Credit Exploratory Data Analysis

Submitted By:

Harsh Deep Jaggi

DSC 57

Objective

Credit Risk Analysis will help the company make a decision for loan approval based on the applicant's profile, which controls the loss of business to the company and avoids Financial Loss for the company.

Problem Statement

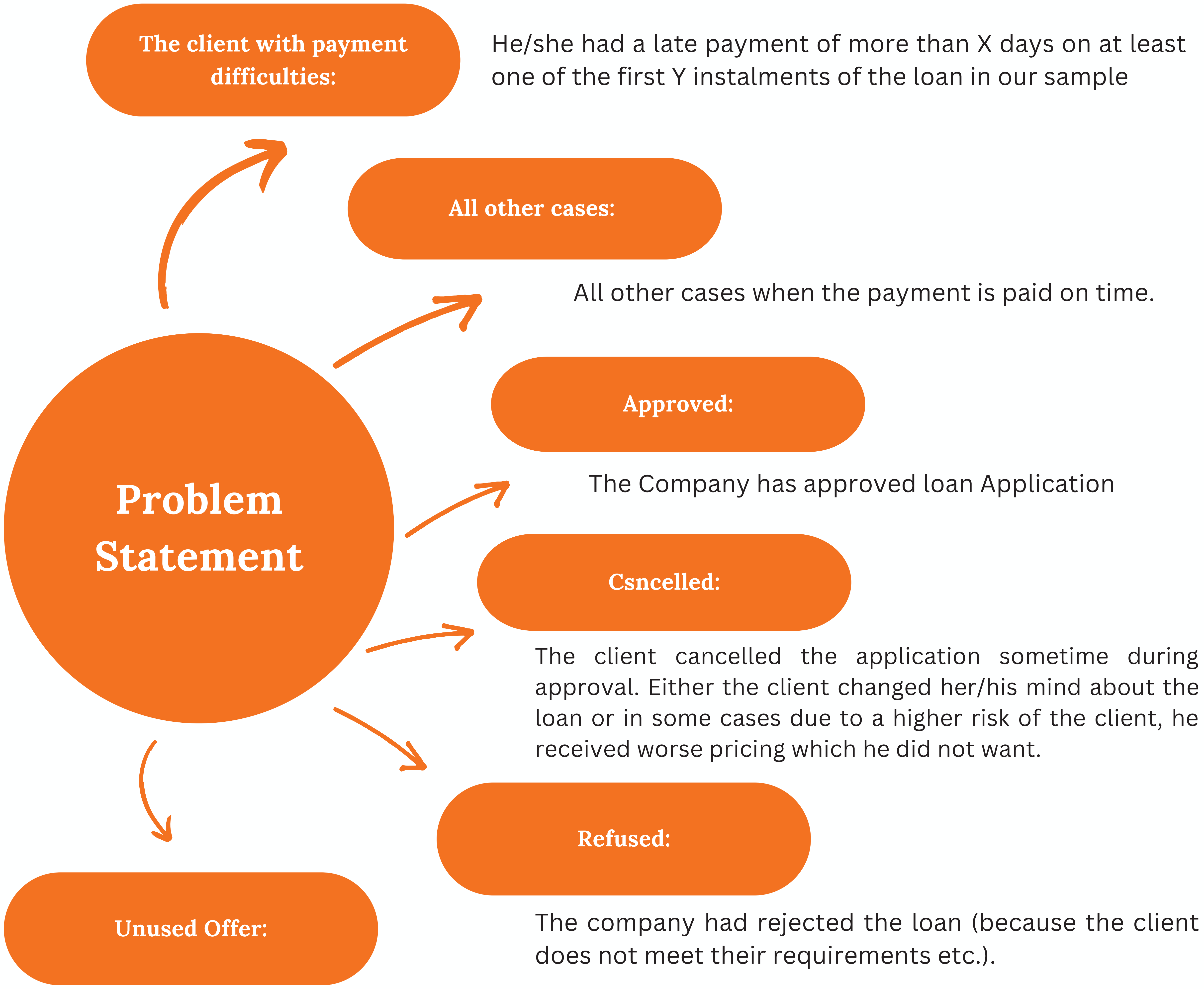
The Loan-providing companies find it hard to give loans to people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company that specialises in lending various types of loans 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 contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):



Problem Statement

The client with payment difficulties:

He/she had a late payment of 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.

Approved:

The Company has approved loan Application

Csncelled:

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.

Business Objective

This case study aims to identify patterns that 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 the 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 that 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.

Datasets

This dataset has 3 files as explained below:

'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.

'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.

'columns_description.csv'

is data dictionary which describes the meaning of the variables.

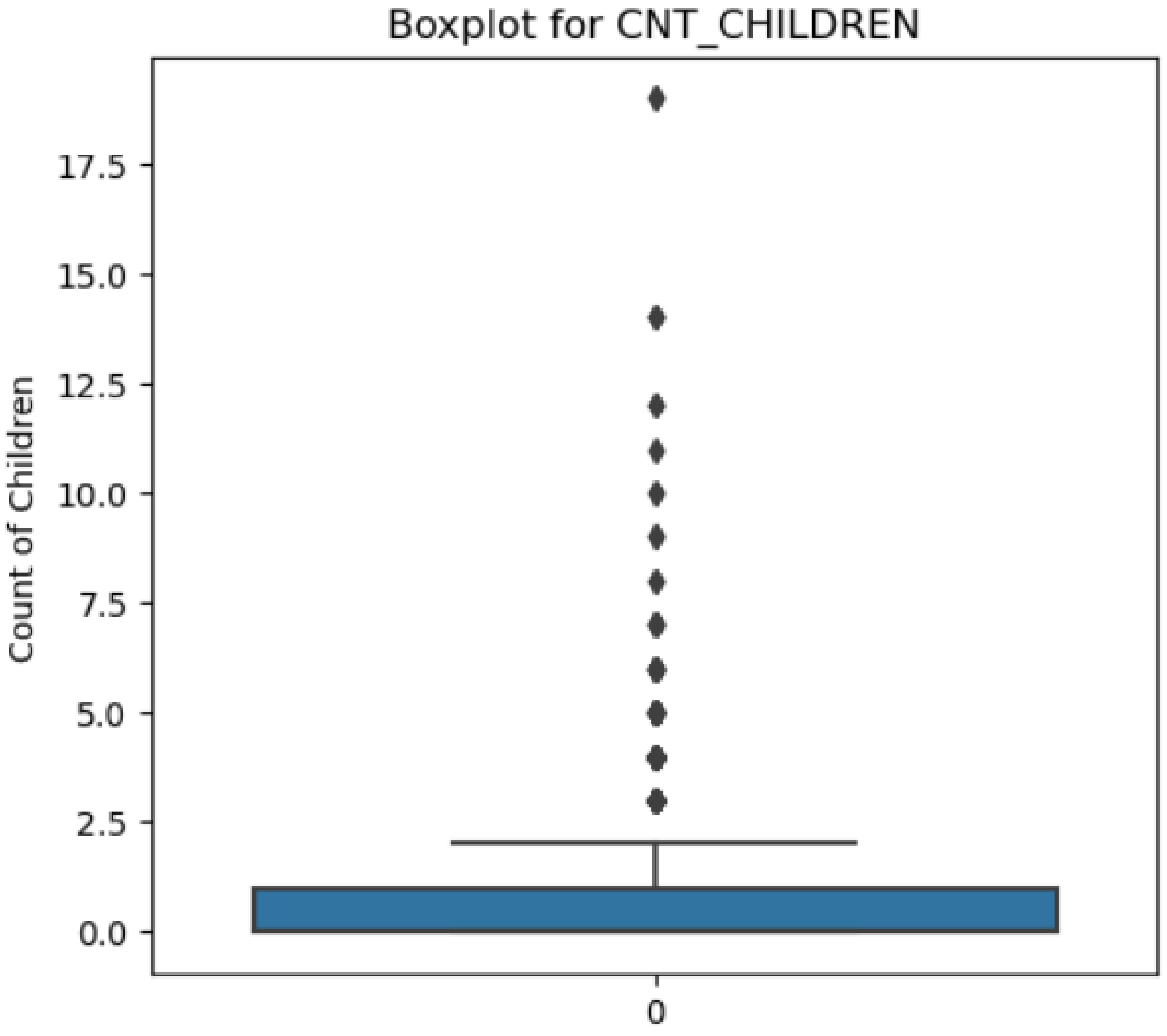
Steps:

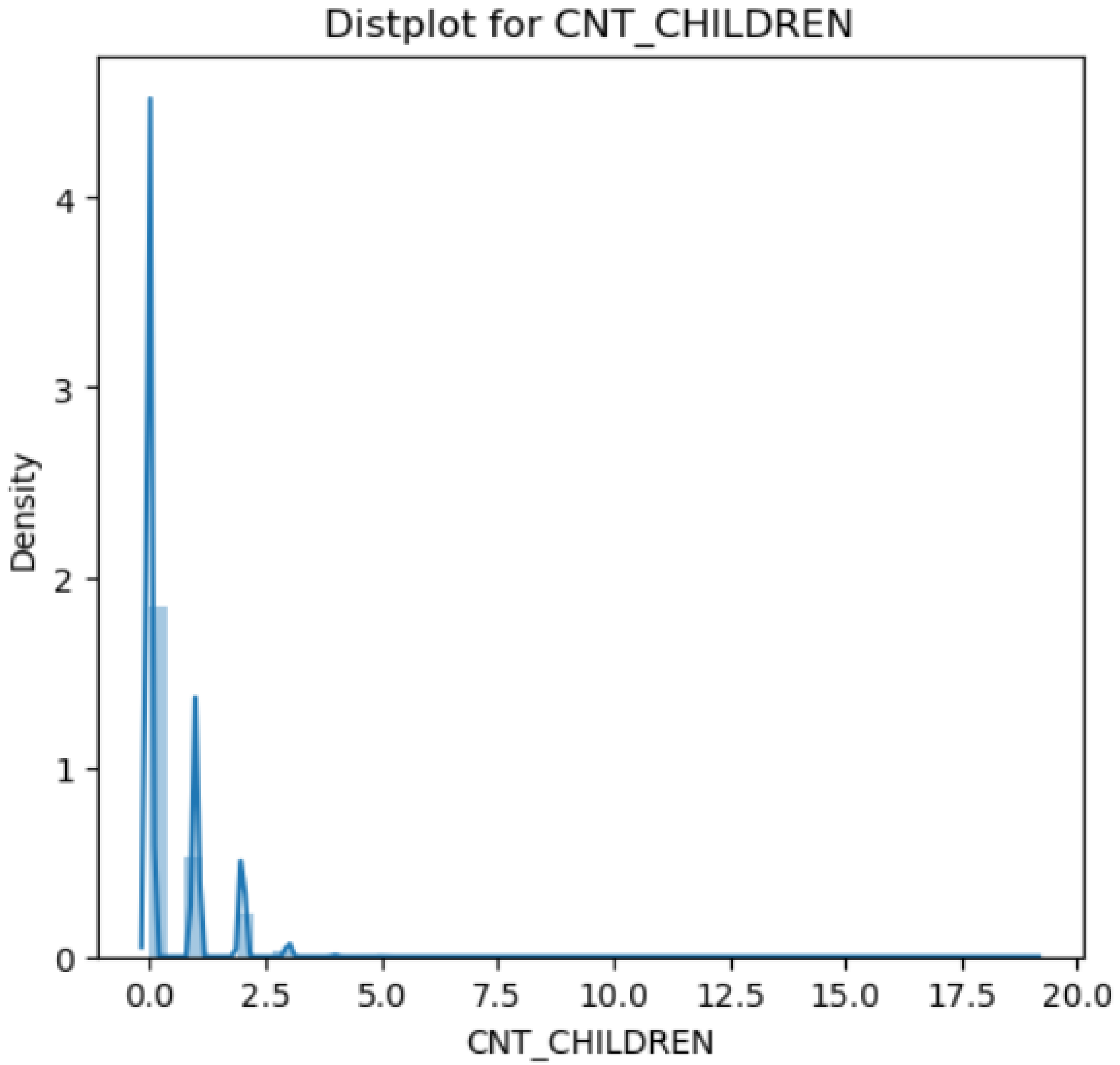
- Understanding business problem
- Import the data
- Understanding the data
- Check the structure of the data
- Data Transformation
- EDA:
    - Univariate Analysis
    - Bivariate Analysis
    - Multivariate Analysis
    - Correlation
- Application of previous data to current data
- Risk & Recommendations
- Conclusion

Application Data Analysis

contains all the information of the client at the time of application.
The data is about whether a client has payment difficulties.

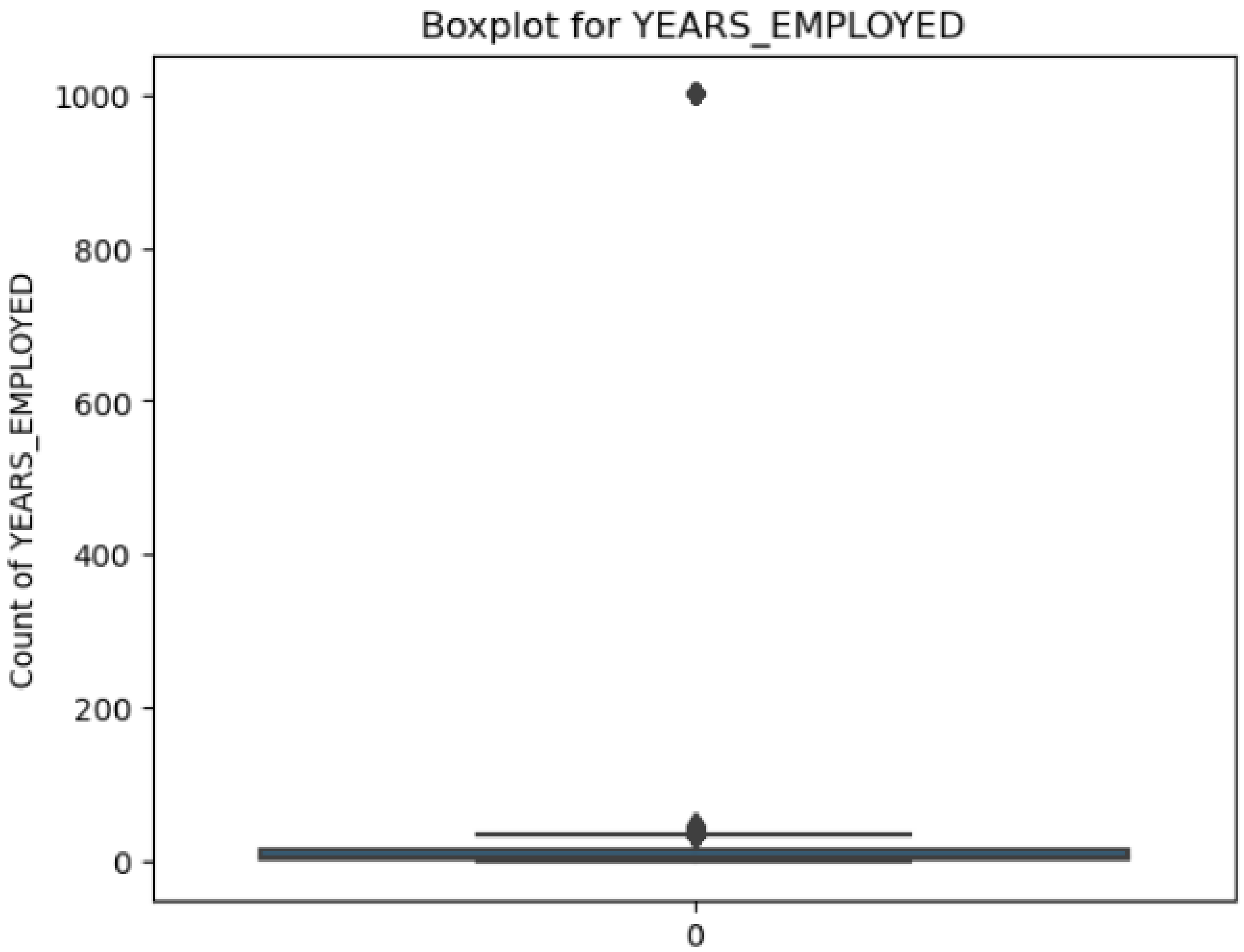
Analysis of CNT_CHILDREN

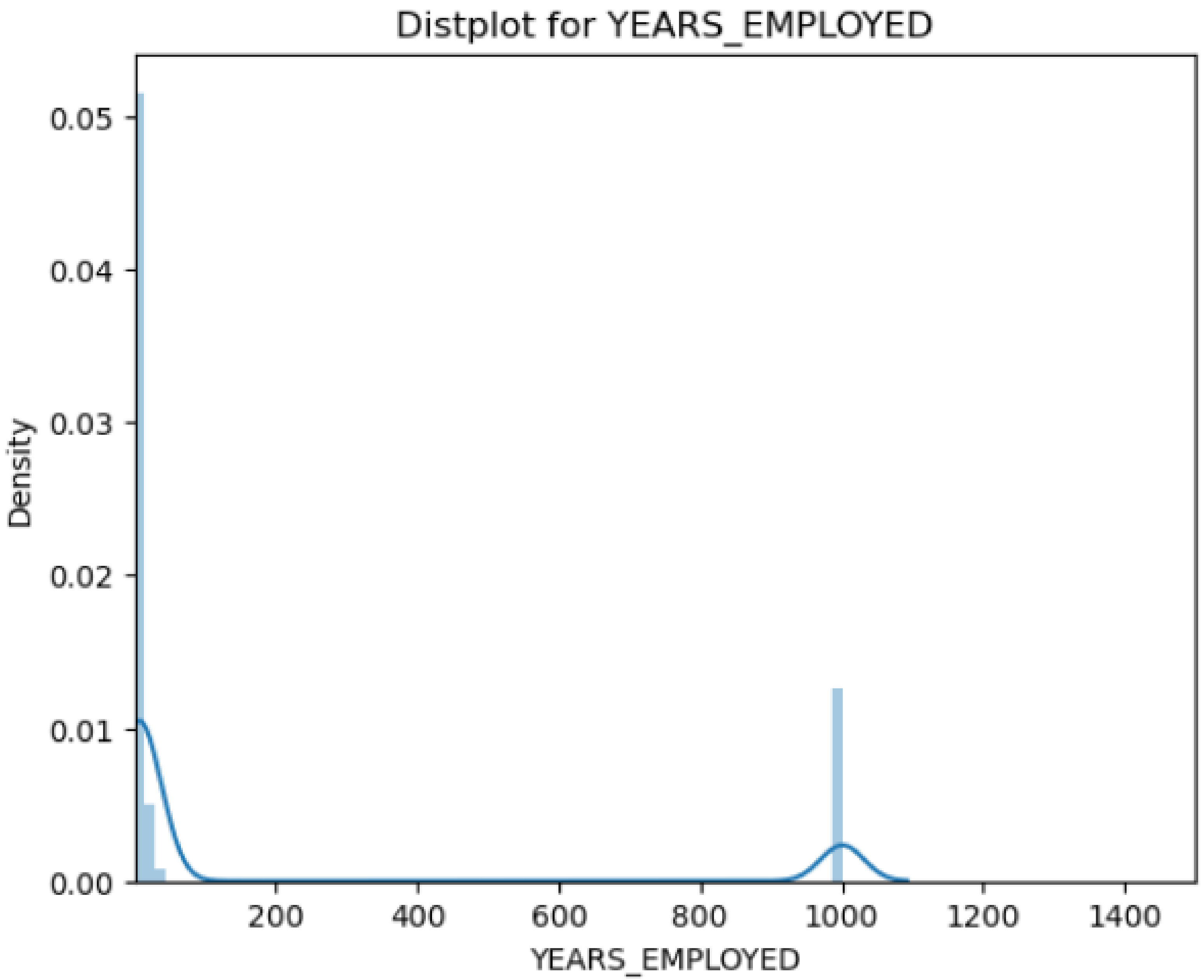

Boxplot for CNT_CHILDREN


Distplot for CNT_CHILDREN

1. Boxplot & Distplot shows that the values above 2.5 are outliers.
2. People who have 3 or more children are the outlier cases.

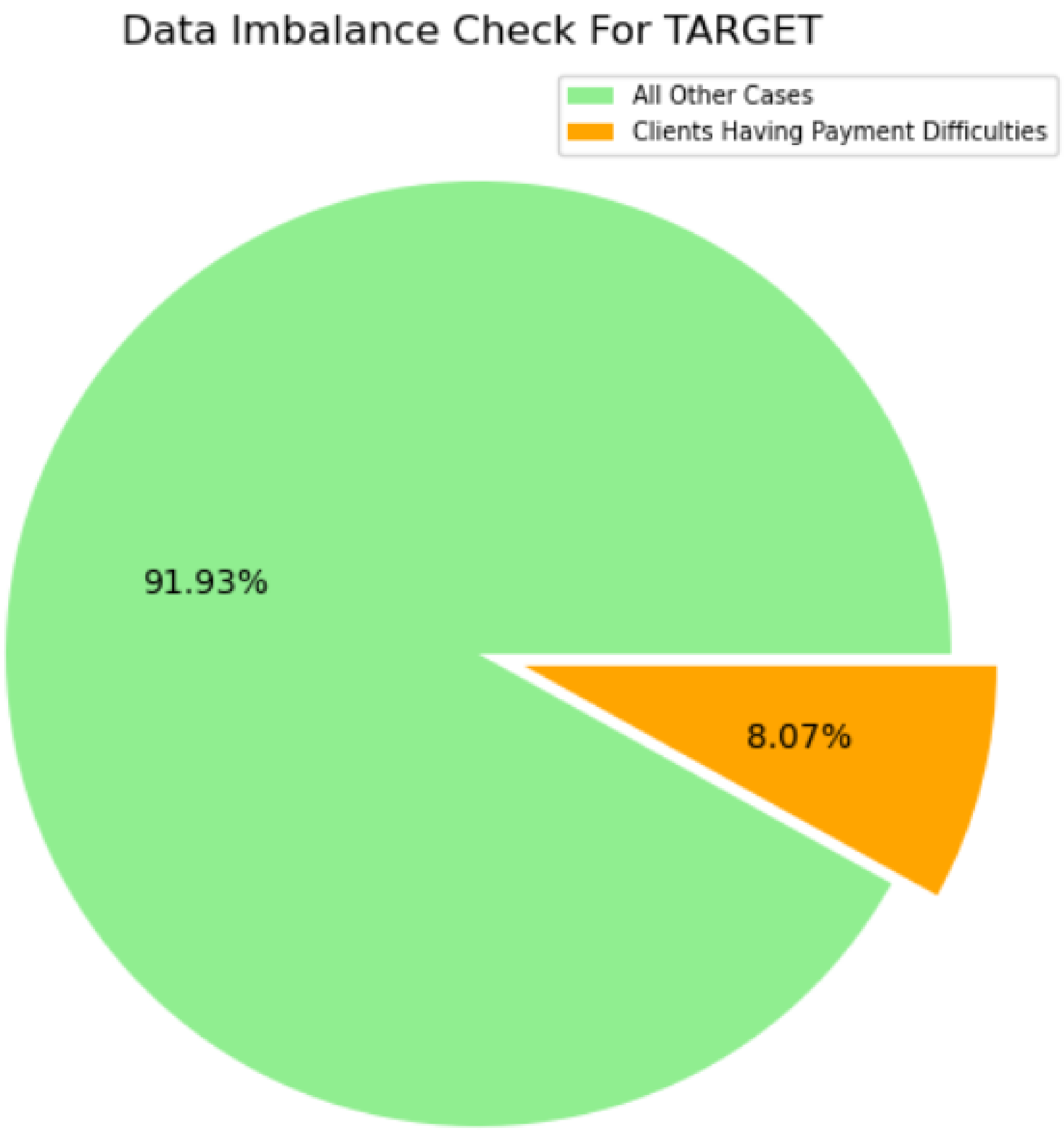
Analysis of YEARS_EMPLOYED

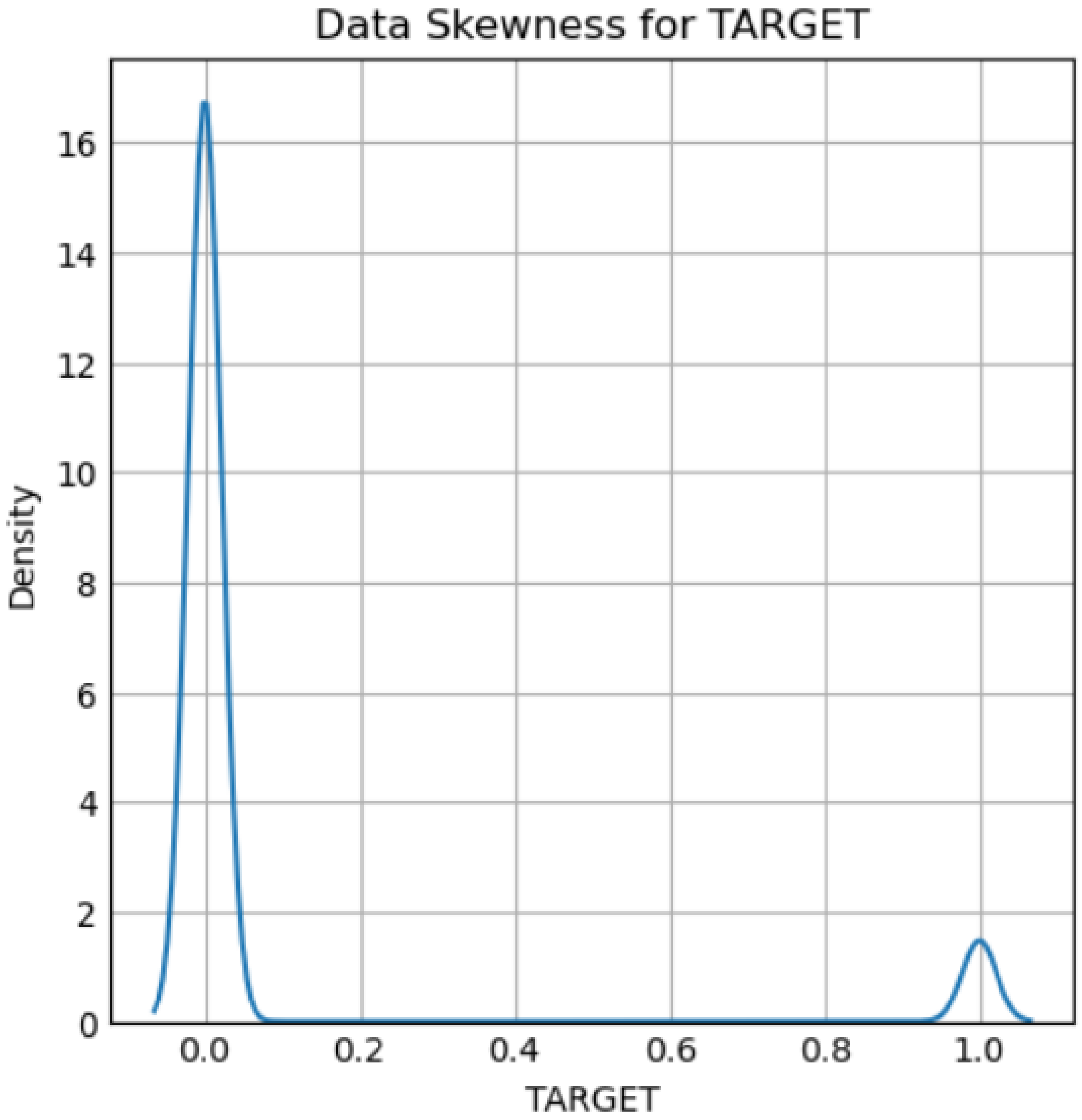




1. 55374 outliers are there in YEARS_EMPLOYED which has 1001 Years which is not possible at all.
2. Any value which is above 49 YEARS_EMPLOYED is an outlier.

