Credit EDA Assignment Case Study

- **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):
- **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, 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**

Download the dataset from below.

- **Dataset**

This dataset has 3 files as explained below:

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

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

- **Problem Statement - II**
- **Results Expected by Learners**

Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.

Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)

Hint: Note that in EDA, since it is not necessary to replace the missing value, but if you have to replace the missing value, what should be the approach. Clearly mention the approach.

Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.

Identify if there is data imbalance in the data. Find the ratio of data imbalance.

Hint: How will you analyse the data in case of data imbalance? You can plot more than one type of plot to analyse the different aspects due to data imbalance. For example, you can choose your own scale for the graphs, i.e. one can plot in terms of percentage or absolute value. Do this analysis for the **'Target variable'** in the dataset (**clients with payment difficulties** and **all other cases**). Use a mix of univariate and bivariate analysis etc.

Hint: Since there are a lot of columns, you can run your analysis in loops for the appropriate columns and find the insights.

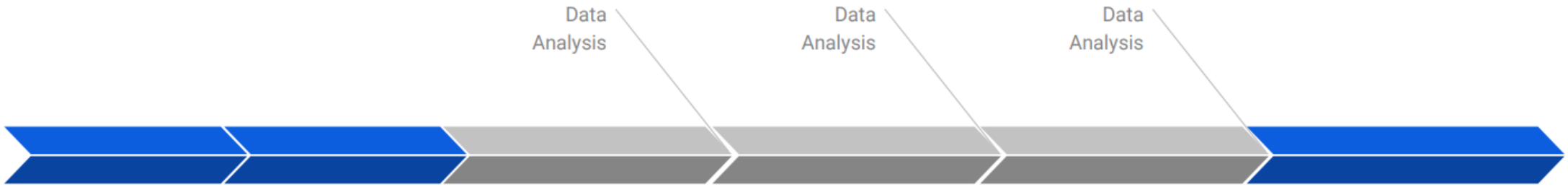
Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.

Find the top 10 correlation for the **Client with payment difficulties** and **all other cases** (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: **Var1, Var2, Var3, Var4, Var5, Target**. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.

Include visualisations and summarise the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the **clients with payment difficulties with all other cases.**

You need to submit one/two Ipython notebook which clearly explains the thought process behind your analysis (either in comments of markdown text), code and relevant plots. The presentation file needs to be in PDF format and should contain the points discussed above with the necessary visualisations. Also, all the visualisations and plots must be done in Python(should be present in the Ipython notebook), though they may be recreated in Tableau for better aesthetics in the PPT file.

Problem solving methodology



Data Cleaning

Removing the null valued columns, unnecessary variables and checking the null value percentage and removing the respective rows.

Data Understanding

Working with the Data Dictionary and getting knowledge of all the columns and their domain specific uses

Univariate Analysis

Analysing each column, plotting the distributions of each column.

Segmented Univariate Analysis

Analysing the continuous data columns with respect to the categorical column

Bivariate Analysis

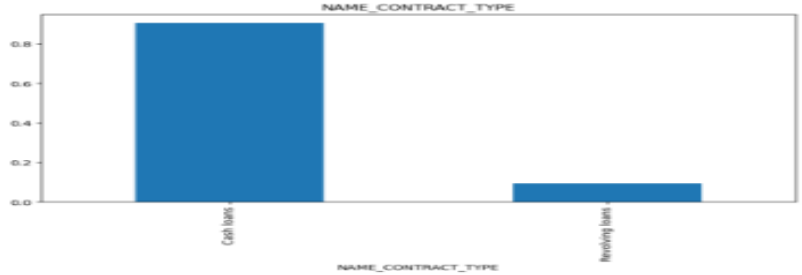
Analysing the two variable behaviour like term and loan status with respect to loan amount.

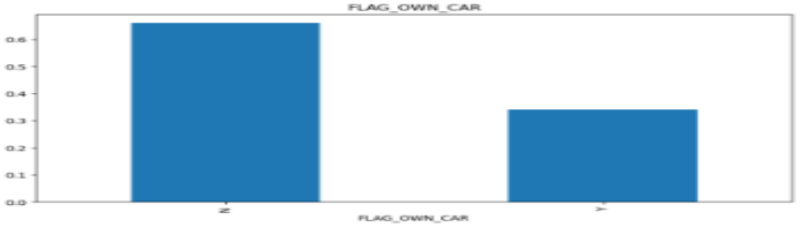
Recommendations

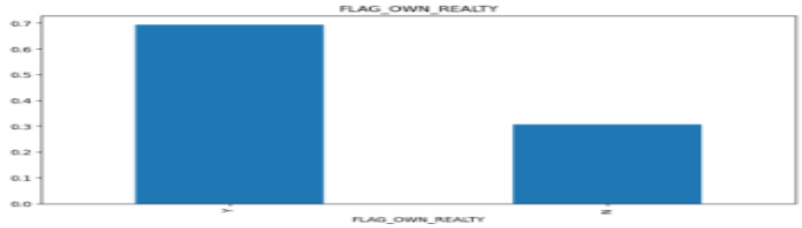
Analysing all plots and recommendations for reducing the loss of business by detecting columns best which contribute to loan defaulters.

