# EDA ASSIGNMENT

Rushmi Roshina M

# Data Handling and Data cleaning

- Columns with more than 40% NULL values are dropped
- Few other Null values are substituted with most suitable values
- Inconsistent data(XNA/XAP) are substituted with appropriate values
- Data with errors(Days columns) are converted to their correct and convenient value.

# Forming Derived Metrics and Bins

- **Age Binning**

  *[18,27,40,50,60,100]=[['Young','Young_Adult',  'Middle_Aged','Old','Very_Old']*

- **Income Binning**

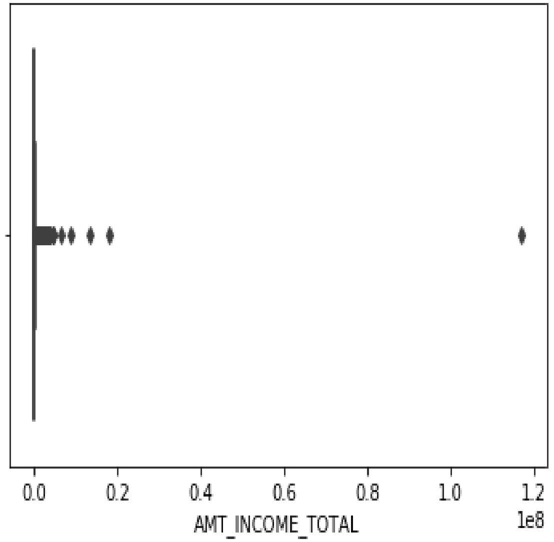
  *[min(income),50000,100000,300000,1000000,3000000,max(income)]=['Below_Poverty','Poverty','Lower_Middle_Class','Upper_Middle_Class','Rich','Very_Rich']*

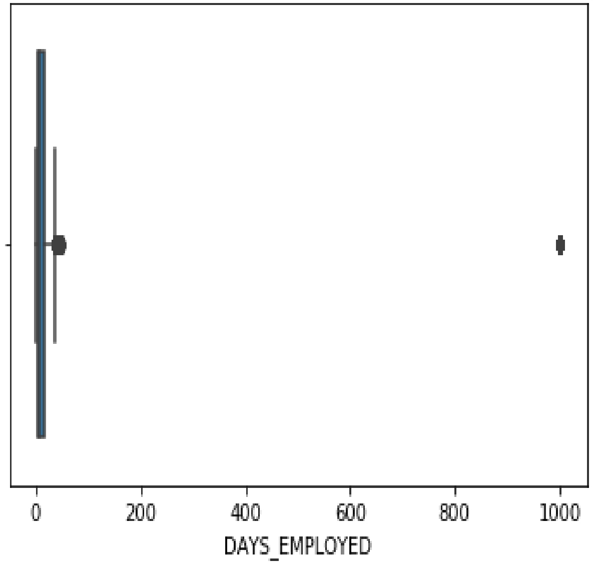
- **CRED_INC_RATIO** =*Ratio of Credit_Amount and Income*
- **CRED_GOODS_RATIO**=*Ratio of Credit_Amount and Goods*

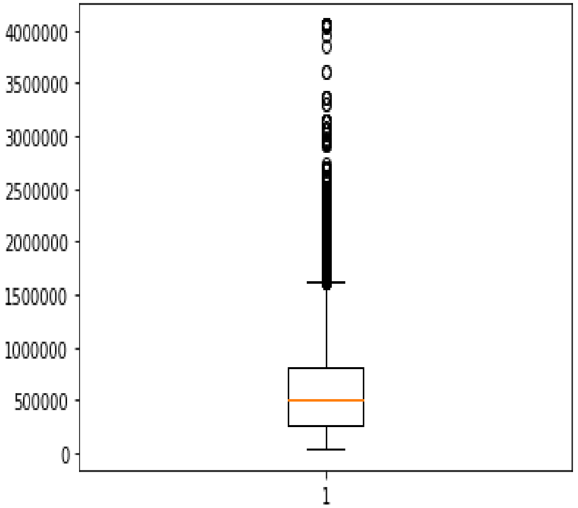
The above Metrics/Bins are created for Application Dataset

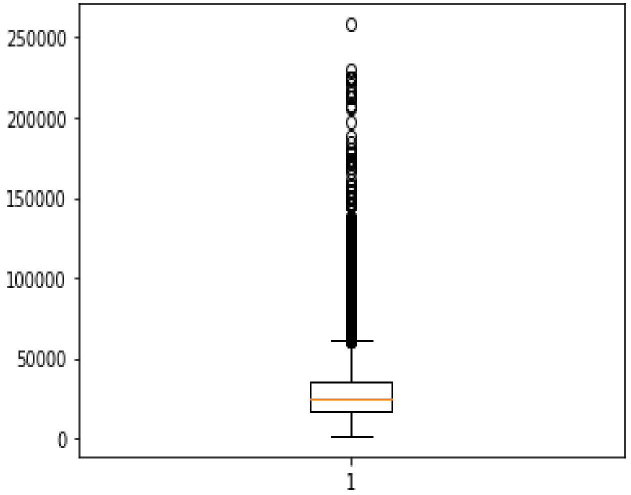
# Outlier Analysis

Outliers are Analyzed for AMT_INCOME_TOTAL,DAYS_EMPLOYED,AMT_CREDIT,AMT_ANNUITY.They seem to have no impact on the data.Hence it is not taken care off.

