upGrad

PG Diploma in Data Science Sep 2020

EDA Case Study

By

Srihari R

Problem Statement

Introduction

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 as their advantage by becoming a defaulter. Using EDA, we have to analyze 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

Dataset

The data given contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

•**The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,

•**All other cases:** All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

1.**Approved:** The Company has approved loan Application

2.**Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.

3.**Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).

4.**Unused offer:** Loan has been cancelled by the client but on different stages of the process

Data Cleaning

Application Data

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR
0	100002	1	Cash loans	M	N
1	100003	0	Cash loans	F	N
2	100004	0	Revolving loans	M	Y
3	100006	0	Cash loans	F	N
4	100007	0	Cash loans	M	N

Shape before cleaning

Rows: 307511
Columns: 122

Shape after cleaning

Rows: 202142
Columns: 50

Previous Applications

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION
0	2030495	271877	Consumer loans	1730.430	17145.0
1	2802425	108129	Cash loans	25188.615	607500.0
2	2523466	122040	Cash loans	15060.735	112500.0
3	2819243	176158	Cash loans	47041.335	450000.0
4	1784265	202054	Cash loans	31924.395	337500.0

Shape before cleaning

Rows: 1670214
Columns: 37

Shape after cleaning

Rows: 1246320
Columns: 26

Merged Data

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE_x	CODE_GENDER	FLAG_OWN_CAR
0	100002	1	Cash loans	M	N
1	100004	0	Revolving loans	M	Y
2	100008	0	Cash loans	M	N
3	100008	0	Cash loans	M	N
4	100008	0	Cash loans	M	N

Shape before cleaning

Rows: 710658
Columns: 79

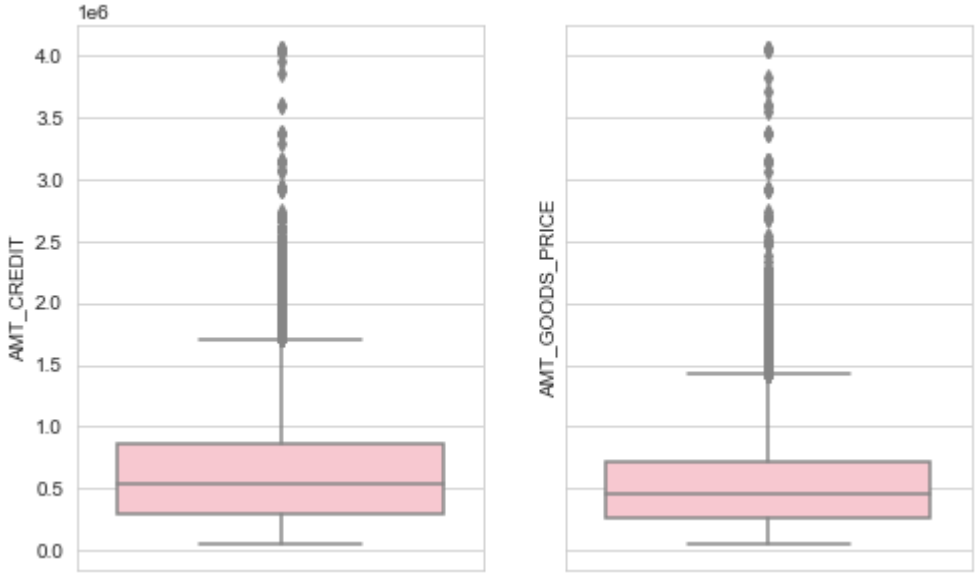
Shape after cleaning

Rows: 42685
Columns: 79

Some basic cleaning like removing columns which has more than 30% null values and removing datapoints which has XNA/XAP in it

Univariate Analysis

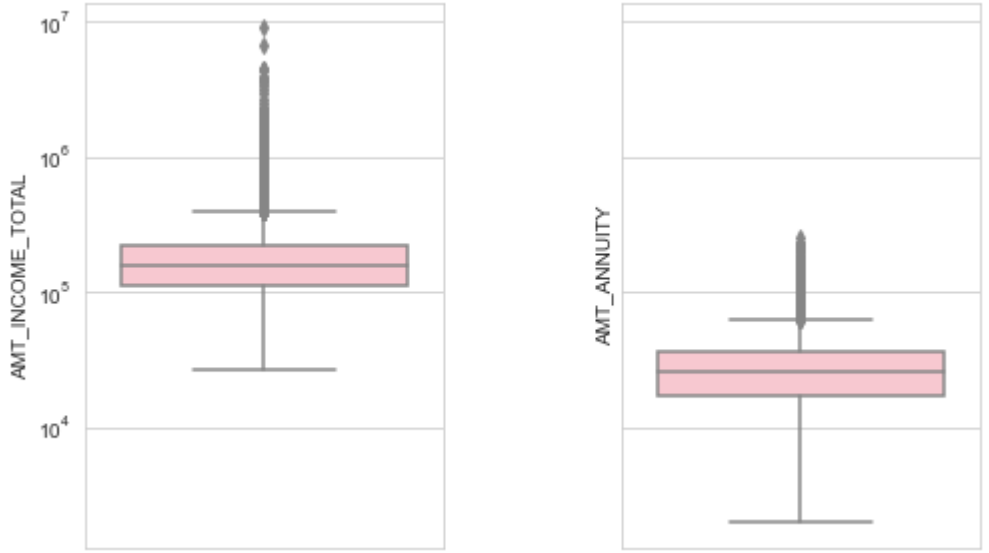
01 Boxplot for goods price and credit amount



Inferences

- Outliers have been identified
- Applicants mostly quote for more amount than the actual Goods price

