CREDIT EDA CASE STUDY

BY-

SAMBIT SEKHAR SAHU

JOSEPH BABU

Data Cleaning and Preparation

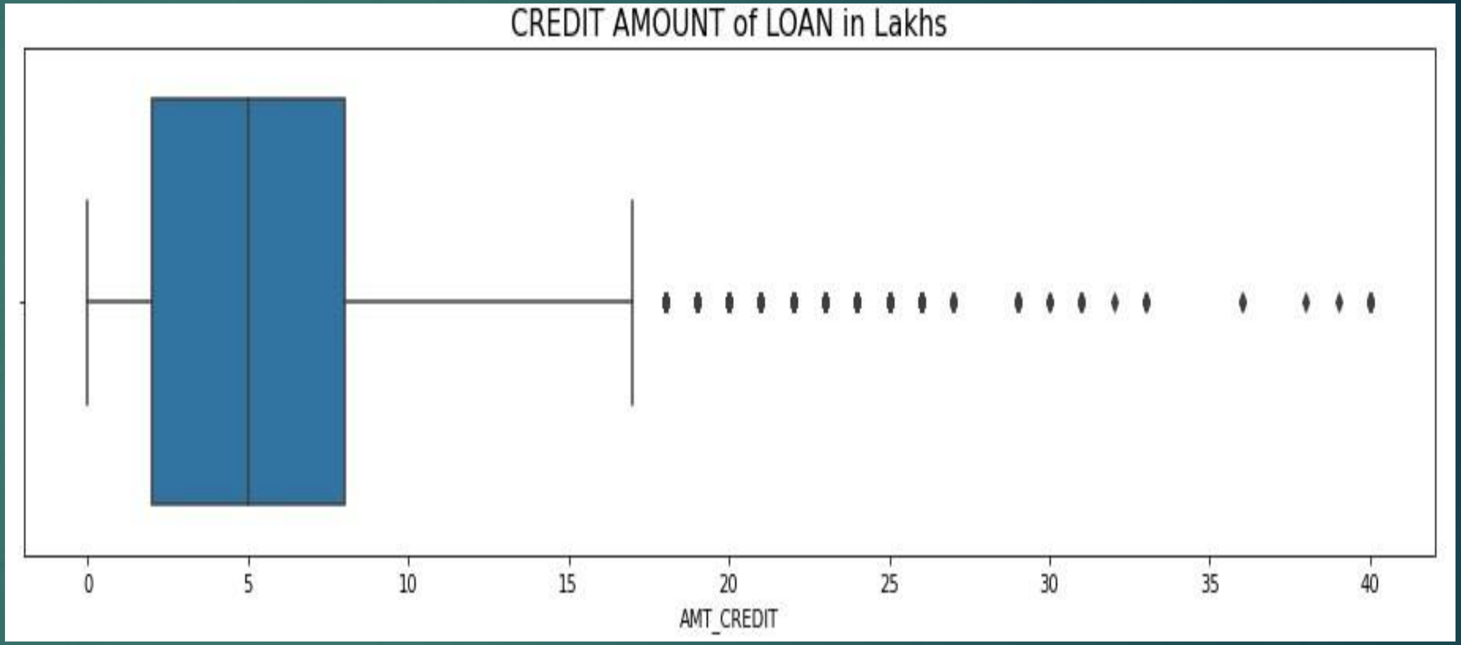
- While working on any data set, the first important step is to clean data and prepare it for further use.
- Columns with null value content of greater than 50% were dropped, according to Industry standards.
- Suitable data imputation techniques were suggested for columns having null values less than or equal to 13%, i.e., Mean/Median for Numerical Columns & Mode for Categorical Columns.

- Then we went on with the data quality checks, where we found Days related column had days in –ve values. So using abs() function –ve values were converted into positive values.
- Missing values of the form ‘XNA’ & ‘XNP’ were replaced with python understood NAN values using np.NAN function.
- Then we moved ahead with binning of continuous variable – Income & Age columns.

Analysis on Application Data

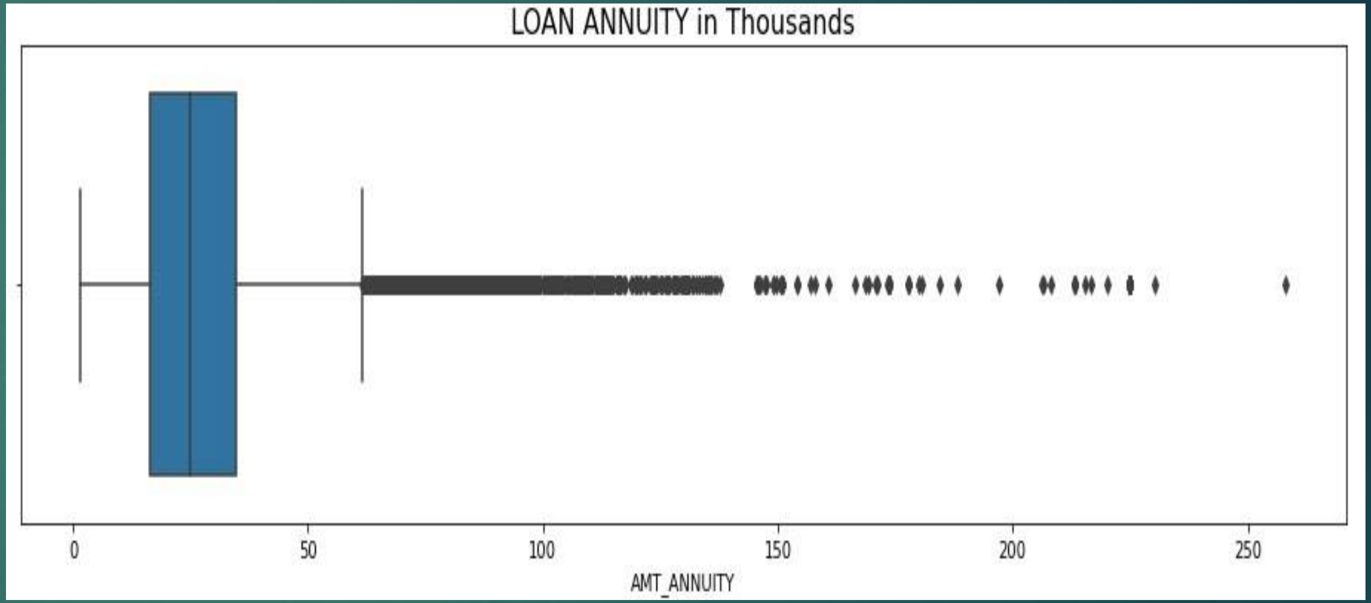
Outlier Analysis of Credit Amount

- We observe outliers are present within the range 20-40 Lakhs



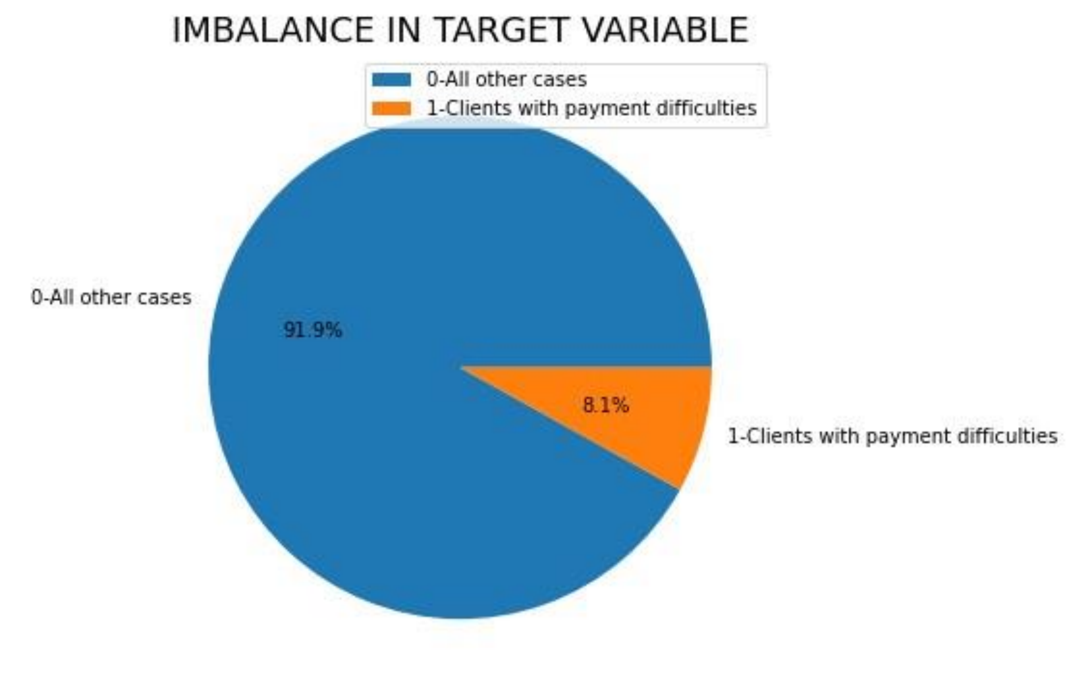
Outlier Analysis of Loan Annuity

- We can observe outliers here, >250k being the highest outlier value.



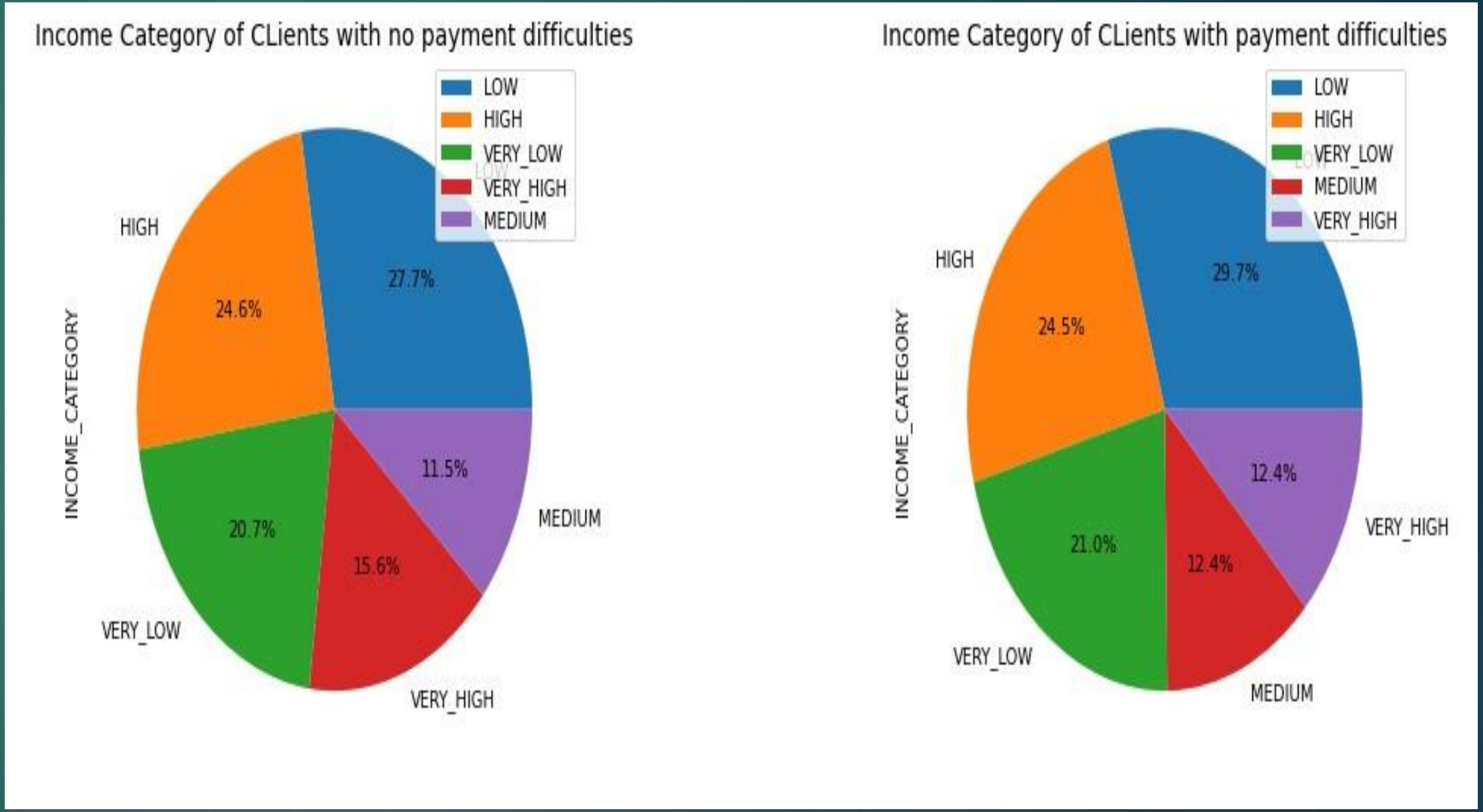
Data Imbalance in Target Variable

- We see the data is not equally distributed in TARGET variable between clients with payment difficulties and all other cases