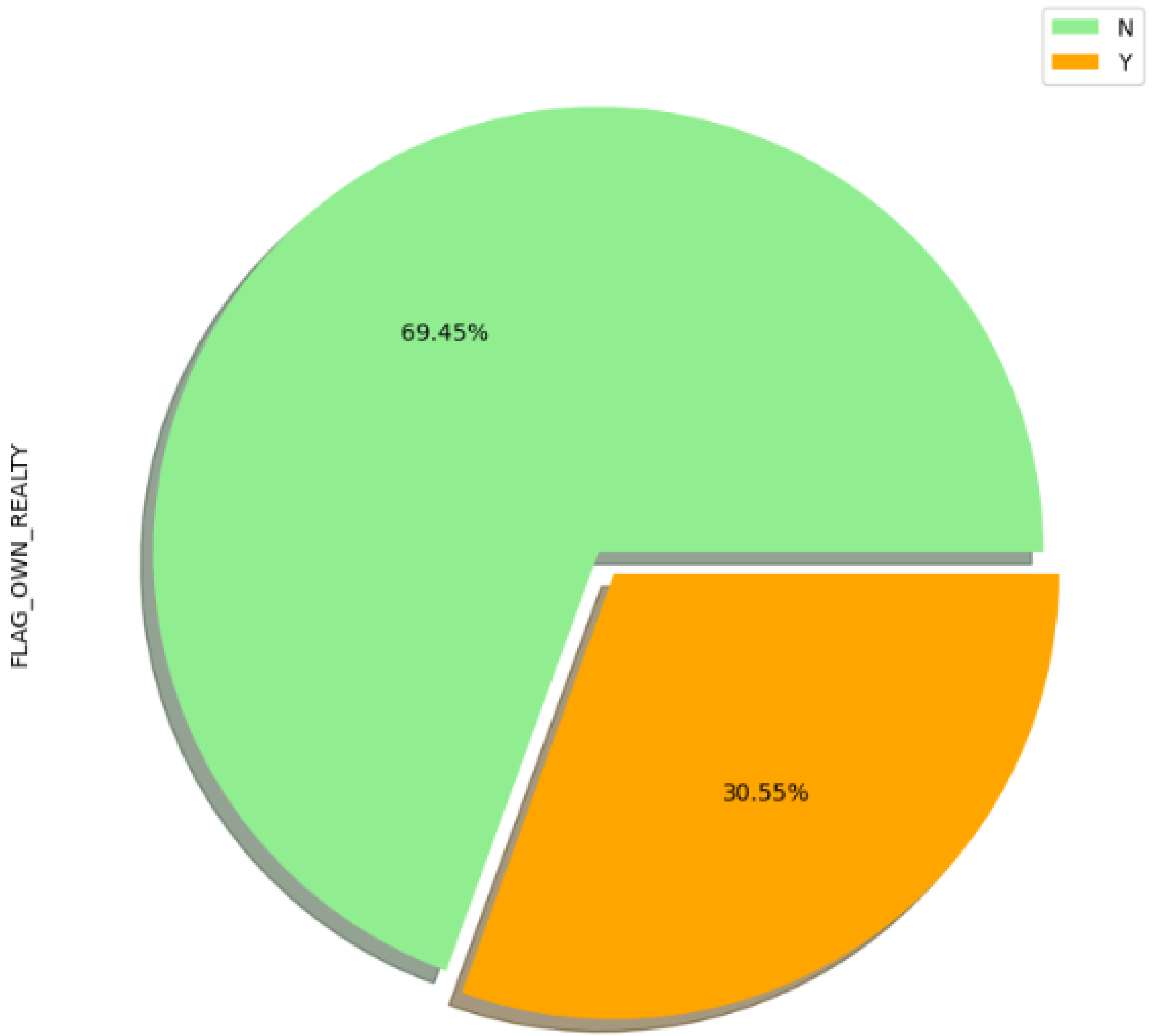
Data Imbalance & Skewness Check



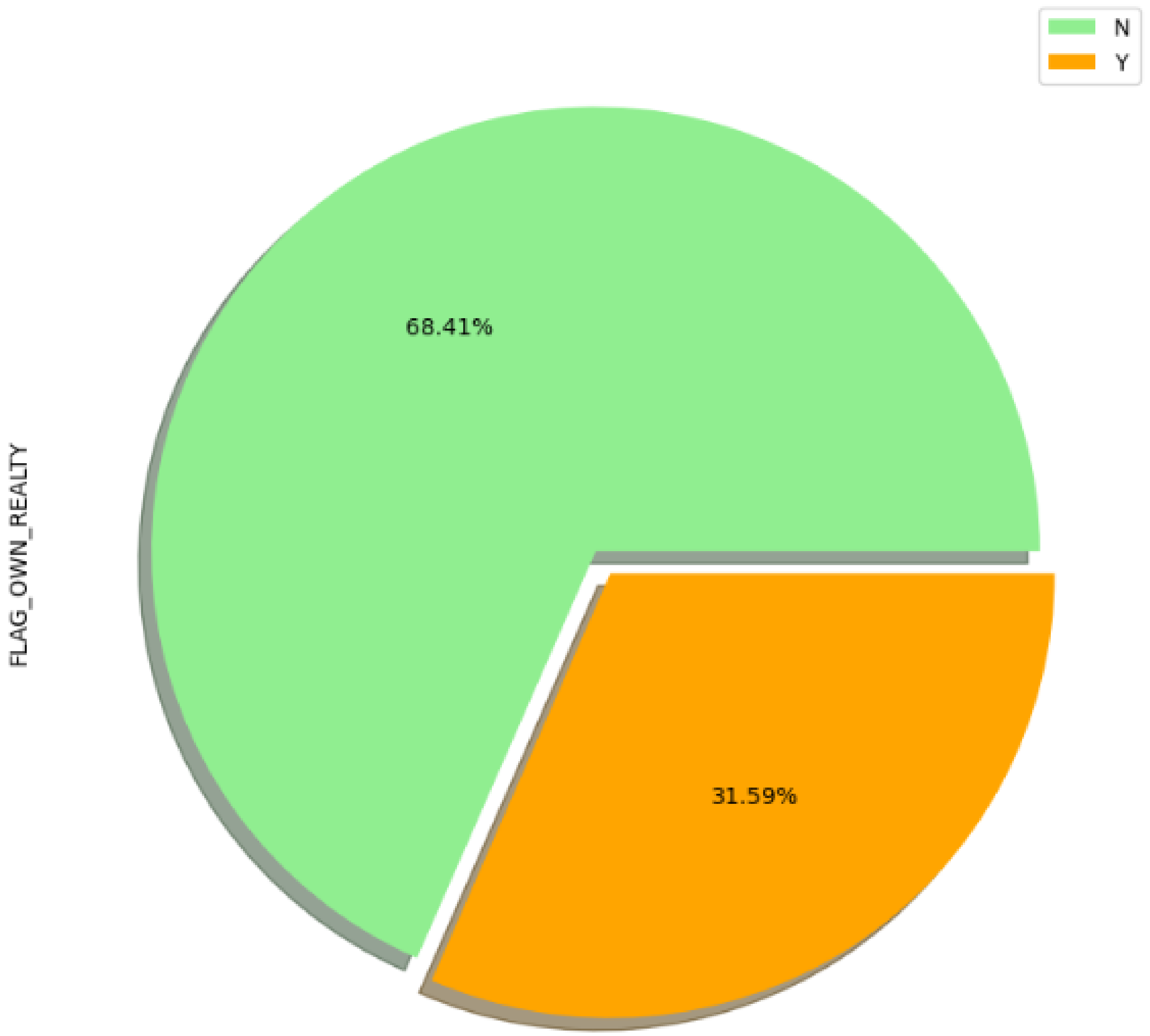


1. The Data Imbalance Ratio is 11.39: 1
2. From the above pie chart we can see that data is imbalanced.
3. It is observed that 8.07% of clients are facing payment difficulties whereas 91.93% of clients paid on time.
4. Data is Right Skewed.

Analysis of FLAG_OWN_REALTY



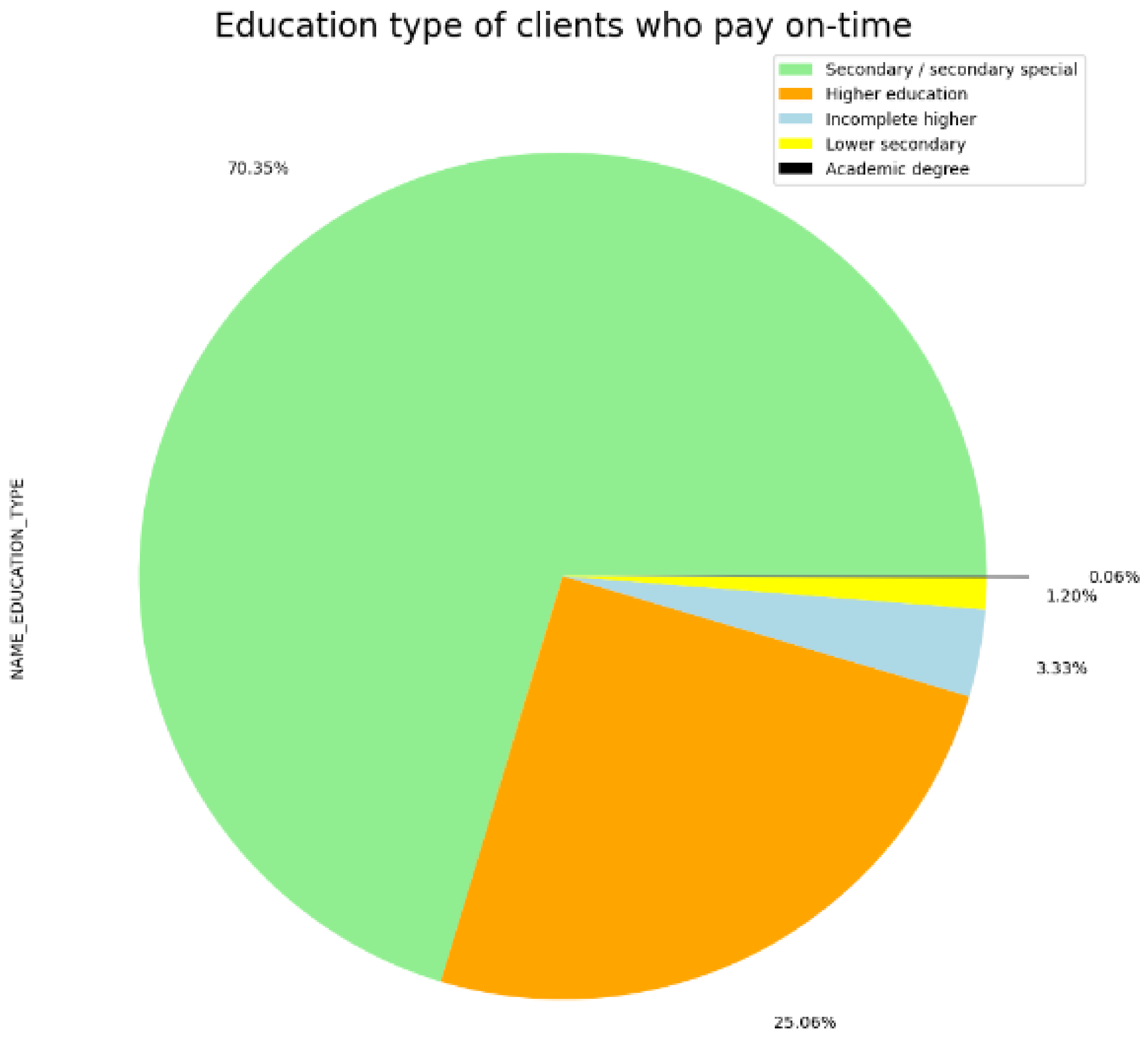
Distribution of Property of clients for all other cases

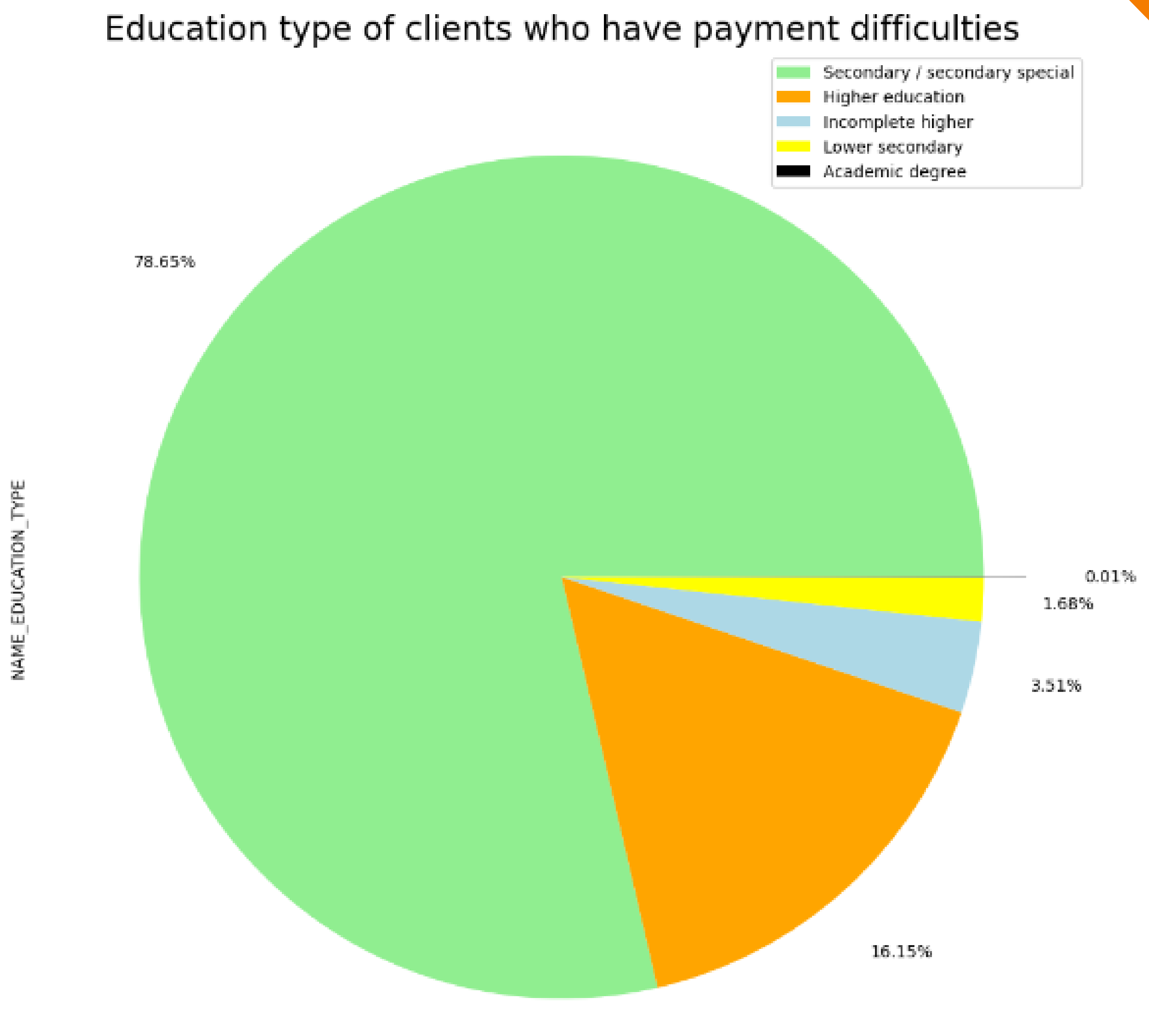


Distribution of Property of clients who have payment difficulties

1 . Approx. 70% of clients don’t own a house.

