PRESENTATION OF EDA CASE STUDY

TEAM DETAILS:
Anishma Kathpal
anishmakathpal1503@gmail.com
Harshitha C
harshithackashyap@gmail.com

TABLE OF CONTENTS

INTRODUCTION

This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

BUSINESS UNDERSTANDING

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

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.

BUSINESS UNDERSTANDING (Continued)

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 on different stages of the process.

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

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.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

DATA UNDERSTANDING

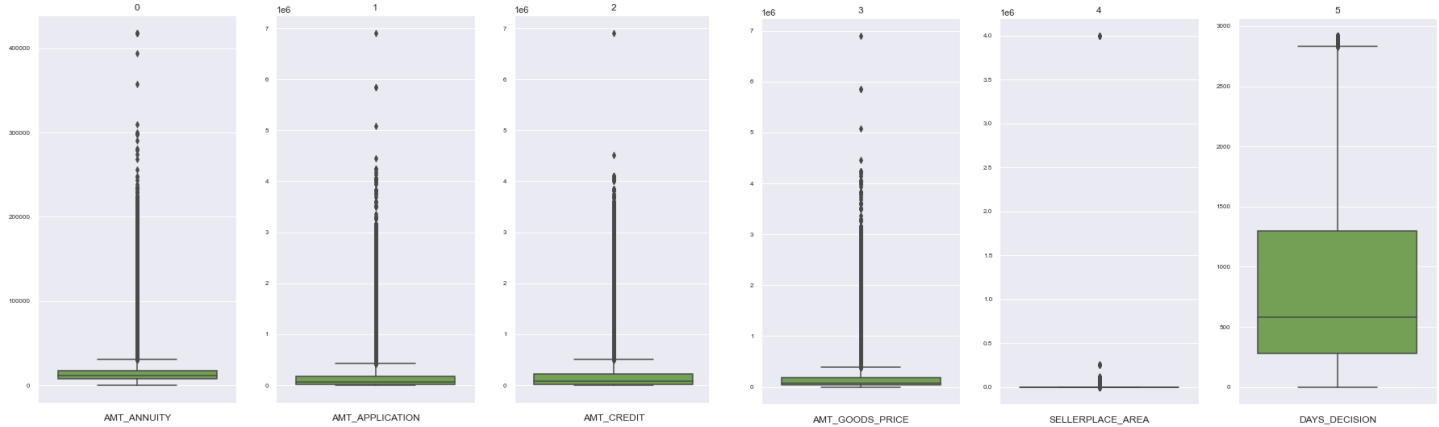
This dataset has 3 files as explained below:

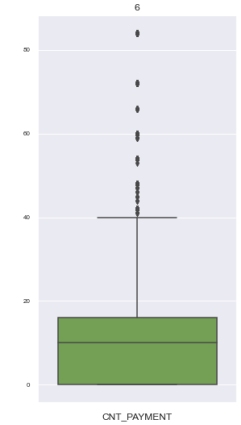
1. *'application_data.csv'* contains all the information of the client at the time of application. The data is about whether a **client has payment difficulties**.

2. *'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.**

3. *'columns_description.csv'* is data dictionary which describes the meaning of the variables.

OUTLIERS



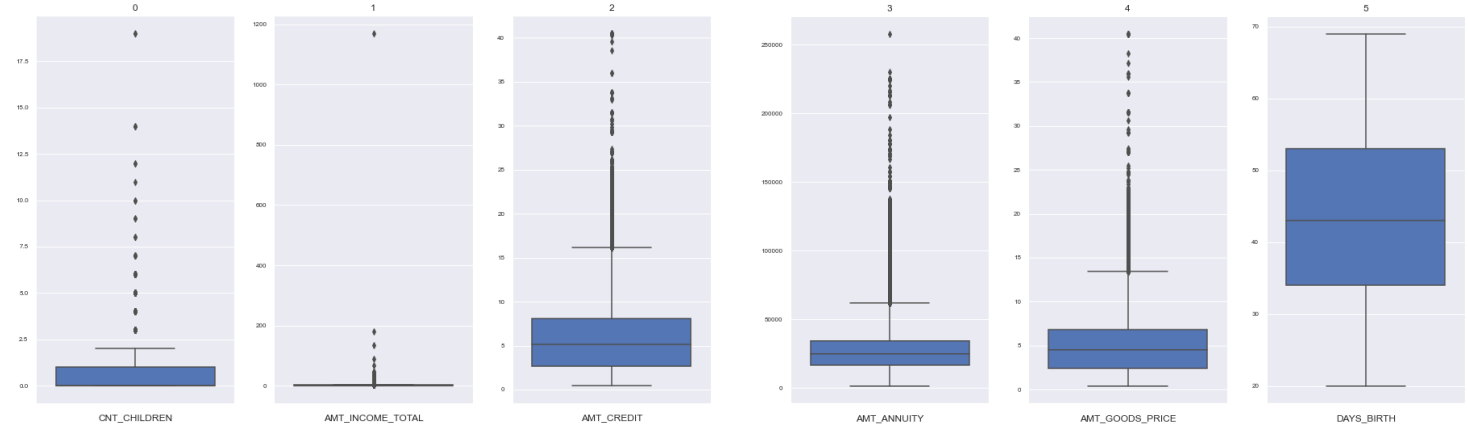


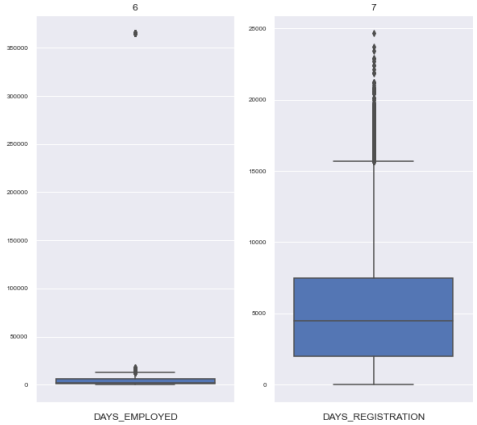
Percentage of Data Loss(outliers) in Predata Dataframe:
37.87634398945285 %

It can be seen that in previous application data

- AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.
- CNT_PAYMENT has few outlier values.
- DAYS_DECISION has little number of outliers indicating that these previous applications decisions were taken long back.

OUTLIERS



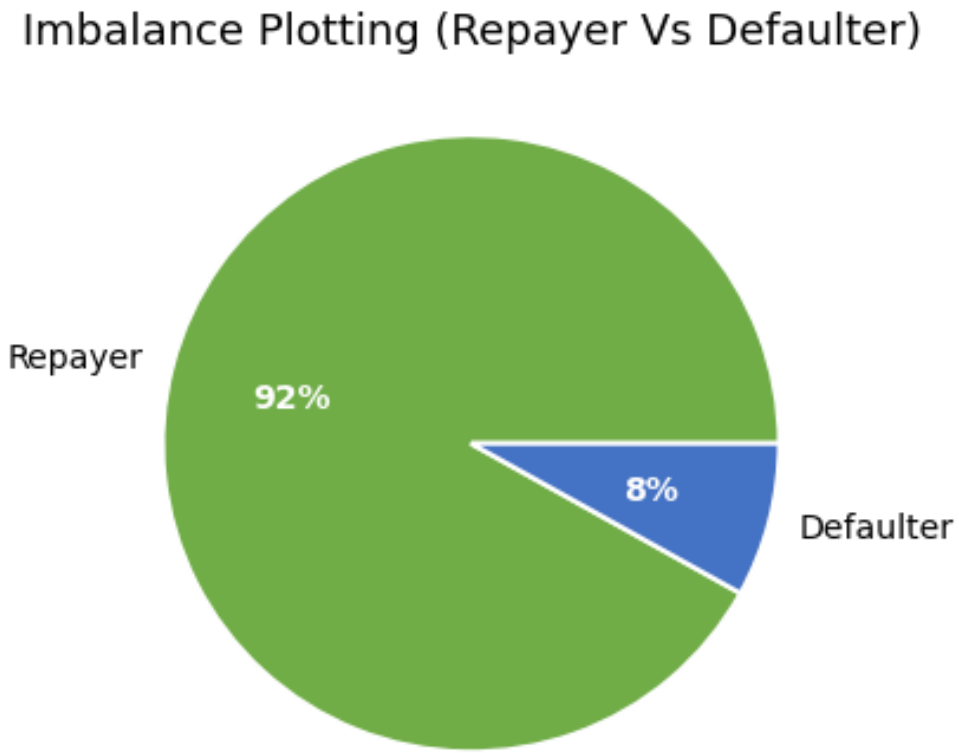


Percentage of Data Loss(outliers) in Appdata Dataframe:
27.211059116584448 %

It can be seen that in current application data

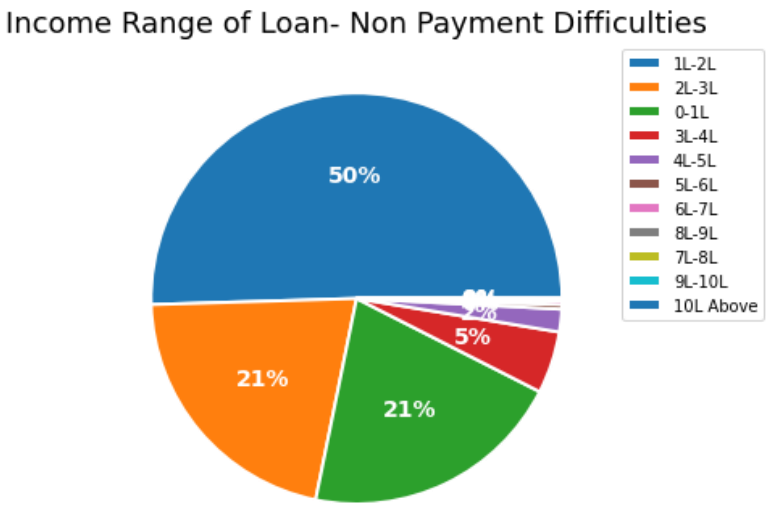
- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.
- AMT_INCOME_TOTAL has huge number of outliers which indicate that few of the loan applicants have high income when compared to the others.
- DAYS_BIRTH has no outliers which means the data available is reliable.
- DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this has to be incorrect entry.

