# CREDIT CASE STUDY
# EDA ASSIGNMENT

# “ BANK LOAN ASSIGNMENT ”

By : Ajit Murugan

# Data Cleaning

# Initial Dropping of Columns

- Data Column for Application data having Rows with ‘Null Values’ which is greater than 50%”
- These mentioned columns will create wrong correlation with target variable . So, dropping this columns
- Image 1 shows the columns name which has null values greater than 50 %

```
# Finding cols in df_a with null data > 50 %
inp_a.isnull().sum()[inp_a.isnull().sum() > 307511//2]
```

```
OWN_CAR_AGE                   202929
EXT_SOURCE_1                  173378
APARTMENTS_AVG                156061
BASEMENTAREA_AVG              179943
YEARS_BUILD_AVG               204488
COMMONAREA_AVG                214865
ELEVATORS_AVG                 163891
ENTRANCES_AVG                 154828
FLOORSMIN_AVG                 208642
LANDAREA_AVG                  182590
LIVINGAPARTMENTS_AVG          210199
LIVINGAREA_AVG                154350
NONLIVINGAPARTMENTS_AVG       213514
NONLIVINGAREA_AVG             169682
APARTMENTS_MODE               156061
BASEMENTAREA_MODE             179943
YEARS_BUILD_MODE              204488
COMMONAREA_MODE               214865
ELEVATORS_MODE                163891
ENTRANCES_MODE                154828
FLOORSMIN_MODE                208642
LANDAREA_MODE                 182590
LIVINGAPARTMENTS_MODE         210199
LIVINGAREA_MODE               154350
NONLIVINGAPARTMENTS_MODE      213514
NONLIVINGAREA_MODE            169682
APARTMENTS_MEDI               156061
BASEMENTAREA_MEDI             179943
YEARS_BUILD_MEDI              204488
COMMONAREA_MEDI               214865
ELEVATORS_MEDI                163891
ENTRANCES_MEDI                154828
FLOORSMIN_MEDI                208642
LANDAREA_MEDI                 182590
LIVINGAPARTMENTS_MEDI         210199
LIVINGAREA_MEDI               154350
NONLIVINGAPARTMENTS_MEDI      213514
NONLIVINGAREA_MEDI            169682
FONDKAPREMONT_MODE            210295
HOUSETYPE_MODE                154297
WALLSMATERIAL_MODE            156341
dtype: int64
```

Image 1

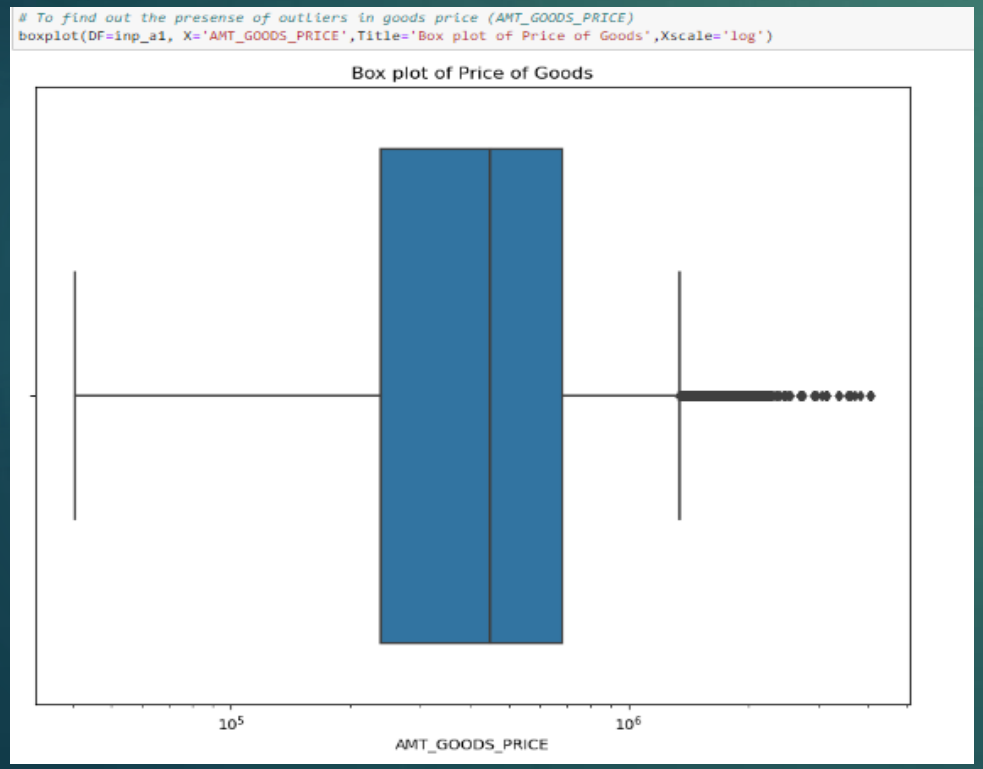
# Outlier Finding Using Box Plot

# Price of Goods

**Evidence**

- Median lies around 4 to 6 Lakhs.
- But most of points lies more than 75%
- Reason is many people have applied for goods ranging less than 6 Lakhs only.

```
# To find out the presense of outliers in goods price (AMT_GOODS_PRICE)
boxplot(DF=inp_a1, X='AMT_GOODS_PRICE',Title='Box plot of Price of Goods',Xscale='log')
```

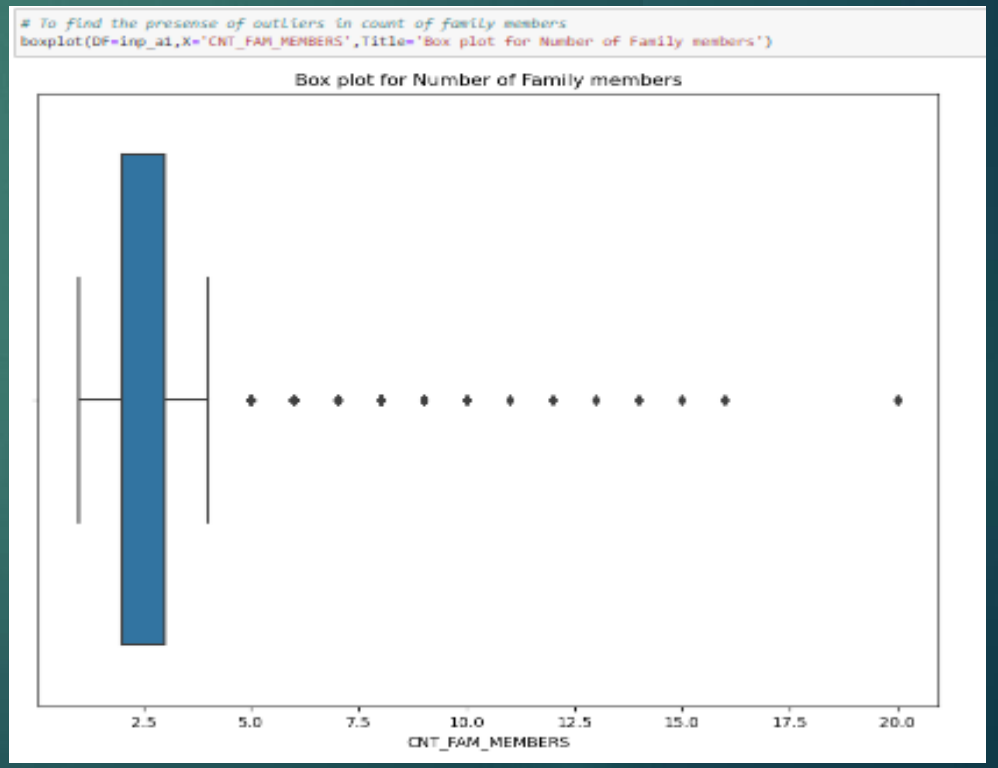



# Family members count

**Evidence**

- Majority  count of family members is less than 4 or 5.
- The count above 5 are like the joint families where its very rare to see such families and hence this is a valid case that this are Outliers

```
# To find the presense of outliers in count of family members
boxplot(DF=inp_a1,X='CNT_FAM_MEMBERS',Title='Box plot for Number of Family members')
```



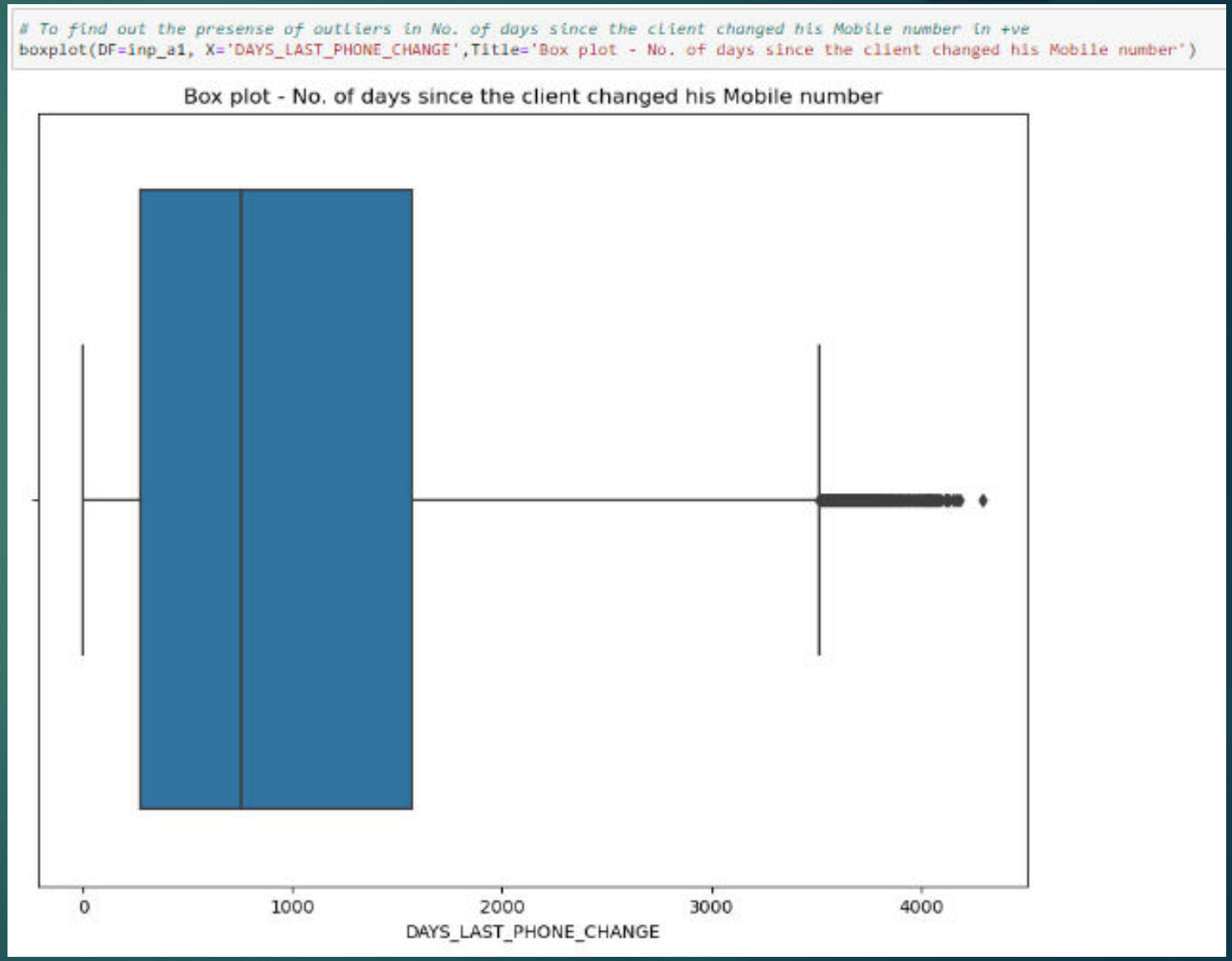
# Number of Days Client changed his Mobile Number

- These column were of negative values which tells us the how many days before number was changed
- For calculation purpose this is changed as positive variable

**Evidence :**

- This Boxplot conveys that for almost all points lies above 3000 Days , which means past 5 years or more than - same number using and the outliers are due to this reason that only few would have not changed the mobile in their life.

```
# To find out the presense of outliers in No. of days since the client changed his Mobile number in +ve
boxplot(DF=inp_a1, X='DAYS_LAST_PHONE_CHANGE',Title='Box plot - No. of days since the client changed his Mobile number')
```

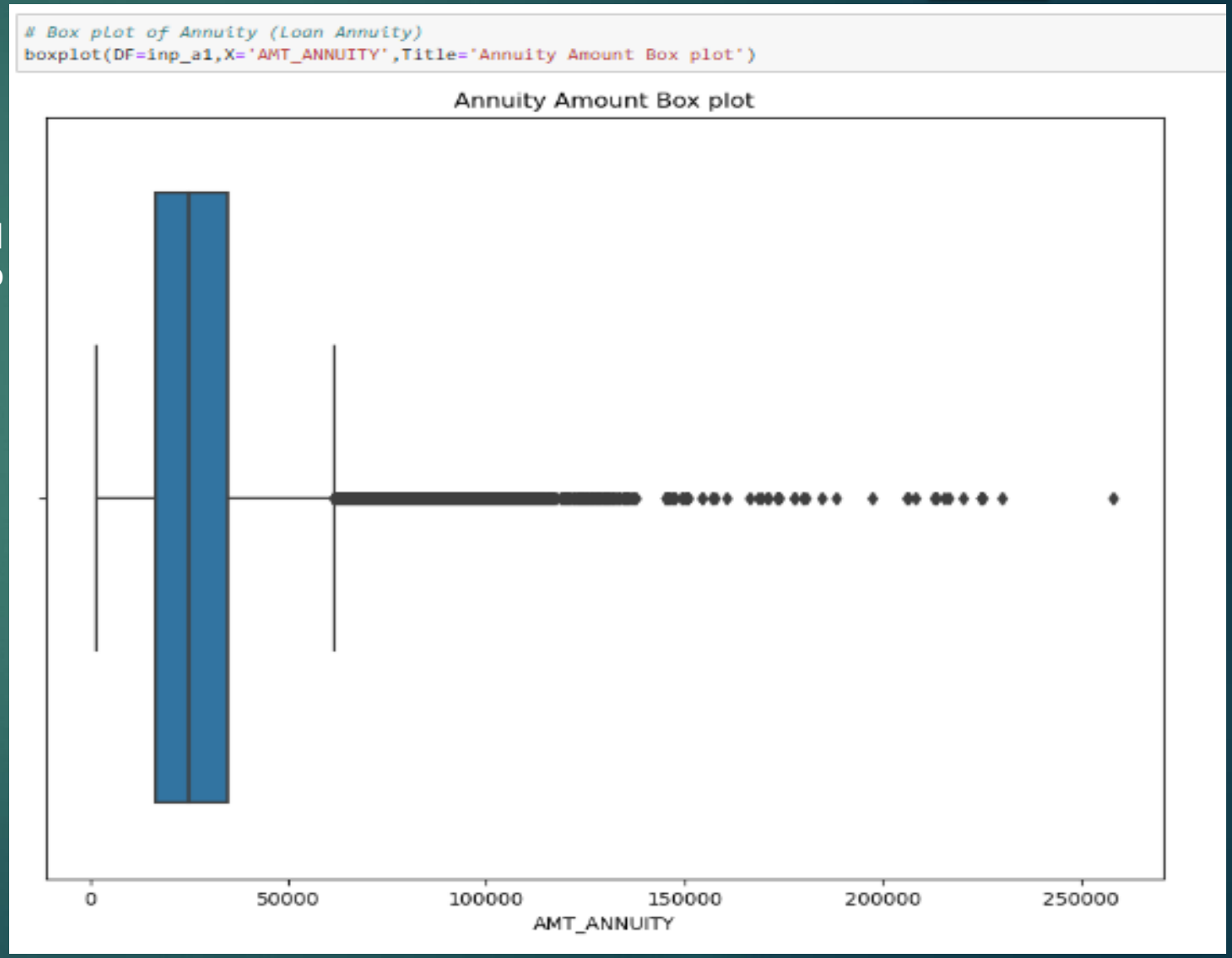


# Annuity Amount

**Evidence :**

- The Annuity seems to have 25000 as  median. there are many outliers in annuity amount
- Most of the lower class and middle  class people will be able to pay only that much as annuity and only few people in the list would have capacity to pay more.

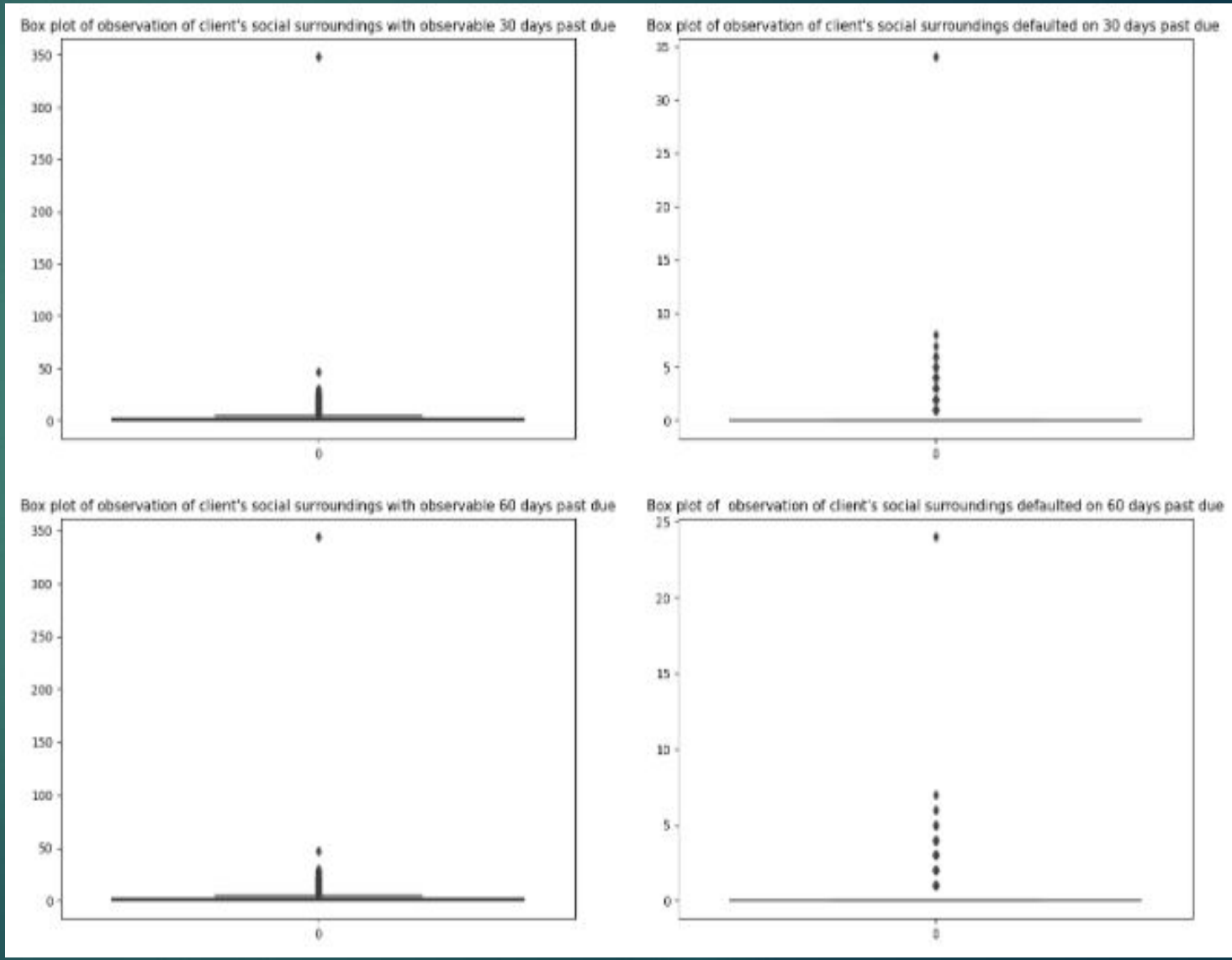
```
# Box plot of Annuity (Loan Annuity)
boxplot(DF=inp_a1,X='AMT_ANNUITY',Title='Annuity Amount Box plot')
```



