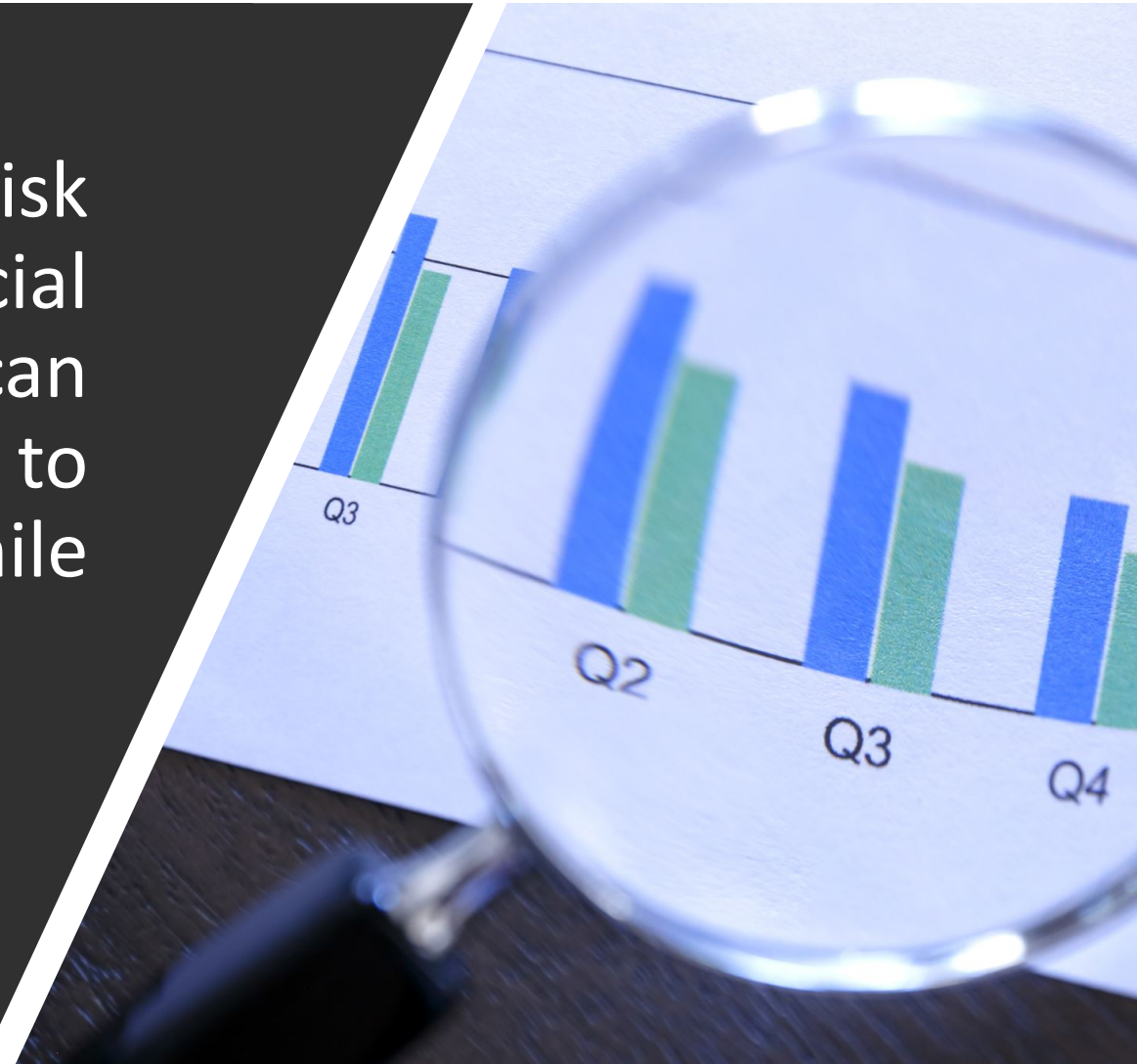
Case Study on Credit EDA

Presenter: - Anhad

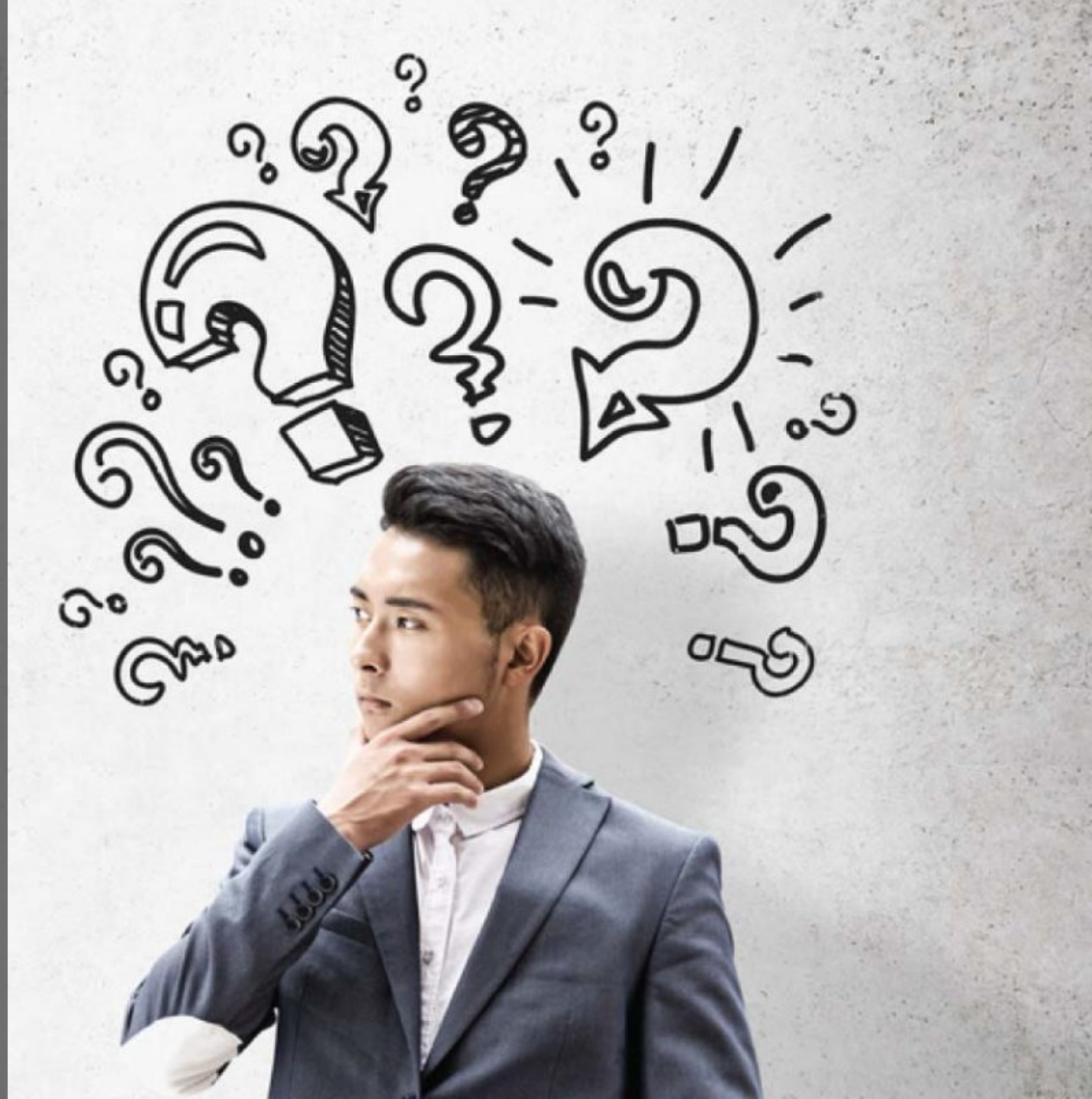
Introduction

This case study aims towards risk analytics in banking and financial services and so that one can understand how data is used to minimise the risk of losing money while lending to customers.



Problem Statement

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter.



Data Structure

The data we have contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
- **All other cases:** All other cases when the payment is paid on time.

Decision Scenarios: -

There are four types of decisions that could be taken by the client/company):

Approved: The Company has approved loan Application

Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.

Refused: The company had rejected the loan (because the client does not meet their requirements etc.).

Unused offer: Loan has been cancelled by the client but on different stages of the process.

Lets Deep dive and fetch solutions and inferences

Research Methodology

- We have taken the secondary bank loan data from UpGrad Portal.
- Initially we'll clean the data and remove the null values as well as convert values according to our research so that we can fetch inference from it.

Research Methodology

- Then we'll drop those particulars which have more than certain percentage of null values.
- Categories the data according to the gender type to get the behavioral inferences from it.
- Change the data type (Qualitative and quantitative) according to the need.

Language and Libraries used

Language — Python

Libraries — Numpy, Pandas

Data visualization tools — Matplotlib, Seaborn

Research

- After importing all the libraries we have called the data in our EDI with the help of python computer language.

- We have tested few basic operations like calling the headers, checking shape of the data, so that we can make sure that the database is working properly.

Null Values Operation

In the null value operation following tasks has been performed

- Cleaning the missing data
- Calculating the total null values in the dataset
- Determining the columns which has more than 30% of the null values

```
In [51]: # Cleaning the missing data
         # Calculating the total null values in the dataset
