

The background of the slide features a dark blue gradient with a faint, stylized financial chart. The chart includes a line graph with circular markers and a series of vertical bars, resembling a candlestick or bar chart, in a lighter blue color. The overall aesthetic is professional and data-oriented.

PRESENTATION ON CREDIT EDA ASSIGNMENT

BY-ANURADHA BHARTI

Introduction

This Report Presents An Exploratory Data Analysis Conducted To Identify Patterns That Indicate If A Customer Is Likely To Have Difficulty Repaying Their Loans. The Analysis Aims To Support Decision-making Processes, Such As Denying A Loan, Reducing The Loan Amount, Or Setting Higher Interest Rates.

Business Understanding


- Understand The Driving Factors That Strongly Indicate Loan Default.
- Provide Insights For Portfolio And Risk Assessment.
- Aid In Improving Business Decisions For Loan Approvals.

The background of the image features a light blue gradient with faint, stylized financial data visualizations. These include several line graphs with circular markers at data points and a bar chart with numerous vertical bars of varying heights. The overall aesthetic is clean and professional, typical of a business or data-related presentation.

Analysis Of Data Of The Client At The Time Of Application

Data understanding, Data Cleaning and Manipulation

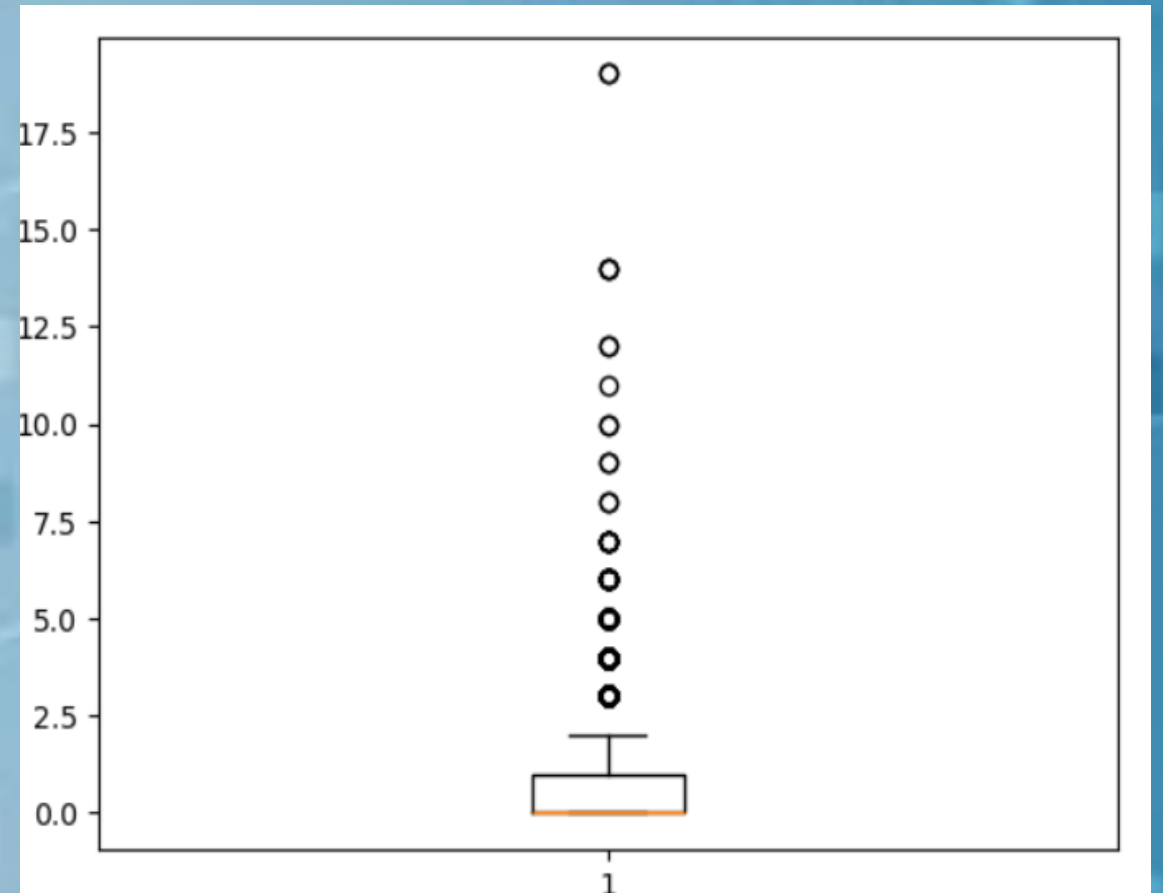
- These Process Involved Data understanding, Data Cleaning, Transformation, And Visualization To Identify Key Patterns And Relationships. Missing Data Was Addressed Using Appropriate Imputation Techniques.
- Example:
 - 'Amt_annuity': Imputed With Median Value (24,903).
 - 'NAME_TYPE_SUITE': Imputed With Mode ('Unaccompanied').
- -'AMT_GOODS_PRICE' imputed with median value(450000.0).
- -in CODE_GENDER, 'XNA' values are handled by replacing with F.
- -in 'NAME_FAMILY_STATUS' , 'unknown' values are replaced by mode('Married')
- Negative Values In Columns Such As 'Days_birth' And 'Days_employed' Were Converted To Positive. Categorical Variables With less Than Three Unique Values Were Converted To Categorical Data Types

The background of the slide features a light blue gradient with faint, stylized financial data visualizations. These include several line graphs with circular markers at data points, some of which are connected by lines. There are also bar charts visible, particularly in the lower half of the image. The overall aesthetic is clean and professional, typical of a business or data science presentation.

Outlier Analysis

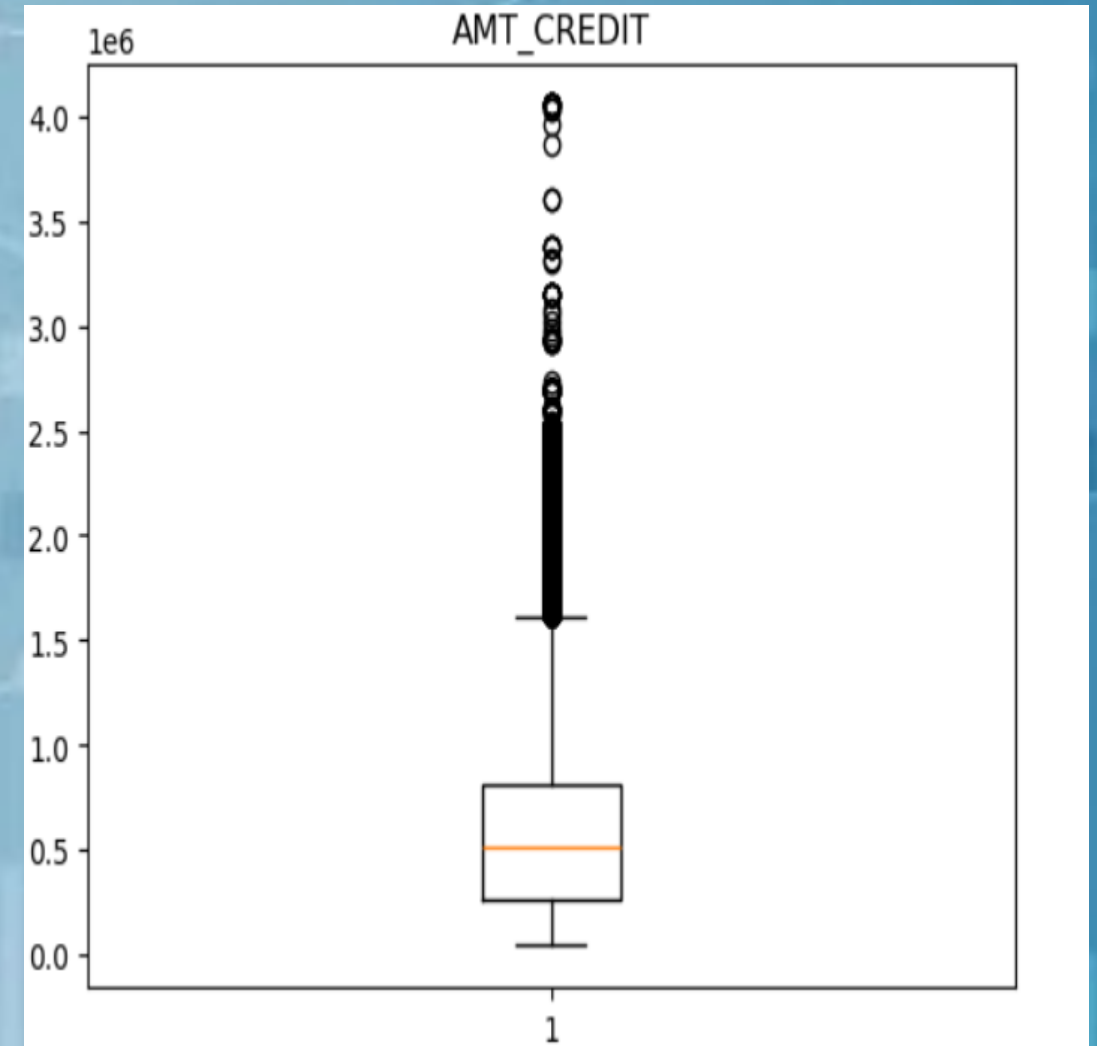
Outlier Analysis CNT_CHILDREN

- Insight- Values Greater Than 2.5 Are Outliers, As a person can't have Children greater than 10. So, we can say count Of Children Greater Than 10 is Outlier.



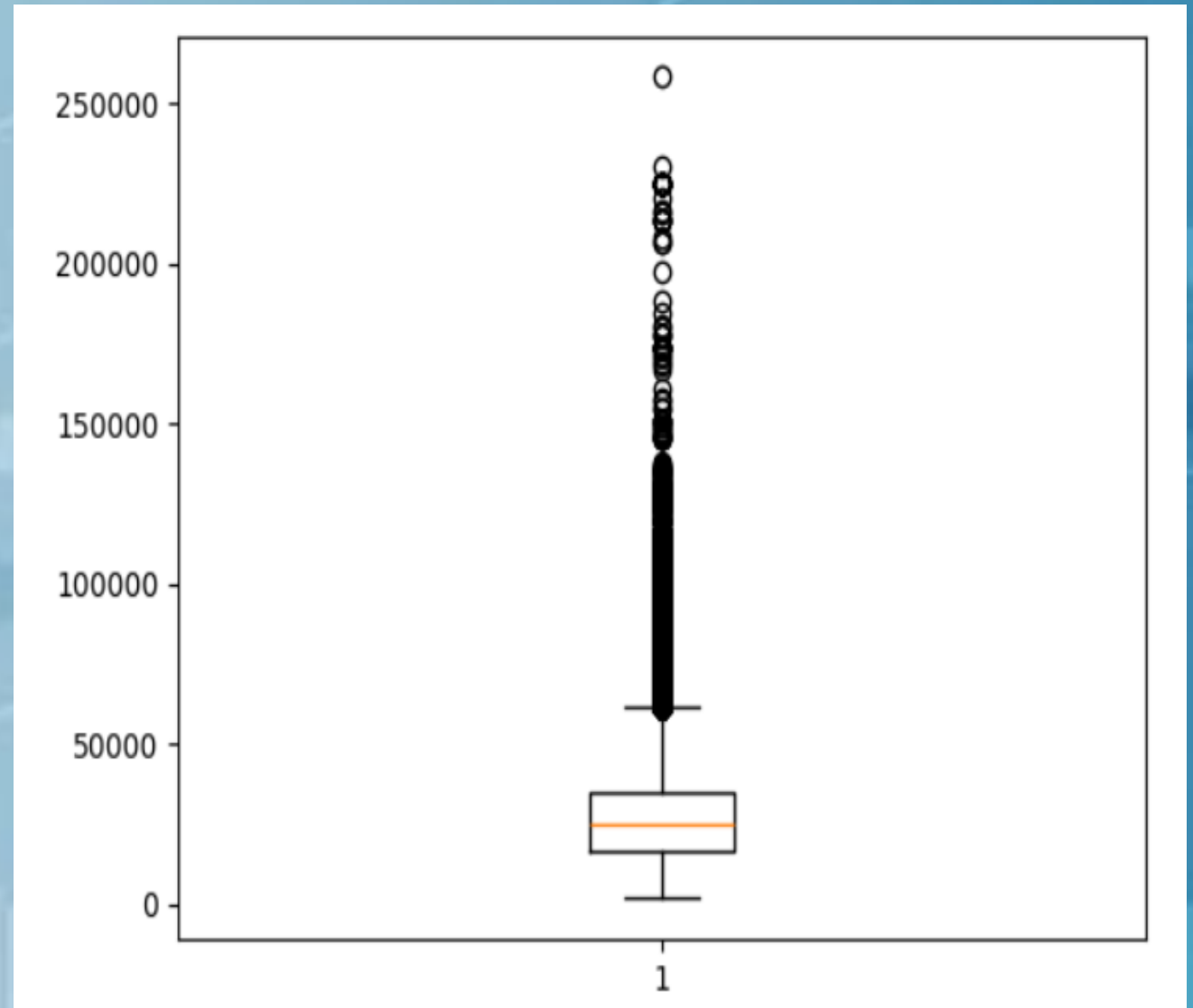
Outlier Analysis of AMT_CREDIT

- Insight- The Values Above The Upper Whisker Are Outliers, As Calculated Using The IQR Method, Value Above 1616625.0 Is An Outlier