02 Boxplot for annual income and annual annuity



Inferences

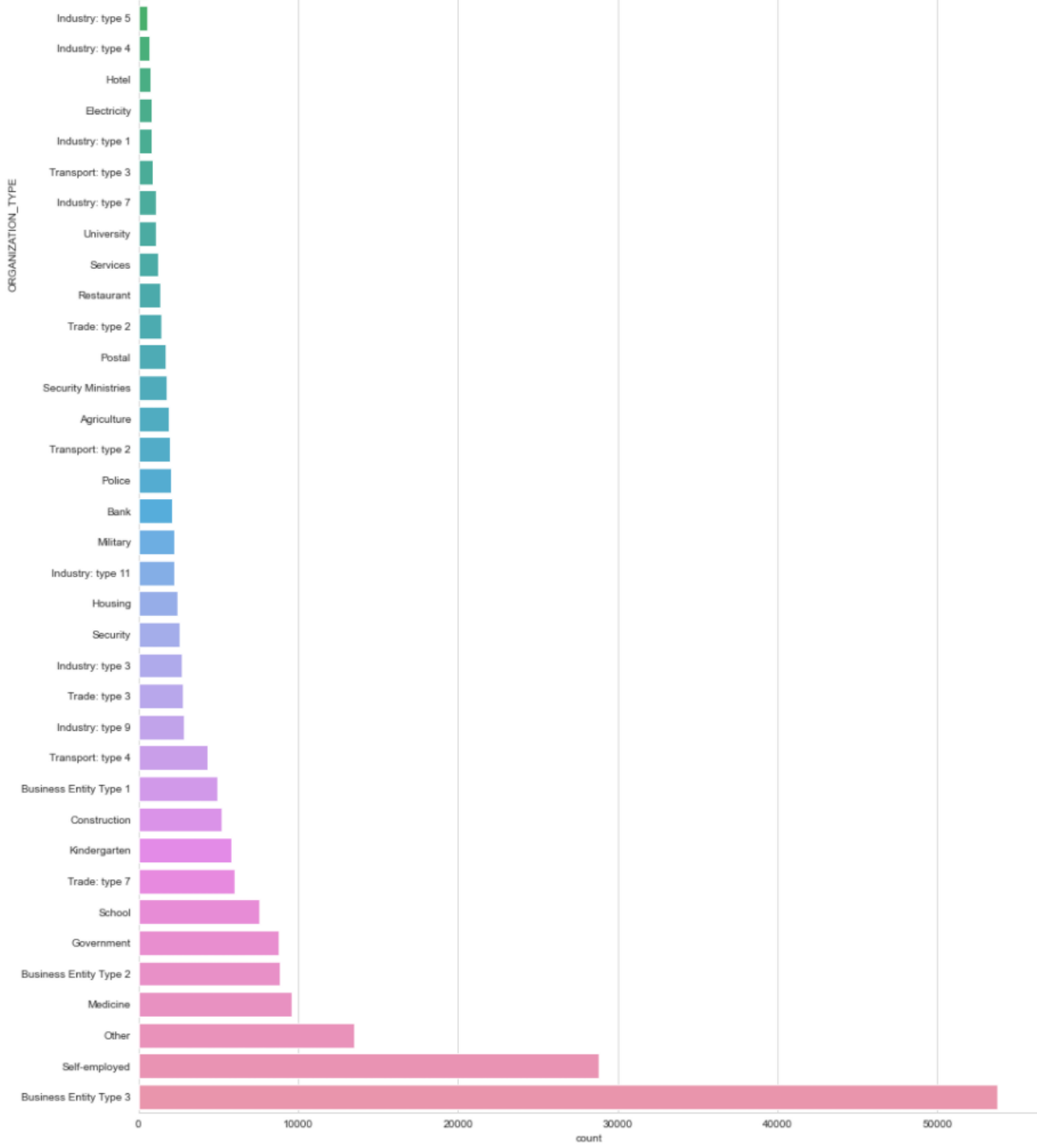
- Outliers have been identified
- Annual annuity is generally lesser than applicant's annual income