         # Determining the columns which has more than 30% of the null values

         nullcols=df.isnull().sum()
         nullcols=nullcols[nullcols.values>(0.3*len(nullcols))]
         len(nullcols)

Out[51]: 64
```

Null value operation result

There are 64 columns having null values greater than 30% in the dataset which can be removed to our analysis

We have dropped those 64 columns which have null values percentage more than 30%

In [52]:
```
# Dropping the 64 columns that have more than 30% null values
nullcols = list(nullcols[nullcols.values>=0.3].index)
df.drop(labels=nullcols,axis=1,inplace=True)
print(len(nullcols))
```

```
64
```

- The column 'AMT_ANNUITY' has very few null values. We can try and impute the missing values.

- There is a possibilities to have outliers. Thus it will be inappropriate to fill the missing values with the mean value. Here we have used the Median value for imputation and fill those missing records.

```
# Filling the missing values with median value

values=df['AMT_ANNUITY'].median()

df.loc[df['AMT_ANNUITY'].isnull(),'AMT_ANNUITY']=values
```

- Similarly we have removed rows having null values greater than or equal to 30% if any

```
# Now we remove rows having null values greater than or equal to 30% if any

nullrows=df.isnull().sum(axis=1)
nullrows=list(nullrows[nullrows.values>=0.3*len(df)].index)
df.drop(labels=nullrows,axis=1,inplace=True)
print(len(nullrows))
```

```
0
```

Data Cleaning

We have removed the irrelevant columns from our dataset which will not play a role in the analysis

```
# We will remove the irrelevant columns from our dataset which will not play a role in the analysis.

unwanted=['FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE',
       'FLAG_PHONE', 'FLAG_EMAIL','REGION_RATING_CLIENT','REGION_RATING_CLIENT_W_CITY','FLAG_EMAIL','CNT_FAM_MEMBERS',
       'REGION_RATING_CLIENT_W_CITY','DAYS_LAST_PHONE_CHANGE', 'FLAG_DOCUMENT_2', 'FLAG_DOCUMENT_3','FLAG_DOCUMENT_4',
       'FLAG_DOCUMENT_7', 'FLAG_DOCUMENT_8', 'FLAG_DOCUMENT_9','FLAG_DOCUMENT_10', 'FLAG_DOCUMENT_11', 'FLAG_DOCUMENT_1
       'FLAG_DOCUMENT_13', 'FLAG_DOCUMENT_14', 'FLAG_DOCUMENT_15','FLAG_DOCUMENT_16', 'FLAG_DOCUMENT_17', 'FLAG_DOCUMEN
       'FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_20', 'FLAG_DOCUMENT_21']

df.drop(labels=unwanted,axis=1,inplace=True)
```

