Credit EDA Case Study

Team Members:
- Tejaswini Kamath
- Ranjiv Sukumaran

Table of Contents

Purpose

Loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history.

Some consumers misuse this and become defaulters.

The following presentation aims to use EDA to analyse the patterns present in the data and infer which criteria need to be considered before deciding on whether to approve or reject any loan application received.

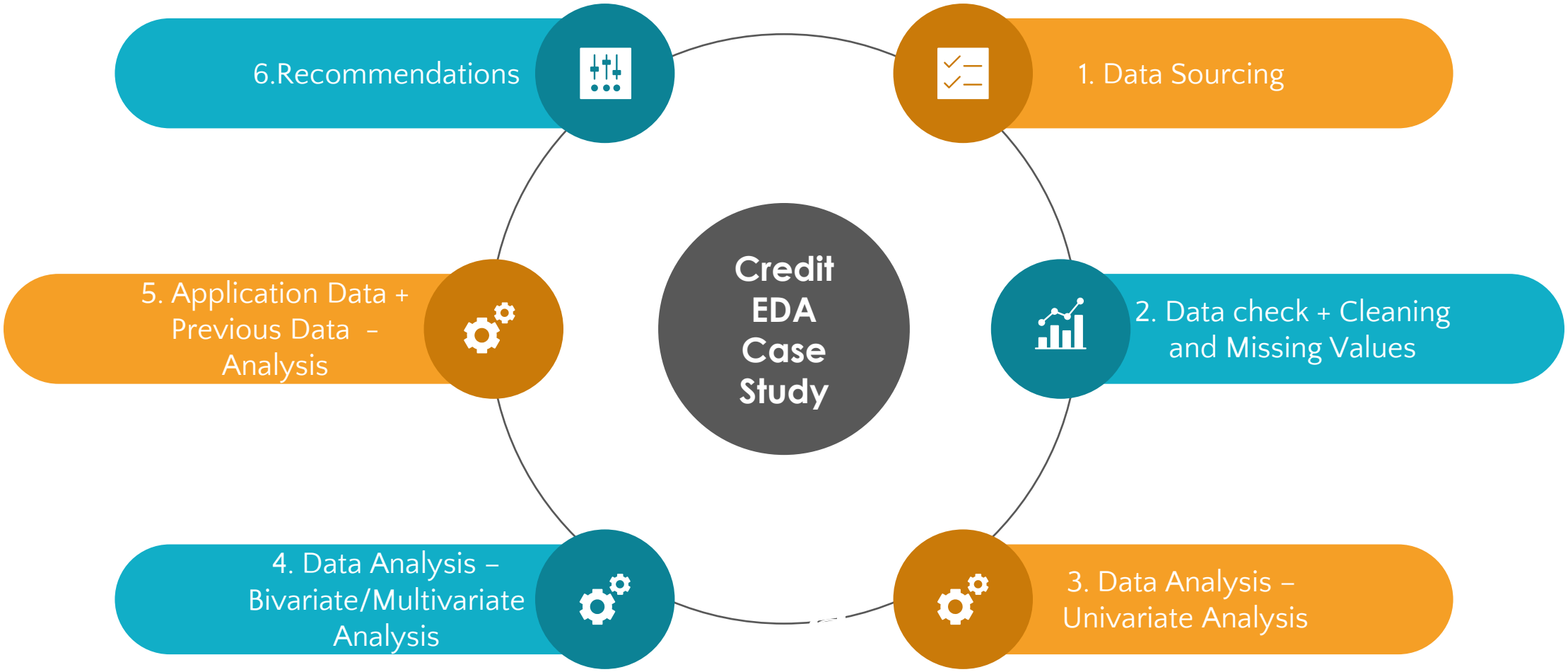
The study will aim to help the company to decide that, when it receives a loan application, how to decide the loan approval based on the applicant's profile.

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.

This presentation will help make such recommendations regarding the decisions.

Credit EDA Case Study Analysis Stages

The analysis will take place in the following 6 stages:



Data Sourcing

- Data is present in two files:

1) Applications data
2) Previous applications data

Application data contains details from current application and indicates whether any of the customers have defaulted for the first X payments as well as different details of the customer.

Previous applications data contains details of the previous applications by the customer and whether they were Approved, Cancelled, Refused or Unused.

Data Cleaning Phases

Step 1:

Routine Checking and analyzing columns for missing data.

Dropping columns as required with respect to missing data percentage

Step 2:

Imputing incorrect values

Handling missing values with very few nulls

Step 3:

Fixing rows and columns as per data required for analysis

Step 4:

Checking for outliers and suggested treatments

Step 5:

Binning the Continuous Variables

Data Cleaning Phase: Step 1

- Routine Checking and analyzing columns for missing data.

1. Finding

- Dropping columns as required with respect to missing data percentage

Data Cleaning Phases

Final targetDataset snapshot (info output):

