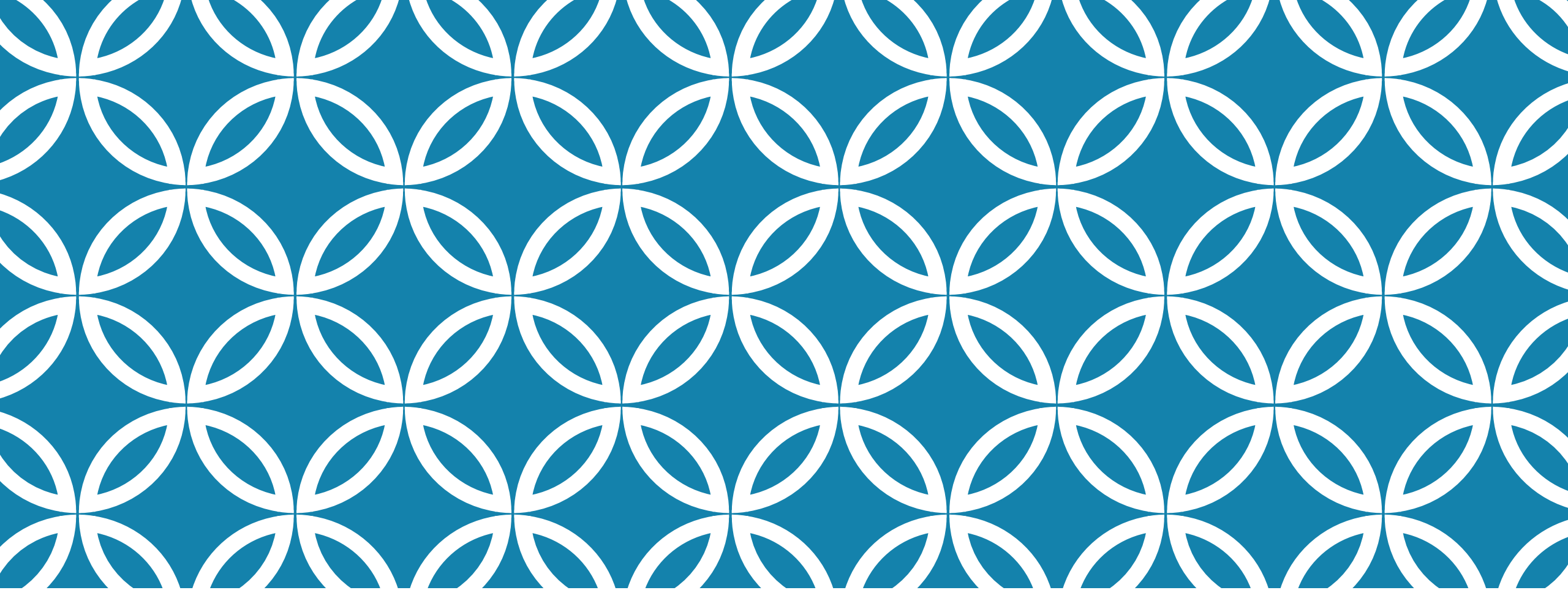


ASSIGNMENT-1 CREDIT EDA

By
Akshath K R

PROBLEM STATEMENT (BUSINESS UNDERSTANDING)

- Usually banking and finance companies make it difficult to provide loans to the people due to their insufficient or non-existent credit history.
- Most of the people, try to become defaulter after taking the loans.
- Here companies have to deal with two types of risk:
  - If the person has the capacity and company/bank not approving the loan will be a loss for the company/bank.
  - If the person is likely to be a defaulter/spammer and lending money to him, will again a loss for the company/bank.
- With the help of EDA analysis, we can find the patterns in the dataset which will ensure, if the clients are able to repay the loans aren't rejected.

BUSINESS UNDERSTANDING (DATASET)

- The dataset, contains two scenario:
  - Client with Payment difficulties (Target Variable with 1): If clients do late payments atleast once
  - Clients with on-time Payments (Target Variable with 0)
- There are 4types of decisions that Bank/Company provide to the client
  - Approved - Bank/Company has accepted the loan application.
  - Cancelled - Client has canceled his loan application.
  - Refused - Bank/Company has reject the loan application.
  - Unused offer - Client has canceled his loan application, during its process.
- Using EDA, lets understand how different attributes influence the tendency to either default or on-time payments

BUSINESS UNDERSTANDING (DATASET)

- There are 3 datasets,
  1. *'application_data.csv'*: This data is all about whether a **client has payment difficulties.**
  2. *'previous_application.csv'*: This data contains whether the previous application of the client had been **Approved, Cancelled, Refused or Unused offer.**
  3. *'columns_description.csv'*: This is the information dataset, that provide the detailed explanation of the attributes/columns in the dataset

BUSINESS OBJECTIVE

- Aim of the study is to determine the patters in the dataset that help with, if the client has difficulties paying loan/installments, by which company can take futher actiosn such as rejecting his/her application, slicing the loan amount, providing to risky applicants with huge interest rates.
- Also to ensure the clients who are capable to repay the loans are not rejected.
- Defining all the attributes and variable that are strong indicators that drive these factors.

OVERVIEW OF ALL THE STEPS

1. Importing the Dataset(*application_data.csv*)
2. *Exploring the Data*
3. *Checking for Null Values*
4. *Dropping columns with more Null Values*
5. *Split Numerical and categorical*
6. *Replacing Incorrect/Missing Values*
7. *Checking Outliers*
8. *Check Data Imbalancing*
9. *Univariate,* Segmented Univariate, *Bivariate and Multivariate analysis of each Target value*
11. Importing the Dataset (*previous_application.csv*)
12. *Repeated the same steps from 2 to 7*
13. *Merged the Datasets*
14. *Confirming the Data Cleaning process*
15. *Univariate,* Segmented Univariate, *Bivariate and Multivariate analysis over the merged data*
16. *Conclusions*

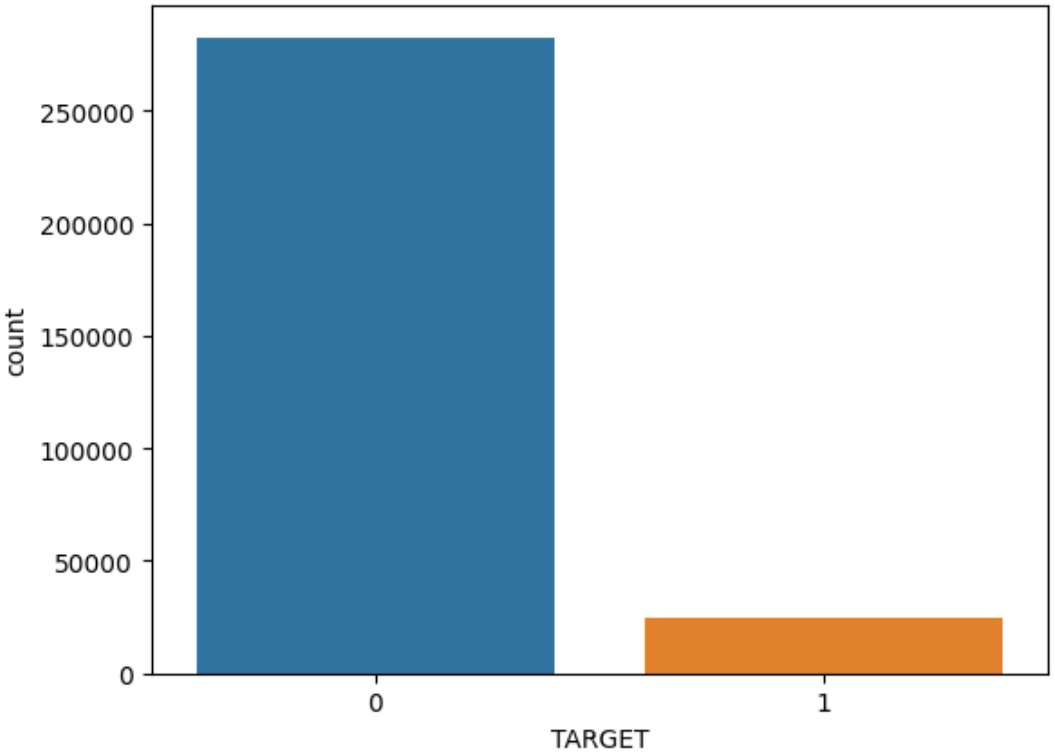
DATA CLEANING PROCESS

1. After **Importing** the dataset(*application_data.csv*), I have explored the data and understood all the attributes. Initially the shape of the dataset was **307511 rows** and **122 columns.**
2. Followed by, Checked for null values, **Null Values** more then **40%** are dropped assuming that they provide zero help to gather insights. Almost **49 columns** was dropped.
3. Later, I got the columns divided into **categorical** and **numerical** columns and ran **for** loop to check the statistics of the same
4. Then checked for **Missing** and **Incorrect** values, and imputed them with appropriate values.
5. Checked the **Outliers**(**Using Boxplot**) and defined **Quantiles** over various ranges to properly impute the **NaN** values and defined the **Lower** and **Upper** fences of the plot.
6. By defining the Lower and Upper fences it was more clear to easily detect the outliers.
7. Few columns that recorded the data in days, was converted to unit Years. This is more convenient to understand the insights better.

Note: All the above steps are detailed out with exact observations in the notebook

DATA IMBALANCE

1. After data cleaning, as per the business objective I had to split he data frame into two based on the target column. One with “Inp_3_TR0” that represent data frame with On-time payments and data frame “Inp_3_TR1” represent clients with payment difficulties.



Target	Value	Percentage(%)
0s (On-Time Payments)	282686	91.92
1s (Payments Difficulties)	24825	8.07

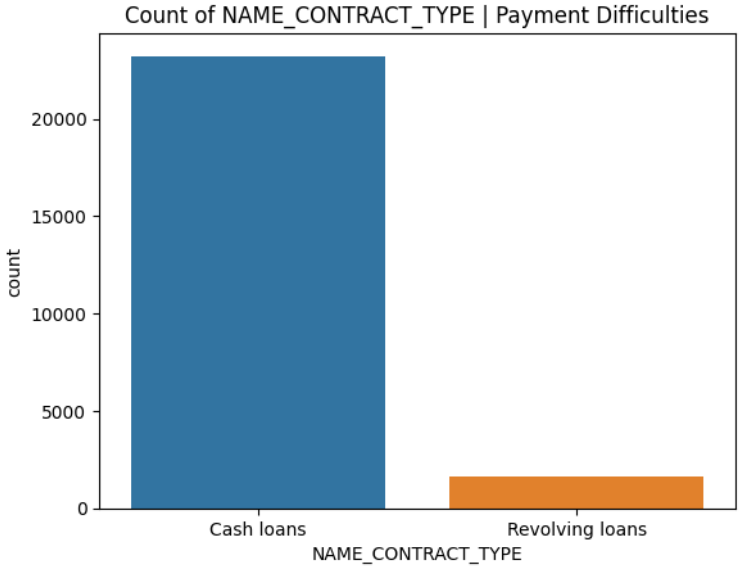
1. We have the count plot based on the total observations of 0s and 1s

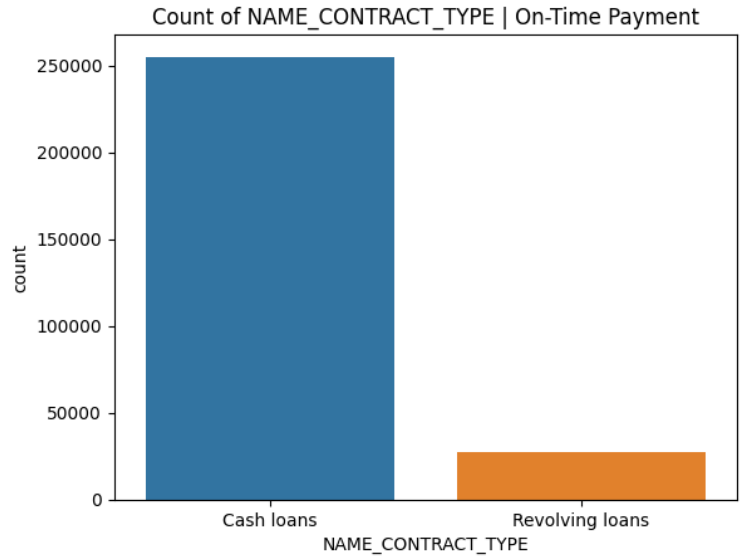
2. Almost 282686 (~91.9%) are 0s i.e. these many clients make on-time payments.

3. Similarly, Almost 24825(~8.1%) are 1s i.e. these many clients have payment difficulties

UNIVARIATE ANALYSIS (CATEGORICAL VARIABLE)

Univariate Analysis on NAME_CONTRACT_TYPE: Identification if loan is cash or revolving



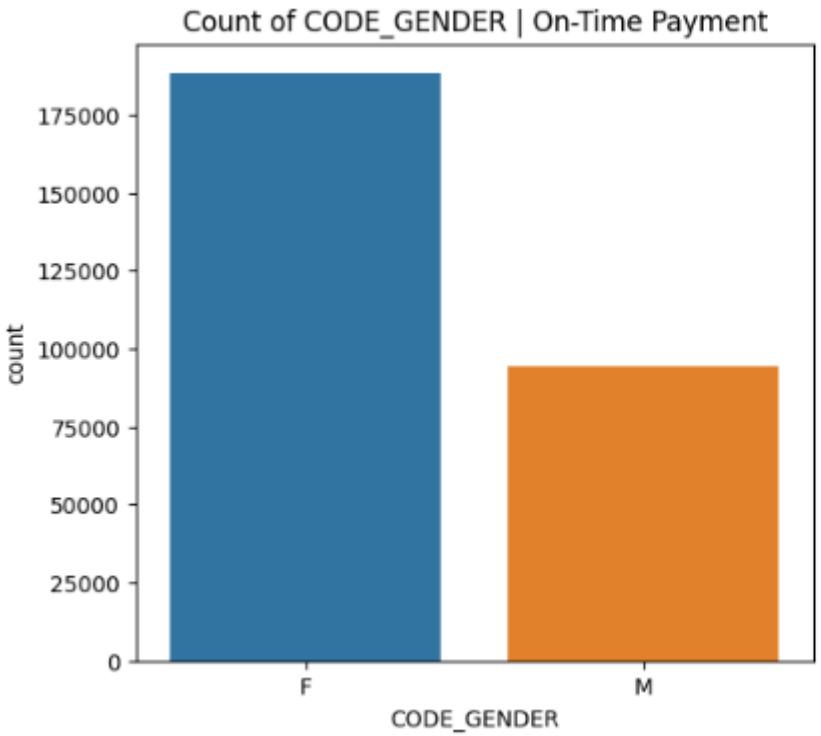


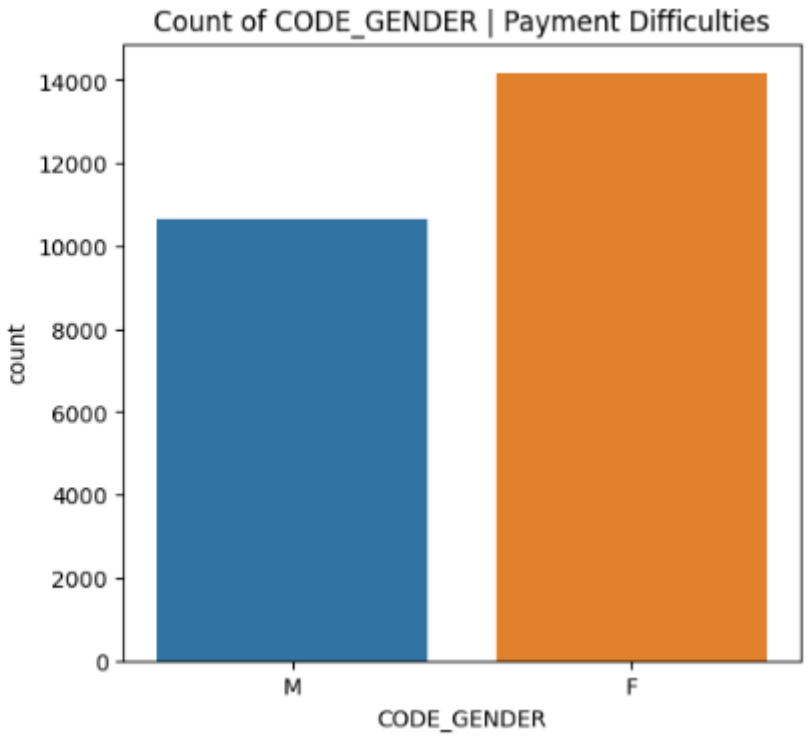
	NAME_CONTRACT_TYPE	Value	Percentage(%)
Payment Difficulties	Cash Loans	23221	93.53
	Revolving loans	1604	6.46
On-Time Payments	Cash Loans	255011	90.20
	Revolving loans	27675	9.8

Most of the clients in both the scenarios have applied for Cash loans.

