EDA PROJECT

TAN HAI LING

BUSINESS UNDERSTANDING

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.

UNDERSTANDING DATA

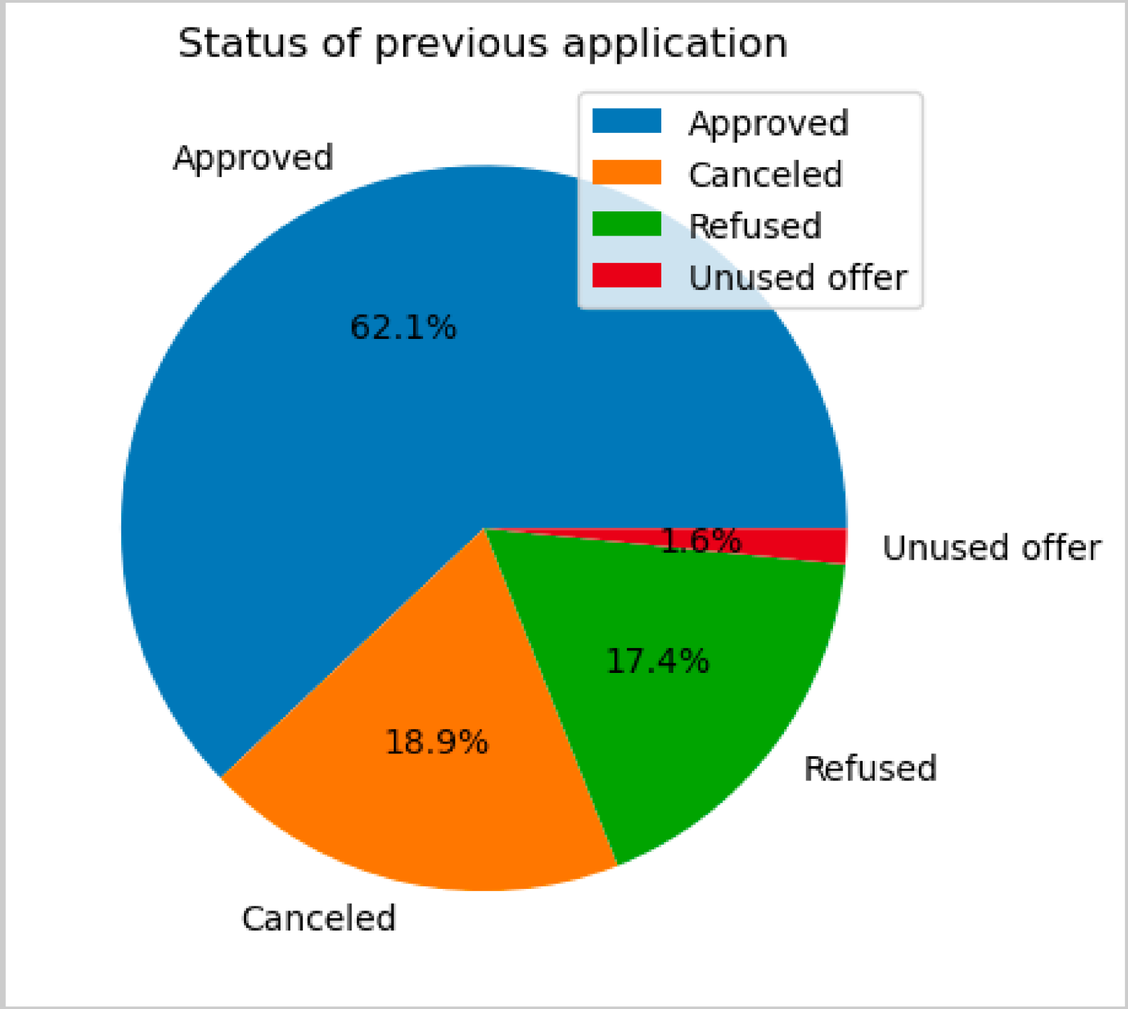
PREVIOUS DATA

TARGET VARIABLE

Findings

- The target variable indicates the status of the client's previous application, which are : Approved, Canceled, Refused and Unused Offer respectively.

- The data is rather imbalanced with 62.1% of applications being approved, 18.9% being canceled, 17.4% being refused and only 1.6% being unused offer.

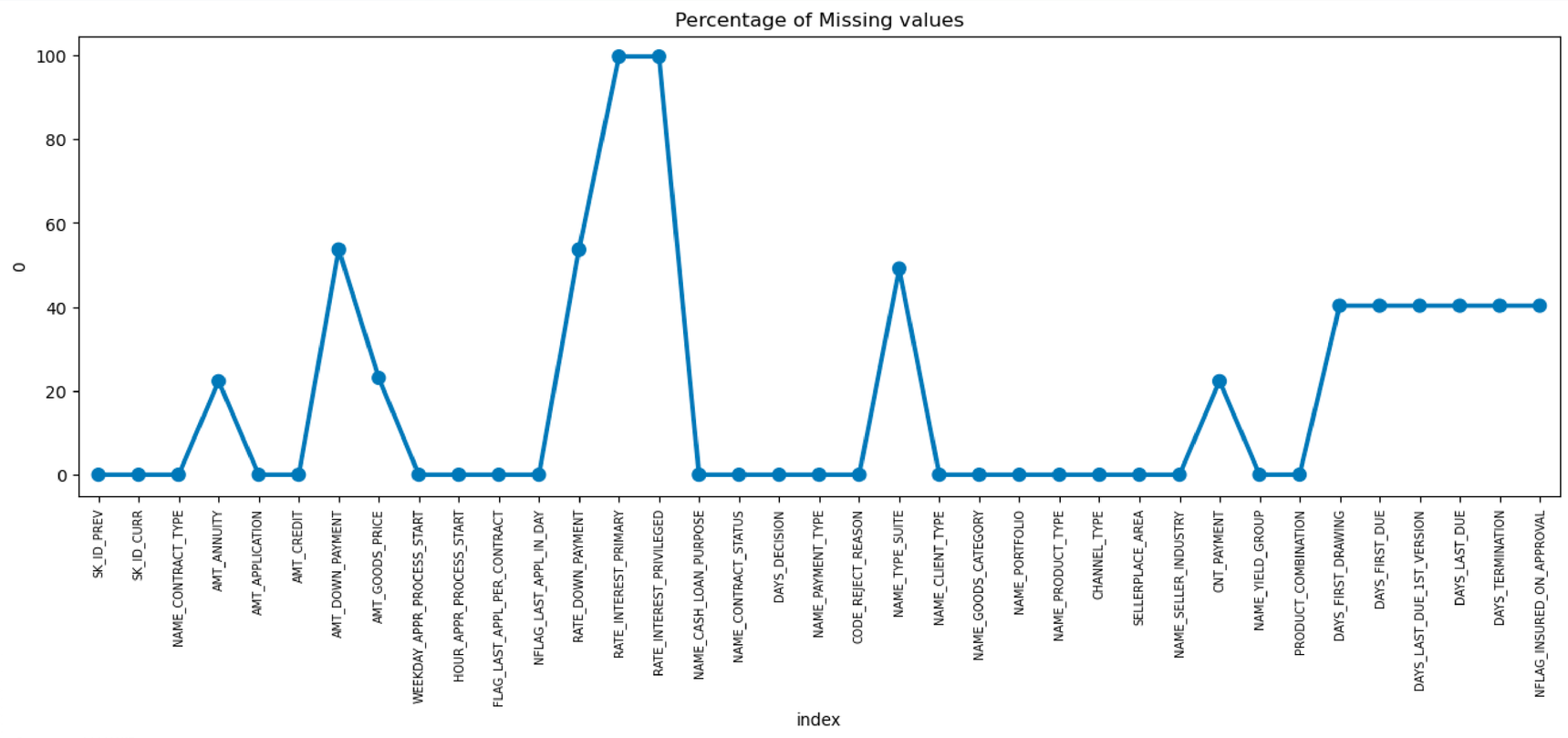


MISSING DATA

Findings

- The percentage of null values identified in the columns ranges from 22.3% to 99.6% (RATE_INTEREST_PRIMARY, RATE_INTEREST_PRIVILEGED)

- Nearly half of the columns contains null values in them.



DATA CLEANING

Findings

We noted that the following columns (**9 columns**) are not relevant to our EDA process as they're basically data for previous applications, hence we have further removed them in our data cleaning process. Those columns include:

(i) NFLAG_INSURED_ON_APPROVAL
(ii) SK_ID_PREV
(iii) SK_ID_CURR
(iv) FLAG_LAST_APPL_PER_CONTRACT
(v) NFLAG_LAST_APPL_IN_DAY
(vi) NAME_TYPE_SUITE
(vii) SELLERPLACE_AREA
(viii) WEEKDAY_APPR_PROCESS_START
(ix) HOUR_APPR_PROCESS_START'

Treatment for missing values

More than 40%

We have dropped the following columns: **AMT_DOWN_PAYMENT , RATE_DOWN_PAYMENT , RATE_INTEREST_PRIMARY, RATE_INTEREST_PRIVILEGED, DAYS_FIRST_DRAWING , DAYS_FIRST_DUE , DAYS_LAST_DUE_1ST_VERSION , DAYS_LAST_DUE , DAYS_TERMINATION** as columns with large amount of missing data wouldn't provide useful insights on the analysis.

