Credit EDA Assignment

By:- Prateek Raj Srivastava

Data Cleaning

First we have to find the null values and then we have to fill them or drop them according to their impact on our analysis.

```
null_count=ass.isnull().sum()
null_count
```

```
SK_ID_CURR                      0
TARGET                          0
NAME_CONTRACT_TYPE              0
CODE_GENDER                     0
FLAG_OWN_CAR                    0
                               ...
AMT_REQ_CREDIT_BUREAU_DAY       41519
AMT_REQ_CREDIT_BUREAU_WEEK      41519
AMT_REQ_CREDIT_BUREAU_MON       41519
AMT_REQ_CREDIT_BUREAU_QRT       41519
AMT_REQ_CREDIT_BUREAU_YEAR      41519
Length: 122, dtype: int64
```

The data which more than 32% of data is missing are useless show we will drop those data.

```
drop_col=ass.isnull().sum()/len(ass)
drop_col=list(drop_col[drop_col.values>=0.32].index)
ass.drop(labels=drop_col,axis=1,inplace=True)
ass.shape
```

```
(307511, 73)
```

Now, the columns like OCCUPATION_TYPE we can fill these missing values with zero.

```
ass["OCCUPATION_TYPE"].fillna(0,inplace=True)
```

```
ass["AMT_REQ_CREDIT_BUREAU_YEAR"].fillna(0,inplace=True)
```

```
ass["AMT_REQ_CREDIT_BUREAU_MON"].fillna(0,inplace=True)
ass["AMT_REQ_CREDIT_BUREAU_WEEK"].fillna(0,inplace=True)
ass["AMT_REQ_CREDIT_BUREAU_DAY"].fillna(0,inplace=True)
ass["AMT_REQ_CREDIT_BUREAU_HOUR"].fillna(0,inplace=True)
ass["AMT_REQ_CREDIT_BUREAU_QRT"].fillna(0,inplace=True)
```

The columns like AMT_GOODS_PRICE we can its missing value through the median of the column. The column CODE Gender we are going to fill the missing values with the values which has occurred most.

```
ass["AMT_GOODS_PRICE"].fillna(ass.AMT_GOODS_PRICE.median(),inplace=True)
ass["AMT_ANNUITY"].fillna(ass.AMT_ANNUITY.median(),inplace=True)
```

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

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

```
ass.loc[ass['CODE_GENDER']=='XNA','CODE_GENDER']='F'
ass['CODE_GENDER'].value_counts()
```

Now, we categorized the column AMT_INCOME_RANGE, AMT_INCOME_TOTAL and DAYS_BIRTH as bins and add three more columns to the data frame.

```
bins=[0,25000,50000,75000,100000,125000,150000,175000,200000,225000,250000,275000,300000,325000,350000,375000,400000,425000,45
slot=['0-25000','25000-50000','50000-75000','75000-100000','100000-125000','125000-150000','150000-175000','175000-200000','20

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

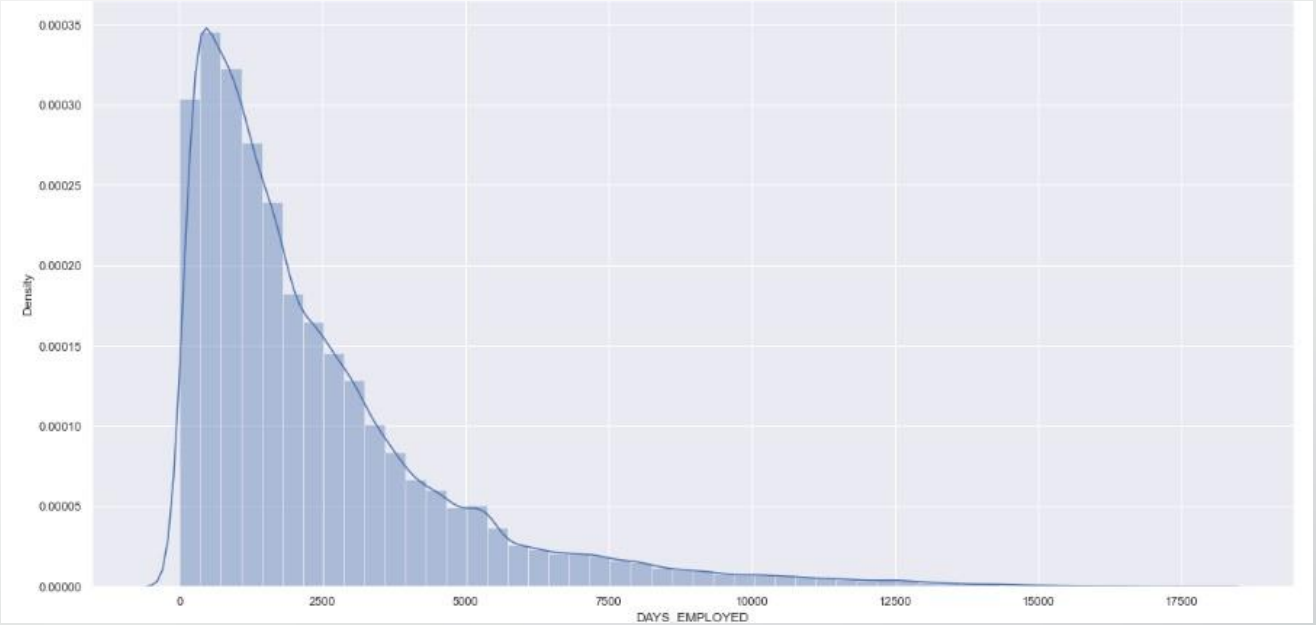
```
bins = [0,150000,200000,250000,300000,350000,400000,450000,500000,550000,600000,650000,700000,750000,800000,850000,900000,1000
slots = ['0-150000', '150000-200000','200000-250000', '250000-300000', '300000-350000', '350000-400000','400000-450000',
       '450000-500000','500000-550000','550000-600000','600000-650000','650000-700000','700000-750000','750000-800000',
       '800000-850000','850000-900000','900000 and above']

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

At last we have seprated the target value according to 0 and 1

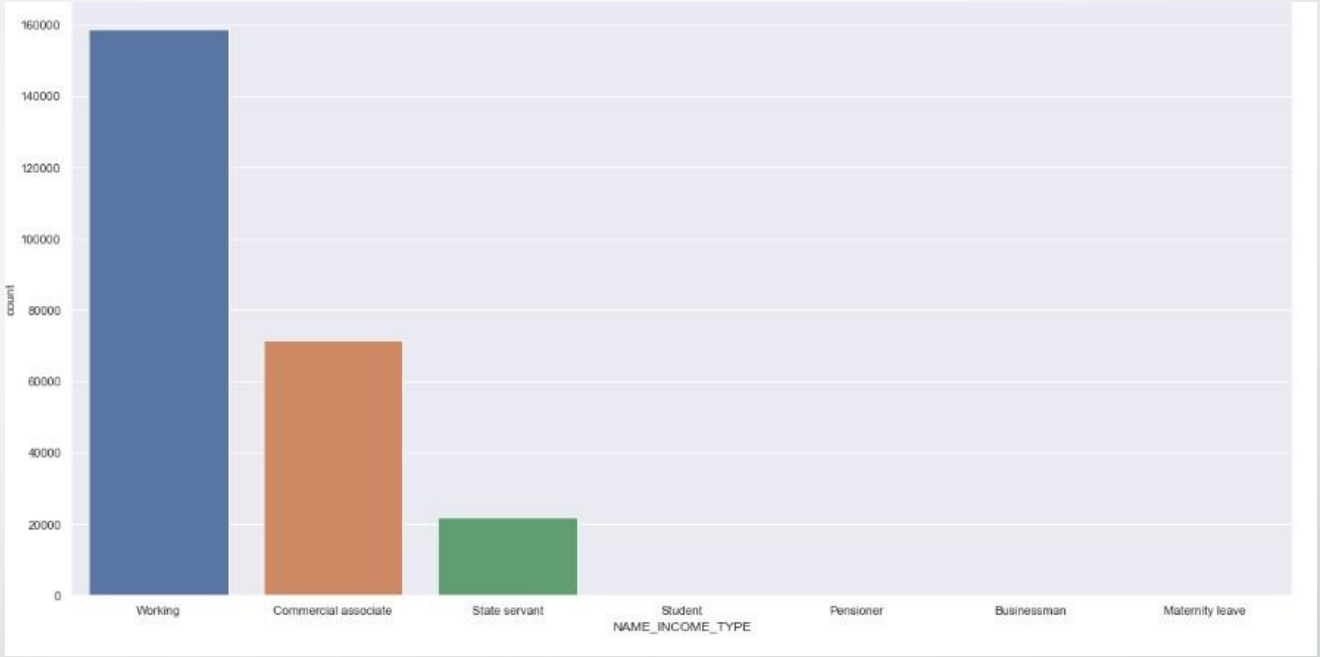
```
t0=ass.loc[ass["TARGET"]==0]
t1=ass.loc[ass["TARGET"]==1]
```

UNIVARIATE ANALYSIS

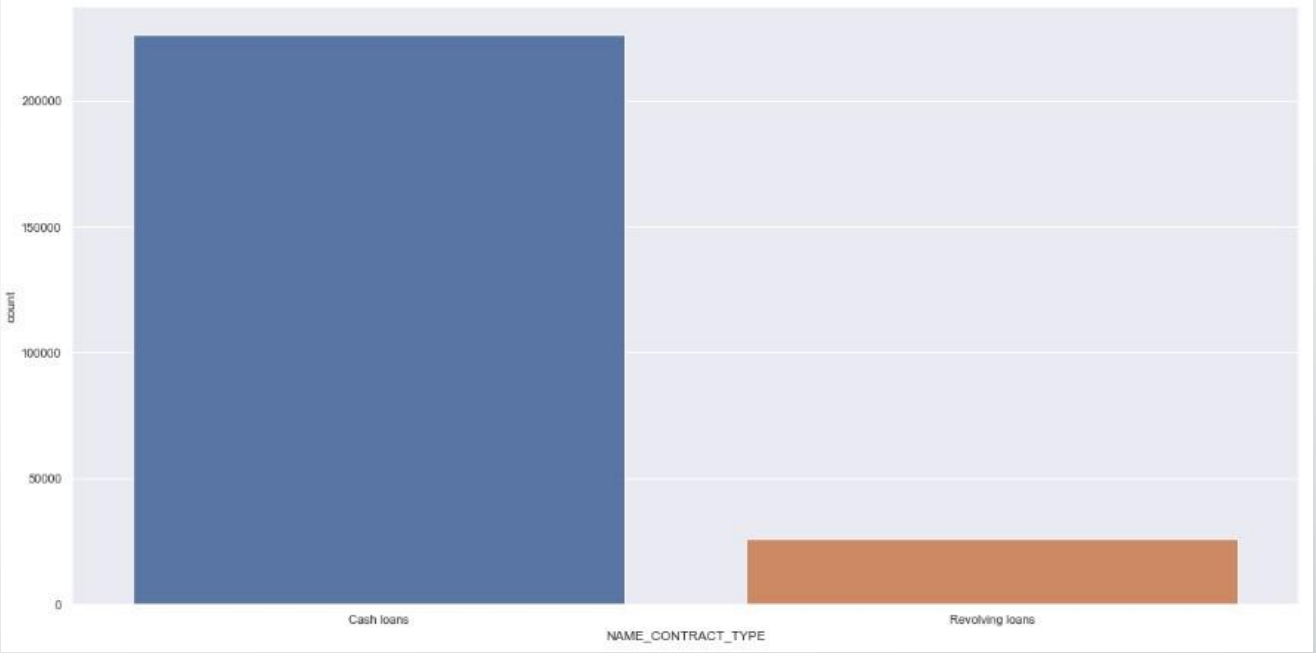
According to the graph we can see that there are few people which have been working from a long time. Mostly are less experienced according to there employment days.



