Bank Loan Case Study

KORADA SAIKIRAN

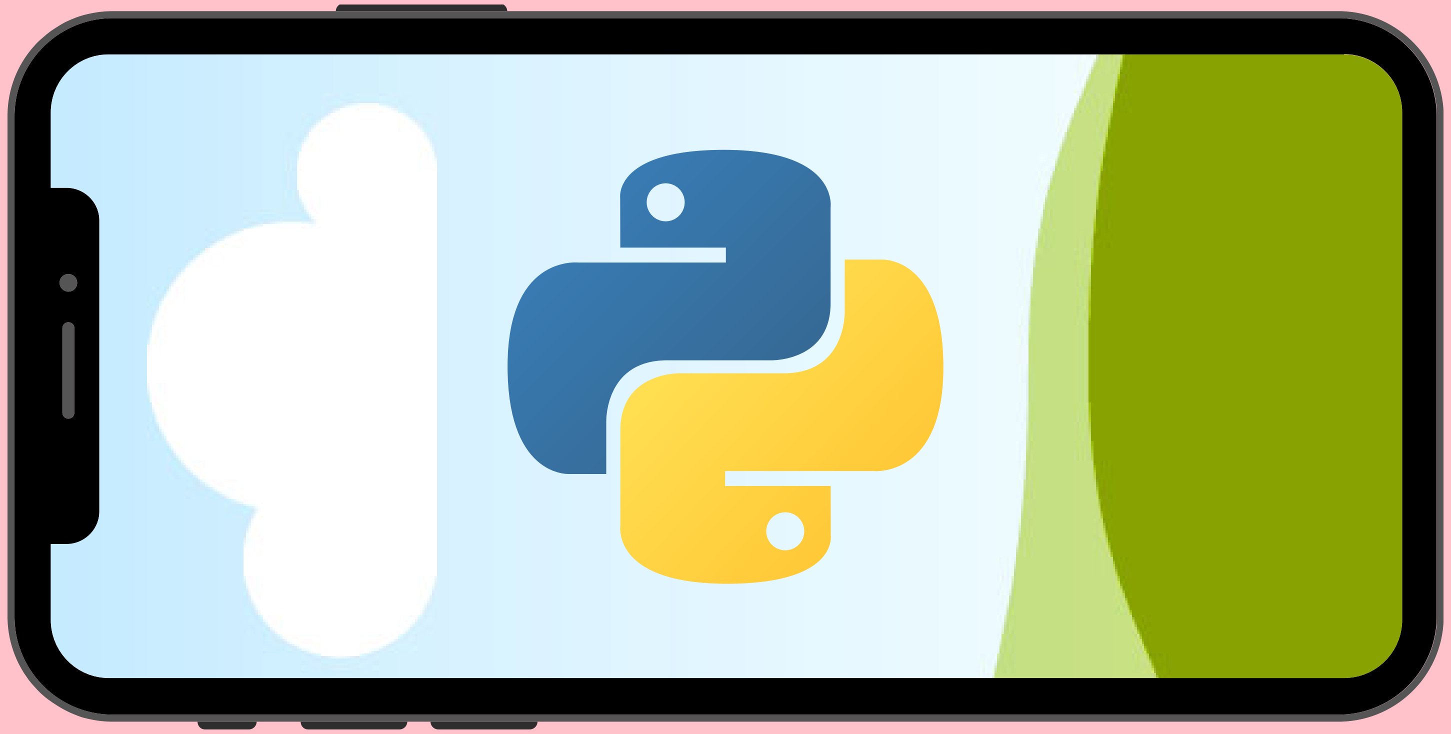
Problem statement

Objective is to analyze the data given to us and reduce the risk of biasing in approval of loan to the applicants between the potential payers and others

Business objectives

Identification of type of applicants using EDA, so that we can take actions like hiking interest rates, decreasing credit limit, declining the loan to the client who are facing payment difficulties. And also to reduce the risk of approving loan to the non potential clients instead of the potential clients (biasing)

Tech stack used



APPROACH

Framed a 4 step analysis process

U

Understanding data

TRIED TO UNDERSTAND THE DATA GIVEN TO START ANALYZING THE PROBLEM

E

Performing EDA in excel

Analyzing the data using variate analysis which is a part of EDA, to recognize the patterns and understand relation between features, and also data cleaning.

P

Python usgae

Also used python for EDA for extensive understanding

S

Summarize insights

Summarizing the insights from the analysis made

Data cleaning

As we are given a huge dataset having lakhs of records(observations), so dropping each record would be a mess and prolonged process that too huge data dealing with excel. So I calculated the missing value percentage feature for each given feature in the dataset so that, we can delete the highest missing percentage feature. And also by observing the impact made by feature on the target which can help to drop the unnecessary features in the dataset. Similar type of features and useless features are also dropped. (used both excel and python to perform this cleaning process as its a huge dataset)

As this is a huge dataset, I just calculated the percentage of missing values per column and delete the column which has missing percentage > 55%
To check the columns with highest percentage,

in application_data

As its a huge dataset, It'll be better to calculate the missing value percentage for each feature

```
In [7]: total_record = df.shape[0]

        for col in df.columns:
            missing_record = df[col].isnull().sum()
            print("Name of the feature is: ",col)
            percentage_missing = (missing_record/total_record)*100
            print("the missing percentage of the feature: ", percentage_missing)
```

in previous_data, dropped features are

RATE_INTEREST_PRIMARY
RATE_INTEREST_PRIVILEGED

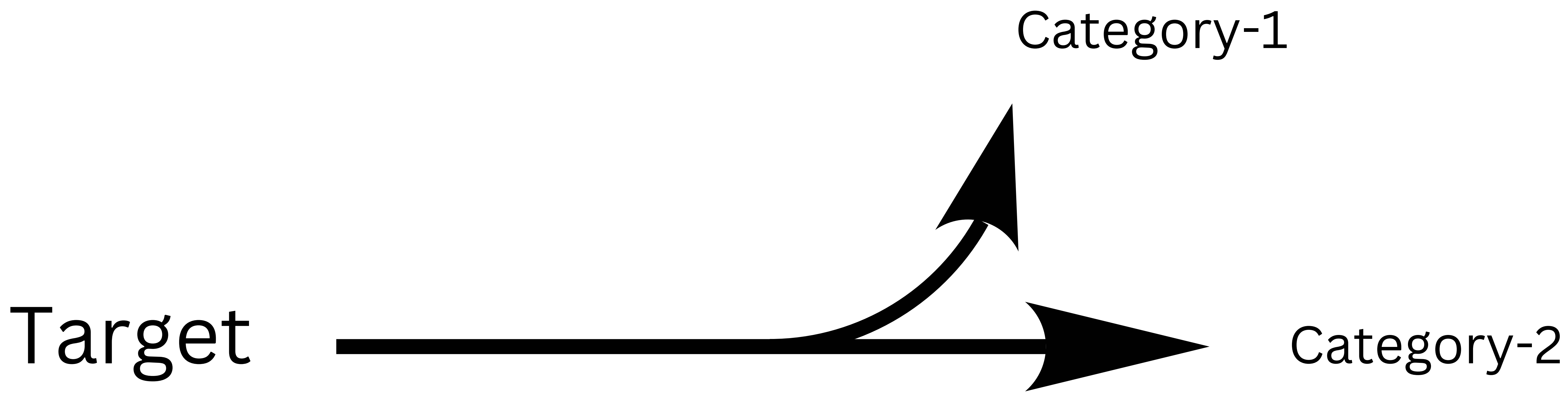
Above features have missing percentage more than 90%

column_name	missing_value_percentage
AMT_ANNUITY	0.0039022
AMT_GOODS_PRICE	0.0904032
NAME_TYPE_SUITE	0.4201475
OWN_CAR_AGE	65.99081
OCCUPATION_TYPE	31.345545
APARTMENTS_AVG	50.749729
BASEMENTAREA_AVG	58.5159555
YEARS_BUILD_AVG	66.497783
COMMONAREA_AVG	69.872297
ELEVATORS_AVG	53.295979
ENTRANCES_AVG	50.348768
FLOORSMIN_AVG	67.848629
LANDAREA_AVG	59.376737
LIVINGAPARTMENTS_AVG	68.354953
NONLIVINGAPARTMENTS_AVG	69.4329633
NONLIVINGAREA_AVG	55.179164
BASEMENTAREA_MODE	58.515955
YEARS_BUILD_MODE	66.497783
COMMONAREA_MODE	69.872297
FLOORSMIN_MODE	67.8486298
LANDAREA_MODE	59.3767377
LIVINGAPARTMENTS_MODE	68.3549531
NONLIVINGAPARTMENTS_MODE	69.432963
NONLIVINGAREA_MODE	55.179164
BASEMENTAREA_MEDI	58.515955
YEARS_BUILD_MEDI	66.4977838
COMMONAREA_MED	69.87229
FLOORSMIN_MEDI	69.87229
LANDAREA_MEDI	59.376737
LIVINGAPARTMENTS_MEDI	68.354953
NONLIVINGAPARTMENTS_MEDI	69.432963
NONLIVINGAREA_MEDI	55.17916432
FONDKAPREMONT_MODE	68.386171

The columns which have been dropped

Category-1 (1 valued) people are the clients who have late payment more than x days on at least one of the Y installments

Category-2 (0 valued) people are the clients who come under other cases like capable clients who can pay on time.

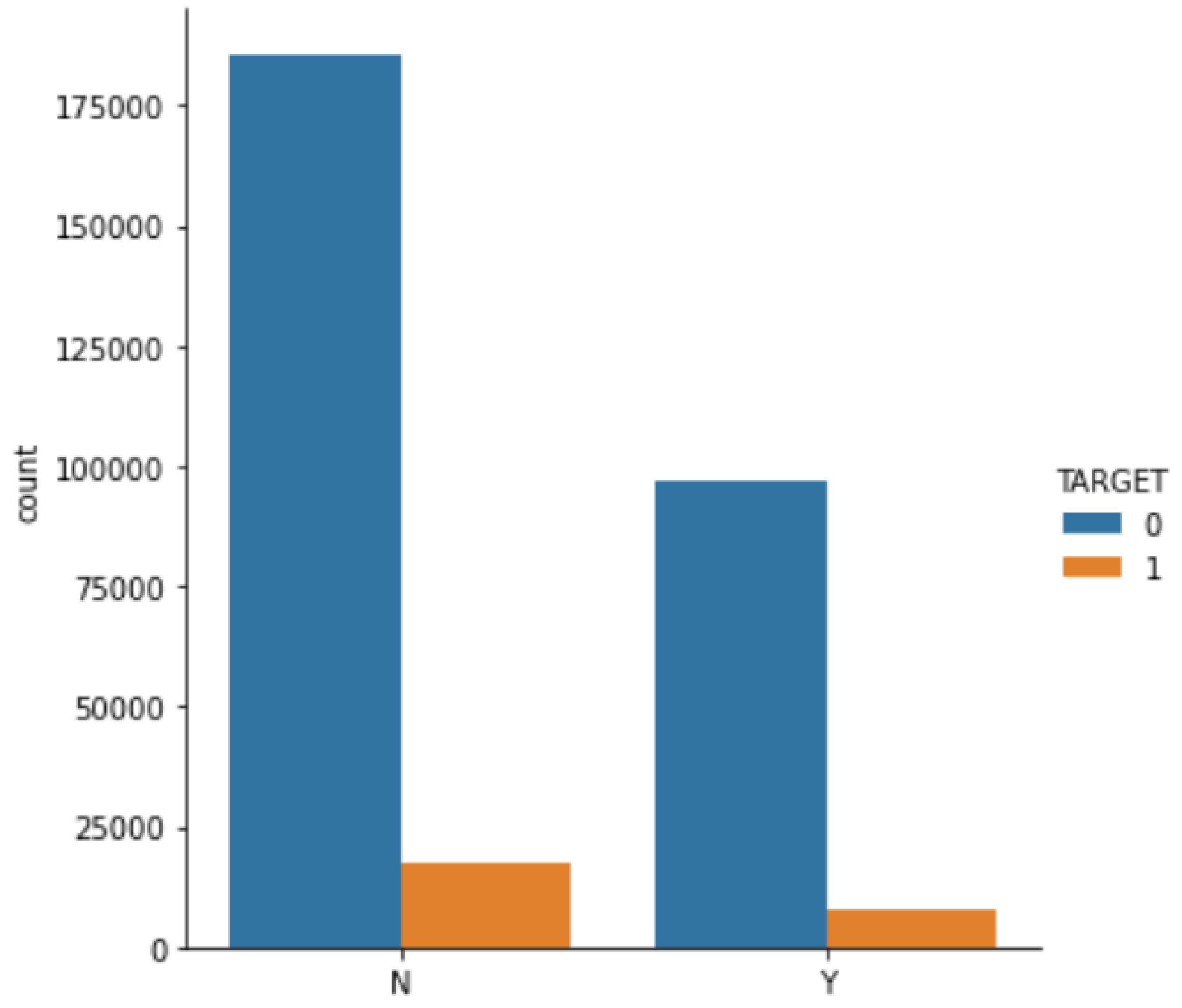


Analyzing all the FLAG variables, these features are heavily imbalance and not useful differentiator for the TARGET

To reduce the data imbalance, this step is to be performed

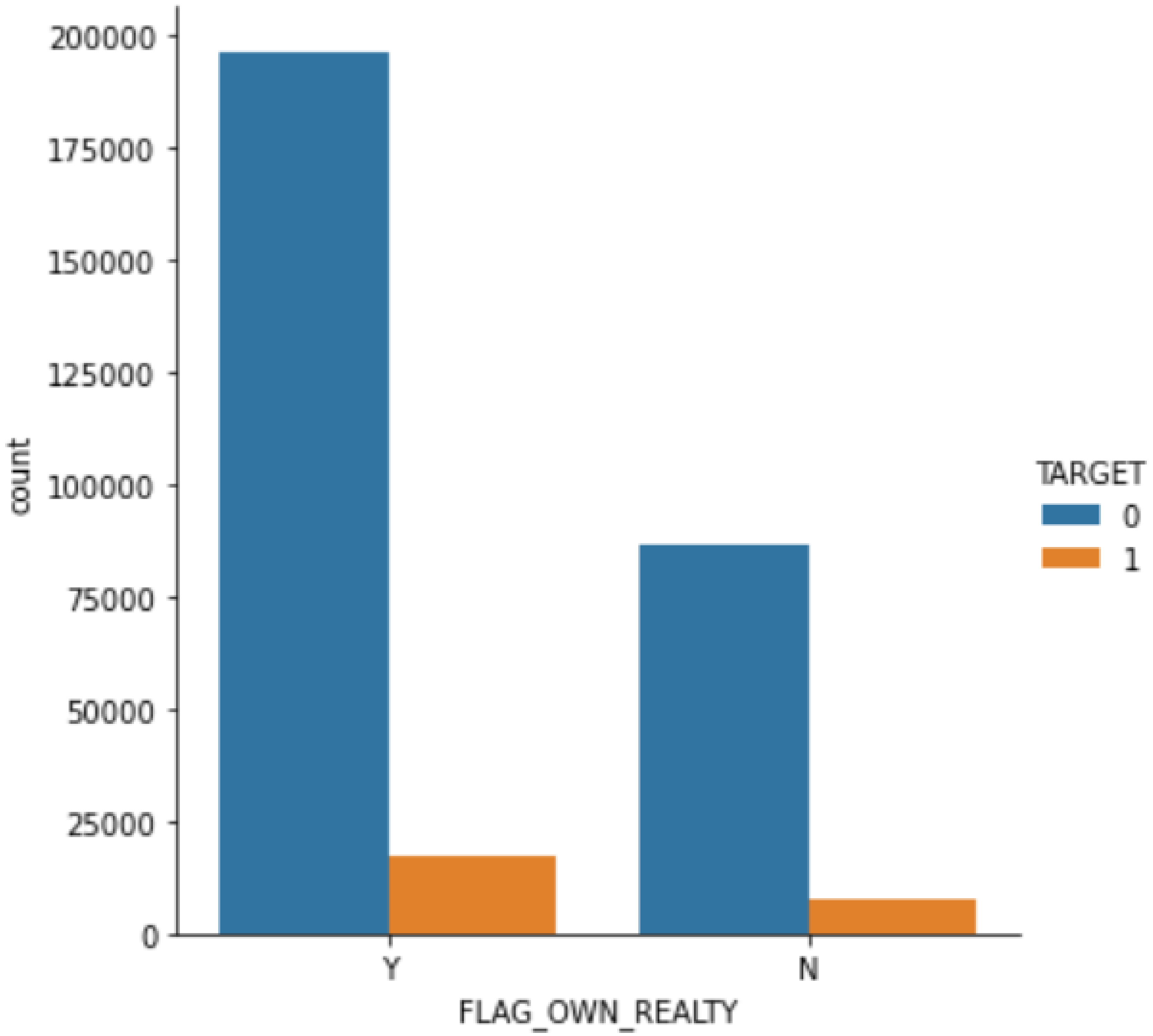
```
sns.catplot(data=df, x = 'FLAG_OWN_CAR',  hue='TARGET', kind="count" )
```

```
<seaborn.axisgrid.FacetGrid at 0x20c1eceeca0>
```

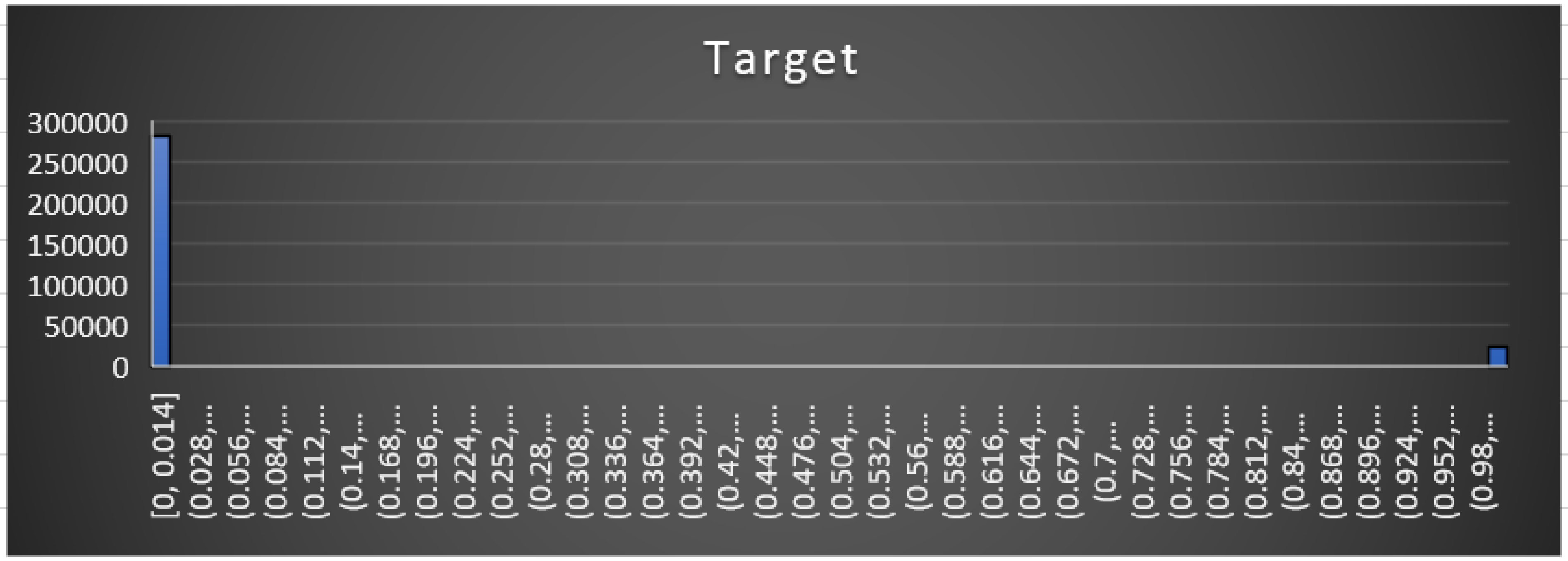


```
sns.catplot(data=df, x = 'FLAG_OWN_REALTY',  hue='TARGET', kind="count" )
```

```
<seaborn.axisgrid.FacetGrid at 0x20c1f82cc10>
```



