CREDIT EDA CASE STUDY

Presented by:
- Praveen Kumar Shah (praveenshah231@gmail.com)

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.
- Top 10 correlation for the **Client with payment difficulties** and **all other cases** (Target variable).

UNDERSTANDING OF DATA SETS

- This dataset has 3 files as shown below:
- 'application_data.csv' contains all the information of the client at the time of application shows whether a client has payment difficulties.
- 'previous_application.csv' contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.
- 'columns_description.csv' is data dictionary which describes the information for the variables.

DATA CLEANING

- Firstly application_data.csv was loaded into Jupyter notebook
- Inspect the dataframe for dimensions, null-values, and summary of different numeric columns.
- Look for columns that makes sense and drop the unnecessary columns.
- Check the number of rows and columns in the dataframe
- Check the column-wise info of the dataframe
- Check the summary for the numeric columns
- Checking the data types of their variables

DATA CLEANING

- Dropping the irrelevant columns
- Initially 50 % of null values were dropped but it throw some irrelevant columns
- So 45 % is selected for dropping
- Later 13 % null value analyzation

COLUMNS THAT HAVE MISSING VALUES PERCENTAGE GREATER THAN 13%

OCCUPATION_TYPE	31.35
EXT_SOURCE_3	19.83
AMT_REQ_CREDIT_BUREAU_HOUR	13.50
AMT_REQ_CREDIT_BUREAU_DAY	13.50
AMT_REQ_CREDIT_BUREAU_WEEK	13.50
AMT_REQ_CREDIT_BUREAU_MON	13.50
AMT_REQ_CREDIT_BUREAU_QRT	13.50
AMT_REQ_CREDIT_BUREAU_YEAR	13.50

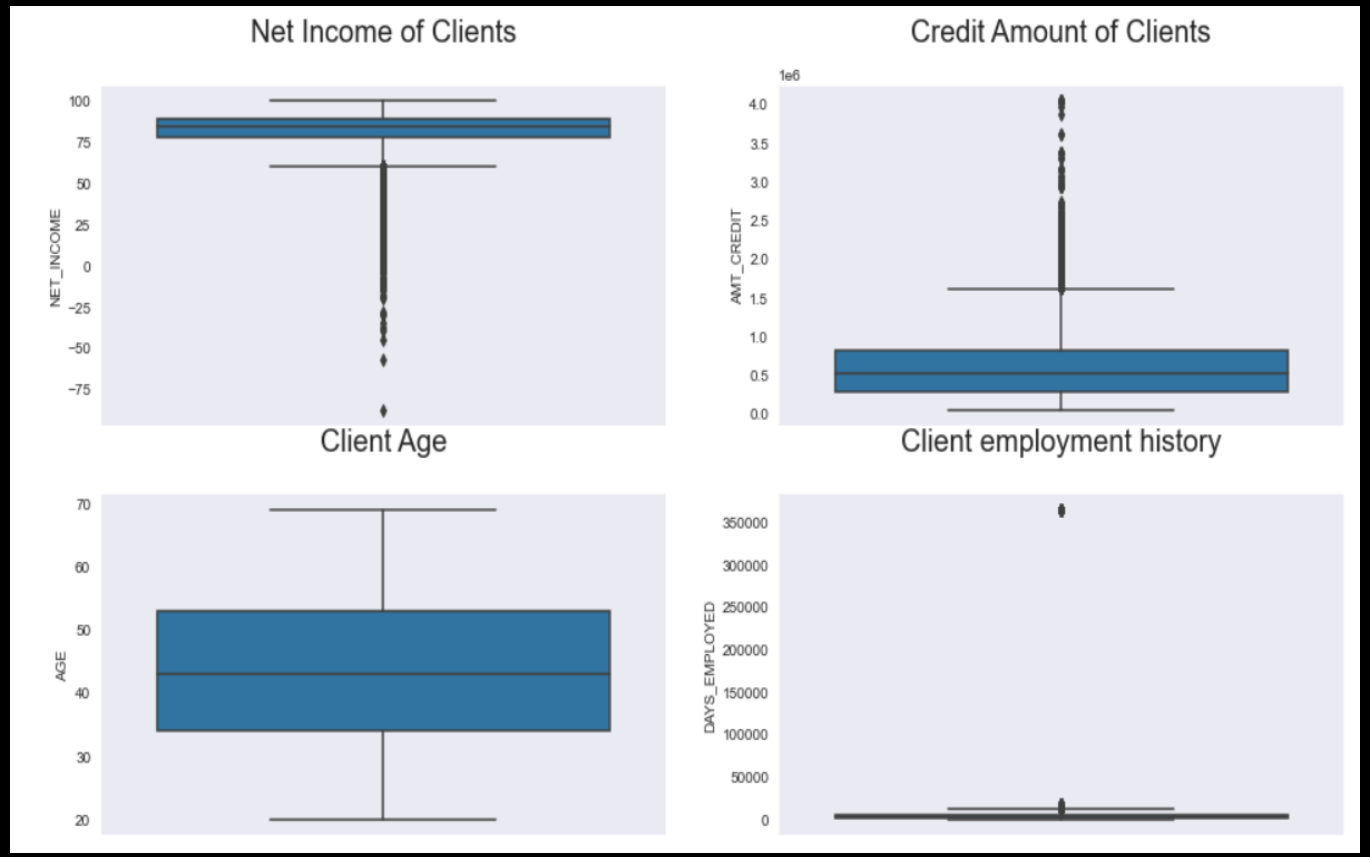
dtype: float64

- Only Occupation_type, ext_source_3 and AMT_REQ_CREDIT_BUREAU_YEAR were found relevant for further use.
- 31% of null values, though a categorical variable and will help during the data analysis. So, we can create new occupation_type as 'Others' and impute the null values.
- The data is evenly distributed close proximity to a normal distribution and null values of this column should be imputed with mean which comes out to be 0.51 for null values for EXT_SOURCE_3
- For AMT_REQ_CREDIT_BUREAU_YEAR we cannot use mean() in this case because this column only takes whole numbers in float form but mean is 1.9 however we can impute null values of this column using median which is 1.0

DAYS EMPLOYED, AMT_INCOME_TOTAL, DAYS_BIRTH, AMT_REQ_CREDIT_BUREAU_YEAR ATTRIBUTE

- New column net_income was created from AMT_INCOME_TOTAL - AMT_ANNUITY.
- Day employed was in negative before so changed it to positive.
- Day since birth has negative values so changed it to positive because it may be a data quality issue and created a new column age.
- So, Making all the values as positive.
- In AMT_REQ_CREDIT_BUREAU_YEAR was float64
- So, converted data type from float64 to int64

UNIVARIATE ANALYSIS



- Analyzing the attributes we can see that there are some extreme outliers points in Client Employment history and credit amount of Clients
- However the average Client Age before the loan application is found to around 44 years of age
- That signifies that mature people applies for more loan than the younger generation.

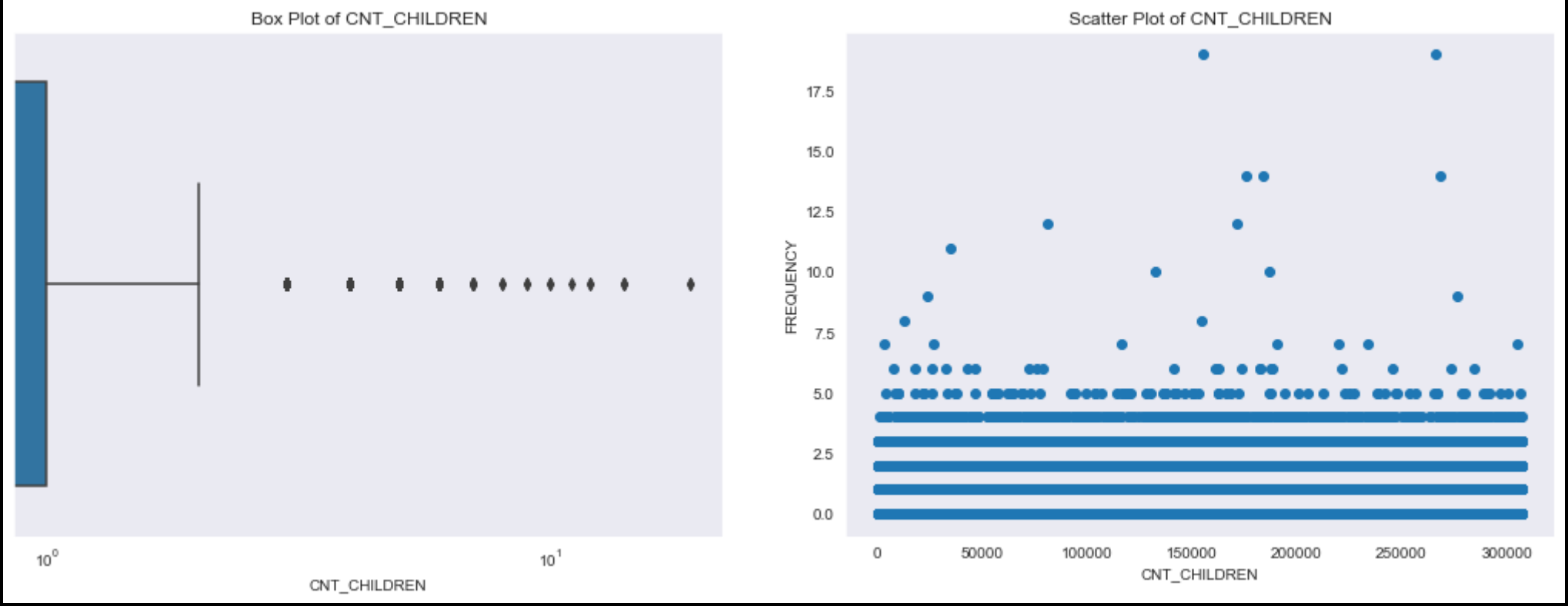
IMBALANCE IN DATA

- Clients with Payment Difficulties: 8.07%
- Other Clients: 91.93%
- Analyzed the imbalance in the target variable
- Response percentage: 8.8%
- Ratio of imbalance is 11.39
- 8.07% represents Clients with Payment Difficulties and 91.03% represents 'Other Clients' shows the imbalance percentage.
- So, imbalance ratio is 11.39, which means Other Clients are 11.39 times more than Clients with Payment Difficulties.

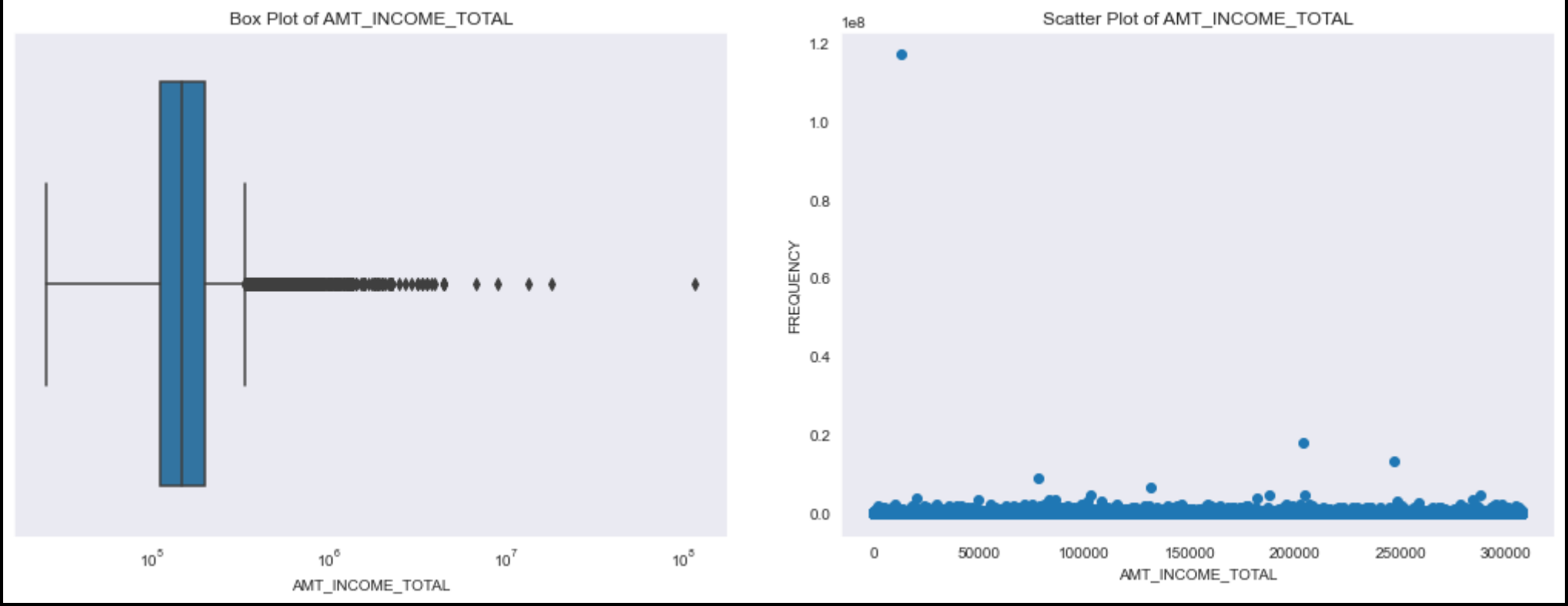
UNIVARIATE ANALYSIS FOR NUMERICAL COLUMNS

- Create two different data frames for target variable 1 and 0
- Analyzing numerical columns for outliers
- 'CNT_CHILDREN',
- 'AMT_INCOME_TOTAL',
- 'AMT_CREDIT',
- 'AMT_ANNUITY',
- 'DAYS_EMPLOYED',
- 'EXT_SOURCE_3',
- 'AGE‘ were selected columns.

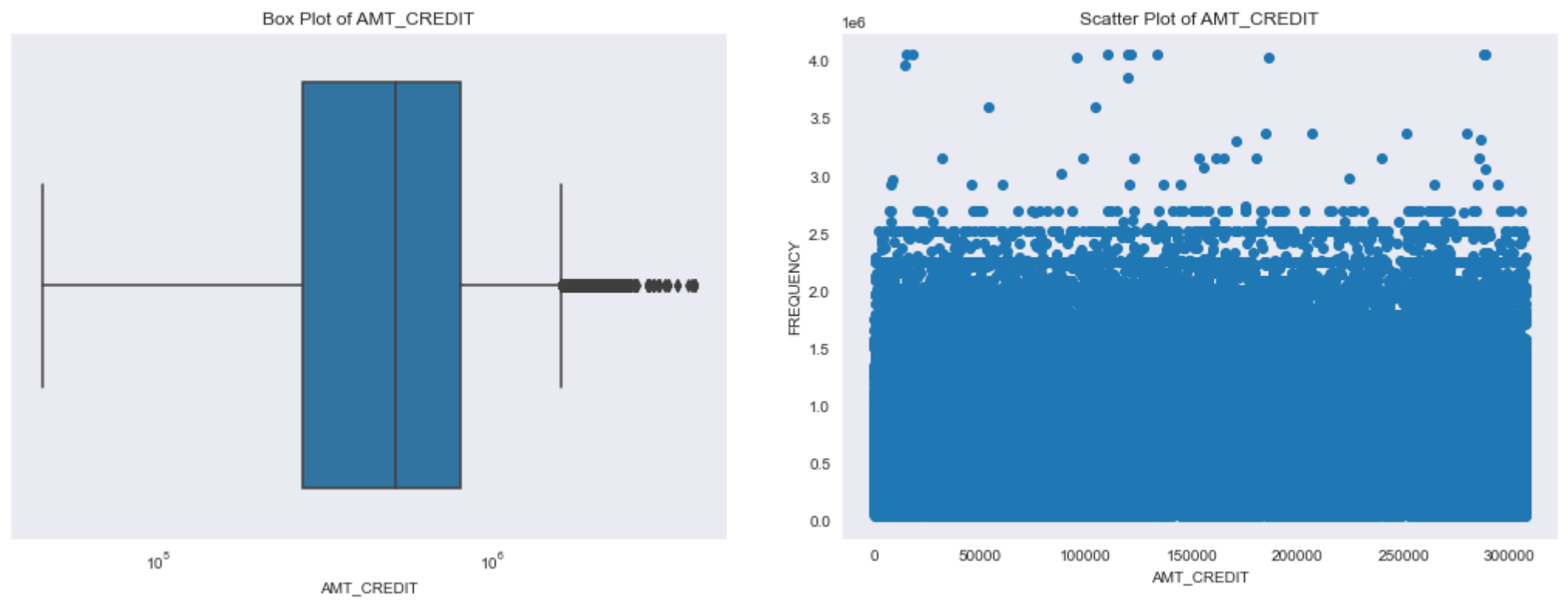
UNIVARIATE ANALYSIS FOR NUMERICAL COLUMNS



