CREDIT EDA CASE STUDY

By Dong Nguyen

TABLE OF CONTENTS

01

INTRODUCTION

OBJECTIVES

- In this Credit EDA, apart from applying the techniques that I have learnt in the EDA module, I 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.
- This Credit EDA 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.

DATA EXPLAINATION

- The data given 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 at different stages of the process.
- In this case study, I will use EDA to understand how consumer attributes and loan attributes influence the tendency to default.

02
DATA CLEANING AND HANDLING OUTLIERS

DATA CLEANING - CURRENT APPLICATION

AMT_ANNUITY	0.003902
AMT_GOODS_PRICE	0.090403
NAME_TYPE_SUITE	0.420148
OCCUPATION_TYPE	31.345545
CNT_FAM_MEMBERS	0.000650
EXT_SOURCE_2	0.214626
EXT_SOURCE_3	19.825307
OBS_30_CNT_SOCIAL_CIRCLE	0.332021
DEF_30_CNT_SOCIAL_CIRCLE	0.332021
OBS_60_CNT_SOCIAL_CIRCLE	0.332021
DEF_60_CNT_SOCIAL_CIRCLE	0.332021
DAYS_LAST_PHONE_CHANGE	0.000325
AMT_REQ_CREDIT_BUREAU_HOUR	13.501631
AMT_REQ_CREDIT_BUREAU_DAY	13.501631
AMT_REQ_CREDIT_BUREAU_WEEK	13.501631
AMT_REQ_CREDIT_BUREAU_MON	13.501631
AMT_REQ_CREDIT_BUREAU_QRT	13.501631
AMT_REQ_CREDIT_BUREAU_YEAR	13.501631

- I check the columns that have more than 40% of missing values and drop these columns for better analysis
- The image show the columns that have less than 40% of missing values and more than 0% of missing values.
- The “OCCUPATION TYPE” column has 31% missing values. Since i think the column is important for further analysis, I would like to keep the column. I would keep the missing values as a "Missing occupation" incase some other expert in the industry will know what type of Opccupation it is.

DATA CLEANING - CURRENT APPLICATION

AMT_ANNUITY	0.003902
AMT_GOODS_PRICE	0.090403
NAME_TYPE_SUITE	0.420148
OCCUPATION_TYPE	31.345545
CNT_FAM_MEMBERS	0.000650
EXT_SOURCE_2	0.214626
EXT_SOURCE_3	19.825307
OBS_30_CNT_SOCIAL_CIRCLE	0.332021
DEF_30_CNT_SOCIAL_CIRCLE	0.332021
OBS_60_CNT_SOCIAL_CIRCLE	0.332021
DEF_60_CNT_SOCIAL_CIRCLE	0.332021
DAYS_LAST_PHONE_CHANGE	0.000325
AMT_REQ_CREDIT_BUREAU_HOUR	13.501631
AMT_REQ_CREDIT_BUREAU_DAY	13.501631
AMT_REQ_CREDIT_BUREAU_WEEK	13.501631
AMT_REQ_CREDIT_BUREAU_MON	13.501631
AMT_REQ_CREDIT_BUREAU_QRT	13.501631
AMT_REQ_CREDIT_BUREAU_YEAR	13.501631

- Let's choose one column among the columns of the "AMT_REQ_CREDIT_BUREAU".
- We need a column with the time frame that is not too short but also not too long.
- I will choose the month column for analysis and drop the hour, day, week, quarter, and year columns.
- The most common value of AMT_REQ_CREDIT_BUREAU_MON is 0, so let's impute the missing value by 0.
- The missing values of columns "AMT_ANNUITY, AMT_GOODS_PRICE, CNT_FAM_MEMBERS, EXT_SOURCE_2" are less than 5%, and as in statistical language, if the number of the cases is less than 5% of the sample, then I can drop them.

DATA CLEANING - CURRENT APPLICATION

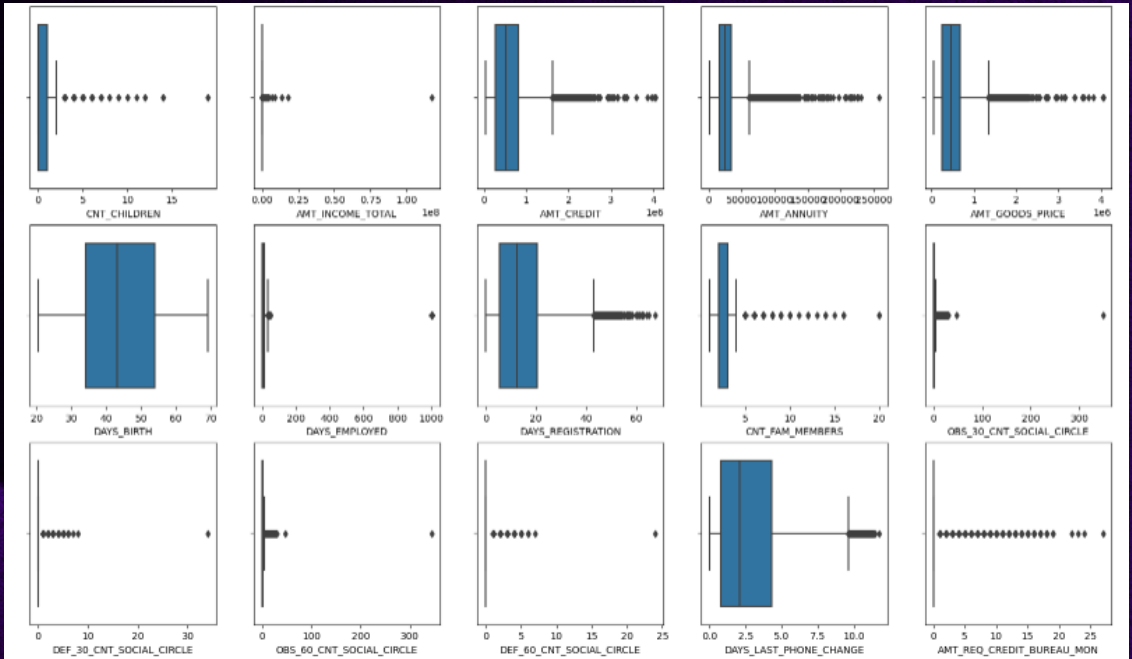
```
count    246546.000000
mean          0.510853
std           0.194844
min           0.000527
25%           0.370650
50%           0.535276
75%           0.669057
max           0.896010
Name: EXT_SOURCE_3, dtype: float64
```

- As we can see, the mean and median of columns "EXT_SOURCE_3" are very close, so I will impute the missing values of this columns with the median.
- We have cleared all the missing values of the current application data set

DATA CLEANING – PREVIOUS APPLICATION

- I repeated the same procress with previous data application which are indentifying columns that have more than 40% of missing values and dropping these columns.

HANDLING OUTLIERS

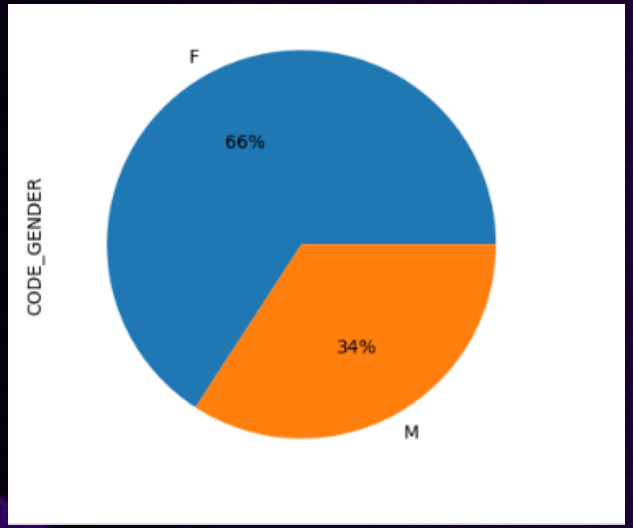


- We could identify outliers by plotting boxplots. All the outliers are beyond the maximum.
- Since there are a lot of columns, I will run a boxplot in loops for the columns that have outliers.
- As we can see, all of these columns have outliers except for "DAYS_BIRTH" column.

