EDA ASSIGNMENT ON BANK DATA SET

SUBMITTED BY: MANDANNA M S
BATCH – DS – C47 PROGRAM : upGrad & IIITB | Data Science Program – AUGUST 2022

Problem Statement

Introduction

This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

Business Understanding

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Problem Statement

Business Objectives

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.

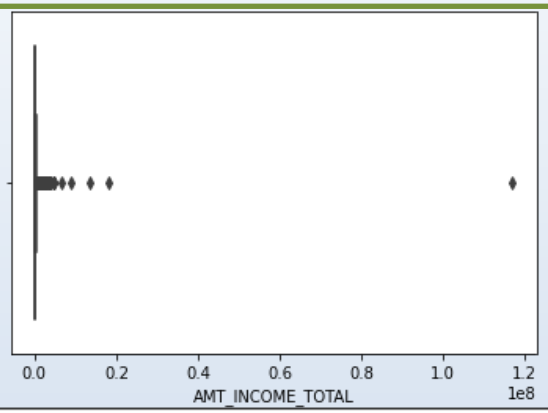
In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.
To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

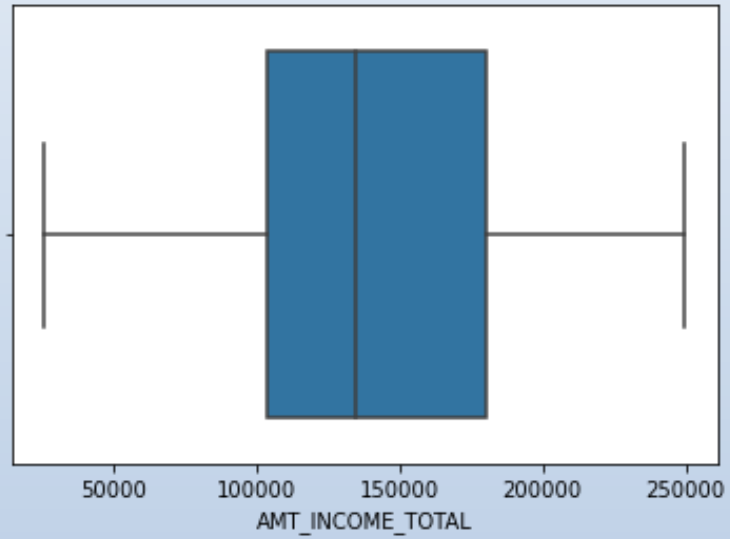
Understanding & cleaning of data:

- ✓ Necessary checks are done on the data using functions like info(), describe(), .shape, .head() etc on both the data set.
- ✓ Checked for numeric variables of dataset.
- ✓ Data cleaning process started on the data set.
- ✓ Check for null values and necessary action is then taken on those.
- ✓ Dropping the column of large missing values and Imputation is done on some of the column features.
- ✓ Check for correct data types and standardizing data .

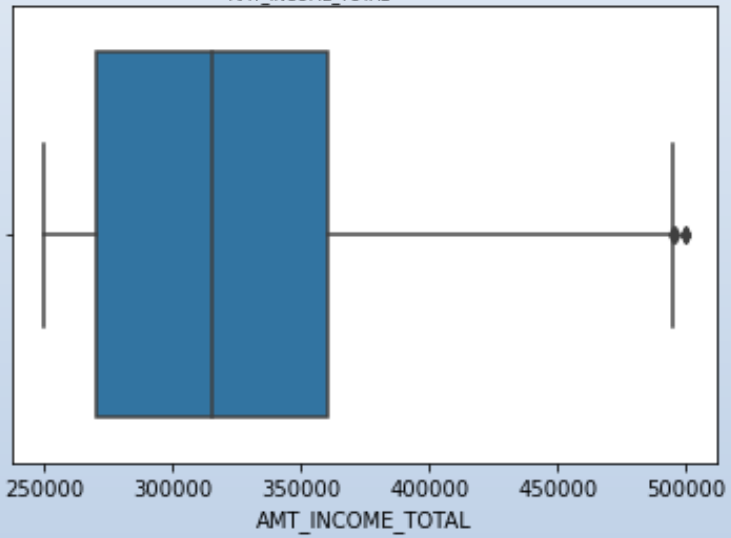
HANDLING OUTLIER

1. AMT_INCOME_TOTAL VARIABLE

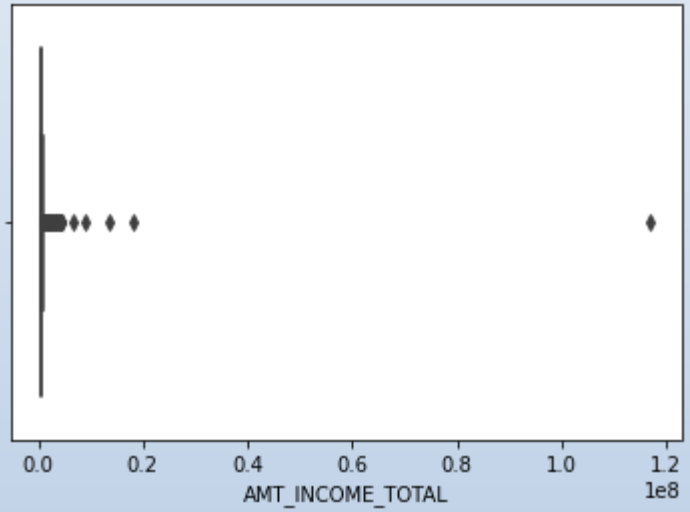




check for income in range < 250000



check for income in range > 250000 &< 500000

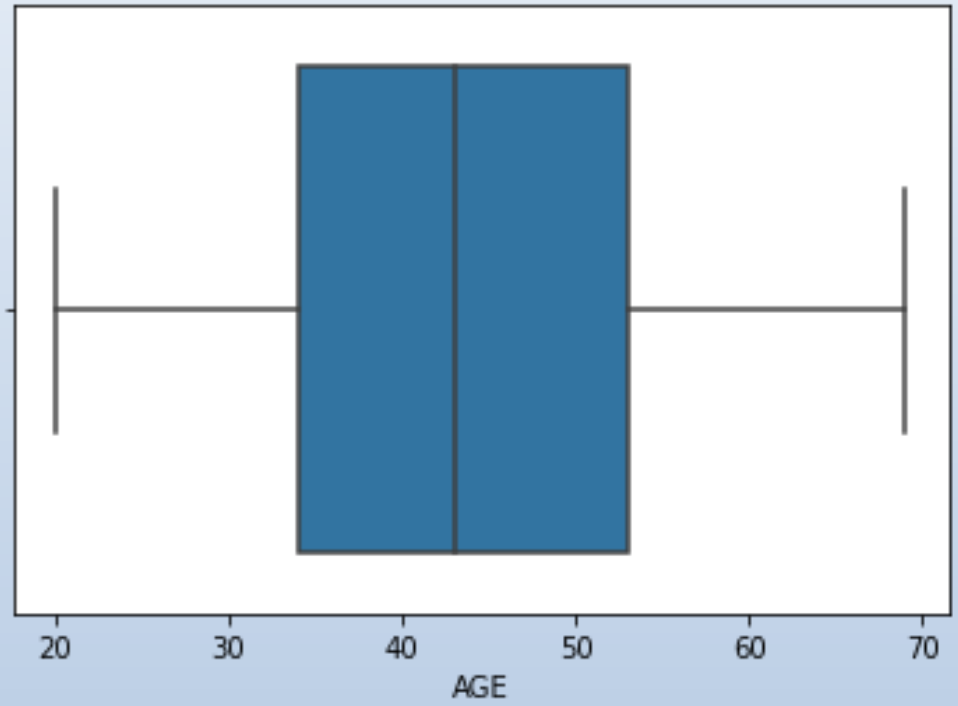


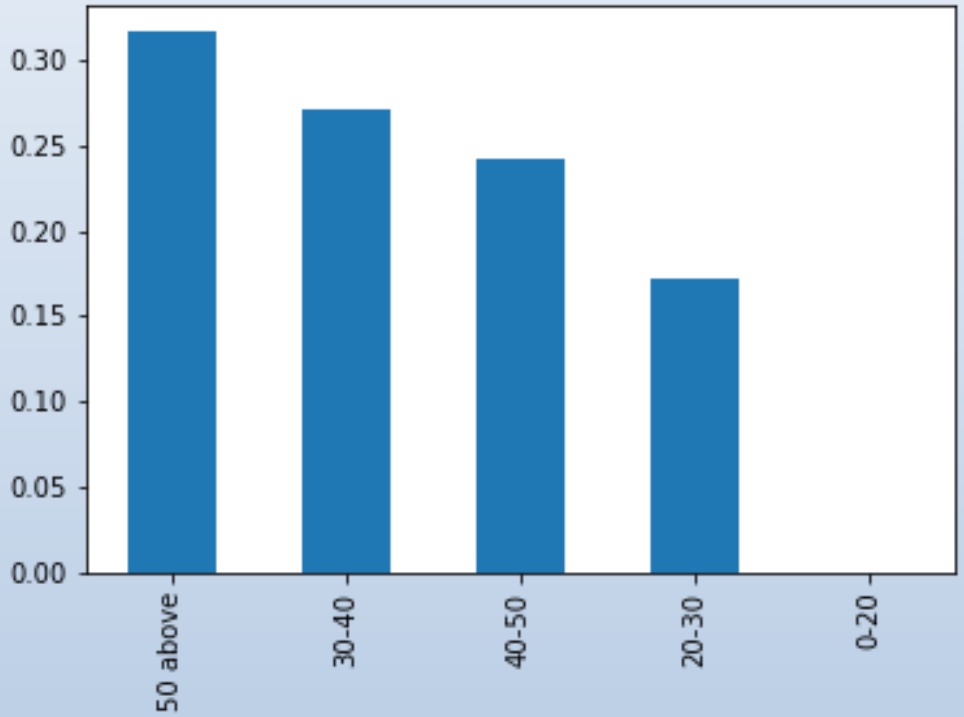
check for income in range >500000

- Created a buckets for income based on above observation as LOW,MEDIUM AND HIGH due to very high outliers observed .

