Risk analytic in banking and financial services - Case Study

Submitted By : Rahul Sen

Abstract

- 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.

Methodology

Data Sourcing

- Importing the necessary libraries.
- Reading the Dataset.
- Checking the Heads, Data types, shapes, basic stats for numeric values etc.

Data Cleaning

- Checking the null value percentage
- Removing the null valued columns (>40%)
- Drop unnecessary variables
- Drop rows with very less % of null values

Univariate Analysis

- Analysing Single Variable i.e. each column,
- plotting the distributions of each column.

Bivariate Analysis

- Analysing Two Variable i.e. Target Vs Age
- Plotting them to find different Insights.
- Plotting numeric - numeric , Numeric categorical etc.

Multivariate Analysis

- Analysing three or more variable Variable i.e. Target Vs Age Vs gender
- Plotting them to find different Insights.

Conclusion

- Analysing all plots and recommendations for reducing the loss of business by detecting columns best which contribute to loan defaulters.

Step 1 - Data Sourcing

- It includes importing the useful libraries for Data analysis and plotting.

```
#import the useful libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

- Reading the data set and assign it to a variable.

```
#read the data set of "previous_application" in pre_app.
pre_app = pd.read_csv("previous_application.csv")
```

```
#read the data set of "application_data" in app_data.
app_data = pd.read_csv("application_data.csv")
```

- Checking the Heads, Data types, shapes, basic stats for numeric values describing the data set etc.

Step 2 - Data Cleaning

Data Cleaning includes the following 6 types of cleaning :

1. Fix rows and columns : Handling Incorrect rows, Missing Column Names etc
2. Missing Values : Significant number of Missing values in a row/column
3. Standardise Numbers : Remove outliers
4. Standardise Text: Non-standard formats
5. Fix Invalid Values : Negative values
6. Filter Data: Incorrect data types

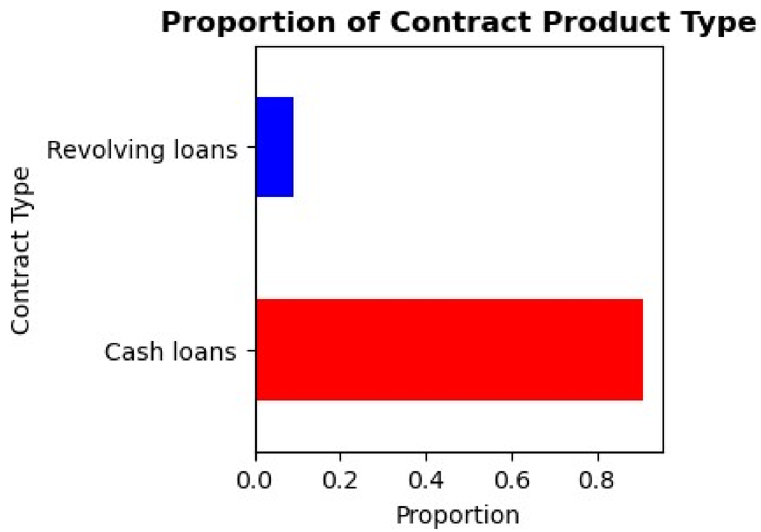
Step 2 - Data Cleaning (contd..)

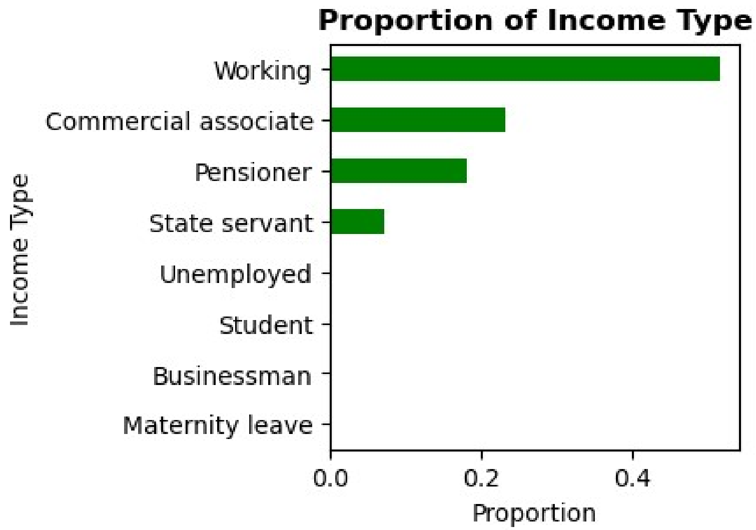
In our study the following considerations are made while cleaning the data:

1. Dropping all the columns having missing value more than 40%.
2. Dropping the irrelevant columns (ex: "AMT_REQ_CREDIT_BUREAU_HOUR" , "AMT_REQ_CREDIT_BUREAU_DAY" , "AMT_REQ_CREDIT_BUREAU_WEEK" etc)
3. droping the rows having less than 1% missing values.
4. For numerical Missing values with no outliers filled with mean and With outliers filled with Median. for Categorical missing values filled with Mode.
5. Binning has been done for some numeric columns (AMT_ANNUITY, GRP_AMT_GOODS_PRICE etc).
