CREDIT EDA CASE STUDY

Purpose of the case study

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:

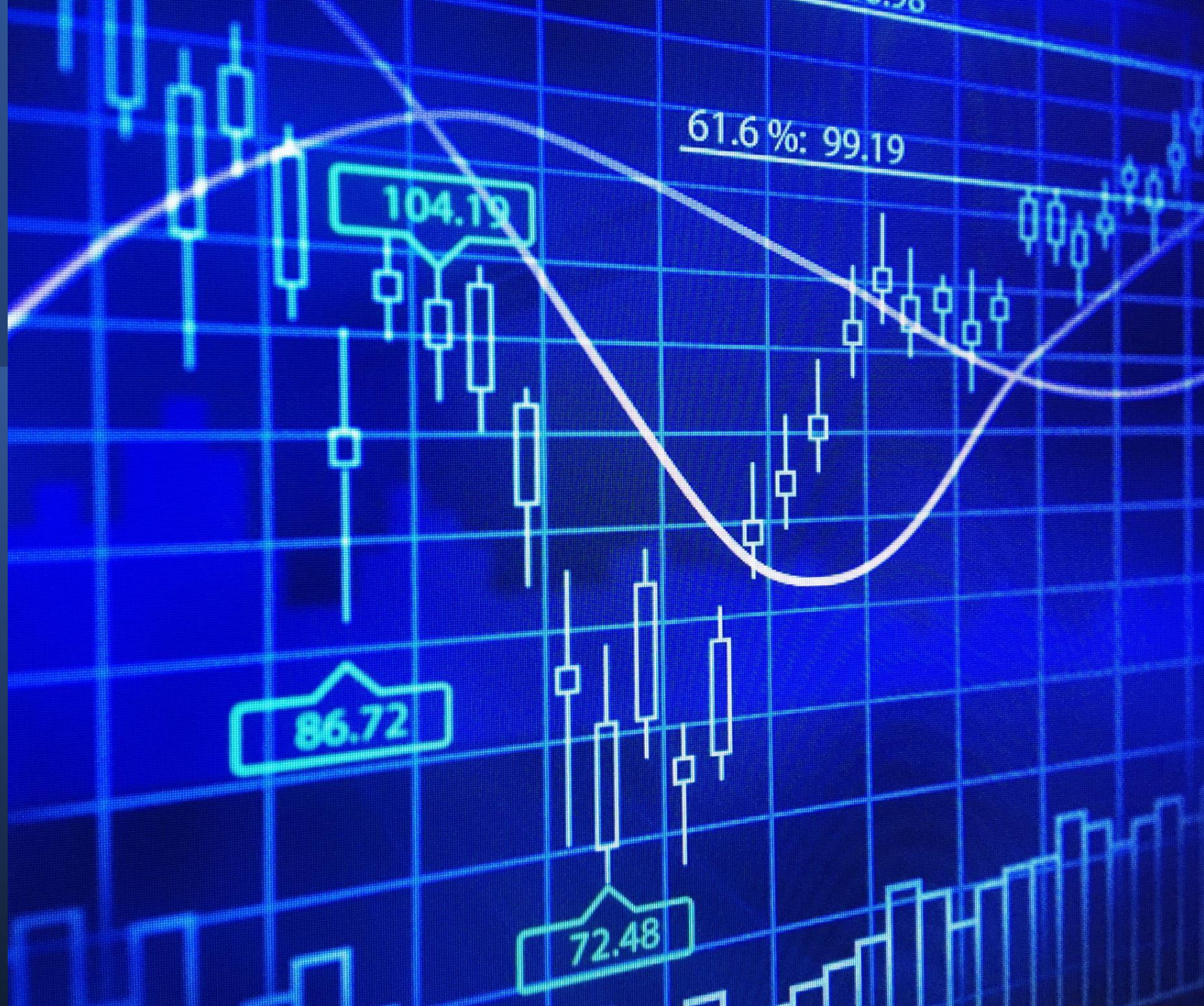
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

This EDA Analysis helps identify the patterns present in the applicant's data and their previous loan history.

- **Analysis 1:** Aims to identify patterns that 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.

- **Analysis 2:** Identify applicants capable of repaying the loan are not rejected (using the previous application data)

Data Handling



Null Value Handling

- Columns with more than 40% null values are dropped after ensuring there are no columns that would impact the inferences if dropped.

- Null value handling for remaining columns is summarized below.

AMT_ANNUITY	0.003902 %	Impute the missing values with the median based on the skewness of the numerical variable
AMT_GOODS_PRICE	0.090403 %	Impute the missing values with the median based on the skewness of the numerical variable
CNT_FAM_MEMBERS	0.000650 %	Impute the missing values with the median based on the skewness of the numerical variable
DAYS_LAST_PHONE_CHANGE	0.000325 %	Impute the missing values with the median based on the skewness of the numerical variable
NAME_TYPE_SUITE	0.420148 %	Impute the missing values with the mode since it is categorical variable
AMT_REQ_CREDIT_BUREAU_HOUR	13.501631 %	Impute the missing values with the mode since it is categorical variable
AMT_REQ_CREDIT_BUREAU_DAY	13.501631 %	Impute the missing values with the mode since it is categorical variable
AMT_REQ_CREDIT_BUREAU_WEEK	13.501631 %	Impute the missing values with the mode since it is categorical variable
AMT_REQ_CREDIT_BUREAU_MON	13.501631 %	Impute the missing values with the mode since it is categorical variable
AMT_REQ_CREDIT_BUREAU_QRT	13.501631 %	Impute the missing values with the mode since it is categorical variable
AMT_REQ_CREDIT_BUREAU_YEAR	13.501631 %	Impute the missing values with the mode since it is categorical variable

Binning & Outlier Detection

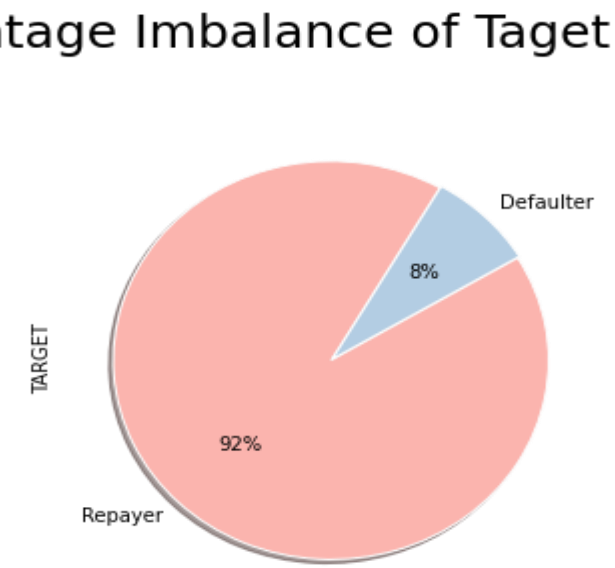
Binning

- Changed the number of days into years.
- New column AGE_GROUP is created which contains bins with labels '<30','30-40','40-50','50-60','60+' for the values in DAYS_BIRTH column.
- New column AMT_INCOME_GROUP is created which contains bins labeled 'VERY_LOW', 'LOW', "MEDIUM", 'HIGH', 'VERY_HIGH' for AMT_INCOME_TOTAL divided in percentiles of [0, 0.2, 0.5, 0.7, 0.9, 1]

Outlier Detection (using Boxplots)

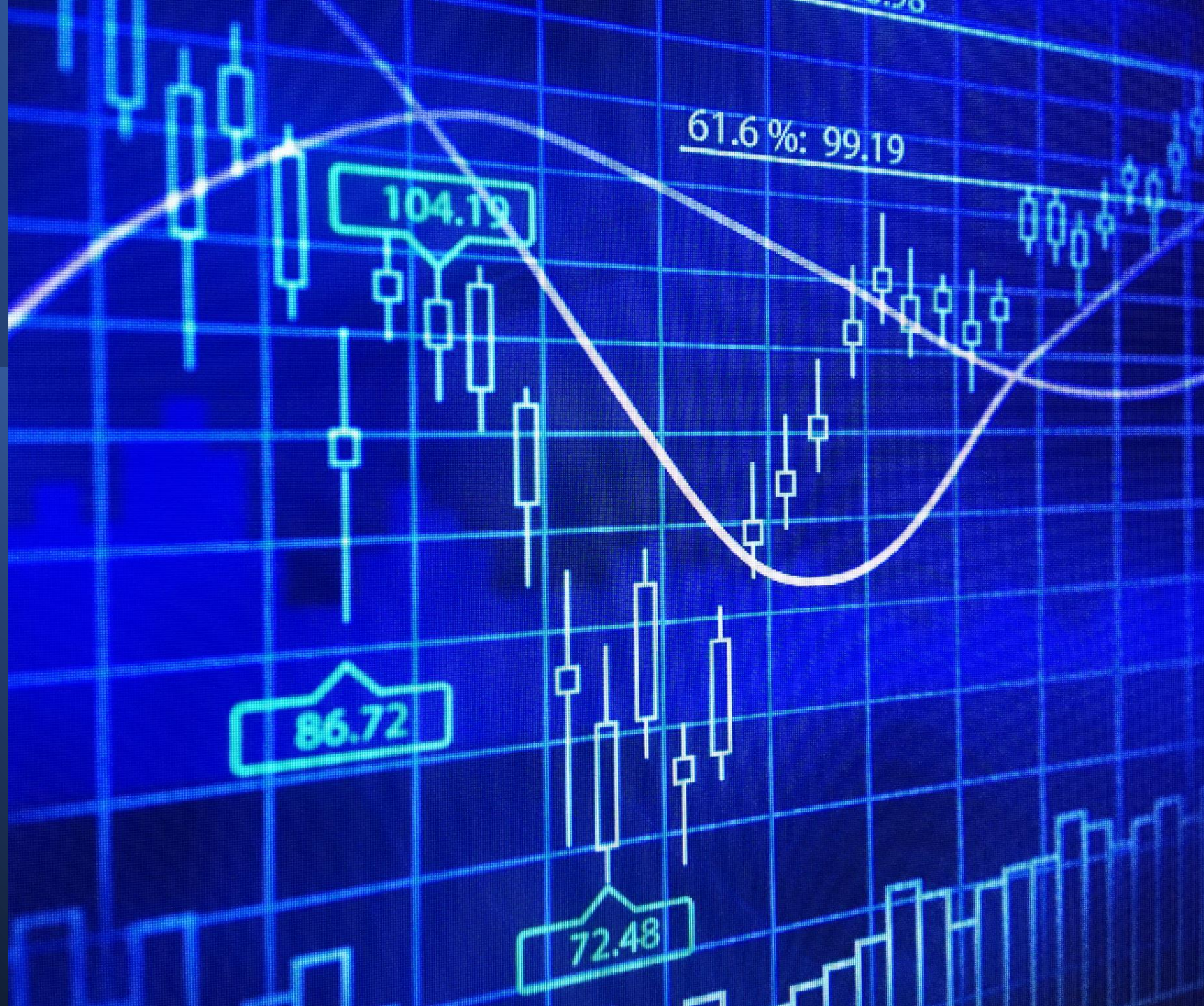
- An outlier was identified with AMT_INCOME_TOTAL - 117000000.0, OCCUPATION_TYPE – Laborers and is in the AGE_GROUP - 30-40.
- An outlier was identified with AMT_ANNUITY=258025.5 which was significantly standing out from the rest of the values in the boxplot.
- An outlier was identified with AMT_GOODS_PRICE whose maximum value was 4050000.0
- An outlier was identified whose DAYS_EMPLOYED was approximately 1000 Years which is unrealistic.

Few important inferences drawn from Univariate Analysis

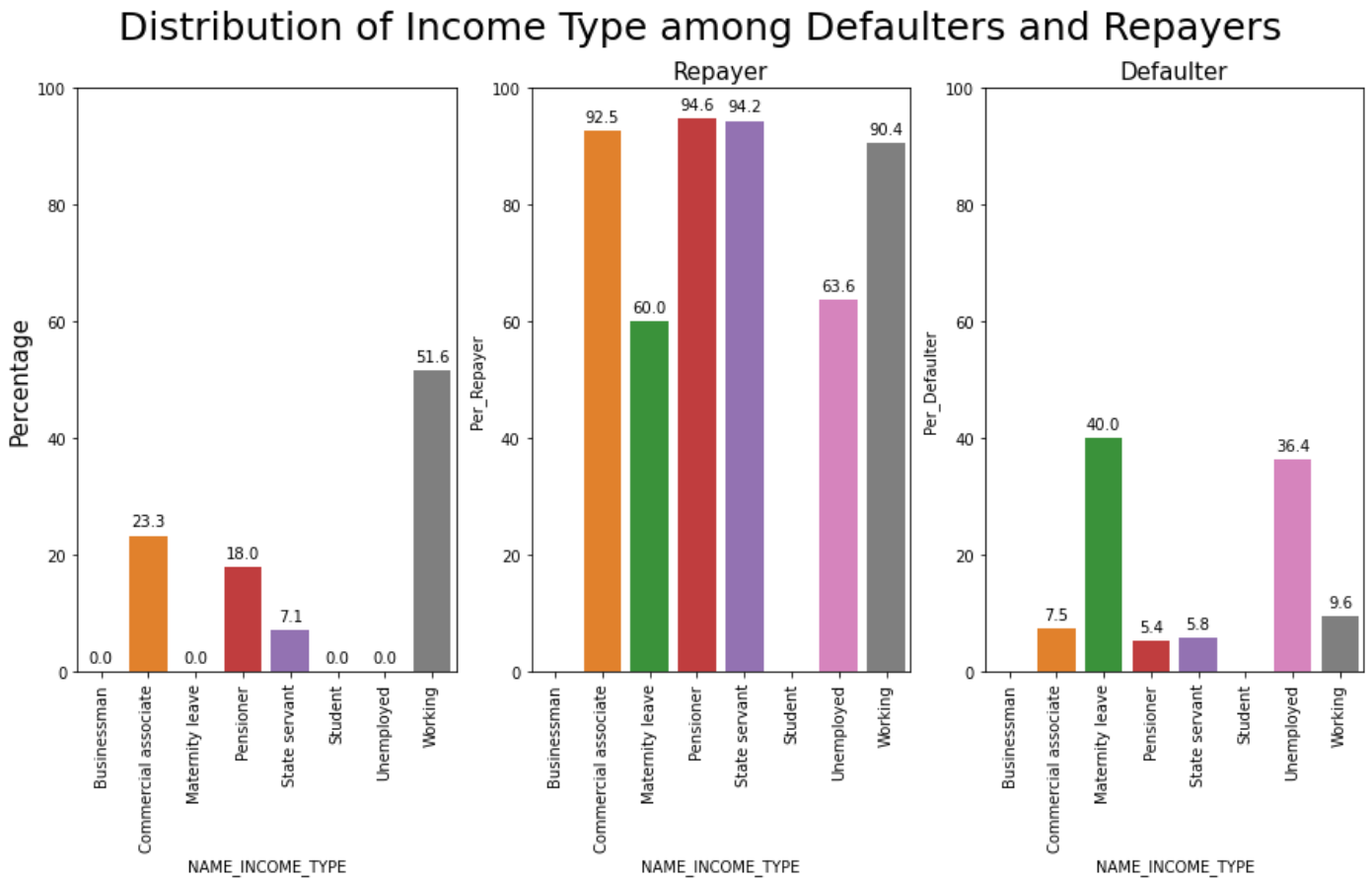


- As the goods price increases, there is decrease in defaulters compared to repayers.
- There are more defaulters around age 30.
- Defaulters are slightly higher over repayers between registration years 0-20. But the trend reverses after 20.
- Majority of defaulters and repayers are from working class.
- Percentage of pensioners are more among repayers than defaulters.
- Percentage of people with secondary education are more defaulters.
- Percentage of people with higher education are more among repayers than defaulters.
- Percentage of married people are more repayers.
- Percentage of single and civil marriage people are more among defaulters than repayers.
- Percentage of people living in house/apartment are more repayers.

