CREDIT EDA CASE STUDY

BY MAYANK TUSHAR & SUHANI SHUKLA


L O A N

WHAT IS THE PURPOSE OF CREDIT RISK ANALYSIS?

- The main purpose of credit risk analysis is to quantify the level of credit risk that the borrower presents to the lender.

- The three factors that lenders use to quantify credit risk include the probability of default, loss given default, and exposure at default.

EXPLORATORY DATA ANALYSIS (EDA)

- Approach to analyzing data sets to summarize their main characteristics, often with visual methods.

Following are the different steps involved in EDA :

- 1. **Data Collection** is the process of gathering information in an established systematic way that enables one to test hypothesis and evaluate outcomes easily.
- 2. **Data Cleaning** is the process of ensuring that your data is correct and useable by identifying any errors
- in the data, or missing data by correcting or deleting them.
- 3. **Data Preprocessing** is a data mining technique that involves transforming raw data into an understandable format
- 4. **Data Visualizations** is the graphical representation of information and data.

Business Understanding:

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. SUPPOSE YOU WORK FOR A CONSUMER FINANCE COMPANY WHICH SPECIALIZES IN LENDING VARIOUS TYPES OF LOANS TO URBAN CUSTOMERS. YOU HAVE TO USE EDA TO ANALYZE THE PATTERNS PRESENT IN THE DATA. THIS WILL ENSURE THAT THE APPLICANTS CAPABLE OF REPAYING THE LOAN ARE NOT REJECTED. WHEN THE COMPANY RECEIVES A LOAN APPLICATION, THE COMPANY HAS TO DECIDE FOR LOAN APPROVAL BASED ON THE APPLICANT'S PROFILE. TWO TYPES OF RISKS ARE ASSOCIATED WITH THE BANK'S DECISION:

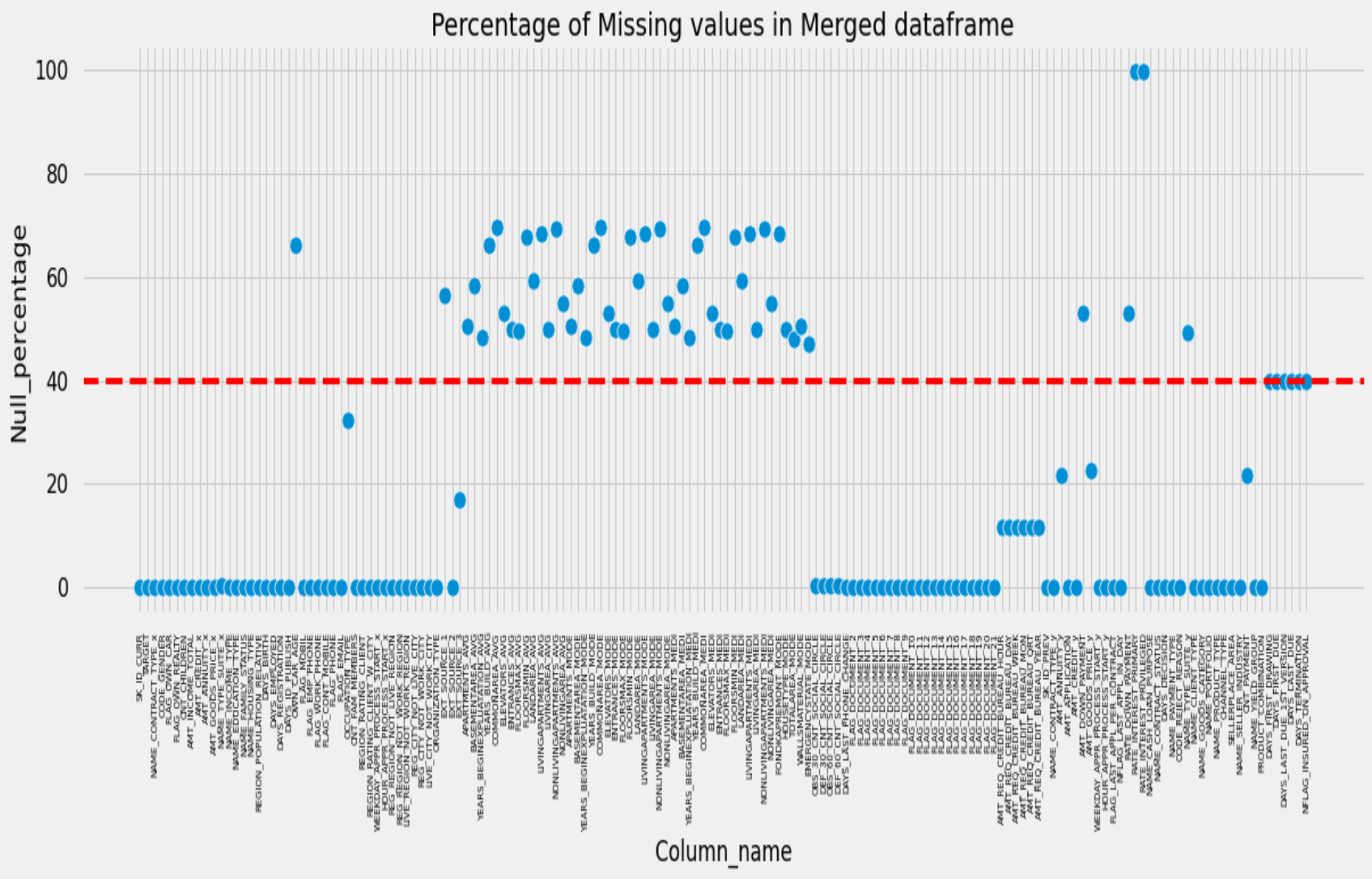
Business Objective:

This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study. In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

DATA CLEANING AND MANIPULATION

Points to be concluded from the graph on the right side.

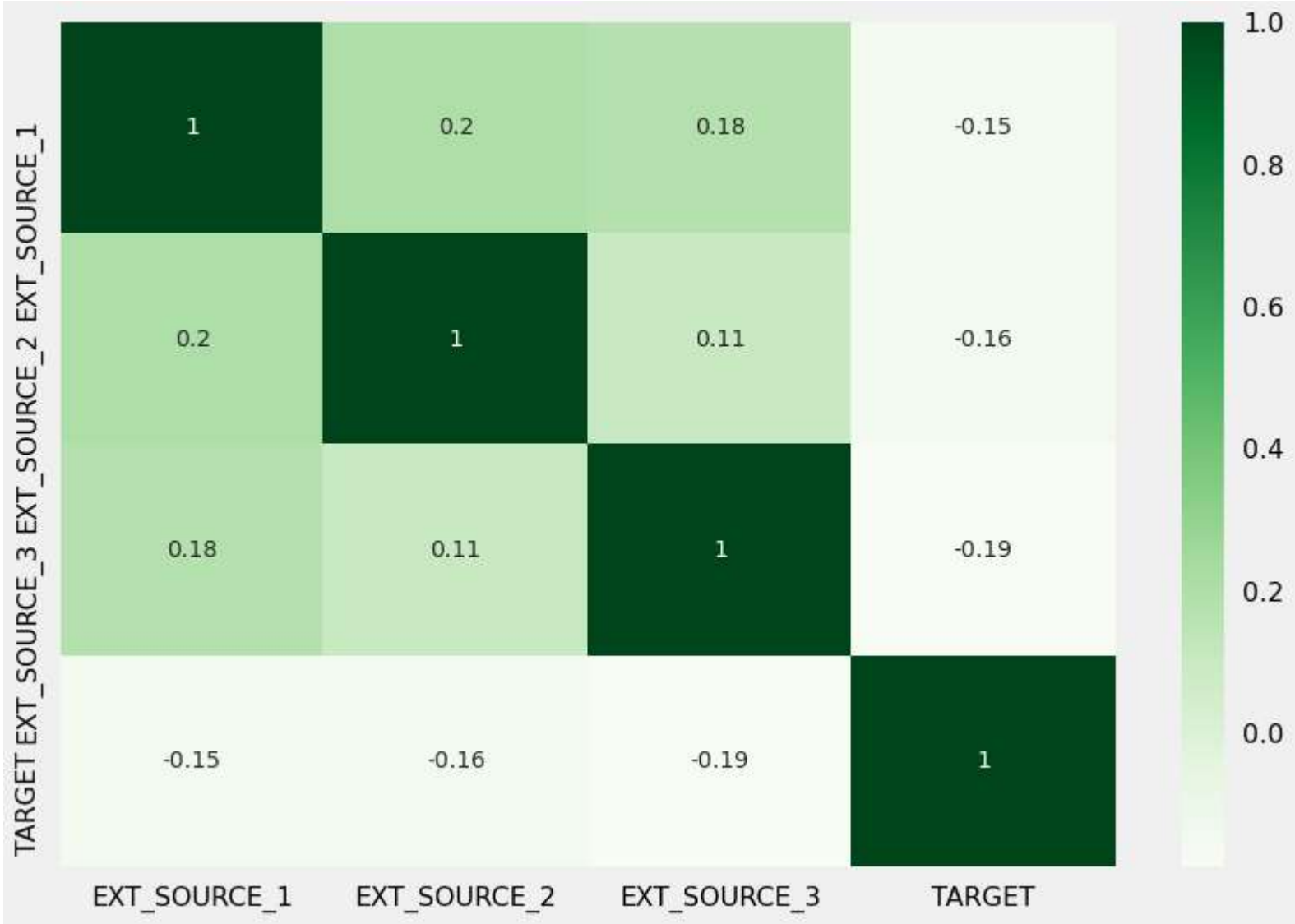
- calculating the percentage of null values in each column, rounded off to 2 decimal places
- we have tried to plot all the columns with respect to the percentage of null values present in them to a greater extent. So if we concentrate on those points which lie above the 40% null marker then we can understand that there are around 50 columns where the percentage of null values is present in more than 40% ratio. So, ultimately we need to treat these columns in order to normalize the dataframe for further analysis.



ANALYSIS AND HANDLING THE COLUMNS WITH NULL VALUES OR LOW CORELATIONS

Points to be concluded from the graph on the right.

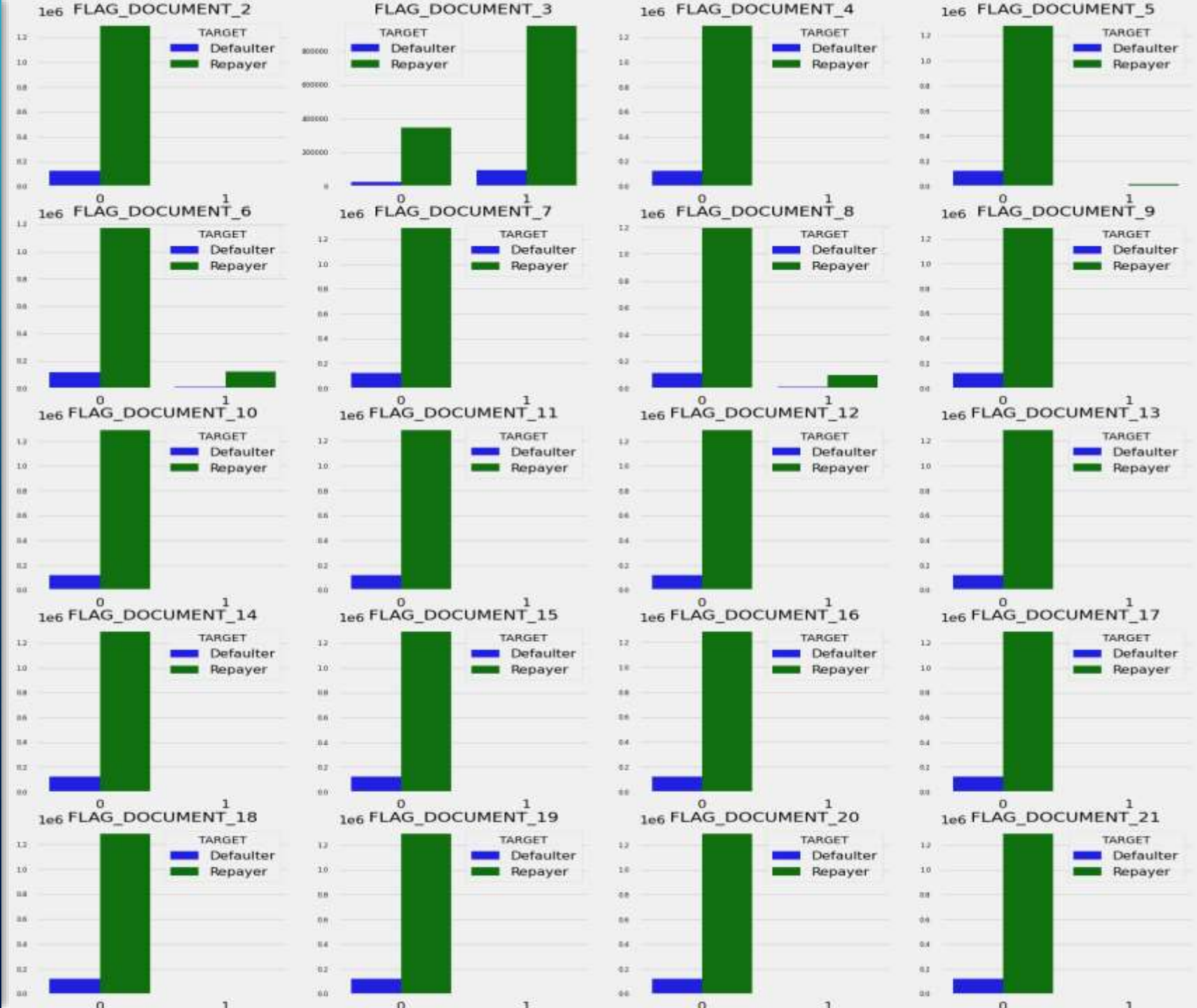
- If we compare EXT_SOURCE_1 to EXT_SOURCE_3 with respect to TARGET Column then we see that the corelation between them Falls on the negative side of the Scale which represents that there is very weak corelation between the EXT fields and the TARGET fields.
- Statistically speaking we see that EXT_SOURCE_1 has around 56% null values, EXT_SOURCE_3 has around 17% null values, and their corelation with TARGET field is less than zero respectively, thus we reach the conclusion that we can easily drop the 3 EXT_SOURCE_X columns from the merged dataframe df.



FLAG DOCUMENT FIELDS CORELATION ANALYSIS

Points to be concluded from the graph on the right.

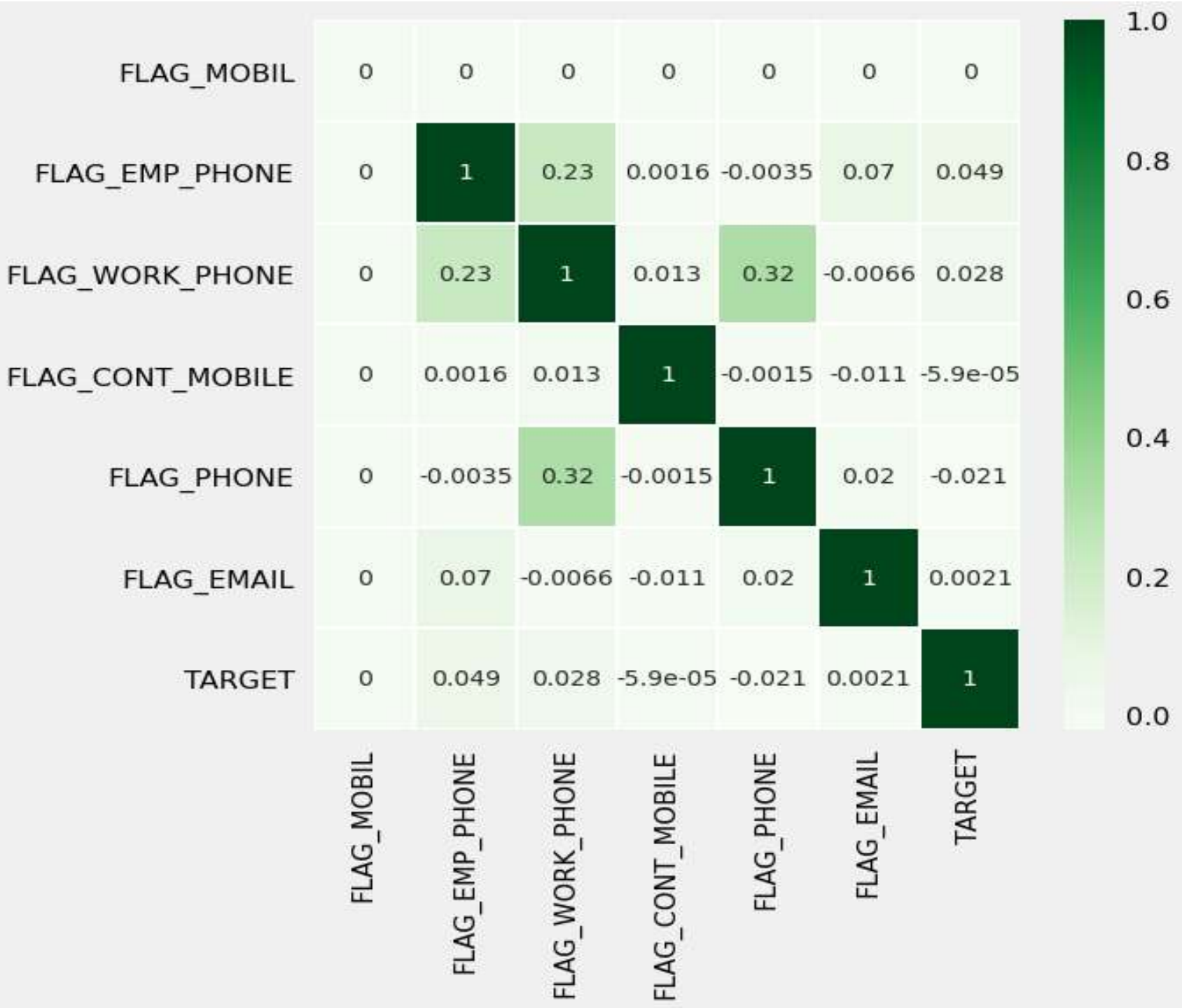
- checking the relevance and corelation of FLAG columns of the merged dataframe by creating a new dataframe df_flag.



CONTACT PARAMETERS CORRELATION ANALYSIS

Points to be concluded from the graph on the right.

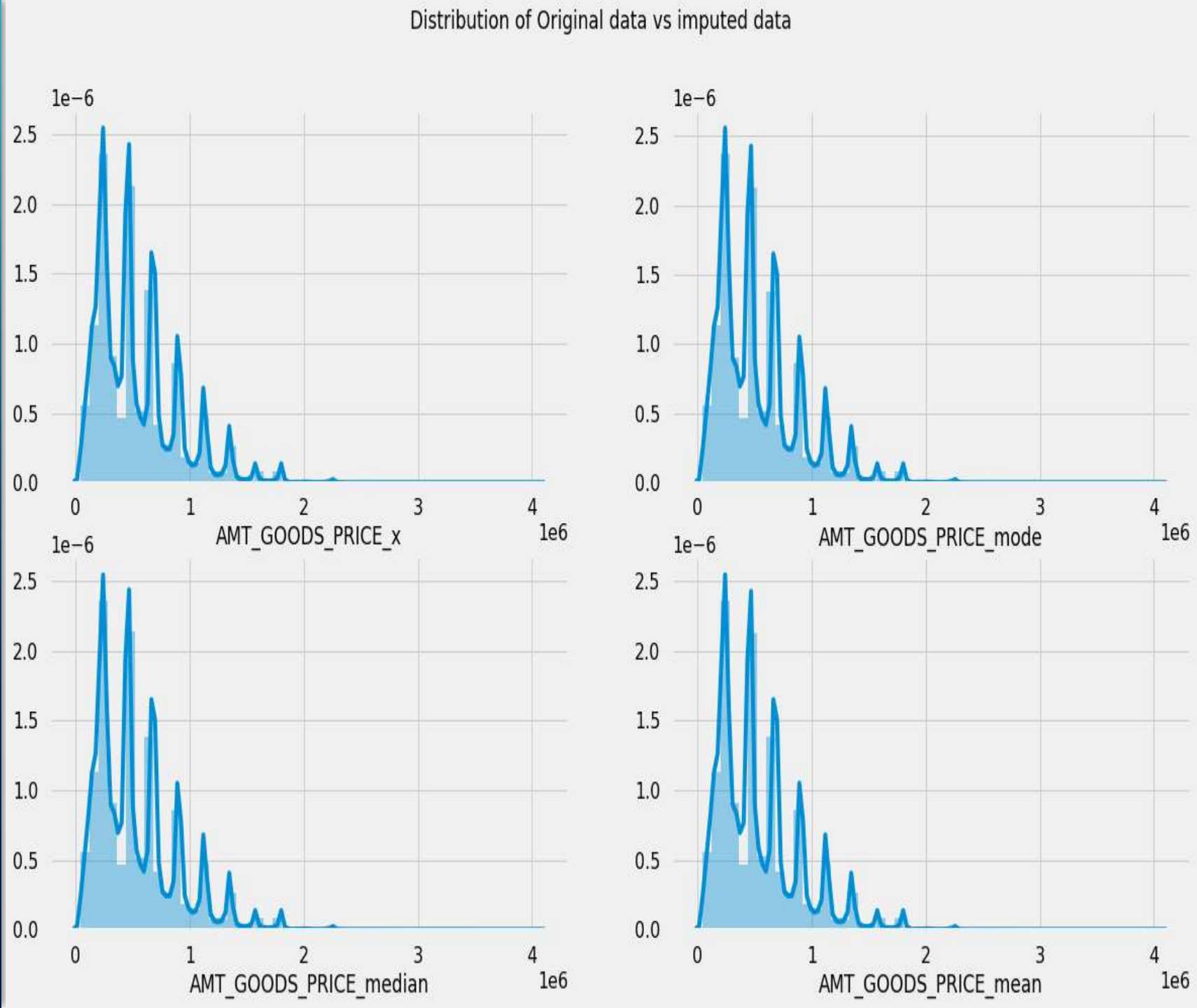
- There are no green colored fields, which means there is no such strong correlation between the contact parameters and the TARGET column.
- Thus, we can drop these columns from the merged dataframe.



FILLING IN ZERO FOR NAN VALUES FOR NULL PERCENTAGE COLUMNS BETWEEN 30 TO 40

Points to be concluded from the graph on the right side.

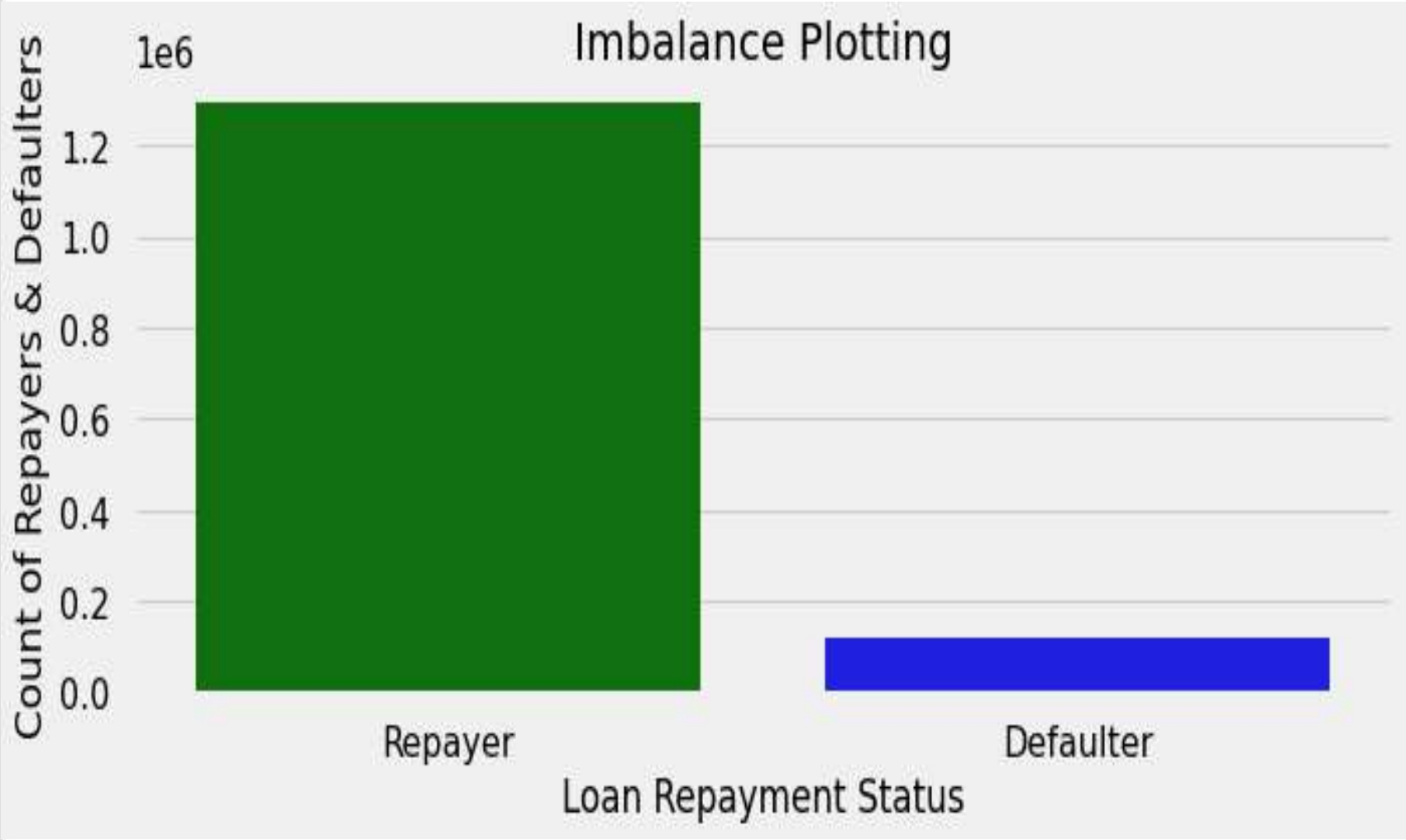
- Finding out percentage of null values in each column of merged dataframe df.
- when we fill the column AMT_GOODS_PRICE_x with mean, median and mode values, then after plotting the graph we can clearly see that the graphs of respective fillings is exactly similar to the one where there is nothing filled in the place of NaN values.
- The mean, median and mode values for given column is biased/skewed to the left side and is thus filled with 0, as they all are equal.



DATA ANALYSIS
IMBALANCE ANALYSIS

Points to be concluded from the graph on the right side.

- Creating a new dataframe named Imbalance to hold the Target column value counts data without index.
- Total values that are present in the TARGET column, if we take 1 as defaulter and 0 as repayer, the there is an imbalance in the values. 0 taken precendence over the 1 value in this column in huge proportion.
- The proptionate analysis is done below, where we can see the difference in proportion in the values present in the TARGET column



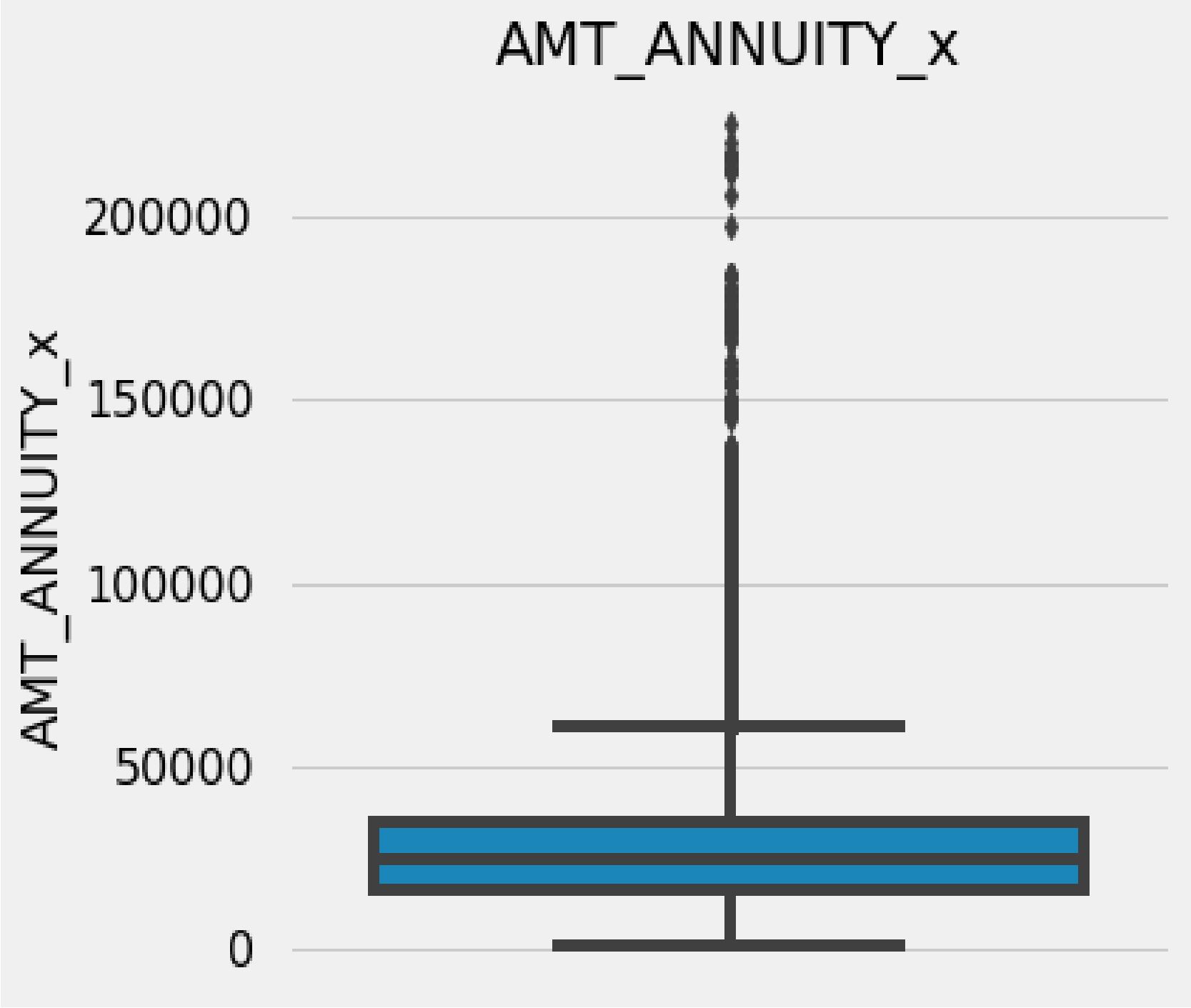
UNIVARIATE ANALYSIS

- This type of data consists of **only one variable**. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. The example of a univariate data can be height.

AMT_ANNUITY

Points to be concluded from the graph on the right.

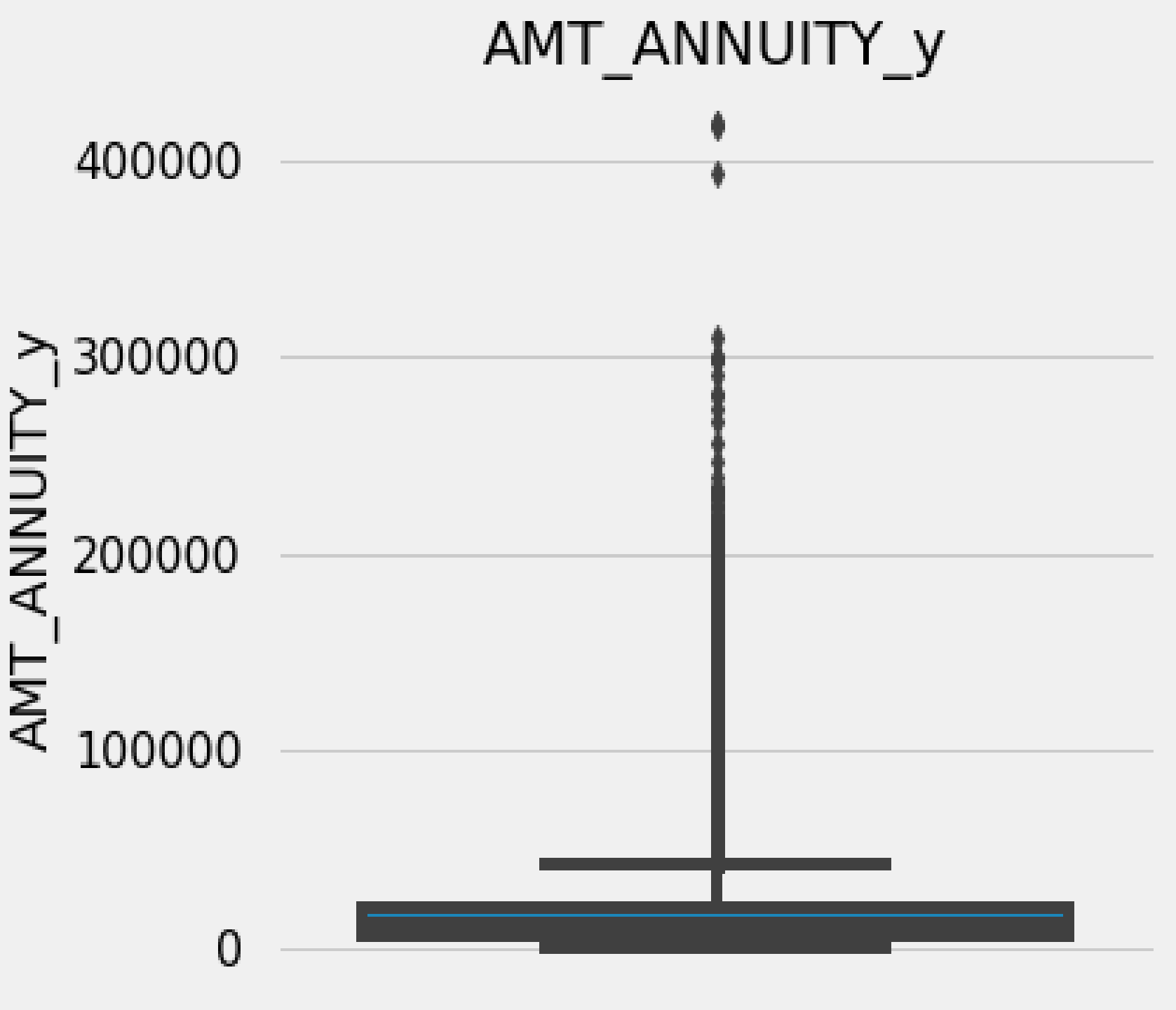
- There are a huge number of outliers present in this data for AMT_ANNUITY_x (ADF), and the median lies somewhere around 25000.
- The upper fence lies around 60000 and lower fence of the data lies at 0.



