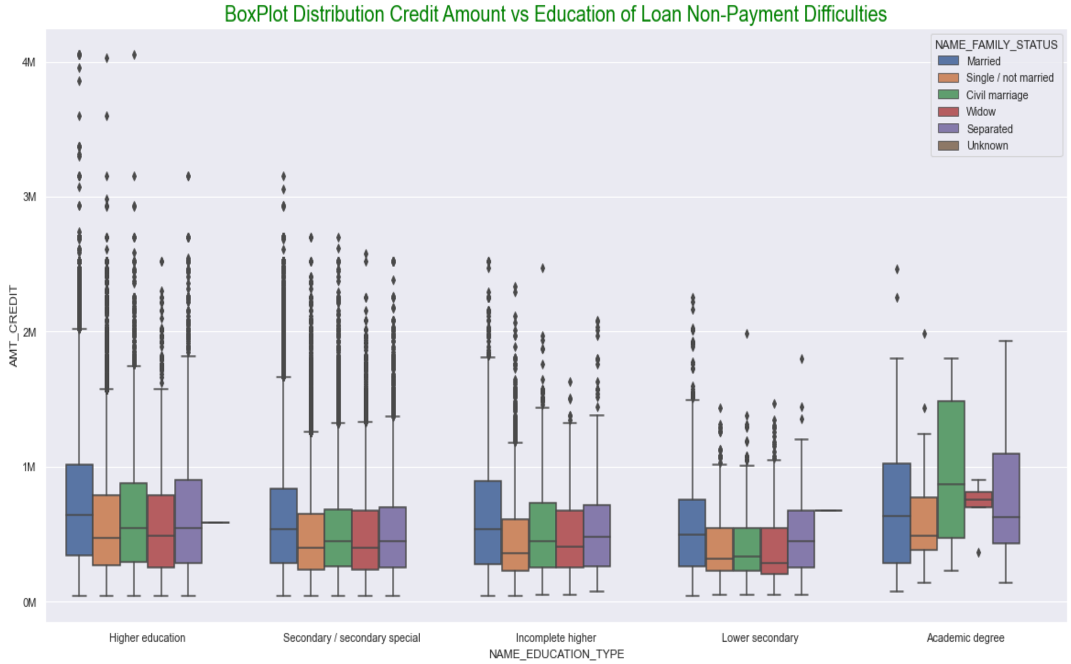


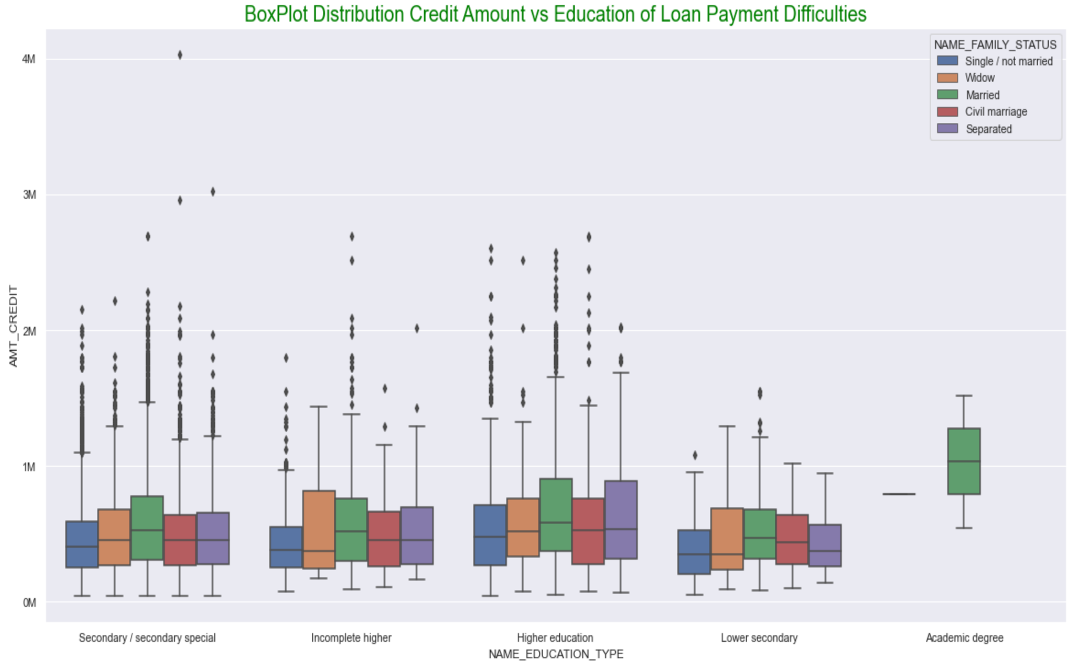
Inference:

- 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.

