EDA

LOAN DATA ANALYSIS

TO THI ANH TUYET

Nov 2024

Business Background

A consumer finance company provides various types of loans to customers. The company needs to analyse the pattern in the **application data & credit history** of loan applicants to **evaluate the loan approval and minimize the financial risks,** including loss of not approving the loan to capable applicants and loss of approving the loan to highly default applicants.

The company's made decisions on each application classified into 4 types: **Approved/ Cancelled / Refused / Unused offer**, as well as **defined the current applicants with their payment difficulties** (a target variable predicting defaulting probabilities)

Problem Statement

BUSINESS OBJECTIVE

- Identify **driving factors behind the loan default and non-default**. The pattern will help the company to understand and build appropriate models on the loan processing.

ANALYSIS OBJECTIVE

- Identify the **missing data** and use appropriate method to deal with it.
- Identify **outliers** to detect abnormalities
- Identify **data imbalance** and find the ratio of data imbalance.
- Explain the **analysis results** of univariate, segmented univariate, bivariate analysis, etc. in business terms.
- Find the **top 10 correlation with the target variable** (Client with payment difficulties and all other cases)
- Visualisations and summarise the most important insights

Assumption

1% value_counts presence - assumed to be insignificant to consider

20% missing value - benchmark to drop and eliminate the data

Co-applicants - who was accompanying client when he was applying for the loan

Data Approach

MISSING VALUES

- **CREDIT SCORE:** (EXT_SOURCE): Normalized credit score from 3 sources. Important data. Not all applicants have scores from all 3 sources. => Create **a mean Credit Score,** drop the individual columns & small (172 records) missing values.
- **CREDIT BUREAU INQUIRIES:** missing & fragmented into by hour, week, month,...-y => Consolidate into Total inquiries & Last-3-month inquiries
- **VARIABLE WITH SMALL MISSING PERCENTAGE (<20%)**:
  - Impute with mode for categorical data and mean for numerical data (For ex: Co-applicants, family members, product combination
  - Impute with median for outlier numerical data (For ex: Annuity)
  - Impute with '0' for Goods_pricing as NA are those not applying for consumption loan
- **VARIABLE WITH SIGNIFICANT MISSING PERCENTAGE (>20%)** as they lack too much data to deduce information
- **IRRELEVANT DATA:** not useful to evaluate applicant credit, dropping to make the data clean & focused (For ex: document data, contact data: phone, email, processing day and hour, days of phone change, living/working area infor)
- **DATA VALUES "XNA"** - filled in for many data columns - treated as missing data
  - Few missing: Replace with mode for categorical data such as Gender
  - Huge amount of missing: keep it and re-labeled with "Not specified" category (For ex: Organization Type
- BINNING: Loan term is often a range such as 3, 6, 12... months has both missing values & value ='0' (even approved status) => Binning the data into ranges & using a label "Not specified" for both missing and "0".

Data Approach

STANDARDIZE THE DATA

- Y/N and 1/0 for Boolean data: convert all into 1/0 format for consistency (For ex: Car & house ownership, last application flag)
- Fix the mismatching data type as some should be int instead of float (For ex: count of family members or social DPD)

OUTLIERS

- Missing values will be dealt with mode. However there are many non-missing values with significant abnormalities (Income) as the applicants has a group of extremely high income ones => Keep the outliers & deal with them when analyzing
- Abnormal data due to error is dropped (For ex: Days_employed > 350K ~ 950 years working)

GROUPING DATA

- Grouping similar labels (based on label name) to reduce the too fragmented cat labels (For ex: Organization Type, Product Combination)
- Grouping labels with value count percentage is too small (<1%) together into "Others" to reduce fragmented cat label (For ex: Goods category, Loan purpose

Data Approach

CALCULATED COLUMNS: New columns computed for more useful data crossing such as

- No. of dependents: including children and spouse (if married)
- Percentage of given Credit Amount over Goods-pricing (for consumption loan); or over Applied Amount
- Convert "Days" data into "Year": Employment evaluated by Years, Age range or Month Decision better than Days

