Presented By –
Rishav B
Rajat R

CASE STUDY

CREDIT RISK ANALYSIS

Problem Statement

Credit company has customers who apply for a loan

Customers might also have an ongoing loan with the Company

Customers who have difficulty in repaying their loan are flagged as defaulters

Goal is to analyse what are the factors that is leading to default in payment

Business Advantage

Identifying applicants who are likely to repay the loan, hence reducing loss of business to the company.

Reducing the financial loss of the company by not approving the loans for the defaulters.

Approach

Data Overview

1. Import the data
2. Check number of rows and columns
3. Select particular columns to perform analysis
4. Get complete data description (mean, standard deviation etc.)

Analysis

1. Find missing data
2. Suggesting ways of handling the missing data by imputing, deleting or keeping the rows with missing values.
3. Data Manipulation and creating different metrics for analysis.
4. Plotting visualizations and generating insights from them.

Conclusion

1. Conclusions generated from the analysis.

Data Description

Application data

- Data length = 307511, Number of columns = 100
- Number of columns selected for analysis = 25

Previous application data

- Data length = 1670213, Number of columns = 37
- Number of columns selected for analysis = 37

Flow of the Presentation

Application Data Analysis → Previous Data Analysis → Combined Data Analysis

Data Analysis

APPLICATION DATA

Data Overview

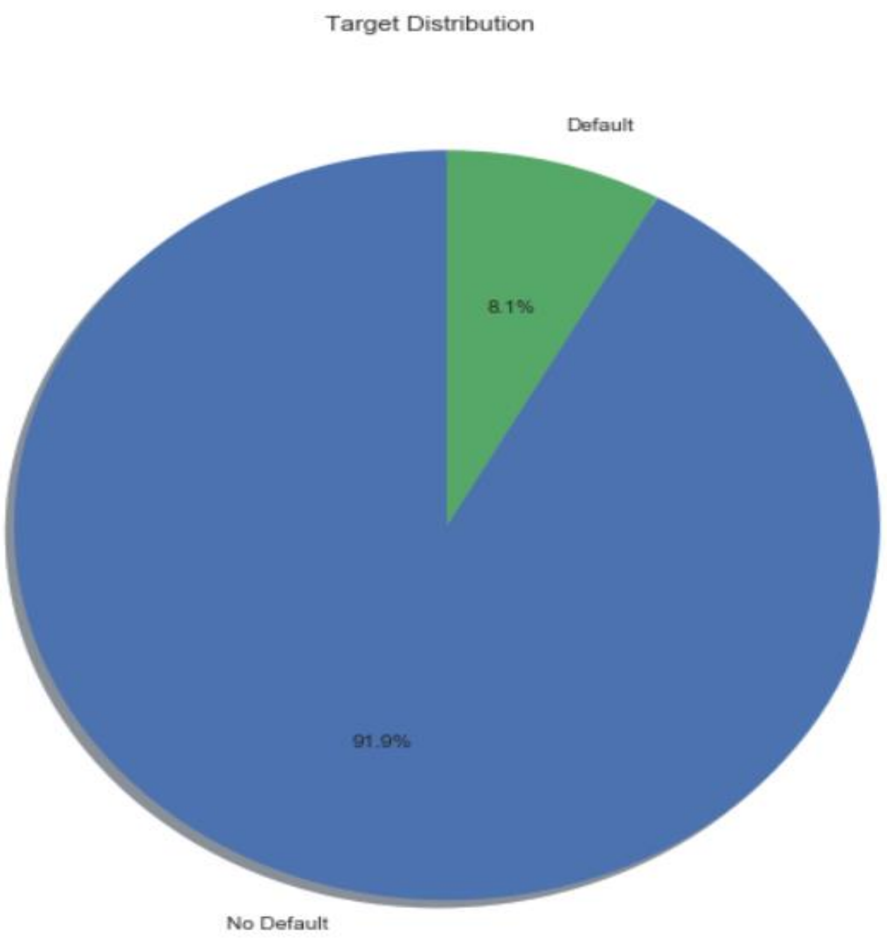
	SK_ID_CURR	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS
count	307511	307511	307511	307499	307233	307511	307511	307511	307511	307511	307509
mean	278180.5	168797.9	599026	27108.57	538396.2	0.020868	-16037	63815.05	-4986.12	-2994.2	2.152665
std	102790.2	237123.1	402490.8	14493.74	369446.5	0.013831	4363.989	141275.8	3522.886	1509.45	0.910682
min	100002	25650	45000	1615.5	40500	0.00029	-25229	-17912	-24672	-7197	1
25%	189145.5	112500	270000	16524	238500	0.010006	-19682	-2760	-7479.5	-4299	2
50%	278202	147150	513531	24903	450000	0.01885	-15750	-1213	-4504	-3254	2
75%	367142.5	202500	808650	34596	679500	0.028663	-12413	-289	-2010	-1720	3
max	456255	1.17E+08	4050000	258025.5	4050000	0.072508	-7489	365243	0	0	20

Missing Data for Current Application

Columns	% Missing	Type
OCCUPATION_TYPE	31.34	object
AMT_GOODS_PRICE	0.90	float64
AMT_ANNUITY	0.003	float64
CNT_FAM_MEMBERS	0.00065	float64

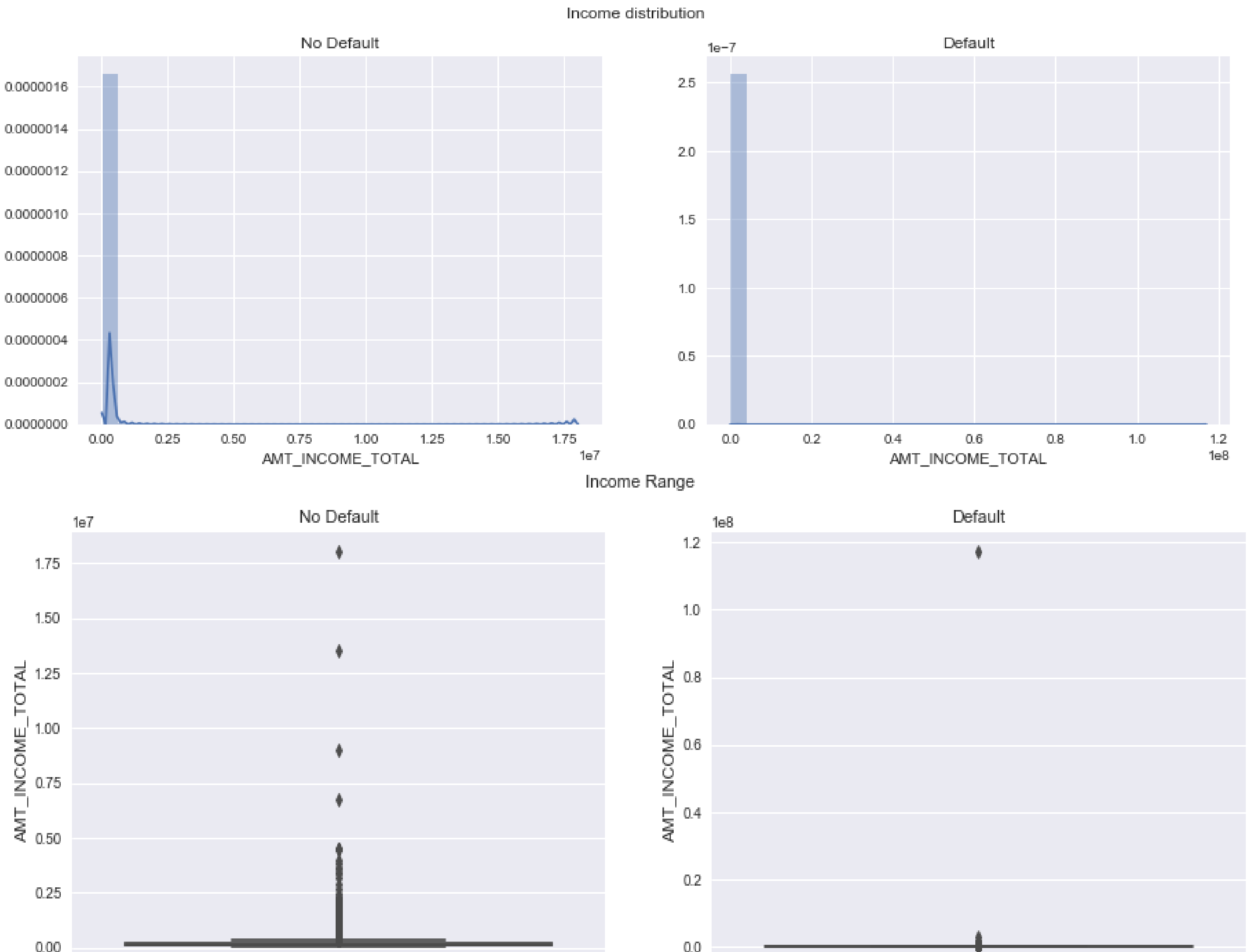
Here we can see that Occupation type has the maximum number of missing data. We can impute the occupation type categories by making income/salary bins of all the occupations and assign their occupation category on the basis of the bin their salary lies.

Data Overview



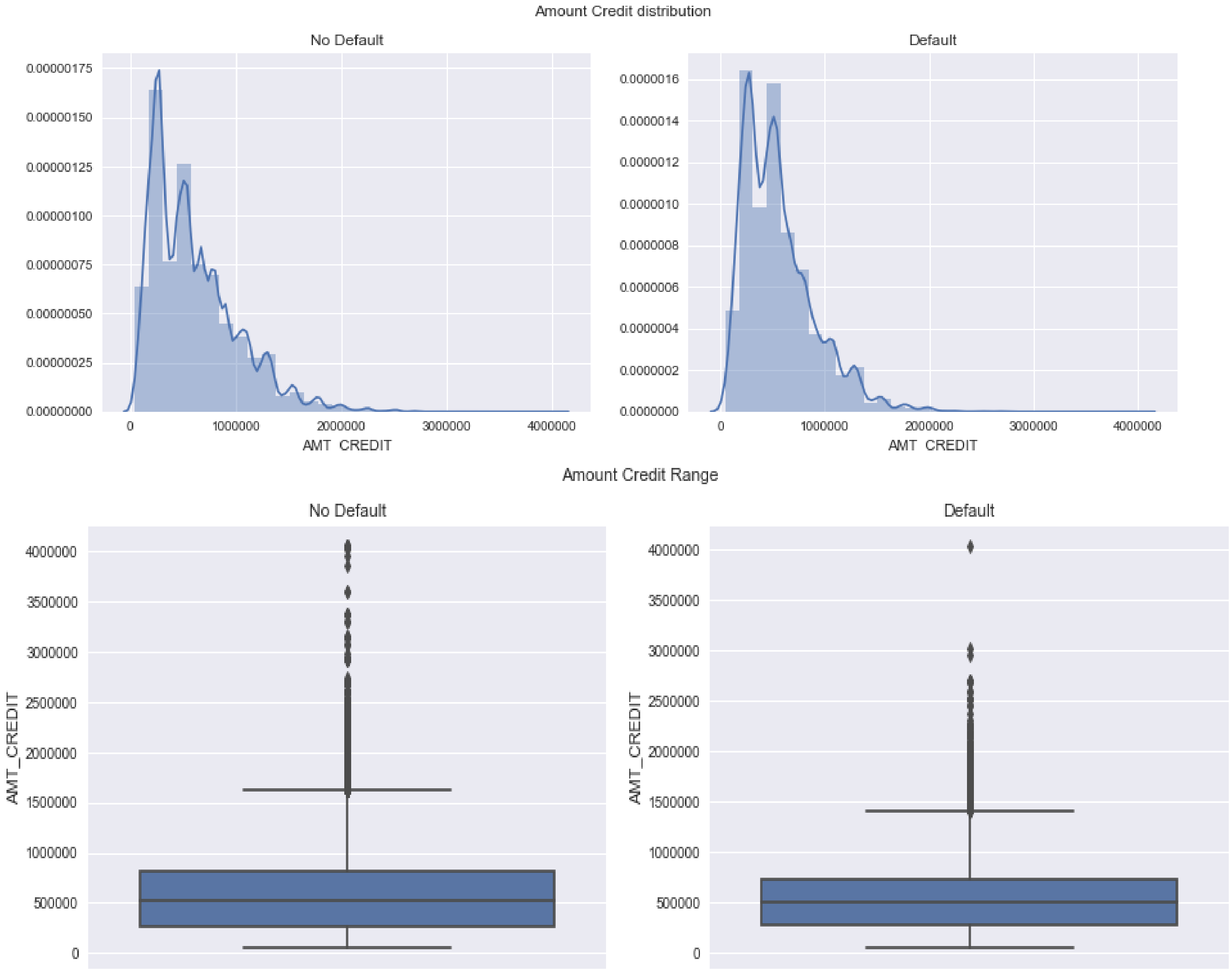
We have 92% of data as No Default and 8% data as Default. Data is highly imbalanced

Univariate Analysis(Numerical Variables) Income



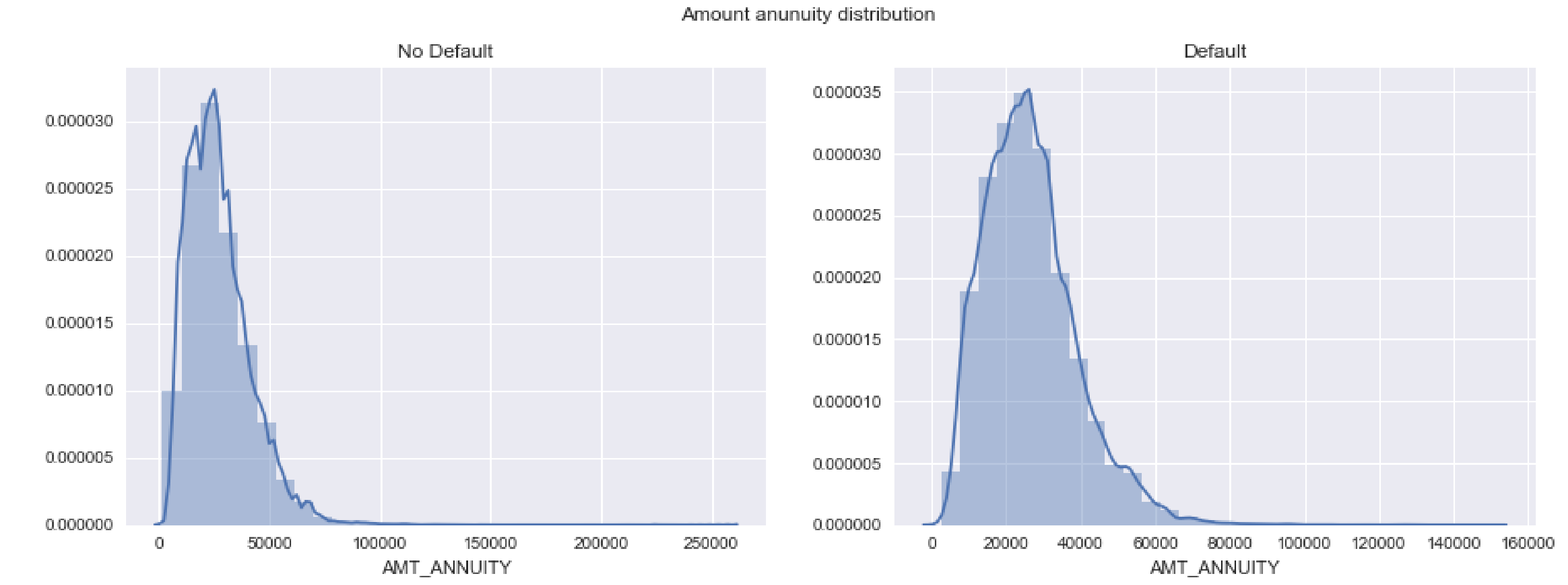
Income distribution is very skewed towards the lower side with very few outliers for the defaulter category but in significant amount for non-defaulters which shows clients having higher income range always repay their loans. The problem lies only for the perople lying in the low income range.

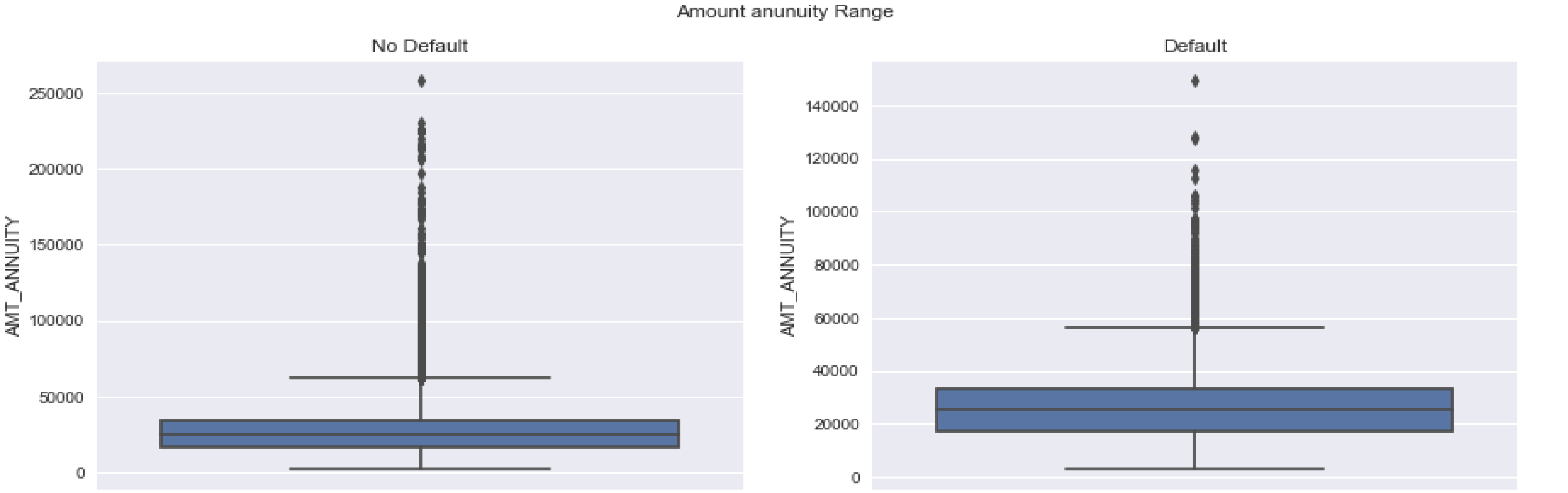
Credit Amount



For defaulter and Non-Defaulter the mean value is alomost equal with more defaulters lying between 0 and 1 million amount for credit.