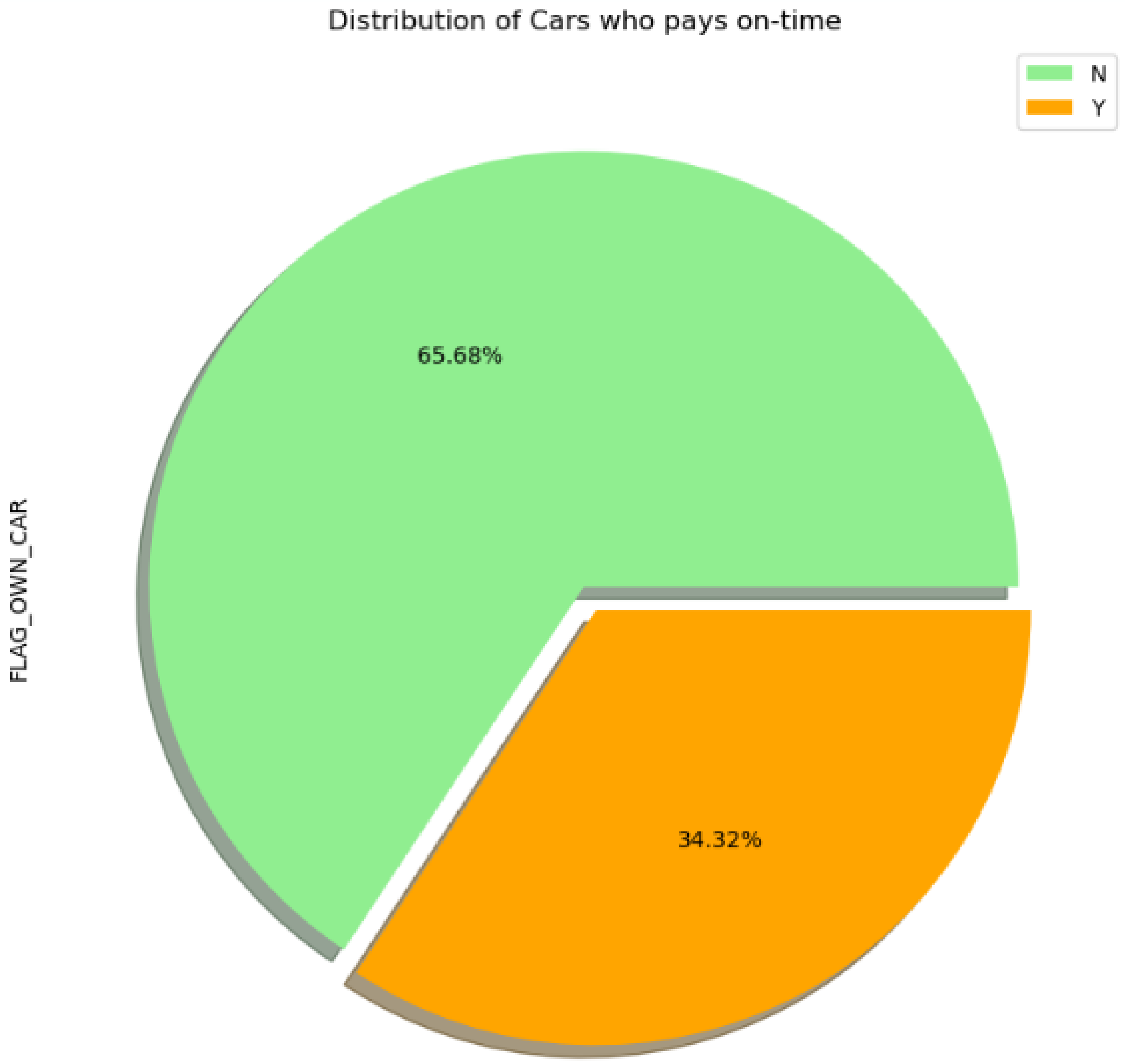
Analysis of NAME_EDUCATION_TYPE


Education type of clients who pay on-time

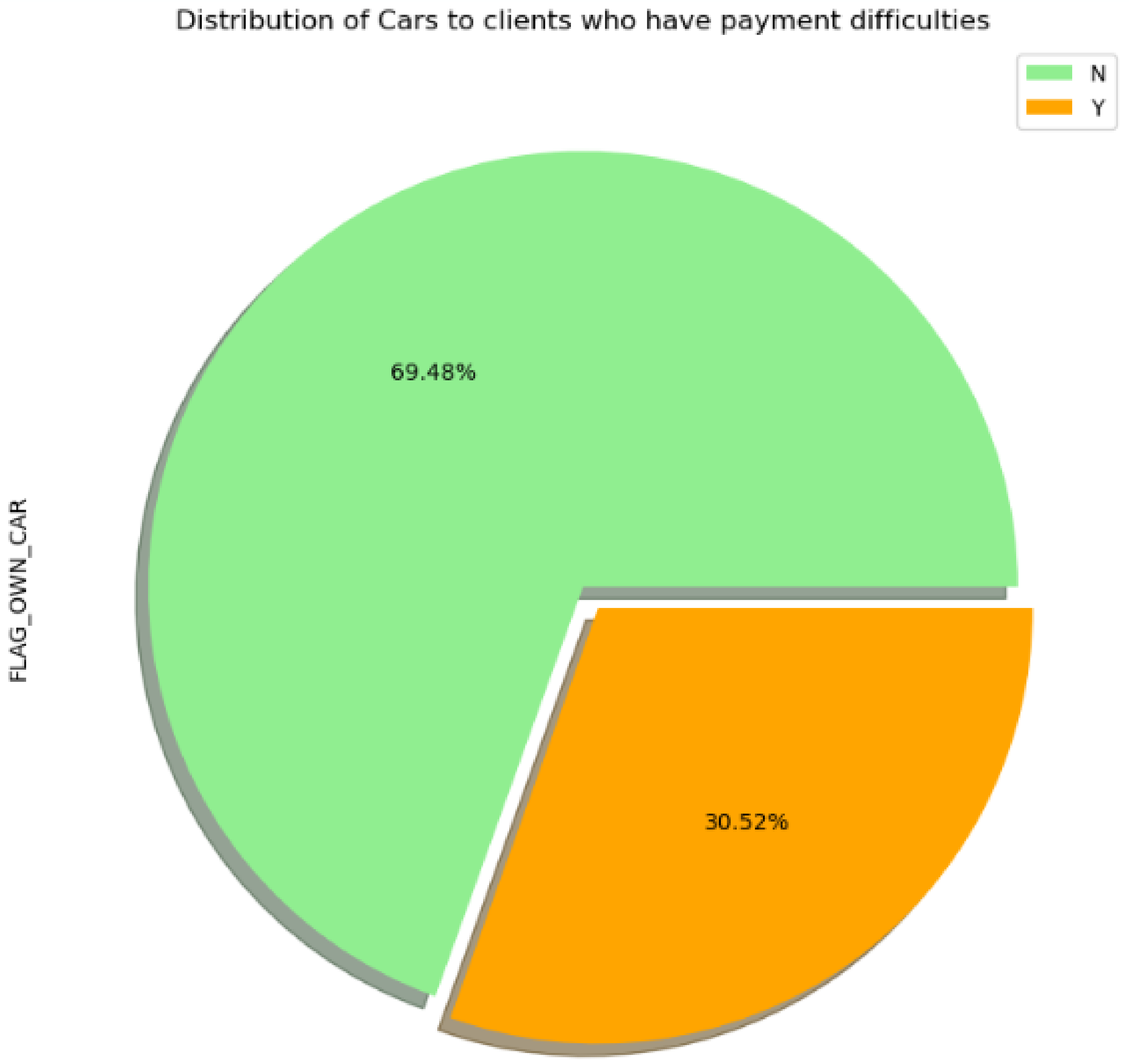

Education type of clients who have payment difficulties

1. Client's with Secondary / Secondary special have more payment difficulties followed by Higher Education.

Analysis of FLAG_OWN_CAR



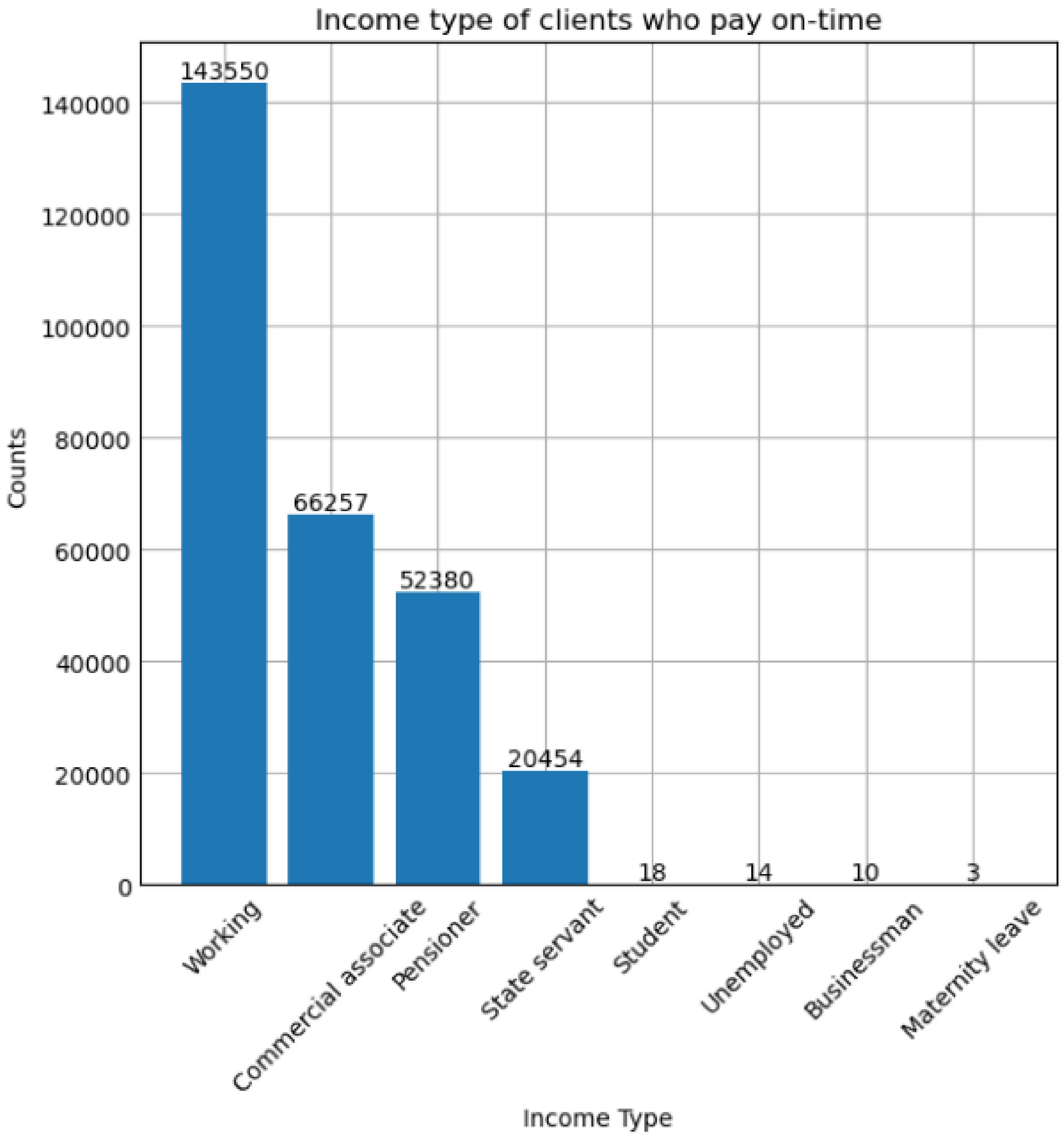
Distribution of Cars who pays on-time

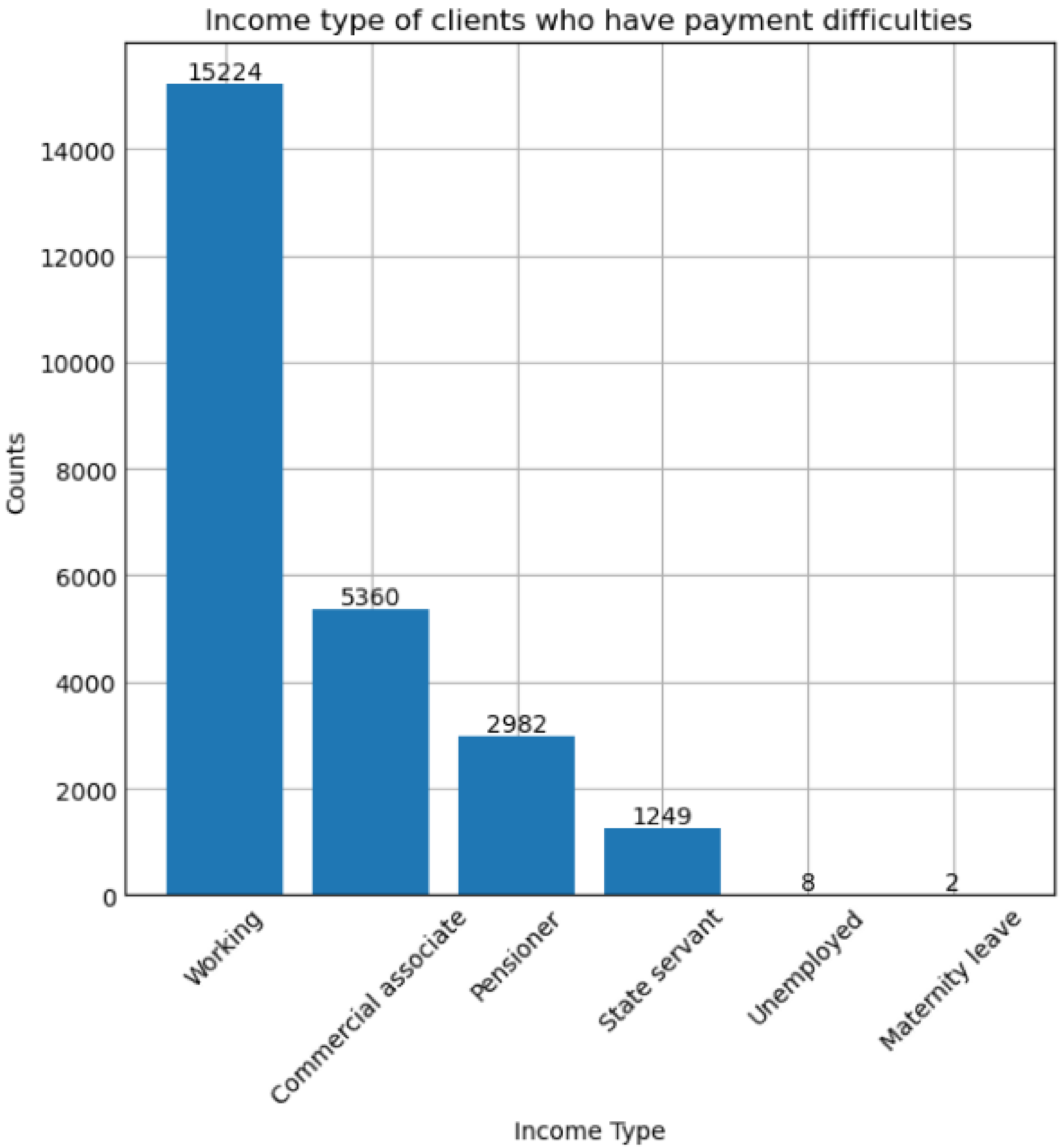


Distribution of Cars to clients who have payment difficulties

1. 69.48% are those clients who do not own a car and defaulted.

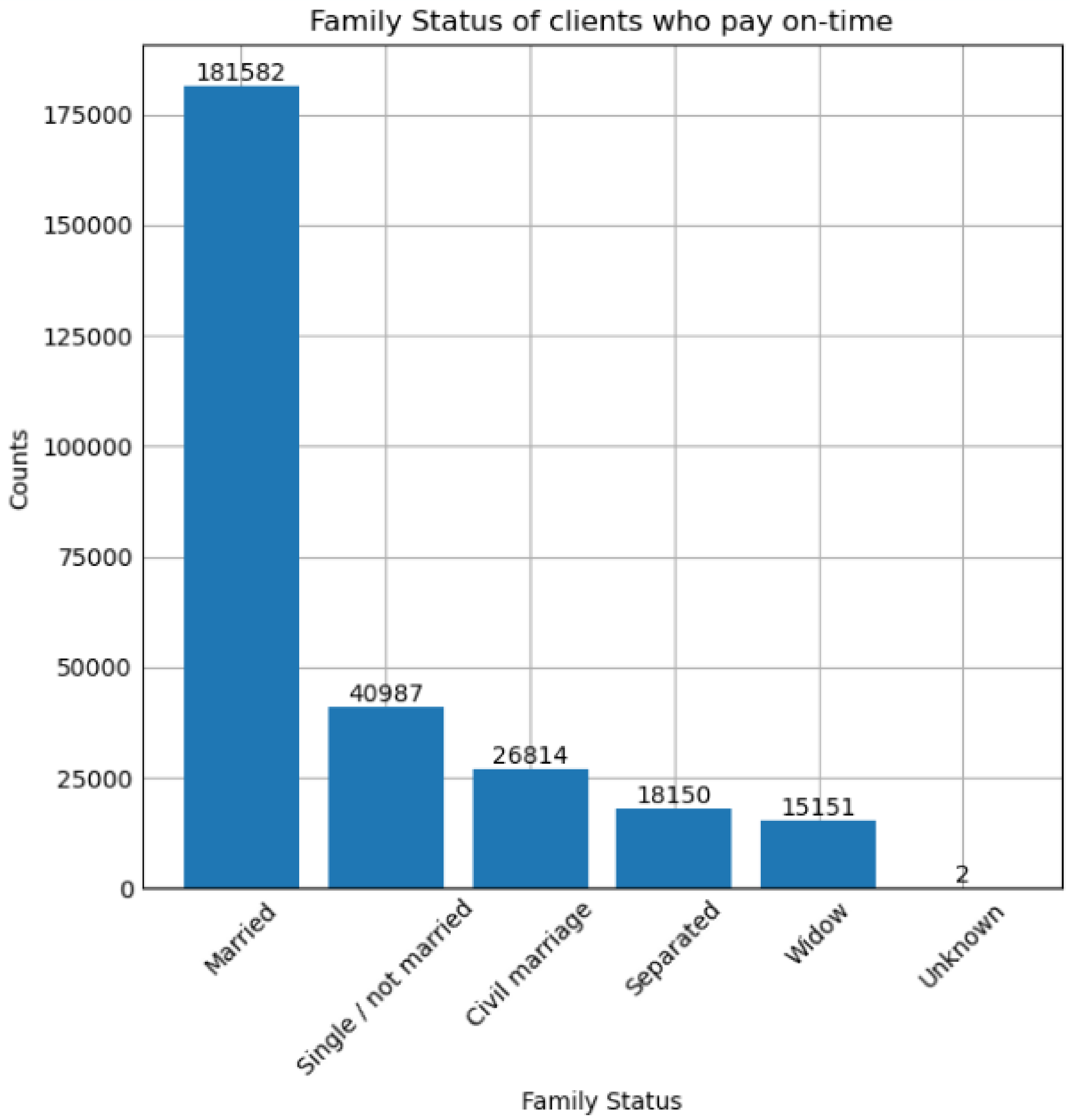
Analysis of NAME_INCOME_TYPE

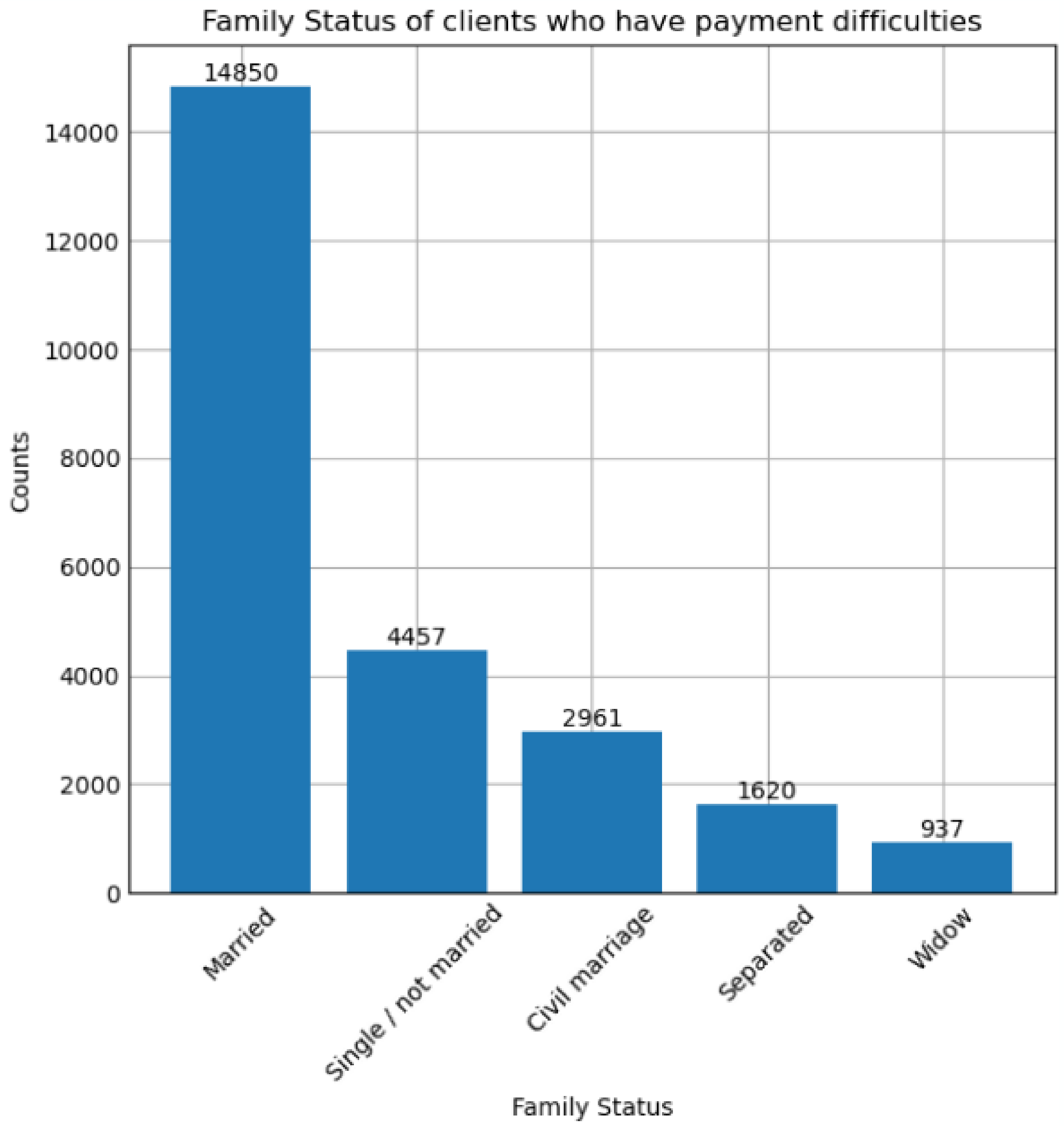

Income type of clients who pay on-time


Income type of clients who have payment difficulties

1. Working professionals are having more payment difficulties.
2. Student & Businessmen pay their amount on time.

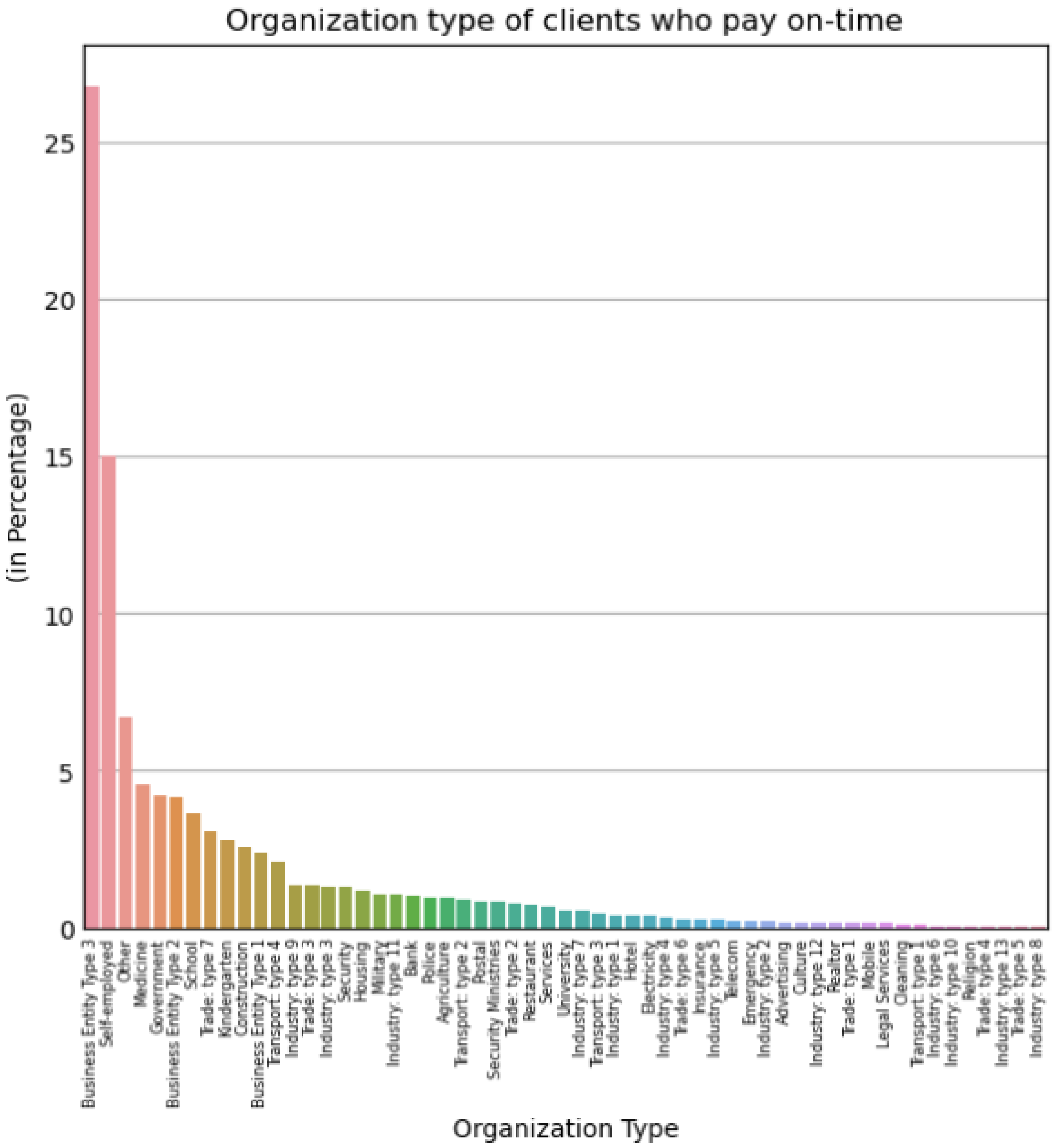
Analysis of NAME_FAMILY_STATUS

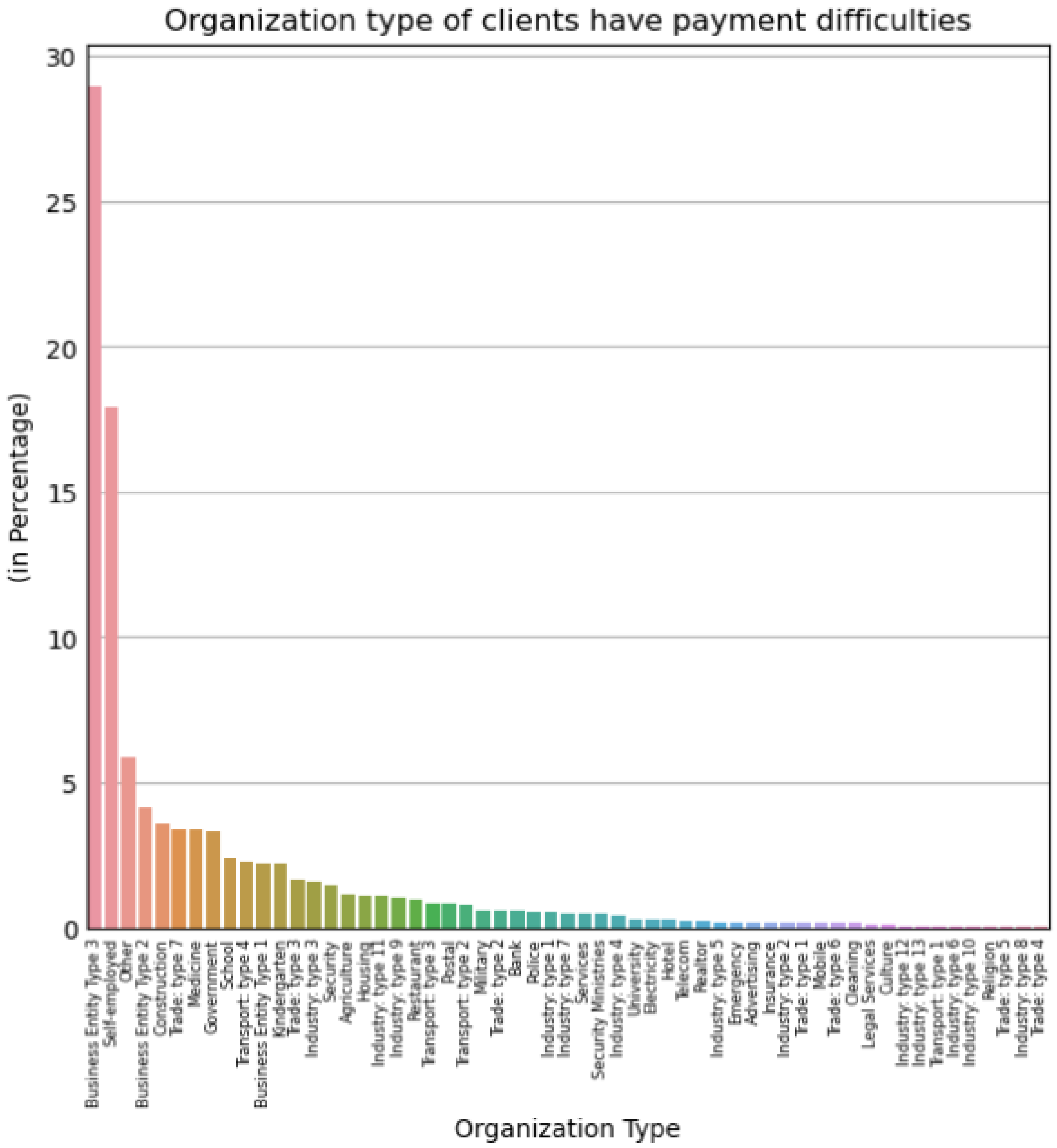




1. Married people defaulted the most followed by Single/ Not Married.
2. Widow defaulted the least.
3. 'Married' OR 'Widow' do on-time payments better. However, this is a Weak Correlation.