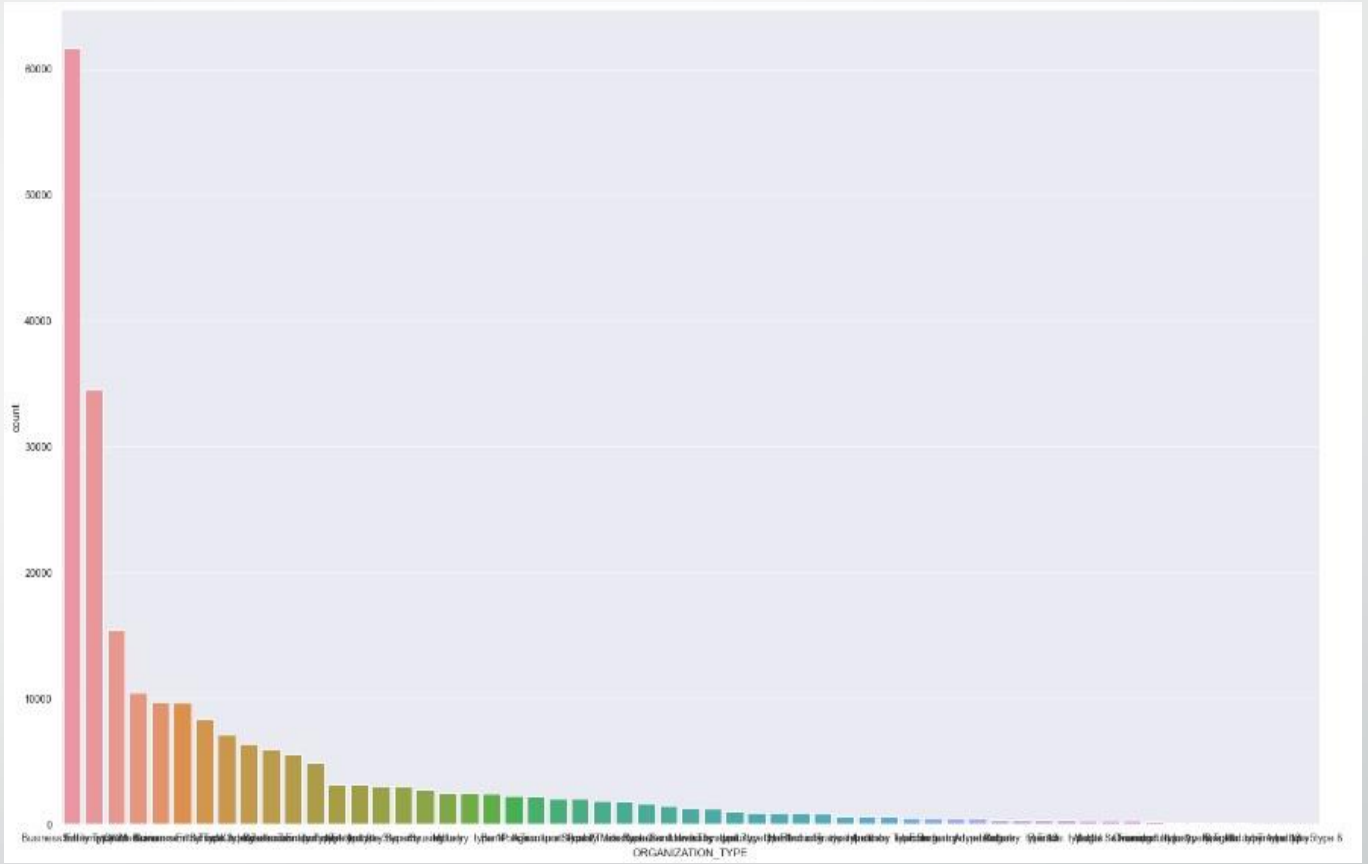
Now in this graph we can see that the most the people which have been contacted are working.



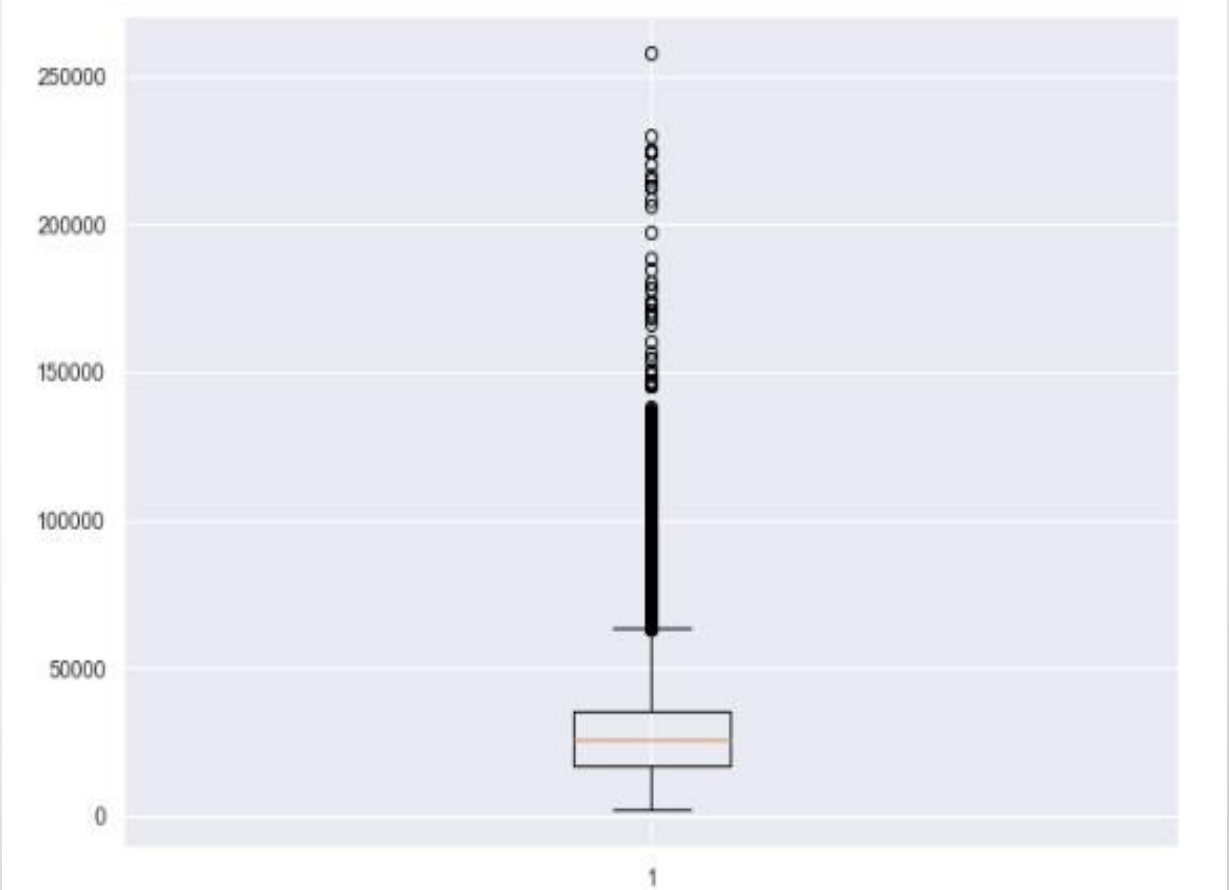
In this graph we can see that peoples are mostly taking cash loans.



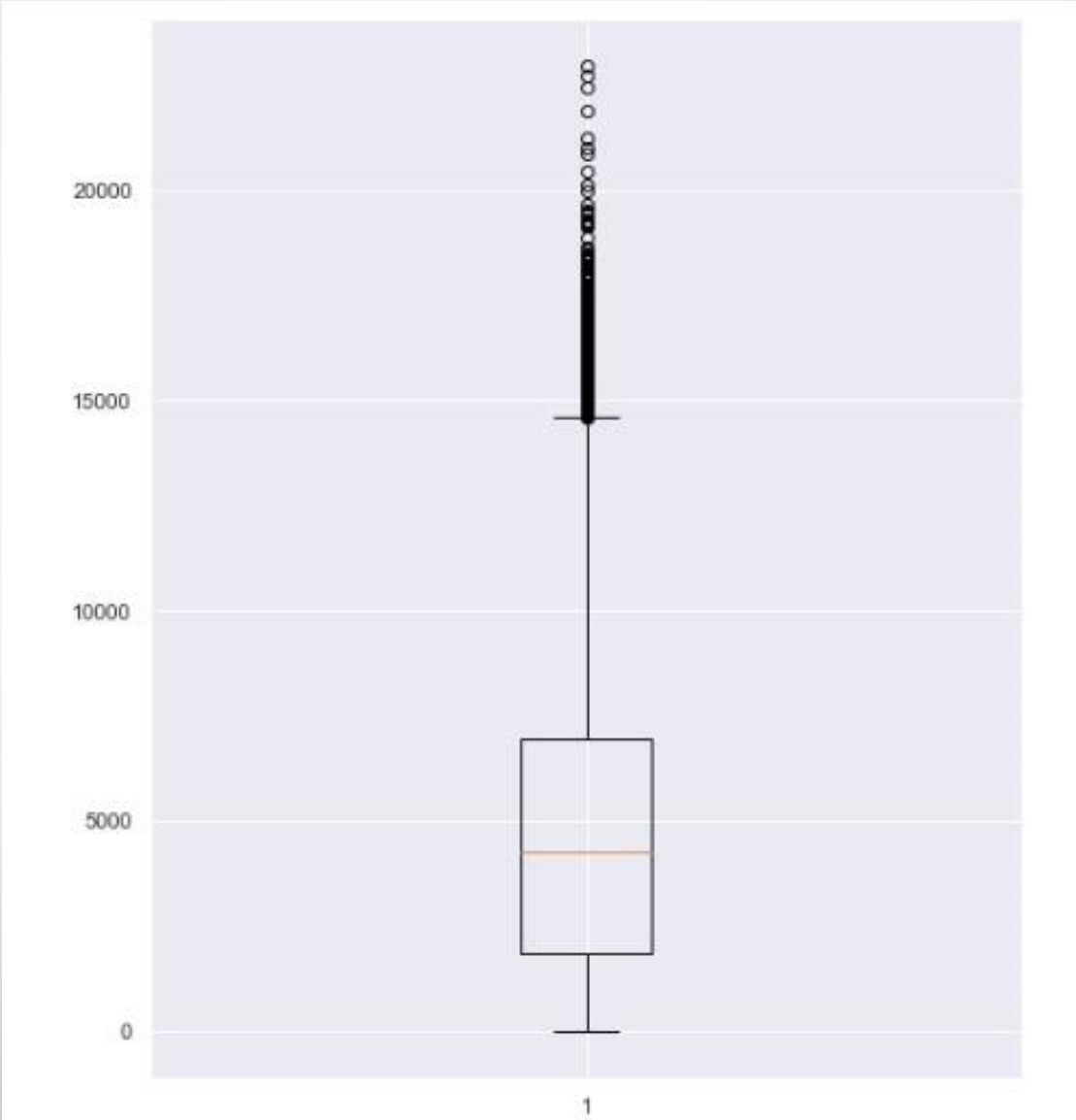
This graph is of Organization Type. To see that most of the people which have been contacted work in which organization.