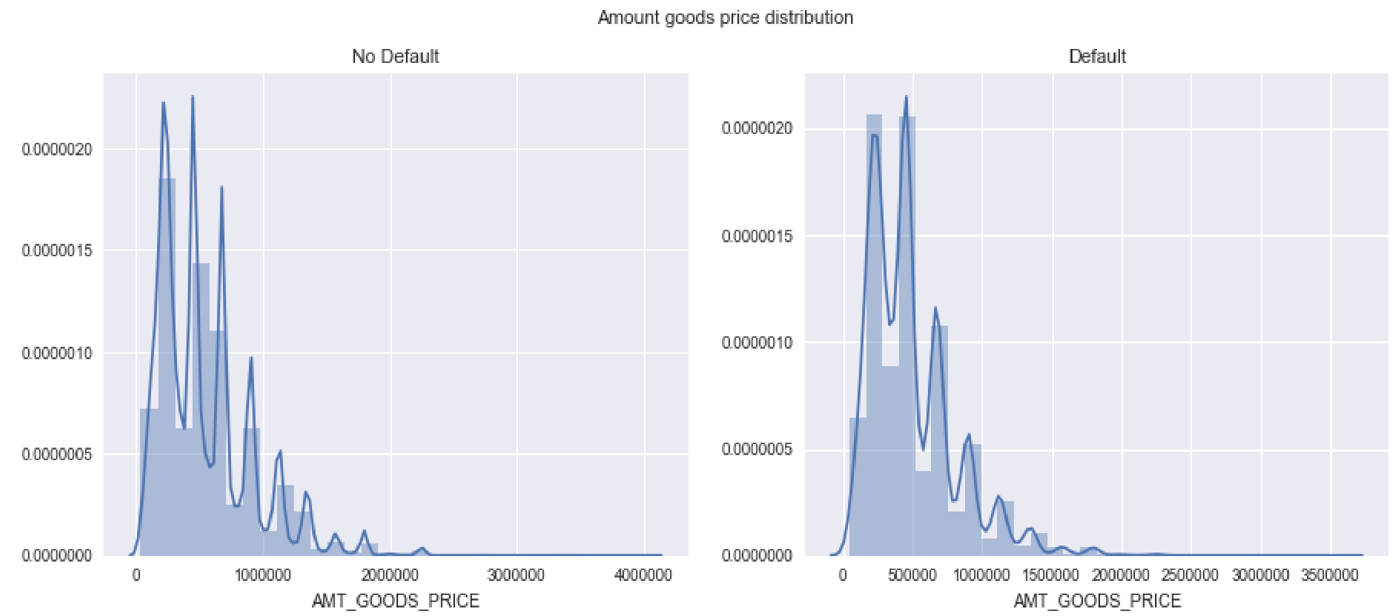
Annuity Amount

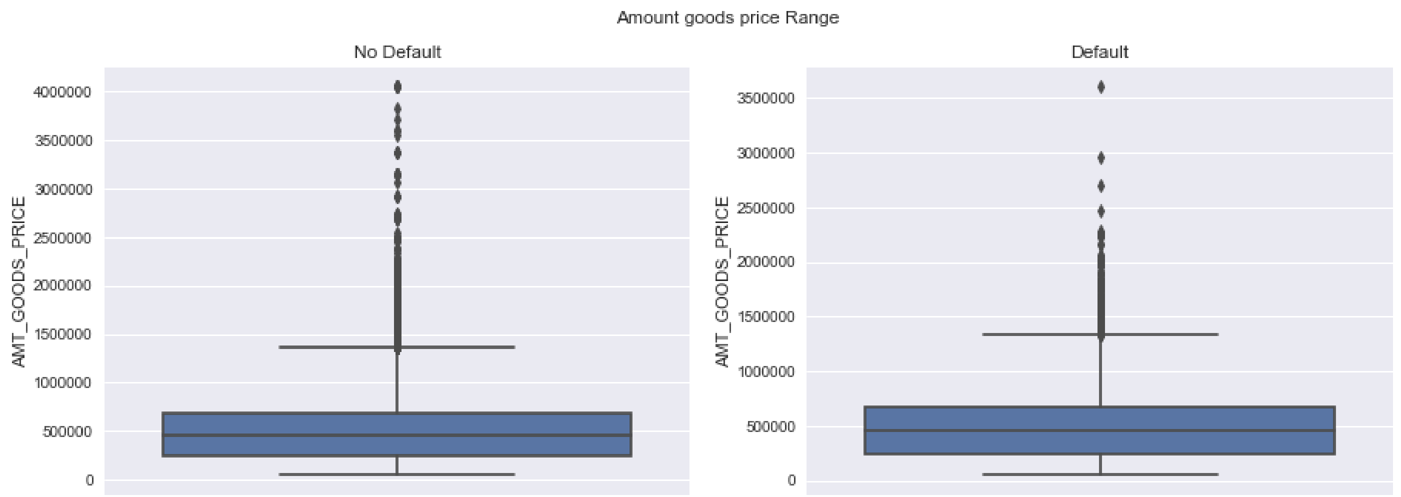




Defaulters have a higher mean value for amount annuity than the non-defaulters with most number of defaulters lying between 20,000 to 40,000 annuity amount.

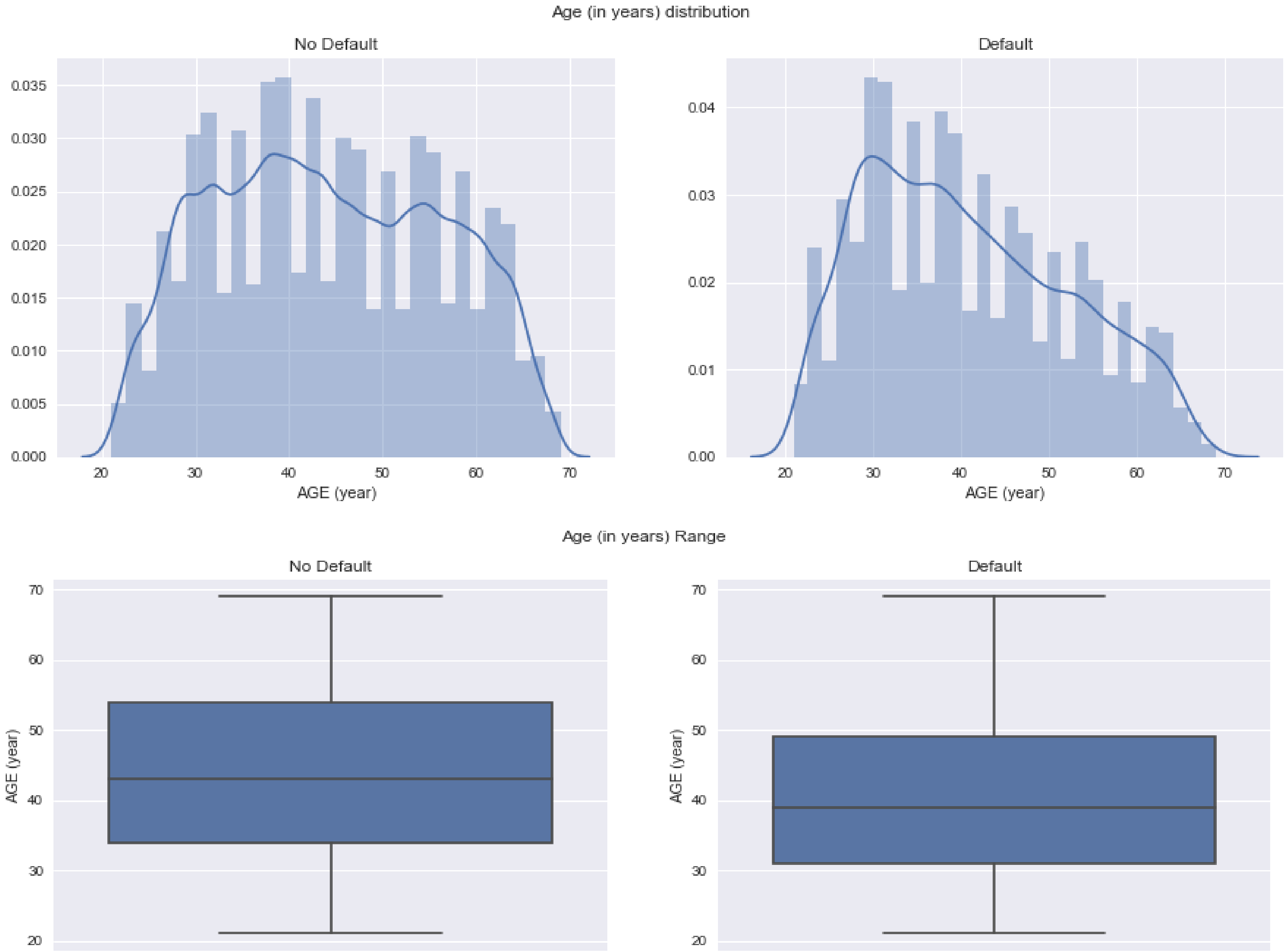
Goods Price





Goods price is almost similarly distributed among the defaulters and Non-Defaulters

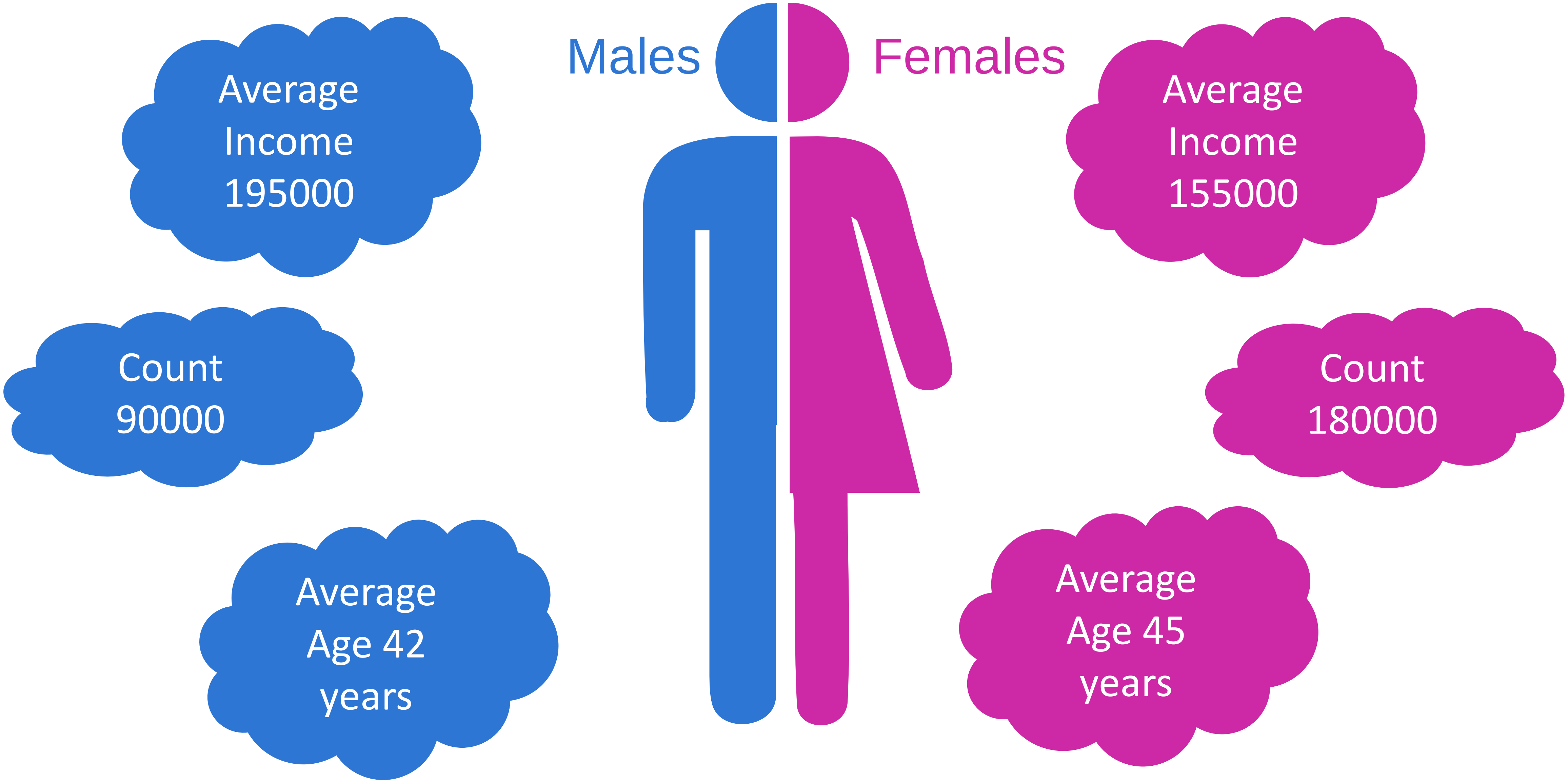
Age(in Years)



Defaulter average age is 40 whereas Non-Defaulter average age is 44. Quite close,difficult to distinguish

