Credit EDA Assignment by HY Tso

Table of Contents

1. Analysis Objectives
2. Task
3. Software Tool Used
4. Approach
5. Insights & Result
6. Conclusion

Analysis Objectives

- This analysis would like to understand the driving factors behind loan default. This is to help the company getting to know their client portfolio and risk assessments.

- Indicating which group among their clients repaying their loan and which group not and the variables would help us to identify such groups

Task steps throughout the analysis

1. Identifying Missing Value and Handling it
2. Identify Outliers in the dataset especially columns that's related to our objectives of analysis.
3. Analyze Data Imbalance and its ratio
4. Perform univariate, segmented univariate, bivariate analysis, finding variables vs TARGET column or between variables and using them as evident.
5. Identify Top Correlations

Software Tool Used

Jupyter Notebook

Python 3.11

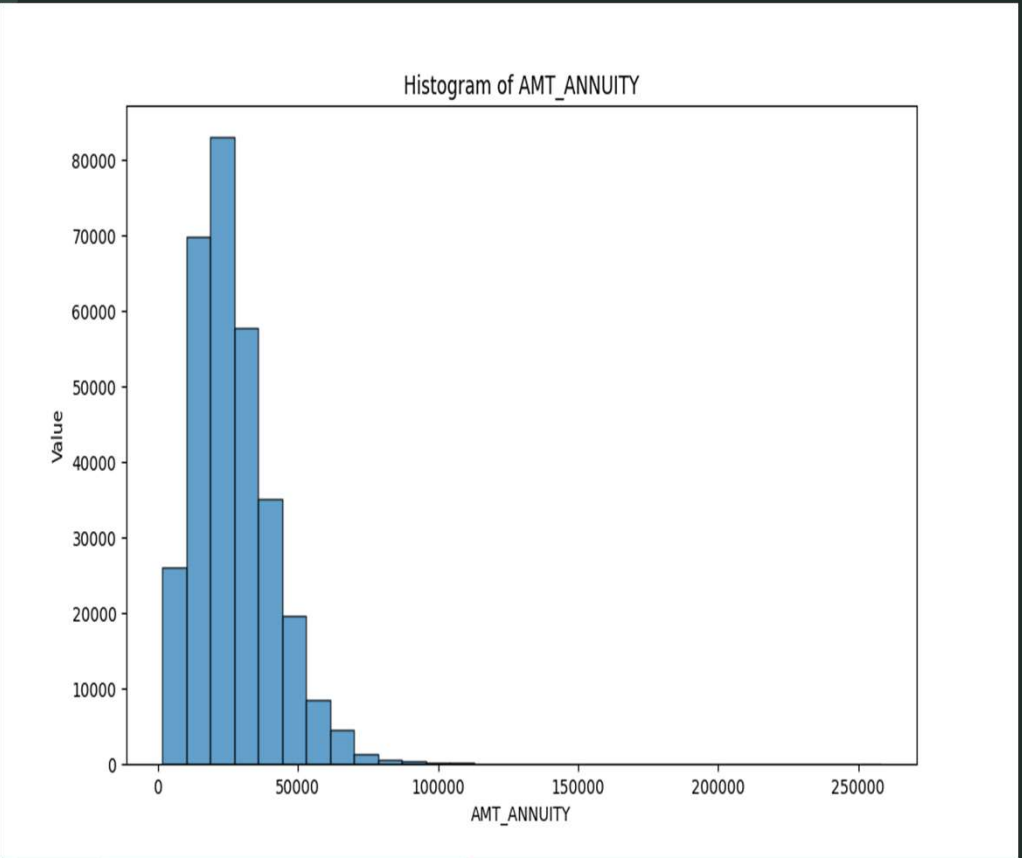
Microsoft PowerPoint

Approach used in this analysis

- Data Acquisition and Understanding – Datasets were downloading from the source, gain understating from file columns_description
- Data Cleaning and Analysis – cleaning process is conducted in Jupyter notebook under Python environment. Visualization will be run in Python and images would be saved
- Report Presenting – explain process how data had been collected and visualized. Insights from preparer would be presented here.

Insights and Result

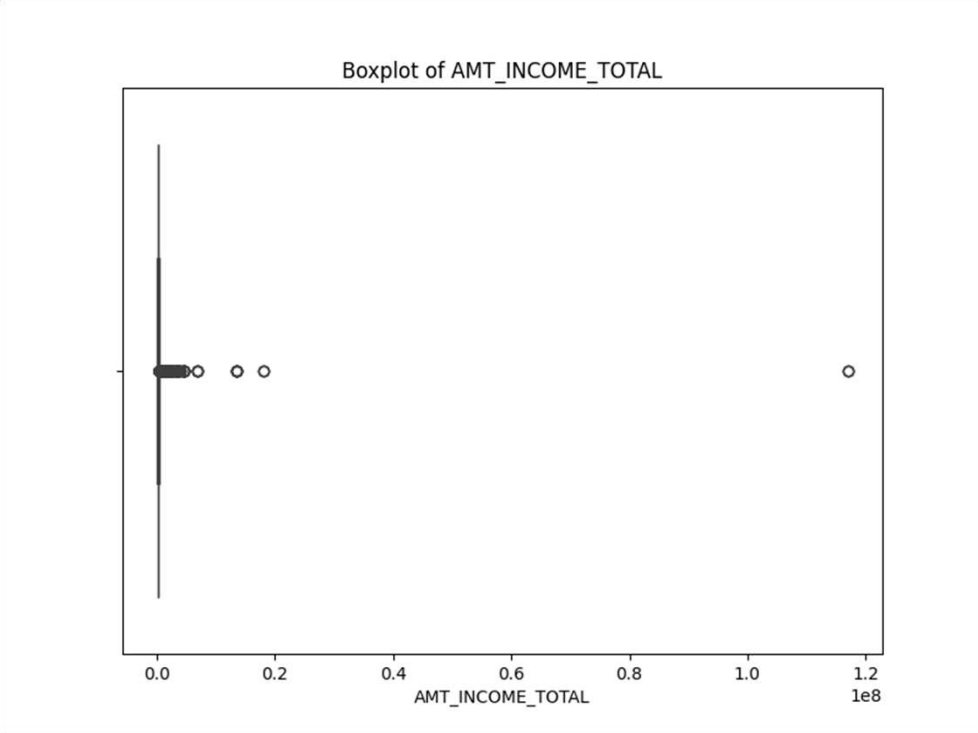
Finding Missing Values and Cleaning Data



After how many missing values in percentage were identify, I dropped those that are more than 30%. I decide replacing values with median, mode or mean either by creating histogram graph. Categorical columns replace missing values with mode. I cleaned both datasets and merged it at the end.

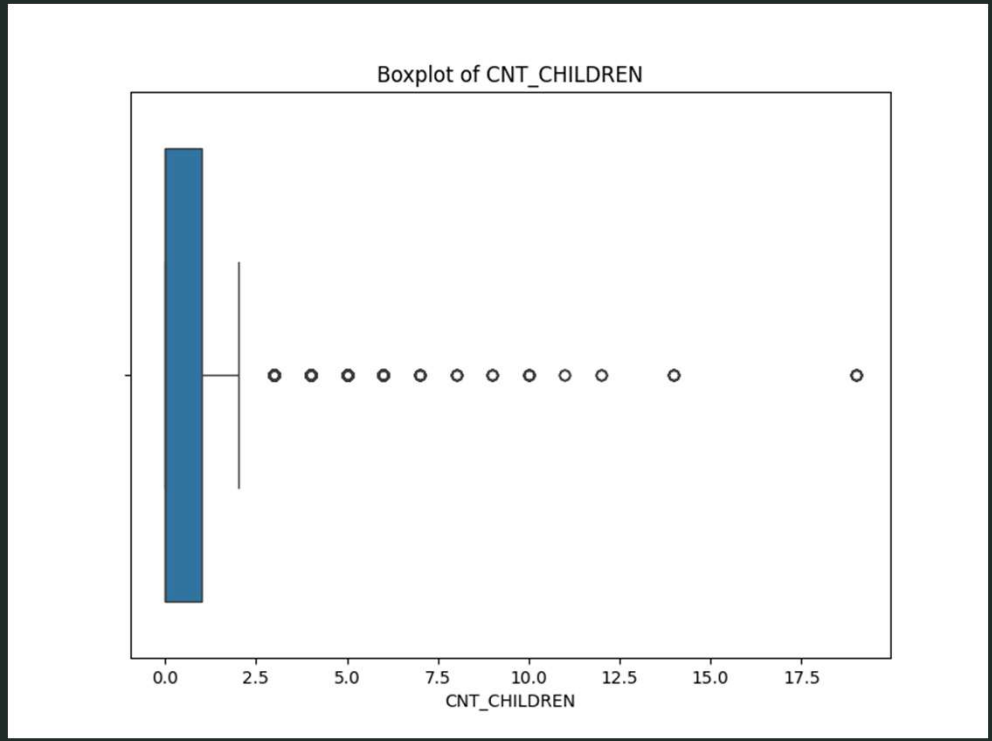
Identify Outliers

- This income Total plot indicating there is an outlier clients with higher income compared to other average clients.
- It is worth investigating that bank can come up financial products catering for clients with higher income



