PRESENTATION

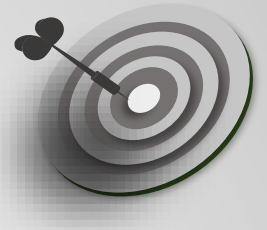
Exploratory Data Analysis

Goka Keerthana
Upgrad Learner

CONTENTS

- PROBLEM STATEMENT
- UNDERSTANDING DATASET
- ASSUMPTIONS
- DATA ANALYSIS APPROACH
- INFERENCES
- CONCLUSION

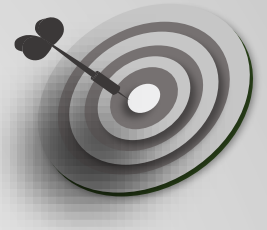
PROBLEM STATEMENT



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.

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.

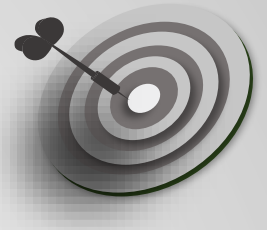
UNDERSTANDING DATASET



The given dataset of a loan providing company which has 3 files as explained below:

1. **'application_data.csv'** contains all the information of the client at the time of application.
The data is about whether a client has payment difficulties.
2. **'previous_application.csv'** contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.
3. **'columns_description.csv'** is data dictionary which describes the meaning of the variables.
ASSUMPTIONS

ASSUMPTIONS

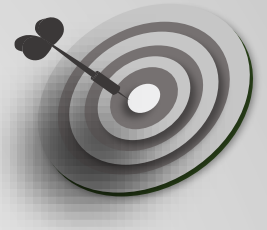


- The value ‘365243’ is assumed as unemployed in the Occupation Type attribute of application dataset
- The values ‘XNA’ & ‘XAP’ are assumed to be as null values in the application dataset and previous dataset
- In previous dataset, Higher number of Region rating is considered higher rating

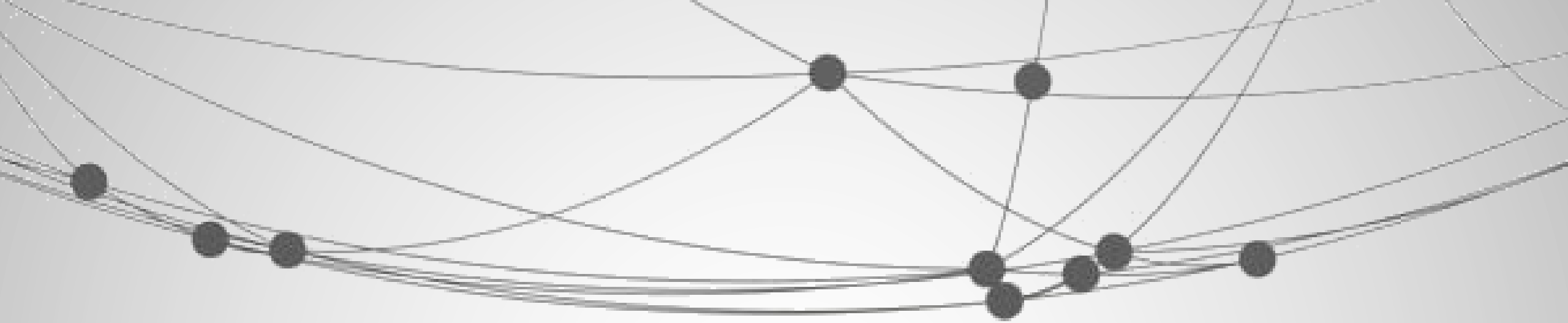
DATA IMBALANCE

Ratio Of Data Imbalance in the target variable is 11.38

APPROACH



- **Data Loading:** Read The Given Data Sets Application_data And Previous_application In Jupyter Application.
- **Data Sanity Checks**
- **Identification Of Missing Values & Their Handling:** Drop Columns That Have More Than 35% Of Missing Data.
- **Selection Of 25 Attributes** Relavent To The Problem Statement In Application Data And Group them into 5 Each
- **Performing Univariant, Bivariant, Segmented Univariant Analysis** With These Attributes
- **Jotting Down Inferences** Based On The Analysis Plots



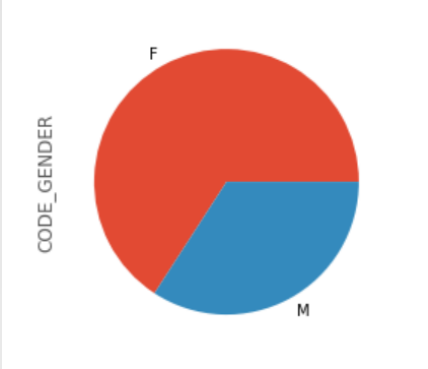
INFERENCES

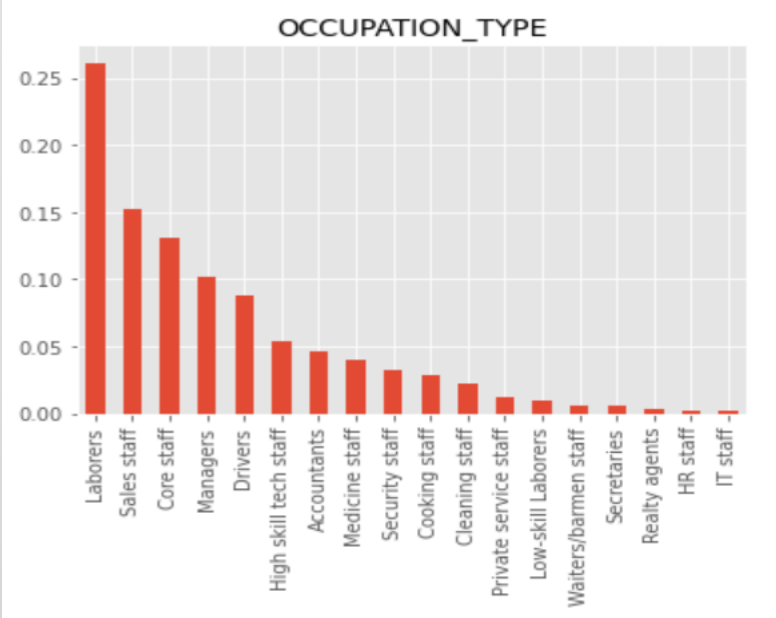
APPLICATION DATA -SET 1

UNIVARIANT ANALYSIS

SET 1

PLOTS





INFERENCES

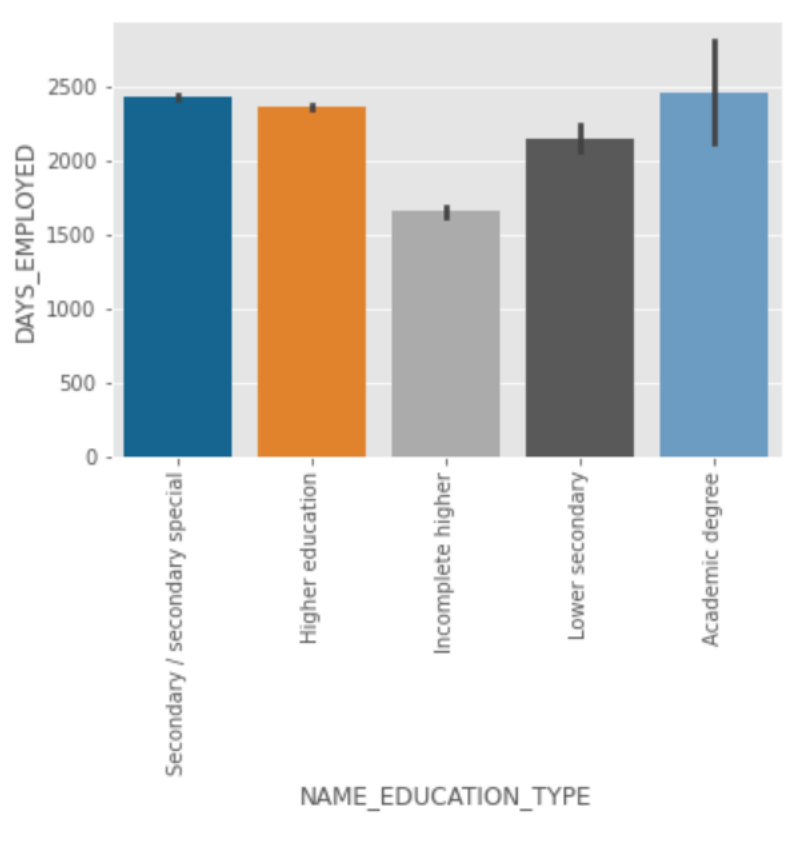
\+ Number of Female consumers are more than the Male consumers

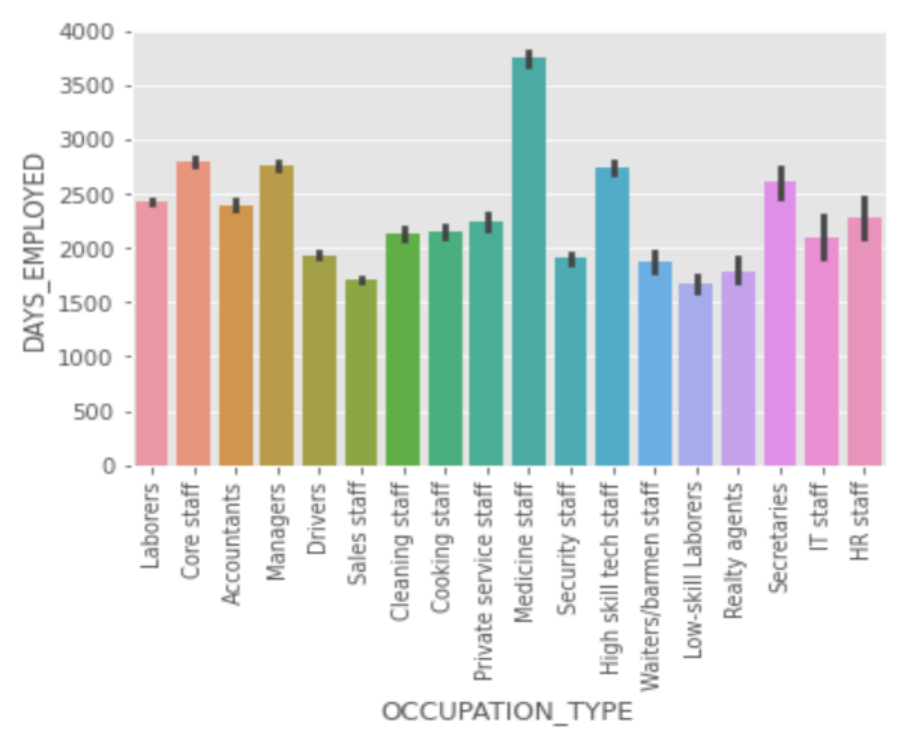
\- Around 25% of the values of Occupation type are Laborers, followed by Sales staff with around 15%.