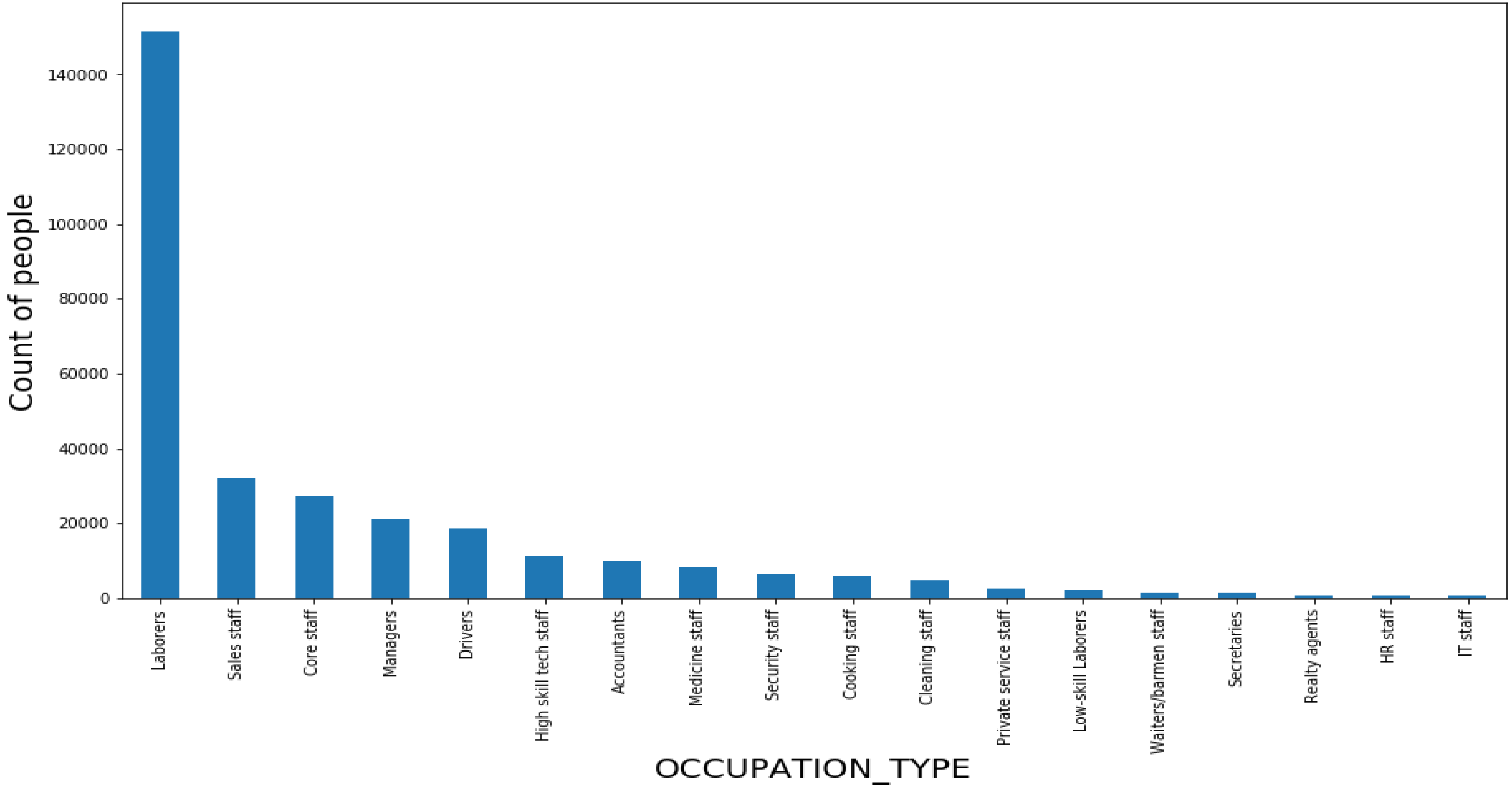


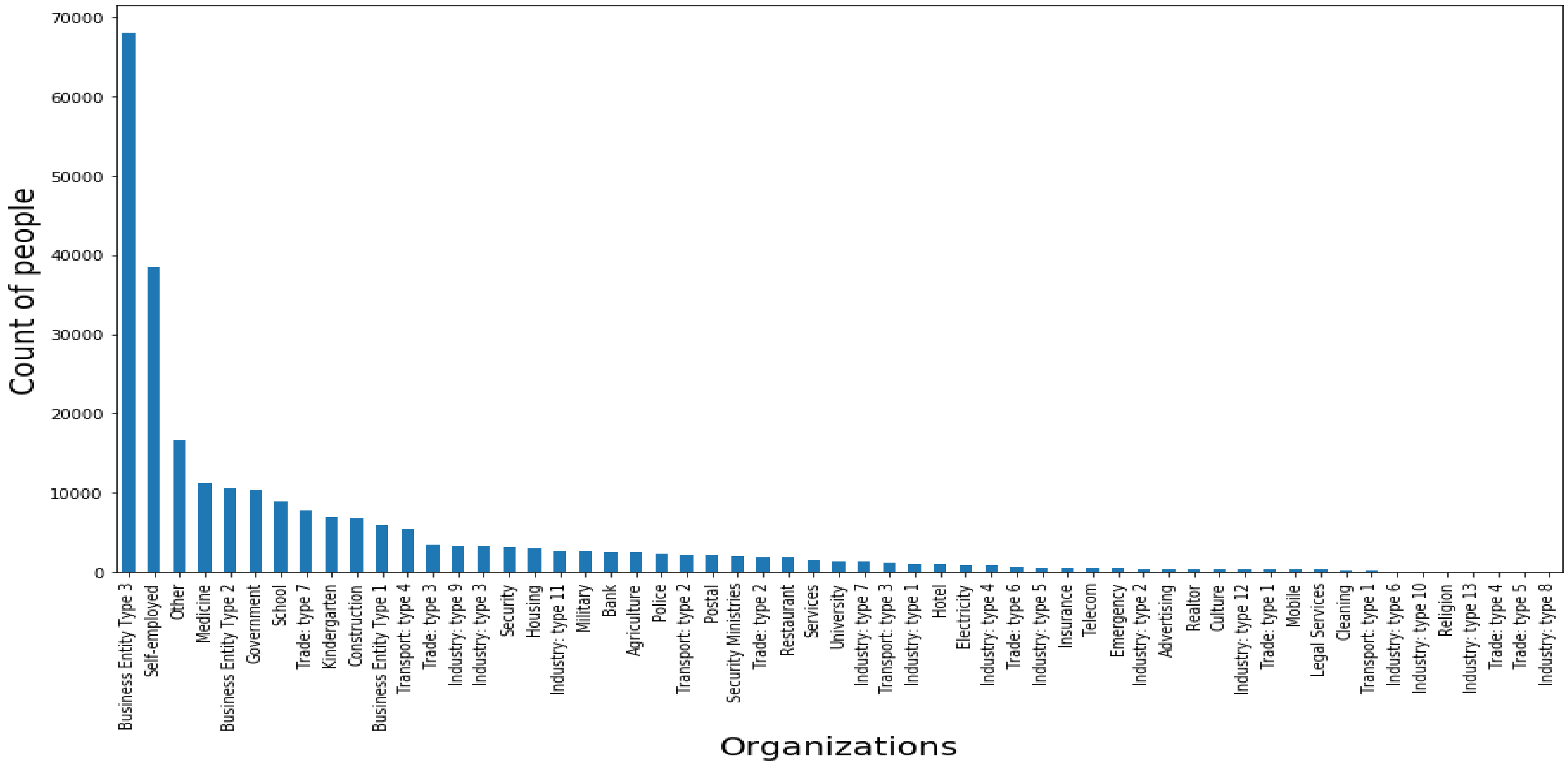



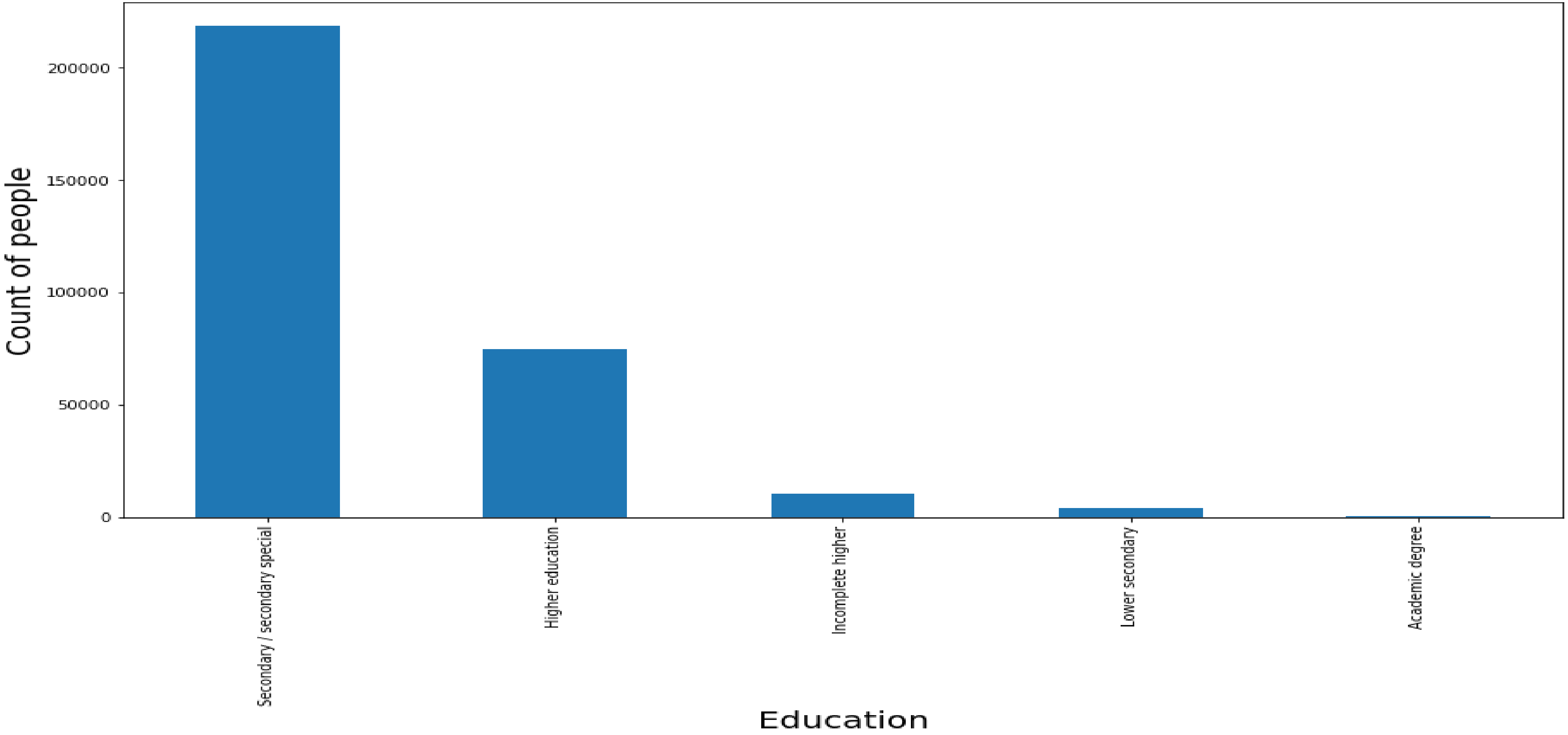
# Data Distribution on Occupation type



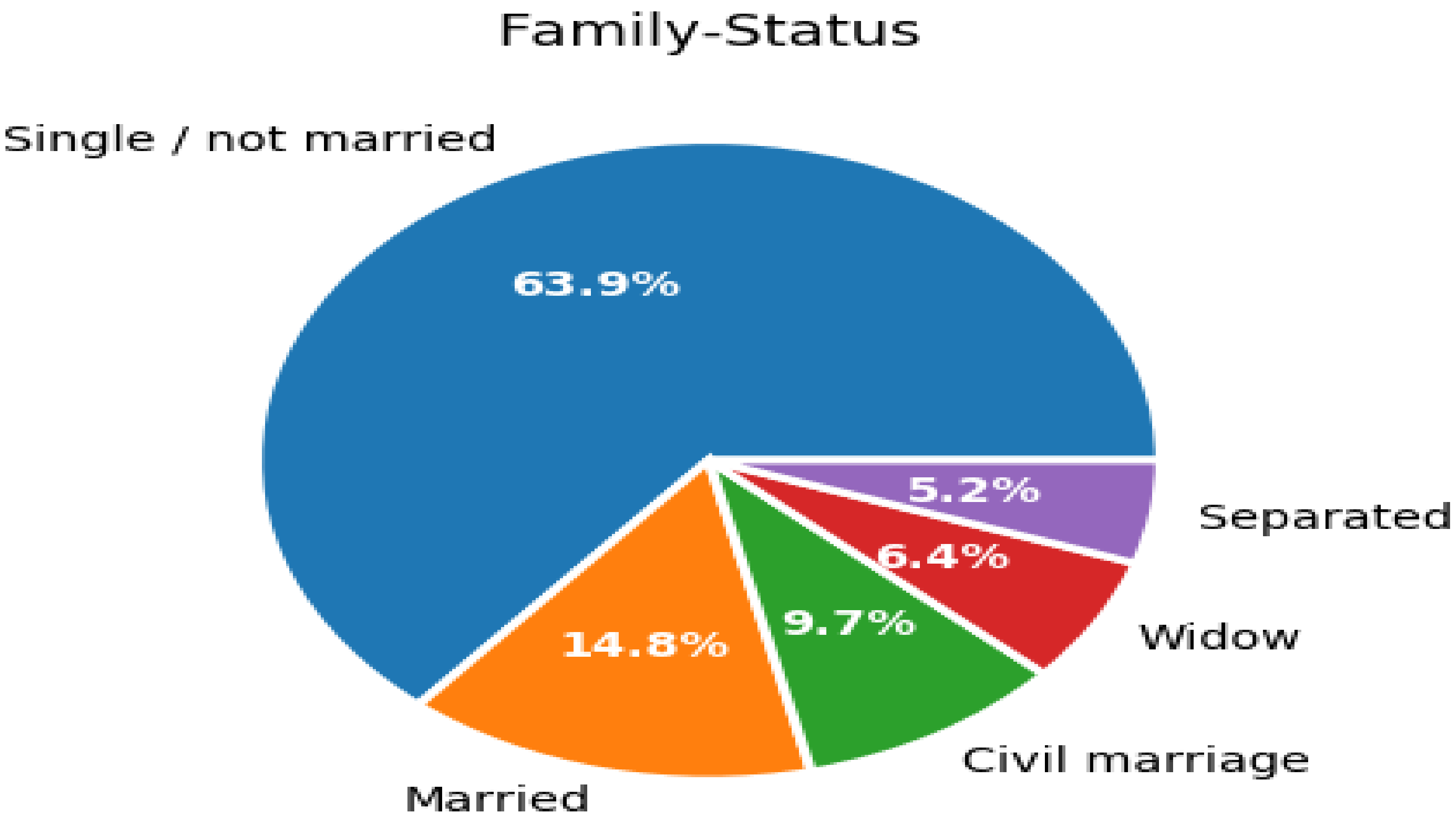
# Data Distribution on Organization type



