CREDIT –EDA Case Study

Umang Rana

Contents:

- Problem Statement
- Data Cleaning Approaches
- Univariate/Bivariate Analysis
- Segmented Univariate/Bivariate Analysis
- Correlations, Heatmaps & Top 10 Target variables
- Analysis on Previous application data
- Merged-Data(Current+Previous application) Analysis
- Conclusion & Recommendations

Problem Statement

Introduction

This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, 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 specializes in lending various types of loans to urban customers. You have to use EDA to analyze 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.

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):

1.Approved: The Company has approved loan Application

2.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.

3.Refused: The company had rejected the loan (because the client does not meet their requirements etc.).

4.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 Objectives

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).

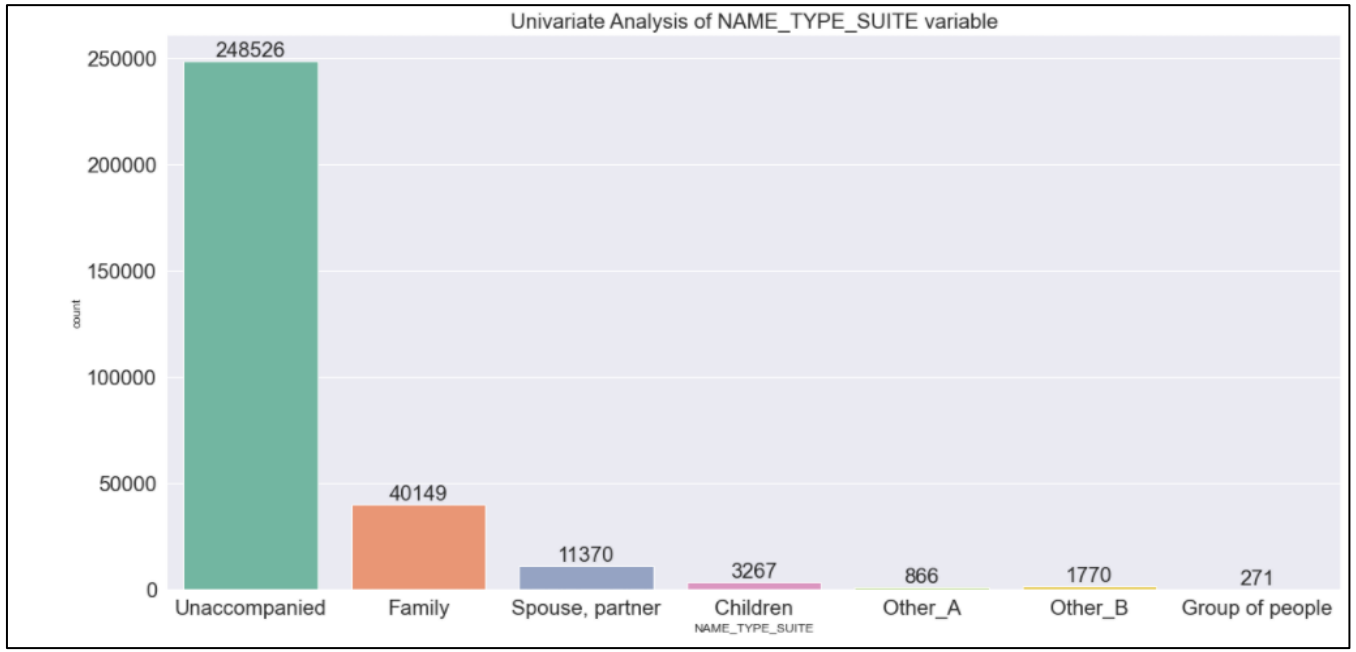
Data Cleaning Approaches

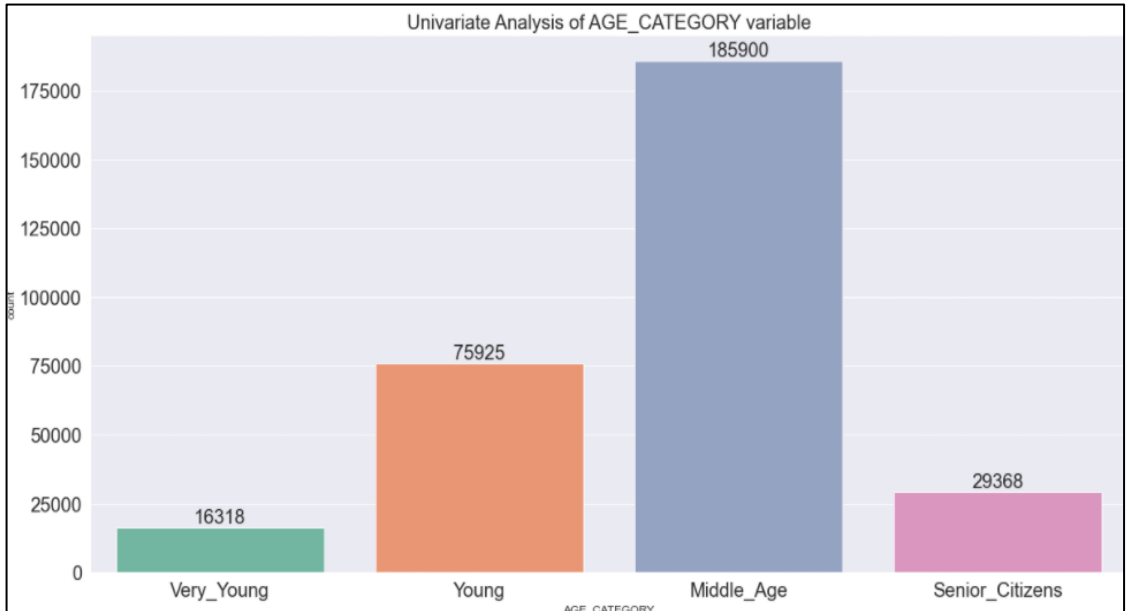
In order to clean the data:

- I have dropped columns having null values greater than 45%, for the remaining columns , I ran missing value check and imputed them few categorical columns with ‘Mode’ of the respective columns.
- Few columns with greater importance had few missing values, and imputing/dropping might have caused a bias Example:OCCUPATION_TYPE, so I have ignored those missing values as they will be ignored by charts and analysis.
- Also few columns like CODE_GENDER & ORGANIZATION_TYPE had ‘XNA’ values, I have imputed/replaced those with ‘NaN’ values as they are missing values.
- The Miscellaneous columns have been dropped and existing columns have been used to create additional new columns for the purpose of analysis using Binning. Example: AGE_CATEGORY & INCOME_CATEGORY

Univariate Analysis

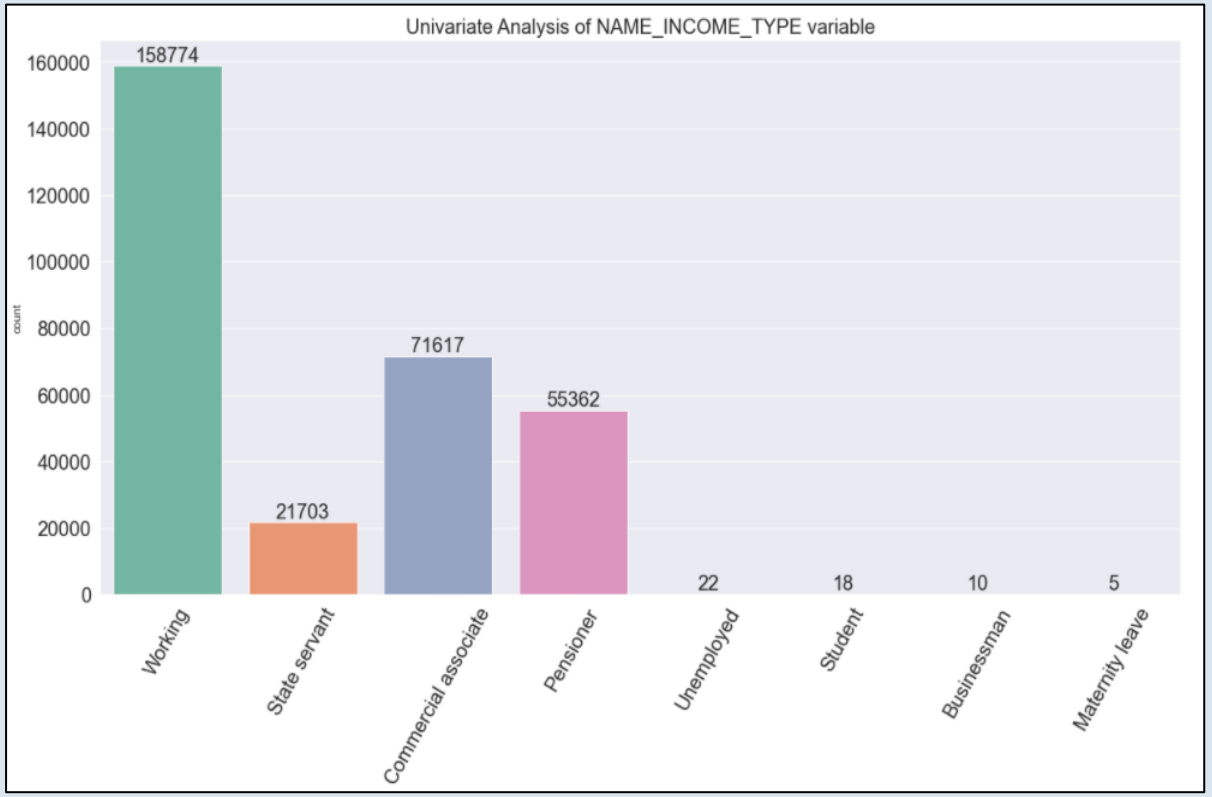
- From the given chart on NAME_TYPE_SUITE variable we could see that most of the applicants are Unaccompanied.

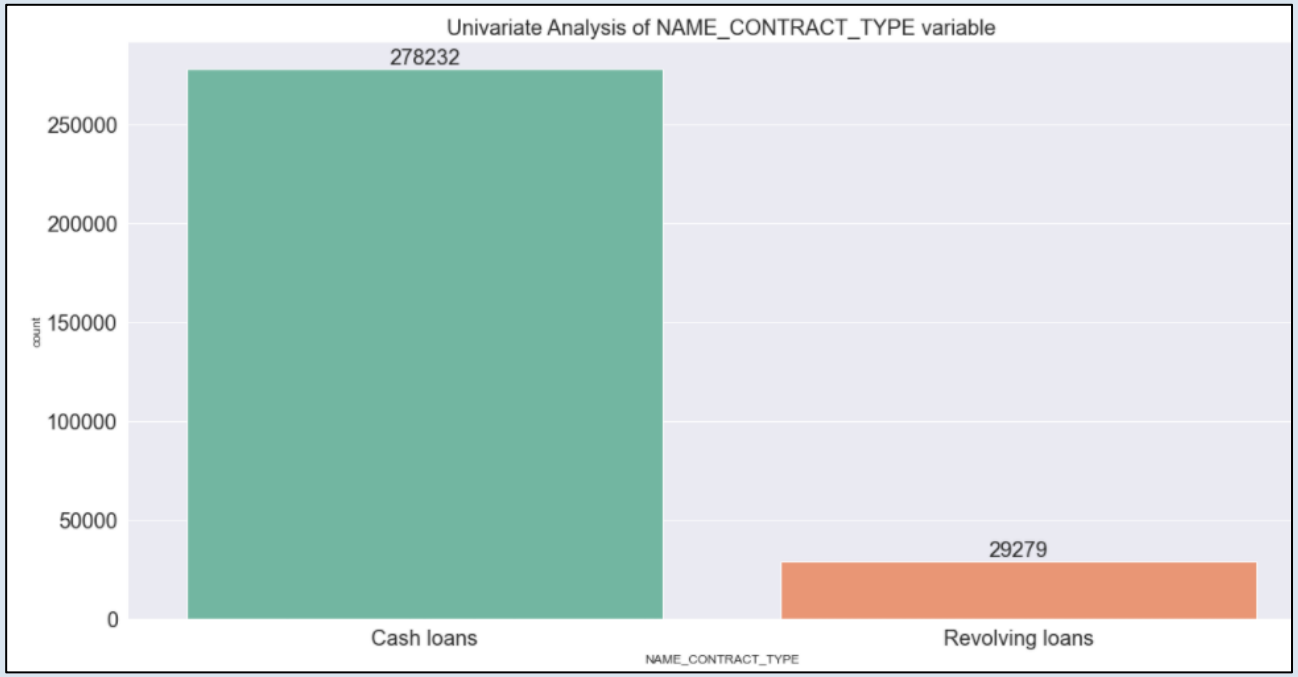


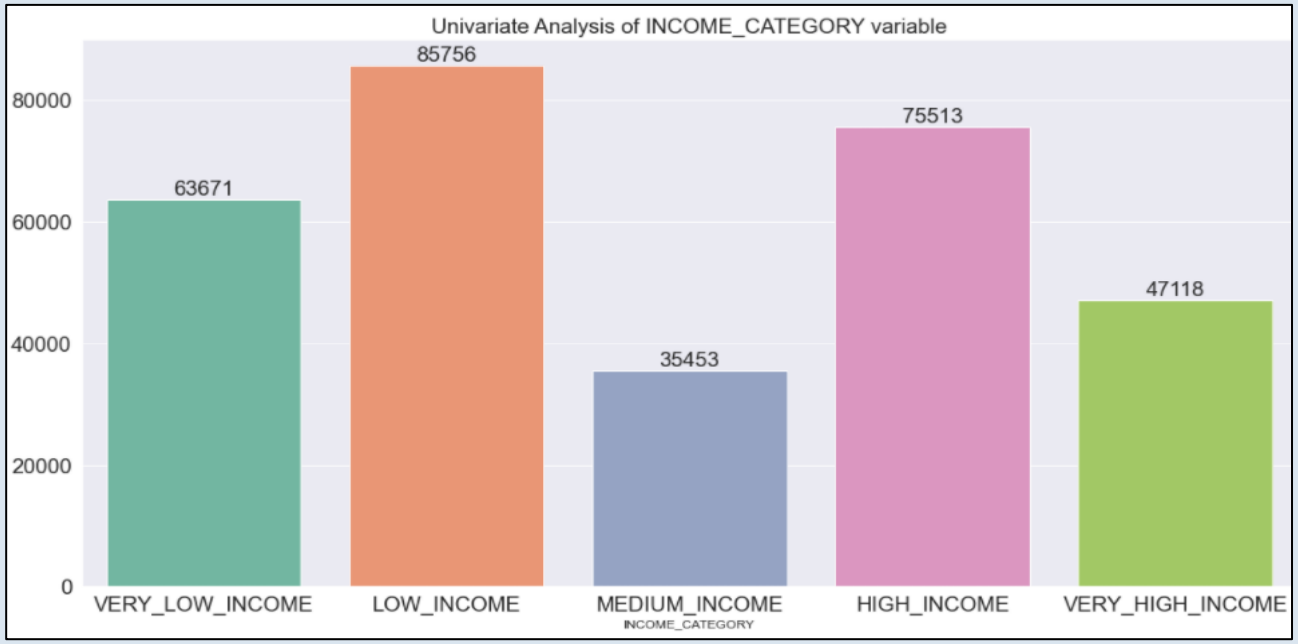


- From the given chart on AGE_CATRGORY variable we could see that most of the applicants belong to Middle_Age category.

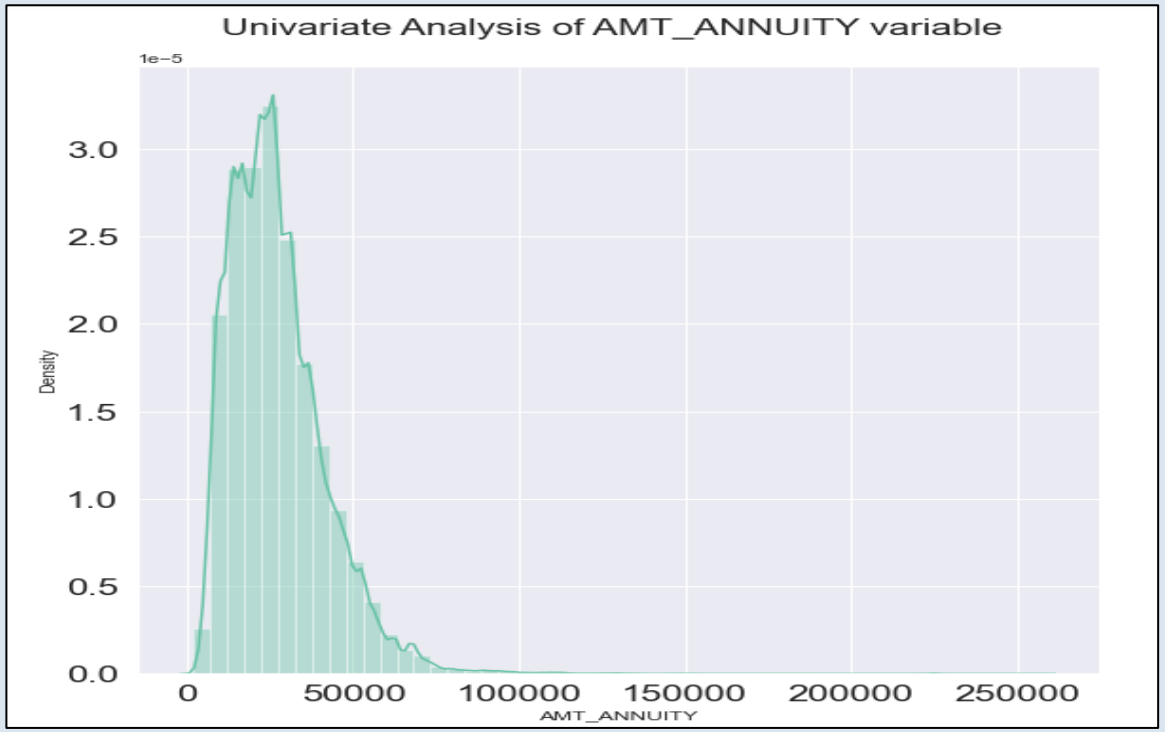
- From the given charts we can depict that the majority of the loan applicants are ‘Working’.
- Majority of the applicants are have having ‘low income’ followed by ‘high income’.
- Although maximum applicants have applied for Cash loans when compared to revolving loans.