AMT_ANNUITY_Y

Points to be concluded from the graph on the right.

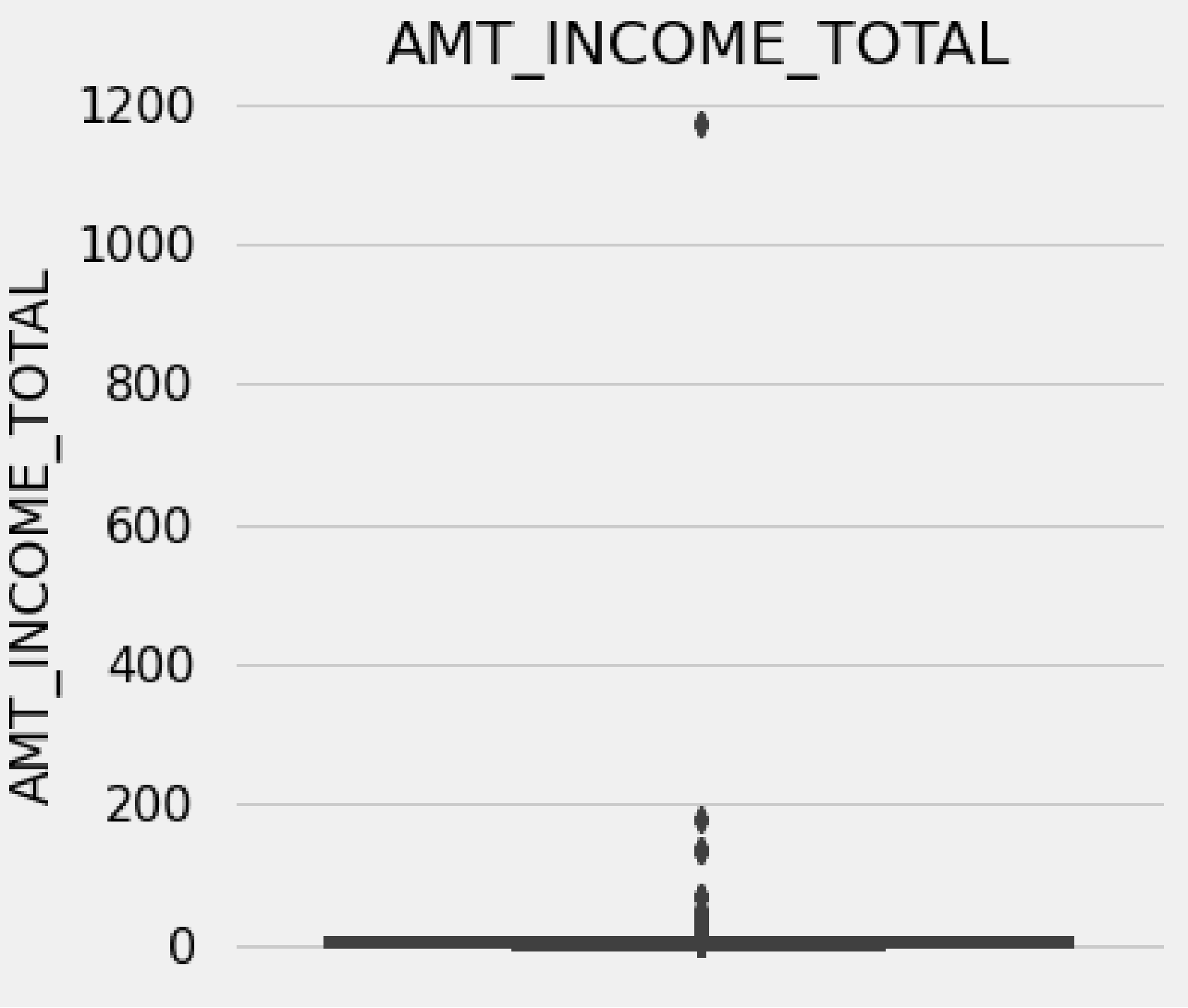
- There are a whole lot of outliers present in the data, where there are some which are extremely outlying.
- There are only few applicants which can pay the annual annuity amount equal to 400000 approximately.
- Other outliers lie between the range of 50000-300000. Median value is around 25000.



AMT_INCOME_TOTAL

Points to be concluded from the graph on the right.

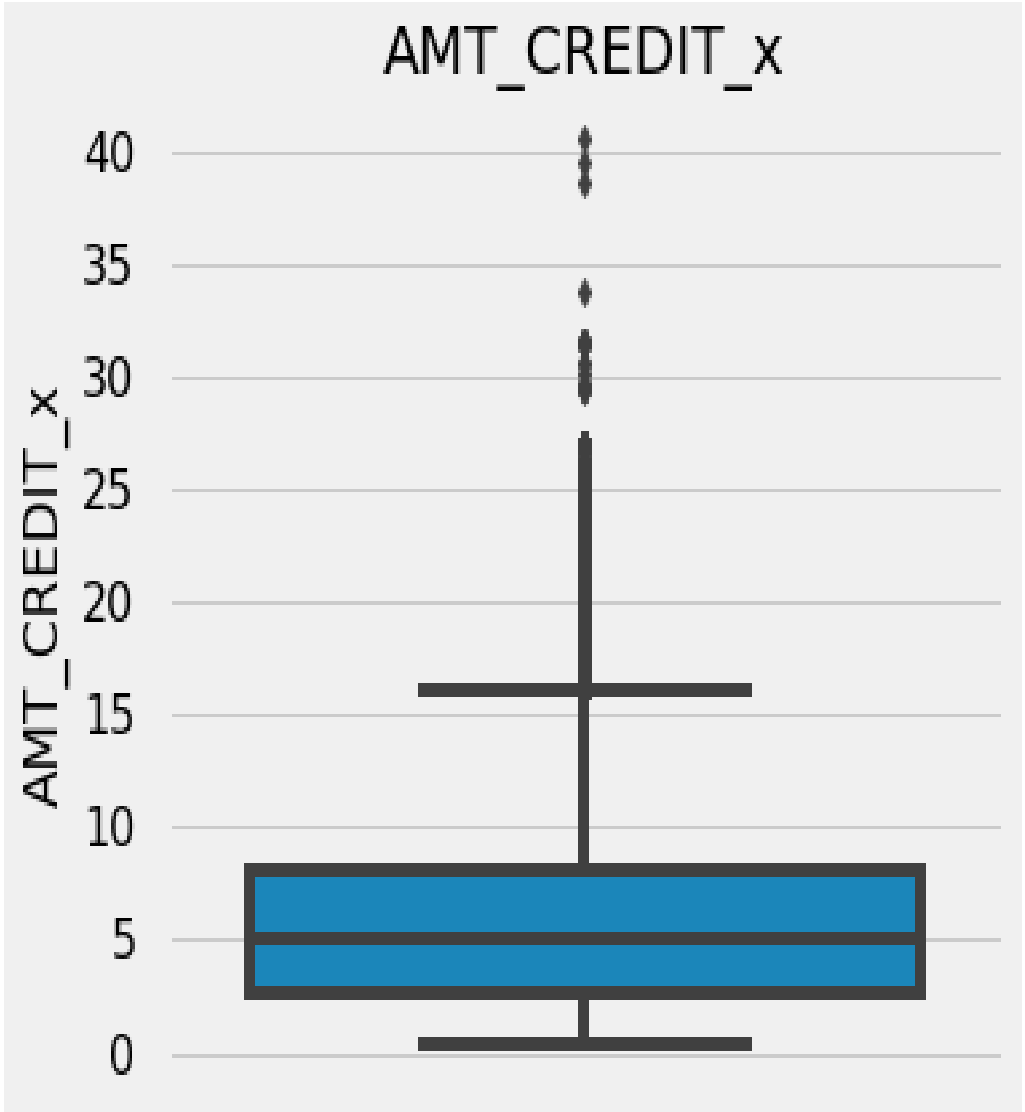
- There are not many outliers. There are some extremely present outliers, which have salary around 1200K, which seem like an impossible salary or income to have because of the gap present in the outlying salaries/incomes. Most of the incomes are present below 200K only.
- The lower and upper fences along with median value are very close to zero.



AMT_CREDIT

Points to be concluded from the graph on the right.

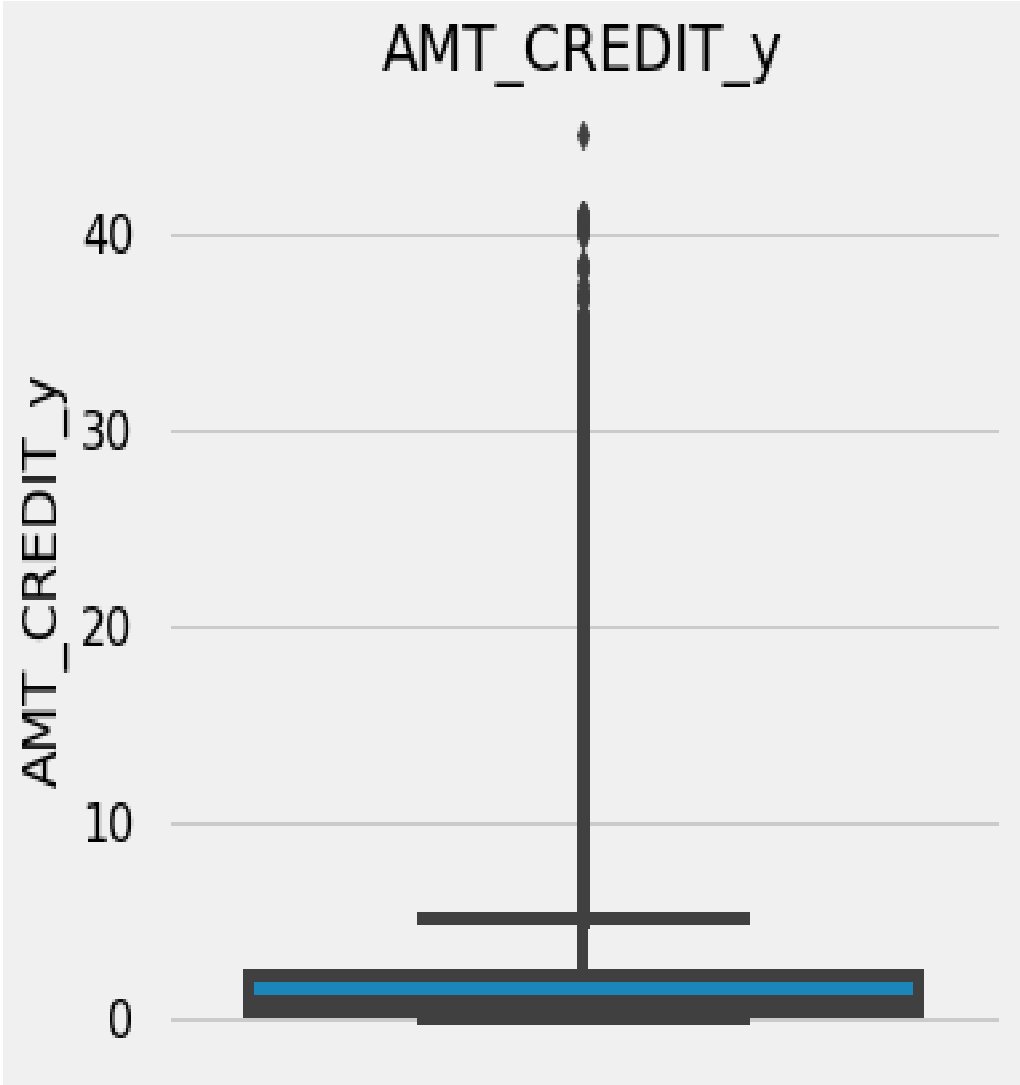
- There are again a lot of outliers present, where majority of the credit amounts passed between the range of 0-1500K.
- The majority of outliers are also present in the range of 1500K to 3500K, where there are only some applicants who have been sanctioned a credit limit of the same range.
- The extreme most credit loan amounts are very rare and are found above 3700K approximately.
- The upper fence lies at around 1500K whereas the median value is lying at 500K. So we can say that the most number of loans passed for this year lies in the range of 0-1500K.



AMT_CREDIT

Points to be concluded from the graph on the right.

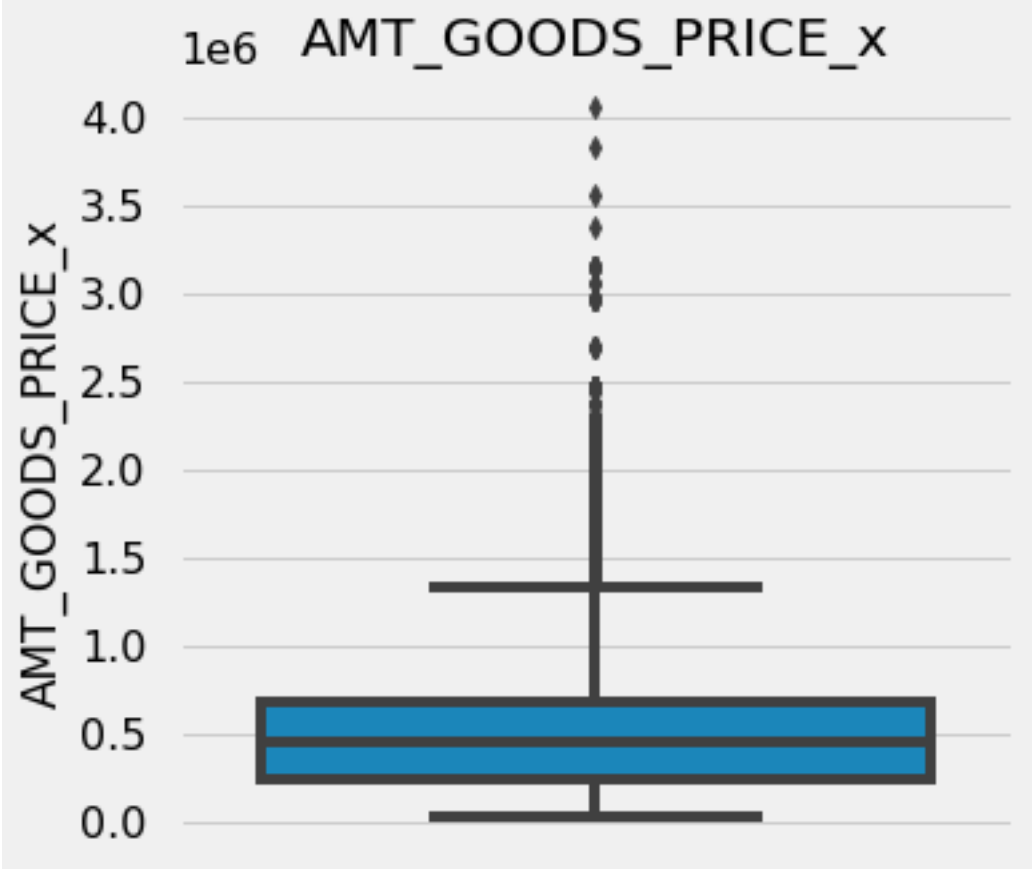
- There are a huge number of outliers for the previous year credit loan amount data. The majority of loans passed are in the range of 0-500K.
- The extreme value lies for 4000K or more credit loan amount.



AMT_GOODS_PRICE_X

Points to be concluded from the graph on the right.

- The majority of the goods bought using the credit loan amount lies in the range of 0 to 150K,
- Along with a huge number of outliers where extreme values are present after 250K in a very rare manner for the current fiscal year.



AMT_GOODS_PRICE_Y

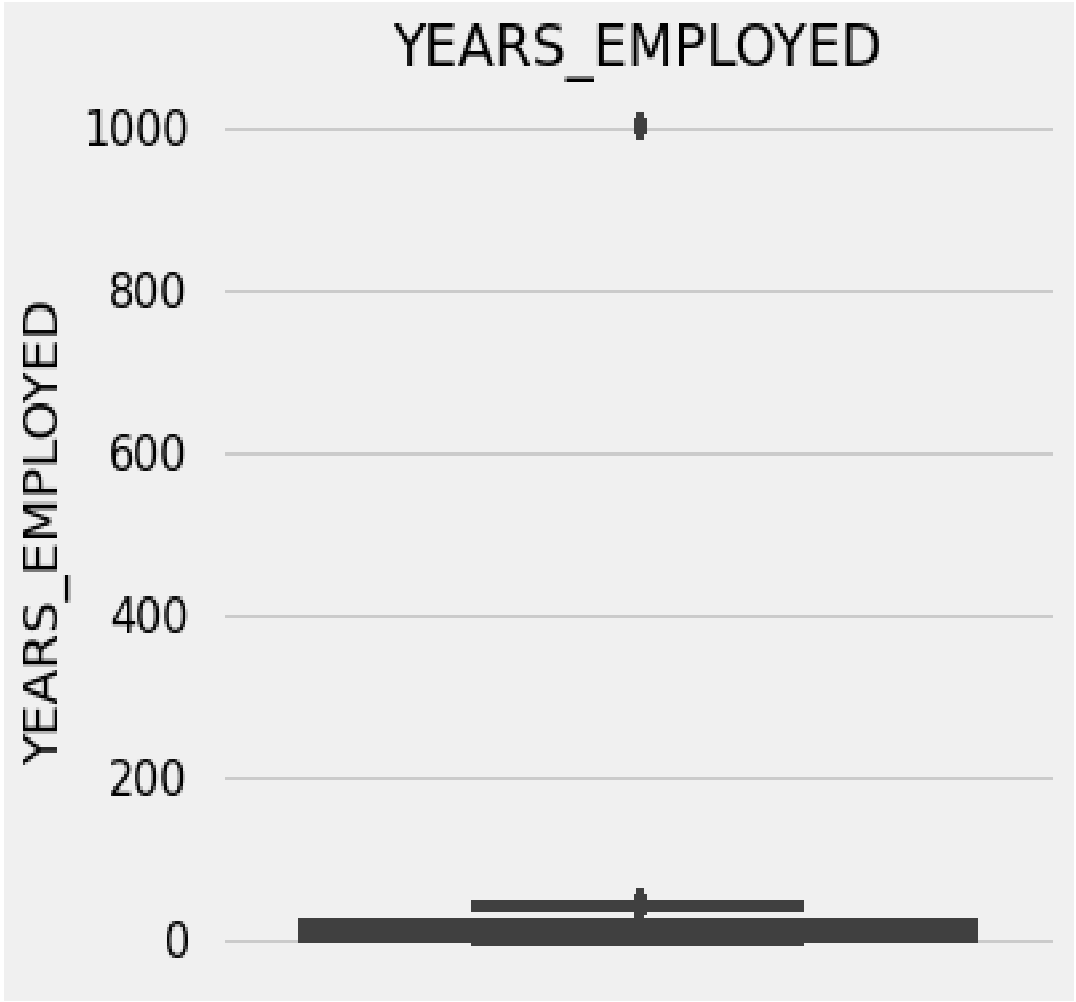
Points to be concluded from the graph on the right.

- For the previous year, we see that the majority of the goods bought using the credit loan amount lies in the range of 0-0.5k, which seems like a valid value.
- Whereas we also see that there are huge number of outliers, where the most extreme falls at around 600K, which also seems like a valid value if the largest credit released is also around the same.

YEARS_EMPLOYED

Points to be concluded from the graph on the right.

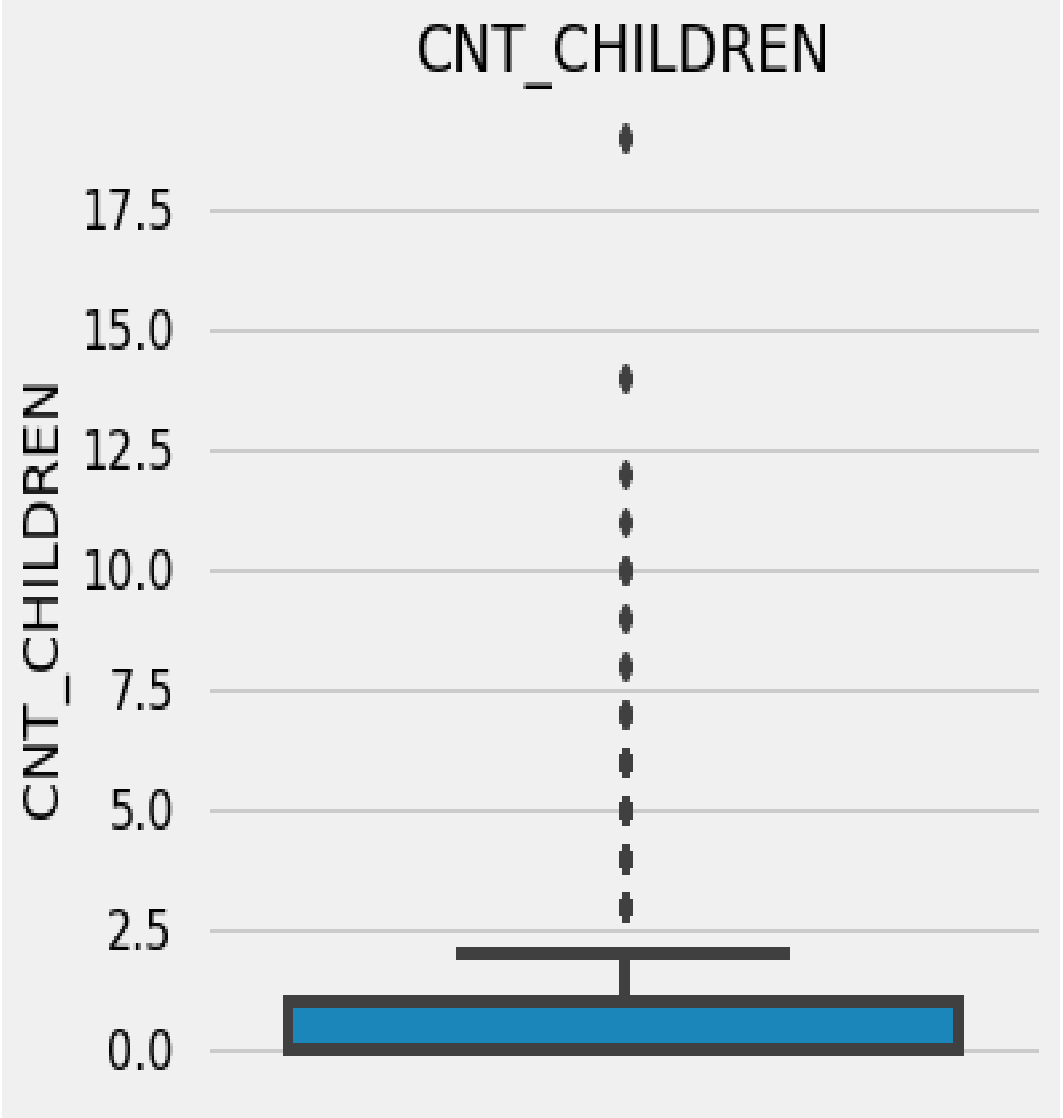
- There are not much outliers present, but the one which is present in clearly an invalid value since no body can be employed for around a 1000 years,
- Thus skewing the data plot, and we should ignore that in our future comparisons and analysis.
- The majority of time period for which an applicant has been employed falls in the valid range of 0-30 years.



CNT_CHILDREN

Points to be concluded from the graph on the right.

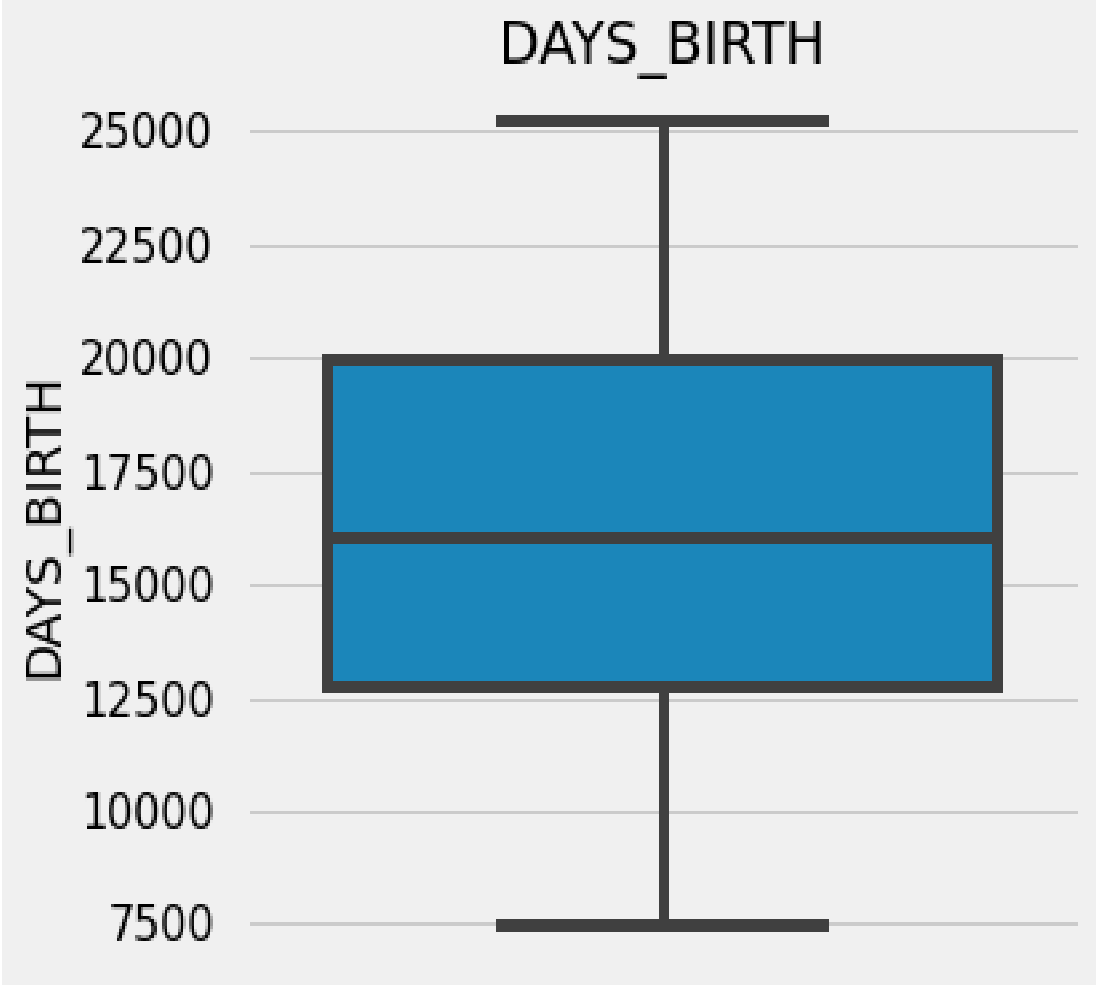
- The volume of the box lies in the region of 0-2 children, which means that most of the families which applied for the loans this year have had 2 children.
- Further there are other number of outliers present, which go the extreme of 18 children or more, which seem almost impossible.
- Majority of the outliers fall in the region of 2 children to 14 children.



DAYS_BIRTH

Points to be concluded from the graph on the right.

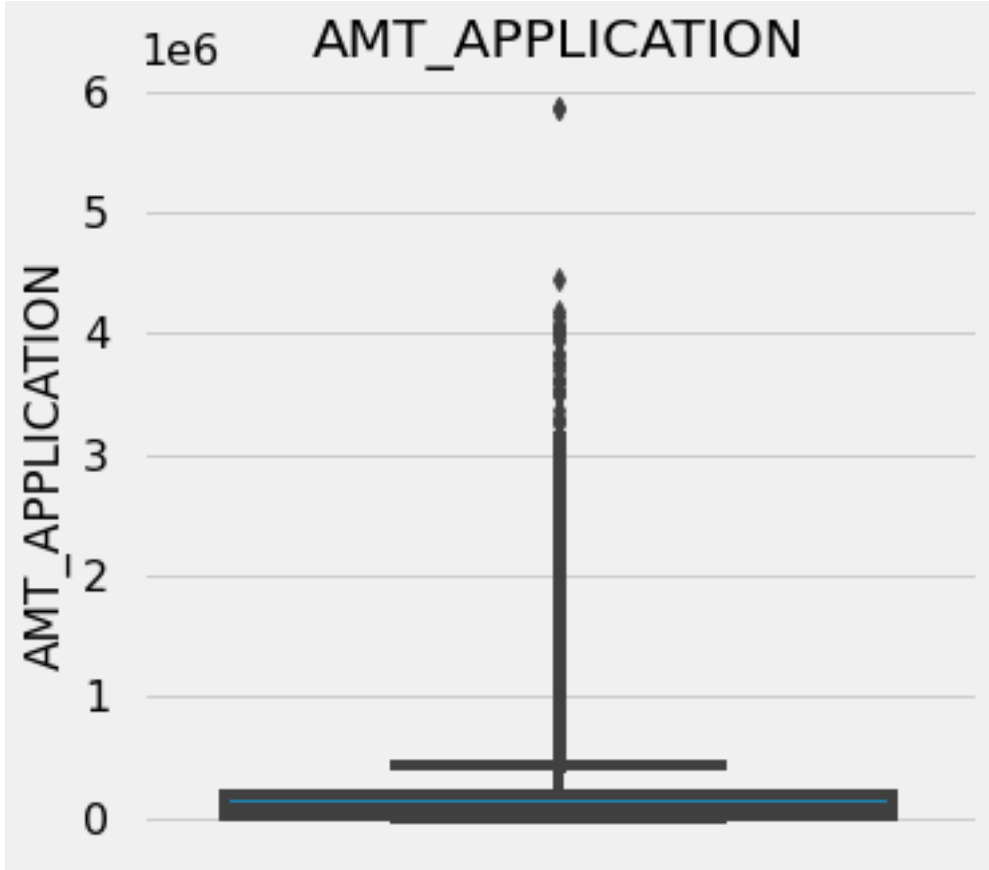
There are no outliers present in the data, and all values are valid values in this case



AMT_APPLICATION

Points to be concluded from the graph on the right.

- There are lots of outliers prsent in the data with extreme values present at around 600k.
- The majority of data falls in the range of 0-50k.

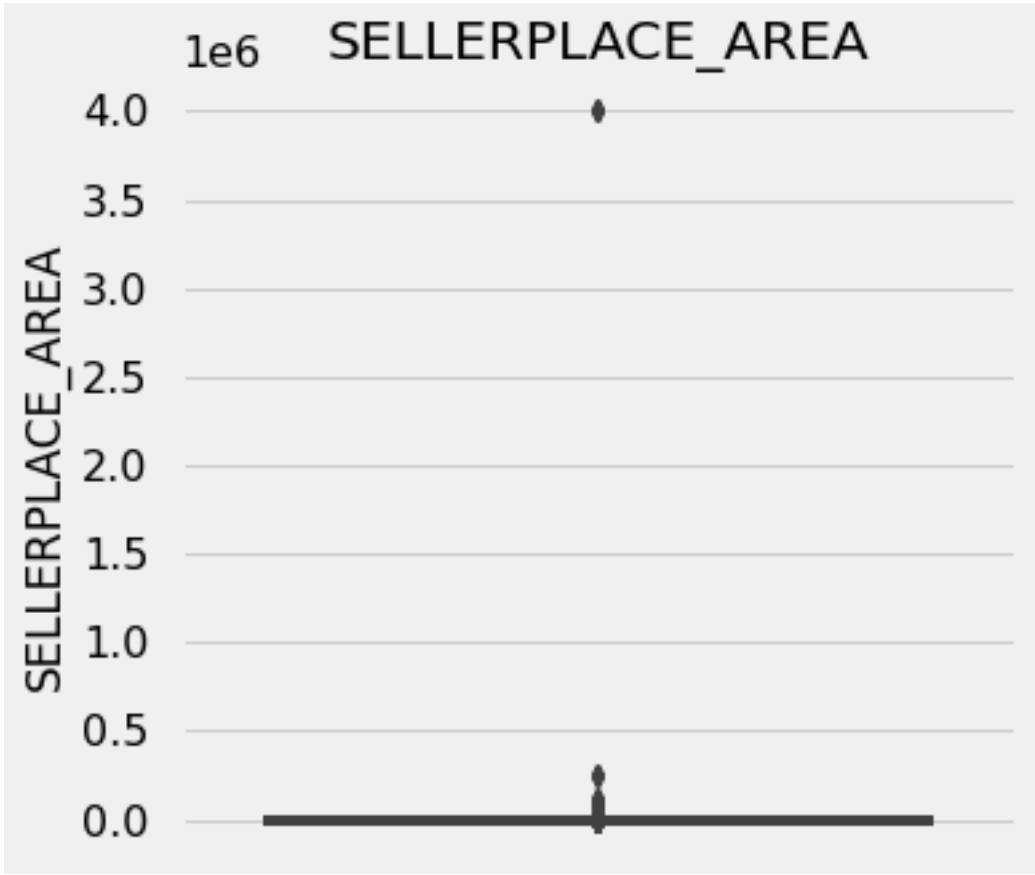


SELLERPLACE_AREA

Points to be concluded from the graph on the right.

