CREDIT EDA CASE STUDY

By:PRATIBHA VERMA

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 to 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.

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 at different stages of the process.

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

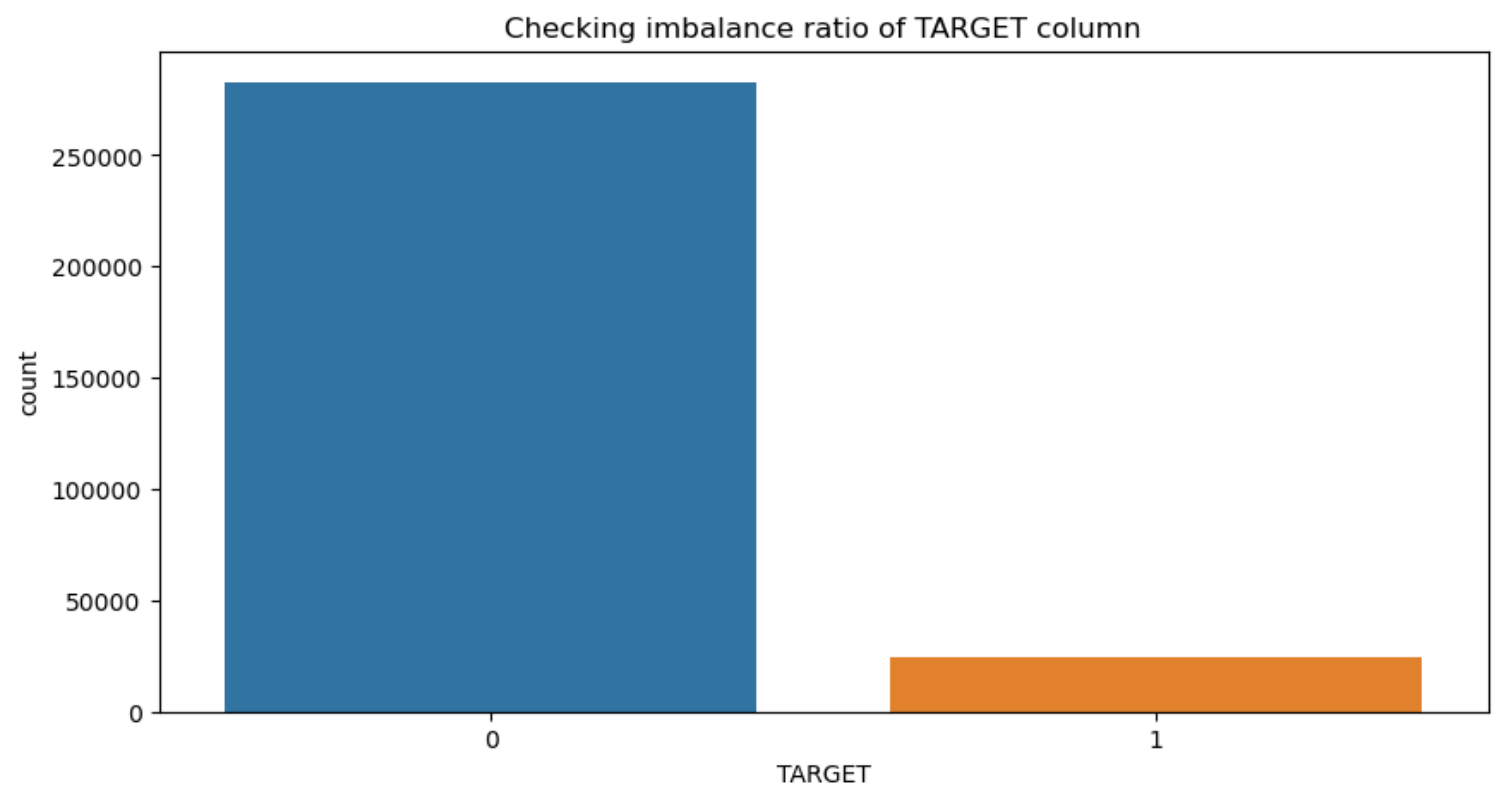
1. 'application_data.csv' contains all the information of the client at the time of application.
The data is about whether a **client has payment difficulties.**

2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data on 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.

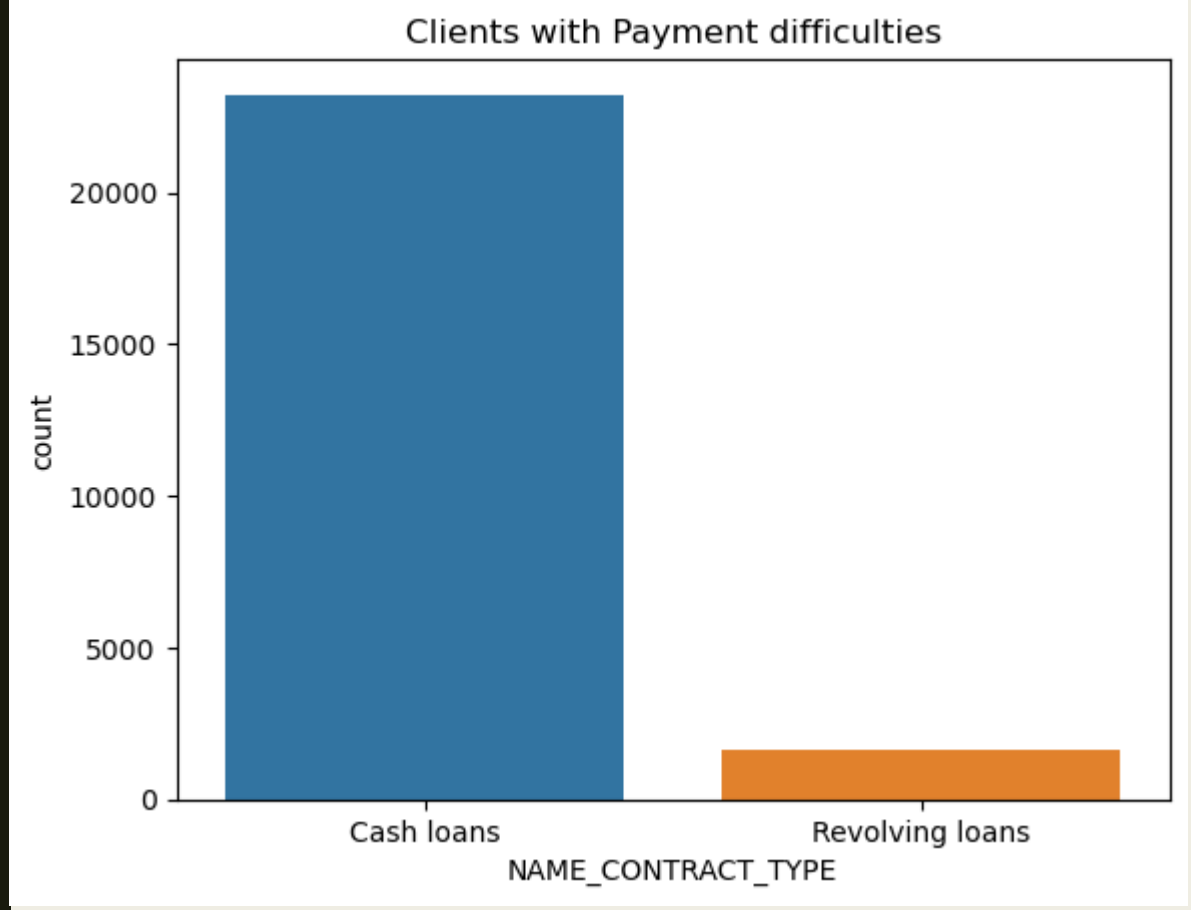
CHECKING IMBALANCE FOR TARGET VARIABLE

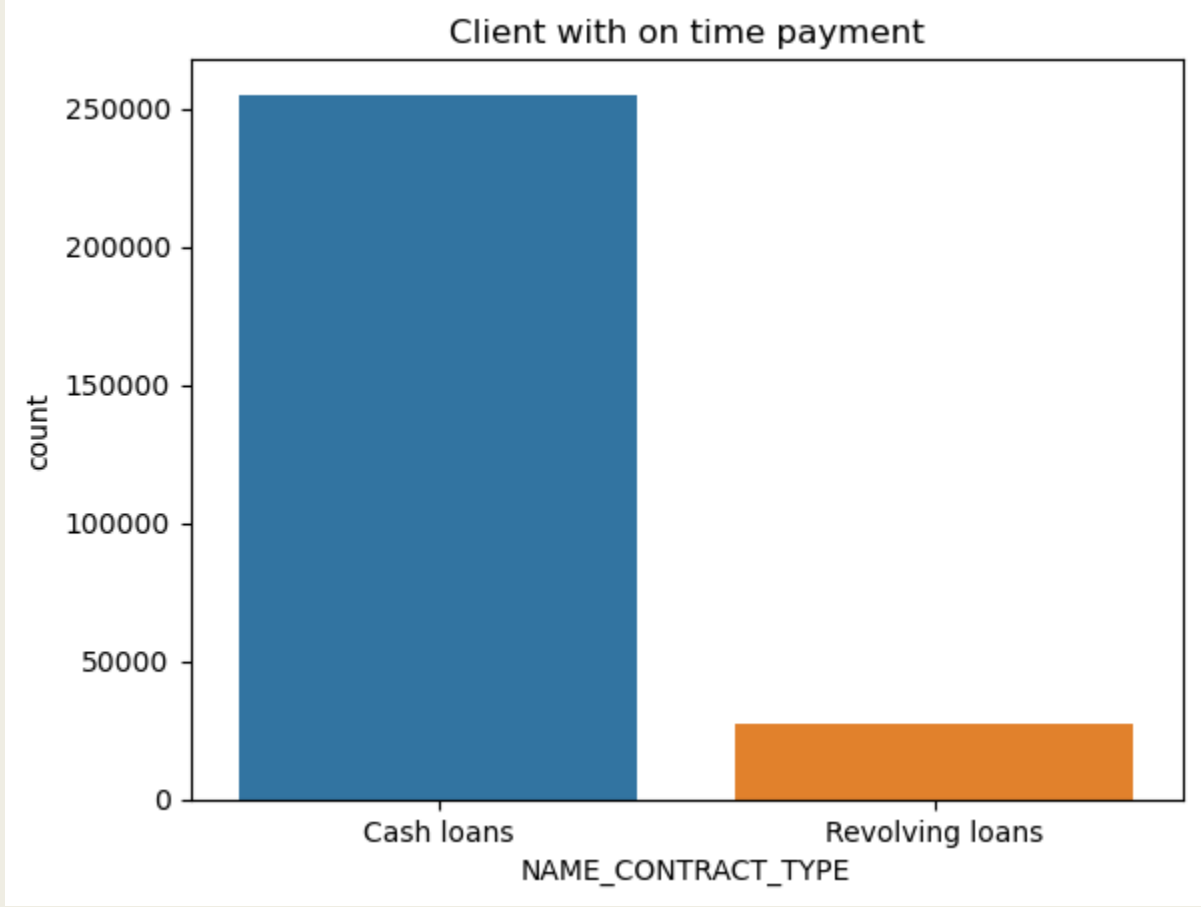
ANALYSIS OF IMBALANCE FOR "TARGET"



- TARGET value 1 represents 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). This is only 8.07% of the data.
- TARGET value 0 represents all other cases than 1. This is 91.93% of the data.

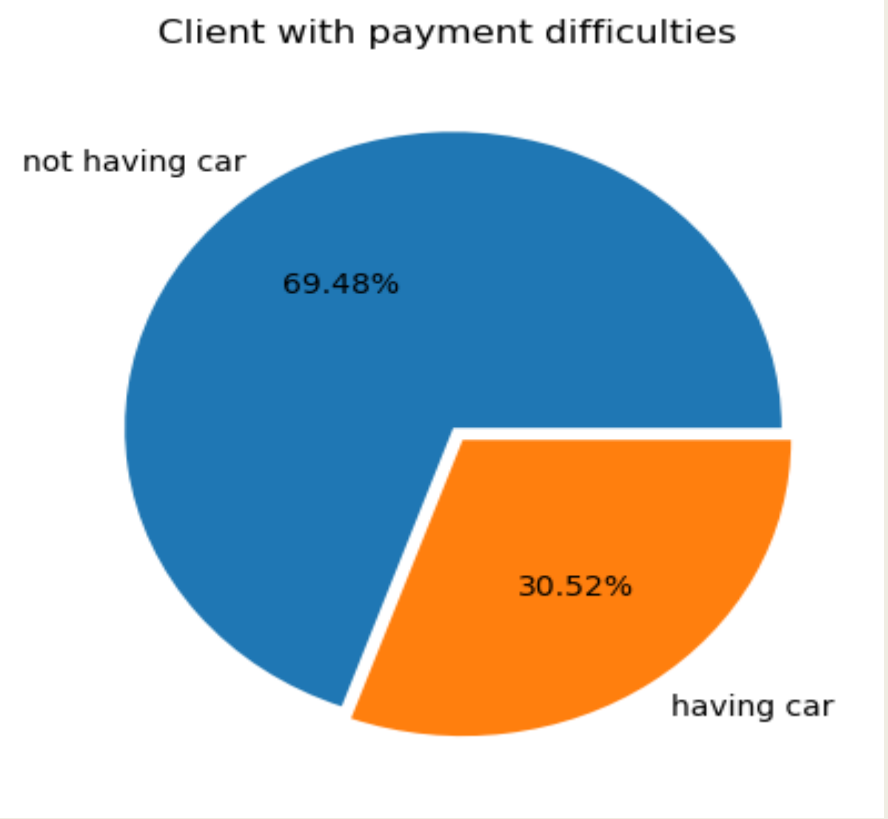
CATEGORICAL UNIVAIRATE ANALYSIS FOR TARGET

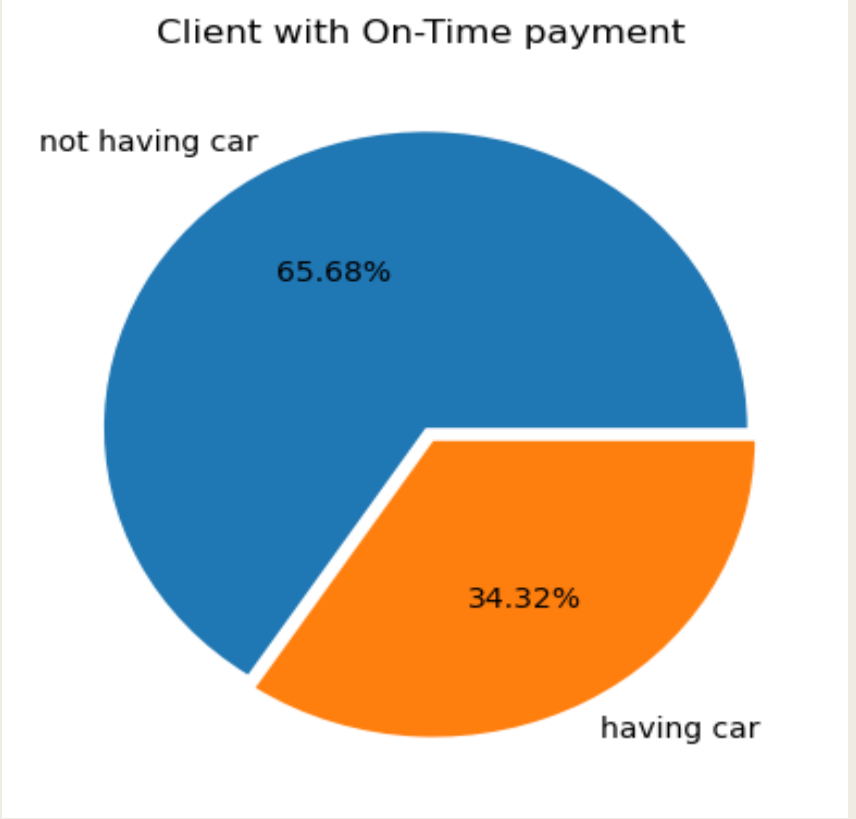




ANALYSIS OF "NAME_CONTRACT_TYPE"

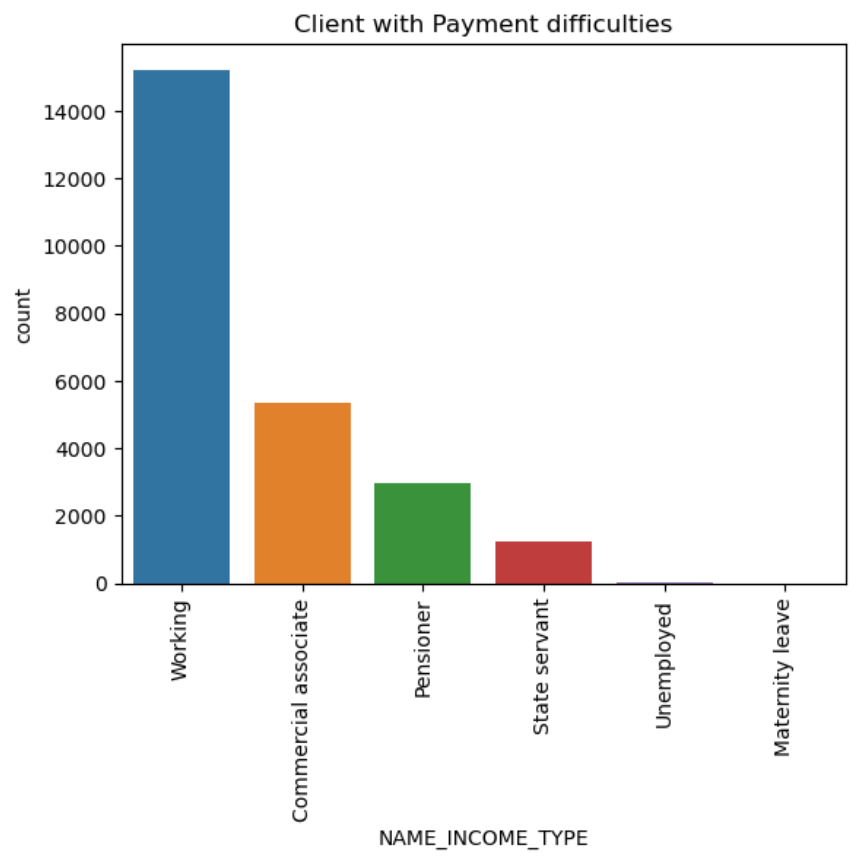
•- `NAME_CONTRACT_TYPE` column does not provide any conclusive evidence in favor of clients with payment difficulties OR on-time payment

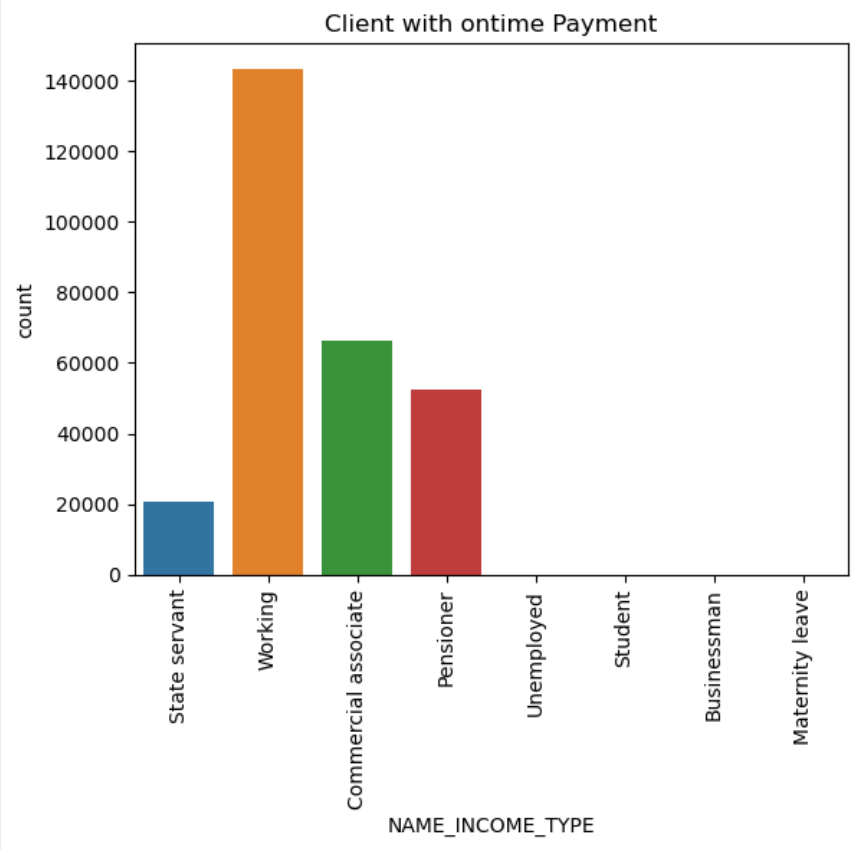




ANALYSIS OF "FLAG_OWN_CAR"

•- `FLAG_OWN_CAR` column does not provide any conclusive evidence in favor of clients with payment difficulties OR on-time payment



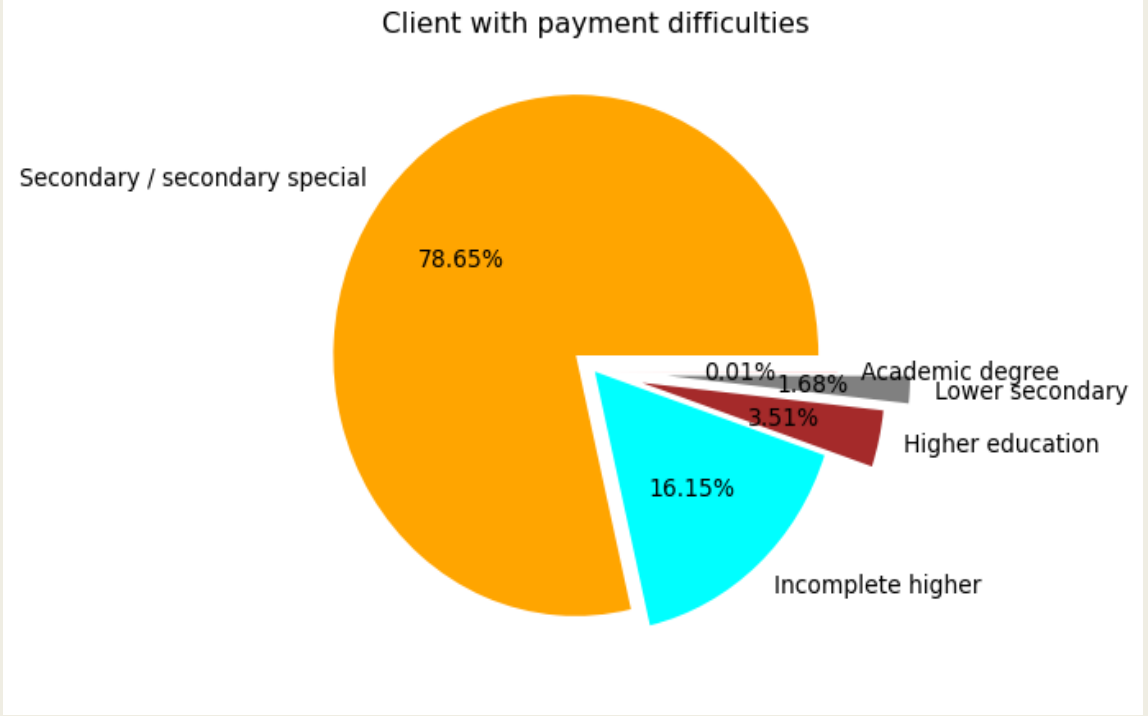


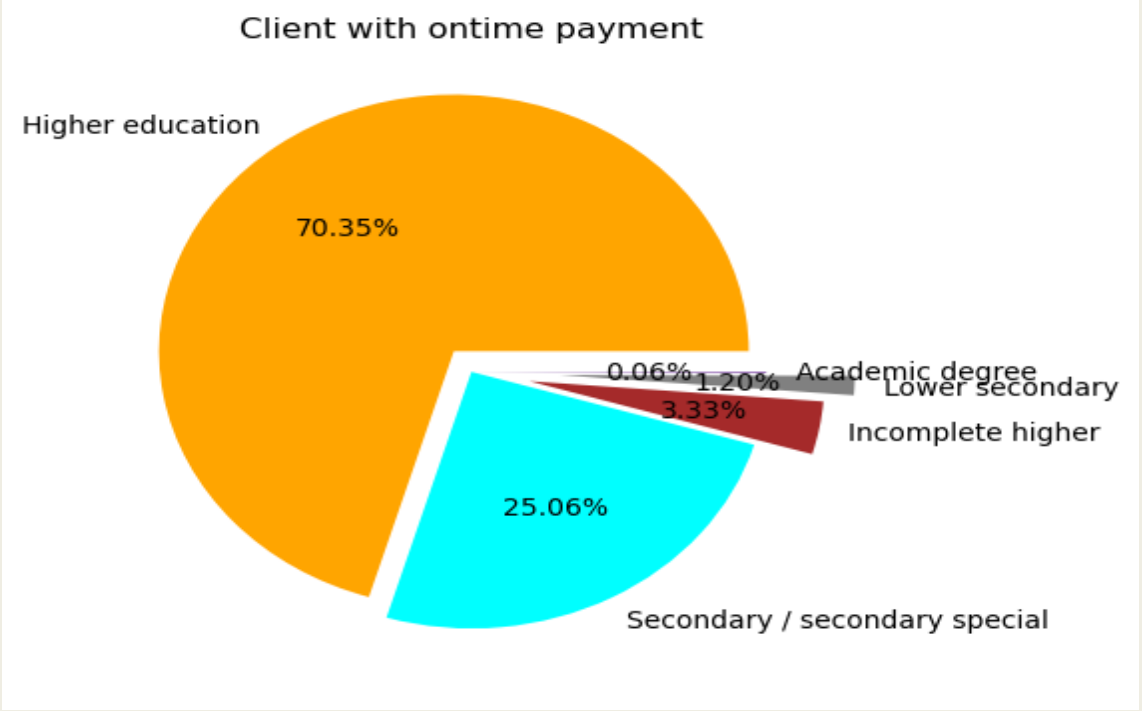
ANALYSIS OF "NAME_INCOME_TYPE

•- Pensioners have better on-time payments. This is a weak correlation.

•- Students don't have Payment difficulties. In this case, total students have only 18 observations and should be treated as a weak correlation .

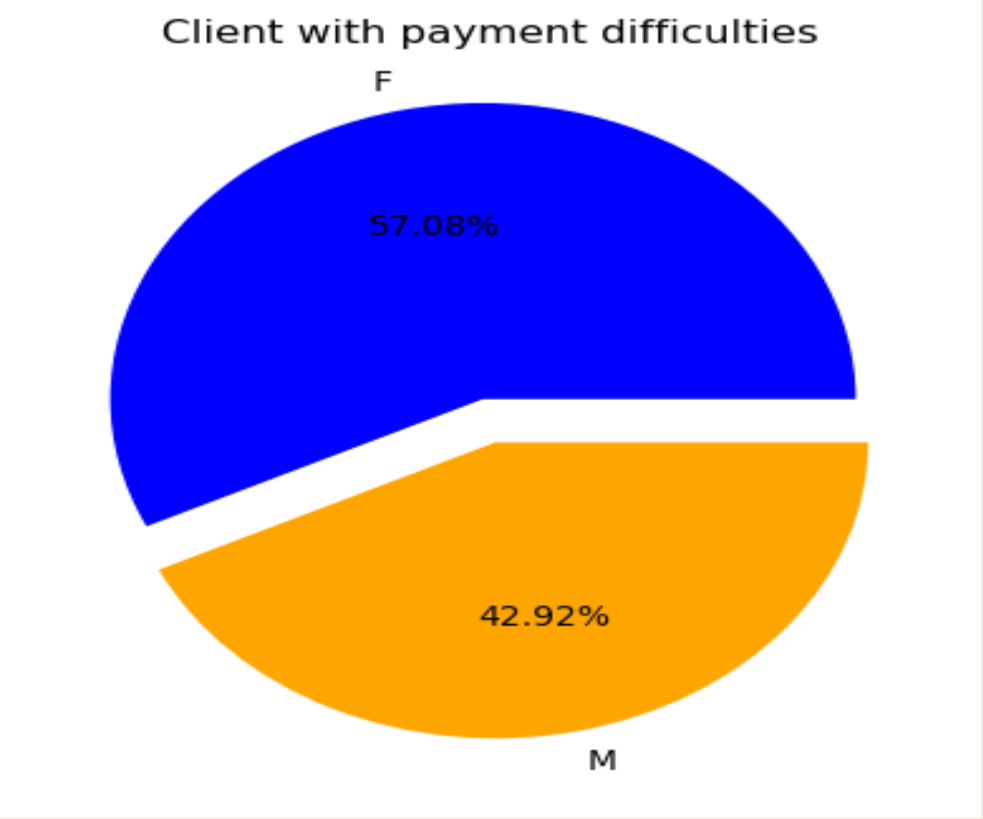
•- Businessmen don't have Payment difficulties. In this case, Businessmen have only 10 observations and should be treated as a weak correlation.