BIVARIANT ANALYSIS

SET 1

PLOTS





INFERENCES

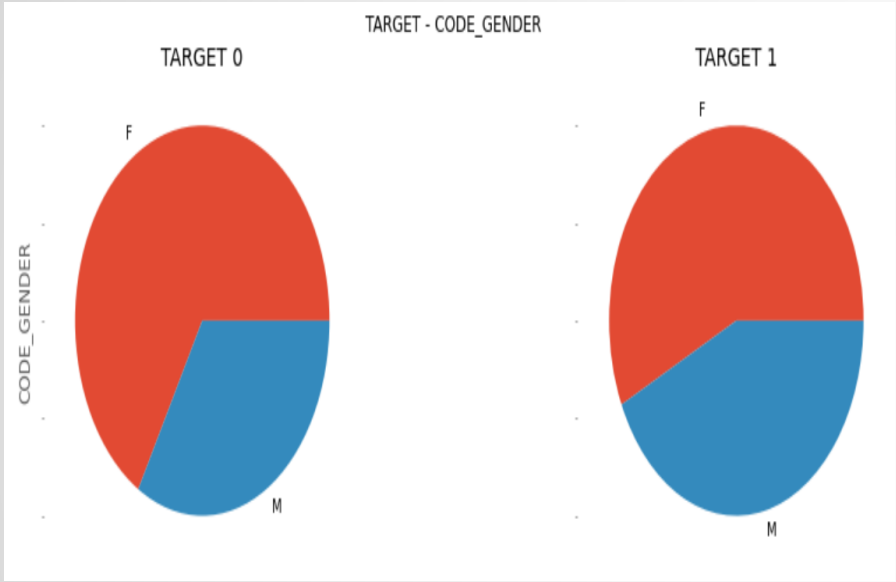
\+ Mean of no of employment days is almost same for all consumers except for Incomplete higher education type and lower secondary education type.

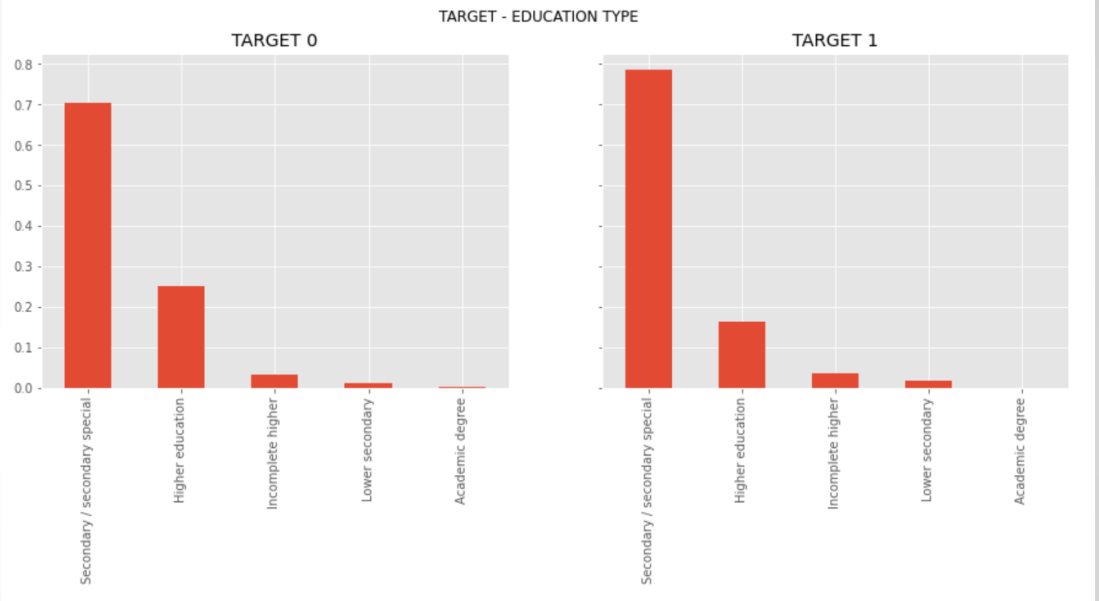
\- Medicine Staff has the highest mean days of employment even though the mean age of them is not the highest

SEGMENTED UNIVARIANT ANALYSIS

SET 1

PLOTS





INFERENCES

\+ Percentage of male consumers is higher in defaulters of loans than in others.

\- Percentage of Secondary/Secondary special Education type is slightly more in defaulters when compared to consumers in others

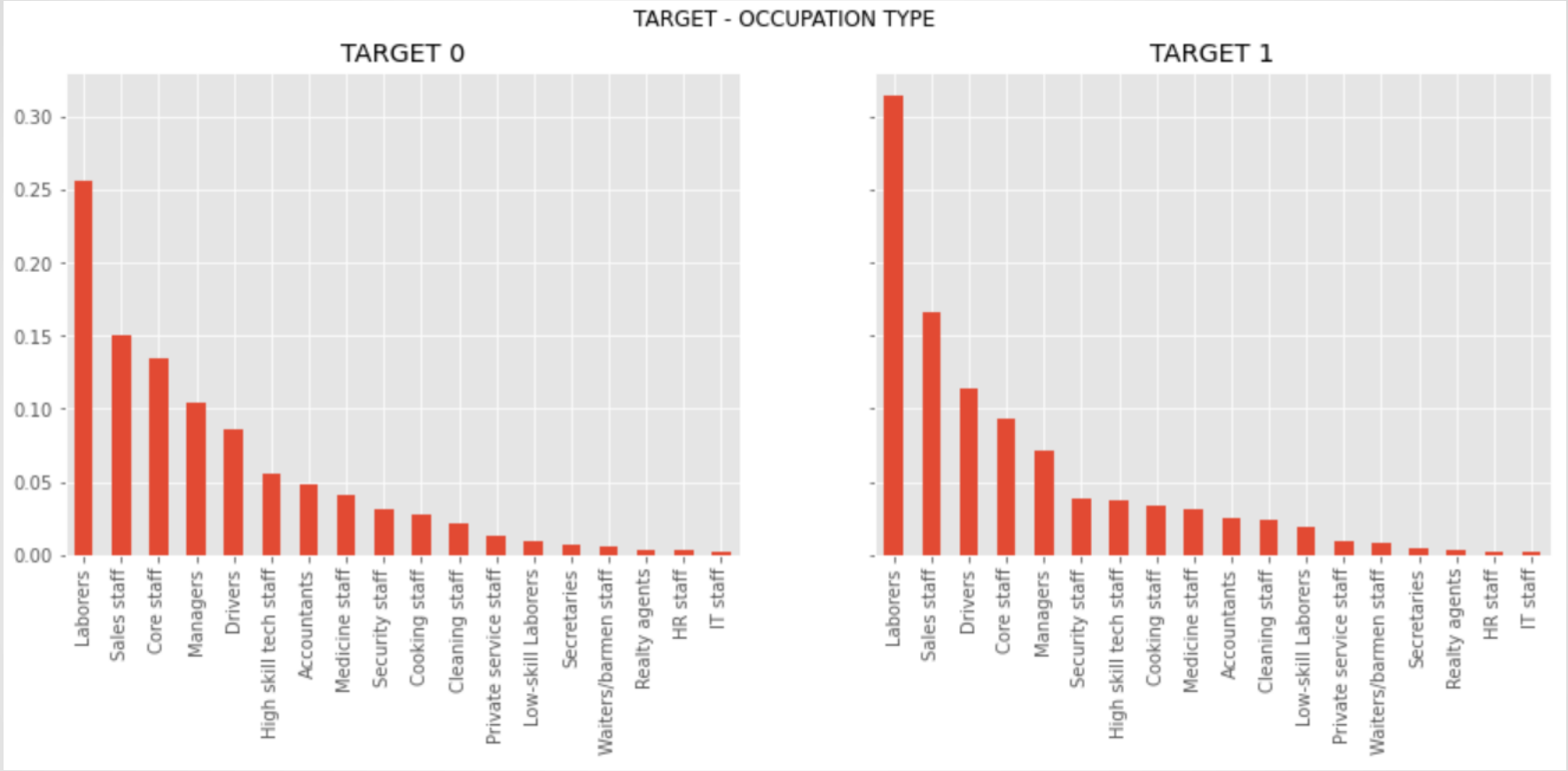
Percentage of Higher Education type is slightly less in defaulters when compared to consumers in others.

Hence Higher education type are to be given more preference

SEGMENTED UNIVARIANT ANALYSIS

SET 1

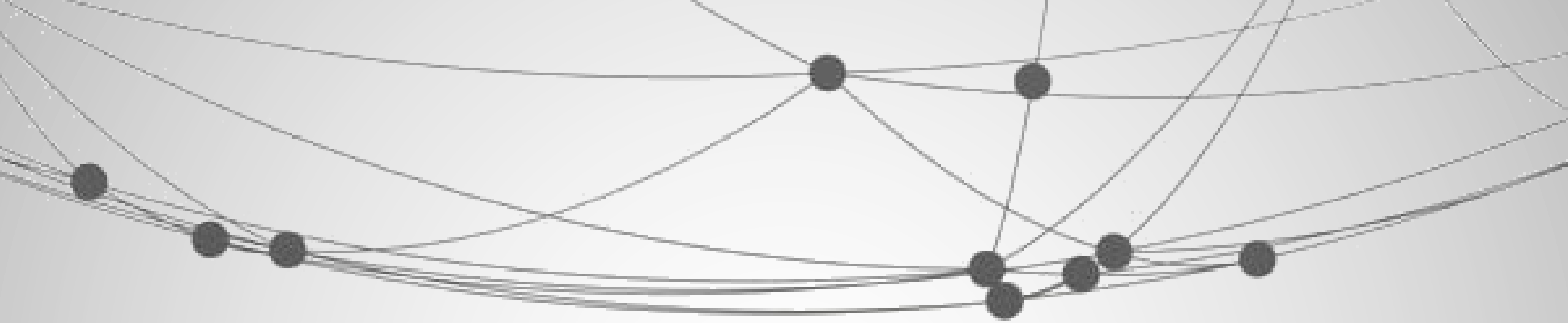
PLOTS



INFERENCES

+ Percentage of Laborers, Drivers, Sales Staff in Occupation type is slightly more in defaulters when compared to consumers in others. Hence these categories may be less preferred

Percentage of Core staff, Managers in Occupation type is slightly less in defaulters when compared to consumers in others. Hence these categories may be more preferred