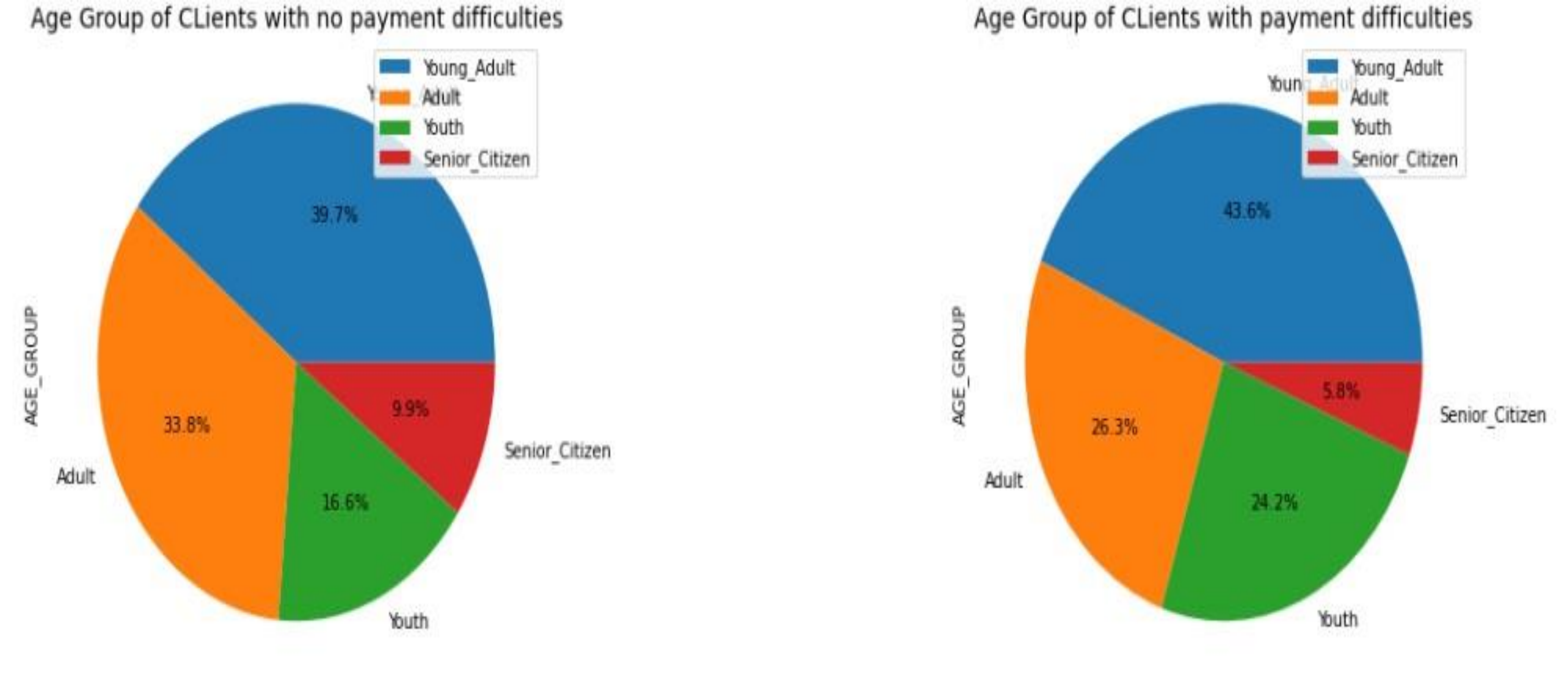
Income CATEOGORY with respect to target variable

- For the clinets falling in the low income category, the possibility of dafaulting has an increased by 2%.
- While clients falling in the Very High income category, the possibility of defaulting has reduced by 3.2%



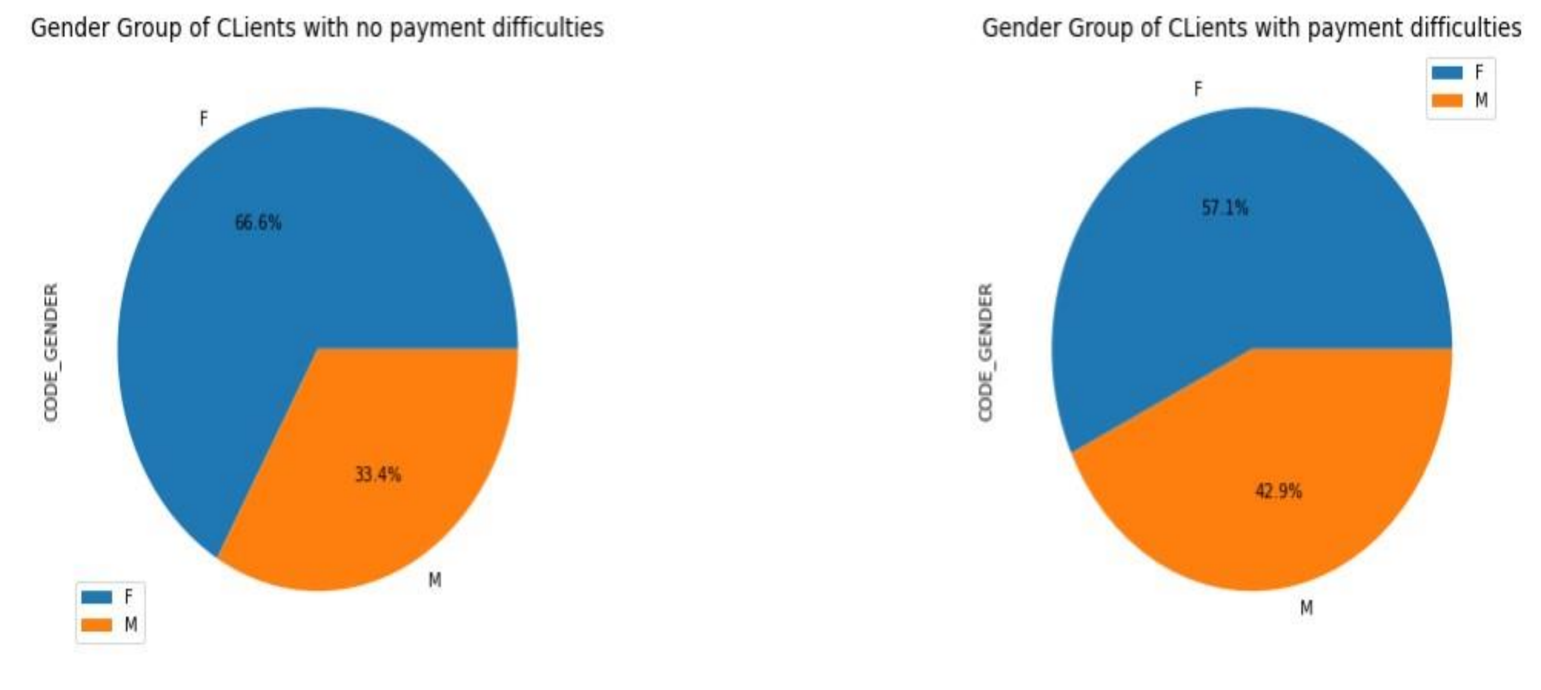
Age Group with respect to target variable

- In Young Adults, the possibility of defaulting has an increased by 3.9%.
- While in Senior Citizens category, the possibility of defaulting has decreased by 4.1%



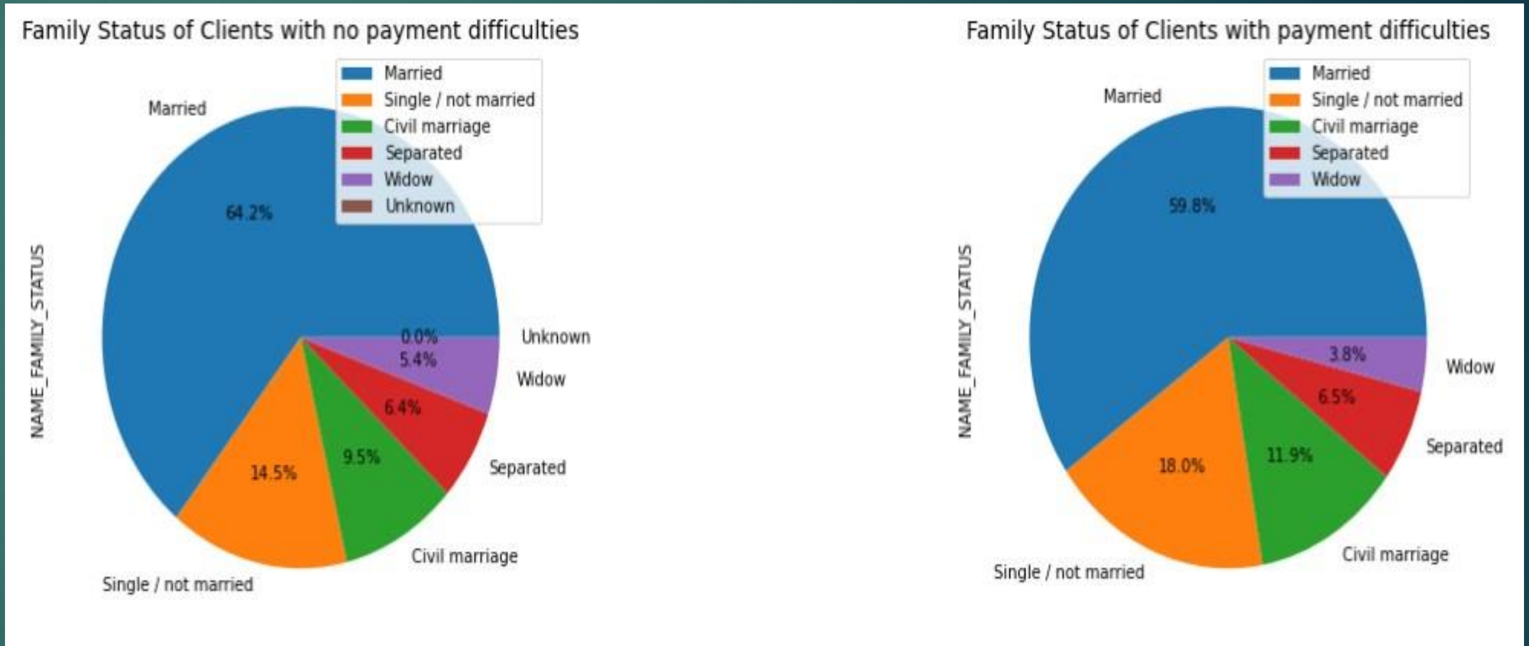
Gender with respect to target variable

- In Females the possibility of defaulting has reduced by 9.5%. While on the other hand the possibility of dafaulting has
- increased by 9.5%



Family Status with respect to target variable

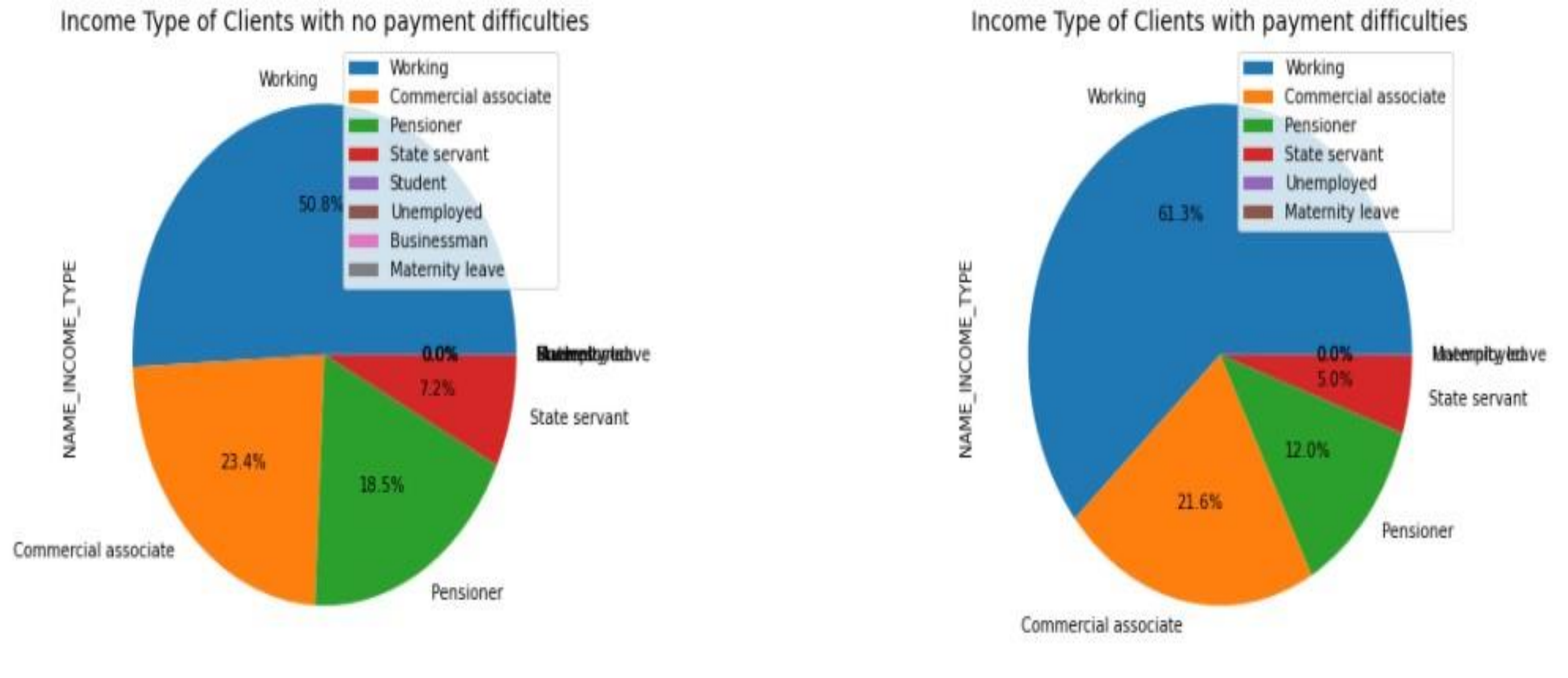
- In Married people payment difficulties has reduced by 4.4% & in widow category it has decreased by 1.6%.
- While in other categories we can see increase in payment difficulties whith Not Married category being highest at 3.5%