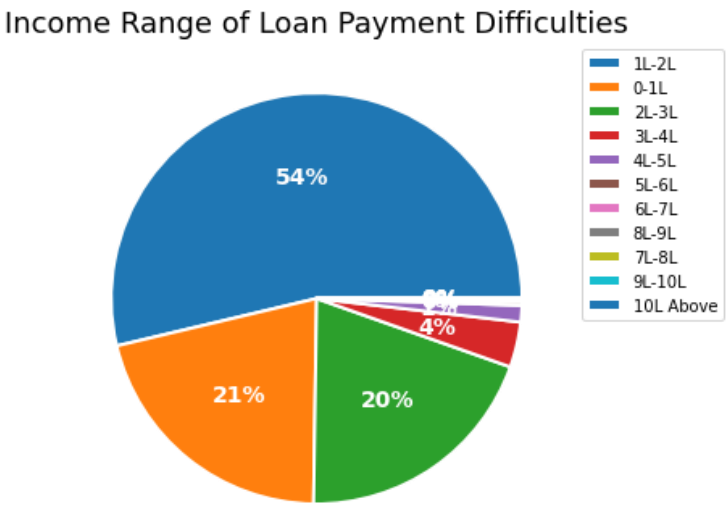
IMBALANCE DATA



Imbalance Ratio with respect to Repayer and Defaulter is given: **11.39:1 (approx.)**

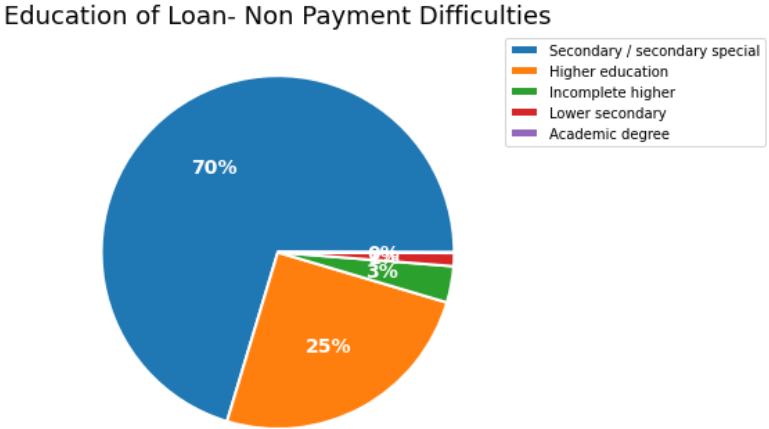
UNIVARIATE ANALYSIS

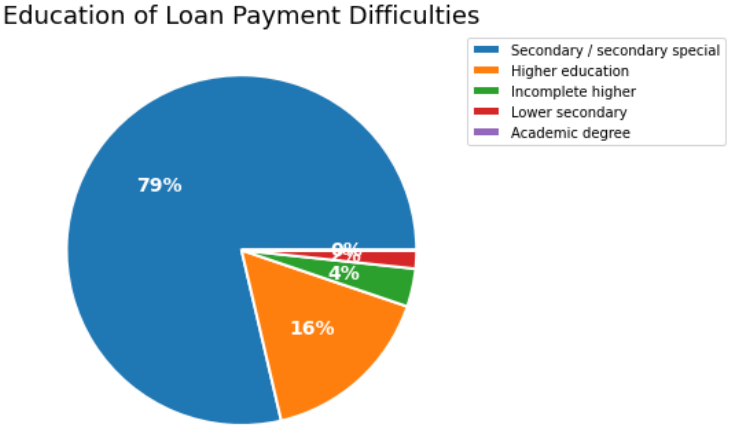




92% Loan applicants for ***non payment difficulties*** have income range below 3 lakh – with 50% in the range of 1L – 2L
95% Loan applicants for ***payment difficulties*** have income range below 3 lakh - with 54% in the range of 1L – 2L

UNIVARIATE ANALYSIS

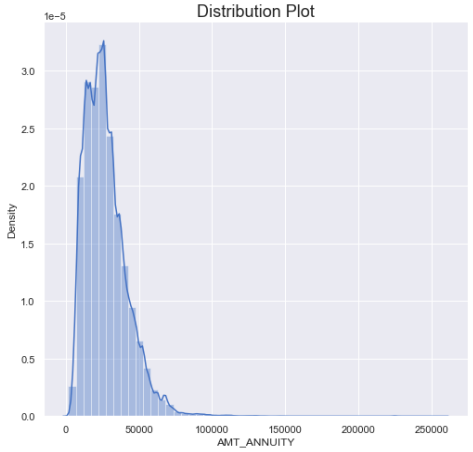


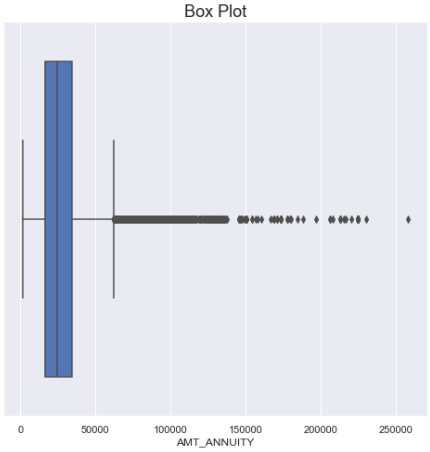


25% Loan applicants for ***non payment difficulties*** have pursued Higher education (**9% more education** as compared to ***loan payment difficulty*** applicants)

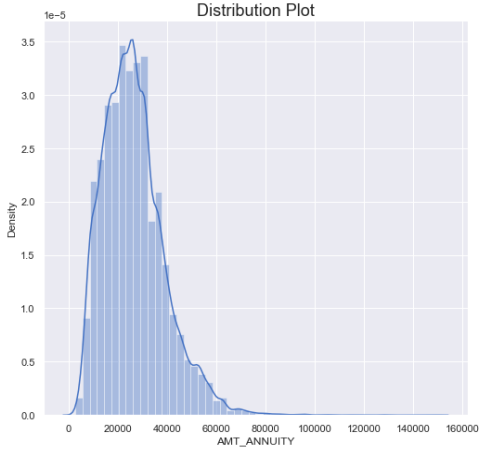
79% Loan applicants for ***payment difficulties*** have pursued Secondary education (**9% more education** as compared to ***non loan payment difficulty*** applicants)

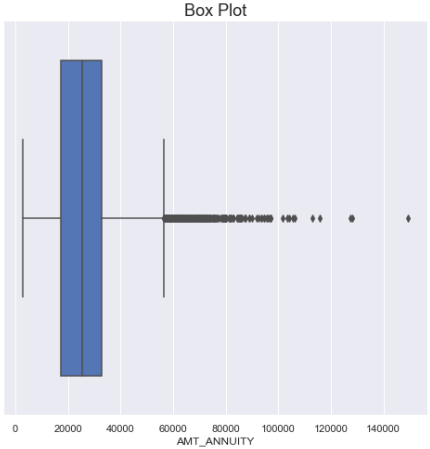
SEGMENTED UNIVARIATE ANALYSIS





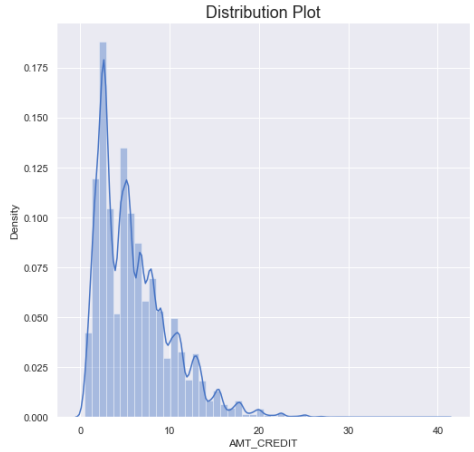
'AMT_ANNUITY' for Loan Non-Payment Difficulties

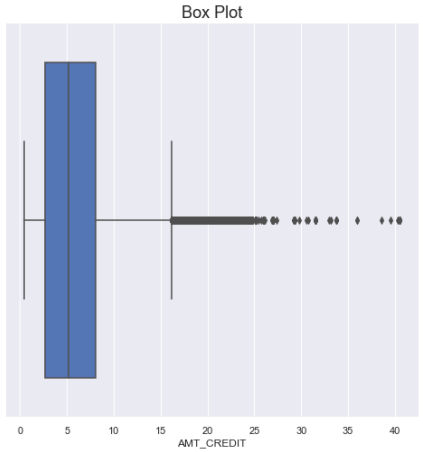




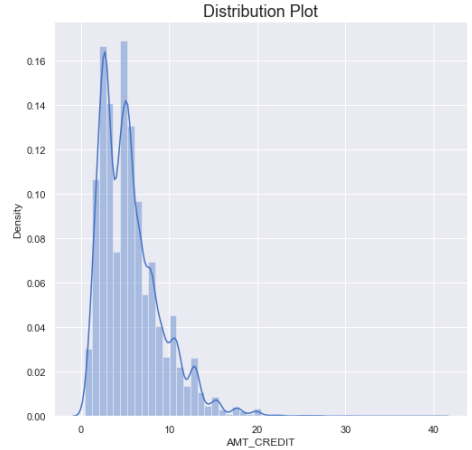
'AMT_ANNUITY' for Loan Payment Difficulties

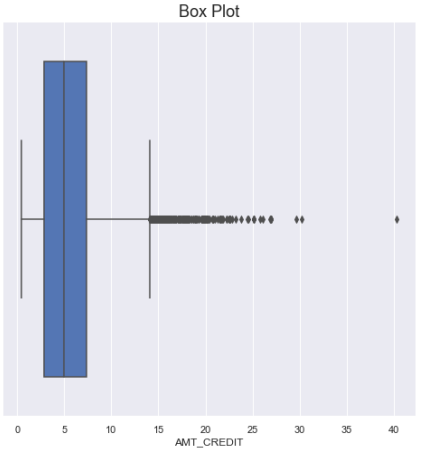
SEGMENTED UNIVARIATE ANALYSIS





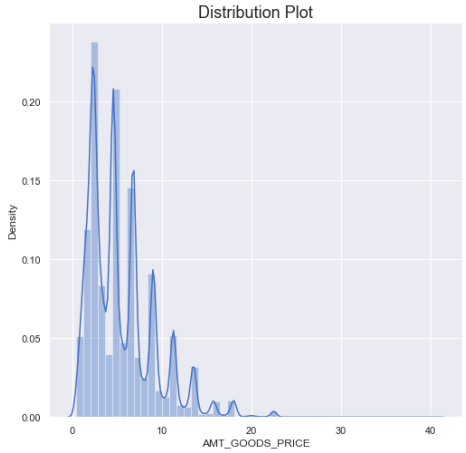
'AMT_CREDIT' for Loan Non-Payment Difficulties