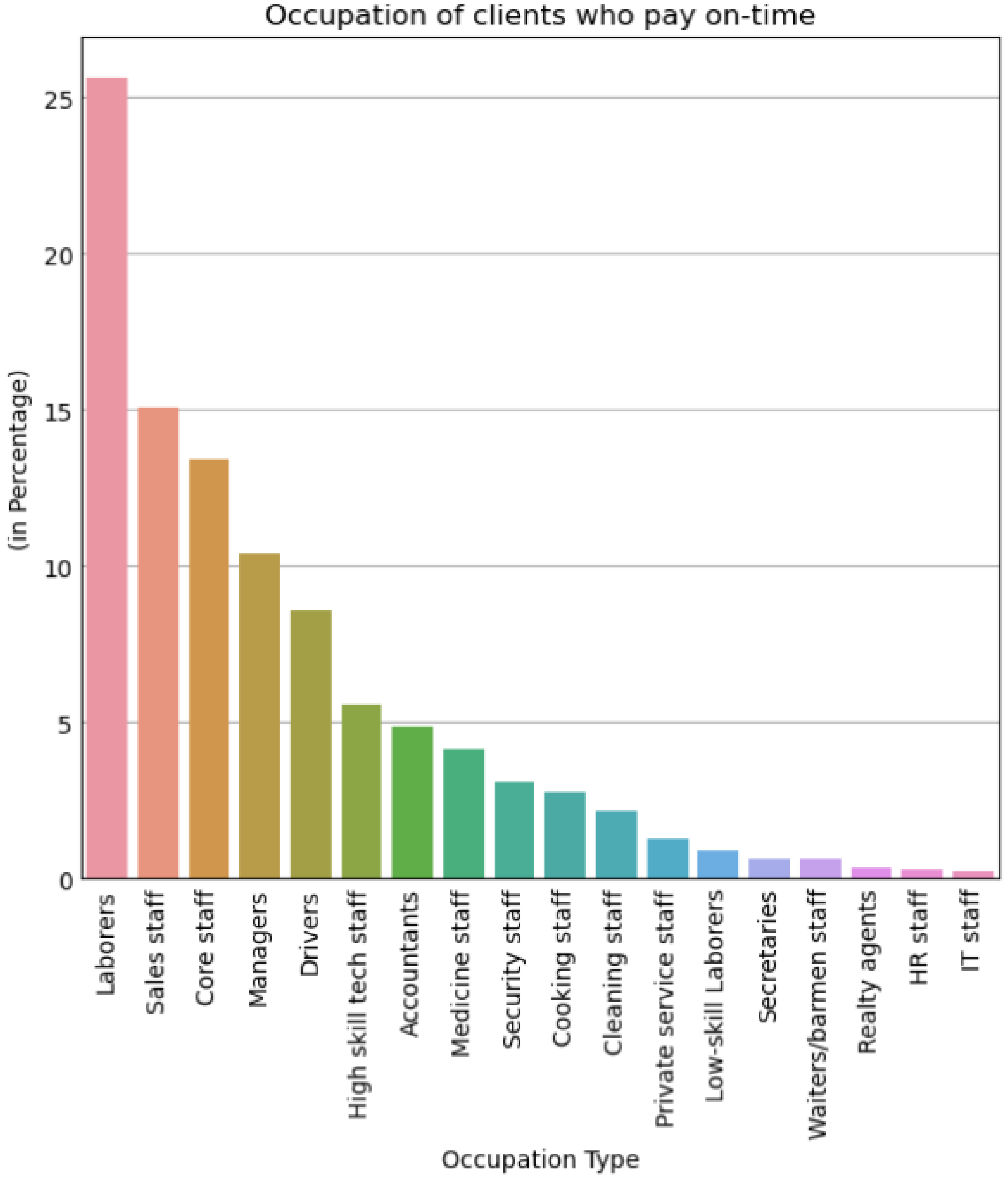
Analysis of ORGANIZATION_TYPE

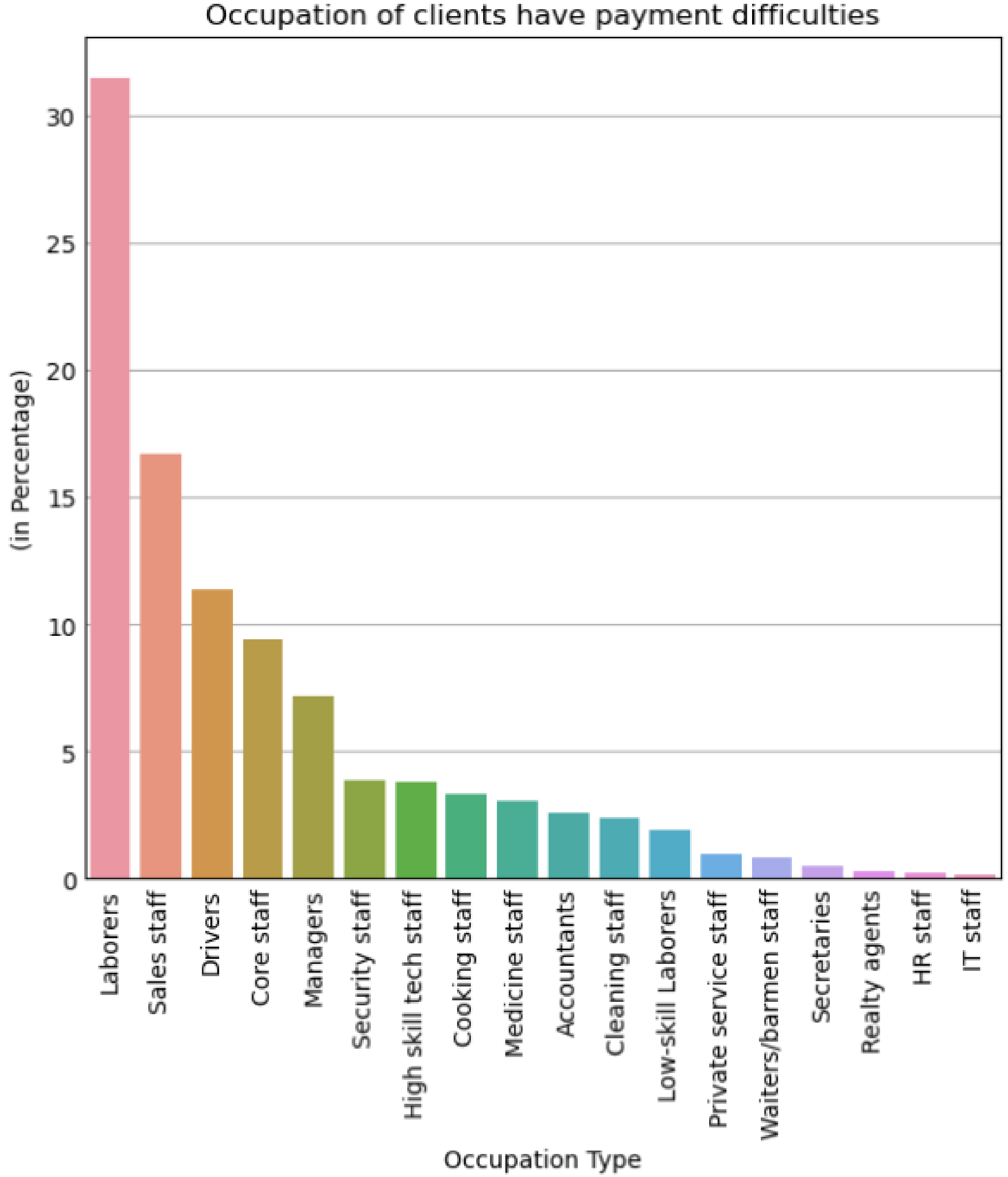

Organization type of clients who pay on-time


Organization type of clients have payment difficulties

1. Approx. 29% of People who are of "Business Entity Type 3" in the Organization column are defaulted the most.
2. 17-18% of those who are "Self Employed" are on the 2nd while analysing the data.

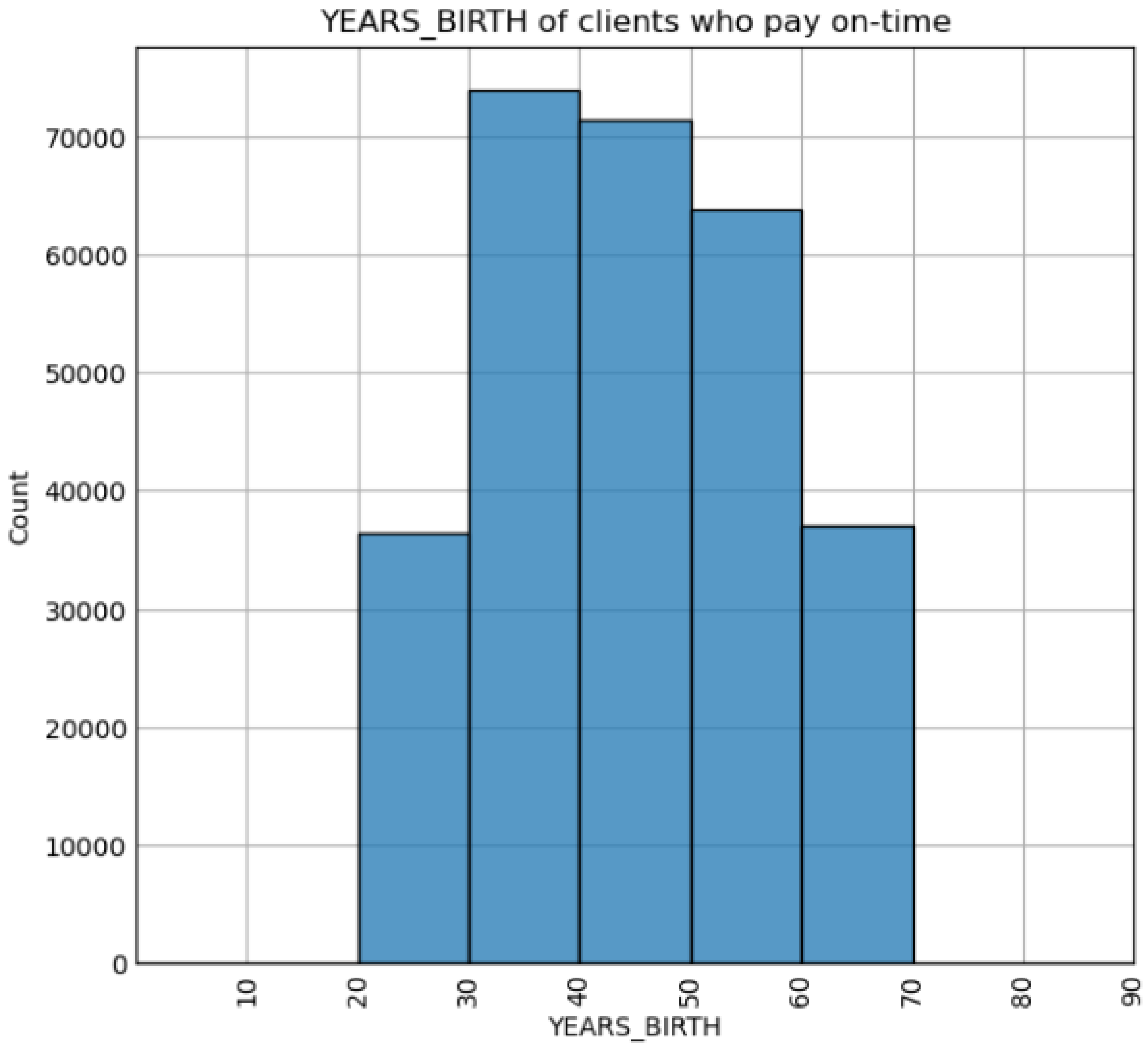
Analysis of OCCUPATION_TYPE

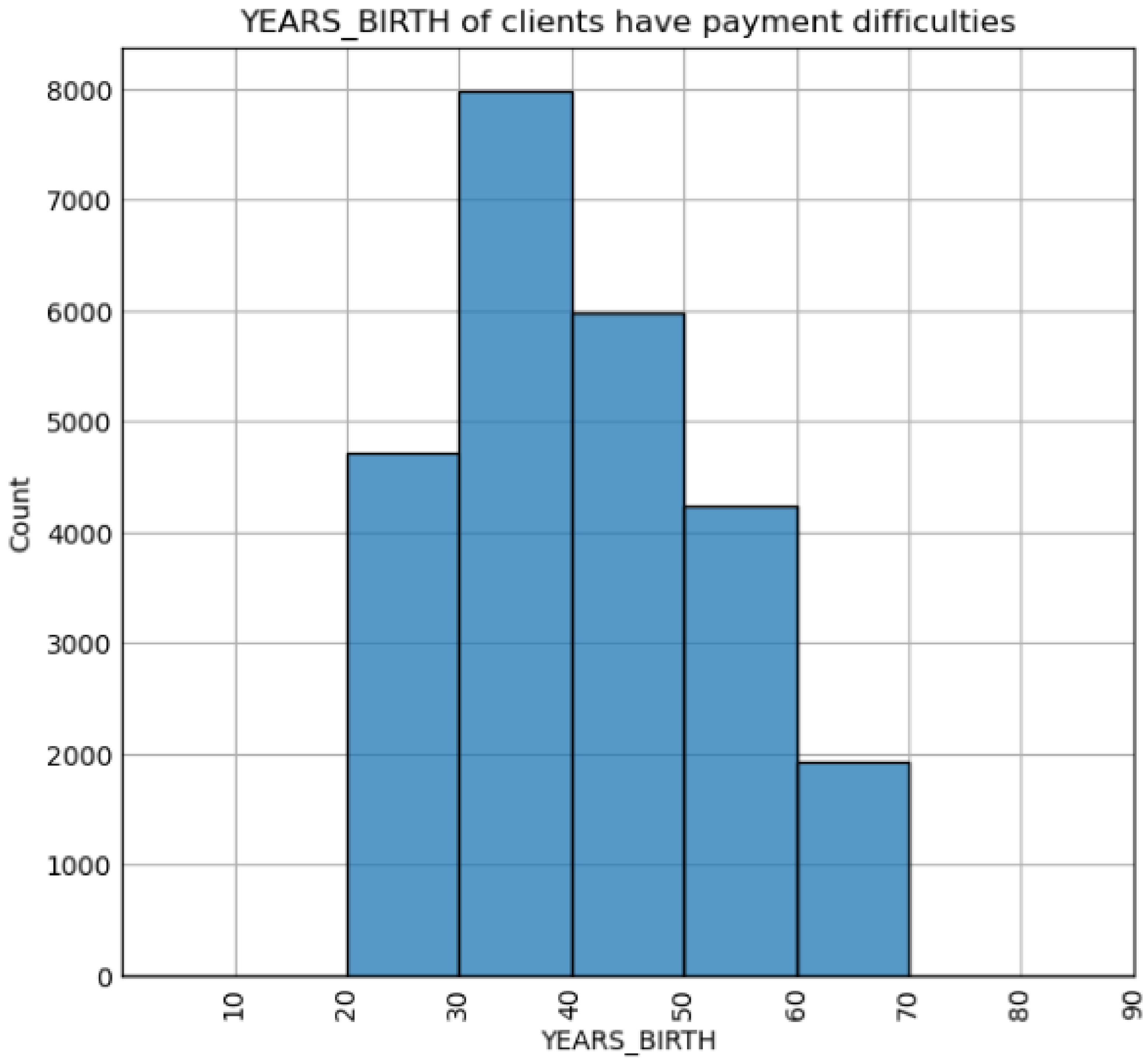

Occupation of clients who pay on-time


Occupation of clients have payment difficulties

1. Labourers are the most defaulted occupation type among all occupations, with approx. 31%, whereas sales staff have approx. 17%, and drivers have approx. 11% defaulters.
2. IT staff are the least defaulters.