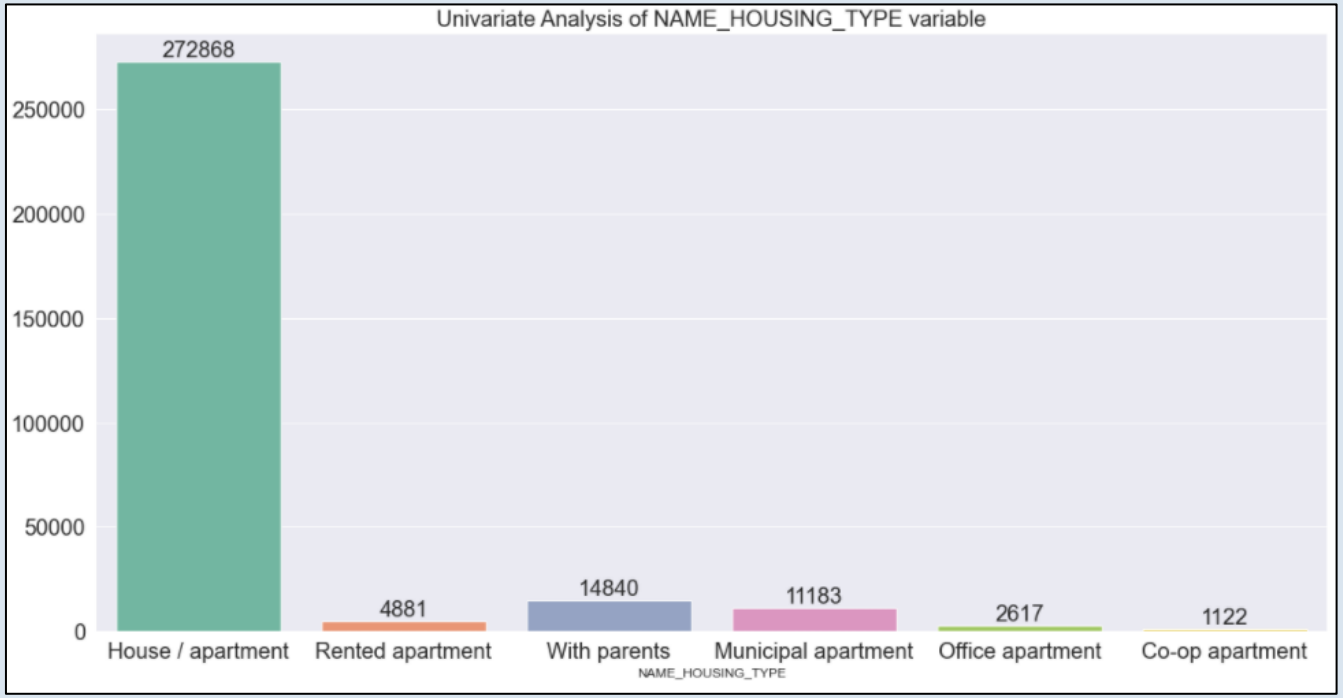


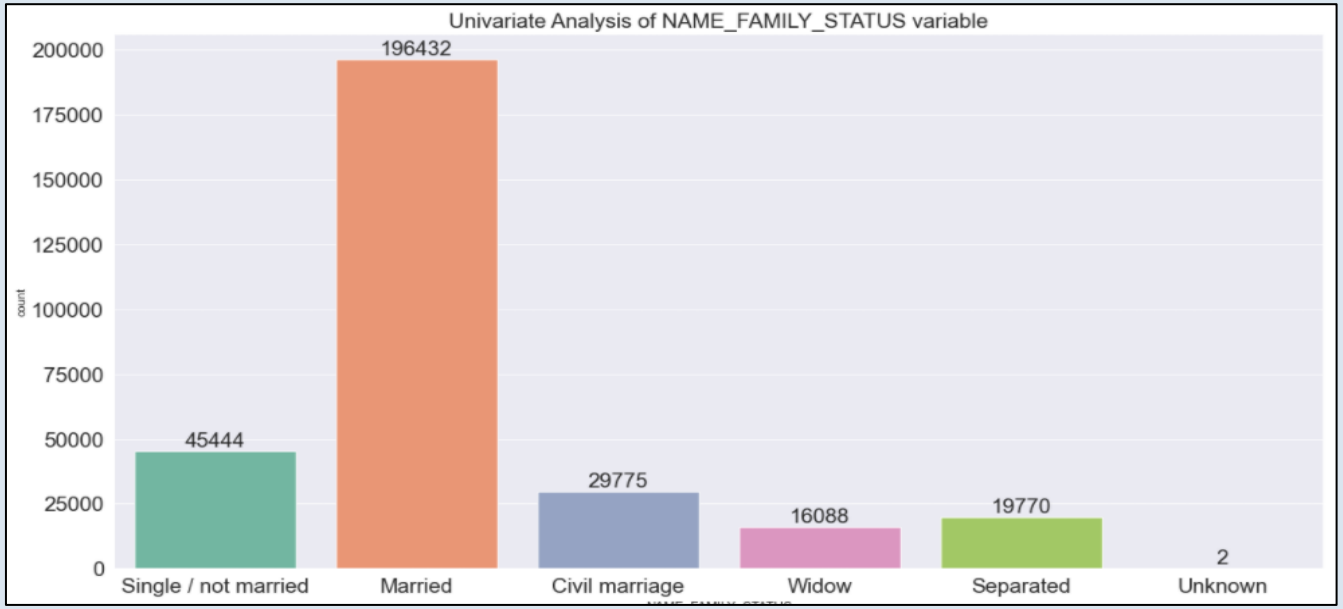




- From the given charts we can depict that maximum applicants reside in House/Apartment and are married.
- Also the Numeric variable ‘AMT_ANNUITY’ shows that major number of applicants have Loan annuity between 20k-40k.

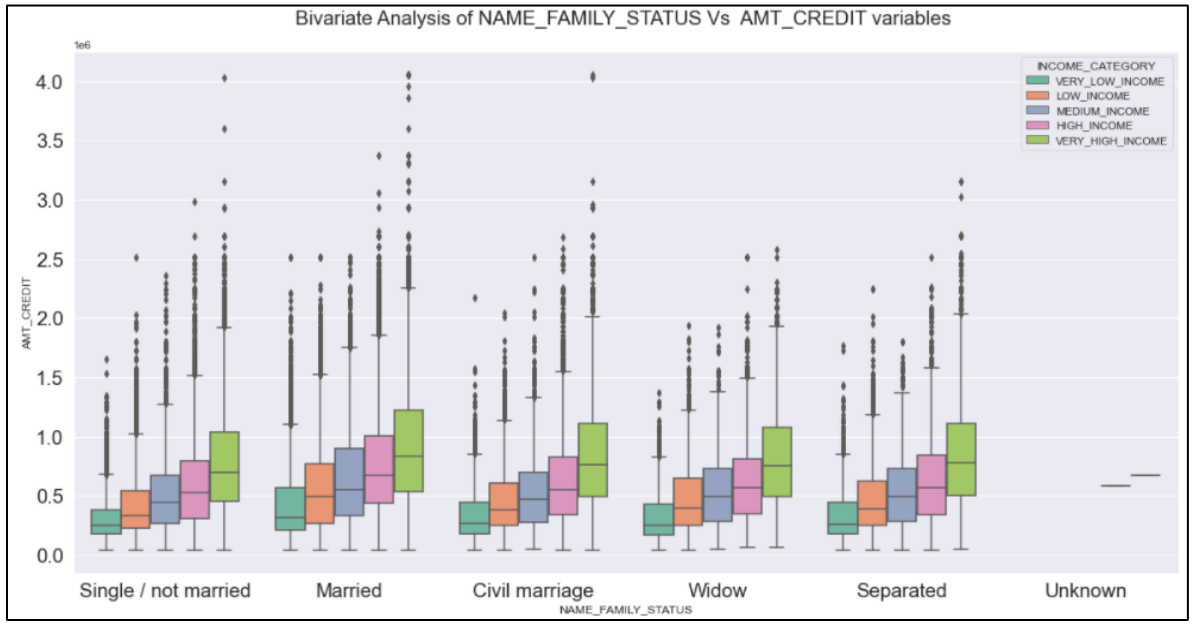


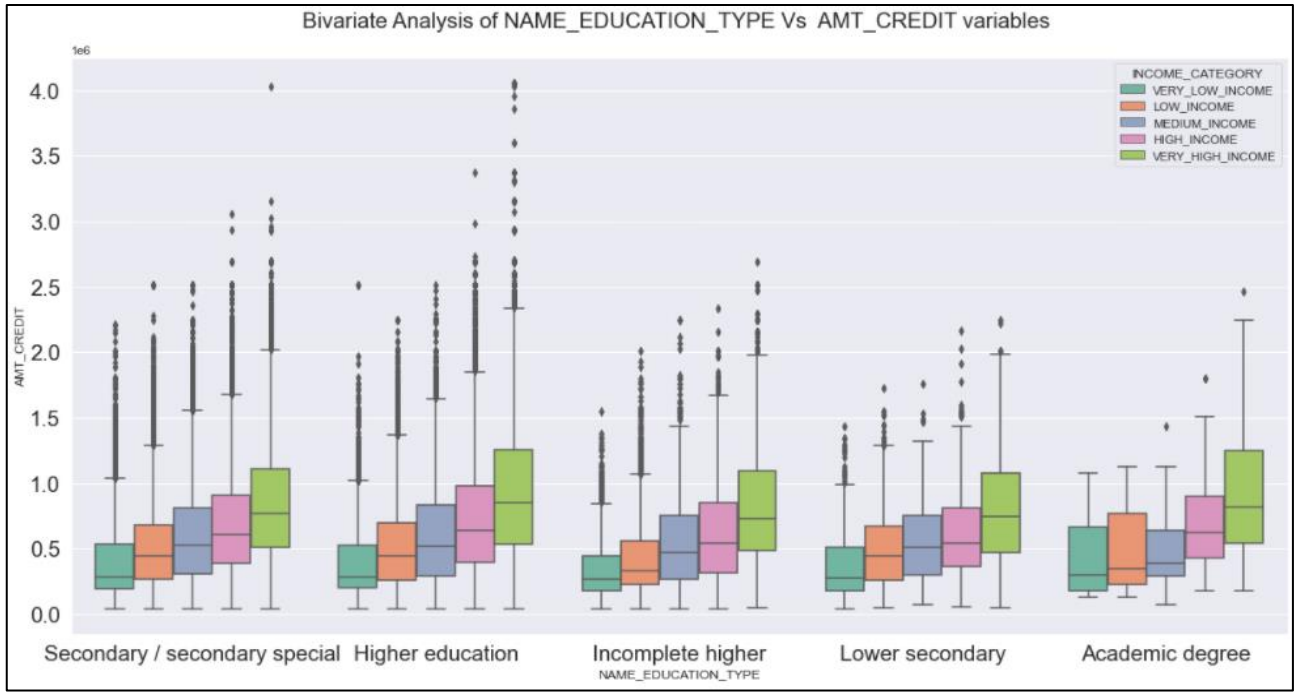




Bivariate Analysis

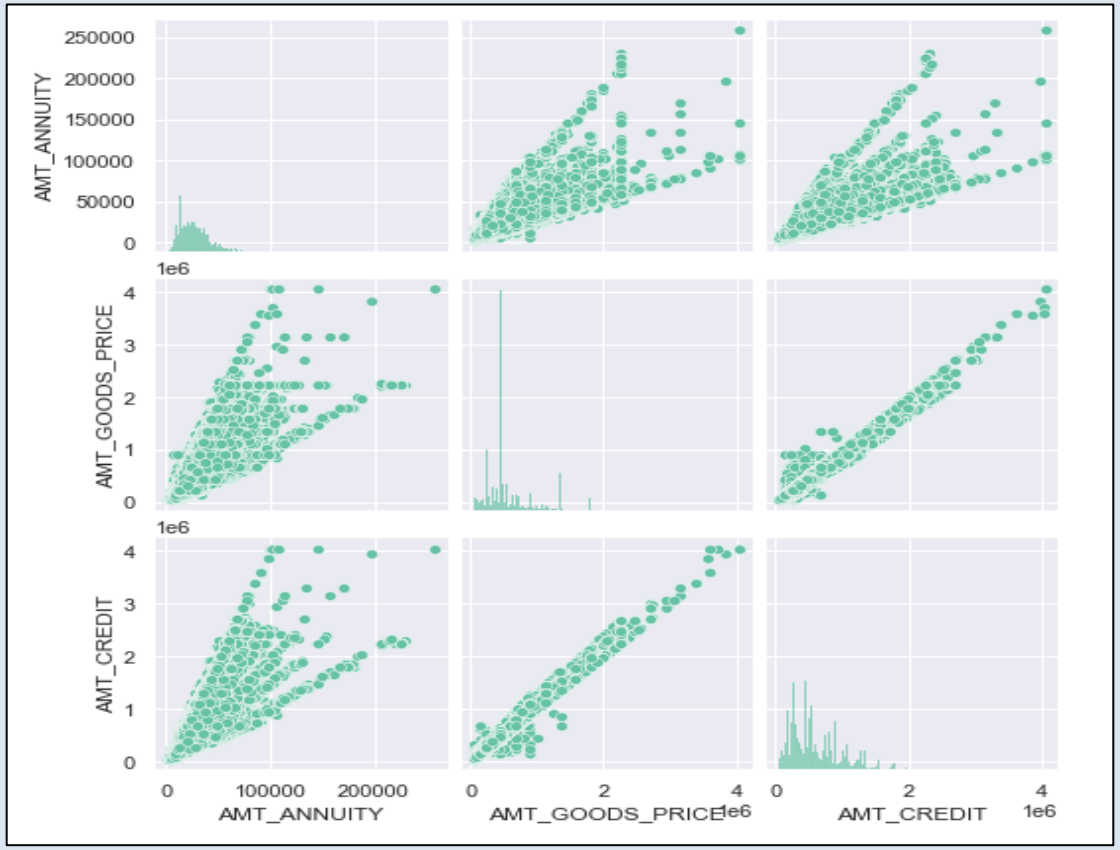
- From the given chart on NAME_FAMILY_STATUS variable we could see that ‘Married’ applicants have high credit allotted across all income ranges.

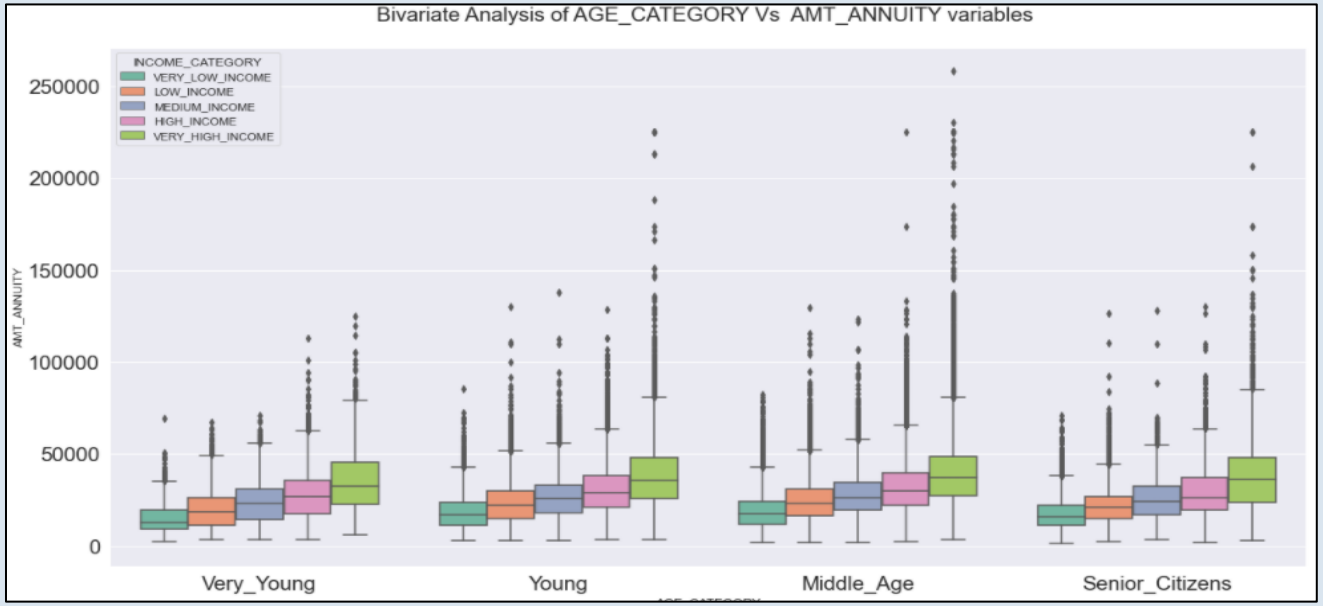


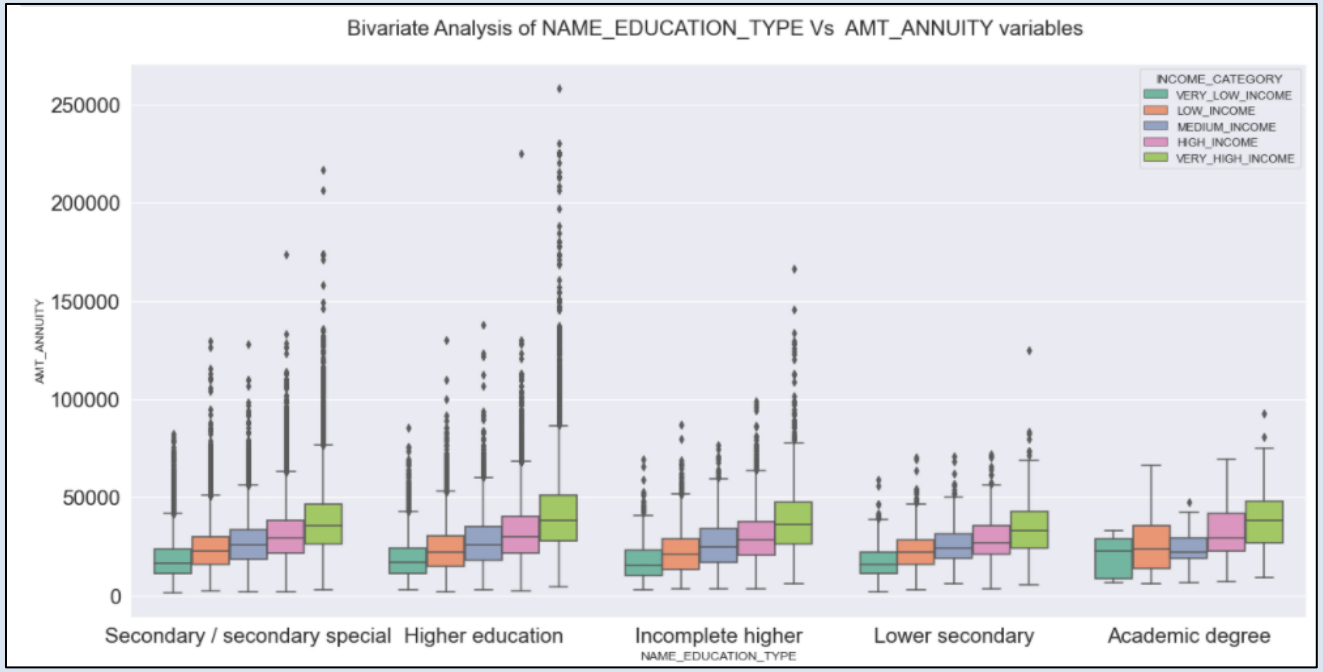


- From the given chart on NAME_EDUCATION_TYPE variable we could see that Higher Education and Academic degree categories have higher credit amounts allot across all income ranges.