# Client’s social surrounding observation

**Evidence :**

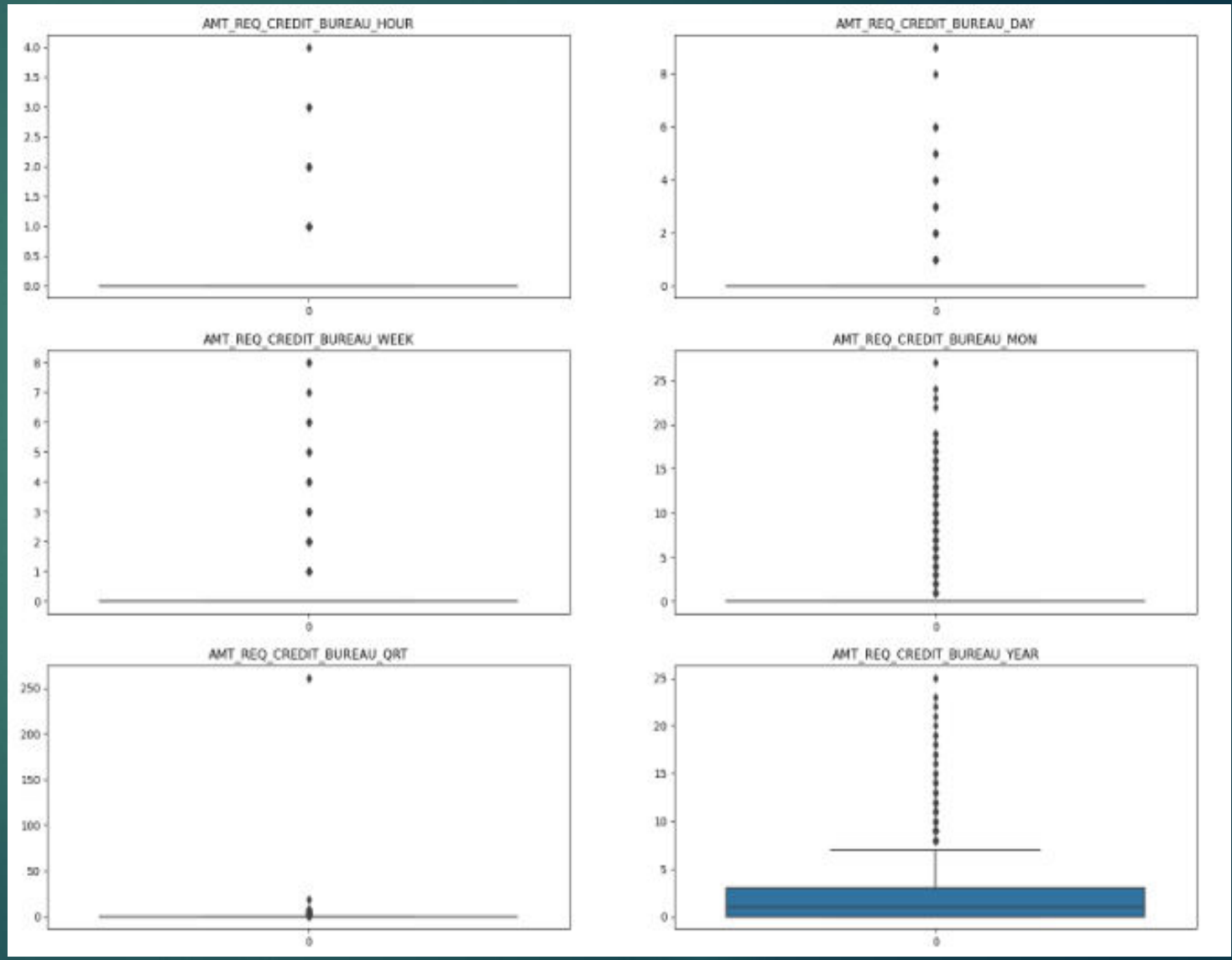
- Since all the median of column are near to 0.
- There is no proper evidence to use this column



# Number of enquiries to Credit Bureau

**Evidence :**

- Median of all column is zero.
- The 6th plot also have median close to zero along with high amount of outliers.
- The null values were filled with 0 (mode). Still it's the same as before.
- These columns will not be contributing to the Target variable

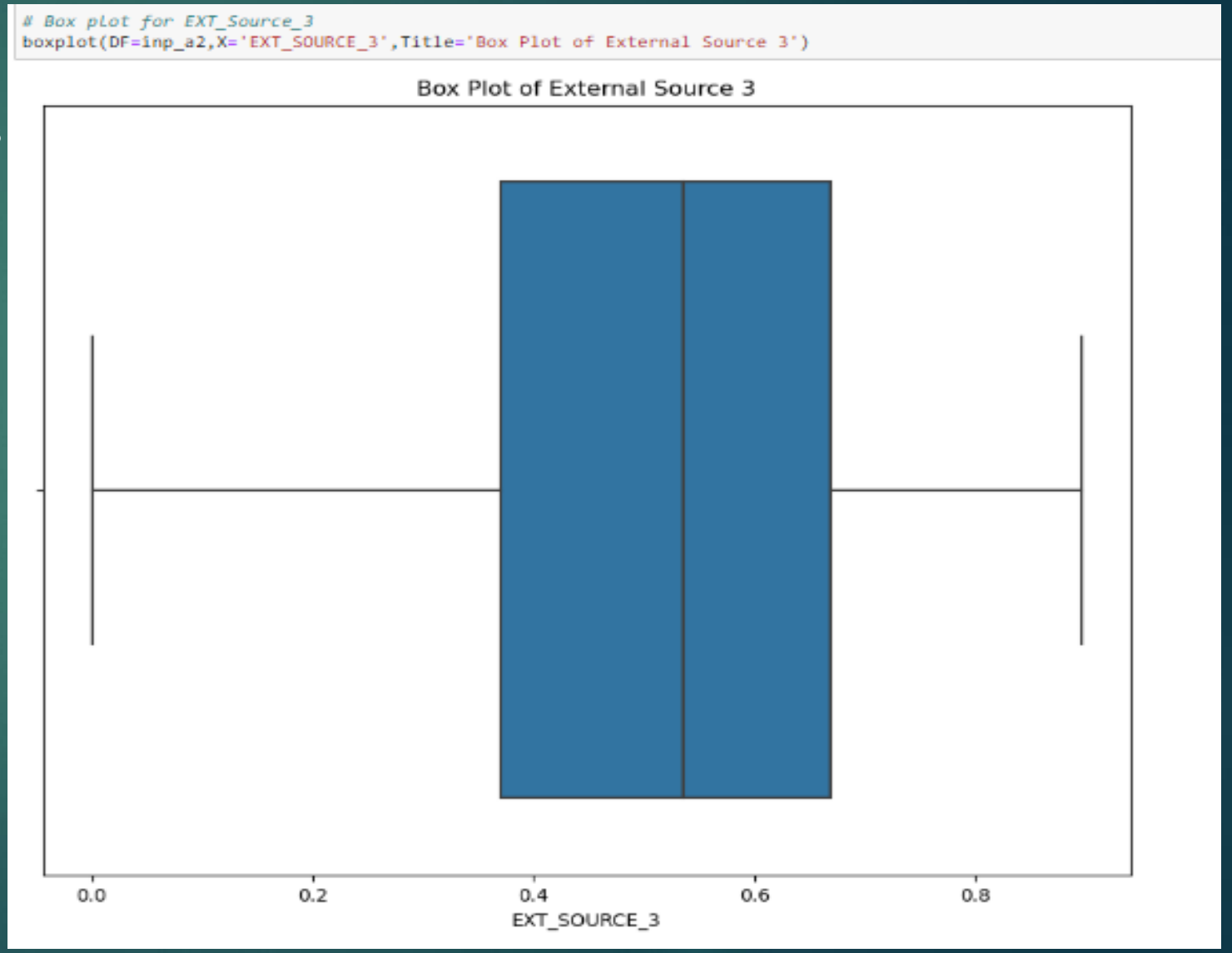


# Box Plot for External source

**Evidence :**

- Only of the external data sources would have given reviews as low and that's the reason it was treated as outlier.

```
# Box plot for EXT_Source_3
boxplot(DF=inp_a2,X='EXT_SOURCE_3',Title='Box Plot of External Source 3')
```

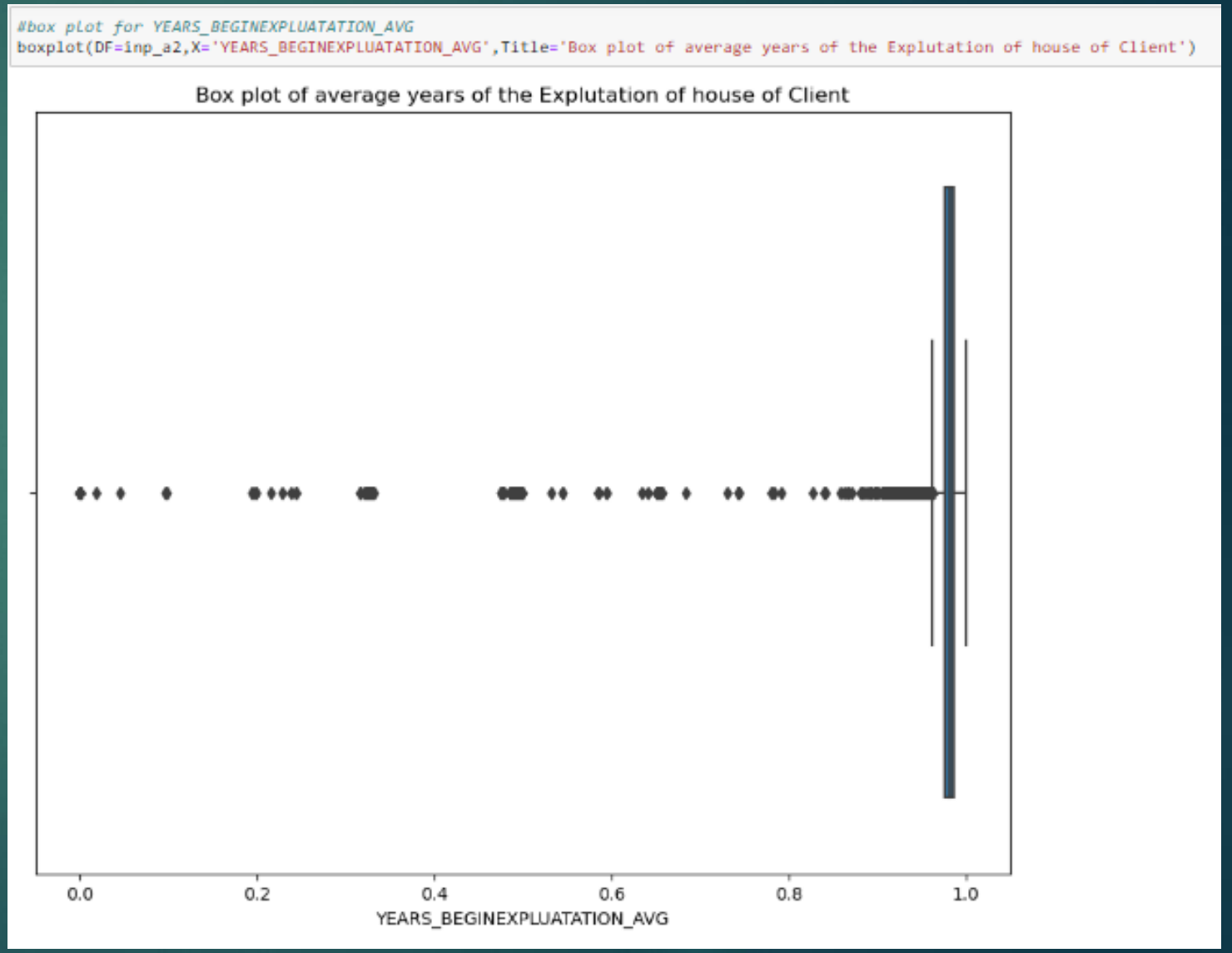


# Average Expluatation of building of Client

**Evidence**:

- As Most of data point  as plotted as outlier, so there is no use of building Expluatation.

```
#box plot for YEARS_BEGINEXPLUATATION_AVG
boxplot(DF=inp_a2,X='YEARS_BEGINEXPLUATATION_AVG',Title='Box plot of average years of the Explutation of house of Client')
```

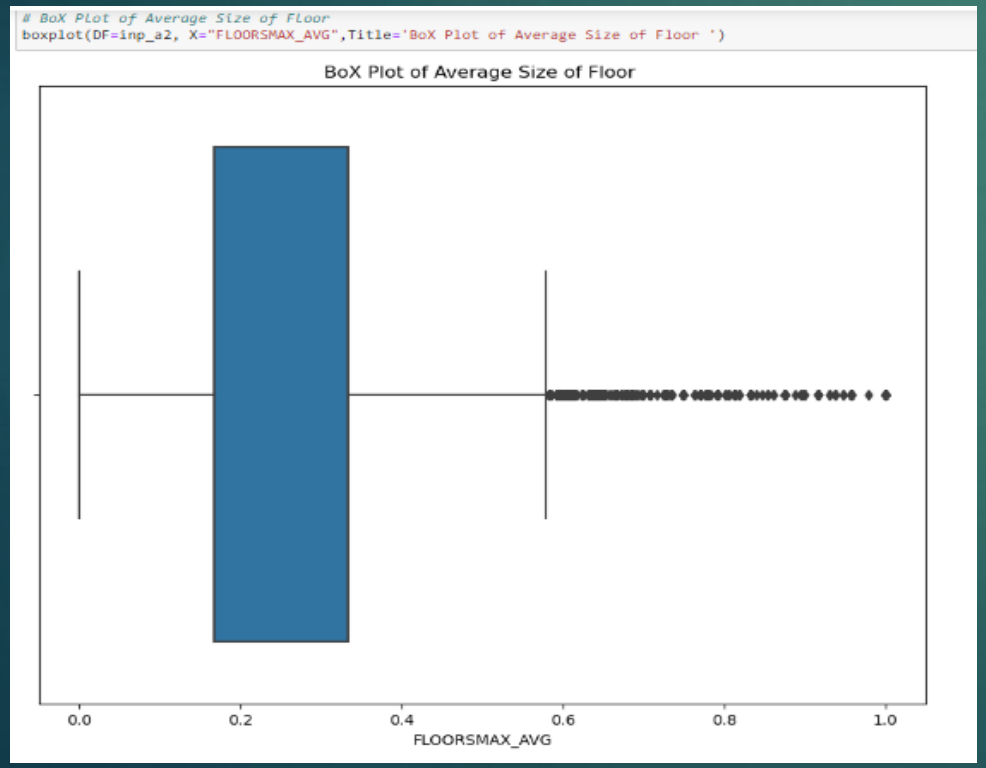


# Average Floormax (Floor Size ) Data

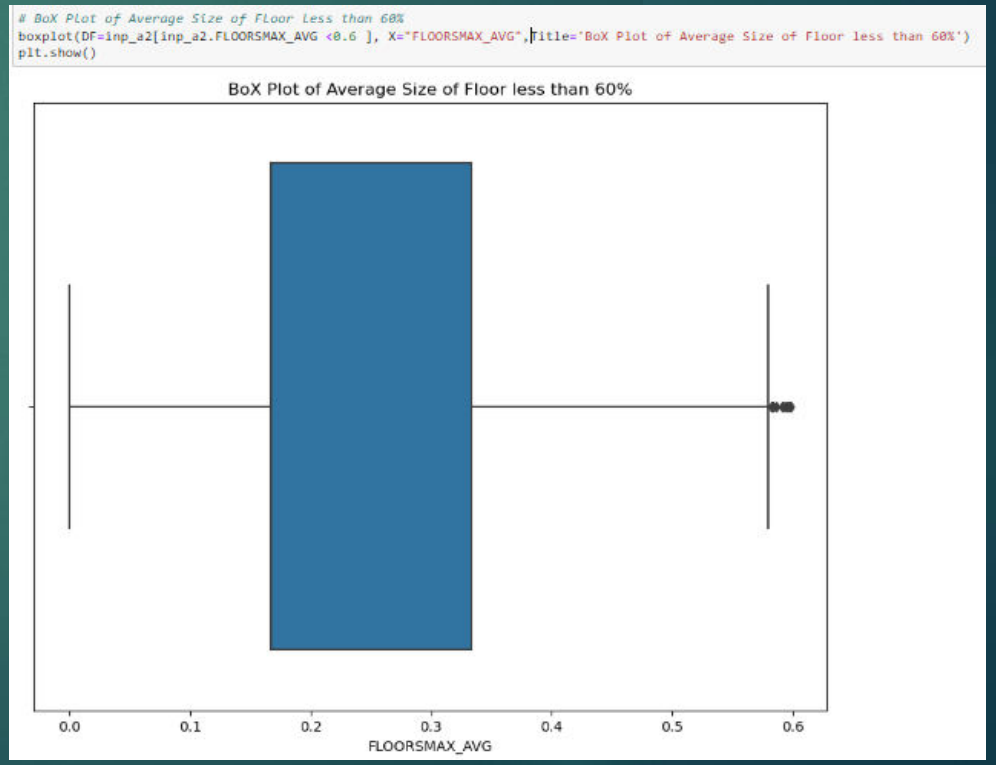
**Evidence** :

- Average floor size has more outliers since most of the houses will be of medium sizes which is depended on the monthly income and only few will be of upper class.
- Hence if we have to remove top 40% of data we are able to get this column with almost no outliers.

```
# BoX Plot of Average Size of Floor
boxplot(DF=inp_a2, X="FLOORSMAX_AVG",Title='BoX Plot of Average Size of Floor ')
```




```
# BoX Plot of Average Size of Floor less than 60%
boxplot(DF=inp_a2[inp_a2.FLOORSMAX_AVG <0.6 ], X="FLOORSMAX_AVG",Title='BoX Plot of Average Size of Floor less than 60%')
plt.show()
```

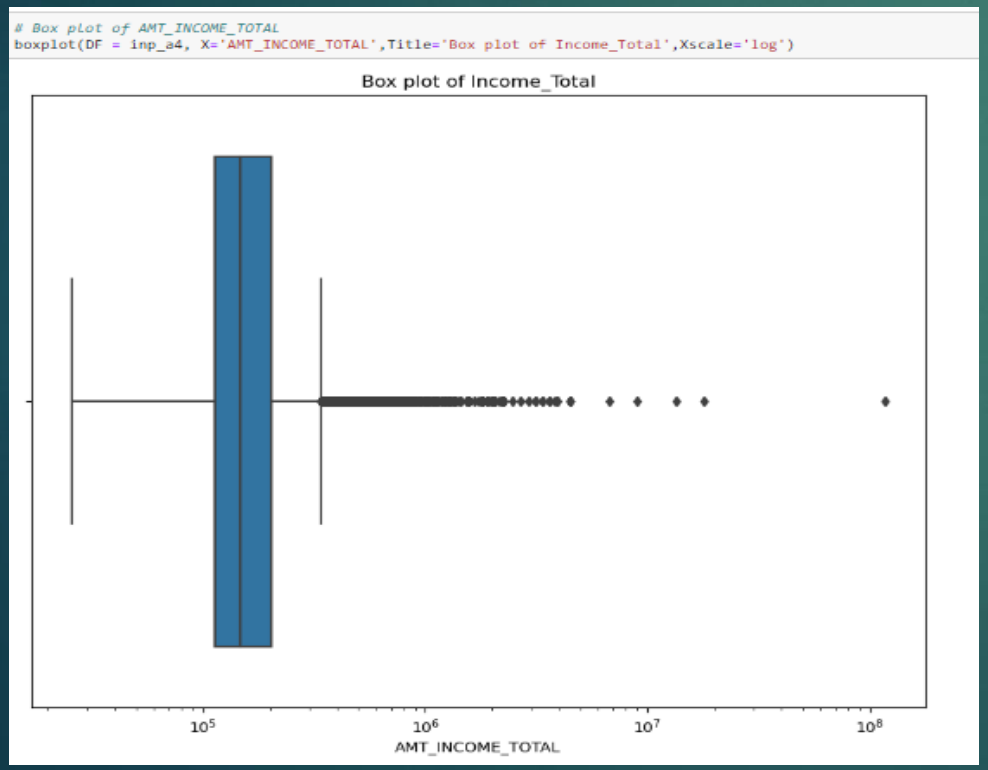


# Income of the Client

**Evidence :**

- Most of the people will be getting lower salary only (applying for loans)
- Very few only will be getting a high Salary. Hence the outlier case.

```
# Box plot of AMT_INCOME_TOTAL
boxplot(DF = inp_a4, X='AMT_INCOME_TOTAL',Title='Box plot of Income_Total',Xscale='log')
```




# Credit Amount

**Evidence** :

- Credit amount is some what having mean and median as same.

```
#Describe CREDIT
inp_a4.AMT_CREDIT.describe()
```

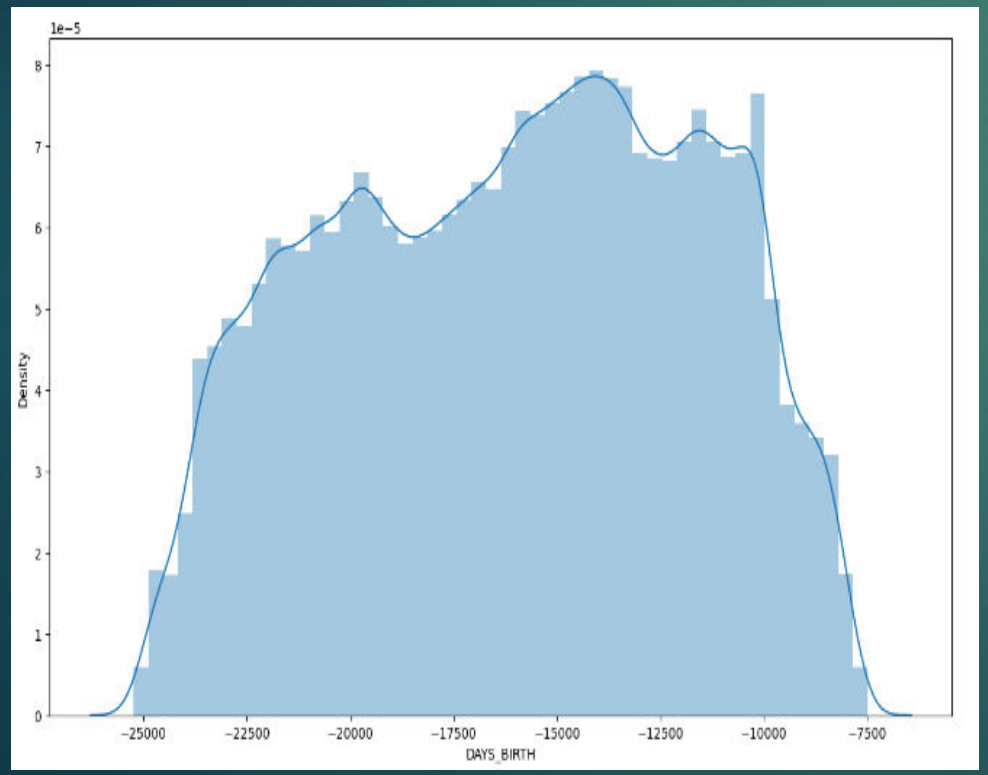
```
count     307511.00000
mean      599025.99971
std       402490.77700
min        45000.00000
25%       270000.00000
50%       513531.00000
75%       808650.00000
max      4050000.00000
Name: AMT_CREDIT, dtype: float64
```