Bivariate Analysis

Bivariate Analysis

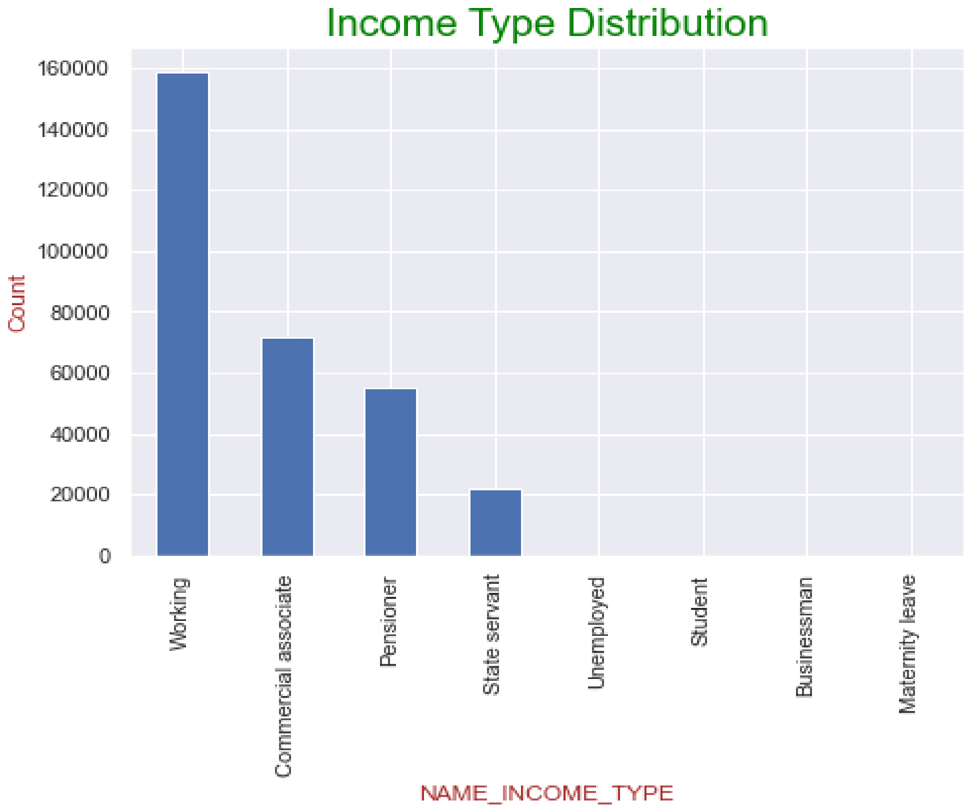


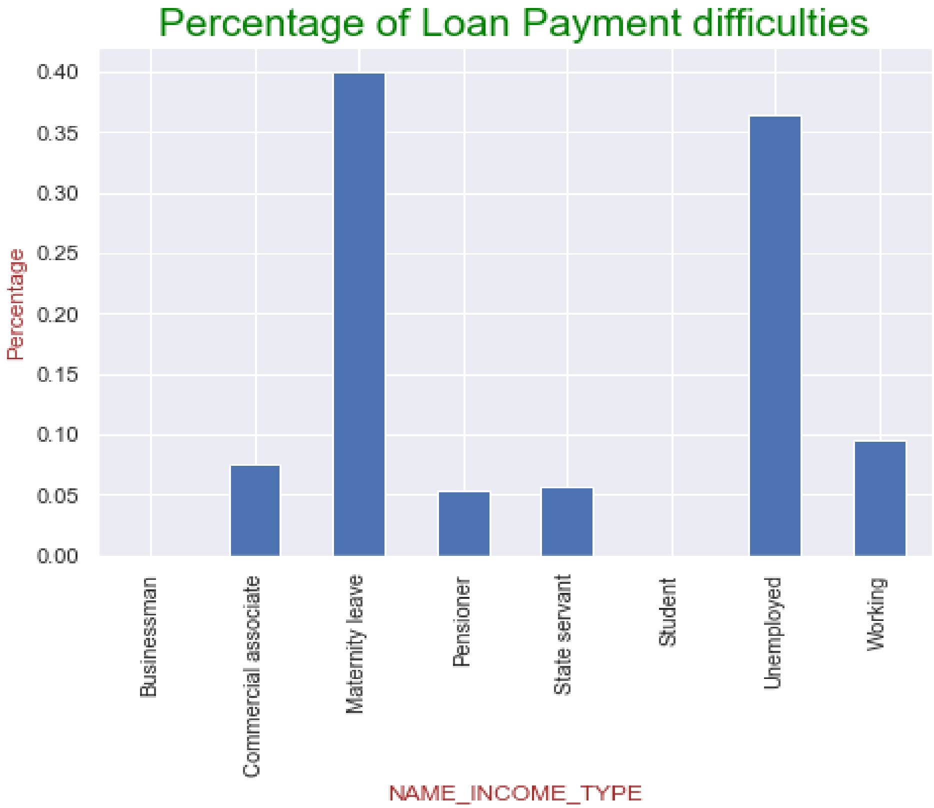


Inference:

- The graphs for Loan Payment Difficulties and Loan Non-Payment Difficulties appears to be similar.
- 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

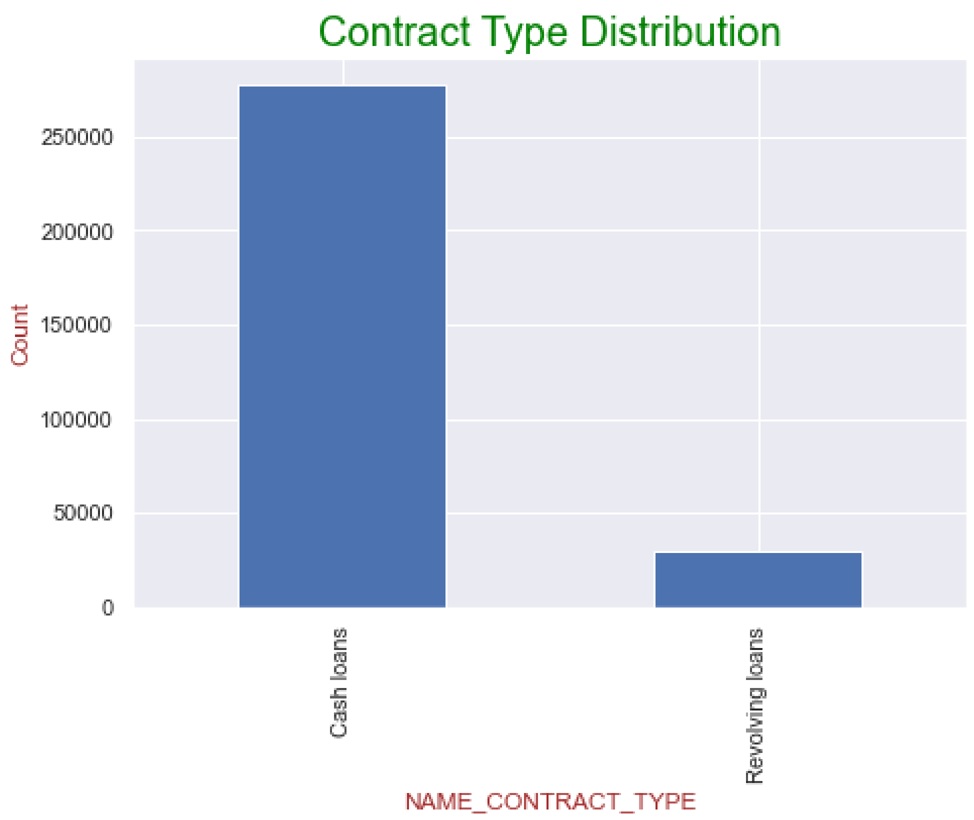


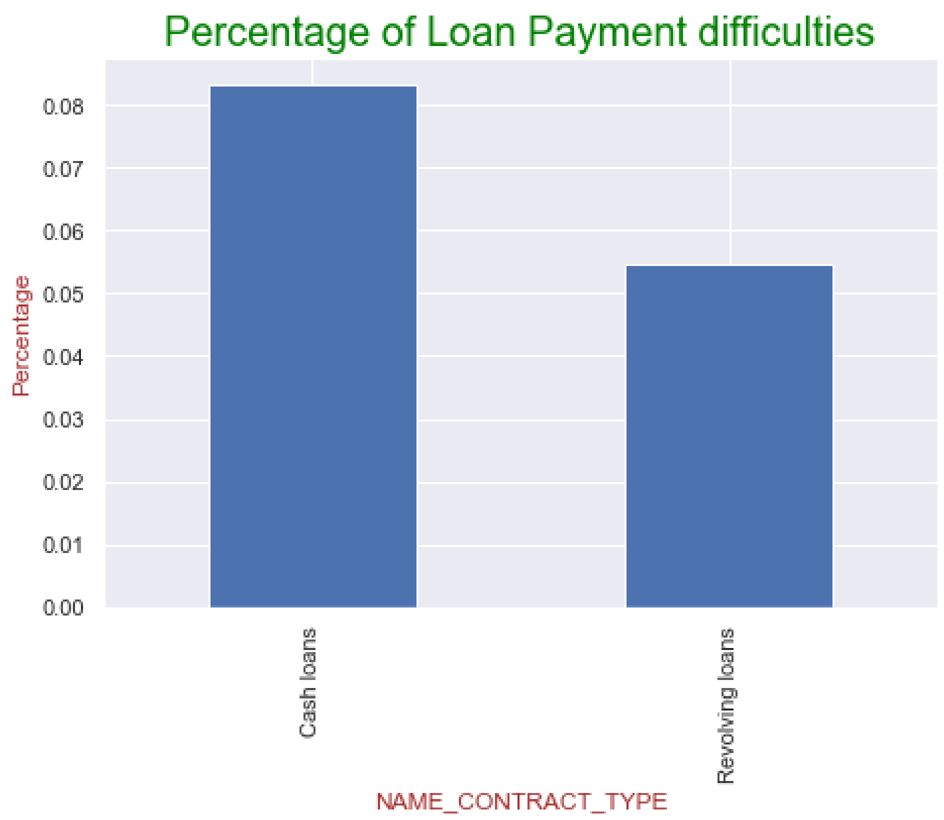


Inference:

- **The clients who are on Maternity Leave and Unemployed have the highest percentage of Loan Payment Difficulties.**

Bivariate Analysis

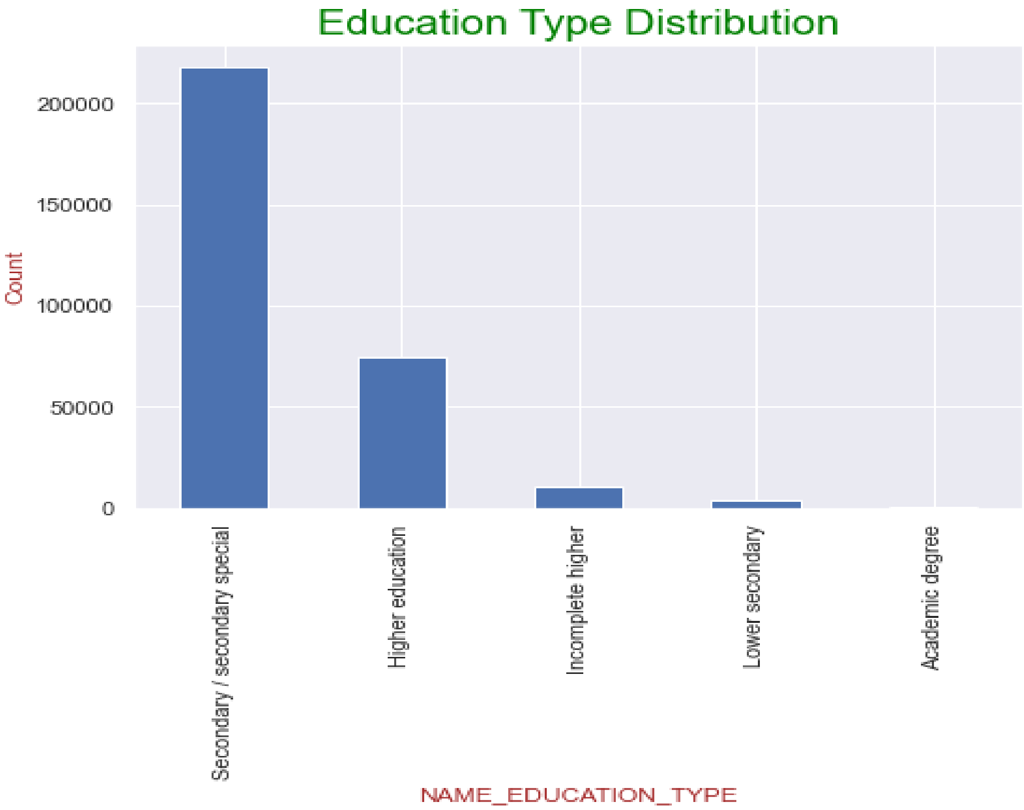


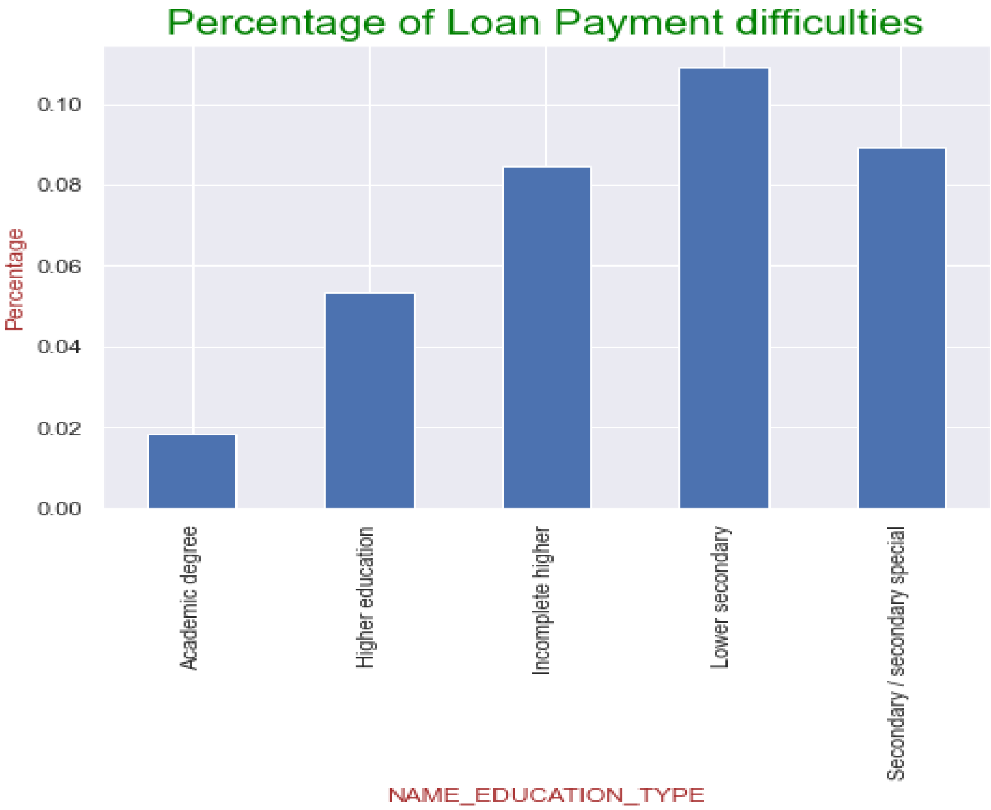


Inference:

- The clients who want Cash Loans have the highest percentage of Loan Payment Difficulties.