As you can see the target variable itself is having huge imbalance



As you can see the target variable itself is having huge imbalance

clients with payment difficulties are so less in number to say, whereas the other category are very high in number comparitively

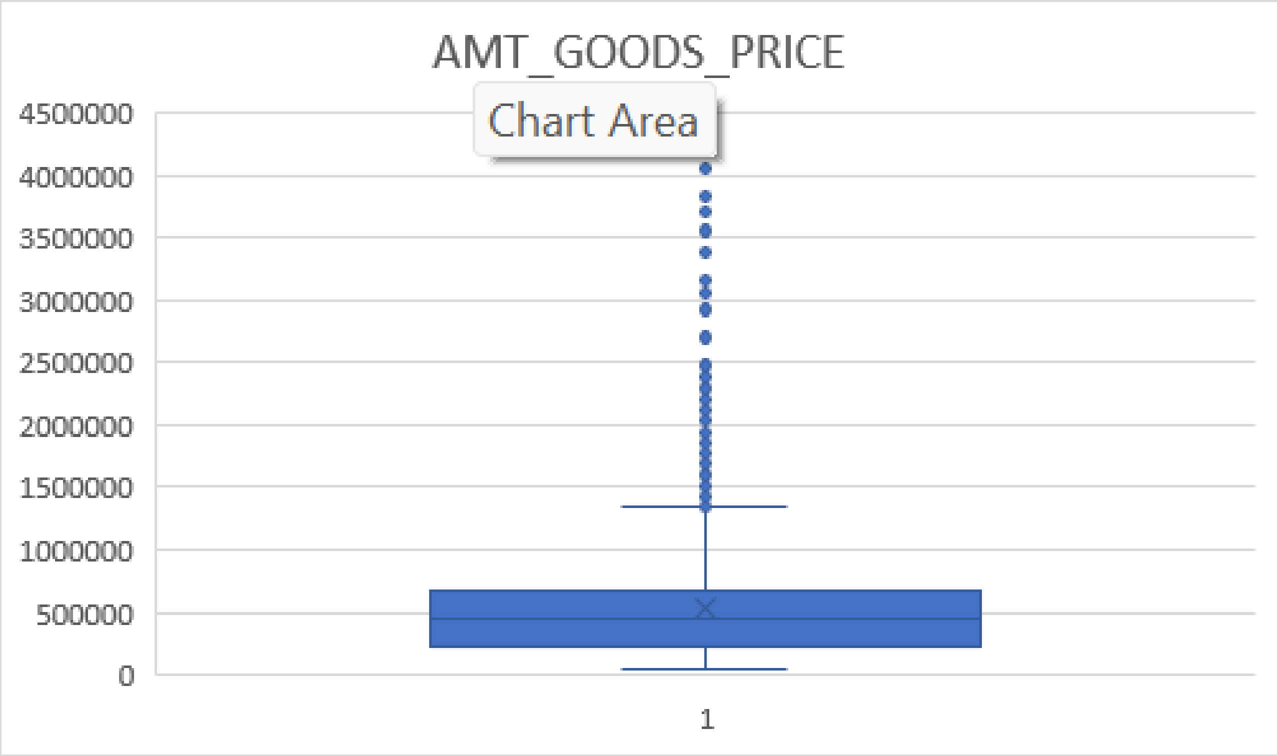
clients with payment difficulties are so less in number to say, whereas the other category are very high in number comparitively

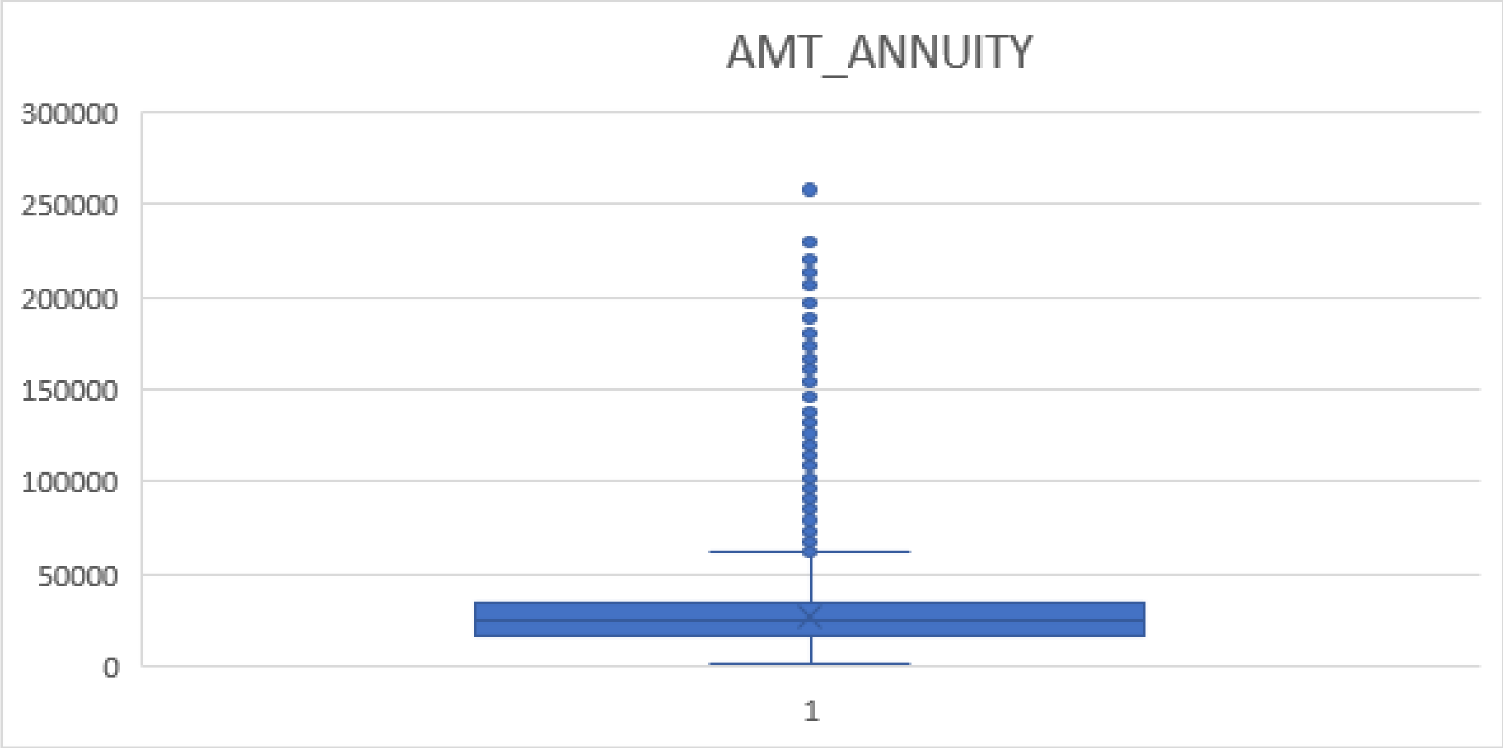
Reg_city_not_work_only_city and live_city_not_work_city columns are way too identical. Thus one of them are dropped

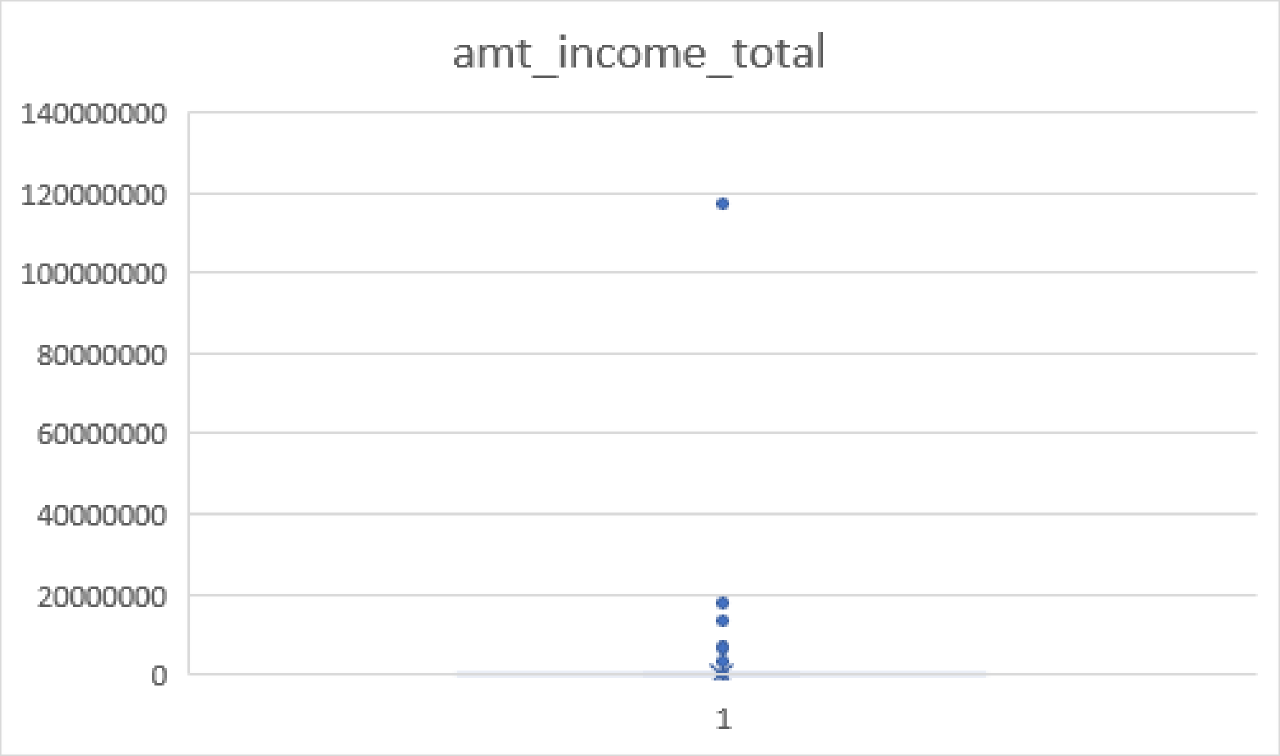
OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE columns are way too identical. Thus one of them are dropped

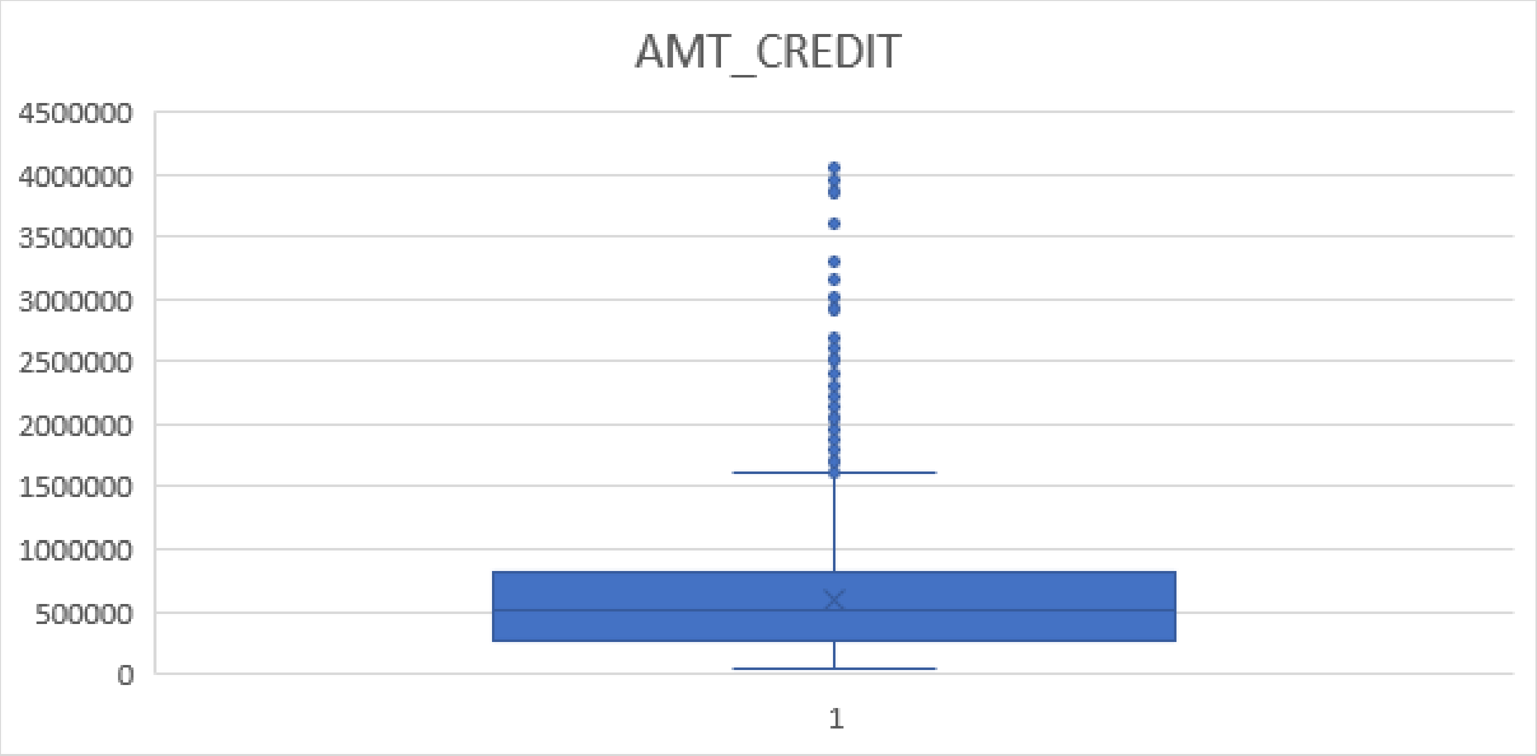
Also except for Document_3, no documents were provided by the applicant. So Document_1, Document_2 were dropped and even FLAG_DOCUMENT_3 is showing similar trend for both categories of target feature