Family Status of Clients with no payment difficulties

Family Status of Clients with payment difficulties

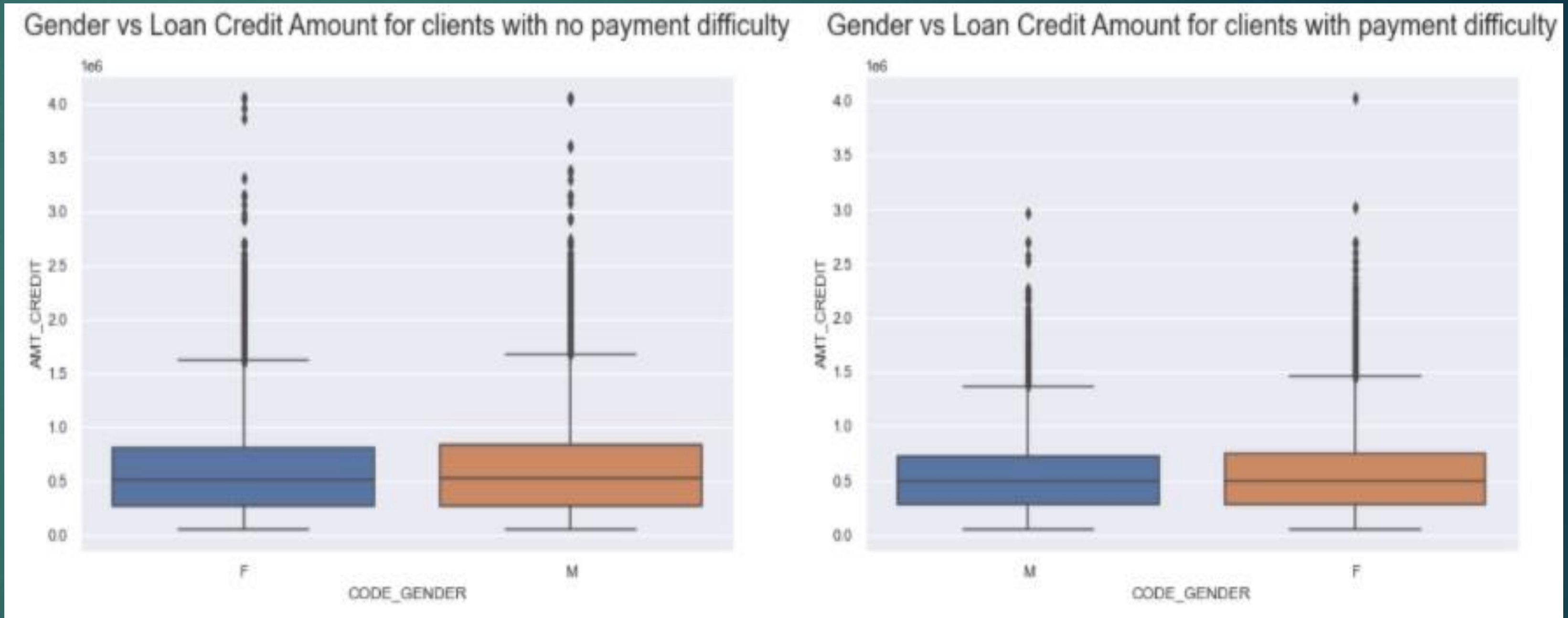
Income Type with respect to target variable

- We see payment difficulties has significantly incresed to 10.5% for Working category clients. While in other categories
- we see reduction in payment difficulties with 6.5% being highest in the Pensioner Group of clients



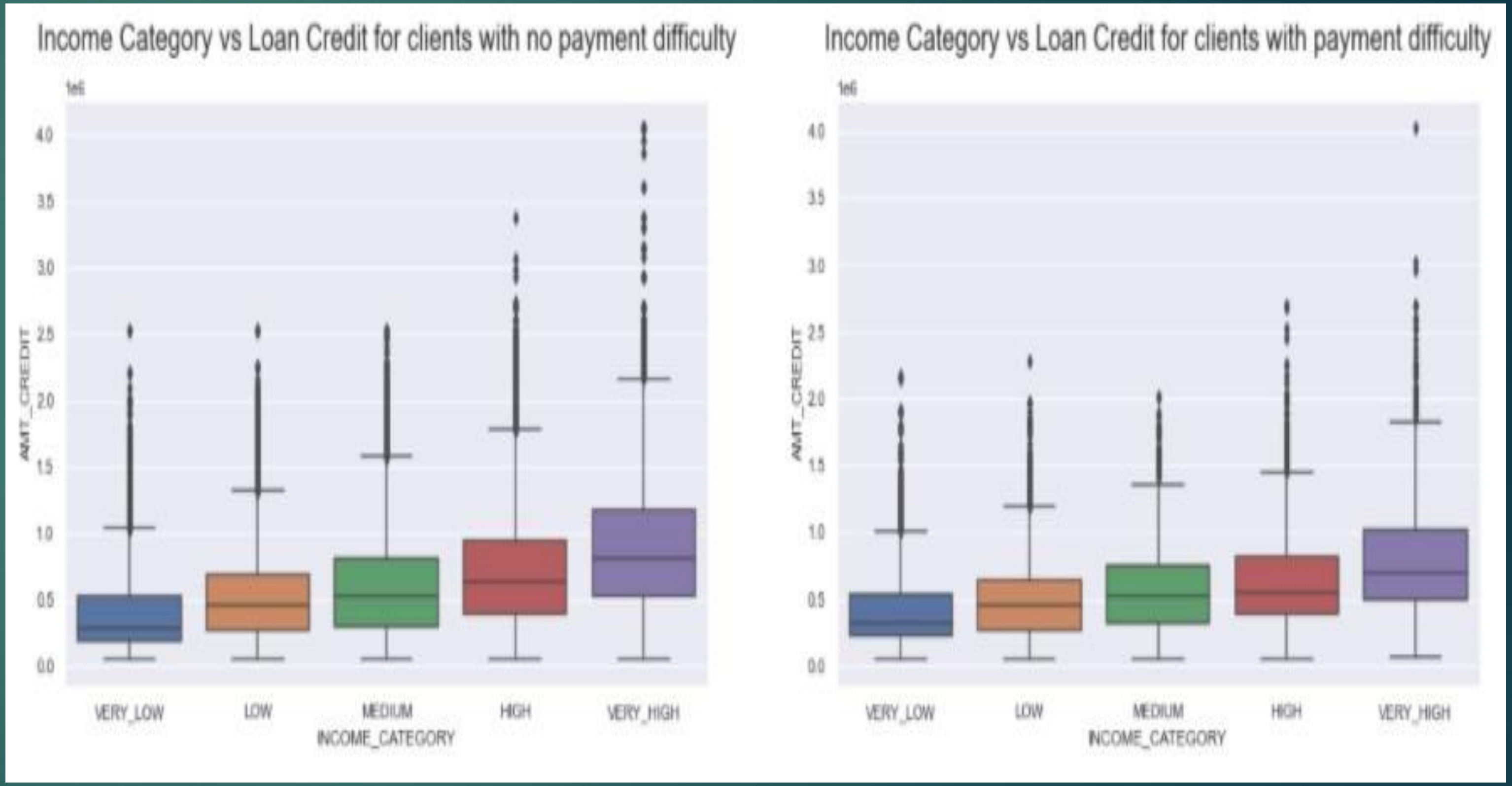
Analysis how Loan Credit amount varies with Gender for both target 0 & 1 type clients

- Both genders seems to perform almost equally



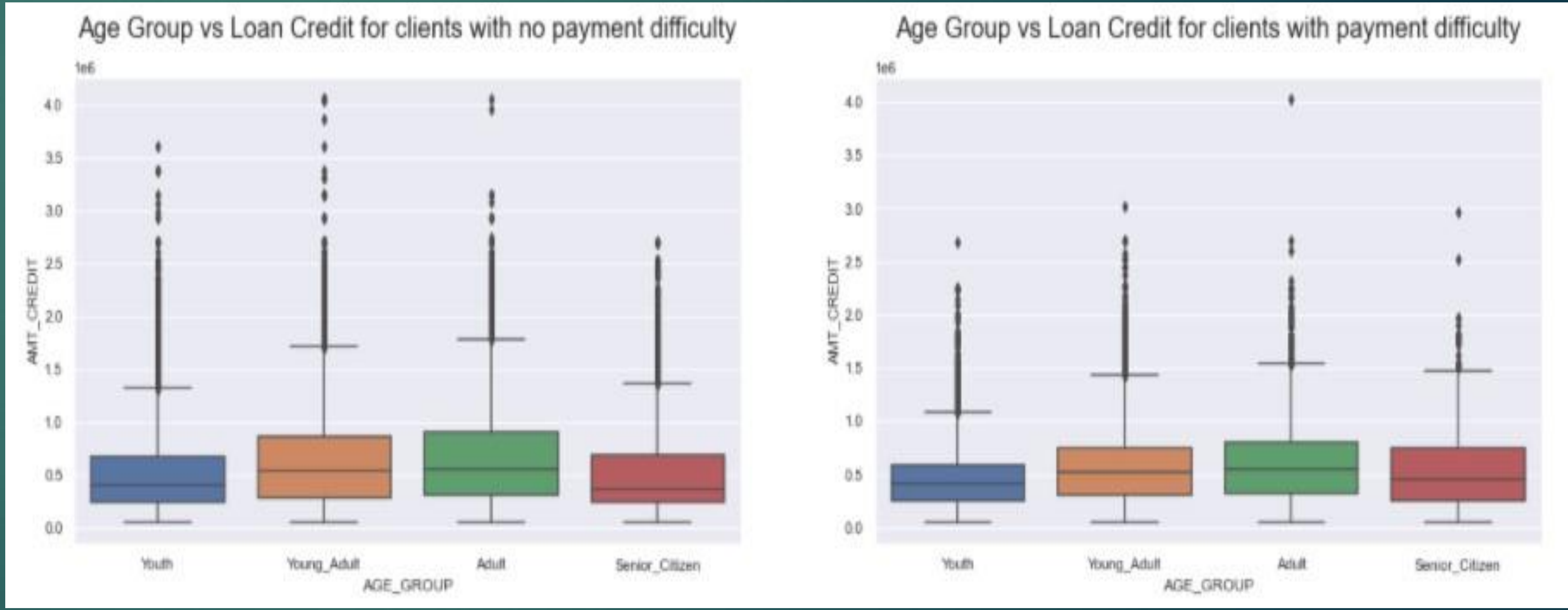
Analysis how Income Category varies with Loan Credit for both target 0 & 1 type clients

- The credit amount increases with each increasing income category.



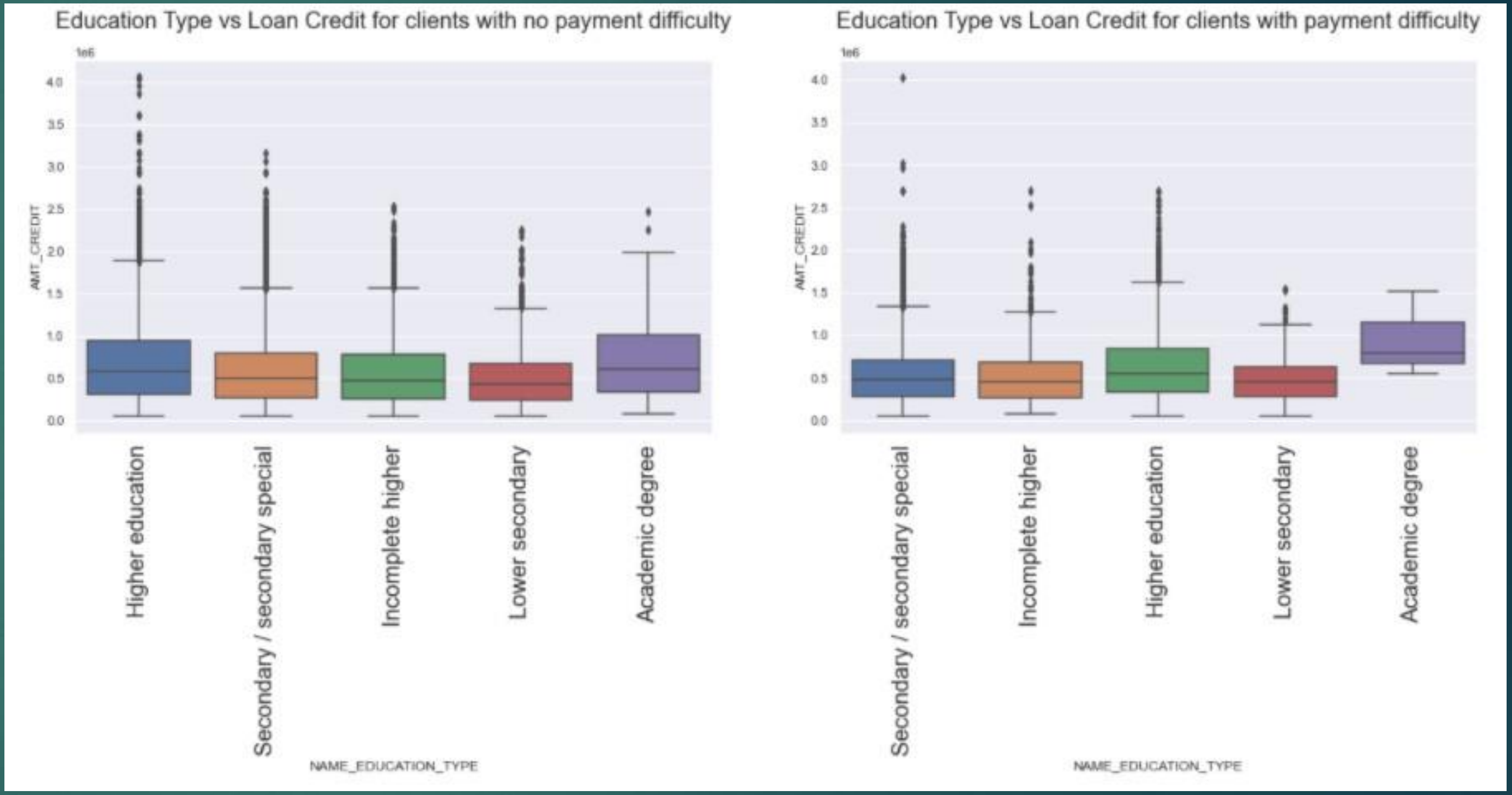
Analysis how Age Group varies with Loan Credit for both target 0 & 1 type clients

- In both cases clients falling under Young_Adult age group & Adult age group applied for or have more Loan Credit Amount



