EDA Assignment: Loan Default Analysis

RITIKA MEHRA (BATCH C47)

Problem Statement

- Loans are the main source of income for banks. The bank generates its revenue from difference between the interest rates offered on deposits and loans.
- Therefore, bank cannot accept or reject all the loan applications that it receives.
- The bank may incur losses, if it rejects a loan application of a client that is not likely to default or if it accepts the loan application of a client that is likely to default.
- The task at hand is to find patterns in current and historic data for the clients who are likely to default.

Assumptions

- Assuming XNA and XAP as null values.
- XNA, XAP, and Missing category in OCCUPATION_TYPE have been ignored while mentioning observations and insights.
- For better readability of pie charts, very low percentages have been ignored as we are only concerned about highest percentage. Percentage cutoff is highlighted in comments.

Overall Approach

1. Data understanding

1.1 Loading dataset and describing them

1.2 Identifying missing values

1.3 Dropping columns having higher percentage of missing values [>40%]

1.4 Data Type correction

2. Data formatting

2.1 Data formatting [Columns representing days and having negative values]

Overall Approach

3. Data cleaning

3.1 Handling missing data [Creating category or imputation with central tendency measures]

3.2 Identifying outliers

3.3 Handling outliers [Creating range columns for corresponding amount columns using binning]

3.4 Checking for data issues [Outlier value in DAYS_EMPLOYED for Unemployed and Pensioners]

3.5 Data Imbalance Ratio [91% data is for non-defaulters and only 9% data for defaulters]

4. Data Analysis