Countplot based on Organisation type

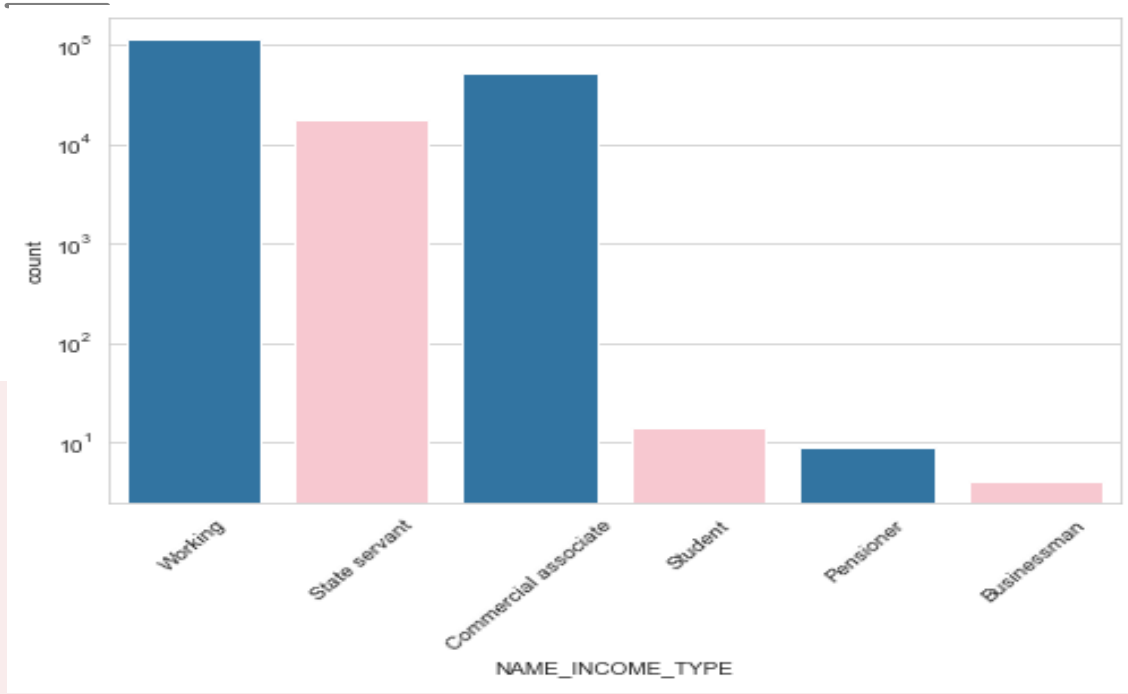
Inferences

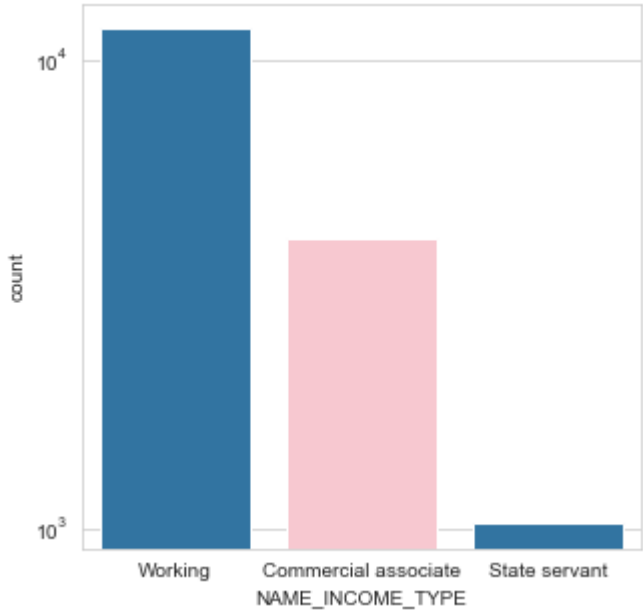
- Most of our applicants work in Business type 3
- Second highest count for the applicants is Self Employed
- very less applicants are from industry type 8

Note: Since this feature has over 57 unique types, not all could be shown in this single page. The entire plot is shown in the python notebook



Countplot based on income type

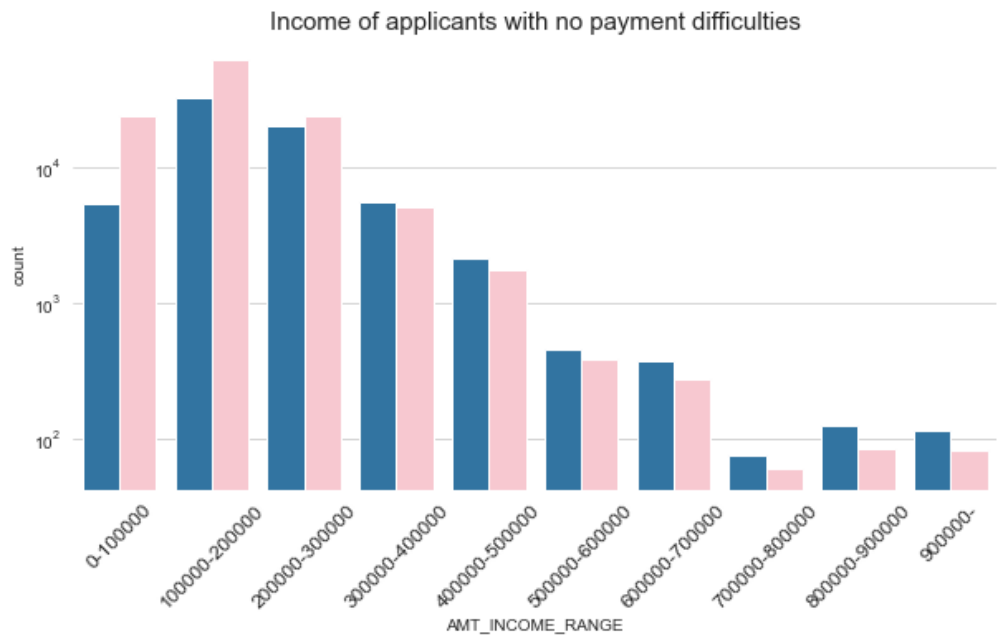


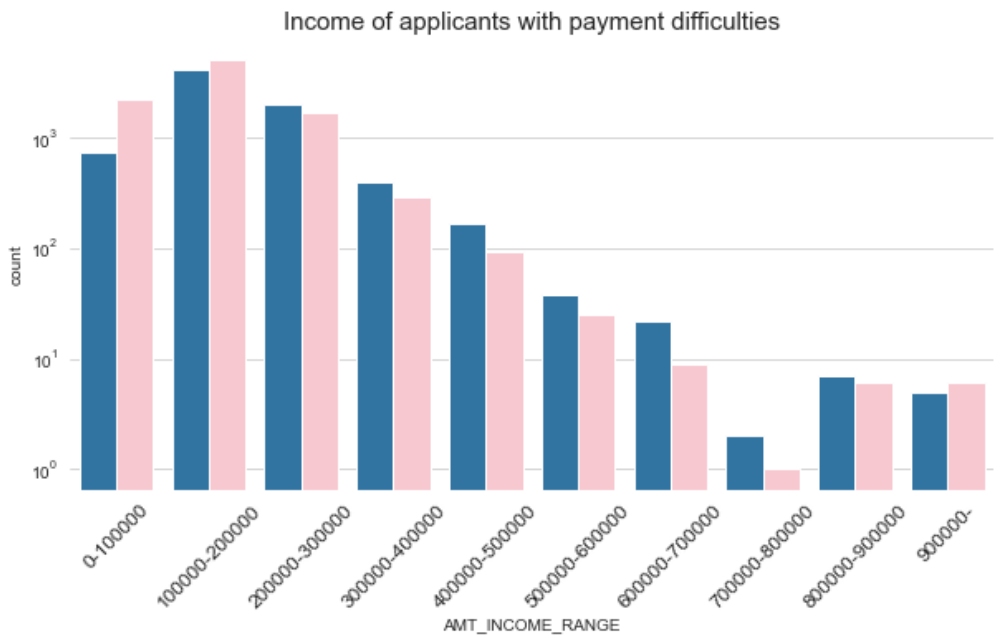


Inferences

- It seems that Students, Unemployed and Businessmen seems to have no trouble in payments
- Working professionals have higher count when compared to other