UNIVARIATE ANALYSIS (CATEGORICAL VARIABLE)

Univariate Analysis on CODE_GENDER: Gender of the client



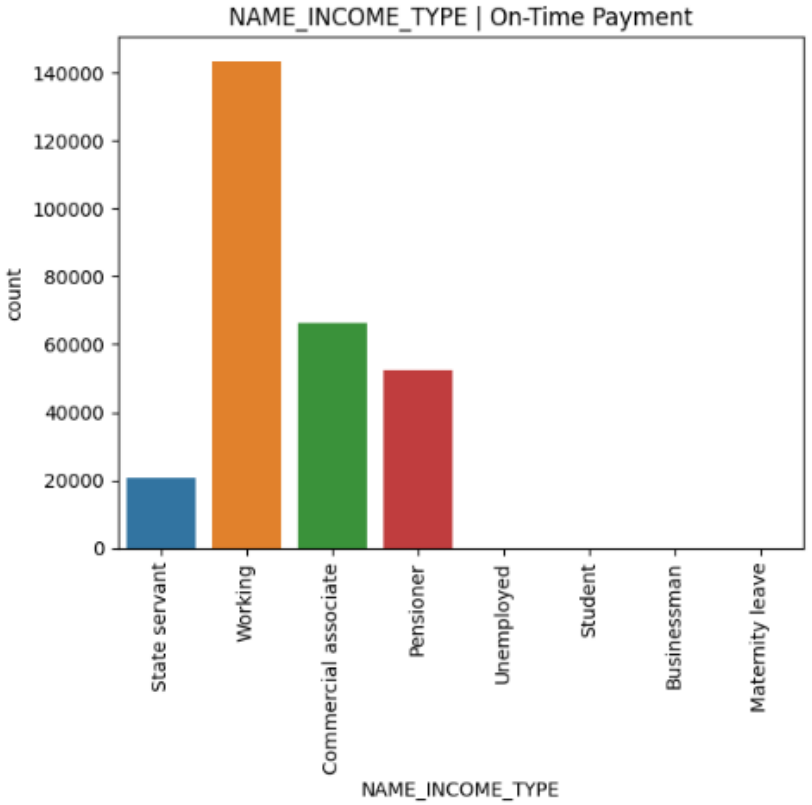


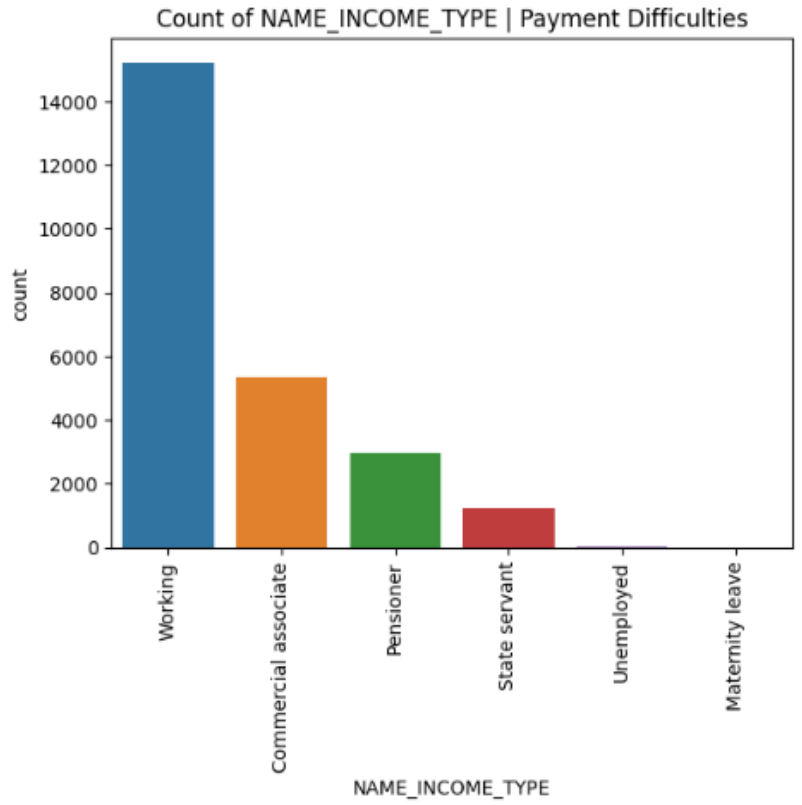
By looking at the graph and the statistics, we can clearly see that females are ~9.5% more efficient in payments, but men is ~9.5% more in payment difficulties

Male have more difficulties in payments

UNIVARIATE ANALYSIS (CATEGORICAL VARIABLE)

Univariate Analysis on NAME_INCOME_TYPE: Clients income type (businessman, working, maternity leave,…)

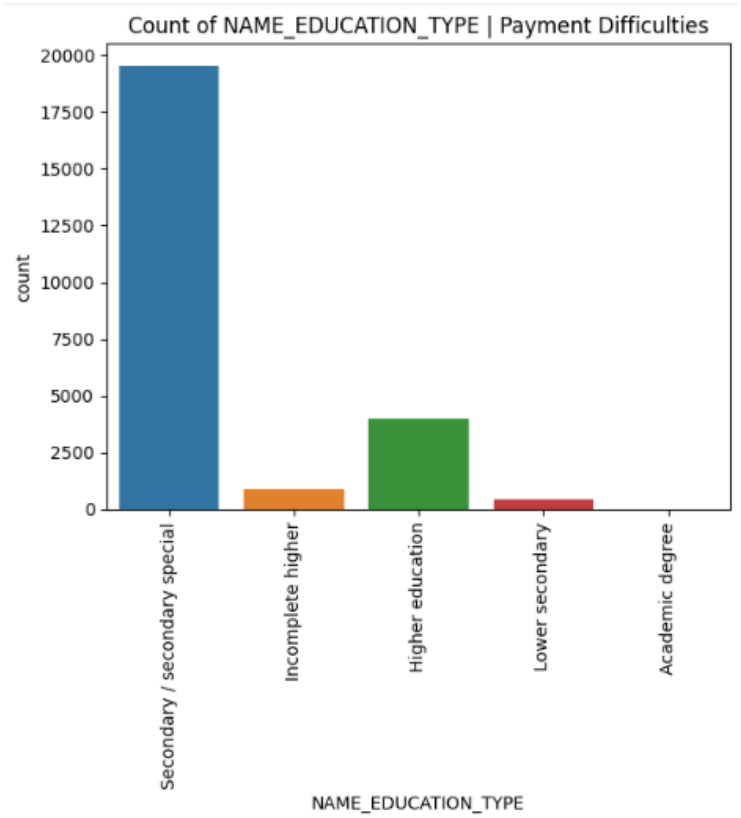


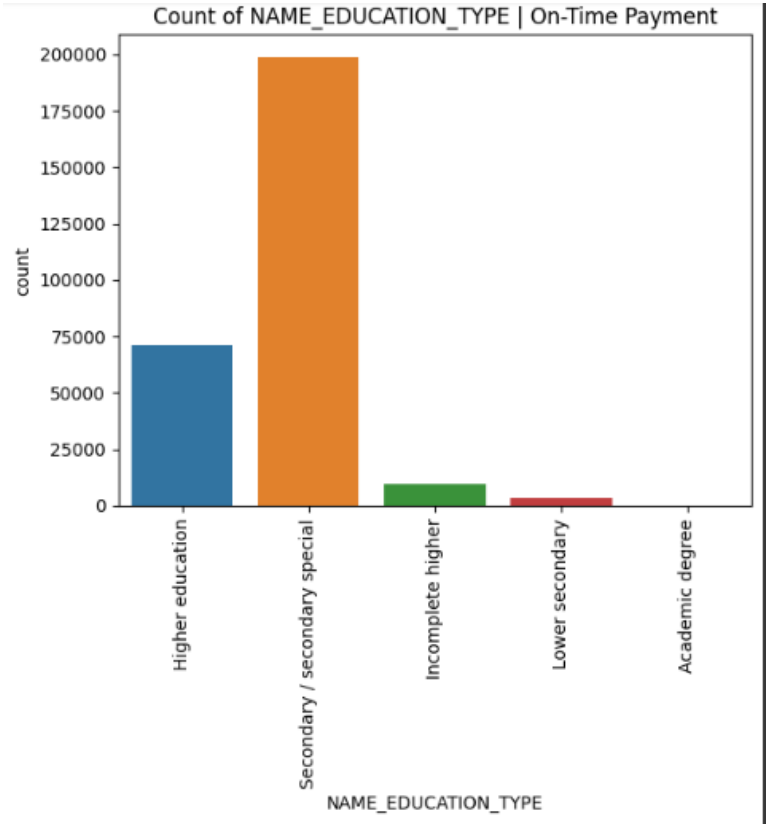


1. Students and Business man have no payment difficulties
2. Commercial Associates have better on-time payments
3. Pensioners are quite better in making payments

UNIVARIATE ANALYSIS (CATEGORICAL VARIABLE)

Univariate Analysis on NAME_EDUCATION_TYPE: Level of highest education the client achieved





UNIVARIATE ANALYSIS (CATEGORICAL VARIABLE)

Univariate Analysis on NAME_EDUCATION_TYPE: Level of highest education the client achieved

	NAME_CONTRACT_TYPE	Value	Percentage(%)
Payment Difficulties	Secondary / secondary special	19524	78.6
	Higher education	4009	16.1
	Incomplete higher	872	3.5
	Lower secondary	417	1.6
	Academic degree	3	0.01

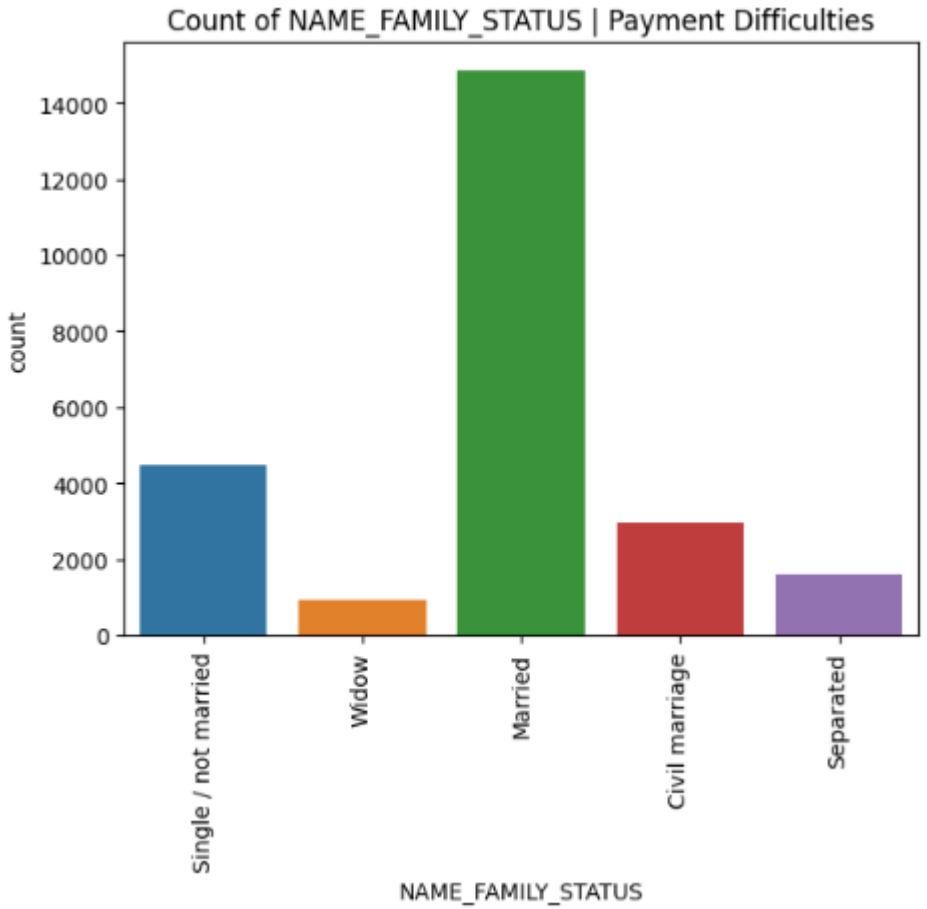
	NAME_CONTRACT_TYPE	Value	Percentage(%)
On-Time Payments	Secondary/secondary special	198867	70.34
	Higher education	70854	25.0
	Incomplete higher	9405	3.3
	Lower secondary	3399	1.2
	Academic degree	161	0.05

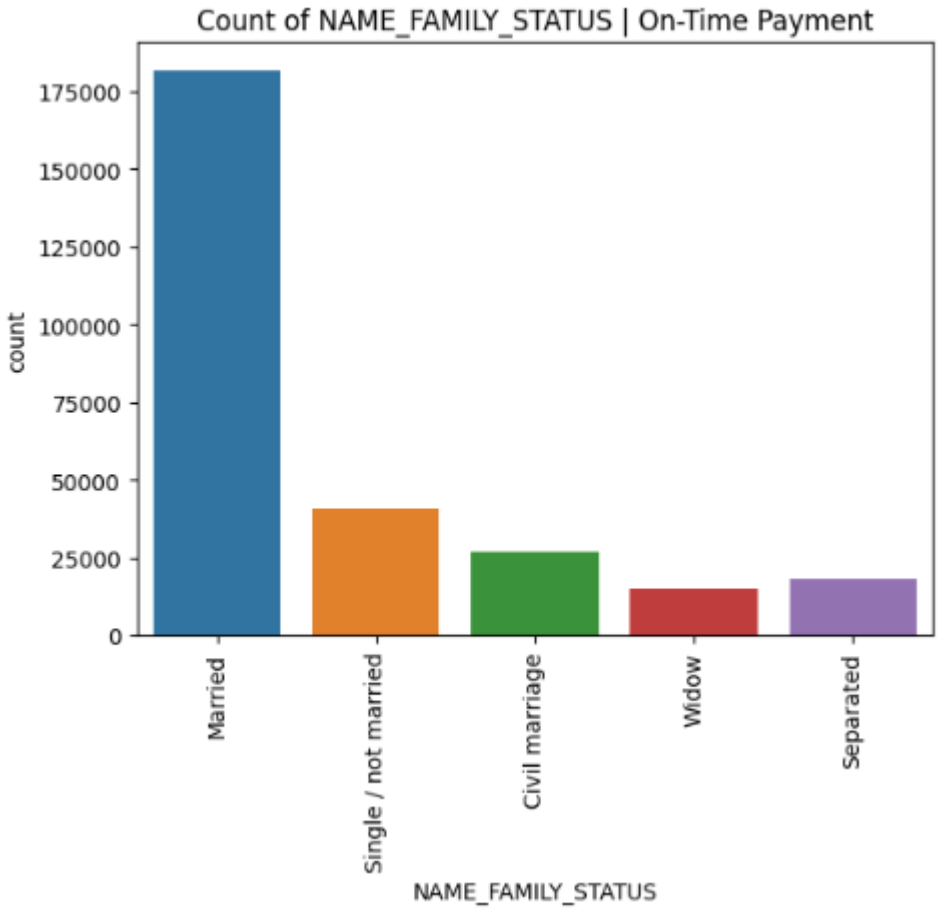
1. We can clearly see from the graphs and the statistics that academic degree guys are quite efficient in payment

2. Clients with higher education population is ~10% more in doing ontime payments

UNIVARIATE ANALYSIS (CATEGORICAL VARIABLE)

Univariate Analysis on NAME_FAMILY_STATUS: Family status of the client





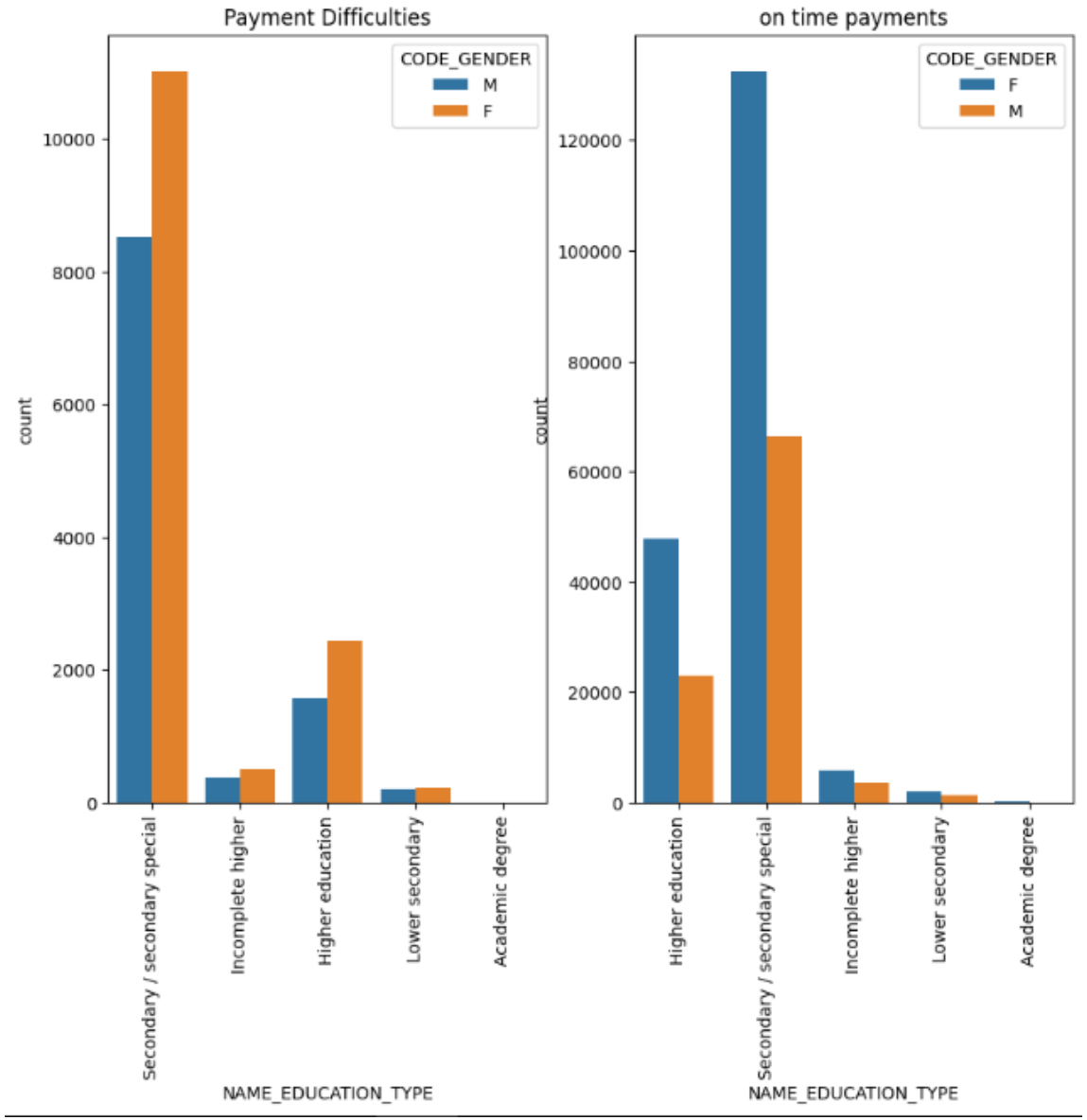
1. It is clear that approximately married clients are 64.2% in making the clear payments and 58.9% clients have payment difficulties