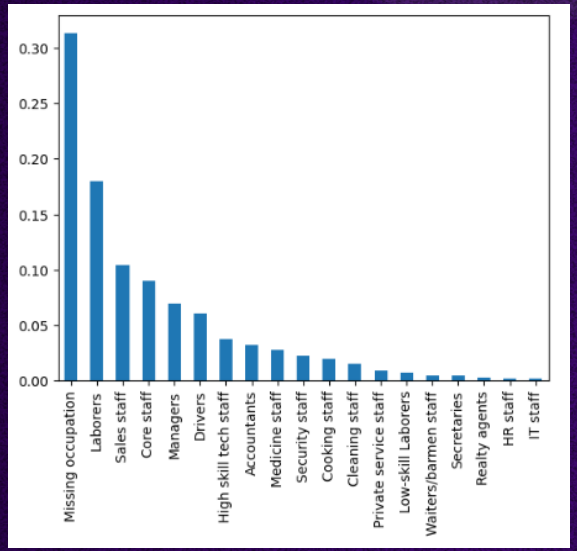
03

CURRENT APPLICATION AND PREVIOUS APPLICATION ANALYSIS

UNIVARIABLE ANALYSIS - CURRENT APPLICATION

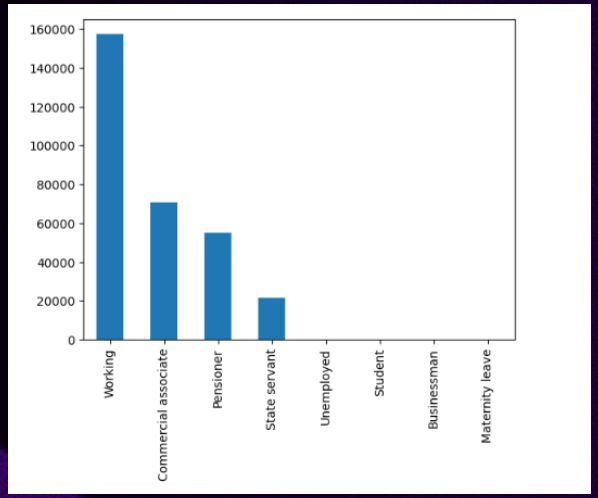


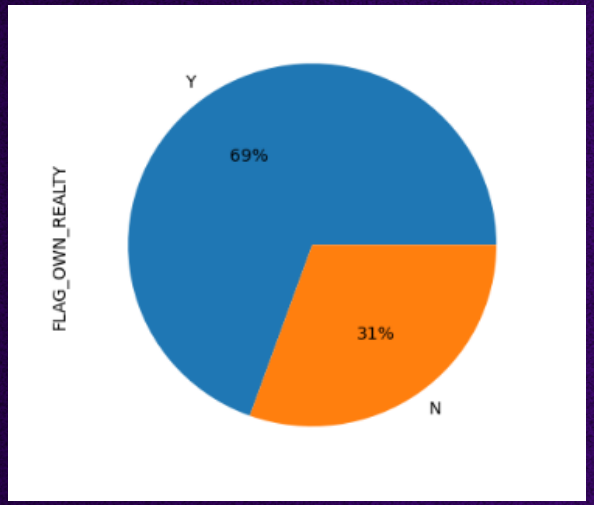
- There are more female clients who apply for the loan than male clients. The Ratio of F:M is 2:1



- There are more than 30% of clients who has missing occupation applying for the loan. The second highest occupation is Laborers with less than 20%.
- HR Staff, IT Staff and Realty agents have least interest in taking loans.

UNIVARIABLE ANALYSIS - CURRENT APPLICATION

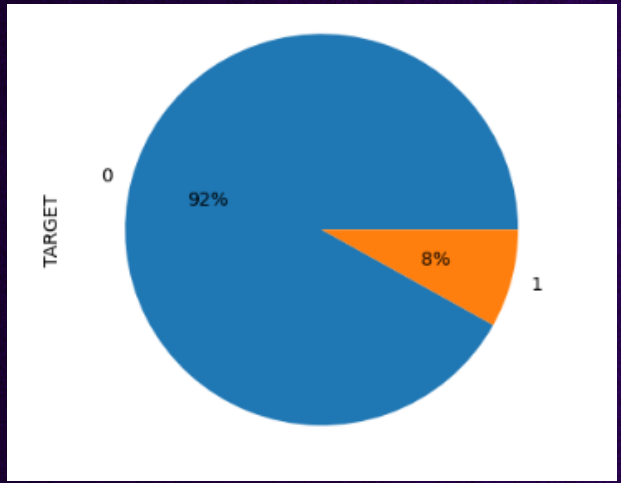




- For the client income type, most of client are working, and then come the commercial associate group.

- There are 69% of clients who apply for the loan own a a house or flat

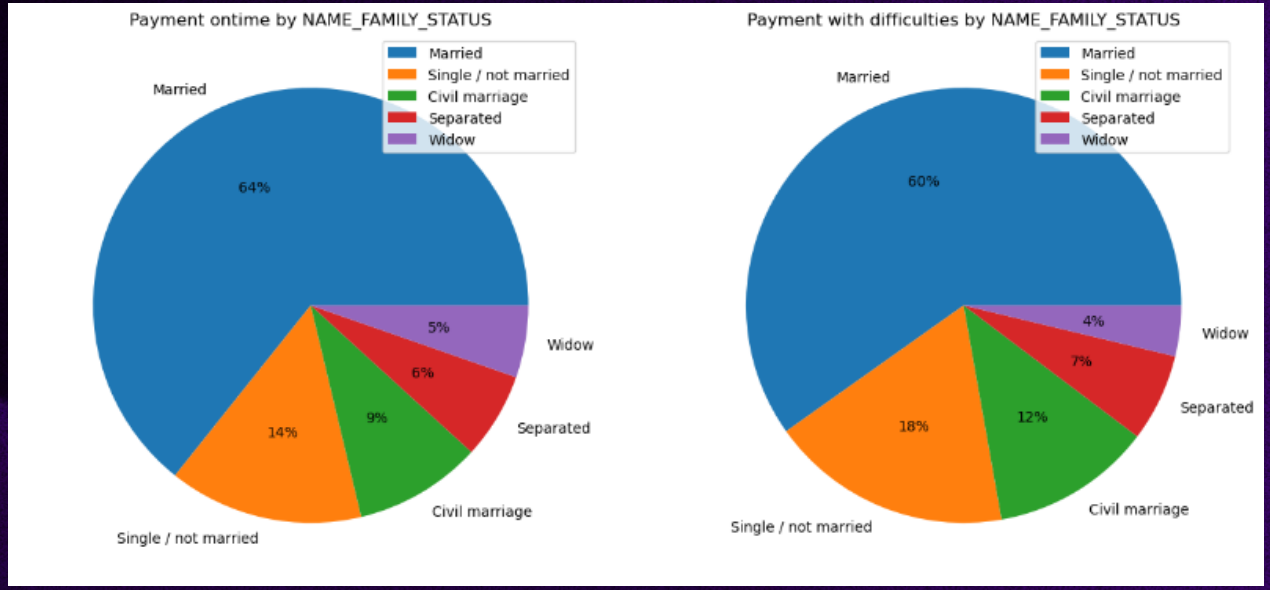
DATA INBALANCE - CURRENT APPLICATION ANALYSIS



- There are imbalance in the ‘Target variable’ columns in the dataset. It shows that most of the customers made the payment on time (92%) and only a few of customers with payment difficulties(8%)
- The ratio of the data imbalance is 11.345:1

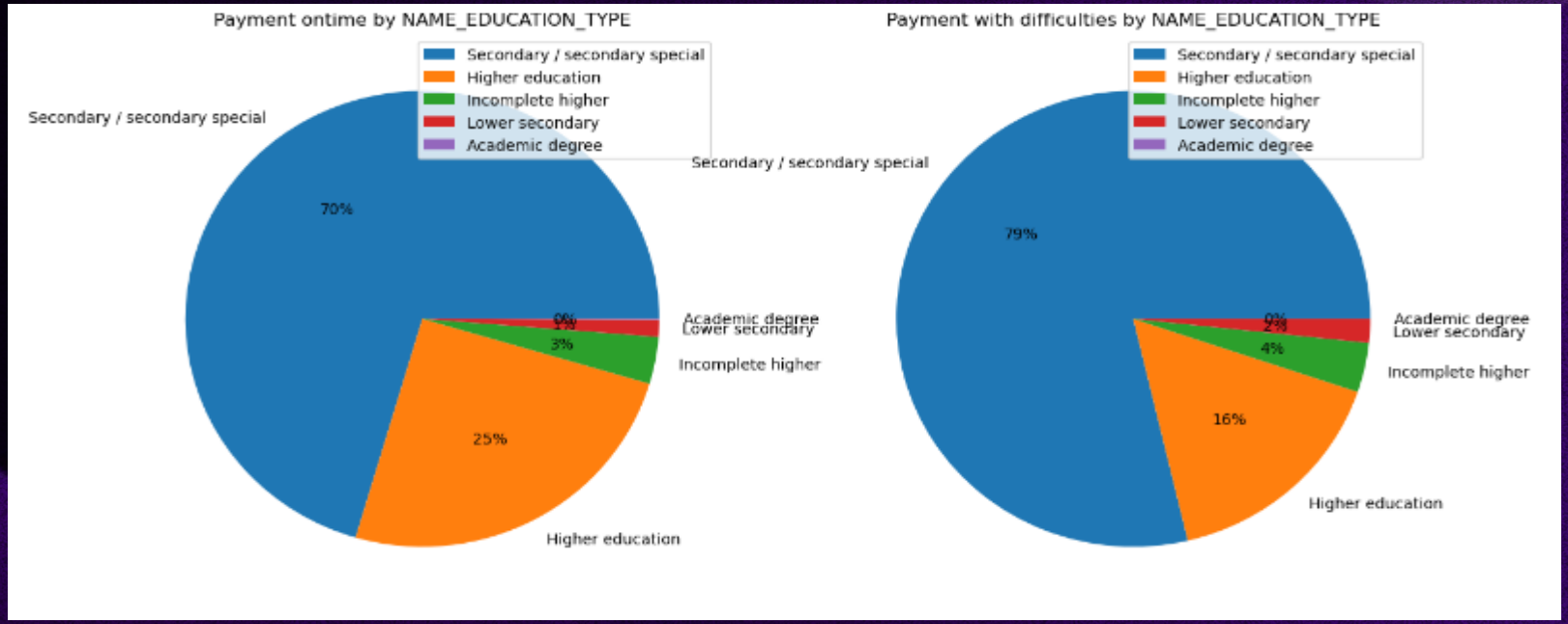
UNIVARIABLE ANALYSIS - CURRENT APPLICATION

