

Exploratory Data Analysis

Risk Analysis on Bank Loans

Jibin Baby

Problem Statement

This assignment focuses on finding patterns indicating clients' difficulty in repaying loans. The exploration involves understanding risk analytics and variable significance in loan assessments.

The given information includes three datasets:

- **application_data.csv** : which contains the current applicant's information
- **previous_application.csv** : which contains the information about applicant's previous loan history
- **columns_description.csv** : which contains the detailed information for each column for the both datasets.

Assumptions

- **Applicants Provide Accurate Information:**

I'm assuming that loan applicants provide accurate information about their financial status.

- **Consistent Loan Approval Criteria:**


I'm assuming that the criteria for loan approval remain relatively consistent throughout the analysis, and situations like people foreclosing loans are not considered.

Approach & Methodology




Data
Understanding
& Domain
Knowledge

1. Understand columns
2. Get Domain knowledge
3. Get variable importance



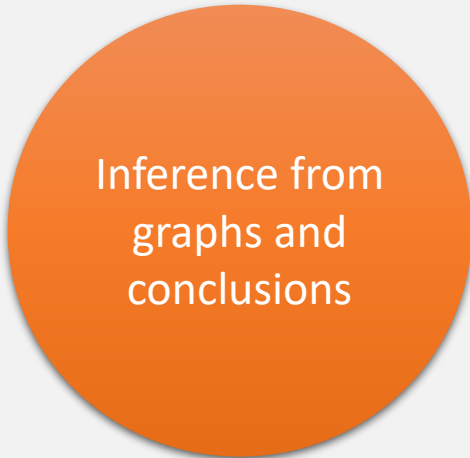
Clean Data
and handle
outliers.

1. Handle missing data
2. Standardize Data
3. Handle Outliers



Univariate,
Bivariate and
Multi Variate
Analysis

1. Perform univariate, bivariate and multivariate analysis on the both data sets, by graphs, plots and heatmaps.



Inference from
graphs and
conclusions

1. Gather insights from the graph and make necessary inferences.
2. Combine these inferences and create analysis report and recommendations

1. Data Cleaning Report

-
- application_data.csv
 - **307511** rows and **122** columns
 - Dropped **77** columns which had either high missing values or were not useful.
 - previous_application.csv
 - **1670214** rows and **37** columns
 - Dropped **13** columns which had either high missing values or were not useful

2. Handling Missing values and Outliers

1. Imputed Mean, Median or Mode

1. Imputed median for values where there are outliers, since the difference between 75th percentile and max value was huge
2. Imputed mean for missing values where there were no big deviation from the max and 75th percentile and the data doesn't have much outliers.
3. Imputed mode for categorical missing values, since the occurrence of the value was high.

2. Removed rows

1. Removed 1 row where the DAYS_PHONE_CHANGE data was empty, since the impact is very low negligible.

3. Binning outliers.

1. Tried not to remove any values for outliers, since they might be valid. For example, age, income and number of children. Therefore, binned the values into different ranges

4. Data Standardization

1. Converted negative values into absolute like DAYS_BIRTH.

3. Graphs and Insights



Univariate Analysis

Categorical Variables In Application Data

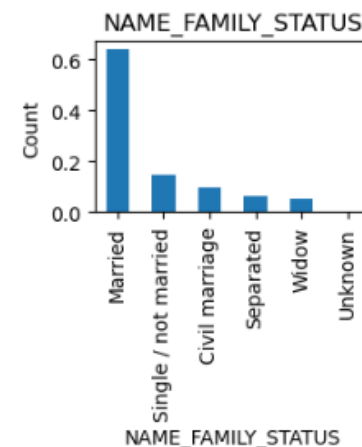
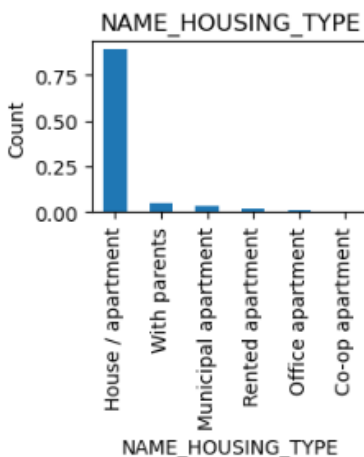
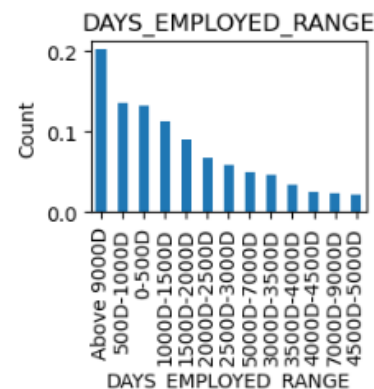
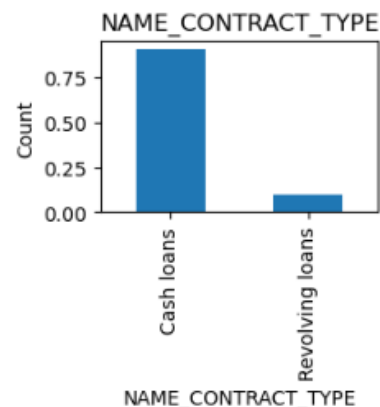
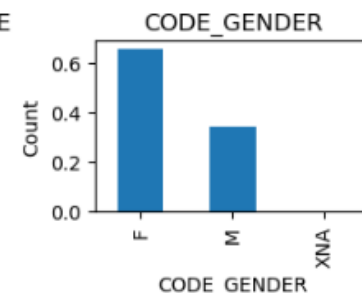
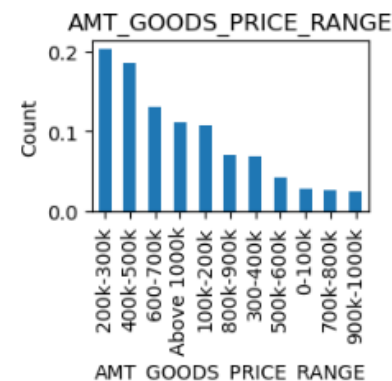
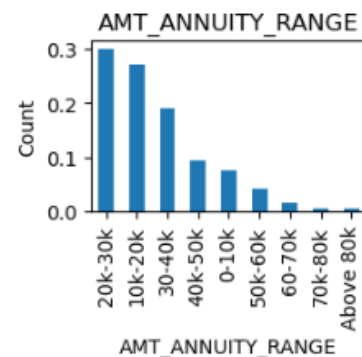
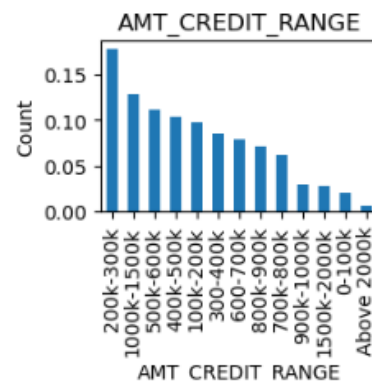
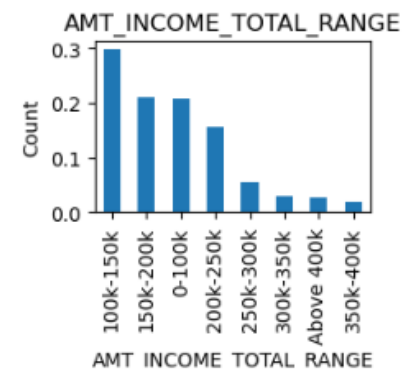
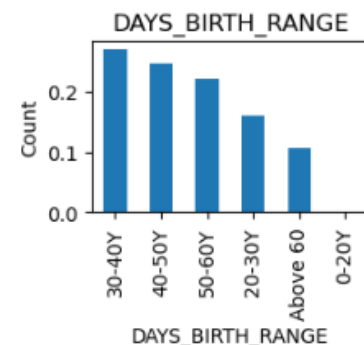
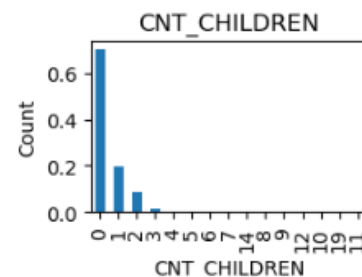
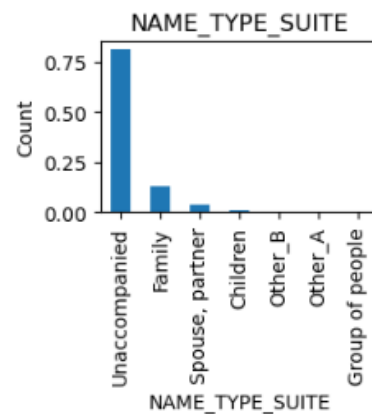
Inference from the graphs

People who applied for the most loan are:

1. People with 0 children (70 percent aprox)
2. People with 100k-150k annual salary (~30%)
3. People with 20k-30k annuity range. (~30 percent)
4. Females with around 65 %.
5. With the age group 40-50 Years. (25 %)
6. People with house/apartment and are married.

EXTRA

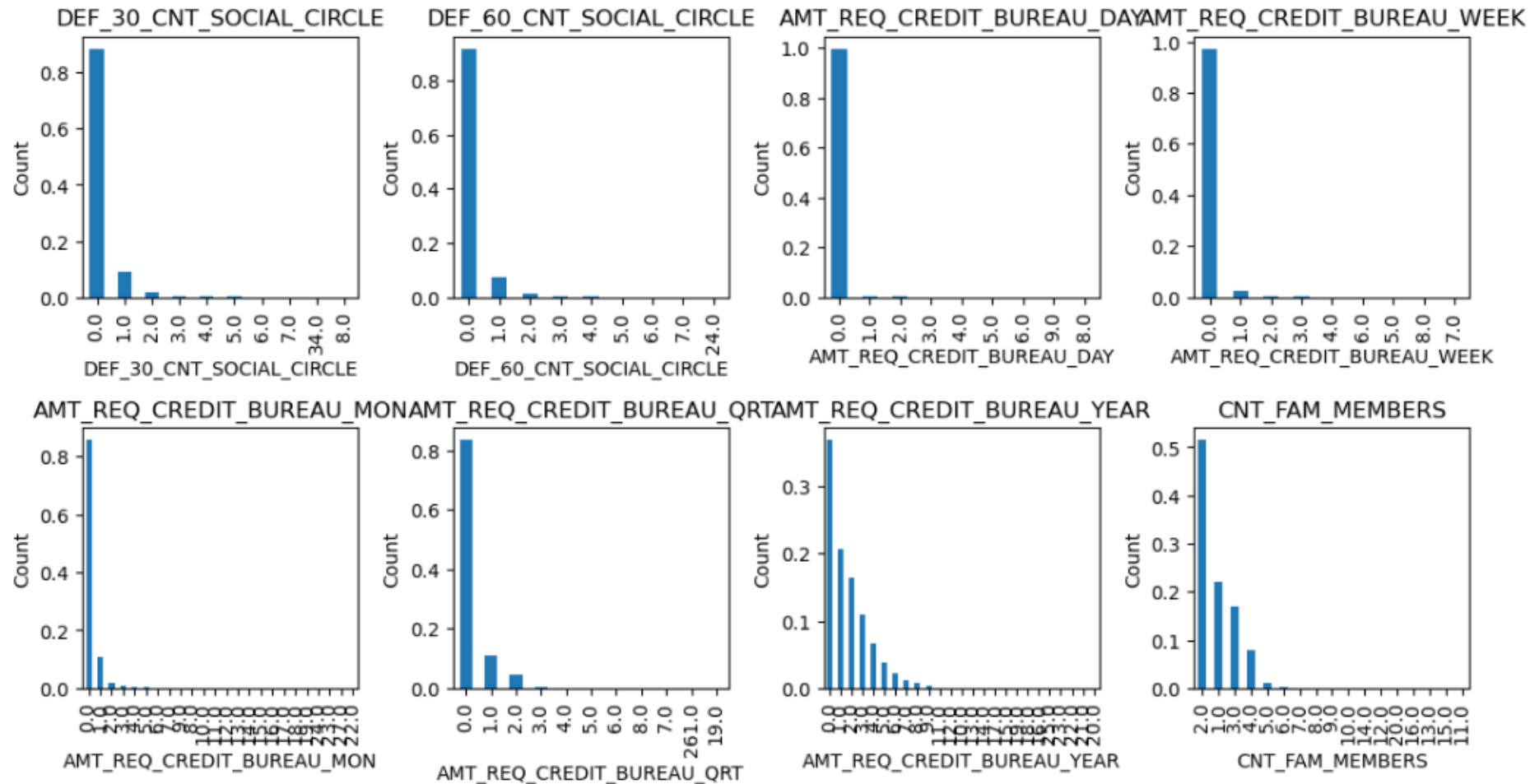
1. Also the average requested credit and goods price is 200k-300k (~18 % and ~20%)
2. Cash loans are the most taken type of loans (90 %)
3. People who are 'Unaccompanied' took the most loan (80%)