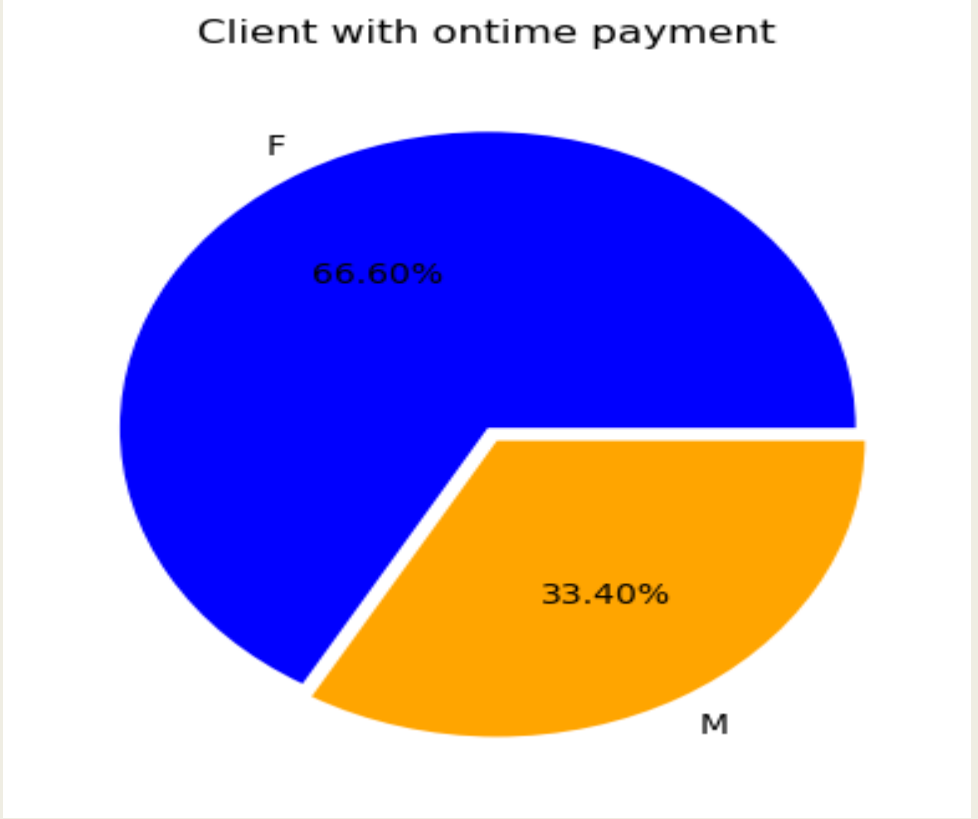




Analysis of "NAME_EDUCATION_TYPE"

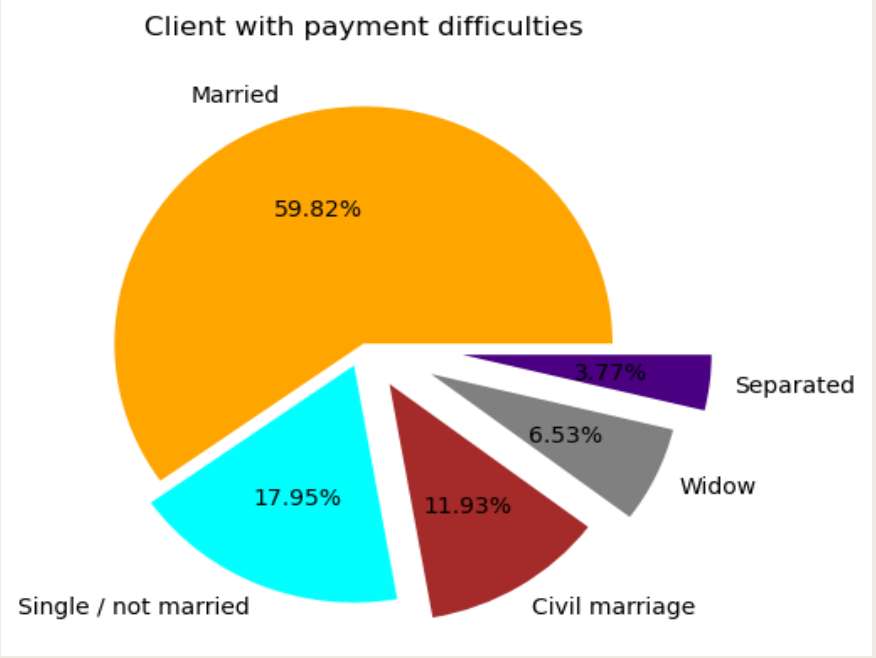
•- Clients with 'Higher education' have less payment difficulties. However, this is a weak correlation

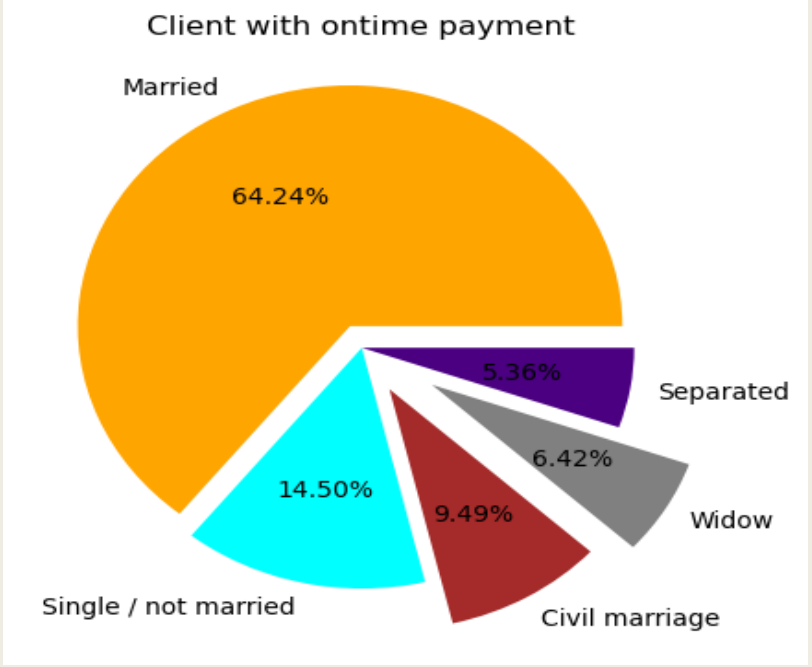




Analysis of "CODE_GENDER"

- - `CODE_GENDER` column provides a weak inference that "Male" clients have more payment difficulties

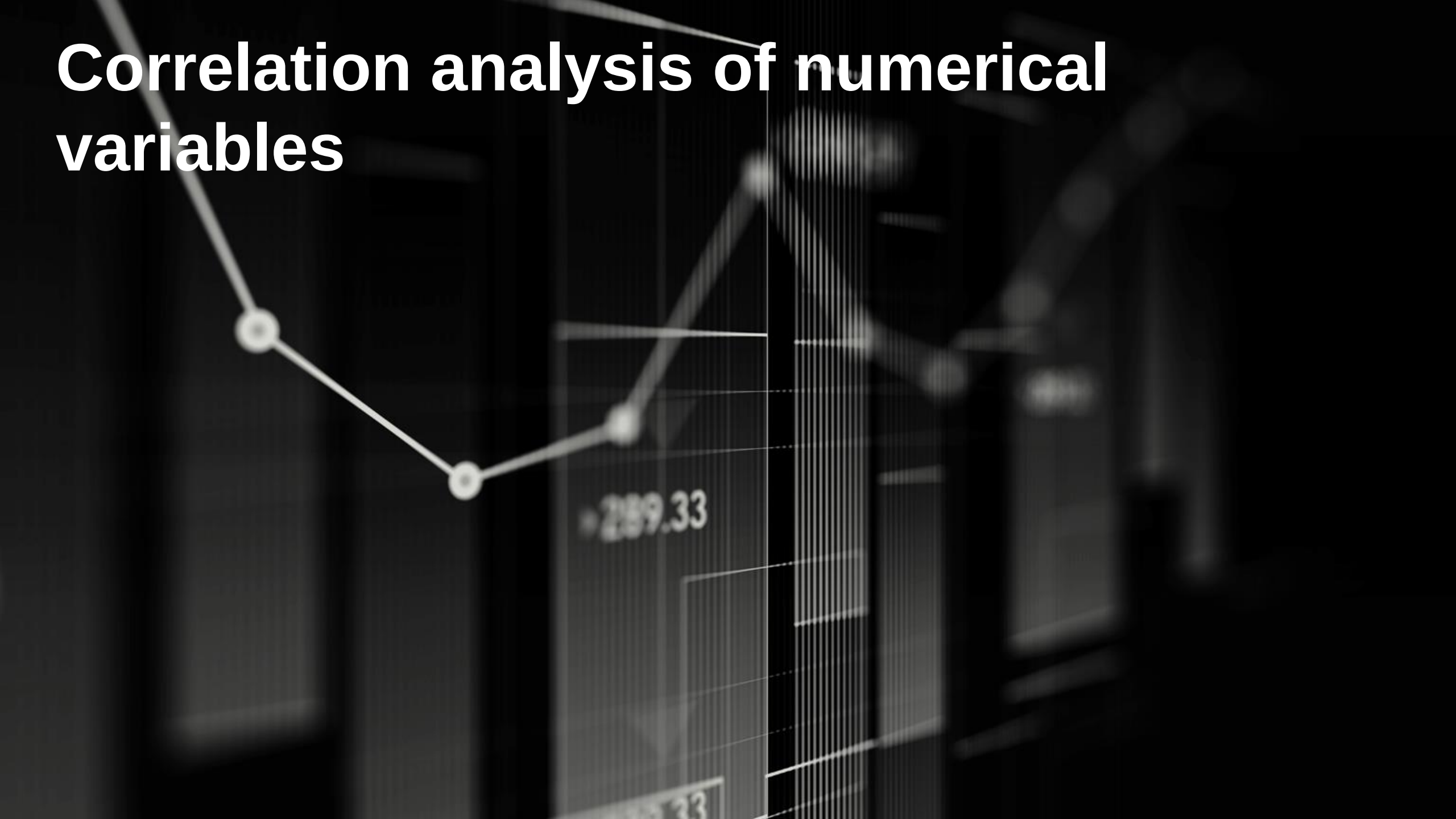




Analysis of "NAME_FAMILY_STATUS"

•- Clients who are 'Married' are 59.8% with payment difficulties and 64.2% with on-timepayments .
•- Clients who are 'Widow' are 3.8% with payment difficulties and 5.4% with ontimepayments .
•- Clients who are 'Single/not married' are 18.0% with payment difficulties and 14.5% with on-timepayments .
•- Clients who are 'Married' OR 'Widow' do on-time payments better comparatively. However, this is a weak correlation.
•- Clients who are 'Single/not married' have more difficulties with on-time payments comparatively. However, this is a weak correlation.

Correlation analysis of numerical variables



"AMT_GOODS_PRICE" with "AMT_CREDIT" 0.98

"REGION_RATING_CLIENT" with "REGION_RATING_CLIENT_W_CITY" 0.96

"CNT_FAM_MEMBERS" with "CNT_CHILDREN" 0.89

"DEF_60_CNT_SOCIAL_CIRCLE" with "DEF_30_CNT_SOCIAL_CIRCLE" 0.87

"REG_REGION_NOT_WORK_REGION" with "LIVE_REGION_NOT_WORK_REGION" 0.85

"LIVE_CITY_NOT_WORK_CITY" with "REG_CITY_NOT_WORK_CITY" 0.78

"AMT_ANNUITY" with "AMT_GOODS_PRICE" 0.75

"AMT_ANNUITY" with "AMT_CREDIT" 0.75

"DAYS_EMPLOYED" with "FLAG_DOCUMENT_6" 0.62

"DAYS_BIRTH" with "DAYS_EMPLOYED" 0.58

With Payment difficulties

"AMT_GOODS_PRICE" with "AMT_CREDIT" 0.99

"REGION_RATING_CLIENT" with "REGION_RATING_CLIENT_W_CITY" 0.95

"CNT_FAM_MEMBERS" with "CNT_CHILDREN" 0.88

"REG_REGION_NOT_WORK_REGION" with "LIVE_REGION_NOT_WORK_REGION" 0.86

"DEF_30_CNT_SOCIAL_CIRCLE" with "DEF_60_CNT_SOCIAL_CIRCLE" 0.86

"LIVE_CITY_NOT_WORK_CITY" with "REG_CITY_NOT_WORK_CITY" 0.83

"AMT_ANNUITY" with "AMT_GOODS_PRICE" 0.78

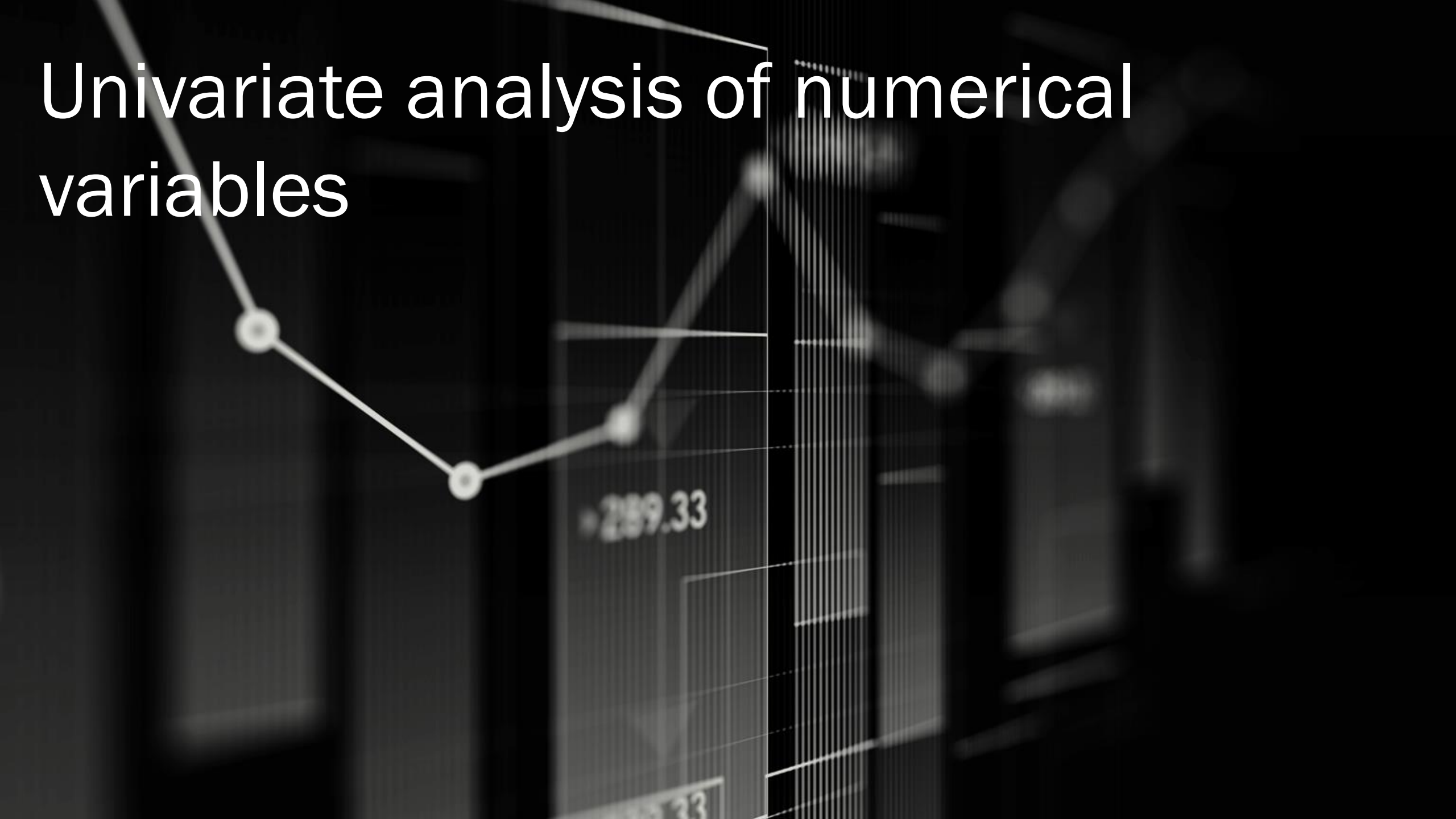
"AMT_ANNUITY" with "AMT_CREDIT" 0.77

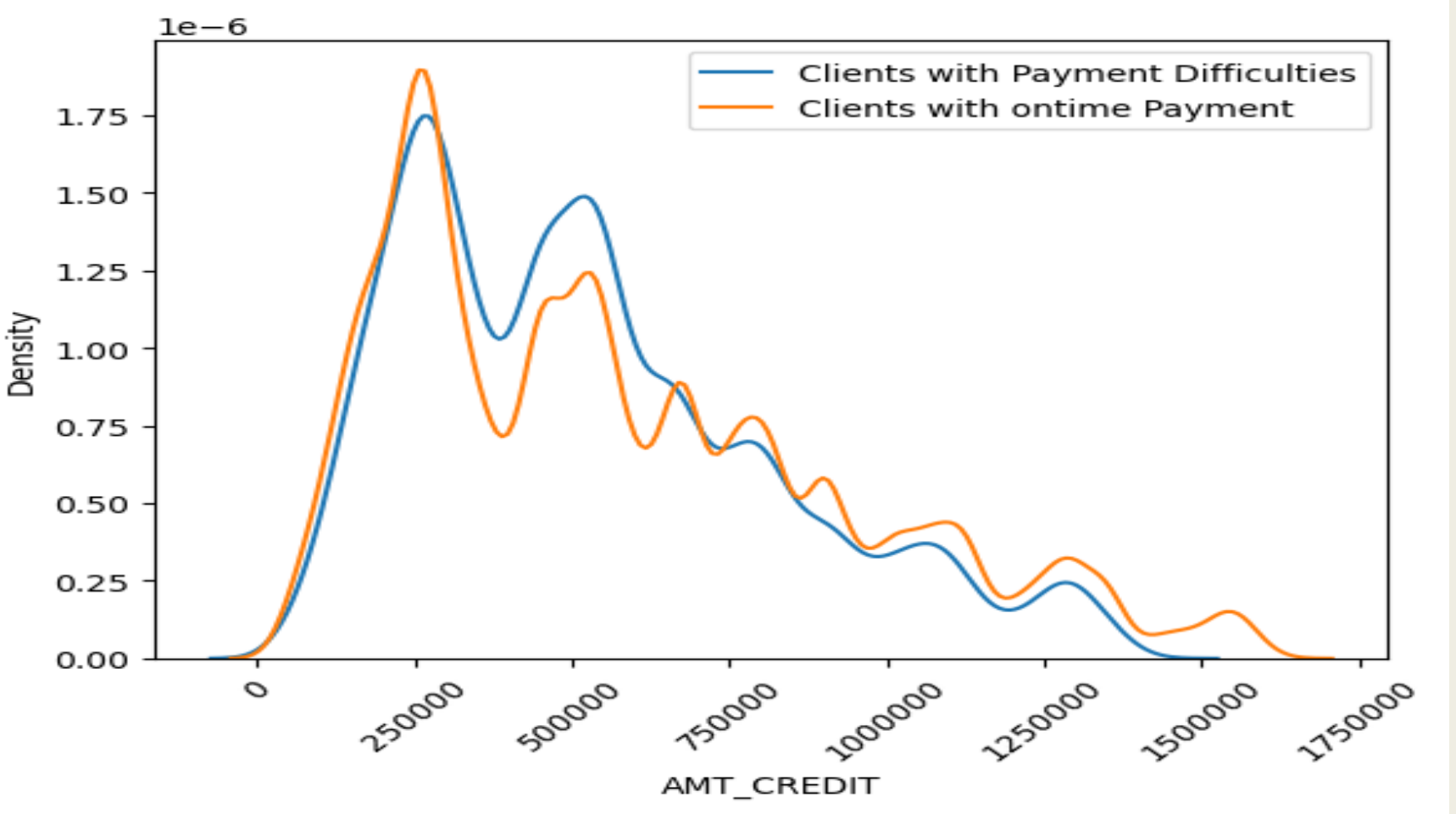
"DAYS_BIRTH" with "DAYS_EMPLOYED" 0.63

"DAYS_EMPLOYED" with "FLAG_DOCUMENT_6" 0.60

On-Time payments

Univariate analysis of numerical variables

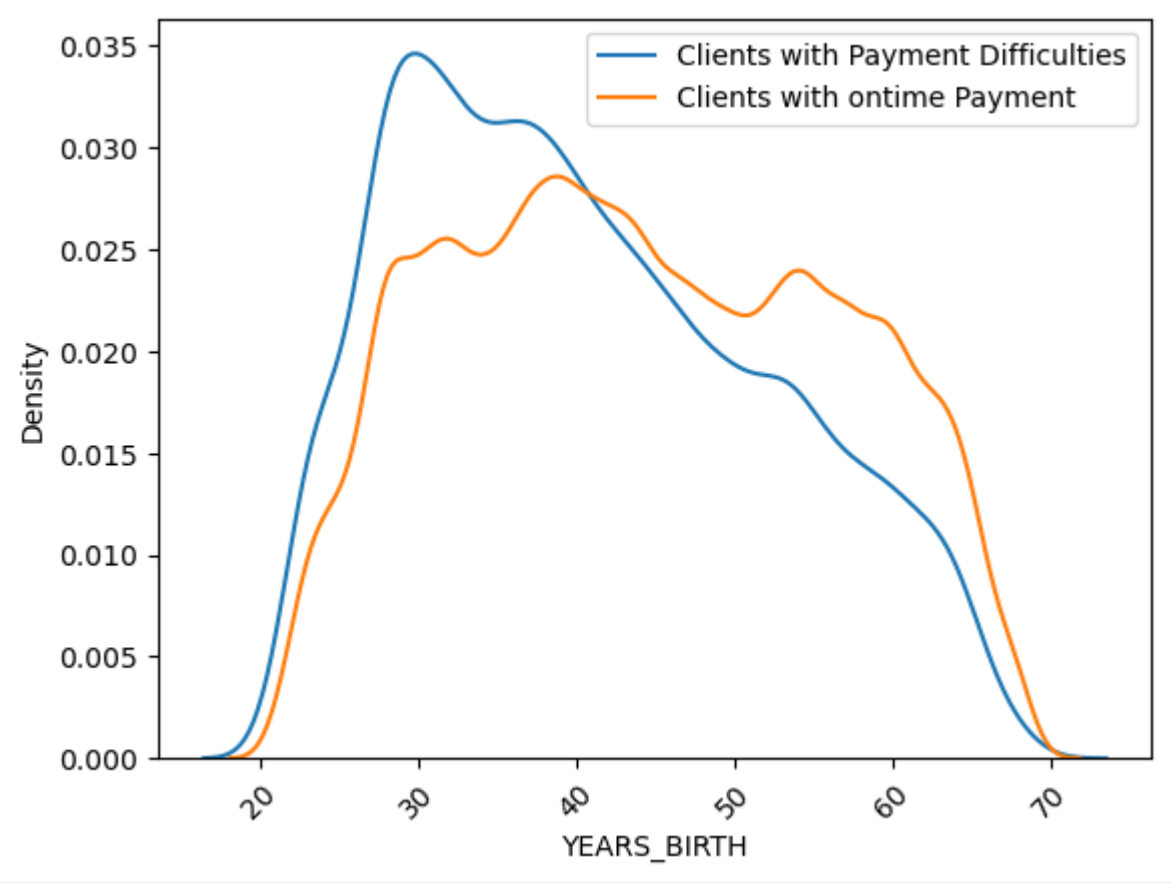




Analysis of "AMT_CREDIT"

- - For `AMT_CREDIT` between 250000 and approximately 650000, there are more clients with Payment difficulties

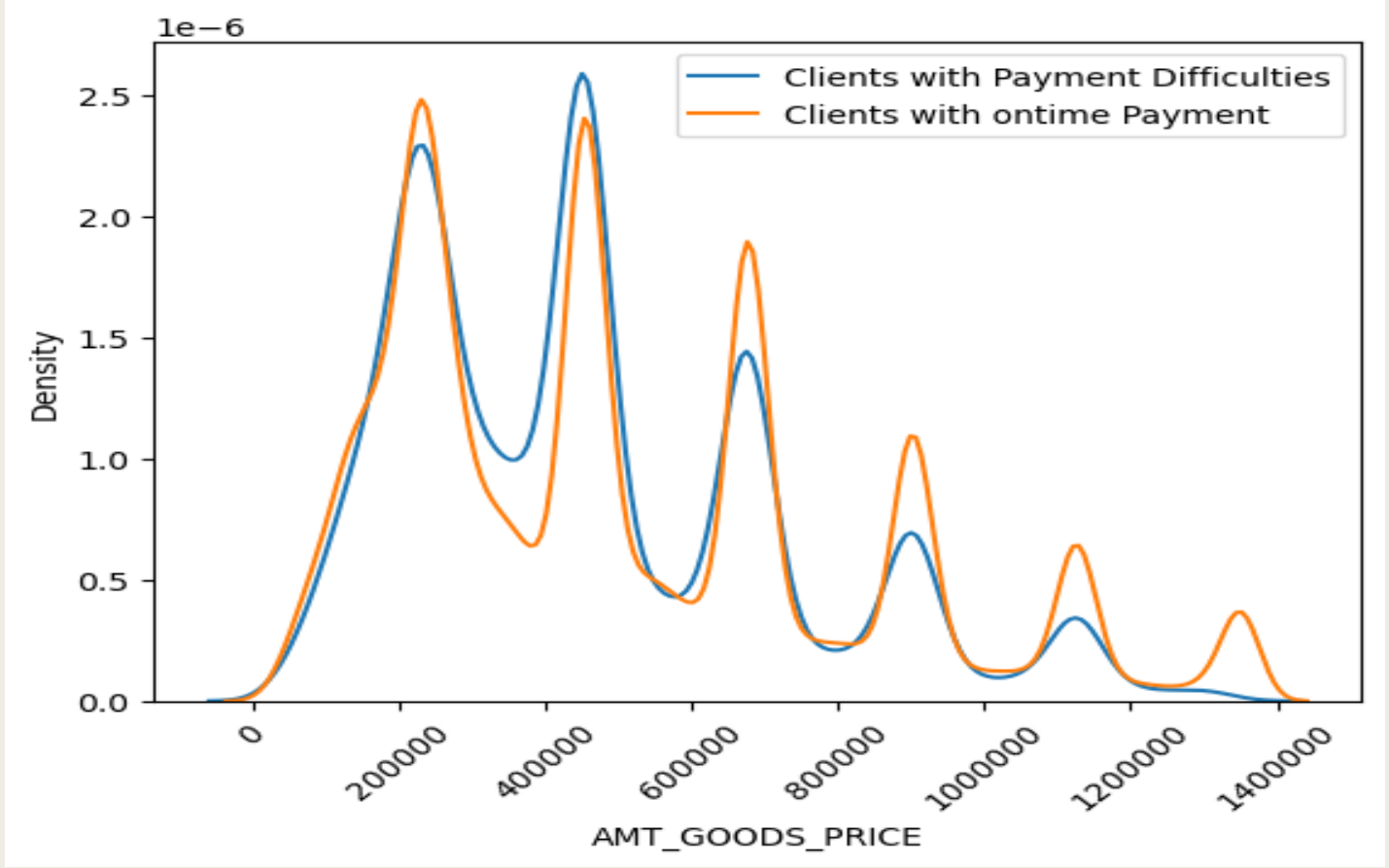
- - For `AMT_CREDIT` > 750000 , there are more clients with On -Time Payments



Analysis of "YEARS_BIRTH"

- - For "YEARS_BIRTH" between 20 and 40, there are more clients with Payment difficulties

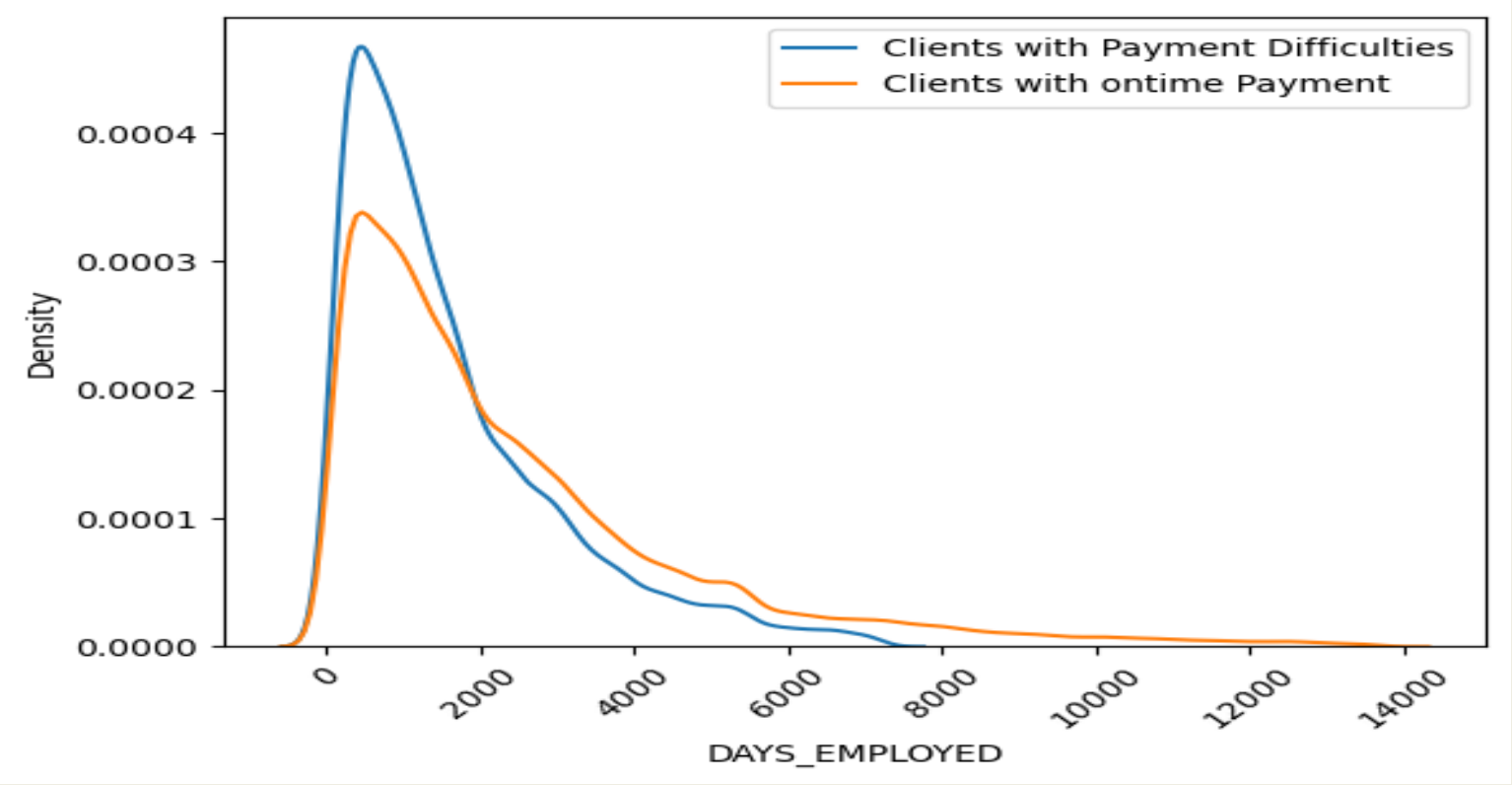
- - Conversely, for "YEARS_BIRTH"> 40 , there are more clients with On - Time Payments



Analysis of "AMT_GOODS_PRICE"

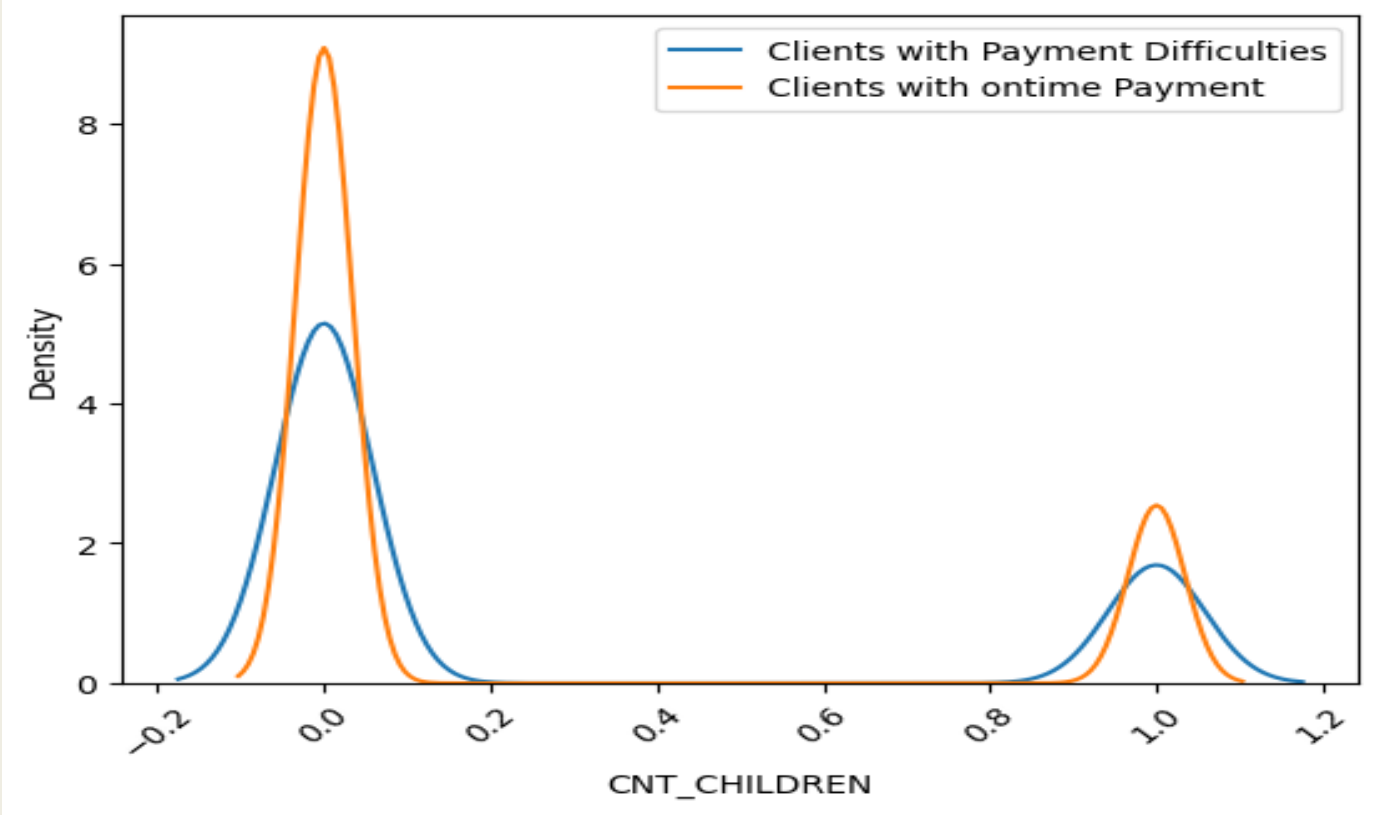
•- For `AMT_GOODS_PRICE` between ~250000 and ~550000, there are more clients with Payment difficulties .

•- Otherwise there are spikes on and off but they don't show any conclusive observations



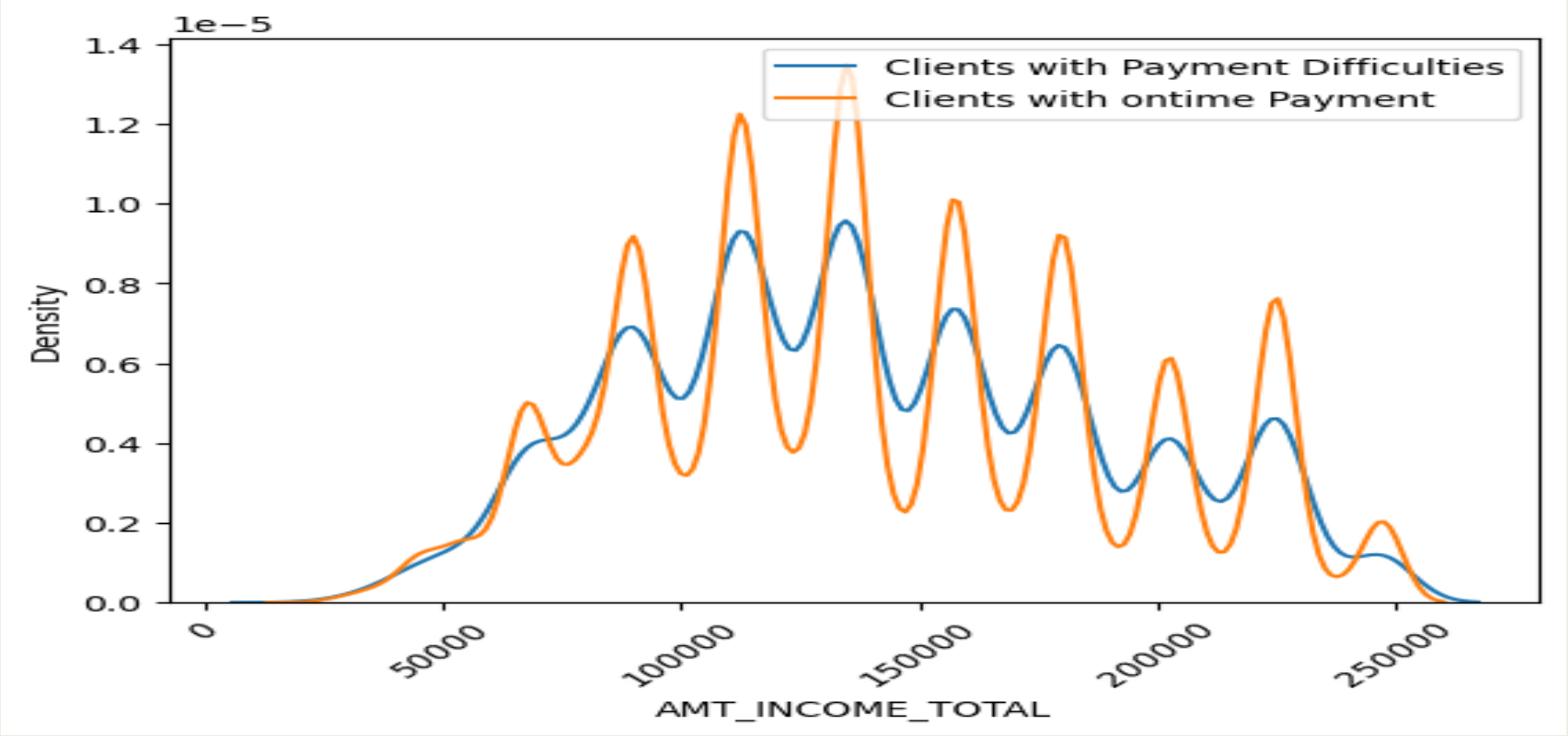
ANALYSIS OF "DAYS_EMPLOYED"

- For DAYS_EMPLOYED less than 2000, there are more clients with Payment difficulties.
- Conversely, for DAYS_EMPLOYED > 2000 , there are more clients with On-Time Payments.
- This means that those who are employed longer have better chances of repaying the loan.