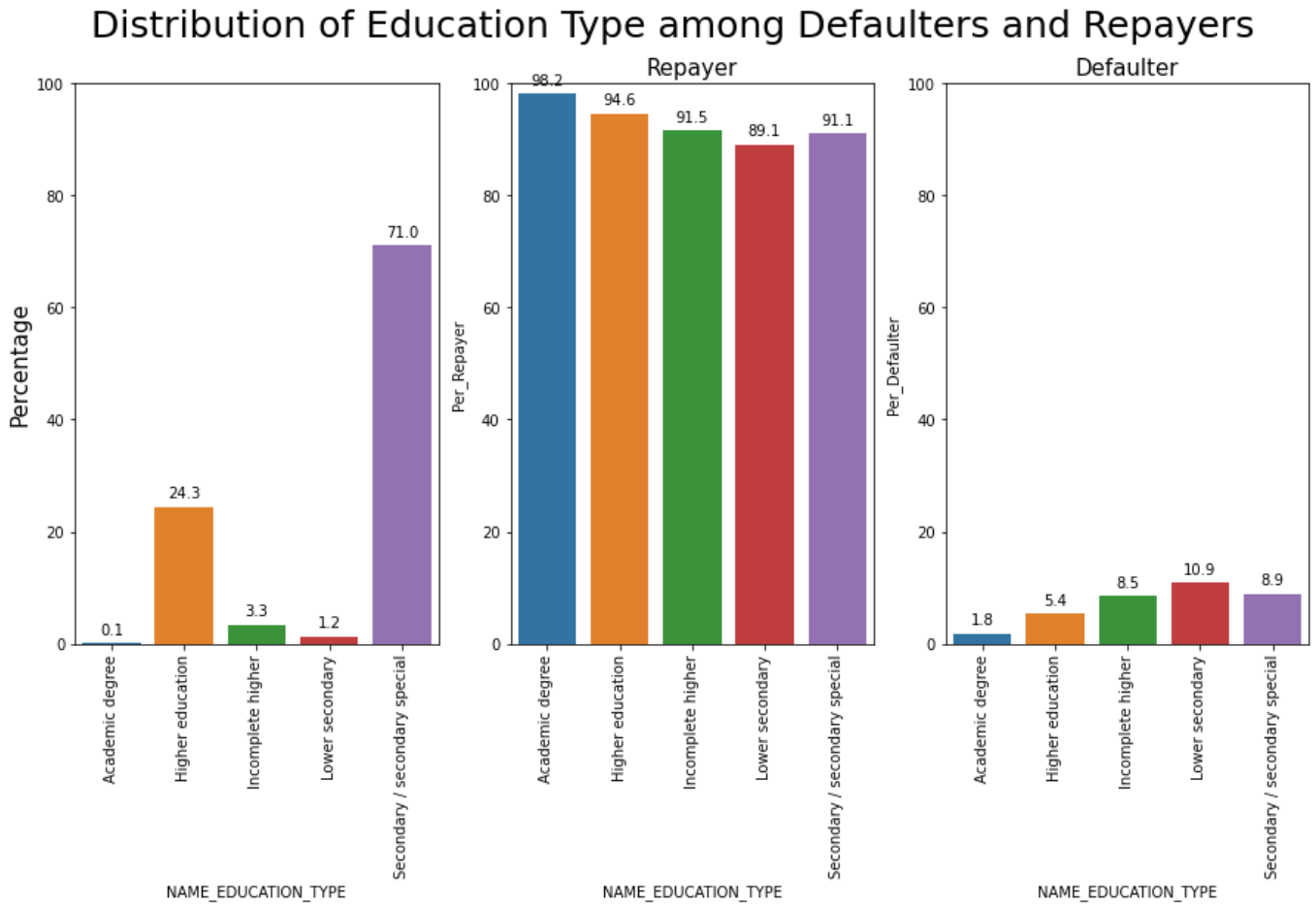
Application Data Analysis



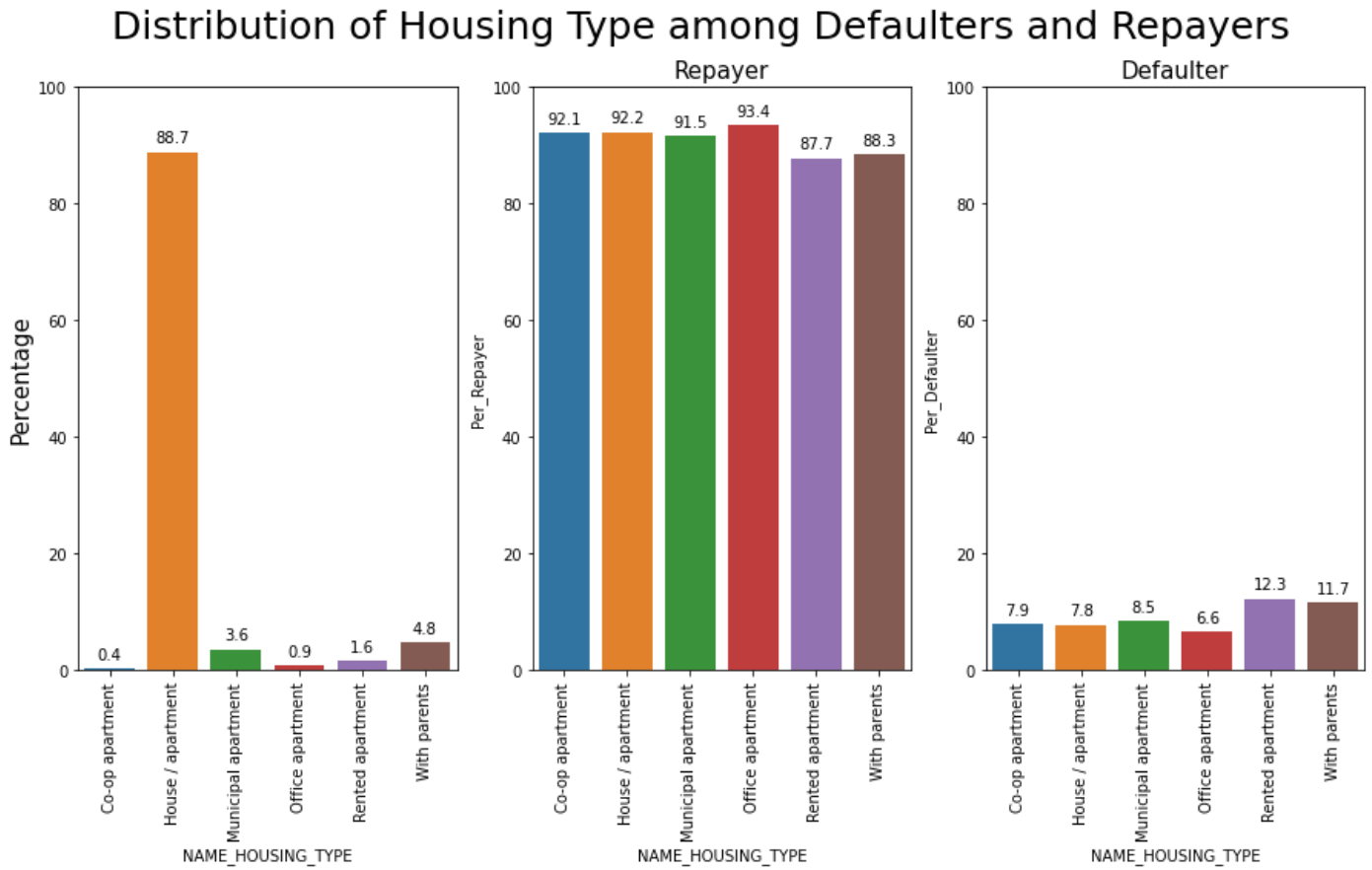
- The percentage of “Maternity Leave” applicants in “NAME_INCOME_TYPE” is very less, but forms 40% of defaulters. Hence, it can be one of the driving factors for loan defaulters.



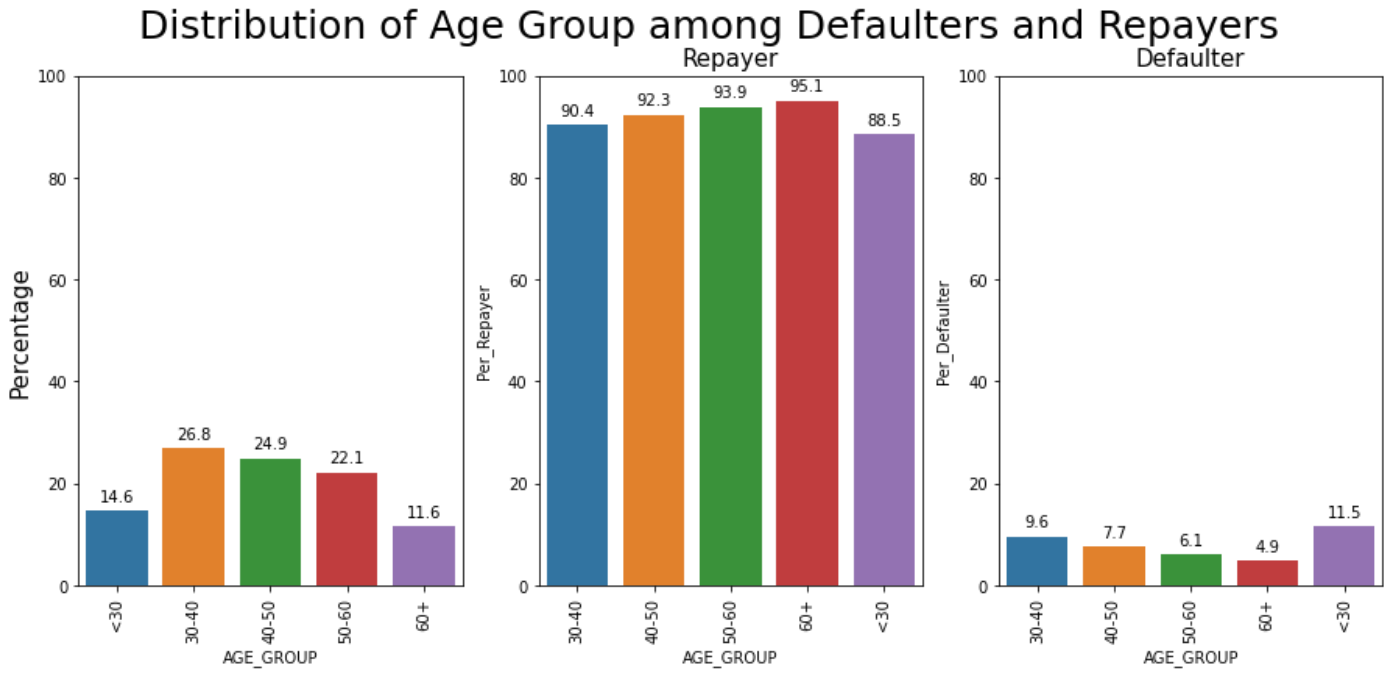
- The percentage of applicants with “Lower Secondary” education is very less, but it has maximum percentage of defaulters, which is 10.9%. Hence, this can be a driving factor for loan defaulter.