6. Columns with negative values DAYS_BIRTH, DAYS_EMPLOYED etc, has been converted to positive values to find actual age in years, Employment in years.

Step 3 - Univariate Analysis

Sec-1 : Categorical unordered univariate analysis

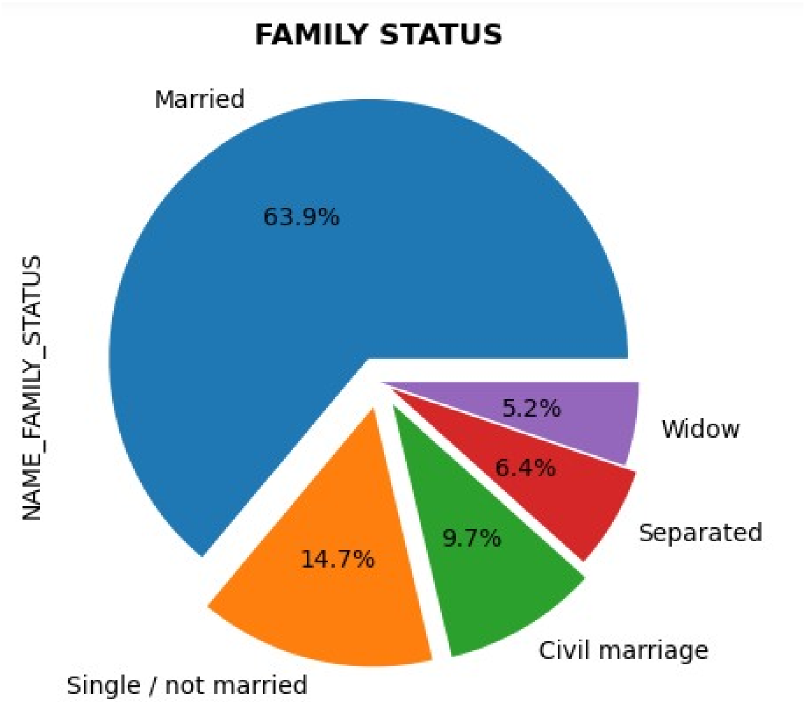


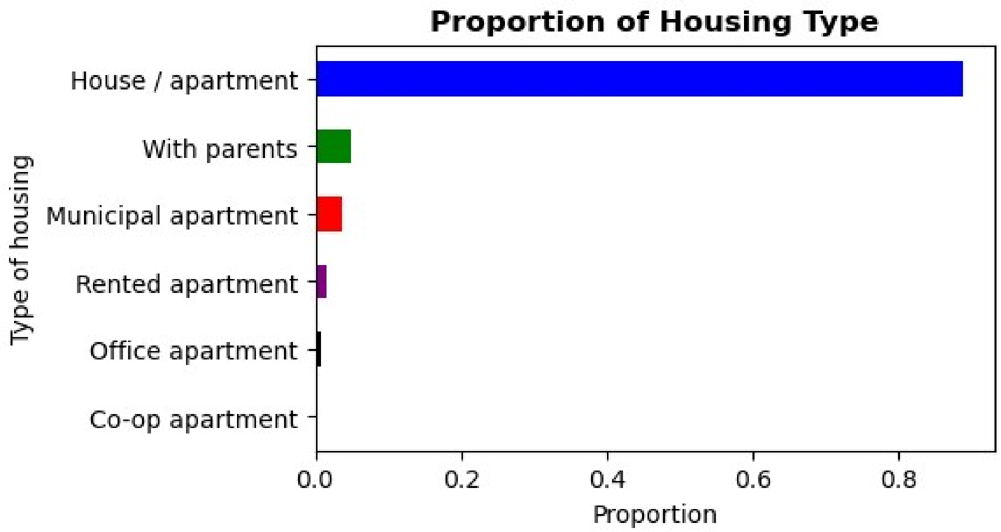


Observation : Almost 91% of loans are of Cash type and only 9 % are of Revolving type. That means there is data imbalance with a ratio of 91 : 9

Observation : Most of the clients are working professionals followed by Commercial associates. The proportion of unemployed,student , businessman,Maternity leave is very less.

Step 3 - Univariate Analysis (contd..)

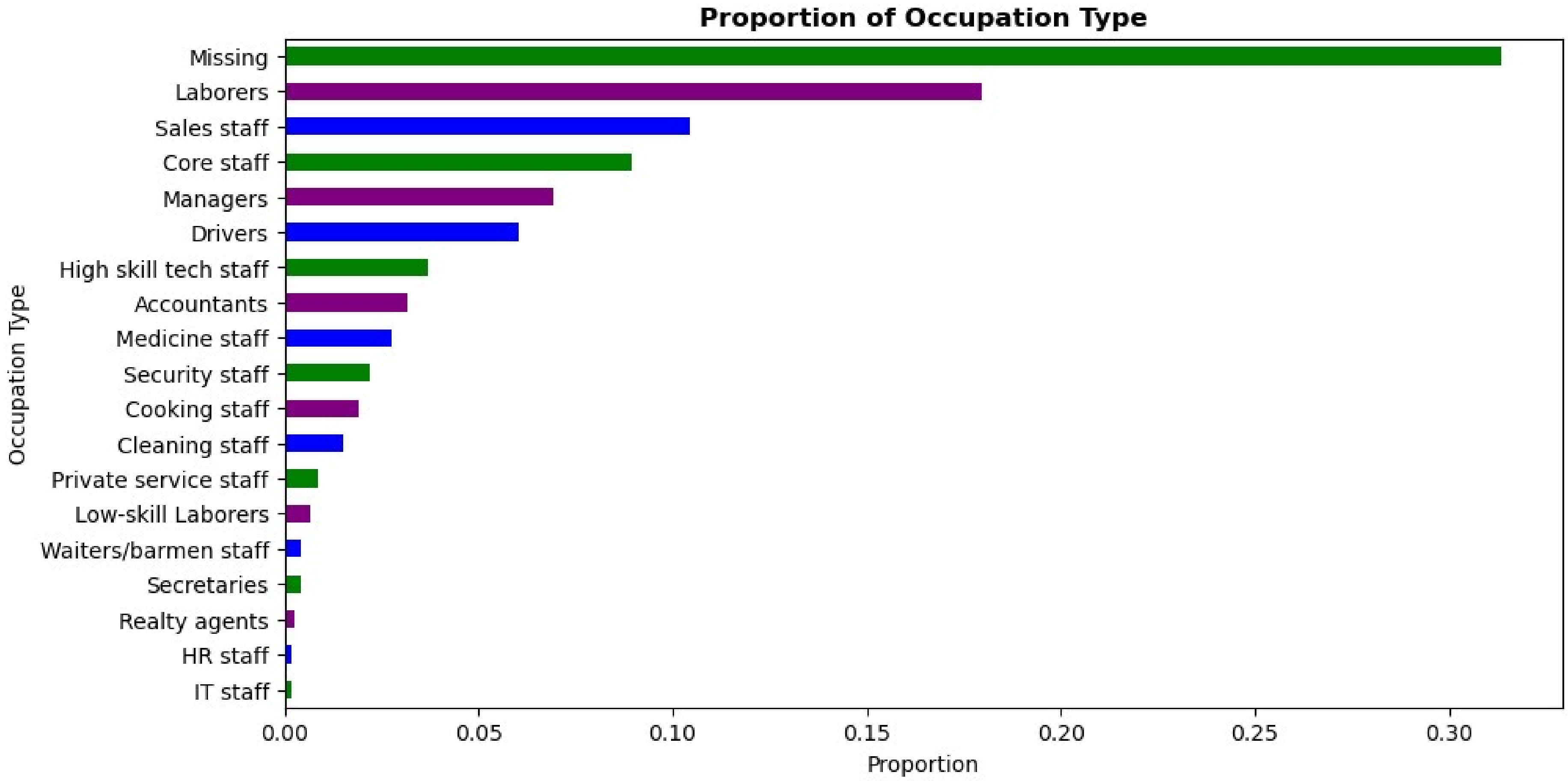




Observation : Most of the clients are Married followed by single / not married.

Observation : Most of the clients having housing type of House / Apartment. The data is heavily imbalance.

Step 3 - Univariate Analysis (contd..)

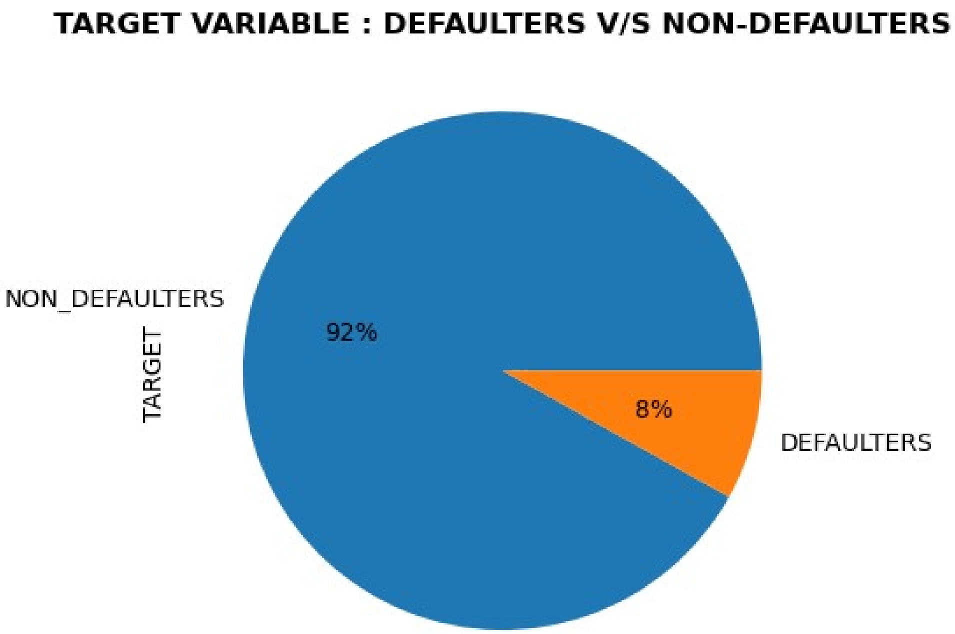


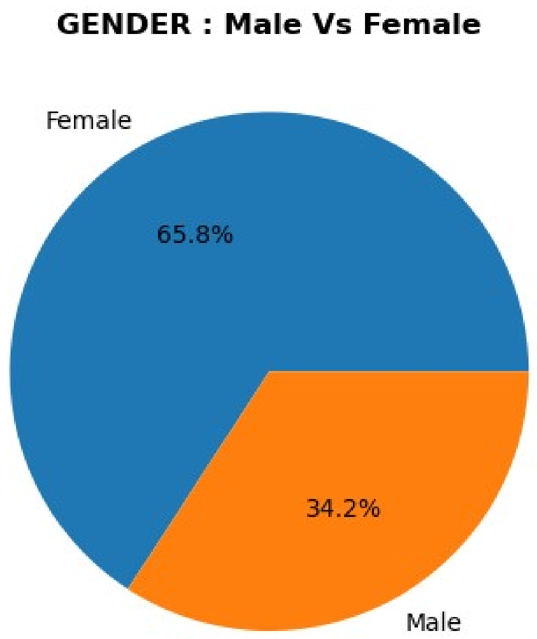
Observation : Occupation Type is having 31.3 % of Missing values. (”Missing” data has been imputed)

Step 3 - Univariate Analysis (contd..)

UpGrad

Sec-2 : Categorical ordered univariate analysis



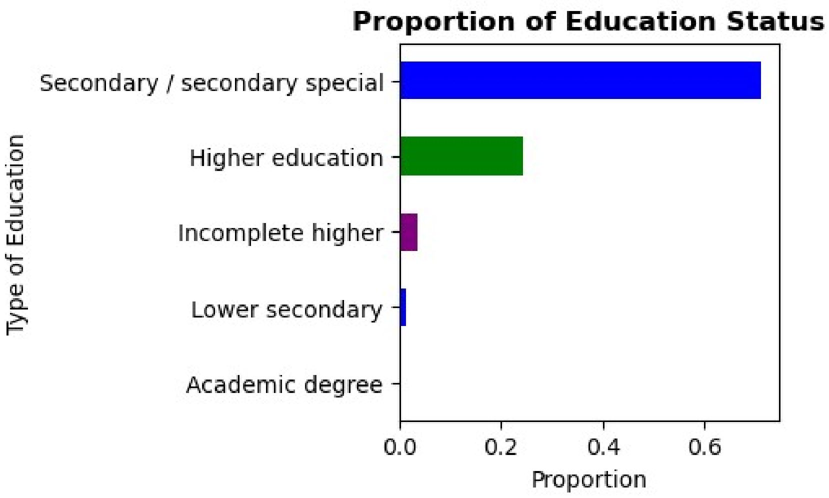


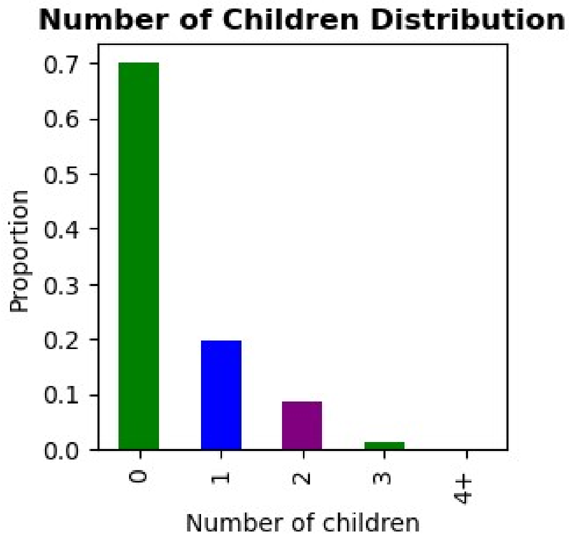
Observation : Almost 92% are NON-DEFAULTERS and only 8% are DEFAULTERS.

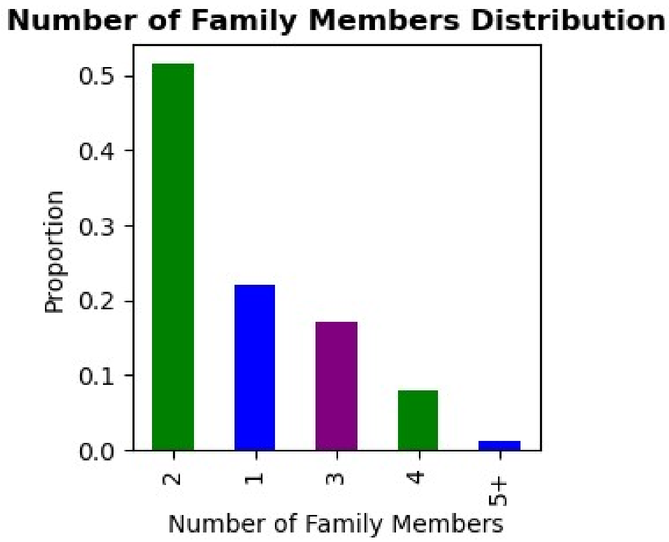
That means there is data imbalance with a ratio of 23 : 2