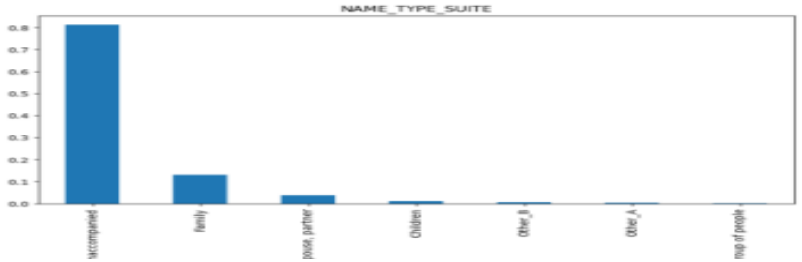
Application Dataset

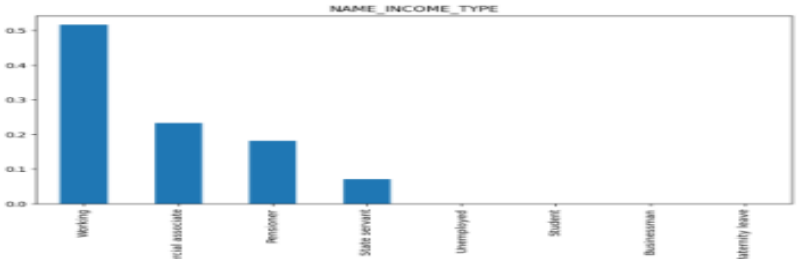
Univariate Analysis :-

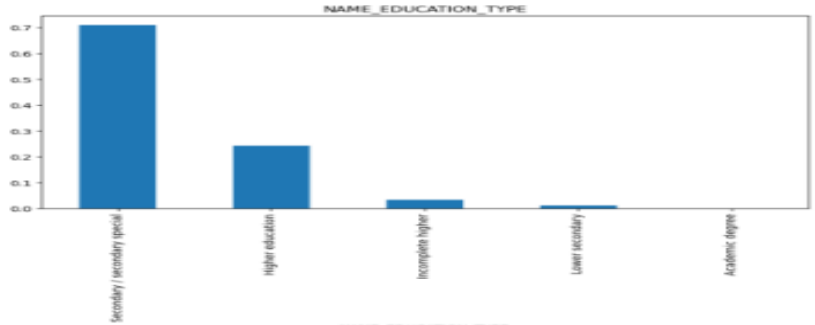


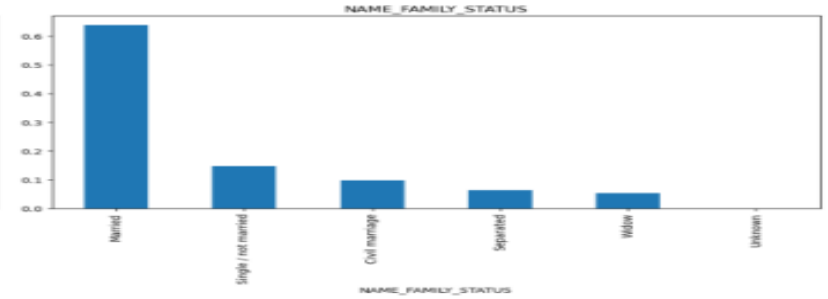


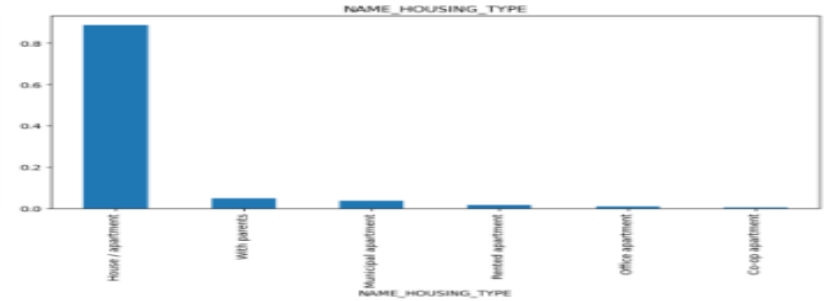


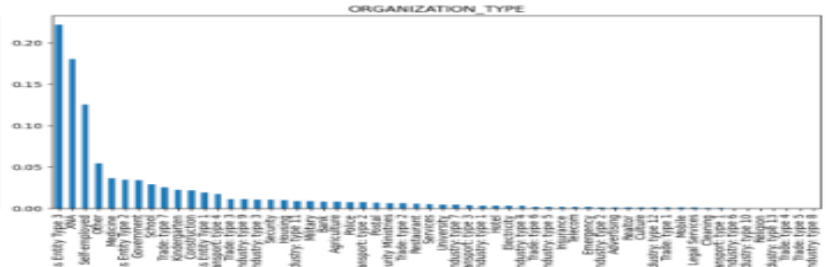












Point to be noticed:

1. cash loans offer more than revolving loans around 88% more.

2. more than male, female taking loans. So we can contact more to females for the loans as company.

3. almost 65% people are not owning the cars. So we can offer them car loan.

4. most people living there own house or apartment, So we can morgagae the property on behalf of loans.

5. around 81 % unaccompanied came for loan process.

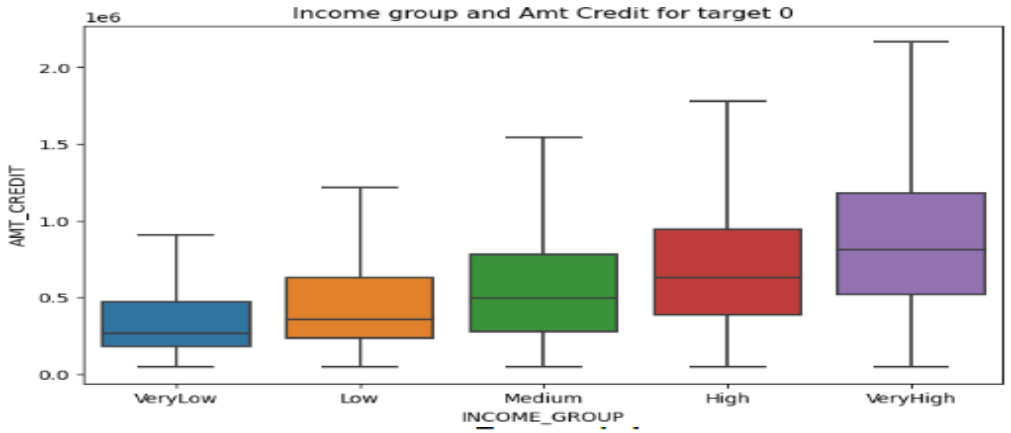
6. 51% people are working, 23% are Commercial associate, 18% pensioner, 7% state serveant and remaining other apply for the loan.

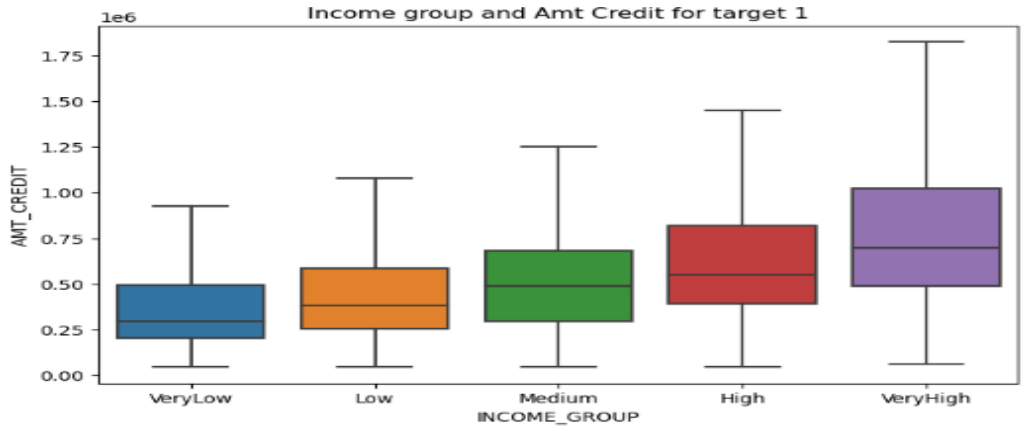
7. 71% have secondary and 24% have higher education.

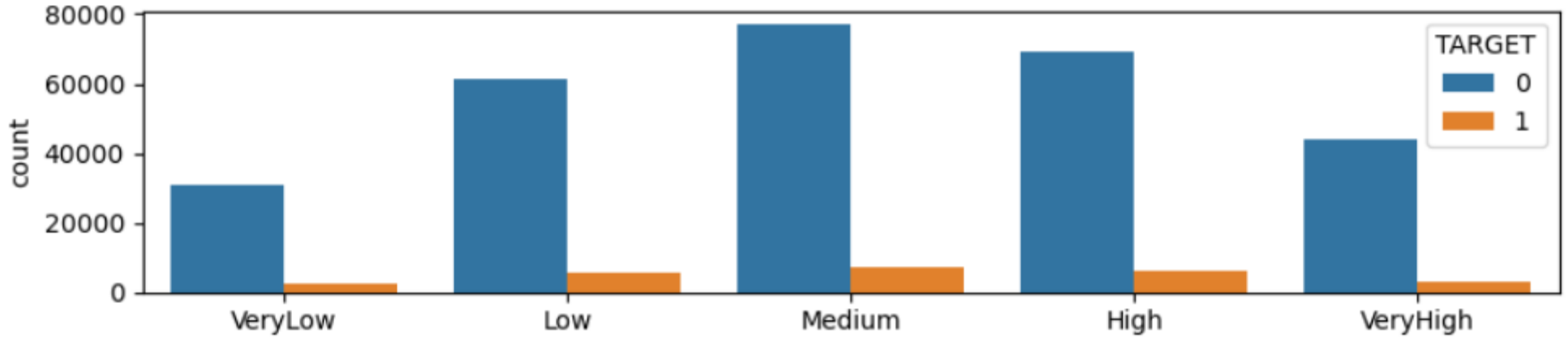
8. 64% are married and 15% are not married.