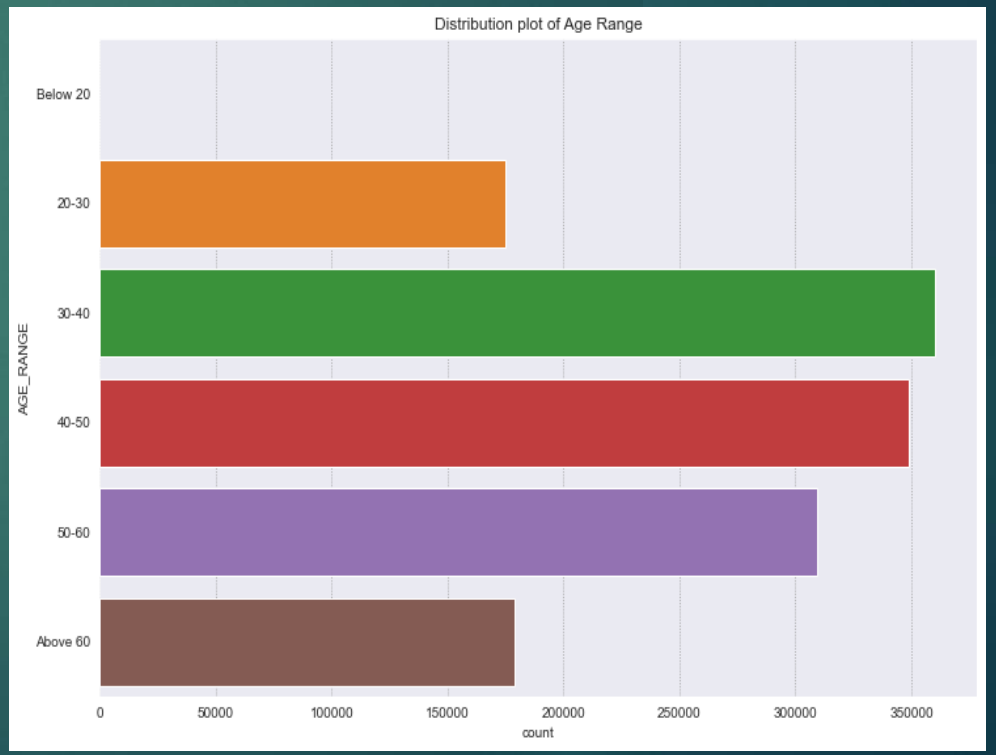
# Distribution Plots for Data Columns

# Age Range – Distribution Plots

**Evidence :**

- Values in Age column has negative values since it describes like how many days before the person has born.
- Visualization purpose we are converting it to positive values.
- Using Histogram,  grouping it into range of values



# Income Range – Distribution Plots

**Evidence :**

- Many of the people will be having income ranging between 50,000 and 2,25,000. (Image 2 )
- But it has a wide range of distribution from 25000 to 70,00,000 (Image 3 )

```
# Distribution plot of Income range
plt.figure(figsize=(12,8))
plt.title('Distribution plot of Income Range')
plt.yscale('linear')
sns.countplot(y=inp_a4.INCOME_RANGE)
plt.show()
```

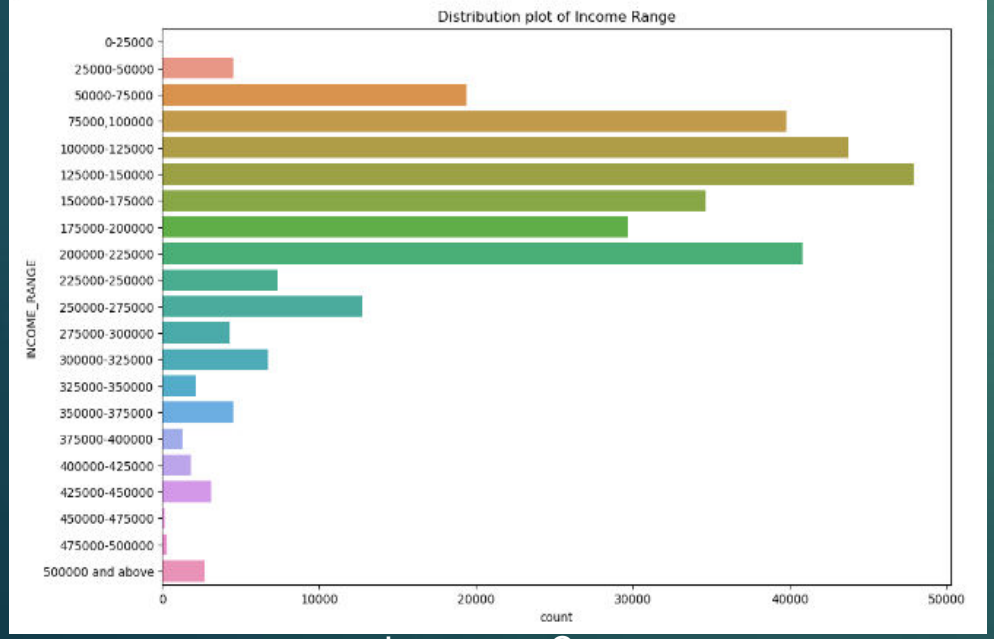



Image 2

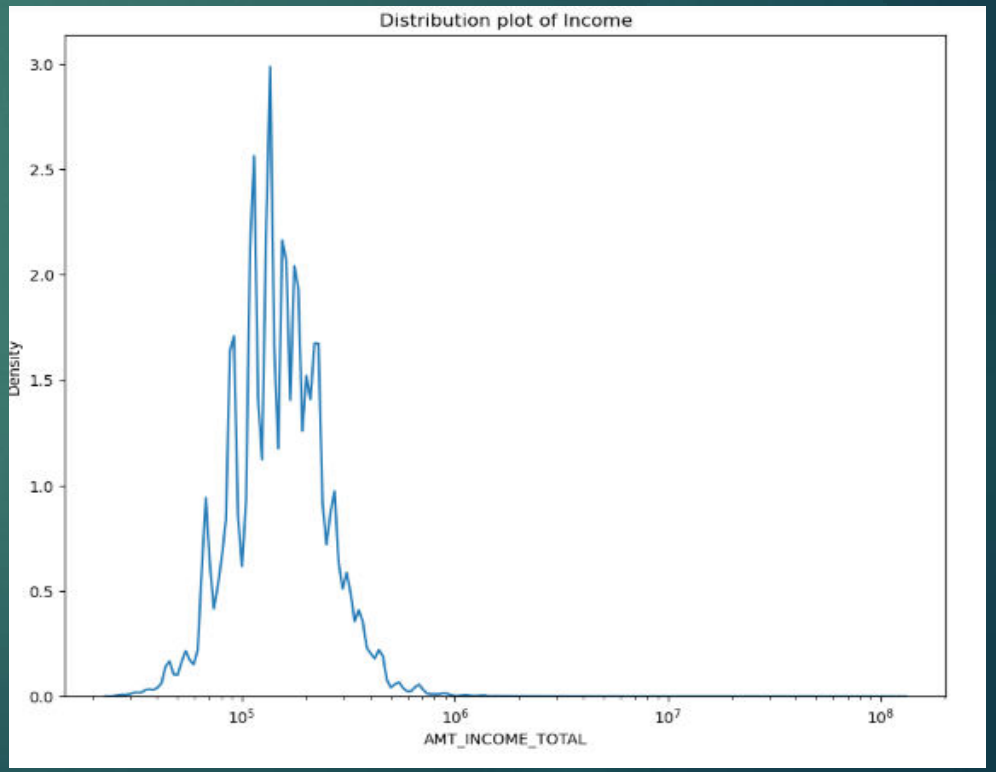



Image 3

# Credit Amount Range – Distribution Plots

**Evidence :**

- Most of the people have been credited by the loan amount ranging between 1,00,000 and 9,00,000. (Image 4 )
- But it has a wide range of distribution from 1,00,000 to 1,50,00,000. (Image 5 )

```
# Distribution plot of Credit amount range
plt.figure(figsize=(12,8))
plt.title('Distribution plot of Credit Amount Range')
plt.yscale('linear')
sns.countplot(y=inp_a4.CREDIT_RANGE)
plt.show()
```

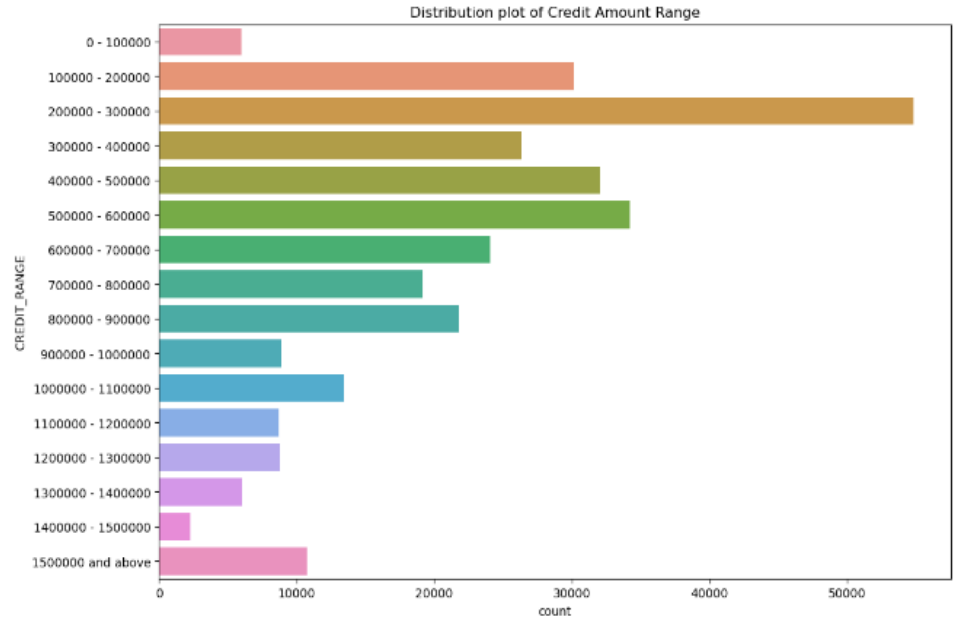



Image 4

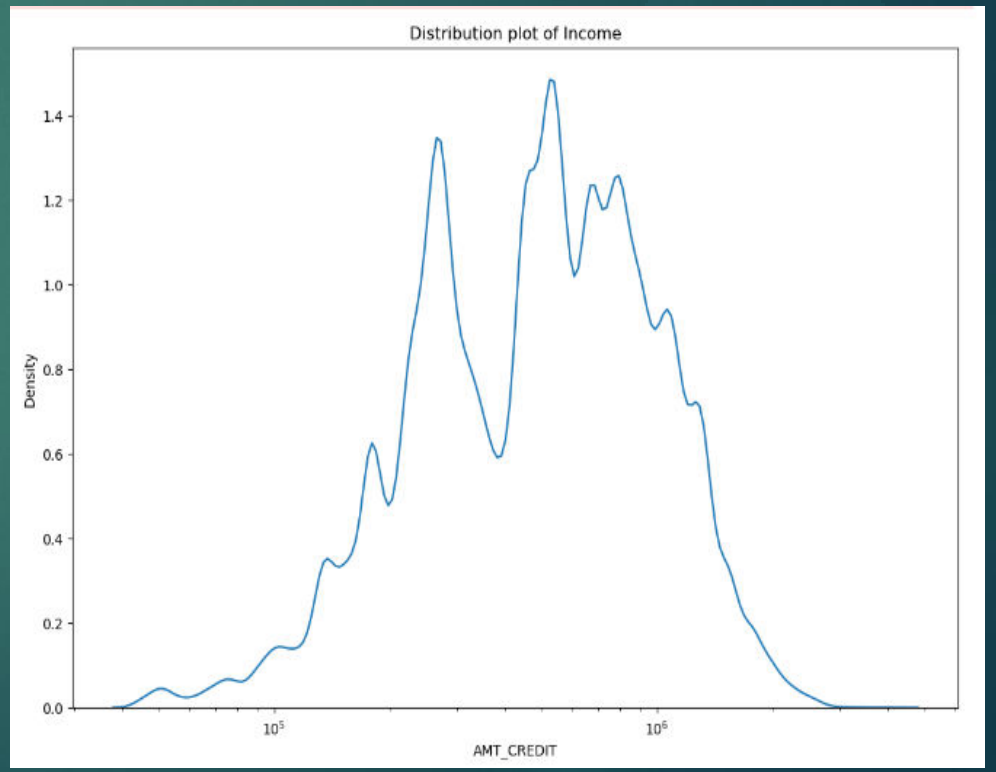


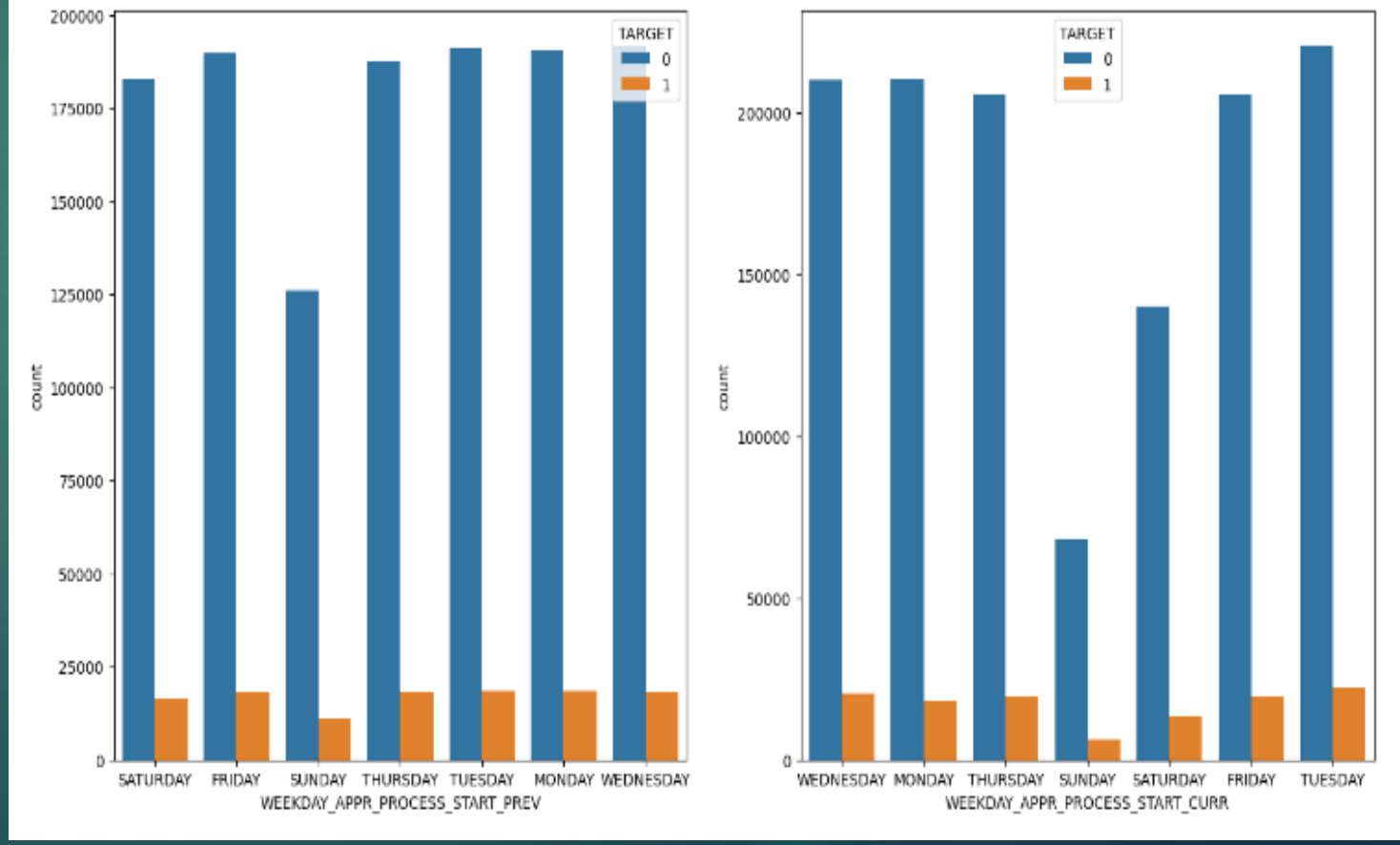

Image 5

# Day of Application for Loan

**Evidence :**

- Both Condition (Previous and Current) plot found to be same and direct correlation for target columns

```
# Count of application of loan day wise (Current and Prev)
plt.figure(figsize=(16,8))
plt.subplot(1,2,1)
sns.countplot(x = inp['WEEKDAY_APPR_PROCESS_START_PREV'],hue=inp.TARGET)
plt.subplot(1,2,2)
sns.countplot(x = inp['WEEKDAY_APPR_PROCESS_START_CURR'],hue=inp.TARGET)
plt.show()
```



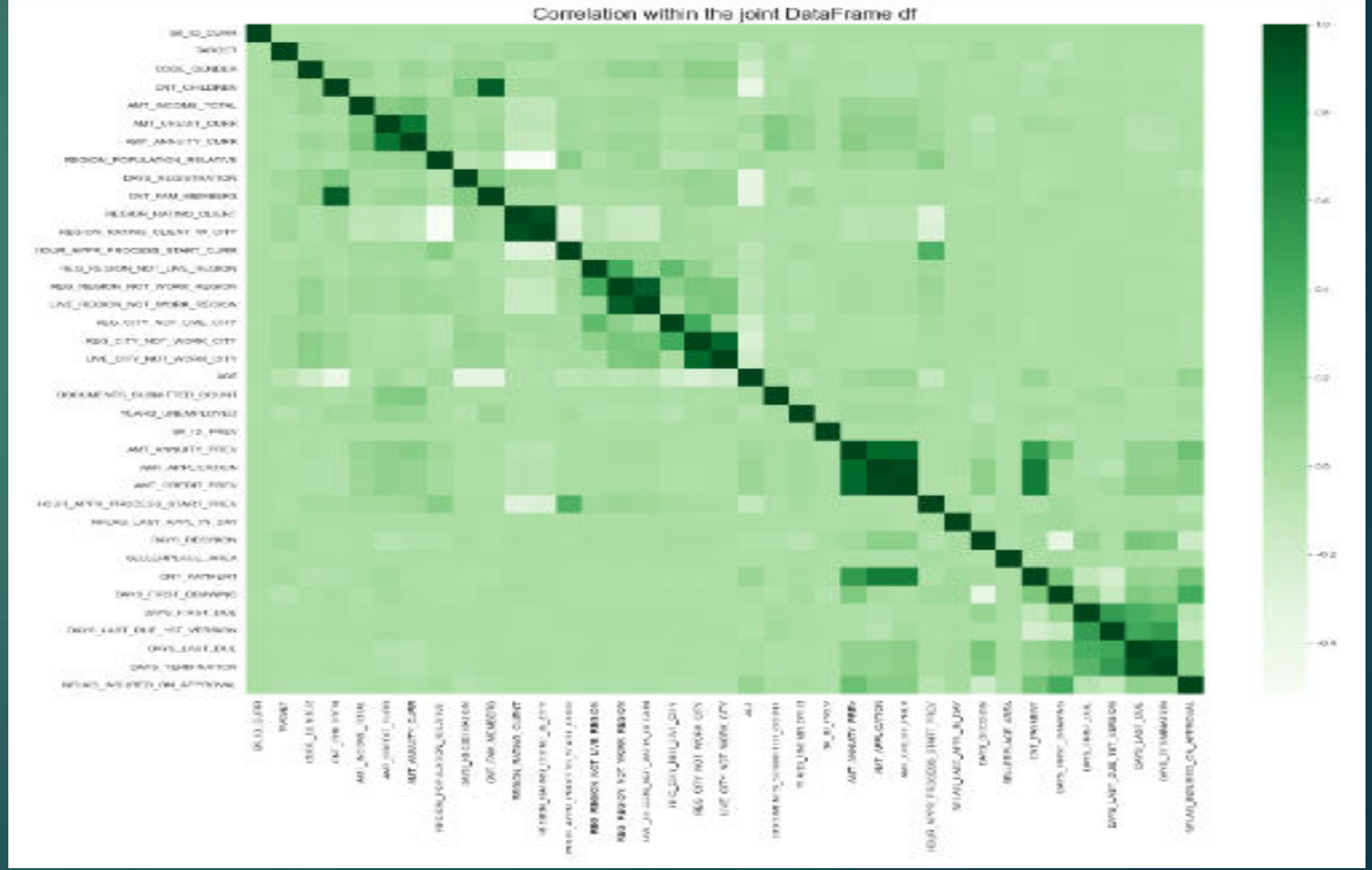
# Correlation with Joint Data Frame

**Evidence :**

- There is greater correlation between the region living status in the center of the heat map.
- Target depends on Count of family members, Region rating and days registration(little).
- The Amount annuity also depends on Credit amount and income and vice versa which makes more sensible.

```
# Correlation of the total joint dataframe
plt.figure(figsize=(20,15))
sns.heatmap(data=inp.corr(), cmap = 'Greens')
plt.title('Correlation within the joint DataFrame df',fontdict={'fontsize': 20})
plt.show()
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_10492\1679585244.py:3: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  sns.heatmap(data=inp.corr(), cmap = 'Greens')
```