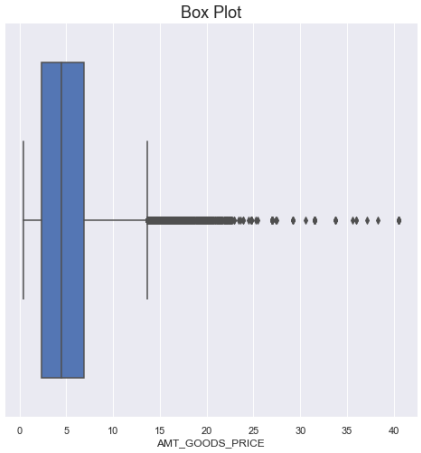




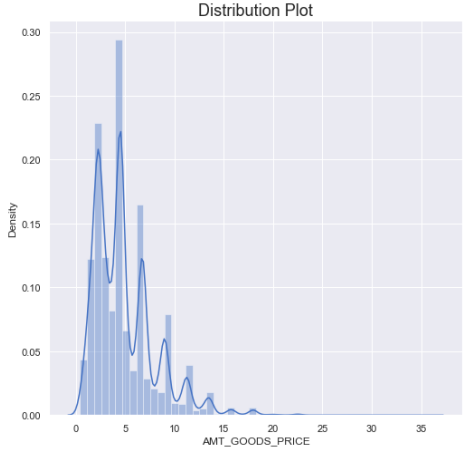
'AMT_CREDIT' for Loan Payment Difficulties

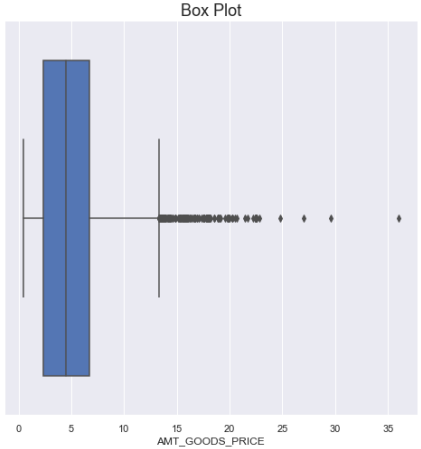
SEGMENTED UNIVARIATE ANALYSIS





'AMT_GOOD_PRICE' for Loan Non-Payment Difficulties





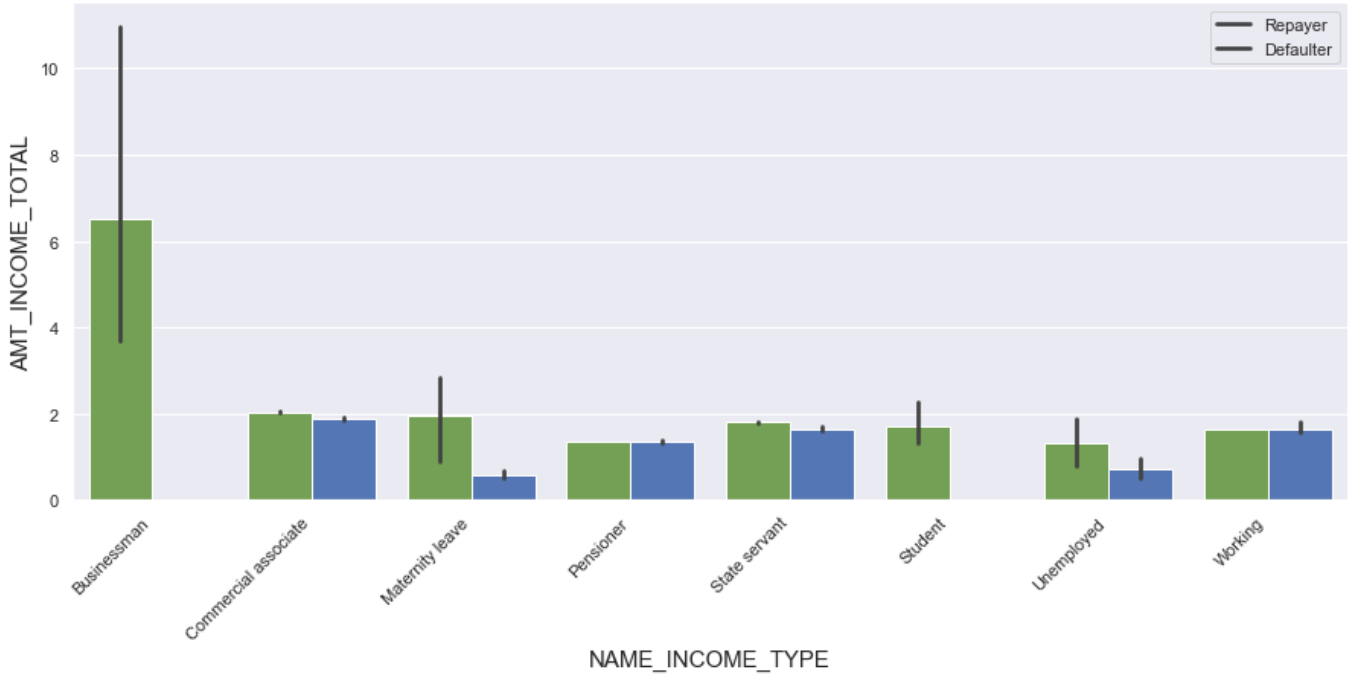
'AMT_GOOD_PRICE' for Loan Payment Difficulties

INSIGHTS

Segmented Univariate Analysis

- Most no. of loans are given for goods price below 10 lakhs
- Credit amount of the loan is mostly less then 10 lakhs
- Most people pay annuity below 50K for the credit loan

BIVARIATE ANALYSIS



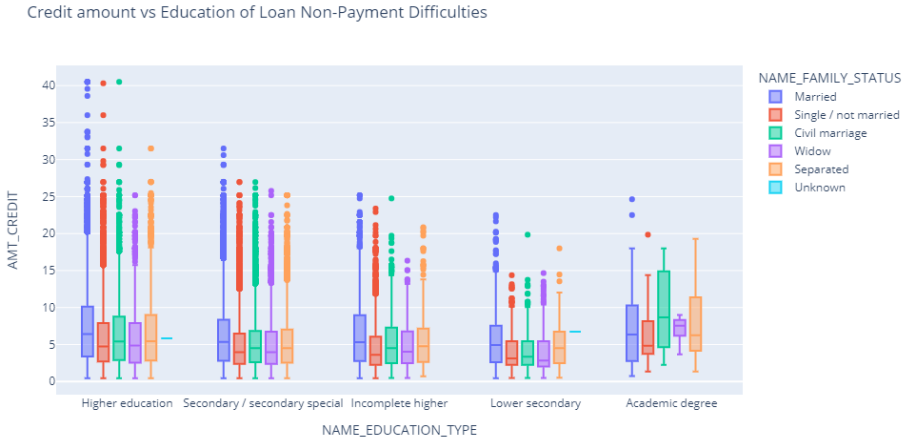
Income type vs Income Amount Range on a Seaborn Barplot

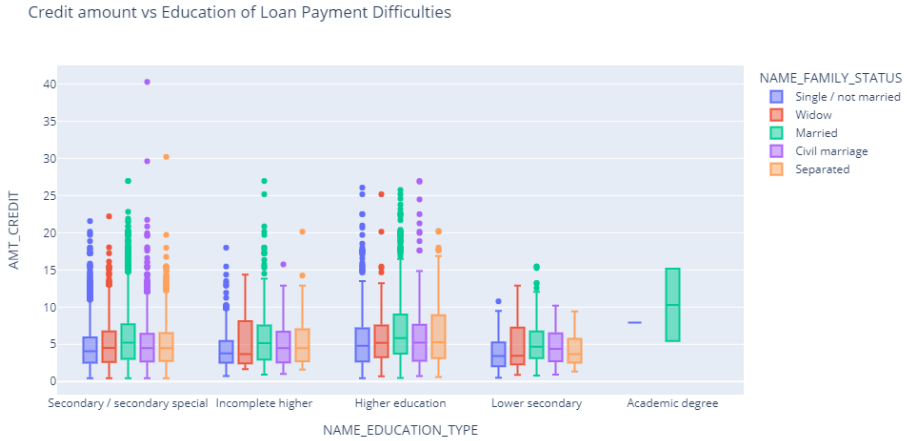
INSIGHTS

Bivariate Analysis

It can be seen that Businessman income is the highest and the estimated range seem to indicate that the income of a Businessman could be in the range close to 4 lakhs and slightly above 10 lakhs.

BIVARIATE ANALYSIS OF CATEGORICAL VS NUMERICAL VARIABLES





INSIGHTS

Bivariate Analysis Of Categorical VS Numerical Variables

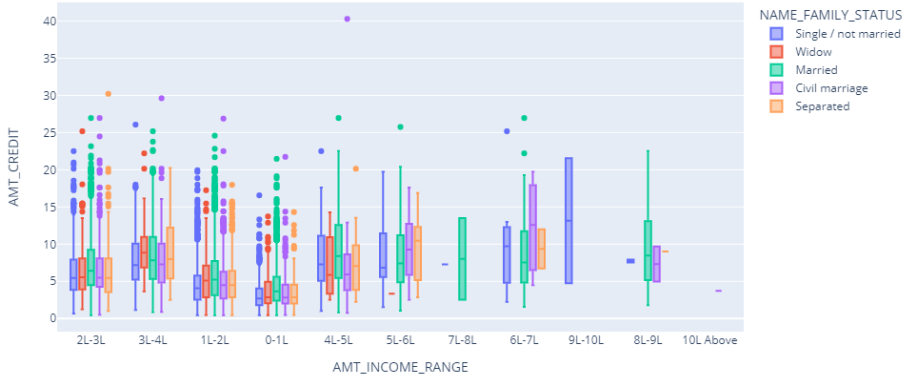
The graphs ***Credit Amount VS Education*** for Loan Payment Difficulties and Loan Non-Payment Difficulties appears to be similar.