•There are some outliers present in the data, with some extreme values like 400000 which cannot be true or valid.

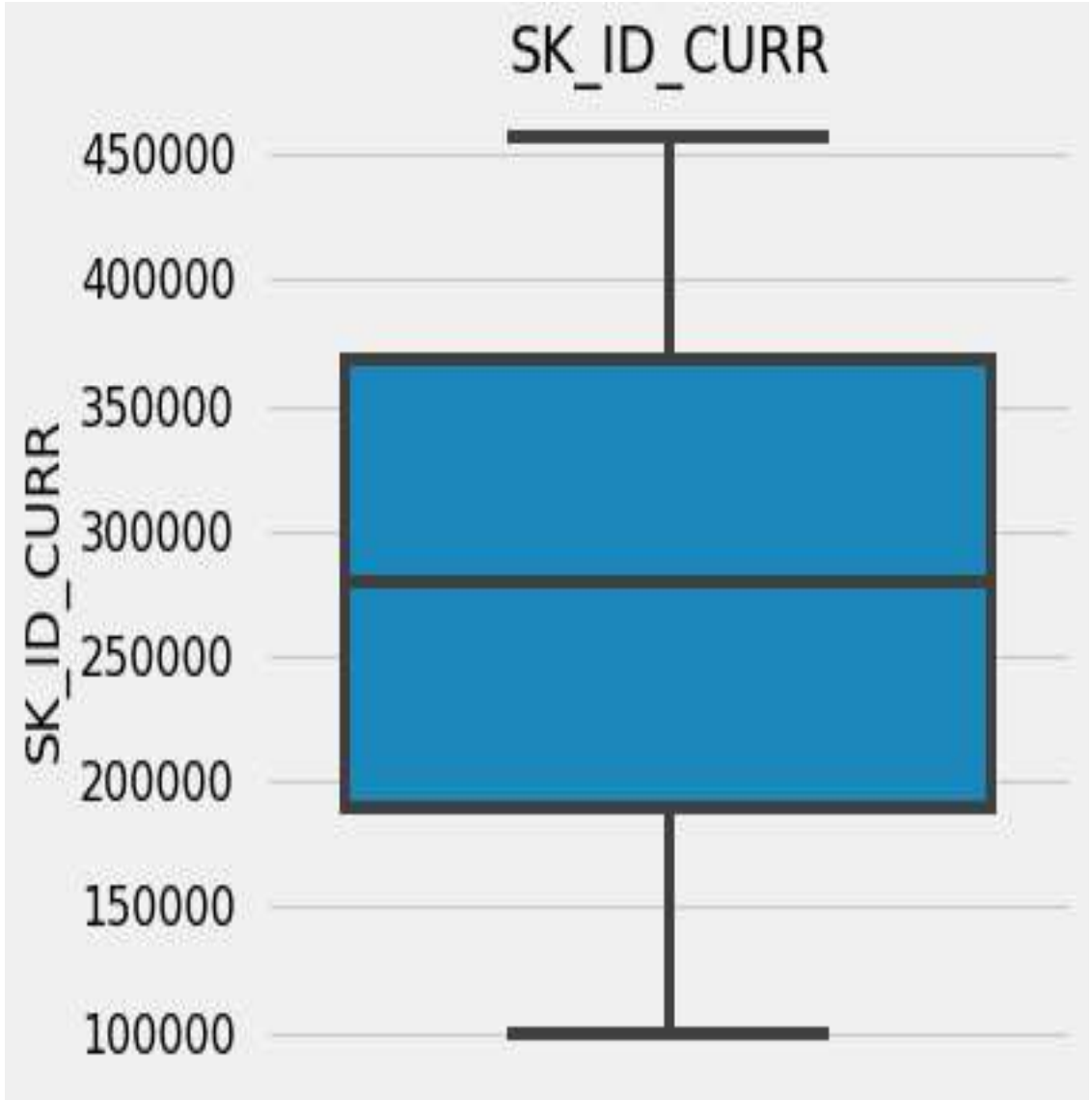
•The majority of the data lies near 0 as the upper limit, lower limit and median value, all are plotted near zero.



SK_ID_CURR

Points to be concluded from the graph on the right.

•There are no outliers present in the data, and all values are valid values in this case.



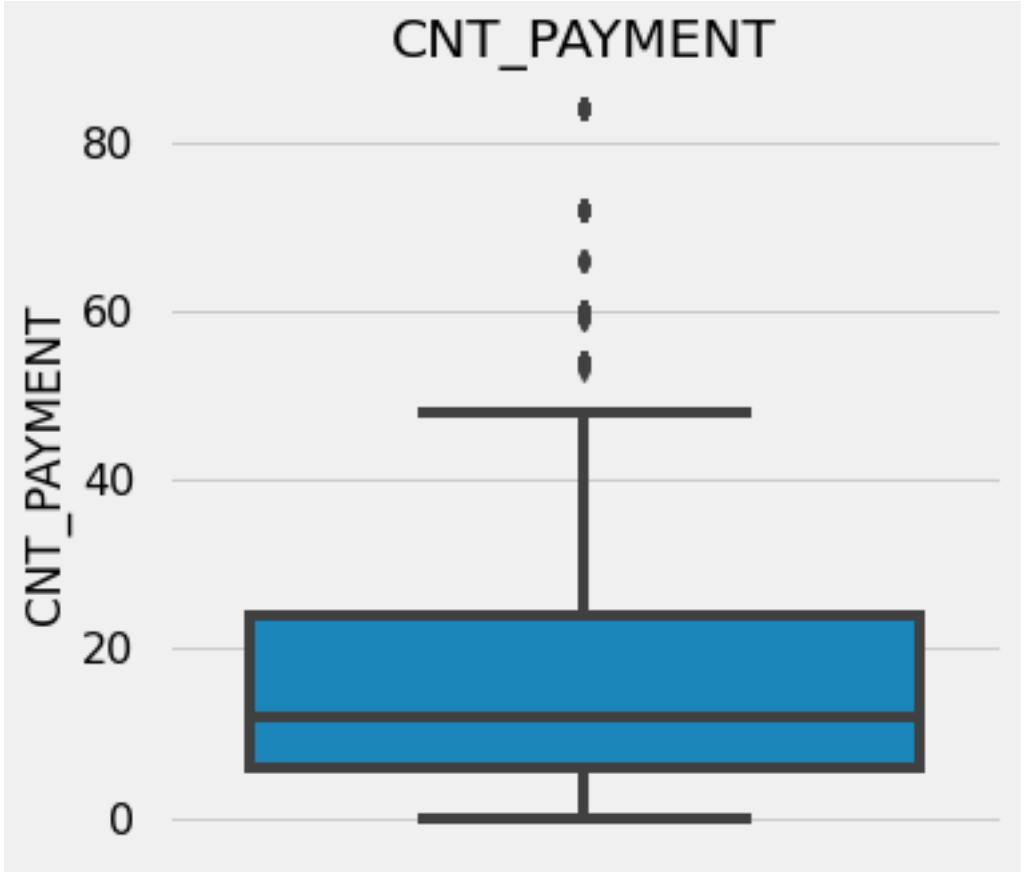
CNT_PAYMENT

Points to be concluded from the graph on the right.

• : There are some outliers present in the data, where the outlier value ranges between 50 and 90.

•The majority of the payment counts at the time of loan application lie in the range of 0-50, with median value at around 10.

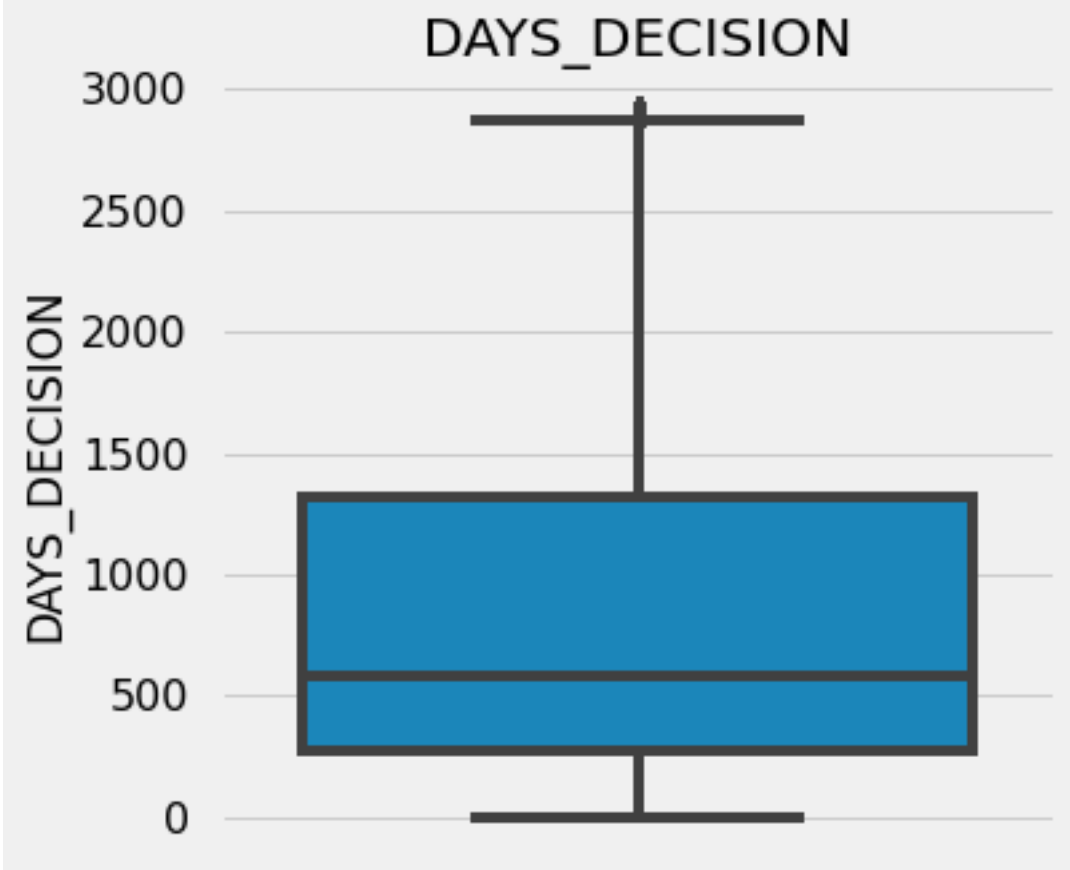
•Thus, most of the loan applicants applied for loan, when their loan repayment counts were within 50, and only some applied when the count of repayments lie in the range 50-90.



DAYS_DECISION

Points to be concluded from the graph on the right.

- There are a very less number of outliers present in the data,
- but that shows that there are some cases where days taken to confirm the decision to take credit loan is more than the majority values.
- The majority of the values lies in the range 300 to 1400 days, with median value around 600 days.

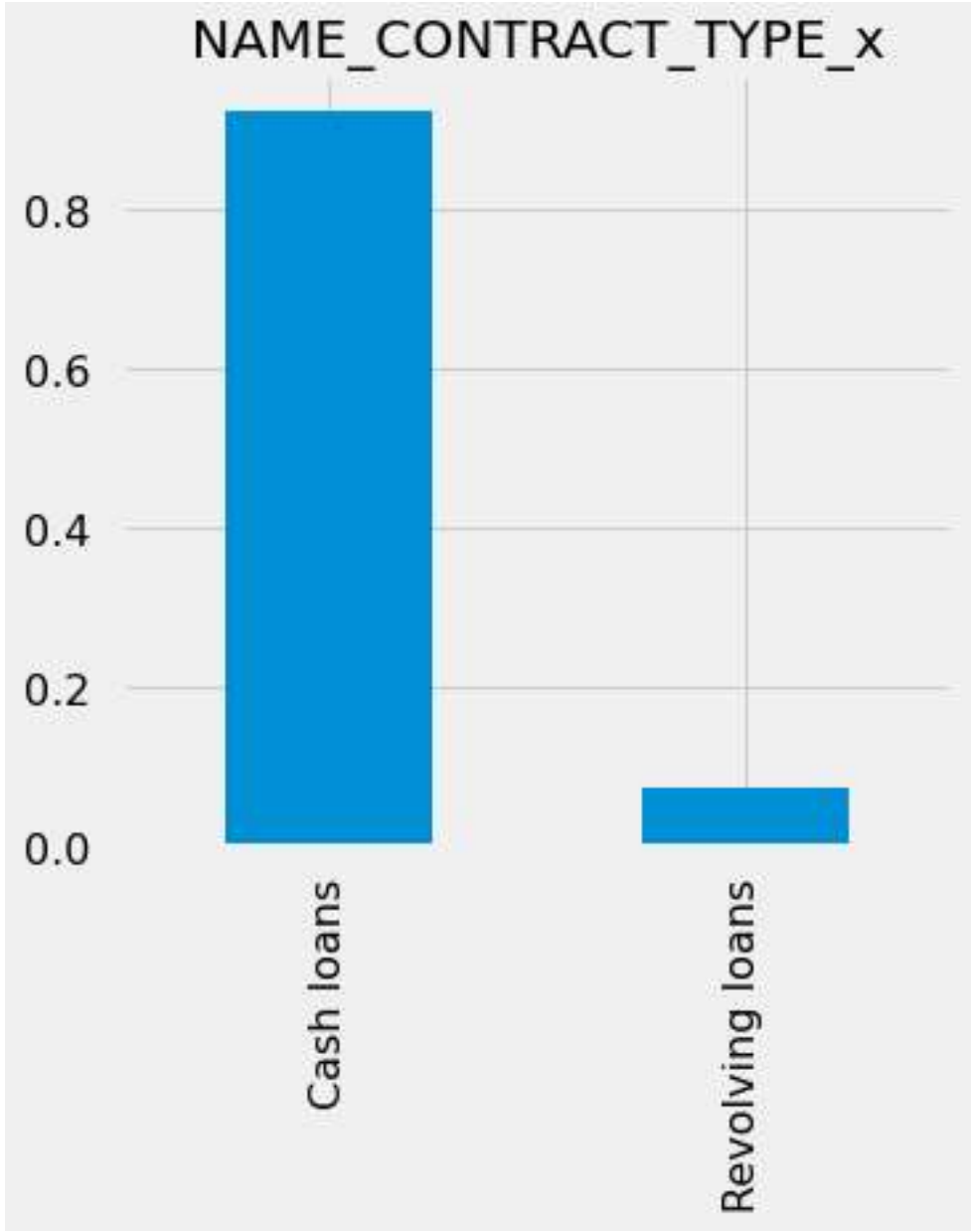


CATEGORICAL UNIVARIATE ANALYSIS

NAME_CONTRACT_TYPE_X

Points to be concluded from the graph on the right.

- From the above Bar plot, we can see that the number of Cash loans is far more than the total number of revolving loans (92 percent vs 7 percent).

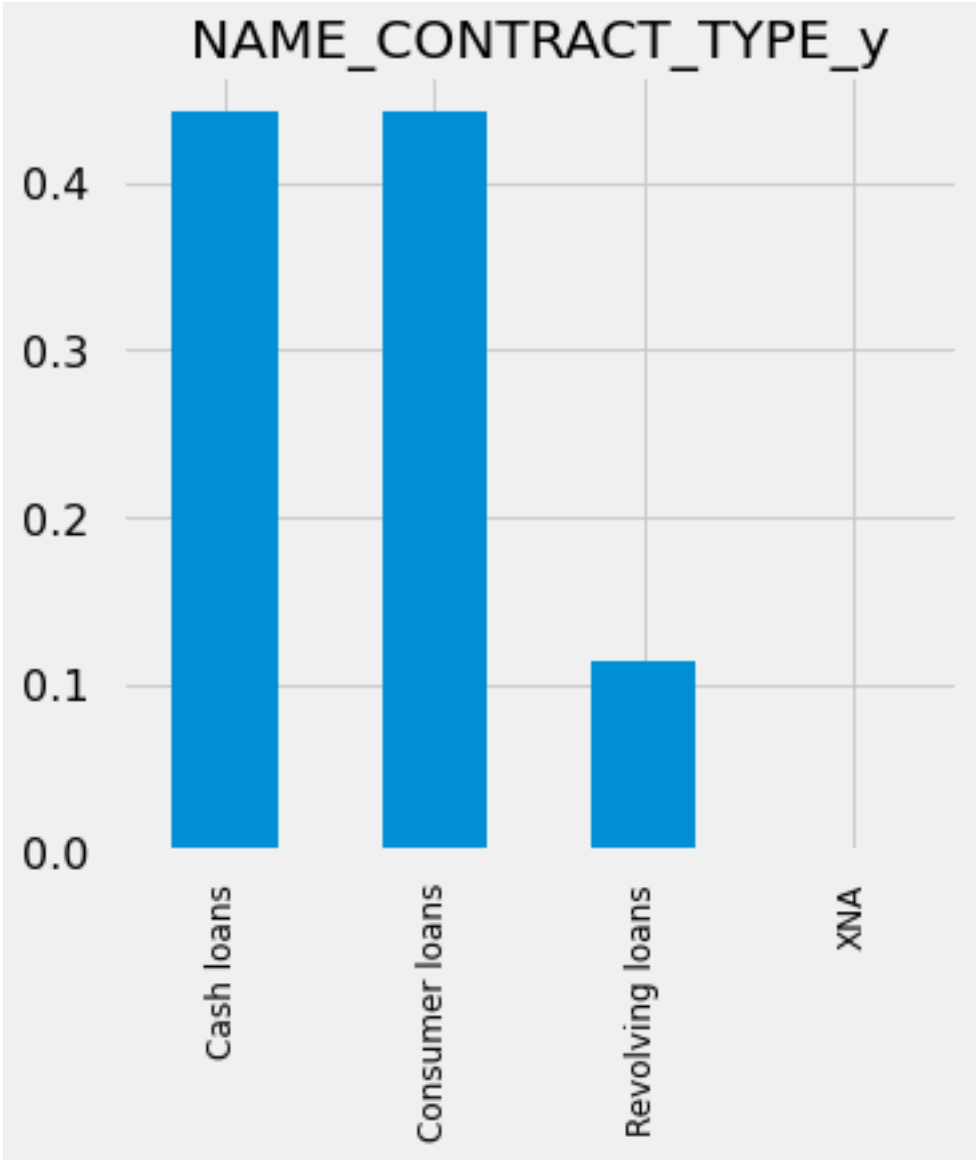


NAME_CONTRACT_TYPE_Y

Points to be concluded from the graph on the right.

- By taking a look at the above bar chart, we can conclude that the maximum number of applications have contract type as "Cash loans" and "Consumer loans" at 44 percent each and there is no data in "XNA" contract type.

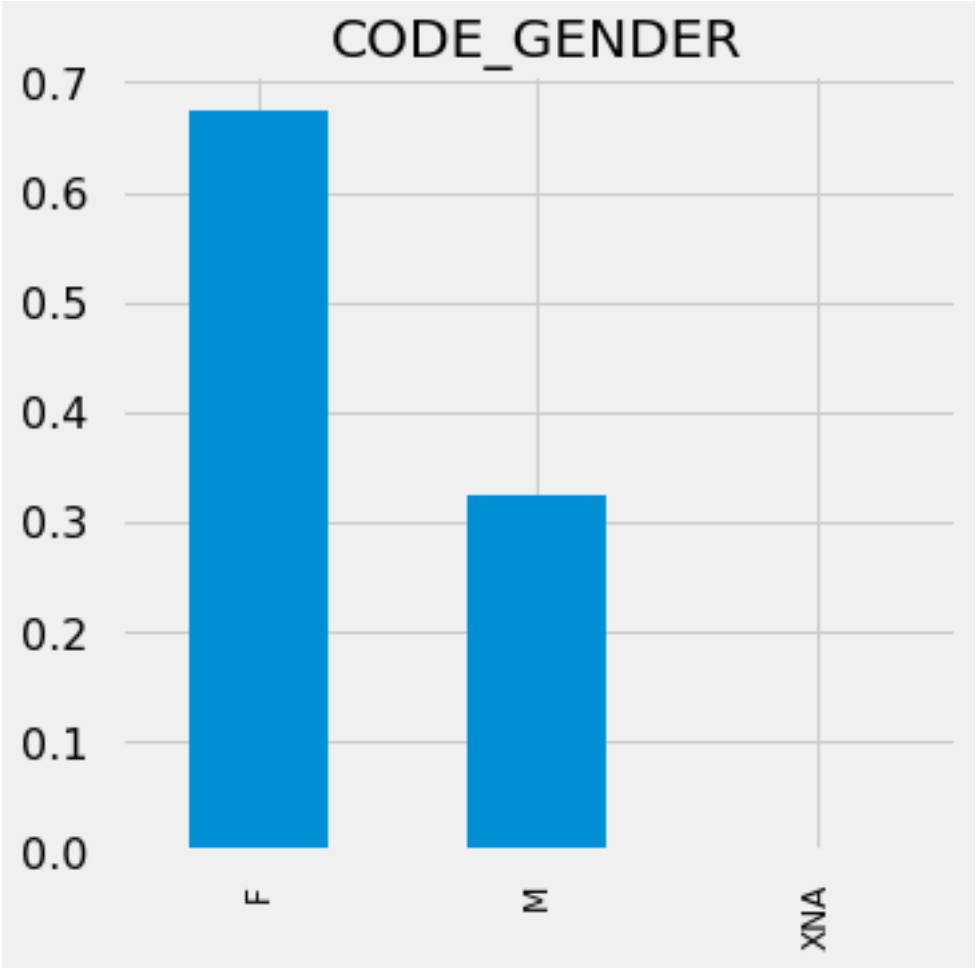
- The least number of applications are for the name contract type "Revolving loans" at 11 percent.



CODE_GENDER

Points to be concluded from the graph on the right.

- The majority of the applicants this year have been from the Female gender category and they have been given credit loans.
- Thus it is simply obvious that there is lesser risk in giving loans to Female category applicants than male category applicants who are at 32 percent(latter) in comparison to 67 percent of the former.

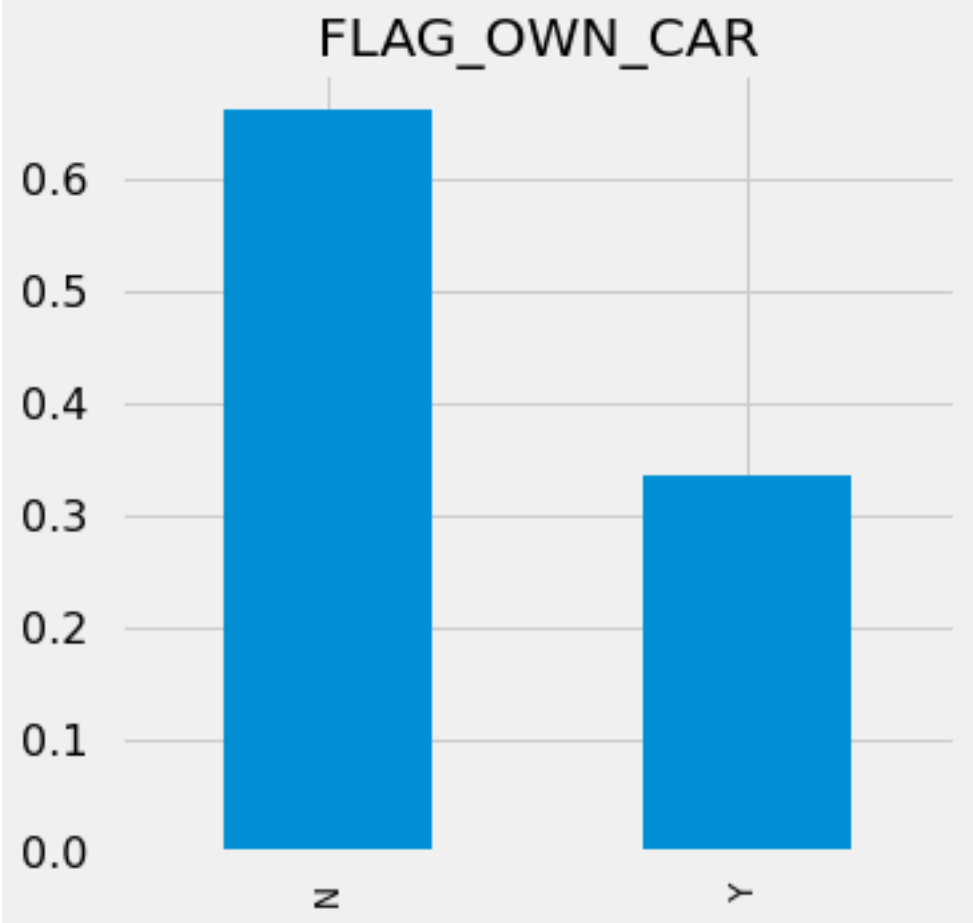


FLAG_OWN_CAR

Points to be concluded from the graph on the right.

- Total applicants which applied for loans and received credit, have a major portion as 'N' in terms of owning a car at 66 percent, in comparison to those who don't own a car at 33 percent.

- Thus, it seems that there is lesser risk at giving loans to those who don't own a car.

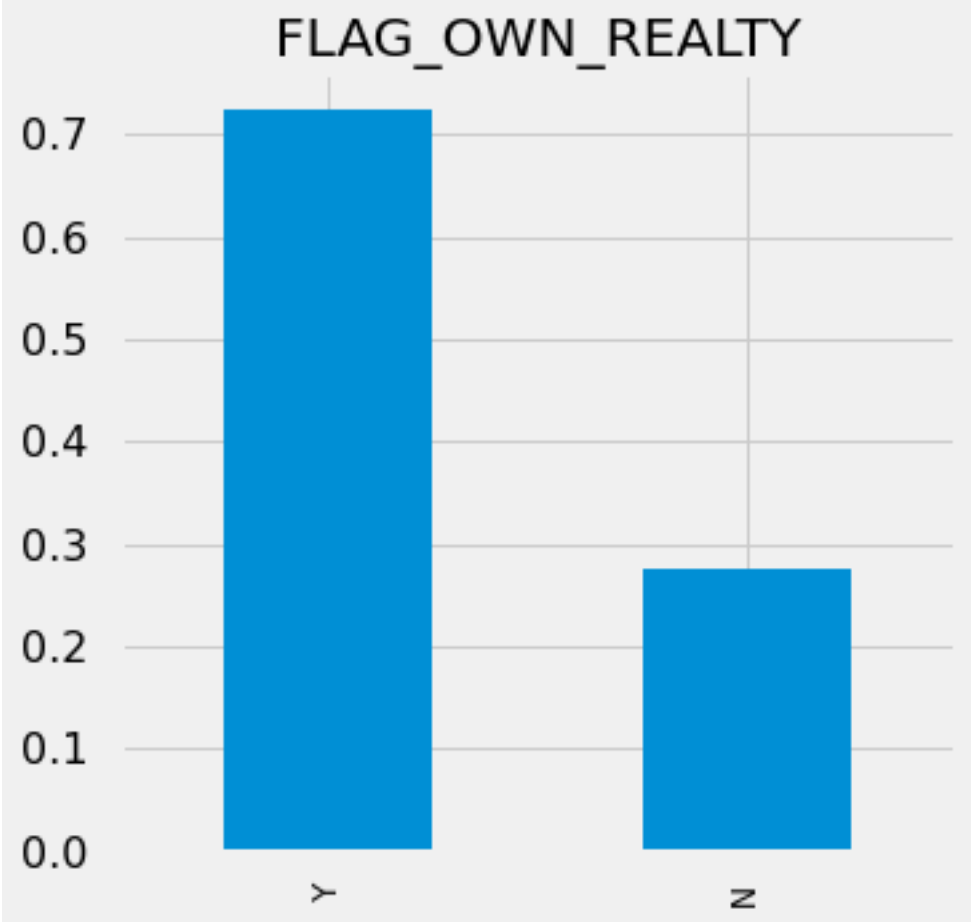


FLAG_OWN_REALTY

Points to be concluded from the graph on the right.

- There is a majority of those applicants who own a realty or real estate as 72 percent, in comparison to those applicants who don't own a house or flat at 27 percent.

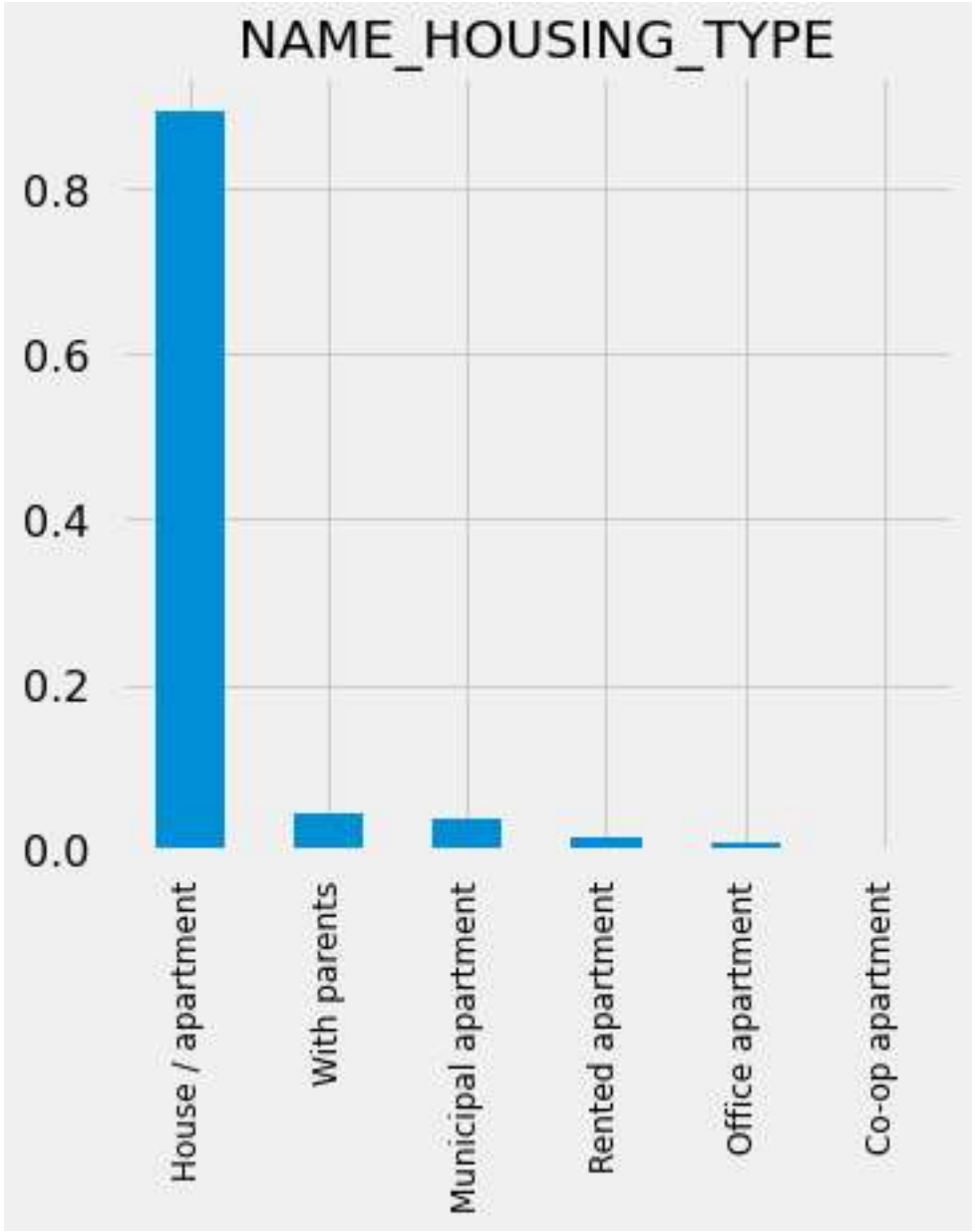
- Thus, there is lesser risk in giving loans to those who own a flat or house than those who don't.



NAME_HOUSING_TYPE

Points to be concluded from the graph on the right.