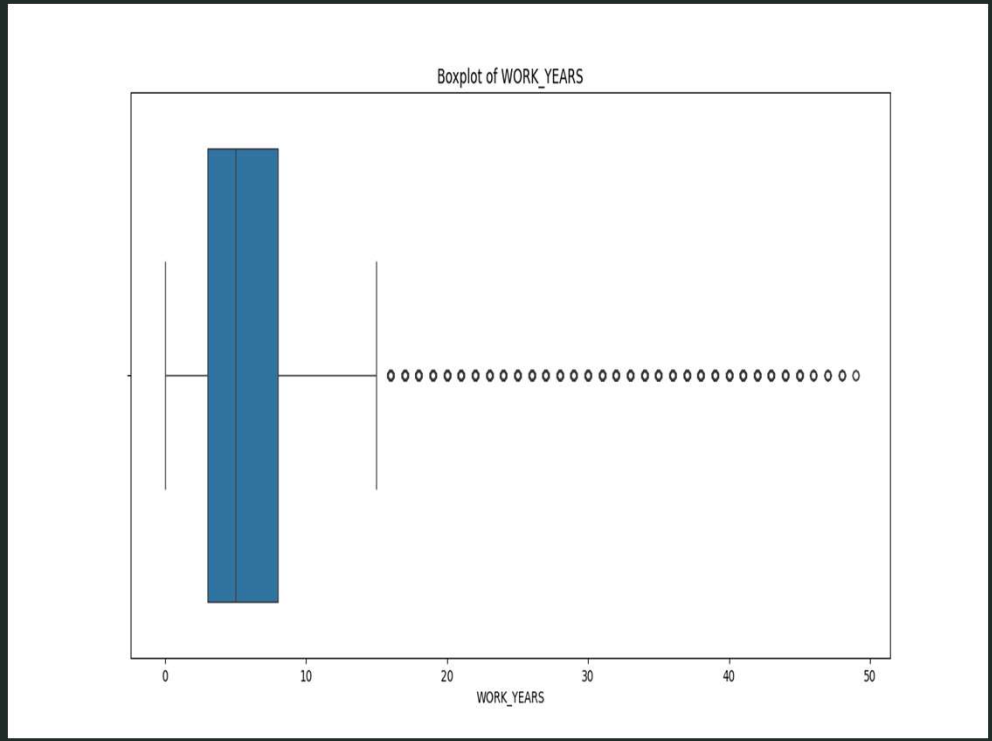
Identify Outliers

- This Children number plot indicating there is an outlier clients with big family like more than 17 children.
- It is worth investigating that bank can come up financial products catering for clients with big families.



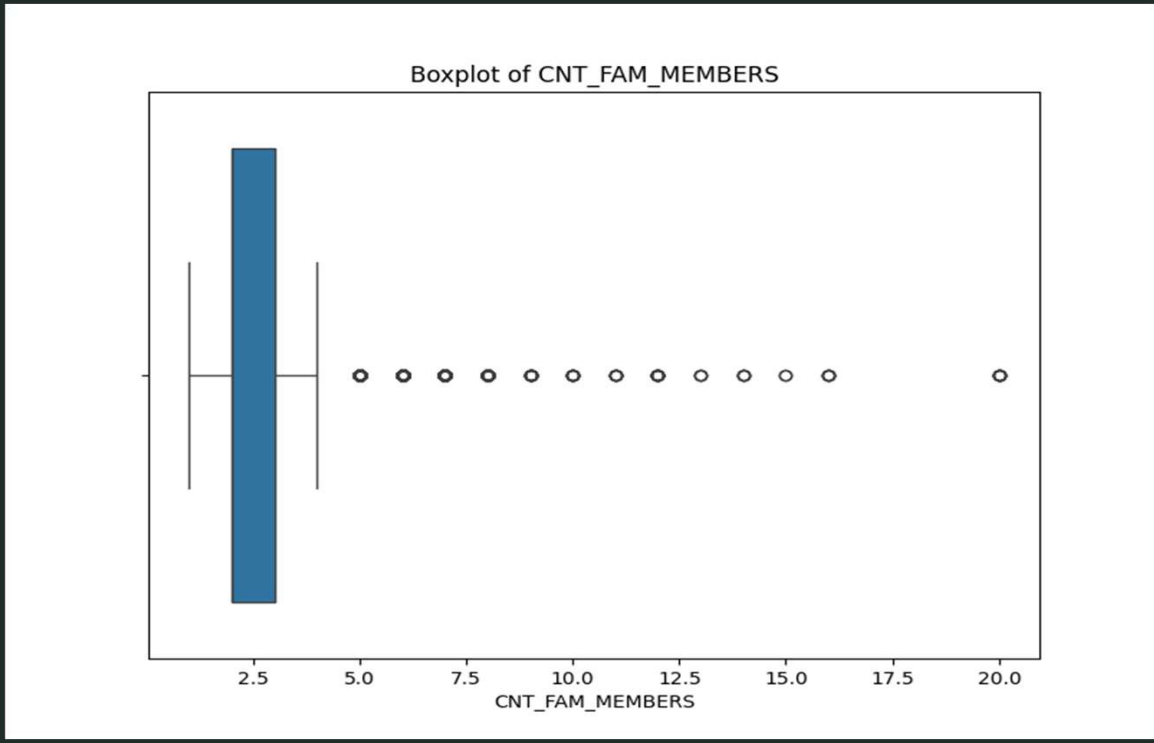
Identify Outliers

- This WORK YEARS plot indicating there are clients being employed for many years.
- It is worth investigating that bank can come up financial products catering for clients with stable career.



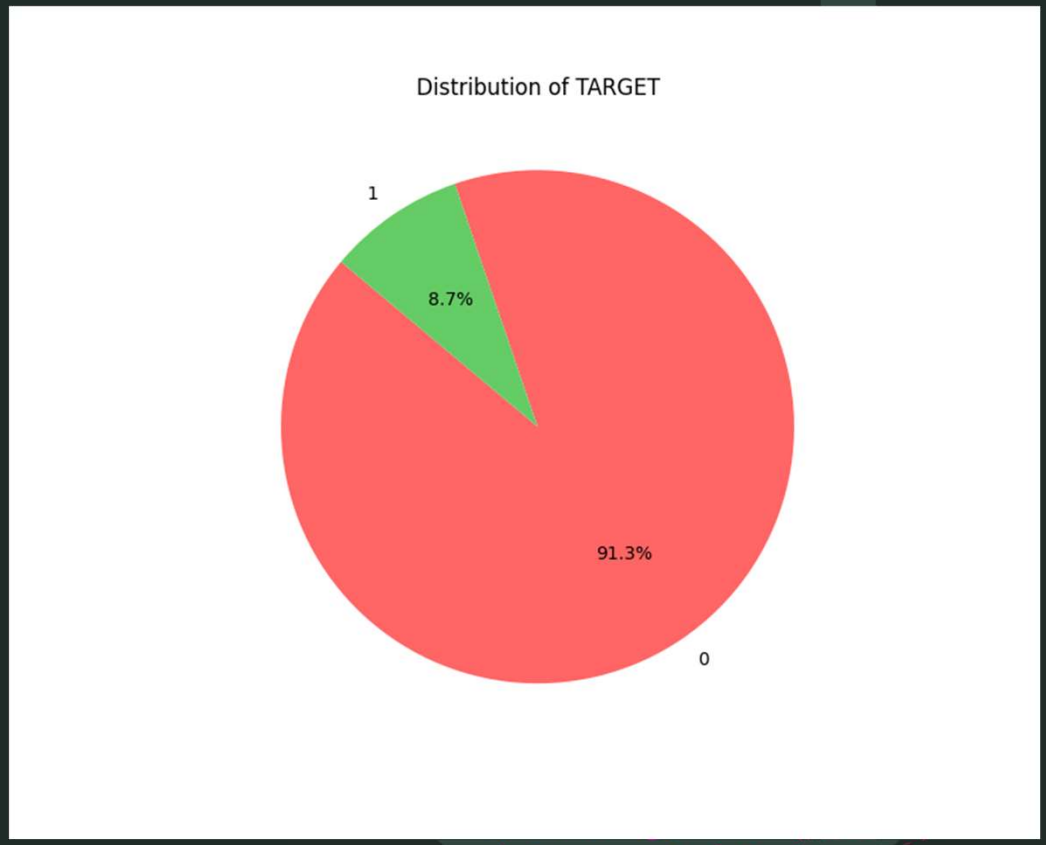
Identify Outliers

- This Family member number plot indicating there are clients having many family members.
- It is worth investigating that clients like this if they have more than one income in the household. If company can design financial product for such household, it can open up potential clientele.



Data Imbalance

- The ratio show dataset is class imbalance significantly and skewedly distributed.
- 0 indicating no payment difficulty, 1 indicating payment difficulty
- Payment difficulty: No payment difficulty ratio is 1:11.
- We can see client with payment difficulty is only 8.7% however company needs to investigate if this percentage still affecting their cashflow as a business and how to improve.