- AGE_CATEGORY: Very_Young and Young applicants have slightly higher loan annuity.
- NAME_EDUCATION_TYPE: We can conclude that Higher Education and Secondary Education categories have higher loan annuity allotted.
- Pairplot: we can see there is slightly higher positive correlation amongst the AMT_GOODS_PRICE & AMT_CREDIT variables compared to other numeric variables.

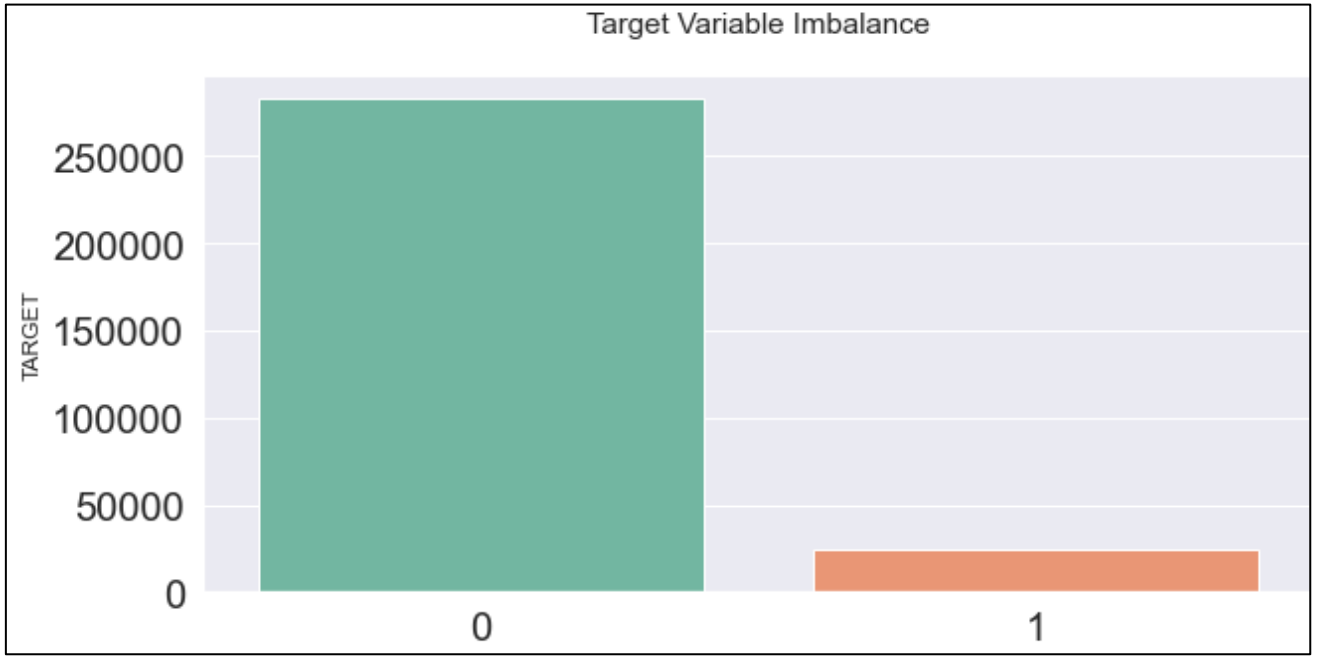






Segmented Analysis

- Before diving deep into segmented analysis I have drawn out the chart for target variable to check the data imbalance as we can see 90% of the applicants are non-defaulters and 10% are defaulters thus, clear data imbalance.

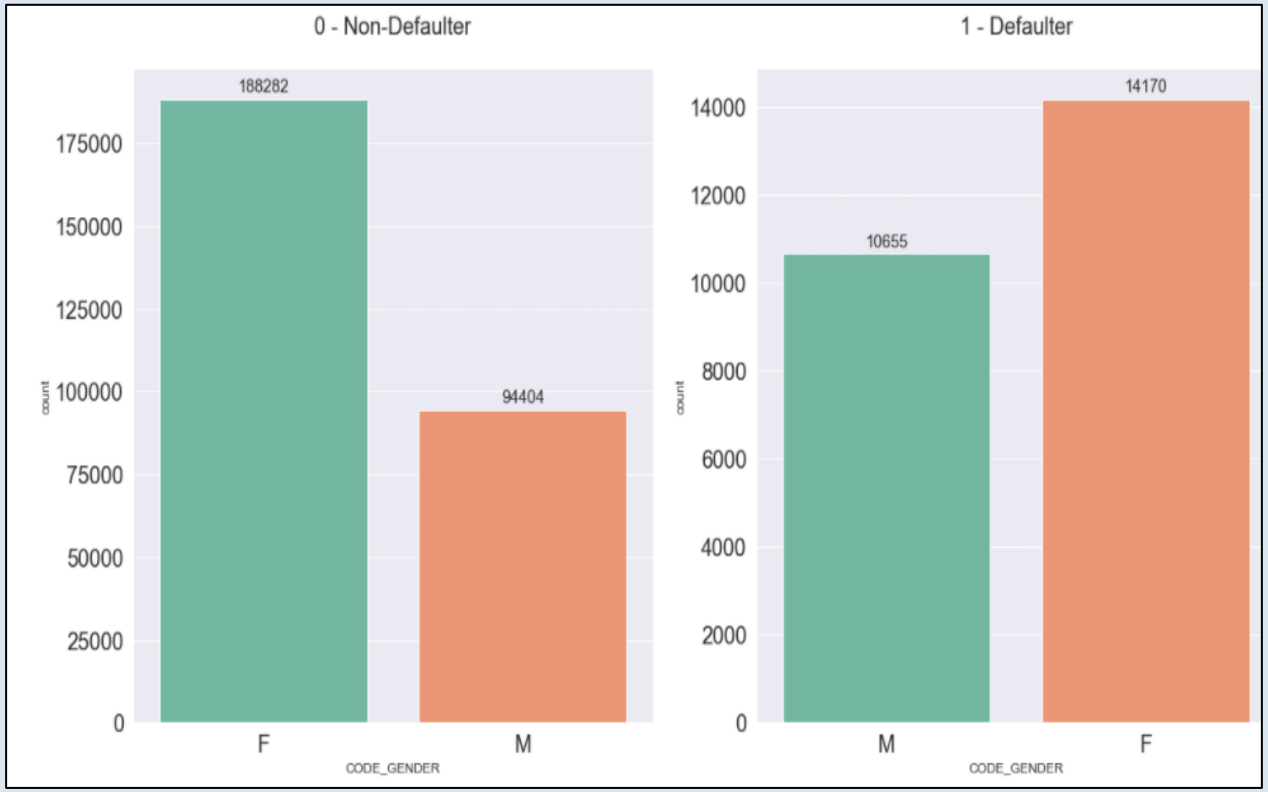


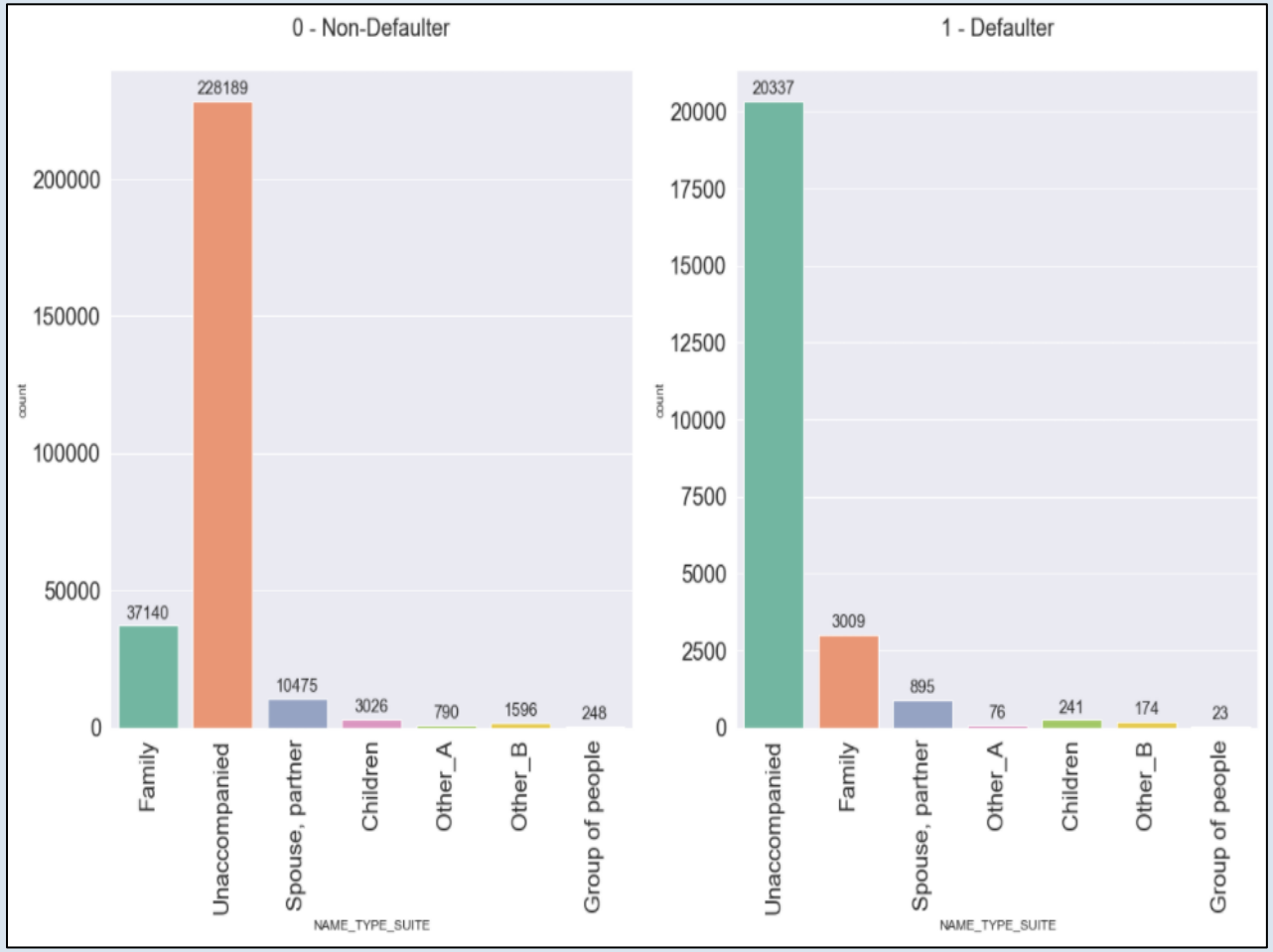
- **<u>Target Variable Imbalance</u>**:

Target variable: (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)

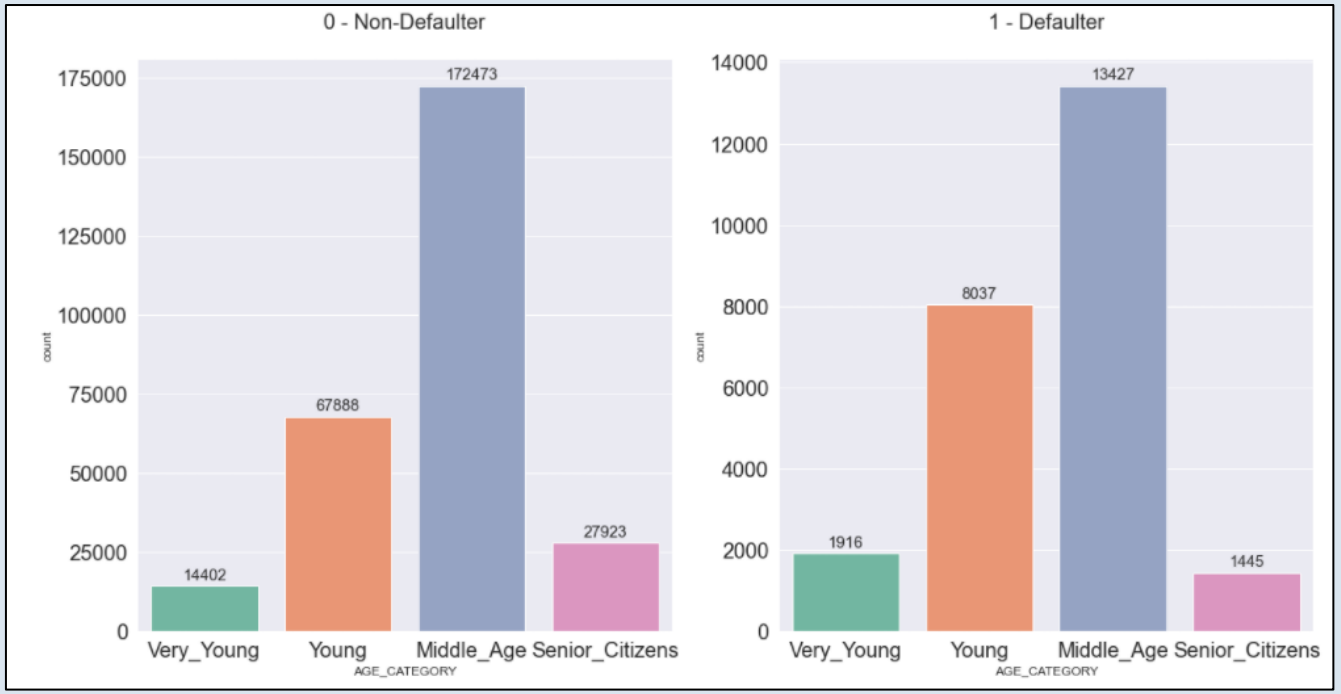
Segmented-Univariate Analysis

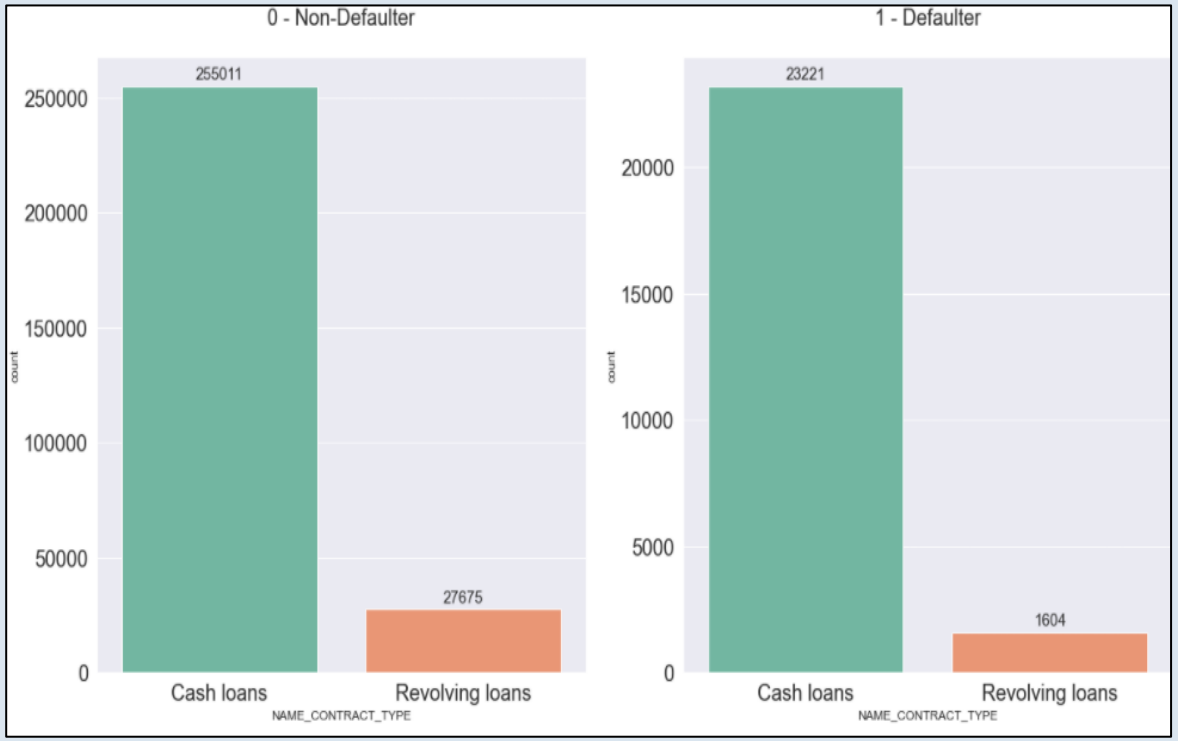
- From the given charts we can see that "Unaccompanied" applications are high defaulters & non-defaulters.
- From the below CODE_GENDER chart we can see that Female applications are high in both Defaulters & Non-defaulters but Male applicants are higher in Defaulters.