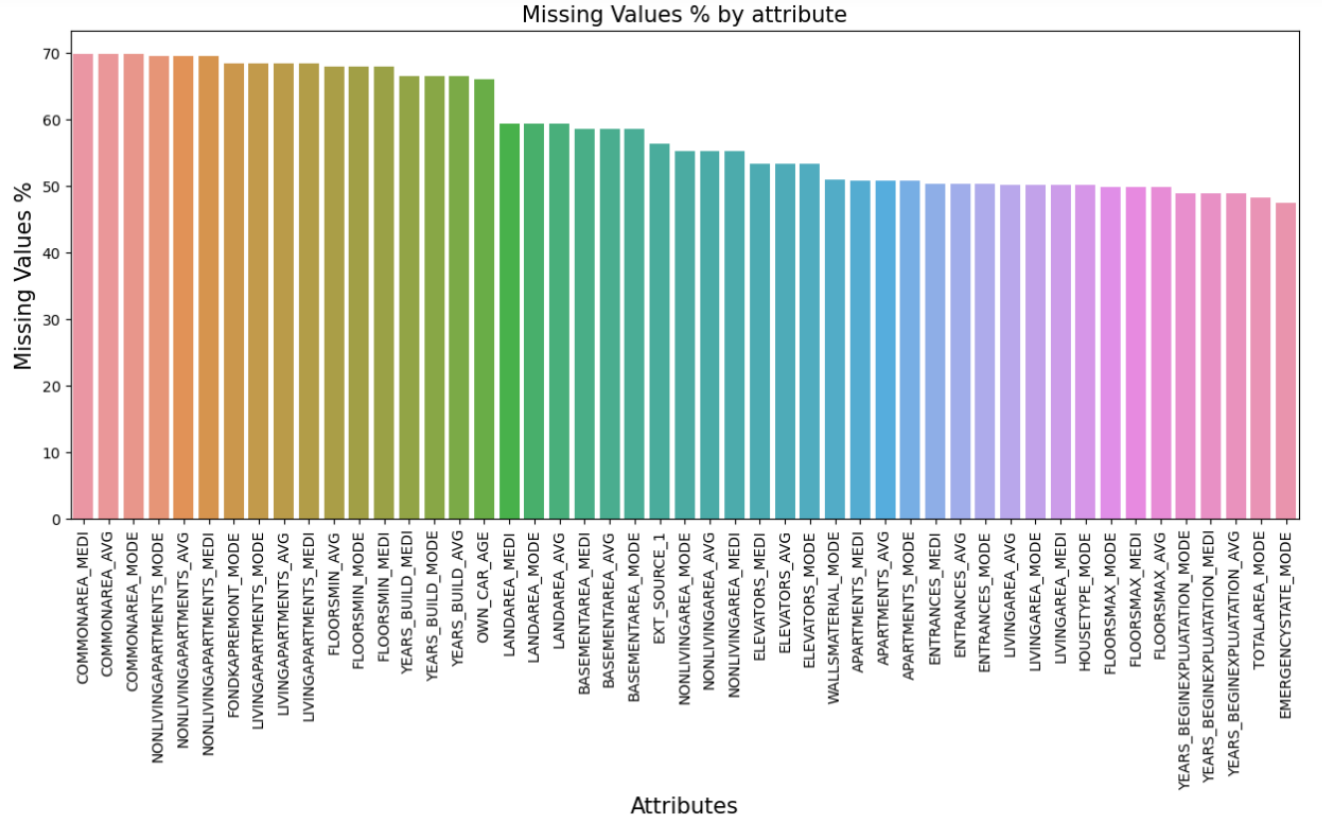
4.1 Univariate Analysis

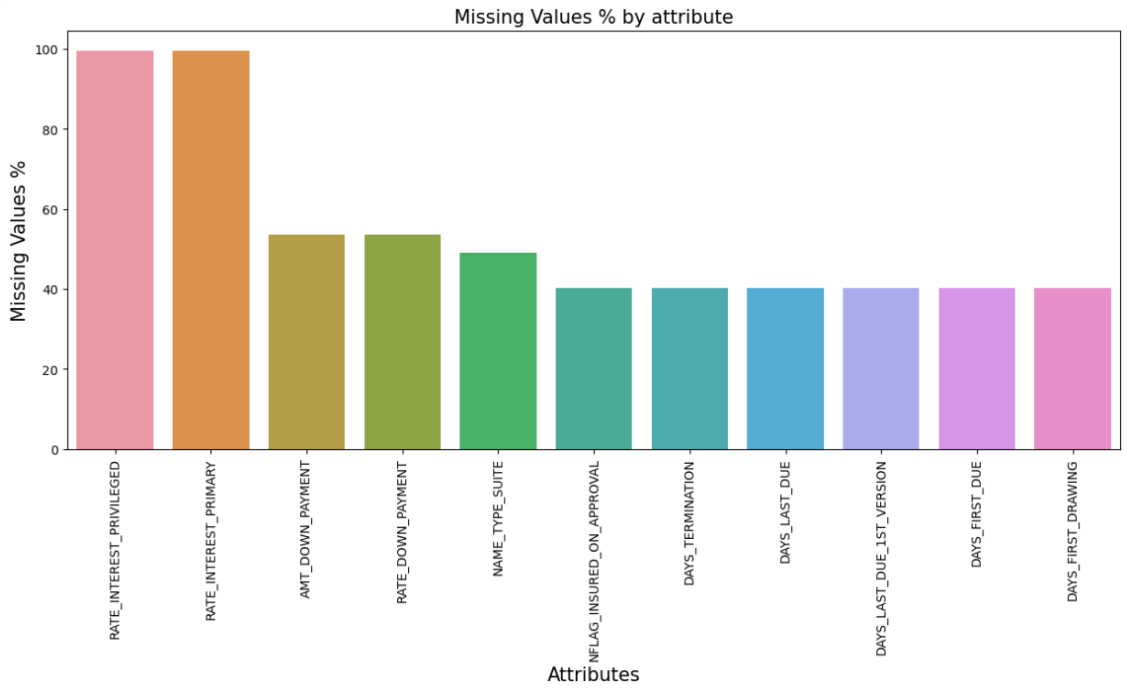
4.2 Bivariate Analysis

4.3 Multivariate Analysis

4.4 Top Correlations

Missing Values Treatment



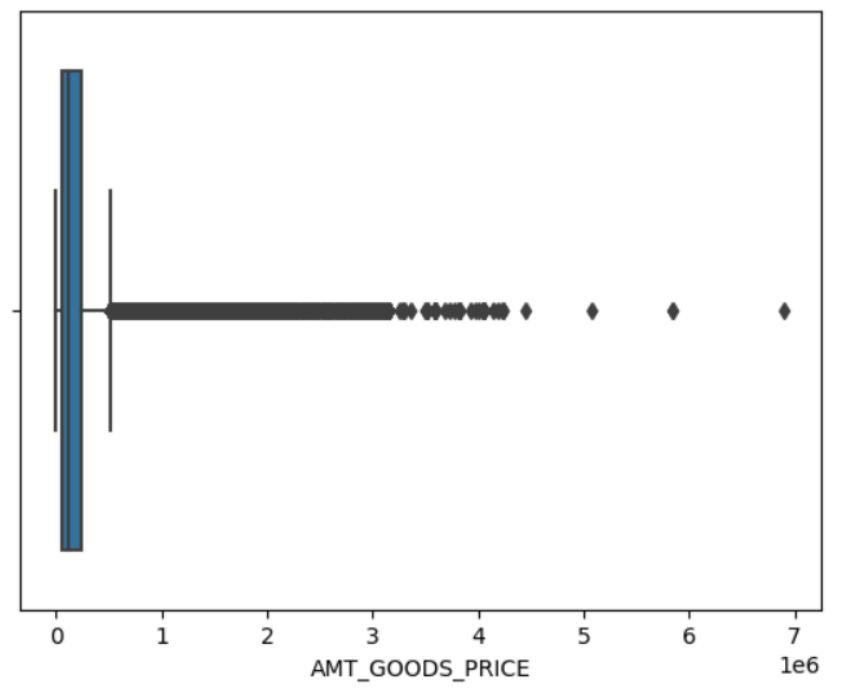


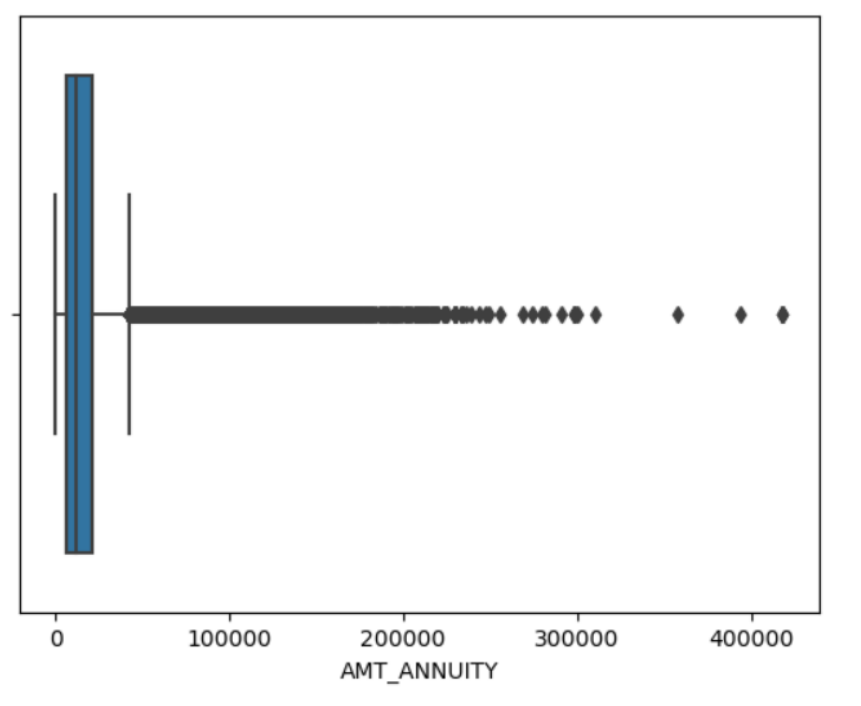
Note: The bar charts represent only those attributes that have greater than 40% missing values.

For this case study, the threshold of >40% missing values is considered for dropping attributes. Other missing value treatments used:

- Creating separate category for missing values for the attribute – OCCUPATION_TYPE.
- Imputing missing values with central tendency measures based on presence of outliers.

Outlier Treatment





Above box plots show examples of outliers. The outliers in amount related columns have been treated by creating corresponding range columns using binning.