Between 20%-40%

The 3 columns (**AMT_ANNUITY','AMT_GOODS_PRICE', 'CNT_PAYMENT**) that falls within this range have been replaced using the **mean** as these columns contains numerical data.

Below 20%

As the remaining columns with null values fall below 1%, hence suggest to leave for further processing.

UNIVARIATE ANALYSIS

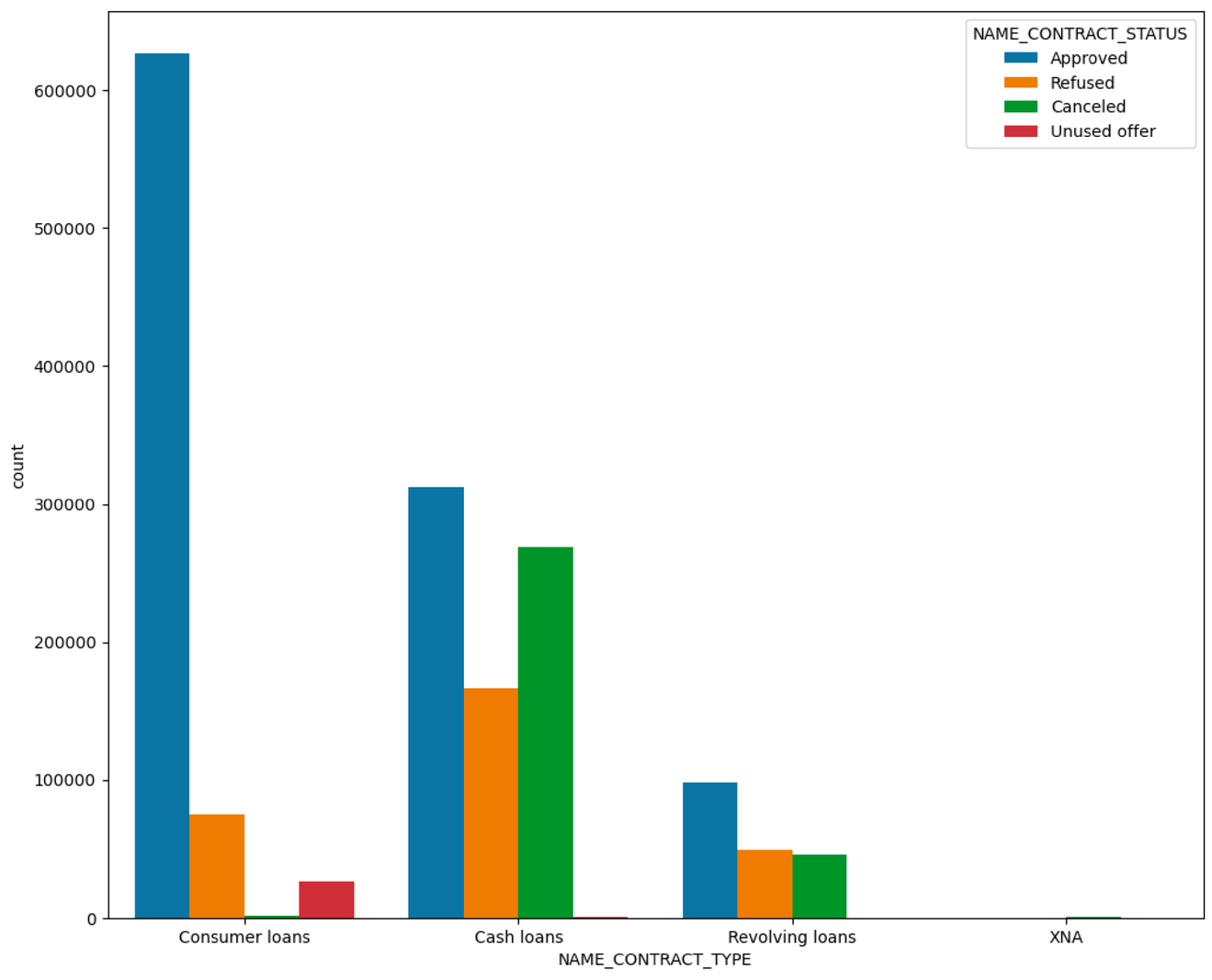
Findings

- We used univariate analysis to analyse categorical data. We identified the type of data in the columns by running '***.info()***' and '***.nunique()***'.

- By analysing single variables, the information that we can obtain is quite limited. However, we did get some very obvious insights for certain columns.

- Since there are so many columns being analysed, we have chosen the top 3 most informative ones to present here. Please refer to the following slides for further information.

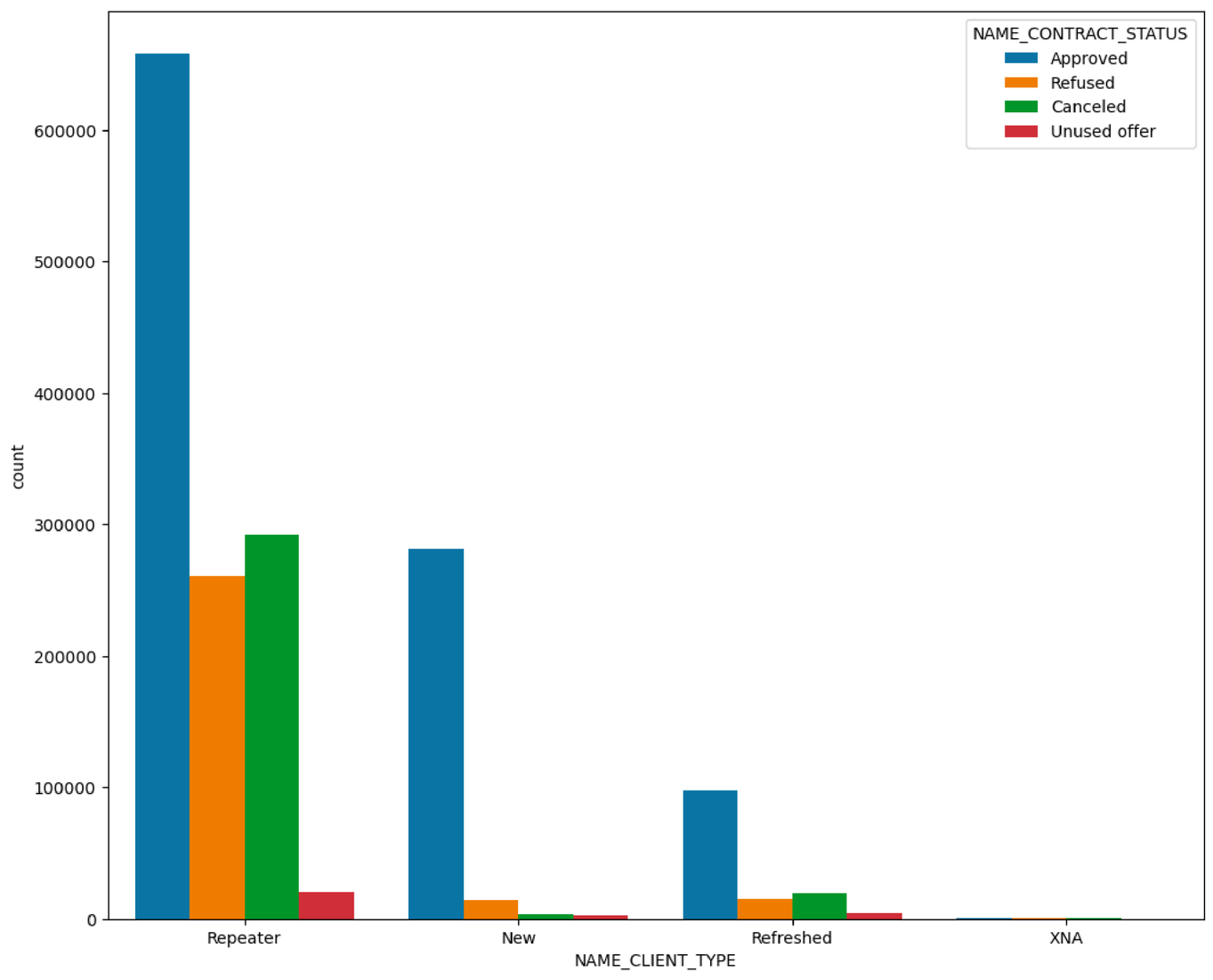
Contract type vs target variable



Findings

- Most previous applications were approved, with consumer loans being the most approved loans followed by cash loans.
- Most cash loans were either canceled or refused in aggregate, if compared to the number of approved cash loans.
- Little to no consumer loans were canceled, while unused offer only appeared under consumer loans.