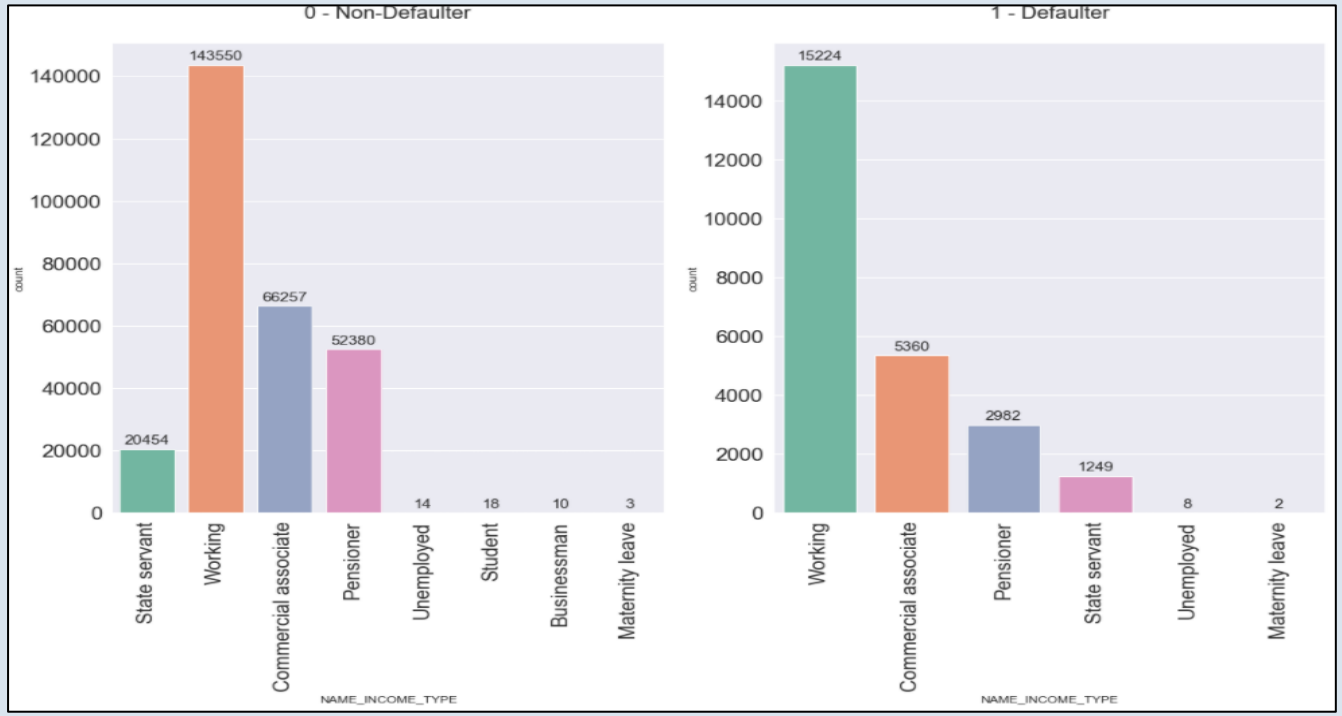




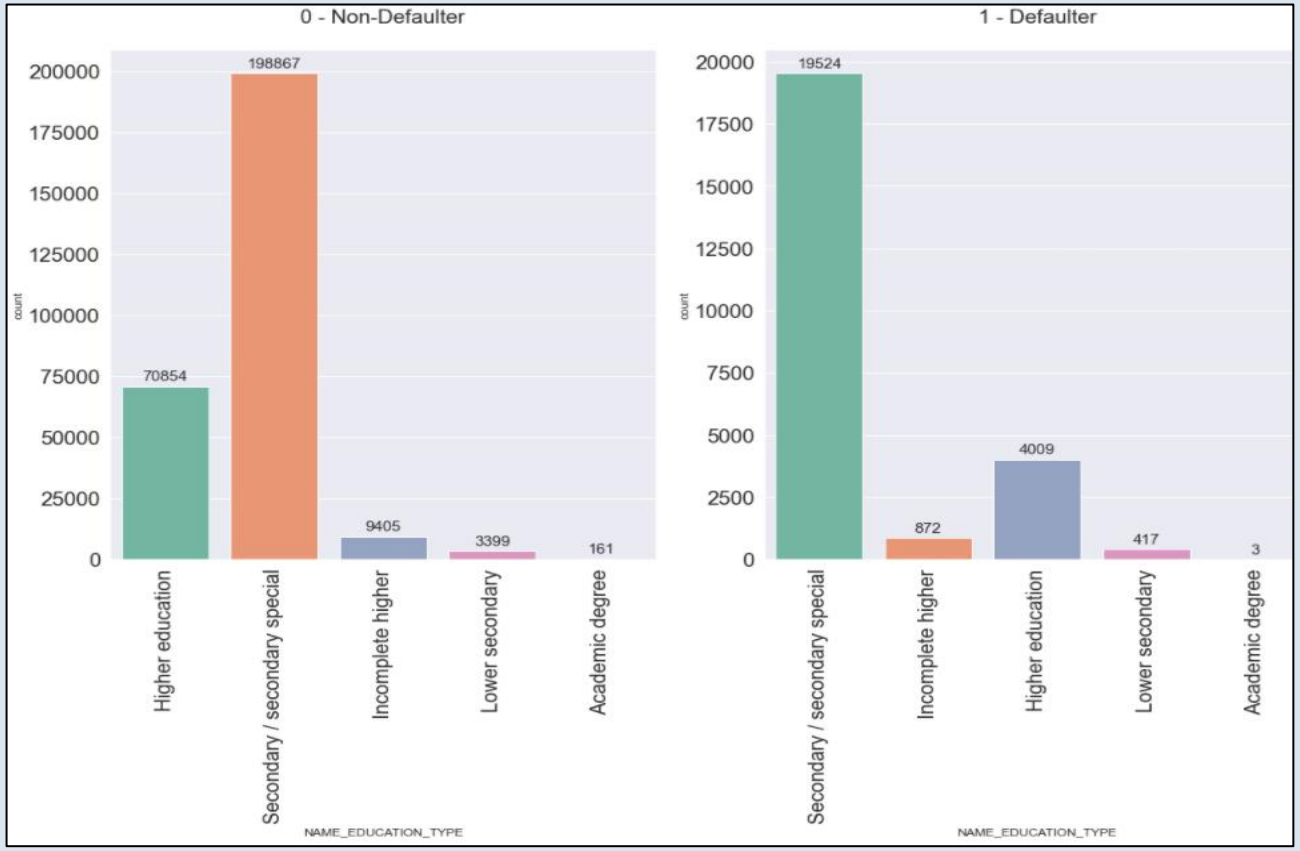
- From the given graphs we can conclude that Young & Middle_Age applicants have high possibility of payment difficulties.
- Moreover more defaulters can be expected for applications for cash loans vs revolving loans.
- We can see that there is increase in Working applicants tend to default but there is a decline in Pensioner applicants to default

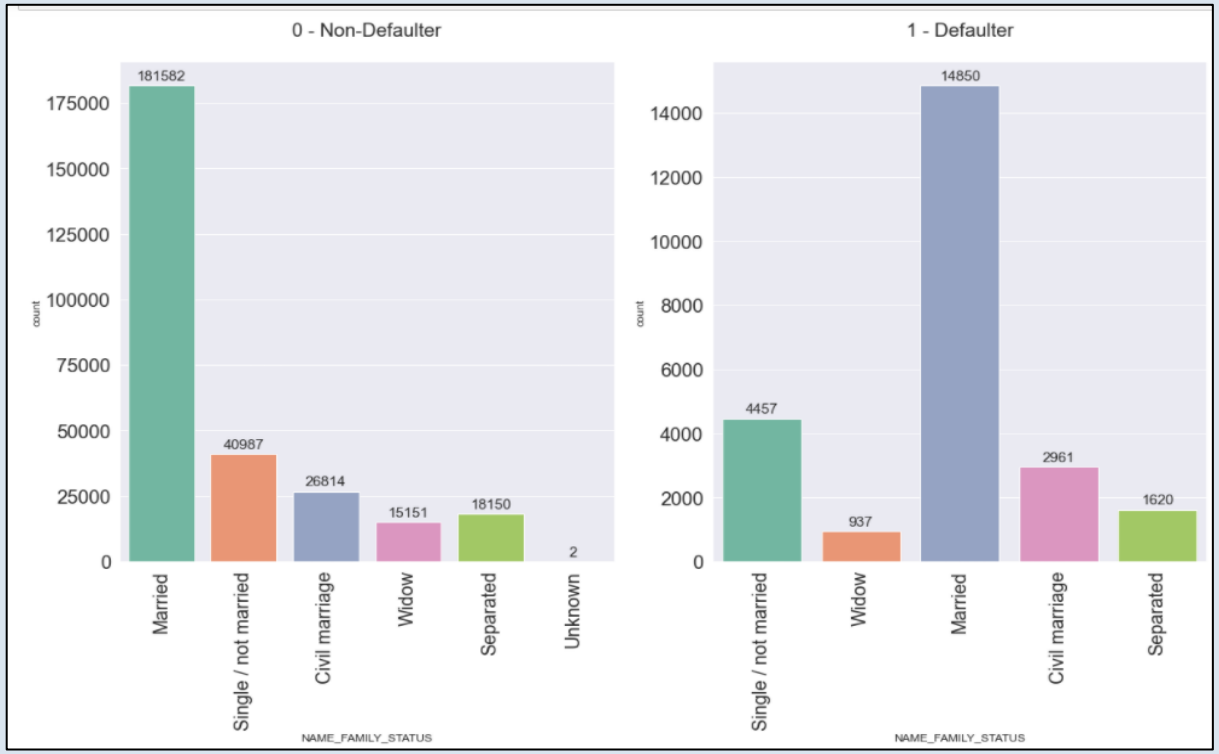


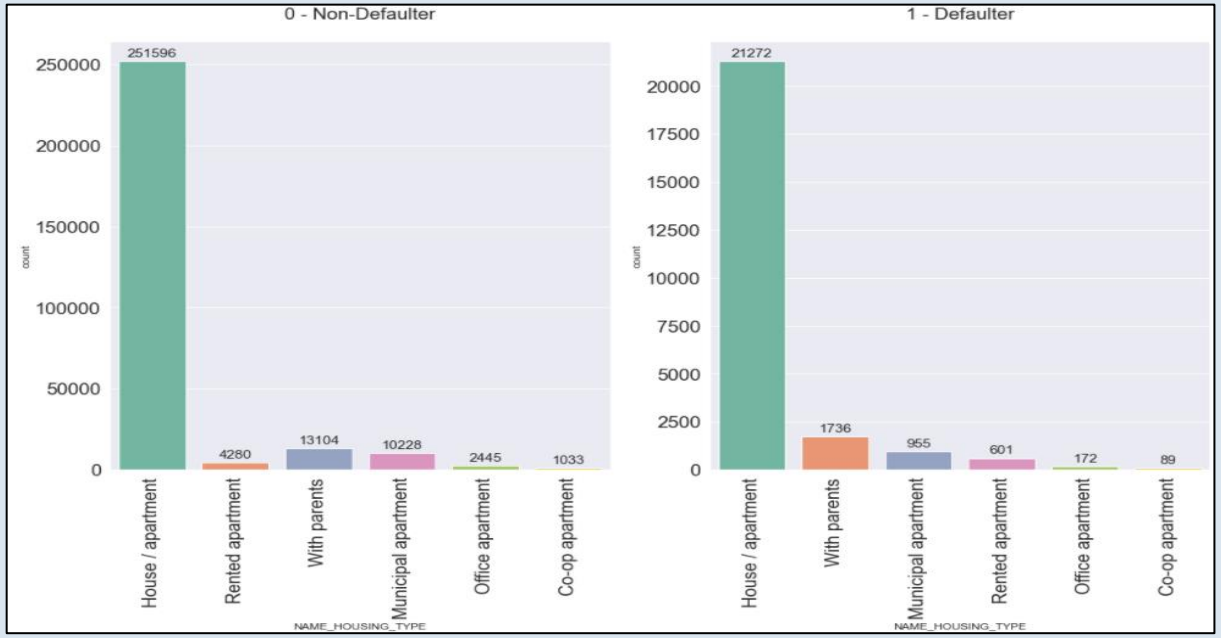




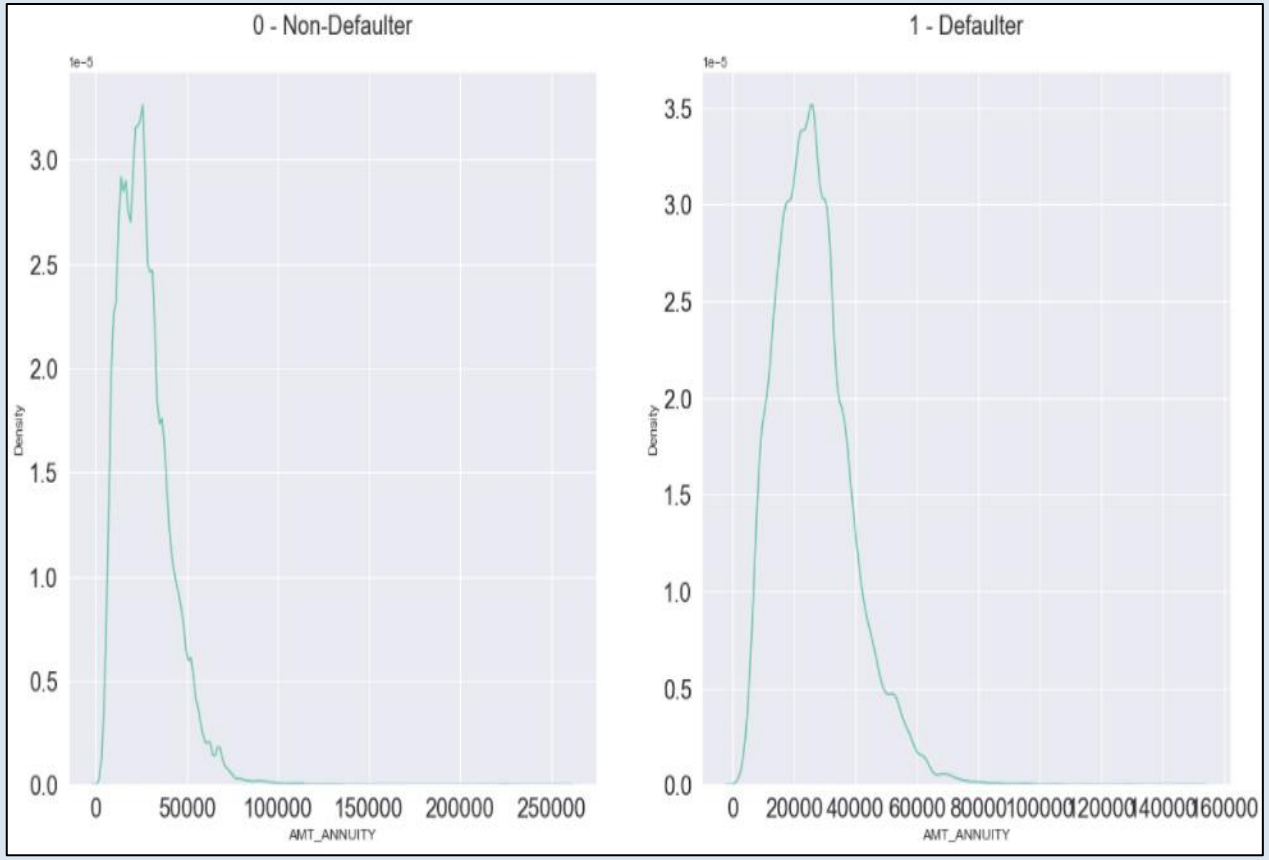
- From the given graphs we can see applicants who are having secondary education will tend to default or might have payment difficulties.
- From the above observation, There is a slight increase in "Civil marriage" applicants to default the loan, and slight decrease in "Widow" to default the loan, apart from the "Married“.
- We can see that there is increase in applicants to default who live with their parents.