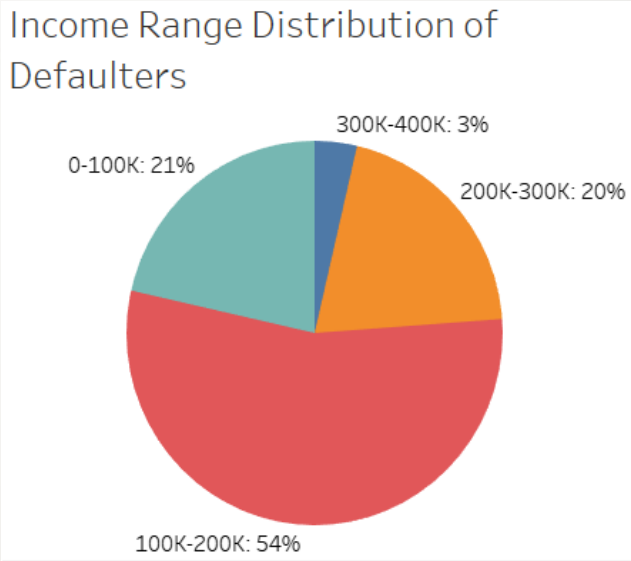
Driver Variables

6 Strong driver variables identified as follows-

AGE

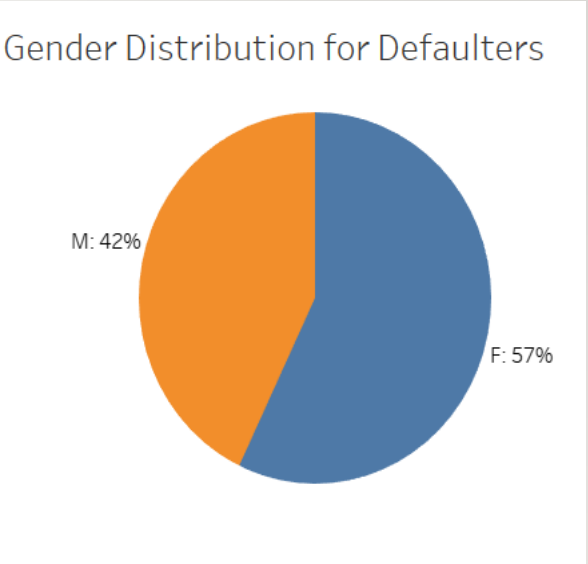
Insights: Clients in the age-group 30-40 are most likely to have payment difficulties (defaulters) as this age group covers most of the expense driven groups such as married and have kids, or having other loans.

INCOME_RANGE



Insights: Clients having income range 100K-200K are most likely to have payment difficulties (defaulters) due to low income.

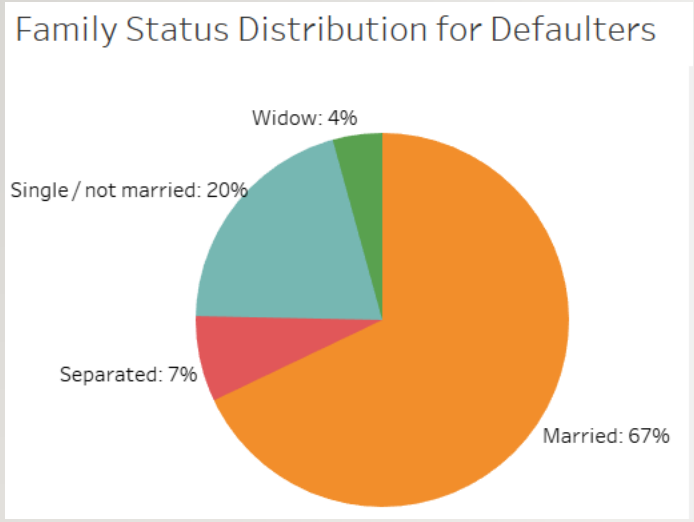
GENDER



Insights: Female Clients are most likely to have payment difficulties (defaulters).

