# EXPLORATORY DATA ANALYSIS ASSIGNMENT

FROM MRINMOY CHOUDHURY

# PROBLEM STATEMENT

- 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 to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse 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:
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

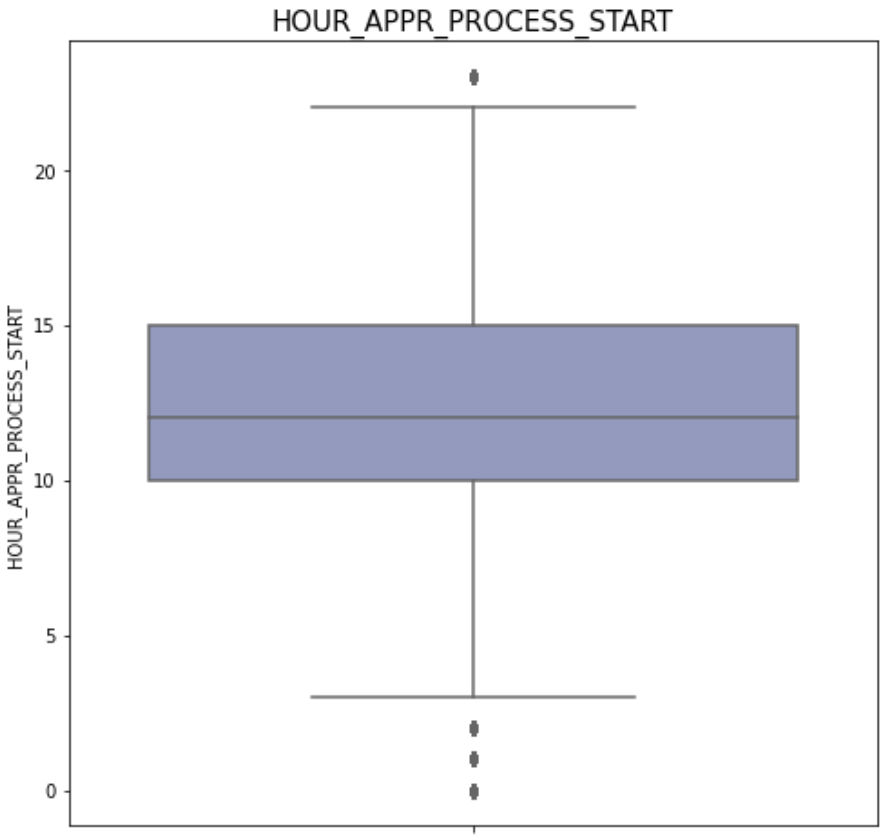
# APPROACH

- Loading both data sets Application data and previous application data.
- Handling NULL values: Checking for columns having null values greater than 50 % and dropping them. For values less than 50% we are imputing the null values with either mean, median or mode depending on the situation.
- Identifying Outliers: We have identified outliers using Boxplot. Outliers are the values that lie above the upper hinge(Q3+1.5*IQR) and below the lower hinge(Q1-1.5*IQR).Q1 IS $25^{TH}$ Percentile and Q3 is $75^{th}$ Percentile. IQR =Q3-Q1

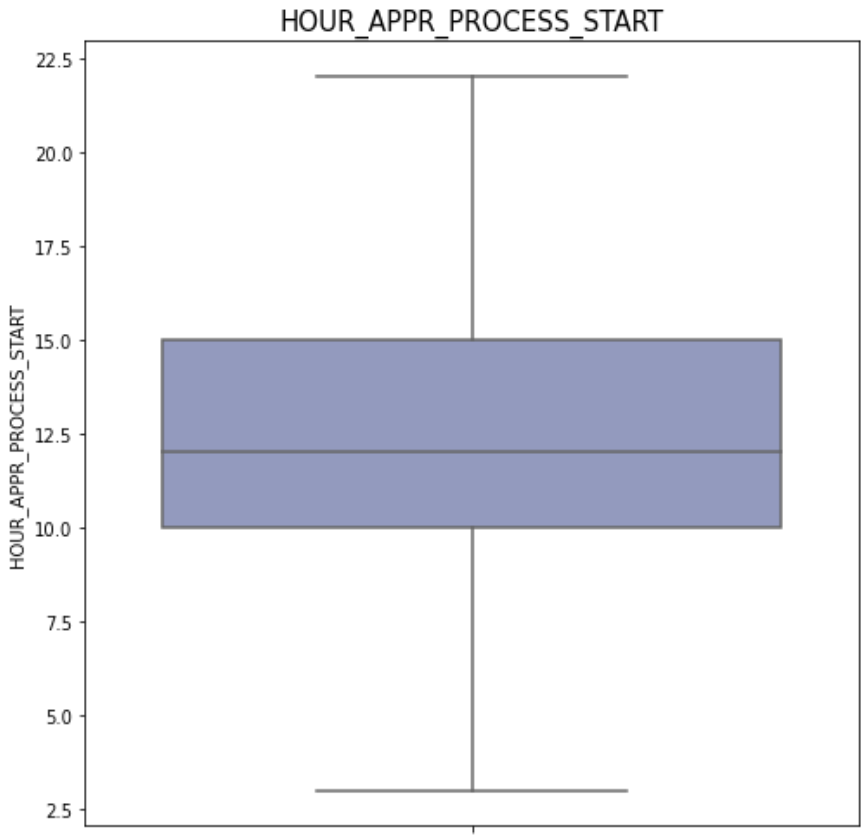
# APPROACH(Cont..)

- UNIVARIATE ANALYSIS of both application and previous application data: We have gathered some insights by performing univariate analysis on various columns.
- BIVARIATE ANALYSIS using the target variable with the other columns in Application data and NAME_CONTRACT_STATUS with the other columns in PREVIOUS APPLICATION DATA
- SEGMENTED UNIVARIATE ANALYSIS of the Application data: We will divide application data into two data frames: 1)TARGET=0 2)TARGET=1
- Finally we have merged application data and previous application data to gain further insights.