There were some columns where the value is mentioned as 'XNA' which means 'Not Available'. So we must find the number of rows and columns and implement suitable techniques on them to fill those missing values or to delete them if necessary.

```
# let's find these categorical columns having these 'XNA' values

# For Gender column`

df[df['CODE_GENDER']=='XNA'].shape
```

```
(4, 28)
```

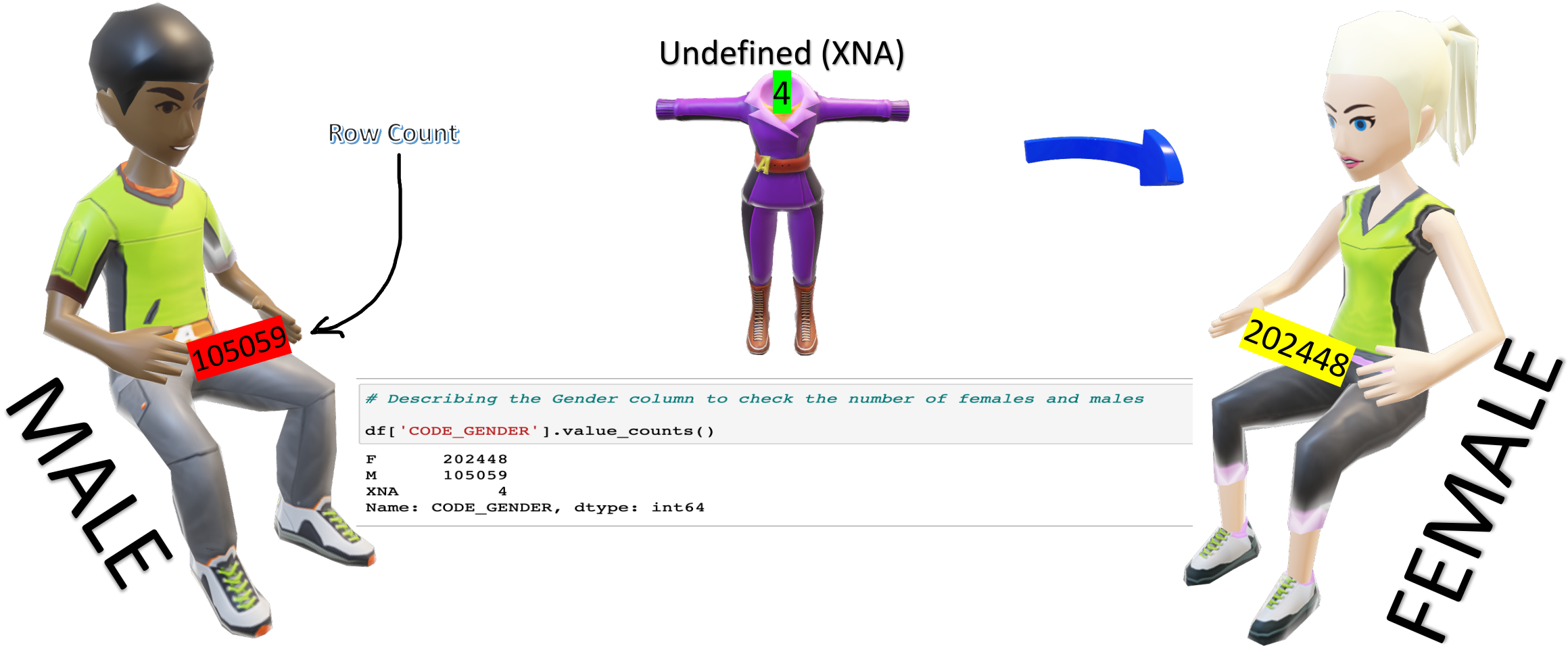
```
# For Organization column

df[df['ORGANIZATION_TYPE']=='XNA'].shape
```

```
(55374, 28)
```

So, there are 4 rows from Gender column and 55374 rows from Organization type column that contains the 'XNA'

Classification of data according to the gender of target customers




```
# Describing the Gender column to check the number of females and males

df['CODE_GENDER'].value_counts()
```

```
F      202448
M      105059
XNA         4
Name: CODE_GENDER, dtype: int64
```

Since, Female gender is having the majority and only 4 rows are having 'XNA' values, we can update those columns with Gender 'F'.

Metrics derived from the Cleaned Dataset

Created bins for continuous variable categories i.e. 'AMT_INCOME_TOTAL' and 'AMT_CREDIT' in which we have subcategories data into specific slots: -

['0-25000', '25000-50000', '50000-75000', '75000,100000', '100000-125000', '125000-150000', '150000-175000', '175000-200000', '200000-225000', '225000-250000', '250000-275000', '275000-300000', '300000-325000', '325000-350000', '350000-375000', '375000-400000', '400000-425000', '425000-450000', '450000-475000', '475000-500000', '500000 and above']

In [71]:
```
# Creating bins for Income Amount

bins = [0,25000,50000,75000,100000,125000,150000,175000,200000,225000,250000,275000,300000,325000,350000,375000,400000,
slot = ['0-25000', '25000-50000','50000-75000','75000,100000','100000-125000', '125000-150000', '150000-175000','175000
       '200000-225000','225000-250000','250000-275000','275000-300000','300000-325000','325000-350000','350000-375000',
       '375000-400000','400000-425000','425000-450000','450000-475000','475000-500000','500000 and above']

df['AMT_INCOME_RANGE']=pd.cut(df['AMT_INCOME_TOTAL'],bins,labels=slot)
```

```
# Creating bins for Credit Amount

bins = [0,150000,200000,250000,300000,350000,400000,450000,500000,550000,600000,650000,700000,750000,800000,850000,900
slots = ['0-150000', '150000-200000','200000-250000', '250000-300000', '300000-350000', '350000-400000','400000-450000'
       '450000-500000','500000-550000','550000-600000','600000-650000','650000-700000','700000-750000','750000-800000'
       '800000-850000','850000-900000','900000 and above']

df['AMT_CREDIT_RANGE']=pd.cut(df['AMT_CREDIT'],bins=bins,labels=slots)
```