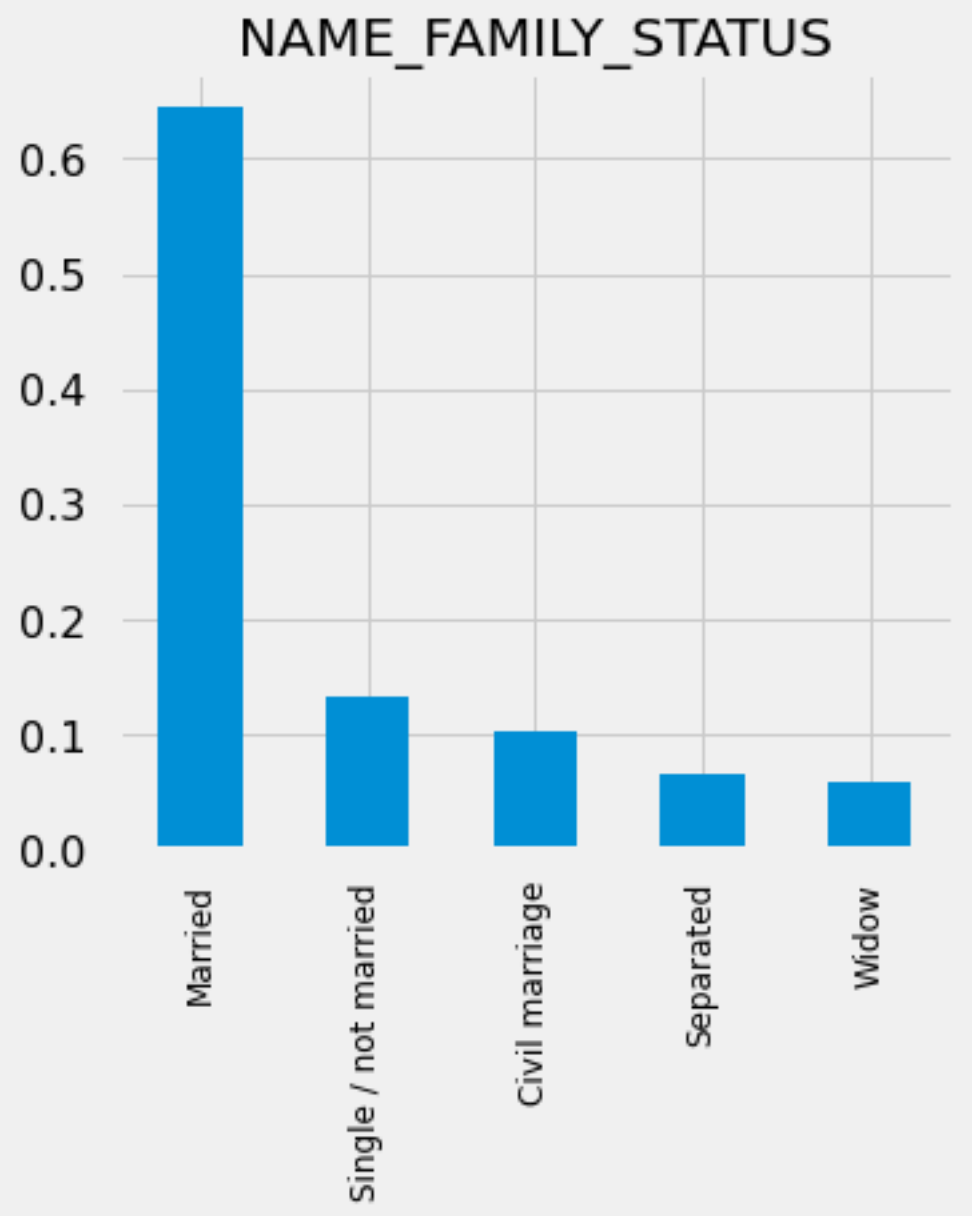
- Maximum number of loan applicant live in a house/apartment



NAME_FAMILY_STATUS

Points to be concluded from the graph on the right.

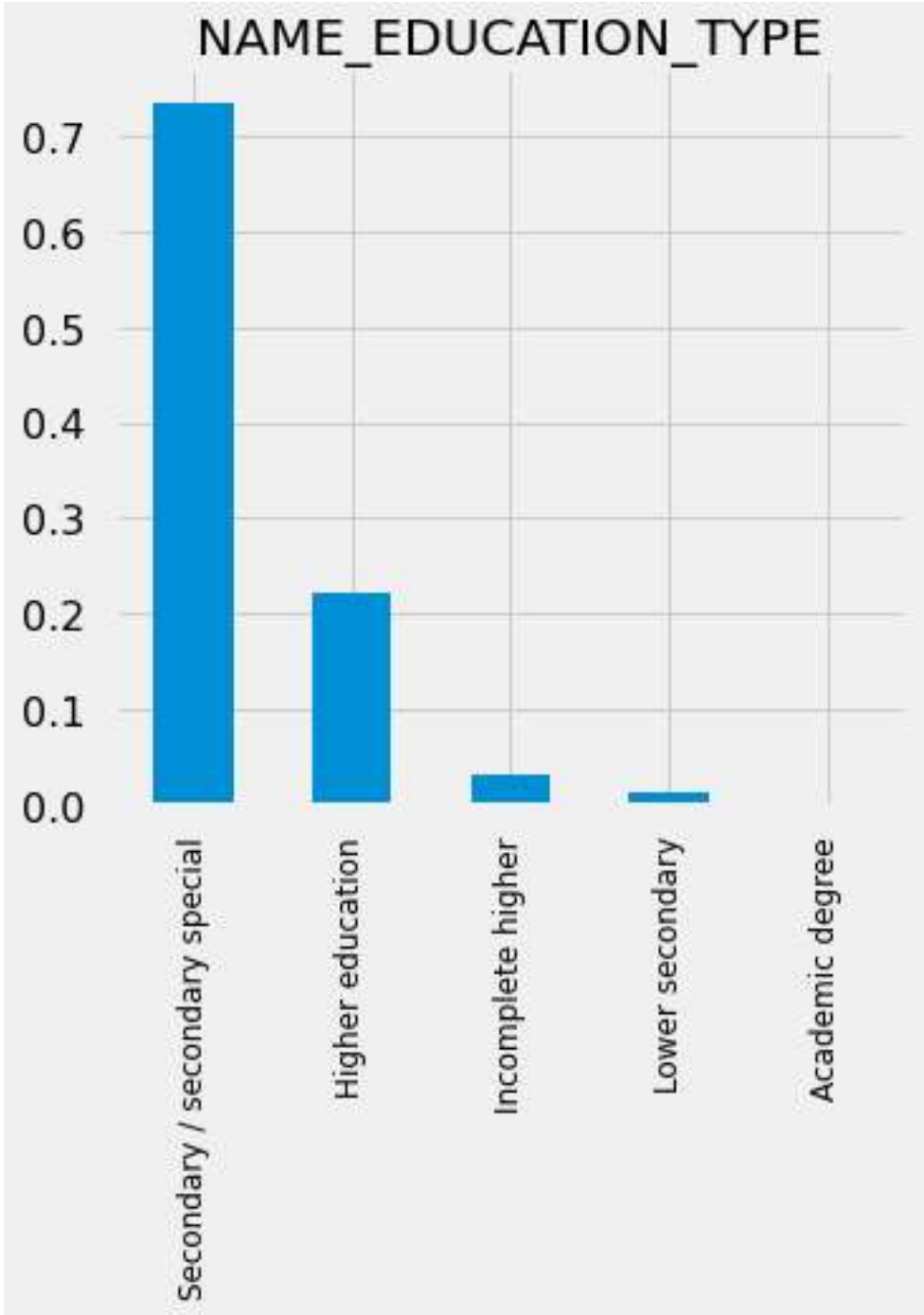
- Maximum number of loan applicant are Married in comparison to the least number of loan applicant over widowed



NAME_EDUCATION_TYPE

Points to be concluded from the graph on the right.

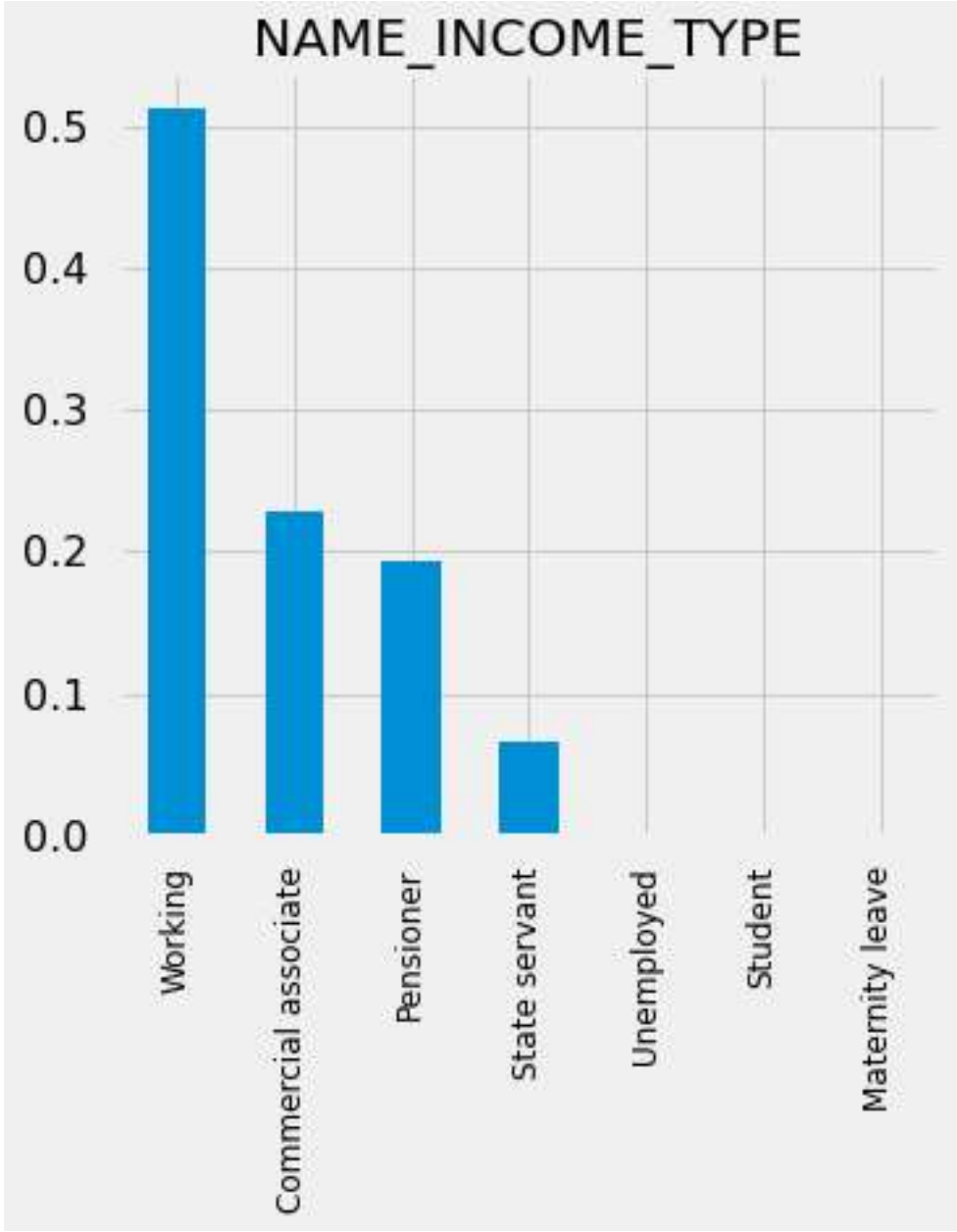
- The maximum number of loan applicant have completed their secondary/secondary special education.



NAME_HOUSING_TYPE

Points to be concluded from the graph on the right.

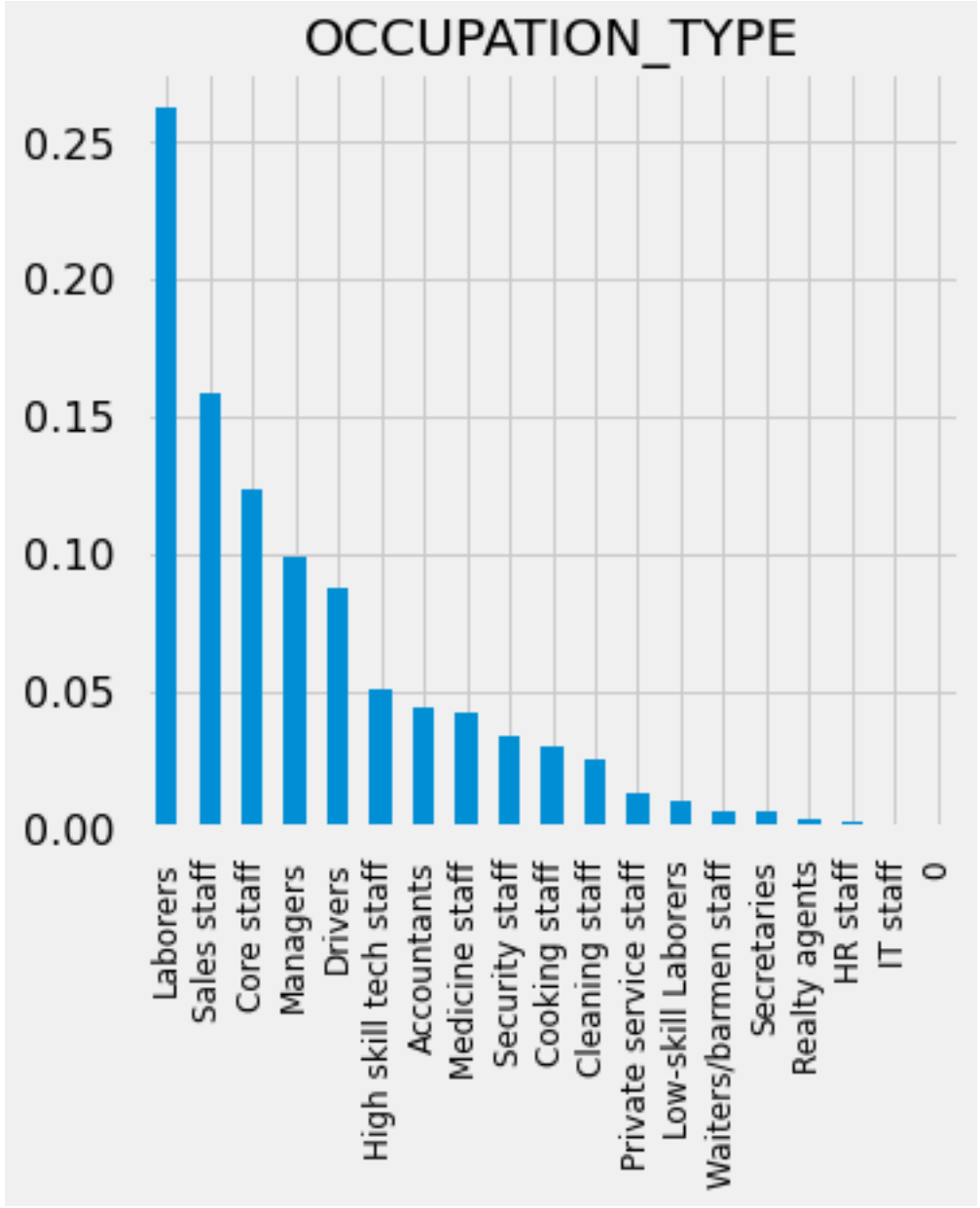
- The maximum number of loan applicants are from the working class next to which are commercial associate , pensioner and state servant.



OCCUPATION_TYPE

Points to be concluded from the graph on the right.

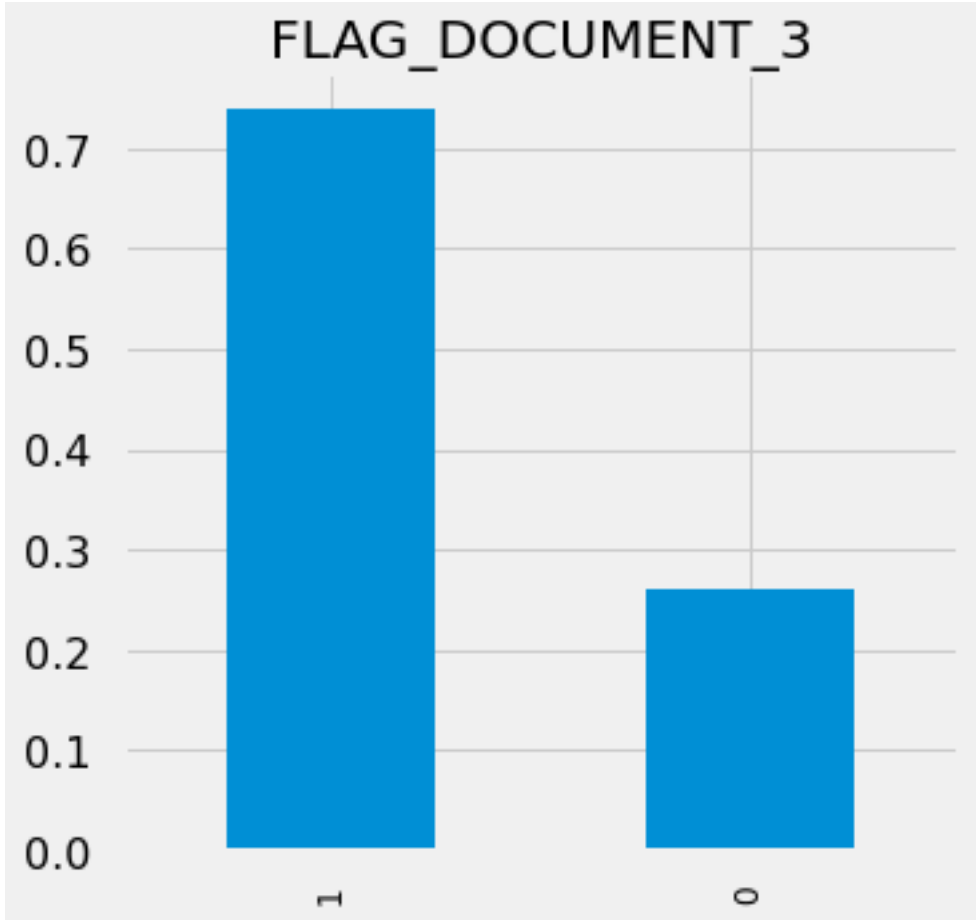
- The maximum number of loan applicants are laborers.



FLAG_DOCUMENT_3

Points to be concluded from the graph on the right.

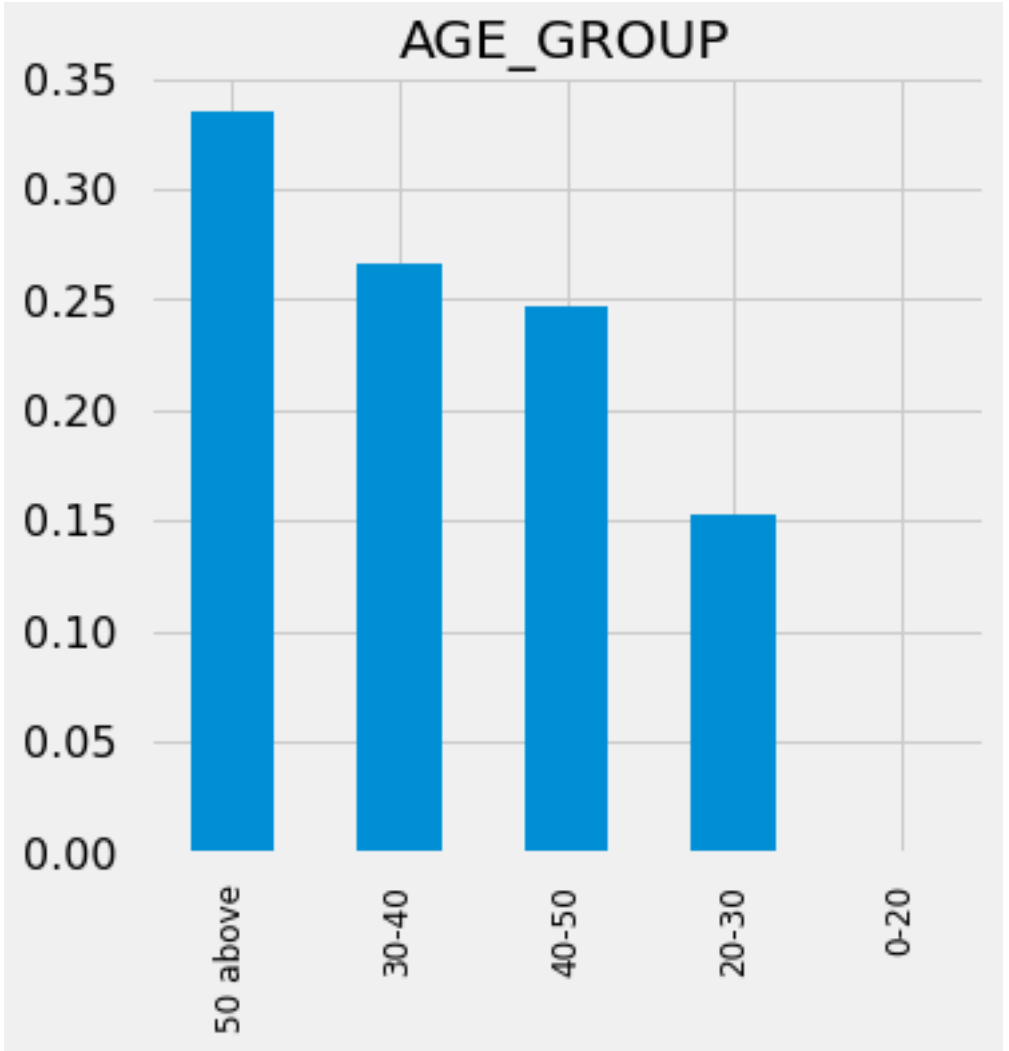
- maximum number of loan applications have submitted FLAG_DOCUMENT_3 and thus there loan applications have less risk factor associated with them in comparison to those applications where the value is 0 and no FLAG_DOCUMENT_3 has been submitted.



AGE_GROUP

Points to be concluded from the graph on the right.

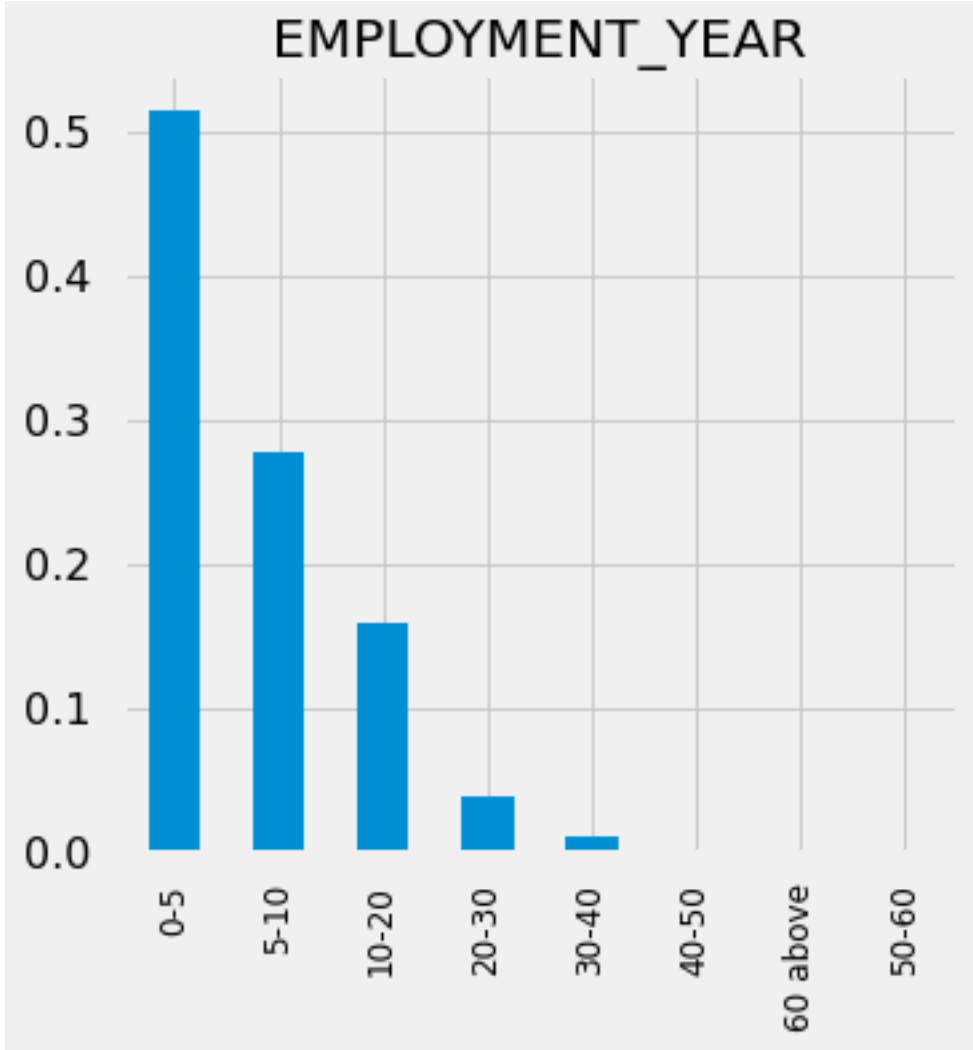
- The maximum number of loan applicants are laborers.



EMPLOYMENT_YEAR

Points to be concluded from the graph on the right.

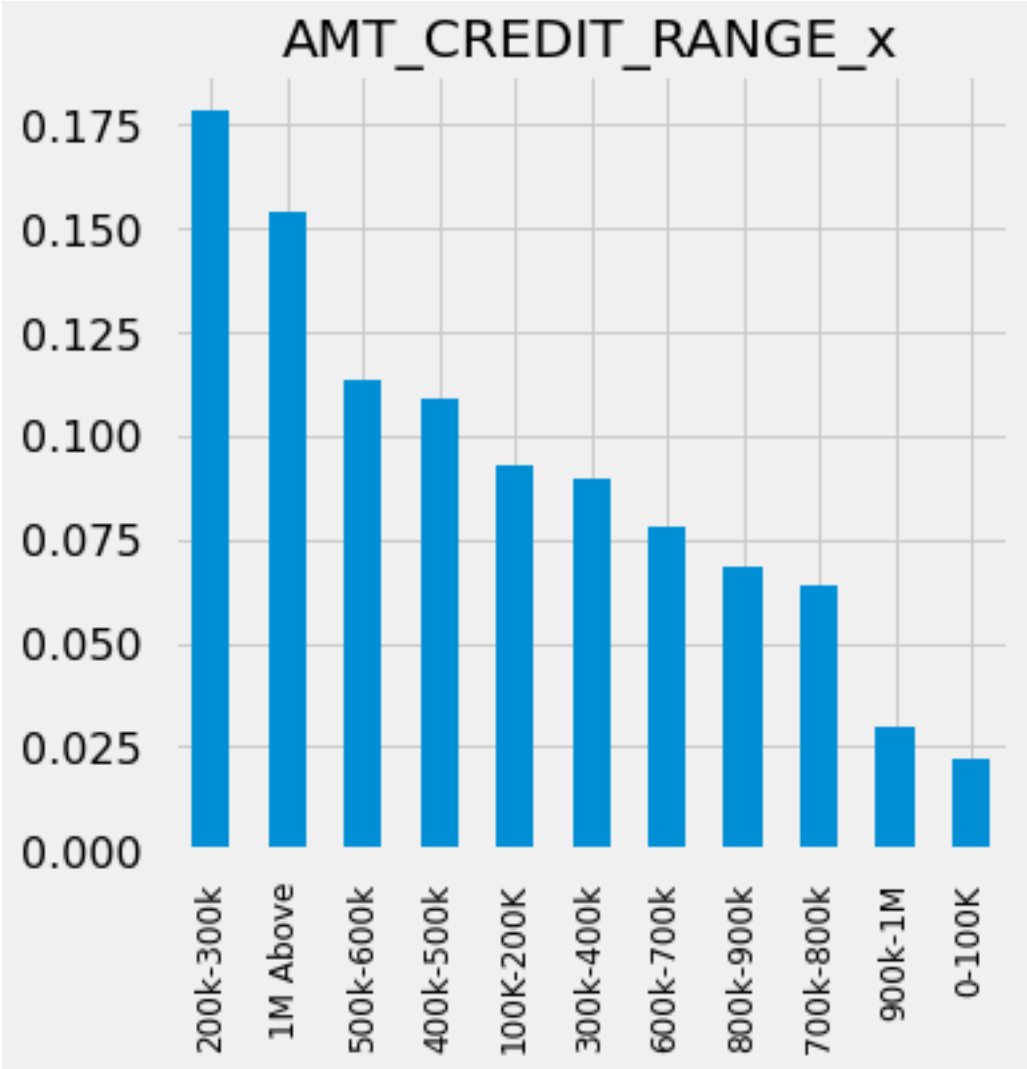
- The maximum number of loan applicants belong to the employment period of 0-5 years. As the employment years increases the number of loan applications submitted and approved decreases.
- This means that bank considers the low experienced more for approving loan applications.



AMT_CREDIT_RANGE_X

Points to be concluded from the graph on the right.

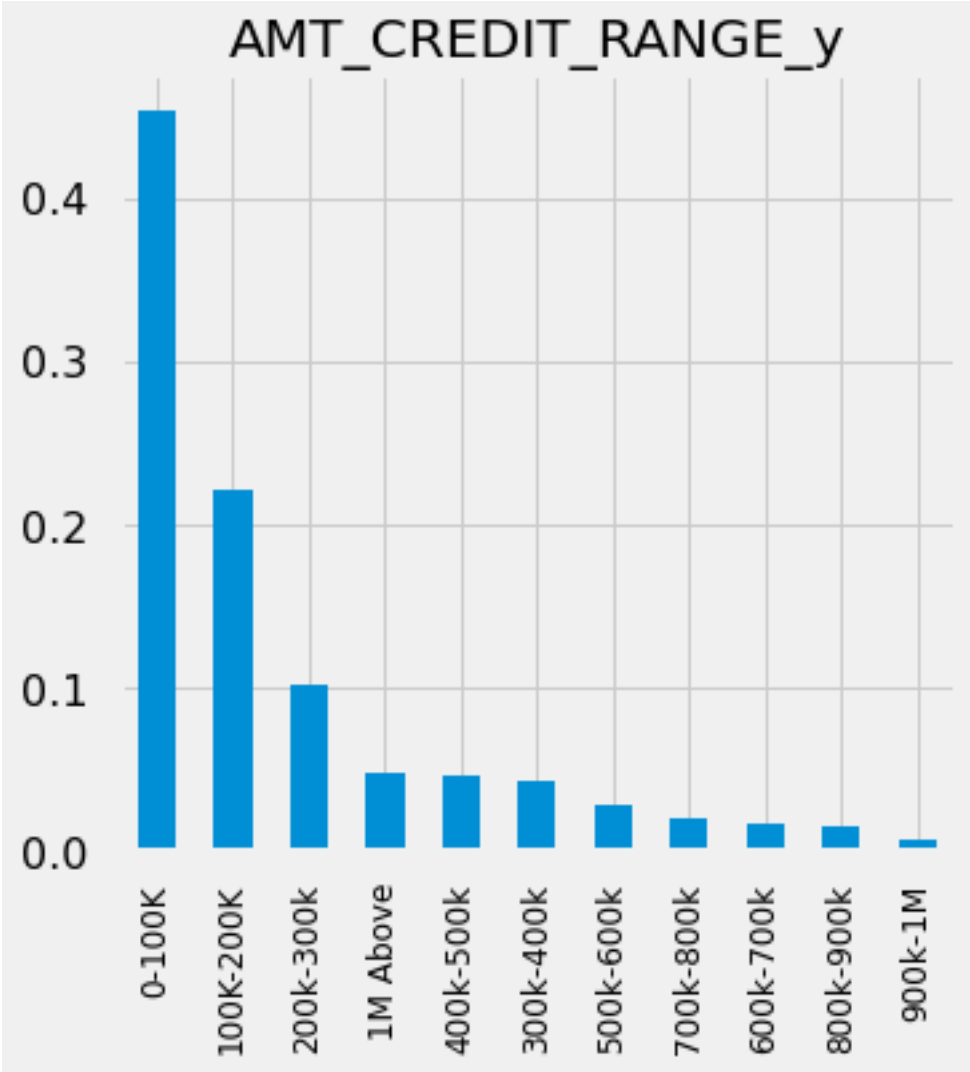
- The maximum amount of loan Sanctioned this Year belongs to the 200k-300k range.



AMT_CREDIT_RANGE_Y

Points to be concluded from the graph on the right.

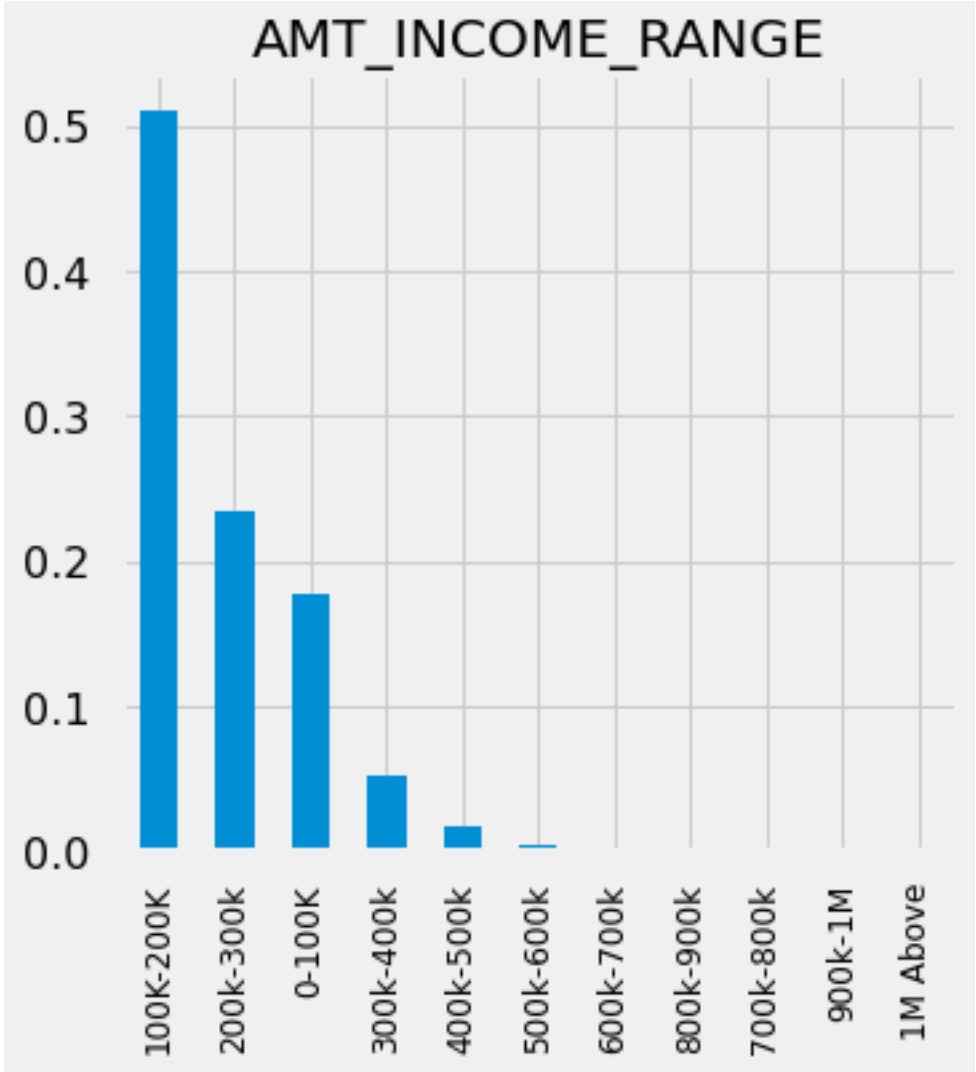
- The amount of loan sanctioned maximum number of times, last year, belongs to 0-100k range.
- This means that last year bank saw less risk associated with sanctioned of low amounts in credit than the current year.