# Identifying Outliers

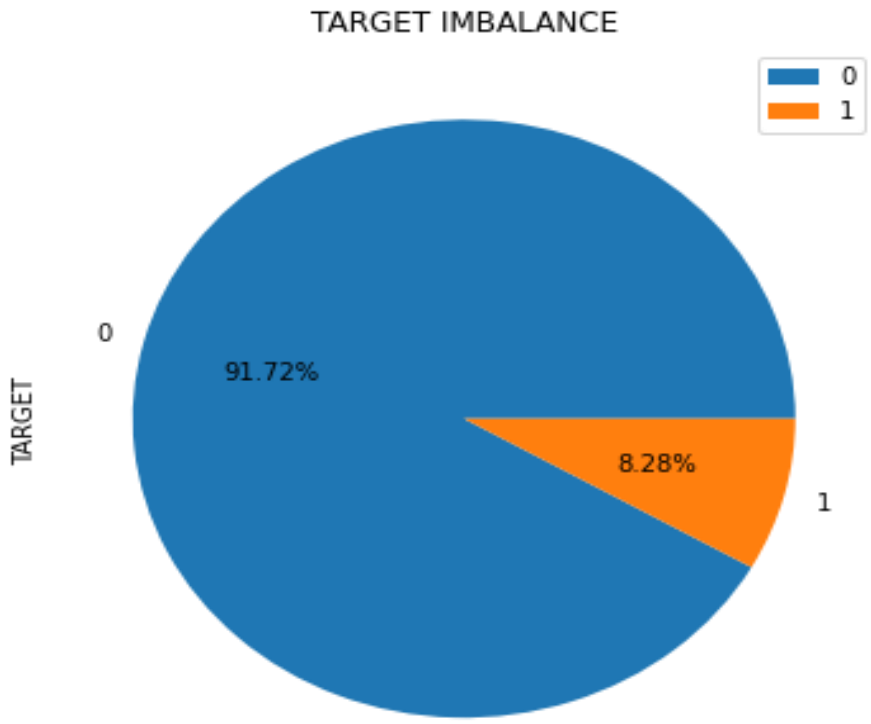



Before handing Outliers




After handling Outliers

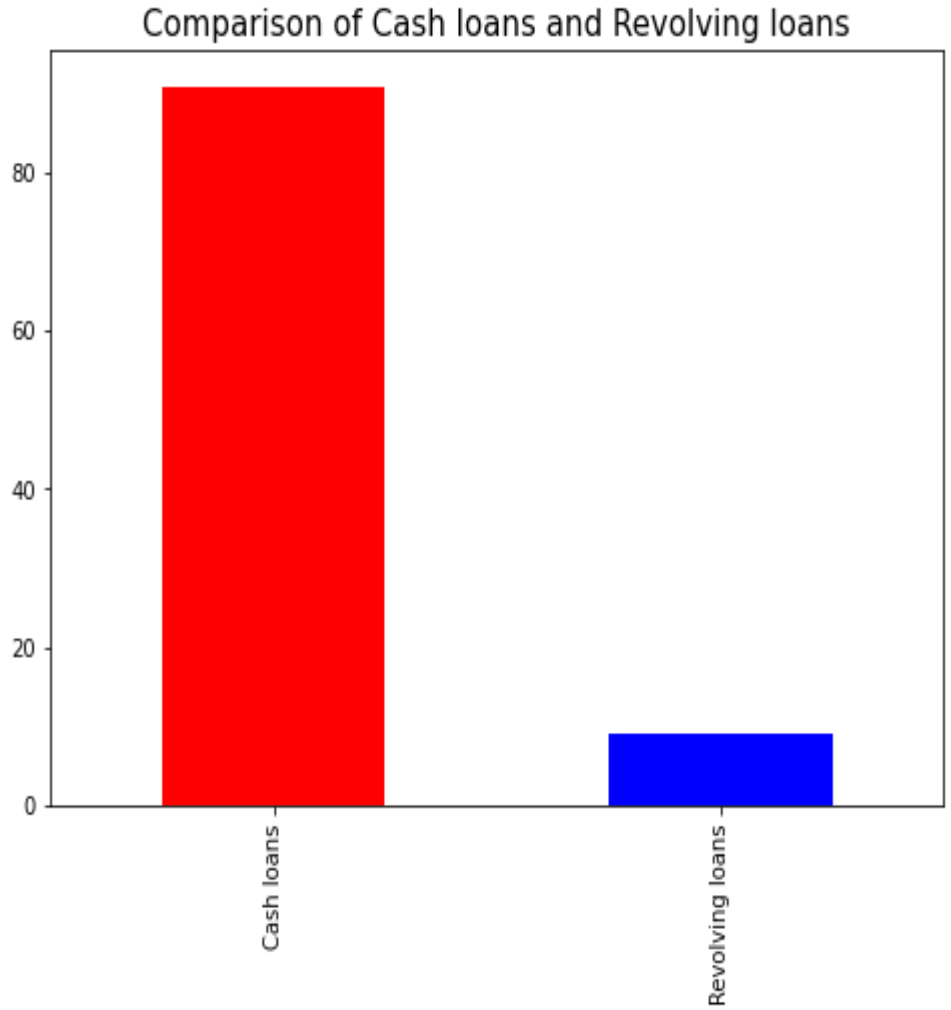
# TARGET IMBALANCE

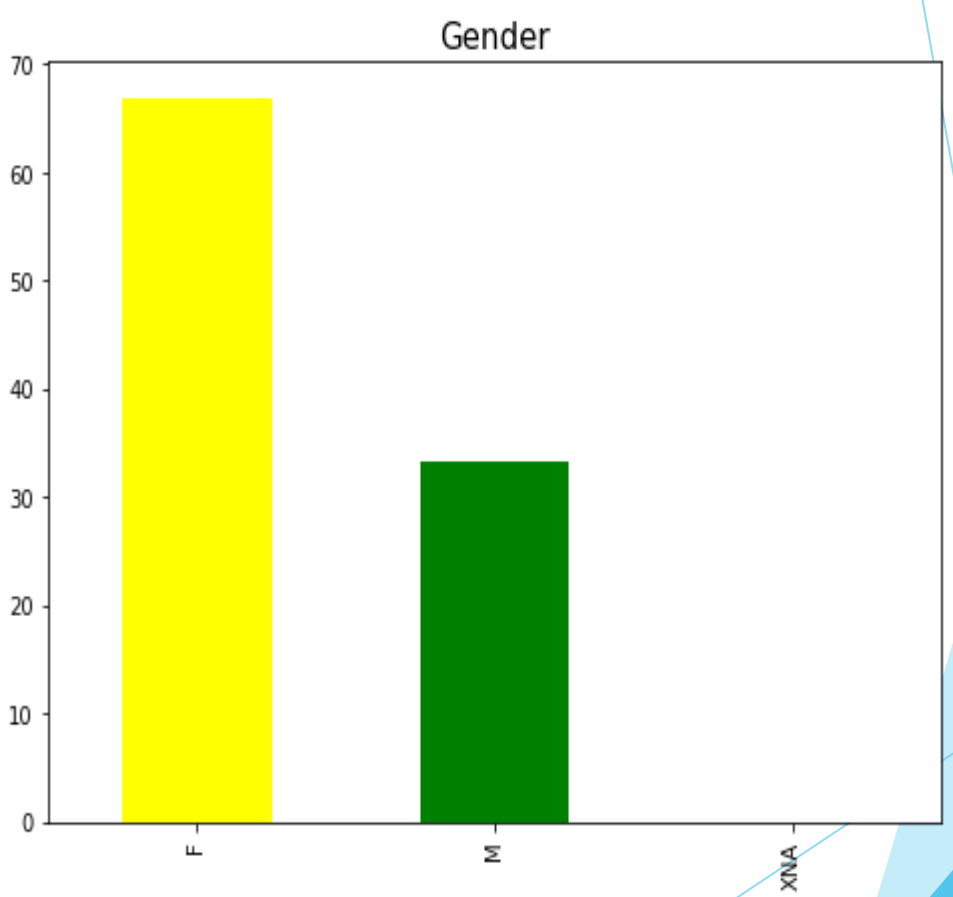



0-Non Defaulters
1-Defaulters

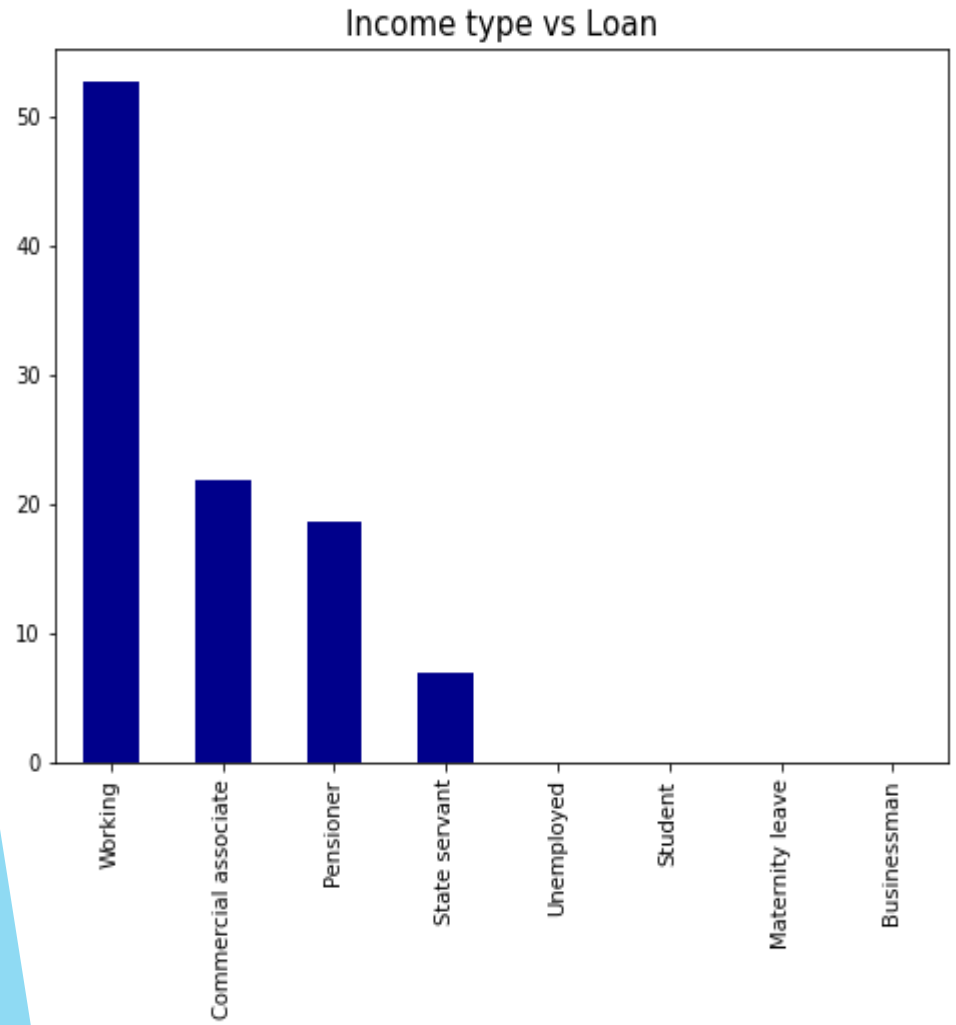
We can observe that there is target imbalance.
There are 91.72% non defaulters and 8.28% defaulters.

