CREDIT EDA CASE STUDY

Sweta Kumari Shaw
Sukeerthi G

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

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 Objectives

- This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.
- 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.

APPLICATION DATA

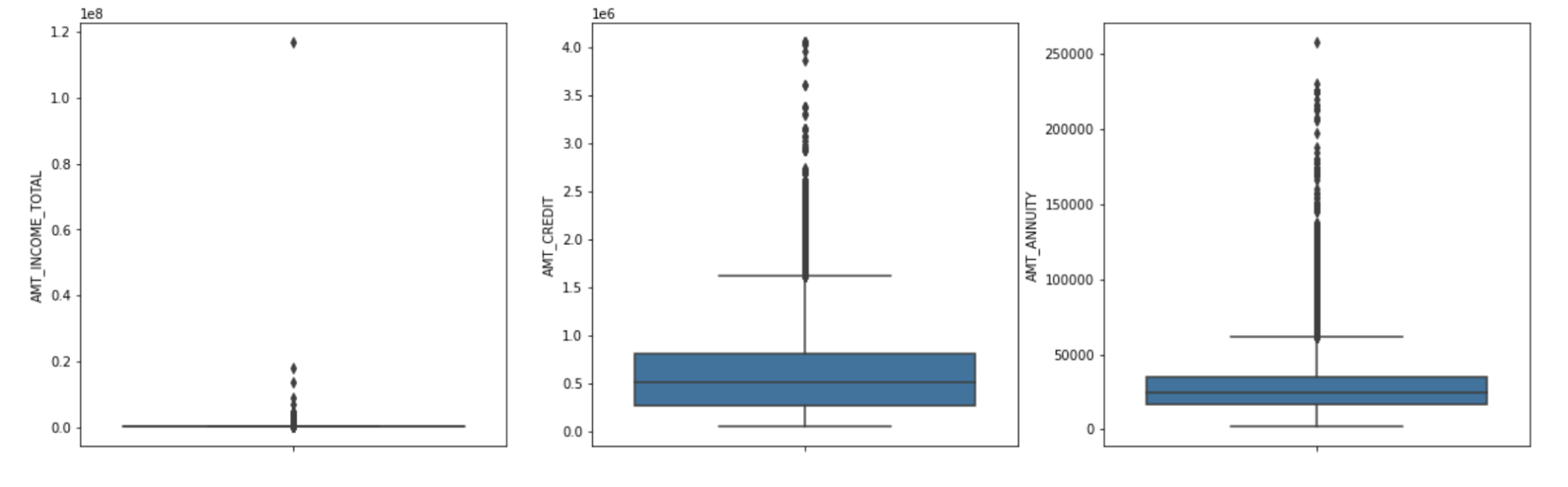
Data Cleaning

Data Cleaning

- Data Sourcing: Correctly read the dataset given ‘application_data.csv’
- Missing Values handling: Removed all columns having missing value greater than 30%(except for Occupation type due to its importance in analysis)
- Removed all columns that we felt contribute the least in the analysis
- Imputation of values: Imputed values by checking the presence of outliers and used either mean, median or mode based in the outliers
- Checked the datatypes of the columns and standardized them wherever required
- Fixed invalid entries
- Working with the outliers in columns like AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, CNT_CHILDREN, DAYS_EMPLOYED, CNT_FAM_MEMBERS

Outliers Analysis

- AMT_INCOME_TOTAL, AMT_CREDIT,AMT_ANNUITY

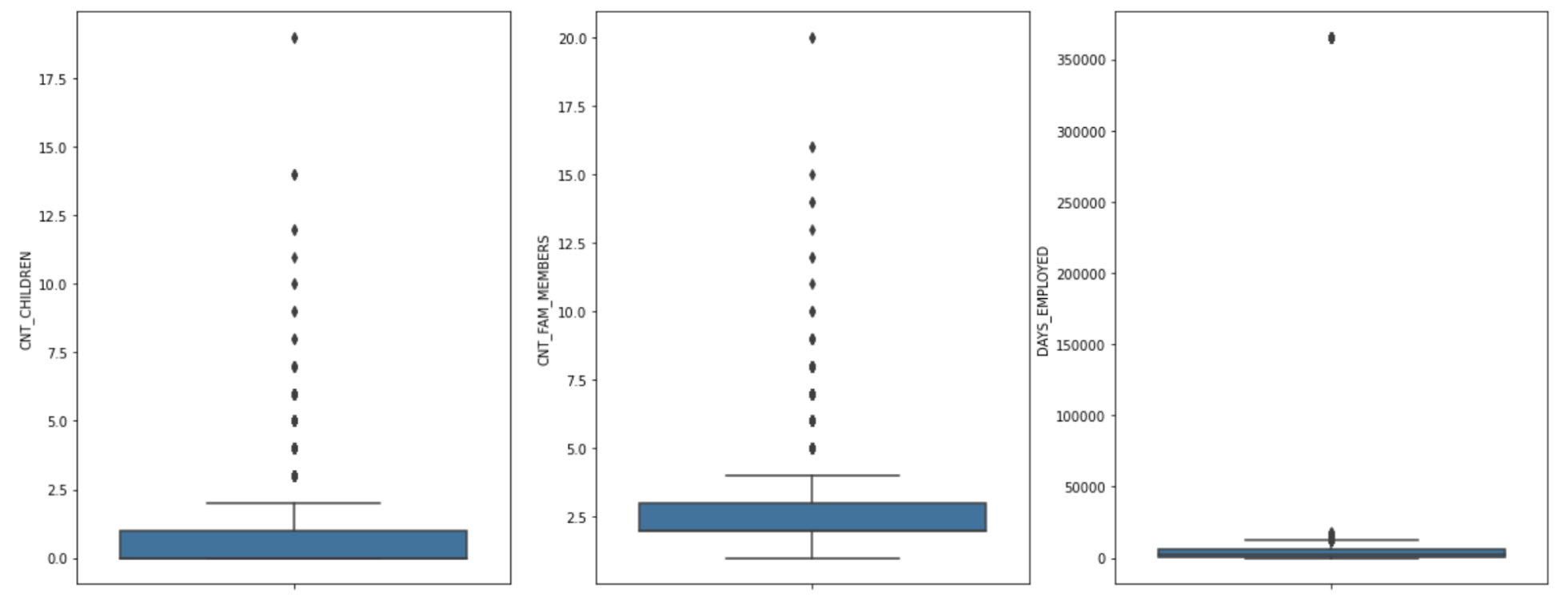


Insights:

- AMT_INCOME_TOTAL column is having outliers with large difference which indicates that some of loan applicant is having very much high salary as compare to others loan applicants.
- AMT_CREDIT,AMT_ANNUITY are also having outliers .

Outliers Analysis

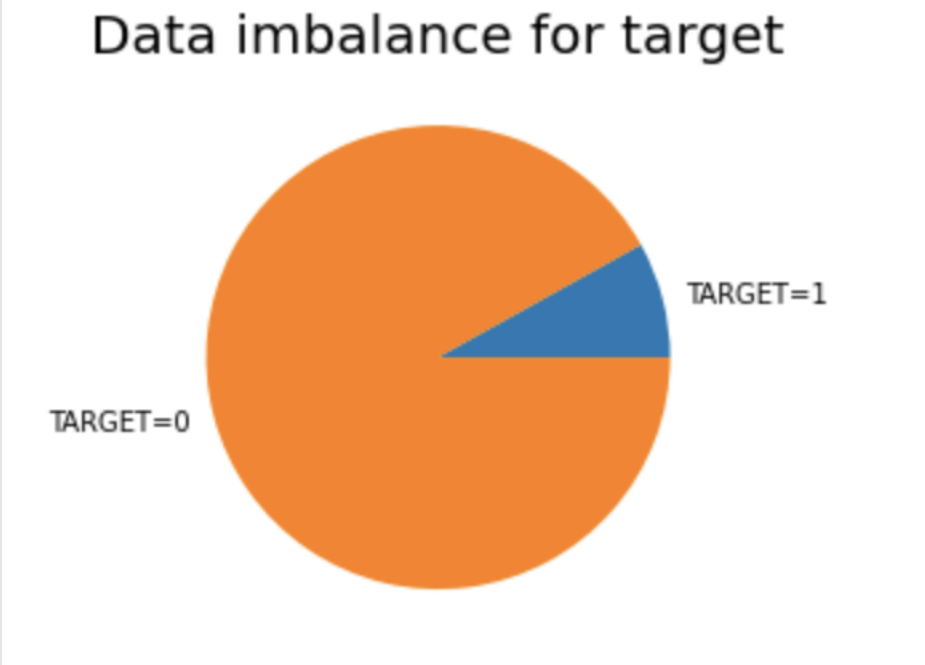
- CNT_CHILDREN, DAYS_EMPLOYED, CNT_FAM_MEMBERS



Insights:

- CNT_CHILDREN is having outliers above 2.5.
- CNT_FAM_MEMBERS having value more than 5 are all outliers.
- DAYS_EMPLOYED is having some outlier values around 350000(days) i.e. 958 years which is not possible. Hence we can conclude these as wrong entries.

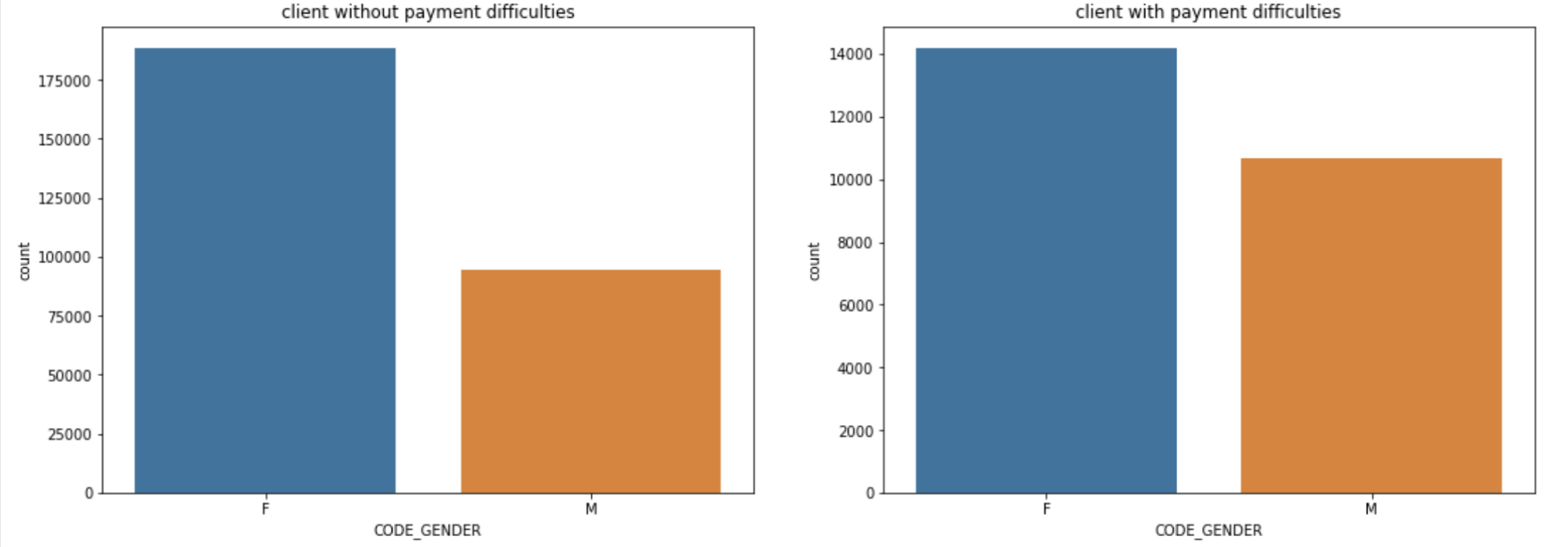
Checking Imbalance for ‘TARGET’ column



- Here we can see that application data is having high imbalance of Defaulted population and Non-defaulted population.
- Majority is target 0 and minority is target 1
- Imbalance ratio is 11.39

UNIVARIATE ANALYSIS OF CATEGORICAL VARIABLES

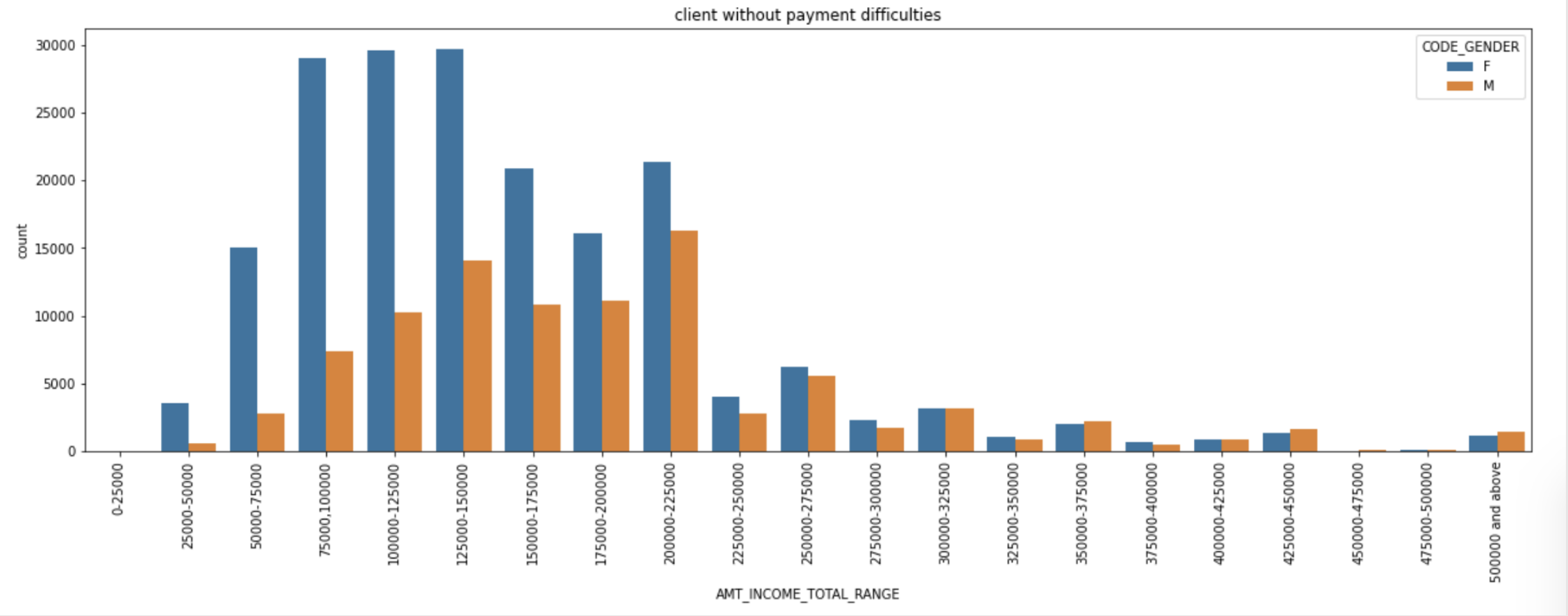
Analysis based on ‘CODE_GENDER’



- We see that there are more number defaulters than the repayors' amongst the ‘Male’

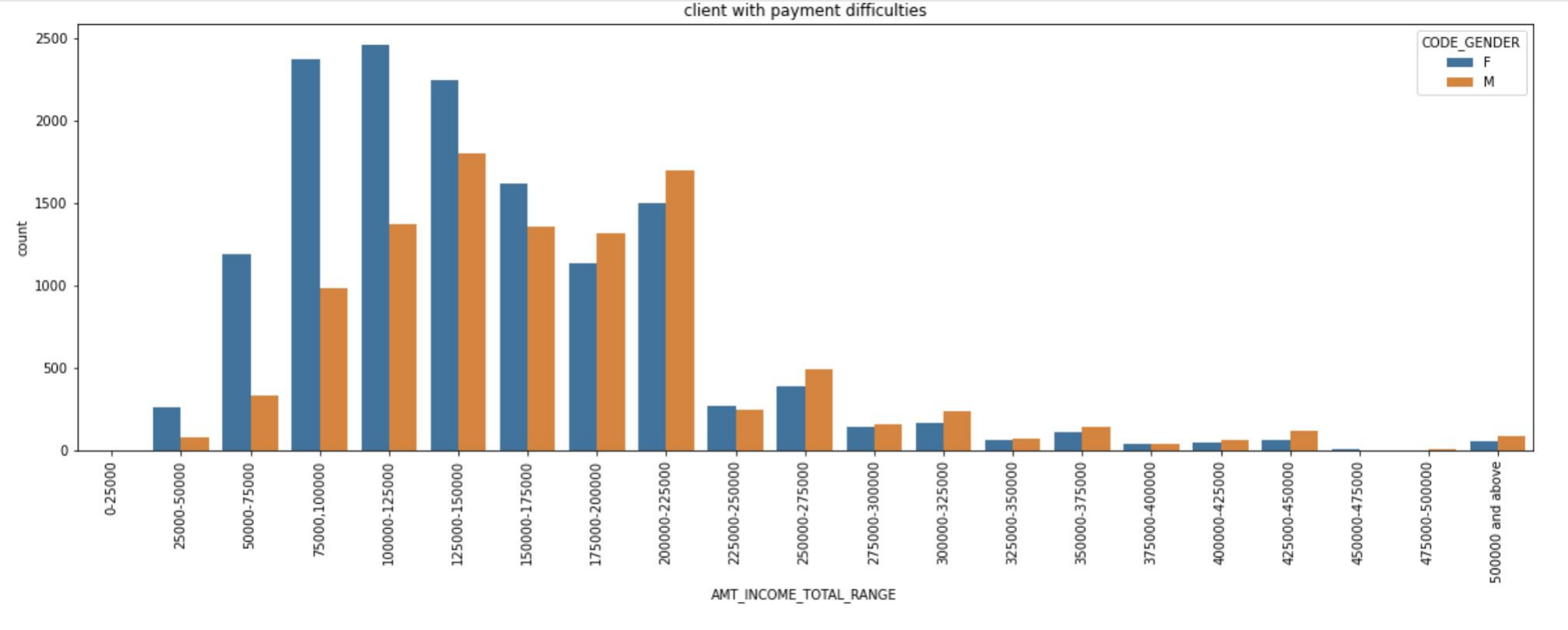
Analysis based on ‘AMT_INCOME_TOTAL_RANGE’