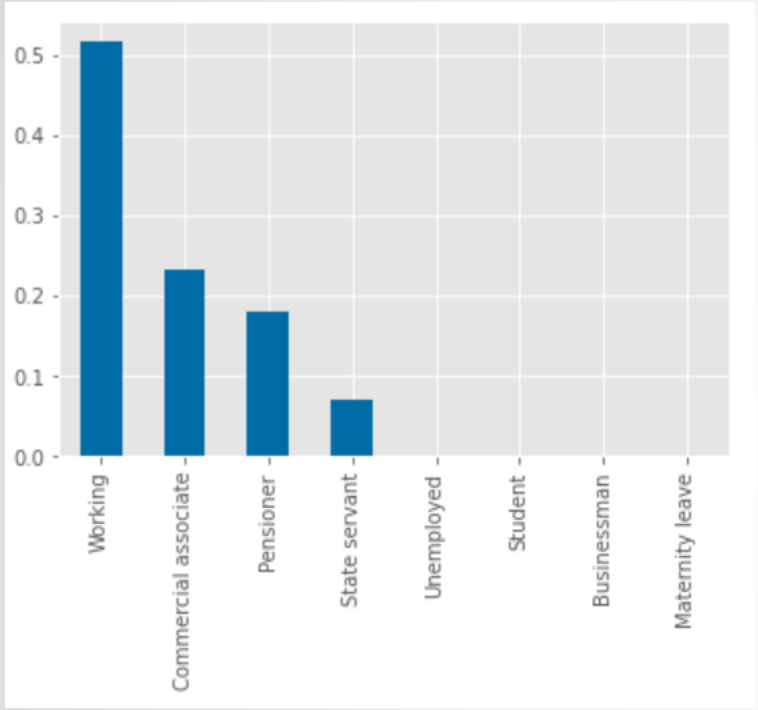
INFERENCES

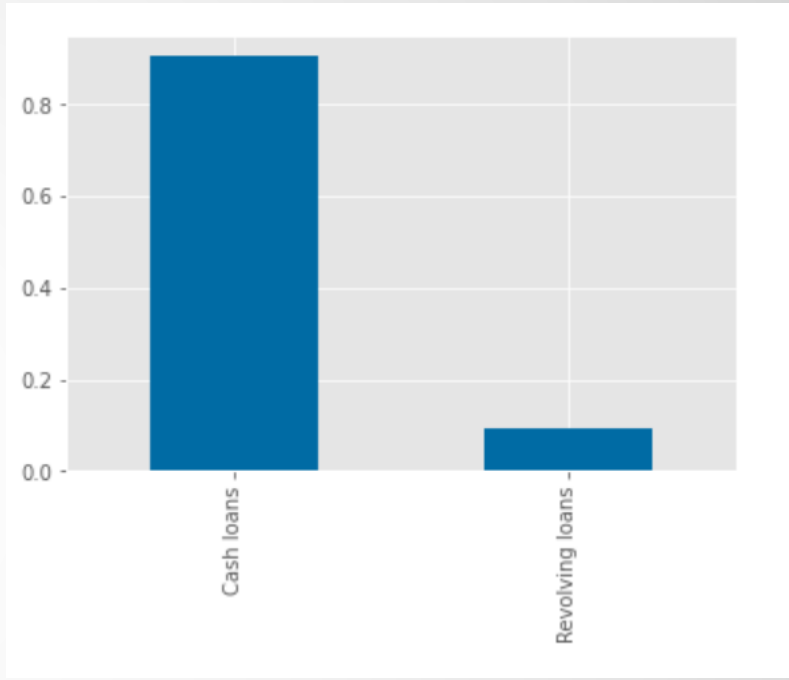
APPLICATION DATA -SET 2

UNIVARIANT ANALYSIS

SET 2

PLOTS





INFERENCES

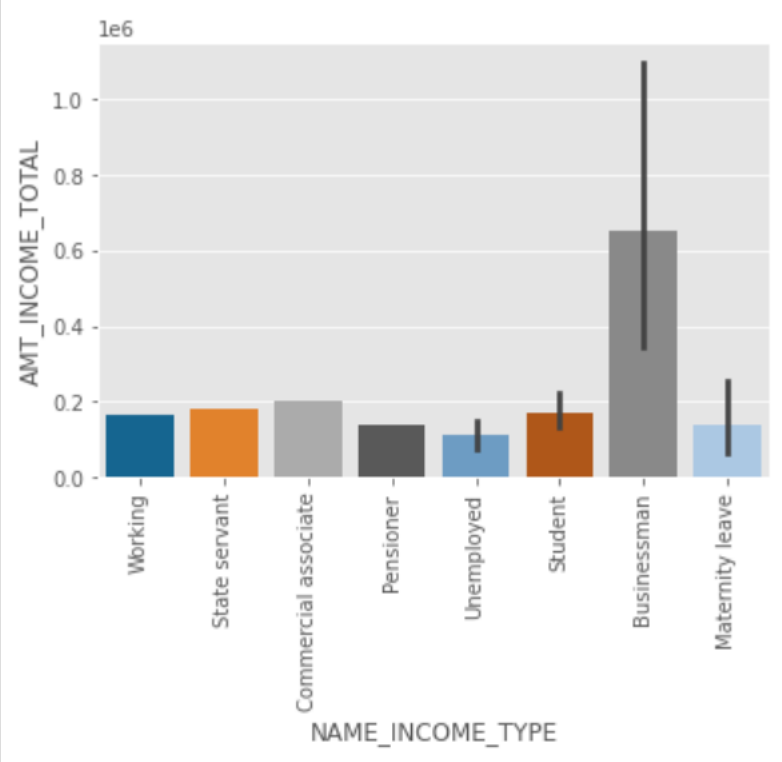
+ 50% of the income type are working

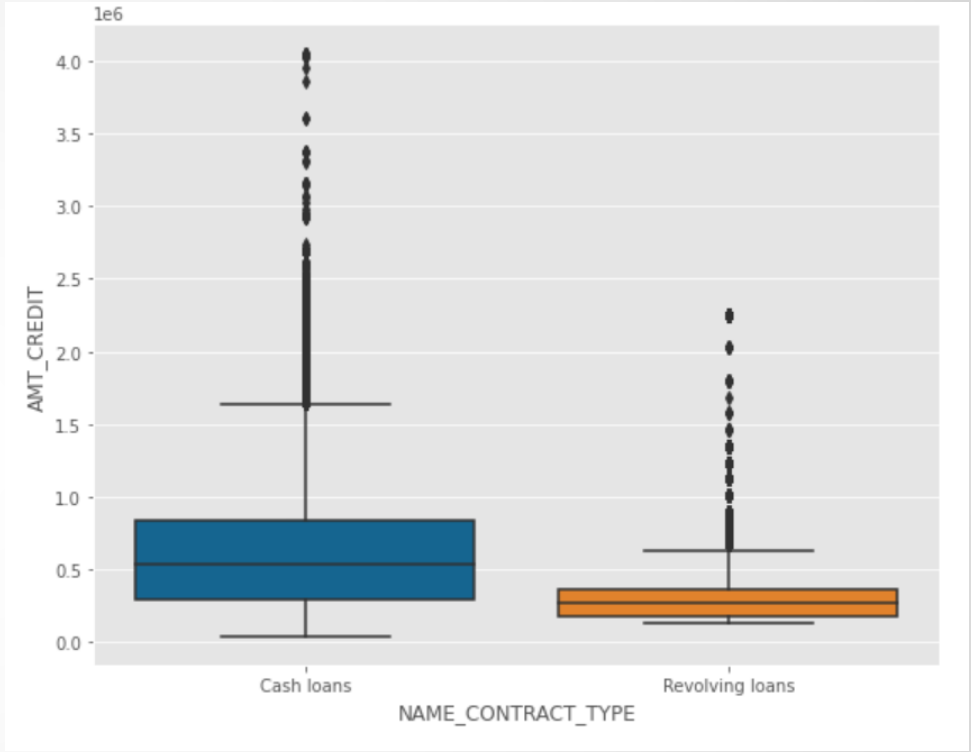
- 80% of the loans are Cash loans

BIVARIANT ANALYSIS

SET 2

PLOTS





INFERENCES

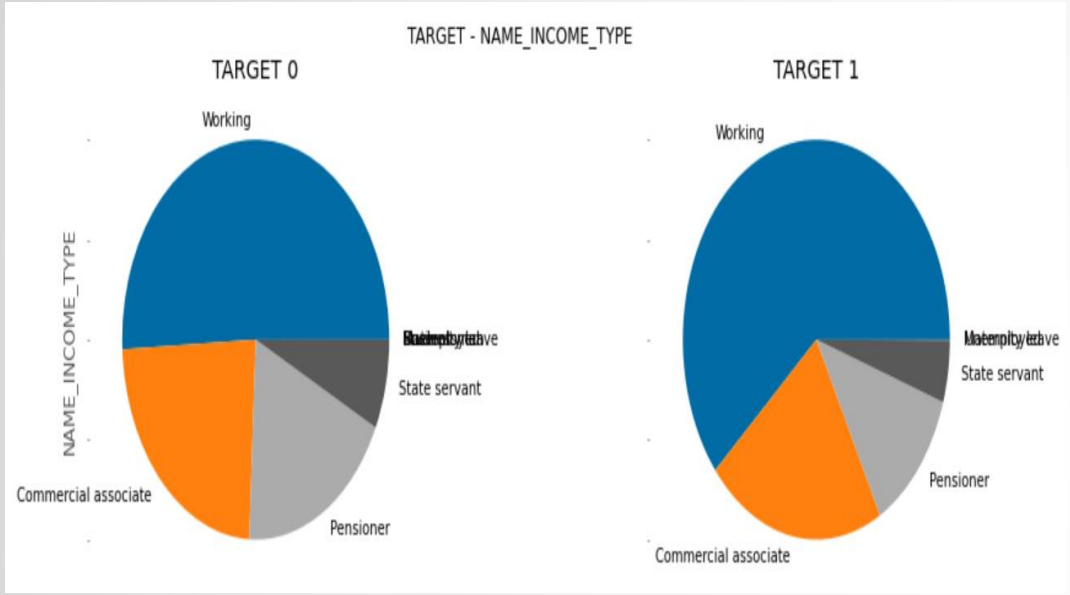
+ The Total Income for the Businessman Income type is almost 3 times higher than any othe Income type.

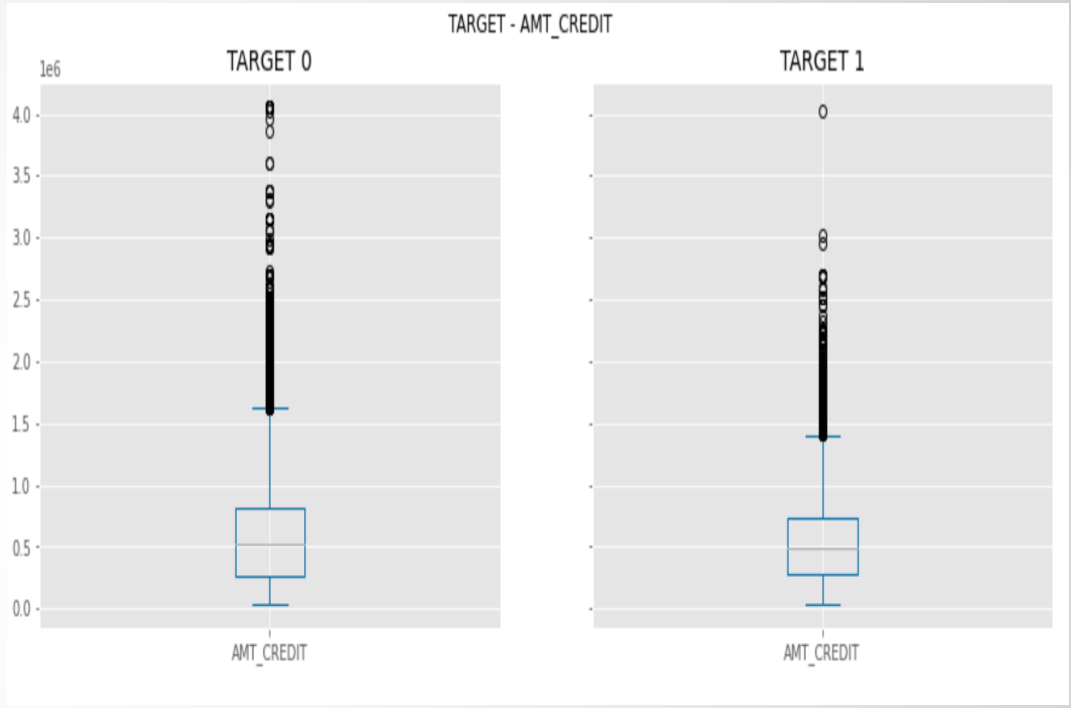
- Median of Credit amount is higher in Cash loans than in Revolving Loans

