# Exploratory Data Analysis

Loan Application Data of a Consumer Finance Company

# Exploratory Data Analysis of Loan Application data

**Problem Statement:**

Consumer finance company which is specialised in lending various types of loans to its customer would like to mitigate the risk associated while lending loans. It has two datasets namely Application data and Previous Application data. Using Exploratory Data Analysis we have to identify the key customer attributes and loan attributes which increases the tendency of default of loan.

**Risk:**

When a company receives a loan application, it has to decide for loan approval/rejection based on applicant's profile and his/her application attributes. Approving without rejecting a non-defaulter and rejecting a defaulter is a key risk to be handled else it could cause loss of business and loss of revenue.

**EDA Objectives**:

The objective is to mitigate the risk. To mitigate we have to identify the consumer attributes and loan attributes which influence the tendency of default and ensure that the consumers capable of repaying the loan are not rejected.

# Approach

We analyse application data containing customers data with payment difficulties and all other cases. we also analyse previous application data. To Identify consumer attributes and loan attributes which influence the tendency of default, we do following necessary steps on the data.

1. Data Inspection
2. Impute/Removing missing Values
3. Identifying Outliers
4. Univariate and Segmented Univariate Analysis
5. Bivariate Analysis
6. Multivariate Analysis

# Impute/Removing missing Values

```
# columns to be dropped with missing values greater than 47% in the application data
col_drop = list(AD_missing[AD_missing.values>47].index)
print(col_drop)
```

```
['COMMONAREA_MEDI', 'COMMONAREA_AVG', 'COMMONAREA_MODE', 'NONLIVINGAPARTMENTS_MODE', 'NONLIVINGAPARTMENTS_AVG', 'NONLIVINGAPARTMENTS_MEDI', 'FONDKAPREMONT_MODE', 'LIVINGAPARTMENTS_MODE', 'LIVINGAPARTMENTS_AVG', 'LIVINGAPARTMENTS_MEDI', 'FLOORSMIN_AVG', 'FLOORSMIN_MODE', 'FLOORSMIN_MEDI', 'YEARS_BUILD_MEDI', 'YEARS_BUILD_MODE', 'YEARS_BUILD_AVG', 'OWN_CAR_AGE', 'LANDAREA_MEDI', 'LANDAREA_MODE', 'LANDAREA_AVG', 'BASEMENTAREA_MEDI', 'BASEMENTAREA_AVG', 'BASEMENTAREA_MODE', 'EXT_SOURCE_1', 'NONLIVINGAREA_MODE', 'NONLIVINGAREA_AVG', 'NONLIVINGAREA_MEDI', 'ELEVATORS_MEDI', 'ELEVATORS_AVG', 'ELEVATORS_MODE', 'WALLSMATERIAL_MODE', 'APARTMENTS_MEDI', 'APARTMENTS_AVG', 'APARTMENTS_MODE', 'ENTRANCES_MEDI', 'ENTRANCES_AVG', 'ENTRANCES_MODE', 'LIVINGAREA_AVG', 'LIVINGAREA_MODE', 'LIVINGAREA_MEDI', 'HOUSETYPE_MODE', 'FLOORSMAX_MODE', 'FLOORSMAX_MEDI', 'FLOORSMAX_AVG', 'YEARS_BEGINEXPLUATATION_MODE', 'YEARS_BEGINEXPLUATATION_MEDI', 'YEARS_BEGINEXPLUATATION_AVG', 'TOTALAREA_MODE', 'EMERGENCYSTATE_MODE']
```

# Impute/Removing missing Values

1. Handling categorical column with missing values >13 % we impute them as 'Others’

```
# handling categorical column with missing values >13 % we impute them as 'Others'
AD.OCCUPATION_TYPE = AD.OCCUPATION_TYPE.fillna('Others')
```

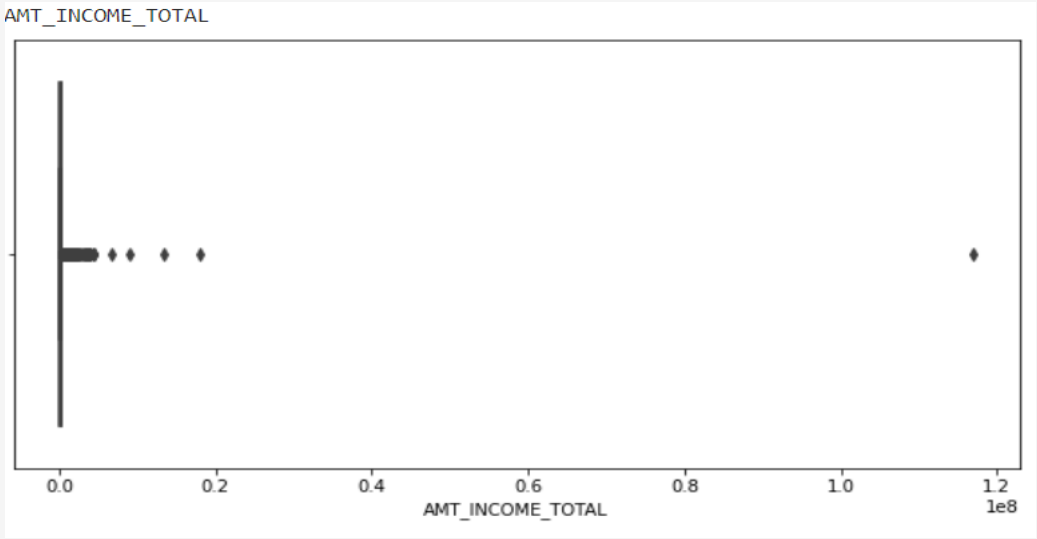
2. Handling categorical columns with missing values <13 % we impute them with mode

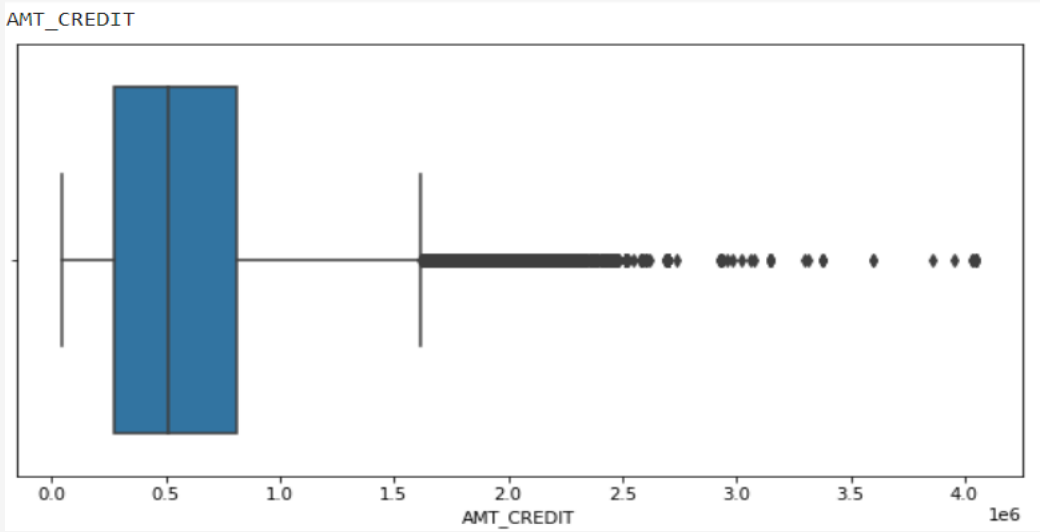
```
# handling categorical columns with missing values <13 % we impute them with mode
AD.NAME_TYPE_SUITE= AD.NAME_TYPE_SUITE.fillna(AD.NAME_TYPE_SUITE.mode()[0])
```

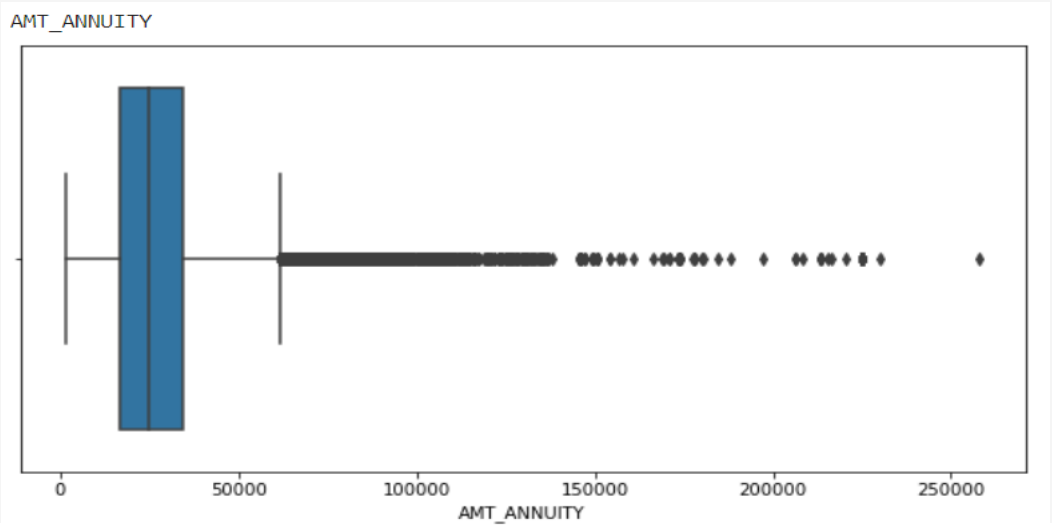
3. Percentage of XNA is CODE_GENDER is 0.0013 we can impute the XNA values with mode as it is categorical column

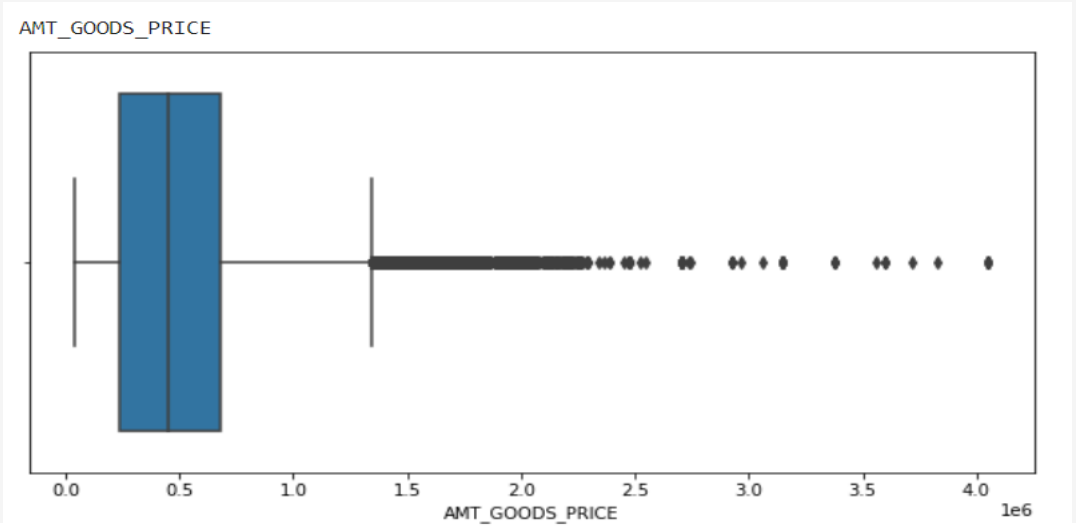
```
# percentage of XNA is CODE_GENDER is 0.0013 we can impute the XNA values with mode as it is categorical column
AD.CODE_GENDER.replace('XNA',AD.CODE_GENDER.mode()[0],inplace = True)
```