2. 5.35% of Widow are making clear payments, whereas 3.77% have payment difficulties

3. ~15% Singles make payments, ~18% have difficulties

UNIVARIATE ANALYSIS (CATEGORICAL VARIABLE)

Univariate Analysis on NAME_FAMILY_STATUS and CODE_GENDER

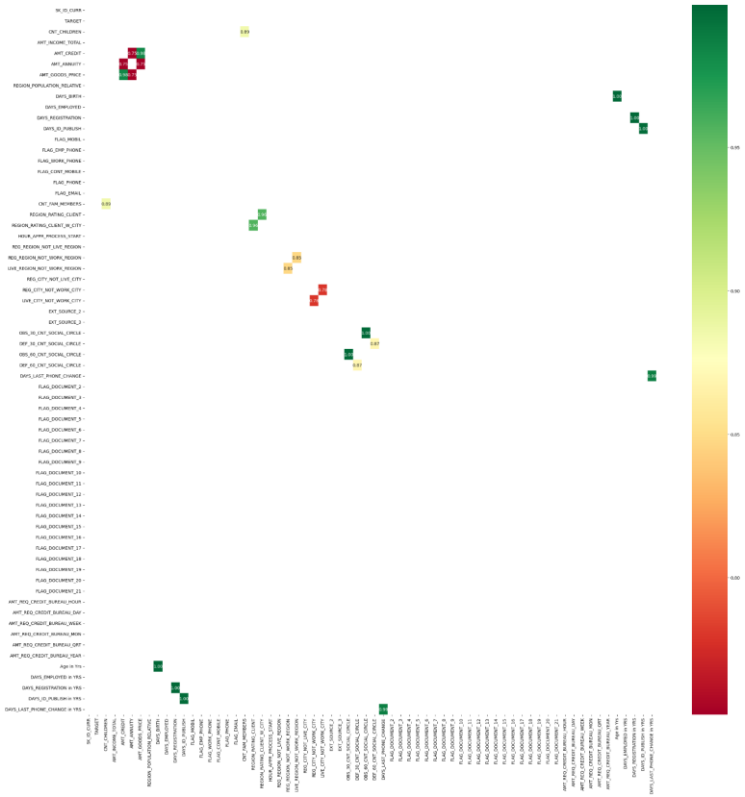


From the plot it is clear that female clients with Higher education and secondary education have more tendency to clear loans

CORRELATION ANALYSIS

Correlation for all the numerical column was found and visualized using Heatmap and Top 10 correlations were extracted

Payment Difficulties



DAYS_BIRTH	Age in Yrs	0.999691
Age in Yrs	DAYS_BIRTH	0.999691
DAYS_REGISTRATION in YRS	DAYS_REGISTRATION	0.999479
DAYS_REGISTRATION	DAYS_REGISTRATION in YRS	0.999479
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998270
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998270
DAYS_ID_PUBLISH	DAYS_ID_PUBLISH in YRS	0.997531
DAYS_ID_PUBLISH in YRS	DAYS_ID_PUBLISH	0.997531
DAYS_LAST_PHONE_CHANGE in YRS	DAYS_LAST_PHONE_CHANGE	0.988086
DAYS_LAST_PHONE_CHANGE	DAYS_LAST_PHONE_CHANGE in YRS	0.988086
AMT_GOODS_PRICE	AMT_CREDIT	0.982783
AMT_CREDIT	AMT_GOODS_PRICE	0.982783
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.956637
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637
CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
CNT_CHILDREN	CNT_FAM_MEMBERS	0.885484
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.869016
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.869016
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.847885
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778540
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.778540
AMT_ANNUITY	AMT_GOODS_PRICE	0.752295
AMT_GOODS_PRICE	AMT_ANNUITY	0.752295
AMT_CREDIT	AMT_ANNUITY	0.752195

dtype: float64

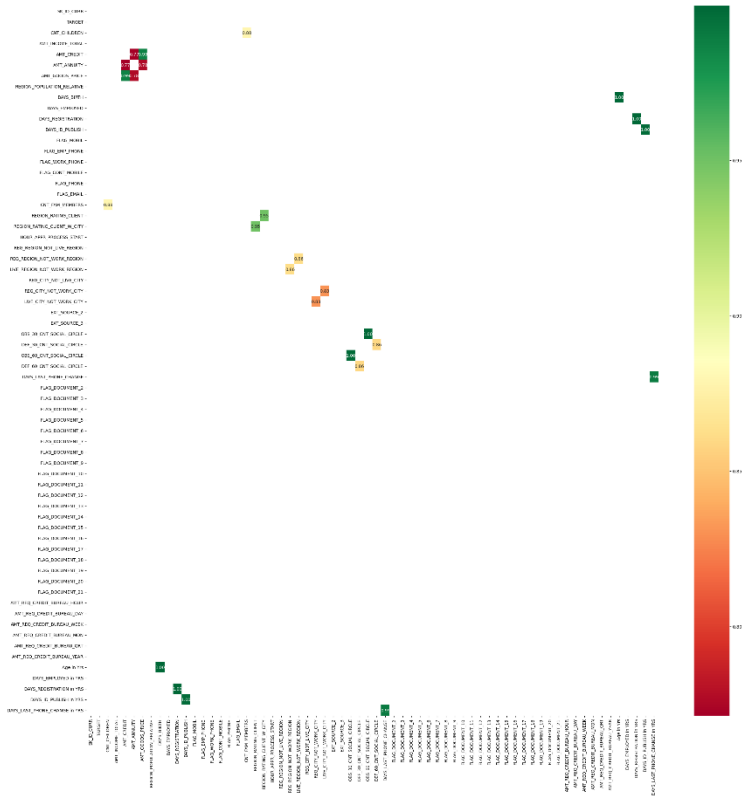
Enlarge the Image to view clearly

Note: There are some duplicates, but still the data is crystal clear

CORRELATION ANALYSIS

Correlation for all the numerical column was found and visualized using Heatmap and Top 10 correlations were extracted

On-Time Payments



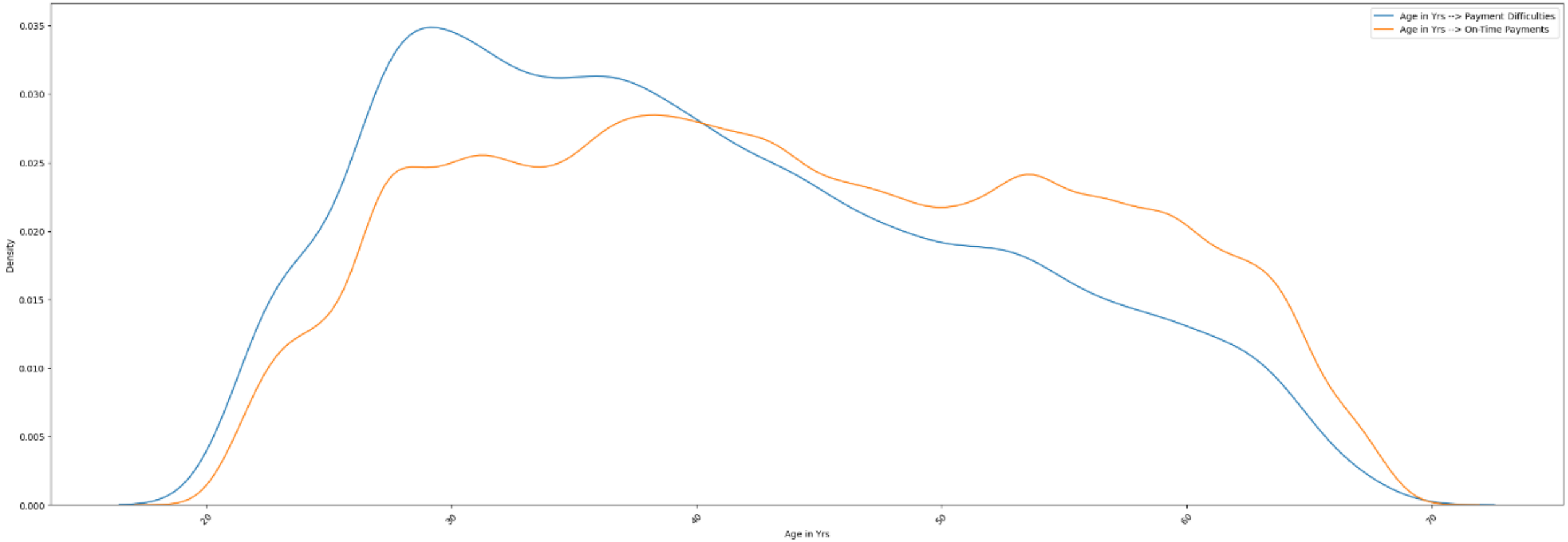
Enlarge the Image to view clearly

DAYS_BIRTH	Age in Yrs	0.999711
Age in Yrs	DAYS_BIRTH	0.999711
DAYS_REGISTRATION in YRS	DAYS_REGISTRATION	0.999554
DAYS_REGISTRATION	DAYS_REGISTRATION in YRS	0.999554
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998510
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998510
DAYS_ID_PUBLISH in YRS	DAYS_ID_PUBLISH	0.997518
DAYS_ID_PUBLISH	DAYS_ID_PUBLISH in YRS	0.997518
DAYS_LAST_PHONE_CHANGE	DAYS_LAST_PHONE_CHANGE in YRS	0.990258
DAYS_LAST_PHONE_CHANGE in YRS	DAYS_LAST_PHONE_CHANGE	0.990258
AMT_CREDIT	AMT_GOODS_PRICE	0.987022
AMT_GOODS_PRICE	AMT_CREDIT	0.987022
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950149
CNT_CHILDREN	CNT_FAM_MEMBERS	0.878571
CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.861861
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859371
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.859371
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.830381
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830381
AMT_ANNUITY	AMT_GOODS_PRICE	0.776371
AMT_GOODS_PRICE	AMT_ANNUITY	0.776371
AMT_ANNUITY	AMT_CREDIT	0.771248

Note: There are some duplicates, but still the data is crystal clear

UNIVARIATE ANALYSIS (NUMERICAL VARIABLE)

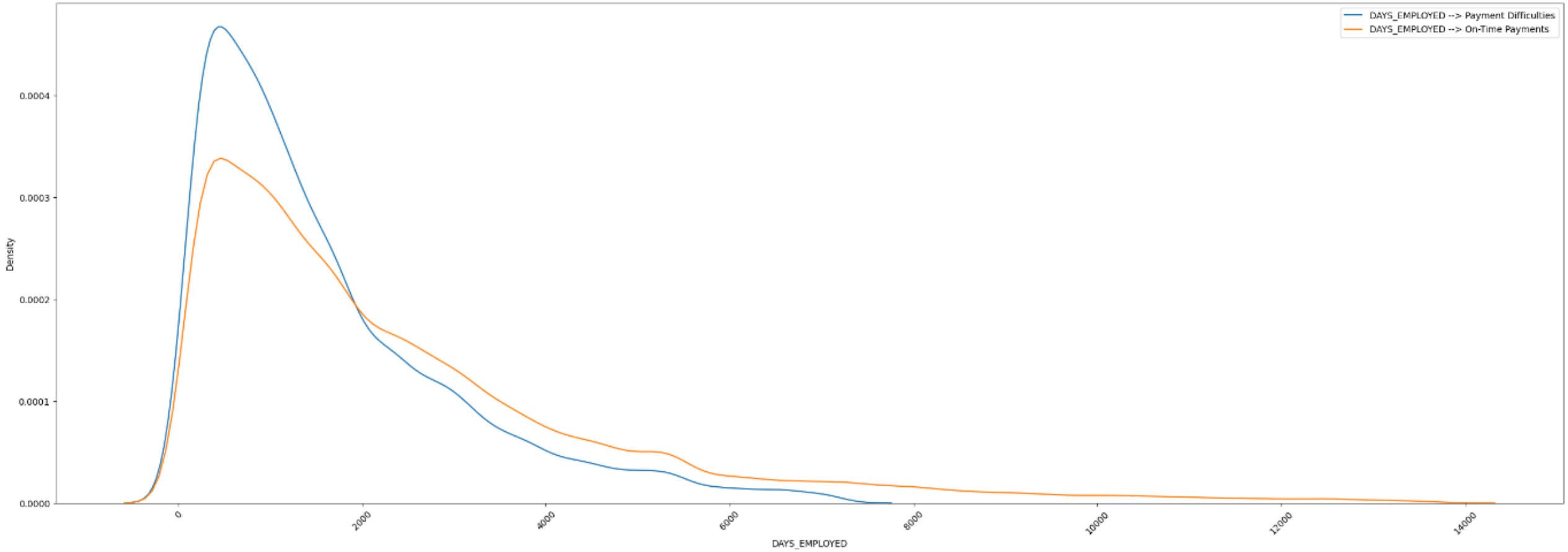
Age in Yrs



From the graph we can easily note that, people aged between 21-45 have more payment difficulties
In contrast, people aged >40 tend to have done more on time payments

UNIVARIATE ANALYSIS (NUMERICAL VARIABLE)

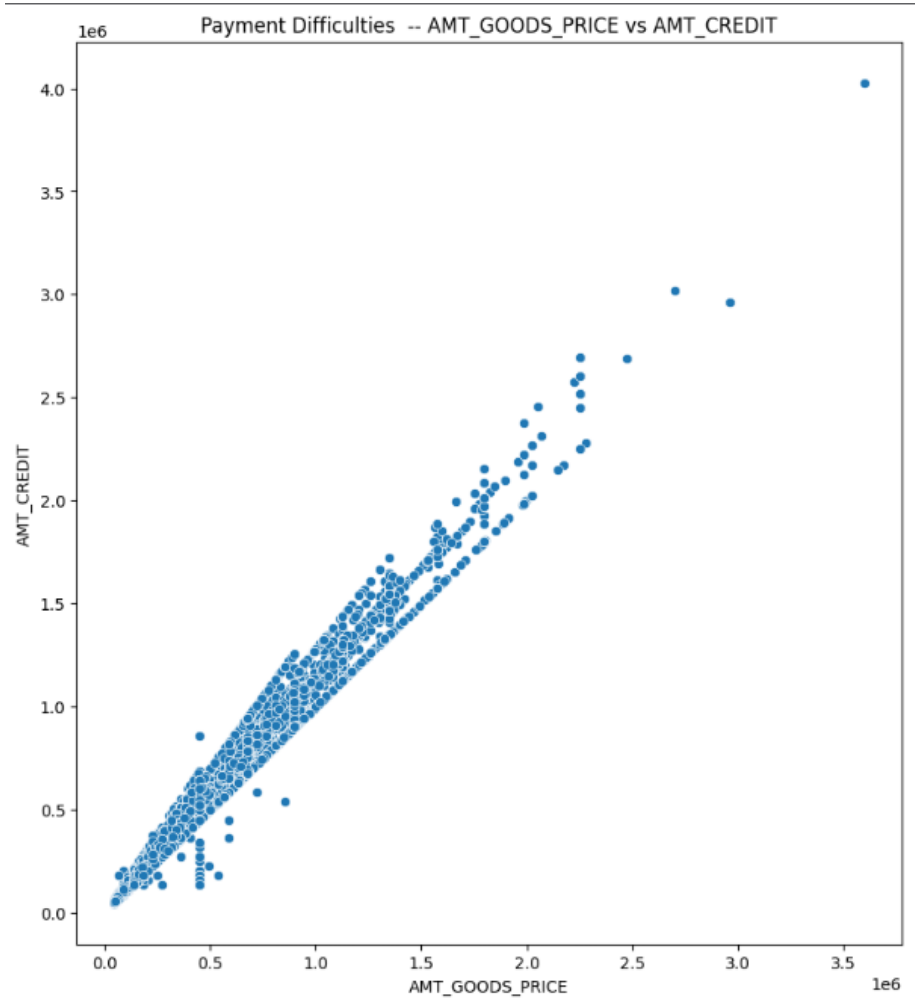
DAYS_EMPLOYED: How many days before the application the person started current employment

