CREDIT EDA CASE STUDY

By Chirag Rana

PROBLEM STATEMENT

- 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 utilise this knowledge for its portfolio and risk assessment.

APPROACH

- Importing Libraries and Loading data.
- Describe Data, View info, Datatypes.
- Check for NULL values and REMOVE columns with > 50% NULL values.
- Analyse columns and drop them or impute them accordingly.
- Drop columns which seems to be insignificant.
- Find incorrect values in columns by using value_counts() and impute them.
- Find outliers in the data and impute or drop them.
- For Numerical data, Plot a scatter plot and if outliers, then impute null values with median and if no outliers, then impute the null values with mean.
- For Categorical data, mode of the column can be used to impute.
- Plot different types of graphs for univariate analysis.
- Segment the data into 2 dataset on Target variable for better analysis.
- Plot graphs for both Defaulters and Non-Defaulters for comparison and analysis.
- Then perform bivariate analysis on different variables.

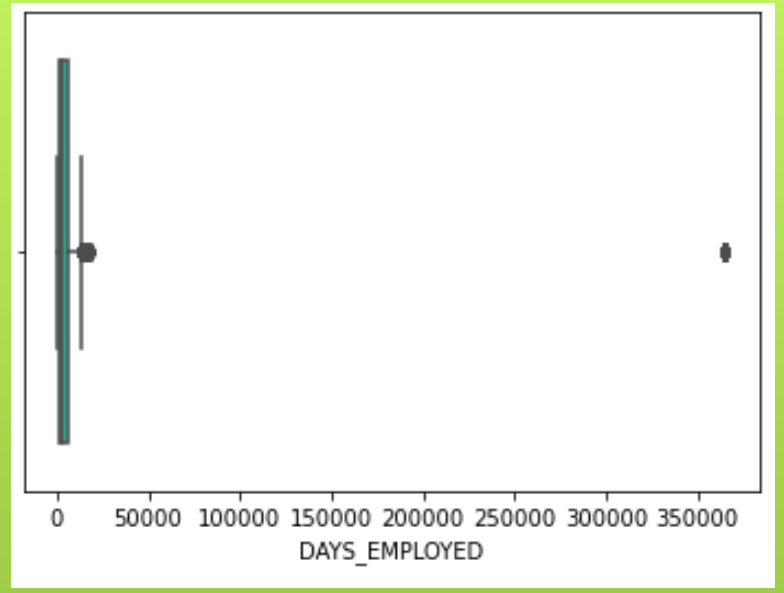
APPROACH

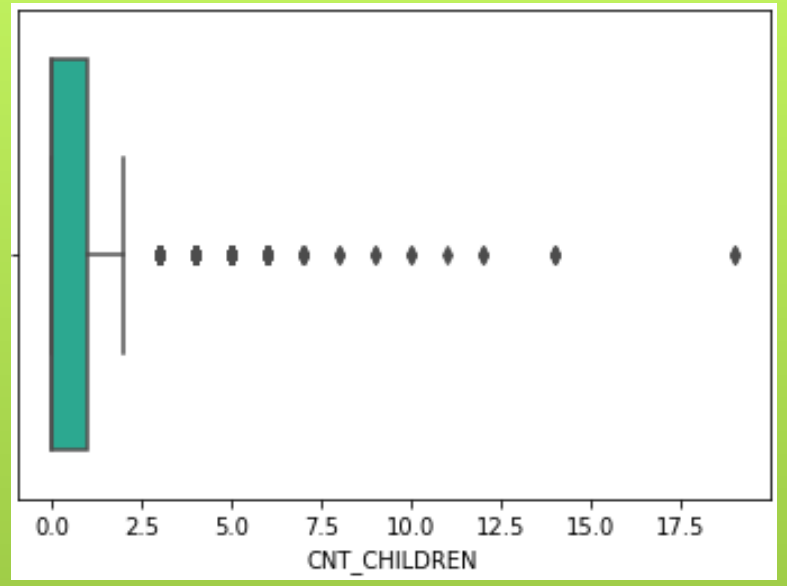
- Bar graphs and Box plots will be the best approaches to analyse the insights and comparison between different variables for categorical analysis.
- For Numerical vs Numerical analysis, Pair Plot can be referred.
- Heat Maps can be used for finding the correlation between variables.
- Then make insights from the plots about the defaulters.
- Load the Previous Dataset and all these operations can be performed on this dataset also like cleaning the data and dropping and imputing null values and handling outliers.
- Then Merge the two dataset on Application ID and do some cleaning if required.
- Perform analysis by plotting graphs and get insights for defaulters and what factors affect the person from defaulting and which factors are considered good for non difficulty payments.

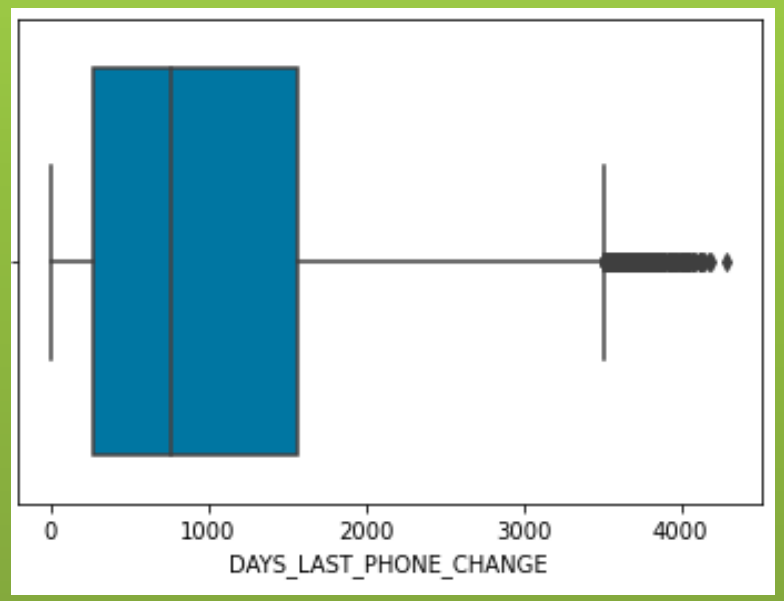
HANDLING OUTLIERS

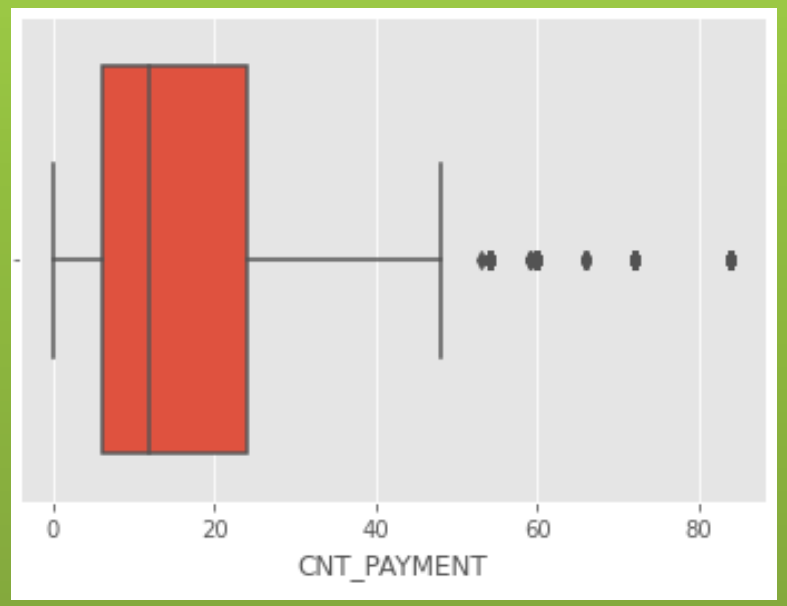
Points to be concluded from the pie-chart on the right side.