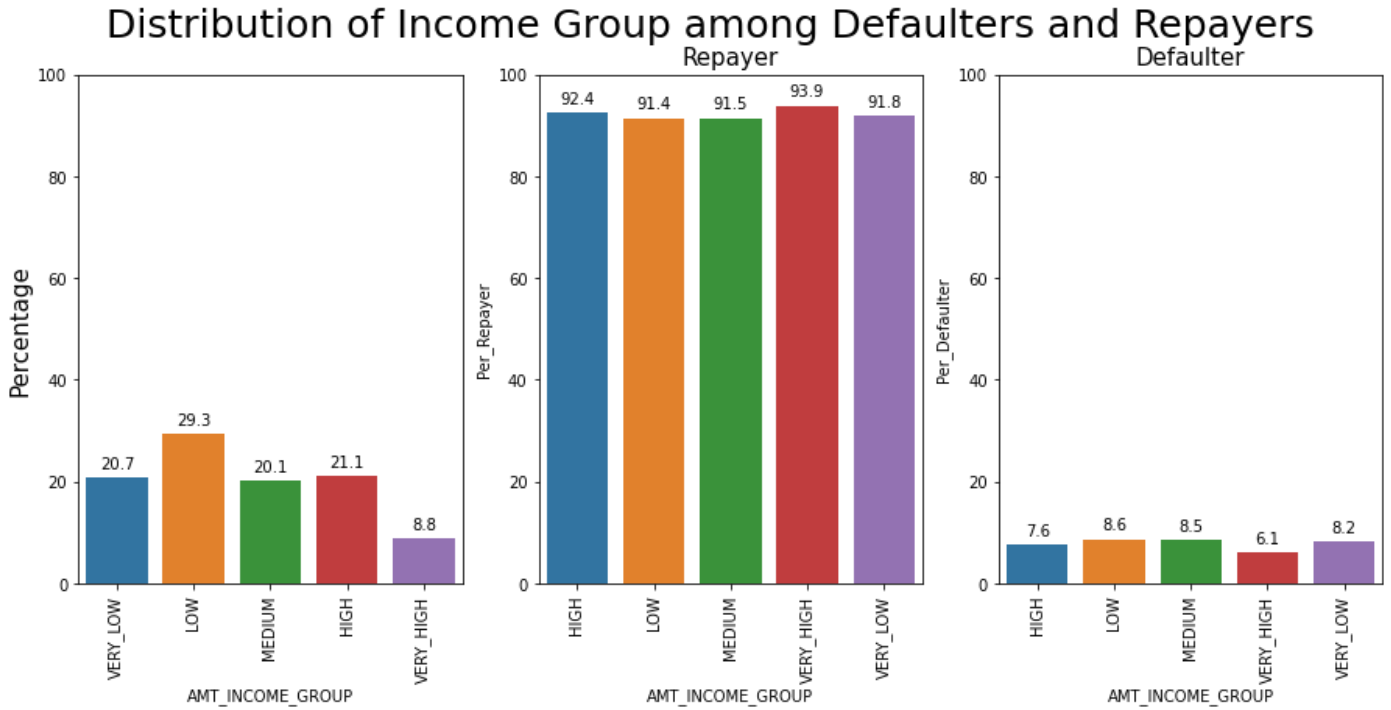
- The percentage of “Office Apartment” applicants is quite less in the Housing type but they are the most repayers. Therefore, they can be one of the factors to identify loan repayers.
- Applicants living in rented apartment form highest percentage of defaulters.



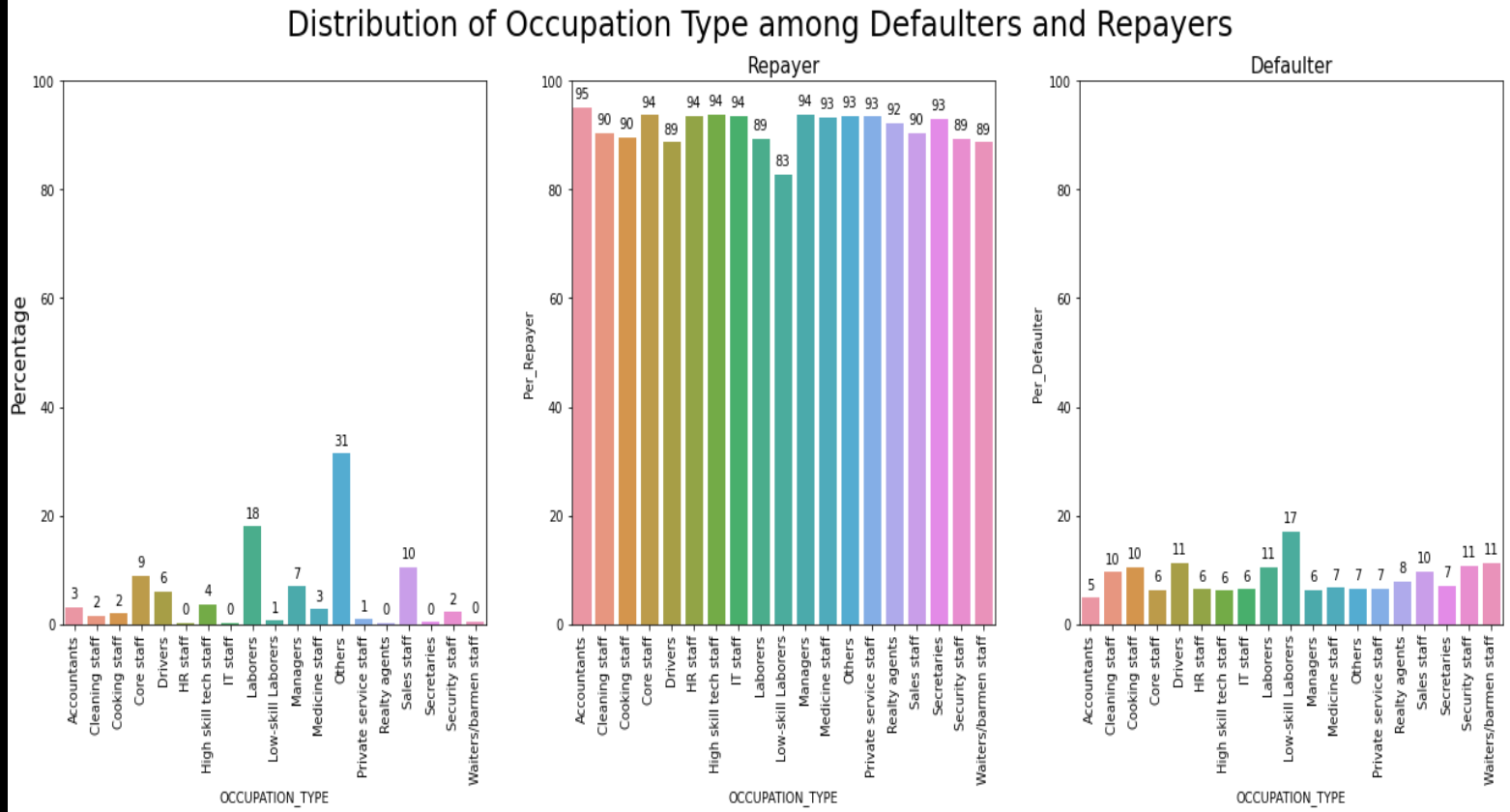
- There are very less applicants above 60 years of age but form the maximum percentage of repayers which is 95.1%. So, this can be one of the driving factors for loan repayers.
- The applicants less than 30 years old are maximum defaulters which can be one of the driving factors for identifying loan defaulters.



- Majority of the applications came in from the low-income group. And they tend to be the major defaulters which can be a driving factor for identifying loan defaulters.
- Applications from very high-income group are major repayers. This factor can help identify the loan repayers.

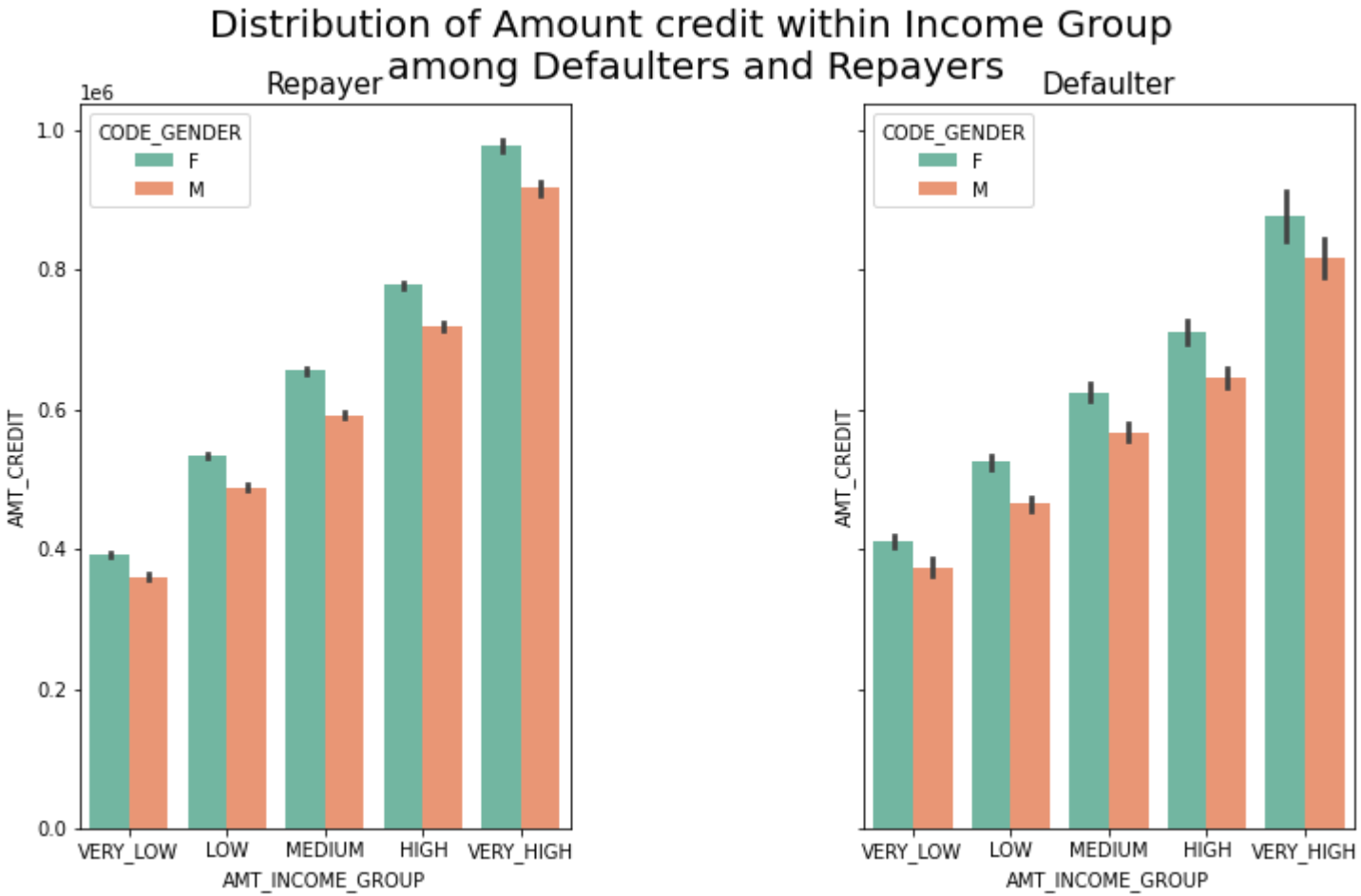


- Although there are a smaller number of applications from low-skill laborers, they form the maximum percentage of defaulters. Hence this can be one of the factors to look at when offering loans and make informed decisions.

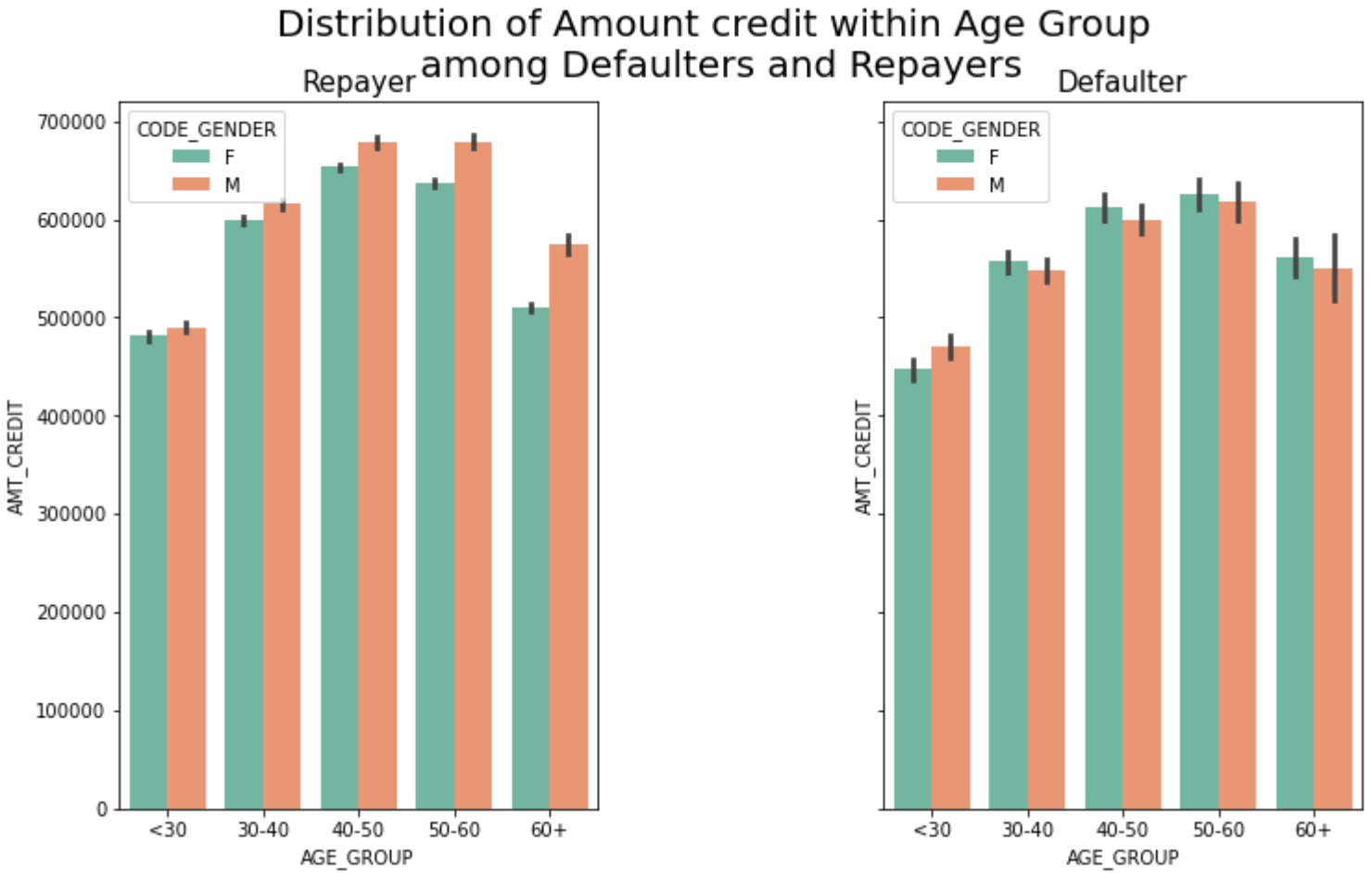


Applicants with very high salary and credit amount are more among repayers than defaulters.