Bivariate Analysis for application dataset :-

Checking INCOME_GROUP vs AMT_CREDIT for both data frames



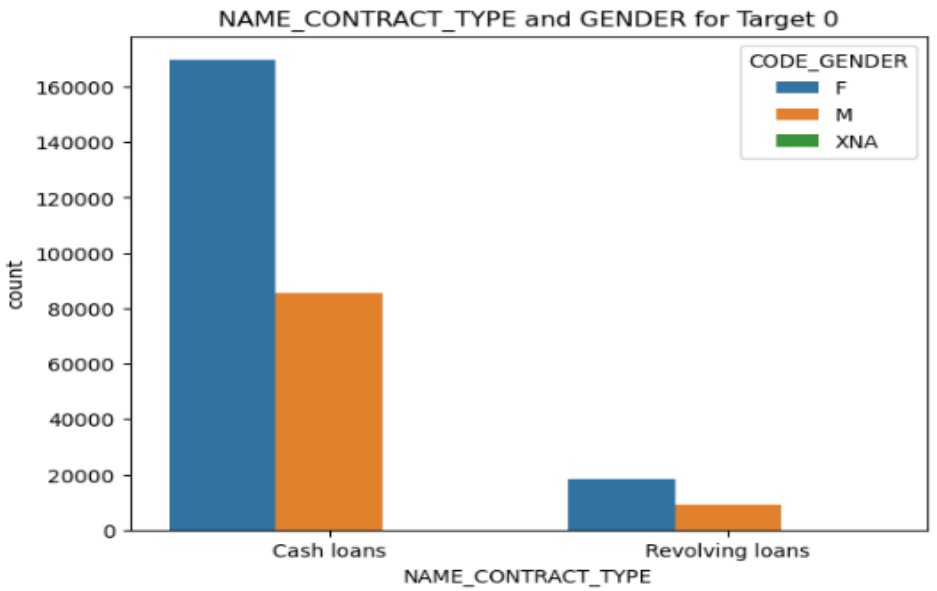


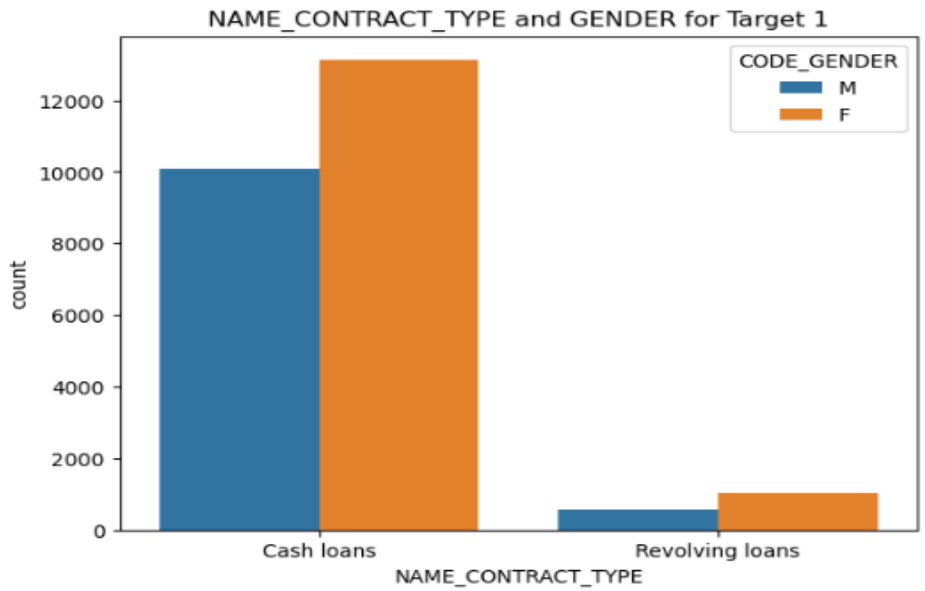


Point to be noticed :-

1. Here I can see maxium number of loan taken by medium group.
2. But the value per loan is higher for high and veryhigh group as AMT_CREDIT is higher to them.
3. Therfore the company get affected due to higher amount not being paid back.
4. Company needs to change policies for higher income group loans.