ANALYSIS OF "CNT_CHILDREN"

- For CNT_CHILDREN 0 (those with no children), there are lots of clients with On-Time Payments.
- For CNT_CHILDREN with 1 OR 2 (those with 1 or 2 children), there are few more clients with On-Time Payments

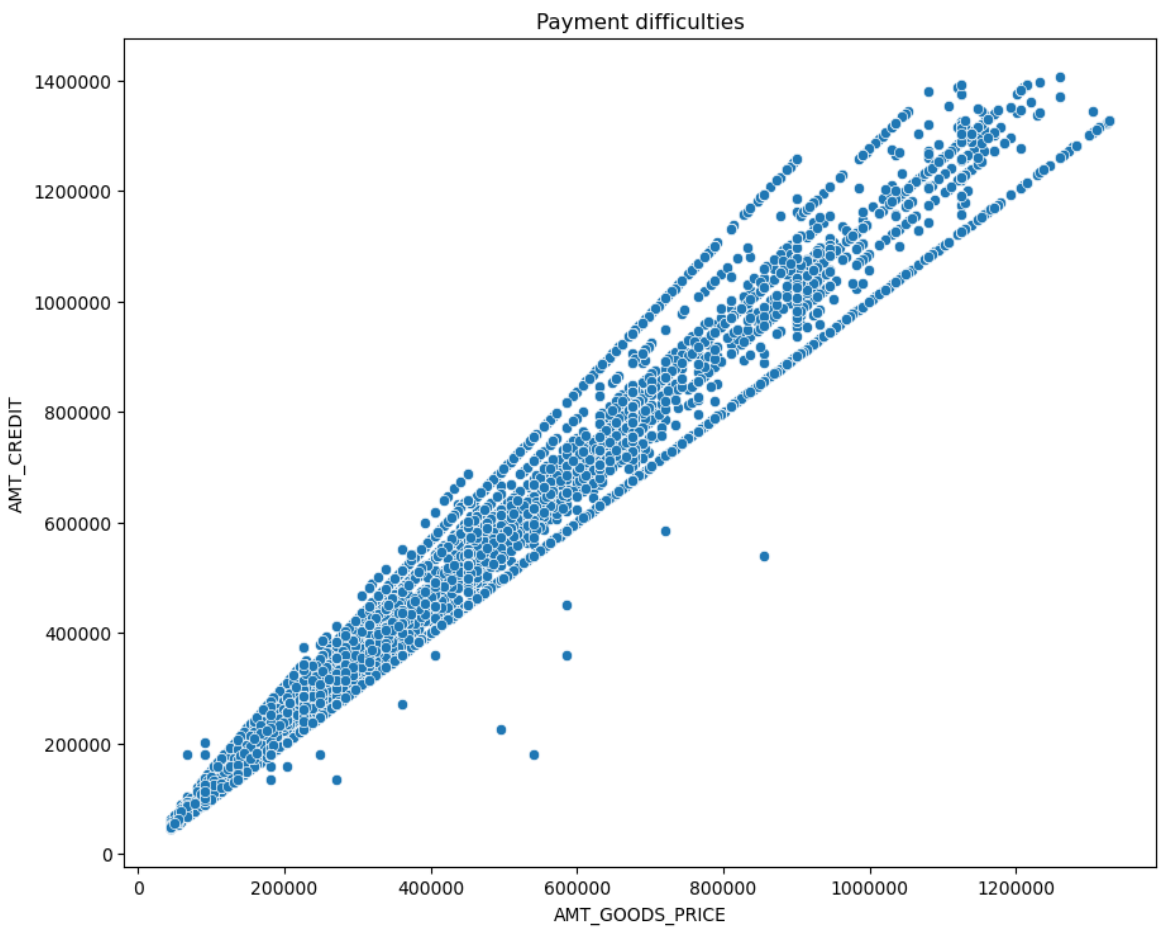


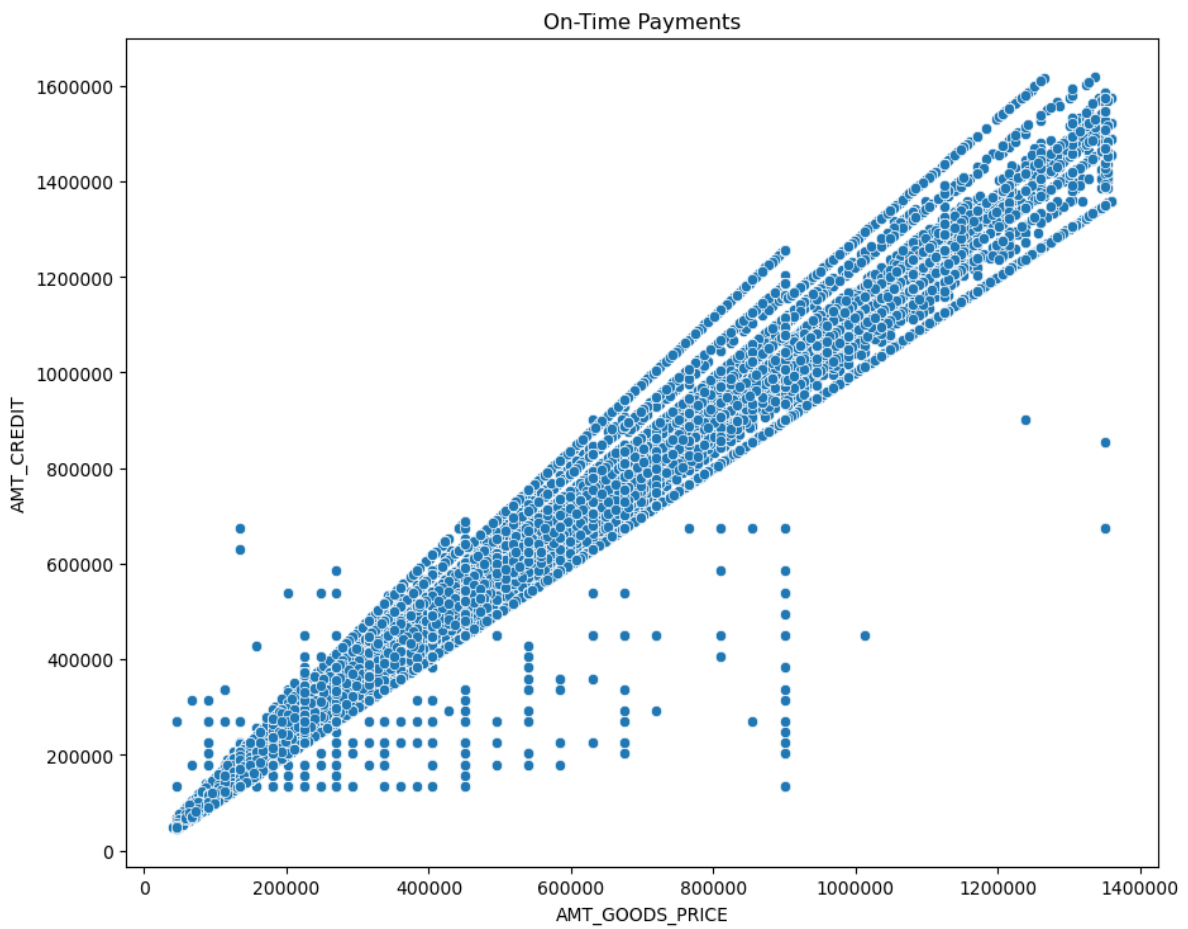
ANALYSIS OF "AMT_INCOME_TOTAL"

- Based on AMT_INCOME_TOTAL, for clients with Payment difficulties, the distribution resembles a normal distribution approximately
- But for clients with On-Time Payments, there are erratic spikes in the distribution which doesn't give any valid observations

BIVARIATE/MULTIVARIATE ANALYSIS

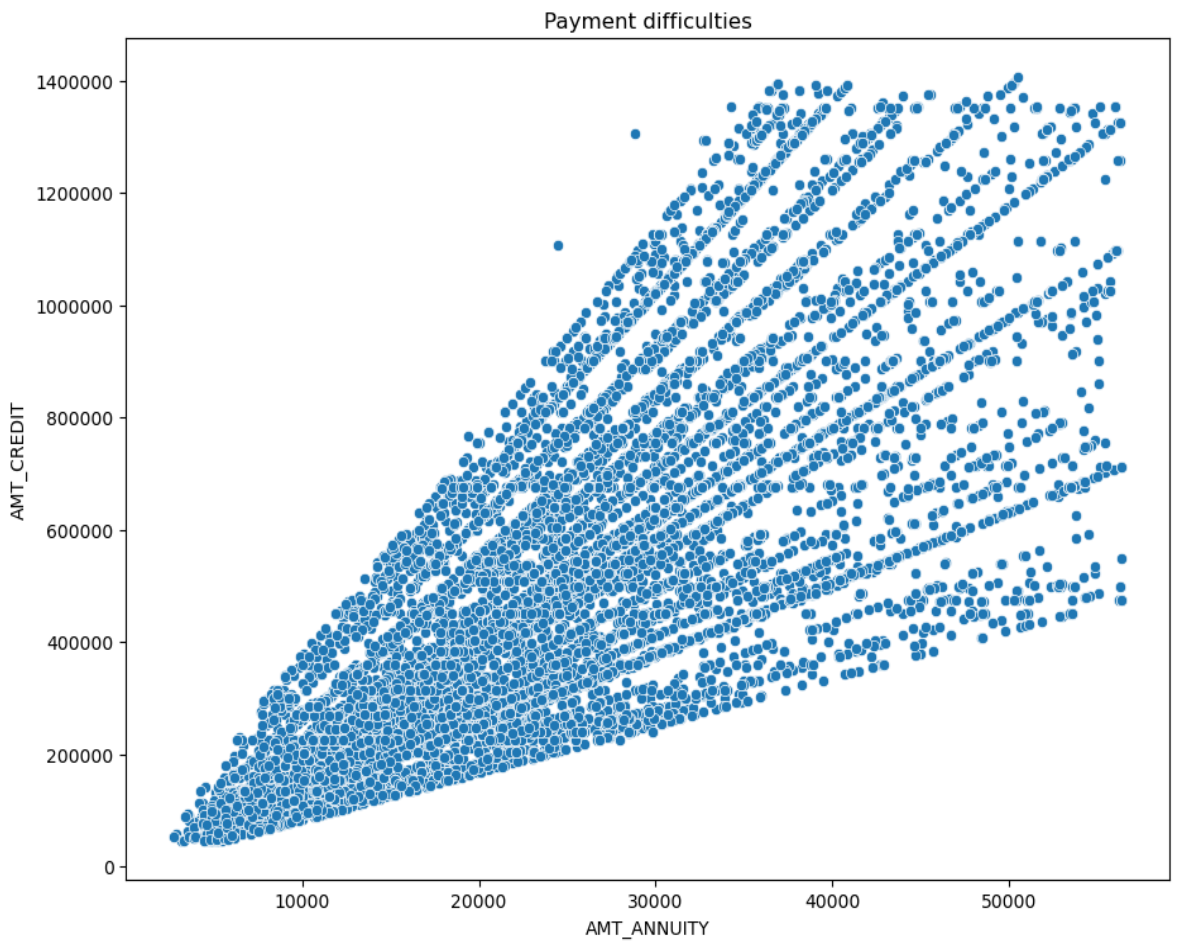
----------Continuous V/S Continuous variables

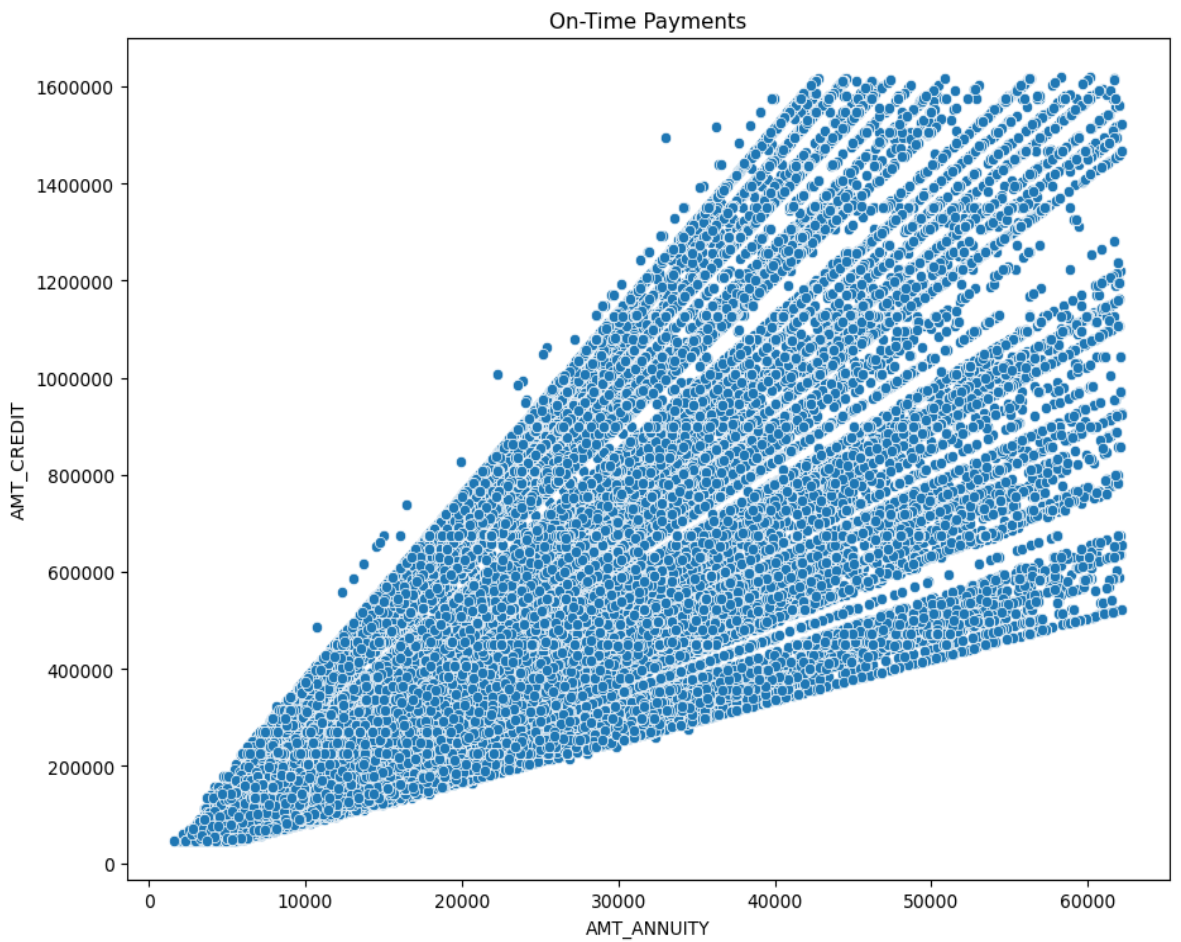




Analysis of "AMT_GOODS_PRICE" V/S "AMT_CREDIT"

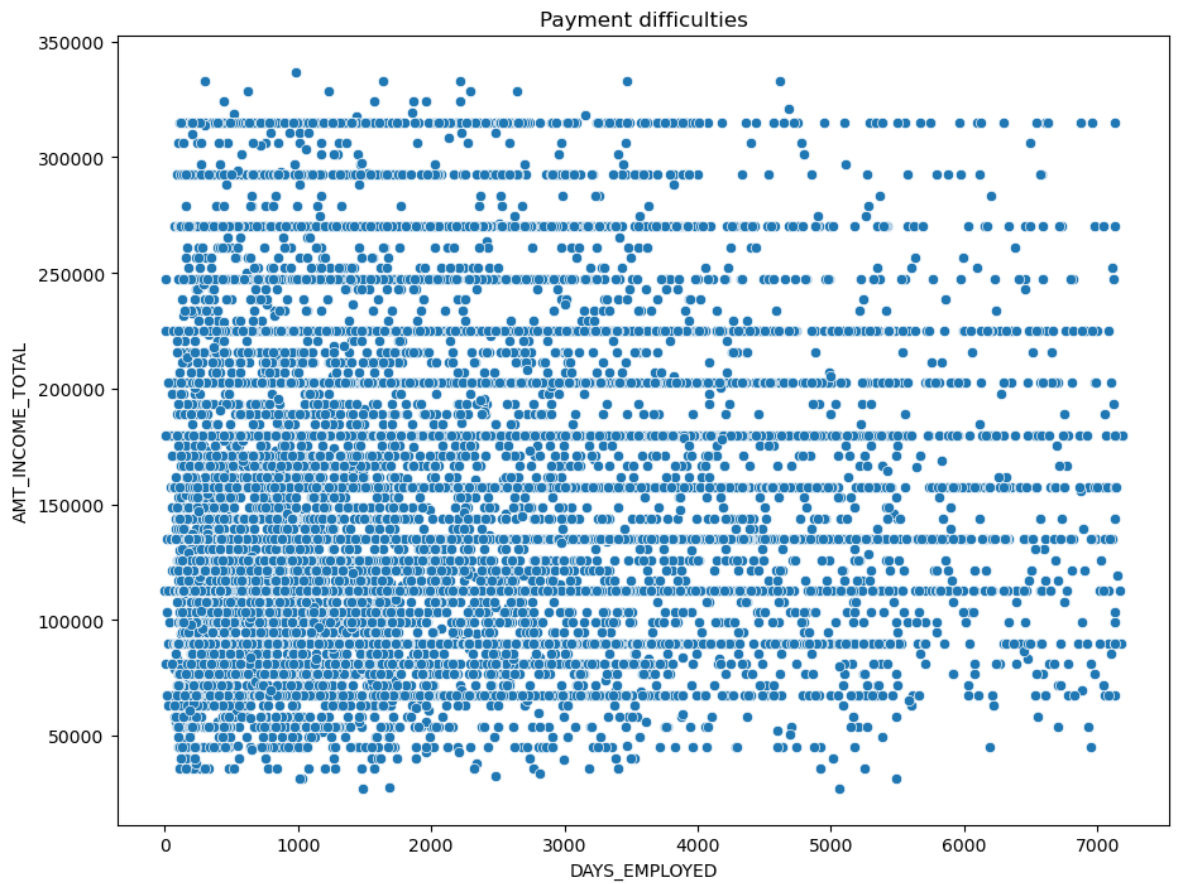
•- "AMT_GOODS_PRICE" and "AMT_CREDIT" have strong positive correlation. This means that as Goods price increases, so does Credit Amount .





Analysis of "AMT_ANNUITY" V/S "AMT_CREDIT"

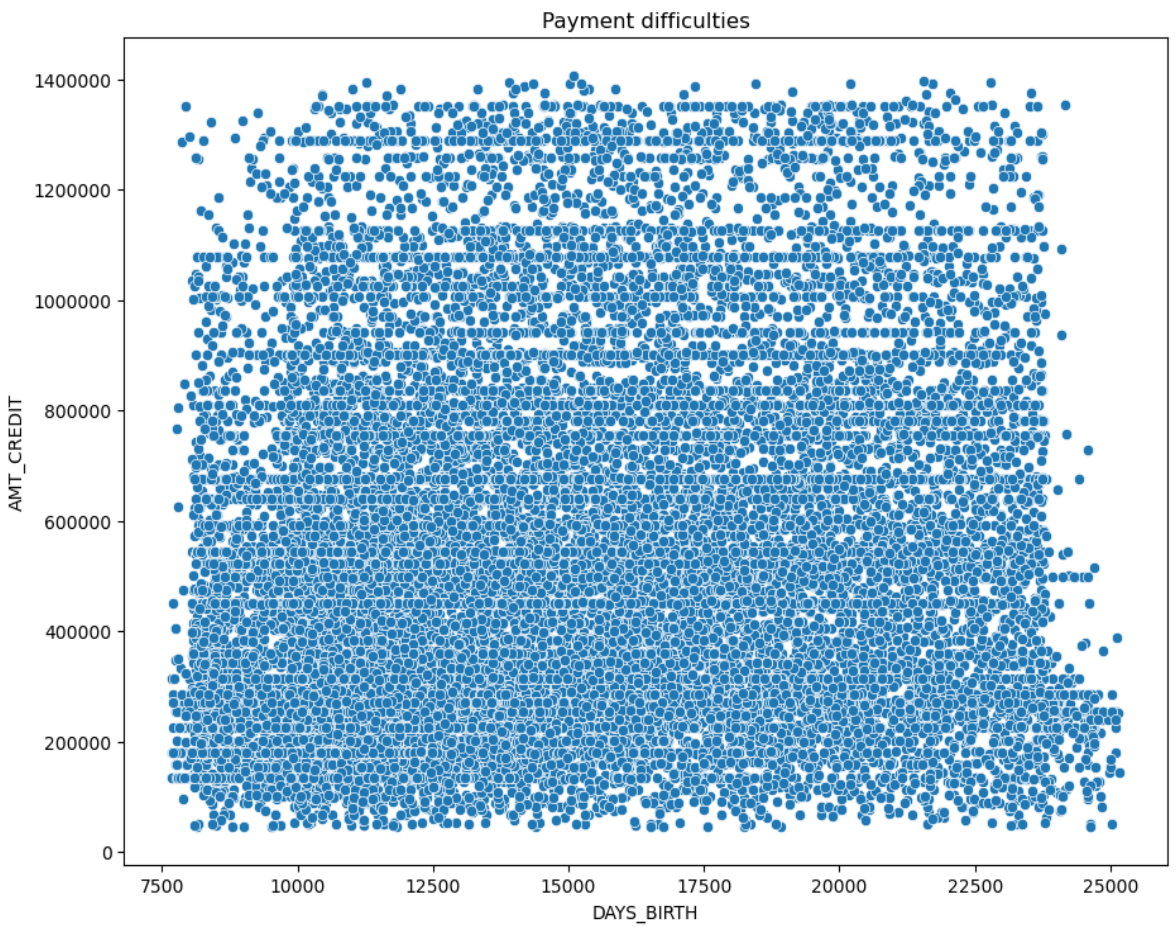
- AMT_ANNUITY and AMT_CREDIT have strong positive correlation. This means that as Annuity Amount increases, so does Credit Amount

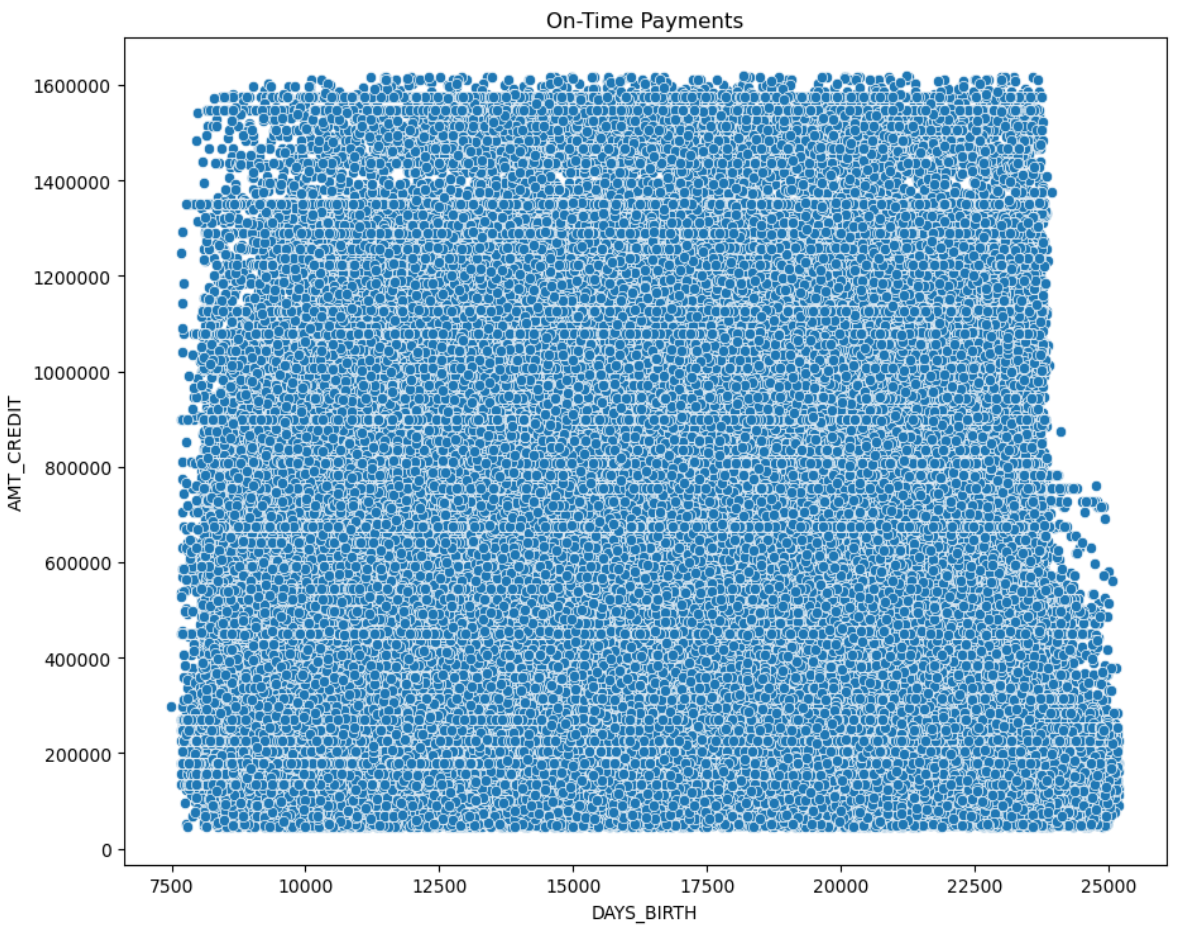


Analysis of "DAYS_EMPLOYED" V/S "AMT_INCOME_TOTAL"

•- Clients who are employed for a long time (>7000) days or 19 years are making their payments on-time but these category of clients do not exist in Payments difficulties group.

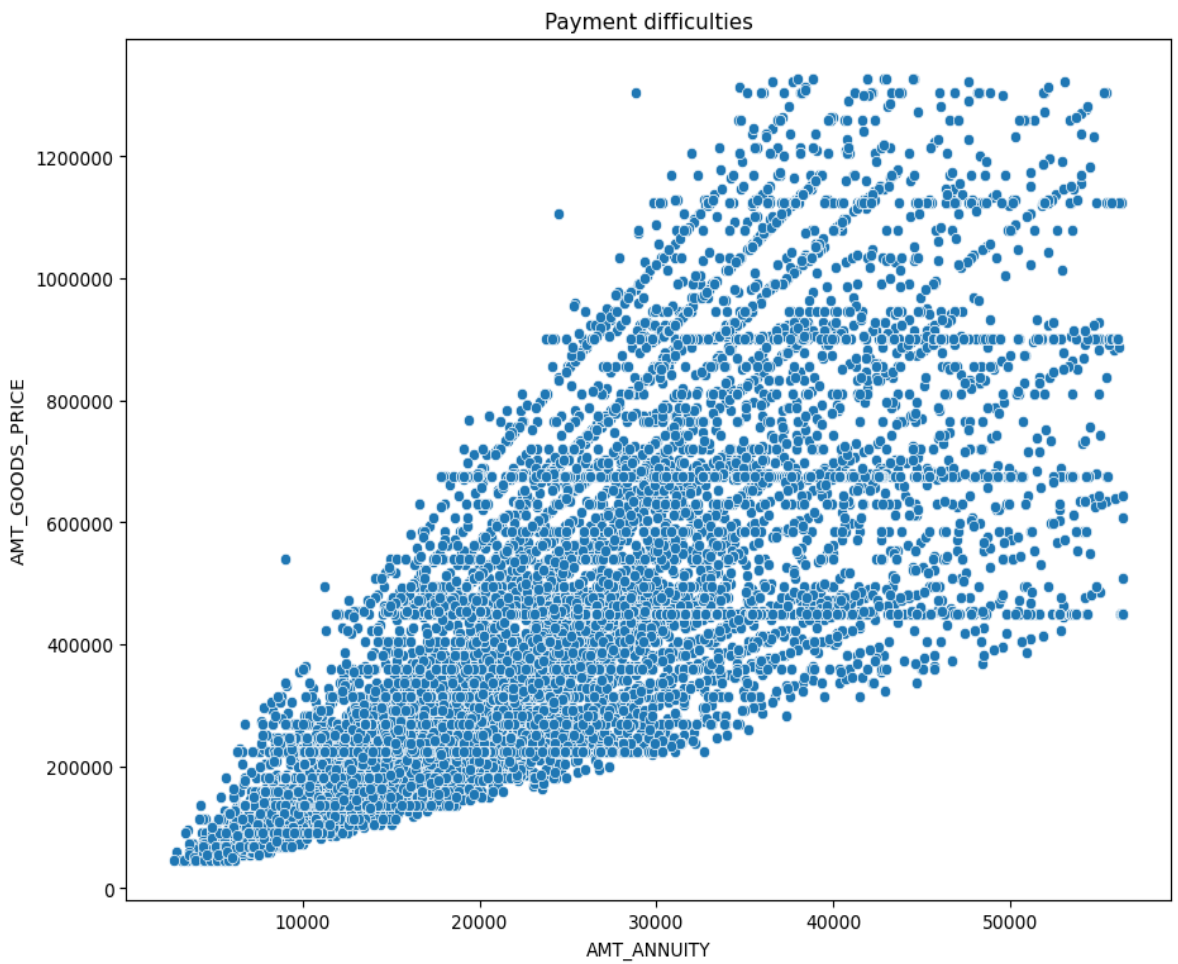
•- Even looking at Payment difficulties group, clients with more than 4000 days of employment are sparse

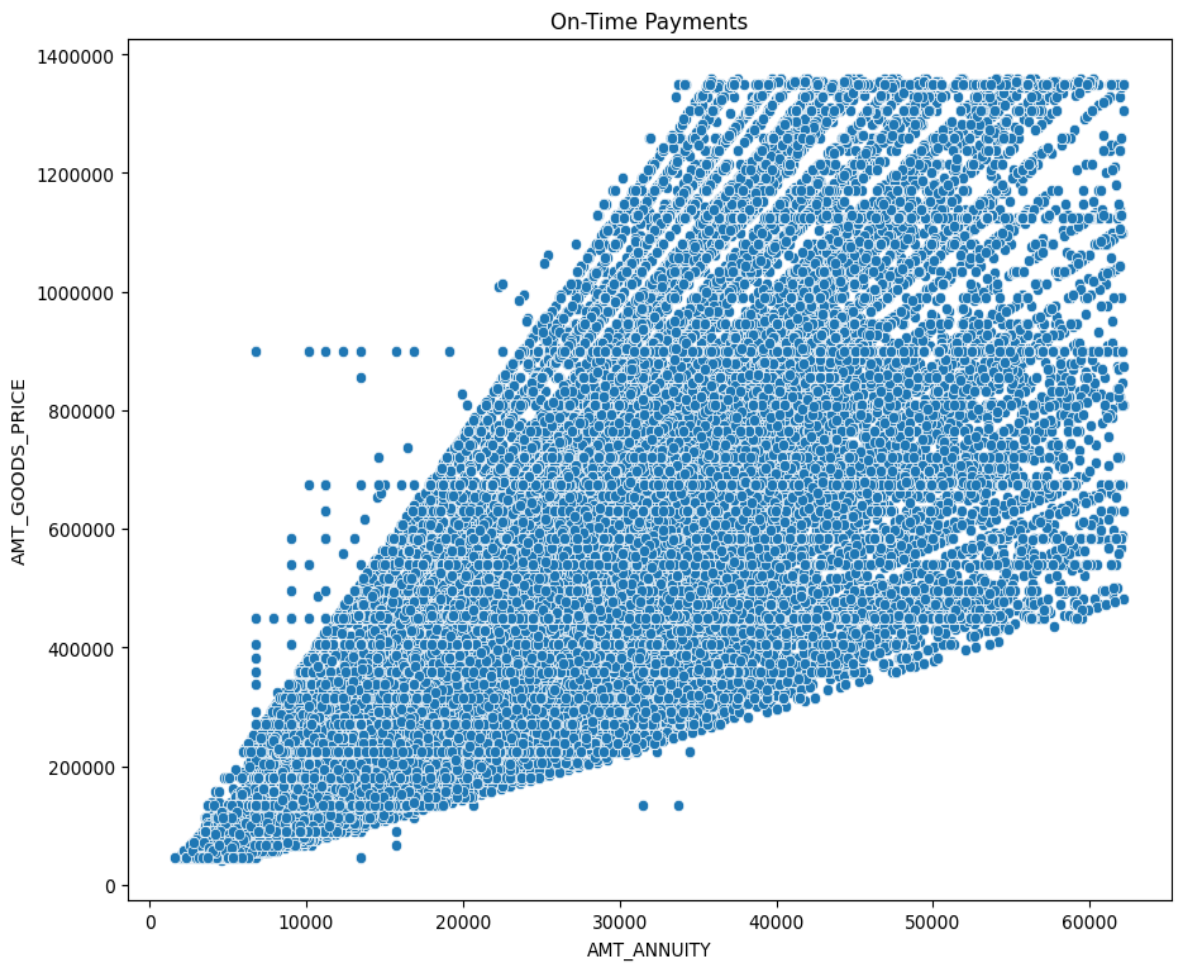




ANALYSIS OF "AMT_CREDIT" AND "DAYS_BIRTH"

- There is no visible correlation between the two.