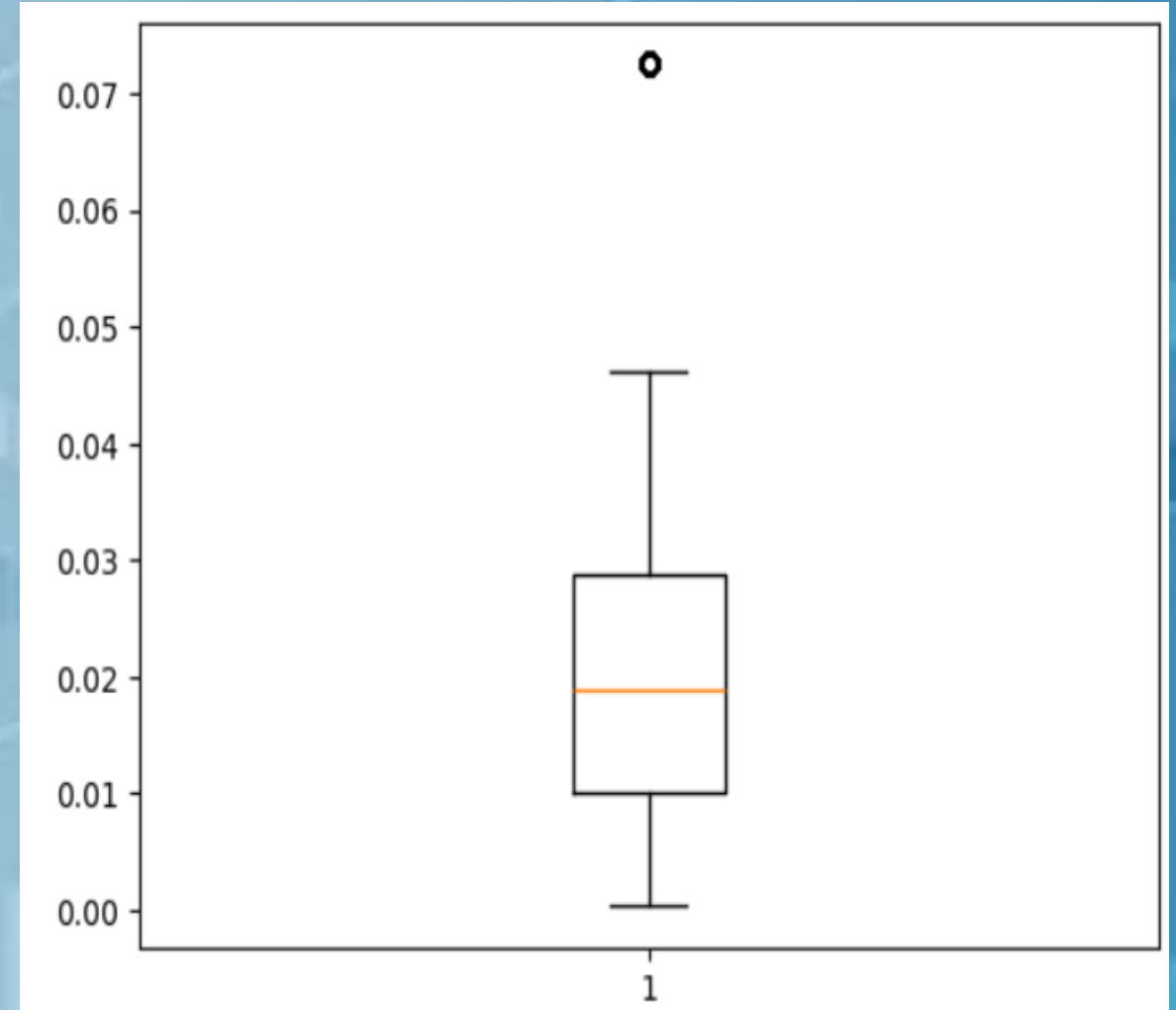
Outlier Analysis Of AMT_ANNUITY

- Insight- Population Relative Count Above Upper Whisker (More Than 61704.0) Is An Outlier



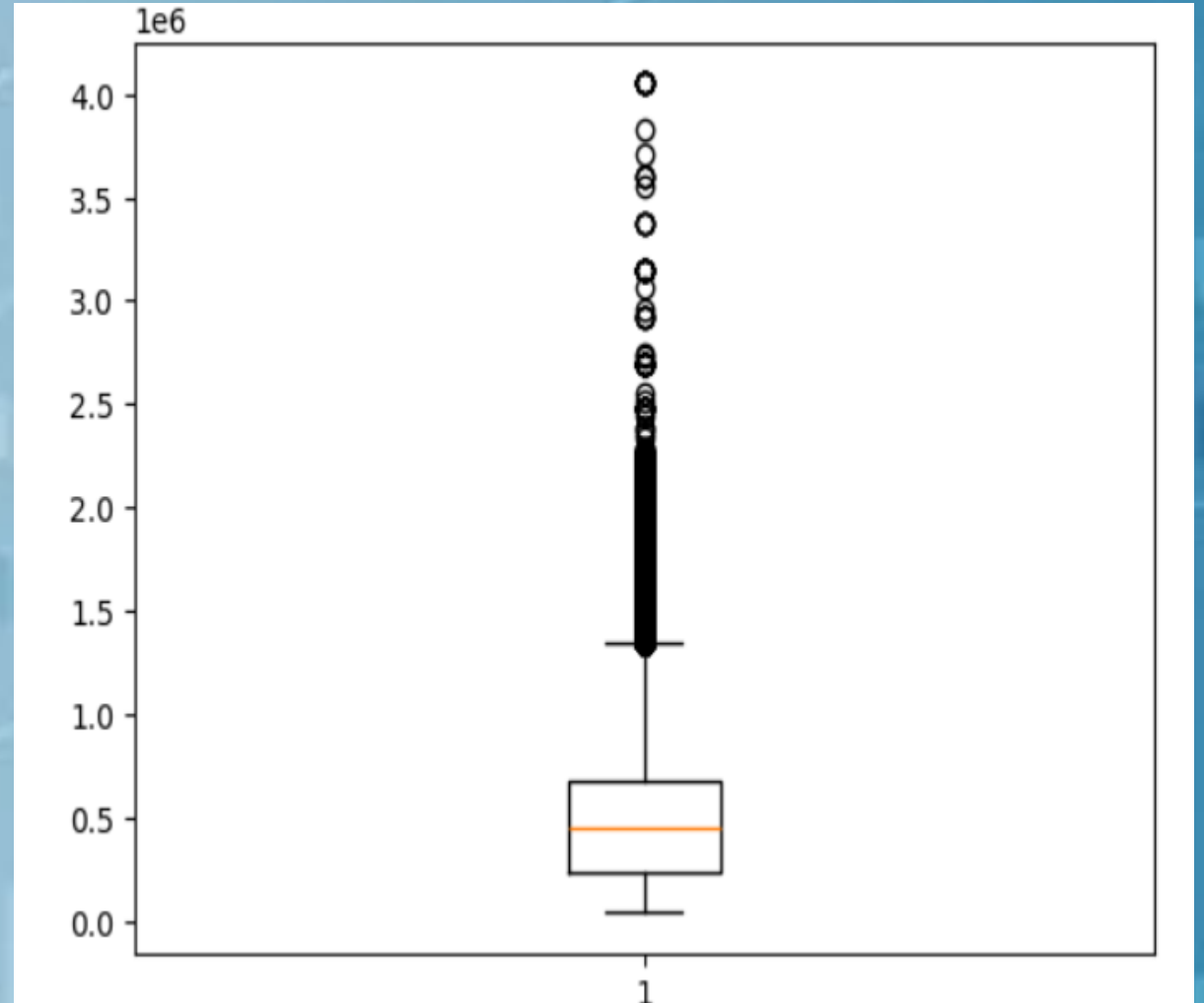
Outlier Analysis Of REGION_POPULATION_RELATIVE

- Insight- Population Relative Count More Than 0.056648500000000004 Is Outlier



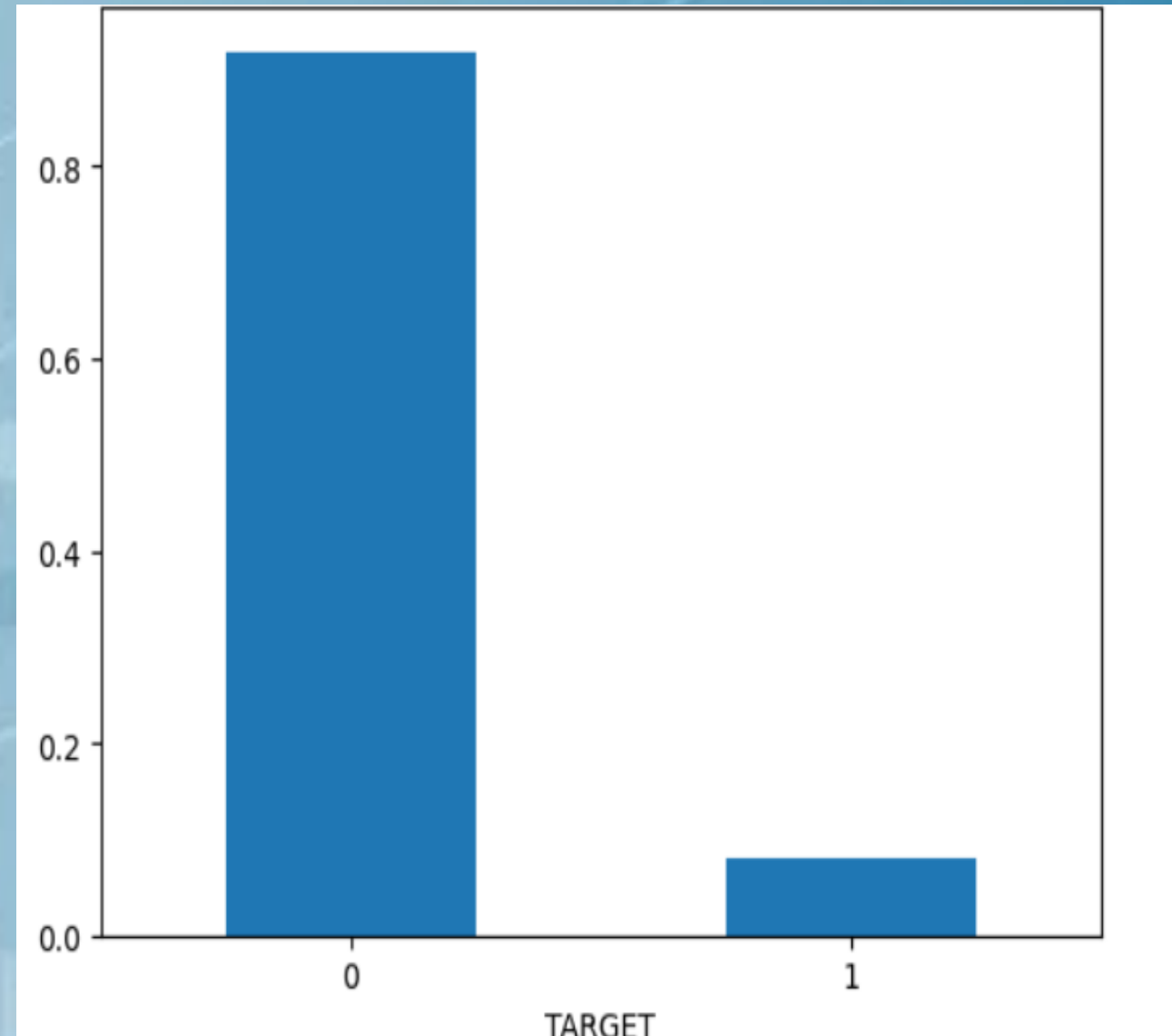
Outlier Analysis Of AMT_GOODS_PRICE

- Insight- AMT_GOODS_PRICE Greater Than 1341000.0 Is An Outlier As Calculated From IQR



Imbalance Percentage Of Target

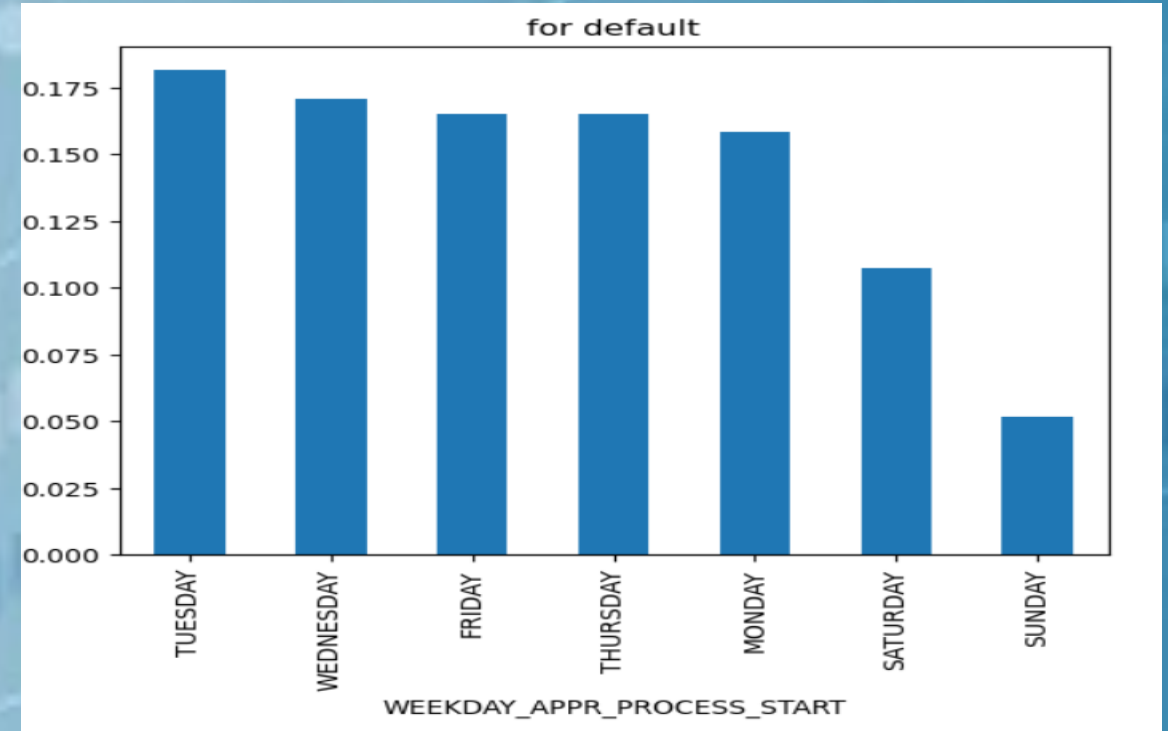
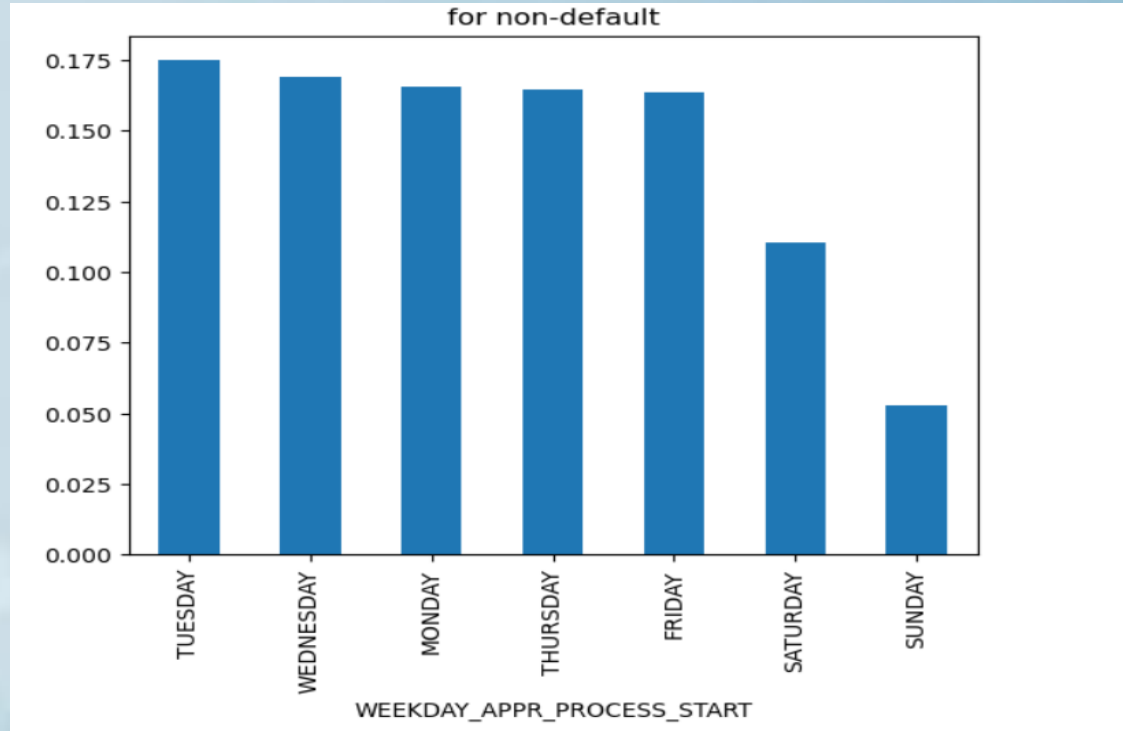
- Insight- In Application_df There Exists 91.927118% Of "Not Default" And 8.072882% Of "Default" Customers. So, We Can Say This Is Not A Balanced Data Set