# Identifying Outliers



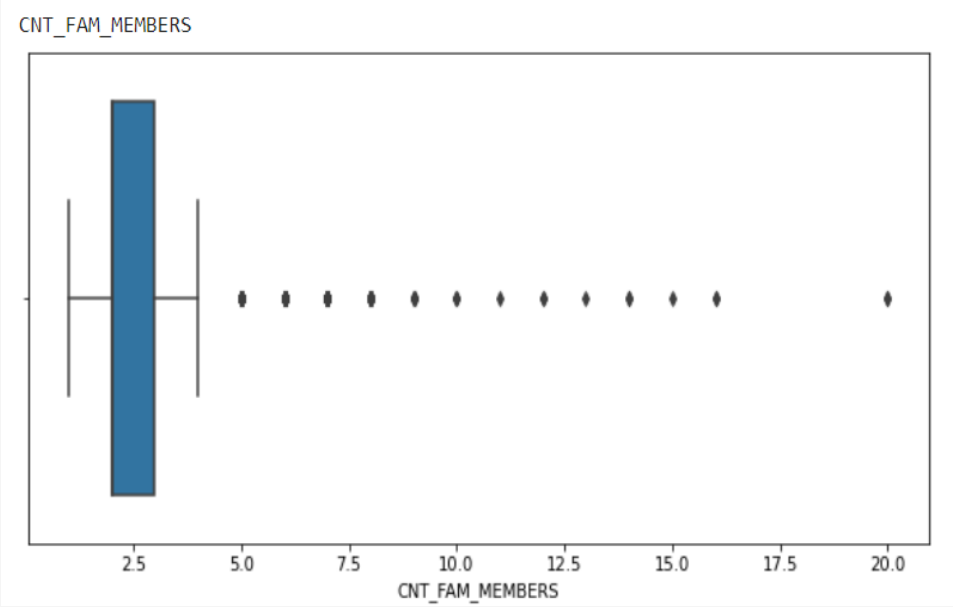



# Identifying Outliers (Continued…)

1. **What is an Outlier :** In the above 5 box plots we saw that beyond the upper bound whiskers there are values and these are data points which are beyond or not in the normal range as other data points




**To handle outliers we can use several ways – Treatment not done in my analysis**

1. Imputation of the outlier. These outliers may be due to errors or veryhigh values which may not occur in general we can treat them similarly like how we treat missing values.

2. Delete the outlier

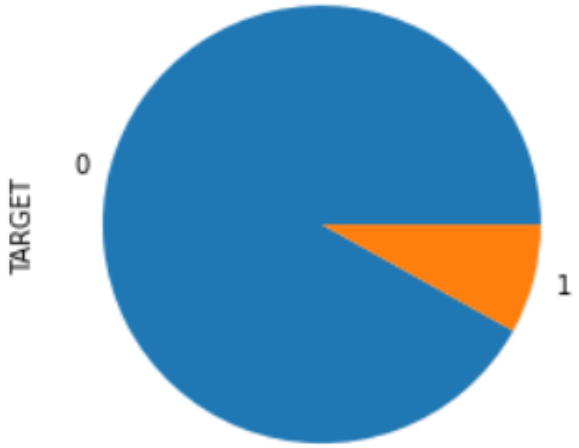
3. Using binning technique we can group those outliers as high values but can keep in our data set

4. We can put a cap on the data point which is almost like an outlier but by including that cap point so that we are not skewing the output but essentially omitting outlier.

5. Also we can use box plots and quantiles for studying the outliers and determine if 90%ile or 95%ile or 99%ile data is sufficient for our analysis and omit the outliers.

# Data Imbalance

[73]:
```
# Data imbalance - visualization of TARGET Column
AD.TARGET.value_counts(normalize = True).plot.pie()
plt.show()
```




```
# Data imbalance in percentage of TARGET Column
AD.TARGET.value_counts(normalize = True)
```

```
0    0.919271
1    0.080729
Name: TARGET, dtype: float64
```

```
AD_0.shape[0]/AD_1.shape[0]
```

```
11.387150050352467
```

**Ratio of Data Imbalance**

**Analysis Result : There is 1 applicant with payment difficulties for every 11 applicants**

# Univariate Analysis - Categorical Columns for TARGET = 0 and TARGET =1

1. 30-40 years Age group have the highest number of loan Applicants with about 27%
2. 1,00,000-2,00,000 income group has the highest number of loan Applicants
3. There is 1 defaulter for every 11.38 loan applicants which amounts to 8.07% of the total loan applicants.
4. ‘CODE_GENDER,’ 'NAME_EDUCATION_TYPE',’ NAME_FAMILY_STATUS’,’ NAME_HOUSING_TYPE’,’ OCCUPATION_TYPE’ are the categorical columns chosen for univariate analysis.
5. About 65% of the loan applicants are females
6. 71% of loan applicants have education types as Secondary /Secondary Special and 24% of the loan applicants have Higher education. Lower the Level of education of the client achieved the higher the default chances.
7. Married category has 63% of the loan applicants.
8. 88% of the housing situation of loan applicants is Housing/Apartments.
9. In OCCUPATION_TYPE missed values termed as others are more in number of applicants. Labourers, Drivers, cooking staff, cleaning staff are more defaulters than compared to Managers, High skill tech staff, core staff.

**Recommendations:**

1. Target more customers who has the Occupation types as Managers, High skill tech staff, core staff.
2. Target more customers who has the education type as Highest education

# Univariate Analysis - Continuous Columns for TARGET = 0 and TARGET =1

1. Below Numerical_columns have outliers in the values 'AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY','AMT_GOODS_PRICE','AGE','WORK_EXPERIENCE','CNT_FAM_MEMBERS' we can observe them by using boxplots and quantile function.

2. 99 percentile of the loan applicants has the income of 4,72,500. Also clients Income is driving the defaults ,lower the income higher the default.

3. 80 percentile of the loan applicants asked for the credit of less than 9,00,000

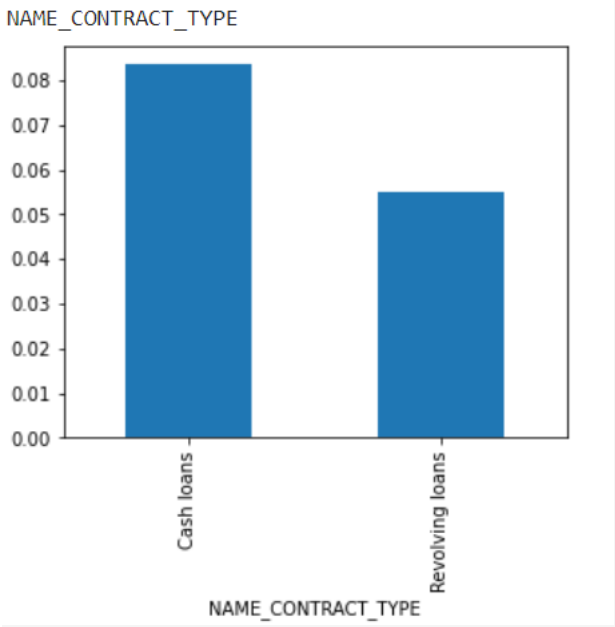
4. 80 percentile of the loan applicants has work experience of less than 25 years

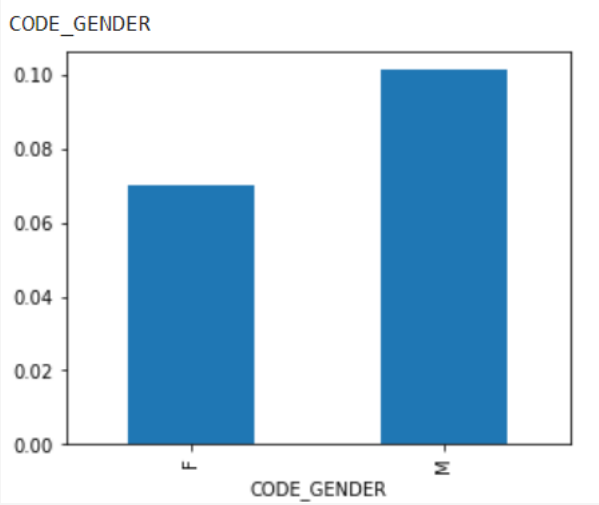
# ## findings from CORRELATION between Numerical Columns : Application Data

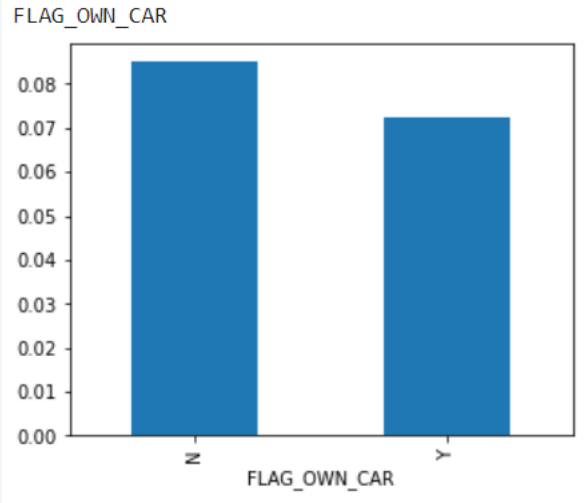
1. There is high positive correlation between AMT_CREDIT and AMT_GOODS_PRICE
2. There is high positive correlation between AMT_GOODS_PRICE and AMT_ANNUITY
3. There is high positive correlation between AMT_ANNUITY and AMT_CREDIT
4. There is high positive correlation between CNT_FAM_MEMBERS and CNT_CHILDREN
5. There is positive correlation between AGE and WORK_EXPERIENCE

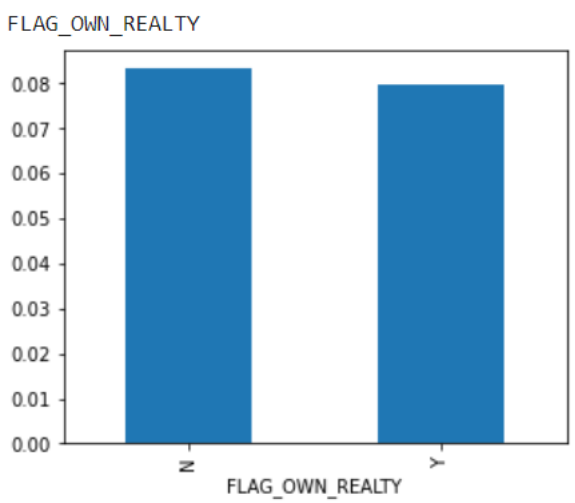
```
sns.heatmap(AD_CORR,annot = True,cmap='RdYlGn',center = 0.080729)
plt.show()
```