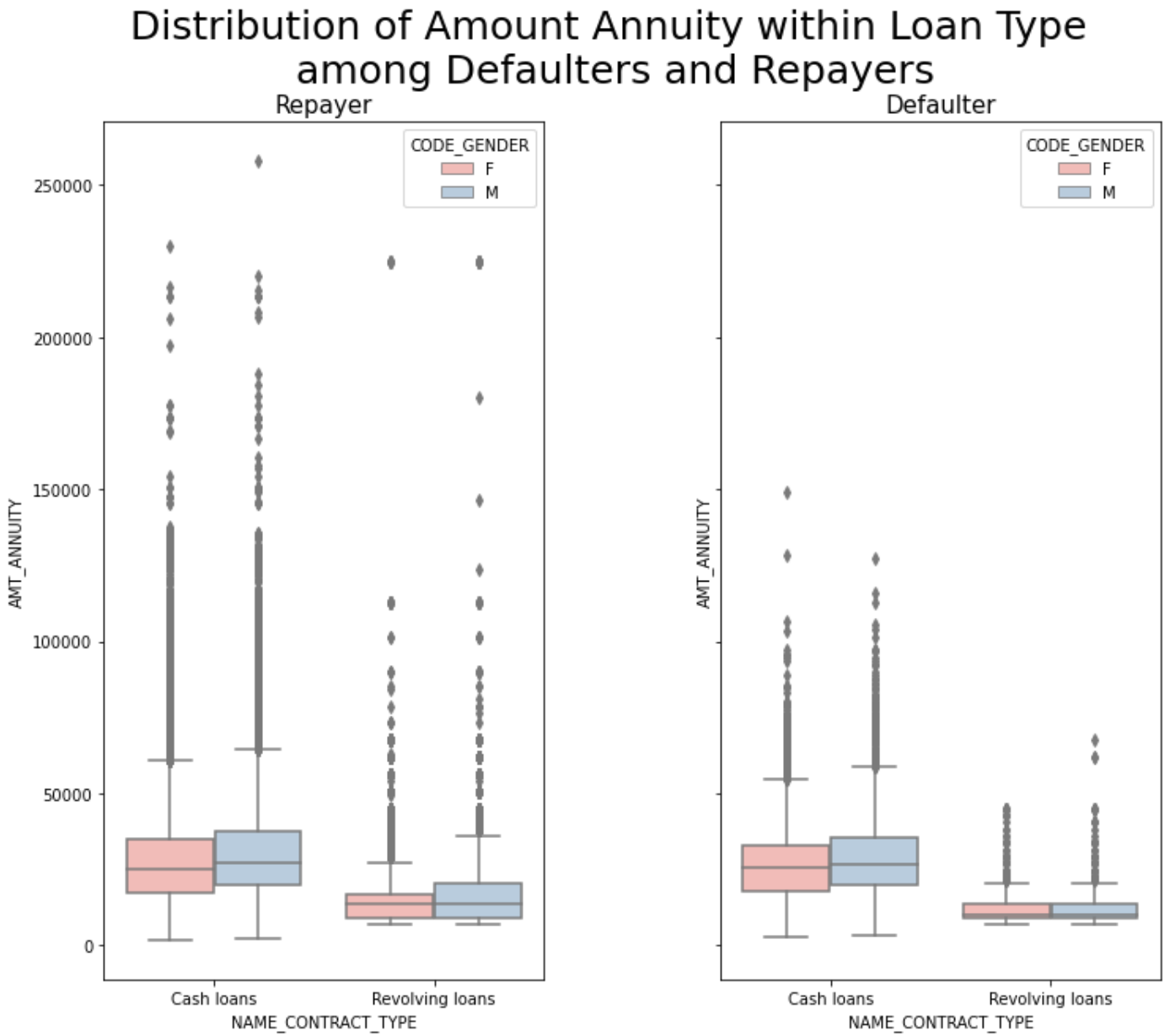
This inference can act as a factor in identifying loan repayers.



- Repayers count appears to be higher among Male compared to female.
- Female above 30 years of age are found to be more defaulters compared to Male.
- People in age group 30-50 with high credit amount are more likely to be repayers .

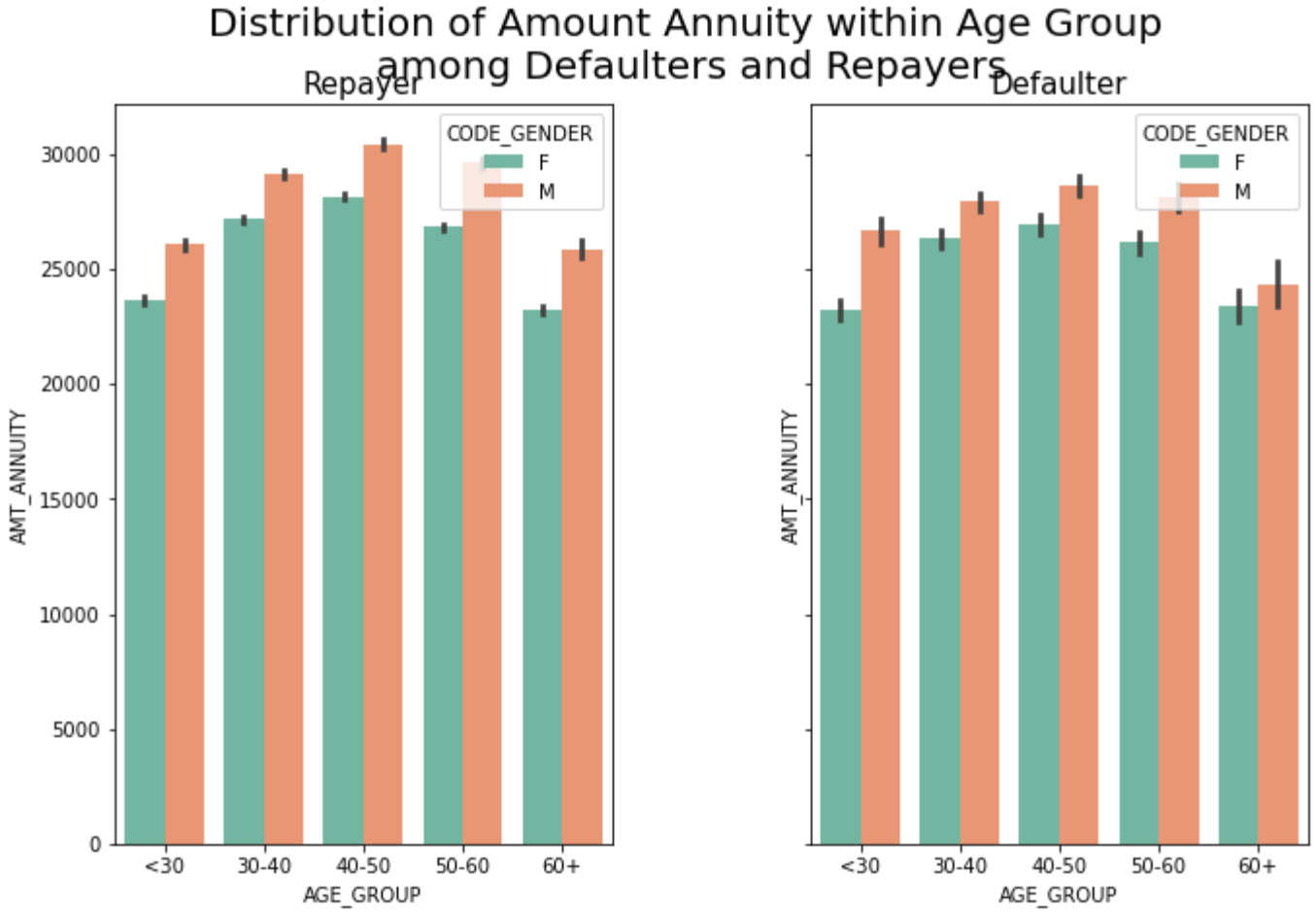


People with high loan annuity amount for cash loans are repayers.



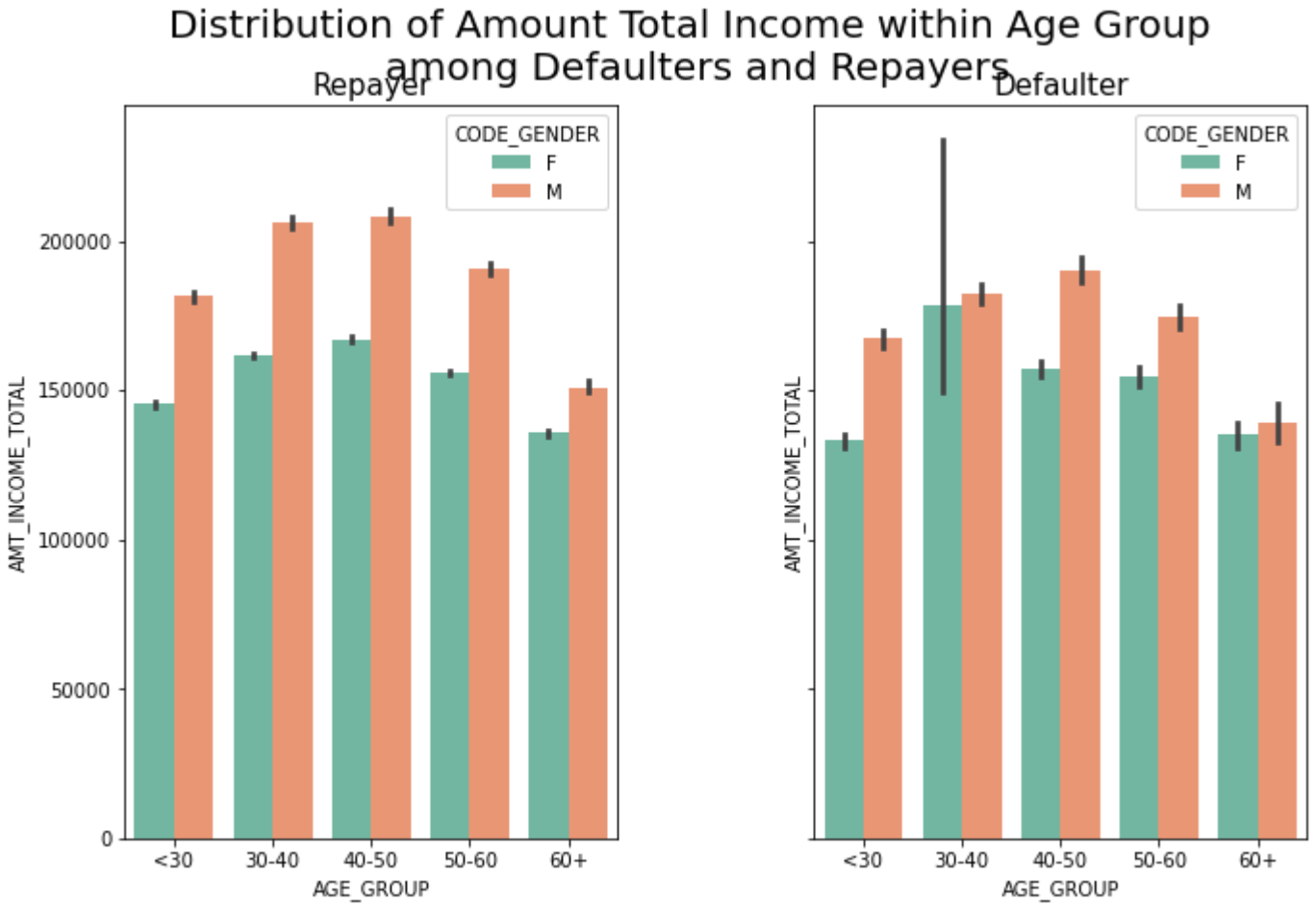
Males of age group 40-50 are more repayers with high loan annuity.

Percentage of males defaulters below age 30 is significantly more than female below age 30. This inference can help as a factor when identifying potential defaulters.