SEGMENTED UNIVARIANT ANALYSIS

SET 2

PLOTS

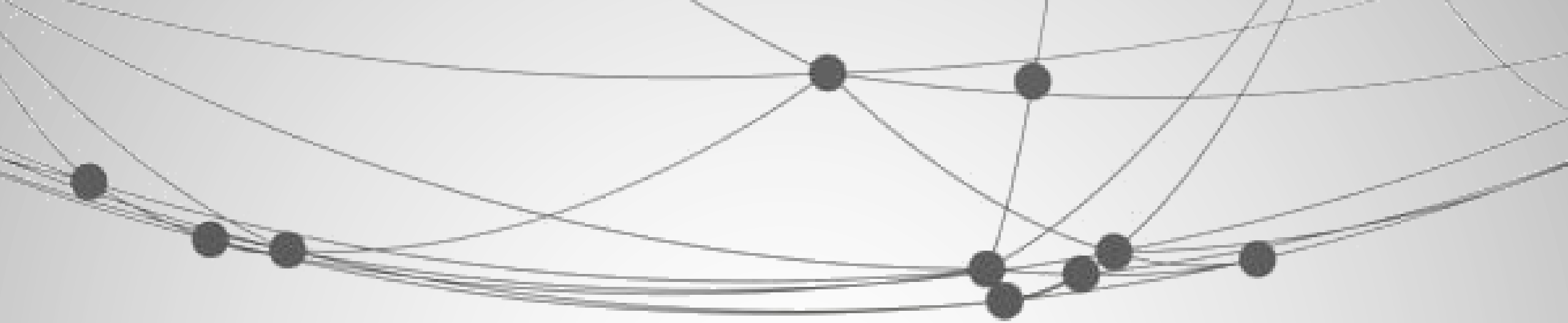




INFERENCES

\+ Percentage of Working Income type are higher in defaulters than in others.Hence less preferable
Percentage of Pensioner Income type are lesser in defaulters than in others. Hence more preferable

\- Lower no of outliers in Total credit amount of defaulters category than in others category. Interestingly Very High credit amounts have lower defaulters.



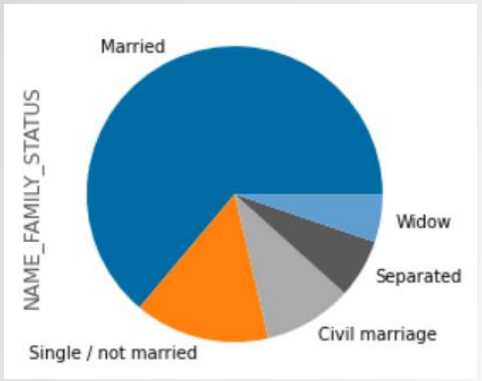
INFERENCES

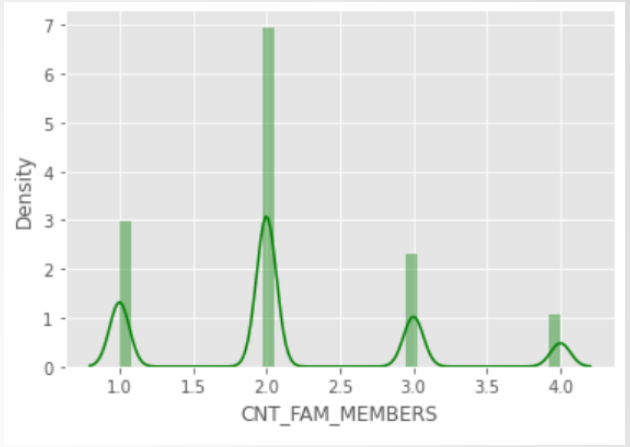
APPLICATION DATA -SET 3

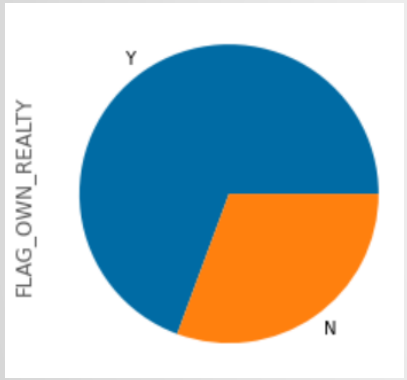
UNIVARIANT ANALYSIS

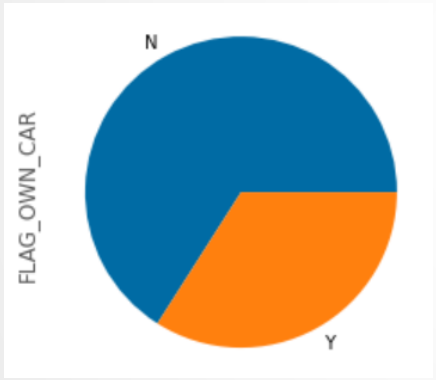
SET 3

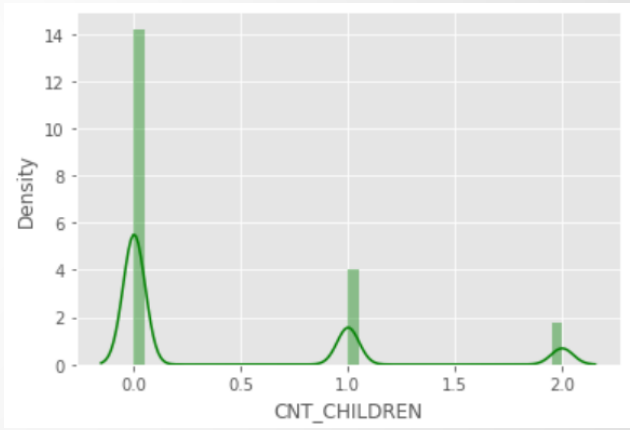
PLOTS











INFERENCES

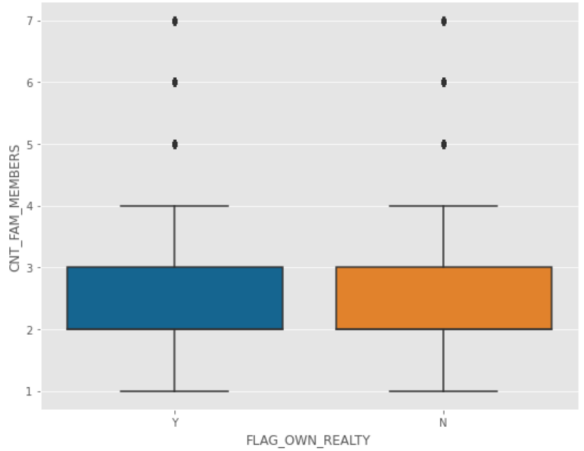
\+ #Approx 65% are married in the data set
#Approx 35% own a car in the data set
#Approx 70% Own a Realty in the data set

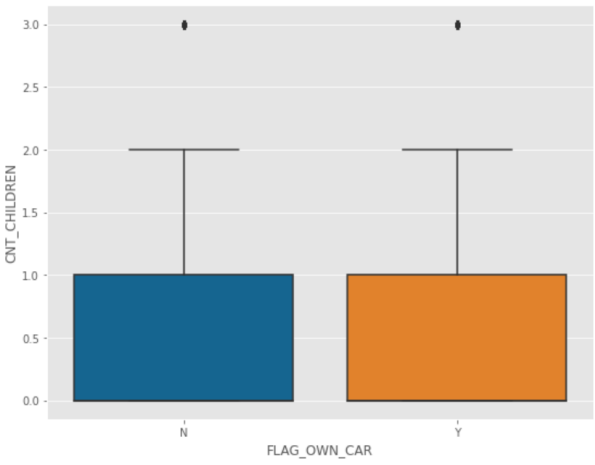
\- 50% of consumers have 2 family members and 50% of consumers do not have children

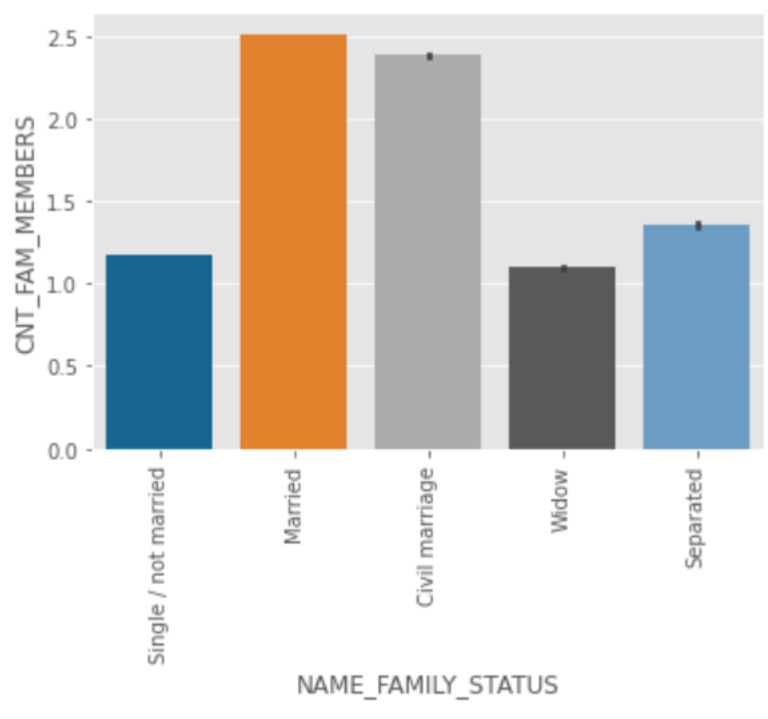
BIVARIANT ANALYSIS

SET 3

PLOTS







INFERENCES

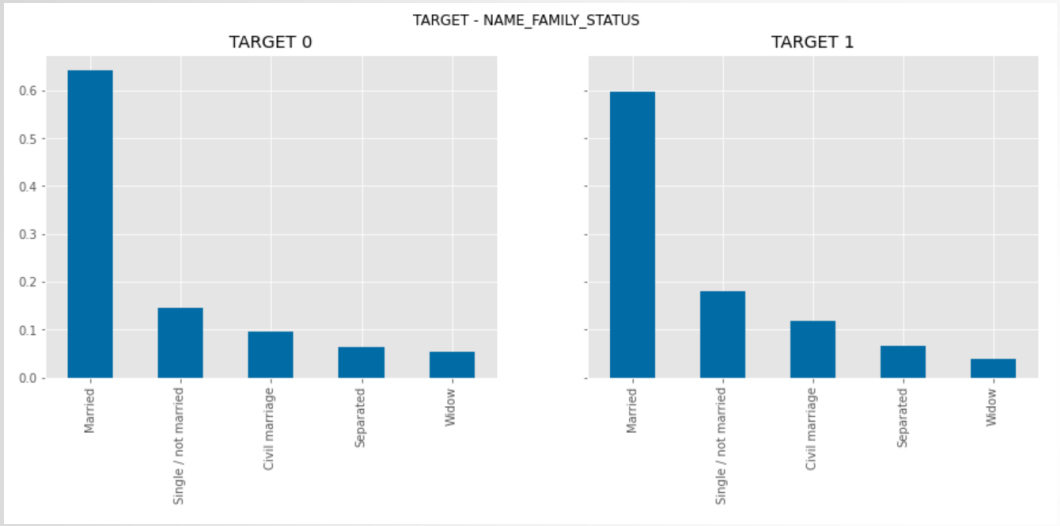
\+ Irrespective of Owning a realty, count of family members is the same
Irrespective of Owning a car, count of children is the same

\- Married and living together couple have 2 or more family members.