Client Type vs target variable



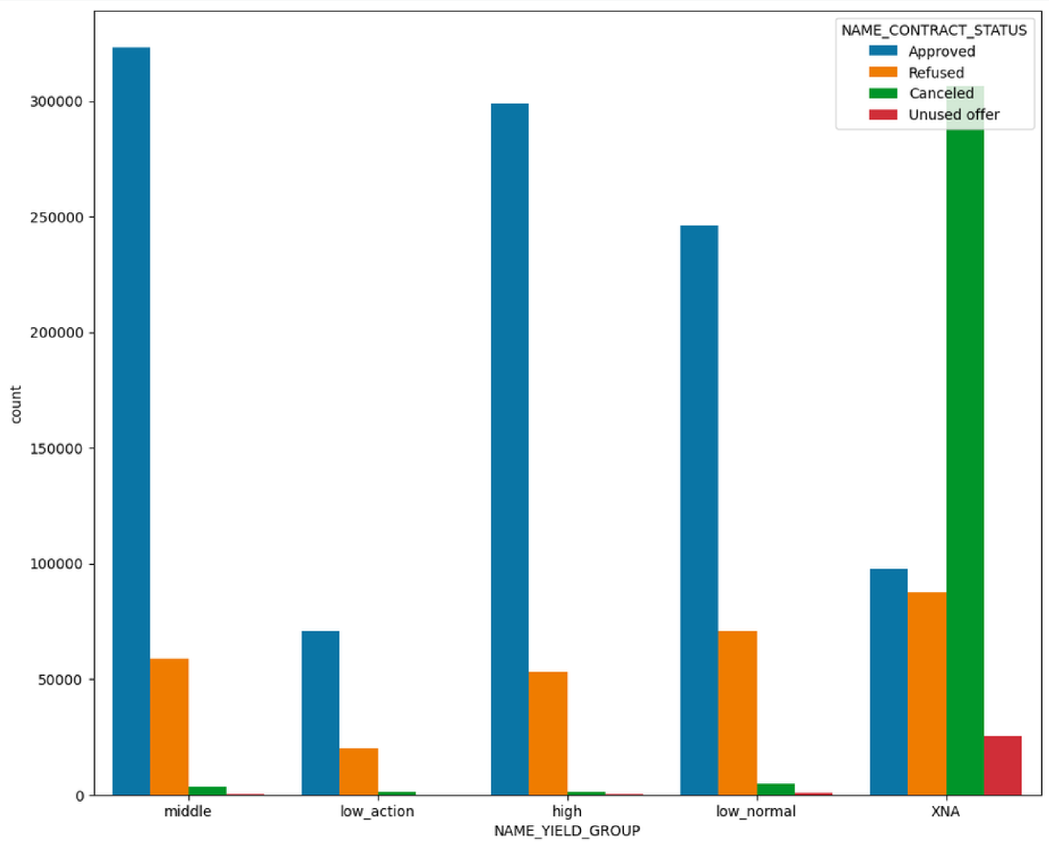
Findings

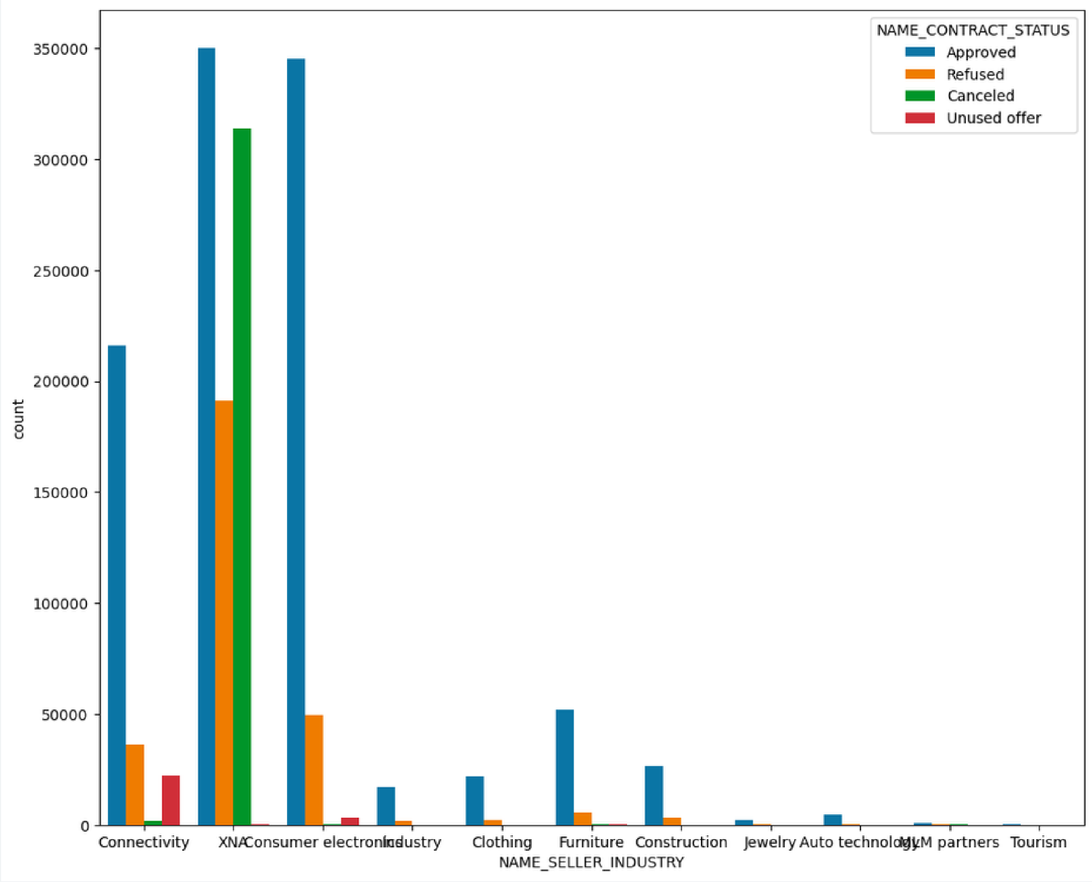
- Most clients are repeaters, followed by new and refreshed clients.
- Most previous applications were being approved, with repeater clients having the highest number followed by new clients.
- Refreshed clients has a higher number of canceled applications than refused loans, which is similar to repeater clients. New clients on the other hand have more applications getting refused than canceled.

XNA vs canceled loans

Findings

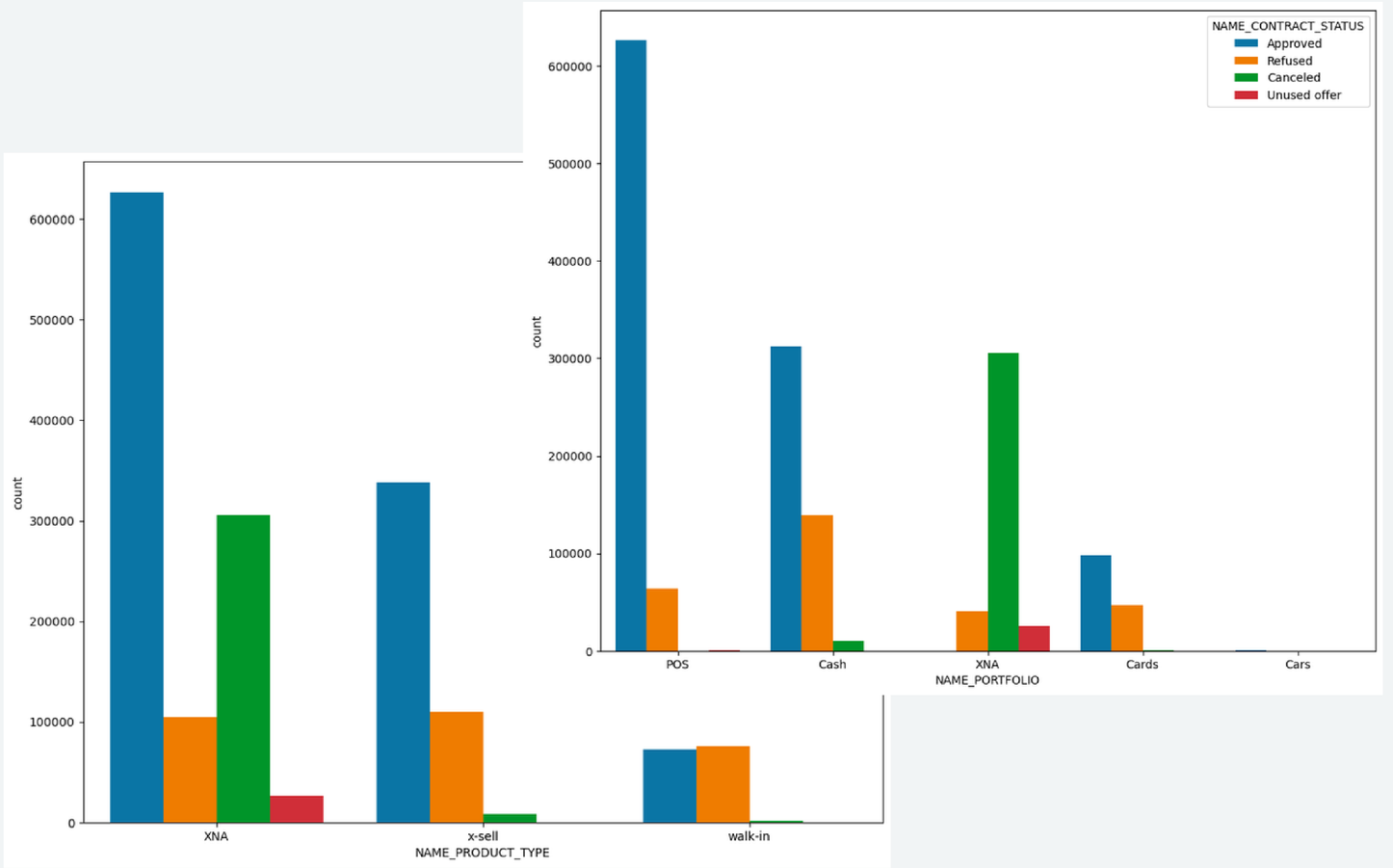
- When analysing these 5 columns : (**NAME_PAYMENT_TYPE, NAME_SELLER_INDUSTRY, NAME_YIELD_GROUP, NAME PORTFOLIO, NAME_PRODUCT_TYPE**), it has been noted that most of the canceled applications fall under XNA category.

