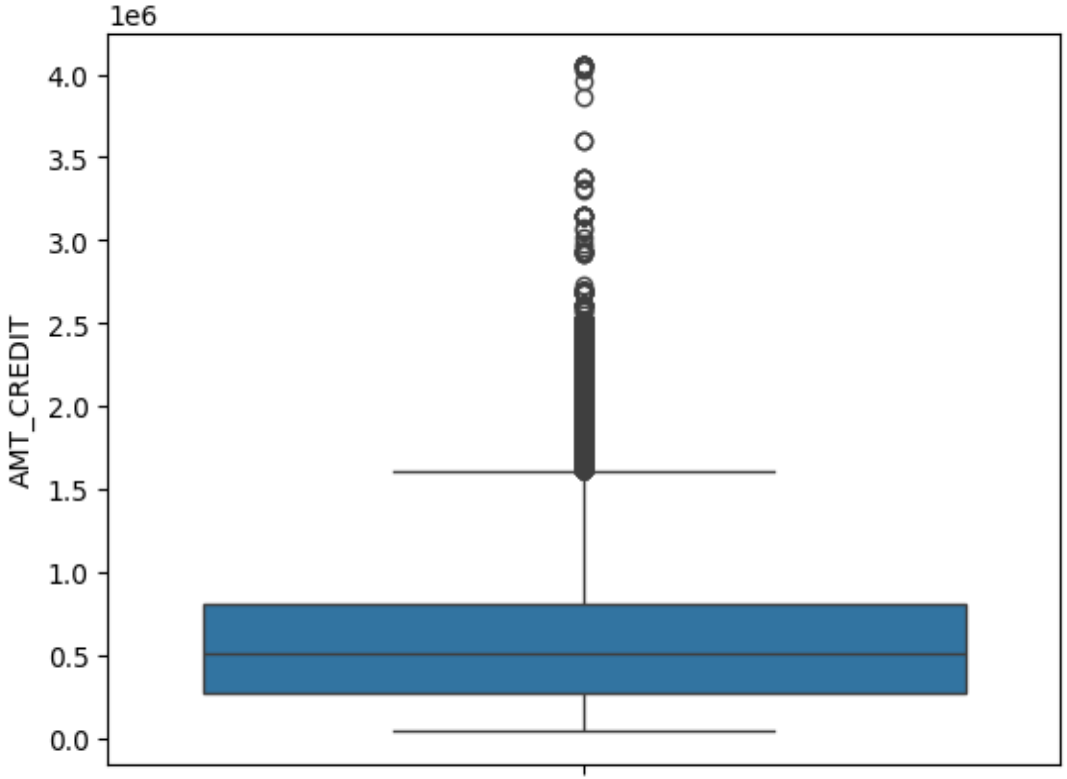
Credit EDA Case Study

By Anshul Mishra

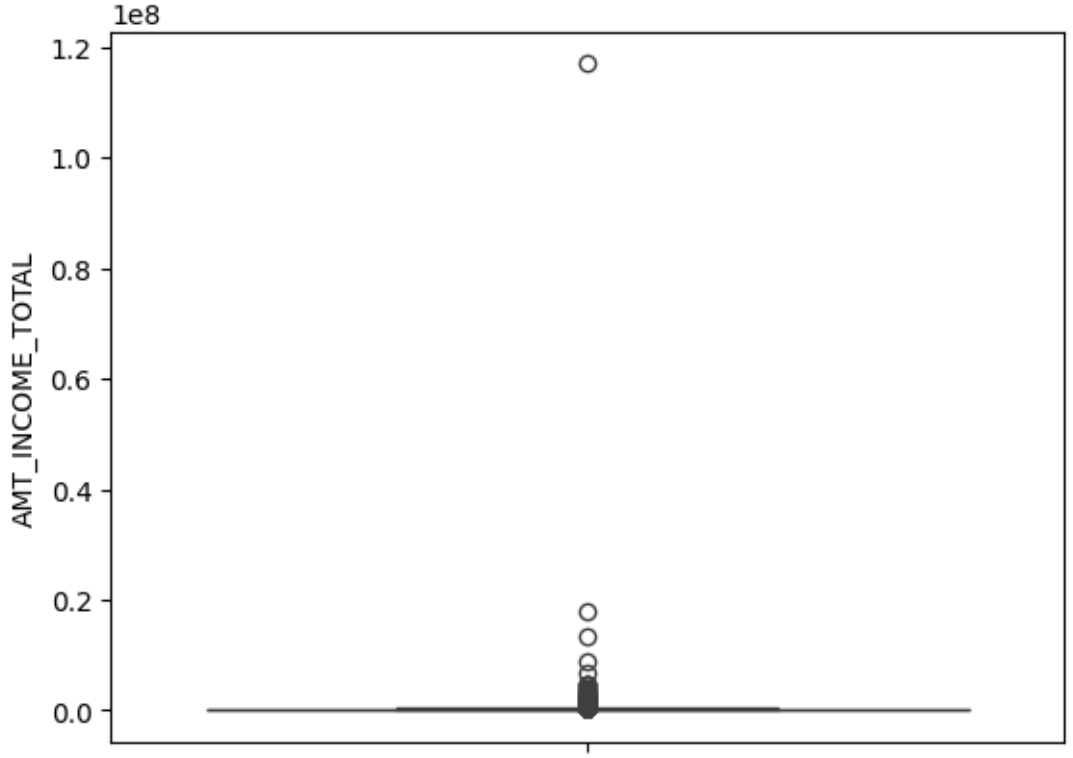
BOXPLOT FOR AMT_CREDIT

1) On analysing the AMT_CREDIT column we notice various outliers.
2) We notice that many outliers lie above the upper quartile.
3) The median is somewhere around 0.5(1e6 in scale) .
4)We also notice that first quartile is very thin as compared to third quartile which means majority of the customers lie in the third quartile region.