# Data Distribution on Education type

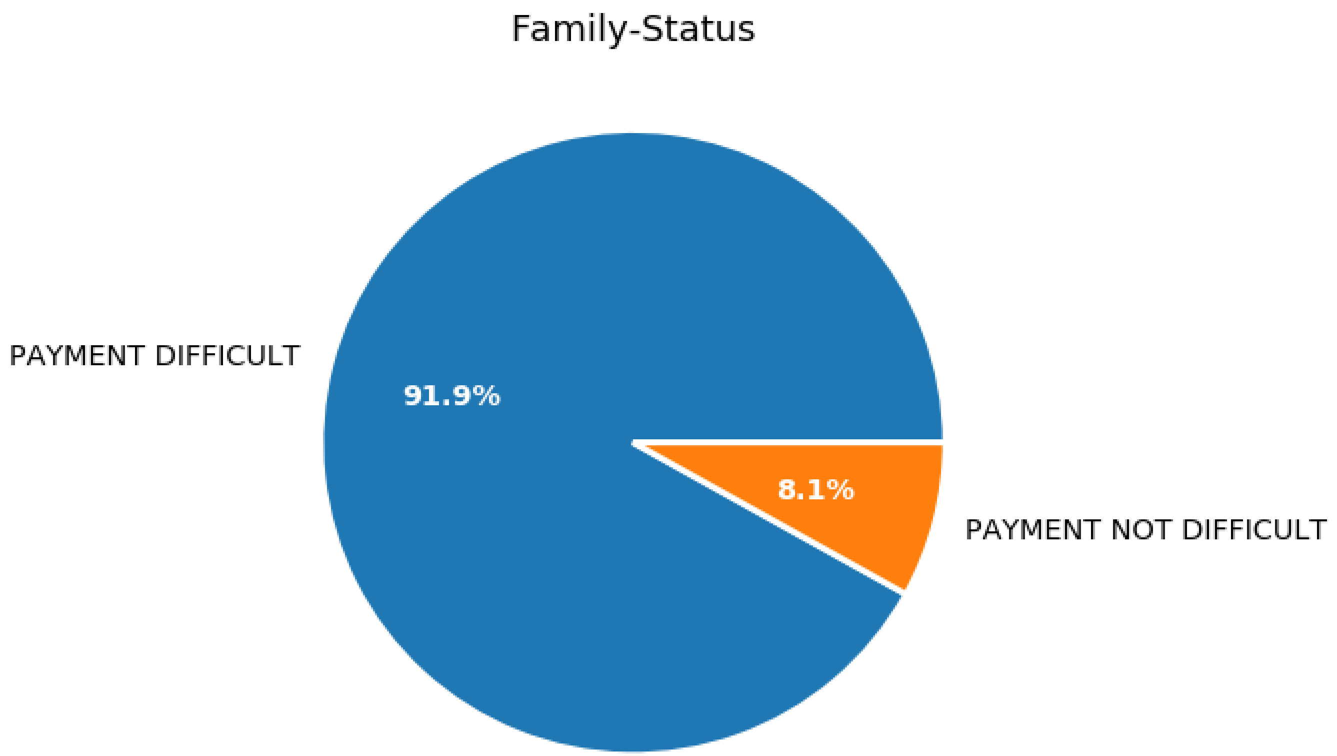


# Proportion of Family-status

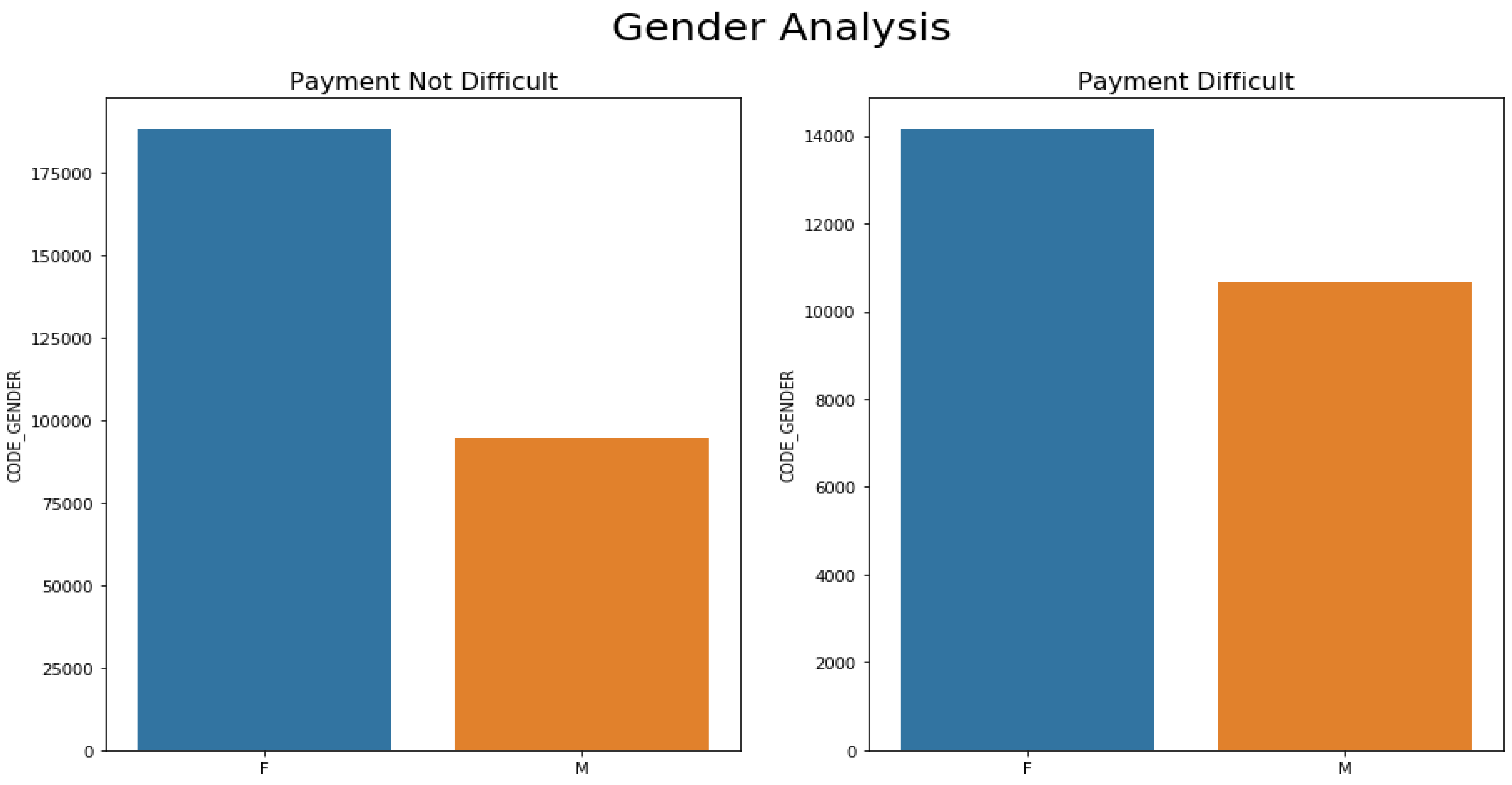


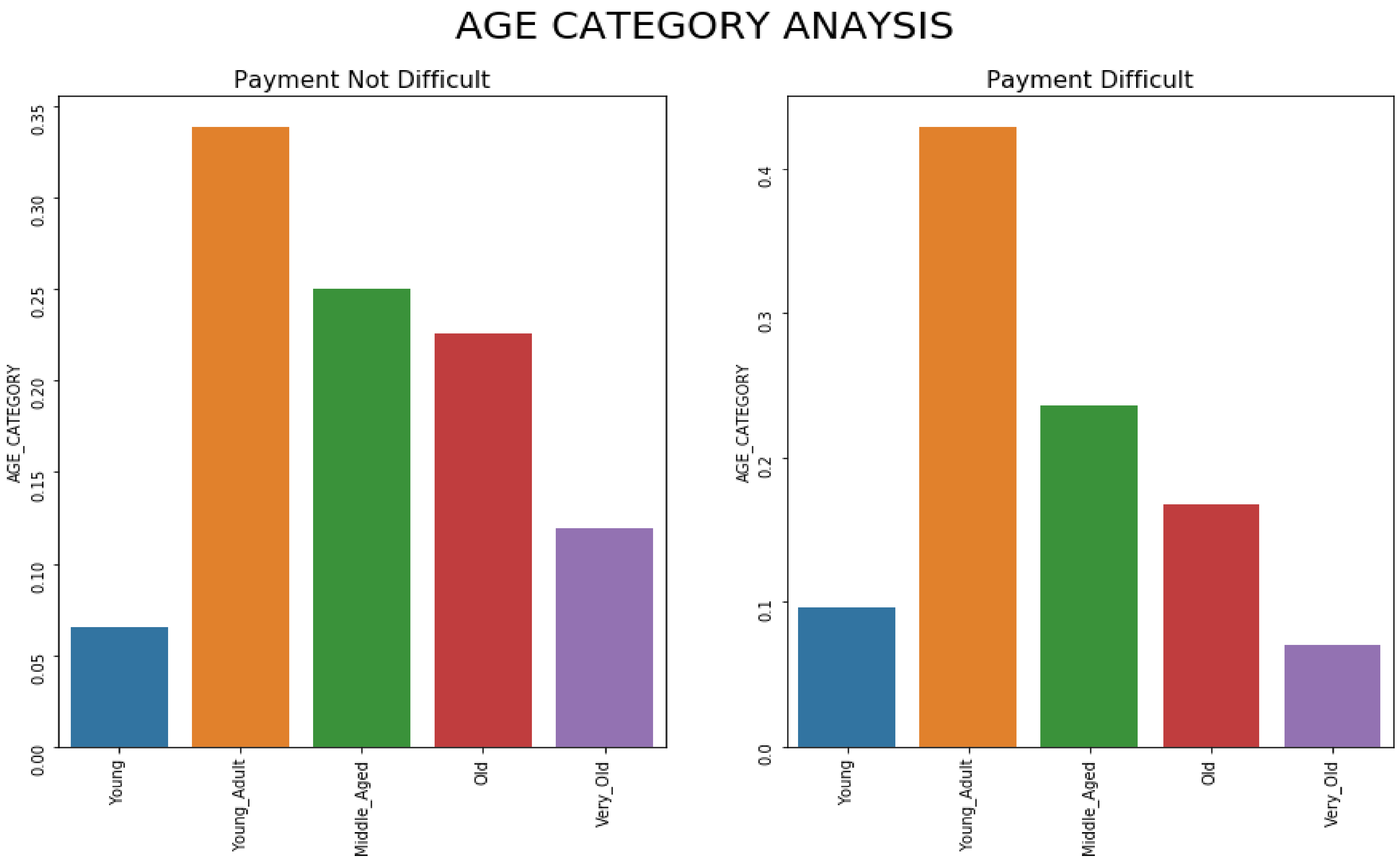
# Data Imbalance

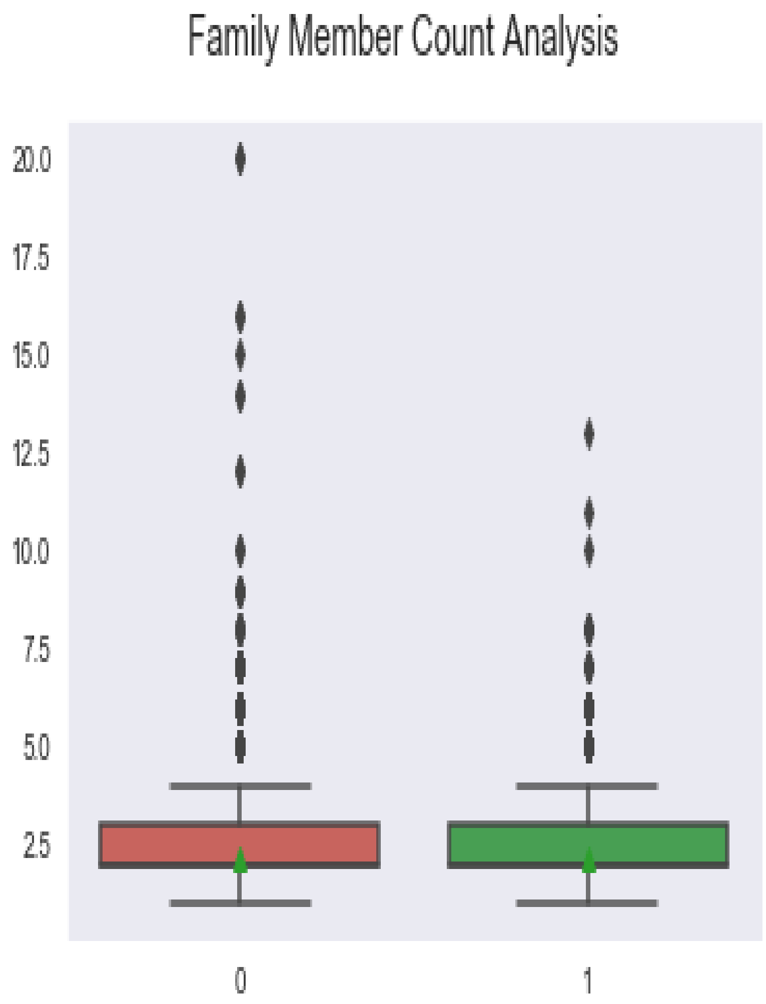
## Most of the Applications have difficulties with payment