# Univariate Analysis with Target Column

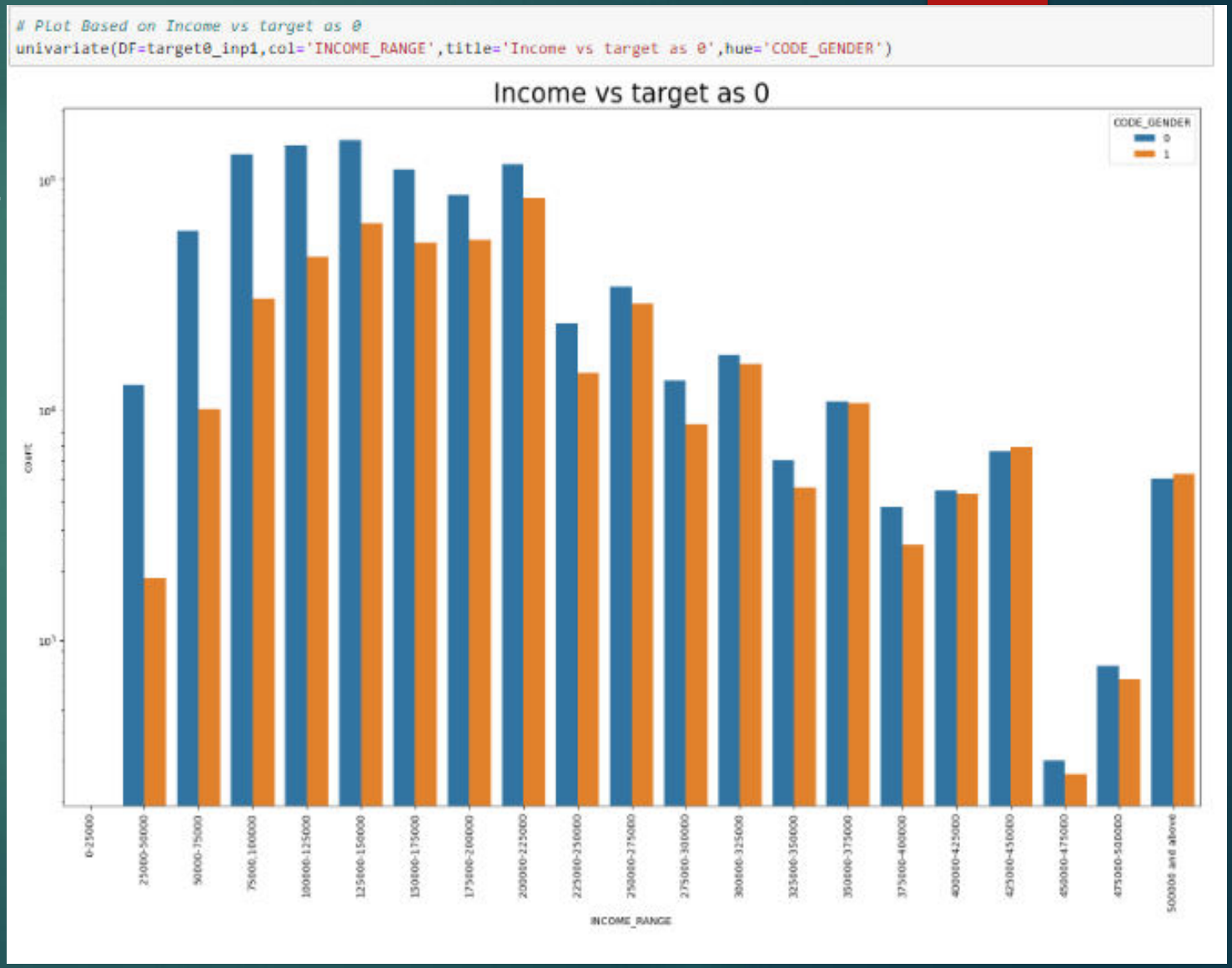
# Income range with Gender Wise

**Evidence :**

1. Seems like the Female gender having high counts and this can be expected since the data has high Female counts like twice as Male count.
2. Both Male and Female gets same, when Salary gets higher
3. The count gets reduced since for higher salary people will be able to repay the loan.

Notification: Code 0 – Female
1 – Male

```
# Plot Based on Income vs target as 0
univariate(DF=target0_inp1,col='INCOME_RANGE',title='Income vs target as 0',hue='CODE_GENDER')
```



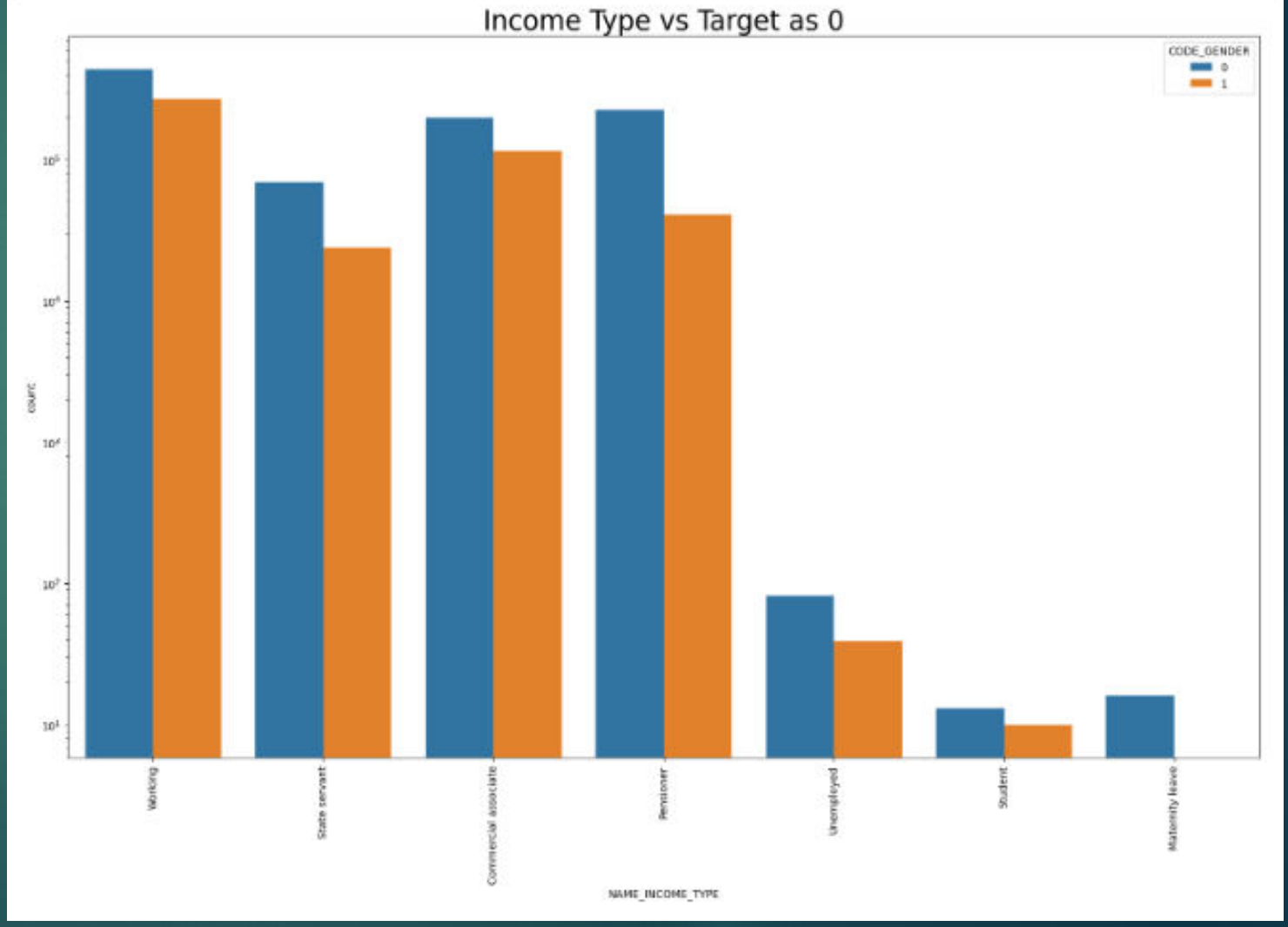
# Income Type VS Target as 0

**Evidence** :

1. State Servant found 4th largest in graph and the chances of repaying difficulties are low (They might get many benefits from Government ).
2. For student loans its not easily predictable, only prediction we could do is with their current education status.
3. Pensioner the Income type is high and most of Pensioner will be aged people and spent less amount for survival.

Notification: Code 0 – Female
1 – Male

```
# Plot Based on NAME_INCOME_TYPE vs target as 0
univariate(DF=target0_inp1,col='NAME_INCOME_TYPE',title='Income Type vs Target as 0',hue='CODE_GENDER')
```



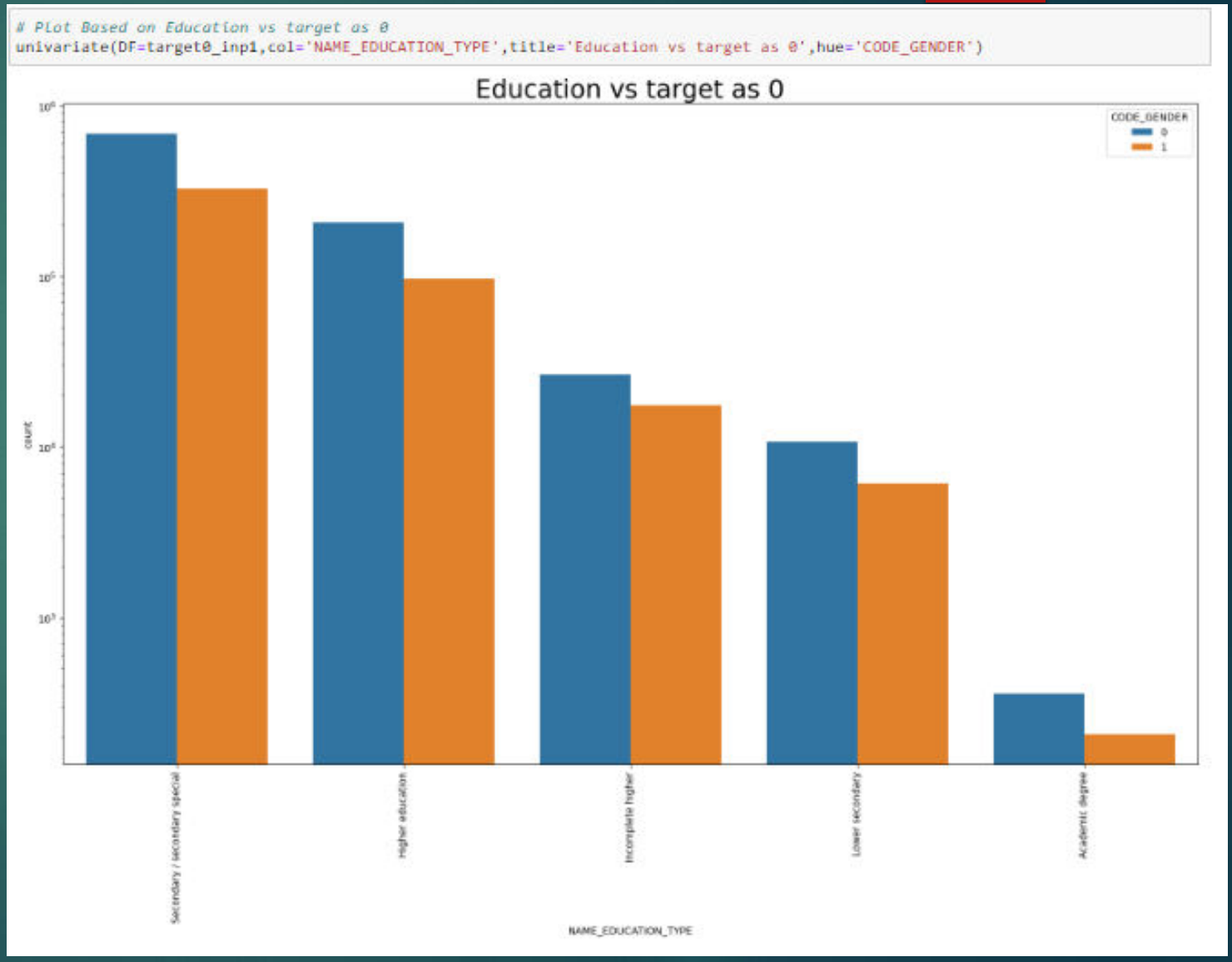
# Education VS Target as 0

**Evidence** :

1. Repayment difficulties are low for the people studying or completed Academic degree.
2. While the Secondary degree will have difficulties higher than Acedemic degree which is very much relatable and acceptable.

Notification: Code 0 – Female
1 – Male

```
# Plot Based on Education vs target as 0
univariate(DF=target0_inp1,col='NAME_EDUCATION_TYPE',title='Education vs target as 0',hue='CODE_GENDER')
```



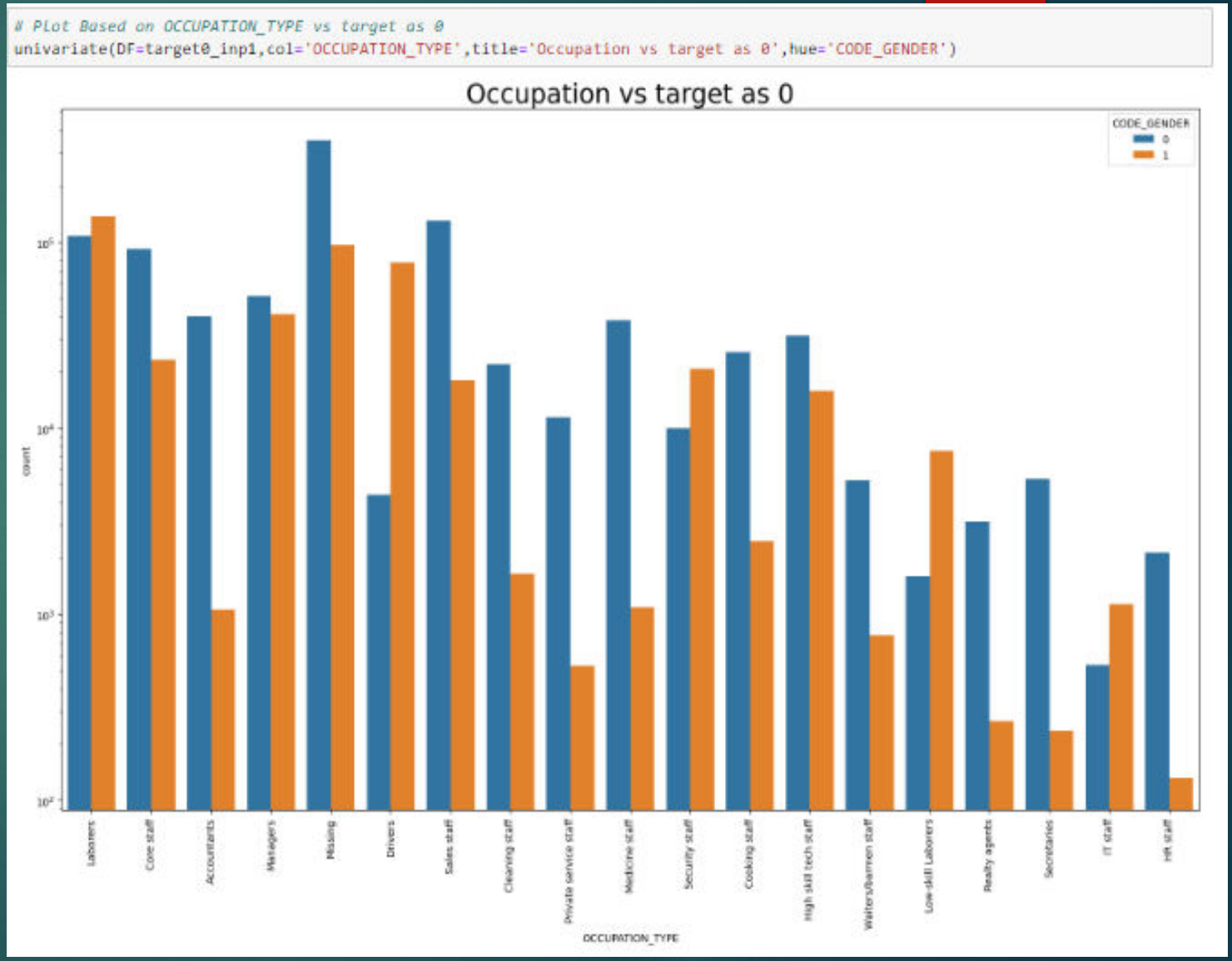
# Occupation VS Target as 0

**Evidence** :

1. Female have  high Missing Occupation because of  homemaker
2. Drivers there might be more men in that occupation than women.
3. The Laborers, Sales staff (Highest ) have difficulties in repaying since their  annual income be less when compared to others.

Notification: Code 0 – Female
1 – Male

```
# Plot Based on OCCUPATION_TYPE vs target as 0
univariate(DF=target0_inp1,col='OCCUPATION_TYPE',title='Occupation vs target as 0',hue='CODE_GENDER')
```



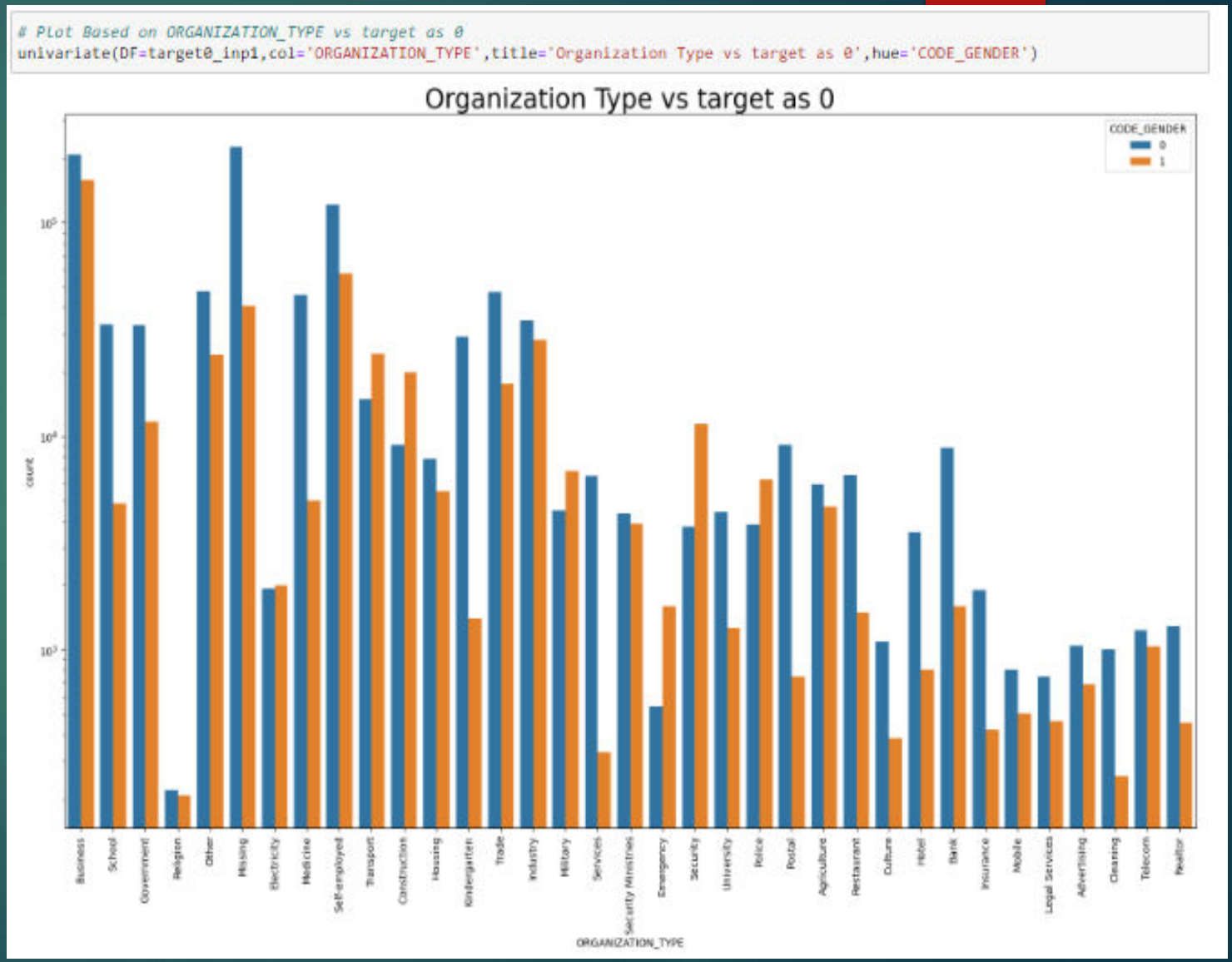
# Organization Type VS Target as 0

**Evidence:**

- Business type found highest difficult to pay Loan as Business type is not stable income based on every month income.
- Self-Employed is same case as business also.
- For other kind where they can get stable income to repay loan