- 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 3rd Quartile.

BIVARIATE ANALYSIS OF CATEGORICAL VS NUMERICAL VARIABLES

Income range vs Credit amount of Loan Non- Payment Difficulties

Income range vs Credit amount of Loan Payment Difficulties



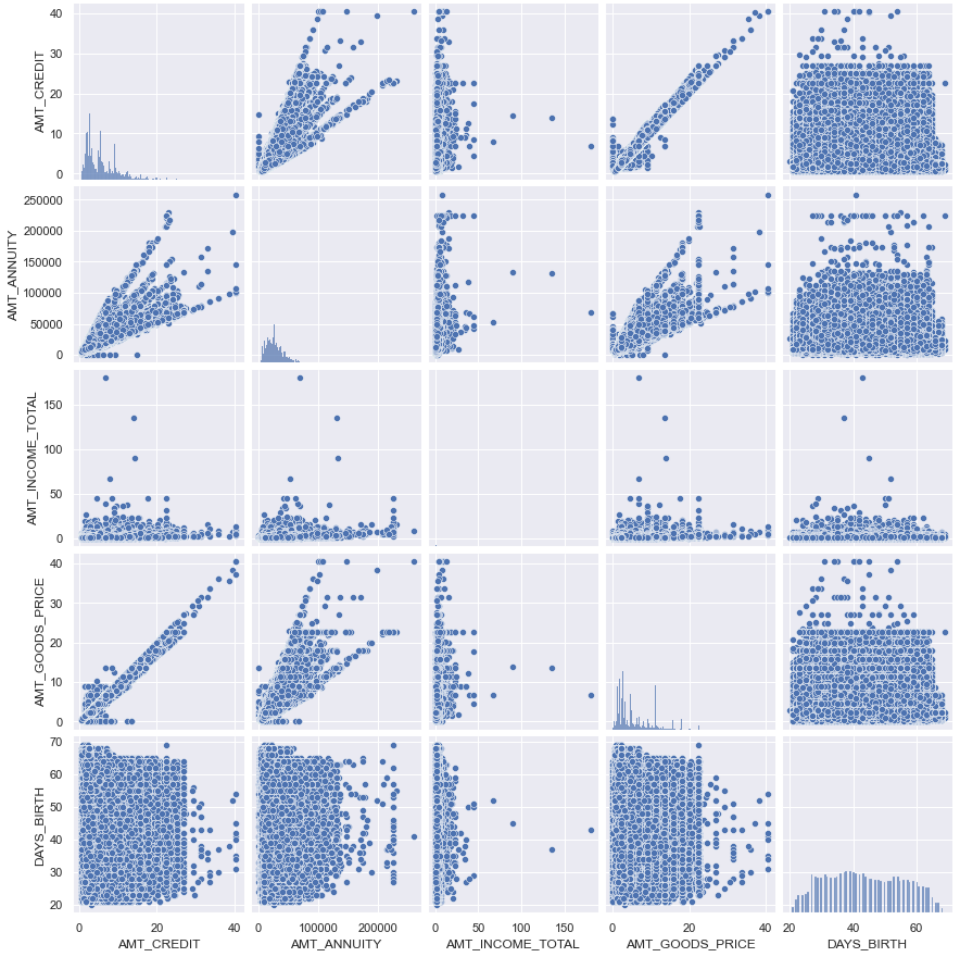
INSIGHTS

Bivariate Analysis Of Categorical VS Numerical Variables

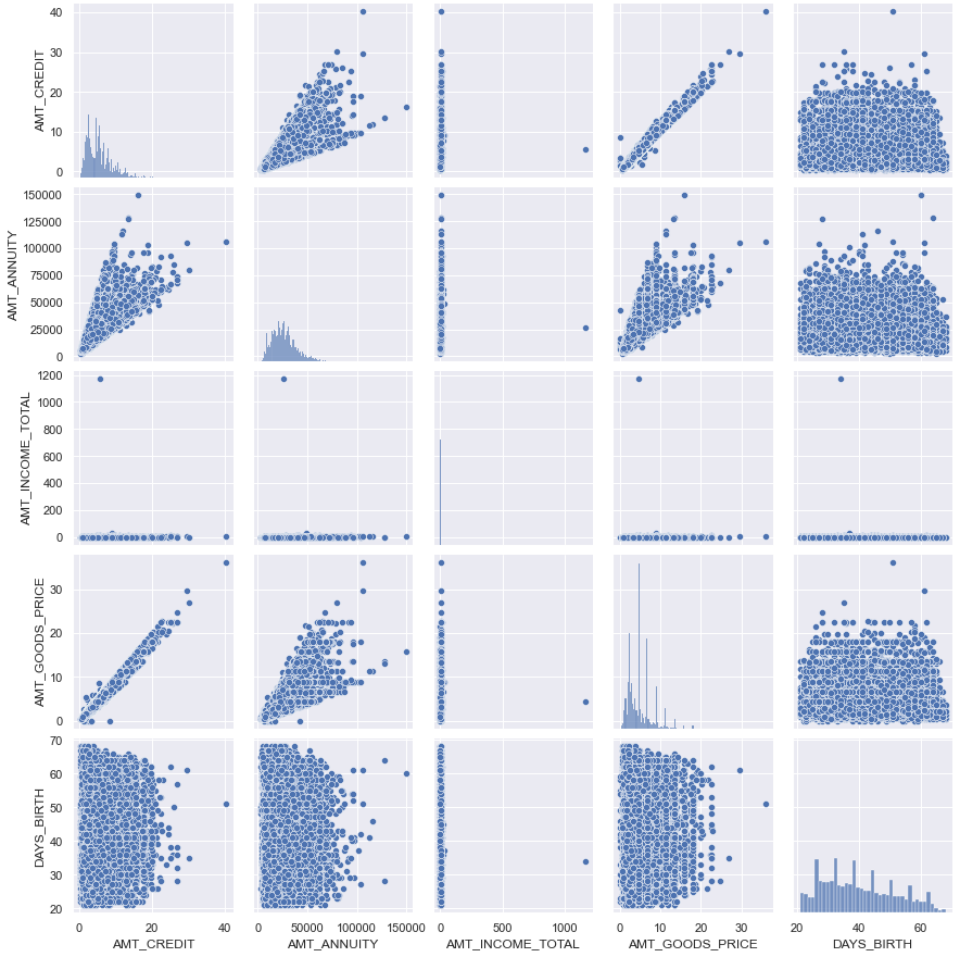
The graphs ***Credit Amount VS Income Range*** for Loan Payment Difficulties and Loan Non-Payment Difficulties appears to be similar.

- We observe that 'Single/ not Married' have the highest credit with income range of 9L – 10L
- Maximum outliers are for 'Married' Applicant in the range of income 1L – 2L
- 'Single' Applicants with Loan Payment Difficulties with Credit > 20 have income range of 9L to 10L in 3rd Quartile
- 'Separated' Applicants with Loan Non-Payment Difficulties with Credit > 25 have income range of 10L & Above in 3rd Quartile

LOAN NON PAYMENT DIFFICULTIES



LOAN PAYMENT DIFFICULTIES



TOP 10 CORRELATIONS DEFAULTERS AND REPAYERS

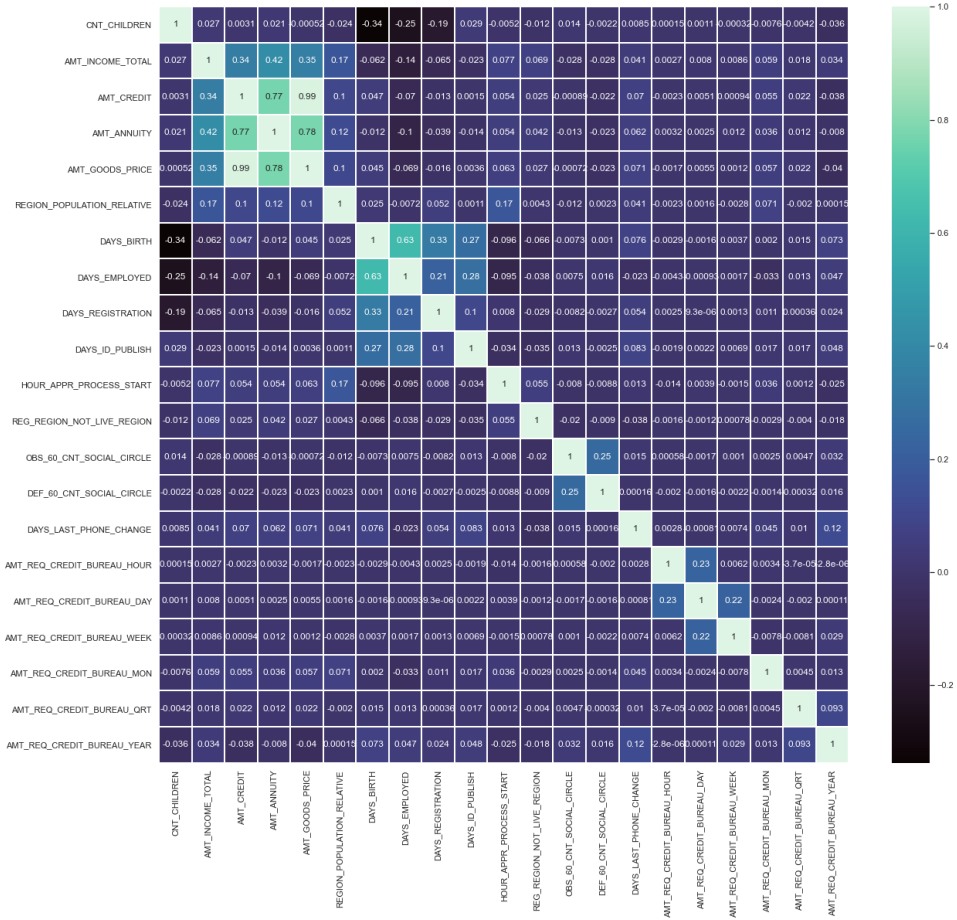
REPAYERS

	VAR1	VAR2	Correlation
86	AMT_GOODS_PRICE	AMT_CREDIT	0.987250
87	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686
65	AMT_ANNUITY	AMT_CREDIT	0.771309
153	DAYS_EMPLOYED	DAYS_BIRTH	0.626028
64	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953
85	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349462
43	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799
126	DAYS_BIRTH	CNT_CHILDREN	0.336907
174	DAYS_REGISTRATION	DAYS_BIRTH	0.333025
196	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.276663