Countplot based on income type





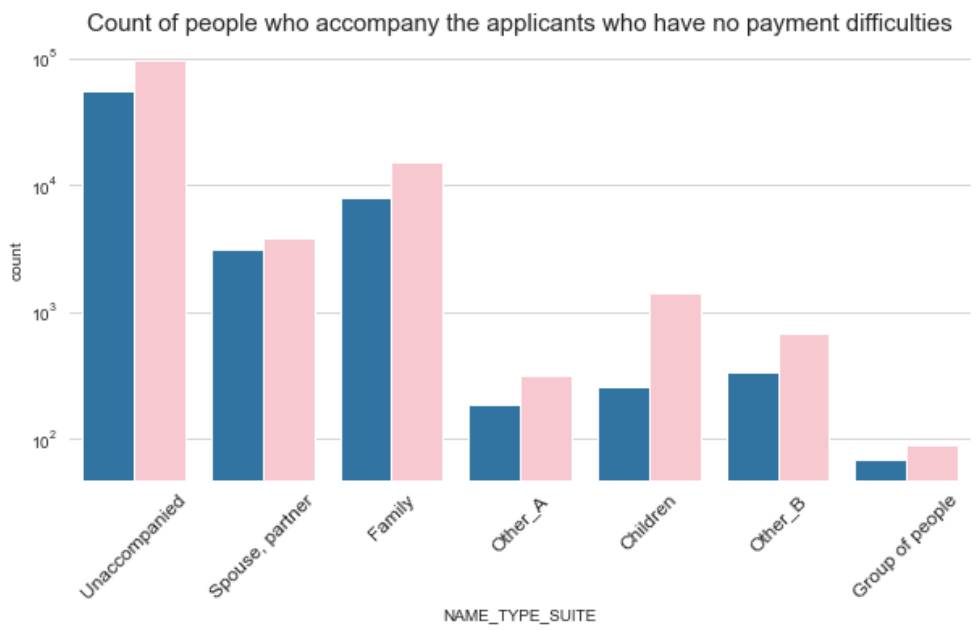
Inferences

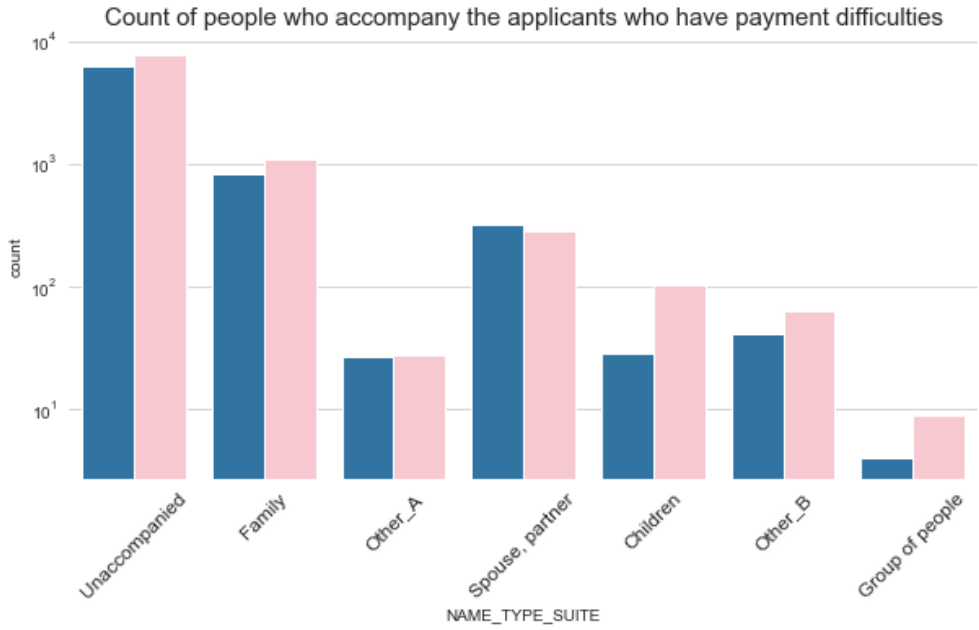
- It seems that Students, Unemployed and Businessmen seems to have no trouble in payments
- Working professionals have higher count when compared to other

Male

Female

Countplot based on people who accompanied the applicant



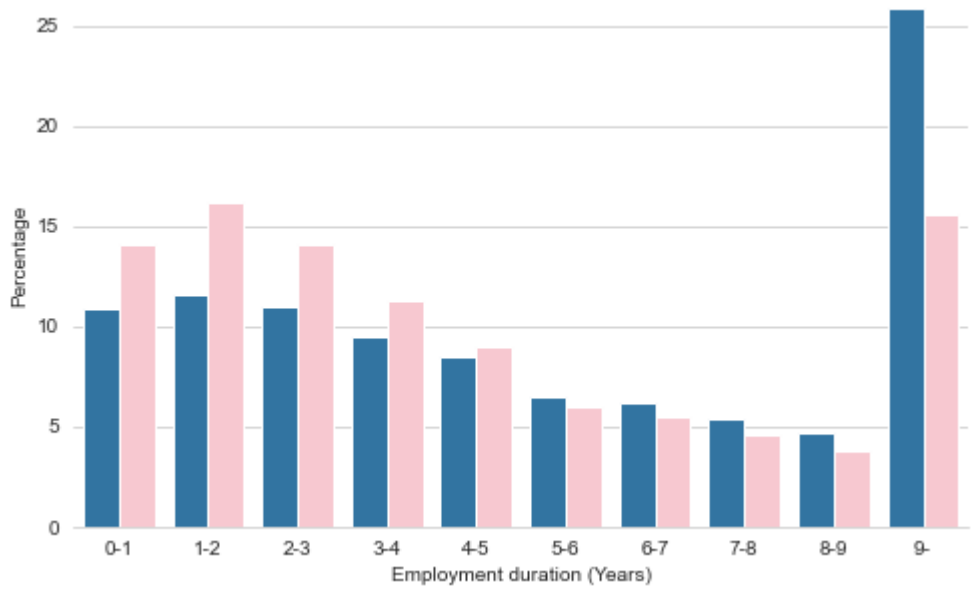


Inferences

- Most applicants come "Unaccompanied" at the time of applying
- Men who accompany with Spouse/partner tend to have payment difficulties when compared to women