- There were outliers observed in the dataset.
- They were DROPPED in some cases like DAYS_EMPLOYED
- Some were treated by imputing outliers with mean, median etc.
- They were treated by 1.5IQR Method in some cases and replaced with a fixed value.







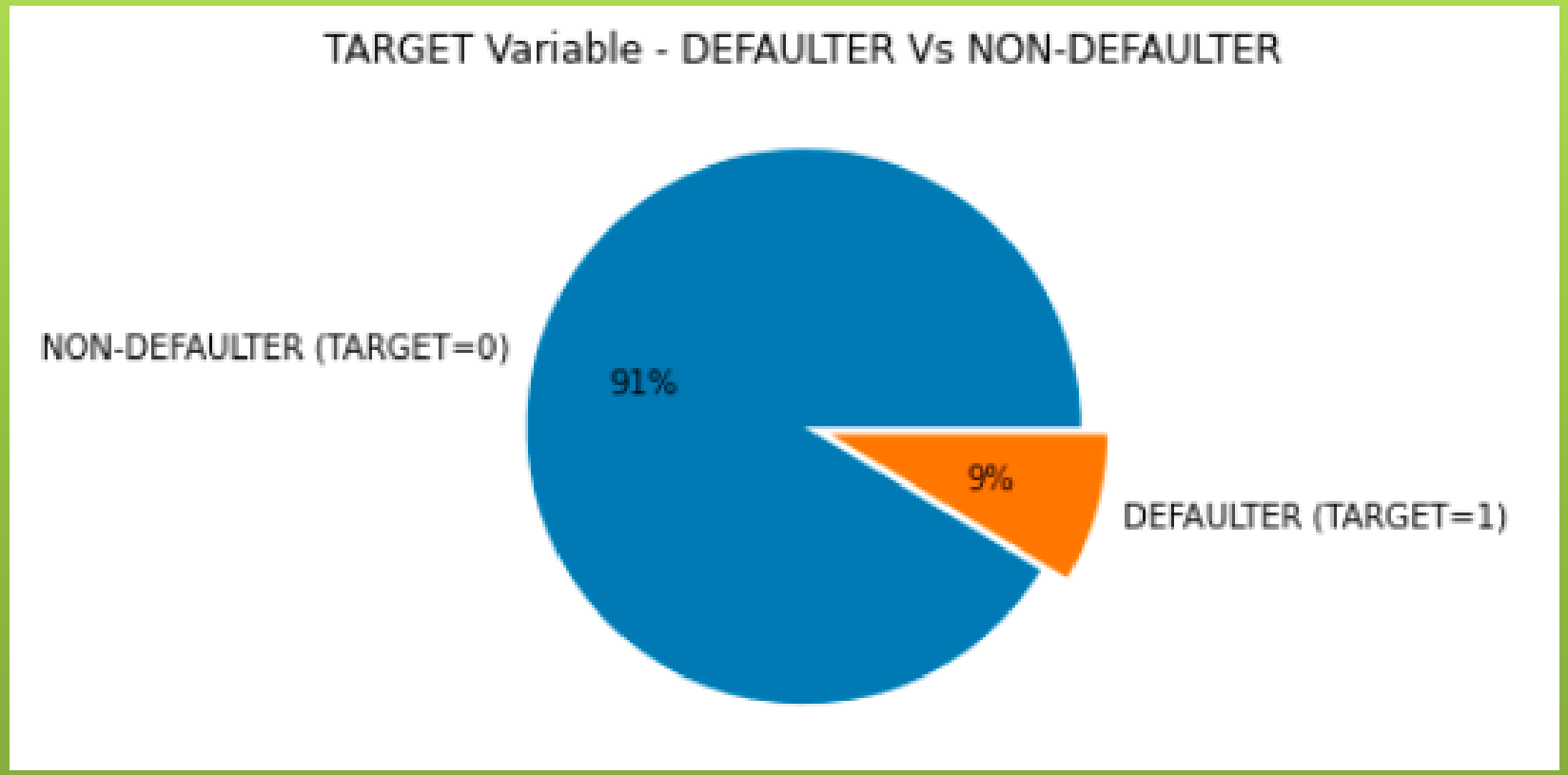


RESULTS

EXPLORATORY DATA ANALYSIS

Points to be concluded from the pie-chart on the right side.

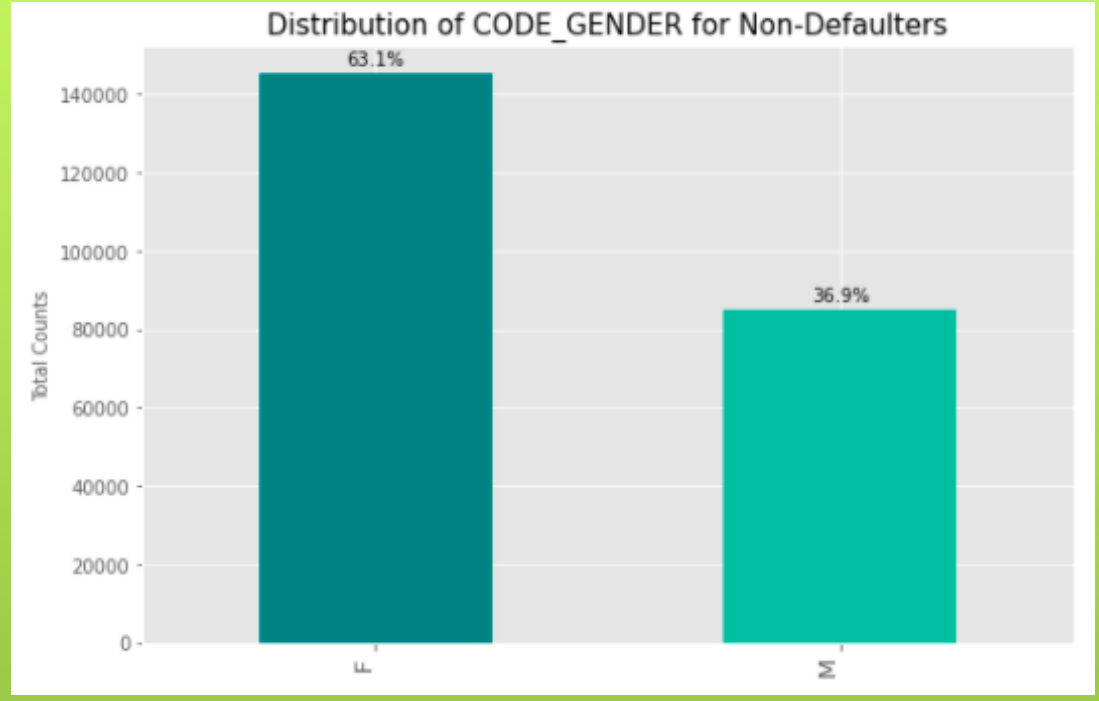
- There is an imbalance in the dataset.
- Number of NON-DEFAULTERS has very high percentage.
- Number of DEFAULTERS have very less percentage.
- The dataset needs to be segmented for correct analysis.