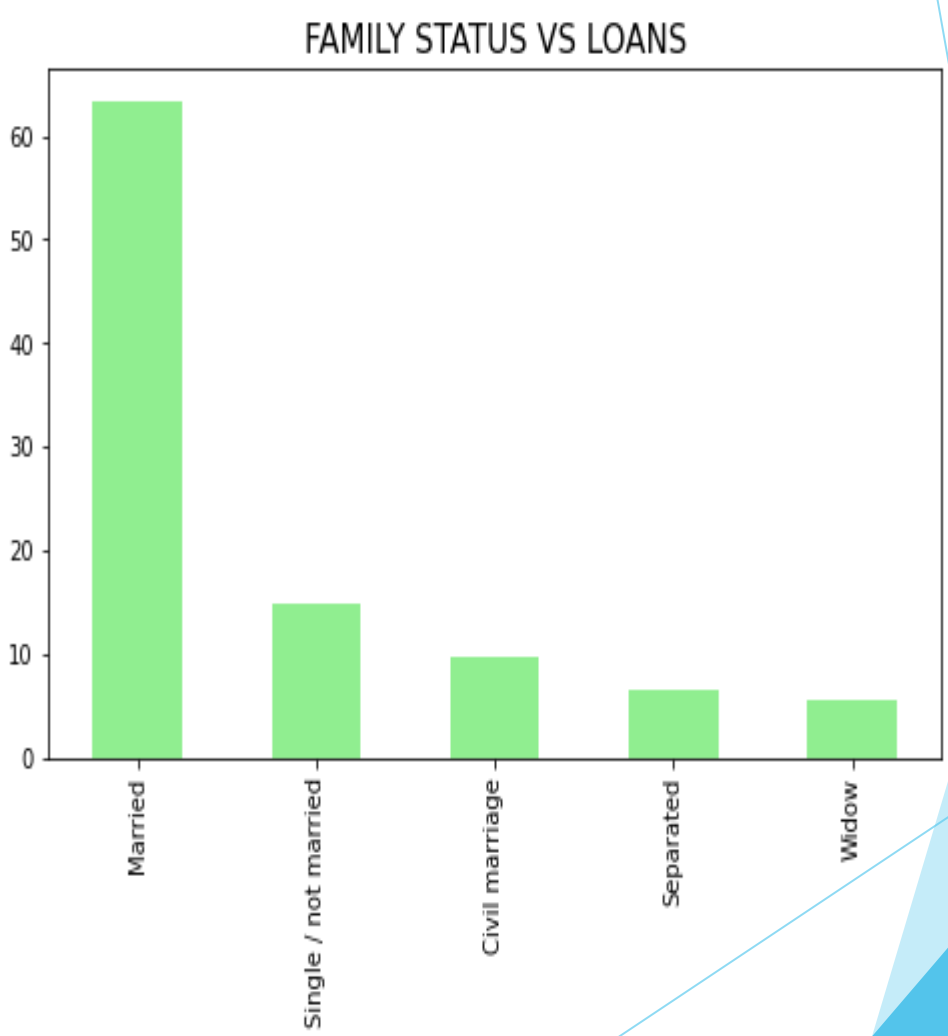
# UNIVARIATE ANALYSIS



# UNIVARIATE ANALYSIS(Cont..)

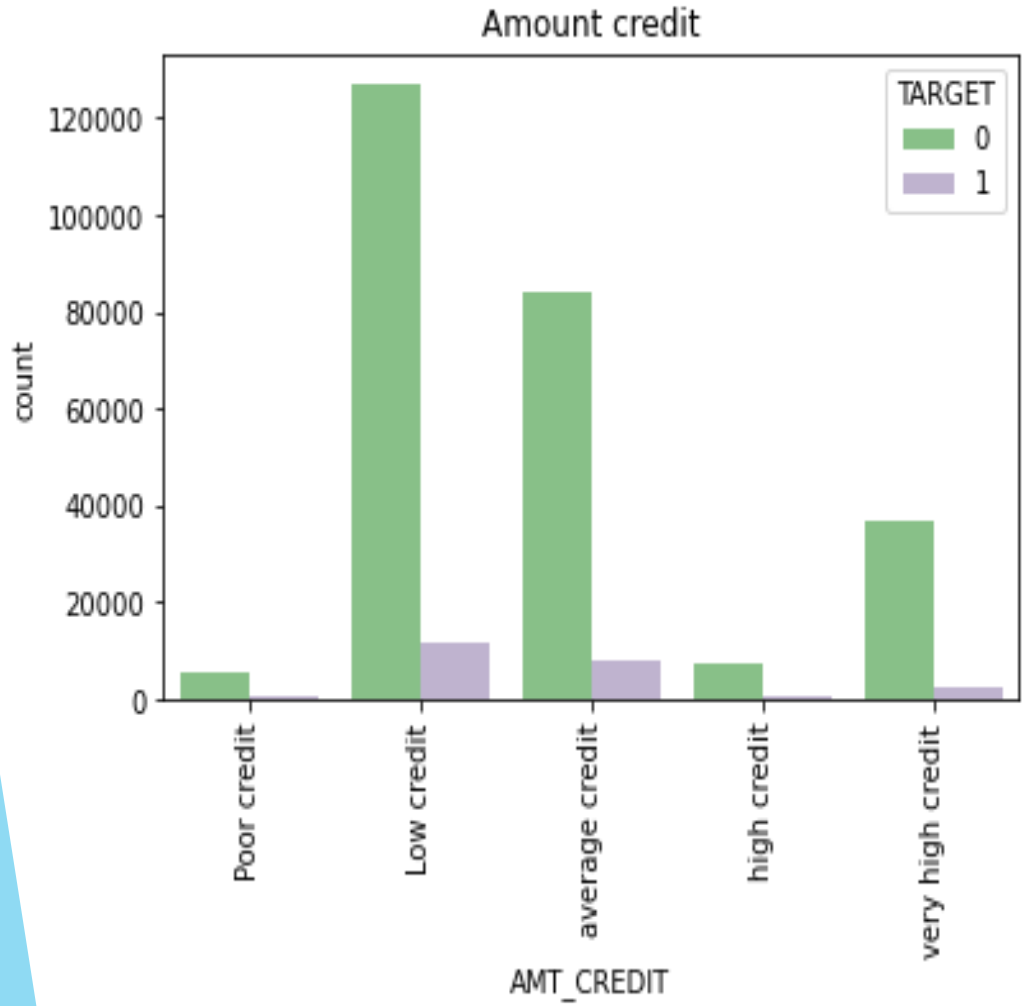


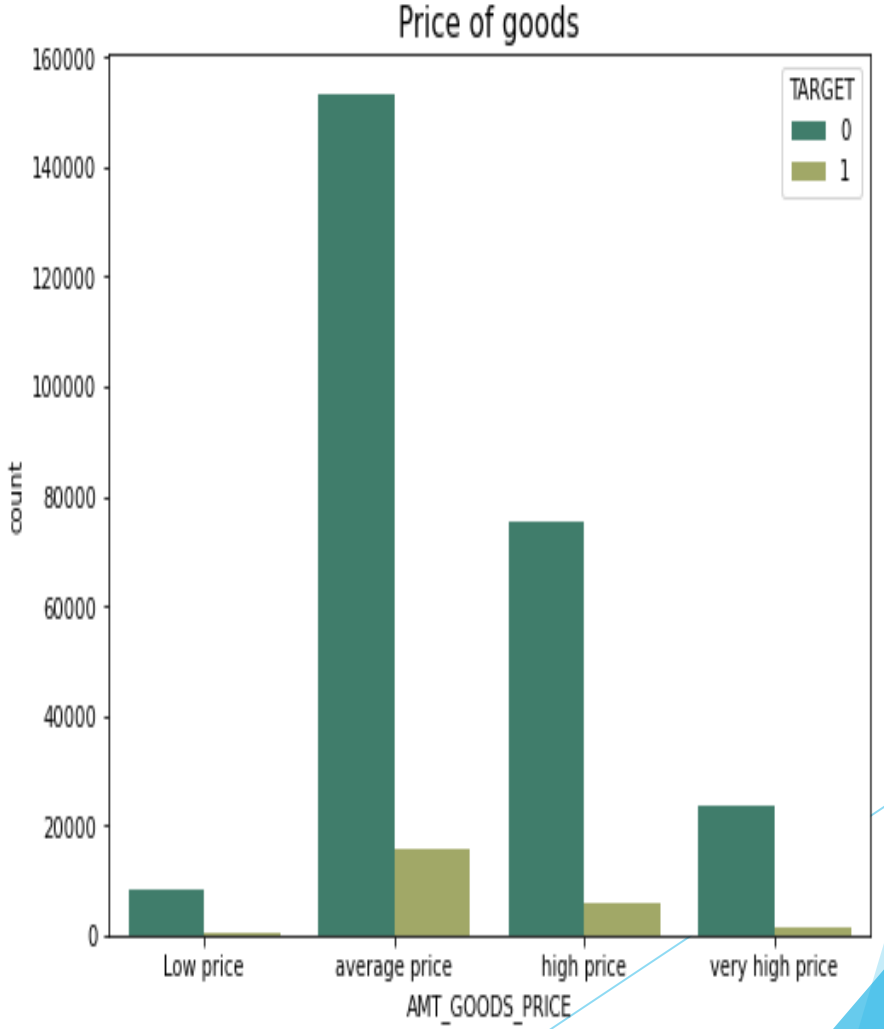



# Some INSIGHTS from Univariate Analysis

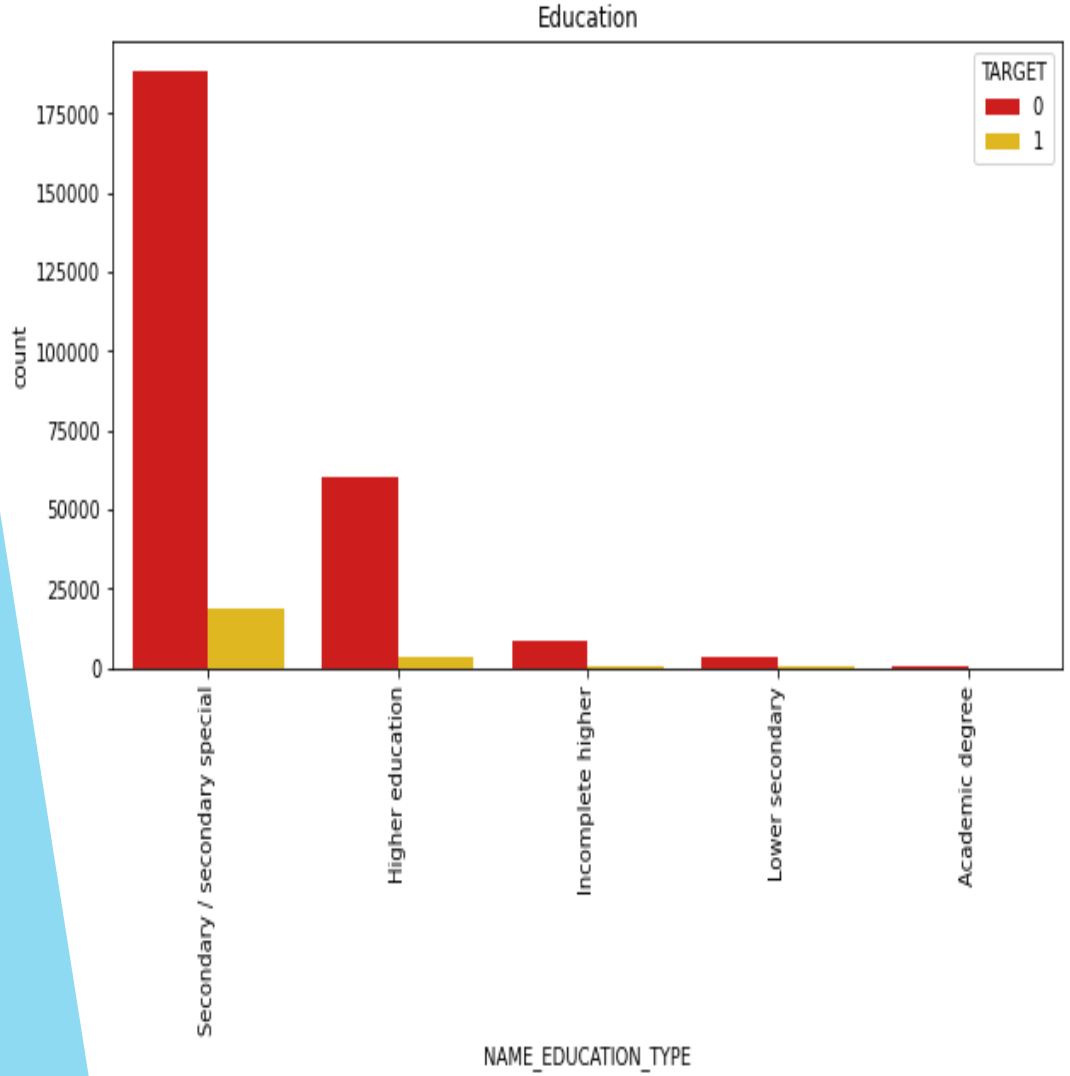
- Customers have taken more cash loans than revolving loans.
- Females have applied more loans than males.
- Working people have applied for more loans.
- Married Customers have taken more loans.

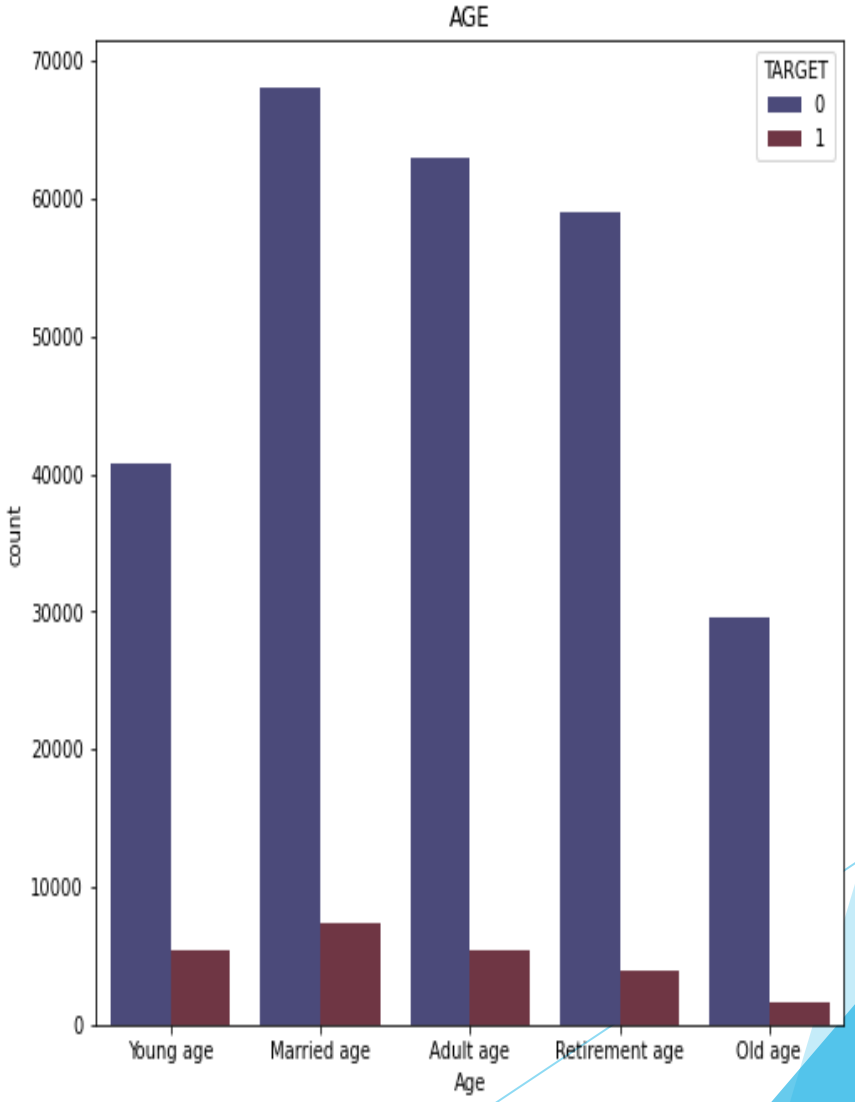
# BIVARIATE ANALYSIS USING TARGET VARIABLE