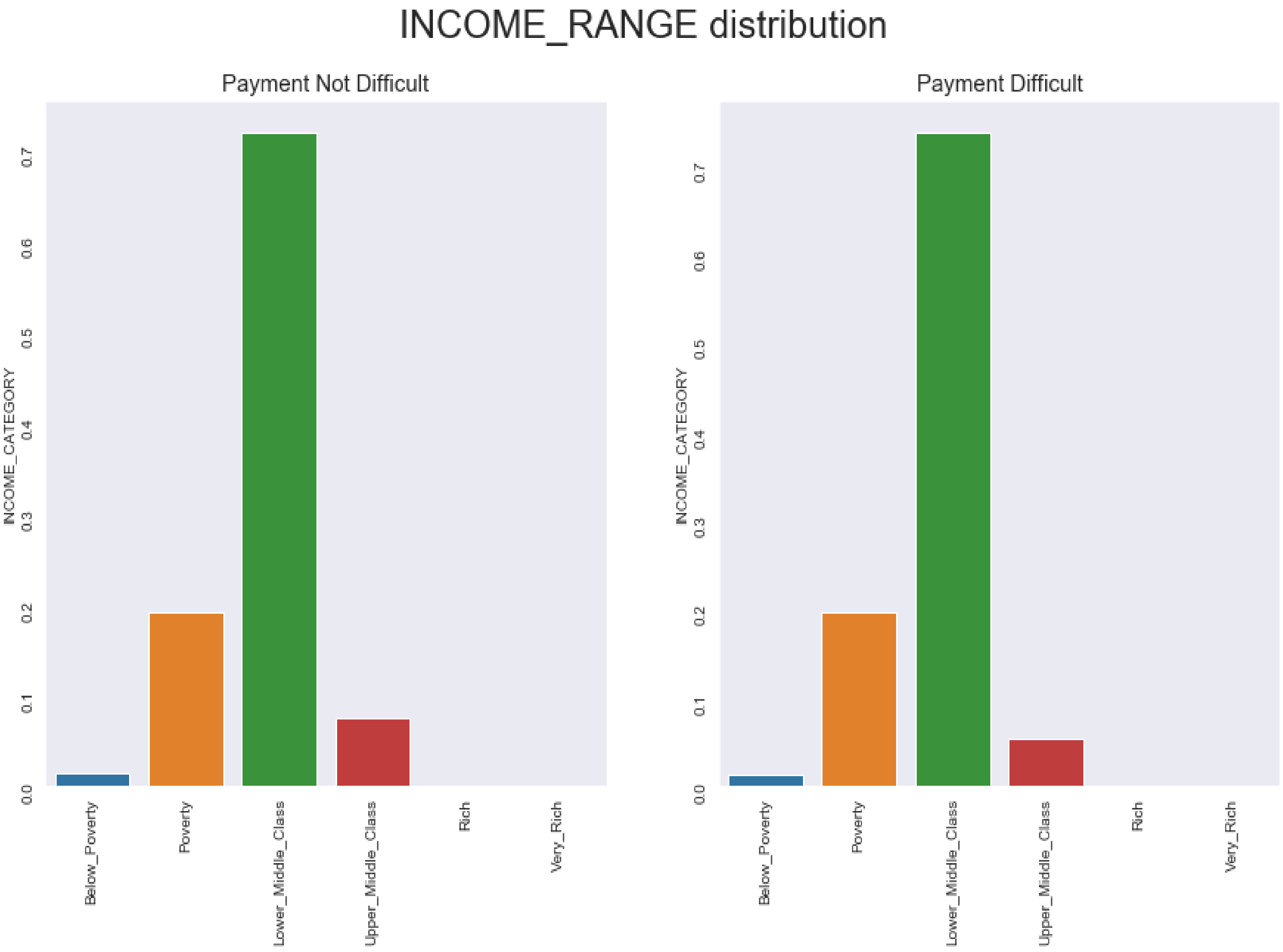


# Bi-variate Analysis

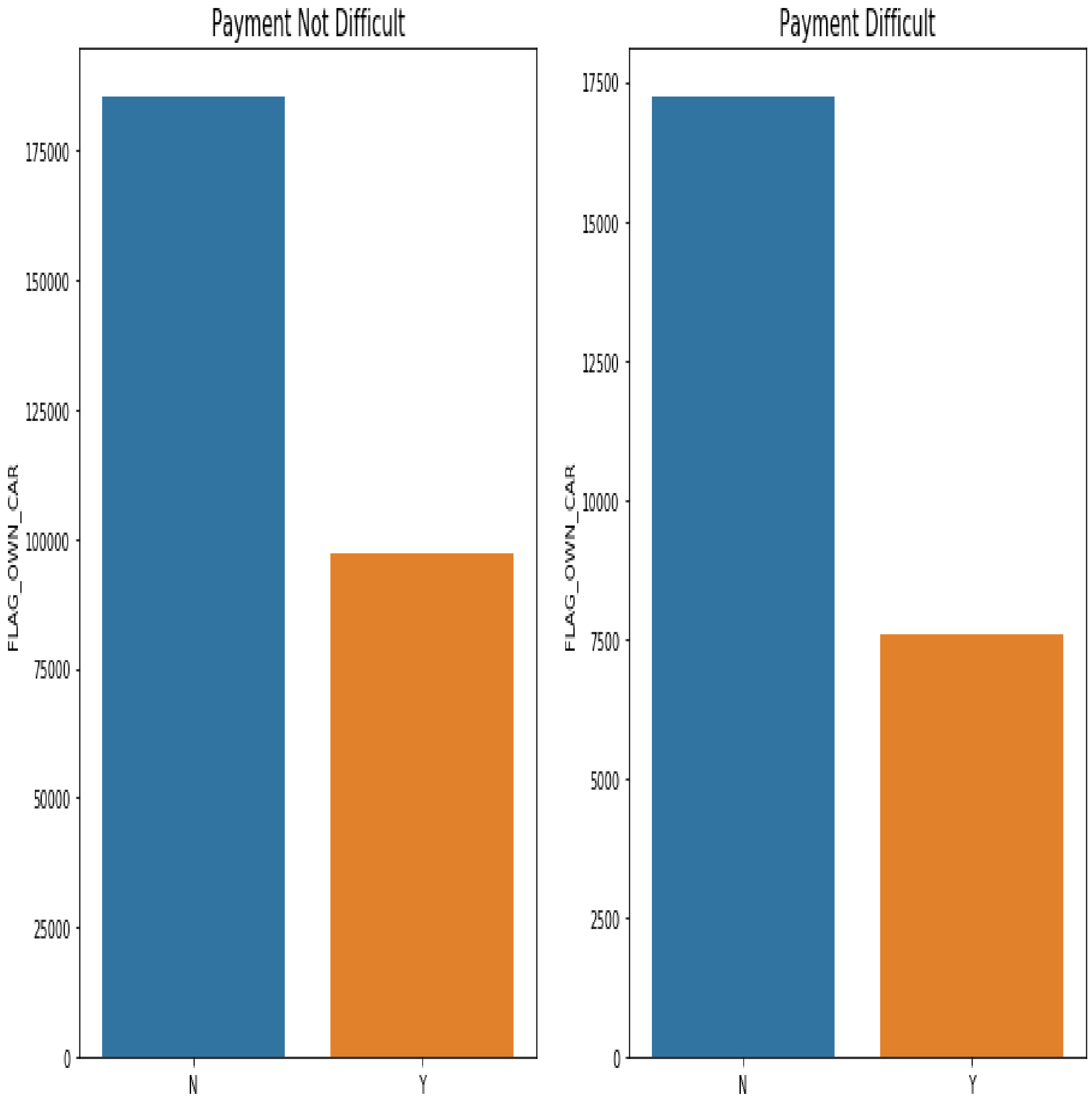


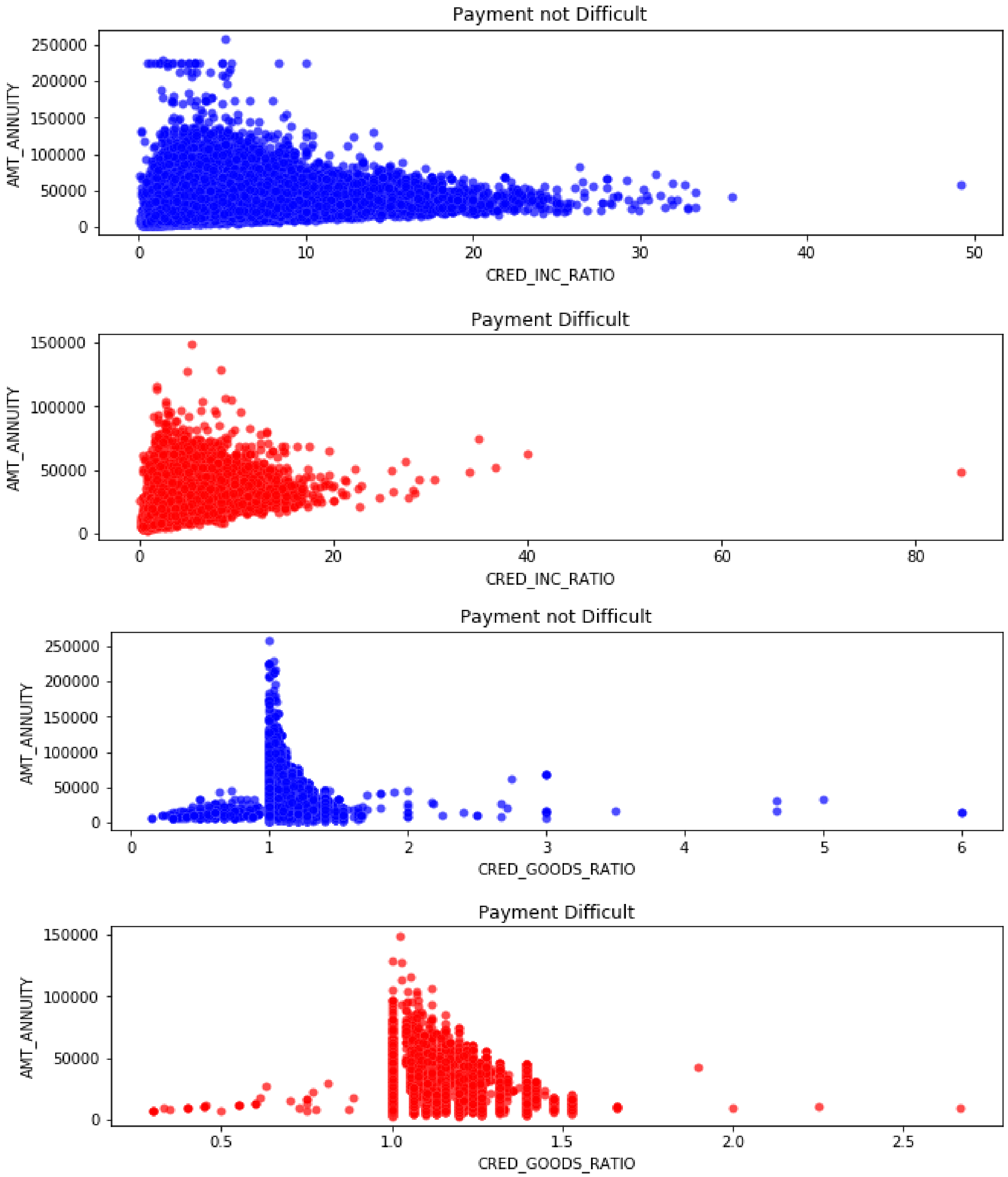

AGE CATEGORY ANAYSIS
Payment Not Difficult
Payment Difficult
AGE_CATEGORY
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.0
0.1
0.2
0.3
0.4
Young
Young_Adult
Middle_Aged
Old
Very_Old


Family Member Count Analysis
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
1