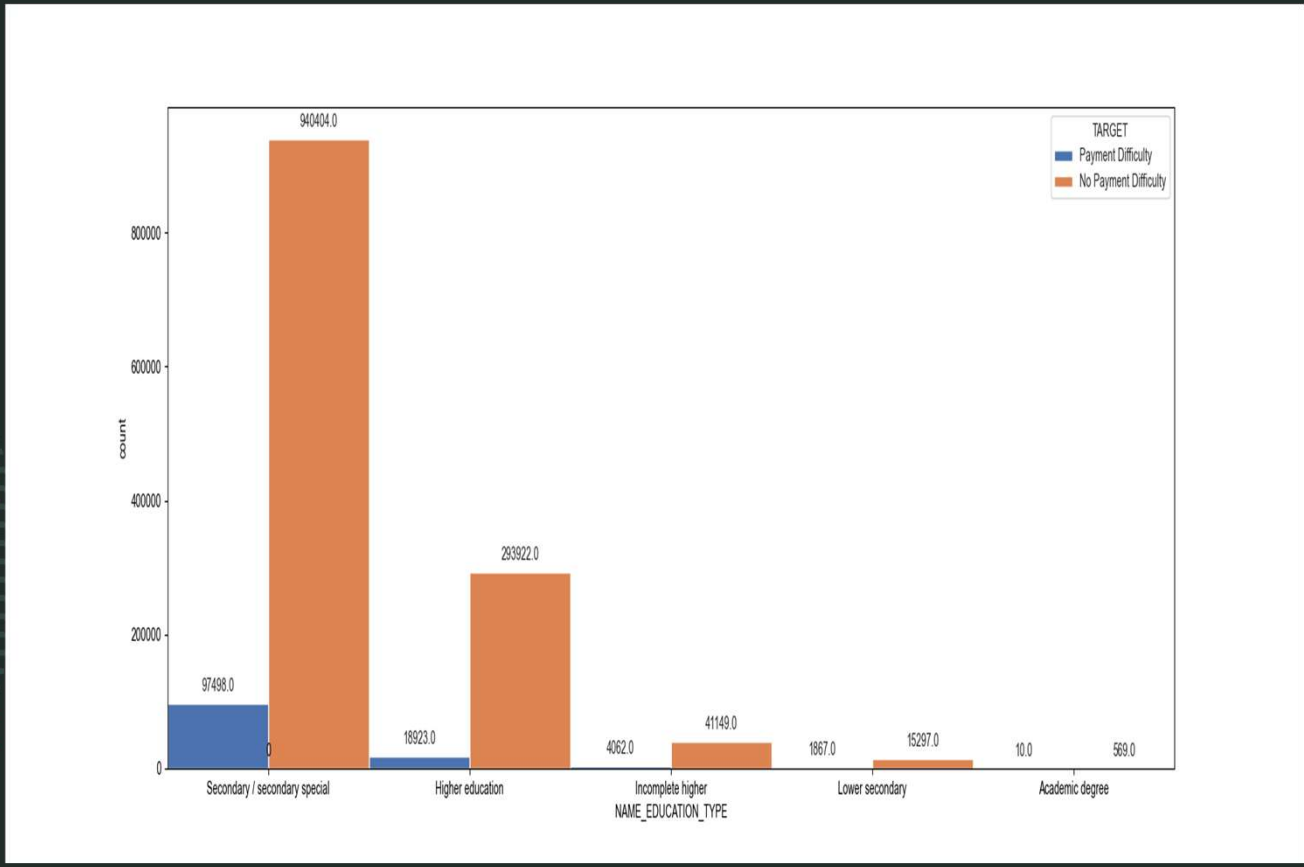


Analysis of Univariate, Segmented Univariate, Bivariate Analysis

Educational Background

- Through this histogram we can see company's clients are with secondary and higher educational background.

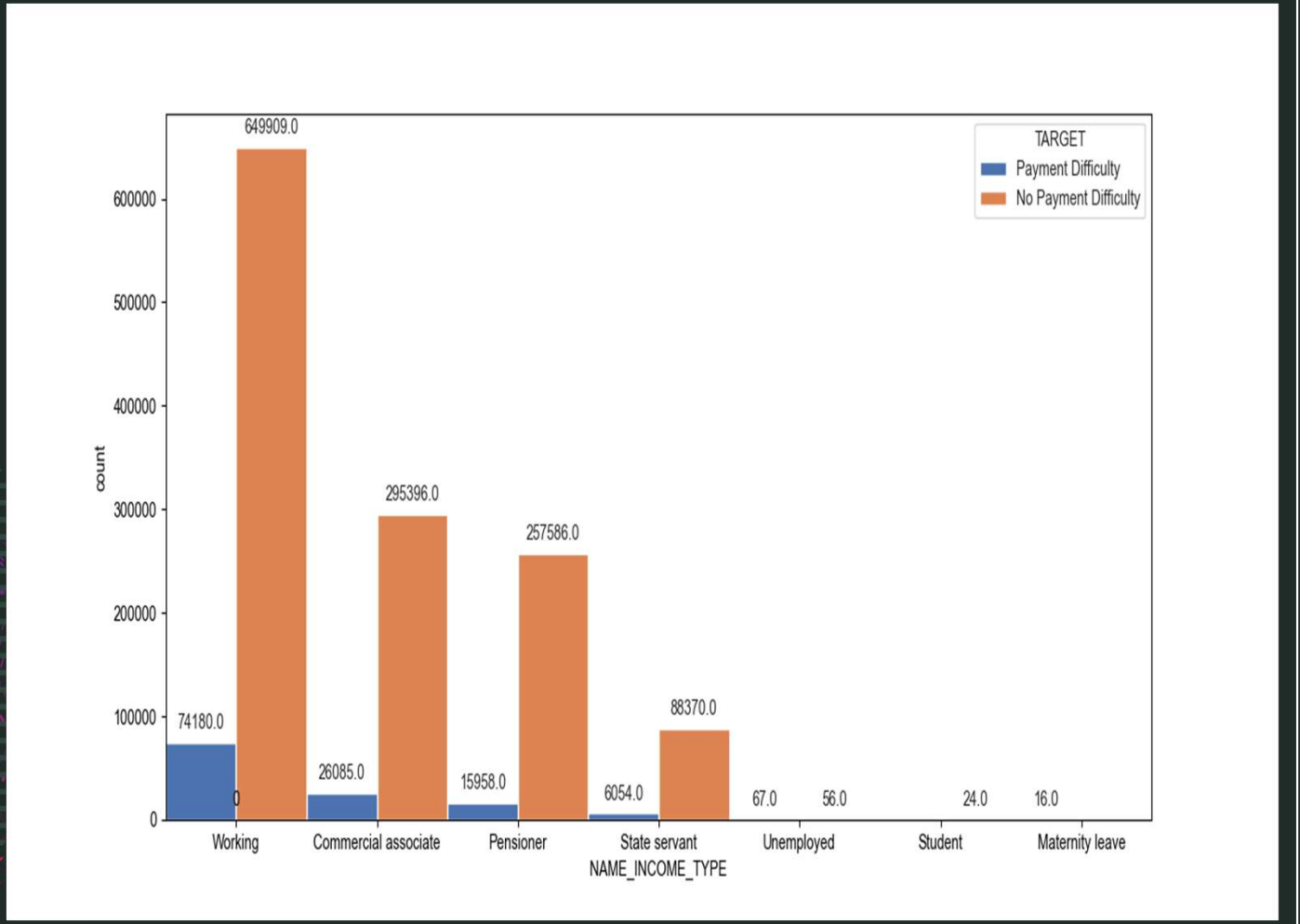
- Clients with secondary educational background has more payment difficulty. Company can investigate the reason.



Income Type

- Working clients are a big clientele for this company and come with commercial associates and pensioner.

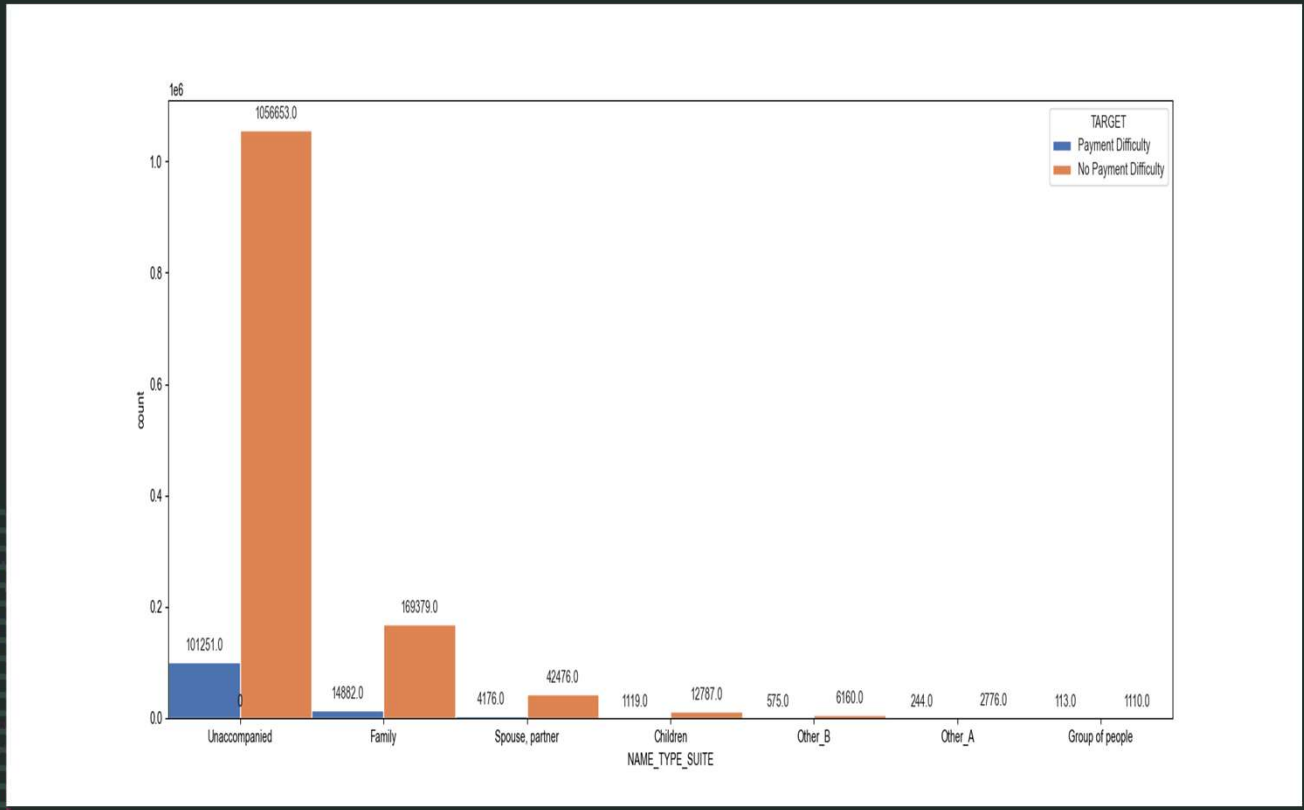
- Working clients' payment difficult ratio is very closed to 1:11. Company can investigate how to promote and attract commercial associates and pensioner and develop their clientele.