# BIVARIATE ANALYSIS USING TARGET VARIABLE(Cont..)

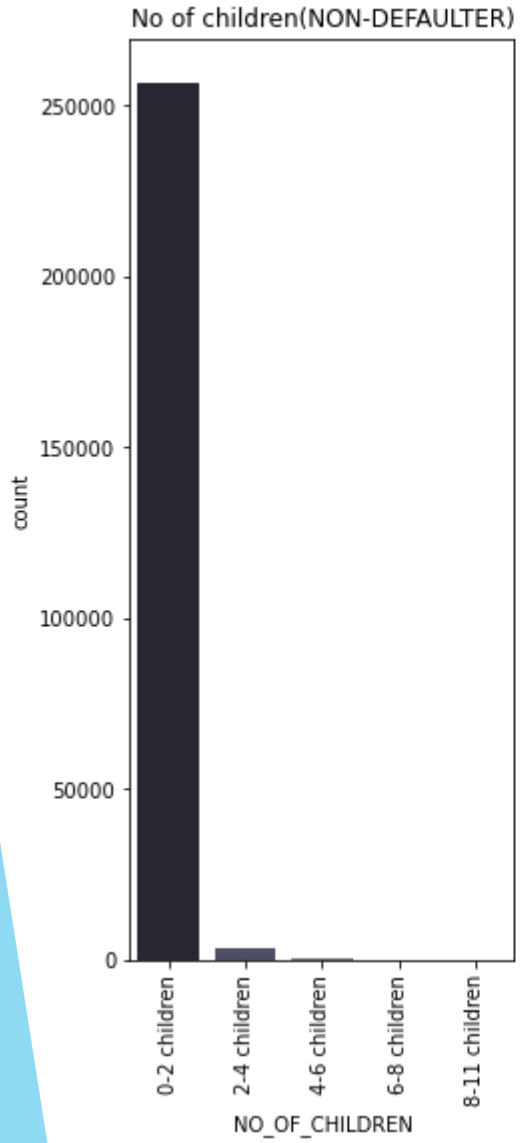


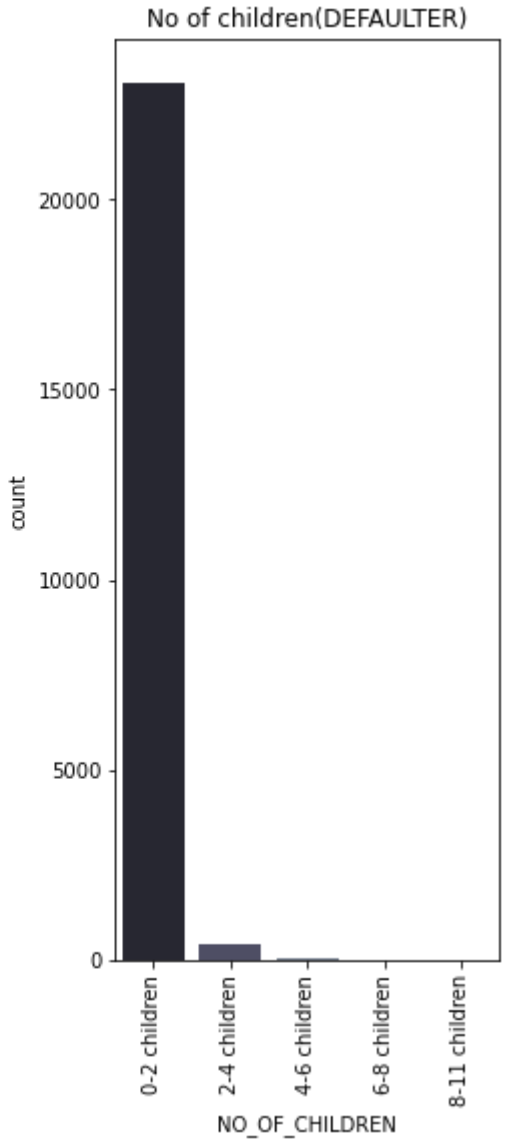



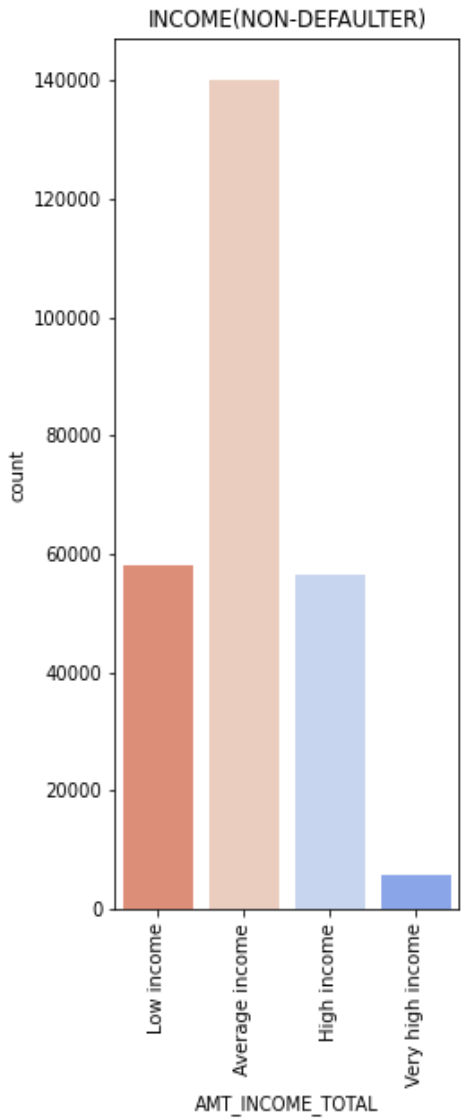
# INSIGHTS FROM BIVARIATE ANALYSIS

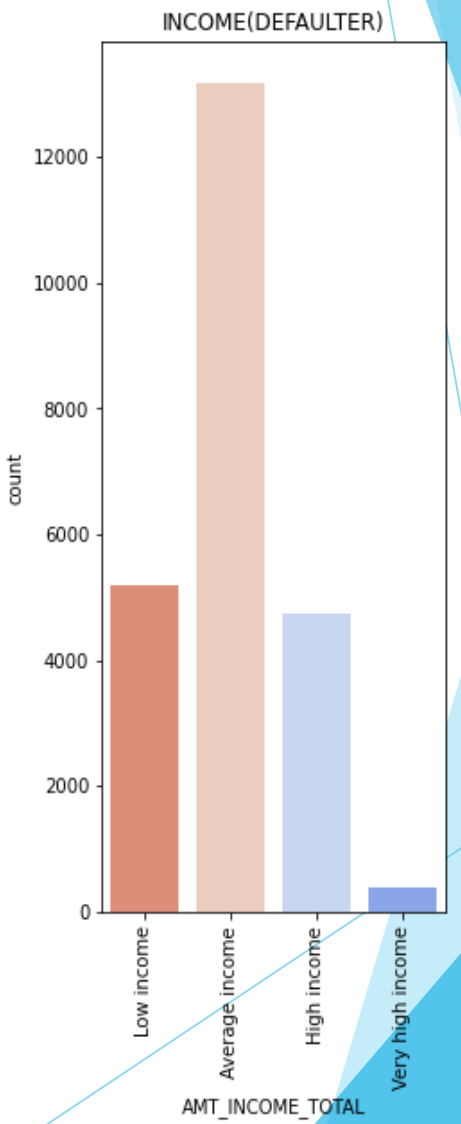
- Customers with low credit have taken more loans.
- Most loans were taken for average price of goods.
- Customers with minimum Secondary education have taken most loans
- Customers with married age(30-40) took most loans.