SEGMENTED UNIVARIANT ANALYSIS

SET 3

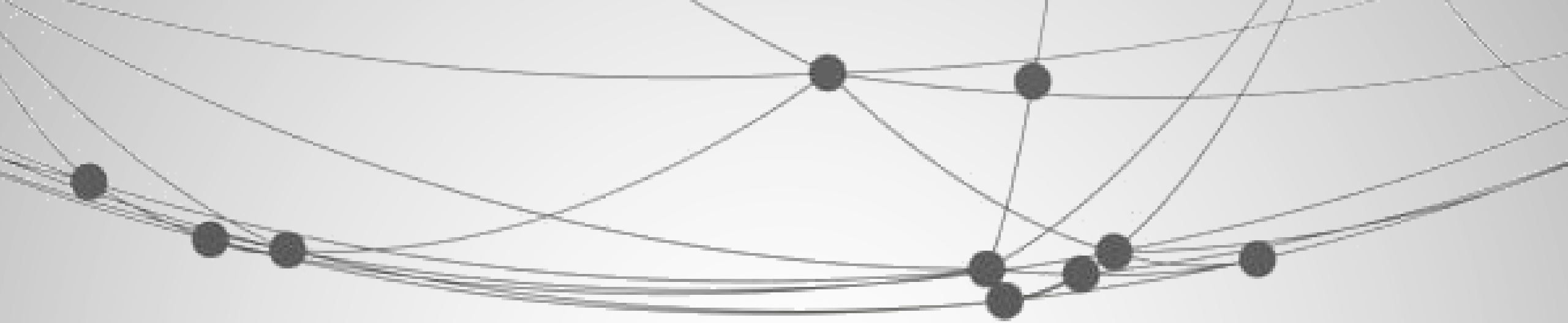
PLOTS



INFERENCES

\+ Proportions of Single/not married & Civil marriage consumers are higher in defaulters than in others. Hence these category consumers may be less preferred for loans

\- Clearly families with higher members are in the defaulter category. Hence 5+ family members may not to be preferred for loans.



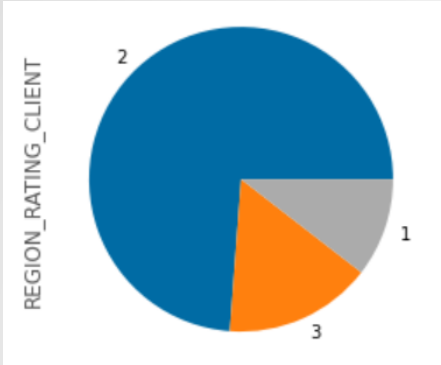
INFERENCES

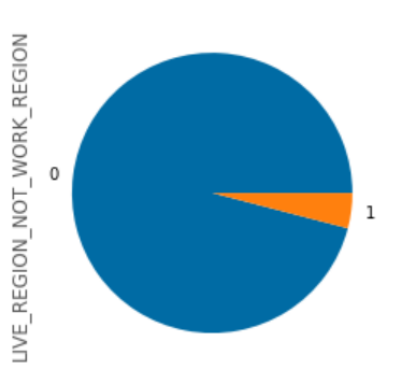
APPLICATION DATA -SET 4

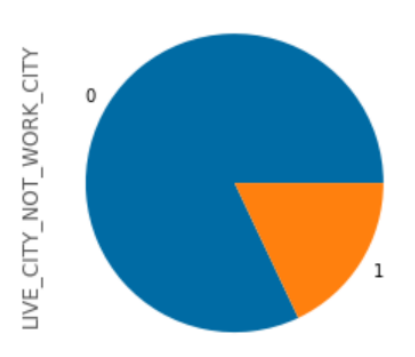
UNIVARIANT ANALYSIS

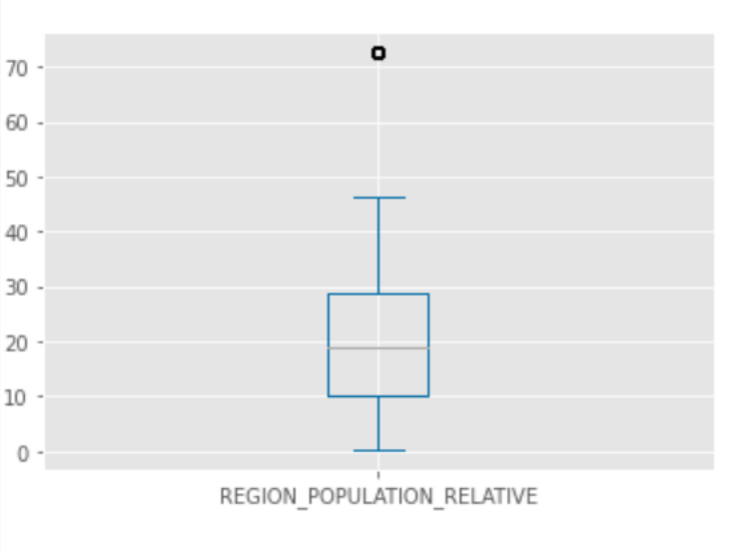
SET 4

PLOTS









INFERENCES

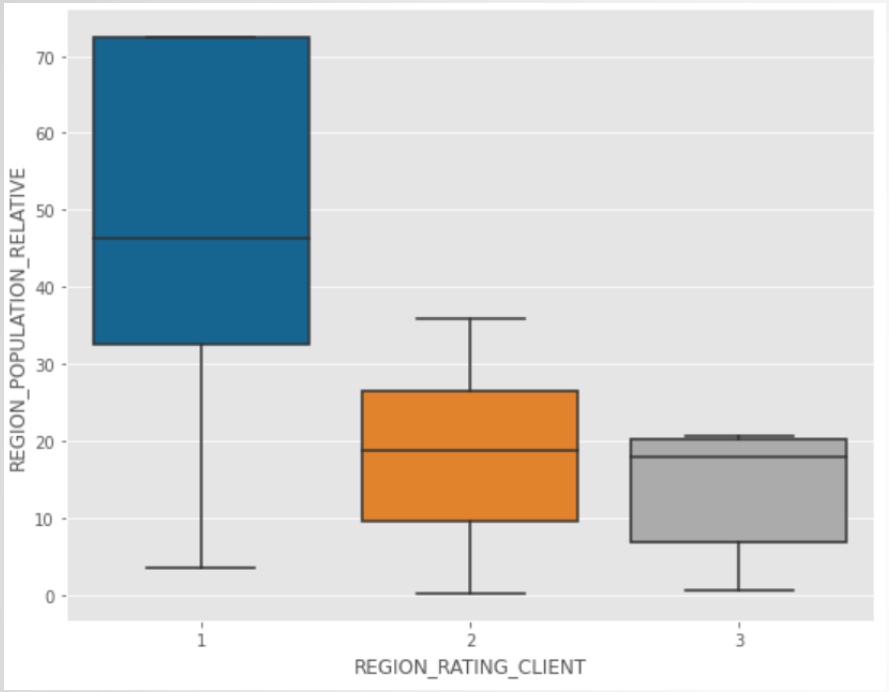
\+ 20% of the consumers do not live in the city they work
Less than 5% of the consumers do not live in the region they work
70% of consumers live in Region with region Rating 2

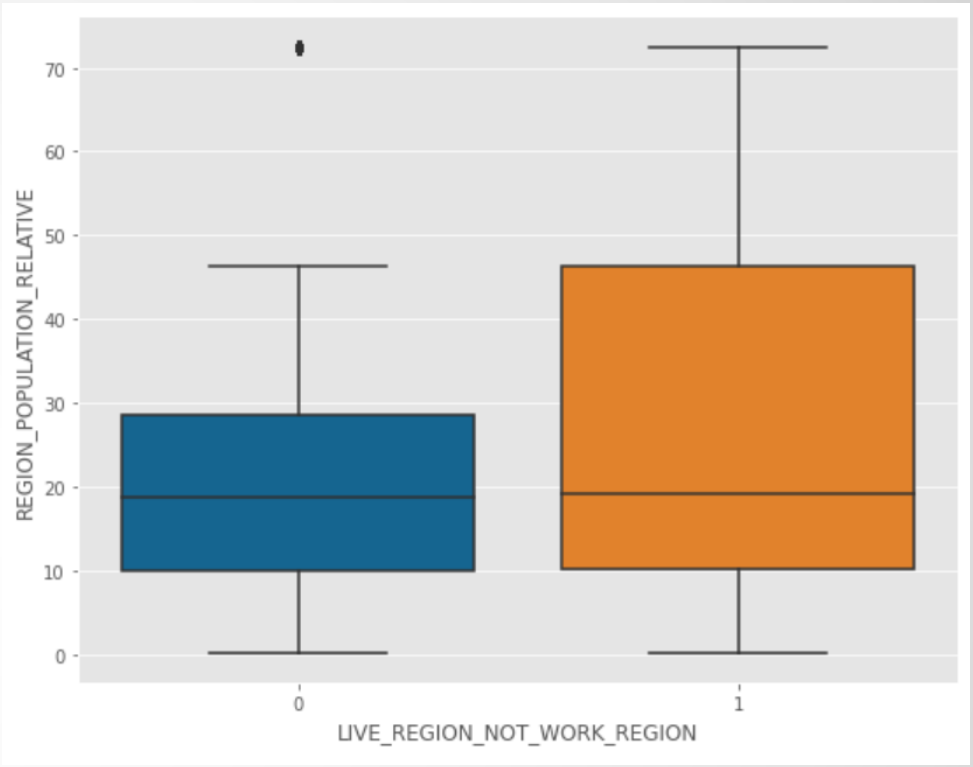
\- Outlier present with region population relative around 70. This might be values of specific regions with higher region population relative like metros.

BIVARIANT ANALYSIS

SET 4

PLOTS





INFERENCES

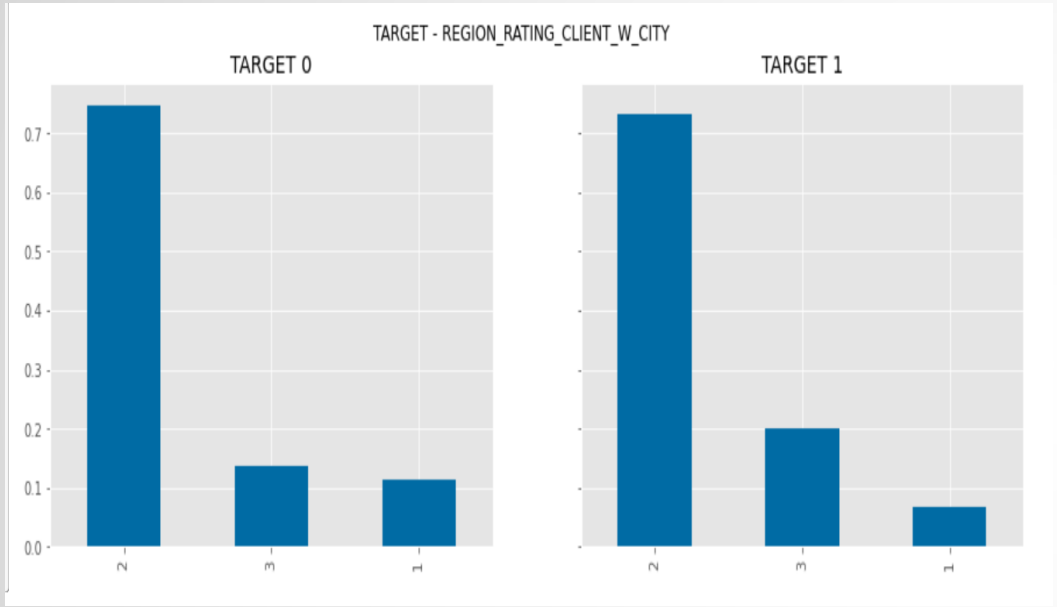
\+ Higher the region rating, relatively lesser the population of the region.