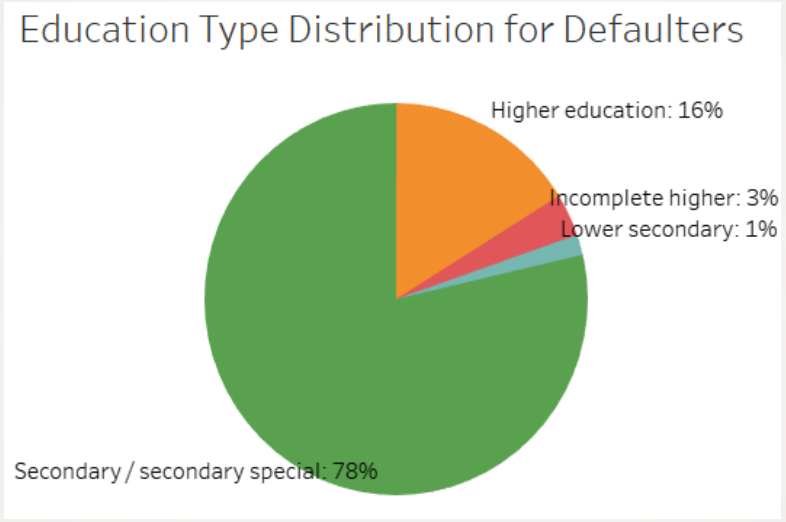
Driver Variables

FAMILY_STATUS



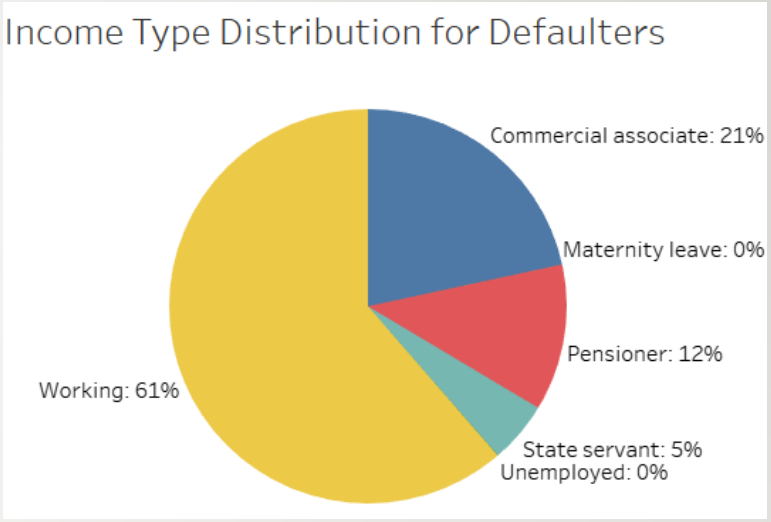
Insights: Clients who are married are most likely to have payment difficulties (defaulters) due to increased expenses of providing for family.

NAME_EDUCATION_TYPE



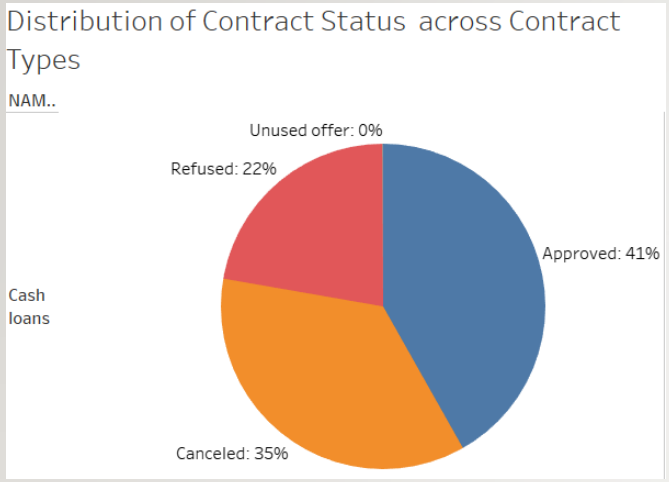
Insights: Clients who have secondary/secondary special education are most likely to have payment difficulties (defaulters).