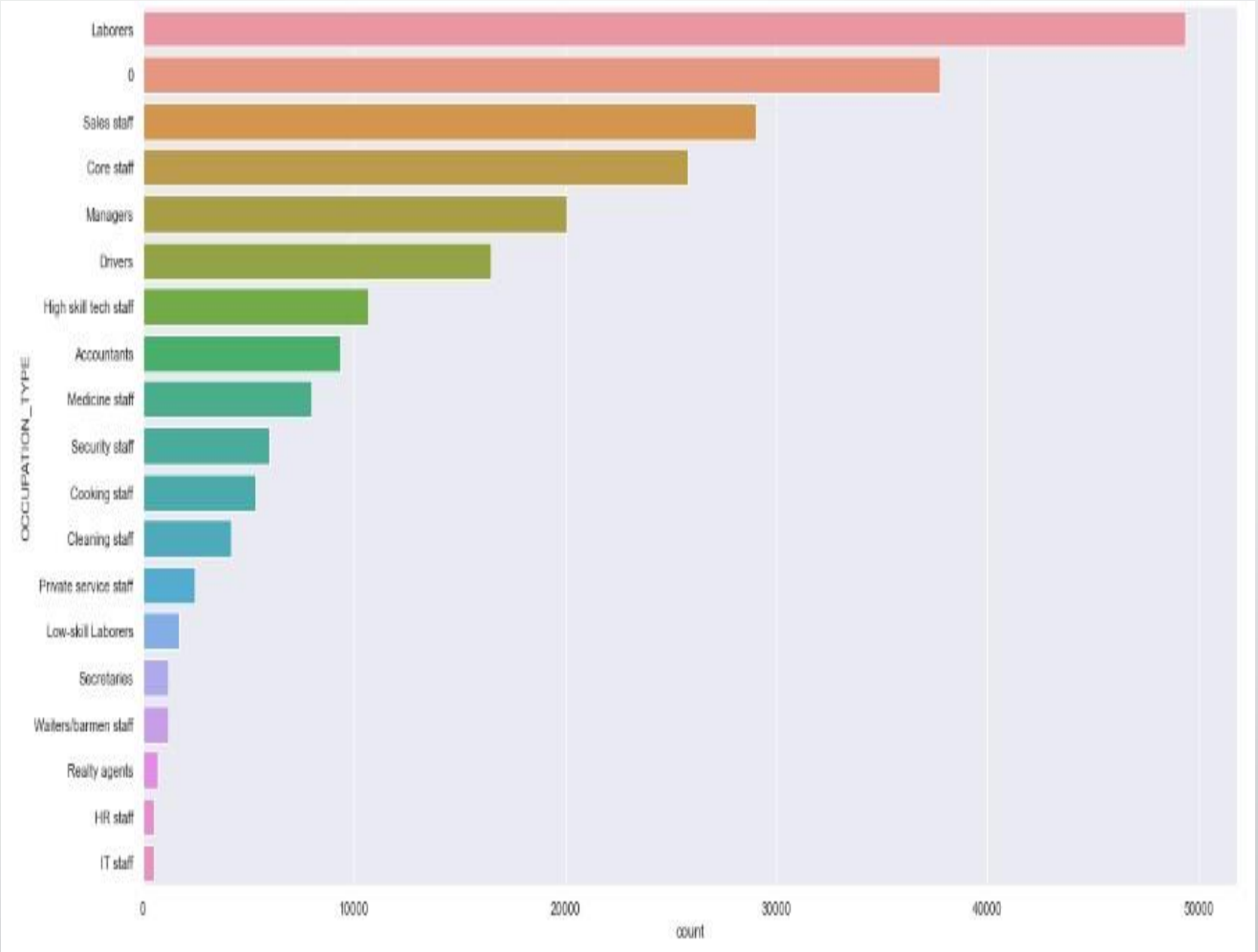
This is the graph of Amt Annuity. In this graph we can see that the there are many outliers in this column.



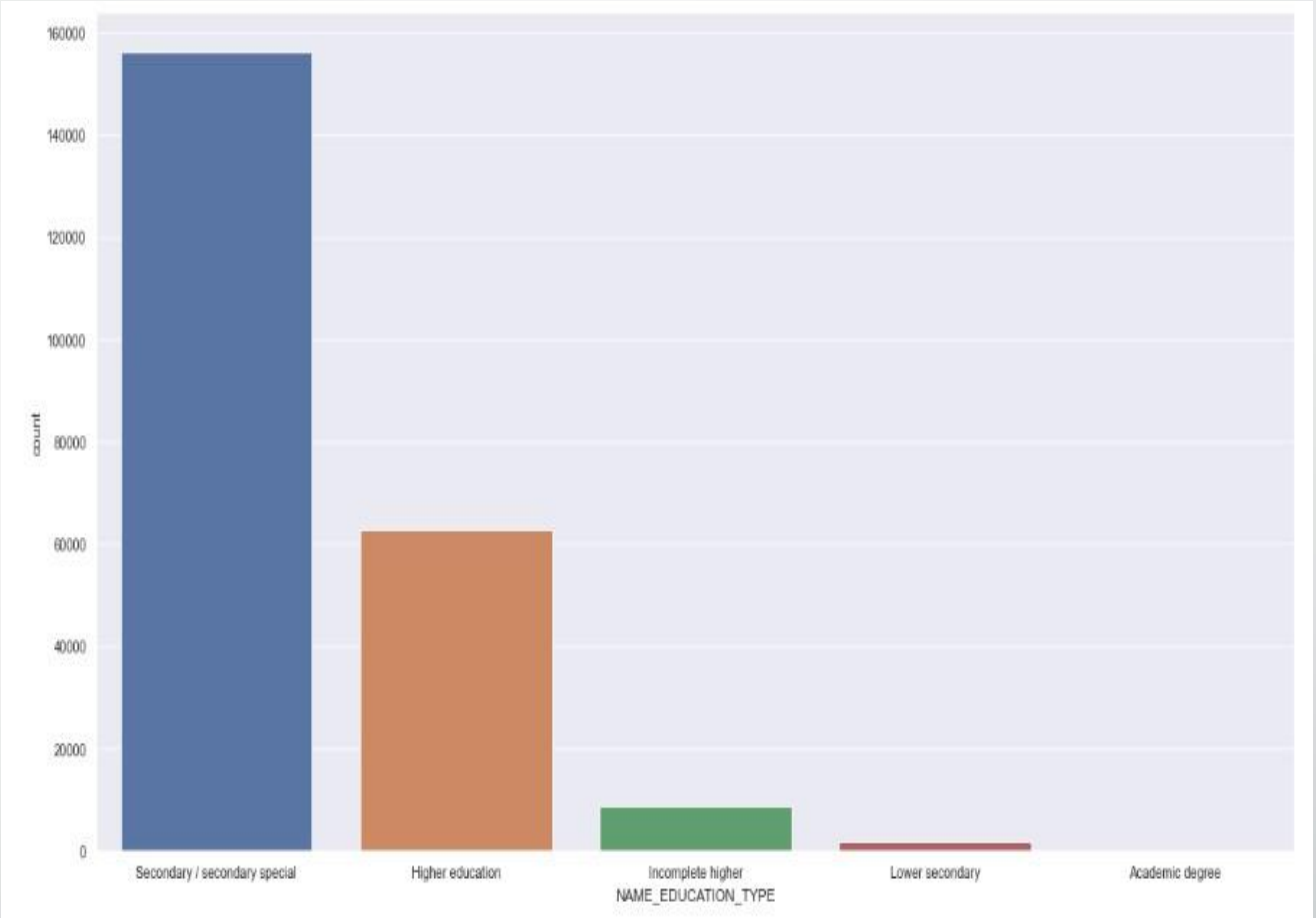
This is the graph of days of registration. According to this graph median is near 50,000.



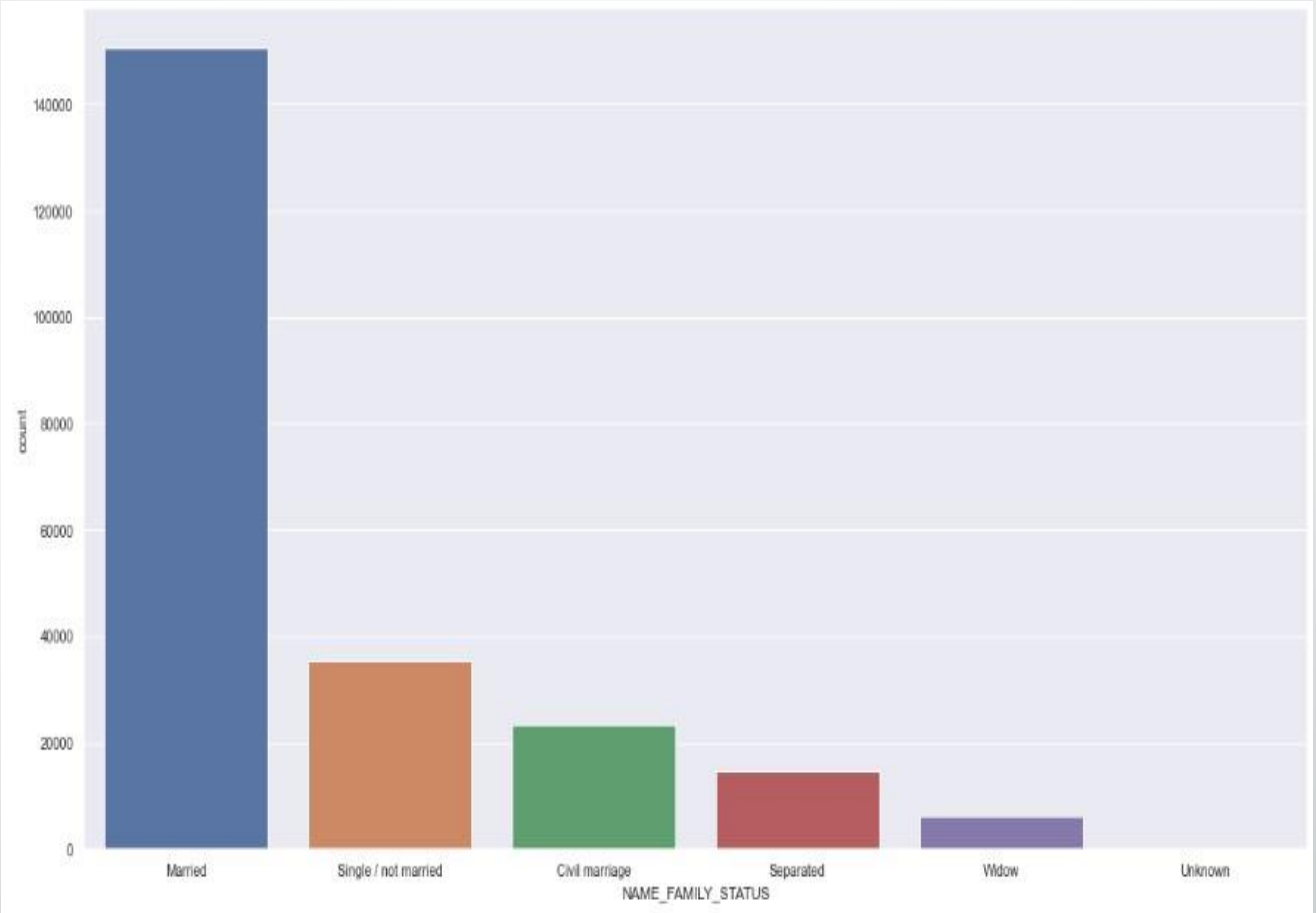
This graph is of occupation type. According to this labor is the occupation of the most peoples and most of the values are missing.