Accompanion

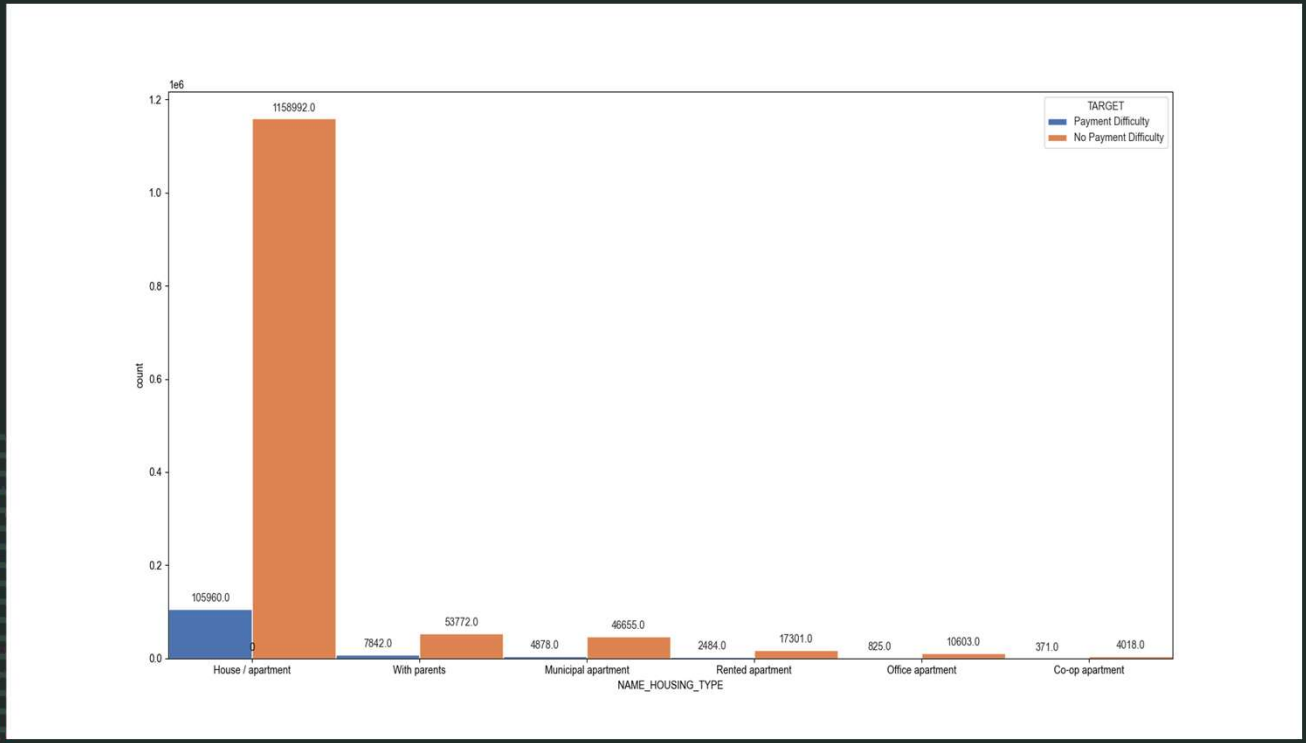
- Most clients come to apply loan unaccompanied, so they are also having a bigger payment difficulty group too.

- Unaccompanied clients has 9.5% of them having payment difficulty and family accompanied has 8.1%. Clients accompanied with family seems to be more responsible with their repayment.



Housing Type

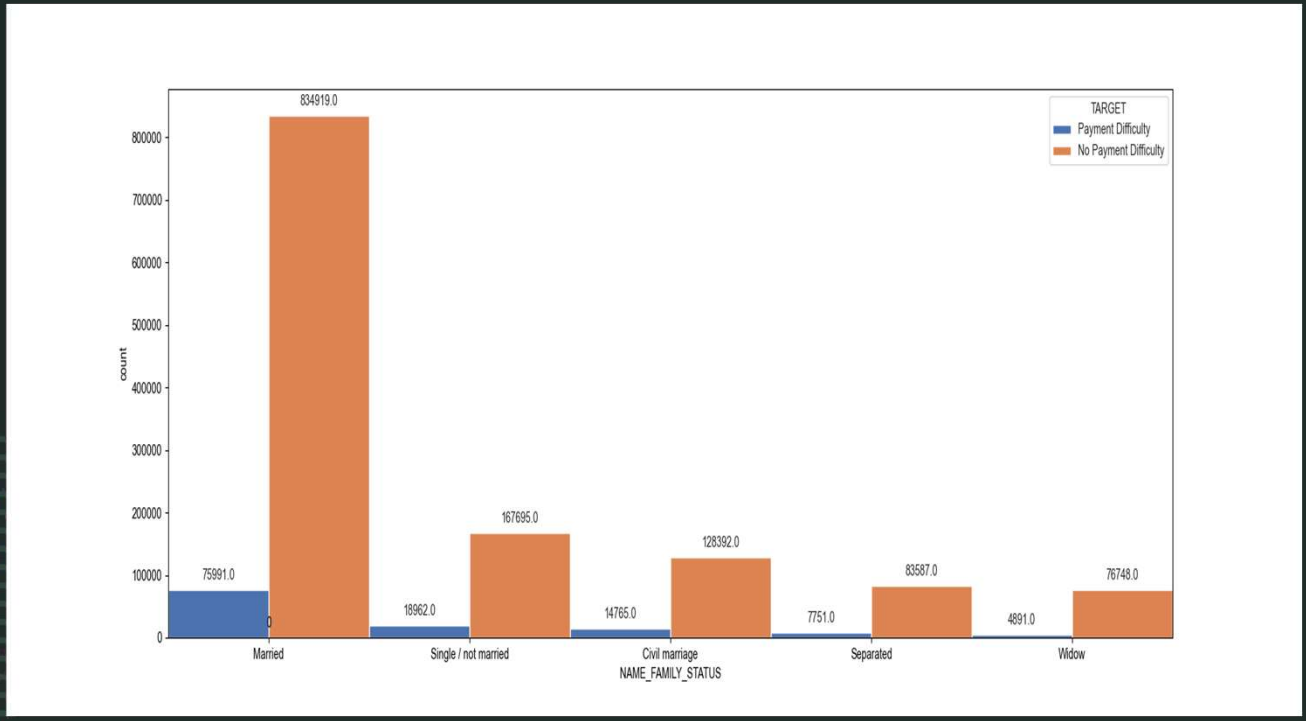
- Clients living in house/apartment tends to be more responsible with their repayments compared to clients living with their parents.
- This can be an indicator company can investigate what kind of group of their clients living with their parents.



Family Status

- Married clients tends to take out more loan which indicating having a family does have more expenses.

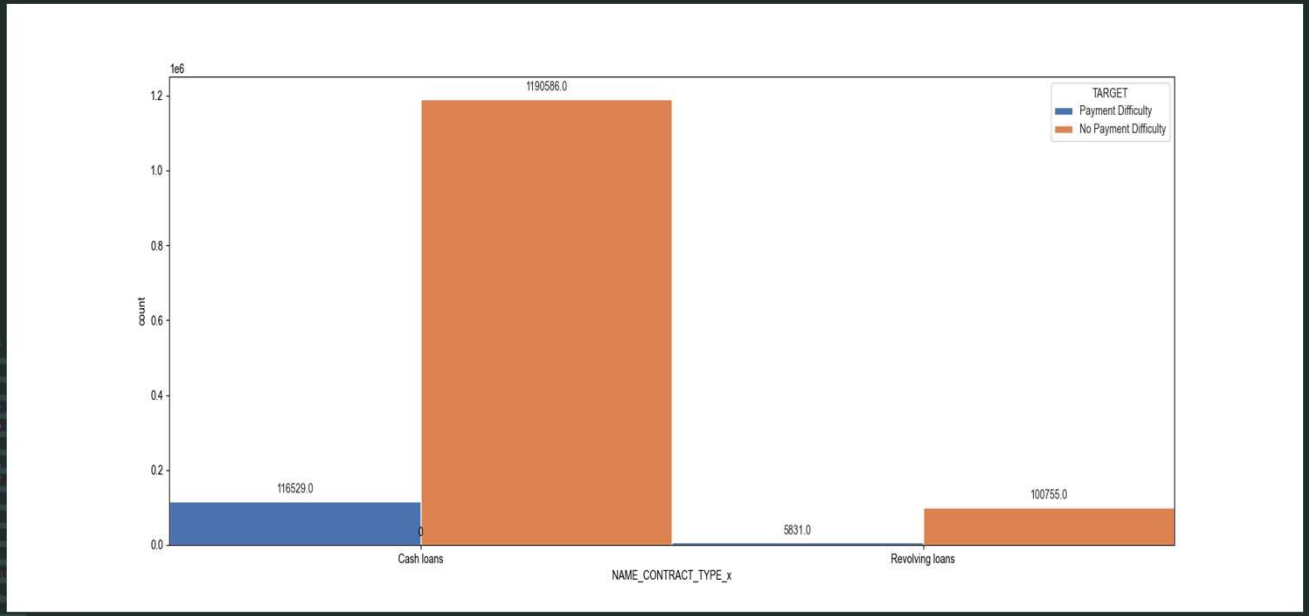
- Percentage wise single clients have more difficulty with their repayment. This can help company to investigate reasons why single clients have difficulty with repayment.



Contract Type

- Clients tend to have cash loan but they tend to have payment difficulty this indicating clients are having liquidation issues.