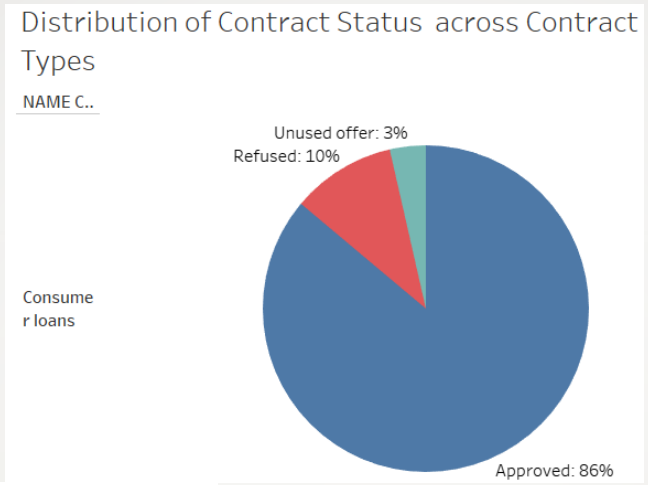
NAME_INCOME_TYPE

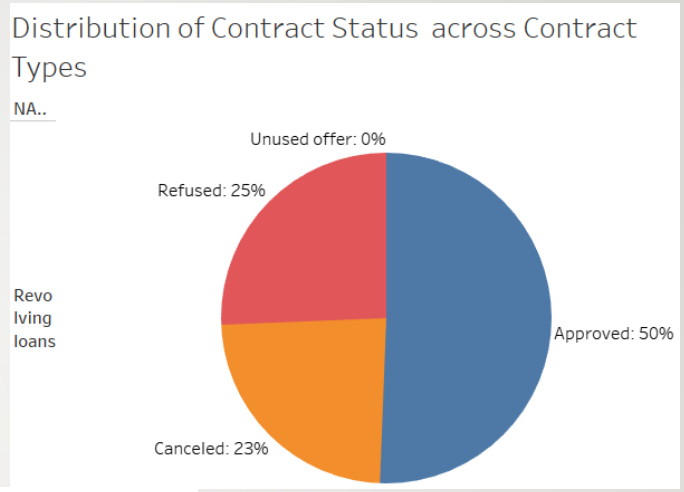


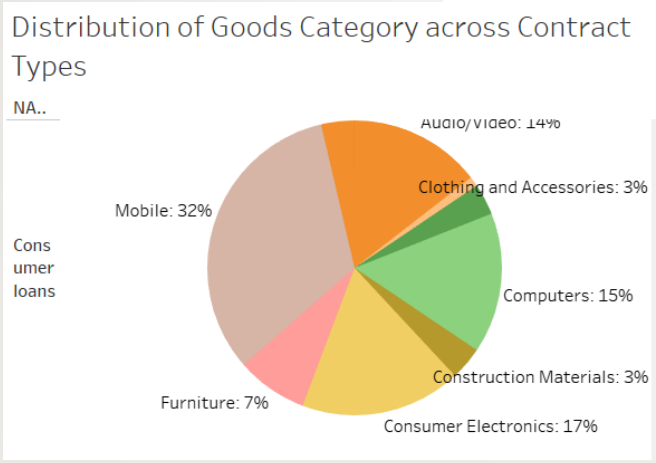
Insights: Clients who are working are most likely to have payment difficulties (defaulters) as there is very less job security in private sector.