If the Mean age is more than 40 years there is low chances of default

Gender Distribution in Application data with Target as No Default



Gender Distribution in Application data with Target as Default

Males

Average Income 177000

Count 11000

Average Age 39 years

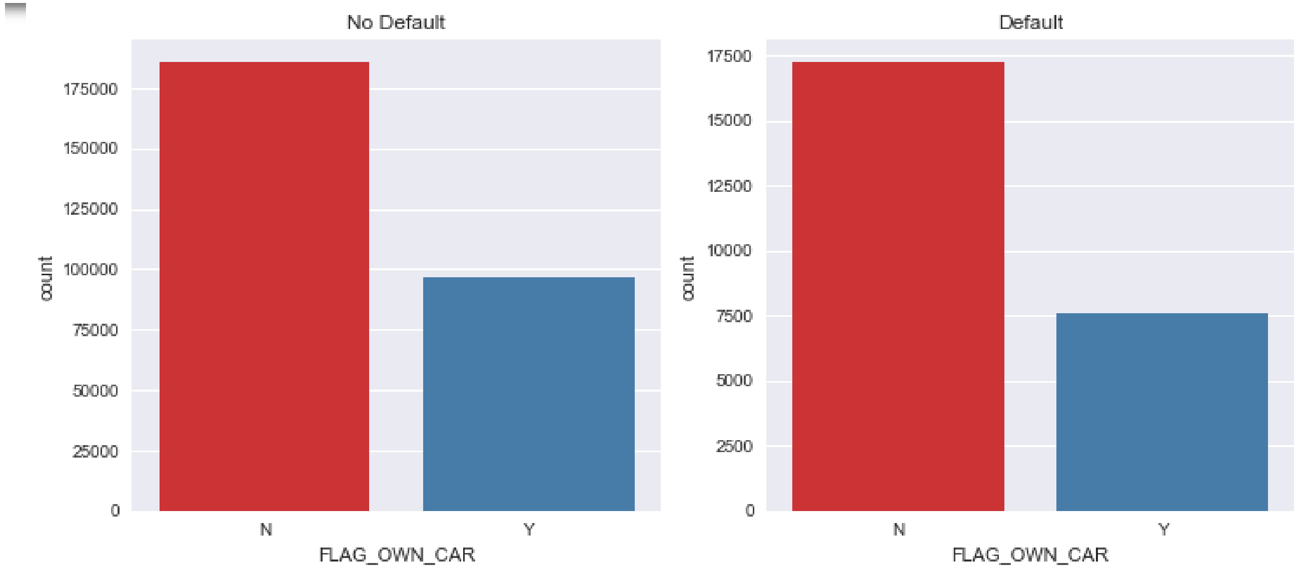
Females

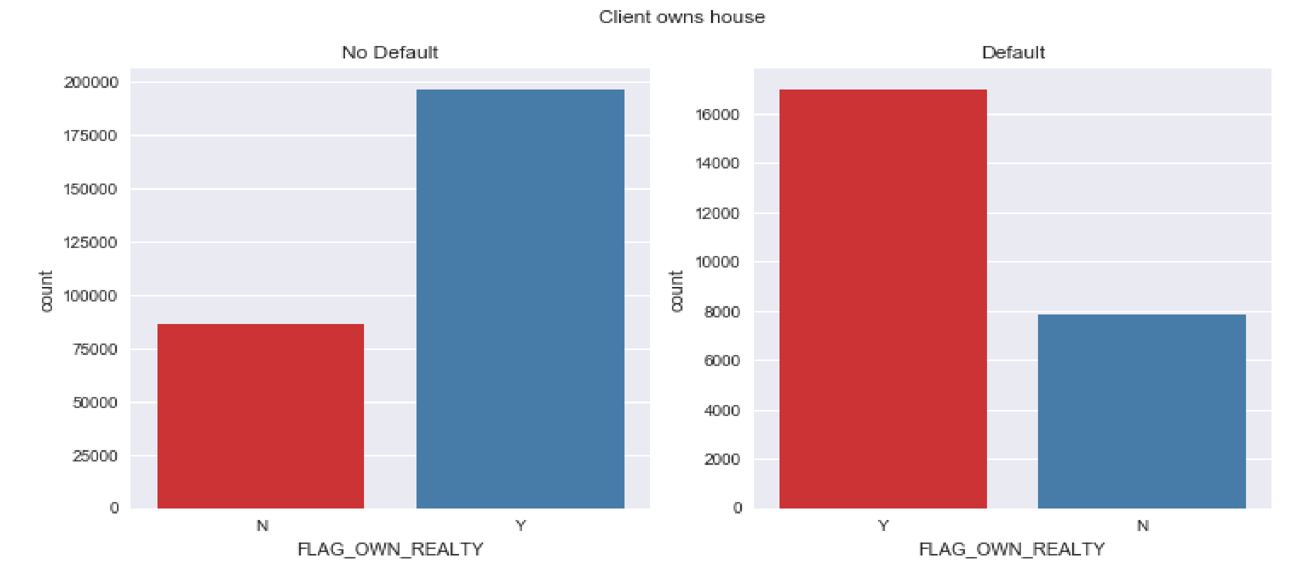
Average Income 156000

Count 14000

Average Age 41 years

Categorical Variables for Current Application
Own Car and Own Realty

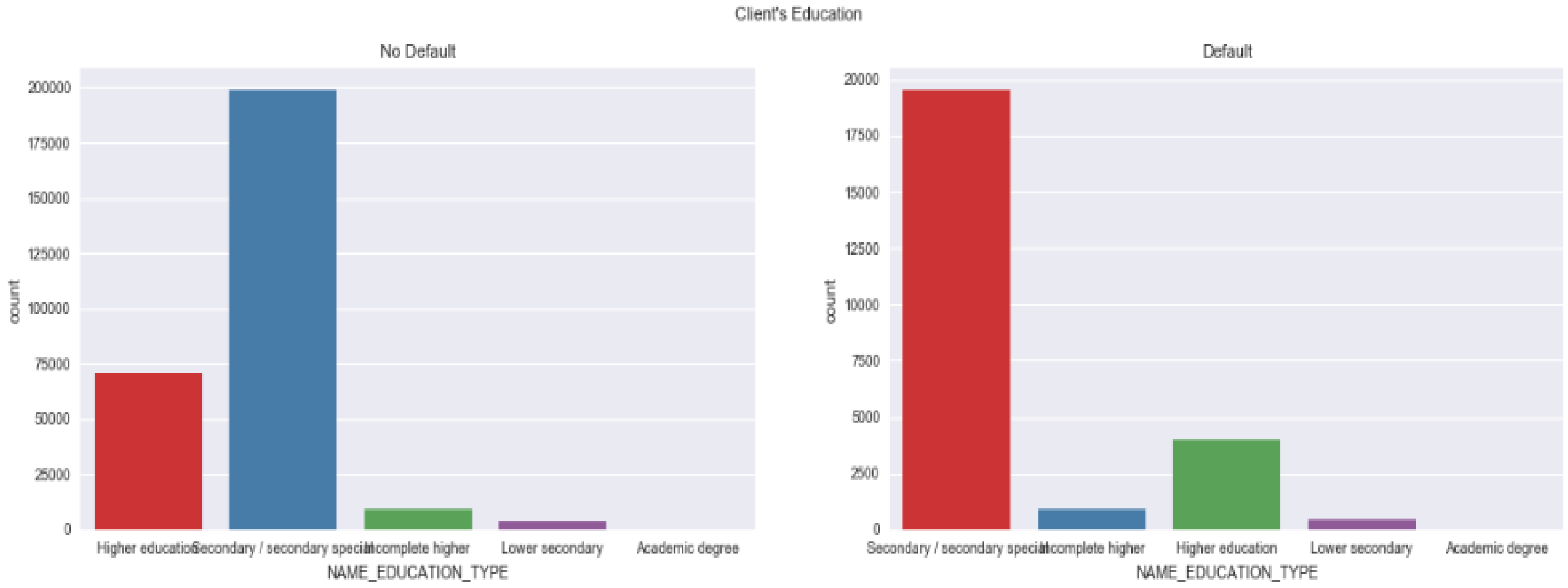


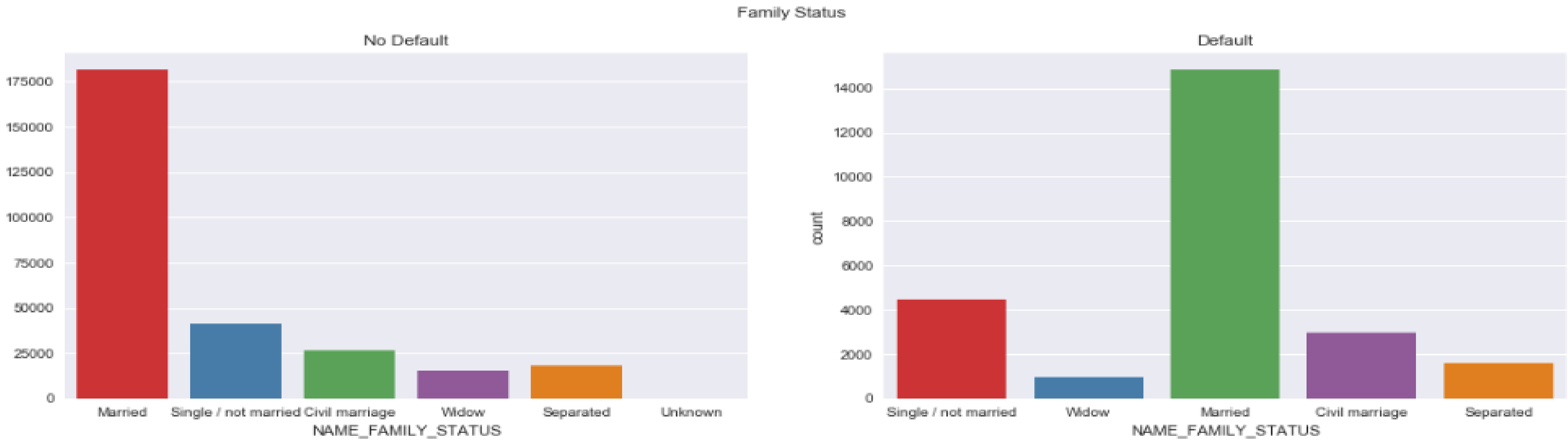


High number of people applied for a loan who don't own a car

High number of people applied for a loan who own a house

Education and Family





Correlation in Data

DEFAULTERS

	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	AGE (year)
AMT_CREDIT	1.000000	0.752195	0.983103	0.038131	0.135318
AMT_ANNUITY	0.752195	1.000000	0.752699	0.046421	0.014249
AMT_GOODS_PRICE	0.983103	0.752699	1.000000	0.037583	0.135744
AMT_INCOME_TOTAL	0.038131	0.046421	0.037583	1.000000	-0.002872
AGE (year)	0.135318	0.014249	0.135744	-0.002872	1.000000

NON DEFAULTERS

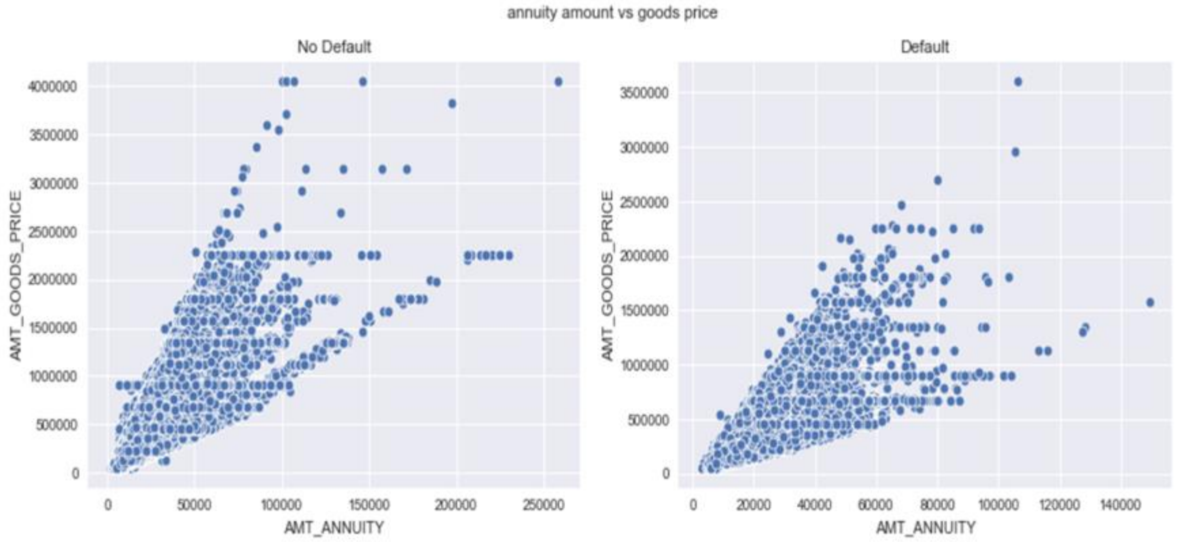
	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	AGE (year)
AMT_CREDIT	1.000000	0.771309	0.987250	0.342799	0.047426
AMT_ANNUITY	0.771309	1.000000	0.776686	0.418953	-0.012202
AMT_GOODS_PRICE	0.987250	0.776686	1.000000	0.349462	0.044601
AMT_INCOME_TOTAL	0.342799	0.418953	0.349462	1.000000	-0.062597
AGE (year)	0.047426	-0.012202	0.044601	-0.062597	1.000000

Top 2 correlations for continues numerical variables amoung defaulters and Non-defaulters are

- Amount Goods Price and Amount Credit
- Amount Goods Price and Amount Annuity

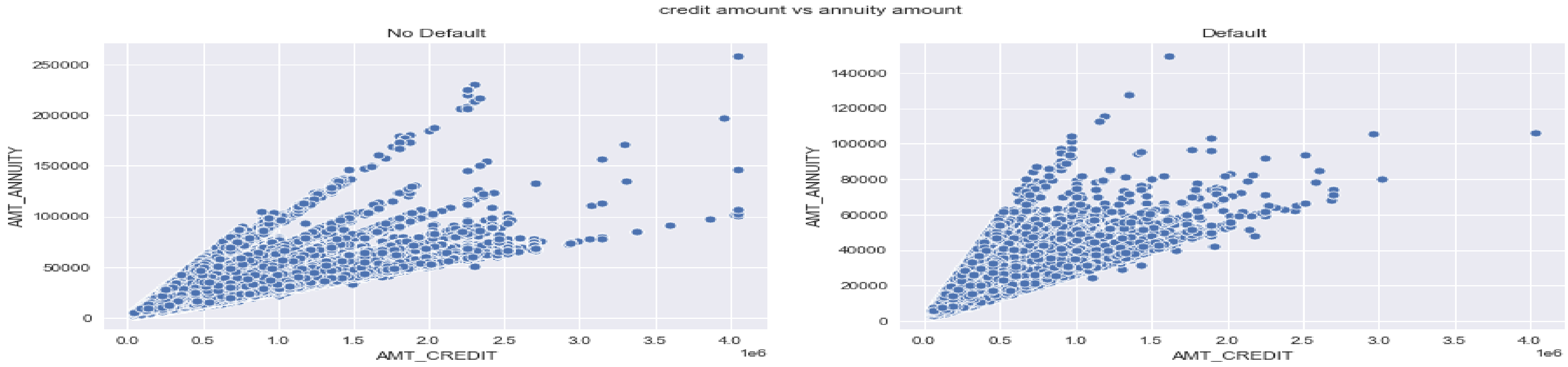
Age and Income has negative relationship which tells us that people with lower age which is 20-30(maximum) in our case has higher income as they are working professionals whereas older people are retired ones and they have less salary.

Scatter Plots for Correlation



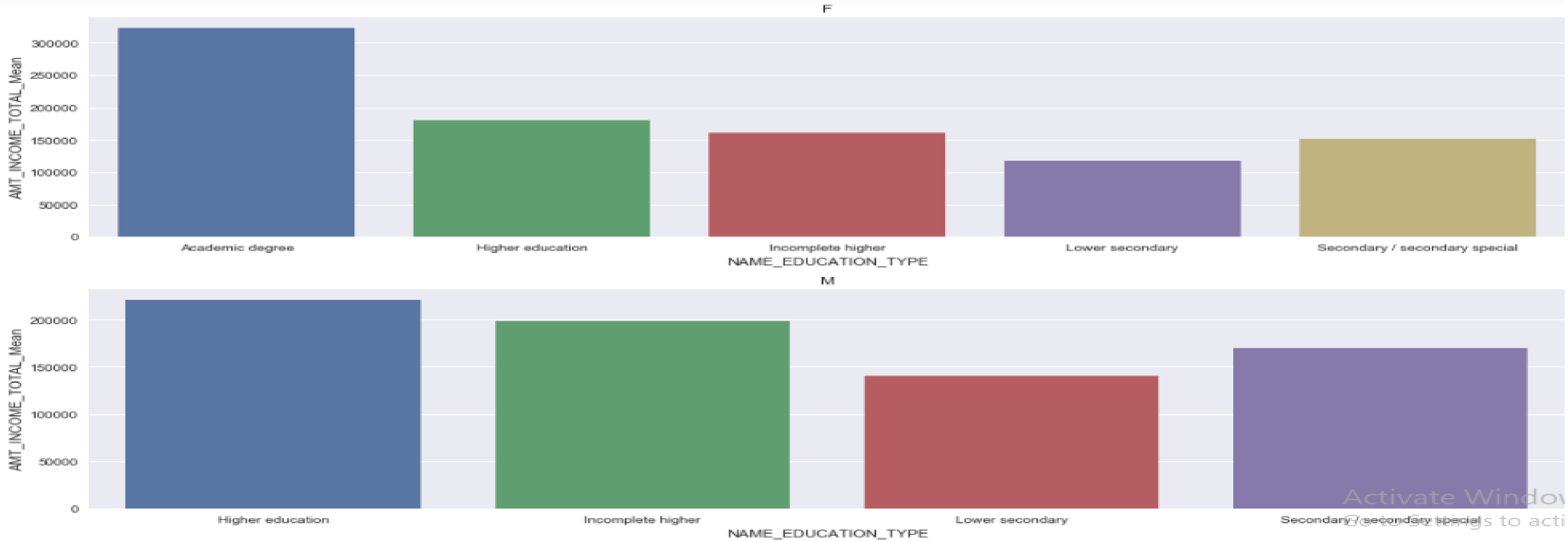
AMT_CREDIT & AMT_GOODS_PRICE has some linear relation

AMT_ANNUITY & AMT_GOODS_PRICE has some linear relation



AMT_CREDIT & AMT_ANNUITY has some linear relation

Income and Education type for Male and Female(Defaulters)



Data Analysis

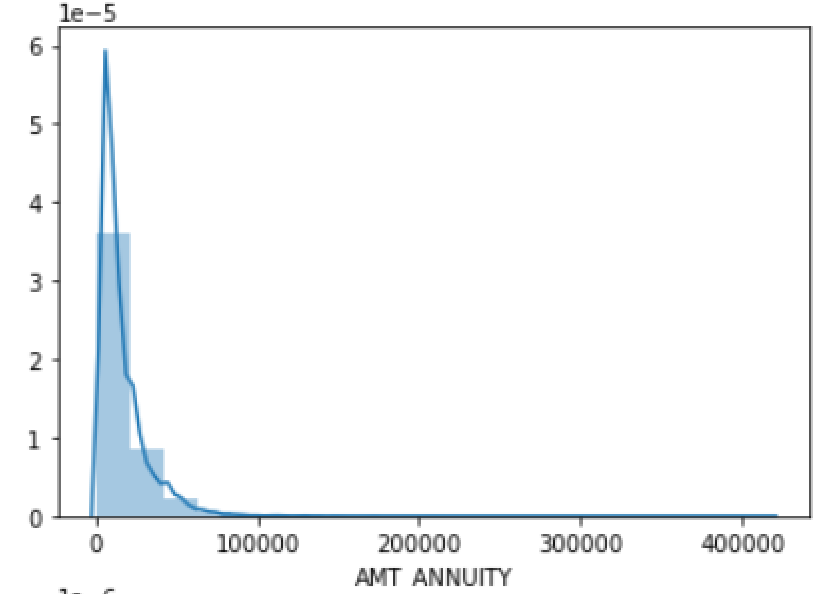
PREVIOUS APPLICATION

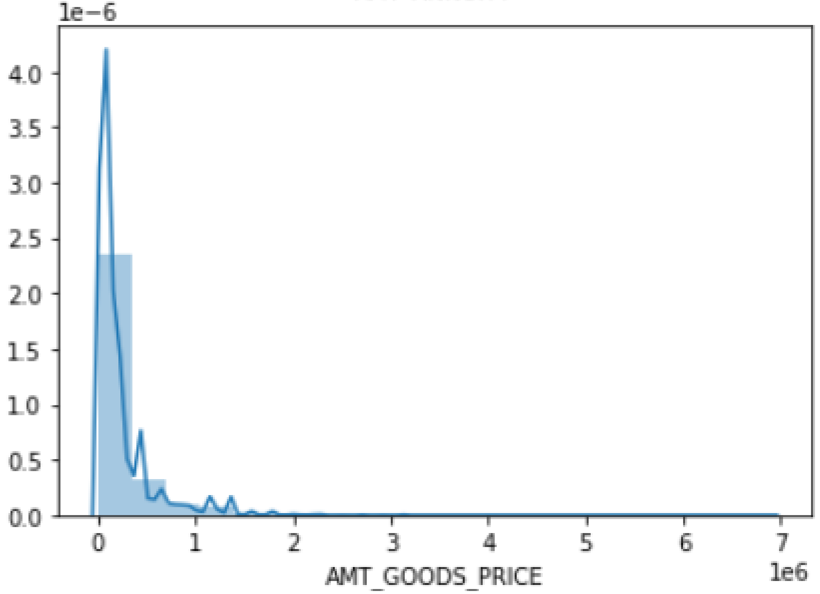
Data Overview

Columns	% Missing
NAME_TYPE_SUITE	49.11
NFLAG_INSURED_ON_APPROVAL	40.29
DAYS_TERMINATION	40.29
DAYS_LAST_DUE	40.29
DAYS_LAST_DUE_1ST_VERSION	40.29
DAYS_FIRST_DUE	40.29
DAYS_FIRST_DRAWING	40.29
AMT_GOODS_PRICE	23.08
AMT_ANNUITY	22.28
CNT_PAYMENT	22.28
PRODUCT_COMBINATION	0.02

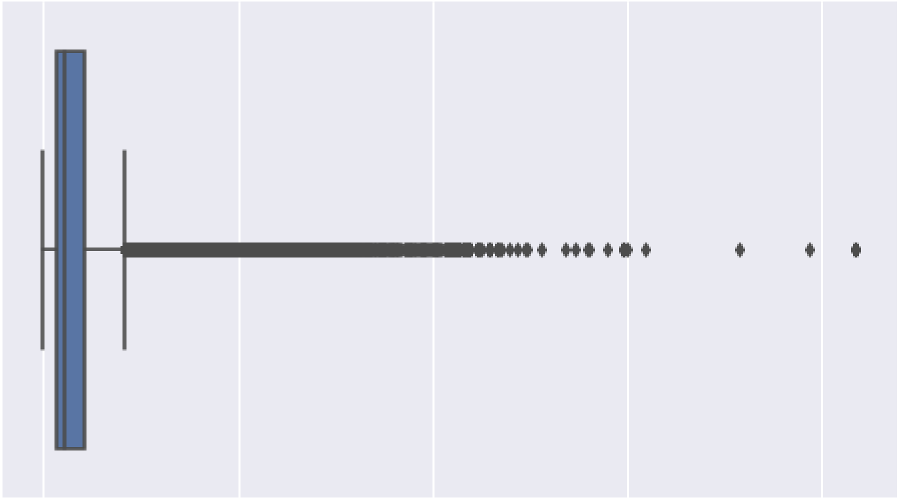
```
RangeIndex: 1670214 entries, 0 to 1670213
Data columns (total 37 columns):
SK_ID_PREV                     1670214 non-null int64
SK_ID_CURR                     1670214 non-null int64
NAME_CONTRACT_TYPE             1670214 non-null object
AMT_ANNUITY                    1297979 non-null float64
AMT_APPLICATION                1670214 non-null float64
AMT_CREDIT                     1670213 non-null float64
AMT_DOWN_PAYMENT               774370 non-null float64
AMT_GOODS_PRICE                1284699 non-null float64
WEEKDAY_APPR_PROCESS_START     1670214 non-null object
HOUR_APPR_PROCESS_START        1670214 non-null int64
FLAG_LAST_APPL_PER_CONTRACT    1670214 non-null object
NFLAG_LAST_APPL_IN_DAY         1670214 non-null int64
RATE_DOWN_PAYMENT              774370 non-null float64
RATE_INTEREST_PRIMARY          5951 non-null float64
RATE_INTEREST_PRIVILEGED       5951 non-null float64
NAME_CASH_LOAN_PURPOSE         1670214 non-null object
NAME_CONTRACT_STATUS           1670214 non-null object
DAYS_DECISION                  1670214 non-null int64
NAME_PAYMENT_TYPE              1670214 non-null object
CODE_REJECT_REASON             1670214 non-null object
NAME_TYPE_SUITE                849809 non-null object
NAME_CLIENT_TYPE               1670214 non-null object
NAME_GOODS_CATEGORY            1670214 non-null object
NAME_PORTFOLIO                 1670214 non-null object
NAME_PRODUCT_TYPE              1670214 non-null object
CHANNEL_TYPE                   1670214 non-null object
SELLERPLACE_AREA               1670214 non-null int64
NAME_SELLER_INDUSTRY           1670214 non-null object
CNT_PAYMENT                    1297984 non-null float64
NAME_YIELD_GROUP               1670214 non-null object
PRODUCT_COMBINATION            1669868 non-null object
DAYS_FIRST_DRAWING             997149 non-null float64
DAYS_FIRST_DUE                 997149 non-null float64
DAYS_LAST_DUE_1ST_VERSION      997149 non-null float64
DAYS_LAST_DUE                  997149 non-null float64
DAYS_TERMINATION               997149 non-null float64
NFLAG_INSURED_ON_APPROVAL      997149 non-null float64
```