Bivariate Analysis

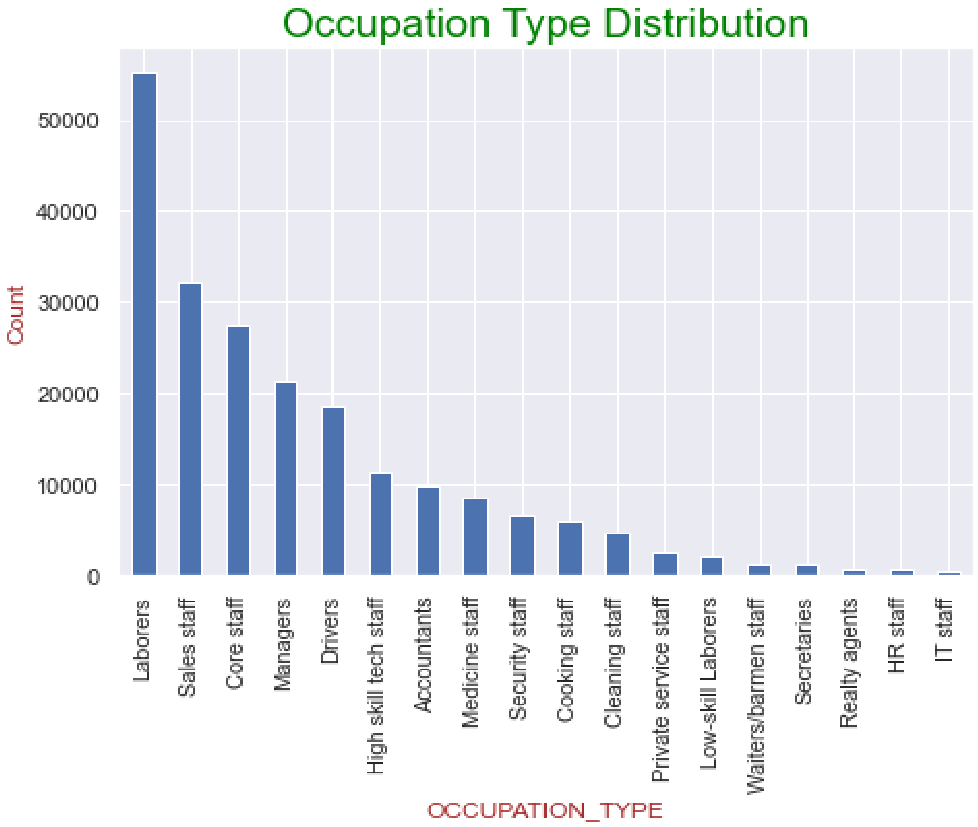


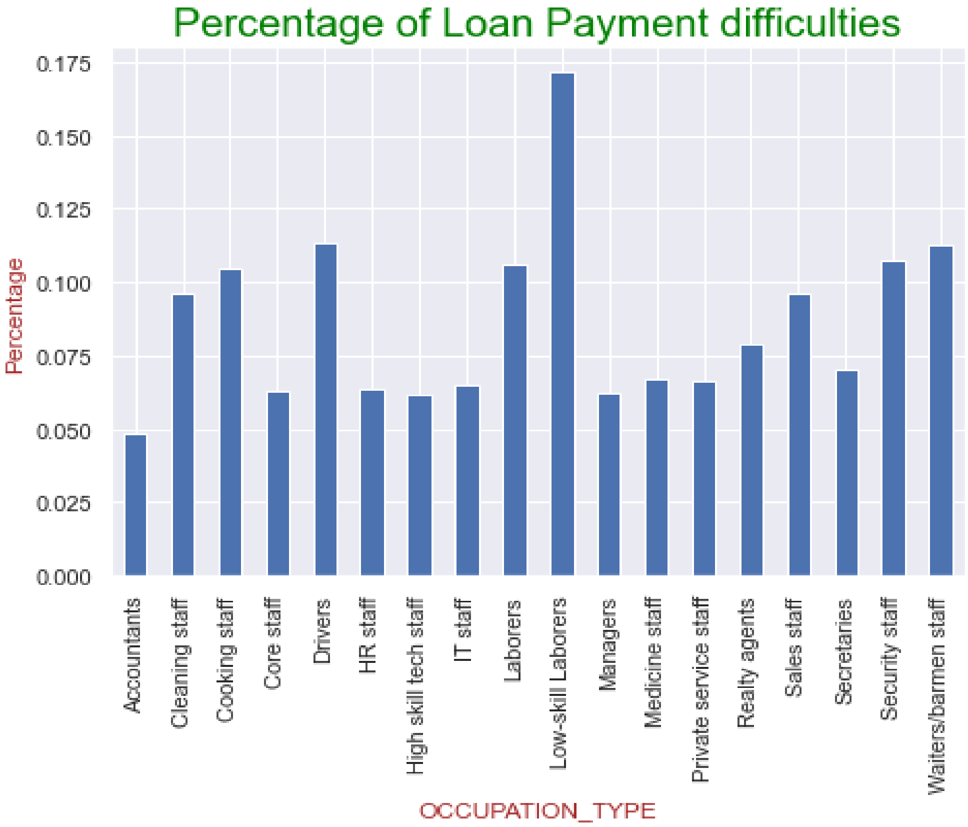


Inference:

- **The clients with Lower Secondary and Secondary/Secondary Special Education Type have the highest percentage of Loan Payment Difficulties.**

Bivariate Analysis

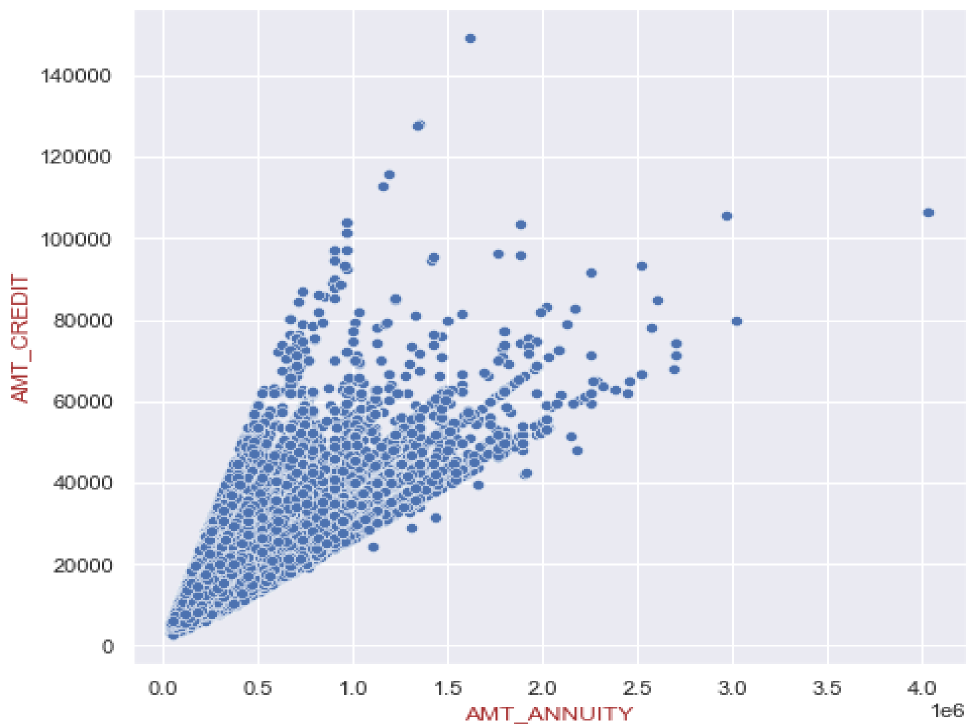




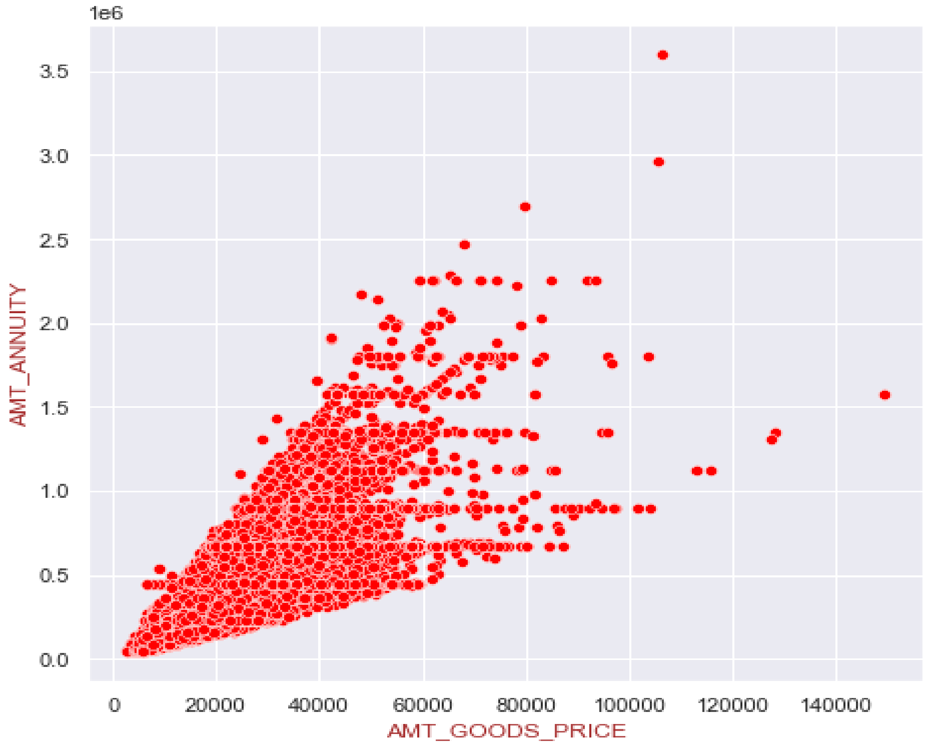
Inference:

- **The clients who are Low-Skill Laborers have the highest percentage of Loan Payment Difficulties.**

Bivariate Analysis



AMT_CREDIT vs AMT_ANNUITY



AMT_ANNUITY vs AMT_GOODS_PRICE

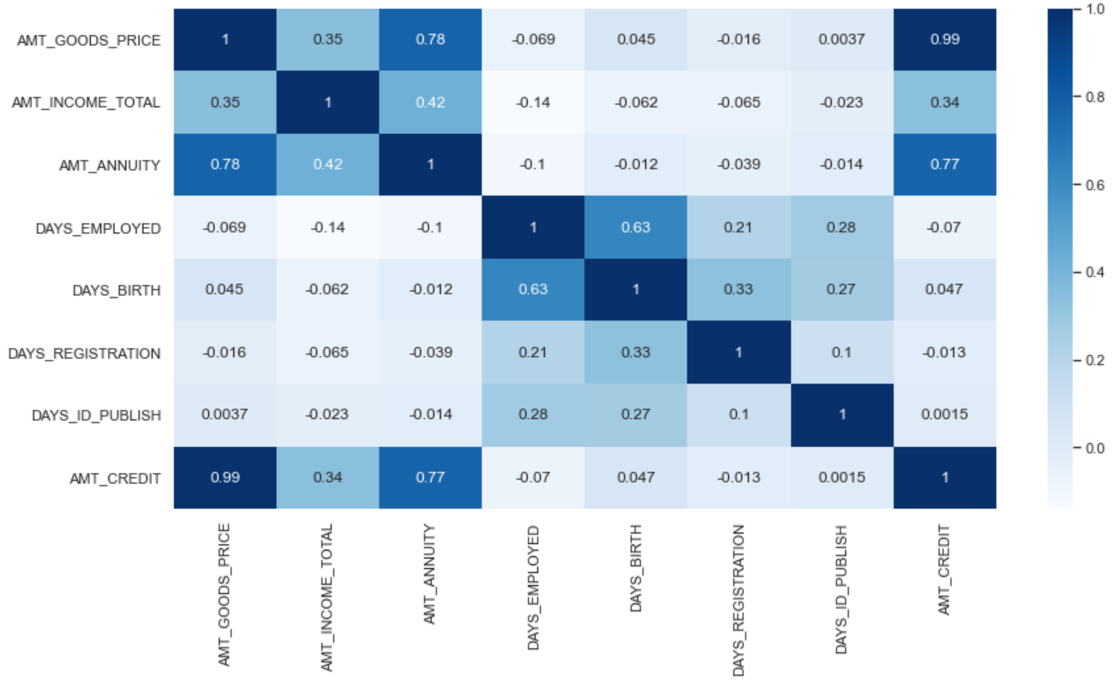
Inference:

- **AMT_CREDIT VS AMT_ANNUITY and AMT_ANNUITY VS AMT_GOODS_PRICE representing Bivariate Analysis for Target Variable.**

