EDA REPORT FOR CREDIT CARD DEFAULTERS ANALYSIS

Muhammad Hanafi Bin Mohd Sani

BUSINESS UNDERSTANDING & OVERVIEW

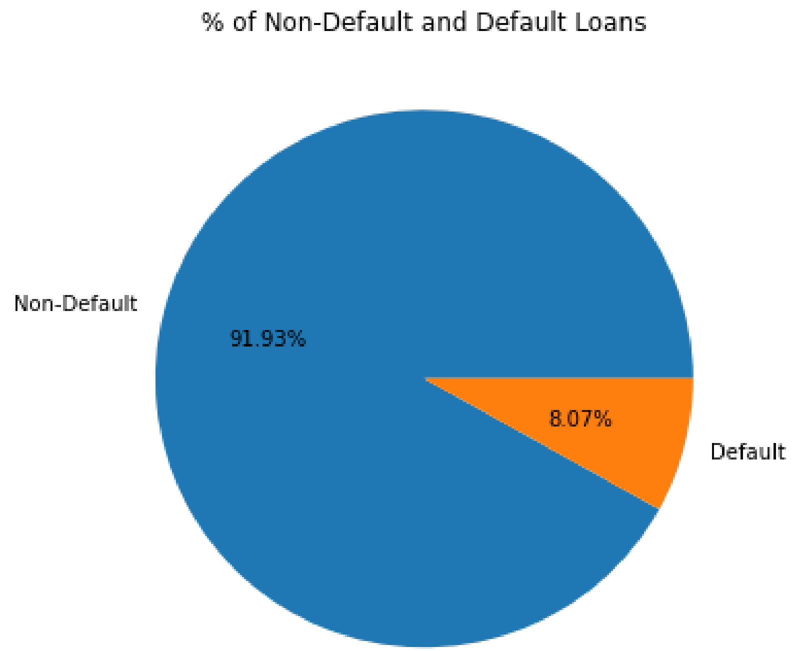
Understanding the cause of behind the loan defaults

By performing the EDA, we can pinpoint the cause based on the insight generated from the analysis

UNDERSTANDING THE DATA

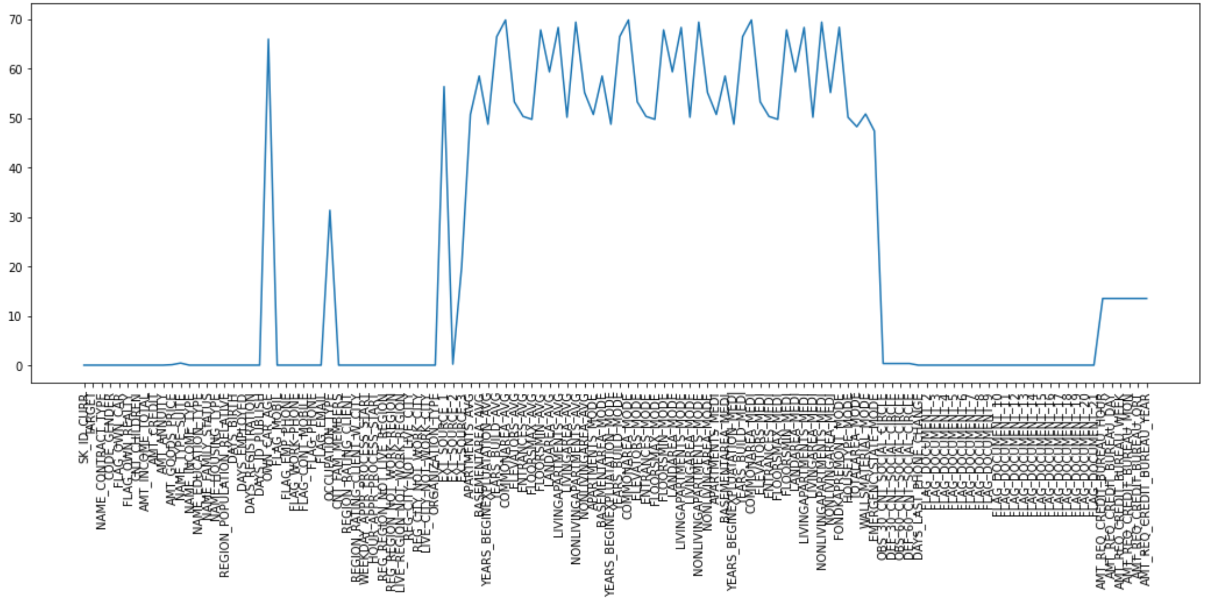
TARGET VARIABLE:FINDINGS

- Target variables are highly unbalanced: Only 8.07% for Default and a whopping 91.93% for non-Defaults.
- Meaning, most of the loans are paid on time (non-default)
- Further analysis required



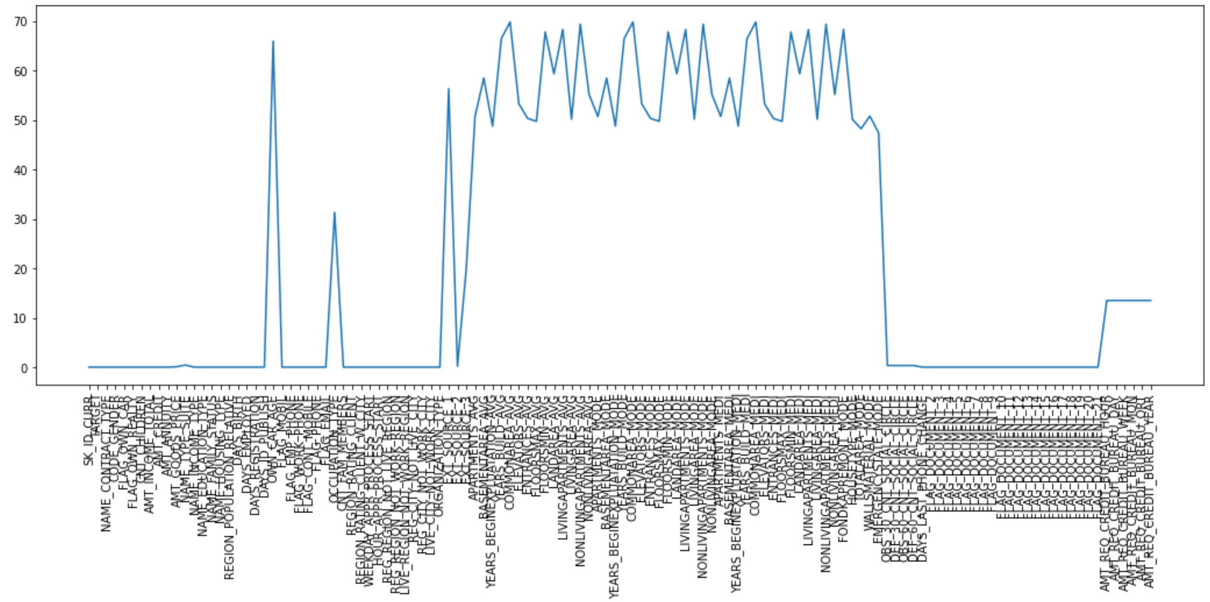
Initial Analysis From The Data

Missing Data



- Total Columns with >40% null values:64
- There is no thumb rule for dropping any column with 40% of data, but for example we can use the where the most column description are Normalized and its description are not clear. Hence we remove the columns.
- Most of the column doesn’t have any null data
- The REGION_RATING_CLIENT_W_CITYREGION_RATING_CLIENT are probably useful for data analysis as it shows the location rating(as in how many people paid their loan at that place) of the client in determining loanability (refer columns description for confirmation)
- CAR_AGE does not have any relation to the loanability

Initial Analysis From The Data (Cont.)