HANDLING OUTLIER

2. AGE VARIABLE

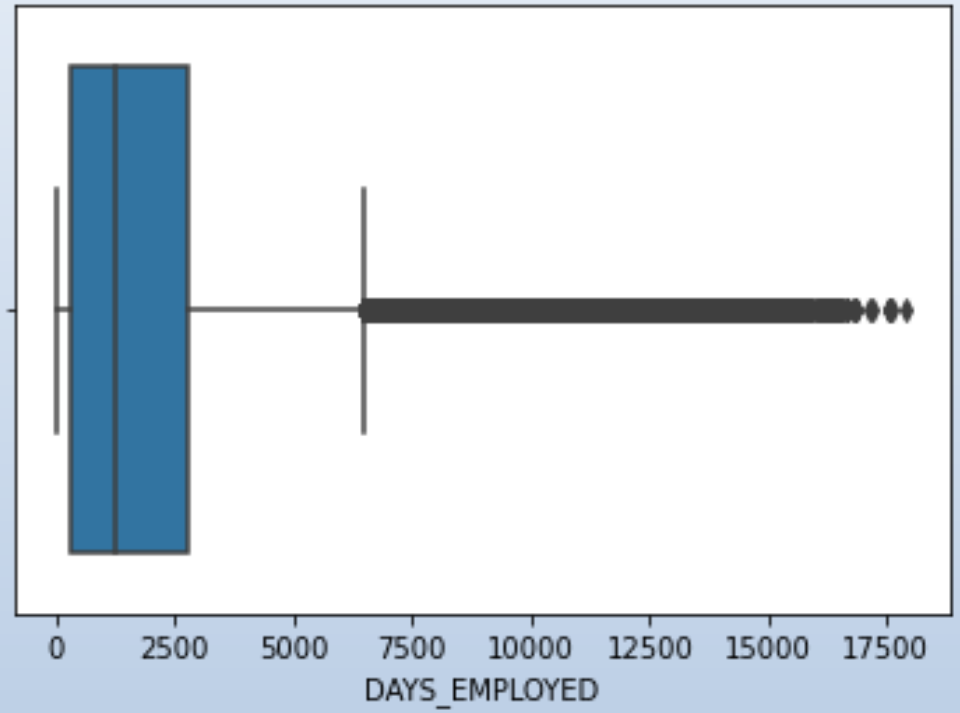


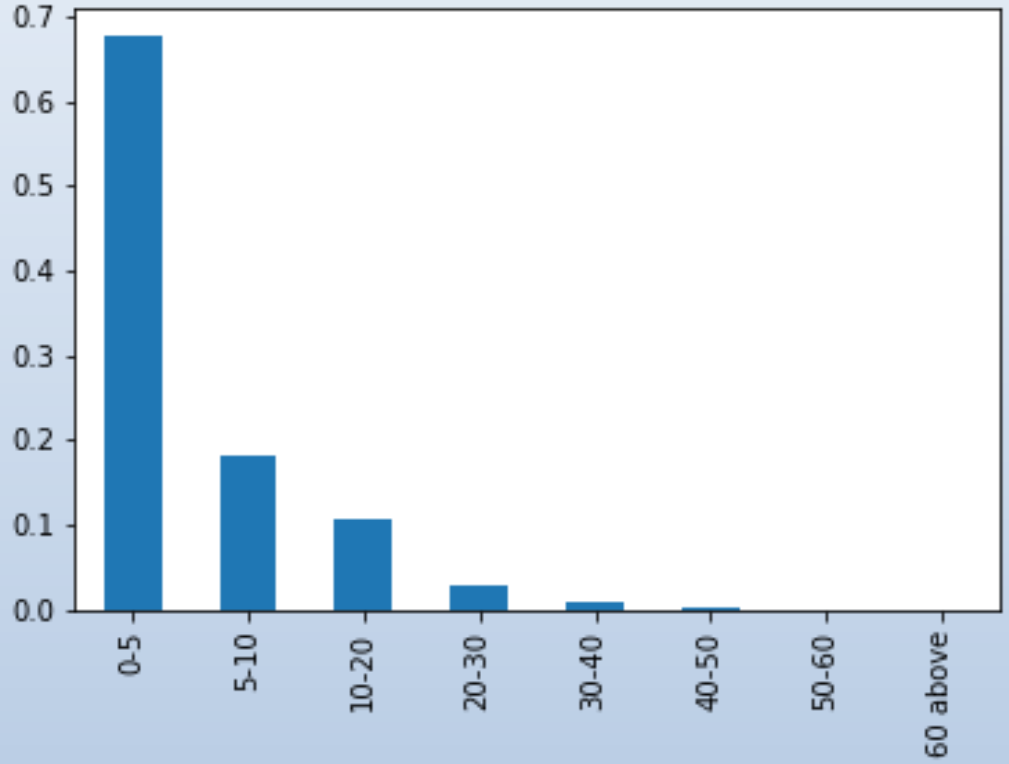


- Age variable had no any outliers but buckets of age range was created for better analysis

HANDLING OUTLIER

3. DAYS_EMPLOYED

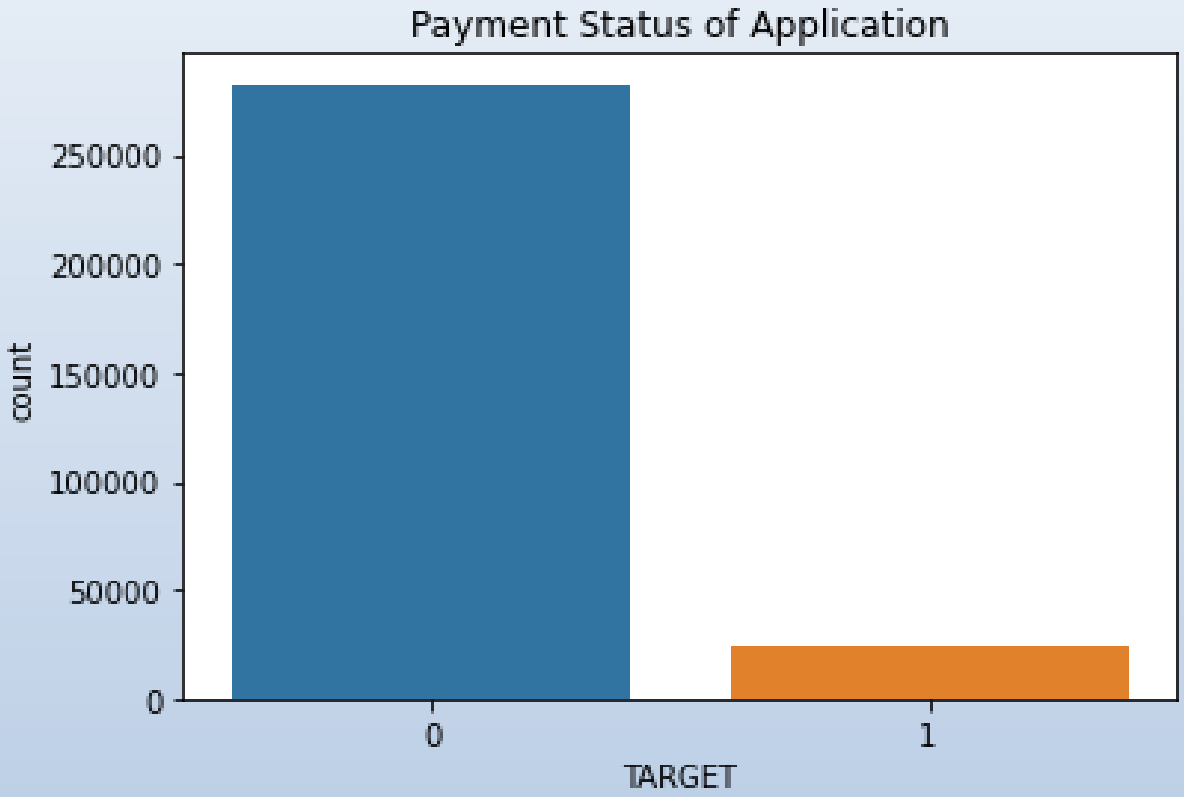




- DAYS_EMPLOYED converted into years to have a better view for analysis and them YEARS_EMPLOYED were bucketed to range and found that 0-5 years employed showed the spike

ANALYSIS OF DATA

Checking Data Imbalance



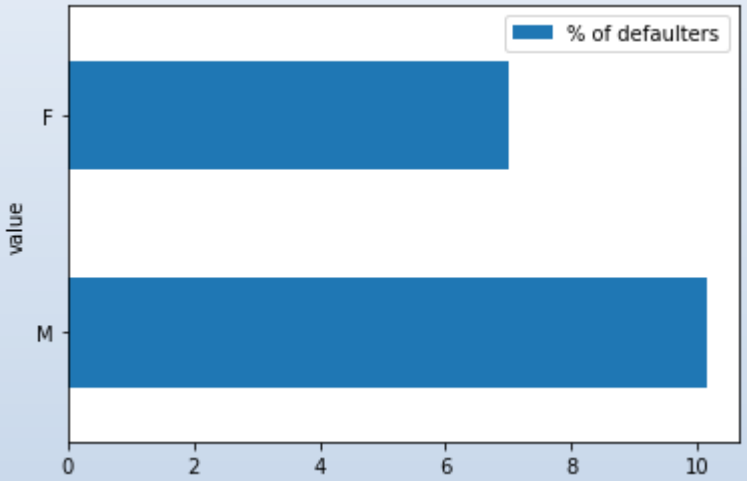
As the Target column indicates the defaulters and repayers we just need to separate the 2 data frames one with 1's and another with 0's

- The data is highly imbalance as defaulters to Repayers ratio is high i.e defaulters are very less in total data.
- Imbalance ratio - Defaulters: Repayers = 8:92

UNIVARIATE ANALYSIS

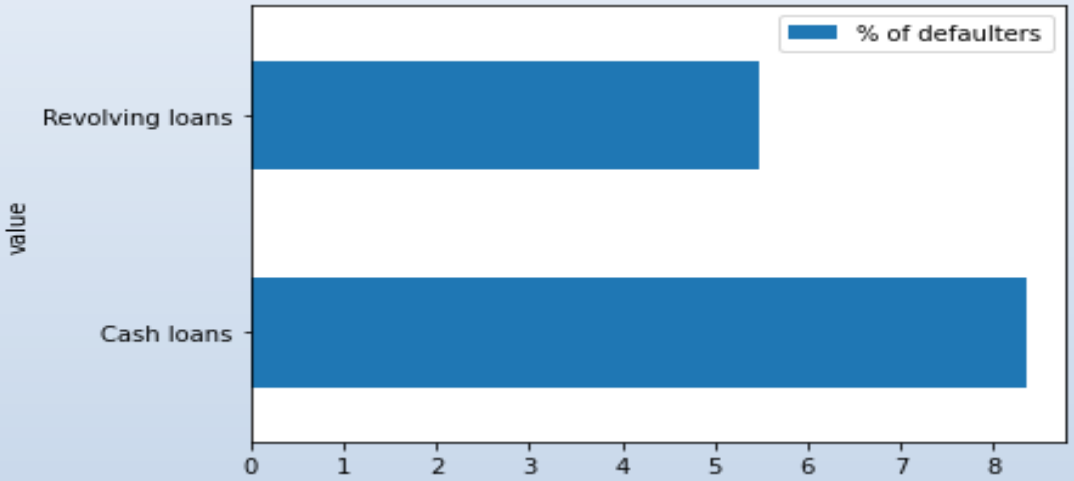
Categorical segmented Univariate Analysis

1. CODE_GENDER



1) Female candidates are more likely to repay the loan than male candidates.