I will compare some of the variables between client with on-time payments and client with payment difficulties



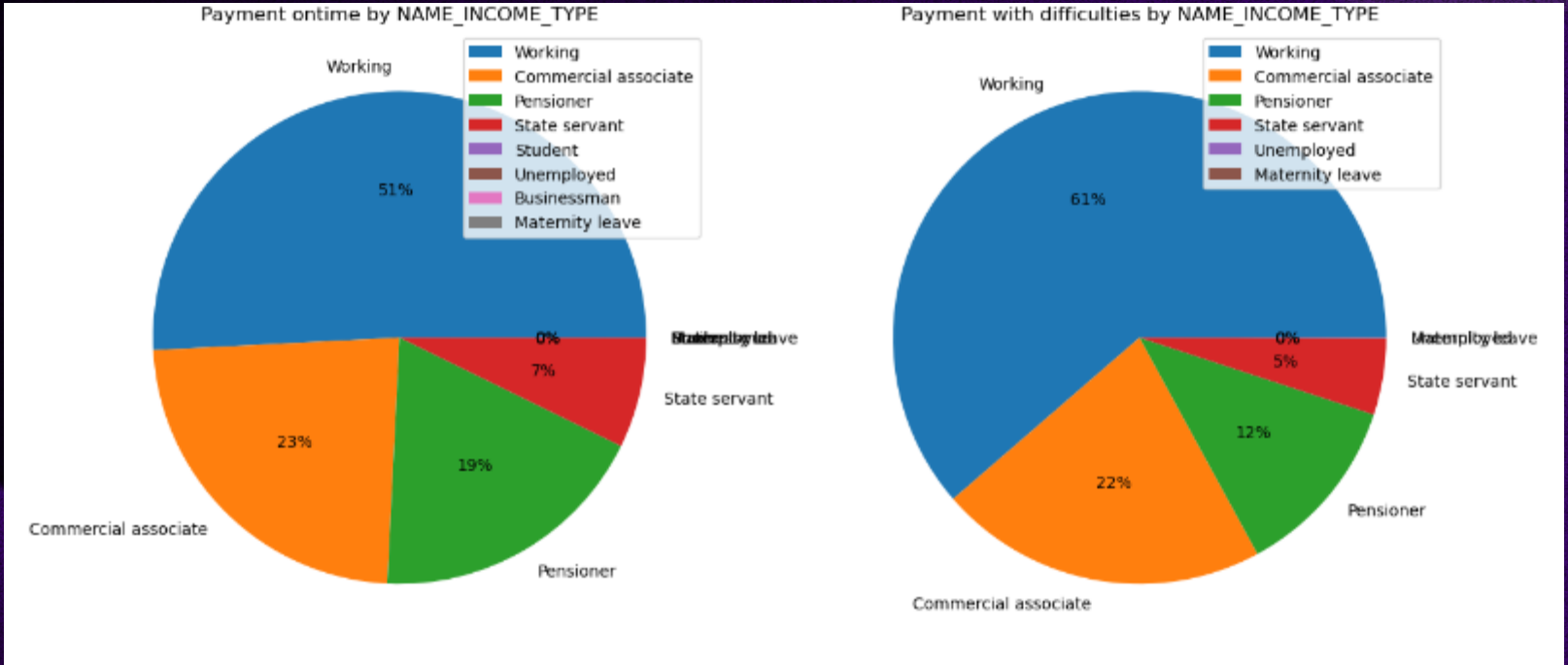
— Clients who are single or civil marriage is more likely to have payment with difficulties

UNIVARIABLE ANALYSIS - CURRENT APPLICATION



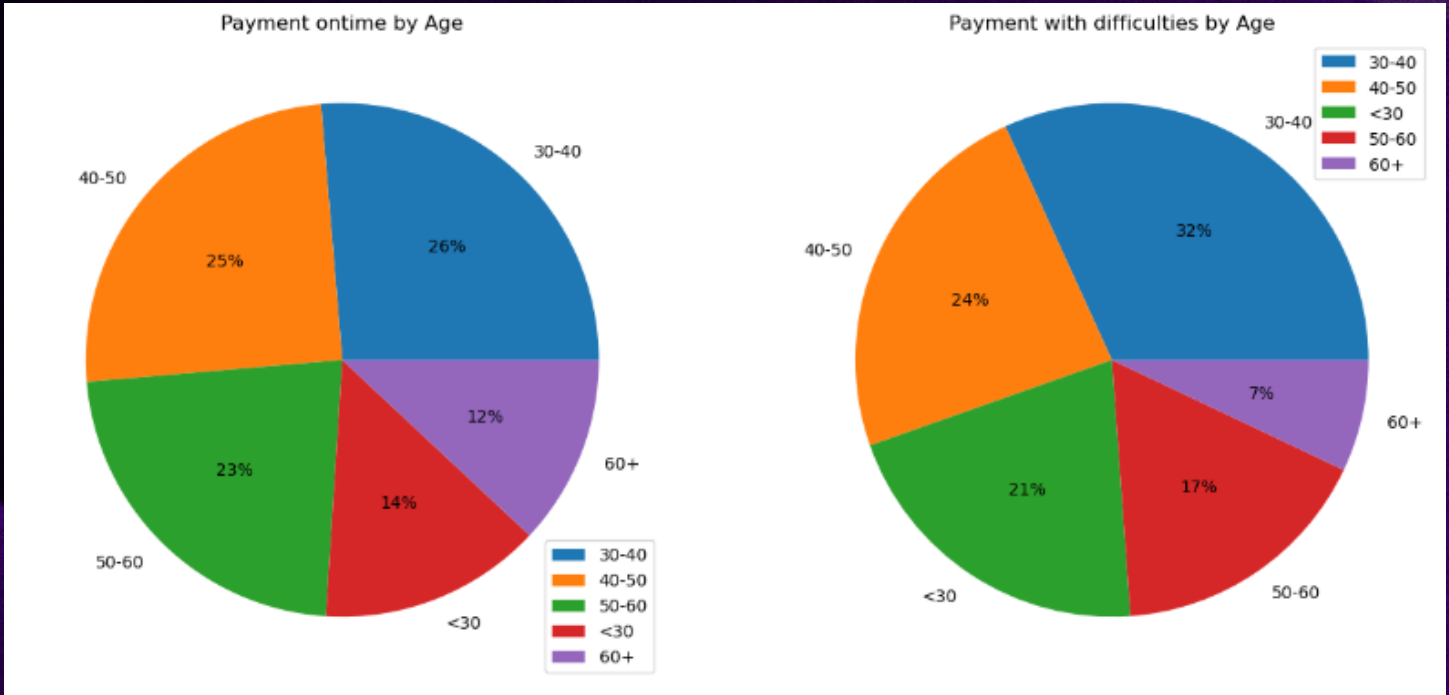
- Secondary/secondary special has more payment with difficulties.
- Higher education group has better on-time payment.

UNIVARIABLE ANALYSIS - CURRENT APPLICATION



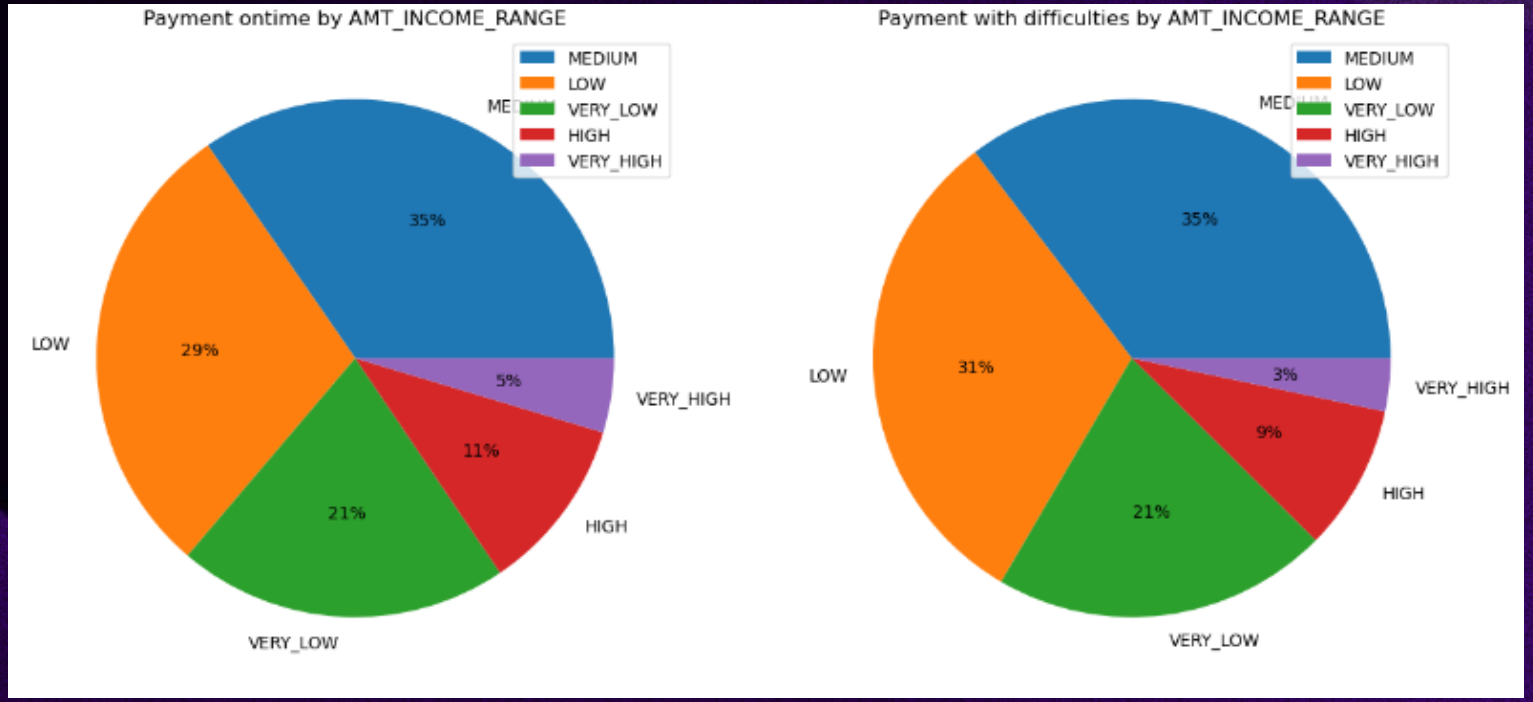
- Clients with working income type has more % of payment with difficulties compared to payment on-time.
- Pensioner has better on-time payment.