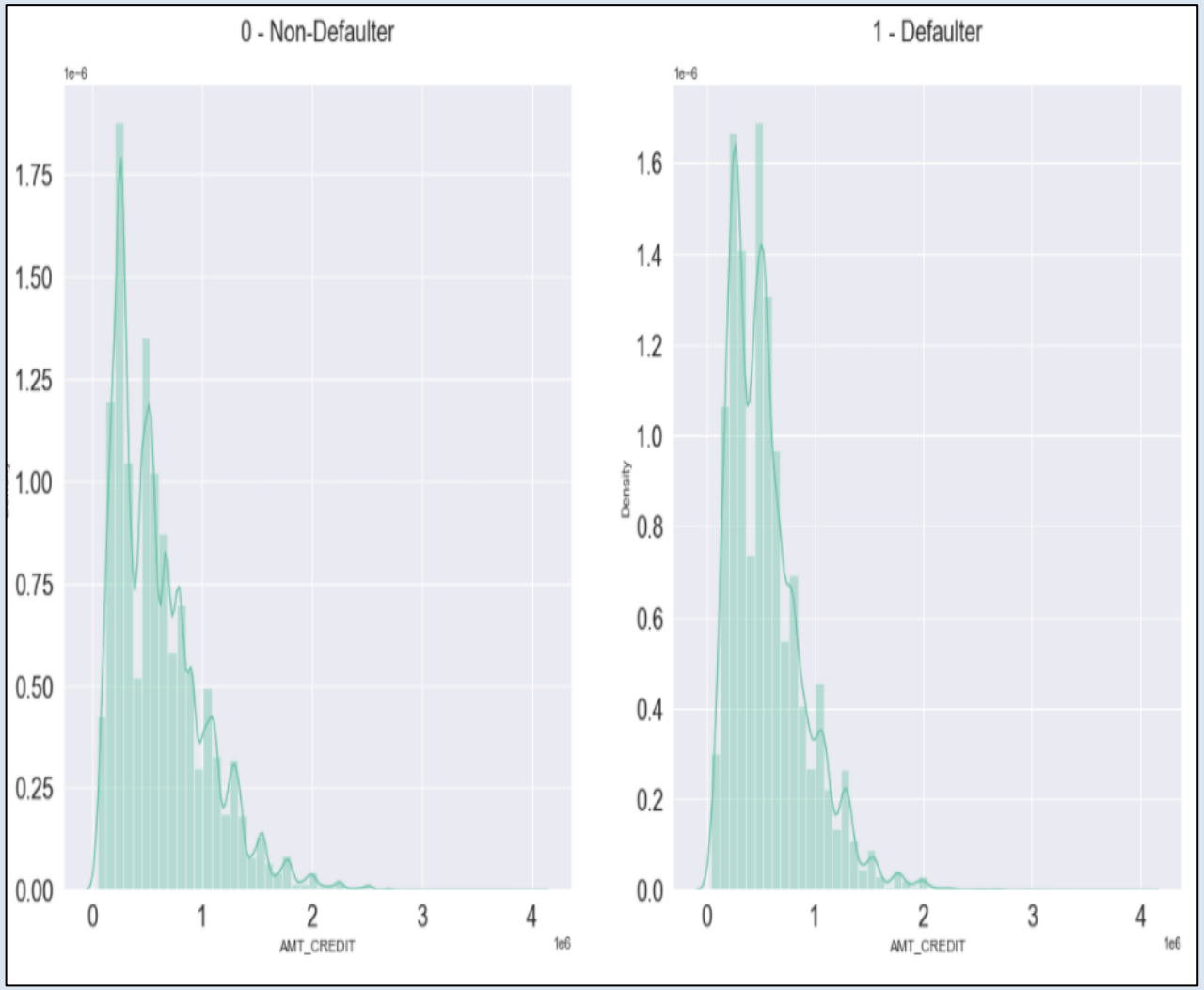






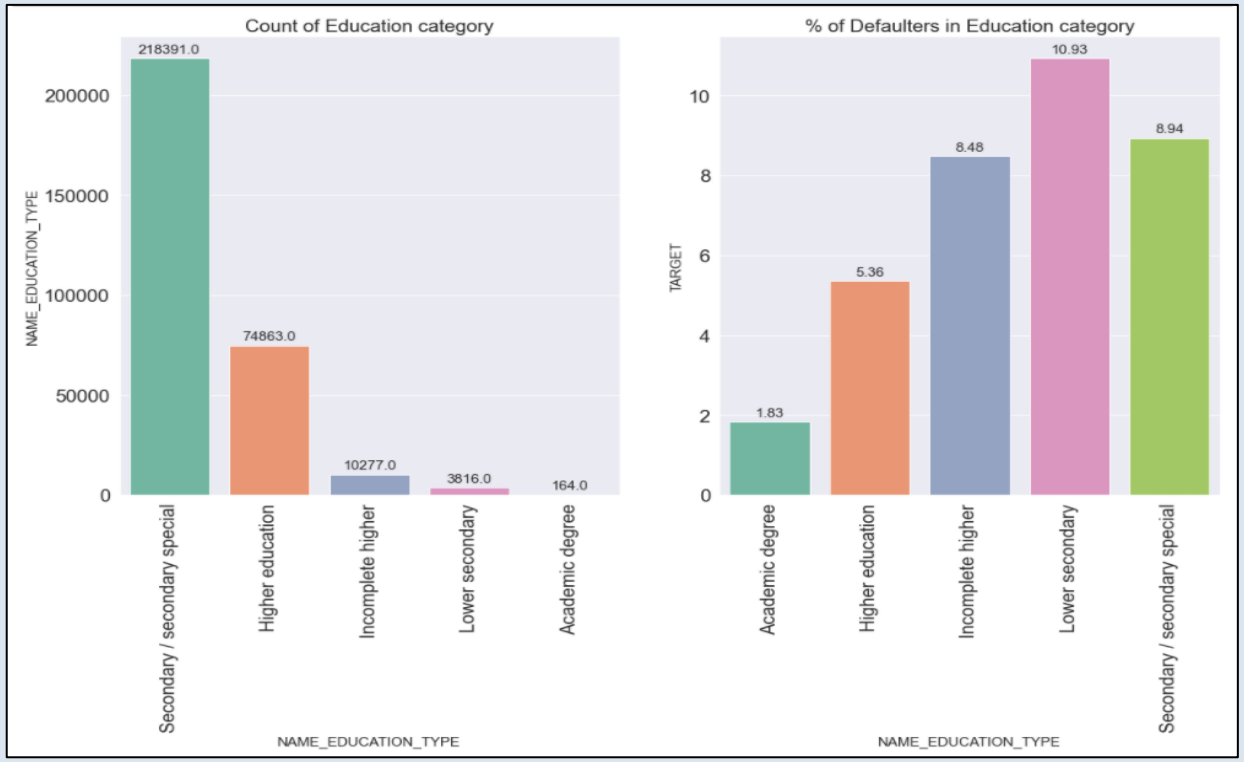
- From the below chart, we can observe that more defaulters exist between 20k to 40k bin on loan annuity and more applicants with no payment difficulties exist at loan annuity of 50k approximately.
- From the chart on the right, we can see that the data is left-skewed and not perfectly distributed, more defaulters exist between the 0 & 1 bins possibly at 0.5.

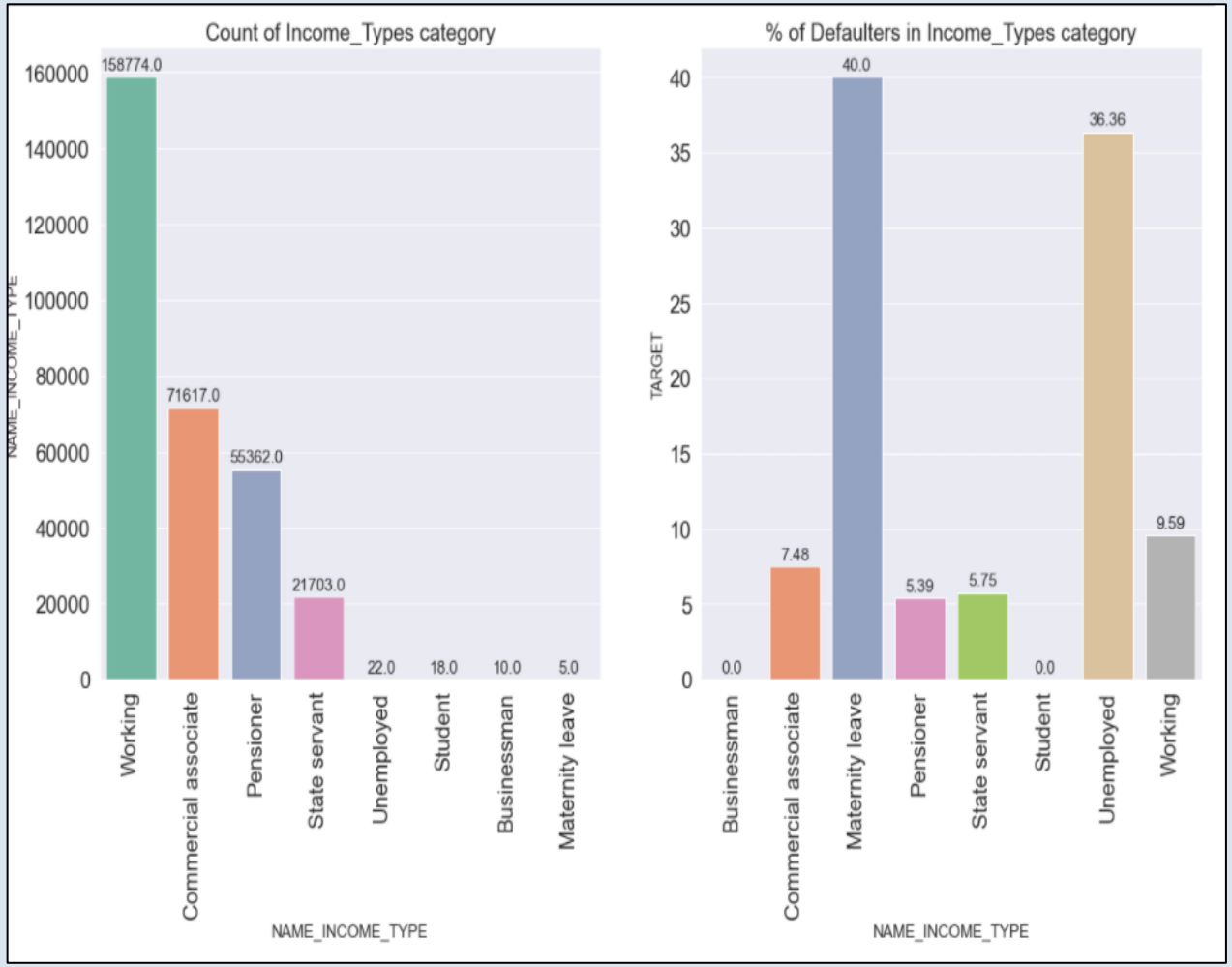


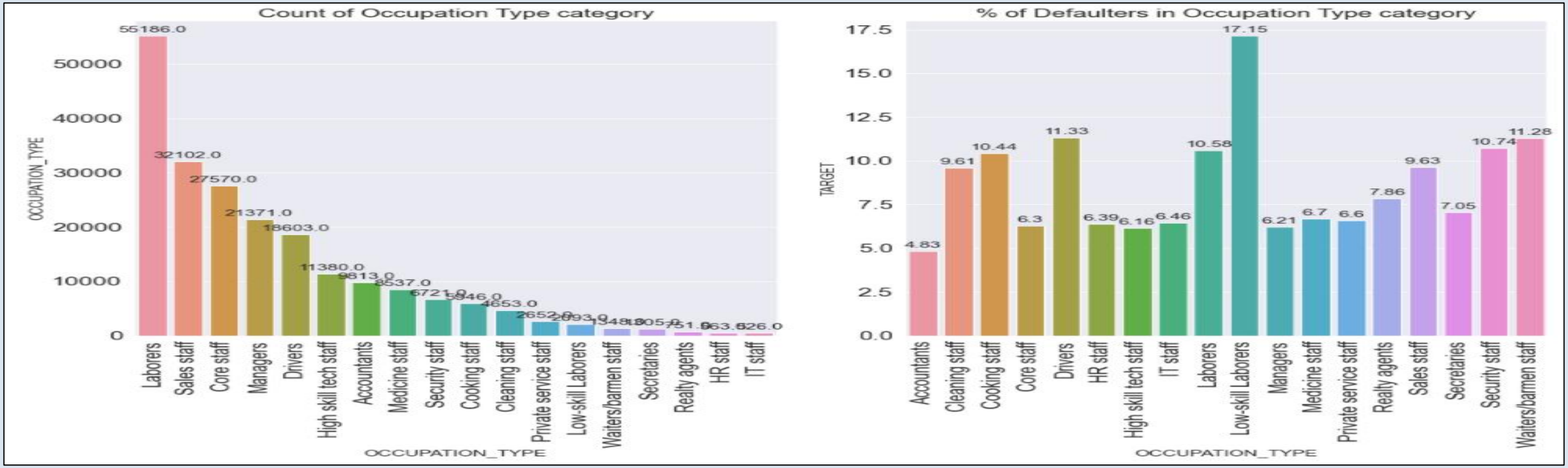


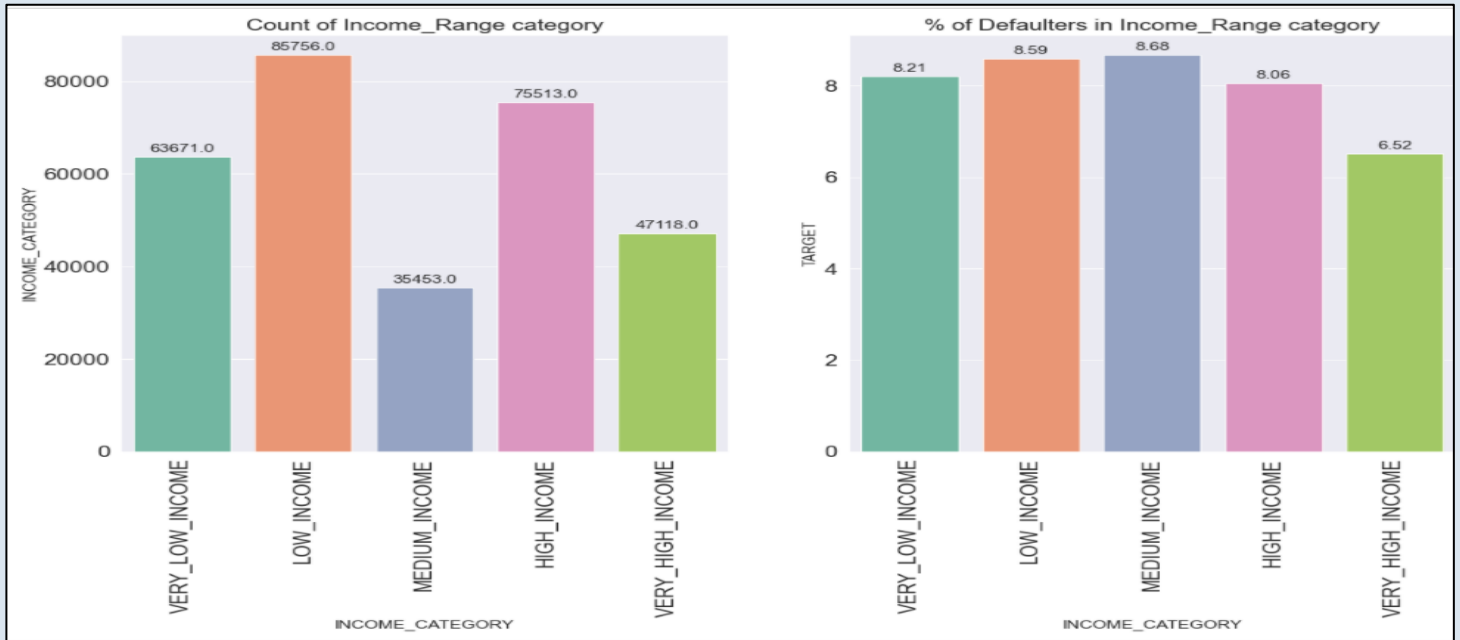
Segmented-Bivariate Analysis

- From the given charts we can observe that Working category has less defaulters inspite of high applications, but Maternity Leave category has maximum defaulters, so is the case with Unemployed category
- From the below, we can determine that Lower Secondary education category has maximum defaulters and minimum applications.



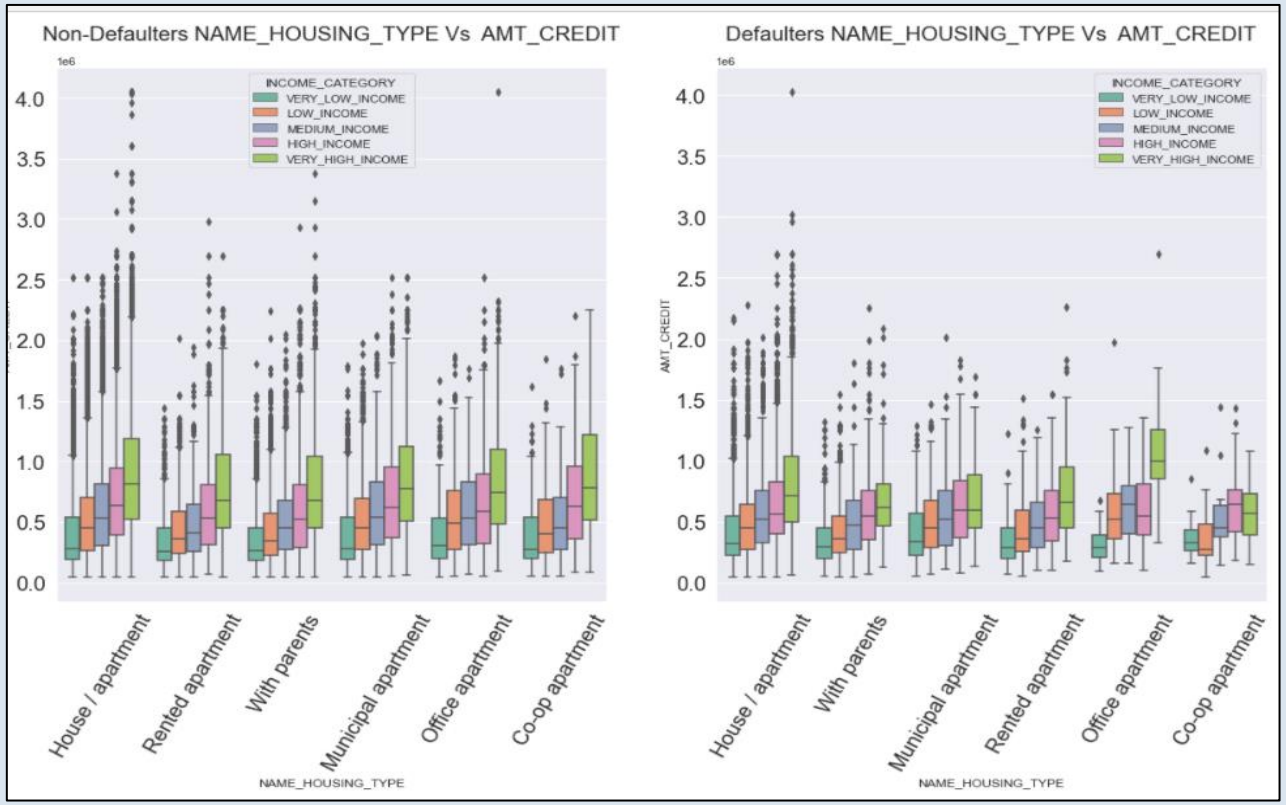


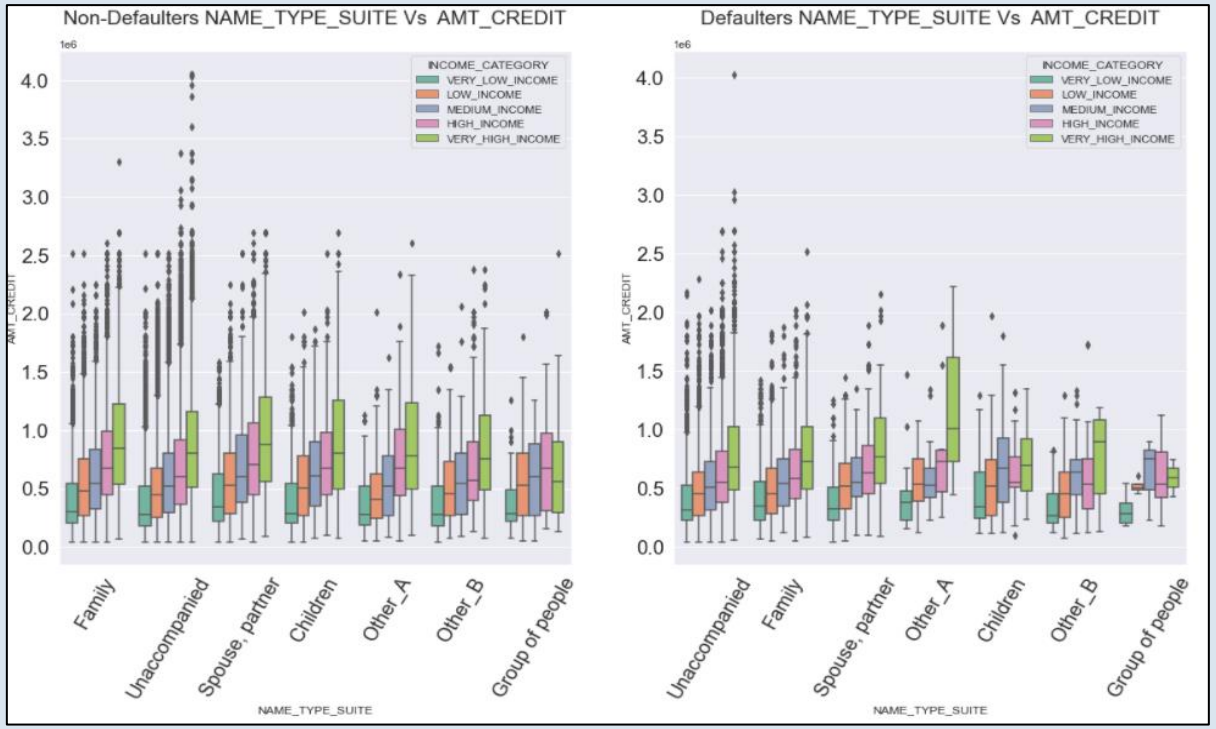




- From the above chart, we can see that Low-Skill Laborers are tend to default or have high difficulties in repaying the loan.
- From the chart on the left, we can see that the MEDIUM_INCOME & LOW_INCOME categories have maximum defaulters with high payment difficulties..