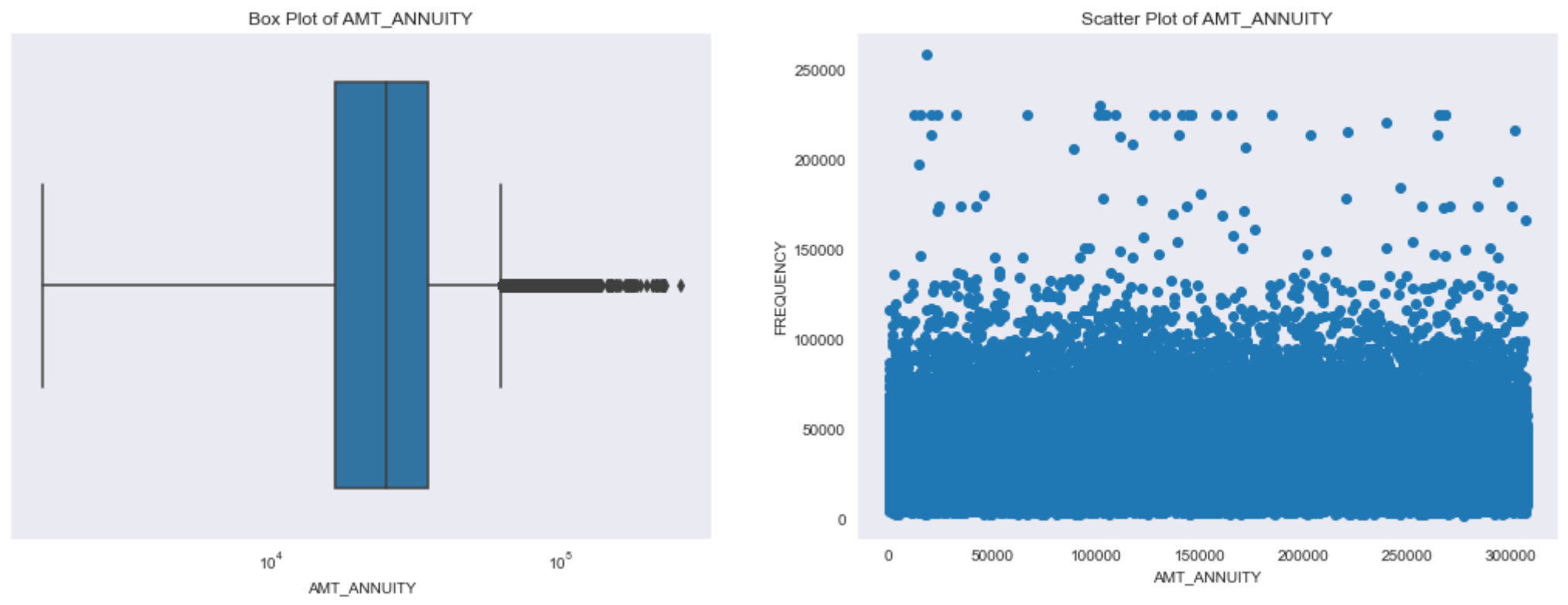
UNIVARIATE ANALYSIS FOR NUMERICAL COLUMNS



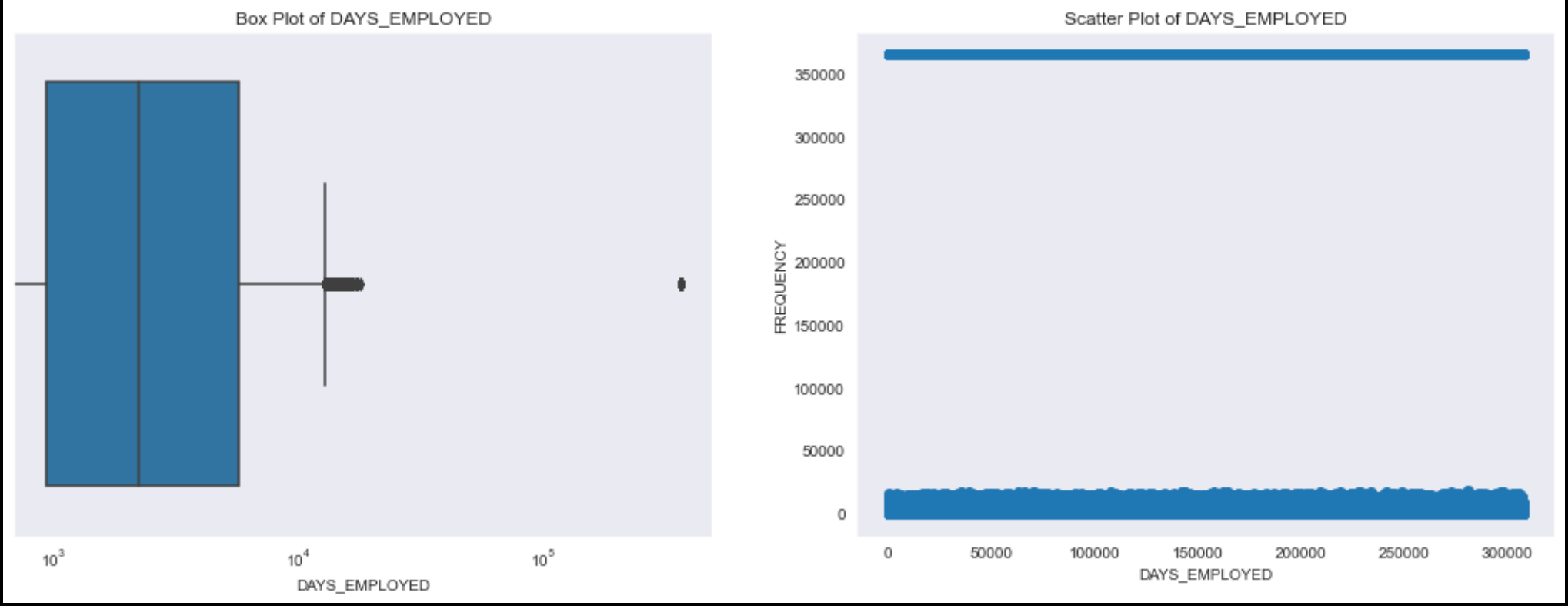
UNIVARIATE ANALYSIS FOR NUMERICAL COLUMNS



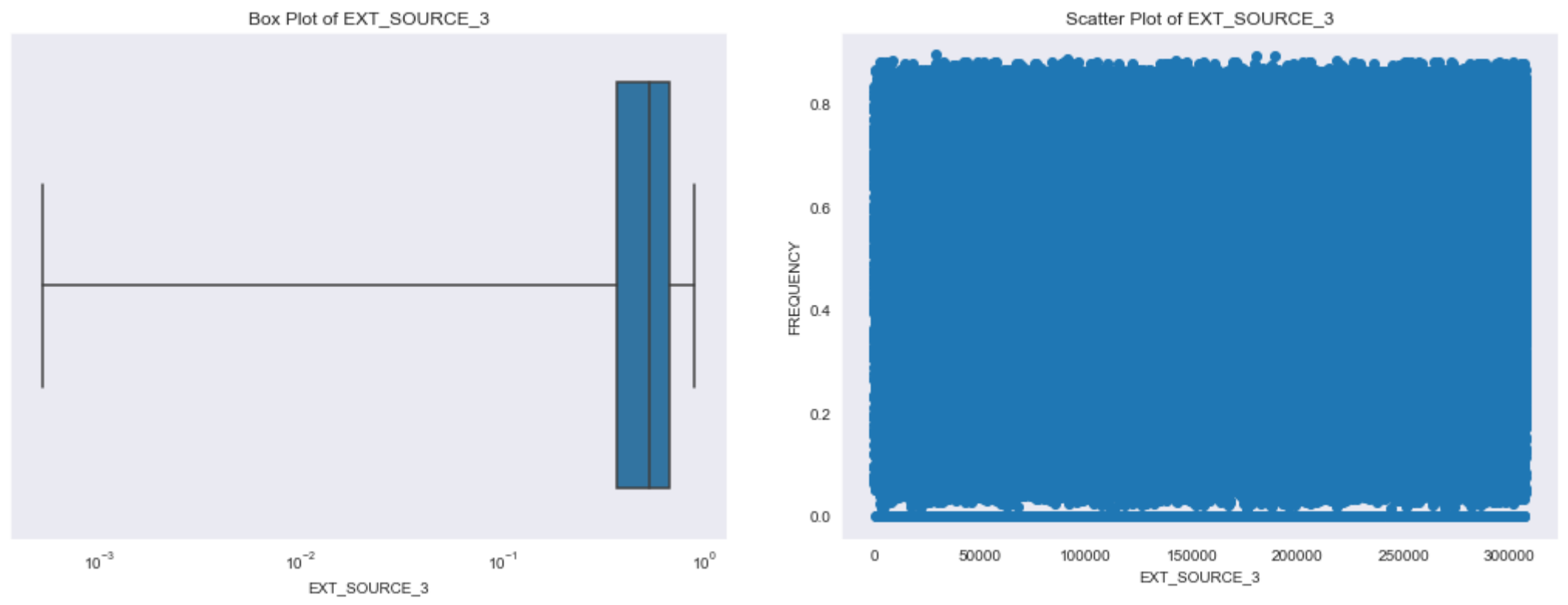
UNIVARIATE ANALYSIS FOR NUMERICAL COLUMNS



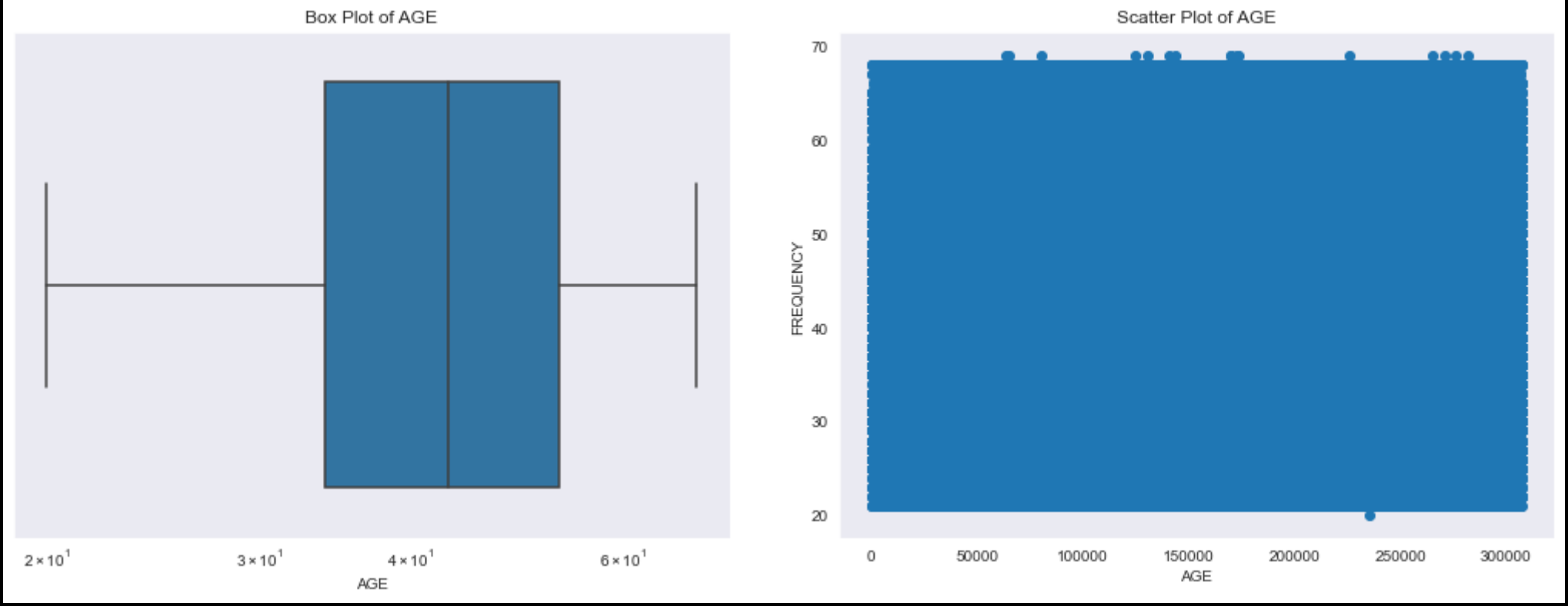
UNIVARIATE ANALYSIS FOR NUMERICAL COLUMNS



UNIVARIATE ANALYSIS FOR NUMERICAL COLUMNS



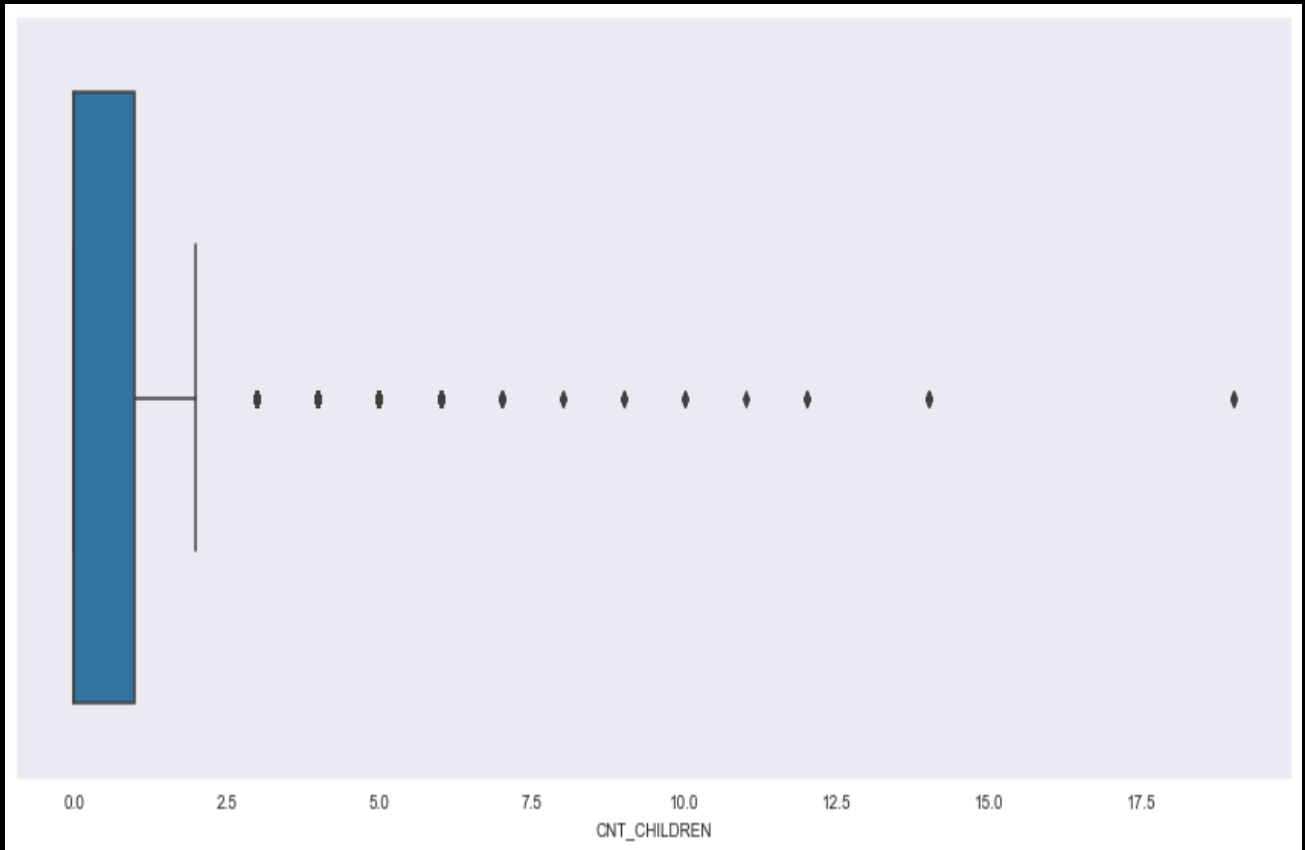
UNIVARIATE ANALYSIS FOR NUMERICAL COLUMNS