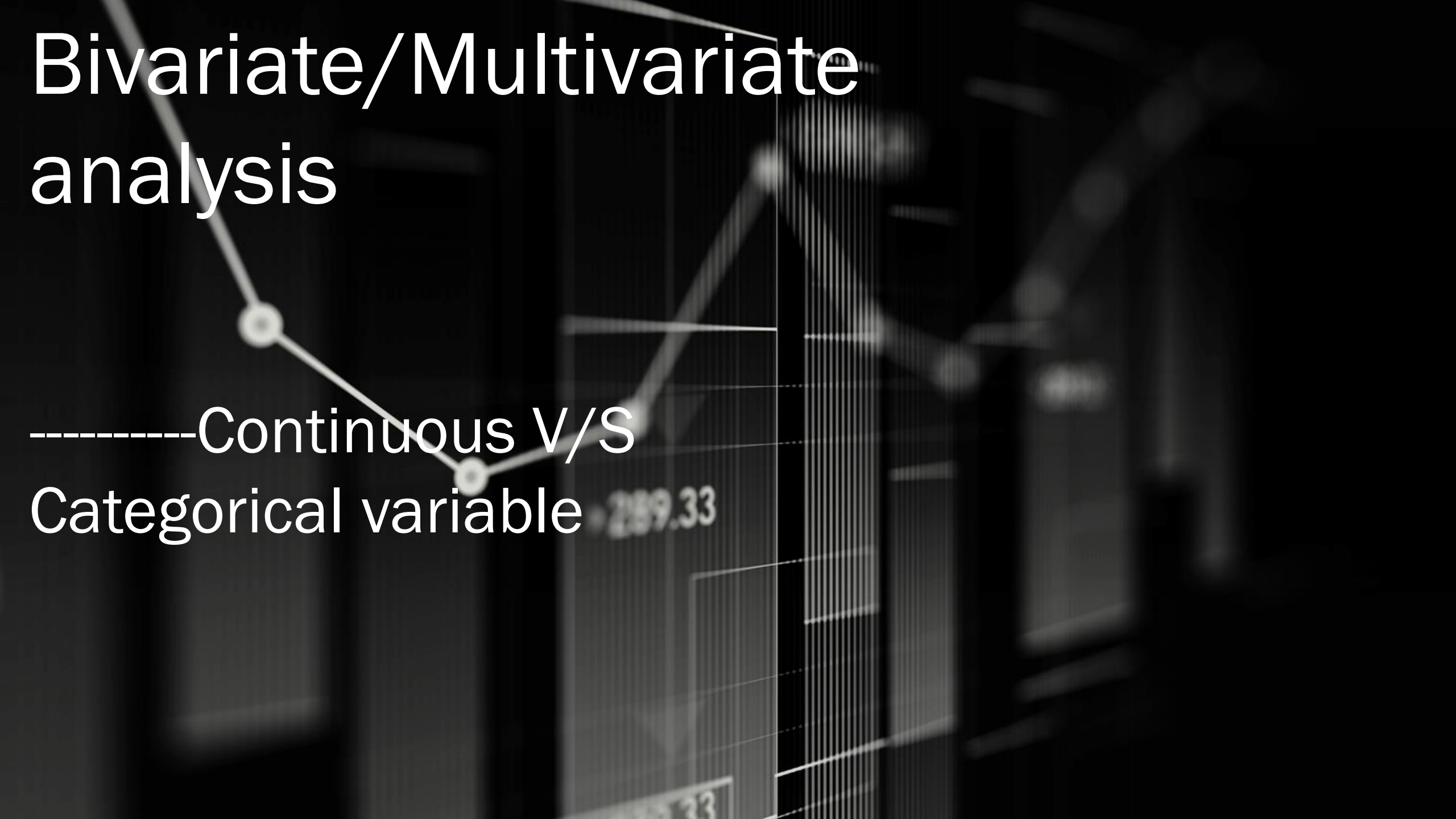


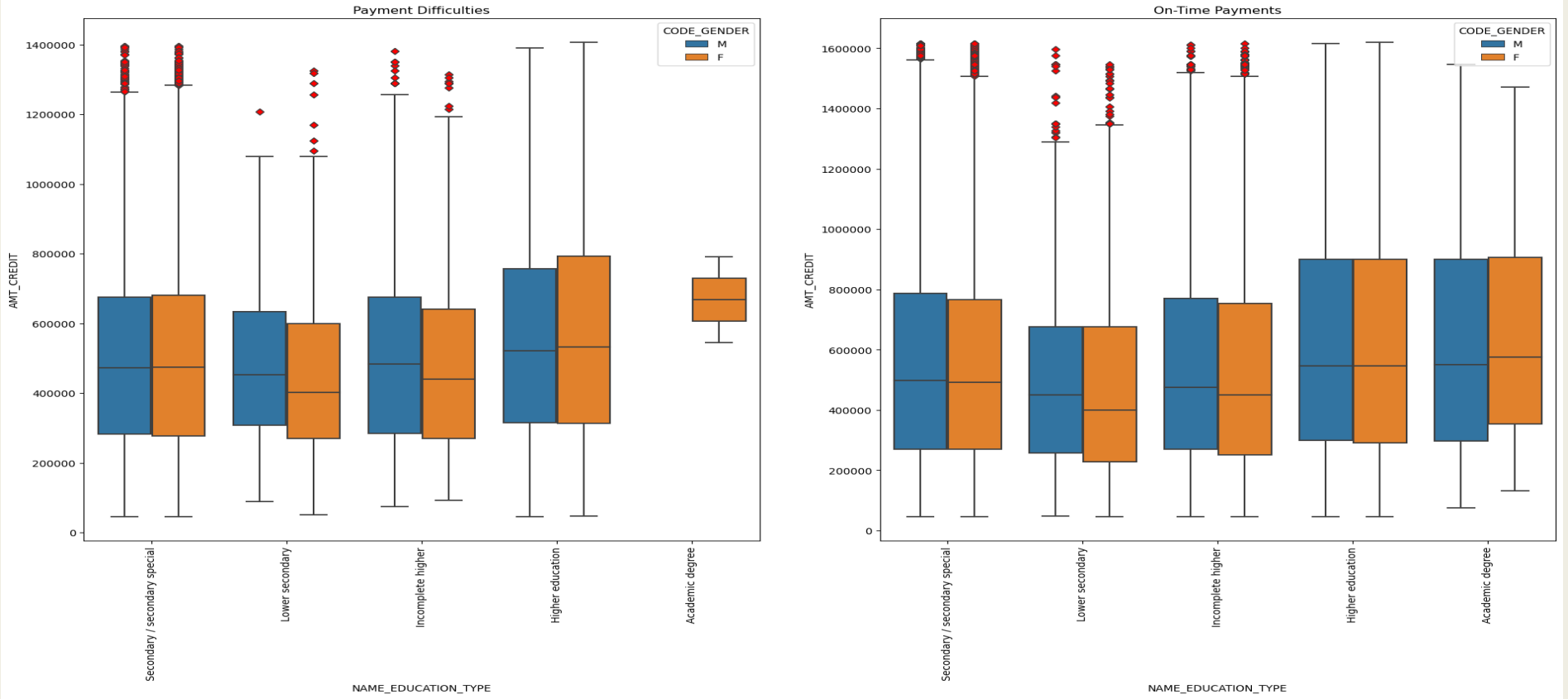
ANALYSIS OF "AMT_GOODS_PRICE" AND "AMT_ANNUITY"

- AMT_ANNUITY and AMT_GOODS_PRICE shares a strong correlation.As annuity increases so does the price.

Bivariate/Multivariate analysis

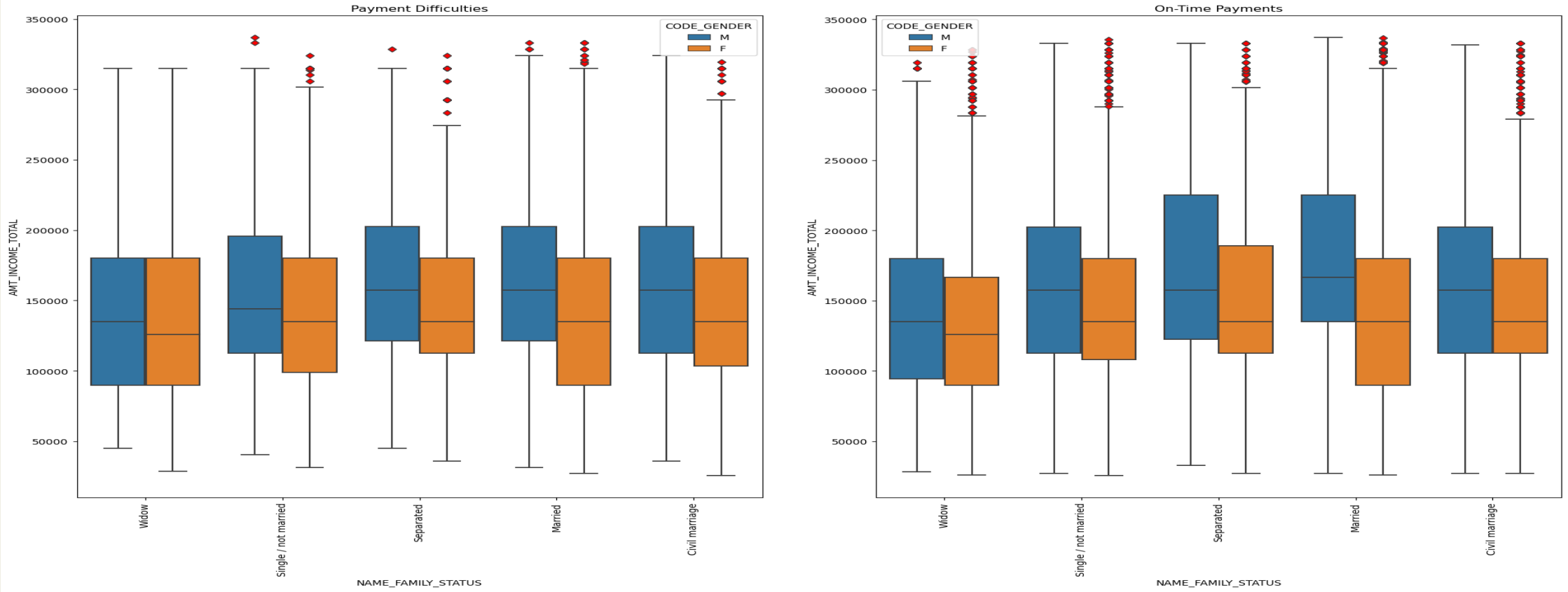
---------Continuous V/S Categorical variable





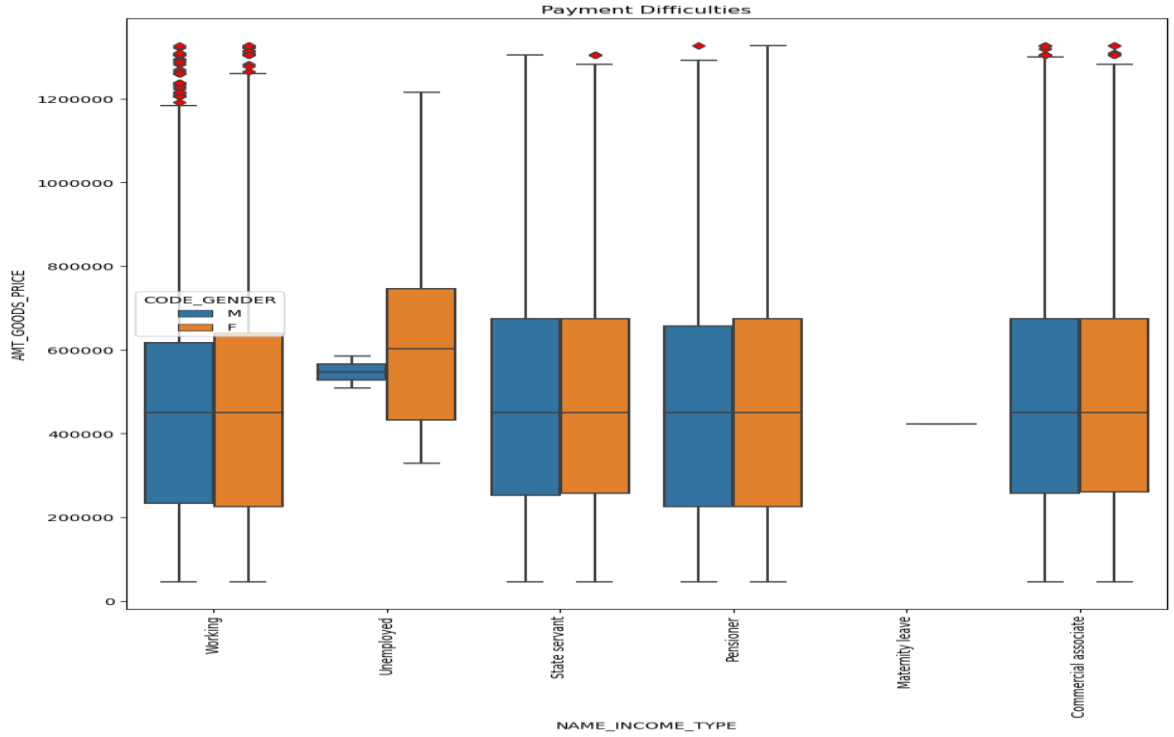
Analysis of "NAME_EDUCATION_TYPE" V/S "AMT_CREDIT" V/S "CODE_GENDER"

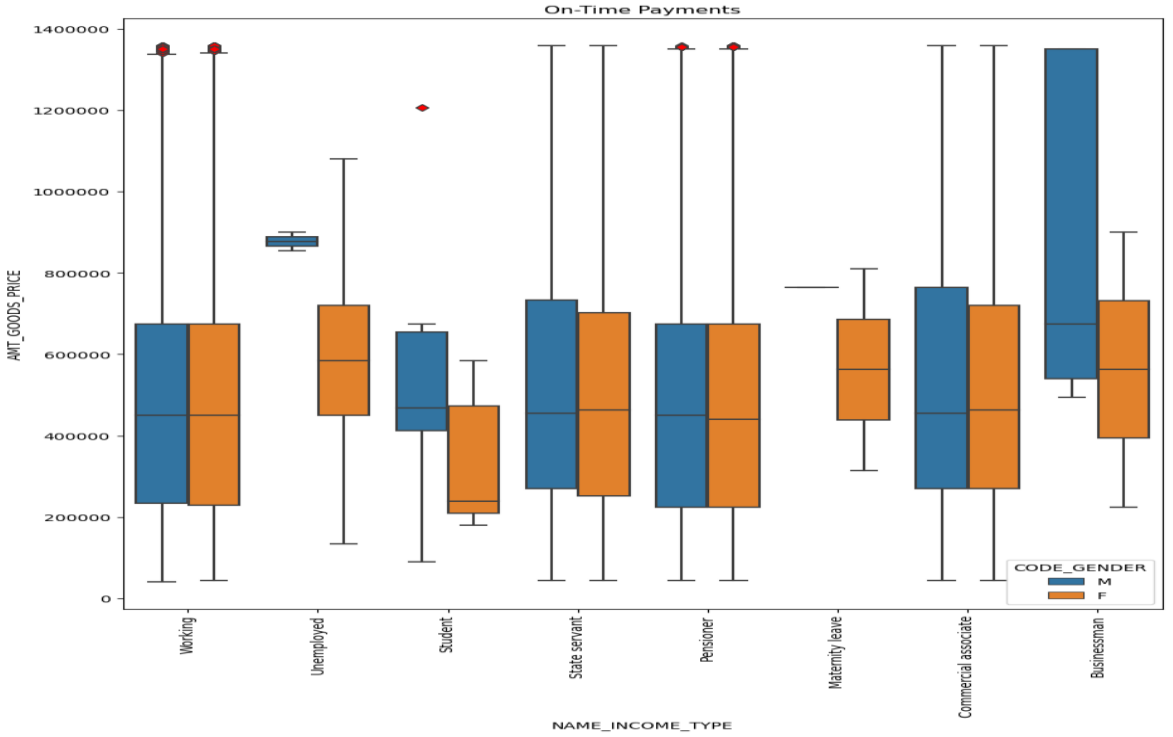
- Clients with `Academic Degree` have a wide range of credits for OnTime Payments whereas the range is much lower for ones with Payment difficulties
- Looking at summary statistics, Clients with `Academic Degree` and Payment difficulties take mean and median credit at a much higher range than On-Time Payment clients
- `Male` clients with `Academic Degree` always pay the loan on-time



Analysis of "NAME_FAMILY_STATUS" V/S "AMT_INCOME_TOTAL" V/S "CODE_GENDER"

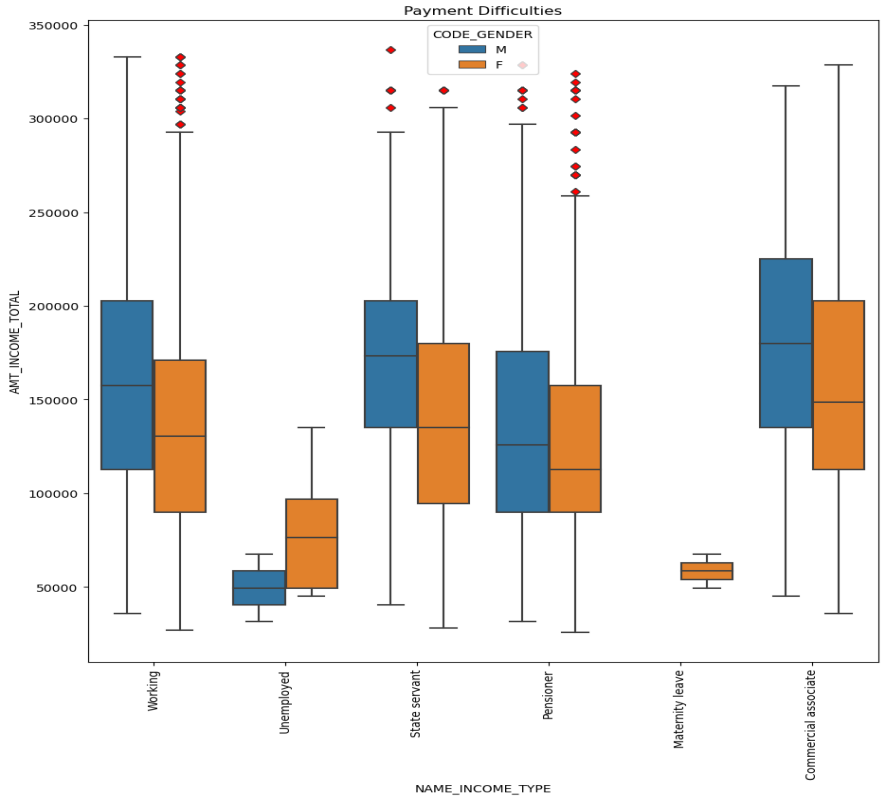
- "Married" clients have a slightly higher mean/median income with OnTime Payments than Payment difficulties category

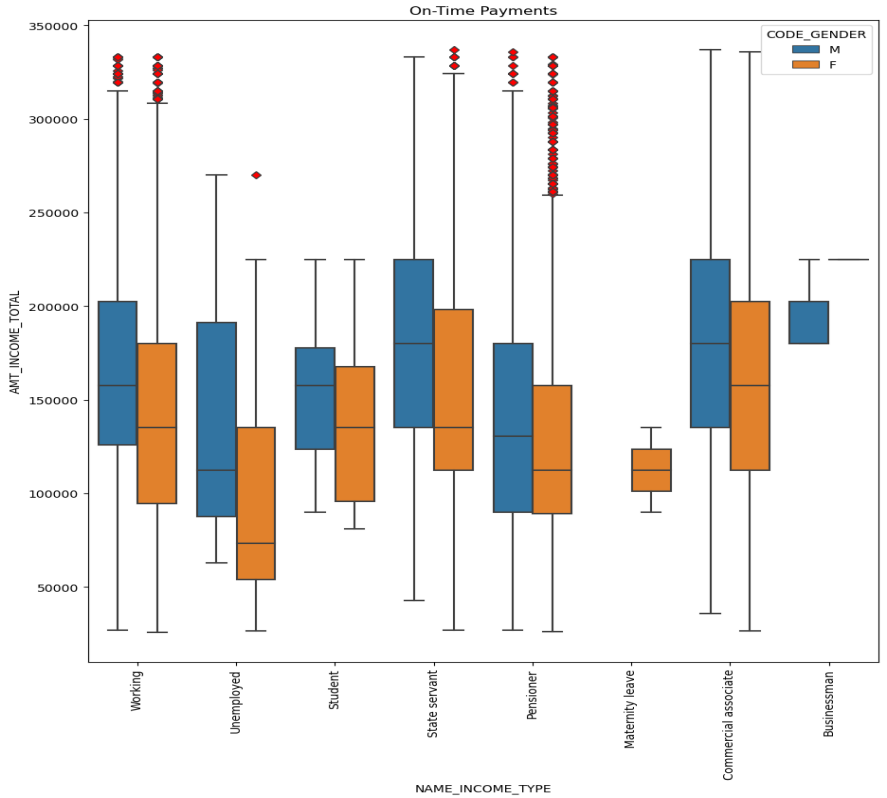




ANALYSIS OF "NAME_INCOME_TYPE" &"AMT_GOODS_PRICE" & "CODE GENDER"

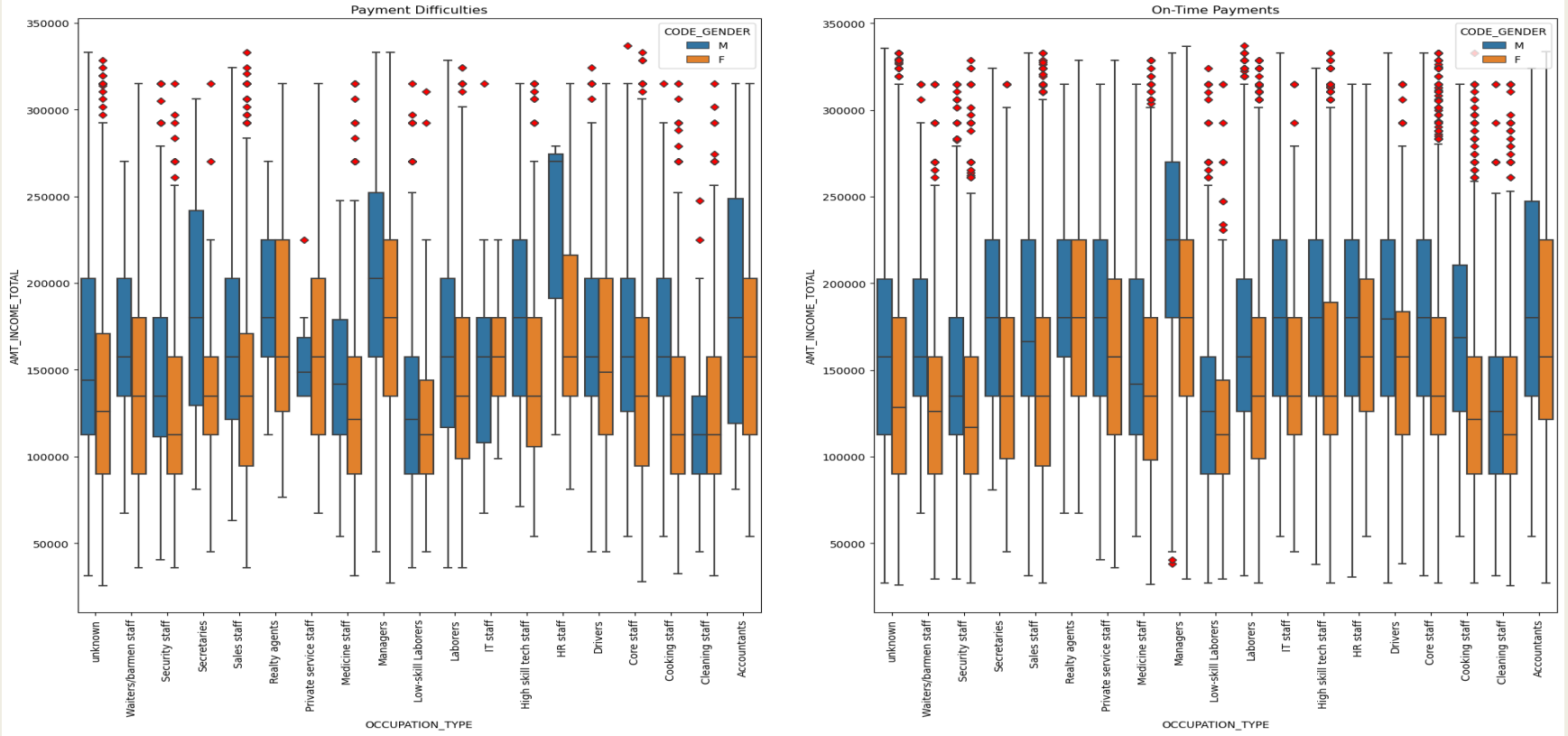
- - Clients who are Unemployed and Male have a very high price of goods in On-Time Payments than Payment difficulties
- - Clients who are Student and either Male OR Female do their payments On-Time. They are completely missing from Payment difficulties category. Student seems to be an attractive category to give loans to.
- - Clients who are Businessman and either Male OR Female do their payments On-Time. They are completely missing from Payment difficulties category. Businessman seems to be an attractive category to give loans to.





Analysis of "NAME_INCOME_TYPE" V/S "AMT_INCOME_TOTAL" V/S "CODE_GENDER"

- Clients who are Unemployed and Male have a very high income in On-Time Payments than Payment difficulties
- Clients who are Student and either Male OR Female do their payments On-Time. They are completely missing from Payment difficulties category. Student seems to be an attractive category to give loans to.
- Clients who are Businessman and either Male OR Female do their payments On-Time. They are completely missing from Payment difficulties category. Businessman seems to be an attractive category to give loans to.

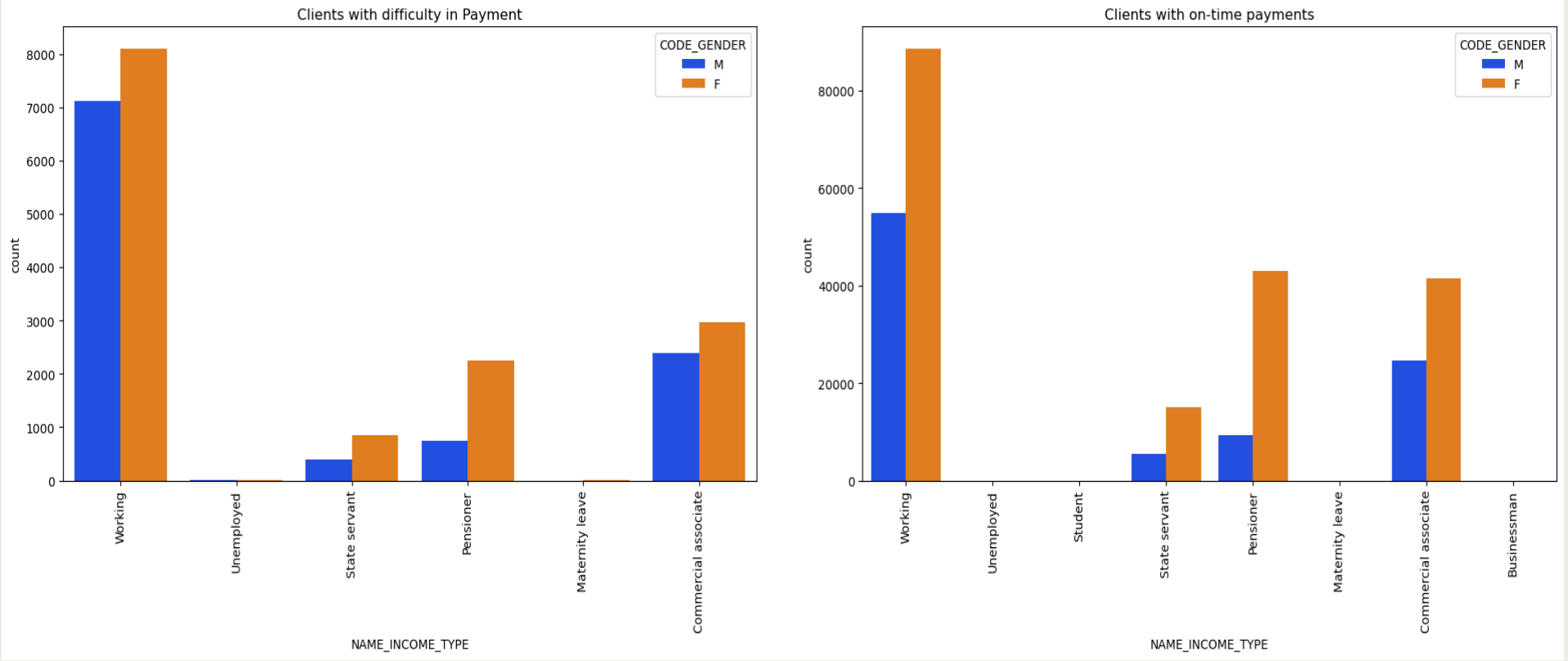


ANALYSIS OF "AMT_INCOME_TOTAL"&"OCCUPATION_TYPE"&"CODE_GENDER"

- - Clients who are Waiters/barment staff and female have less median income in On-Time Payments than Payment difficulties
- - Clients who are Cleaning staff and female have more median income in On-Time Payments than Payment difficulties
- - Clients who are HR Staff and Male have more median income in Payment difficulties than On-Time Payments
- - Clients who are Managers and Male have more median income in On-Time Payments than Payment difficulties

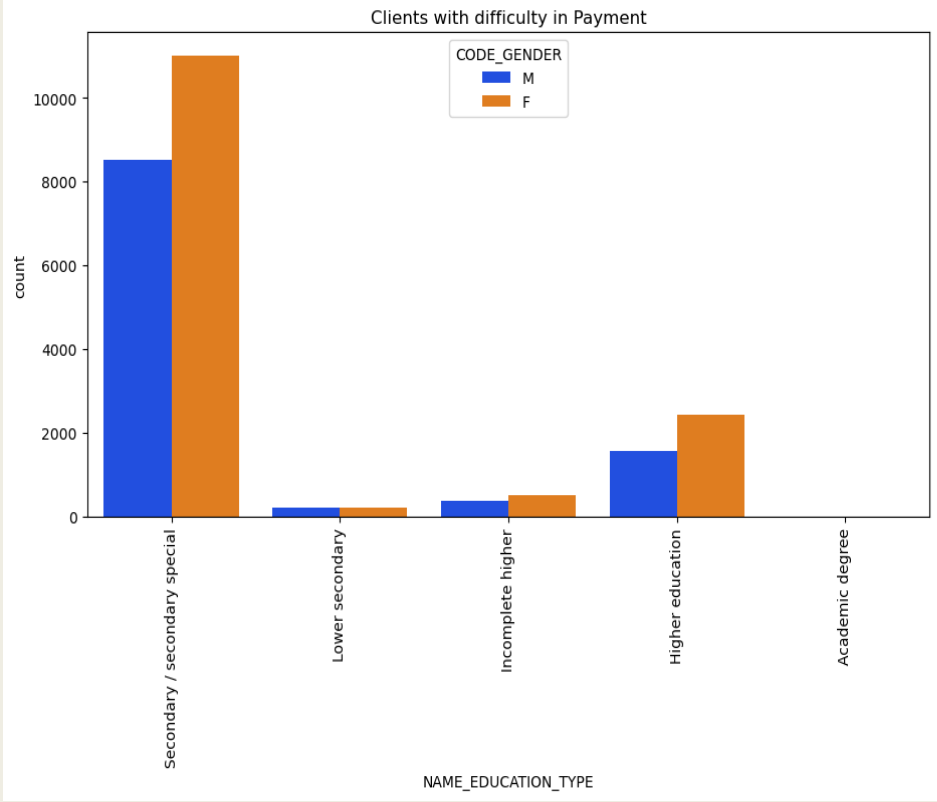
Bivariate/Multivariate analysis

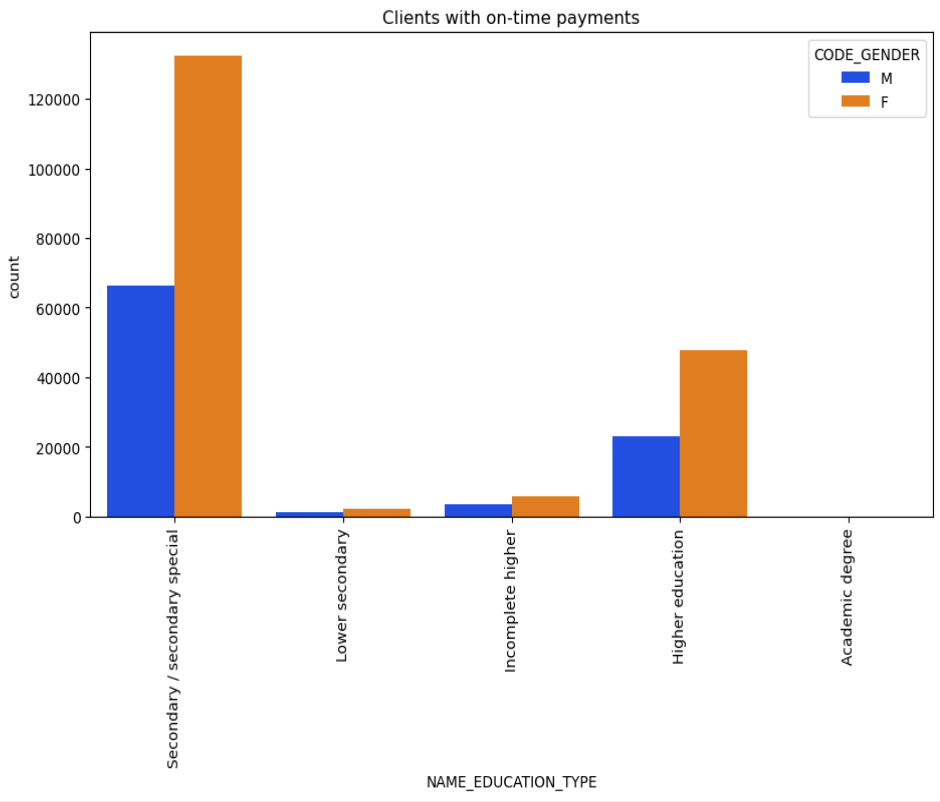
-----Categorical V/S Categorical variables