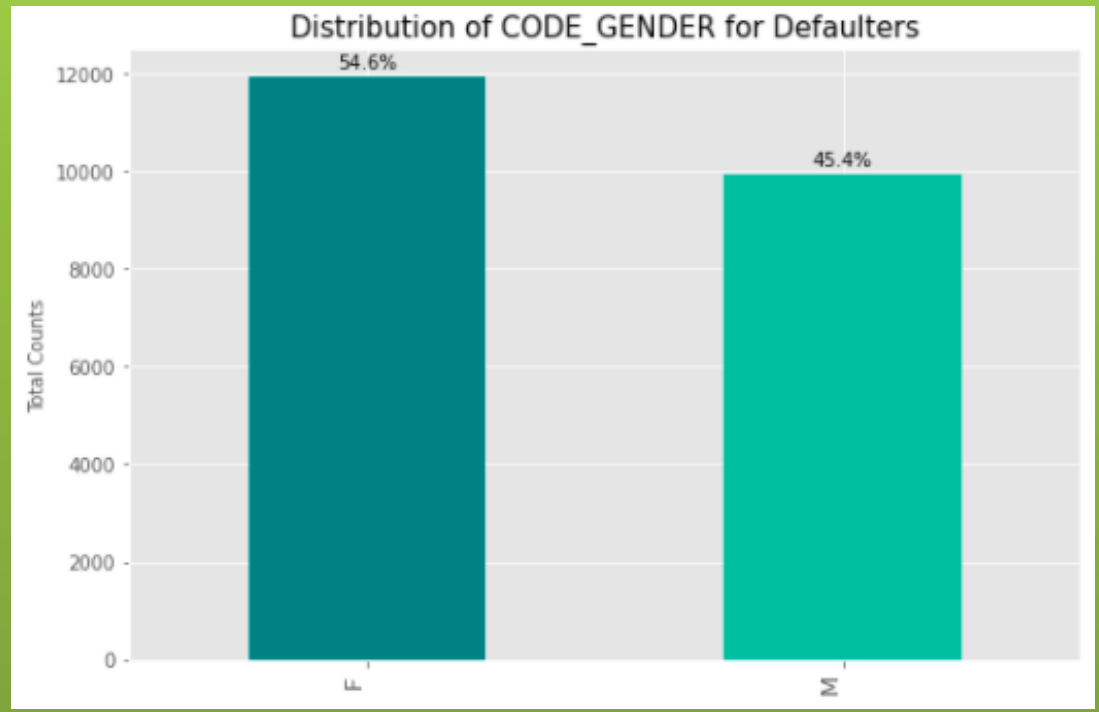


Distribution of GENDER

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

- Female contribute 63% to the non-defaulters and 55% to the defaulters.
- We see more female applying for loans than males and hence the more number of female defaulters as well.
- But the rate of defaulting of FEMALE is much lower as compared to the MALE.

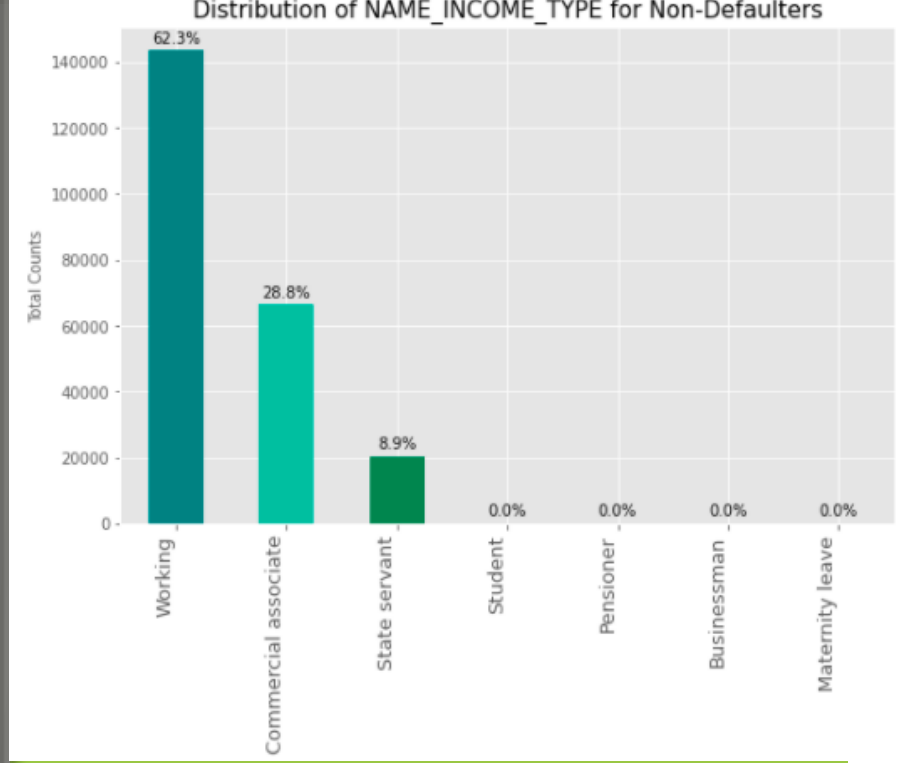


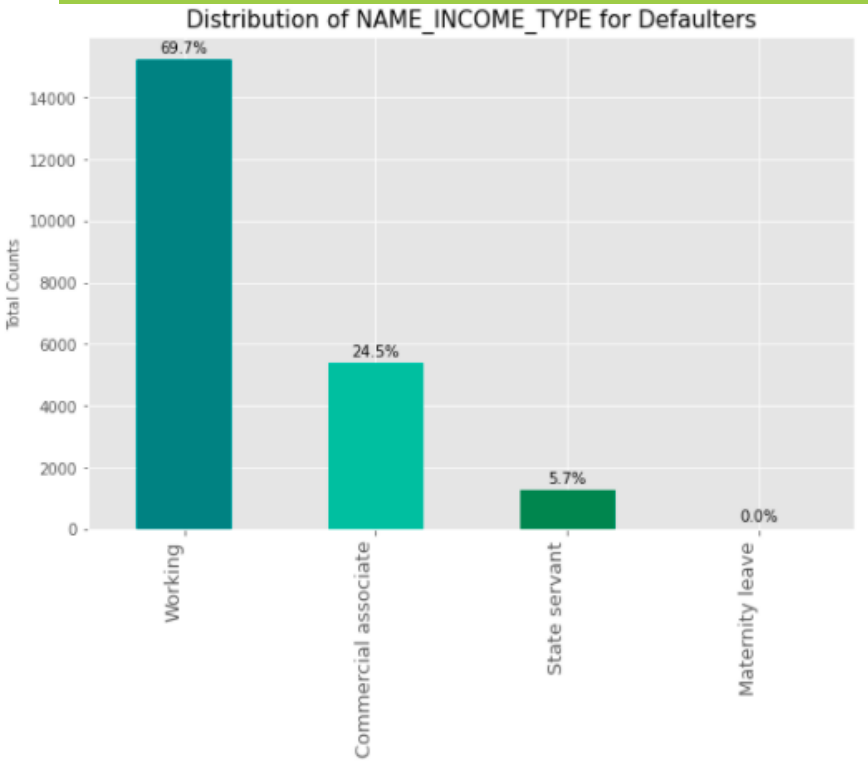


Distribution of INCOME TYPE

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

- There is a decrease in the percentage of Defaulters who are pentioners
- There is a increase in the Defaulters who are working.
- We can notice that the students and BusinessMen don't default.