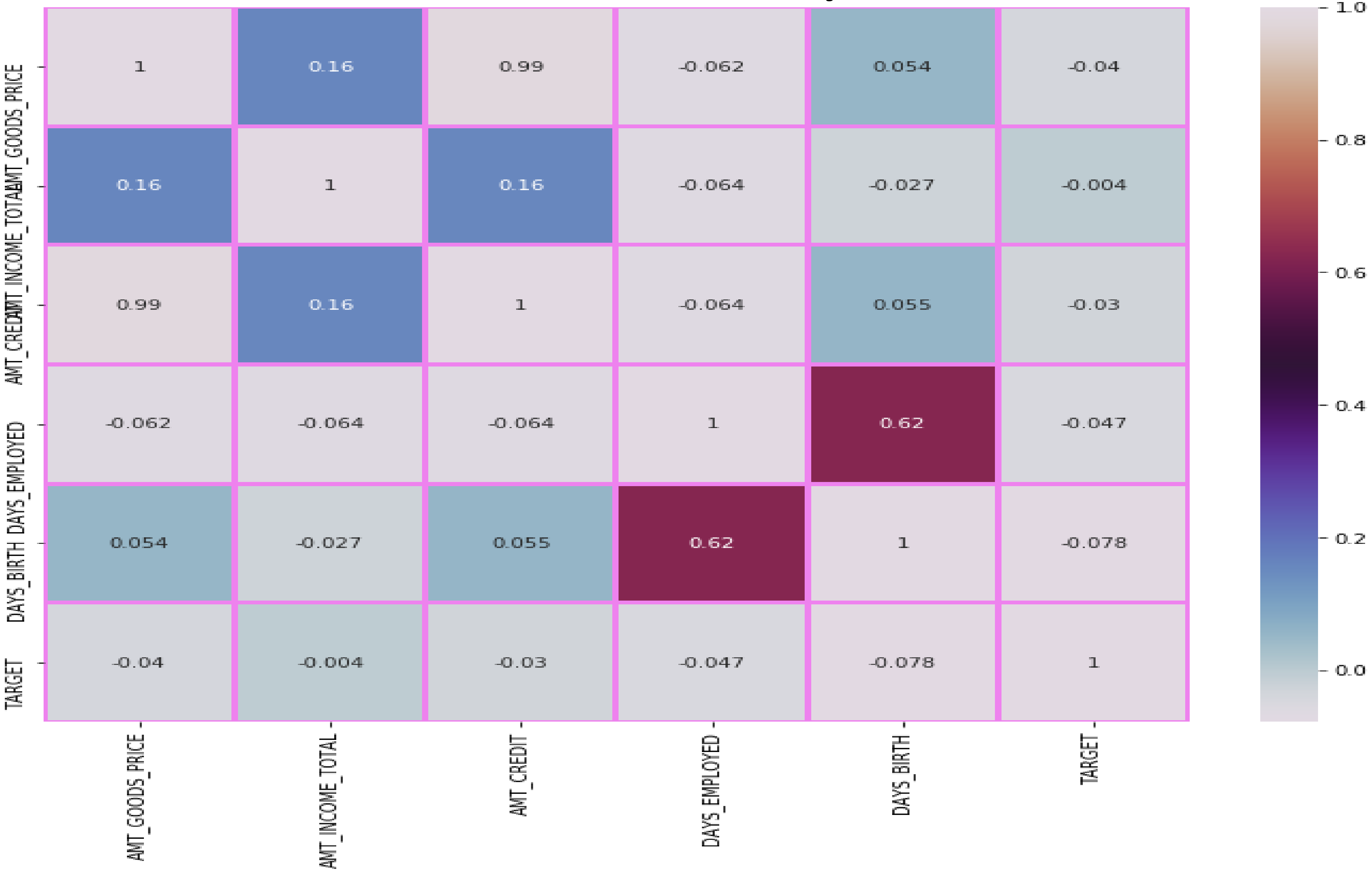

INCOME_RANGE distribution
Payment Not Difficult
Payment Difficult
INCOME_CATEGORY
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Below_Poverty
Poverty
Lower_Middle_Class
Upper_Middle_Class
Rich
Very_Rich