Bivariate categorical and categorical



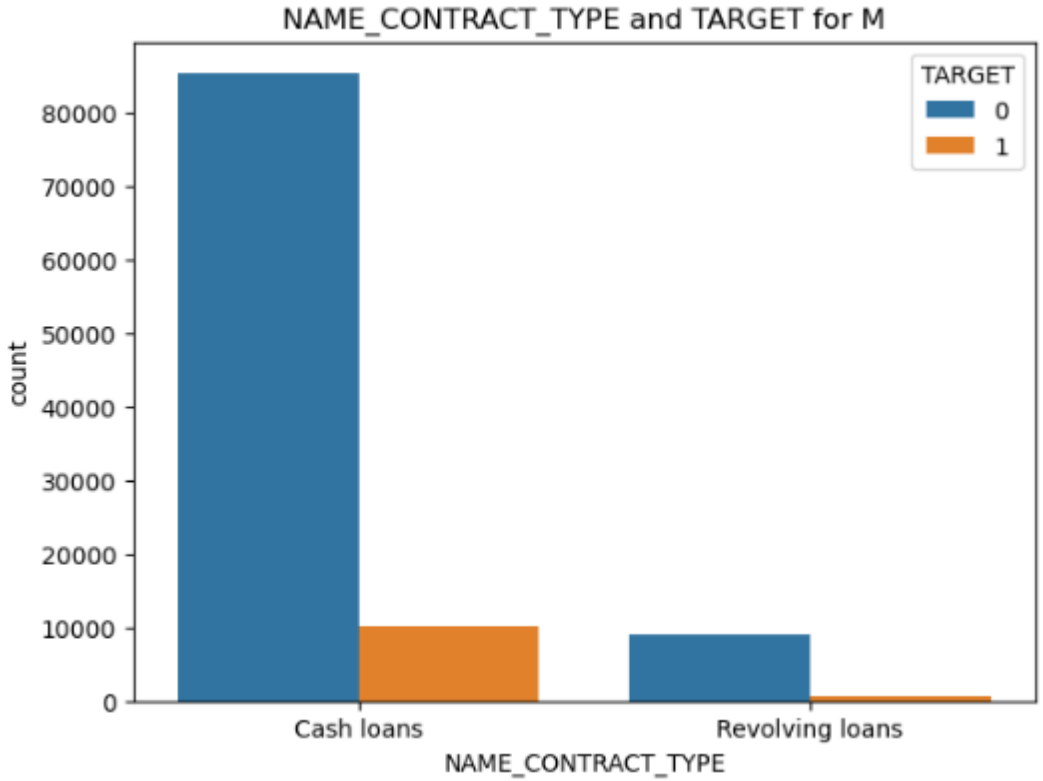


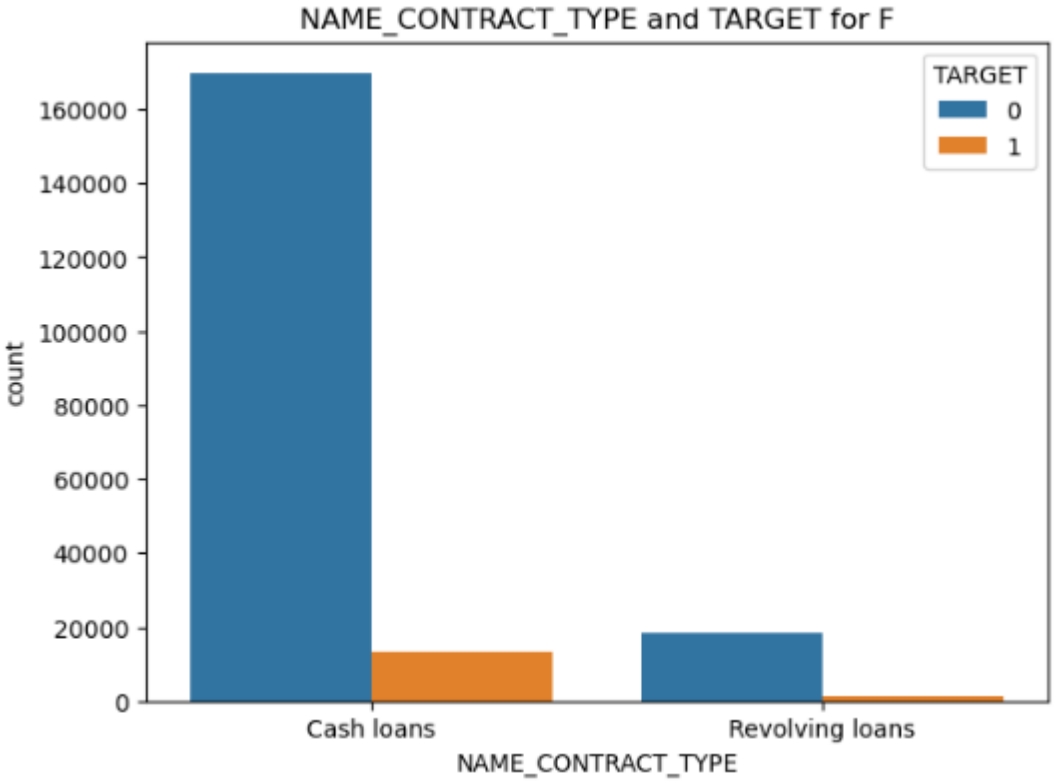
- Point to be notice :¶

1. As noted above data has more females as loan applicant.

2. As seen in plot above, though male applicants are lower, ratio of male applicants deafulting is higher.

Let us check this by another analysis

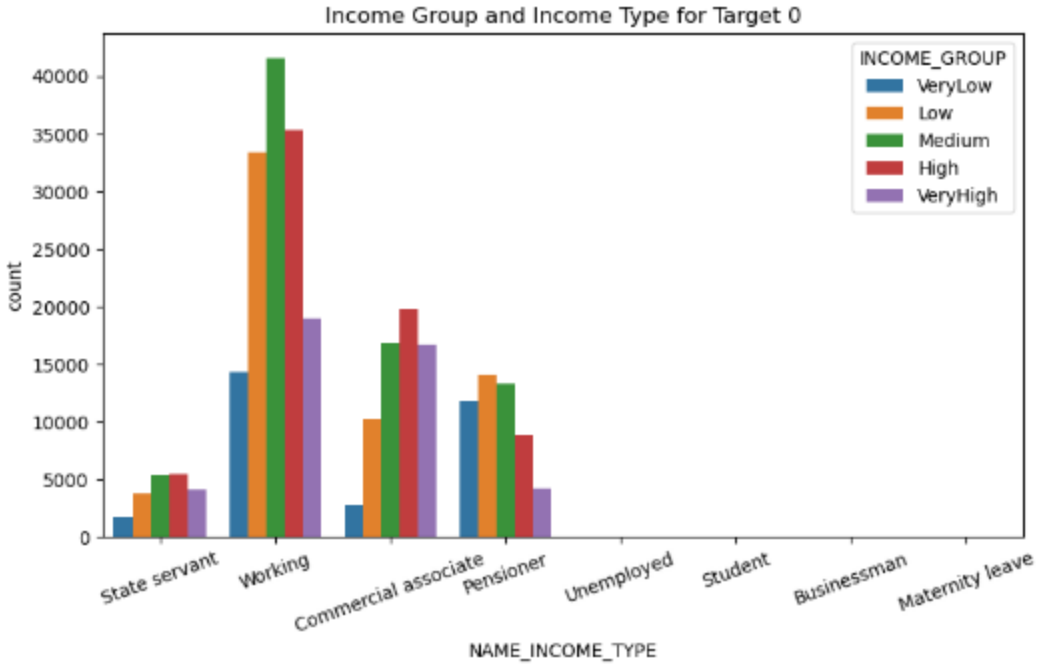


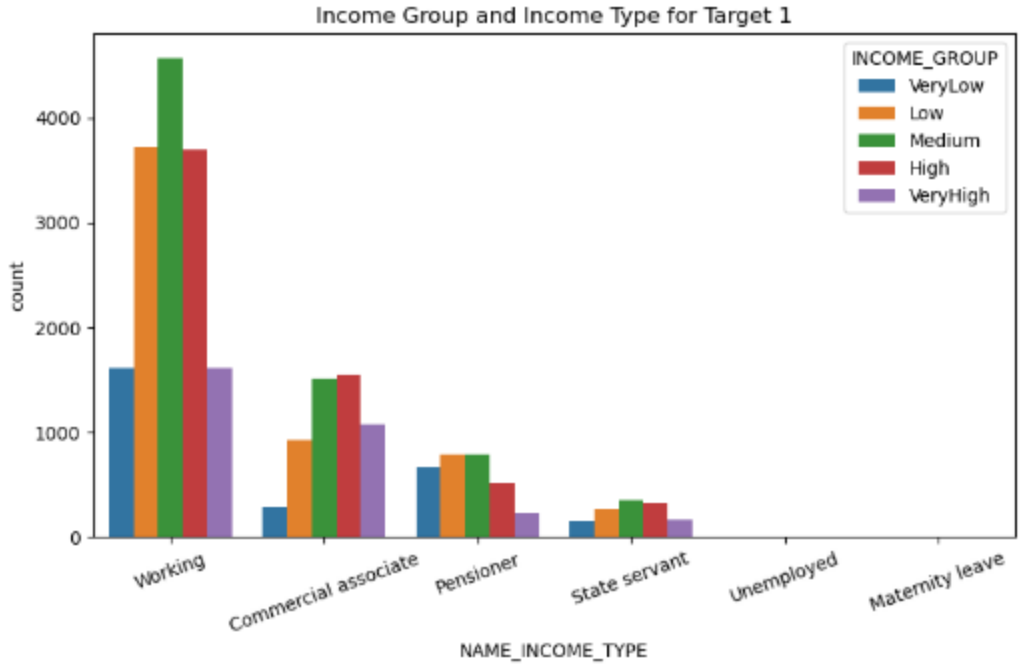


- **Point to be notice :**

- Male applicants are defaulting more than female applicants.

checking NAME_INCOME_TYPE vs INCOME_GROUP for both target =0 & target = 1





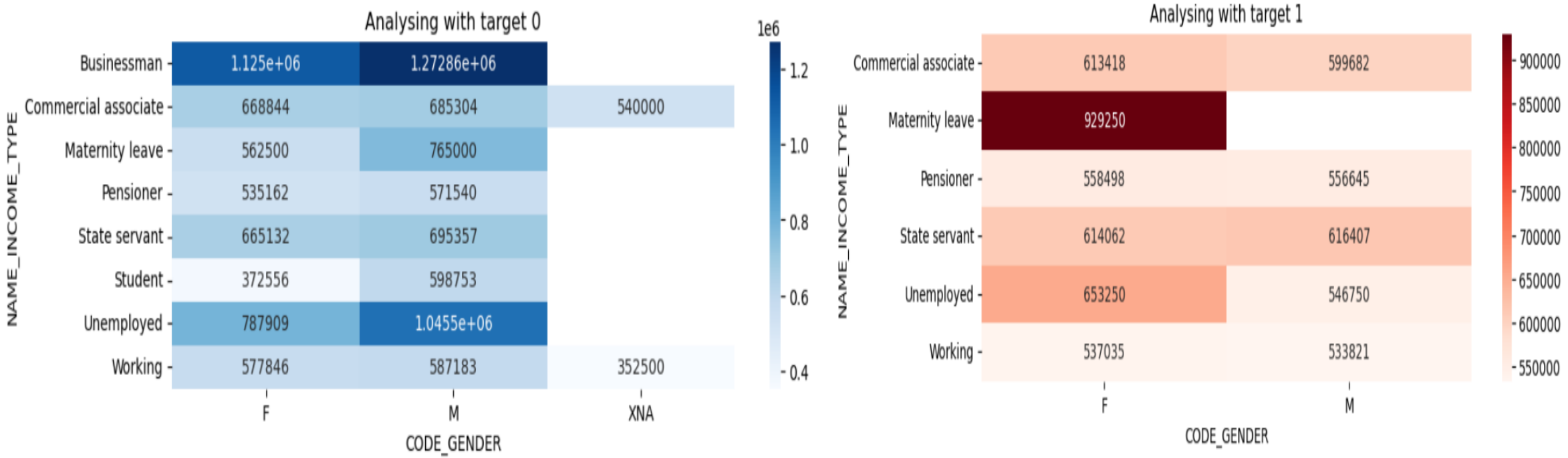
- **Point to notice:**

1 . Medium Income Group who is working applies for loans most of the time.

2. Medium income group with income type has almost 1 in 12 defaults

Multivariate Analysis for application dataset:-

Analysing relationship of AMT_CREDIT with CODE_GENDER and NAME_INCOME_TYPE type for Target=0 and Target=1



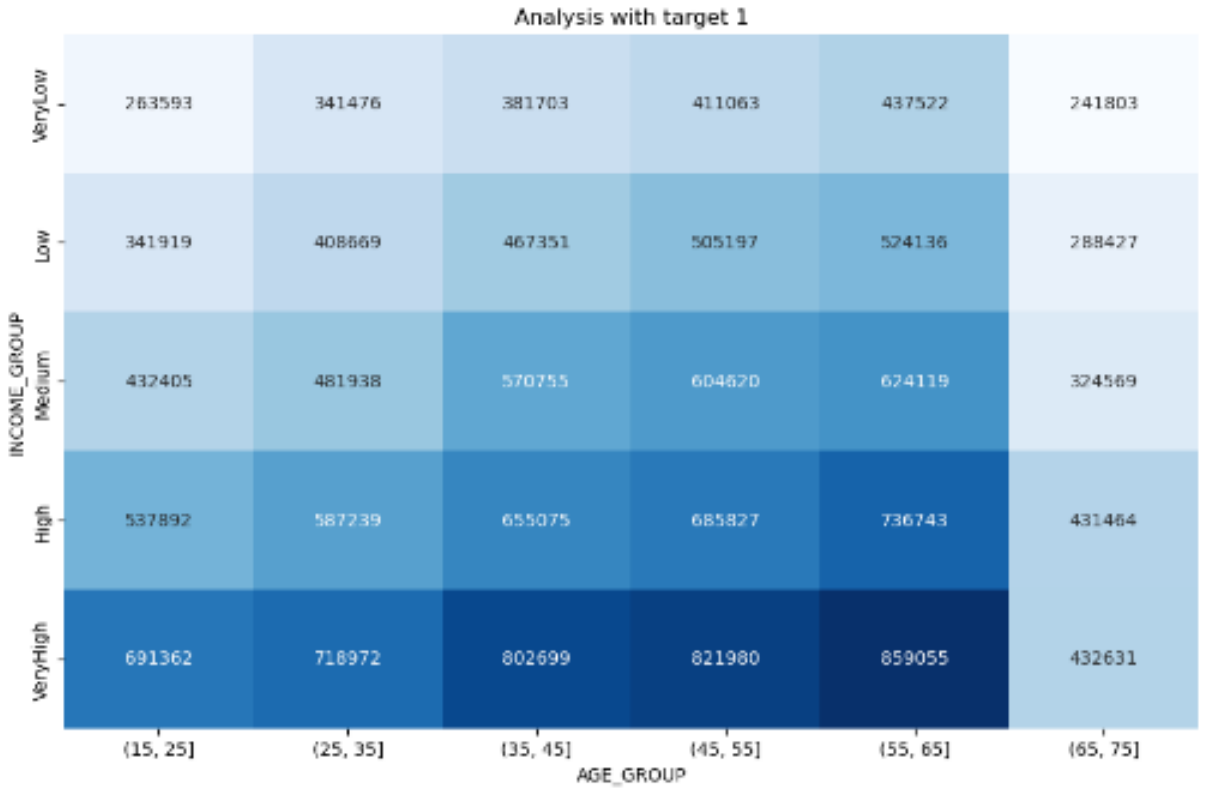
• Point to be notice:

1. Businessman in both gender credited larger loans and have no issues with loan payment¶

2. Female on Maternity leaves have larger loan credited and also they have payment issues i.e defaulting there loans.

3. for Unemployed both females and males have 2nd highest loan credited

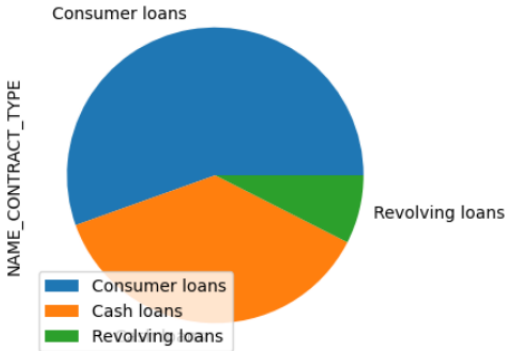
Analysing relationship of INCOME_GROUP and AGE GROUP with AMT_CREDIT



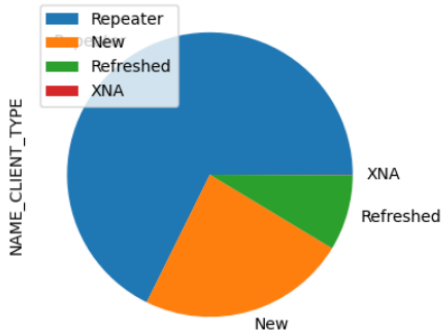
- **Point to be notice :**
- **1. Age Group of 55-65 have higher amount of loan credited.**
- **2. Need to focus on higher loans credited as there are lots of customers having payment issues.**
- **3. Company needs to change the policies and rules for higher group of loans.**
- **4. There is less amount are loans credited for the group of 65-75 and which is true.**

Previous Application Dataset Univariate Analysis :-

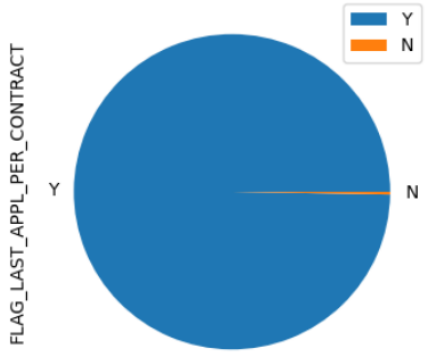
```
Consumer loans     55.477887
Cash loans         37.034068
Revolving loans     7.488045
Name: NAME_CONTRACT_TYPE, dtype: float64
```



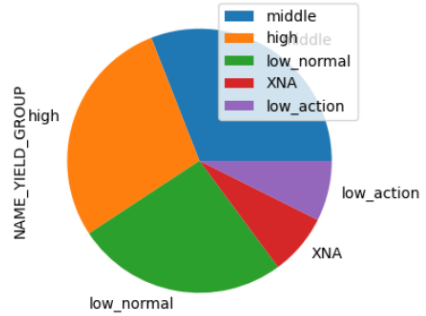

```
Repeater     67.706608
New          23.633417
Refreshed     8.587682
XNA           0.072293
Name: NAME_CLIENT_TYPE, dtype: float64
```



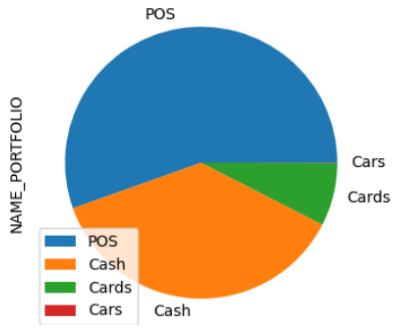

```
Y    99.722222
N     0.277778
Name: FLAG_LAST_APPL_PER_CONTRACT, dtype: float64
```



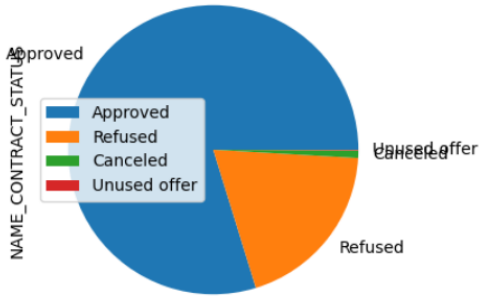

```
middle        30.933548
high          28.349702
low_normal    25.843684
XNA            7.488045
low_action     7.385022
Name: NAME_YIELD_GROUP, dtype: float64
```



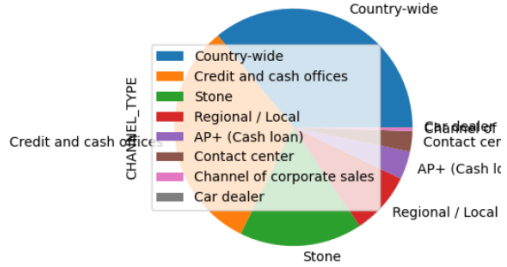

```
POS      55.443787
Cash     37.034068
Cards     7.488045
Cars      0.034100
Name: NAME_PORTFOLIO, dtype: float64
```



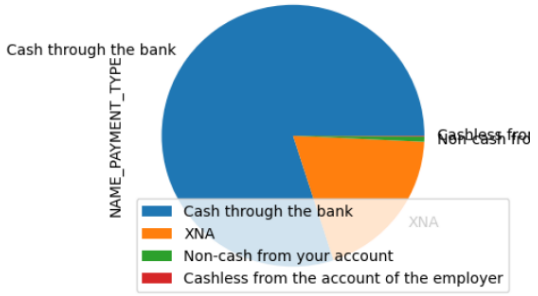

```
Approved        79.749823
Refused         19.334441
Canceled         0.842721
Unused offer     0.073015
Name: NAME_CONTRACT_STATUS, dtype: float64
```




```
Country-wide                35.930178
Credit and cash offices     31.792557
Stone                       16.664500
Regional / Local             8.503113
AP+ (Cash loan)              3.808171
Contact center               2.783555
Channel of corporate sales   0.481658
Car dealer                   0.036267
Name: CHANNEL_TYPE, dtype: float64
```




```
Cash through the bank                        79.990291
XNA                                          19.271937
Non-cash from your account                    0.651919
Cashless from the account of the employer     0.085853
Name: NAME_PAYMENT_TYPE, dtype: float64
```