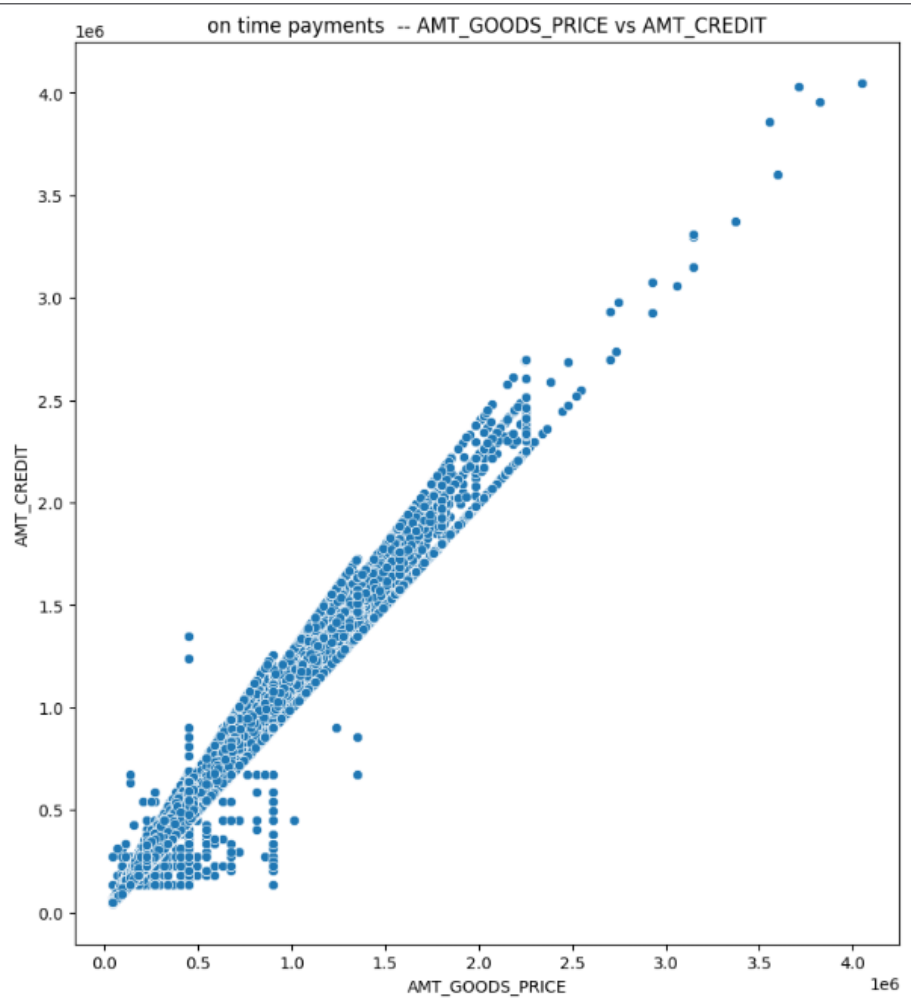


Therefore, a person who started wthin 2500 days and got loan has more payment difficulties
In contrast, a person with more then 2500 days experience have done on time payments

BIVARIATE ANALYSIS (NUMERICAL VARIABLE)

AMT_CREDIT and AMT_GOODS_PRICE

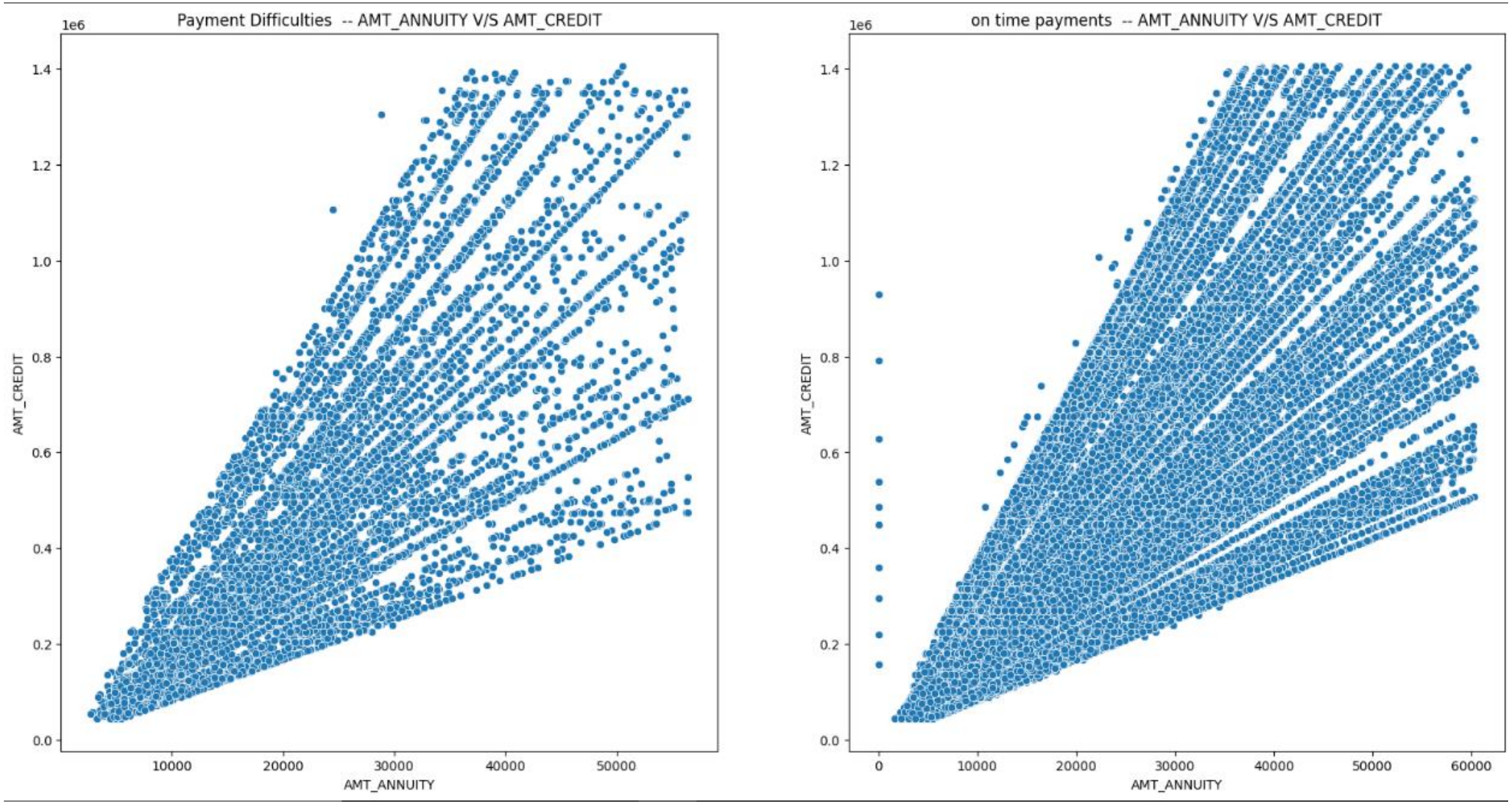




We can clearly see that there is a strong correlation, i.e. increase in AMT_GOODS_PRICE increase AMT_CREDIT. In other words they both are directly proportional.

BIVARIATE ANALYSIS (NUMERICAL VARIABLE)

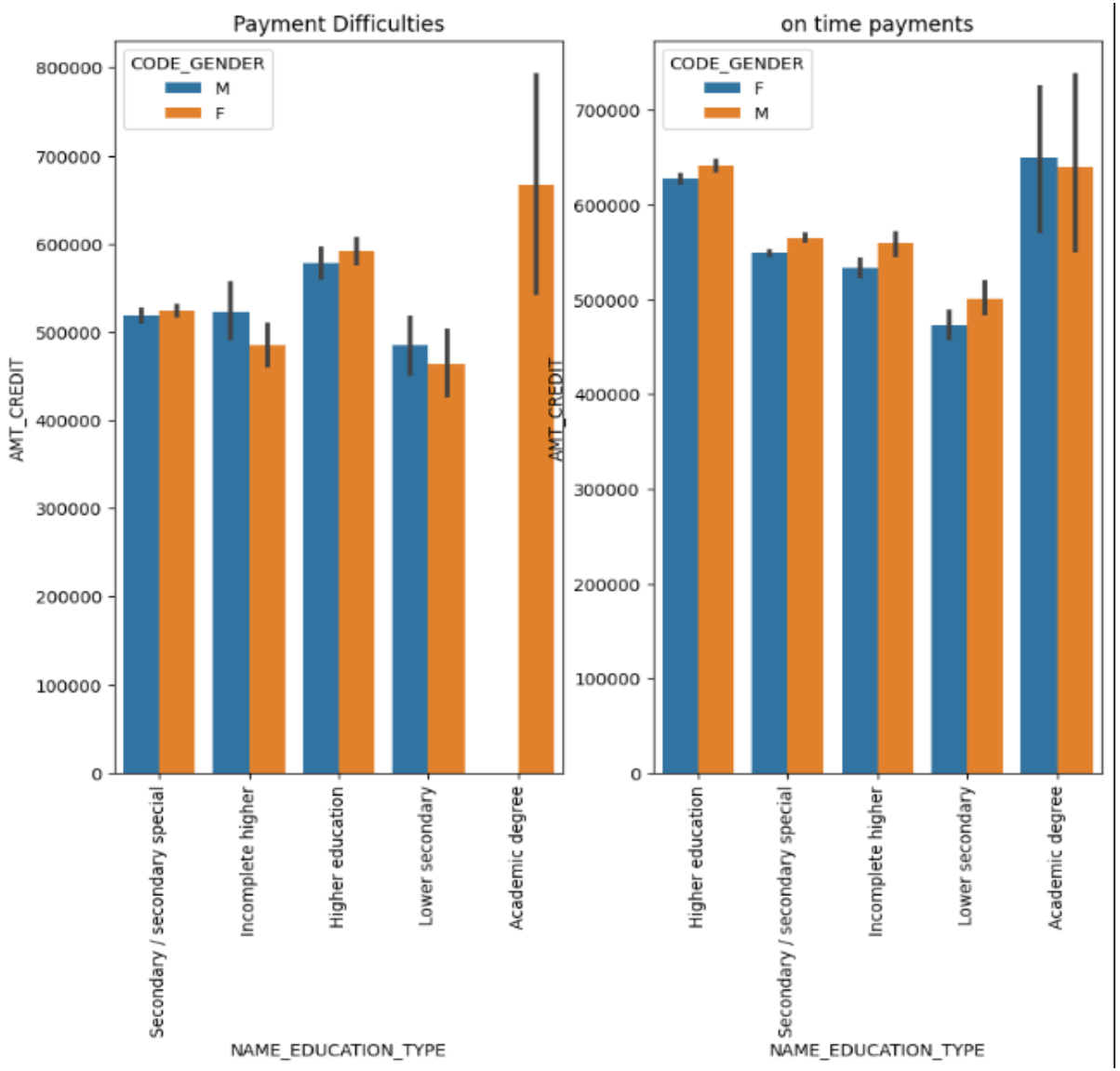
AMT_ANNUITY and AMT_CREDIT



We can clearly see that there is a strong correlation, i.e. increase in AMT_ANNUITY increase AMT_CREDIT. In other words they both are directly proportional.

BIVARIATE ANALYSIS (NUMERICAL VARIABLE)

NAME_EDUCATION_TYPE vs AMT_CREDIT vs CODE_GENDER

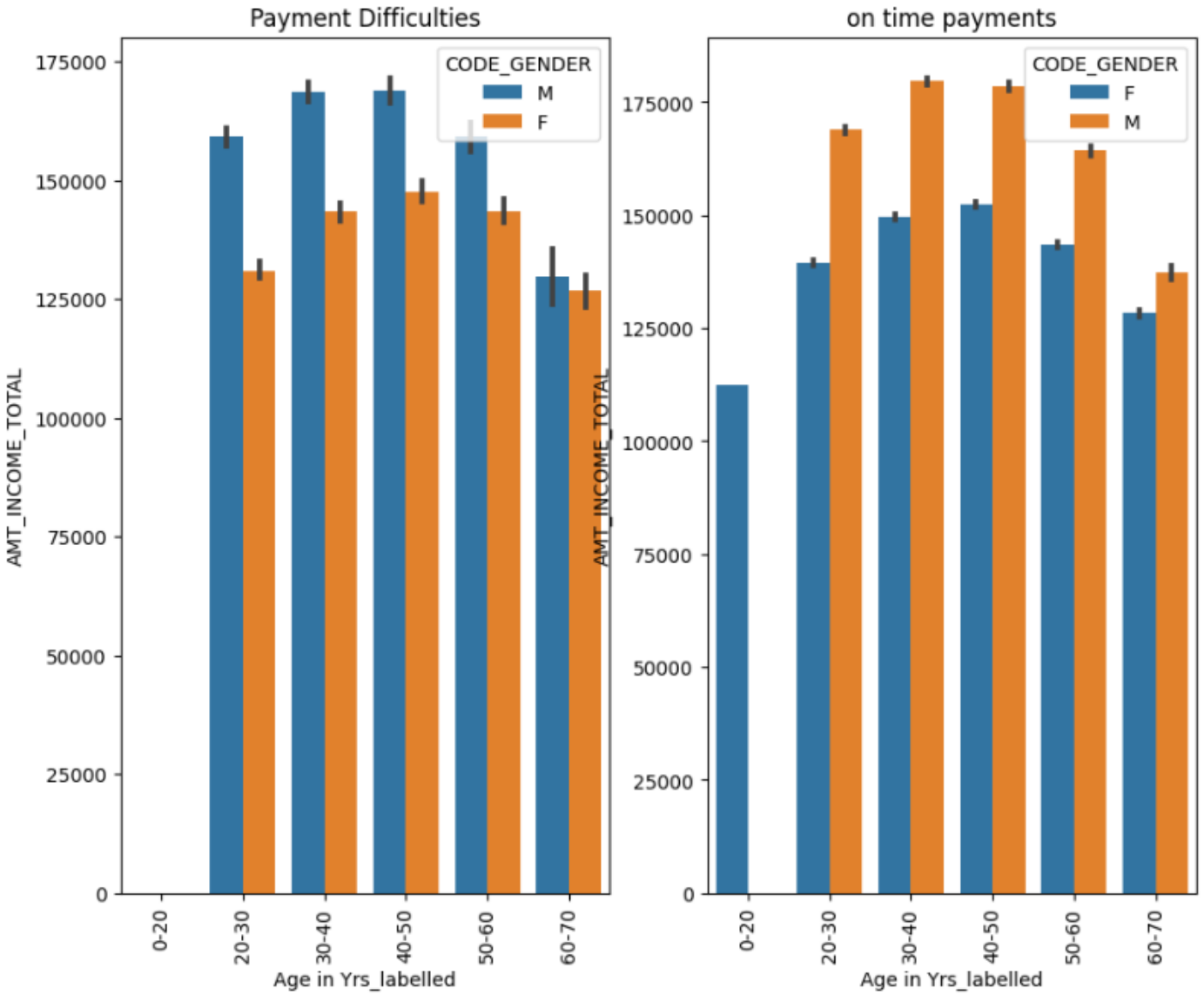


Here, its clear that female clients with Academic degree got more points on AMT_Credit, as well as in our previously analysis Female clients found more efficient in on-time payments

Again, Males and females with higher education got more points on AMT_Credit. Therefor the Male clients with Academic Degree Pay on-time

BIVARIATE ANALYSIS (NUMERICAL VARIABLE)

Age in Yrs_labelled vs AMT_INCOME_TOTAL vs CODE_GENDER



From the graph, it is clear that male guys tend to earn higher then female

Male clients, age between 20-60 and earning more then 175k+ tend to do the payments on time