UNIVARIABLE ANALYSIS - CURRENT APPLICATION



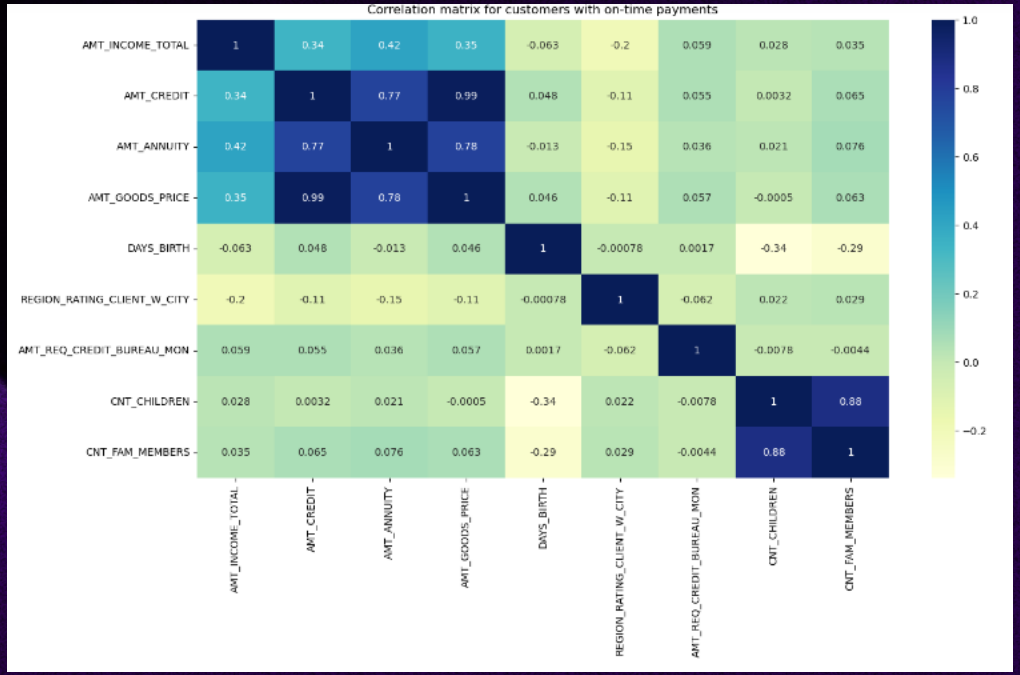
- Clients under 40 are more likely to have payment with difficulties
- Clients over 60 are more likely to have payment on-time

UNIVARIABLE ANALYSIS - CURRENT APPLICATION



— Clients who have high and very high incomes are supposed to make the payment on-time

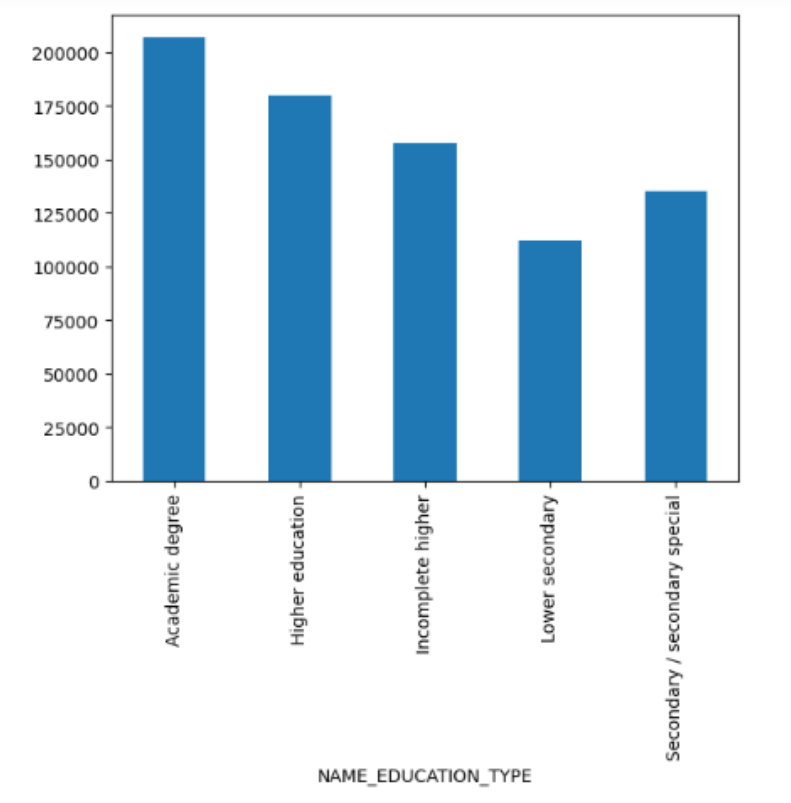
BIVARIABLE ANALYSIS - CURRENT APPLICATION



Quantify using correlation values and correlation heatmap

- As we can see, among customers with on-time payments, the correlation is very high between AMT_CREDIT and AMT_GOODS_PRICE with a correlation coefficient of 0.99.
- Also, there is a high correlation coefficient of 0.77 between AMT_CREDIT and AMT_ANNUITY, 0.78 between AMT_ANNUITY and AMT_GOODS_PRICE, 0.88 between CNT_CHILDREN and CNT_FAM_MEMBERS

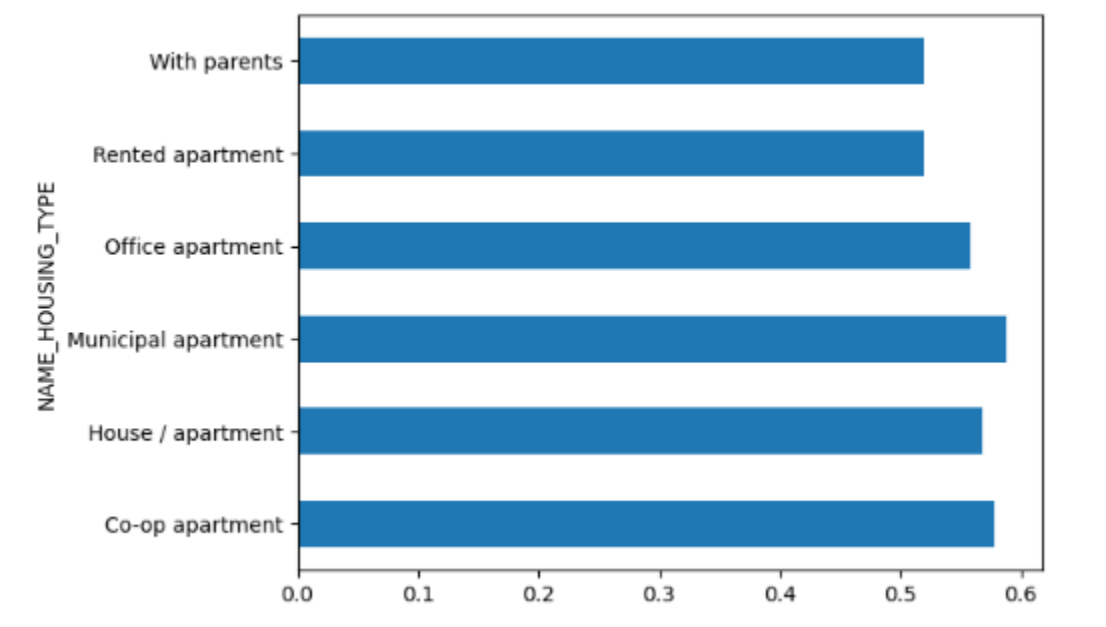
BIVARIABLE ANALYSIS - CURRENT APPLICATION



AMT_INCOME_TOTAL vs NAME_EDUCATION_TYPE

- Clients with academic degree have the highest income among others.
- Clients have Lower secondary recieve the least income in the data set.