Dividing the dataset into two categories and storing them in different datasets

Target=1 (client with payment difficulties)
Target=0 (rest of the data)

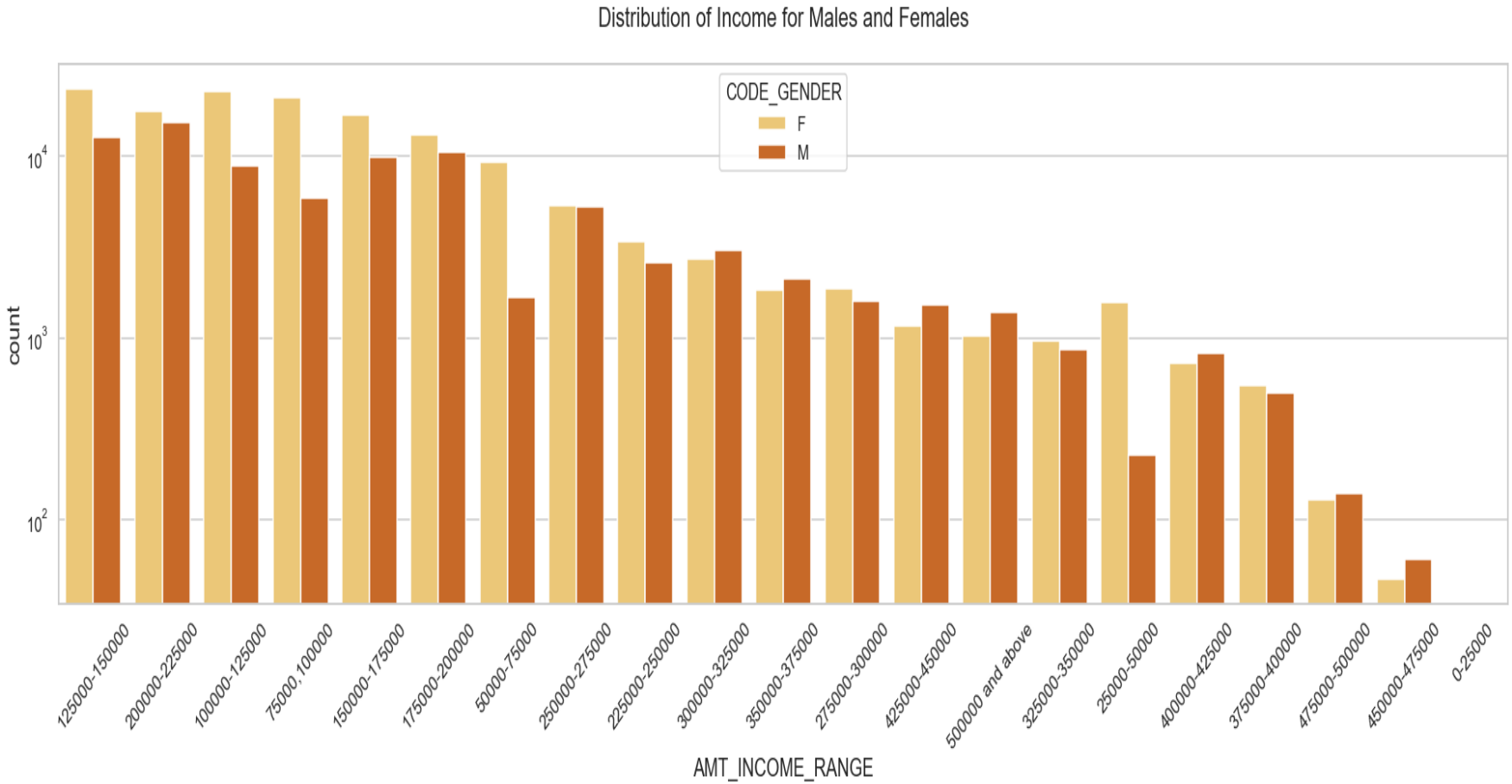
Calculating Imbalance percentage between target0 and target1. target0 has higher number of records

round(len(target0_df)/len(target1_df),2)