Observation : The ratio of applicant Female : Male is in the ratio of 2:1 (approx)

Step 3 - Univariate Analysis (contd..)



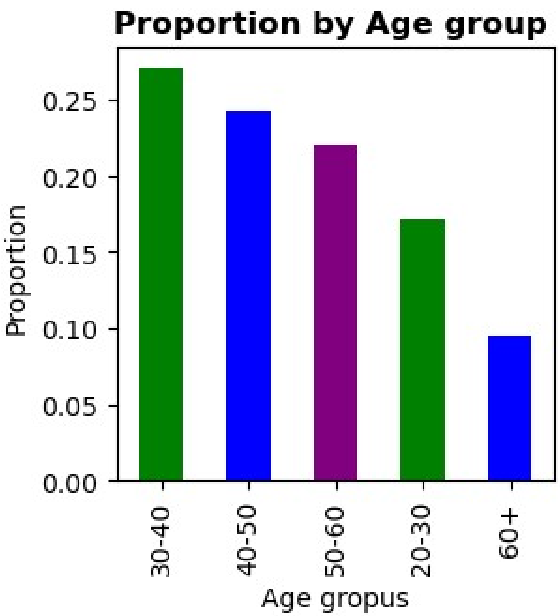


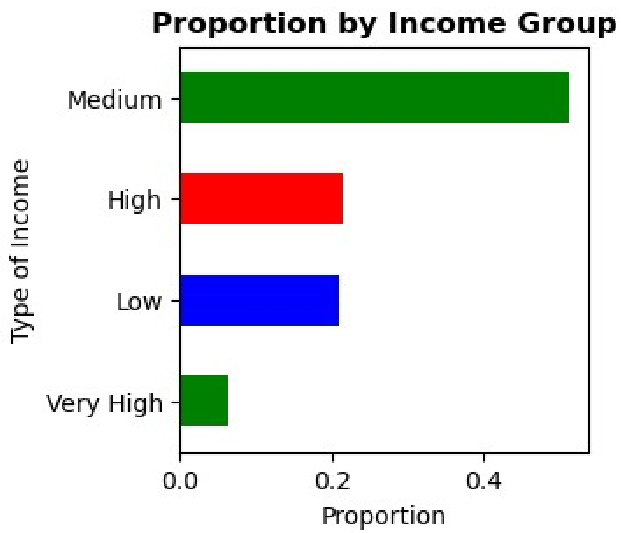


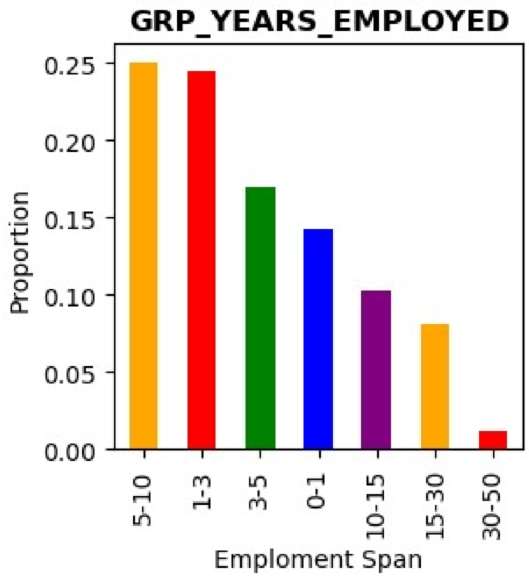
Observation : Almost 71 % of the clients are having highest education of Secondary / secondary special.

Observation : Almost 70 % of the clients are having no children and 50 % of the client are of 2 member family only.

Step 3 - Univariate Analysis (contd..)





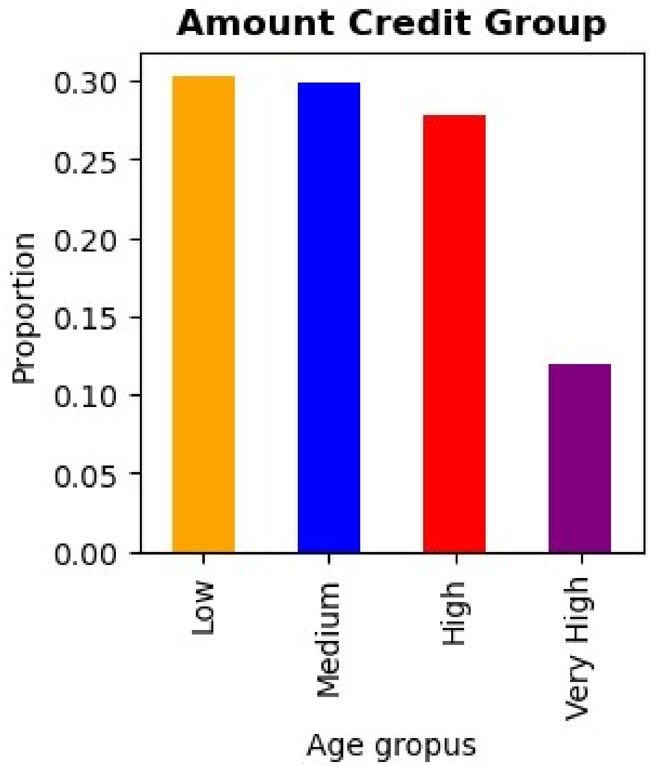


Observation : The age group of 30-40 ,40-50,50-60 are proportioned by 27% , 24% & 22% respectively

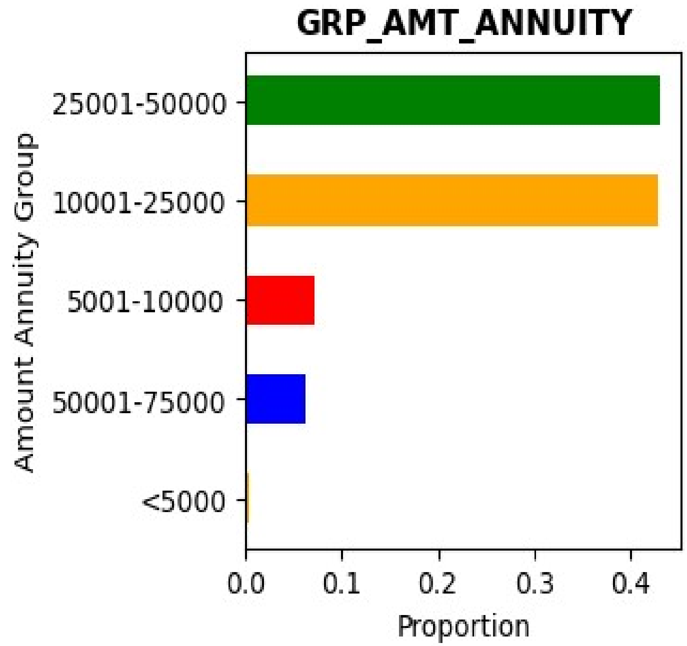
Observation : Most of the Clients belongs to the medium income group

Observation : Most of the Clients having work experience of 5-10 and 1-3 years.

Step 3 - Univariate Analysis (contd..)

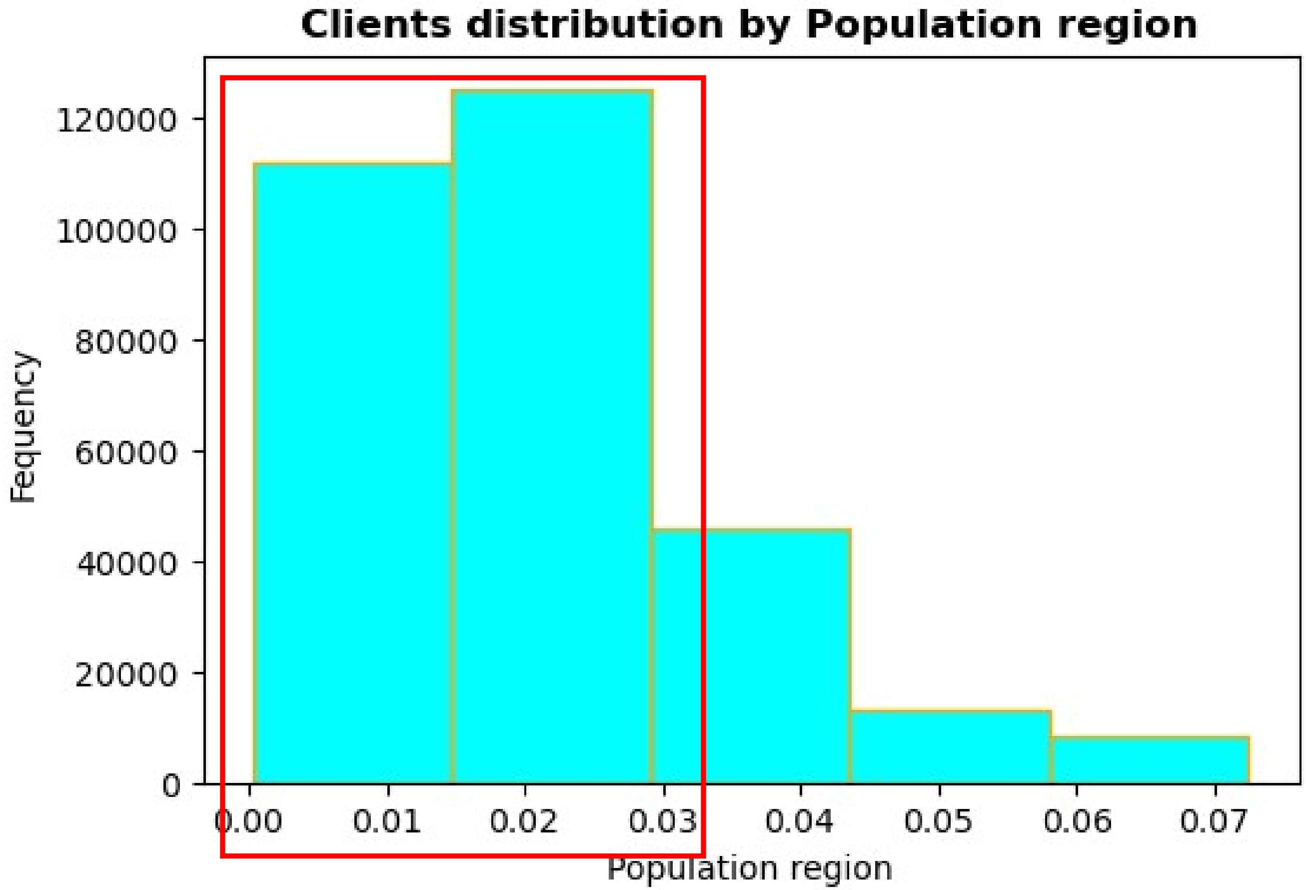


Observation :



Observation :

Step 3 - Univariate Analysis (contd..)

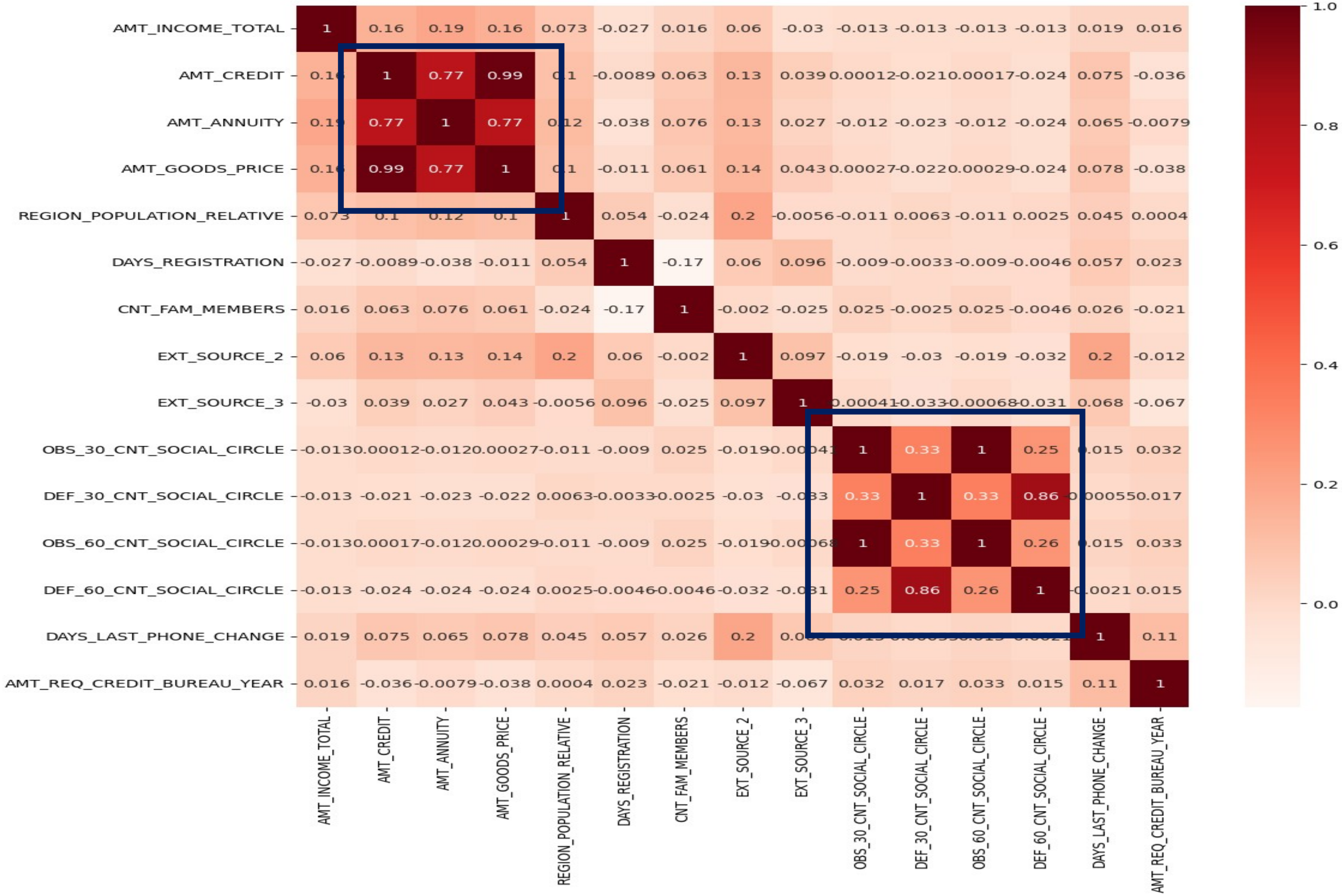


Observation : Most of the clients are belongs from low populated region.

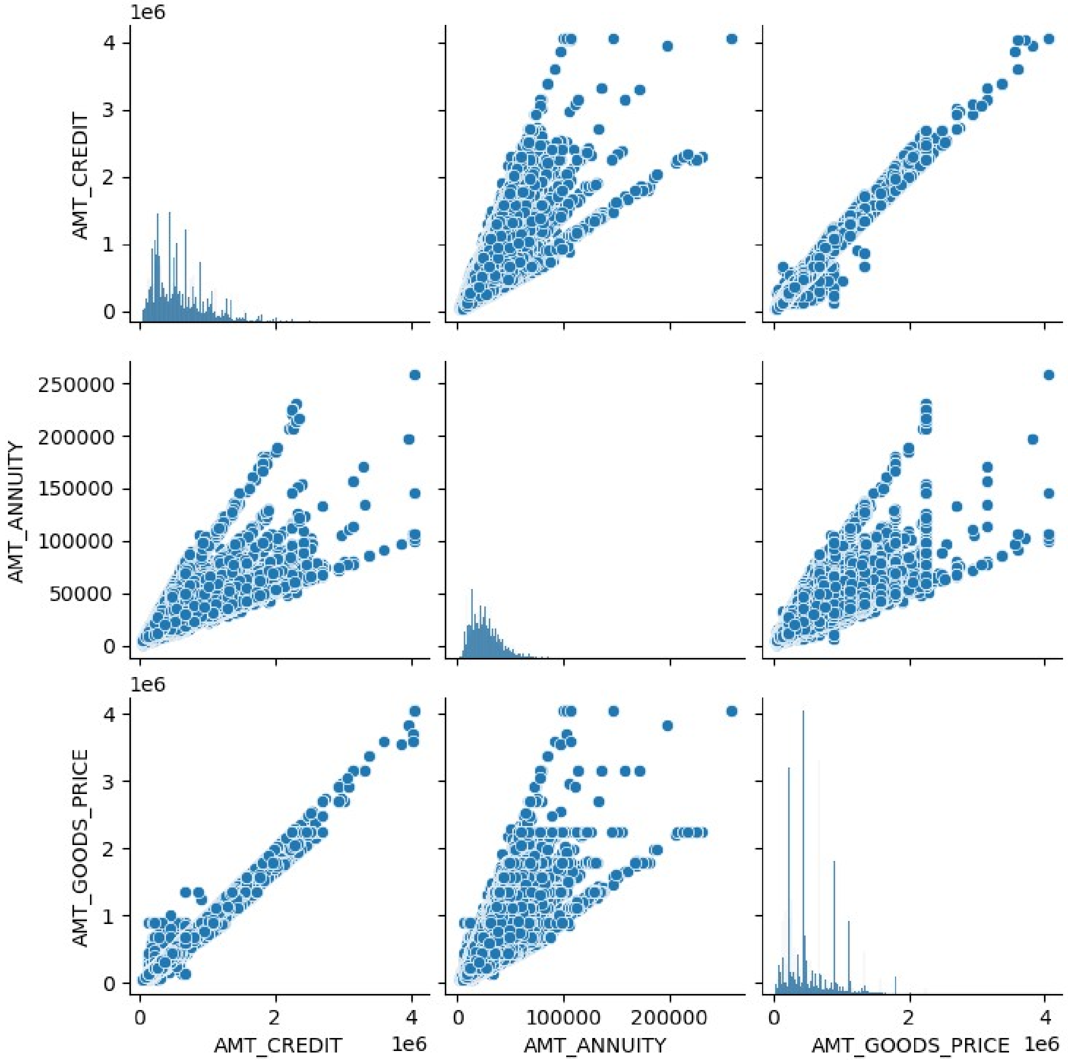
Step 4 - Bivariate Analysis

Sec-1 : Numeric-numeric analysis

Observation : The heat has been between the numerical features of the data set. Some linearly co related zones are marked in the plot.