- Resulting columns left after removal:
  'SK_ID_CURR', 'TARGET', 'NAME_CONTRACT_TYPE’, 'CODE_GENDER','FLAG_OWN_REALTY','FLAG_OWN_CAR','CNT_CHILDREN', 'AMT_INCOME_TOTAL’, 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE’, 'NAME_FAMILY_STATUS','NAME_HOUSING_TYPE', 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION’, 'DAYS_ID_PUBLISH', 'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_CONT_MOBILE’, 'FLAG_PHONE','FLAG_EMAIL', 'OCCUPATION_TYPE', 'ORGANIZATION_TYPE’, 'CNT_FAM_MEMBERS','REGION_RATING_CLIENT_W_CITY', 'REGION_RATING_CLIENT’
- FLAG_WORK_PHONE is a duplicate of FLAG_PHONE

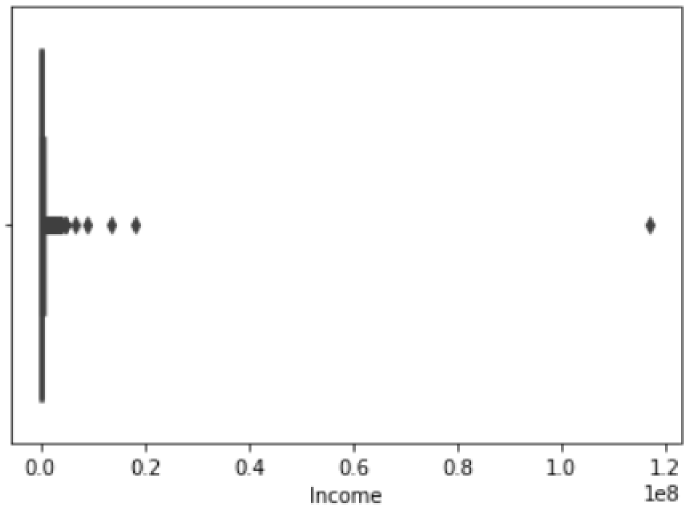
Incorrect Data

- Some data like ORGANIZATION_TYPE has “XNA” in it instead of NULL VALUES. To fix this: we will remove the row or convert the XNA into mean, median or mode in the columns depending on the situation.
- Data that should such as DAYS_EMPLOYED and DAYS_REGISTRATION are in negative values and should be positive.
- Some of data such as AMT_INCOME_TOTAL and have such an extreme value that caused skewedness in analysis. To fix: remove them. Removal of the extreme data does not affected the analysis.
- Data such as ['AMT_ANNUITY' , 'AMT_APPLICATION' , 'AMT_CREDIT', 'AMT_GOODS_PRICE’] are converted into thousands.

```
Business Entity Type 3    67992
XNA                       55374
Self-employed             38412
Other                     16683
Medicine                  11193
Business Entity Type 2    10553
Government                10404
School                     8893
Trade: type 7              7831
Kindergarten               6880
Name: ORGANIZATION_TYPE, dtype: int64
```

```
 365243    55374
-200         156
-224         152
-230         151
-199         151
Name: DAYS_EMPLOYED, dtype: int64
```

```
-1.0    113
-7.0     98
-6.0     96
-4.0     92
-2.0     92
Name: DAYS_REGISTRATION, dtype: int64
```

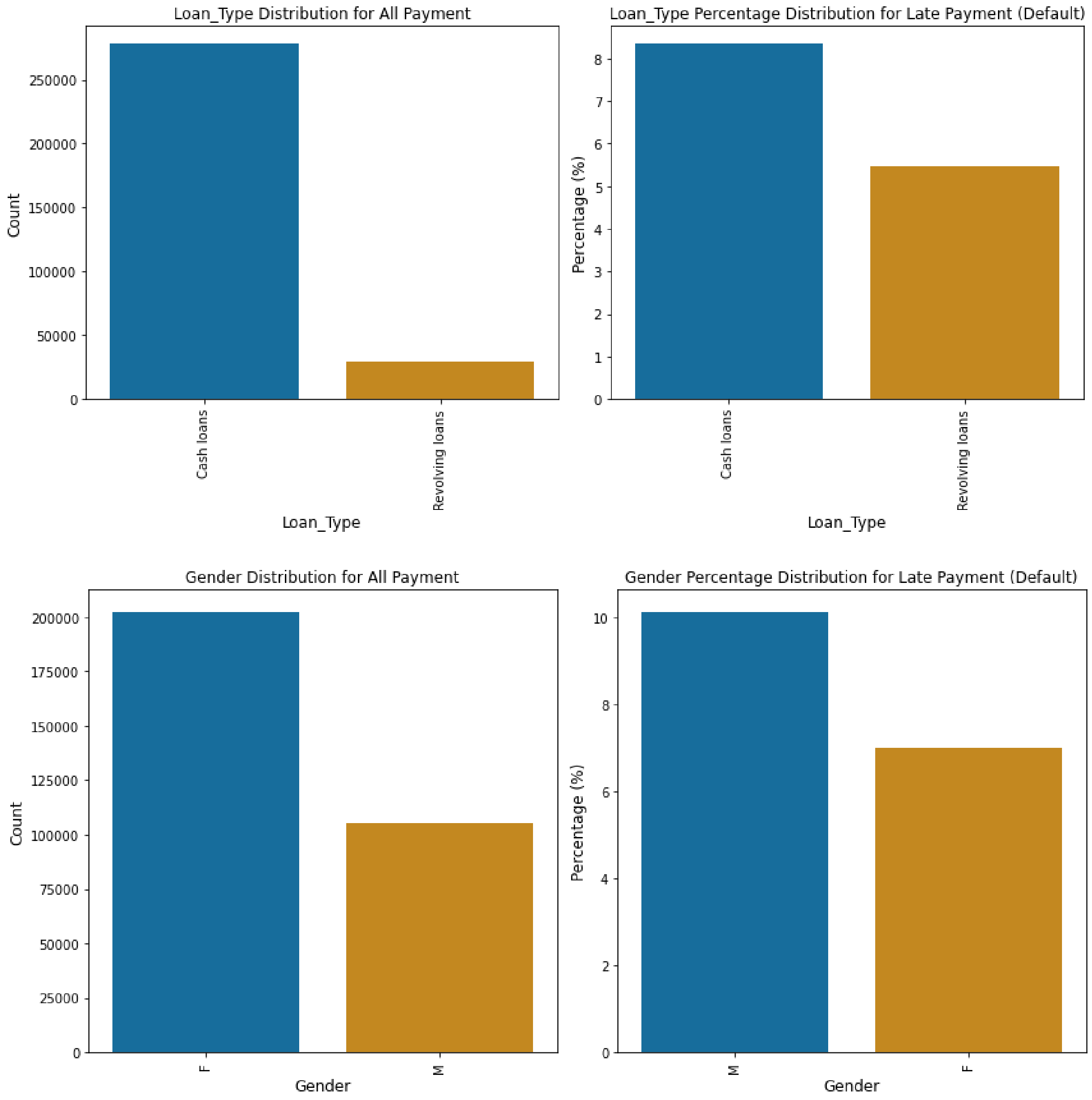


CATEGORICAL ANALYSIS

Loan Type, Gender vs. Target Variable

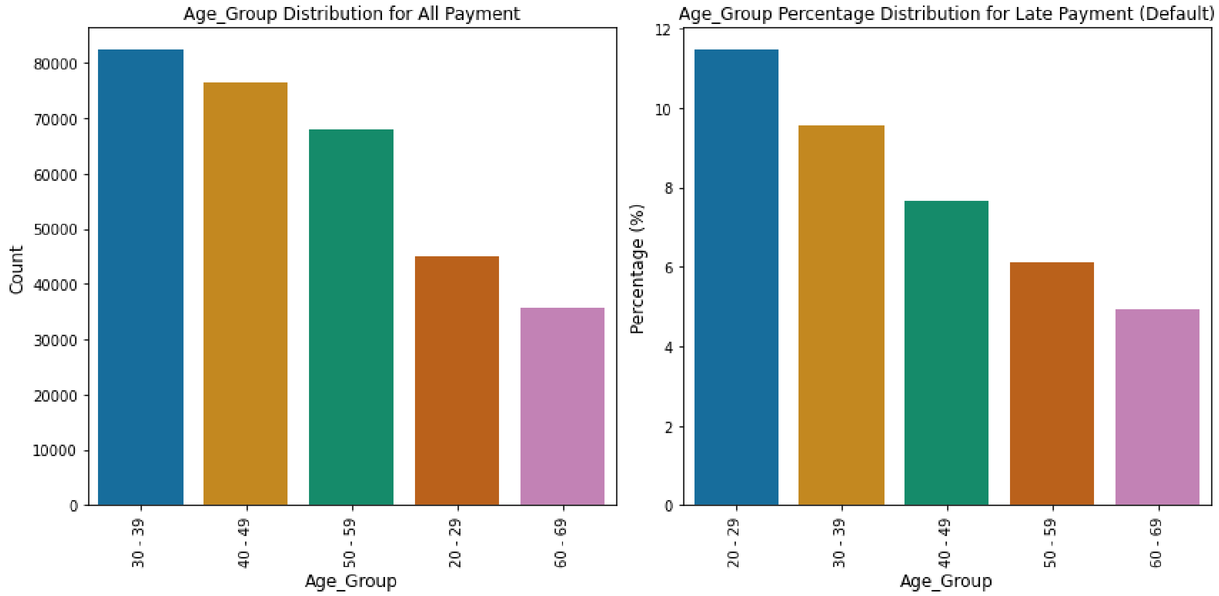
Findings

- Cash Loan are the majority for both loans type and contributors to defaults.
- We can see that % of loan non-default of males are low but the defaulted % for Males are higher, this shows that males generally have higher chance of becoming a defaulter than females.