INFERENCES DRAWN

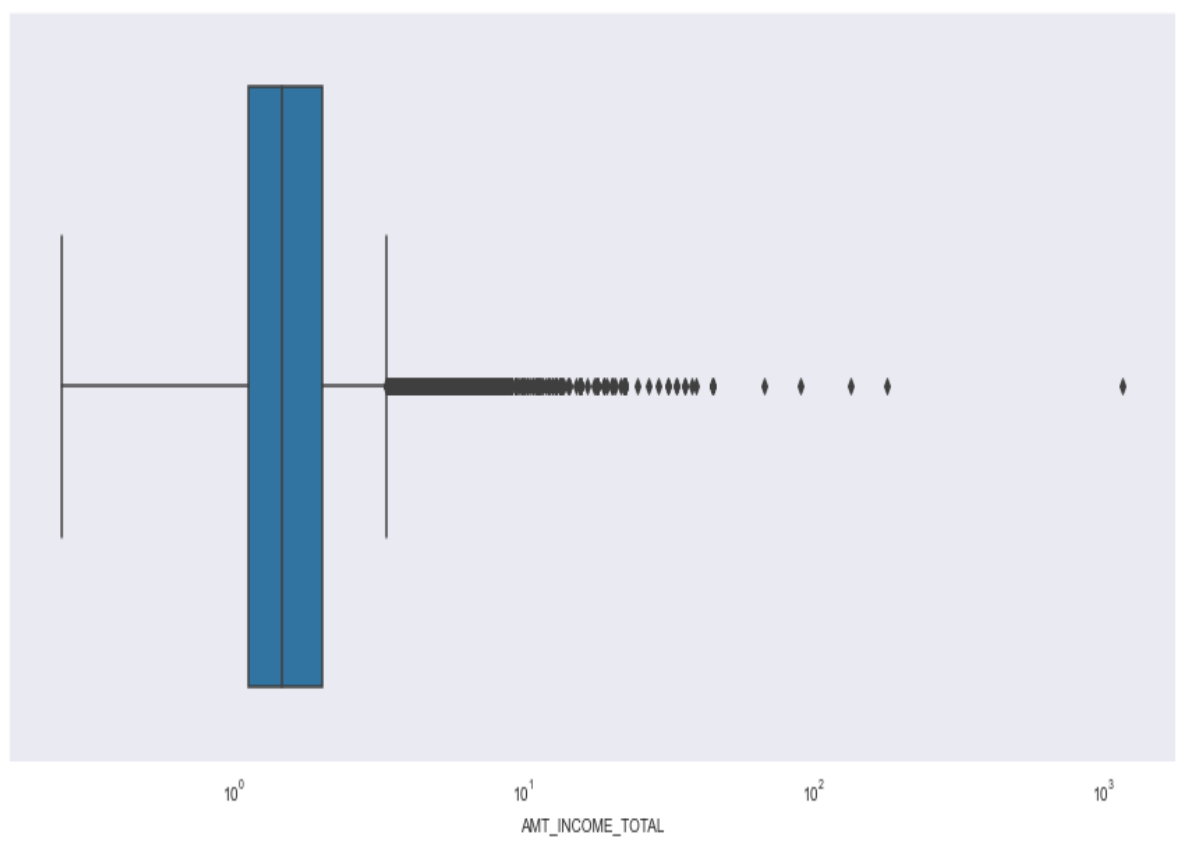
- From the above plots,
- DAYS_EMPLOYED column could not be concluded to have outliers.
- CNT_CHILDREN has outliers
- AMT_INCOME_TOTAL has outliers
- DAYS_EMPLOYED has outliers

TREATING OUTLIERS



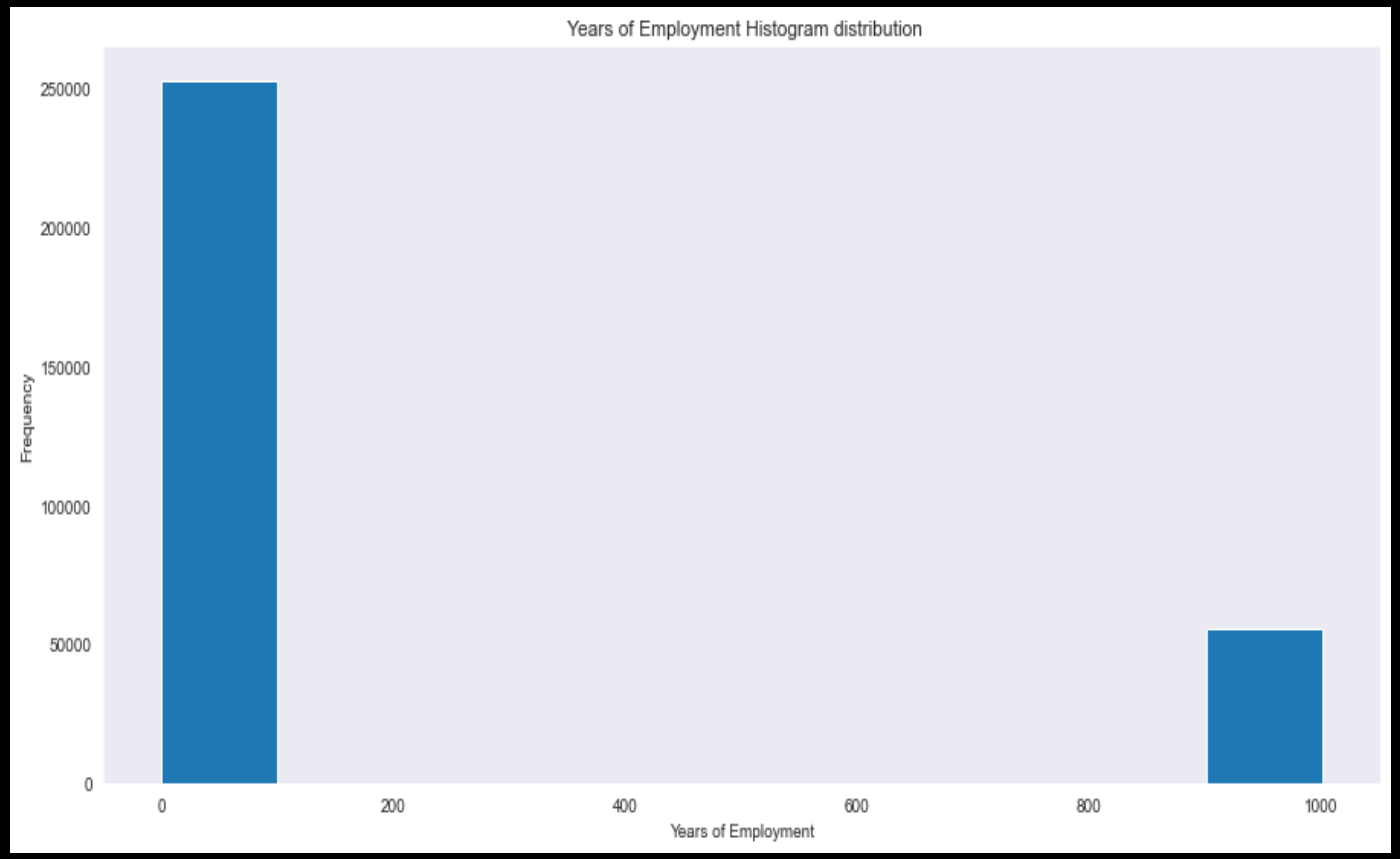
Box plot for column CNT_CHILDREN column does show some outliers must be for people who have number of children as high as 19 . So no need to treat outliers in this case

TREATING OUTLIERS



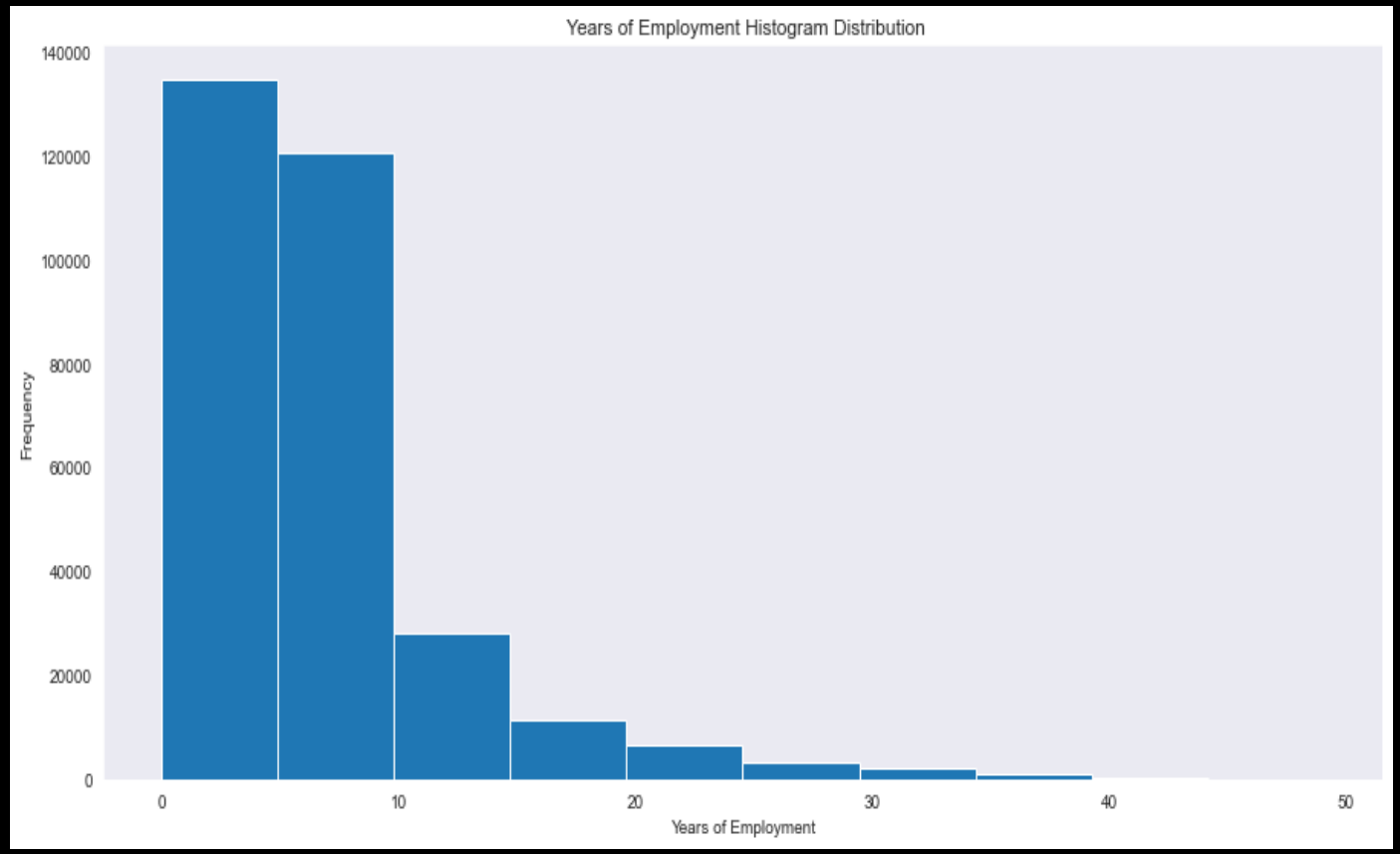
- Even if we don't delete the outlier it shouldn't affect your analysis.
- On the other hand, since this column represents the income of the applicants, such a high value of income is expected.

TREATING OUTLIERS



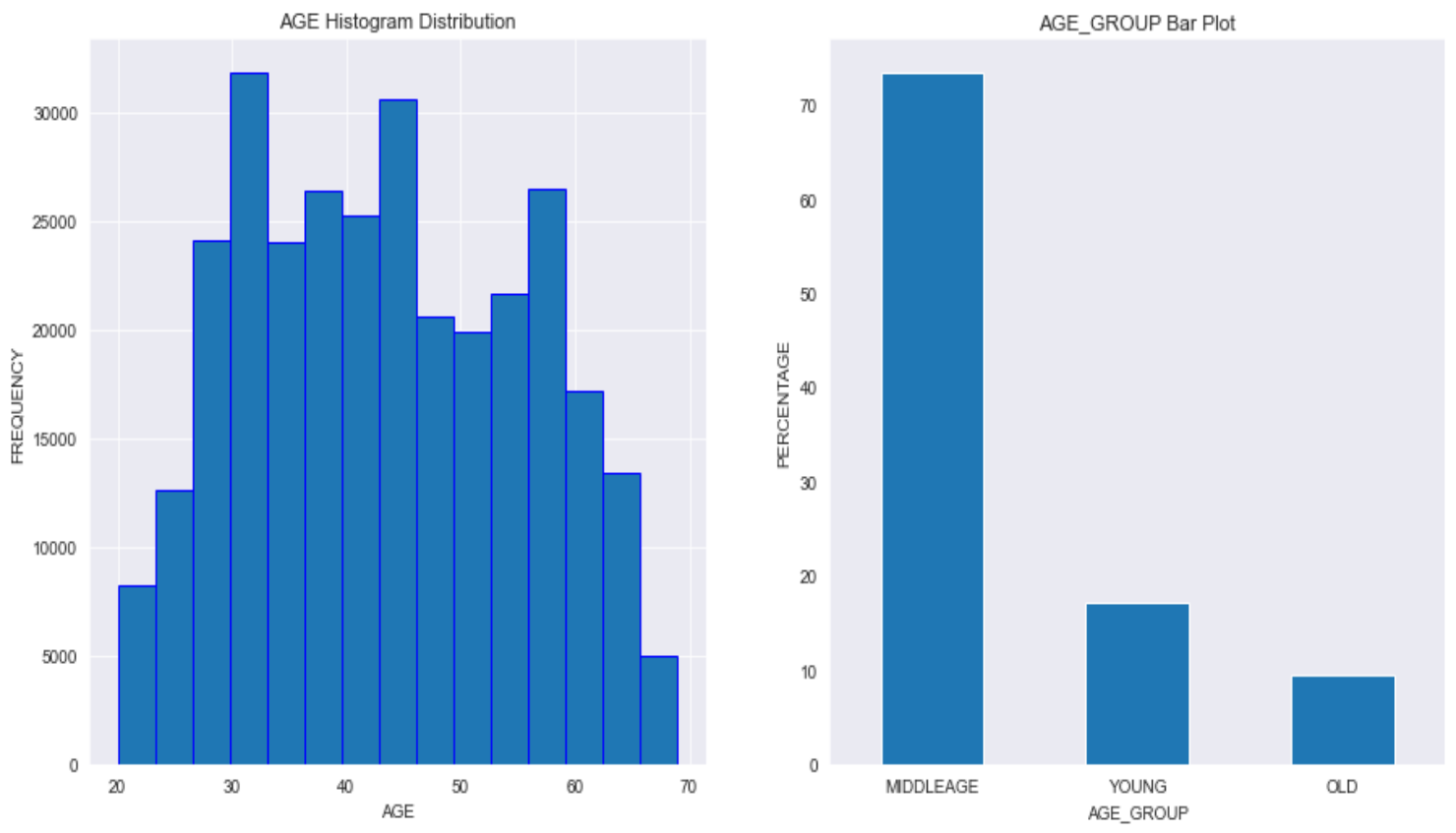
- Since there are outliers, mean is not the right choice for replacing the outliers.
- In this case, median will be suitable.