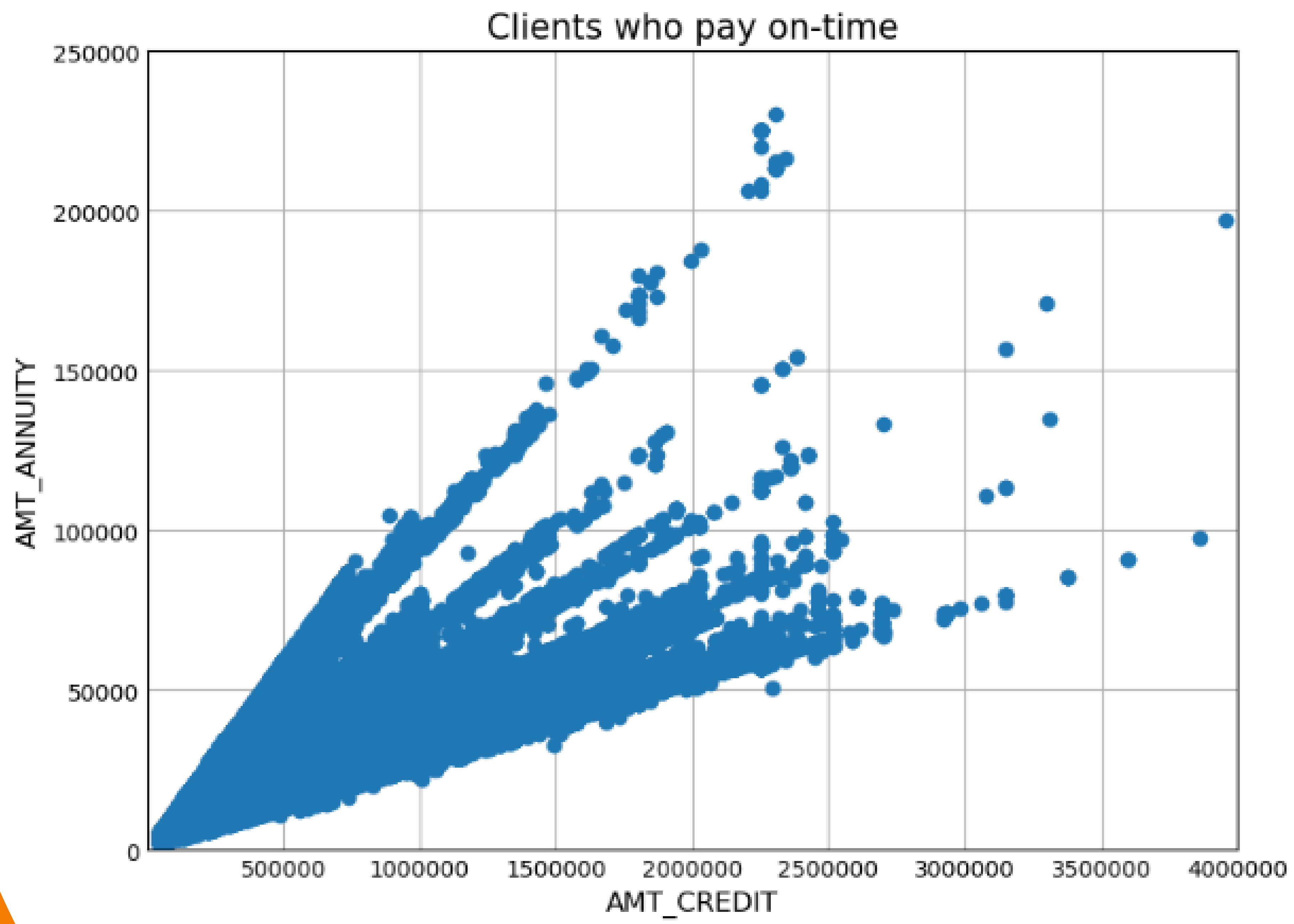
Analysis of YEARS_BIRTH

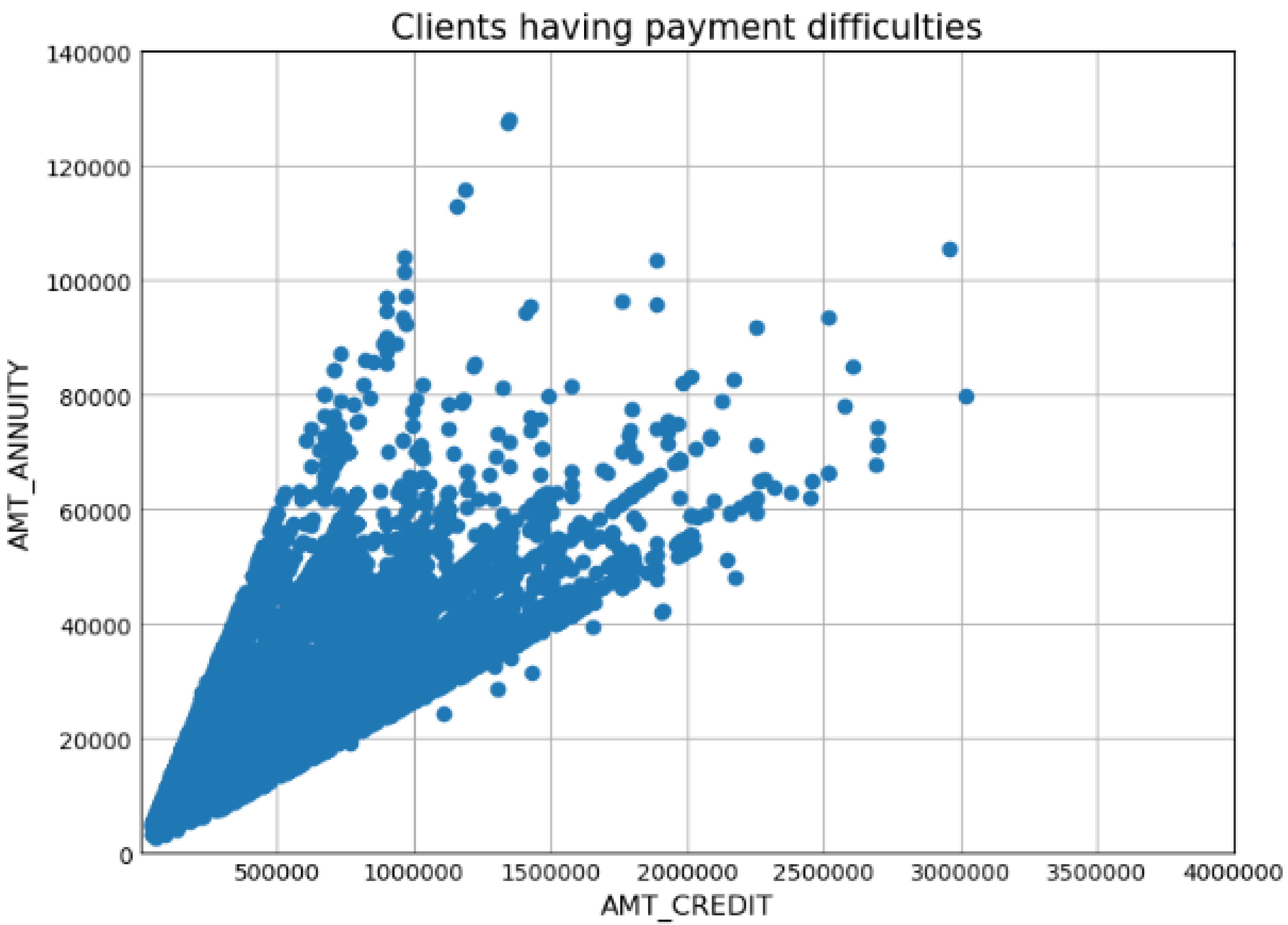




1. People between 30-40 years have more payment difficulties.
2. People over the age of 60 make timely payments.

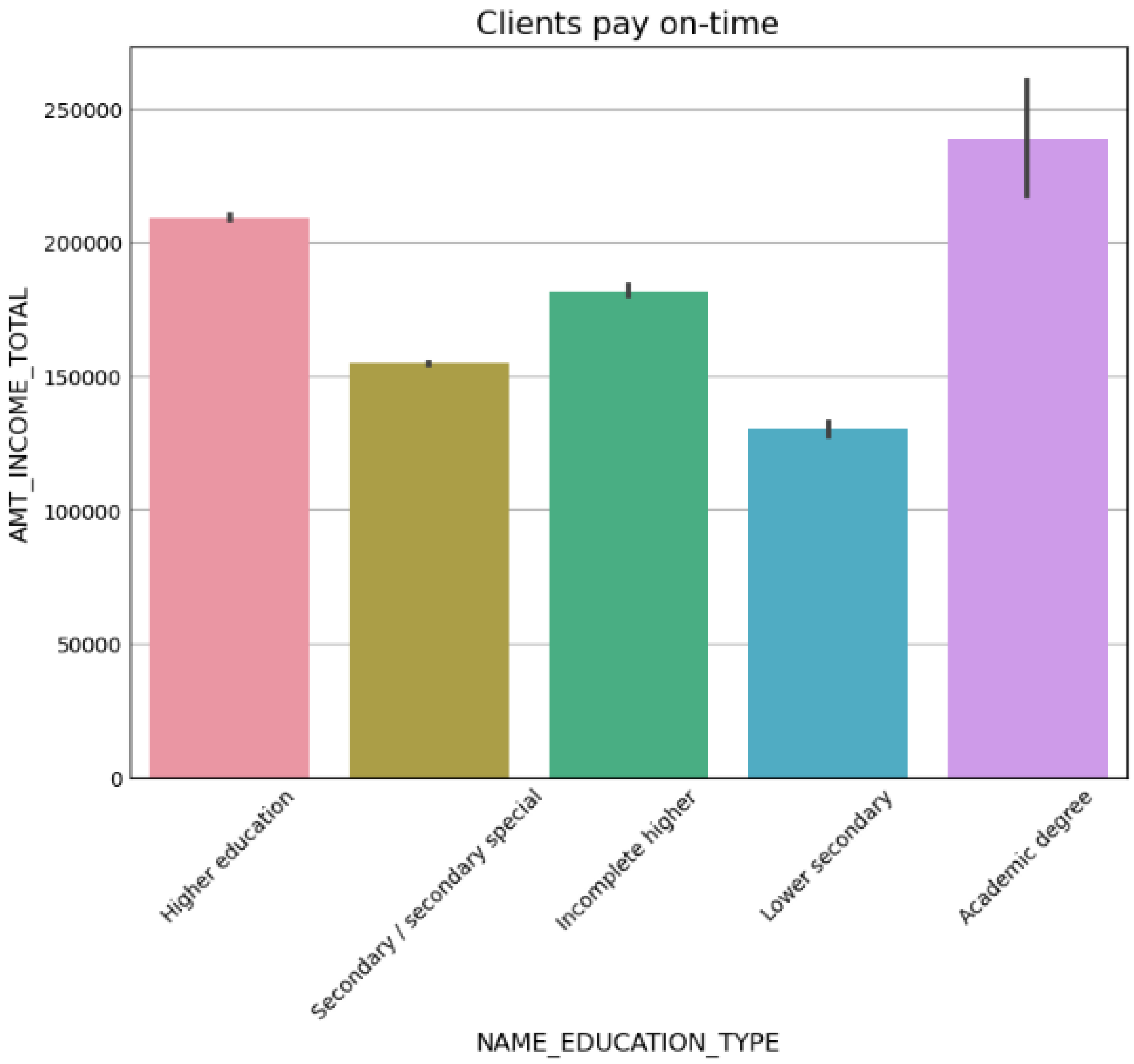
AMT_ANNUITY V/s AMT_CREDIT

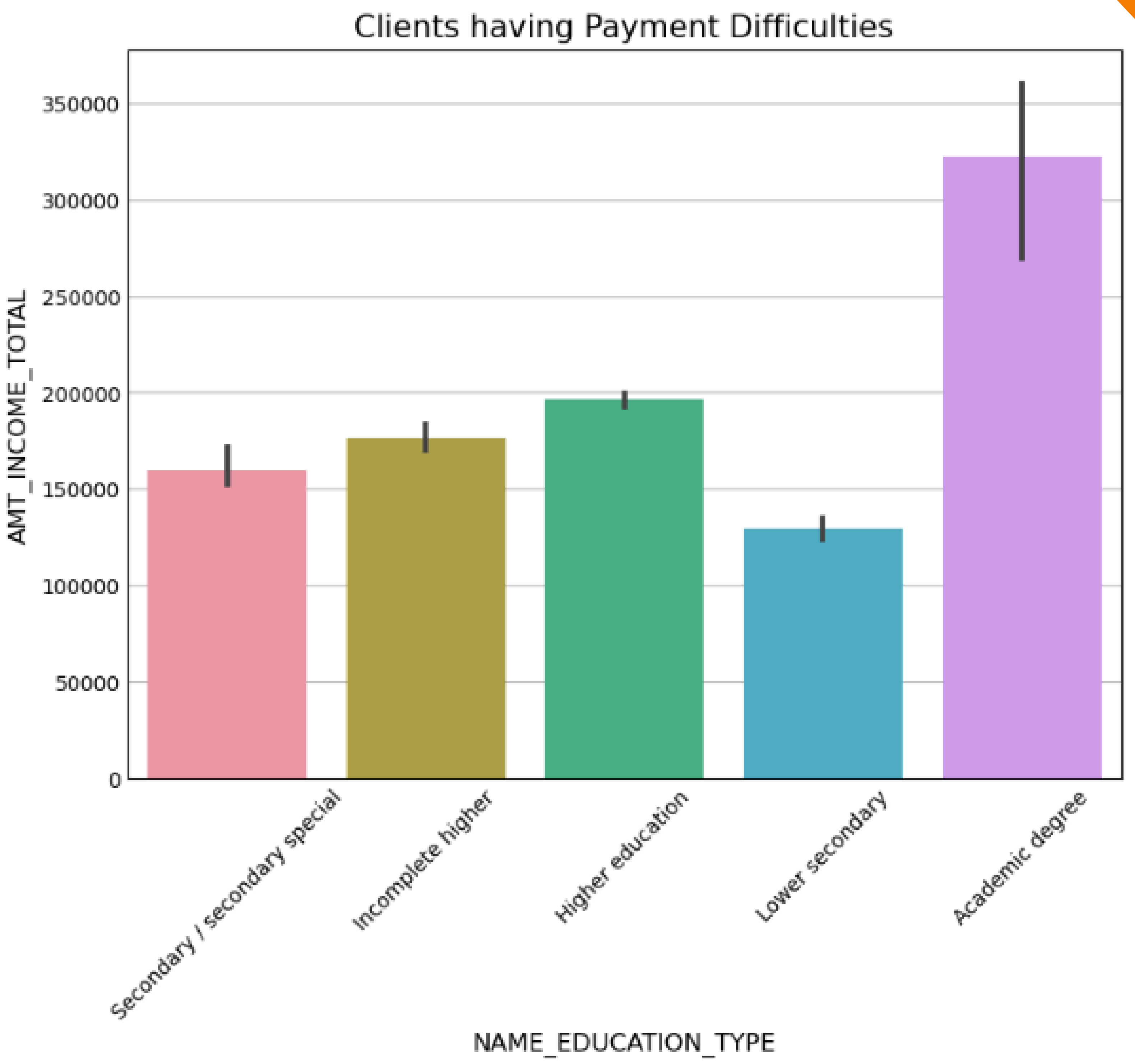




1. "AMT_ANNUITY" & "AMT_CREDIT" have strong positive correlation.
2. If the amount of credit increases then the amount of annuity also increases.

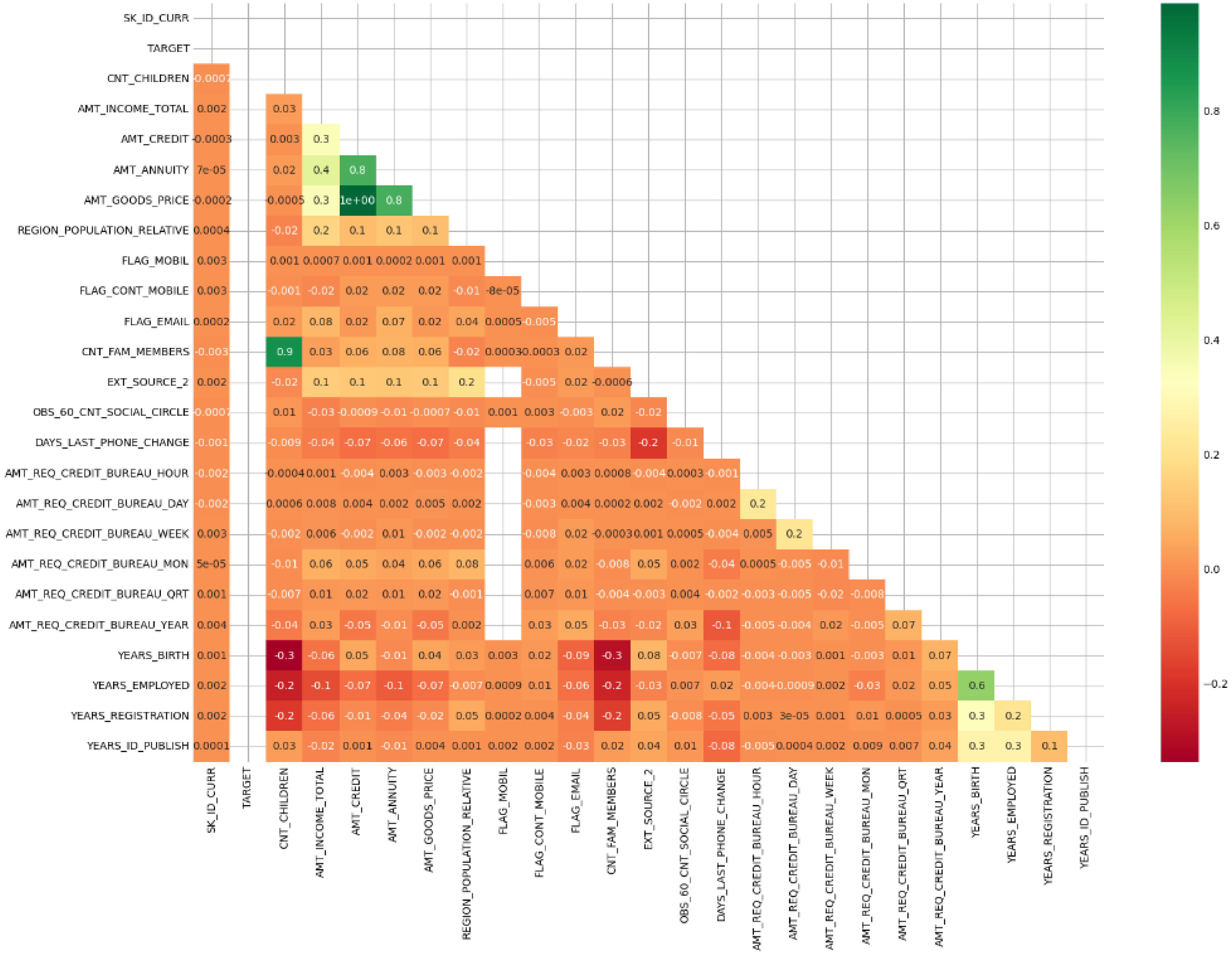
AMT_INCOME_TOTAL V/s NAME_EDUCATION_TYPE





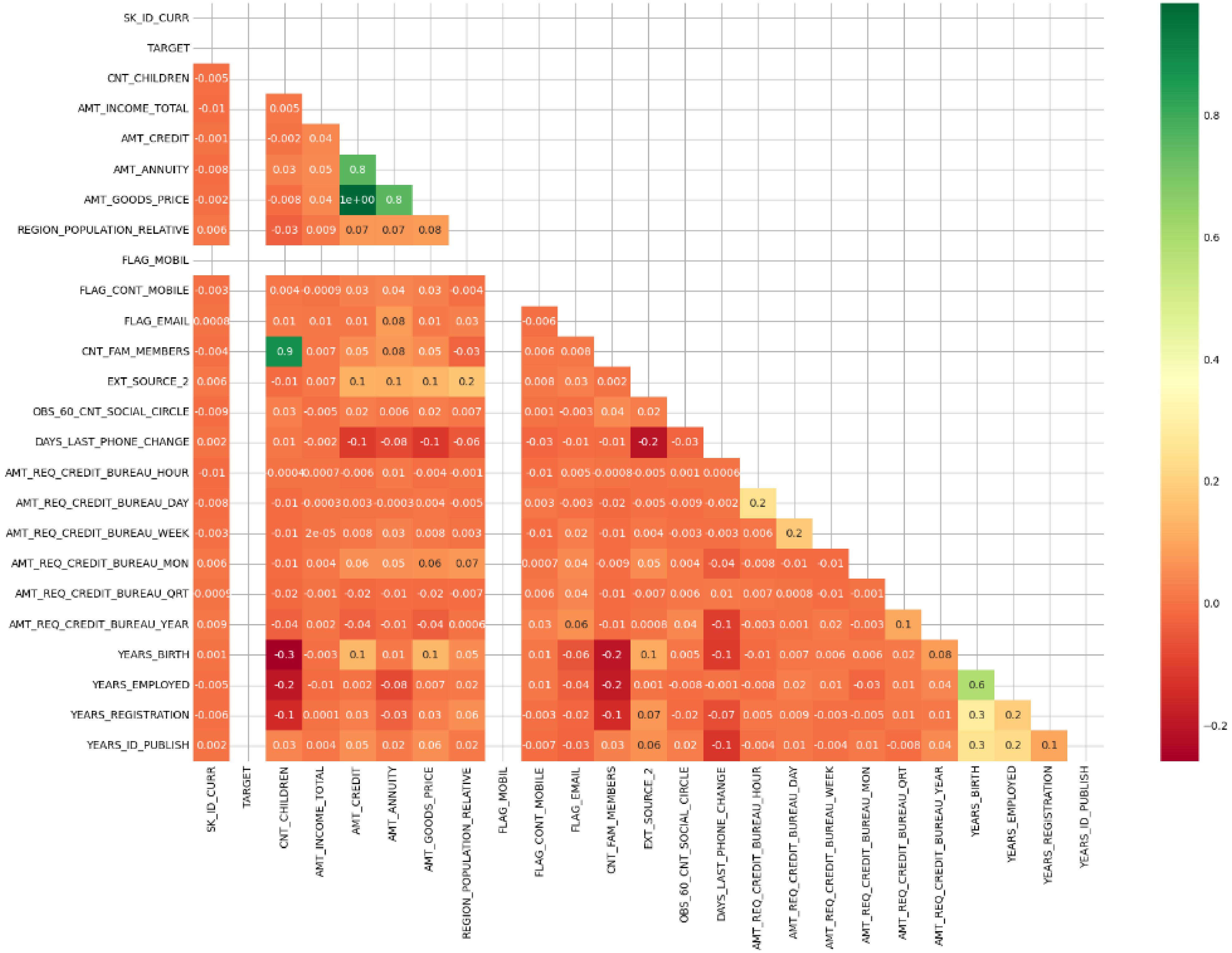
1. A client with an Academic Degree faces more payment difficulties.
2. Clients with Lower Secondary education are the least defaulted.

Analysis of df0 (Non-Defaulters)



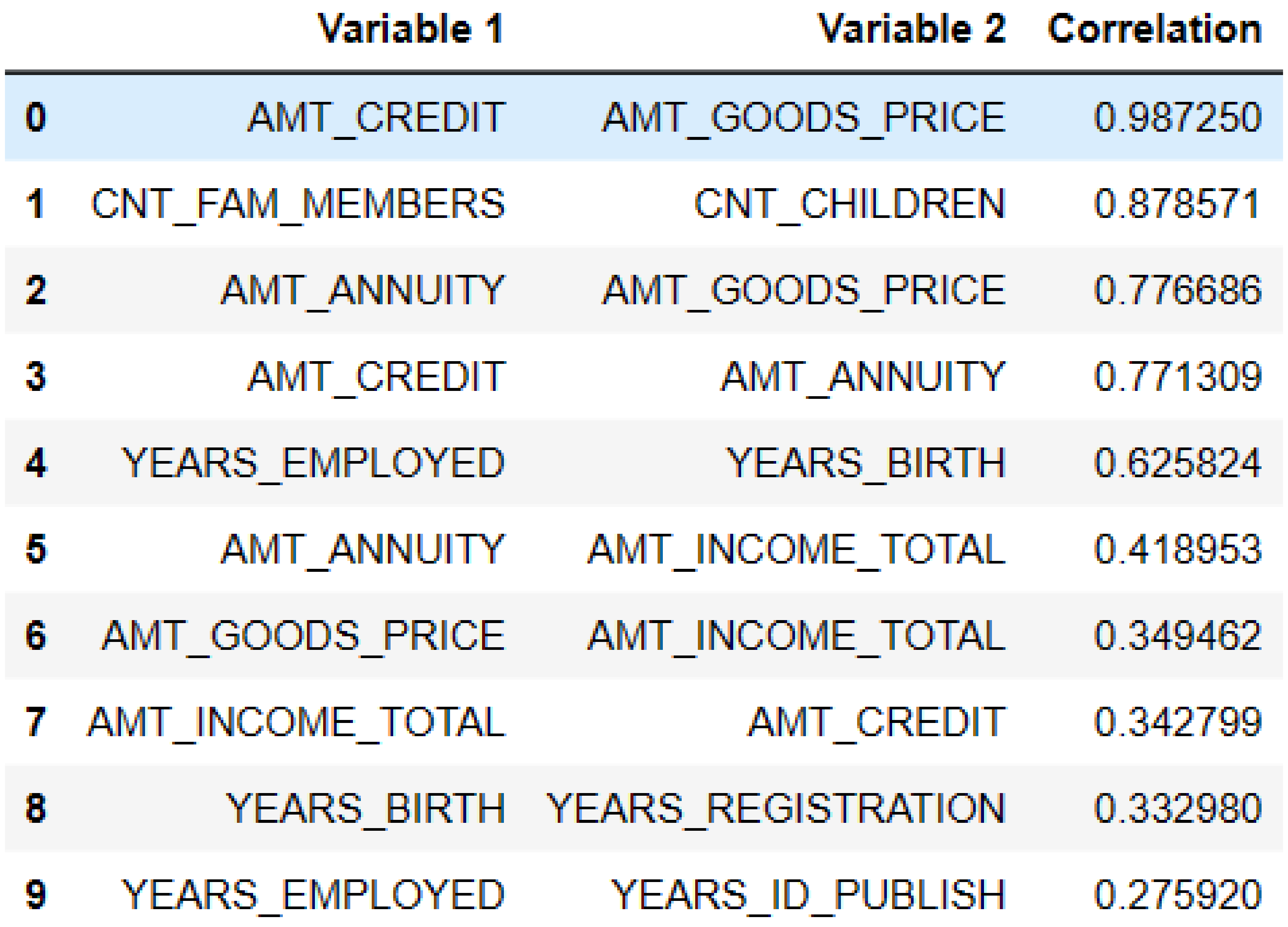
1. AMT_ANNUITY and AMT_CREDIT ; AMT_ANNUITY and AMT_GOOD_PRICE ; AMT_GOODS_PRICE and AMT_CREDIT ; CNT_FAM_MEMBERS and CNT_CHILDREN have strong positive correlation.
2. YEARS_EMPLOYED and YEARS_BIRTH have positive correlation.
3. YEARS_BIRTH and CNT_CHILDREN ; YEARS_EMPLOYED and CNT_CHILDREN ; YEARS_REGISTRATION and CNT_CHILDREN ; YEARS_BIRTH and CNT_FAMILY_MEMBERS ; YEARS_EMPLOYED and CNT_FAMILY_MEMBERS ; YEARS_REGISTRATION and CNT_FAMILY_MEMBERS have strongly negative correlation.

Analysis of df1 (Defaulters)



1. AMT_ANNUITY and AMT_CREDIT ; AMT_ANNUITY and AMT_GOOD_PRICE ; AMT_GOODS_PRICE and AMT_CREDIT ; CNT_FAM_MEMBERS and CNT_CHILDREN have strong positive correlation.
2. YEARS_EMPLOYED and YEARS_BIRTH have positive correlation.
3. YEARS_BIRTH and CNT_CHILDREN ; YEARS_EMPLOYED and CNT_CHILDREN ; YEARS_REGISTRATION and CNT_CHILDREN ; YEARS_BIRTH and CNT_FAMILY_MEMBERS ; YEARS_EMPLOYED and CNT_FAMILY_MEMBERS ; YEARS_REGISTRATION and CNT_FAMILY_MEMBERS have strongly negative correlation.

Top 10 Correlation of df0 (Non-Defaulters)

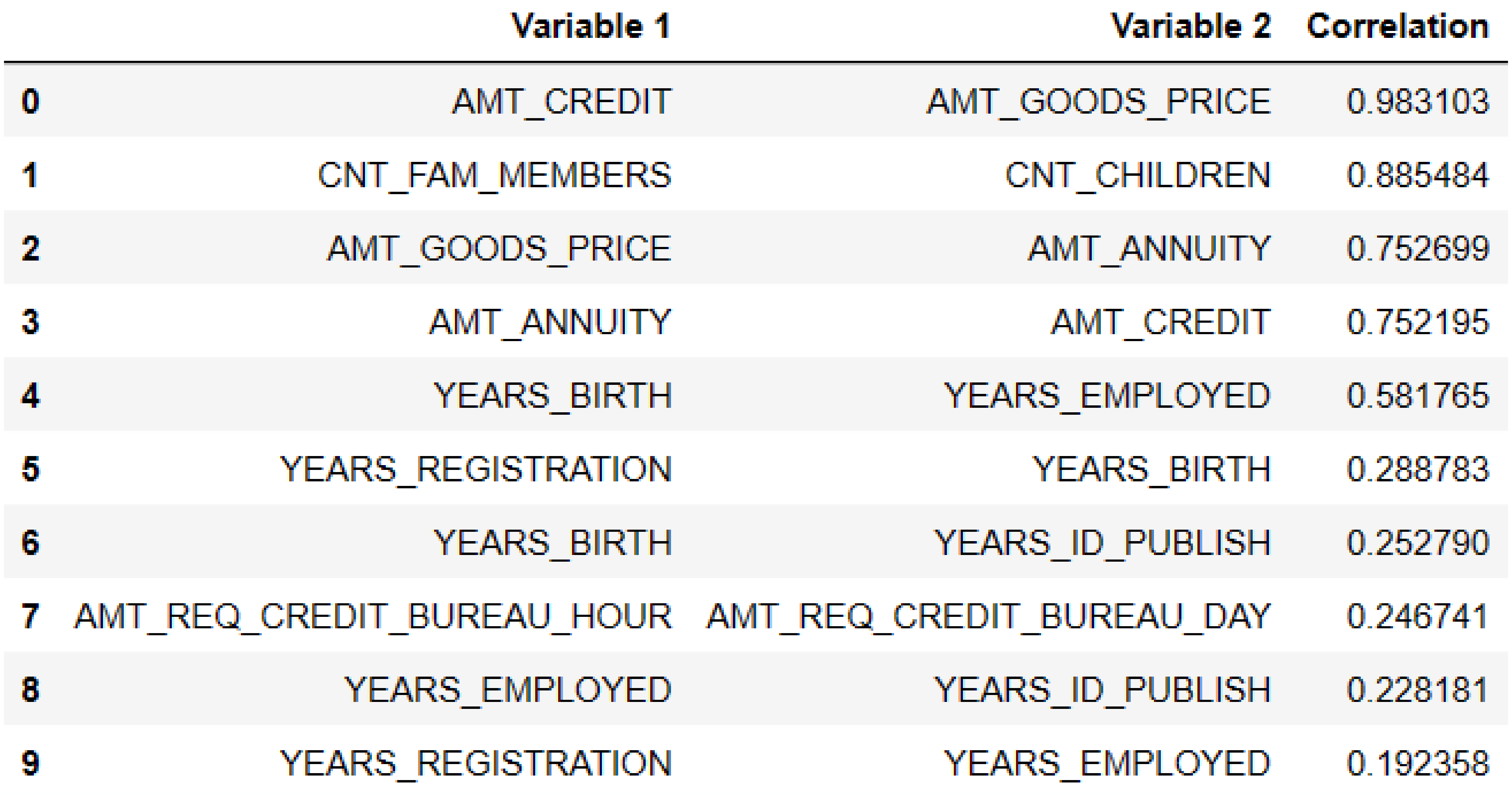


	Variable 1	Variable 2	Correlation
0	AMT_CREDIT	AMT_GOODS_PRICE	0.987250
1	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
2	AMT_ANNUITY	AMT_GOODS_PRICE	0.776686
3	AMT_CREDIT	AMT_ANNUITY	0.771309
4	YEARS_EMPLOYED	YEARS_BIRTH	0.625824
5	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953
6	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349462
7	AMT_INCOME_TOTAL	AMT_CREDIT	0.342799
8	YEARS_BIRTH	YEARS_REGISTRATION	0.332980
9	YEARS_EMPLOYED	YEARS_ID_PUBLISH	0.275920