Age Groups vs. Target Variable

Findings
• people that are in their 20s and 30s shows higher % in the default graph compared to other age groups. This shows that this group are likely to become a defaulter.

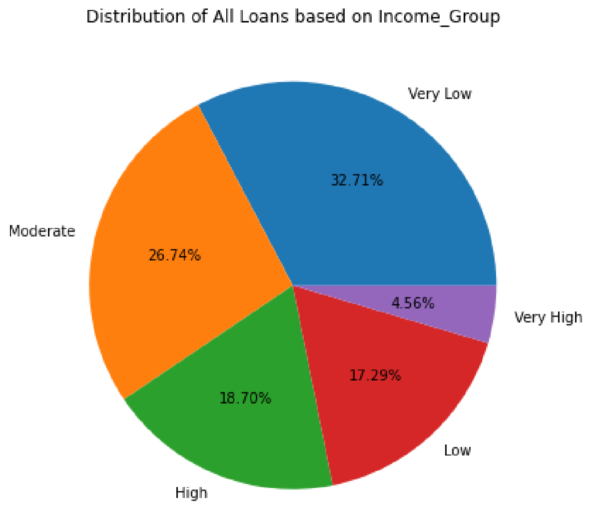


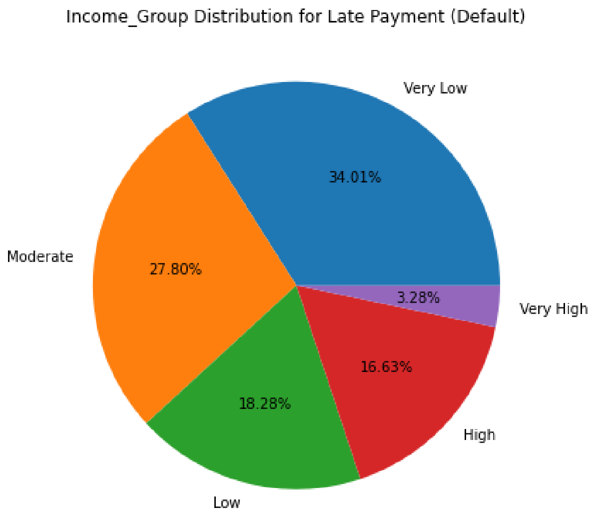
Income and Loan Groups vs. Target Variable

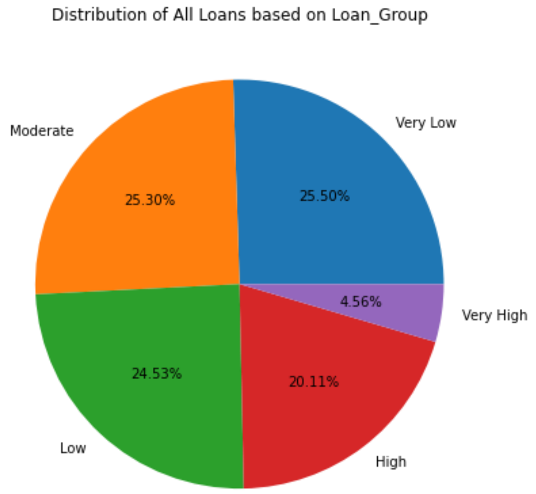
Findings

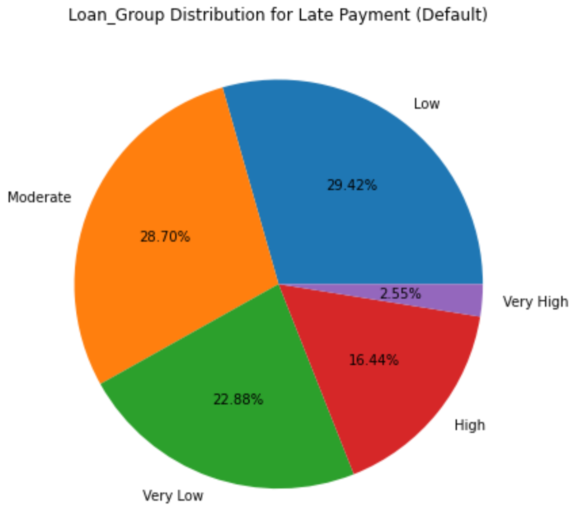
- As you can see from the 2 top pie charts it’s clear that major % of loans and its defaulters falls in the Very Low Income Group followed by Moderate.
- In the 2 bottom pie charts, moderate and very low loan group almost shares majority %, while a different case in default, where Low loan group have the majority % followed by Moderate.

Note: The groups are divided as follows:
Very Low: Bottom 25%
Low: 25th - 50th percentile
Moderate: 50th - 75th percentile
High: 75th - 95th percentile
Very High: Top 5%

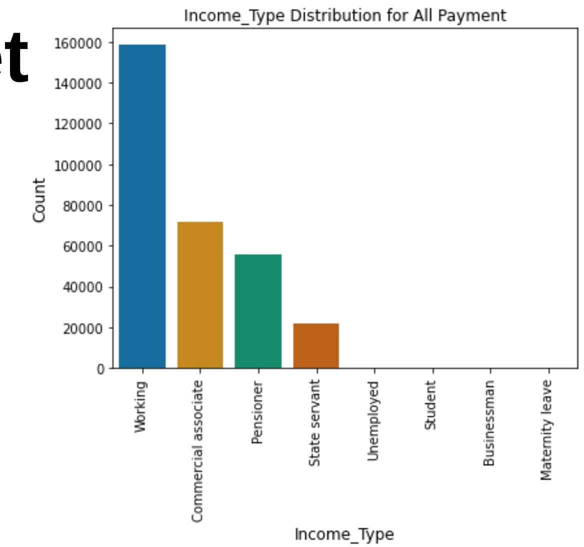


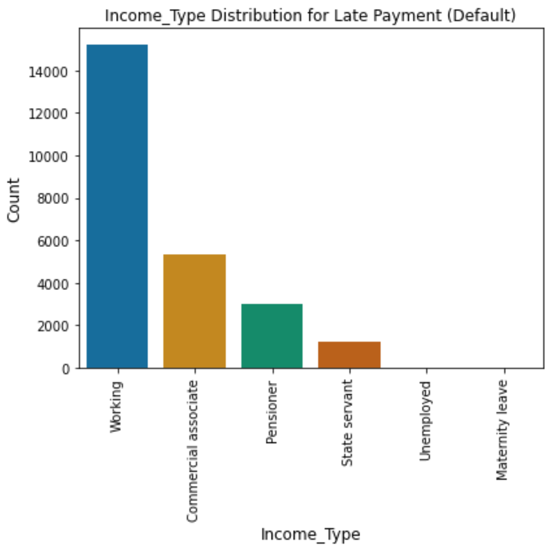






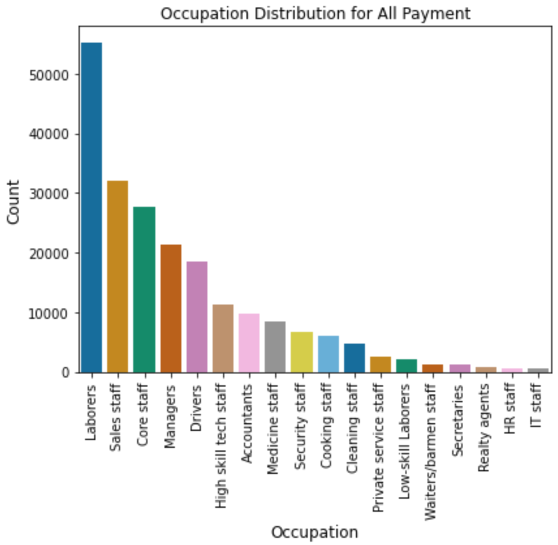
Income and Occupation vs. Target Variable

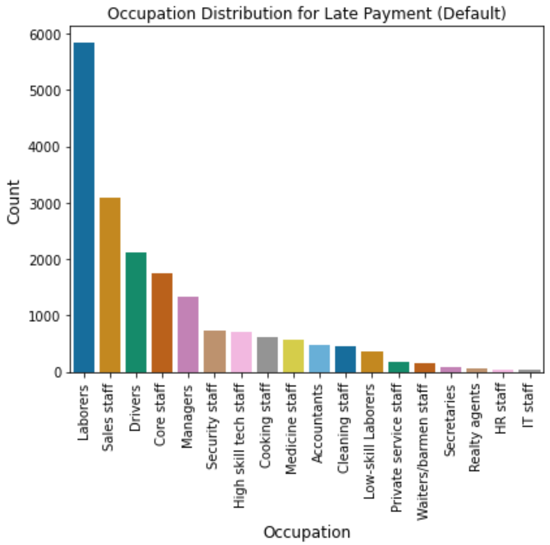




Findings

- Top Bar charts: Working people dominates the chart of amount of Loans and defaults: Working people are at the most risk to default
- From bottom bar charts: Laborers are the majority of the loaners and also the most who are defaulted.