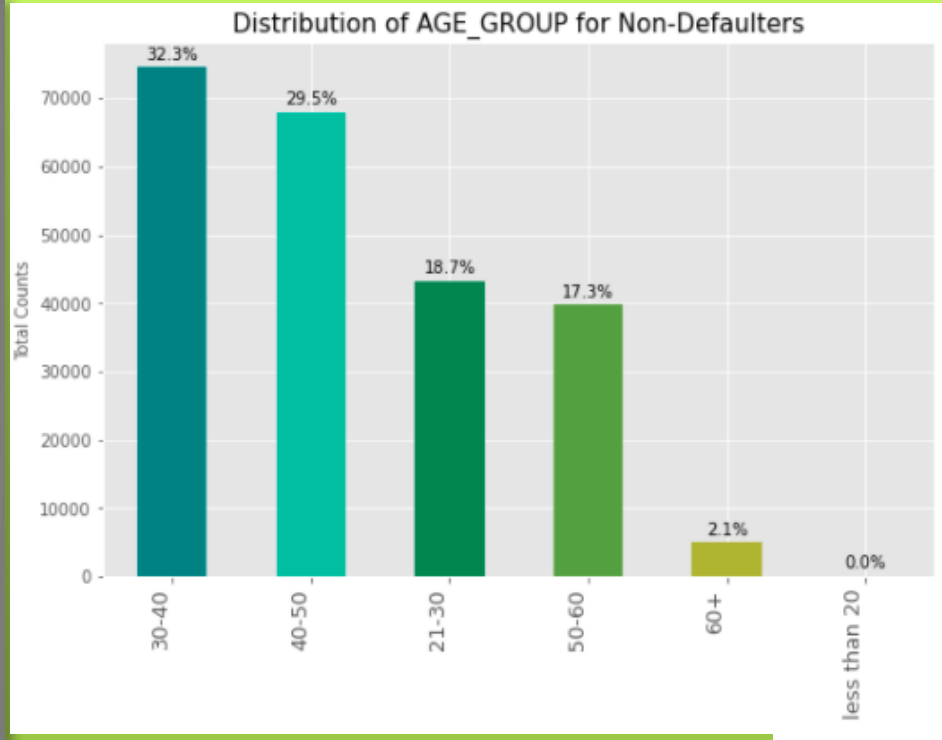


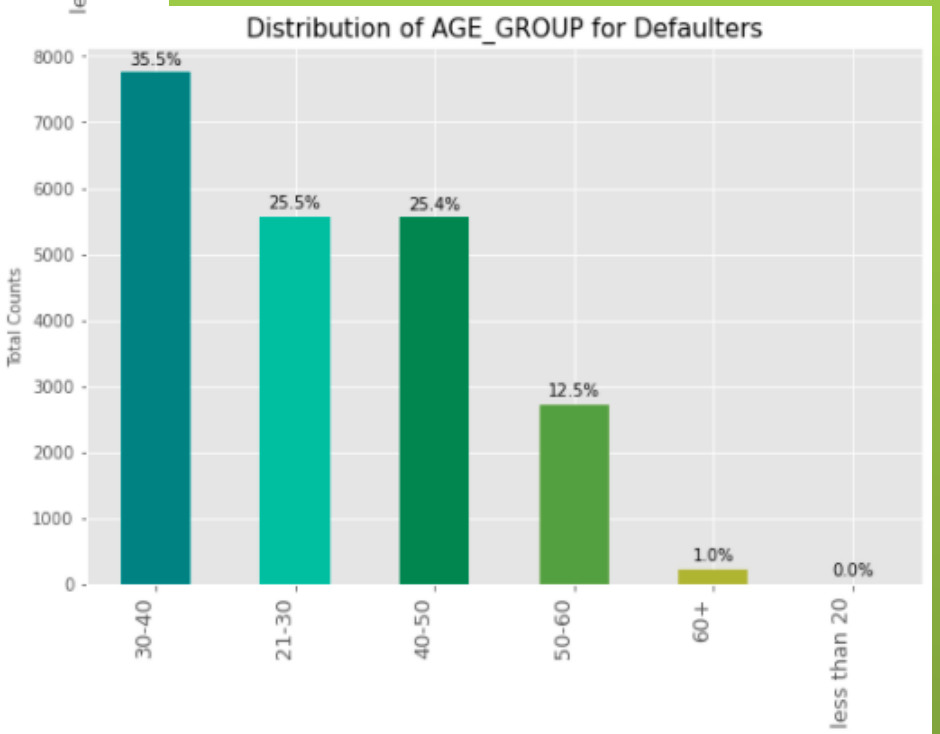


Distribution of AGE GROUP

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

- People in age between 20 - 40 age group tend to default more often as there is an increase in the percentage of defaulters.
- So they are the riskiest people to loan to.