Annuity Amount and Goods Price

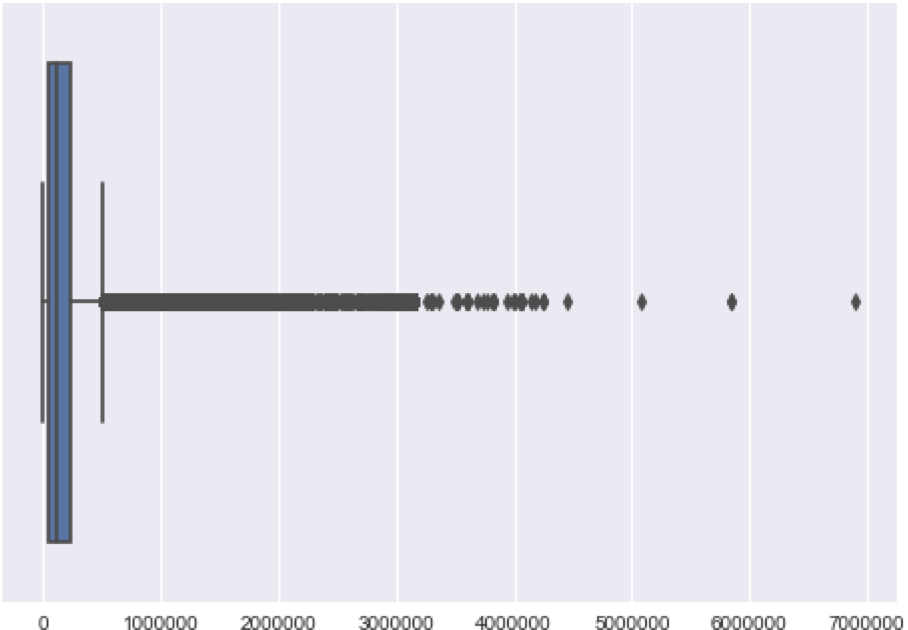




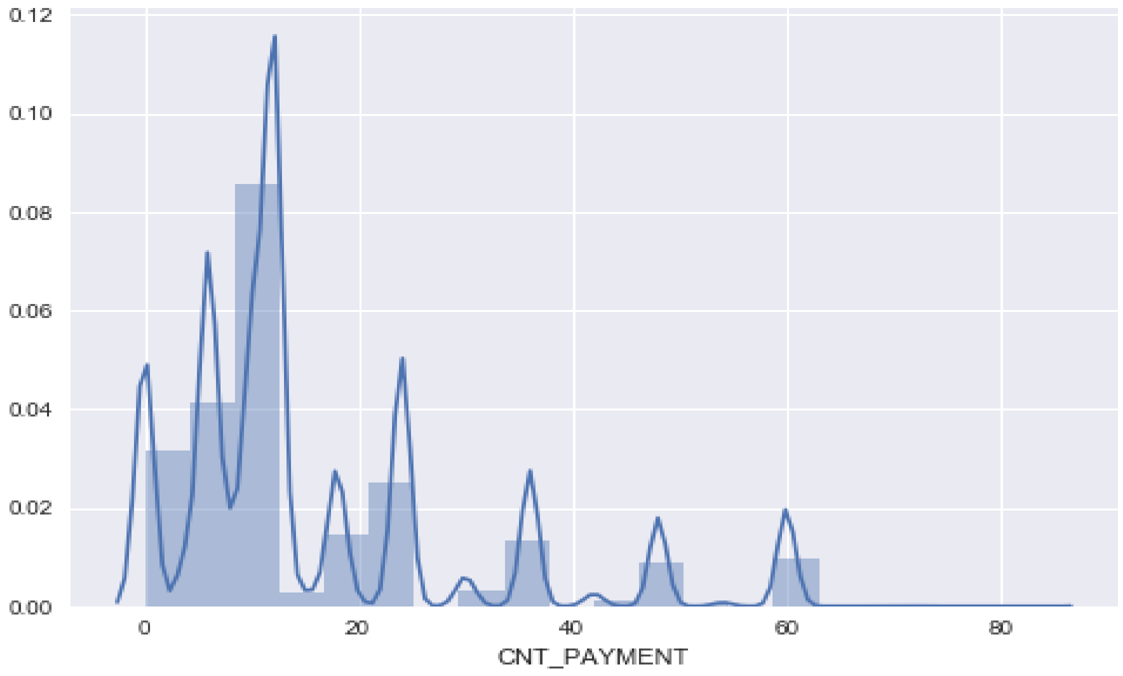
Annuity for previous application

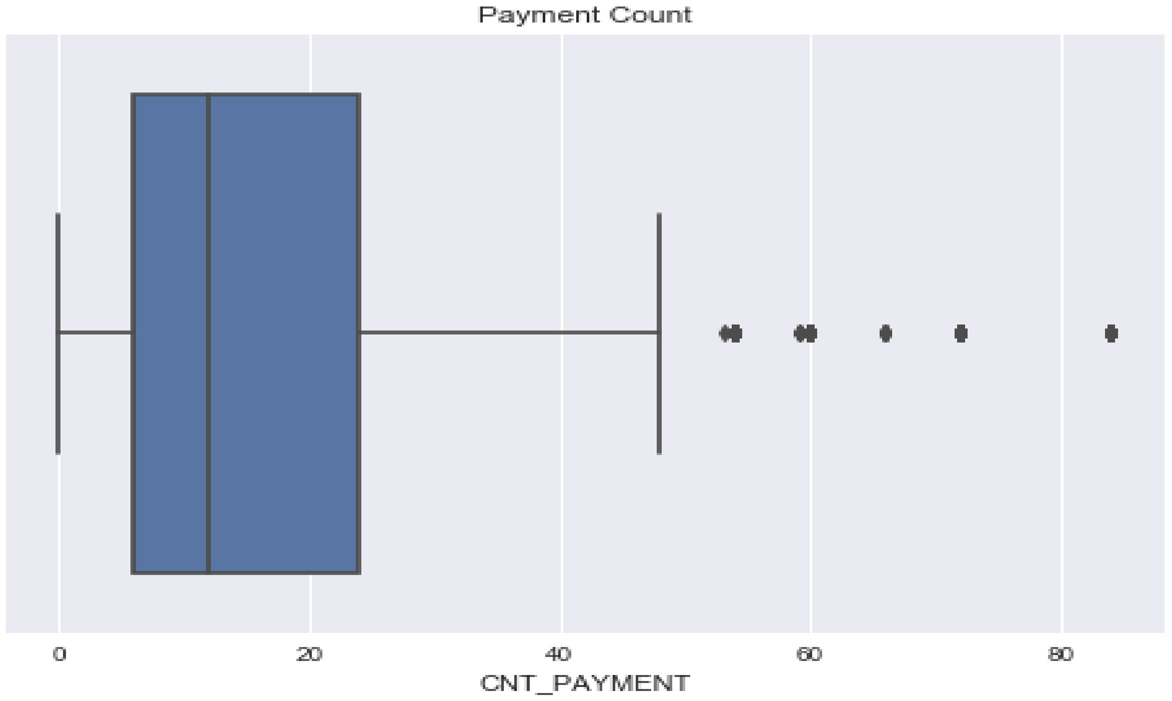


Goods price of previous application



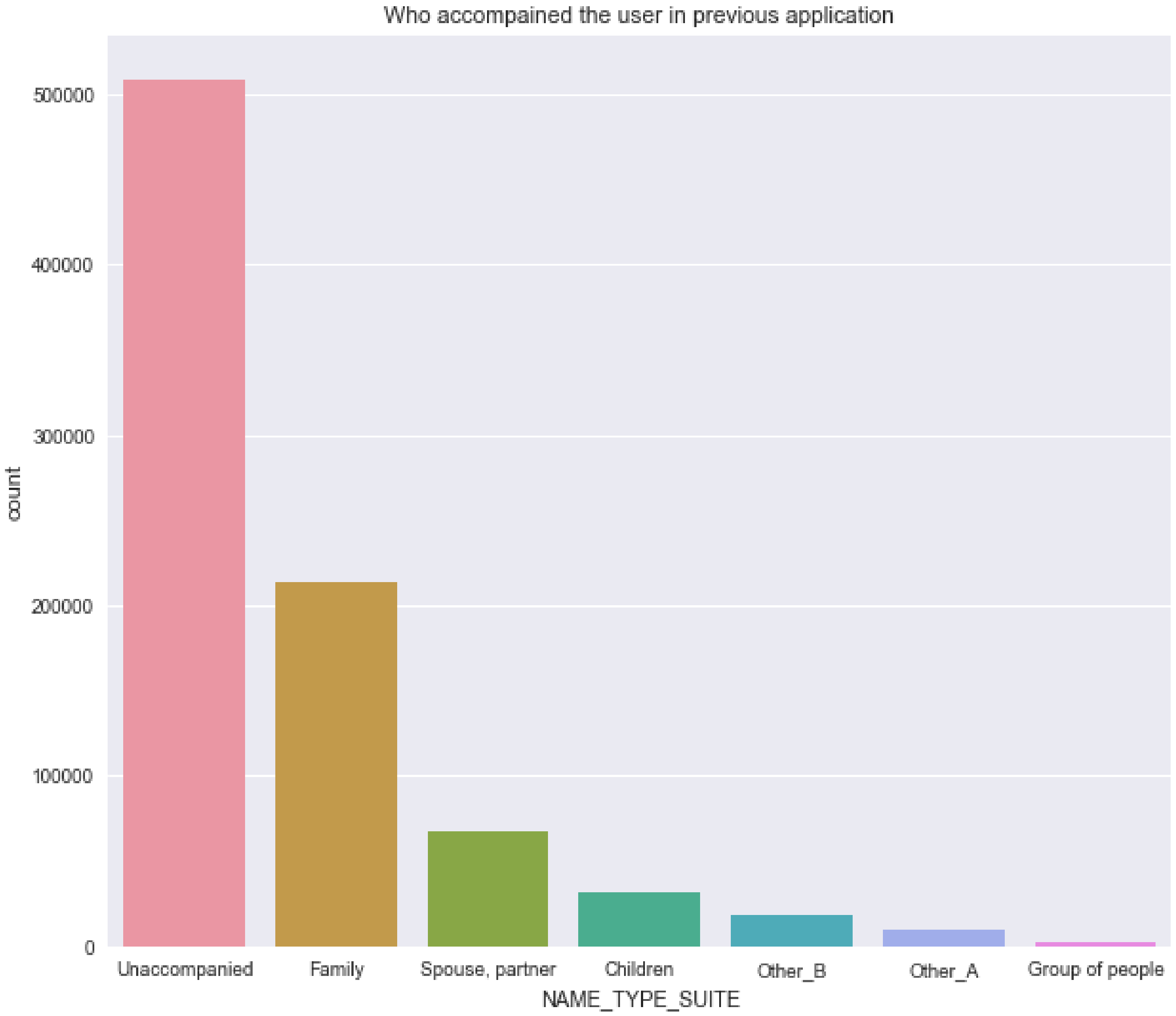
Payment Count





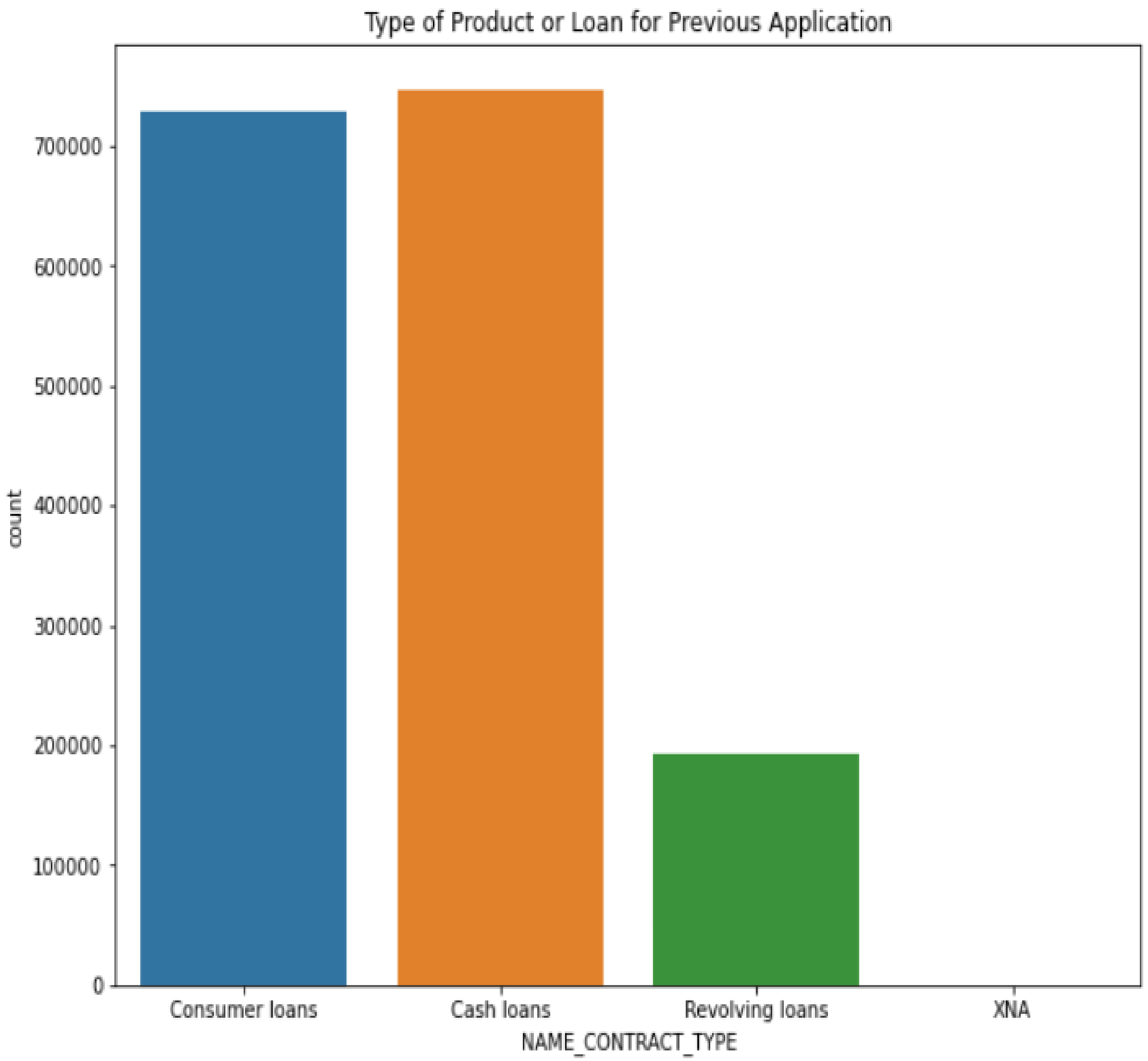
Data is not skewed and the spread is also high therefore the missing values can be impted with mean value

Who accompanied the user in previous application



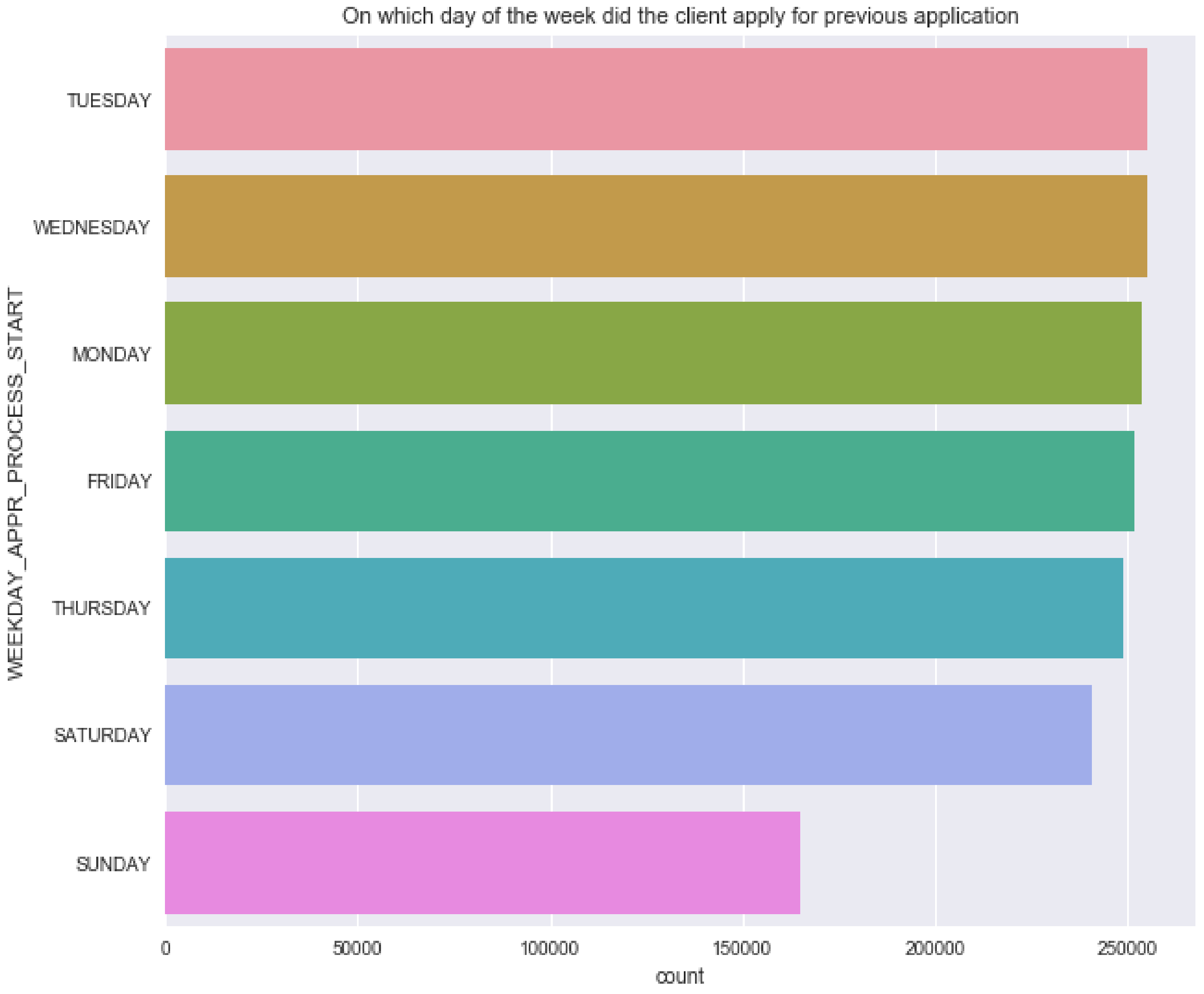
As we see that most of the times client came unaccompanied in the previous application followed by Family and spouse or partner