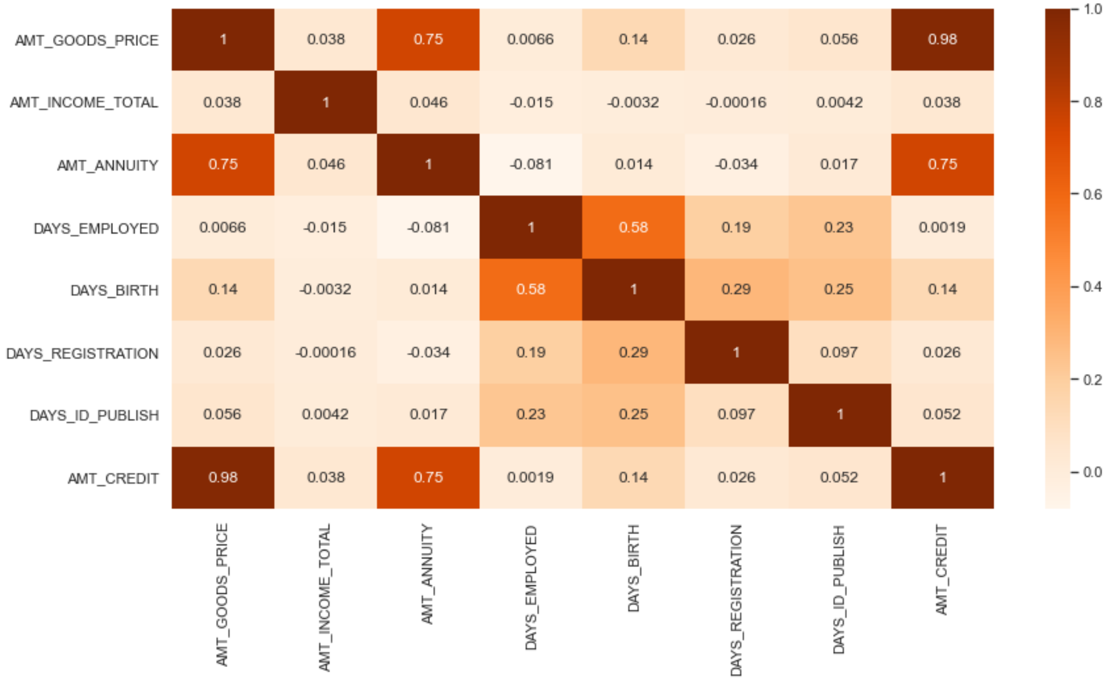
Correlation Analysis

Correlation Analysis

Correlation Heatmap of Loan Non-Payment Difficulties



Correlation Heatmap of Loan Payment Difficulties

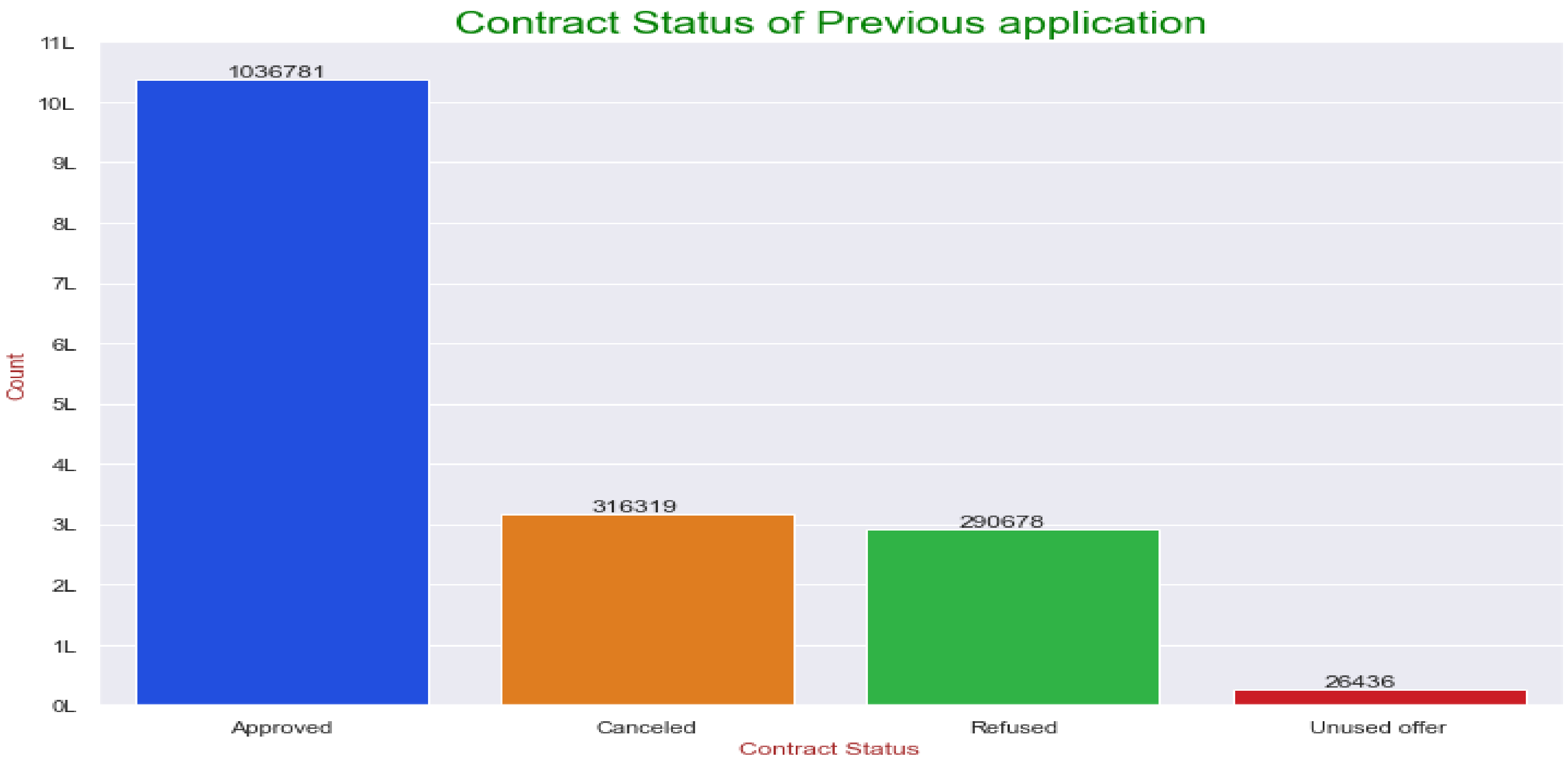


Inference:

- **High correlation between Credit Amount and Goods Price.**
- **Some deviance appearing in the correlation of Loan Payment Difficulties and Loan Non-Payment Difficulties such as Credit Amount vs Income.**

COMPARISON WITH PREVIOUS DATA

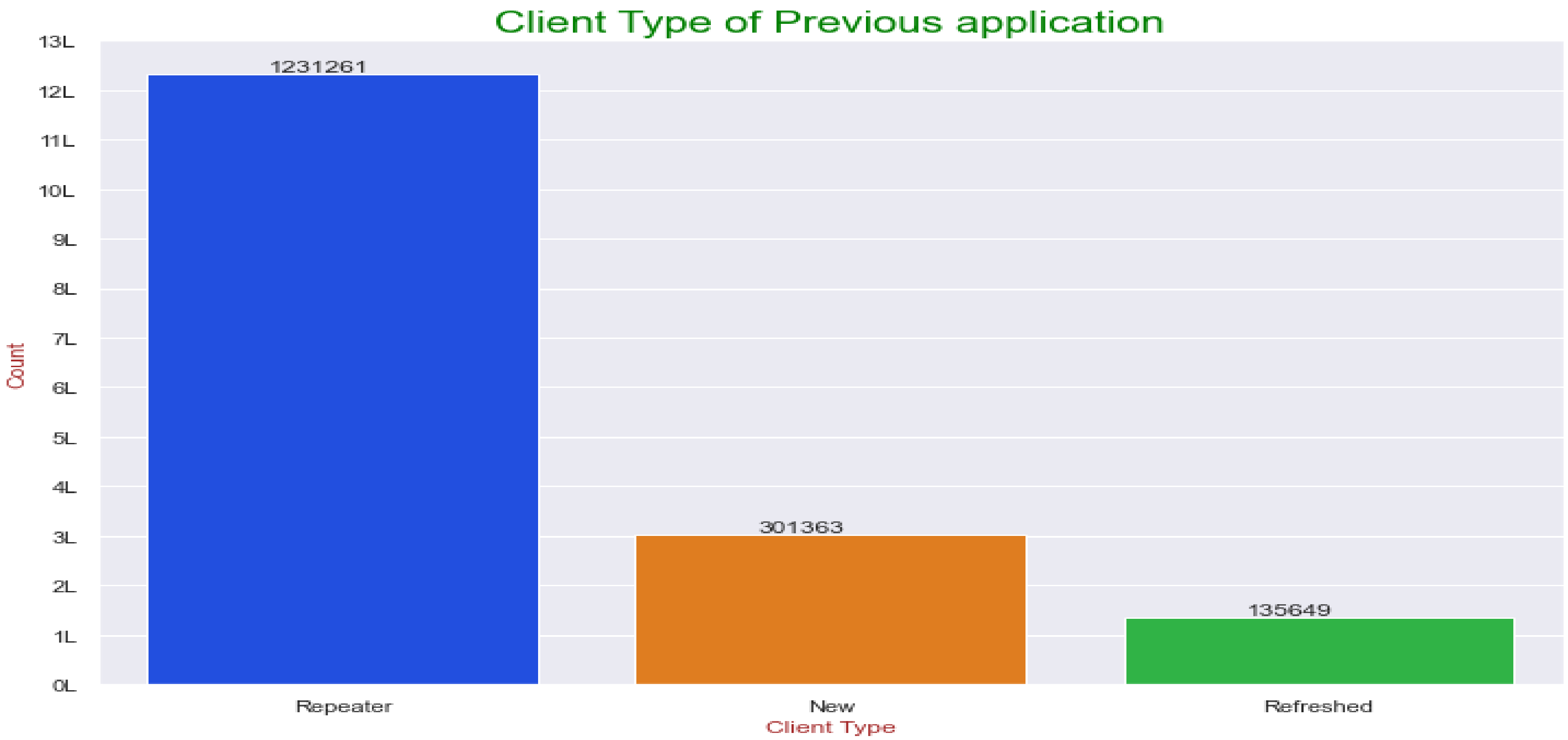
COMPARISON WITH PREVIOUS DATA



Inference:

- Majority of the loans were Approved
- Very less percentage of loans are Unused Offer