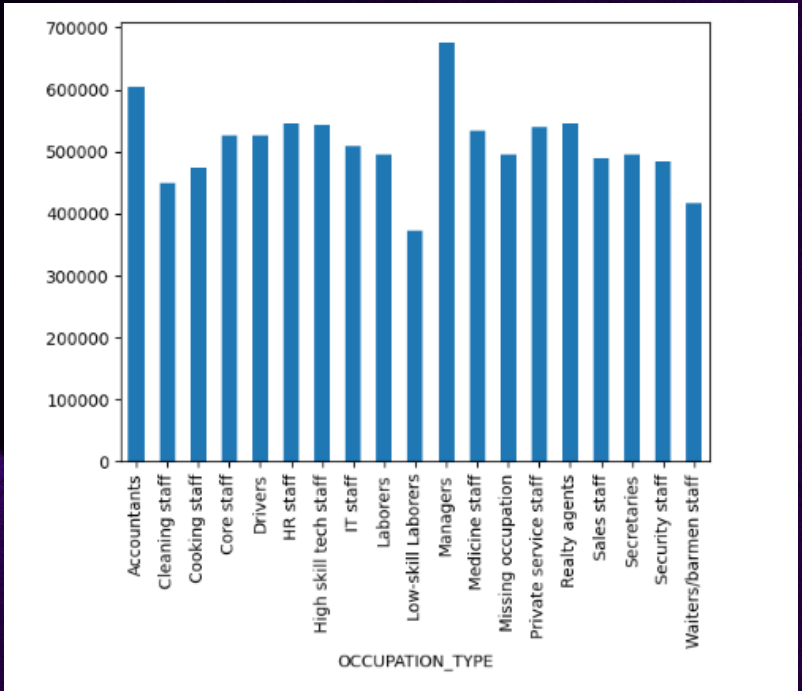
BIVARIABLE ANALYSIS - CURRENT APPLICATION



NAME_HOUSING_TYPE vs EXT_SOURCE_2

- Clients who have municipal apartment are rated highest points according to normalized score from external data source.
- Clients who live with their parents or live in rented apartment are Top 2 lowest rated.

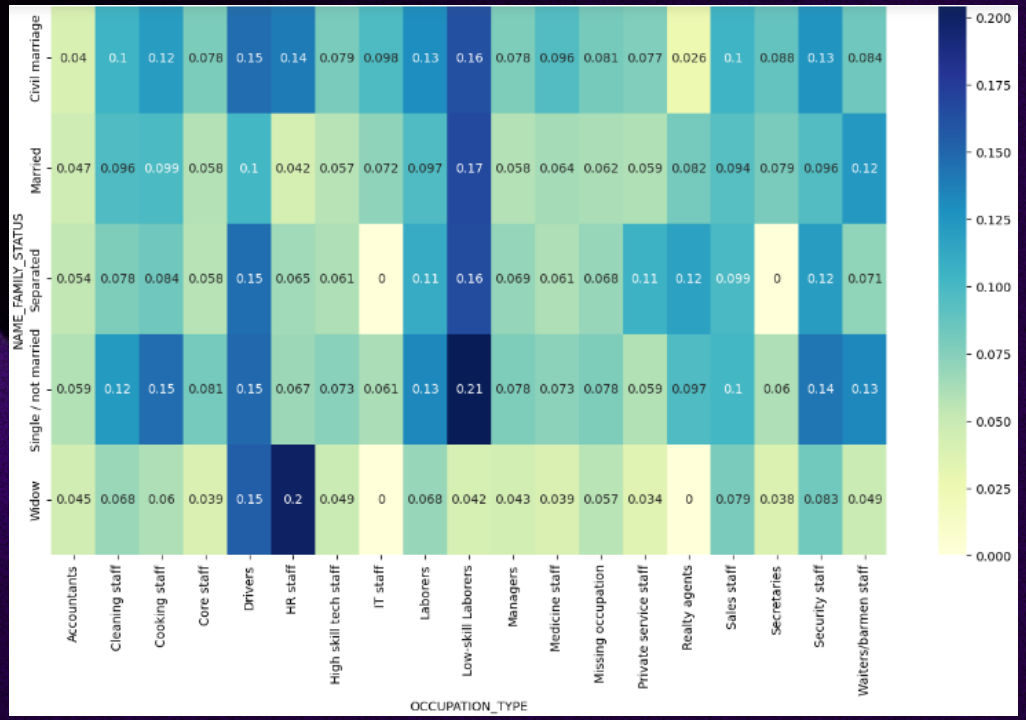
BIVARIABLE ANALYSIS- CURRENT APPLICATION



AMT_CREDIT vs OCCUPATION_TYPE

- Clients who are manager having the highest credit amount of the loan
- Clients who are Low-Skill Laborers and Waiters/barmen staff having the lowest credit amount of the loan.

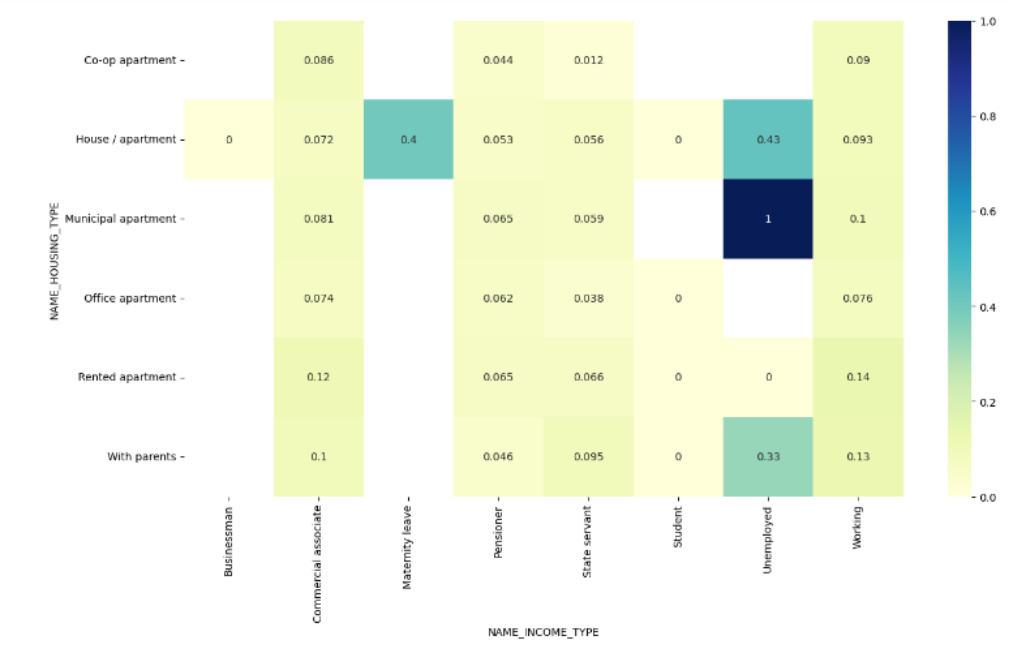
MULTIVARIABLE ANALYSIS - CURRENT APPLICATION



NAME_FAMILY_STATUS vs OCCUPATION_TYPE vs TARGET

- As 0 is customers with on-time payments and 1 is customers with payment difficulties, the higher correlation the more chance of customers have payment difficulties.
- Low-skill Laborers who are Single/Not married seems to be the group has the most payment difficulties. The same applies to HR staffs who are Widdow.
- Secretaries, IT staffs, accountants and Core staffs, are clients who have on-time payments.

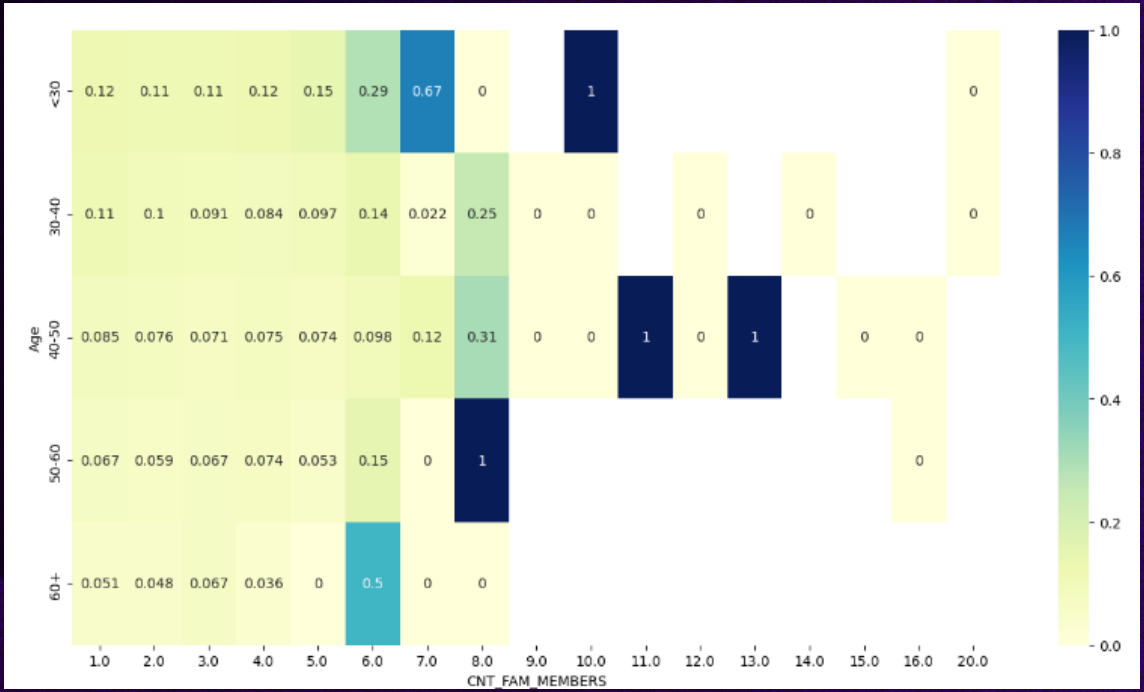
MULTIVARIABLE ANALYSIS - CURRENT APPLICATION



NAME_HOUSING_TYPE vs NAME_INCOME_TYPE vs TARGET

- Unemployed clients who own municipal apartment are very likely to have payment difficulties.
- Clients who are student has no trouble with payment on-time.