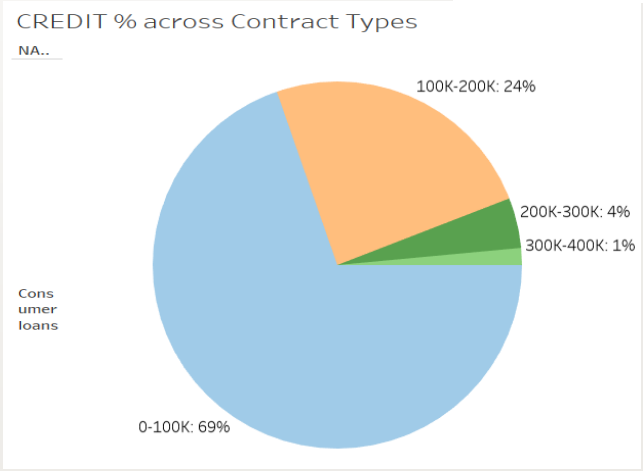
Previous Application Data Analysis





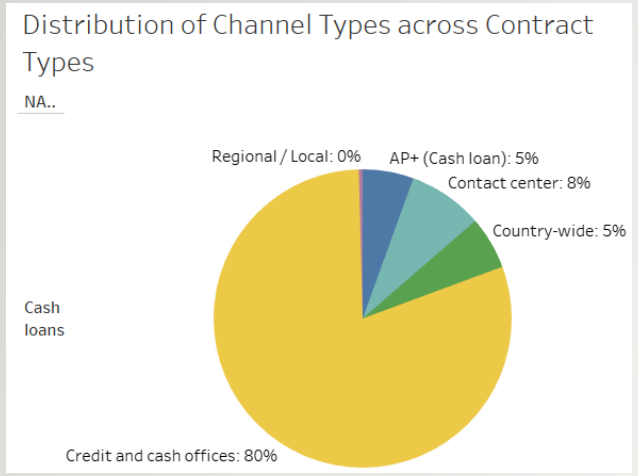


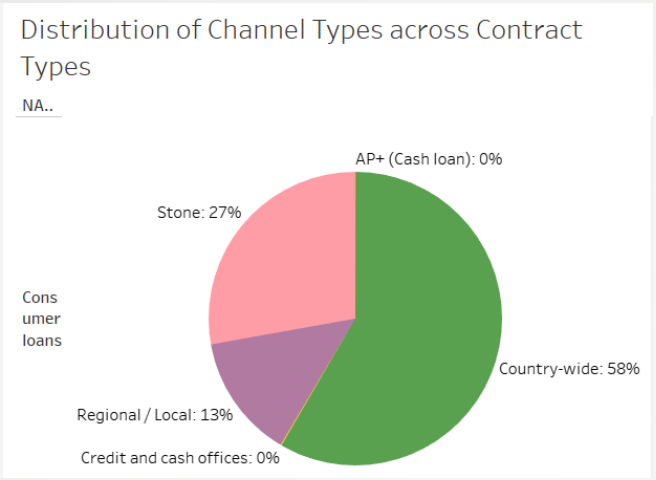


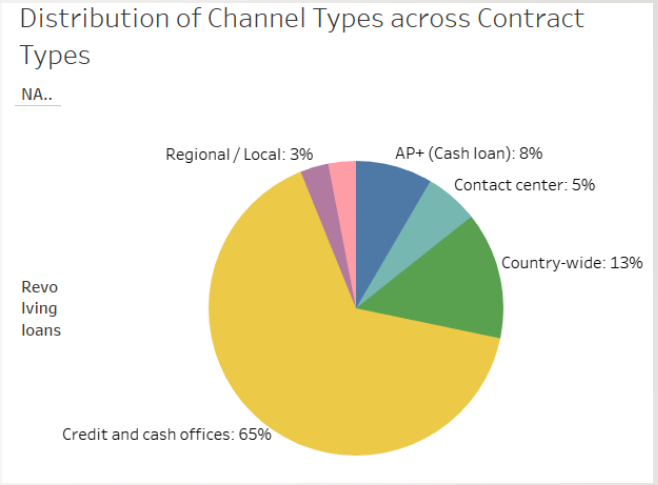


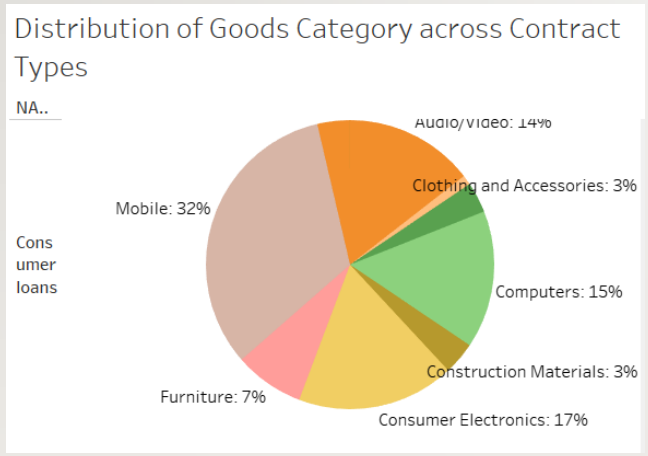
Insights: Maximum Approval Rate is in Consumer Loan as most loans are for mobiles and for low credit range 0-100K.