Male
Female

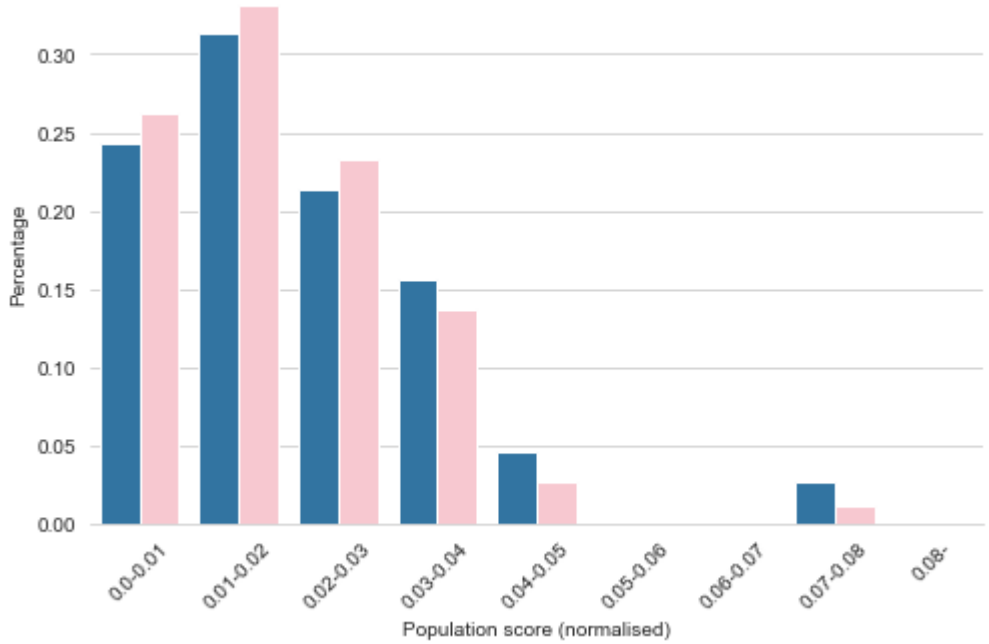
Count plot of days employed



Inferences

- applicants who are working for more than 6 years are less likely to default
- applicants who have changed employment within 0-5 years of time defaulted more.

Count plot of Population Score



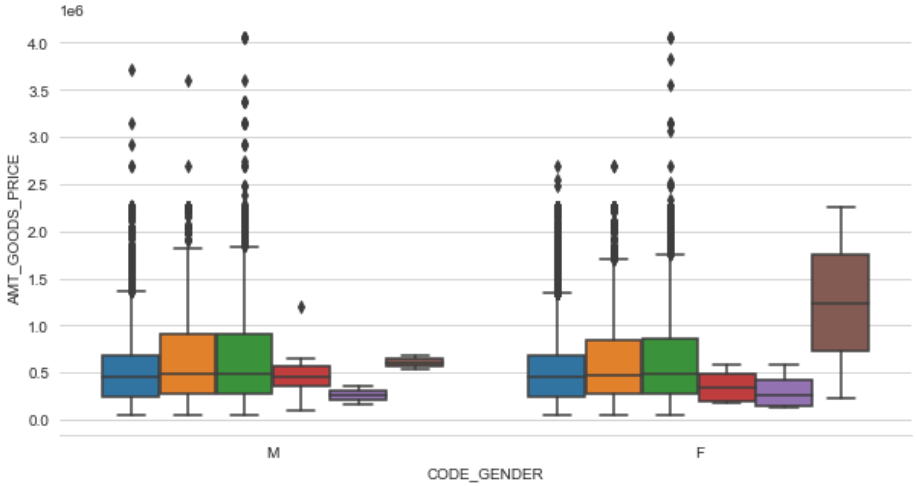
Inferences

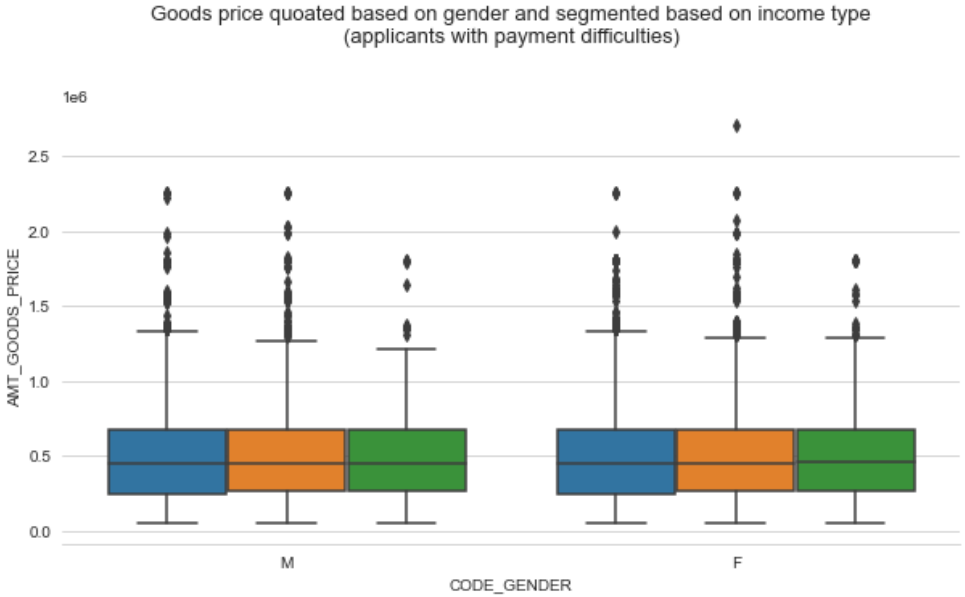
- Applicants from less populated areas seems to have more defaulters
- Higher the population density, lesser the defaulter applicants
- We have more applicants where the population density is in the range of 0.01 to 0.02