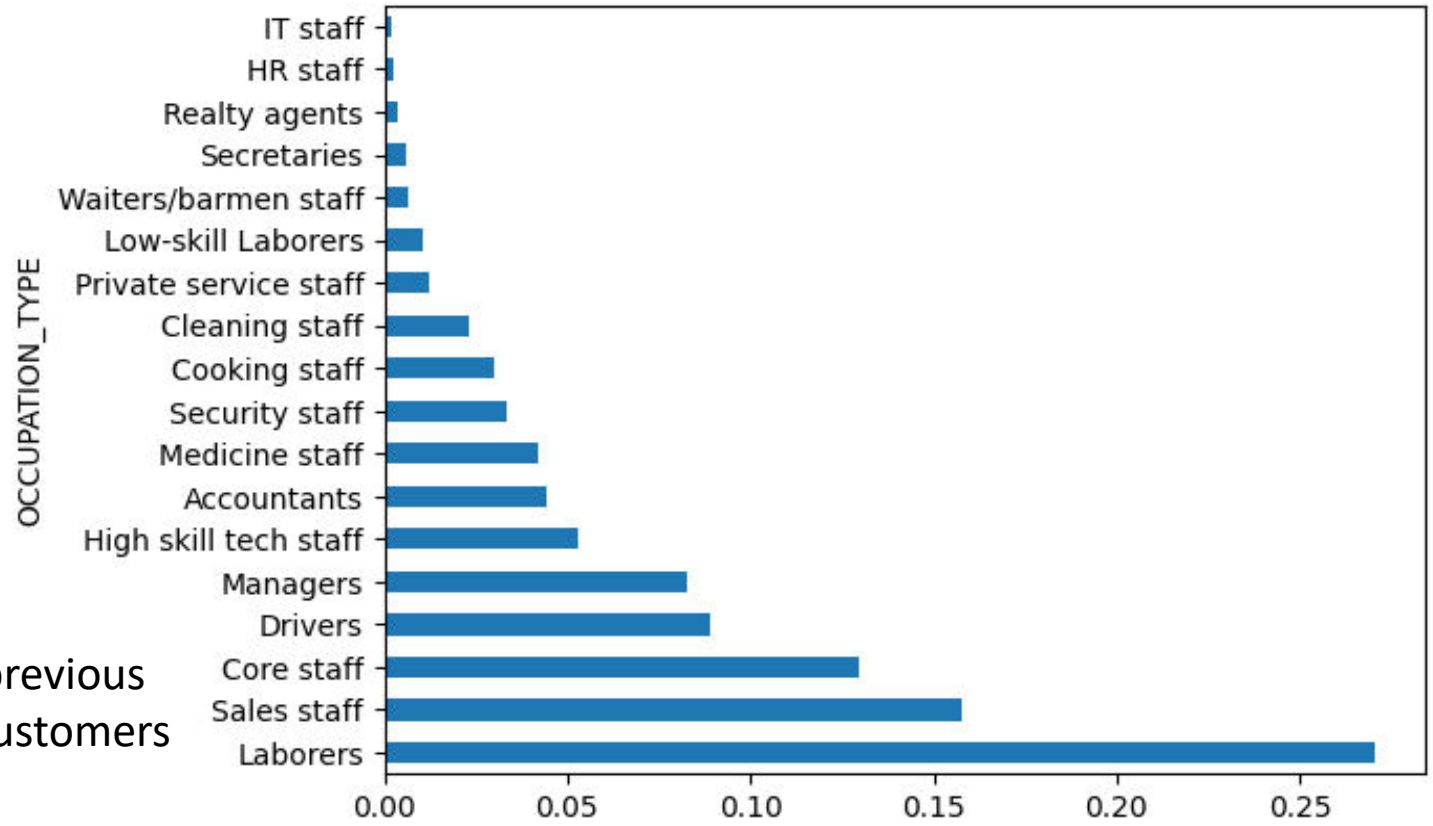


BOXPLOT FOR AMT_INCOME_TOTAL

1) On analysing the AMT_INCOME_TOTAL we notice there are many outliers in the boxplot.
2) First and third quartiles are very thin.