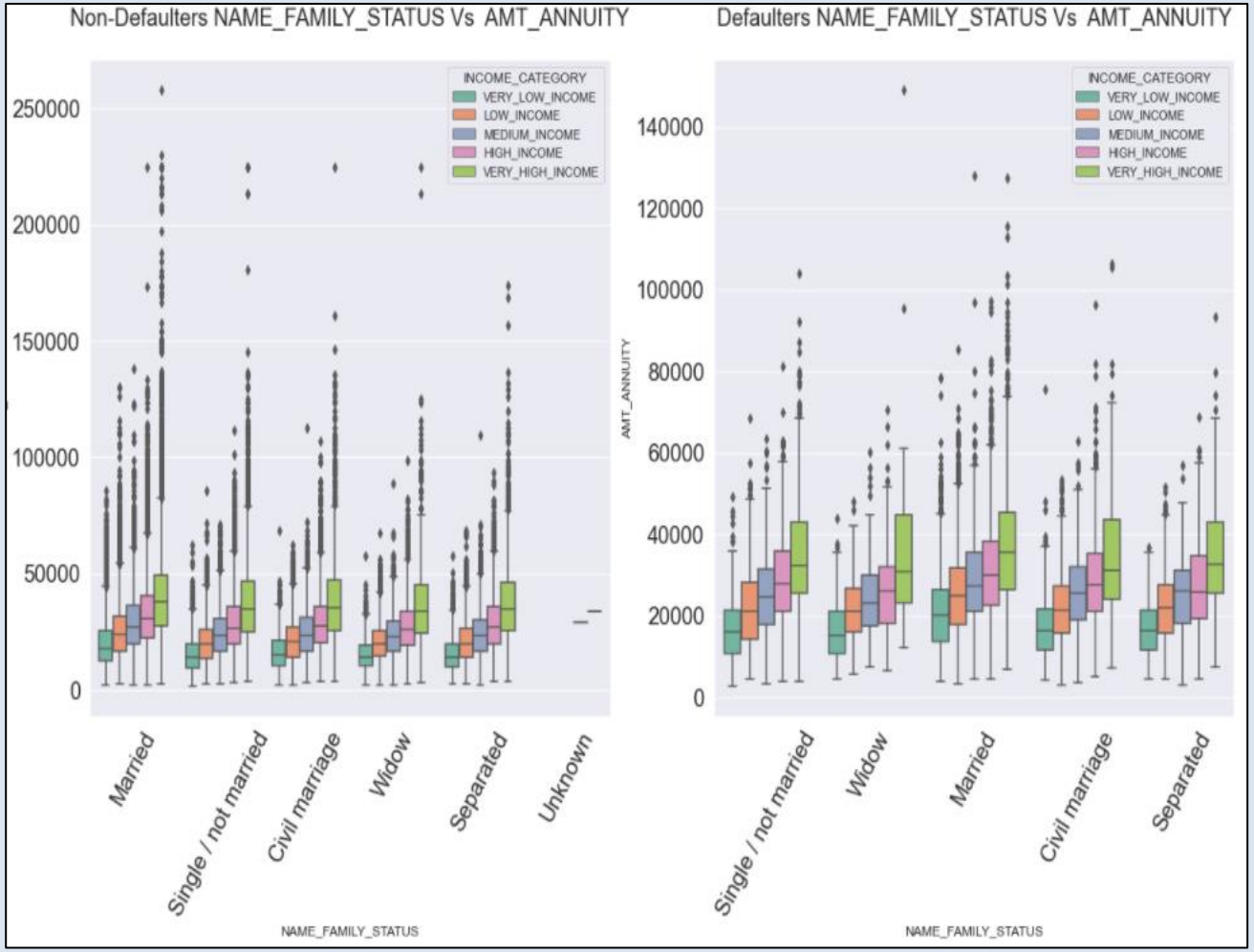
- From the graph on right, we can depict that across all income categories, the most of non-defaulters reside in House/Apartment & CO-OP apartment are given higher credits, but, higher credits are assigned to defaulters tend to reside in House/Apartment and Rented apartments .

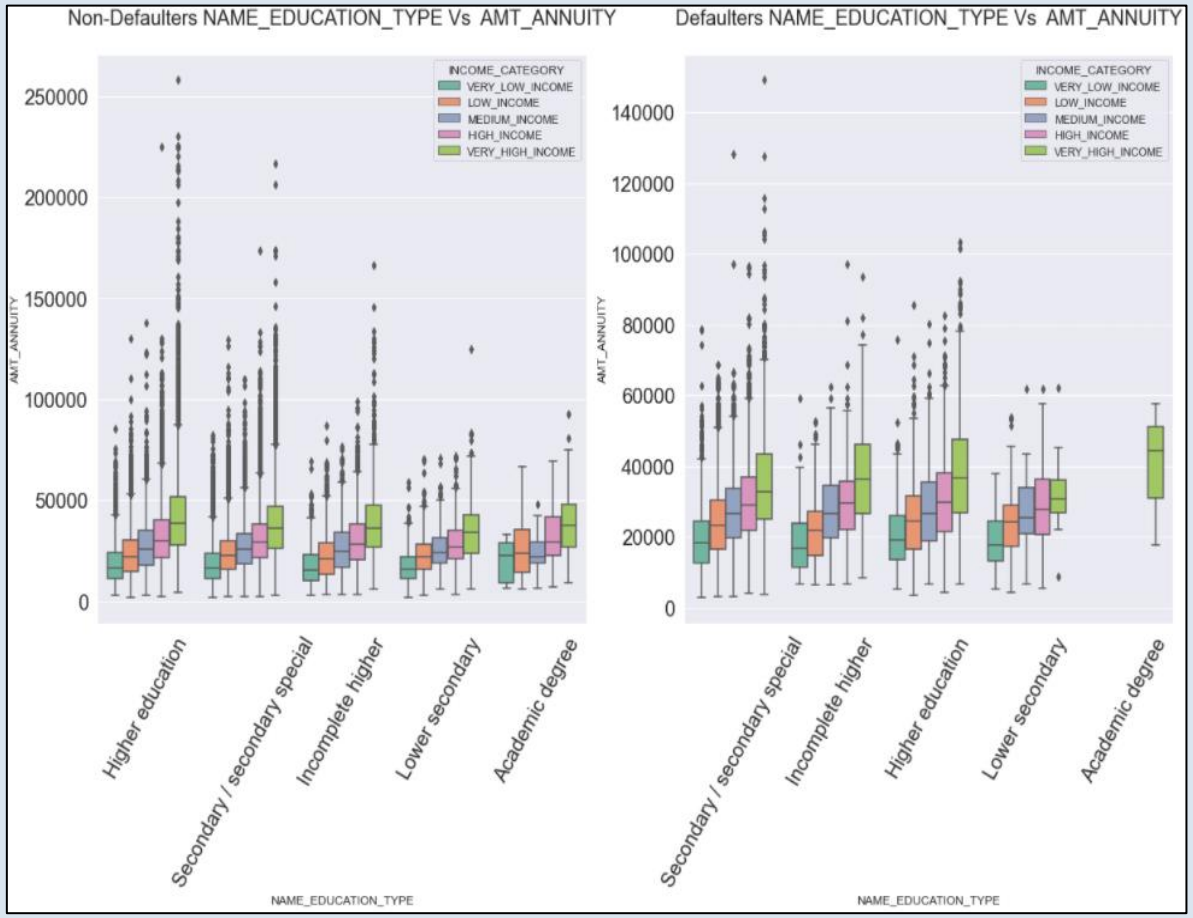




- From the graph on left, we can see that Major Non-Defaulter applicants which are accompanied by Spouse, Partner & Family are given high credits , but, Major Defaulters are given more credits which are Unaccompanied or Other_A categories across all income categories.

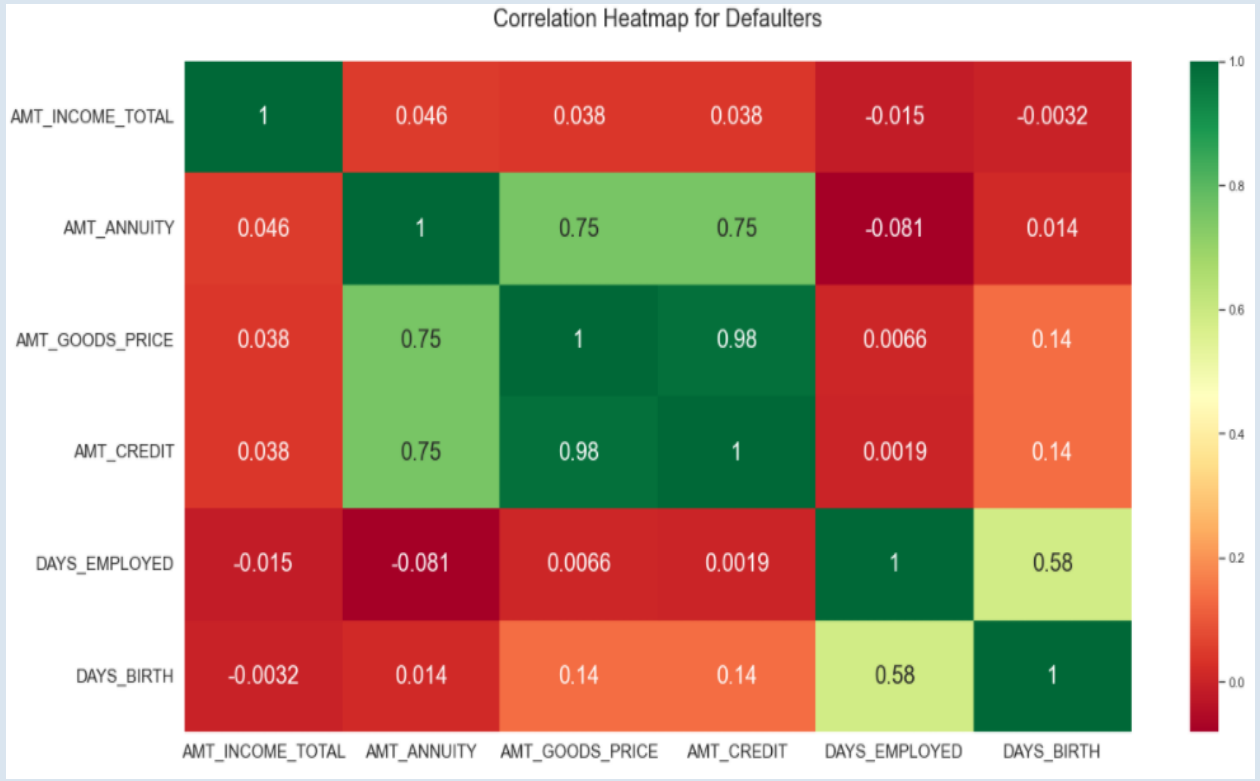
- From the graph on right, we can see that Widow-Defaulters are having comparatively higher Loan annuity and Non-defaulter's category is almost similar for all income ranges.

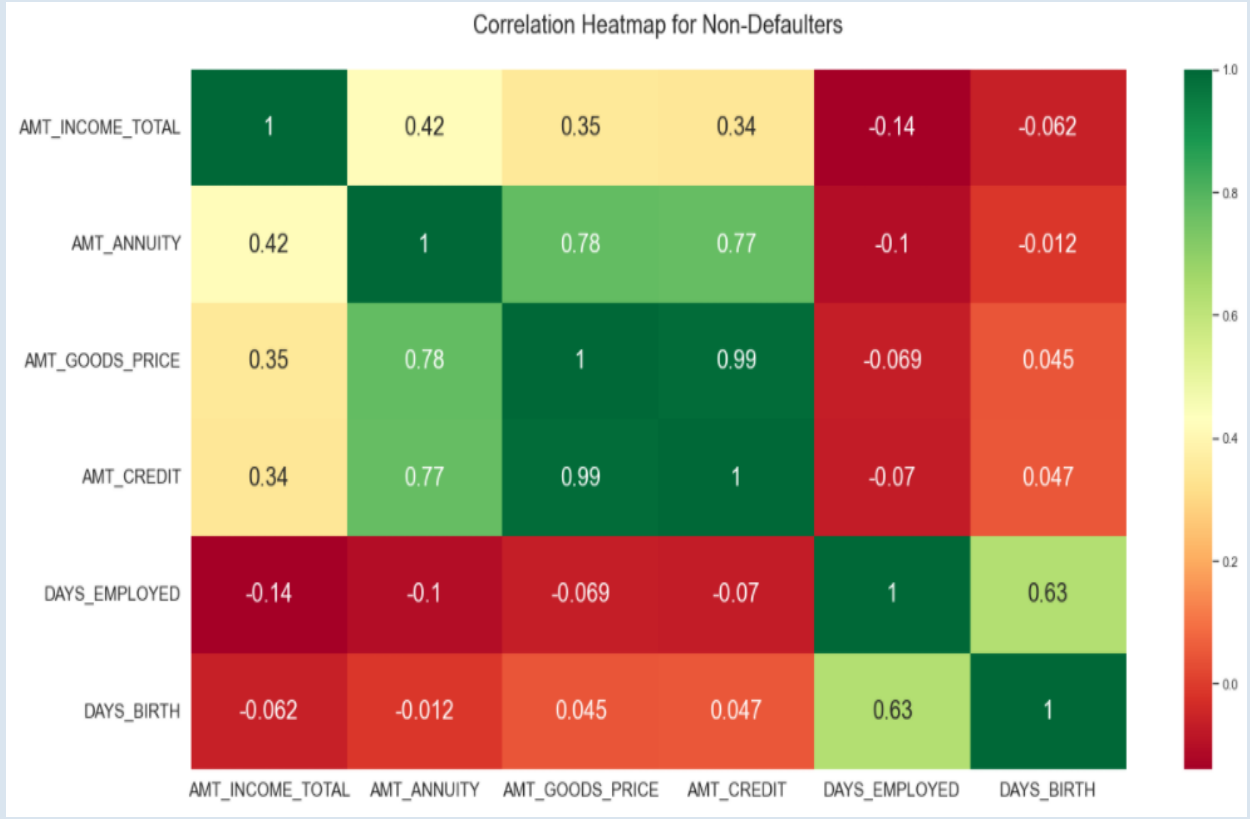




- From the graph on left, if you look closely, the Incomplete-higher & Higher education categories are assigned major loan annuity which has more defaulters across all income categories.

Correlations & Heatmaps





- From both the heatmap comparisons we can conclude that there is certain deviation for Goods_Price Vs Income_Total and Credit_Amount vs Income_Total, although there is a high positive correlation between Credit_amount & Goods_Price .

Top 10 Target Variables

	Attribute_1	Attribute_2	Correlation_Score
87	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637
10	AMT_ANNUITY	AMT_GOODS_PRICE	0.752699
53	DAYS_BIRTH	DAYS_EMPLOYED	0.582441
65	DAYS_REGISTRATION	DAYS_BIRTH	0.289116
63	DAYS_REGISTRATION	DAYS_EMPLOYED	0.192455
50	DAYS_BIRTH	AMT_GOODS_PRICE	0.135603
94	LIVE_CITY_NOT_WORK_CITY	CNT_CHILDREN	0.053515
21	AMT_INCOME_TOTAL	AMT_ANNUITY	0.046421
84	REGION_RATING_CLIENT_W_CITY	CNT_CHILDREN	0.043185
74	REGION_RATING_CLIENT	CNT_CHILDREN	0.040680

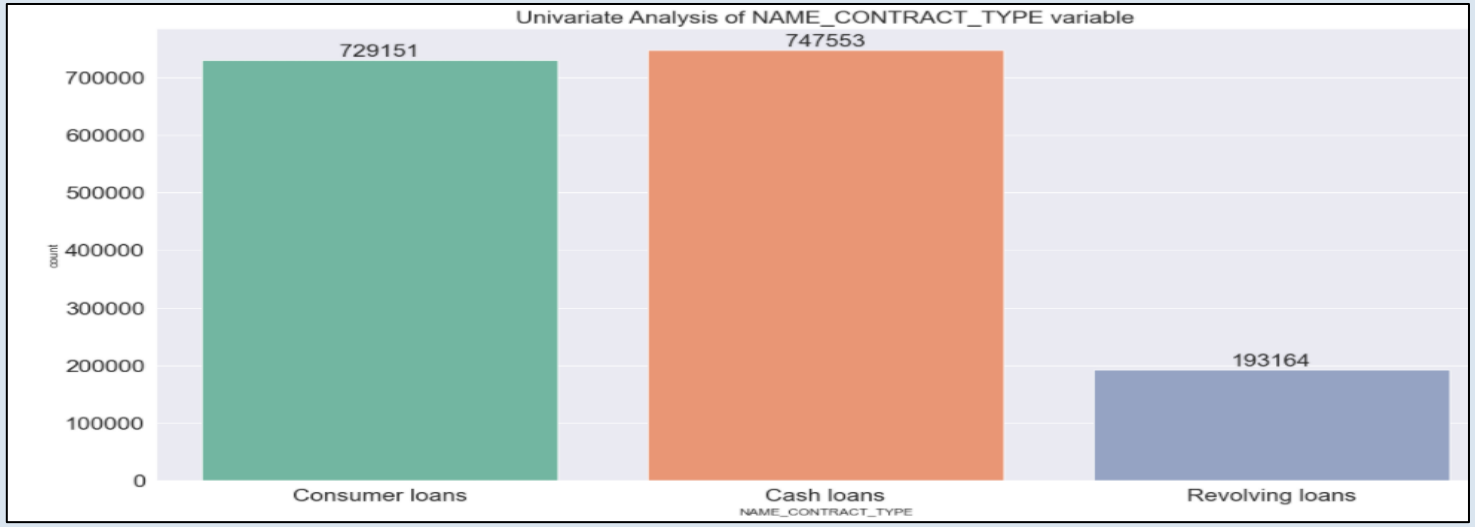
- So based on the left table, we can conclude the Top 10 Correlations for defaulters or clients with payment difficulties