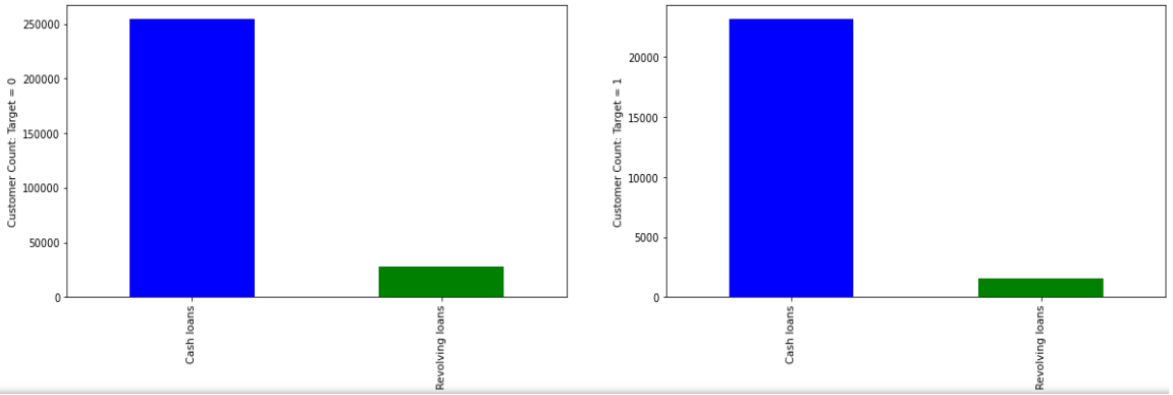
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Data columns (total 61 columns):
 #   Column                       Non-Null Count   Dtype
---  ------                       --------------   -----
 0   SK_ID_CURR                   307511 non-null  int64
 1   TARGET                       307511 non-null  int64
 2   NAME_CONTRACT_TYPE           307511 non-null  object
 3   CODE_GENDER                  307511 non-null  object
 4   FLAG_OWN_CAR                 307511 non-null  object
 5   FLAG_OWN_REALTY              307511 non-null  object
 6   CNT_CHILDREN                 307511 non-null  int64
 7   AMT_INCOME_TOTAL             307511 non-null  float64
 8   AMT_CREDIT                   307511 non-null  float64
 9   AMT_ANNUITY                  307499 non-null  float64
 10  AMT_GOODS_PRICE              307233 non-null  float64
 11  NAME_TYPE_SUITE              306219 non-null  object
 12  NAME_INCOME_TYPE             307511 non-null  object
 13  NAME_EDUCATION_TYPE          307511 non-null  object
 14  NAME_FAMILY_STATUS           307511 non-null  object
 15  NAME_HOUSING_TYPE            307511 non-null  object
 16  REGION_POPULATION_RELATIVE   307511 non-null  float64
 17  DAYS_BIRTH                   307511 non-null  int64
 18  DAYS_EMPLOYED                252137 non-null  float64
 19  DAYS_REGISTRATION            307511 non-null  float64
 20  DAYS_ID_PUBLISH              307511 non-null  int64
 21  FLAG_MOBIL                   307511 non-null  int64
 22  FLAG_EMP_PHONE               307511 non-null  int64
 23  FLAG_WORK_PHONE              307511 non-null  int64
 24  FLAG_CONT_MOBILE             307511 non-null  int64
 25  FLAG_PHONE                   307511 non-null  int64
 26  FLAG_EMAIL                   307511 non-null  int64
 27  OCCUPATION_TYPE              211120 non-null  object
 28  CNT_FAM_MEMBERS              307509 non-null  float64