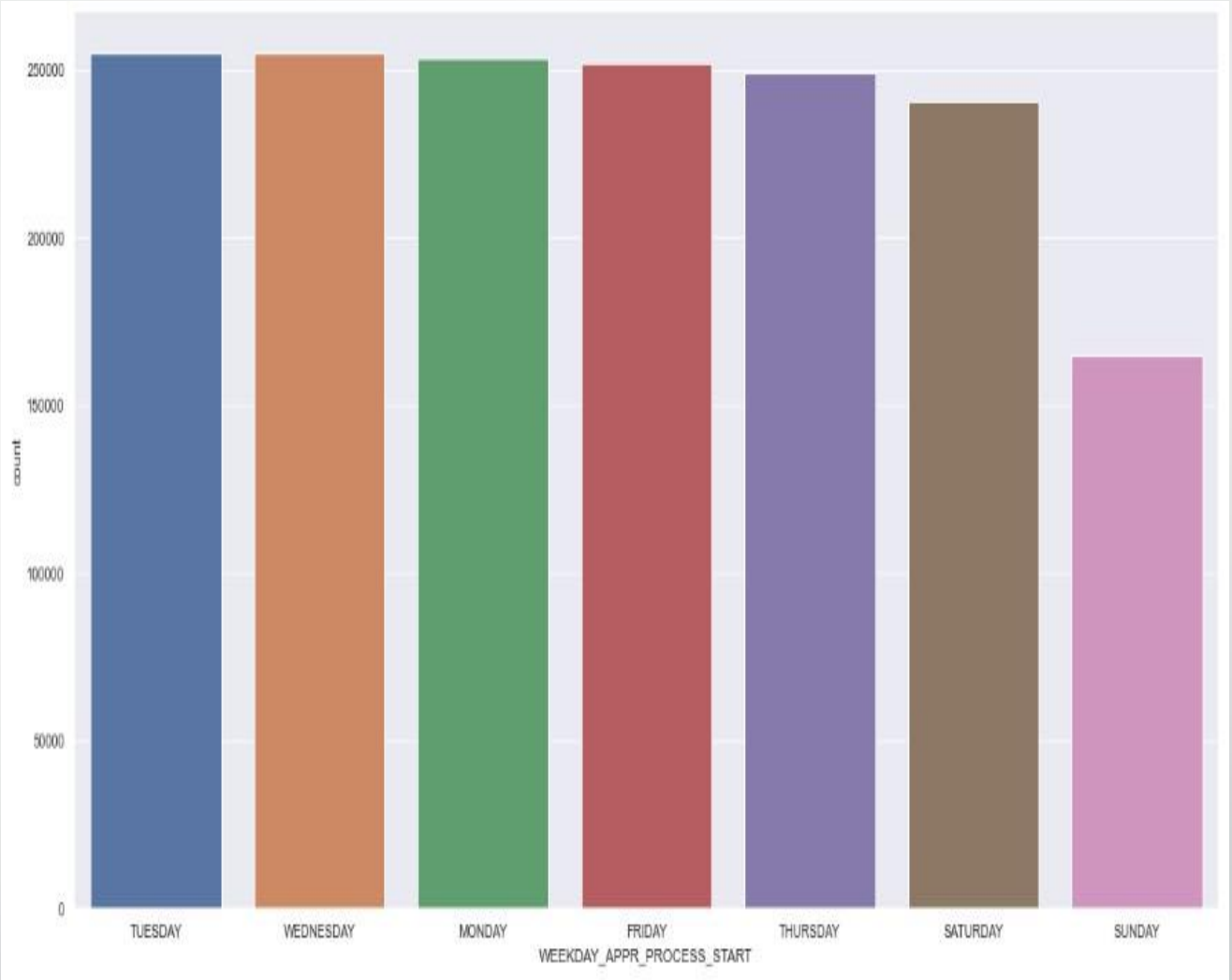
This graph is of Education type. This graph shows that most of the people have secondary or special education.



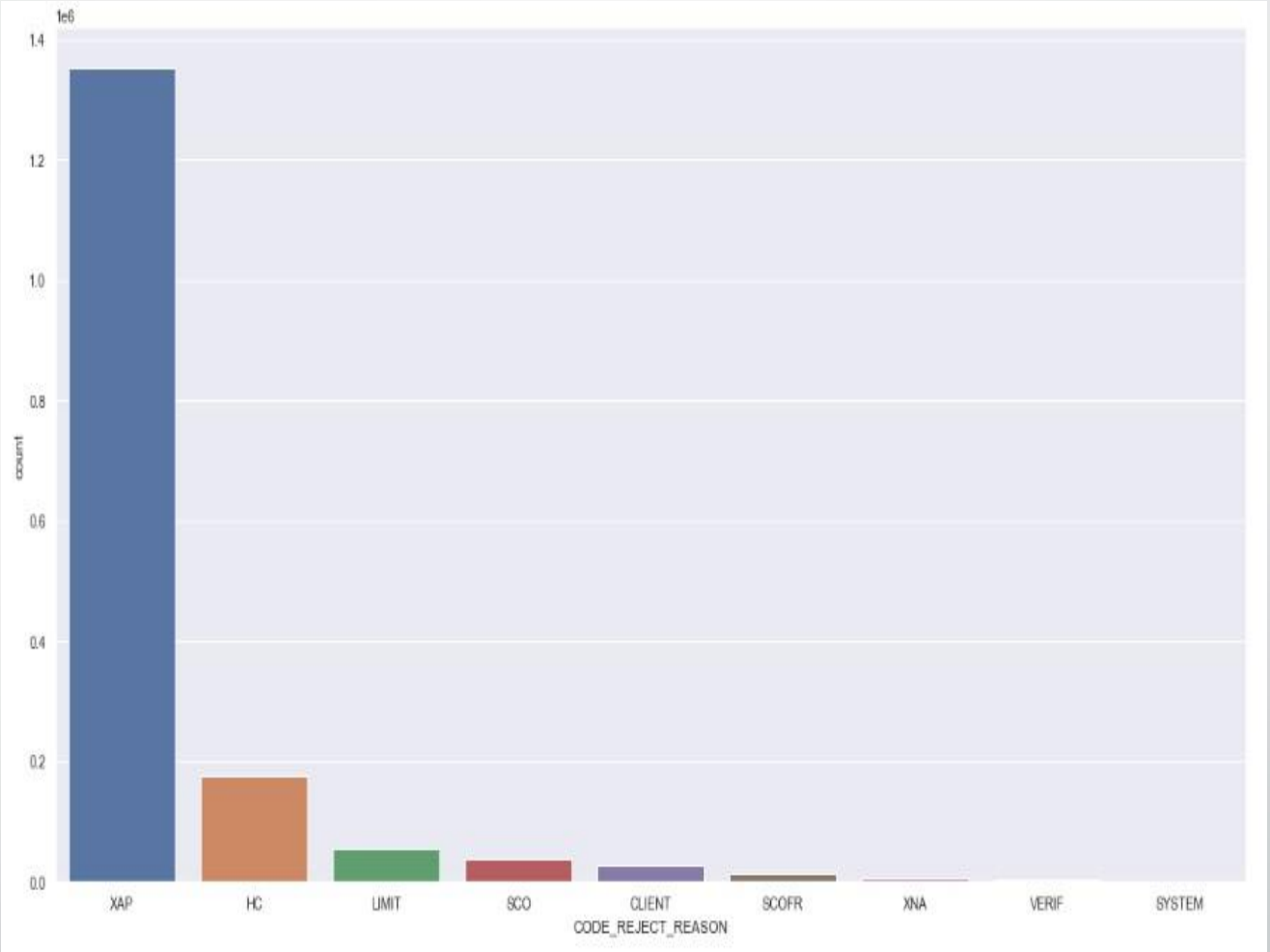
This graph is of Family status in which we can see that most of the people are married who have been contacted.



This graph is from previous application of column weekday application start process which shows less process start on Sunday.