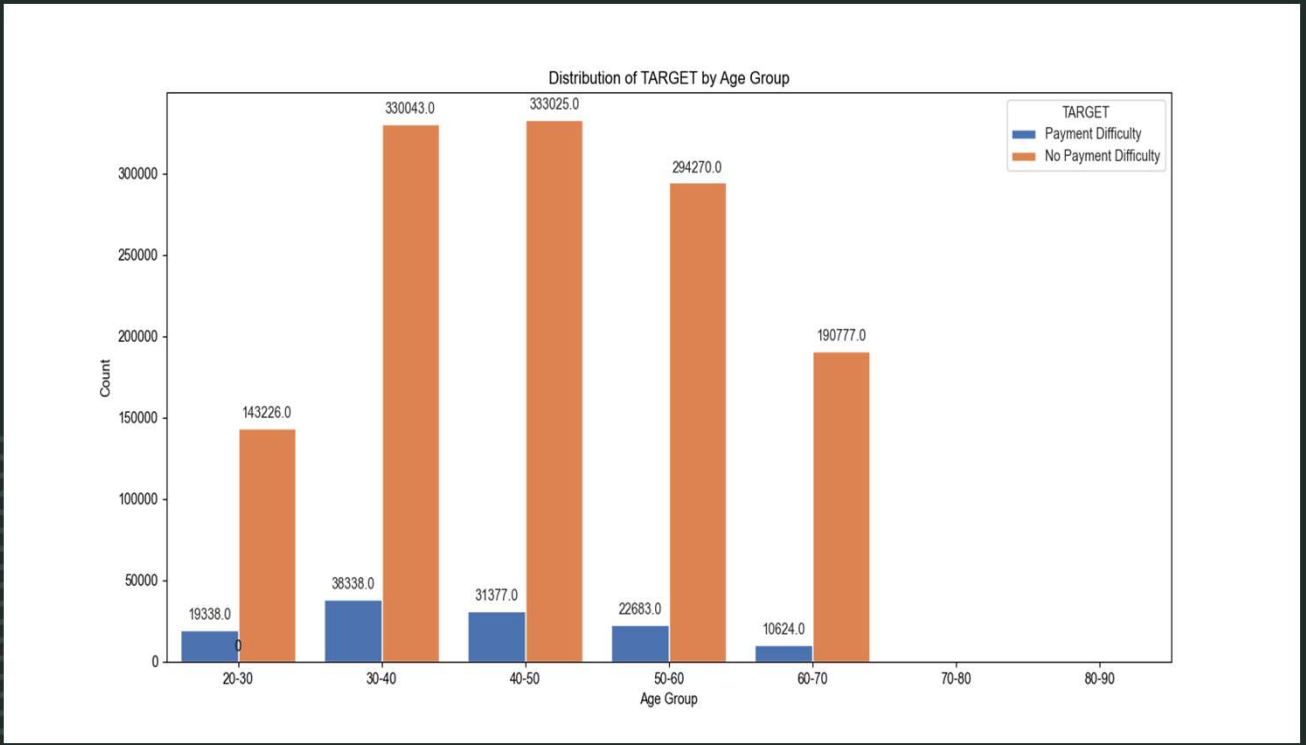
- Clients with revolving loan tend to not having payment difficulty. Company can investigate if they perform the same screening process between these two contract types.



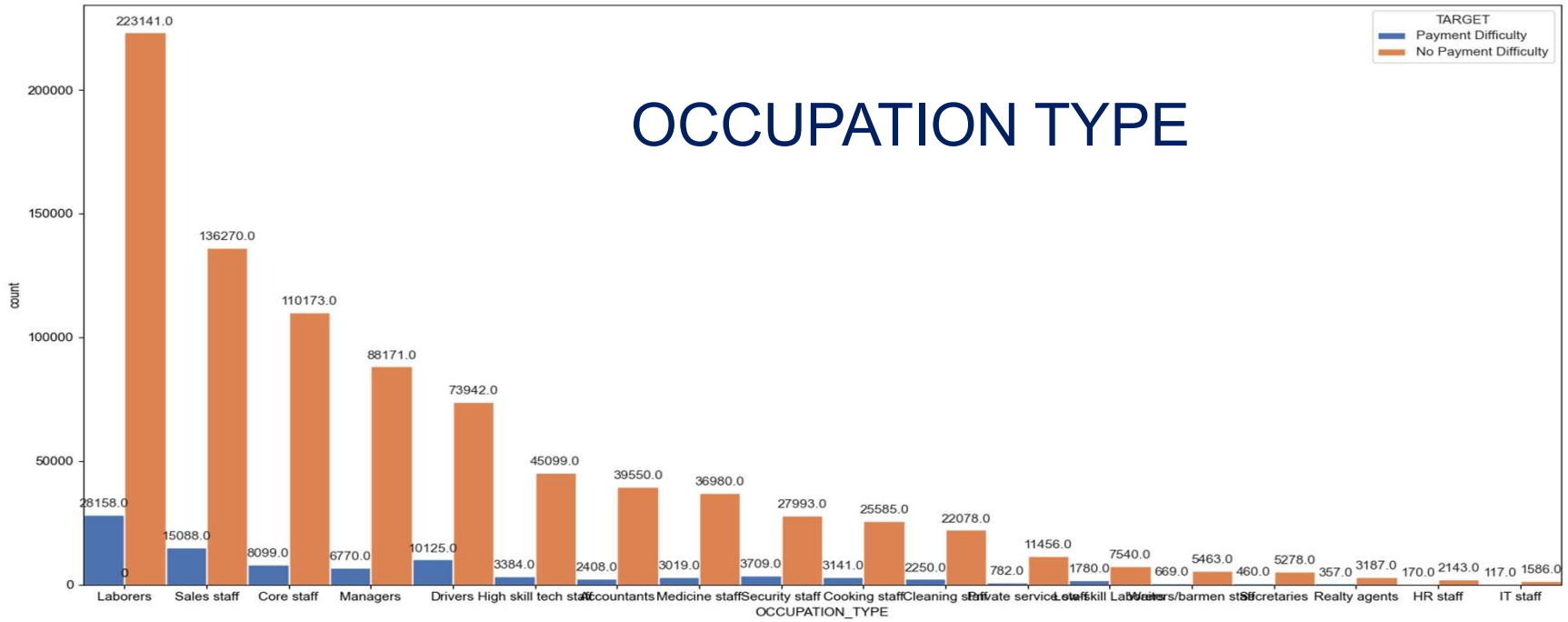
Age Group

- Clientele fall under age 30 – 50 is a big group but clientele fall under age 20 – 40 tend to have payment difficulty.

- This can indicate young clients are still building their career and might not have high income compared to middle-aged client.