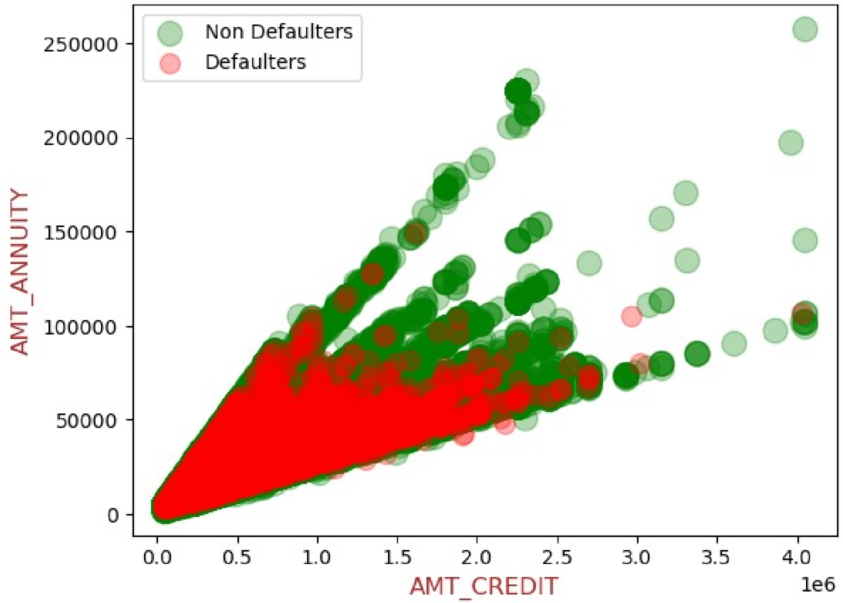
Step 4 - Bivariate Analysis(Contd..)



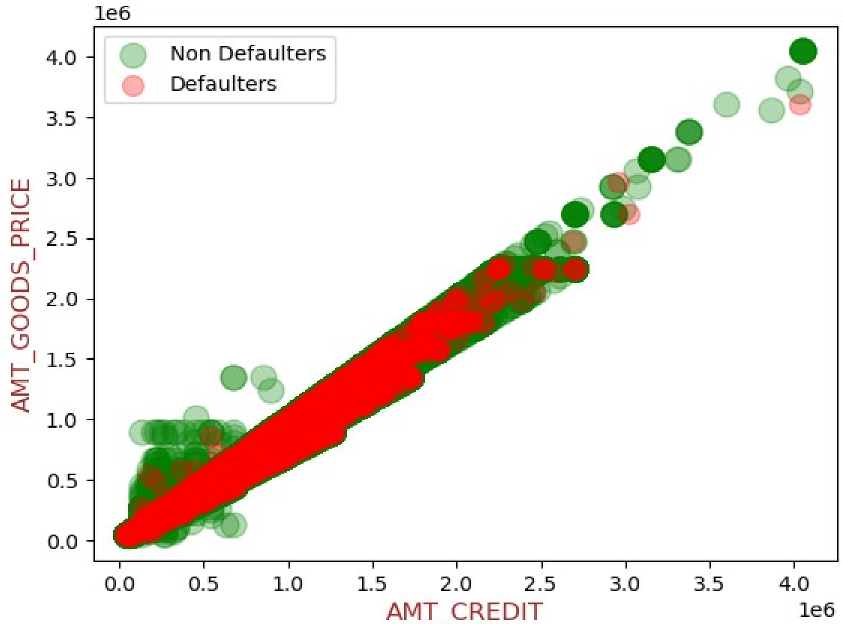
Observation : Pair plot has been done to visualize the type of relationship exist in between the variables AMT_CREDIT , AMT_ANNUITY & AMT_GOODS_PRICE

Step 4 - Bivariate Analysis(Contd..)

AMT_CREDIT Vs AMT_ANNUITY across Target varriable



AMT_CREDIT Vs AMT_GOODS_PRICE across Target varriable

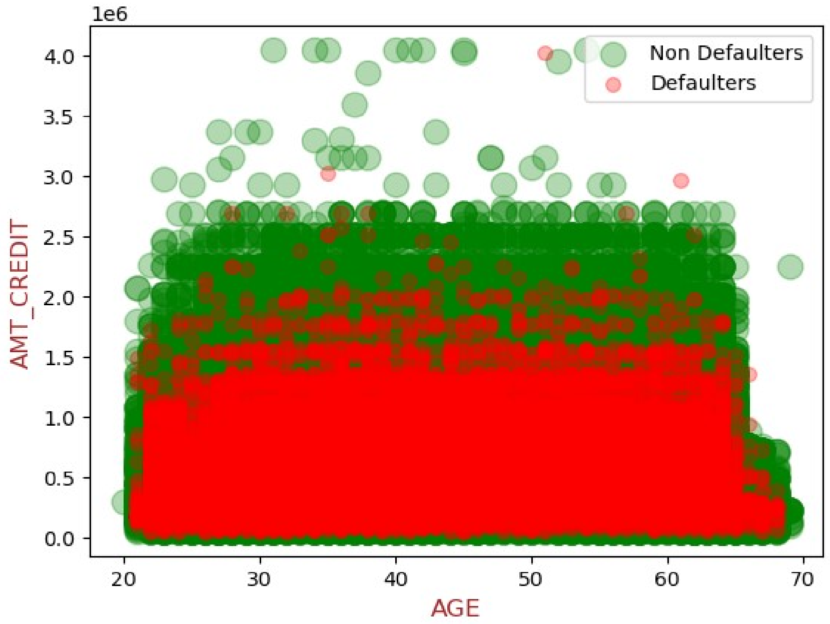


Observation : The scatter plot shows that as the annuity amount goes beyond 100000 & Credit amount of the loan goes beyond 200000 the defaulters decreases drastically.

Observation : Credit amount of the loan is having a linear correlation with the Goods price variable

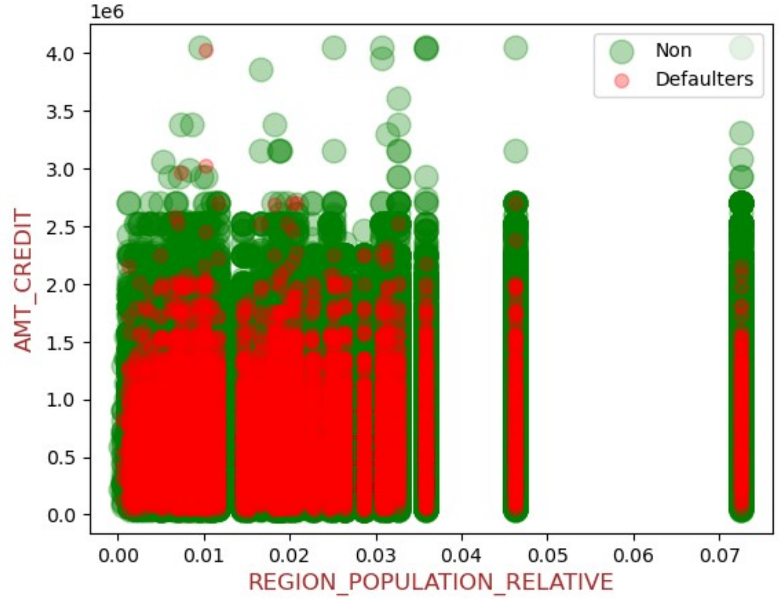
Step 4 - Bivariate Analysis(Contd..)

AGE versus AMT_CREDIT across Target varriable



Observation : The credit amount of the loan is less of aged people beyond 65 age group. Also as the Credit amount of the loan goes beyond 200000 the defaulters decreases drastically for all age groups

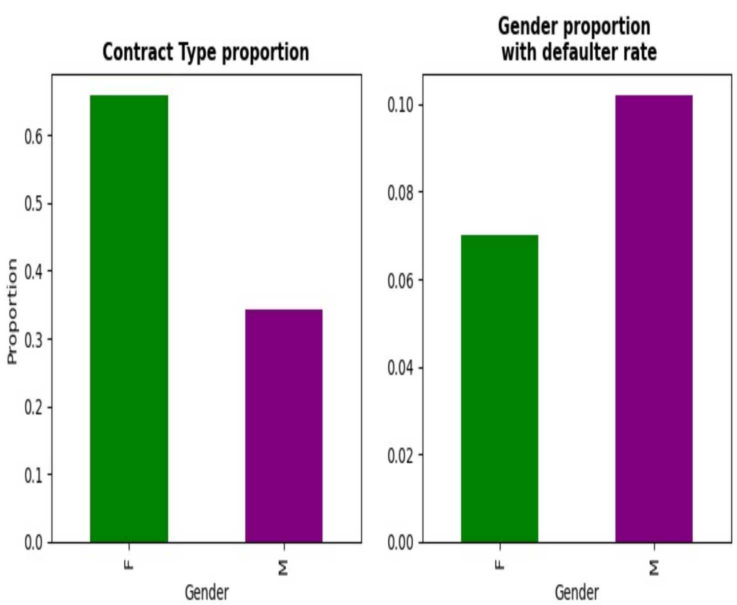
REGION_POPULATION_RELATIVE Vs AMT_CREDIT across Target varriable

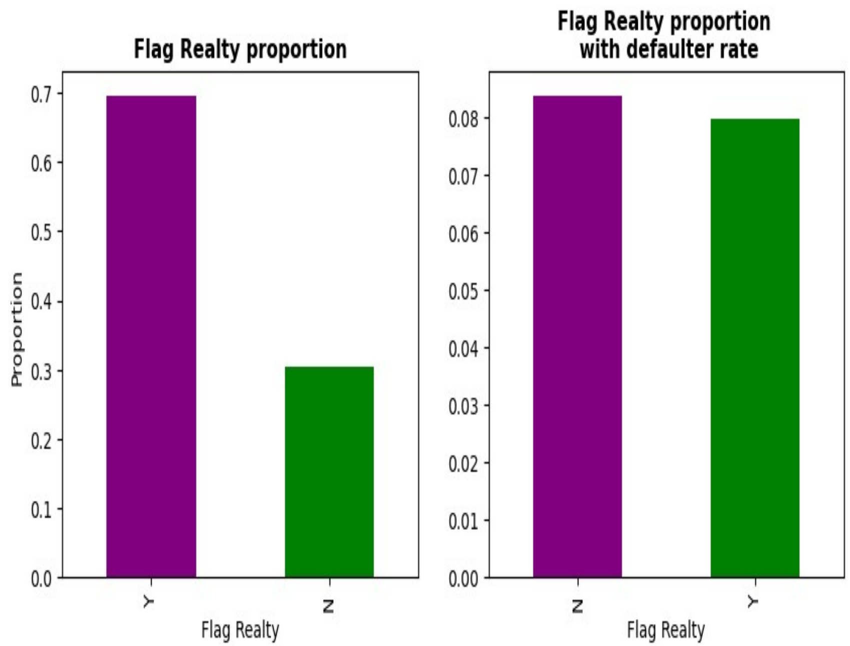


Observation : Most of the clients are belongs from low populated region. Also it is safe to provide loans for high density population (>0.07)people beyond credit amount of 200000.