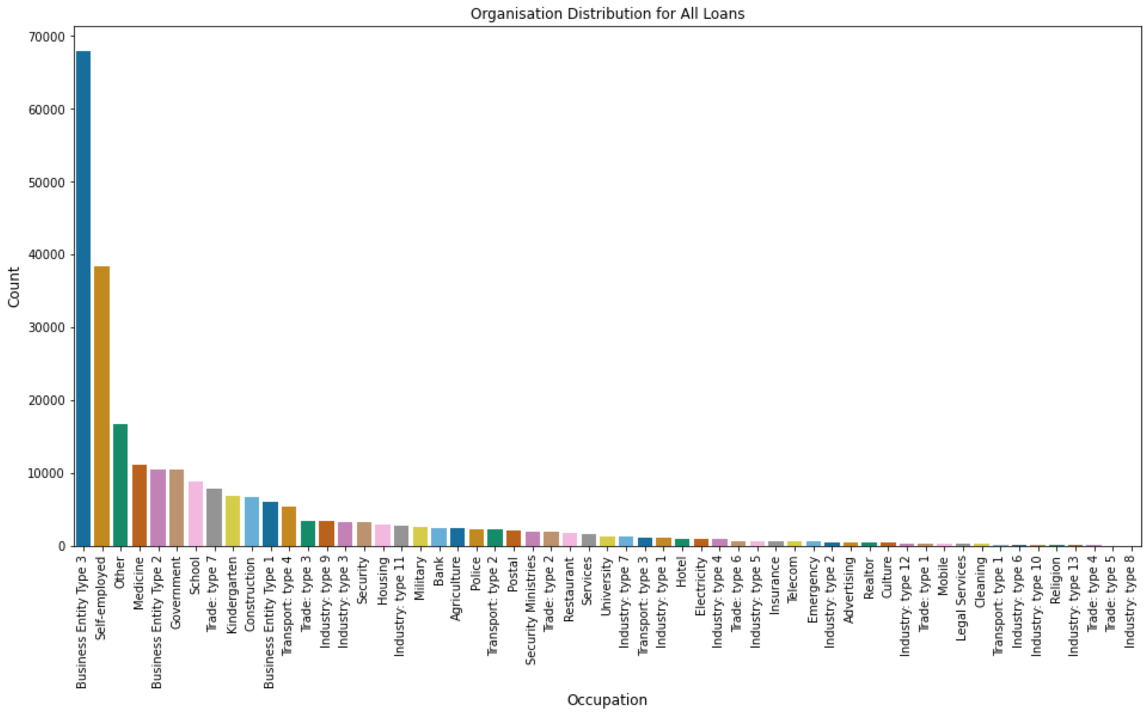


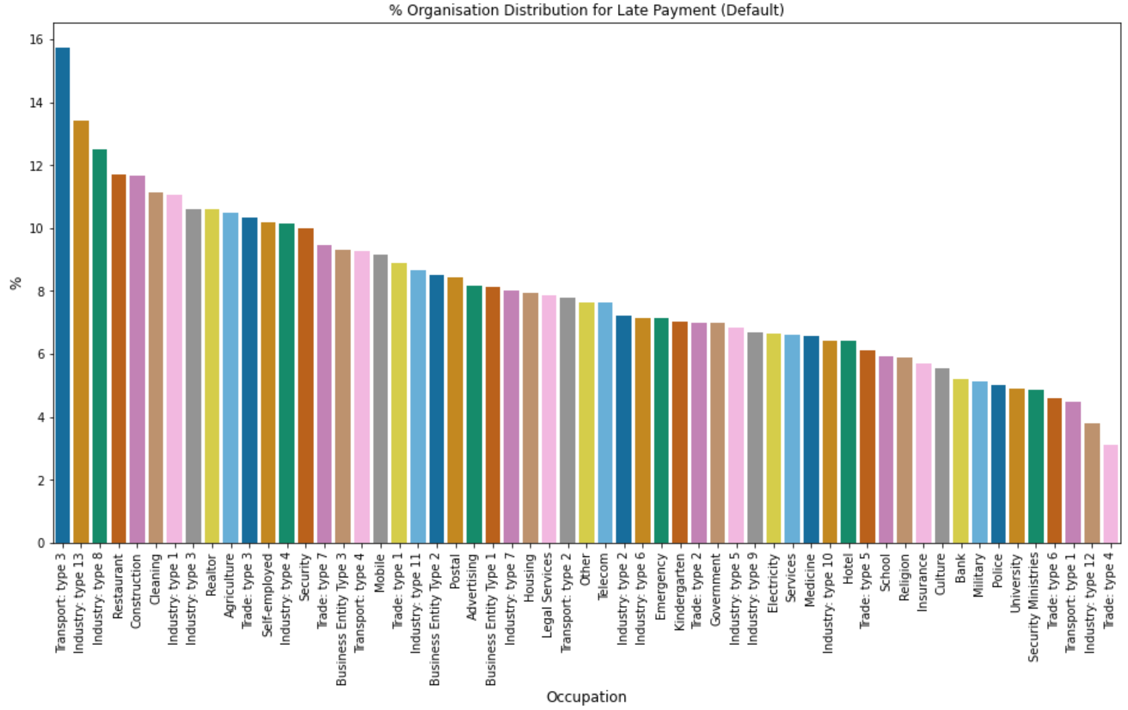


Organisation Type vs. Target Variable

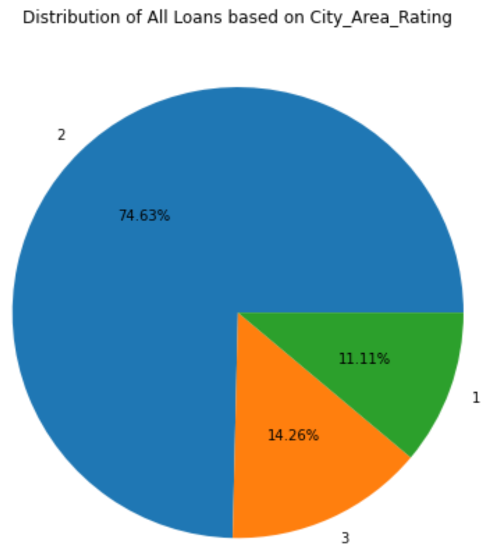
Findings

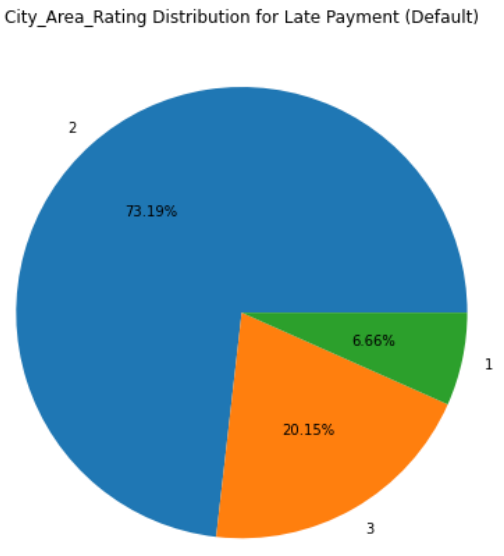
- Organisations with the highest % of Defaulters are Transport: Type3 (16%), Industry: Type 13(13.5%) and Industry: Type 8(12.5%).
- Business Entity: Type 3, Trade: type 4 and Industry Type: 12 are the most reliable to be given loan to, as they have a relatively low default compared to loans or low % of loan defaults