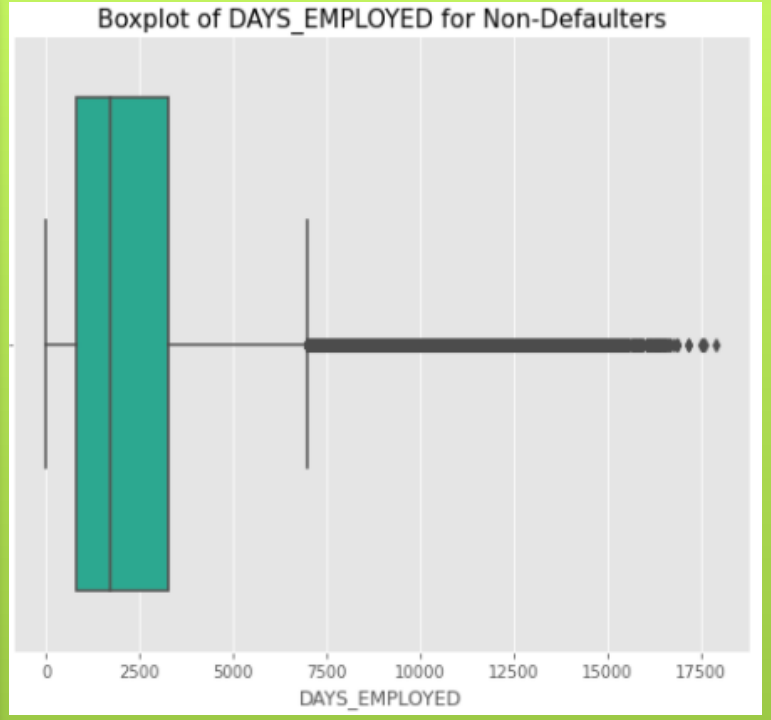


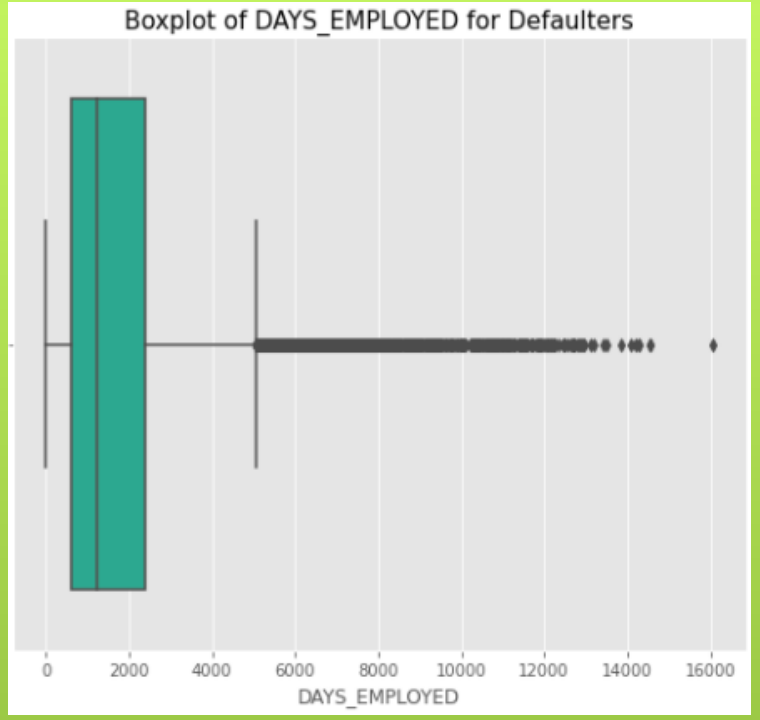


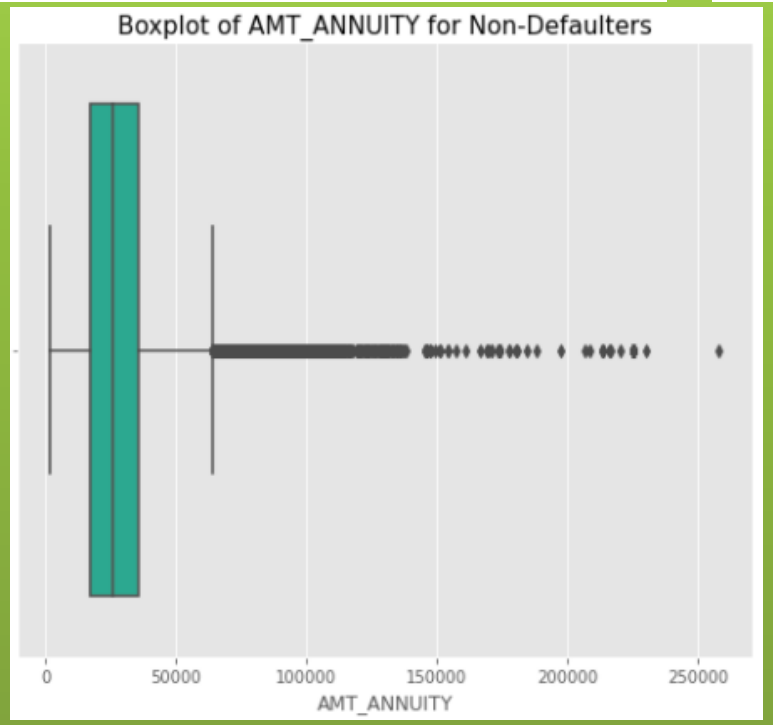
BOX-PLOT for Defaulter Percentage for DAYS EMPLOYED and AMOUNT ANNUITY

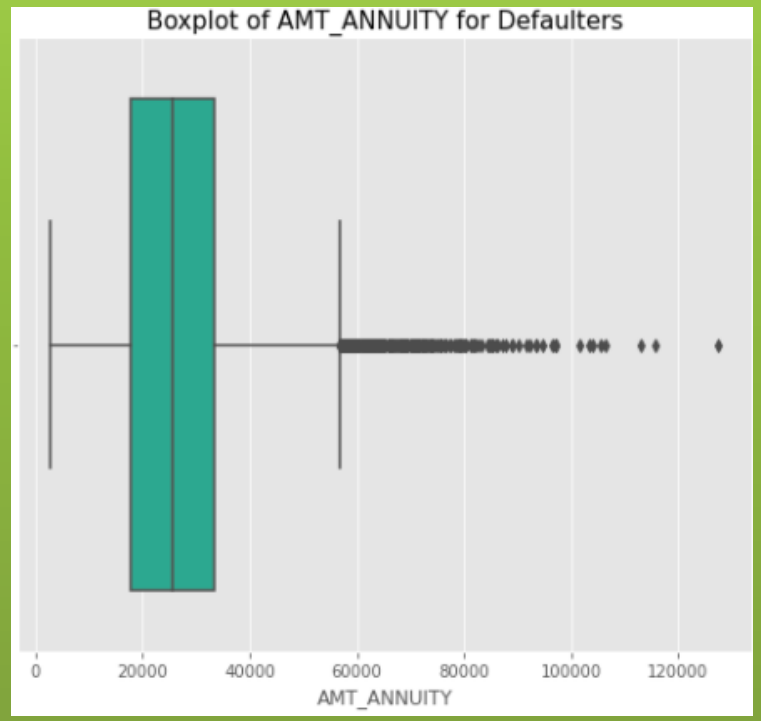
Points to be concluded from the graph on the right side.

- There are some outliers and the third quartile is bigger than first quartile which means most of the client's employment years are from third quartile.
- There are some outliers and the first quartile is bigger than third quartile which means most of the credit amount falls in first quartile