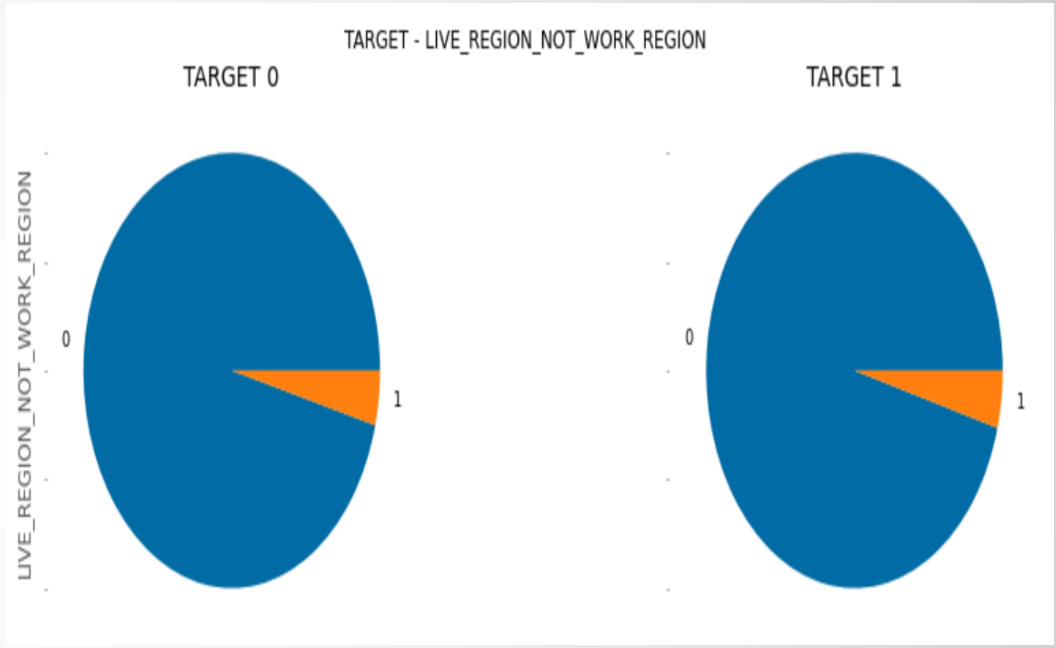
\- Population of the city is relatively higher if the living region is not the work region. Its reasonable as some people commute to highly populated regions but do not stay due to high living costs

SEGMENTED UNIVARIANT ANALYSIS

SET 4

PLOTS

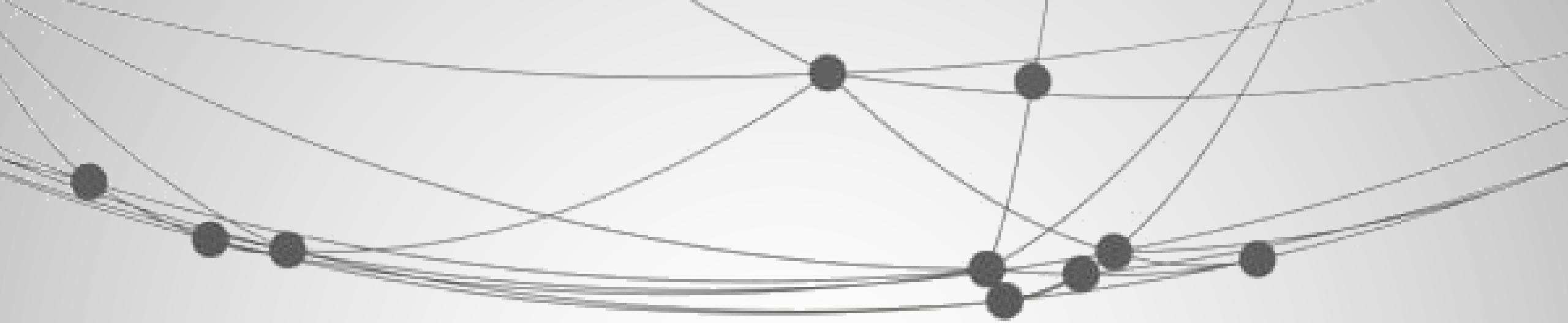




INFERENCES

\+ Slightly higher chance of defaulting by the consumers who do not live in the region with rating 3 w.r.t. city.

\- chance of defaulting the loan amounts is independent whether the consumers lives in the work region or not.



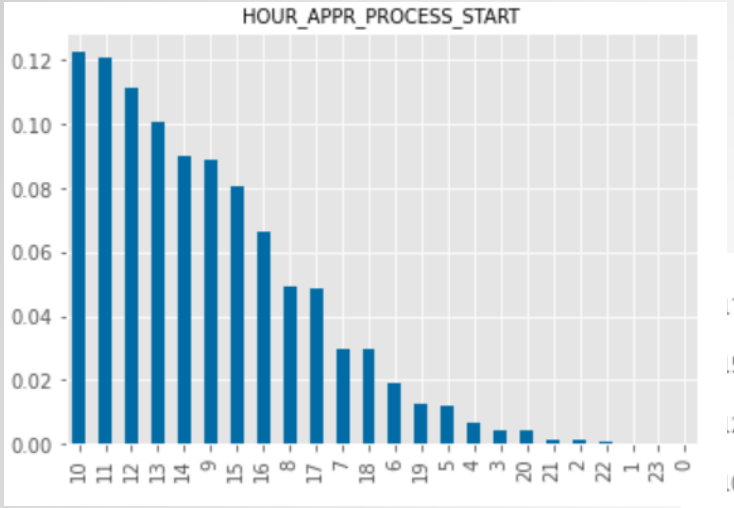
INFERENCES

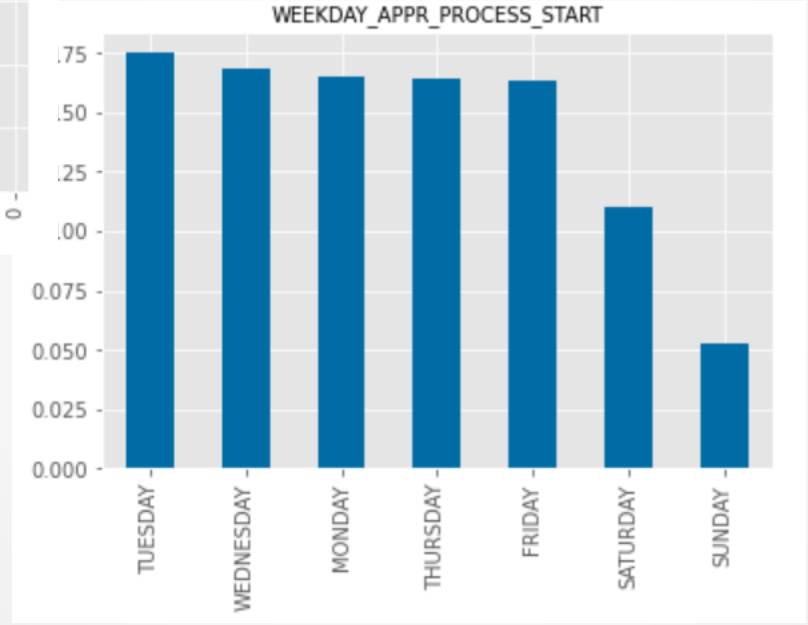
APPLICATION DATA -SET 5

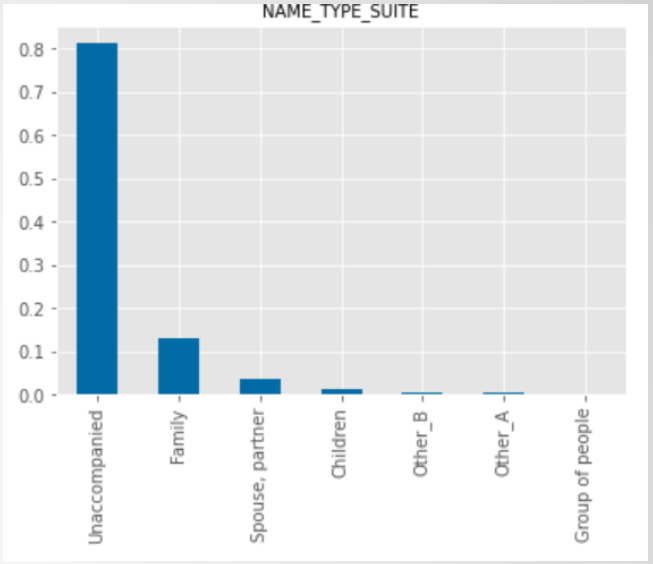
UNIVARIANT ANALYSIS

SET 5

PLOTS







INFERENCES

+

#Saturday & Sunday have the least no of applications started

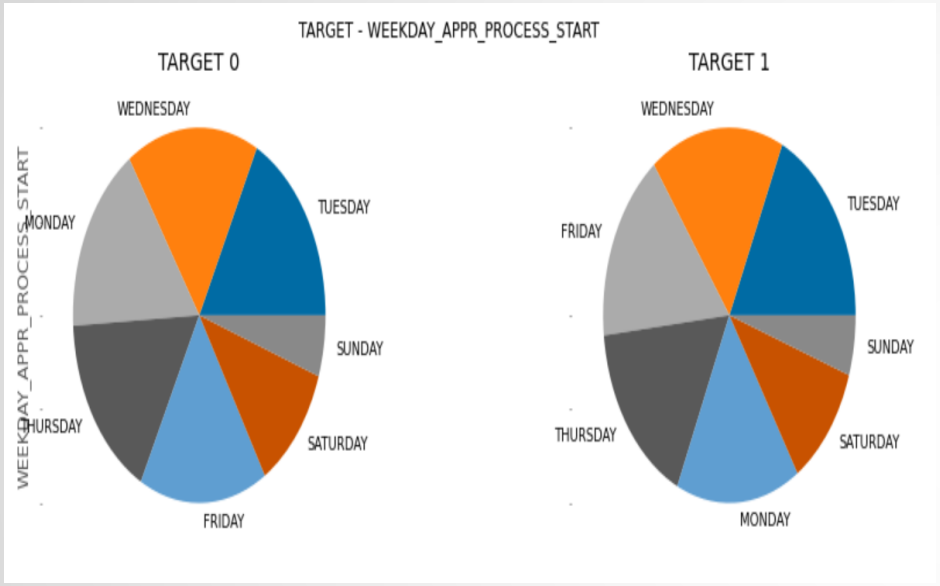
#8am to 5pm have most of the applications started