GROUPING DATA

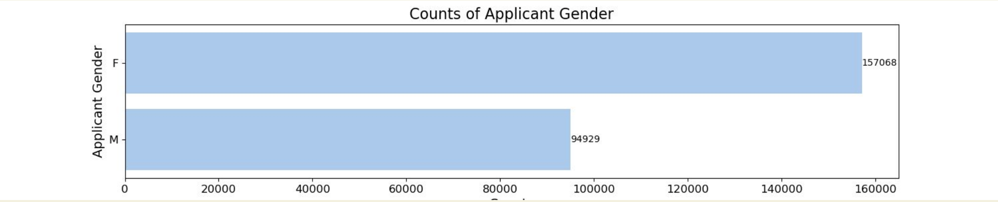
- Grouping similar labels (based on label name) to reduce the too fragmented cat labels (For ex: Organization Type, Product Combination)
- Grouping labels with value count percentage is too small (<1%) together into "Others" to reduce fragmented cat label (For ex: Goods category, Loan purpose
- Binning numerical data into categorical data with ranges for more helpful insight (For ex: Ranges of income, of Credit, of Good-Pricing, of Application Amount, of Age, of Employment Periods, and of Decision Range)

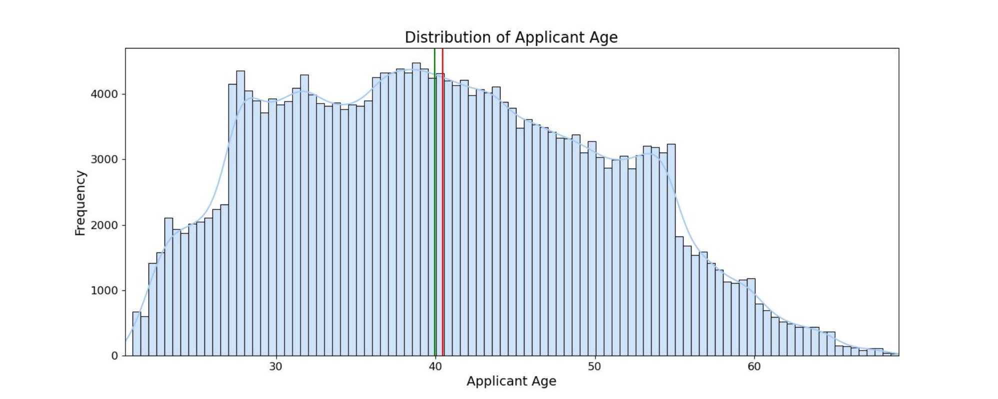
DATA DUPLICATION

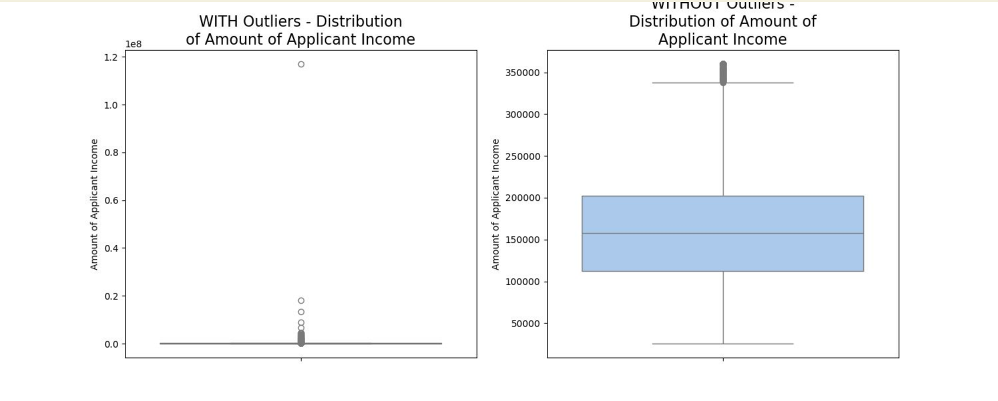
- Possibility of duplicated application records for one contract due to human mistakes - reflected by FLAG_LAST_APPL_PER_CONTRACT => Filtering data to get only applications for one contracts which is the last processed one

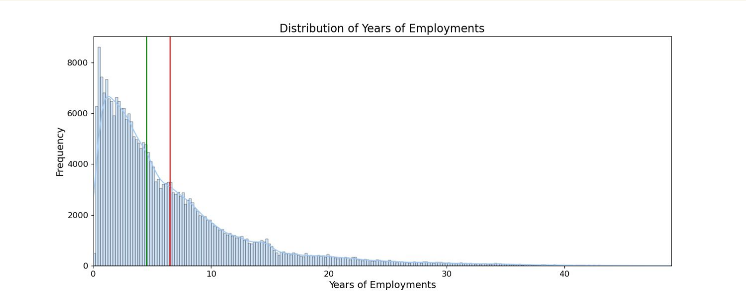
Typical Current Applicant Profile

- Majority if female (almost double male), applied with co-members
- Married with 1-2 dependents at 40s
- Income ranges 100-200K,, mainly from working and commercial associate in the business organization or self-employed.
- Working with less-than 10 year employed
- Secondary education is dominant
- Owning a house/apartment