Female clients, age between 30-60 have more tendency to pay the loans

BIVARIATE ANALYSIS

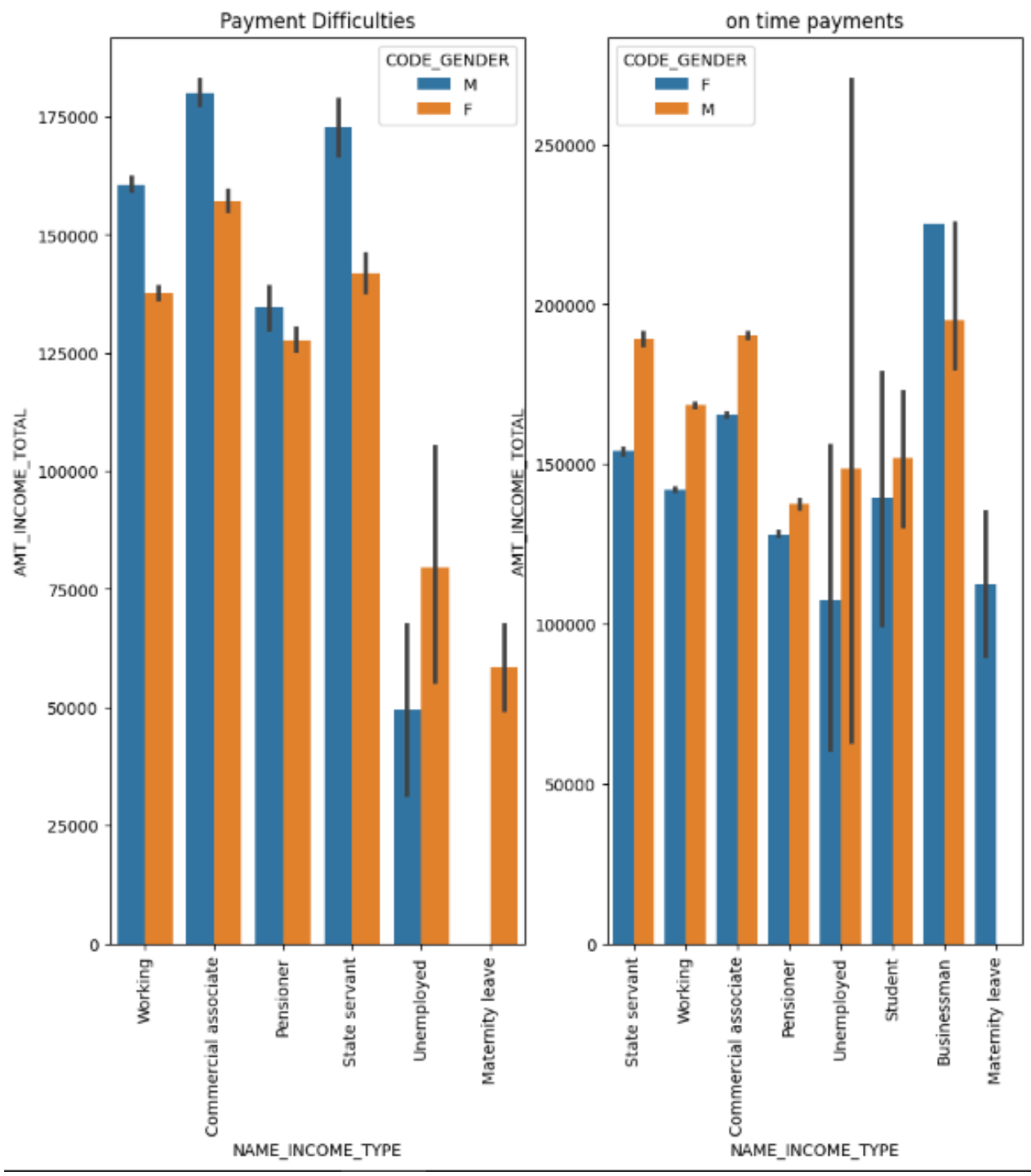
NAME_INCOME_TYPE and AMT_INCOME_TOTAL and CODE_GENDER

From both the graphs its clear that students and business man (both male and female) have high tendency to pay loans on time as well as they are the ones who earn more compared to any other income type.

Unemployed males and earning more then 80k average pays on time

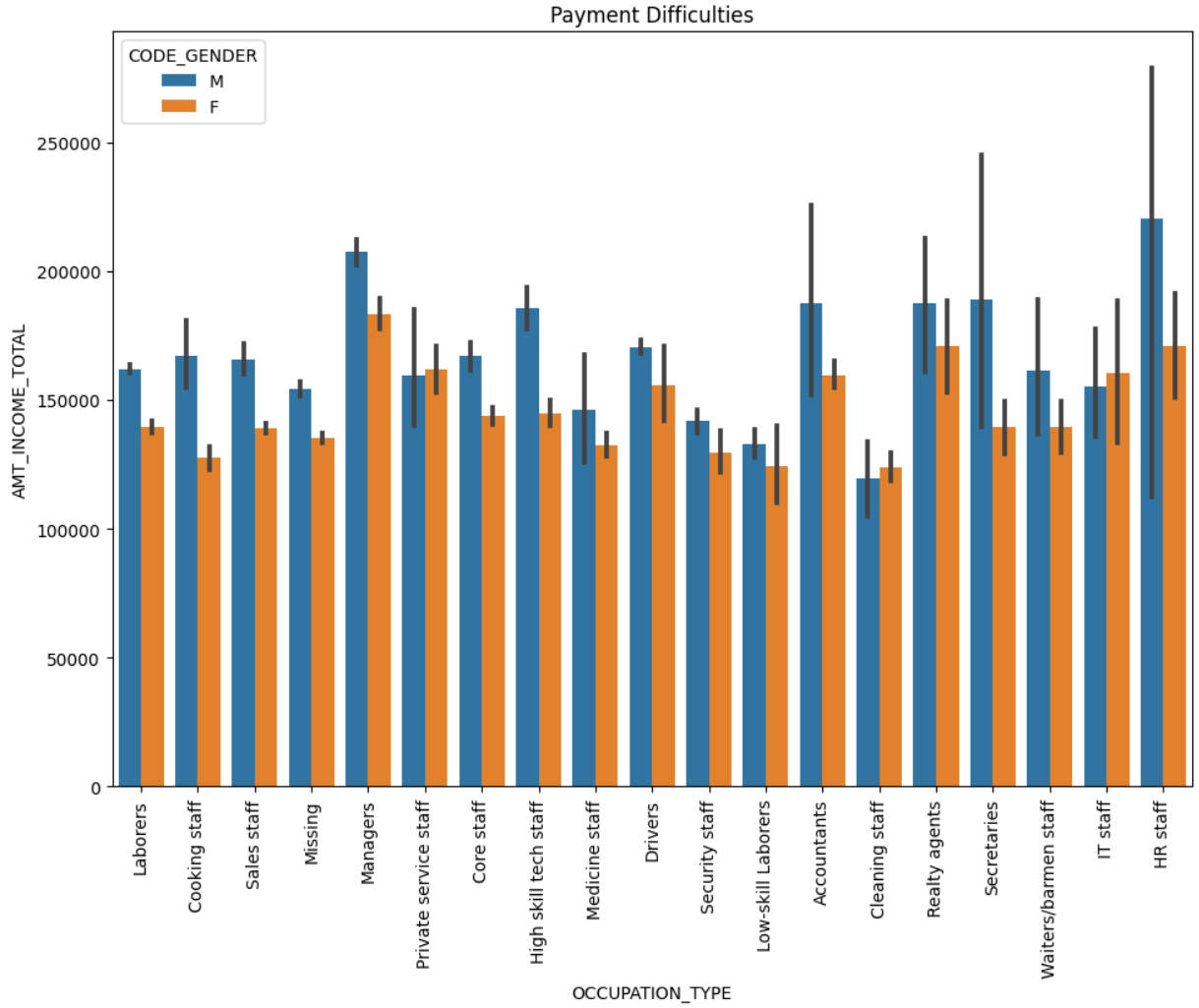
Unemployed females and earning more then 50k average pays on time

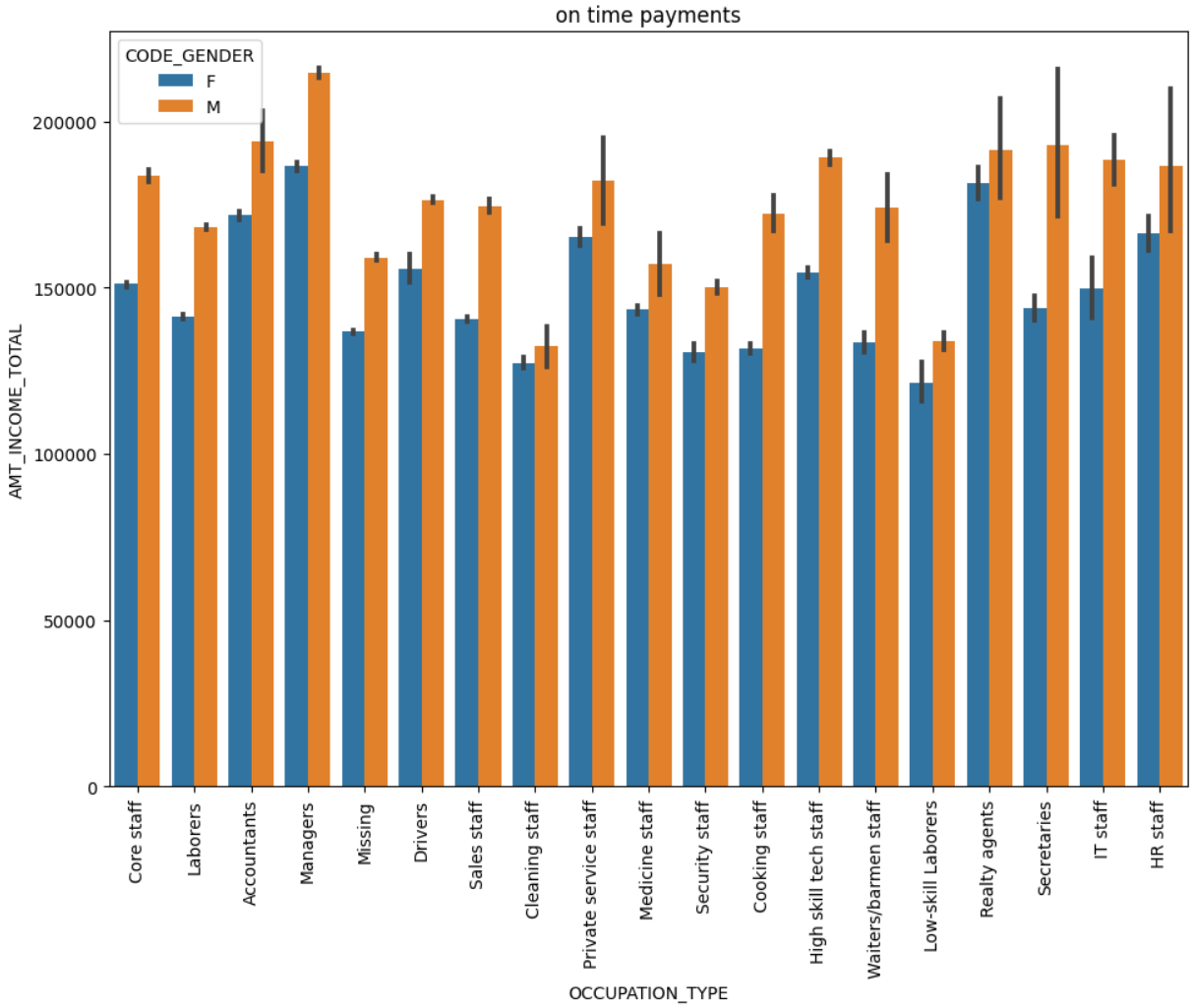
Most of the Females who are in maternity leave pays on time



BIVARIATE ANALYSIS

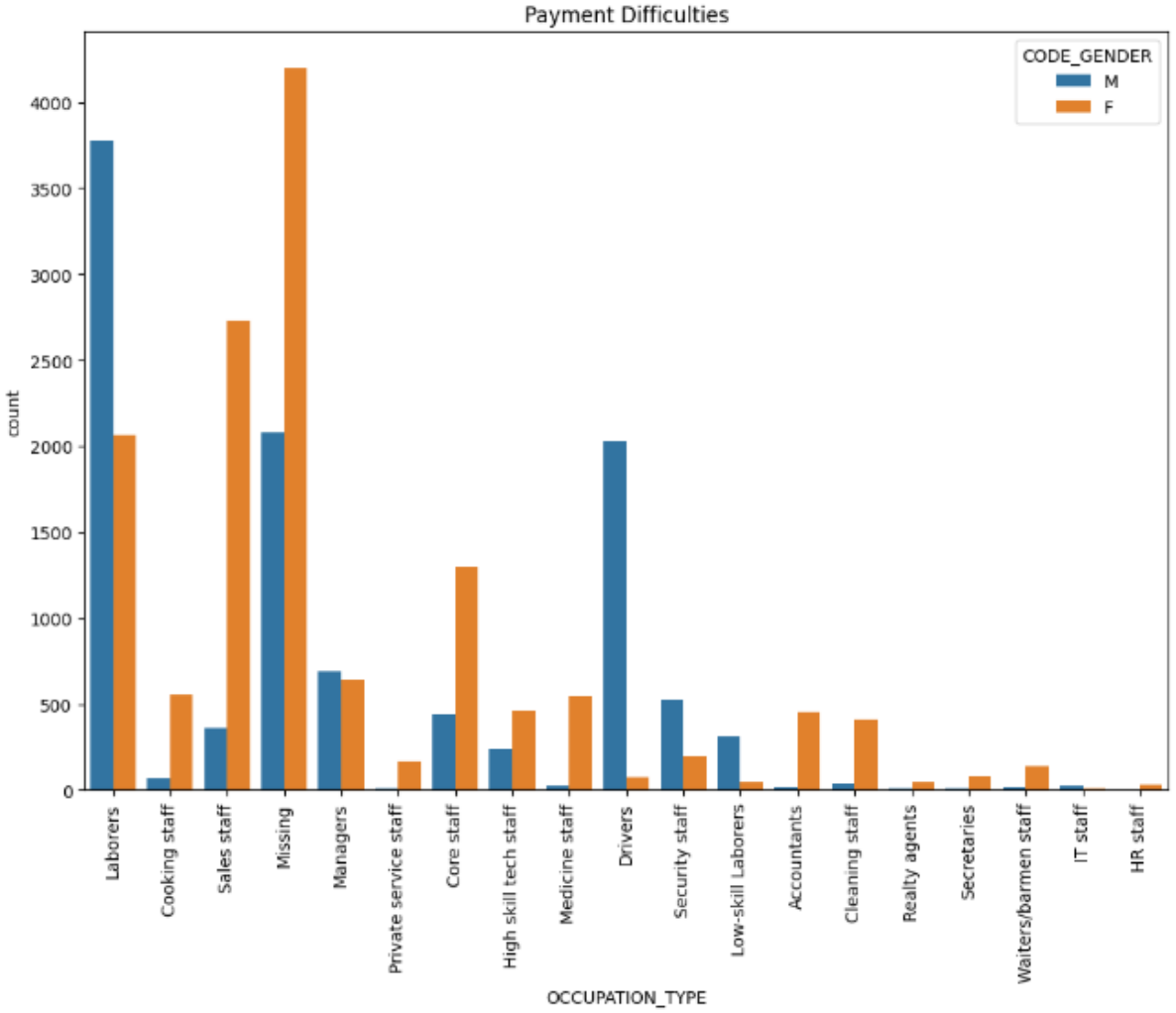
OCCUPATION_TYPE and AMT_INCOME_TOTAL and CODE_GENDER

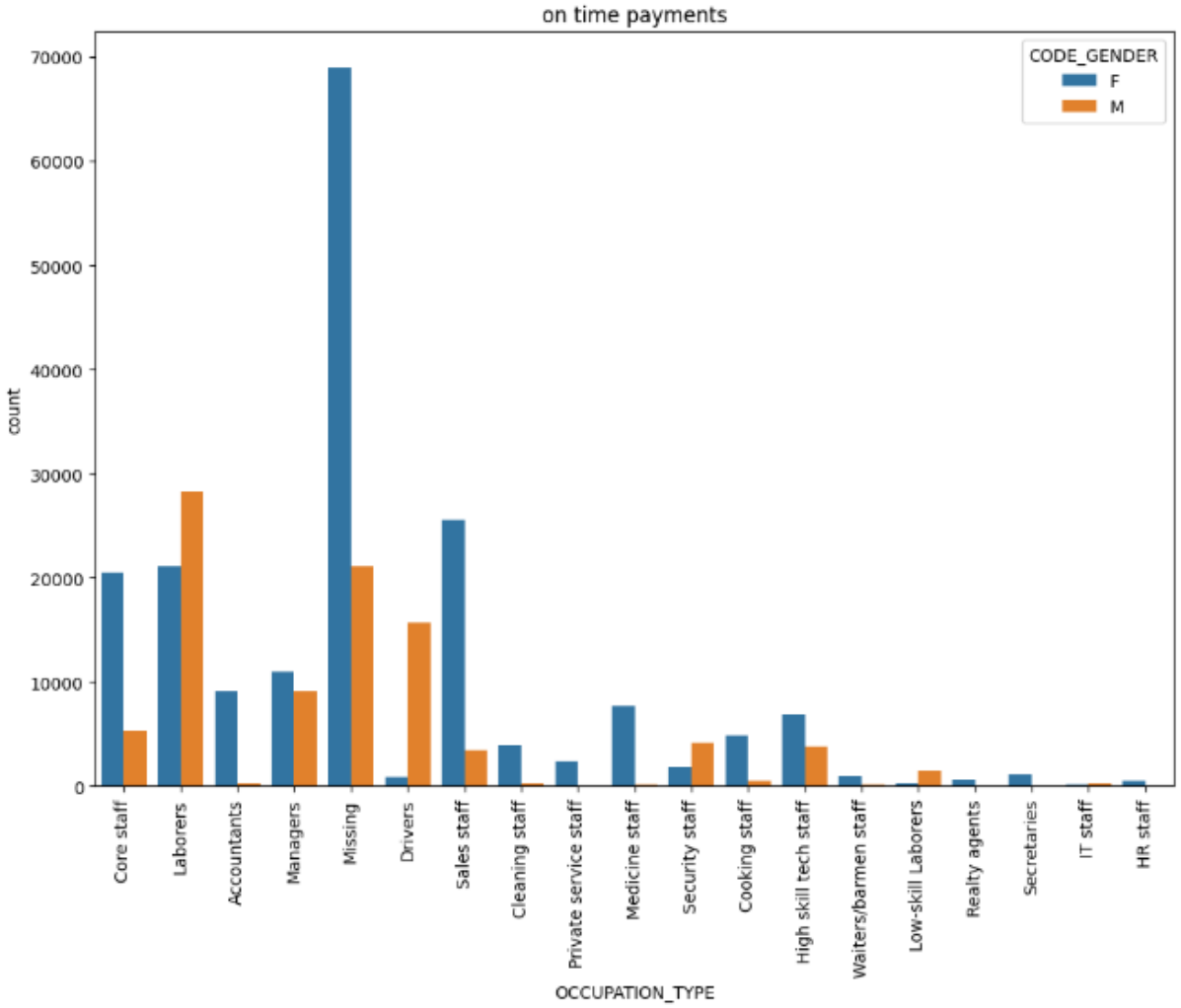




BIVARIATE ANALYSIS

OCCUPATION_TYPE and AMT_INCOME_TOTAL and CODE_GENDER





BIVARIATE ANALYSIS

OCCUPATION_TYPE and AMT_INCOME_TOTAL and CODE_GENDER

Its clear that male clients who's average income is more then 140k have high tendency in repaying