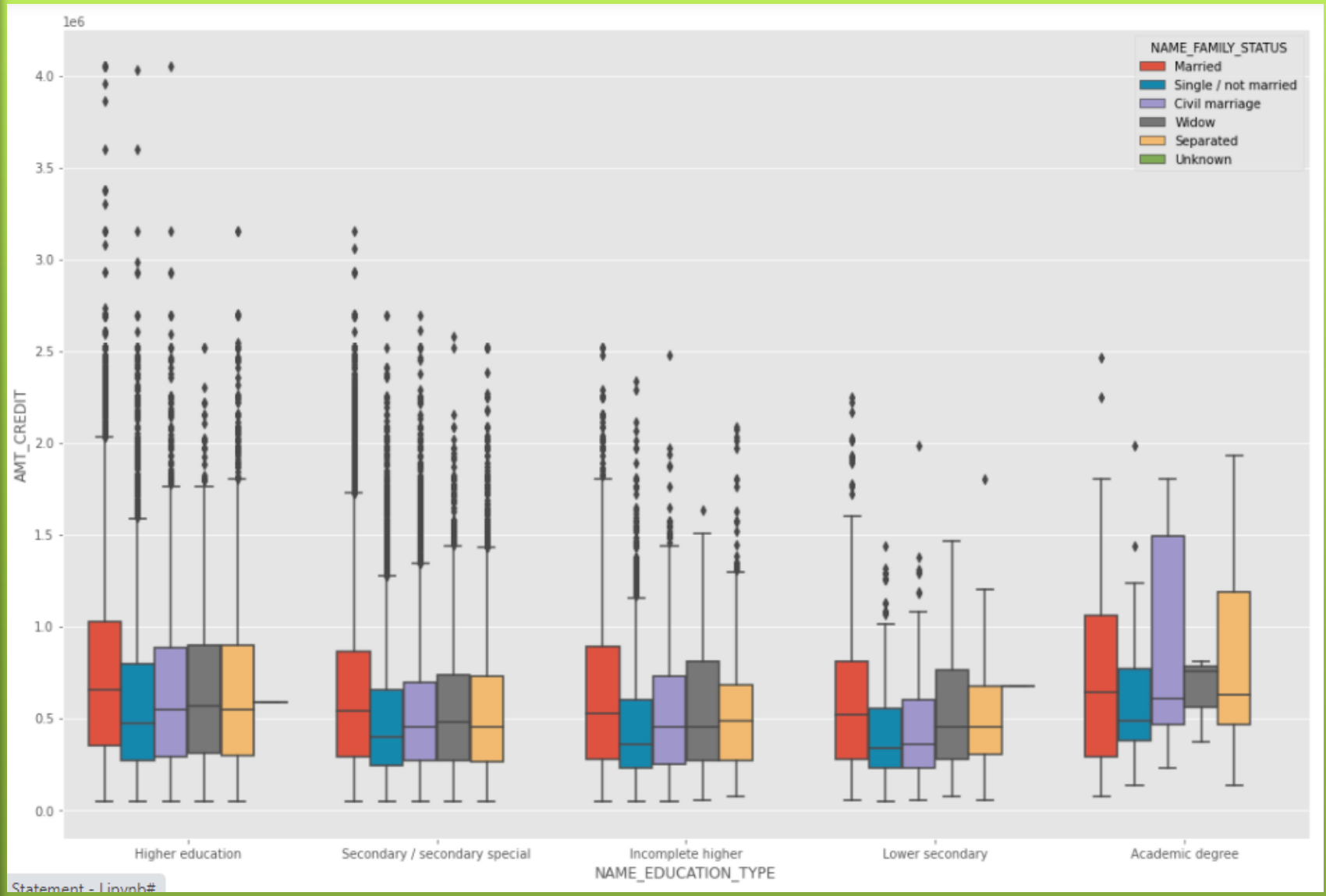




Bivariate Analysis of EDUCATION vs CREDIT AMOUNT

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

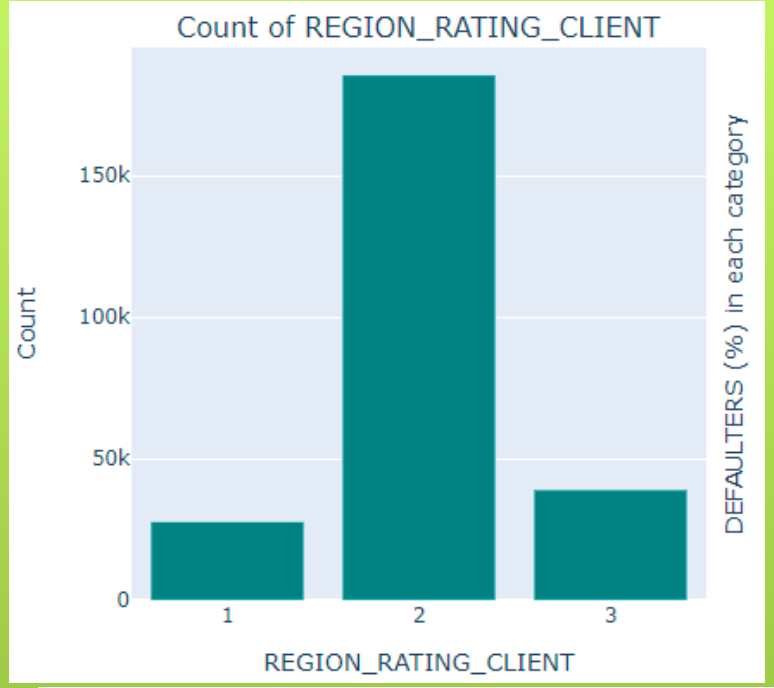
- Higher education of family status of 'marriage', 'single' and 'civil marriage' are having more outliers.
- Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others.
- Civil marriage for Academic degree is having most of the credits in the third quartile.

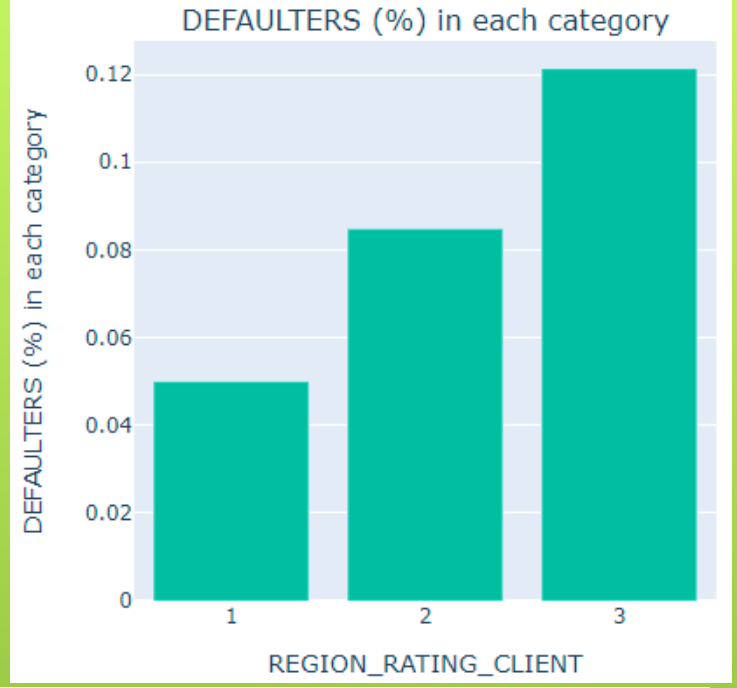


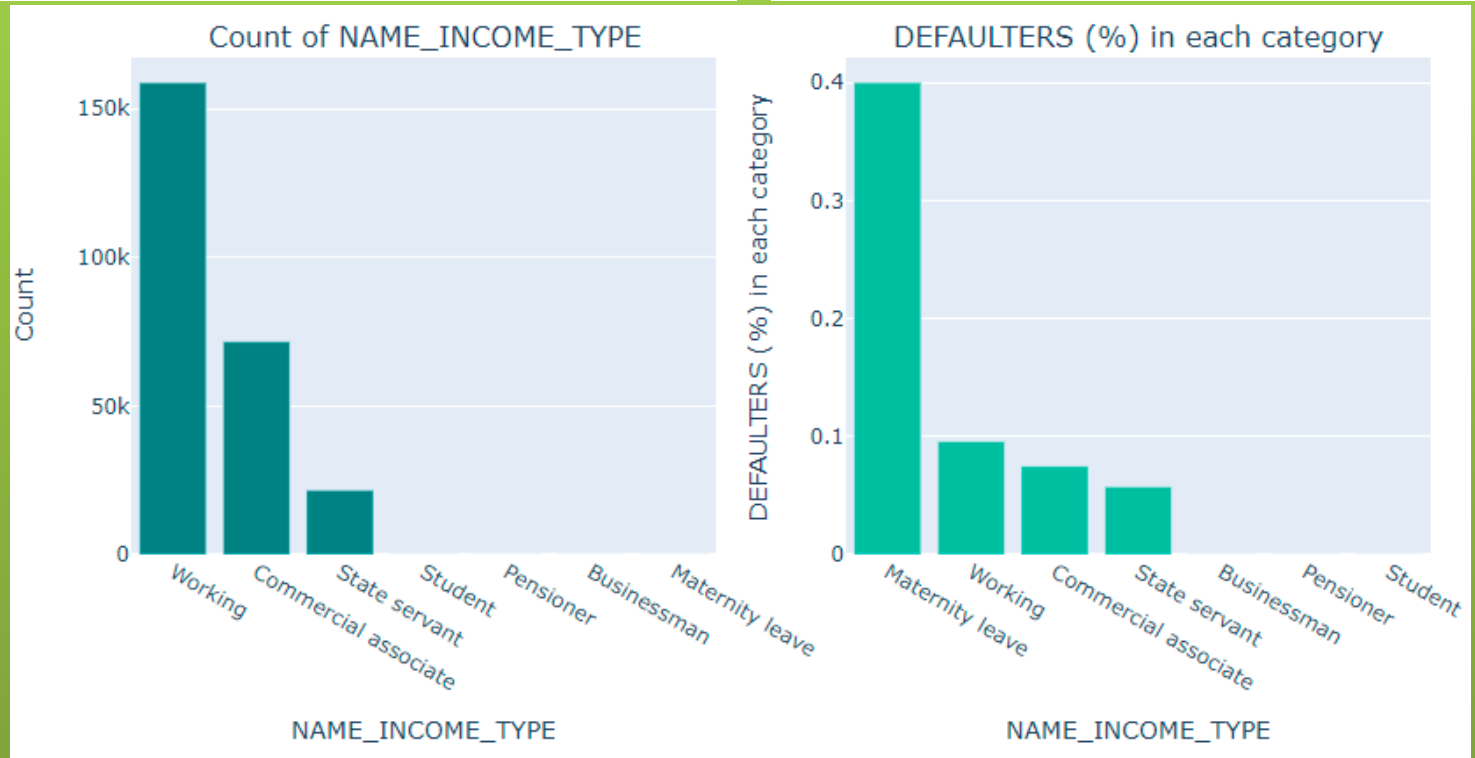
Defaulter Percentage for REGION RATING and INCOME TYPE