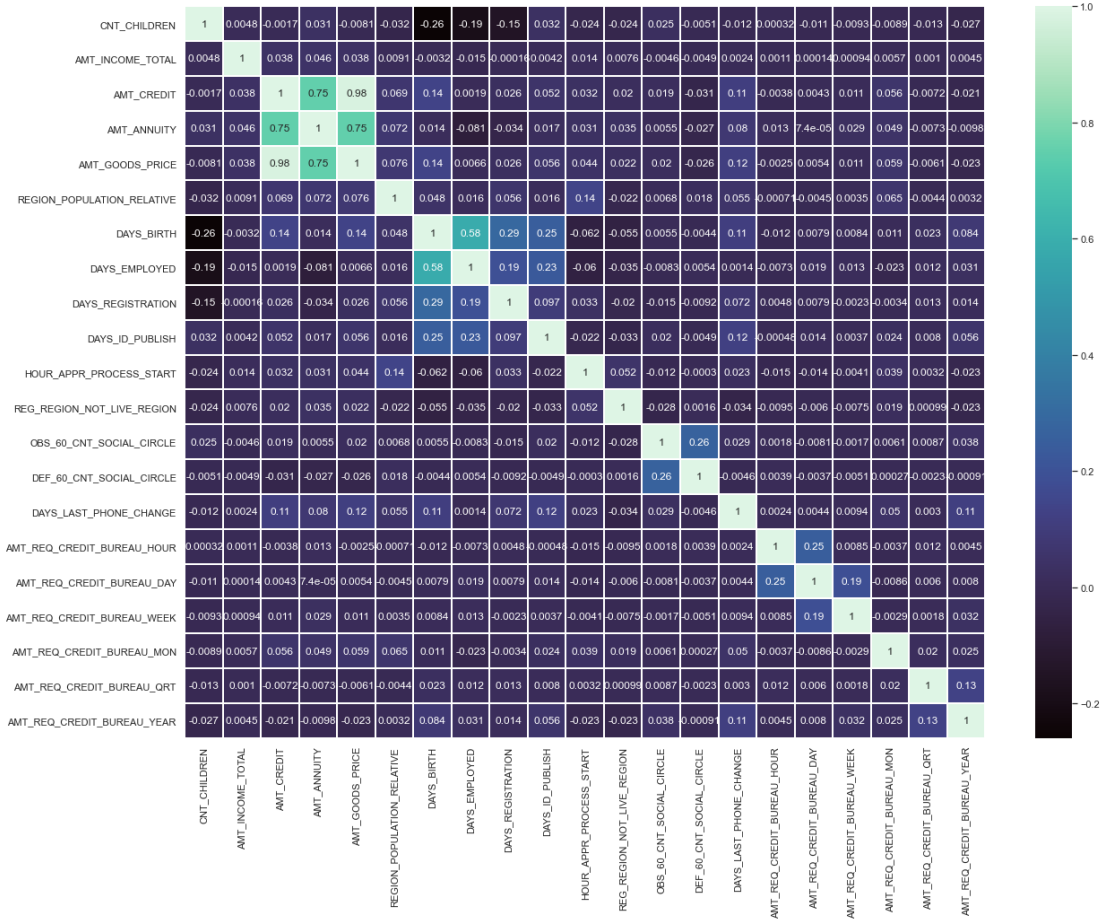
DEFAULTERS

	VAR1	VAR2	Correlation
86	AMT_GOODS_PRICE	AMT_CREDIT	0.983103
87	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699
65	AMT_ANNUITY	AMT_CREDIT	0.752195
153	DAYS_EMPLOYED	DAYS_BIRTH	0.582441
174	DAYS_REGISTRATION	DAYS_BIRTH	0.289116
285	DEF_60_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.264159
126	DAYS_BIRTH	CNT_CHILDREN	0.259222
195	DAYS_ID_PUBLISH	DAYS_BIRTH	0.252256
351	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_HOUR	0.247511
196	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.229090

HEATMAP TO SEE LINEAR CORRELATION AMONG REPAYERS



HEATMAP TO SEE LINEAR CORRELATION AMONG DEFAULTERS

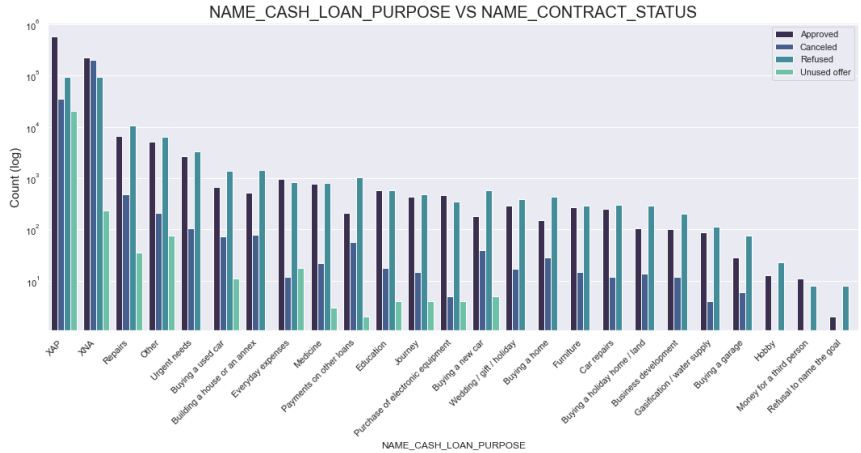


INSIGHTS

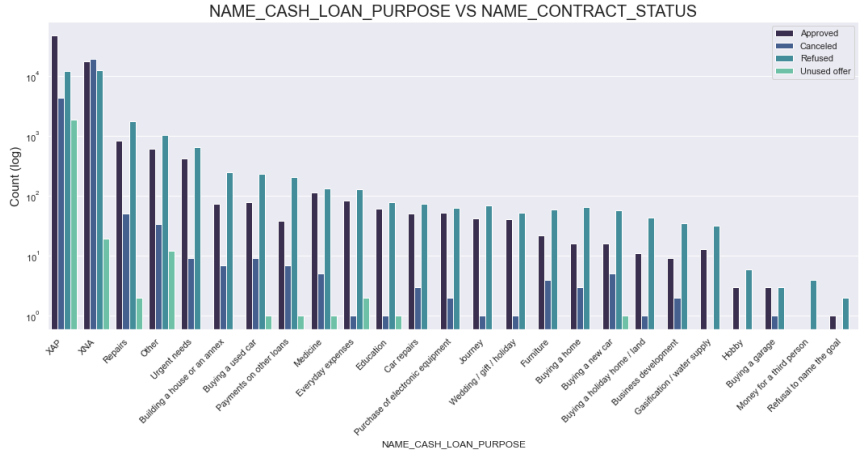
Heatmap To See Linear Correlation Among Defaulters & Repayers

- Credit amount is highly correlated with good price amount which is same as repayers.
- Loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayers(0.77).
- We can also see that repayers have high correlation in number of days employed(0.62) when compared to defaulters(0.58).
- There is a slight increase in defaulted to observed count in social circle among defaulters(0.264) when compared to repayers(0.254)

MERGED DATA ANALYSIS
STATUS OF LOAN APPLICATION
VS PURPOSE



Loan Non-Payment Difficulties



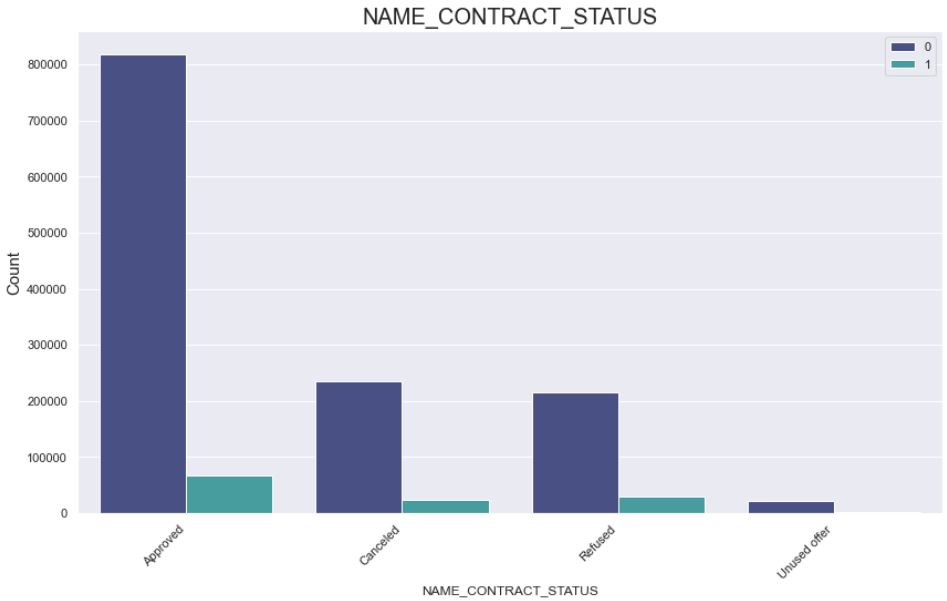
Loan Payment Difficulties

INSIGHTS

Merged Data Analysis Status Of Loan Application Vs Purpose

- Loan purpose has high number of unknown values (XAP, XNA)
- Loan taken for the purpose of Repairs looks to have highest default rate
- Huge number application have been rejected by bank or refused by client which are applied for Repair or Other. from this we can infer that repair is considered high risk by bank. Also, either they are rejected or bank offers loan on high interest rate which is not feasible by the clients and they refuse the loan.