Top Product for Previous Application



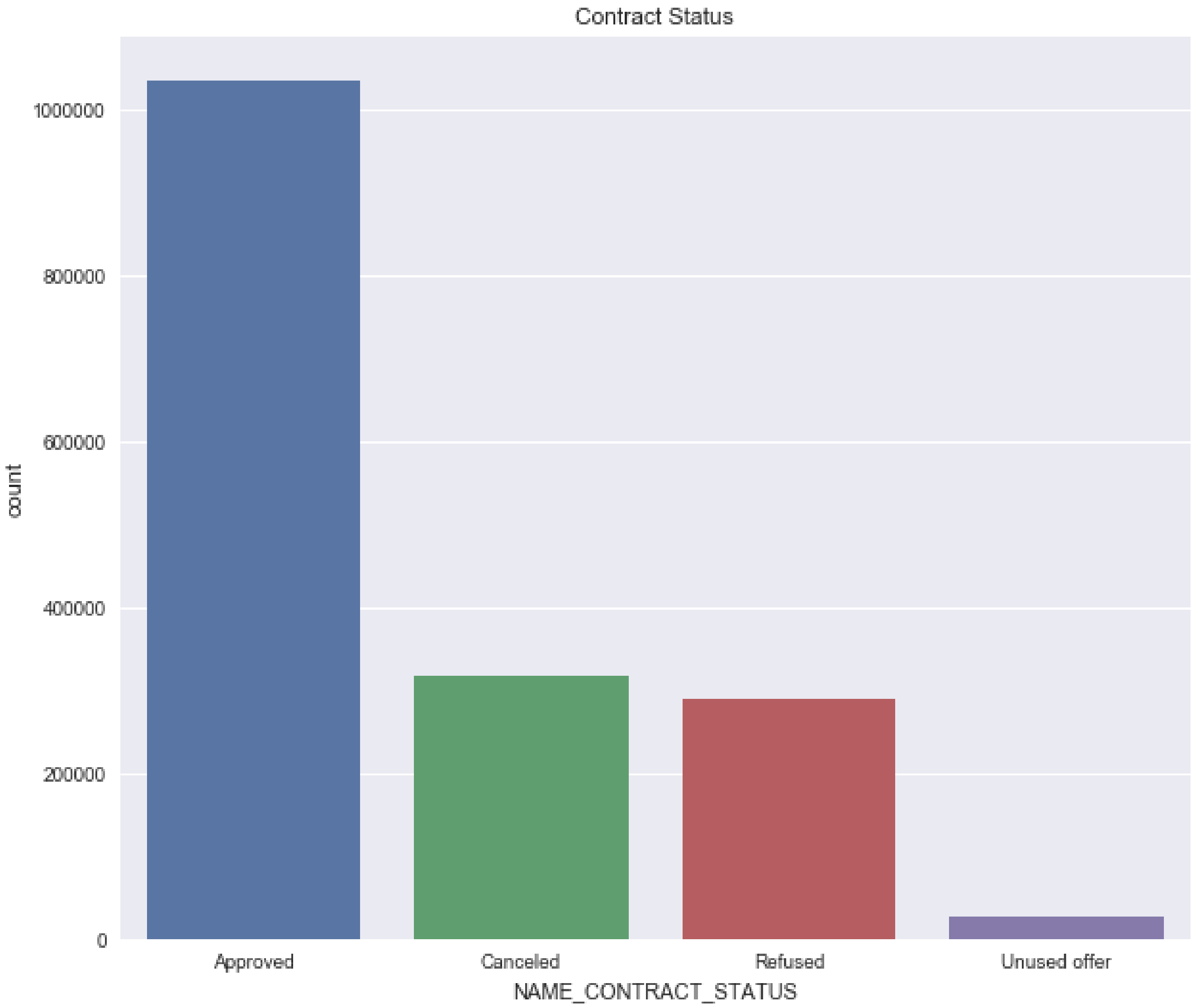
Here we can say that most common contract type for previous application is Cash Loans and Consumer Loans in the previous application

Day of the week



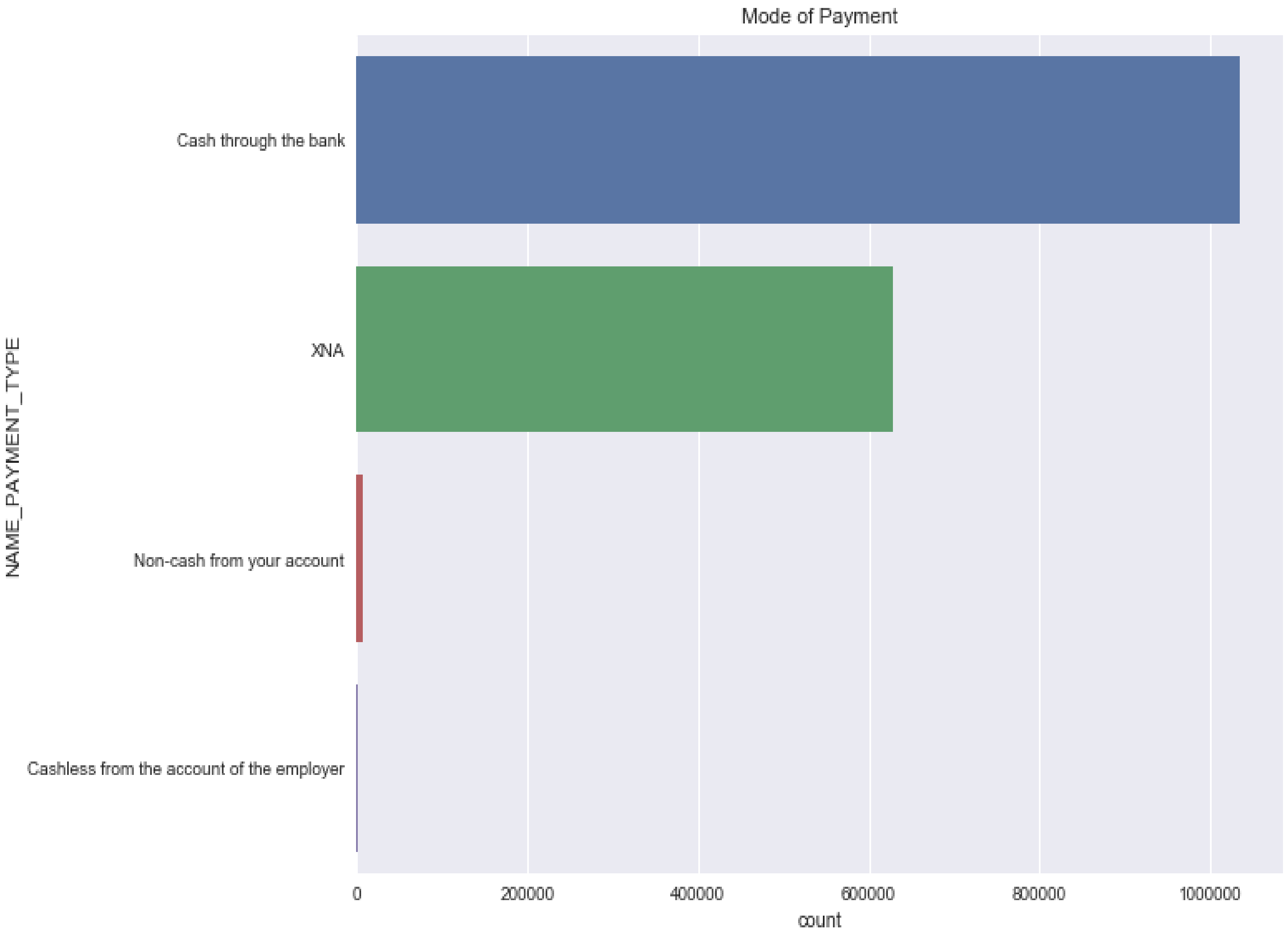
Tuesday followed by Wednesday and Monday show maximum number of previous client application and Sunday being lowest because the loan providing company has mostly sundays as holidays.

Contract Status



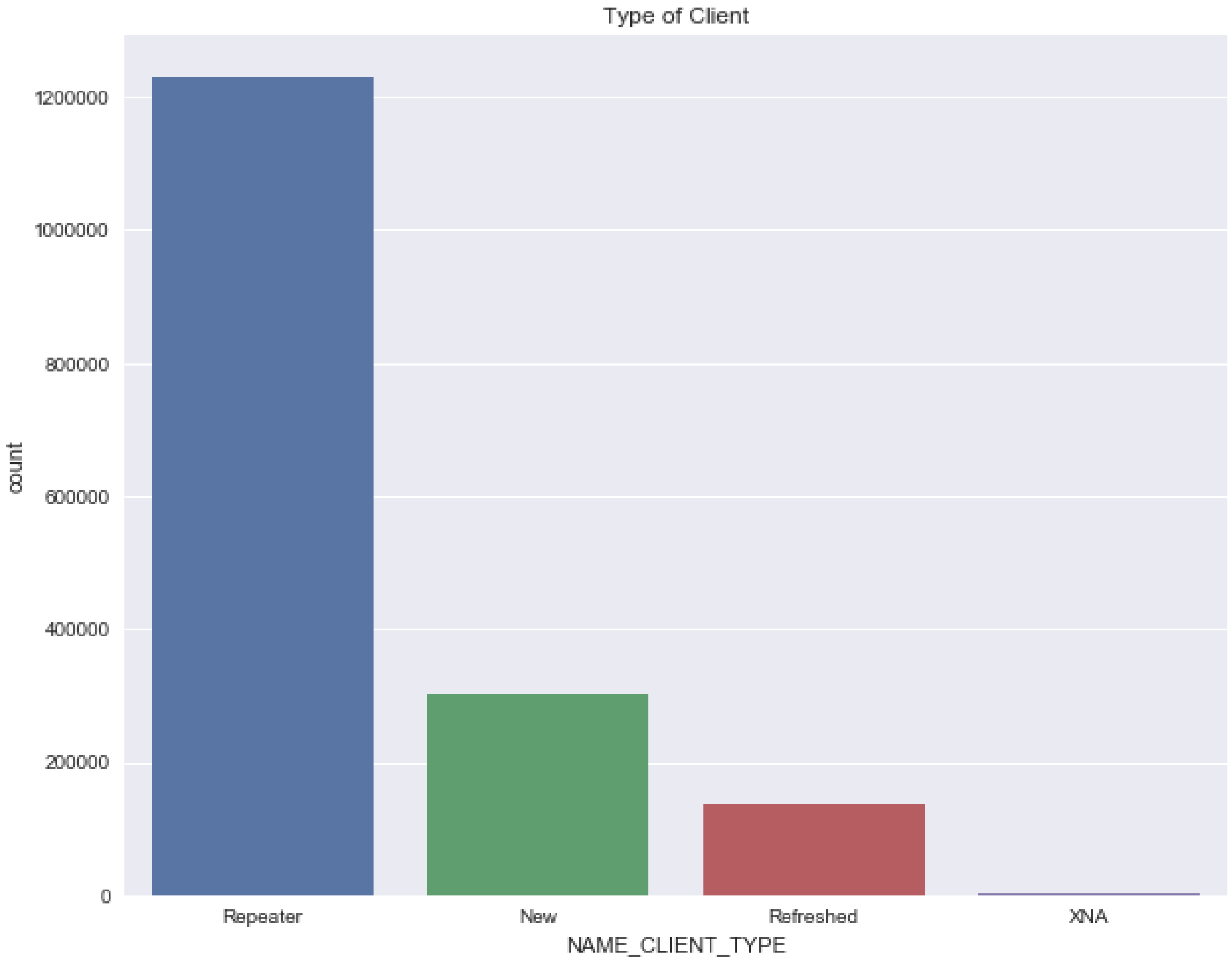
Most of the contracts got approved by the company in the previous applications

Mode of Payment



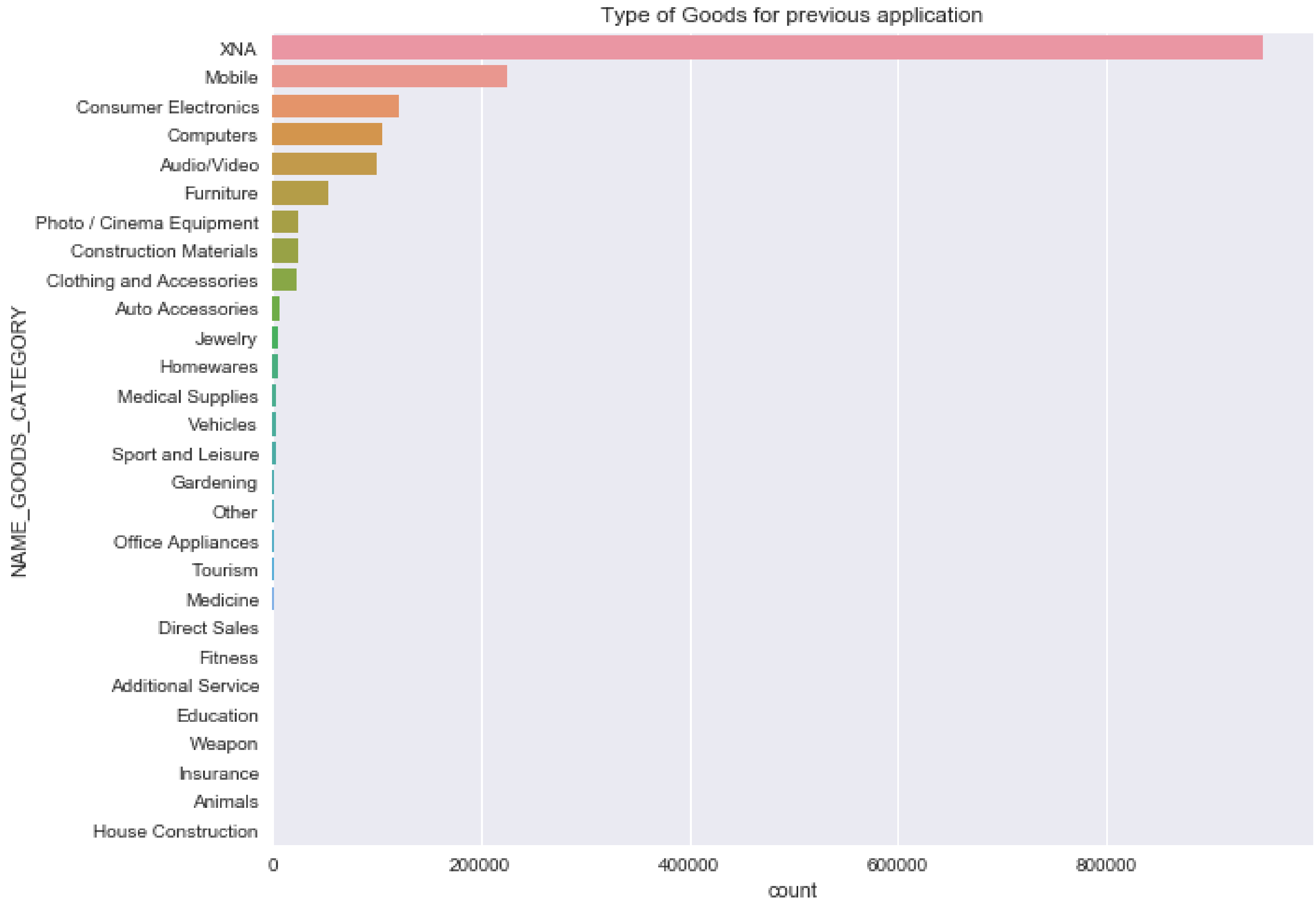
For most of the previous applications mode of payment has been cash from the bank with 'XNA' denoting cash from untraceable resource.

Type of client



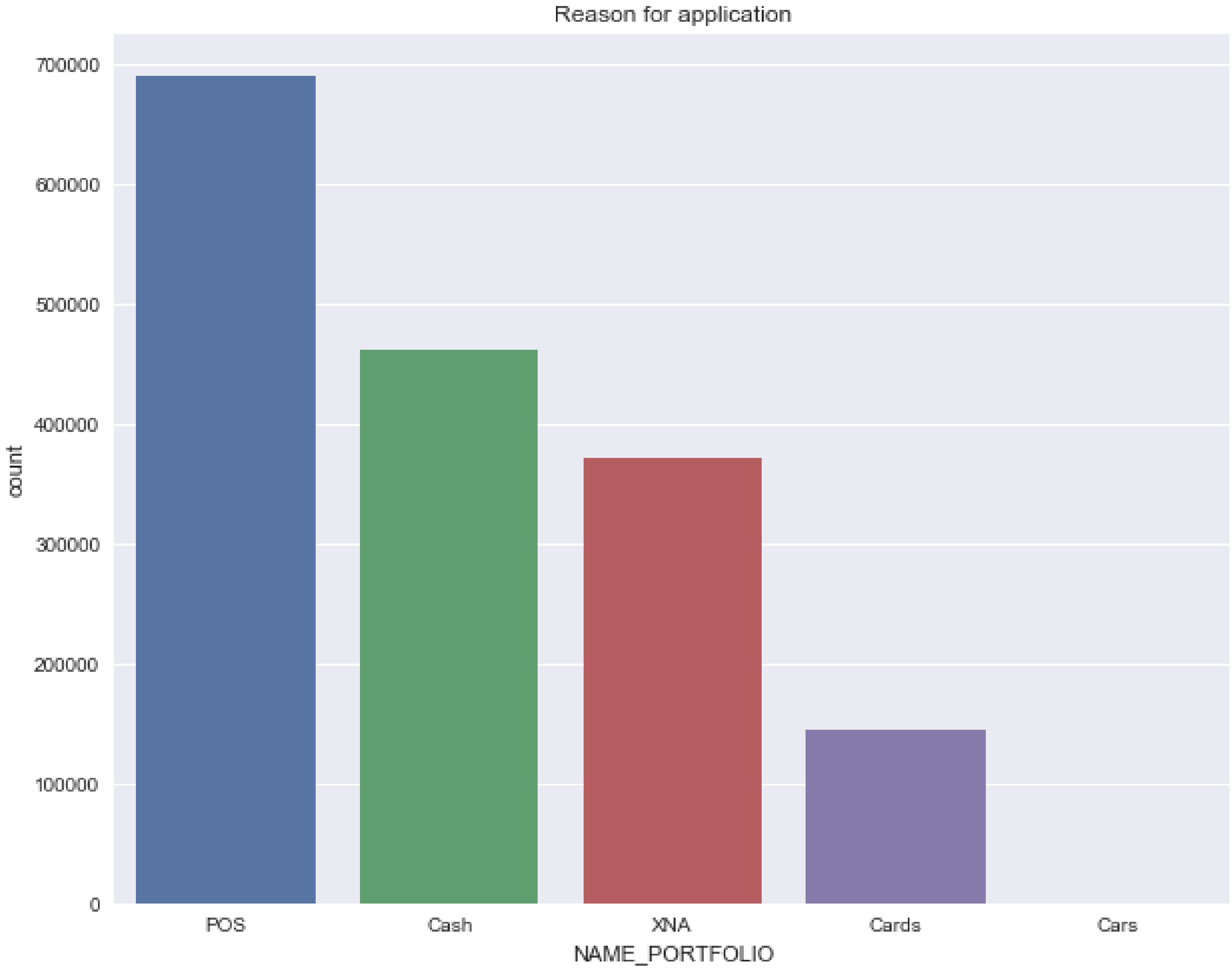
There is very high number of clients from the repeater batch followed by New clients showing a very high difference in number. Therefore we can say that the Company's Retention rate is very high but with a low acquisition rate in case of clients in the previous applications.