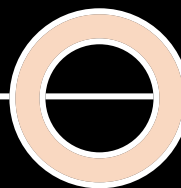



Numerical Variables in Application data

Inference :

- Most defaulters observation is 0 which is a good indicator
- When looking at the number of requests to credit bureau, most of the count is 0 but, in a year, the count is varying
- Most loans were taken by people with 2 family members.





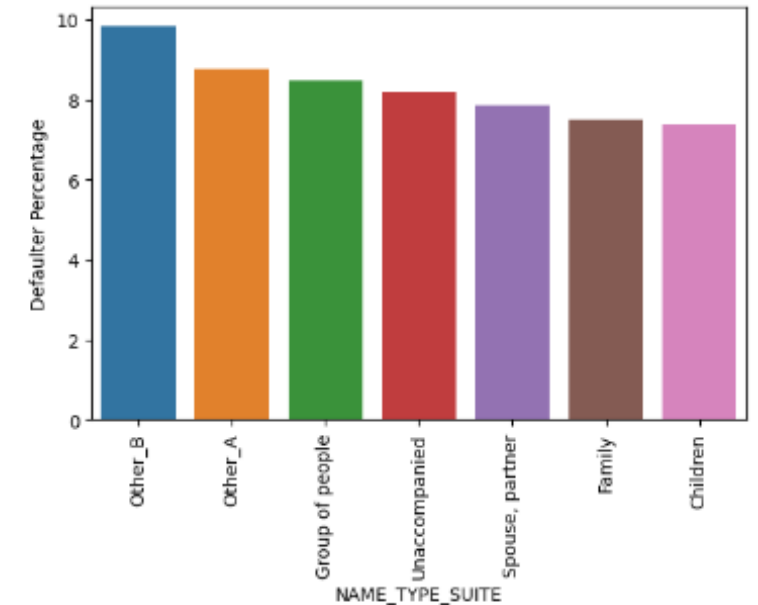
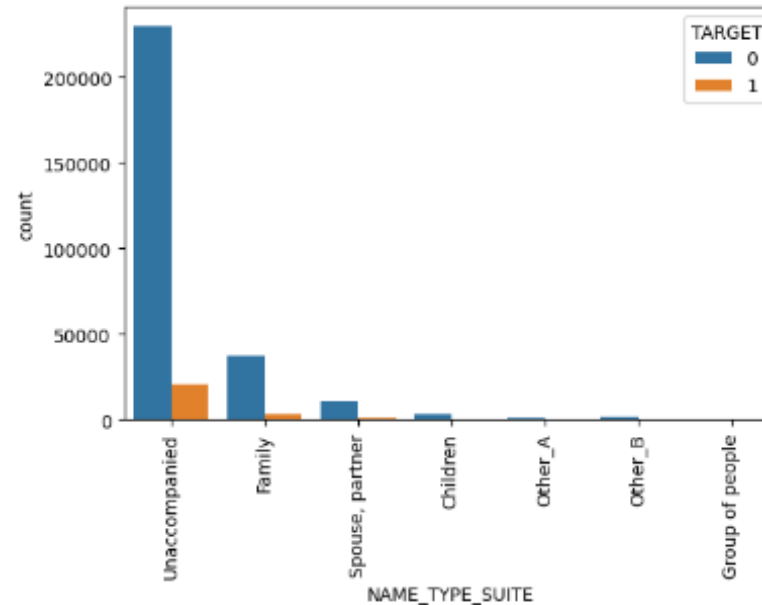
Bi - Variate and Multivariate Analysis

Bivariate analysis on application data

Analysis on Categorical Variables

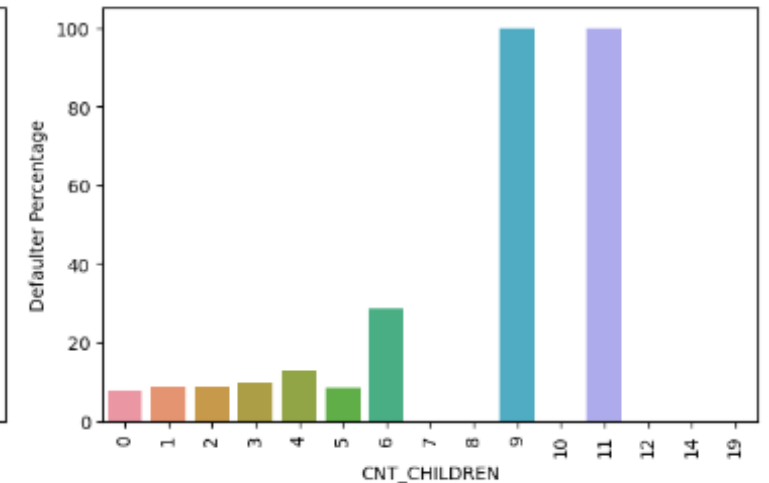
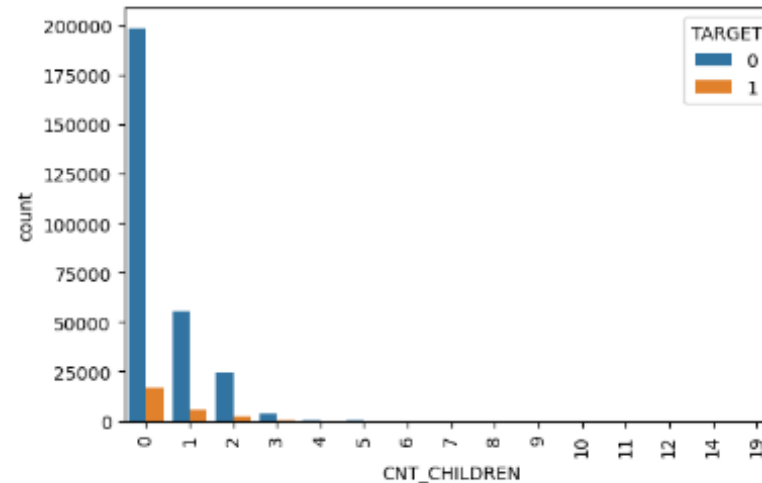
1. NAME_TYPE_SUITE :

- Its evident that unaccompanied took the most loan, and their default rate is ~ 8 %



2. CNT_CHILDREN

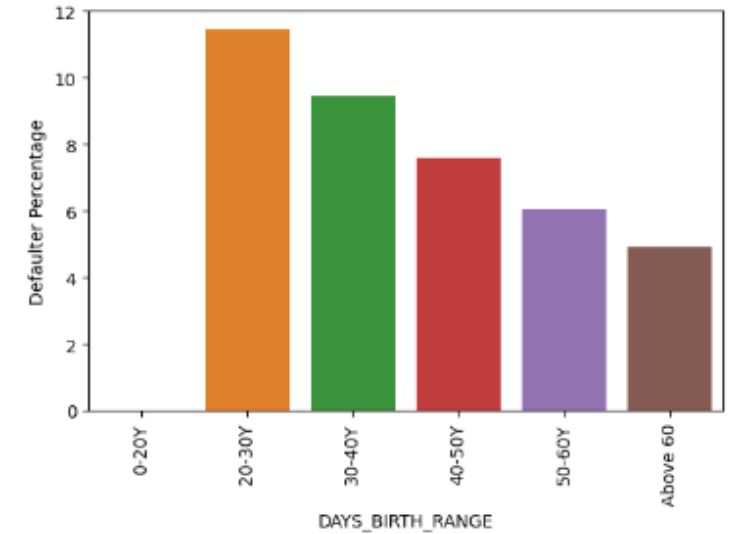
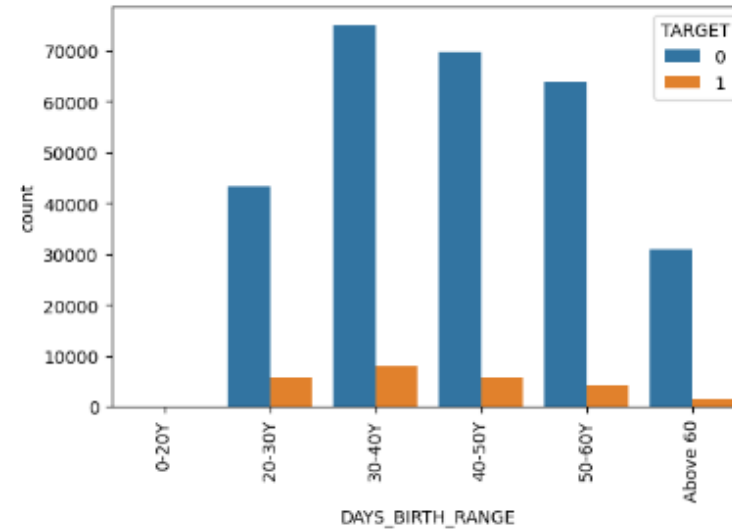
- People with 0 children took the most loan the default rate is low, therefore, they are safe to give loans.
- People with 9 or 11 children are the ones who are most likely to turn out to be a defaulter.