- Client Without payment difficulties



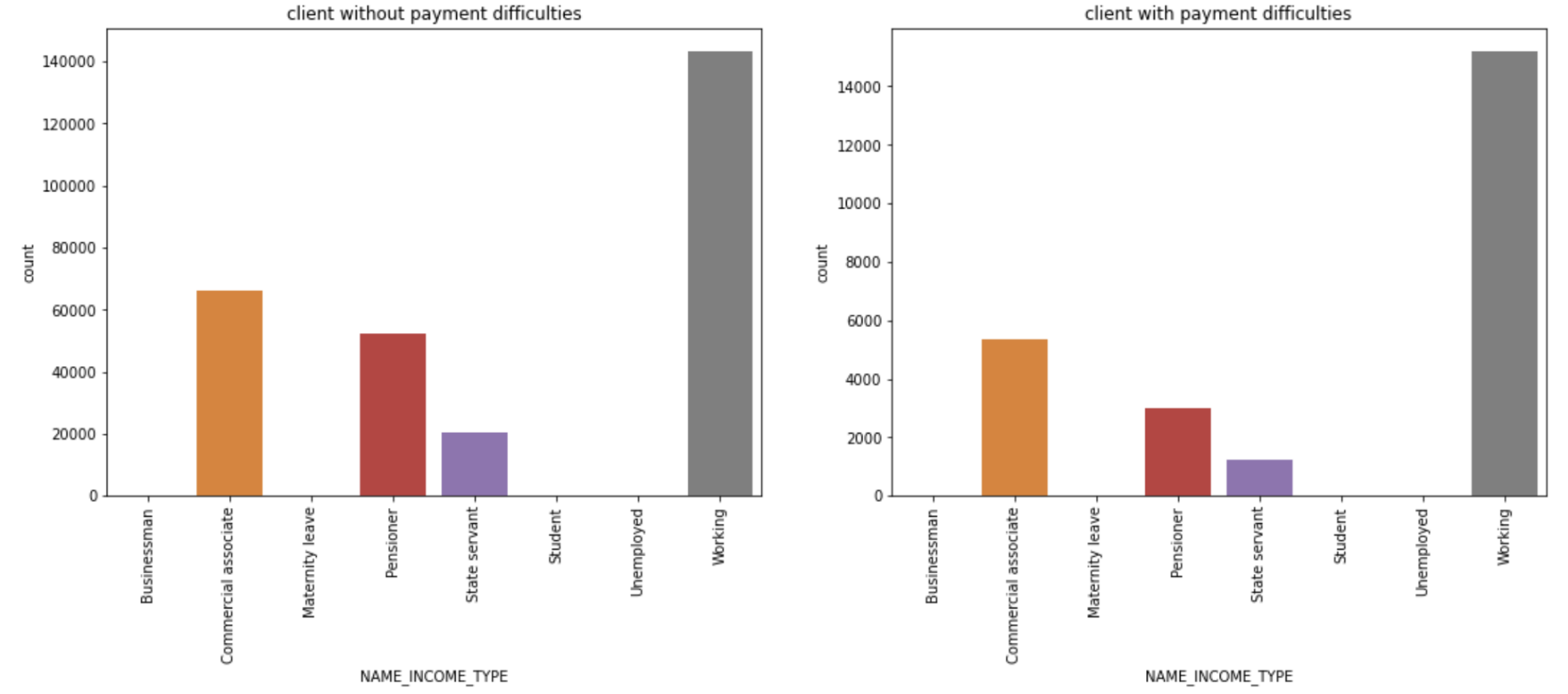
Analysis based on ‘AMT_INCOME_TOTAL_RANGE’

- Client With payment difficulties

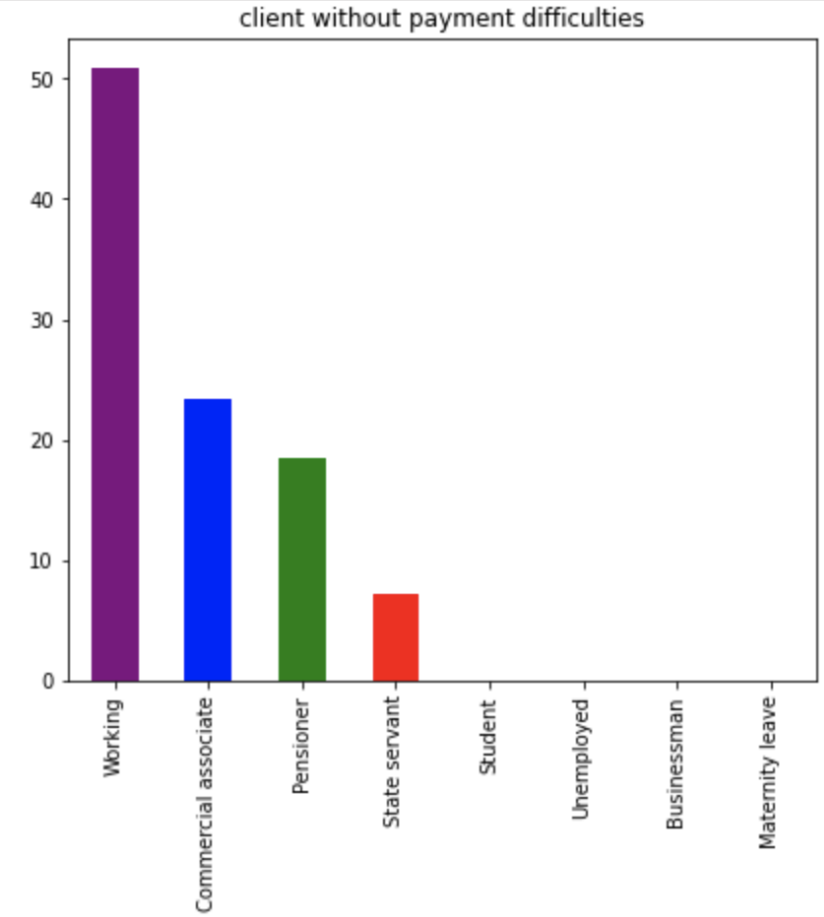


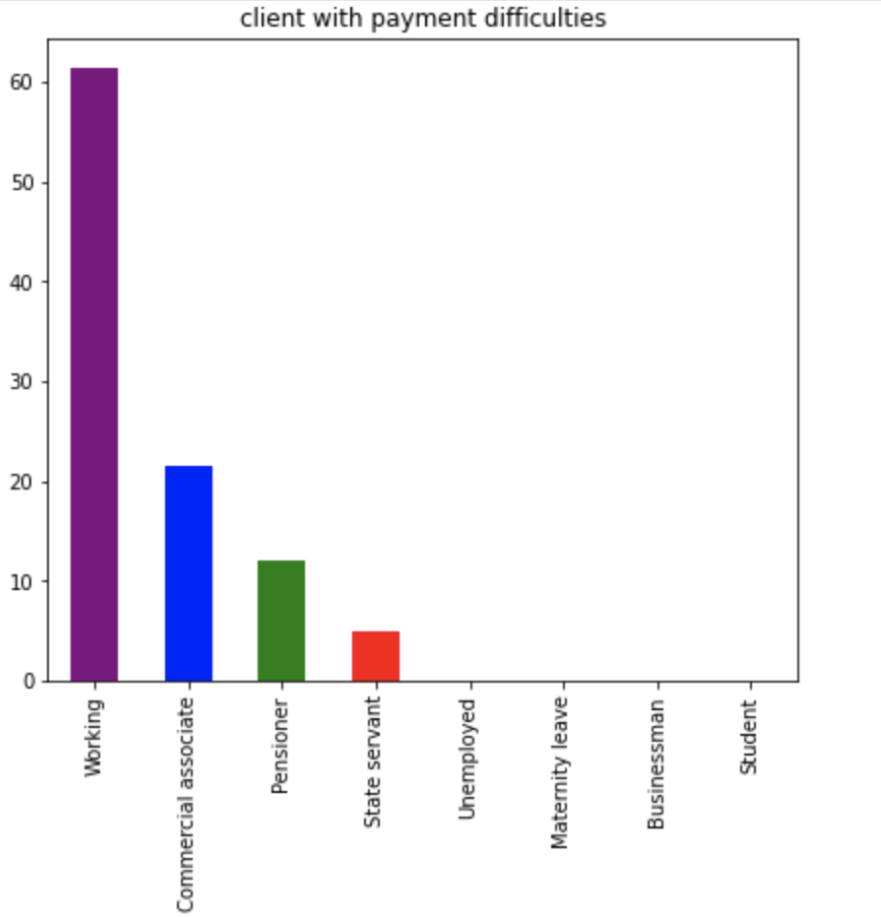
- Most of applications have Income total less than 3 Lakhs.
- Application with Income less than 3 Lakhs has high probability for being defaulter.
- Applicant with Income more than 4.5-5 Lakhs are less likely to default.

Analysis based on ‘NAME_INCOME_TYPE’



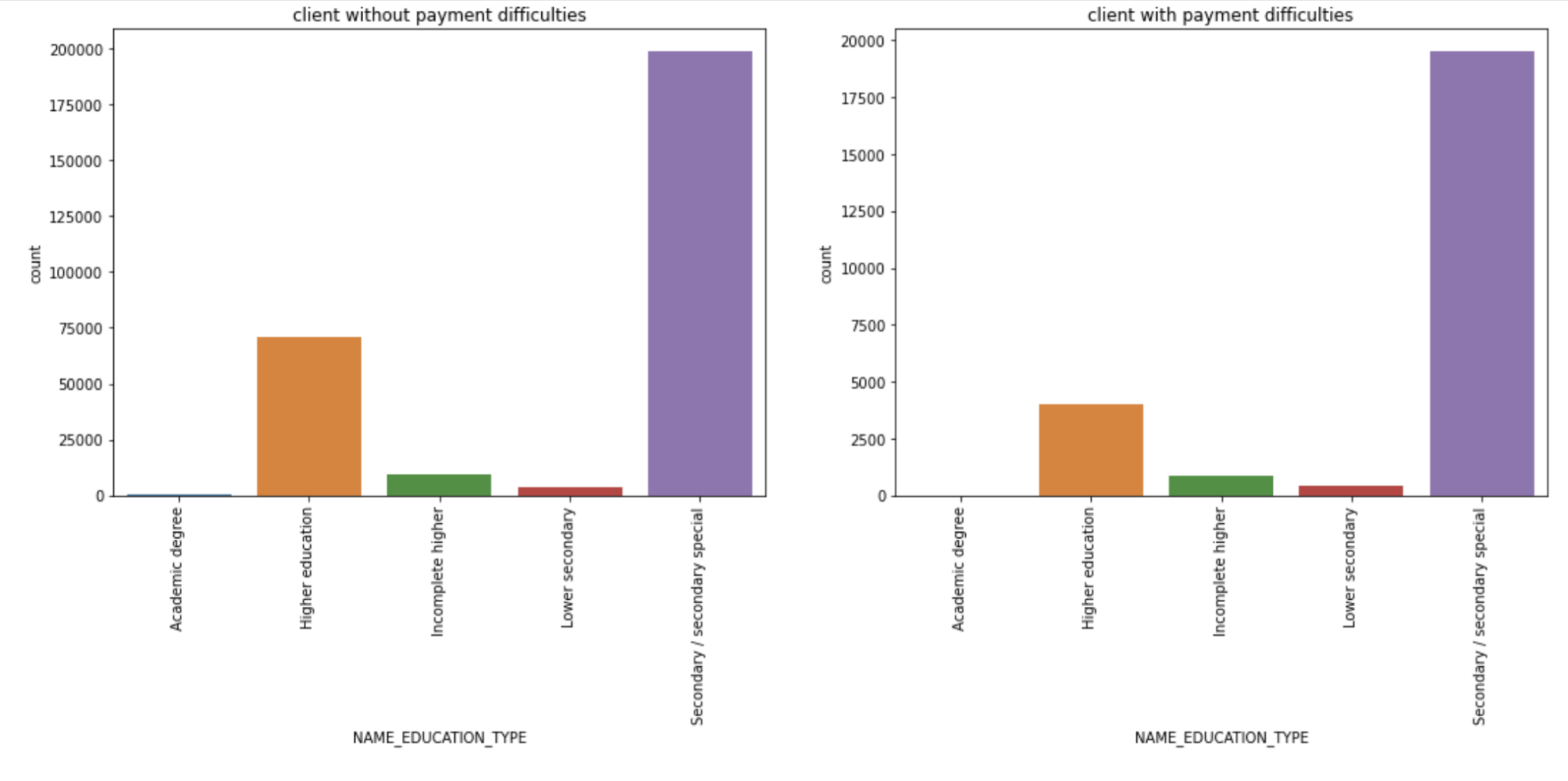
Analysis based on ‘NAME_INCOME_TYPE’





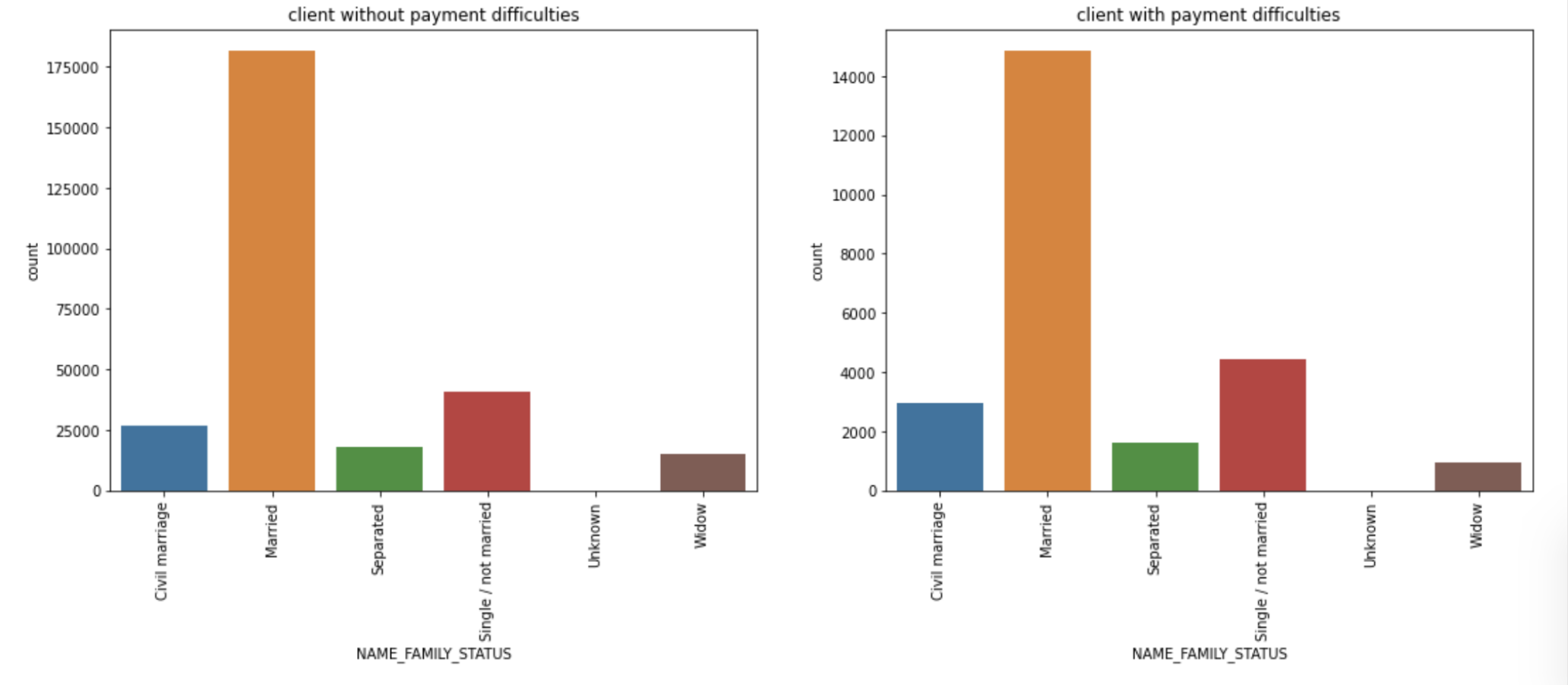
- Pensioners are having better on-time payments
- Students & Businessmen don't have Payment difficulties

Analysis based on ‘NAME_EDUCATION_TYPE’

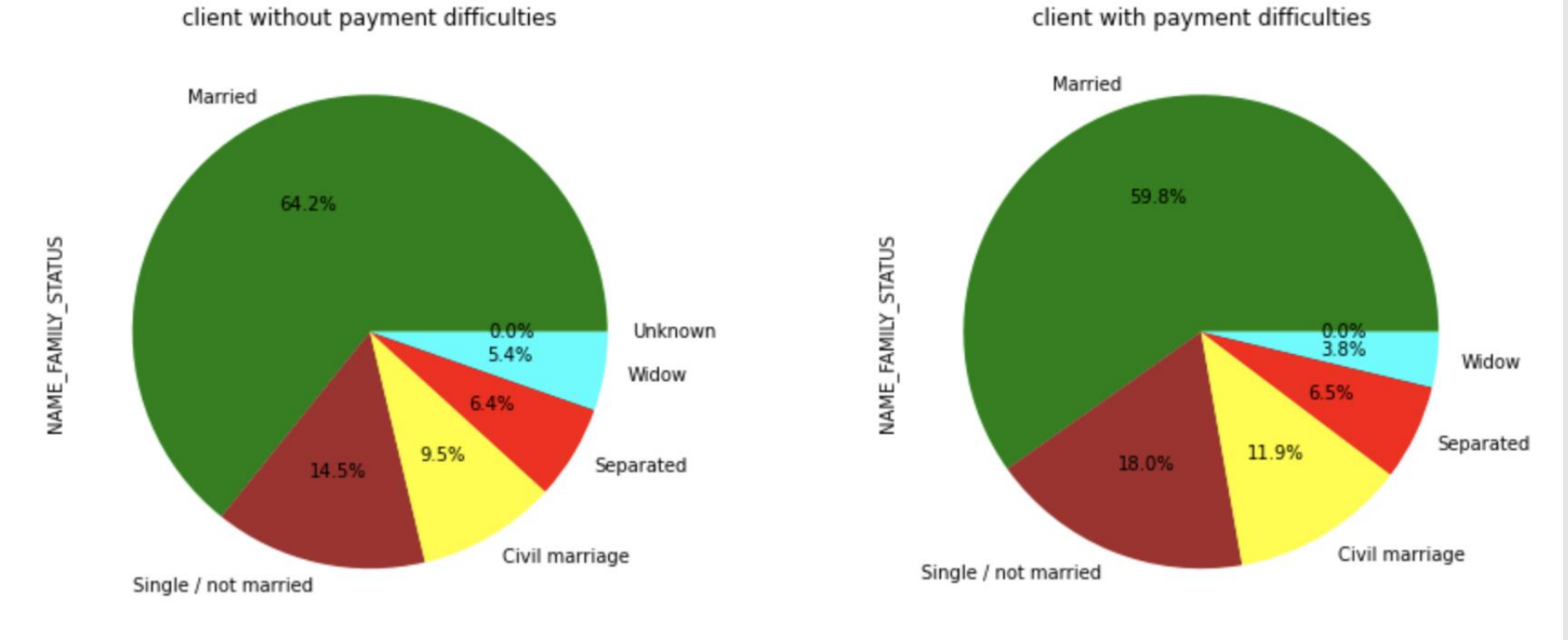


- Applicants with 'Higher education' have less payment difficulties.