Clients with no Payment Difficulties Clients with payment Difficulties

Bivariate Analysis

Boxplot on goods price and gender segmented on income type

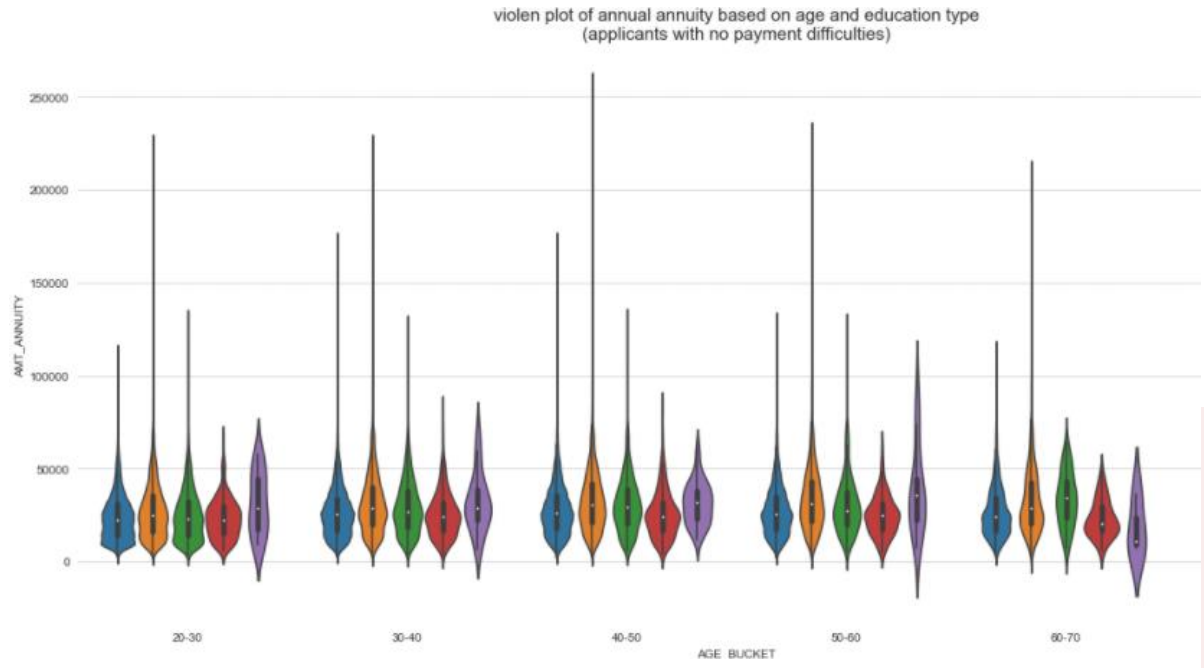


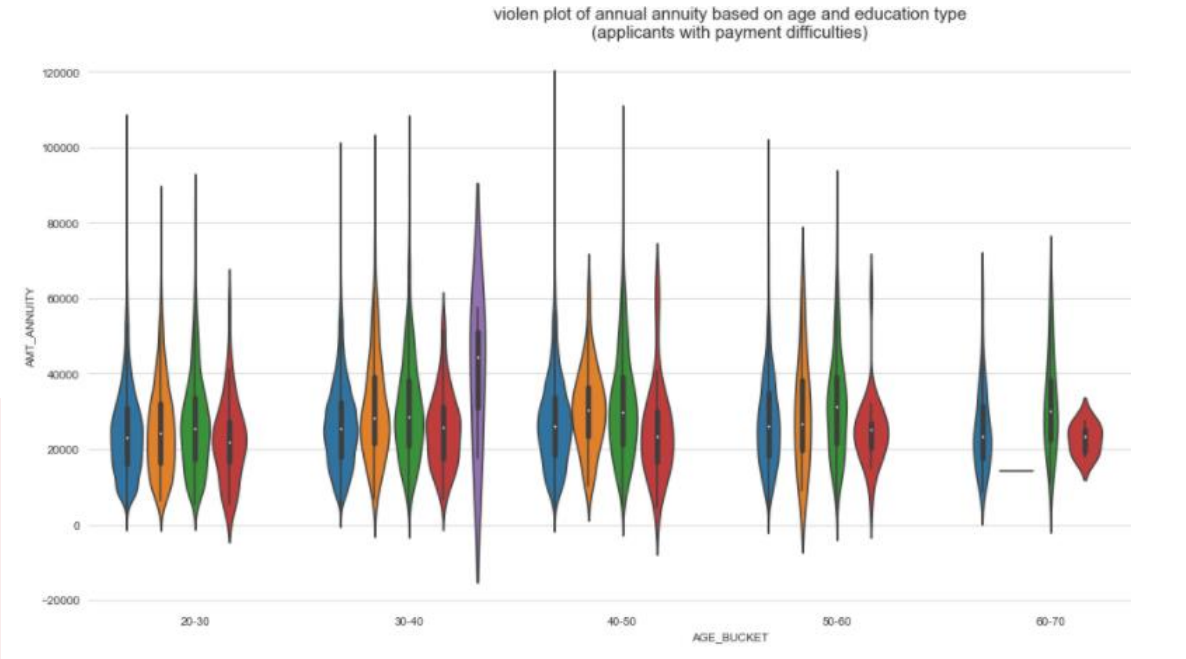


- Comparing both the plots, we can find no Business, Students or Unemployed applicants with payment difficulties
- Businesswomen have a very wide range in Goods Price compared to Businessmen
- Unemployed men tend have very less range in Goods price but quote higher compared to women, where they have slightly bigger spread but price lesser than men
- applicants with payments difficulties tends to have very equal range of goods price in all sectors of income
- Commercial Associate applicants tend to have higher goods price compared to all