# SEGMENTED UNIVARIATE ANALYSIS

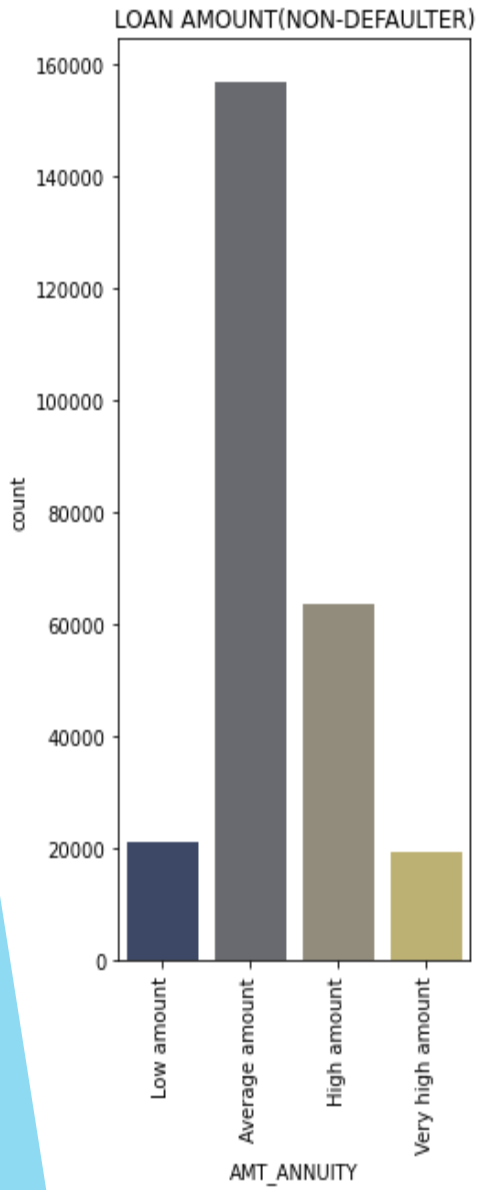


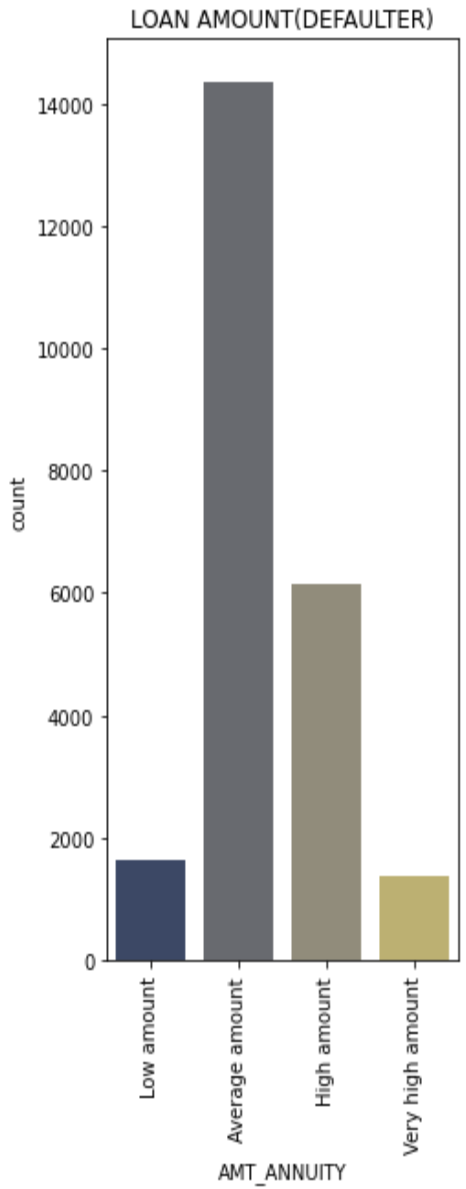



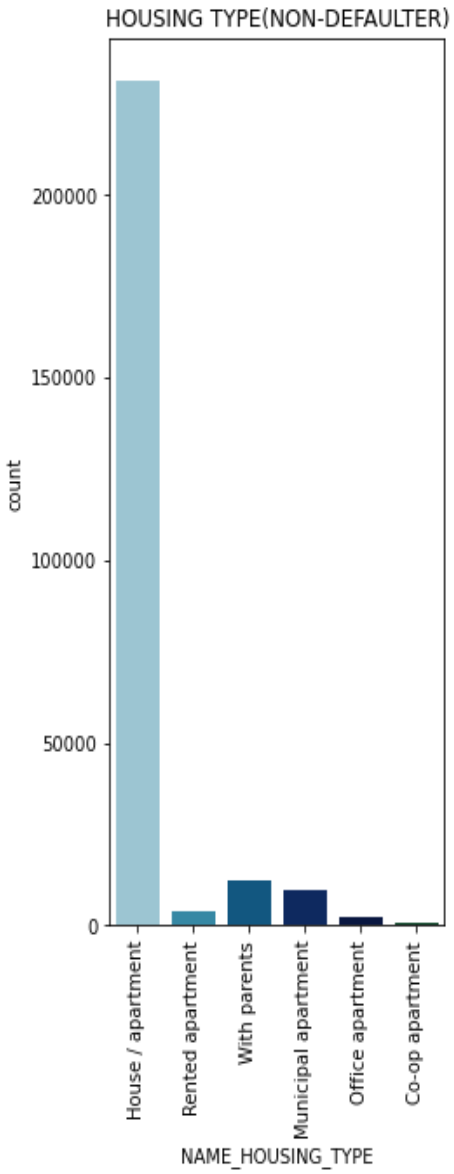



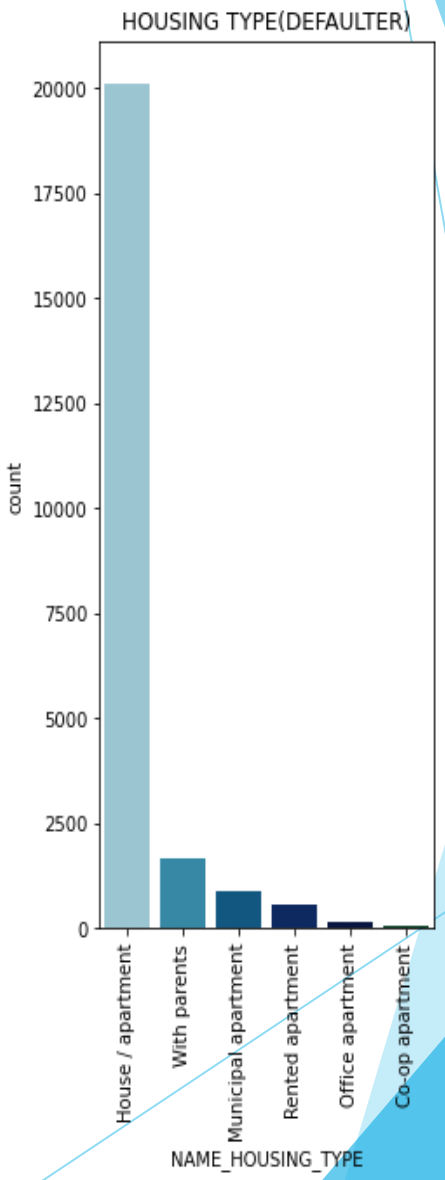



# SEGMENTED UNIVARIATE ANALYSIS(Cont..)



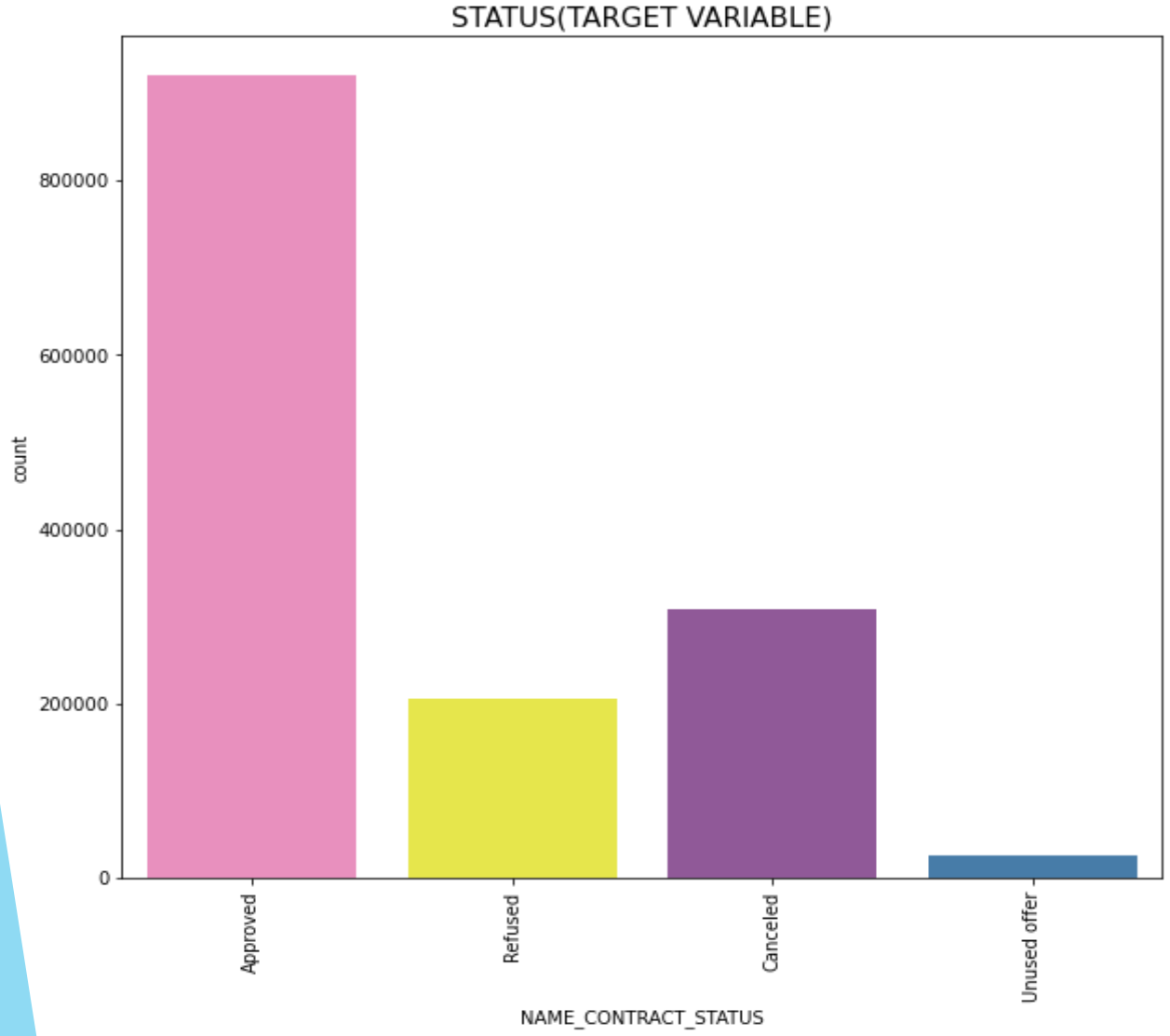



# INSIGHTS FROM SEGMENTED UNIVARIATE ANALYSIS

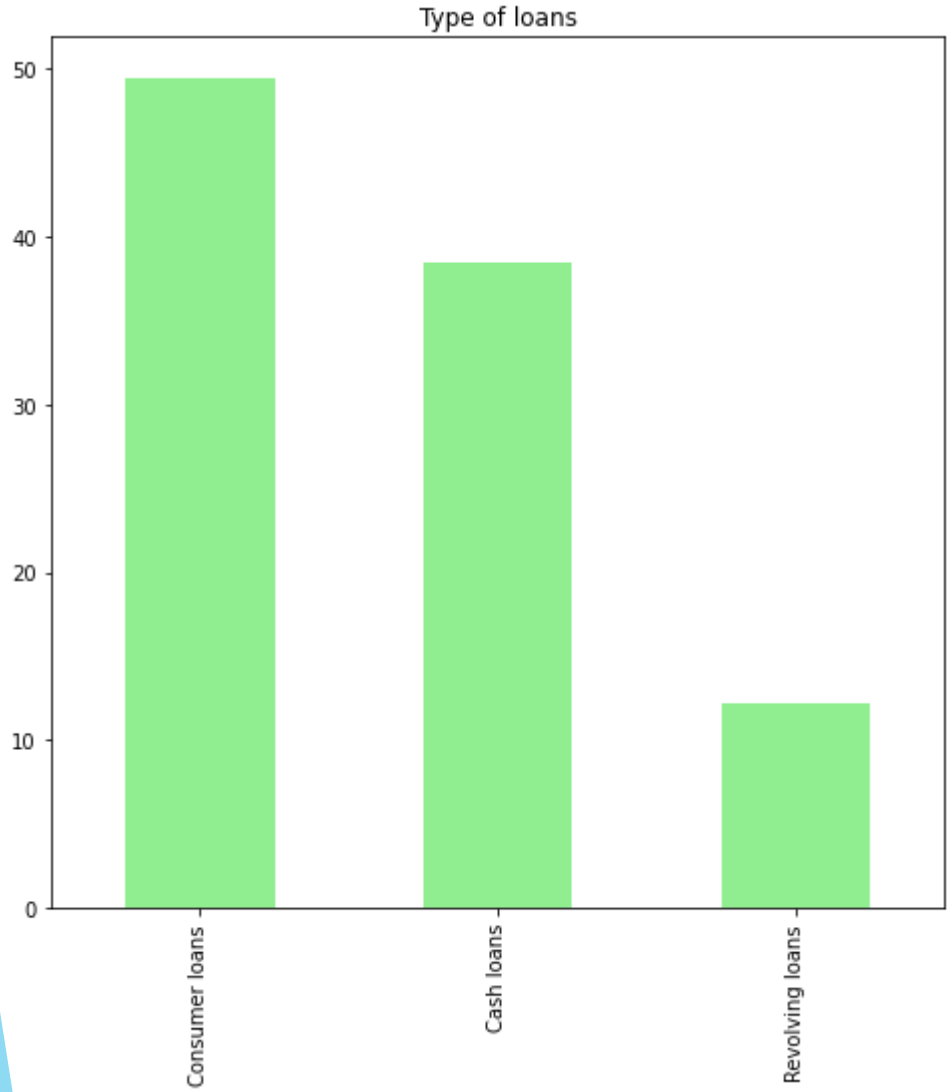
- Most loans were taken by customers who had 0-2 children.
- Most loans were taken by customers who had average income.
- Most customers took loan with average amount of loan.
- Most loans were taken by customers who stayed in a house/apartment.

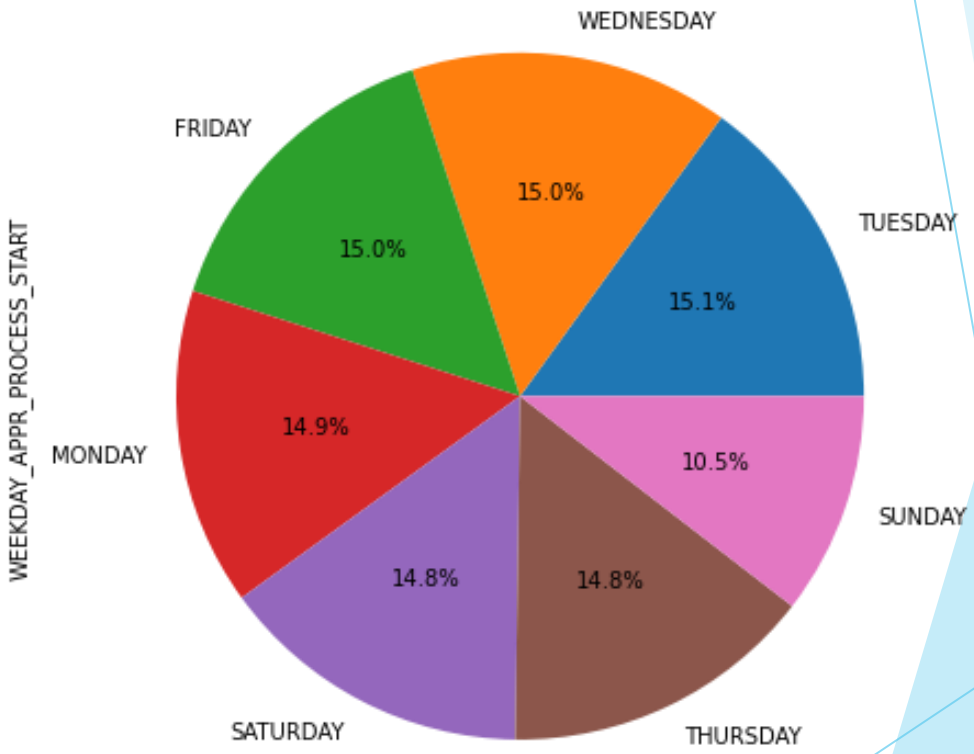
# UNIVARIATE ANALYIS OF TARGET VARIABLE(NAME_CONTRACT_STATUS)



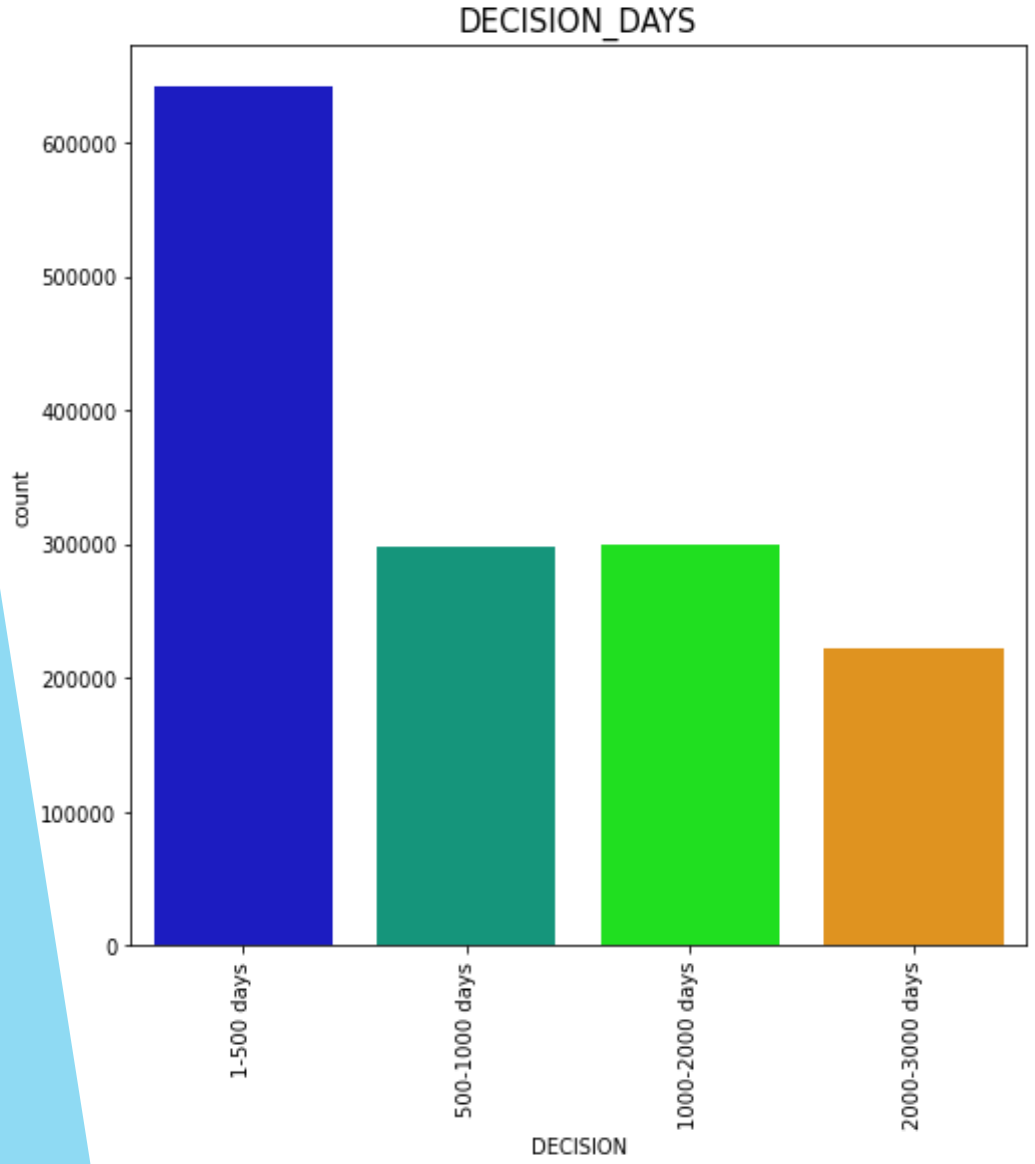

We can observe that most loans were approved.
More loans were cancelled than refused or unused.