Analysis based on ‘NAME_FAMILY_STATUS’

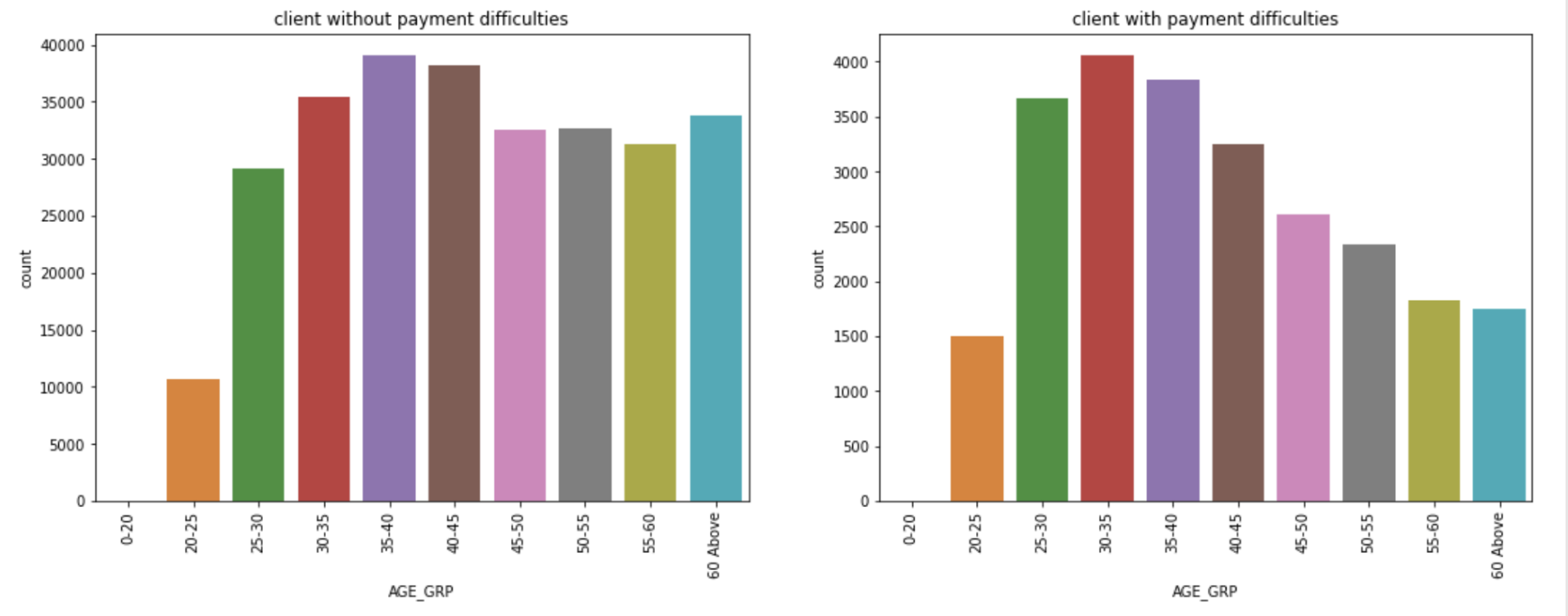


Analysis based on ‘NAME_FAMILY_STATUS’



- Applicants who are belonging to 'Married' or 'Widow', are having better on-time payments compare to others.
- Applicants who are 'Single/not married' have difficulties with on-time payment.

Analysis based on ‘AGE_GRP’



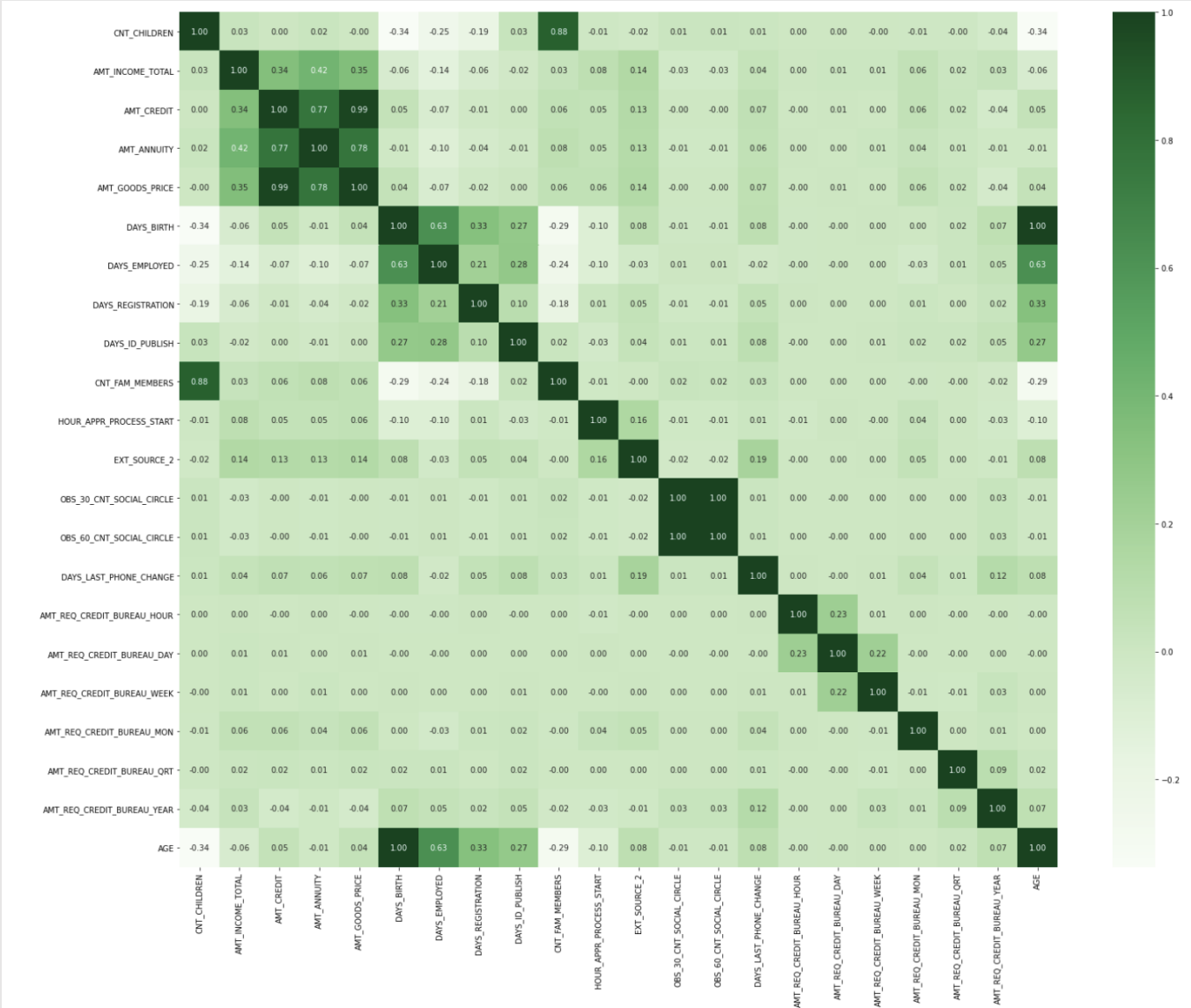
- Applicants having age less than 40 are having payment difficulties.
- Applicants having age more than 45 , they are having better on time payment.

Summary- Univariate Analysis Of Categorical Variables

- We see that there are more number defaulters than the repayors' amongst the ‘Male’
- Application with Income less than 3 Lakhs has high probability for being defaulter and applicants with Income more than 4.5-5 Lakhs are less likely to default
- Pensioners are having better on-time payments where as students & businessmen don't have Payment difficulties
- Applicants with 'Higher education' have less payment difficulties
- Applicants who are 'Single/not married' have difficulties with on-time payment
- Applicants having age less than 40 are having payment difficulties

CORRELATION OF NUMERICAL COLUMNS

Target 0 (client without payment difficulties)



Target 0 (client without payment difficulties)

	Column1	Column2	Correlation	Abs_Correlation
467	AGE	DAYS_BIRTH	1.000	1.000
298	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.999	0.999
90	AMT_GOODS_PRICE	AMT_CREDIT	0.987	0.987
198	CNT_FAM_MEMBERS	CNT_CHILDREN	0.879	0.879
91	AMT_GOODS_PRICE	AMT_ANNUITY	0.777	0.777
68	AMT_ANNUITY	AMT_CREDIT	0.771	0.771
468	AGE	DAYS_EMPLOYED	0.626	0.626
137	DAYS_EMPLOYED	DAYS_BIRTH	0.626	0.626
67	AMT_ANNUITY	AMT_INCOME_TOTAL	0.419	0.419
89	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349	0.349
45	AMT_CREDIT	AMT_INCOME_TOTAL	0.343	0.343
462	AGE	CNT_CHILDREN	-0.337	0.337
110	DAYS_BIRTH	CNT_CHILDREN	-0.337	0.337
159	DAYS_REGISTRATION	DAYS_BIRTH	0.333	0.333
469	AGE	DAYS_REGISTRATION	0.333	0.333

Target 0 (client without payment difficulties)

- Top 10 Correlations (after removing ‘AGE’ as it is a custom column added for analysis)

OBS_60_CNT_SOCIAL_CIRCLE OBS_30_CNT_SOCIAL_CIRCLE 0.999 0.999

AMT_GOODS_PRICE AMT_CREDIT 0.987 0.987

CNT_FAM_MEMBERS CNT_CHILDREN 0.879 0.879

AMT_GOODS_PRICE AMT_ANNUITY 0.777 0.777

AMT_ANNUITY AMT_CREDIT 0.771 0.771

DAYS_EMPLOYED DAYS_BIRTH 0.626 0.626

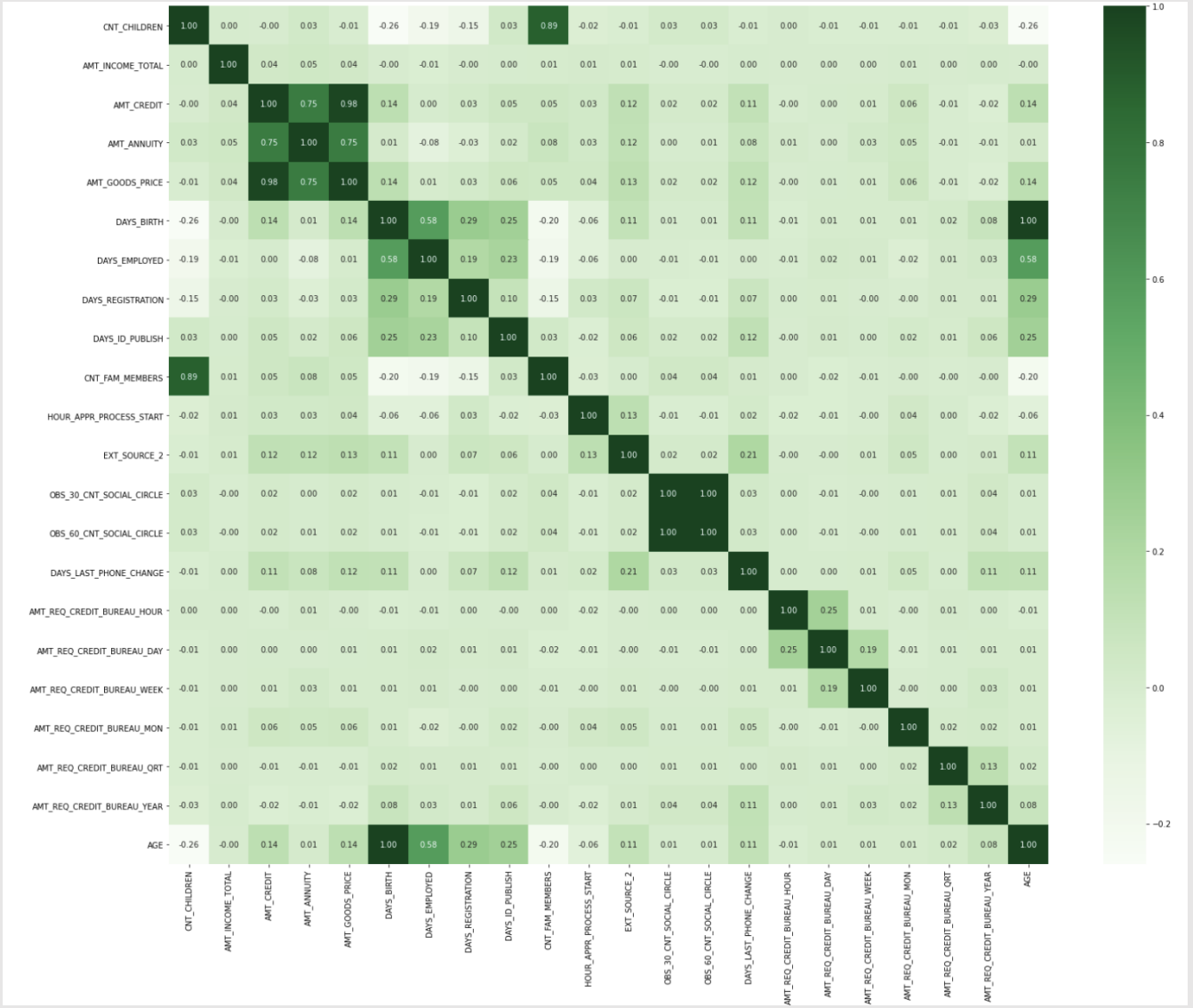
AMT_ANNUITY AMT_INCOME_TOTAL 0.419 0.419

AMT_GOODS_PRICE AMT_INCOME_TOTAL 0.349 0.349

AMT_CREDIT AMT_INCOME_TOTAL 0.343 0.343

DAYS_BIRTH CNT_CHILDREN -0.337 0.337

Target 1 (client with payment difficulties)



Target 1 (client with payment difficulties)

	Column1	Column2	Correlation	Abs_Correlation
467	AGE	DAYS_BIRTH	1.000	1.000
298	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998	0.998
90	AMT_GOODS_PRICE	AMT_CREDIT	0.983	0.983
198	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885	0.885
91	AMT_GOODS_PRICE	AMT_ANNUITY	0.753	0.753
68	AMT_ANNUITY	AMT_CREDIT	0.752	0.752
137	DAYS_EMPLOYED	DAYS_BIRTH	0.582	0.582
468	AGE	DAYS_EMPLOYED	0.582	0.582
159	DAYS_REGISTRATION	DAYS_BIRTH	0.289	0.289
469	AGE	DAYS_REGISTRATION	0.289	0.289
110	DAYS_BIRTH	CNT_CHILDREN	-0.259	0.259
462	AGE	CNT_CHILDREN	-0.259	0.259
181	DAYS_ID_PUBLISH	DAYS_BIRTH	0.253	0.253
470	AGE	DAYS_ID_PUBLISH	0.253	0.253
367	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_HOUR	0.248	0.248

Target 1 (client with payment difficulties)

- Top 10 Correlations (after removing 'AGE' as it is a custom column added for analysis)