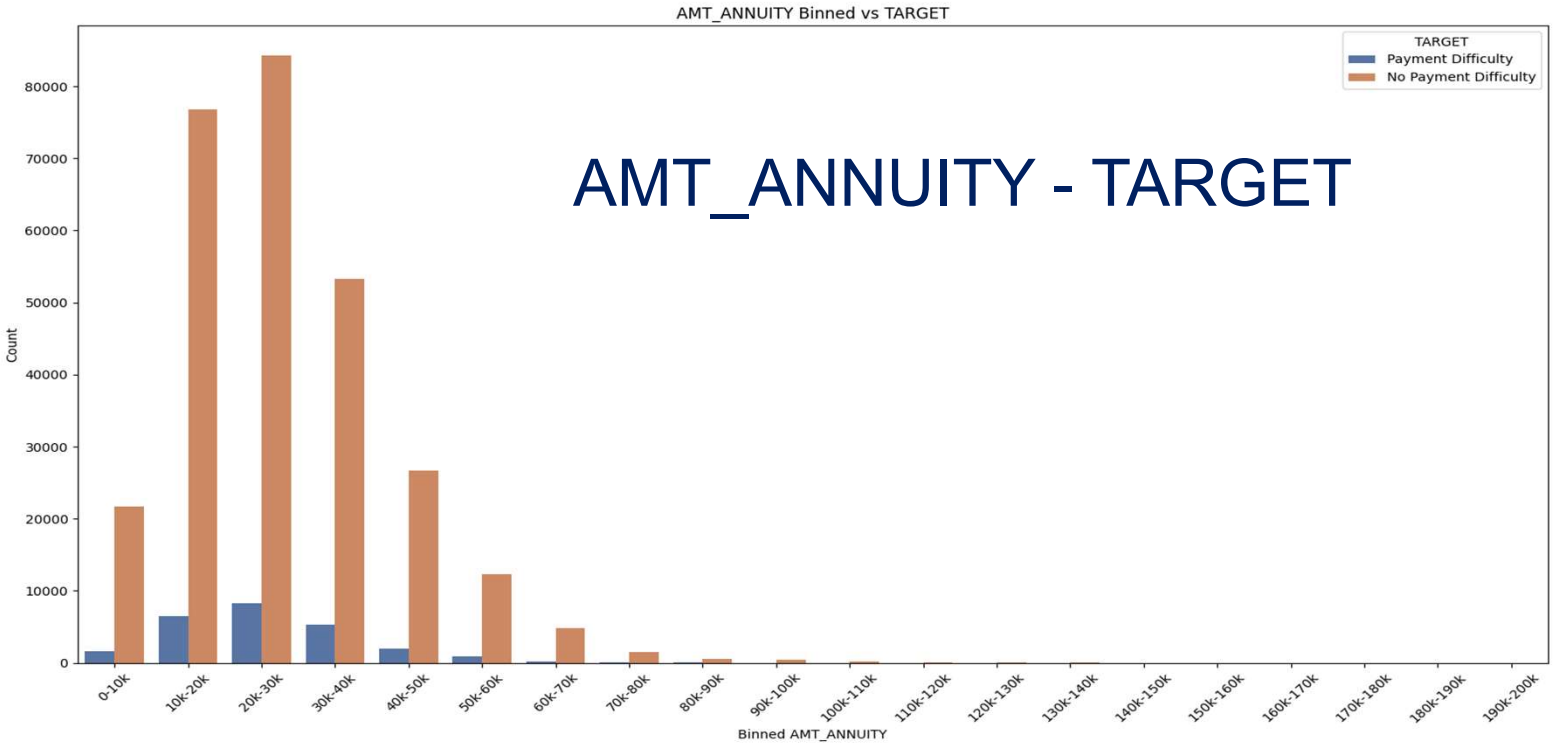


OCCUPATION TYPE

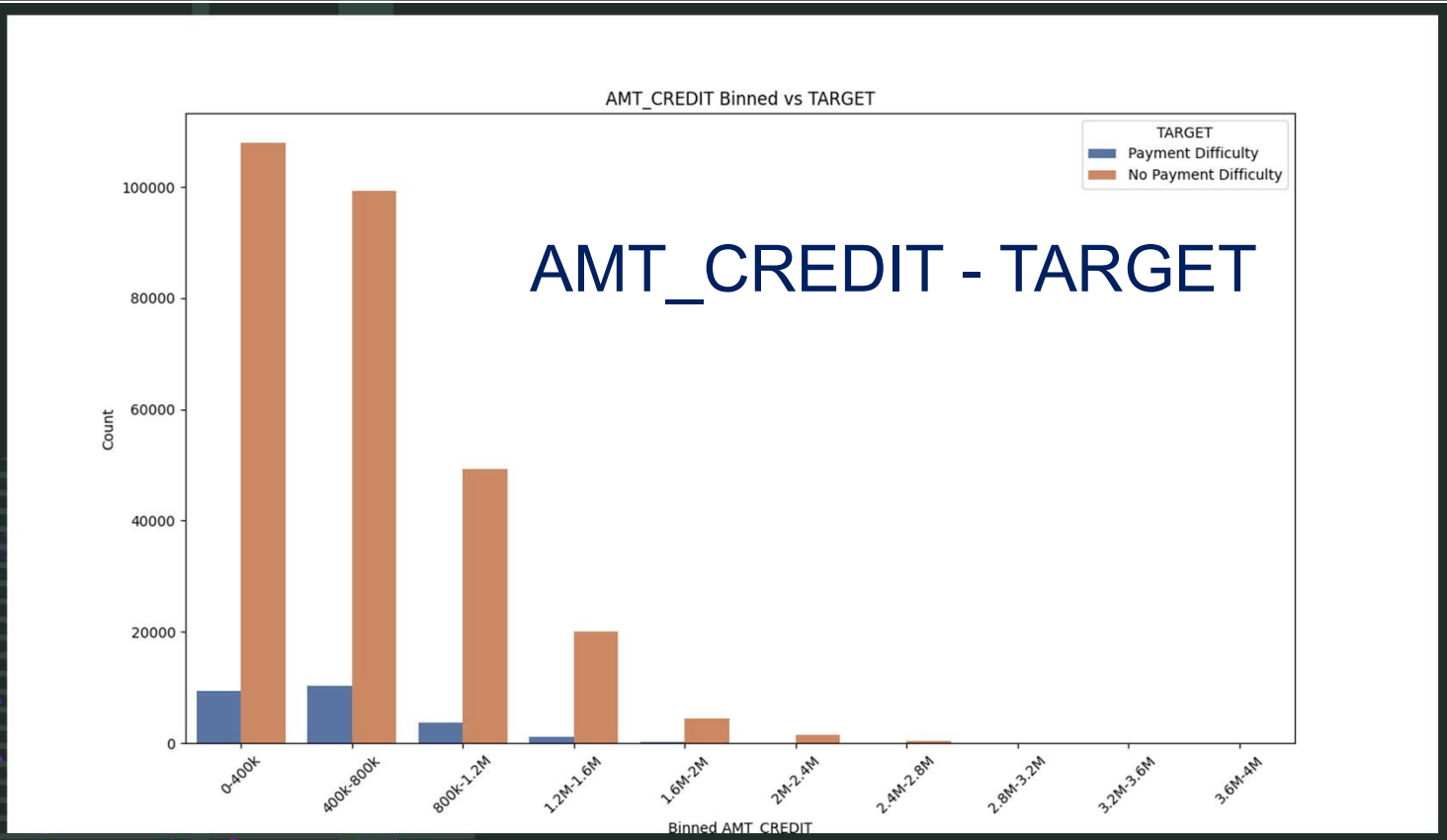


Company has a big clientele who are labourers but they also tend to have payment difficulty and sales staff come the second place. Company should examine their screening process if they did take occupation types into consideration.

AMT_ANNUITY - TARGET



Clients with 10k – 30k annuity tends to have payment difficulty and they are a big clientele to the company. This is indicating company need to screen client's spending through screening process.

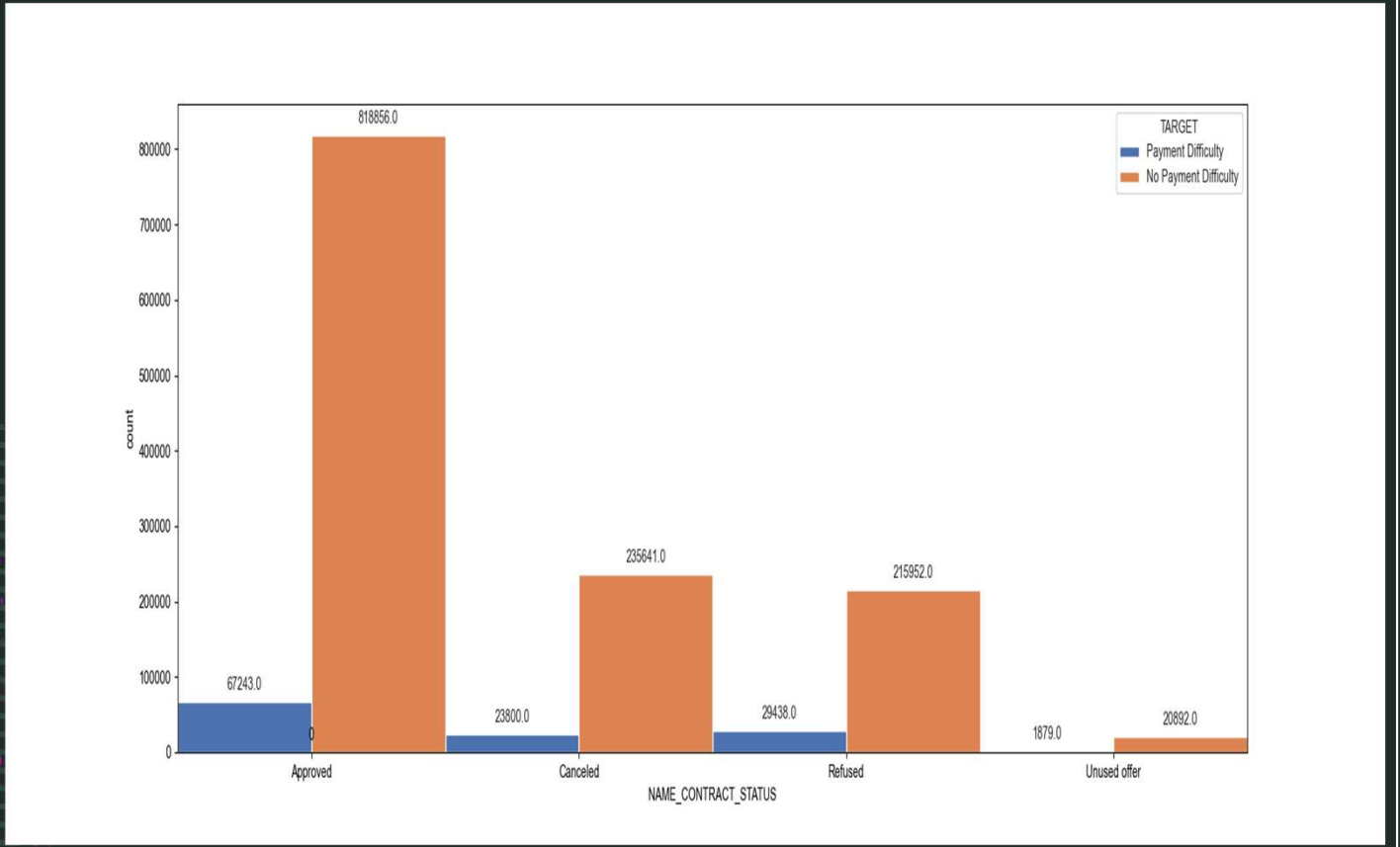


Clients with payment difficulty their credit amount fall between 400k – 800k.

Application Status

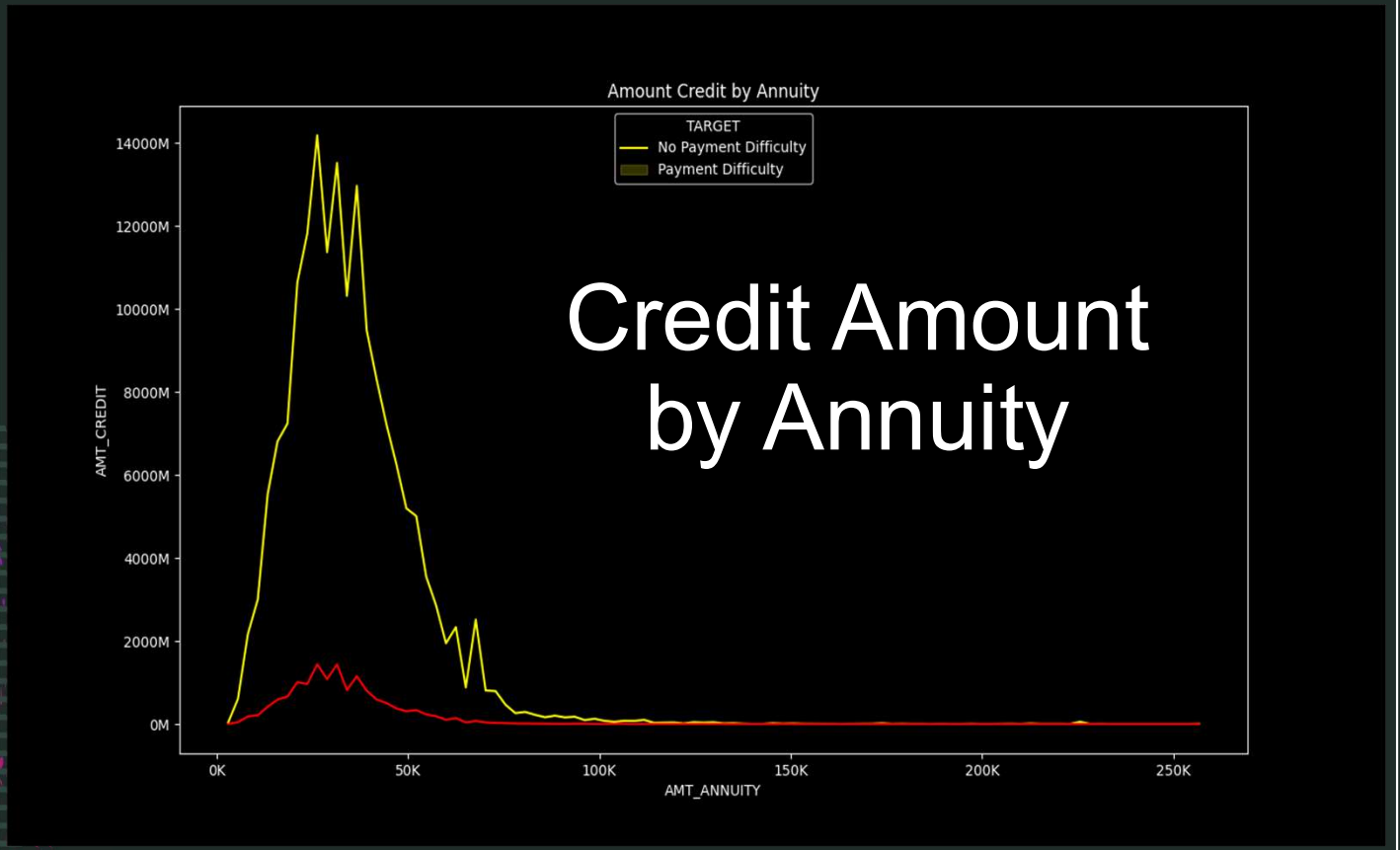
Most loans get approved, but some clients with payment difficulties are first refused for specific reasons. This is worth to be investigated.