The background of the slide features a light blue gradient with faint, stylized financial data visualizations. These include several line graphs with circular markers at data points and a bar chart with numerous vertical bars of varying heights. The overall aesthetic is clean and professional, typical of a business or academic presentation.

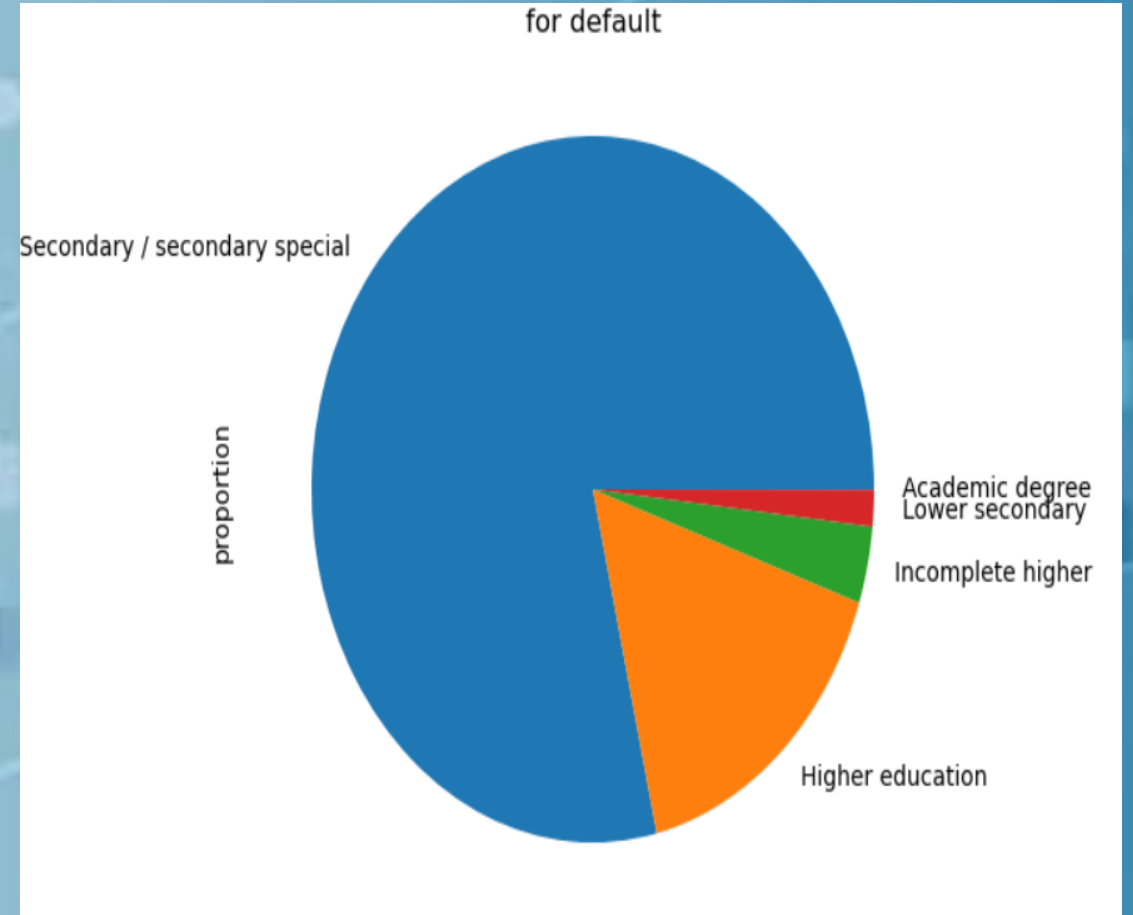
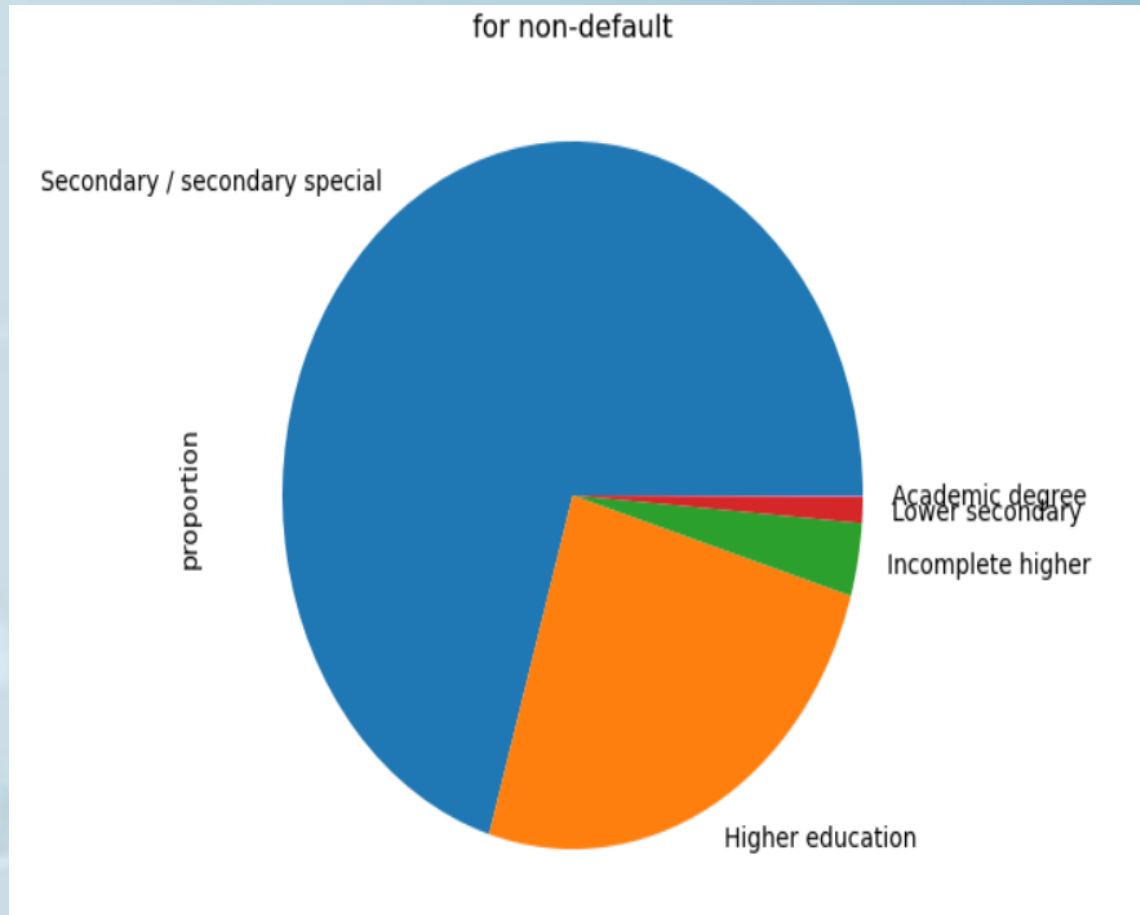
Univariate Analysis/Segmented Univariate Analysis

Univariate Analysis of WEEKDAY_APPR_PROCESS_START



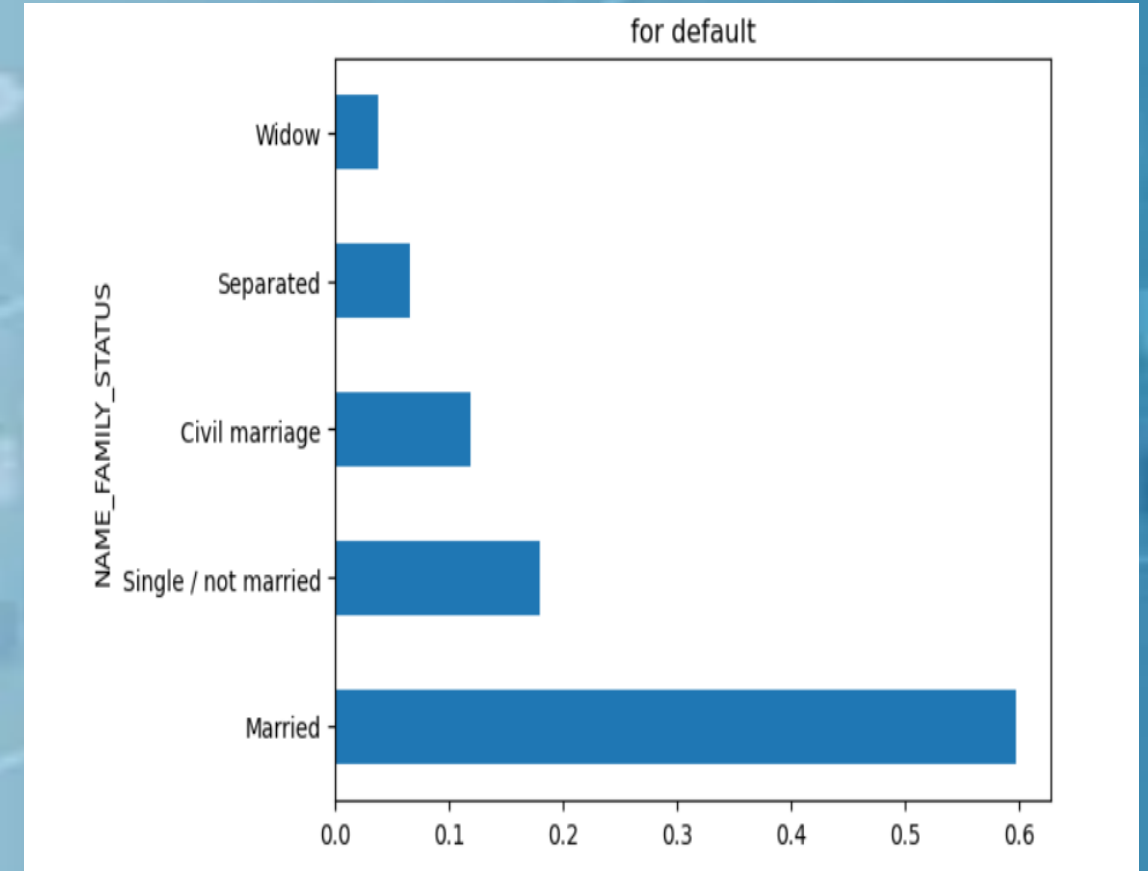
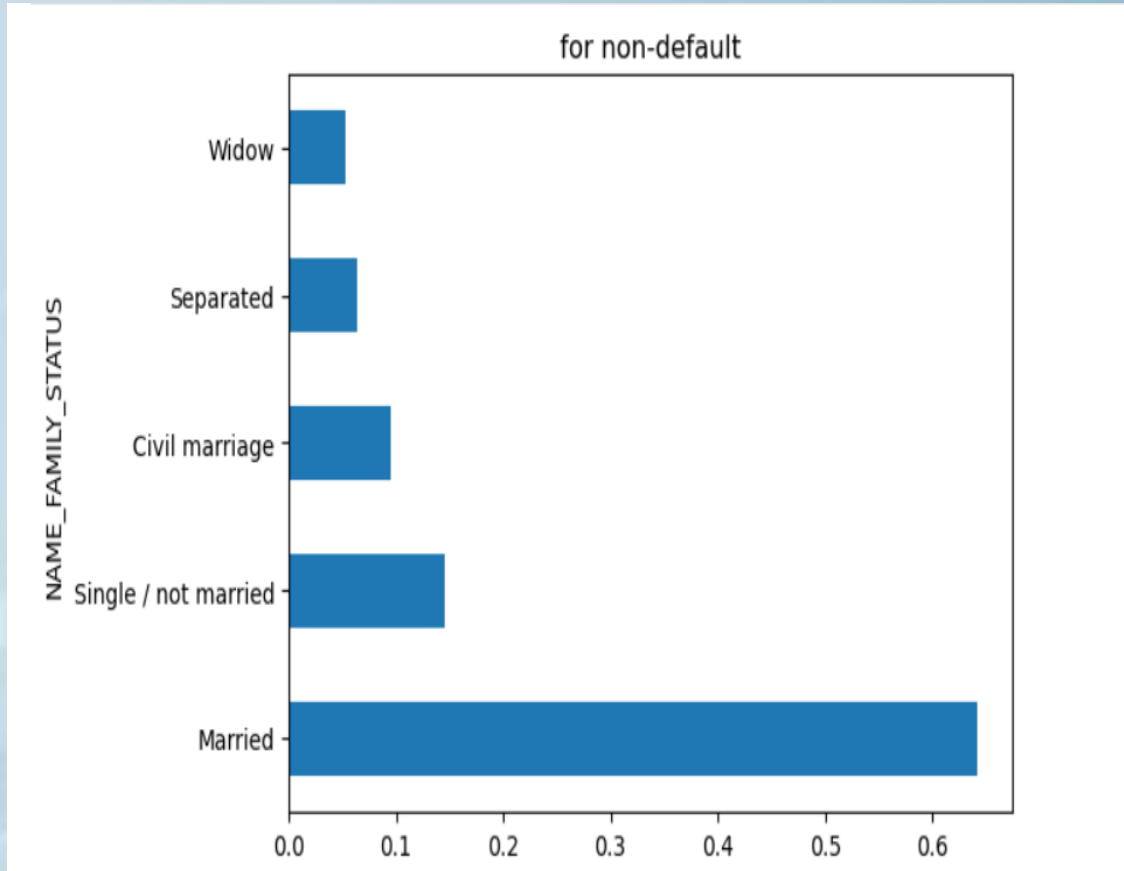
- Insight- From The Above Graph It Can Be Conclude That Application Starting Processes Are Less In Saturday & Sunday for both target-0 & target-1.