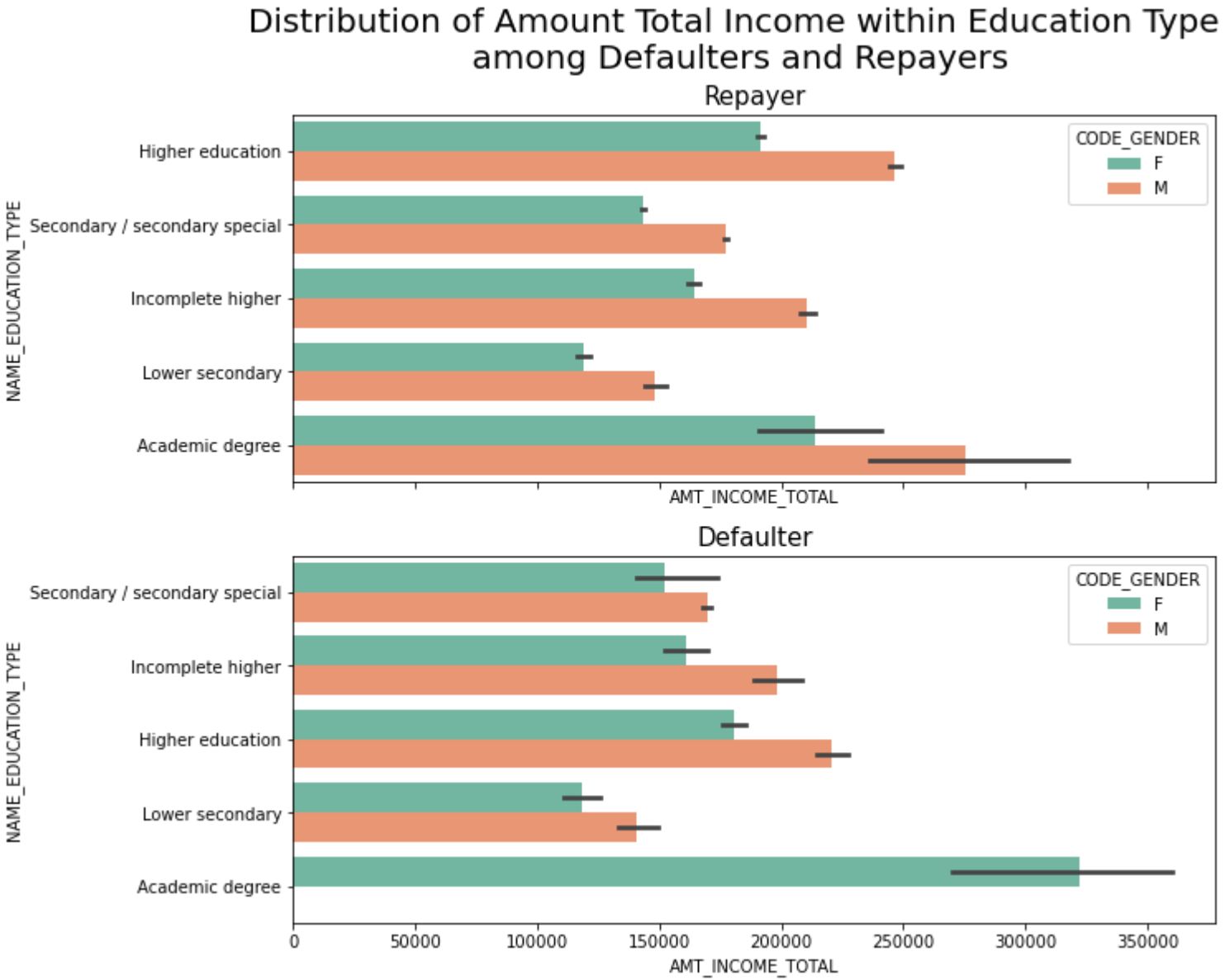


Female in the age group of 30-40 with high salary appear to be potential defaulters.

Male in the age group 30-50 are repayers with decent salary.

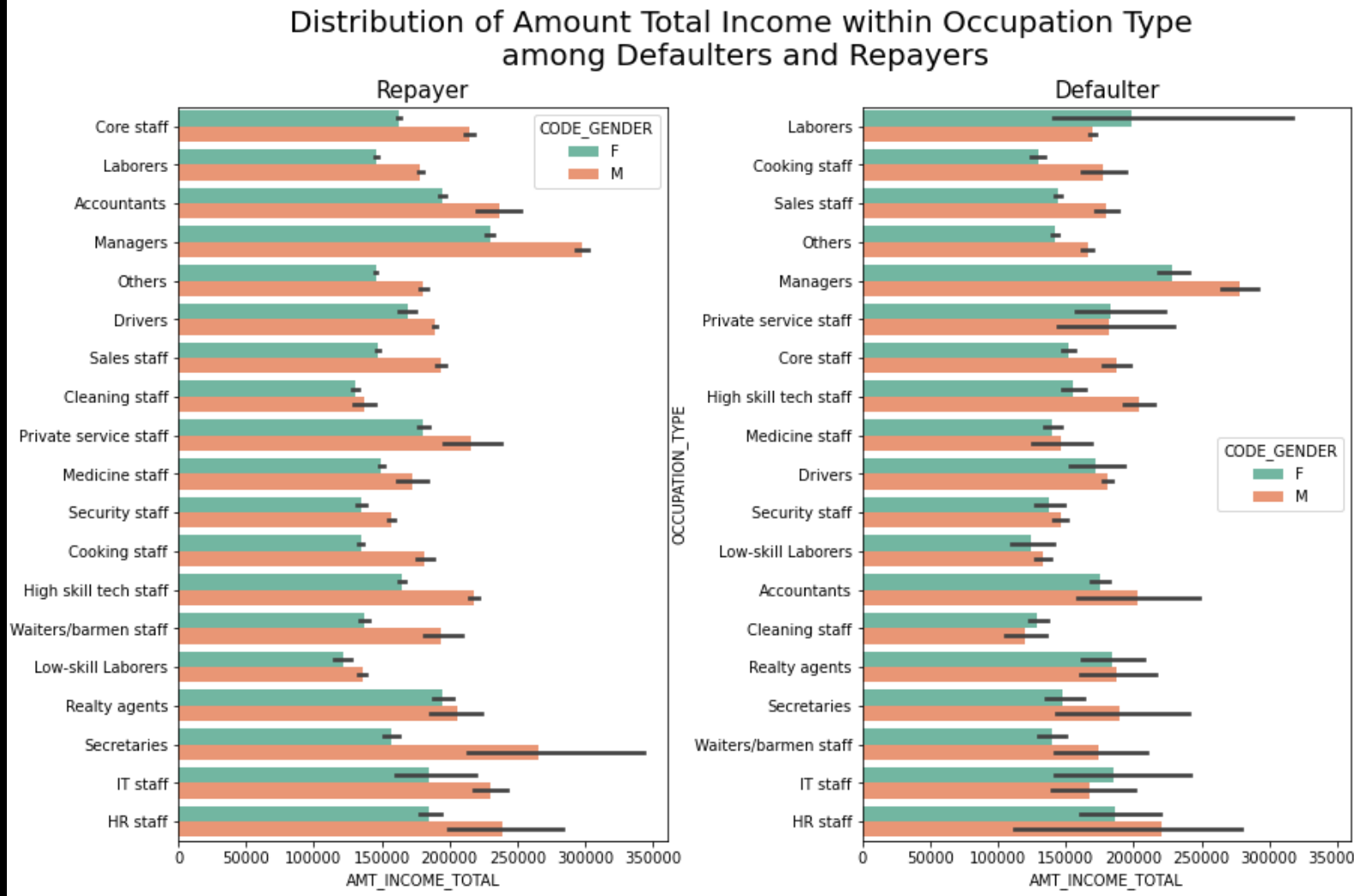


- Female with an academic degree and high income are defaulters whereas Male with Academic degree and high salary appear to be more repayers.
- There are no male defaulters with an academic degree.
- People with lower secondary education appear to be less repayers compared to any other education type.



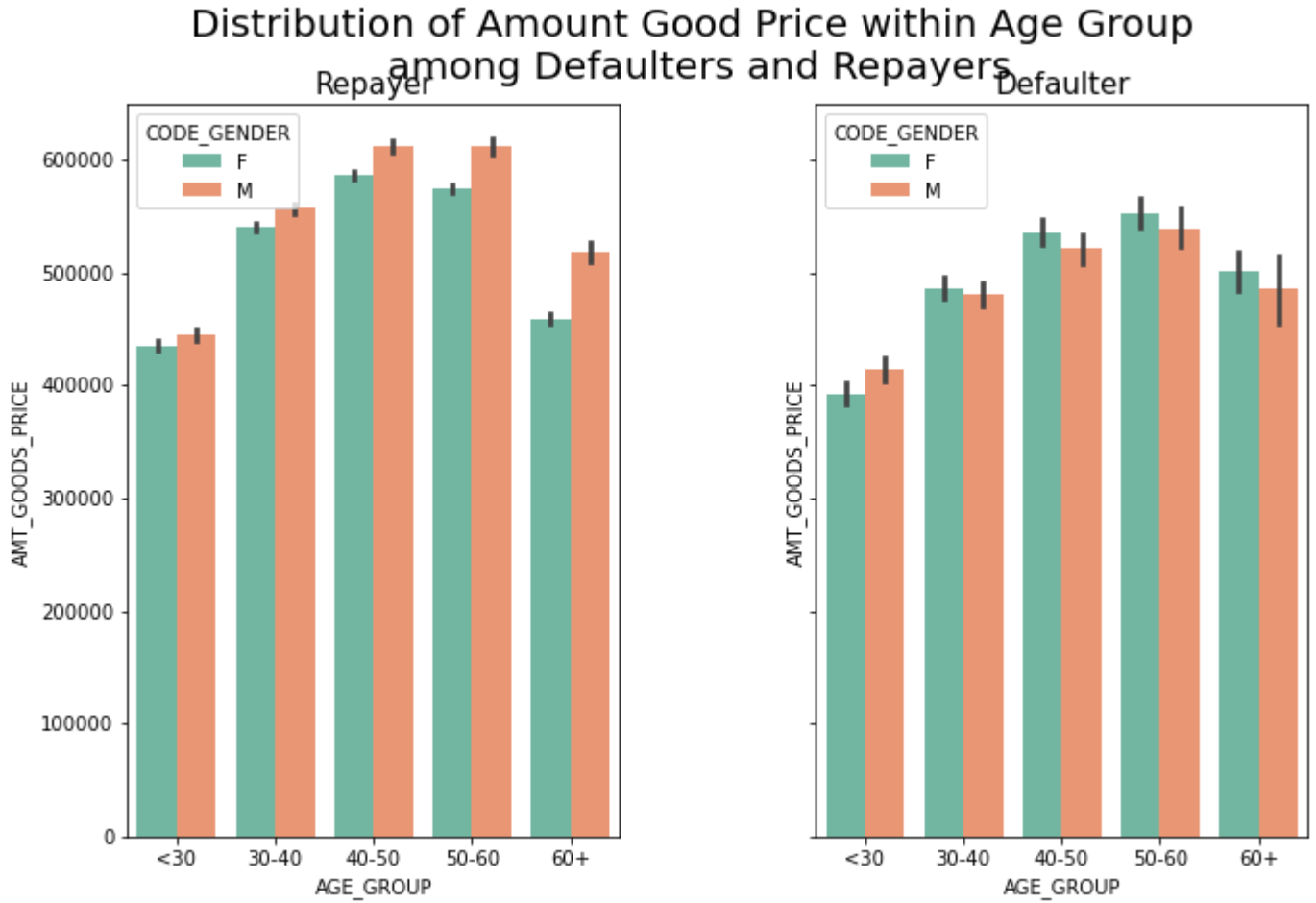
Among laborers, female with high salary are more defaulters. These records were also found in the outlier detection.

Among Secretaries, Male with high salary are more repayers.



As identified in one of the previous inferences (Distribution of Amt_Credit with Age group), female above age 30 are more defaulters compared to Male. However, Female below age 30 appear to be less defaulters compared to Male.

When goods price is very high Male are more repayers than female.



Top 10 Correlation for Defaulter & Repayer clients (Application Data)

Top 10 Correlation for defaulter clients (with payment difficulties)

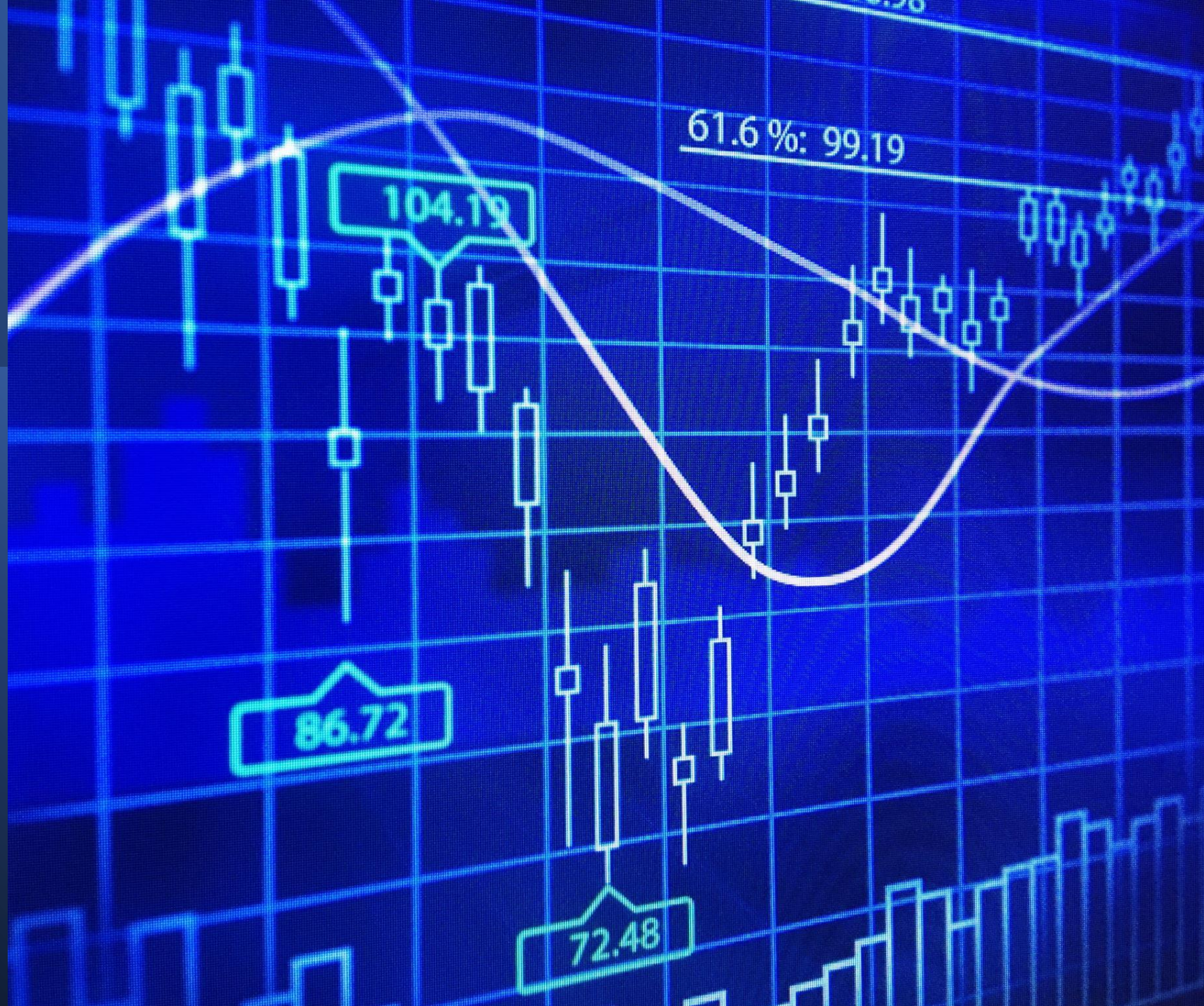
	Var1	Var2	Correlation
28	AMT_GOODS_PRICE	AMT_CREDIT	0.98
19	AMT_ANNUITY	AMT_CREDIT	0.75
29	AMT_GOODS_PRICE	AMT_ANNUITY	0.75
49	DAYS_EMPLOYED	DAYS_BIRTH	0.58
58	DAYS_REGISTRATION	DAYS_BIRTH	0.29
67	DAYS_ID_PUBLISH	DAYS_BIRTH	0.25
68	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.23
59	DAYS_REGISTRATION	DAYS_EMPLOYED	0.19
37	DAYS_BIRTH	AMT_CREDIT	0.14
39	DAYS_BIRTH	AMT_GOODS_PRICE	0.14