After fixing these issues, their loans are reviewed again and eventually approved.



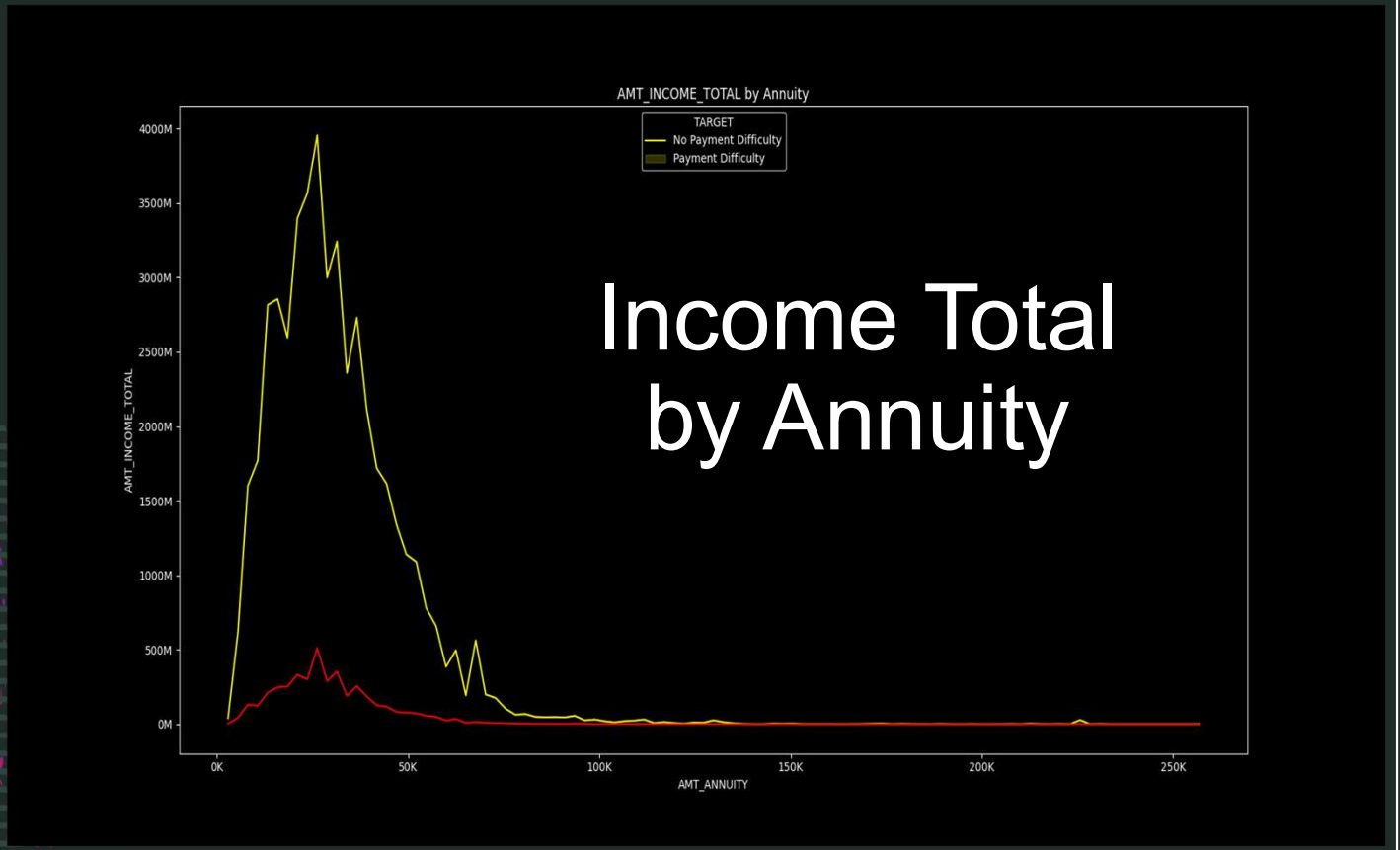
High credit amount client (14189M) their annuity is 25k.

Company can investigate client whose annuity between 20k - 36k, they are the group having payment difficulty.

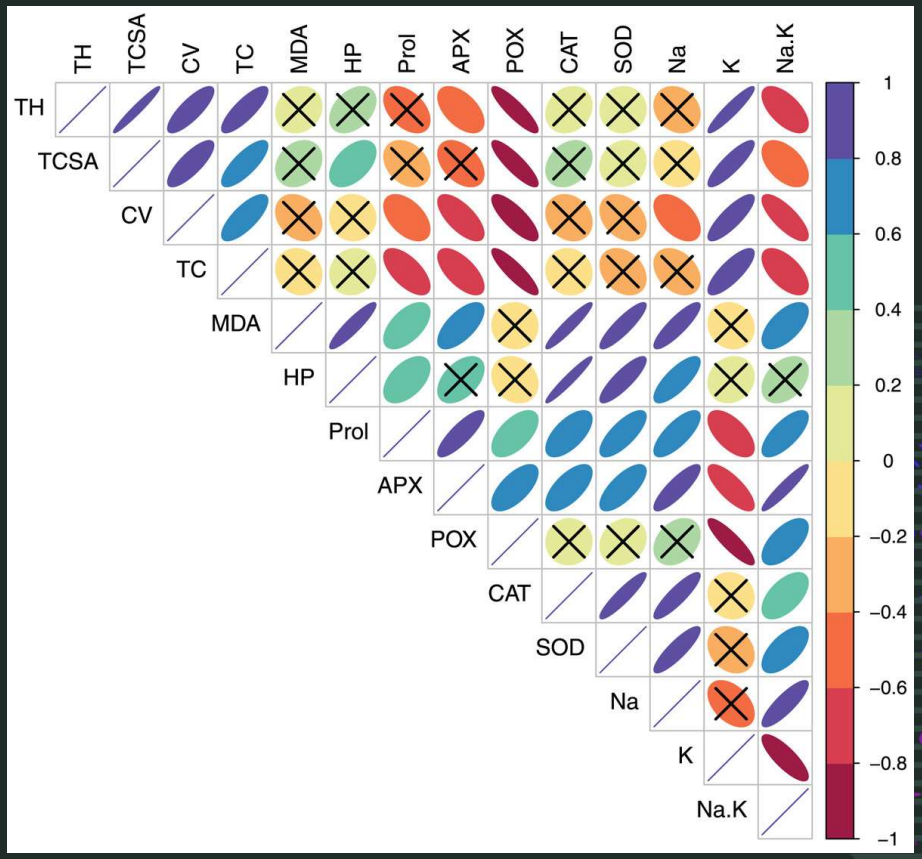


Clients having payment difficulty their income total falls under 500m, this is worth investigating.

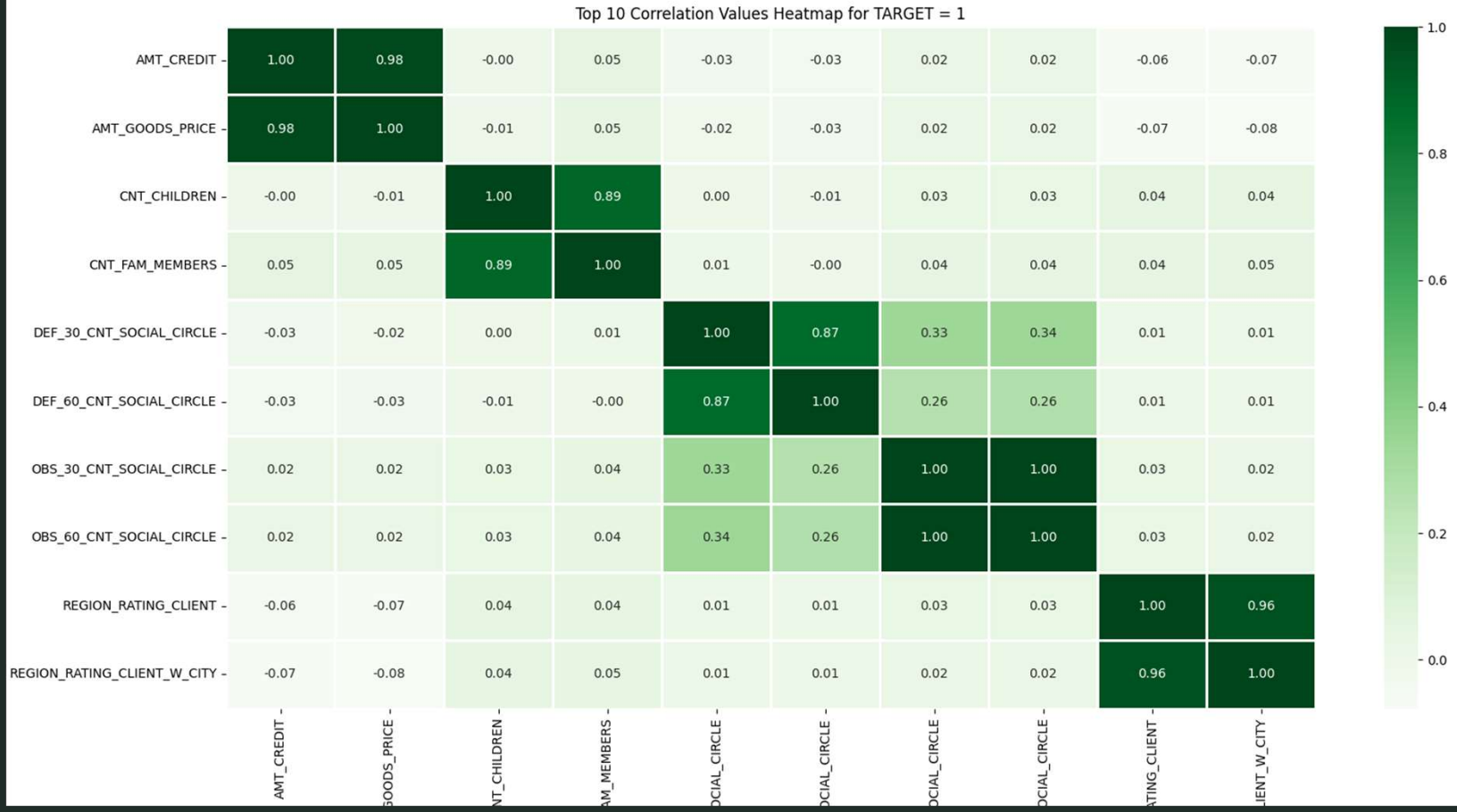
Clients with higher incomes tend to have higher annuities.