2. NAME_CONTRACT_TYPE

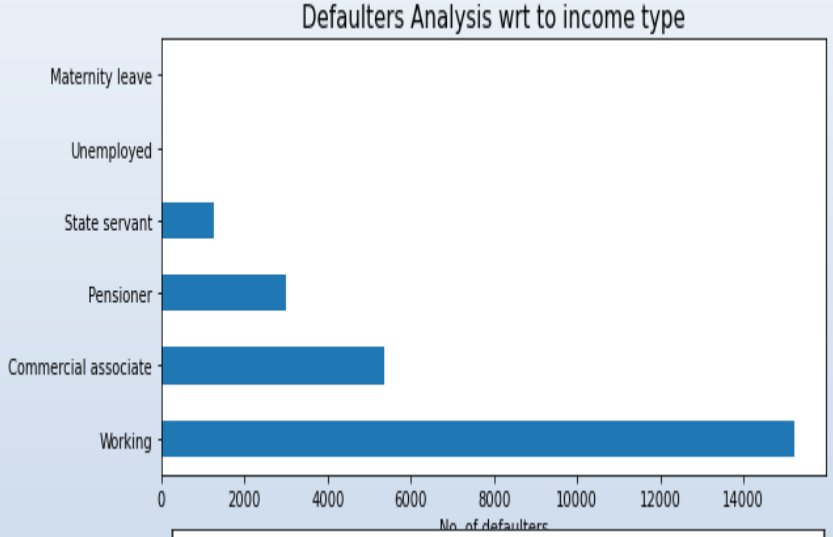


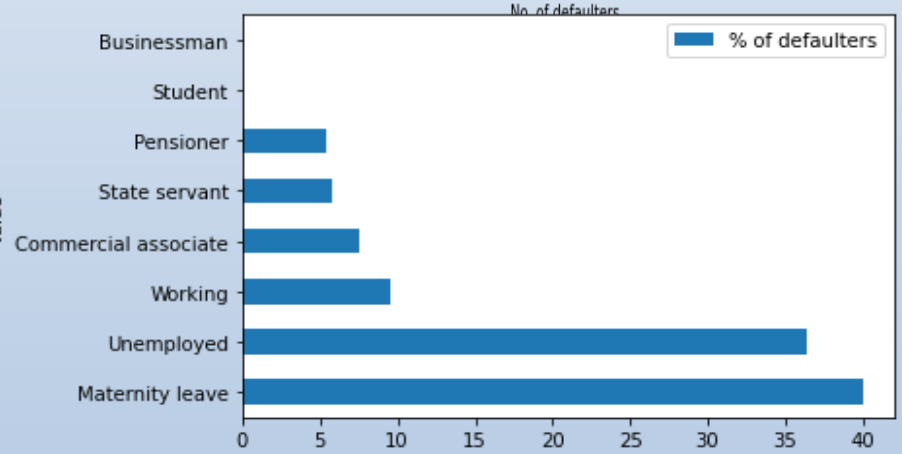
1) Candidates who have taken cash loans are more trouble repaying the loan

UNIVARIATE ANALYSIS

Categorical segmented Univariate Analysis

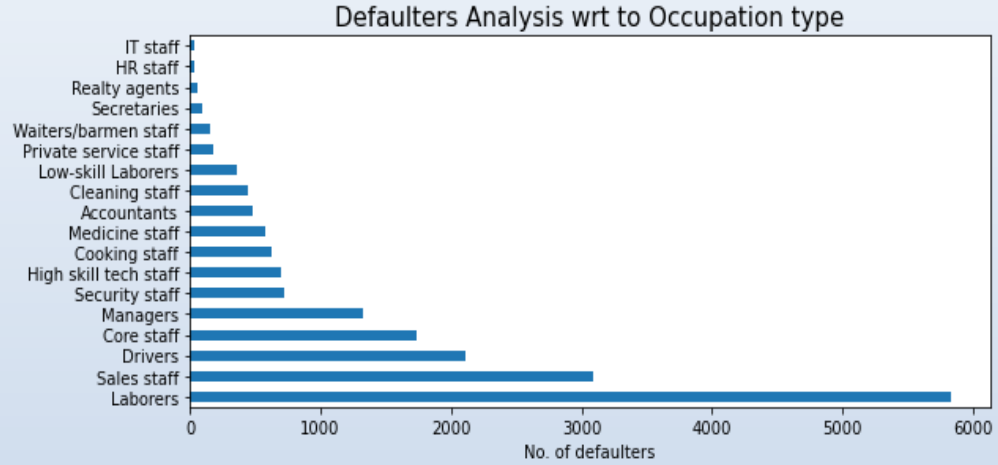
3. NAME_INCOME_TYPE

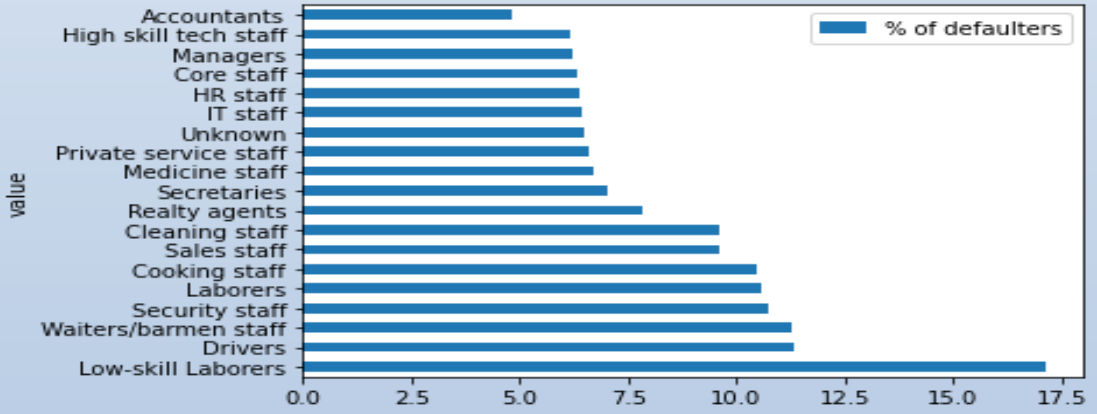




1) Most application are from working professionals
2) But applicants with maternity leave type are more likely not to repay the loans then followed by unemployed category.
3) Businessman and Student with least defaults

4. OCCUPATION_TYPE





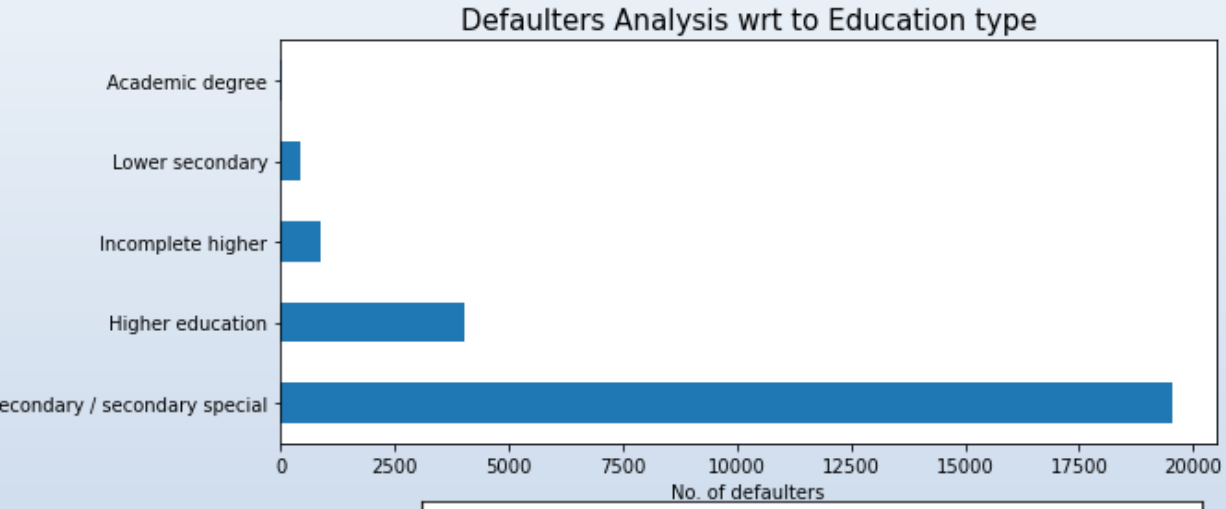
1) Most application are from labor occupation

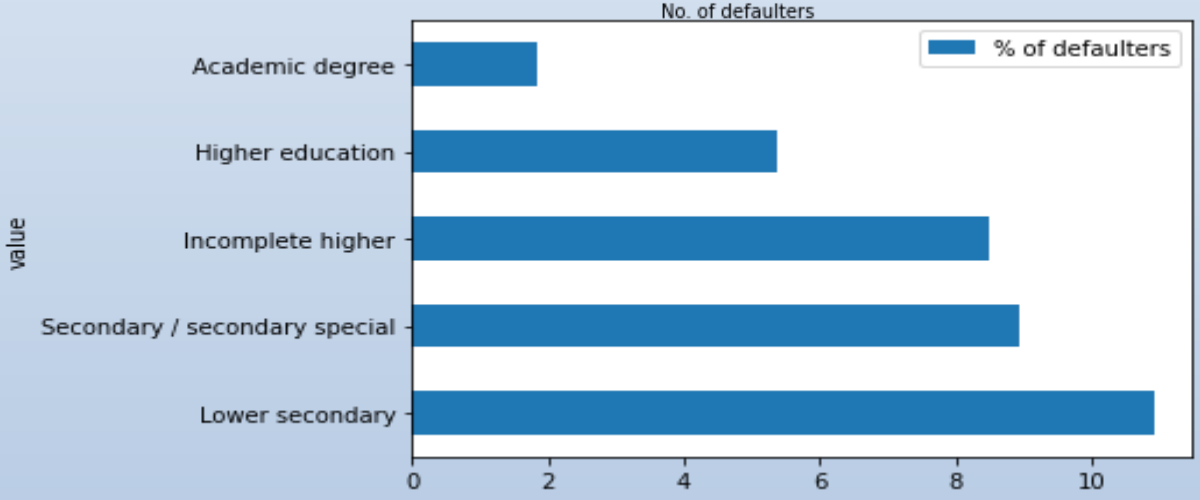
2) But applicants of low-skill laborers occupation are more likely not to repay the loans.