ANALYSIS OF "NAME_INCOME_TYPE" & "CODE_GENDER"

- Clients who are Working and Male have more Payment difficulties compared to On-Time Payments
- Clients who are Pensioner and Female have more Payment difficulties compared to On-Time Payments
- Clients who are Businessman and Students do their payments On-Time though their record count is low

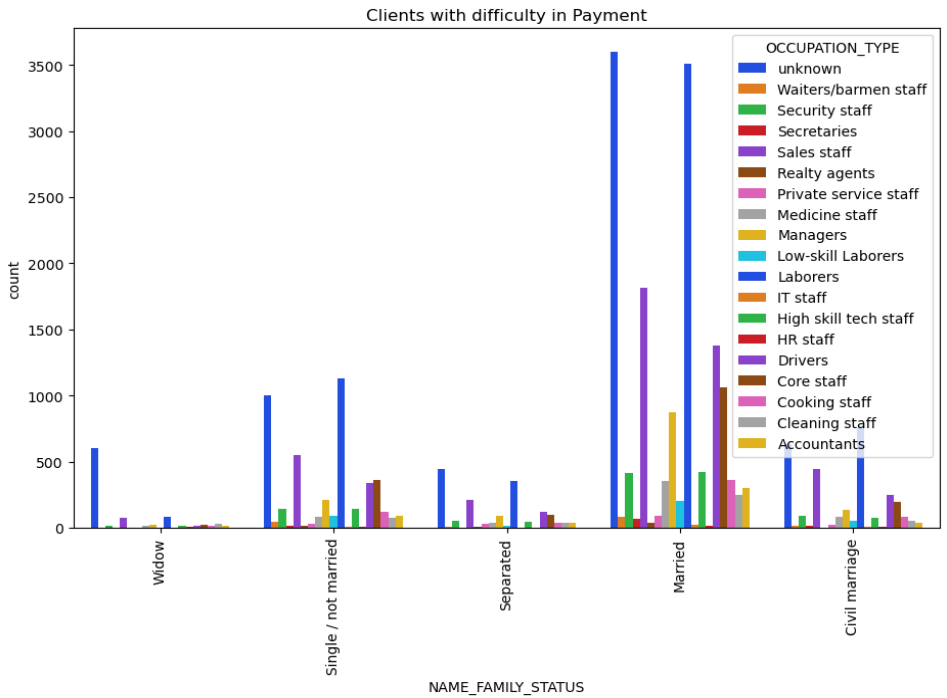


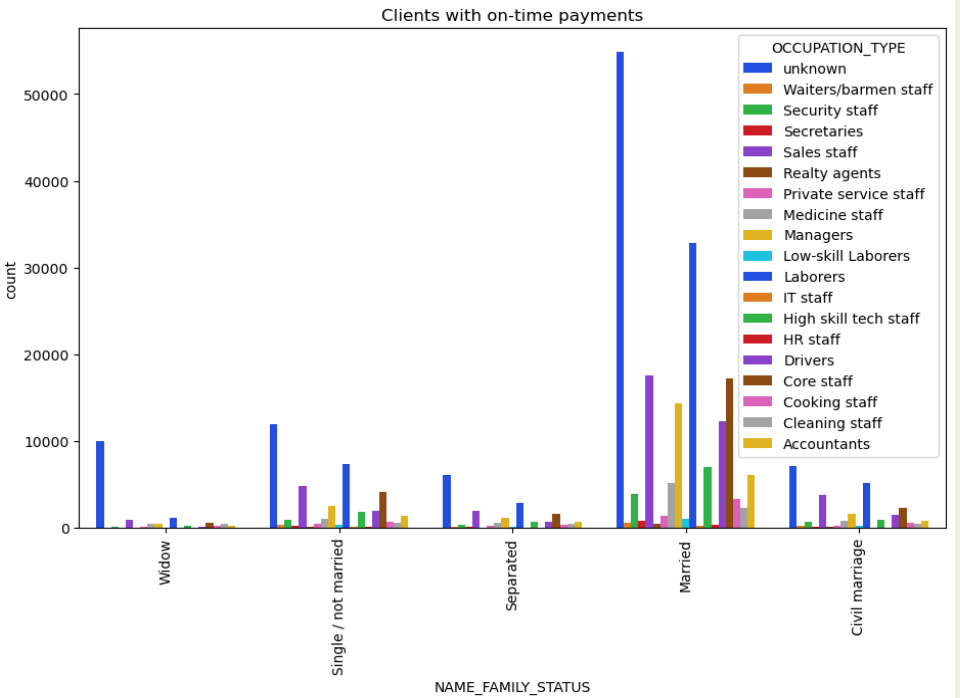


ANALYSIS OF "NAME_EDUCATION_TYPE" & "CODE_GENDER"

•- Clients who have Secondary/Secondary special education and Male have more Payment difficulties compared to On-Time Payments

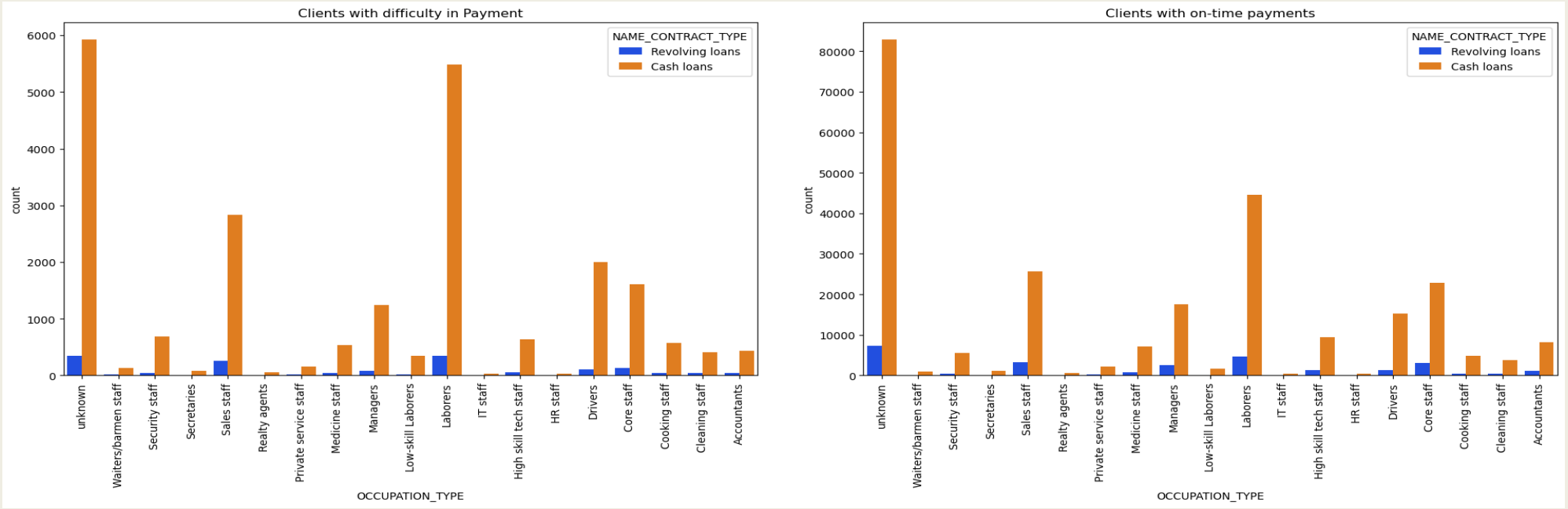
•- Clients who have Higher education and Female have more On-Time Payments compared to Payment difficulties





ANALYSIS OF "NAME_FAMILY_STATUS" &"OCCUPATION_TYPE"

- - Clients who are Single/not married, Married & Civil marriage and are Waiters/barmen staff have more Payment difficulties compared to On-Time Payments
- - Clients who are Single/not married & Married and are Laborers have more Payment difficulties compared to On-Time Payments
- - Clients who are Married and are Drivers have more Payment difficulties compared to On-Time Payments
- -Married and Accountants have better On-Time Payments

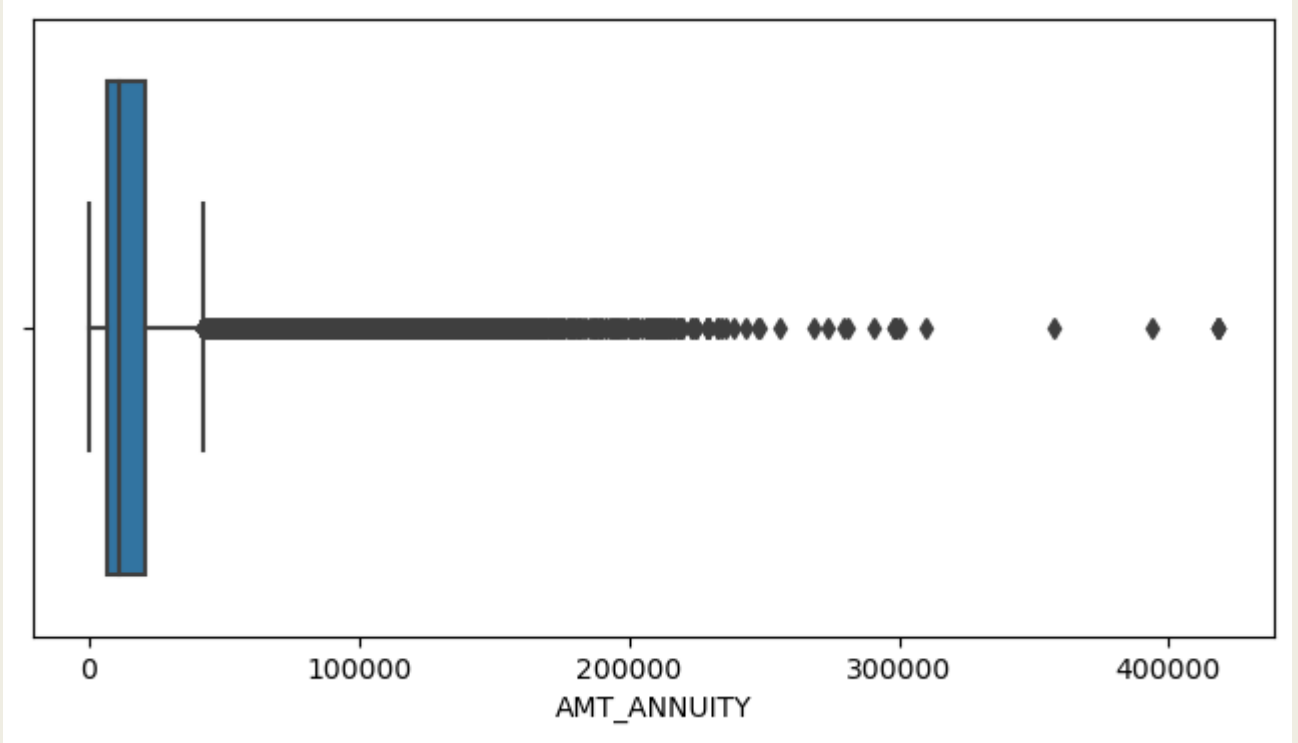


ANALYSIS OF "OCCUPATION_TYPE"&"NAME _CONTRACT_TYPE"

- Clients who are Sales staff,Laborers,Drivers and have Cash loans have more Payment difficulties compared to On-Time Payments.

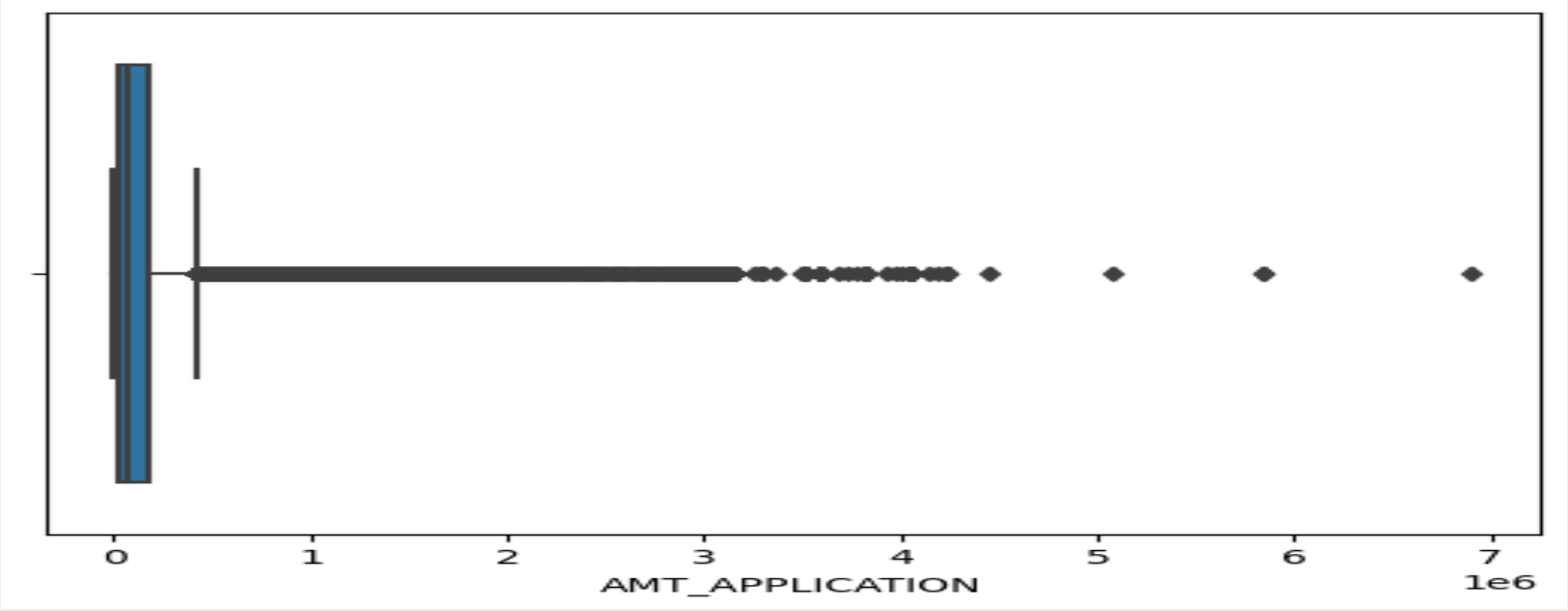
Analysis of information about the client’s previous loan data

Outlier analysis



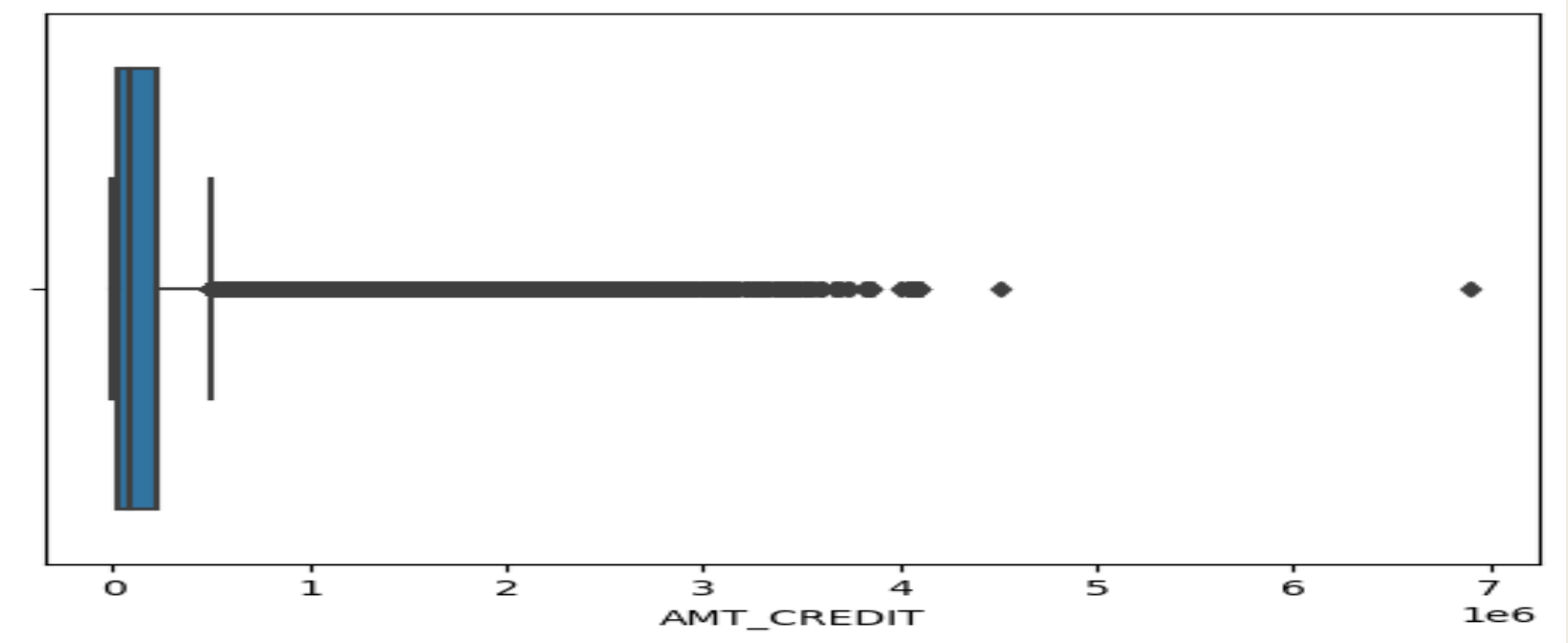
ANALYSIS OF "AMT_ANNUITY"

- - `AMT_ANNUITY` values above 42163.38 are outlier



ANALYSIS OF 'AMT_APPLICATION' COLUMN

- - `AMT_APPLICATION` values above 422820.0 are outlier

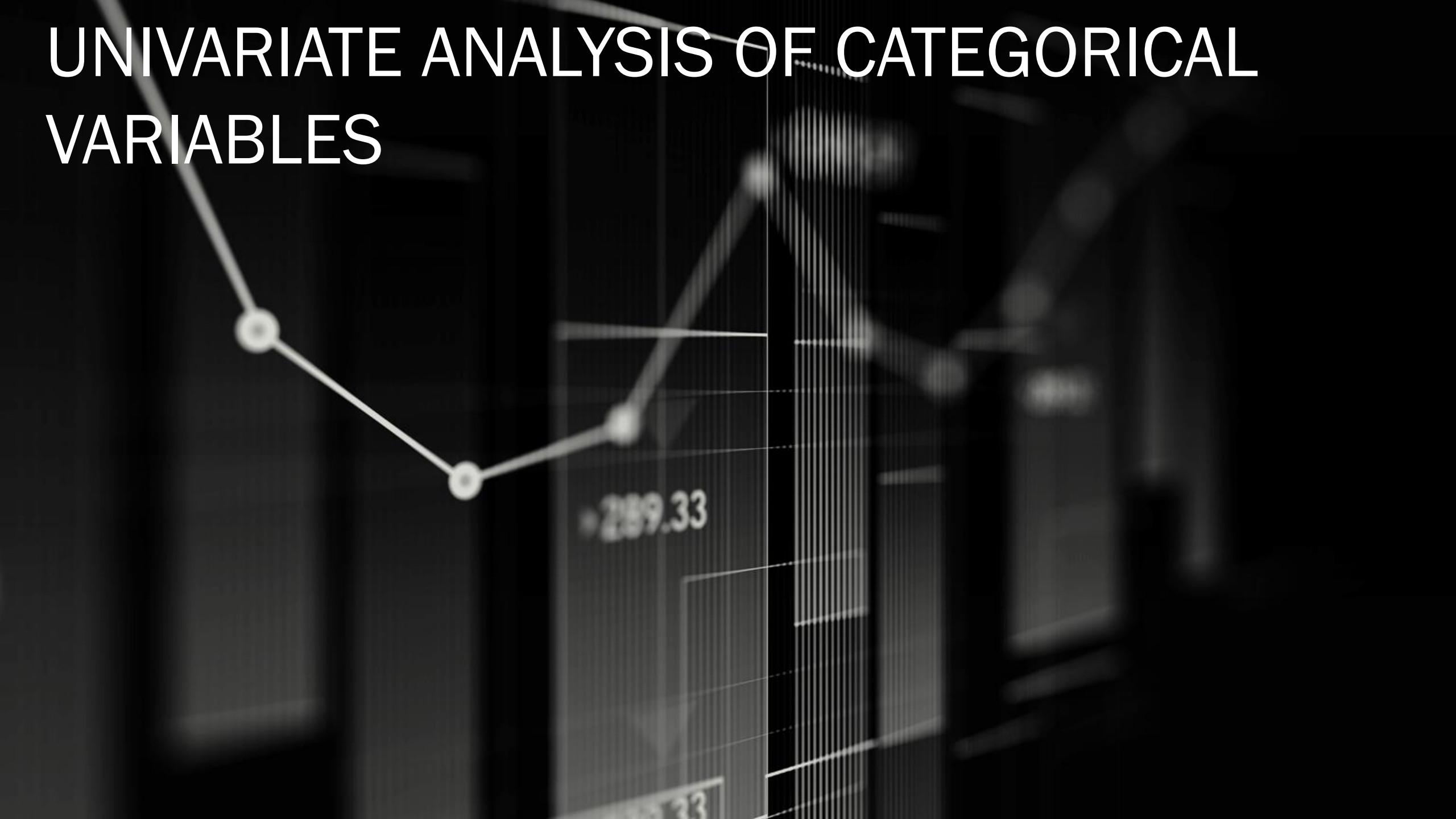


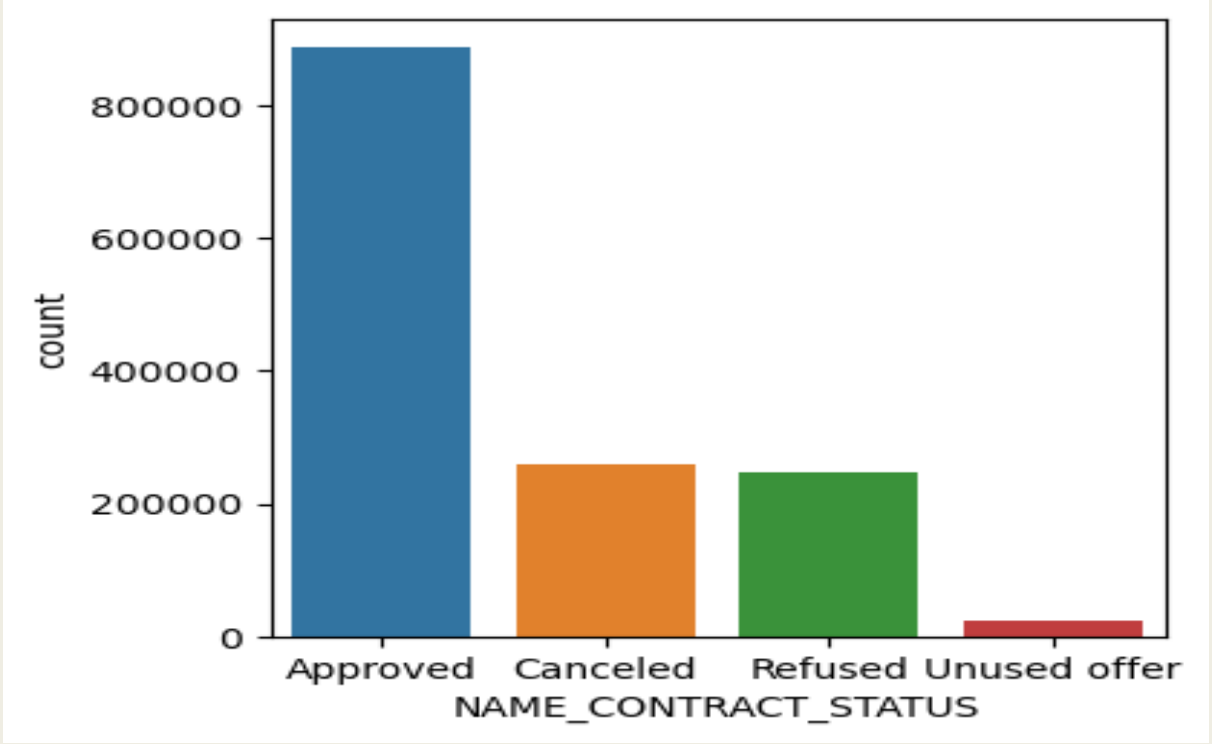
ANALYSIS OF "AMT_CREDIT" COLUMN

- `AMT_CREDIT` values above 504805.5 are outlier

Analysis of merged information about the client’s previous loan data and current loan application

UNIVARIATE ANALYSIS OF CATEGORICAL VARIABLES

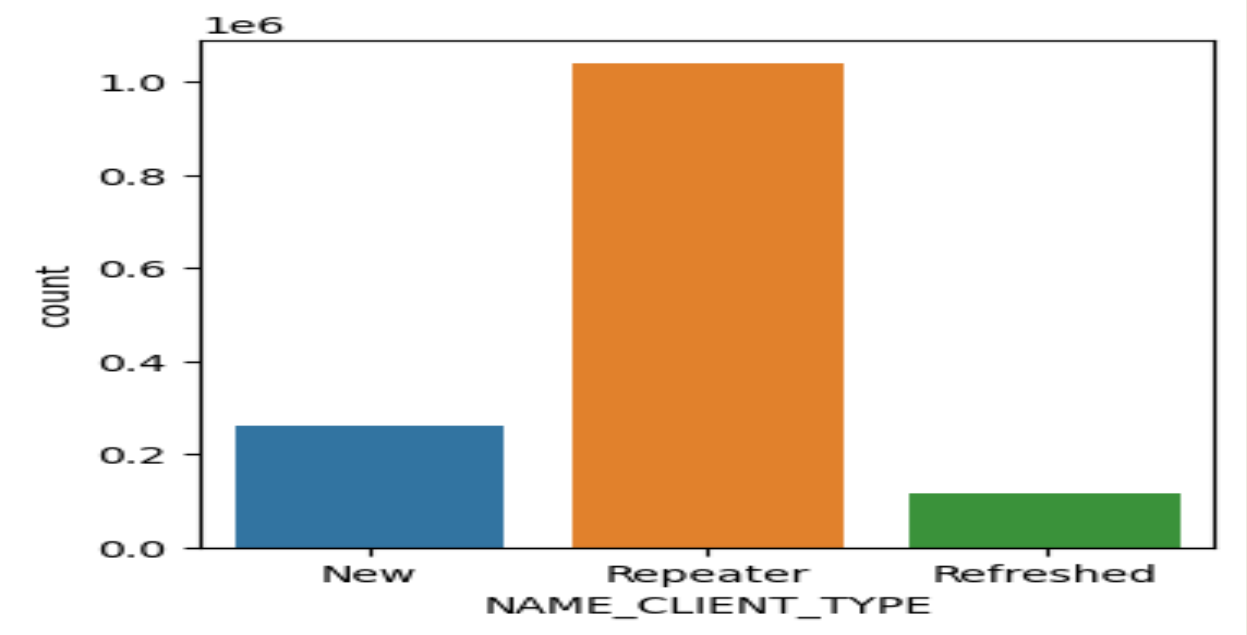




Analysis of 'NAME_CONTRACT_STATUS'

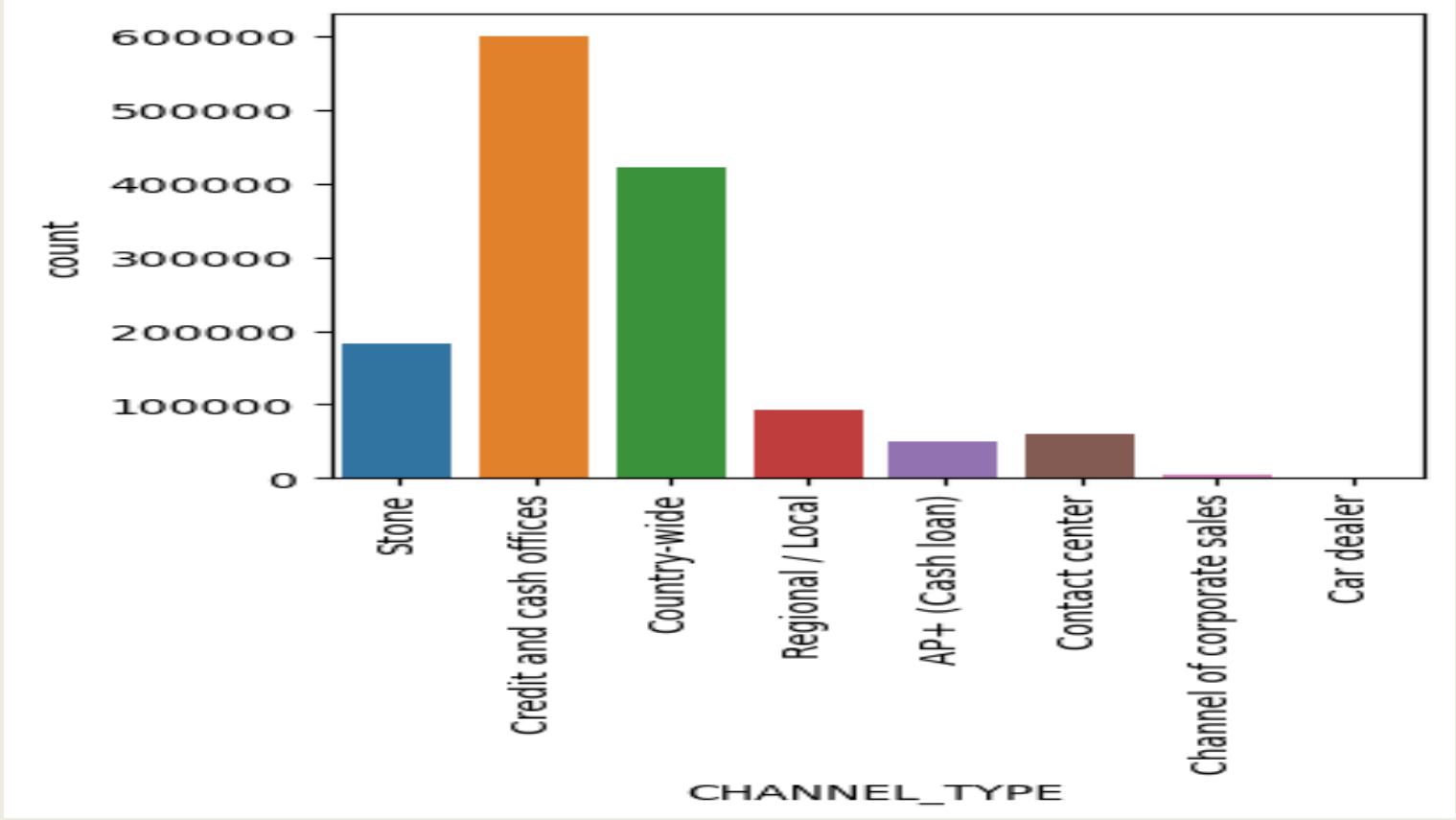
- - Approved loan status is the highest among all loan applications

- - Canceled loan status is the second highest among all loan application



ANALYSIS OF "NAME_CLIENT_TYPE"