Top Correlations for Different Scenarios



TARGET = 1


Top 10 Correlation Values Heatmap for TARGET = 1
AMT_CREDIT
AMT_GOODS_PRICE
CNT_CHILDREN
CNT_FAM_MEMBERS
DEF_30_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
OBS_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
1.00 0.98 -0.00 0.05 -0.03 -0.03 0.02 0.02 -0.06 -0.07
0.98 1.00 -0.01 0.05 -0.02 -0.03 0.02 0.02 -0.07 -0.08
-0.00 -0.01 1.00 0.89 0.00 -0.01 0.03 0.03 0.04 0.04
0.05 0.05 0.89 1.00 0.01 -0.00 0.04 0.04 0.04 0.05
-0.03 -0.02 0.00 0.01 1.00 0.87 0.33 0.34 0.01 0.01
-0.03 -0.03 -0.01 -0.00 0.87 1.00 0.26 0.26 0.01 0.01
0.02 0.02 0.03 0.04 0.33 0.26 1.00 1.00 0.03 0.02
0.02 0.02 0.03 0.04 0.34 0.26 1.00 1.00 0.03 0.02
-0.06 -0.07 0.04 0.04 0.01 0.01 0.03 0.03 1.00 0.96
-0.07 -0.08 0.04 0.05 0.01 0.01 0.02 0.02 0.96 1.00
1.0
0.8
0.6
0.4
0.2
0.0

Top 10 Correlation Columns – TARGET = 1

The top 10 correlations with TARGET = 1 show important connections in the data. The numbers that track how often a client's friends or family had trouble paying bills over 30 days and 60 days are almost exactly the same (0.998). This means if someone's friends had trouble paying bills in the last 30 days, they likely had the same trouble in the last 60 days too.

Similarly, there's a strong link (0.983) between how much money a client borrows (AMT_CREDIT) and the price of the things they buy (AMT_GOODS_PRICE). This shows that when clients borrow more money, they usually buy more expensive items.

The ratings for where a client lives also show a strong connection (0.957). If a client lives in a highly-rated city, their region is probably highly rated too.

Top 10 Correlation Columns – TARGET = 1

The size of a client's family is closely related to how many children they have (0.885). More children usually mean a bigger family.

Lastly, when looking at clients with defaults, the troubles they face over 30 days are closely connected to those over 60 days (0.869). This means that recent payment issues are often linked to problems that happened a bit earlier.

These connections help us understand how different things are related when clients have trouble paying their bills, which can be useful for further analysis and understanding of their behaviour.

TARGET = 0


Top 10 Correlation Values Heatmap for TARGET = 0
AMT_CREDIT
AMT_GOODS_PRICE
CNT_CHILDREN
CNT_FAM_MEMBERS
VE_REGION_NOT_WORK_REGION
OBS_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
EG_REGION_NOT_WORK_REGION
1.00 0.99 0.00 0.06 0.05 -0.00 -0.00 -0.10 -0.11 0.05
0.99 1.00 -0.00 0.06 0.05 -0.00 -0.00 -0.10 -0.11 0.05
0.00 -0.00 1.00 0.88 0.02 0.01 0.01 0.02 0.02 0.01
0.06 0.06 0.88 1.00 0.01 0.02 0.02 0.03 0.03 0.01
0.05 0.05 0.02 0.01 1.00 -0.02 -0.02 -0.15 -0.14 0.86
-0.00 -0.00 0.01 0.02 -0.02 1.00 1.00 0.03 0.03 -0.03
-0.00 -0.00 0.01 0.02 -0.02 1.00 1.00 0.03 0.03 -0.03
-0.10 -0.10 0.02 0.03 -0.15 0.03 0.03 1.00 0.95 -0.14
-0.11 -0.11 0.02 0.03 -0.14 0.03 0.03 0.95 1.00 -0.14
0.05 0.05 0.01 0.01 0.86 -0.03 -0.03 -0.14 -0.14 1.00
AMT_CREDIT
GOODS_PRICE
NT_CHILDREN
AM_MEMBERS
ORK_REGION
OCIAL_CIRCLE
OCIAL_CIRCLE
ATING_CLIENT
LIENT_W_CITY
ORK_REGION
1.0
0.8
0.6
0.4
0.2
0.0

Top 10 Correlation Columns – TARGT = 0

The top 10 correlations with TARGET = 0 (clients who don’t have payment difficulties) show important connections in the data. The numbers that count how often a client’s friends or family had trouble paying bills over 30 days and 60 days are almost exactly the same (0.999). This means that if someone’s friends had trouble paying bills in the last 30 days, they probably had the same trouble in the last 60 days too.

There’s also a strong link (0.987) between how much money a client borrows (AMT_CREDIT) and the price of the things they buy (AMT_GOODS_PRICE). This shows that when clients borrow more money, they usually buy more expensive items.

The ratings for where a client lives are closely connected (0.950). If a client lives in a highly-rated city, their region is probably highly rated too.

Top 10 Correlation Columns – TARGT = 0

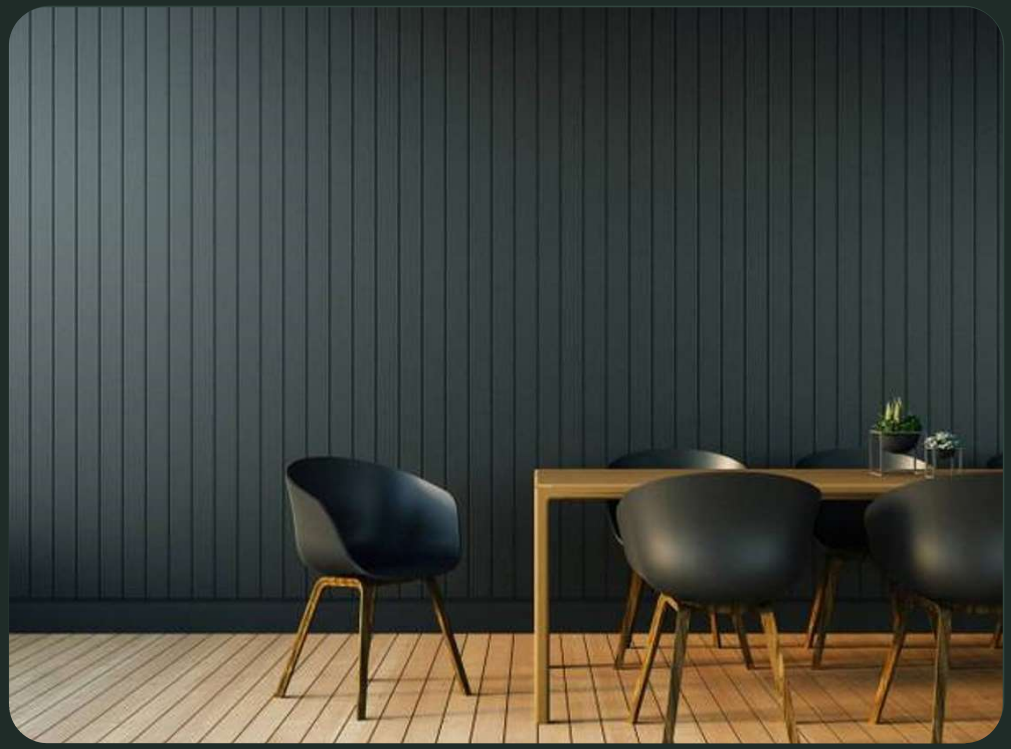
The size of a client's family is closely related to how many children they have (0.879). More children usually mean a bigger family.

Lastly, when looking at where clients live and work, there's a strong connection (0.862) between living in a different region than where they work. This means that if someone lives in a different region from where they work, they likely also live in a different city from where they work.

Lastly, the numbers showing if a client lives and works in different regions are also closely related (0.862). If someone lives in a different region than they work, there's a good chance they also live in a different city than where they work.

These connections help us understand how different things are related when clients don't have trouble paying their bills.

Conclusion



In conclusion, the insights from this project help us better understand different characteristics and behaviors of people applying for loans. When we looked at total incomes, we found some people who make much more money than most others. This shows how important it is to think about income differences when evaluating the risk of lending money and creating financial products for different income levels.

Overall, these insights help financial institutions understand loan applicants better. This allows them to create loan products that fit different needs, assess the risk of lending money more accurately, and offer financial solutions that are fair for everyone. By using these insights, lenders can make better decisions, encourage responsible lending, and help improve their customers' financial well-being.