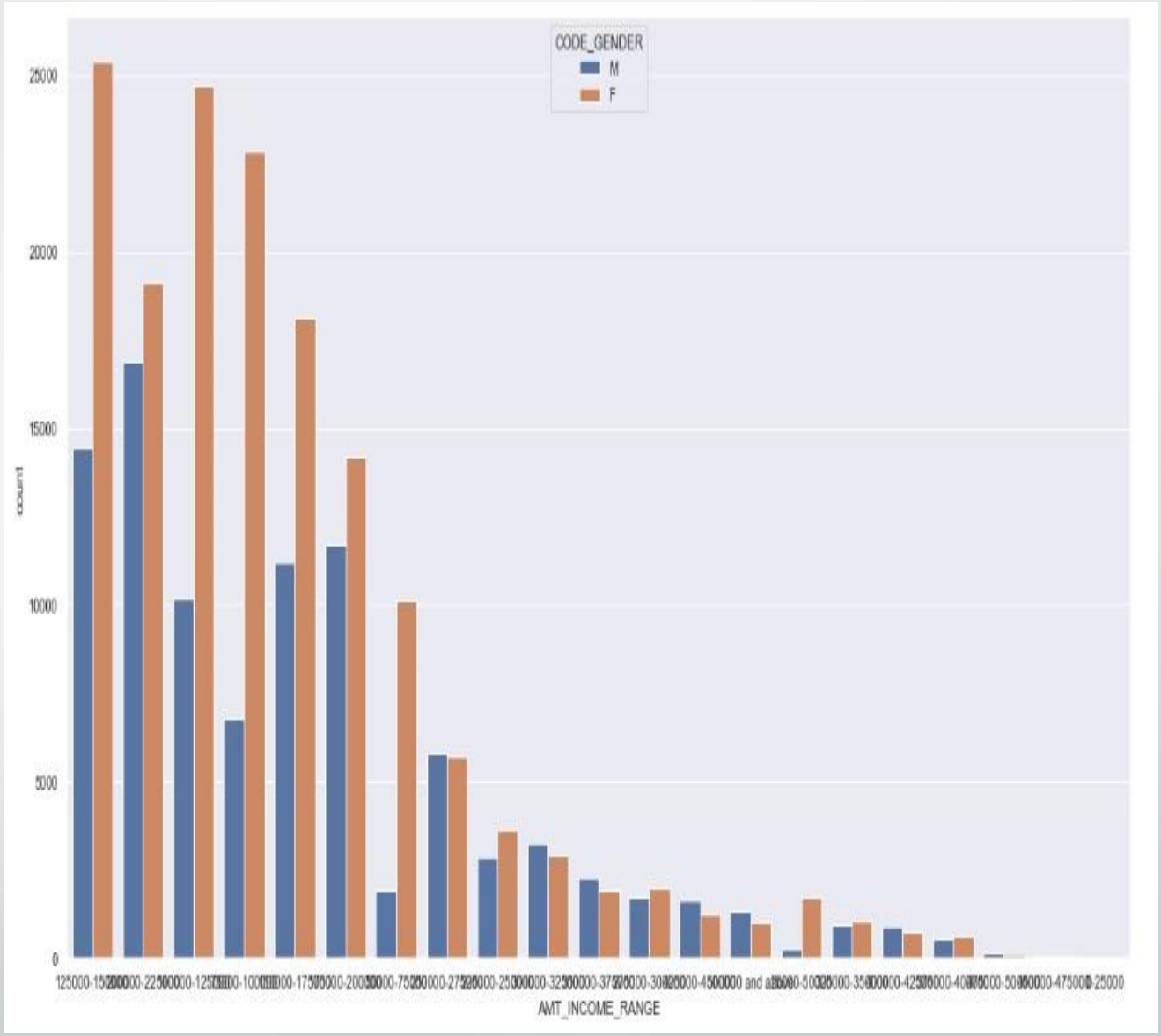


In this graph we can see the rejection reason of the most of the loan.

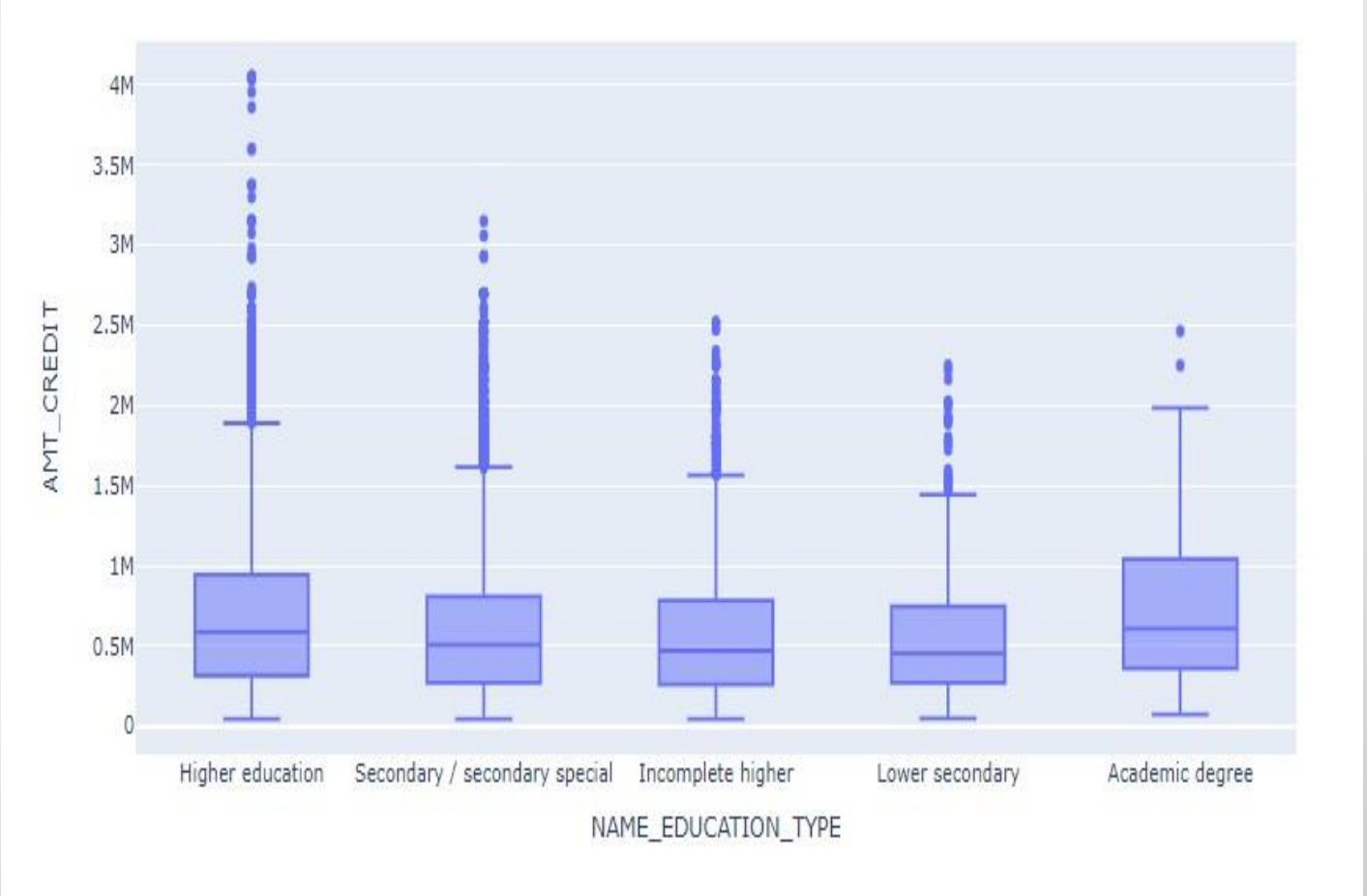


Bivariate Analysis

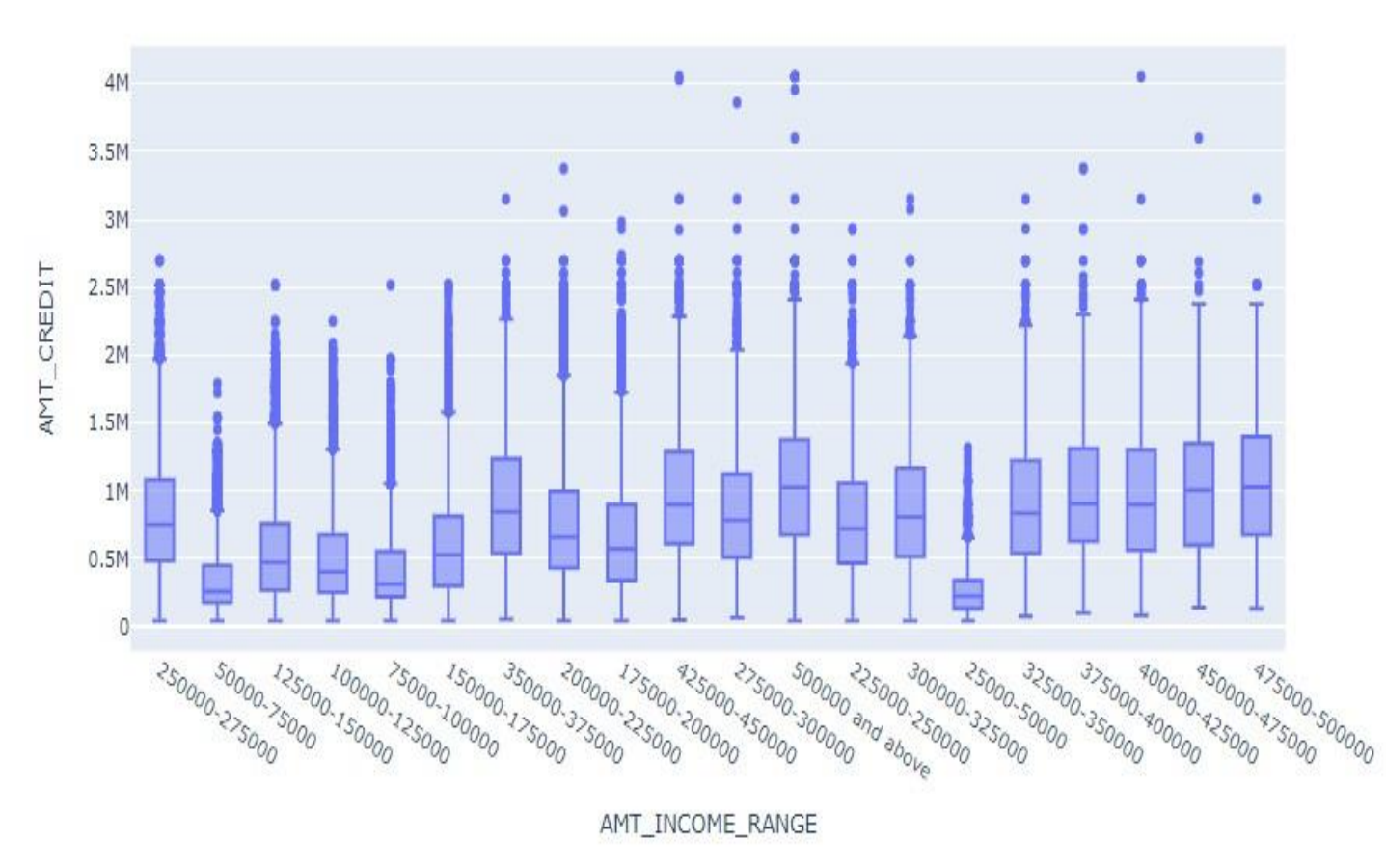
This graph is drawn between the code gender and amt income range. As we can see females are earning more than males.