Type of Goods for previous application



Here we can see that most of the values are 'XNA'. It can be that most of the clients didn't choose to fill the name of the Goods category,it can be optional. Rest all other goods category belongs to Electronic devices.

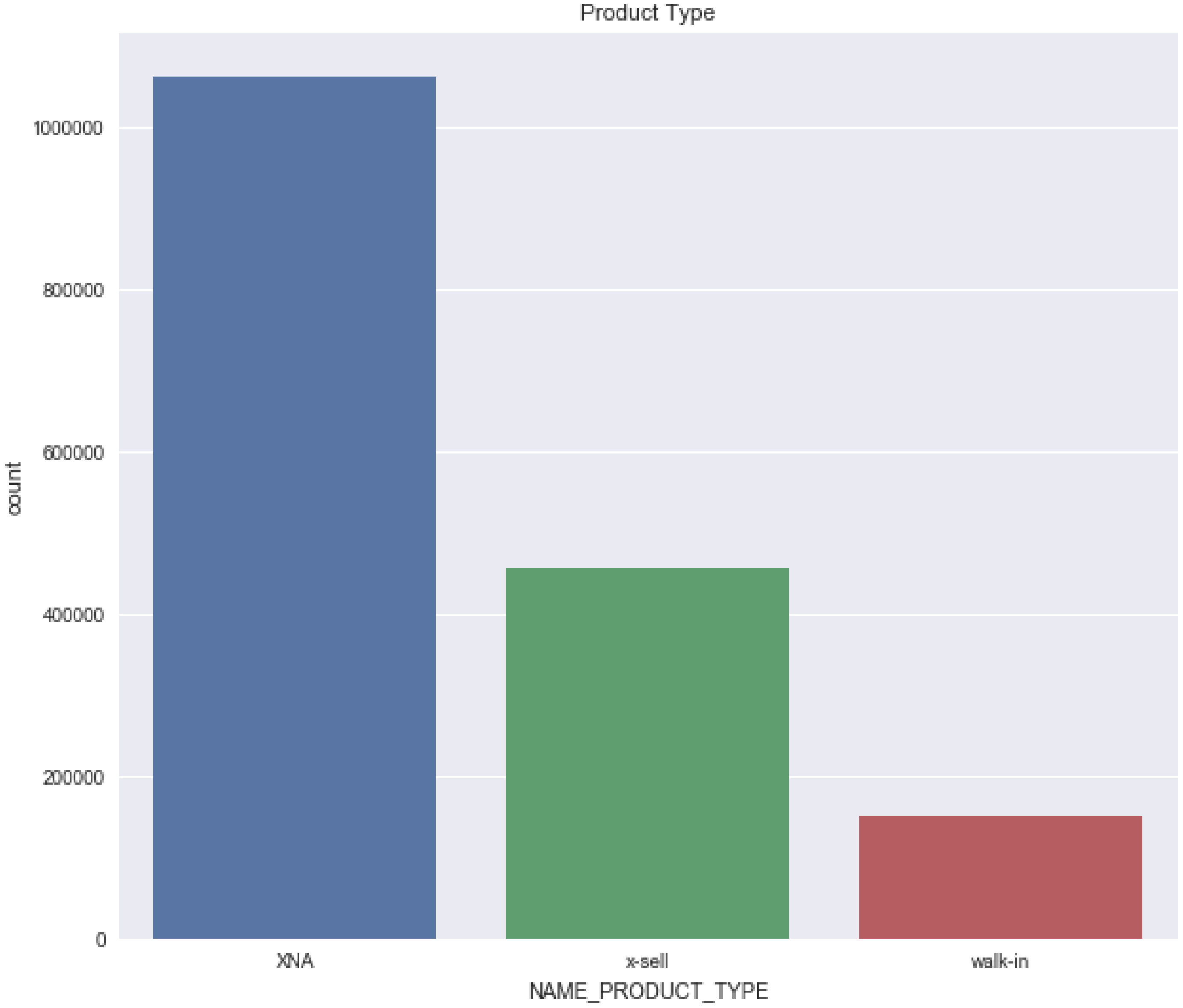
Reason for application



The Reason for loan application for previous application which is most common is 'POS'(Point Of Sale) which means when the merchant offers their customers a financial solution at the point of purchase, in order to assist them in buying the product or service. POS financing is a type of consumer finance and refers to open loop credit cards, closed loop store cards and installment loans followed by Cash loans.

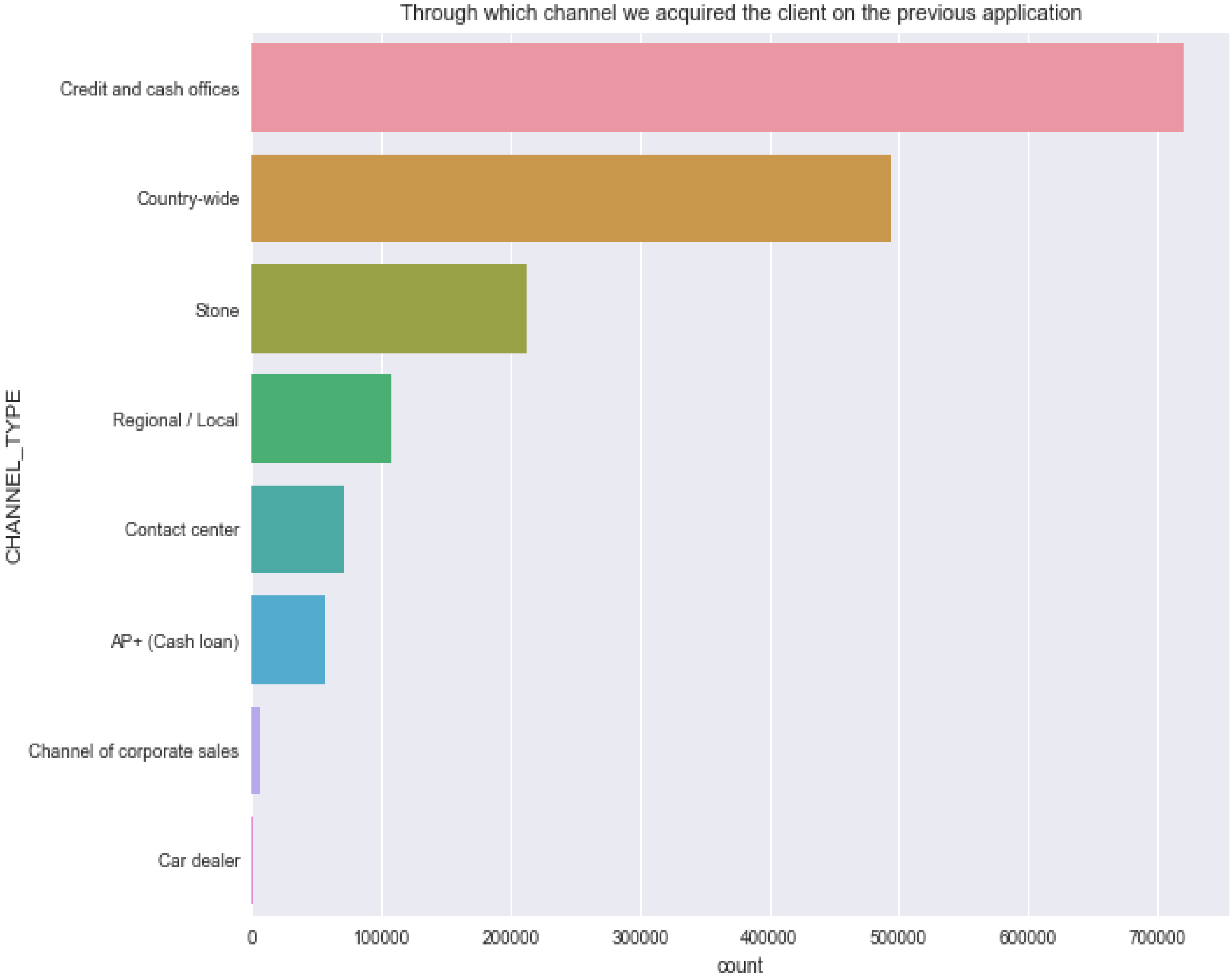
'XNA' denotes that many applicants didn't choose to give the reason in previous applications.

Product Type



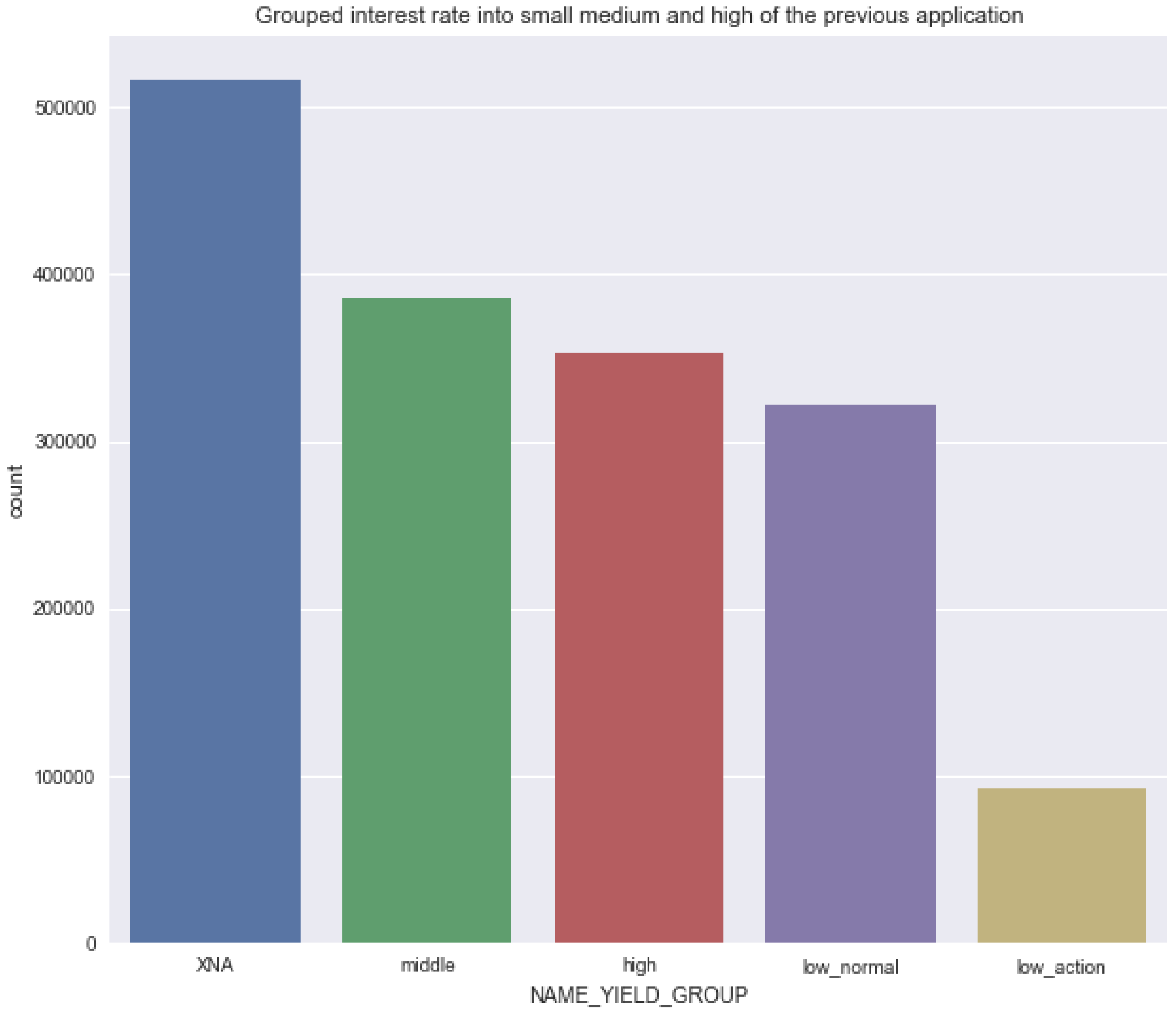
Most of the values show 'XNA' values which can be imputed with the 2nd highest that is x-sell rest others are walk-in.
'x-sell' is a short form for 'cross sell'. Means that the buyer had already brought some other product from the company and then this credit was sold as a second related or similar product to the buyer.
'Walk-in' means - the customer walked into Home credit branch on his own and applied for the credit.

Channel type



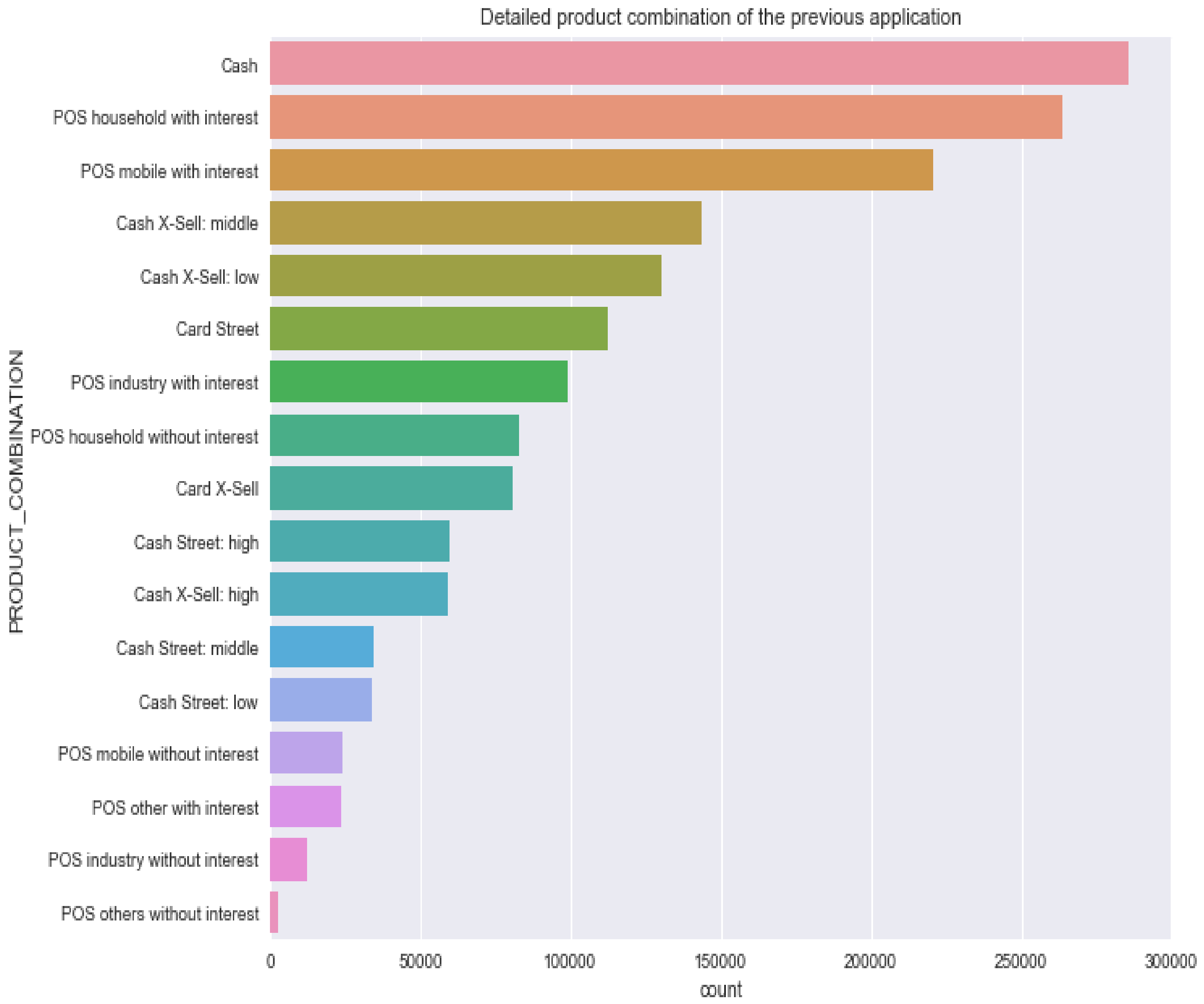
The most common channel the company acquired the client on the previous application is Credit and Cash offices and Country wide.