Analysis on Categorical Variables

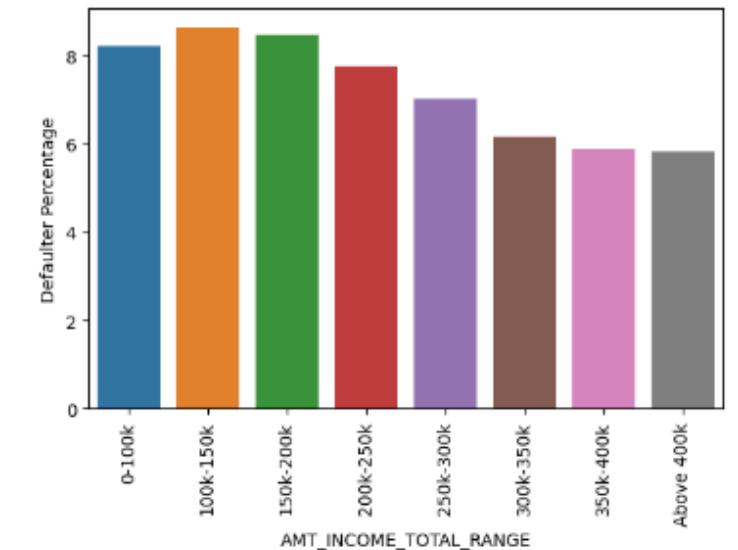
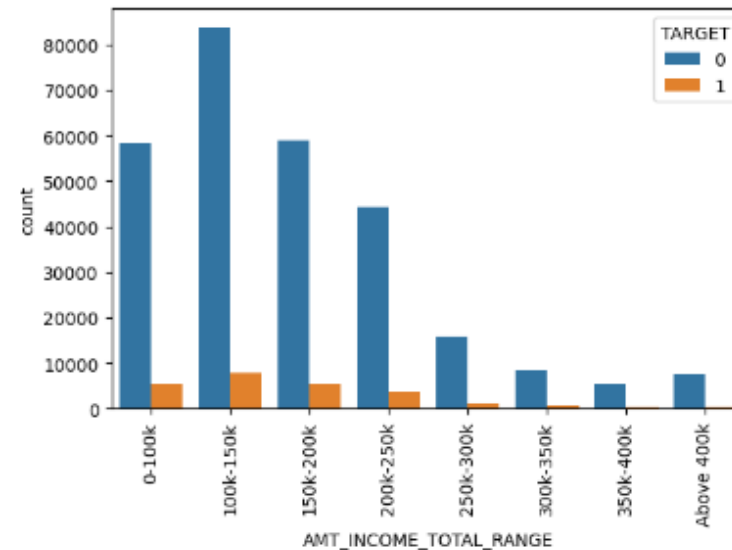
3. DAYS_BIRTH

- People in 20-30 Years of age has the highest default rate
- One more observation is that the default rate is decreasing with increase in age.



4. AMT_INCOME_TOTAL

- People with 100-150k income are the ones who took the most loans but their default rate is also high (~8.5%)
- People with 200-250k income are safe to give loans since their default rate is comparatively low



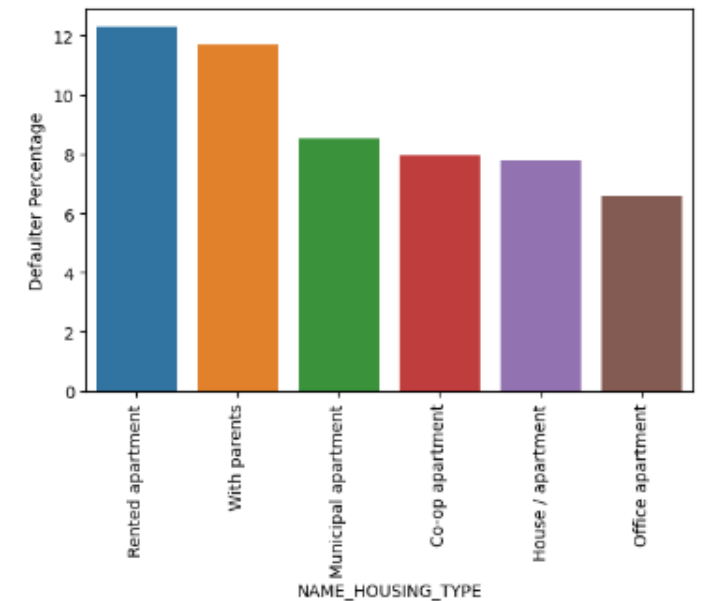
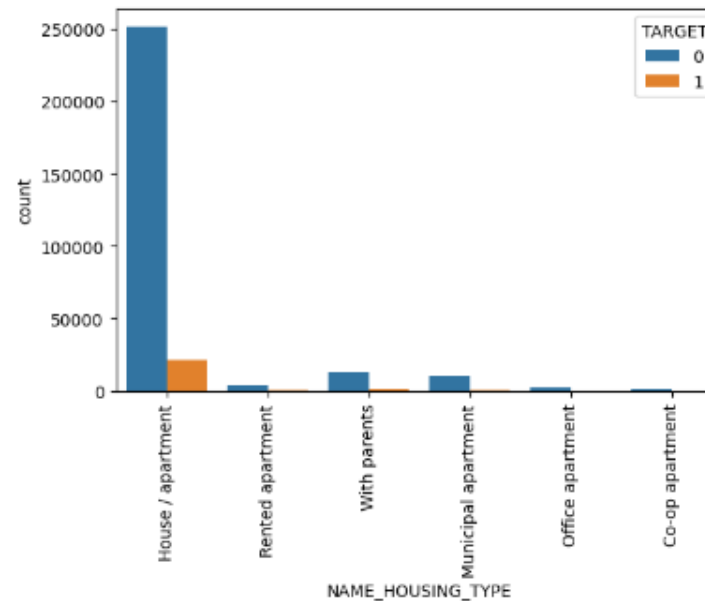
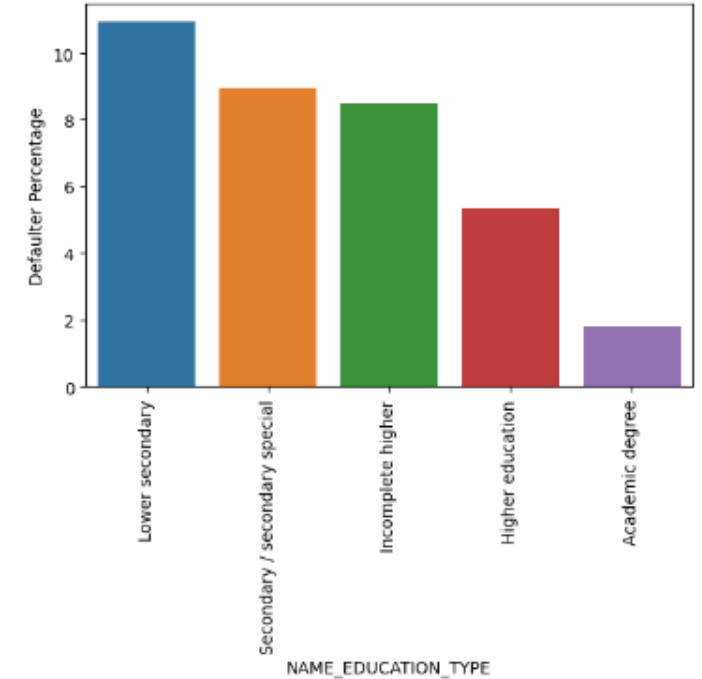
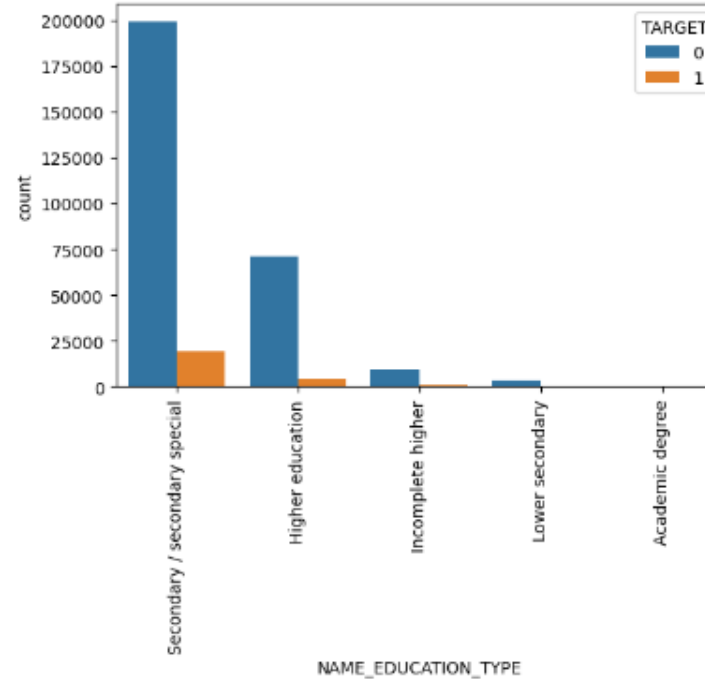
Analysis on Categorical Variables

5. EDUCATION TYPE

- People with Higher education has very low default rate
- Secondary education took the most loan, but they have high default rates.

6. HOUSING TYPE

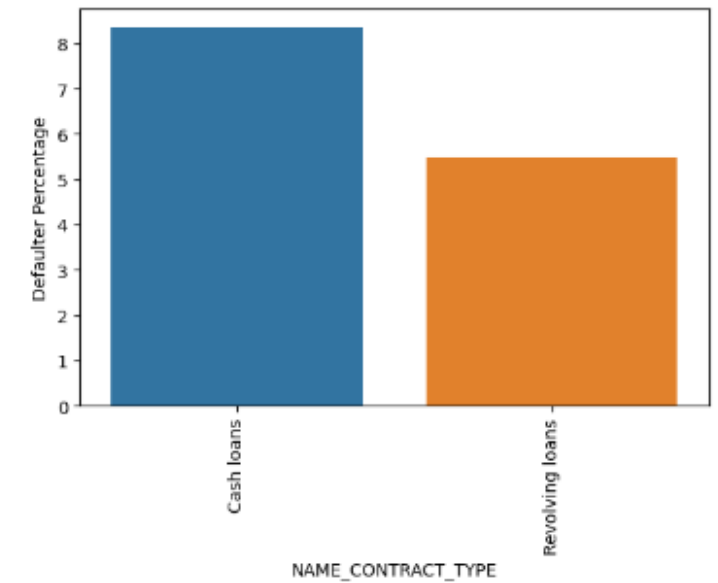
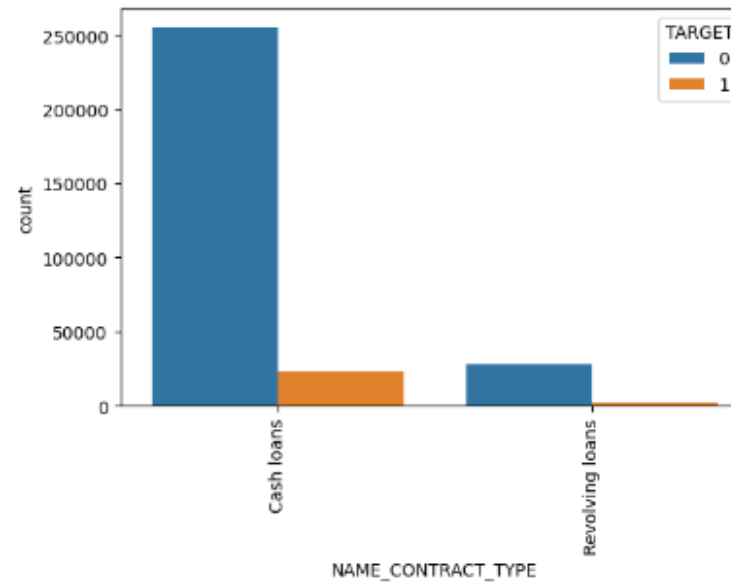
- People with House or apartment took the most loan and they have very low default rate as well
- People with rented apartment have the highest defaulters followed by people who live with parents.