Some Unique correlations in df0 are:

- AMT_ANNUITY and AMT_INCOME_TOTAL 0.418953
- AMT_GOODS_PRICE and AMT_INCOME_TOTAL 0.349462
- AMT_INCOME_TOTAL and AMT_CREDIT 0.342799

Top 10 Correlation of df1 (Defaulters)



	Variable 1	Variable 2	Correlation
0	AMT_CREDIT	AMT_GOODS_PRICE	0.983103
1	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
2	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699
3	AMT_ANNUITY	AMT_CREDIT	0.752195
4	YEARS_BIRTH	YEARS_EMPLOYED	0.581765
5	YEARS_REGISTRATION	YEARS_BIRTH	0.288783
6	YEARS_BIRTH	YEARS_ID_PUBLISH	0.252790
7	AMT_REQ_CREDIT_BUREAU_HOUR	AMT_REQ_CREDIT_BUREAU_DAY	0.246741
8	YEARS_EMPLOYED	YEARS_ID_PUBLISH	0.228181
9	YEARS_REGISTRATION	YEARS_EMPLOYED	0.192358

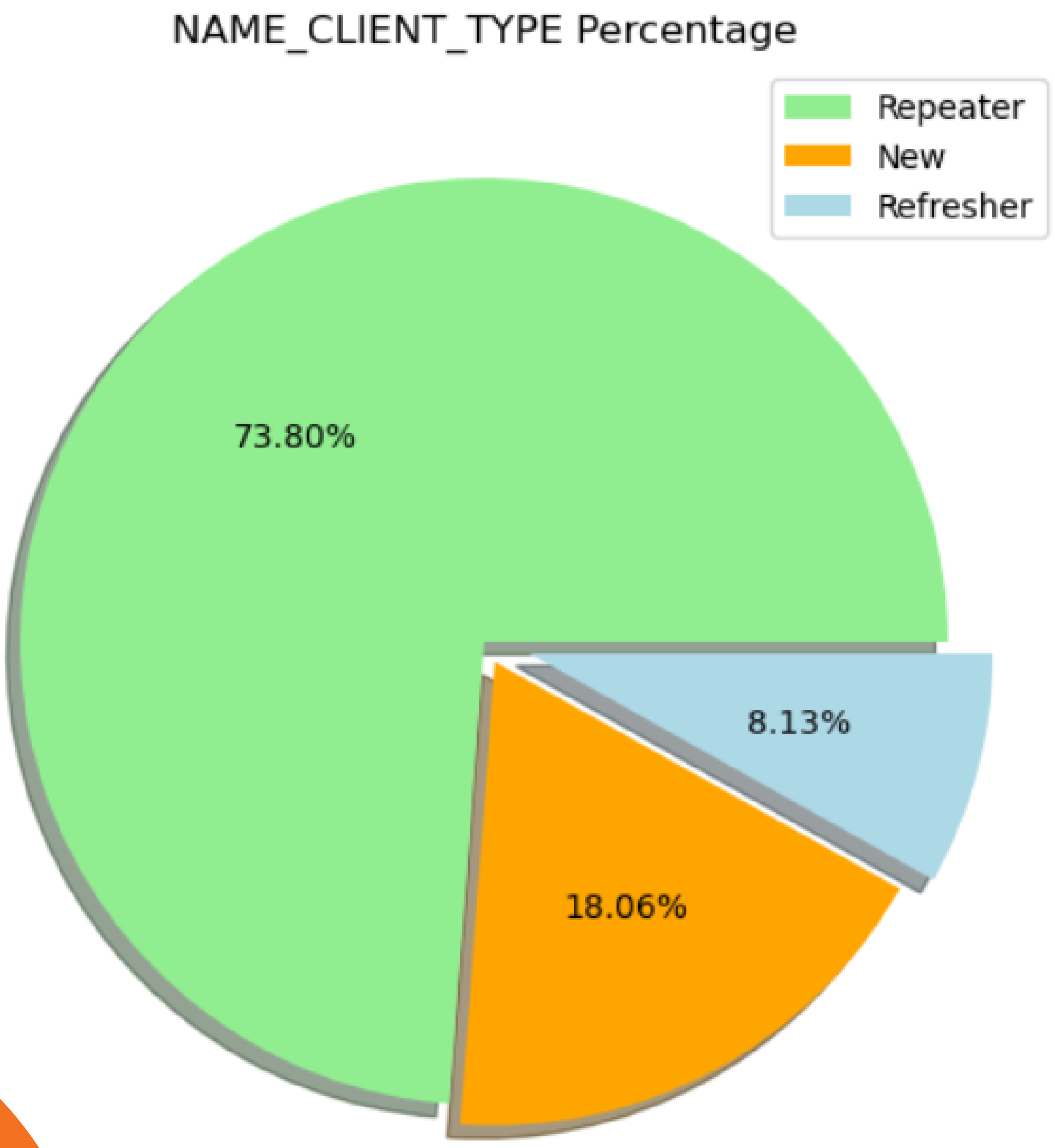
Some Unique correlations in df1 are:

- YEARS_BIRTH and YEARS_ID_PUBLISH	0.25279
- AMT_REQ_CREDIT_BUREAU_HOUR and AMT_REQ_CREDIT_BUREAU_DAY	0.246741
- YEARS_REGISTRATION and YEARS_EMPLOYED	0.192358

Previous Application Data Analysis

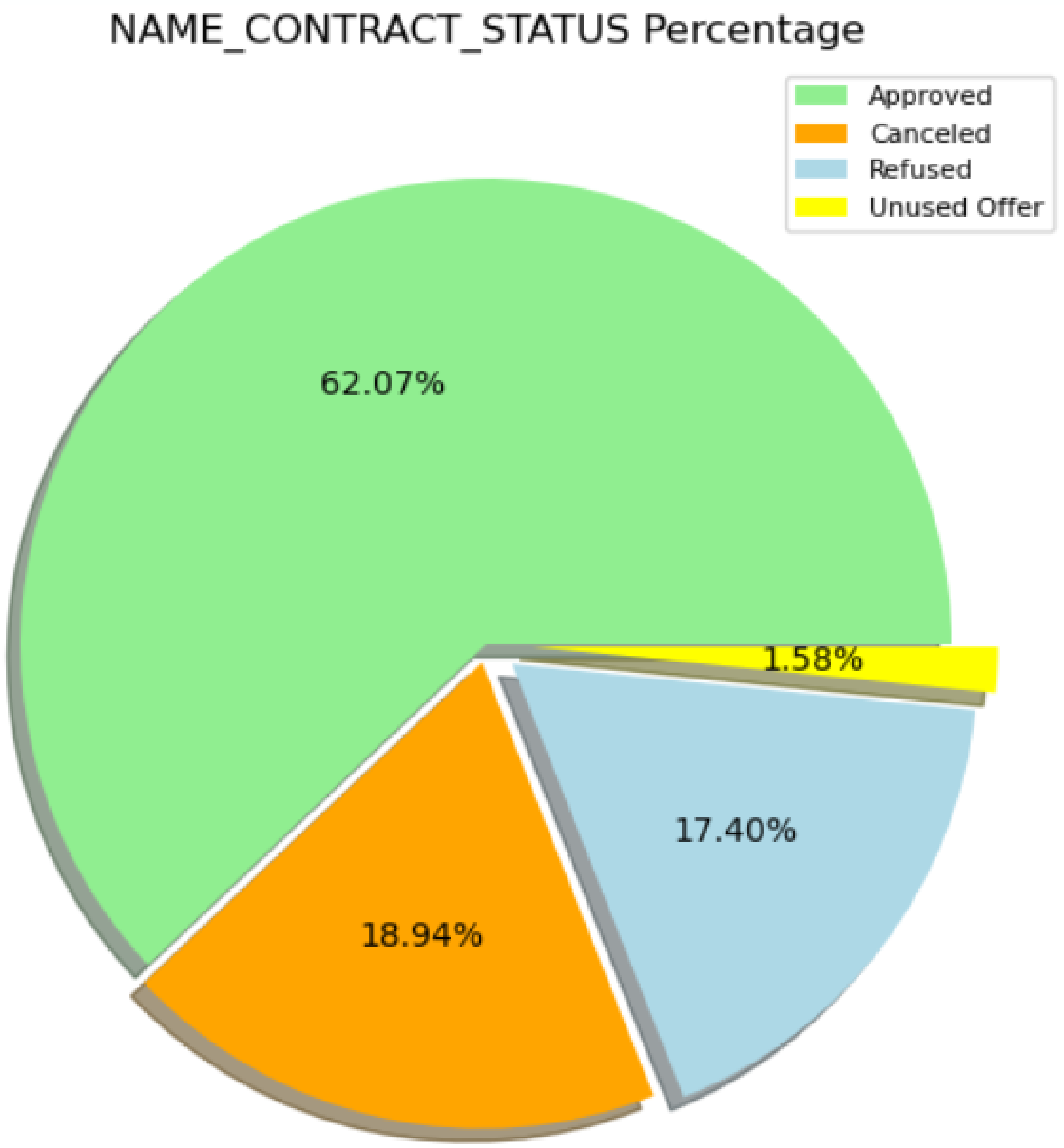
contains information about the client’s previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.

Analysis of NAME_CLIENT_TYPE



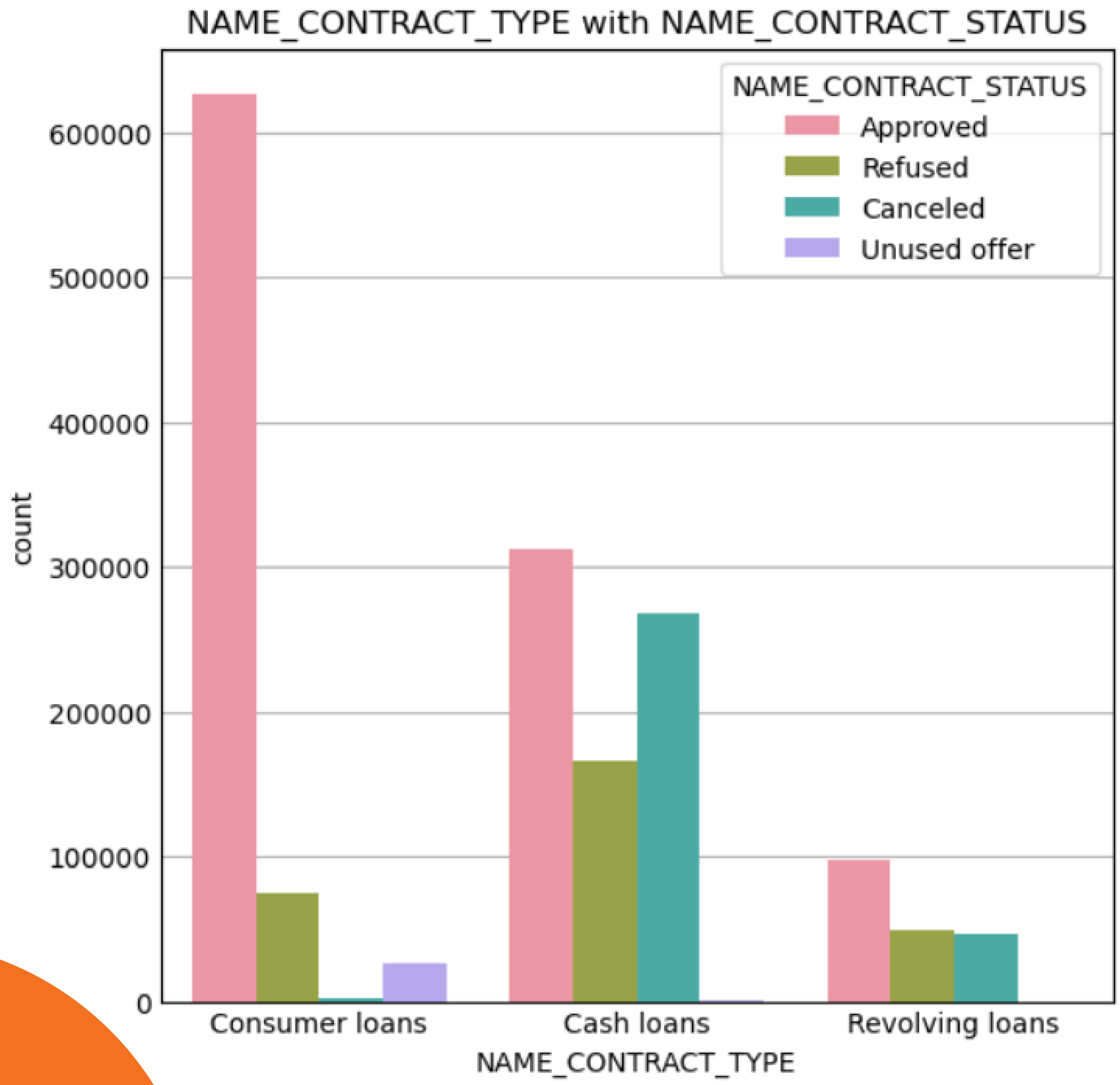
1. The "Repeater" client type is the highest among all loan applications with 73.80%
2. "New" client type has 18.06%

Analysis of NAME_CONTRACT_STATUS



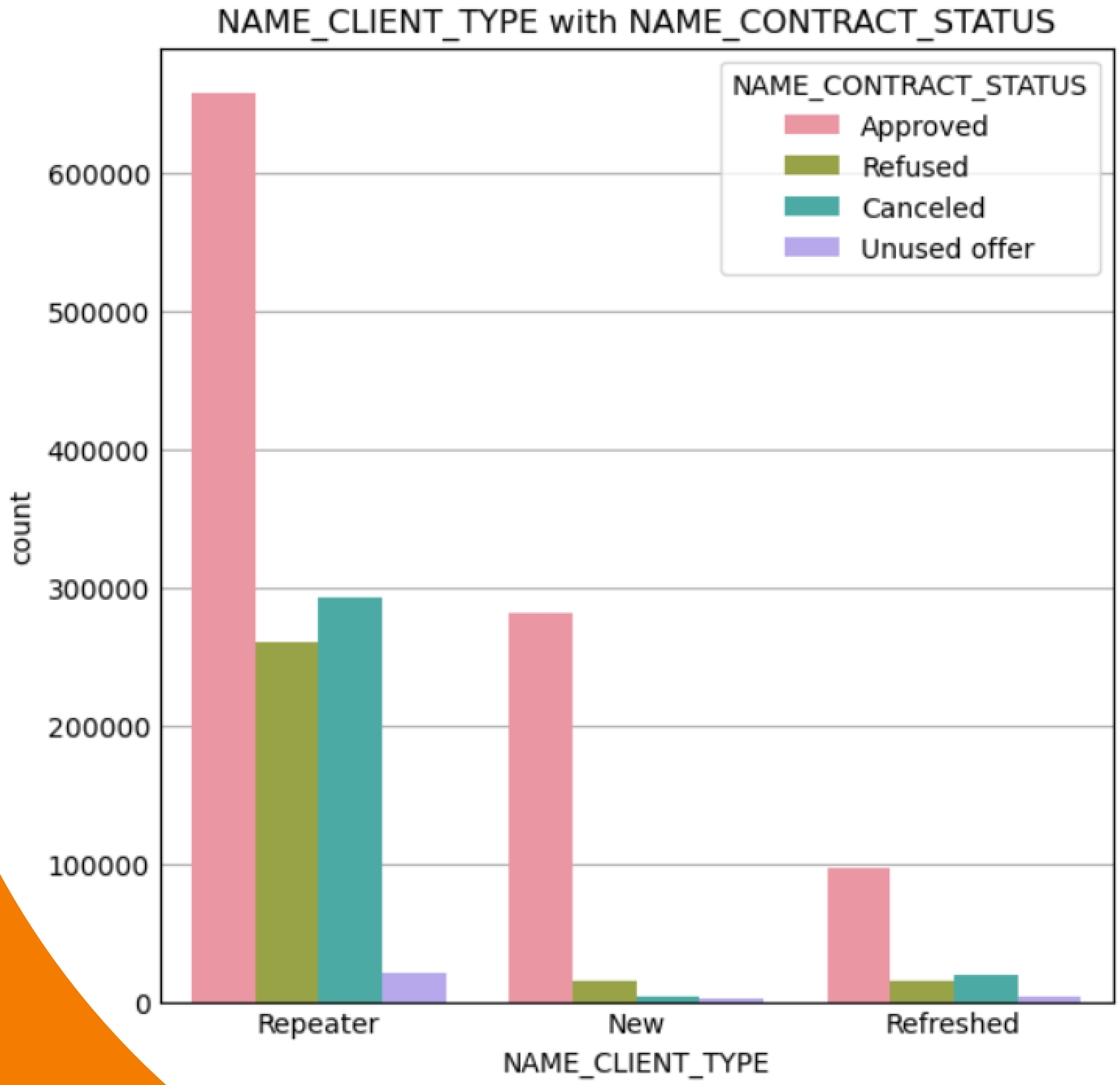
1. "Approved" loan status is the highest among all loan applications with 62.07% followed by "Canceled" with 18.94%
2. 1.58% of clients have unused offers.

NAME_CONTRACT_TYPE with NAME_CONTRACT_STATUS



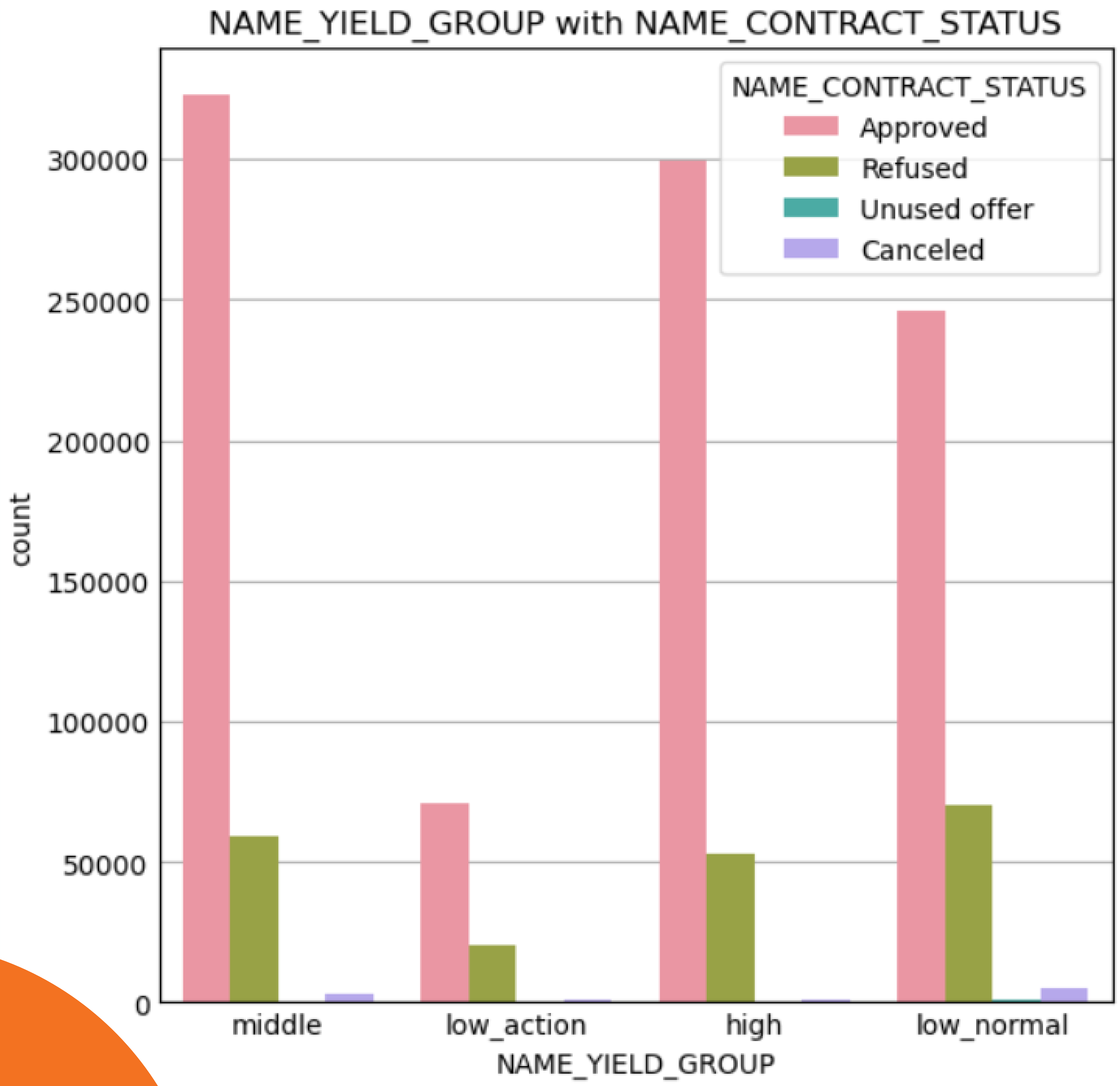
1. The "Consumer Loans" Category has the highest number of applicants with approval.
2. The "Cash Loans" category has more Refused loans.
3. Consumer Loans are the least cancelled.
4. There is no cancelled loan in the Cash Loan category.

NAME_CLIENT_TYPE with NAME_CONTRACT_STATUS



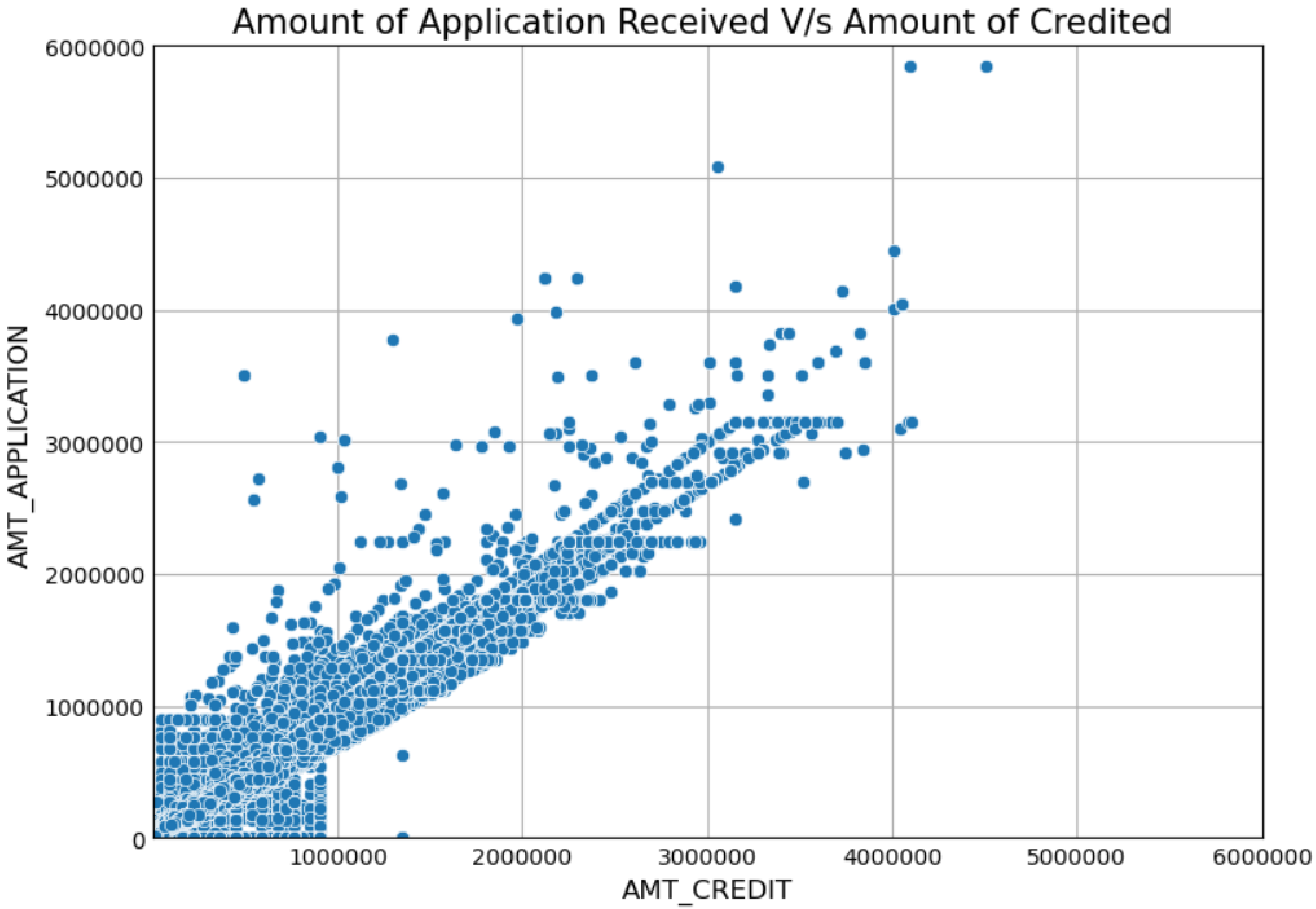
1. The "Repeater" Client type has the highest number of all types of contract statuses including Approved loans.