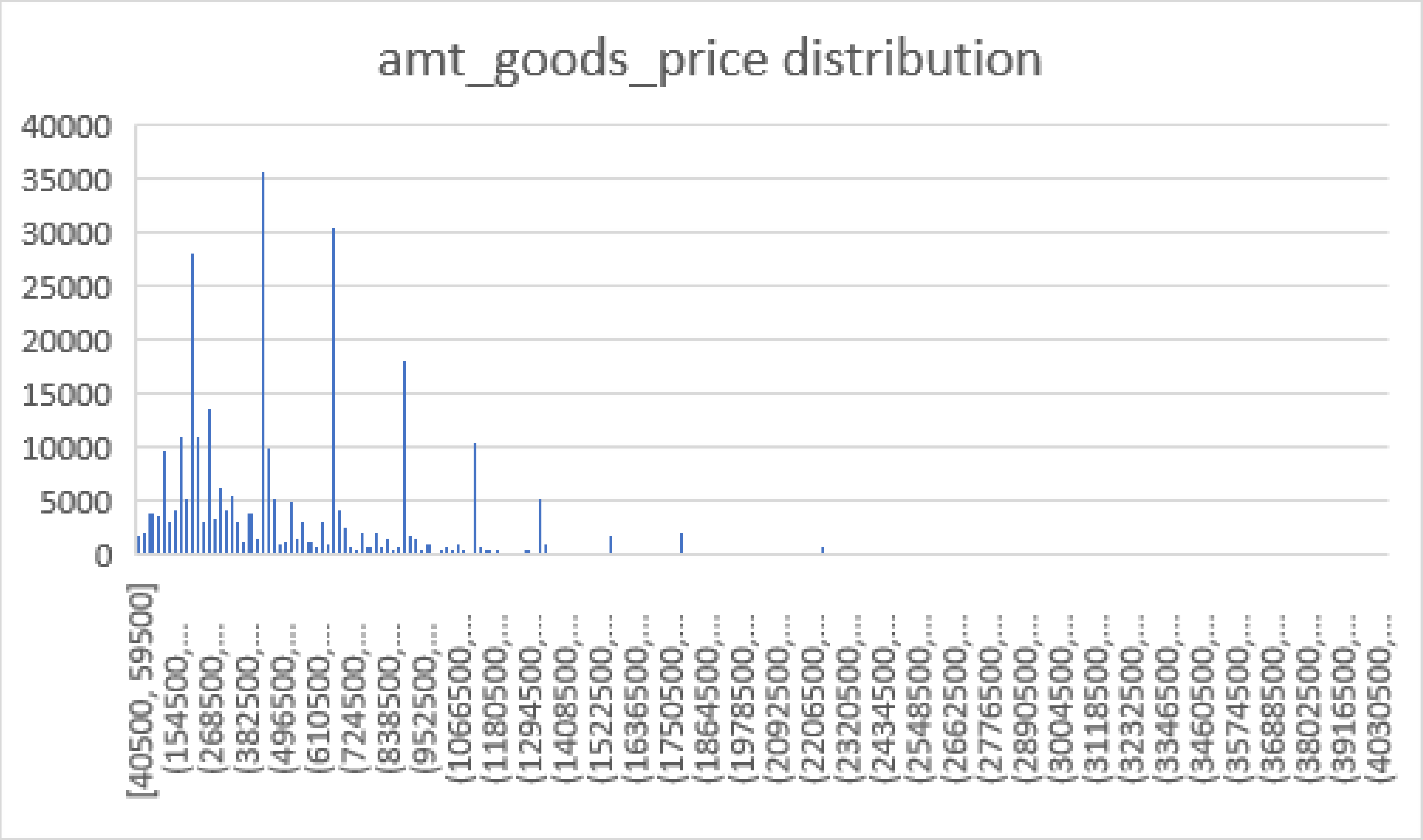
Outlier Analysis





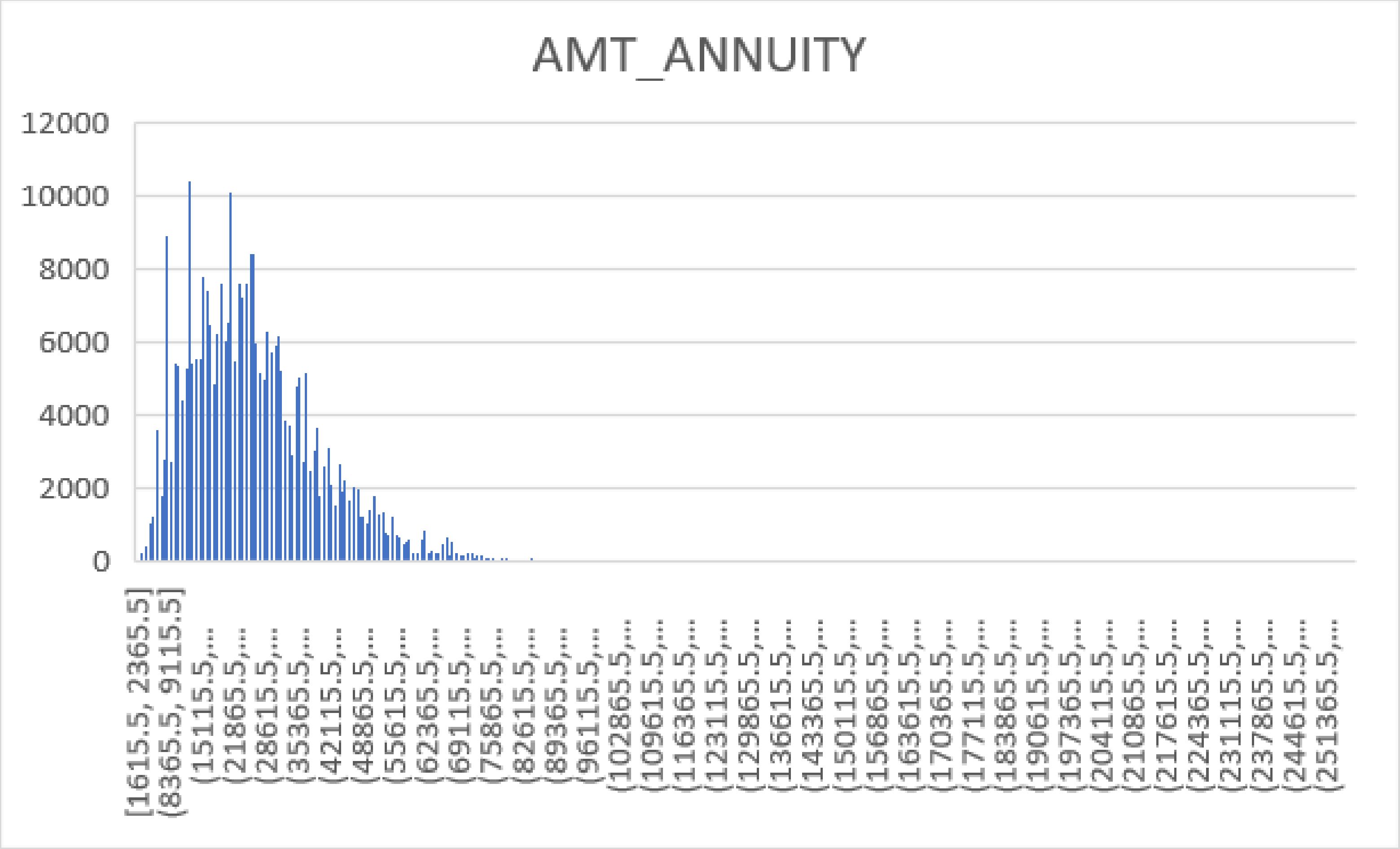




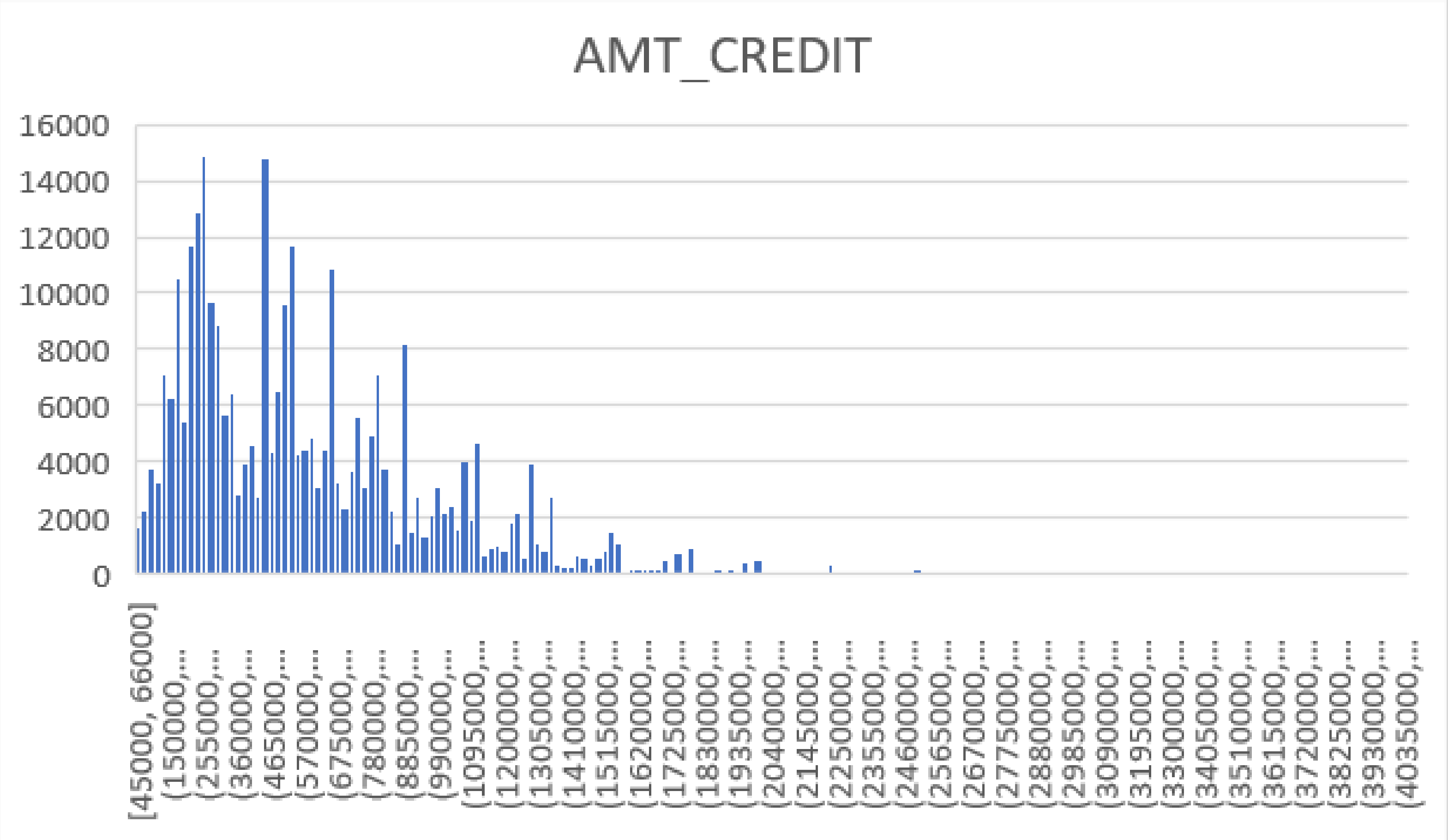


The data distribution is skewed(positively skewed), most of the data is at the minimum side, outliers on maximum side

This means, most of the goods price (for which the loan is given) are lesser in price,
there are only few goods whose price is larger, as we can also see from the boxplot

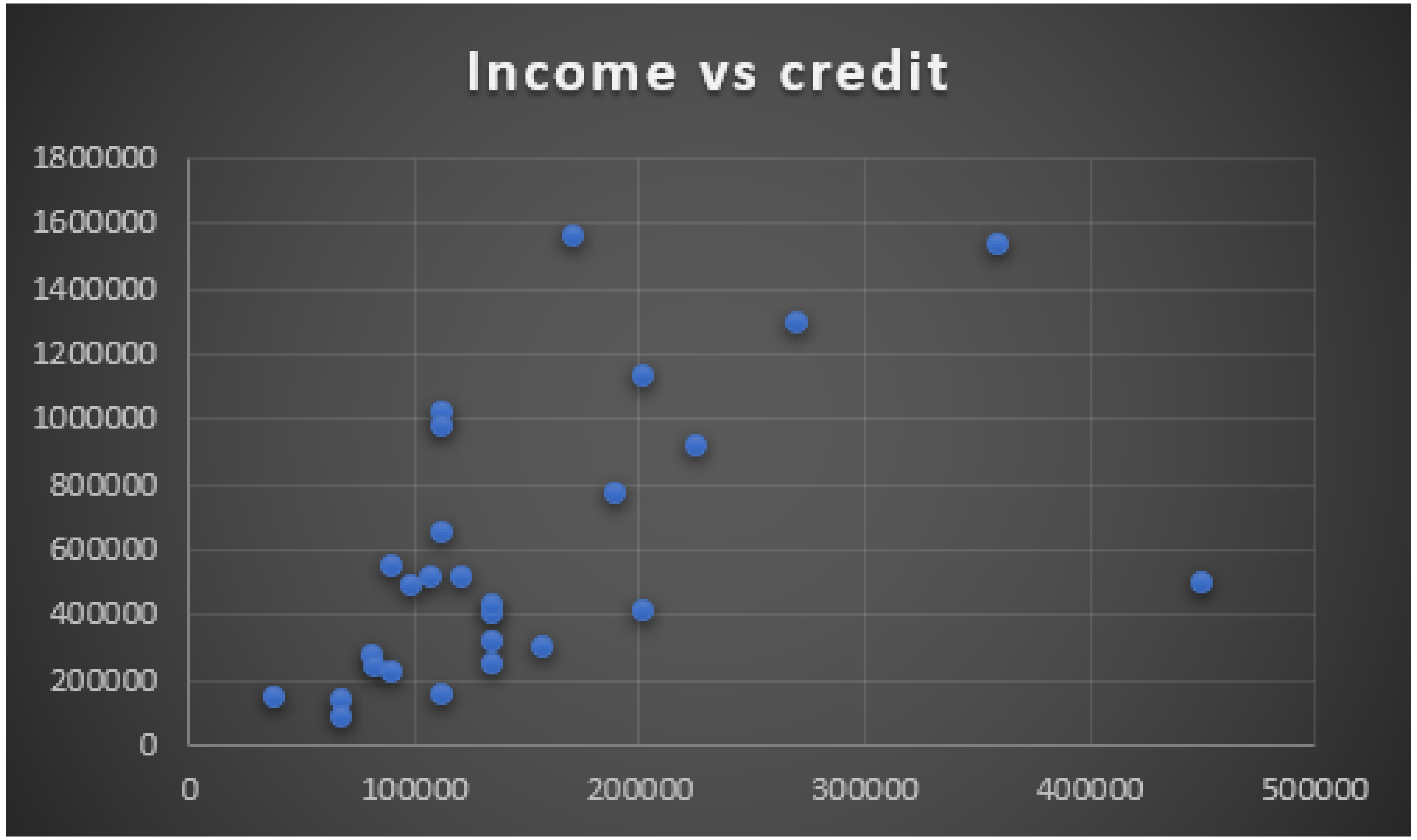


Annuity loan values are distributed well except for outliers on maximum side, these outliers tell us that some of the annuity loans are not as anticipated because, these annuity loans values are very higher comparitively to the most of the distribution.

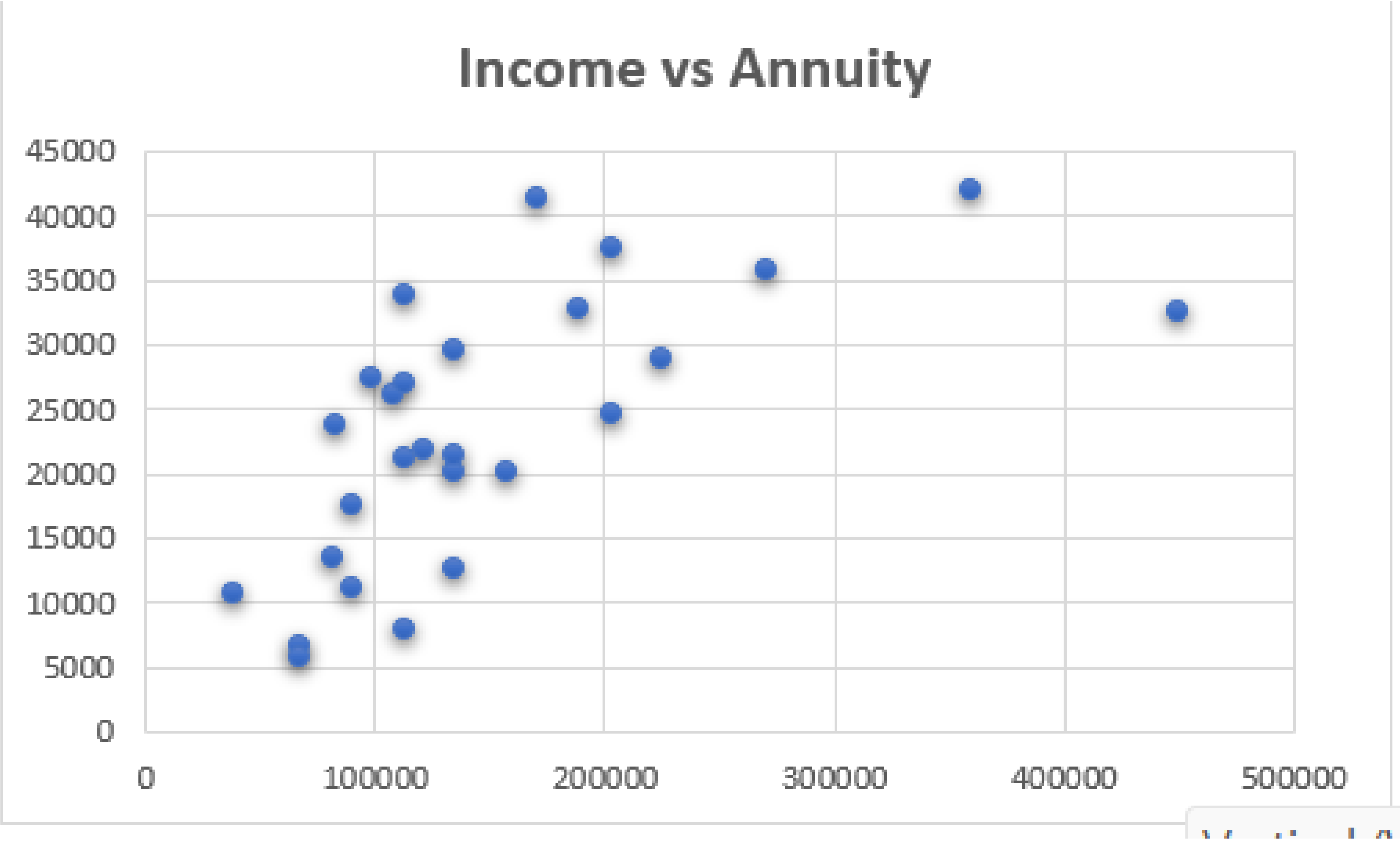


The outlying Credit amount of the loan values are not so often are the credit amount that to be issued to the applicant, they're not similar to the most of the credit amount of loan issued to the applicant.

Variate Analysis

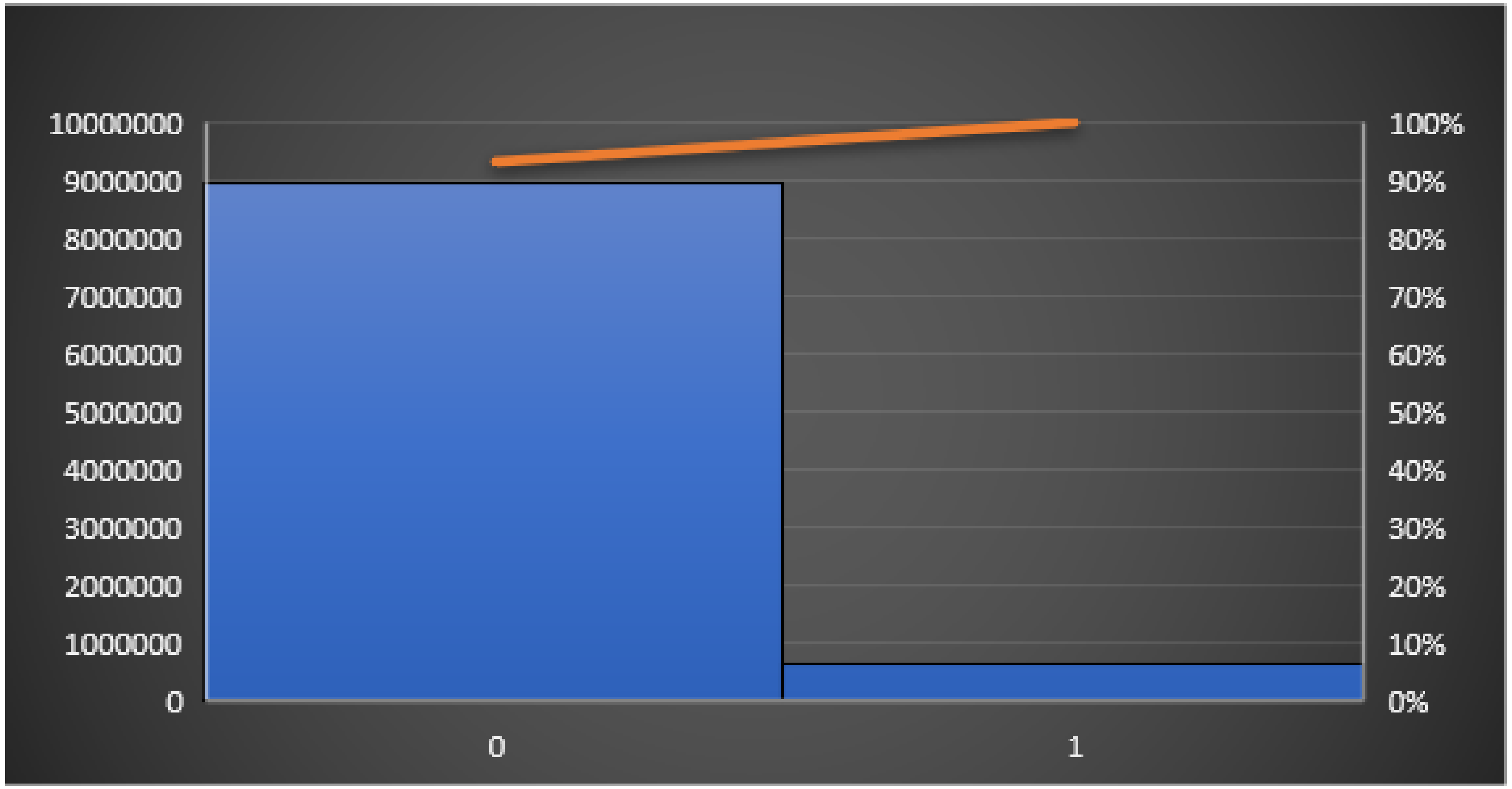


As the Income is increasing, the credit amount also seems to be increasing



Also for the Annuity loan, annuity seems increasing as income of the applicant increases.