Analysis on Categorical Variables

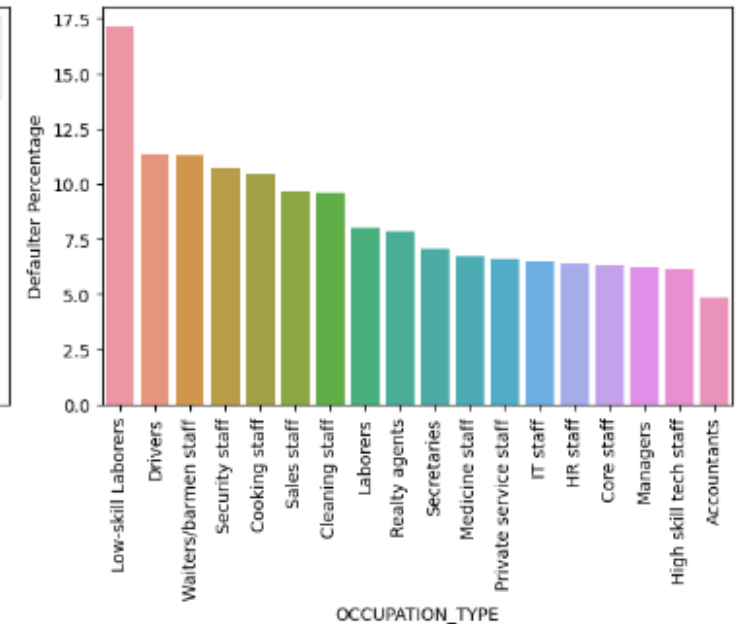
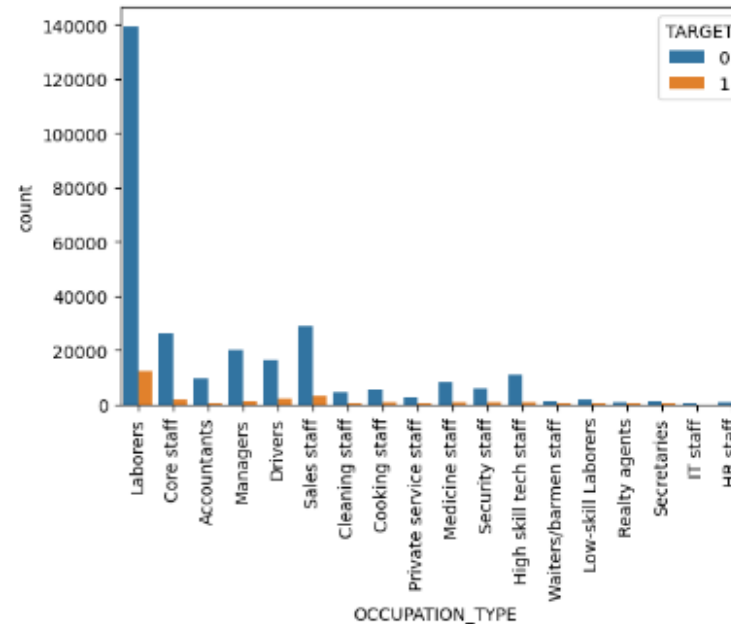
7. CONTRACT TYPE

- Cash Loans are having the highest default rate at ~8.5 %
- Most of the loans are also taken as CASH LOANS.



8. OCCUPATION

- Laborers are the most loan takers, and their default rate is comparatively low at 9 percent
- Low Skilled laborers have the highest default rate



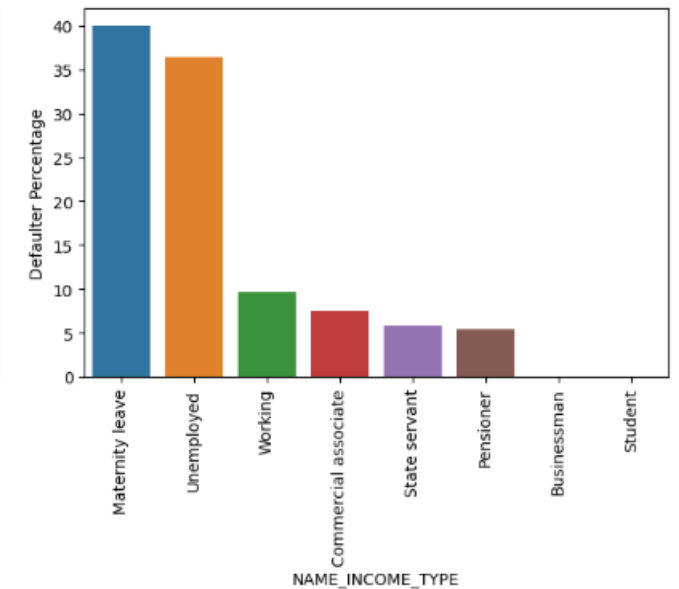
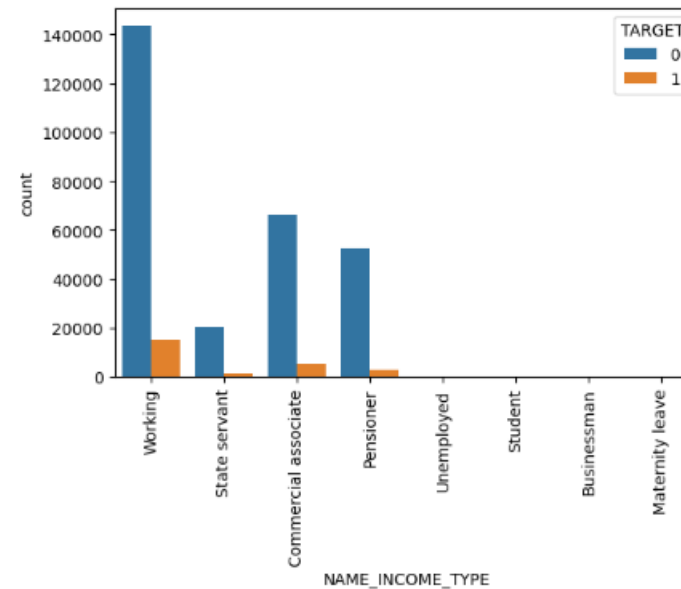
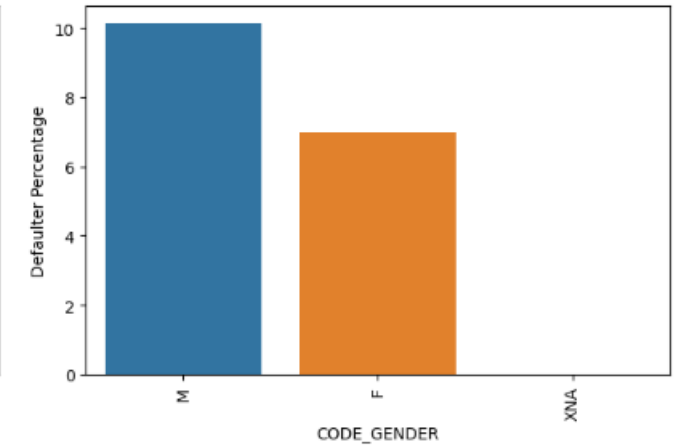
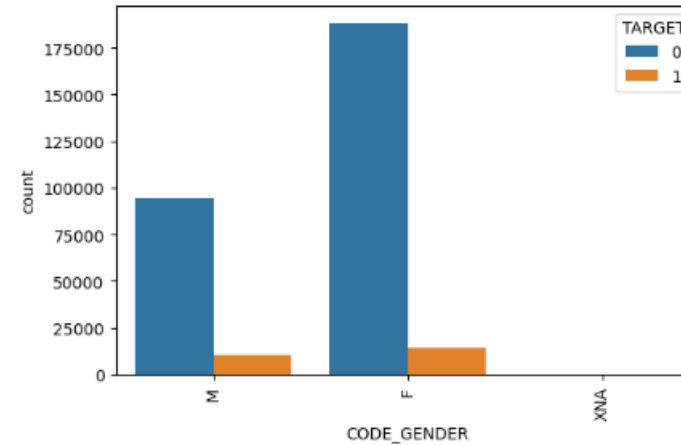
Analysis on Categorical Variables

7. CODE_GENDER

- Males have the highest default rate at ~10% while female default rate is at ~7%

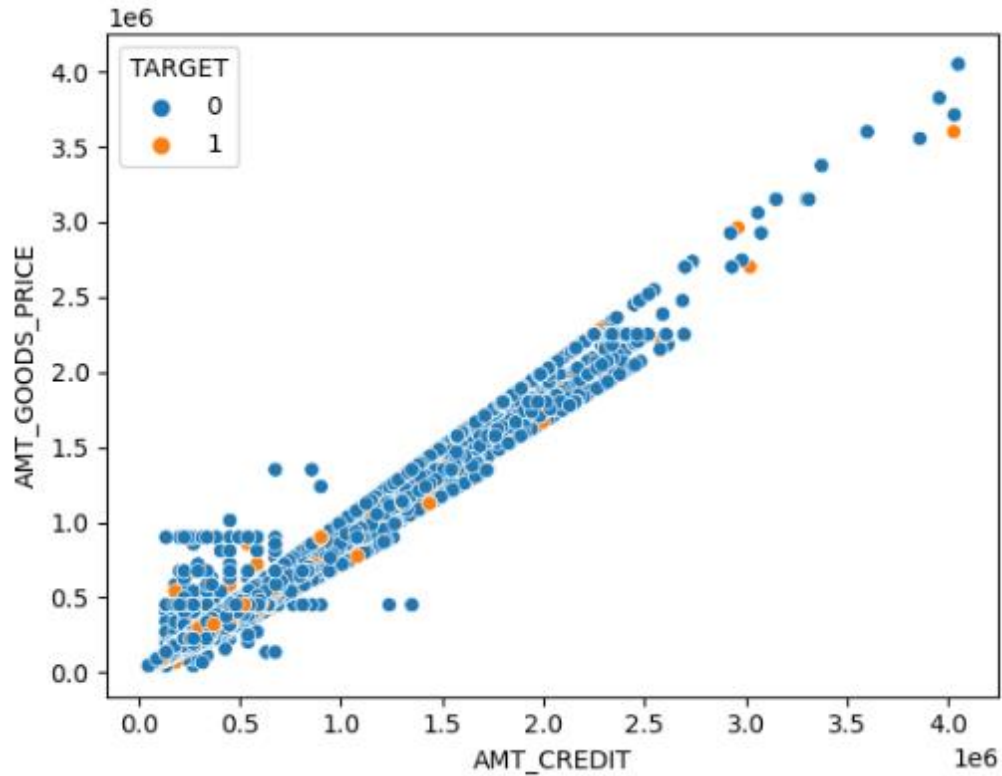
8. NAME_INCOME_TYPE

- People who belong to "working" as income type took the most loan and their default rate is comparatively low
- Commercial associates are also safe to give loans as their default rate is at ~8%
- People with "Maternity leave" as income type has the most defaulters, but they have taken the least loans.
- Unemployed category also has more default rate, so not safe to give them loans.