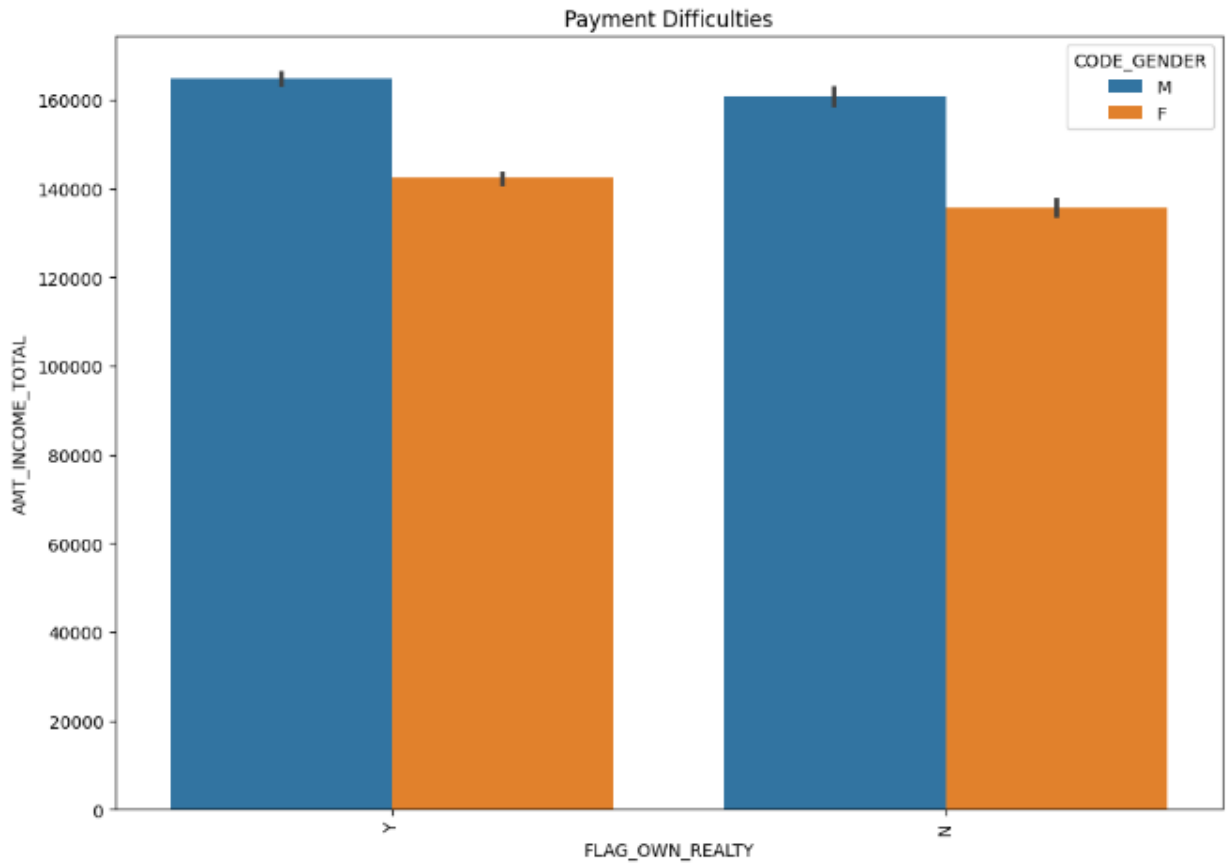
Similarly female clients who's average income is more then 120k have high tendency in repaying

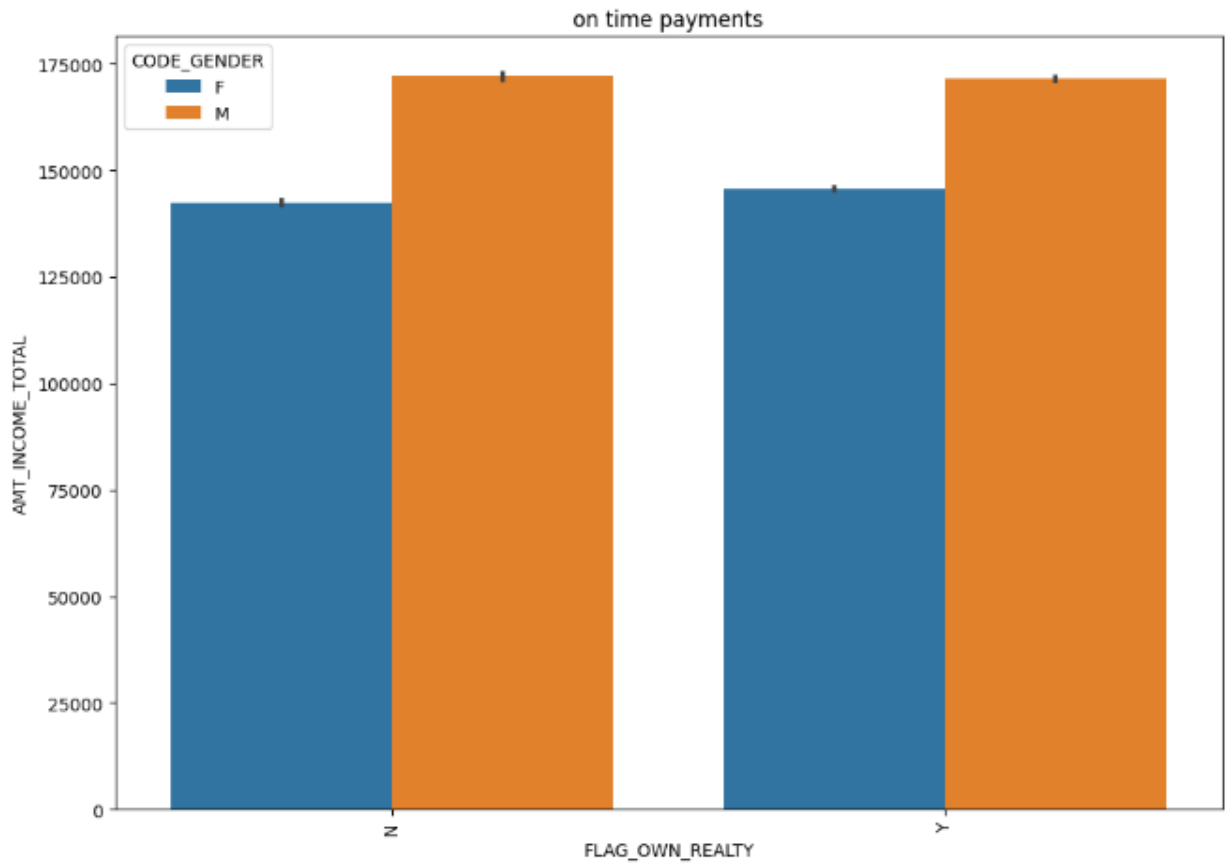
More concentration can be given towards Managers, Accountants, High skill tech staff, reality agent and secretary and earning more then 150k

Sales staff, laborers, Drives, security staff have difficulties paying the loans

BIVARIATE ANALYSIS

FLAG_OWN_REALTY and AMT_INCOME_TOTAL and CODE_GENDER

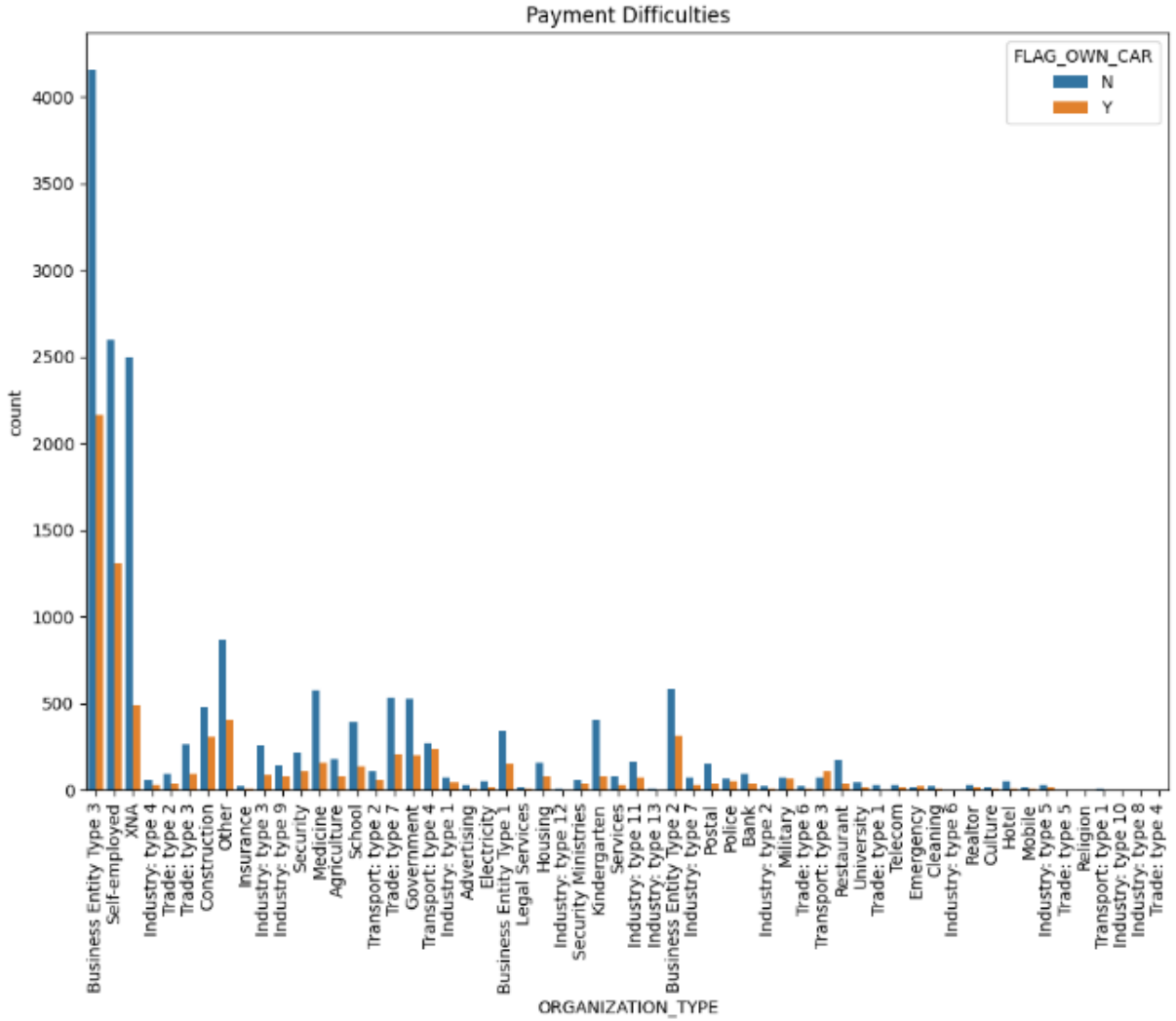


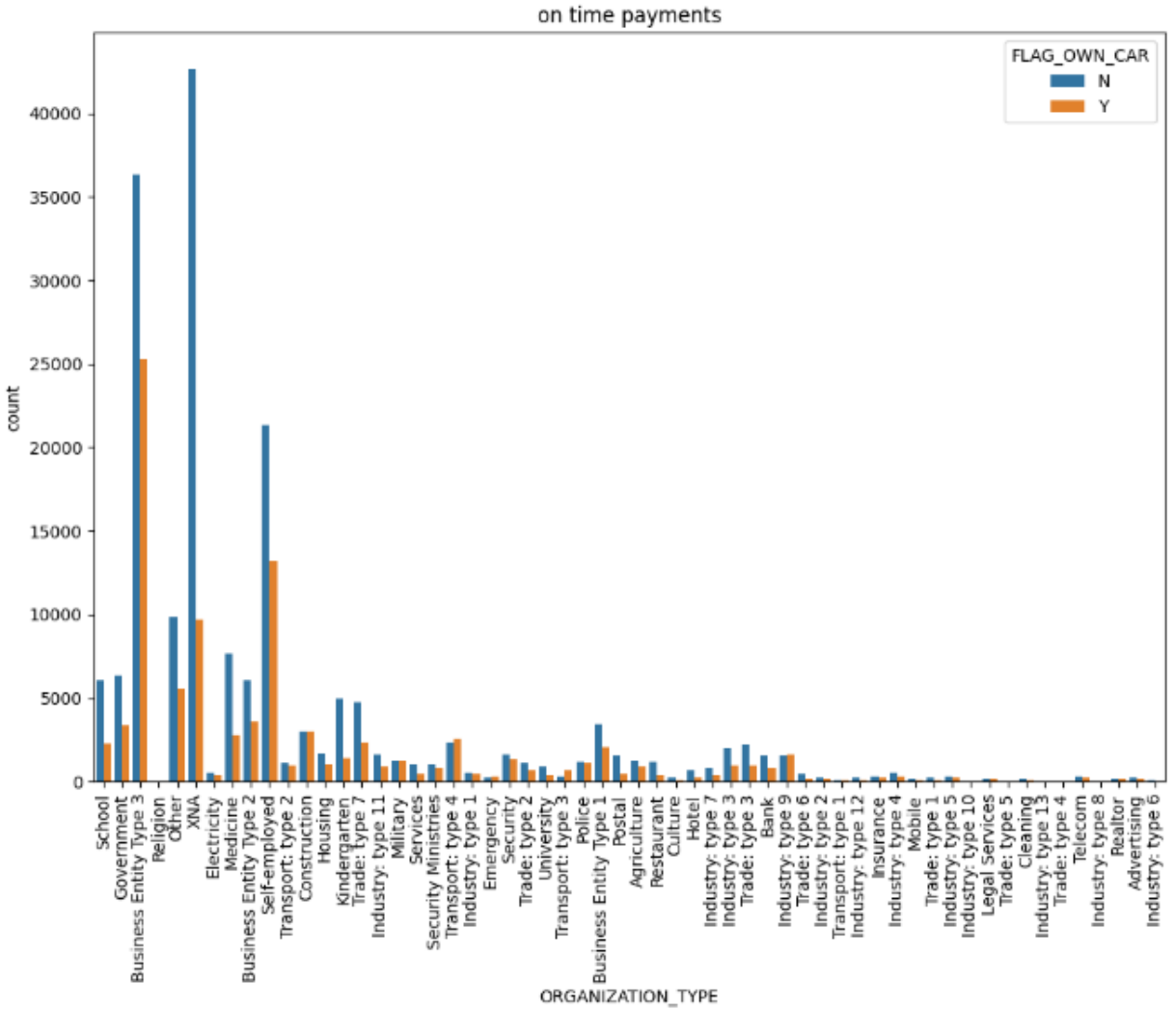


From the graph it is clear that male clients who are earning more then 150K+ and own either a house/flat, tend pay the loans on time