Ratio : - 10.55

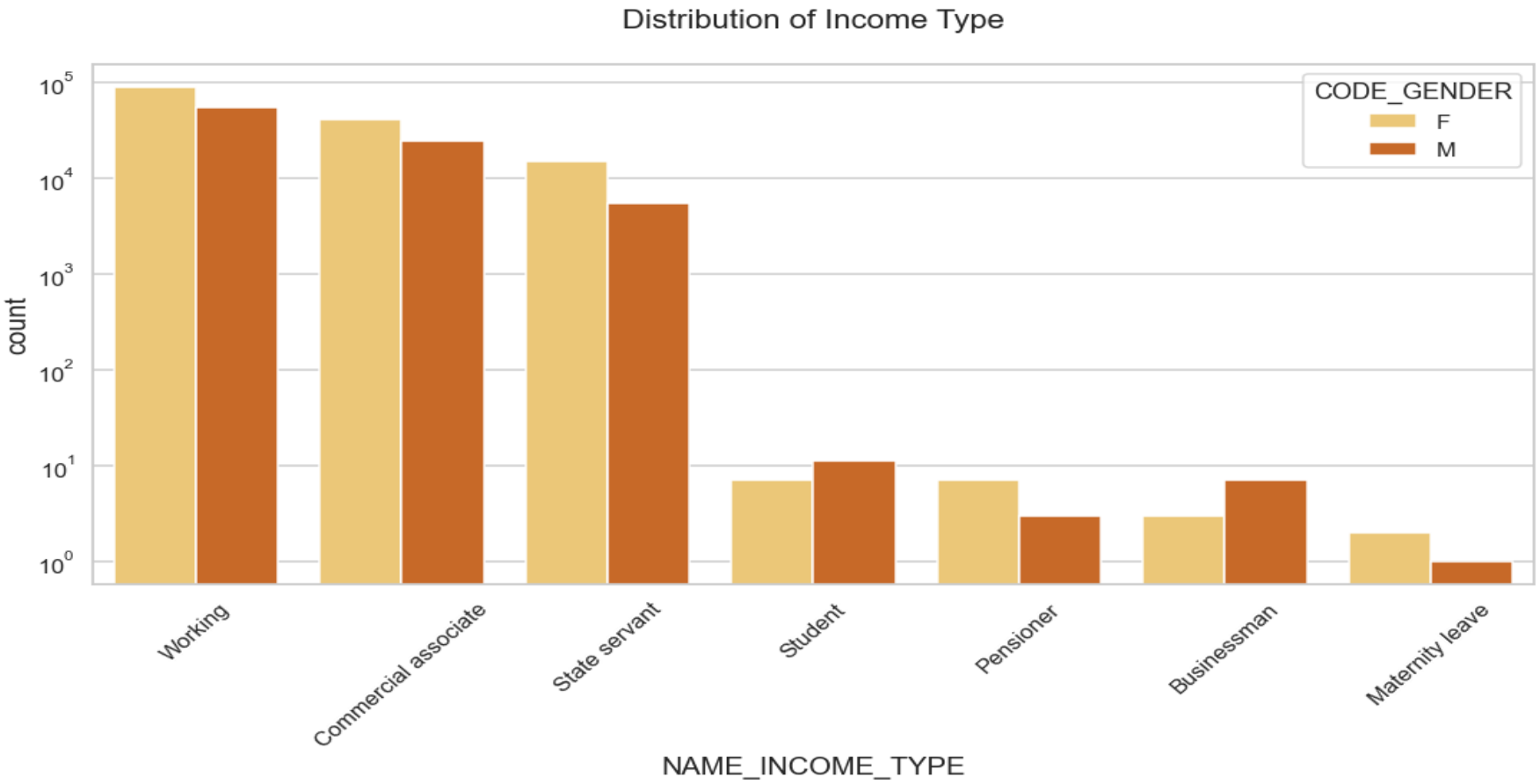
Output: - The Imbalance ratio between target0 and target1 is 10.55. Target 0 customers are 10.55 times more than Target 1 customer

FINDINGS



Points that can be inferred from the above graph.

1. The count of Females are greater than that of Males.
2. Income range from 125000 to 200000 is having significantly more number of credits than others.
3. Number of people with income of 400000 and above are very less.

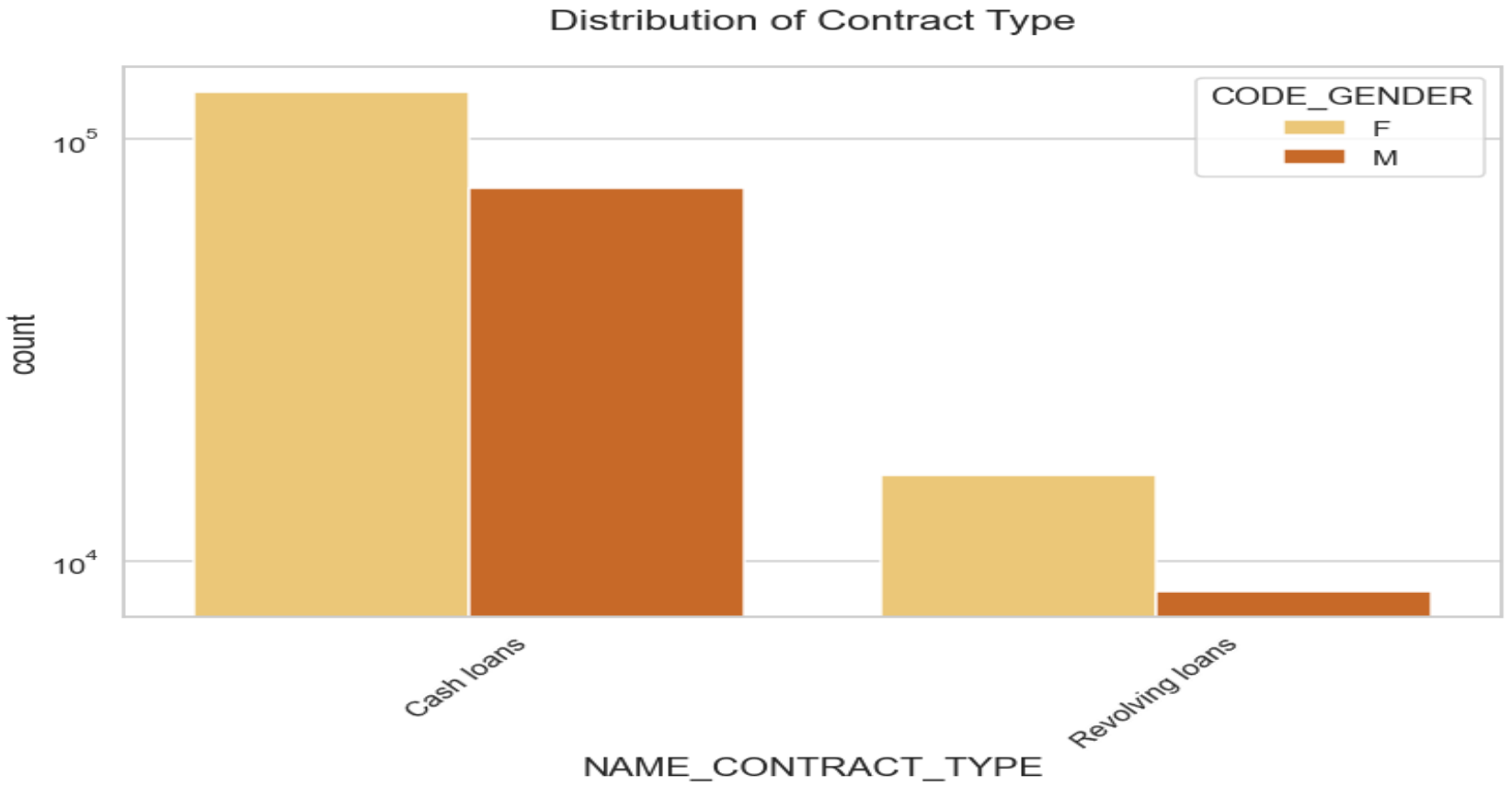


Points inferred from the above graph.