Univariate Analysis Of NAME_EDUCATION_TYPE



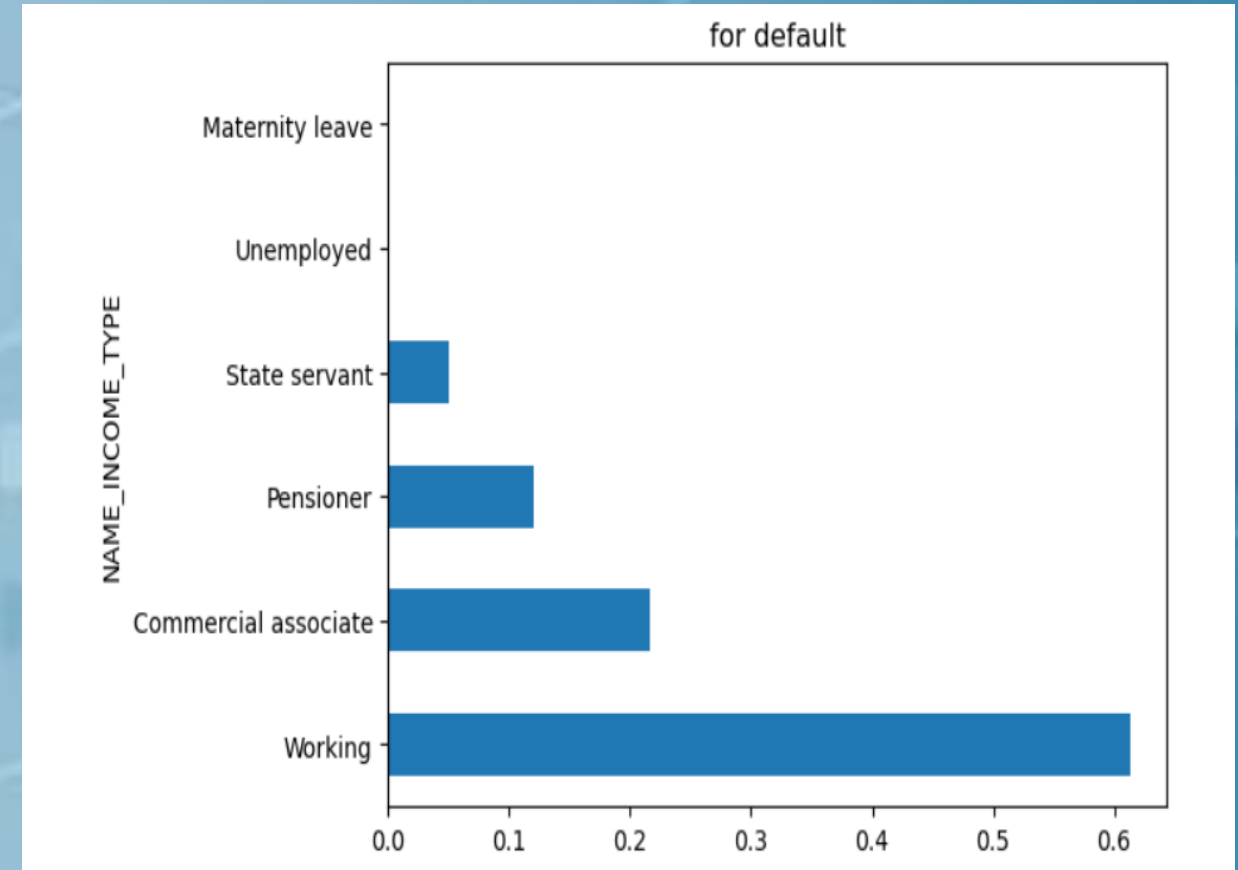
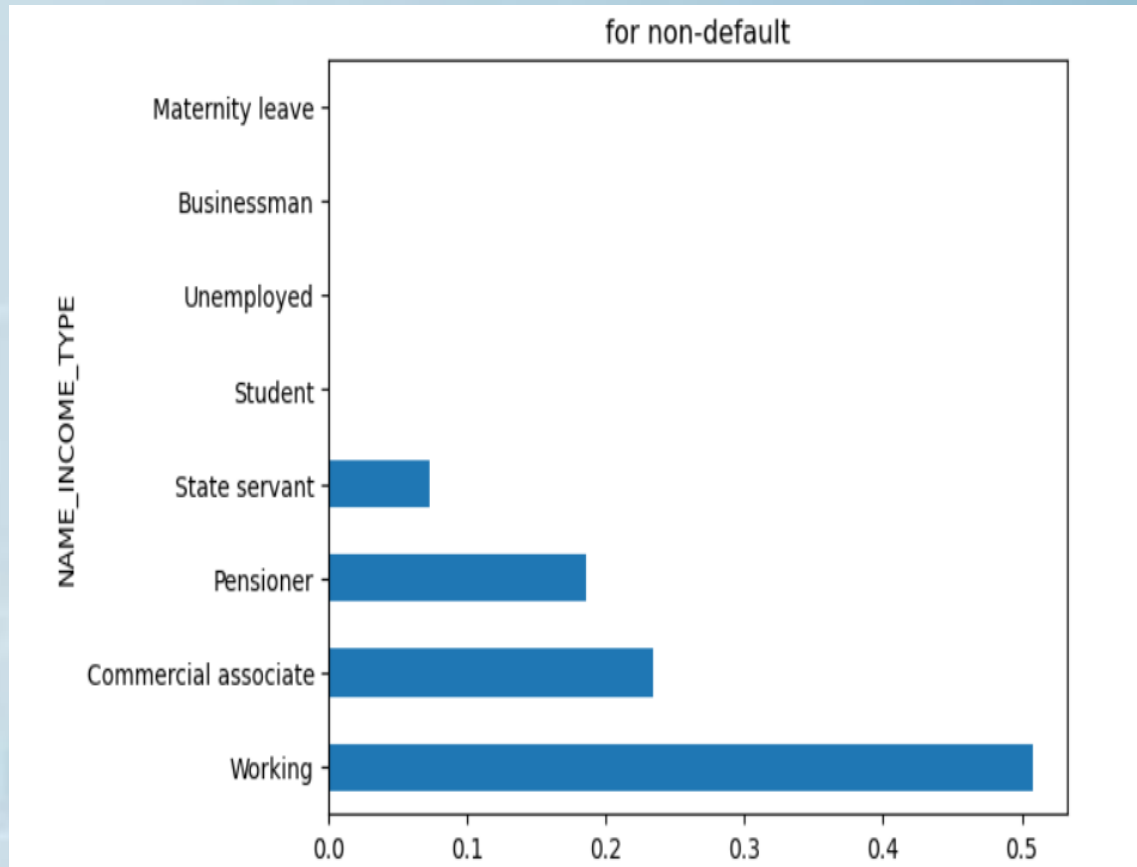
- Insight- From The Plot Above, It Can Be Concluded That Secondary/Special Educated People Are More Applying For Loans .And Academic Degree Educated People Are Less Applying For Loan .For Both Target= 0 And 1

Univariate Analysis Of NAME_FAMILY_STATUS



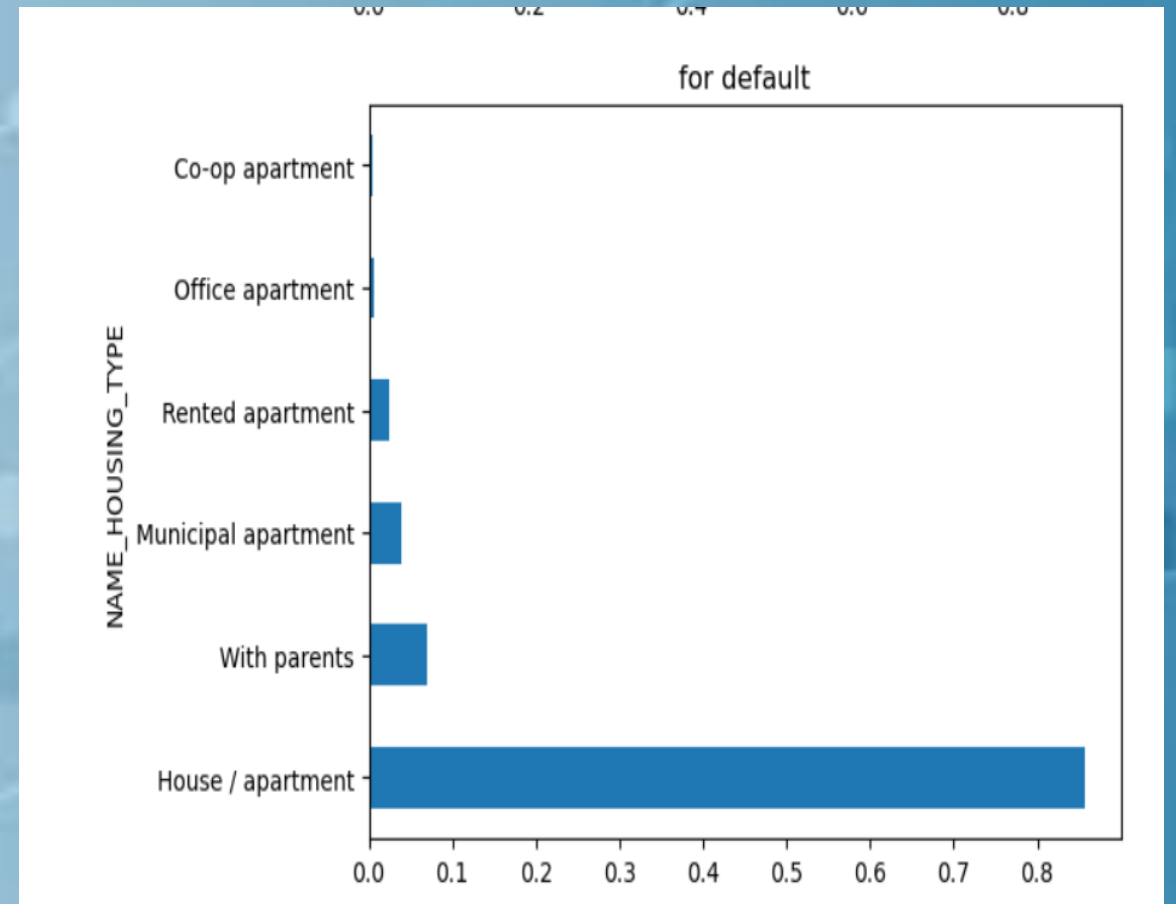
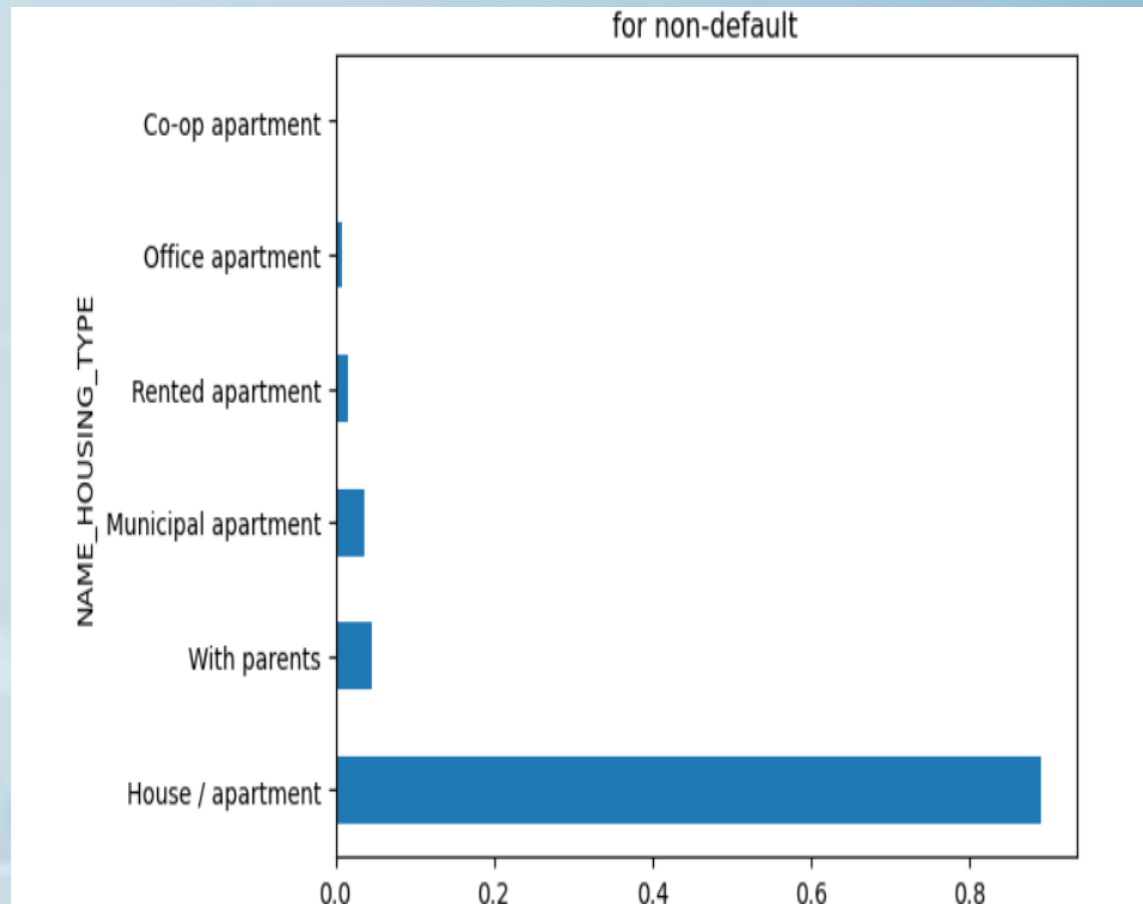
- Insight- The Order Of Both The Cases (Defaulters And Non- Defaulters) Is Same .
- It Can Be Said That Married People Take More Loan As Compared To The Other Categories. It Can Be Said That Being Married Is Not Impacting Default And Not Defaulting People.