BIVARIATE ANALYSIS

ORGANIZATION_TYPE and FLAG_OWN_CAR





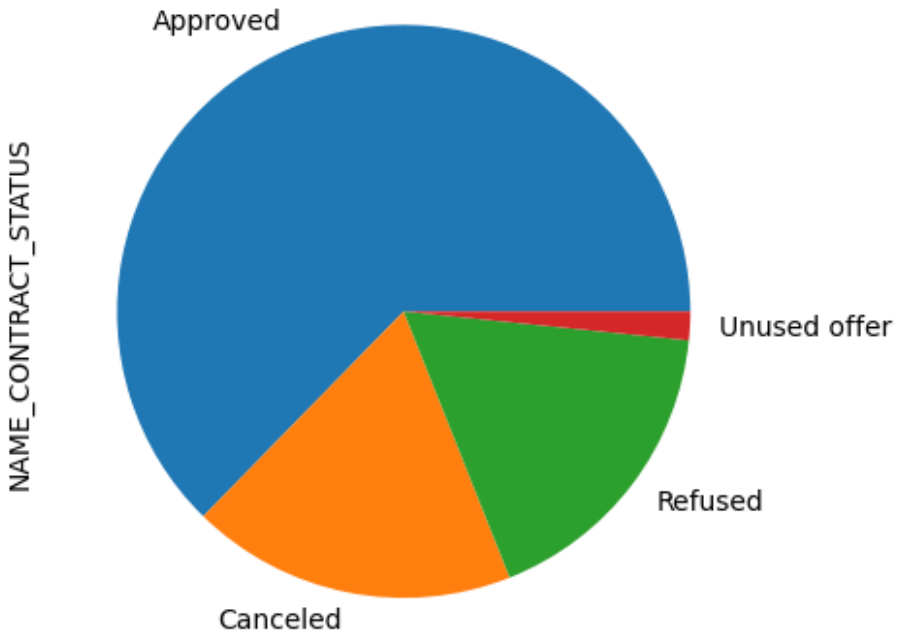
Clients who are Business Equity Type 3 and self employed and doesn't own a car have difficulty paying the loans

ANALYSIS PREVIOUS_APPLICATION.CSV

1. After **Importing** the dataset(previous_application.csv), I have explored the data and understood all the attributes. Initially the shape of the dataset was **1670214 rows** and **37 columns.**
2. All the Data Cleaning Process mentioned in **Slide 7** is Followed.
3. After all the outlier analysis, This dataset "previous_application.csv" is merged to our initial dataset "*application_data.csv*".
4. *From the next slide all the visualizations, Insights are drawn from merged data*

UNIVARIATE ANALYSIS

Univariate Analysis on NAME_CONTRACT_STATUS: Contract status (approved, cancelled, ...) of previous application

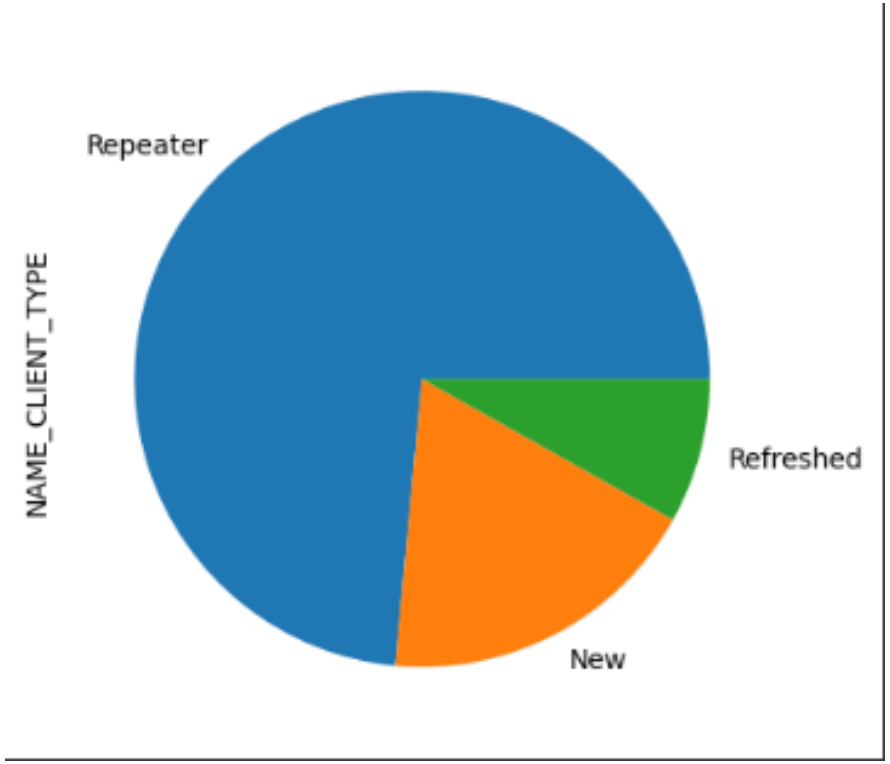


NAME_CONTRACT_TYPE	Value	Percentage(%)
Approved	886099	63
Canceled	259441	18.35
Refused	245390	17.35
Unused offer	22771	16.1

It is clearly visible that most of the applications got approved, but even canceled percentage is quite huge too.

UNIVARIATE ANALYSIS

Univariate Analysis on NAME_CLIENT_TYPE: Was the client old or new client when applying for the previous application

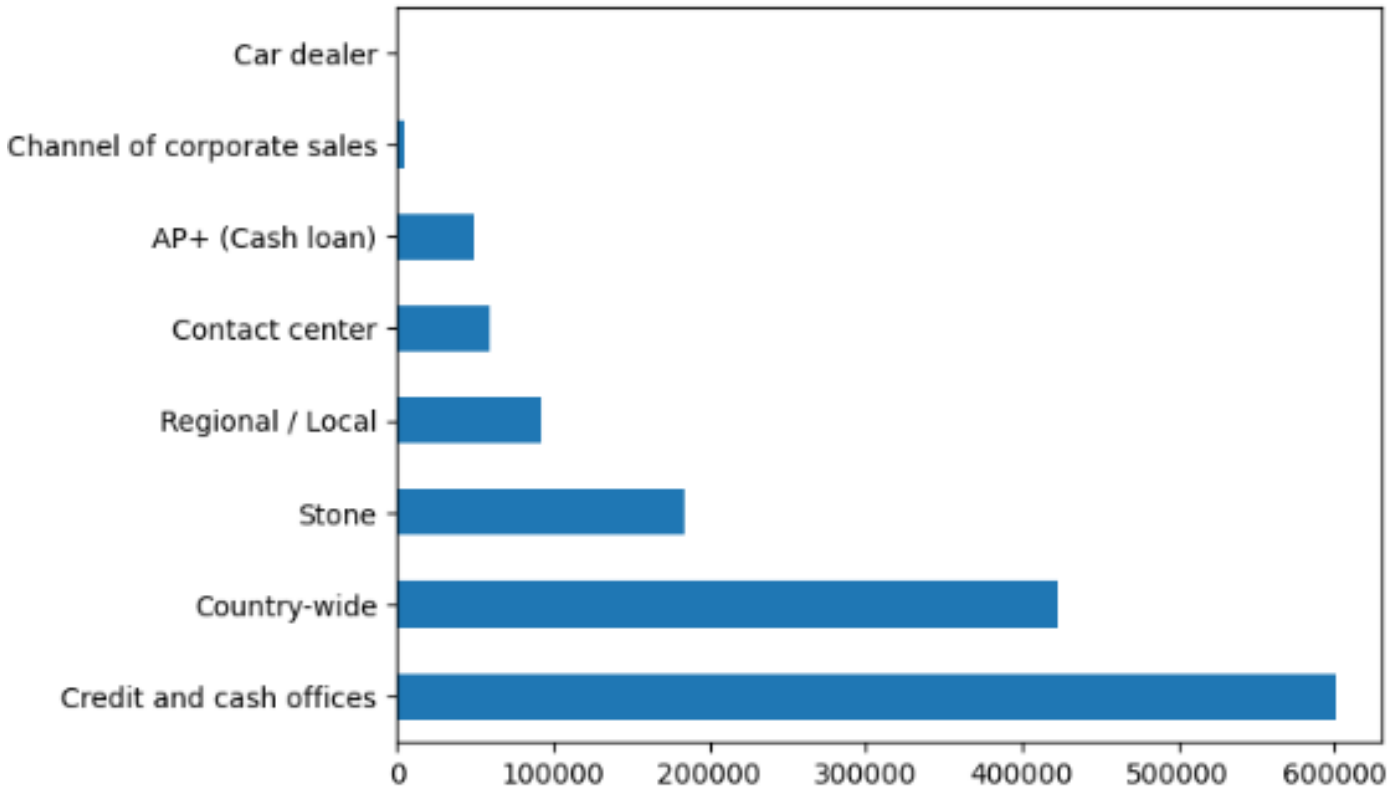


NAME_CLIENT_TYPE	Value	Percentage(%)
Repeater	1039225	73.5
New	259540	18.35
Refreshed	114936	8.13

Most of the clients are repeater as well as New clients are in quite high percent too

UNIVARIATE ANALYSIS

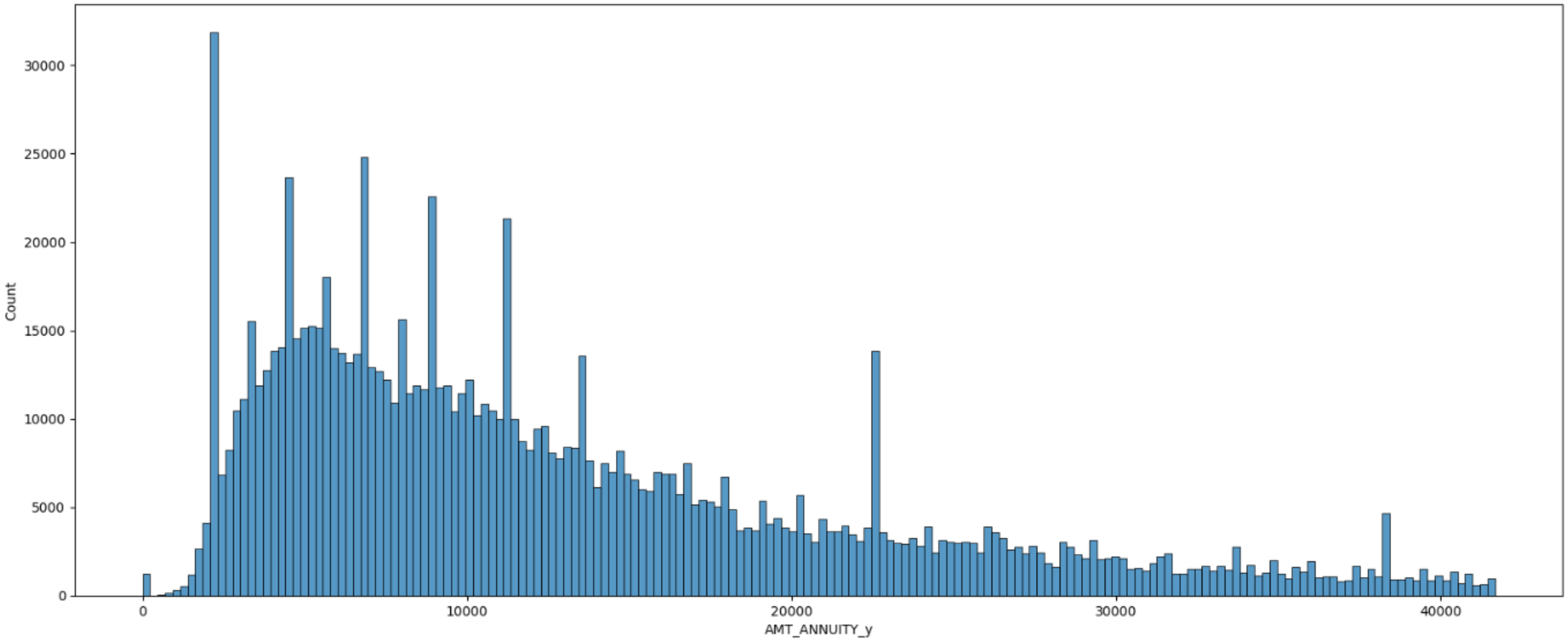
Univariate Analysis on CHANNEL_TYPE: Was the client old or new client when applying for the previous application



Its clearly visible that most of the clients were acquired from "Credit and cash offices " followed by country-wide.

UNIVARIATE ANALYSIS

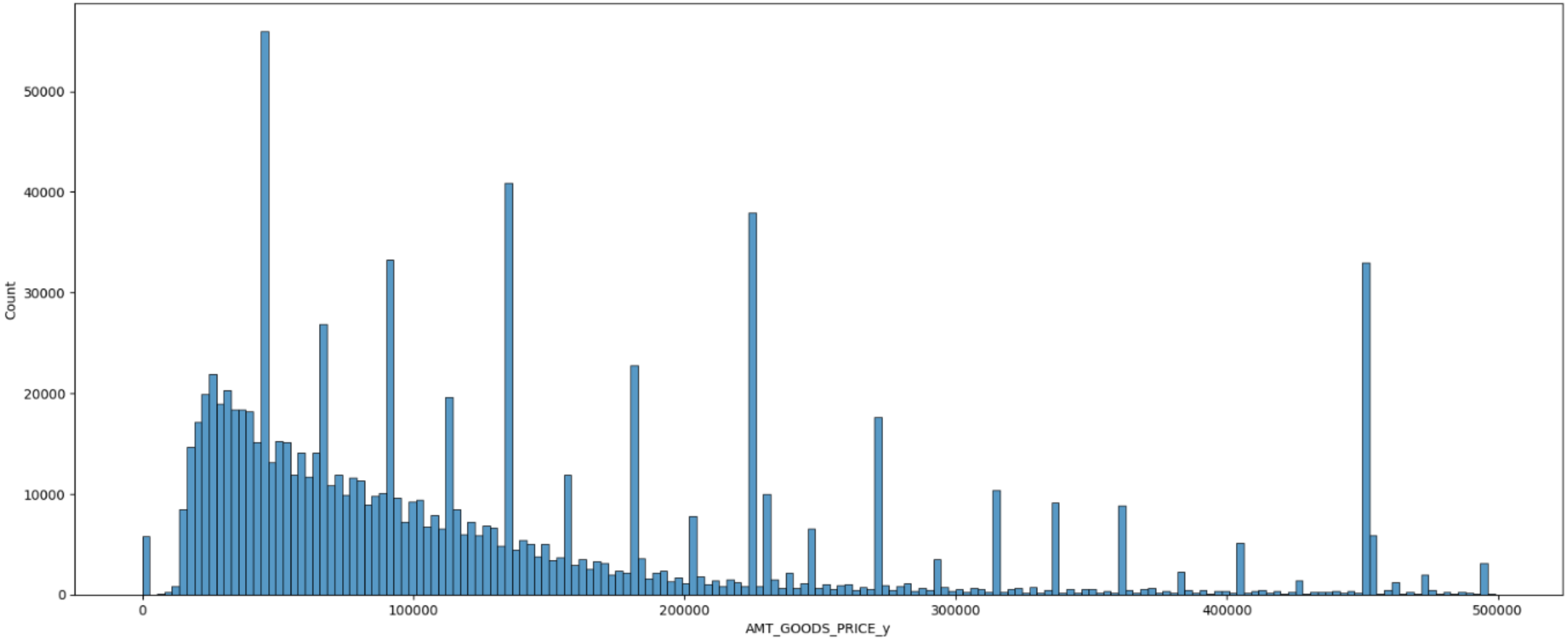
Univariate Analysis on AMT_ANNUITY from previous_application.csv



Most of the previous loan annuity is less then 15k

UNIVARIATE ANALYSIS

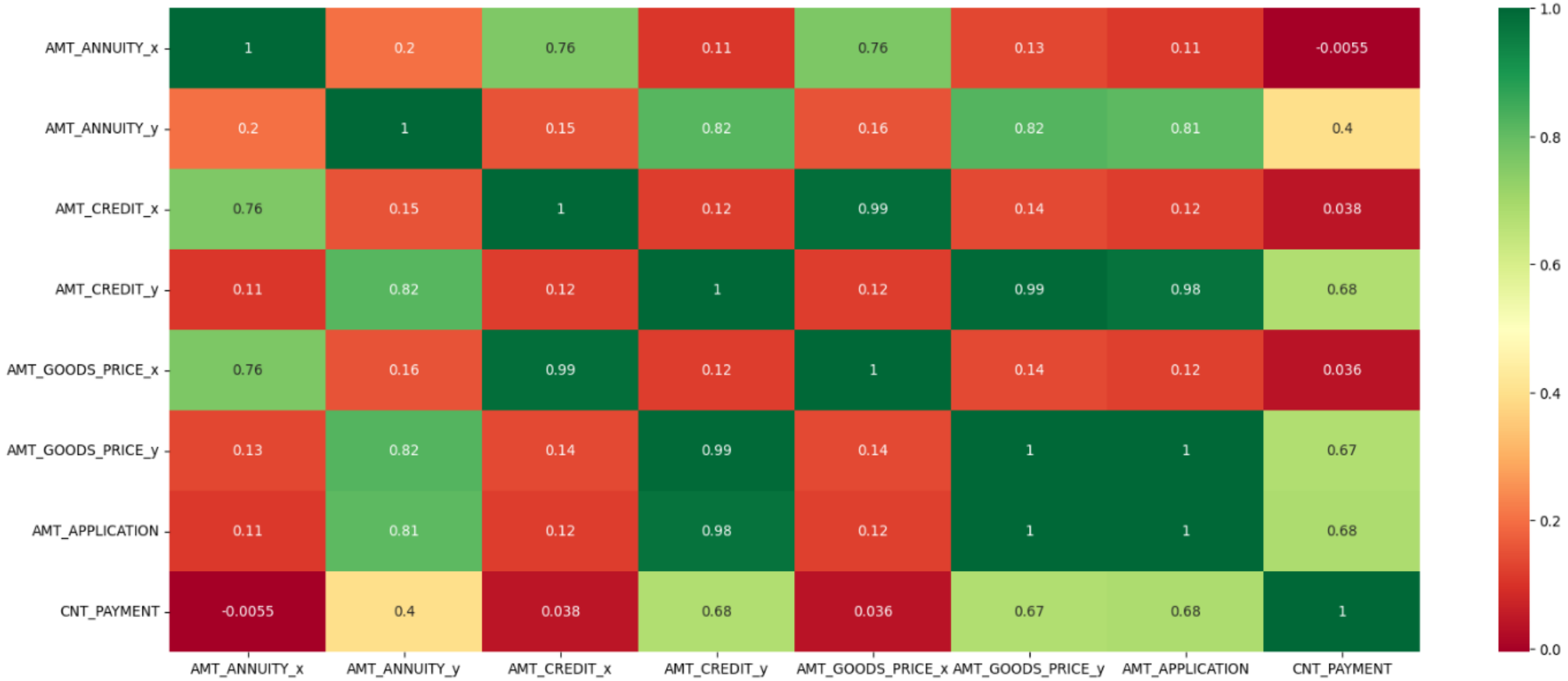
Univariate Analysis on AMT_GOODS_PRICE from previous_application.csv