As Income increasing, the goods price on which loan is given seems increasing

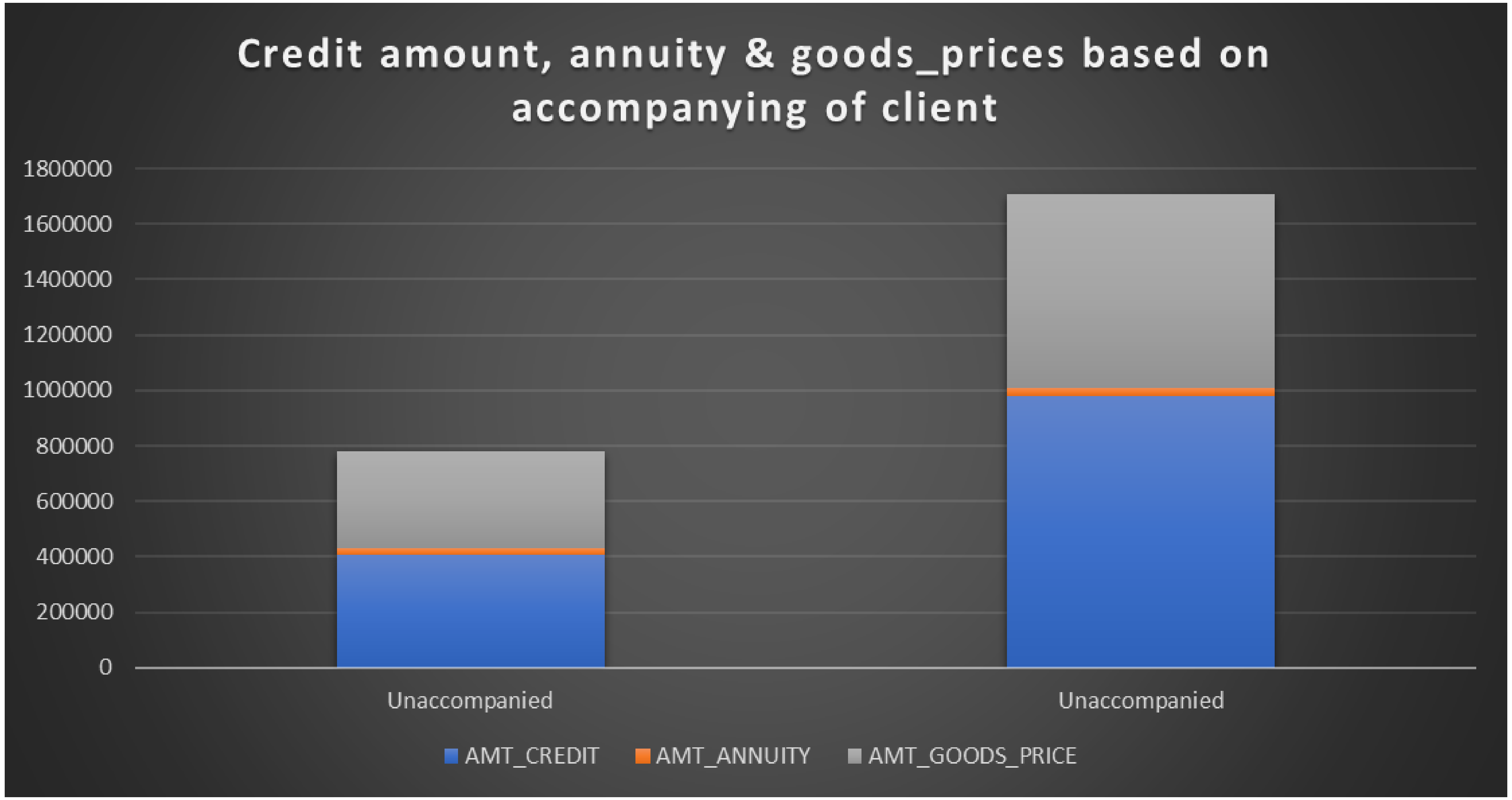


1

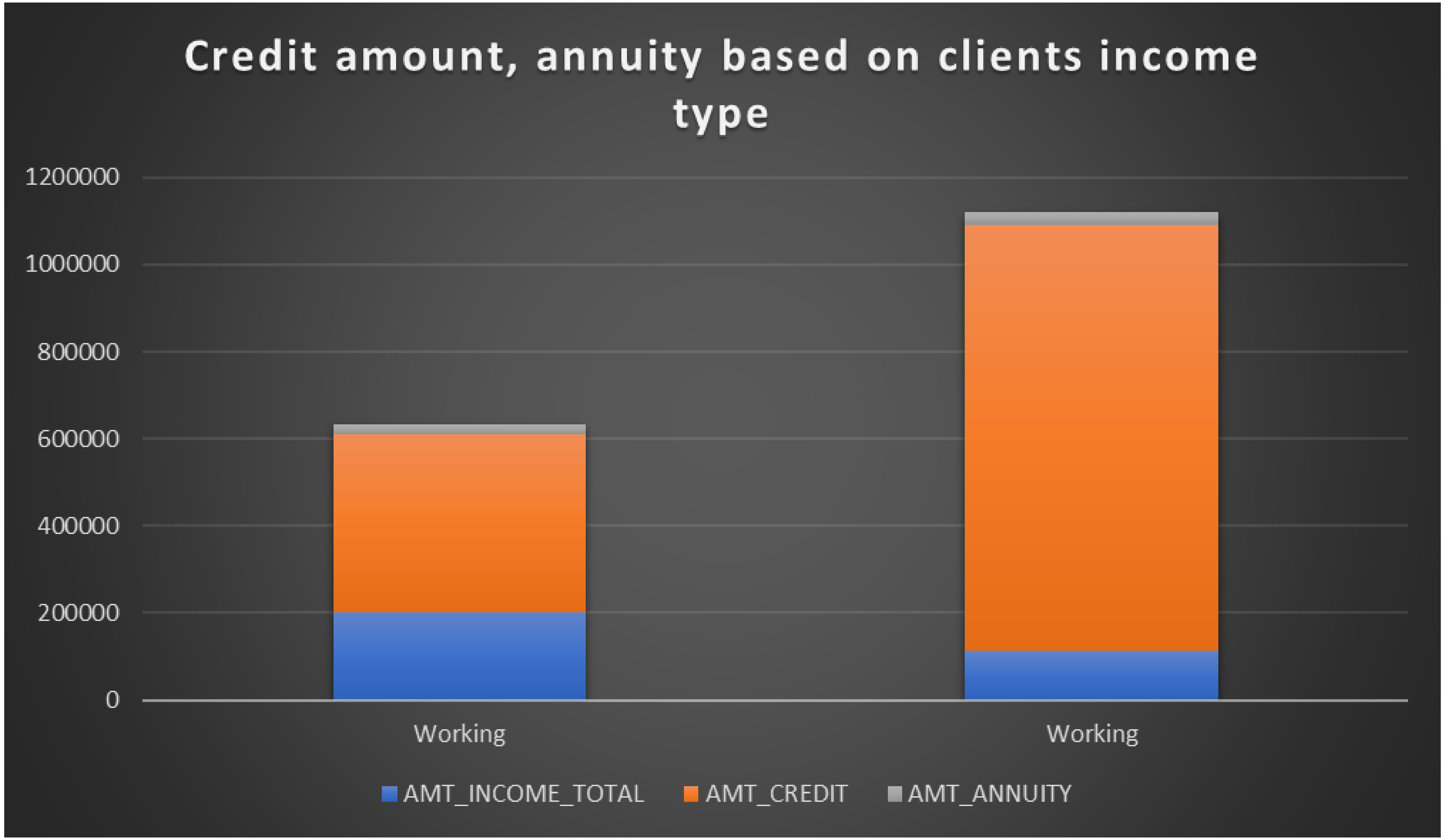
client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample

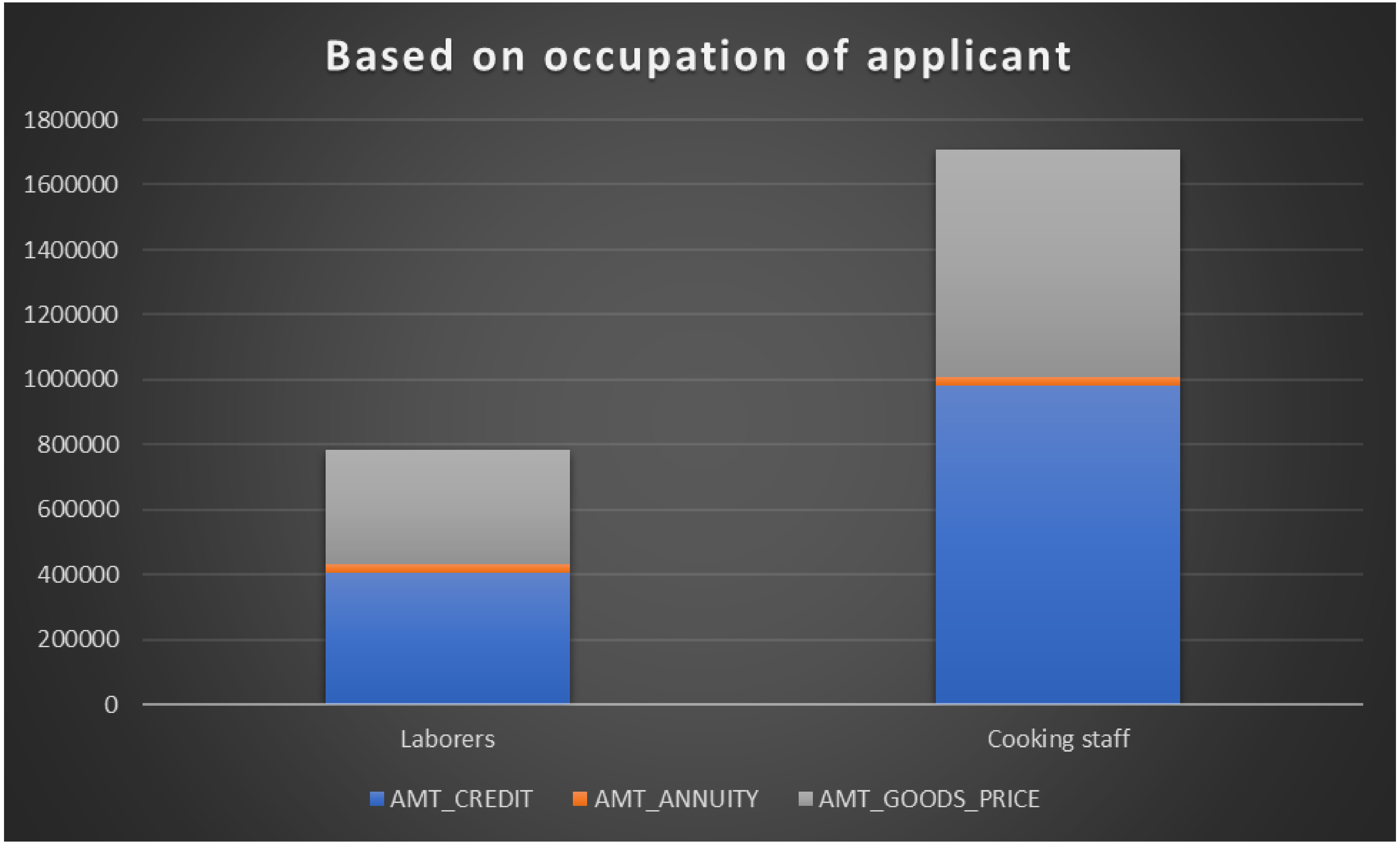
0

All other cases

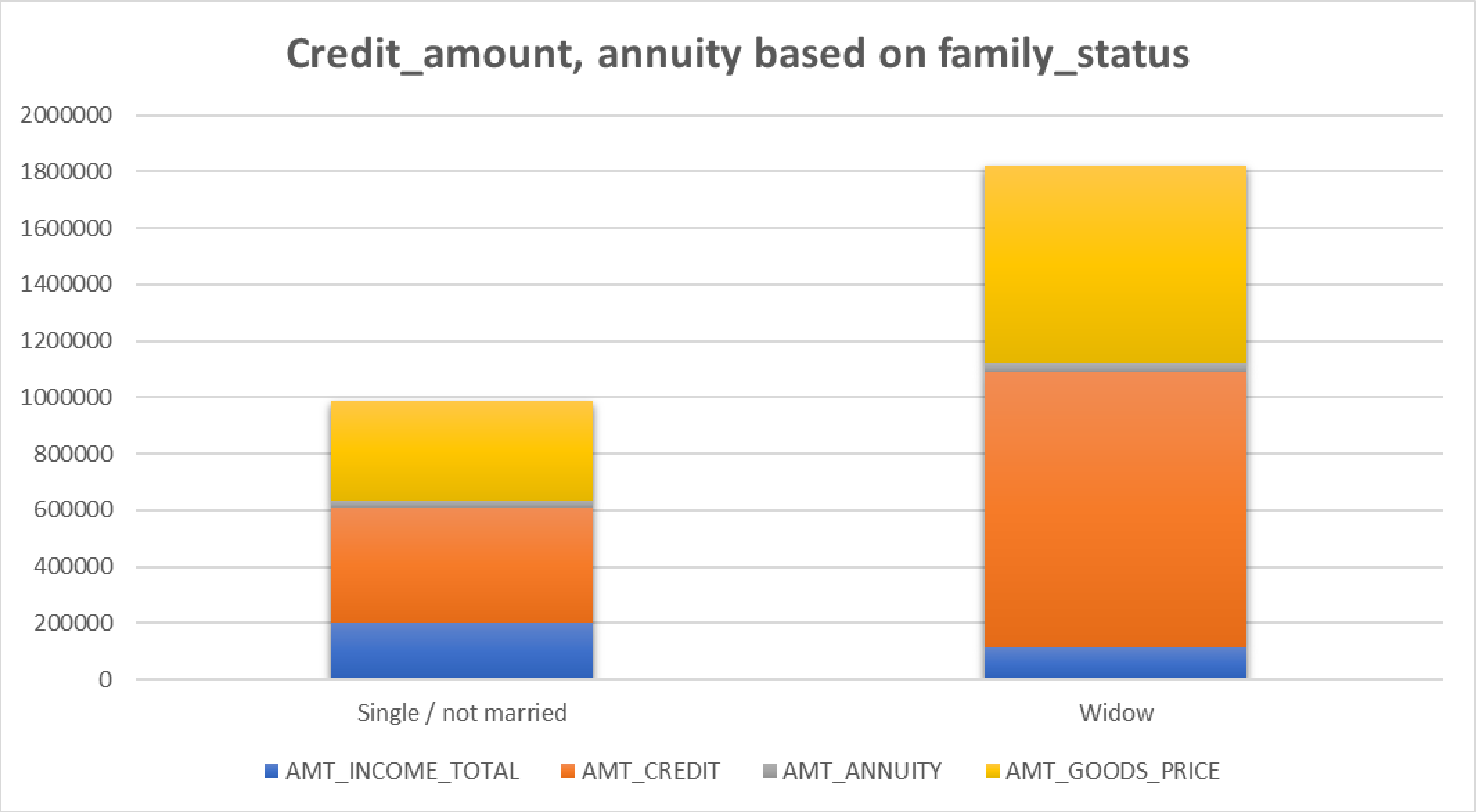


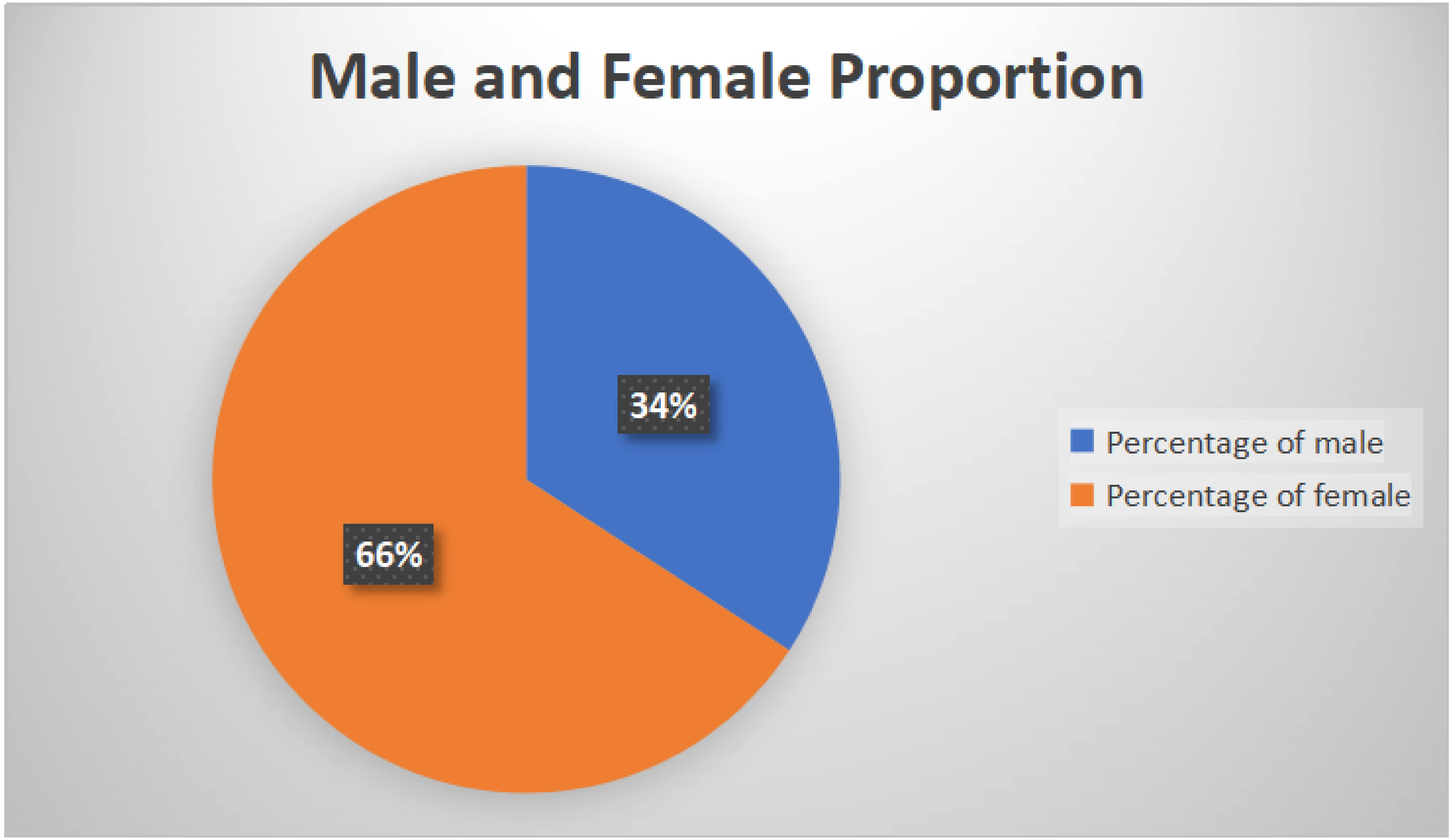
Loan annuity seems to be higher for the applicants who are unaccompanied