OBS_60_CNT_SOCIAL_CIRCLE OBS_30_CNT_SOCIAL_CIRCLE 0.998 0.998

AMT_GOODS_PRICE AMT_CREDIT 0.983 0.983

CNT_FAM_MEMBERS CNT_CHILDREN 0.885 0.885

AMT_GOODS_PRICE AMT_ANNUITY 0.753 0.753

AMT_ANNUITY AMT_CREDIT 0.752 0.752

DAYS_EMPLOYED DAYS_BIRTH 0.582 0.582

DAYS_REGISTRATION DAYS_BIRTH 0.289 0.289

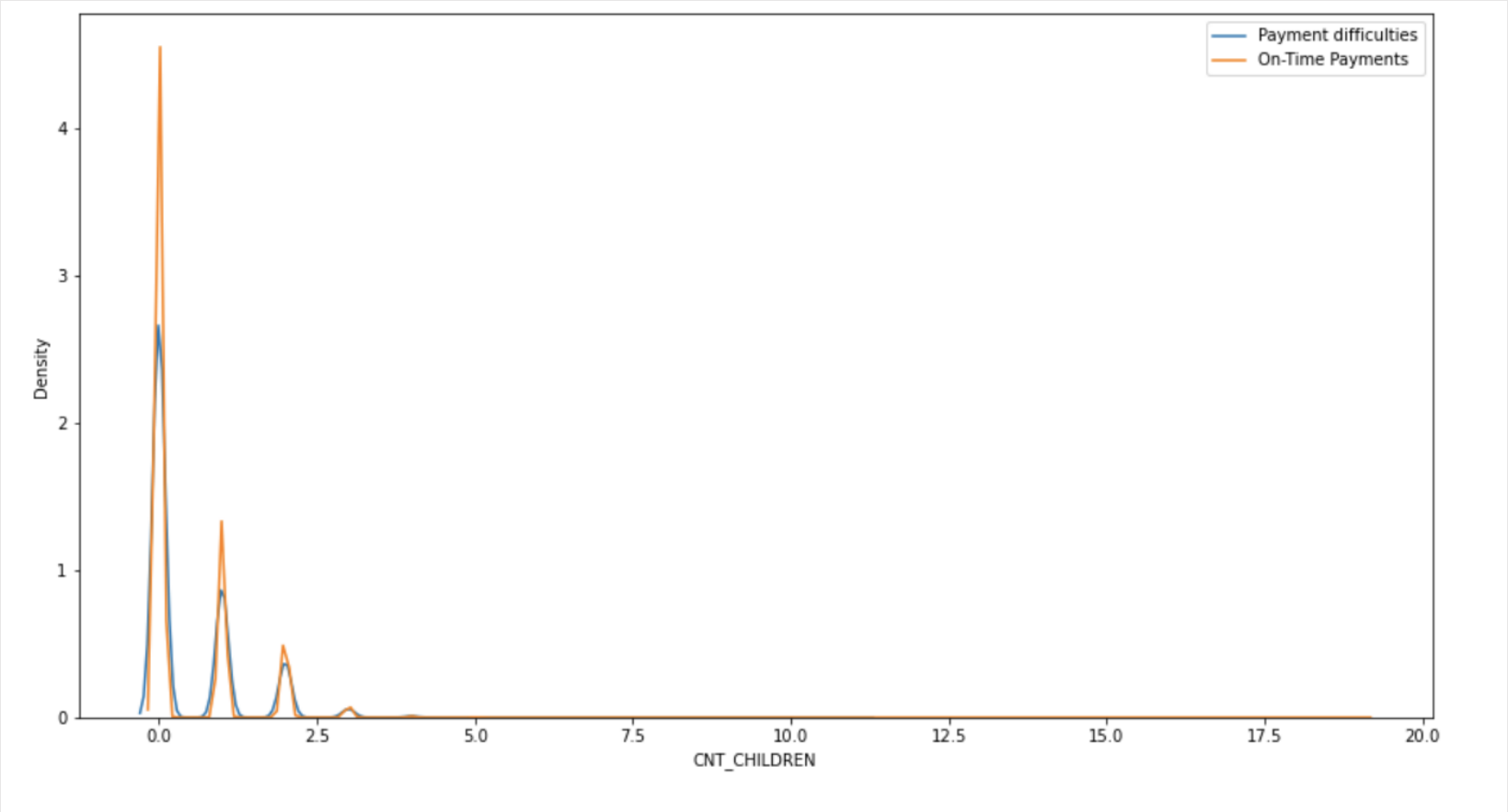
DAYS_BIRTH CNT_CHILDREN -0.259 0.259

DAYS_ID_PUBLISH DAYS_BIRTH 0.253 0.253

AMT_REQ_CREDIT_BUREAU_DAY AMT_REQ_CREDIT_BUREAU_HOUR 0.248 0.248

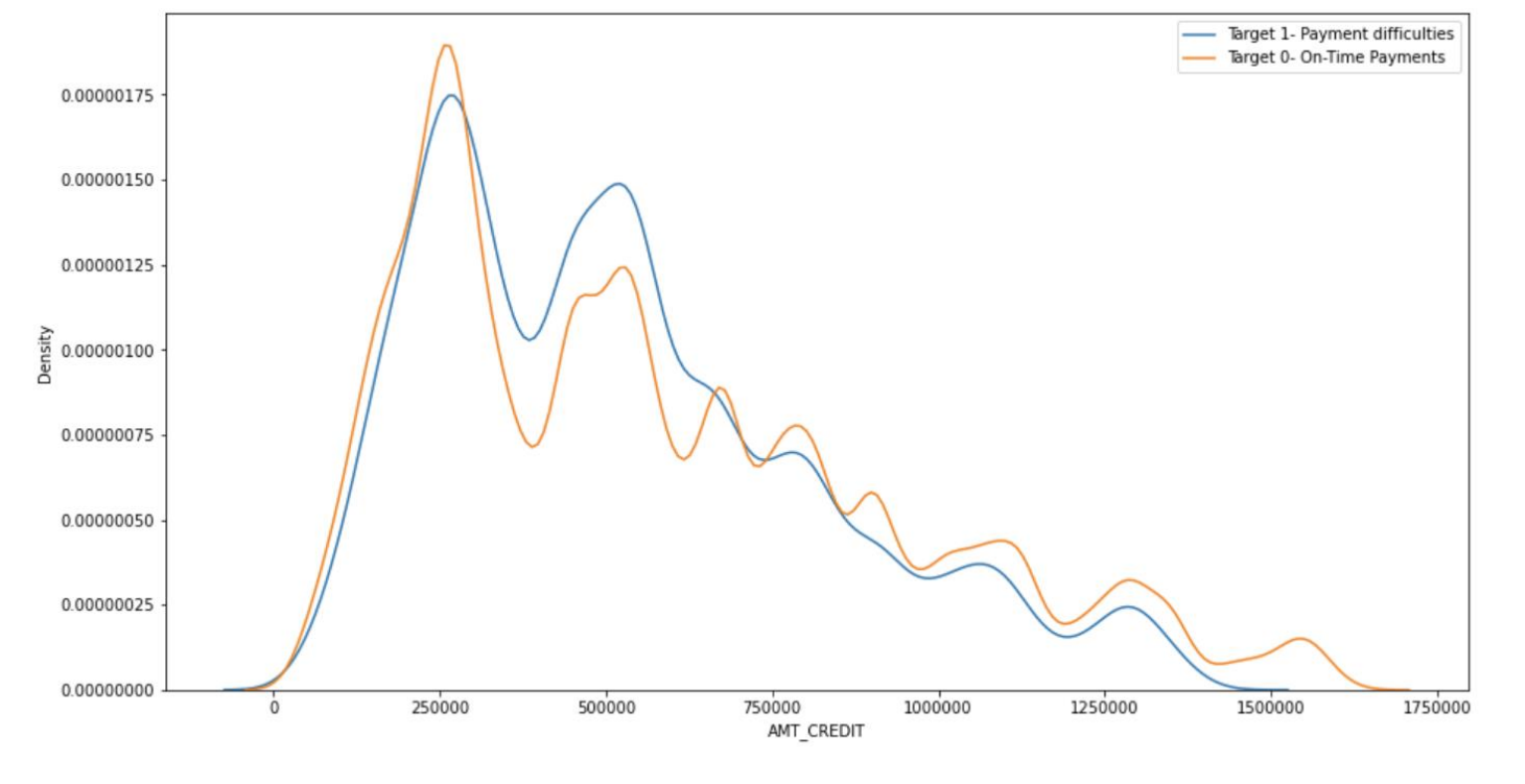
UNIVARIATE ANALYSIS OF NUMERICAL VARIABLES

Analysis based on ‘CNT_CHILDREN’



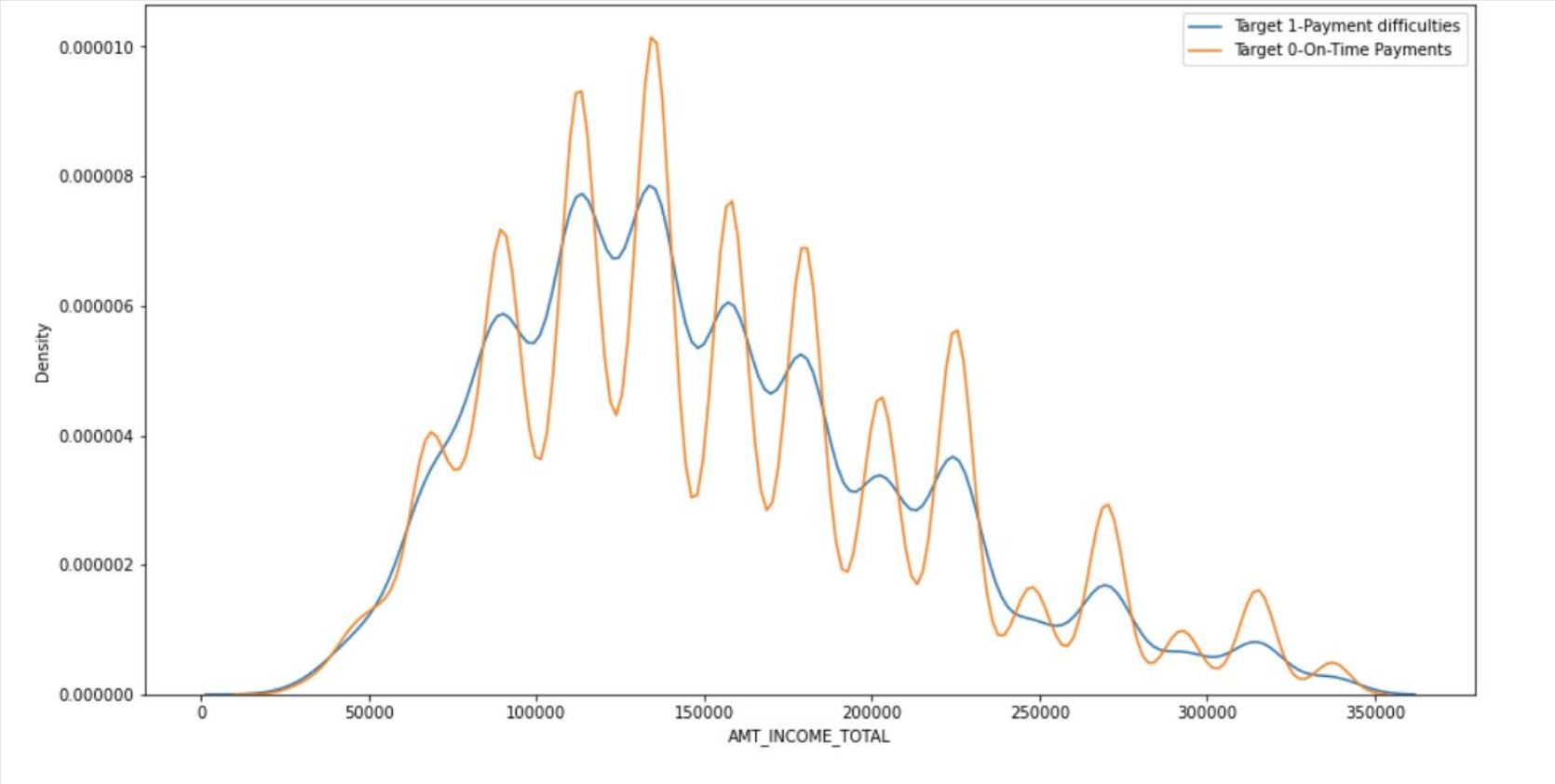
- For clients those are having no children are more efficient for On-Time Payments
- For clients those are with 1-2 children, there are few more clients with On-Time Payments as compare to clients with payment difficulties.

Analysis based on ‘AMT_CREDIT’



- For clients having AMT_CREDIT > 750000 , they are more likely for doing On-Time Payments.
- For clients having AMT_CREDIT between 200000 and 650000, they are having Payment difficulties

Analysis based on ‘AMT_INCOME_TOTAL’



- For clients with Payment difficulties, AMT_INCOME_TOTAL value distribution seems a normal distribution but in case of clients with On-Time Payments, we don't see any valid pattern

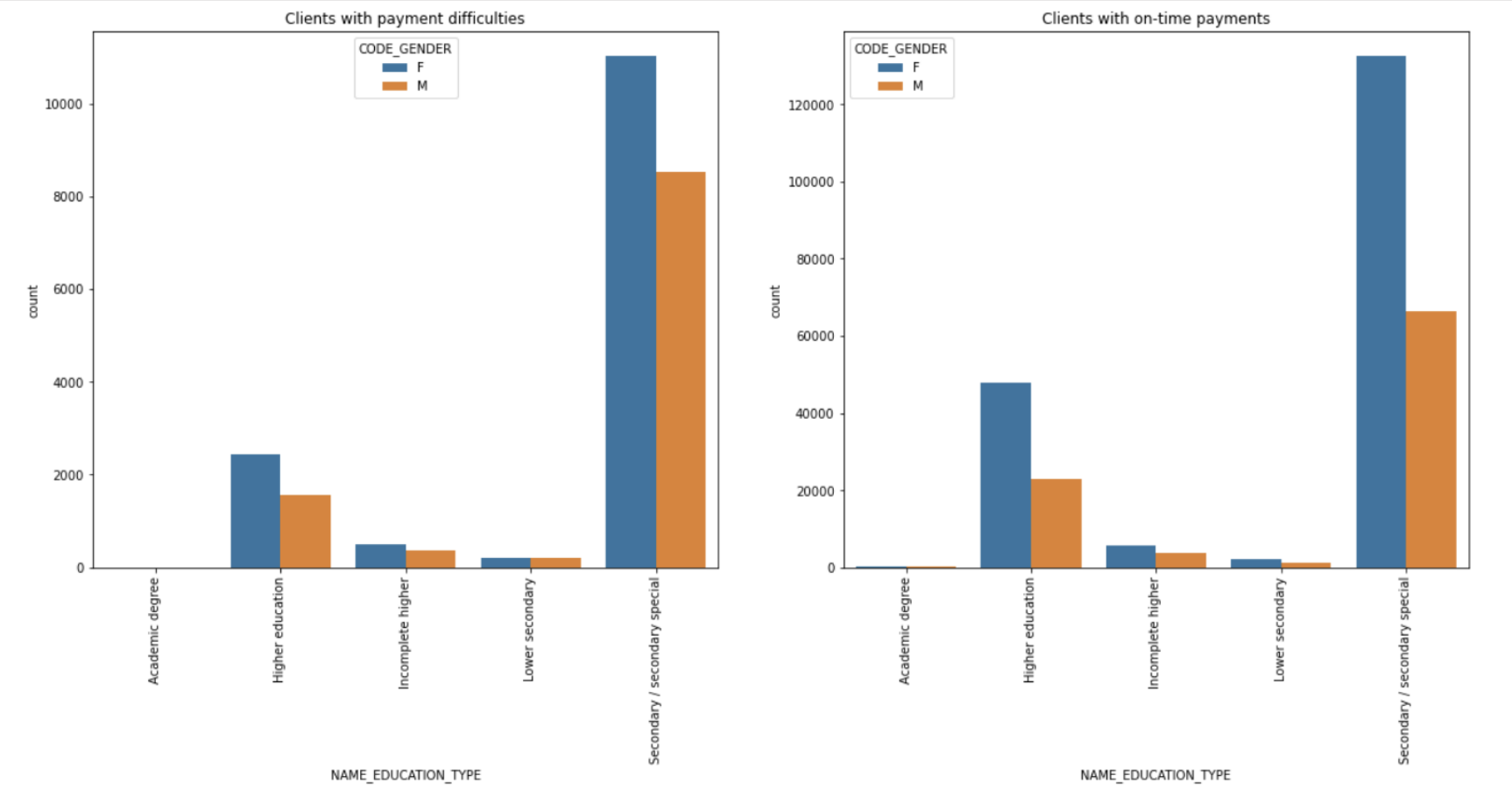
Summary- Univariate analysis of numerical variables

- For clients those are with 1-2 children, there are few more clients with On-Time Payments as compare to clients with payment difficulties.
- For clients having AMT_CREDIT > 750000 , they are more likely for doing On-Time Payments and clients having AMT_CREDIT between 200000 and 650000, they are having Payment difficulties
- For clients with Payment difficulties, AMT_INCOME_TOTAL value distribution seems a normal distribution but in case of clients with On-Time Payments, we don't see any valid pattern

BIVARIATE OR MULTIVARIATE ANALYSIS

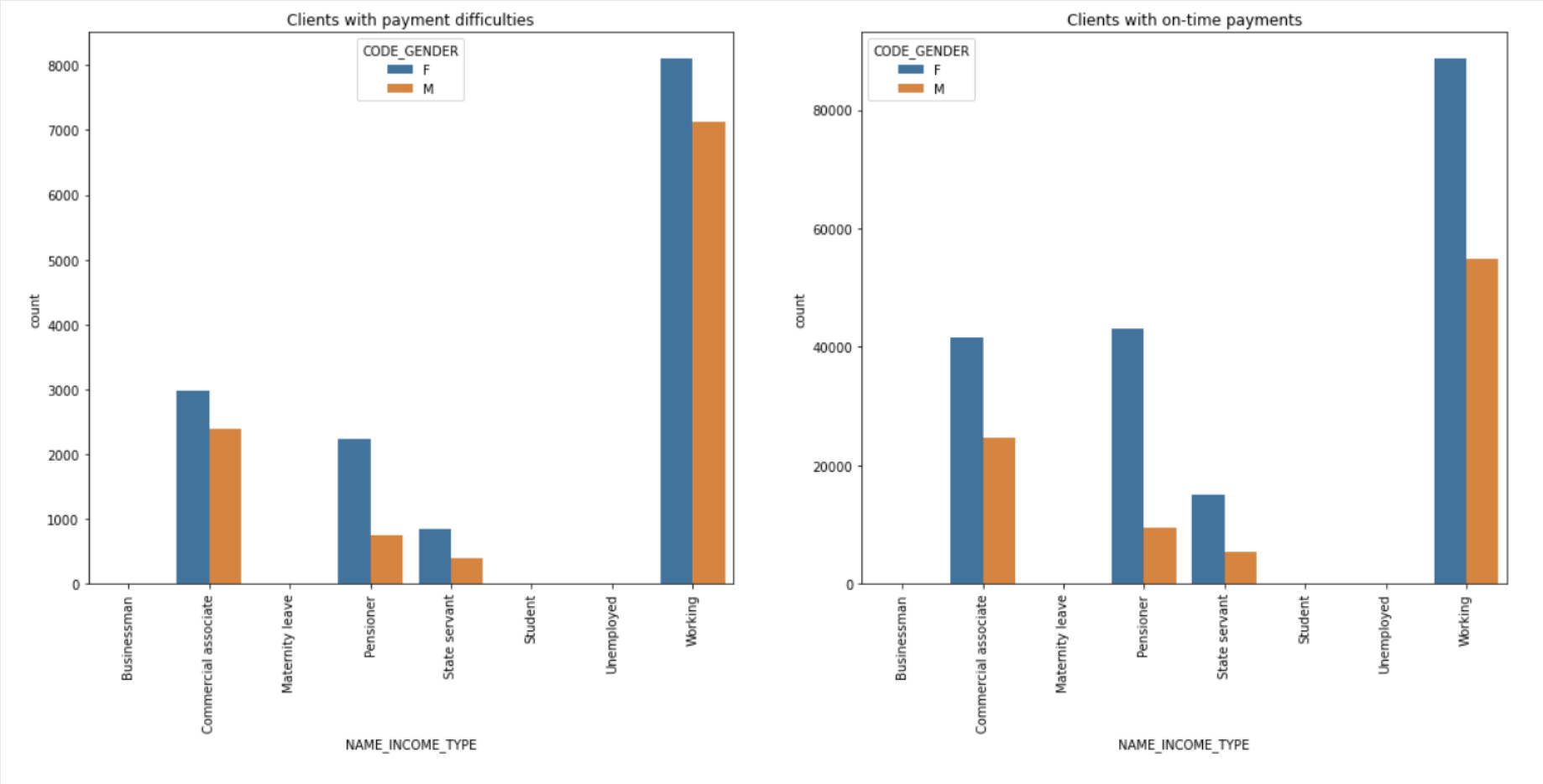
Between Categorical variables

Analysis on NAME_EDUCATION_TYPE vs CODE_GENDER



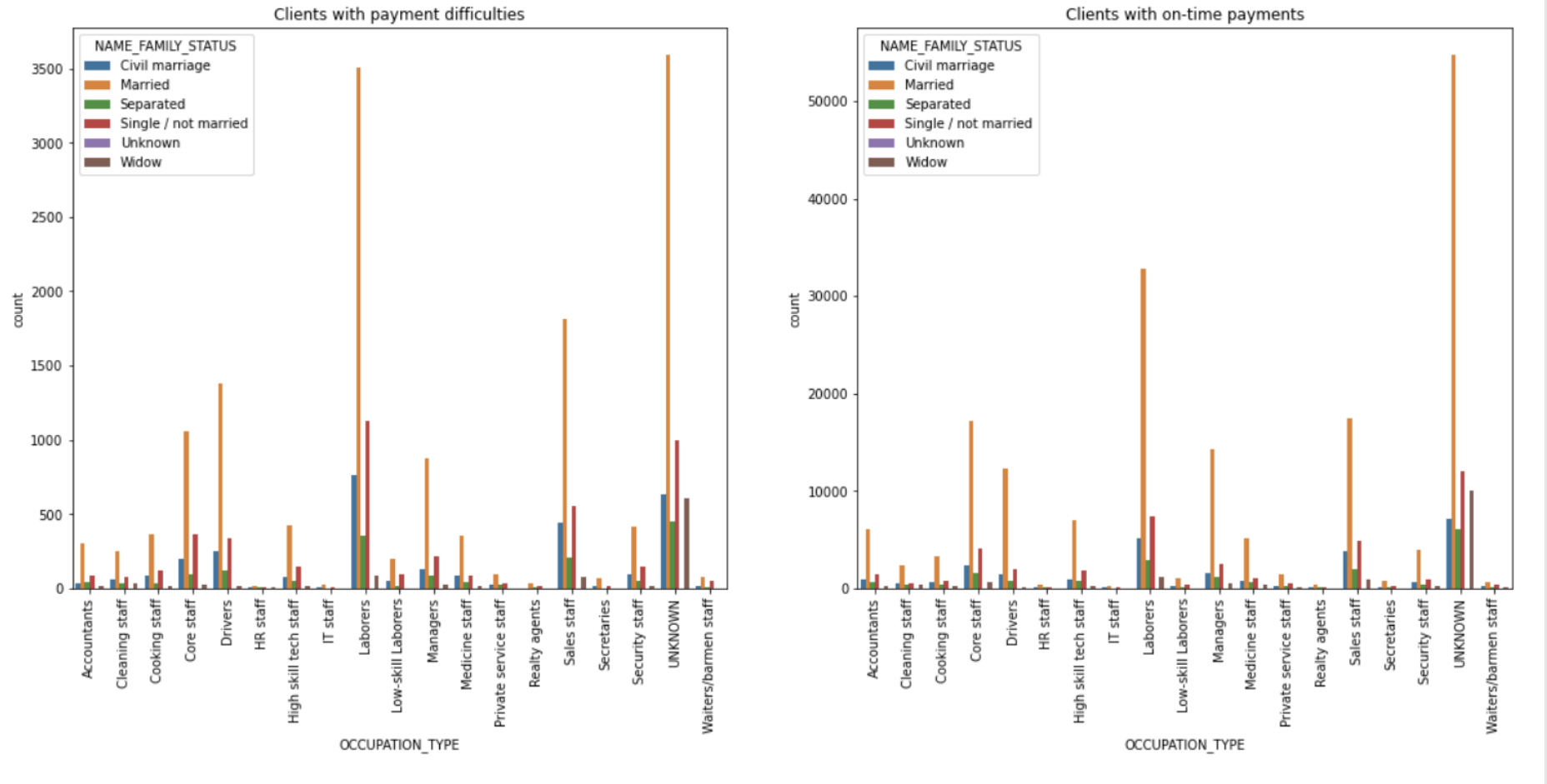
- Clients with Secondary/Secondary special education and Males are having more Payment difficulties as compared to On-Time Payments.
- Clients with Higher education and Females are having more On-Time Payments as compared to Payment difficulties.