TREATING OUTLIERS



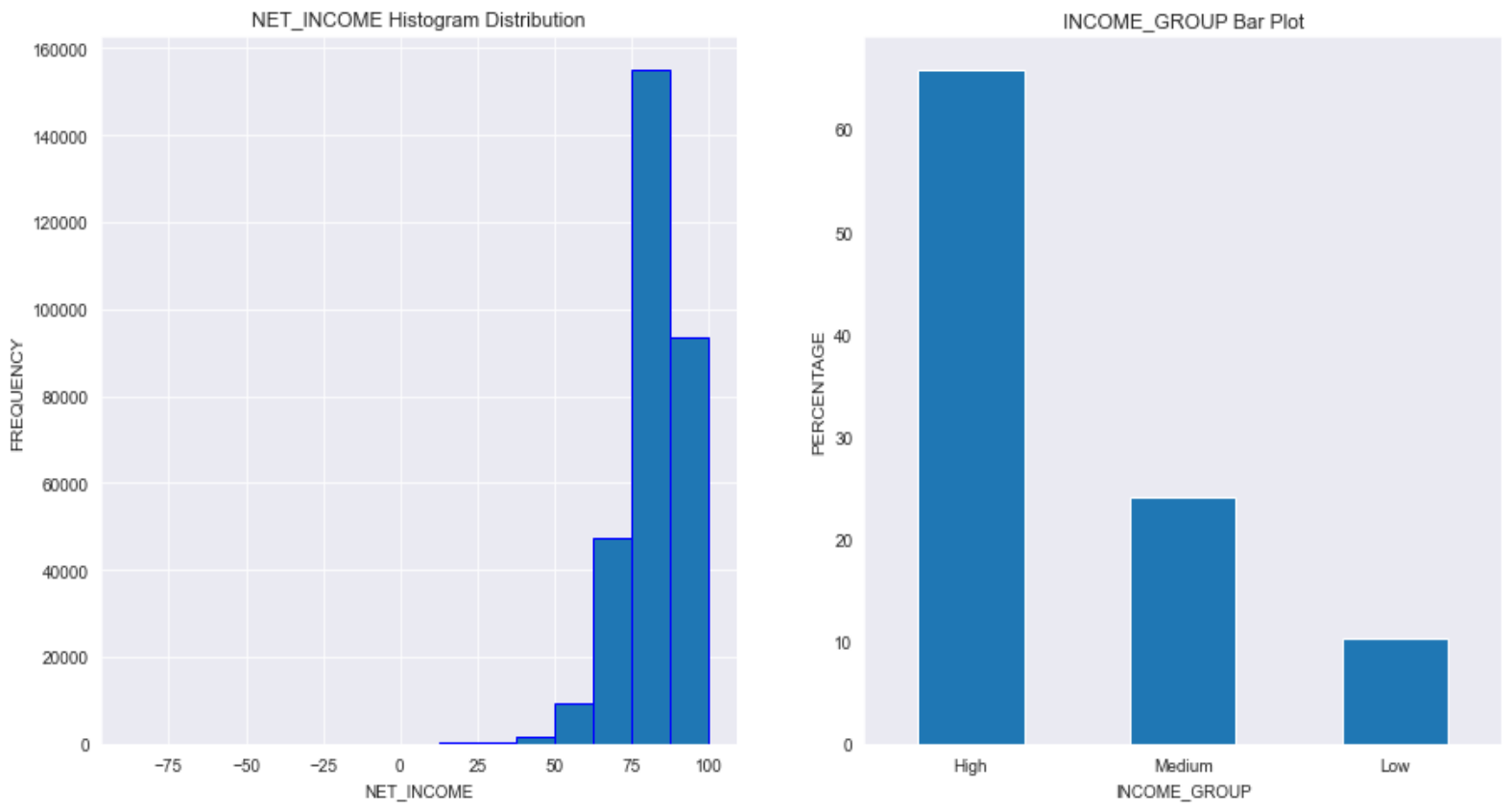
We have treated the outliers in DAYS_EMPLOYED by replacing them with median, which is 2219.0

BINNING CONTINUOUS VARIABLES



AGE_GROUP column has been created from AGE column using Binning.

BINNING CONTINUOUS VARIABLES

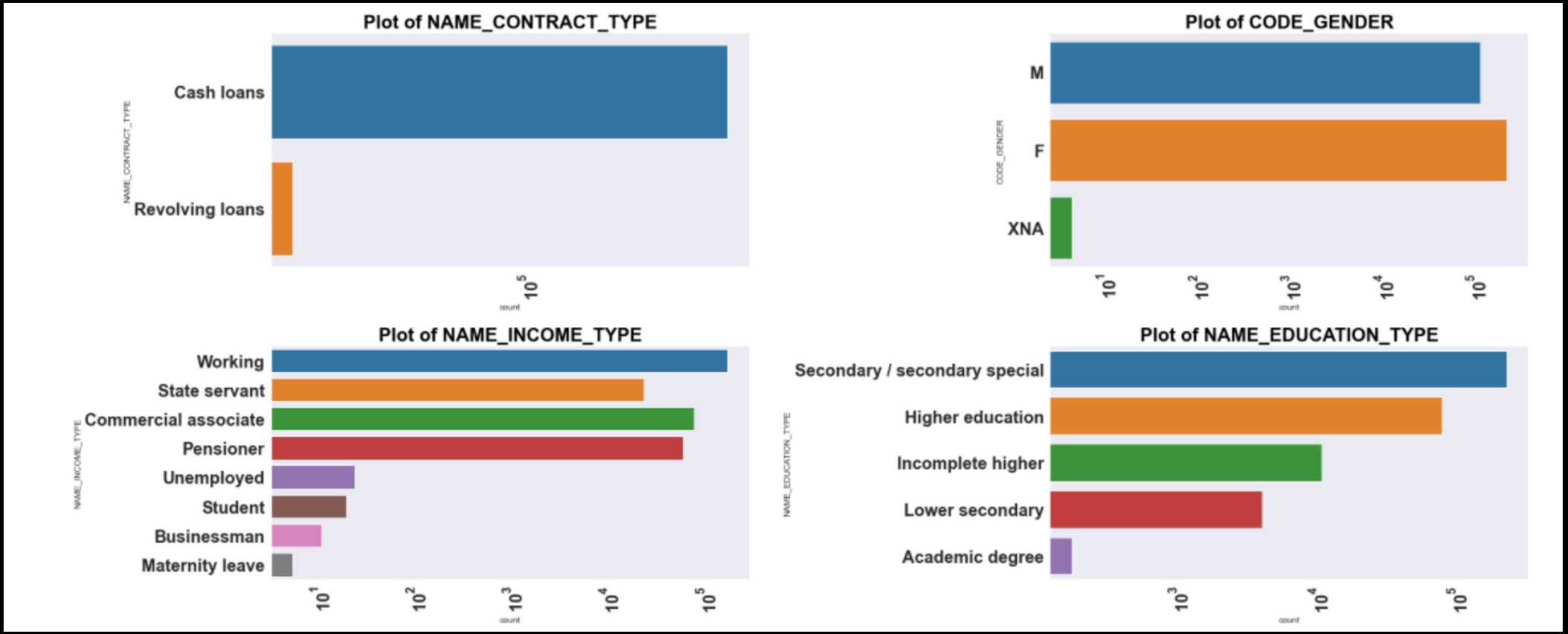


INCOME_GROUP column has been created AMT_INCOME_TOTAL and AMT_ANNUITY columns using Binning.

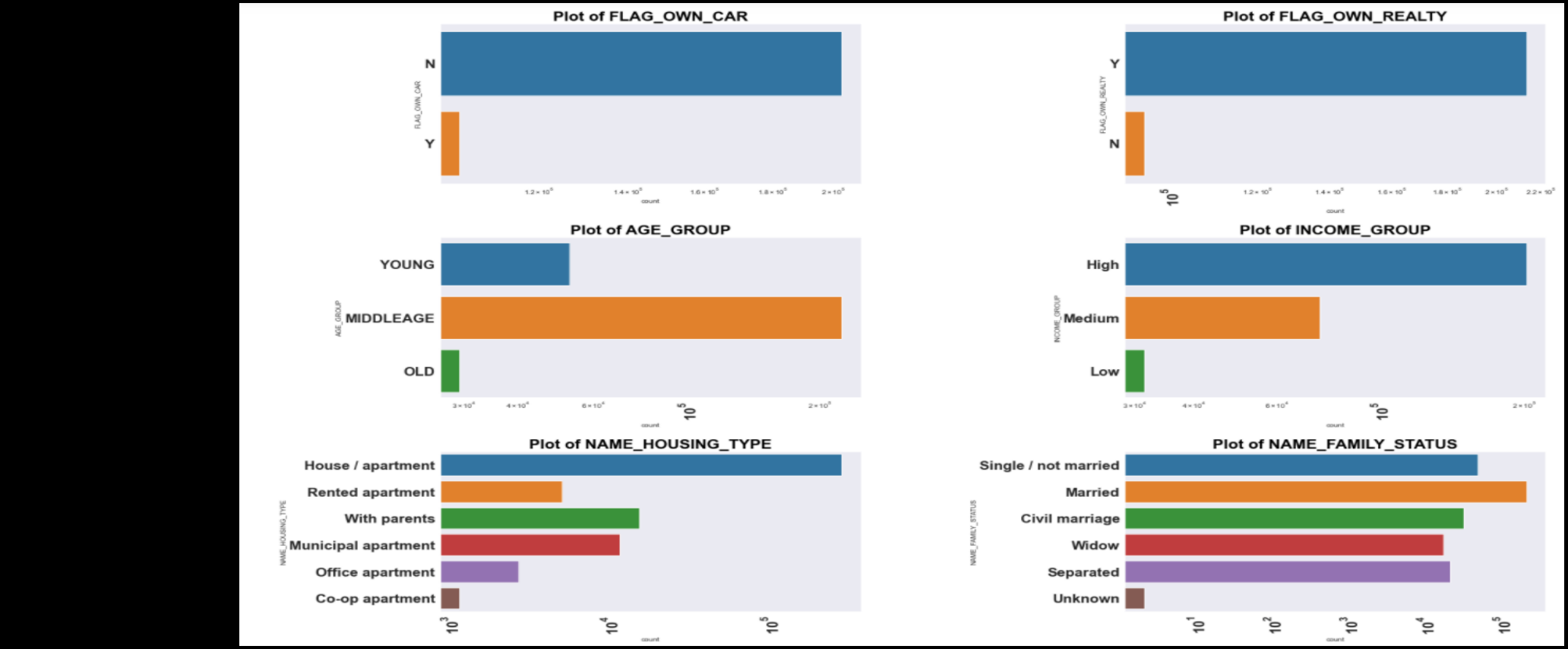
UNIVARIATE ANALYSIS FOR CATEGORICAL COLUMNS

- 'NAME_CONTRACT_TYPE',
- 'CODE_GENDER',
- 'NAME_INCOME_TYPE',
- 'NAME_EDUCATION_TYPE',
- 'FLAG_OWN_CAR',
- 'FLAG_OWN_REALTY',
- 'AGE_GROUP',
- 'INCOME_GROUP',
- 'NAME_HOUSING_TYPE',
- 'NAME_FAMILY_STATUS' were selected columns

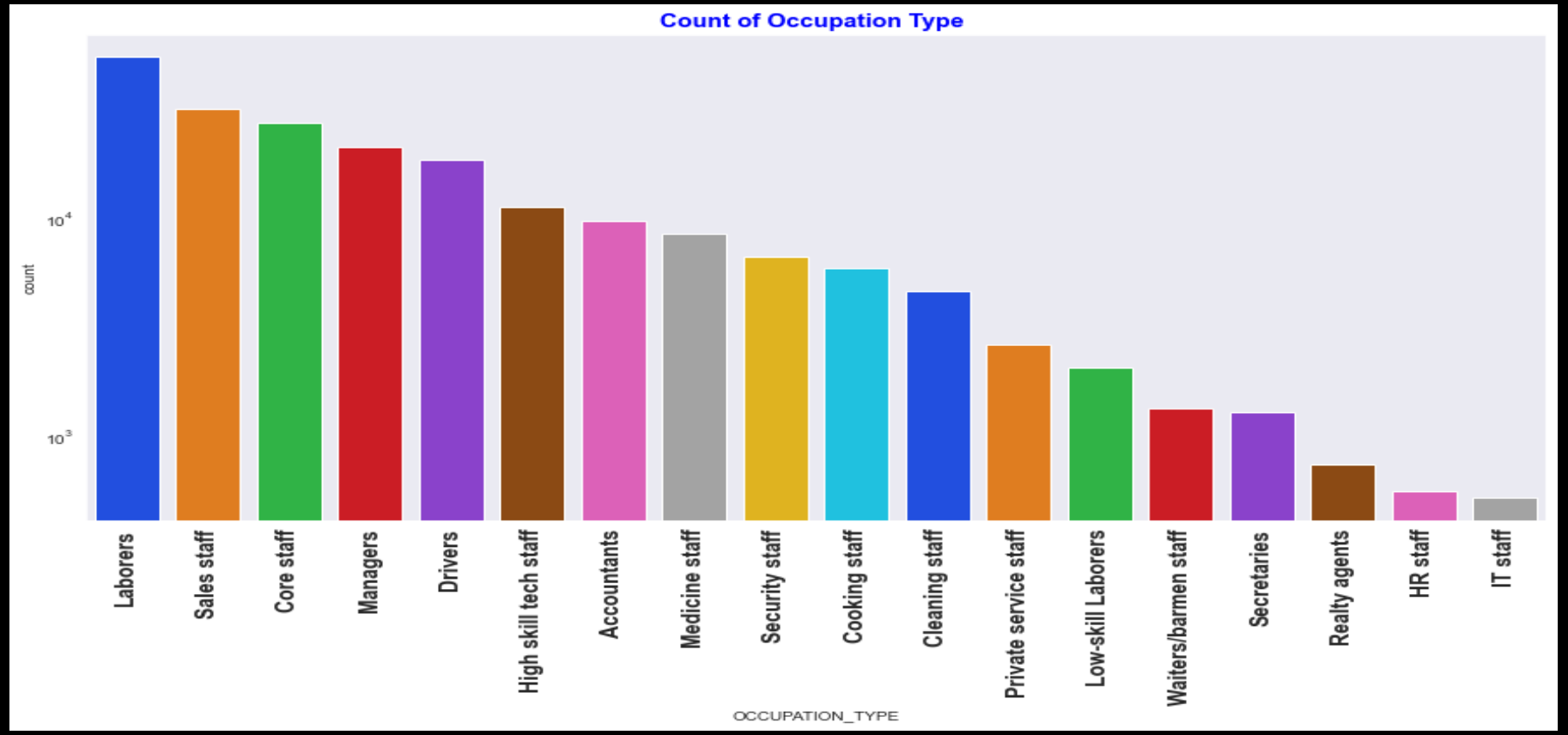
UNIVARIATE ANALYSIS FOR CATEGORICAL COLUMNS



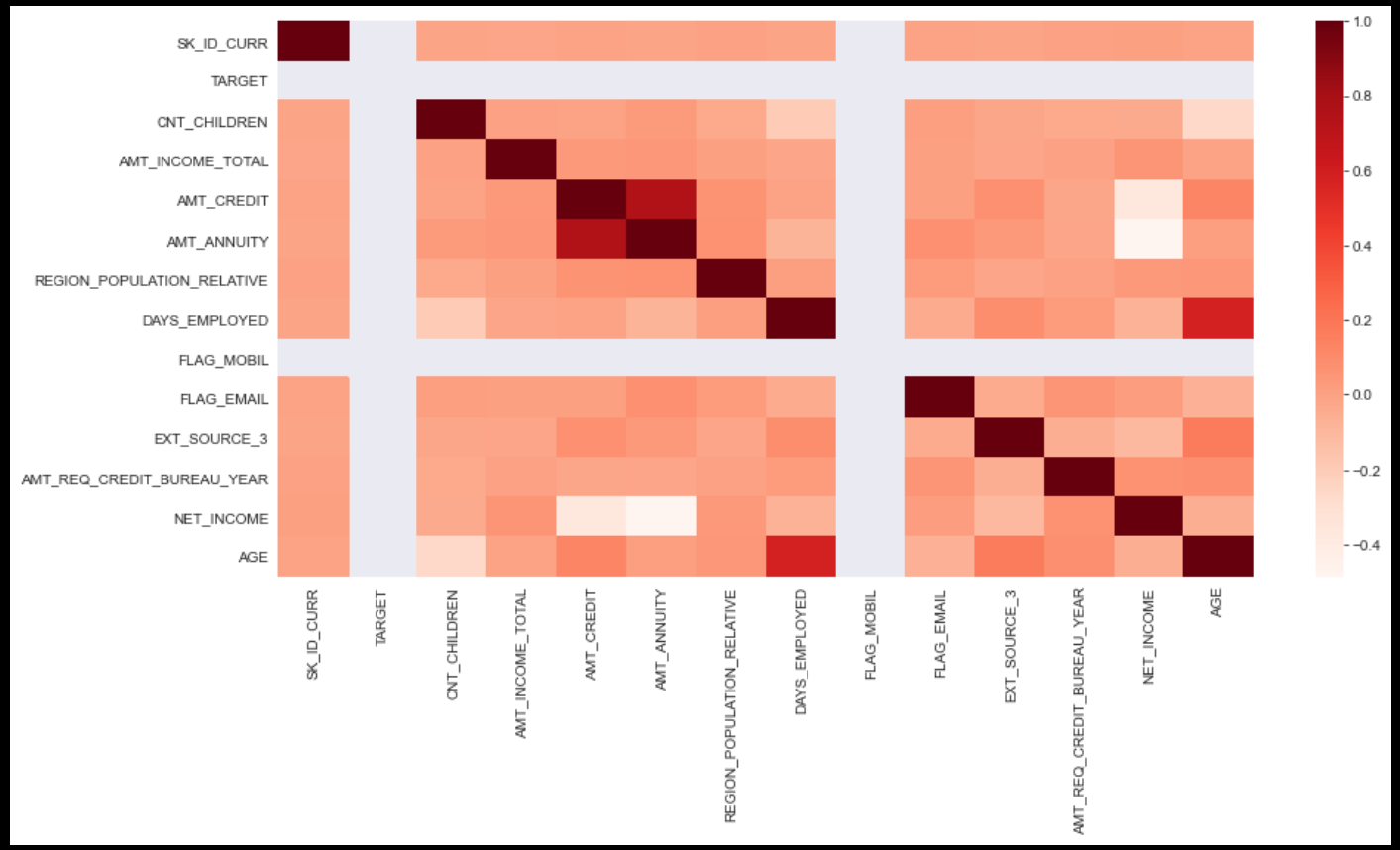
UNIVARIATE ANALYSIS FOR CATEGORICAL COLUMNS