 29  REGION_RATING_CLIENT         307511 non-null  int64
 30  REGION_RATING_CLIENT_W_CITY  307511 non-null  int64
 31  WEEKDAY_APPR_PROCESS_START   307511 non-null  object
 32  HOUR_APPR_PROCESS_START      307511 non-null  int64
 33  REG_REGION_NOT_LIVE_REGION   307511 non-null  int64
 34  REG_REGION_NOT_WORK_REGION   307511 non-null  int64
 35  LIVE_REGION_NOT_WORK_REGION  307511 non-null  int64
 36  REG_CITY_NOT_LIVE_CITY       307511 non-null  int64
 37  REG_CITY_NOT_WORK_CITY       307511 non-null  int64
 38  LIVE_CITY_NOT_WORK_CITY      307511 non-null  int64
 39  ORGANIZATION_TYPE            307511 non-null  object
 40  EXT_SOURCE_2                 306851 non-null  float64
 41  EXT_SOURCE_3                 246546 non-null  float64
 42  DAYS_LAST_PHONE_CHANGE       307510 non-null  float64
 43  AMT_REQ_CREDIT_BUREAU_HOUR   265992 non-null  float64
 44  AMT_REQ_CREDIT_BUREAU_DAY    265992 non-null  float64
 45  AMT_REQ_CREDIT_BUREAU_WEEK   265992 non-null  float64
 46  AMT_REQ_CREDIT_BUREAU_MON    265992 non-null  float64
 47  AMT_REQ_CREDIT_BUREAU_QRT    265992 non-null  float64
 48  AMT_REQ_CREDIT_BUREAU_YEAR   265992 non-null  float64
 49  Age                          307511 non-null  float64
 50  Tenure                       252137 non-null  float64
 51  YearRegistran                307511 non-null  float64
 52  YearID                       307511 non-null  float64
 53  TenureBins                   242385 non-null  category
 54  Incomebins                   307511 non-null  category
 55  AnnuityBins                  307499 non-null  category
 56  CreditBins                   307511 non-null  category
 57  AGEBINS                      307511 non-null  category
 58  RegionPopulationBins         307511 non-null  category
 59  RegistrationBins             298504 non-null  category
 60  IDBins                       299650 non-null  category