Analysis on NAME_INCOME_TYPE vs CODE_GENDER



- Though the count is very low but clients types Businessman and Students are doing their payments On-Time.
- Clients with combination of Working and Male, is having more Payment difficulties compared to On-Time Payments.

Analysis on OCCUPATION_TYPE vs NAME_FAMILY_STATUS



- Clients who are Laborers and are Single/not married & Married, they are having more Payment difficulties as compared to On-Time Payments.
- Clients who are Drivers and are Married, they are having more Payment difficulties compared to On-Time Payments.
- clients who are Married and are Accountants have better On-Time Payments.

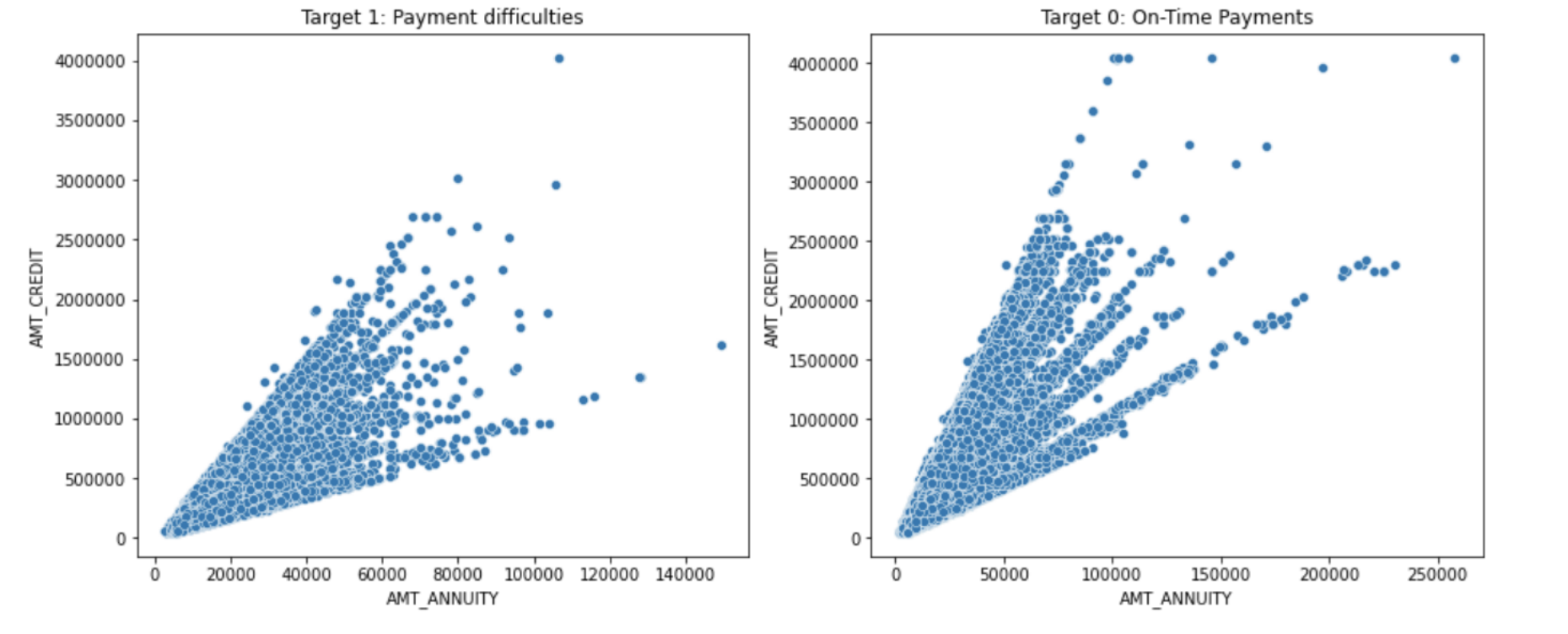
Summary-**Bivariate or Multivariate Analysis [Between Categorical variables]**

- Clients with Secondary/Secondary special education and Males are having more Payment difficulties as compared to On-Time Payments.
- Clients with Higher education and Females are having more On-Time Payments as compared to Payment difficulties.
- Though the count is very low but clients types Businessman and Students are doing their payments On-Time.
- Clients with combination of Working and Male, is having more Payment difficulties compared to On-Time Payments.
- Clients who are Laborers and are Single/not married & Married, they are having more Payment difficulties as compared to On-Time Payments.
- Clients who are Drivers and are Married, they are having more Payment difficulties compared to On-Time Payments.
- clients who are Married and are Accountants have better On-Time Payments.

BIVARIATE OR MULTIVARIATE ANALYSIS

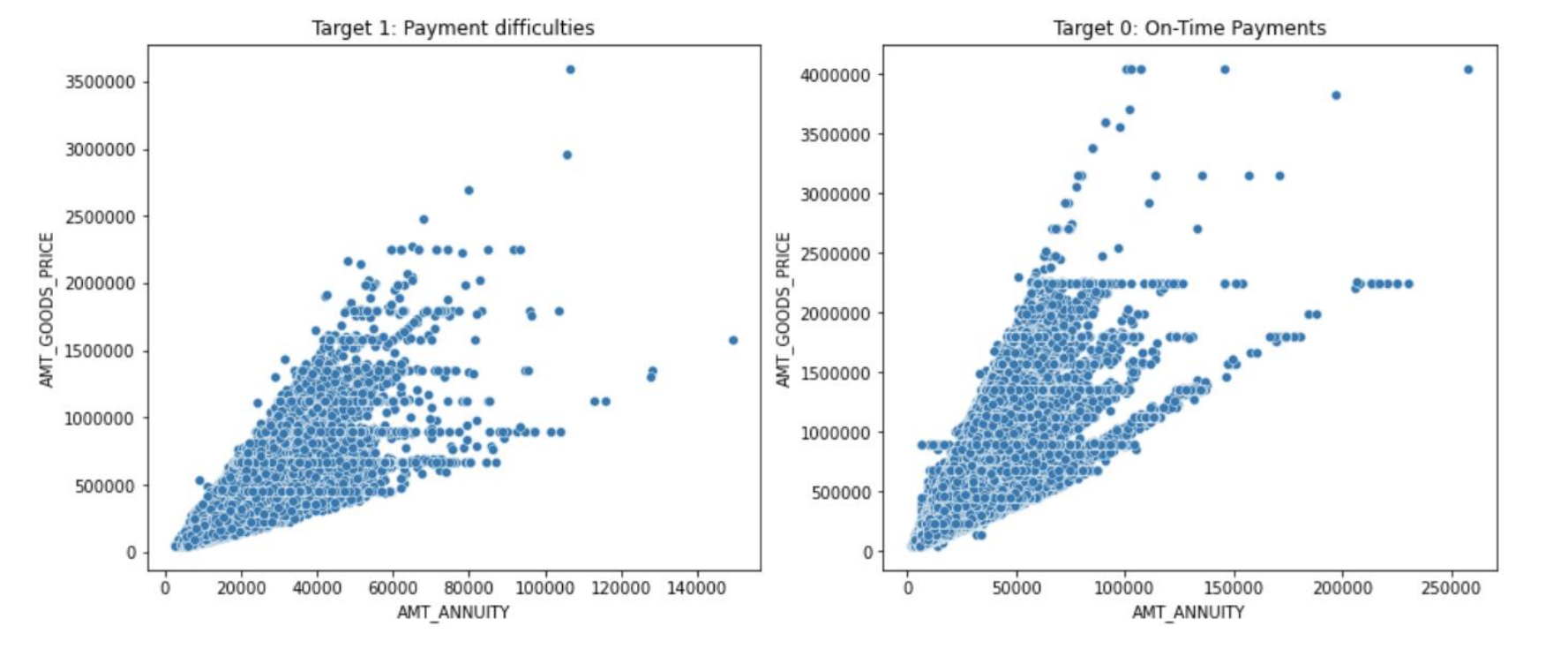
Between Continuous Variables

Analysis on AMT_ANNUITY vs AMT_CREDIT



- We can see strong positive correlation b/w AMT_ANNUITY and AMT_CREDIT.
- This imply that as Annuity will increase, Credit Amount will also increase.

Analysis on AMT_ANNUITY vs AMT_GOODS_PRICE



- We can see strong positive correlation b/w AMT_ANNUITY and AMT_GOODS_PRICE.
- This imply that as Annuity will increase , Goods price will also increase.

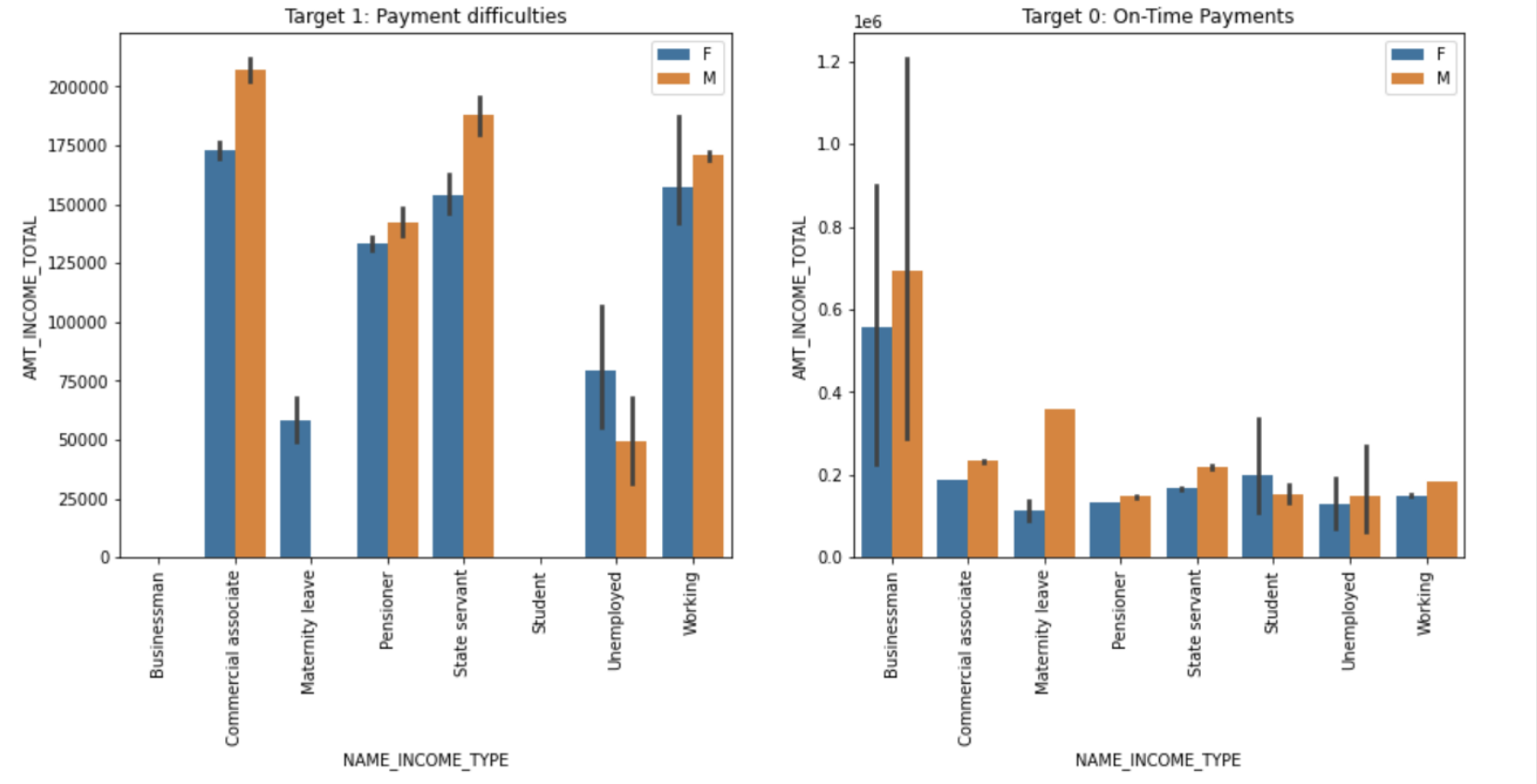
Summary- Bivariate or Multivariate Analysis [Between Continuous Variables]

- We can see strong positive correlation b/w AMT_ANNUITY and AMT_CREDIT.
- This imply that as Annuity will increase, Credit Amount will also increase.
- We can see strong positive correlation b/w AMT_ANNUITY and AMT_GOODS_PRICE.
- This imply that as Annuity will increase , Goods price will also increase.

BIVARIATE OR MULTIVARIATE ANALYSIS

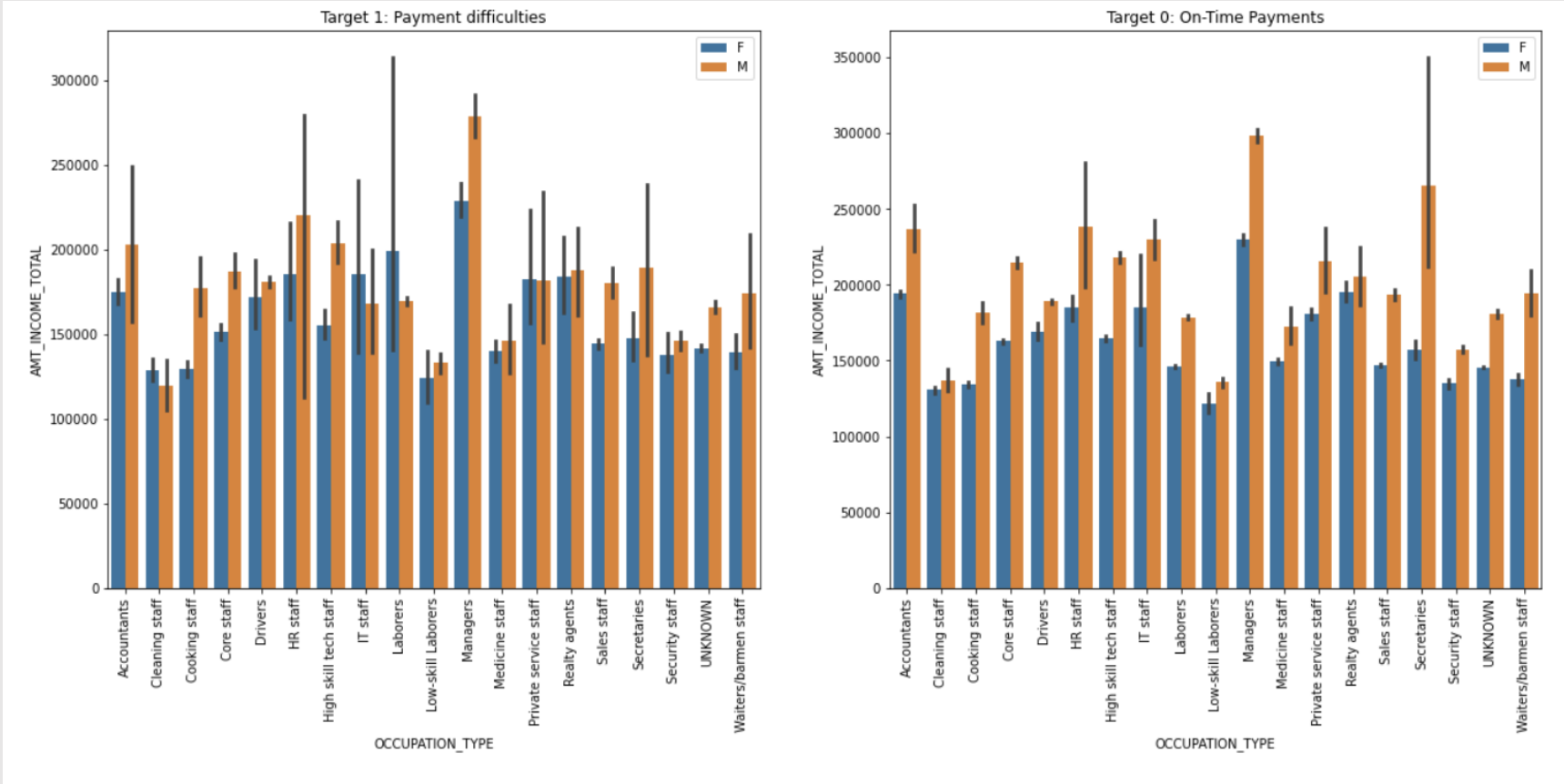
Continuous vs Categorical Variables

Analysis of AMT_INCOME_TOTAL V/S NAME_INCOME_TYPE V/S CODE_GENDER



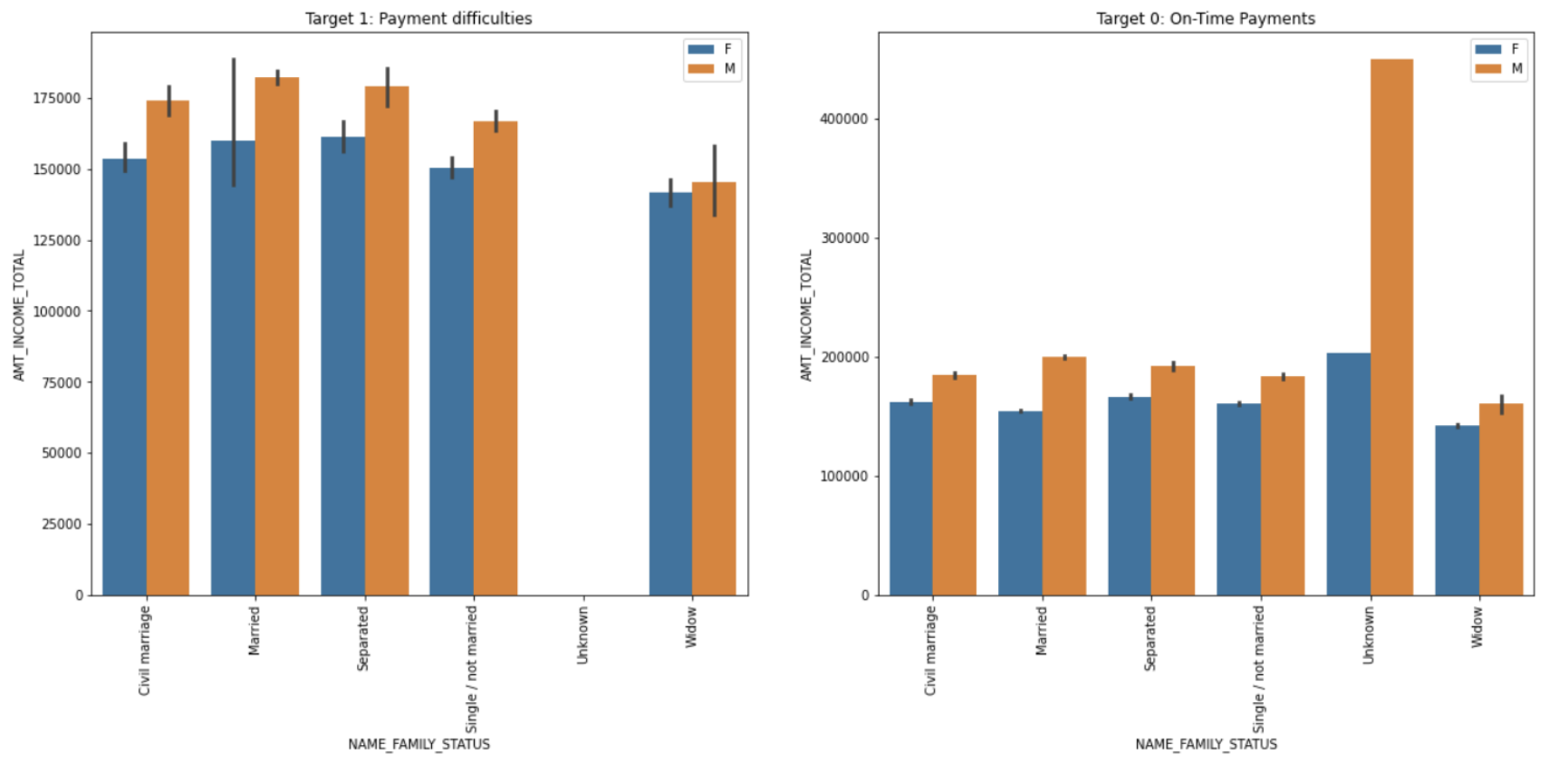
- Clients who are 'Female' and in 'Maternity Leave', they are having a very high income with On-Time Payments than Payment difficulties.
- Clients who are 'Male' and 'Unemployed' , they are having a very high income with On-Time Payments than Payment difficulties.
- Clients who are 'Businessman' and either 'Male' OR 'Female' , they are doing their payment on time.