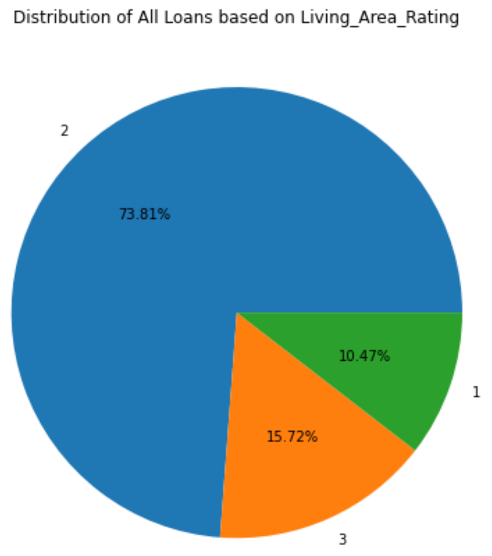


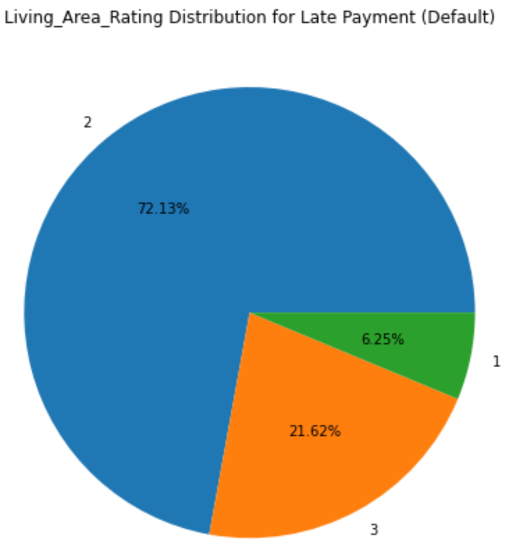


Area Rating vs. Target Variable









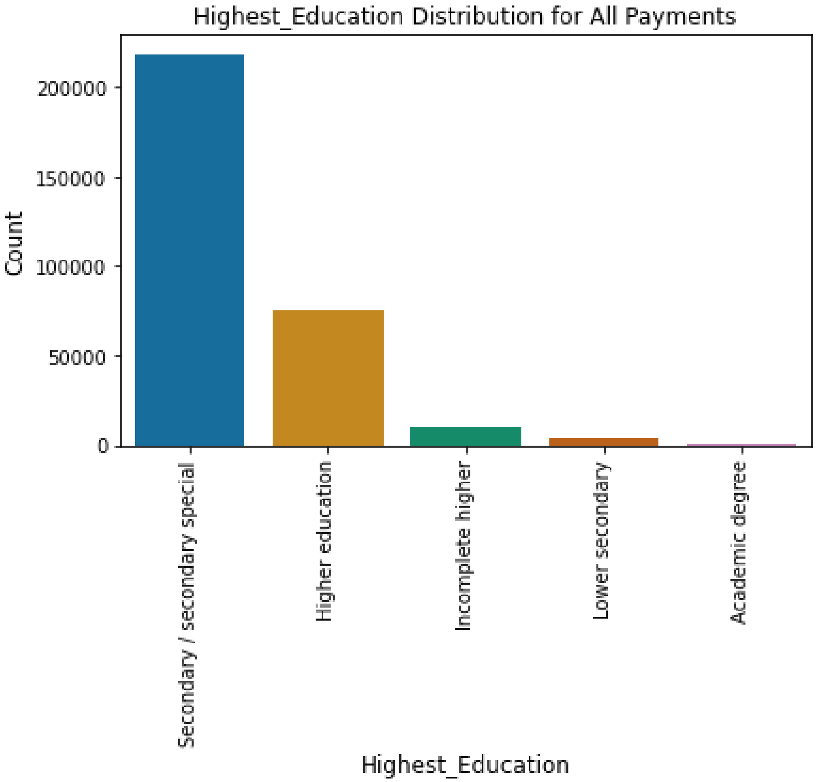
Findings

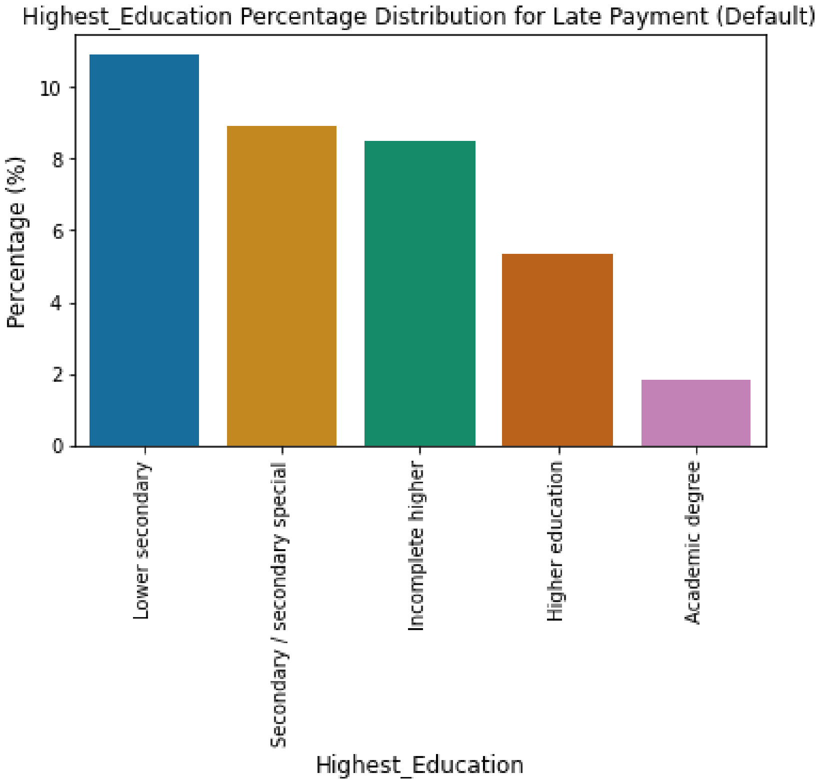
- City rating category 2 take the most loan and also have the most default.
- The default % of rating category 3 is higher than In the all loans
- Category 3 City area rating are risky.

Findings

- Living are rating category 2 take the most loan and also have the most default, the default % of rating category 3 is higher than In the all loans
- Category 3 living area rating are risky.

Education vs. Target Variables





Findings

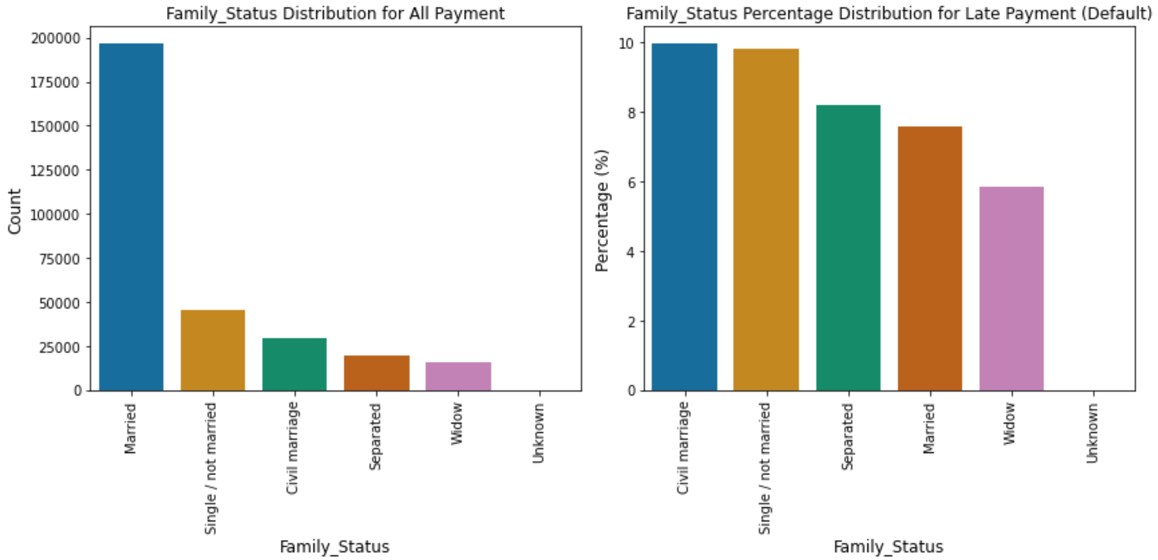
- clients with a secondary education contribute the most to default (10%)