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

- People from Region with rating 3 are more likely to default
- Working people are the least on defaulter side and People on Maternity leave are more tend to default.



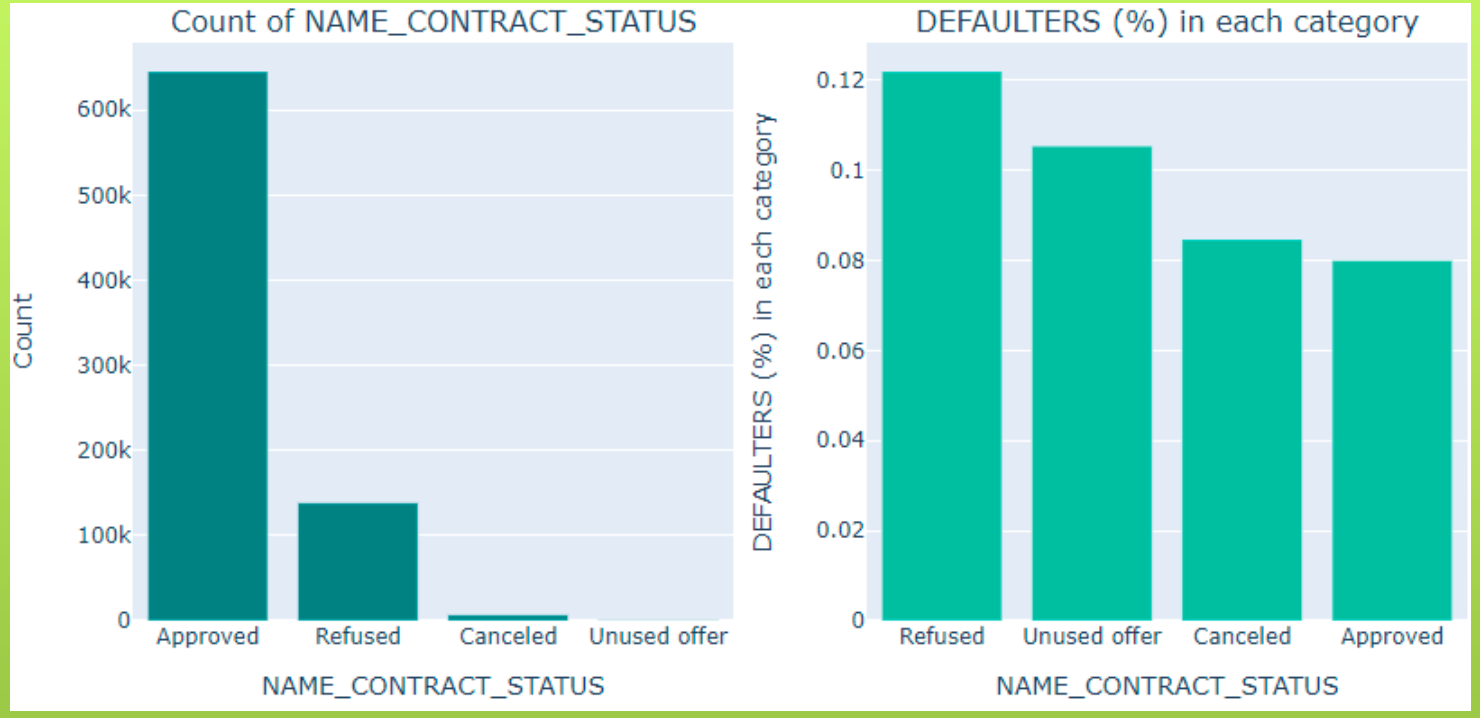


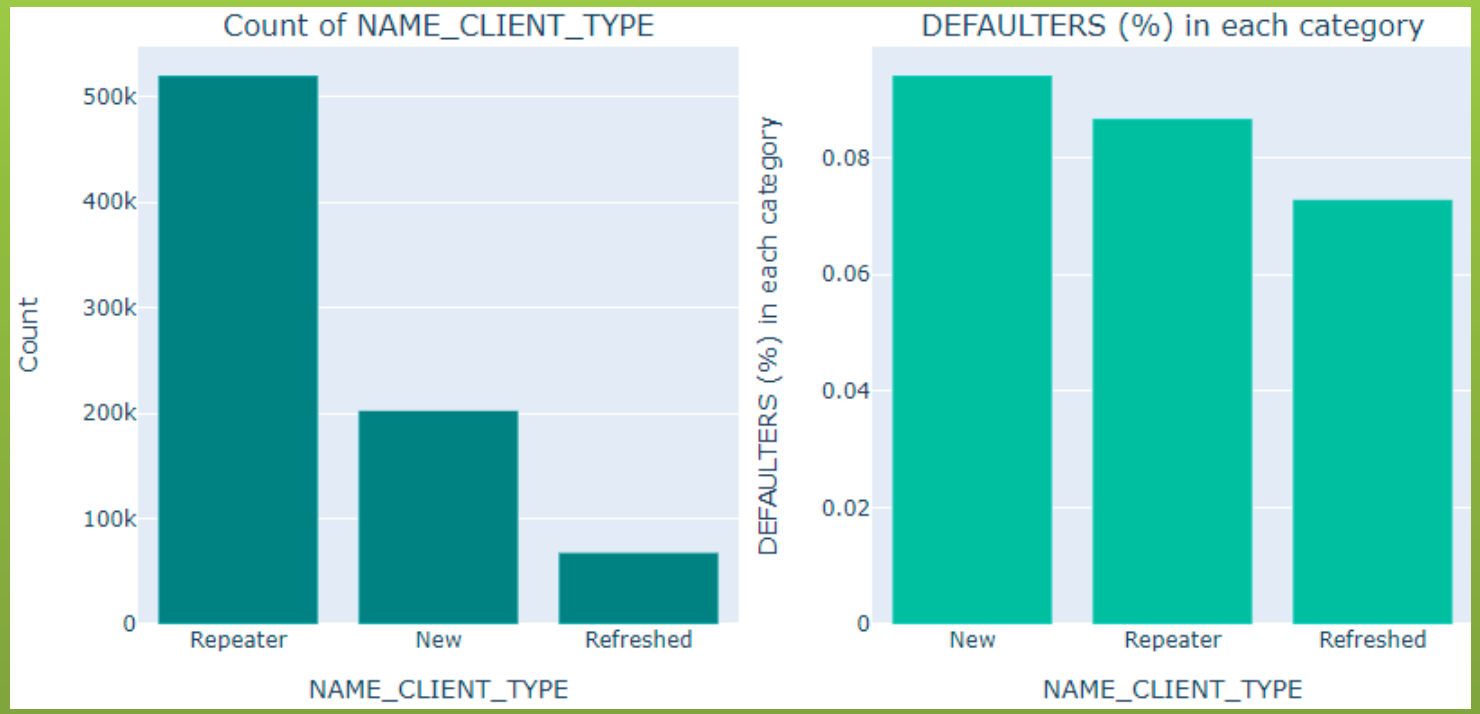


Defaulter Percentage for CONTRACT STATUS and CLIENT TYPE

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

- Approved applications have a higher weightage in previous application.
- People with refused offer have defaulted more
- Whereas people with approved offer have defaulted less.
- 'New' and 'Refreshed' clients tends to default much more than 'Repeated' clients.

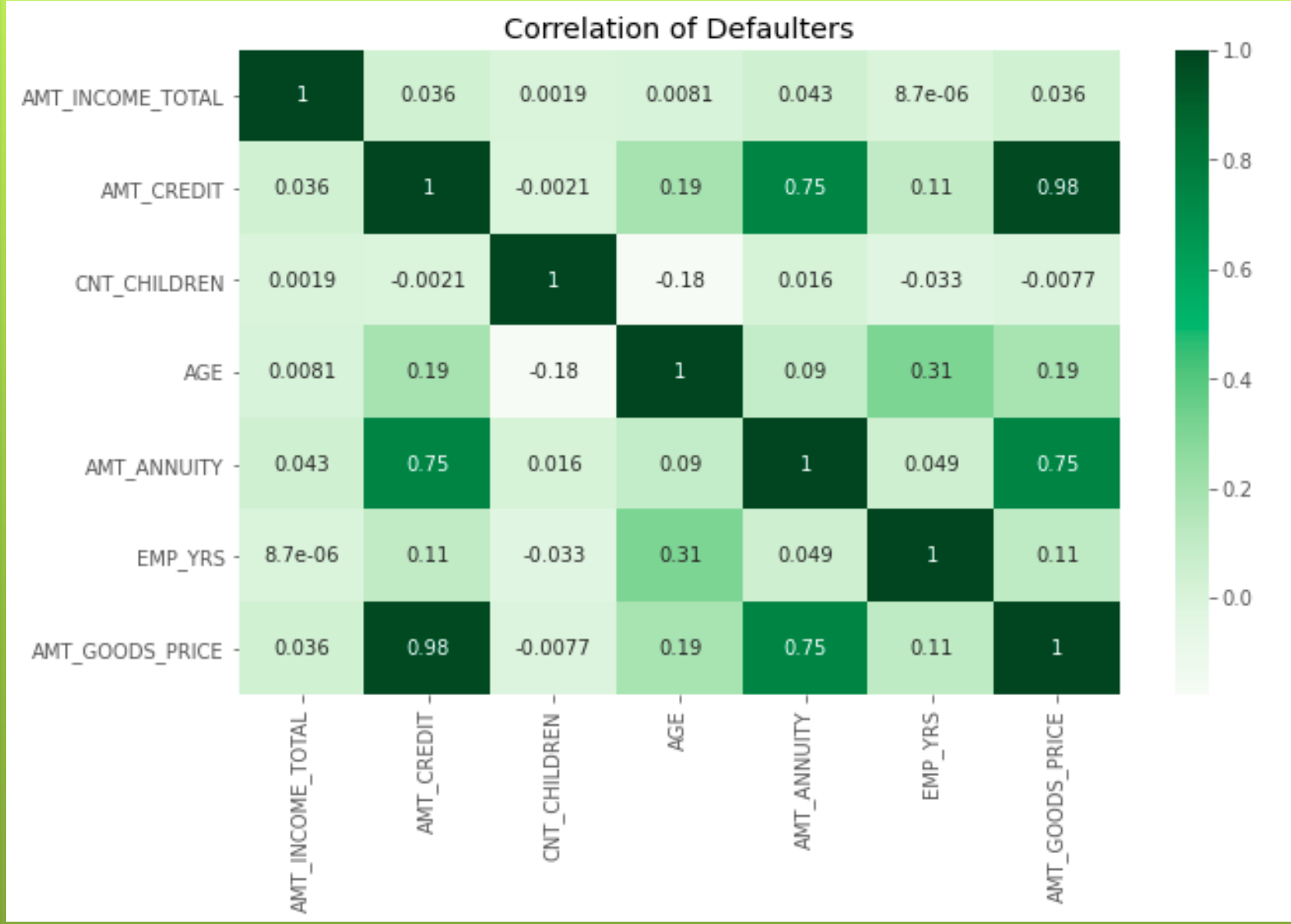




CORRELATION of DEFAULTERS

Points to be concluded from the heat-map on the right side.

- We discover high correlation between-
- 'AMT_CREDIT' vs 'AMT_GOODS_PRICE'
- 'AMT_GOODS_PRICE' vs 'AMT_ANNUITY'
- 'AMT_CREDIT' vs 'AMT_ANNUITY'



CONCLUSION

- Banks should focus more on contract type ‘Student’ ,’pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ for successful payments.
- Banks should focus less on income type ‘Working’ as they are having most number of unsuccessful payments.