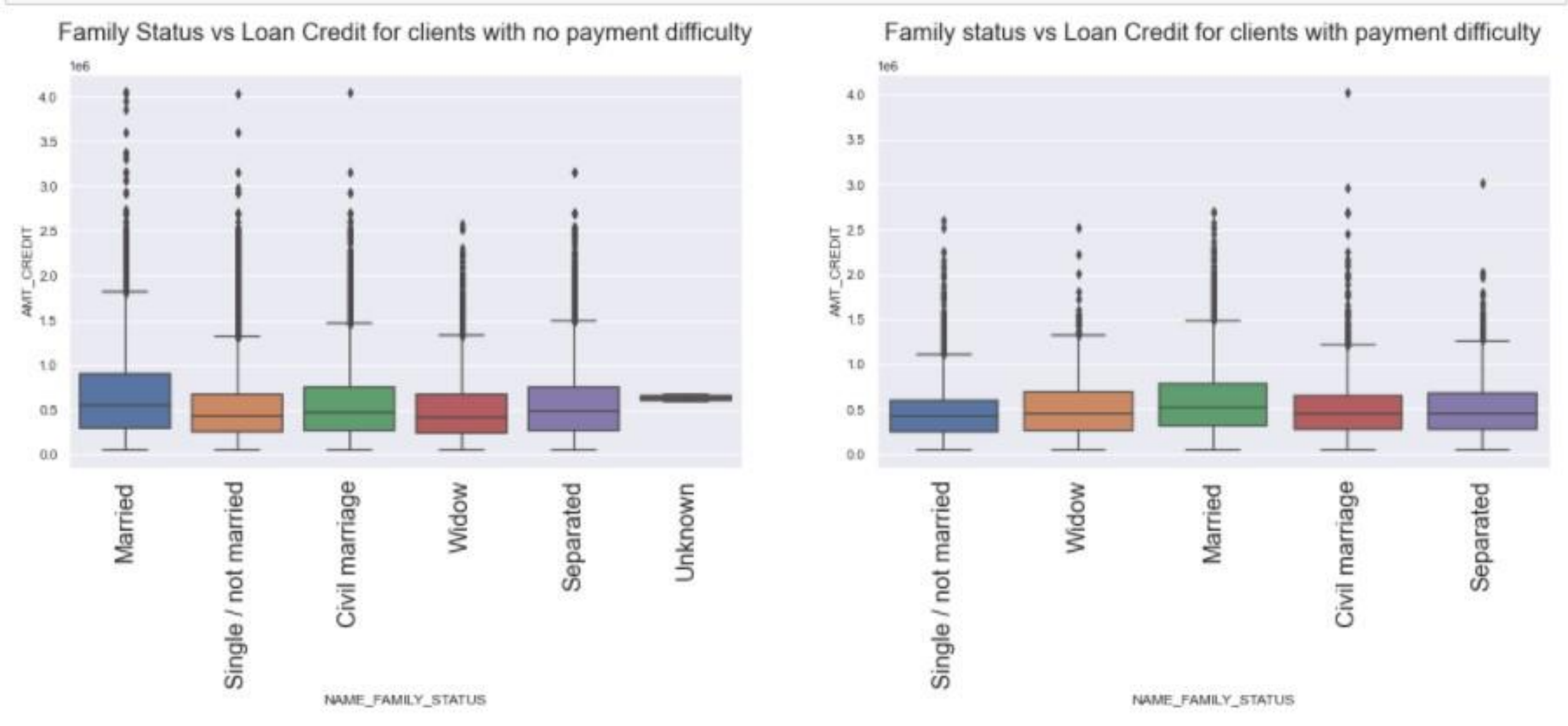
Analysis how Education Type varies with Loan Credit for both target 0 & 1 type clients

- In both the graphs Clients with Academic degree & Higher Education are more likely to take loans
- And also Academic degree category of clients are most likely to default in loans, followed by higher education category.



Analysis how Family Status varies with Loan Credit for both target 0 & 1 type clients

- We can see that Married Clients are more likely to apply for loans and also most likely to default on loans among the rest

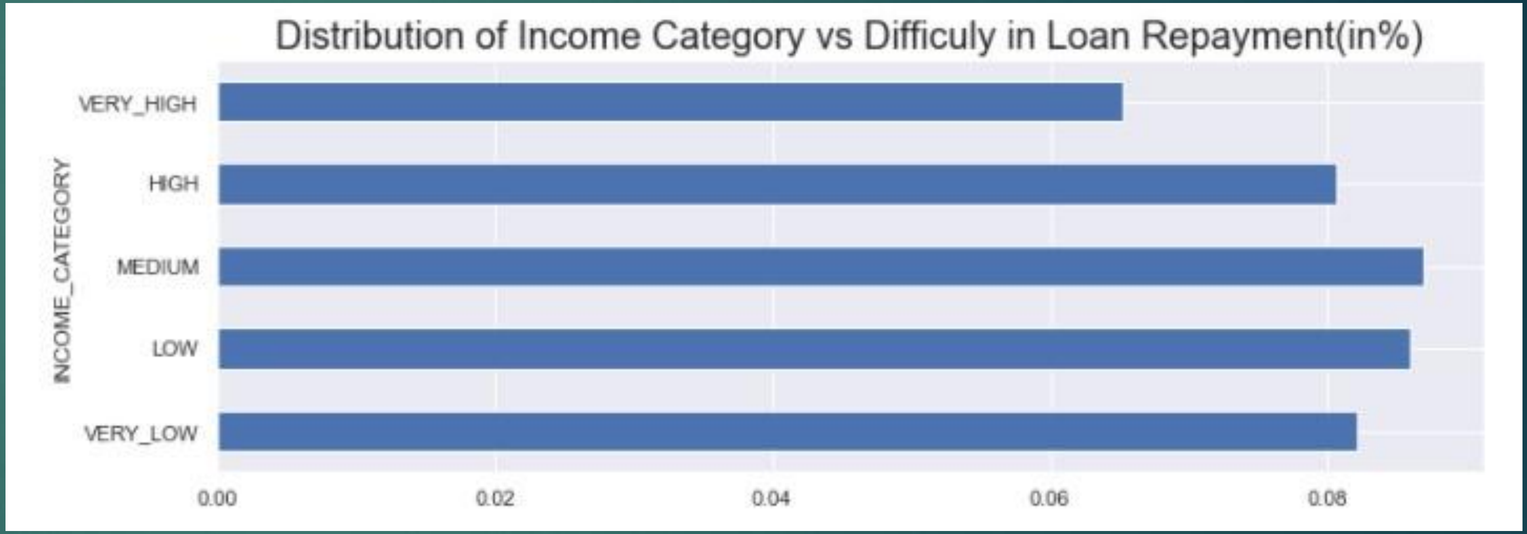


Family Status vs Loan Credit for clients with no payment difficulty

Family status vs Loan Credit for clients with payment difficulty

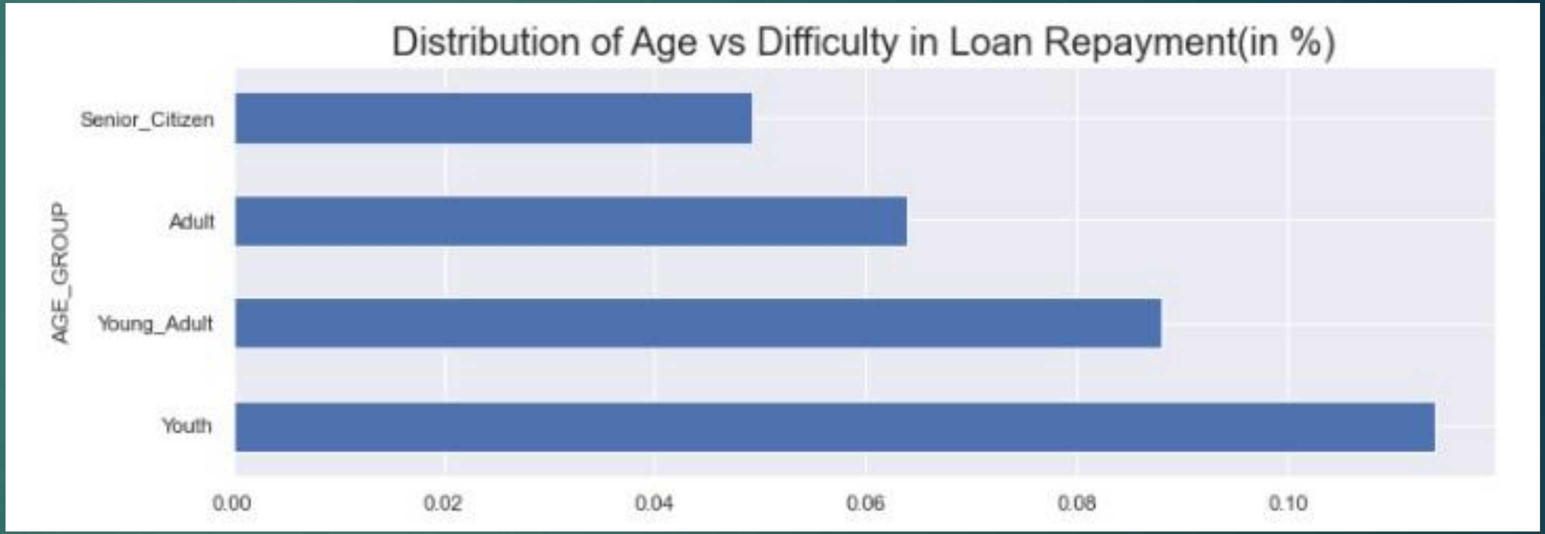
Distribution of Income category with respect to Difficulty in Loan Repayment

- We can observe that Medium Income range clients are more likely to default, followed by Low Income range clients



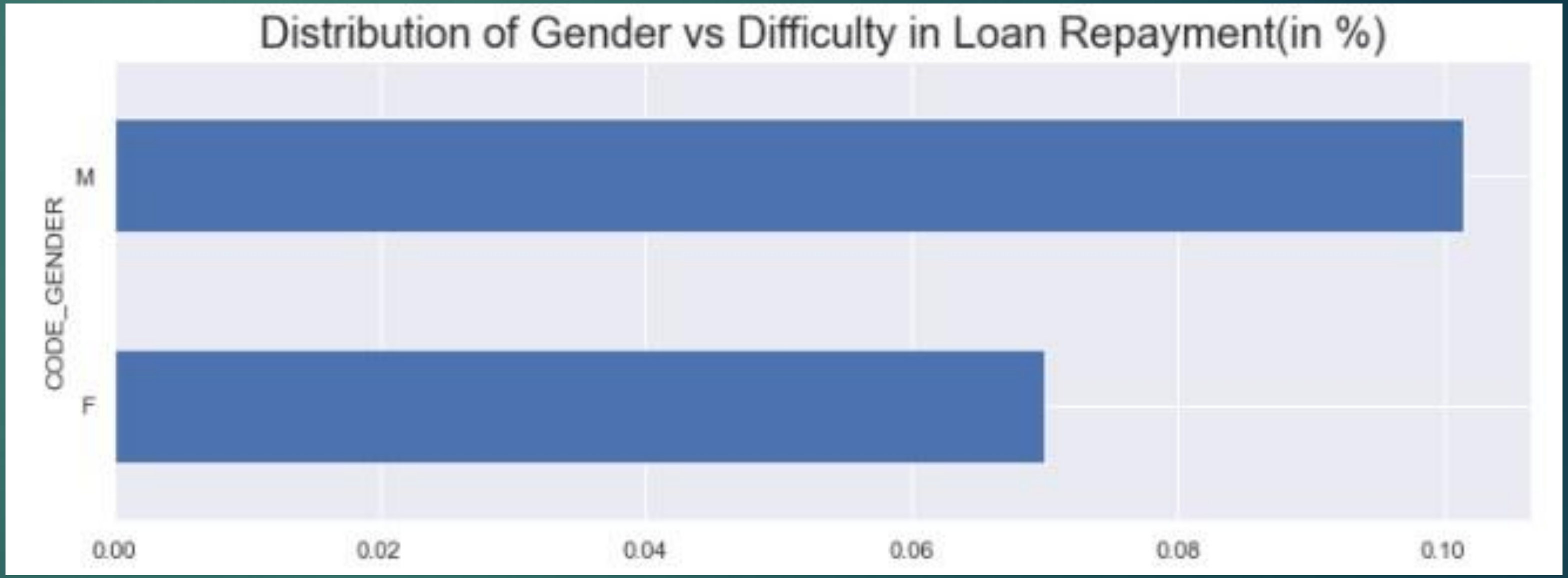
Distribution of Age Group with respect to Difficulty in Loan Repayment

- We observe that clients falling in Youth age category is most likely to default, followed by Young Adults



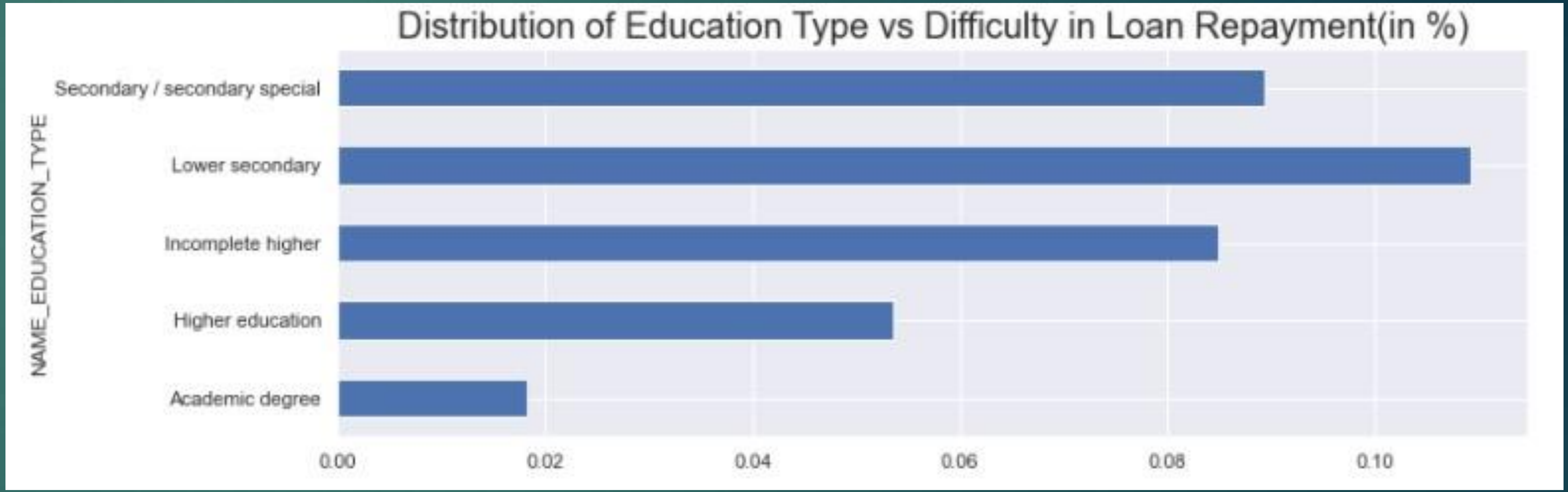
Distribution of Age Group with respect to Difficulty in Loan Repayment