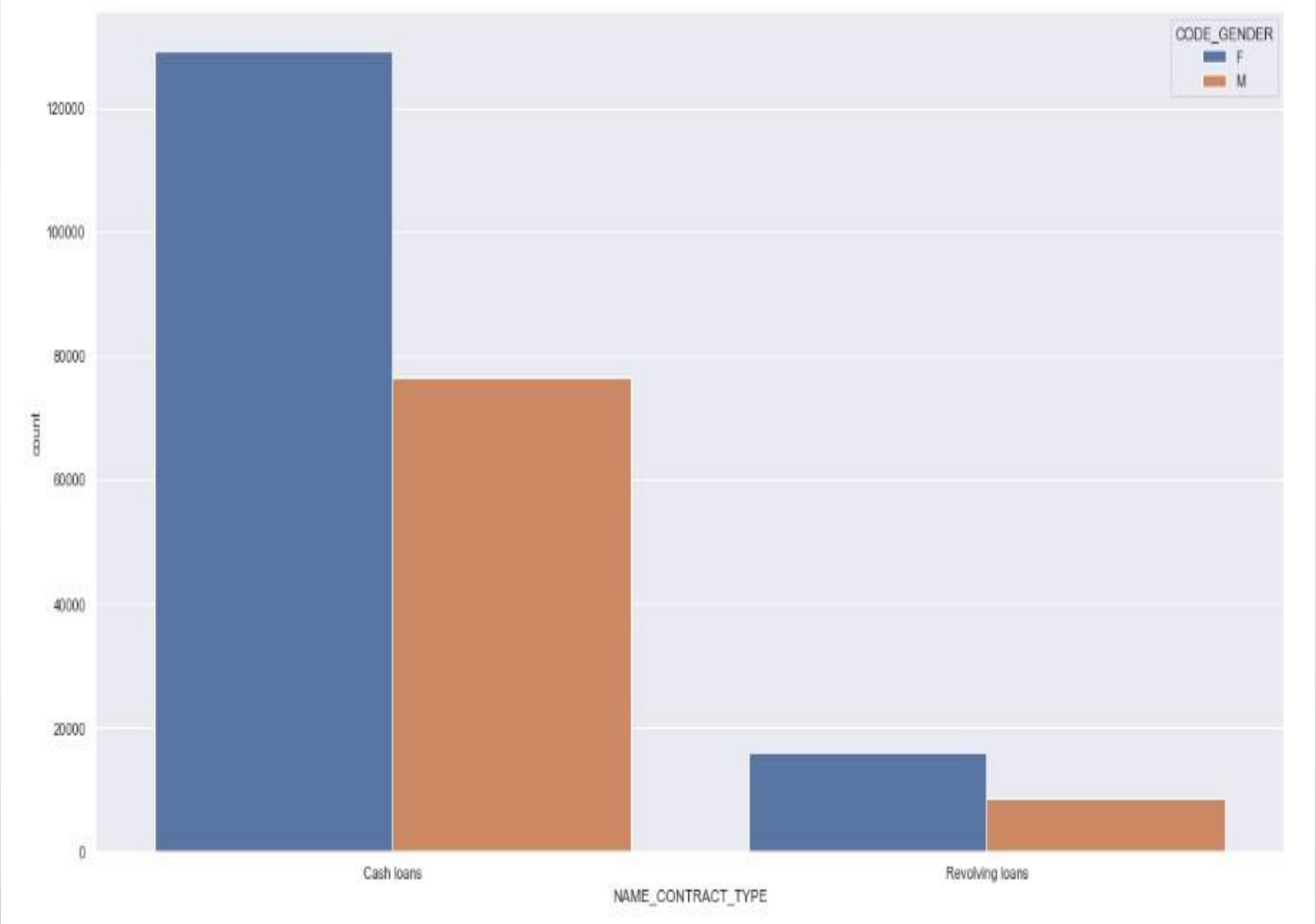
This graph is between amt credit and name education type. In this graph we can see that Academic degree peoples have more credit amount.

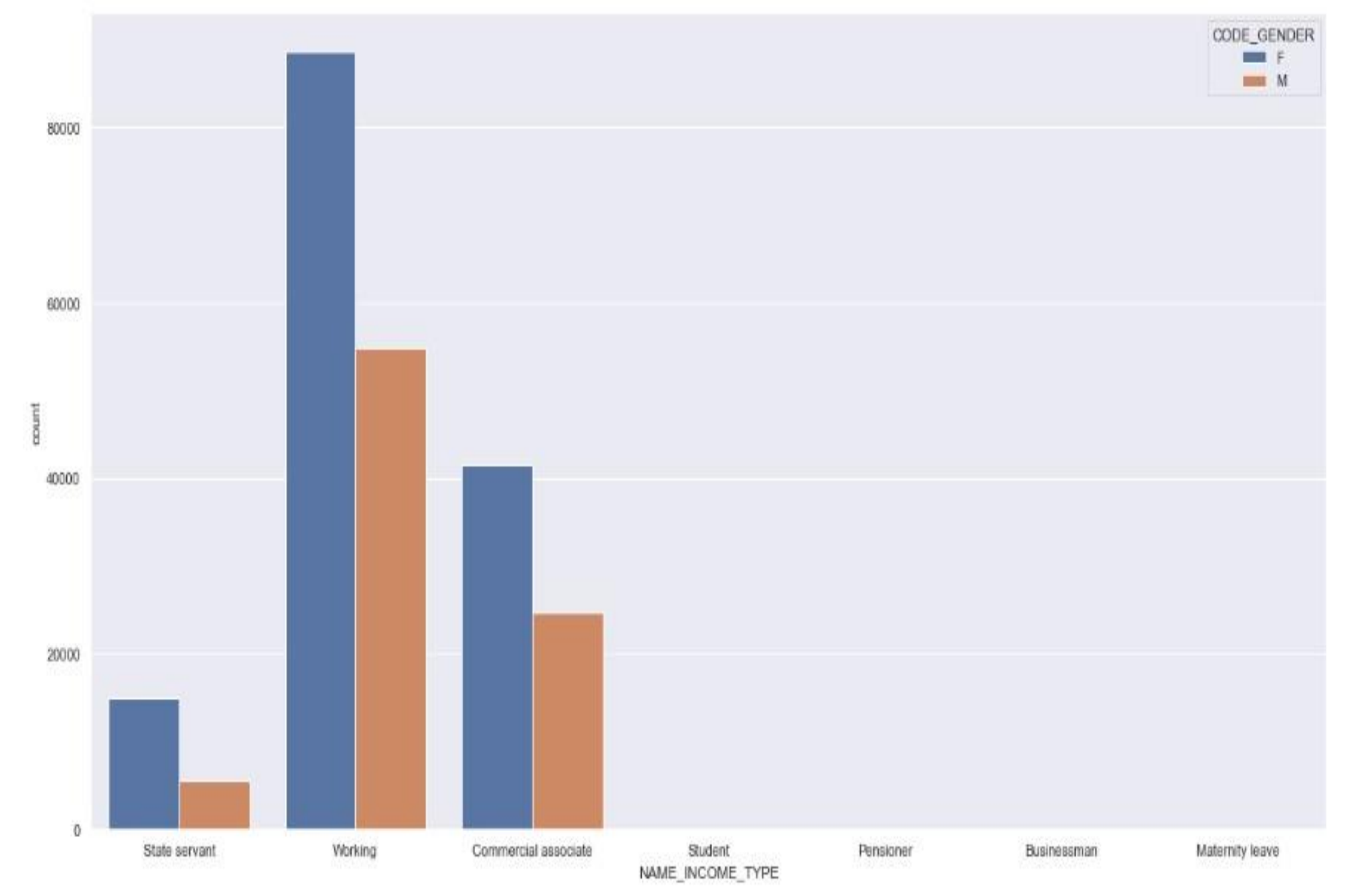


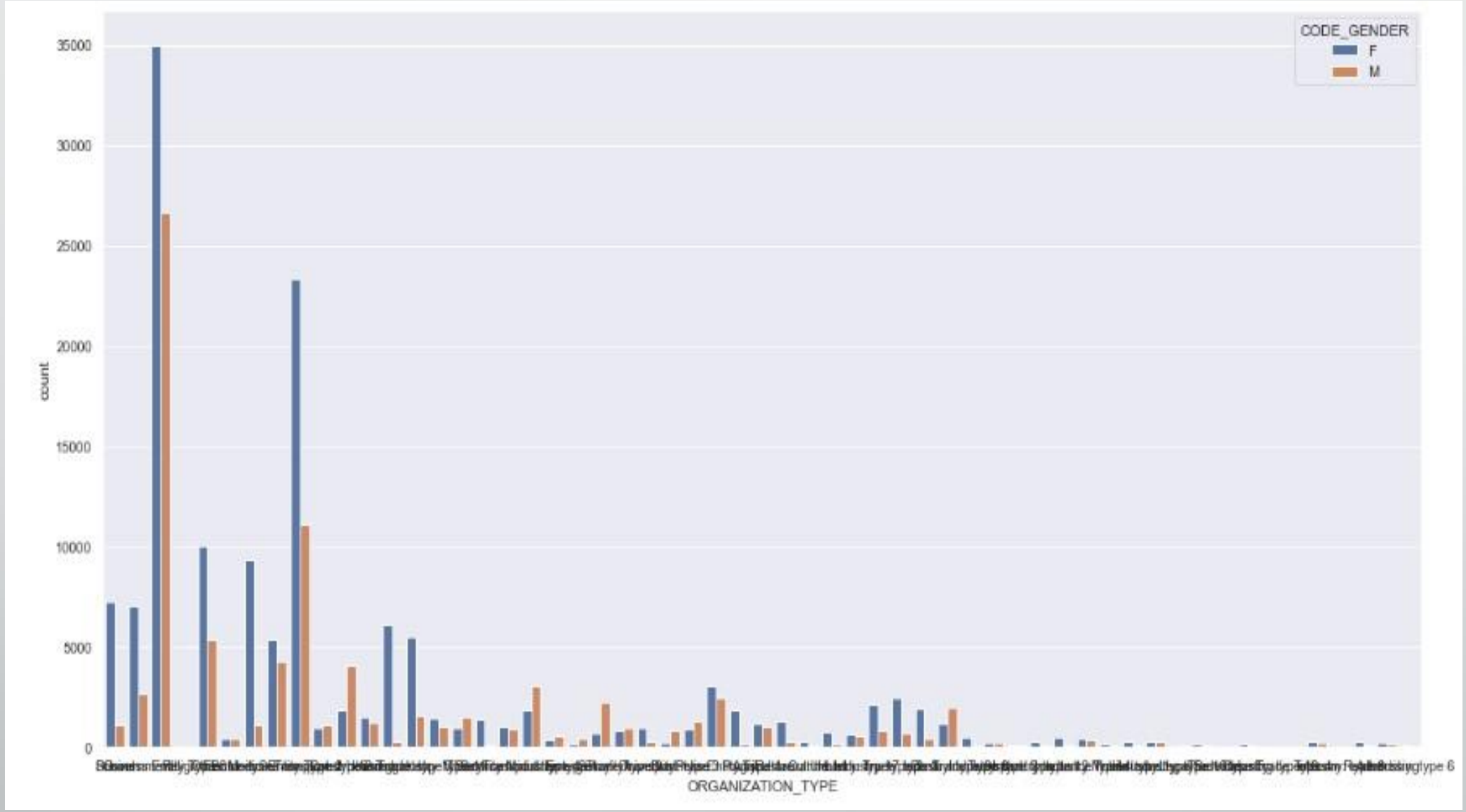
This graph is between amt credit vs amt income range