dtypes: category(8), float64(21), int64(20), object(12)
memory usage: 126.7+ MB
```

Data Analysis

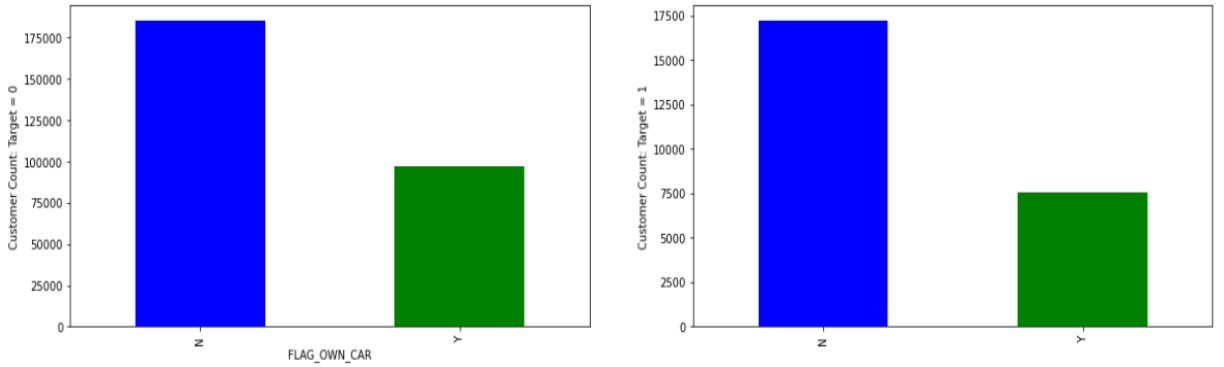
IMBALANCED DATASET : It can be observed that Target value distribution as well as value counts shows a major imbalance in data with values with Target = 0 being over 10 times that of data with Target = 1

Data Analysis – Univariate Analysis

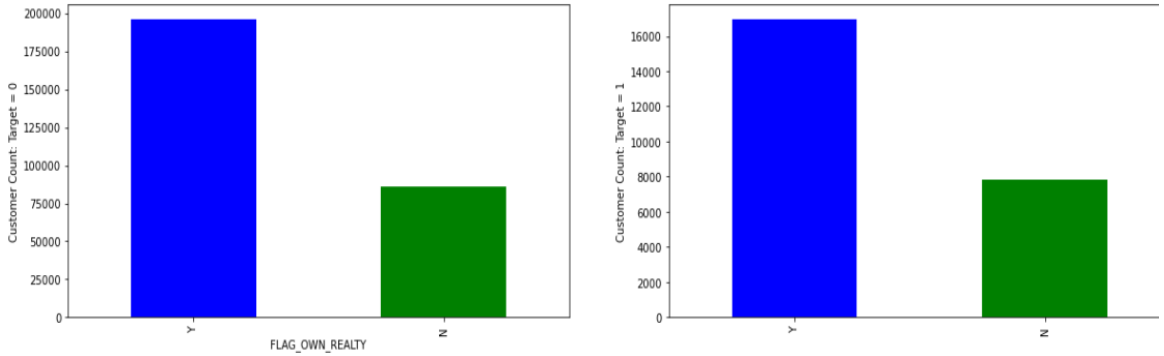
Comparing Target 0 i.e. situations where Clients have payment difficulties with Target 1 where there have been no defaults observed



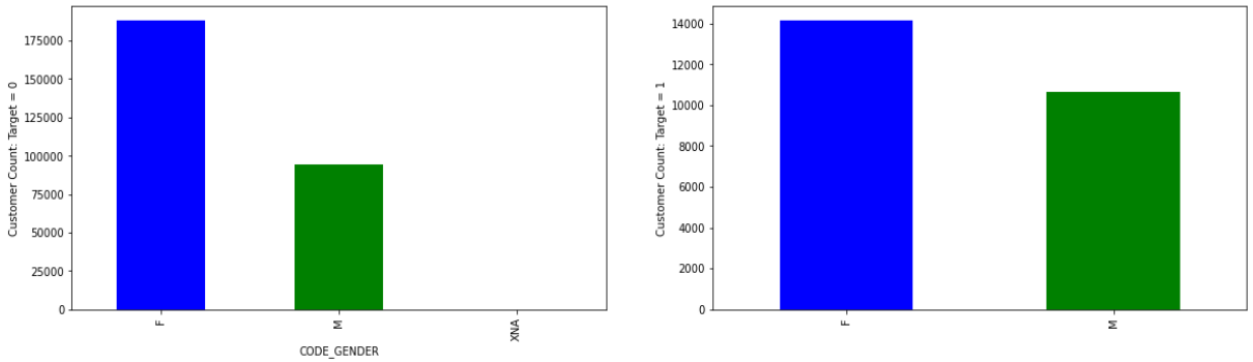
Its observed that in both sets Cash Loans are higher as compared to Revolving Loans.



Its observed that in both sets a higher number of Customers do not have their own cars



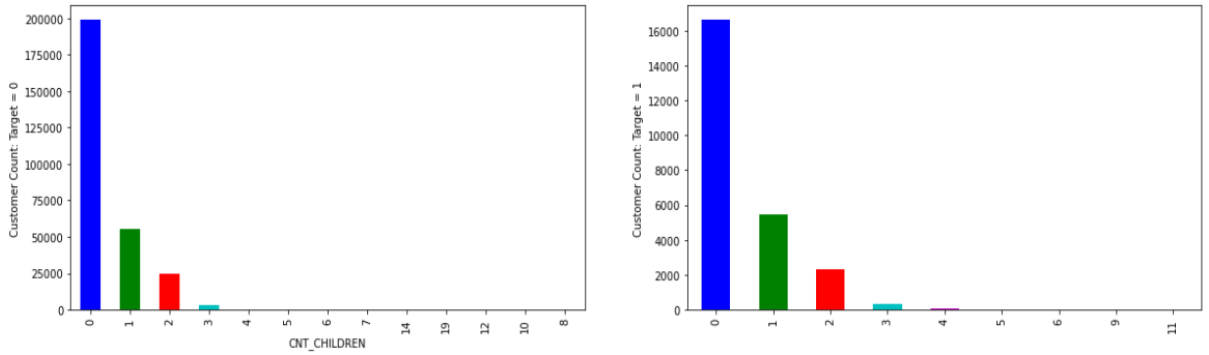
Its observed that in both sets a higher number of Customers do not have their own realty



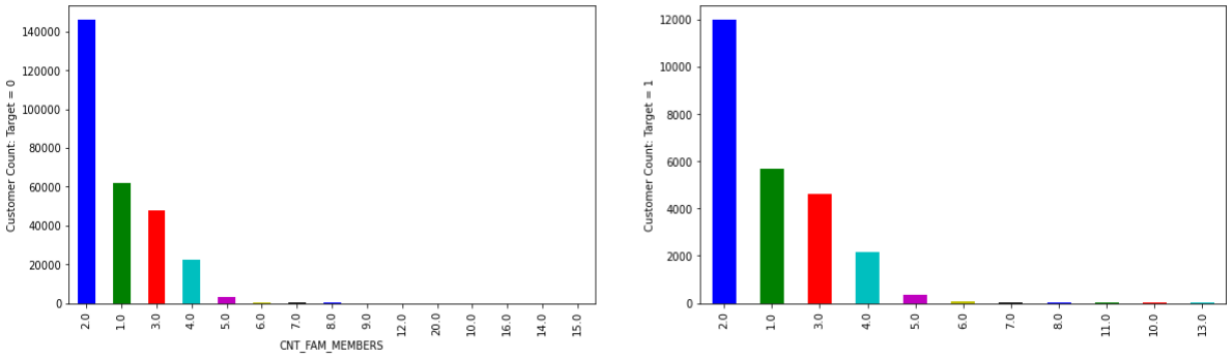
Its observed that in both sets a higher number of female Customers is observed

Data Analysis – Univariate Analysis

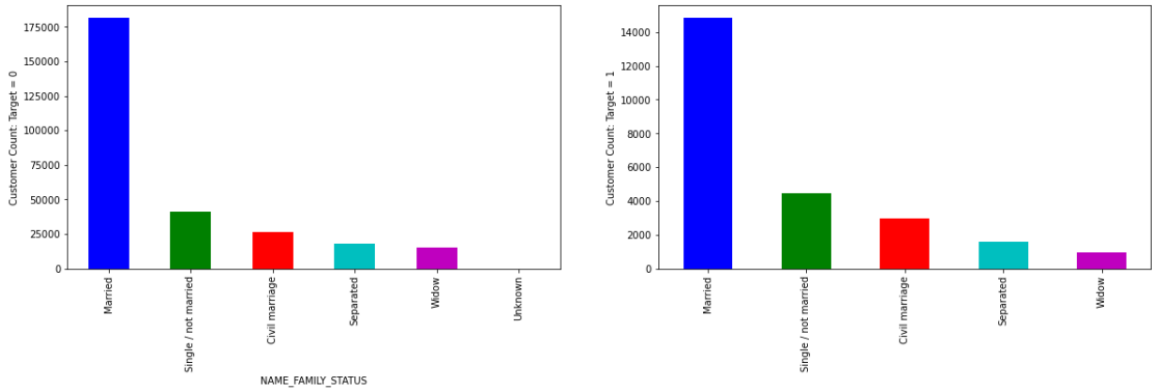
Comparing Target 0 i.e. situations where Clients have payment difficulties with Target 1 where there have been no defaults observed



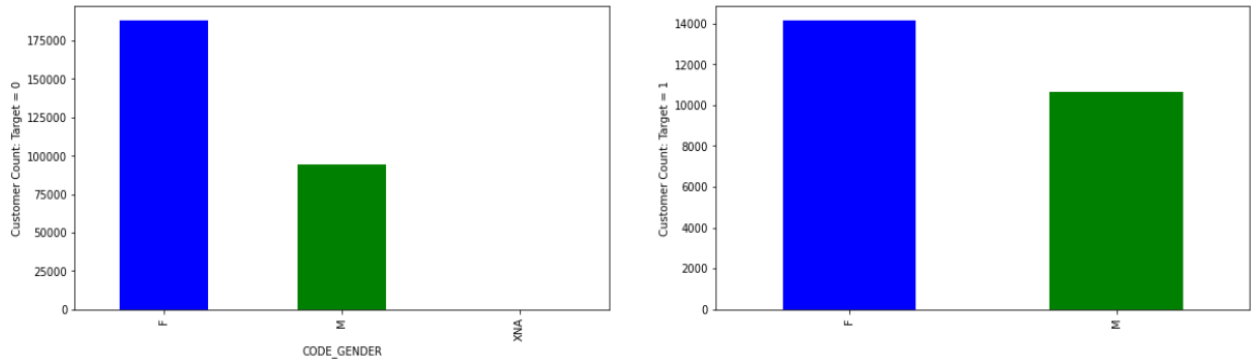
Its observed that in both sets a higher number of Customers have no children



Its observed that in both sets a higher number of Customers have family count of 2



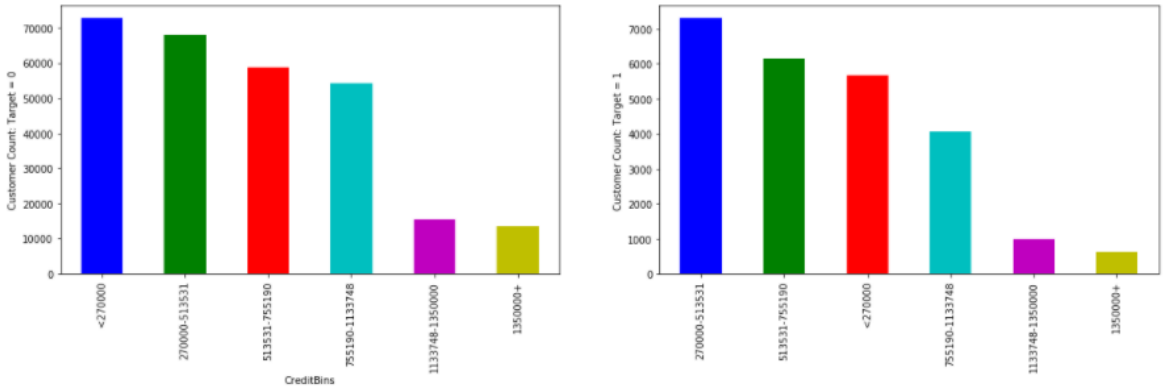
Its observed that in both sets a higher number of Customers have a Family Status of 'Married'. This also matches the previous observation indicating highest customers in both set having Family Size = 2



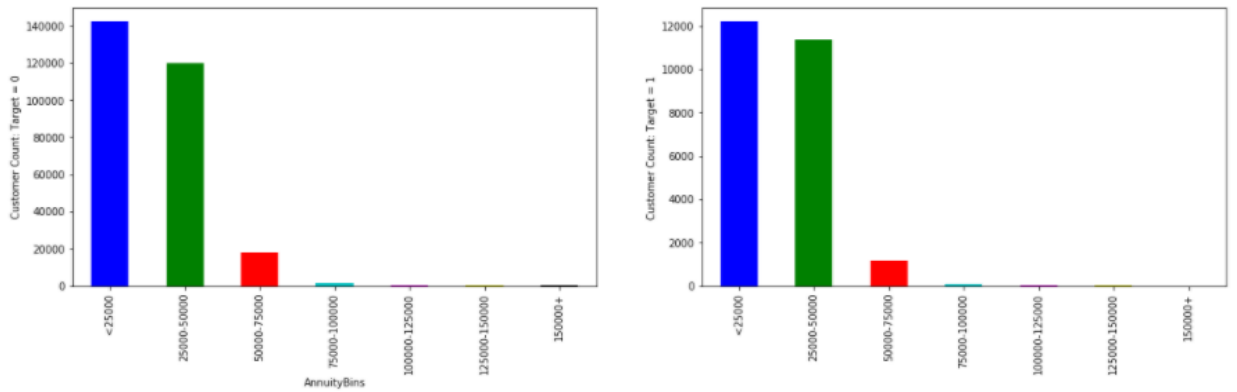
Its observed that in both sets a higher number of female Customers is observed

Data Analysis – Univariate Analysis

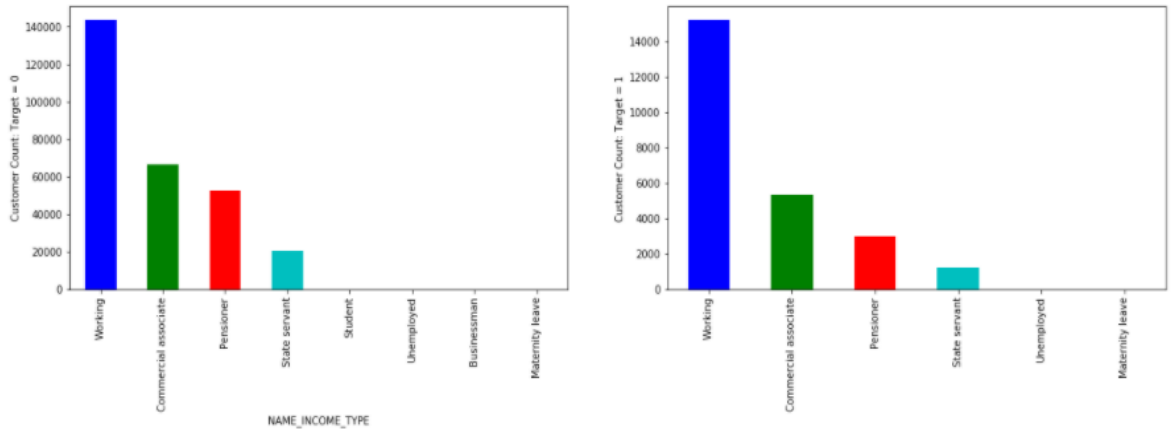