- People with loan purpose ‘Repair’ is having higher number of unsuccessful payments on time.
- Clients from housing type ‘With parents’ are having least number of unsuccessful payments, so they can me focused on.
- ‘Maternity Leave’ in ‘NAME_INCOME_TYPE’ has very less count and it also has maximum % of payment difficulties, such cases can be avoided.
- ‘Low skilled Laborers’ in ‘OCCUPATION_TYPE’ has comparatively less count and it also has maximum % of payment difficulties, such cases can be avoided.

CONCLUSION

- ‘Lower Secondary’ in ‘NAME_EDUCATION_TYPE’ has comparatively very less count and it also has maximum % of payment difficulties, such cases can be avoided.
- People with very low income are tend to default more, so focus on higher salary slab people.
- Higher educated people are less likely to default. People with 'Lower' and 'Incomplete Higher' education are more in the defaulter list
- People living in areas - Rating 3 tend contribute more to the defaulters
- People applying for cash loans are tend to default more
- People from age group of 21-30 and 60+ are more likely to default
- We discover high correlation between 'AMT_CREDIT' vs 'AMT_GOODS_PRICE' and 'AMT_GOODS_PRICE' vs 'AMT_ANNUITY' and 'AMT_CREDIT' vs 'AMT_ANNUITY'

THANK YOU