# Owns Car Analysis



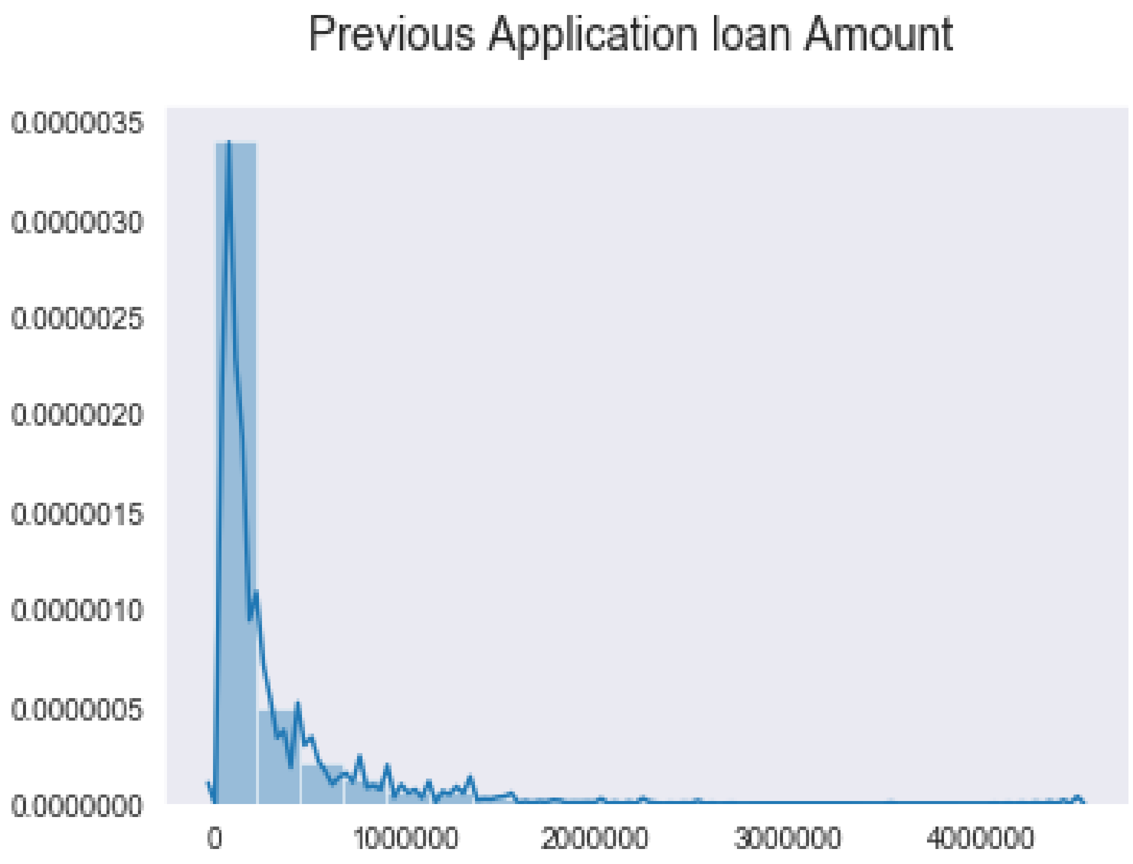



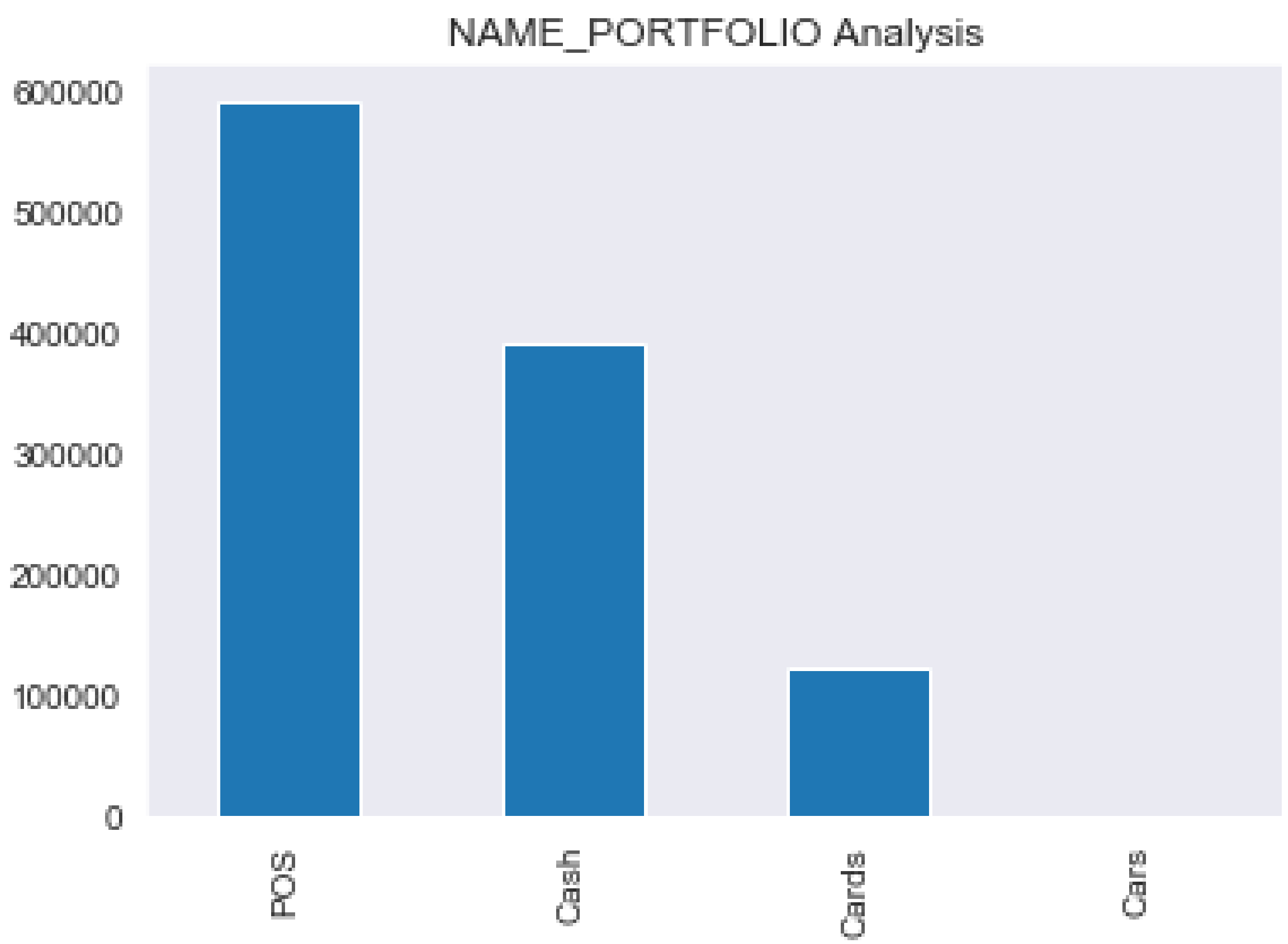
# Multivariate Analysis

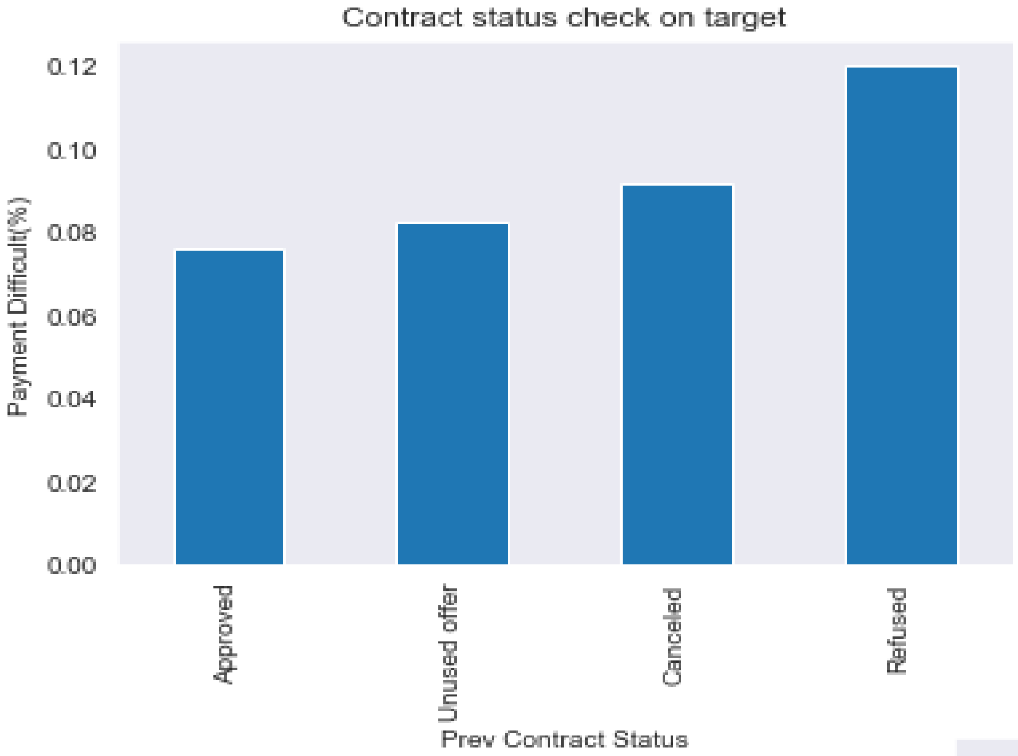


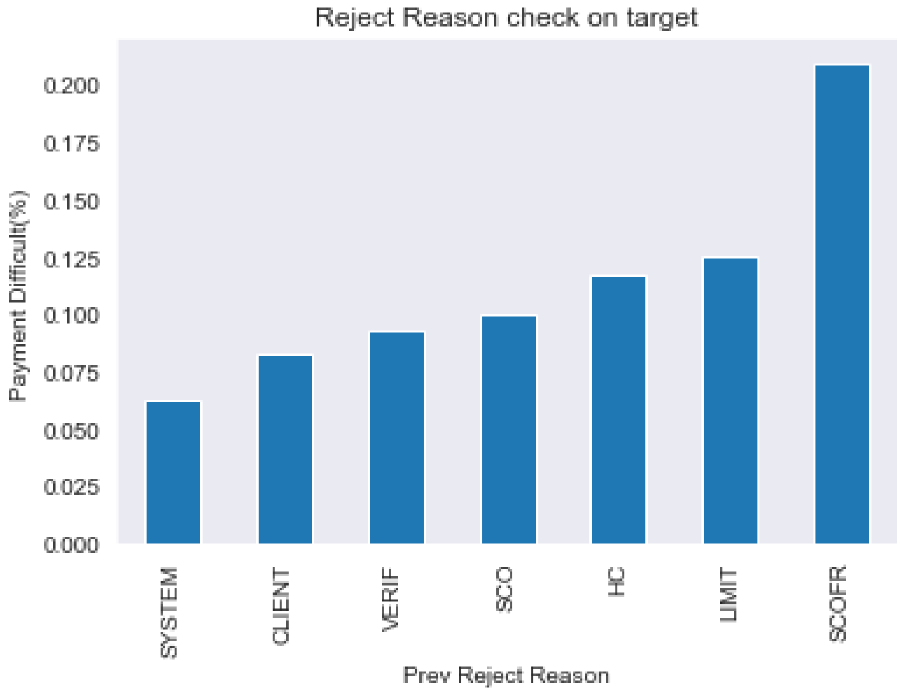
# Insights on Application Data

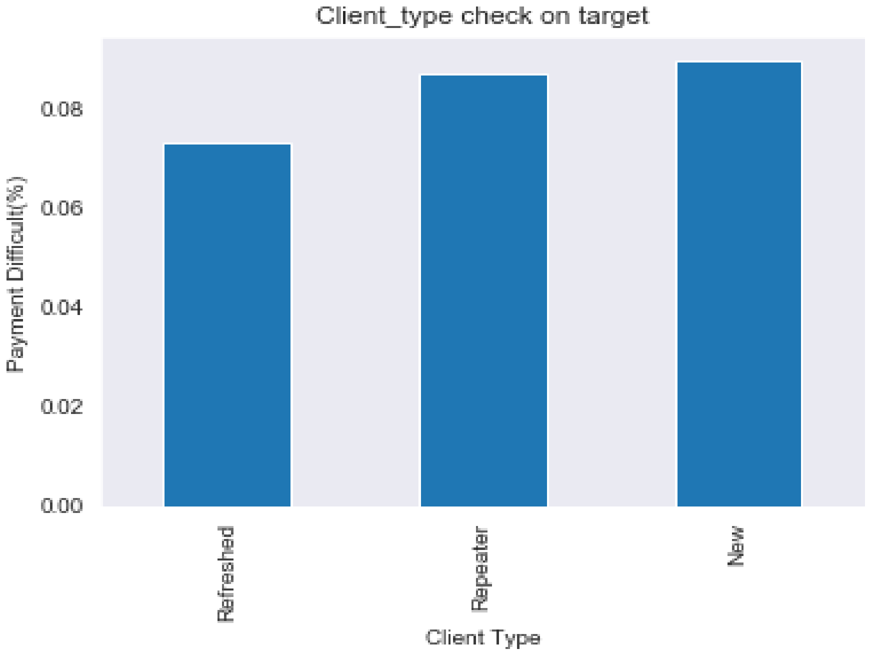
- Most of the Male Applicants have  Payment Difficulty
- People of age less than 40 have more payment difficulty
- Family Member count has no big impact on the Target variable
- All the Income_Category people are equally distributed on the Target variable
- People who don't  own car have more difficulty with payment than the people who own car


Previous Application loan Amount
0.0000035
0.0000030
0.0000025
0.0000020
0.0000015
0.0000010
0.0000005
0.0000000
0
1000000
2000000
3000000
4000000


NAME_PORTFOLIO Analysis
600000
500000
400000
300000
200000
100000
0
POS
Cash
Cards
Cars


Contract status check on target
Payment Difficult(%)
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Approved
Unused offer
Canceled
Refused
Prev Contract Status

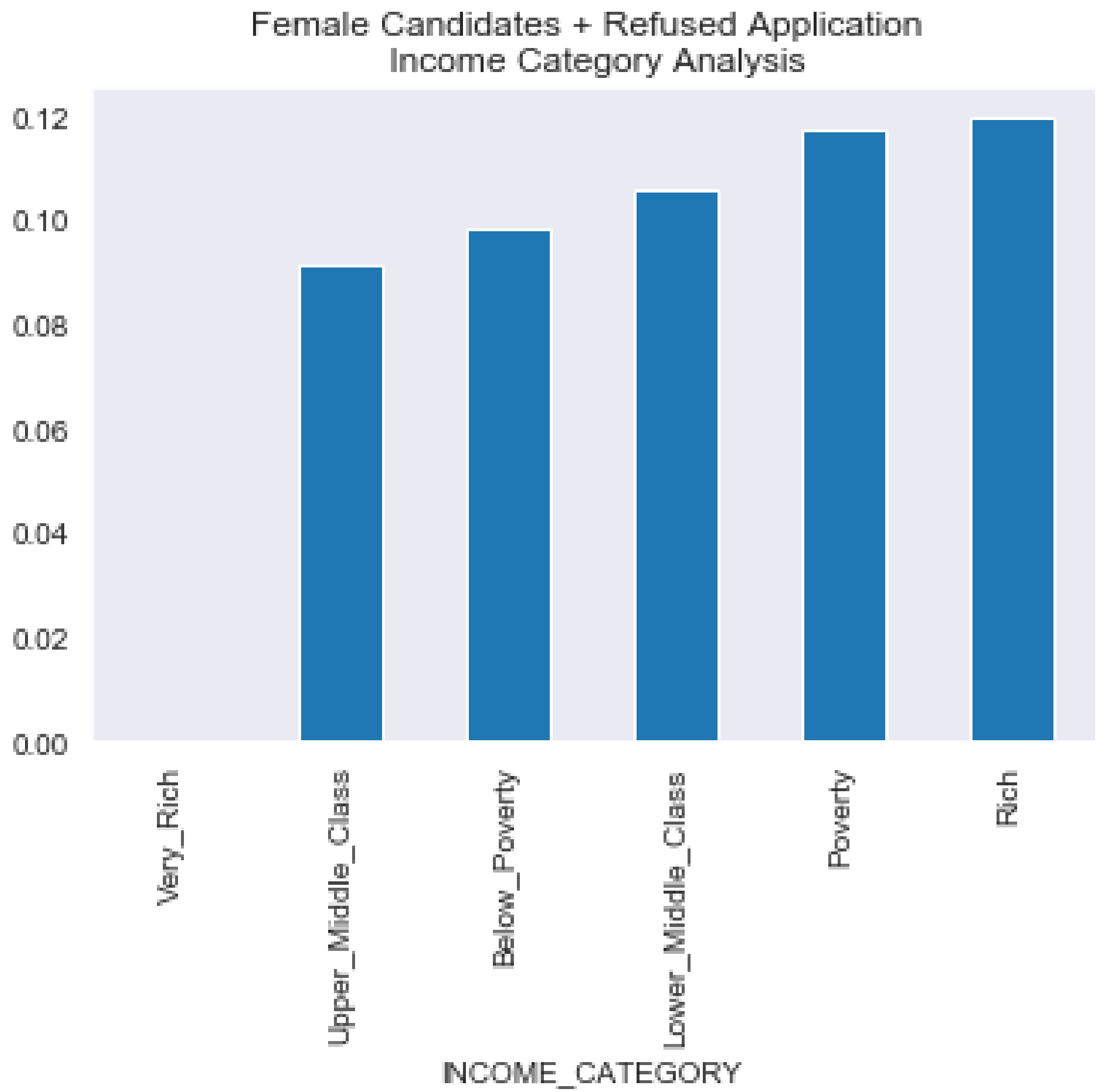

Reject Reason check on target
Payment Difficult(%)
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
SYSTEM
CLIENT
VERIF
SCO
HC
LIMIT
SCOFR
Prev Reject Reason

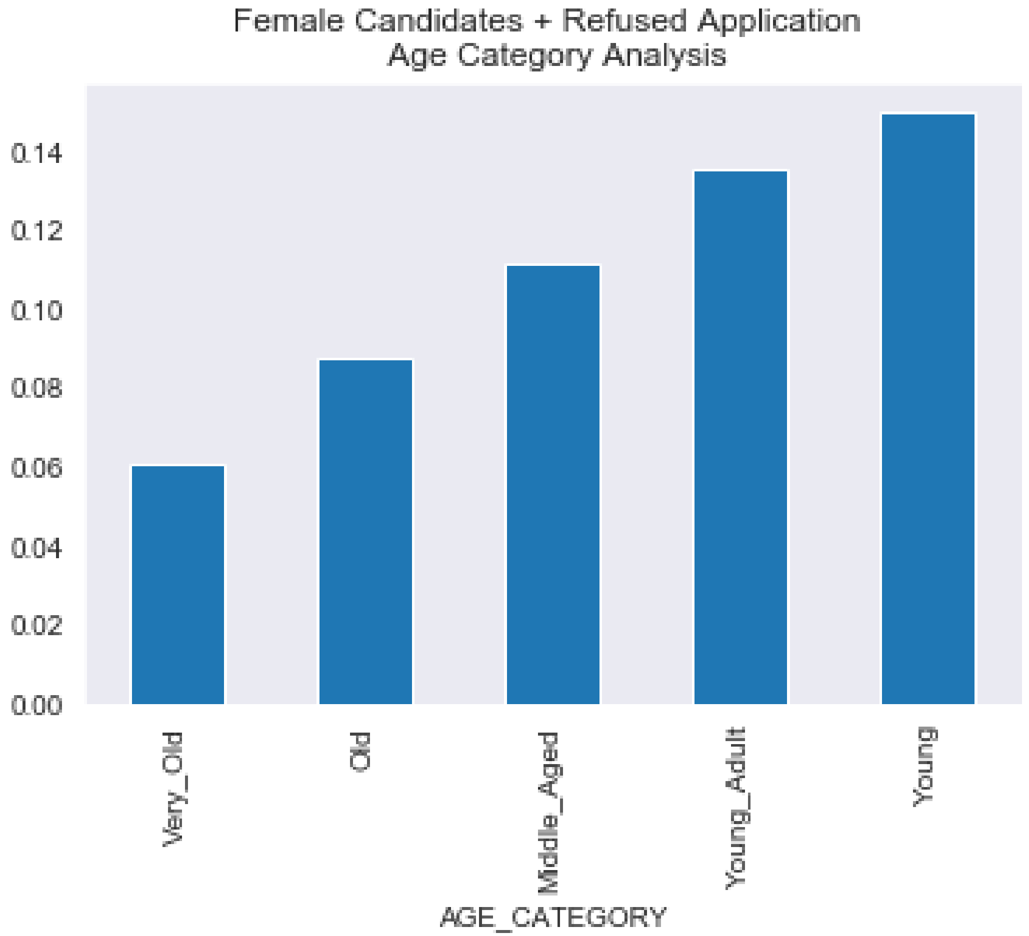

Client_type check on target
Payment Difficult(%)
0.08
0.06
0.04
0.02
0.00
Refreshed
Repeater
New
Client Type