UNIVARIATE ANALYSIS OF OCCUPATION_TYPE



BIVARIATE ANALYSIS - NUMERICAL TO NUMERICAL

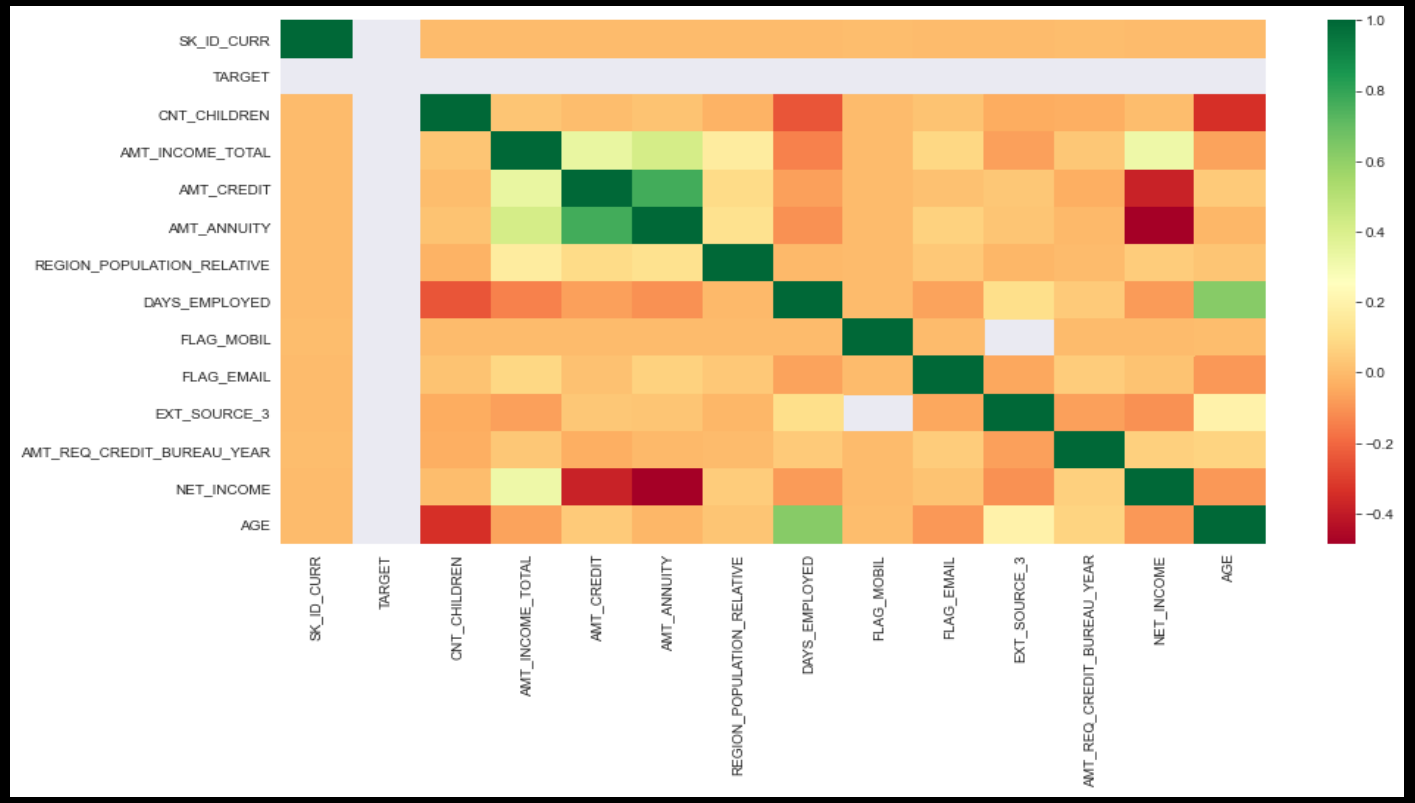


AMT_CREDIT	AMT_ANNUITY	0.752
AMT_ANNUITY	AMT_CREDIT	0.752
DAYS_EMPLOYED	AGE	0.582
AGE	DAYS_EMPLOYED	0.582
EXT_SOURCE_3	AGE	0.172
AGE	EXT_SOURCE_3	0.172
	AMT_CREDIT	0.135
AMT_CREDIT	AGE	0.135
EXT_SOURCE_3	DAYS_EMPLOYED	0.096
DAYS_EMPLOYED	EXT_SOURCE_3	0.096
AGE	AMT_REQ_CREDIT_BUREAU_YEAR	0.084
AMT_REQ_CREDIT_BUREAU_YEAR	AGE	0.084
AMT_ANNUITY	FLAG_EMAIL	0.078
FLAG_EMAIL	AMT_ANNUITY	0.078
EXT_SOURCE_3	AMT_CREDIT	0.078
AMT_CREDIT	EXT_SOURCE_3	0.078
AMT_REQ_CREDIT_BUREAU_YEAR	NET_INCOME	0.077
NET_INCOME	AMT_REQ_CREDIT_BUREAU_YEAR	0.077
REGION_POPULATION_RELATIVE	AMT_ANNUITY	0.072
AMT_ANNUITY	REGION_POPULATION_RELATIVE	0.072

dtype: float64

Top 10 Correlation Pairs of Columns for Target = 1 (Clients with Payment Difficulties or Defaulters)

BIVARIATE ANALYSIS - NUMERICAL TO NUMERICAL

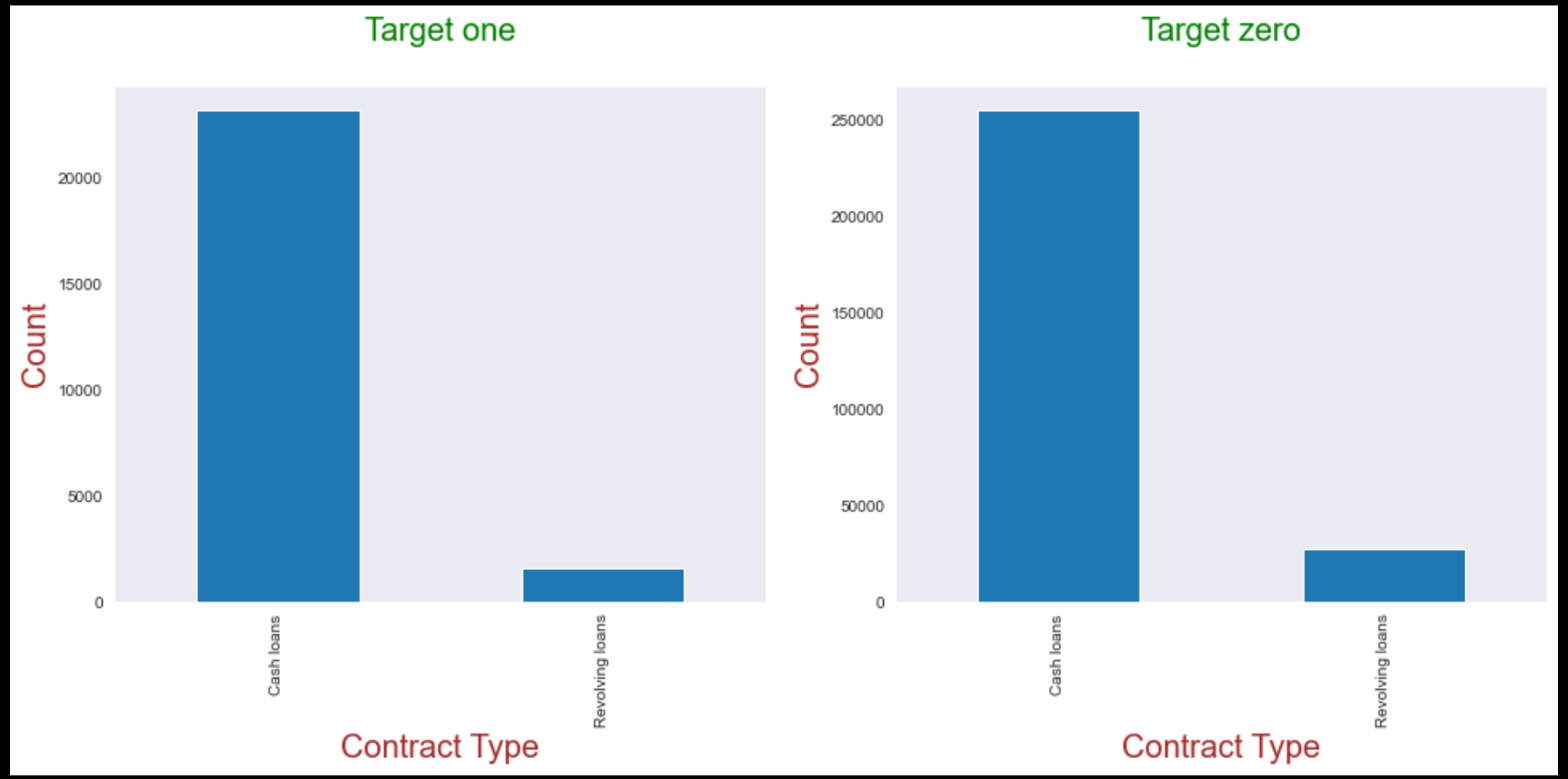


AMT_ANNUITY	AMT_CREDIT	0.771
AMT_CREDIT	AMT_ANNUITY	0.771
AGE	DAYS_EMPLOYED	0.626
DAYS_EMPLOYED	AGE	0.626
AMT_INCOME_TOTAL	AMT_ANNUITY	0.419
AMT_ANNUITY	AMT_INCOME_TOTAL	0.419
AMT_INCOME_TOTAL	AMT_CREDIT	0.343
AMT_CREDIT	AMT_INCOME_TOTAL	0.343
AMT_INCOME_TOTAL	NET_INCOME	0.322
NET_INCOME	AMT_INCOME_TOTAL	0.322
AGE	EXT_SOURCE_3	0.197
EXT_SOURCE_3	AGE	0.197
AMT_INCOME_TOTAL	REGION_POPULATION_RELATIVE	0.168
REGION_POPULATION_RELATIVE	AMT_INCOME_TOTAL	0.168
	AMT_ANNUITY	0.121
AMT_ANNUITY	REGION_POPULATION_RELATIVE	0.121
DAYS_EMPLOYED	EXT_SOURCE_3	0.112
EXT_SOURCE_3	DAYS_EMPLOYED	0.112
REGION_POPULATION_RELATIVE	AMT_CREDIT	0.101
AMT_CREDIT	REGION_POPULATION_RELATIVE	0.101

dtype: float64

Above observation shows top 10 pairs of columns have strong correlation for Other Clients (TARGET = 0):

BIVARIATE ANALYSIS - CATEGORICAL TO CATEGORICAL



Clients are clearly not paying cash loans on time rather than revolving loans.

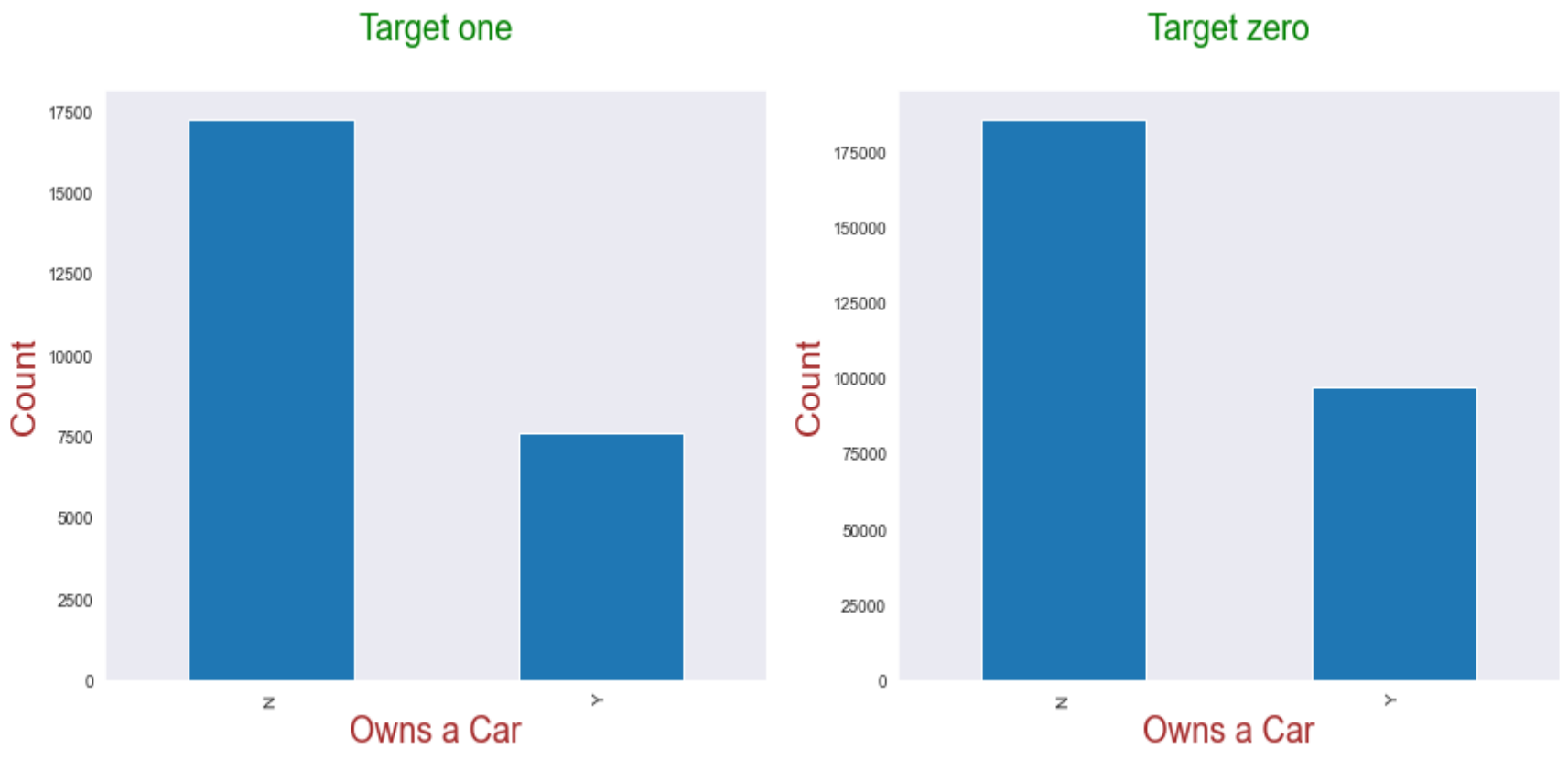
BIVARIATE ANALYSIS - CATEGORICAL TO CATEGORICAL

- In both the cases, the female clients are not paying loans to companies on time rather than male employees.
- Possibly the salary difference could be the reason.