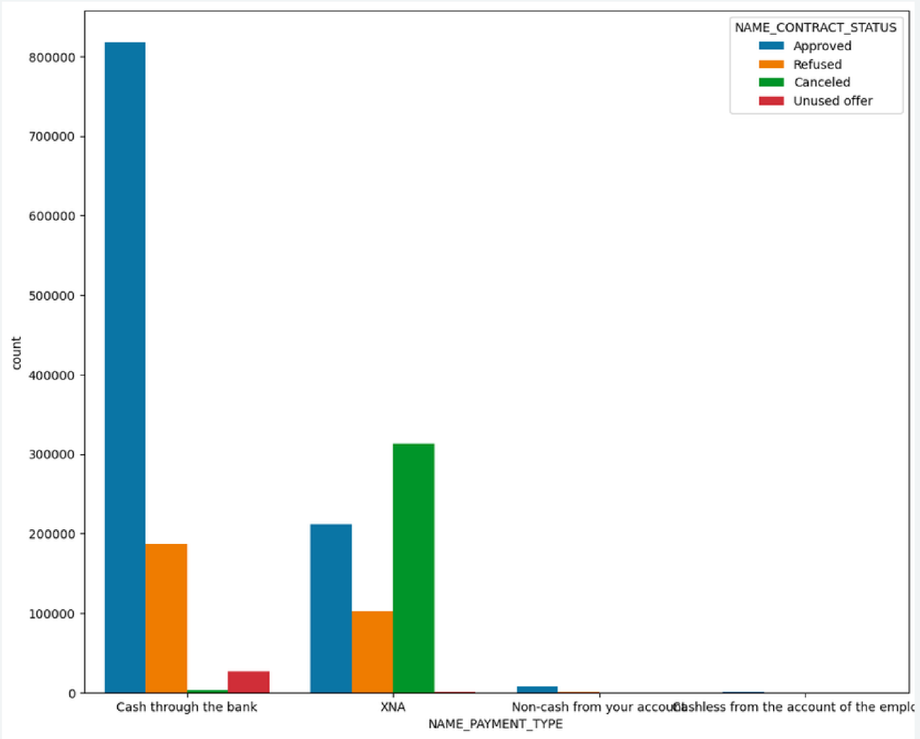




XNA vs canceled loans (Continued)

- XNA category also sees the 2nd highest amount of refused applications as compared to other categories, namely for columns **NAME_PAYMENT_TYPE** and **NAME_PRODUCT_TYPE**.





NUMERICAL ANALYSIS

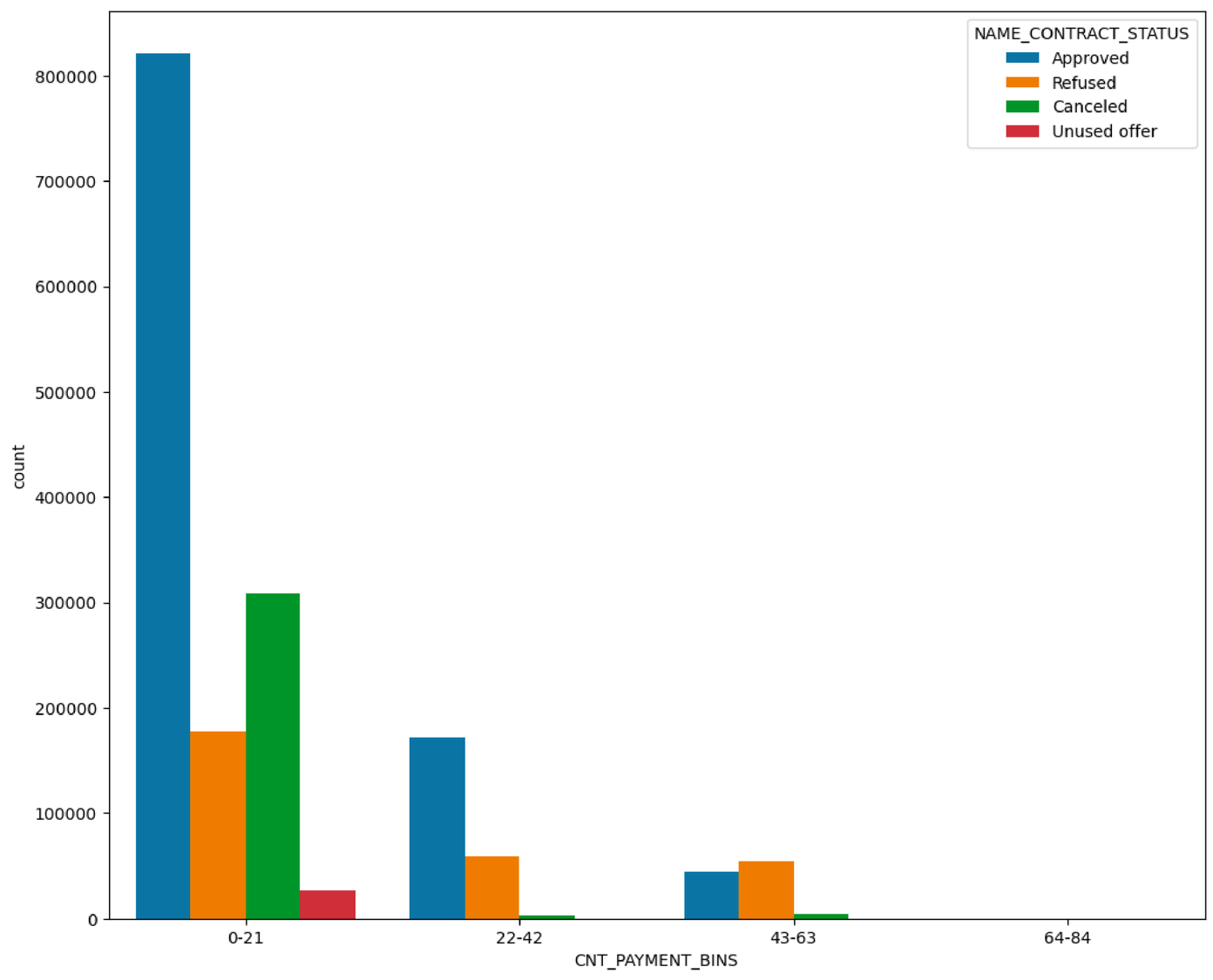
Findings

- We used numerical analysis to analyse numerical data. We identified the type of data in the columns by running '***.info()***' and '***.nunique()***'.

- We have utilised the boxplot function to visualise the numerical columns as they contain too many outliers.

- Subsequent to using the boxplot function, we filtered the columns based on 5th-95th percentile while running the KDE plots to obtain a better visualisation as the remaining data are outliers.

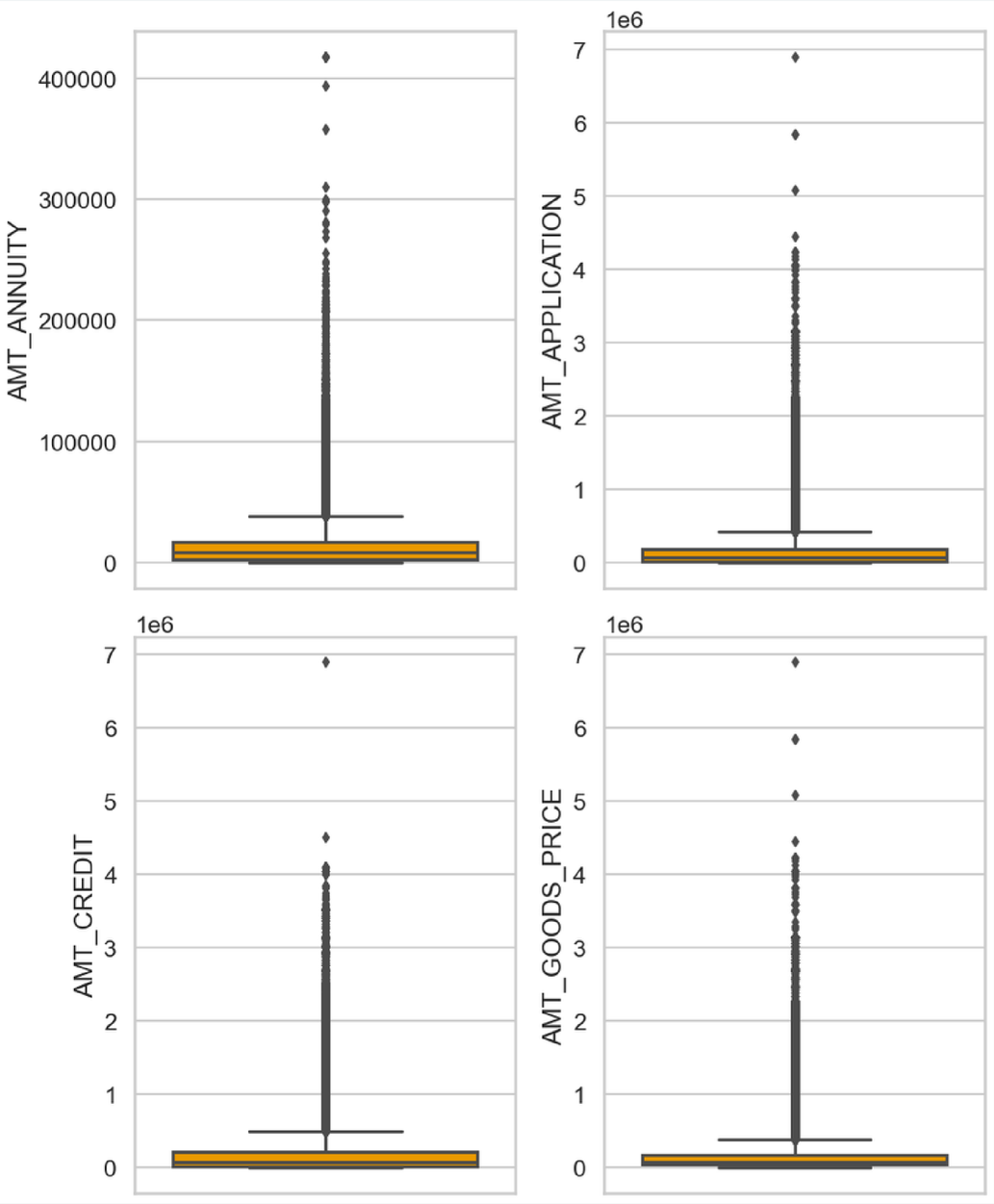
Term of previous credit at application of the previous application