Credit amount, annuity based on clients income type
1200000
1000000
800000
600000
400000
200000
0
Working
Working
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY

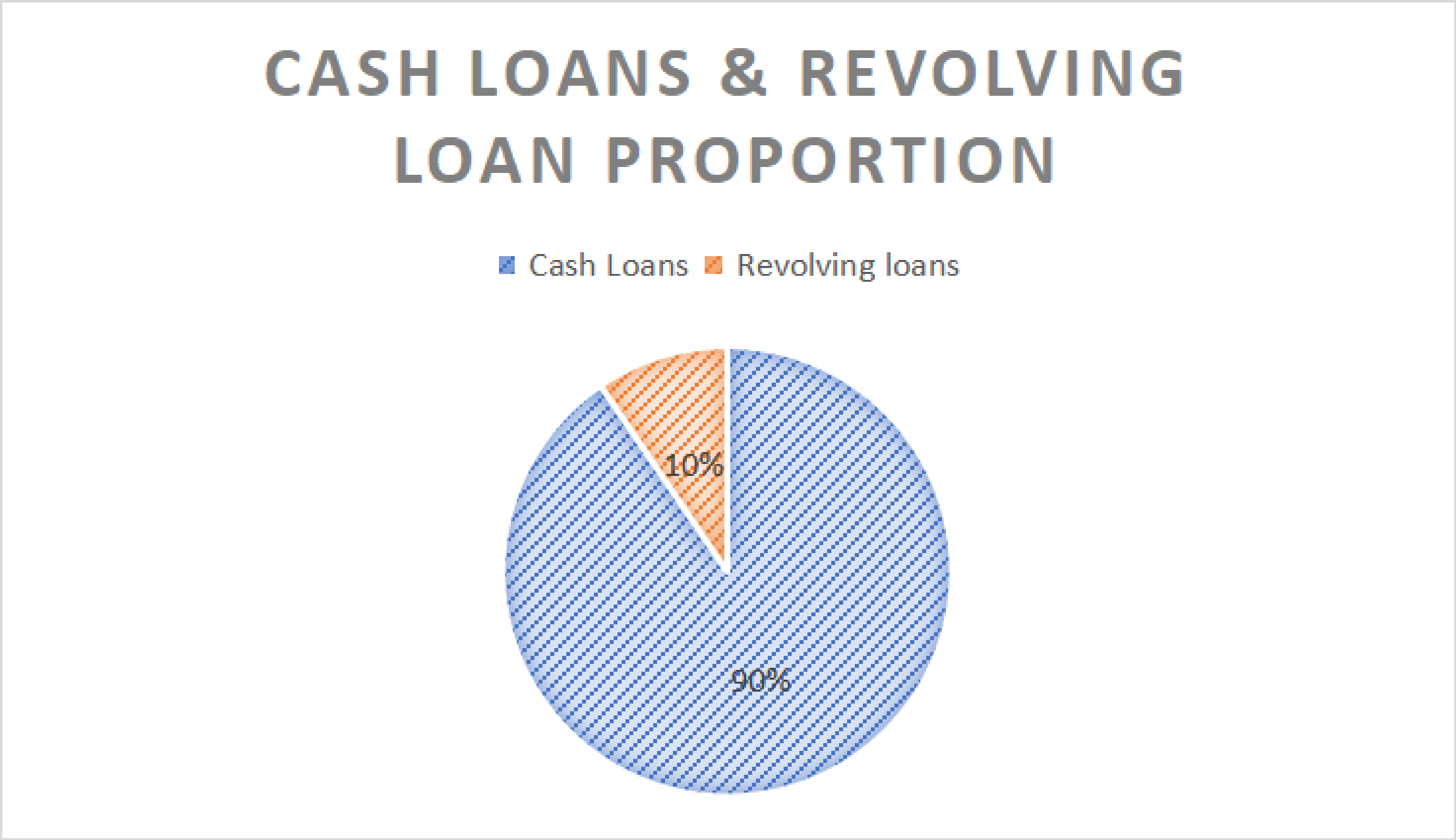


Loan annuity based on occupation of the loan applicant

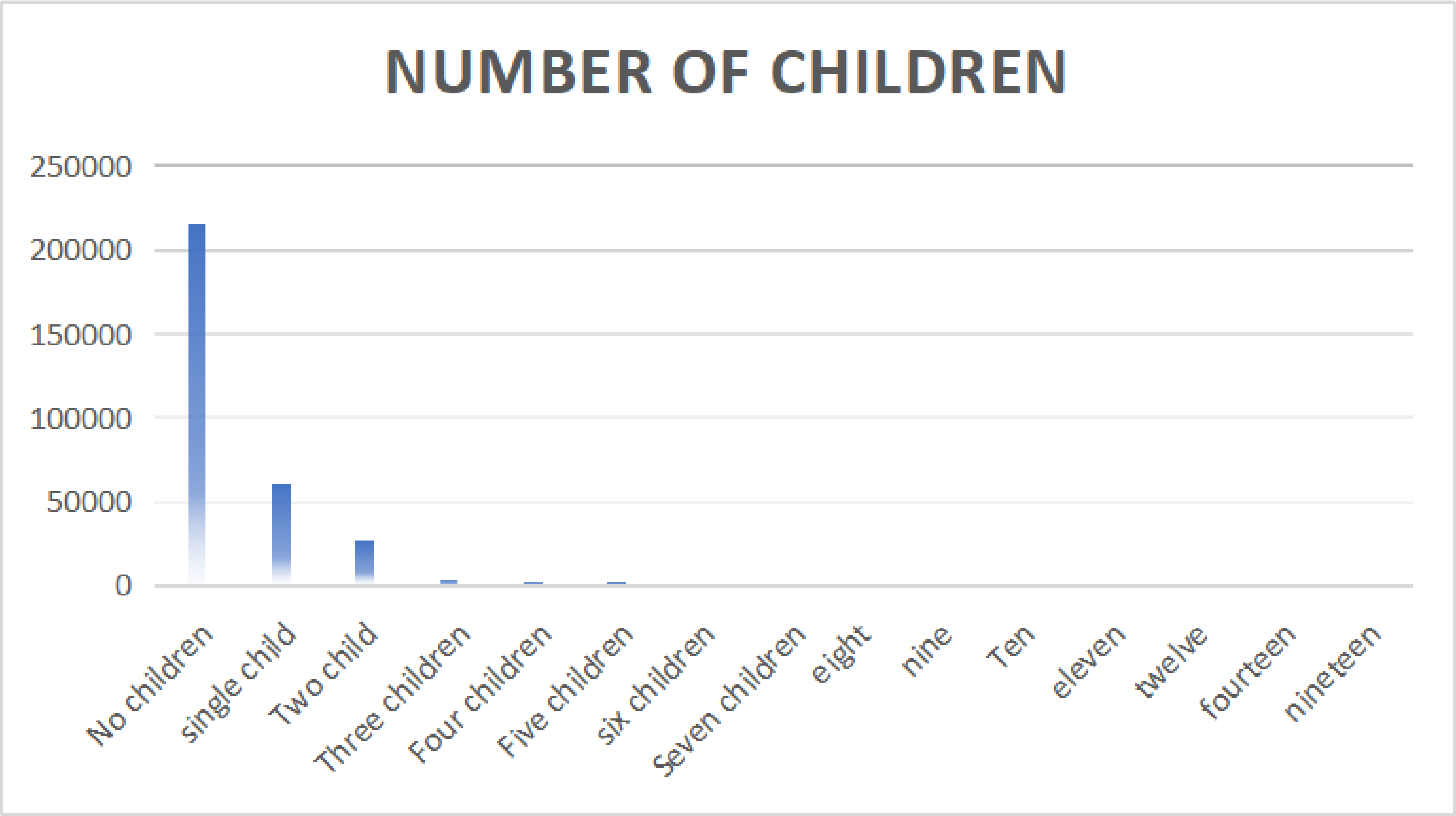

Credit_amount, annuity based on family_status
2000000
1800000
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
Single / not married
Widow
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE



percentage of female applicants are more

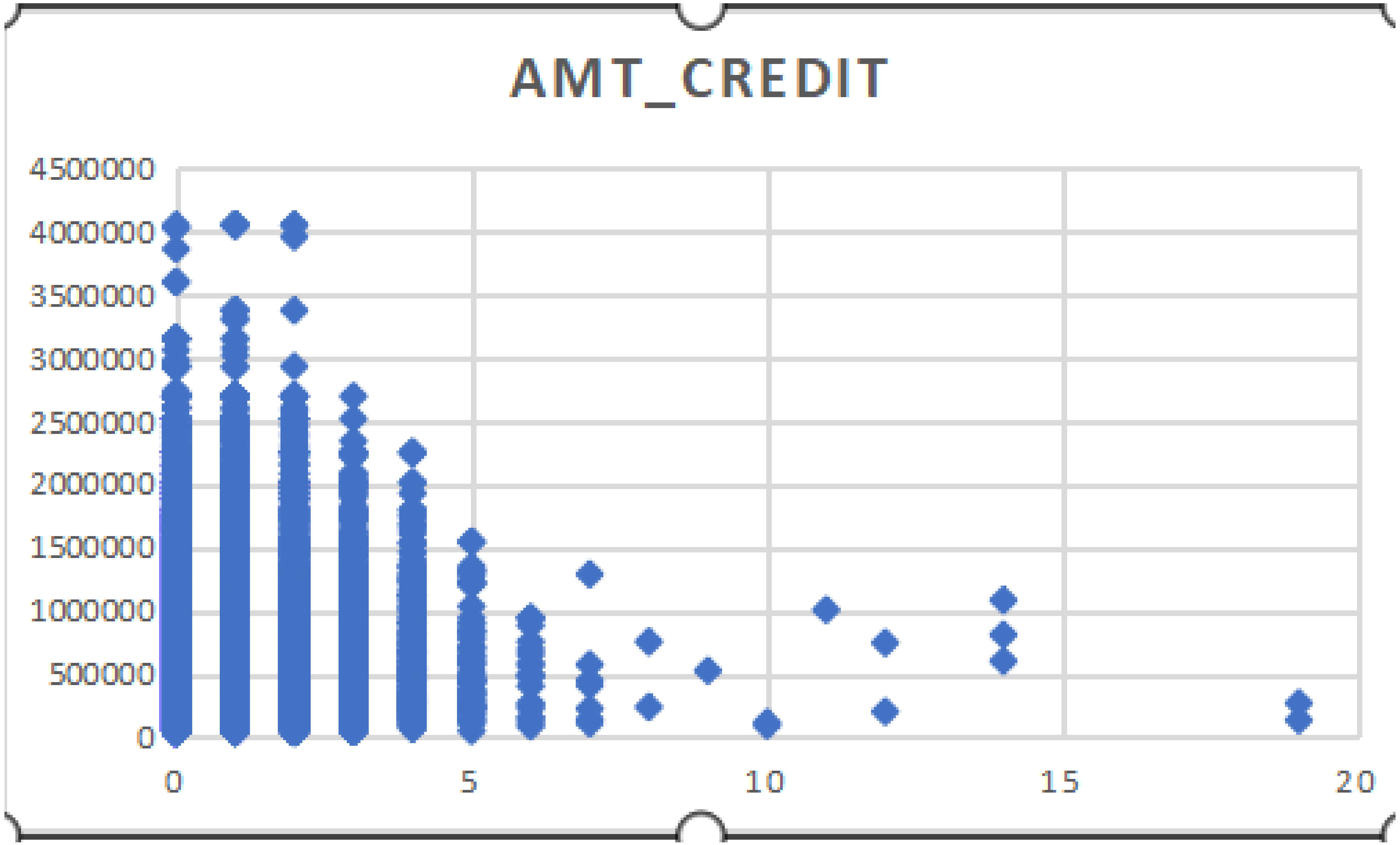


90% of the loans are of 'cash loan' type



There are more applicants who mostly have no children

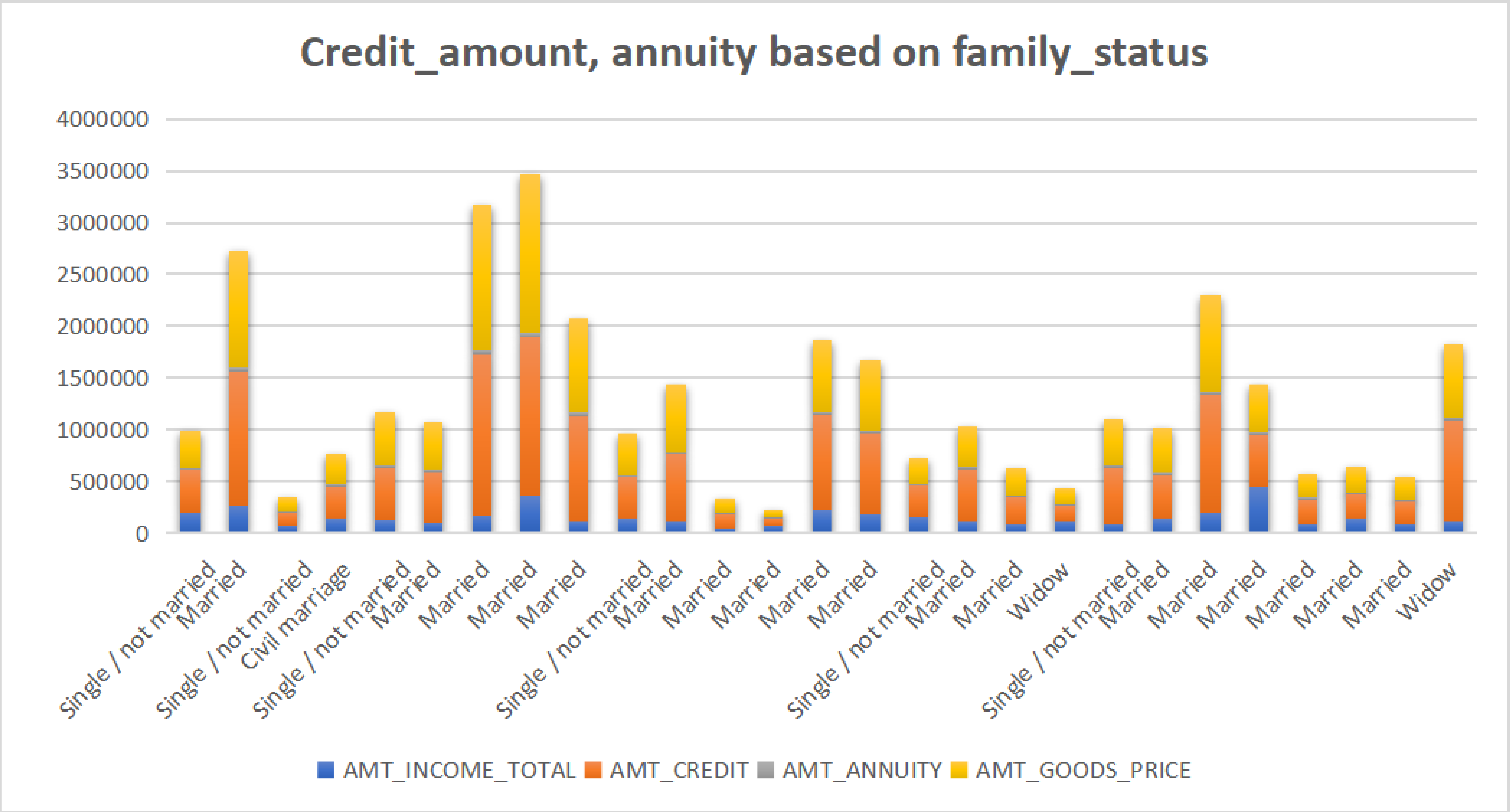
As we can see there's a huge imbalance between the categories of the target feature. The ratio of data imbalance is 8:92



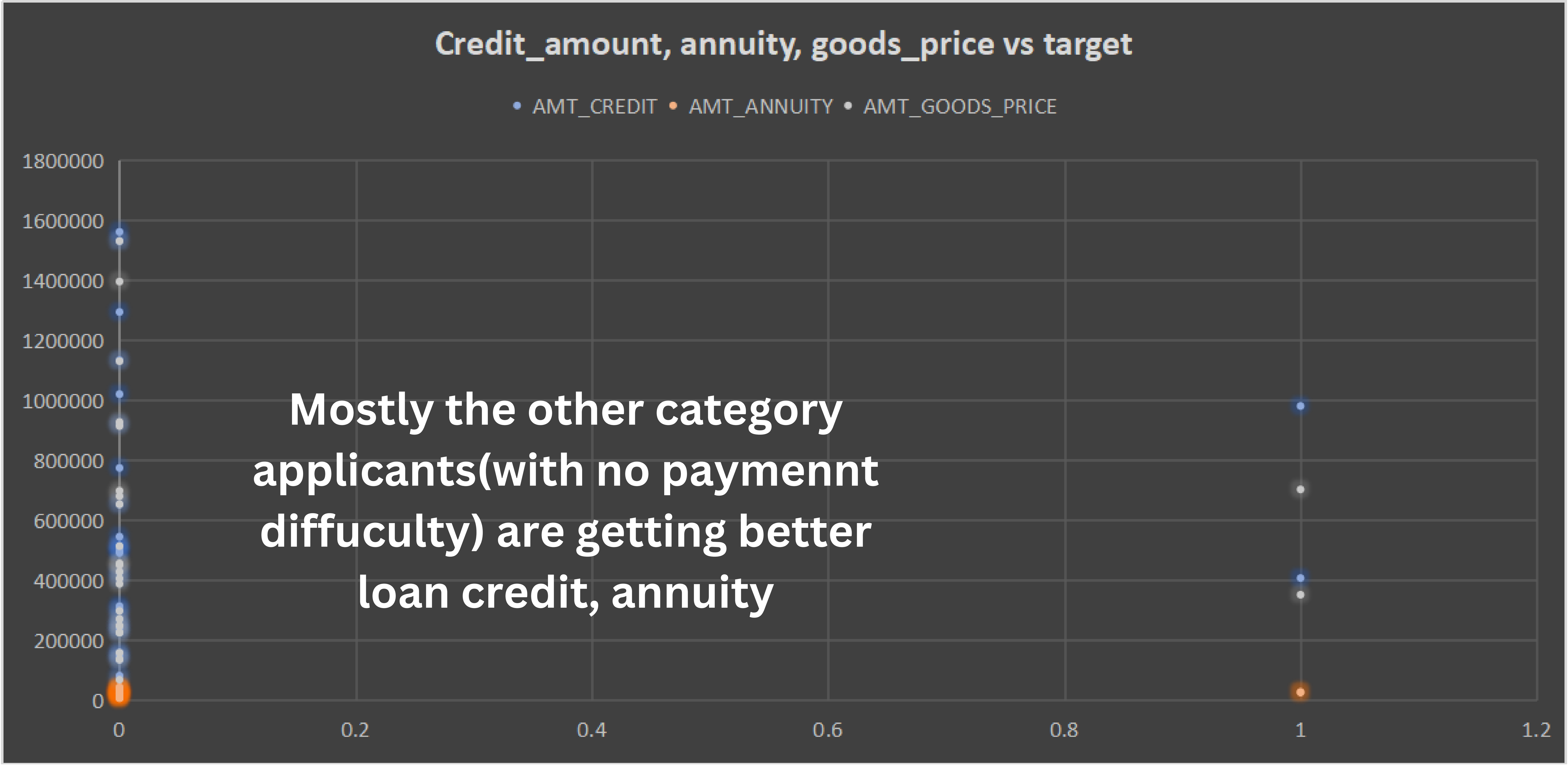
Credit amount based on the number of children