BIVARIATE ANALYSIS - CATEGORICAL TO CATEGORICAL

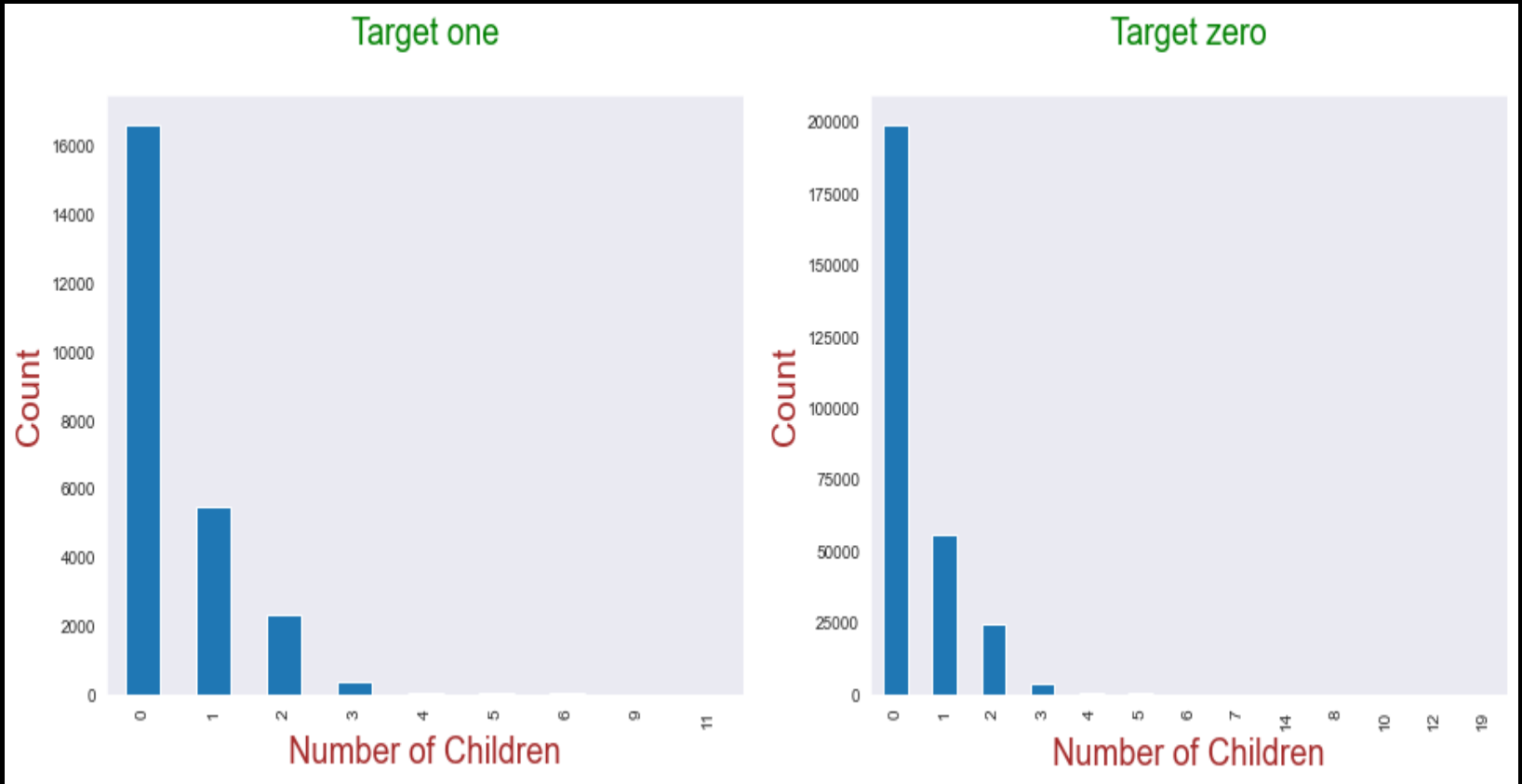
- Clients who own a house don't seem to pay loan on time
- Mostly, they are not able to manage house loans with company loans.

CLIENT BEHAVIOR TOWARDS LOAN WHO OWNS A CAR



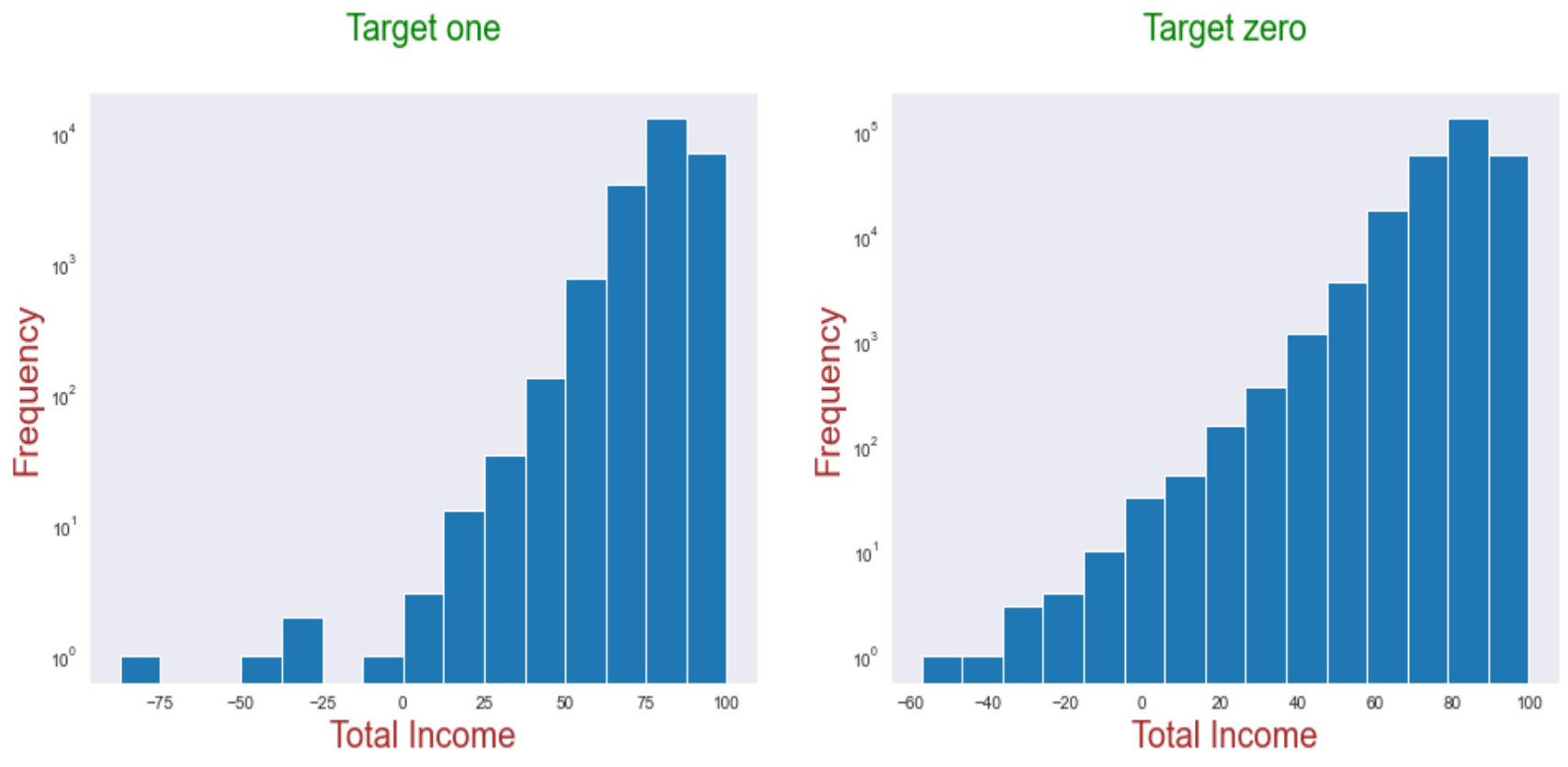
- Client who don't own a car clearly don't pay loans on time
- Because giving loans to clients without cars means they are not financially sound
- In this modern age, car is the basic requirement

CLIENT ATTITUDE WITH CHILDREN TOWARDS COMPANY LOAN



- Client with higher number of children are paying loans on time compared to clients with clients with no children.
- It could be because taking care of children increases the sense of responsibility in us rather than people without children.

CLIENT ATTITUDE WITH INCOME TOWARDS COMPANY LOAN

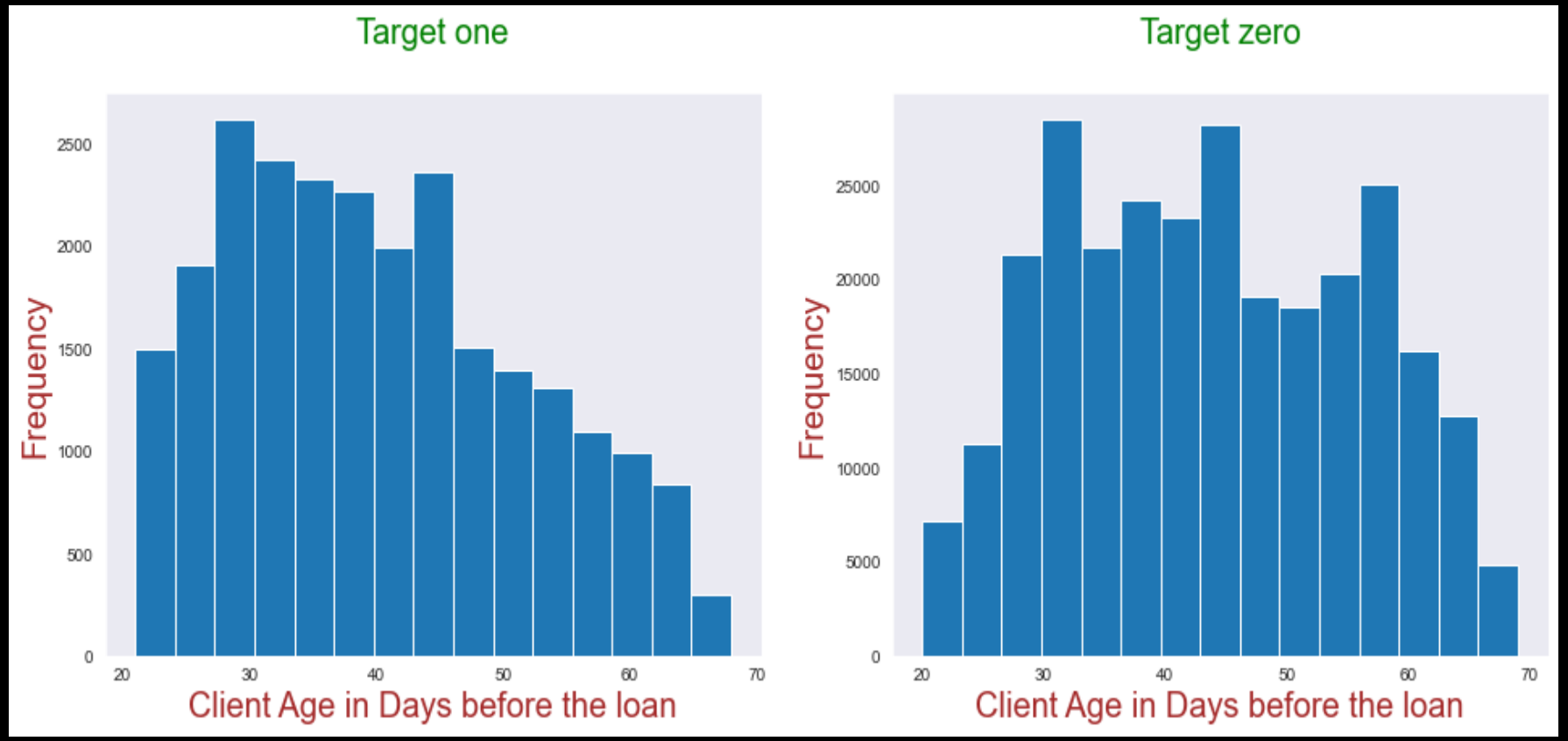


- Clearly client with low income are not paying loans on time rather than clients with higher income.
- Clients with low income could possibly be students or fresher's who have very less income.
- Company should have a complete background check before giving loans.

CLIENT ATTITUDE WITH LOAN AMOUNT TOWARDS COMPANY LOAN

- Its very amazing to see that people with higher loans are paying loans not on time rather than people with low credits

CLIENT ATTITUDE WITH AGE TOWARDS COMPANY LOAN



- Middle age Clients paying loans on time rather than younger and senior citizens.