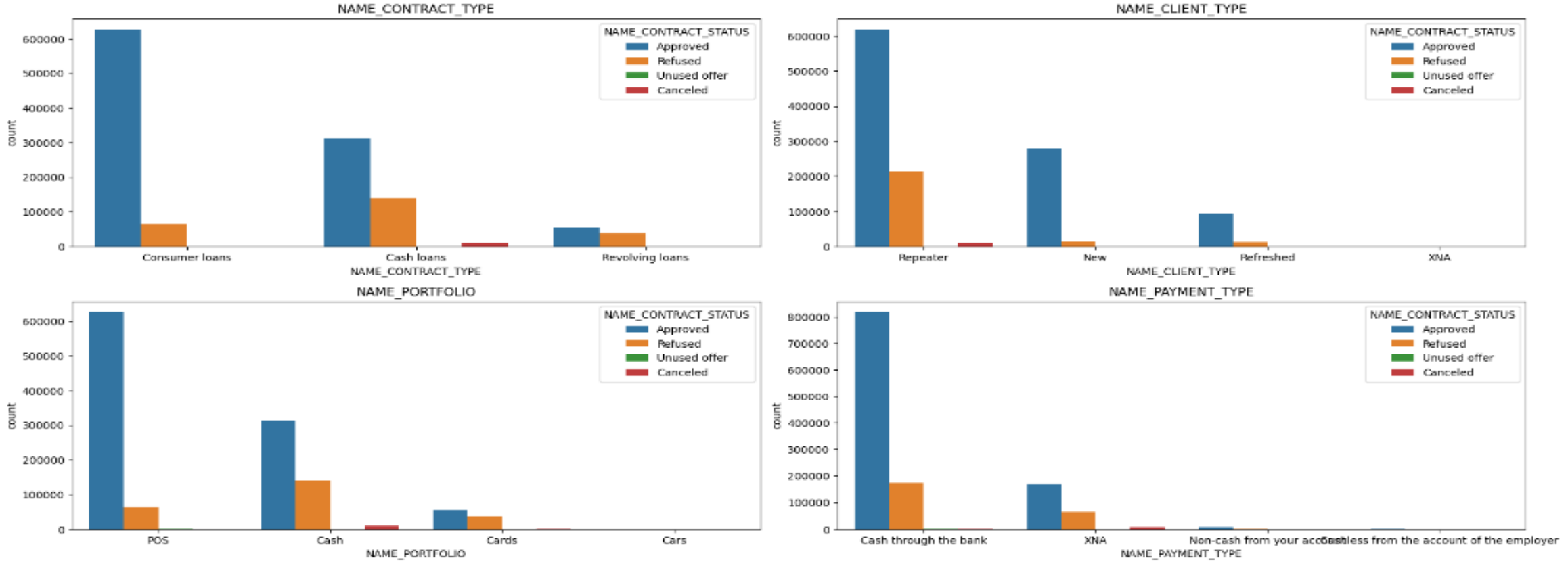
Point to be noticed here :

1. there is 55% of consumer loan , 37% of car loans and rest is revolving loans.
2. Approved loans are almost 80% and rest is other like refuse, cancel or unused.
3. There is 67% of new repeat client and 23% new client.
4. There is 55% of loan taken for POS purchase and 37% in cash.
5. country wide and credit & cash offices are the place where most of the loan takes place

Bivariate Analysis :-

Categorical vs Categorical

Analysis for which loans are approved or not along client type , porfolio, etc



- **point to be notice :**

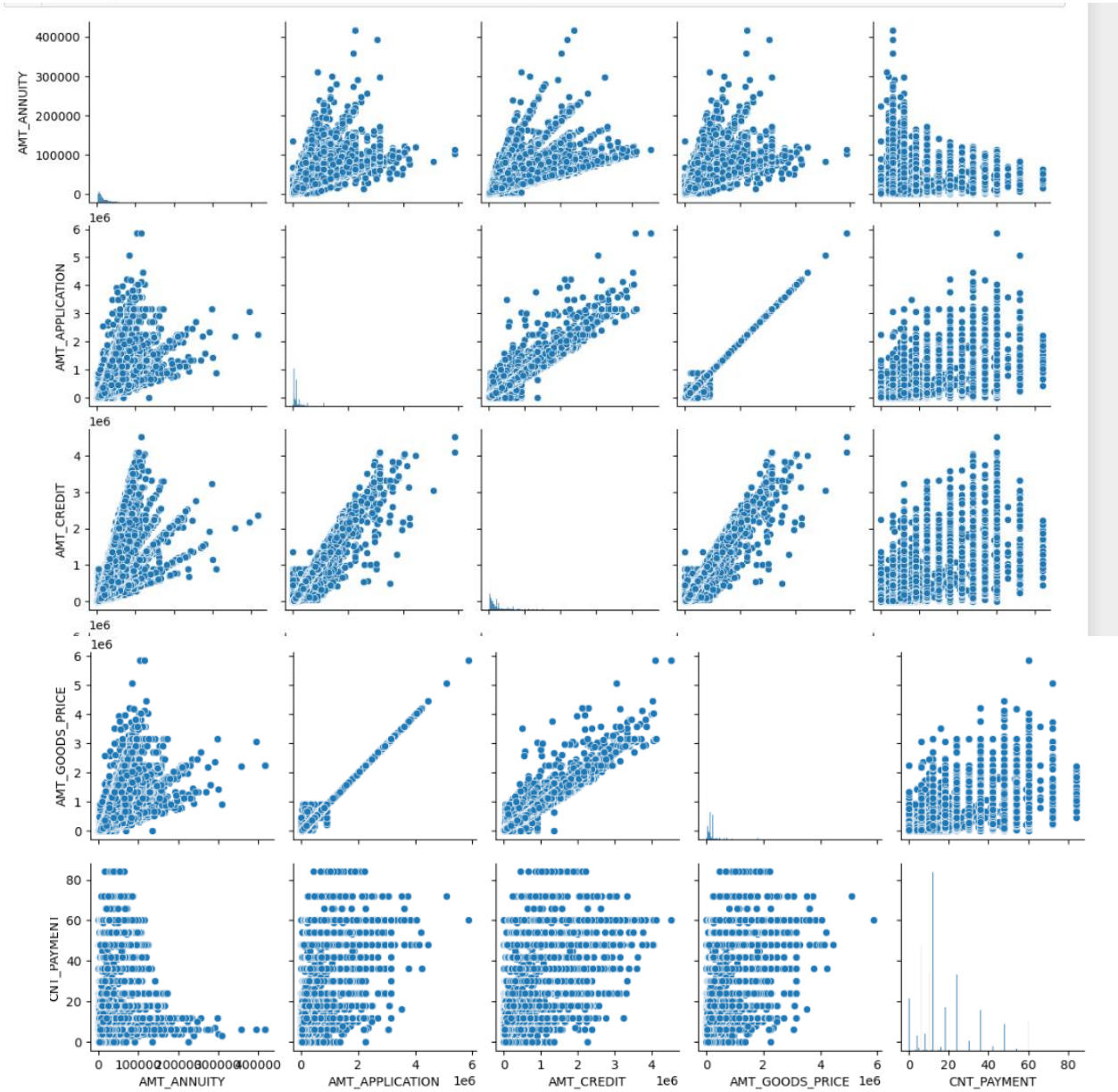
1. the higher number of approved loans is consumer loans.
2. most of repeater clients are here, therefore approved as well as refused loans for repeater clients.
3. for the new clients the loans is approved mostly as compaired to refused.
4. POS portfolio seems to be consumer loan, Also more cash portfolio are refused than POS.
5. cash through the bank payments have more transcations and approved also.

Numerical vs Numerical Bivariate Analysis :-

Analysis column name which has a numerical datatype

point to be notice :