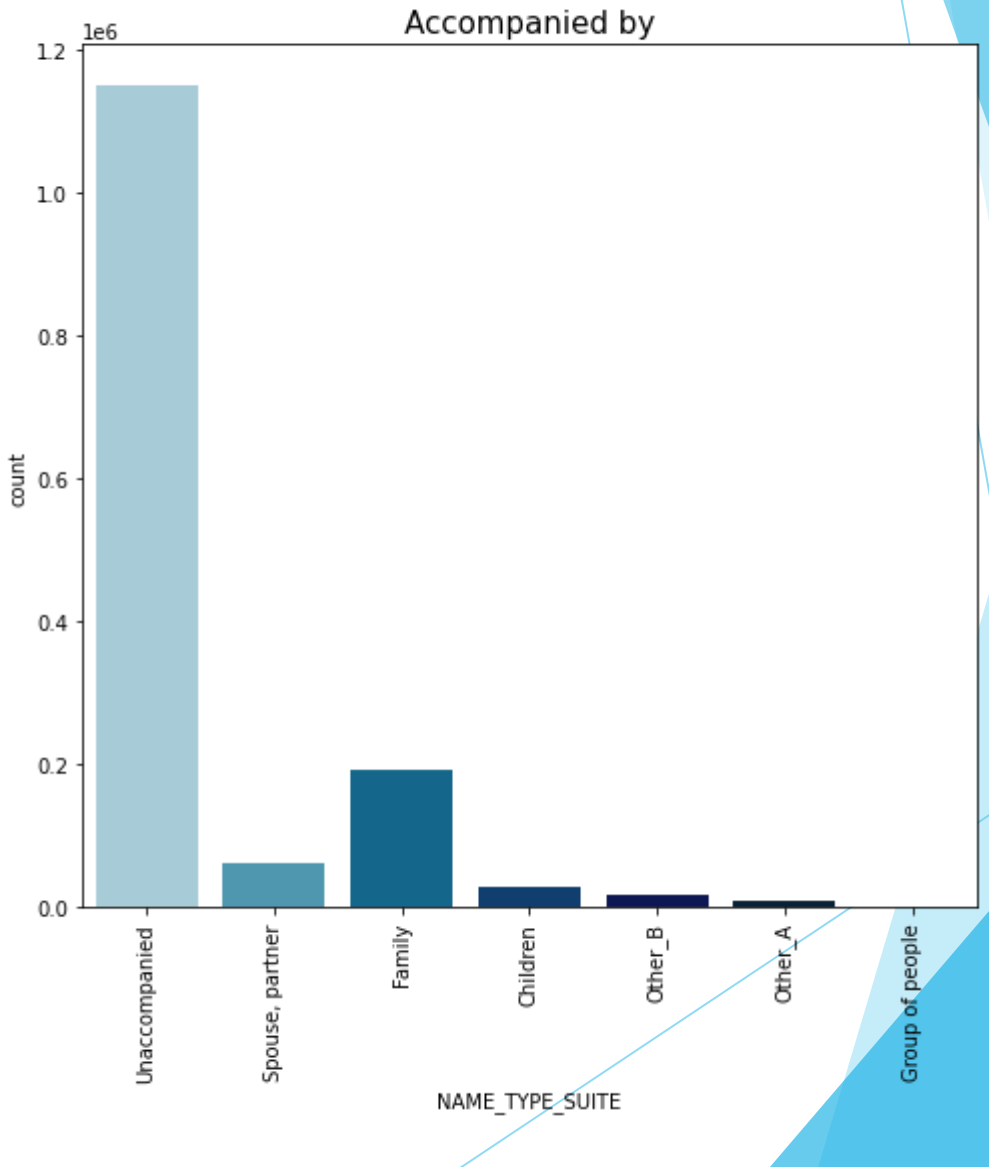
# UNIVARIATE ANALYSIS OF PREVIOUS APPLICATION DATA