UNIVARIATE ANALYSIS

Categorical segmented Univariate Analysis

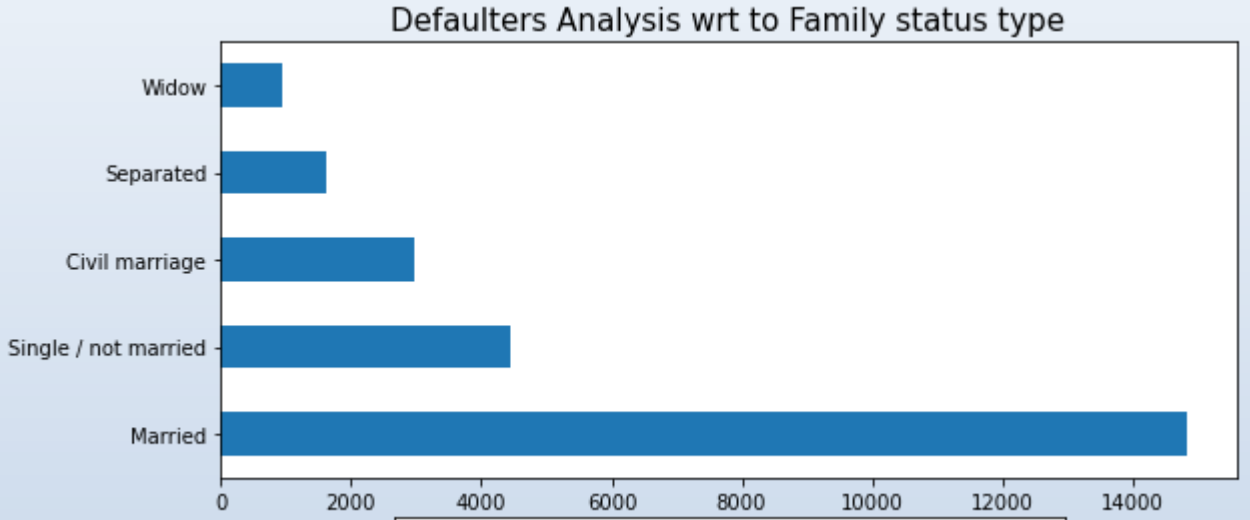
5. NAME_EDUCATION_TYPE

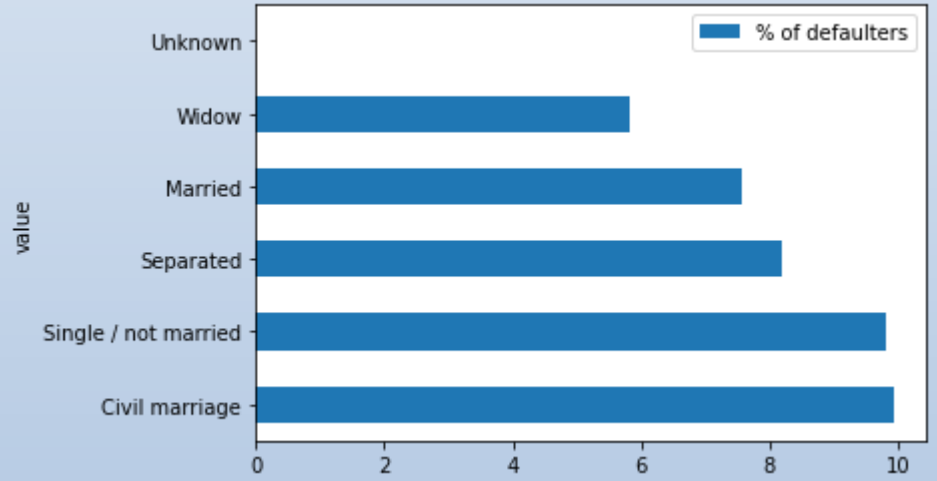




1) Most application are having secondary education.
2) But applicants with lower secondary education are more likely not to repay the loans then followed by secondary education.
3) Academic degree and higher education people are less likely of default.

6. NAME_FAMILY_STATUS



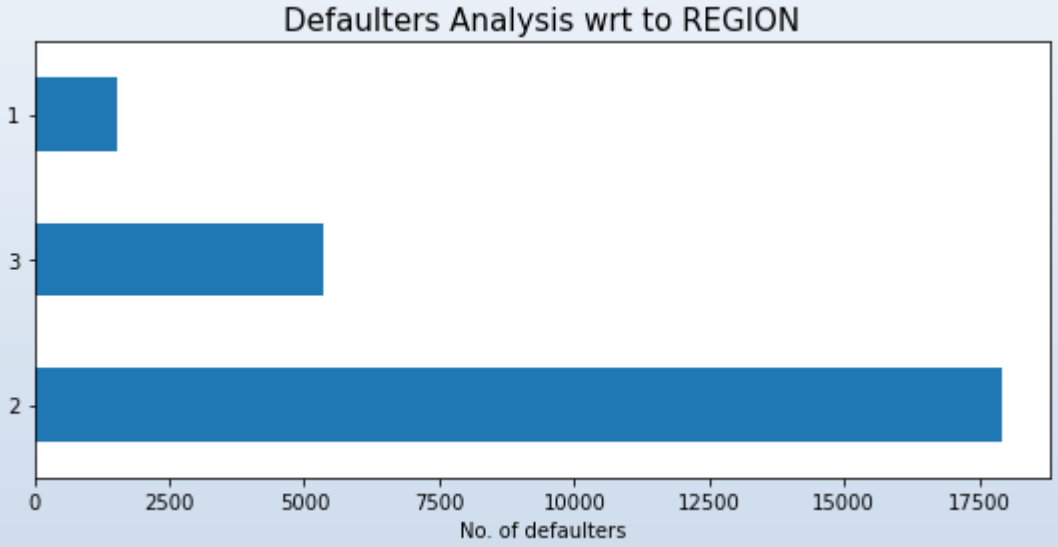


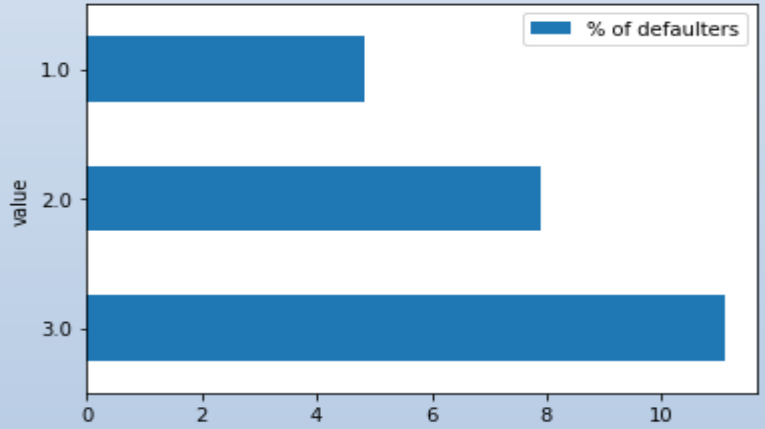
1) Most application are from Married people.
2) But applicants with civil marriage and single are more likely not to repay the loans.
3) Widow are more likely repay.

UNIVARIATE ANALYSIS

Categorical segmented Univariate Analysis

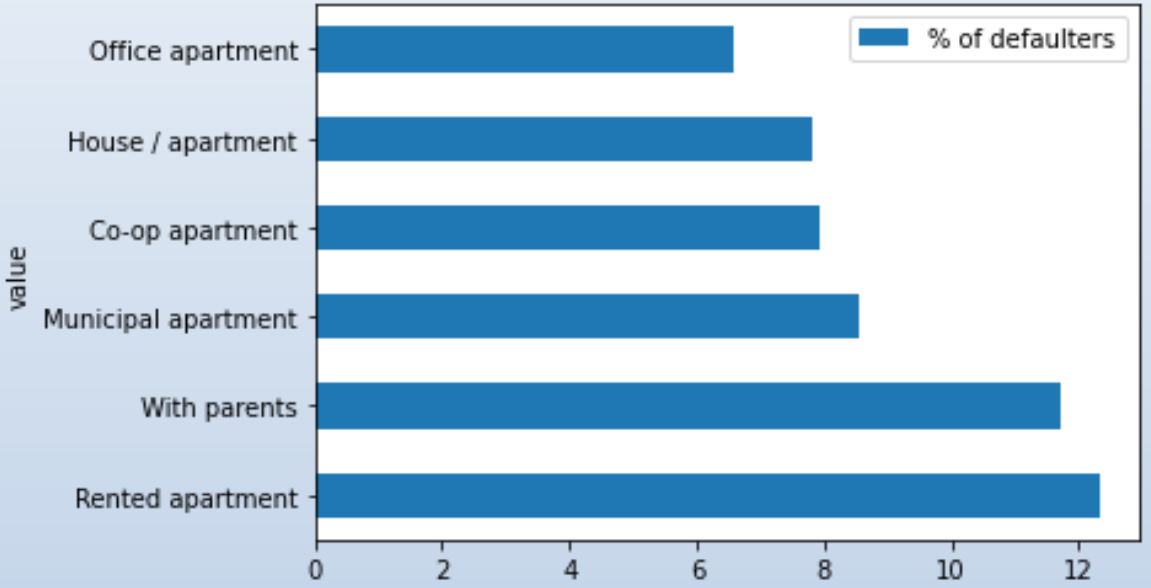
7. REGION_RATING_CLIENT





1) Most application are from Region 2 .
2) But applicants from region 3 are more likely not to repay the loans.
3) Region 1 applicants are more likely repay.