CHECKING CONTRACT STATUS BASED ON LOAN REPAYMENT STATUS WHETHER THERE IS ANY BUSINESS LOSS OR FINANCIAL LOSS

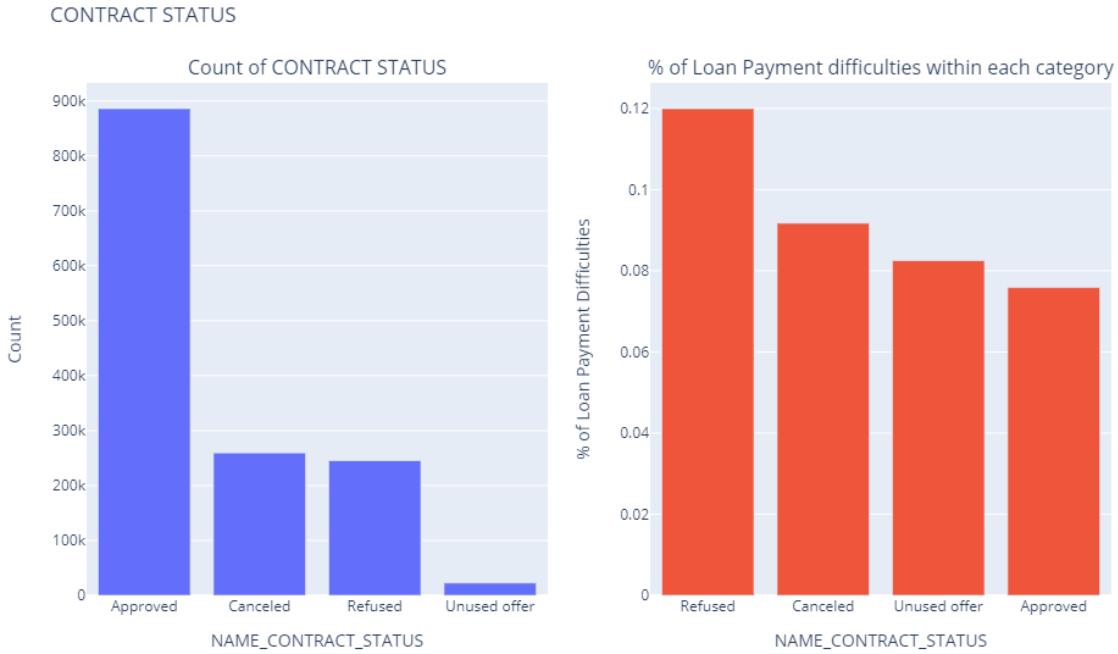


NAME_CONTRACT_STATUS	TARGET	Counts	Percentage
Approved	0	818856	92.41%
	1	67243	7.59%
Canceled	0	235641	90.83%
	1	23800	9.17%
Refused	0	215952	88.0%
	1	29438	12.0%
Unused offer	0	20892	91.75%
	1	1879	8.25%

INSIGHTS

1. 90% of the previously cancelled client have actually repayed the loan. Revising the interest rates would increase business opportunity for these clients
2. 88% of the clients who have been previously refused a loan has payed back the loan in current case.

DISTRIBUTION OF CONTRACT STATUS AND ITS CATEGORY WITH MAXIMUM % OF LOAN-PAYMENT DIFFICULTIES



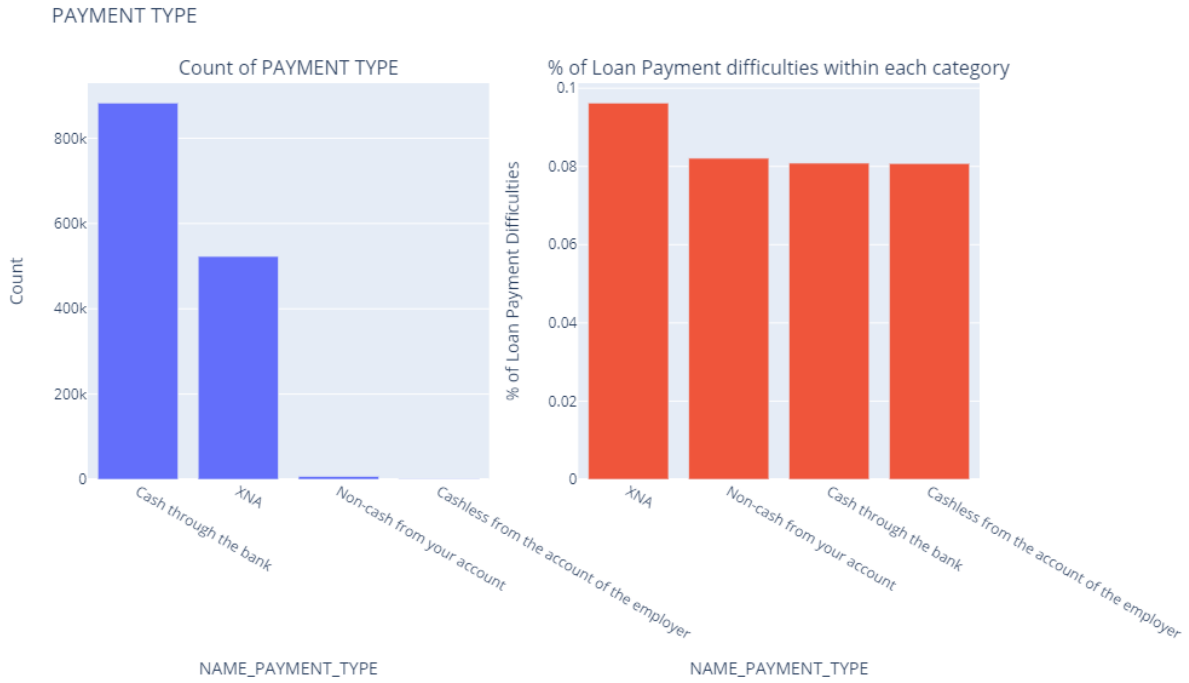
INSIGHTS

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

1.'Revolving Loans' contracts from previous application are the ones who have maximum % of Loan-Payment Difficulties from current application.

2.'Consumer loans' contracts from previous application are the ones who have minimum % of Loan-Payment Difficulties from current application.

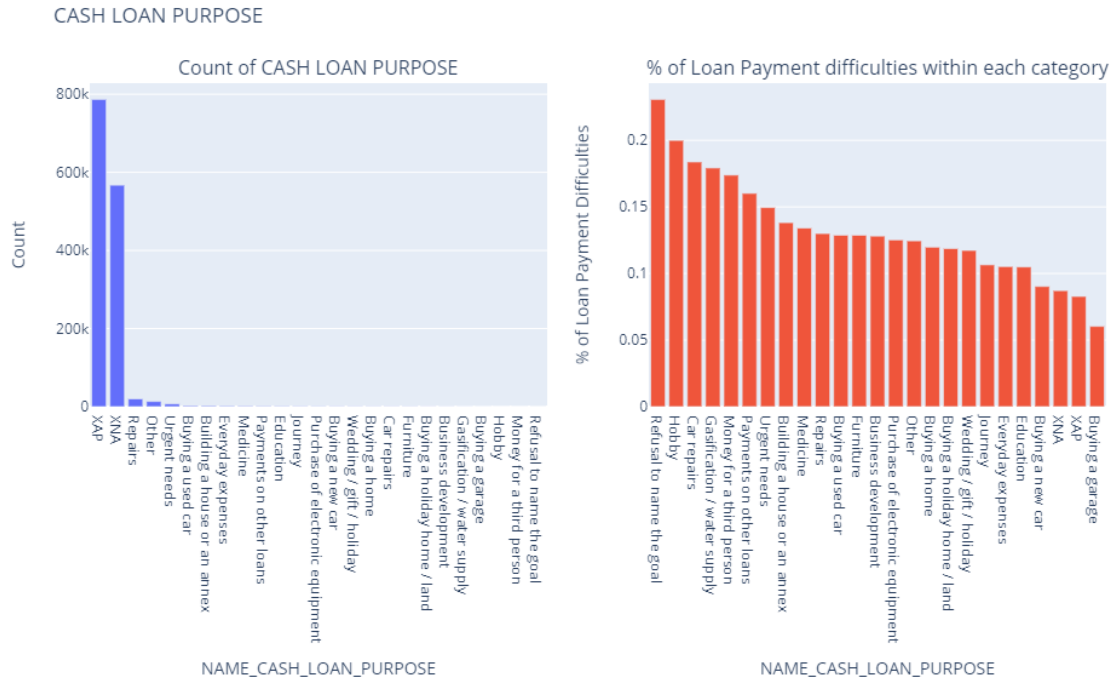
DISTRIBUTION OF PAYMENT TYPE AND ITS CATEGORY WITH MAXIMUM % OF LOAN-PAYMENT DIFFICULTIES



INSIGHTS

From the first graph it can be seen that most of the payment type from previous application was 'Cash through bank'. It can be clearly seen from the second graph that all three types of payments from the previous application have almost same % of Loan-Payment Difficulties from current application

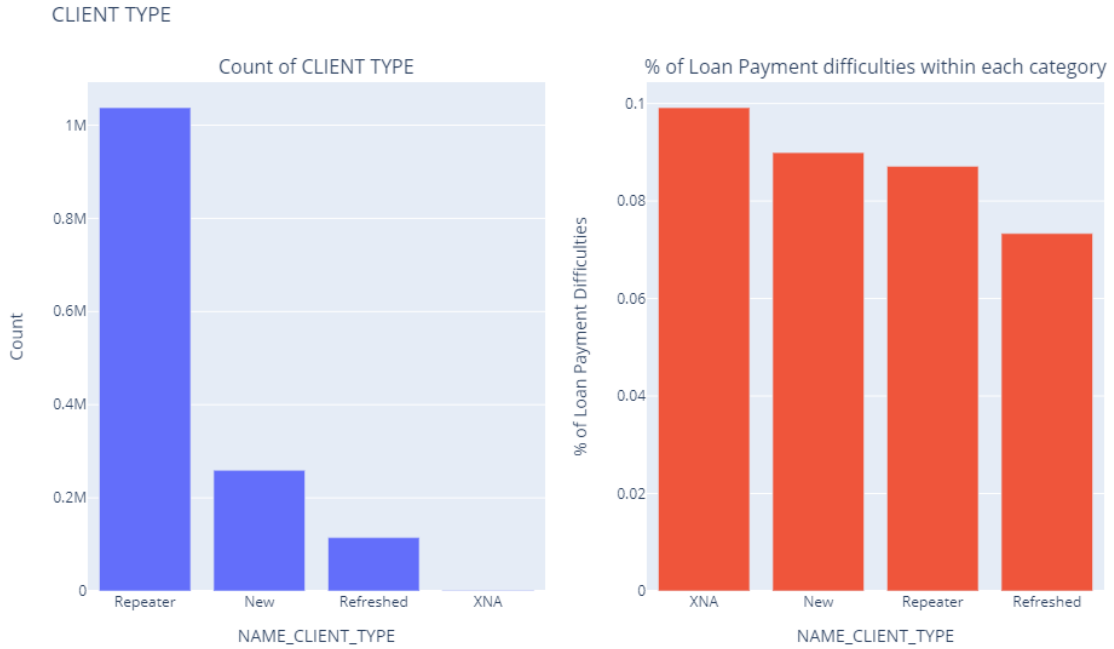
DISTRIBUTION OF CASH LOAN PURPOSE AND ITS CATEGORY WITH MAXIMUM % OF LOAN-PAYMENT DIFFICULTIES



INSIGHTS

From the first graph it can be seen that purpose of cash loan from previous data was maximum for 'Repairs'. It can be clearly seen from the second graph that the 'Refusal to name the goal' for cash loan from previous application are the ones who have maximum % of Loan-Payment Difficulties from current application.