AMT_INCOME_RANGE

Points to be concluded from the graph on the right.

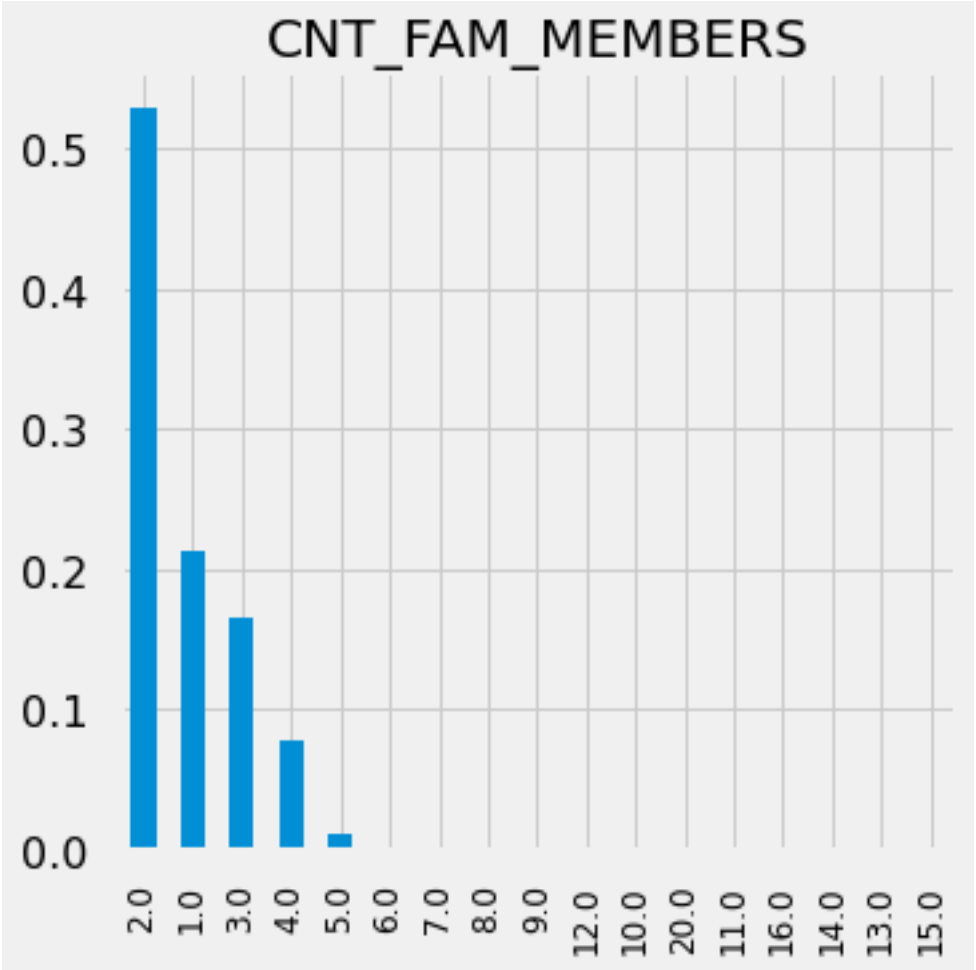
- The salary earned by maximum number of applicants this year lies in the range of 100k-200k.



CNT_FAM_MEMBERS

Points to be concluded from the graph on the right.

- The maximum number of loan applicants have a family that contains two members in total.



BIVARIATE ANALYSIS

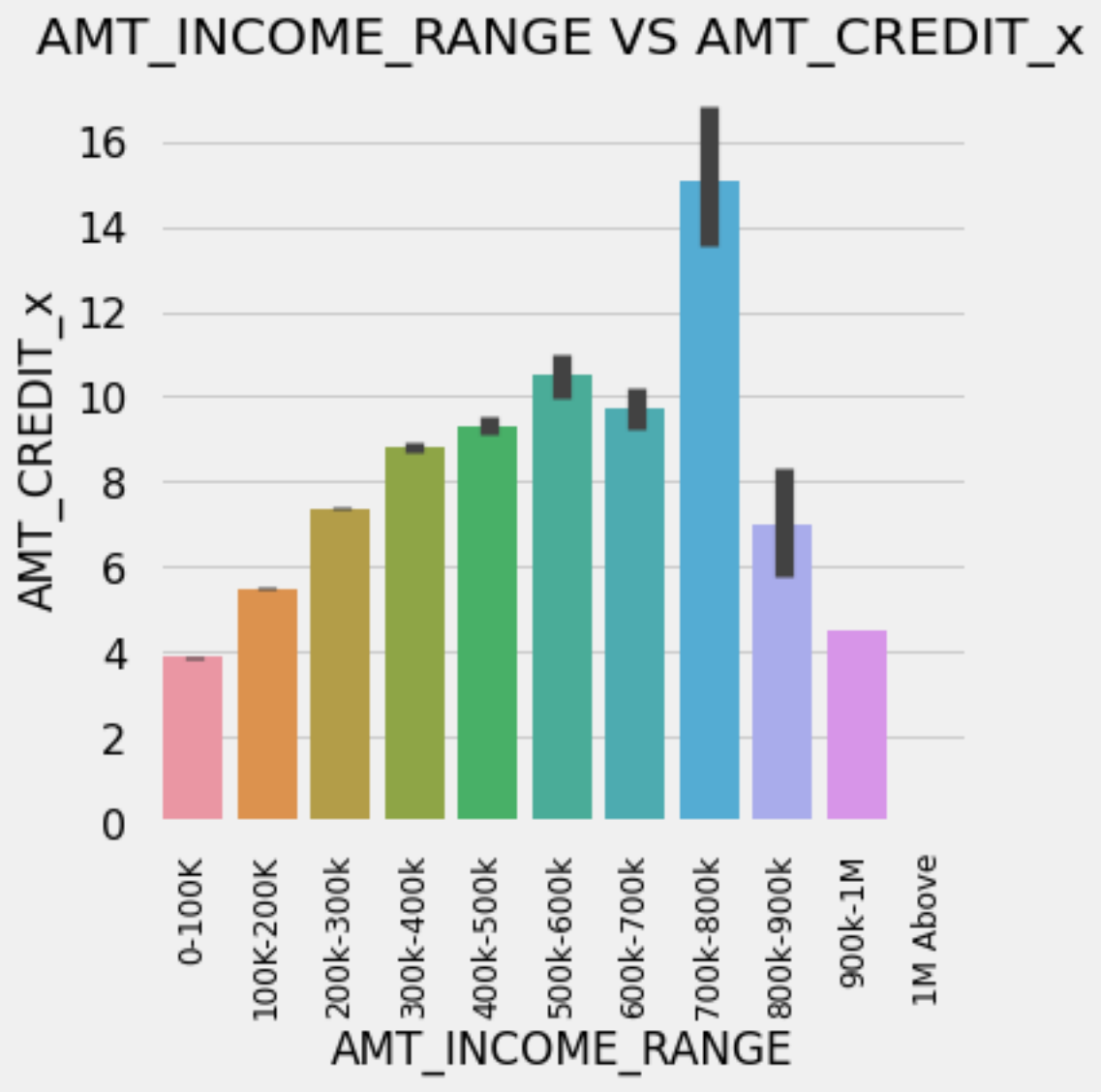
- This type of data involves **two different variables**. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.
- Example of bivariate data can be temperature and ice cream sales in summer season.

CATEGORICAL BIVARIATE ANALYSIS

1.AMT_INCOME_RANGE VS AMT_CREDIT¶

Points to be concluded from the graph on the right.

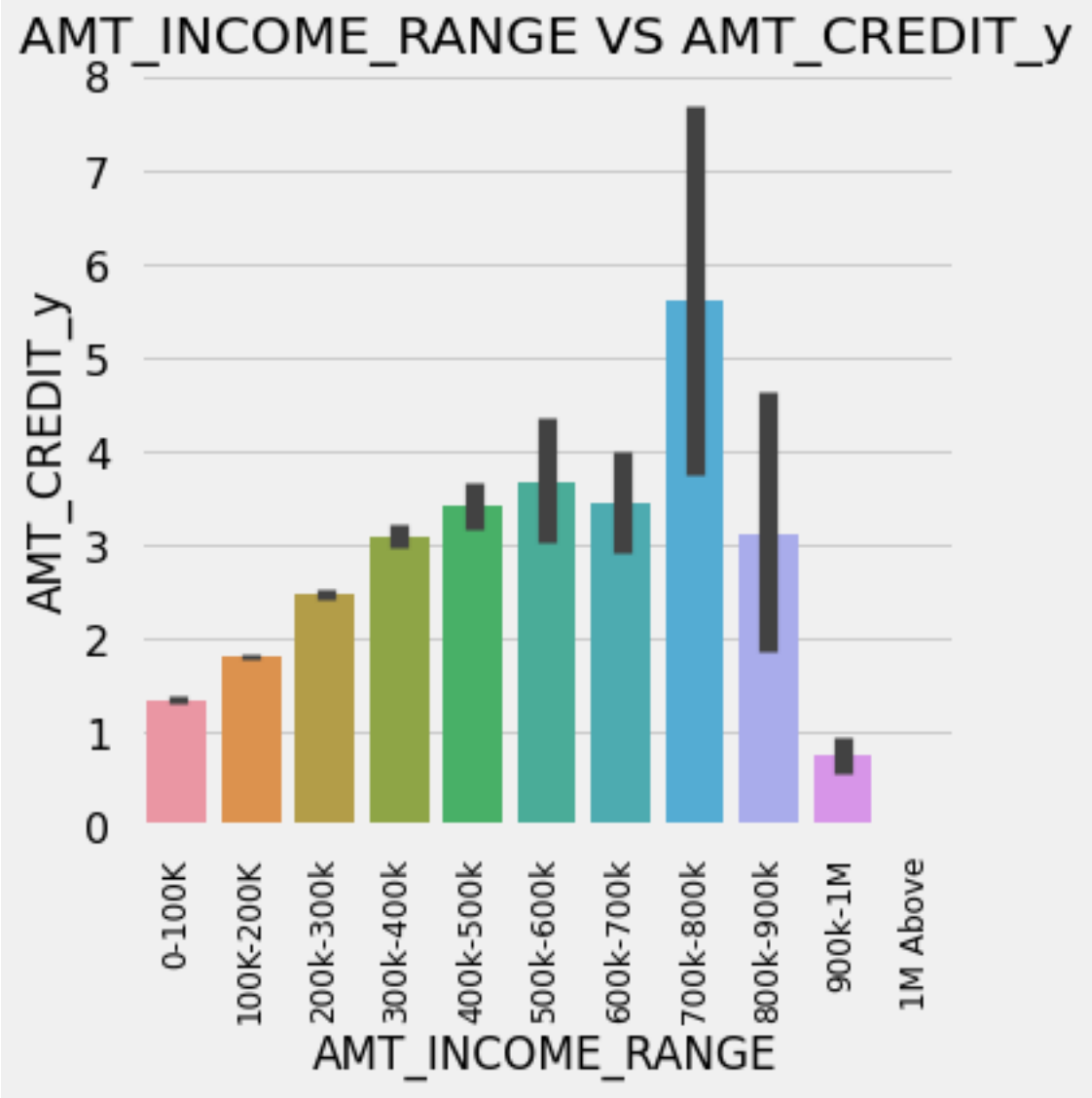
- The income range seems to be increasing in a linear fashion with respect to credit amount.
- So, as the income increases the amount credit also increases.
- From 0 to 800k income increases so credit increases as well and then it starts dropping for this year



AMT_INCOME_RANGE VS AMT_CREDIT_Y

Points to be concluded from the graph on the right.

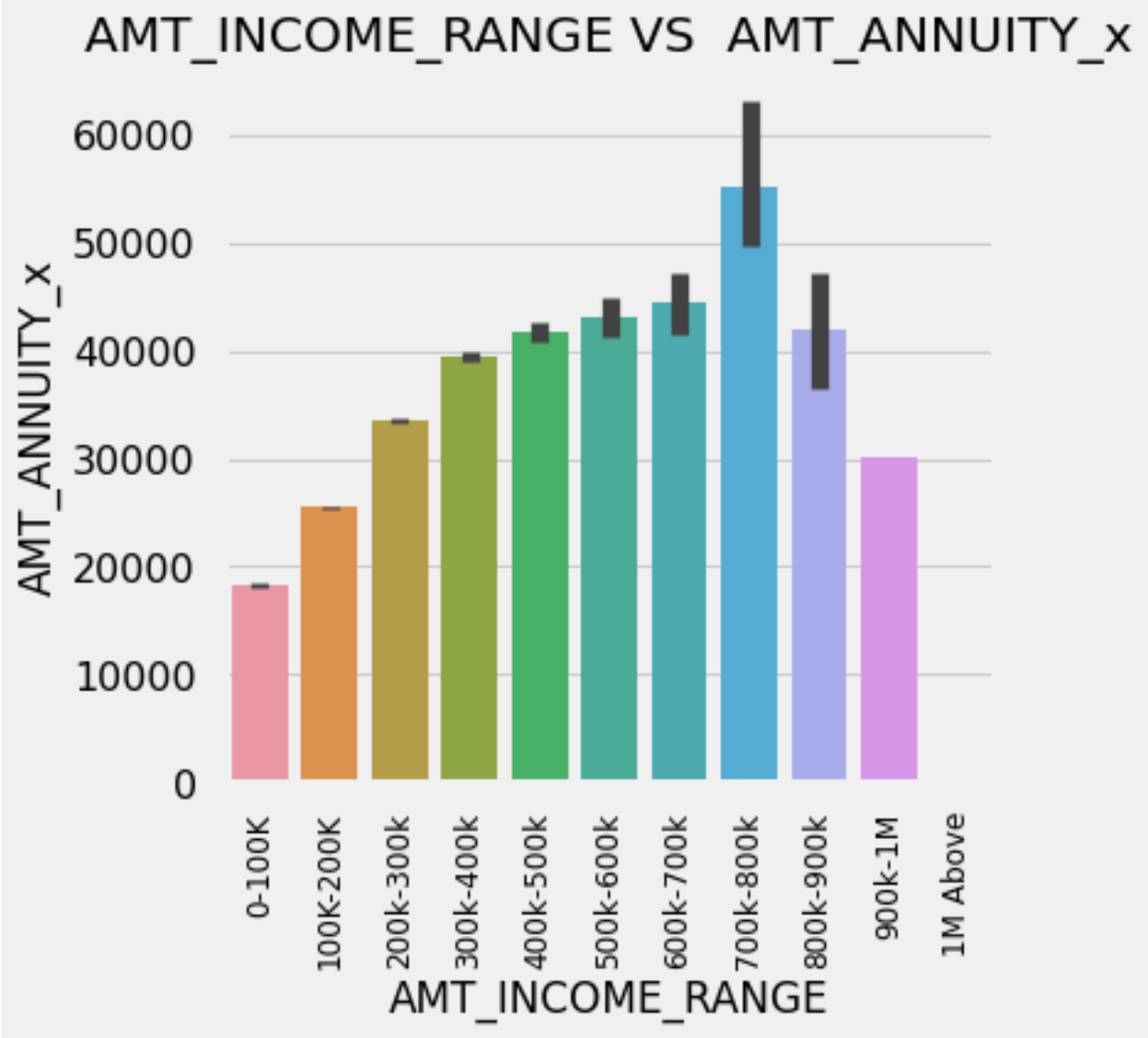
- The credit amount increases in a linear fashion as income increases for the previous year.
- The highest proportional relationship is at 700k to 800k income range.



AMT_INCOME_RANGE VS AMT_ANNUITY_X

Points to be concluded from the graph on the right.

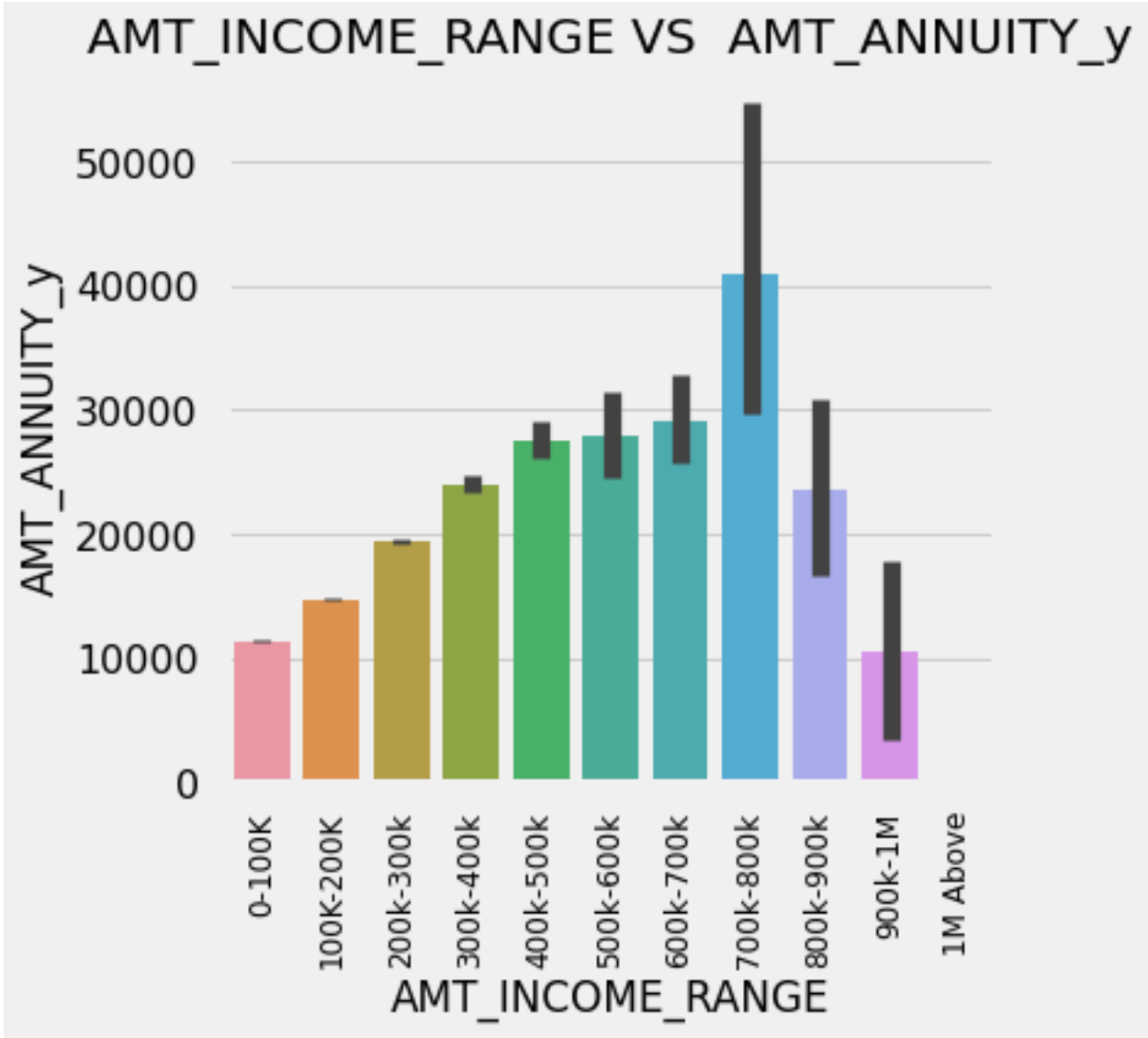
- The annuity increases linearly up to 800k income since it depends on income range and after that it starts dropping for this year.



AMT_INCOME_RANGE VS AMT_ANNUITY_Y

Points to be concluded from the graph on the right.

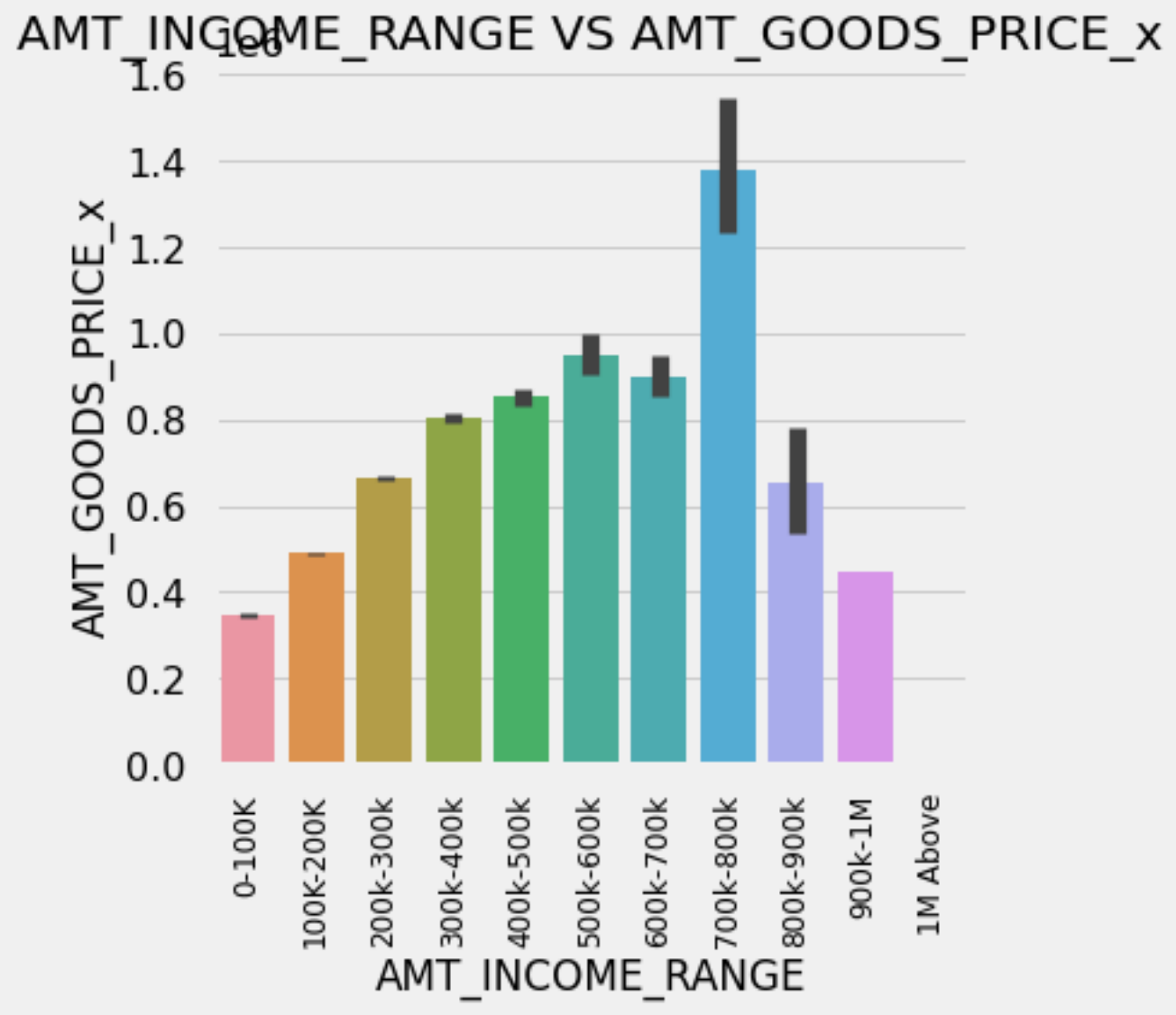
- The annuity increases linearly upto 800k income since it depends on income range and after that it starts dropping for the previous year.



AMT_INCOME_RANGE VS AMT_GOODS_PRICE

Points to be concluded from the graph on the right.

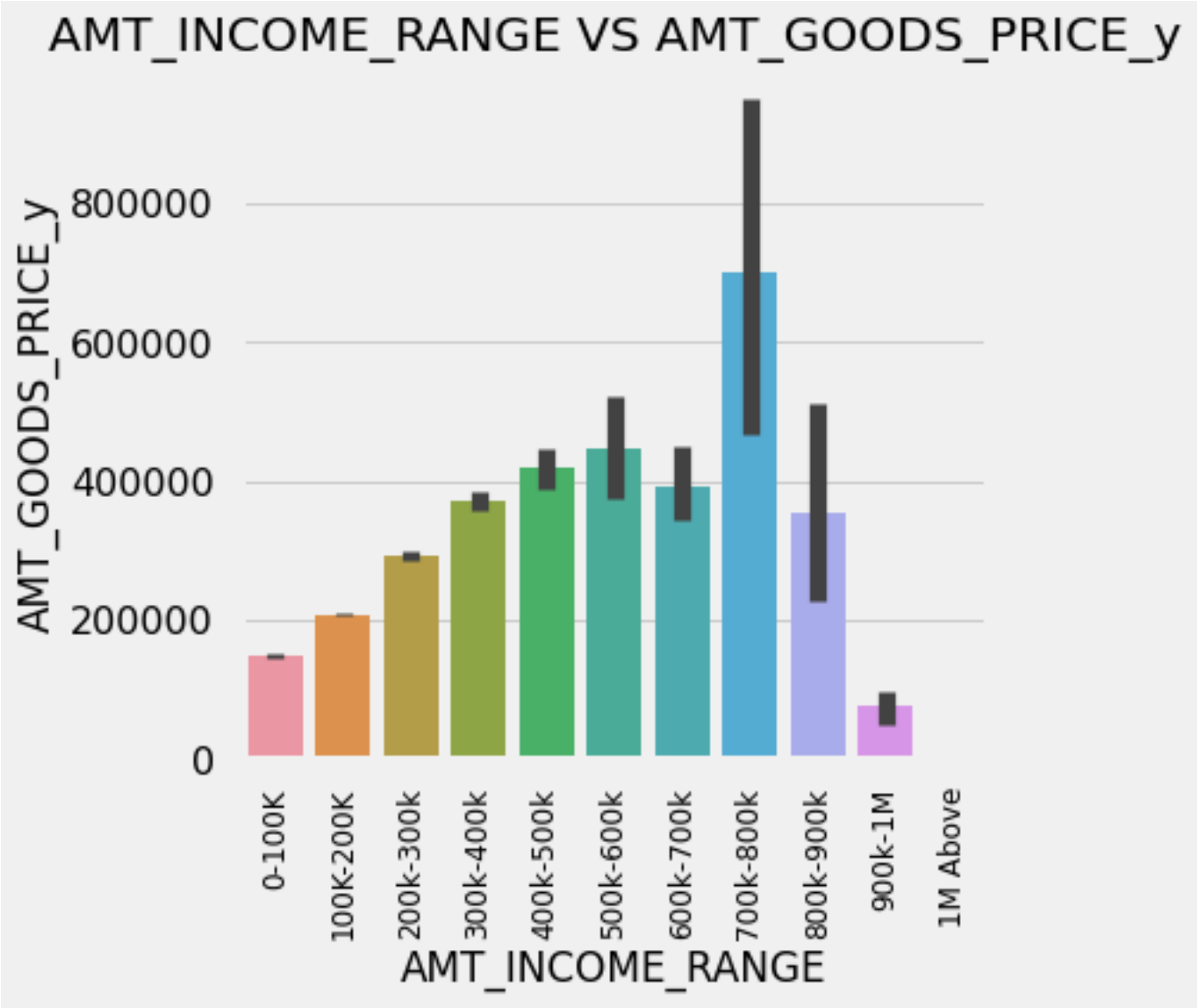
- The goods price increases in the linear fashion upto 800k since it depends on income range and after that it starts dropping for this year



AMT_INCOME_RANGE VS AMT_GOODS_PRICE_Y

Points to be concluded from the graph on the right.

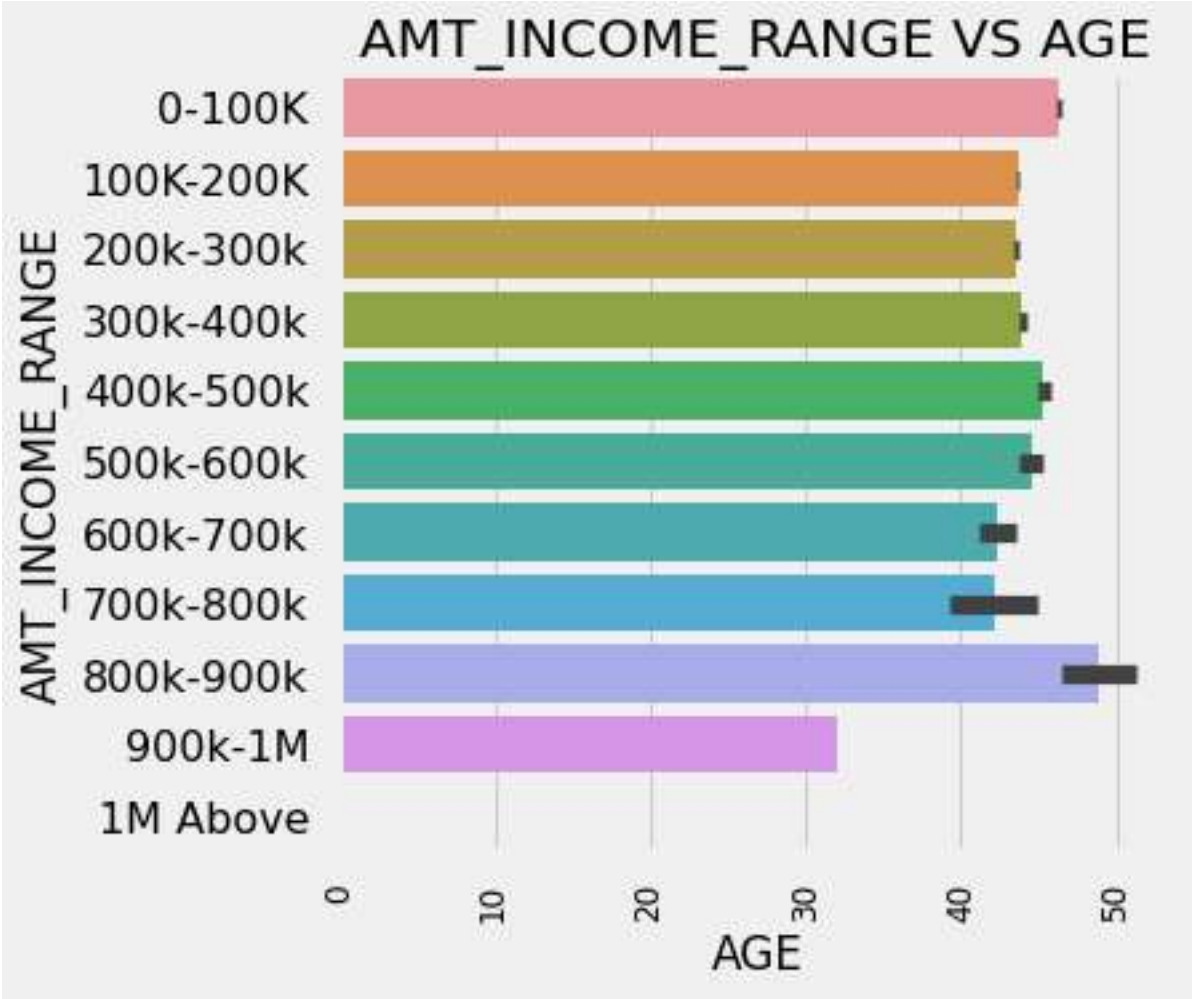
- The goods price increases in the linear fashion upto 800k since it depends on income range and after that it starts dropping for the previous year.



AMT_INCOME_RANGE VS AGE

Points to be concluded from the graph on the right.

- The amount of income earned doesn't depend upon the age of the applicants for this year's applications.