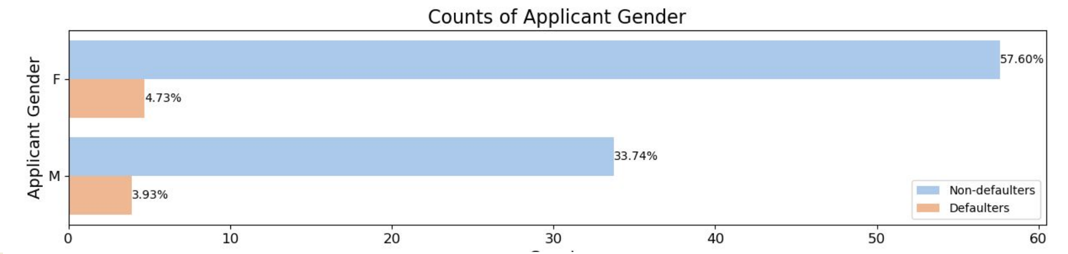


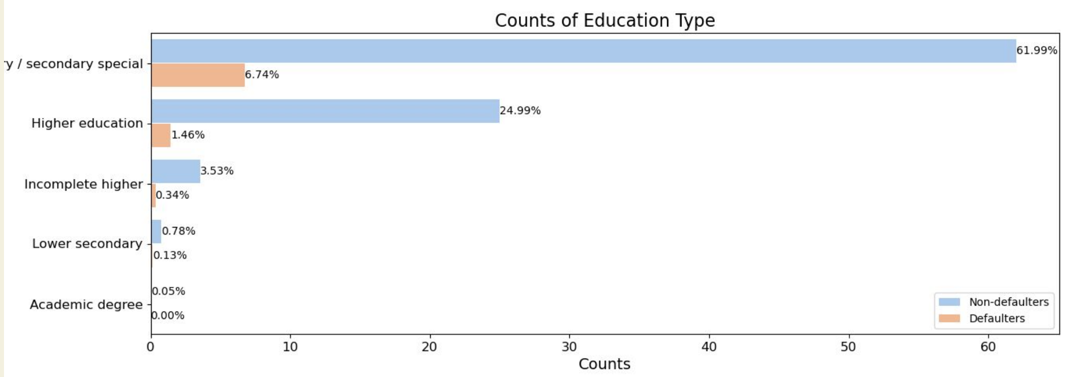


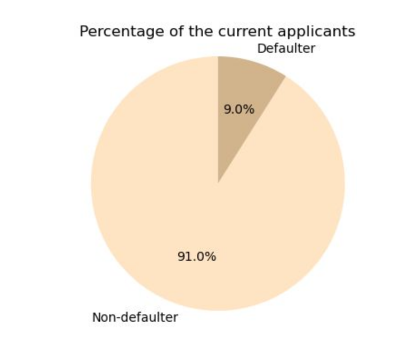


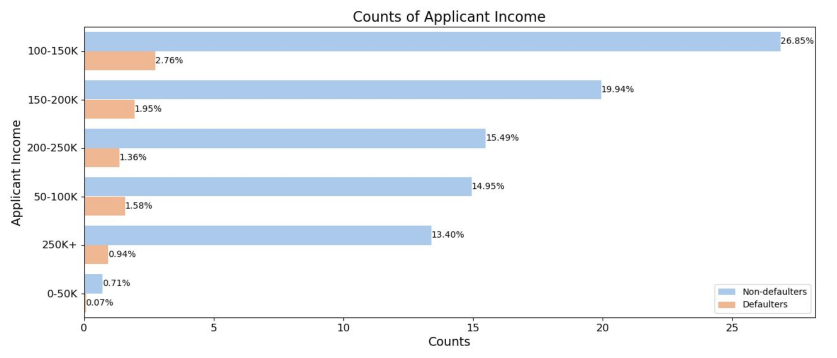
Defaulters vs. Non-defaulters

- Data imbalance toward 91% non-defaulting
- While female is much higher in the group of non-default, a relatively similar ratio of male/female in the group of defaults.
- Applicants with higher education relatively less likely to default
- Defaults fall into middle income range, rather than focus on low income range.