Analysis of AMT_INCOME_TOTAL V/S OCCUPATION_TYPE V/S CODE_GENDER



- Clients who are 'Female' and 'Cleaning staff', they are having more income with On-Time Payments than Payment difficulties
- Clients who are 'Male' and 'IT staff' , they are having a very high income with On-Time Payments than Payment difficulties.
- Clients who are 'Male' and 'Laborers', they are having a very high income with On-Time Payments than Payment difficulties.

Analysis of AMT_INCOME_TOTAL V/S NAME_FAMILY_STATUS V/S CODE_GENDER



- Clients who are 'Male' and 'Married', they are having more income with On-Time Payments than Payment difficulties

PREVIOUS APPLICATION DATA

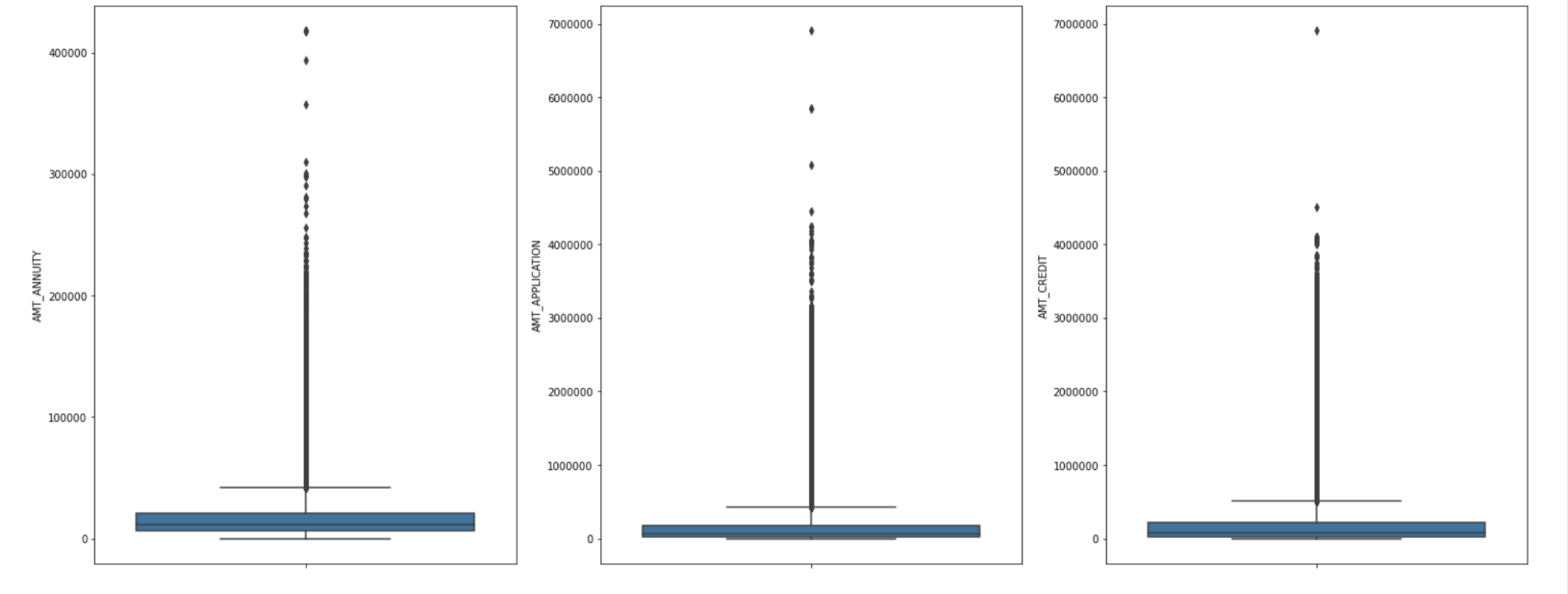
Data cleaning

Data Cleaning

- Data Sourcing: Correctly read the dataset given ‘previous_application.csv’
- Missing Values handling: Removed all columns having missing value greater than 40%
- Imputation of values: Imputation done on AMT_GOODS_PRICE, AMT_ANNUITY, CNT_PAYMENT, PRODUCT_COMBINATION
- Checked the datatypes of the columns and standardized them wherever required
- Working with the outliers in columns like AMT_ANNUITY,AMT_APPLICATION,AMT_CREDIT

Outlier Analysis

- AMT_ANNUITY,AMT_APPLICATION,AMT_CREDIT



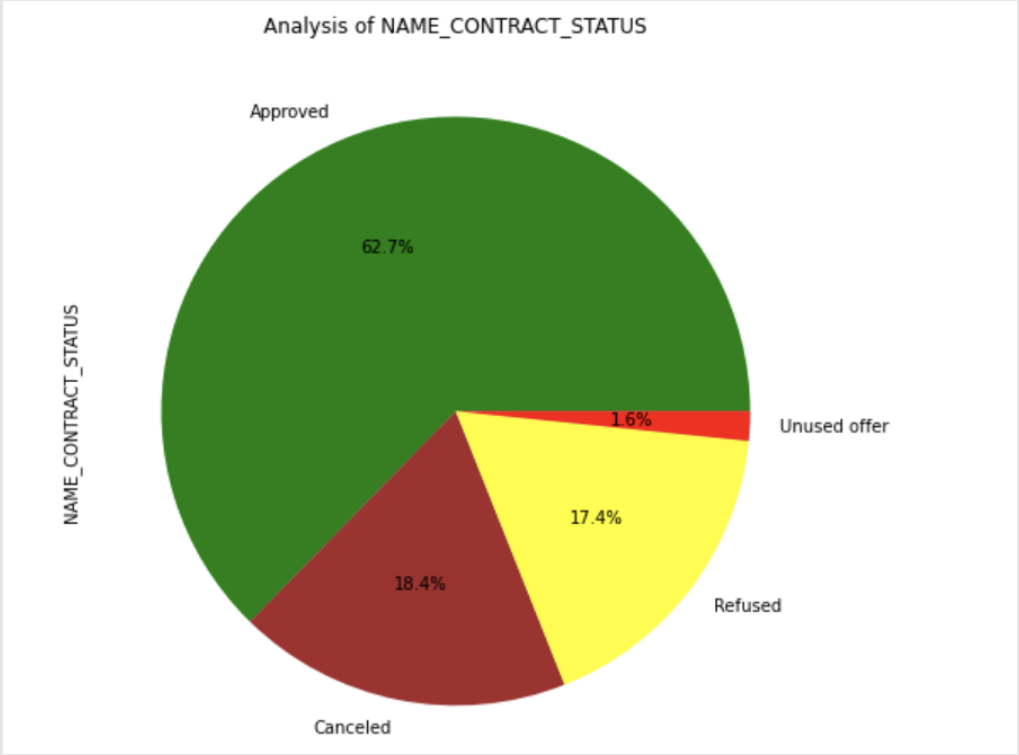
•All three columns are having good number of outliers

MERGED DATASET HANDLING

Merging the application data and previous application

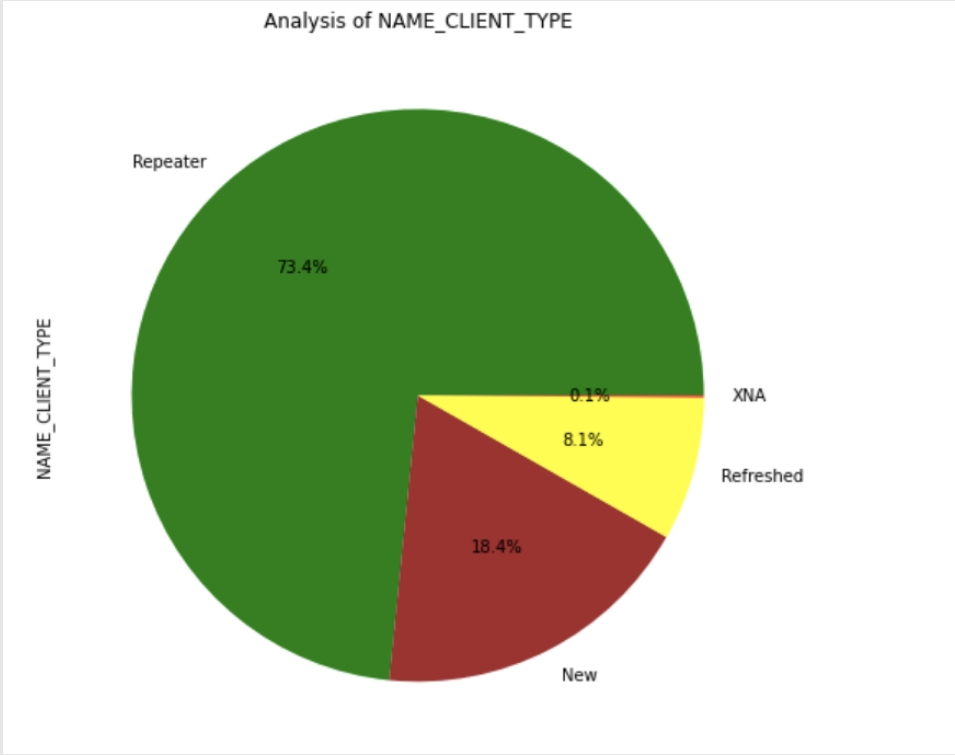
UNIVARIATE ANALYSIS OF CATEGORICAL VARIABLES

Analysis of ‘NAME_CONTRACT_STATUS’



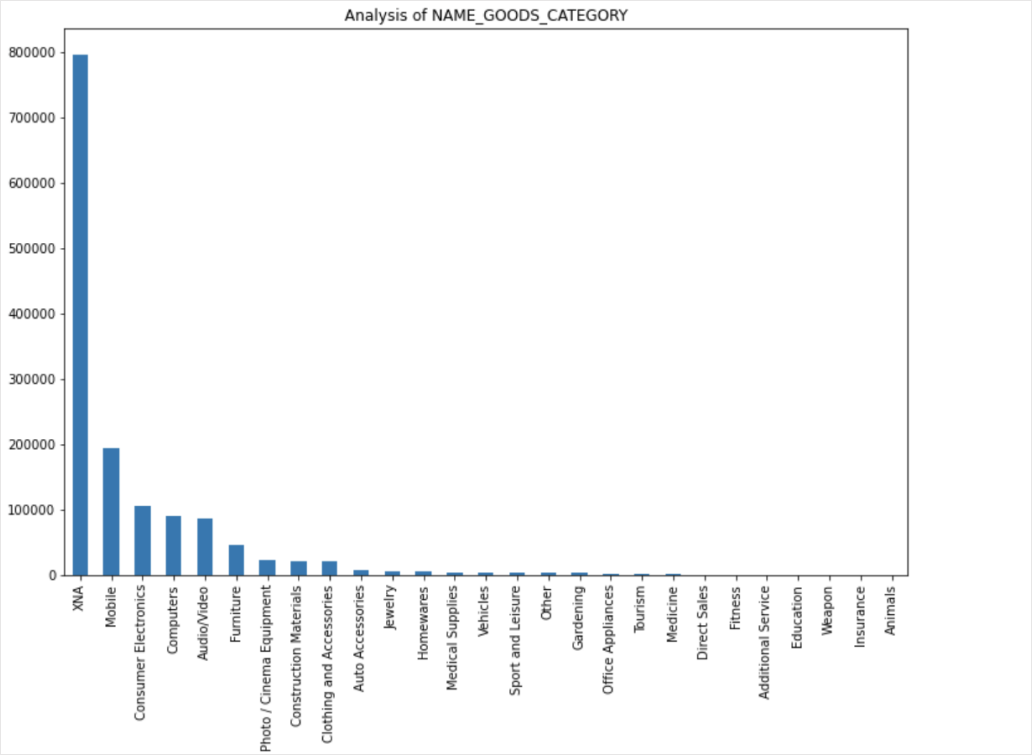
- We have 'Approved' loan status highest among all applications followed by 'Cancelled'.

Analysis of ‘NAME_CLIENT_TYPE’



- We have 'Repeater' client type highest among all applications followed by 'New'.

Analysis of ‘NAME_GOODS_CATEGORY’



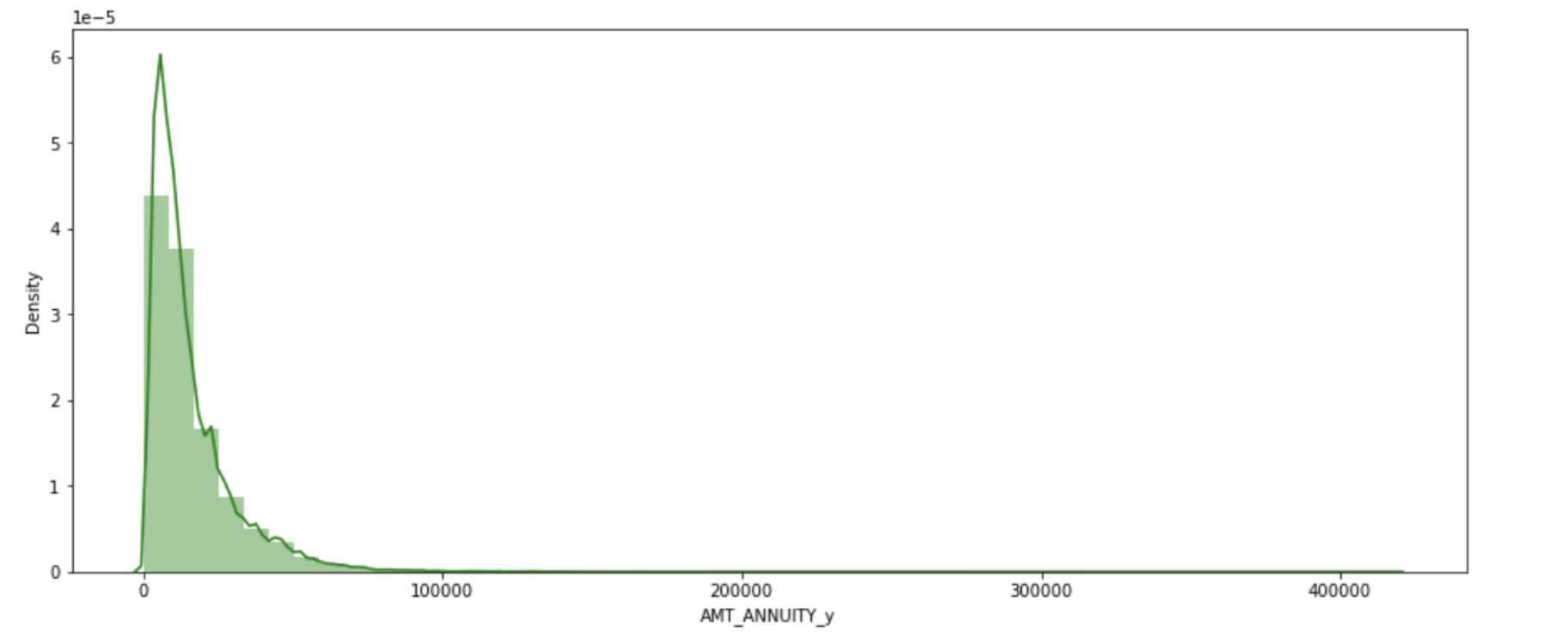
- We have 'XNA' GOODS_CATEGORY highest among all applications followed by 'Mobile'.

Summary- Univariate analysis of categorical variables

- We have 'Approved' loan status highest among all applications followed by 'Cancelled'.
- We have 'Repeater' client type highest among all applications followed by 'New'.
- We have 'XNA' GOODS_CATEGORY highest among all applications followed by 'Mobile'.