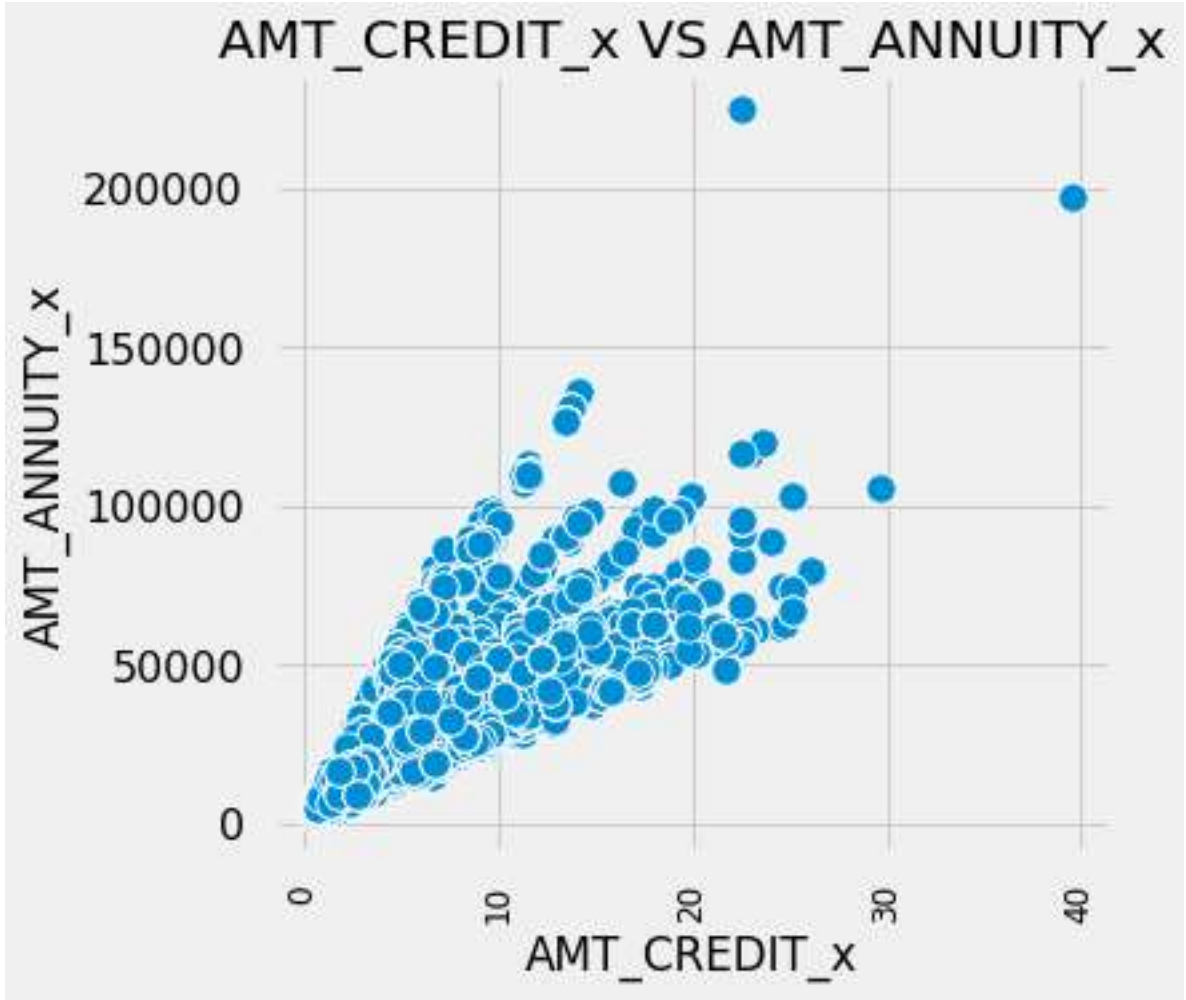


AMT_CREDIT VS AMT_ANNUITY

AMT_CREDIT_X VS AMT_ANNUITY_X

Points to be concluded from the graph on the right.

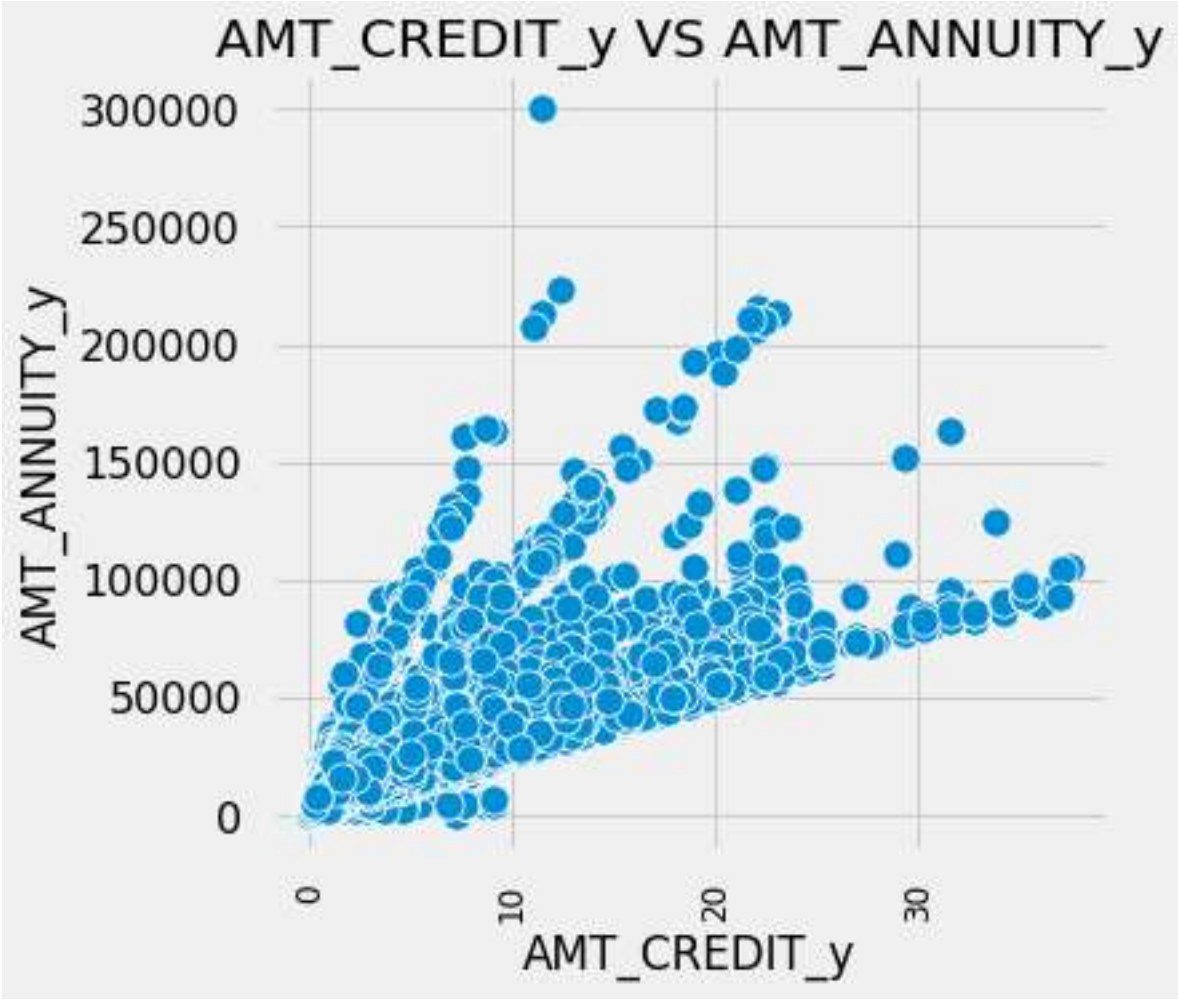
•. The amount annuity is directly proportional to the amount credit for this year



AMT_CREDIT_Y VS AMT_ANNUITY_Y

Points to be concluded from the graph on the right.

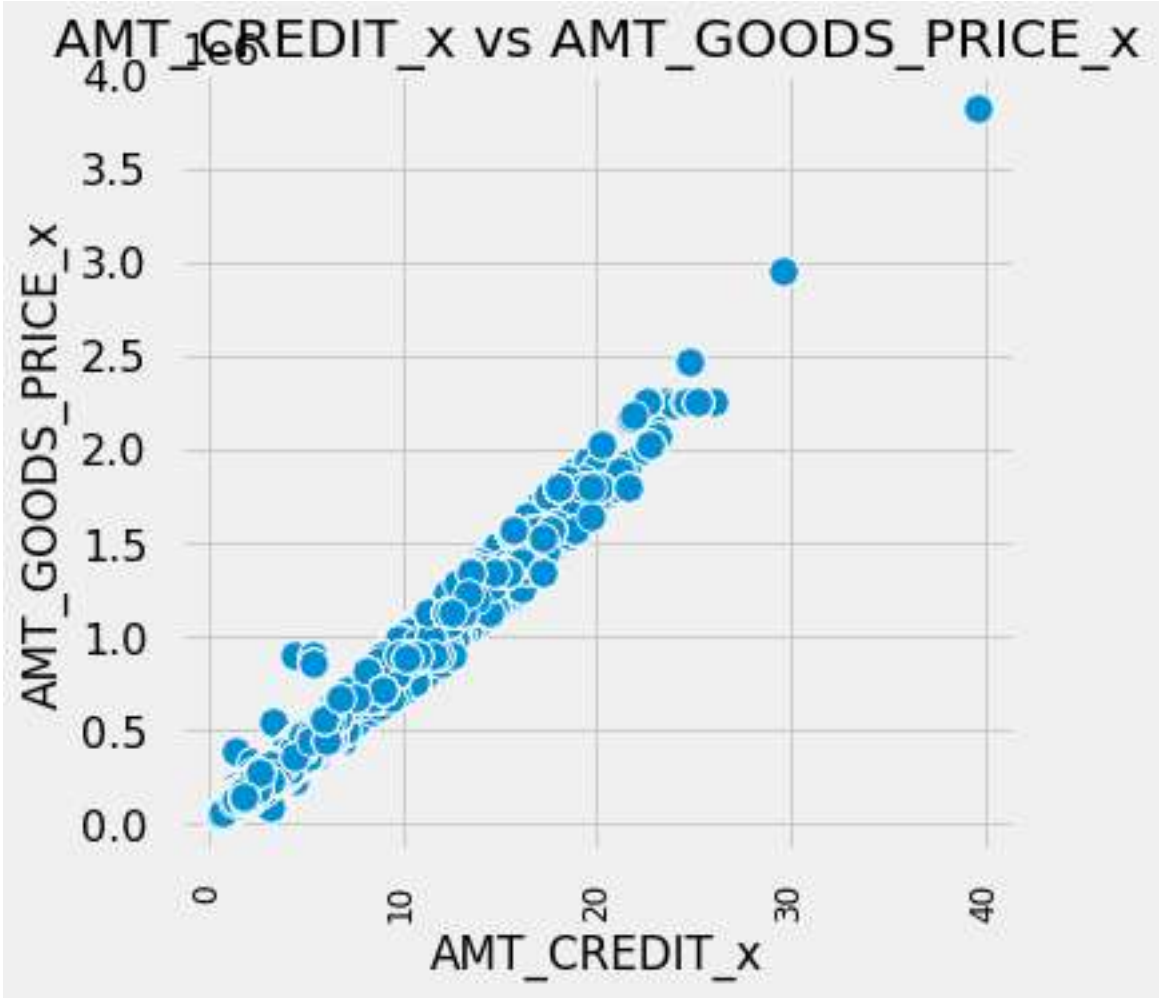
•. The annuity amount is directly proportional to the credit amount for the previous year's applications.



AMT_CREDIT_X VS AMT_GOODS_PRICE_X

Points to be concluded from the graph on the right.

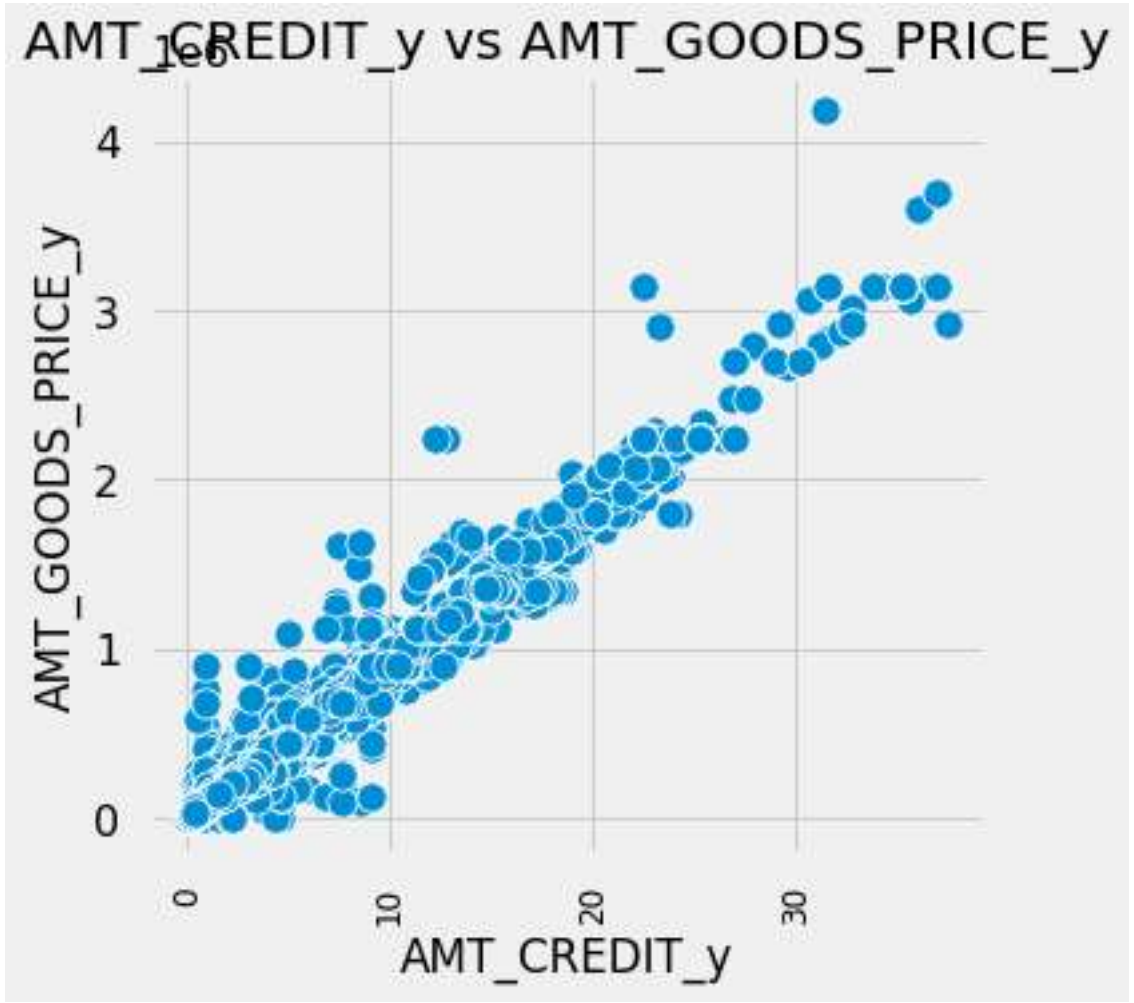
•. The goods price amount is directly proportional to the credit amount for the current year's applications.



AMT_CREDIT_Y VS AMT_GOODS_PRICE_Y

Points to be concluded from the graph on the right.

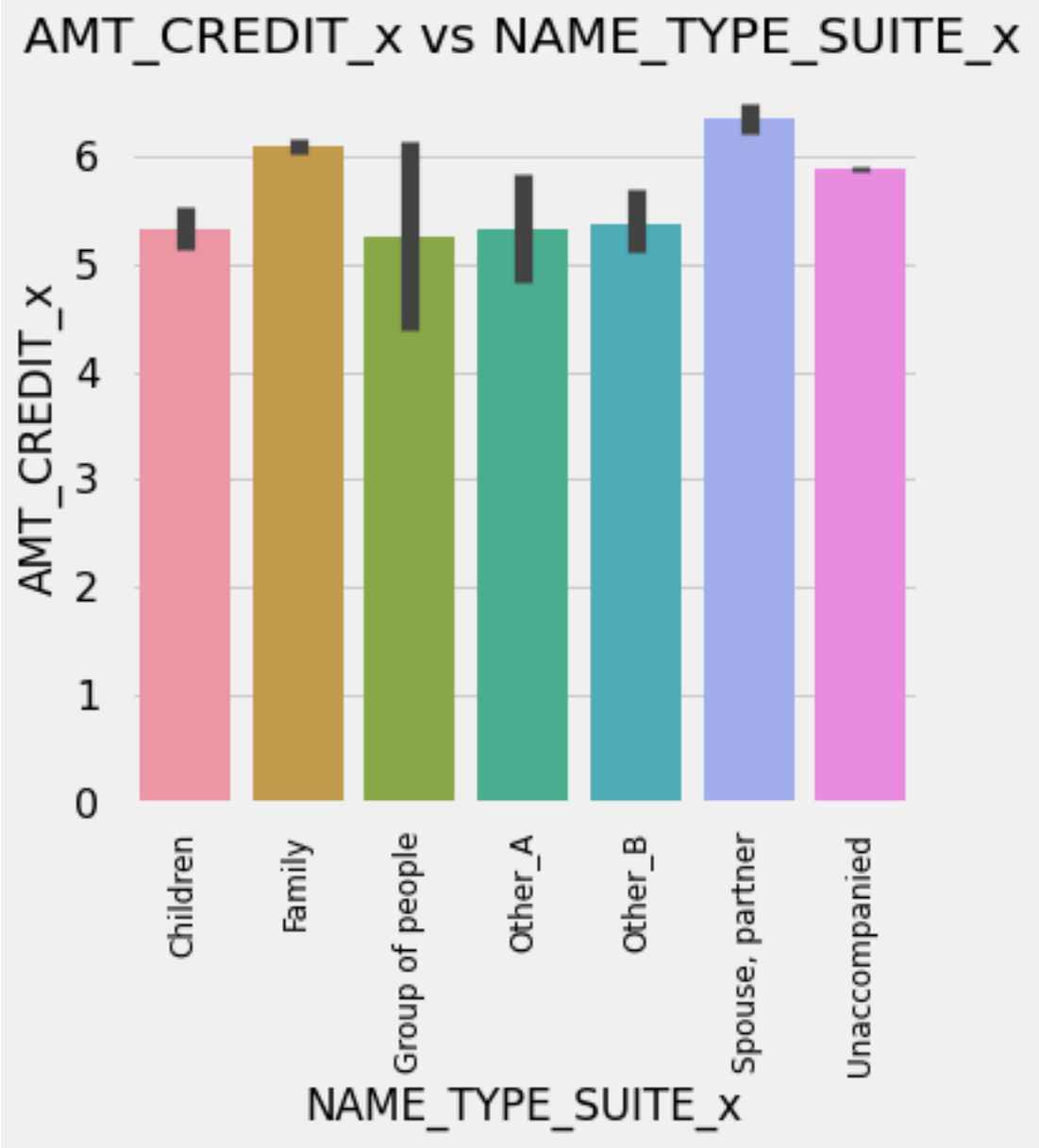
•. The goods price amount is directly proportional to the credit amount for the previous year's applications.



AMT_CREDIT_X VS NAME_TYPE_SUITE_X

Points to be concluded from the graph on the right.

•. The maximum amount of credit was passed for the incoming group which contained Spouse,partner after which comes the family as the next best credit loan group for the current year.

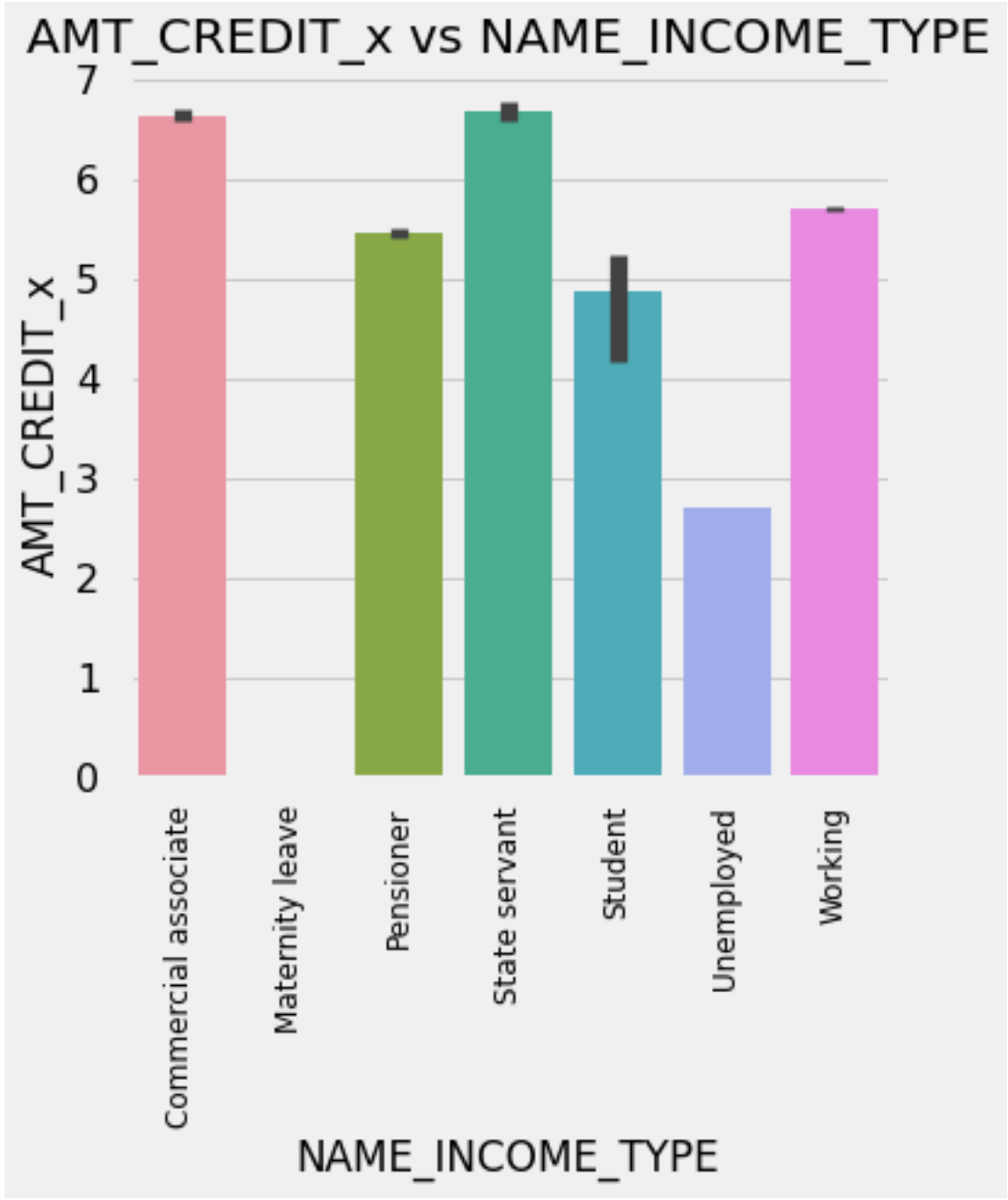


AMT_CREDIT VS NAME_INCOME_TYPE

AMT_CREDIT_X VS NAME_INCOME_TYPE

Points to be concluded from the graph on the right.

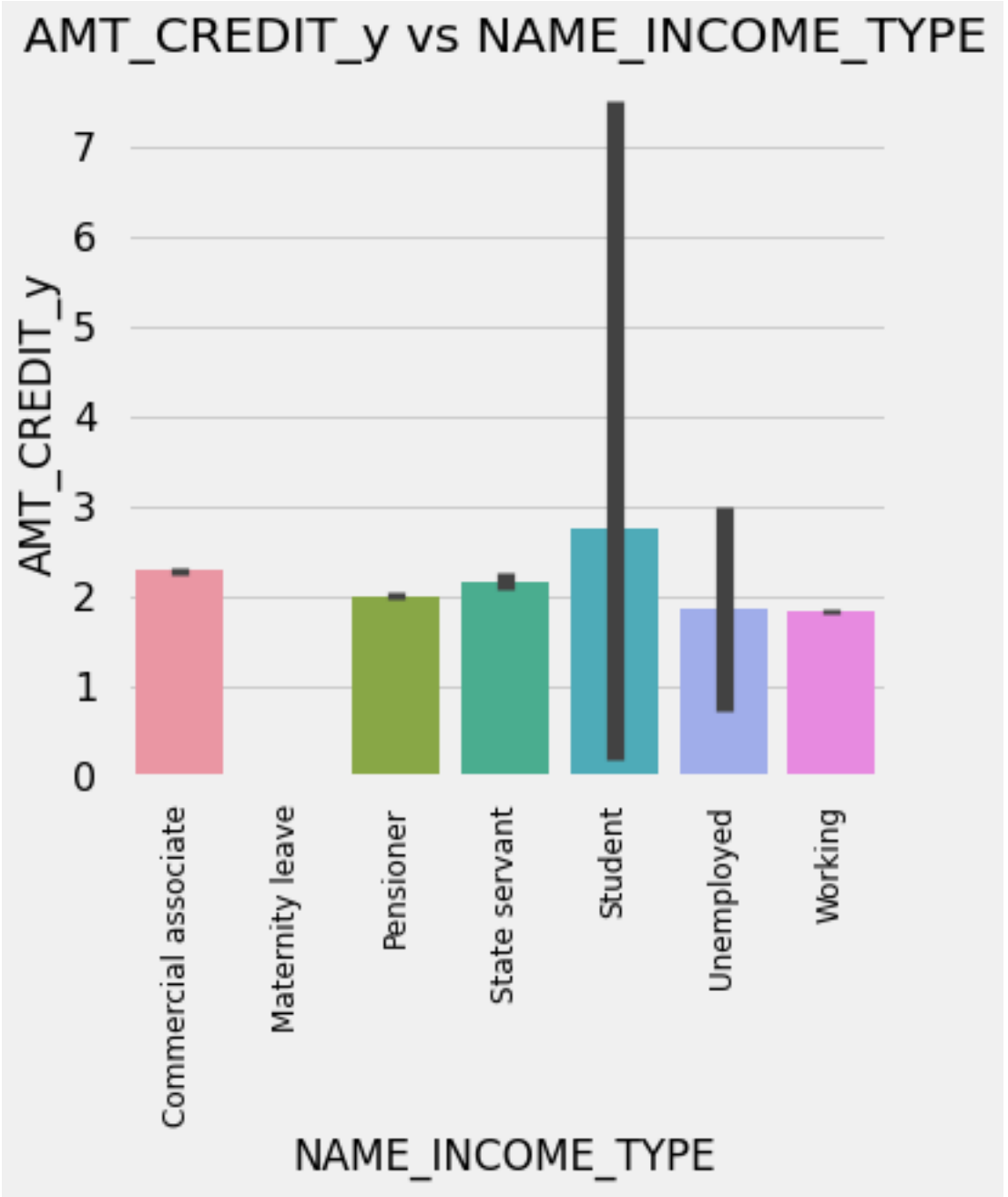
•. The maximum loan amount was credited to Commercial associate and state servant in comparison to other income types for current year.



AMT_CREDIT_Y VS NAME_INCOME_TYPE

Points to be concluded from the graph on the right.

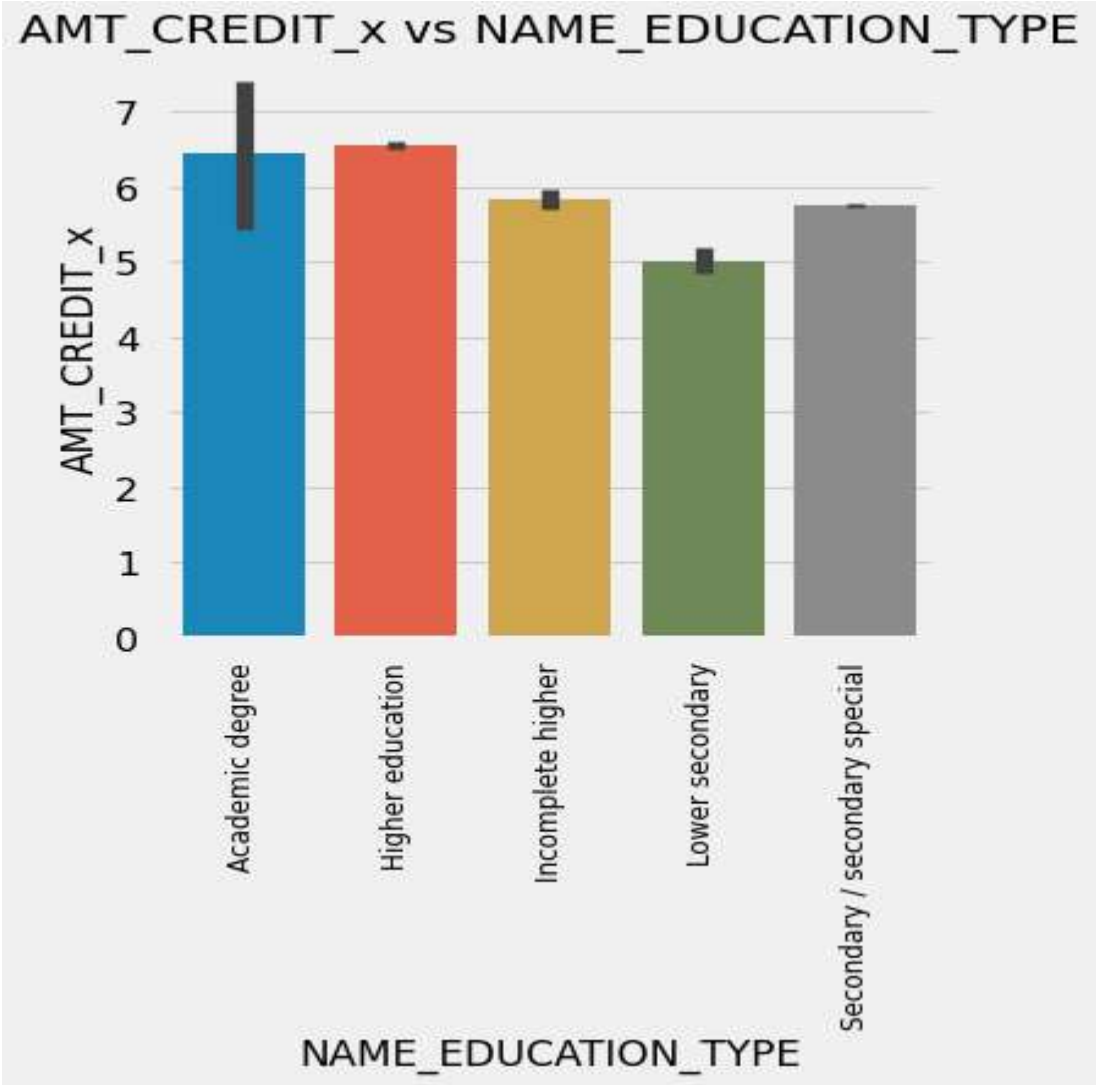
•. The maximum loan amount was passed for Student income type in comparison to others in the previous year.



AMT_CREDIT_X VS NAME_EDUCATION_TYPE

Points to be concluded from the graph on the right.

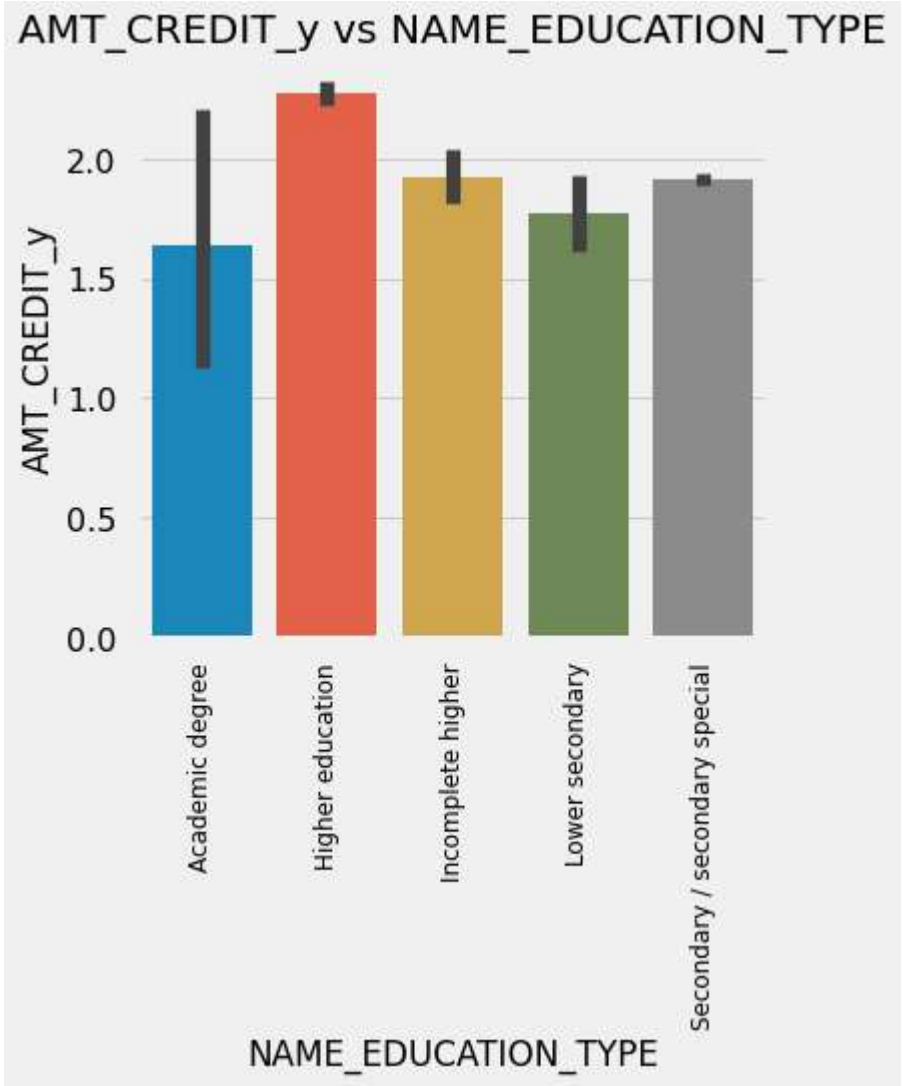
•. The maximum amount of credit was sanctioned for applicants with higher education degree after which the next best education type was Academic degree for the current year applications



AMT_CREDIT_Y VS NAME_EDUCATION_TYPE

Points to be concluded from the graph on the right.