As we can observe that, the credit amount is high when the number of children are less

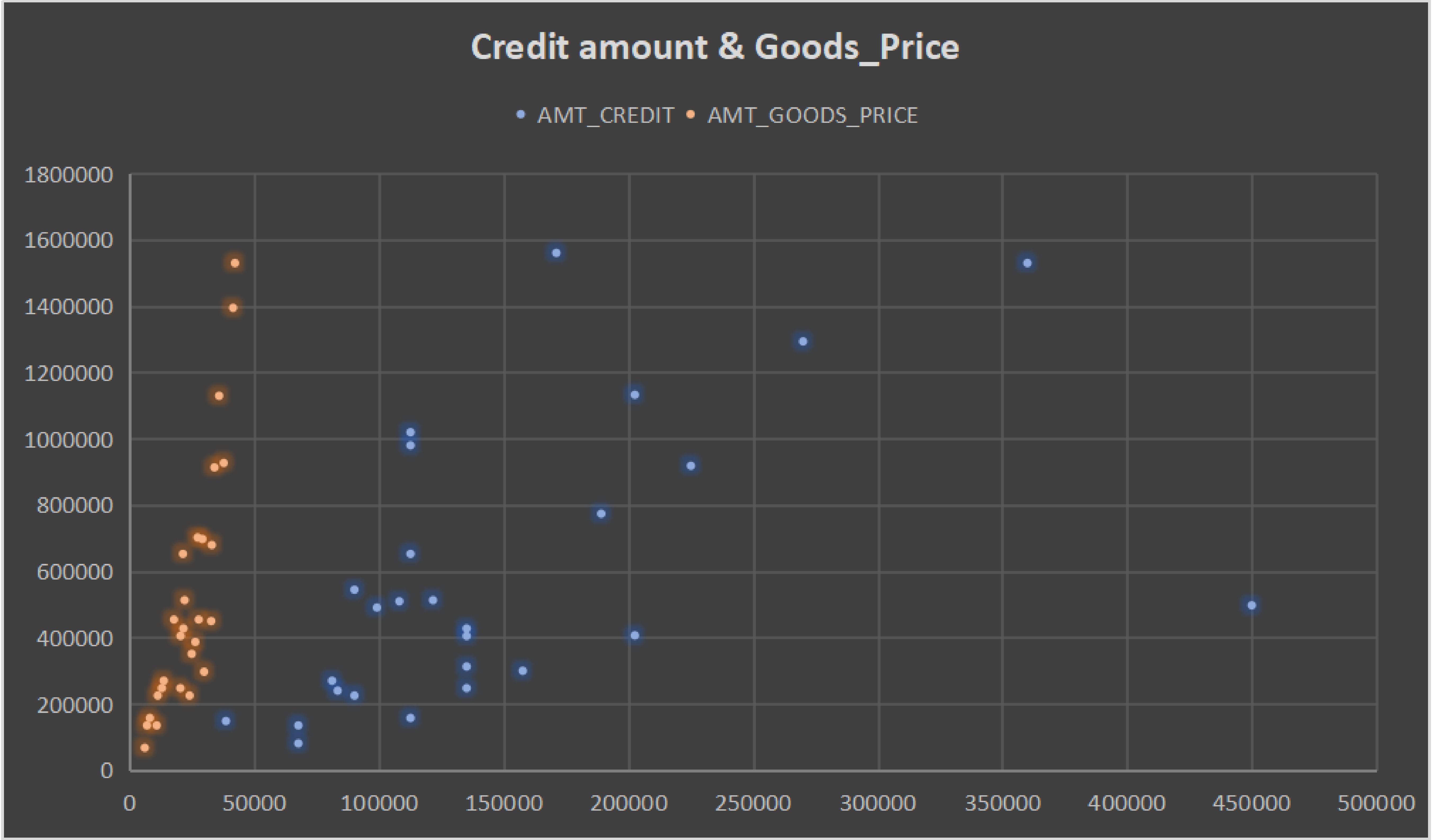
It is easy to observe because the count of children on the horizontal axis is discrete

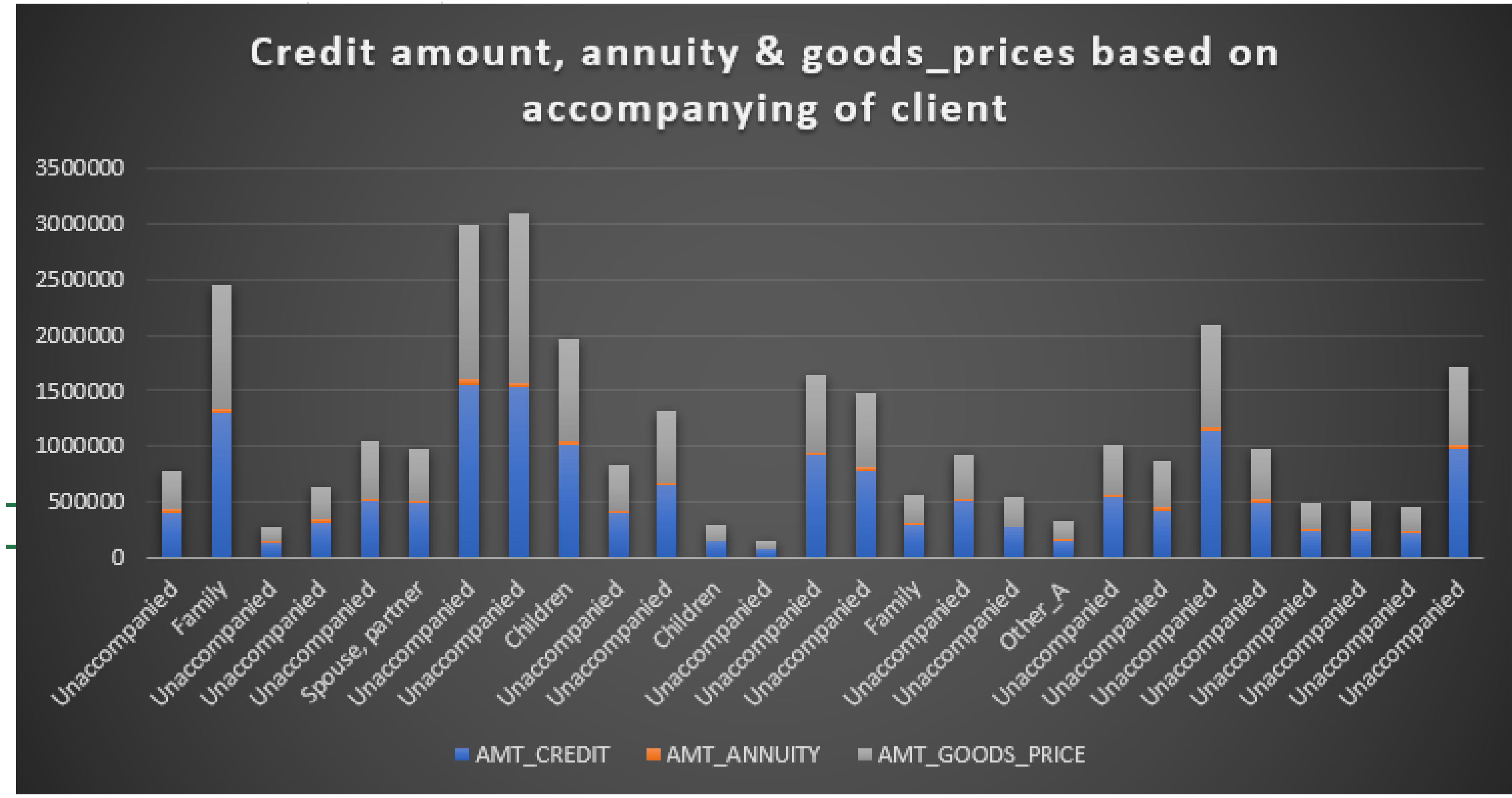


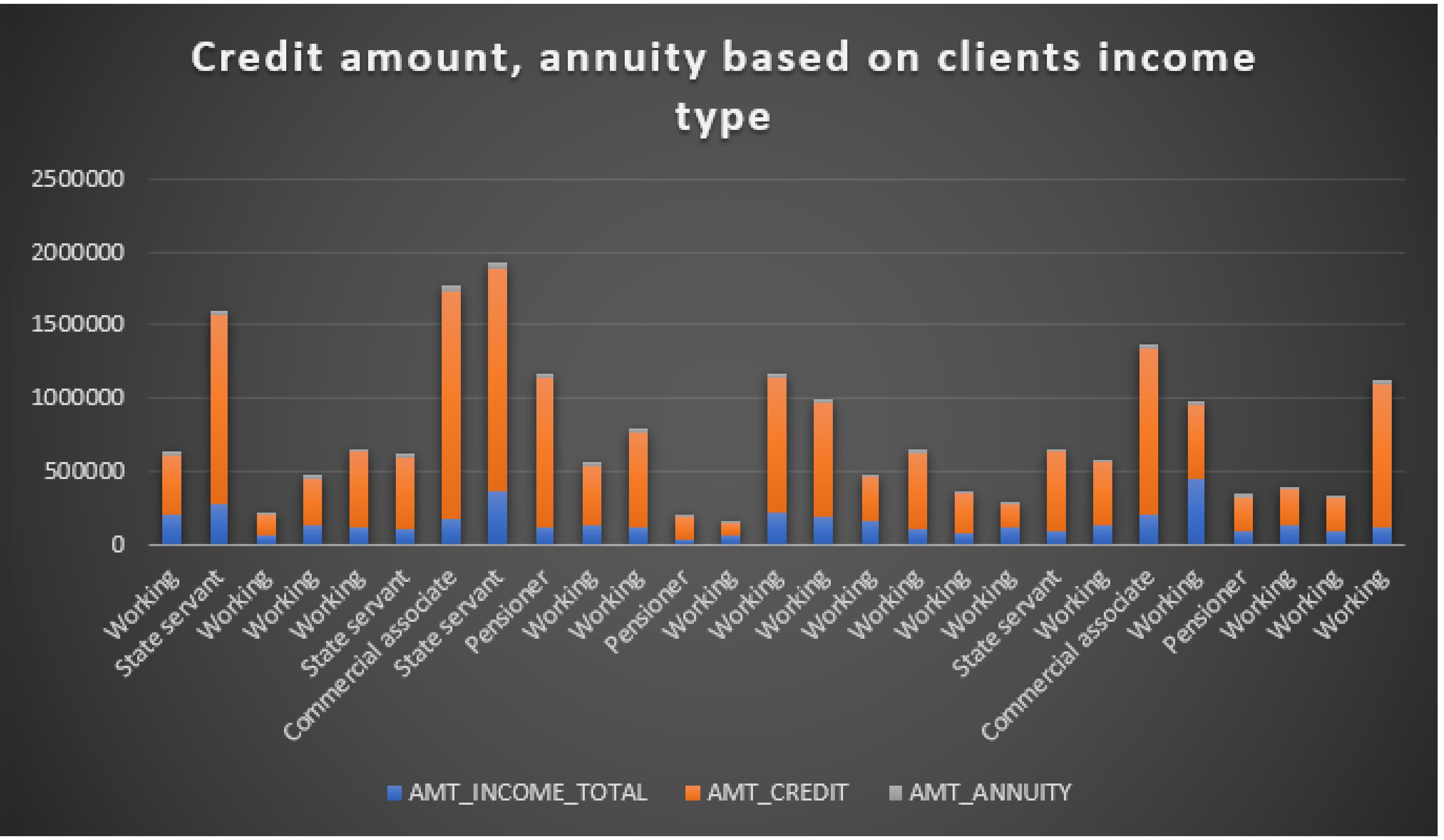
Mostly applicants with marital status married are getting better loan credit amount, and annuity loans



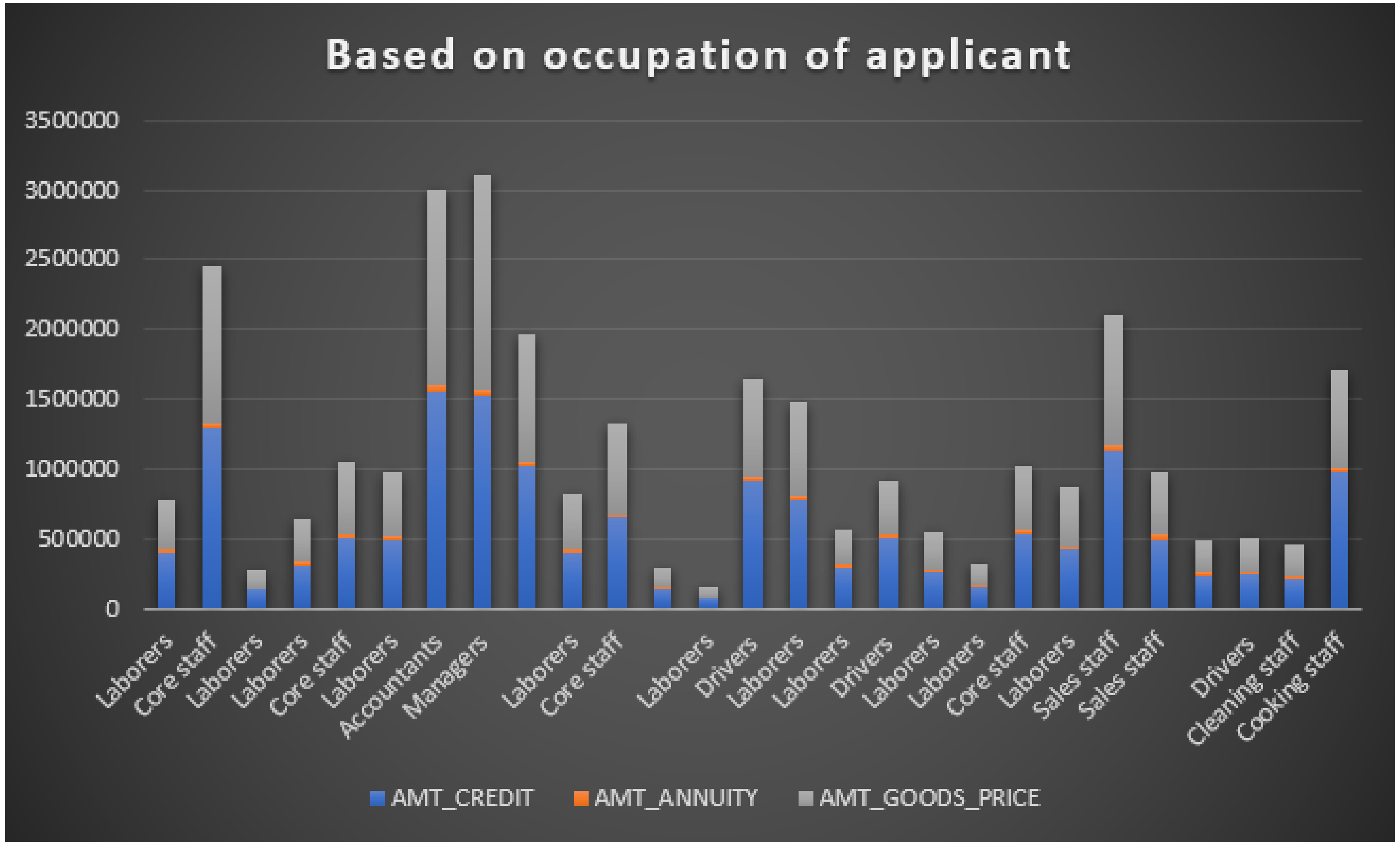
Mostly the other category applicants(with no paymennt diffuculty) are getting better loan credit, annuity


Credit amount & Goods_Price
AMT_CREDIT
AMT_GOODS_PRICE
1800000
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
0
50000
100000
150000
200000
250000
300000
350000
400000
450000
500000