Findings

- When we look at the previous application's credit term data, it has shown that majority of the previous credit application's term falls within the 0-21 month bucket
- There were more canceled applications as compared to refused applications for 0-21 month bucket.
- Little to zero applications fall under the 64-84 month bucket.

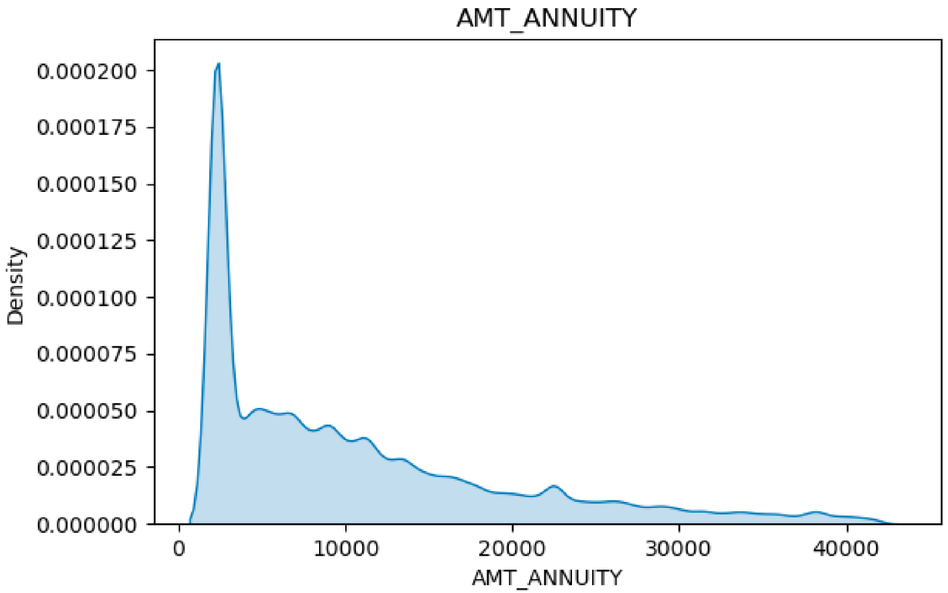
Boxplot for numerical columns

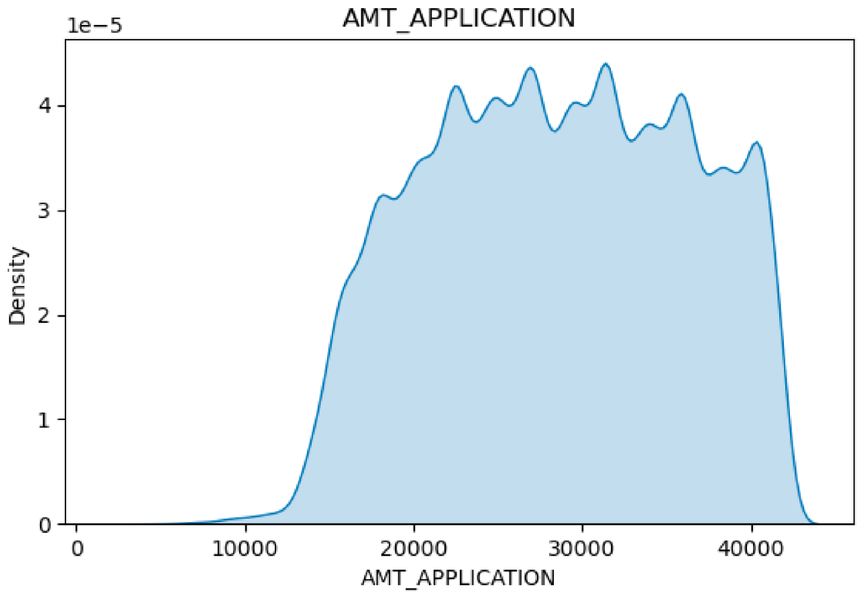


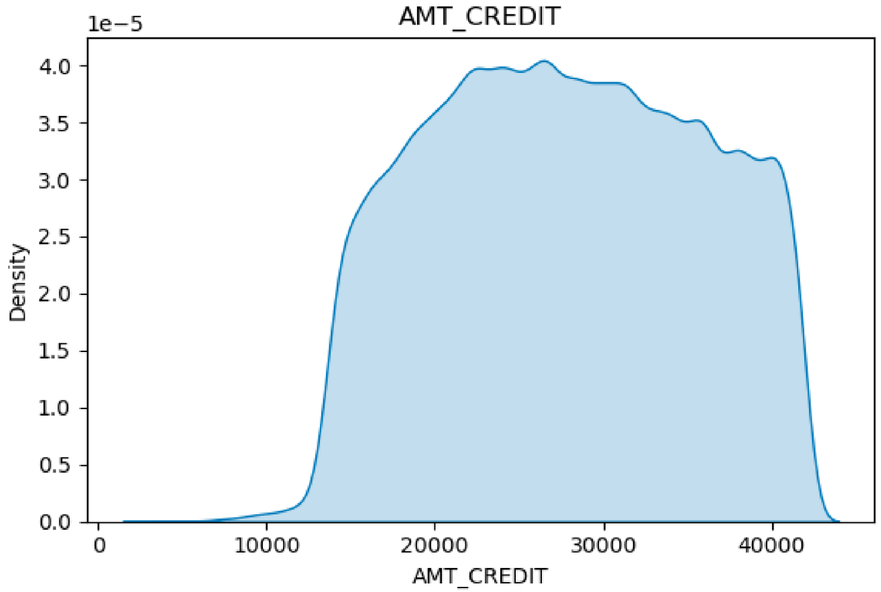
Findings

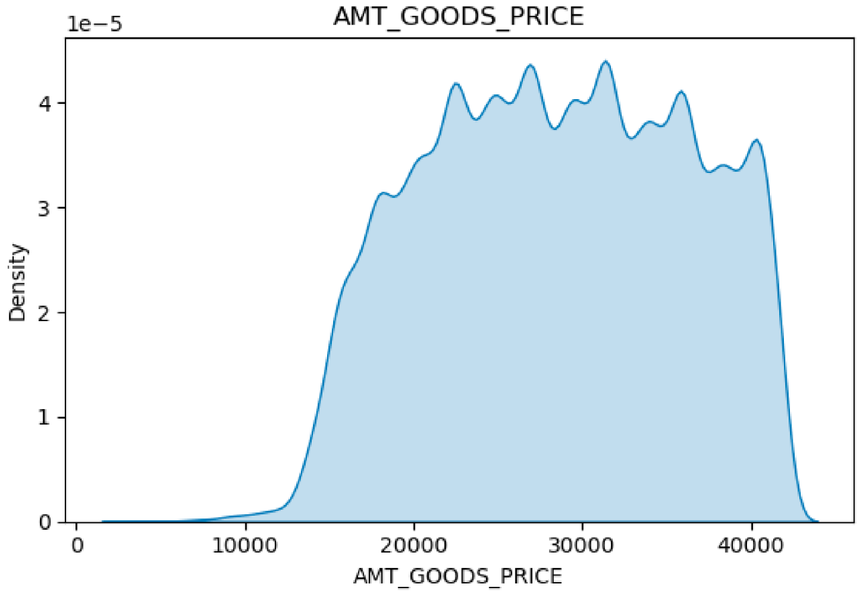
- We noted that most numerical columns contain large amount of outliers. Hence, we have utilised the boxplot function to visualise how the outlier values are being spread out.
- The columns affected are : **AMT_ANNUITY , AMT_APPLICATION , AMT_CREDIT , AMT_GOODS_PRICE** respectively.
- Without using this function, the KDE plot visualisations look rather imbalanced and no useful data can be drawn from the plots.