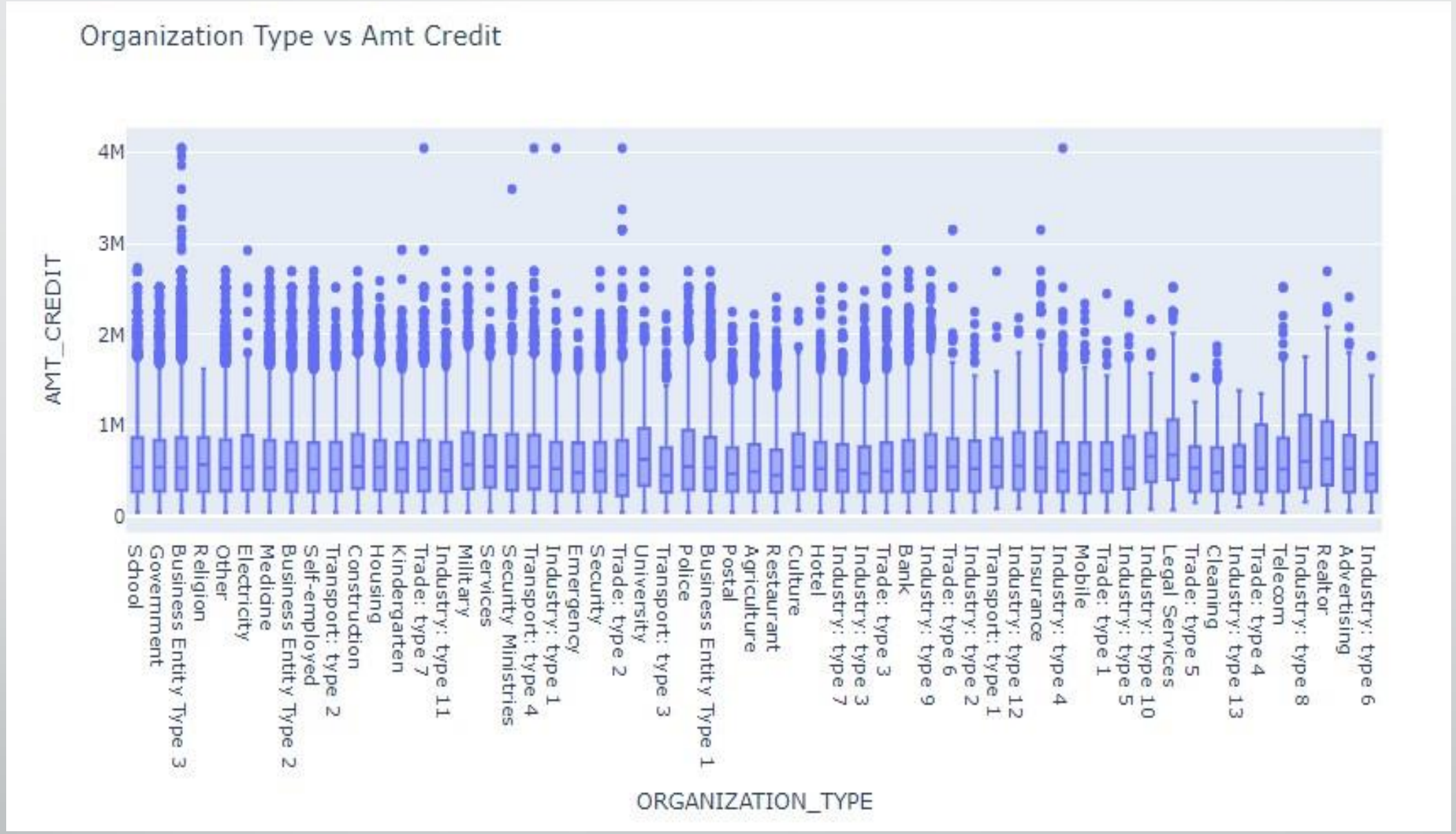


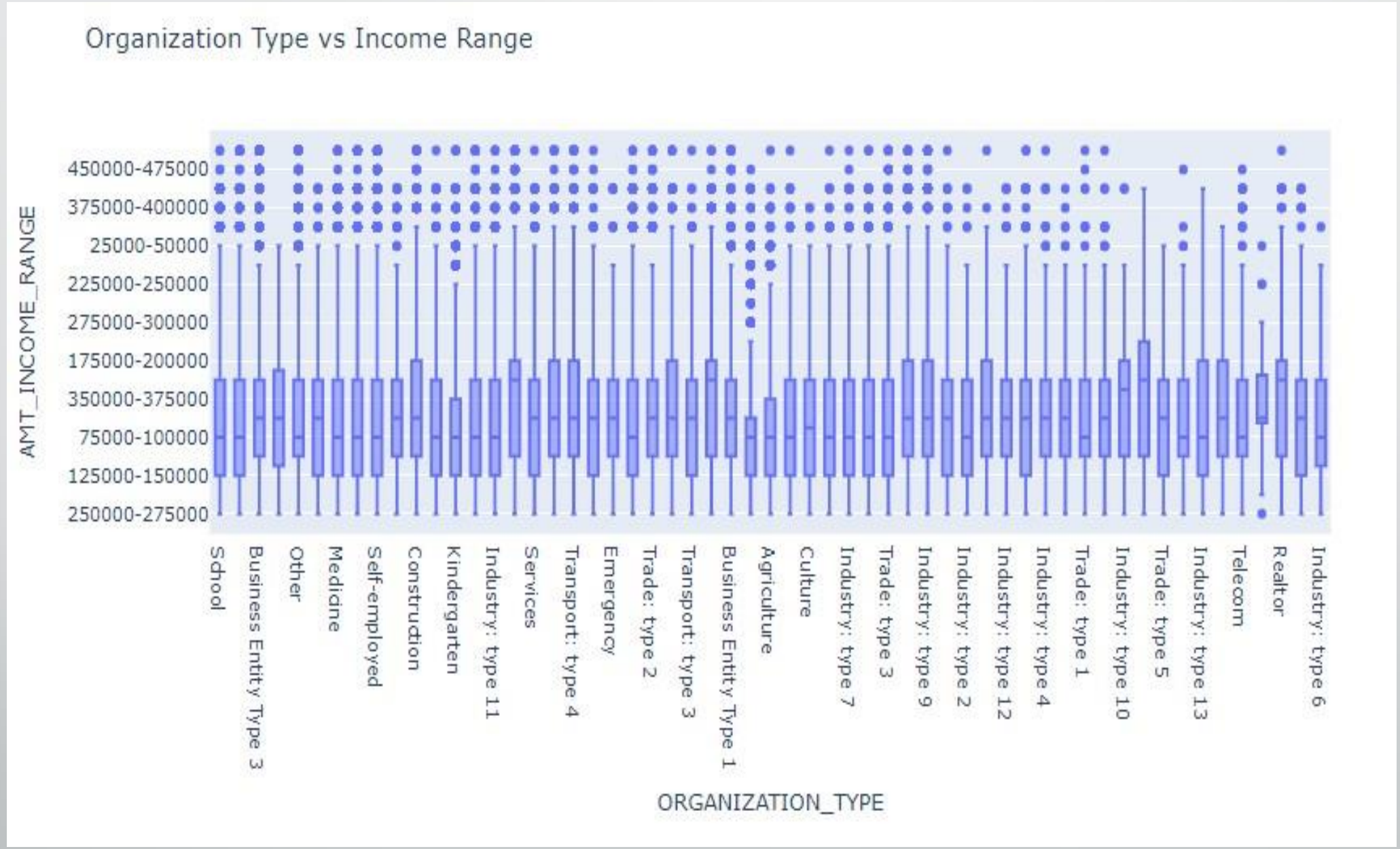
This contract type and code gender and we can see males take less loans then females

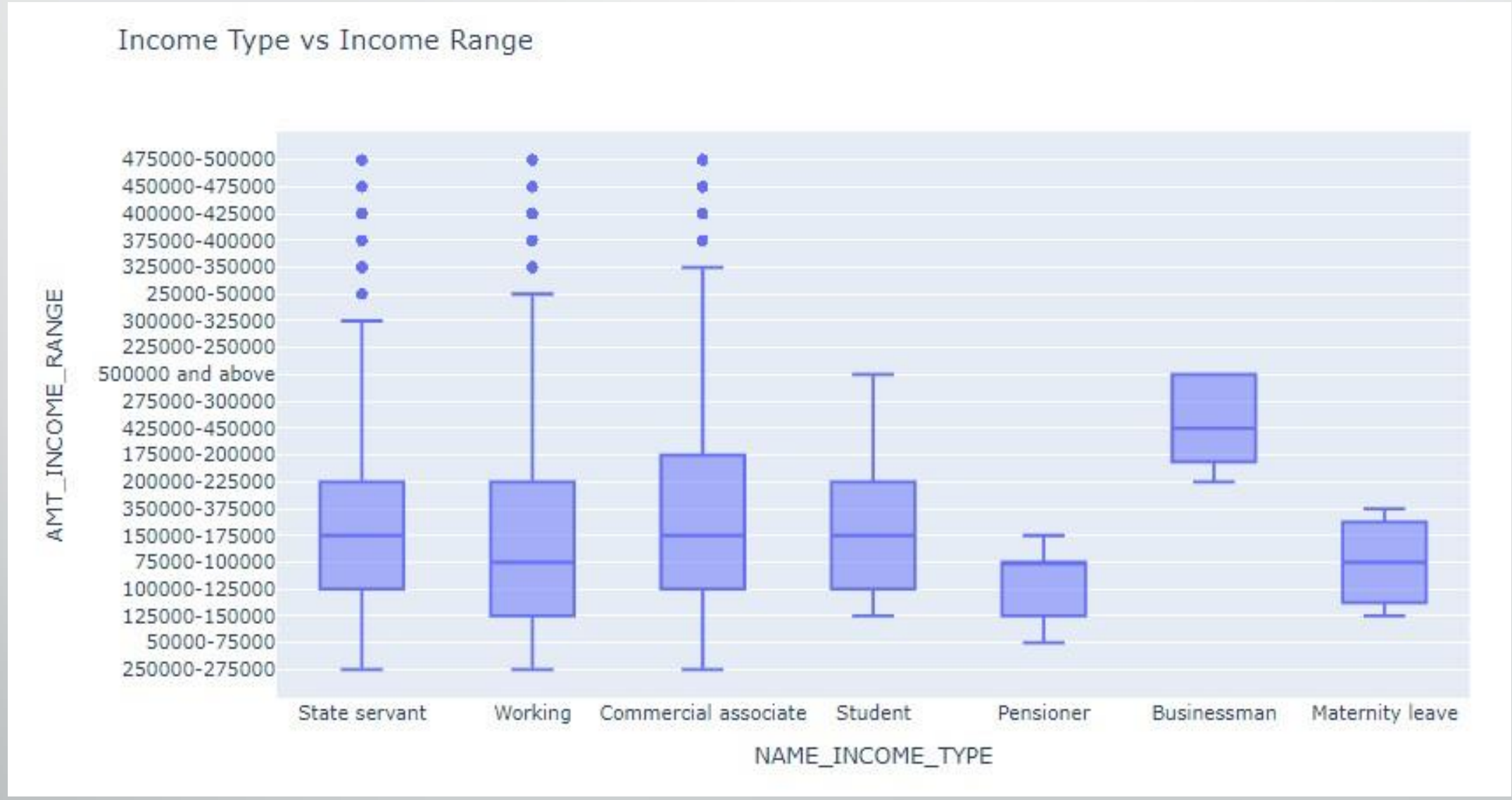



CODE_GENDER
F
M
80000
60000
40000
20000
0
count
State servant
Working
Commercial associate
Student
Pensioner
Businessman
Maternity leave
NAME_INCOME_TYPE