Univariate Analysis Of NAME_INCOME_TYPE



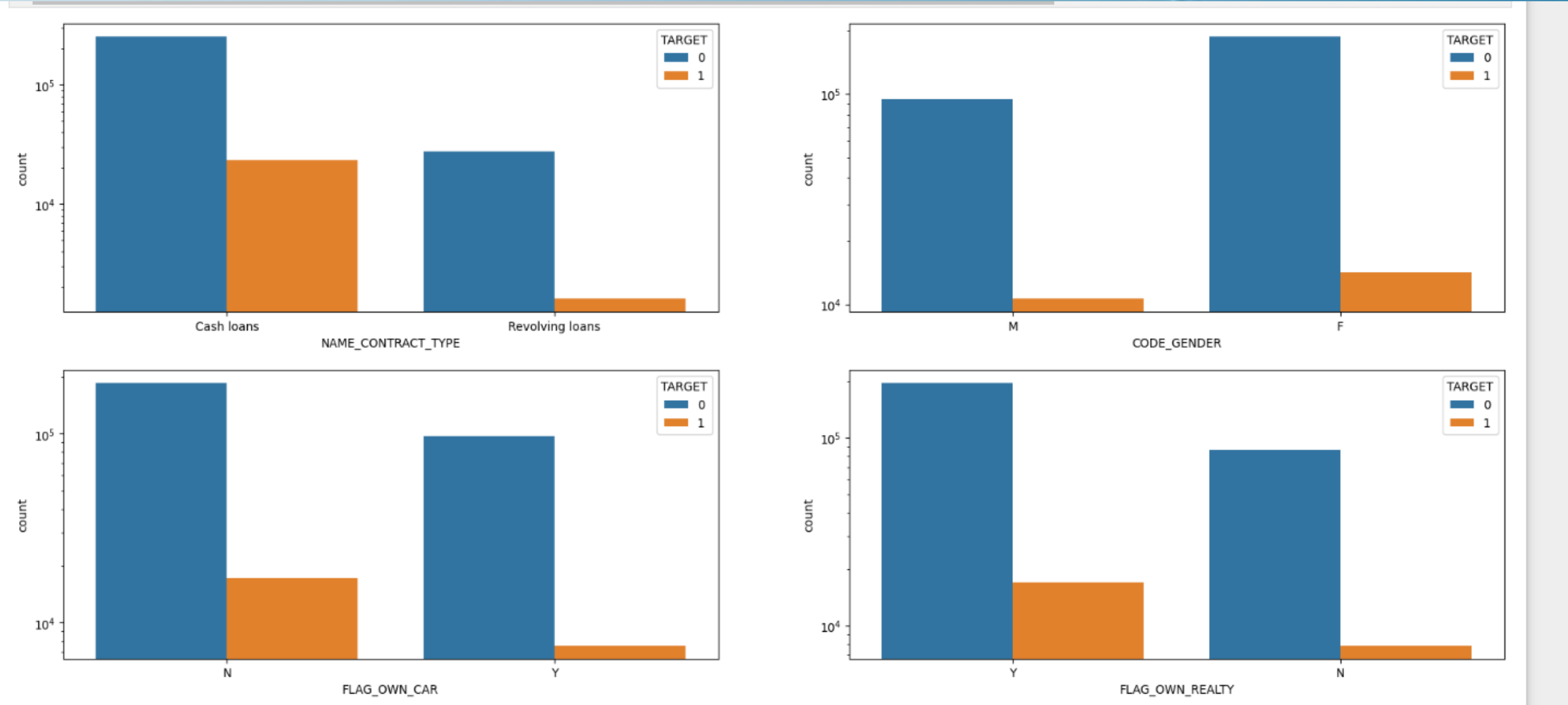
- Insight- From The Graphs Above , It Can Be Concluded That Pensioner Of Not Default Cases Are More Compared To Pensioner Of Default Cases. There Is Both, Loss And Profit Due To Pensioners & commercial associate to the Bank. It Can Also Be Seen That Most Of Defaulter's Income Type Is Working. & At The Same Time From Working People There Is Good Income To Bank.

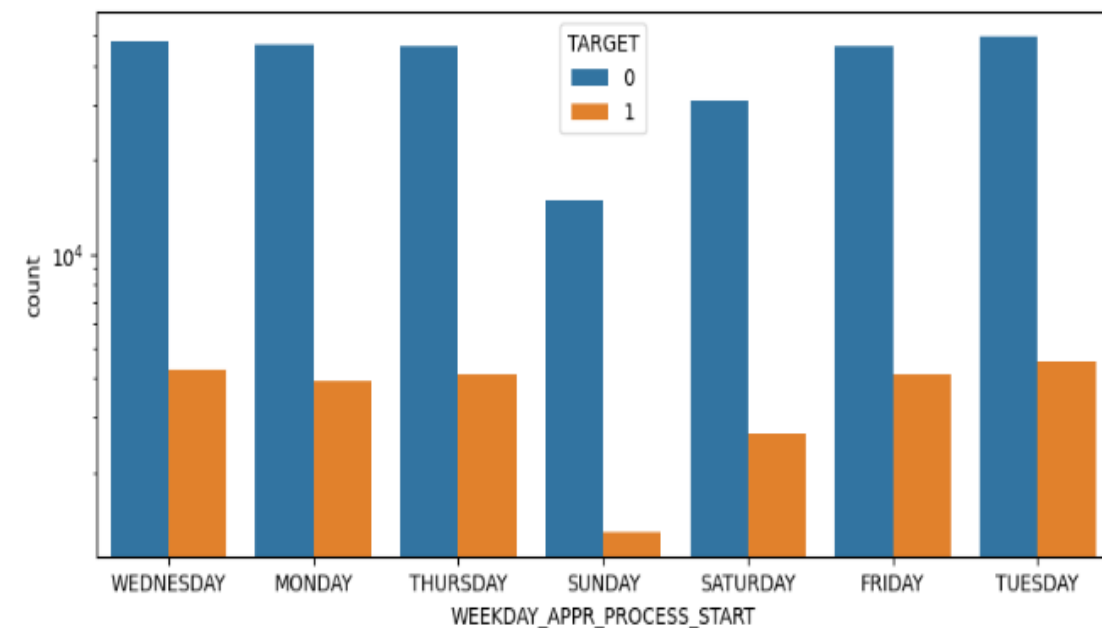
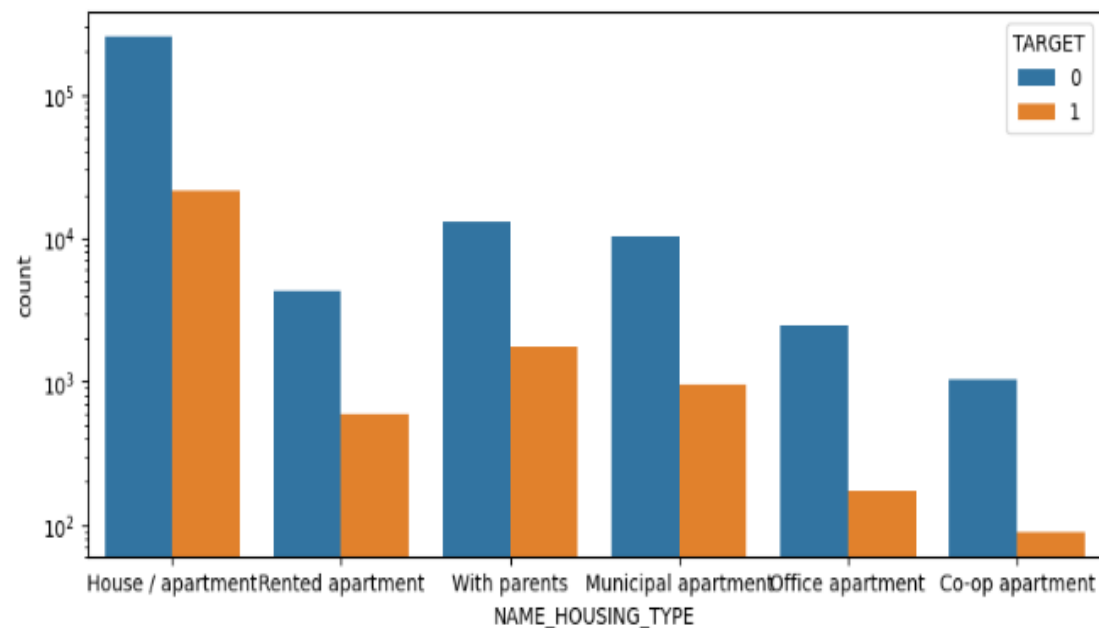
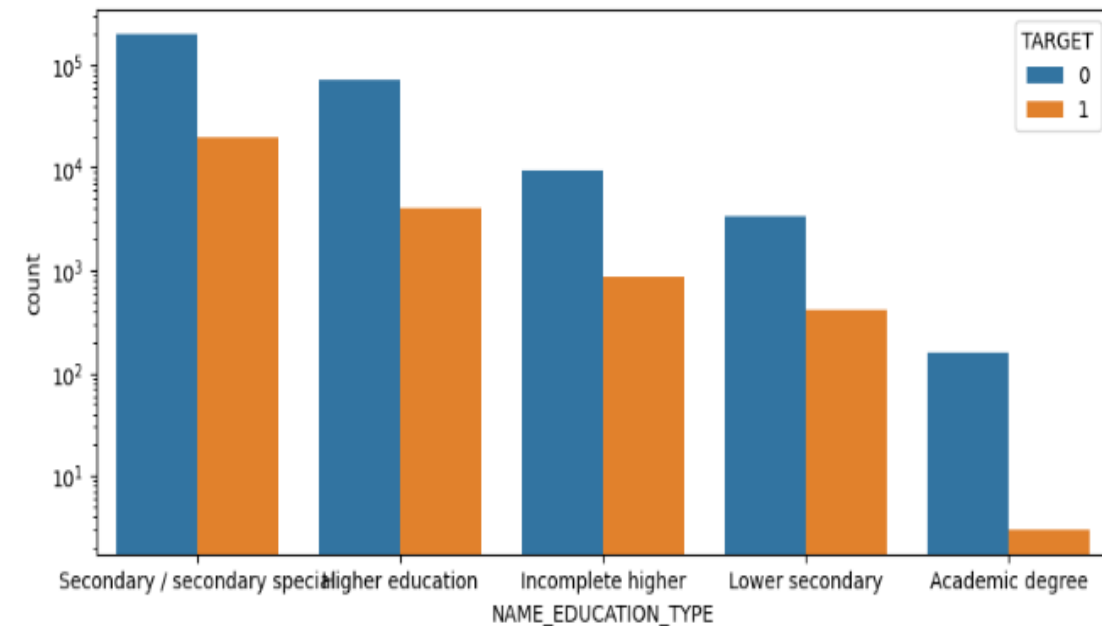
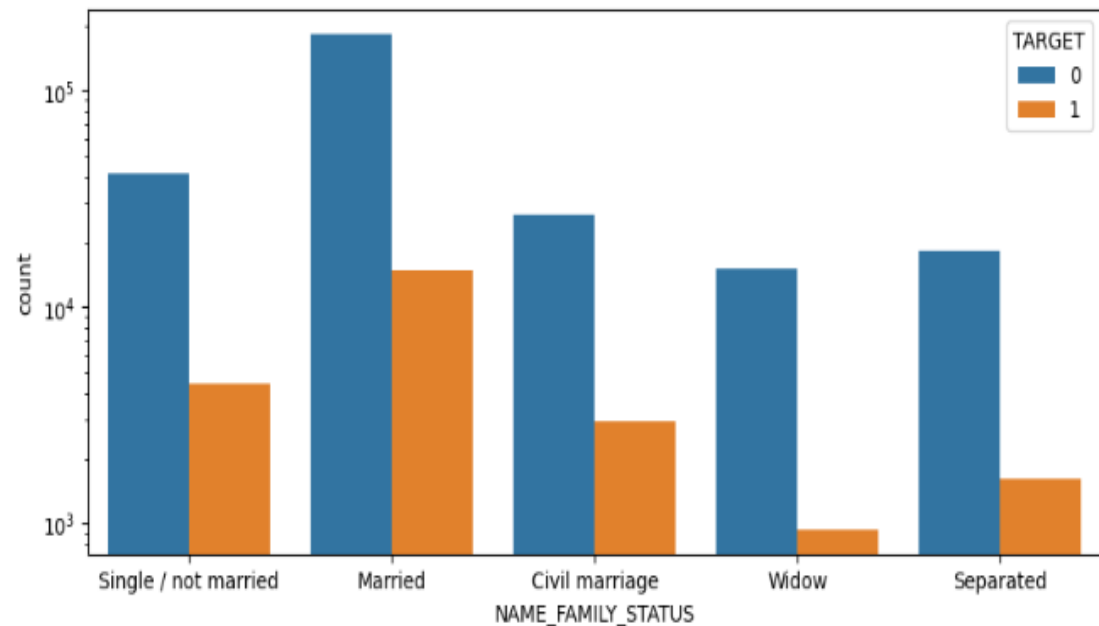
Segmented Univariate Analysis Of NAME_HOUSING_TYPE