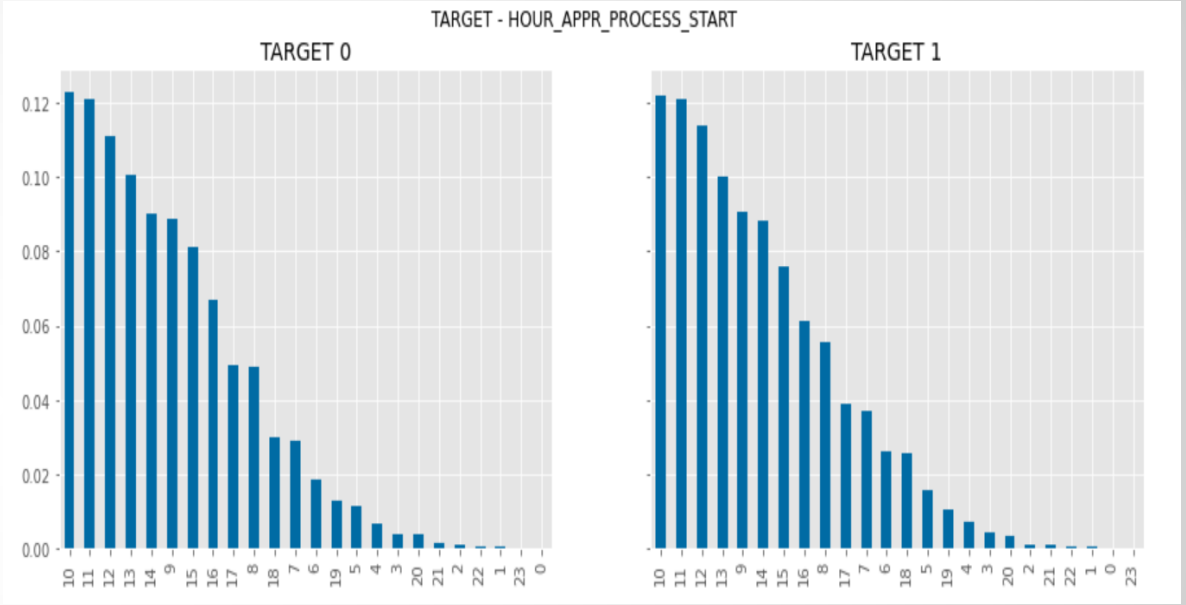
#80% consumers are unaccompanied for start of application processing.

SEGMENTED UNIVARIANT ANALYSIS

SET 5

PLOTS





INFERENCES

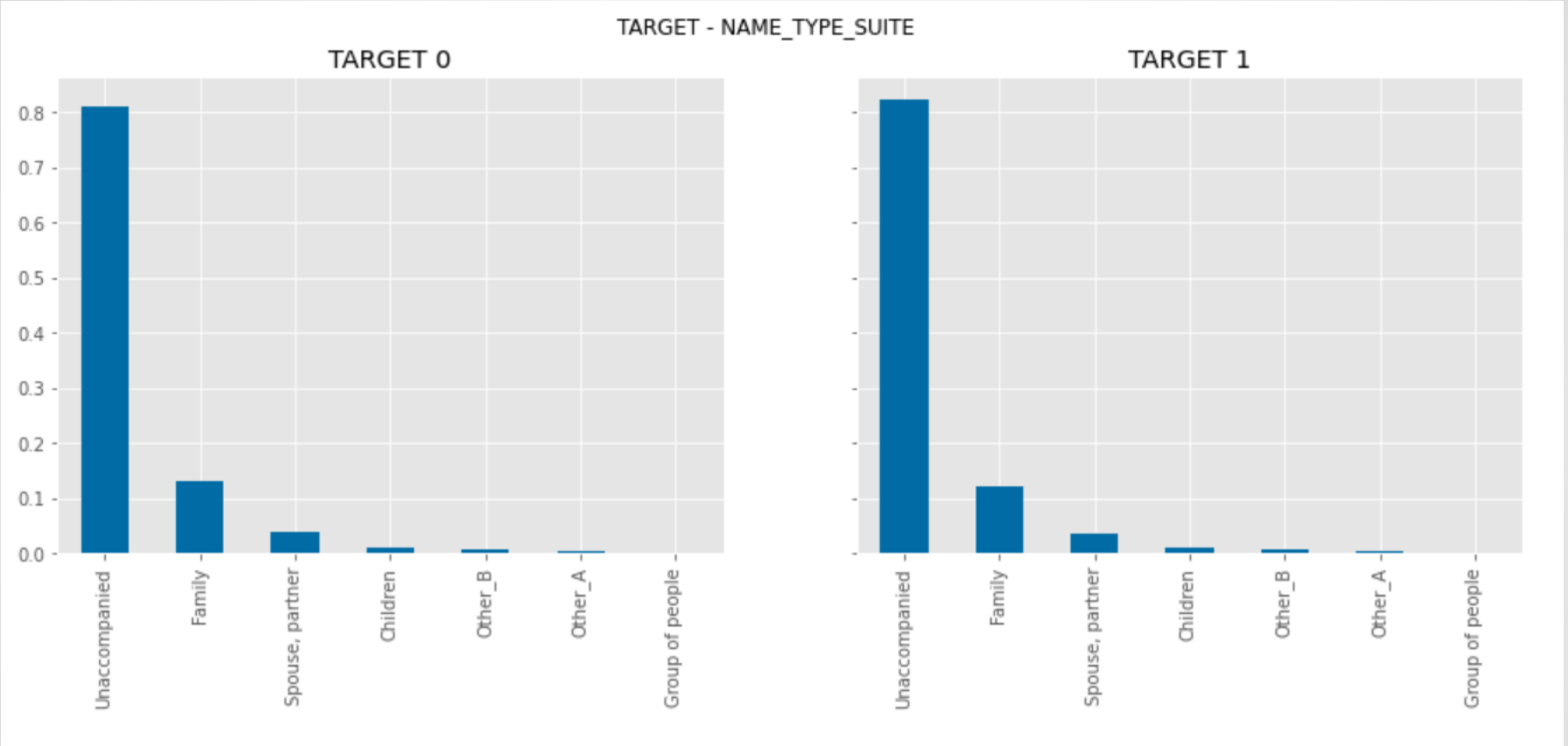
+ Tendency of defaulting loans is independent of the Application start day of the week.

- Tendency of defaulting loans is almost independent of the Application start hour of the day.