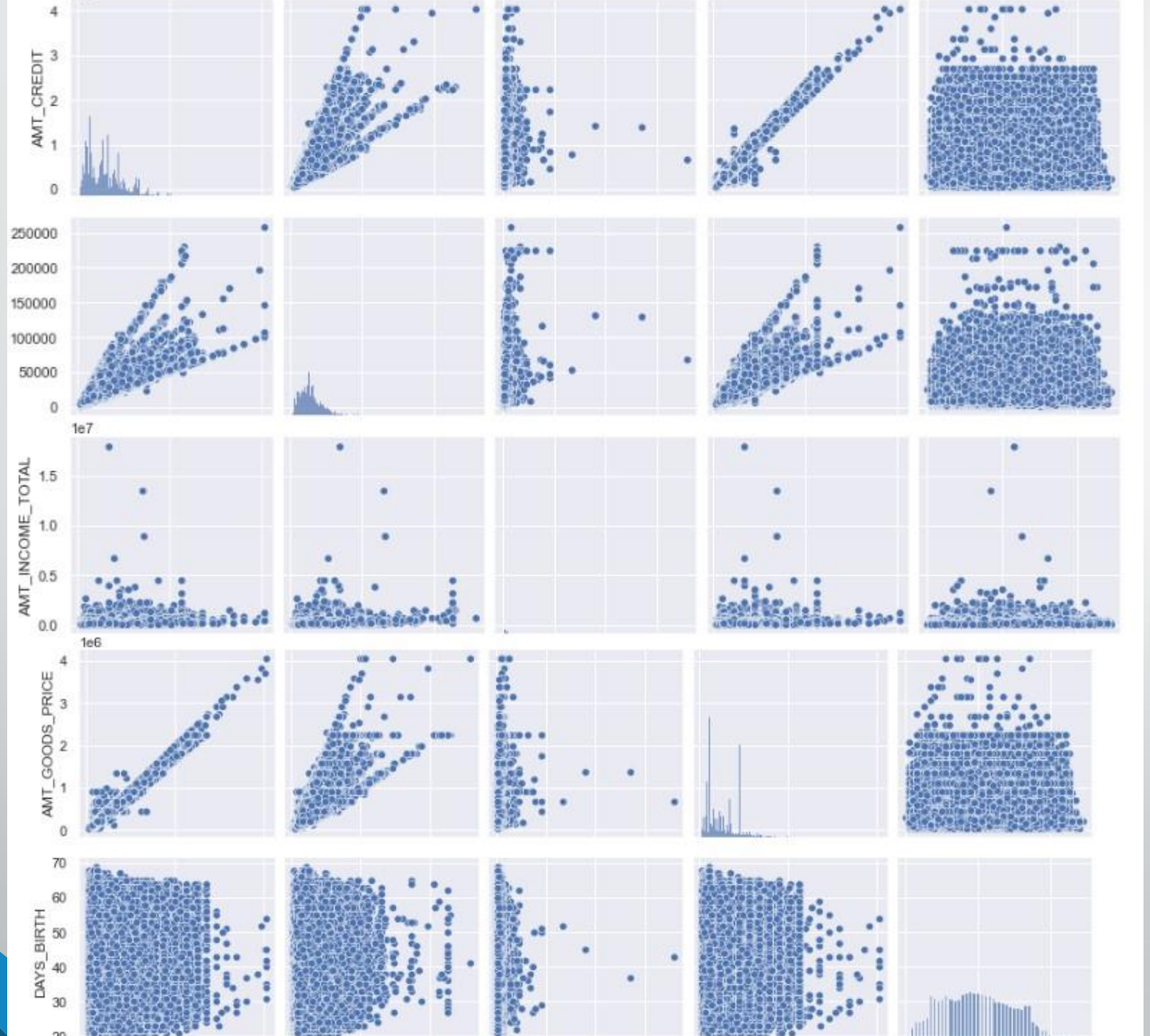

CODE_GENDER
F
M
35000
30000
25000
20000
15000
10000
5000
0
count
ORGANIZATION_TYPE

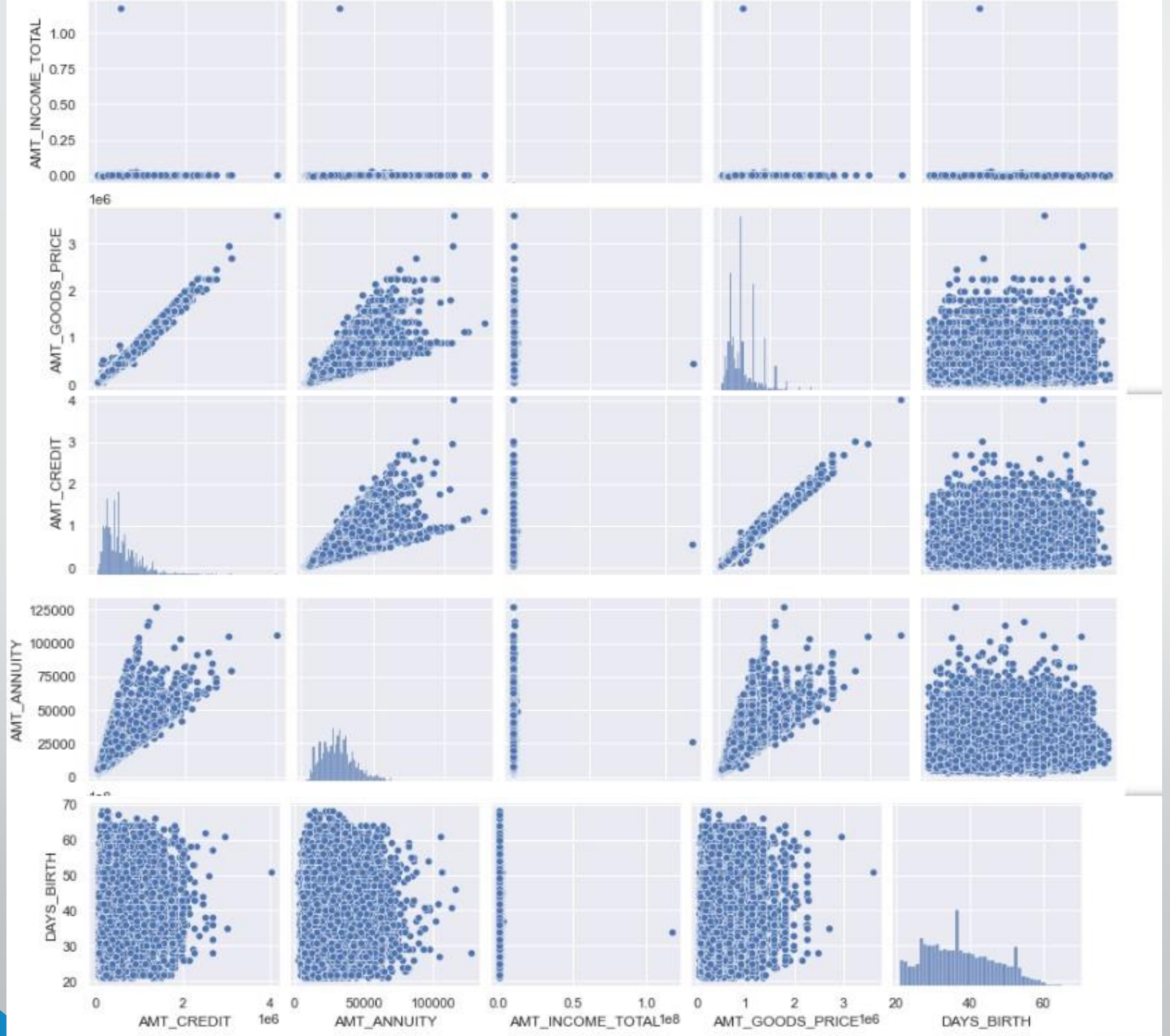

Organization Type vs Amt Credit
AMT_CREDIT
4M
3M
2M
1M
0
Industry: type 6
Advertising
Realtor
Industry: type 8
Telecom
Trade: type 4
Industry: type 13
Cleaning
Trade: type 5
Legal Services
Industry: type 10
Industry: type 5
Trade: type 1
Mobile
Industry: type 4
Insurance
Industry: type 12
Transport: type 1
Industry: type 2
Trade: type 6
Industry: type 9
Bank
Trade: type 3
Industry: type 3
Industry: type 7
Hotel
Culture
Restaurant
Agriculture
Postal
Business Entity Type 1
Police
Transport: type 3
University
Trade: type 2
Security
Emergency
Industry: type 1
Transport: type 4
Security Ministries
Services
Military
Industry: type 11
Trade: type 7
Kindergarten
Housing
Construction
Transport: type 2
Self-employed
Business Entity Type 2
Medicine
Electricity
Other
Religion
Business Entity Type 3
Government
School
ORGANIZATION_TYPE


Organization Type vs Income Range
AMT_INCOME_RANGE
450000-475000
375000-400000
25000-50000
225000-250000
275000-300000
175000-200000
350000-375000
75000-100000
125000-150000
250000-275000
School
Business Entity Type 3
Other
Medicine
Self-employed
Construction
Kindergarten
Industry: type 11
Services
Transport: type 4
Trade: type 2
Emergency
Transport: type 3
Business Entity Type 1
Agriculture
Culture
Industry: type 7
Trade: type 3
Industry: type 9
Industry: type 2
Industry: type 12
Industry: type 4
Trade: type 1
Industry: type 10
Trade: type 5
Industry: type 13
Telecom
Realtor
Industry: type 6
ORGANIZATION_TYPE


Income Type vs Income Range
AMT_INCOME_RANGE
475000-500000
450000-475000
400000-425000
375000-400000
325000-350000
25000-50000
300000-325000
225000-250000
500000 and above
275000-300000
425000-450000
175000-200000
200000-225000
350000-375000
150000-175000
75000-100000
100000-125000
125000-150000
50000-75000
250000-275000
State servant
Working
Commercial associate
Student
Pensioner
Businessman
Maternity leave
NAME_INCOME_TYPE

Co-relation Analysis


AMT_CREDIT
AMT_INCOME_TOTAL
AMT_GOODS_PRICE
DAYS_BIRTH
1e7
1e6


AMT_INCOME_TOTAL
AMT_GOODS_PRICE
AMT_CREDIT
AMT_ANNUITY
DAYS_BIRTH
1e6
1e8
AMT_CREDIT
AMT_ANNUITY
AMT_INCOME_TOTAL
AMT_GOODS_PRICE
DAYS_BIRTH

Conclusion

- Banks should focus on “Pensioners”, “Student” and “Businessman” for successful payments.
- Banks should less focus on Co-op apartment living peoples.
- Banks should less on income type ‘Working’ because they have less successful payments.
- Get more customer from housing type ‘With parents’ because they have less unsuccessful payments.

Thank You