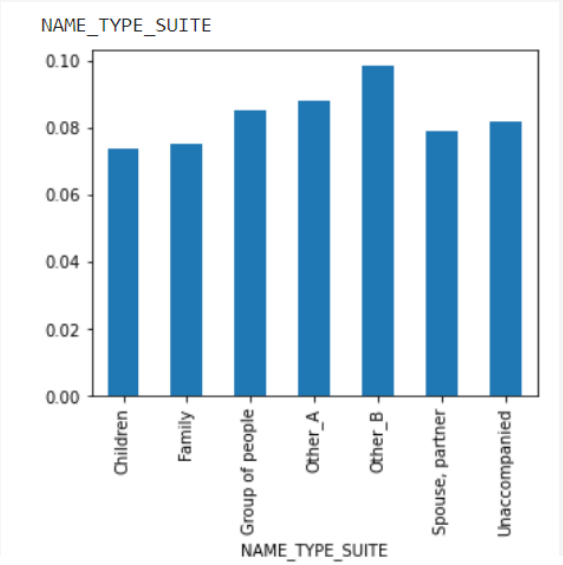
# Bivariate Analysis: Application Data

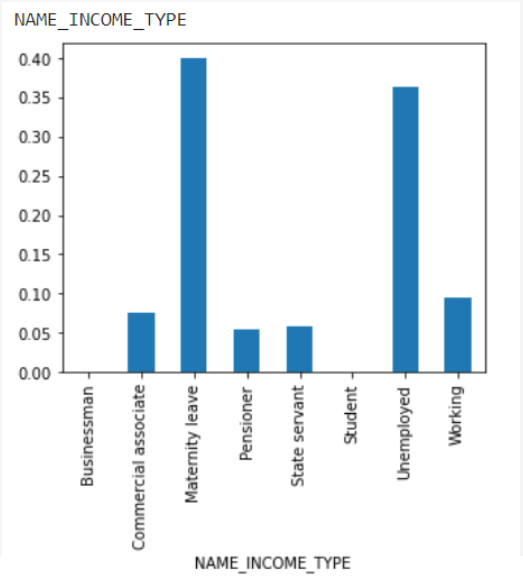


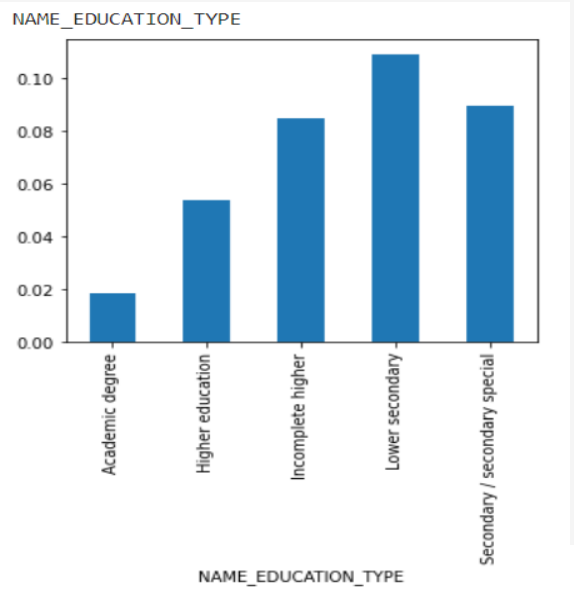



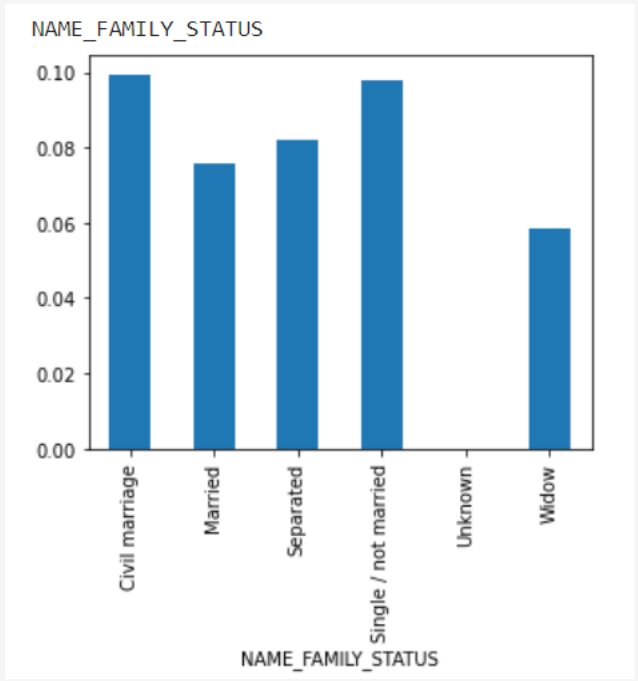



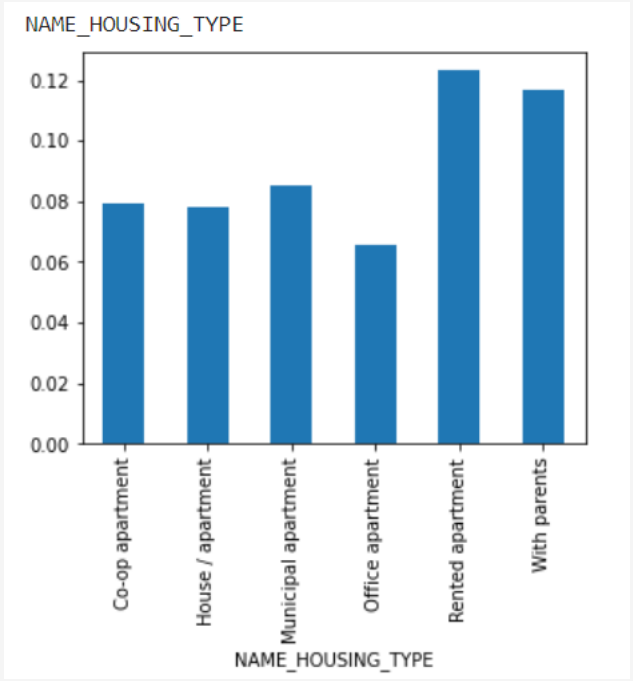



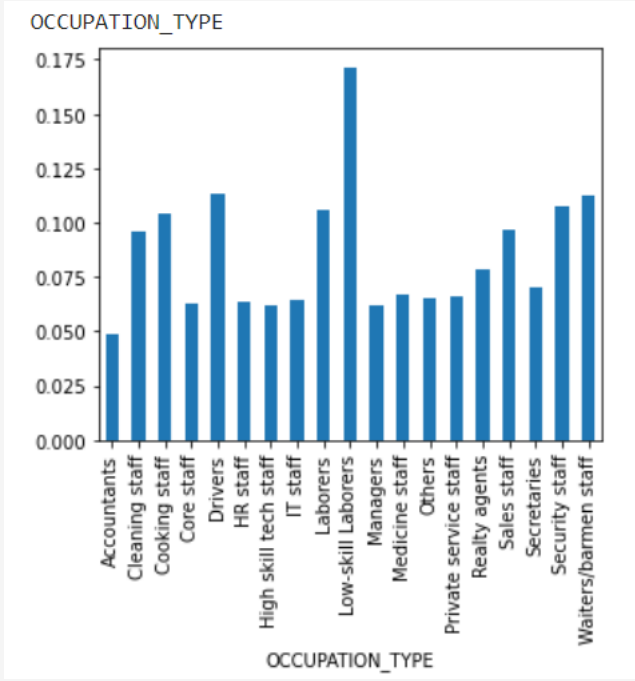



# Bivariate Analysis: Application Data



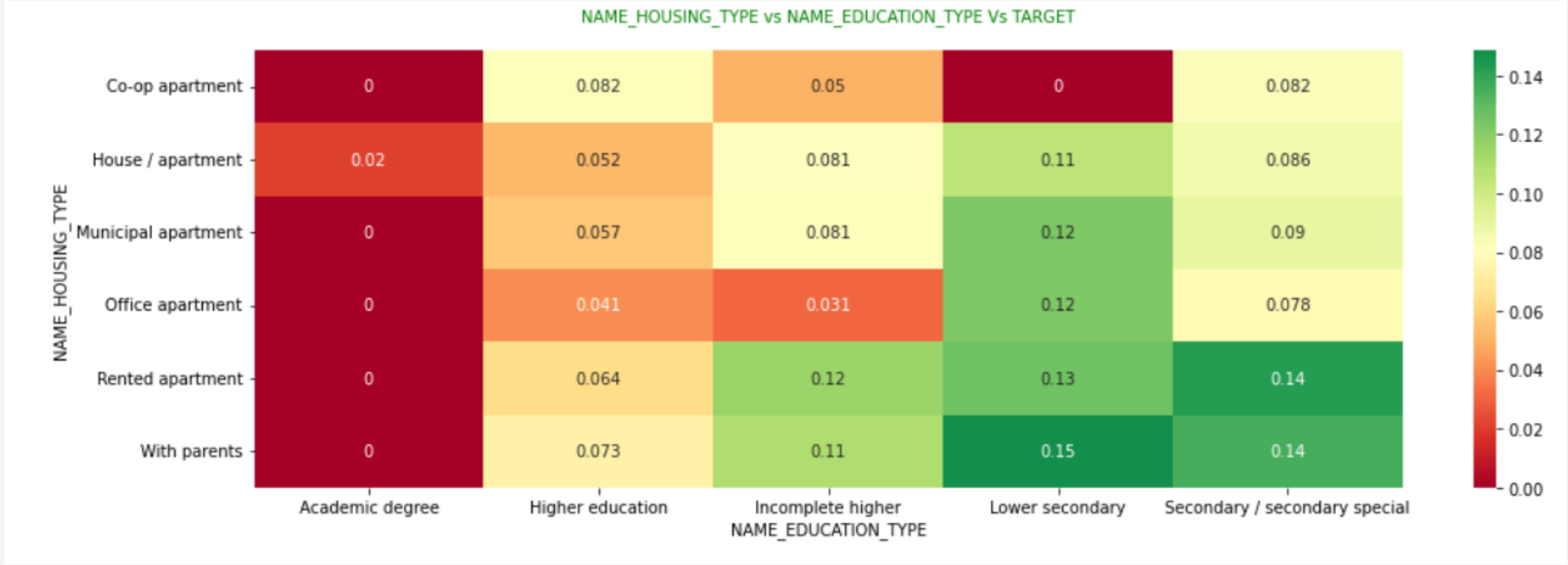



# Bivariate Analysis: Application Data

- Cash loans tend to default more compared to revolving loans
- Males tend to default more compared to females
- When loan applicants are accompanied by Other_B and Other_A are more likely to default.
- When clients who are in maternity leave and those who are unemployed are more likely to default
- Lower secondary and secondary/secondary special  education types are more likely to default
- Clients in rented apartments are more likely to default
- Low skill labourers, labourers, waiters/barmen staff ,drivers are more likely to default
- Here are the top 3 organization types where clients are working tend to default more and those are Transport type 3 , Industry type 13, Industry type 8