Step 4 - Bivariate Analysis(Contd..)

Sec-2 : categorical categorical variable

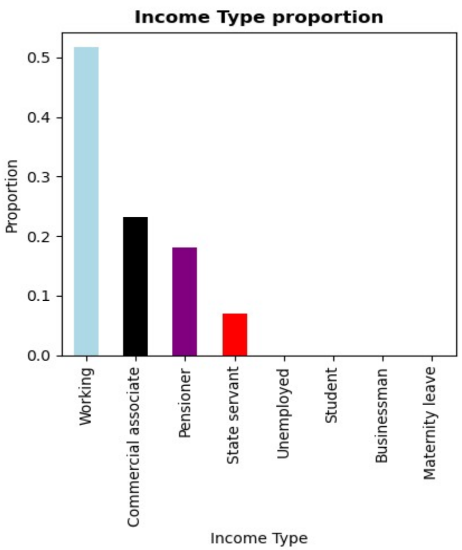


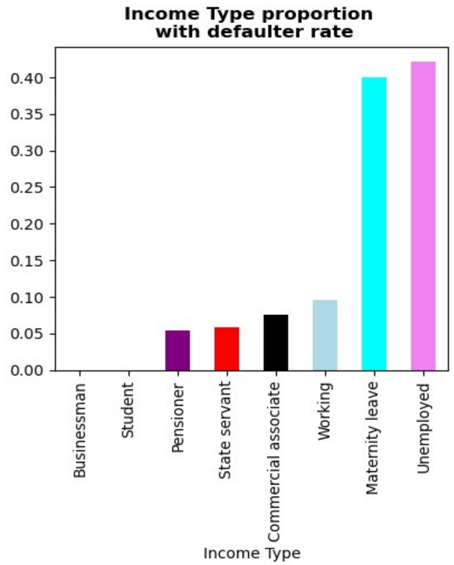


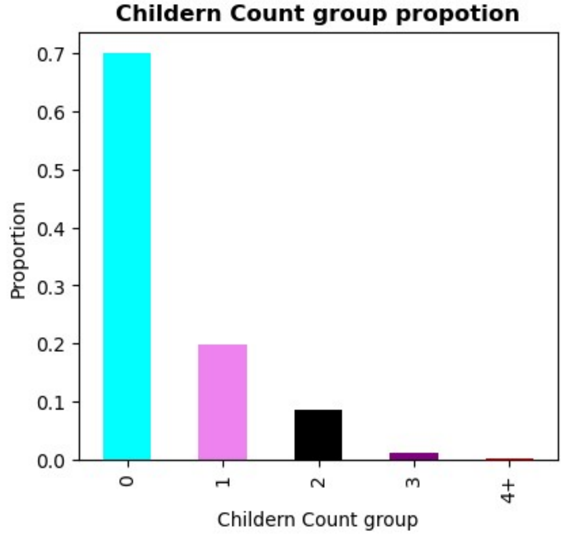
Observation : Females are the most Highest loan takers but males are highest in terms of defaulter rate

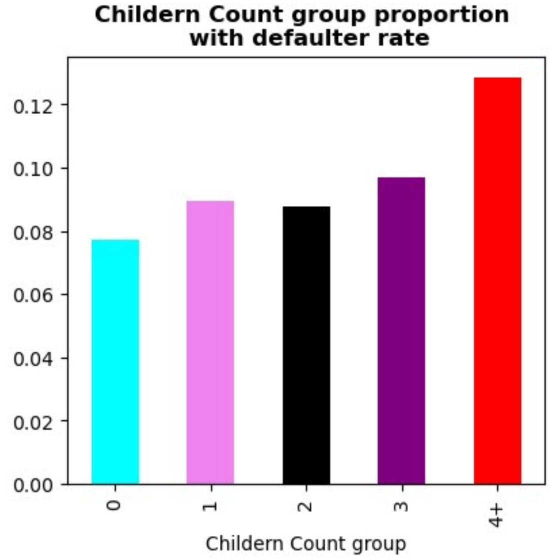
Observation : The defaulter rate is almost same for flag reality.

Step 4 - Bivariate Analysis(Contd..)





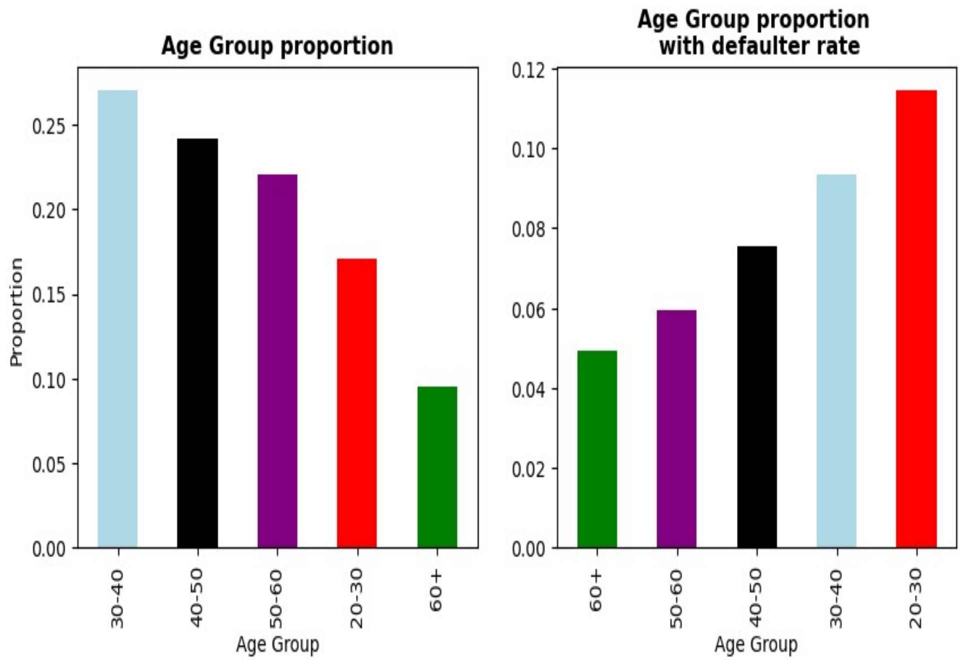


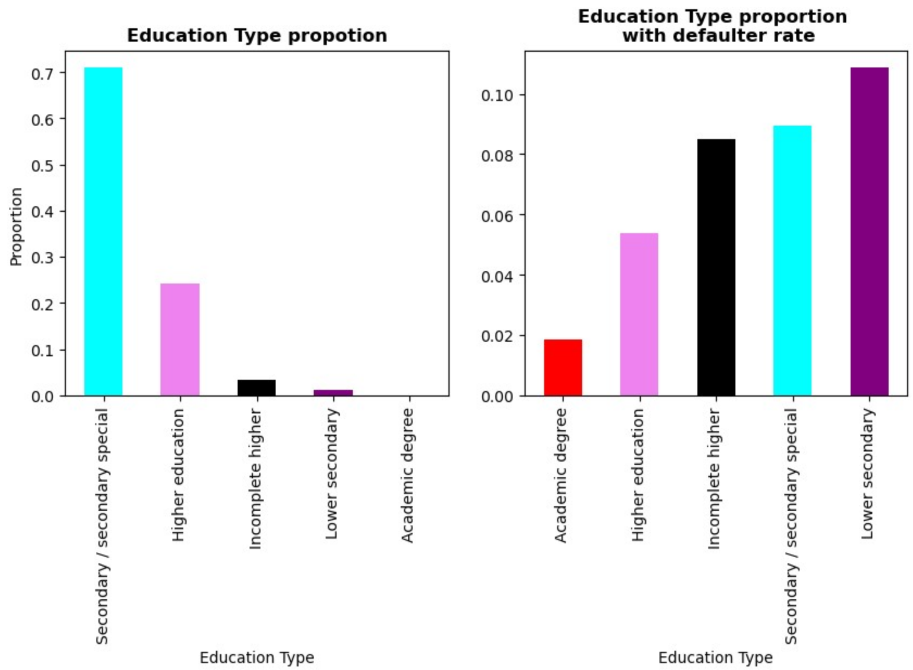


Observation : No businessman and student has defaulted and all the Maternity leave income type client has defaulted.

Observation : Client having more than 4 children having the highest defaulter rate.

Step 4 - Bivariate Analysis(Contd..)

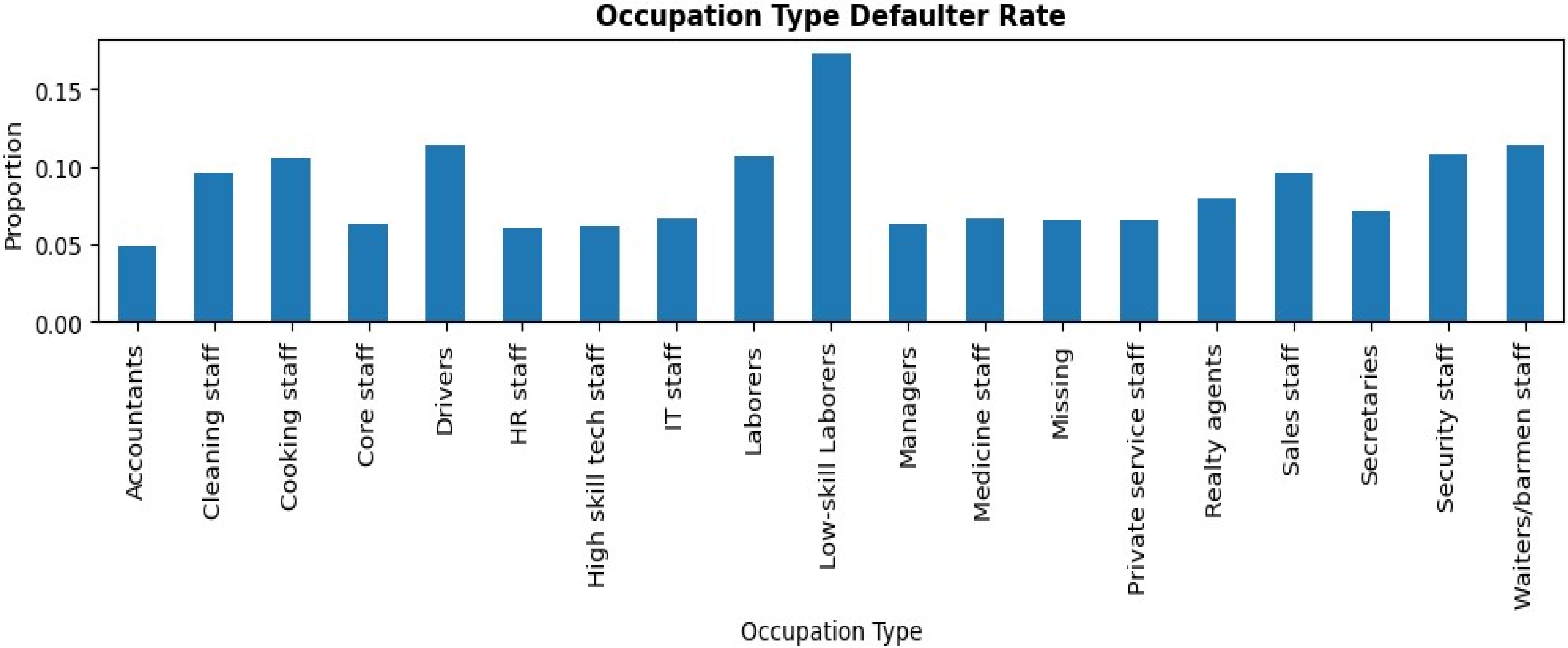




Observation : The 60+ age group is having lowest proportion in taking loans as well as lowest defaulter rate. 20 -30 age grp is having the highest Defaulter rate.

Observation : The proportion of lower academy and Academy degree is very low but compared to defaulter rate it is high. The secondary defaulter rate is also high.

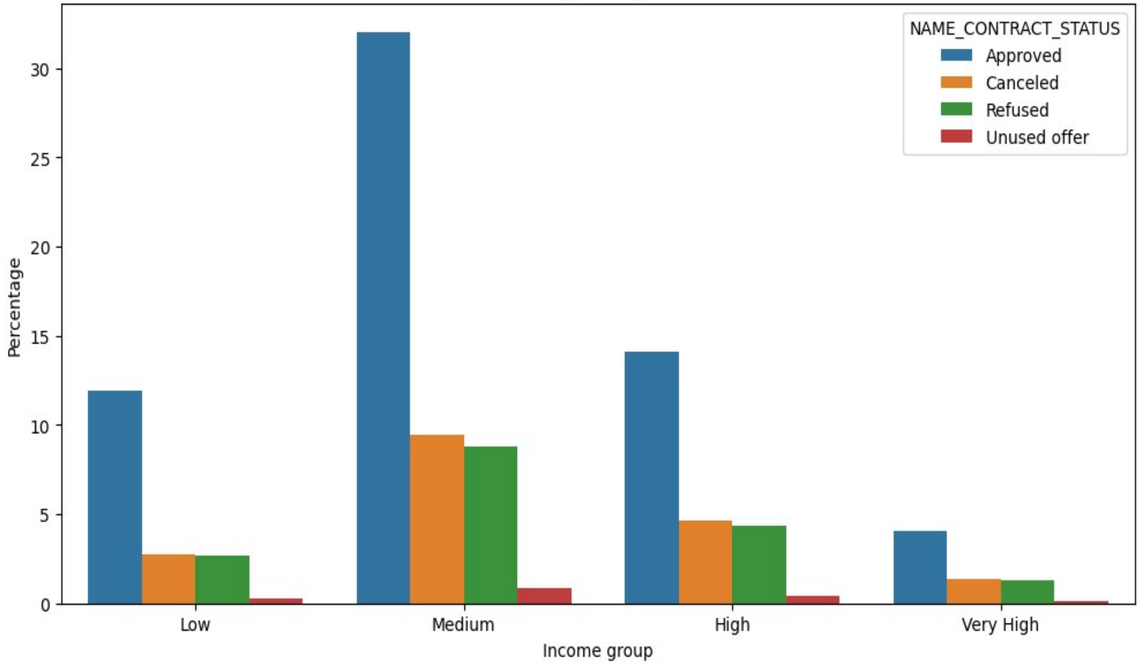
Step 4 - Bivariate Analysis(Contd..)



Observation : Low Skill Laborers have the highest default rate among all the occupation type

Step 4 - Bivariate Analysis - Merged