8. NAME_HOUSING_TYPE

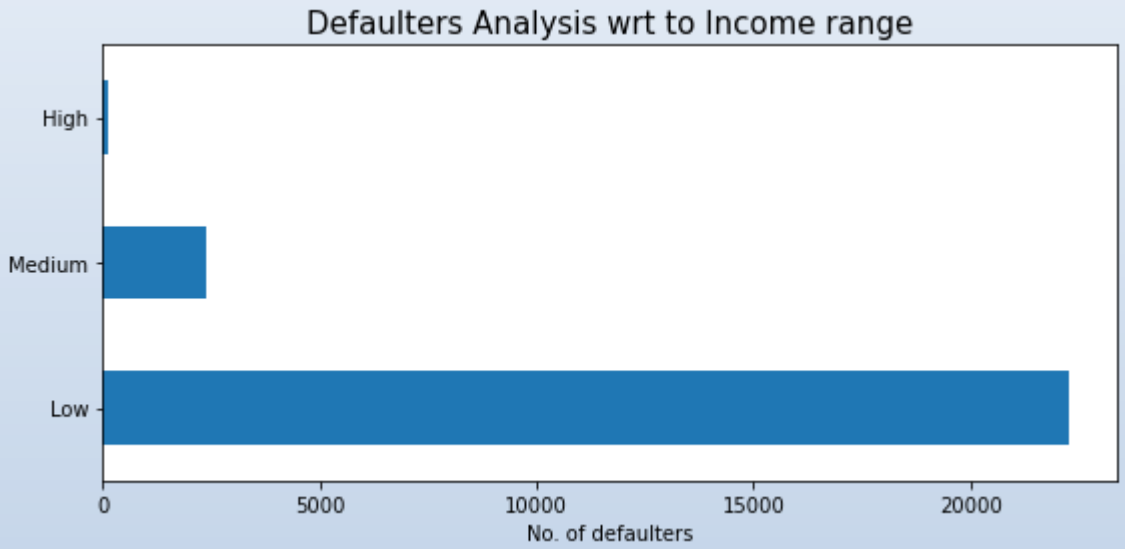


1. people who stay in rented house and with parents are more likely to default.

UNIVARIATE ANALYSIS

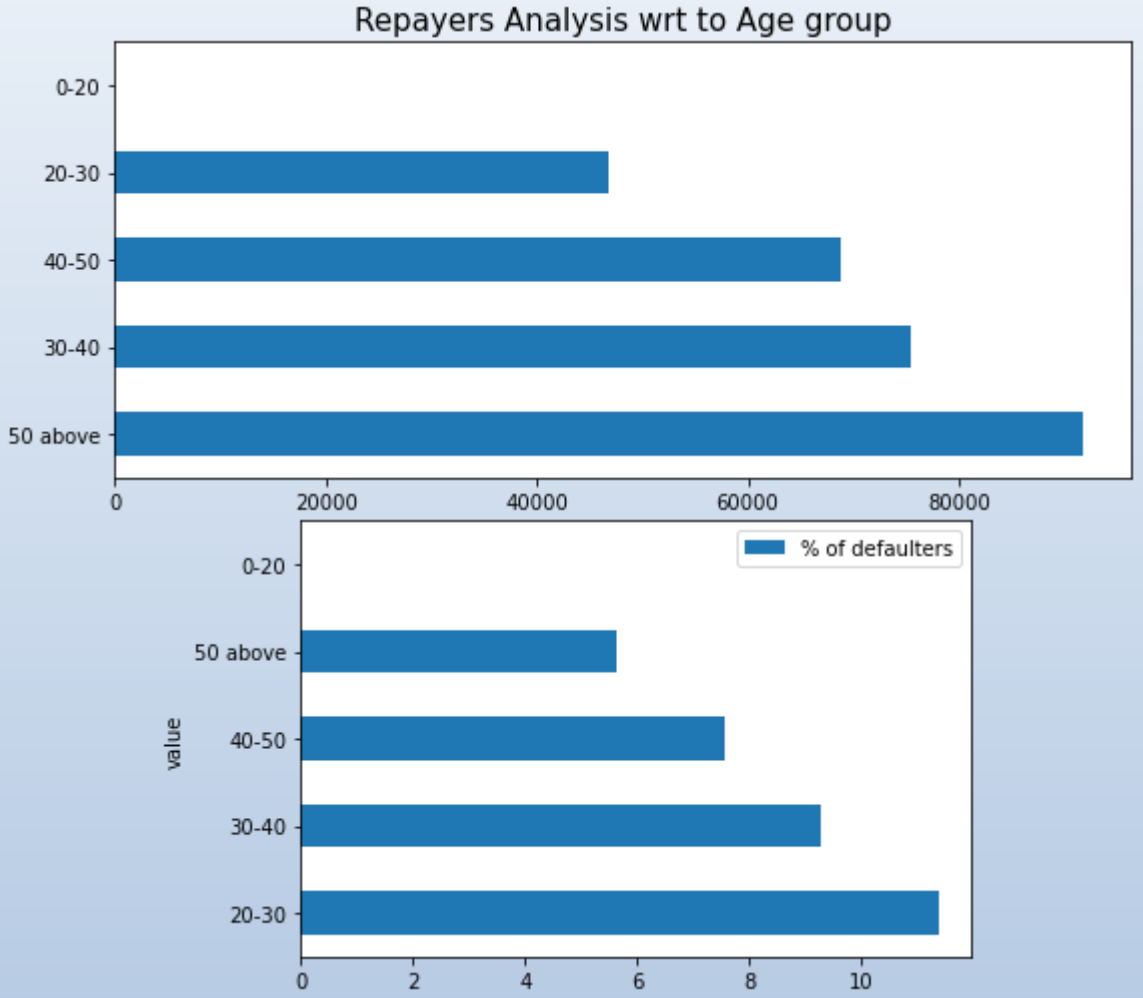
Categorical ordered univariate analysis

1. INCOME_RANGE



1) Major applications are from low income range (0-250000).
2) Applicants whose income is in between 0 to 250000 are default.
.

2. AGE_GROUP

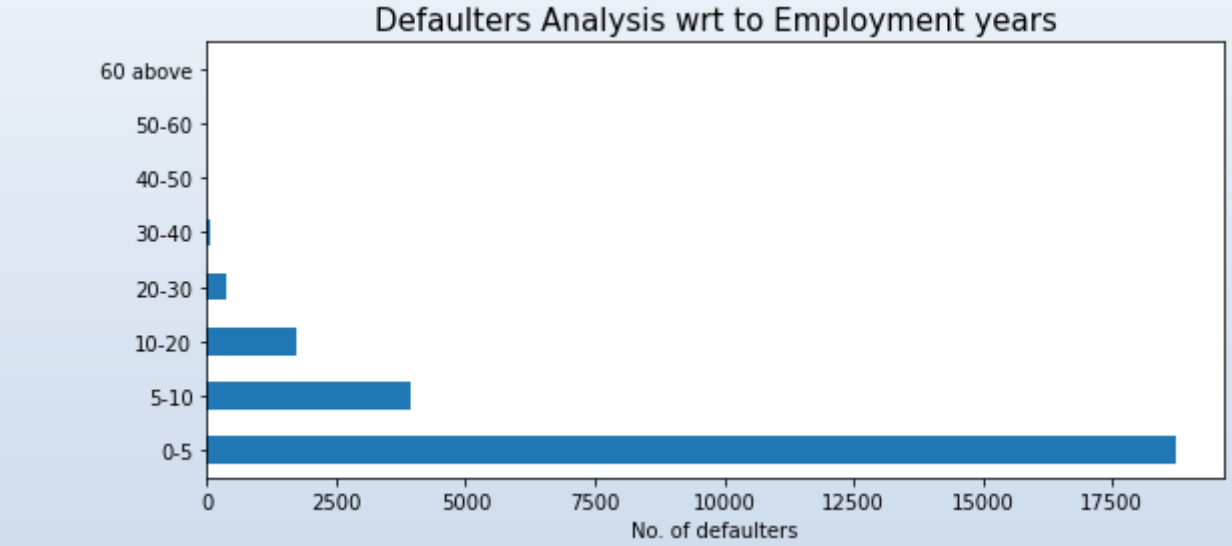


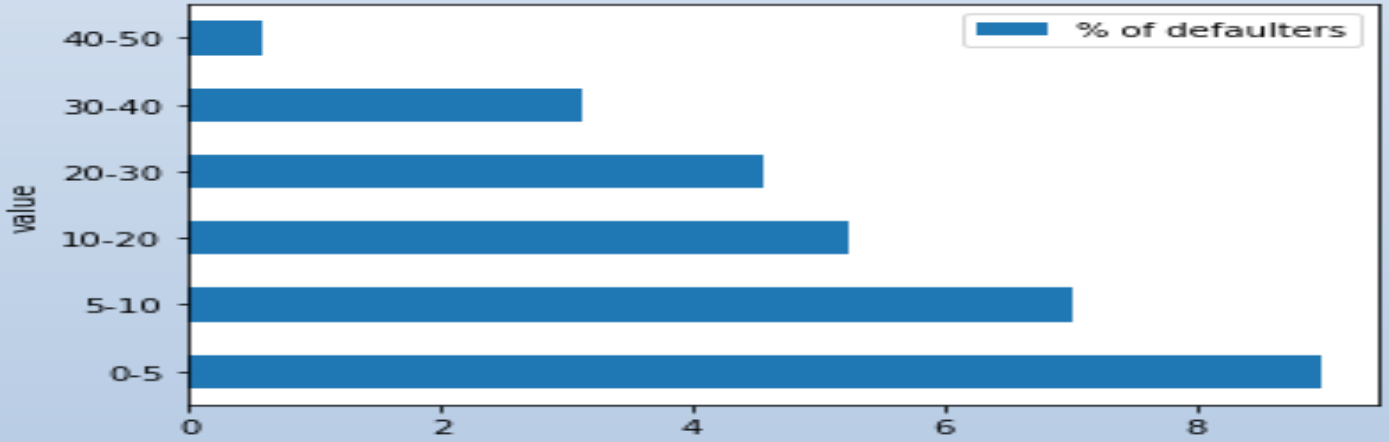
1) Applicants with age group 20-30 are more likely to default.
2) But applicants who are above 50 are very good repayers.