Notification: Code 0 – Female
1 – Male

```
# Plot Based on ORGANIZATION_TYPE vs target as 0
univariate(DF=target0_inp1,col='ORGANIZATION_TYPE',title='Organization Type vs target as 0',hue='CODE_GENDER')
```



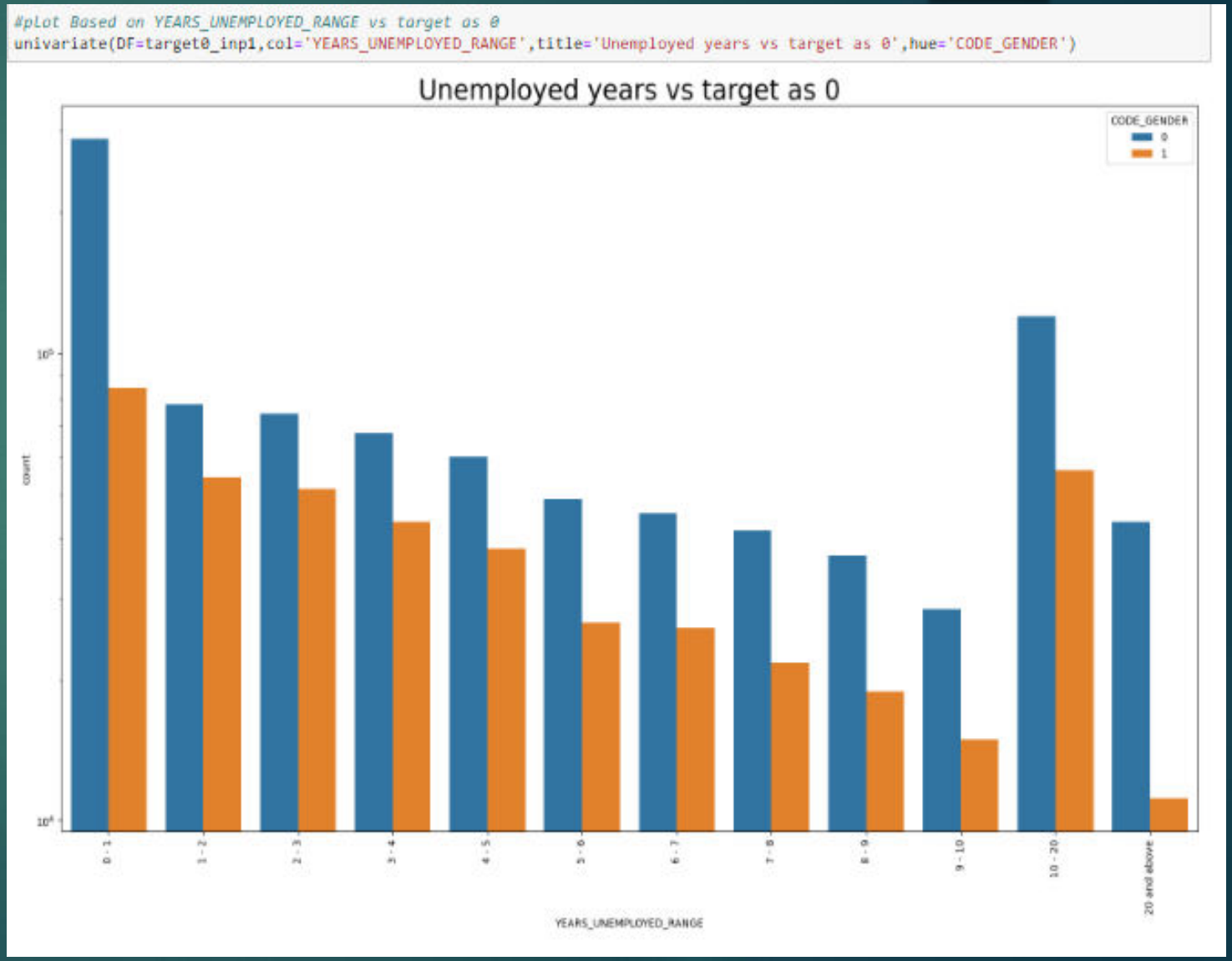
# Unemployed Year Type VS Target as 0

**Evidence :**

- In the Year column – less than 1 year has the highest count which tells difficult to repay loan
- In this graph found , female are more in unemployed category

Notification: Code 0 – Female
1 – Male

```
#plot Based on YEARS_UNEMPLOYED_RANGE vs target as 0
univariate(DF=target0_inp1,col='YEARS_UNEMPLOYED_RANGE',title='Unemployed years vs target as 0',hue='CODE_GENDER')
```



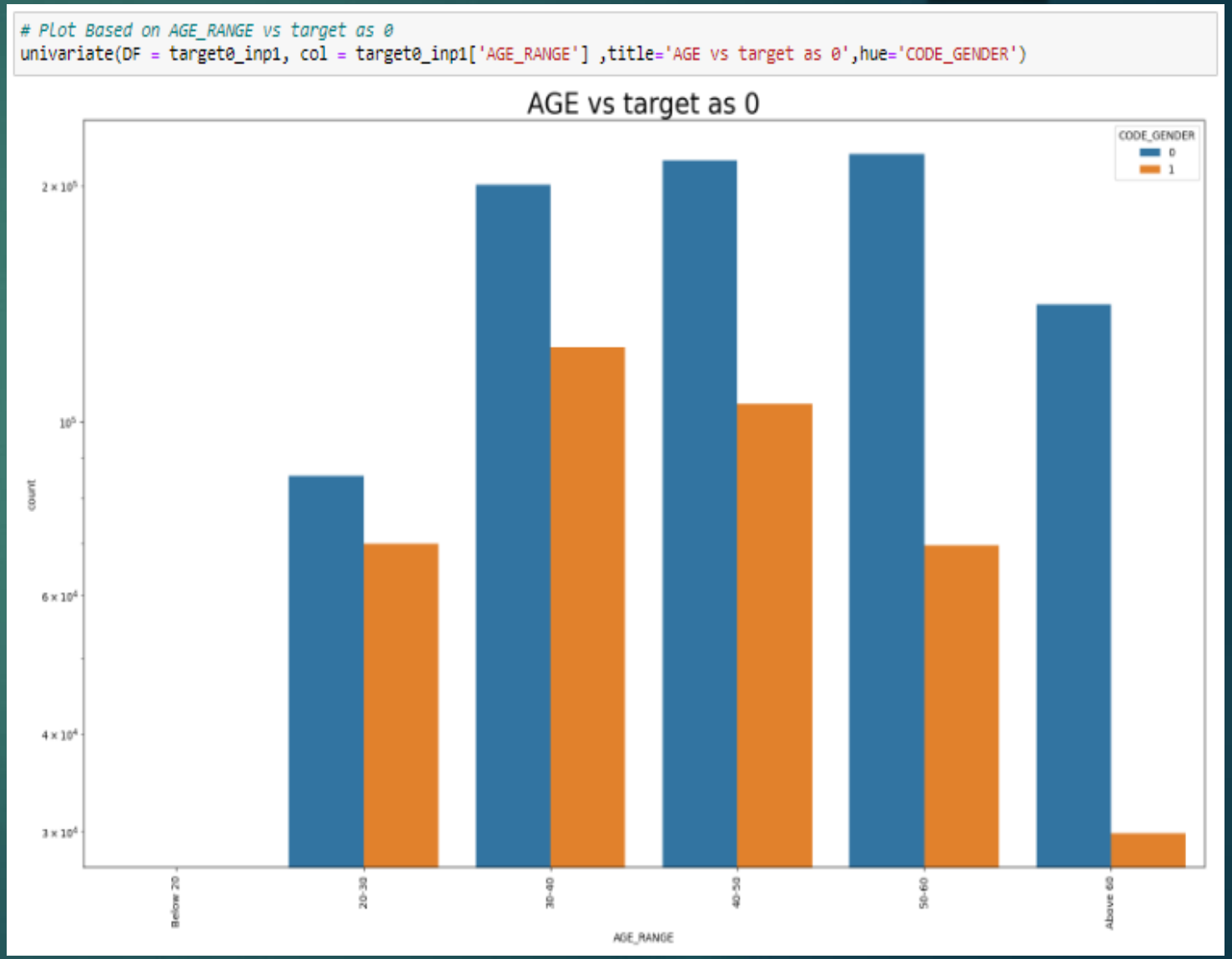
# Age VS Target as 0

**Evidence :**

- Age plot defines that Middle aged person have high risk of paying loan
- In this graph, both male and female has same percentage with Age range in difficult to pay loan

Notification: Code 0 – Female
1 – Male

```
# Plot Based on AGE_RANGE vs target as 0
univariate(DF = target0_inp1, col = target0_inp1['AGE_RANGE'] ,title='AGE vs target as 0',hue='CODE_GENDER')
```



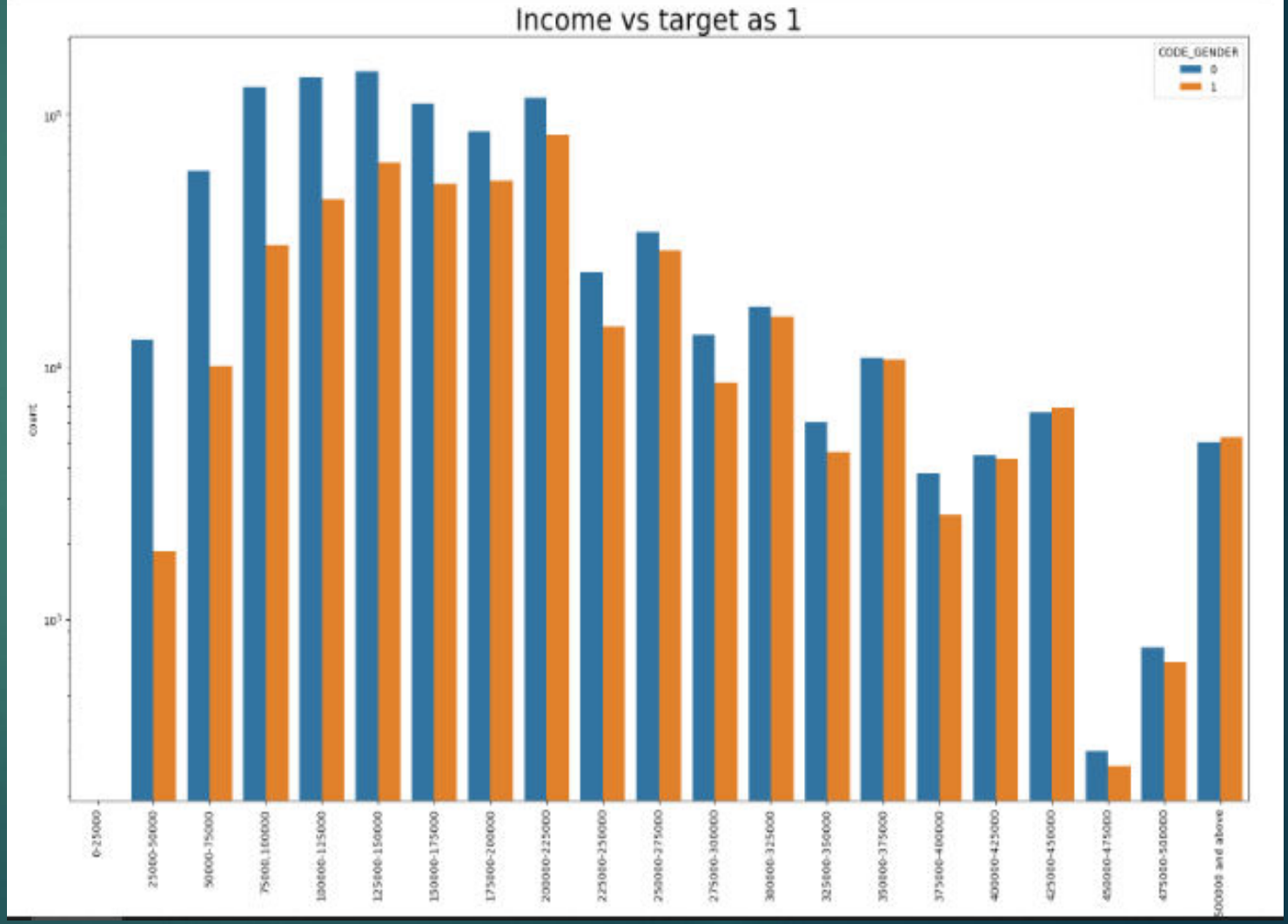
# Income VS Target as 1

**Evidence:**

- The Highest category who earn  between 1Lakh and 2.5Lakh,  They will repay loan successfully and have more time to pay loan
- People above 4 lakhs do need to loan

Notification: Code 0 – Female
1 – Male

```
# Plot Based on Income vs target as 1
univariate(DF=target1_inp1,col='INCOME_RANGE',title='Income vs target as 1',hue='CODE_GENDER')
```



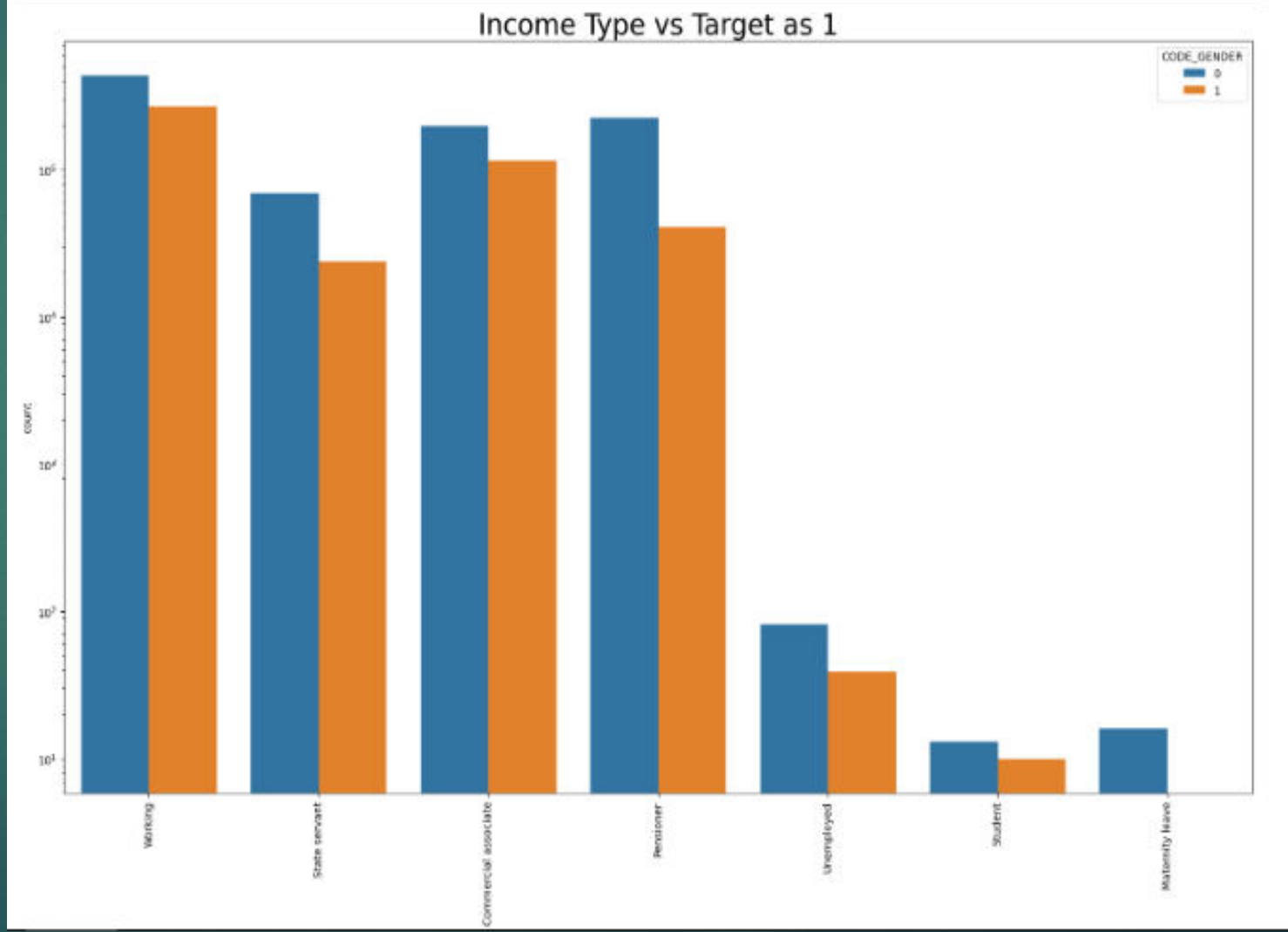
# Income Type VS Target as 1

**Evidence:**

- Working people have highest that they repay because of stable income.
- Other hand unemployed found lowest difficult to pay loan back

Notification: Code 0 – Female
1 – Male

```
# Plot Based on NAME_INCOME_TYPE vs target as 1
univariate(DF=target1_inp1,col='NAME_INCOME_TYPE',title='Income Type vs Target as 1',hue='CODE_GENDER')
```

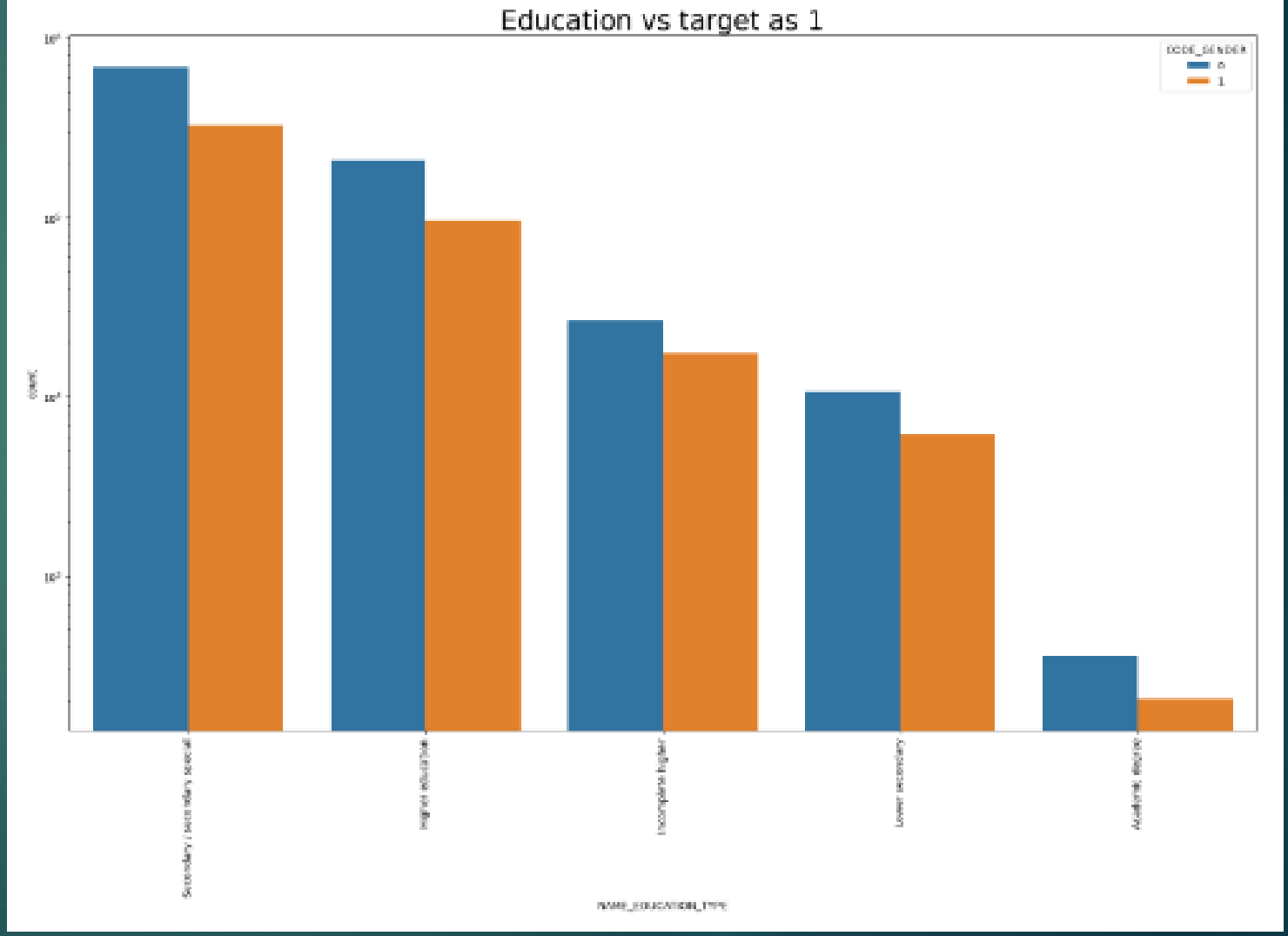


# Education VS Target as 1

**Evidence :**

- The people applying for loan would have been at least completed secondary

```
# Plot Based on Education vs target as 1
univariate(DF=target1_inp1,col='NAME_EDUCATION_TYPE',title='Education vs target as 1',hue='CODE_GENDER')
```