KDE plot for numerical columns









Findings

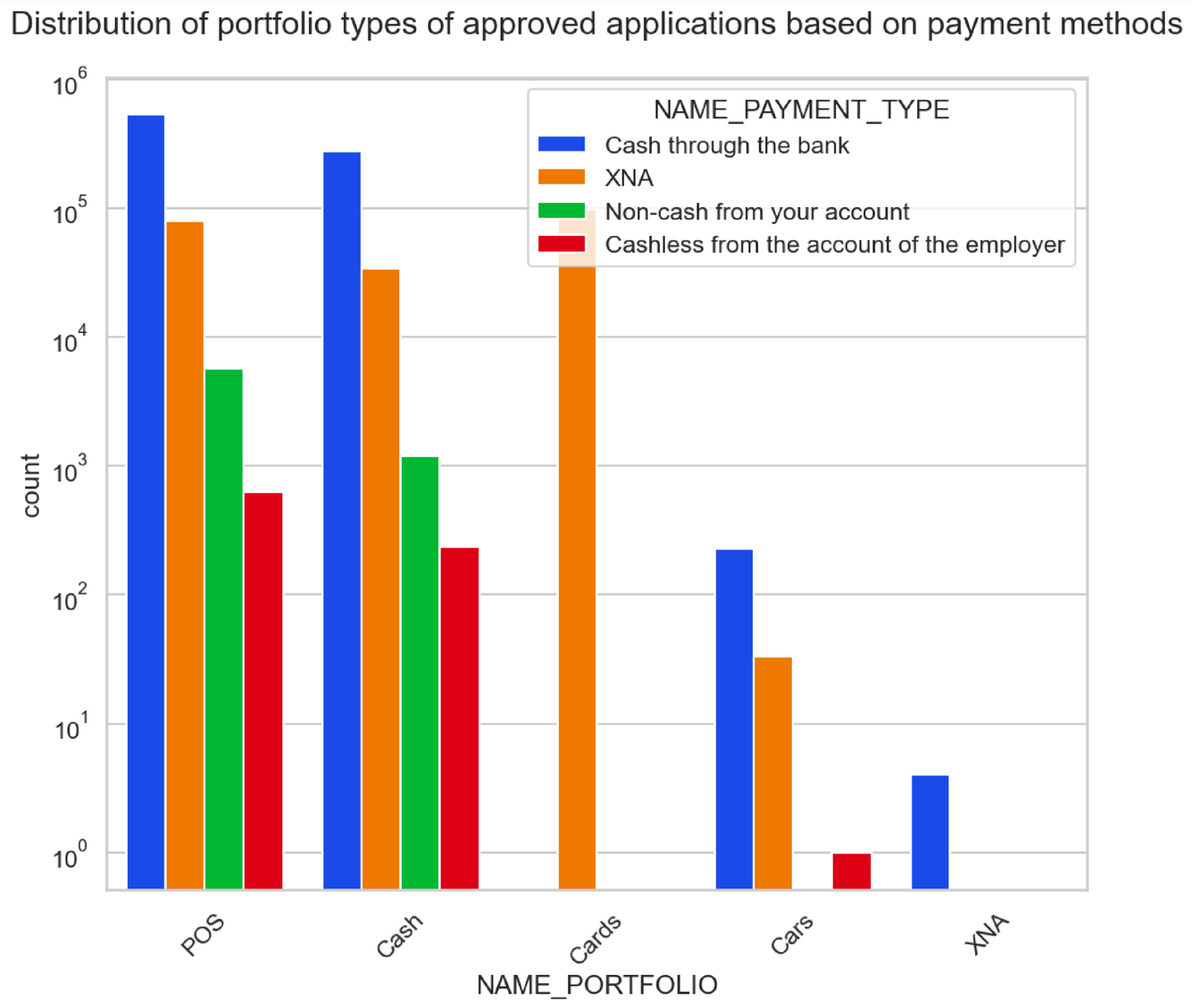
- This is how the KDE plots looks like for the numerical columns after filtering the columns by using the 5th & 95th percentile
- The distribution looks more normalised after filtering out the outliers.

BIVARIATE ANALYSIS

Findings

- We used a heat-map to run the correlation between all categories using dummy data in the multivariate analysis.
- We have inspected the correlation for the respective target variable columns (status of previous application), namely Approved, Canceled, Refused and Unused Offer.
- However, as the data shown in the heat-map is too large, we will only run 1-2 bivariate analysis for each target variable category based on the highest/ lowest correlation identified for each category in the heat-map.

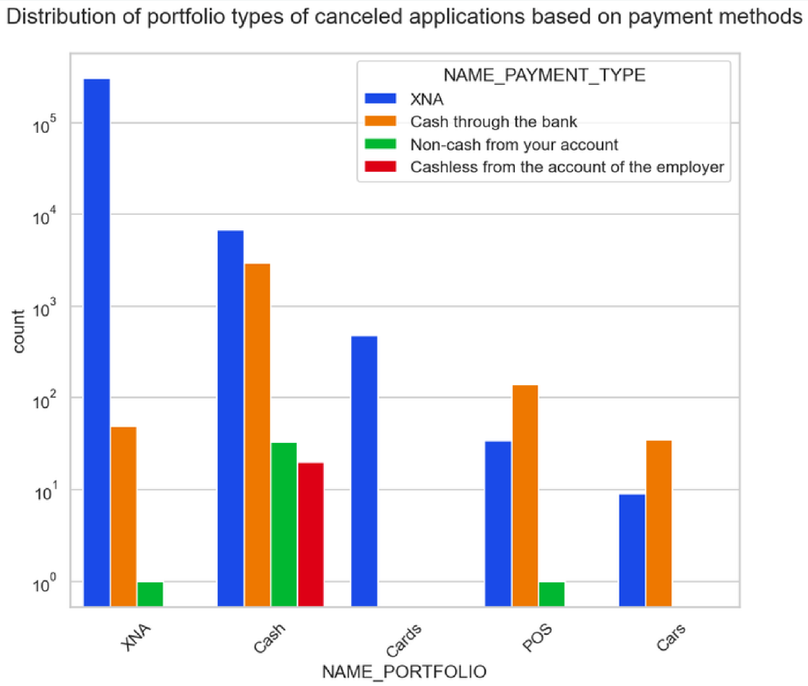
Approved applications

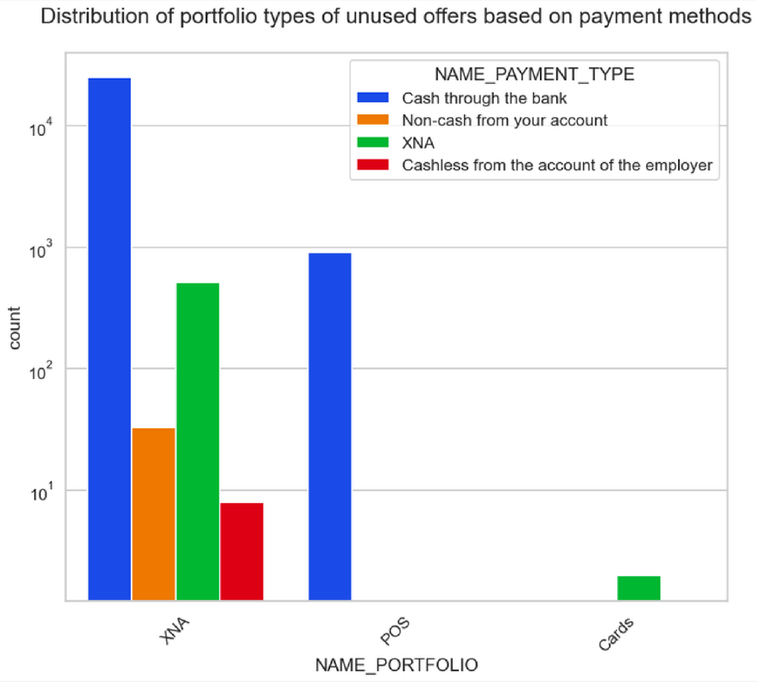


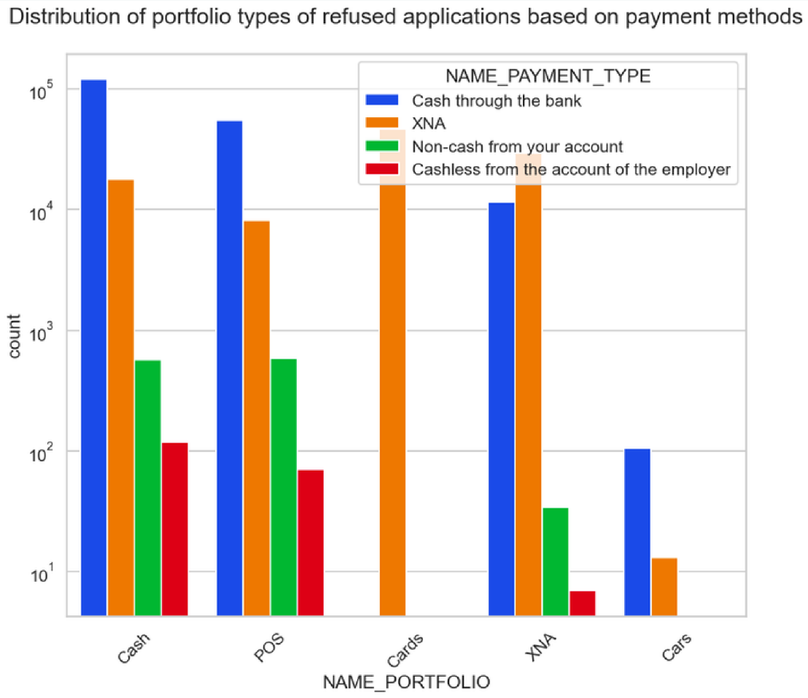
Findings

- According to the heat-map, approved applications has the lowest correlation with **NAME_PORTFOLIO: XNA,** which can be seen from the histogram on the left whereby there is only 1 payment type available for XNA.
- Most approved applications fall under the POS and Cash portfolio.
- Cards portfolio only consists on 1 payment type, which is XNA.
- Cars portfolio does not consist any payment type from "non-cash from your account".