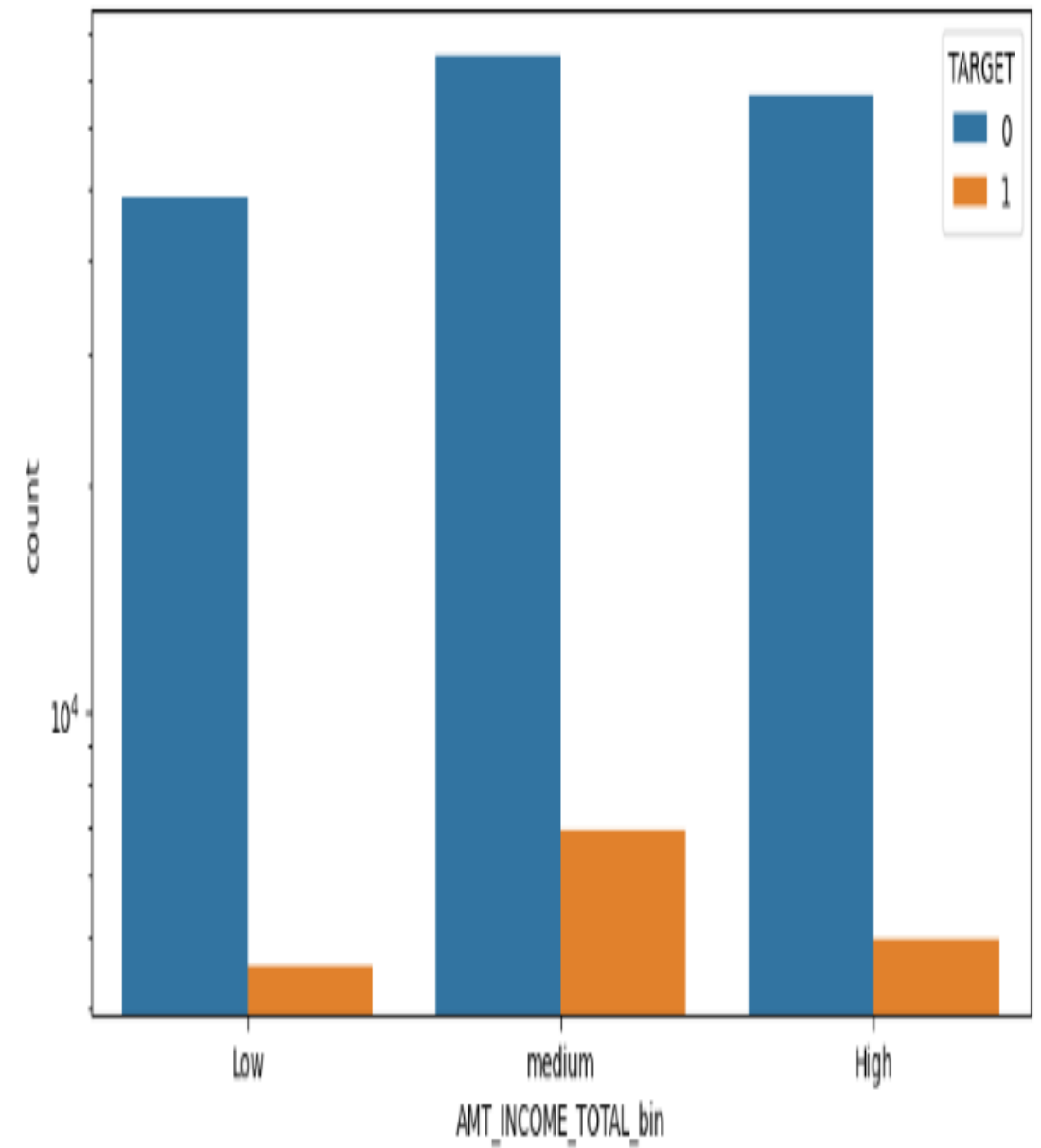
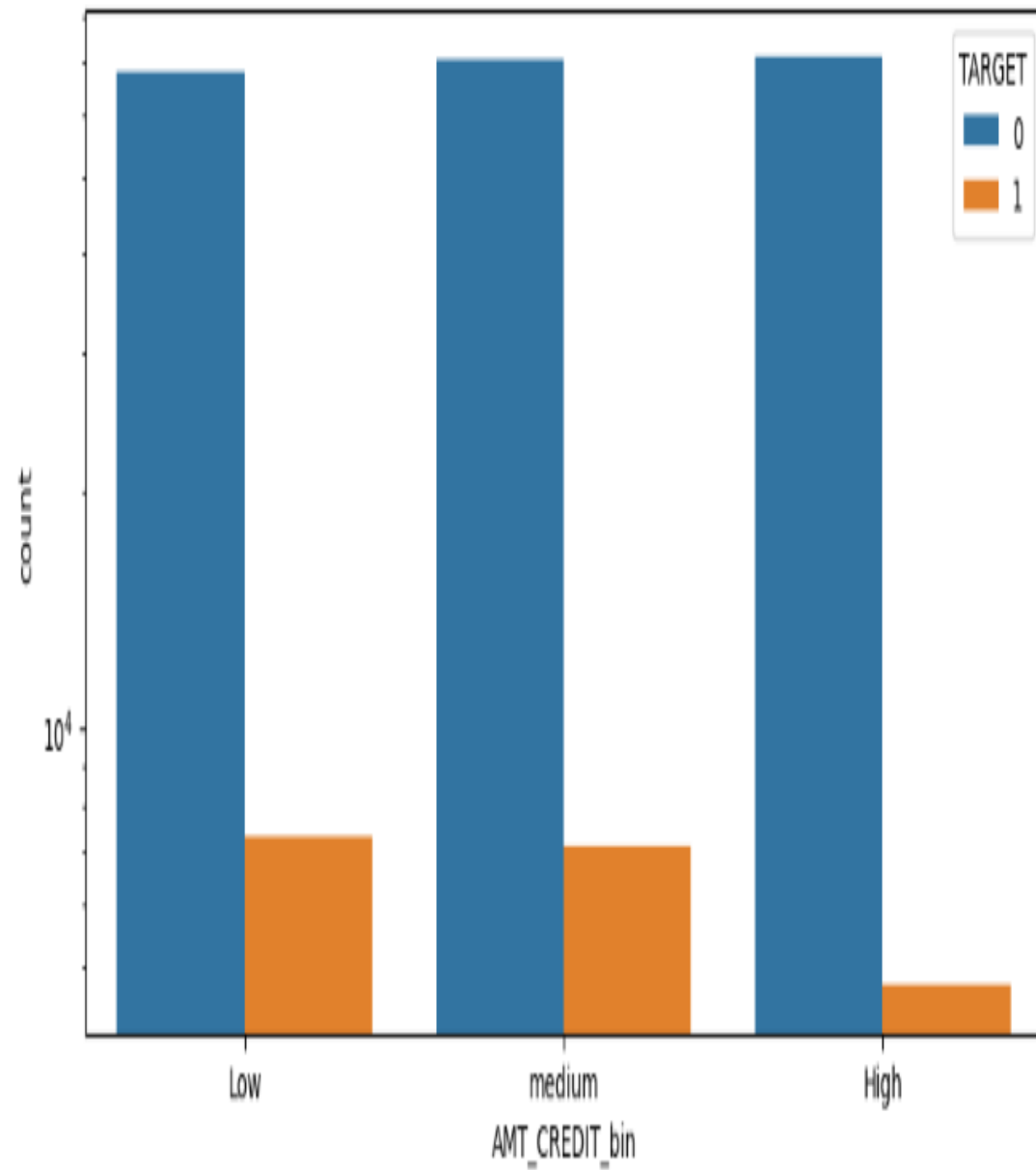


- Insight- From Graph It Can Be Concluded That There Exists People Who Have Own House Lies In Both Defaulters And Non -Defaulters.

Target Variable Across The Categories Of Categorical Variables Against Target 0 And 1