# UNIVARIATE ANALYSIS OF PREVIOUS APPLICATION DATA(Cont..)

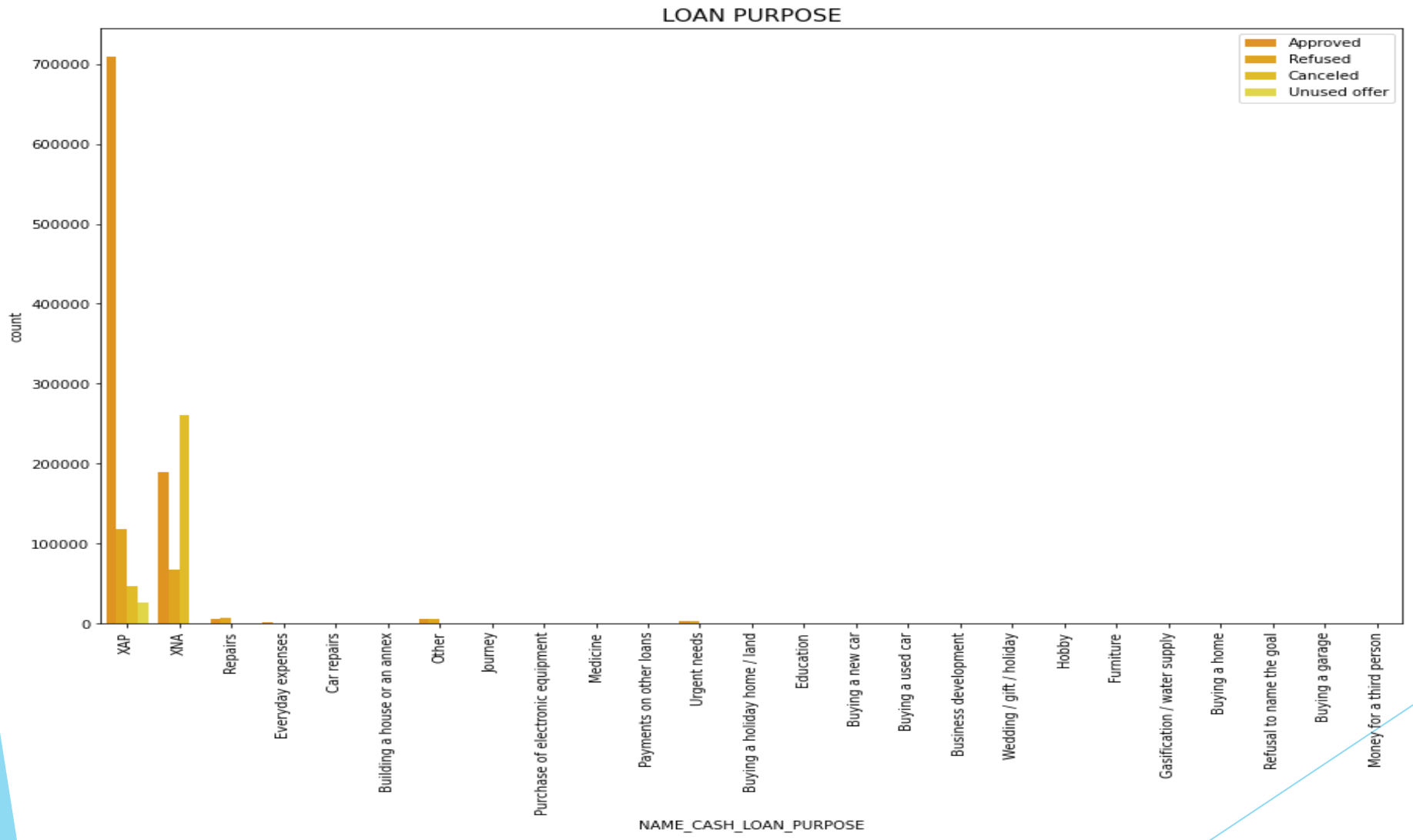


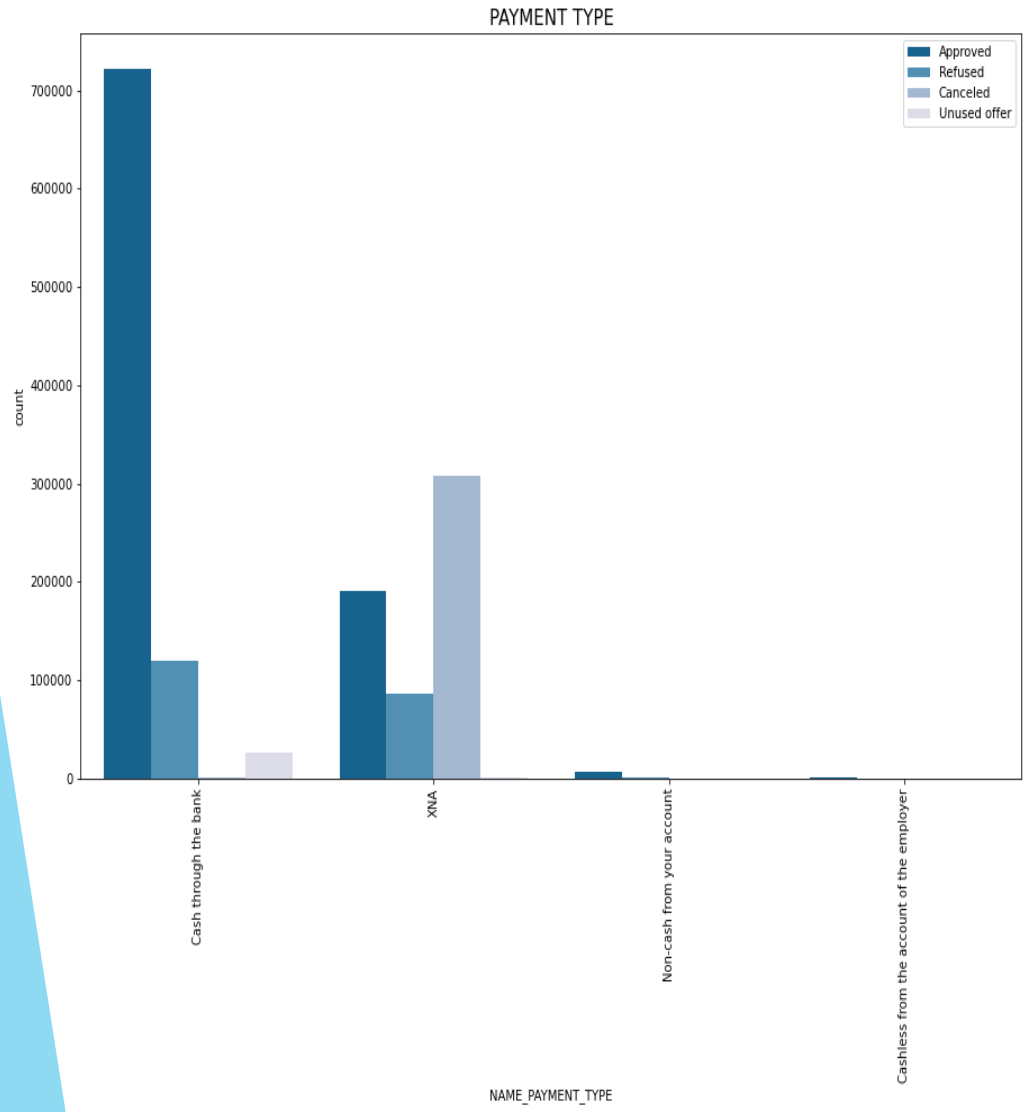
# INSIGHTS FROM UNIVARIATE ANALYSIS

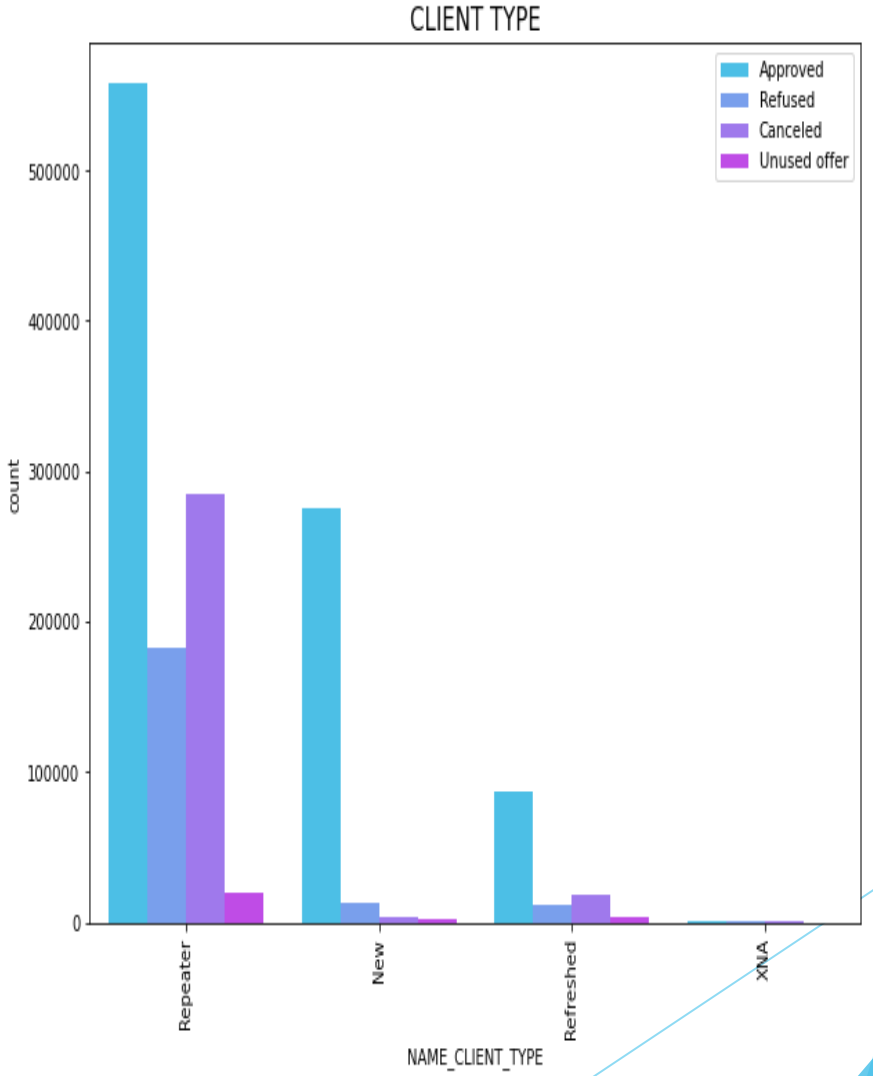
- We can see that most loans were consumer loans followed by cash loans. Least loans were revolving loans.
- We can see that most loans were applied on Tuesday.
- The average time to taken decision of a loan is 1-500 days.
- Most loans were taken by customers who were accompanied by no one.