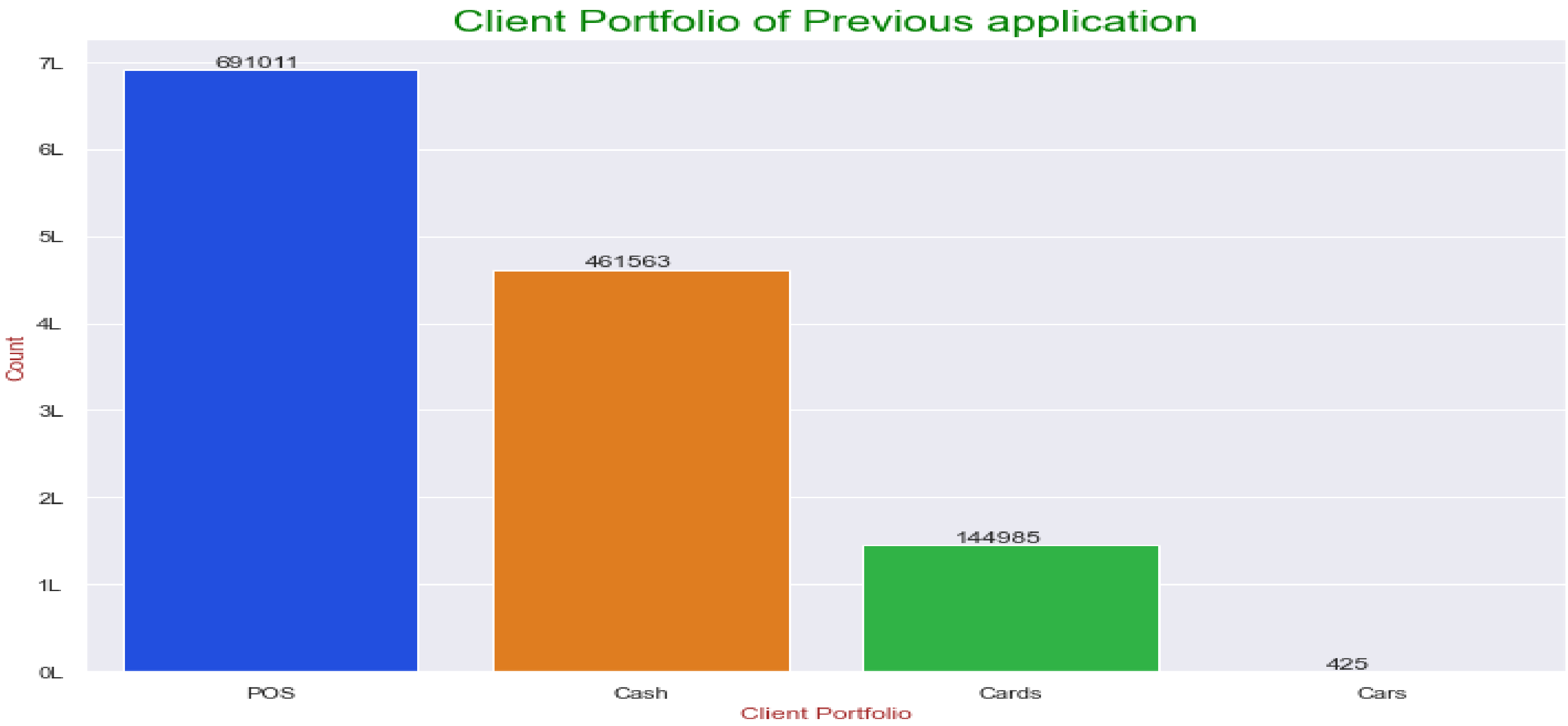
COMPARISON WITH PREVIOUS DATA



Inference:

- Majority of the loans were taken by repeated clients.

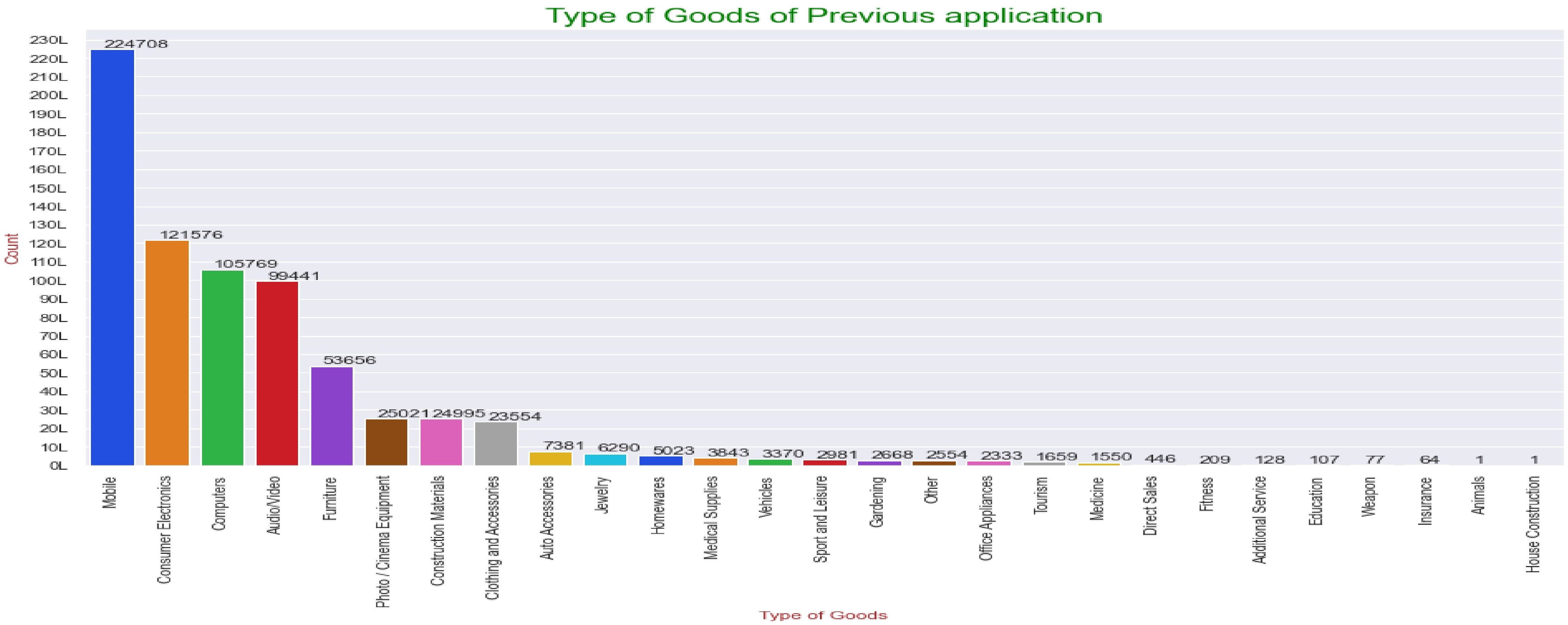
COMPARISON WITH PREVIOUS DATA



Inference:

- Majority of the Portfolio was POS and good amount of Cash.