Comparing Target 0 i.e. situations where Clients have payment difficulties with Target 1 where there have been no defaults observed



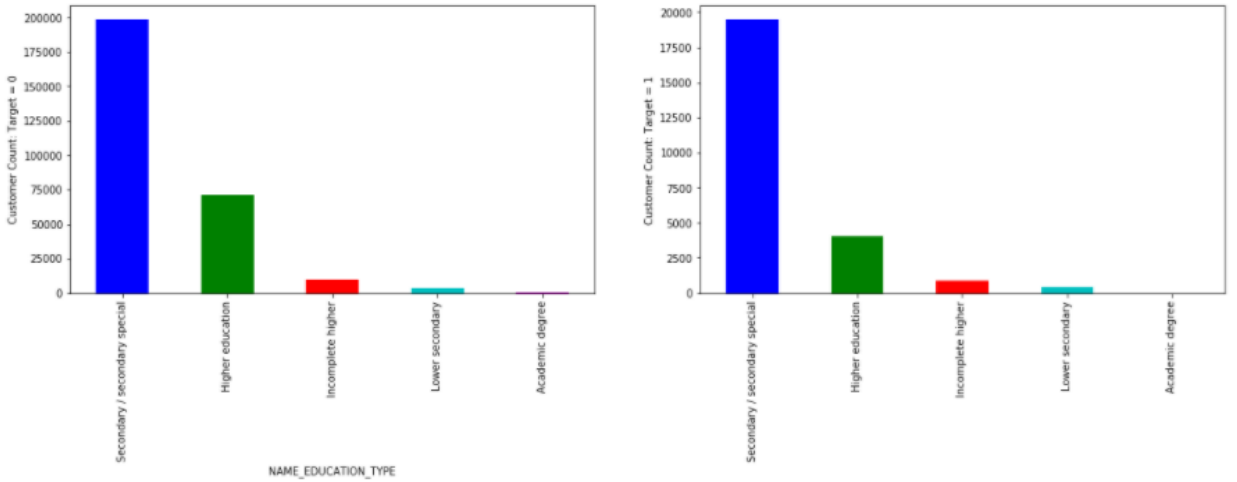
While most non-defaulting customers appear to have an income <270000, most defaulting customers appear to have an income in the range '270000-517788'



Its observed that in both sets a higher number of Customers have <25000 as annuity amount



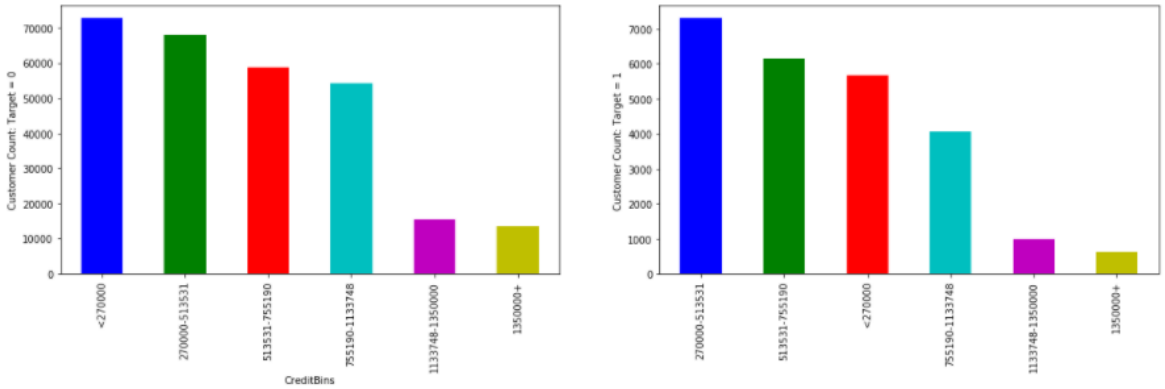
Its observed that in both sets a higher number of Customers have an Income Type = Working



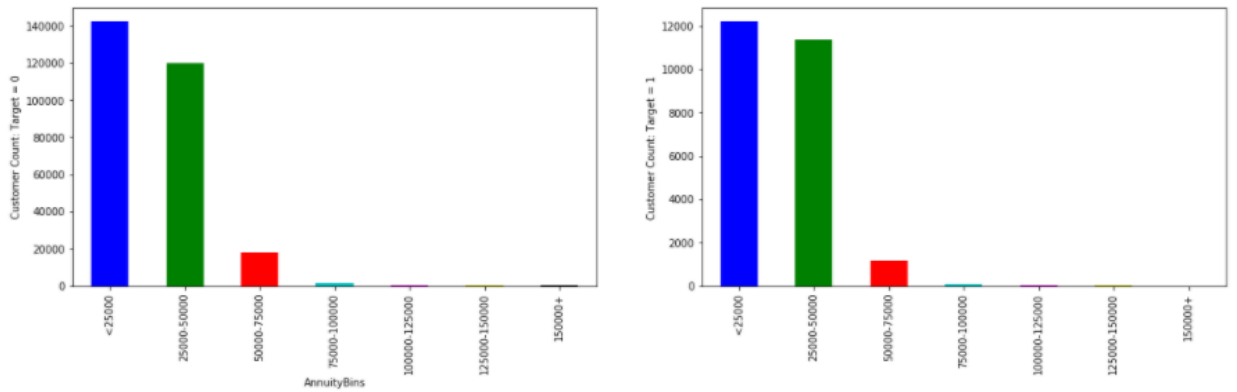
Its observed that in both sets a higher number of Customers have secondary/secondary special education

Data Analysis – Univariate Analysis

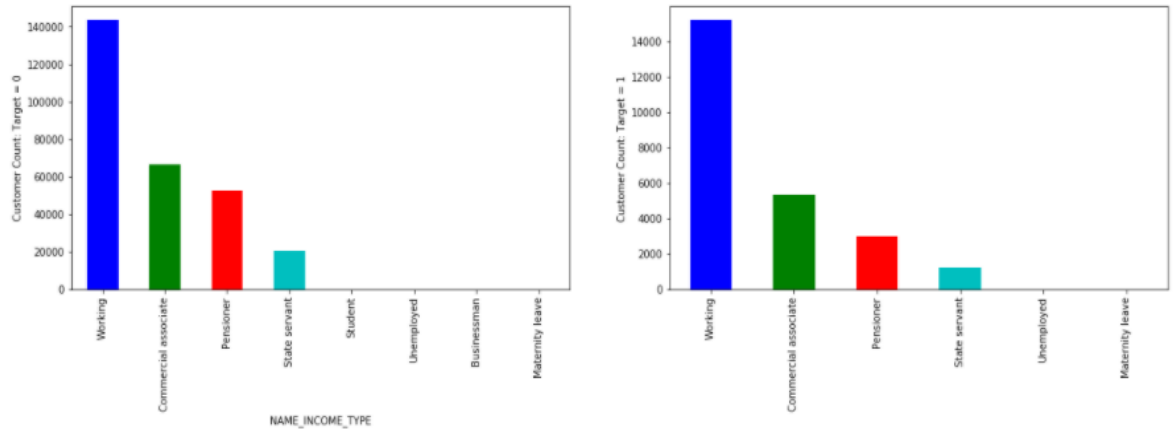
Comparing Target 0 i.e. situations where Clients have payment difficulties with Target 1 where there have been no defaults observed



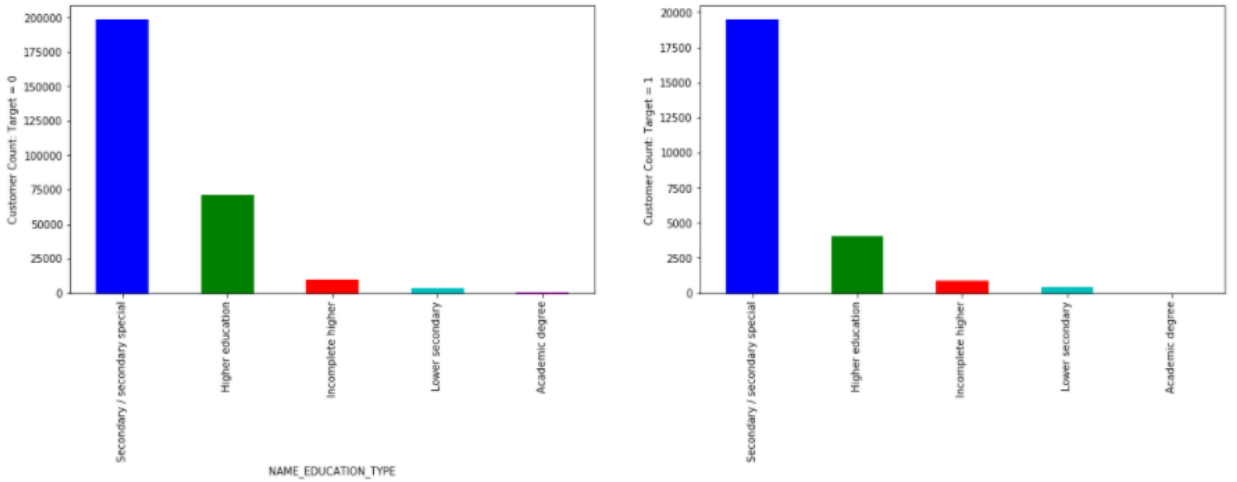
While most non-defaulting customers appear to have an income <270000, most defaulting customers appear to have an income in the range '270000-517788'



Its observed that in both sets a higher number of Customers have <25000 as annuity amount



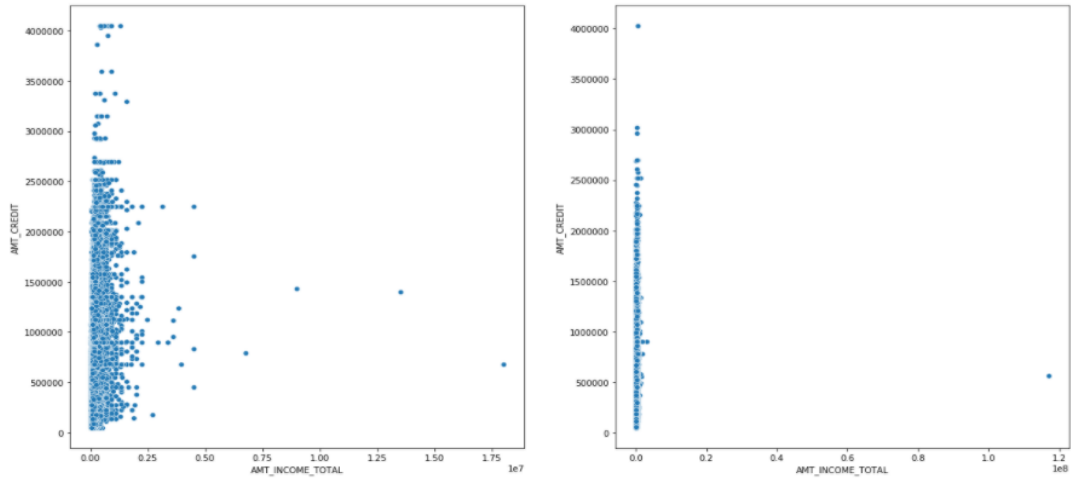
Its observed that in both sets a higher number of Customers have an Income Type = Working



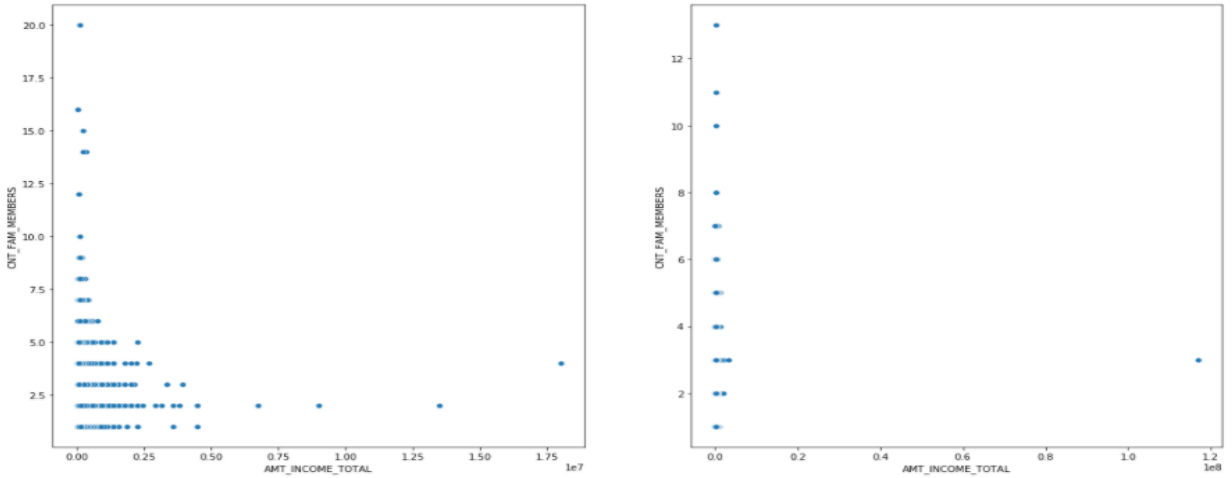
Its observed that in both sets a higher number of Customers have secondary/secondary special education

Data Analysis – Bivariate Analysis

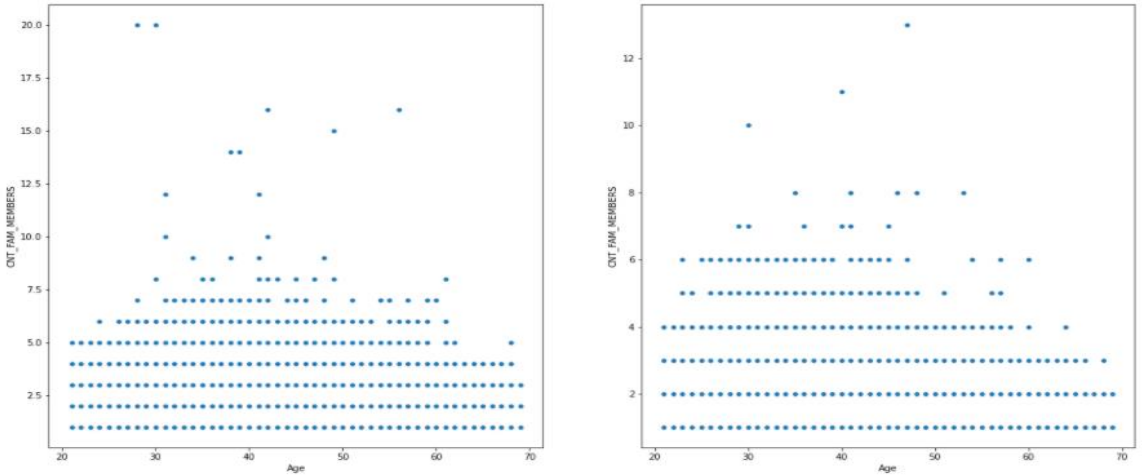
Comparing Target 0 i.e. situations where Clients have payment difficulties with Target 1 where there have been no defaults observed



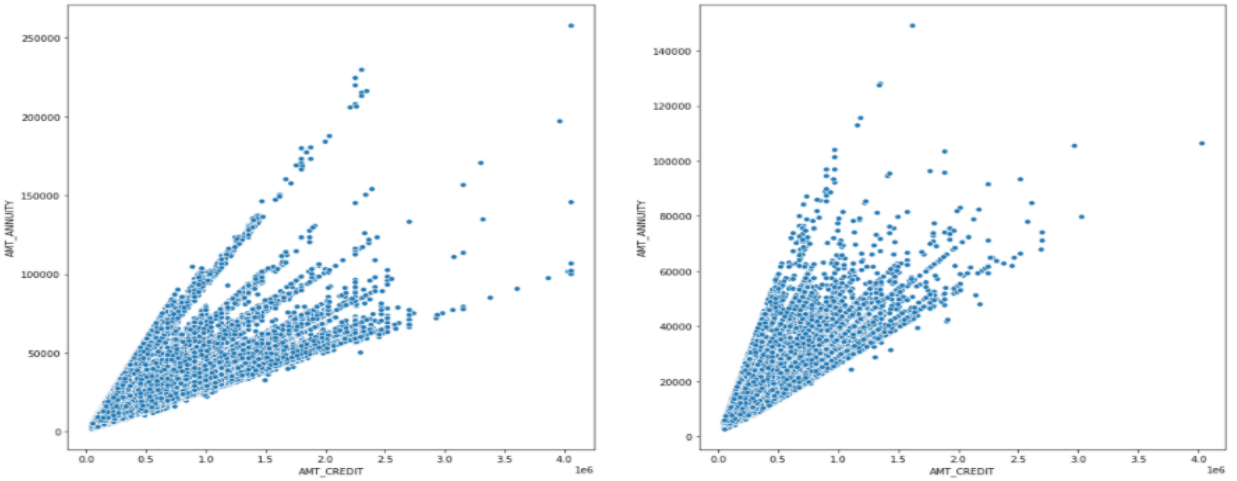
Both datasets indicate a high correlation between the two variables - AMT_INCOME_TOTAL and AMT_CREDIT. However there appear to be more outliers for non-defaulting customers in comparison to defaulting customers