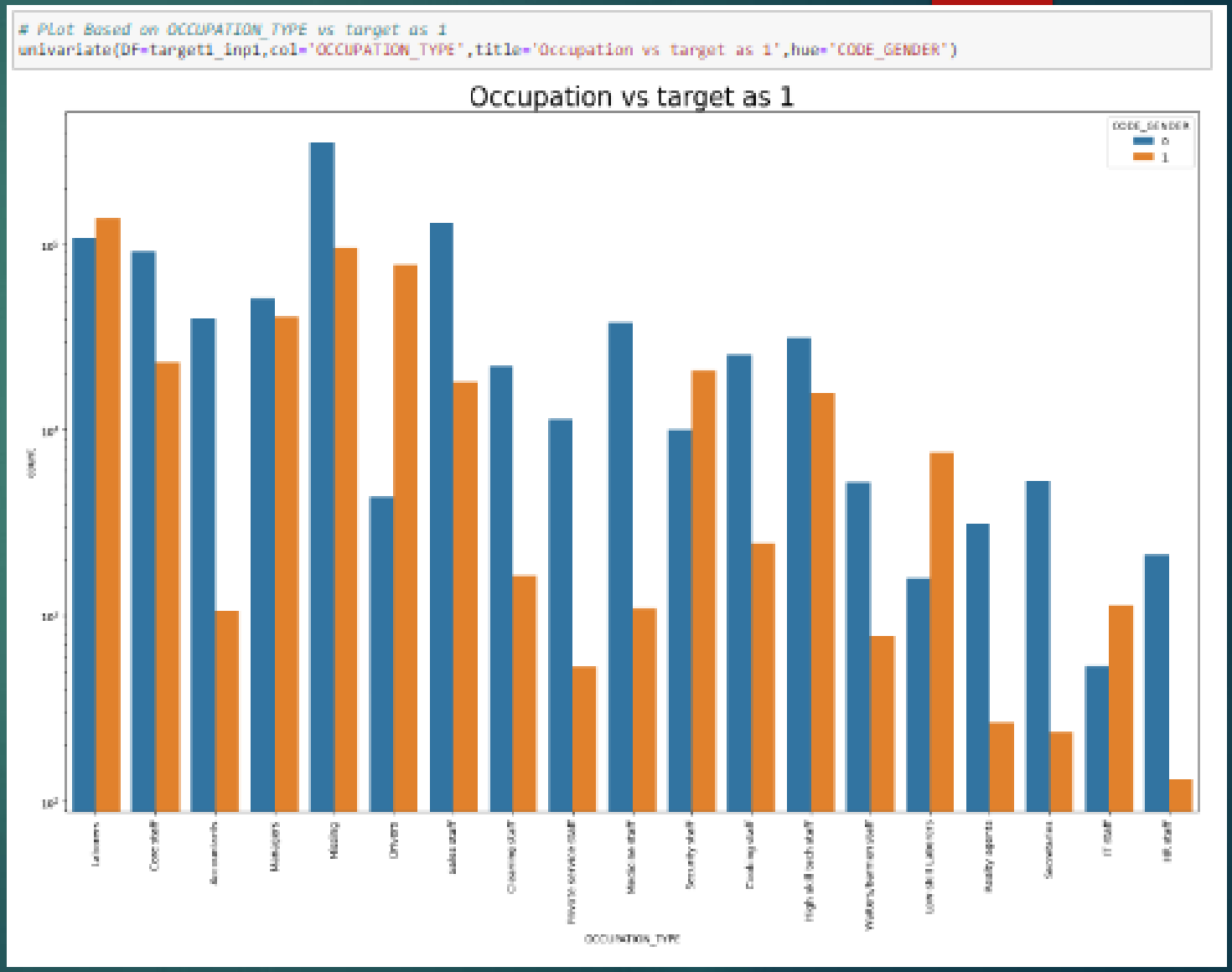


# Occupation VS Target as 1

**Evidence** :

- In this plot found lot of missing data, so neglecting this missing data
- This plot is same as Occupation VS Target as 0

```
# Plot Based on OCCUPATION_TYPE vs target as 1
univariate(DF=target1_inp1,col='OCCUPATION_TYPE',title='Occupation vs target as 1',hue='CODE_GENDER')
```

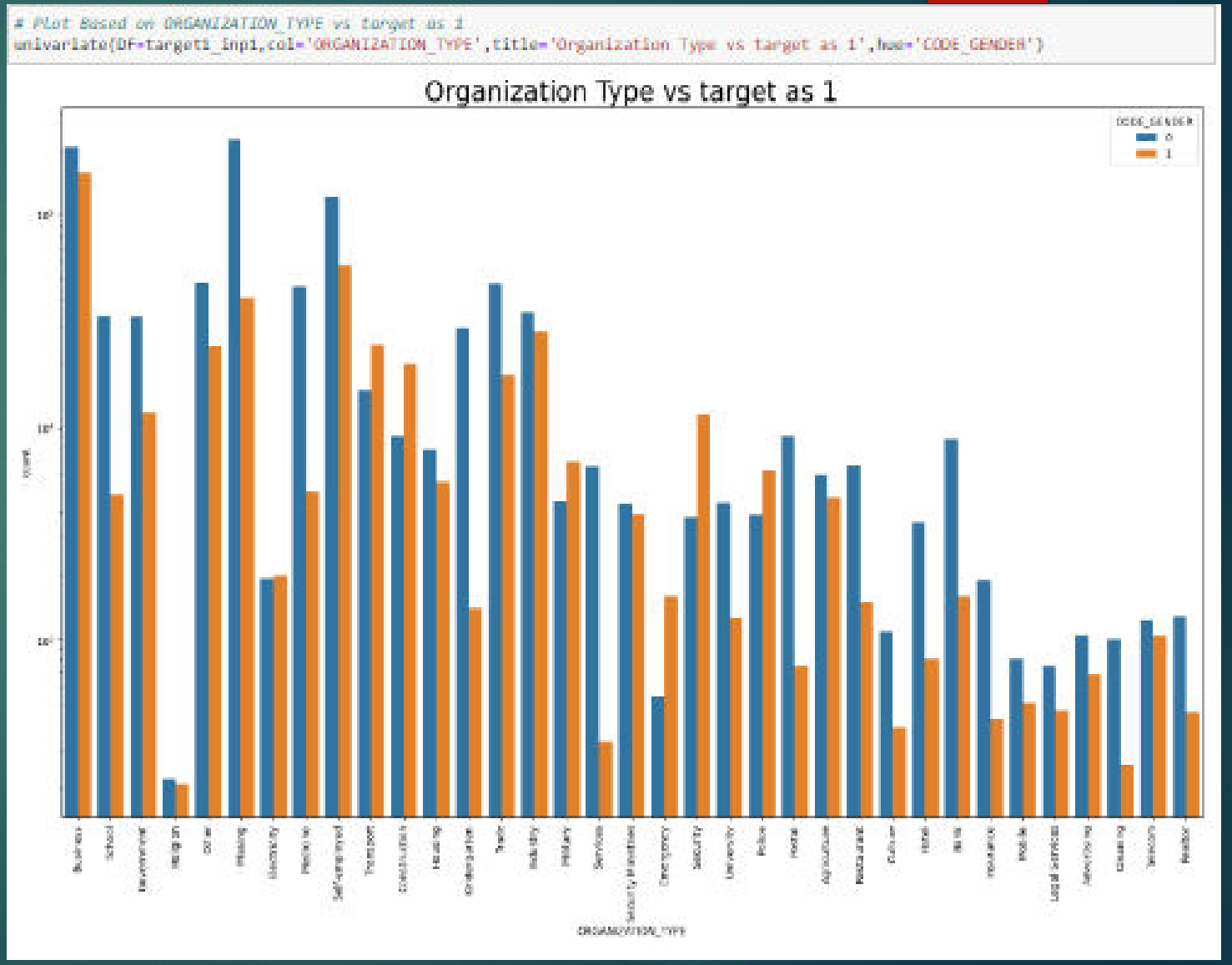


# Organization Type VS Target as 1

**Evidence :**

- Business people have high chance to pay loan.
- People working in Industry will  be getting a fixed income and they find difficult to repay loan

```
# Plot Based on ORGANIZATION_TYPE vs target as 1
univariate(DF=target1_Inp1,col='ORGANIZATION_TYPE',title='Organization Type vs target as 1',hue='CODE_GENDER')
```

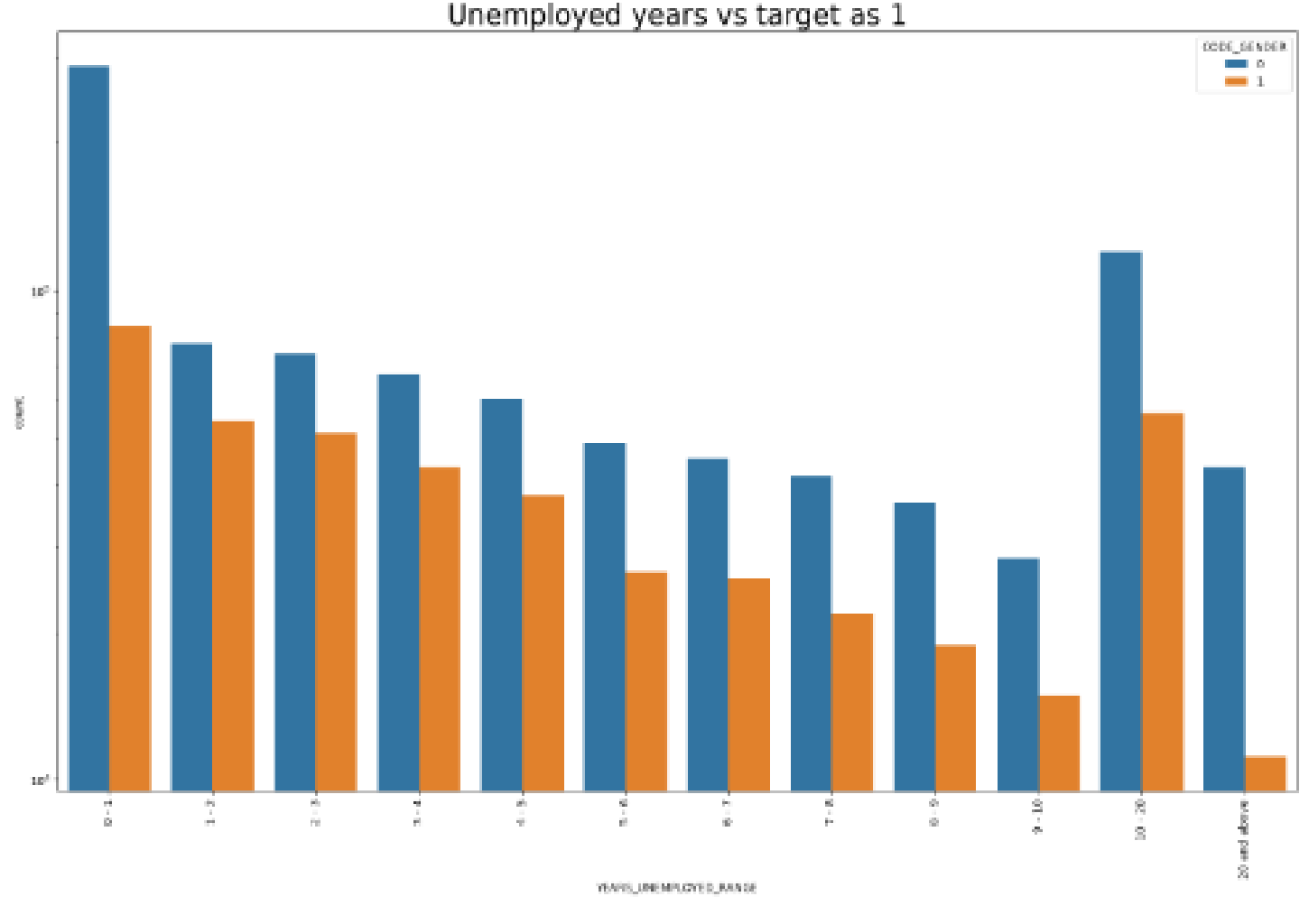


# Unemployed Year VS Target as 1

**Evidence :**

- Plot defines that increase in number of years increase in difficulty in repaying the loan

```
# Plot Based on YEARS_UNEMPLOYED_RANGE vs target as 1
univariate(DF=target1_inp1,col='YEARS_UNEMPLOYED_RANGE',title='Unemployed years vs target as 1',hue='CODE_GENDER')
```

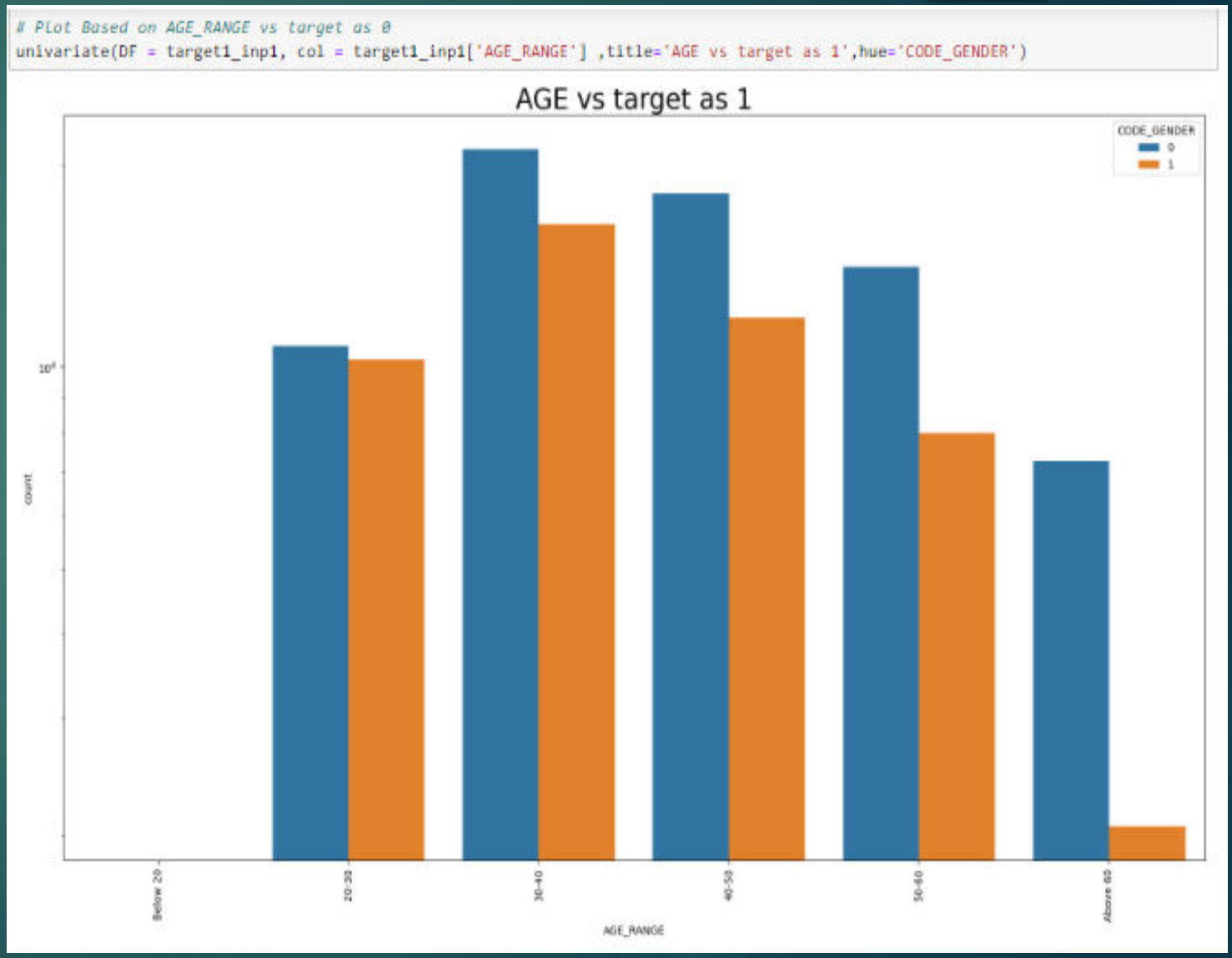


# Age VS Target as 1

**Evidence :**

- Plot defines that Middle age people has highest count in graph and they have low need of loan

```
# Plot Based on AGE_RANGE vs target as 0
univariate(DF = target1_inp1, col = target1_inp1['AGE_RANGE'] ,title='AGE vs target as 1',hue='CODE_GENDER')
```

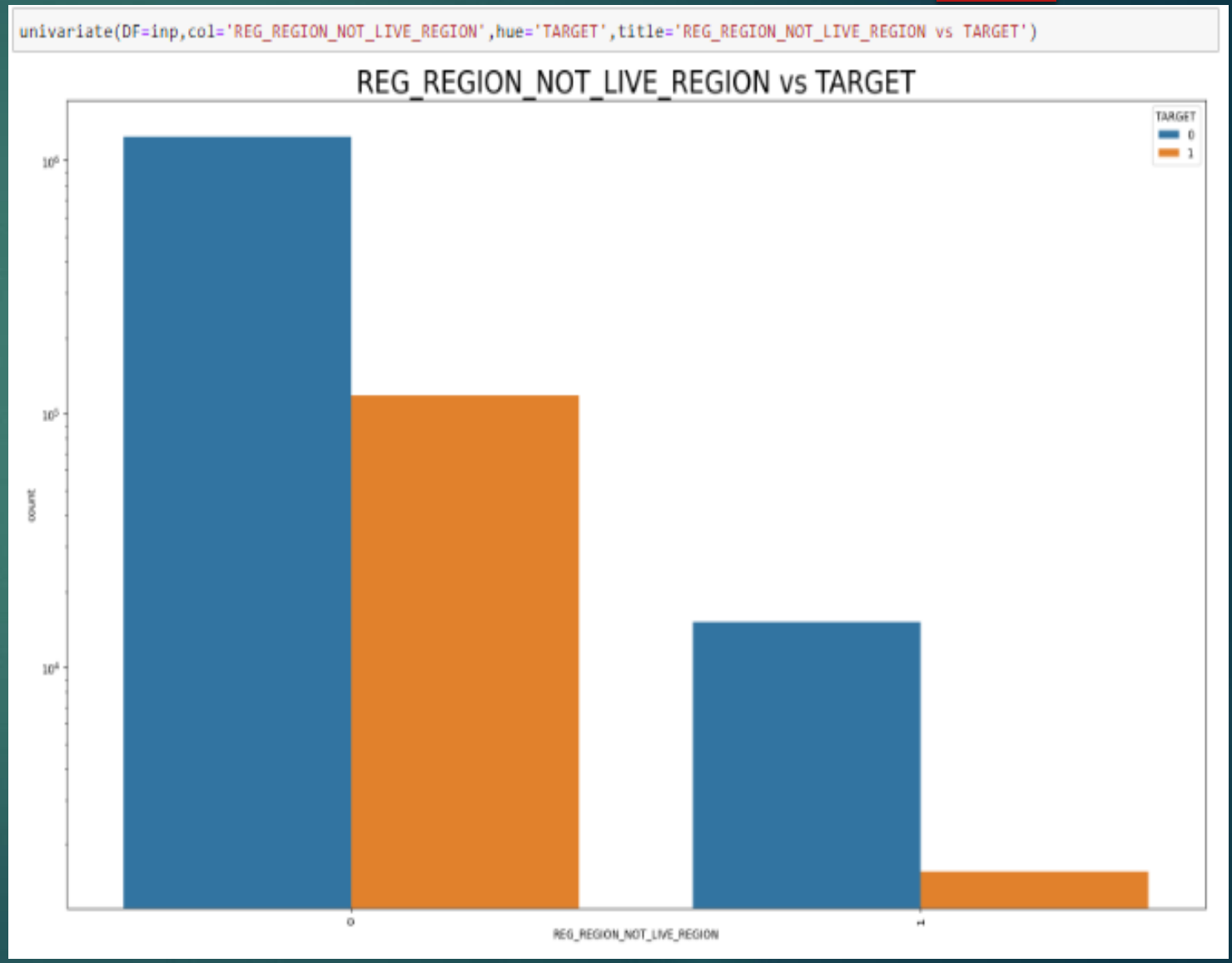


# Target VS Columns in Data

# Given Address not same as current address

**Evidence :**

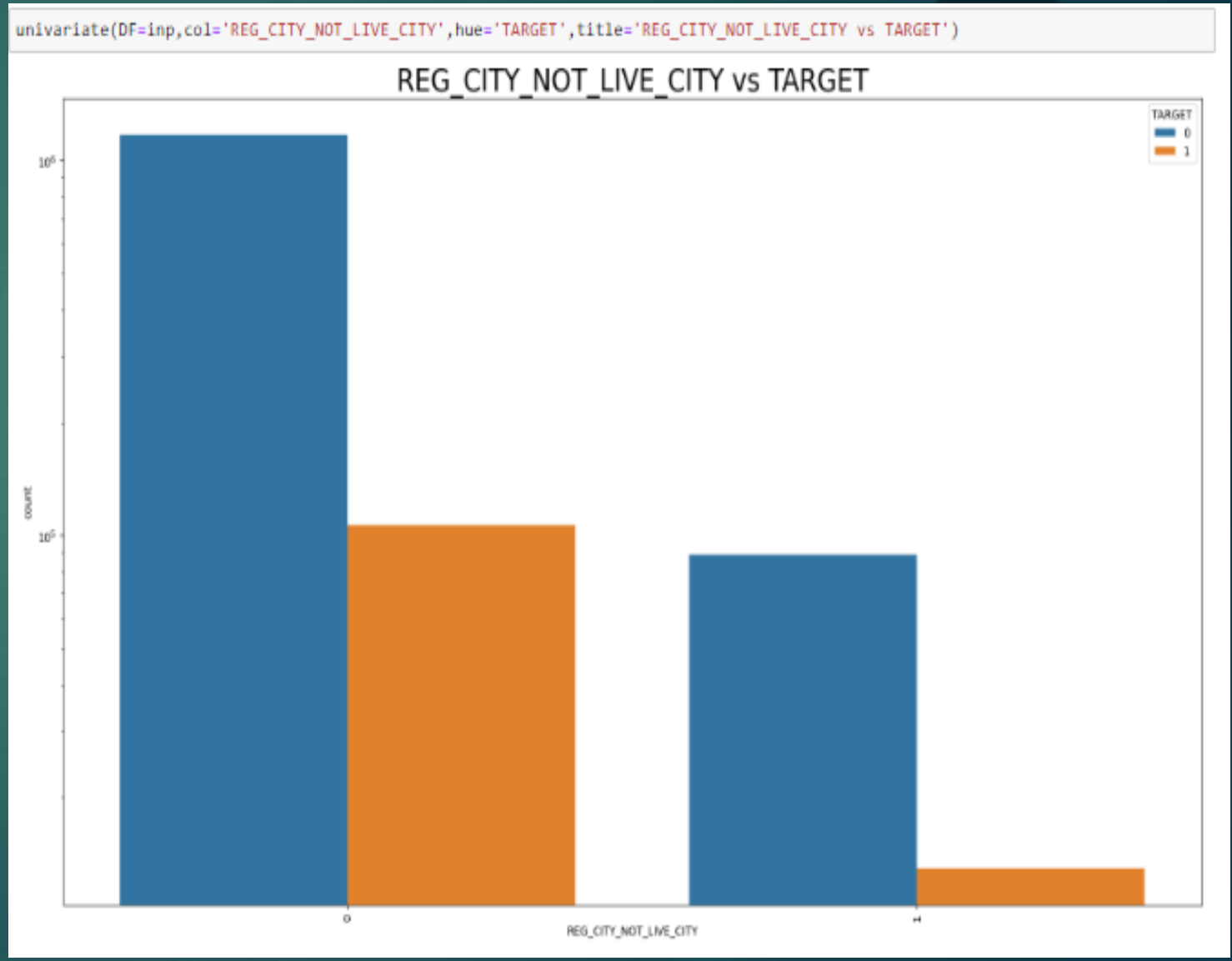
- This plot defines that 0’s is more than 1’s.
- That tells that current address is not same as living address



# Given City not same as current city

**Evidence :**

- Plot defines that there is high risk for loan repaying when Permanent address (Region/City) is not the contact address .