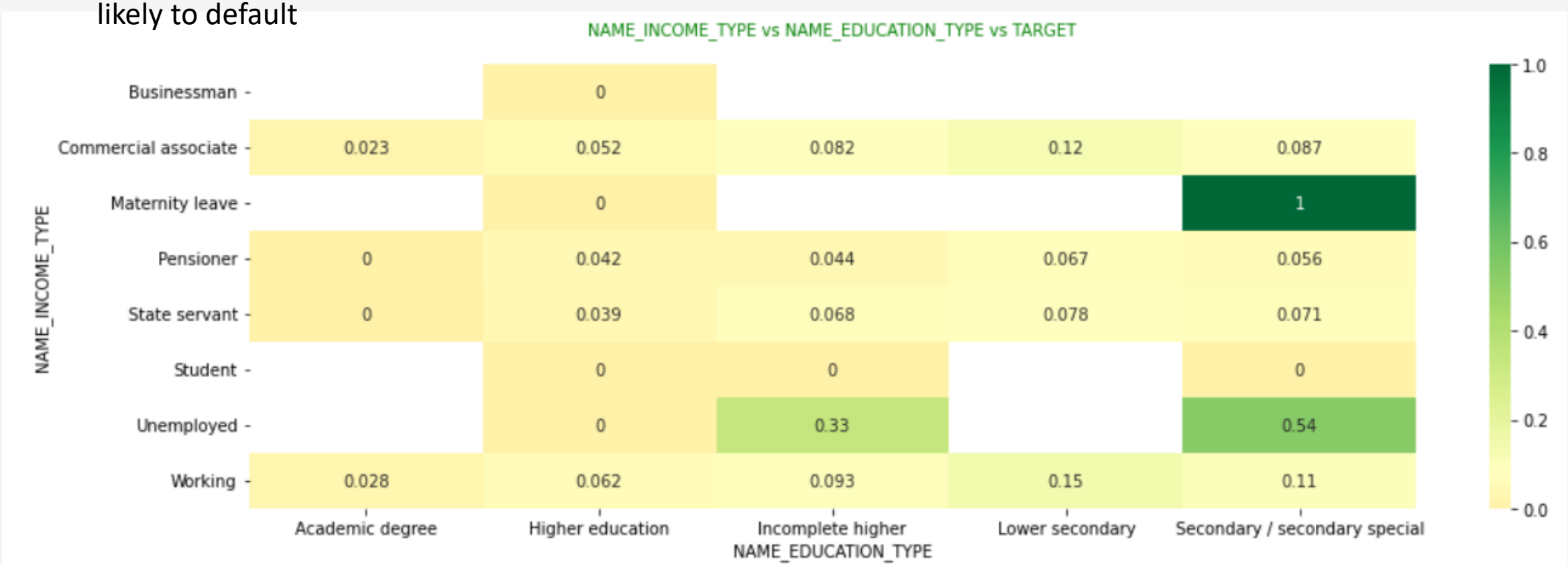
# Multi variate Analysis:  Application Data

1. Clients who are on maternity leave and having education type as secondary/secondary special are more likely to default
2. Clients who are unemployed and education type as incomplete higher and secondary/secondary special are more likely to default



# Multi variate Analysis:  Application Data

1. Clients living with parents or in rented apartment or office apartment or municipal apartment or house/apartment having education type as lower secondary are more likely to default
2. Clients living with parents or rented apartments having education type as secondary/secondary special are more likely to default



# Exploratory Data Analysis

Previous Loan Application Data

# Approach

We analyse Previous application data containing loan attributes. We merge the application data and previous application data to determine how consumer attributes and loan attributes influence the tendency of default, we do following  necessary steps on the data.

1. Data Inspection
2. Impute/Removing missing Values
3. Identifying Outliers
4. Univariate and Segmented Univariate Analysis
5. Bivariate Analysis
6. Multivariate Analysis

# Impute/Removing missing Values

```
# columns to be dropped with missing values greater than 50% in the previous application dataset
col_drop = list(PA_missing[PA_missing.values>50].index)
col_drop
```

```
['RATE_INTEREST_PRIVILEGED',
 'RATE_INTEREST_PRIMARY',
 'AMT_DOWN_PAYMENT',
 'RATE_DOWN_PAYMENT']
```

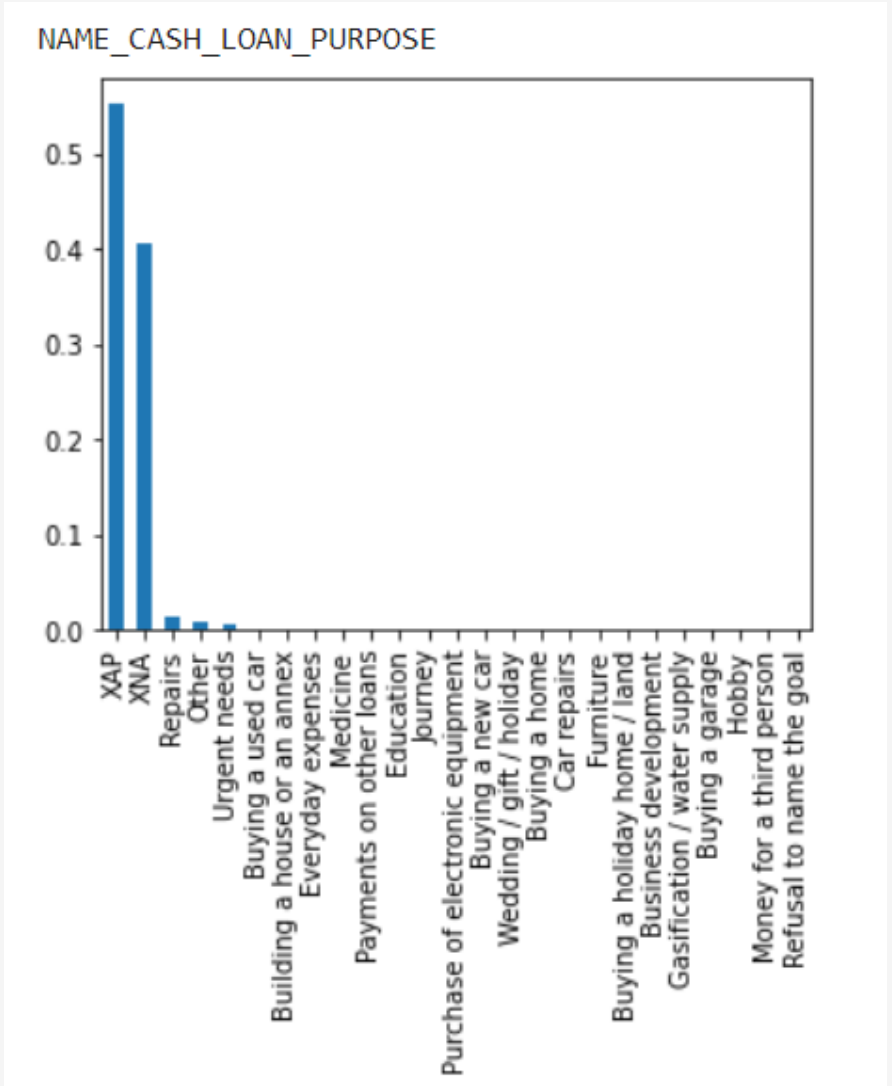
## Handling Missing values in Categorical Columns

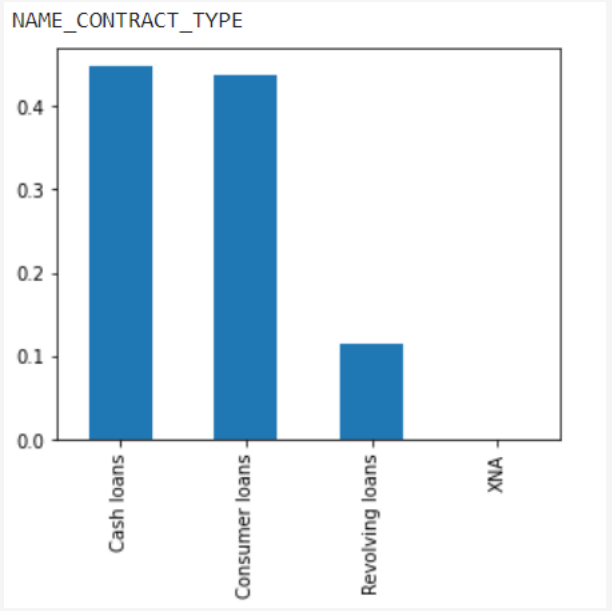
```
# for catgeorical columns we impute missing values as 'missing' if percentage of missing values is >15%
#and clearly in NAME_TYPE_SUITE has 49%
PA.NAME_TYPE_SUITE = PA.NAME_TYPE_SUITE.fillna('missing')
# for catgeorical columns we impute missing values with median if percentage of missing values is <15%
#and PRODUCT_COMBINATION has 0.02%
PA.PRODUCT_COMBINATION = PA.PRODUCT_COMBINATION.fillna(PA.PRODUCT_COMBINATION.mode()[0])
```

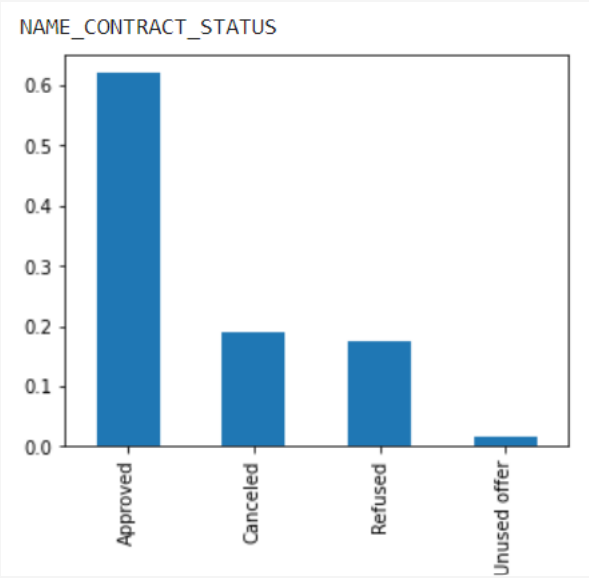
# Impute/Removing missing Values

```
# Imputing the missing values in numerical columns with median as there are outliers
# This we can observe the mean and median values as we saw above.
for x in PA_missing_col_list:
    PA[x] = PA[x].fillna(PA[x].median())
```

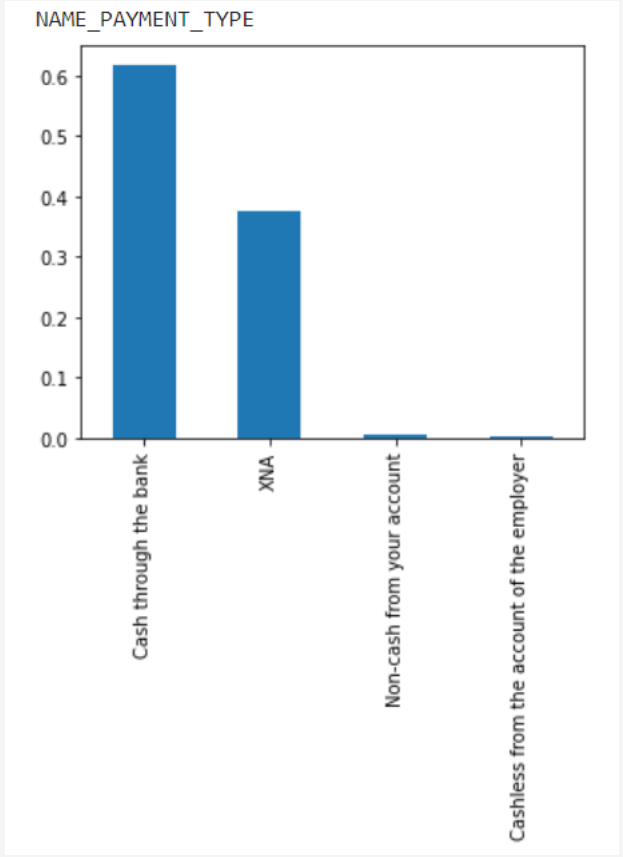
# Univariate Analysis – categorical columns