NAME_PORTFOLIO vs other application status



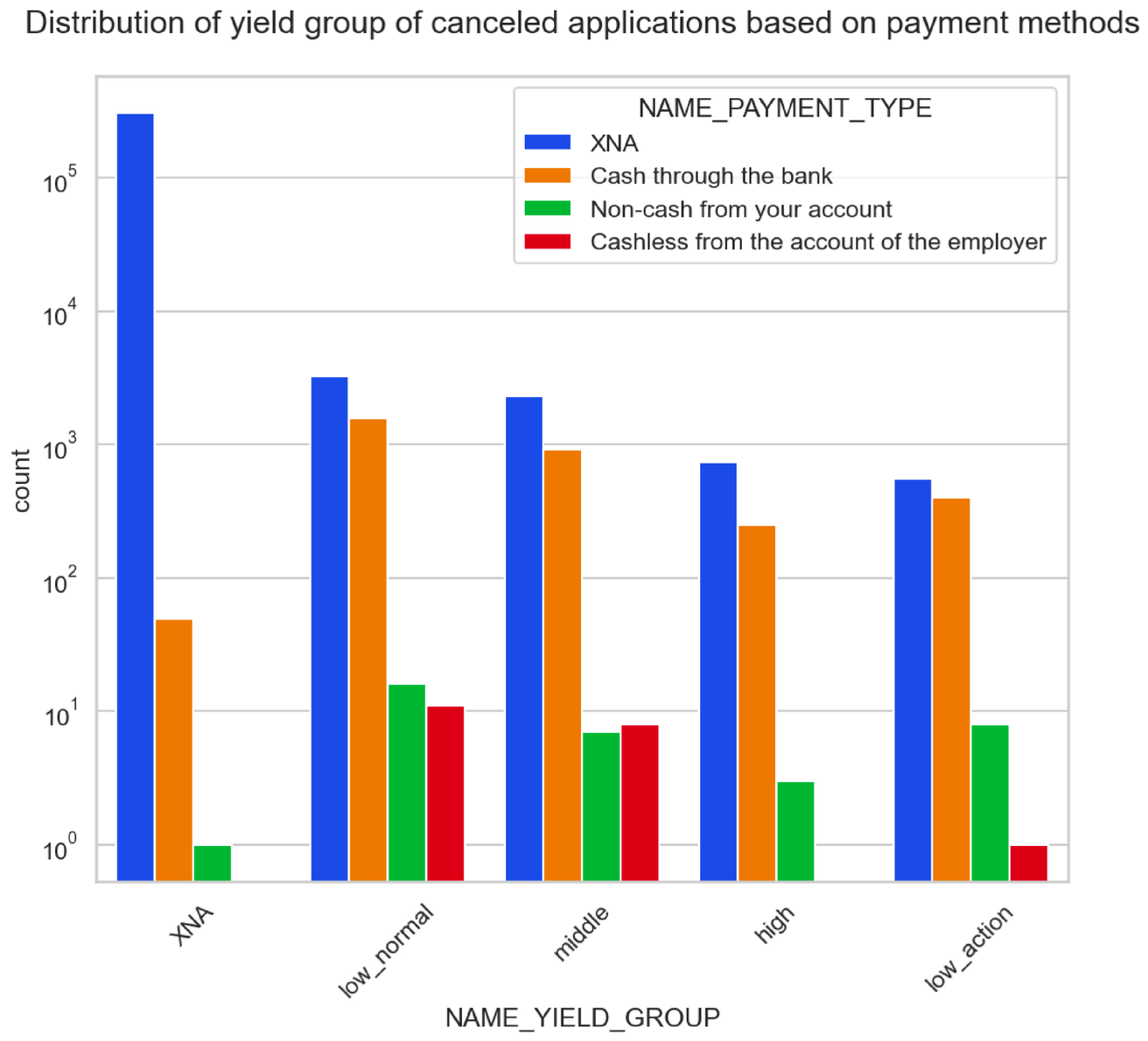




Findings

- Canceled and unused offers has the highest distribution for portfolio XNA; while for refused applications the highest type of portfolio distribution are for Cash and POS.
- Canceled and refused applications consists of all 5 types of portfolio, while unused offer only consisted of 3.

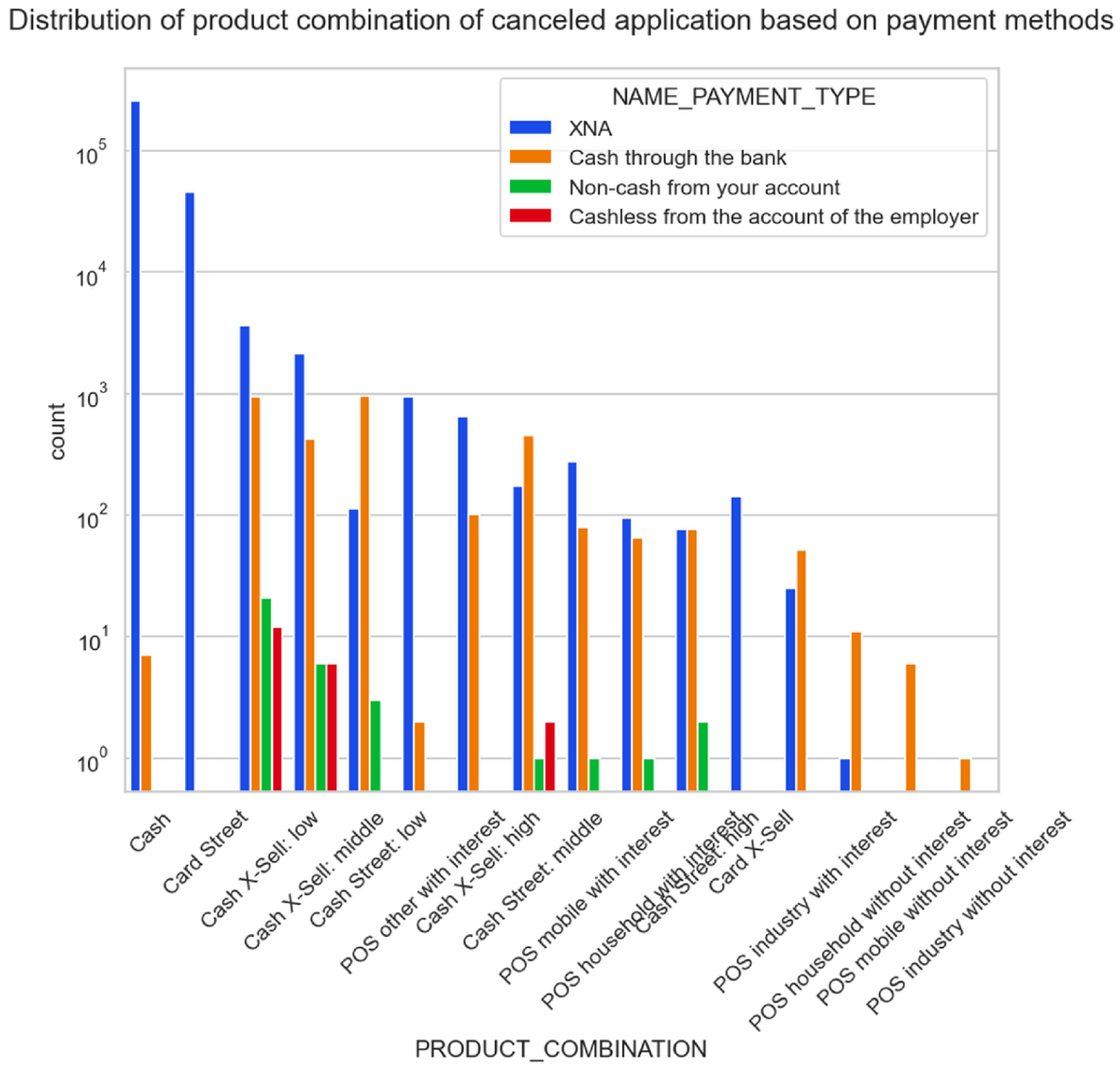
Canceled applications



Findings

- According to the heat-map, canceled applications has high correlation with **NAME_YIELD_GROUP: XNA,** which can be seen from the histogram on the left whereby XNA tops as the payment type for XNA yield group.
- There are 5 different yield group for the previously canceled applications.
- Except for XNA yield group which does not consist of payment type "cashless from the account of the employer", the other yield type consists of all 4 payment types.