Credit_amount , loan annuity based on accompany of client

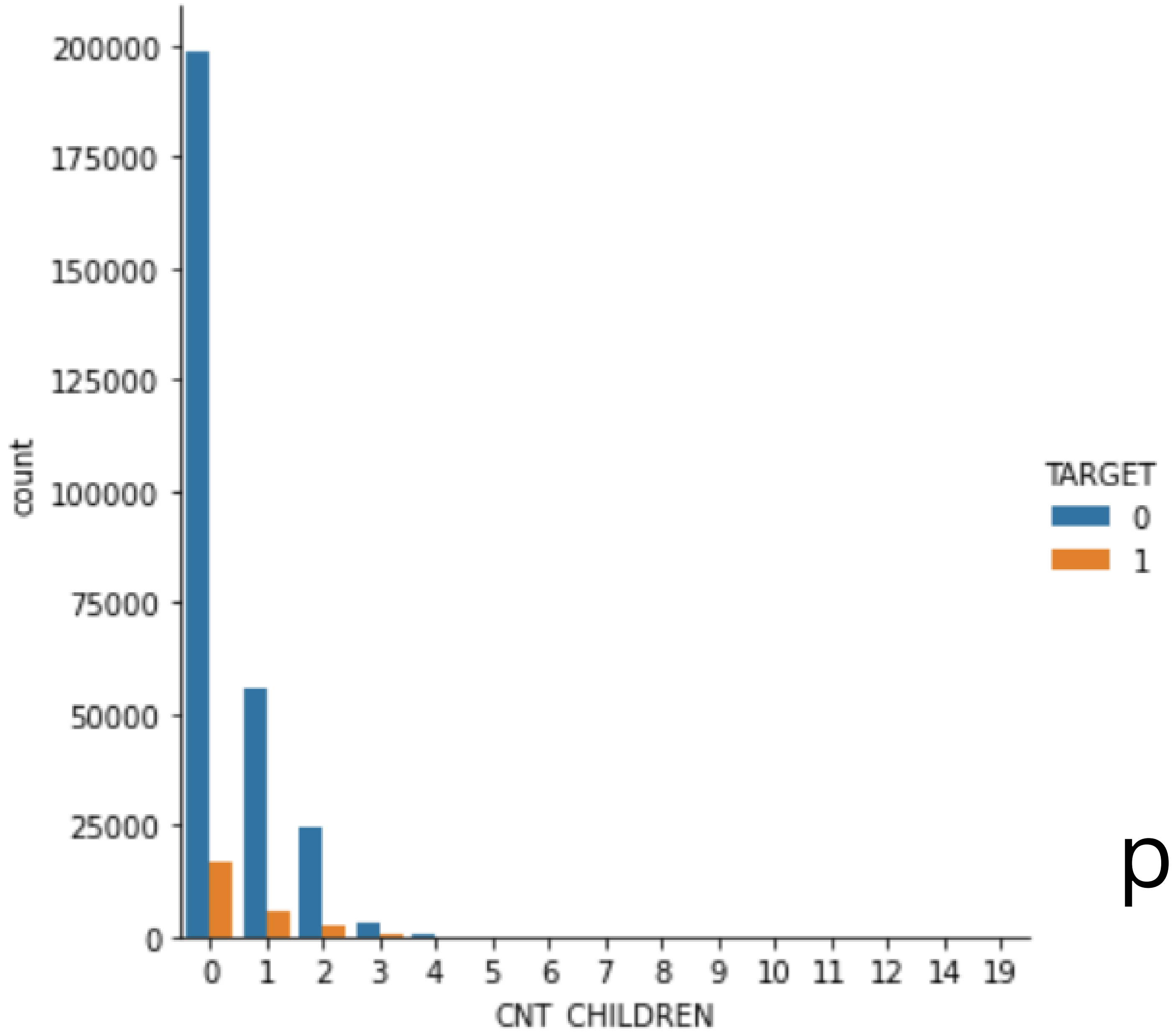


credit amount of loan, annuity based on occupation of client.

mostly the staff based employement(sales staff, core staff, cooking staff) are getting better annuity loan, credit amount comparitive to Laborers, drivers

```
sns.catplot(data = df, x = "CNT_CHILDREN", hue="TARGET", kind = "count")
```

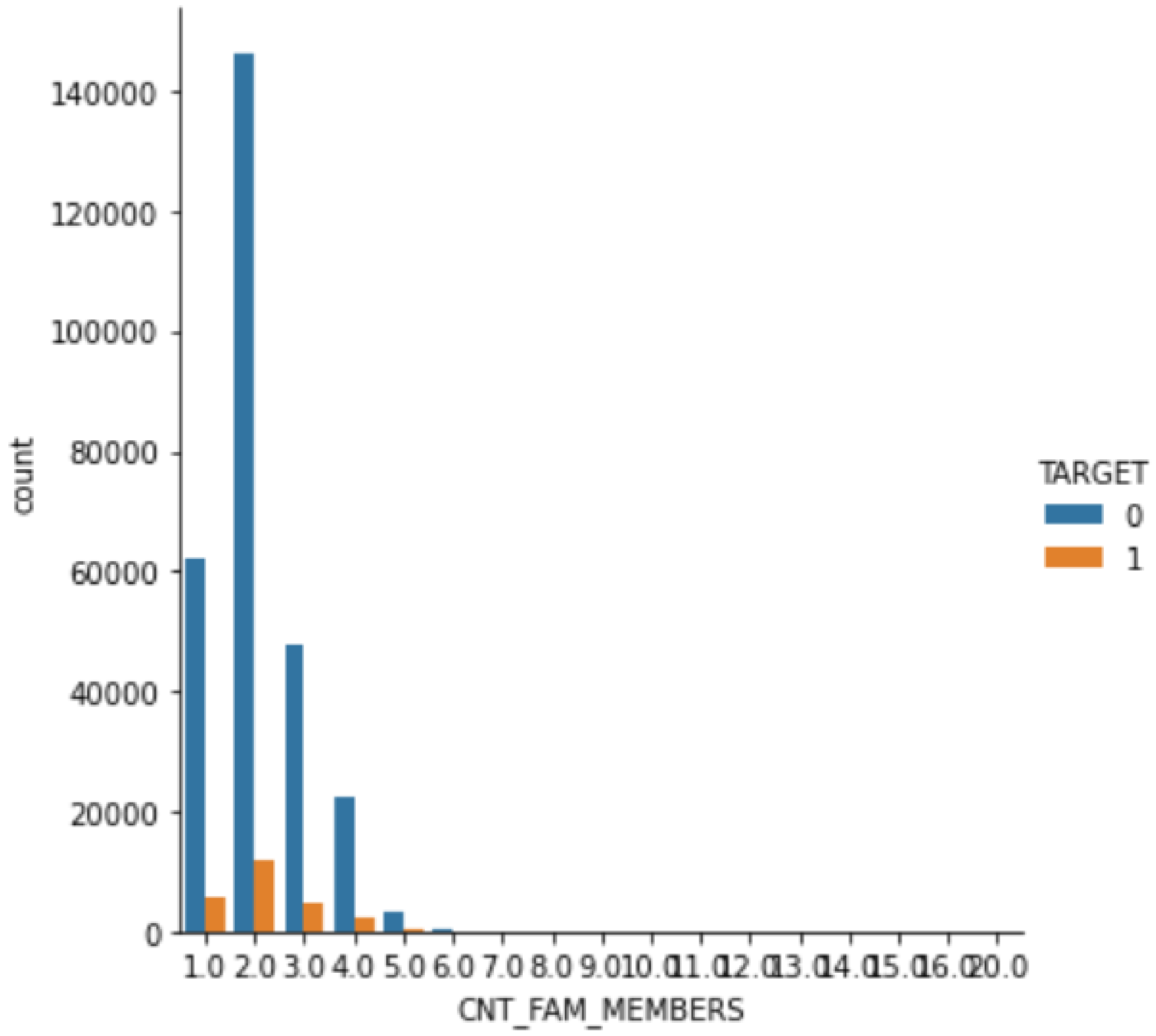
```
<seaborn.axisgrid.FacetGrid at 0x20c09514eb0>
```



Target(category-1 and category-2 people) based on number of children

```
sns.catplot(data = df, x = 'CNT_FAM_MEMBERS', hue = "TARGET", kind = "count")
```

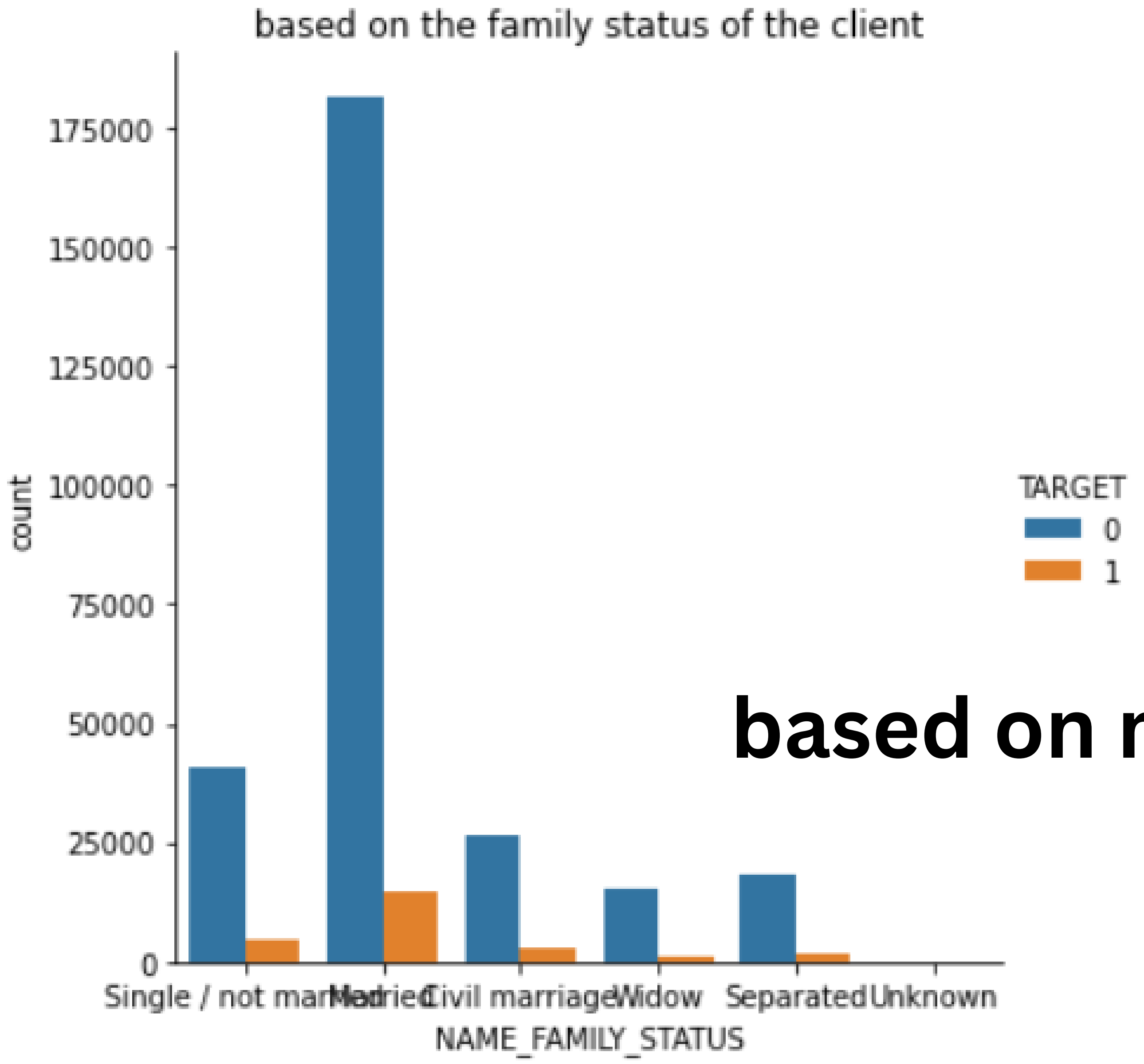
```
<seaborn.axisgrid.FacetGrid at 0x20c1eb49670>
```



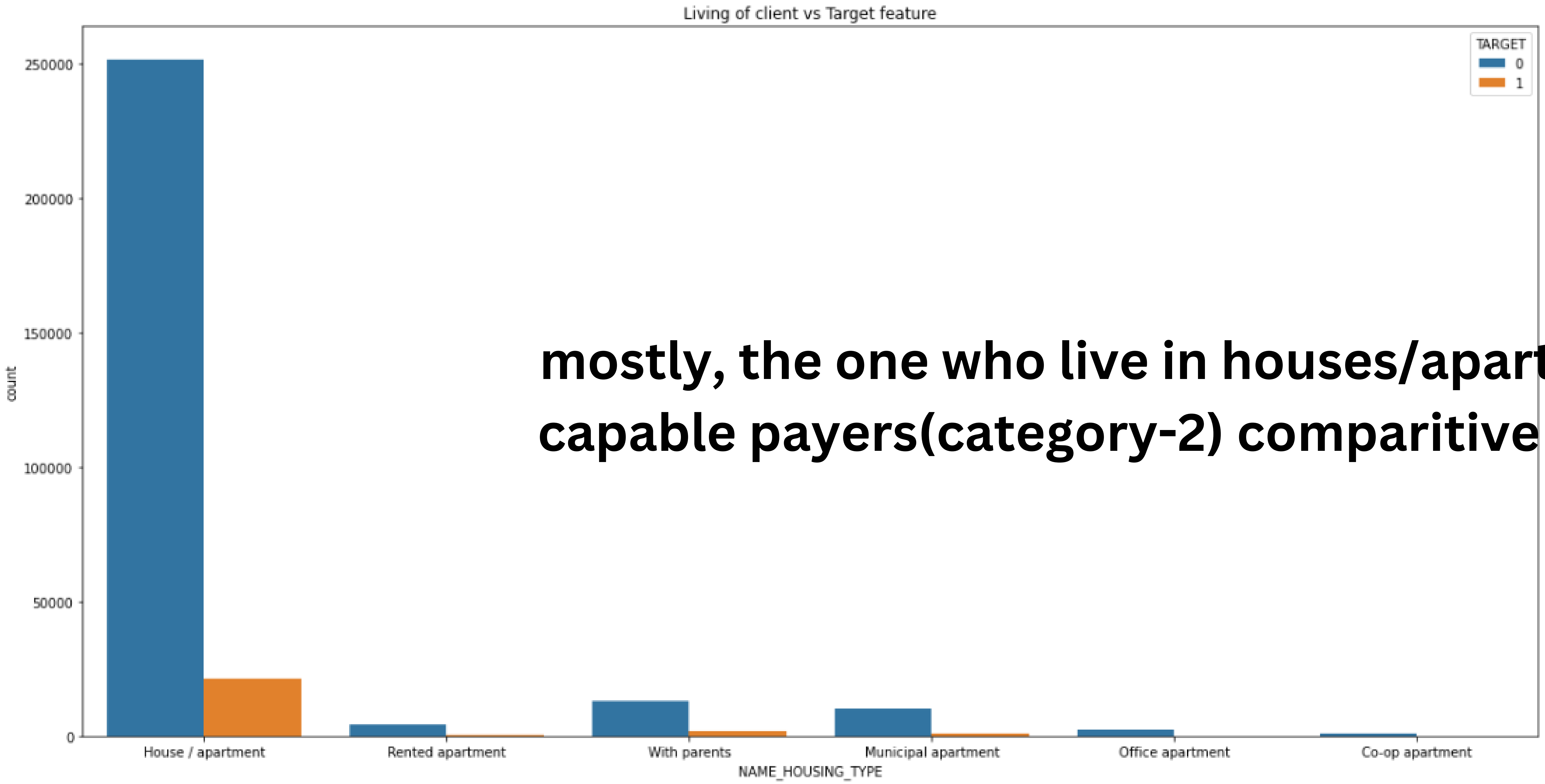
based on number of family members

```
plt.figure(figsize = (25, 20))
sns.catplot(data = df, x = "NAME_FAMILY_STATUS", hue="TARGET", kind = "count")
plt.title('based on the family status of the client')
plt.show()
```

```
<Figure size 1800x1440 with 0 Axes>
```



based on marital status of client, mostly married are capable payers



mostly, the one who live in houses/apartment are capable payers(category-2) comparitive to others

based on housing type of client

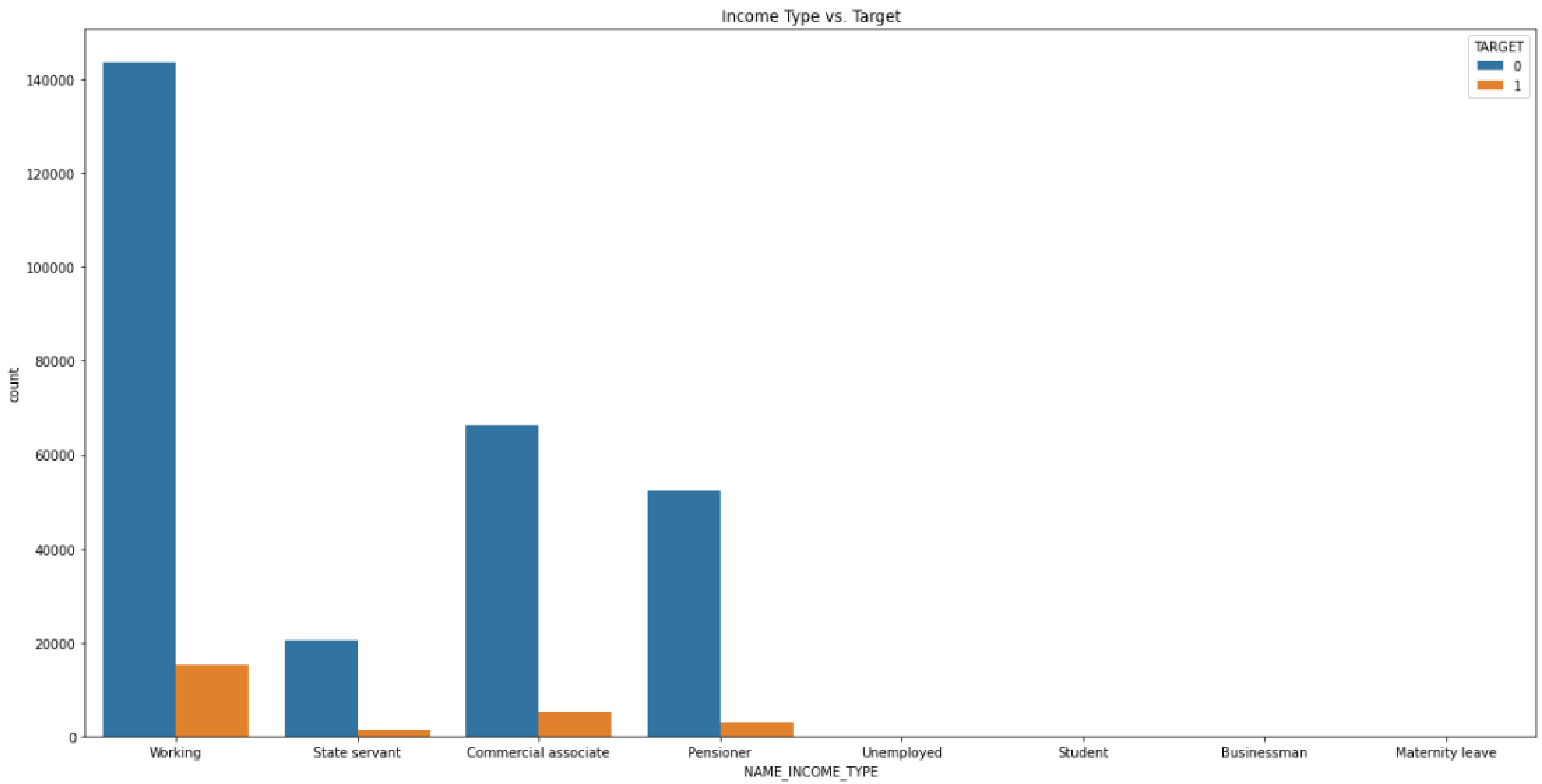
	Value	Percentage of category_1(clients who pay late, facing payment difficulties)
1	Rented apartment	12.313051
2	With parents	11.698113
3	Municipal apartment	8.539748
5	Co-op apartment	7.932264
0	House / apartment	7.795711
4	Office apartment	6.572411