Previous Application Data Analysis





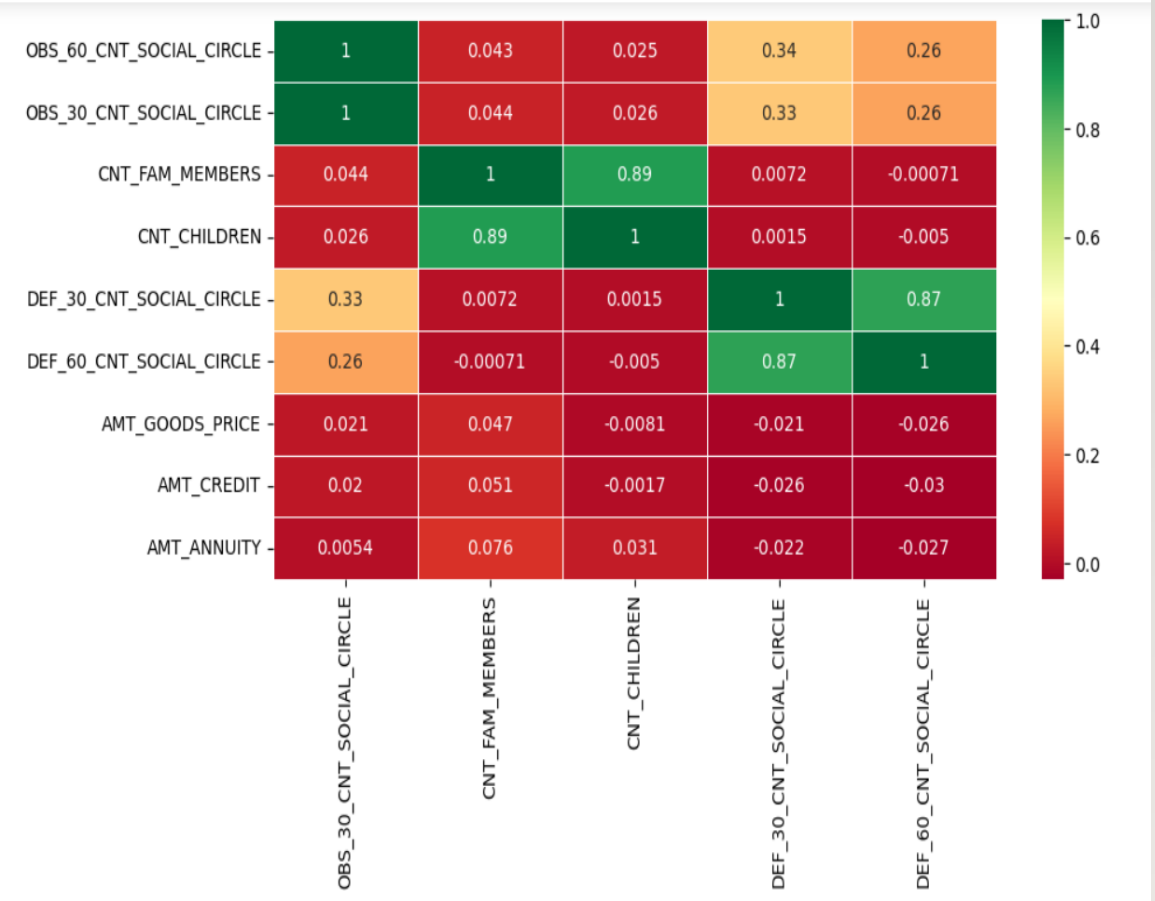




Insights: Credit and cash offices are the biggest source for Cash and Revolving loans, and country-wide is for Consumer loans. As for cash loans, client has to visit the office and but they can get consumer loans online as well.

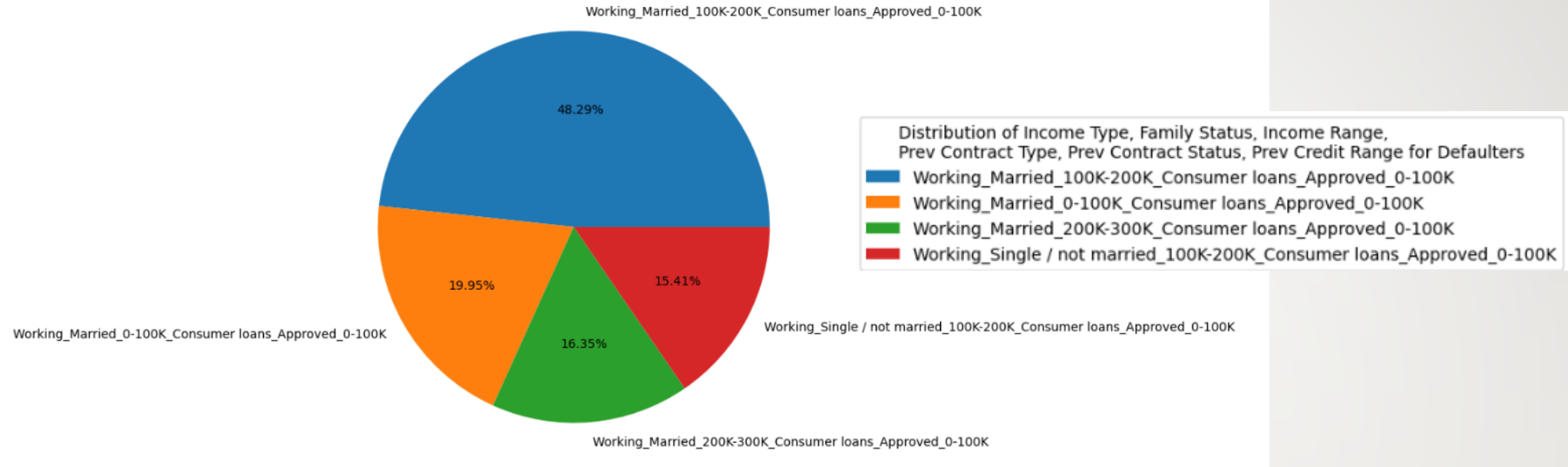
Top Correlations

1. Top correlations for loan defaulters–
    1. OBS_60_CNT_SOCIAL_CIRCLE and OBS_30_CNT_SOCIAL_CIRCLE
    2. AMT_CREDIT and AMT_GOODS_PRICE
    3. CNT_FAM_MEMBERS and CNT_CHILDREN
    4. DEF_30_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE
    5. AMT_GOODS_PRICE and AMT_ANNUITY
    6. AMT_CREDIT and AMT_ANNUITY



Note: The figure shows heatmap only for top correlations. Heatmap for all the variables is available in notebook.

Merged Data Frame Analysis



Insights: Clients who are Working, Married, have income range 100K-200K and have previously got consumer loans approved for credit range 0-100K are most likely to have payment difficulties (defaulters).

Conclusions

- Clients are most likely to default if,
- They are in the age group 30-40 yrs, Working, Married and have income range 100K-200K.
- They have previously got consumer loans approved of credit range 0-100K.