- Males are more likely to default on loans rather than Females



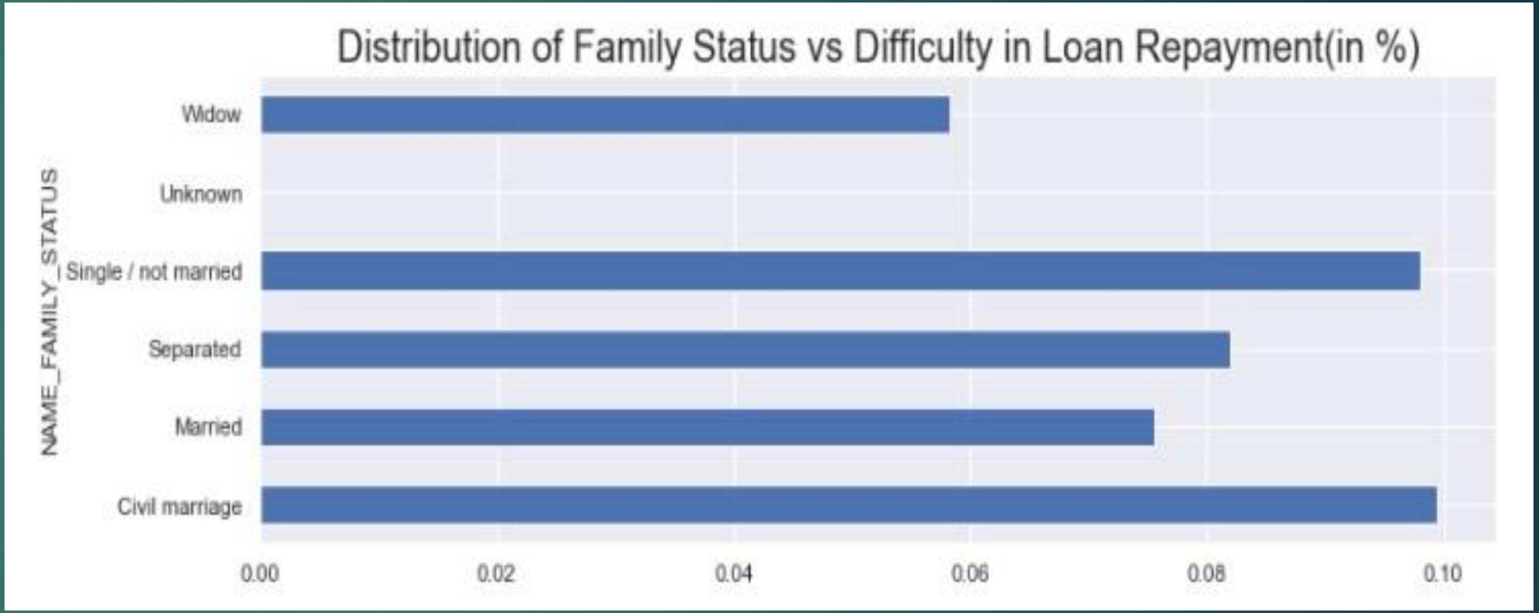
Distribution of Education Type with respect to Difficulty in Loan Repayment

- Clients with Lower Secondary Education Type are most likely to default on loans. followed by Secondary Special



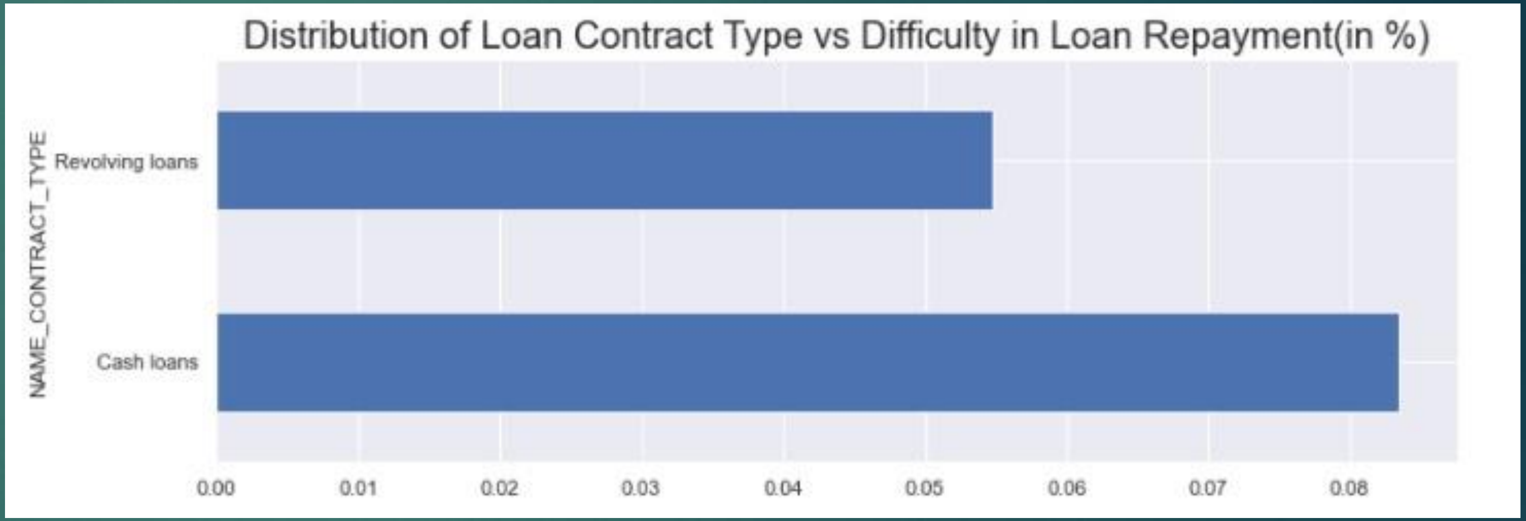
Distribution of Family Status with respect to Difficulty in Loan Repayment

- We can see that clients who had civil marriage or are unmarried are almost equally likely to default on loans, followed by Separated



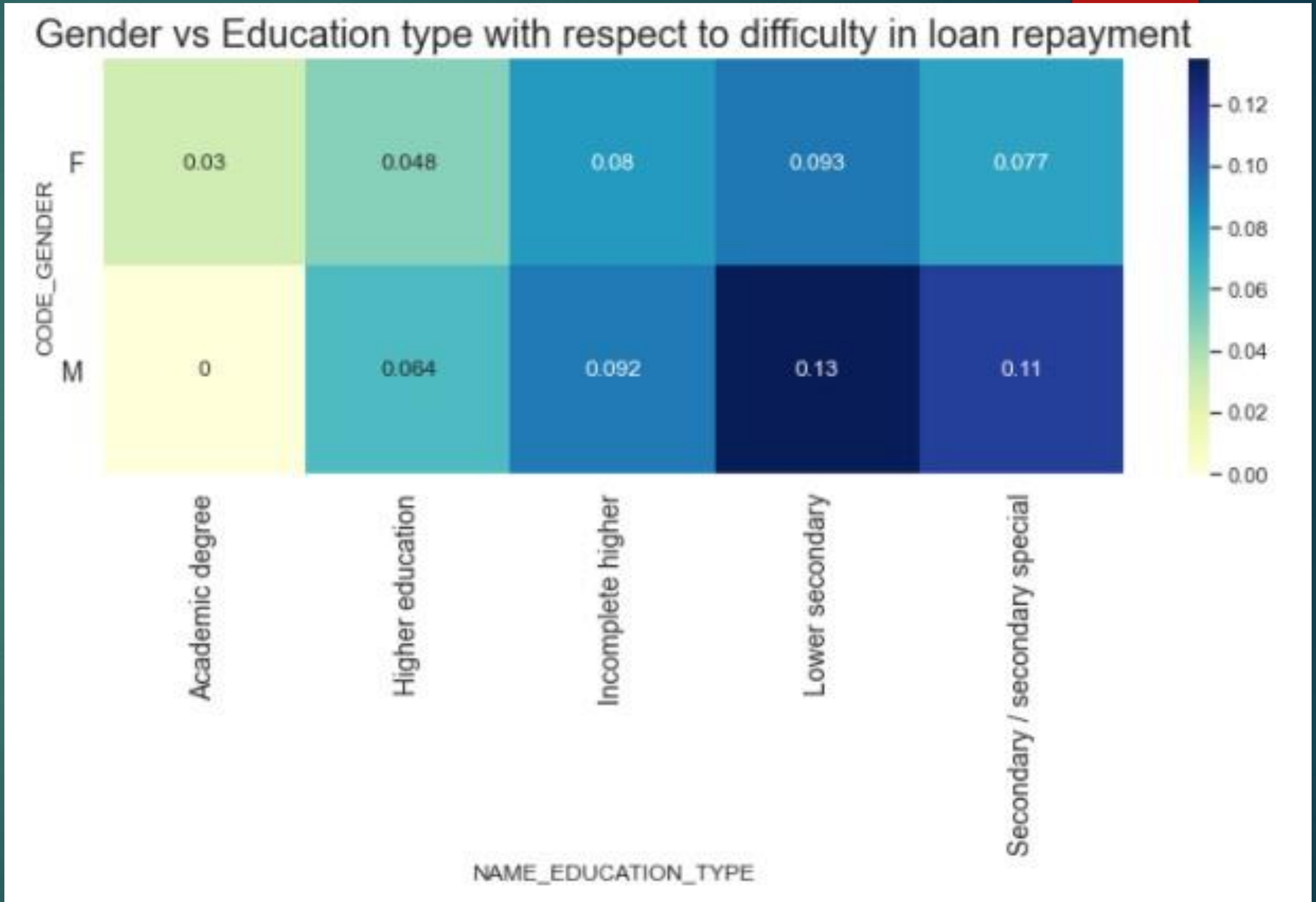
Distribution of Loan Contract Type with respect to Difficulty in Loan Repayment

- Clients are more likely to default on Cash Loans rather than Revolving Loans



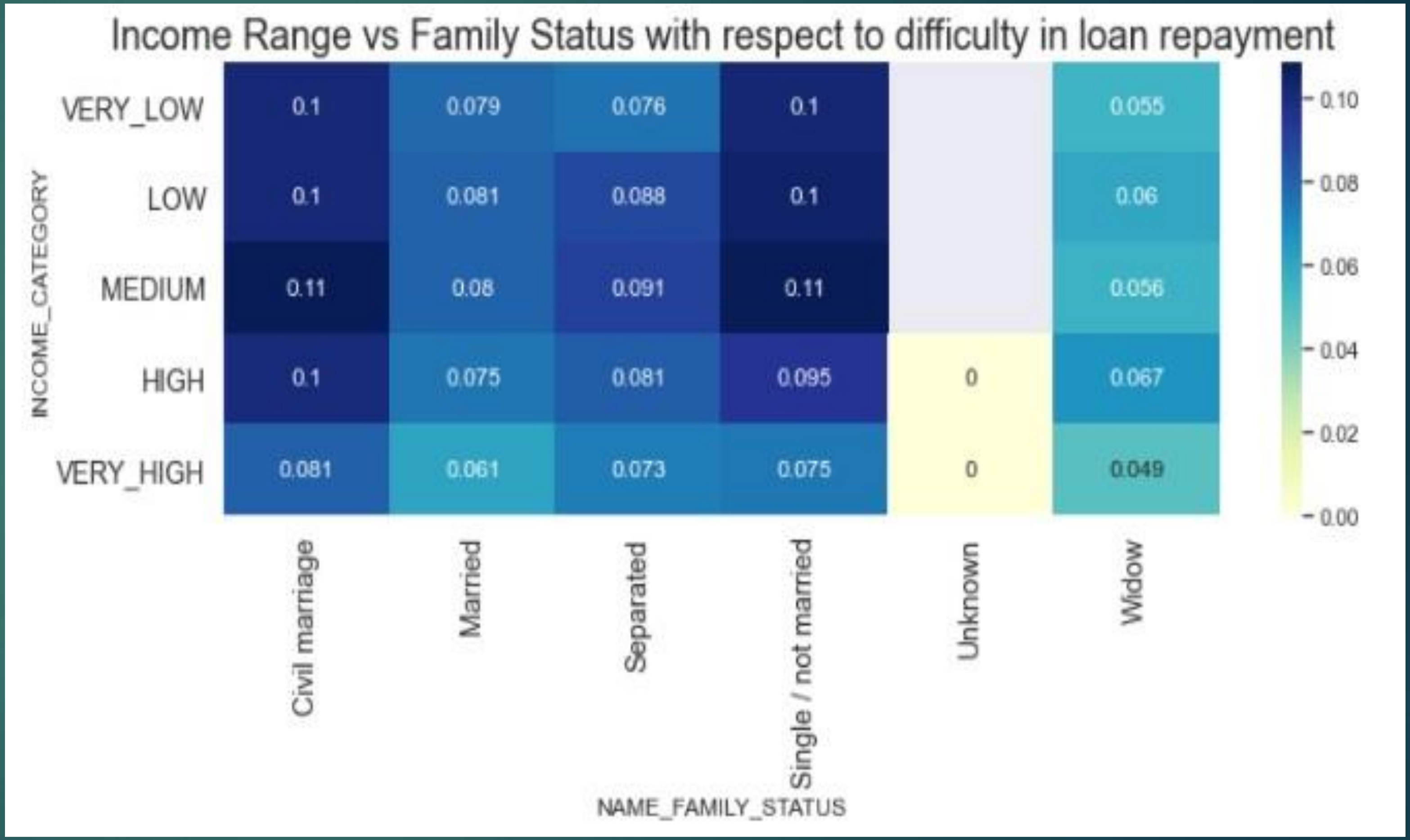
Gender vs Education Type with respect to difficulty in loan repayment

- Both Male & Female Clients having Lower Secondary Education are most likely to default on loans