3) It clearly suggests that majority of the customers neither lie on the first quartile nor on the third quartile.



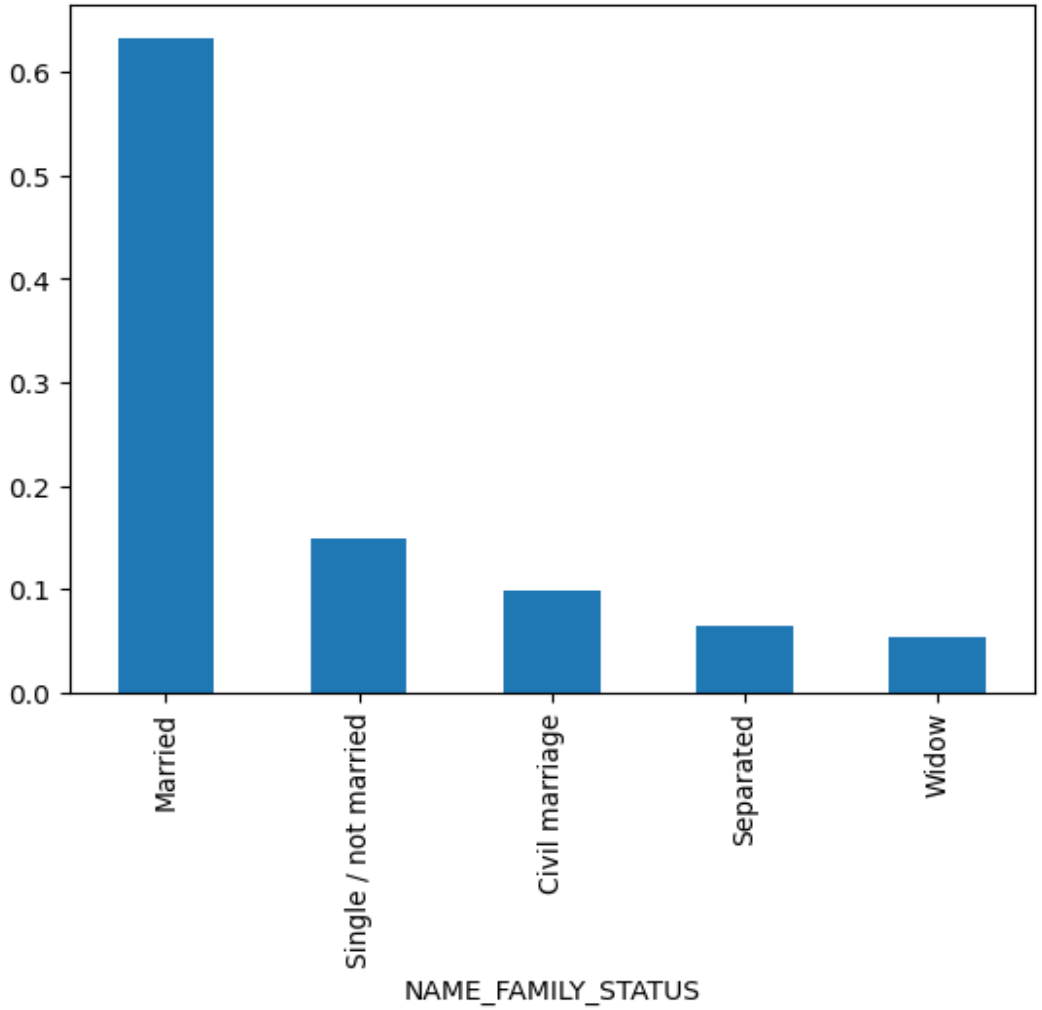
UNIVARIATE UNORDERED CATEGORICAL ANALYSIS