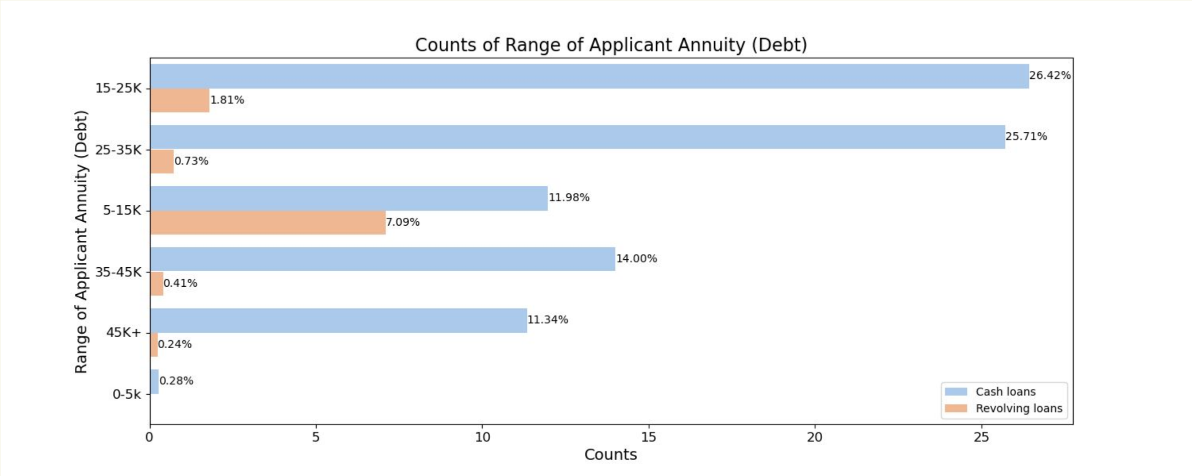


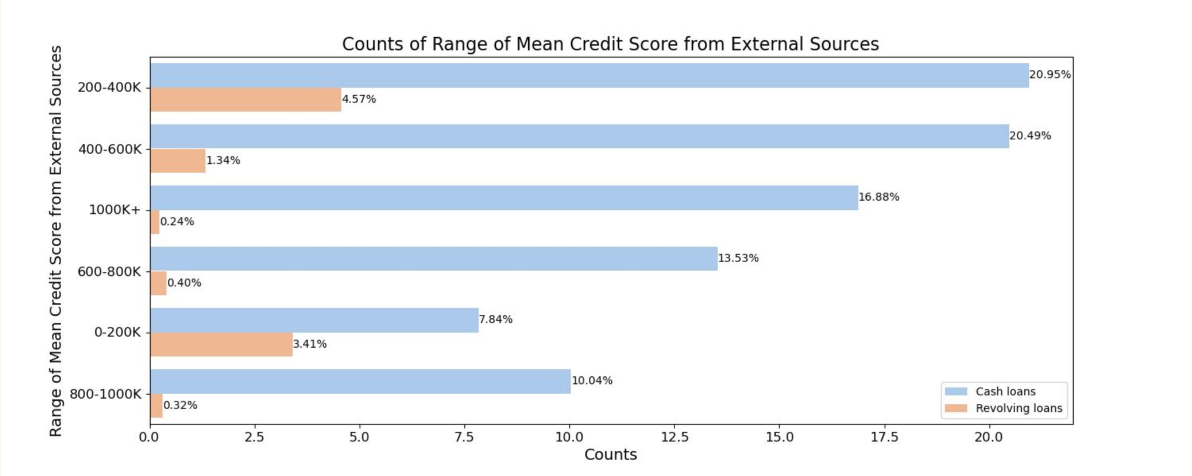




Cash loan Vs. Revolving

- While cash loan applicants has a high range of debts (15+), revolving loan applicants has a smaller annuity range 5-15k.
- Interestingly, revolving loan generally has got lower mean credit score than cash loan applicants