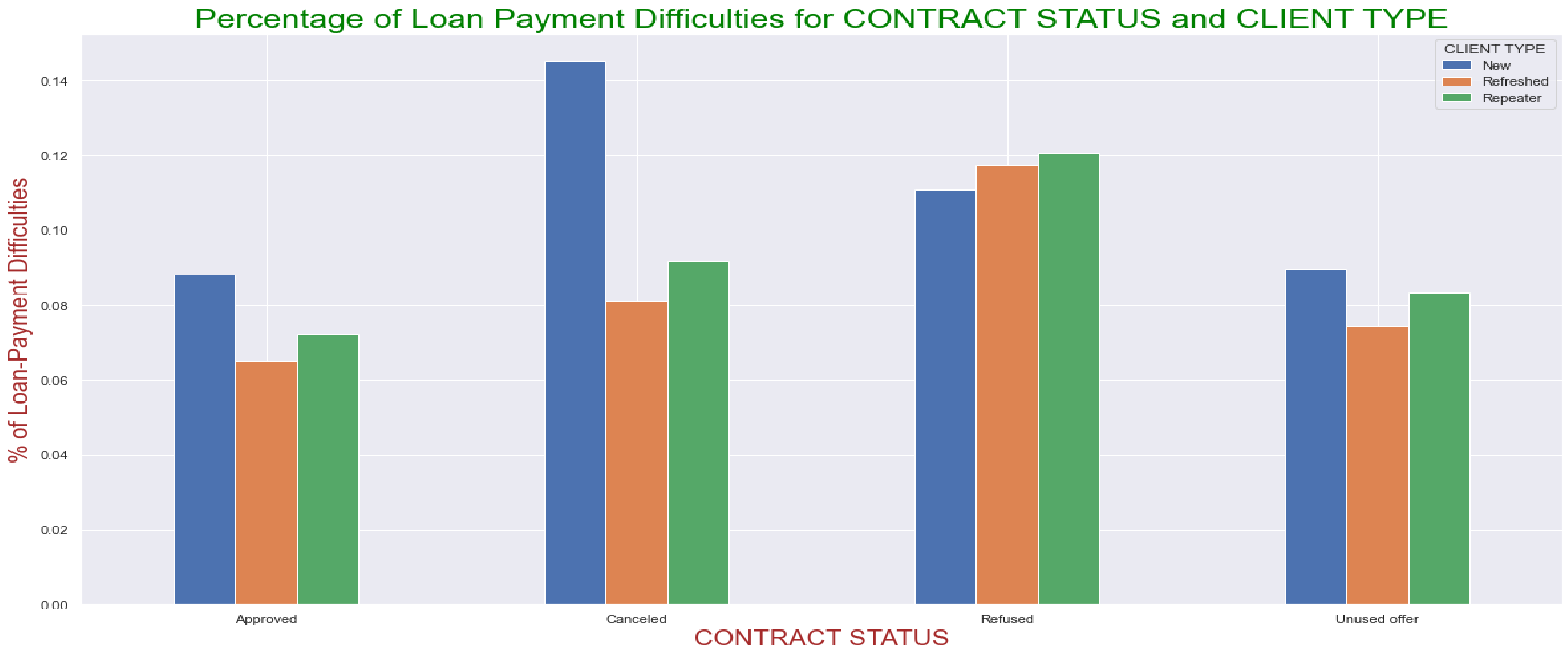
COMPARISON WITH PREVIOUS DATA



Inference:

- Majority of the loans were taken for mobiles, consumer electronics, computers and furniture.

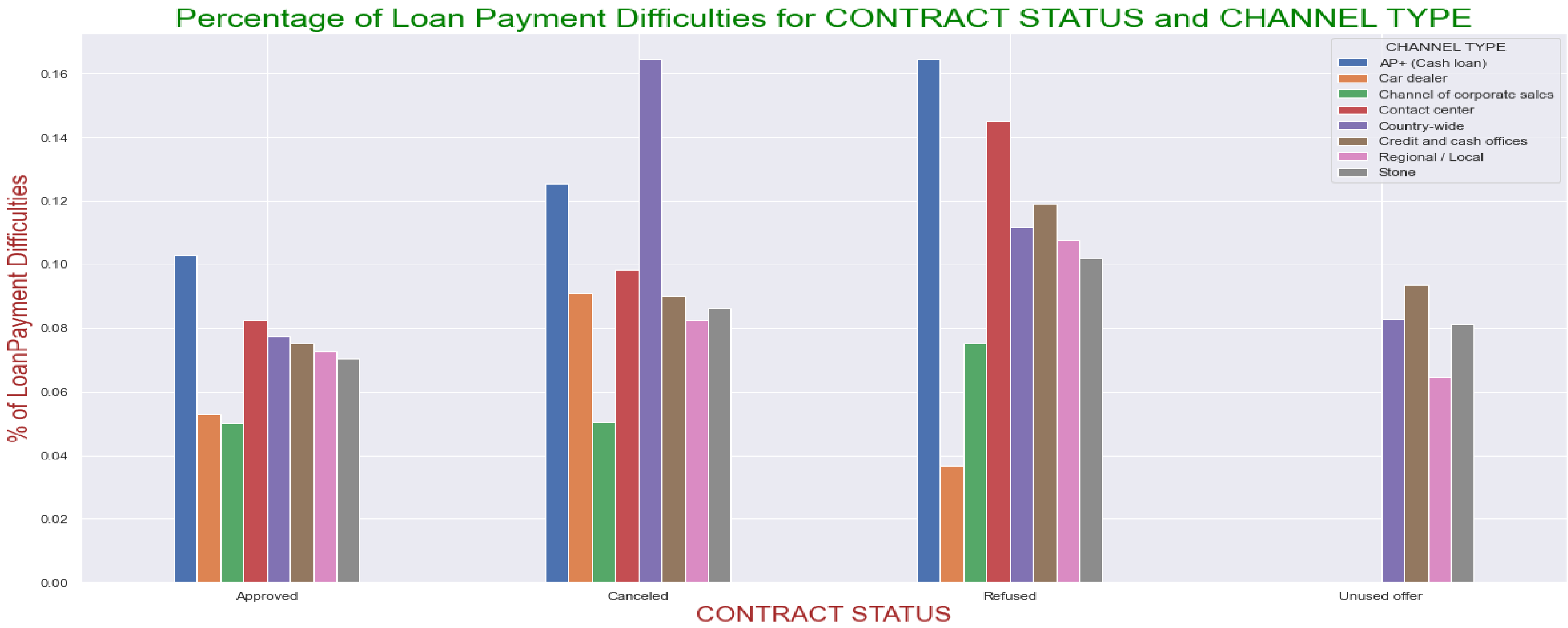
COMPARISON WITH PREVIOUS DATA



Inference:

- Clients who were 'New' and had their previous application 'Cancelled' tend to have higher percentage of Loan Payment Difficulties in current application.

COMPARISON WITH PREVIOUS DATA



Inference:

- Clients whose Channel Type was AP+(Cash Loan) and Country-wide had their Loan cancelled more than anyone else.
- Clients whose Channel Type was AP+(Cash Loan) and Contact center got their Loan refused more than anyone else.

CONCLUSION

Conclusion

- Female clients on maternity leave should NOT be targeted as they have no record of repayments. Hence, client with income type as ‘Maternity leave’ are the driving factors for Loan Defaulters.
- Clients who want Cash Loans have the highest percentage of Loan Payment Difficulties.
- The count of ‘Lower Secondary’ is comparatively very less and it also has maximum % of payment difficulties. Hence, client with education type as ‘Lower Secondary’ are the driving factors for Loan Defaulters.
- The count of ‘Low Skilled Laborers’ is comparatively very less and it also has maximum % of payment difficulties. Hence, client with occupation type as ‘Low skilled Laborers’ are the driving factors for Loan Defaulters.
- Clients who are working and from Low Income group need to be targeted LESS by the bank as they have the highest amount of defaulters.
- Banks SHOULD also target female clients as they have the highest repayment (almost as double as males) amongst both the genders.
- Banks SHOULD target clients who own a car as they more likely to repay.
- Majority of the loans were taken for mobiles, consumer electronics, computers and furniture. So the bank should provide better interest in these kind of loans to attract more clients.
- The bank should provide better interest rates and offers for the repeat clients as they’re more likely to repay.
- Application and verification process for the new clients should be made more seamless so that their applications don’t get rejected.

Thank You