MULTIVARIATE ANALYSIS USING SCATTER AND BAR PLOT

Application Data

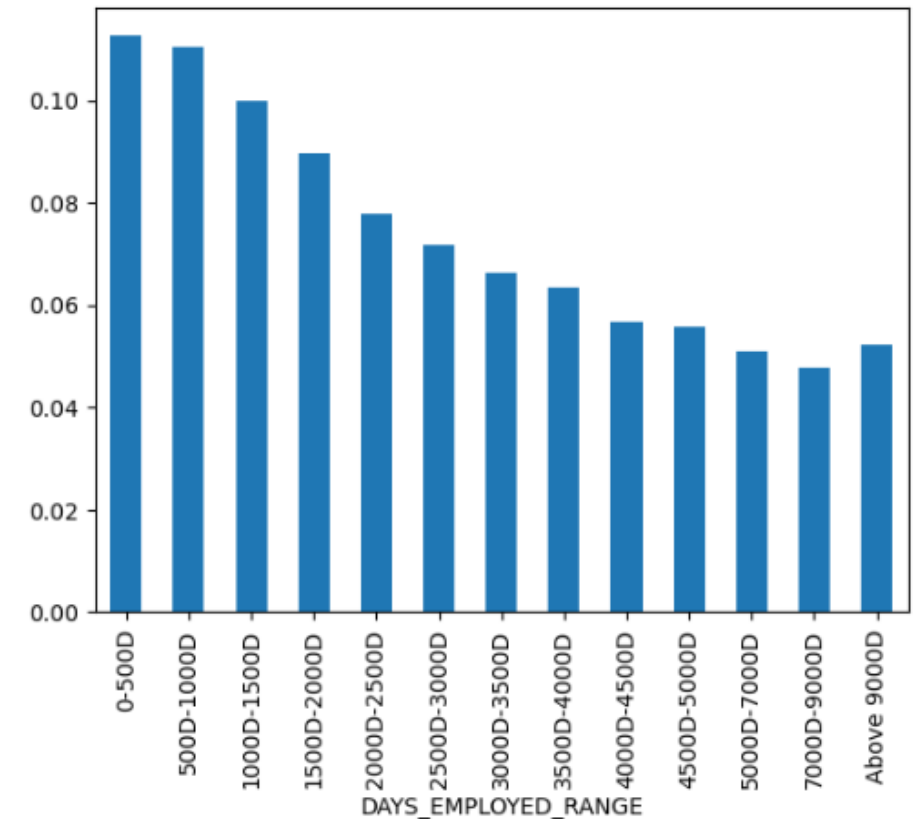


Inference

- AMT_CREDIT and GOODS PRICE have a strong relation and its obvious
- Also, the defaulters are concentrated between 1 and 2 million

Inference

- Default rate is decreasing with increase in DAYS EMPLOYED



MULTIVARIATE ANALYSIS USING HEATMAP

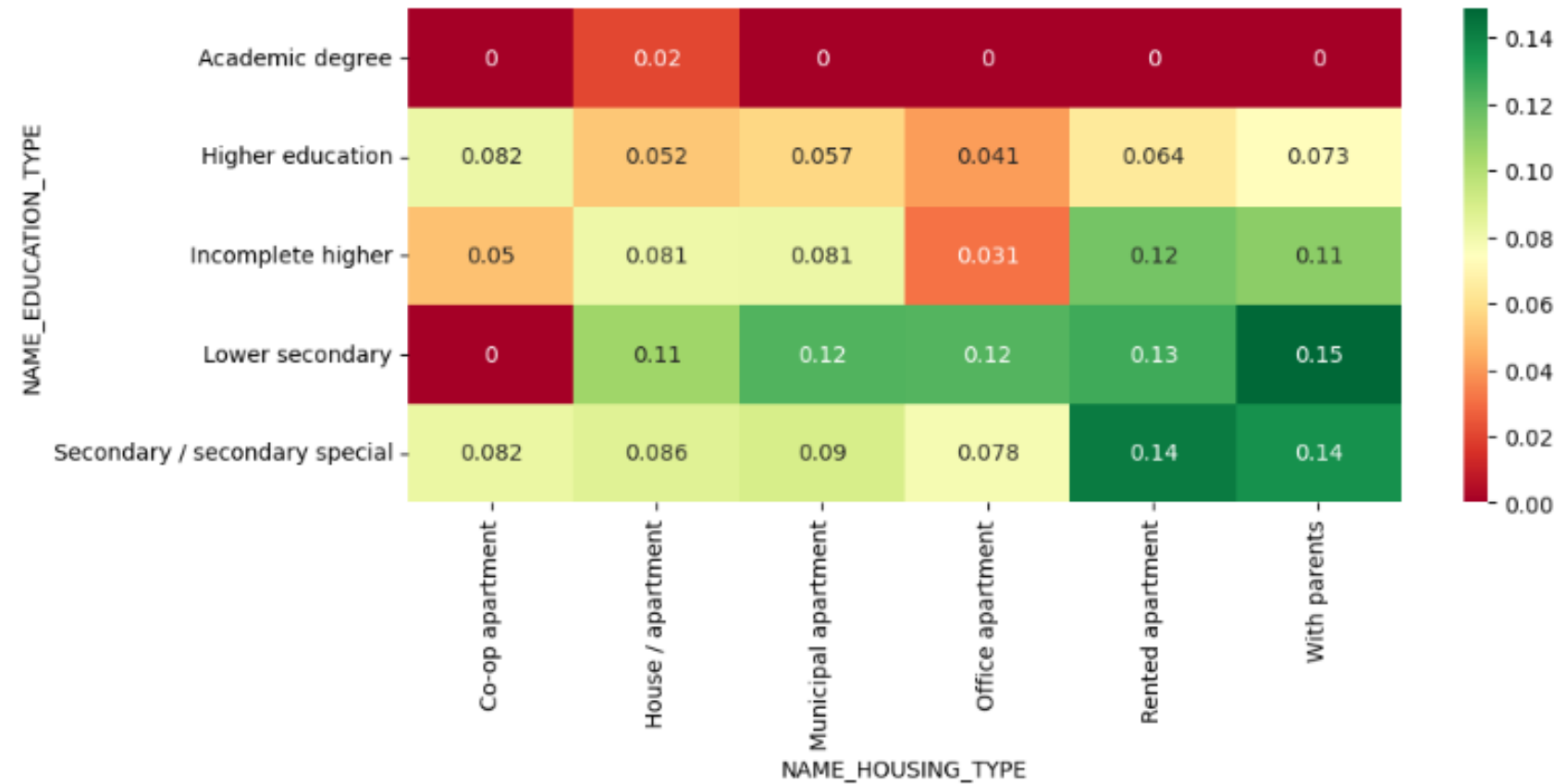
Application Data

- People with lower secondary education and who live with parents have the highest default rate.

- People with Senior Secondary special education with rented apartment also has high default rates.

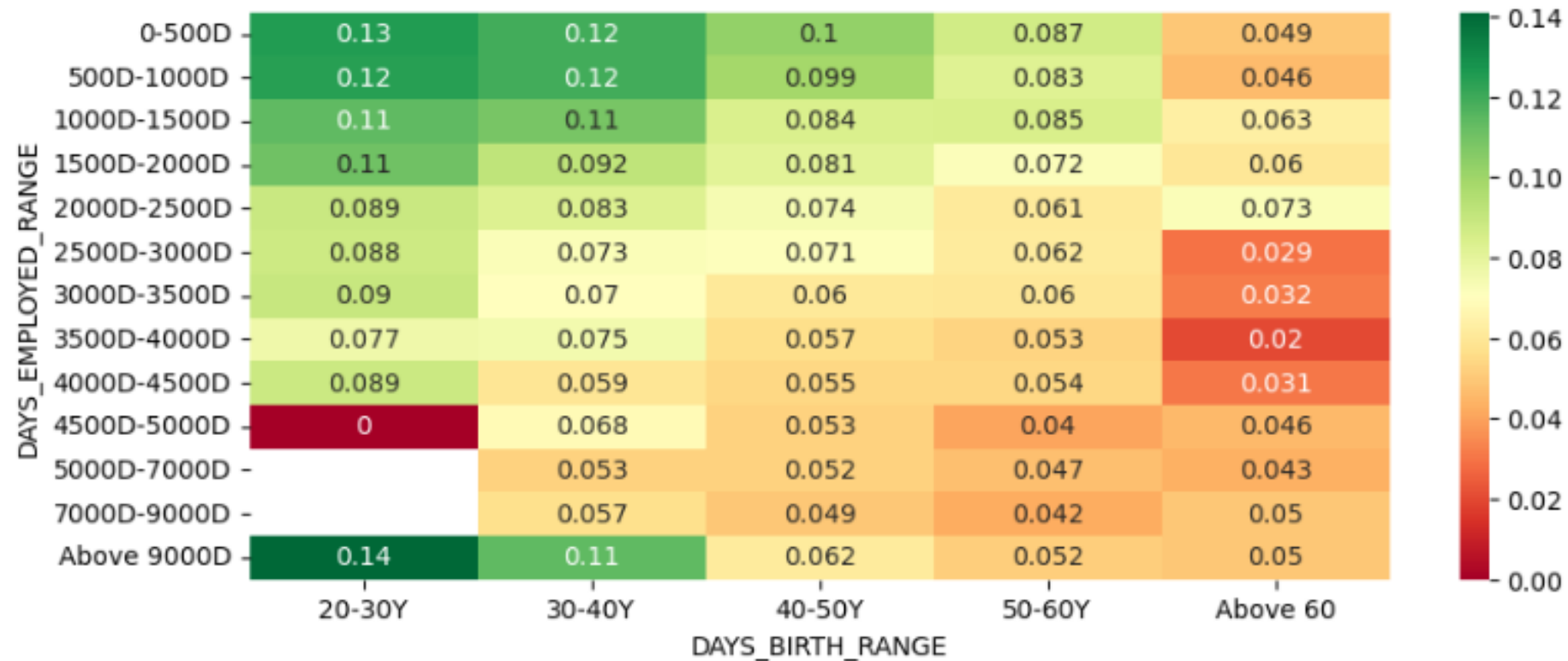
- Generally, people with lower secondary education or secondary education, and with rented apartment or who live with parents tends to show high default rates.

- People with higher education or with academic degree are safe to give loans irrespective of their housing type.