DISTRIBUTION OF CLIENT TYPE AND ITS CATEGORY WITH MAXIMUM % OF LOAN-PAYMENT DIFFICULTIES

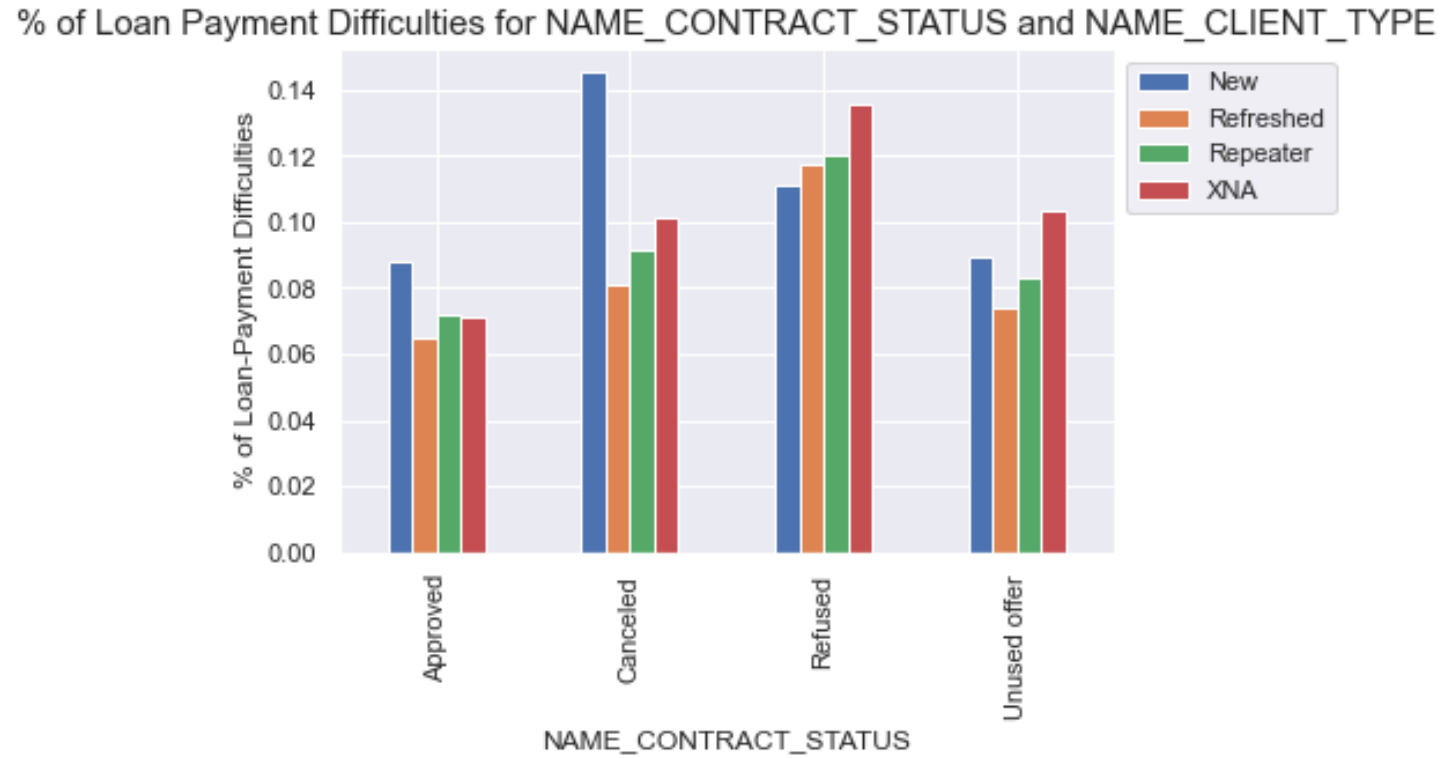


INSIGHTS

From the first graph it can be seen that most of the clients from previous application are 'Repeater'
It can be clearly seen from the second graph that the

1. 'New' clients from previous application are the ones who have maximum % of Loan-Payment Difficulties from current application.
2. 'Refreshed' clients from previous application are the ones who have minimum % of Loan-Payment Difficulties from current application.

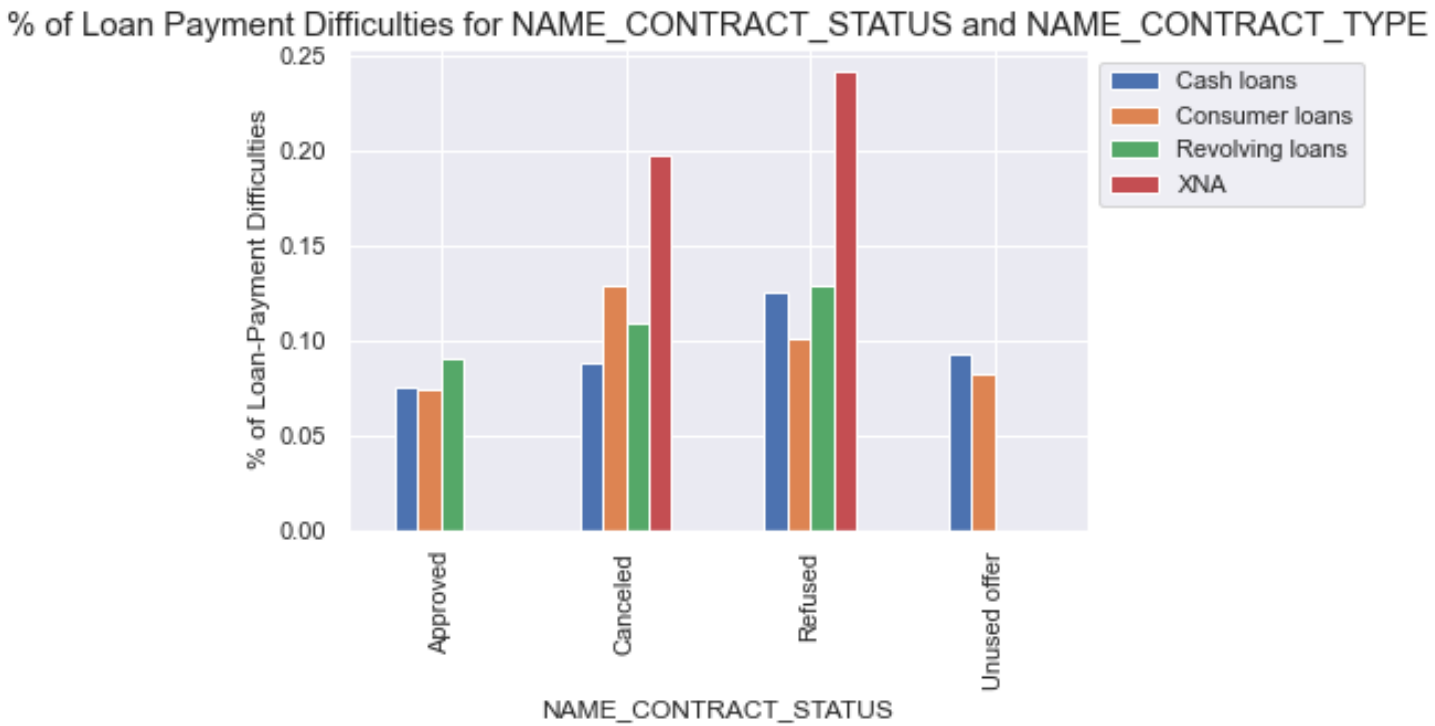
% OF LOAN PAYMENT DIFFICULTIES FOR NAME_CONTRACT_STATUS AND NAME_CLIENT_TYPE



INSIGHTS

It can be observed from the above graph that Client who where 'New' and had 'Cancelled' previous application tend to have more % of Loan-Payment Difficulties in current application