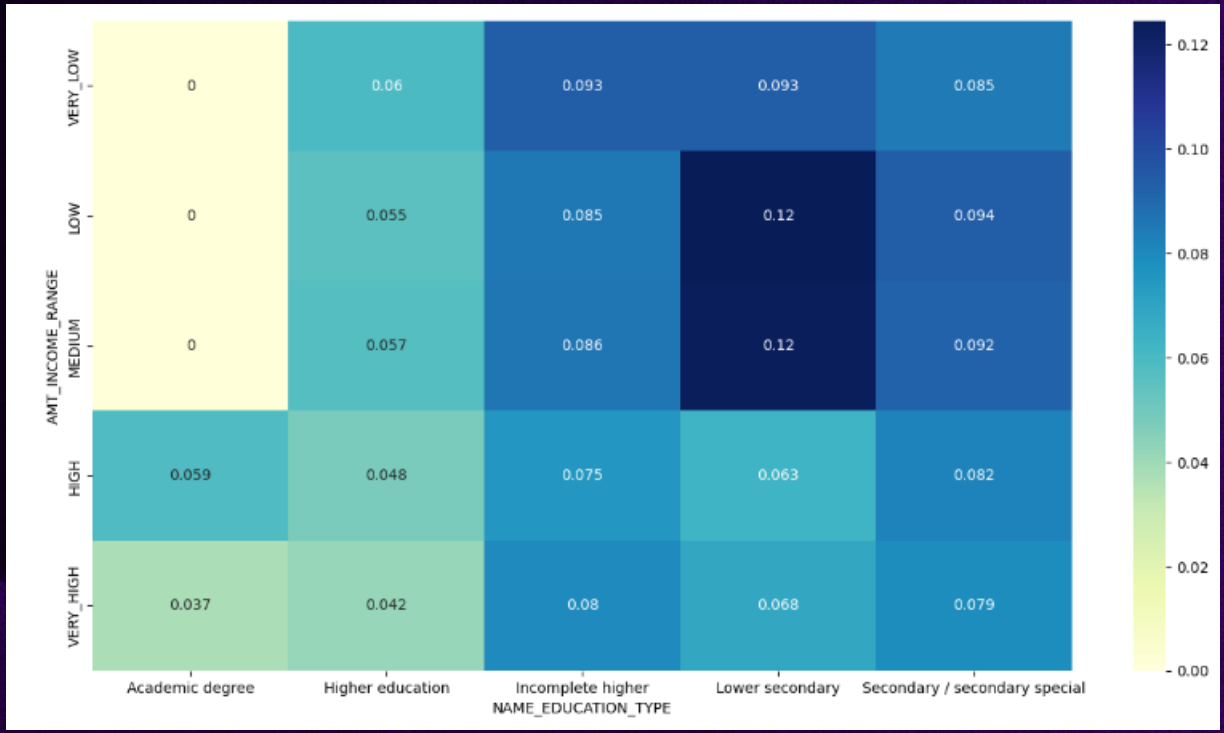
MULTIVARIABLE ANALYSIS - CURRENT APPLICATION



AGE vs CNT_FAM_MEMBERS vs TARGET

— Clients have less than 5 people in their family are more likely to make the payment on-time.

MULTIVARIABLE ANALYSIS - CURRENT APPLICATION

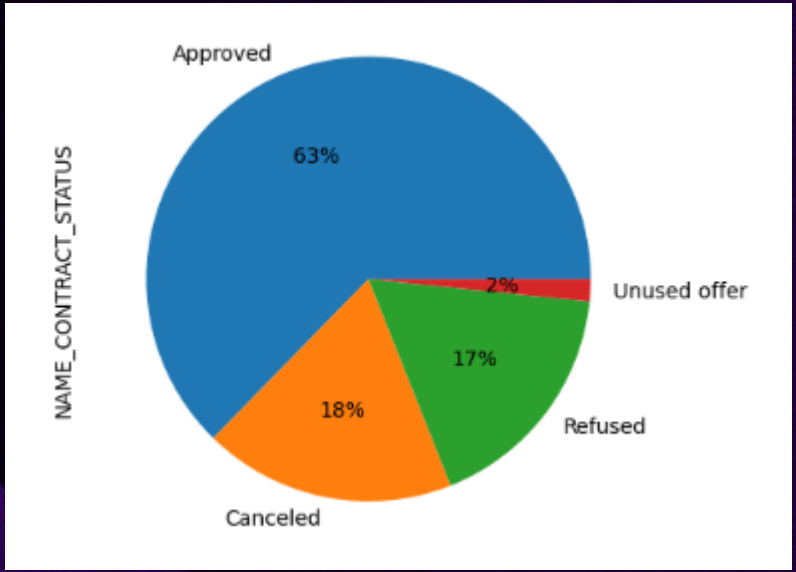


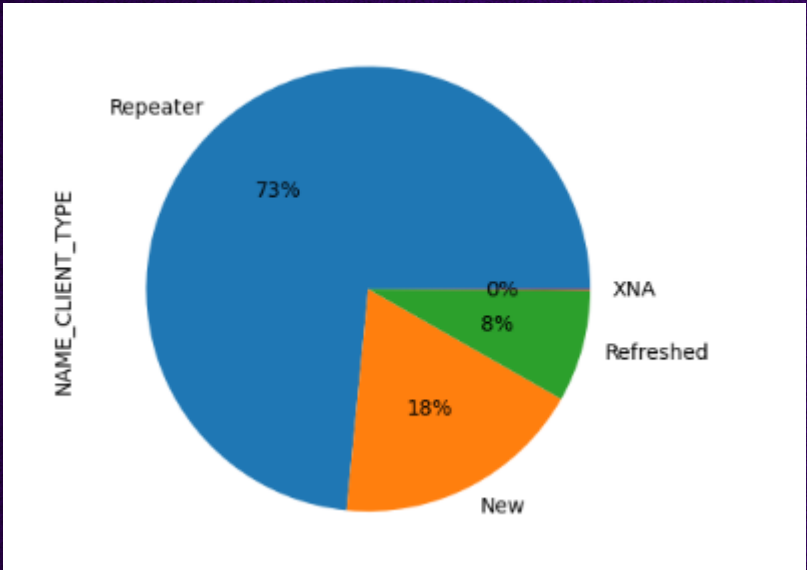
AMT_INCOME_RANGE vs NAME_EDUCATION_TYPE vs TARGET

Clients who holds Academic degree and earn very low - medium income (80% quantiles) are least likely to have payment difficulties.

UNIVARIABLE ANALYSIS – MERGE APPLICATION

I will merge the current application and previous application data sets

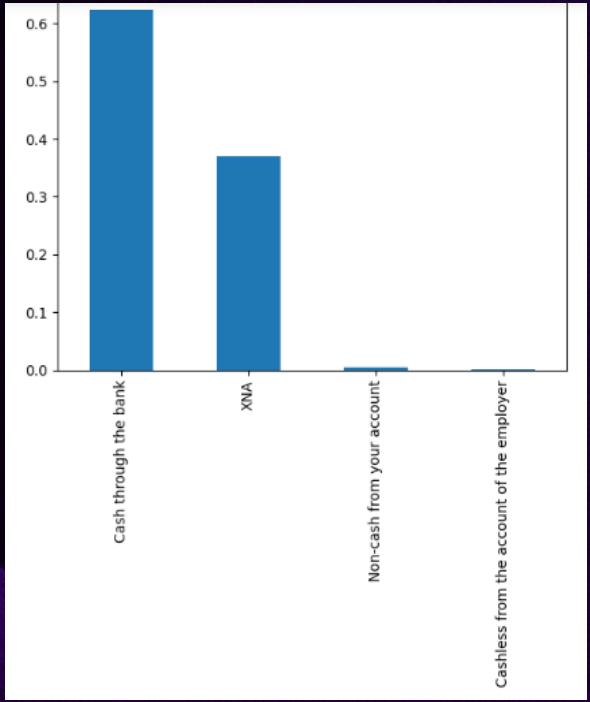


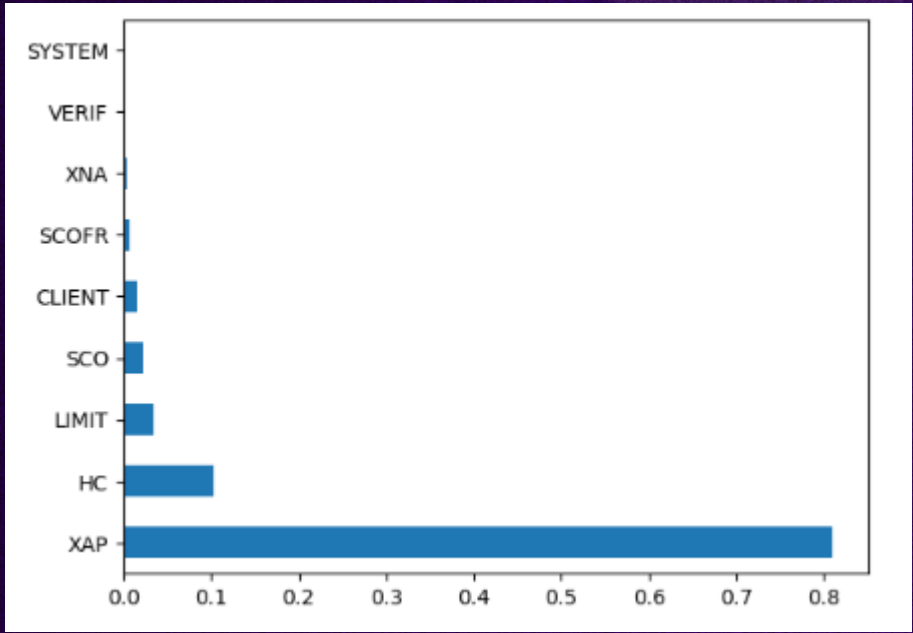


- Among clients who applied for the previous loan, 63% are approved and 17% are refused by the bank.