we can differentiate from the above percent table

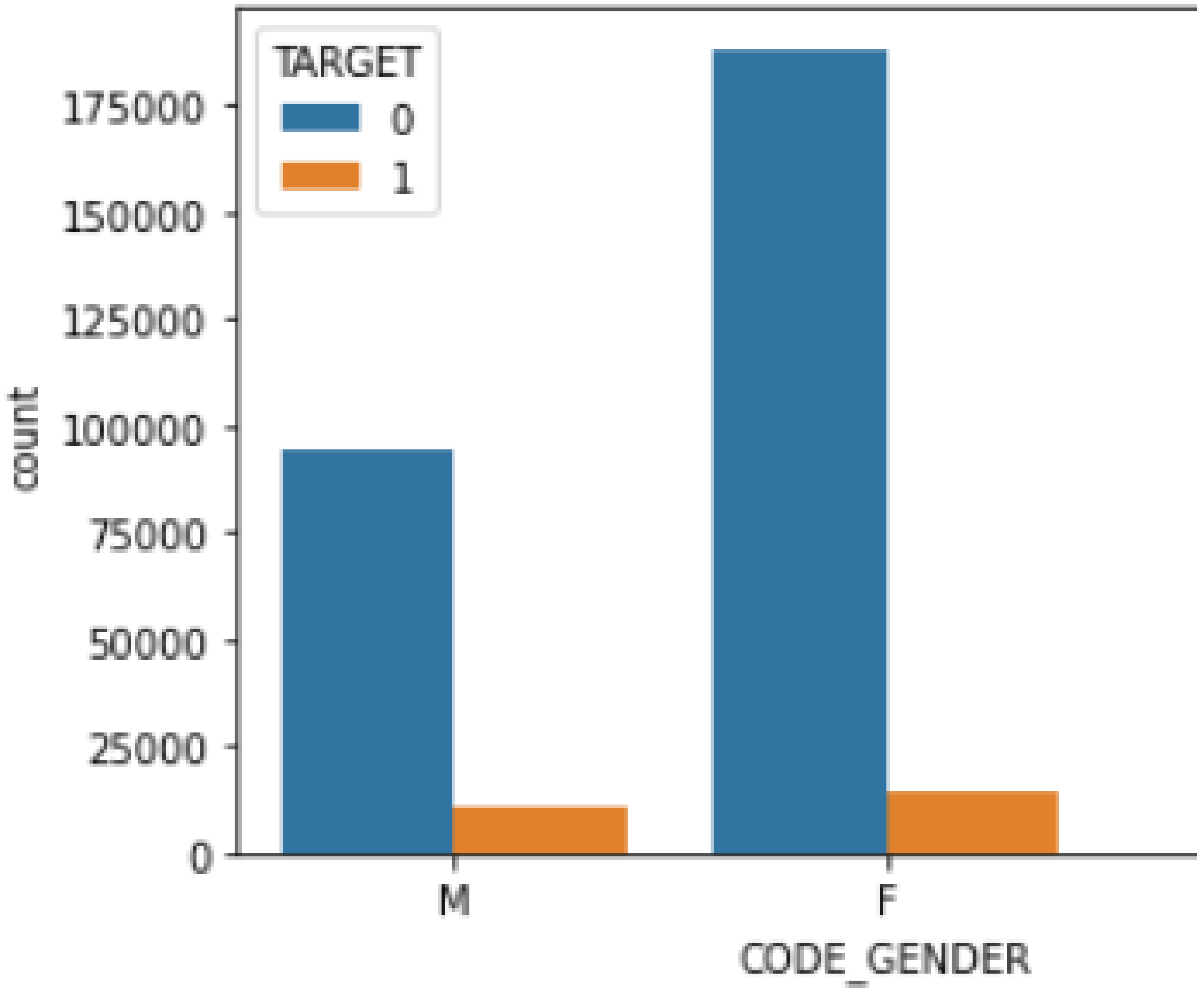
Children_count vs Target(category-1)

	Value	Percentage of category_1(clients who pay late, facing payment difficulties)
9	9.0	100.000000
10	11.0	100.000000
7	6.0	28.571429
4	4.0	12.820513
3	3.0	9.631423
1	1.0	8.923575
2	2.0	8.721821
6	5.0	8.333333
0	0.0	7.711809
5	7.0	0.000000
8	8.0	0.000000
11	12.0	0.000000
12	10.0	0.000000
13	19.0	0.000000
14	14.0	0.000000

here we can see client behaviour based on number of children he/she has



clients who are working (mostly come under cat-2), are not facing any difficulty, paying on time



As we know there are more female clients, even then the pattern is identical b/w the male and female applicants

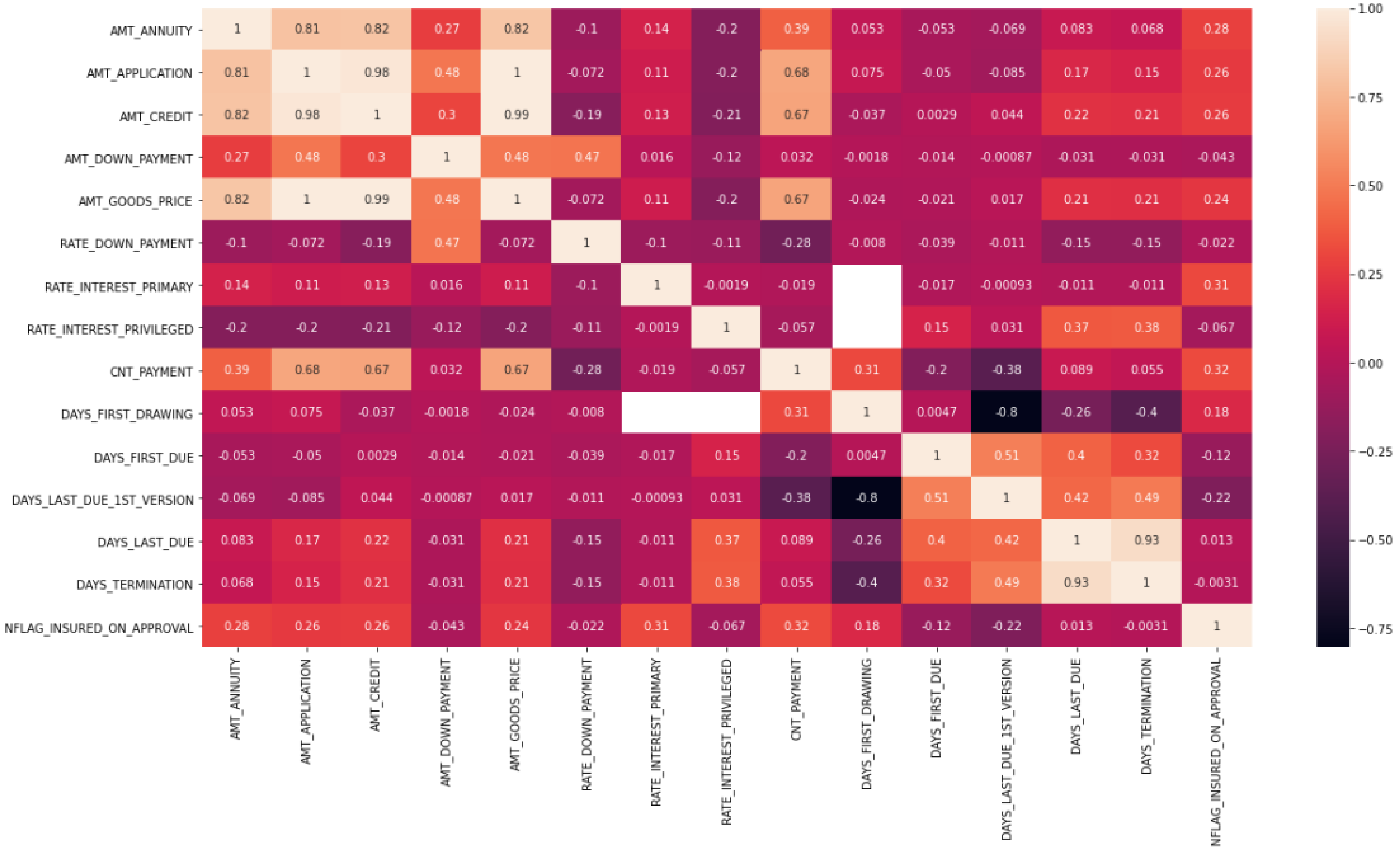
Top-10 correlated features(category-1)

```
SK_ID_CURR                    SK_ID_CURR                    1.000000
OBS_60_CNT_SOCIAL_CIRCLE      OBS_30_CNT_SOCIAL_CIRCLE      0.998269
BASEMENTAREA_AVG              BASEMENTAREA_MEDI             0.998250
COMMONAREA_MEDI               COMMONAREA_AVG                0.998107
YEARS_BUILD_MEDI              YEARS_BUILD_AVG               0.998100
NONLIVINGAPARTMENTS_AVG       NONLIVINGAPARTMENTS_MEDI      0.998075
FLOORSMIN_MEDI                FLOORSMIN_AVG                 0.997825
LIVINGAPARTMENTS_MEDI         LIVINGAPARTMENTS_AVG          0.997668
FLOORSMAX_MEDI                FLOORSMAX_AVG                 0.997187
NONLIVINGAPARTMENTS_MODE      NONLIVINGAPARTMENTS_MEDI      0.997032
ENTRANCES_MEDI                ENTRANCES_AVG                 0.996700
dtype: float64
```

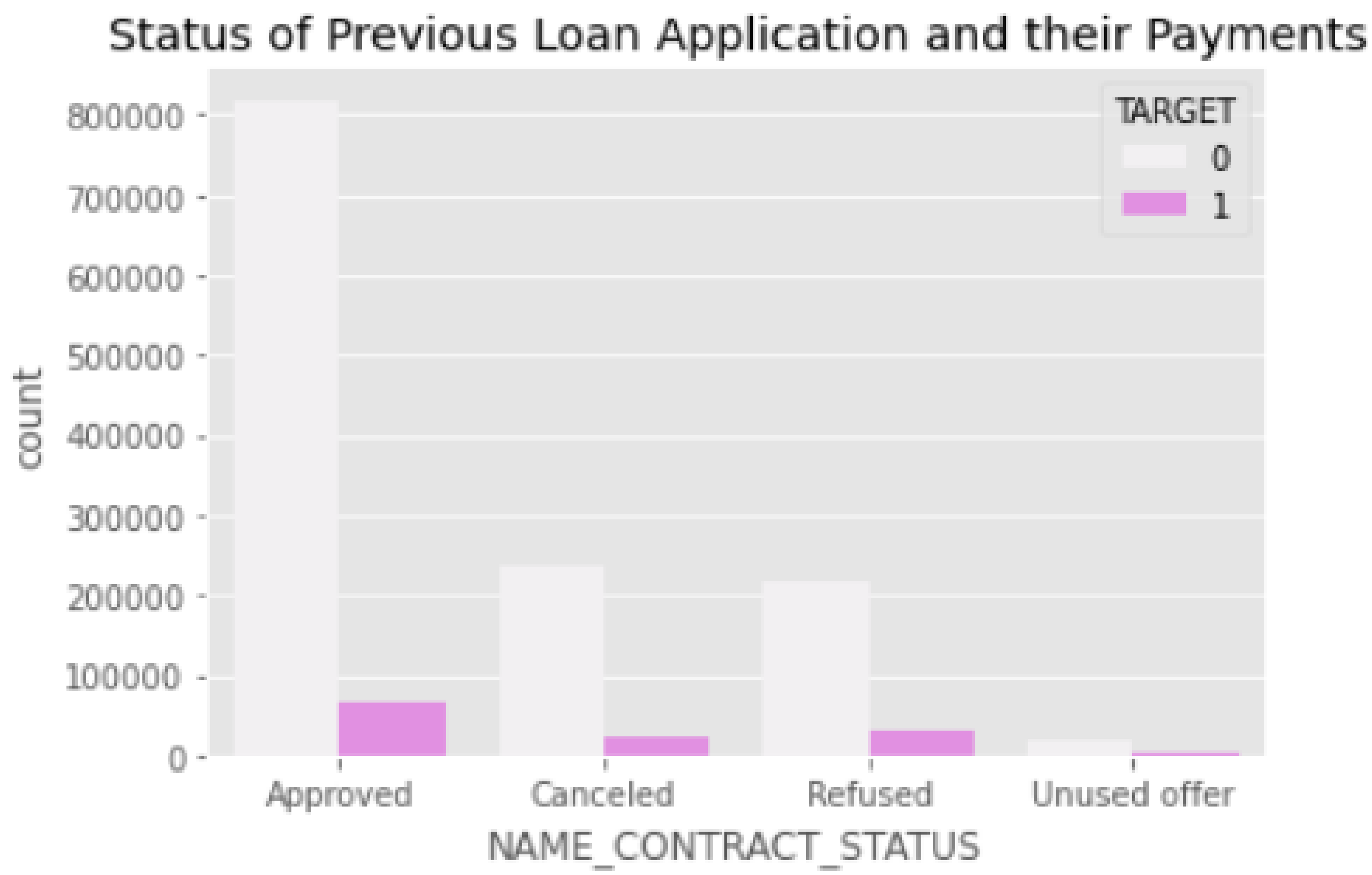
Top-10 correlated features(category-2)

```
SK_ID_CURR                  SK_ID_CURR                  1.000000
YEARS_BUILD_AVG             YEARS_BUILD_MEDI            0.998522
OBS_60_CNT_SOCIAL_CIRCLE    OBS_30_CNT_SOCIAL_CIRCLE    0.998508
FLOORSMIN_MEDI              FLOORSMIN_AVG               0.997202
FLOORSMAX_AVG               FLOORSMAX_MEDI              0.997018
ENTRANCES_AVG               ENTRANCES_MEDI              0.996899
ELEVATORS_AVG               ELEVATORS_MEDI              0.996161
COMMONAREA_AVG              COMMONAREA_MEDI             0.995857
LIVINGAREA_MEDI             LIVINGAREA_AVG              0.995568
APARTMENTS_AVG              APARTMENTS_MEDI             0.995163
BASEMENTAREA_MEDI           BASEMENTAREA_AVG            0.994081
dtype: float64
```

Correlation heatmap in previous data



merged the application and previous application datasets to compare with previous applicants and for better observation for the business objective



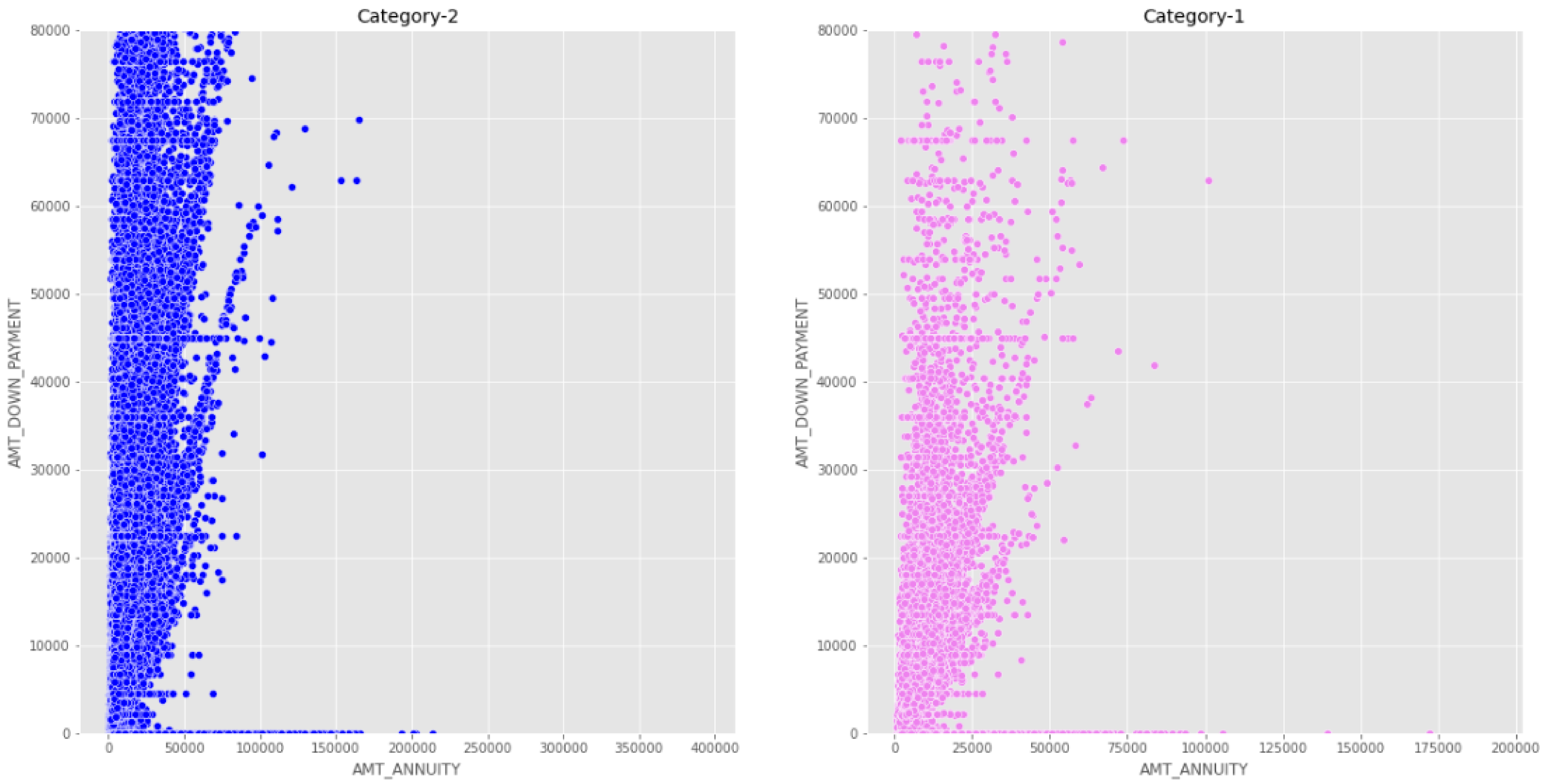
Percentage of previously approved loan applicants that come under category-1(late payment clients) in current loan

Percentage of previously approved loan applicants that came under category-1 in current loan :
7.588655443691958

Percentage of previously refused loan(that came under category-1) applicants that were able to pay current loan

Percentage of previously refused loan applicants that were able to pay current loan :
88.00358612820408

The applicants whose loans were previously approved more likely to pay for the current application (90% chances)



Category-2 applicants are less for larger amount of annuity of previous application. For higher down payment, Category-1 are less.

summary

This data is highly imbalanced as number of category-1 people is very less in total .

‘CNT_FAM_MEMBERS’, ‘CNT_CHILDREN’,’NAME_INCOME_TYPE’, ‘OCCUPATION_TYPE’,CODE_GENDER are the effective features hold importance towards target

Highly correlated features make impact

data has been cleaned and analysed to derive insights

Thank You

By korada saikiran