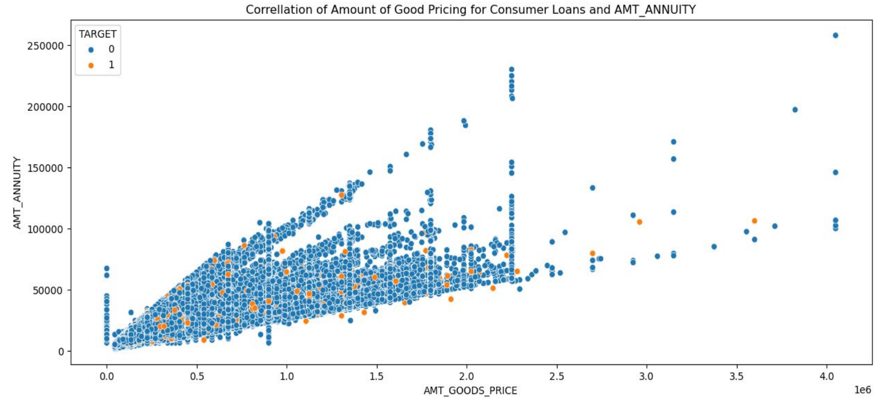


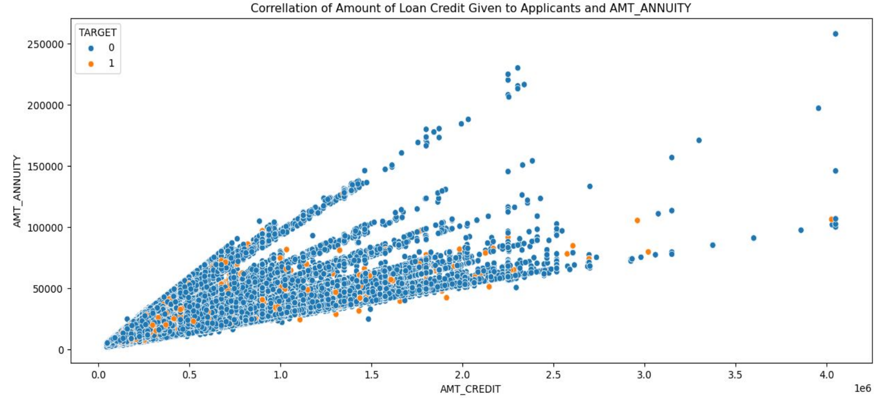


Correlation with Annuity (Debt)

A strong correlation amongst current applicants between

- Amount of annuity vs. good pricing, showing that the higher value their need for consumption loan, the higher debt they've been taking
- Amount of given credit in par with the debt as well, as credit is more often given to people with long and many types of loan.