NAME_YIELD_GROUP with NAME_CONTRACT_STATUS



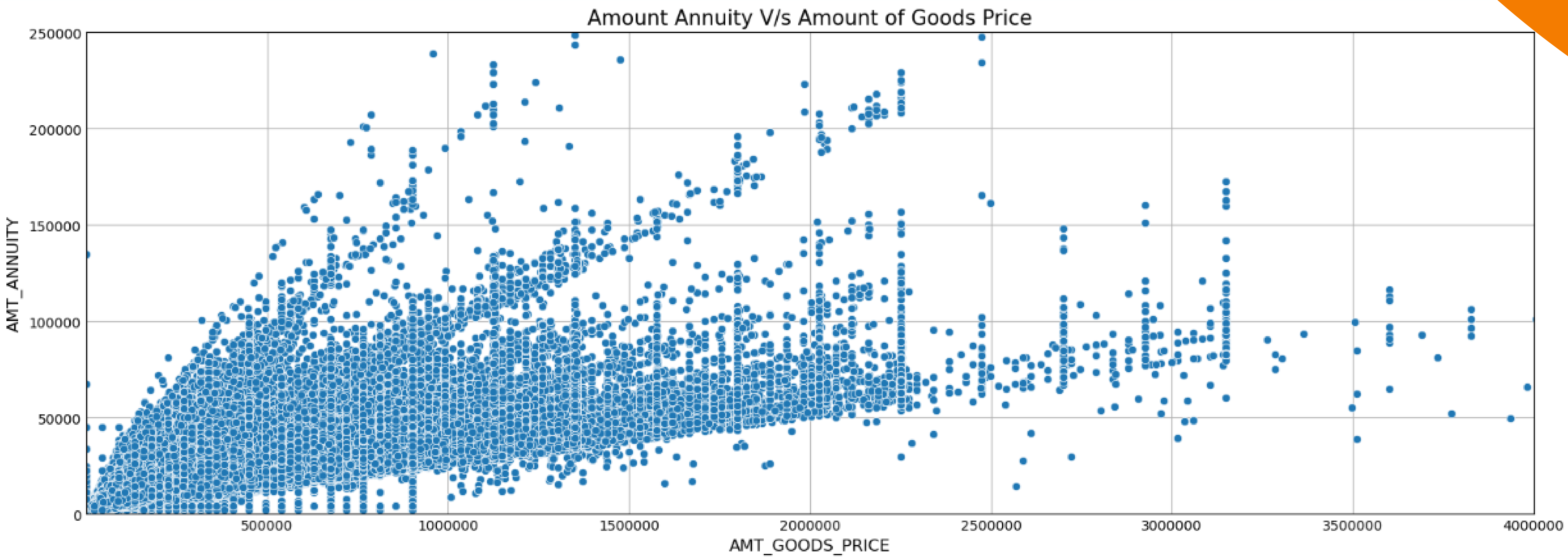
1. Middle and High-interest rates loans are approved more.
2. Low Normal interest rate loan Refused and cancelled the most.

Amount of Application Received V/s Amount of Credit



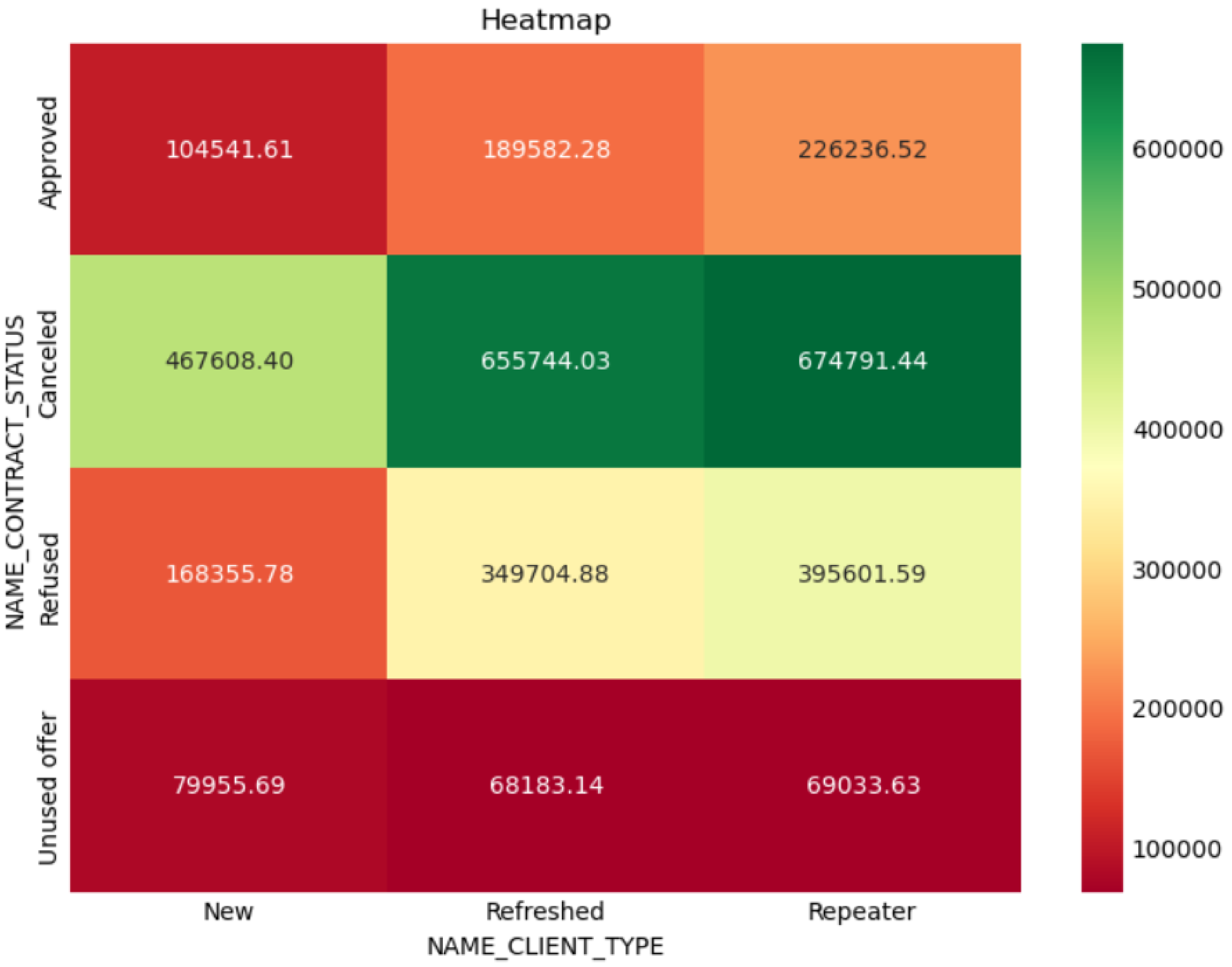
1. "AMT_CREDIT" has a strong positive correlation with "AMT_APPLICATION"

Amount Annuity V/s Amount of Goods Price



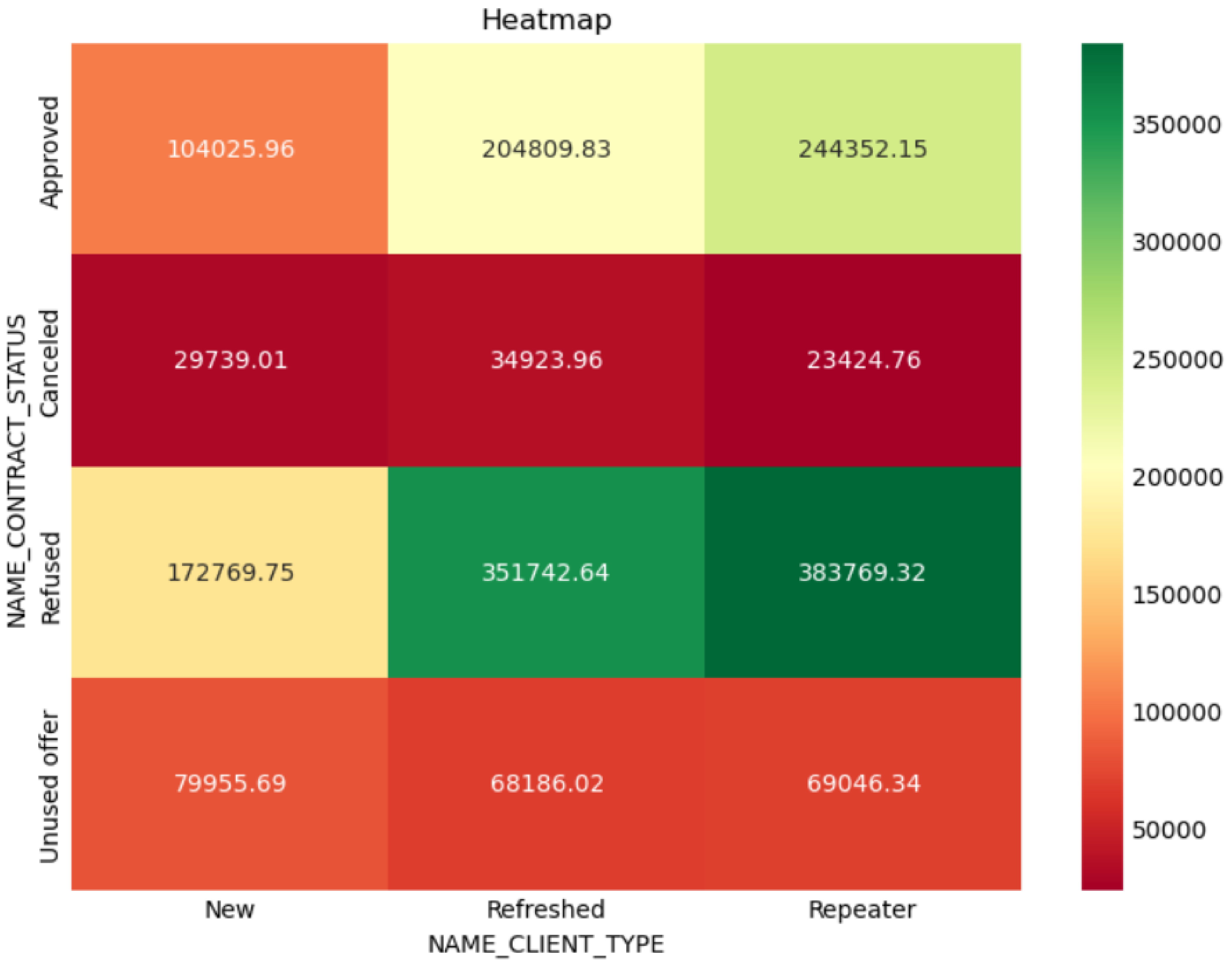
1. "AMT_ANNUITY" has strong positive correlation with "AMT_GOODS_PRICE"

NAME_CONTRACT_STATUS with NAME_CLIENT_TYPE along with AMT_GOODS_PRICE



1. Canceled contract status has a higher value of goods.

NAME_CONTRACT_STATUS with NAME_CLIENT_TYPE along with AMT_CREDIT

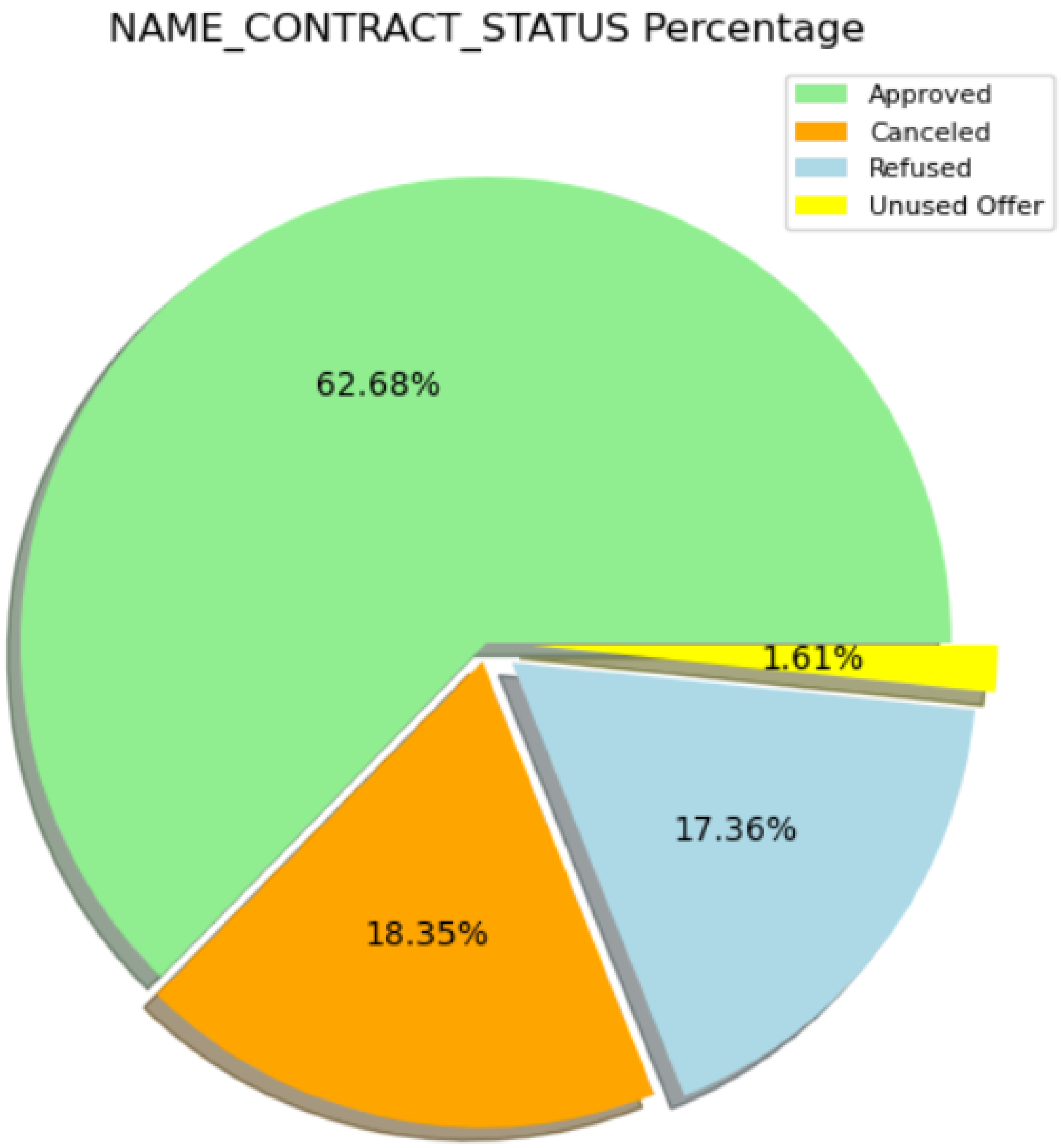


1. Unused offer contract status has a low amount of credit.

Merged Application Data Analysis

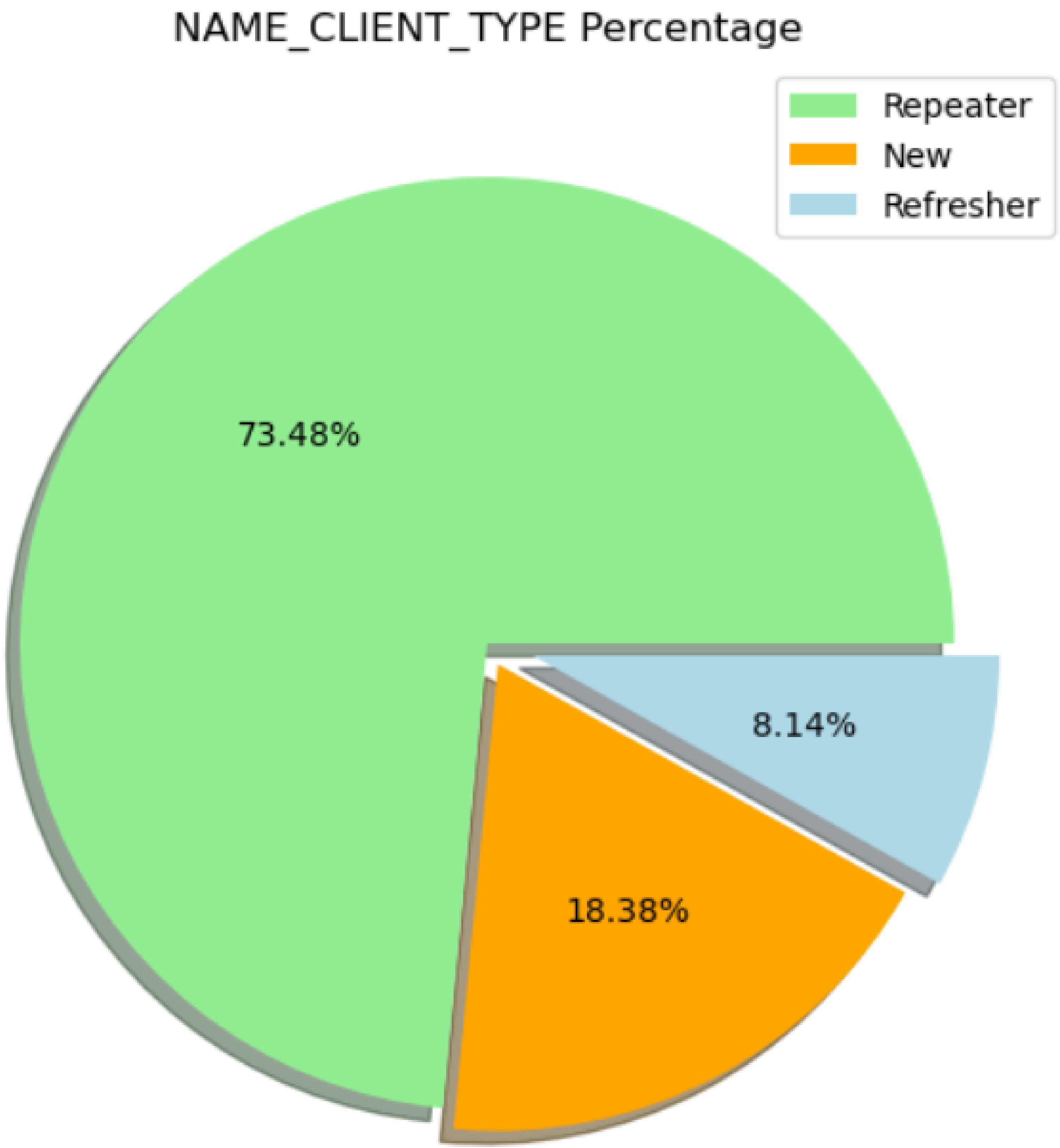
contains information about the client’s records. Current application data merged with previous application data.

Analysis of NAME_CONTRACT_STATUS



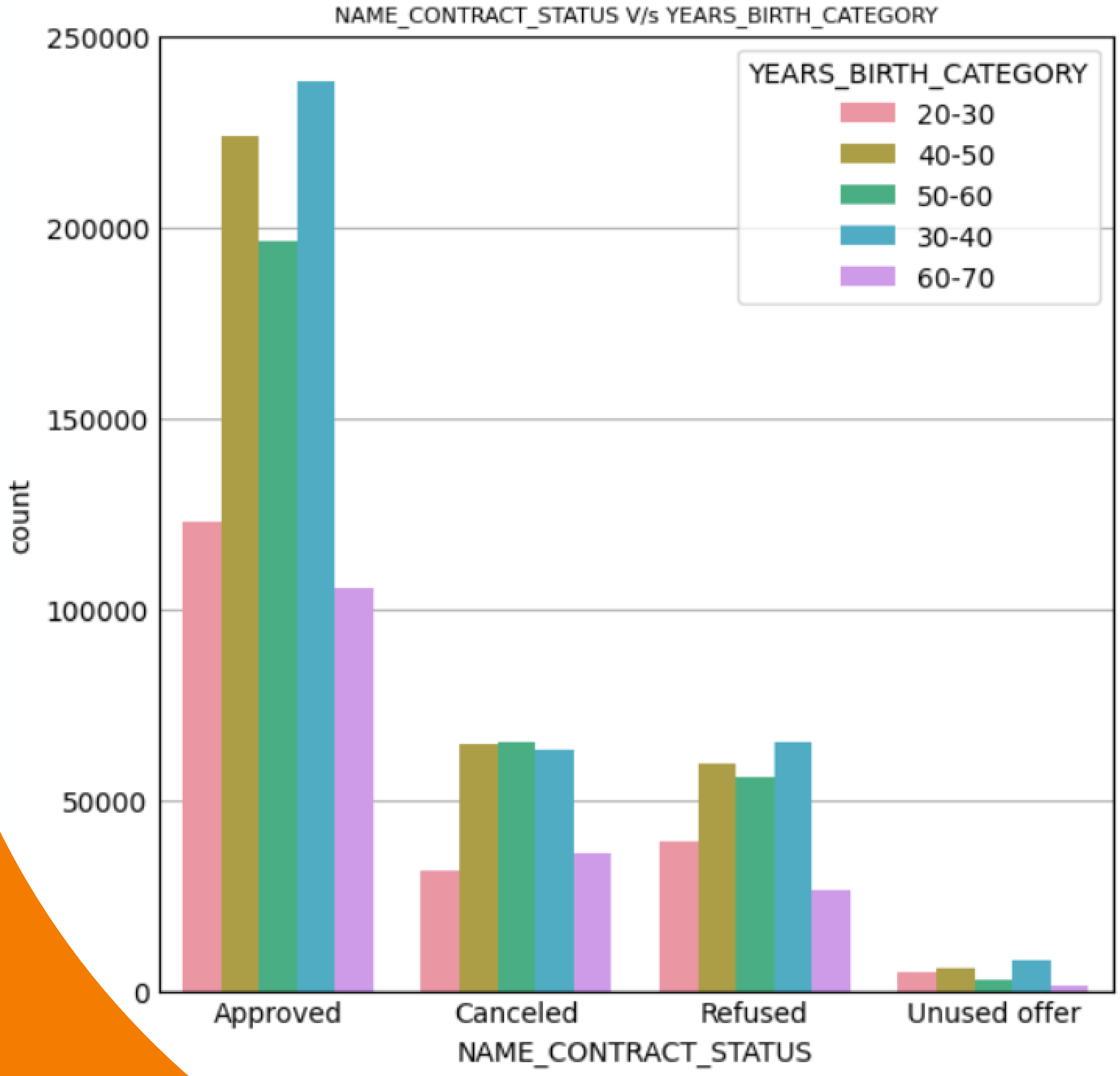
1. Approved status is the highest among all categories followed by Cancelled.