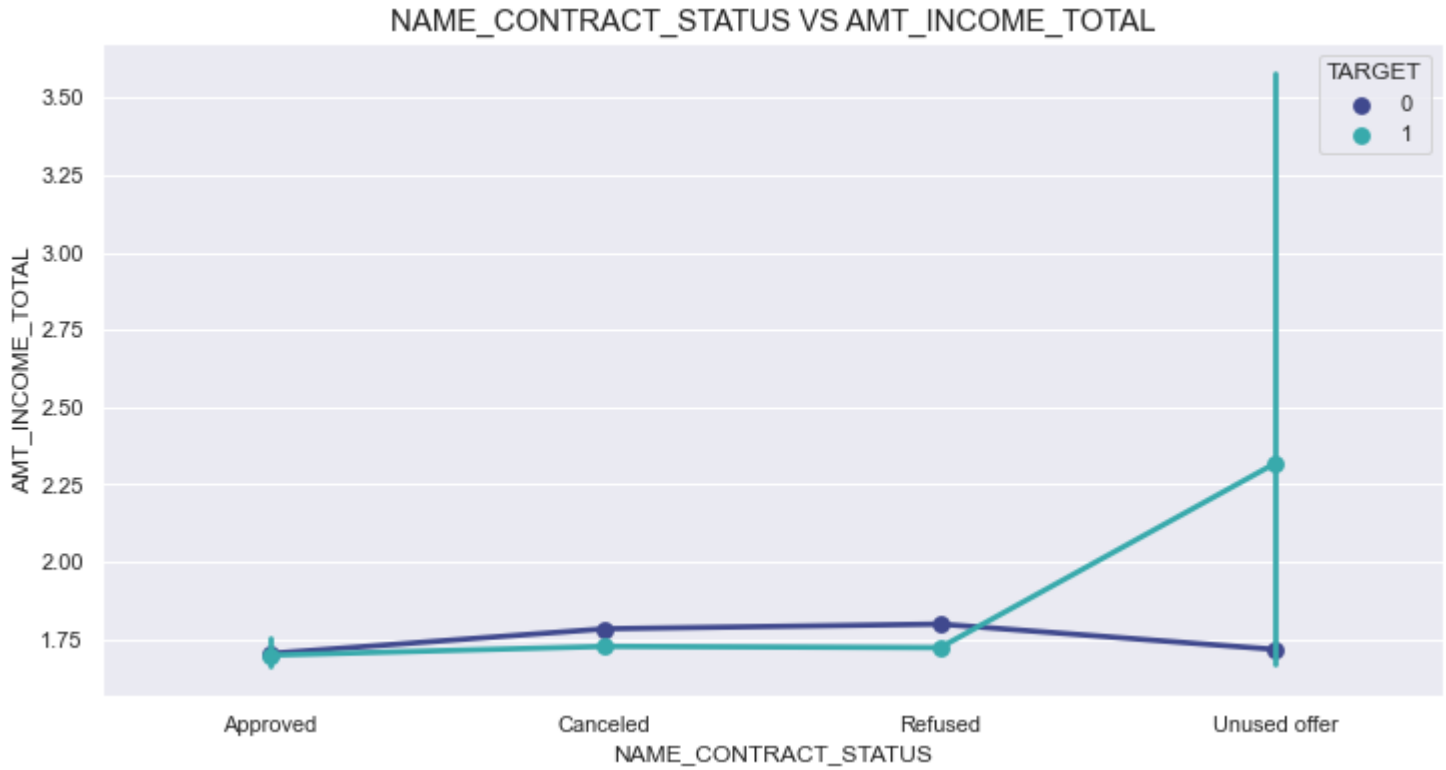
% OF LOAN PAYMENT DIFFICULTIES FOR NAME_CONTRACT_STATUS AND NAME_CONTRACT_TYPE



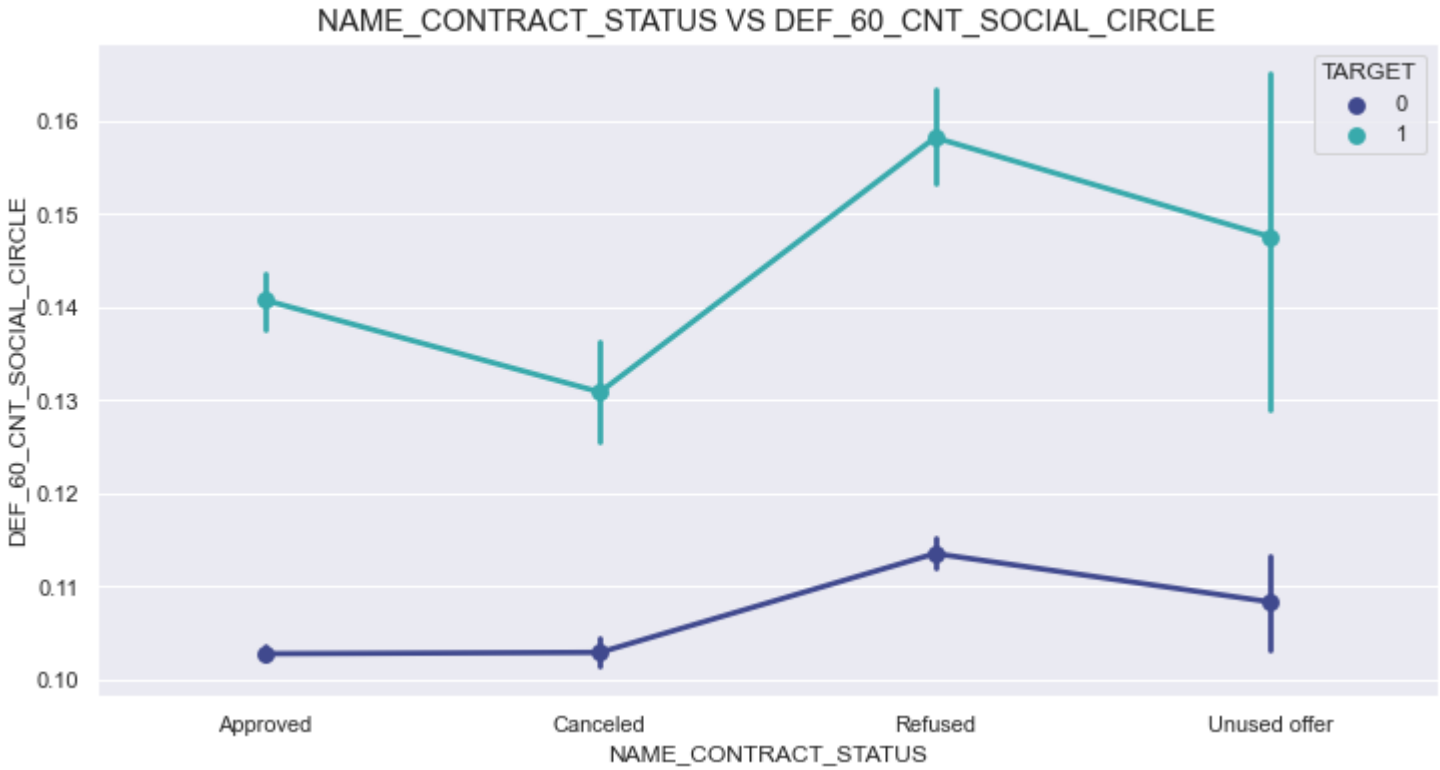
INSIGHTS

It can be observed from the above graph that maximum loans were Refused and lesser loans were approved. It is observed that 'Revolving Loans' have the highest average % of loan-payment difficulties.

RELATIONSHIP BETWEEN INCOME TOTAL AND CONTACT STATUS



RELATIONSHIP BETWEEN PEOPLE WHO DEFAULTED IN LAST 60 DAYS BEING IN CLIENT'S SOCIAL CIRCLE AND CONTACT STATUS



CONCLUSION

After analysing the datasets, there are few attributes of a client with which the bank would be able to identify if they will repay the loan or not. The analysis is consised as below with the contributing factors and categorization:

A. Decisive Factor whether an applicant will be Repayer:

1. NAME_EDUCATION_TYPE: Academic degree has less defaults.
2. NAME_INCOME_TYPE: Student and Businessmen have no defaults.
3. REGION_RATING_CLIENT: RATING 1 is safer.
4. ORGANIZATION_TYPE: Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%
5. DAYS_BIRTH: People above age of 50 have low probability of defaulting
6. DAYS_EMPLOYED: Clients with 40+ year experience having less than 1% default rate
7. AMT_INCOME_TOTAL:Applicant with Income more than 700,000 are less likely to default
8. NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, Buying garage are being repayed mostly.
9. CNT_CHILDREN: People with zero to two children tend to repay the loans.

CONCLUSION

B. Decisive Factor whether an applicant will be Defaulter:

1. CODE_GENDER: Men are at relatively higher default rate
2. NAME_FAMILY_STATUS : People who have civil marriage or who are single default a lot.
3. NAME_EDUCATION_TYPE: People with Lower Secondary and Secondary education
4. NAME_INCOME_TYPE: Clients who are either at Maternity leave OR Unemployed default a lot.
5. REGION_RATING_CLIENT: People who live in Rating 3 has highest defaults.
6. OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as their default rate is huge.
7. ORGANIZATION_TYPE: Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self-employed people have relative high defaulting rate, and thus should be avoided to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.
8. DAYS_BIRTH: Avoid young people who are in age group of 20-40 as they have higher probability of defaulting
9. DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.
10. CNT_CHILDREN & CNT_FAM_MEMBERS: Client who have children equal to or more than 9 default 100% and hence their applications are to be rejected.
11. AMT_GOODS_PRICE: When the credit amount goes beyond 3lakhs, there is an increase in defaulters.

CONCLUSION

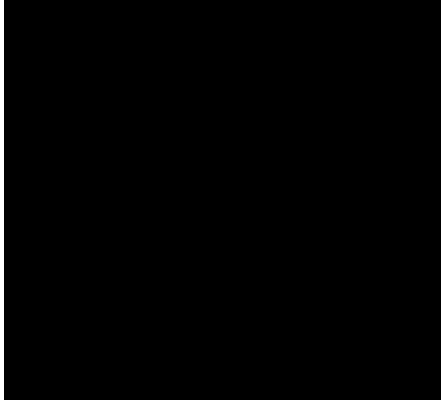
C. Factors that Loan can be given on Condition of High Interest rate to mitigate any default risk leading to business loss:

1. NAME_HOUSING_TYPE: High number of loan applications are from the category of people who live in Rented apartments & living with parents and hence offering the loan would mitigate the loss if any of those default.
2. AMT_CREDIT: People who get loan for 3-6 Lakhs tend to default more than others and hence having higher interest specifically for this credit range would be ideal.
3. AMT_INCOME: Since 90% of the applications have Income total less than 3Lakhs and they have high probability of defaulting, they could be offered loan with higher interest compared to other income category.
4. CNT_CHILDREN & CNT_FAM_MEMBERS: Clients who have 4 to 8 children has a very high default rate and hence higher interest should be imposed on their loans.
5. NAME_CASH_LOAN_PURPOSE: Loan taken for the purpose of Repairs seems to have highest default rate. A very high number applications have been rejected by bank or refused by client in previous applications as well which has purpose as repair or other. This shows that purpose repair is taken as high risk by bank and either they are rejected, or bank offers very high loan interest rate which is not feasible by the clients, thus they refuse the loan. The same approach could be followed in future as well.

TEAM DETAILS:

Anishma Kathpal
+91 70839 82723
anishmakathpal1503@gmail.com

Harshitha C
+91 99456 41660
harshithackashyap@gmail.com



THANK YOU!