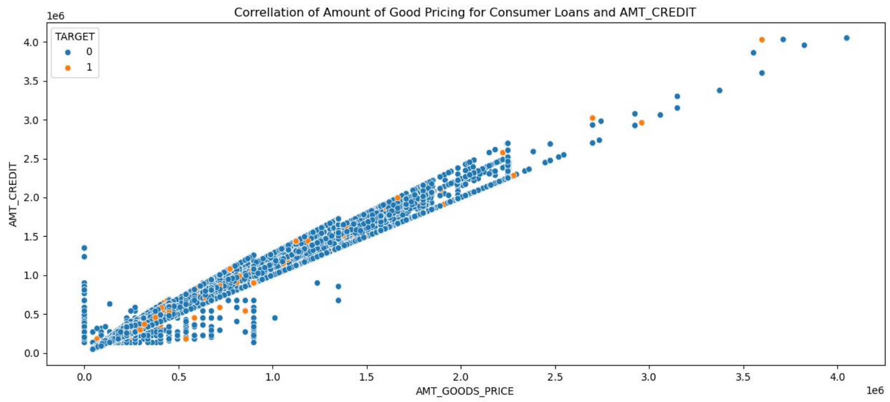


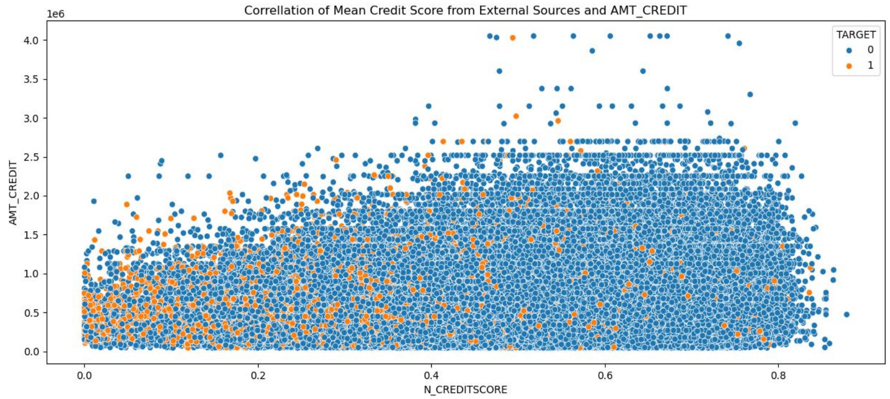


Correlation with Credit

A strong correlation amongst current applicants between

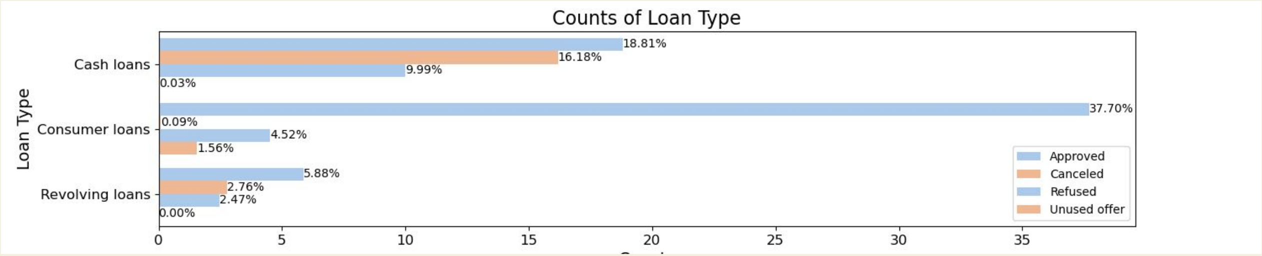
- Amount of given credit with good-pricing, as the credit limit also takes into the account of the percentage of purchased goods - collaterals of the loan
- Amount of given credit, interestingly, not shows a linear trend with the credit score. It seems that the credit limit is decided by other factors than the applicants' credit score.

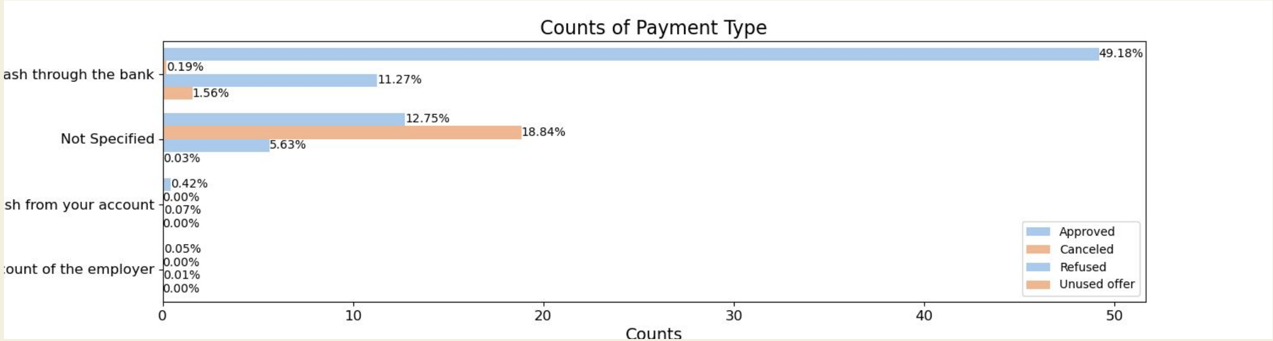


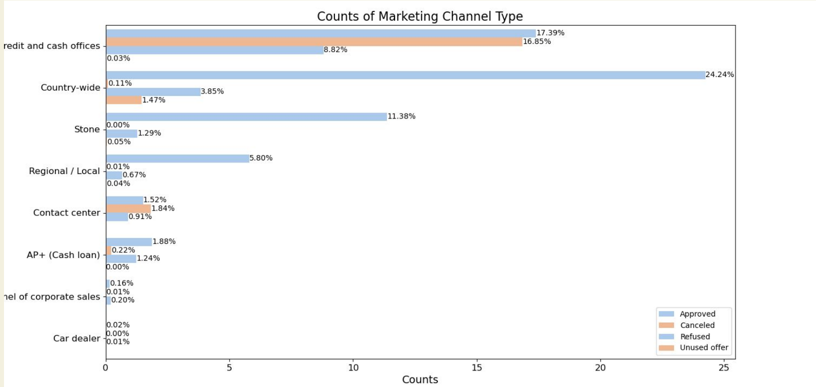


Previous applicants

- Mainly approved for consumer loans, than cash loans
- They get cash through the bank
- They're acquired mainly through the credit and cash offices