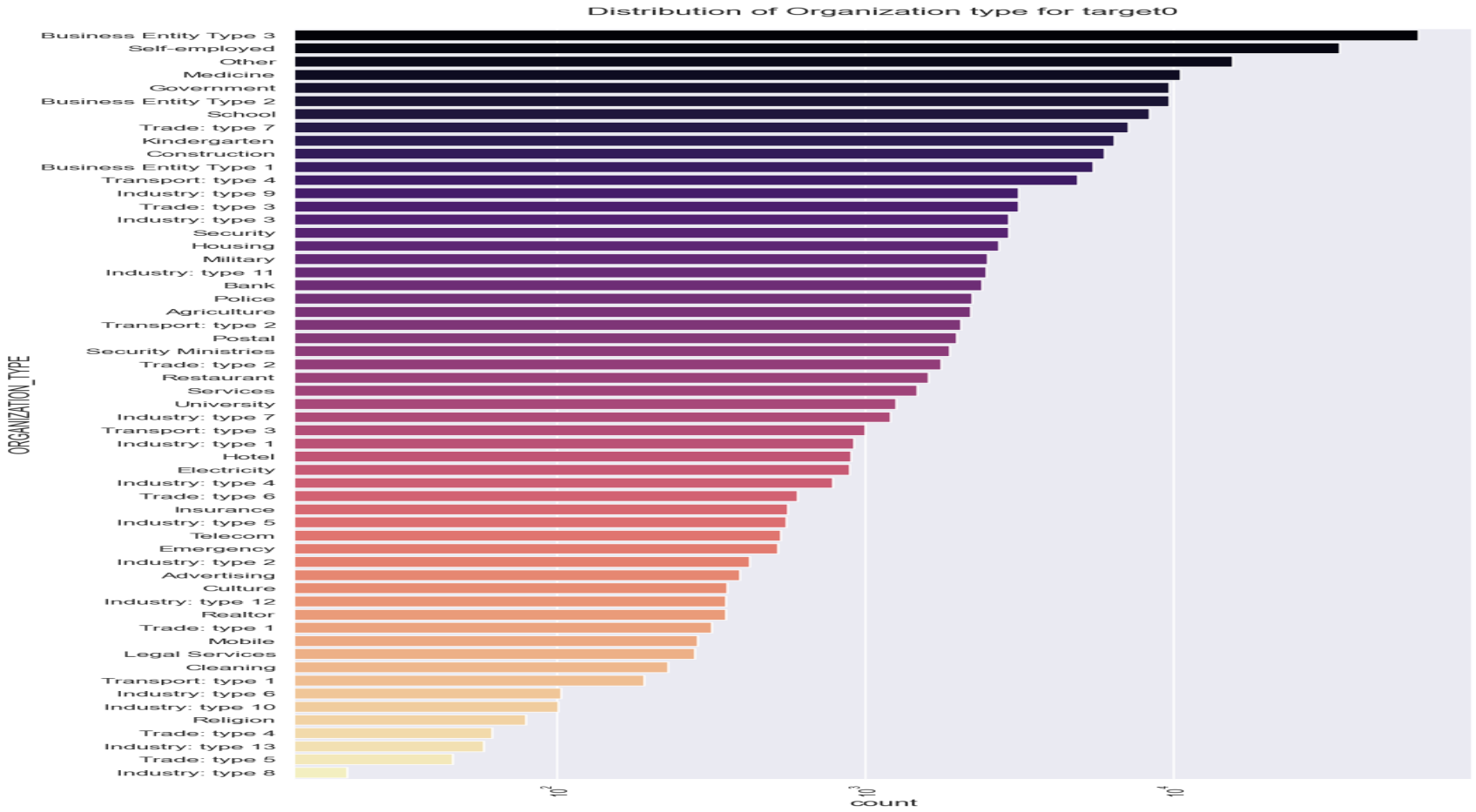
1.For income type ‘Working’, 'Commercial associate', and ‘State Servant’ the number of credits are higher than others.

2.Again Females are having more number of credits than males.