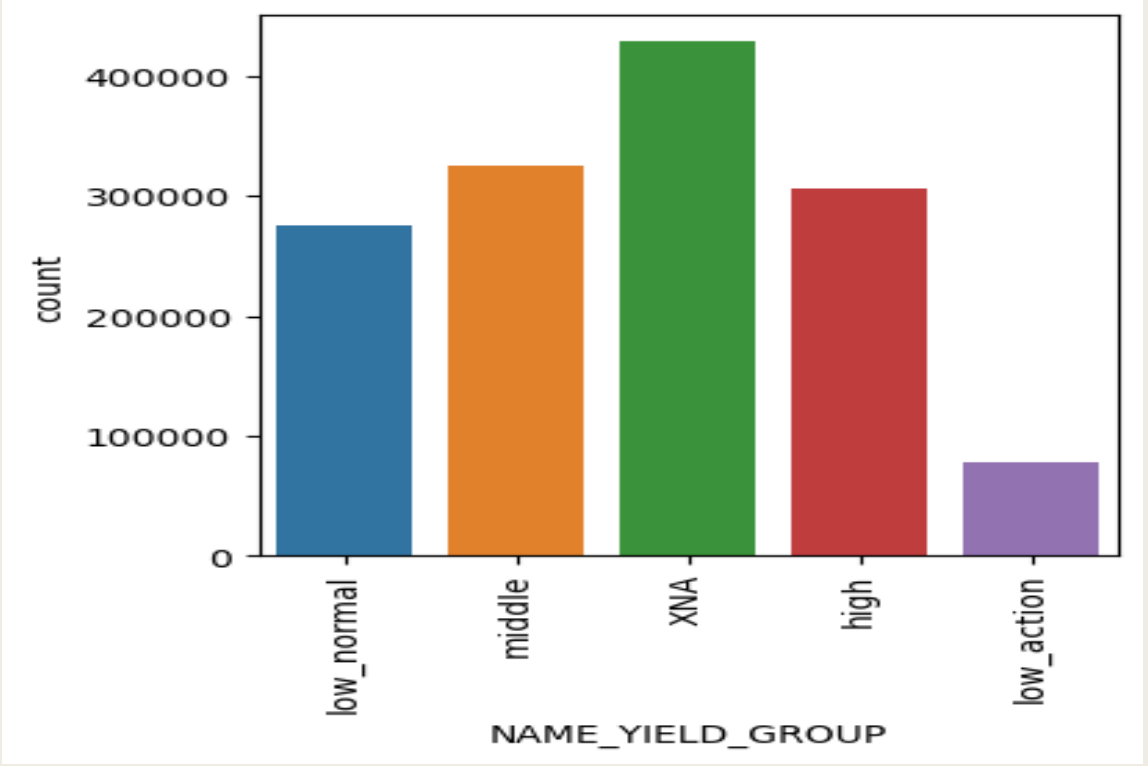
- Repeater loan applicants are higher in number
- New loan applicants are second higher in number



ANALYSIS OF "CHANNEL_TYPE"

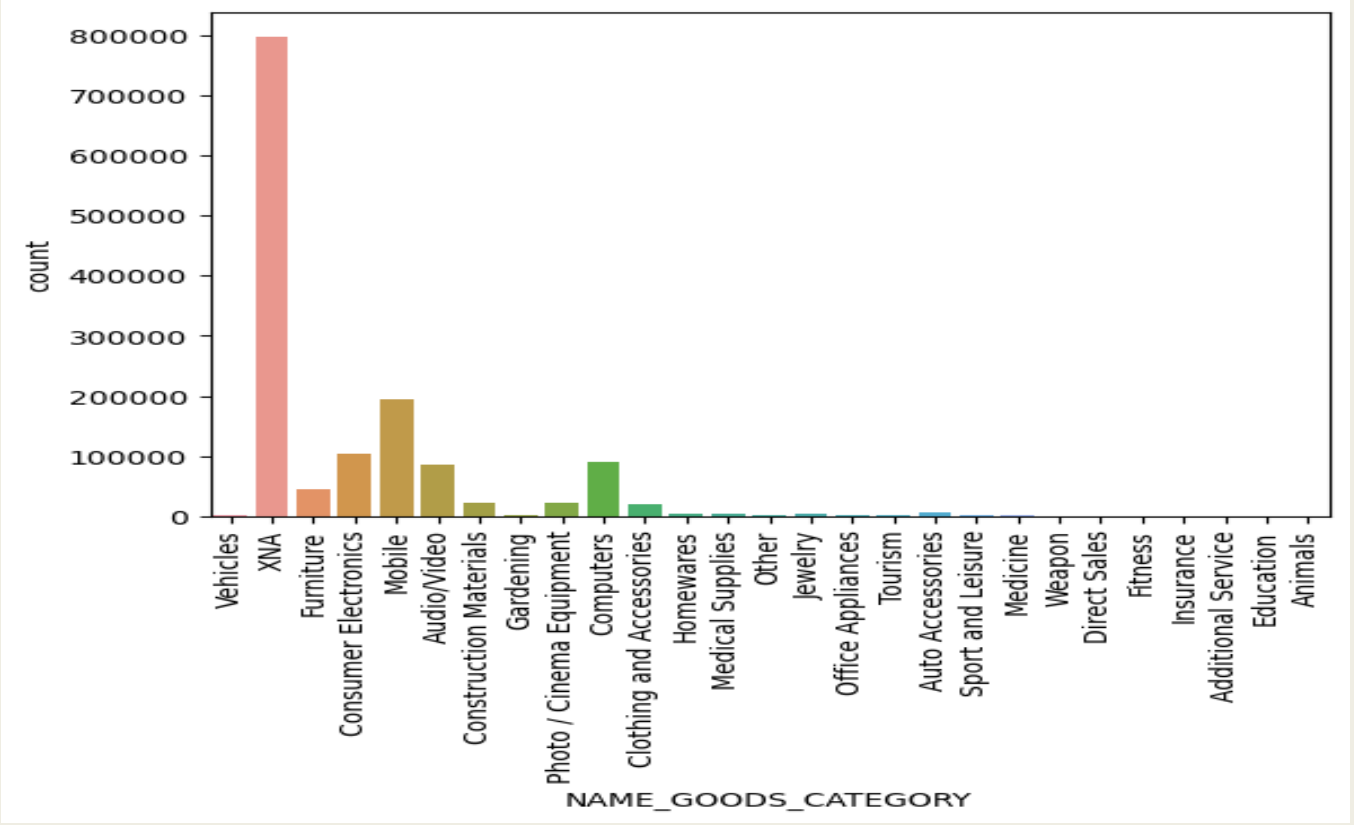
•- "Country-wide" Channel type is the highest among all loan applications

•- "Credit and cash offices" is the second highest Channel Type among all loan applications



ANALYSIS OF "NAME_YIELD_GROUP"

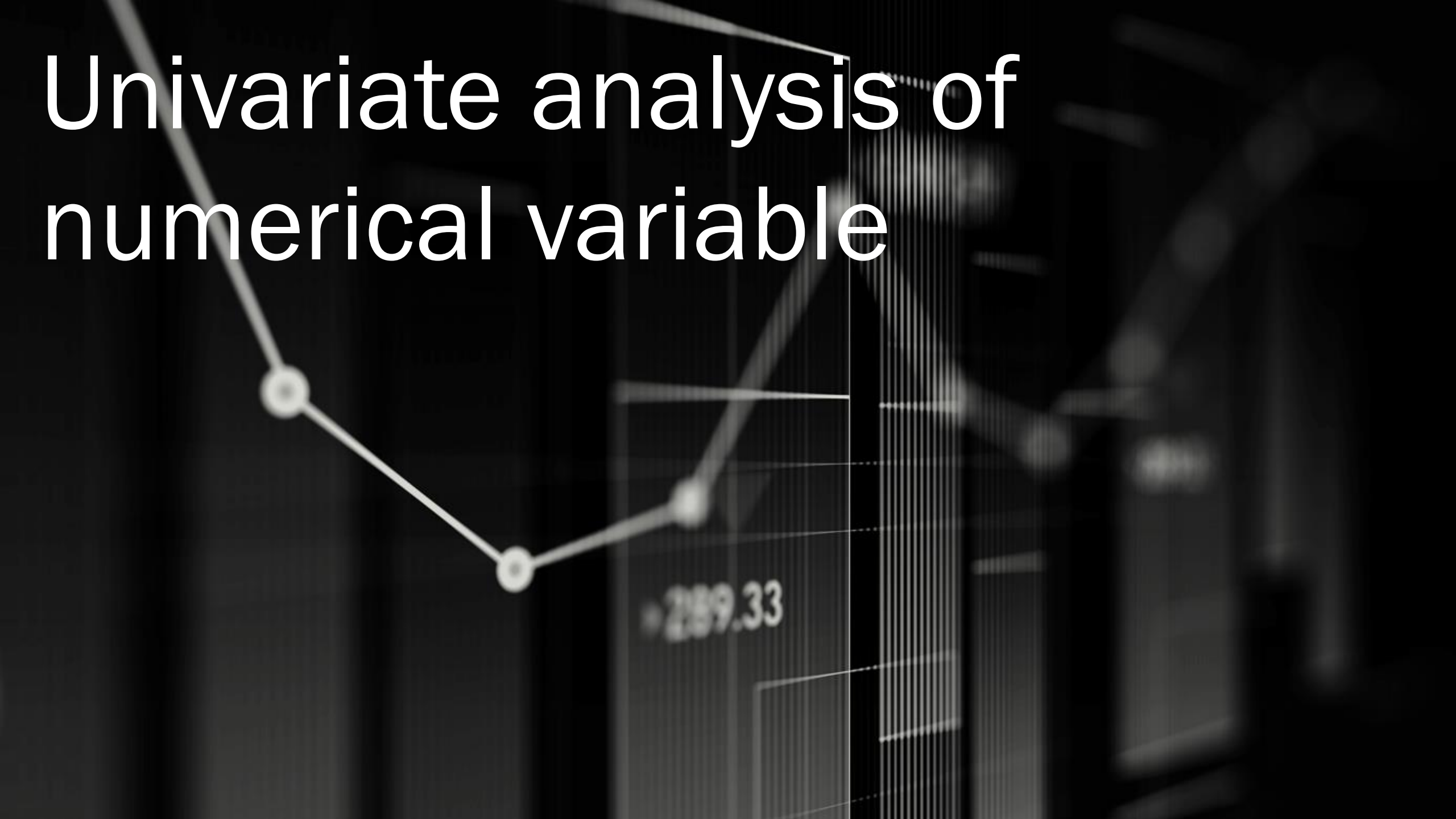
- XNA interest rate is the highest among all loan applications
- middle and high interest rates are the second and third highest among all loan applications

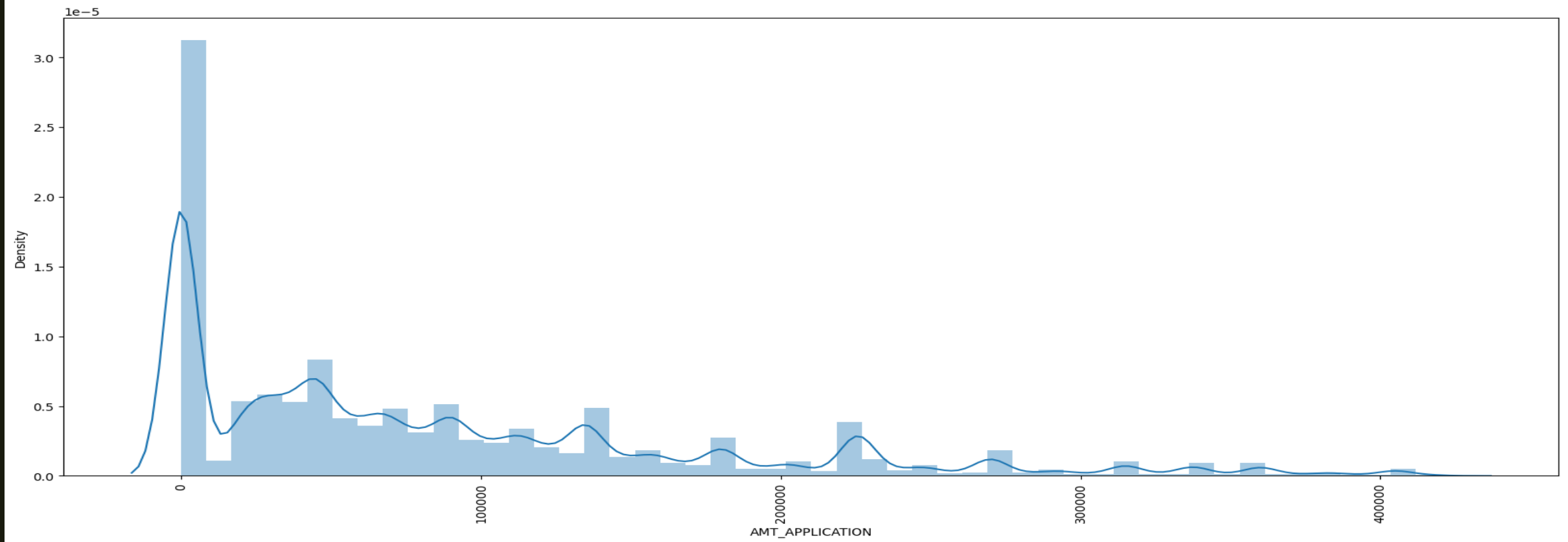


ANALYSIS OF "NAME_GOODS_CATEGORY"

- XNA goods category is the highest among all loan applications
- mobile goods categories the second highest among all loan applications

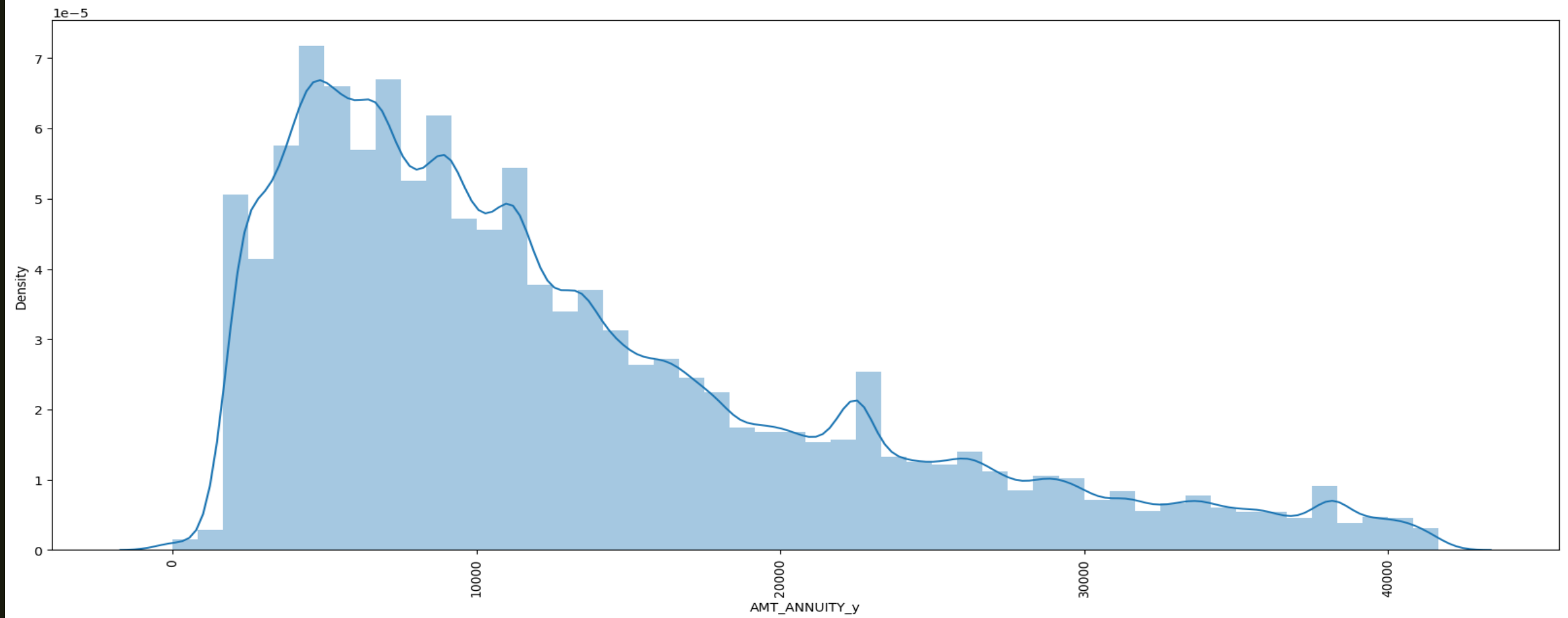
Univariate analysis of numerical variable





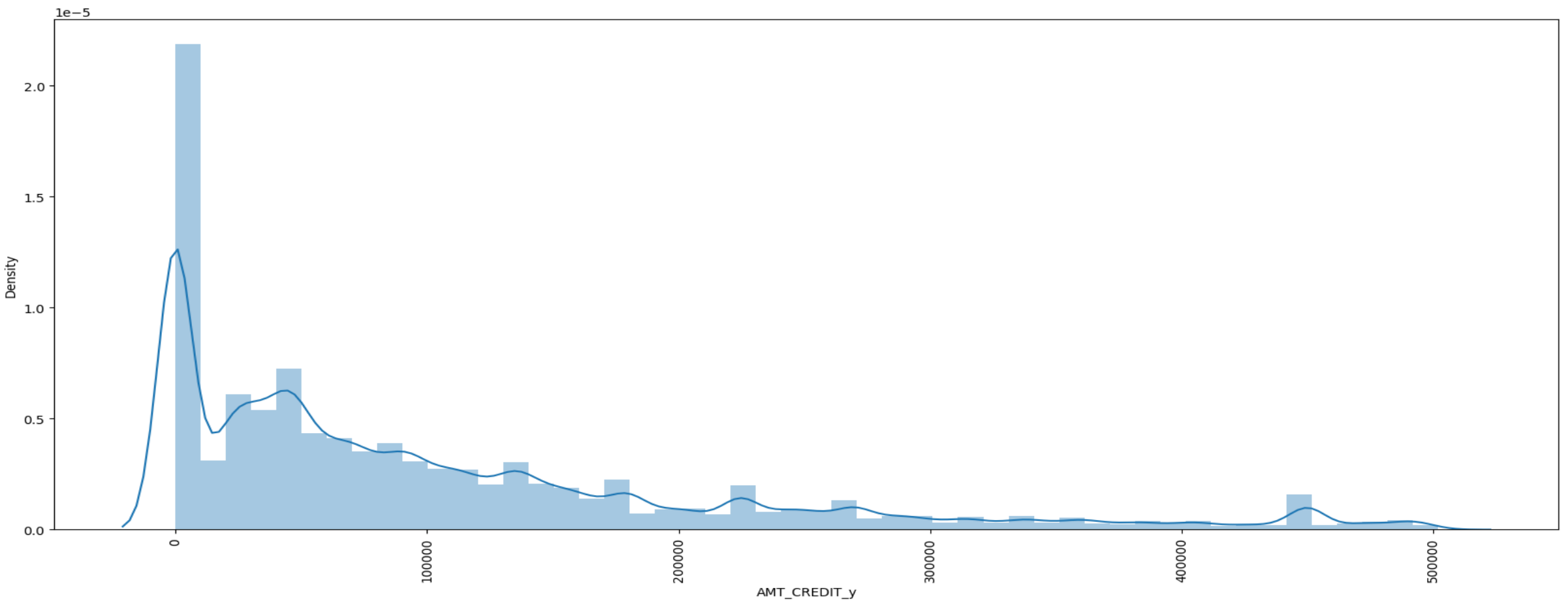
Analysis of 'AMT_APPLICATION'

- Most of the loan amount applied by the clients initially seems to be very small as can be seen from the huge spike at the beginning of the distribution



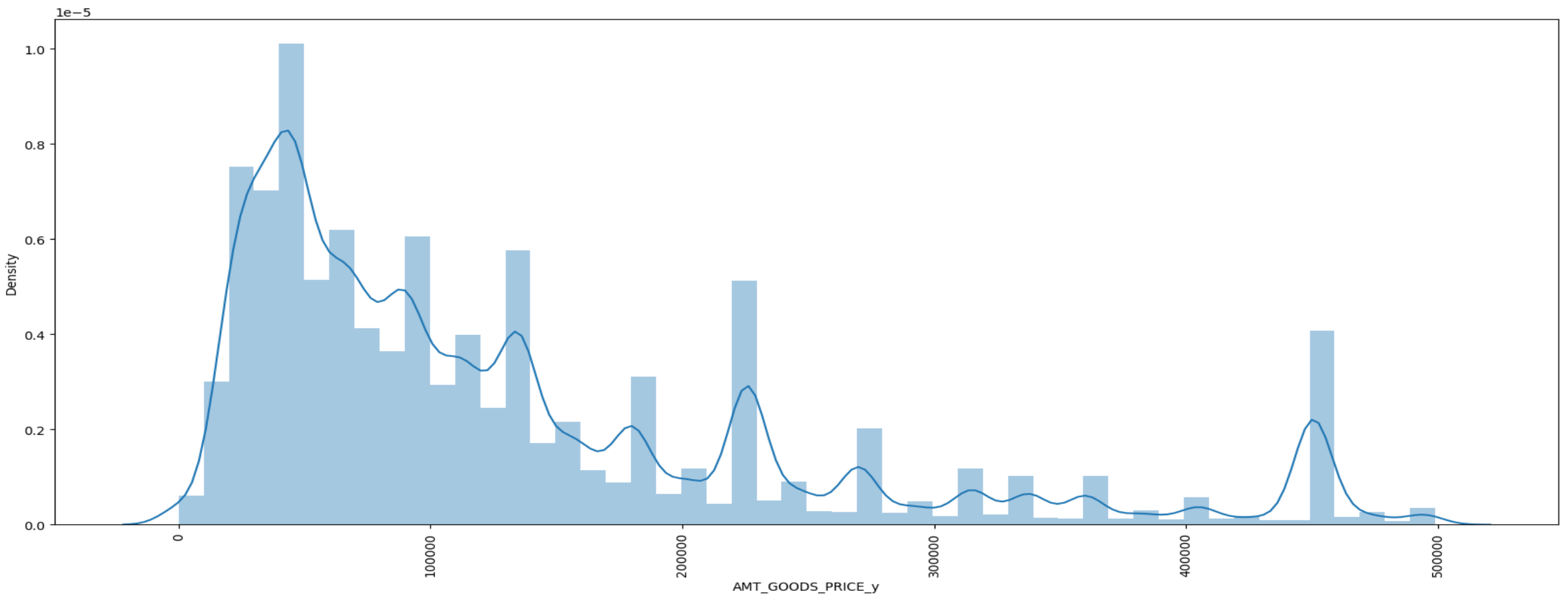
Analysis of 'AMT_ANNUITY_y'

- Most of the previous loan's annuity from the clients is less than 10,000 as the distribution is high here.
- As previous loan's annuity increases, the no. of clients decreases.



Analysis of 'AMT_CREDIT_y'

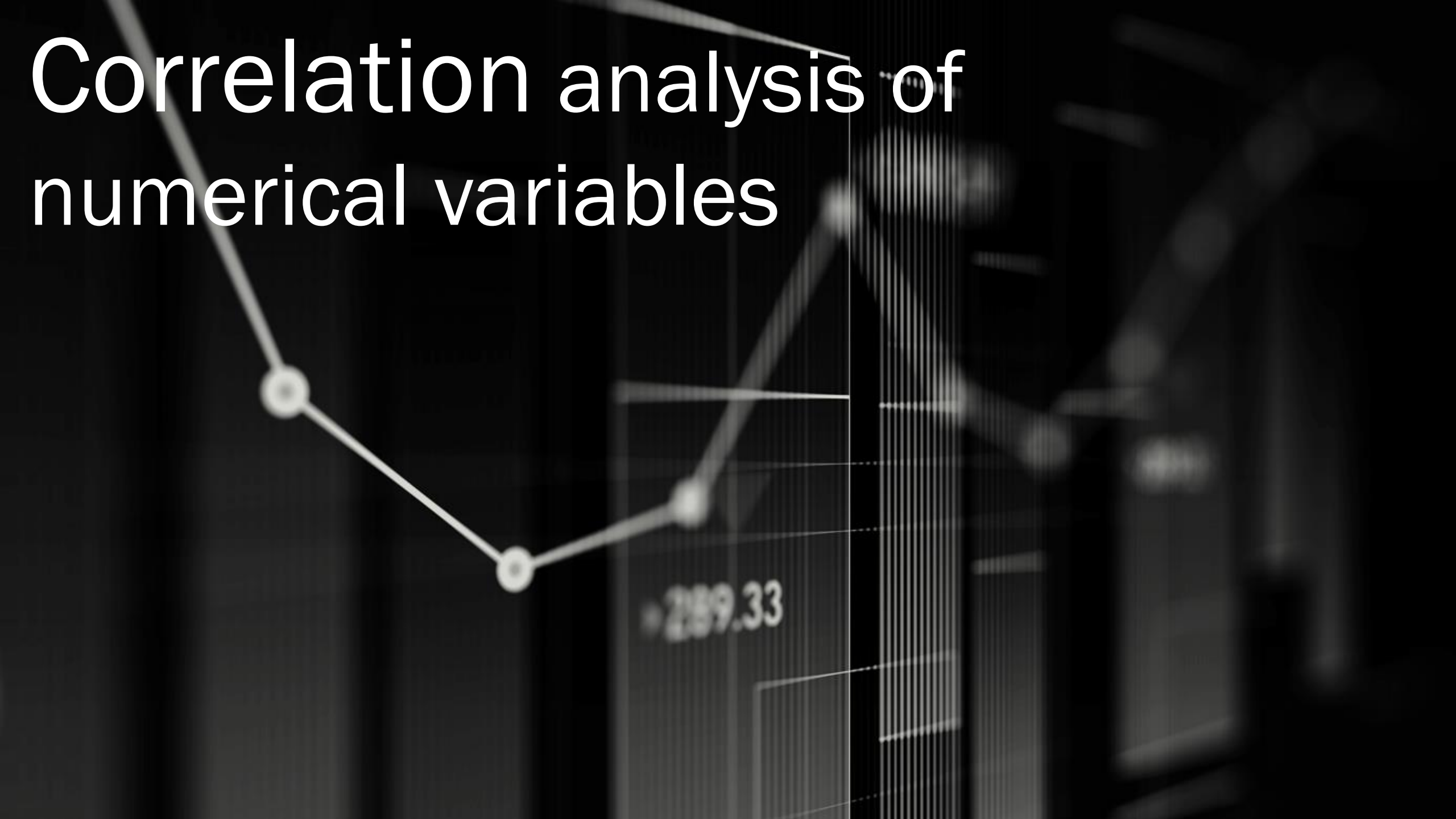
- This distribution very closely resembles that of AMT_APPLICATION. This means that most people received the loan amount that they applied for.

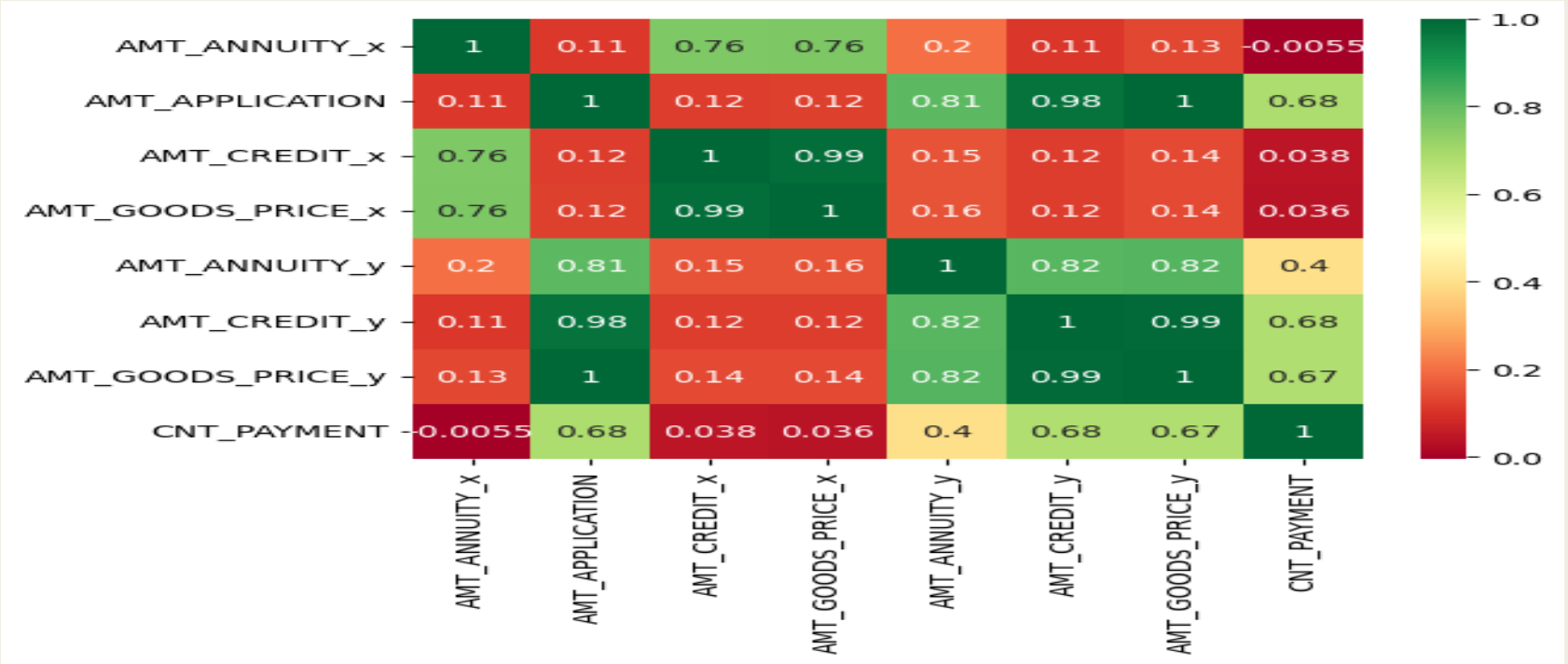


Analysis of 'AMT_GOODS_PRICE_y'

- Most of the goods price asked by clients in previous application is less than 100K

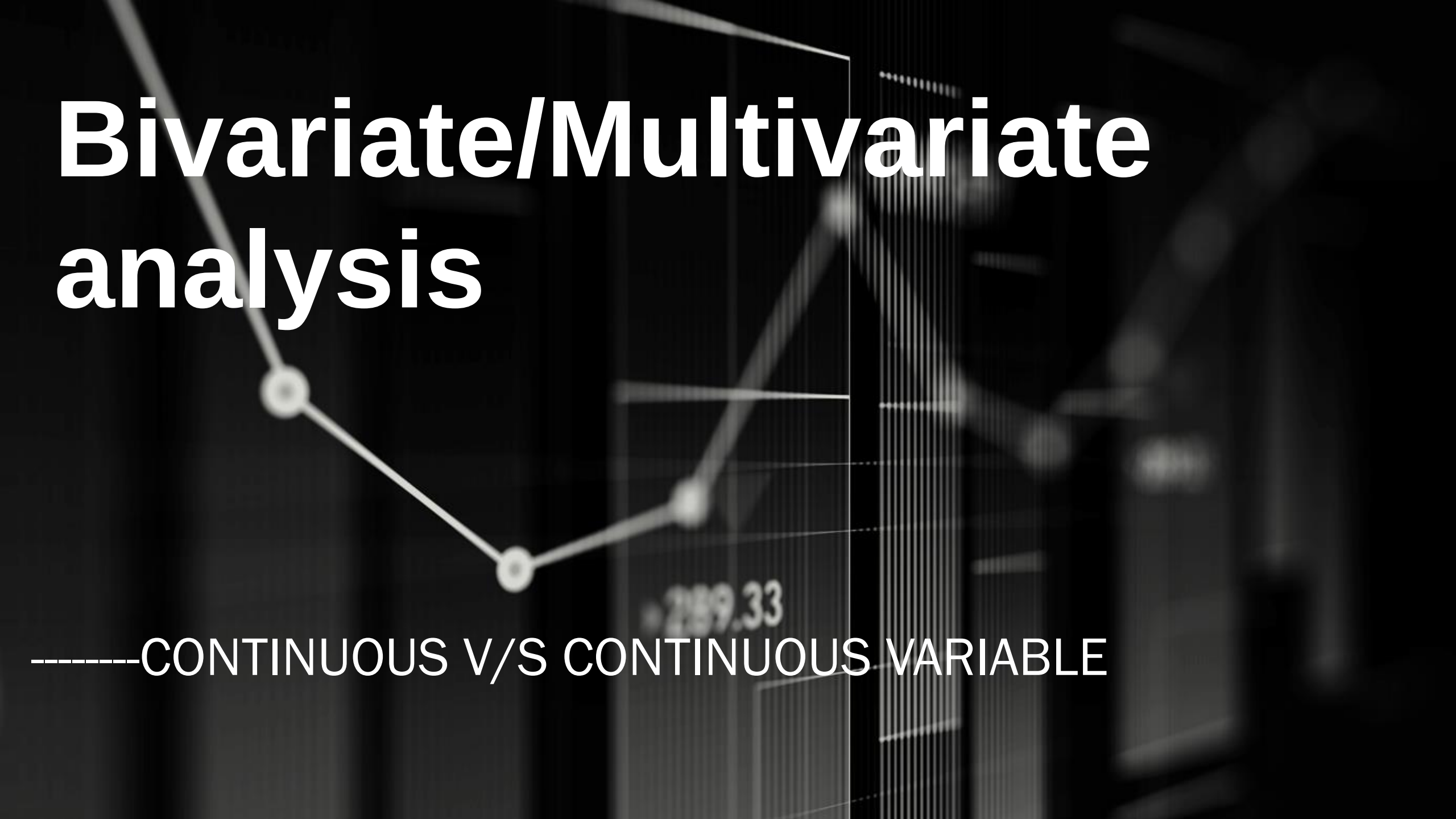
Correlation analysis of numerical variables