•. The maximum amount of credit was sanctioned for applicants with higher education degree after which the next best education types were Incomplete higher and Secondary degree for the previous year applications.

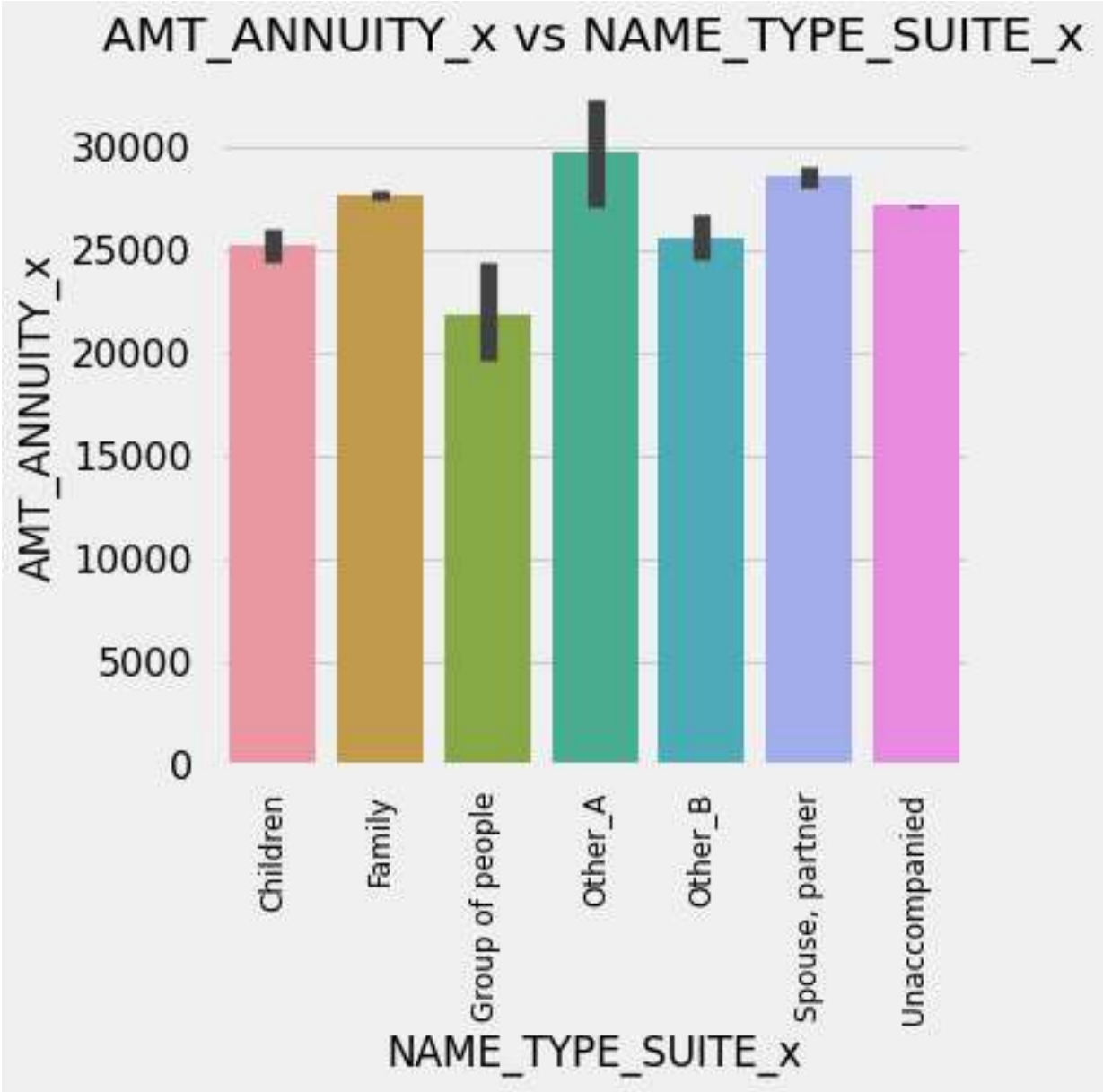


AMT_ANNUITY_X VS NAME_TYPE_SUITE_X

Points to be concluded from the graph on the right.

•. The maximum amount of annuity debits from those applicants who came with Other.

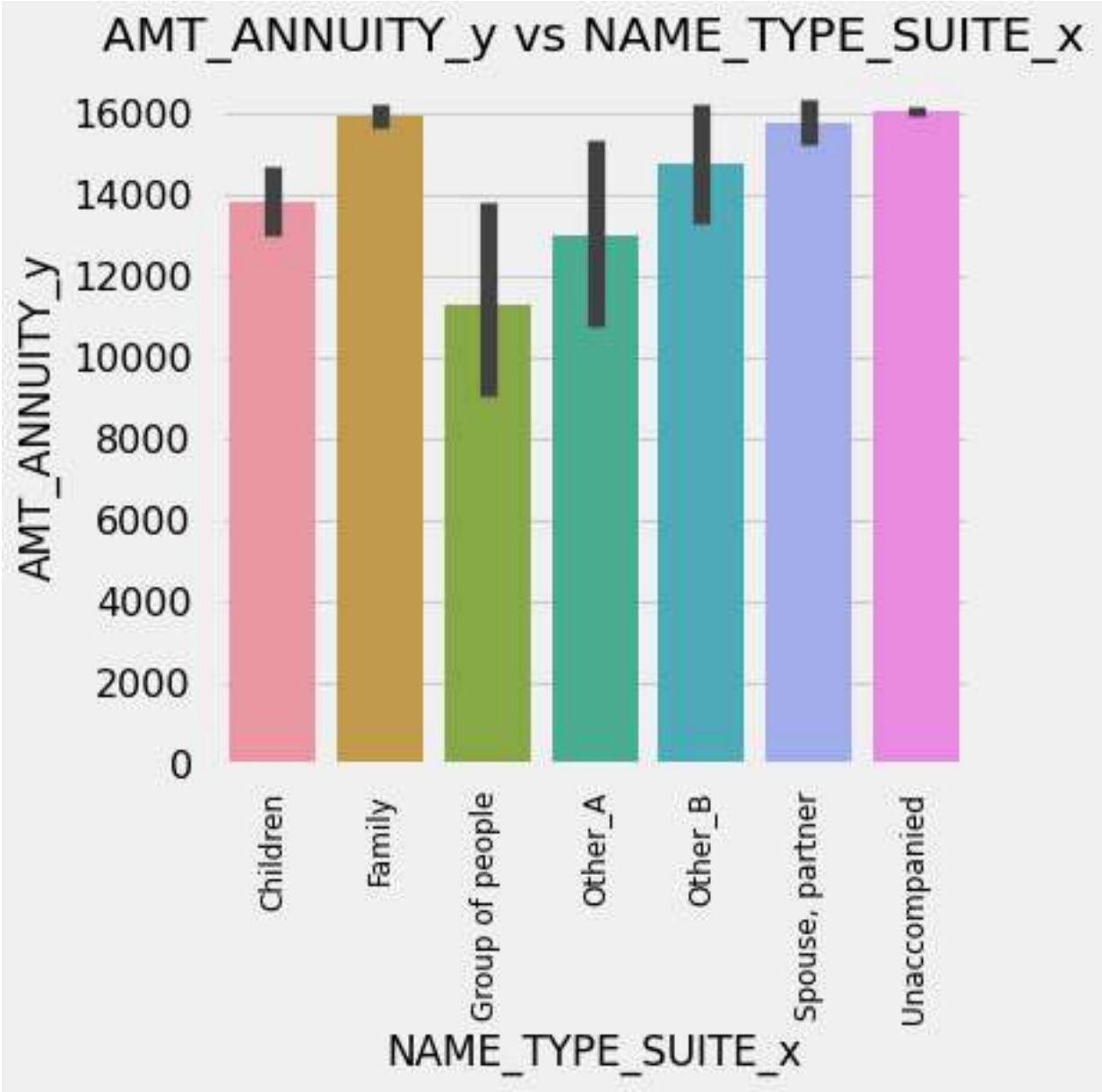
•A type of people at the time of loan application for the current year.



AMT_ANNUITY_Y VS NAME_EDUCATION_TYPE

Points to be concluded from the graph on the right.

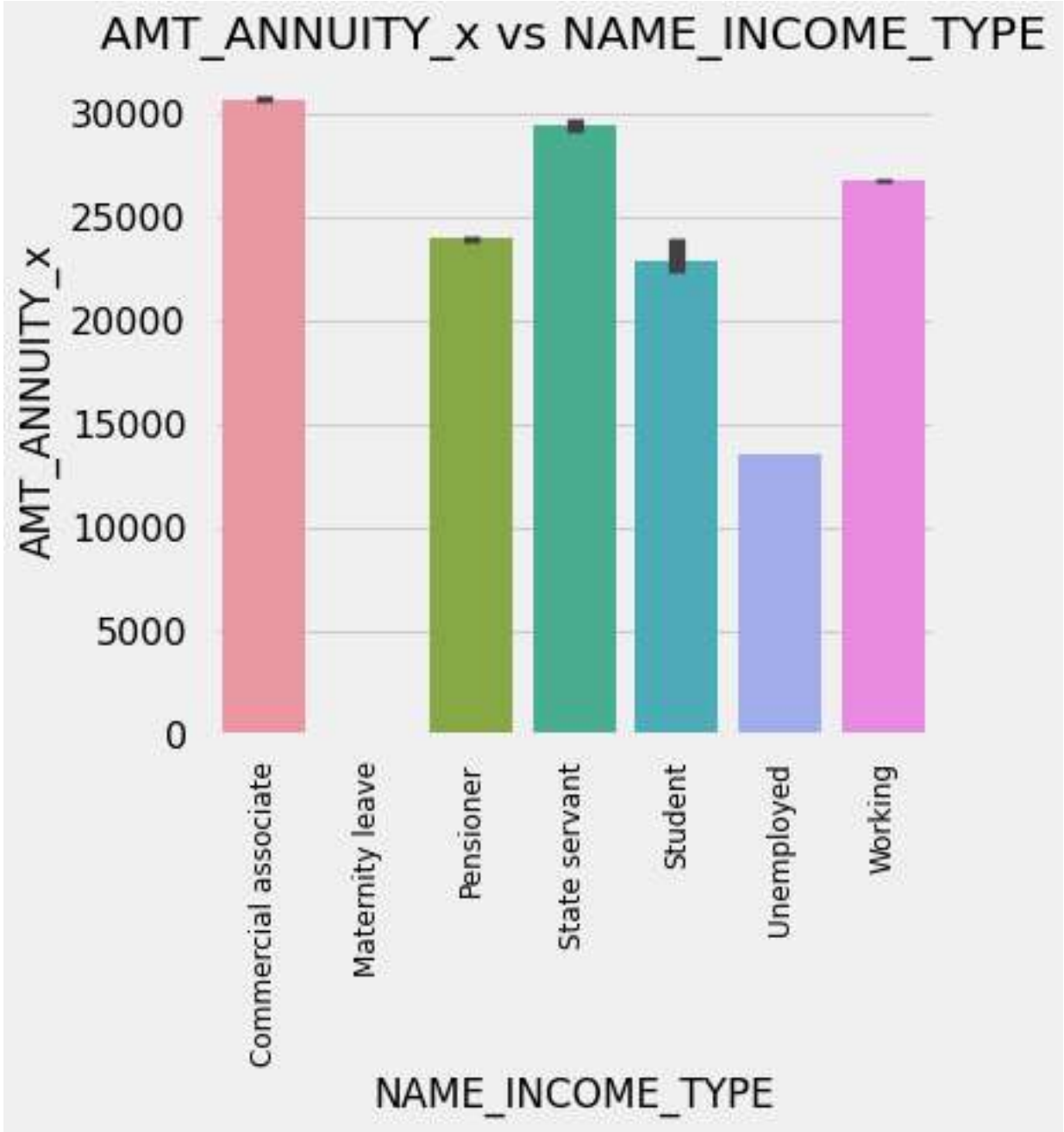
•. The maximum amount of annuity debits from those applicants who came with Family or Unaccompanied type of people at the time of loan application for the previous year



AMT_ANNUITY_X VS NAME_INCOME_TYPE

Points to be concluded from the graph on the right.

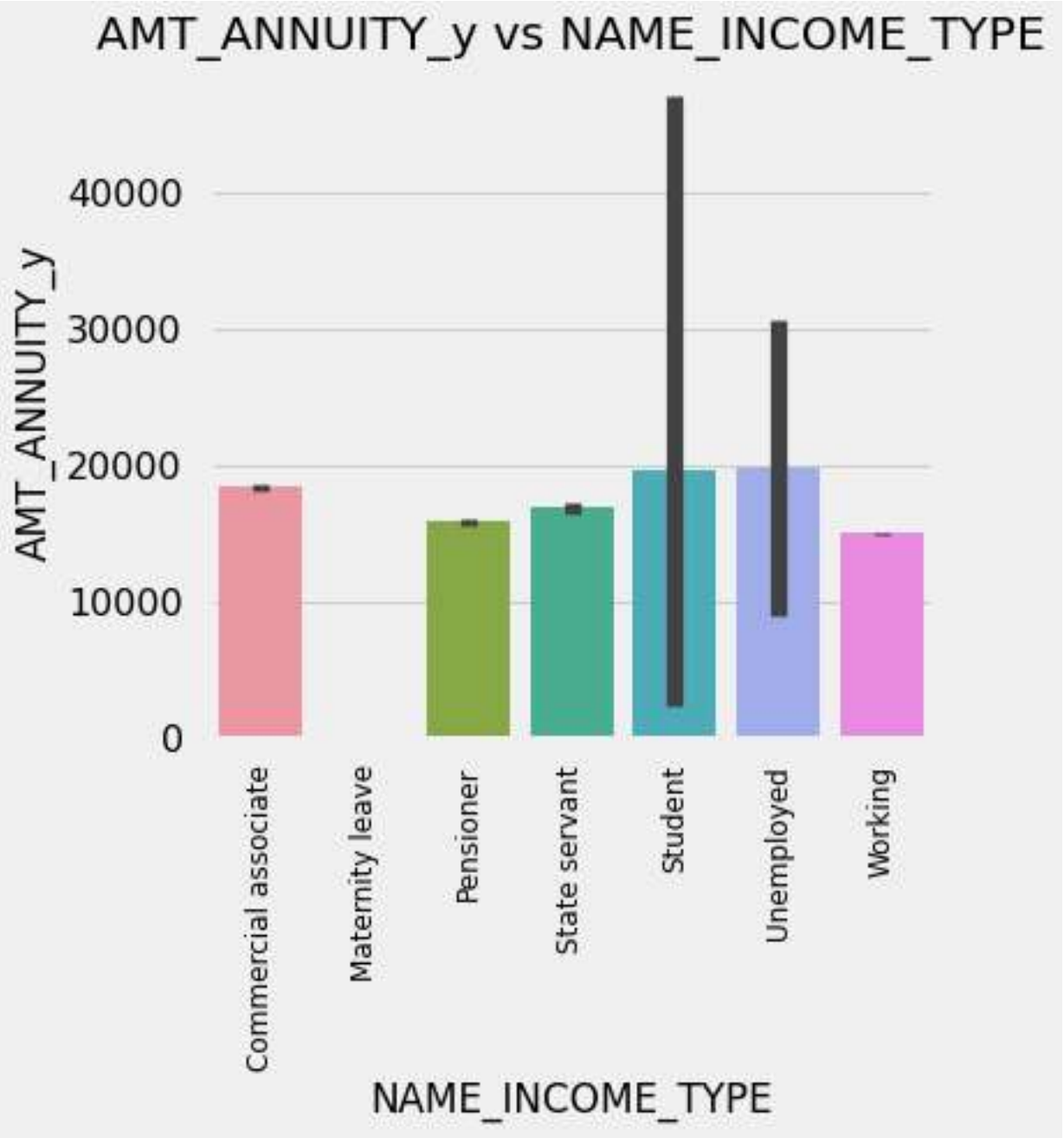
•. The maximum amount of annuity debits from Commerical associate income type people,for the current year.



AMT_ANNUITY_Y VS NAME_INCOME_TYPE

Points to be concluded from the graph on the right.

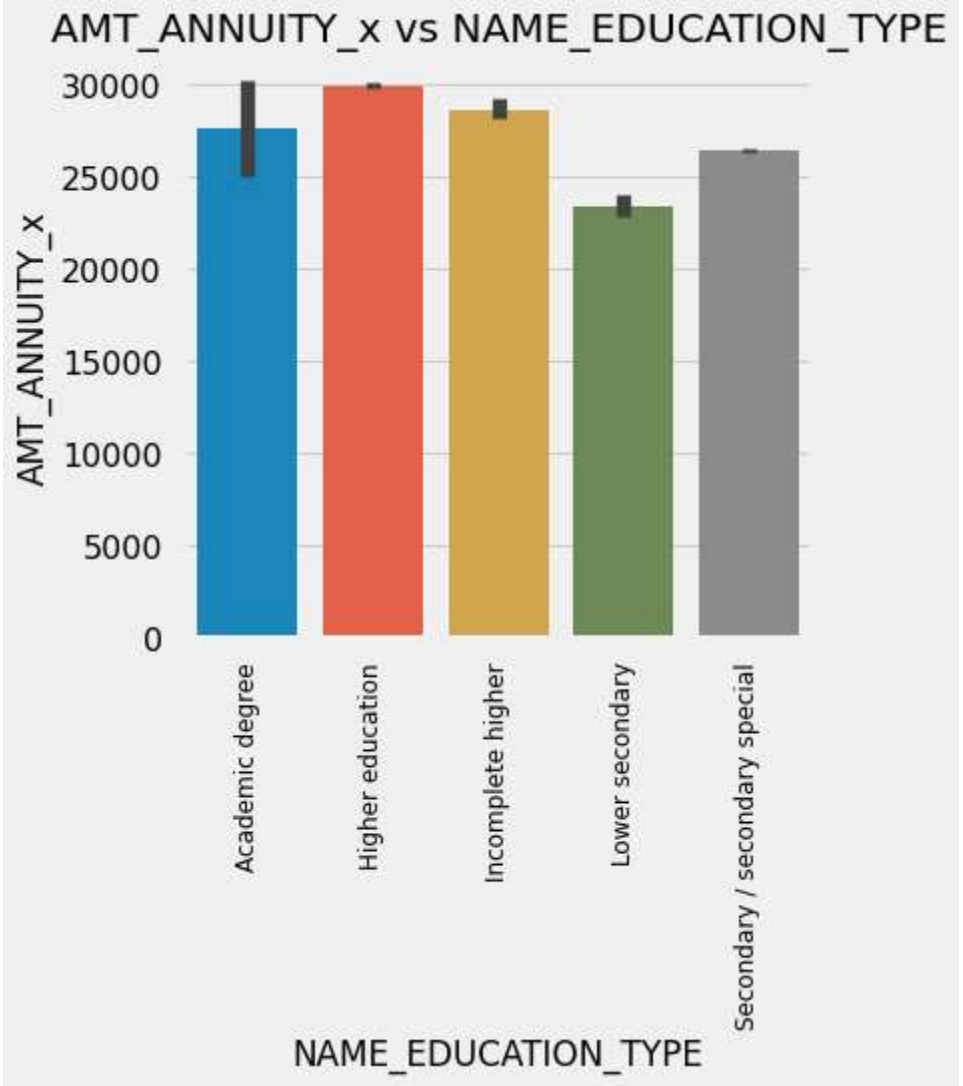
•. The maximum amount of annuity debited from Student and Unemployed income type people, for the previous year.



AMT_ANNUITY_X VS NAME_EDUCATION_TYPE

Points to be concluded from the graph on the right.

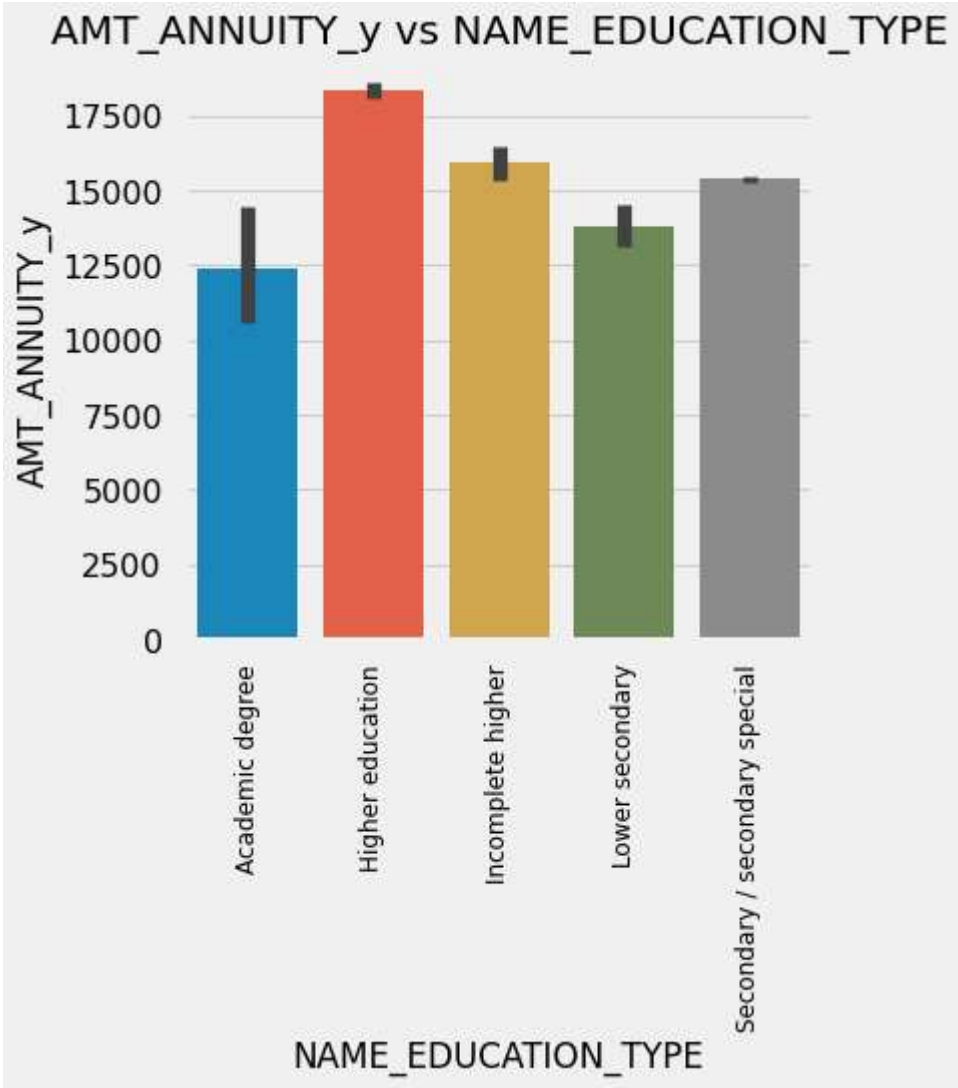
•. The maximum amount of annuity debits from Higher education type applicants for current year.



AMT_ANNUITY_Y VS NAME_EDUCATION_TYPE

Points to be concluded from the graph on the right.

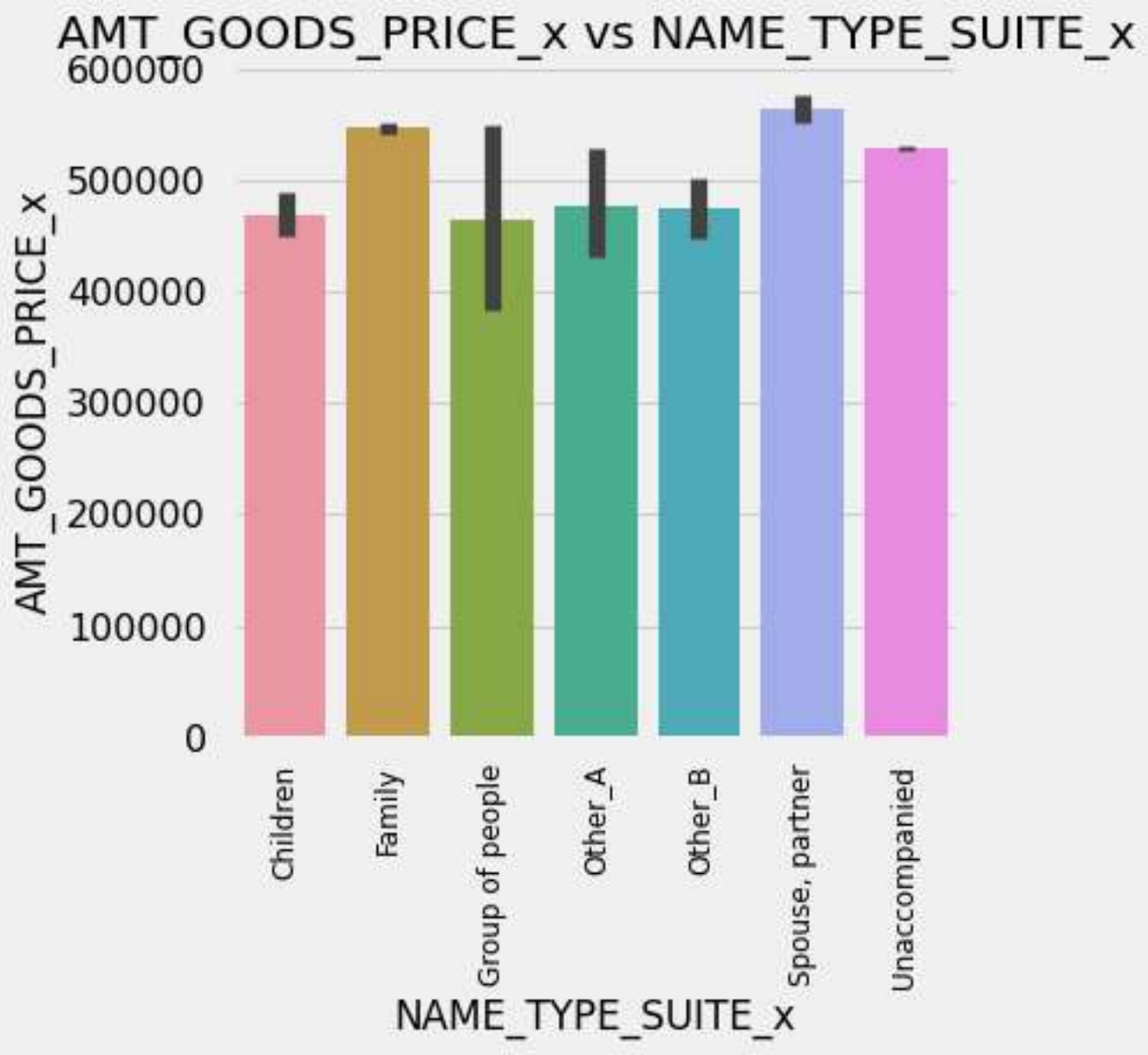
•. The maximum amount of annuity debits from Higher education type applicants for the previous year.



AMT_GOODS_PRICE_X VS NAME_TYPE_SUITE_X

Points to be concluded from the graph on the right.

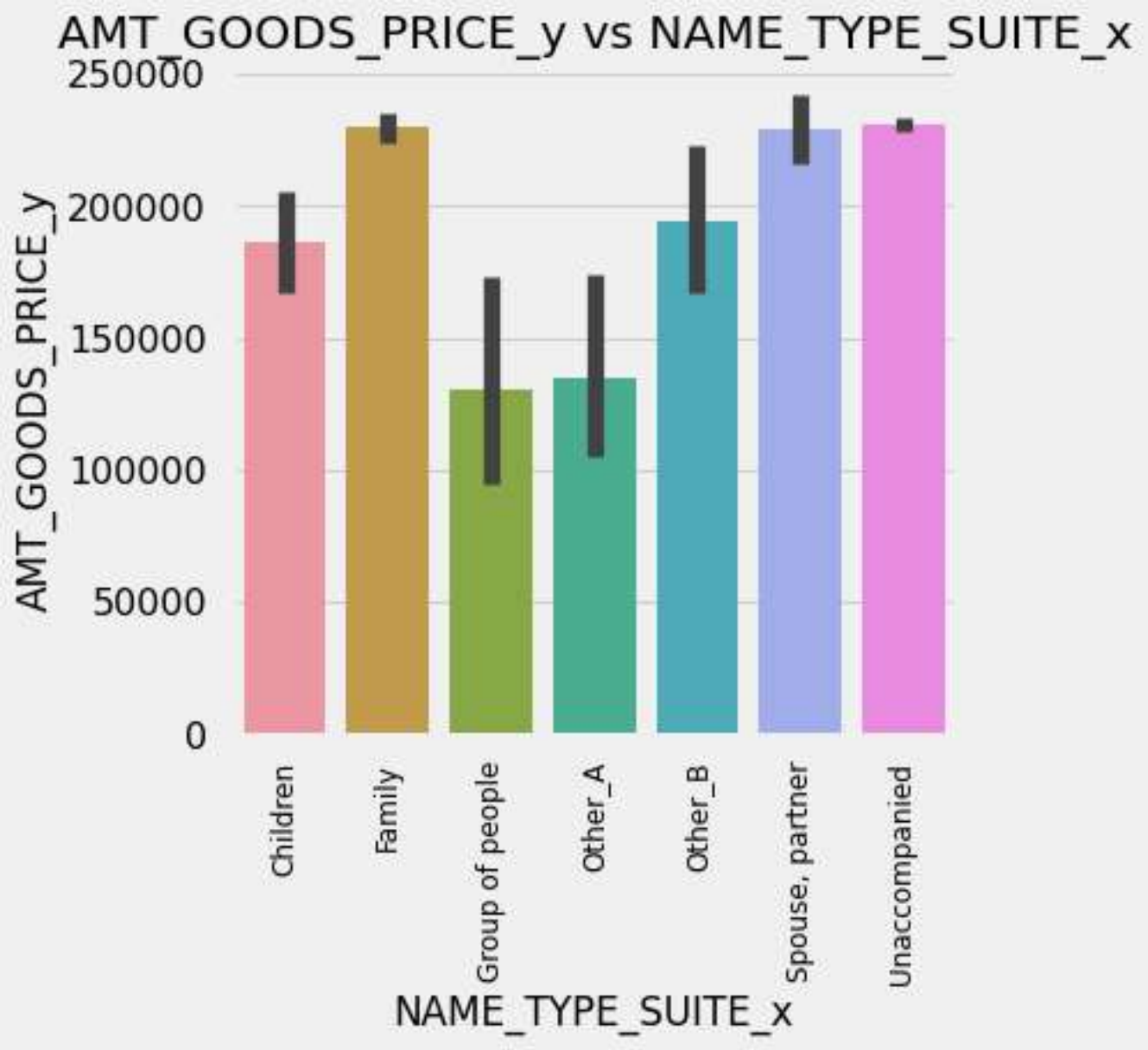
•. The maximum amount of goods price is spent by the Spouse, partner in comparison to other support groups at the time of application in the current year.



AMT_GOODS_PRICE_Y VS NAME_TYPE_SUITE_X

Points to be concluded from the graph on the right.

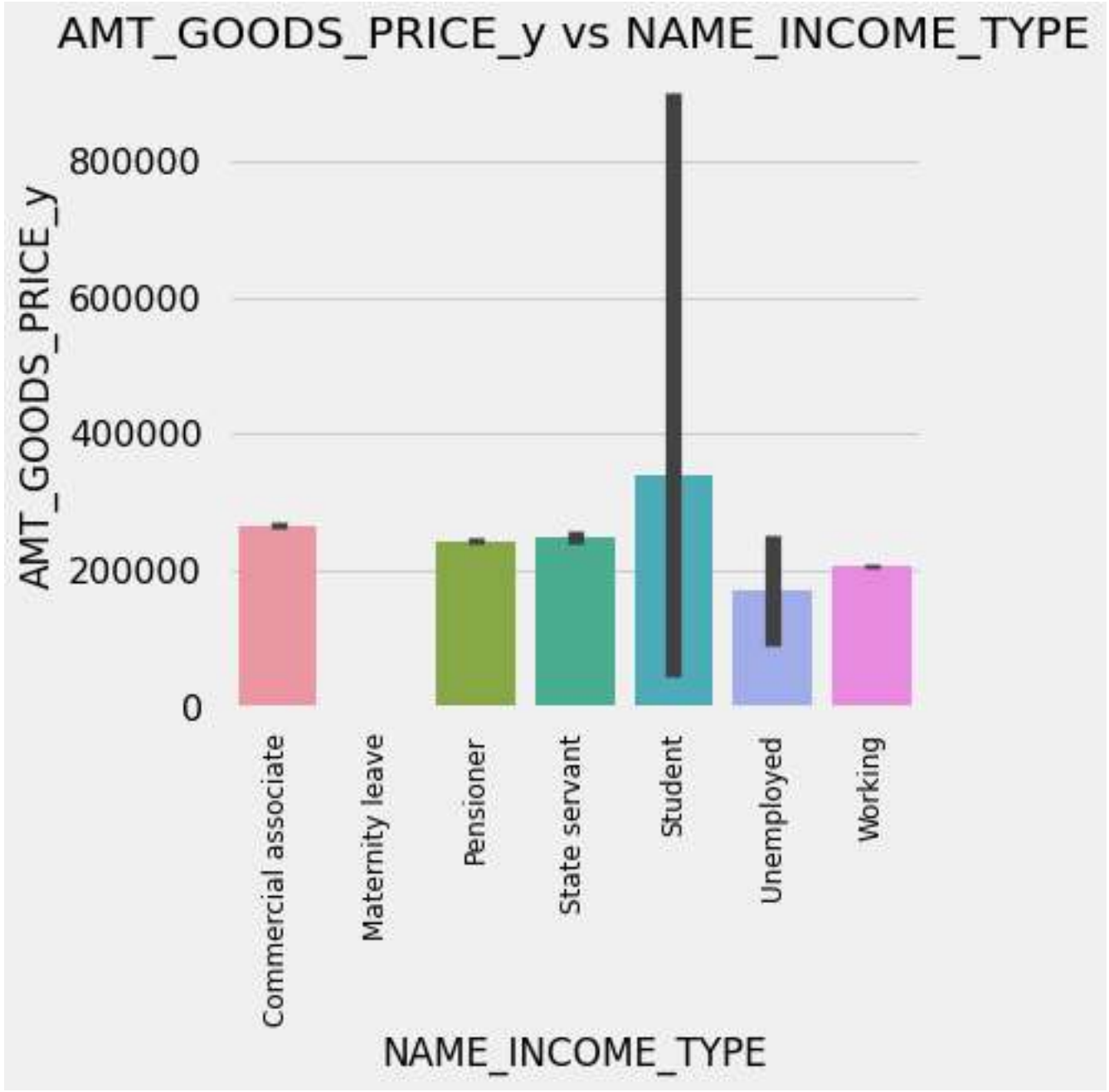
•. The maximum amount of goods price is spent by the Spouse,partner or Unaccompanied or family in comparison to other support groups at the time of application in the previous year.