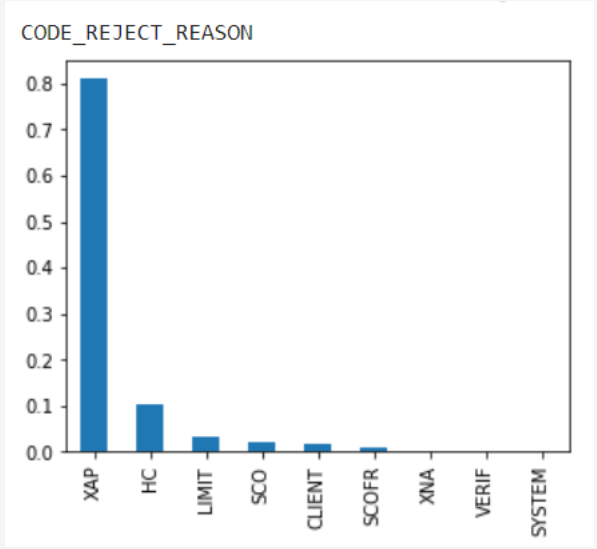


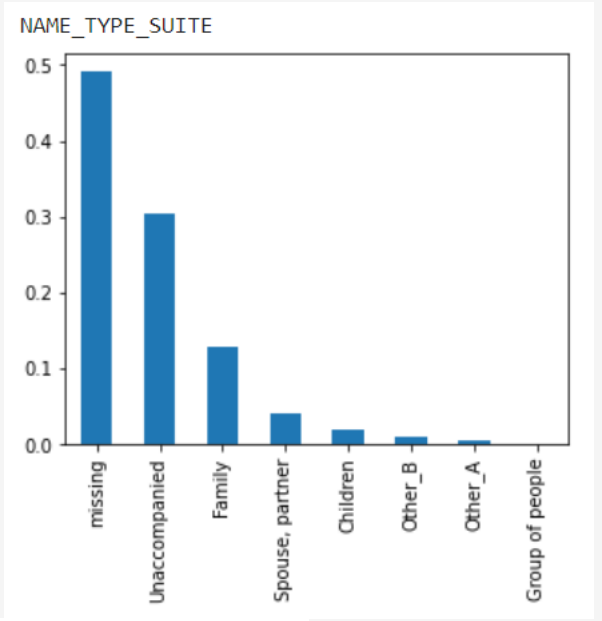



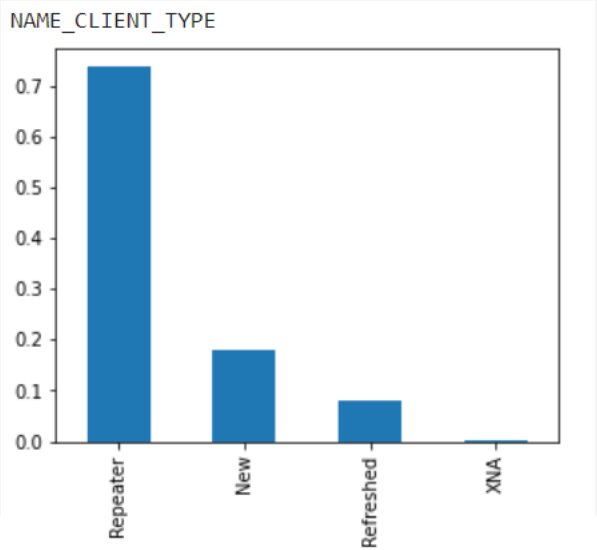



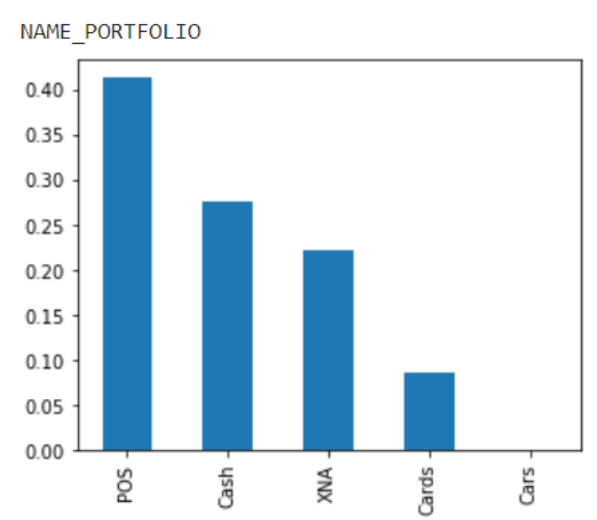
# Univariate Analysis – Catgeorical Columns