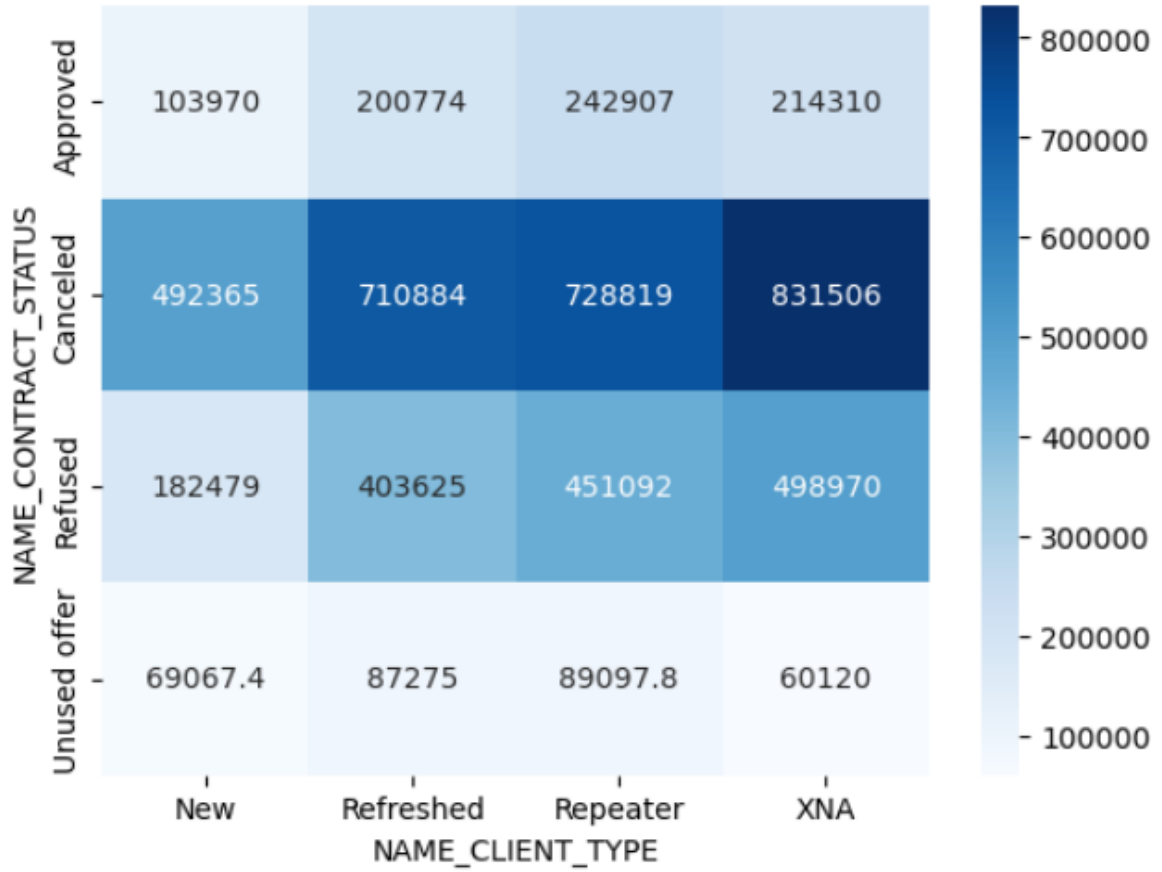
1. AMT_ANNUITY , AMT_APPLICATION, AMT_CREDIT ,AMT_GOODS_PRICE has as expected higher correlaion in between them.¶
2. Column CNT_Payment ideally should have had a high correlation with AMT_credit, i.e higher credit, more the term of loan. But no such correaltion can be seen



Multivarient Analysis :-

1. Checking contract status vs name client type aggregating over application amount

2. Checking contract status vs name client type a



Point to be notice for 1 heatmap :

1. Unused offer application numbers are low.
2. Canceled applications rate is high, in every type of client.
3. Repeater clients are more which means the service or policies proivde by company is good.

Point to be notice for heatmap2:

1. Unused offer CREDIT AMOUNT is low. This may be the reason for customer not using
2. Unable to understand why for cancelled and refused there should be any credit amount?

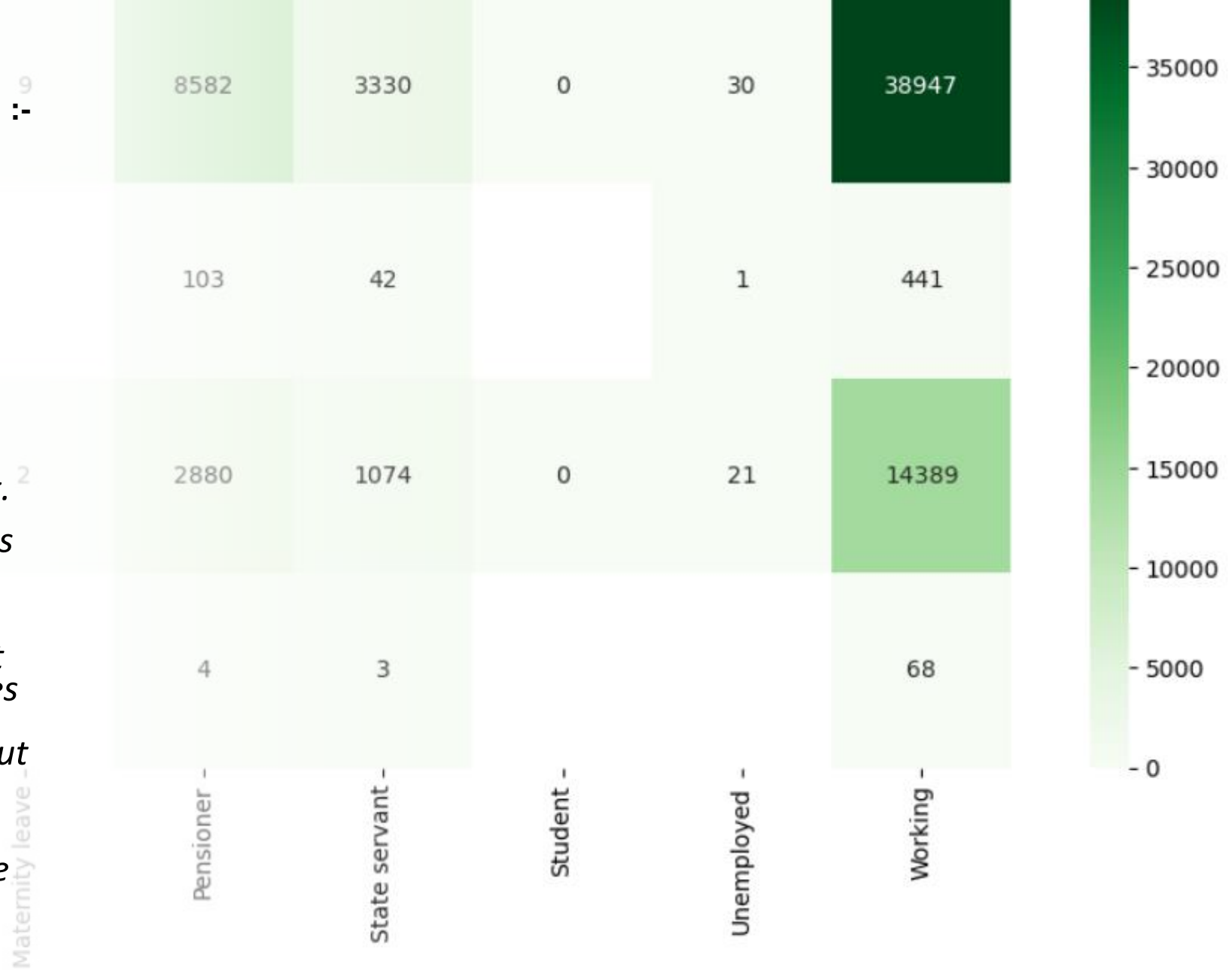
1. Checking contract type vs name client type aggregating over AMT CREDIT
2. Checking contract status vs name client type aggregating over AMOUNT GOOD PRICE

Point to be notice :
1. There is lot of Cash loans Canceled , Approved, Refused.

Point to be noticed:
1. All cancelled and Refused have higher goods than others, so the company needs to focus on it.