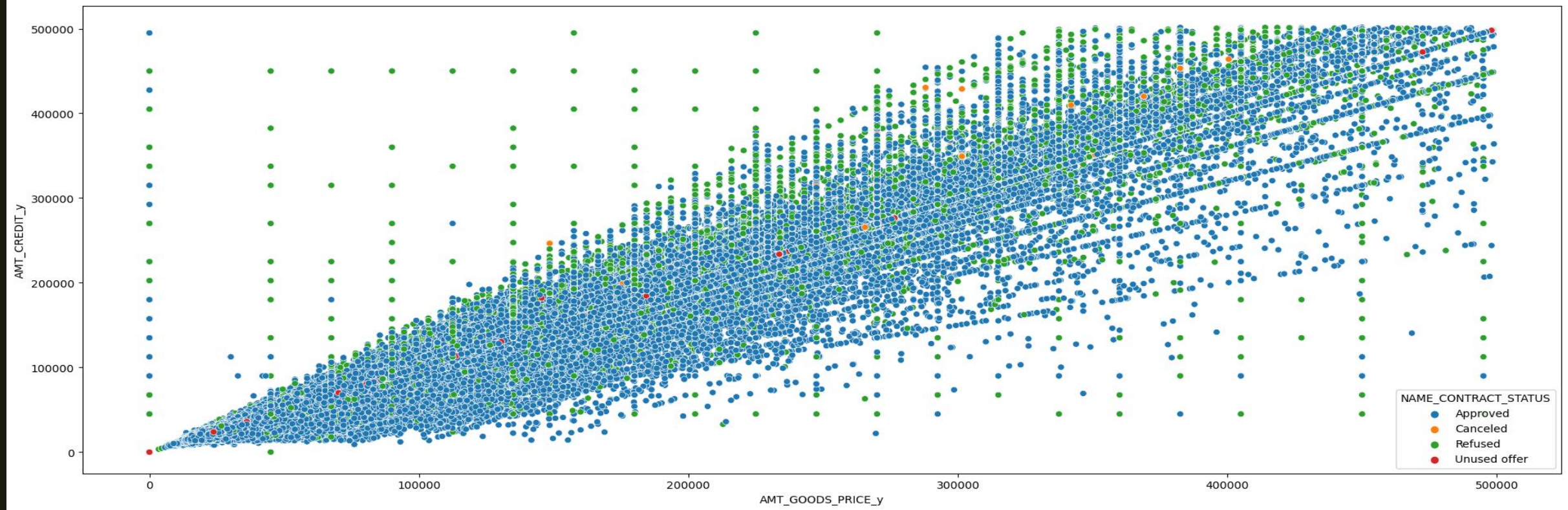
Correlation matrix

- AMT_APPLICATION has a high correlation with AMT_ANNUITY_y,AMT_CREDIT_y,AMT_GOODS_PRICE_y and decent correlation with CNT_PAYMENT
- AMT_GOODS_PRICE_y has a high correlation with AMT_ANNUITY_y,AMT_CREDIT_y,AMT_APPLICATION and decent correlation with CNT_PAYMENT
- AMT_CREDIT_y has a high correlation with AMT_GOODS_PRICE_y and decent correlation with CNT_PAYMENT
- AMT_ANNUITY_x has a high correlation with AMT_GOODS_PRICE_y,AMT_CREDIT_y
- AMT_ANNUITY_x has a high correlation with AMT_GOODS_PRICE_x,AMT_CREDIT_x
- AMT_CREDIT_x has a high correlation with AMT_GOODS_PRICE_x



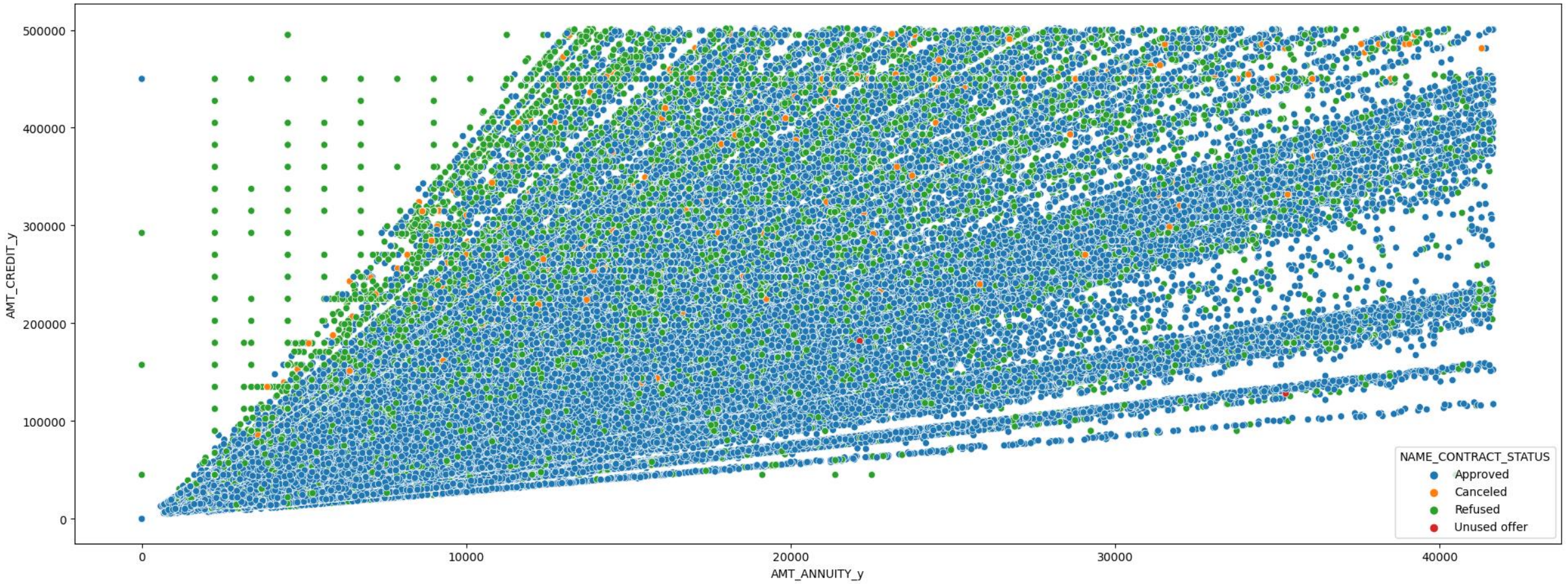
Bivariate/Multivariate analysis

-------CONTINUOUS V/S CONTINUOUS VARIABLE



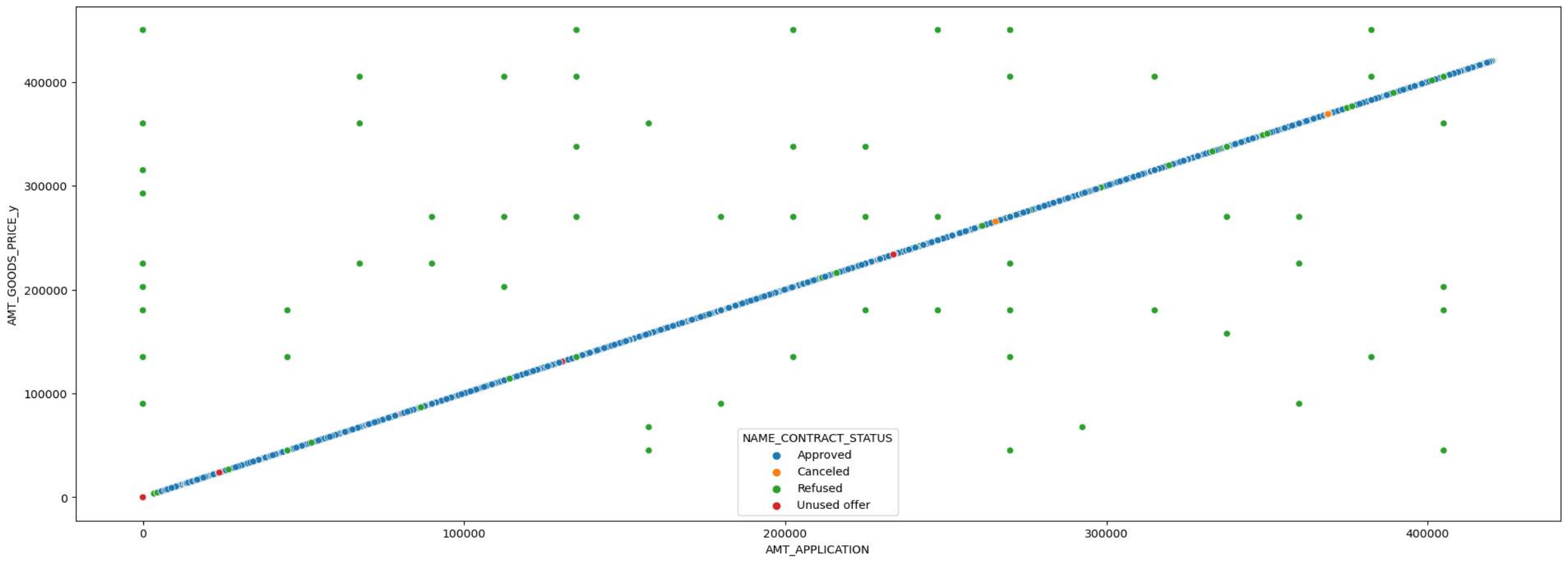
Analysis of 'AMT_GOODS_PRICE_y' V/S 'AMT_CREDIT_y' V/S 'NAME_CONTRACT_STATUS'

- At lower levels of previous application's Goods price < 200K and Credit > 300k, have a chance of getting refused. However, this is a weak correlation as we have less data points to support this.



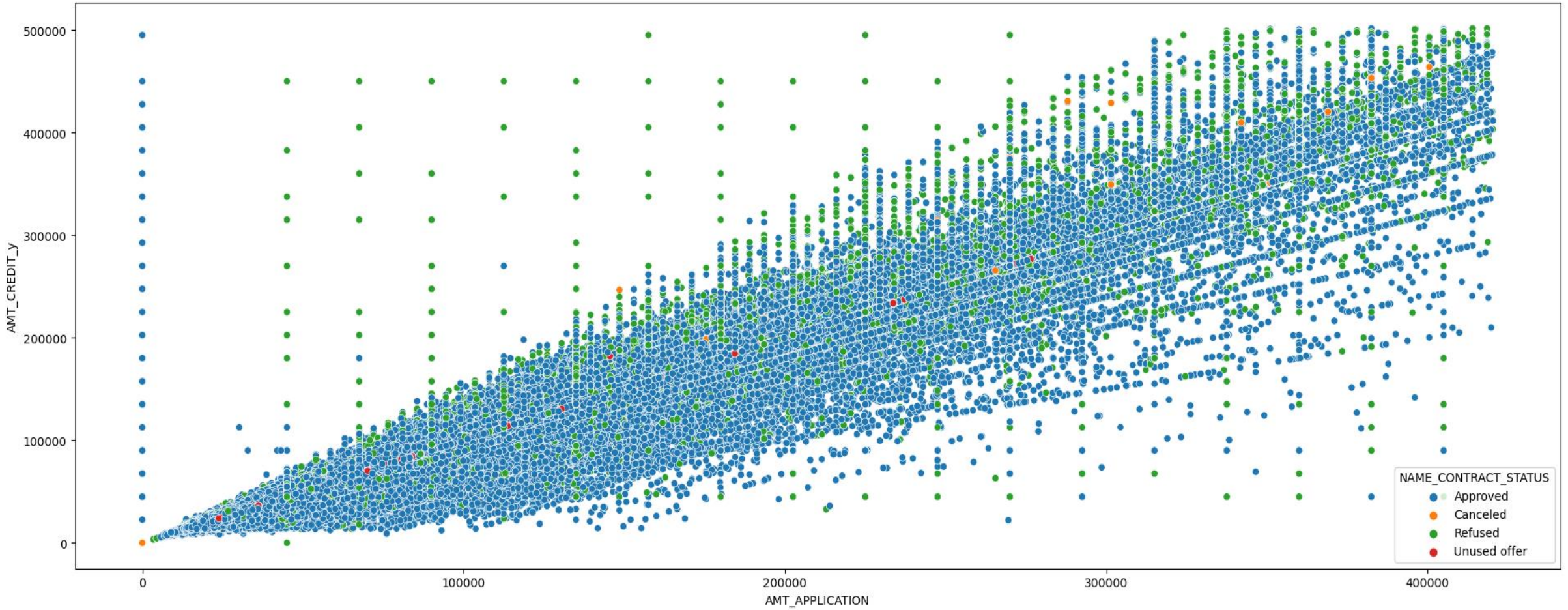
Analysis of 'AMT_ANNUITY_y' V/S 'AMT_CREDIT_y' V/S 'NAME_CONTRACT_STATUS'

- There are lots of refusal observations with Annuity amount < 10000 and Credit amount > ~250K. This might be because higher credit amount should also require higher Annuity from Client to pay it



Analysis of 'AMT_APPLICATION' V/S 'AMT_GOODS_PRICE_y' V/S 'NAME_CONTRACT_STATUS'

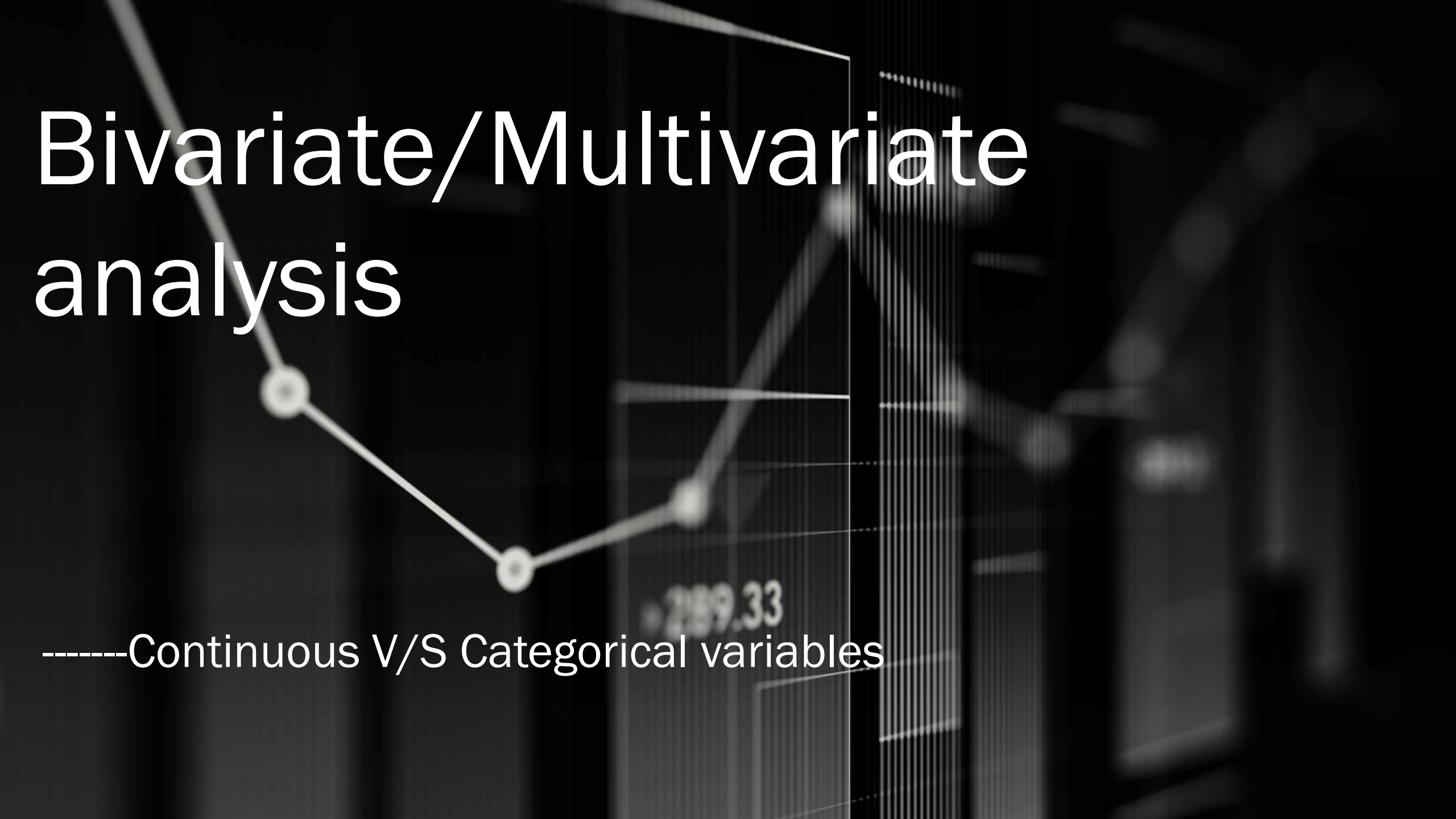
- Application amount has strong positive correlation with Goods price



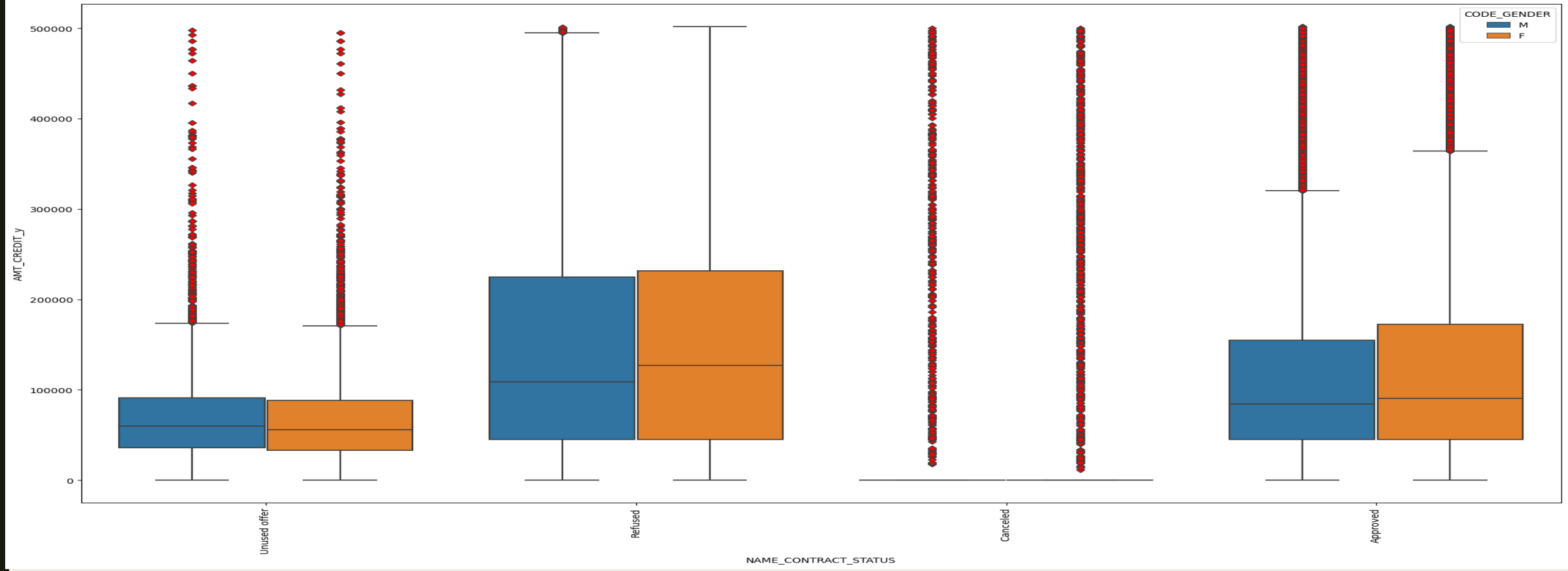
Analysis of 'AMT_APPLICATION' V/S 'AMT_CREDIT_y' V/S 'NAME_CONTRACT_STATUS'

- - Application amount has strong positive correlation with Credit amount

Bivariate/Multivariate analysis

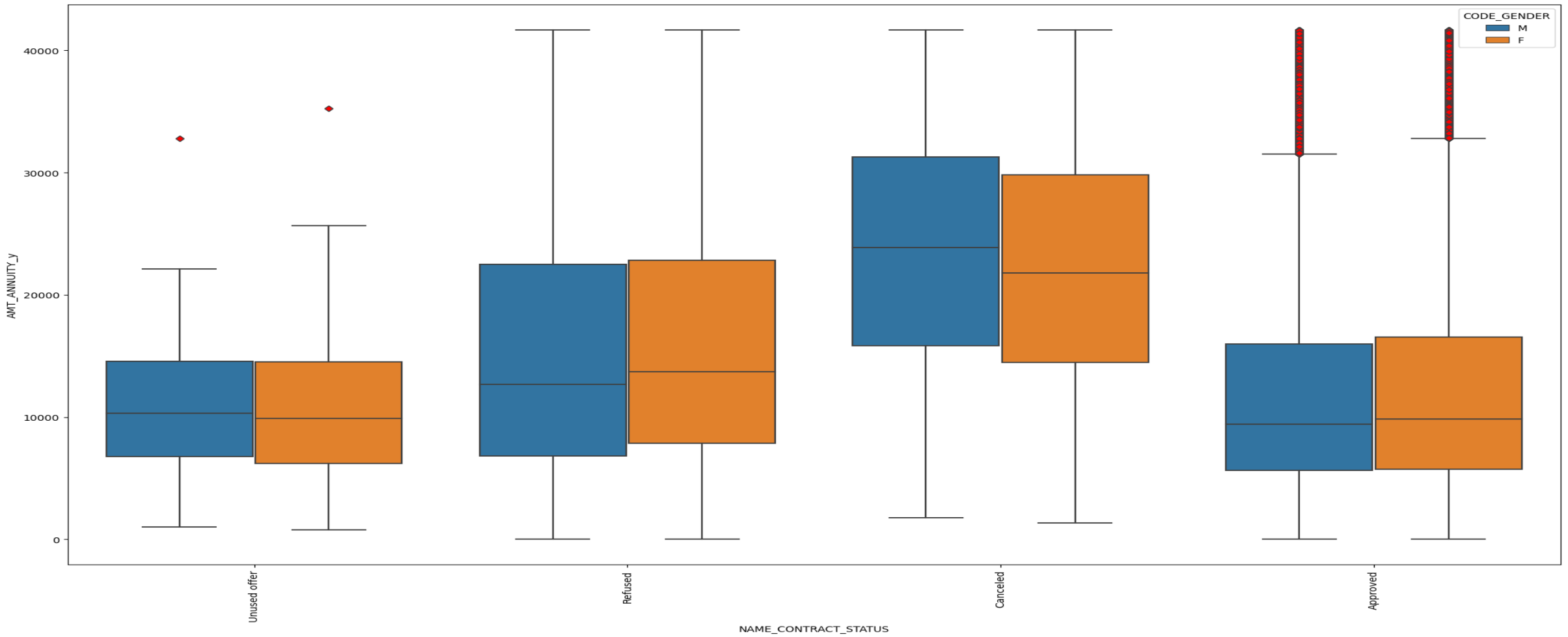


------Continuous V/S Categorical variables



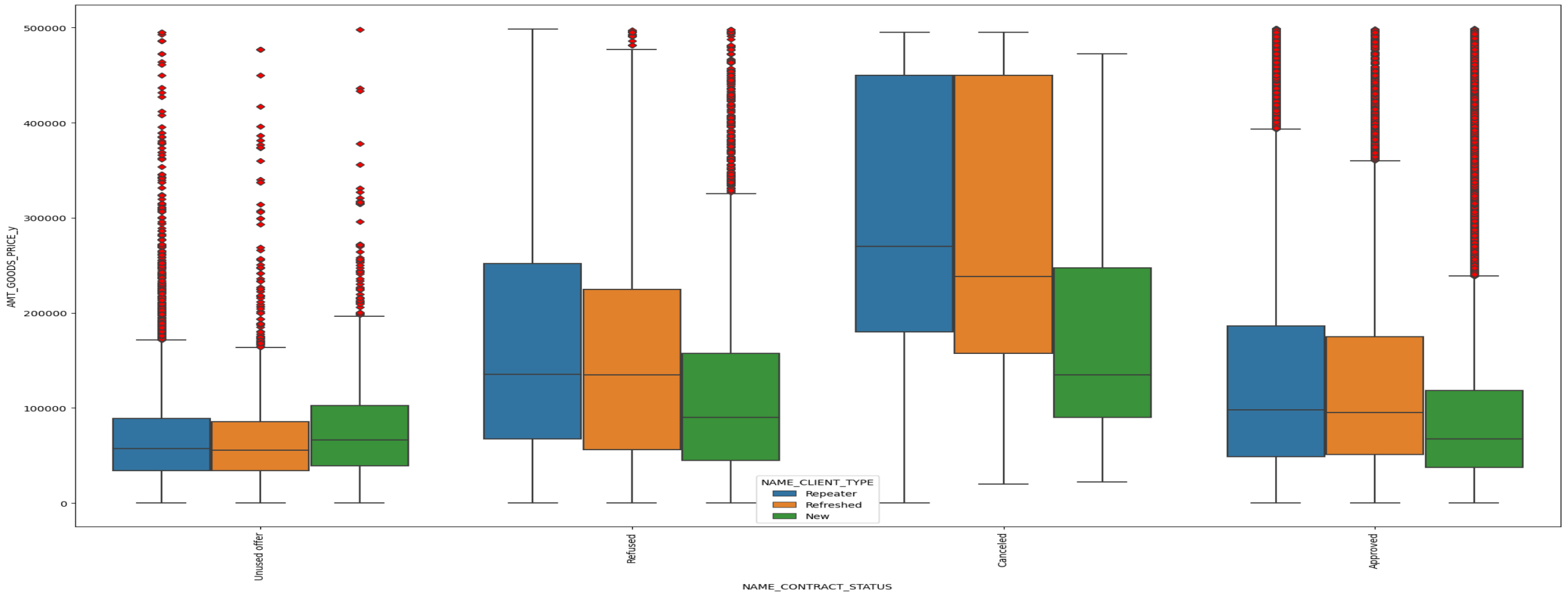
Analysis of 'NAME_CONTRACT_STATUS' V/S 'AMT_CREDIT_y' V/S 'CODE_GENDER'

- Clients who are Refused and Female apply for higher median credit amount than Male



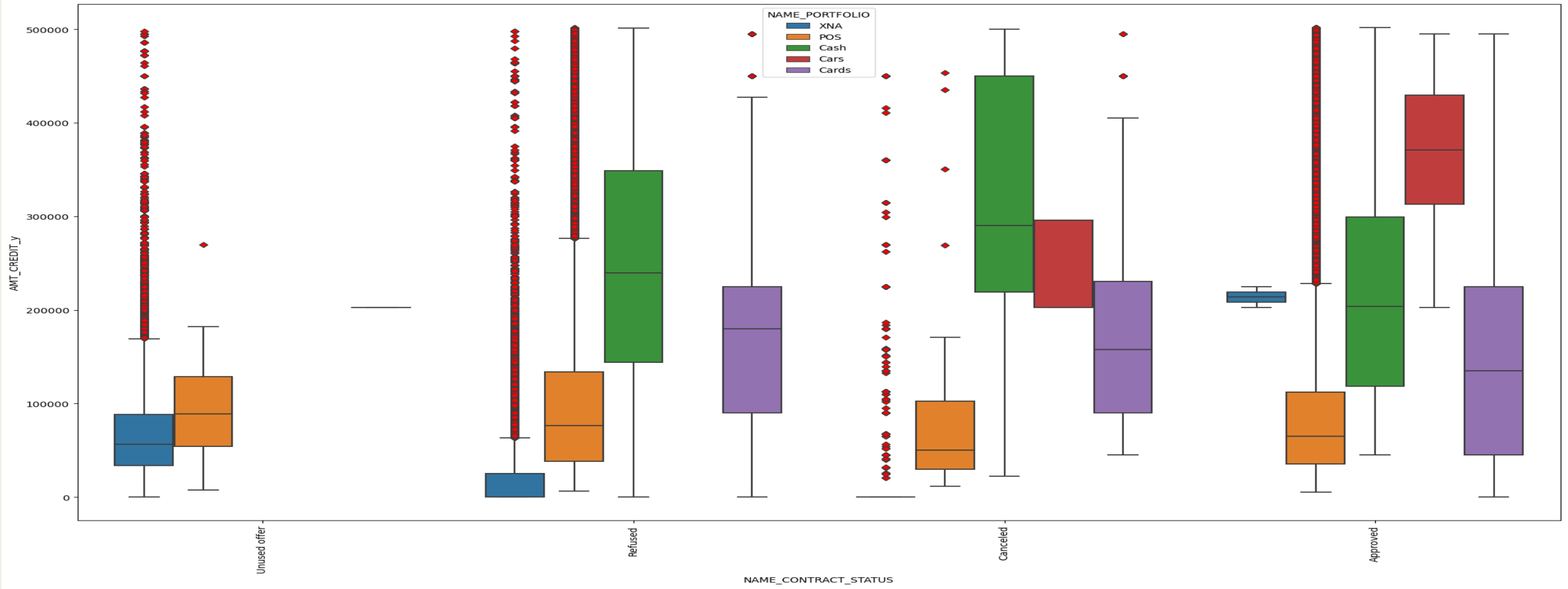
Analysis of 'NAME_CONTRACT_STATUS ' V/S 'AMT_ANNUITY_y' V/S 'CODE_GENDER'

- Clients who got Cancelled and Male paid higher median Annuity than Female
- Clients who got Refused and Female paid higher median Annuity than Male



Analysis of 'NAME_CLIENT_TYPE' V/S 'AMT_GOODS_PRICE_y' V/S 'NAME_CONTRACT_STATUS'

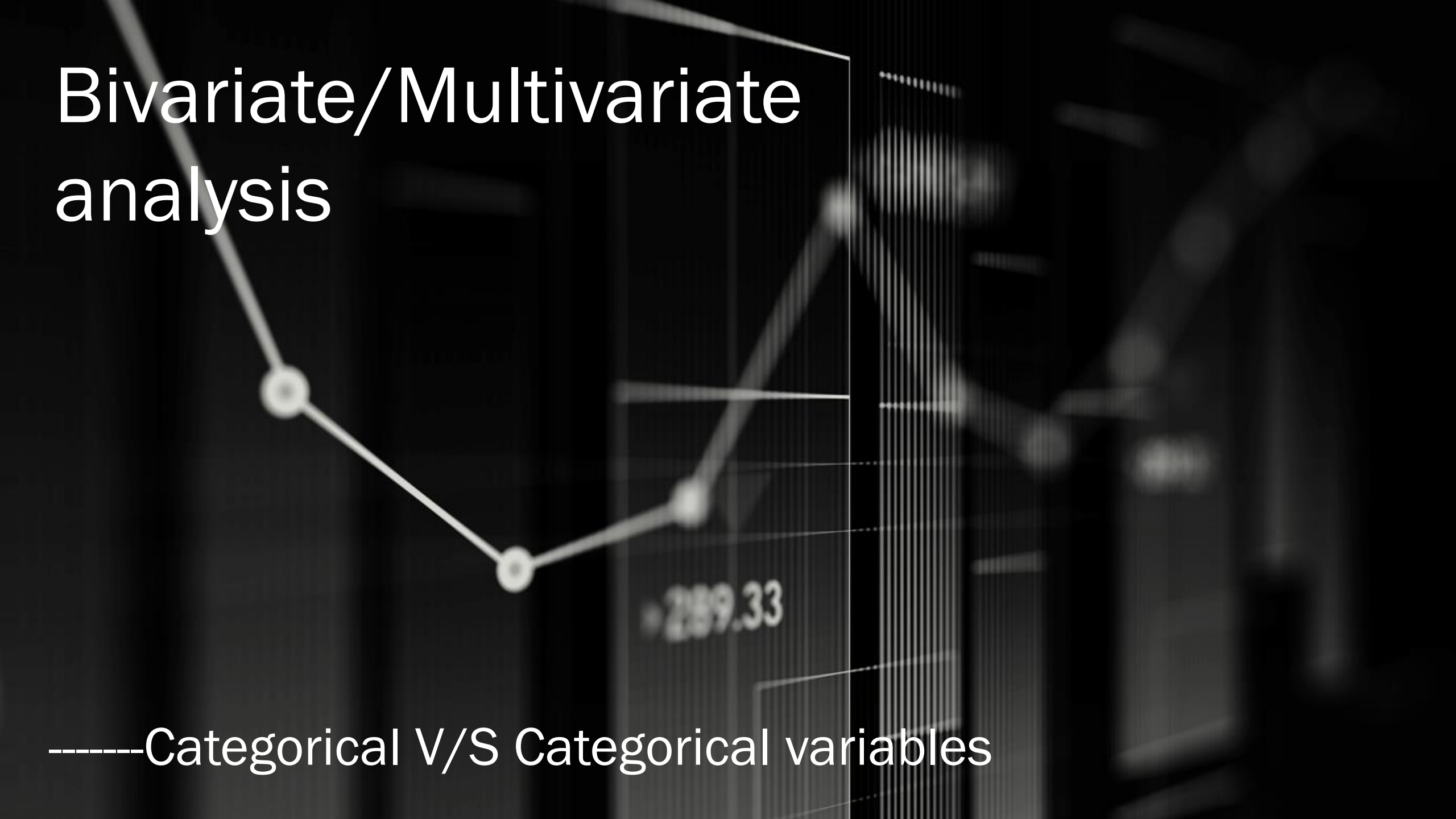
- Clients who are New and Canceled have less median goods price compared to Repeater and Refreshed
- Clients who are Approved and New have less median goods price compared to Repeater and Refreshed