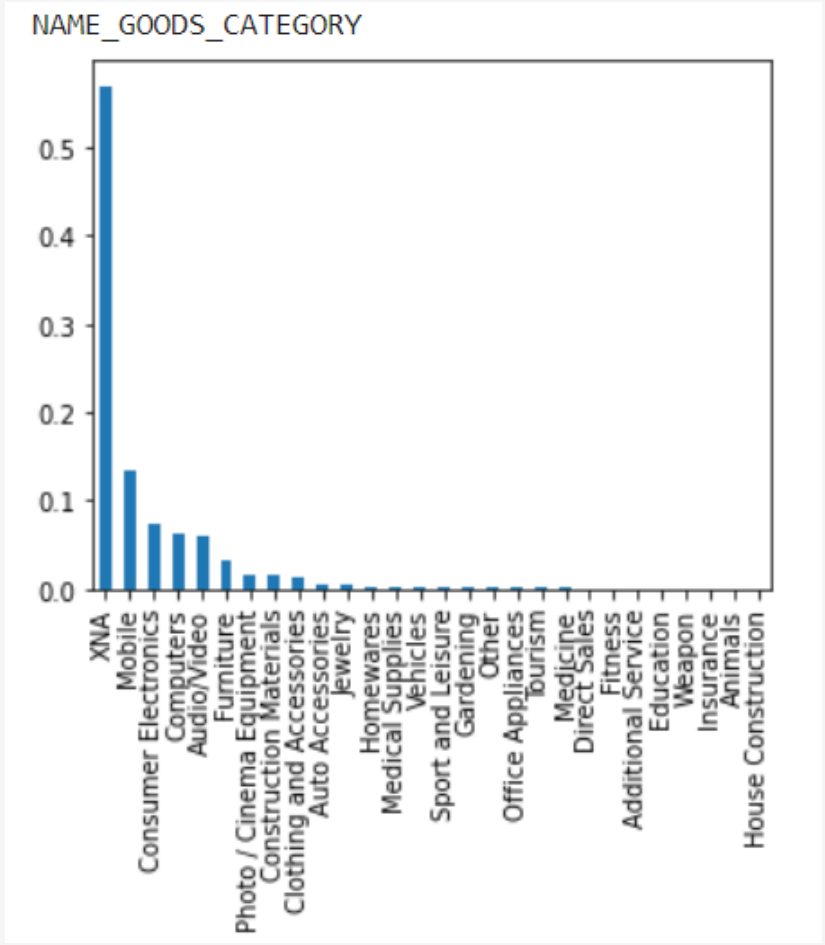



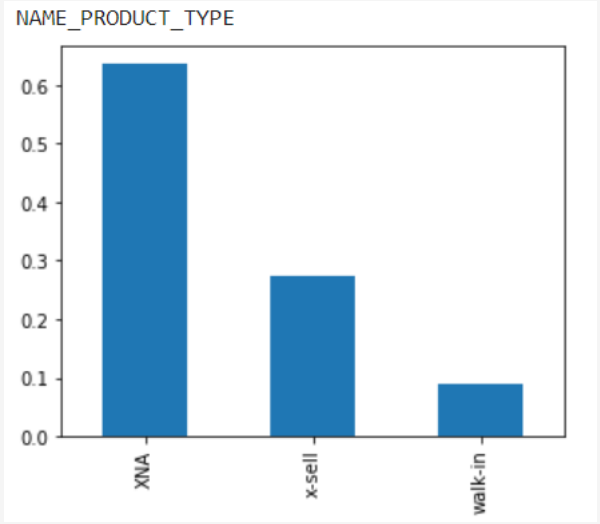



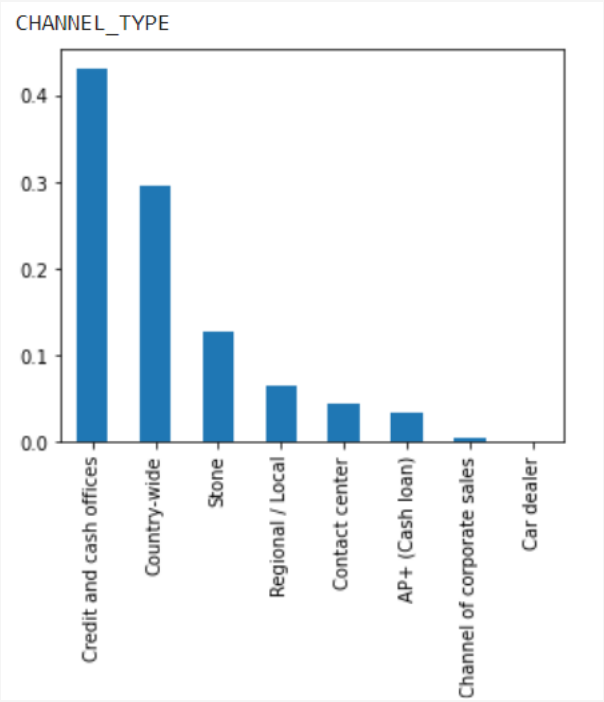



# Univariate Analysis – Catgeorical Columns

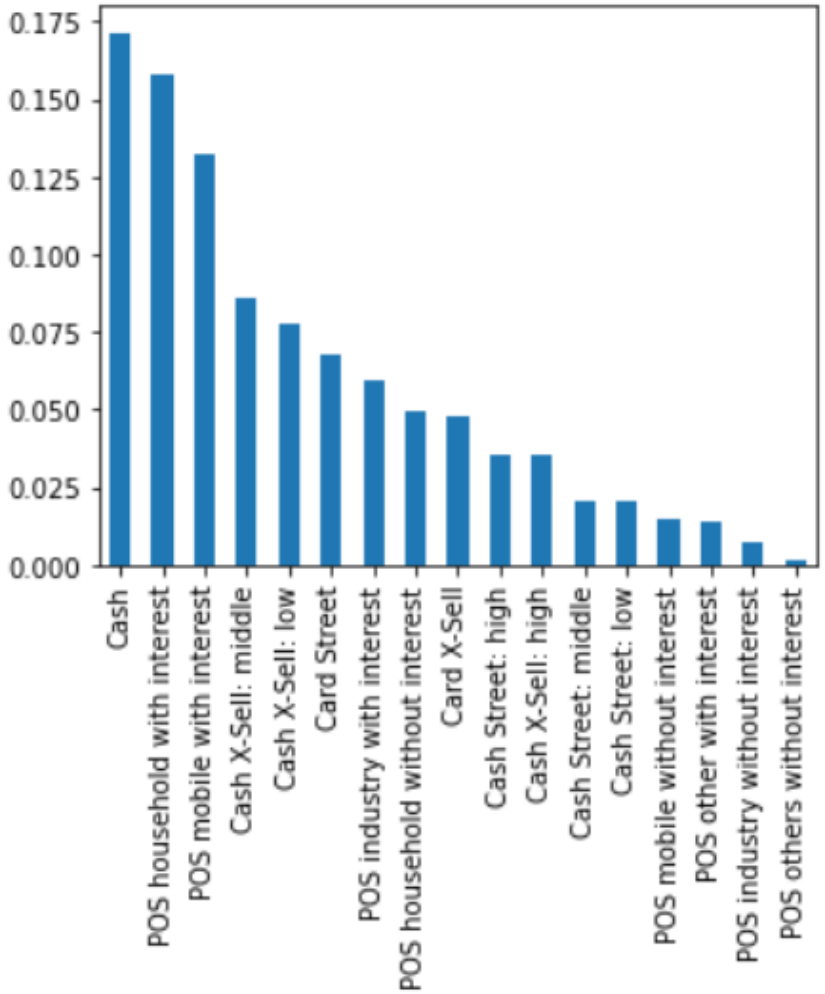


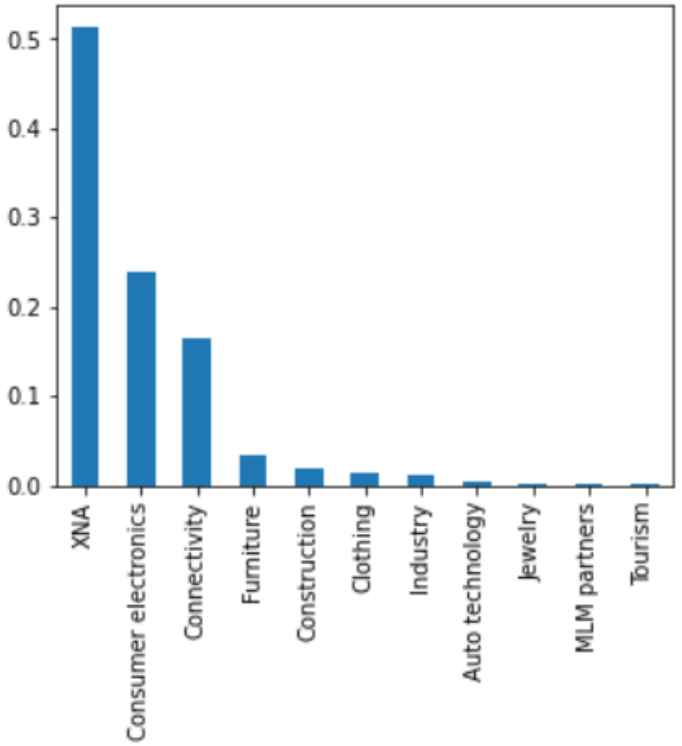



# Univariate Analysis – Catgeorical Columns

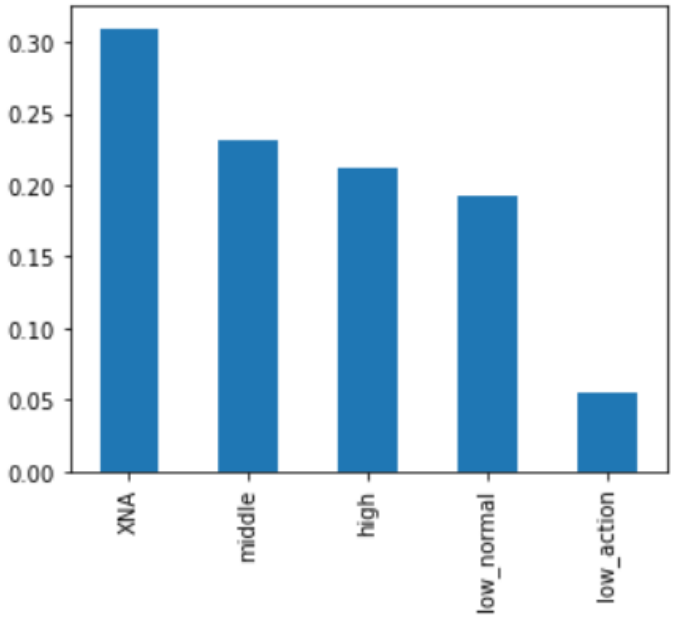
PRODUCT_COMBINATION




NAME_SELLER_INDUSTRY




NAME_YIELD_GROUP



### Univariate Analysis of categorical columns of the Previous Application:

1. Loans are of type cash loans and consumer loans
2. During the week days loans were applied more compared to weekends and Sunday is further less
3. The purpose of the loan was mentioned as XAP and XNA for majority of the cases so we do not know the reasons what is XAP and XNA and rest of the applications the reasons were mentioned.
4. About 62% of the previous loans were approved, 17.4% were refused, 18.9% loans were canceled,1.5% previous applicants have Unused offer
5. Around 60% of previous applicants that chose Cash through the bank as mode of payment.
6. The reason for rejection of loans is maximum with reason as XAP and we do not know what XAP means.
7. For Majority of the applicants the value is “missing” when we are answering who accompanied client when applying for the previous application
8. 73.71% of the previous applicants are repeaters,18.04% are new applicants
9. Kind of goods client apply loan for in the previous application are mentioned as XNA
10. Highest previous application are through POS
11. Previous application is of product type as XNA
12. Majority of the previous applicants came from Credit and Cash offices Chanel type
13. Majority of the seller industry are of type XNA
14. Grouped interest rate as XNA
15. Cash as the highest product combination

## Bivariate Analysis of Previous Application Data

1. Loans are of type cash loans and consumer loans
2. During the week days loans were applied more compared to weekends and Sunday is further less
3. The purpose of the loan was mentioned as XAP and XNA for majority of the cases so we do not know the reasons what is XAP and XNA and rest of the applications the reasons were mentioned.
4. About 62% of the previous loans were approved, 17.4% were refused, 18.9% loans were canceled,1.5% previous applicants have Unused offer
5. Around 60% of previous applicants that chose Cash through the bank as mode of payment.
6. The reason for rejection of loans is maximum with reason as XAP and we do not know what XAP means.
7. For Majority of the applicants the value is “missing” when we are answering who accompanied client when applying for the previous application
8. 73.71% of the previous applicants are repeaters,18.04% are new applicants
9. Kind of goods client apply loan for in the previous application are mentioned as XNA
10. Highest previous application are through POS
11. Previous application is of product type as XNA
12. Majority of the previous applicants came from Credit and Cash offices Chanel type
13. Majority of the seller industry are of type XNA
14. Grouped interest rate as XNA
15. Cash as the highest product combination

## ## CORRELATION between Numerical Columns of previous application

1. There is high positive corelation between AMT_ANNUITY and AMT_APPLICATION
2. There is high positive corelation between AMT_ANNUITY and AMT_CREDIT
3. There is high positive corelation between AMT_ANNUITY and AMT_GOODS_PRICE
4. There is positive corelation between CNT_PAYMENT and AMT_ANNUITY
5. There is positive corelation between CNT_PAYMENT and APPLICATION
6. There is positive corelation between CNT_PAYMENT and CREDIT
7. There is positive corelation between CNT_PAYMENT and AMT_GOODS_PRICE
8. There is high positive corelation between DAYS_FIRST_DRAWING and DAYS_LAST_DUE_1$^{ST}$_VERSION
9. There is high positive corelation between DAYS_TERMINATION and DAYS_LAST_DUE

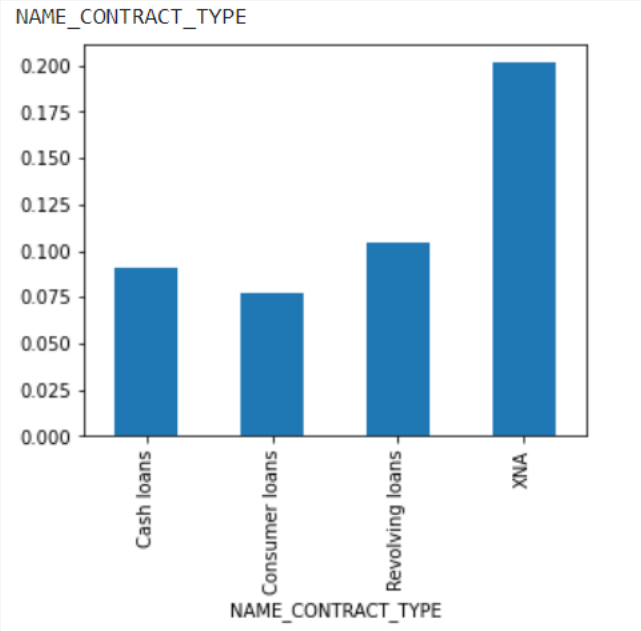
# ## Bi variate Analysis of the merged Data

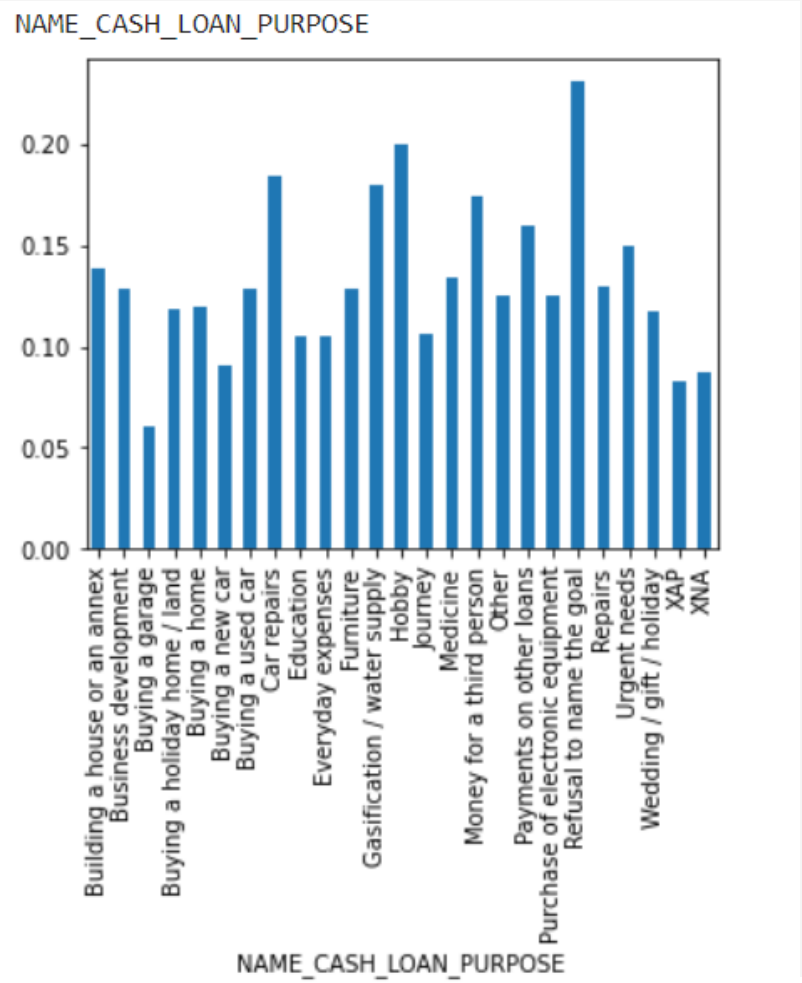
**The defaults are more for following type of previous applications.**