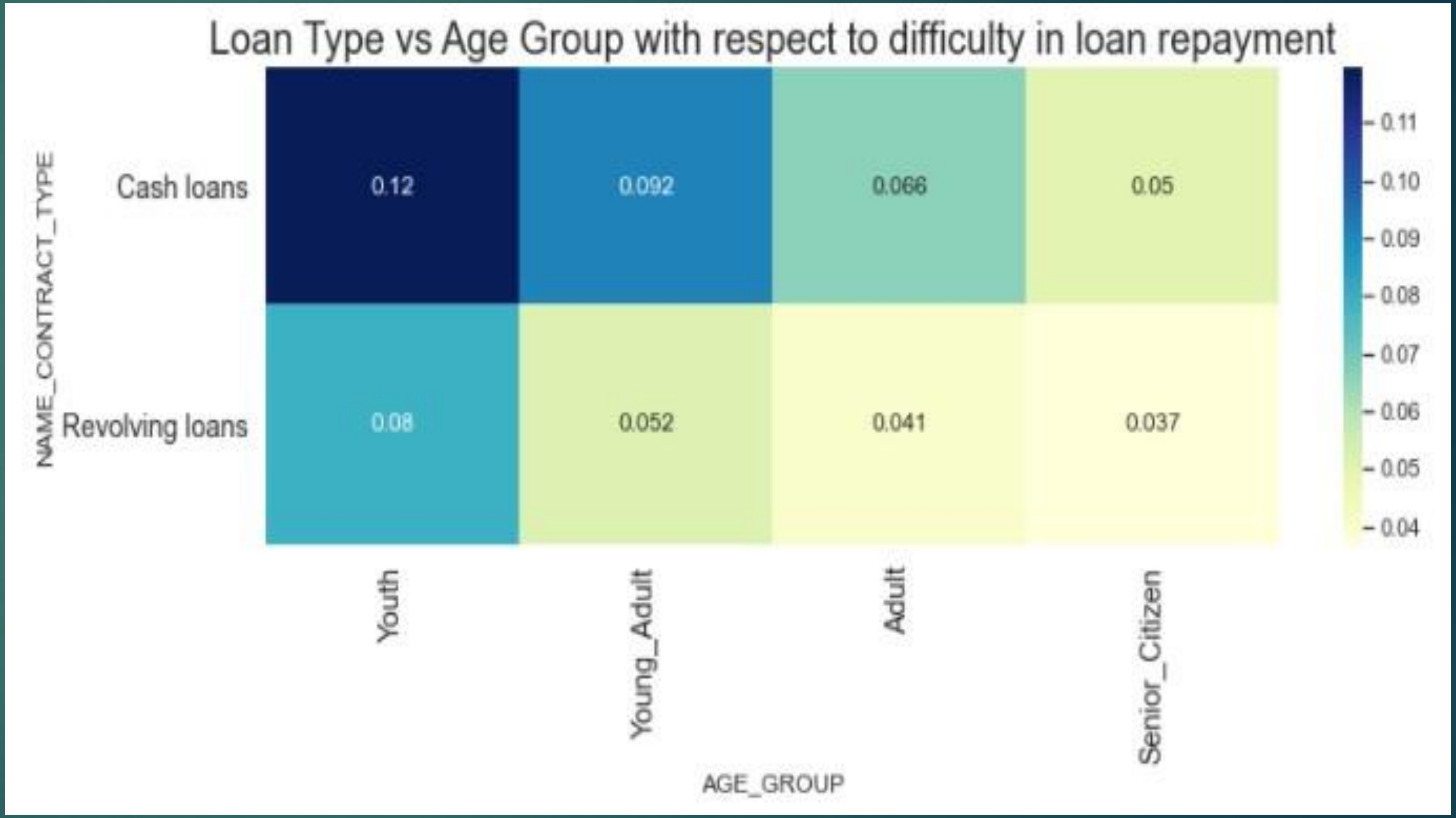
Income Range vs Family Status with respect to difficulty in loan repayment

- Clients with Civil Marriage and Single with Income Category as Medium are most likely to default on loans

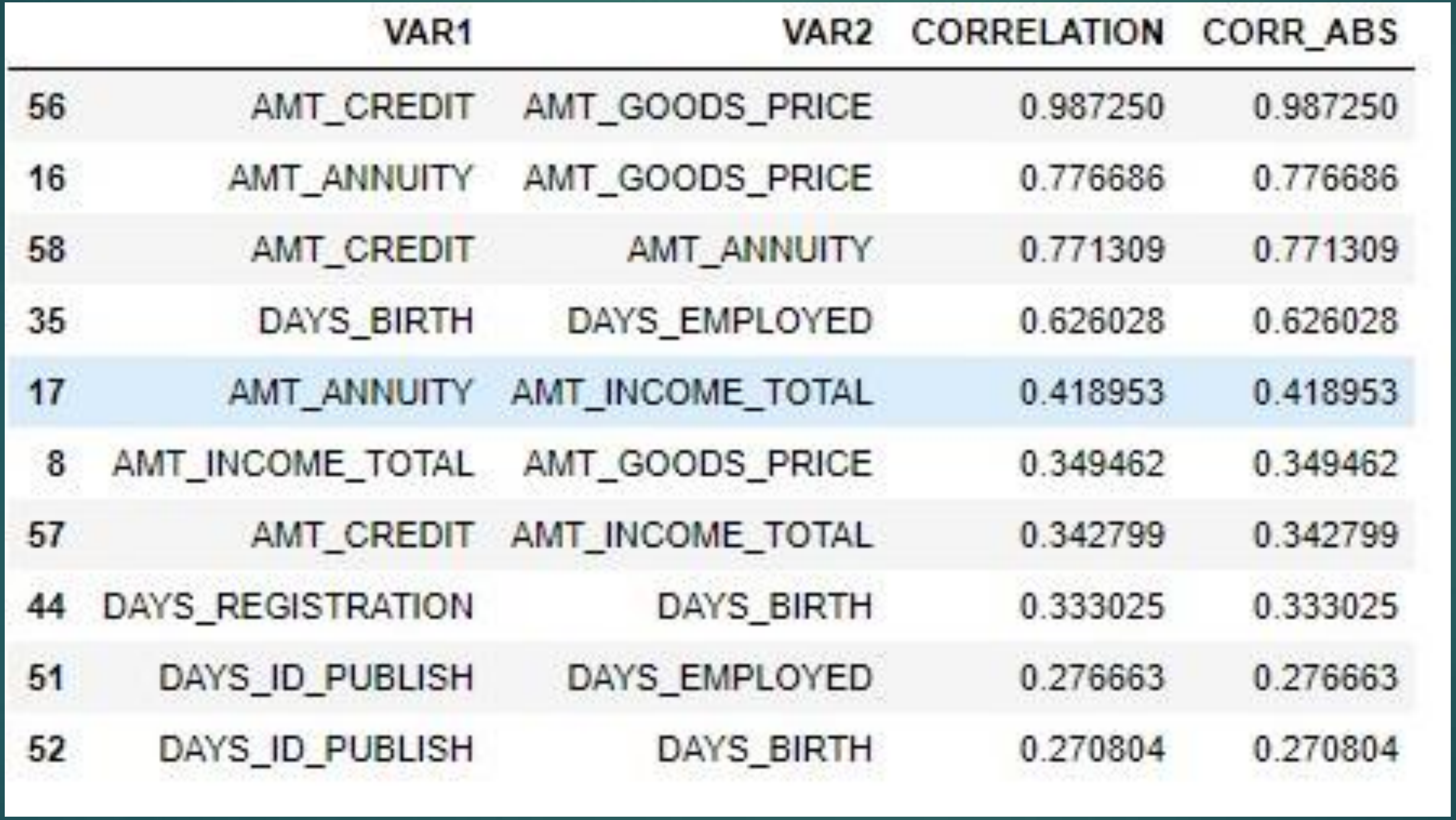


Age group vs Loan Type with respect to difficulty in loan repayment

- Clients falling in Youth Age group category & Opting for Cash Loans are most likely to default

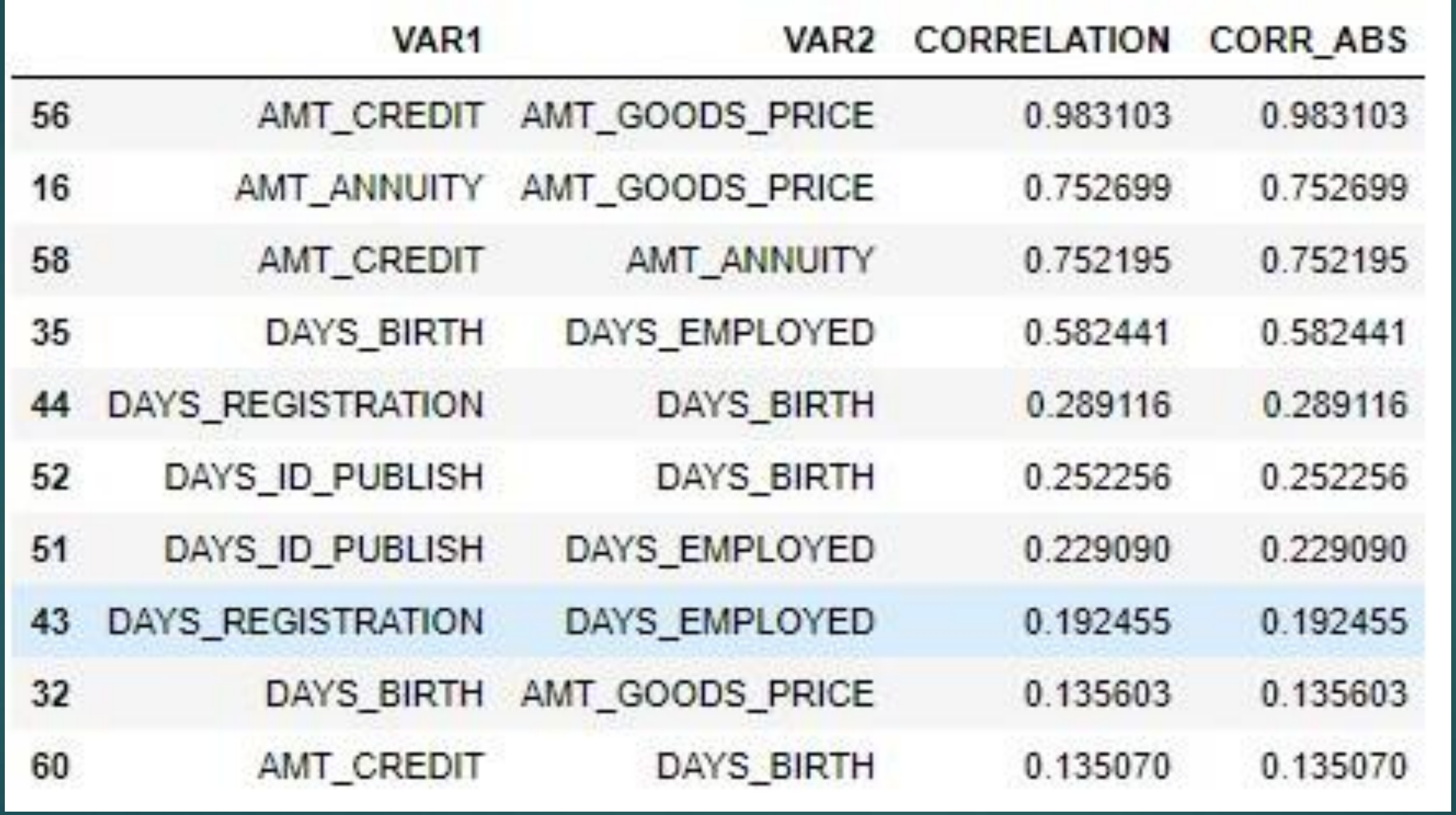


Top 10 Correlation for clients having no repayment difficulties of loan (Target 0)



	VAR1	VAR2	CORRELATION	CORR_ABS
56	AMT_CREDIT	AMT_GOODS_PRICE	0.987250	0.987250
16	AMT_ANNUITY	AMT_GOODS_PRICE	0.776686	0.776686
58	AMT_CREDIT	AMT_ANNUITY	0.771309	0.771309
35	DAYS_BIRTH	DAYS_EMPLOYED	0.626028	0.626028
17	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953	0.418953
8	AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.349462	0.349462
57	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799	0.342799
44	DAYS_REGISTRATION	DAYS_BIRTH	0.333025	0.333025
51	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.276663	0.276663
52	DAYS_ID_PUBLISH	DAYS_BIRTH	0.270804	0.270804

Top 10 Correlation for clients having no repayment difficulties of loan (Target 0)