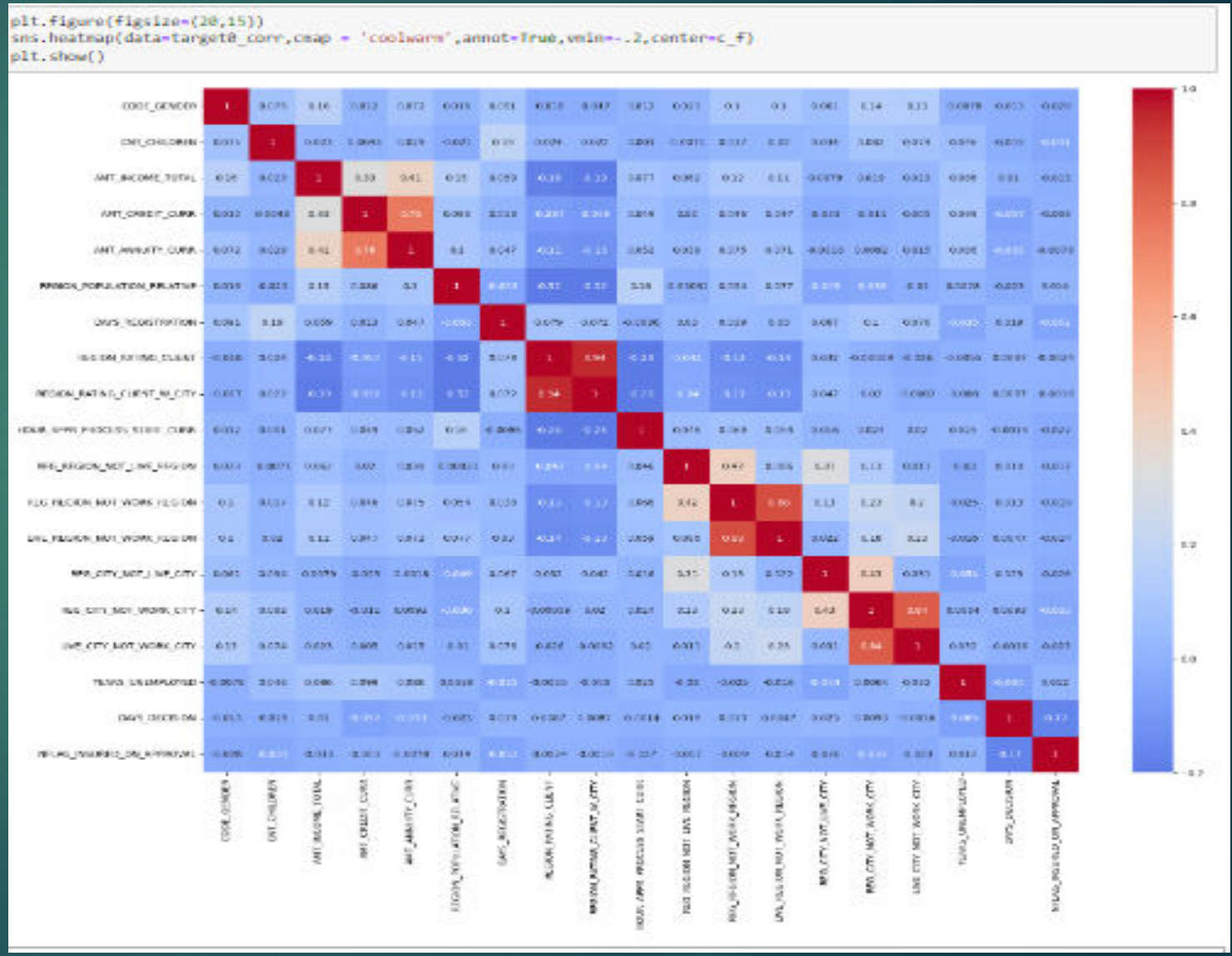
# Correlation of Target 0 Data

As there is Data imbalance in Target column, so the Total data is split into two columns by column target by 0's in one table and by 1's as another table

**Evidence:**

- Income amount, Credit amount and annuity current & previous which create majority of correlation with darker blue color
- The next high correlation is the region column.
- So the above are the factors contributing relationship with each other.
- Since , gender bias not play important role in this data

Note : The Value is on Gender



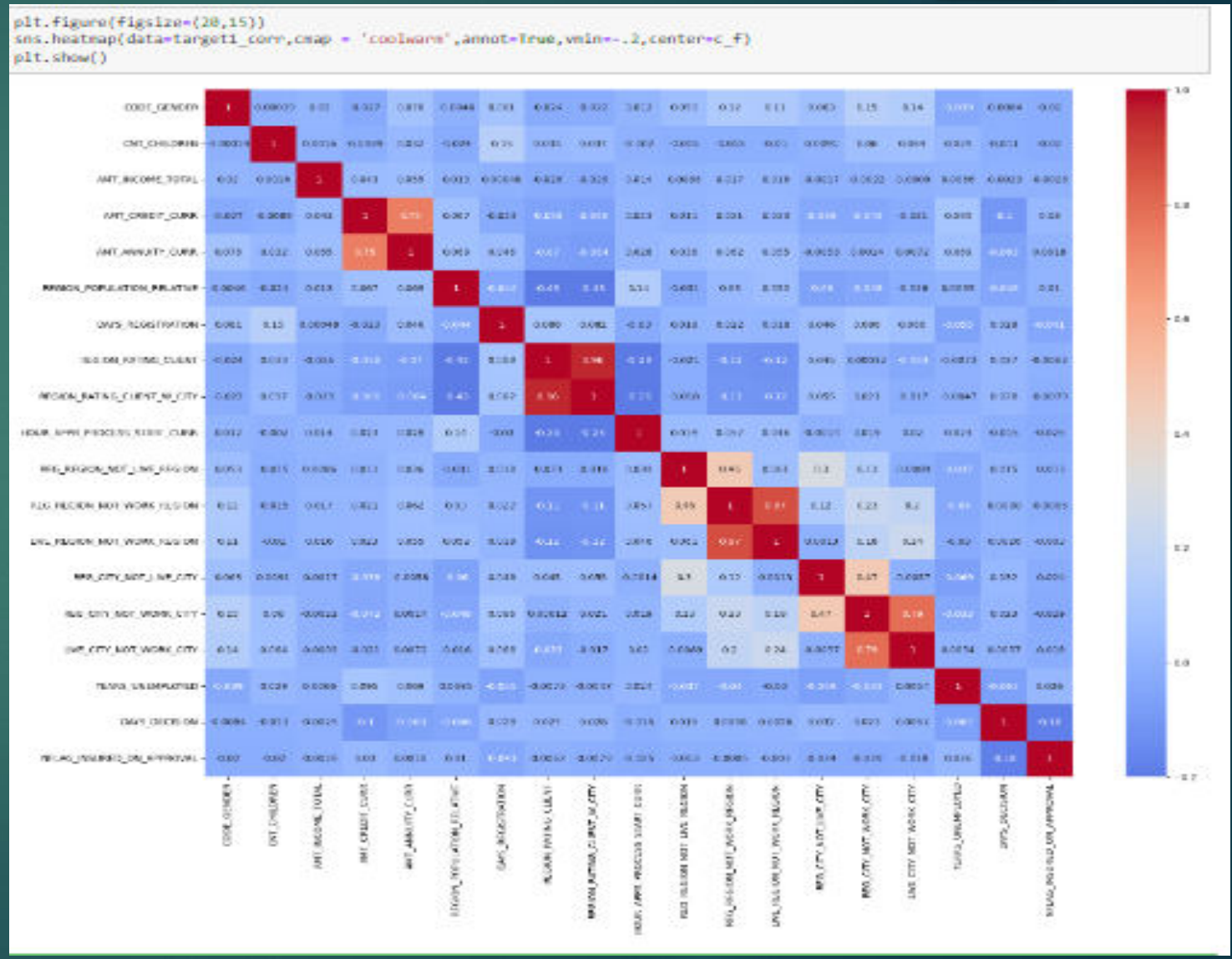
# Correlation of Target 1 Data

As there is Data imbalance in Target column, so the Total data is split into two columns by column target by 0's in one table and by 1's as another table

**Evidence:**

- In this Heat map, the majority of Correlation found with the Region data which is more mentioned in dark blue color
- The next high correlation found to be CREDIT Amount
- So the above are the factors contributing relationship with each other.
- Since , gender bias not play important role in this data

Note : The Value is on Gender



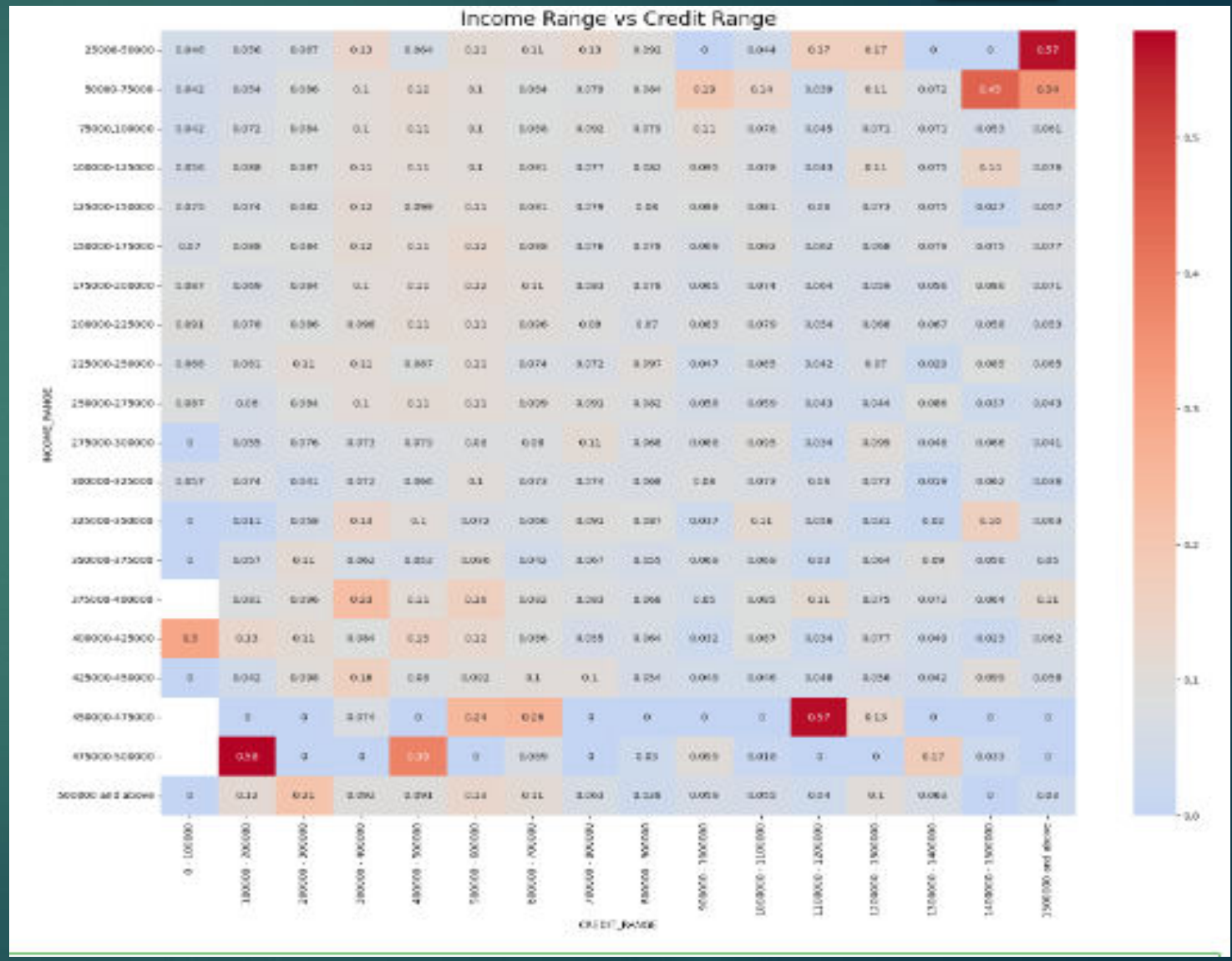
# Multivariate Analysis
# ( More than 2 Variables )

# Income Range VS Credit Range

**Evidence :**

- This Plot defines that Income Range is low with High Credit range shows with high chance that the target is 1.
- Also for the people with high income the Target 1 changes are high.

Notification : Pivot on Target

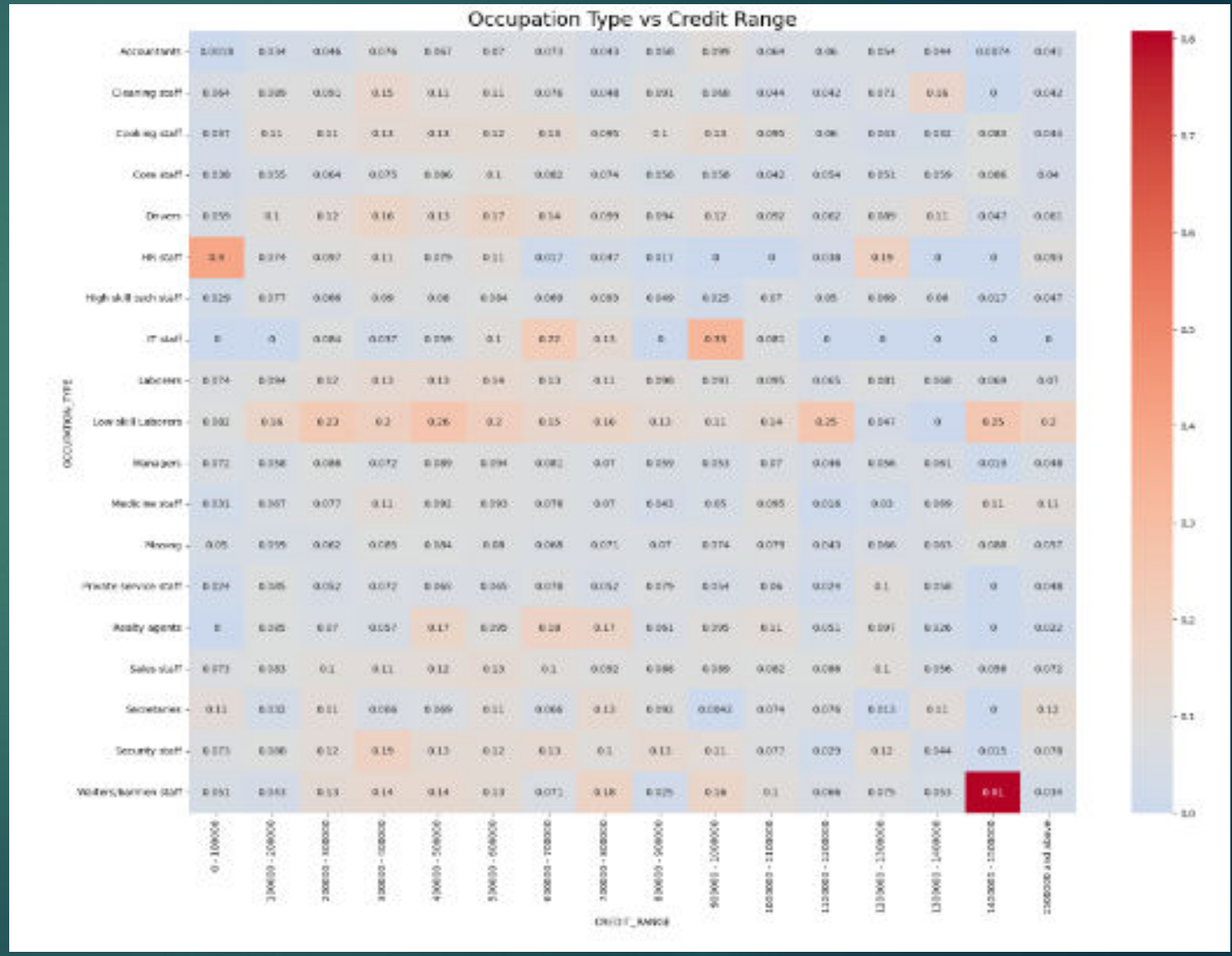


# Occupation Type VS Credit Range

**Evidence :**

- This plot defines that IT staff, and HR staff has little  higher probability that risk is  low.
- Also again the lower the credit rate lower the risk.