UNIVARIATE ANALYSIS

Categorical ordered univariate analysis

3. YEARS_EMPLOYED



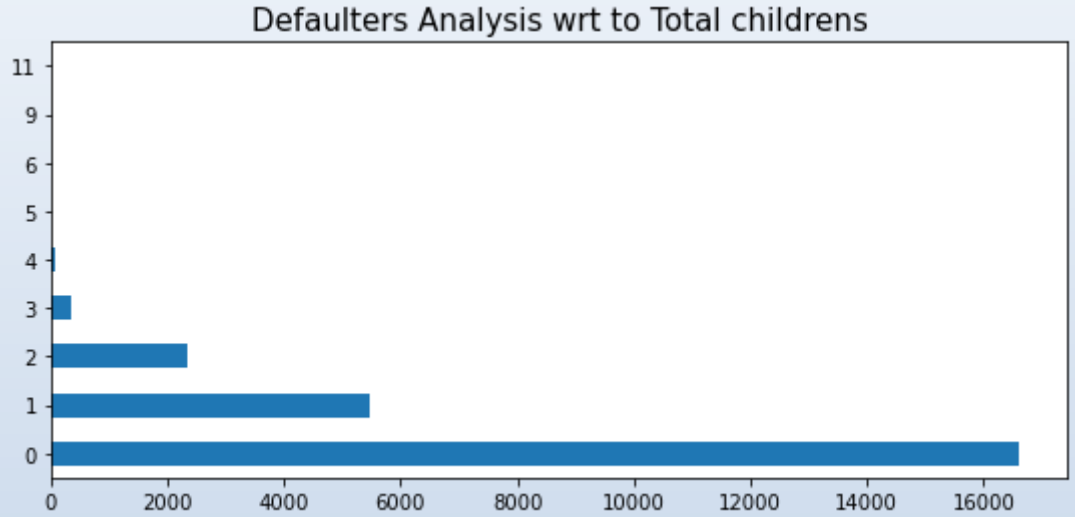


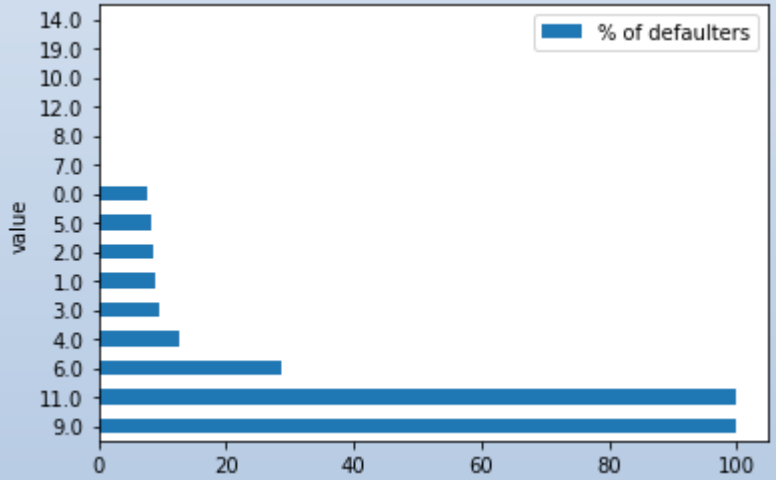
1) Applications are high from work experience 0-5 years.
2) If we observe the % defaulters are also high for applicants work experience 0-5 years.

UNIVARIATE ANALYSIS

Categorical ordered univariate analysis

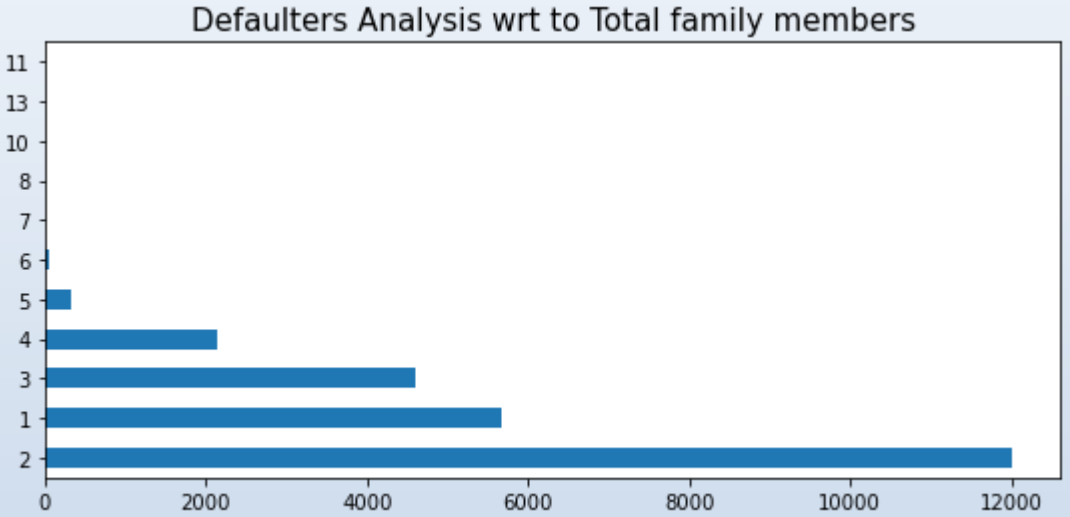
4. CNT_CHILDREN

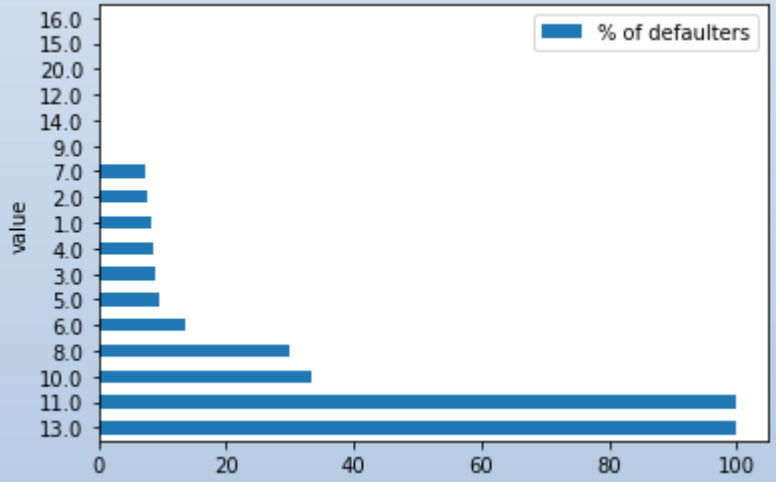




1) Most application are from applicants having no children(or may not be married).
2) But as the children's increases more than 3 the default rate increases (MAX- 9 & 11).
3) But also we need to see that applicants with more than 3 children's are very low.

5. CNT_FAM_MEMBERS



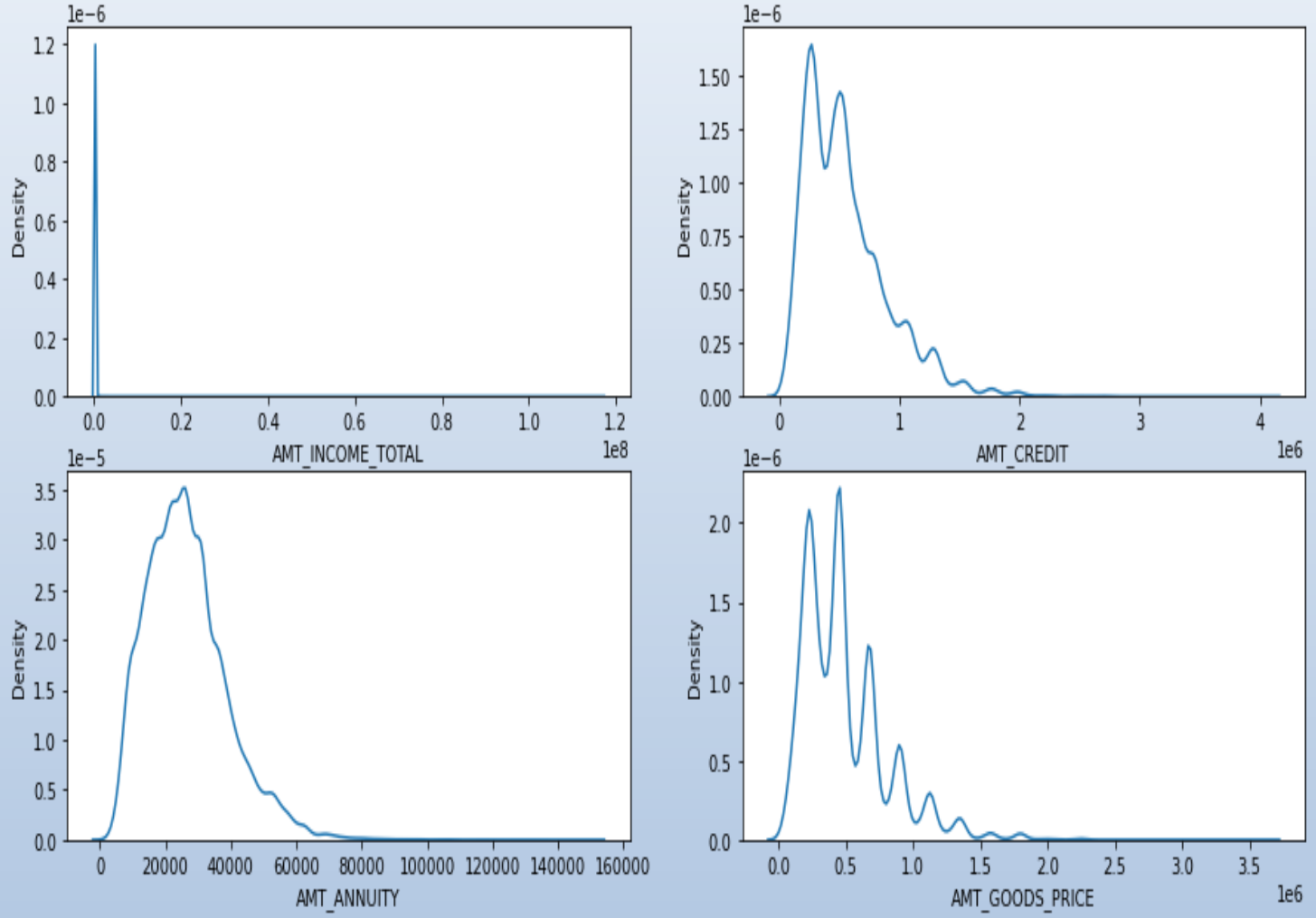


1) Most application are from applicants having 2 family members.
2) But as the family members increases the defaulters % increases.(MAX-13 members)
3) But also we need to see that applicants with more than 5 family members are very low.