# BIVARIATE ANALYSIS OF PEVIOUS APPLICATION DATA



# BIVARIATE ANALYSIS OF PEVIOUS APPLICATION DATA(Cont..)

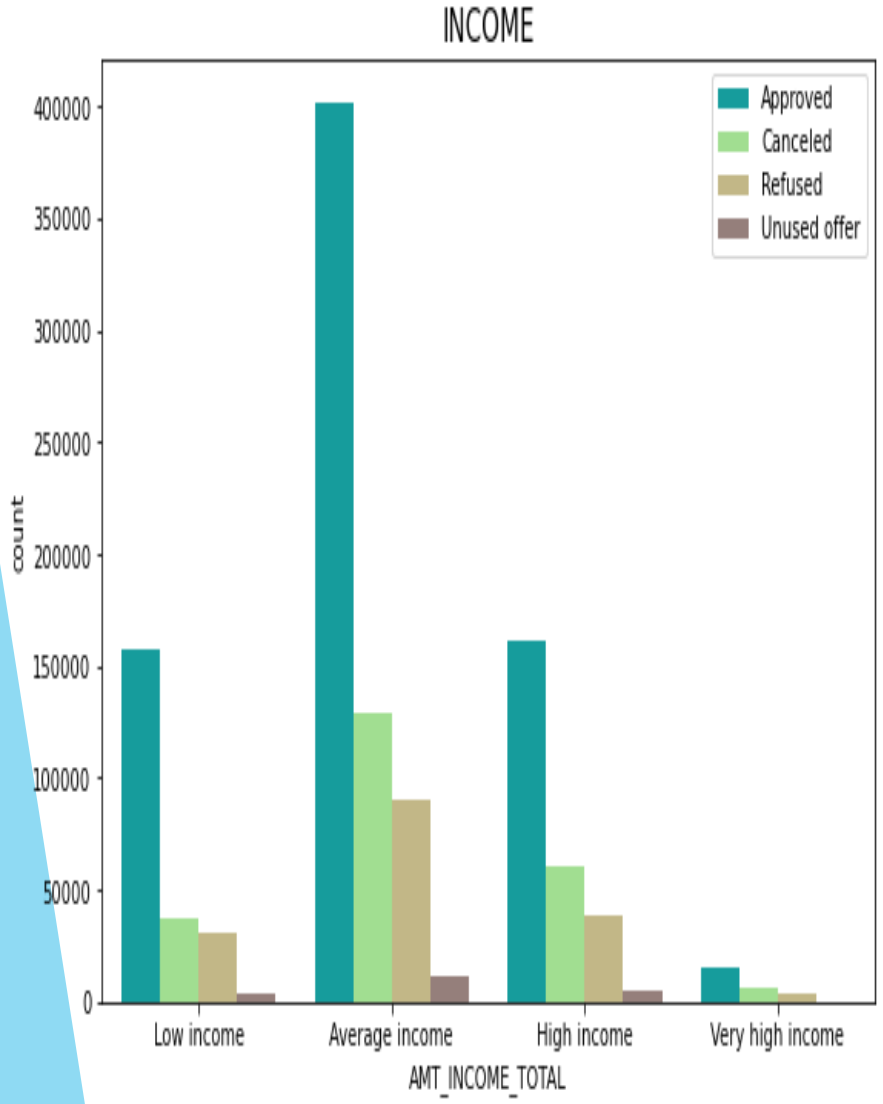


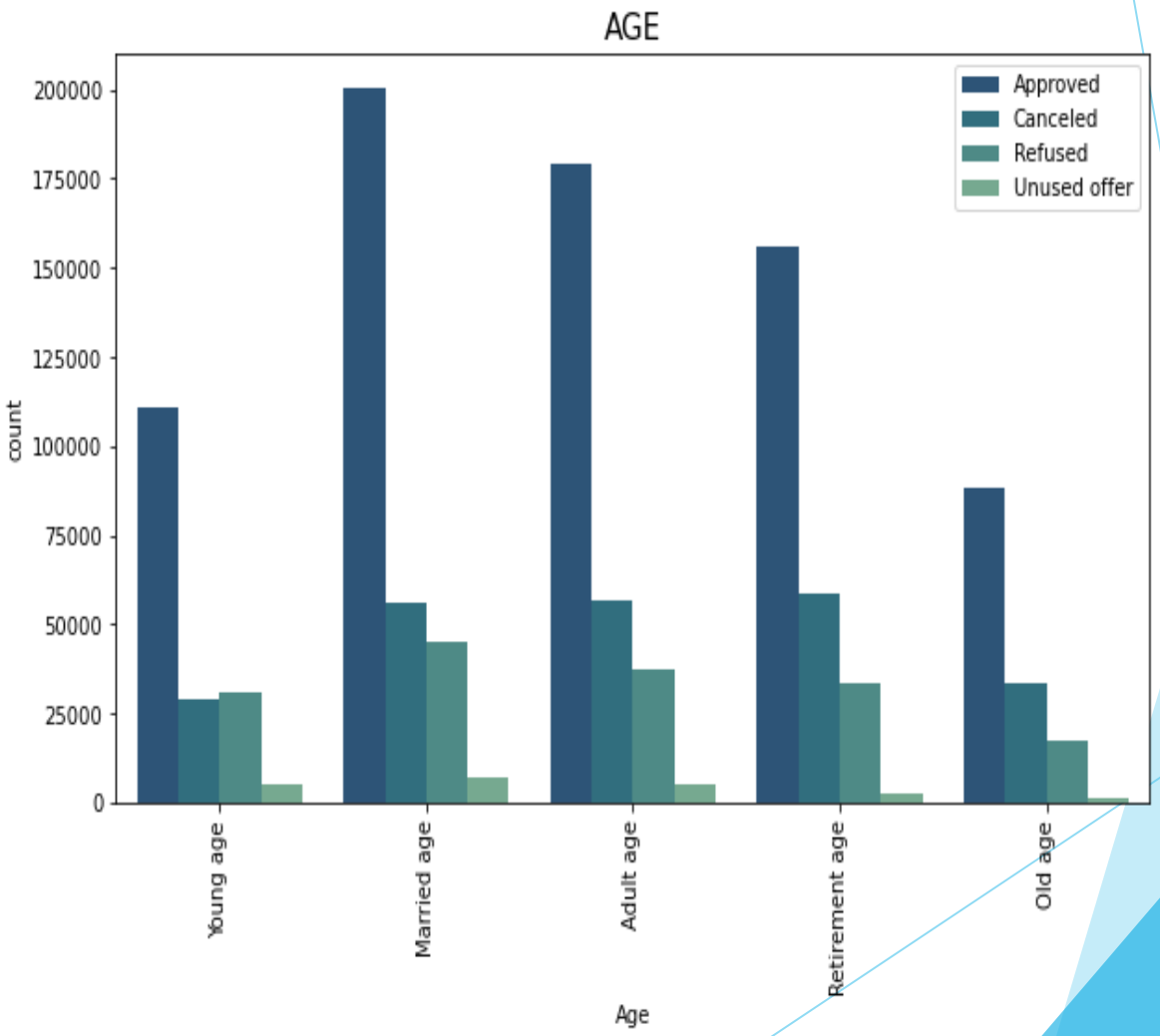



# INSIGHTS from Bivariate Analysis

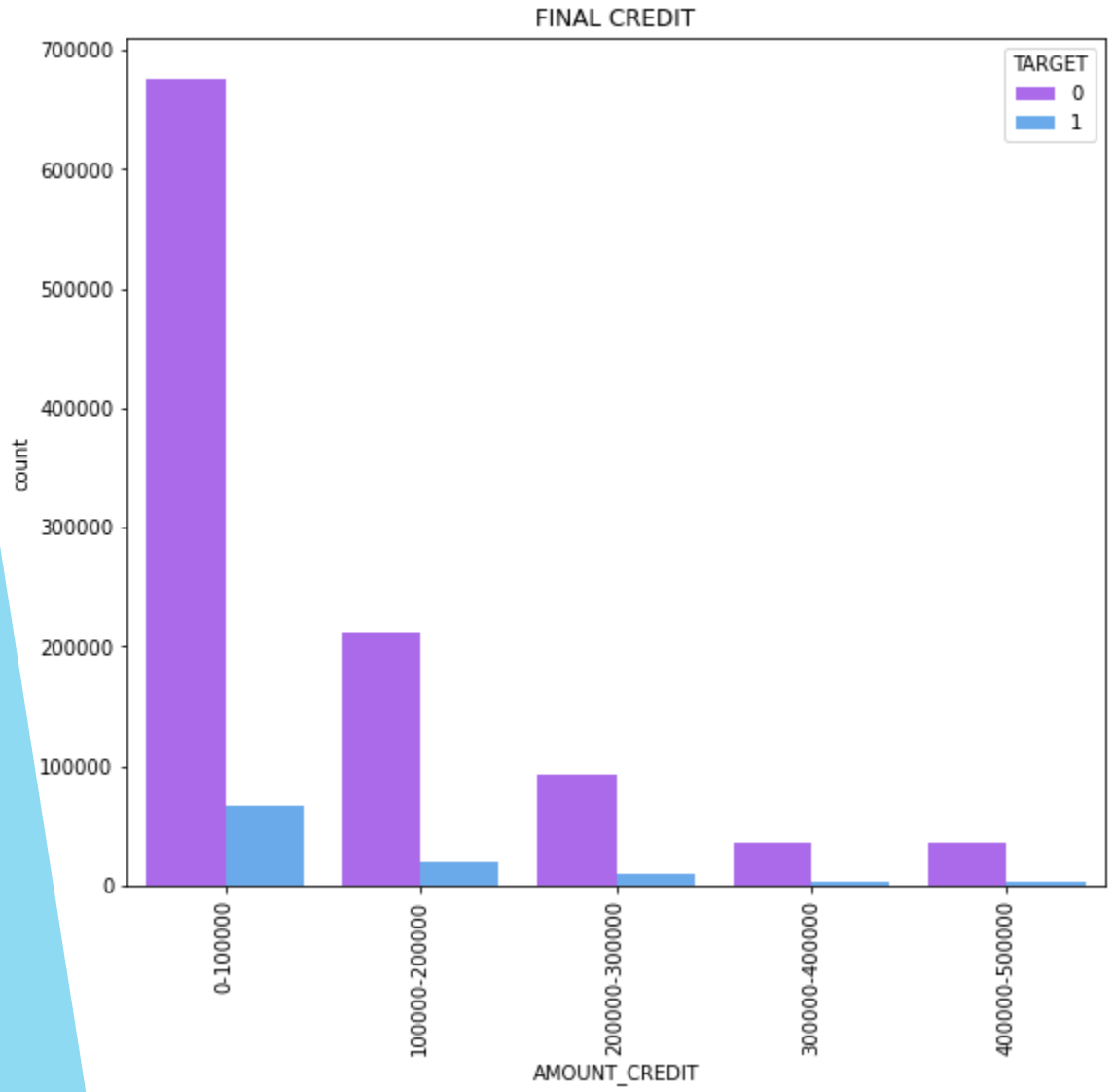
- Most loans approved were for the purpose of XAP.
- Most loans approved were taken from cash through bank.
- Most customers whose loans were approved were repeaters.

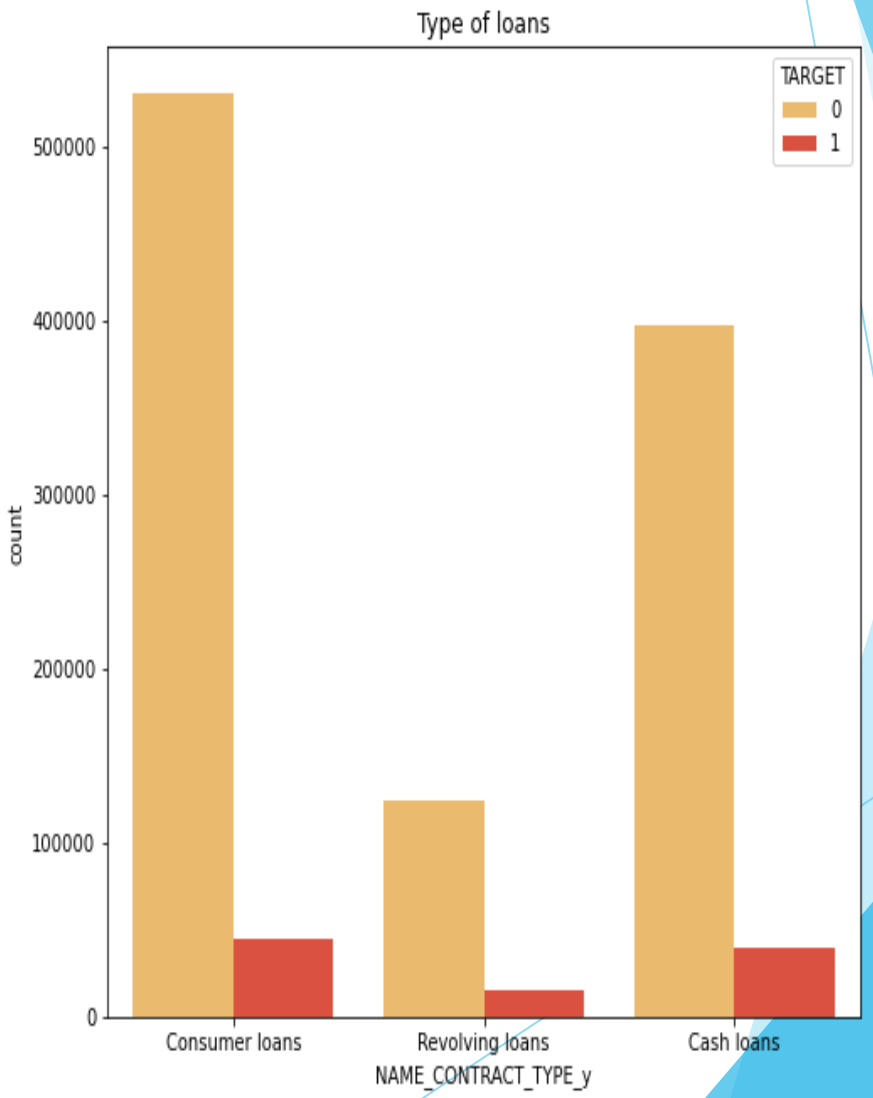
# MERGED DATA BIVARIATE ANALYSIS



# MERGED DATA BIVARIATE ANALYSIS

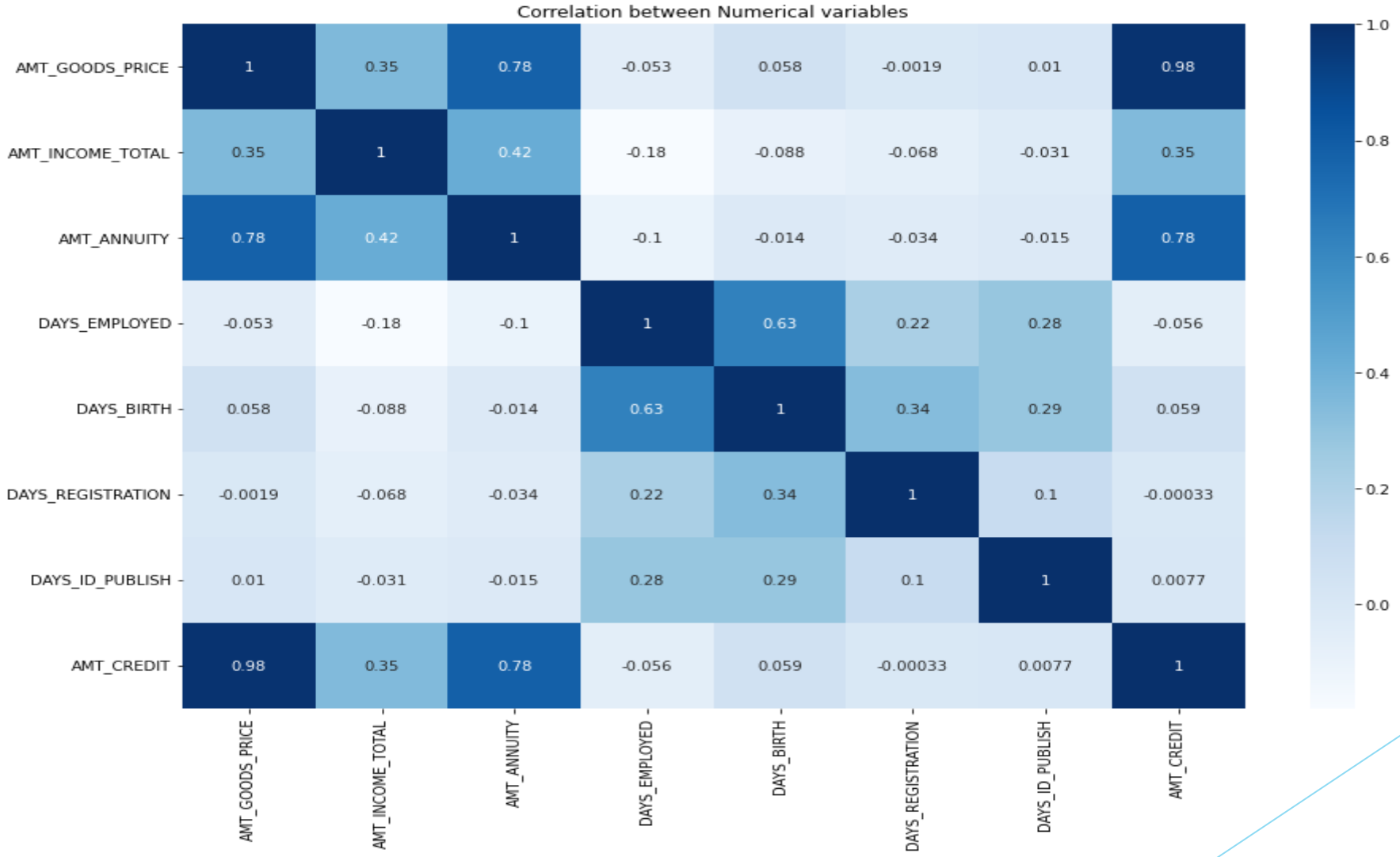


# INSIGHTS

- We can see that most loans were approved for customers having average income.
- Customer of married age(30-40) got most loans.
- Most loans were taken by customers who had final credit of 0-100000.
- Consumer customer took consumer loans followed by cash loans and least loans were revolving loans.

# Correlation of Numerical data



# INSIGHTS of correlated data

- We can see there is a very HIGH correlation between AMT_GOODS_PRICE and AMT_CREDIT.
- There is a good correlation between AMT_CREDIT and AMT_ANNUITY.
- There is very low correlation between DAYS_ID_PUBLISH and DAYS_REGISTERATION.
- There is negative correlation between AMT_CREDIT AND DAYS_EMPLOYED,AMT_CREDIT AND DAYS_REGISTERATION,DAYS_ID_PUBLISH and AMT_INCOME_TOTAL,etc.

# Conclusion

- There is target imbalance. There were 91.72% non defaulters and 8.28% defaulters.
- Most loans taken were consumer loans for previous application. For application data most loans taken were Cash loans.
- There were more females who applied for loans than males.
- Customers who are working applied for most loans.
- Customers who are married applied for most loans.
- Customers with low credit have taken more loans.
- Most loans were taken for average price of goods.
- Customers with minimum Secondary education have taken most loans
- Most loans were taken by customers who had 0-2 children.
- Most customers took loan with average amount of loan.
- Most loans were taken by customers who stayed in a house/apartment.

# Conclusion(Cont..)

- We can observe that most loans were approved. (TARGET VARIABLE ANALYSIS)
- More loans were cancelled than refused or unused
- We can see that most loans were applied on Tuesday.
- The average time to taken decision of a loan is 1-500 days.
- Most loans were taken by customers who were accompanied by no one.
- Most loans approved were for the purpose of XAP.
- Most loans approved were taken from cash through bank.
- Most customers whose loans were approved were repeaters.
- We can see there is a very HIGH correlation between AMT_GOODS_PRICE and AMT_CREDIT.
- There is a good correlation between AMT_CREDIT and AMT_ANNUITY.
- There is very low correlation between DAYS_ID_PUBLISH and DAYS_REGISTERATION.
- There is negative correlation between AMT_CREDIT AND DAYS_EMPLOYED,AMT_CREDIT AND DAYS_REGISTERATION,DAYS_ID_PUBLISH and AMT_INCOME_TOTAL,etc.