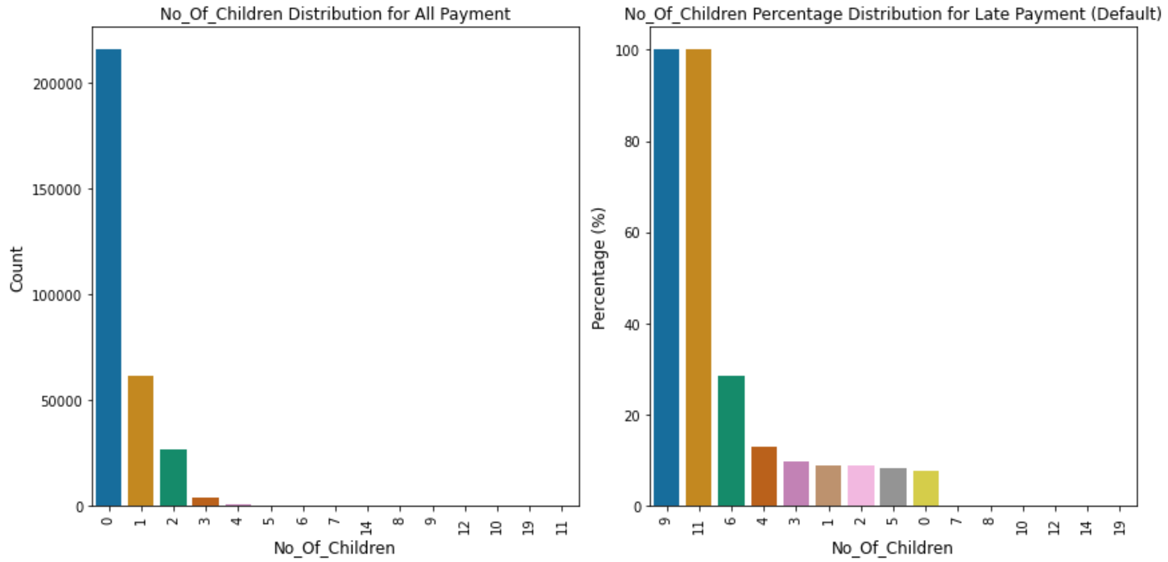
NUMERICAL ANALYSIS

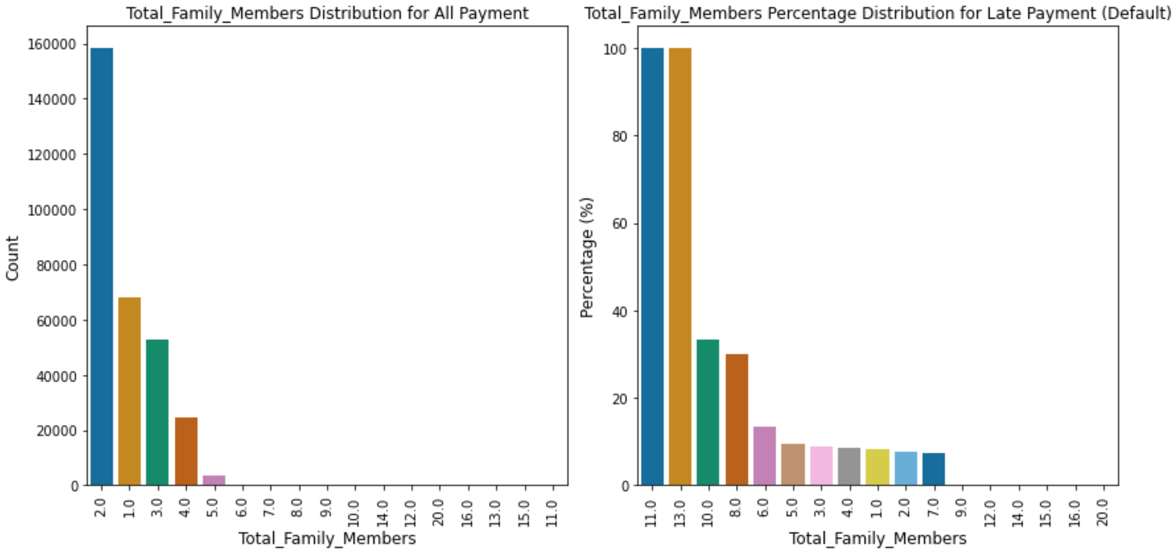
Family Features

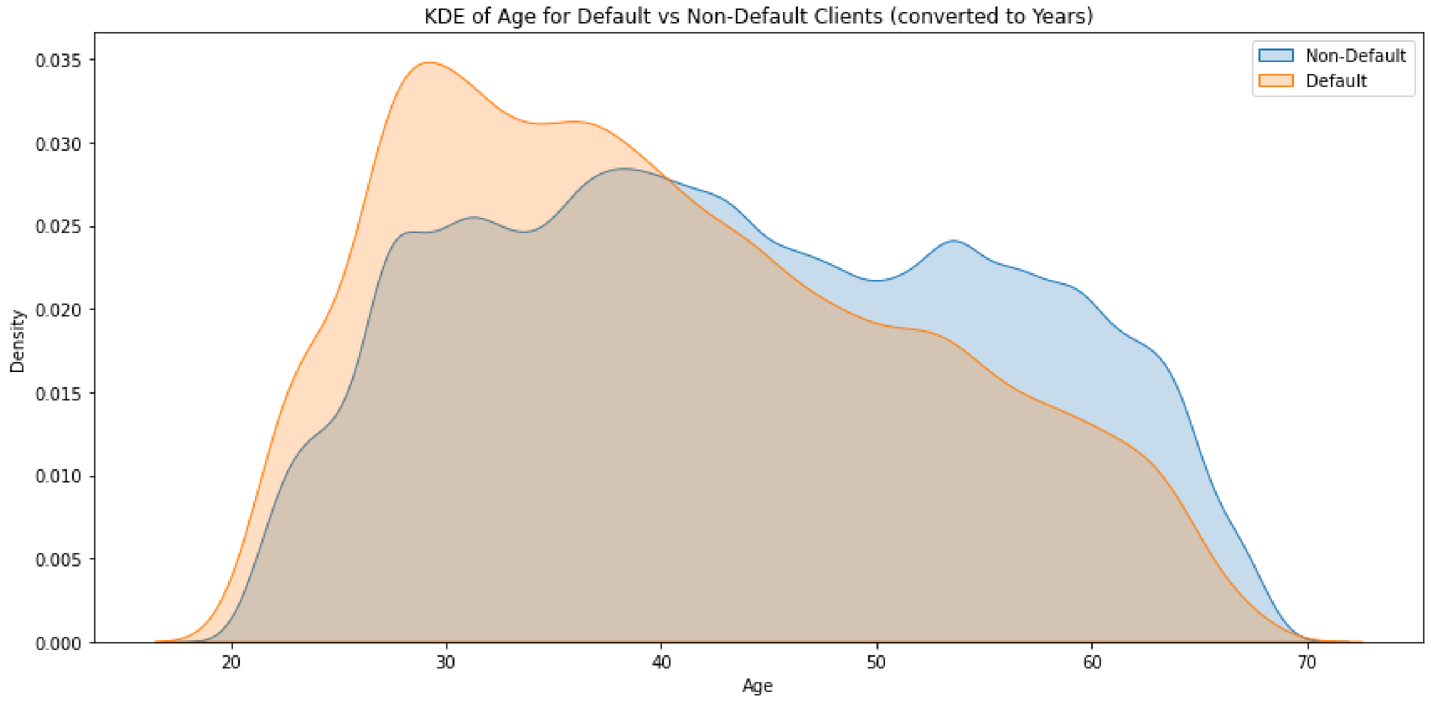
Findings

- Married clients are take the most loans, Civil marriage have the most defaults (10%), followed by single clients (almost 10%)
- Clients with no or few child will likely to pay their loan on time.
- Clients with higher number of children/total family members are riskier.









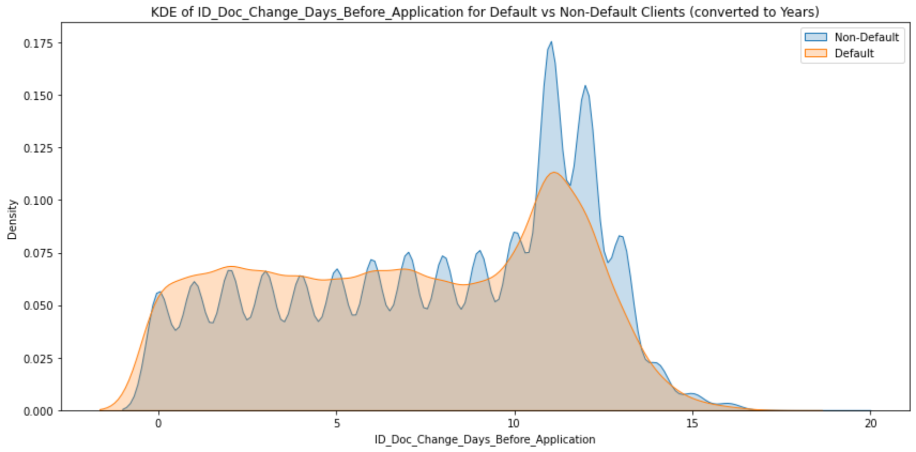
Findings

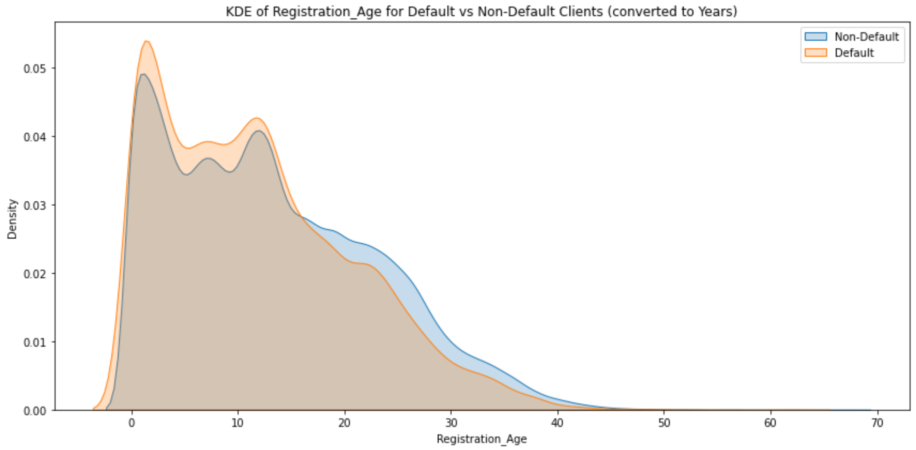
Age distribution:

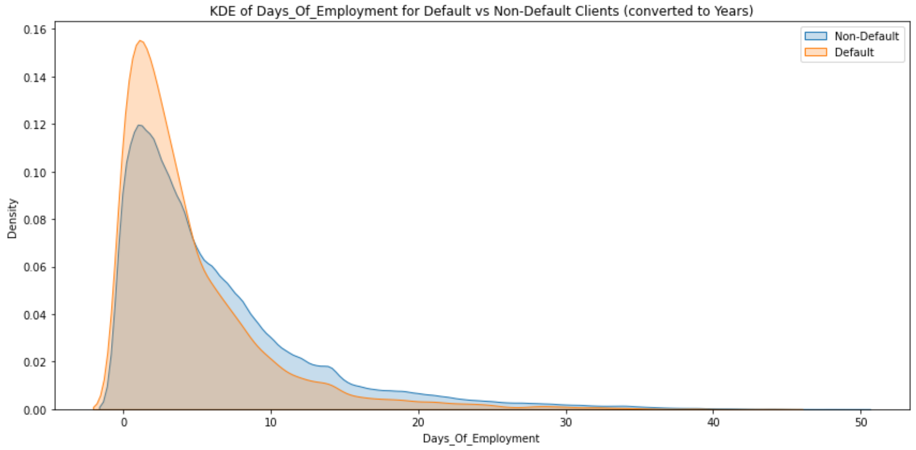
- As you can see, as the clients are getting older, the clients are often pay they loan on time more often.
- Younger client are less likely to pay on time than older clients.

Age Analysis:

- Consistent with our analysis before, where younger age group have higher amount of defaults than older people.
- Maybe is the difference in financial knowledge, guidance for younger people in financing may be required



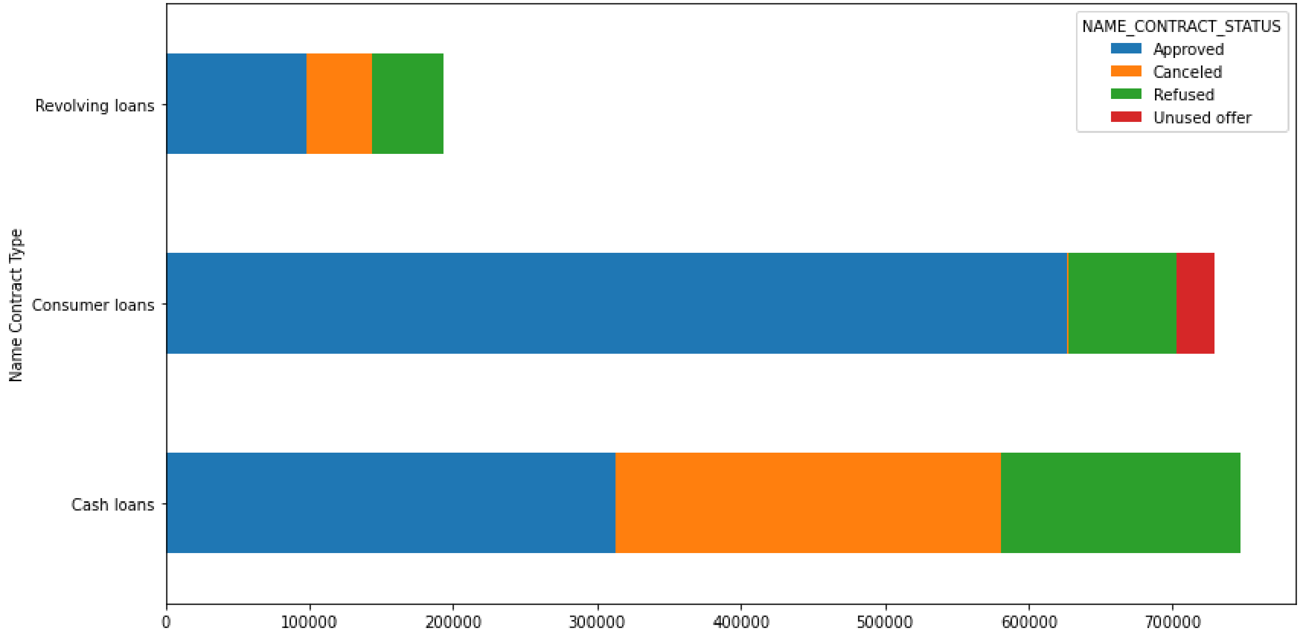


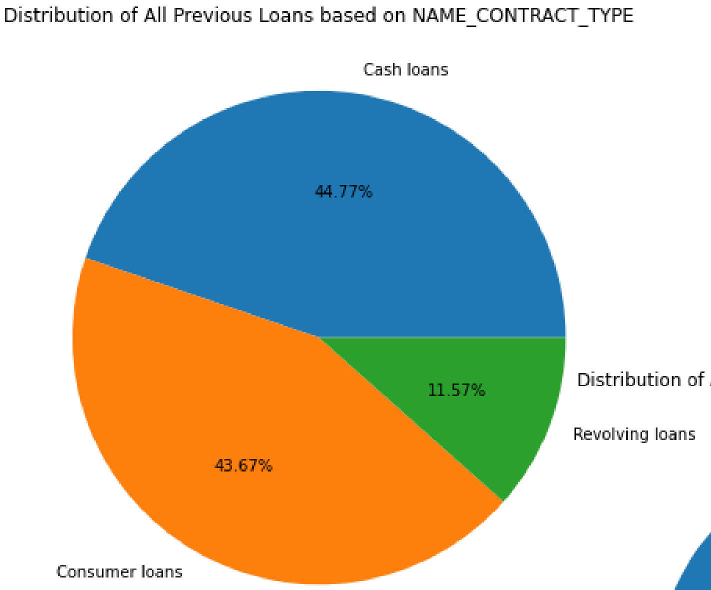


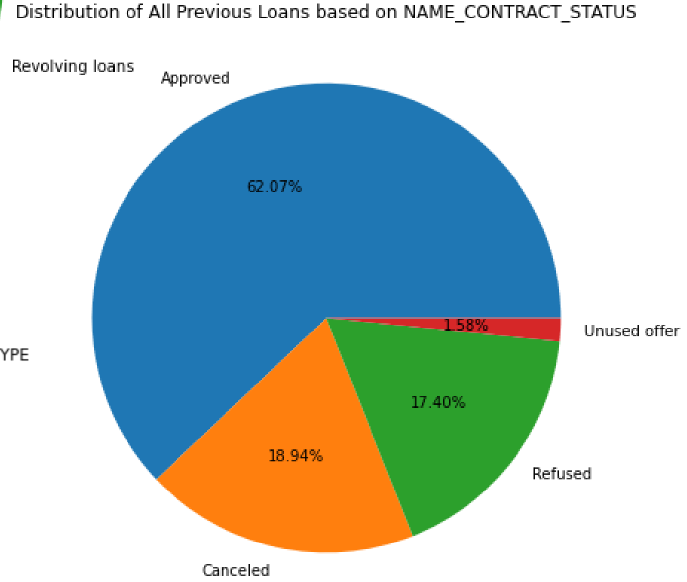
Findings

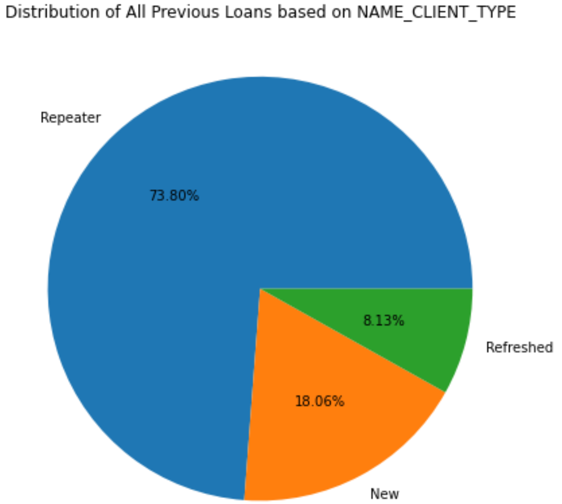
- Clients that change their registration closer to the application are more likely to default

- Clients that change their registration ID closer to the application are less reliable than of those who changed it in advance.
- The more prepared clients, has a higher success in paying back the loan on time.