3.People with 'Maternity Leave' have the least credits

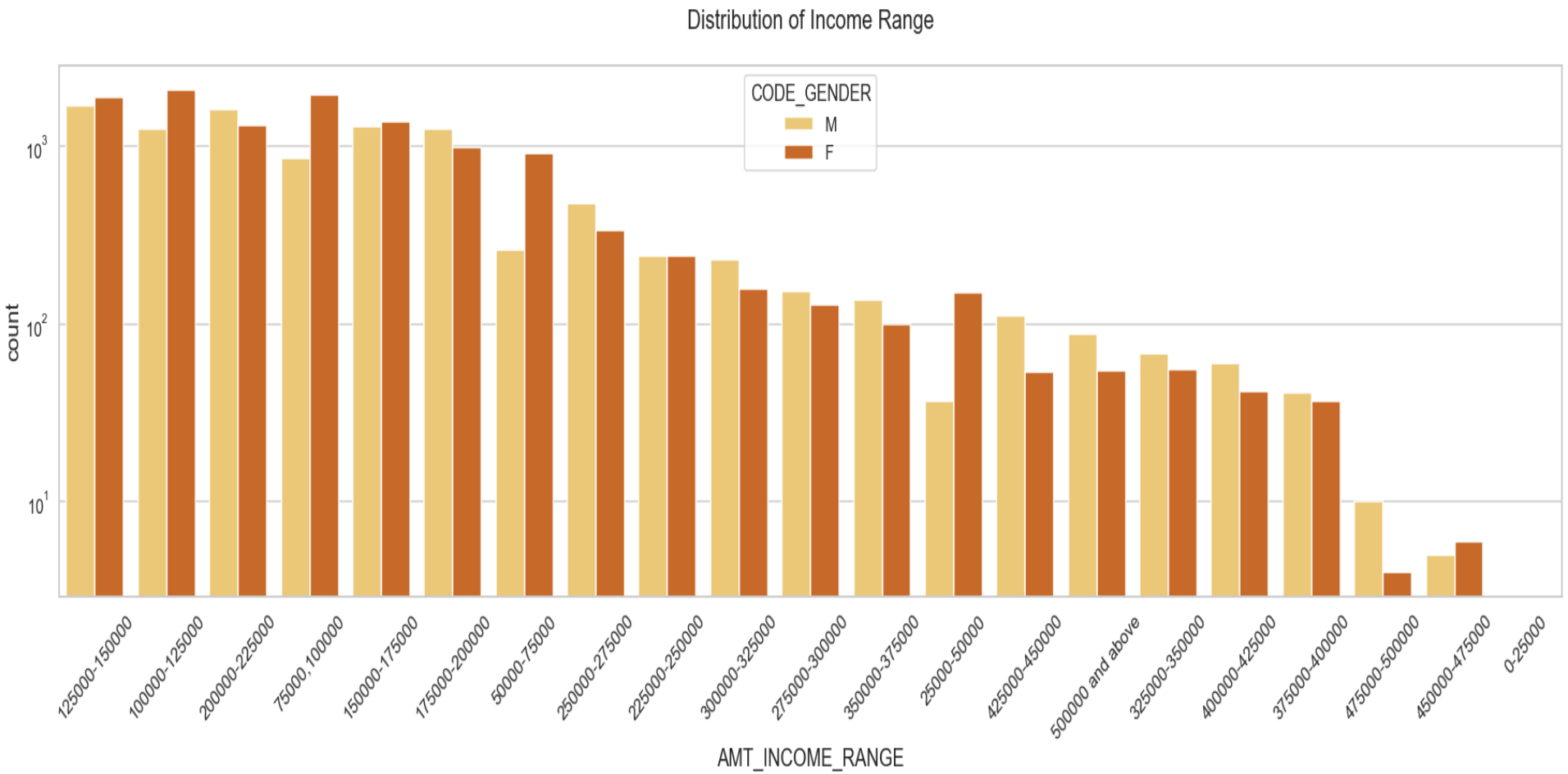


Points inferred from the above graph.
1.‘Cash Loans’ is having higher number of credits than ‘Revolving loans’.
2.Once again, Female is leading for applying credits.



Points inferred from the above graph.

1. People doing Business(entity Type 3) , Self employed, from Medicine and Government sectors have the highest count in applying for credits.
2. People from background as Industry, Trade, Cleaning, Legal Services etc. are very few in counts compared to others