MULTIVARIATE ANALYSIS USING HEATMAP

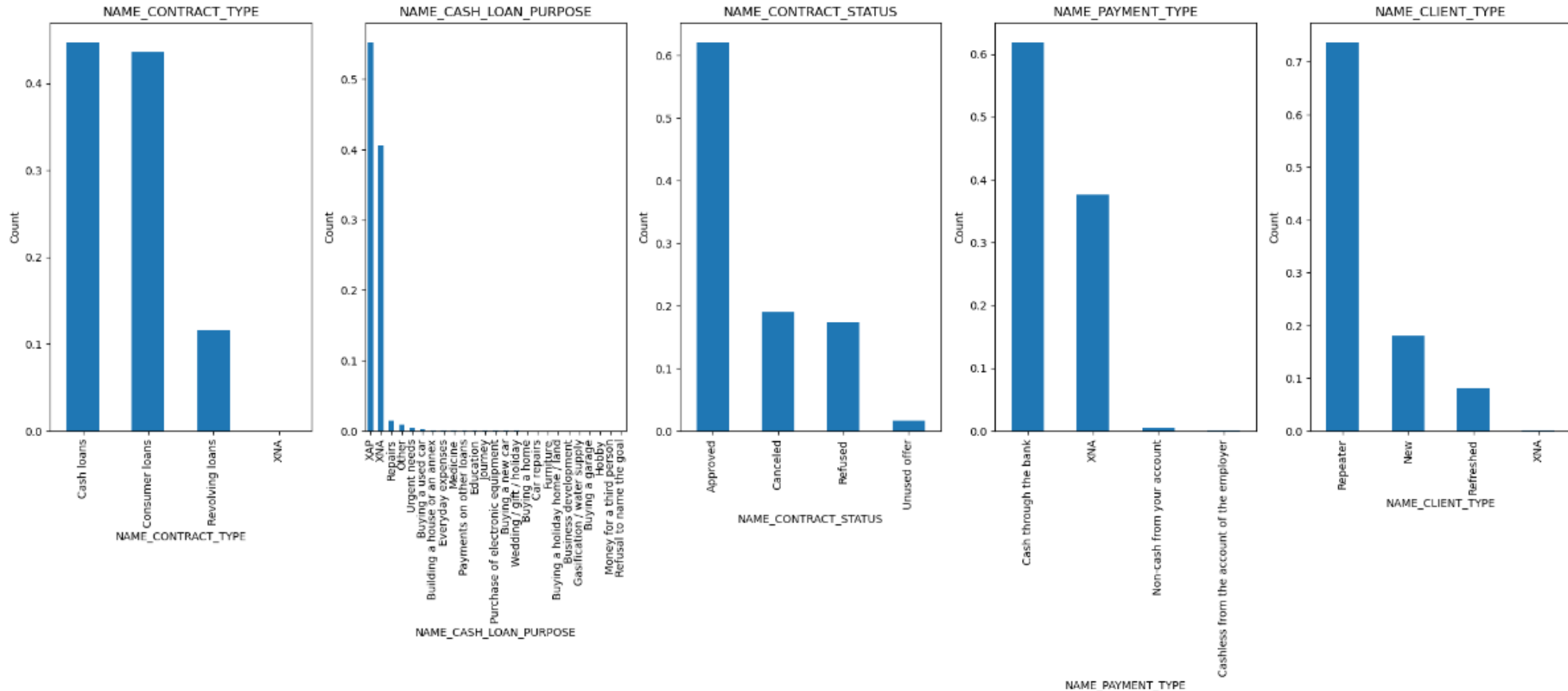
Application Data



- People default rate is decreasing with increase in age as wells increase in days employed.
- Most default rate is for 20-30Y with more than 9000Days of employment (which is impossible)
- People with lower age group and of lower experience are more likely to default.

UNIVARIATE ANALYSIS WITH BARPLOT

Previous Application Data



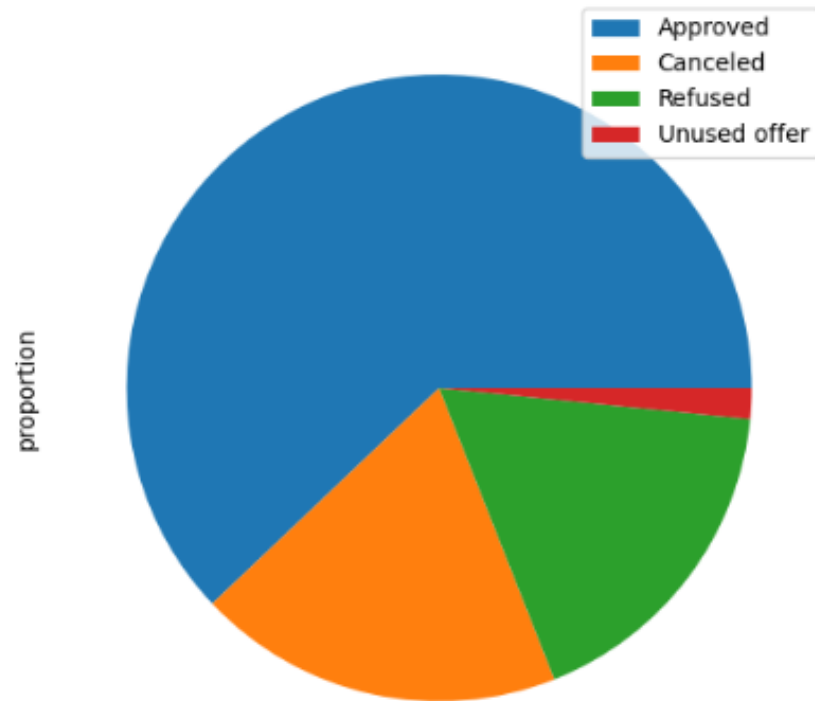
Inference :

- Cash loans and consumer loans were the most taken loans
- XAP category is the major loan purpose followed by XNA (don't really know what this mean by)
- About 62 percent of application were approved , 20 percent cancelled, and 18 percent Refused and around 0.3 percent unused
- Most of the loan takers were repeaters at around 72 percent, and new applications at around 18 percent

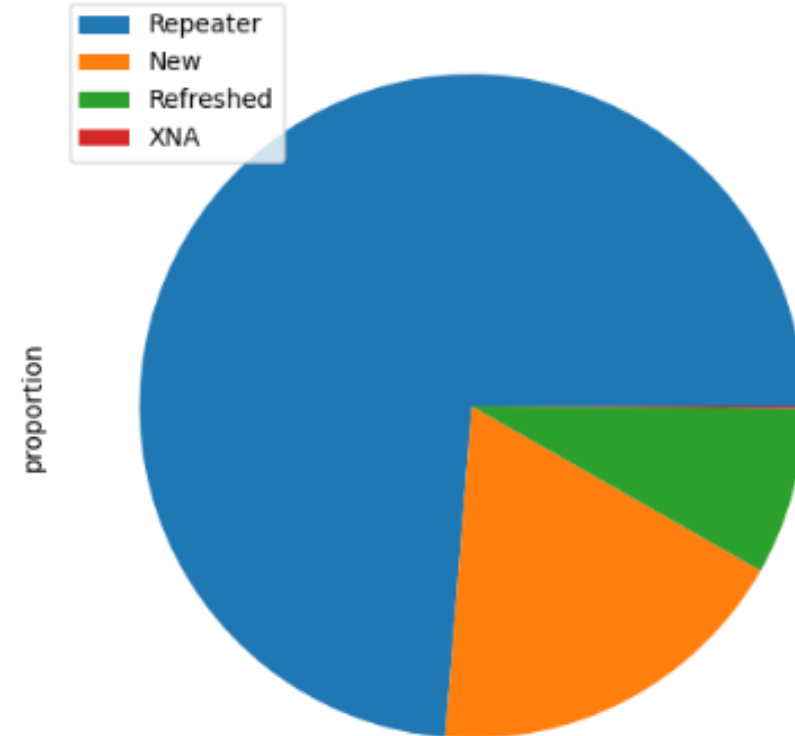
UNIVARIATE ANALYSIS WITH PIECHART

Previous Application Data

```
NAME_CONTRACT_STATUS
Approved      0.621
Canceled      0.189
Refused       0.174
Unused offer  0.016
Name: proportion, dtype: float64
```



```
NAME_CLIENT_TYPE
Repeater      0.737
New           0.180
Refreshed     0.081
XNA           0.001
Name: proportion, dtype: float64
```



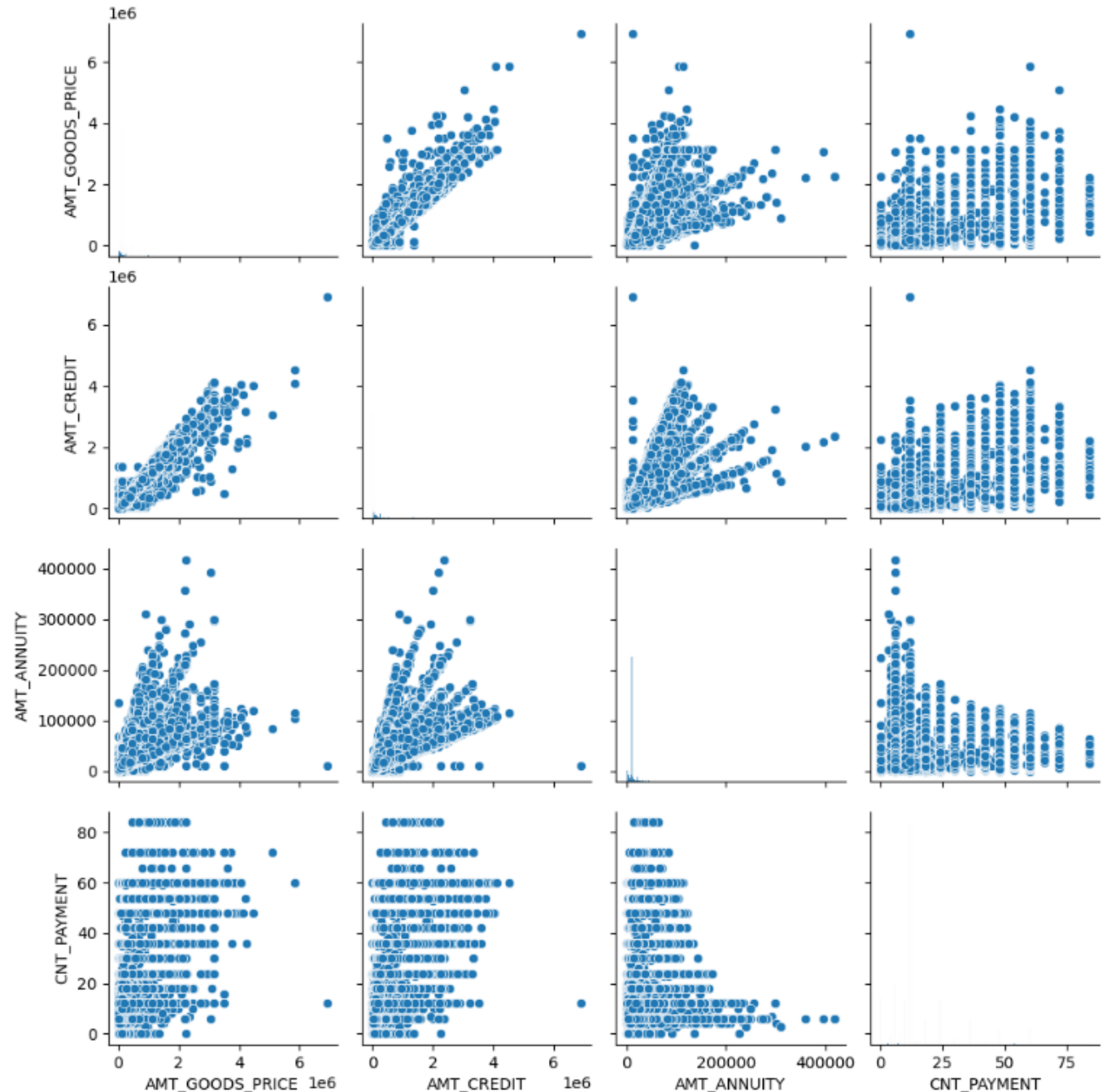
BIVARIATE ANALYSIS WITH PAIRPLOT

Previous Application Data

Inference :

1. AMT_GOODS_PRICE, AMT_ANNUIITY, AMT_APPLICATION - have high correlation and its obvious. Higher the value of good price more there will be need of loan.