SEGMENTED UNIVARIANT ANALYSIS

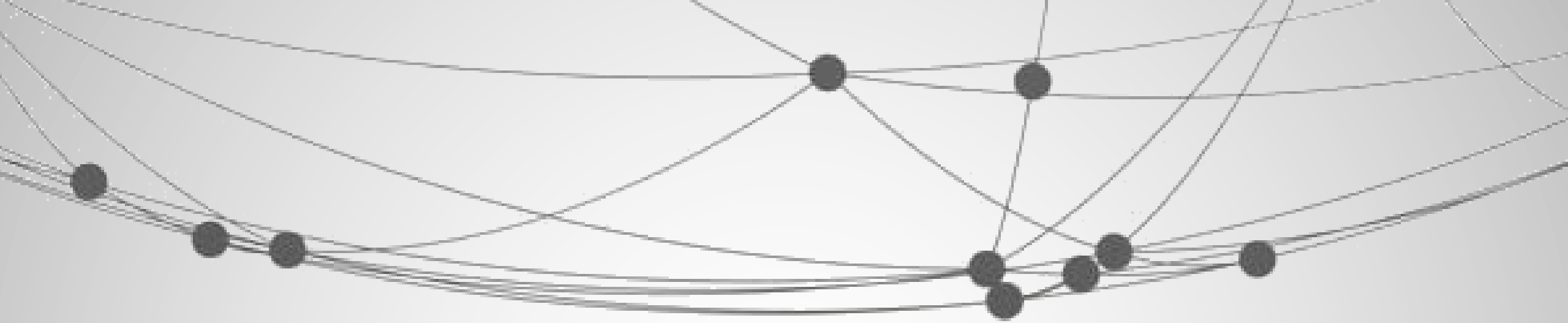
SET 5

PLOTS



INFERENCES

+ Tendency of defaulting a loan is independent of the accompanying person



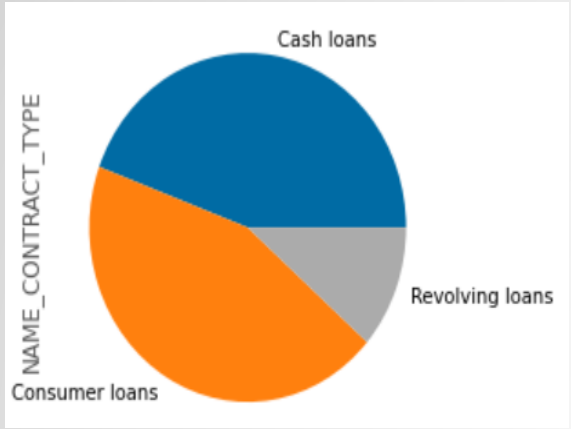
INFERENCES

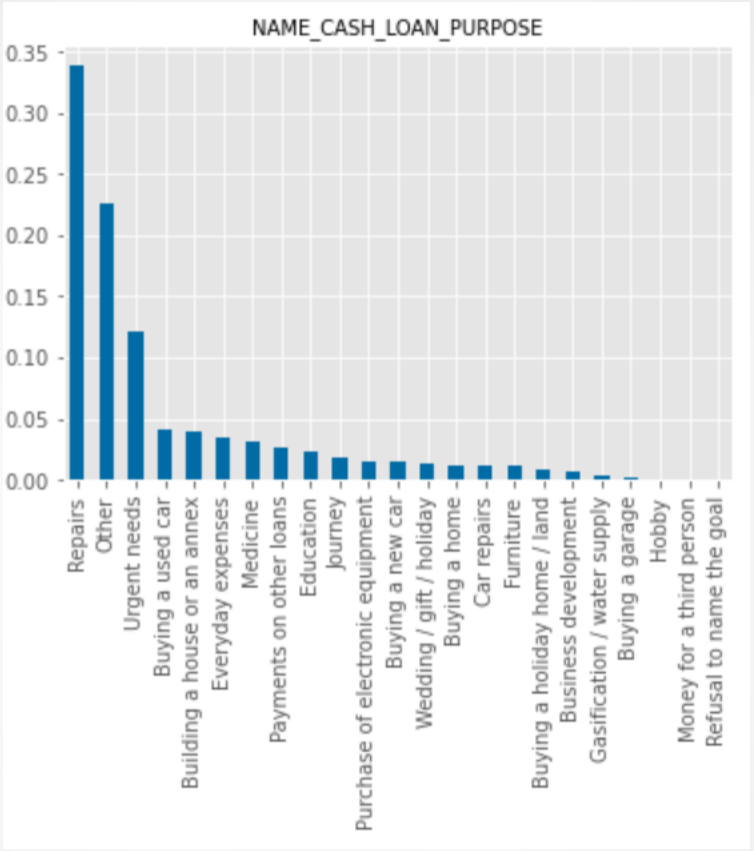
PREVIOUS APPLICATION DATA

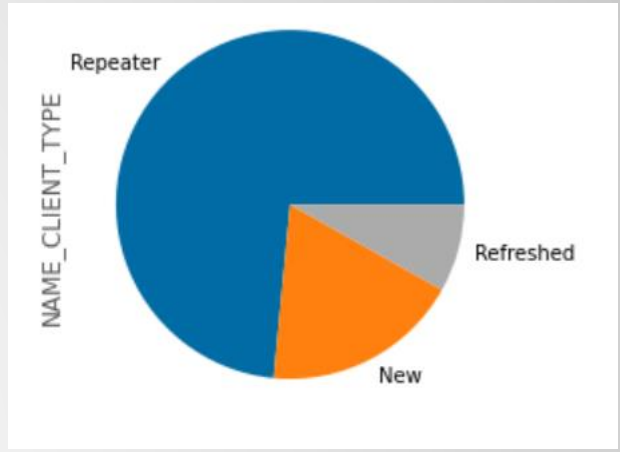
UNIVARIANT ANALYSIS

PREVIOUS APPLICATION DATA

PLOTS







INFERENCES

+

#Consumer & Cash loans are around 40% each

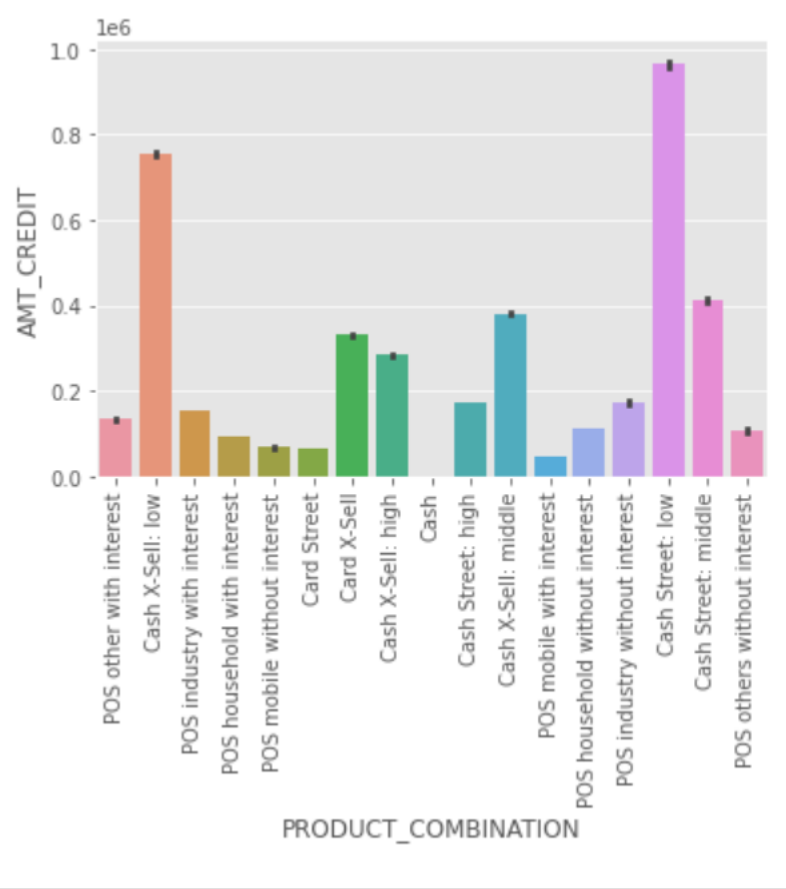
#Repairs is the purpose for approx. 32% of the cash loans

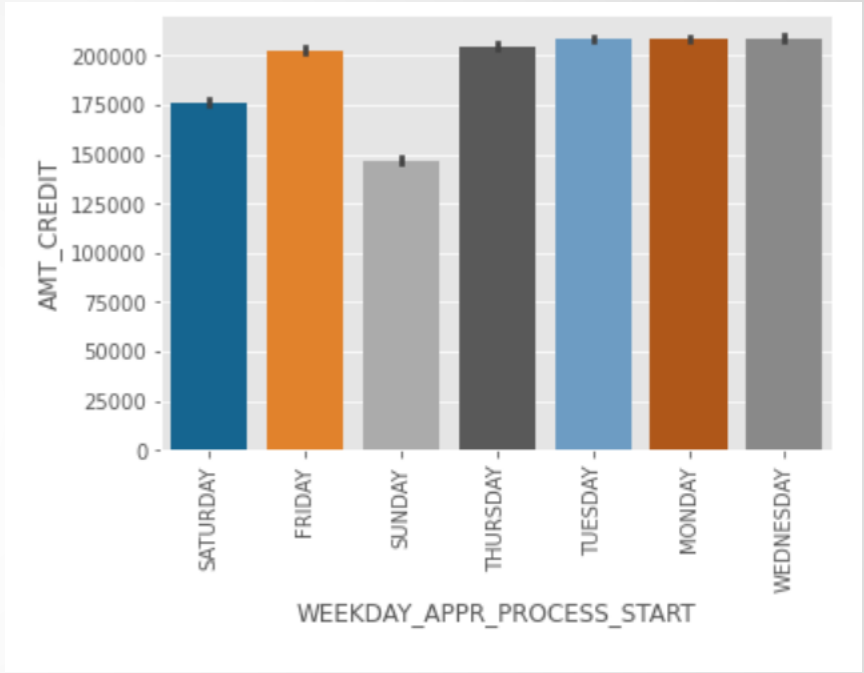
#70% of the clients are repeater

BIVARIANT ANALYSIS

PREVIOUS APPLICATION DATA

PLOTS





INFERENCES

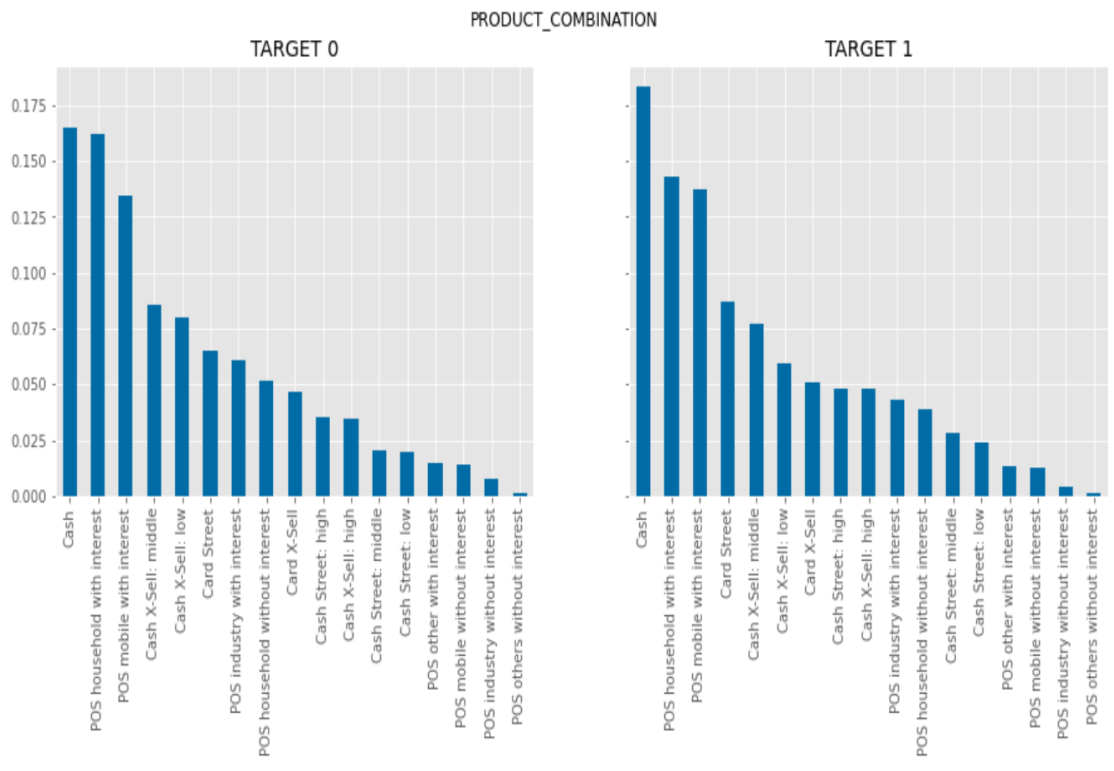
\+ Cash Street: low and Cash X-Sell:low – Product Type has the highest credit amounts

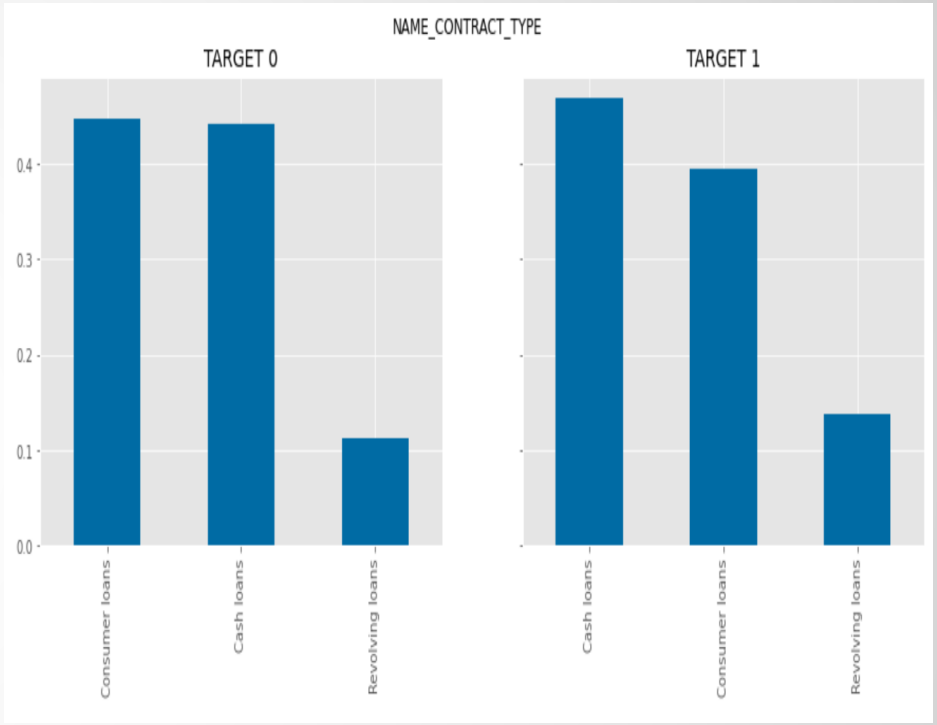
\- Applications started in sundays have the lowest mean credit amounts

SEGMENTED UNIVARIANT ANALYSIS

PREVIOUS APPLICATION DATA

PLOTS



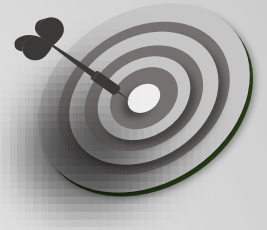


INFERENCES

+ POS household through interest may be slightly more reliable than cash product combination to decrease the risk of defaulters.

- Proportion of Consumer loans is less in defaulters set compared to others. Hence Consumer loans may be more preferred

CONCLUSION



In conclusion, top 10 correlations for the Client with payment difficulties and all other cases as per the inferences of the univariant, Bivariant and Segmented univariant analysis are as follows:

Category of attributes to be more preferred for loan approval
Higher education type
Core staff, Managers, Pensioners
Families with higher no of members

Category of attributes to be less preferred for loan approval
Male consumers
Laborers, Drivers, Sales Staff
Working Income type
Single/not married & Civil marriage consumers

Category of attributes that are almost independent for loan approval
Accompanying person while application processing start
Application start day of the week
Application start hour of the day

Best regards,

Thank you