- 73% of the clients are repeaters when applying for the previous application
- 8% of the clients are refreshed when applying for the previous application

UNIVARIABLE ANALYSIS – MERGE APPLICATION

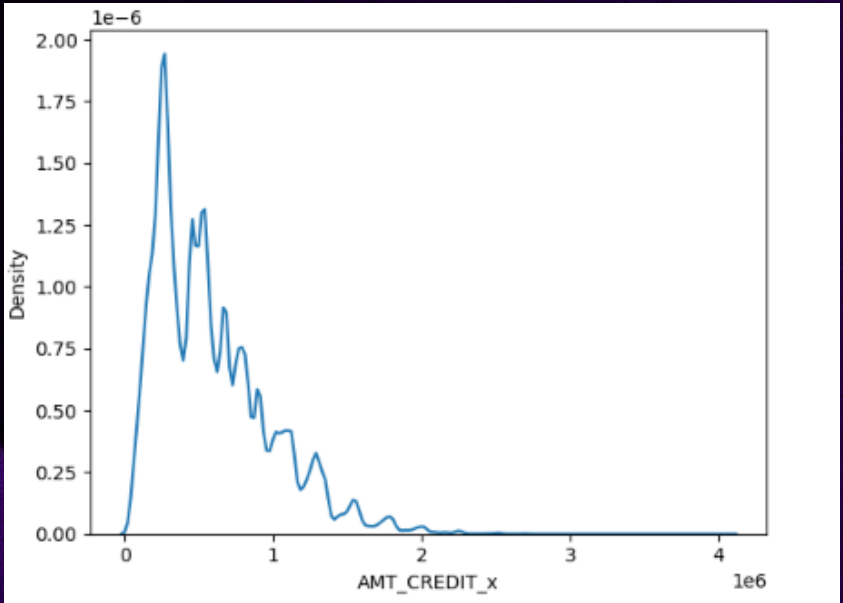


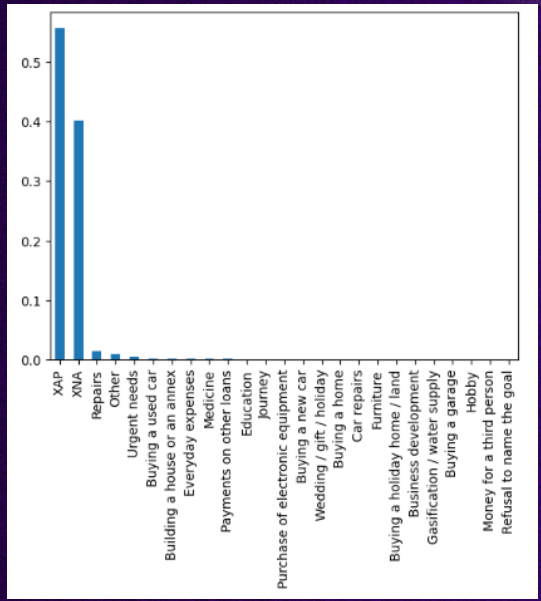


- About 60% of Clients chose to pay cash through the bank for the previous application

- XAP is 80% of the reason why the previous application was rejected

UNIVARIABLE ANALYSIS – MERGE APPLICATION

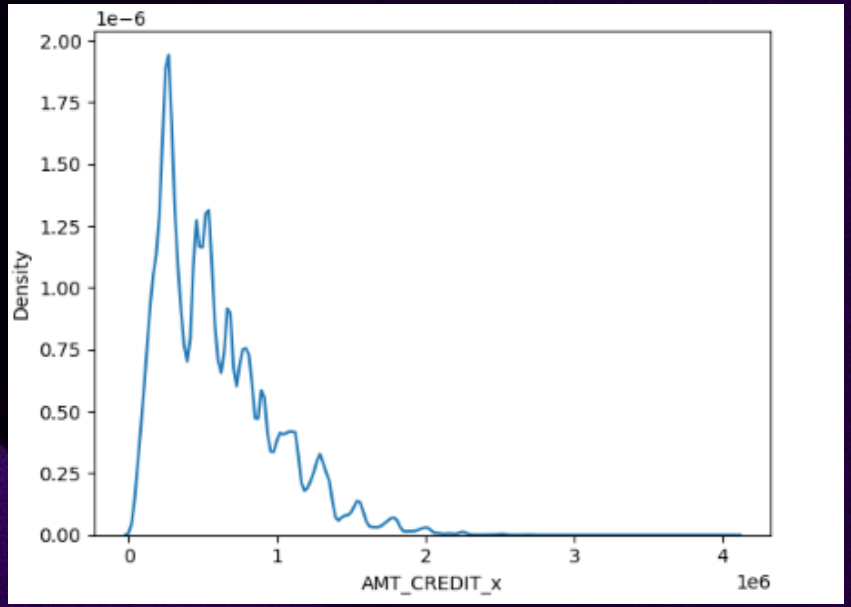


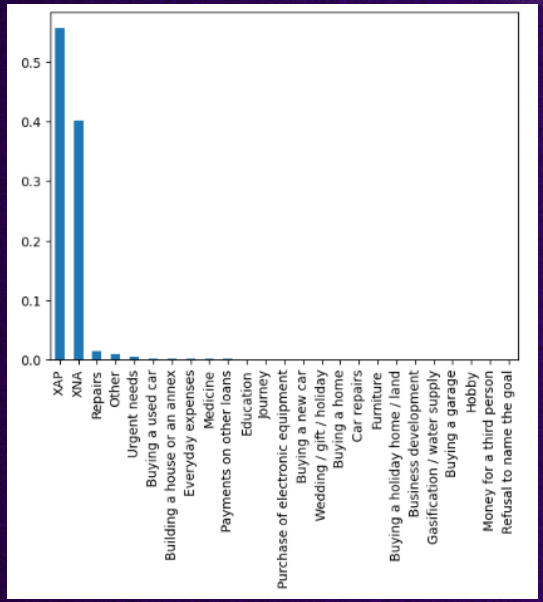


- The distribution of the credited amount of the loan was mostly in range from 0 - 1,000,000

- More than 50% of purpose of the cash loan is XAP

UNIVARIABLE ANALYSIS – MERGE APPLICATION

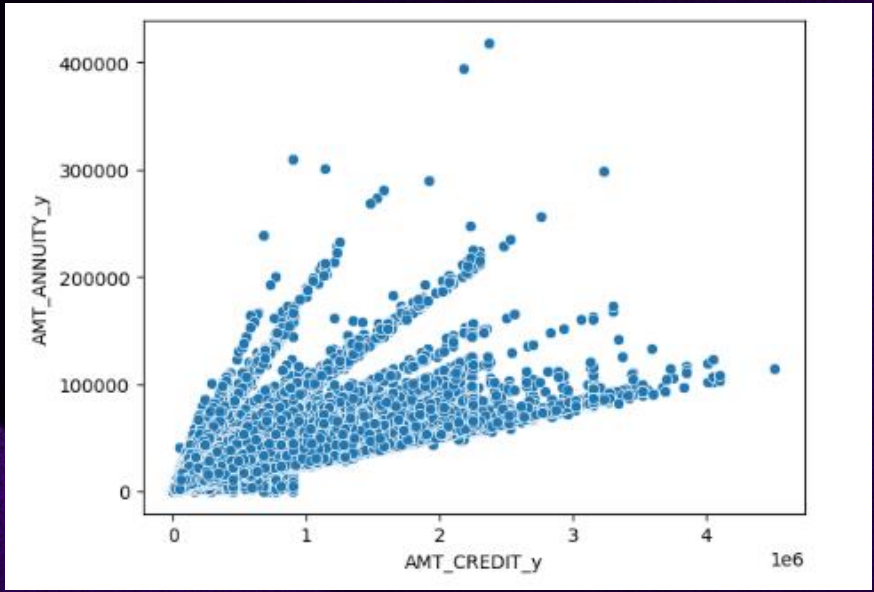




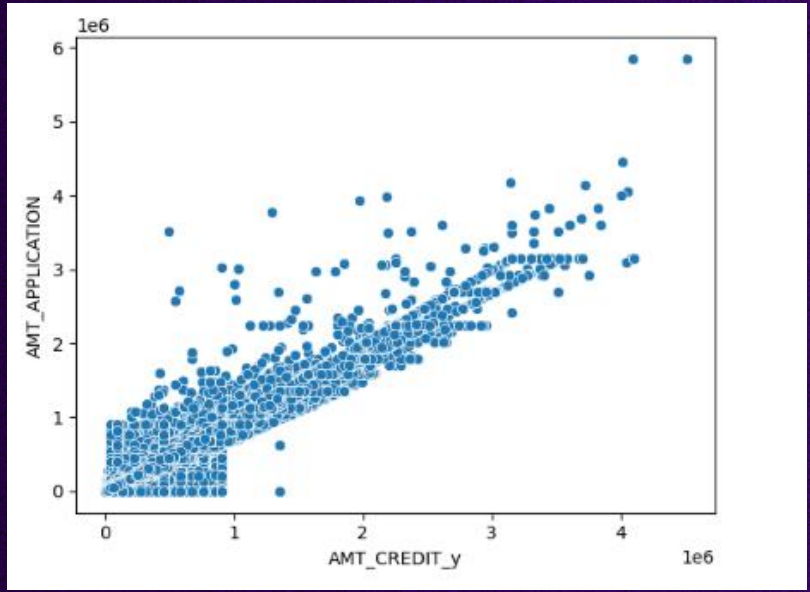
- The distribution of the credited amount of the loan was mostly in range from 0 - 1,000,000