Grouped interest rate into small medium and high of the previous application



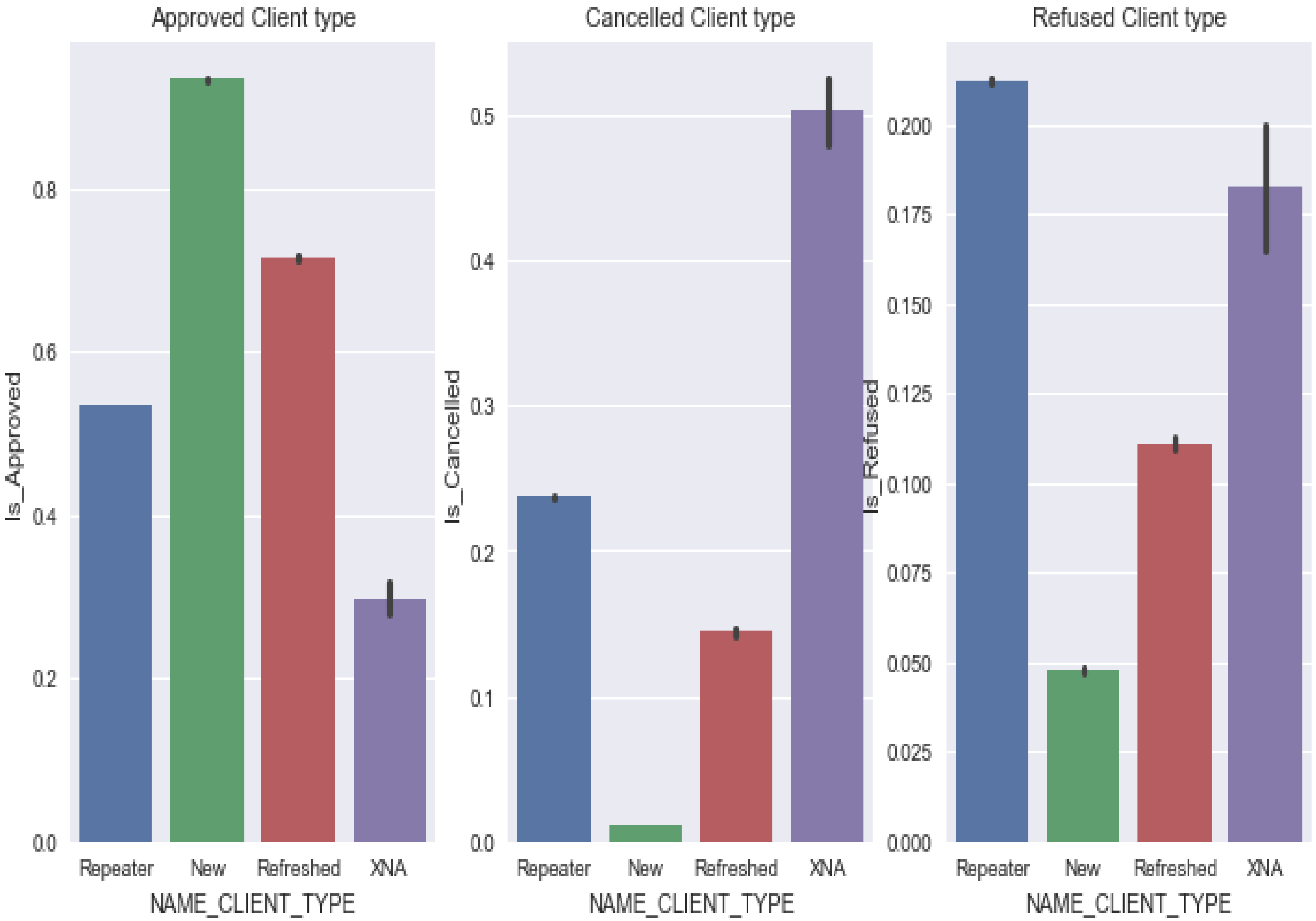
As we can see that there are high number of 'XNA' values which we can impute with 'middle' or 'high' category according to the factors which influence the interest rate and creating a machine learning model on top of that to classify the categories.

Detailed product combination of the previous application



The most common product combination being Cash followed by POS household with interest and POS mobile with interest for previous applications.

Bivariate Analysis



The clients who get approved are maximum from the new batch. So company is trying to improve it's customer acquisition rate by allowing large number of news clients.

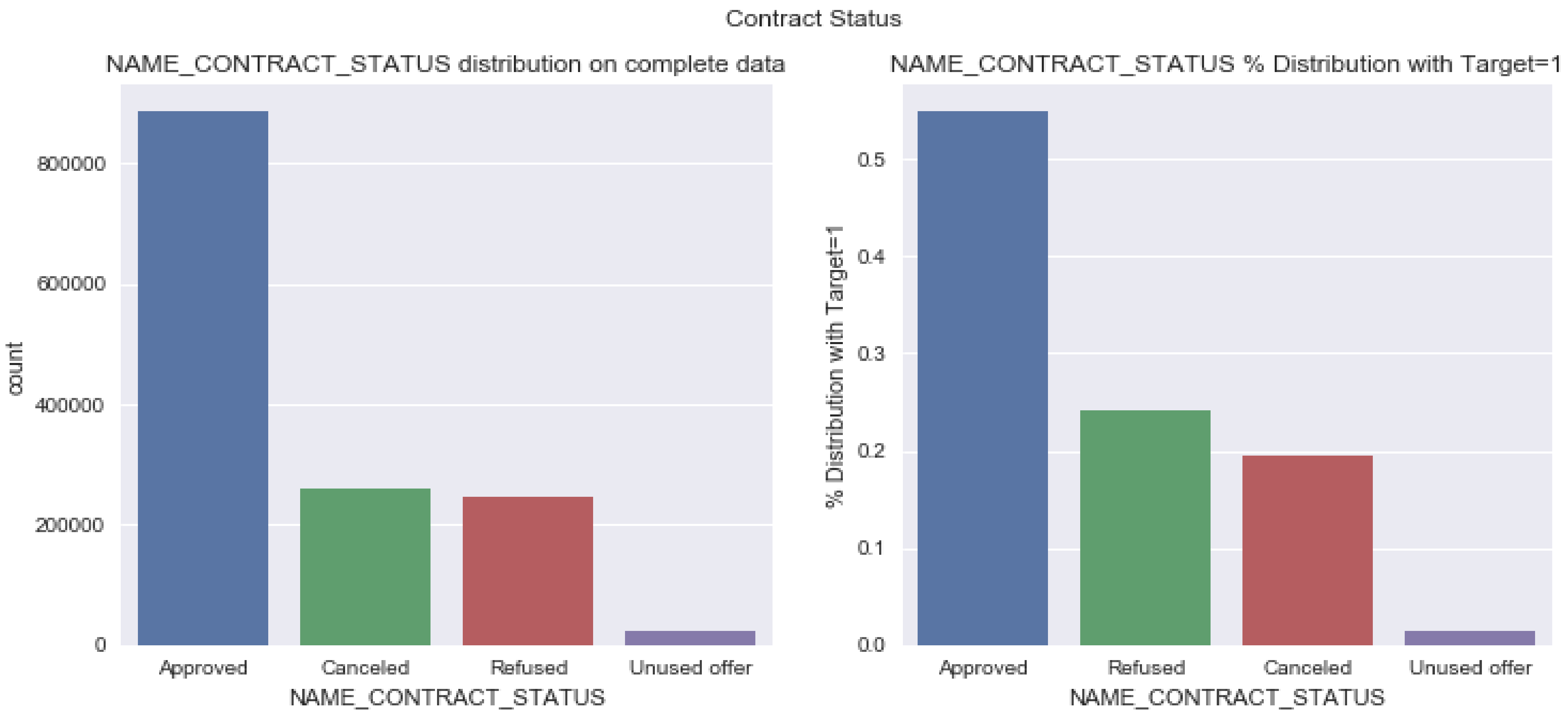
Most of the clients who has canceled their previous applications doesn't belong to any category.

The company has refused most of the clients from the repeater batch in the previous applications and also the clients who doesn't belong to any category.

Data analysis

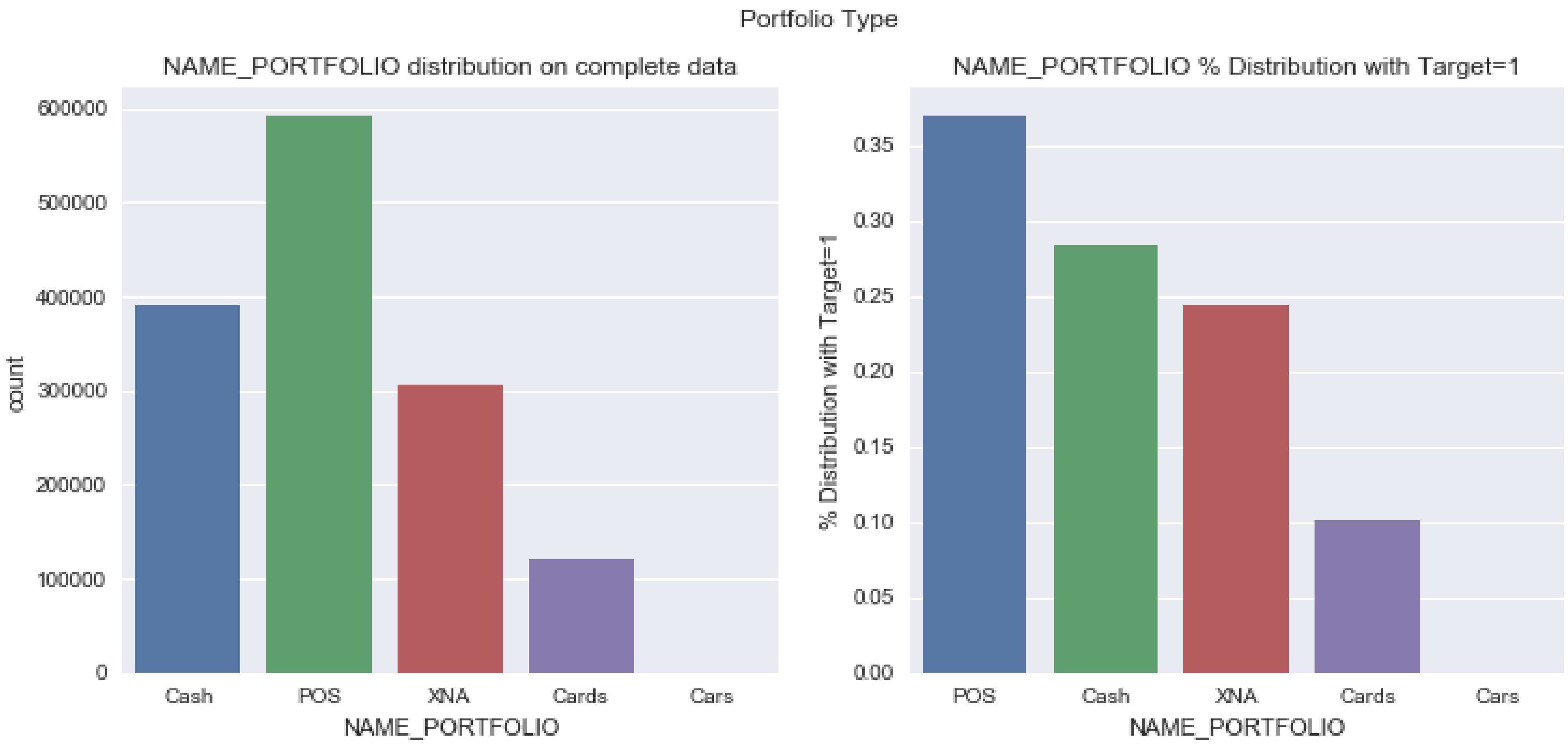
MERGED DATA

Contract Status(Overall and For defaulters)

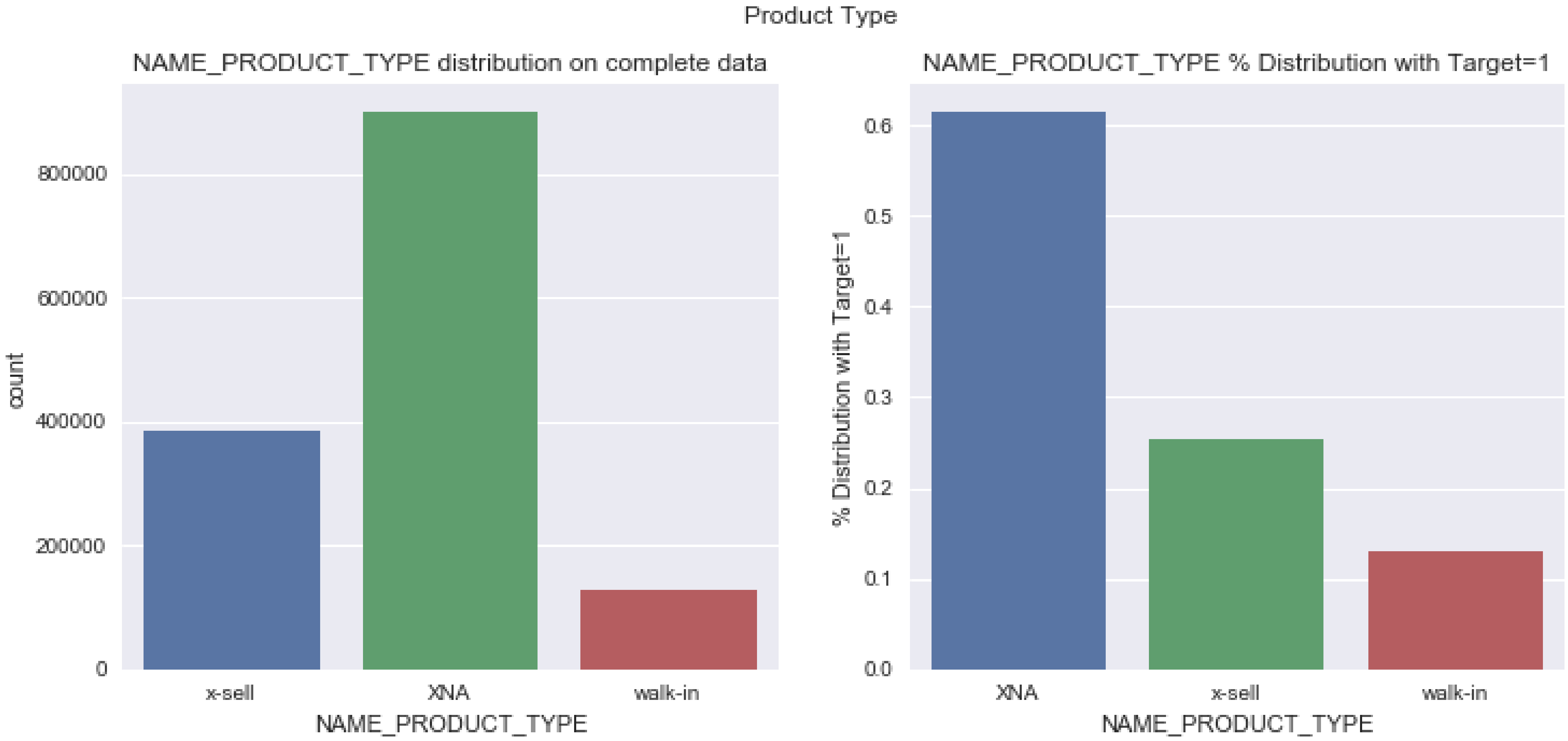


Client type(Overall and For Defaulters)

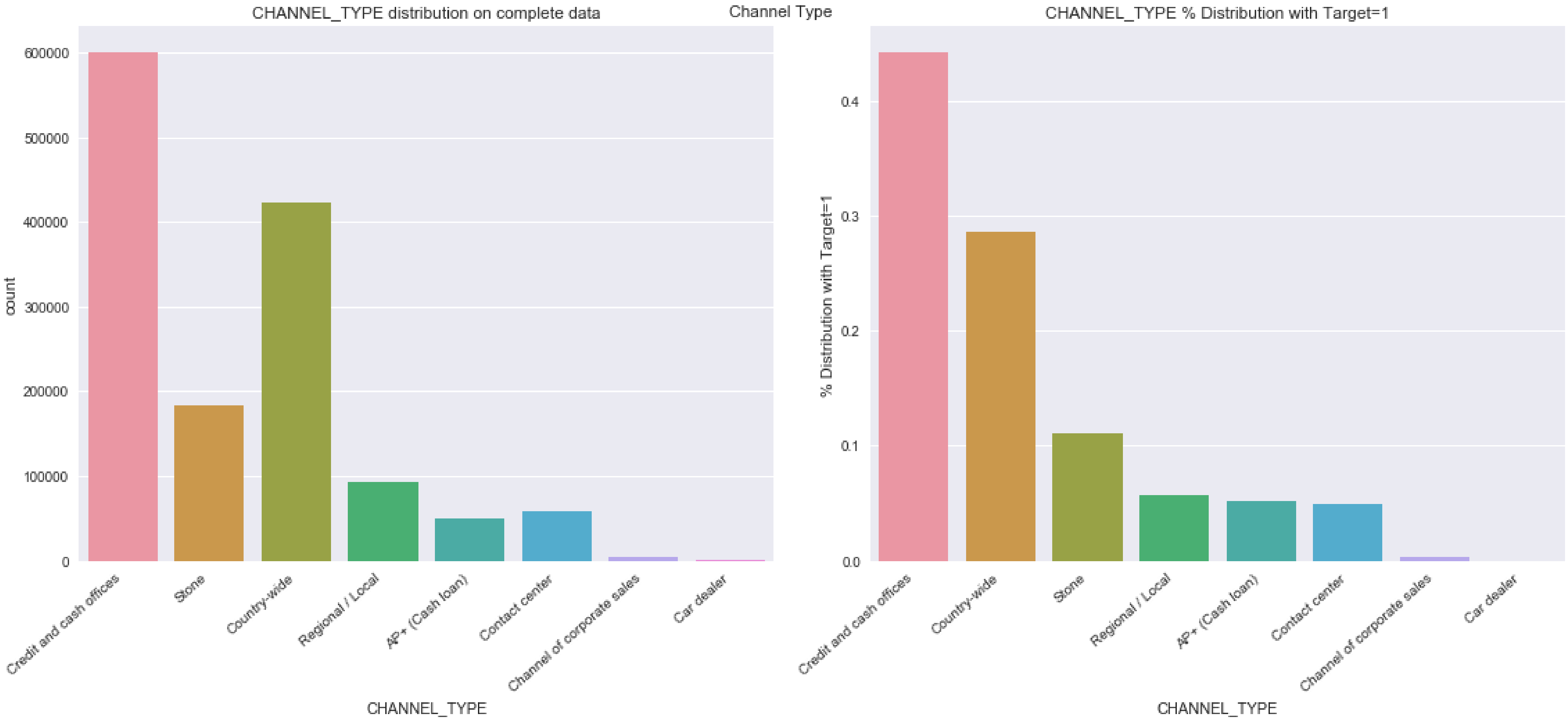
Product Portfolio(Overall and for Defaulters)



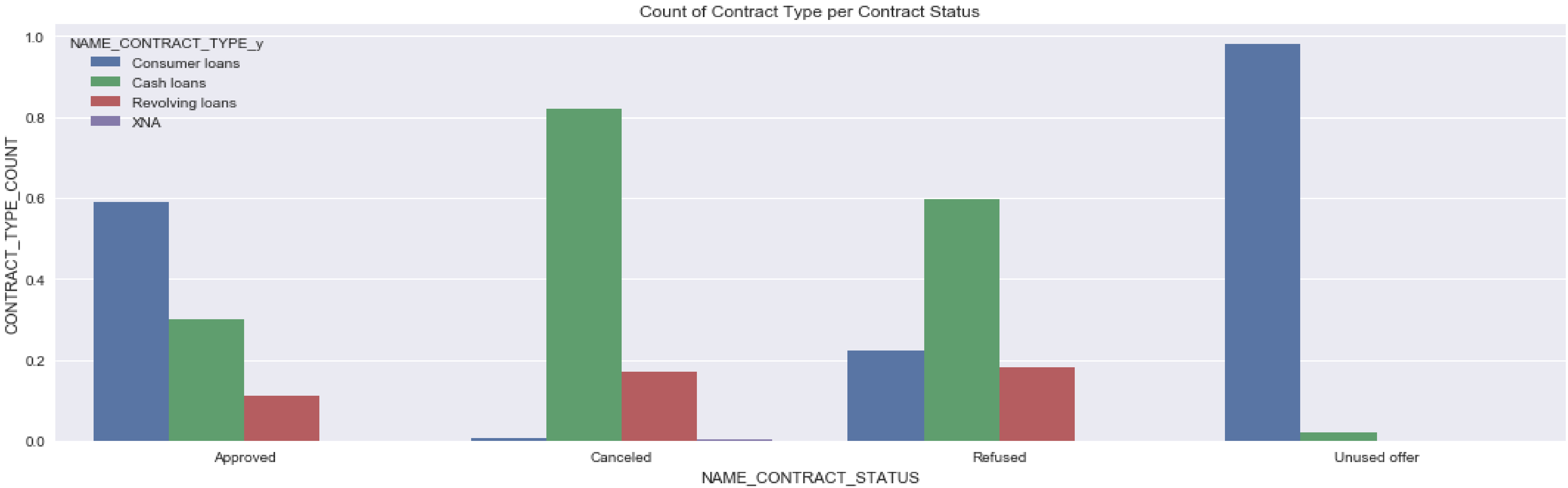
Product type(Overall and for Defaulters)