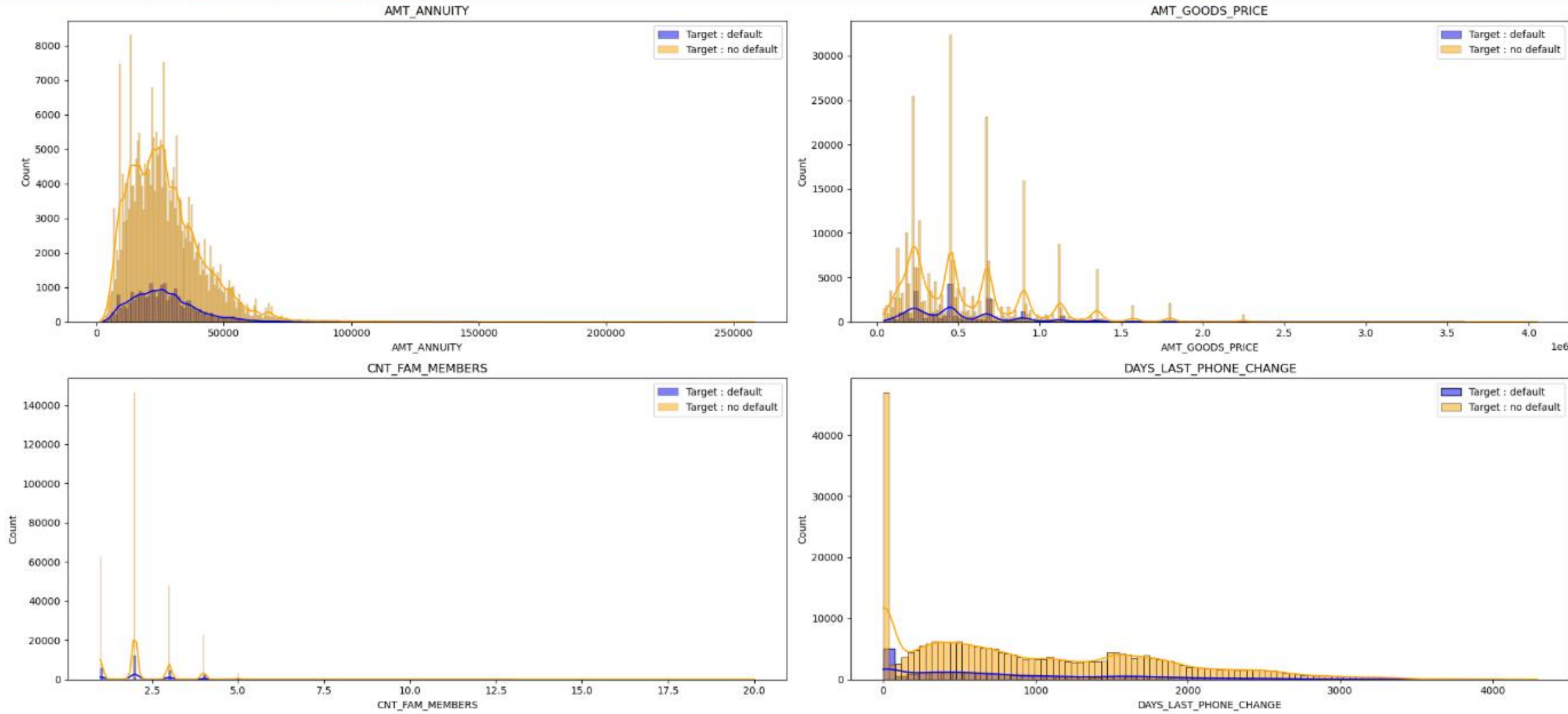


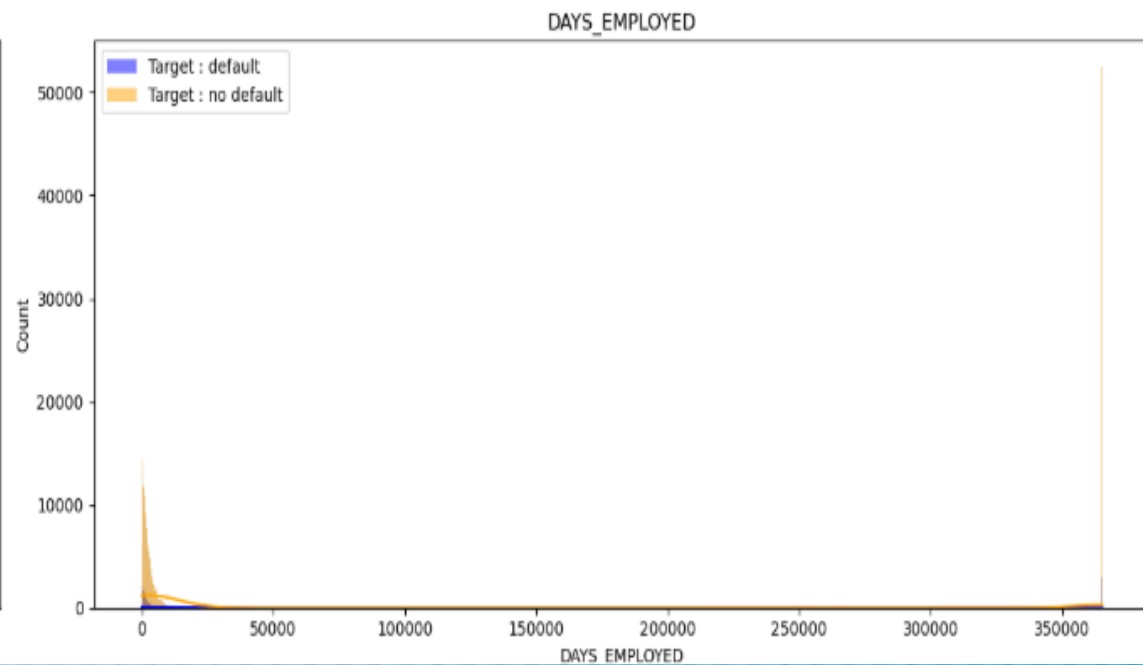
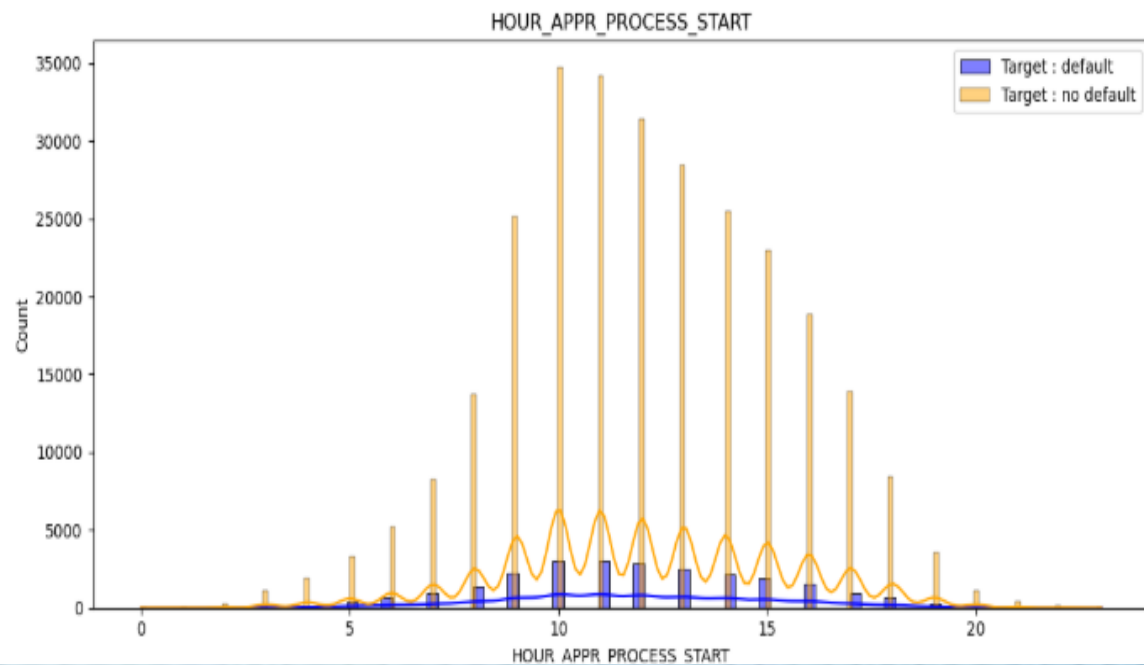
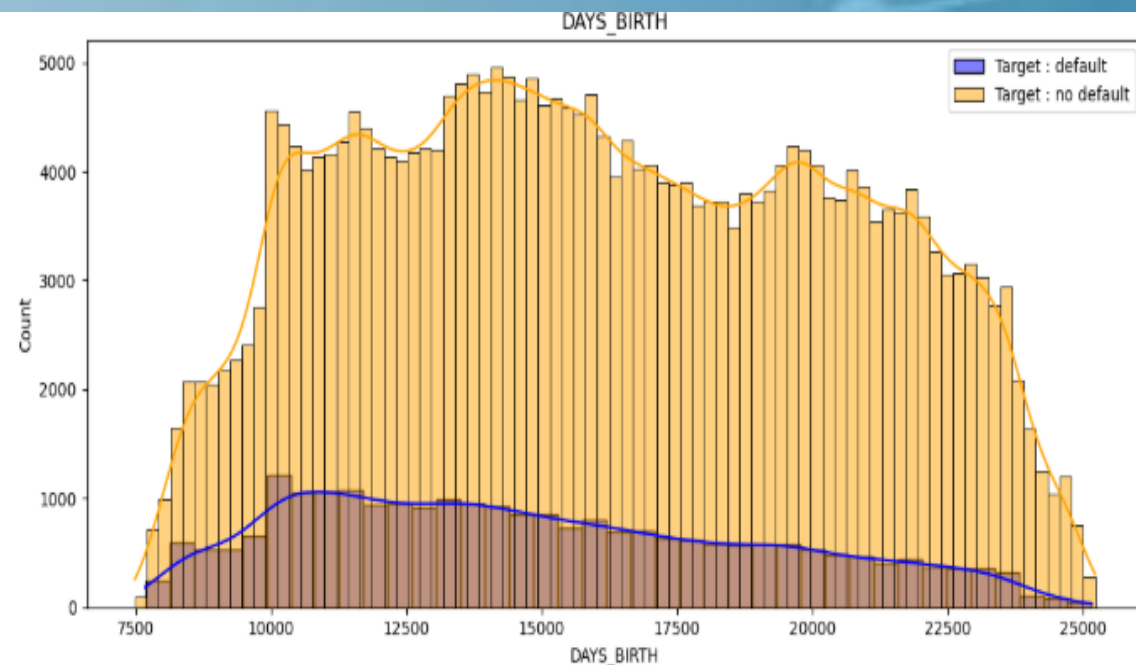
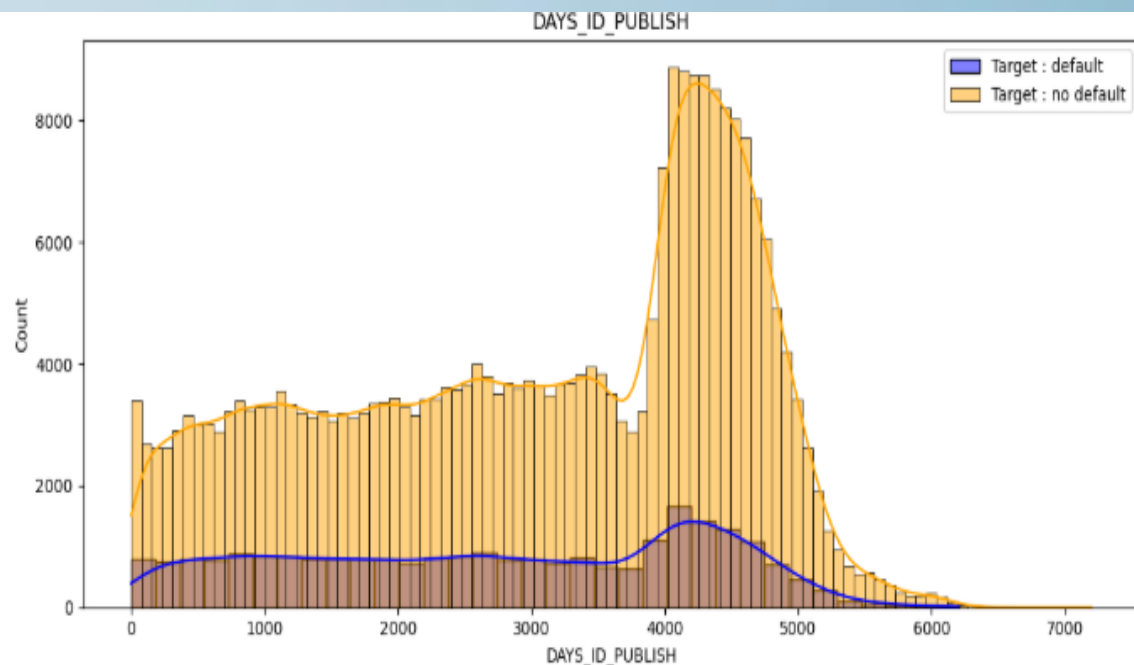
Insights

As We Can See From Above Graphs:-

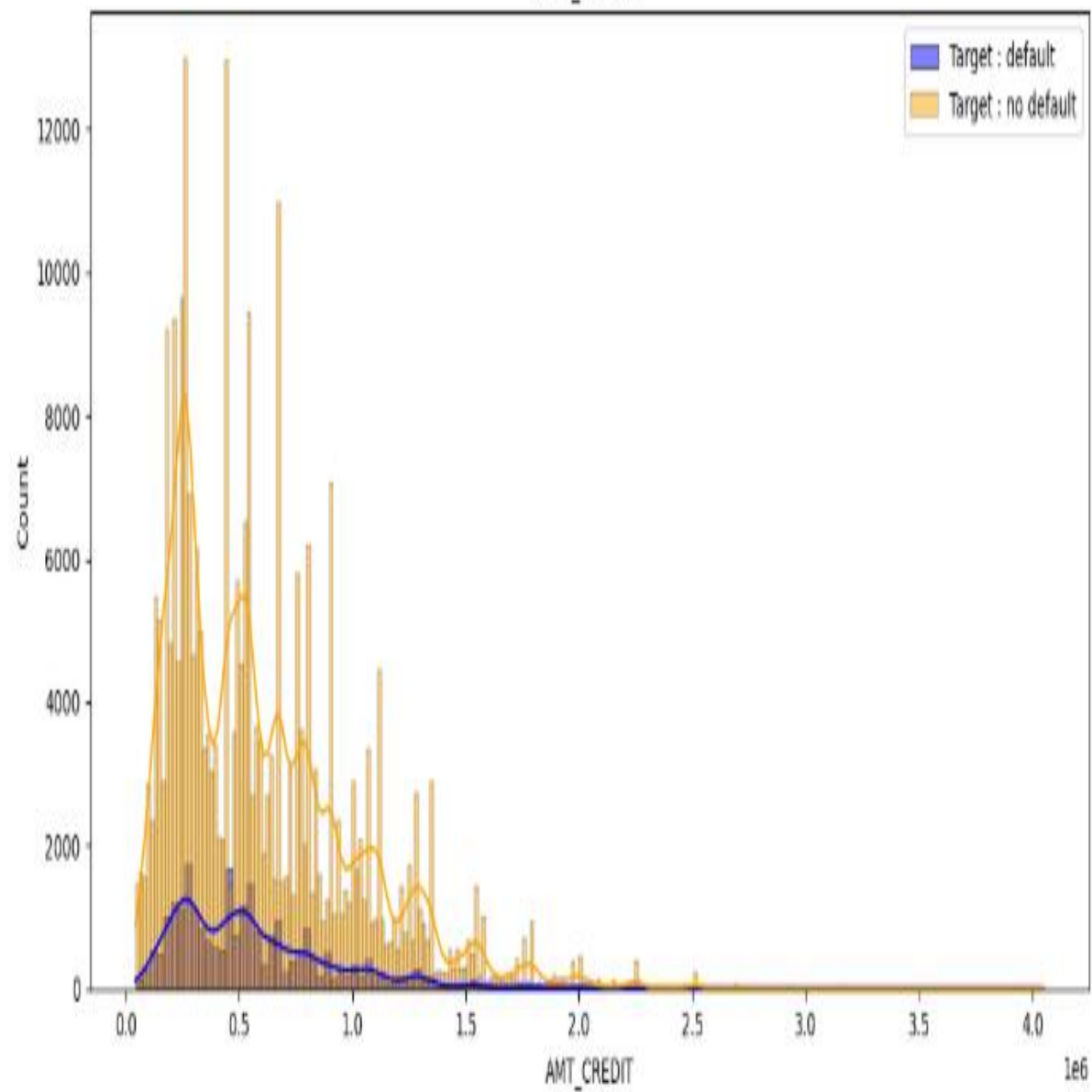
- People With Medium Total Income Default More.
- People With High Credit Amounts Are Less Defaulters.
- People Who Started Application Process On Sunday Are Less Defaulters.
- For Bank In Terms Of Loan Applications Saturday And Sunday Are Less Busy.
- People With House Or Apartment Take More Loans Than Others.
- Married People Take More Loan As Compared To Other Categories
- We Can Conclude That Secondary/Special Educated People More Applying For Loans.
- People With Real Estate Take More Loans.
- People Who Don't Own A Car Take More Loans Than Others.
- Female Customer Take More Loans.
- People Take More Cash Loans.

Target Variable Across The Numerical Variables Against Target 0 And 1

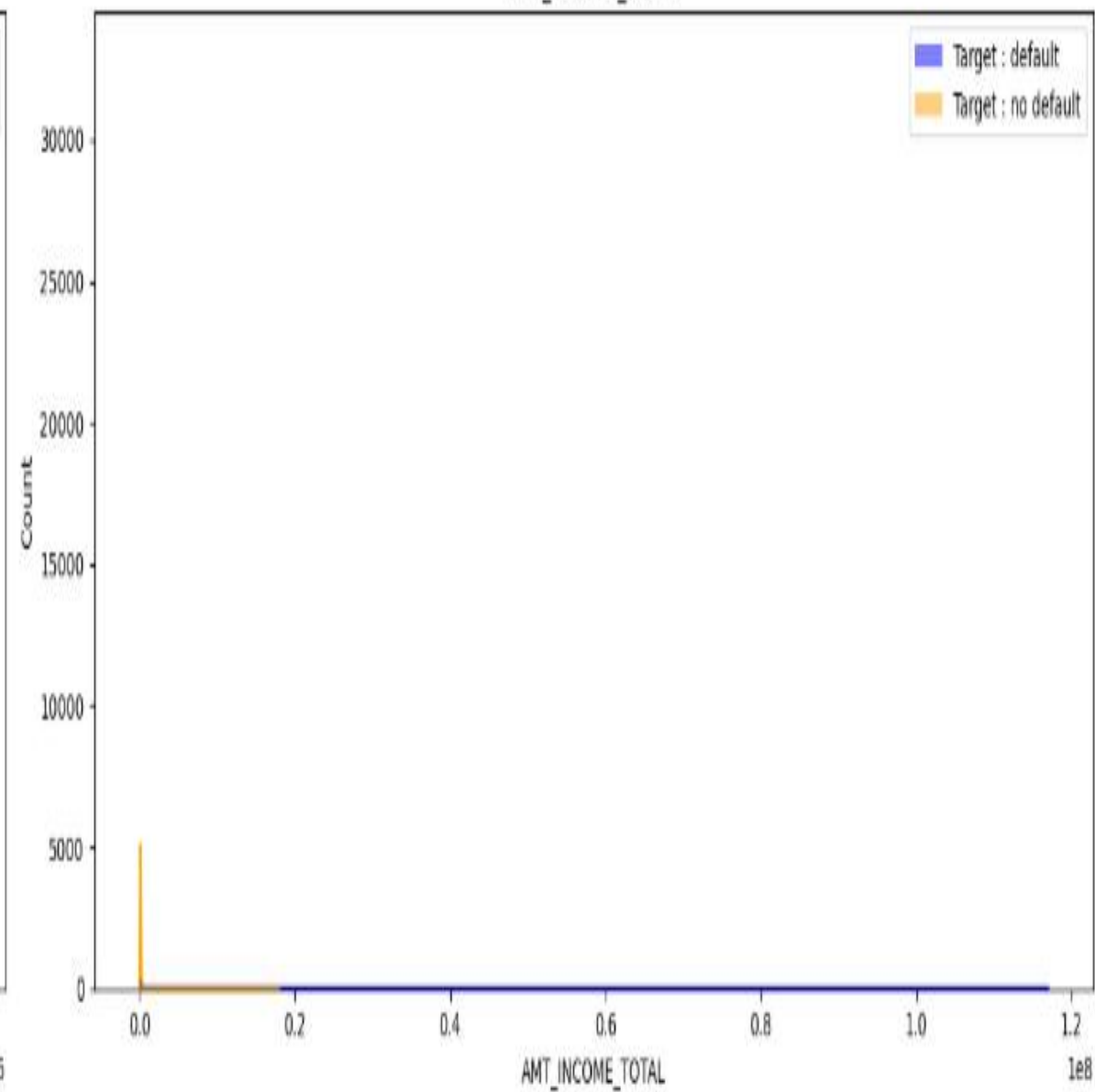




AMT_CREDIT



AMT_INCOME_TOTAL



Insights

As We Can See From Graphs:-

- People With Lower Total Income Are More Defaulters.
- People Who Are Just Employed Take More Loans.
- Retired People Tends To Take More Loans.
- During 10 Am To 2 Pm More Applications Are Filled.
- People Whose Age Are Between 27yrs(10000-days) And 41(15000-days) Yrs Take More Loans.
- People Whose Id(s) Published Between 4000 Days And 5000 Days Ago ,Take More Loans.
- More Loans Taken By Nuclear Family.
- Low A People Take Loans For Less Goods Amount
- Mount Annuity Has Higher Count Of Loans