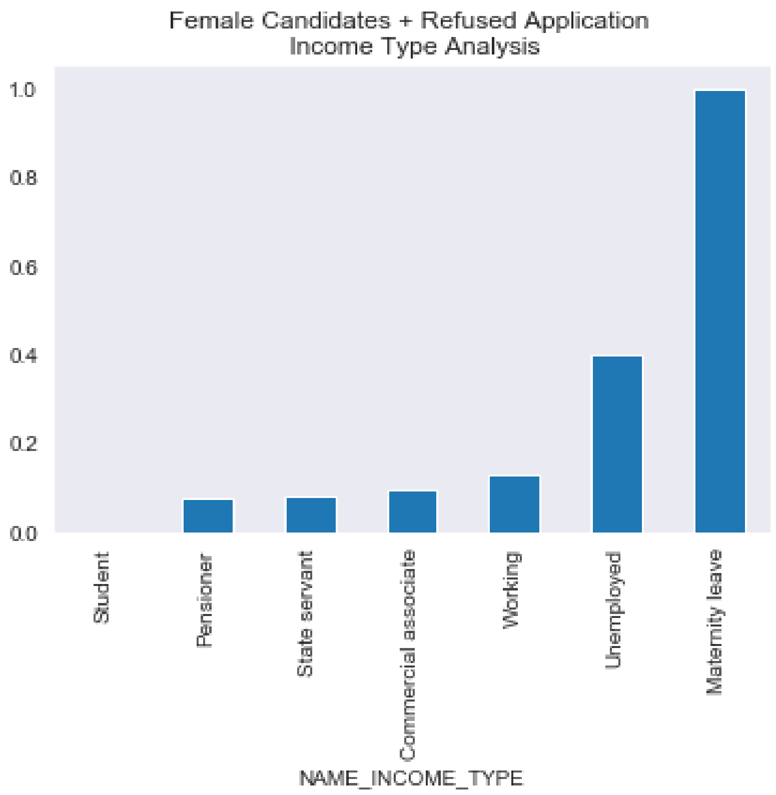
# Insights on Previous Application Dataset

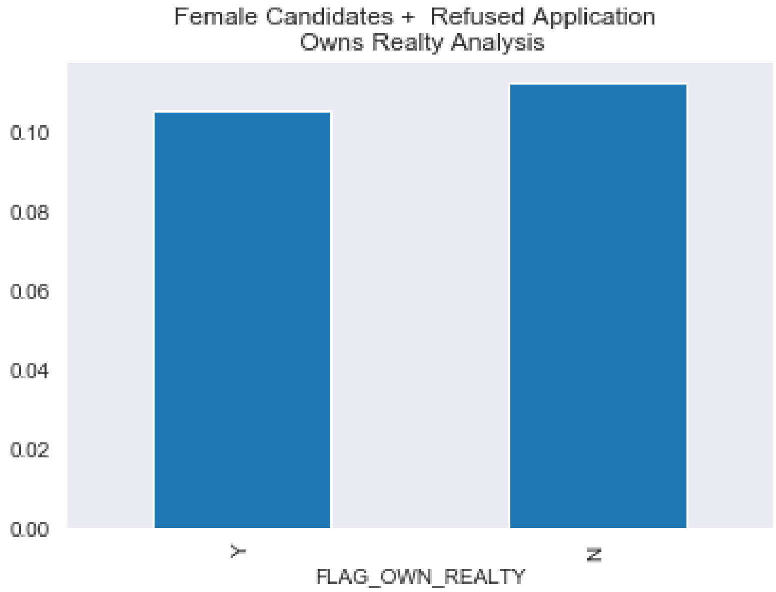
- Most of the Applicant who has difficulties with payment,was previously Refuse loan
- The highest rejected reason for people with payment difficulties was SCOFR
- Most Applicant's loan_amount was below 1000000
- Most of the loan was through POS
- People with payment Difficulties are mostly new loan Applicants

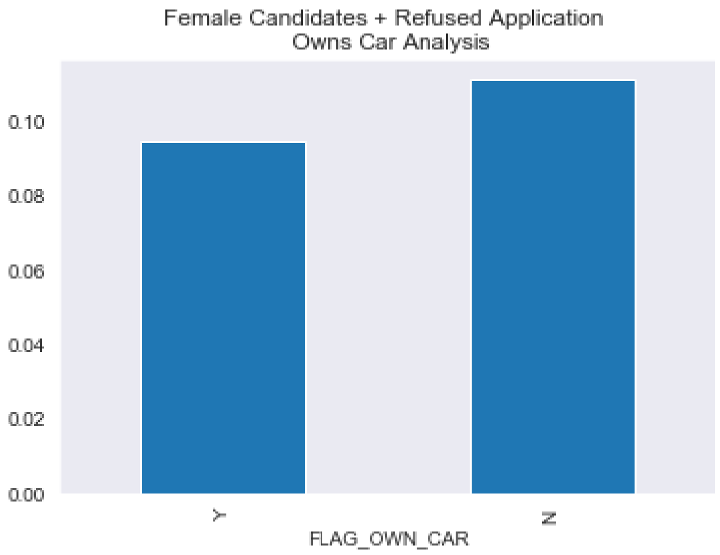
# Combined Analysis(Female)



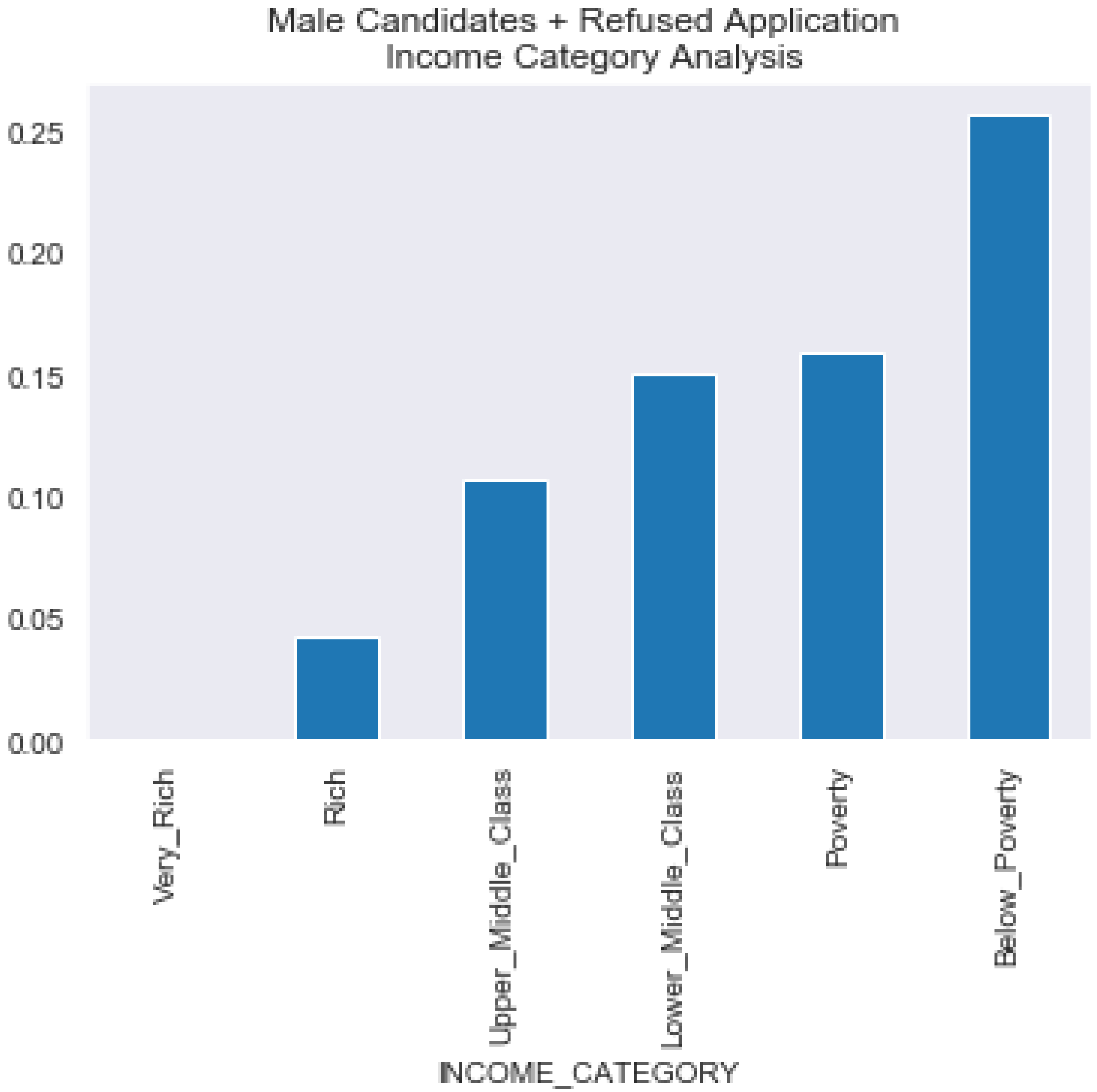



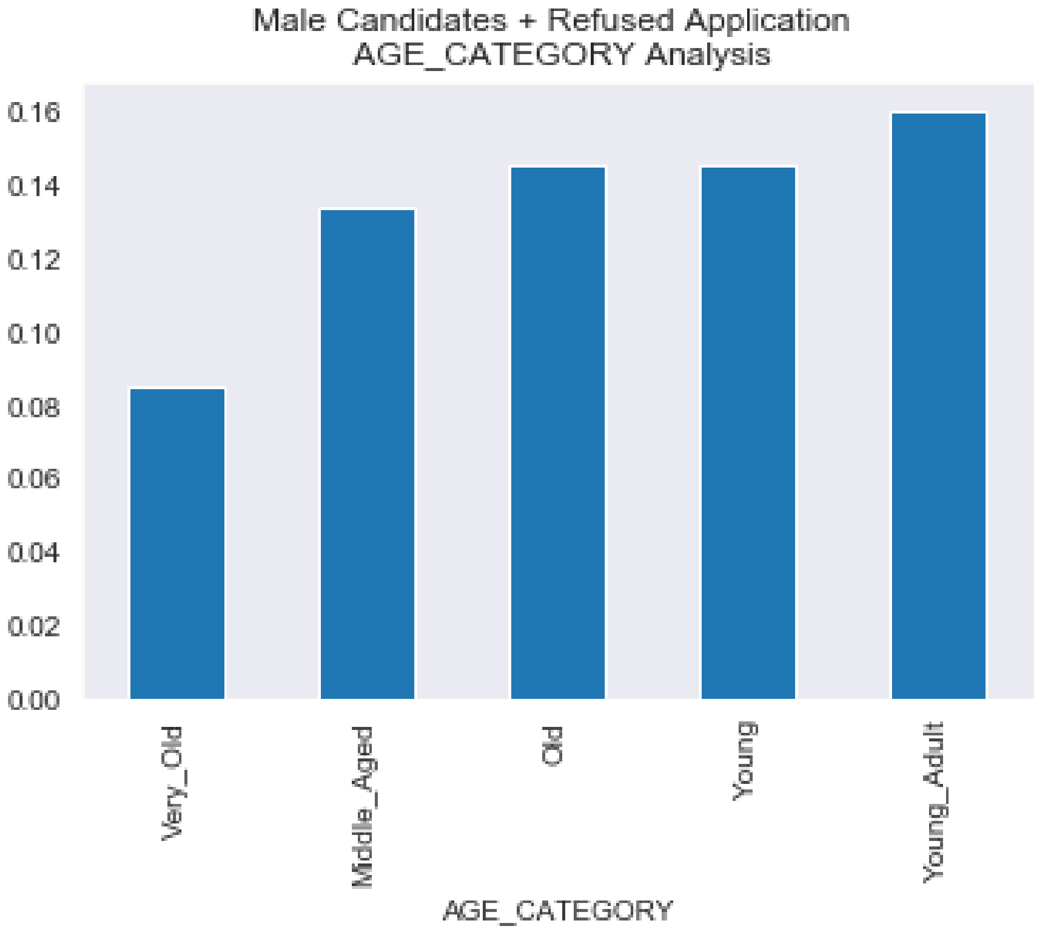

Female Candidates + Refused Application
Income Type Analysis
1.0
0.8
0.6
0.4
0.2
0.0
Student
Pensioner
State servant
Commercial associate
Working
Unemployed
Maternity leave
NAME_INCOME_TYPE