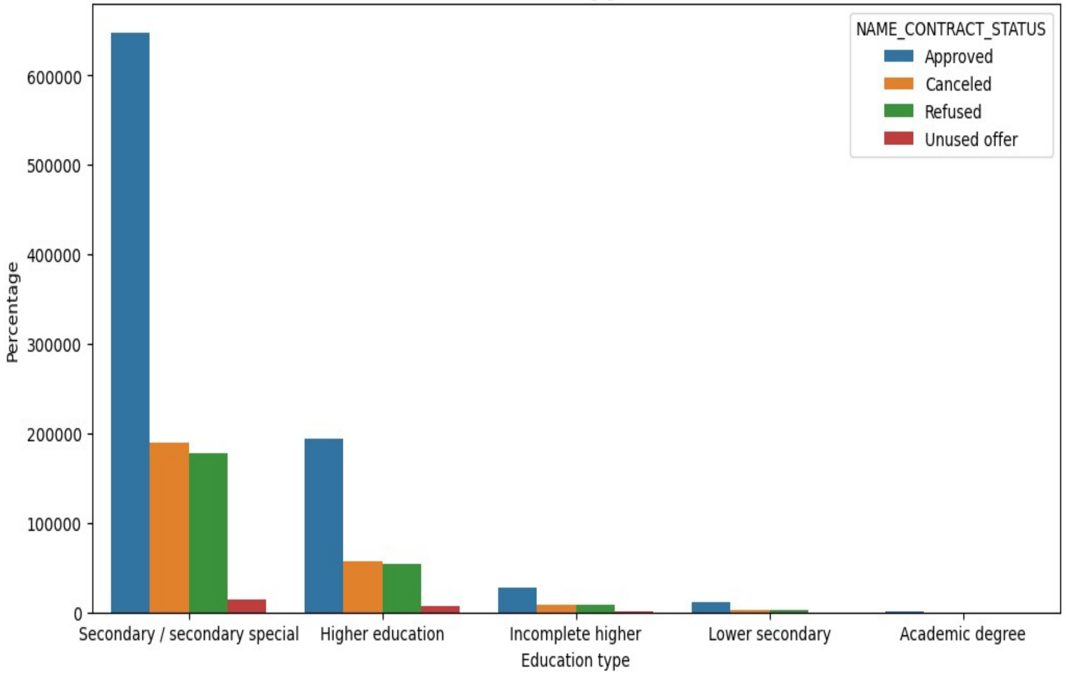
Distribution of Income group WRT Contract Status



Observation : In previous applicants, majority of Approved loans are from Medium income group.

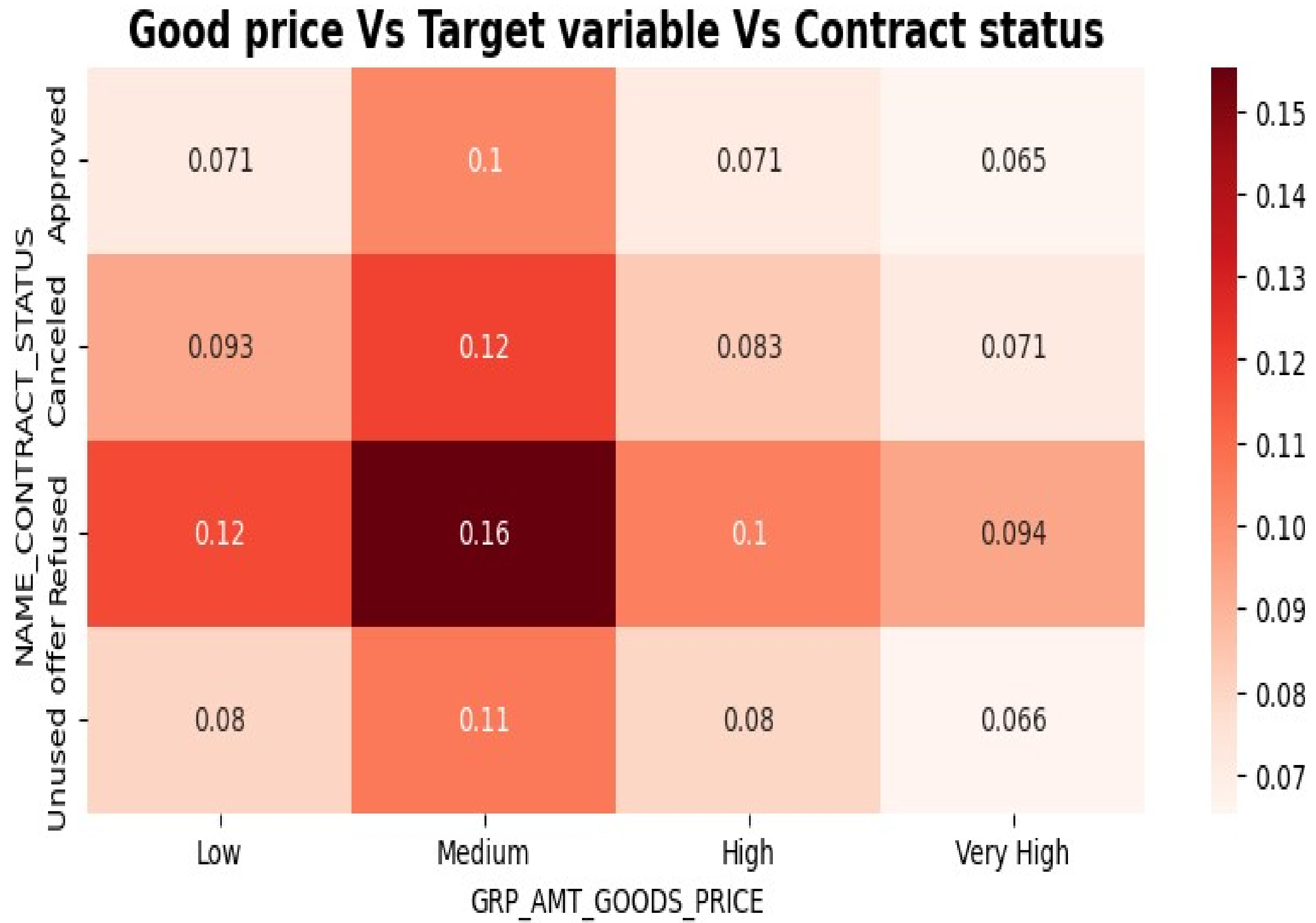
Observation : Secondary/Secondary special education type is having majority of approved loans in previous applicants.

Distribution of Education type WRT Contract status



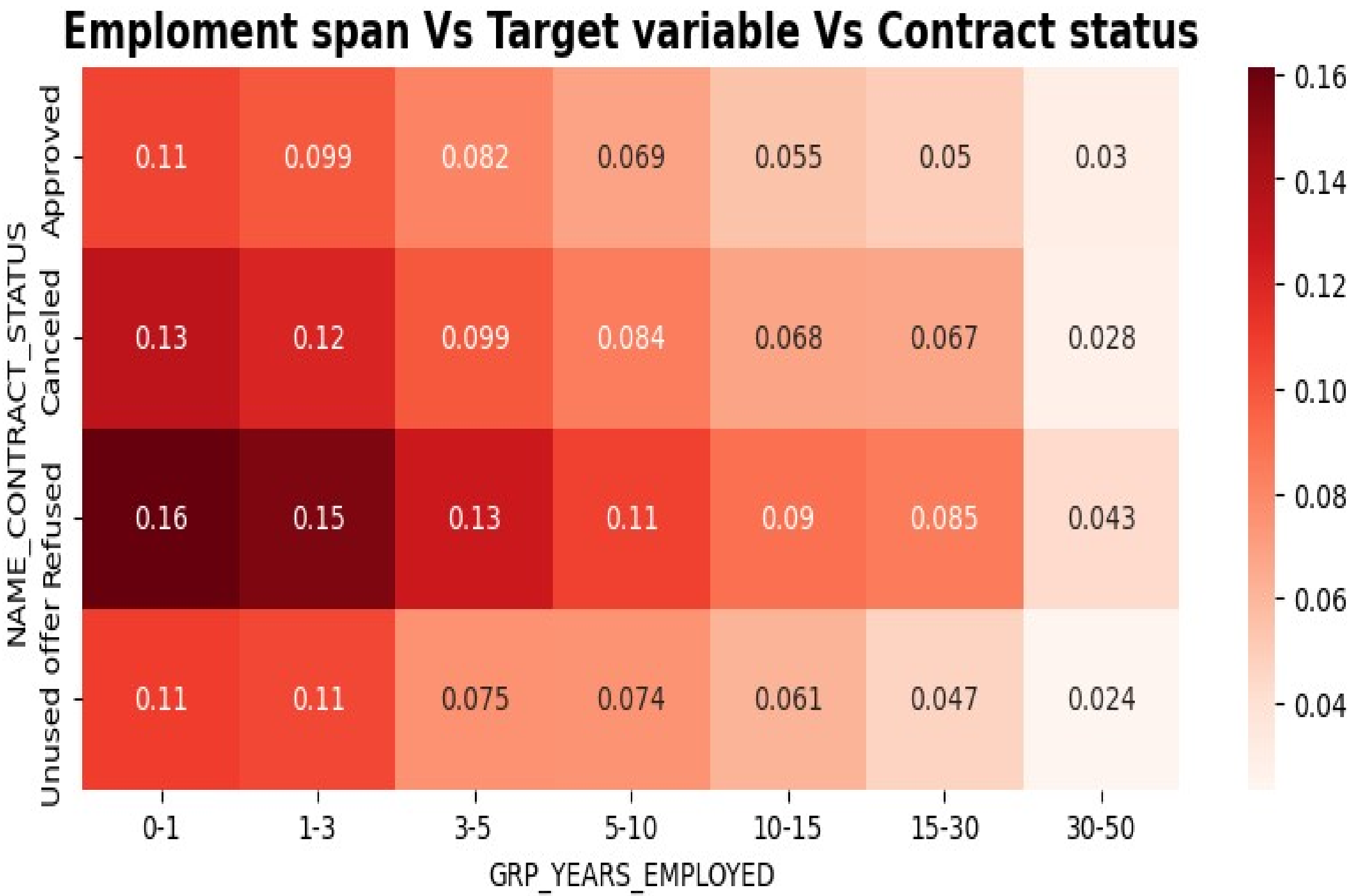
Step 5 - Multivariate Analysis - Merged

Observation: Medium goods price having linear relationship with defaulter rate.



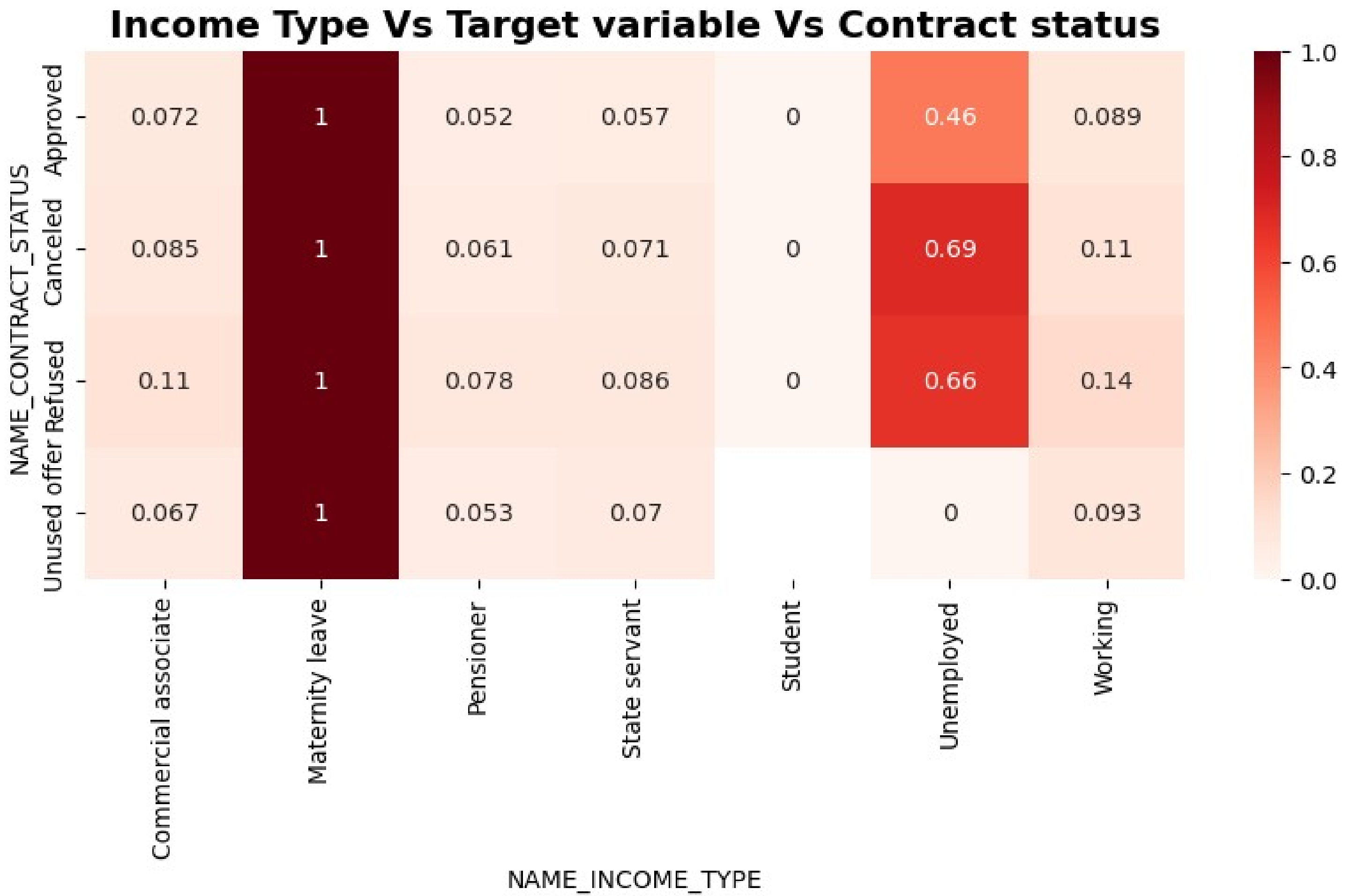
Step 5 - Multivariate Analysis - Merged

Observation: Lesser employment span tends to default more.



Step 5 - Multivariate Analysis - Merged

Observation: Maternity income type is having strong linear relationship with defaulter. Unemployed - approved loans is having medium correlation with defaulter.



#TOP 10 Correlation

#TOP 10 Correlation for New_application data frame.

SK_ID_CURR	SK_ID_CURR	1.000000
DAYS_EMPLOYED	YEARS_EMPLOYED	1.000000
	FLAG_EMP_PHONE	0.999752
YEARS_EMPLOYED	FLAG_EMP_PHONE	0.999751
DAYS_BIRTH	AGE	0.999711
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998495
AMT_CREDIT	AMT_GOODS_PRICE	0.986975
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950608
CNT_FAM_MEMBERS	CNT_CHILDREN	0.879268
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.860710