Huge amount of clients asked the goods price in the previous application is less then 180k

CORRELATION ANALYSIS

Correlation for all the columns Annunity, Credit, Goods Price, Cnt Payment was found and visualized using Heatmap

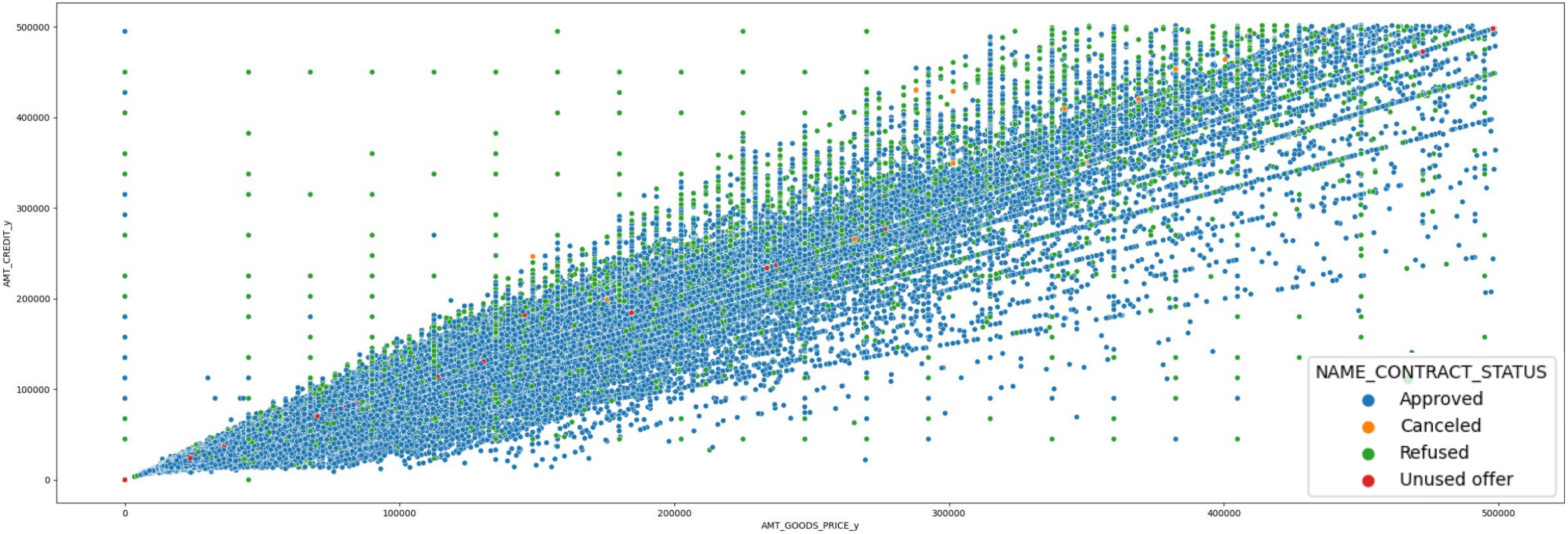


AMT_ANNUITY_x has the decent correlation with AMT_CREDIT_x and AMT_GOODS_PRICE_x, these 3 are directly proportional to each other

AMT_APPLICATION has the high correlation with AMT_CREDIT_y, AMT_GOODS_PRICE_y and decent correlation with CNT_PAYMENT.
This means the bank has given the exact amount or close to the value what client has asked for

MULTIVARIATE ANALYSIS

'AMT_GOODS_PRICE(y)' and 'AMT_CREDIT(y) and 'NAME_CONTRACT_STATUS'

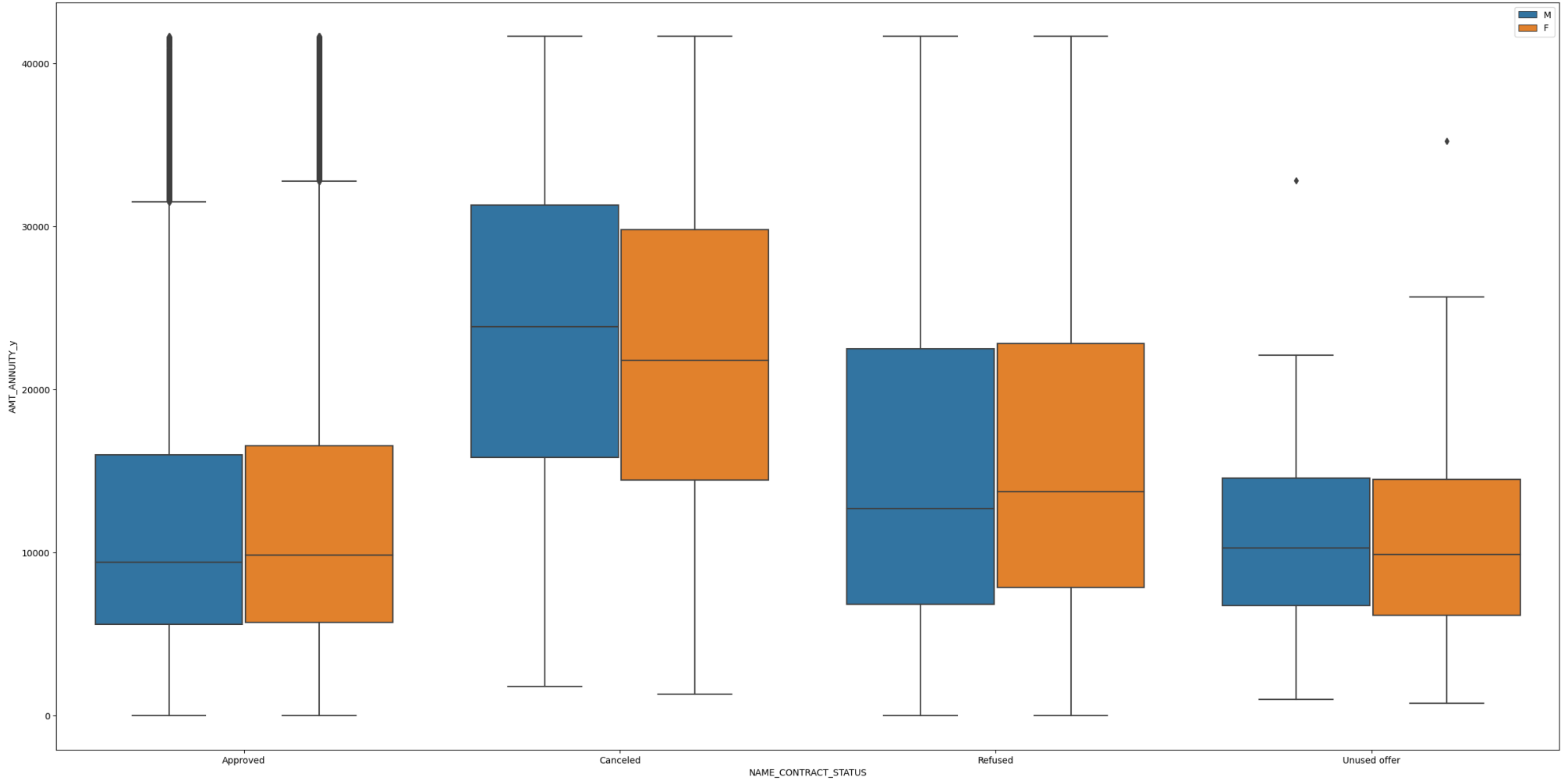


In the previous application dataset, most of the refused applications are represented in the beginning, this is because with less goods price, the client is expecting more loan credit from the bank/companies.

Goods Price less then 300k and credit greater then 300k, in this area more refused applications are found

MULTIVARIATE ANALYSIS

'AMT_GOODS_PRICE(y)' and 'AMT_ANNUITY'(y) and 'NAME_CONTRACT_STATUS'

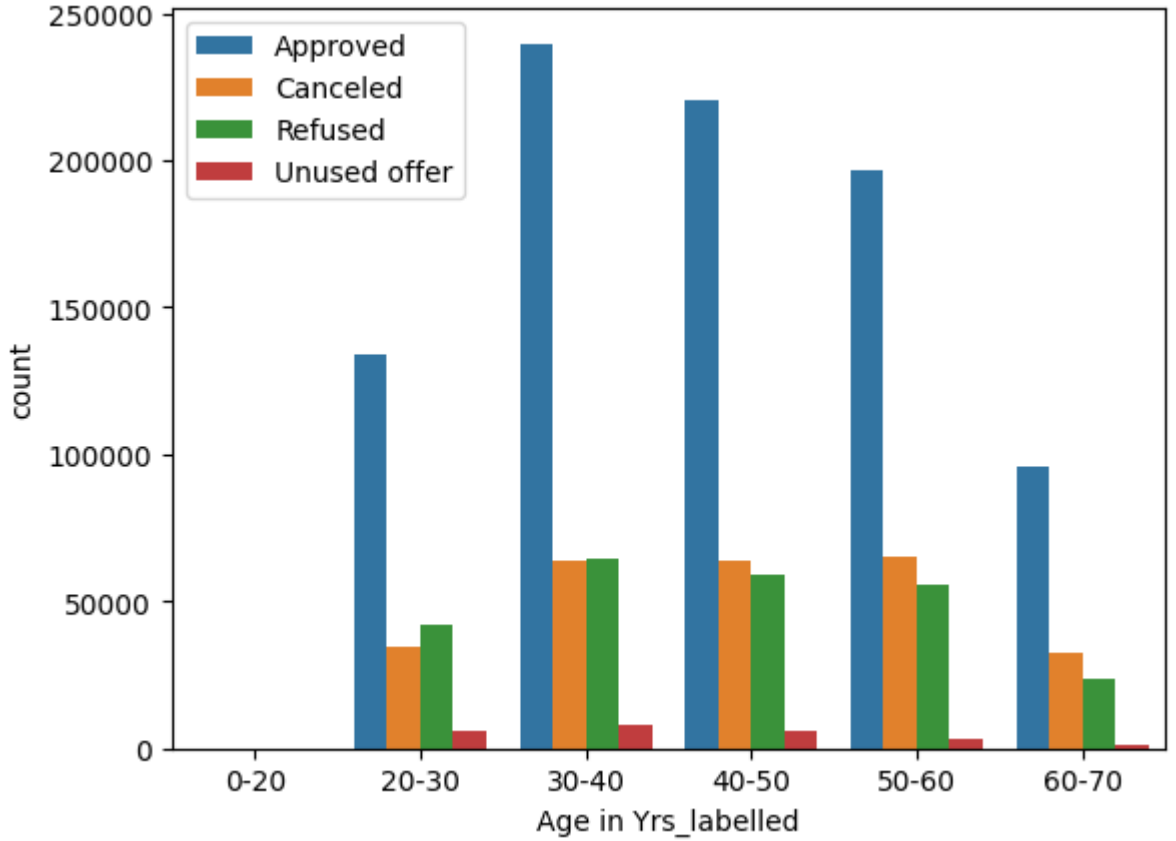


Males clients who got canceled, paid higher median annuity then females

Female clients who got refused, paid higher median annuity

BIVARIATE ANALYSIS

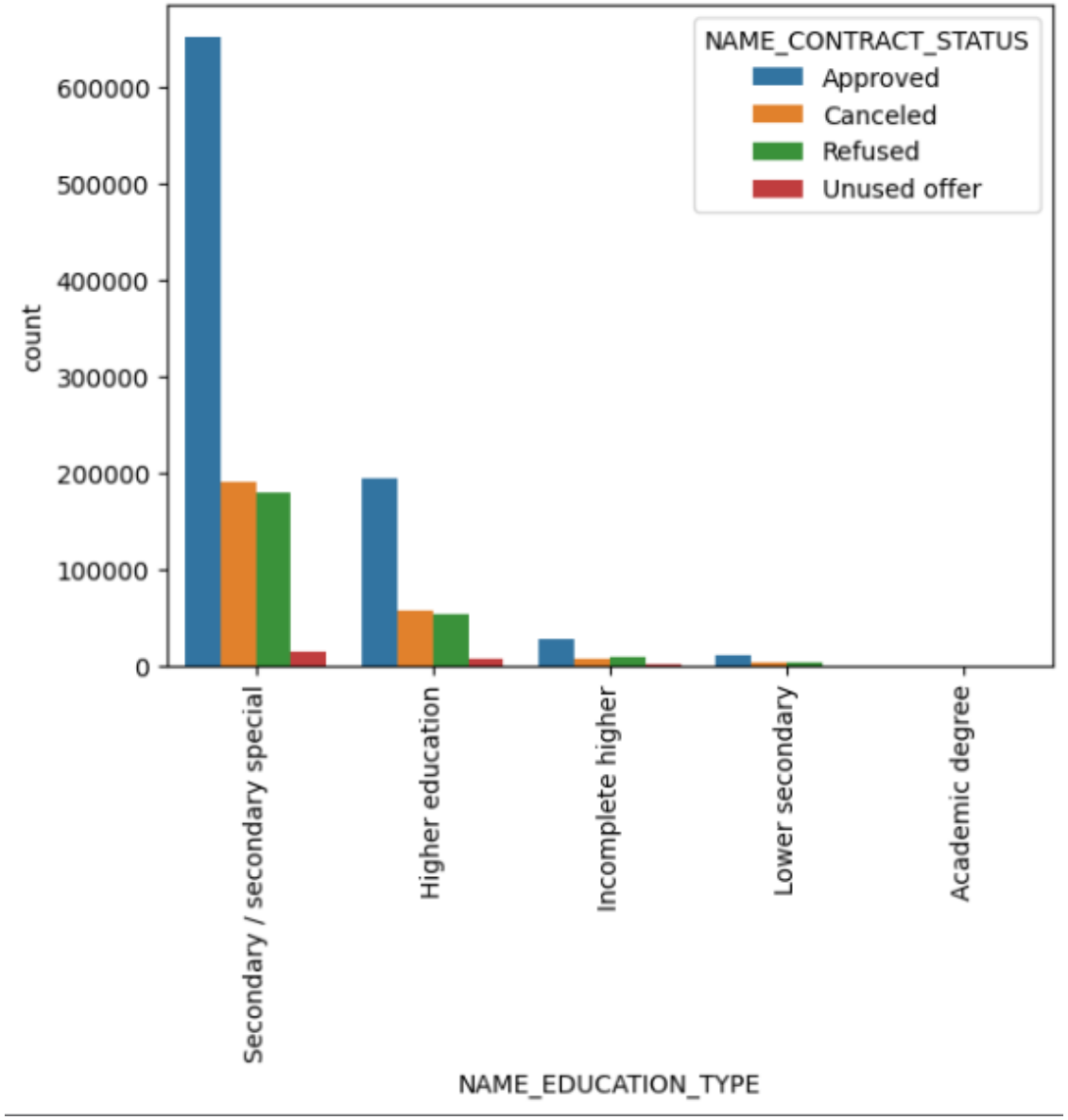
'Age in Yrs_labelled' and 'NAME_CONTRACT_STATUS'



between the age range of 30-60 got more approvals then other age clients

BIVARIATE ANALYSIS

'NAME_EDUCATION_TYPE' and 'NAME_CONTRACT_STATUS'



People with Secondary / Secondary special and Higher education got the most approvals

CONCLUSIONS

Better do the focus on clients,

1. **Female clients** with higher education and secondary education
2. **Business Man** & **Students** have zero records of late payments
3. **Commercial Associates** and **Pensioners** have good payment records
4. **Higher education** and **academic degree** clients
5. **Married clients** make the payments clearly
6. More clear and efficient payments have happened from clients **aged above 40**
7. A person who got his career started before at least **2500 days,** from the time of application

CONCLUSIONS

Better do the focus on clients,

8. Male clients, age between **20-60** and earning more then **175k+** tend to do the payments on time
9. Female clients, age between **30-60** have more tendency to pay the loans
10. **Unemployed males** and earning more then **80k** average pays on time
11. **Unemployed females** and earning more then **50k** average pays on time
12. More concentration can be given towards **Managers, Accountants, High skill tech staff, reality agent** and **secretary** and earning more then **150k**
13. **Government clients** who own car have more tendency playing the loans