2. Similary, AMT_Credit and AMT_GOOD_PRICE also has high correlation.

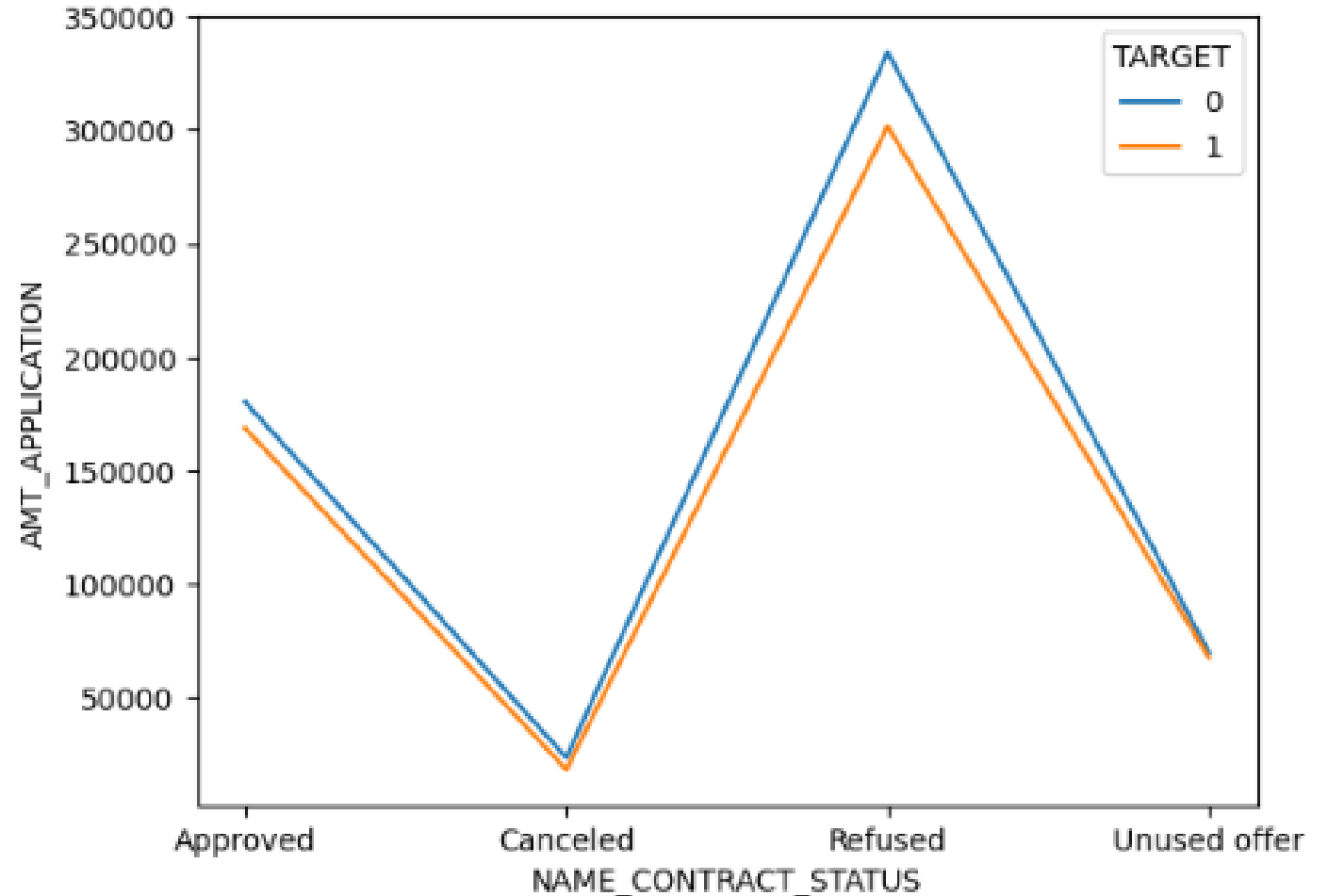


MULTIVARIATE ANALYSIS WITH LINEPLOT

NAME_CONTRACT_STATUS VS AMT_APPLICATION VS TARGET

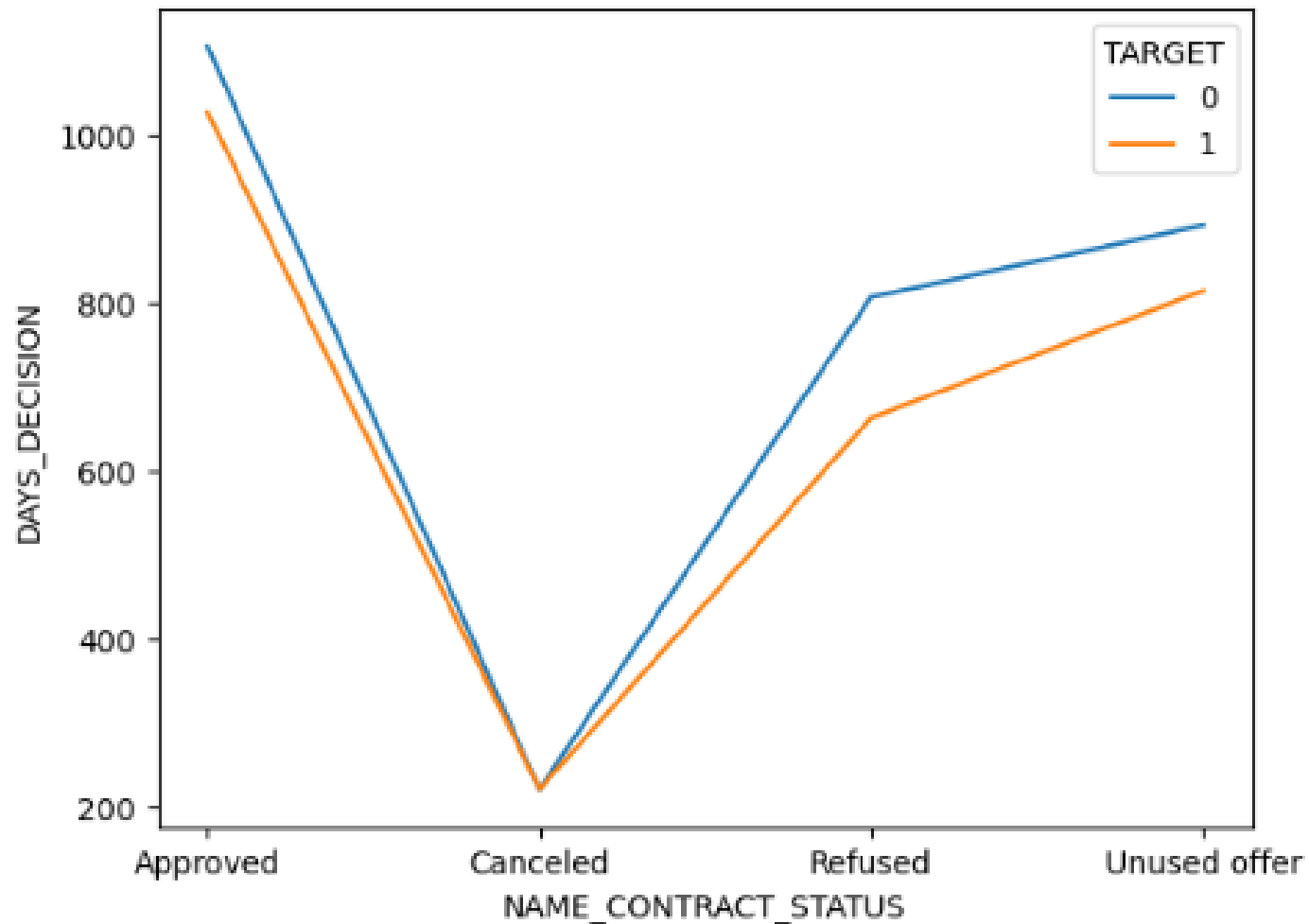
Inference :

- There are more defaulters for higher loan amounts. This could be because borrowers who take out larger loans are more likely to default on their payments.
- The number of defaulters is relatively small compared to the number of re-payers. This suggests that most borrowers are able to repay their loans.
- There are more approved contracts for lower loan amounts. This could be because banks are more likely to approve loans for smaller amounts, as they are seen as less risky.



MULTIVARIATE ANALYSIS WITH LINEPLOT

NAME_CONTRACT_STATUS VS DAYS_DECISION VS TARGET



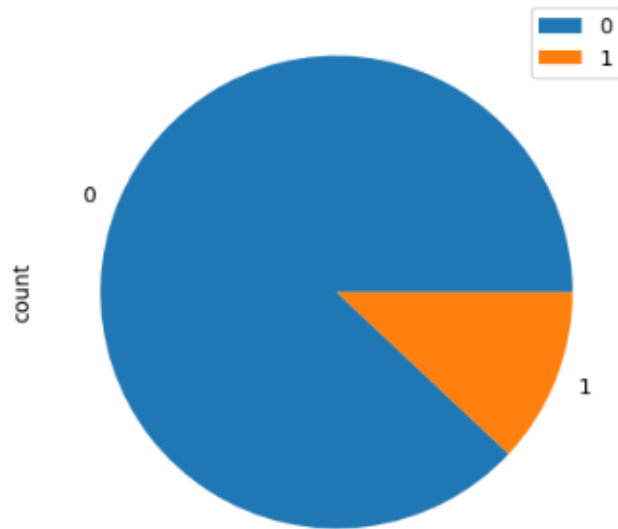
Inference:

The proportion of defaulters increases steadily as the number of days between decisions increases. This may be because the borrowers who have had a longer wait between loan decisions are more likely to default on a new loan.