Analysis of NAME_CLIENT_TYPE



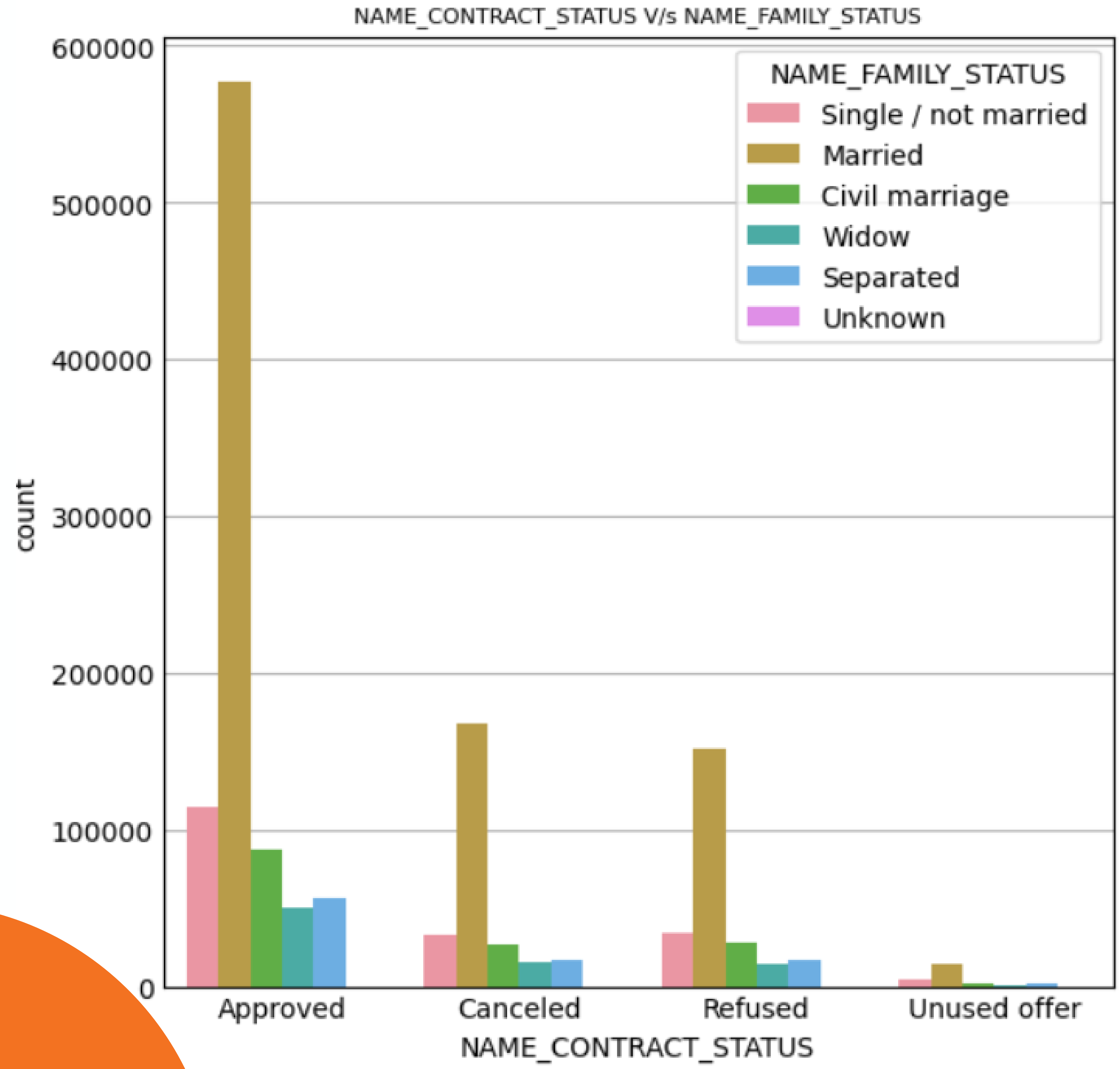
1. Repeater clients are the most with 73.48%
2. New Clients are 18.38%

YEARS_BIRTH_CATEGORY with NAME_CONTRACT_STATUS



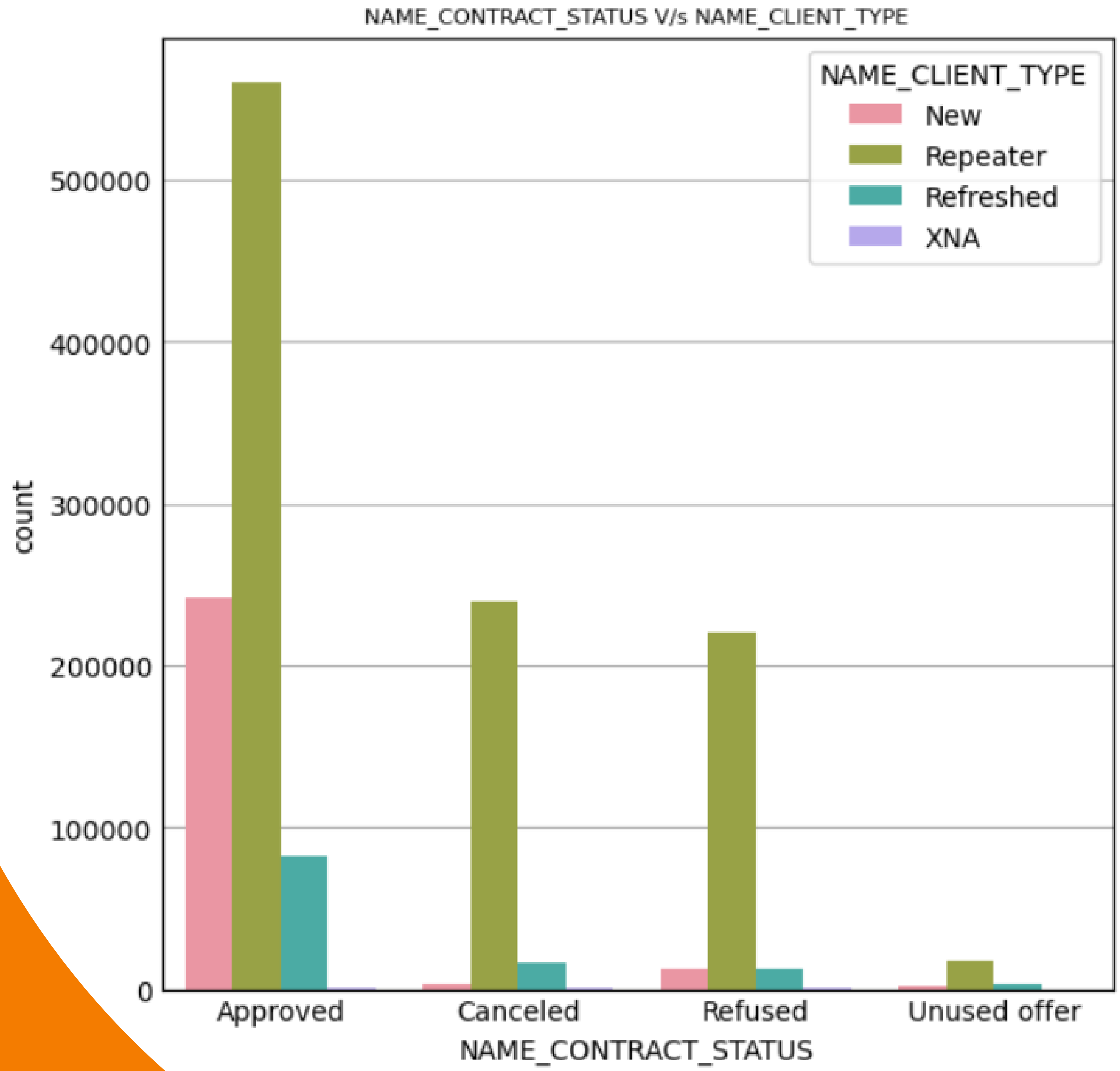
1. Clients in the age range 30-40 get the most approval followed by clients in the 40-50 age range.
2. Clients of the age range 60-70 receive the least refusals followed by the 20-30 age range

NAME_FAMILY_STATUS with NAME_CONTRACT_STATUS



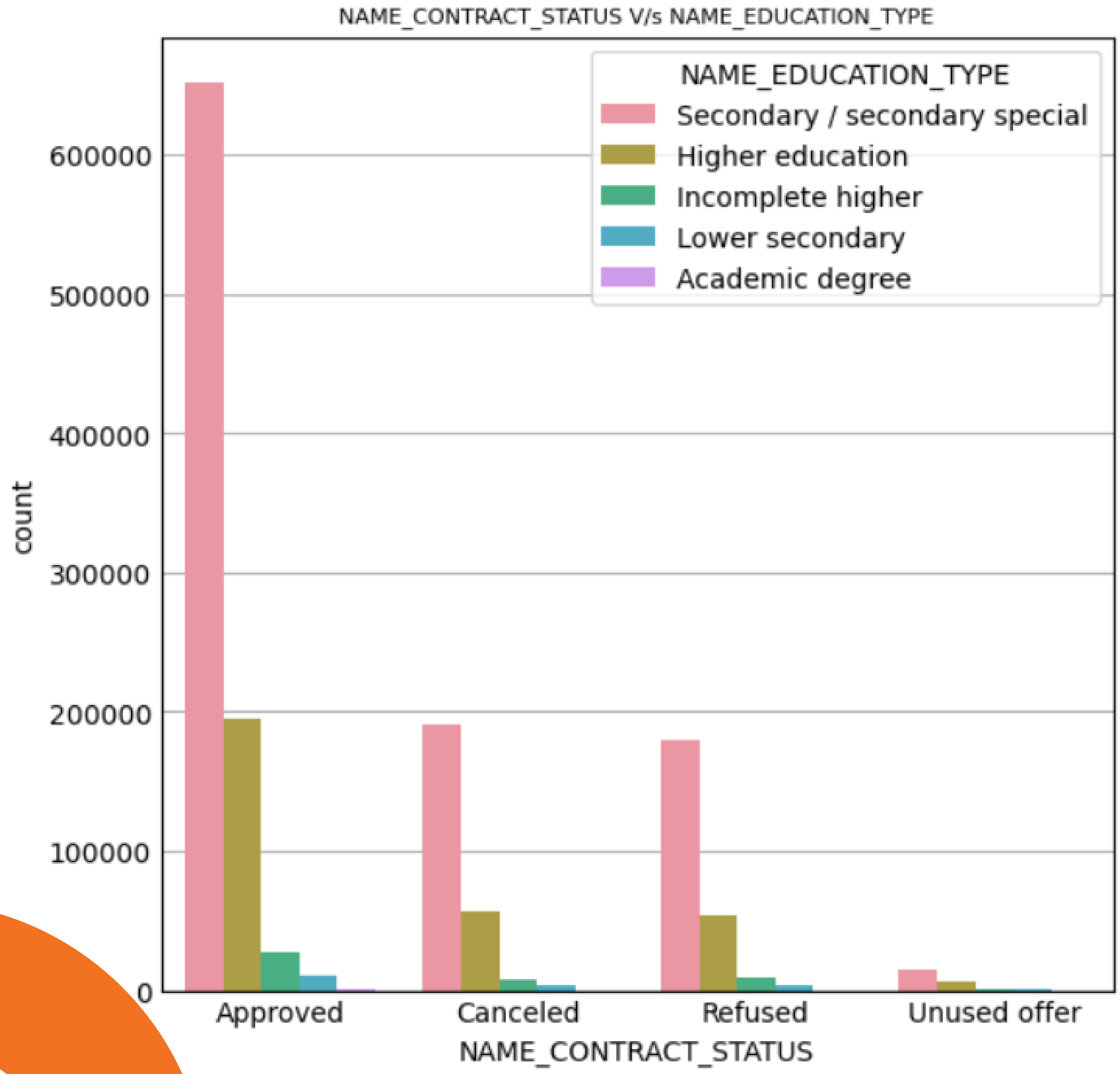
1. Married Client receives the most approval followed by Single/not married.

NAME_CLIENT_TYPE with NAME_CONTRACT_STATUS



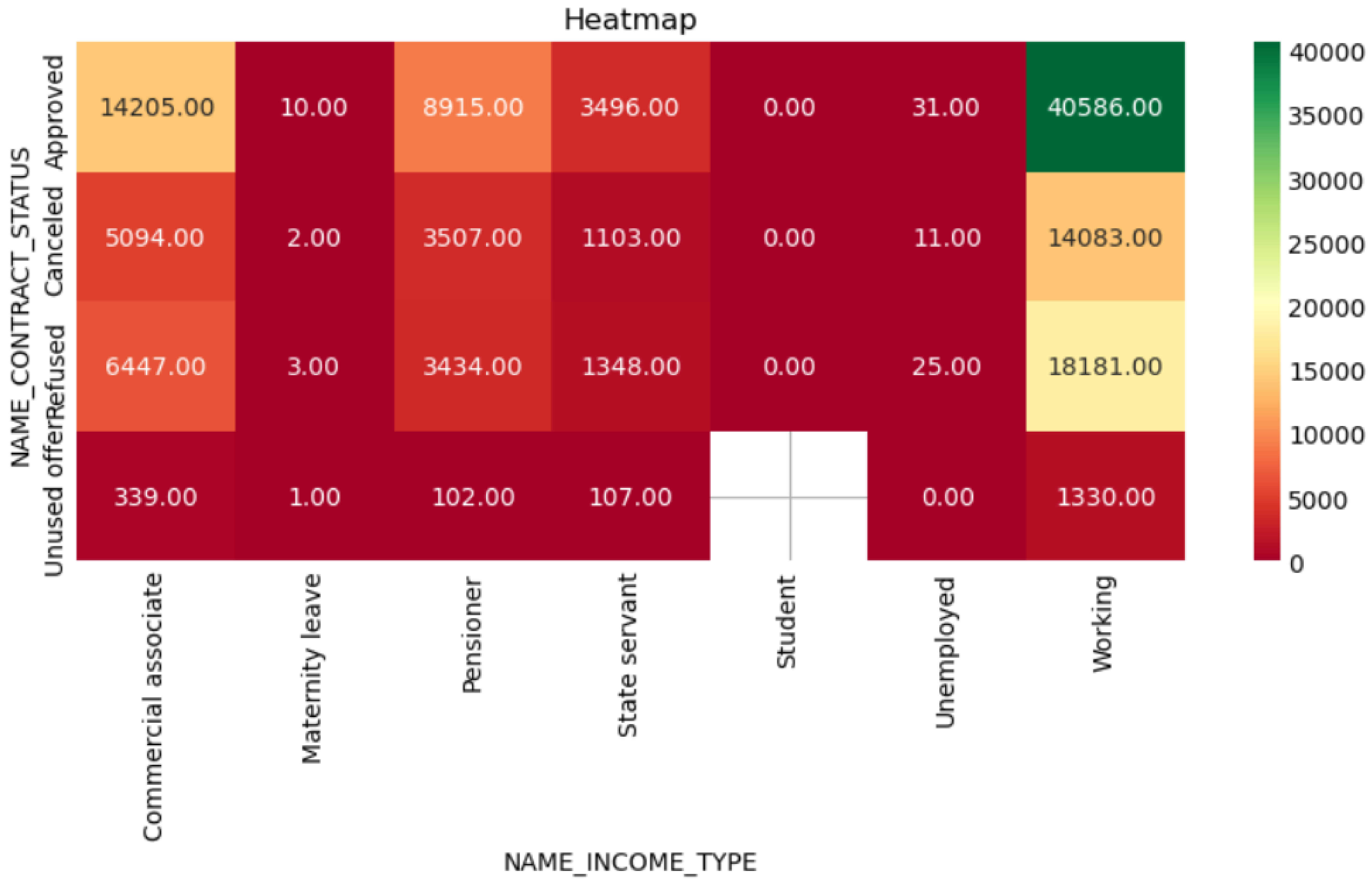
1. Repeaters clients have the highest approval rate followed by New clients.

NAME_EDUCATION_TYPE with NAME_CONTRACT_STATUS



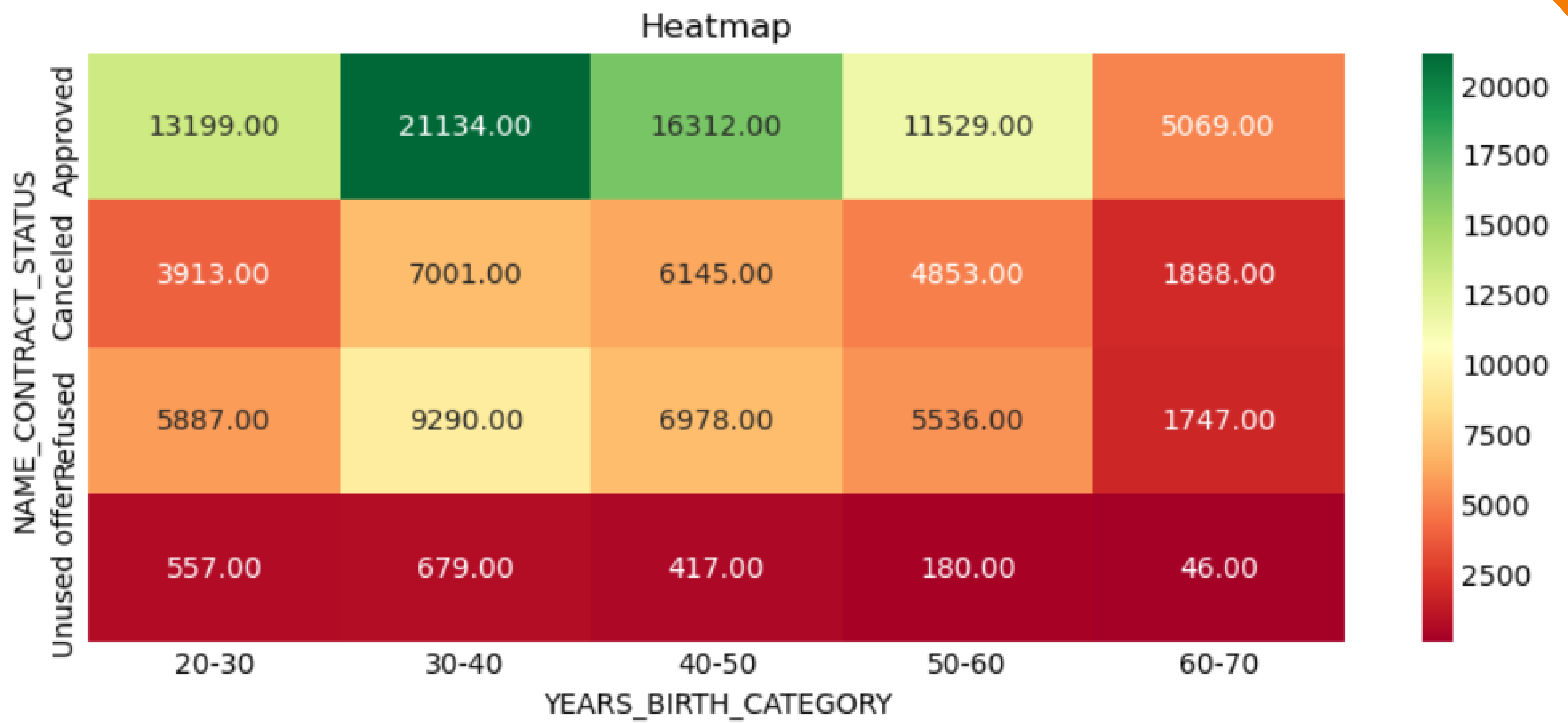
1. Clients with Secondary/ Secondary special education type has more approved loans followed by higher education.

"NAME_CONTRACT_STATUS", "NAME_INCOME_TYPE", aggregating on TARGET



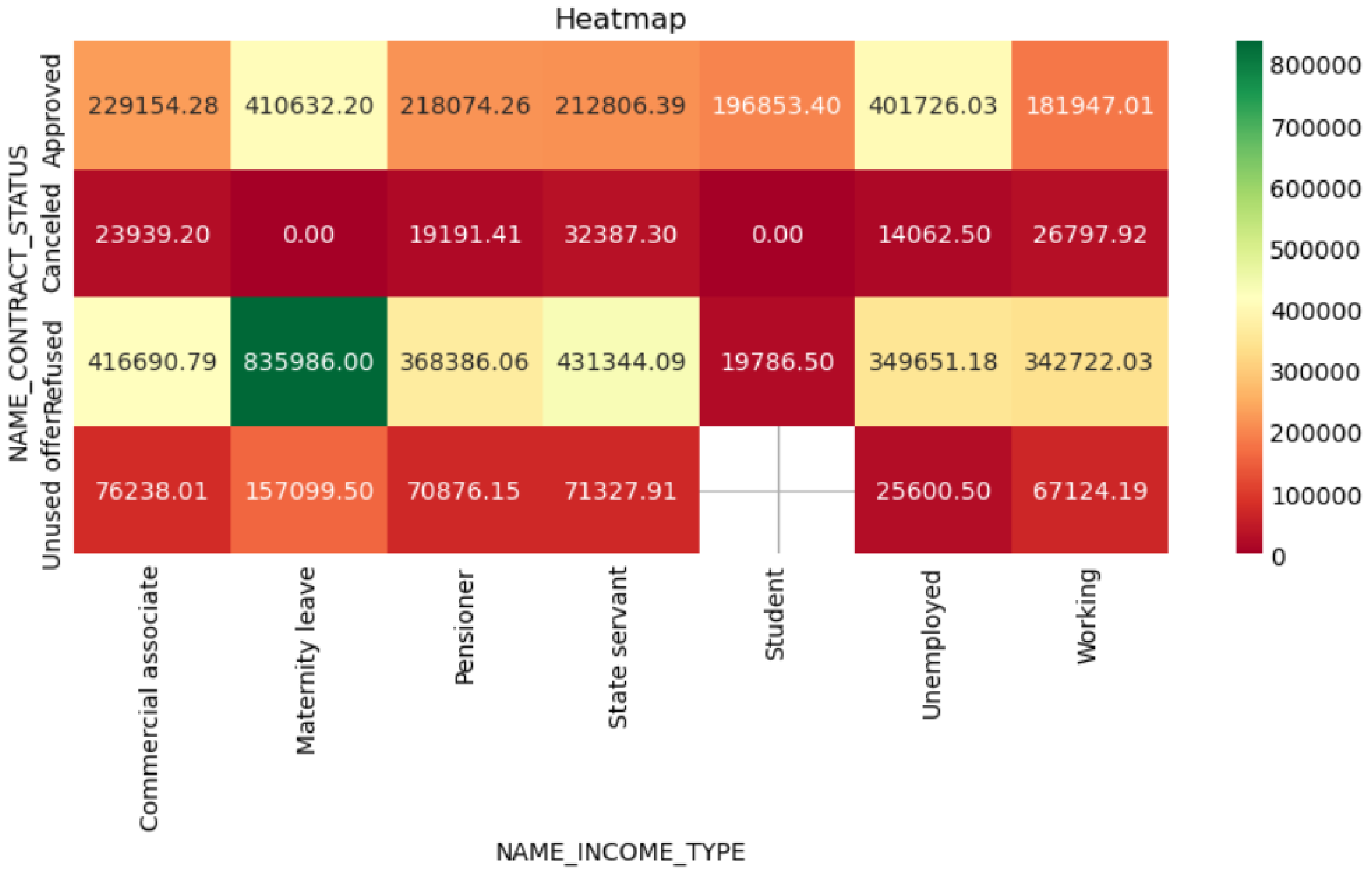
1. Working applicants with Approved status have defaulted in the highest numbers.

"NAME_CONTRACT_STATUS", "YEARS_BIRTH_CATEGORY", aggregating on TARGET



1. Age group 30-40 and 40-50 clients with approved loans defaulted the most.

"NAME_CONTRACT_STATUS", "NAME_INCOME_TYPE",aggregating on AMT_CREDIT



NAME_CONTRACT_STATUS	Commercial associate	Maternity leave	Pensioner	State servant	Student	Unemployed	Working
Approved	229154.28	410632.20	218074.26	212806.39	196853.40	401726.03	181947.01
Canceled	23939.20	0.00	19191.41	32387.30	0.00	14062.50	26797.92
Refused	416690.79	835986.00	368386.06	431344.09	19786.50	349651.18	342722.03
Unused offer	76238.01	157099.50	70876.15	71327.91		25600.50	67124.19

1. Smaller credits are provided to unused offer contract status.

Conclusion

Approved applications having default cases:

The variables/columns mentioned below are for **defaulters**:-

1. In the column **"NAME_INCOME_TYPE"** those who are working are the defaulters. It doesn't mean working clients must be refused. Proper checks of other parameters for them are needed.
2. In the column "**OCCUPATION_TYPE**", **Labourers with 31%**; **Sales Staff with 17%** and **Drivers with 11%** are the defaulters.
3. In the column "**ORGANIZATION_TYPE**", **29%** of clients are defaulters who have "**Business Entity Type 3**".
4. In the column "**OWN_REALTY_FLAG**", **70%** of clients **don't have their own houses**.
5. In the column "**OWN_CAR_FLAG**", **31%** of clients **don't have cars**.
6. In the column "**NAME_YIELD_GROUP**", the **middle** ones are the most defaulters.
7. **Previous application data** also have default cases with application statuses "**Refused, Unused and Cancelled**".

The Company should focus more on the below listed variable/columns to increase its revenue:-

1. Clients with age groups 30-40 & 40-50.
2. Clients with Married Family Status.
3. Repeater clients are more trustworthy.
4. Clients have income types of Businessmen & students.

THANK YOU

Harsh Deep Jaggi
DSC-57 Batch