UNIVARIATE ANALYSIS

Numerical univariate analysis

AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE



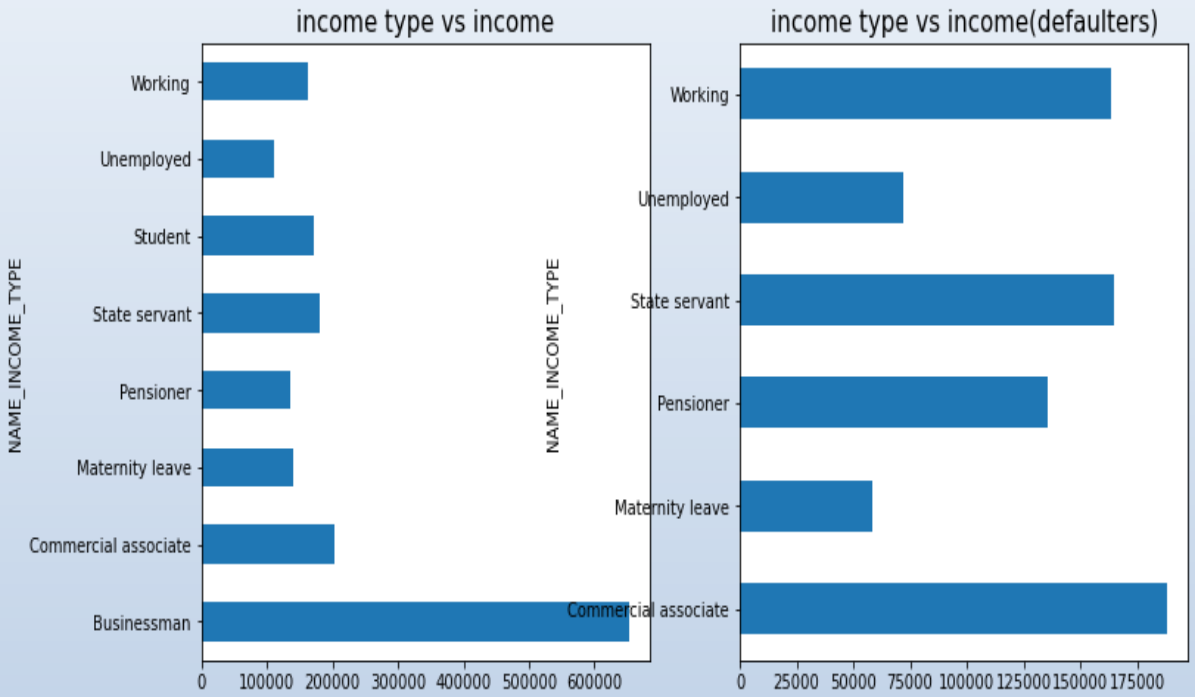
Inputs from numeric analysis of amount .

1) Income we have derived earlier that suggest lower the income greater the default rate .
2) Credit amount of loan is below 10 lakh.
3) Most people pay annuity below 40000 for the credit loan.
4) Most number of loans are given for goods price below 10 lakhs.
5) There are no much insights which are useful for analysis .

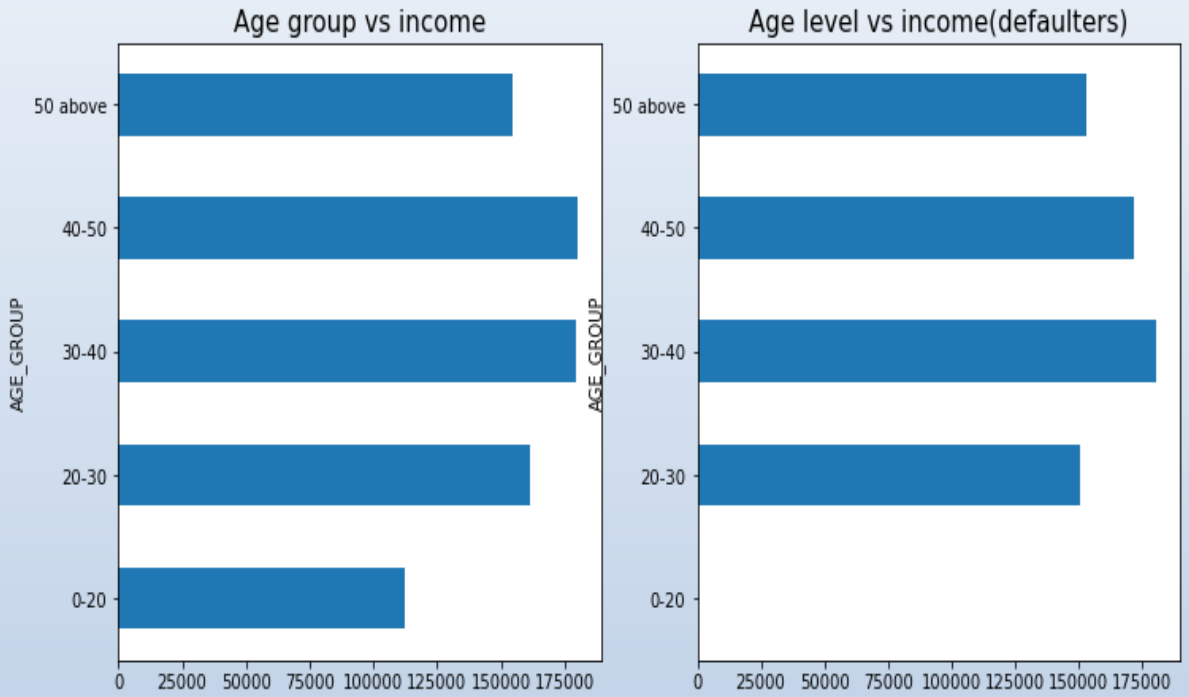
BIVARIATE ANALYSIS

Categorical vs numerical

Income type vs income



Age group vs income

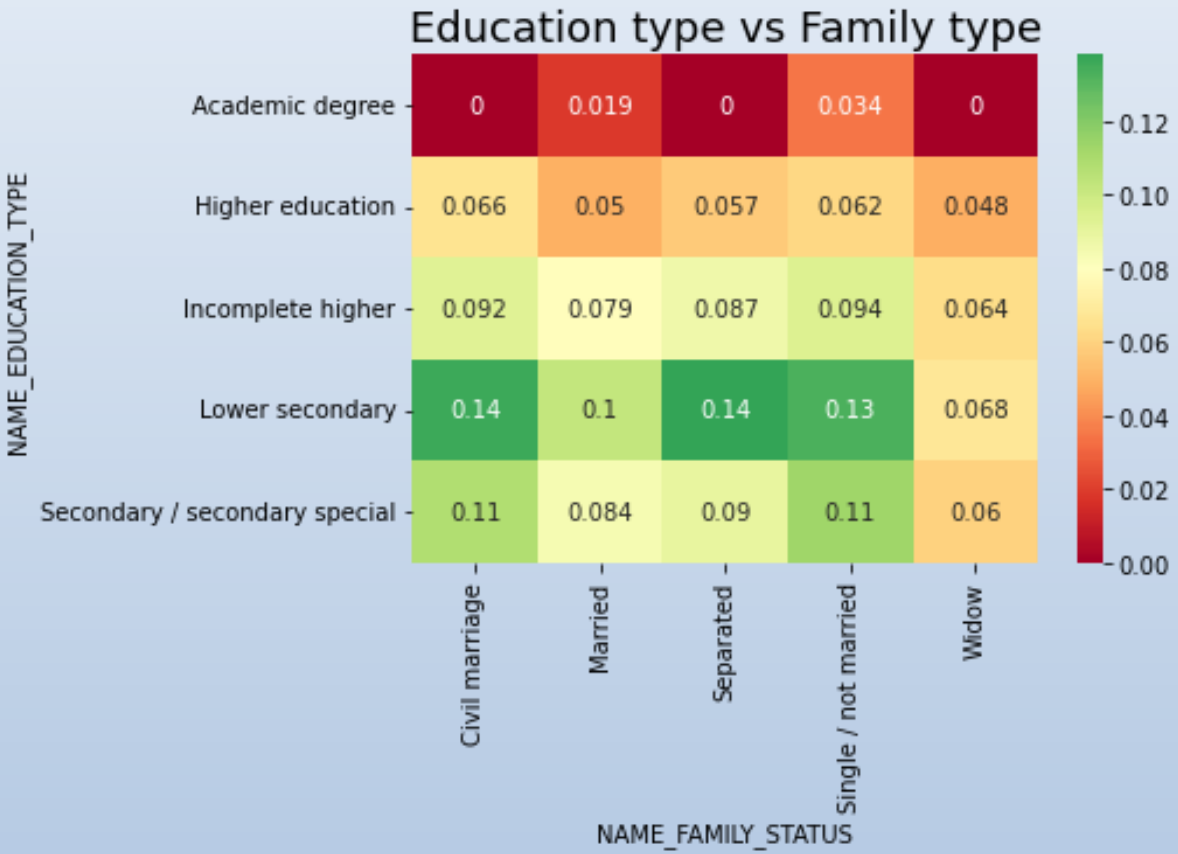


1) Businessman have highest income in category and also they are the once with no defaulters
2) Commercial associates have high defaulters rate

1) People with age group 40-50 have high income .
2) But defaulters are more in 30-40 range with high income

MULTIVARIATE ANALYSIS

Education type vs Family type



1) From heat map higher the correlation higher will be the defaulters rate.