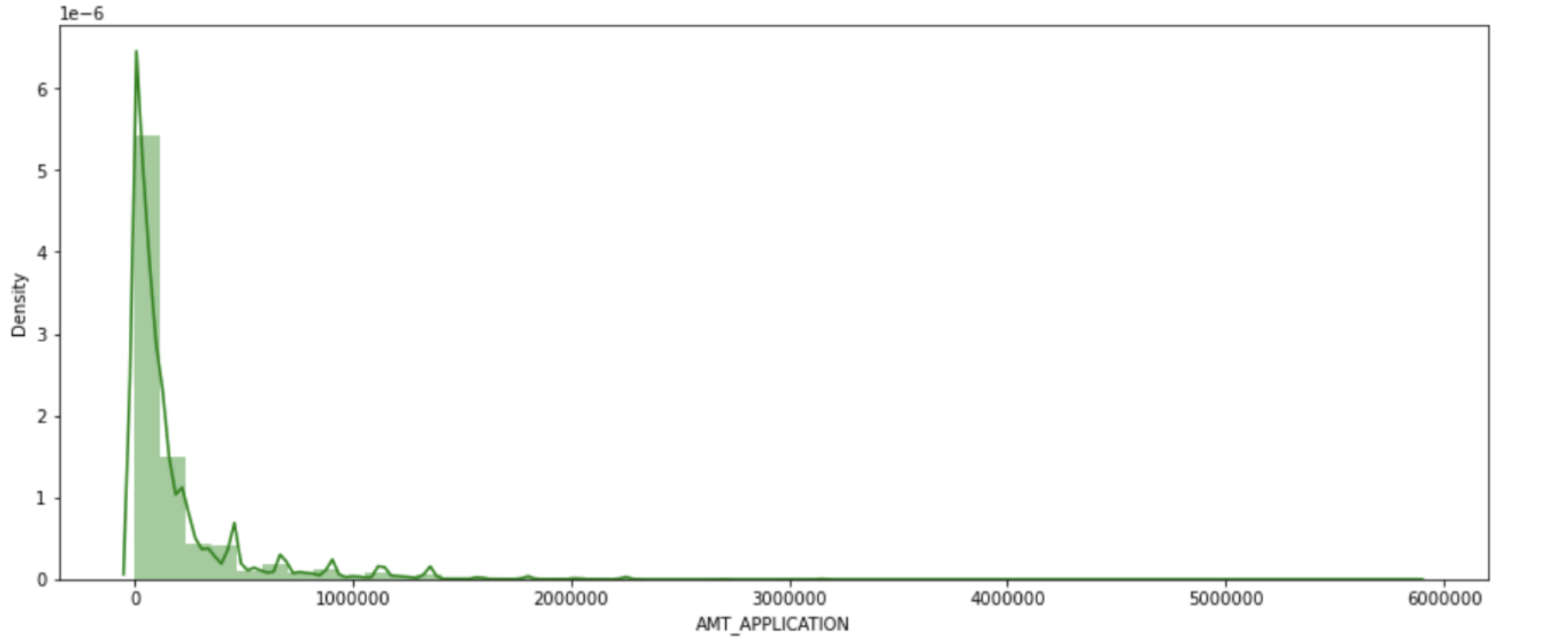
UNIVARIATE ANALYSIS OF NUMERICAL VARIABLES

Analysis of 'AMT_ANNUITY_y'



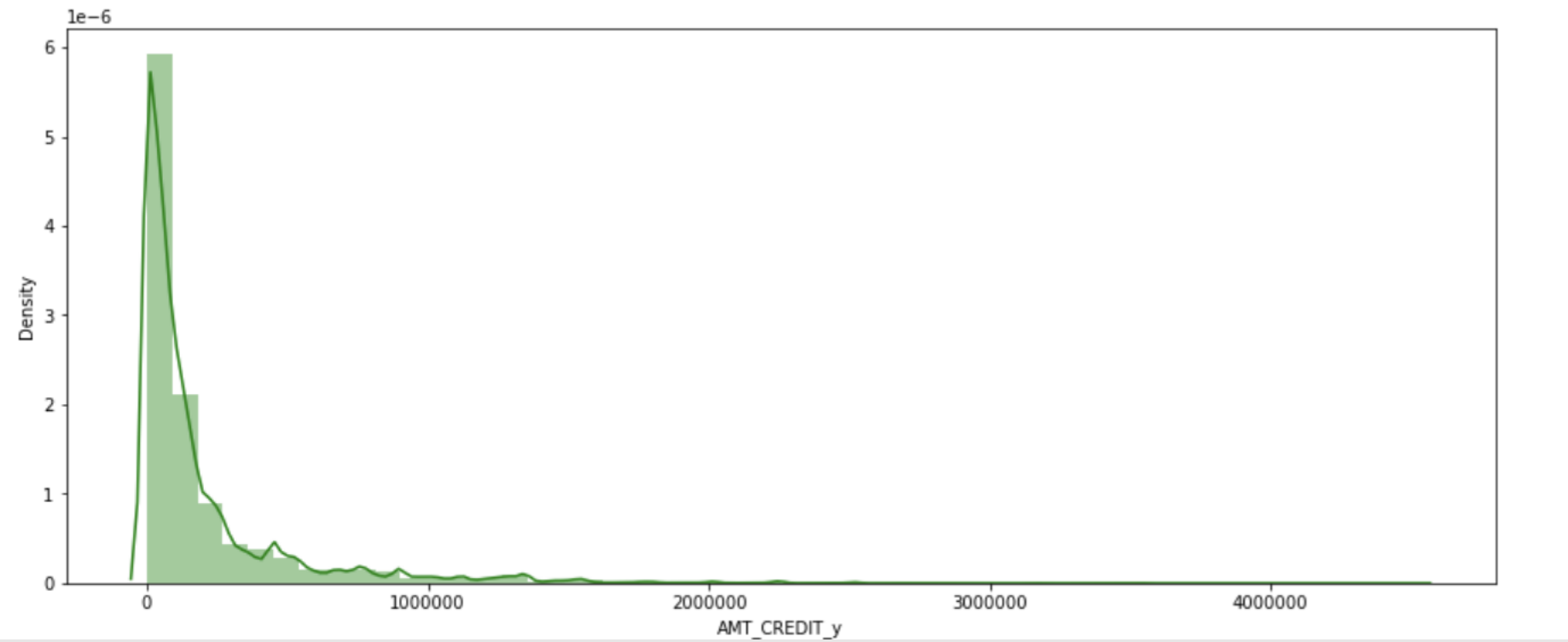
- For above , we can see amt_annuity for previous loan mostly lie less than 100000. This imply that amt_annuity is getting increased, number of clients are getting decreased.

Analysis of 'AMT_APPLICATION'



- For above , we can see most loan amount applied by clients lie less than 200000. This Imply most of the clients are applying for lesser amount.

Analysis of 'AMT_CREDIT_y'



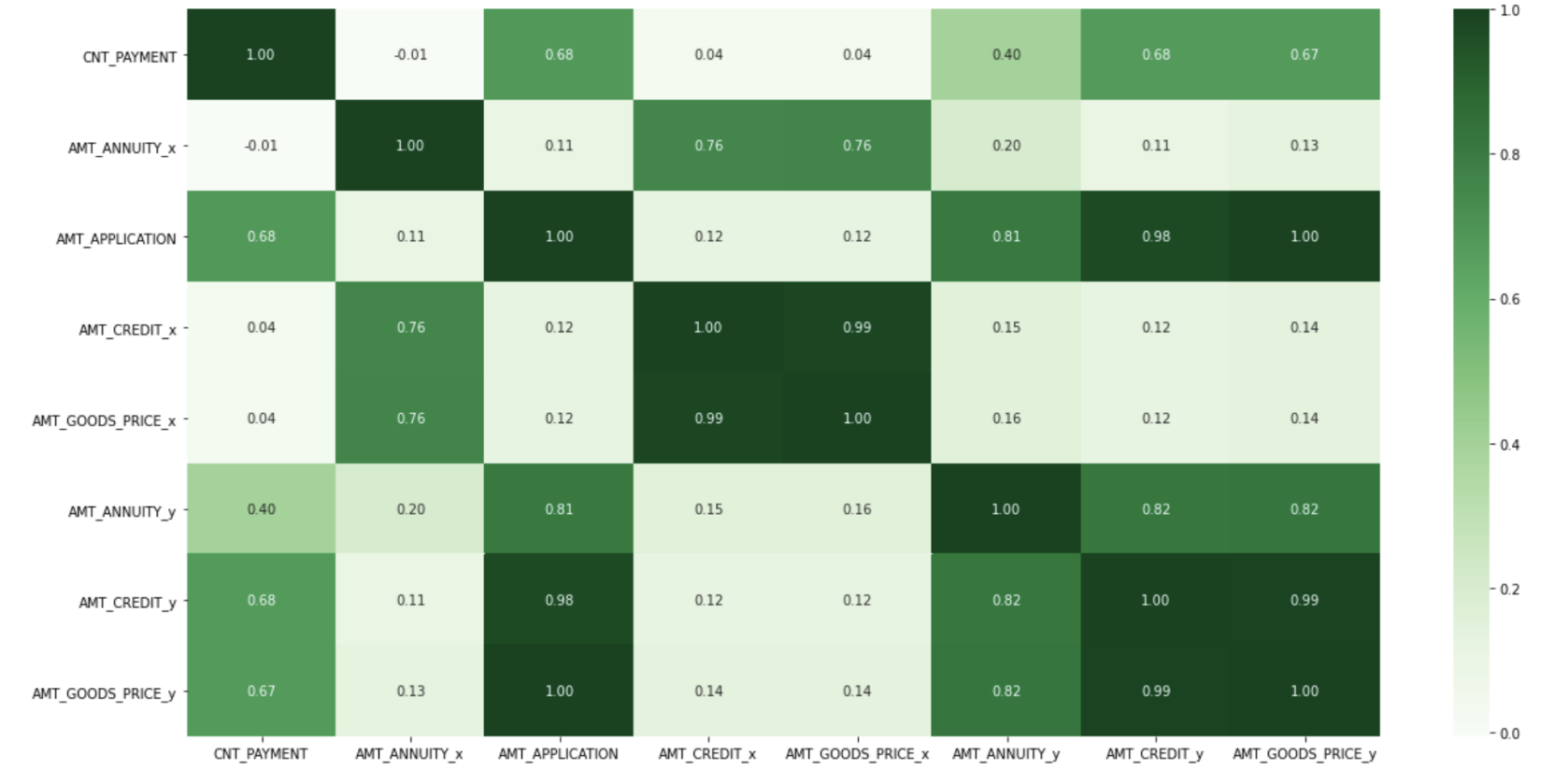
- For above , we can see most of clients received the loan amount they have applied as it is quiet similar distributed as AMT_APPLICATION.

Summary- Univariate analysis of numerical variables

- For above , we can see amt_annuity for previous loan mostly lie less than 100000. This imply that amt_annuity is getting increased, number of clients are getting decreased.
- For above , we can see most loan amount applied by clients lie less than 200000. This Imply most of the clients are applying for lesser amount.
- For above , we can see most of clients received the loan amount they have applied as it is quiet similar distributed as AMT_APPLICATION.

CORRELATION OF NUMERICAL COLUMNS

Amount and payment columns correlations

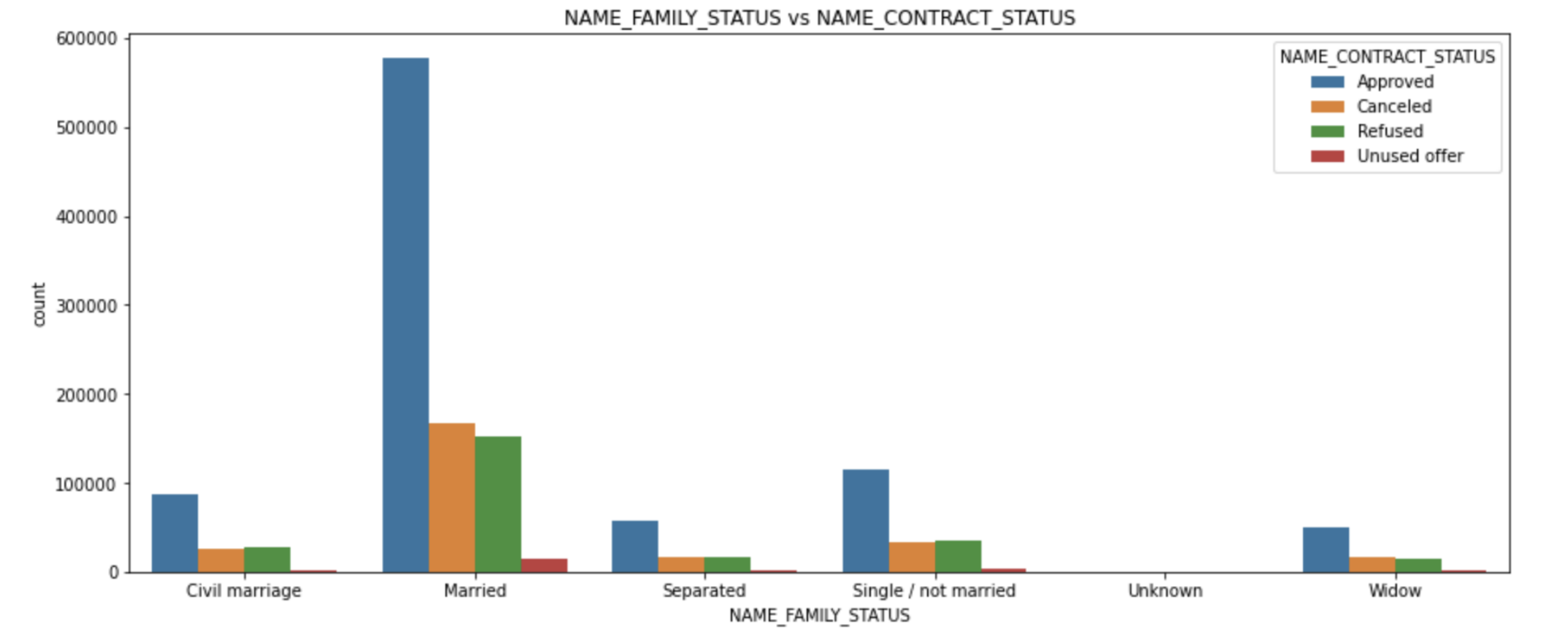


- AMT_ANNUITY_x is highly correlated to AMT_CREDIT_x and and AMT_GOODS_PRICE_x.
- CNT_PAYMENT is highly correlated to AMT_APPLICATION,AMT_CREDIT_y and AMT_GOODS_PRICE_y.
- AMT_CREDIT_y is highly correlated to AMT_APPLICATION,AMT_GOODS_PRICE_y and AMT_ANNUITY_y.
- AMT_GOODS_PRICE_x is highly correlated to AMT_CREDIT_x.

BIVARIATE/ MULTIVARIATE ANALYSIS

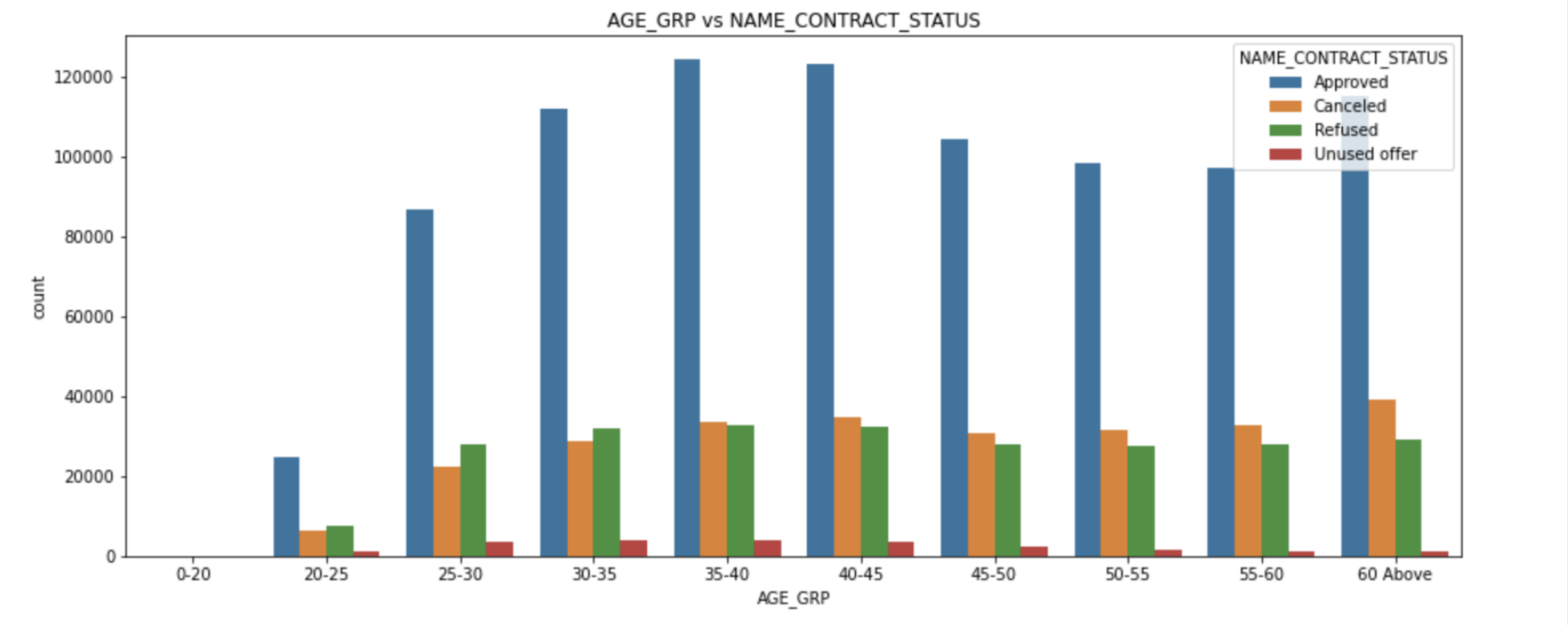
Between Categorical Variables

Analysis on 'NAME_FAMILY_STATUS' vs 'NAME_CONTRACT_STATUS'



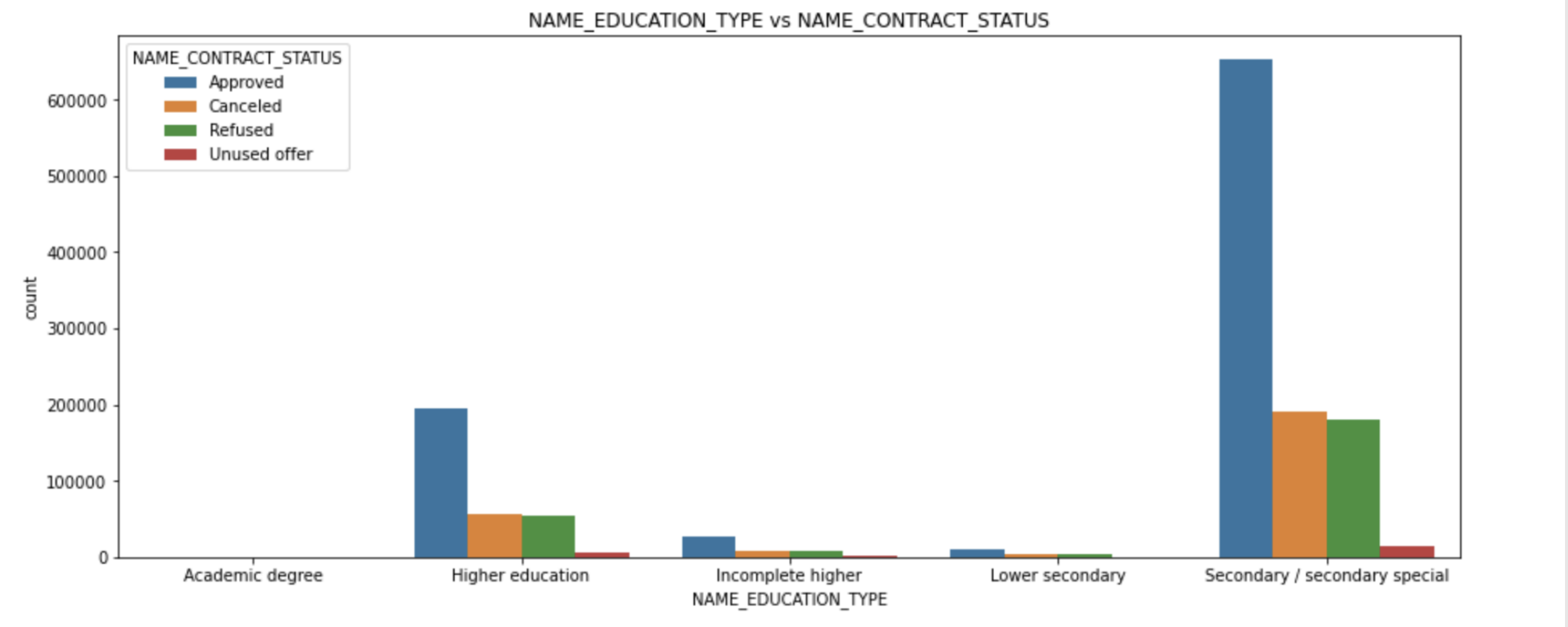
- Clients who are 'Married' , are having most number of loan approvals followed by 'Single/not married' clients.