Channel type(Overall and for Defaulters)



Contract Type Count and Contract Status for Defaulters



TARGET = 1 (Default)

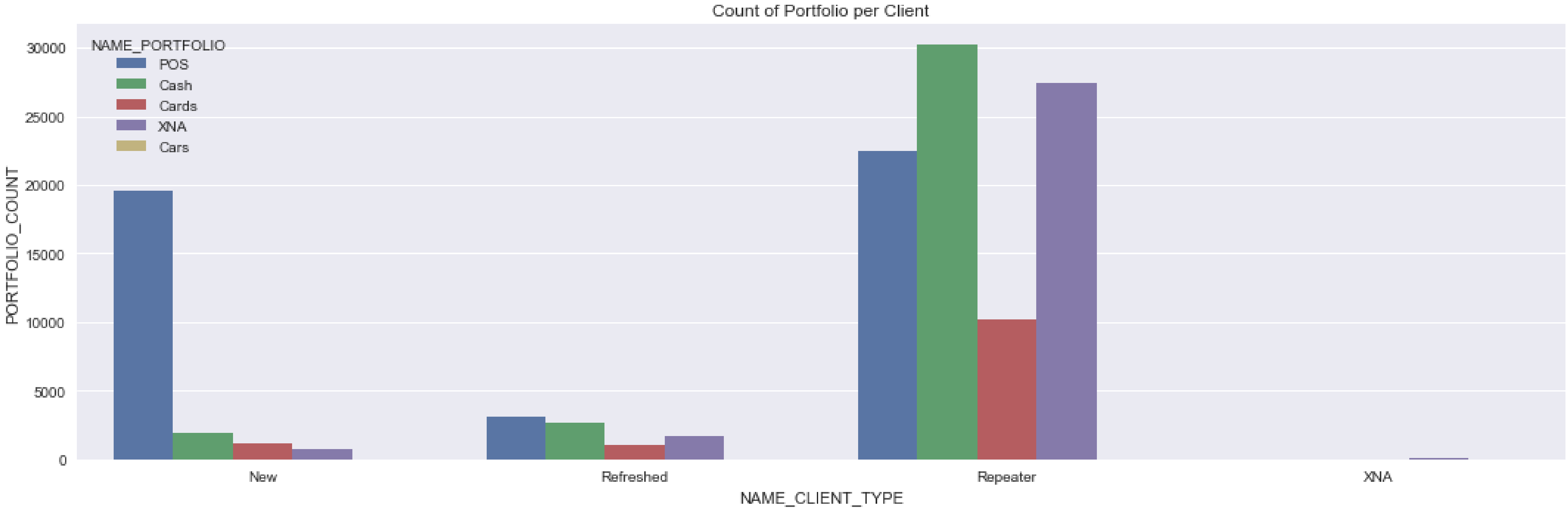
High number of Consumer loans are unused offers

Consumer loans are highly approved

Cash loans have high cancellation rate

Cash loans have high refusal rate

Product Portfolio count and Client Type for Defaulters

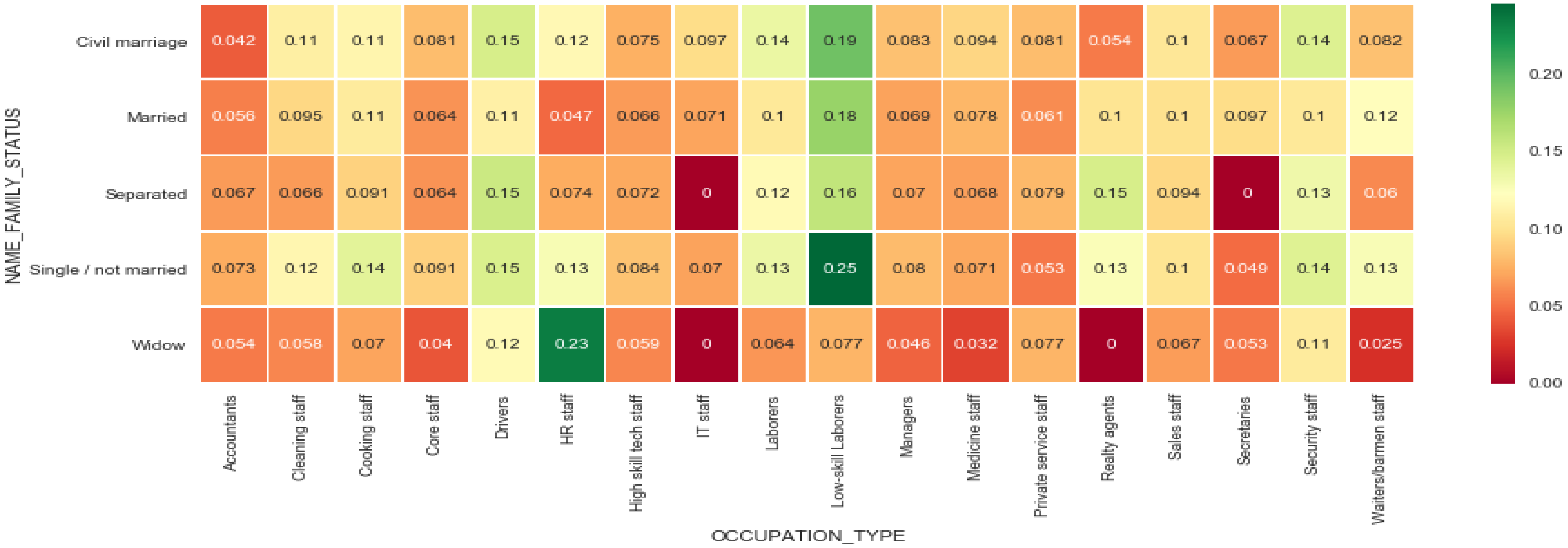


TARGET = 1 (Default)

POS type portfolio have high approval rate

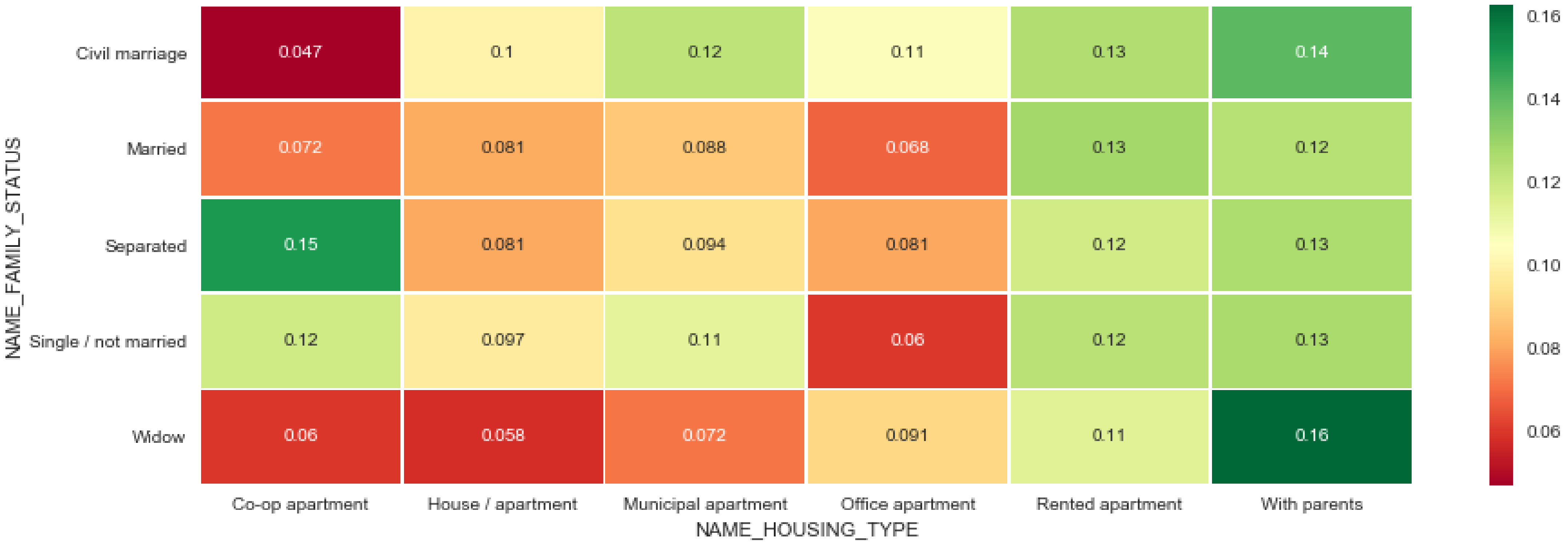
Cash type portfolio have high refusal rate

Correlation between Family Status and Occupation type for Defaulters



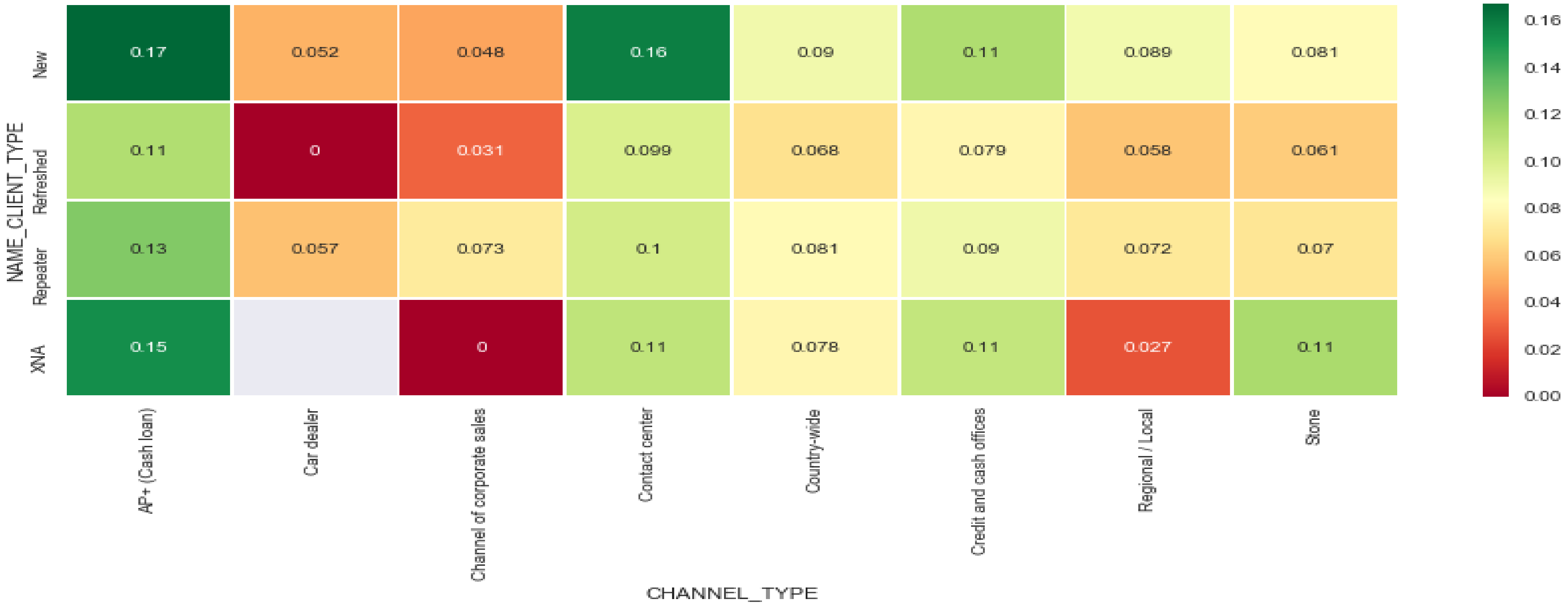
Highest correlation to be found among Low skill laborers who are not married and for HR and Widow

Family Status and Housing Type for Defaulters



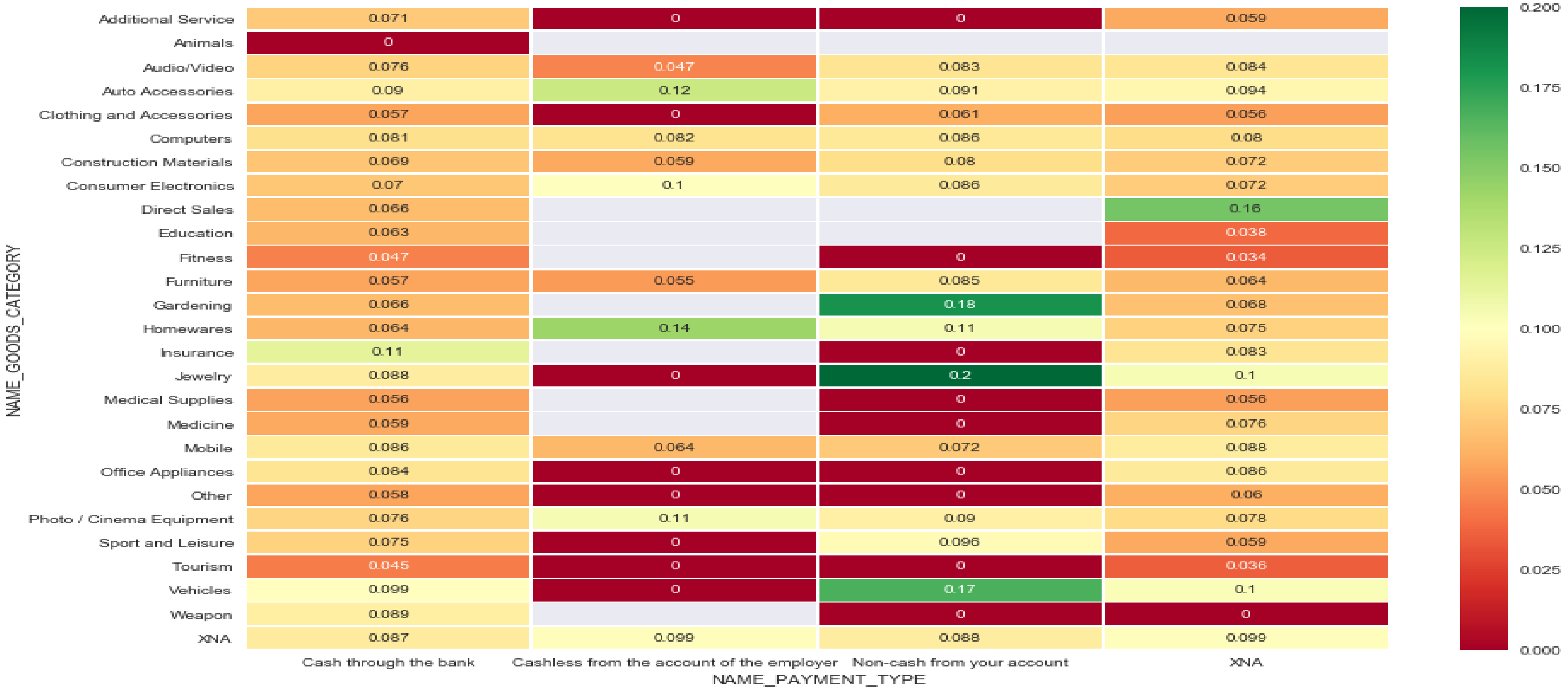
Highest correlation is between Not married people and who are staying with parents. Means teenagers.

Client Type and Channel Type for Defaulters



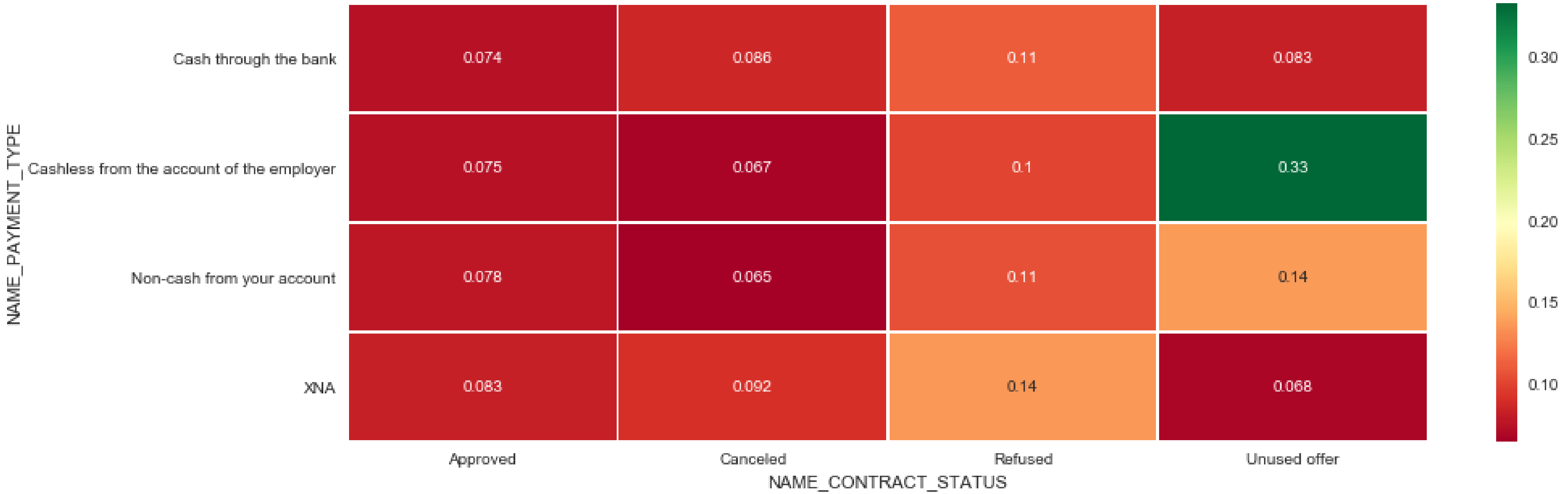
Highest correlation for New clients coming from AP+(Cash Loan) channels followed by New clients and Contact Center Channel

Goods Category and Payment type for Defaulters



Maximum correlation for jewelry and Non-cash from account type of payment method.

Contract Status and Payment type for defaulters



Only and very high correlation among Cashless from the account and Unused offer for defaulters.