BIVARIATE ANALYSIS WITH PIECHART

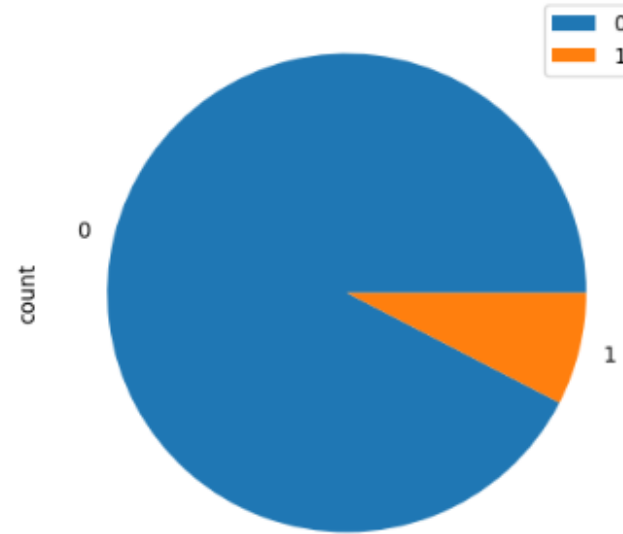
NAME_CONTRACT_STATUS VS TARGET

TARGET Values for Refused Loans



```
TARGET
0    88.004
1    11.996
Name: proportion, dtype: float64
```

TARGET Values for Approved Loans



```
TARGET
0    92.411
1     7.589
Name: proportion, dtype: float64
```

Inference:

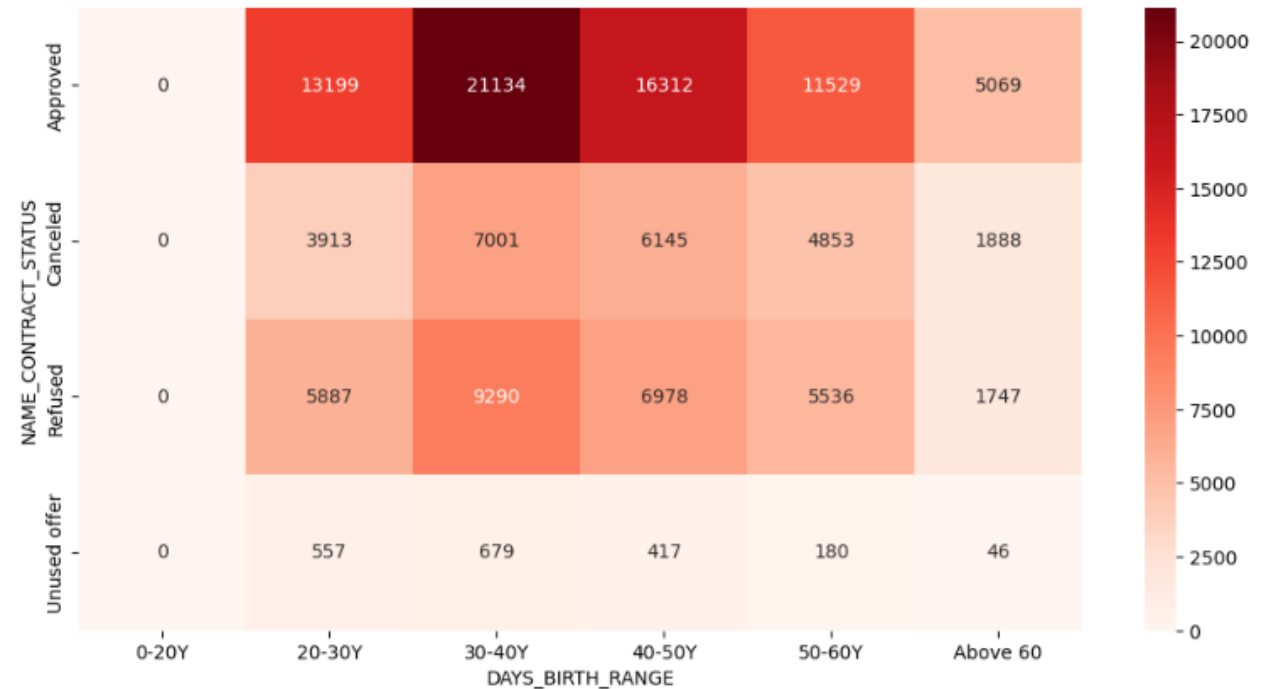
In previously refused loans, there are 12 percent defaulters now. and about 7 percent are defaulters now in previously approved loans

MULTIVARIATE ANALYSIS WITH HEATMAP

NAME_CONTRACT_STATUS VS DAYS_BIRTH VS TARGET

Inference

- -Younger applicants (around 20-30 years old) seem to have the highest default rates (since redder here means more default rate)
- - Default rates generally decrease with age, reaching a minimum around 40-50 years old
- - Also, people with more age gets more easy approval than people with less age



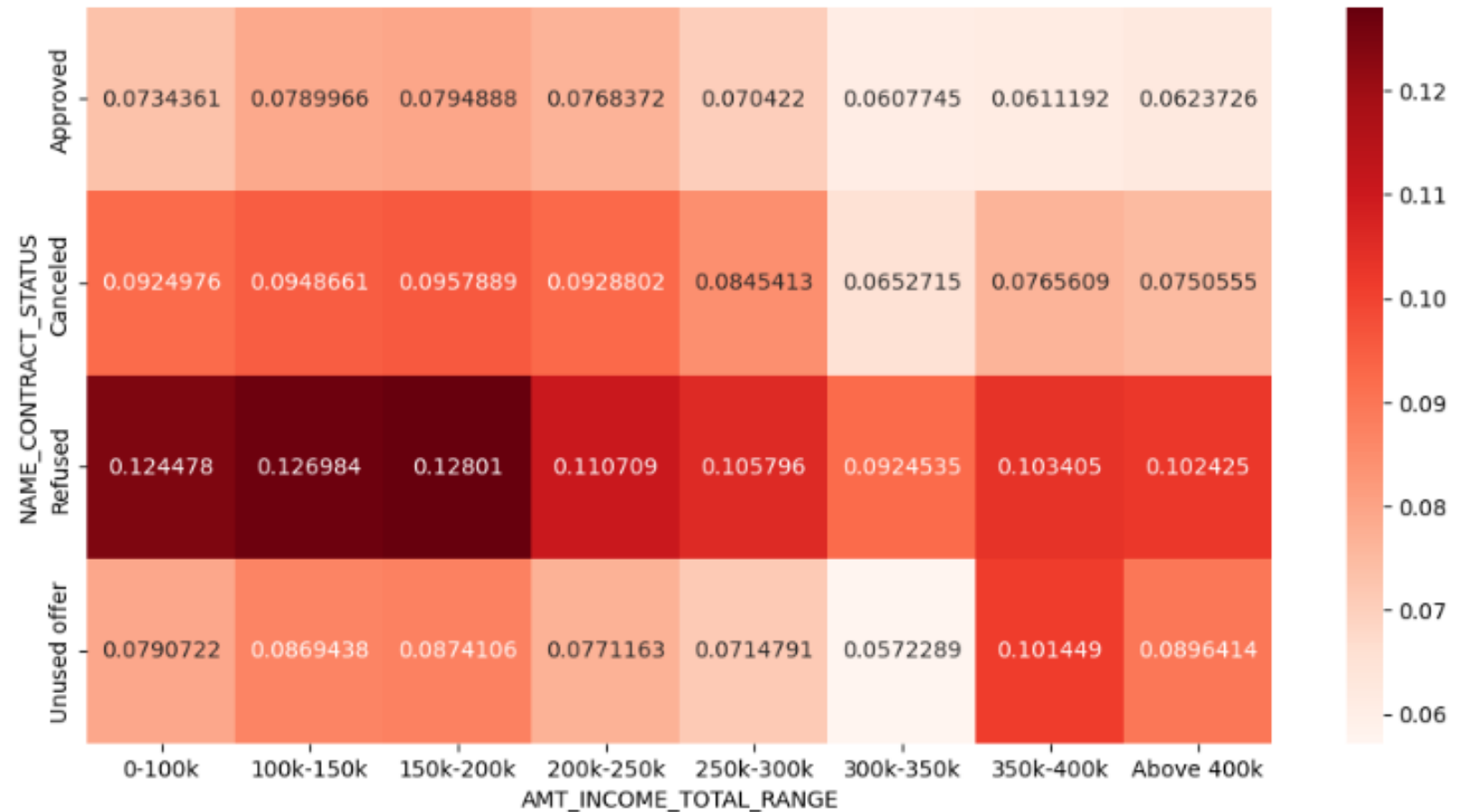
MULTIVARIATE ANALYSIS WITH HEATMAP

NAME_CONTRACT_STATUS VS AMT_INCOME_TOTAL VS TARGET

Inference

Applicants with lower salary ranges (potentially below 100k) seem to have higher default rates

The heatmap might show different approval patterns for different salary ranges.



Strong Correlations

- **AMT_CREDIT vs AMT_GOODS_PRICE**
 - Both are obvious as AMT_CREDIT will increase with AMT_GOODS_PRICE
- **AMT_CREDIT vs AMT_ANNUITY**
 - Both are also obvious since AMT_ANNUITY will increase with increase in AMT_CREDIT
- **AGE vs TARGET**
 - AGE and TARGET variable has a negative correlation, as in , when age increases, the default rate decreases.
- **DAYS_EMPLOYED vs TARGET**
 - As DAYS_EMPLOYED increases, the default rate decreases, negative correlation
- **AMT_INCOME_TOTAL VS DAYS_EMPLOYED**
 - Both are obvious, since a greater number of work experience will result in more salary
- **AMT_INCOME_TOTAL vs TARGET**
 - I could observe that default rate is decreasing with increase in annual income.
- **CREDIT_BUREAU VS TARGET**
 - Default rate increases with increase in the number of requests sent to CREDIT BUREAU
- **DEF_30/60_DPD VS CNT_SOCIAL_CIRCLE**
 - Default rate increases with increase in the number of defaulters in the applicant's surroundings



CONCLUSION AND RECOMMENDATIONS

Conclusion

1. Demographic Factors:

1. Check the number of children. Prefer customers with no children for a lower default risk.
2. Consider age group; prioritize individuals aged 40-50 for better repayment stability.
3. The gender of the applicant, with females having a slightly lower default rate.

2. Financial Factors:

1. Evaluate annual income. While 100k-150k income applicants are common, they have a higher default rate. Prefer applicants with 200k-250k income for lower risk.
2. Assess the credit amount requested. Loans between 200k-300k are more common and have a moderate default rate, but loans in the 400k-500k range are having comparatively high default rate.
3. Purpose of the loan is not clear like, XAP category is the most common, followed by XNA. (no explanation for this in columns_description.csv)

3. Employment, Occupation and Income Type:

1. Prioritize applicants with a stable job ("working" category).
2. Low Skilled workers with Rented apartments or people with no job and staying with parents are very risky.
3. Pay attention to the individuals on "Maternity leave" or marked as "Unemployed," as they pose a higher default risk.

4. Loan Types and Approval Status:

1. Pay attention to cash loans, which are popular but have the highest default rate.
2. Consider the applicant's previous loan records. New loan applications have a higher approval rate.
3. Evaluate the approval status of previous loans, as previously refused loans are associated with higher default rates in the current application.

5. Education and Credit Check:

1. People with Higher Education are more likely to payback loans.
2. Lower and Secondary education people are very risky to give loans as they have high default rates.
3. More requests to Credit Bureau means more default rate.

6. Refused Loans VS current default rate:

1. Note that previously refused loans, when approved in the current application, have a higher default rate (about 7 percent). Carefully assess the risk associated with such cases.
2. Almost all previously refused loans are now associated with a high default rate, especially for applicants with incomes between 0-250k.



Thank You
Jibin Baby