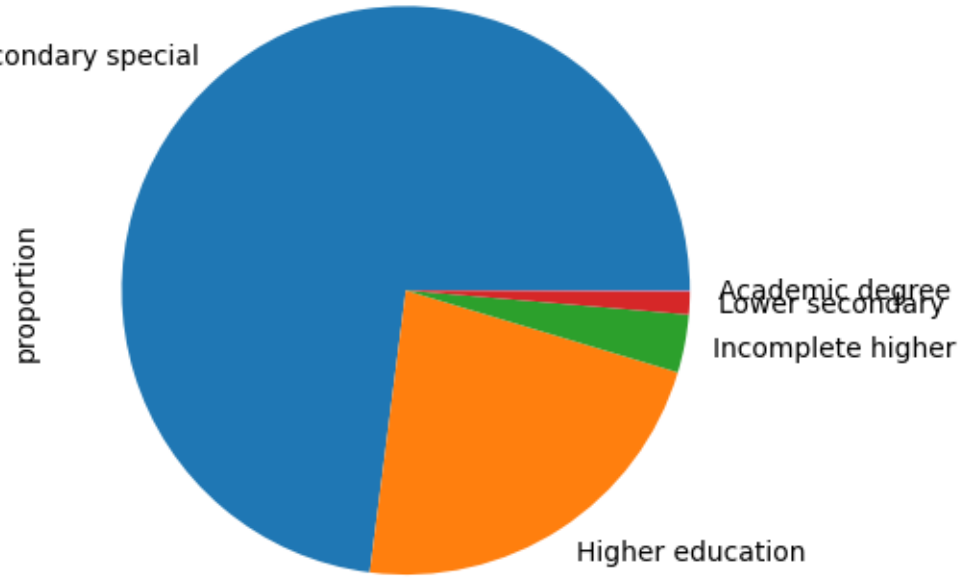
On analysing the OCCUPATION_TYPE column in the previous slide we see that the major occupation type of the customers are Labourers followed by sales staff and core staff.
We notice that majority of them are middle class people or even lower than that .
Hence a proper policy can be taken for these people can be taken based on whether they are defaulter or not and accordingly some amount of loan can be given to them.

UNIVARIATE ORDERED CATEGORICAL ANALYSIS

1) On analysing the NAME_FAMILY_STATUS column the majority lies in married followed by the single and civil marriage.
2) We can take a note of the family status and see whether which of the categories lie on the defaulter list or not.