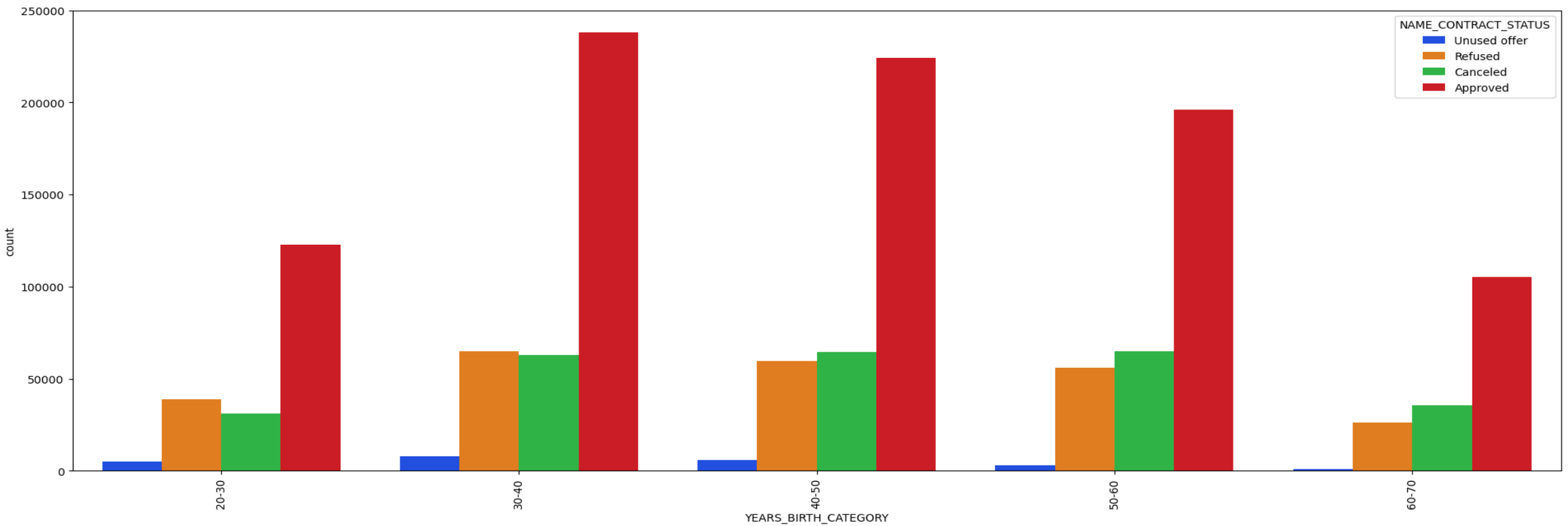
Analysis of 'NAME_CONTRACT_STATUS' V/S 'AMT_CREDIT_y' V/S 'NAME_PORTFOLIO'

- Clients who have Unused offer receive more median credit in POS portfolio
- Clients who are Refused receive more median credit in Cash portfolio
- Clients who are Approved receive more median credit in Cars portfolio

Bivariate/Multivariate analysis

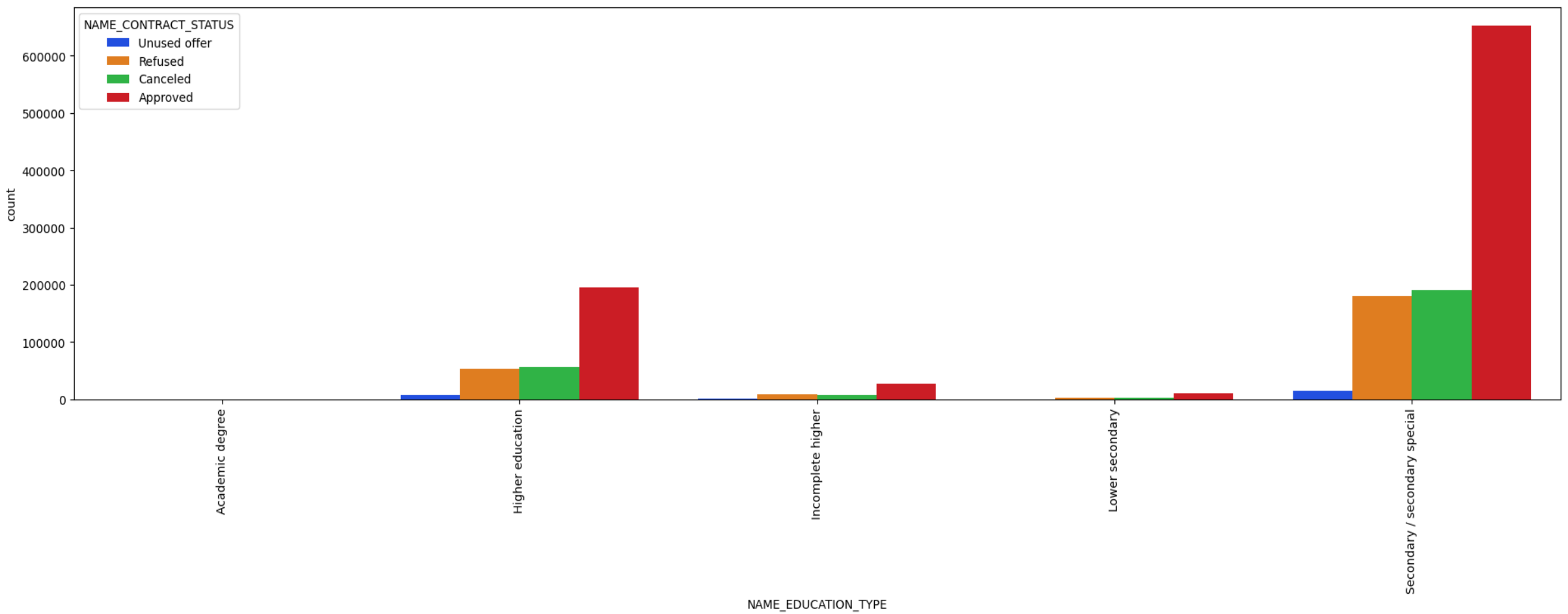


------Categorical V/S Categorical variables



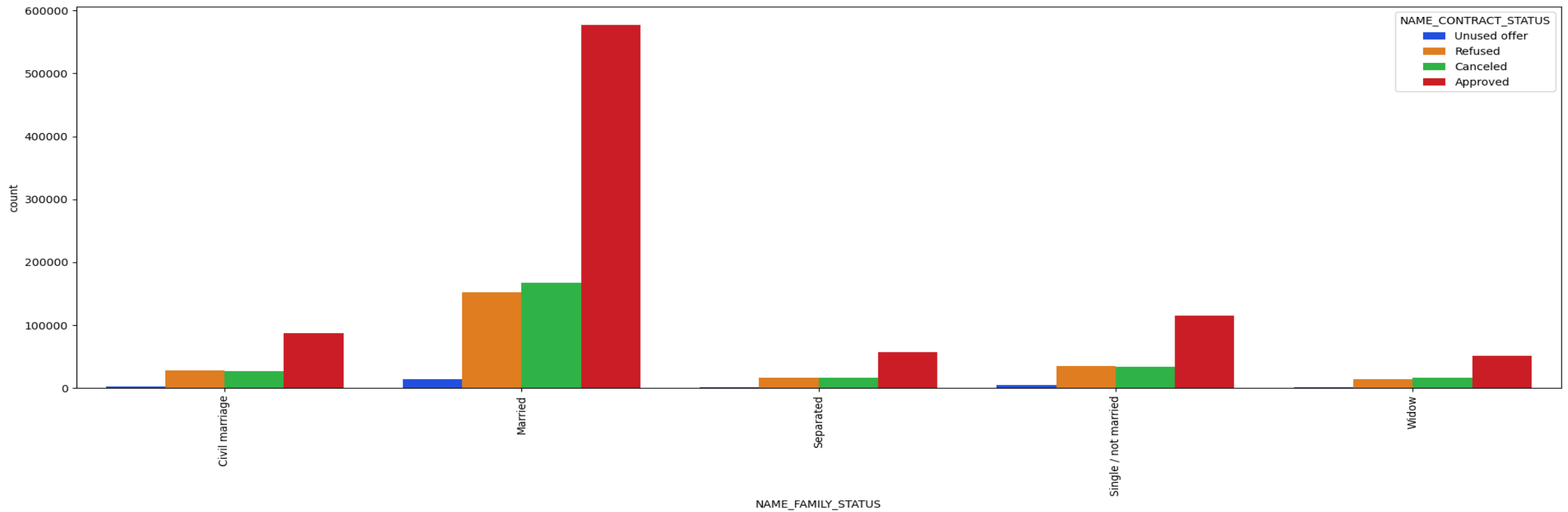
ANALYSIS OF "YEAR_BIRTH_CATEGORY" &"NAME_CONTRACT_STATUS"

- Clients who are in the age range 30-40 get most approval followed by clients in 40-50 age range
- Clients who are in the age range 60-70 receive least refusals followed by 20-30 age range



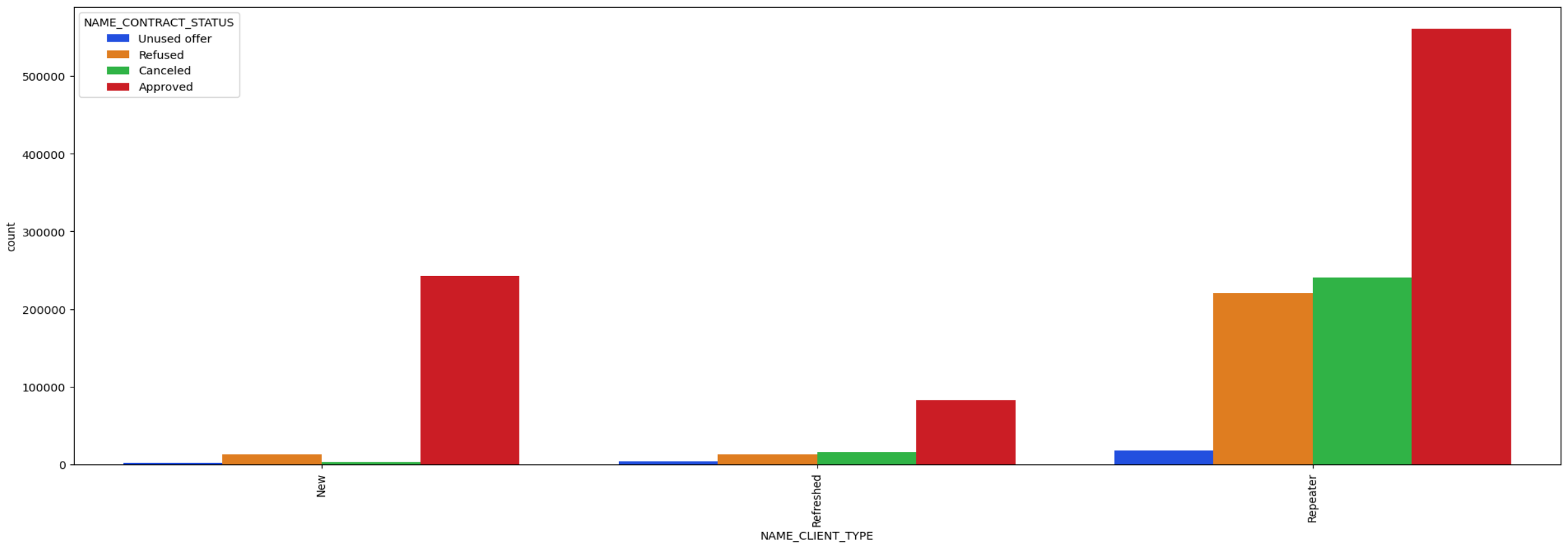
ANALYSIS OF "NAME_EDUCATION_TYPE" & "NAME_CONTRACT_STATUS"

- Clients who have Secondary/secondary special receive the most approvals



Analysis of 'NAME_FAMILY_STATUS' V/S 'NAME_CONTRACT_STATUS'

- Clients who are 'Married' receive the most approvals



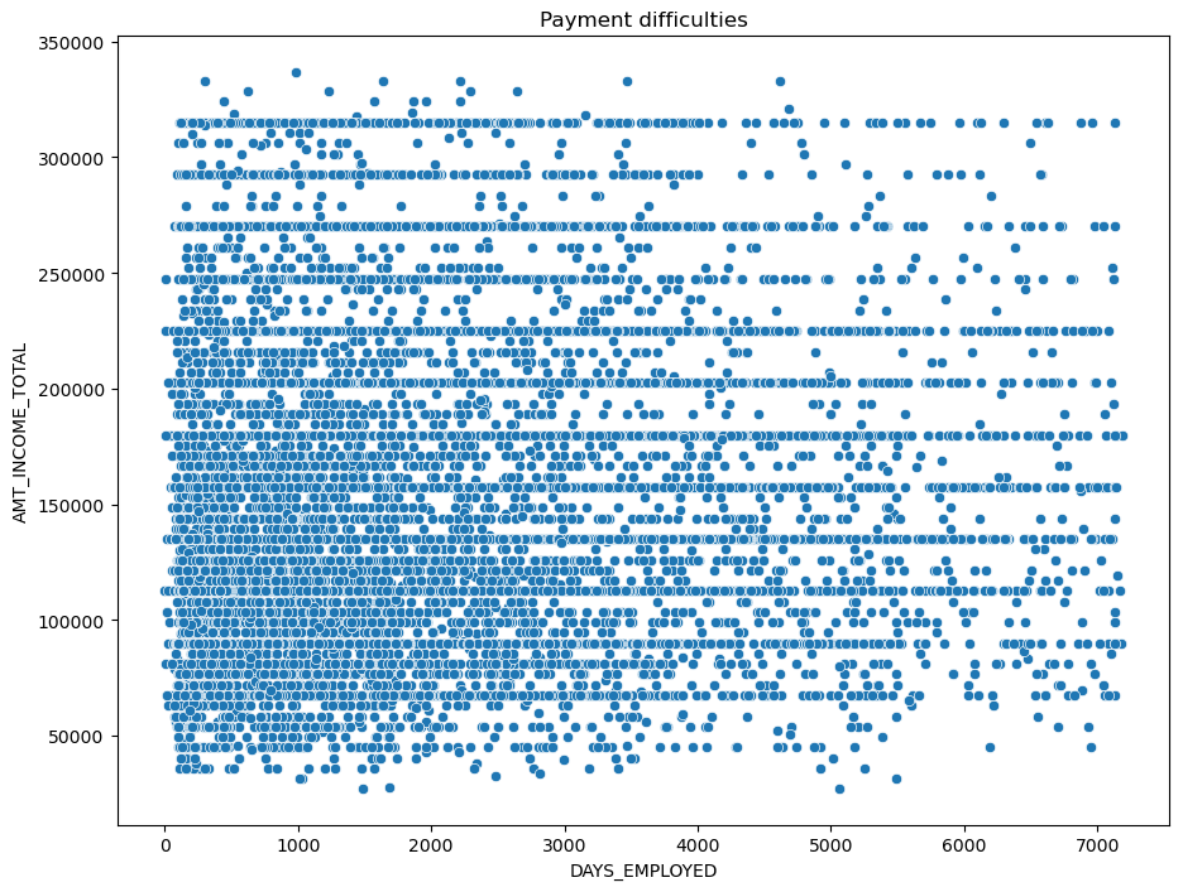
Analysis of 'NAME_CLIENT_TYPE' V/S 'NAME_CONTRACT_STATUS'

- Clients who are Repeaters receive the most approvals followed by New

conclusion:Clients we should target more

- Clients who are employed for more than 19 years
- Clients in the age range 30-40 and 40-50
- Clients who are Married
- Male clients with Academic degree
- Students and Businessman
- Repeater clients

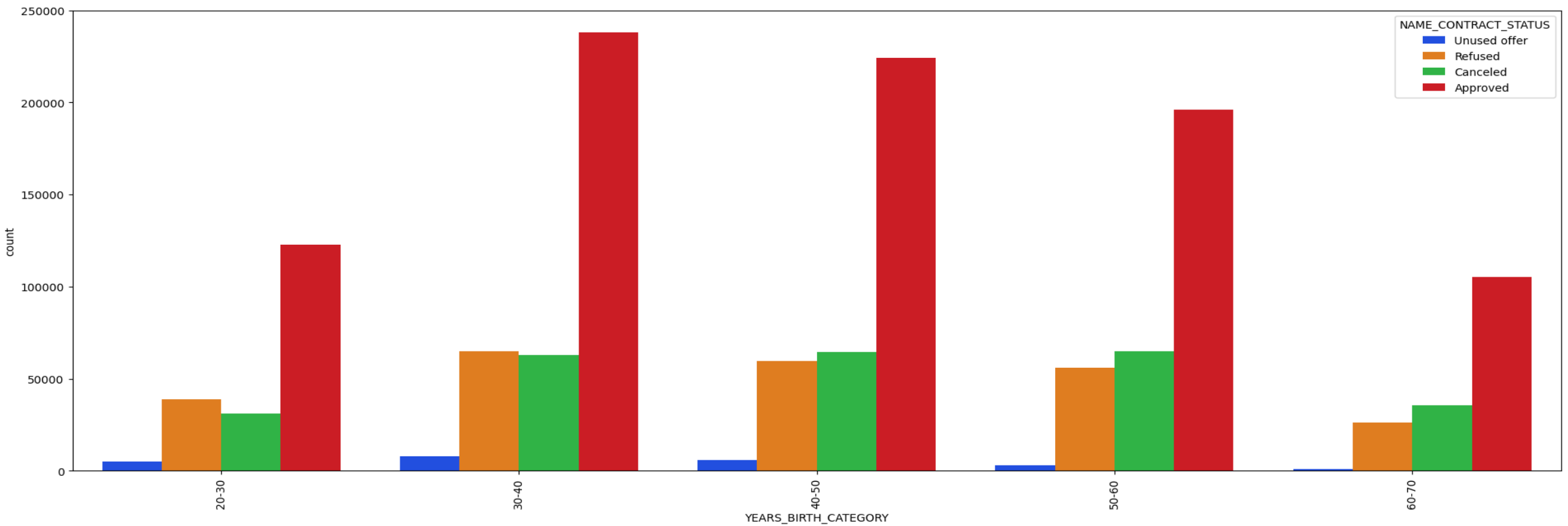
Reference slides used for arriving at conclusion



Analysis of "DAYS_EMPLOYED" V/S "AMT_INCOME_TOTAL"

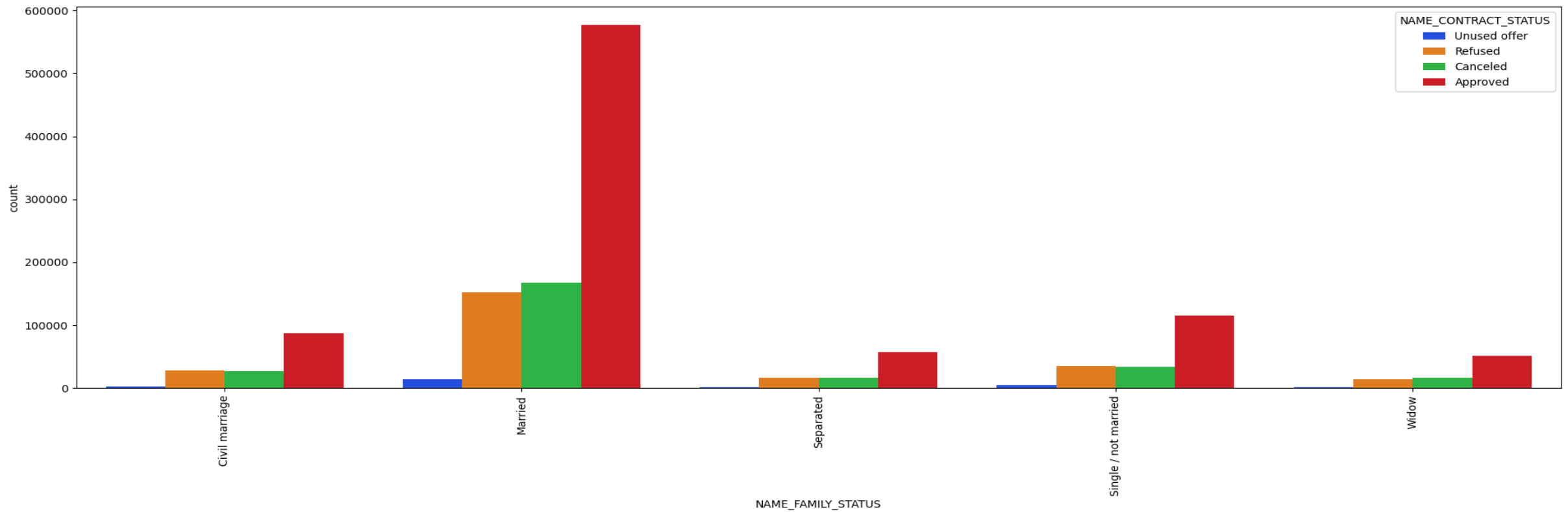
•- Clients who are employed for a long time (>7000) days or 19 years are making their payments on-time but these category of clients do not exist in Payments difficulties group.

•- Even looking at Payment difficulties group, clients with more than 4000 days of employment are sparse



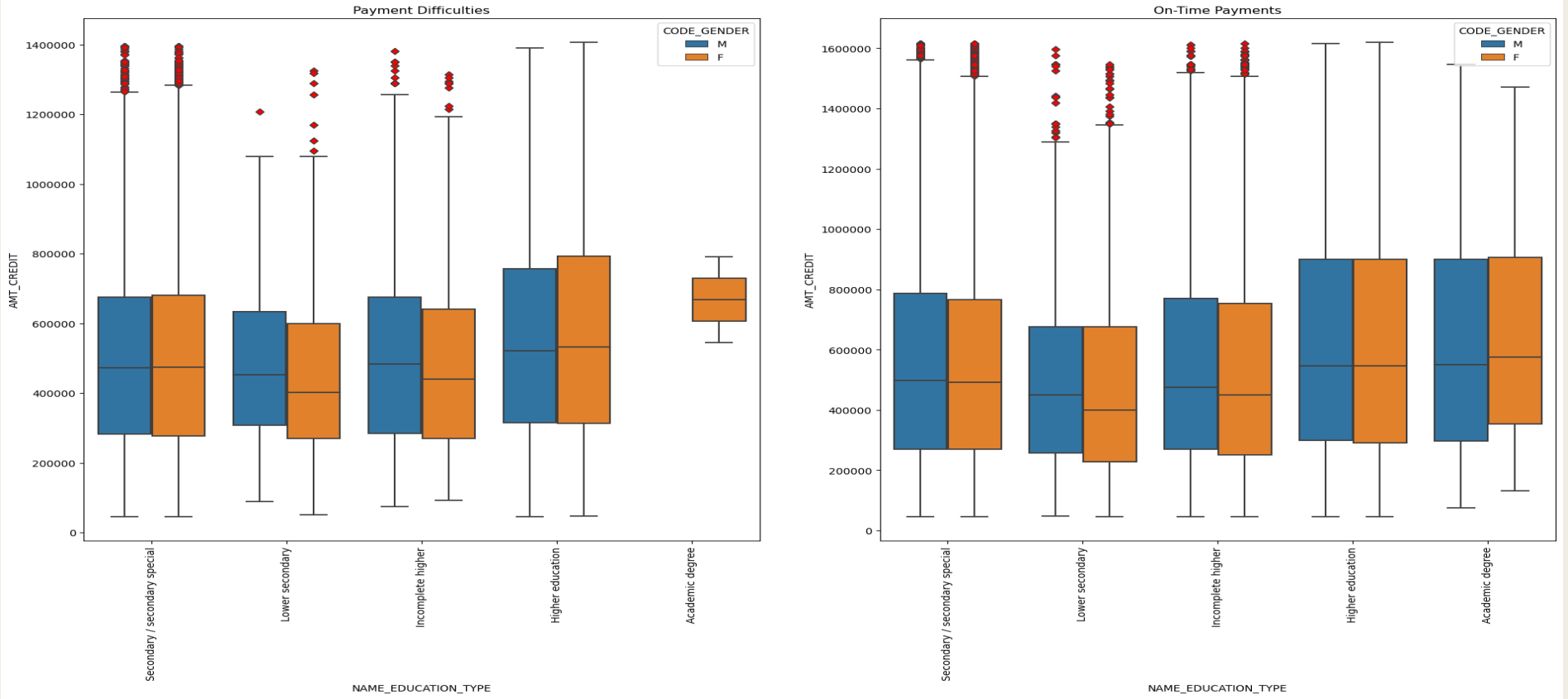
ANALYSIS OF "YEAR_BIRTH_CATEGORY" &"NAME_CONTRACT_STATUS"

- Clients who are in the age range 30-40 get most approval followed by clients in 40-50 age range
- Clients who are in the age range 60-70 receive least refusals followed by 20-30 age range



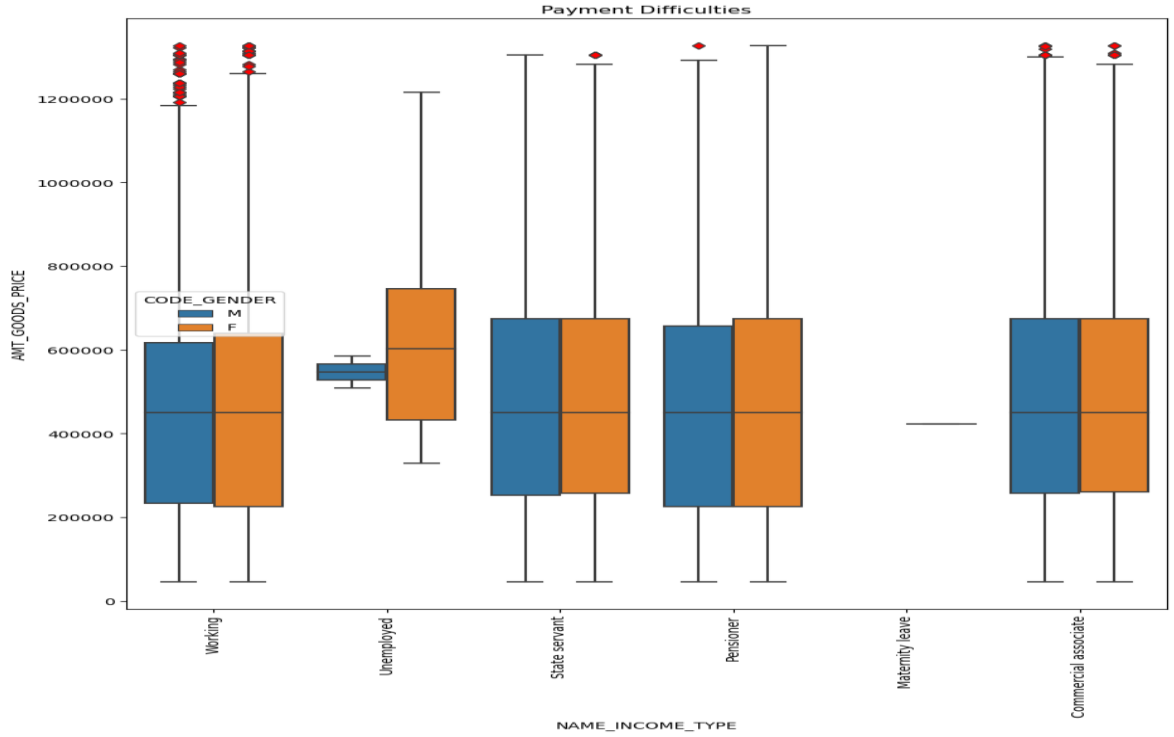
Analysis of 'NAME_FAMILY_STATUS' V/S 'NAME_CONTRACT_STATUS'

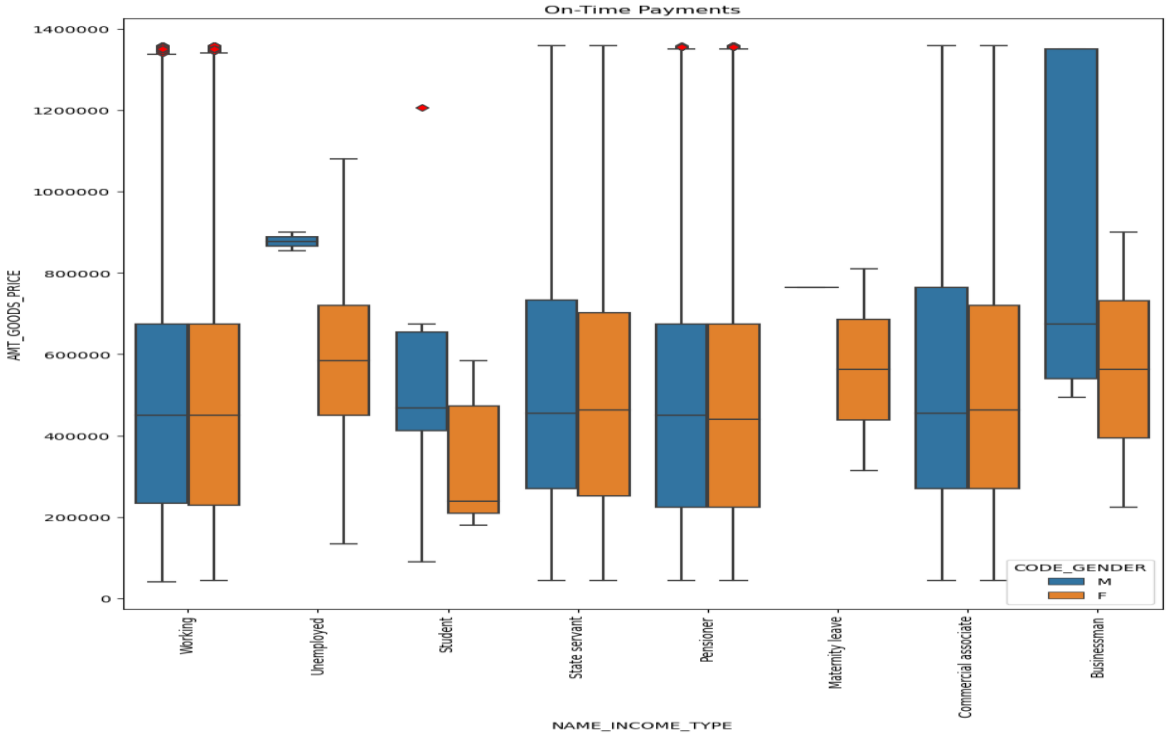
- Clients who are 'Married' receive the most approvals



Analysis of "NAME_EDUCATION_TYPE" V/S "AMT_CREDIT" V/S "CODE_GENDER"

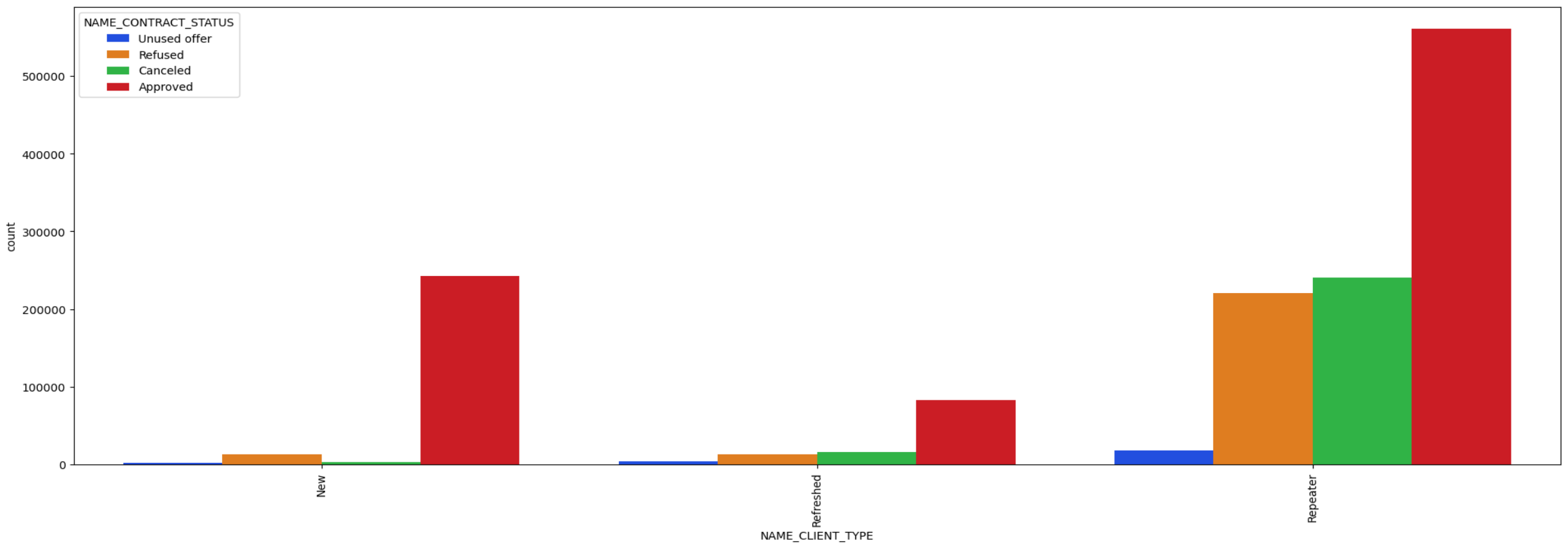
- Clients with `Academic Degree` have a wide range of credits for OnTime Payments whereas the range is much lower for ones with Payment difficulties
- Looking at summary statistics, Clients with `Academic Degree` and Payment difficulties take mean and median credit at a much higher range than On-Time Payment clients
- `Male` clients with `Academic Degree` always pay the loan on-time





ANALYSIS OF "NAME_INCOME_TYPE" &"AMT_GOODS_PRICE" & "CODE GENDER"

- - Clients who are Unemployed and Male have a very high price of goods in On-Time Payments than Payment difficulties
- - Clients who are Student and either Male OR Female do their payments On-Time. They are completely missing from Payment difficulties category. Student seems to be an attractive category to give loans to.
- - Clients who are Businessman and either Male OR Female do their payments On-Time. They are completely missing from Payment difficulties category. Businessman seems to be an attractive category to give loans to.



Analysis of 'NAME_CLIENT_TYPE' V/S 'NAME_CONTRACT_STATUS'

- Clients who are Repeaters receive the most approvals followed by New