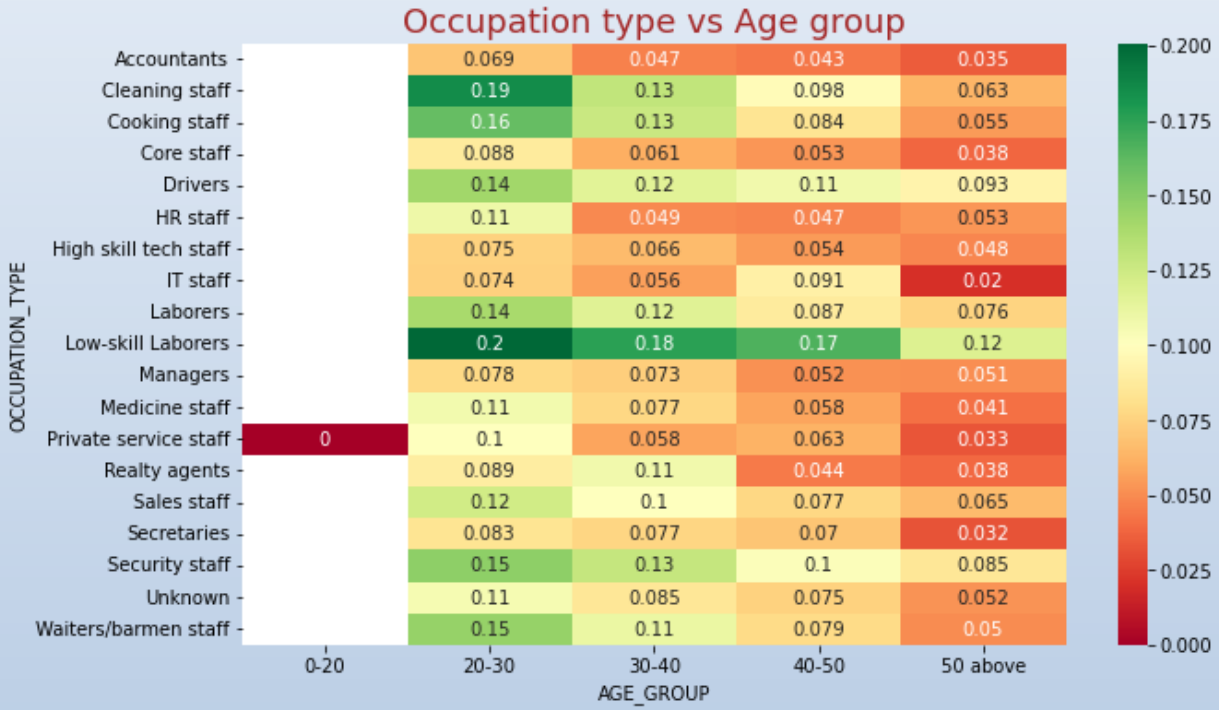
2) We can see that Lower secondary education with any family type except widow are more likely to be defaulters .

3) Even with secondary education of civil marriage abd single type have high default rate.

4) Academic degree people with any family type are very good repayers.

MULTIVARIATE ANALYSIS

Occupation type vs Age group



1) From heat map higher the correlation higher will be the defaulters rate.

2) We can see that people of age group 20-30 are more likely to be defaulters except some job type .

3) But Realty agents,medicine staff,managers,IT staff,High skilled tech staff,core staff,accountants are repayers in all age group.

4) We can see that people of age group 50 and above are more likely to be repayers.

TOP 10 CORRELATIONS OF DEFAULTERS AND REPAYERS

DEFAULTERS CORRELATIONS

OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998269
AMT_CREDIT	AMT_GOODS_PRICE	0.983103
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.956637
CNT_CHILDREN	CNT_FAM_MEMBERS	0.885484
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.868994
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778540
AMT_ANNUITY	AMT_GOODS_PRICE	0.752699
	AMT_CREDIT	0.752195
REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.497937

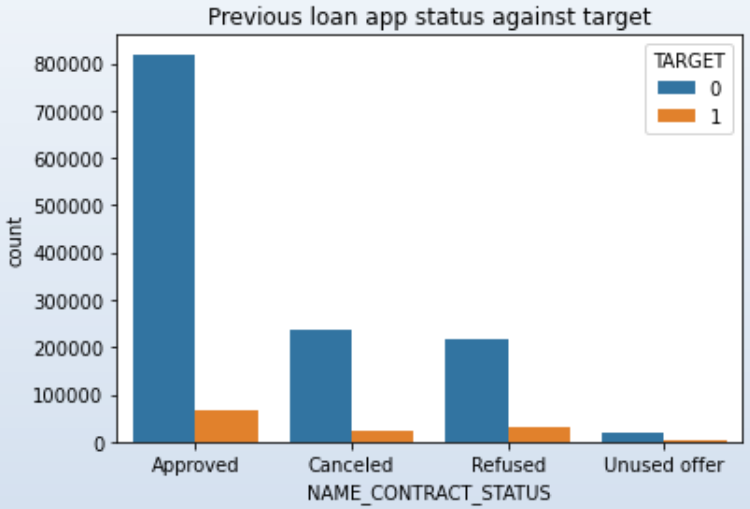
1) Count of children's is highly correlated with count of family members

REPAYERS CORRELATIONS

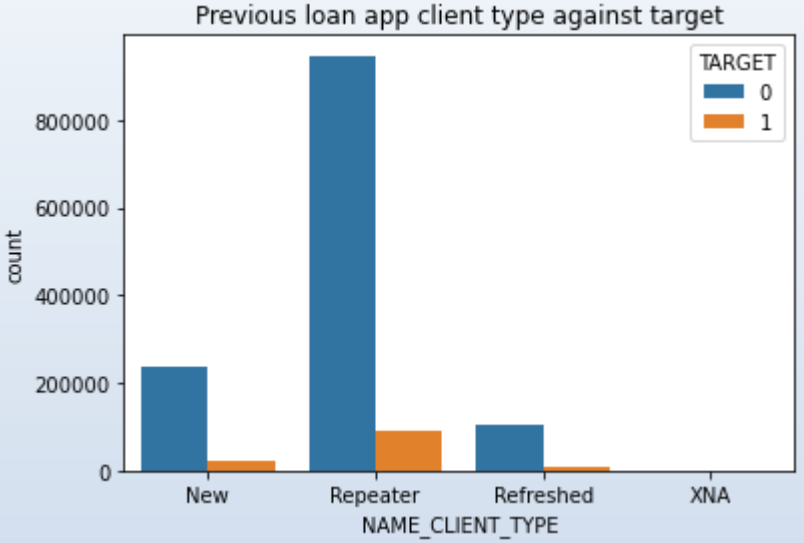
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508
AMT_CREDIT	AMT_GOODS_PRICE	0.987251
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950146
CNT_CHILDREN	CNT_FAM_MEMBERS	0.878574
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.861822
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859331
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830369
AMT_ANNUITY	AMT_GOODS_PRICE	0.776686
AMT_CREDIT	AMT_ANNUITY	0.771309
REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	0.446093

1) Credit amount is highly correlated with amount of goods price , loan annuity, total income

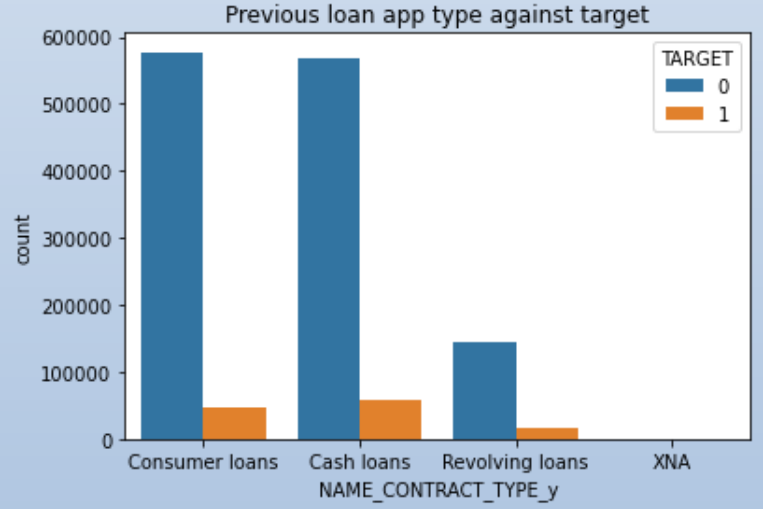
ANALYSIS OF PREV DATA WITH APP DATA



1) We can say that maximum rate of people whose loan was approved previously are likely to repay the loan .
2) Even people whose prev application was canceled and refused previously have turned into potential repayers .



1) We can say that maximum client type are repeaters and this is good for business with less default



1) We can say that maximum loans are applied for consumer and cash type with less default

CONCLUSION FACTORS FOR APPLICANTS WHO ARE REPAYERS & DEFAULTERS

CONCLUSION FACTORS FOR APPLICANTS WHO ARE REPAYERS

1) CODE_GENDER : Female candidates are more likely to repay the loan .
2) NAME_CONTRACT_TYPE : Candidates who have taken Revolving loans are more likely repaying the loan.
3) NAME_INCOME_TYPE :Businessman and Student with least defaults rate.
4) NAME_EDUCATION_TYPE : Academic degree and higher education people are with least defaults rate.
5) REGION_RATING_CLIENT : Region 1 applicants are more likely repay.
6) AMT_INCOME_TOTAL: Applicant with Income more than 700,000 are less likely to default.
7) AGE_GROUP : Applicants who are above 50 are very good repayers.
8) CNT_CHILDREN: People with zero to two children are likely to repay the loans.
9) CNT_FAM_MEMBERS: Lower the family members applicant is more likely to repay the loan
10) YEARS_EMPLOYED: Applicants with 40+ year experience having less default rate

CONCLUSION FACTORS FOR APPLICANTS WHO ARE DEFAULTERS

1) CODE_GENDER : Male candidates are having high default rate.
2) NAME_CONTRACT_TYPE : Candidates who have taken cash loans are more trouble repaying the loan.
3) NAME_INCOME_TYPE :But applicants with maternity leave type are more likely not to repay the loans then followed by unemployed category.
4) OCCUPATION_TYPE: Applicants of low-skill laborers occupation are more likely not to repay the loans
5) NAME_EDUCATION_TYPE : Applicants with lower secondary education are more likely not to repay the loans then followed by secondary education.
6) REGION_RATING_CLIENT : Applicants from region 3 are more likely not to repay the loans.
7) AMT_INCOME_TOTAL:) Applicants whose income is in between 0 to 250000 are more likely to default.
8) AGE_GROUP : Applicants with age group 20-30 are more likely to default.
9) CNT_CHILDREN: The children’s increases more than 3 the default rate increases.
10) CNT_FAM_MEMBERS: The family members increases the defaulters % increases.(
11) YEARS_EMPLOYED: The % defaulters are high for applicants work experience 0-5 years
12) NAME_FAMILY_STATUS: Applicants with civil marriage and single are more likely not to repay the loans.

FACTORS FOR APPLICANTS WHO ARE DEFAULTERS BUT MAJOR CONTRIBUTERS

FACTORS FOR APPLICANTS WHO ARE DEFAULTERS BUT MAJOR CONTRIBUTERS

1) INCOME_RANGE : Applicants whose income is in between 0 to 250000 are relatively more but aslo they are high defaulters so we can provide loan on high interest risk.
2) NAME_HOUSING_TYPE : People who stay in rented house and with parents are more likely to default but they are major applicants for loan so can be considered for high interest loan.
3)CNT_CHILDREN & CNT_FAM_MEMBERS: Clients who have more than 3 children has a very high default rate and hence higher interest can be imposed on their loans.