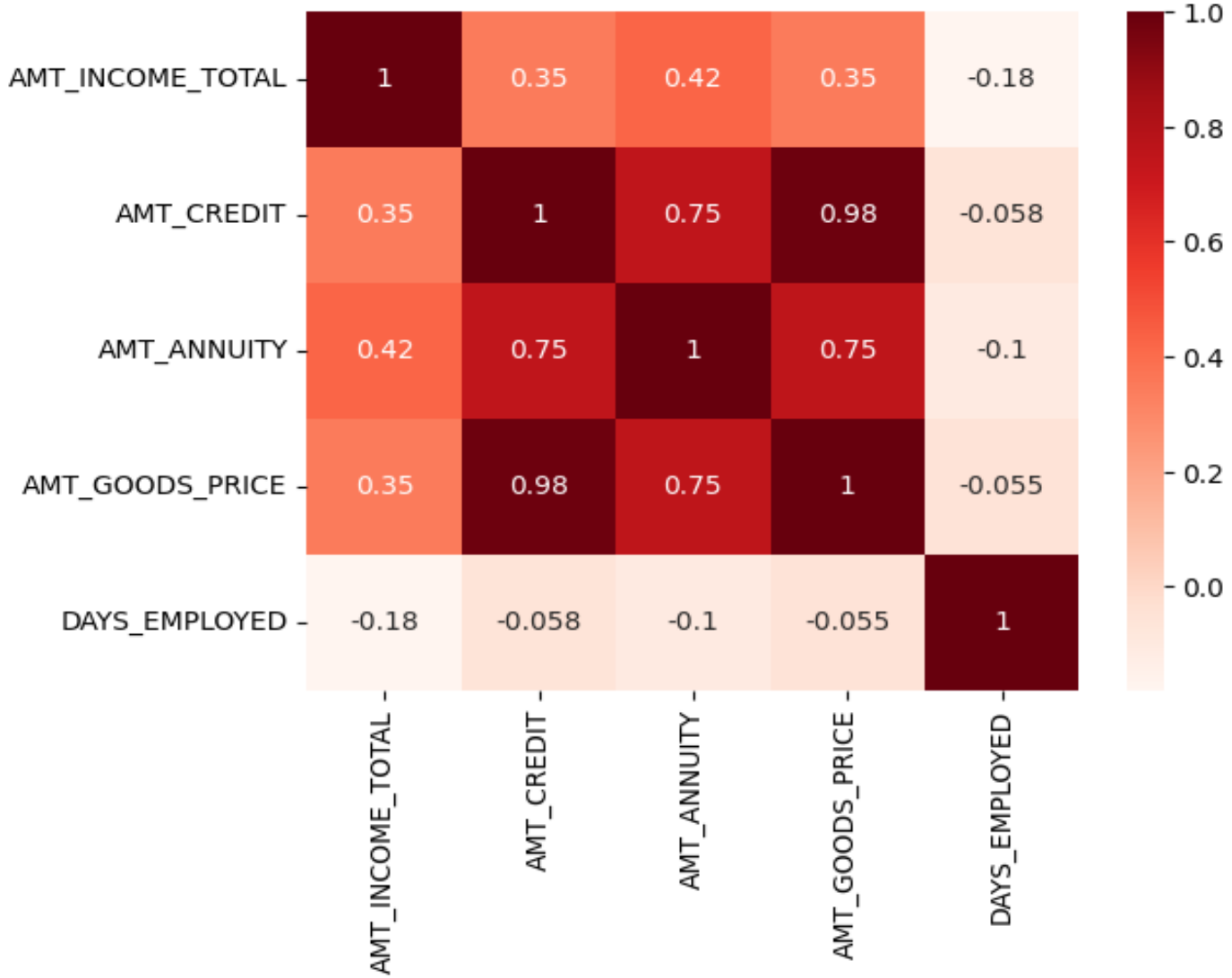


UNIVARIATE CATEGORICAL ORDERED ANALYSIS ON NAME_EDUCATION TYPE COLUMN



Similarly we are performing univariate analysis on the type of education column through the pie chart.
We see the major part here lies in the secondary category followed by the Higher education.
Incomplete Higher and lower secondary category have very less contribution as compared to the other categories.