Analysis on 'AGE_GRP' vs 'NAME_CONTRACT_STATUS'



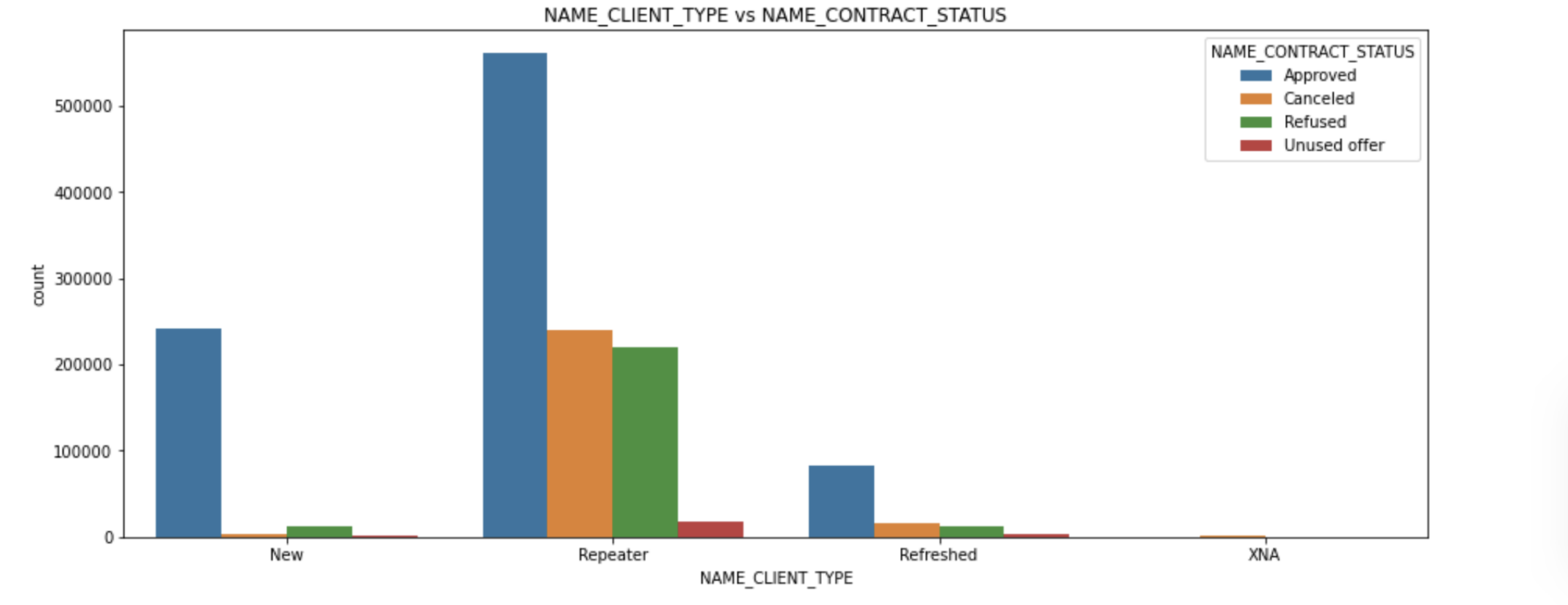
- Clients who are in age group 35-40 ,they are having most number of loan approvals followed by clients having age 40-45.

Analysis on 'NAME_EDUCATION_TYPE' vs NAME_CONTRACT_STATUS'



- Clients who belong to education type 'Secondary/secondary special', they are having most number of loan approvals followed by clients having 'Higher education'.

Analysis on 'NAME_CLIENT_TYPE' vs 'NAME_CONTRACT_STATUS'



- Clients who are 'Repeater', they are having most number of loan approvals followed by 'New' clients.

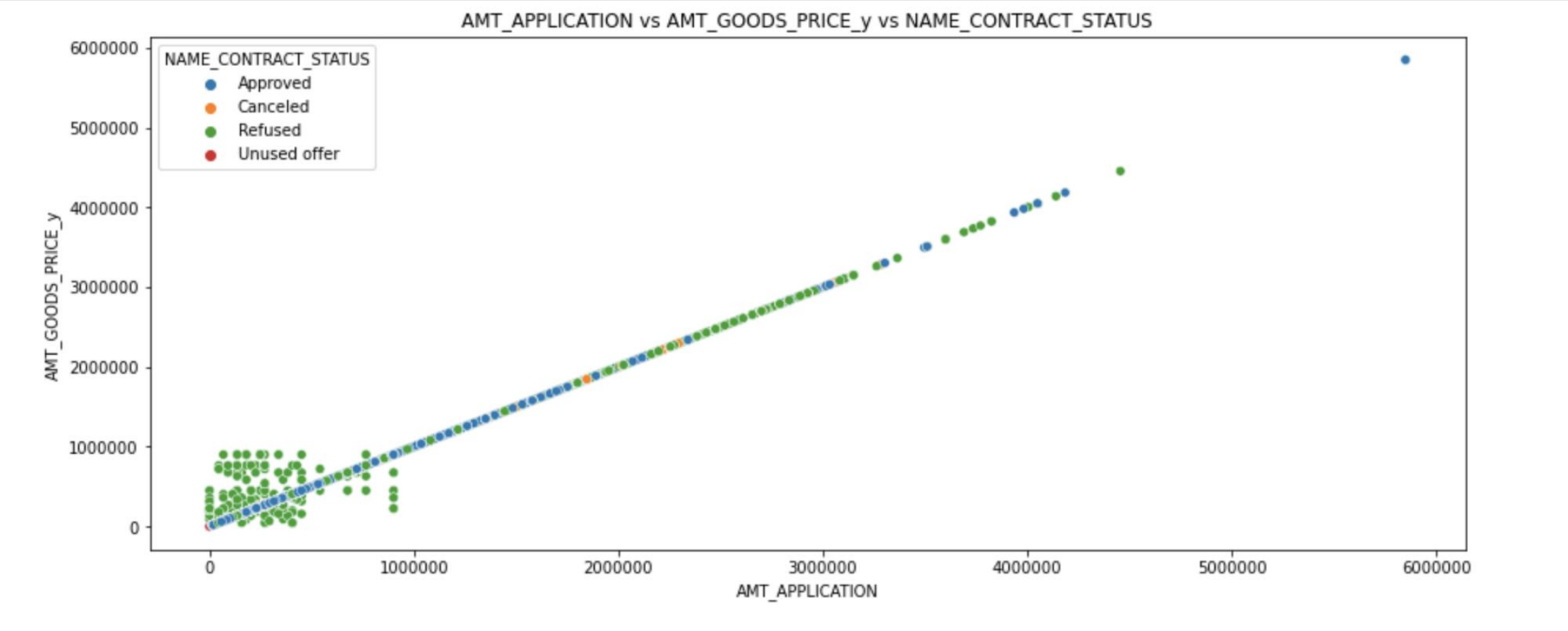
Summary- Bivariate/ Multivariate Analysis [Between Categorical Variables]

- Clients who are 'Married' , are having most number of loan approvals followed by 'Single/not married' clients.
- Clients who are in age group 35-40 ,they are having most number of loan approvals followed by clients having age 40-45.
- Clients who belong to education type 'Secondary/secondary special', they are having most number of loan approvals followed by clients having 'Higher education'.
- Clients who are 'Repeater', they are having most number of loan approvals followed by 'New' clients.

BIVARIATE/ MULTIVARIATE ANALYSIS

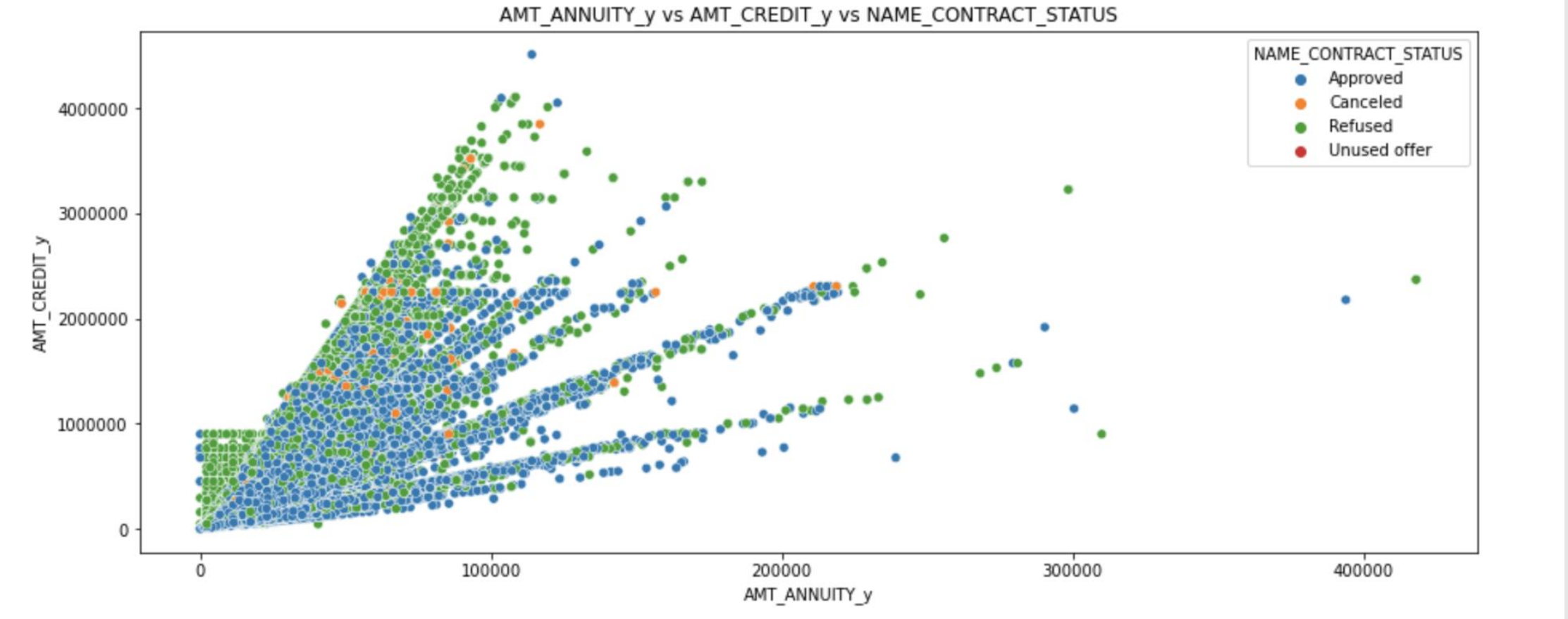
Between Continuous Variables

Analysis on ‘AMT_APPLICATION’ vs ‘AMT_GOODS_PRICE_y’ vs ‘NAME_CONTRACT_STATUS’



- AMT_APPLICATION & AMT_GOODS_PRICE_y are having strong positive correlation.

Analysis on AMT_ANNUITY_y vs AMT_CREDIT_y vs NAME_CONTRACT_STATUS



- AMT_ANNUITY_y & AMT_CREDIT_y are having strong positive correlation.

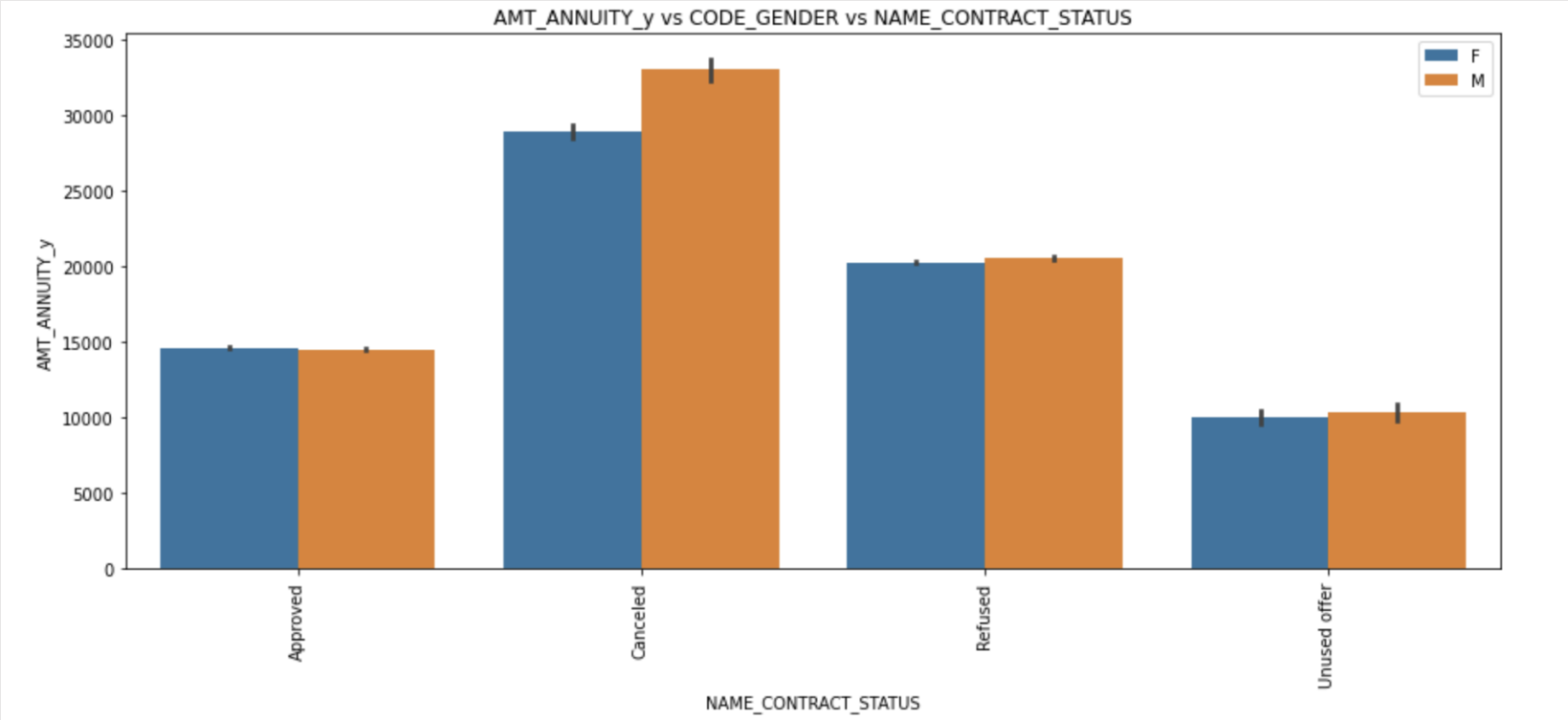
Summary- Bivariate/ Multivariate Analysis [Between Continuous Variables]

- AMT_APPLICATION & AMT_GOODS_PRICE_y are having strong positive correlation.
- AMT_ANNUITY_y & AMT_CREDIT_y are having strong positive correlation.

BIVARIATE/ MULTIVARIATE ANALYSIS

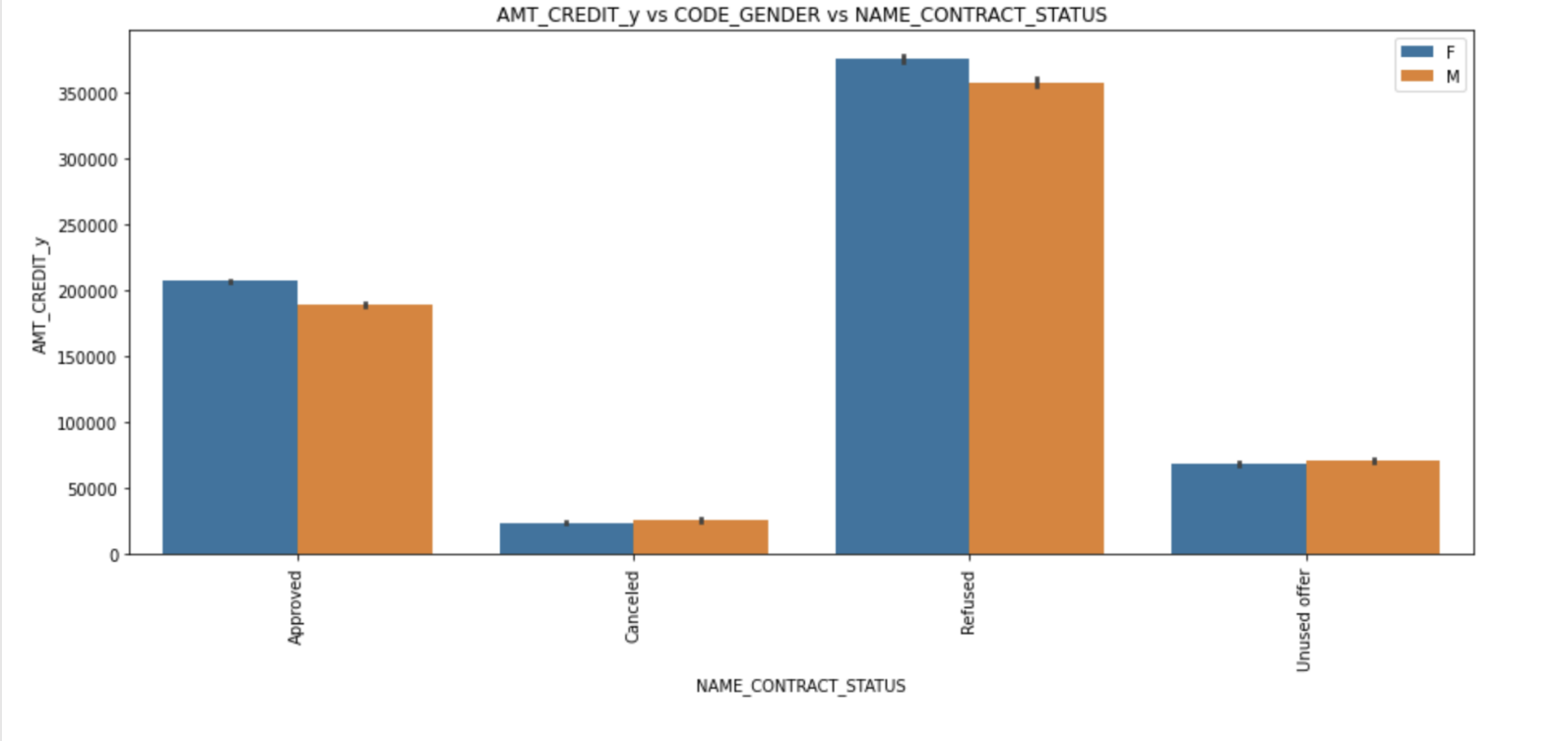
Continuous vs Categorical Variables

'AMT_ANNUITY_y' vs 'CODE_GENDER' vs 'NAME_CONTRACT_STATUS'



- Clients who are "Male" and loan application status is 'Cancelled', they have higher AMT_ANNUITY as compare to female.
- Clients with loan application status 'Approved', they are having almost same AMT_ANNUITY for both male and female.

'AMT_CREDIT_y' vs 'CODE_GENDER' vs 'NAME_CONTRACT_STATUS'



- Clients who are "Female" and loan application status is 'Refused', they have applied higher AMT_CREDIT as compare to male.
- Clients who are "Female" and loan application status is 'Approved', they have applied higher AMT_CREDIT as compare to male

Summary- Bivariate/ Multivariate Analysis [Continuous vs Categorical Variables]

- Clients who are "Male" and loan application status is 'Cancelled', they have higher AMT_ANNUITY as compare to female.
- Clients with loan application status 'Approved', they are having almost same AMT_ANNUITY for both male and female.
- Clients who are "Female" and loan application status is 'Refused', they have applied higher AMT_CREDIT as compare to male.
- Clients who are "Female" and loan application status is 'Approved', they have applied higher AMT_CREDIT as compare to male

CONCLUSION

Below Type of Clients should be targeted while approving loan applications:

- Clients with type Repeater.
- Clients with Occupation type Student and Businessman
- Clients who are married.
- Clients with gender 'Female'
- Clients with higher income.
- Clients with more employment experience.
- Clients with less children.(0-2)
- Clients with age between 35 to 45.
- Clients with academic degree either Secondary/Secondary special or higher education.