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.860057

dtype: float64

#TOP 10 Correlation for Merge_data data frame.

SK_ID_CURR	SK_ID_CURR	1.000000
YEARS_EMPLOYED	DAYS_EMPLOYED	1.000000
AMT_APPLICATION	AMT_GOODS_PRICE_y	0.999870
DAYS_EMPLOYED	FLAG_EMP_PHONE	0.999772
YEARS_EMPLOYED	FLAG_EMP_PHONE	0.999771
AGE	DAYS_BIRTH	0.999708
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998566
AMT_GOODS_PRICE_y	AMT_CREDIT_y	0.993196
AMT_GOODS_PRICE_x	AMT_CREDIT_x	0.986342
AMT_APPLICATION	AMT_CREDIT_y	0.975715
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.945437

dtype: float64

#Recommendation

- The annuity amount beyond 100000 & Credit amount of the loan goes beyond 200000 is safe to approve loans.
- Credit amount of the loan goes beyond 200000 the defaulters decreases drastically for all age groups.
- Client having more than 4 children should be lend with high interest rate.
- No businessman and student has defaulted, so they are safe to approve loans and all the Maternity leave income type client should be rejected. Unemployed income types should be lend with higher interest rate.
- Low Skill Laborers have the highest default rate among all the occupation type, should be lend with higher interest rate.
- **Some of the strong Indicators of defaulters are: age_grp , AMT_CREDIT , AMT_GOODS_PRICE, NAME_INCOME_TYPE, NAME_EDUCATION_TYPE & GRP_CNT_FAM_MEMBERS.**

THANK YOU