- More than 50% of purpose of the cash loan is XAP

BIVARIABLE ANALYSIS – MERGE APPLICATION

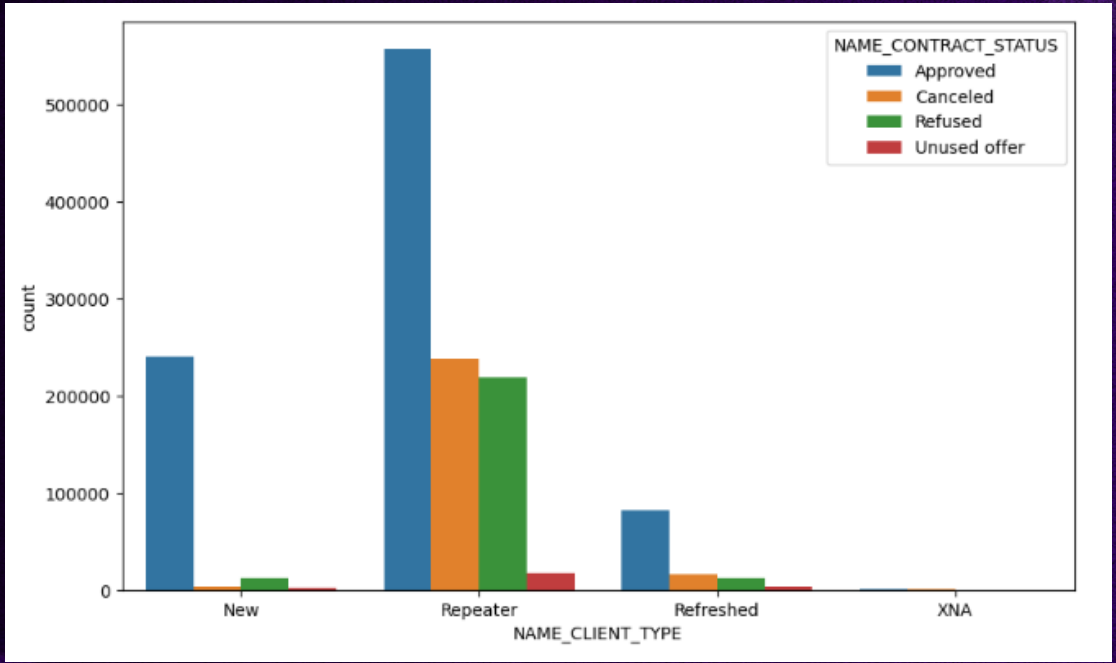


- AMT_CREDIT_x and AMT_ANNUITY_x have a positive linear correlation.



- AMT_CREDIT_y and AMT_APPLICATION have a positive linear correlation.

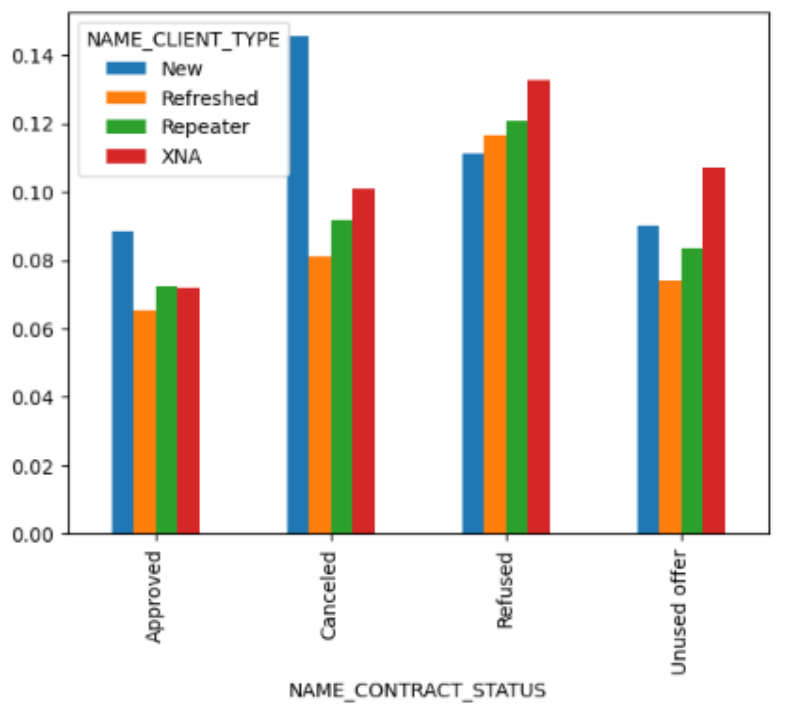
BIVARIABLE ANALYSIS – MERGE APPLICATION



NAME_CLIENT_TYPE vs NAME_CONTRACT_STATUS

- The repeater clients have more approved loans than new and refreshed clients.

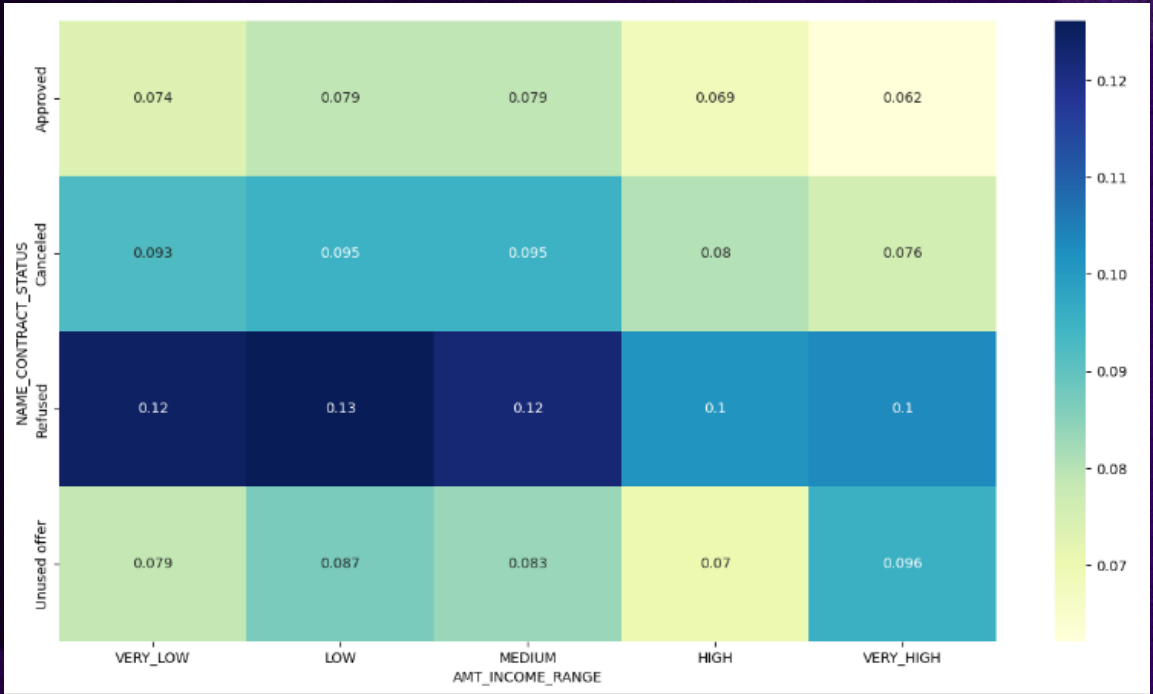
MULTIVARIABLE ANALYSIS – MERGE APPLICATION



NAME_CONTRACT_STATUS vs NAME_CLIENT_TYPE vs TARGET

Among the clients whose contracts got approved, Repeaters and Refreshed customers are more likely to have payment on-time than New customers.

MULTIVARIABLE ANALYSIS – MERGE APPLICATION



NAME_CONTRACT_STATUS vs AMT_INCOME_RANGE vs TARGET
Clients whose contracts are approved have the highest chance of paying the loan on time no matter what their income ranges are.

04

RECOMMENDATION

RECOMMENDATION

- The bank should avoid giving loan to clients who are under 40
- The bank should focus on giving loan to clients who are over 60
- The bank should focus on customers with Pensioner income type and avoid customers with working income type
- The bank should avoid giving loan to Unemployed clients who own municipal apartment
- The bank should avoid customers who are Low-skill Laborers and Single/Not married
- The bank should focus on giving loan to clients who hold academic degree and earn very low - medium income (80% quantiles)
- The bank should focus on giving loan to the Repeaters and Refreshed customers whose contract got approved in the previous application.