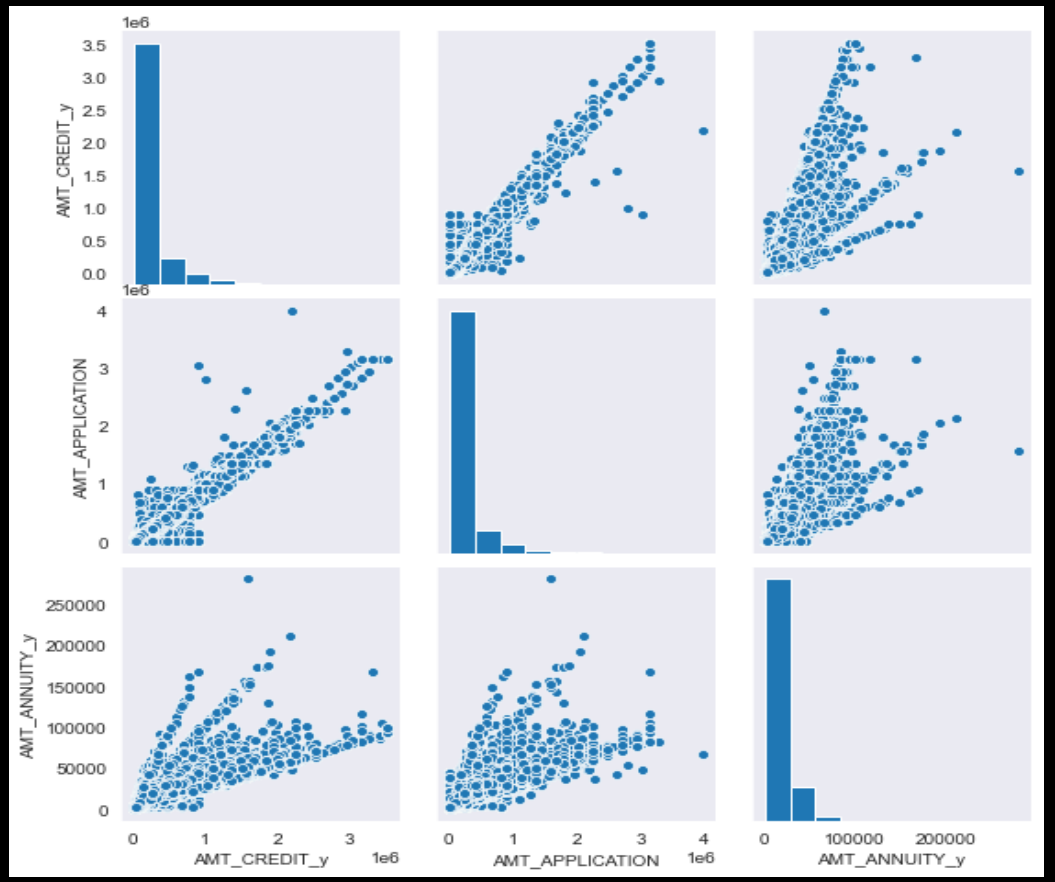
CLIENT ATTITUDE WITH EMPLOYMENT HISTORY TOWARDS COMPANY LOAN

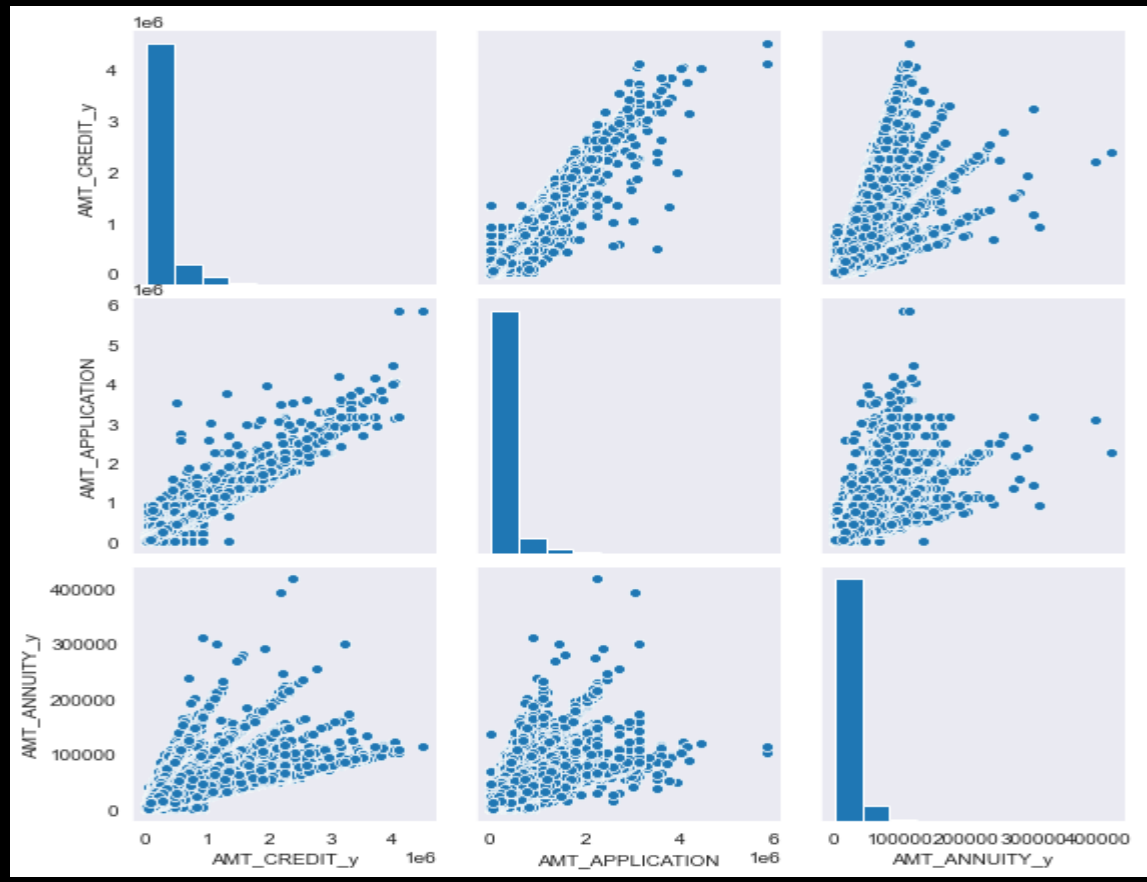
- Clearly fresher's and people with less job experience are not paying loans on time.

READING 'PREVIOUS_DATA' FILE

- checking for null values in a particular column
- Dropping columns that have more than 40 % null values
- checking for null values in a particular column
- Merging the credit_imp data frame & Previous dataframe
- Diving Merged data frame into Target 0 and Target 1
- Determining numerical columns for plotting pair plots which are ['AMT_CREDIT_y','AMT_APPLICATION', 'AMT_ANNUITY_y']

PLOTTING PAIR PLOT FOR TARGET 1AND 0 DATA FRAME

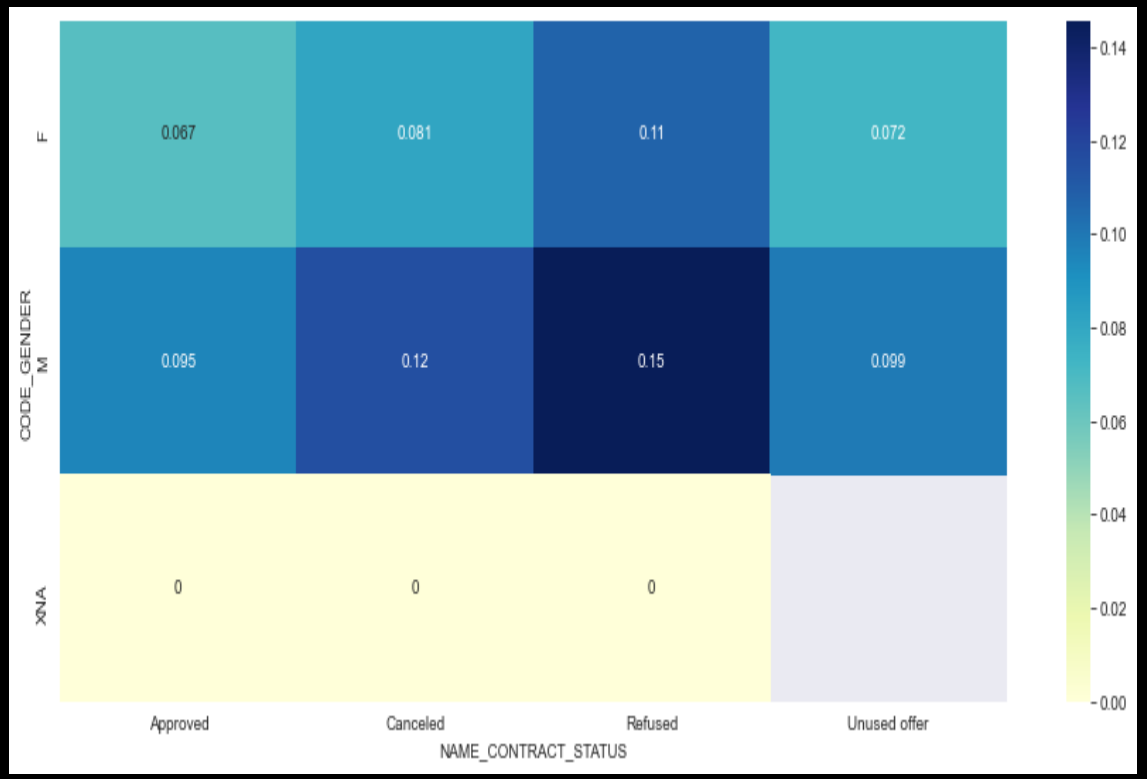


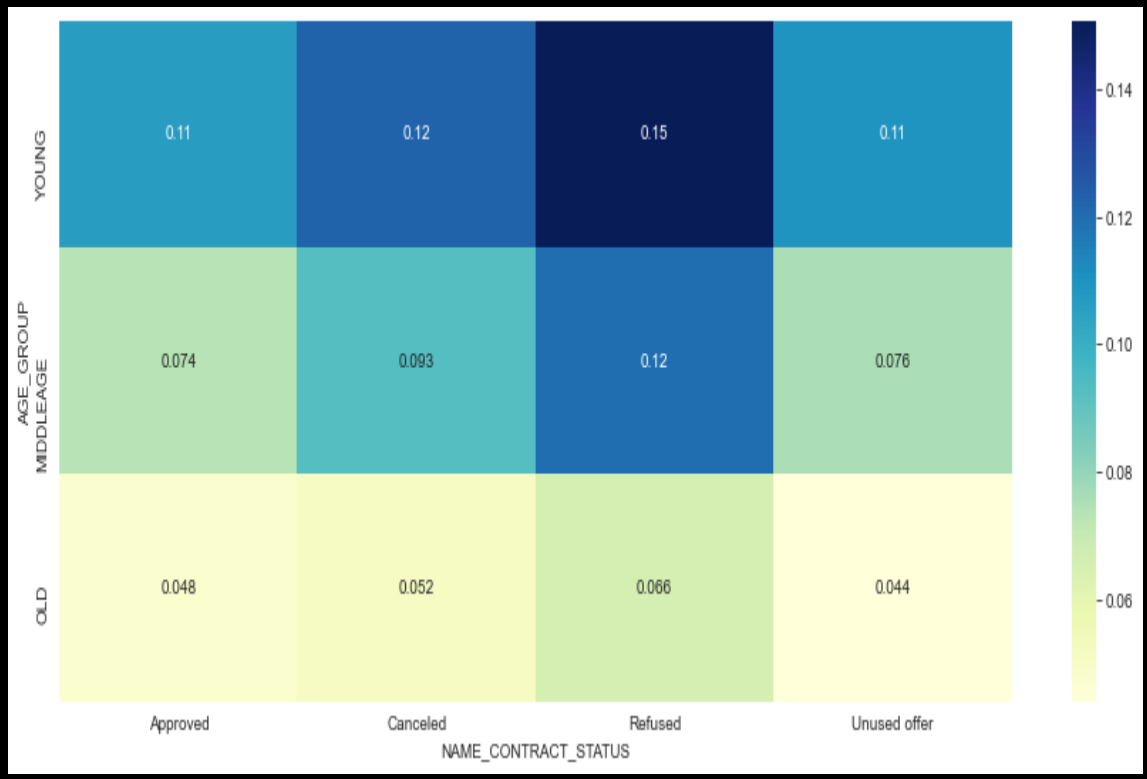


INFERENCES DRAWN

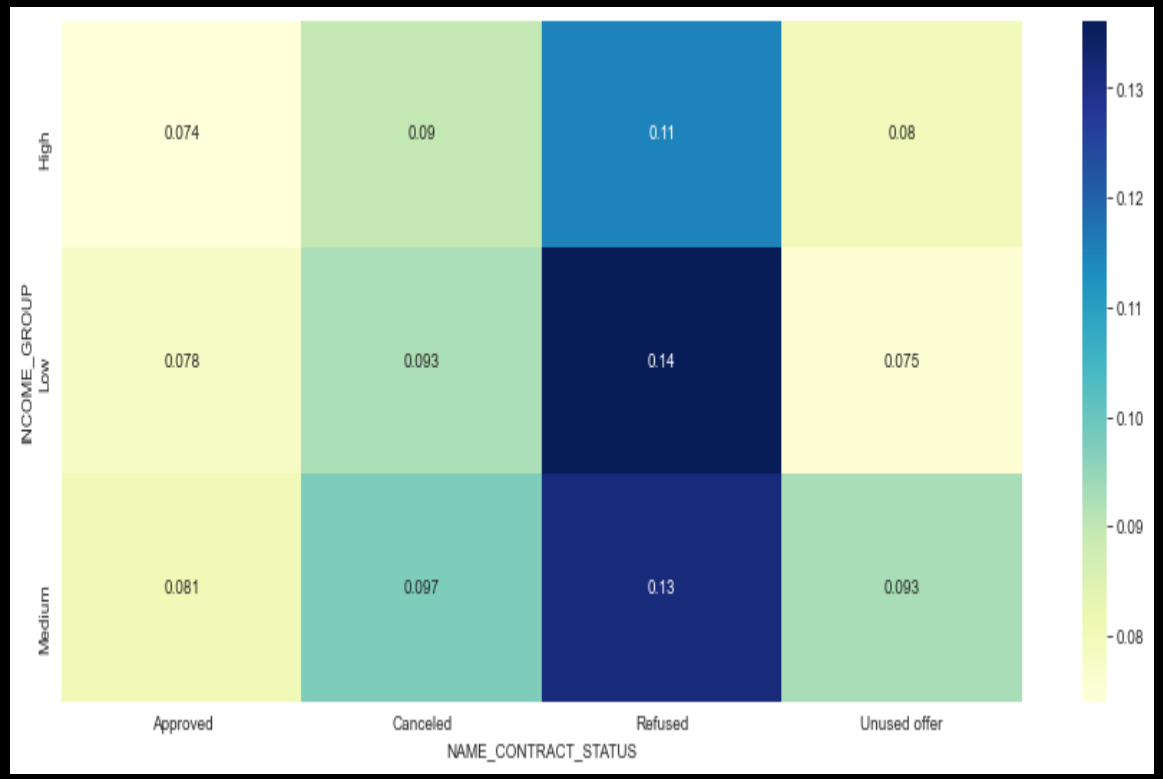
- It is observed that there is a linear relationship between following pairs of columns for both target 0 and target 1 dataframes:
- AMT_CREDIT_y and AMT_APPLICATION
- AMT_ANNUITY_y and AMT_CREDIT_y
- AMT_ANNUITY_y and AMT_APPLICATION