Notification : Pivot on Target

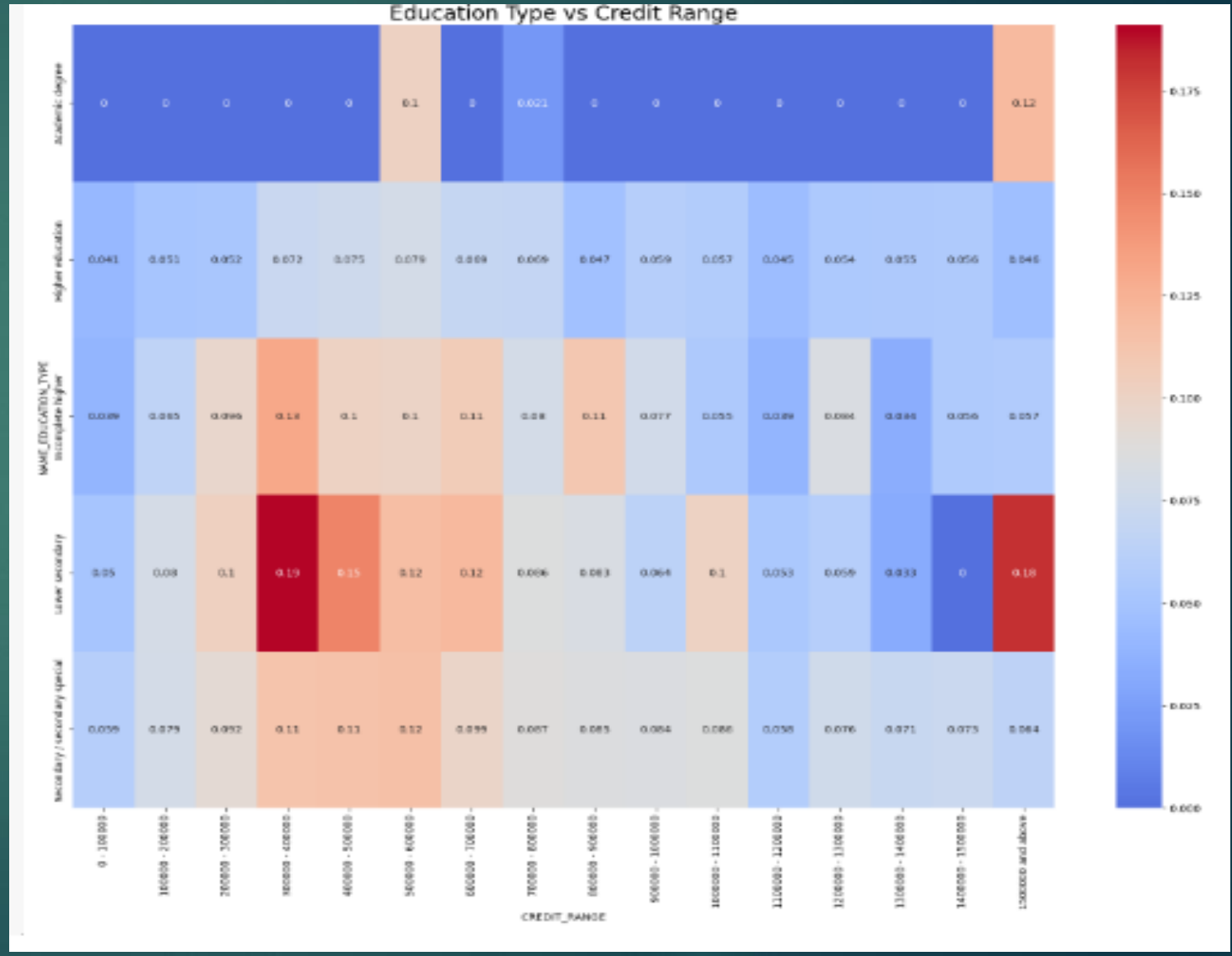


# Education Type VS Credit Range

**Evidence :**

- This plot defines that Lower secondary has high correlation with credit range
- Also the Academic education has low correlation with credit range

Notification : Pivot on Target

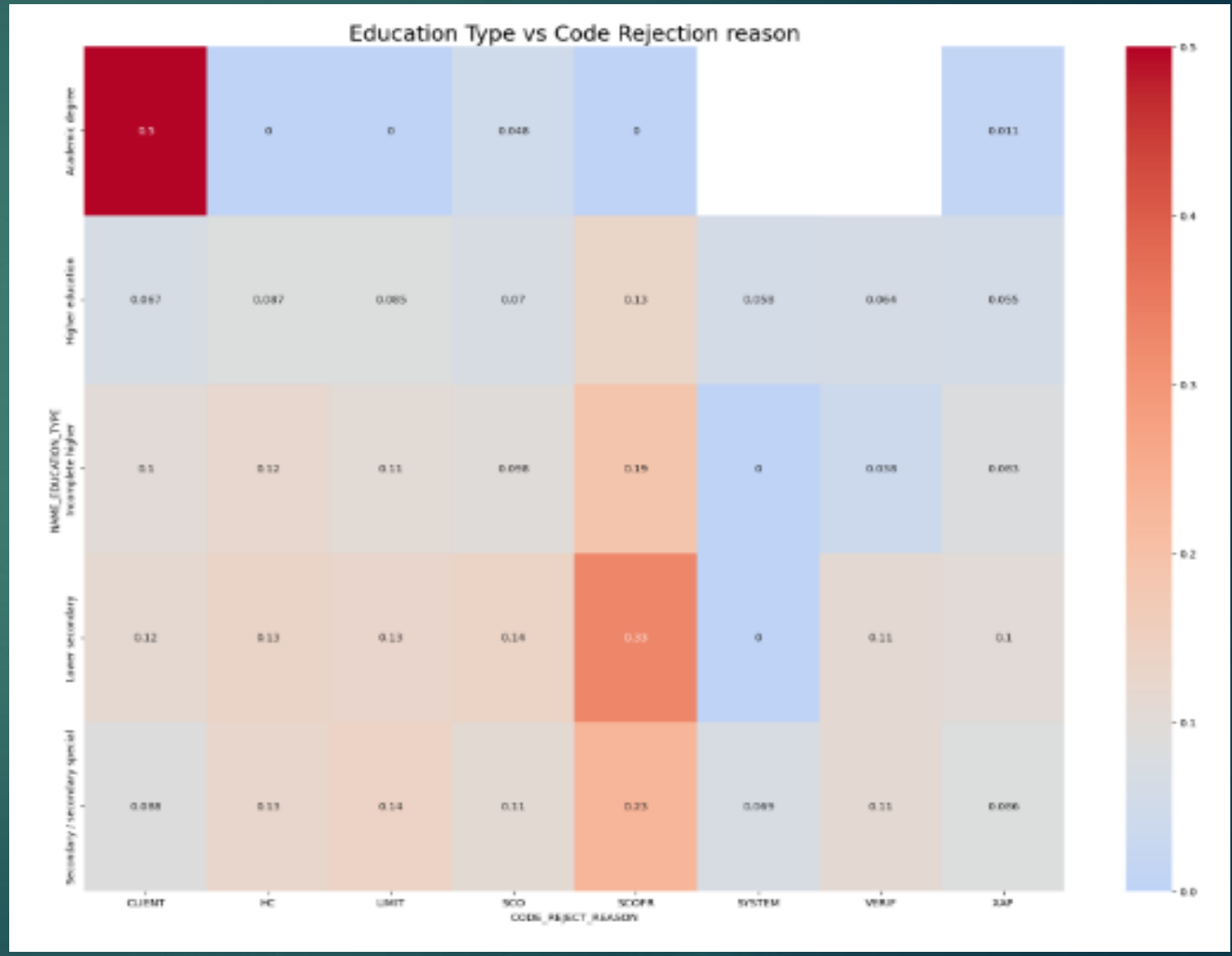


# Education Type VS Code Rejection Reason

**Evidence :**

- This plot defines that Academic education has highest correlation with SCOFR Code rejection Reason
- And after that Lower secondary has second highest correlation with Code Rejection Reason

Notification : Pivot on Target

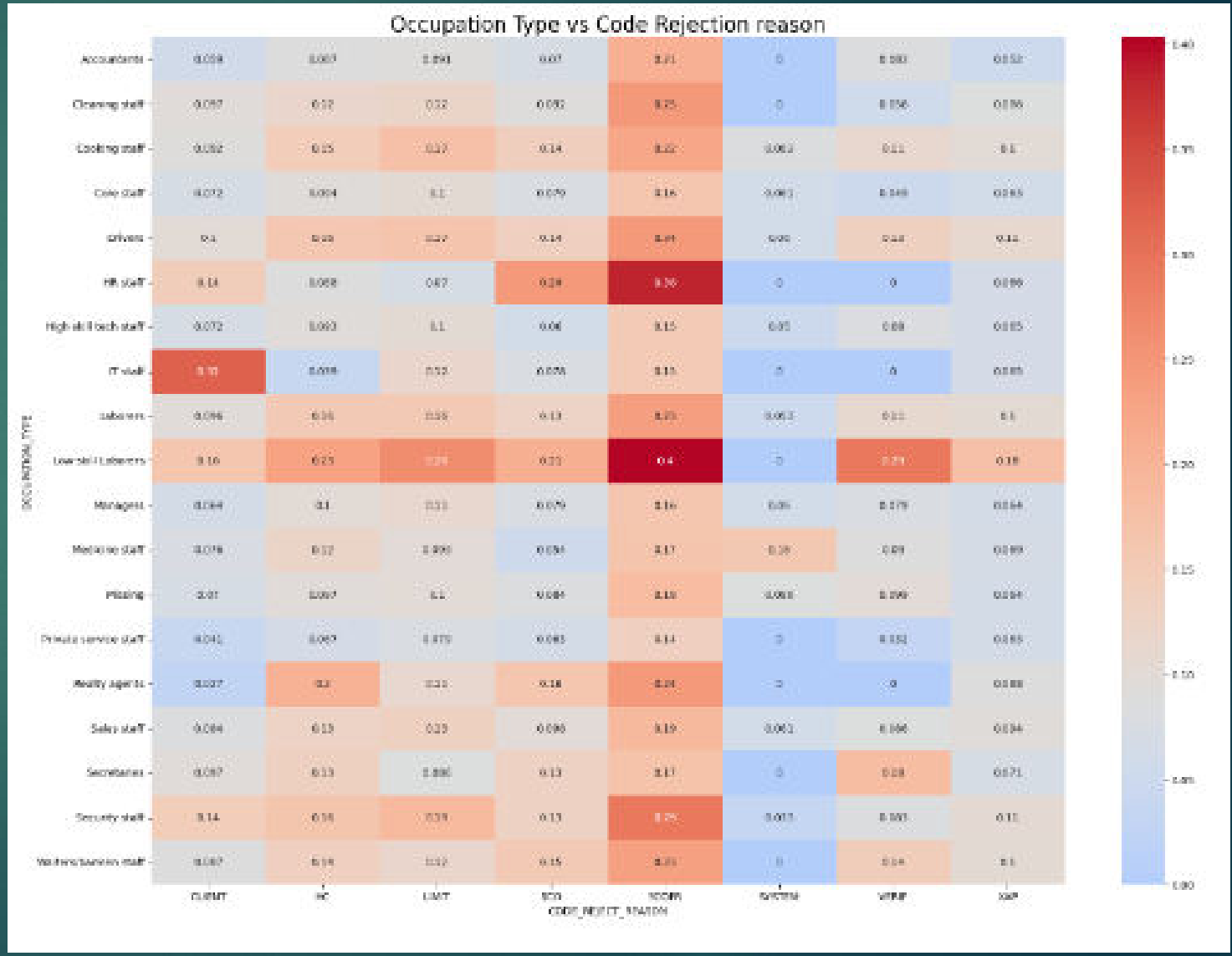


# Occupation Type VS Code Rejection Reason

**Evidence :**

- This plot defines that SCOFR Code rejection has high correlation with all occupation type

Notification : Pivot on Target

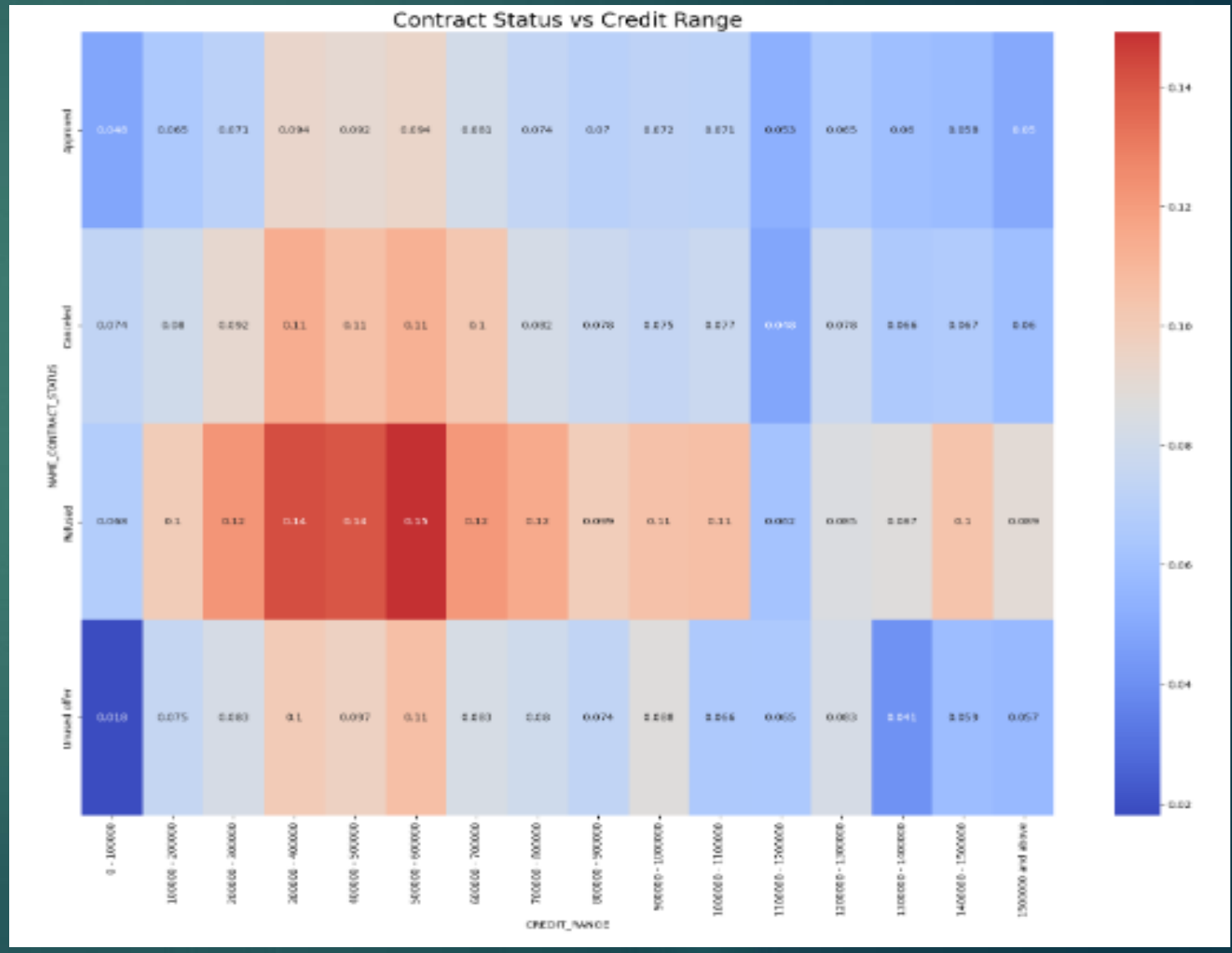


# Contract Status VS Credit Range

**Evidence :**

- This plot defines that Refused Candidate has high correlation with Credit Range
- Next , cancelled candidate has high correlation with Credit Range

Notification : Pivot on Target

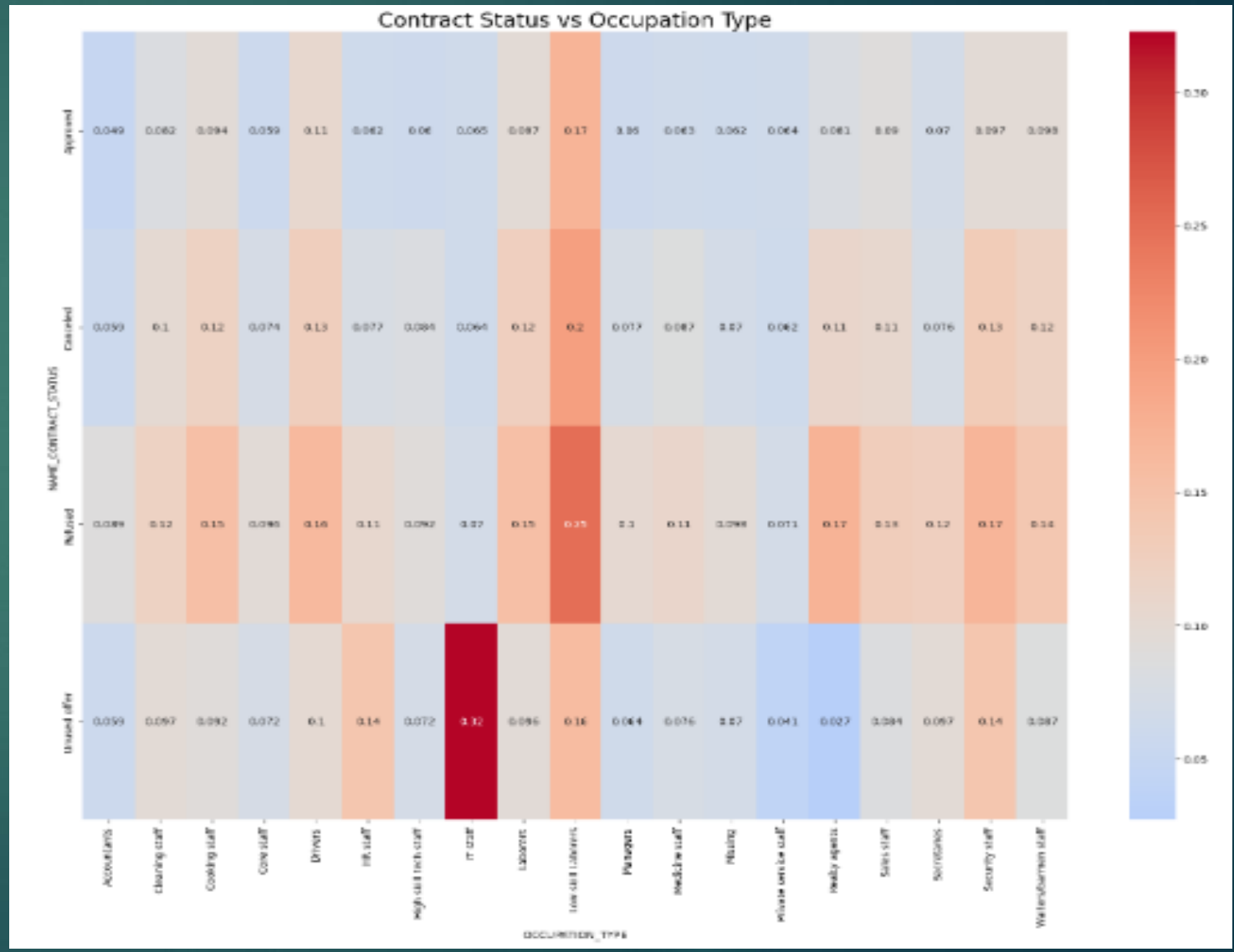


# Contract Status VS Occupation Type

**Evidence :**

- This plot defines that Low skill labor has high correlation with all Contract type
- In Contract type , Refused person having high correlation with all occupation type

Notification : Pivot on Target

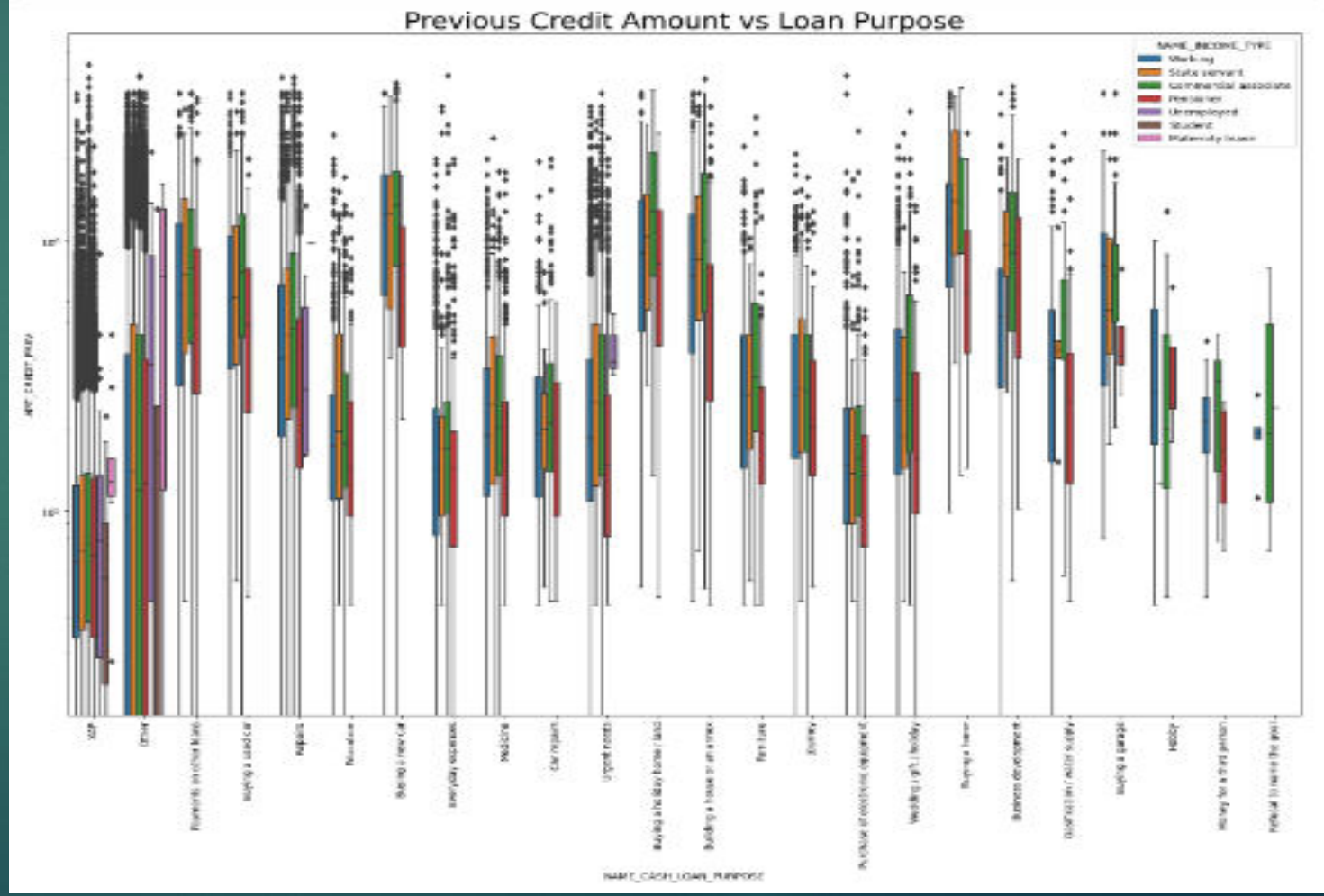


# Previous Credit amount VS Loan Purpose

**Evidence :**

- This plot defines that it has wide spread of amount Credit
- Amount credit range is high for all income type for other category

```
plt.figure(figsize=(20,14))
plt.xticks(rotation=90)
plt.yscale('log')
plt.title('Previous Credit Amount vs Loan Purpose', fontdict={'fontsize':25})
sns.boxplot(data =inpt, x='NAME_CASH_LOAN_PURPOSE',hue='NAME_INCOME_TYPE',y='AMT_CREDIT_PREV')
plt.show()
```



# Final Result : Banker seeking for Loan Approval

- ‘Businessman’, ‘Pensioner’ and ‘Students’ column are more successful payment of Loan.

- In Education column – Academic range education seems to be less loan approval rate . Bank can also consider for successful loan rate incomes.

- Bank should focus less on ‘Income type working’ as they are difficulty in paying the loan.

# Thankyou