Working
State Servant
Commercial associate
Student
Pensioner
Businessman

Violen plot on age and amount annuity segmented by education type

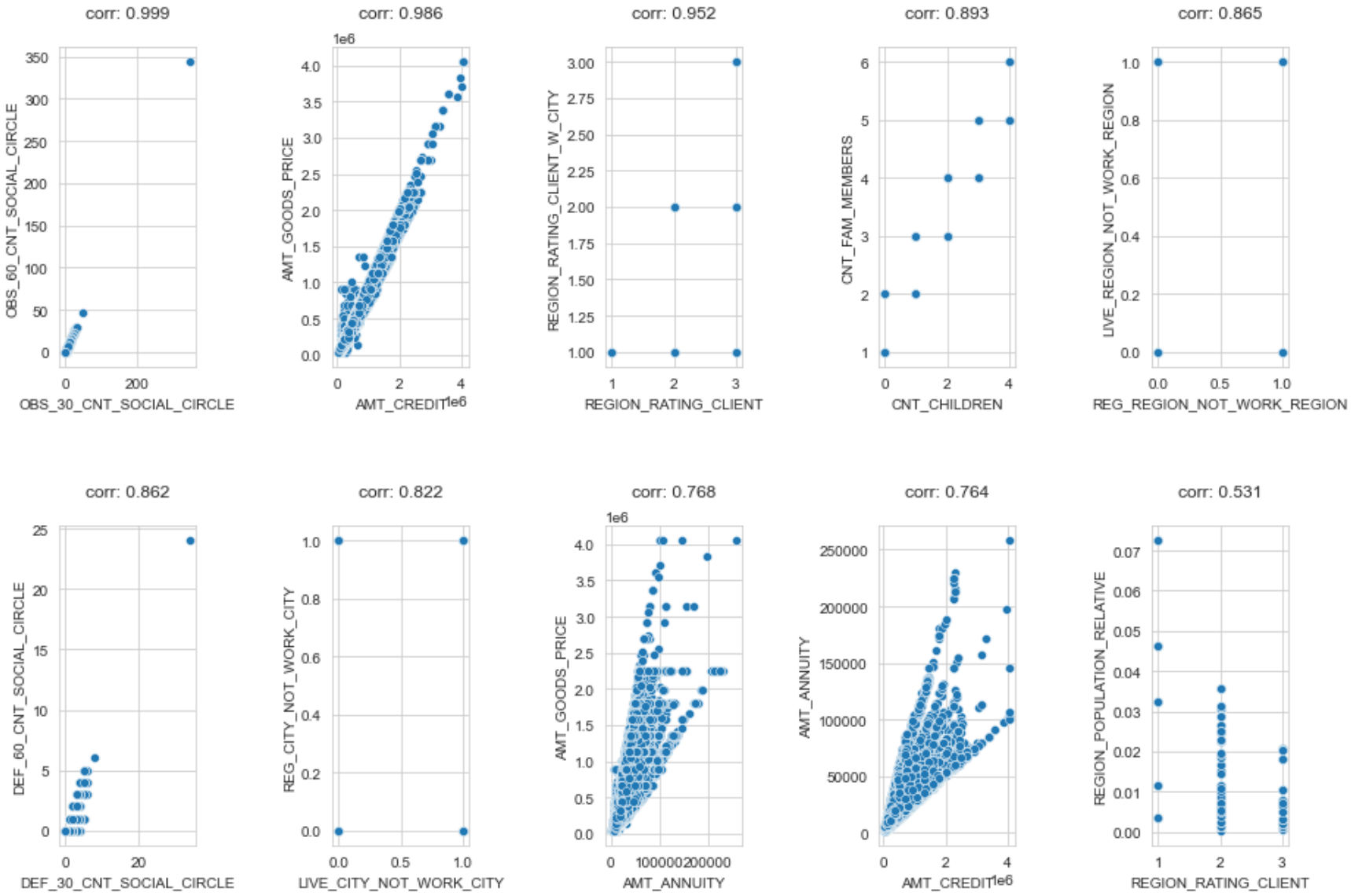




- applicants with Academic Degree seems to face very little issue with payment difficulties
- applicants within the age 60-70, who face difficulty in payments have very less spread of annual annuity compared to applicants who have no difficulty in payment applicants with higher education and with no payment difficulties pay more annual annuity
- There are extremely less applicants within age 60-70 having incomplete higher education face any payment difficulties compared to all age groups and even education type

Secondary/ secondary special
Higher education
Incomplete higher
Lower Secondary
Academic degree

Top 10 Corelated Variables in the Dataset



These are the top 10 corelated variables in the entire dataset. The graphs are ordered in descending order of correlation number.

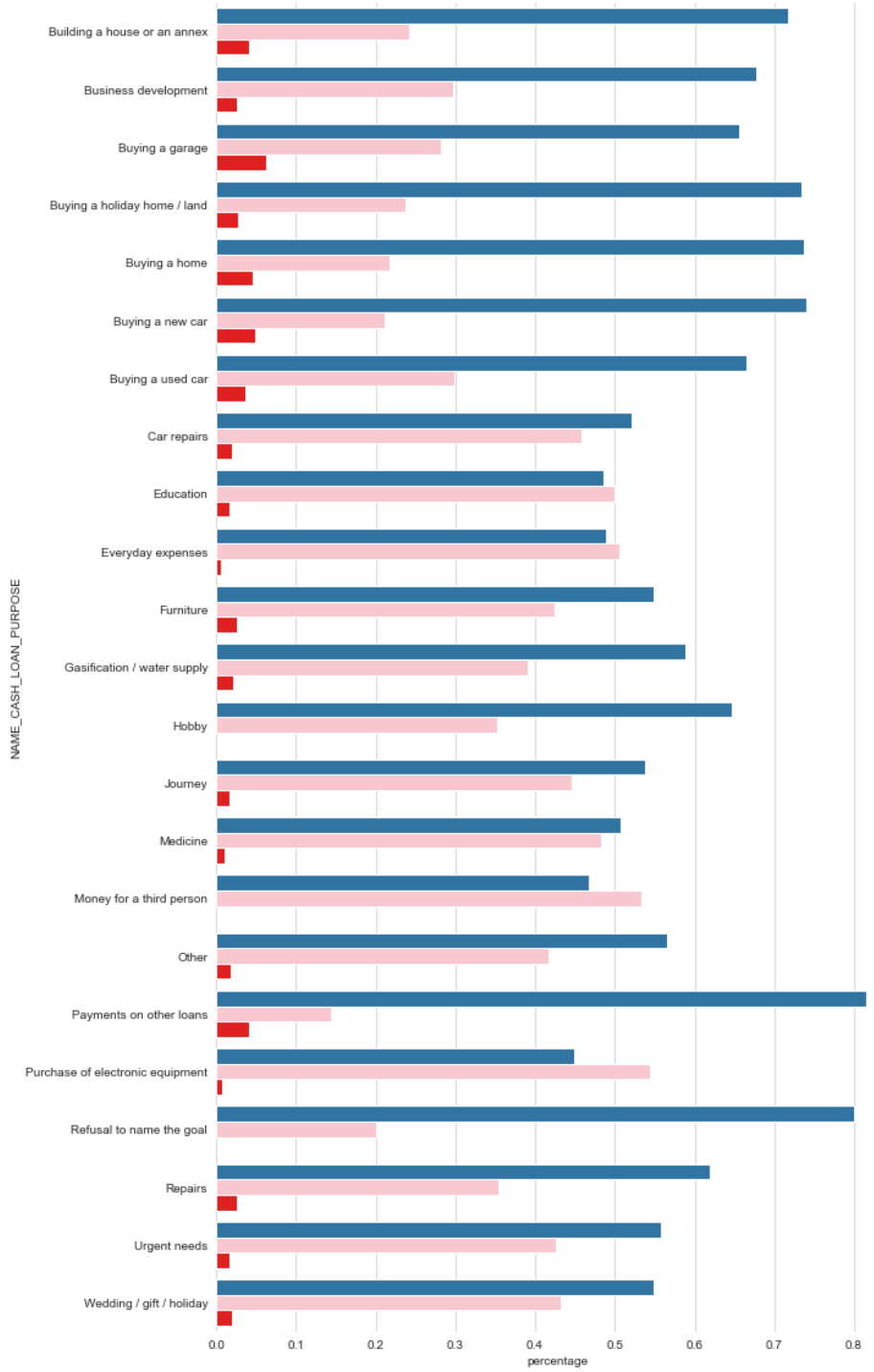
Each graph has its corelation number on top of their respective graph, and the axis label indicates the variable names