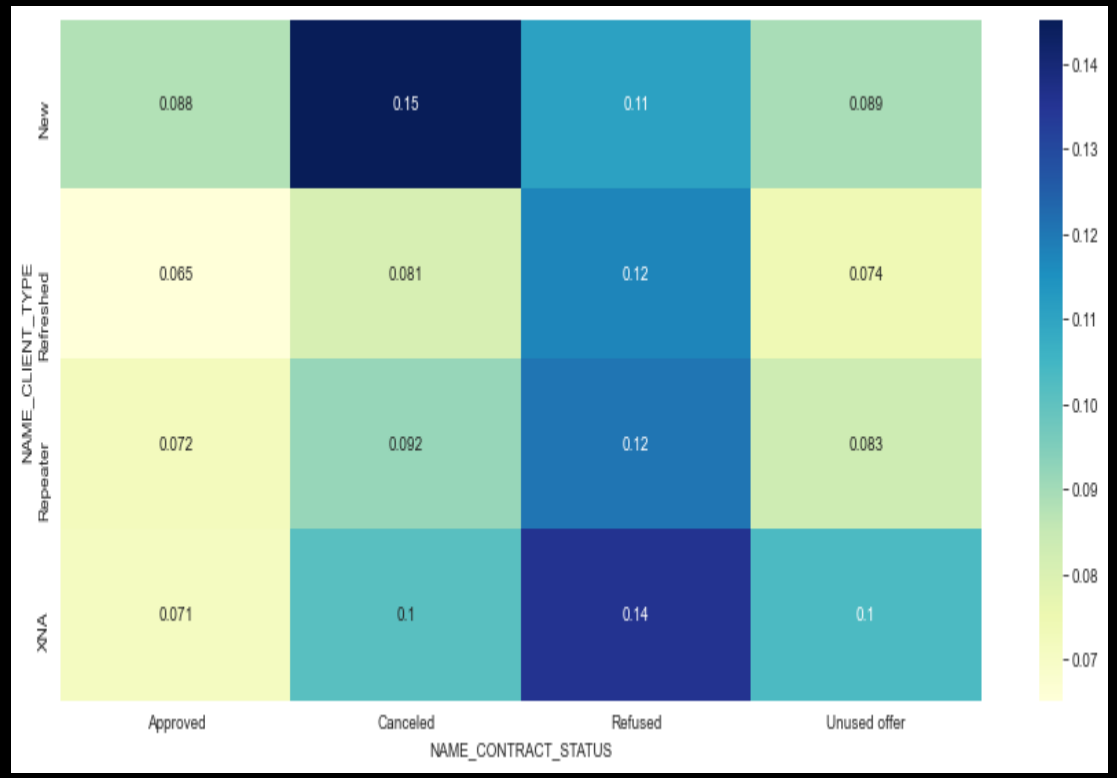
CATEGORICAL - CATEGORICAL OBSERVATIONS



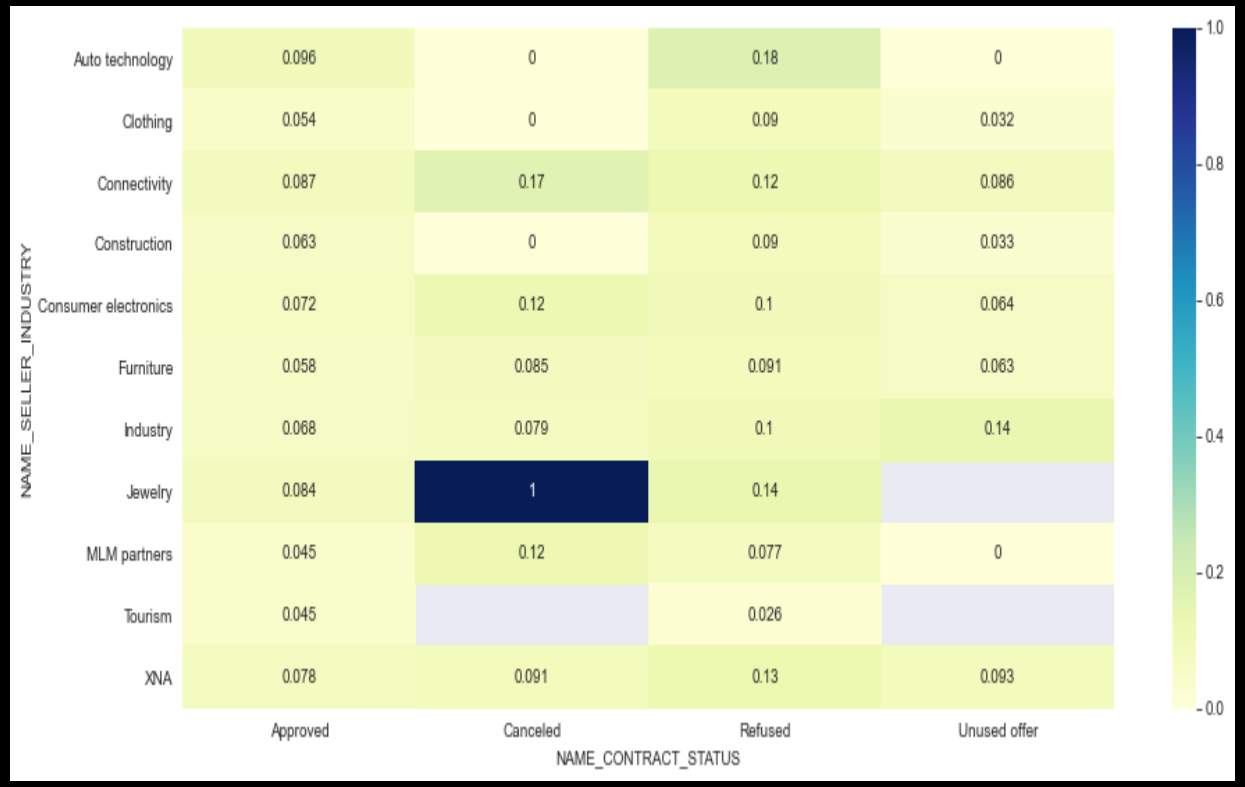


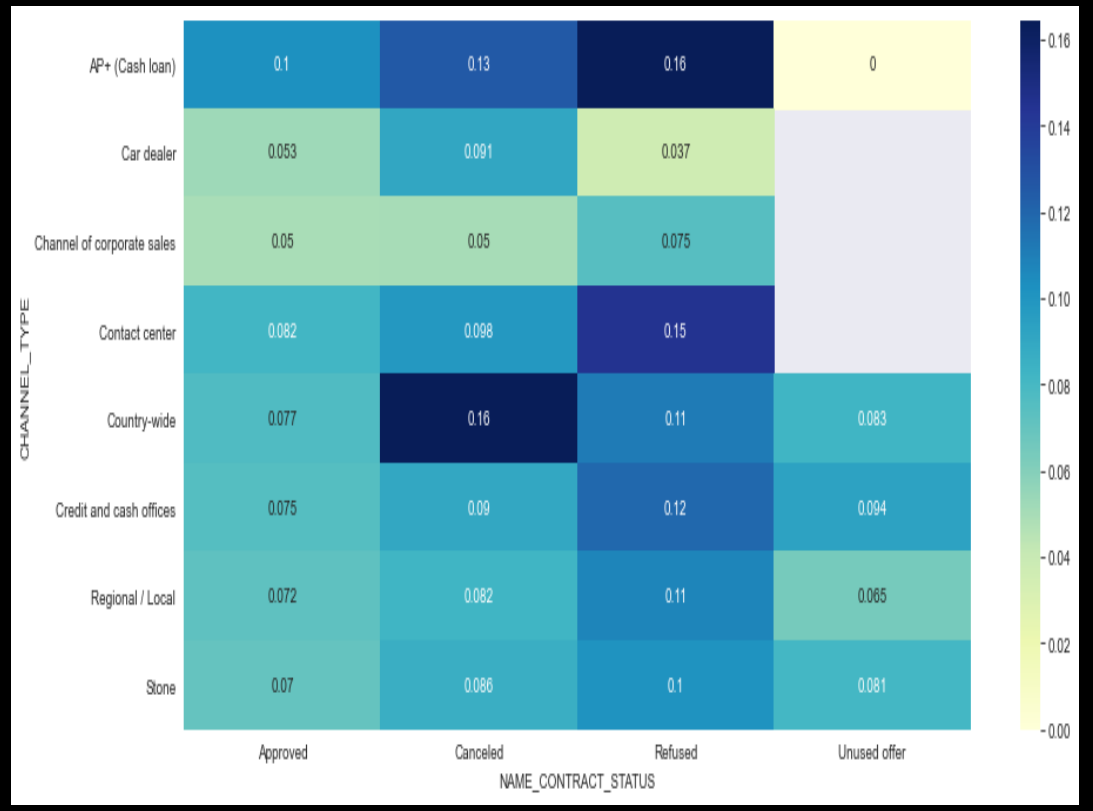
CATEGORICAL - CATEGORICAL OBSERVATIONS





CATEGORICAL - CATEGORICAL OBSERVATIONS

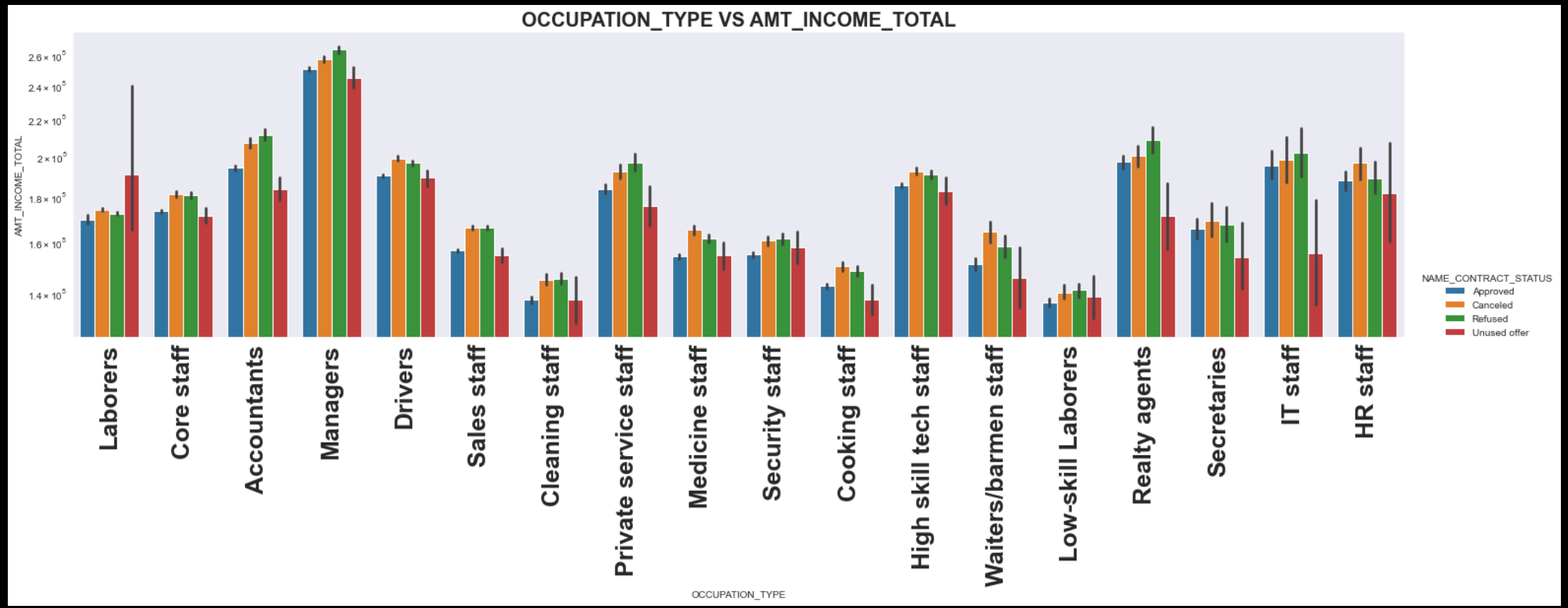




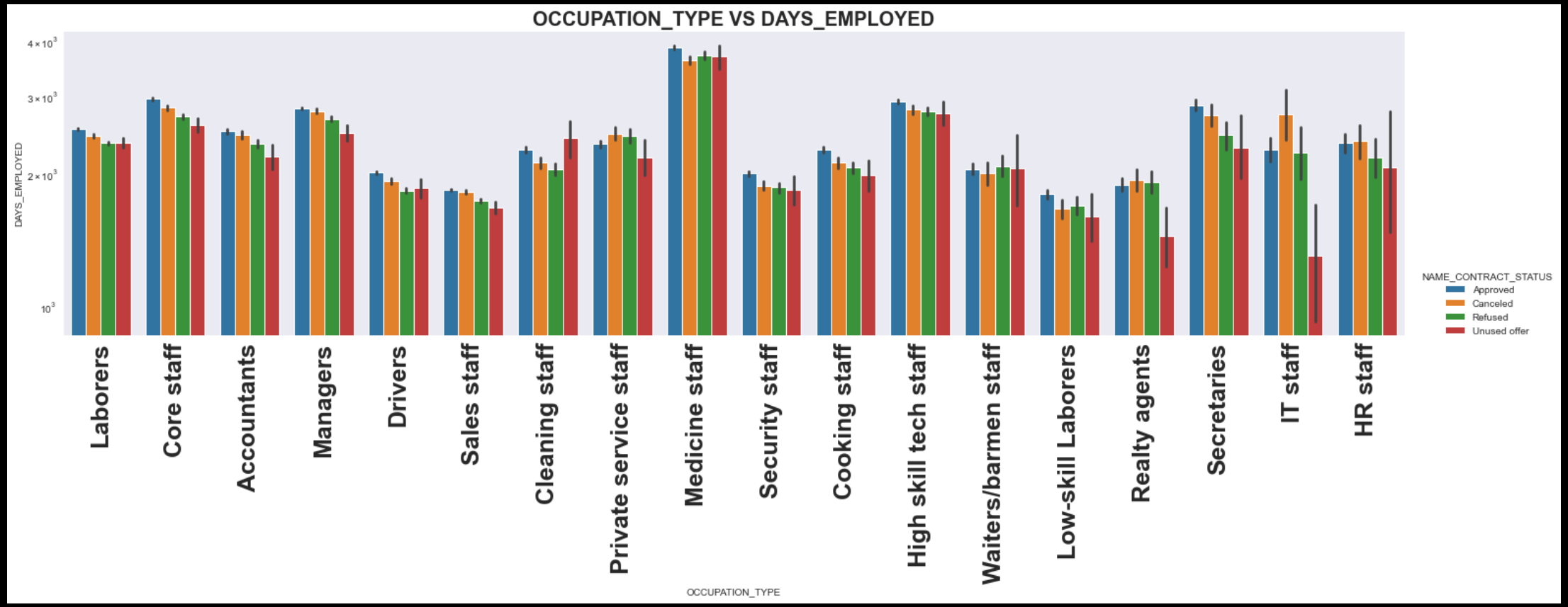
INFERENCES DRAWN

- Clients who are more likely to have payment difficulties:
- Male clients having previous Contract status is refused
- Clients who are young have previous contract status is refused.
- Clients with low income group previous contract status is refused.
- Clients whose Client type is 'New ' and previous contract status as cancelled.
- Clients with 'Seller Industry' as Jewelry and previous contract status as cancelled.
- Clients with channel type as 'AP + (Cash loans)' is refused.

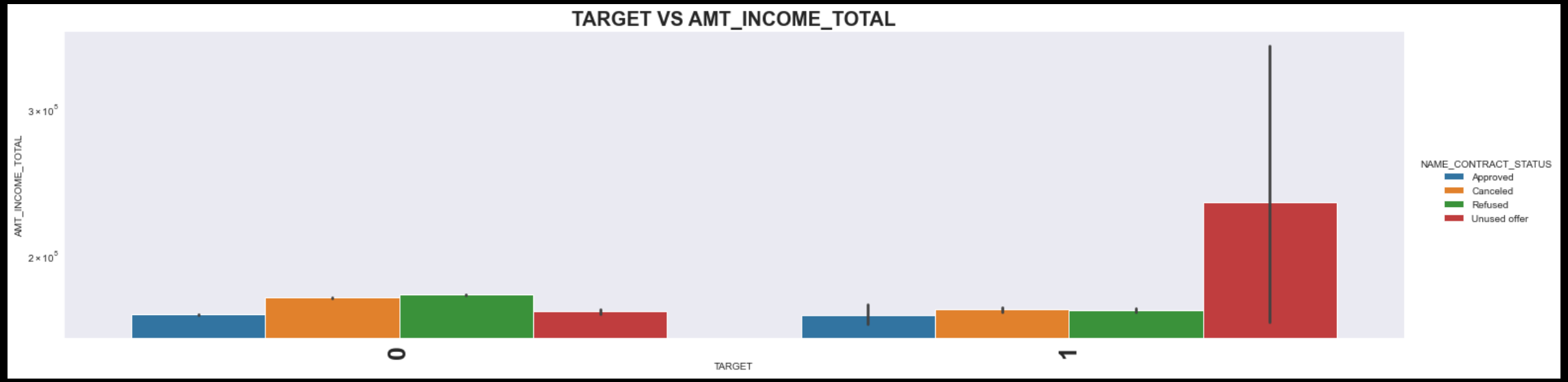
- Clients who are less likely to have payment difficulties:
- Clients who are old have previous contract status as unused offer.
- Clients with client type as 'Refreshed' has previous contract status as approved.
- Clients with previous contract status as cancelled and 'Seller Industry' as Auto technology or Clothing or Construction indicated by zero
- Clients with previous contract status as unused offer and 'Seller Industry' as Auto technology or MLM Partners indicated by zero


OCCUPATION_TYPE VS AMT_INCOME_TOTAL
AMT_INCOME_TOTAL
2.6 × 10^5
2.4 × 10^5
2.2 × 10^5
2 × 10^5
1.8 × 10^5
1.6 × 10^5
1.4 × 10^5
Laborers
Core staff
Accountants
Managers
Drivers
Sales staff
Cleaning staff
Private service staff
Medicine staff
Security staff
Cooking staff
High skill tech staff
Waiters/barmen staff
Low-skill Laborers
Realty agents
Secretaries
IT staff
HR staff
OCCUPATION_TYPE
NAME_CONTRACT_STATUS
Approved
Canceled
Refused
Unused offer

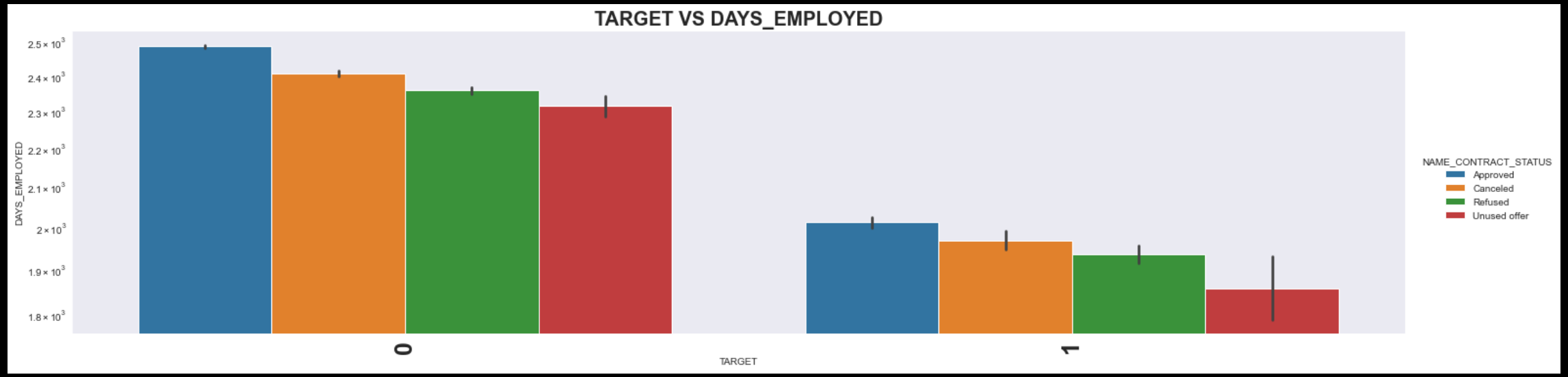
CATEGORICAL - NUMERICAL



CATEGORICAL - NUMERICAL



CATEGORICAL - NUMERICAL



INFERENCES DRAWN

- Clients with more chances of getting loan approved based on previous application data:
- Clients with occupation type as Manager have high income.
- Clients with occupation type as Medicine Staff have good employment history.
- Clients with good employment history don't face payment difficulties.
- Clients with fewer chances of getting loan approved based on previous application data:
- Cleaning Staff or Low skill Laborers have very low income.
- Drivers, Sales, Security and cooking staffs have a average employment history.
- It signifies that they have a tendency to shift from one job to another and face payment difficulties