Top 10 Correlation for repayers clients (with no payment difficulties)

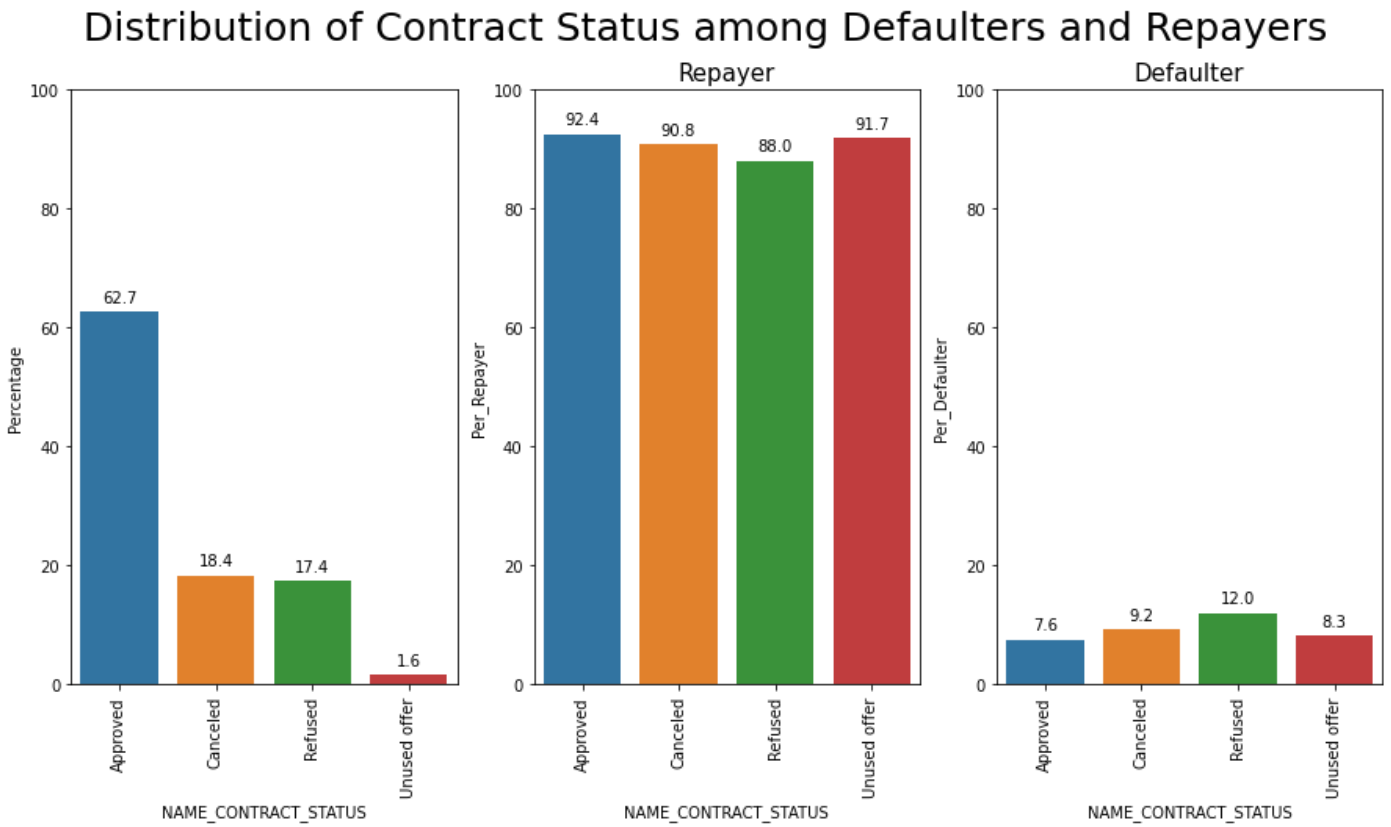
	Var1	Var2	Correlation
28	AMT_GOODS_PRICE	AMT_CREDIT	0.99
29	AMT_GOODS_PRICE	AMT_ANNUITY	0.78
19	AMT_ANNUITY	AMT_CREDIT	0.77
49	DAYS_EMPLOYED	DAYS_BIRTH	0.63
18	AMT_ANNUITY	AMT_INCOME_TOTAL	0.42
27	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.35
9	AMT_CREDIT	AMT_INCOME_TOTAL	0.34
58	DAYS_REGISTRATION	DAYS_BIRTH	0.33
68	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.28
67	DAYS_ID_PUBLISH	DAYS_BIRTH	0.27

- Amount Annuity, Amount Credit and Amount Goods Price are more correlated to Amount Total Income among repayers compared to defaulters.

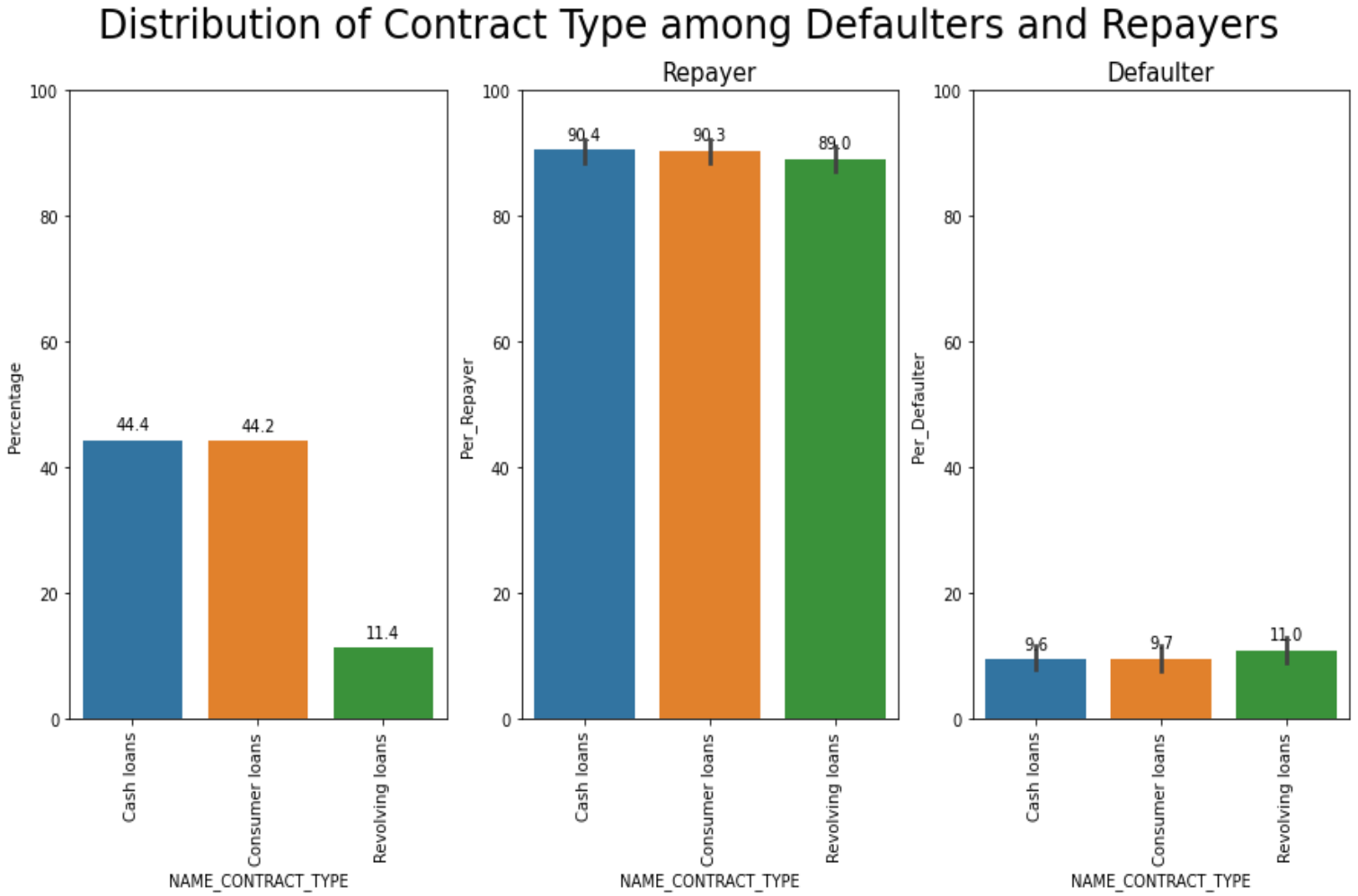
Merged Data Analysis



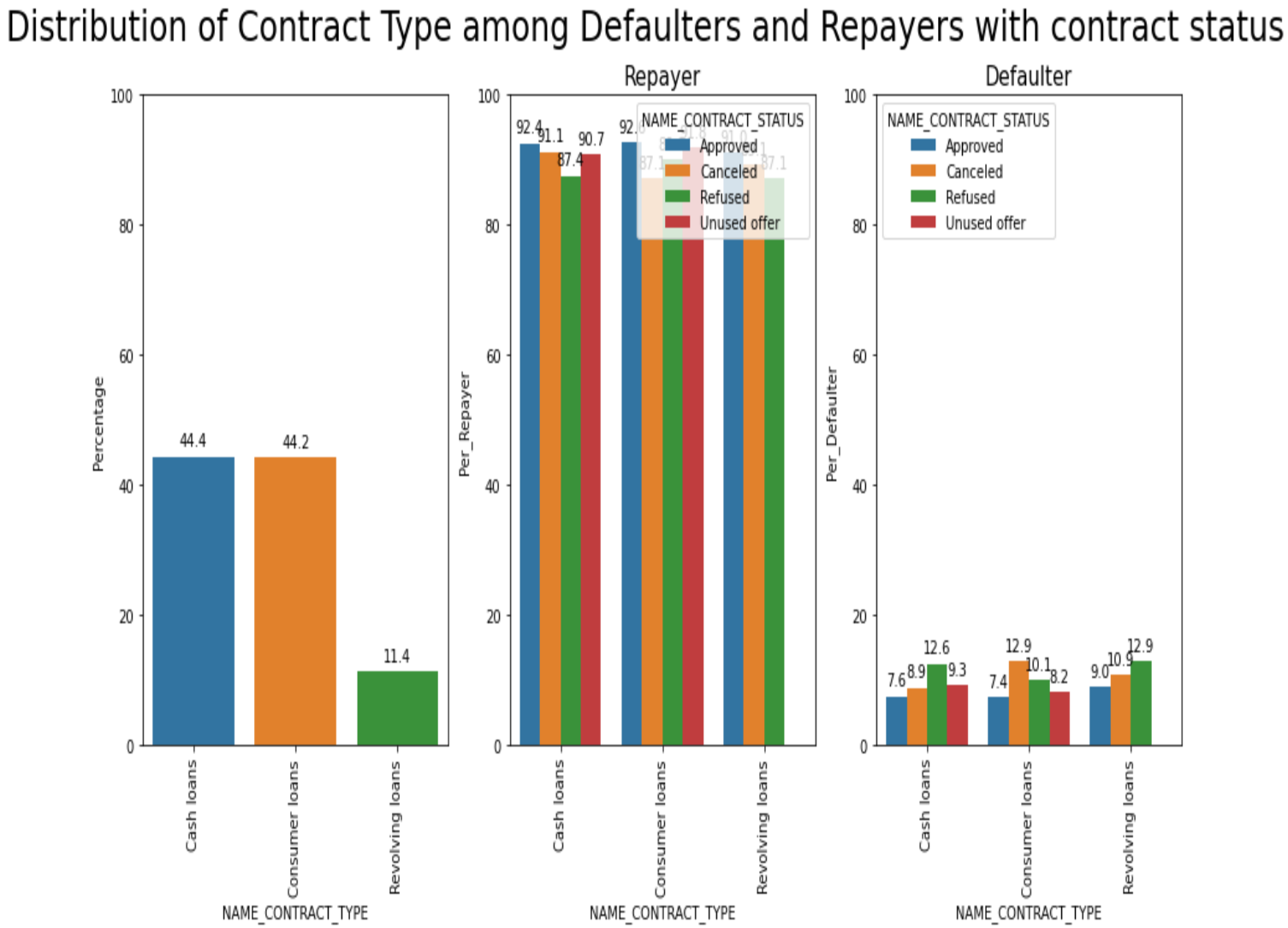
The percentage of applicants with “Refused” NAME_CONTRACT_STATUS in previous application is comparatively less but has maximum percentage of defaulters. This can be a driving factor for loan defaulters.