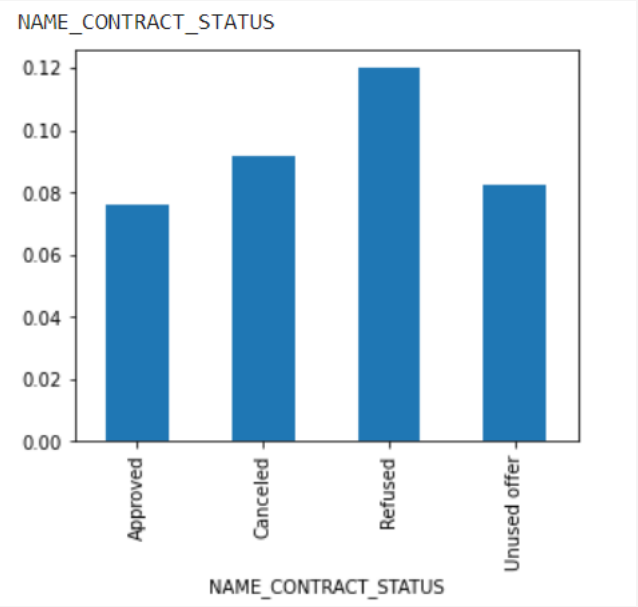
1. In the Previous Application XNA type of Loans are highest.
2. Many of the previous applicants Refused to name of the the goal when asked for purpose of the loan.
3. Many of the previous applicants were refused the loan
4. Payment Type is highest for XNA.
5. SCOFR is the Reason for rejections for many of the previous application.
6. Missing category is highest for NAME_TYPE_SUITE.
7. Client Type of XNA is highest for previous applicants
8. Goods category as Insurance is highest for previous applicants.
9. Cards was highest for the NAME_PORTFOLIO
10. Walkin is highest for NAME_PRODUCT_TYPE
11. On AP+ cash loan Chanel type we acquired the client
12. Auto Technology as the seller industry is highest
13. Cash street : middle is highest for product combination

# ## Bi variate Analysis of the merged Data

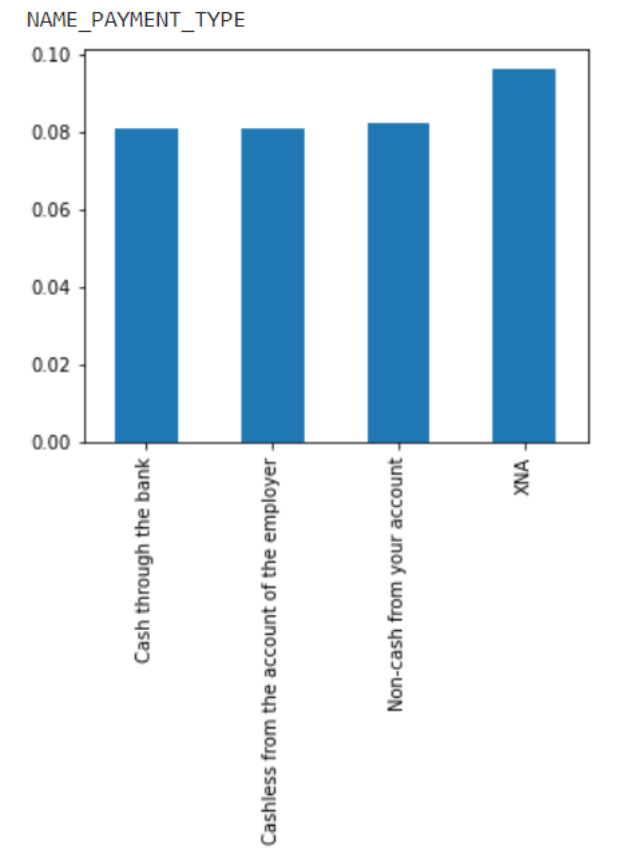
### Categorical columns(ALL) of Previous Application vs Categorical(TARGET column)

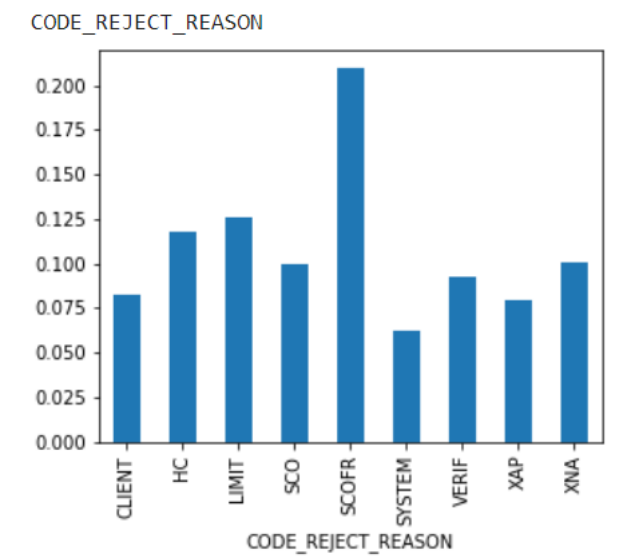


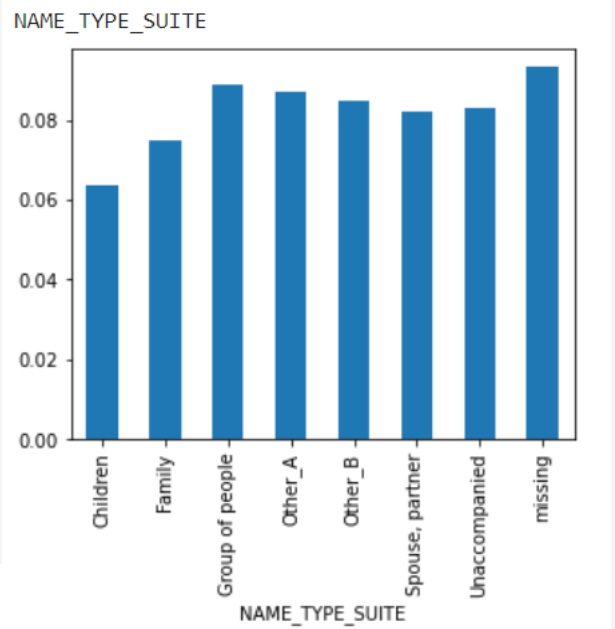



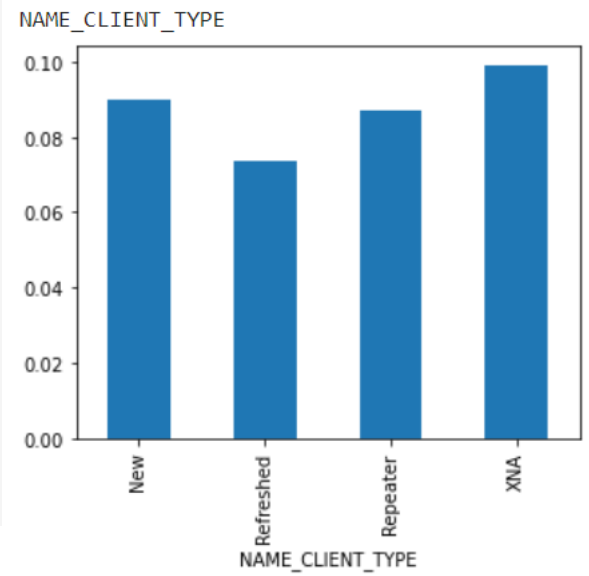



# ## Bi variate Analysis of the merged Data

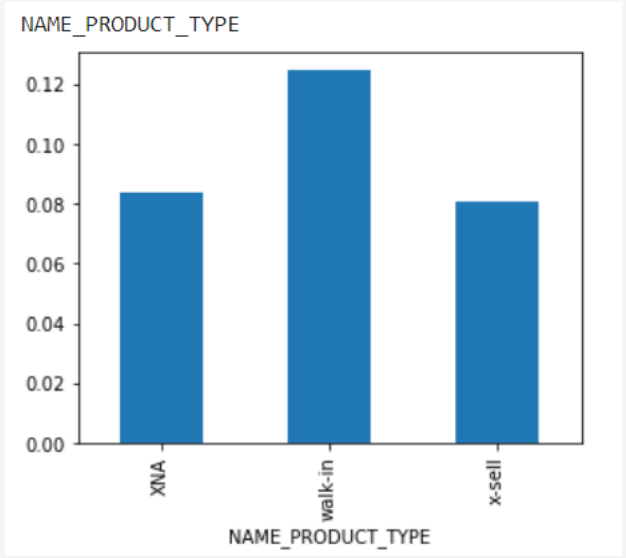


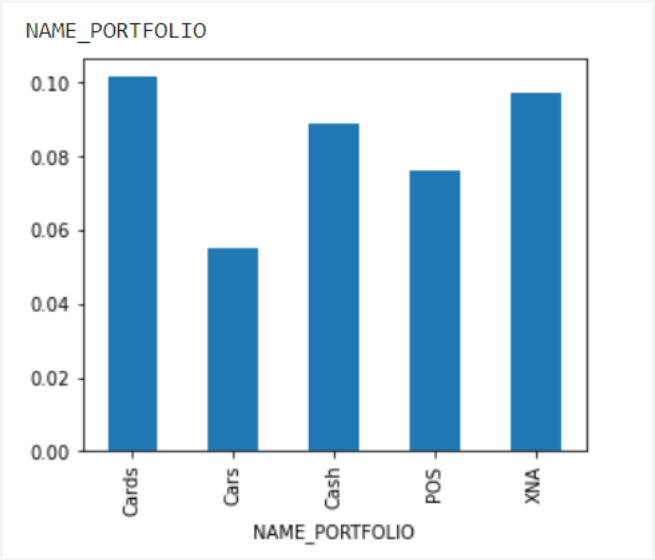



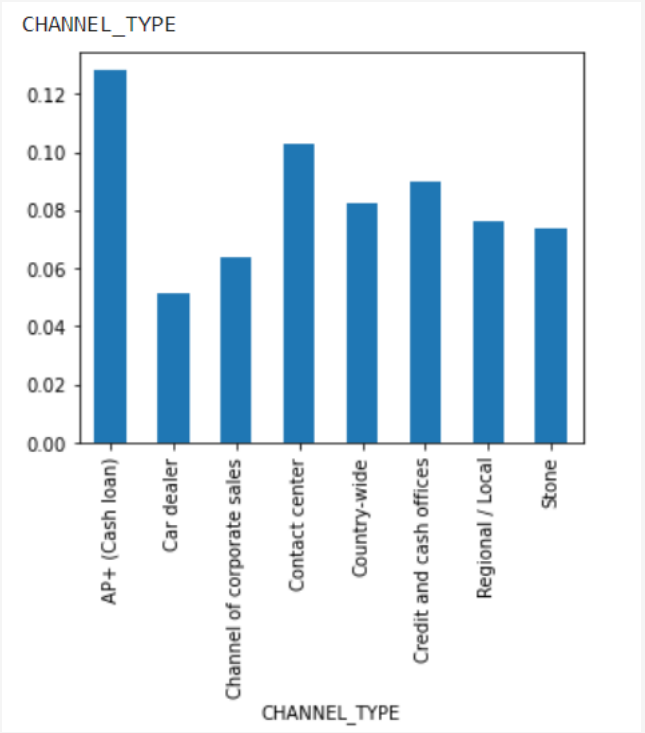



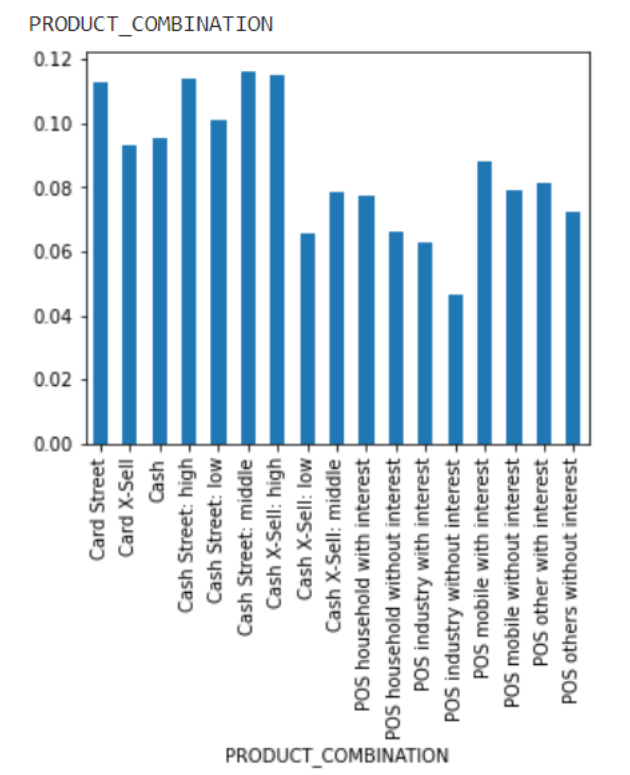
# ## Bi variate Analysis of the merged Data