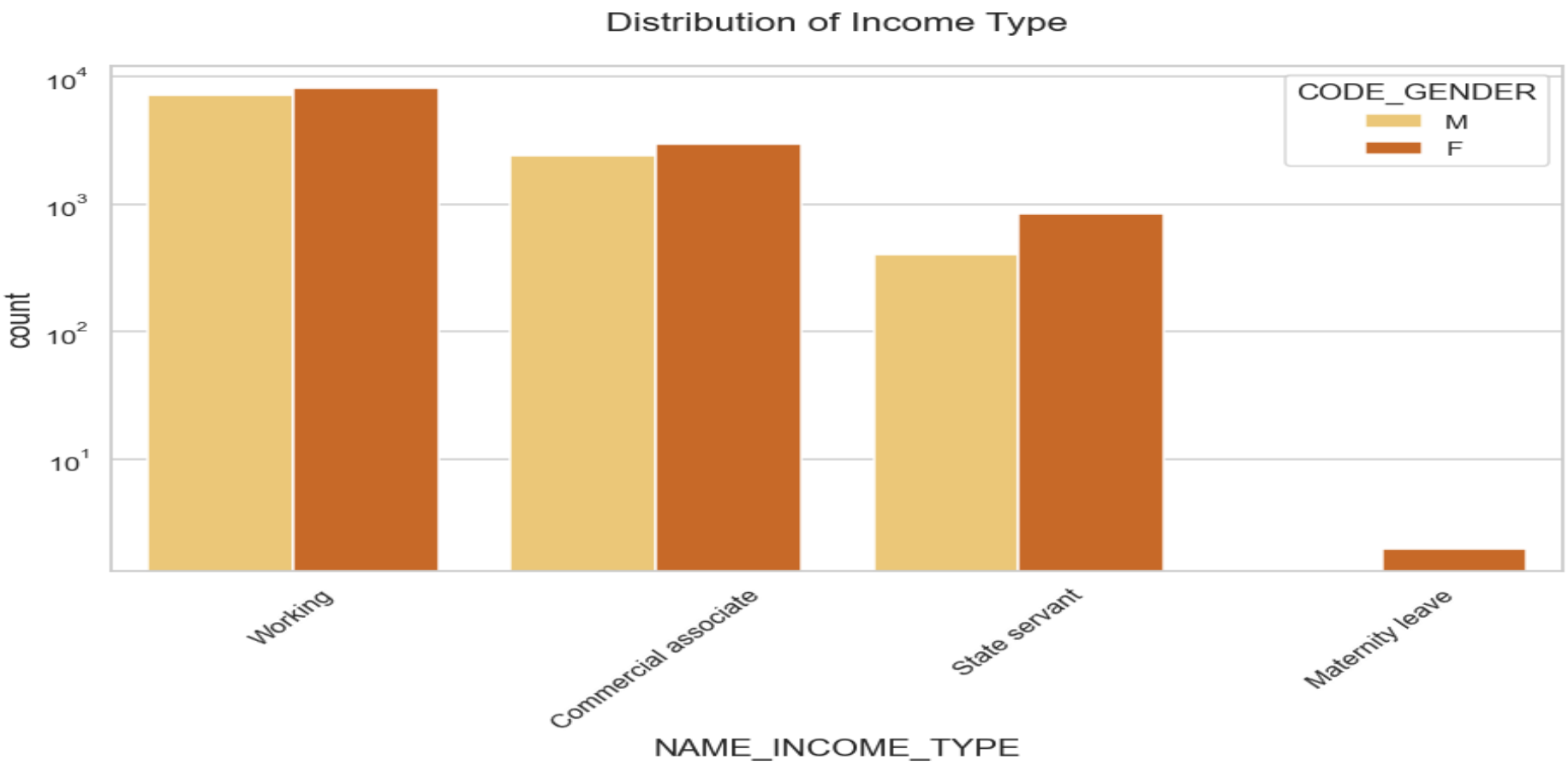


Points inferred from the above graph.

1.Here the number of Males are more than number of females.

2.Income range from 100000 to 200000 is having more number of credits.

3.Once again, very less count for income of 400000 and above.



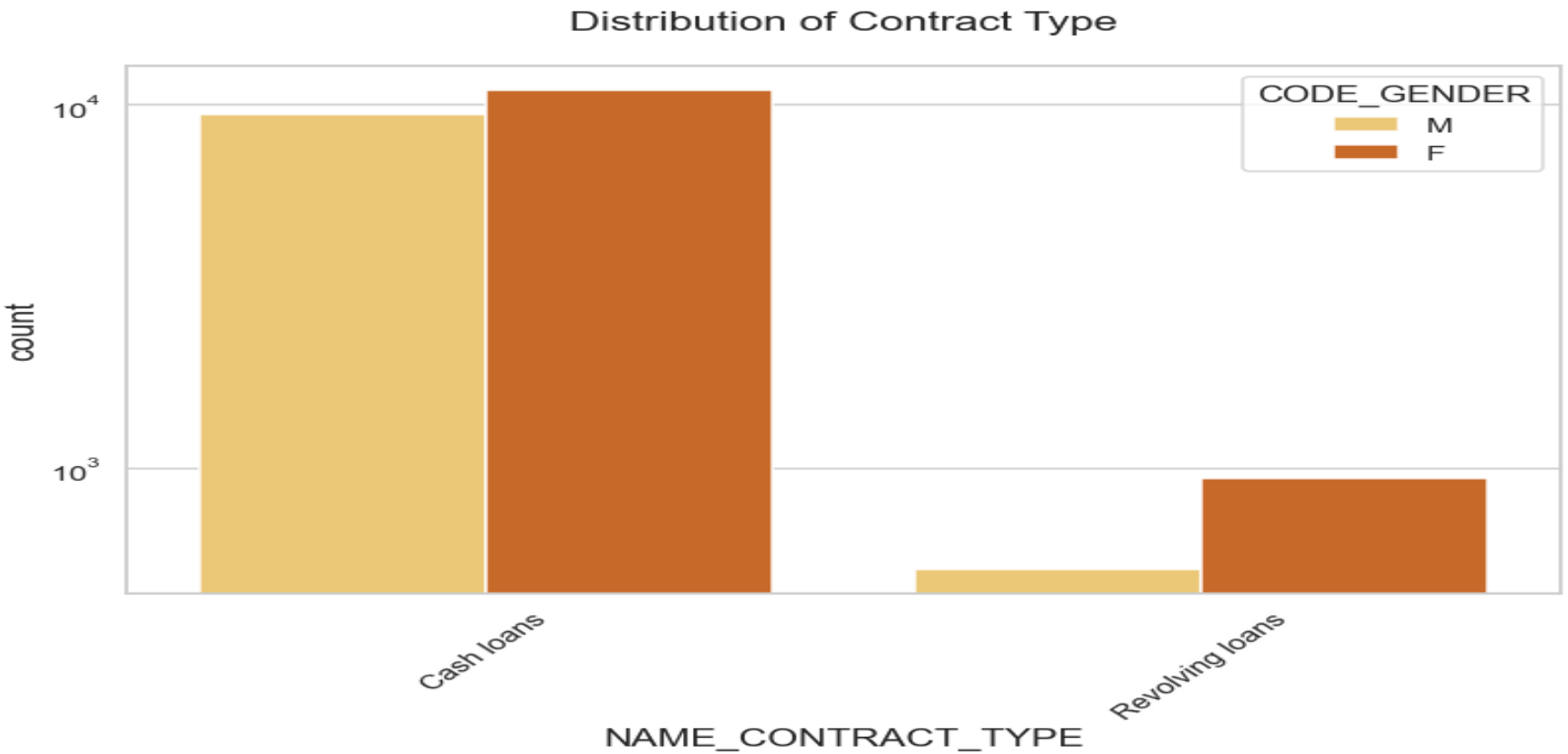
Points inferred from the above graph.

1.For income type ‘working’, ’commercial associate’, and ‘State Servant’ the number of credits are higher than other i.e. ‘Maternity leave.

2.For this Females are having more number of credits than male.

3.Less number of credits for income type ‘Maternity leave’.

4.For type 1: There is no income type for ‘student’ , ’pensioner’ and ‘Businessman’ which means they don’t do any late payments.



Points to be concluded from the above graph.
1.For contract type ‘cash loans’ is having higher number of credits than ‘Revolving loans’ contract type.
2.For this also Female is leading for applying credits.
3.For type 1 : Female Revolving loans is significantly higher than that of Males.