Merged DataFrame Analysis :-

Checking "NAME_CONTRACT_STATUS", "NAME_INCOME_TYPE",aggregating on Target



- **Point to be notice :**

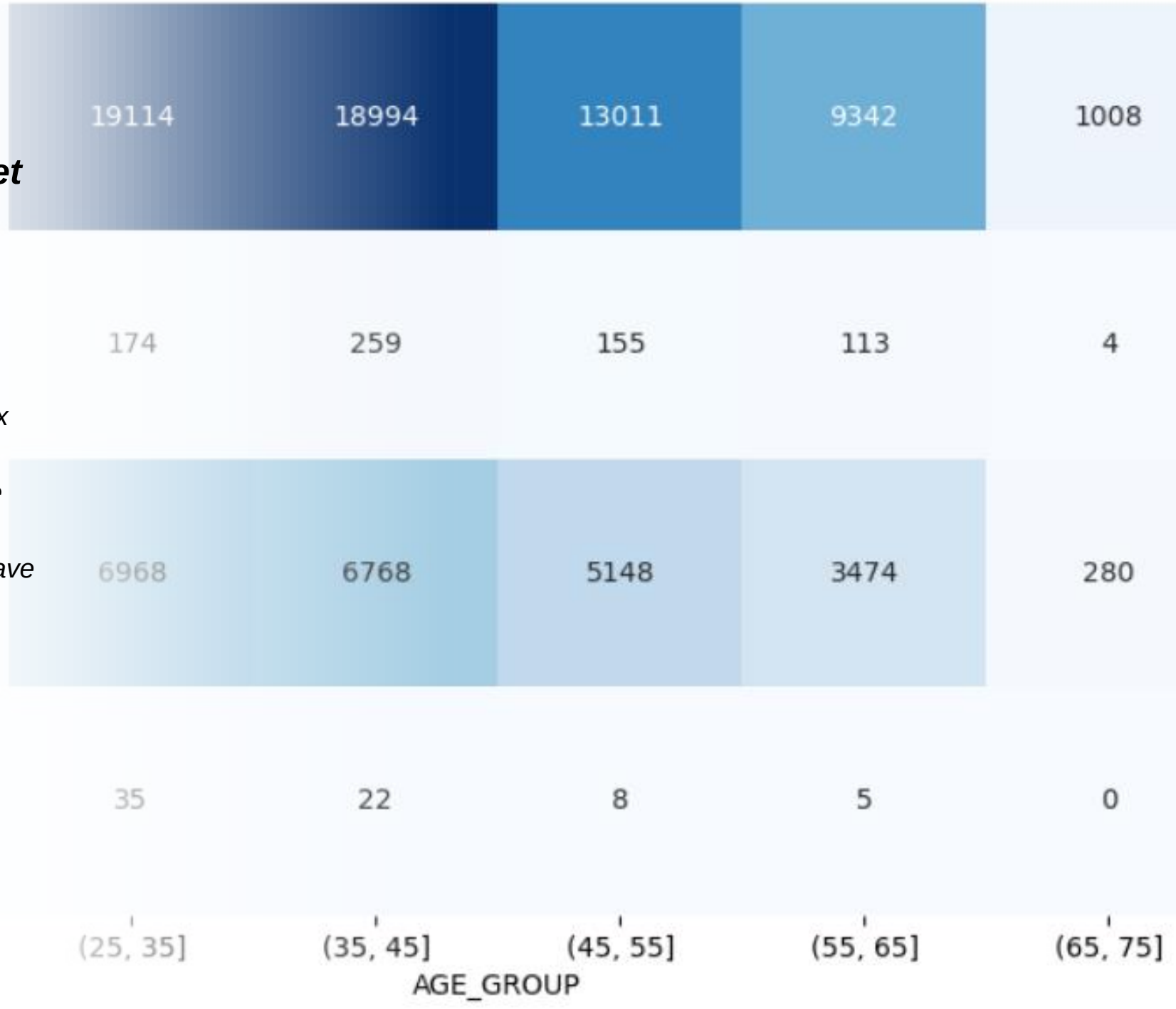
1. Since Target 1 is default, higher on the above matrix shows correlation to default.

2. Working applicant with Approved status have defaulted in highest numbers

3. Previous applications with Refused, Cancelled, Unused loans also have default which is a matter of concern. This indicates that the financial company had Refused/cancelled previous application, but has approved the current and is facing default on these loans.

4. 14,389 applicanst of working class were REFUSED earlier and now have defaulted.

Checking "NAME_CONTRACT_STATUS", "AGE_GROUP",aggregating on Target



- **Point to be notice :**

1. Since Target 1 is default, higher on the above matrix shows correlation to default.

2. Approved loans of age group 25-35 and 35-45 have higher defaults

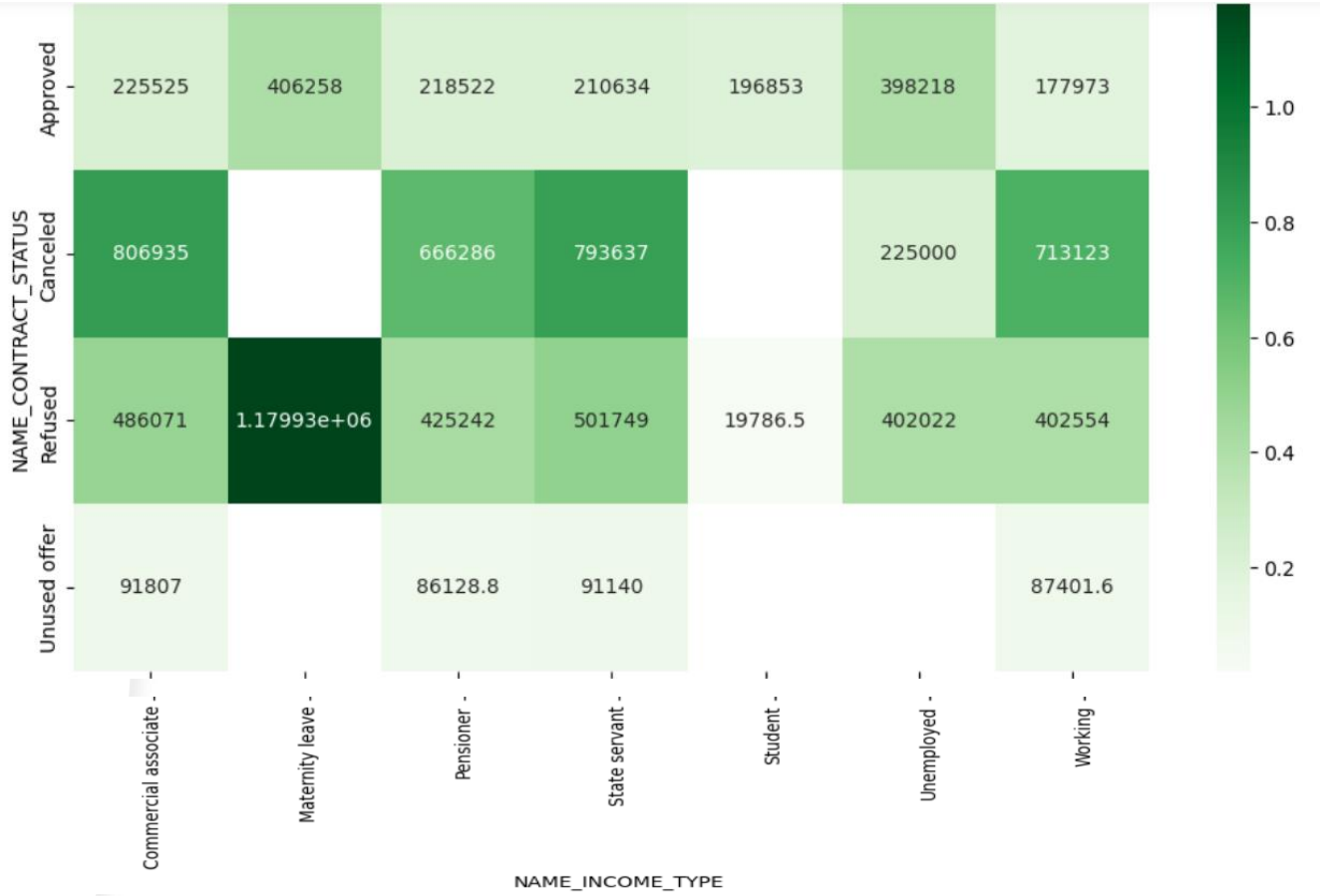
3. Refused, cancelled, loans in previous application have defaulted in current.

Checking "NAME_CONTRACT_STATUS", "AGE_GROUP",aggregating on Target

- **Point to be notice :**

1.Higher credot offered to unemployed, maternity leave is a notable factor

2. Unused offers have smaller credit values and possibly the reason why applicant is not usign them



Case Summary

- ***Defaulters' demography***
- *All the below variables were established in analysis of Application dataframe as leading to default.*
- *Checked these against the Approved loans which have defaults, and it proves to be correct*

  - Medium income

  - 25-35 years age group , followed by 35-45 years age group

  - Male

  - Unemployed

  - Labourers, Salesman, Drivers

  - Business type 3

  - Own House - No
- *Other IMPORTANT Factors to be considered*

  - Days last phone number changed - Lower figure points at concern

  - No of Bureau Hits in last week. Month etc – zero hits is good

  - Amount income not correspondingly equivalent to Good Bought – Income low and good value high is a concern

  - Previous applications with Refused, Cancelled, Unused loans also have default which is a matter of concern.

This indicates that the financial company had Refused/Cancelled previous application but has approved the current and is facing default on these.

Credible Applications refused

- Unused applications have lower loan amount. Is this the reason for no usage?
- Female applicants should be given extra weightage as defaults are lesser.
- 60% of defaulters are Working applicants. This does not mean working applicants must be refused. Proper scrutiny of other parameters needed
- Previous applications with Refused, Cancelled,Unused loans also have cases where payments are coming on time in current application. This indicates that possibly wrong decisions were done in those cases.

END....