AMT_GOODS_PRICE_Y VS NAME_INCOME_TYPE

Points to be concluded from the graph on the right.

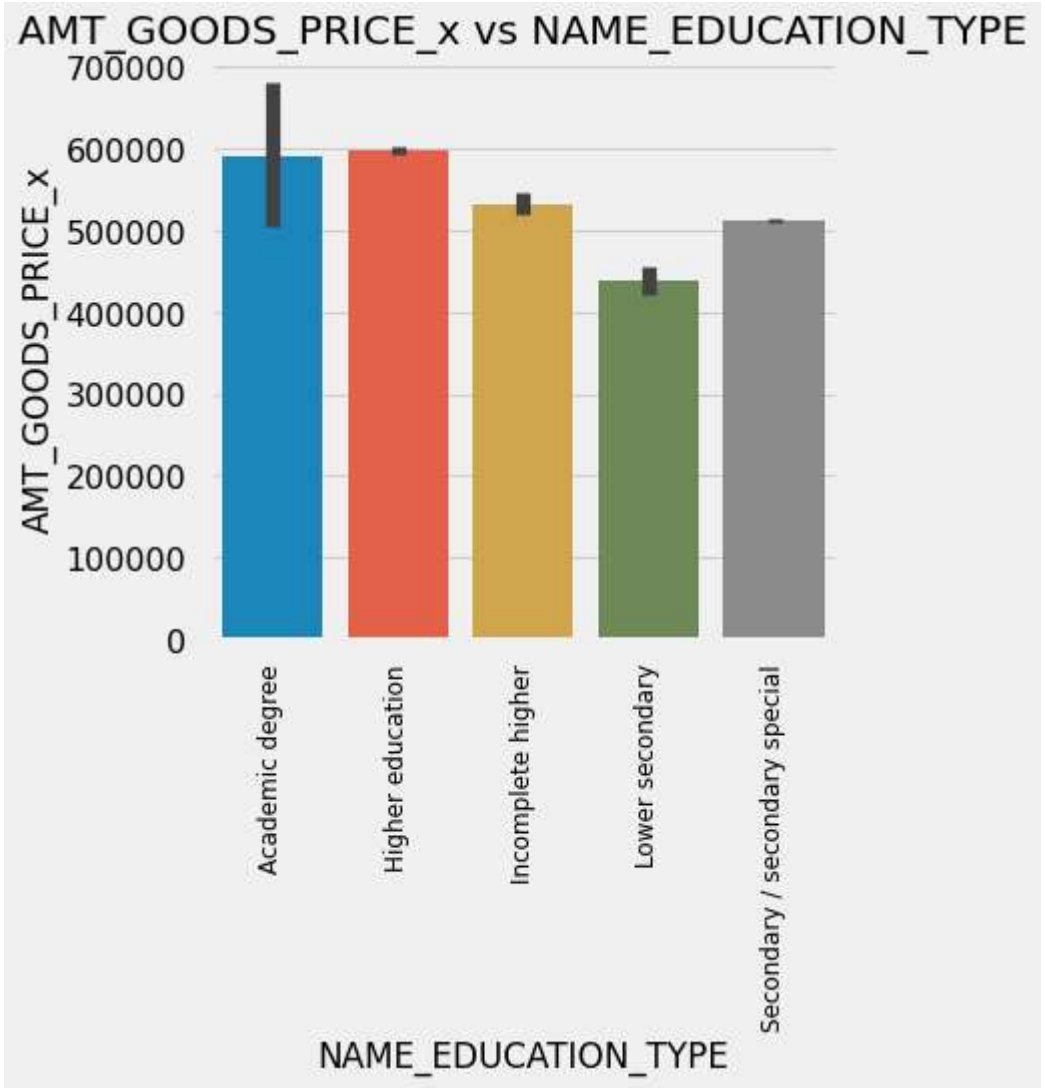
•. The maximum amount of goods price is spent by student income type applicants followed by commercial associate applicants for the previous year.



AMT_GOODS_PRICE_X VS NAME_EDUCATION_TYPE

Points to be concluded from the graph on the right.

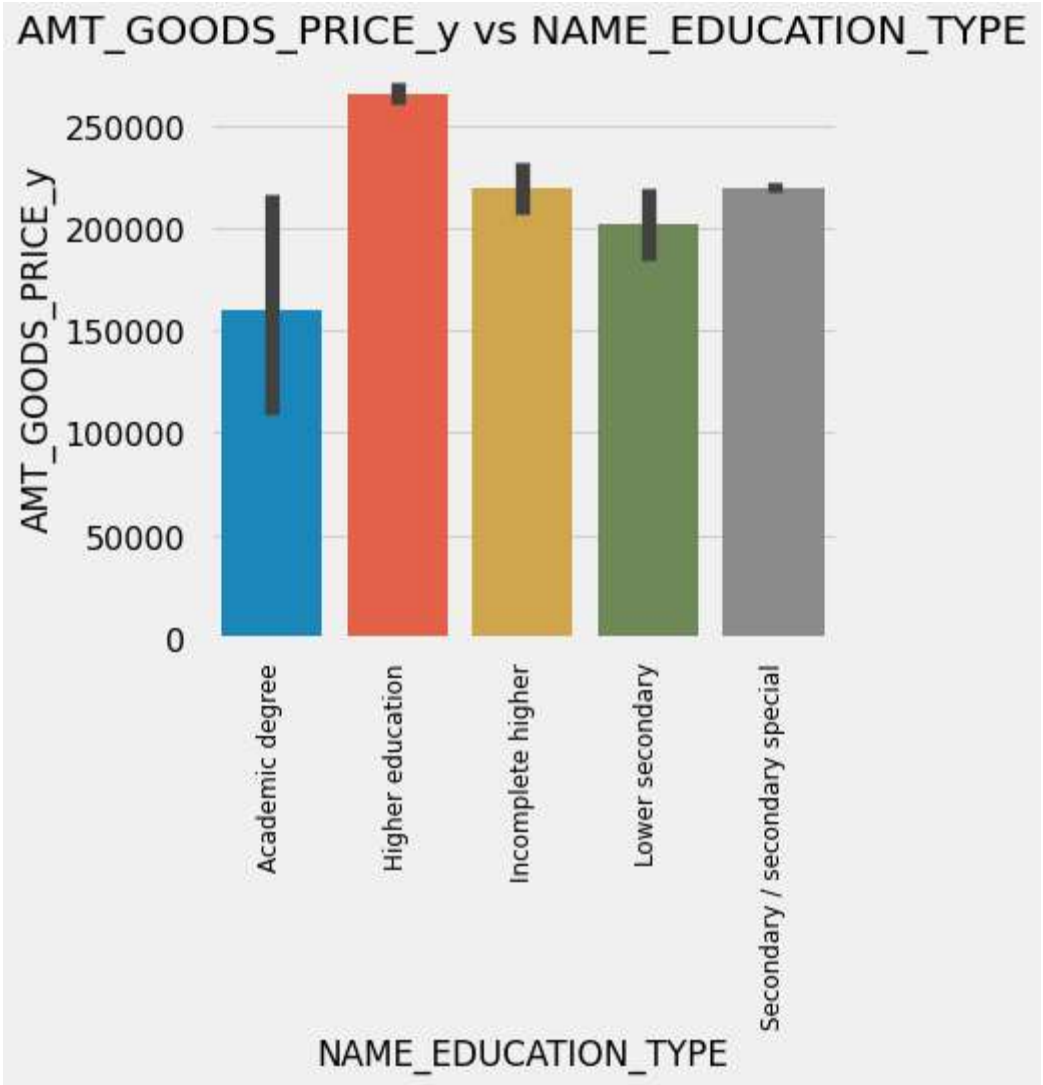
•. The maximum amount of goods price was spent by higher education type applicants followed by academic degree for this year.



AMT_GOODS_PRICE_Y VS NAME_EDUCATION_TYPE

Points to be concluded from the graph on the right.

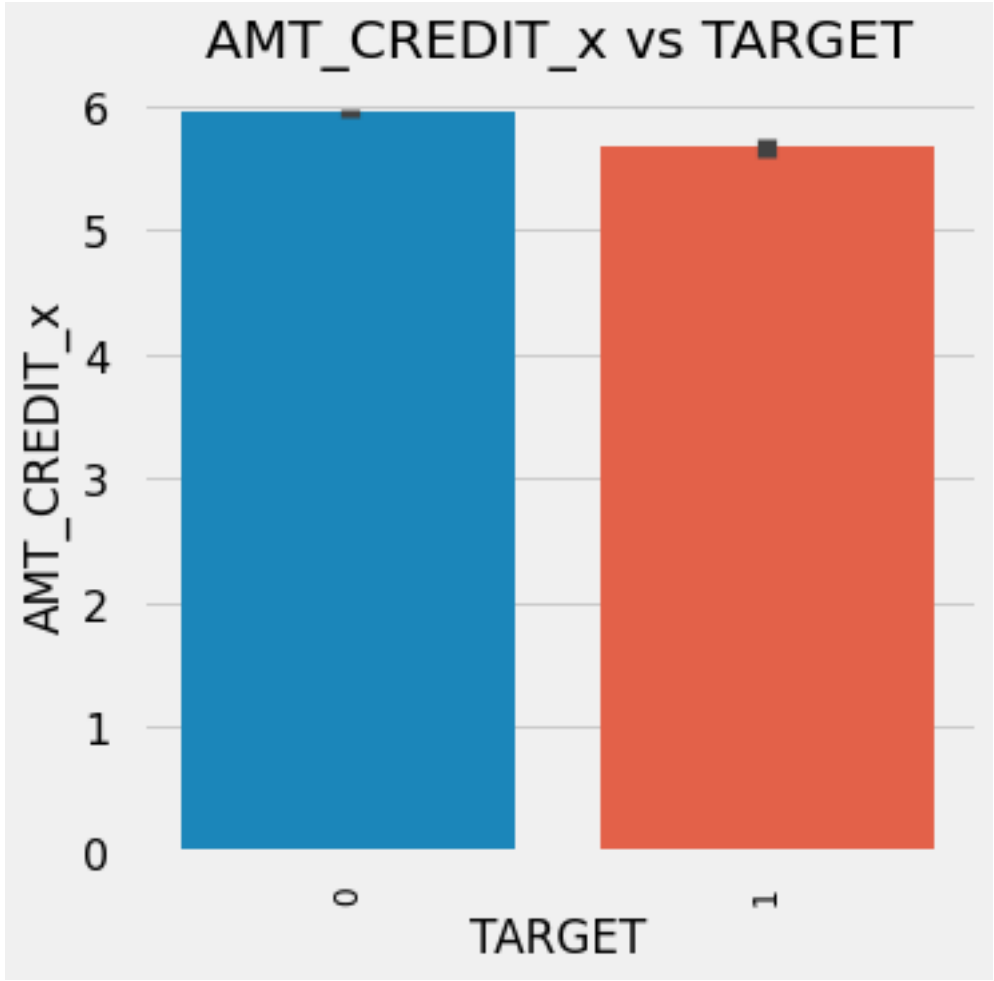
- The maximum amount of goods price was spent by higher education type applicants followed by incomplete higher type for the previous year.



AMT_CREDIT_X VS TARGET

Points to be concluded from the graph on the right.

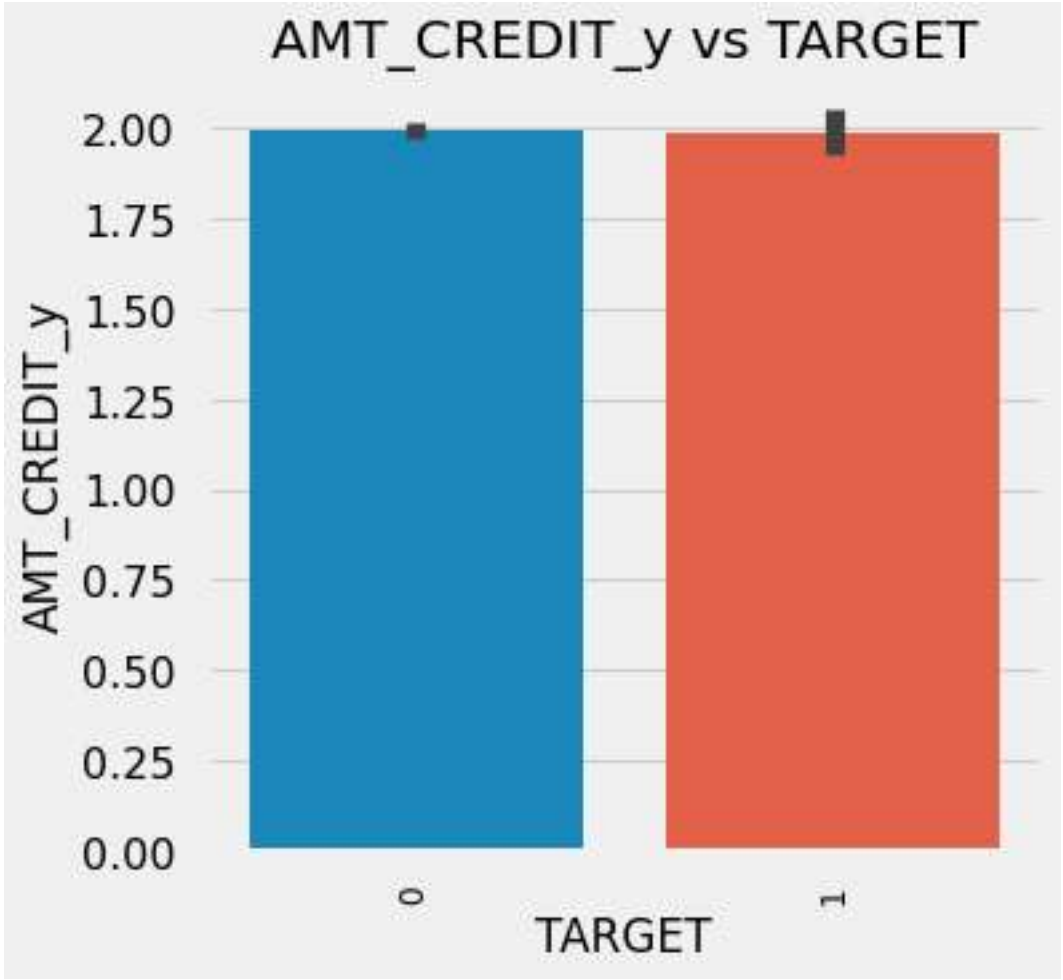
- The repayers of the loan have more loan credit passed than the defaulters for the current year



AMT_GOODS_PRICE_Y VS TARGET

Points to be concluded from the graph on the right.

- The repayers of the loan have almost same loan credit passed as the defaulters for the previous year.



AMT_GOODS_PRICE_X VS TARGET

Points to be concluded from the graph on the right.

- The maximum amount of goods price was spent by the repayers in comparison to defaulters for the current year.

AMT_GOODS_PRICE_Y VS TARGET

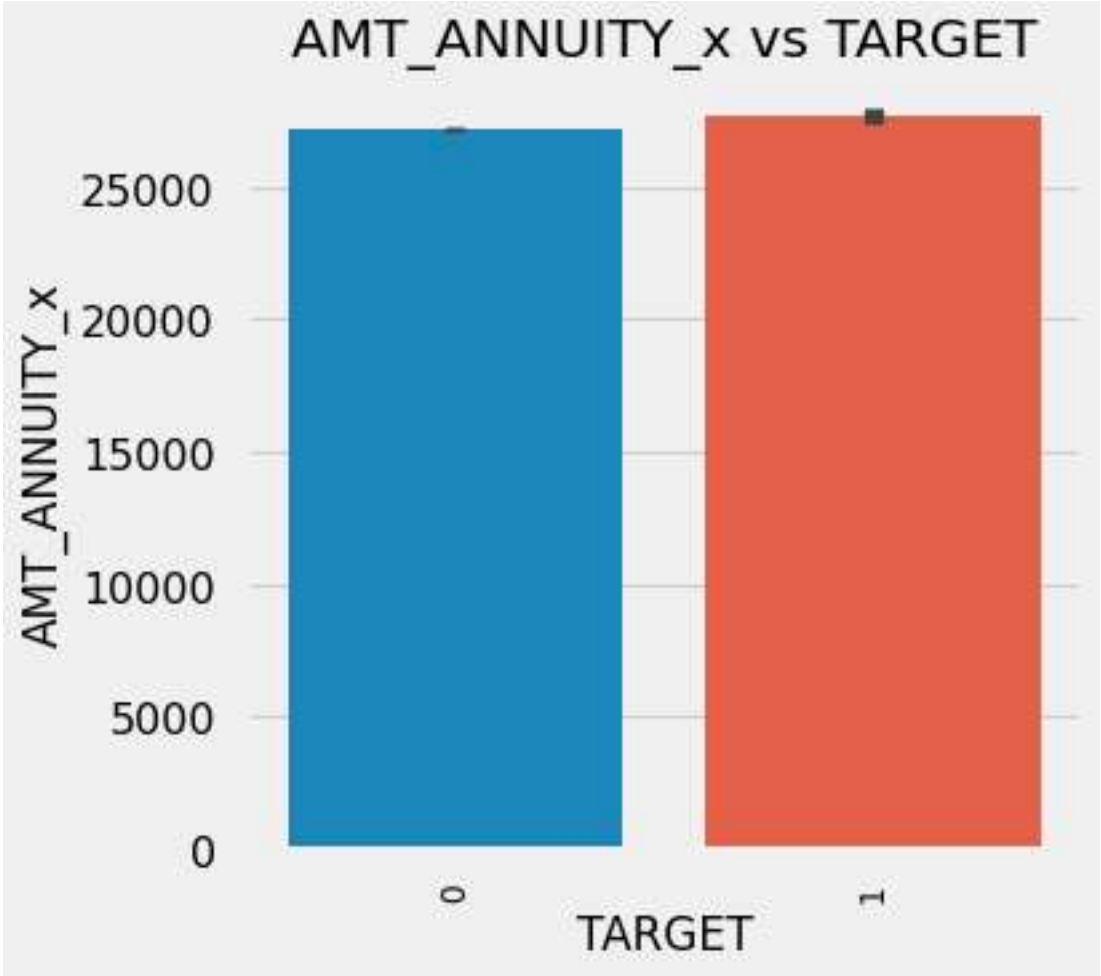
Points to be concluded from the graph on the right.

- The maximum amount of goods price was spent by the defaulters in comparison to repayers for the previous year.

AMT_ANNUITY_X VS TARGET

Points to be concluded from the graph on the right.

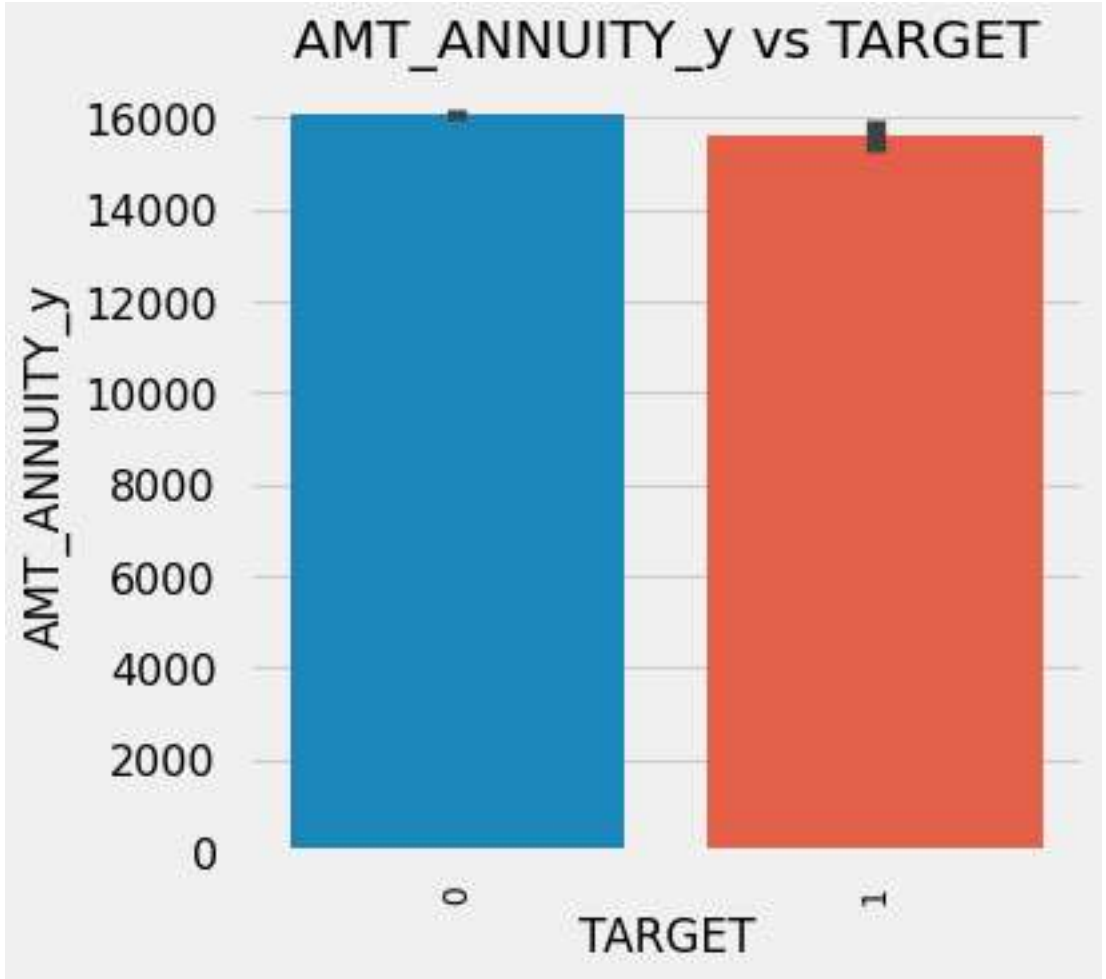
- The amount annuity debited from the repayers and defaulters was almost same for the current year.



AMT_ANNUITY_Y VS TARGET

Points to be concluded from the graph on the right.

•The amount annuity debited from the repayers is more than the defaulters for the previous year.



DAYS_EMPLOYED VS TARGET

Points to be concluded from the graph on the right.

- Those who have been employed more days are repayers in comparison to defaulters.

MULTIVARIATE ANALYSIS

WHEN THE DATA INVOLVES **THREE OR MORE VARIABLES**, IT IS CATEGORIZED UNDER MULTIVARIATE. EXAMPLE OF THIS TYPE OF DATA IS SUPPOSE AN ADVERTISER WANTS TO COMPARE THE POPULARITY OF FOUR ADVERTISEMENTS ON A WEBSITE, THEN THEIR CLICK RATES COULD BE MEASURED FOR BOTH MEN AND WOMEN AND RELATIONSHIPS BETWEEN VARIABLES CAN THEN BE EXAMINED

HEATMAP FOR REPAYERS DATA

Points to be concluded from the graph on the right.

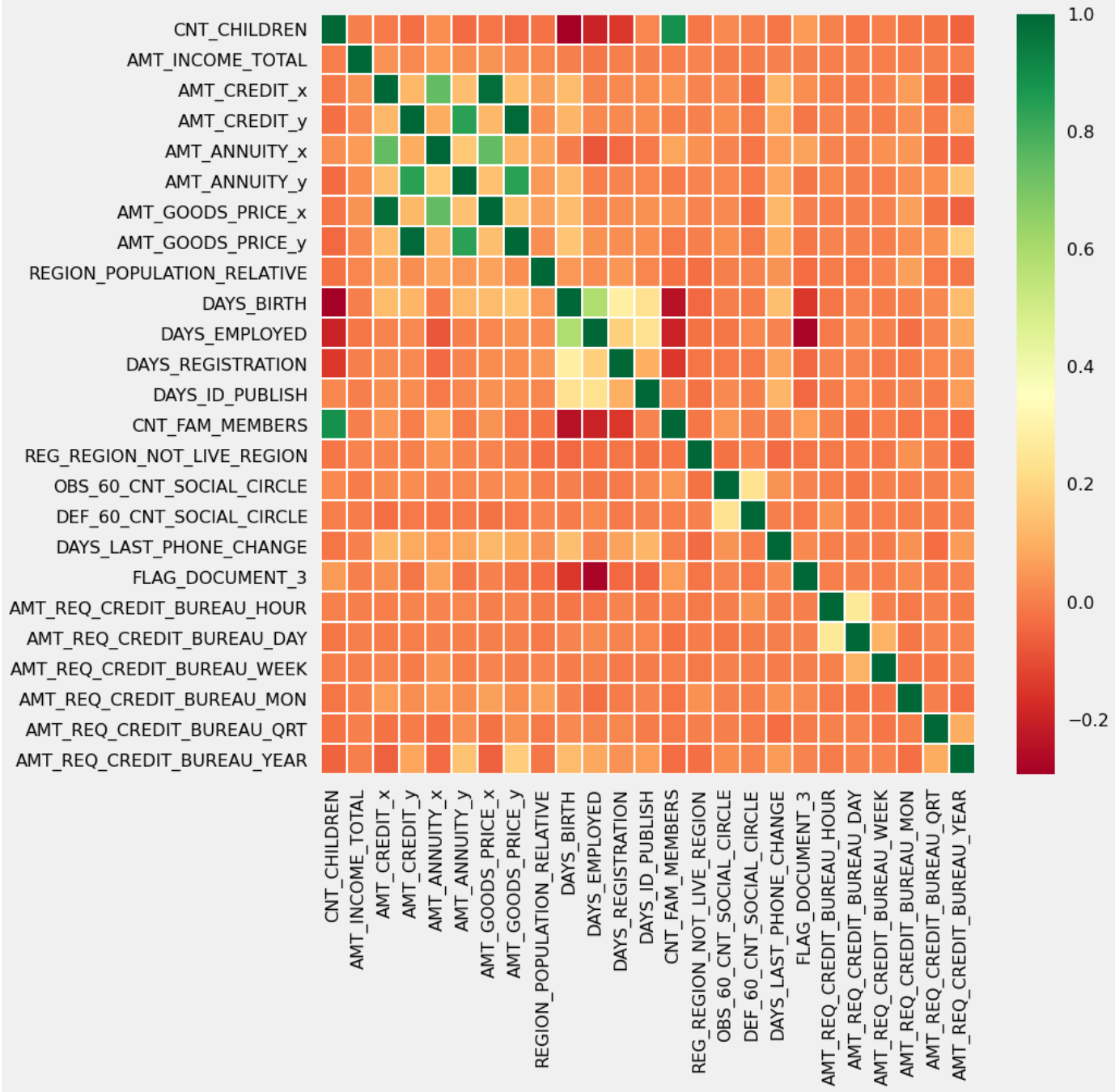
Below fields have high corelation between them with respect to repayers:

1. CNT_FAM_MEMBERS vs CNT_CHILDREN
2. DAYS_EMPLOYED vs DAYS_BIRTH
3. AMT_GOODS_PRICE_x vs AMT_CREDIT_x for current year
4. AMT_GOODS_PRICE_y vs AMT_CREDIT_y for previous year
5. AMT_ANNUITY_x vs AMT_CREDIT_x for current year
6. AMT_ANNUITY_y vs AMT_CREDIT_y for previous year
7. AMT_GOODS_PRICE_x vs AMT_ANNUITY_x for current year
8. AMT_GOODS_PRICE_y vs AMT_ANNUITY_y for previous year

HEATMAP FOR DEFAULTER DATA

Below fields have high corelation between them with respect to defaulters:

- CNT_CHILDREN vs CNT_FAM_MEMBERS
- AMT_ANNUITY_x vs AMT_CREDIT_x for current year
- AMT_ANNUITY_y vs AMT_CREDIT_y for previous year
- AMT_GOODS_PRICE_x vs AMT_CREDIT_x for current year
- AMT_GOODS_PRICE_y vs AMT_CREDIT_y for previous year
- AMT_GOODS_PRICE_x vs AMT_ANNUITY_x for current year
- AMT_GOODS_PRICE_y vs AMT_ANNUITY_y for previous year.



CONCLUSION

- If we check the above Value counts output, then we can be sure that more than 50% of the applicants have Total Income in the range of 100K-200K. The next highest proportion of the people lie in the range of 200K-300K with percentage of 23%.If we aggregate the above percentages and their respective value range, then we observe that around 96% applicants for loan have Income below 400K range.
- we can clearly see that maximum number of loan applicants i.e. 17 percent have been sanctioned a loan in the amount range 200K-300K. The next highest bucket of applicants goes to 15% where the amount of loan sanctioned or the Credit Amount is 1 Million or above for the current year. The highest proportion of the sanctioned loan applications last year were almost 45% where the credit range was between 0-100K for the previous year
- The most number of applicants whose applications have been sanctioned or approved belong to the age group 50 and Above, with proportion equal to 33 percent. The next best group is 30-40 with 26 percent, since this is the medium age group and also the age group which generally earns a wholesome salary.
- we can figure out that the maximum number of applicants, about 51 percent, have been employed since 0-5 years. The other inference that can be extracted from the data is that the minimum proportion of loan applicants belong to the employment group 50 and above.
- we can predict that out of the total values that are present in the TARGET column, if we take 1 as defaulter and 0 as re payer, the there is an imbalance in the values. 0 taken precedence over the 1 value in target column in huge proportion.
- we can see that the number of Cash loans is far more than the total number of revolving loans (92 percent vs 7 percent) for the current year. we can conclude that the maximum number of applications have contract type as "Cash loans" and "Consumer loans" at 44 percent each and there is no data in "XNA" contract type. The least number of applications are for the name contract type "Revolving loans" at 11 percent for the previous year.
- we can observe that the majority of the applicants this year have been from the Female gender category and they have been given credit loans. Thus it is simply obvious that there is lesser risk in giving loans to Female category applicants than male category applicants who are at 32 percent(latter) in comparison to 67 percent of the former.

CONCLUSION

- There is a majority of those applicants who own a realty or real estate as 72 percent, in comparison to those applicants who don't own a house or flat at 27 percent. Thus, there is lesser risk in giving loans to those who own a flat or house than those who don't.
- Maximum number of loan applicant live in a house/apartment.
- Maximum number of loan applicant are Married in comparision to to the least number of loan applicant over widowed.
- By looking at the above graph we can observe that maximum number of loan applications have submitted FLAG_DOCUMENT_3 and thus there loan applications have less risk factor associated with them in comparision to those applications where the value is 0 and no FLAG_DOCUMENT_3 has been submitted.
- The maximum amount of loan Sanctioned this Year belongs to the 200k-300k range. The amount of loan sanctioned maximum number of times, last year, belongs to 0-100k range. This means that last year bank saw less risk associated with sanctioned of low amounts in credit than the current year.
- The salary earned by maximum number of applicants this year lies in the range of 100k-200k.
- The maximum nuber of loan applicants have a family that contains two members in total.

CONCLUSION

- The income range seems to be increasing in a linear fashion with respect to credit amount. So, as the income increases the amount credit also increases. From 0 to 800k income increases so credit increases as well and then it starts dropping for this year. The credit amount increases in a linear fashion as income increases for the previous year. The highest proportional relationship is at 700k to 800k income range.
- The amount annuity is directly proportional to the amount credit for this year. The annuity amount is directly proportional to the credit amount for the previous year's applications.
- The maximum loan amount was credited to Commercial associate and state servant in comparison to other income types for current year.The maximum loan amount was passed for Student income type in comparison to others in the previous year.
- The maximum amount of credit was sanctioned for applicants with higher education degree after which the next best education type was Academic degree for the current year applications.The maximum amount of credit was sanctioned for applicants with higher education degree after which the next best education types were Incomplete higher and Secondary degree for the previous year applications.
- Those who have been employed more days are repayers in comparison to defaulters.
- AMT_GOODS_PRICE and AMT_ANNUITY are directly proportional to AMT_CREDIT field in both current and previous year.

THANK YOU