Both datasets indicate a high correlation between the two variables - AMT_INCOME_TOTAL and CNT_FAM_MEMBERS. However there appear to be more outliers for non-defaulting customers in comparison to defaulting customers



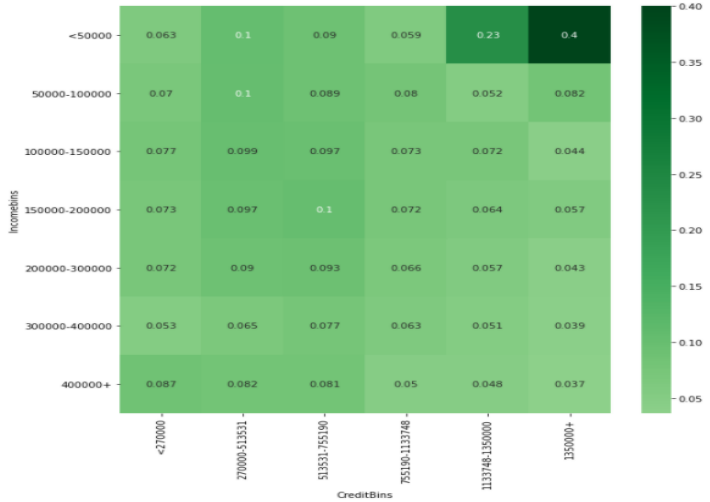
Both datasets indicate that there is no specific correlation between Age and Count of family members



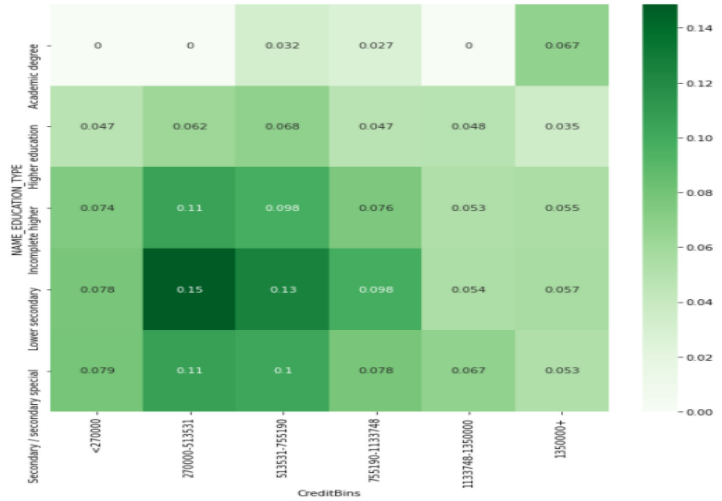
The above plots indicate that as Amount Annuity increases, Amount of credit increases as well. This holds true for both datasets

Data Analysis – Bivariate Analysis

Comparing Target 0 i.e. situations where Clients have payment difficulties with Target 1 where there have been no defaults observed

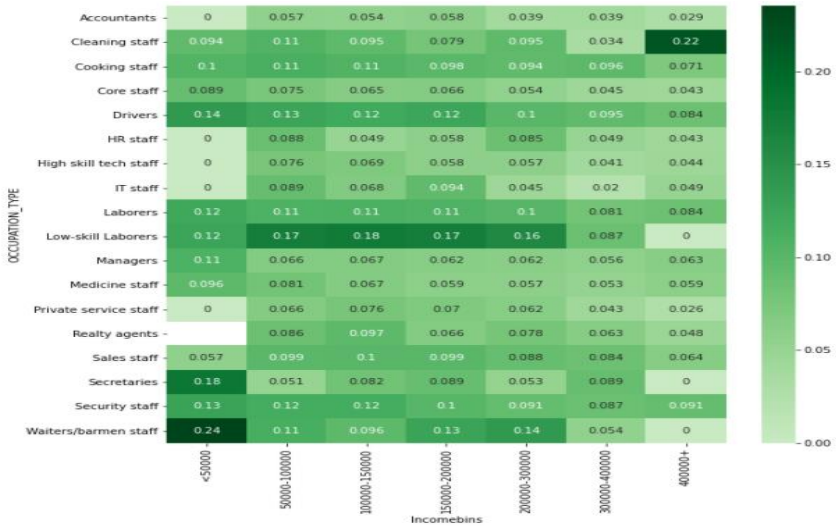


The above heatmap indicates that there is not much of a correlation between Income and Credit bins when plotted against the target database (i.e. both Target values 0 and 1. Maximum correlation observed = 0.4



The above heatmap indicates that there is very low correlation between Education of customer and Credit bins when plotted against the target database (i.e. both Target values 0 and 1. Maximum correlation observed = 0.15

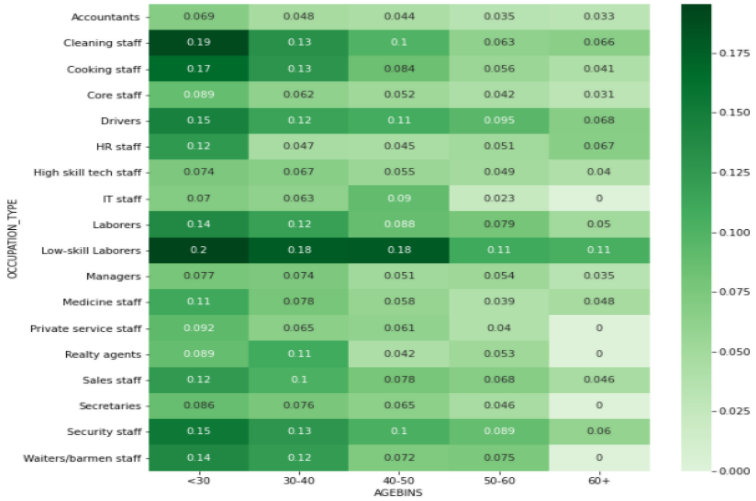
The above heatmap indicates that there is very low correlation between Education of customer and Income bins when plotted against the target database (i.e. both Target values 0 and 1. Maximum correlation observed = 0.13



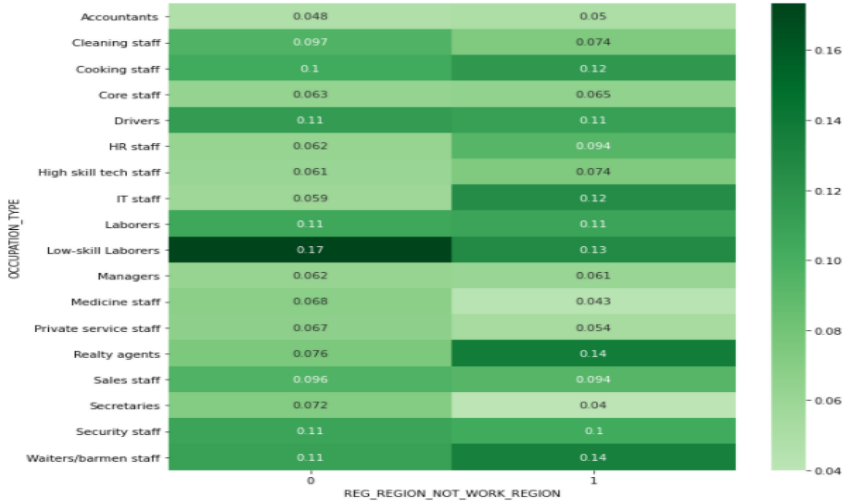
The above heatmap indicates that there is very low correlation between Occupation Type of customer and Income bins when plotted against the target database (i.e. both Target values 0 and 1. Maximum correlation observed = 0.24

Data Analysis – Bivariate Analysis

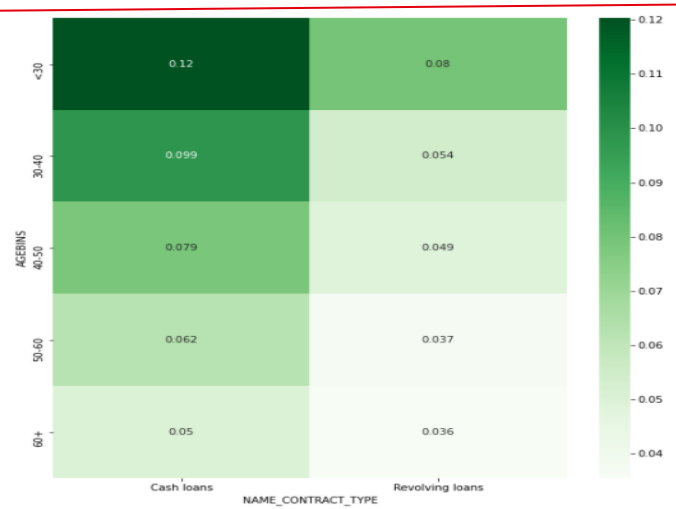