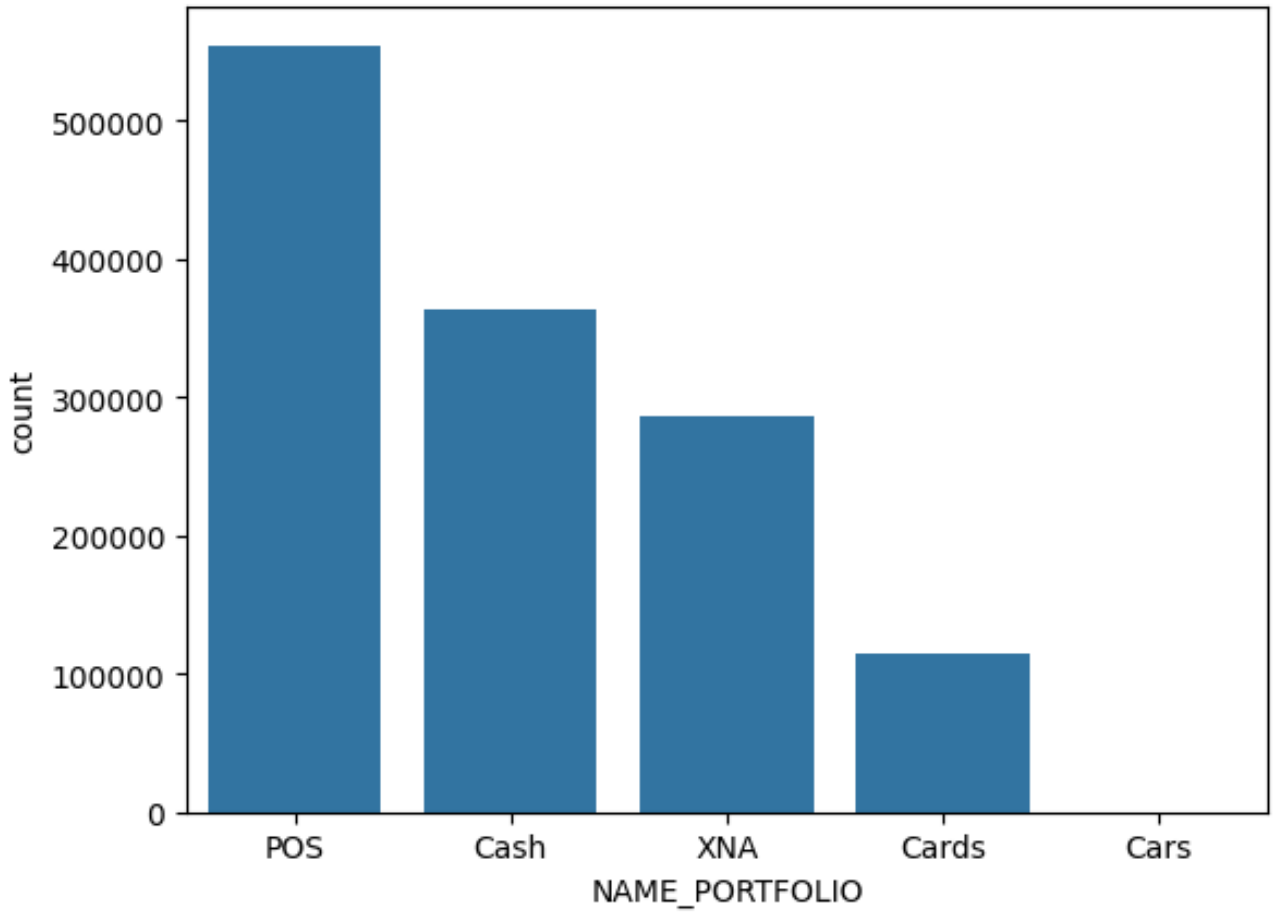
BIVARIATE ANALYSIS THROUGH CORRELATION MATRIX



1) We notice that there is positive value between the AMT_CREDIT and AMT_ANNUITY which suggests that both are directly proportional to each other increase in one parameter will increase the other.