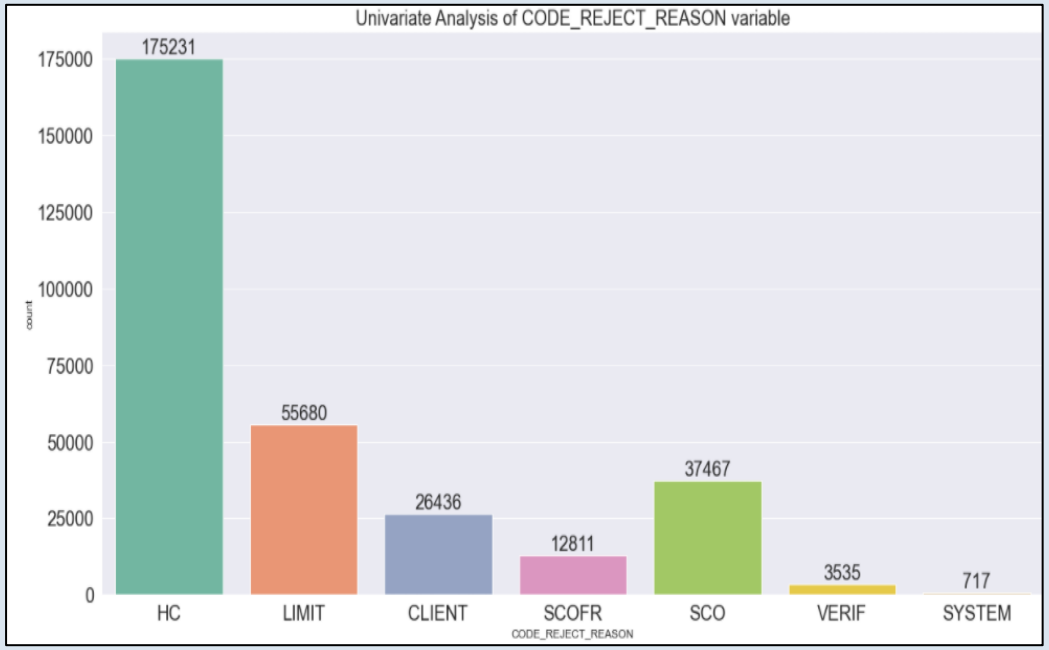
- So based on the right table, we can conclude the Top 10 Correlations for non-defaulters or clients with no payment difficulties

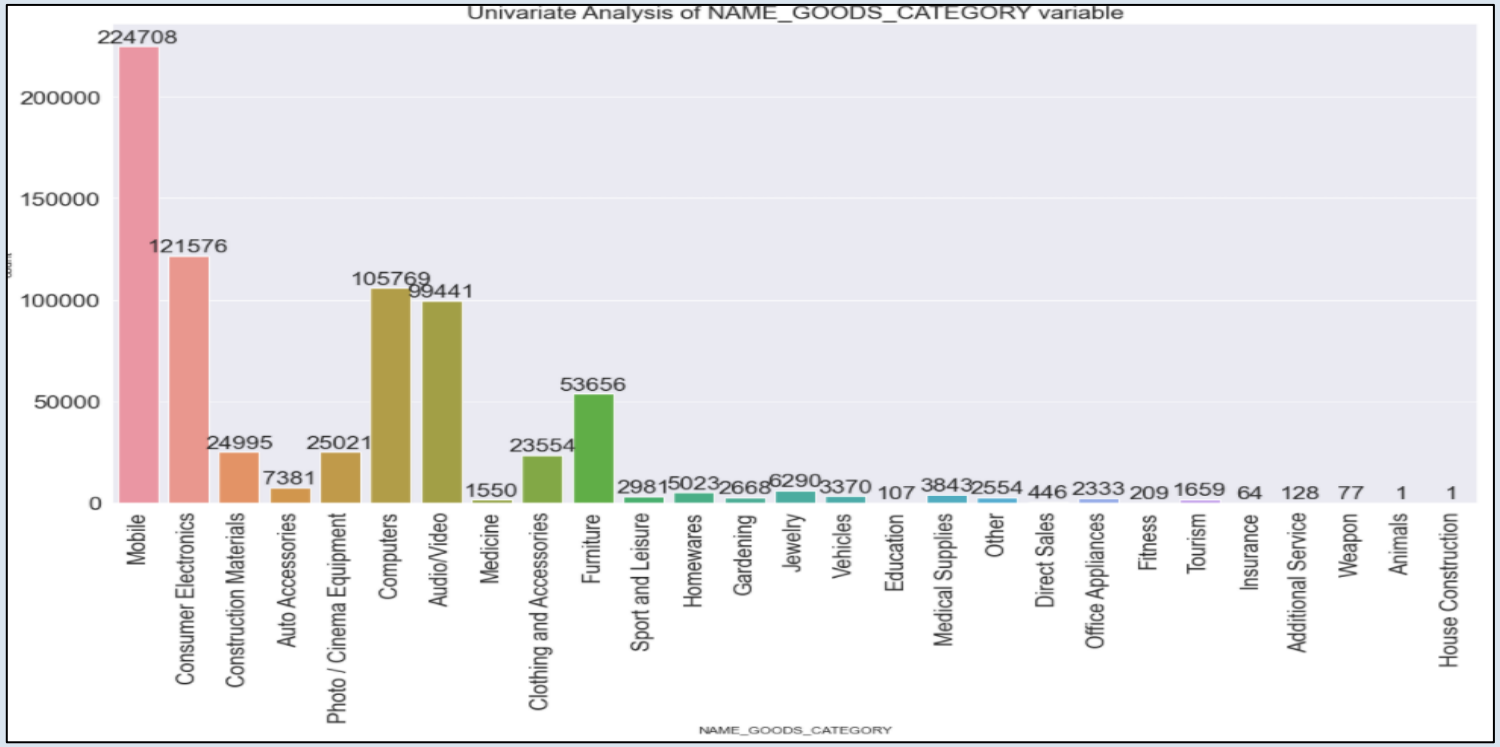
	Attribute_1	Attribute_2	Correlation_Score
87	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149
10	AMT_ANNUITY	AMT_GOODS_PRICE	0.776686
53	DAYS_BIRTH	DAYS_EMPLOYED	0.626028
21	AMT_INCOME_TOTAL	AMT_ANNUITY	0.418953
20	AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.349462
65	DAYS_REGISTRATION	DAYS_BIRTH	0.333025
63	DAYS_REGISTRATION	DAYS_EMPLOYED	0.214511
94	LIVE_CITY_NOT_WORK_CITY	CNT_CHILDREN	0.070988
50	DAYS_BIRTH	AMT_GOODS_PRICE	0.044552
83	REGION_RATING_CLIENT_W_CITY	DAYS_EMPLOYED	0.040461

Analysis on Previous Application data

- We can observe that major loans are for Consumer/Cash loans.
- From the below, we can see that Majority of the applications got rejected because of 'HC'.
- From the bottom right graph, we can conclude that majority of the loan applications are for Goods - Consumer Electronics, Computers & Mobile.

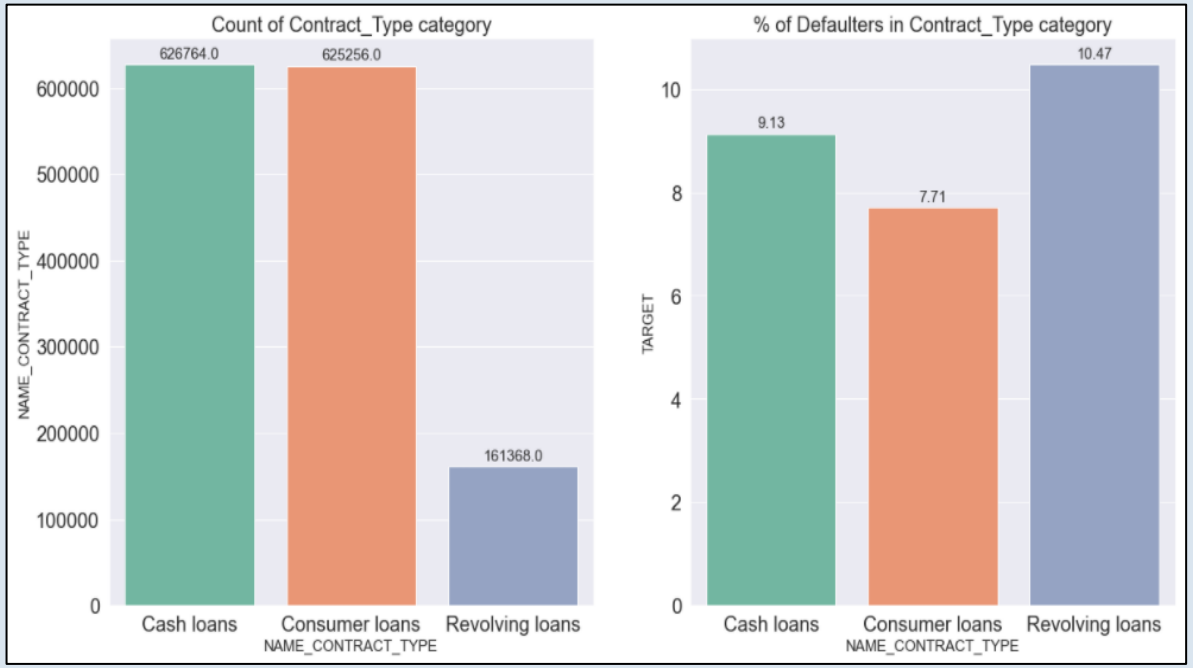


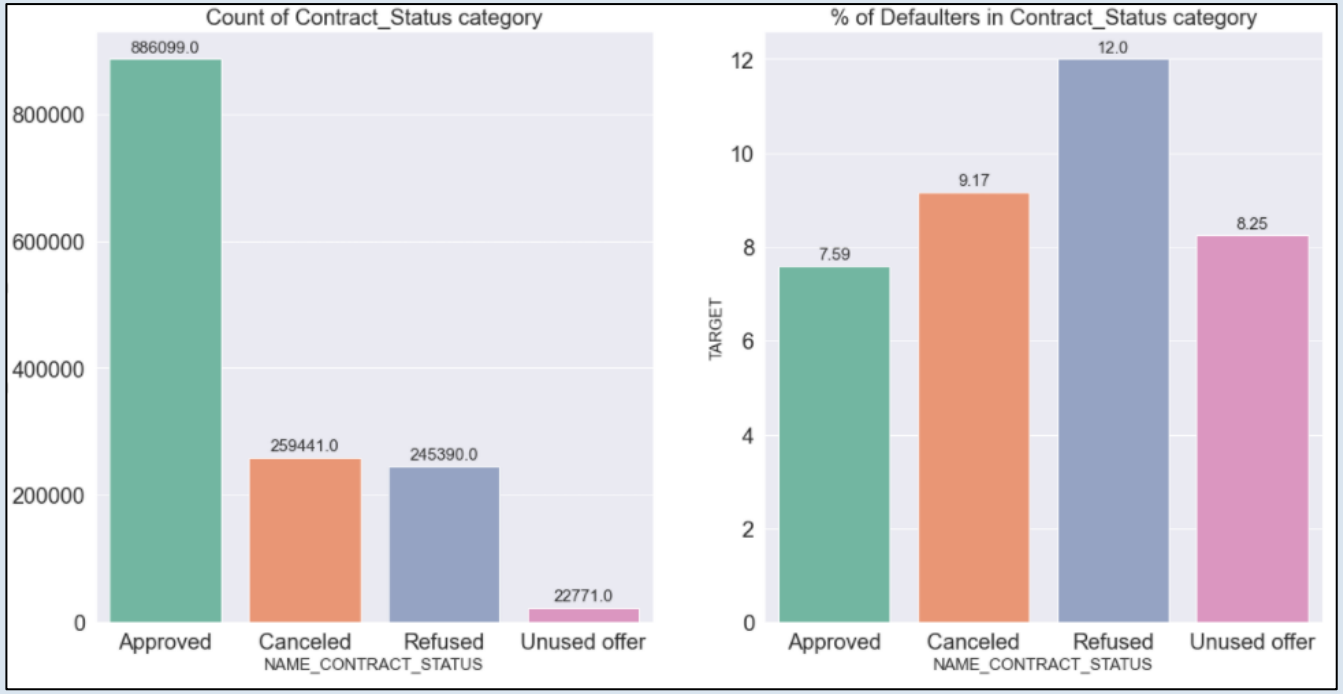




Merged - Data Analysis

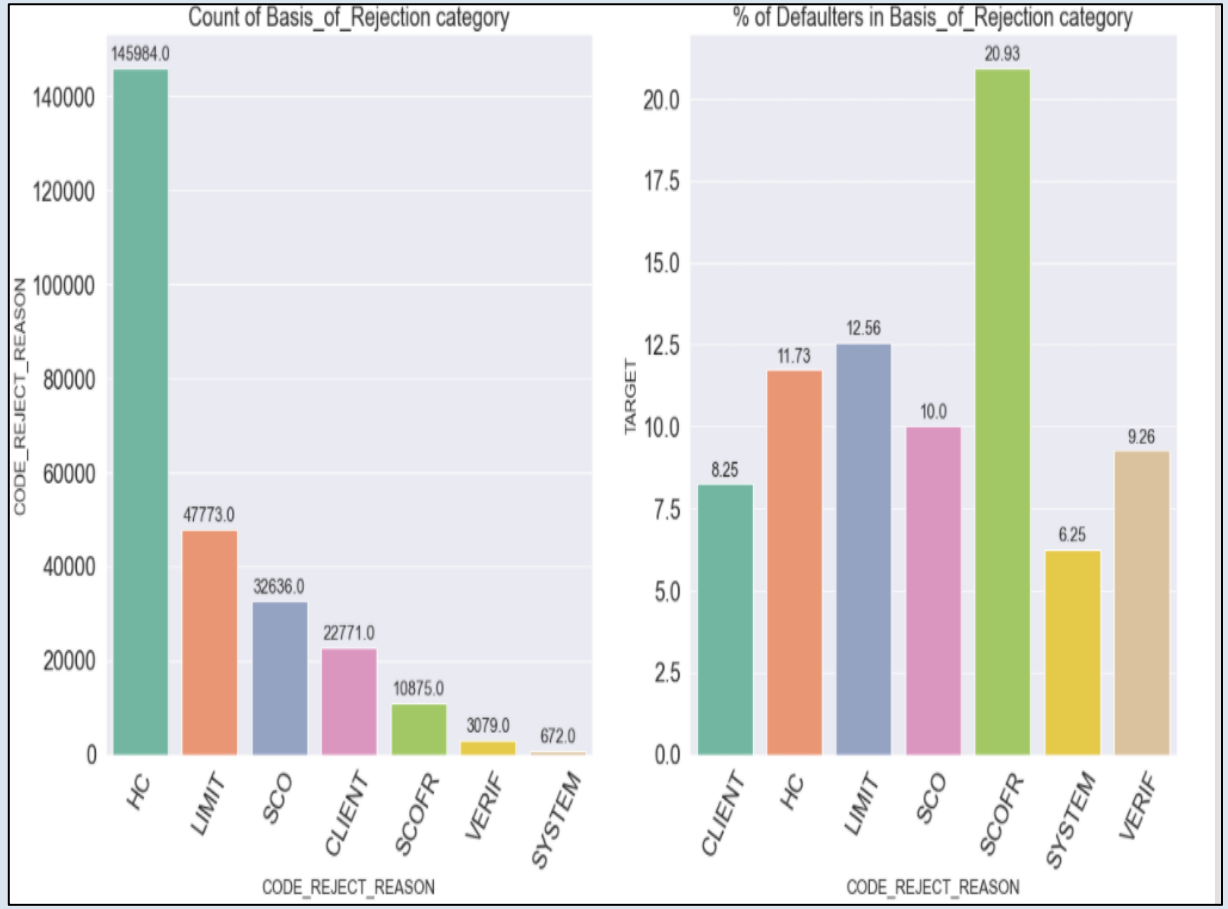
- From the right graph, we can see that the amount of applications for Consumer & Cash loans are pretty close, although we can see that defaulters of current application faced maximum % difficulties for Revolving loans(10.47%) in previous application and minimum % of difficulties in Consumer loans(7.71%).

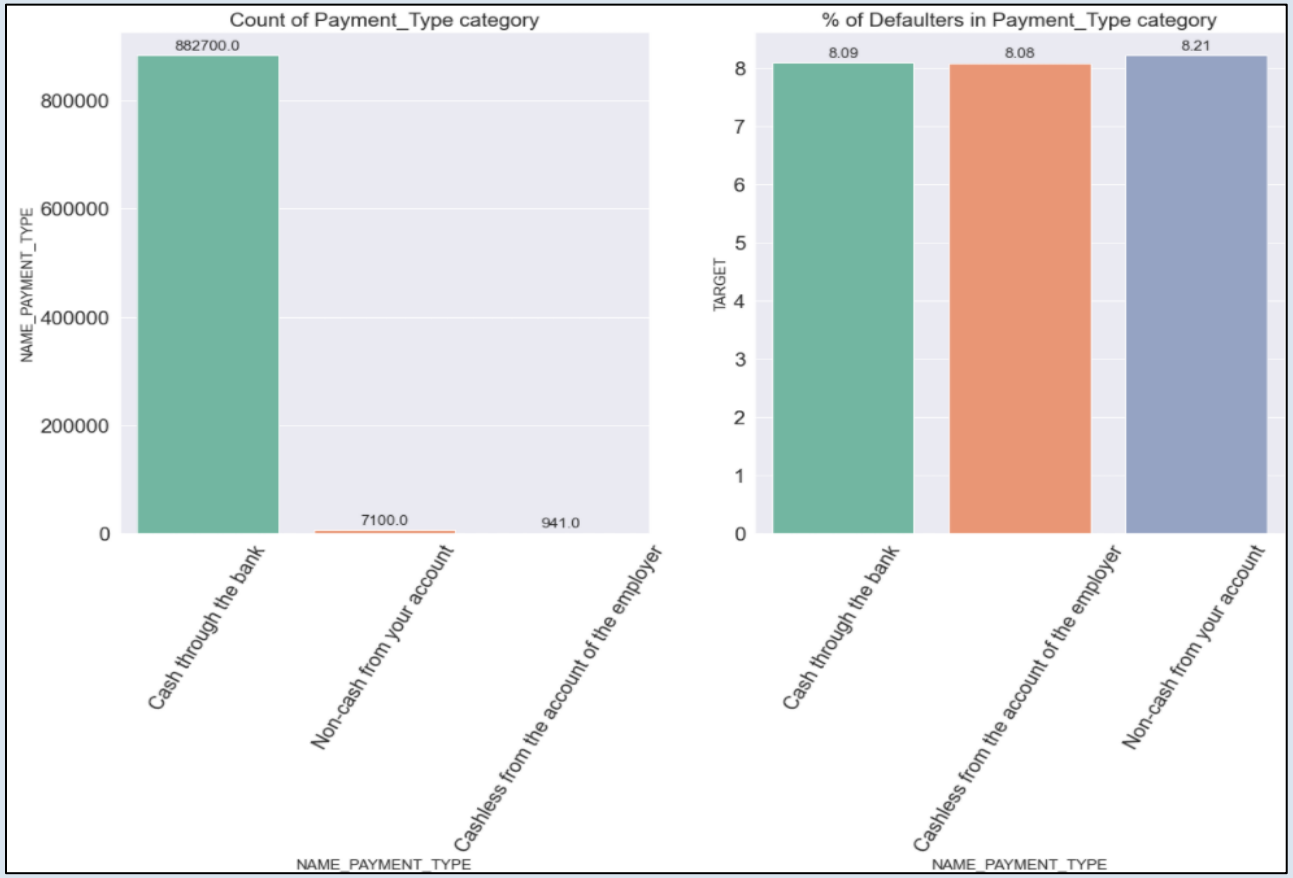




- From the left graph, we can see that majority of the applications have been approved, although we can also observe that the refused % which is 12% is the maximum % of difficulties faced by current applicants in their previous application & Approved % which is 7.59% is the minimum % of difficulties faced by current applicants in their previous application.