	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

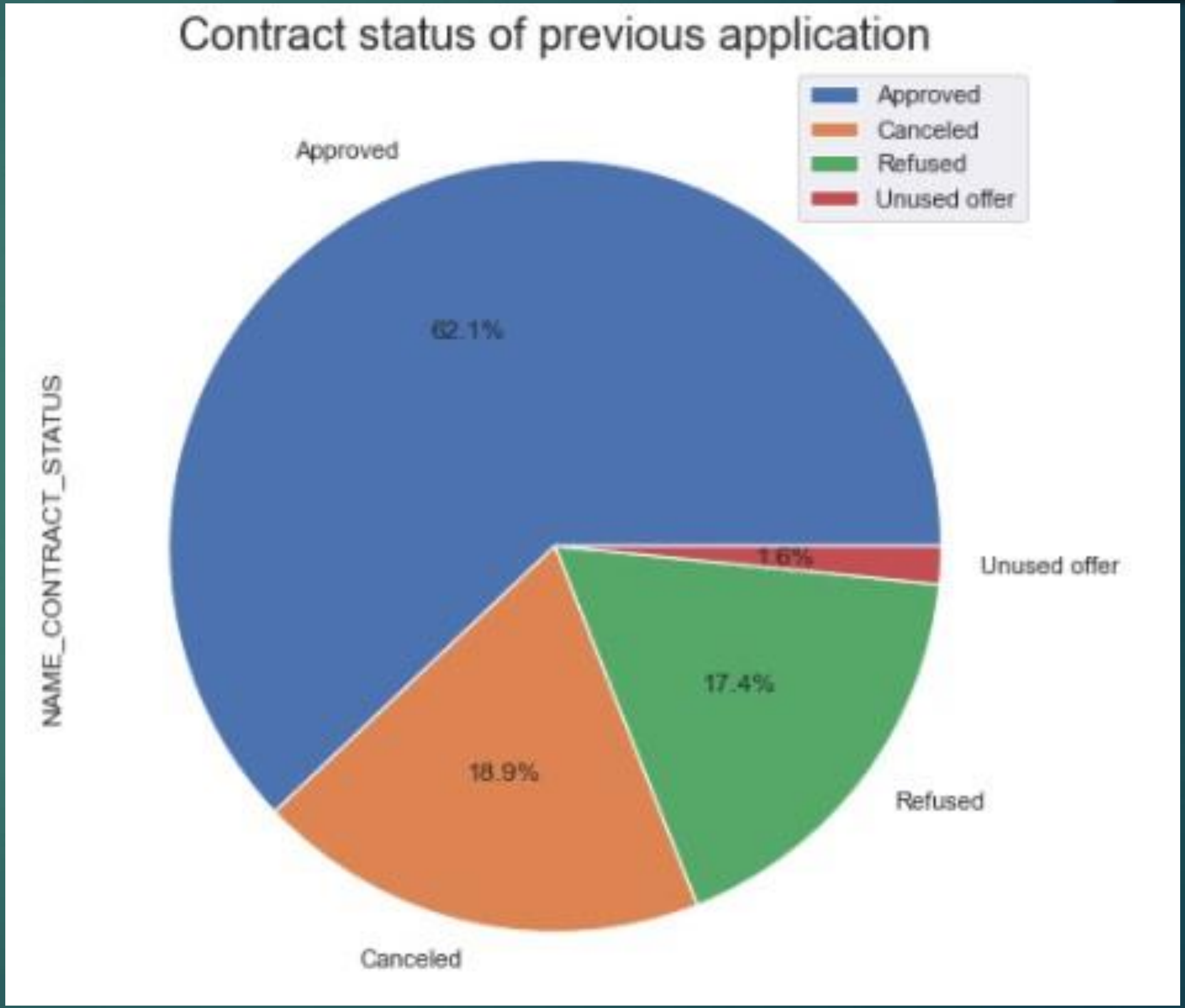
Insights on Top 10 Correlation Pair

- From the above table we can conclude that, for both Target type of clients, the top 4 correlation pair is same.
- 1.AMT_CREDIT & AMT_GOODS_PRICE
- 2.AMT_ANNUITY & AMT_GOODS_PRICE
- 3.AMT_CREDIT & AMT_ANNUITY
- 4.DAYS_BIRTH & DAYS_EMPLOYED

Analysis on Previous Application Data

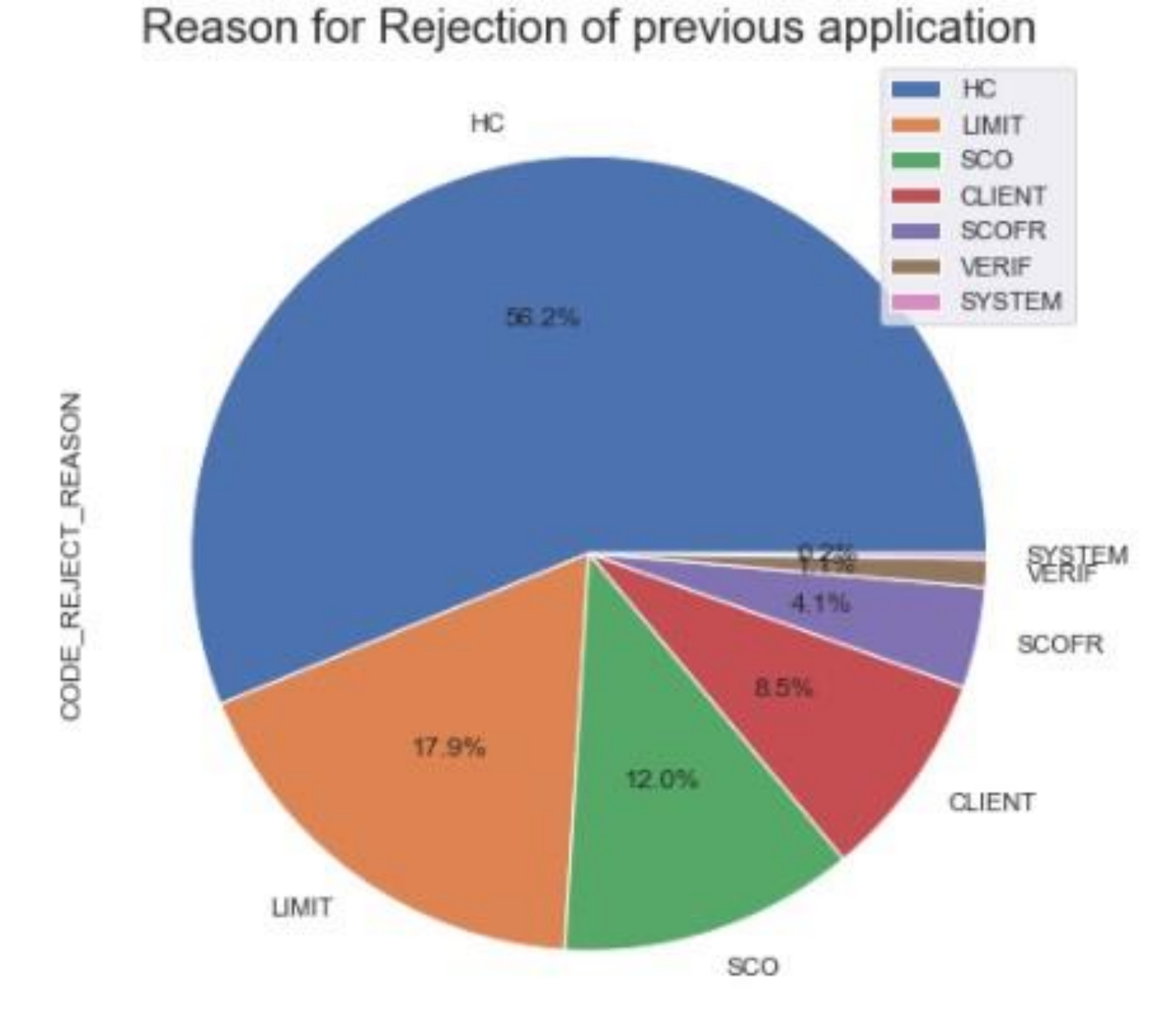
Contract status of previous application

- Majority of the loans were approved & a very less percentage of loans were unused



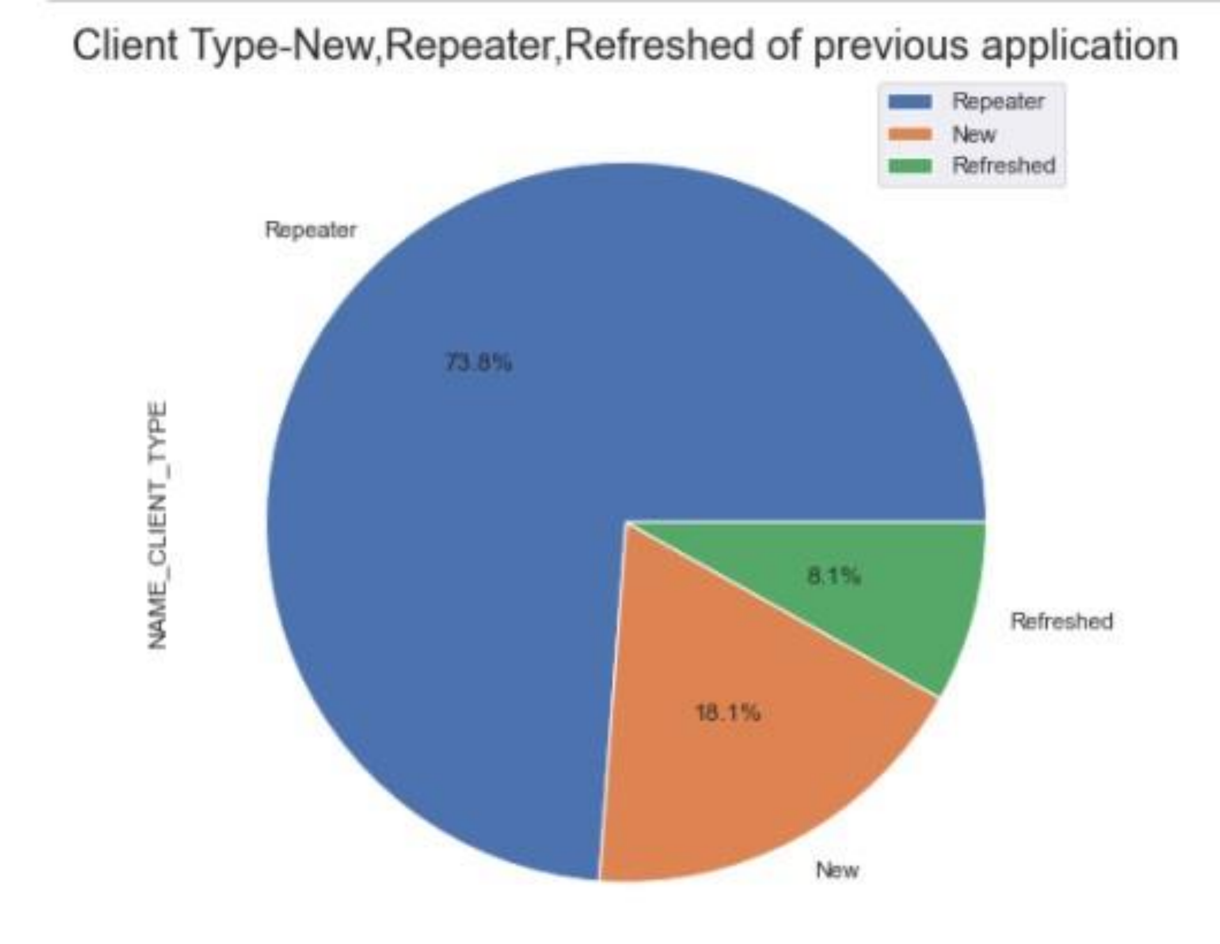
Reason for Rejection of previous application

- HC is the reason for which maximum loans were rejected



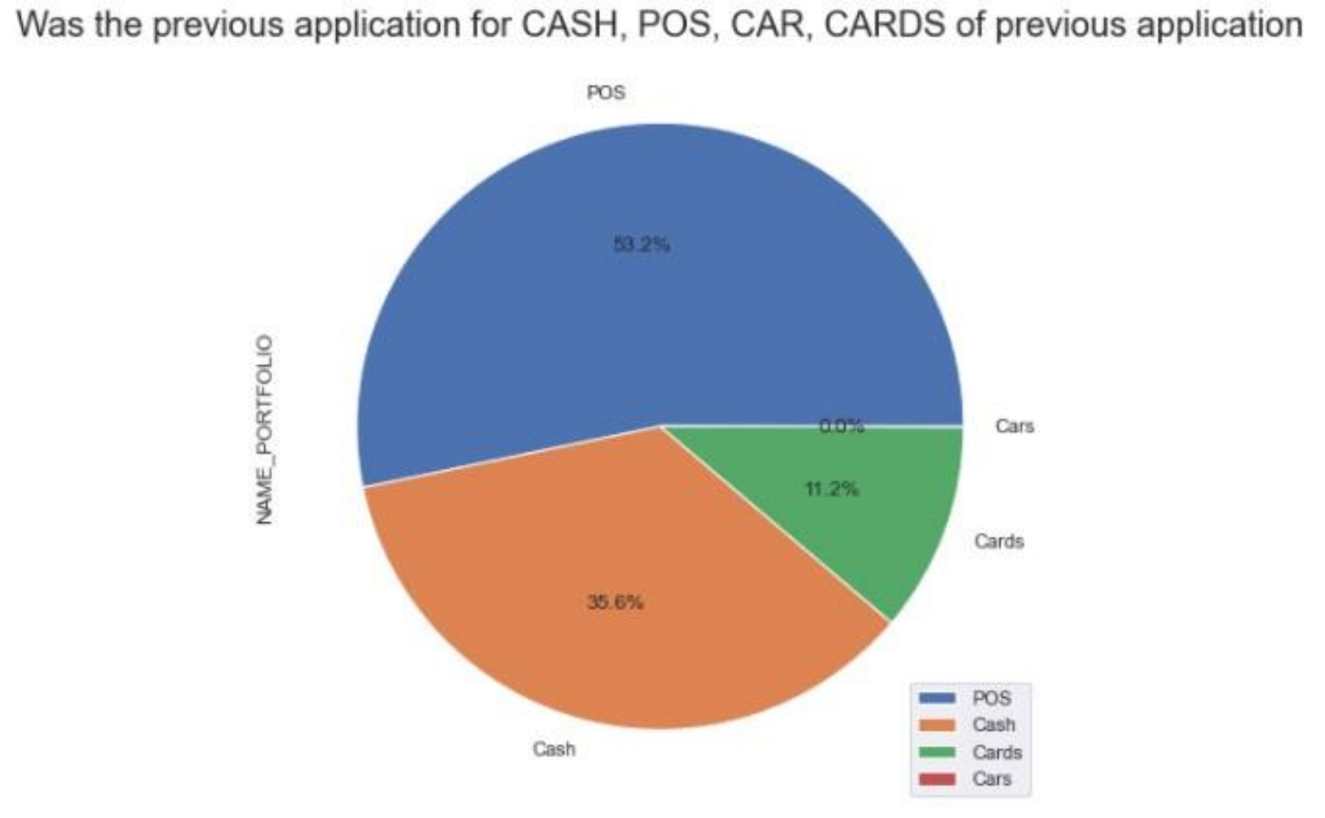
Client Type- New Repeater or Refreshed,of previous application

- Majority of the loan appliers are Repeaters or clients who have taken loan previously