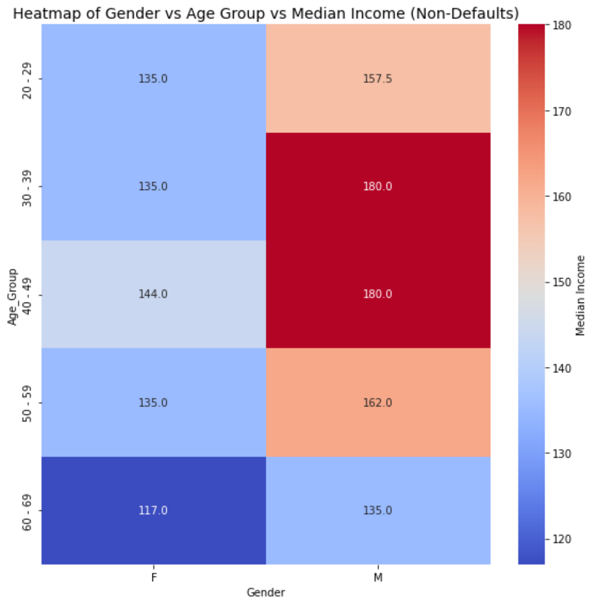


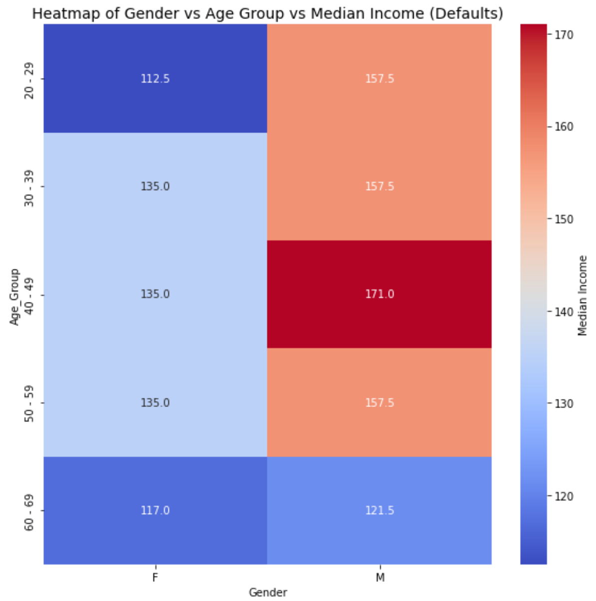


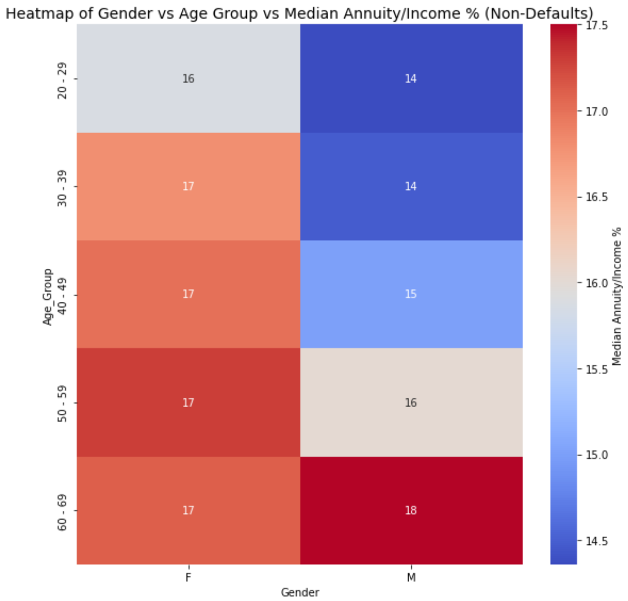


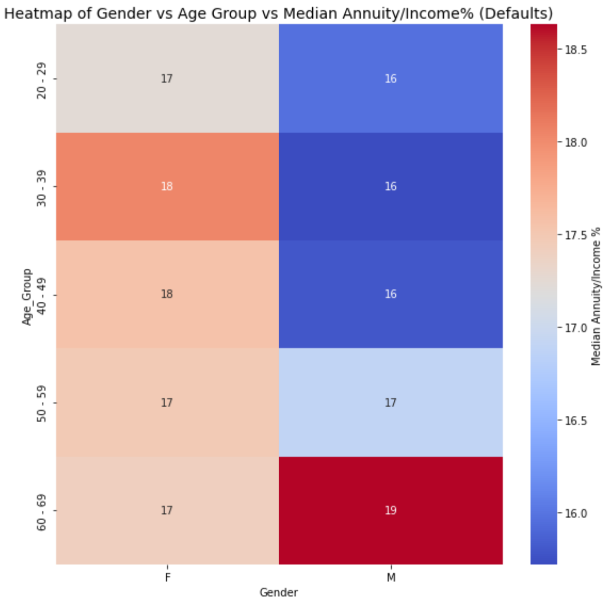
Findings

- Most of the previous loans are approved.
- Most of the previous loan application are Cash Loans followed by Consumer Loans
- Most of the loans are from Repeaters (those who had been loaned before). Only 18.06% are new.
- Consumer loans are most approved and rarely cancel, they are the most reliable types.



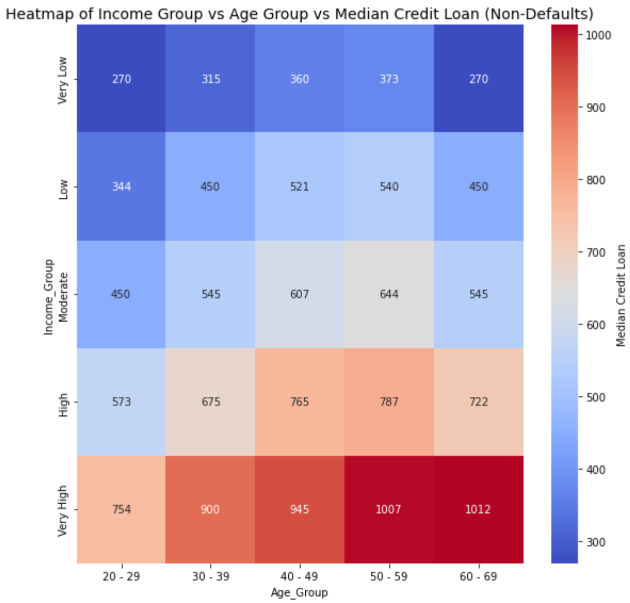


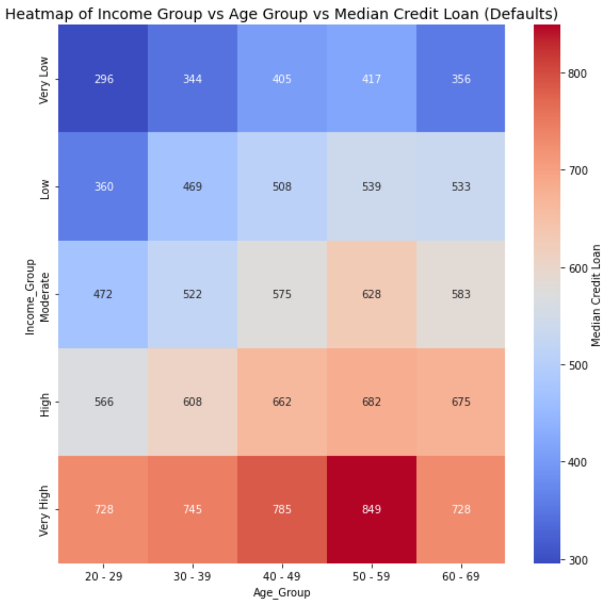




Findings

- As you can see the Median Income for those who are defaulted are lower than those wo are non-defaults.
- The % of Annuity/Income median are higher across Age for defaulters than non-defaulters
- Median Credit Loan of Very low income group of defaulters is high in all age segment as compared to the Non-Default Loans.





CONCLUSION

- Most defaulters are from very low and low income range.
- Younger People are more tend to default.
- Focus more on having client from category 1 & 2 city and living rating area.
- Laborers, Sales Staff, and Drivers are most defaulters.
- Focus less on clients who has Working Income types.
- Attract more repeating clients.
- Approve loan to more prepared clients.
- Clients with more children/family members are likely to default