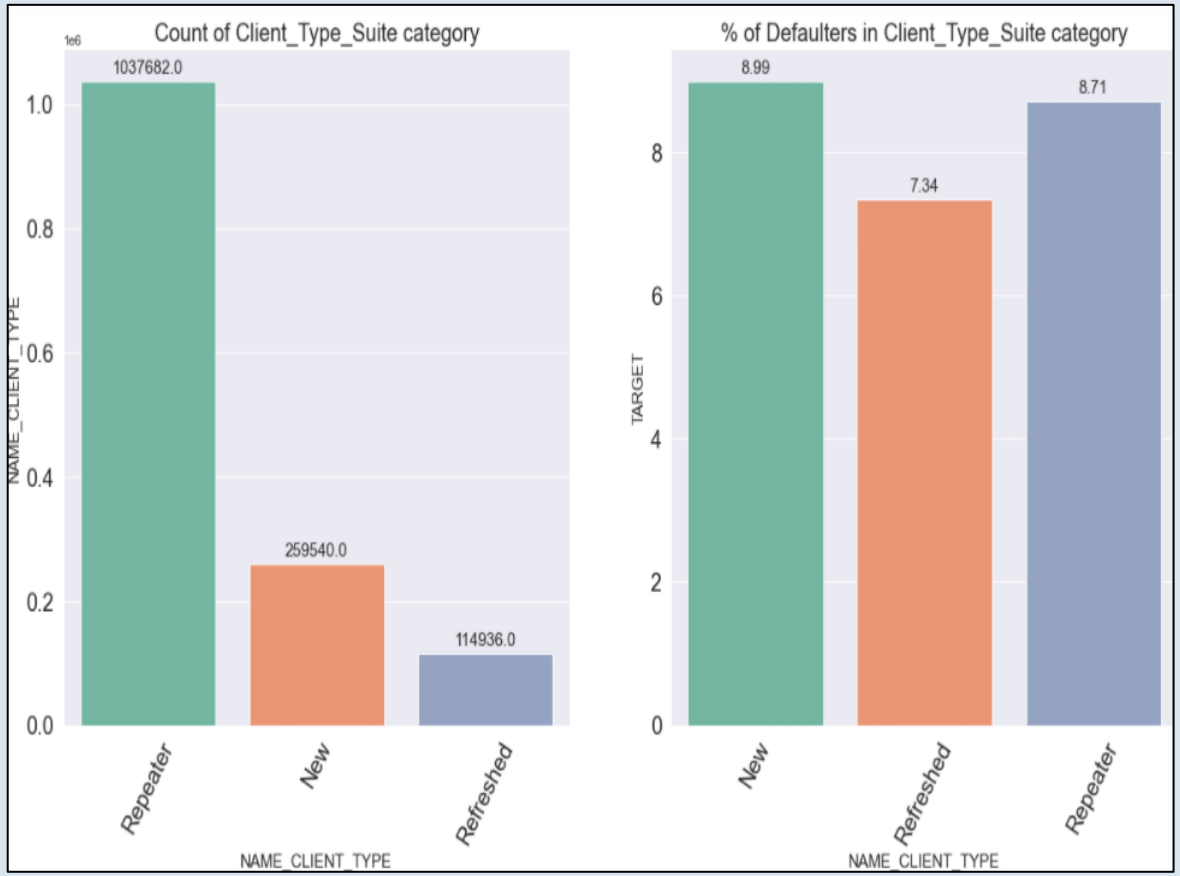
- From the right graph shows majority of applications got rejected for 'HC', but current applicants faced maximum payment difficulties for rejection in 'SCOFR' (20.93%) and minimum payment difficulties in 'SYSTEM'(6.25%) in their previous applications

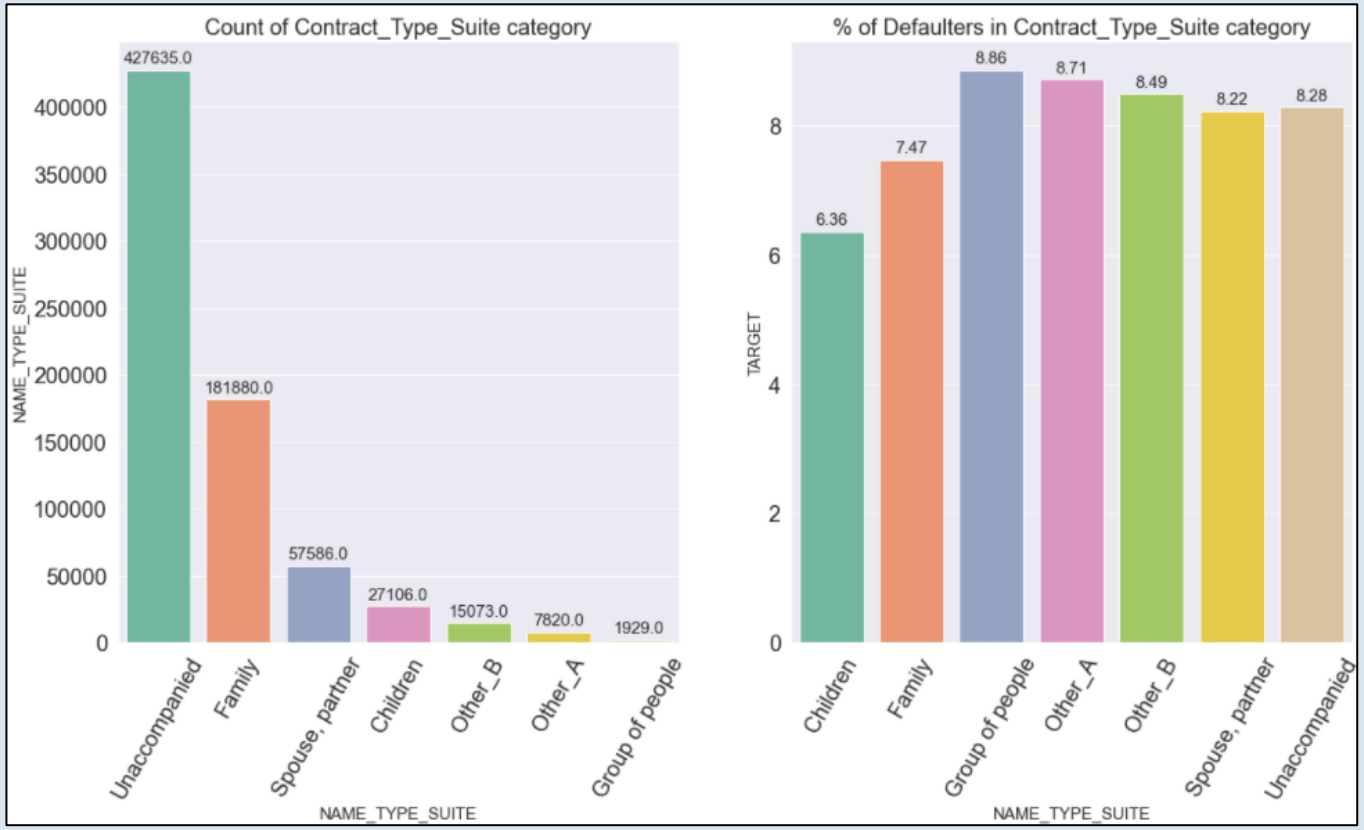




- From the left graph, we can see that majority of payments done by current applicants in their previous applications are through "Cash through the bank" payment-method.

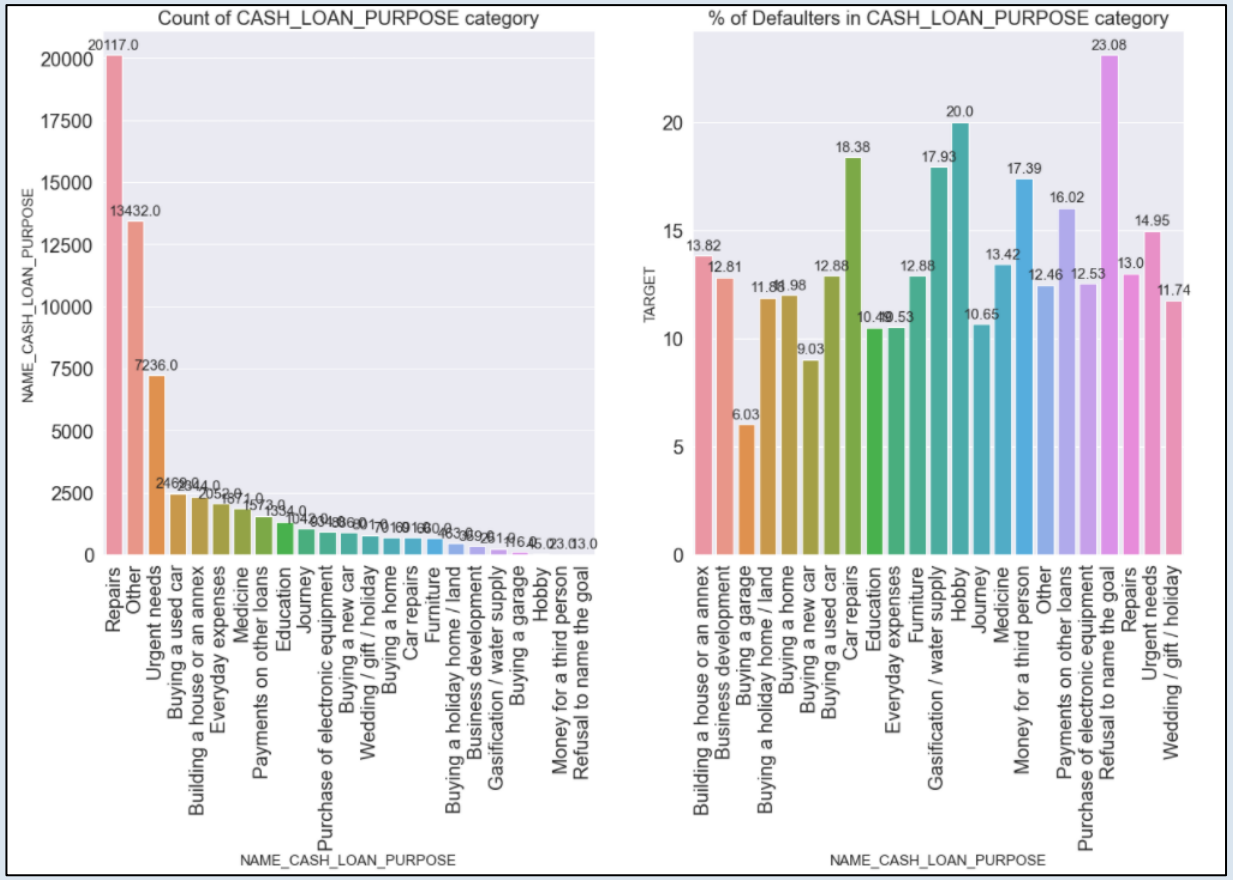
- From the right graph we can see that the graph on left shows most applicants from previous applications are 'Repeaters', whereas, 'Refreshed' clients from previous applications had least % difficulties (7.34%) and 'New' clients from previous applications had most % difficulties(8.99%).

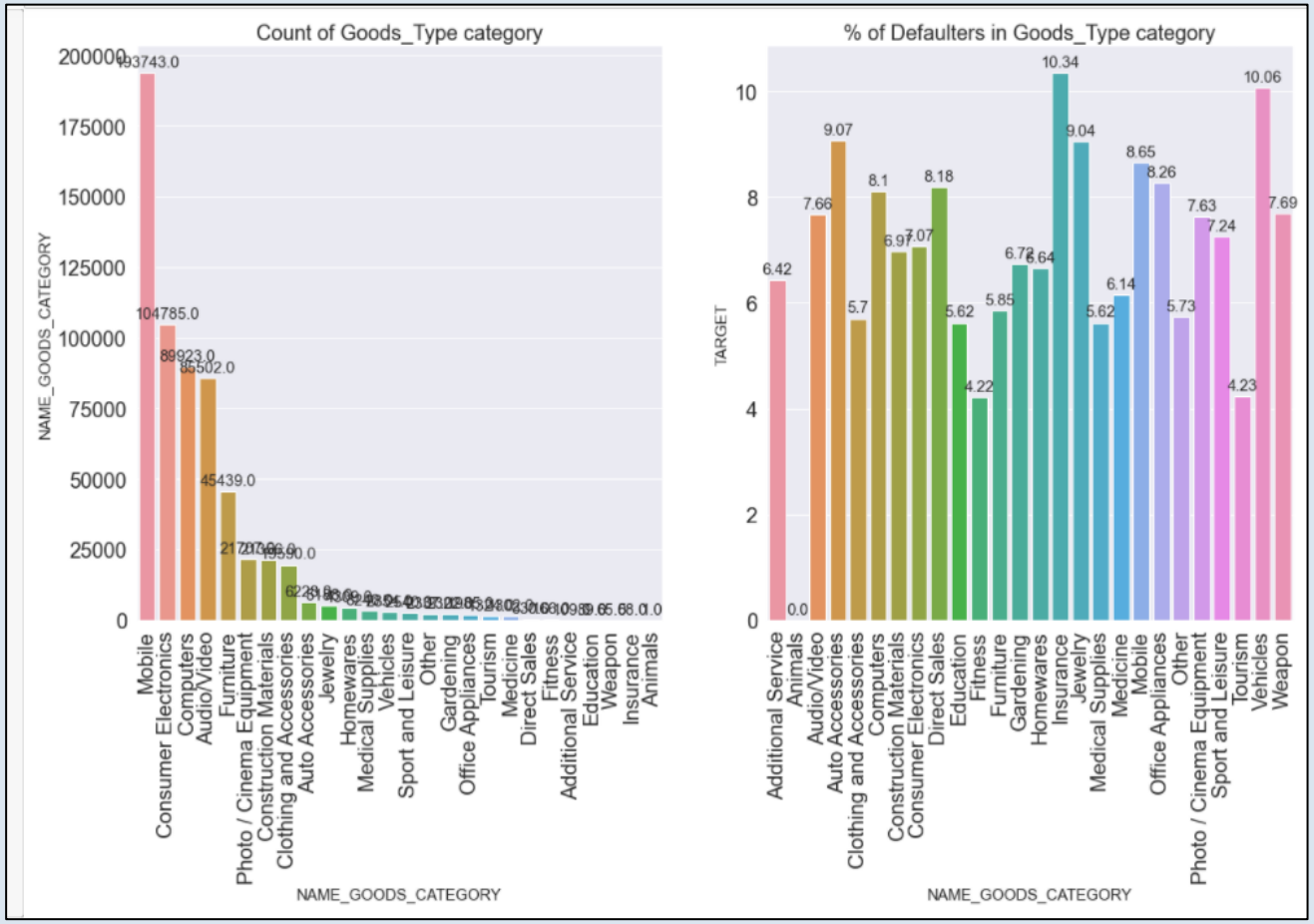




- From the left graph, we can see majority of loans are for 'Unaccompanied' but, current applicants in their previous applications faced maximum difficulties when they applied as 'Group of people' (8.86%).

- From the right graph we can see that most of the applicants have applied for loan for Mobile -goods category, but current applicants in previous application faced maximum % difficulties in goods category of 'Insurance'(10.34%) and 'Vehicles'(10.06%).





- From the left graph, we can see majority of applicants applied for cash loans for Repair category, but they faced maximum % difficulties in 'Refused to name the goal'(23.08%) & minimum difficulties in buying a garage(6.03%).

Conclusion & Recommendations

- Banks should be careful while giving loans to applicants who are of Young age, Middle age & are Unaccompanied as there are high chances for default.
- Cash loans have major applications as well as major defaulters (8.35%). Although there are less applicants for revolving loans but maximum difficulties(10.47%) are faced by clients for repayment of revolving loans.
- Banks can approve loan without much hesitation to applicants who are highly educated or hold academic degree or applicants who are senior citizens & applicants who are pensioners or commercial associates as they are less likely to default.
- Banks should be careful while approving loans for applicants who live with their families and applicants having loan annuity between 20,000 to 40,000 as they are more likely to default.
- The applicants who are single/not married , civil-marriage or widows are likely to default on credit. Alongside it is also observed that applicants in maternity leave and unemployed will have payment difficulties.

Conclusion & Recommendations

- Applicants who are having lower-secondary education, incomplete higher education are most likely to default on payments.
- Banks should be more likely to approve loans for working professionals and applicants who live in house/apartments & co-op apartments or are accompanied by spouse/family as they are less likely to default.
- Applicants who are low-skilled labours, live in rented-house and are accompanied by Other_A categories are most likely to have payment difficulties.
- Applicants who get rejected in previous application faced maximum difficulties (20.93%) for loan repayment when they got rejected due to ‘SCOFR’ rejection-category.
- Applicants who were ‘New’ clients in their previous applications faced maximum difficulties (8.99%) for loan repayment. Thus, bank should be little cautious while granting loans to new applicants or applicants without credit history.

Conclusion & Recommendations

- Applicants who applied loan as ‘Group of people’ faced maximum difficulties (8.86%) in terms of loan repayment.
- Refreshed clients tend to have less difficulties (7.34%) for repayment of loan.
- Applicants faced payment difficulties who applied for loan for goods – Insurance (10.34%) and ‘Vehicles’ (10.06%).
- Applicants who refuse to name the goal or purpose of the loan (23.08%) are most likely to default or have payment difficulties.
- The clients who were refused in previous applications faced maximum difficulties (12%) in loan repayment.

Thank you

Umang Rana