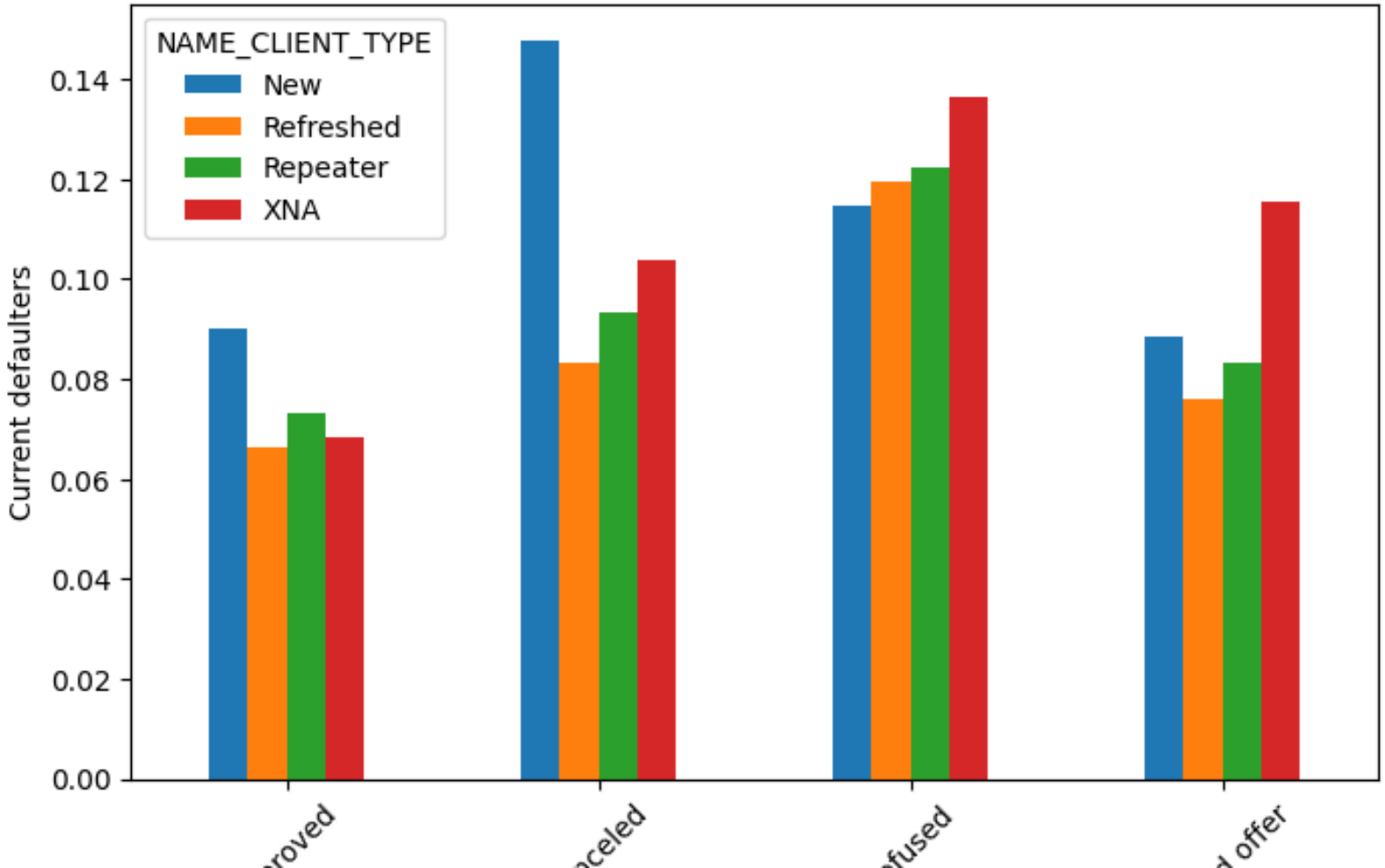
2) We also that there is negative value between the DAYS_EMPLOYED and any other column which suggests that both are inversely proportional to each other increase in one parameter will decrease the other.