The percentage of applicants for “Revolving loans” is less among previous applications but they form the maximum percentage of defaulters. Hence, applicants with revolving loans in previous application can be a contributing factor for loan defaulters

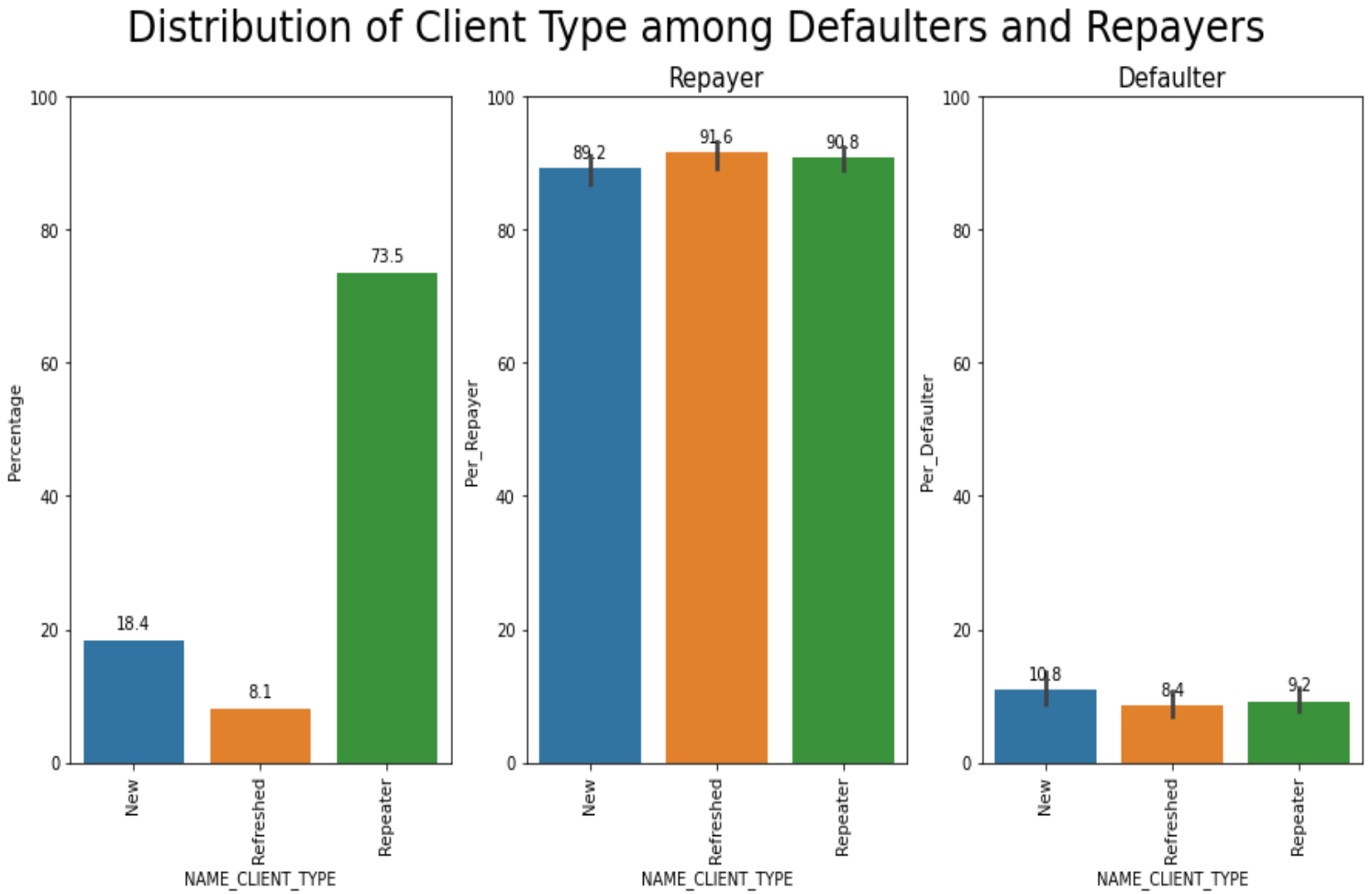


Applicants with 'Revolving loans' and with 'Refused’ contract status on previous applications tend to be more defaulters.



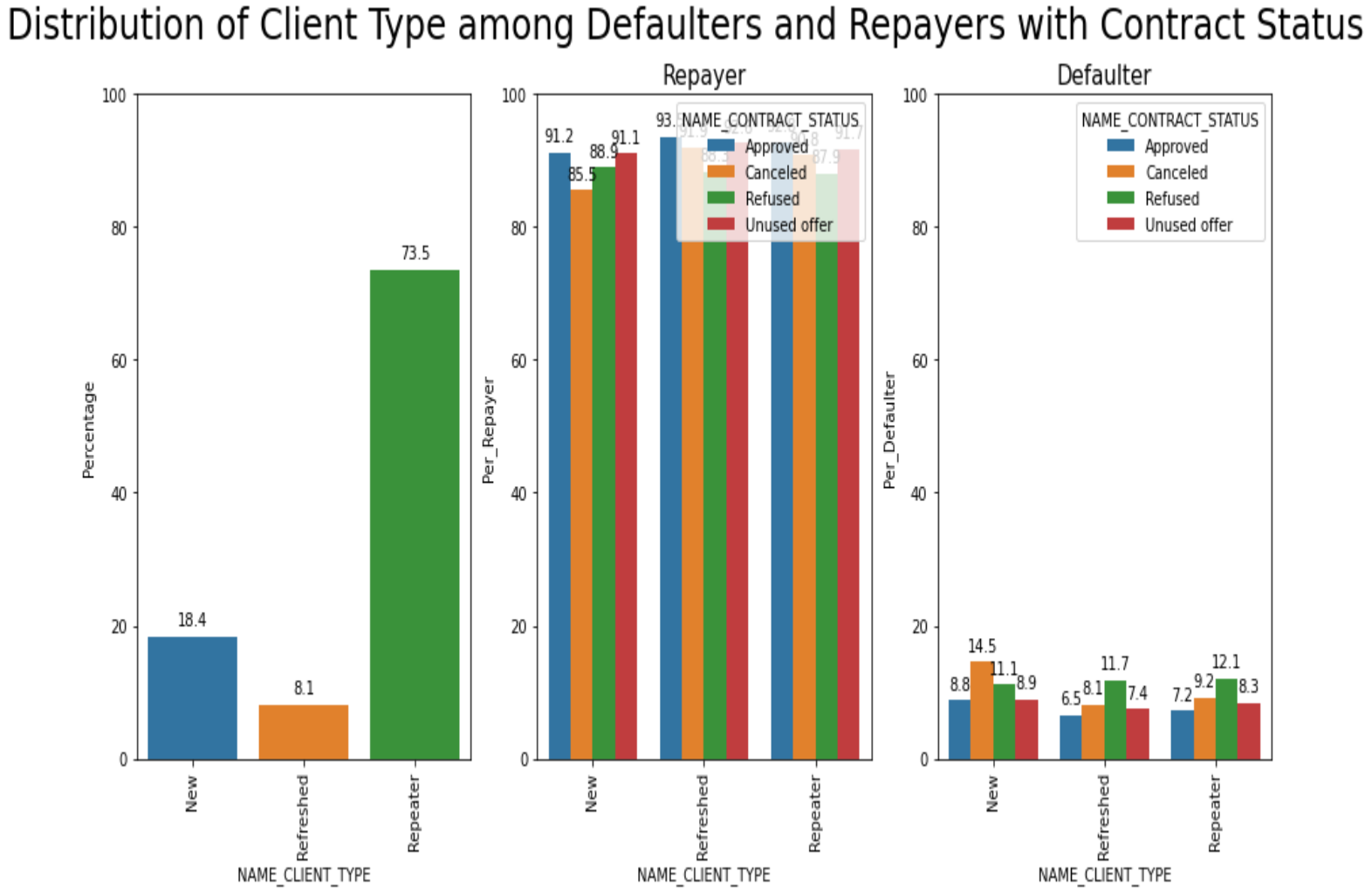
The ‘Refreshed’ clients are less among the previous applications but are highest number of repayers. This inference can be helpful in identifying loan repayers.

‘New’ clients tend to be more defaulters as per the distribution shown.

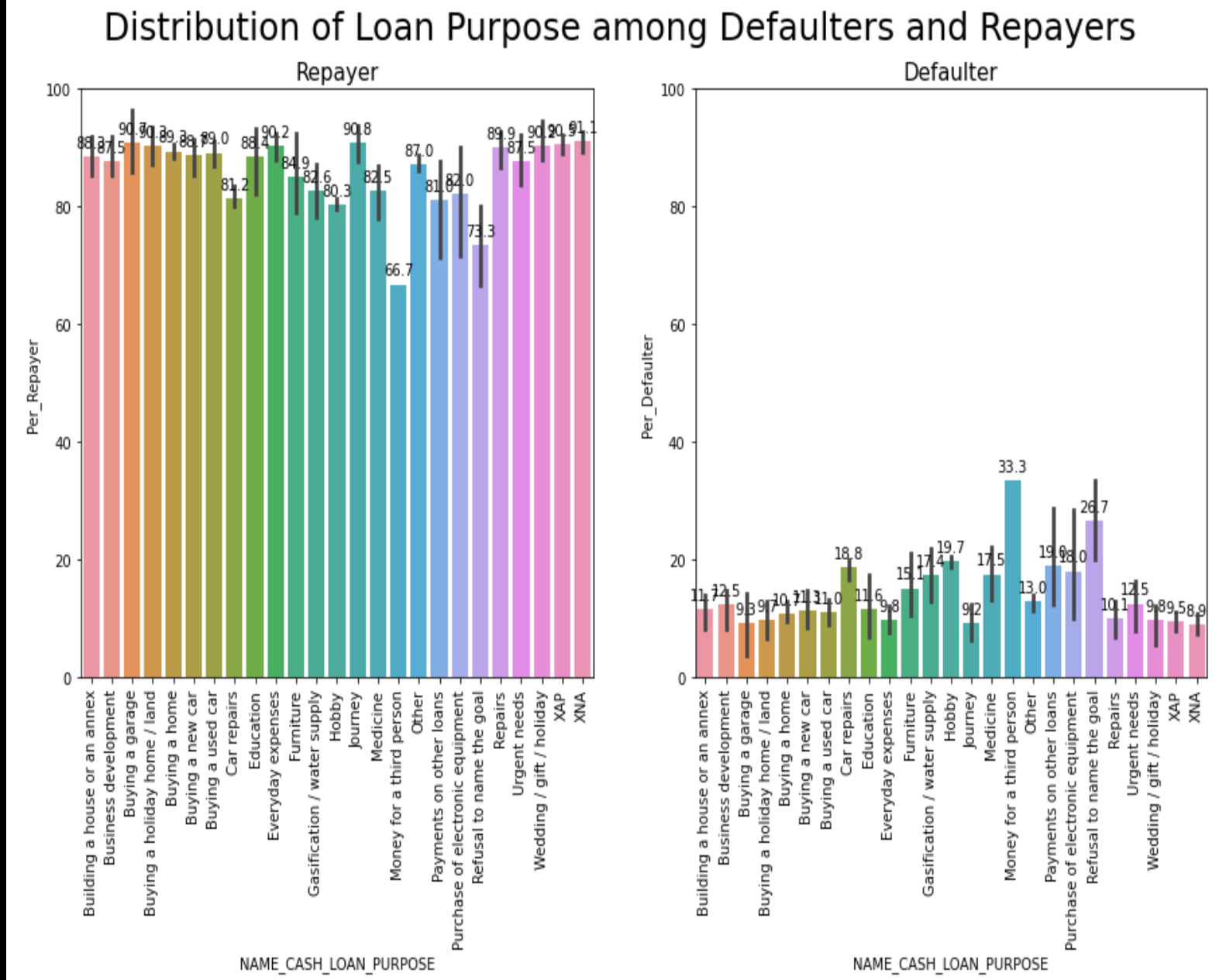


Most of the previous applications were Repeaters. Majority of previous applications of repeaters that are current defaulters were refused.

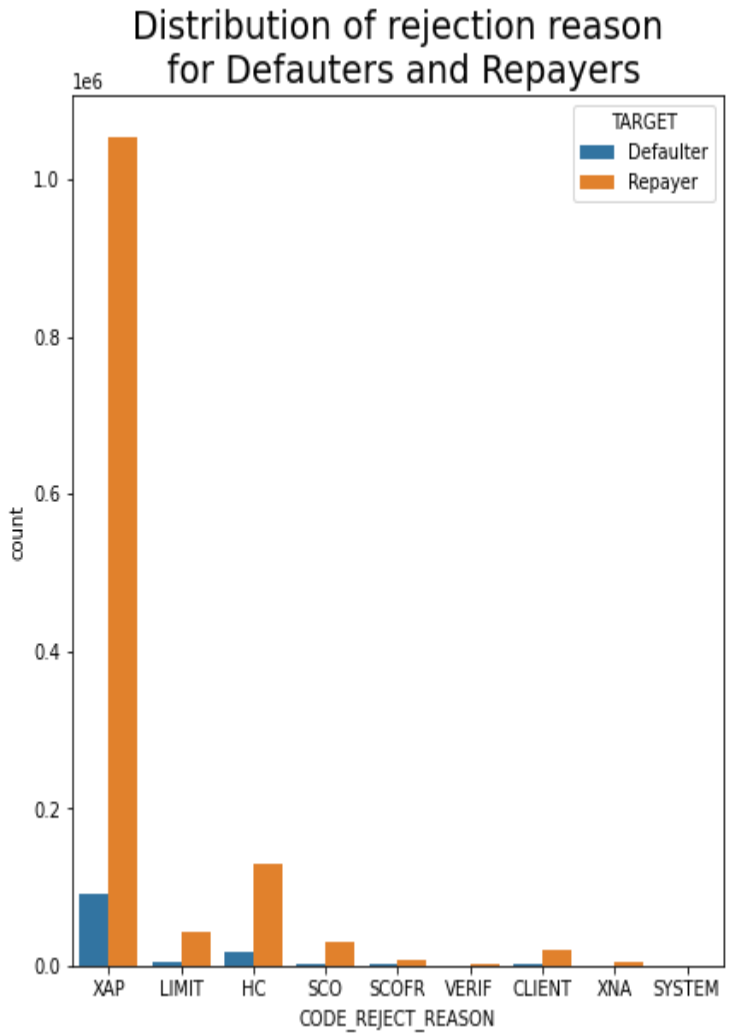
Highest percentage were from cancelled contract type that are current defaulters.