Analysis on merged Dataset

Countplot based on percentage of loan purpose segmented on previous application status

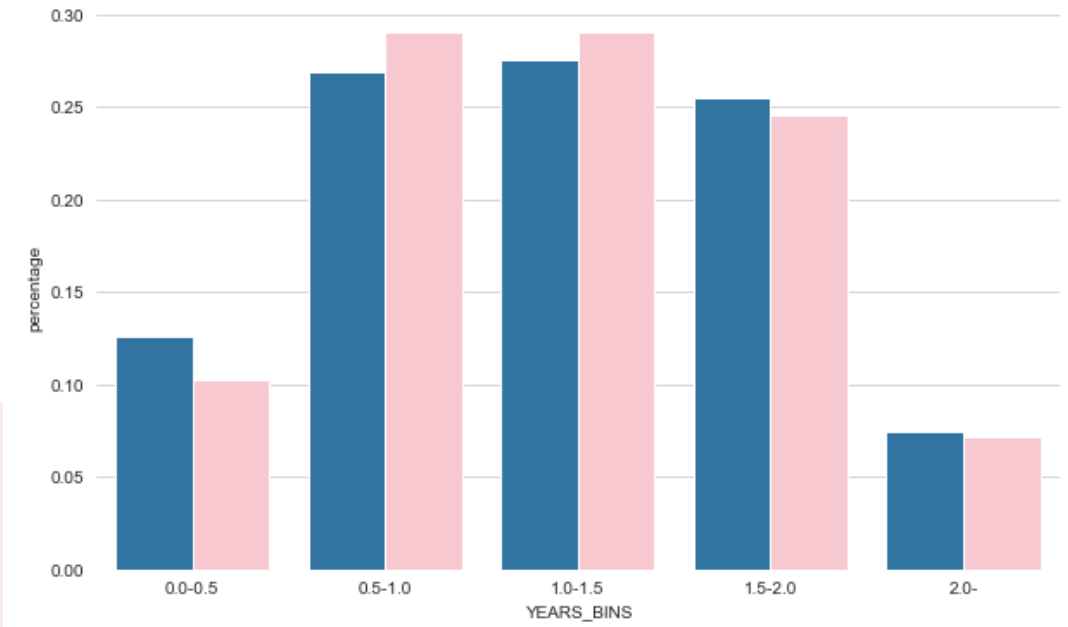
Inferences

- Cash Loan for Electronic Equipment has higher approval rate
- Loans for payments on other loans has the highest reject rate
- Loans for Hobbies, lending money for third person has least amount of rejection rates

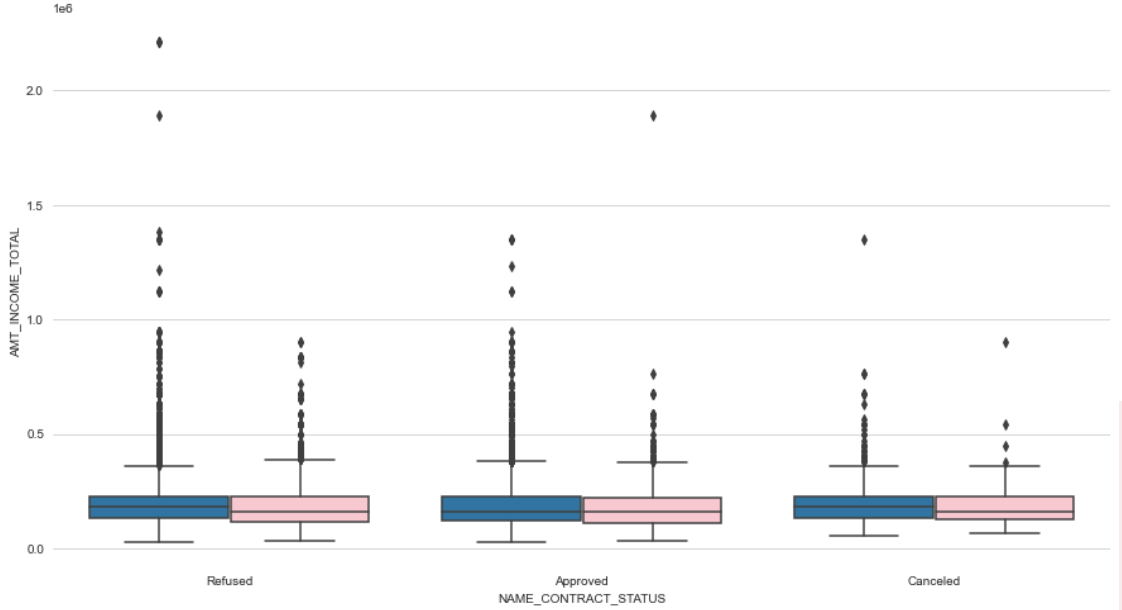


Refused Approved Cancelled

Count plot of difference in time between current and previous application



Boxplot of annual income based on previous application status segmented by target variable



Inferences

- Most of the defaulters who previously applied for a loan, apply now with nearly a gap of 6 months to 1.5 years
- very less people apply for loan with a gap of more than 2 years from the previous application

Inferences

- Outliers are identified
- Income of applicant's who have payment difficulties is lesser than the people who doesn't
- Higher the income of the applicant more the chances he/she refuses the loan offer

Clients with no Payment Difficulties

Clients with payment Difficulties

Final Insights

1. Focus more on people having academic degree
2. Focus more on businesswomen who quotes for higher goods price more than 1 million
3. Applicants with secondary education who are ready to pay annuity of more than 150 thousand are less likely to default
4. Applicants from densely populated areas are less likely to default compared to poorly populated areas
5. Applicants who are working in their current employment for more than 6 years are very less likely to default
6. Women who are into business generally would like to spend more on their goods compared to men

Note: All the code required for Data cleaning and development of visualizations are provided in a well commented python notebook that was submitted along with the presentation

Thank You

- Srihari R (sriharikrishna06@gmail.com)