Comparing Target 0 i.e. situations where Clients have payment difficulties with Target 1 where there have been no defaults observed



The above heatmap indicates that there is very low correlation between Occupation of customer and Age bins when plotted against the target database (i.e. both Target values 0 and 1. Maximum correlation observed = 0.2



The above heatmap indicates that there is very low correlation between Occupation of customer and Registration region is not same as work region when plotted against the target database (i.e. both Target values 0 and 1. Maximum correlation observed = 0.17

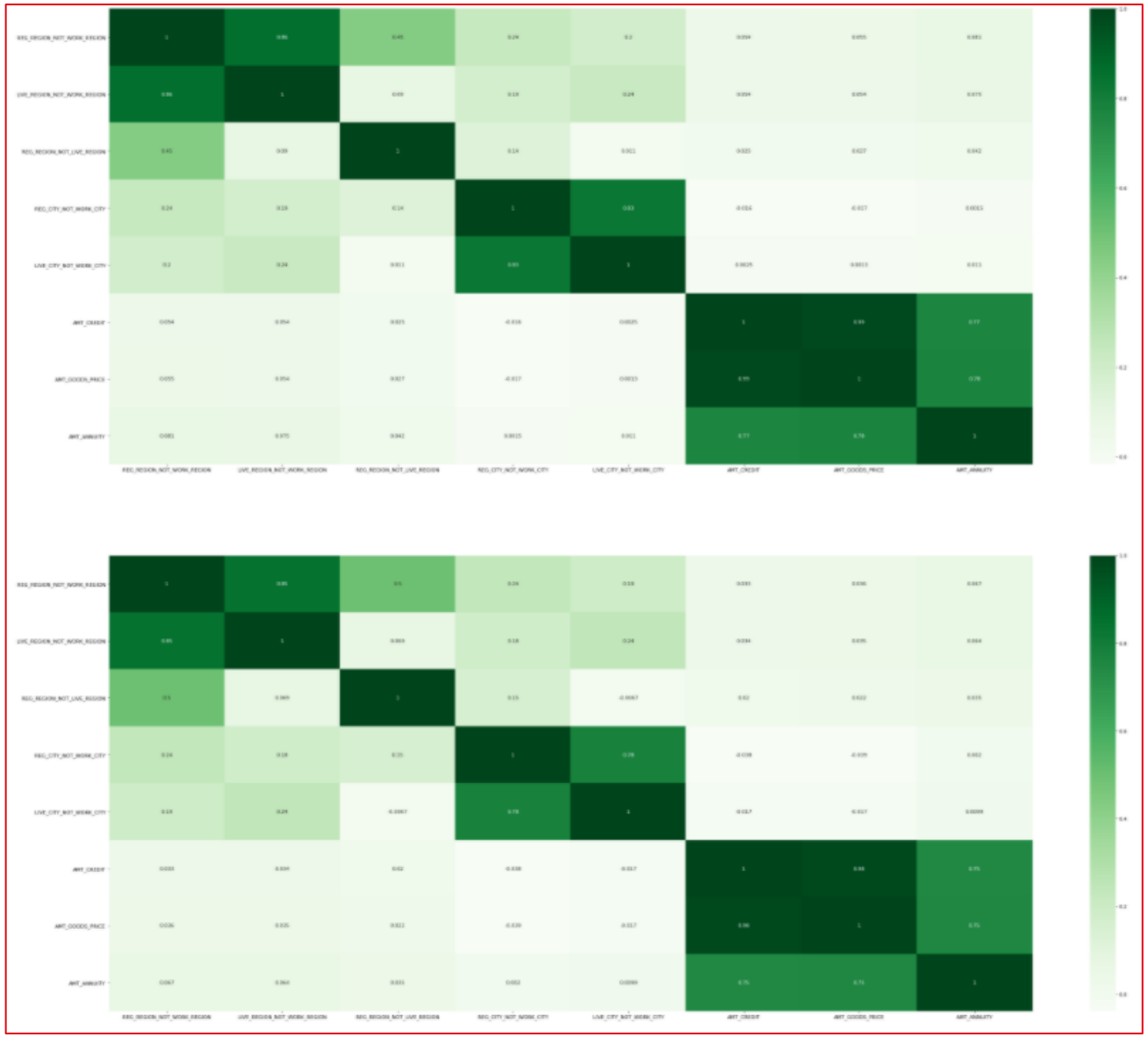


The above heatmap indicates that there is very low correlation between Contract Type and Age bins when plotted against the target database (i.e. both Target values 0 and 1. Maximum correlation observed = 0.12

The above heatmap indicates that there is low correlation between Registration bins and Credit bins when plotted against the target database (i.e. both Target values 0 and 1. Maximum correlation observed = 0.5

Data Analysis – Correlations

Comparing Target 0 i.e. situations where Clients have payment difficulties with Target 1 where there have been no defaults observed

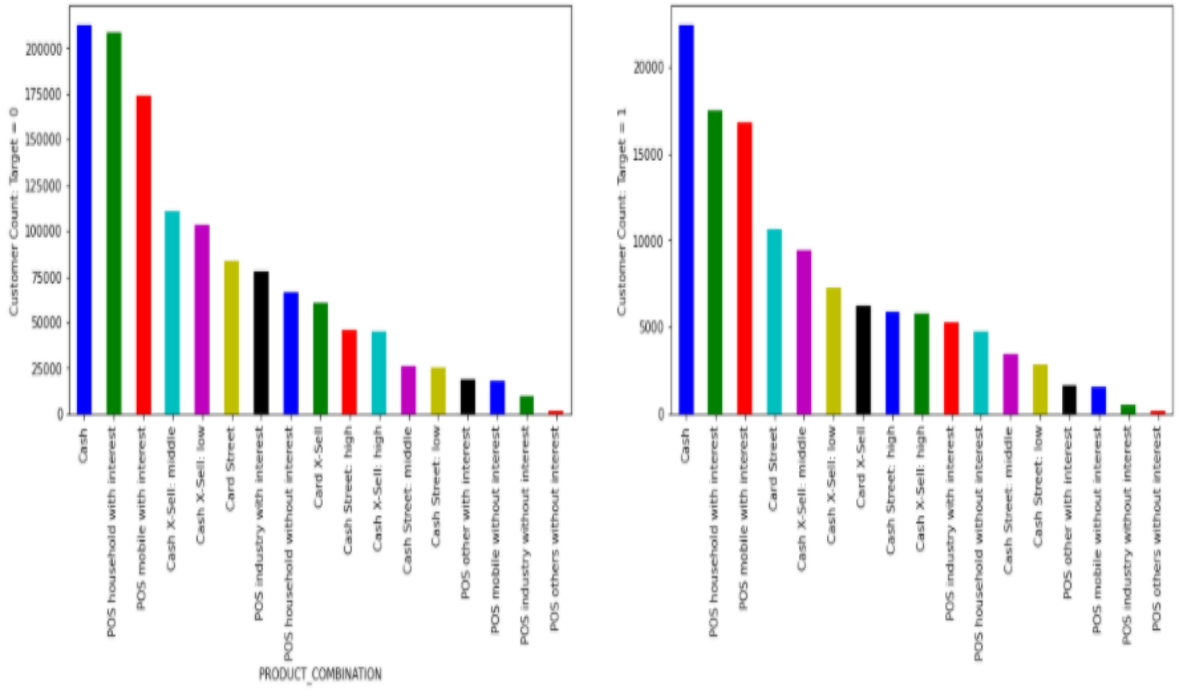


From the above diagram the same fields appear to have similar correlations in both the split data sets

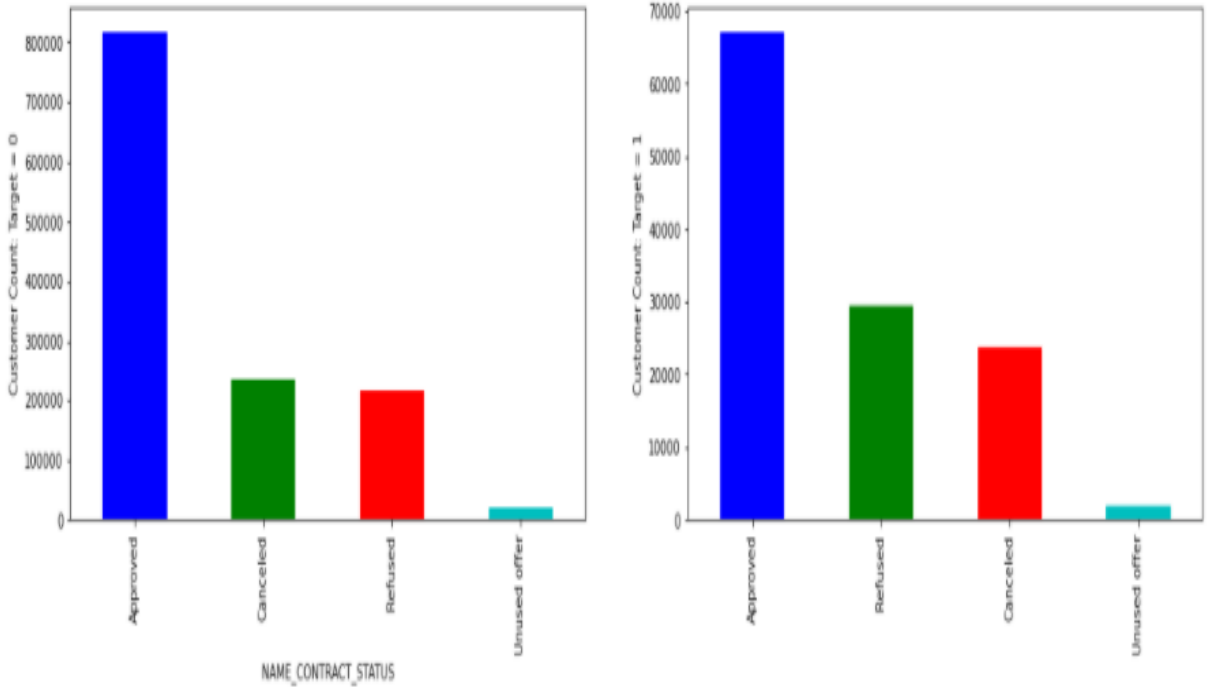
1. REG_REGION_NOT_WORK_REGION vs LIVE_REGION_NOT_WORK_REGION
2. REG_REGION_NOT_LIVE_REGION vs LIVE_REGION_NOT_WORK_REGION
3. REG_CITY_NOT_WORK_CITY vs LIVE_CITY_NOT_WORK_CITY
4. AMT_CREDIT vs AMT_GOODS_PRICE
5. AMT_ANNUITY vs AMT_GOODS_PRICE

Data Analysis – Combined Inferences -Univariate Analysis

Comparing Target 0 i.e. situations where Clients have payment difficulties with Target 1 where there have been no defaults observed



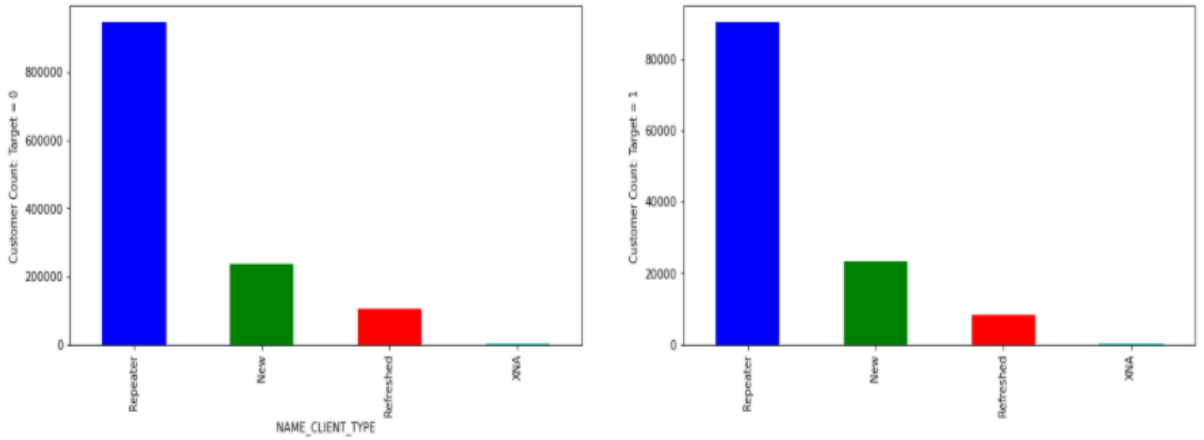
From above plot, both datasets indicate that customers preferred product combination Cash