Was the previous application for CASH, POS, CAR, CARDS,of previous application

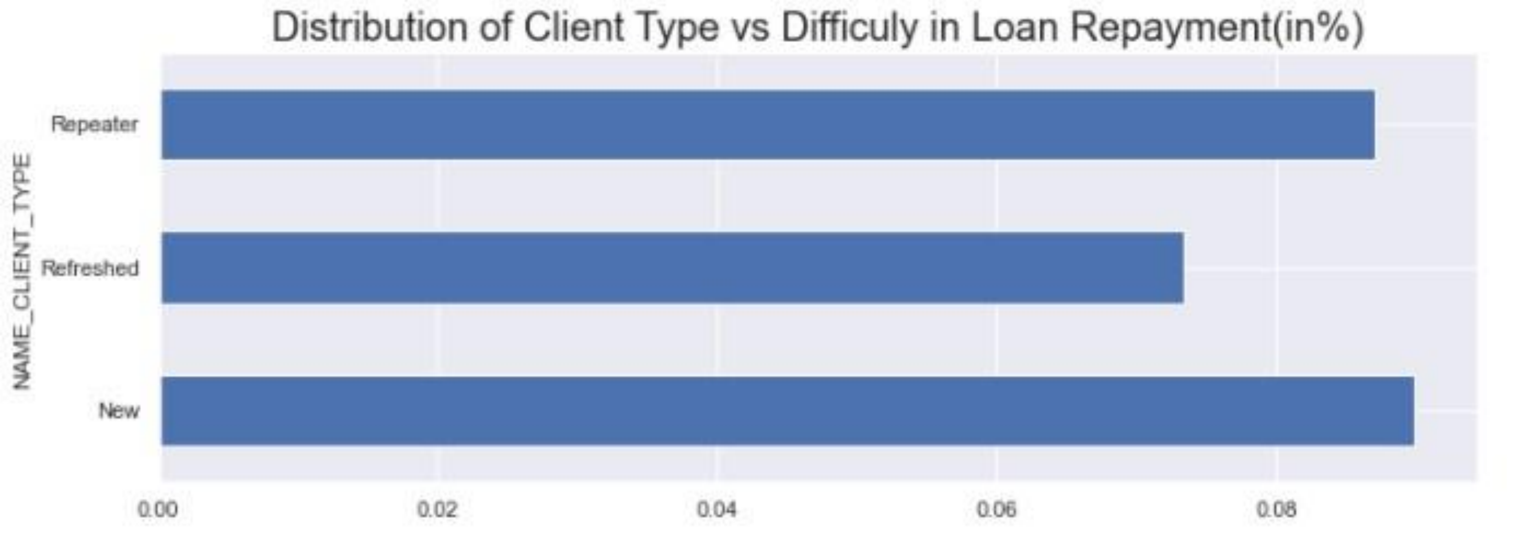
- Majority of the loan application were for POS



Analysis after merging both Application & Previous application

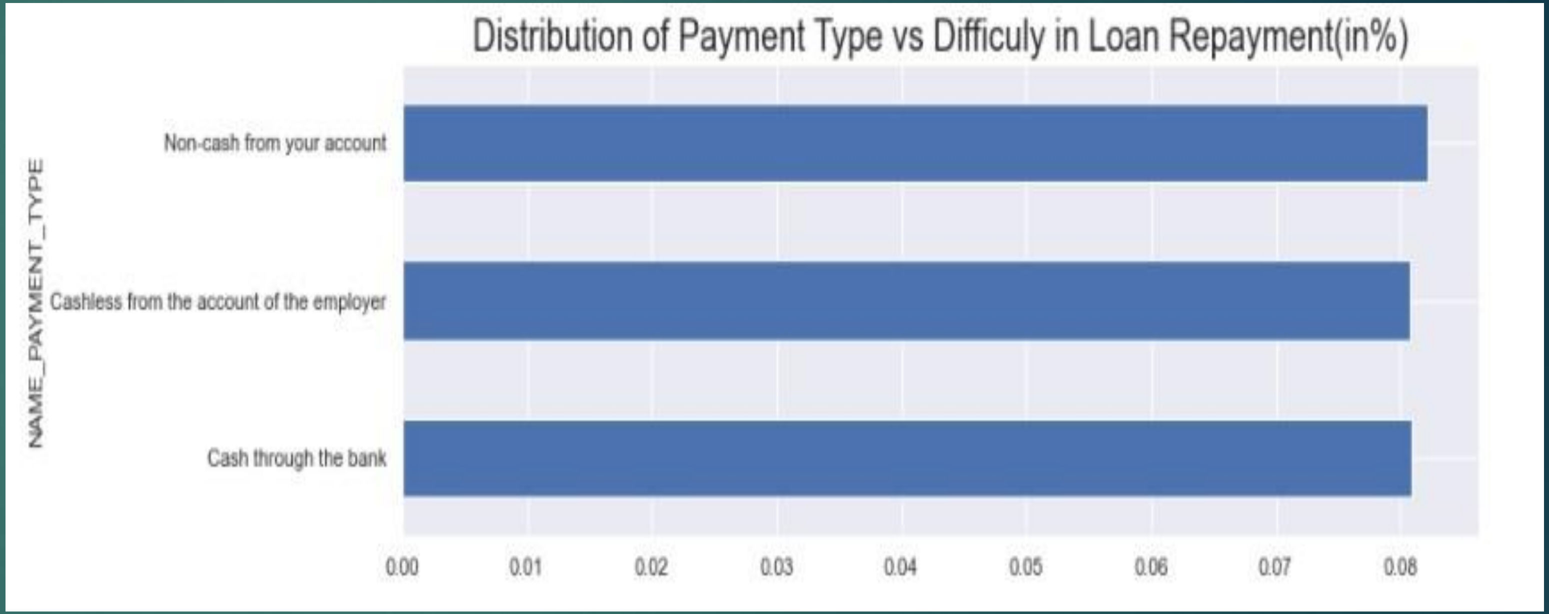
Distribution of Client Type vs Difficulty in Repayment of Loan(in%)

- We can see that clients with no previous loan history have most difficulty in loan repayment



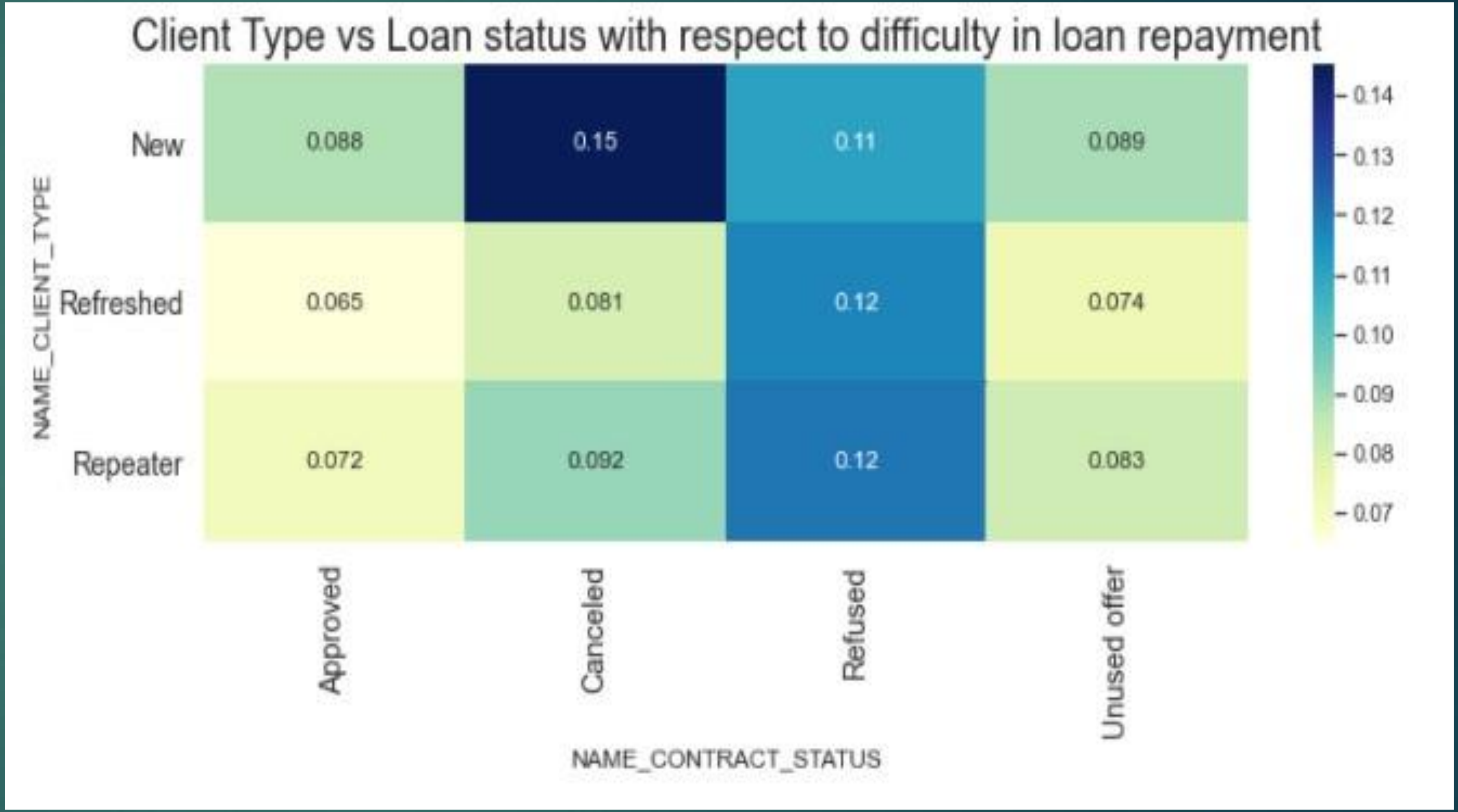
Distribution of Payment Type vs Difficulty in Loan Repayment(in%)

- All payment type perform almost similarly with respect to difficulty in loan repayment



Client Type vs Loan with respect to difficulty in loan repayment

- We can observe that New clients, those with no previous loan history and with previous Loan status cancelled have highest loan repayment difficulty

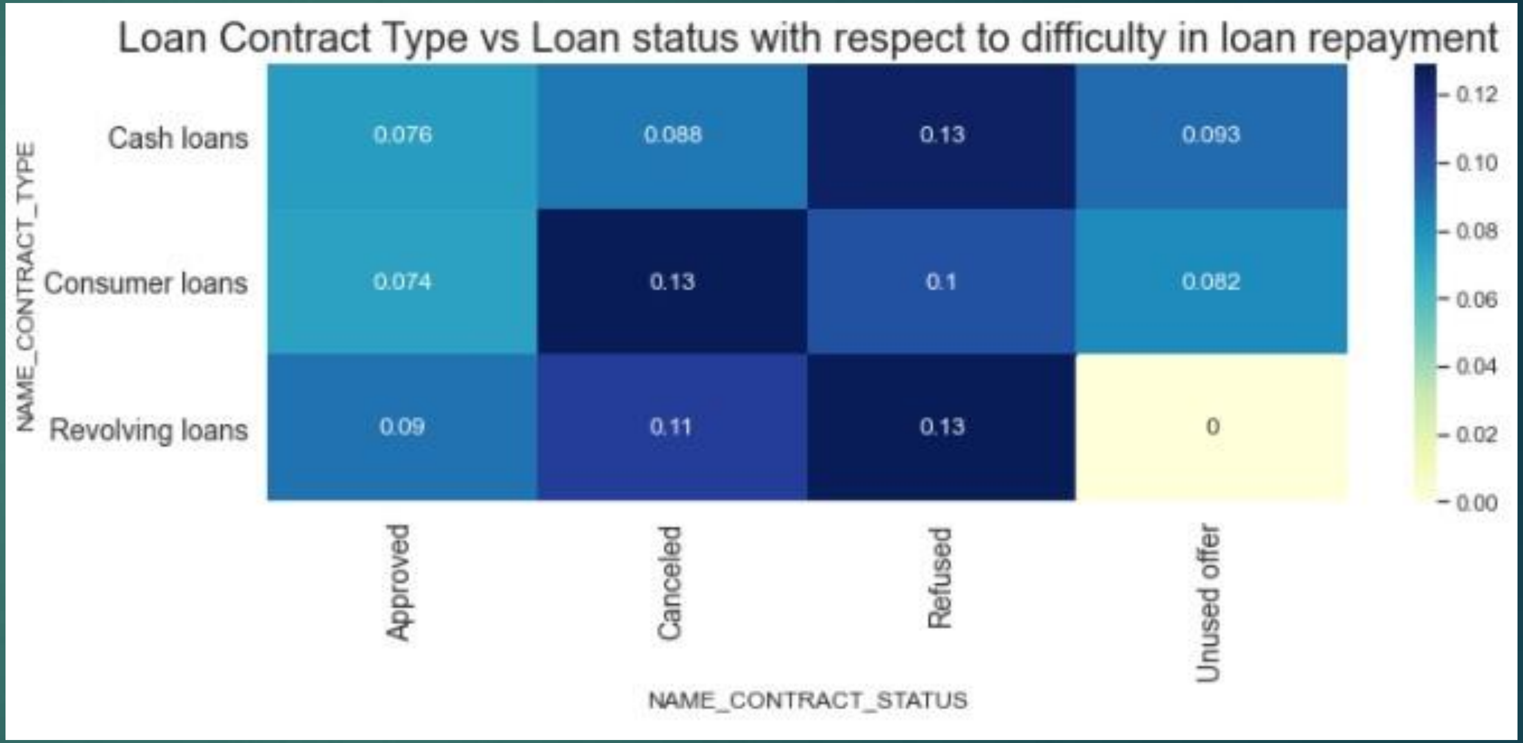


Loan Contract vs Loan with respect to difficulty in loan repayment

- Clients with Cash loans & with previous application status as Refused have most loan repayment difficulties.

And the same goes for-

- Clients with Consumer loans & with previous application status as Cancelled.
- Clients with Revolving loans & with previous application status as Refused.



INSIGHTS- Application Data

- Both Male & Female Clients having Lower Secondary Education are most likely to default on loans
- Clients with Civil Marriage and Single with Income Category as Medium are most likely to default on loans
- Clients falling in Youth Age group category & Opting for Cash Loans are most likely to default
- We can see that Low-Skill Laborers are most likely to default on loans
- Clients are more likely to default on Cash Loans rather than Revolving Loans
- Clients with Lower Secondary Education Type are most likely to default on loans.

INSIGHTS- Previous Application Data

- HC of 'CODE_REJECT_REASON' is the reason for which maximum loans were rejected
- We can observe that New clients, those with no previous loan history and with previous Loan status cancelled have highest loan repayment difficulty

Other Insights

- We also observe outliers in the Credit Amount, which bank should be careful about otherwise it might result of Credit Loss to the bank.