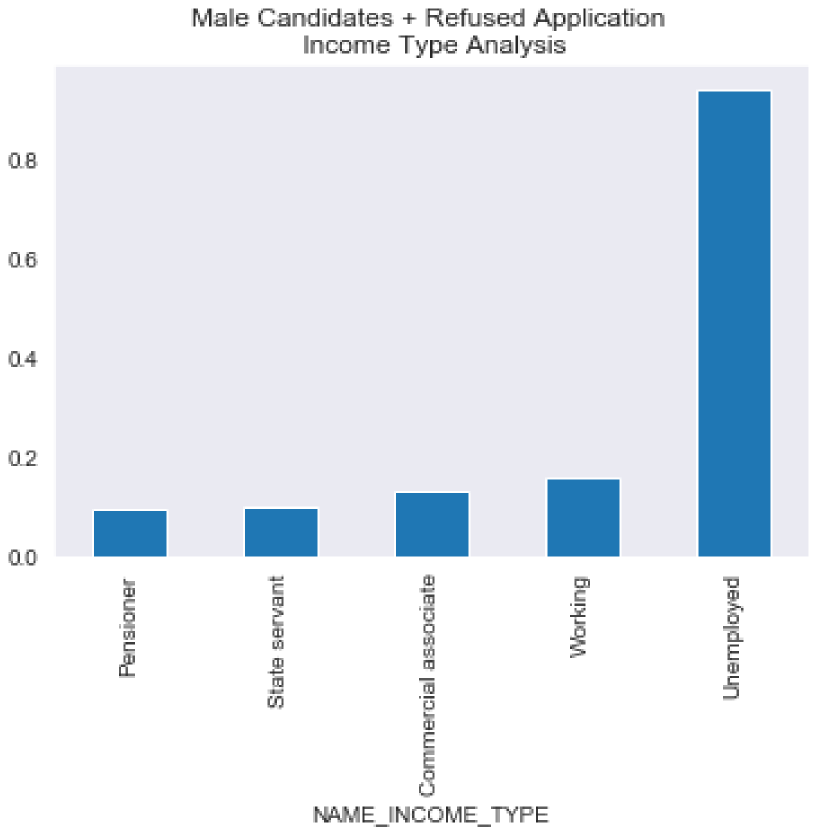

Female Candidates +  Refused Application
Owns Realty Analysis
0.10
0.08
0.06
0.04
0.02
0.00
Y
N
FLAG_OWN_REALTY

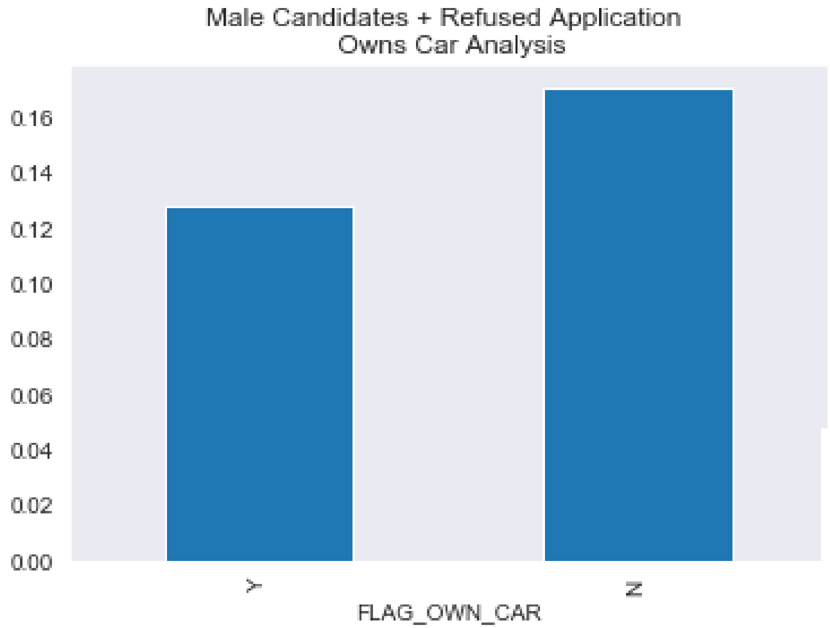

Female Candidates + Refused Application
Owns Car Analysis
0.10
0.08
0.06
0.04
0.02
0.00
Y
N
FLAG_OWN_CAR

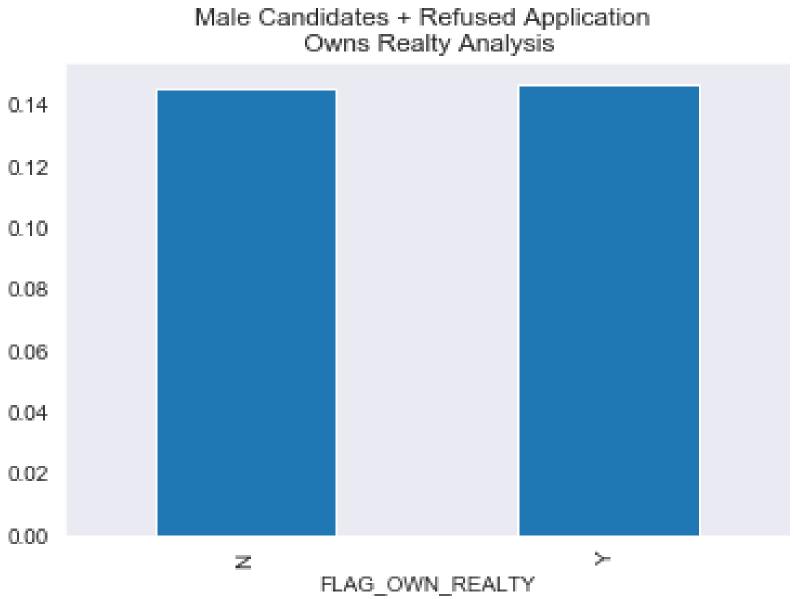
# Combined Analysis(Male)



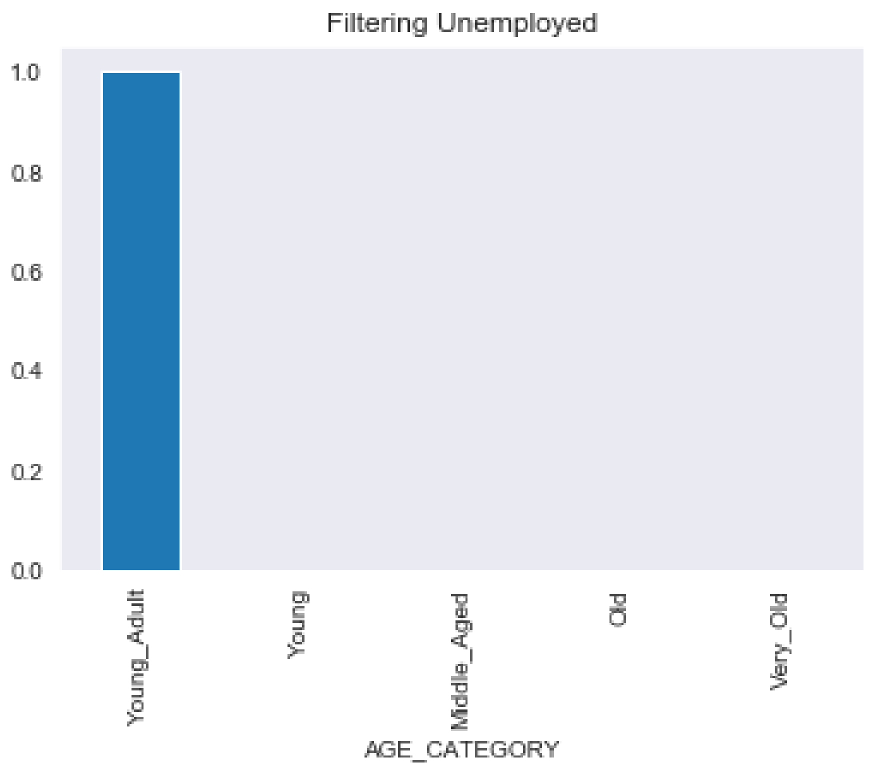




Male Candidates + Refused Application
Income Type Analysis
0.8
0.6
0.4
0.2
0.0
Pensioner
State servant
Commercial associate
Working
Unemployed
NAME_INCOME_TYPE

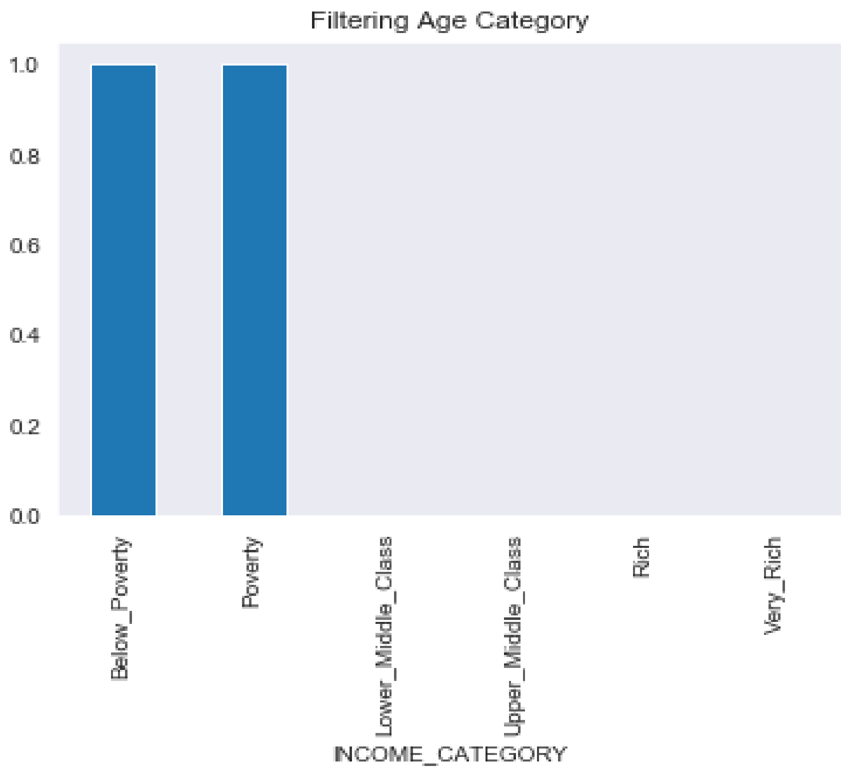

Male Candidates + Refused Application
Owns Car Analysis
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Y
N
FLAG_OWN_CAR


Male Candidates + Refused Application
Owns Realty Analysis
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
N
Y
FLAG_OWN_REALTY

Further these Analysis can be used to filter out rows and fetch the people who have least possibility of paying back the loan




Click to add text



# Combined Insights

- Most people who doesn't own Car or Realty may not be able to payback the loan
- Female Candidates who fall under the category Young and are in Maternity leave may face difficulties in paying back the loan
- Male Candidates who fall under the category Young-Adult and are Unemployed and are in the Below_Poverty range may face difficulties in paying back the loan