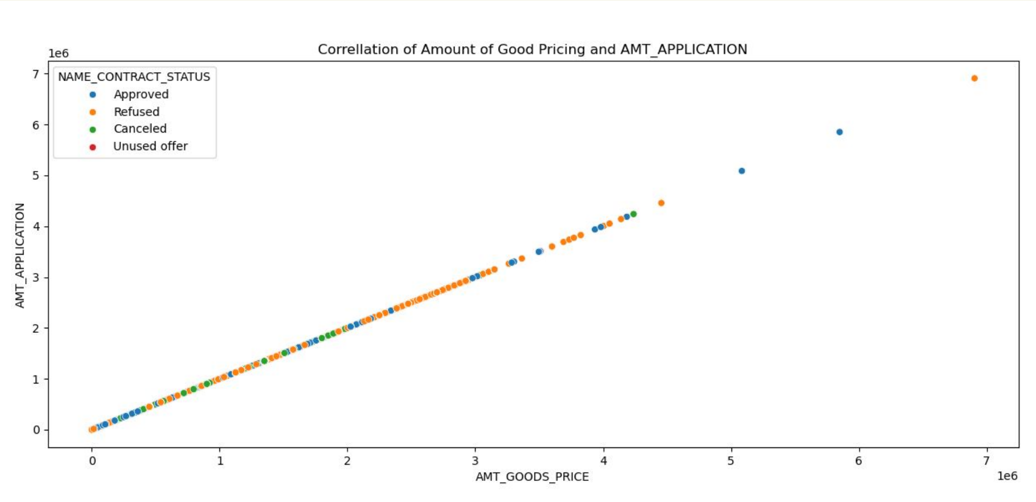


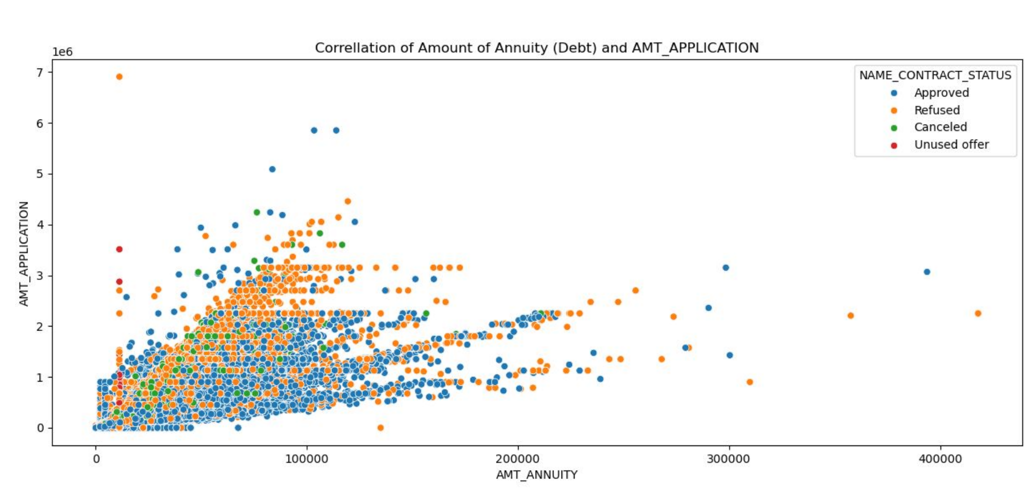




Previous applicants

- Since the previous clients mainly applied for consumer loan:
  + A highly correlated trend between the amount applied for the loan and the goods pricing
  + The correlation with the debt is less, much more scatterly





Recommendations

- Potential clients are married working female, middle-aged. Working for business organization or self-employed with a middle-range income. The marketing team can approach them through credit and cash offices nationwide.
- Men is more likely to default than women, probably women are more careful with their credit.
- Approaching high-income and high education group is also potential too, consider different package for this cohort.
- Consumption loan is strongly aligned to the good pricing, rather than the annuity in the previous clients; however, the current clients are applying more for cash loan with much higher correlation with annuity.
- Credit limit tends to be given more on the collaterals or good pricing rather than neither the applied amount nor the credit score