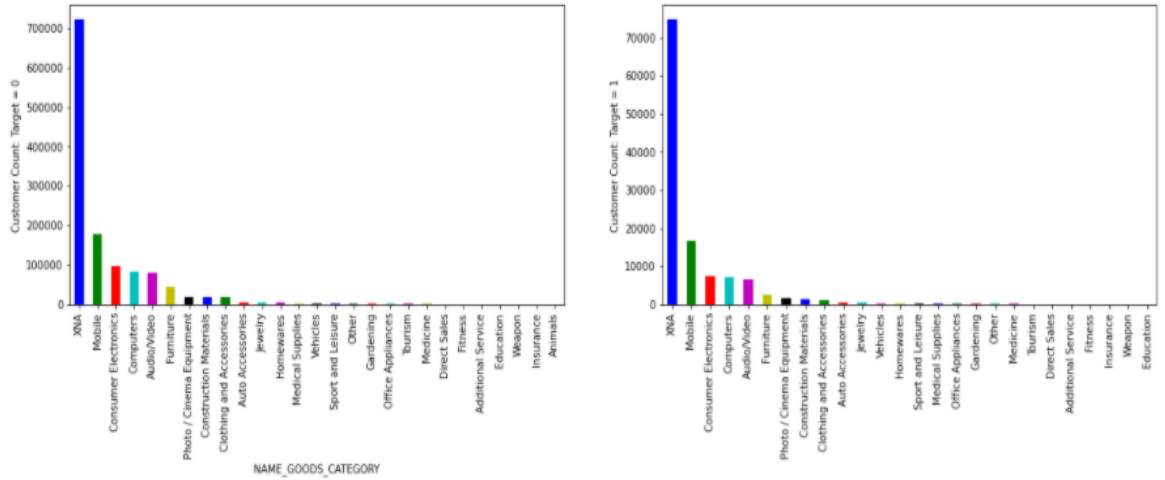
It appears from above plots that more loan applications were approved than refused in both data sets

Data Analysis – Combined Inferences -Univariate Analysis

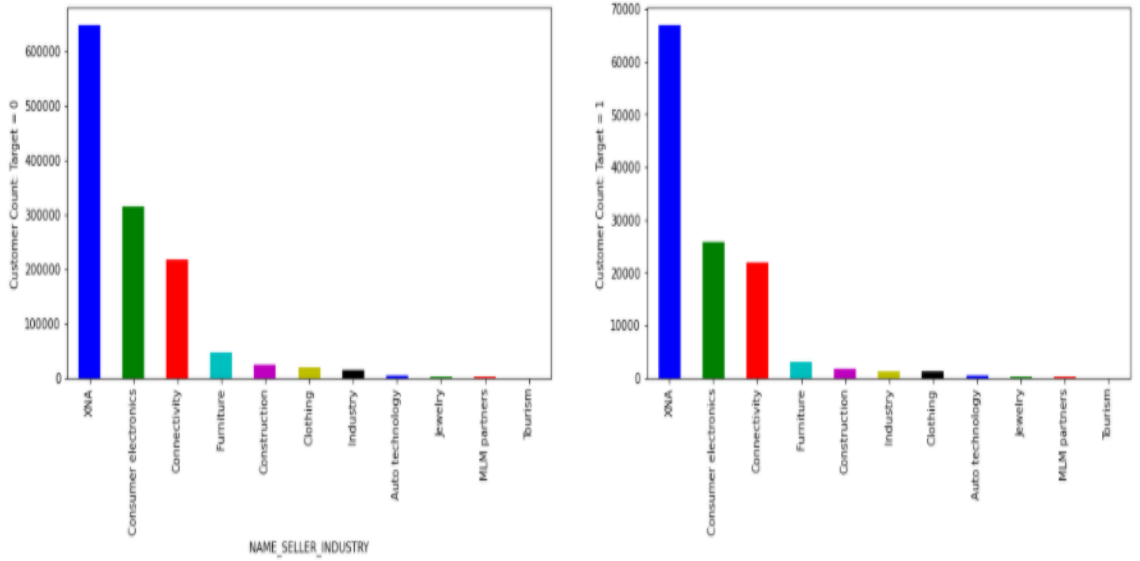
Comparing Target 0 i.e. situations where Clients have payment difficulties with Target 1 where there have been no defaults observed

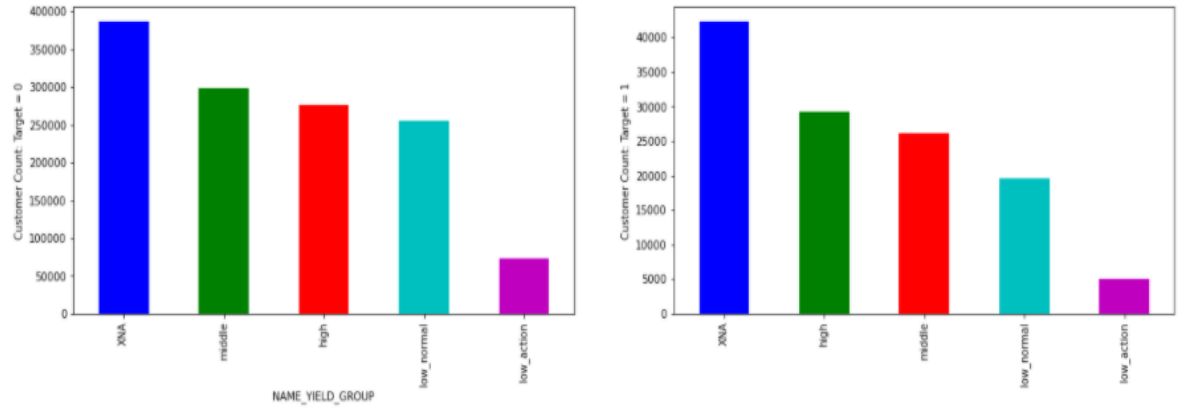


Above plots indicate that higher number of customers for both sets were repeat customers



Above plot indicates that Mobile appears to be the most common Goods category for both data sets

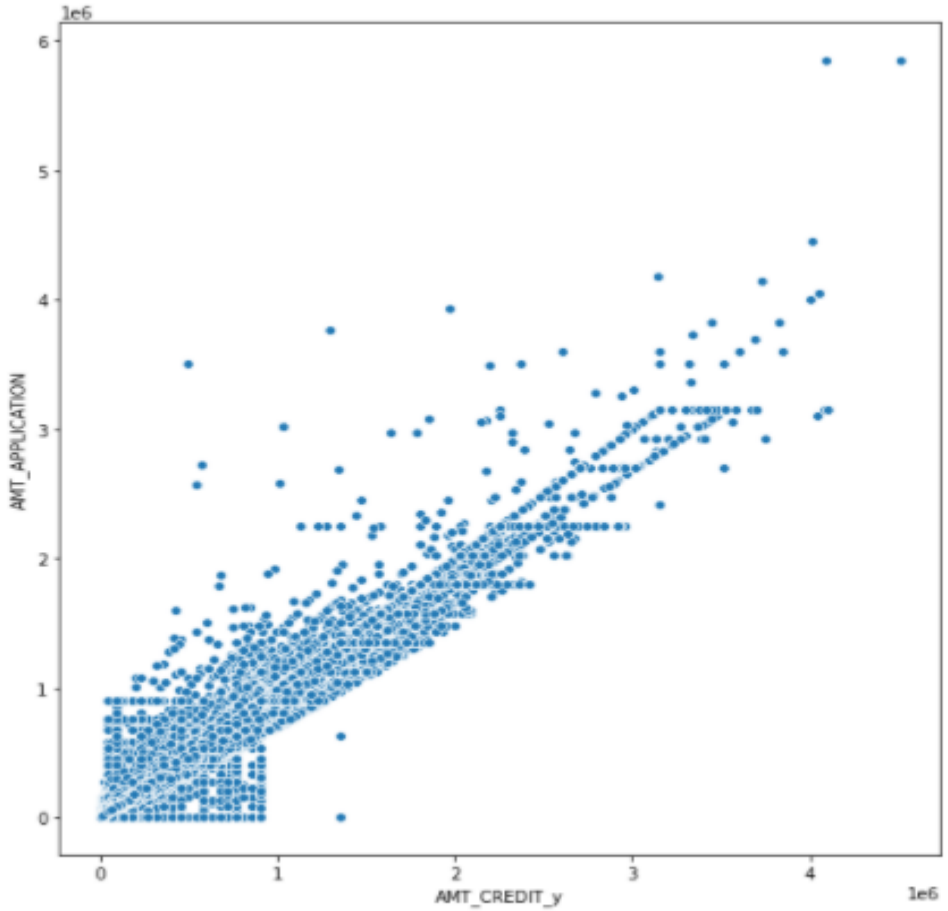


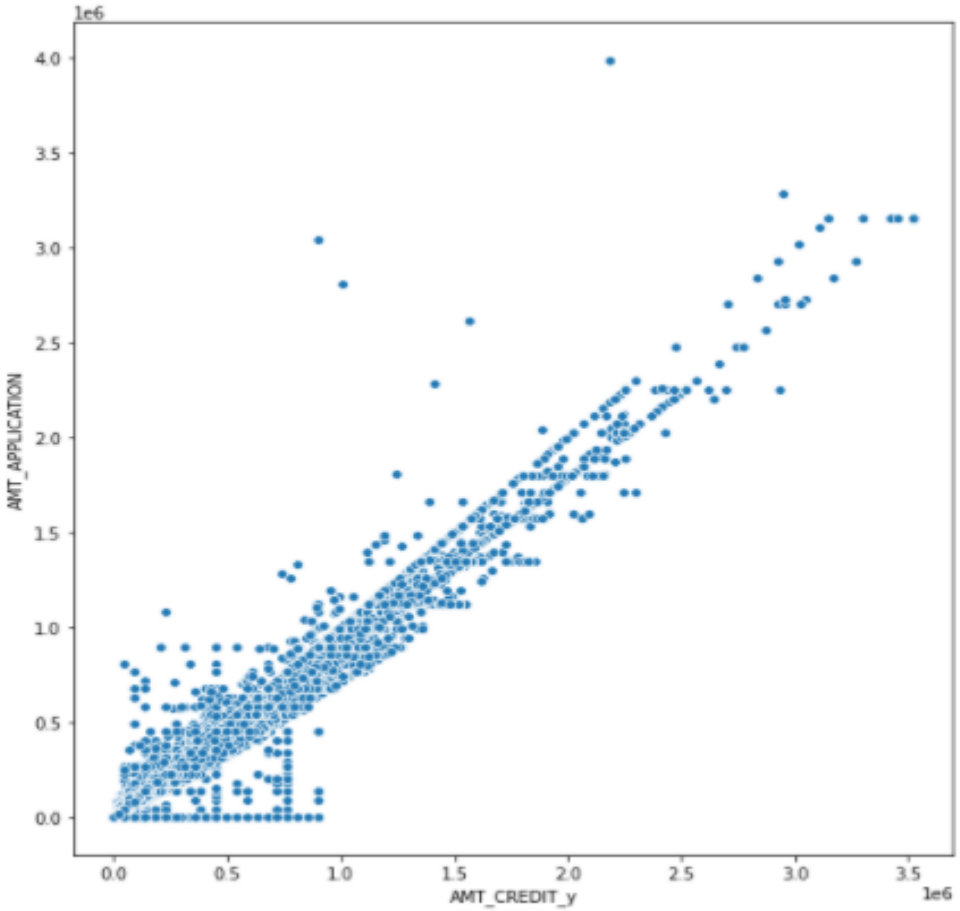


From above plots it appears that while middle yield group are more likely to be non-defaulters while high yield group appear to have a higher likelihood of being defaulters

Data Analysis – Combined Inferences -Bivariate Analysis

Comparing Target 0 i.e. situations where Clients have payment difficulties with Target 1 where there have been no defaults observed

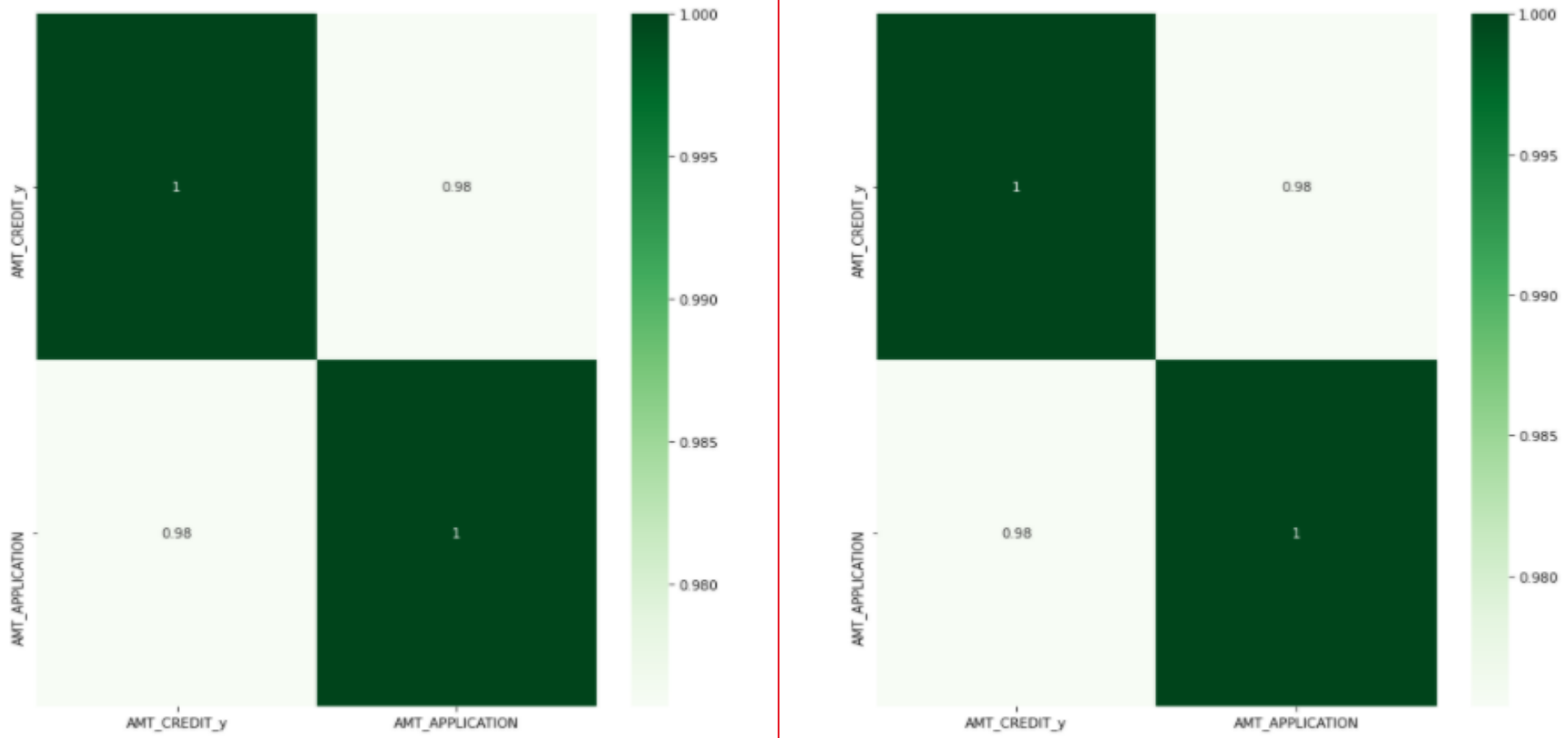




Above plot indicates a positive linear correlation between Amount of application and amount of credit especially post 1000000 credit

Data Analysis – Combined Inferences - Correlations

Comparing Target 0 i.e. situations where Clients have payment difficulties with Target 1 where there have been no defaults observed



The above plot indicates a strong correlation between Amount credit and amount application for both datasets

Data Analysis – Recommendations

A : Its seen that Married couples with no children seem to default more.

As a result for this category :

- If the couples have their own Realty, then this can be attached as collateral
- If the couples do not have their own Realty, then a Guarantor can be requested for the same.

B :


Thank You