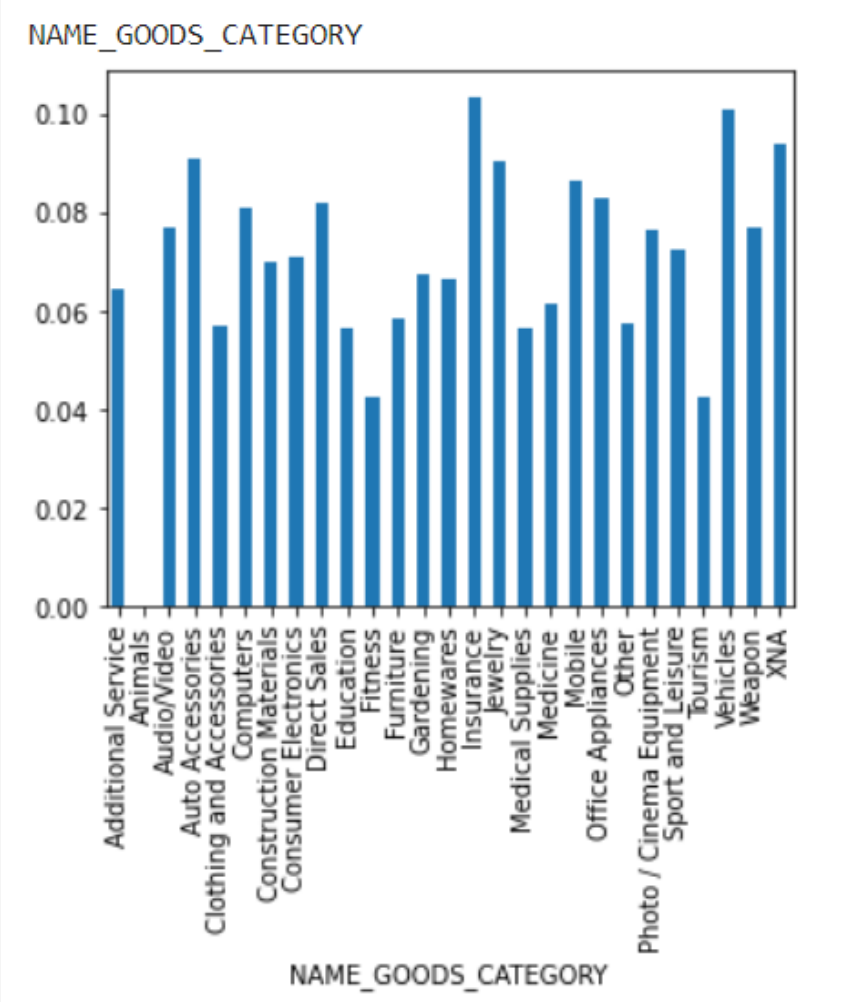


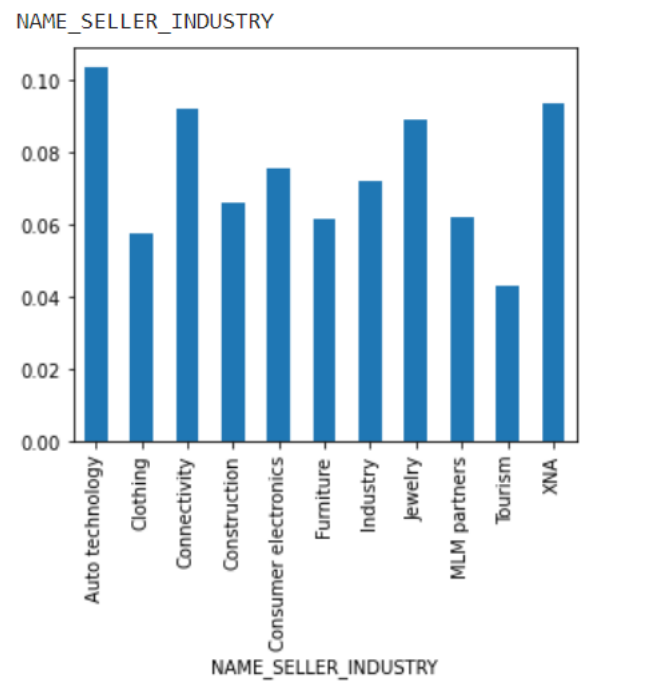



# ## Bi variate Analysis of the merged Data

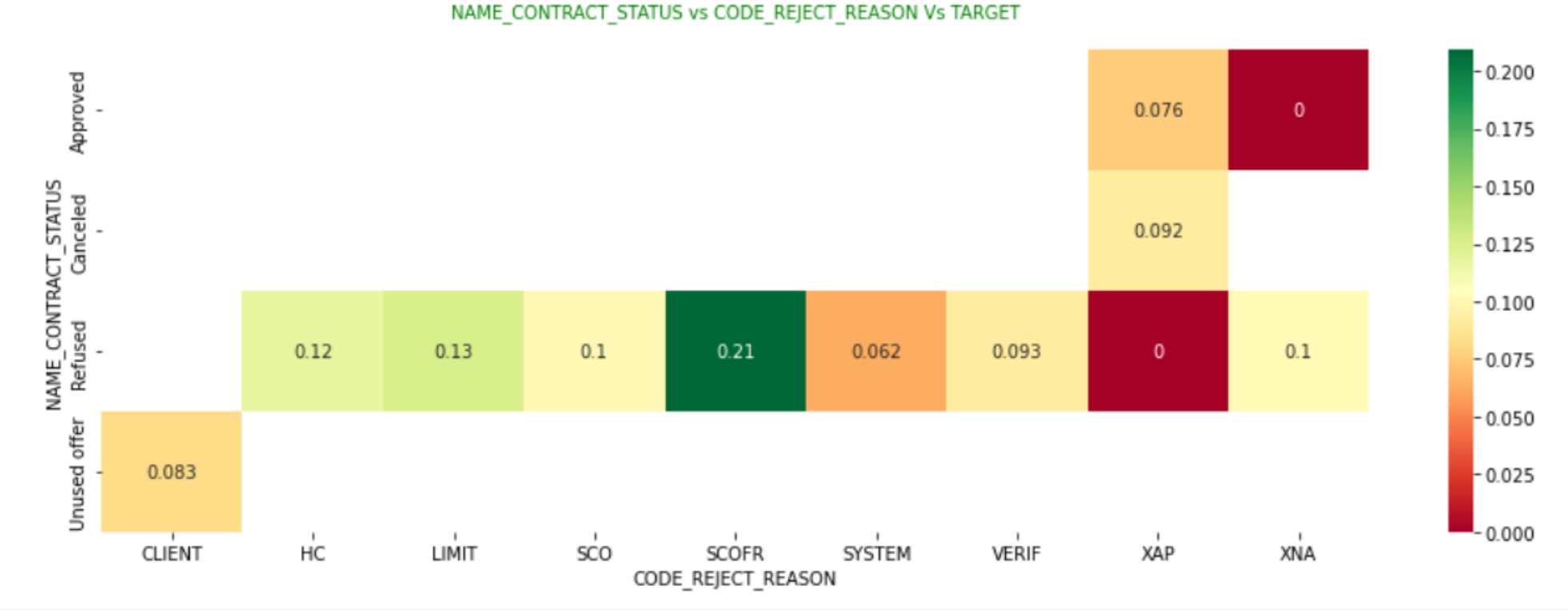


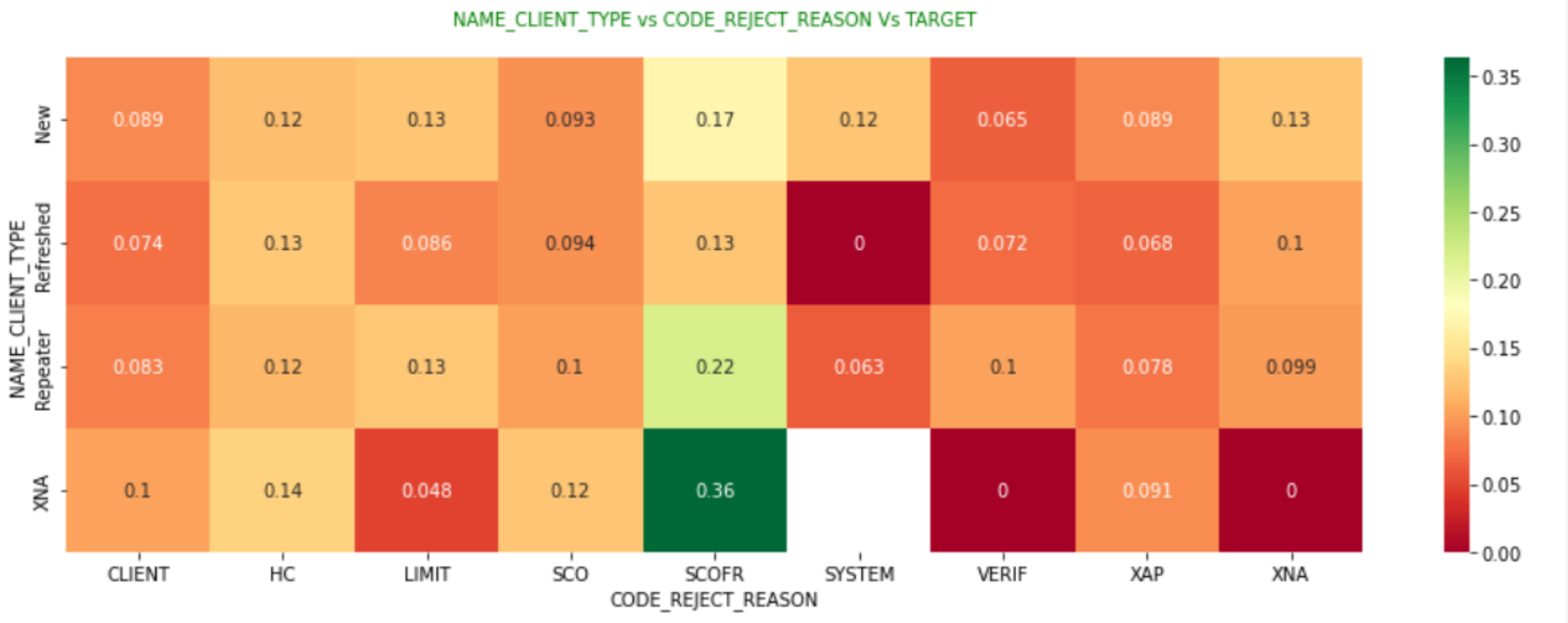
# ## Multivariate Analysis of the merged Data

1. Many Loan applications are refused with SCOFR as the reject reason



# ## Multivariate Analysis of the merged Data

1. XNA Client type have been refused with SCOFR as the reject reason



# ## Multivariate Analysis of the merged Data

1. Cash loans are high defaults with SCOFR as the reject reason and XNA type loans with HC as reject reason.