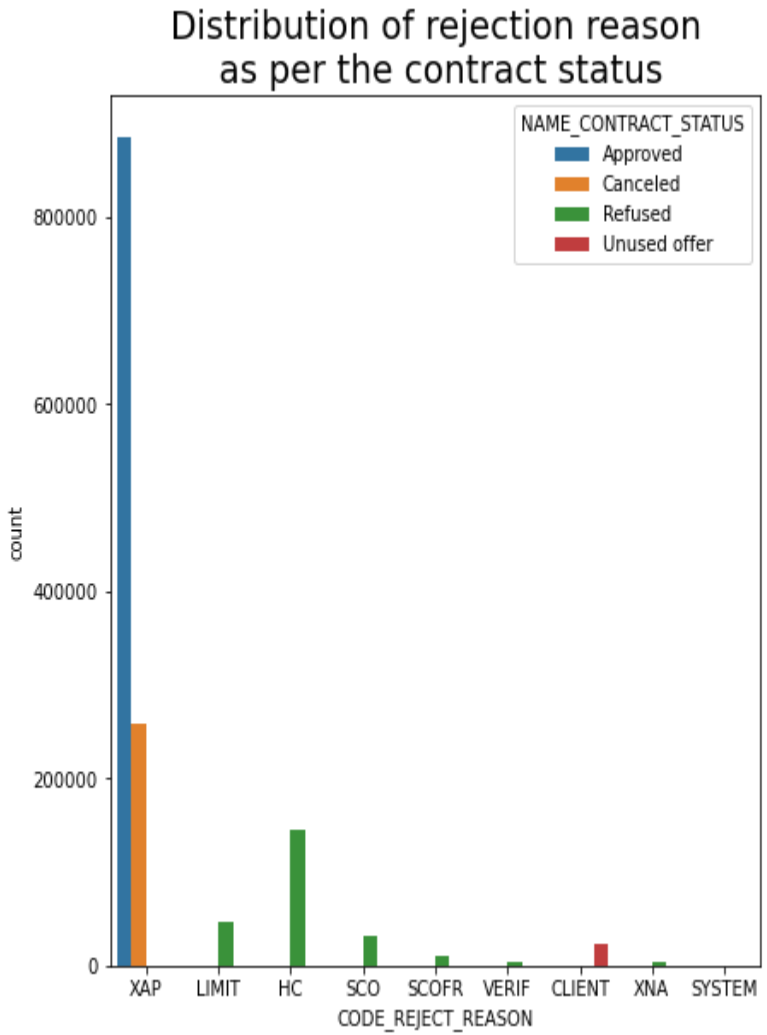


Loan applications applied for third person turned out to be maximum defaulters.



Most of the rejection reason is 'XAP' which is unknown. However, it is identified that majority of 'XAP' values has their contract status as 'Approved'.

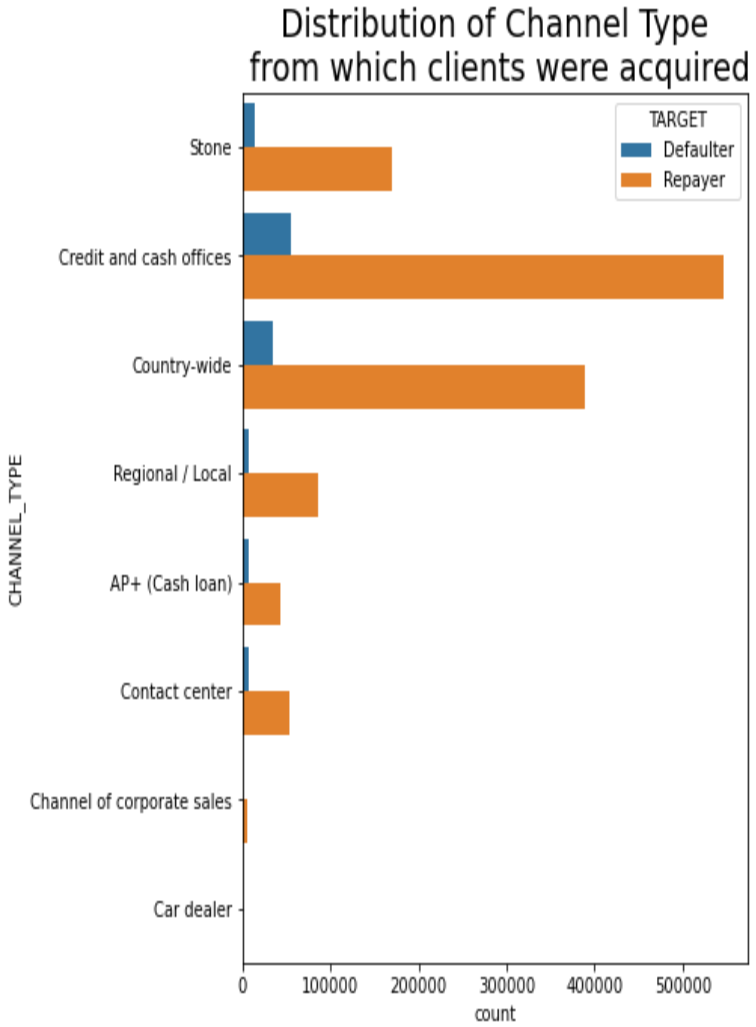


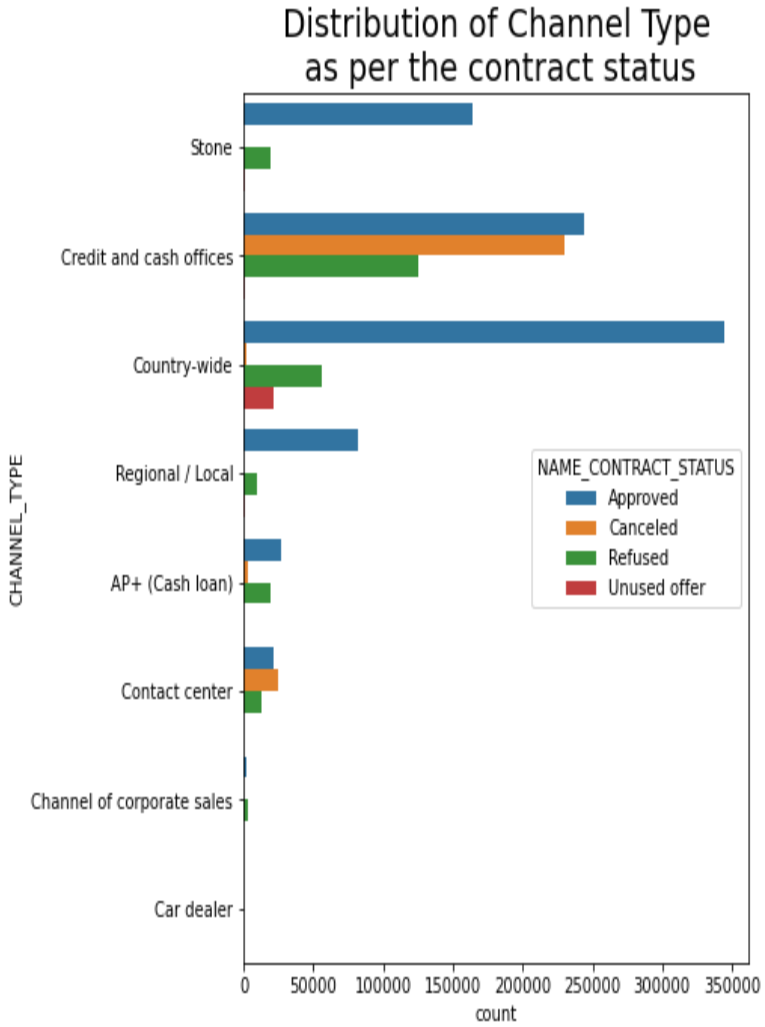


The approval rate from country wide channel was high among previous applications. Hence these applications can be trusted for sanctioning a loan.

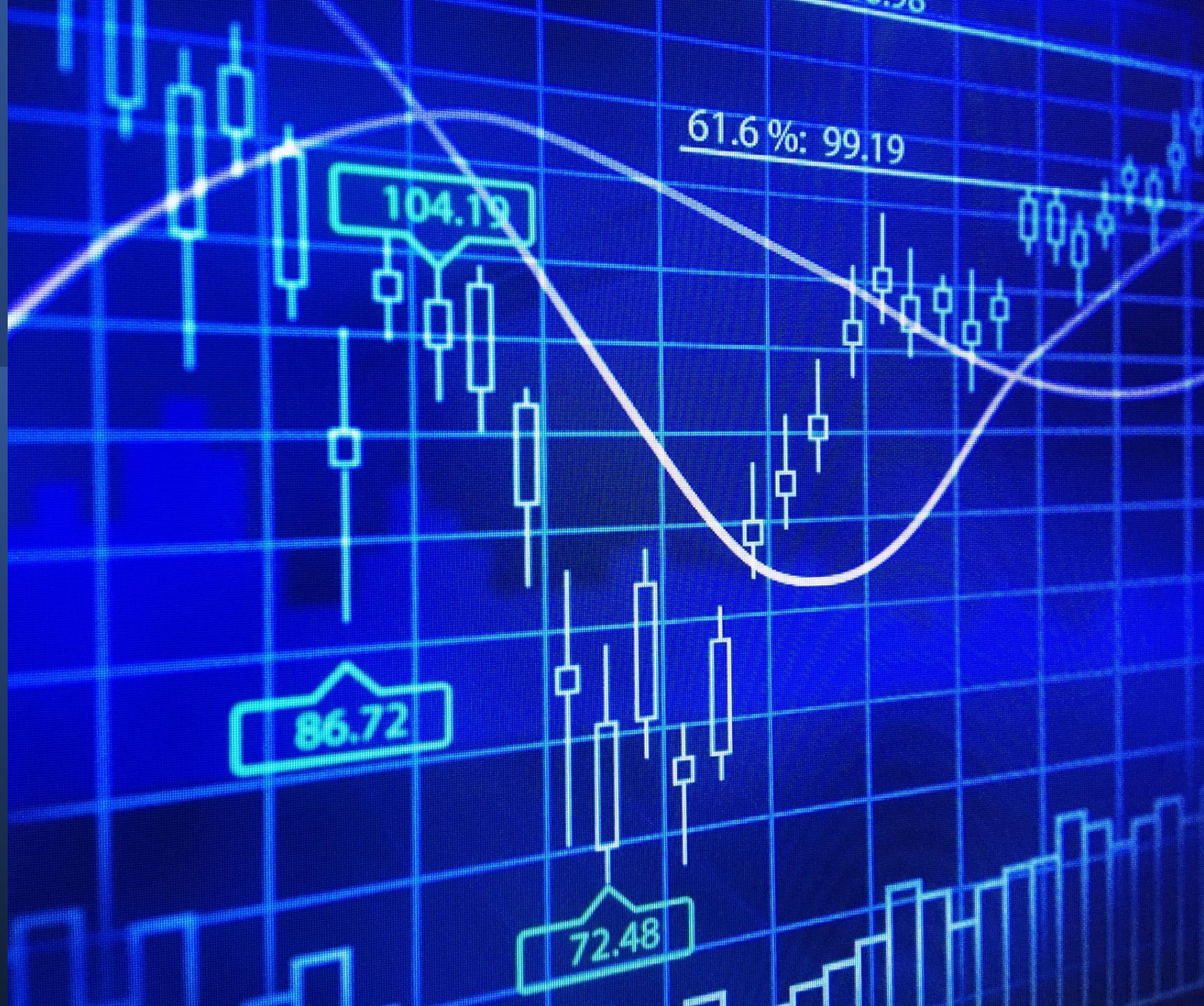
Most of the current repayers are from Credit and Cash Offices. But majority of their previous applications were cancelled or refused.

Although there are more current repayers than defaulters in Regional/Local, no previous applications were approved from this channel type.





Correlation Analysis



- *Among the given variables in the data set, AMT_GOODS_PRICE and AMT_APPLICATION are observed to be highly correlated.*
- *AMT_GOODS_PRICE, AMT_APPLICATION, AMT_CREDIT and AMT_ANNUITY are strongly correlated with each other.*
- *It can be observed that CNT_PAYMENT and AMT_ANNUITY are more correlated among defaulters compared to repayers.*

Correlation between the variables for Defaulters

Correlation between the variables for Repayers

Top 10 Correlation for Defaulter & Repayer clients

Top 10 Correlation for defaulter clients (with payment difficulties)

	Var1	Var2	Correlation
22	AMT_GOODS_PRICE	AMT_APPLICATION	1.00
23	AMT_GOODS_PRICE	AMT_CREDIT	0.99
15	AMT_CREDIT	AMT_APPLICATION	0.98
14	AMT_CREDIT	AMT_ANNUITY	0.84
21	AMT_GOODS_PRICE	AMT_ANNUITY	0.84
7	AMT_APPLICATION	AMT_ANNUITY	0.82
29	CNT_PAYMENT	AMT_APPLICATION	0.70
30	CNT_PAYMENT	AMT_CREDIT	0.69
31	CNT_PAYMENT	AMT_GOODS_PRICE	0.69
28	CNT_PAYMENT	AMT_ANNUITY	0.47

Top 10 Correlation for repayers clients (with no payment difficulties)

	Var1	Var2	Correlation
22	AMT_GOODS_PRICE	AMT_APPLICATION	1.00
23	AMT_GOODS_PRICE	AMT_CREDIT	0.99
15	AMT_CREDIT	AMT_APPLICATION	0.98
21	AMT_GOODS_PRICE	AMT_ANNUITY	0.82
14	AMT_CREDIT	AMT_ANNUITY	0.82
7	AMT_APPLICATION	AMT_ANNUITY	0.81
29	CNT_PAYMENT	AMT_APPLICATION	0.68
31	CNT_PAYMENT	AMT_GOODS_PRICE	0.67
30	CNT_PAYMENT	AMT_CREDIT	0.67
47	DAYS_TERMINATION	DAYS_DECISION	0.39

- AMT_GOODS_PRICE-AMT_ANNUITY and AMT_APPLICATION-AMT_ANNUITY are slightly more correlated among defaulters compared to repayers.