PERFORMING UNIVARIATE ANALYSIS ON PREVIOUS DATASET



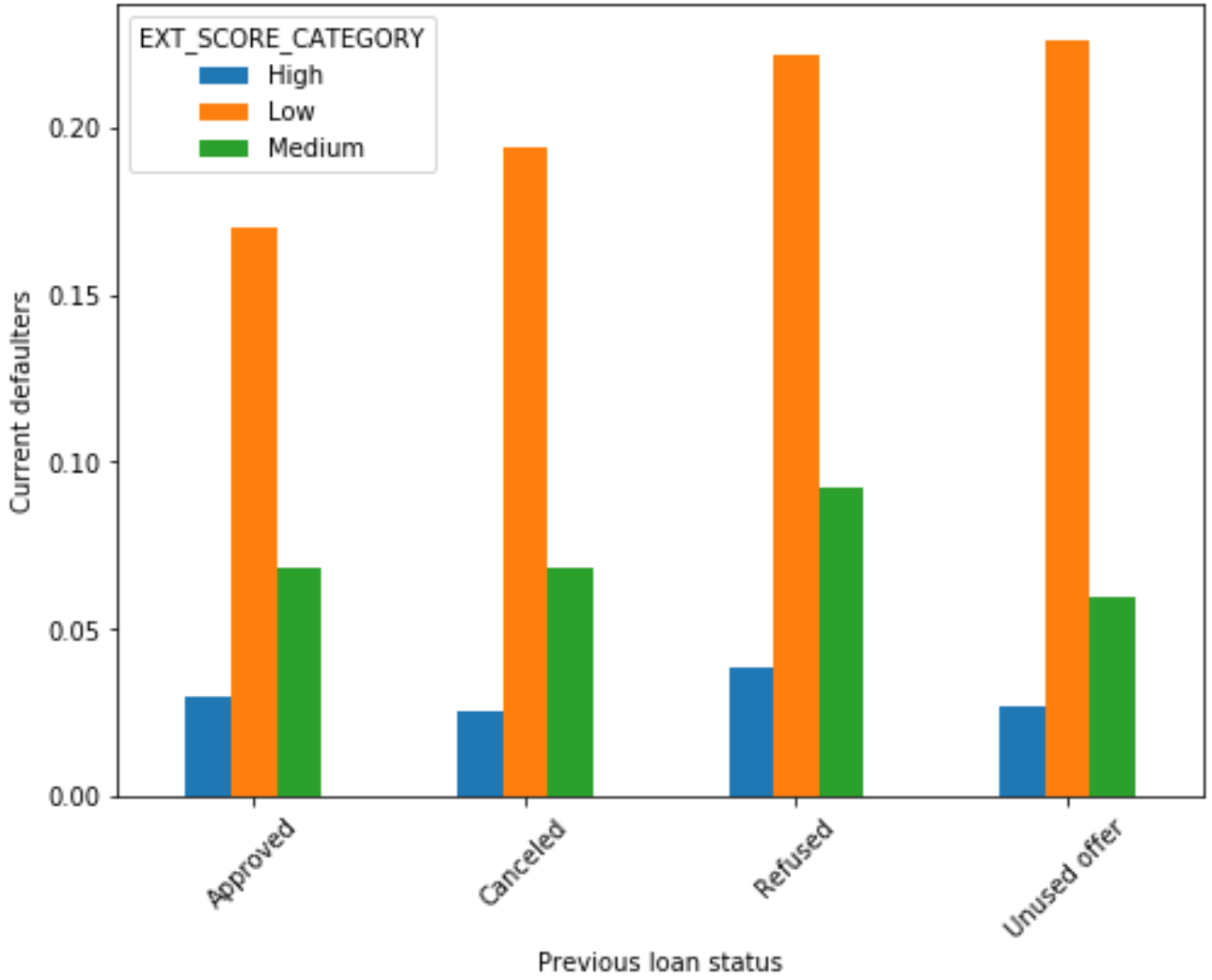
- The things we notice is that the majority of the name portfolio is of the POS type which is followed by the CASH type.
- The next categories which we notice are the XNA and CARDS type portfolio type.
- We can clearly see the preference of the customers as they are least in XNA and CARDS type transaction.
- Priorities are the CASH and POS type columns.

BROAD COMPARISON OF CURRENT LOAN DEFAULTERS WITH RESPECT TO PREVIOUS CUSTOMERS AND THE CATEGORY OF THE CLIENT



- The graph from the previous slide on analysing states that in the approved defaulters the new clients are more in default types.
- Also in the refused defaulters XNA are very large in number and new clients are very less in number.
- In the cancelled defaulters refreshed customers are least in number and again new defaulters are very large in number.

BROAD COMPARISON OF CURRENT LOAN DEFAULTERS WITH RESPECT TO PREVIOUS CUSTOMERS AND THE EXTERNAL SOURCE CATEGORY



- We notice from the previous slide that low external source scores are highly defaulted be it any of the previous loan status.
- High external source scores are still in a better stage as compared to the previous default status.
- We can make a note of that analysis is that we can grant loans to those people who have low external sources on various terms and conditions.
- We can also take a note on the bank status.

THANK YOU!