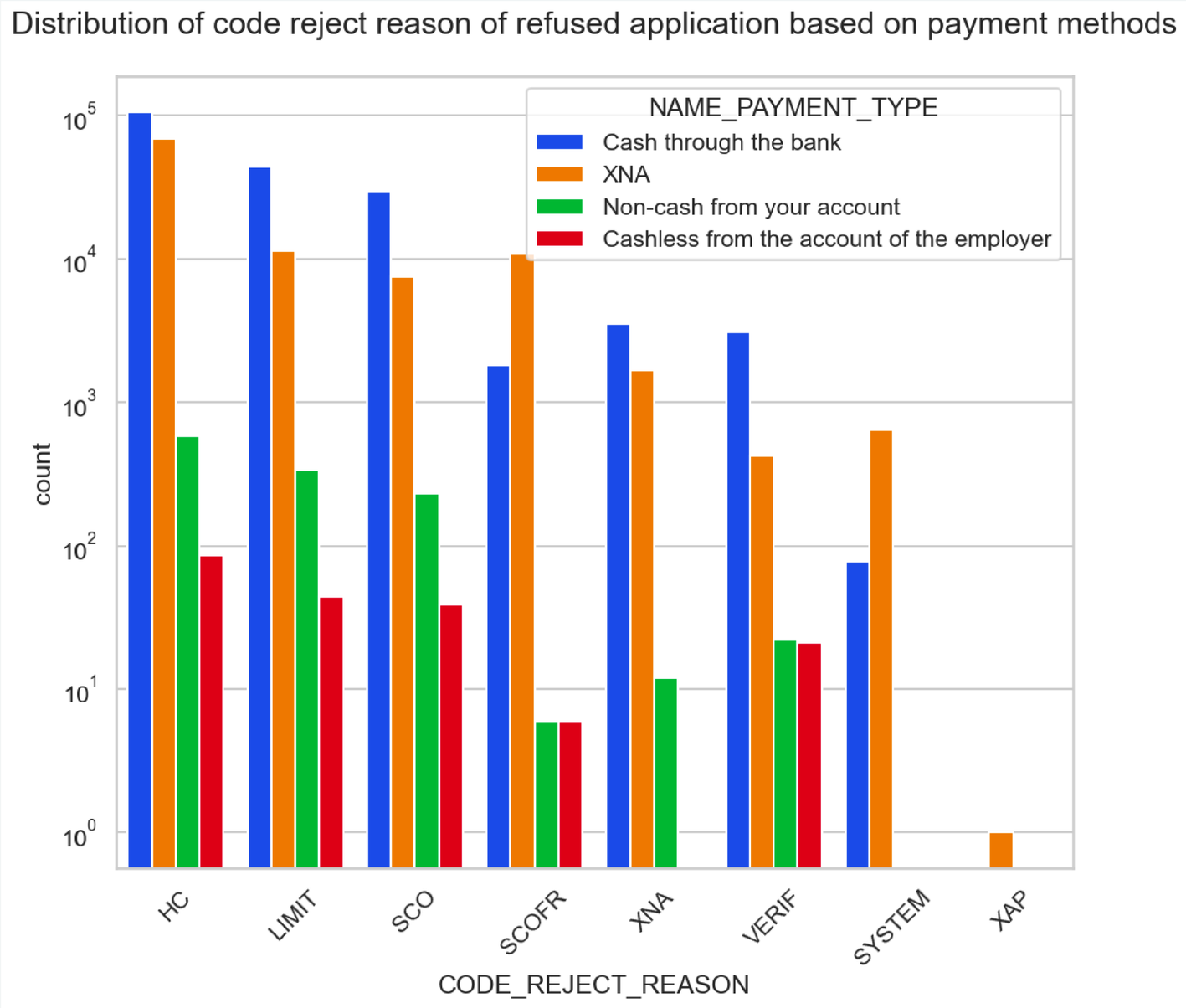
Canceled applications (Continued)



Findings

- According to the heat-map, canceled applications has high correlation with **PRODUCT_COMBINATION,** with cash as the highest correlated category.
- Most categories consists of payment types "cash through bank", except for "Card Street".
- XNA is the most distinct payment type for the product combinations of canceled applications, which is mainly seen from "Cash" and "Card Street".

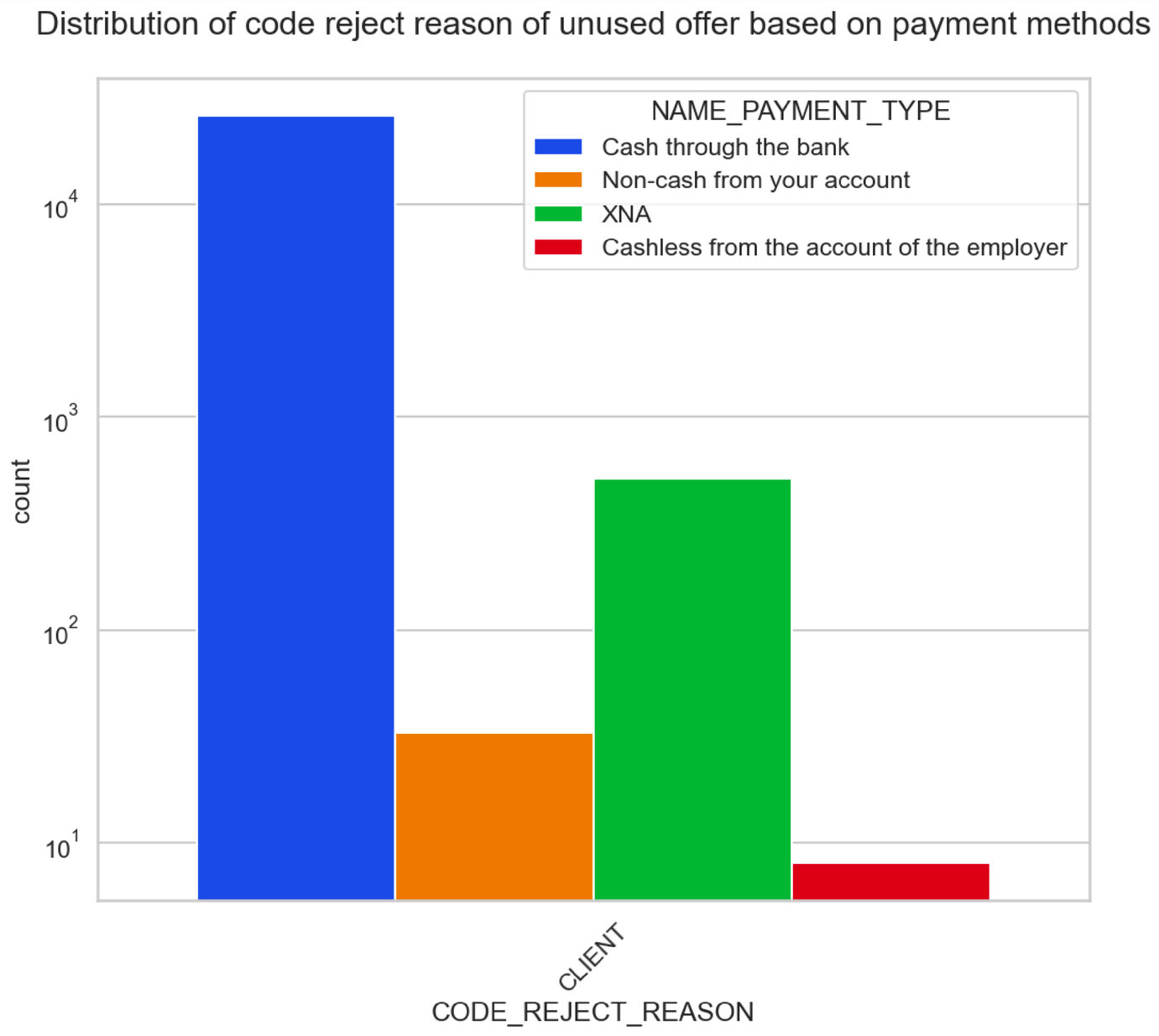
Refused applications



Findings

- According to the heat-map generated, refused applications has seen a high correlation between **CODE_REJECT_REASON**.
- The 2 main categories with high correlation are "**HC**" and "**limit**" respectively.
- Payment type XNA is the highest payment type for categories SCOFR, SYSTEM and XAP, while cash through bank is the main payment type for the remaining categories.

Unused offers



Findings

- According to the heat-map, unused offers for previous applications has seen a rather high correlation with column **CODE_REJECT_REASON: CLIENT**, which can be seen from the histogram on the left as "**CLIENT**" is the only reason for unused offers being rejected.

- Cash through payment tops as the highest payment type chosen, followed by XNA, non-cash from your account and cashless from the account of the employer.

SUMMARY
PREVIOUS DATA

PREVIOUS DATA

- Most refused and canceled loans fall are cash loans
- Most approved loans fall under consumer loans
- Most client types are repeaters, with majority of applications approved
- The XNA category for these 5 columns (NAME_PAYMENT_TYPE, NAME_SELLER_INDUSTRY, NAME_YIELD_GROUP, NAME PORTFOLIO, NAME_PRODUCT_TYPE) consisted the highest number of canceled applications, while other columns normally have "approved applications" as the highest choice of application status.
- Too many outliers for numerical data, hence some filtering is needed (using 5th-95th percentile)
- POS and Cash are the highest distributed portfolio categories for approved and refused applications; XNA for Canceled and unused offers.
- The only reason for unused offers being rejected is "Client" ; top 3 reasons for refused applications are HC, LIMIT, SCO.