Correlation

For Default

Case ID	Var1	Var2	Coorelation
814	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
642	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	1.00
676	FLOORSMAX_MEDI	FLOORSMAX_AVG	1.00
610	FLOORSMAX_MODE	FLOORSMAX_AVG	0.99
678	FLOORSMAX_MEDI	FLOORSMAX_MODE	0.99
168	AMT_GOODS_PRICE	AMT_CREDIT	0.98
644	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.98
576	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.98
364	CNT_FAM_MEMBERS	CNT_CHILDREN	0.89
848	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87

For Non-default

Case ID	Var1	Var2	Coorelation
676	FLOORSMAX_MEDI	FLOORSMAX_AVG	1.00
814	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
168	AMT_GOODS_PRICE	AMT_CREDIT	0.99
642	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.99
678	FLOORSMAX_MEDI	FLOORSMAX_MODE	0.99
610	FLOORSMAX_MODE	FLOORSMAX_AVG	0.99
576	YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.97
644	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.96
364	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
848	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86

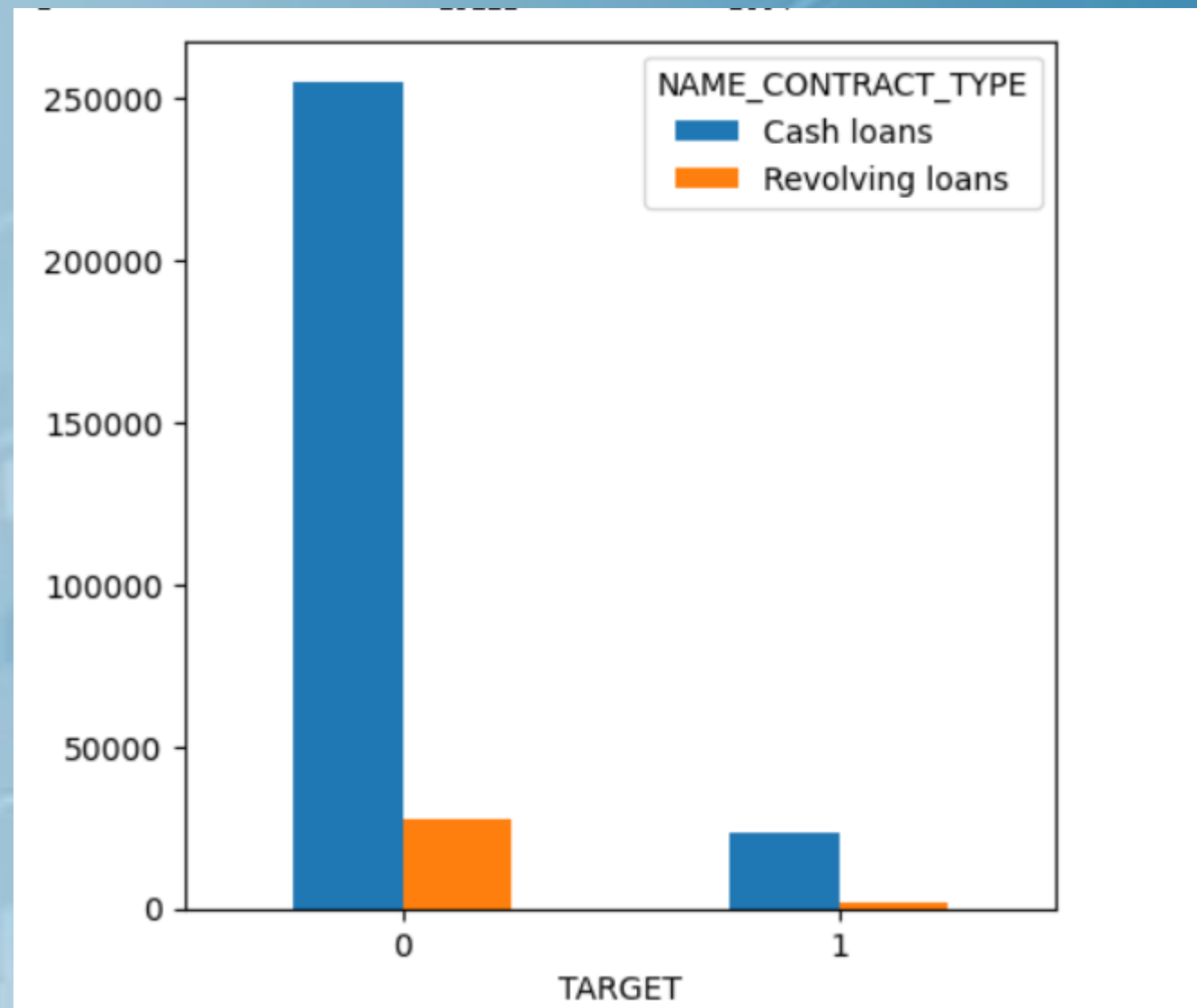
Correlation Observations

- For Default Cases, Major Portion Of Decision Is Taken By- (YEARS_BEGINEXPLUATATION_MEDI AND YEARS_BEGINEXPLUATATION_AVG) (OBS_60_CNT_SOCIAL_CIRCLE AND OBS_30_CNT_SOCIAL_CIRCLE) (FLOORSMAX_MEDI AND FLOORSMAX_AVG)
- For Non-Defaulters, Major Portion Of Decision Is Taken By- (FLOORSMAX_MEDI AND FLOORSMAX_AVG) (OBS_60_CNT_SOCIAL_CIRCLE AND OBS_30_CNT_SOCIAL_CIRCLE)
- The Values - YEARS_BEGINEXPLUATATION_MEDI And YEARS_BEGINEXPLUATATION_AVG Are More Correlated In Case Of Default Than Non-default Cases.

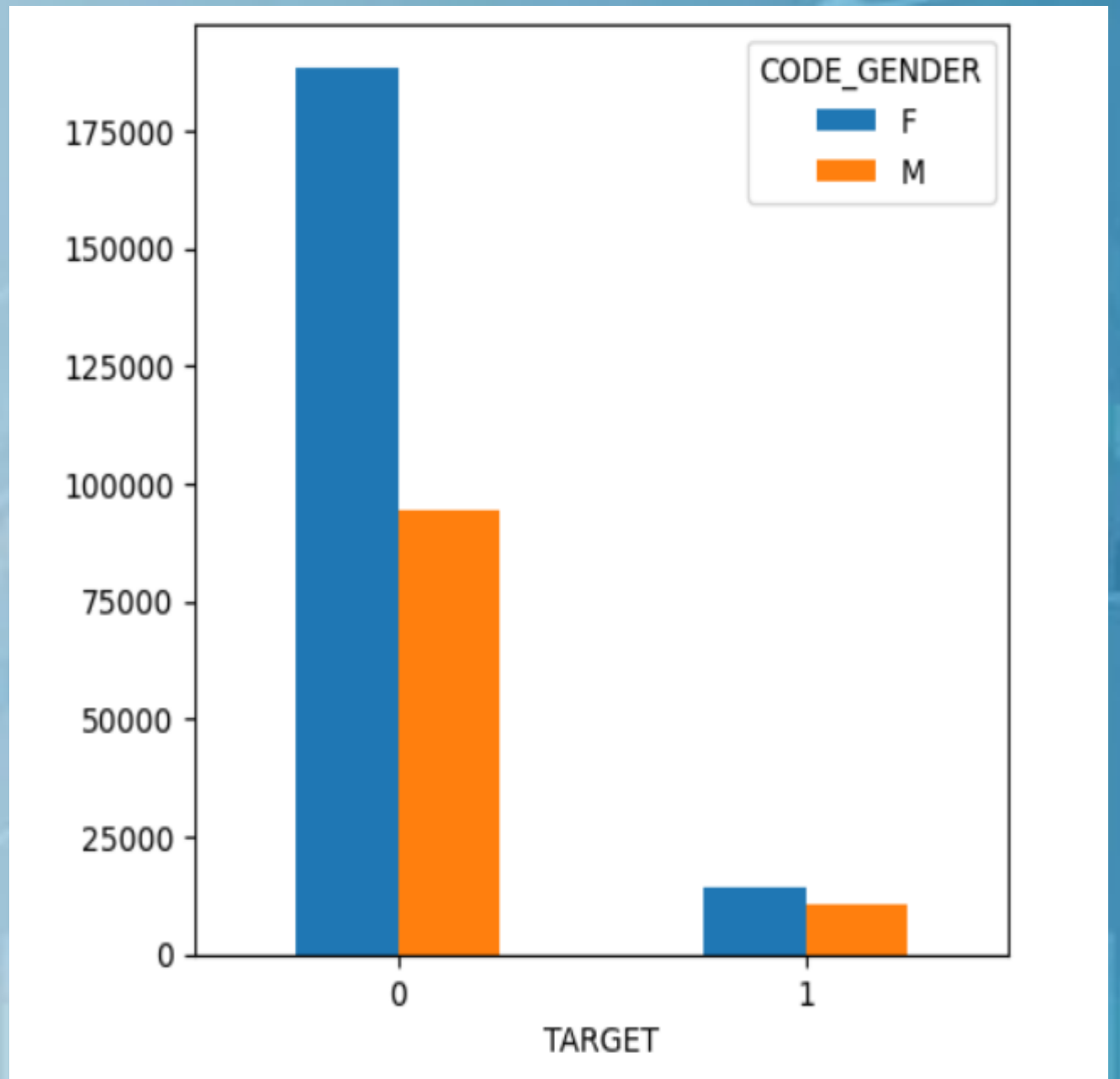
The background of the slide features a light blue gradient with faint, stylized financial data visualizations. These include several line graphs with circular markers at data points, and a series of vertical bars of varying heights, resembling a bar chart or a candlestick chart. The overall aesthetic is clean and professional, typical of a business or academic presentation.

Bi-variate analysis

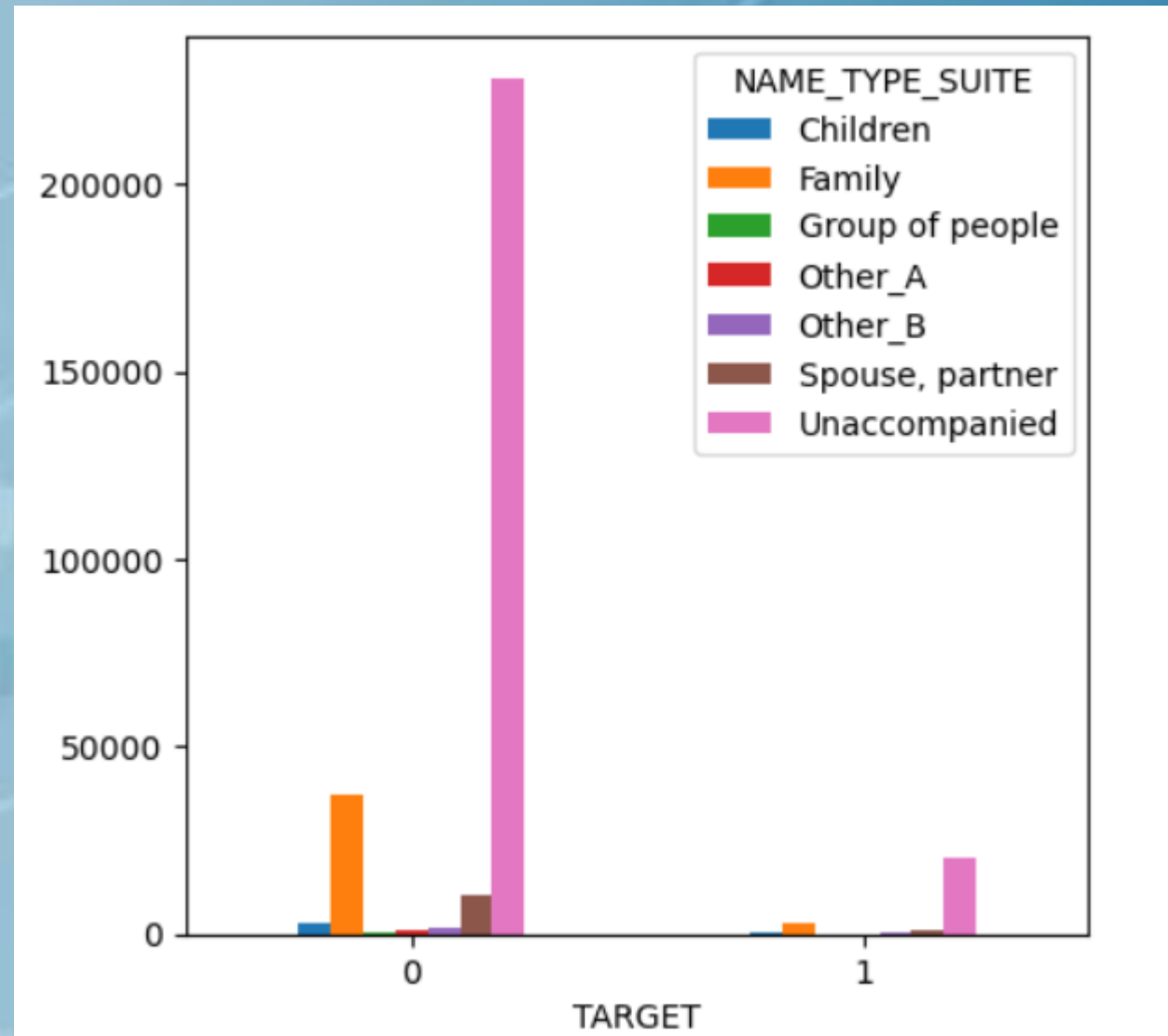
- Insight- High Number Of Cash Loans Than Revolving Loans



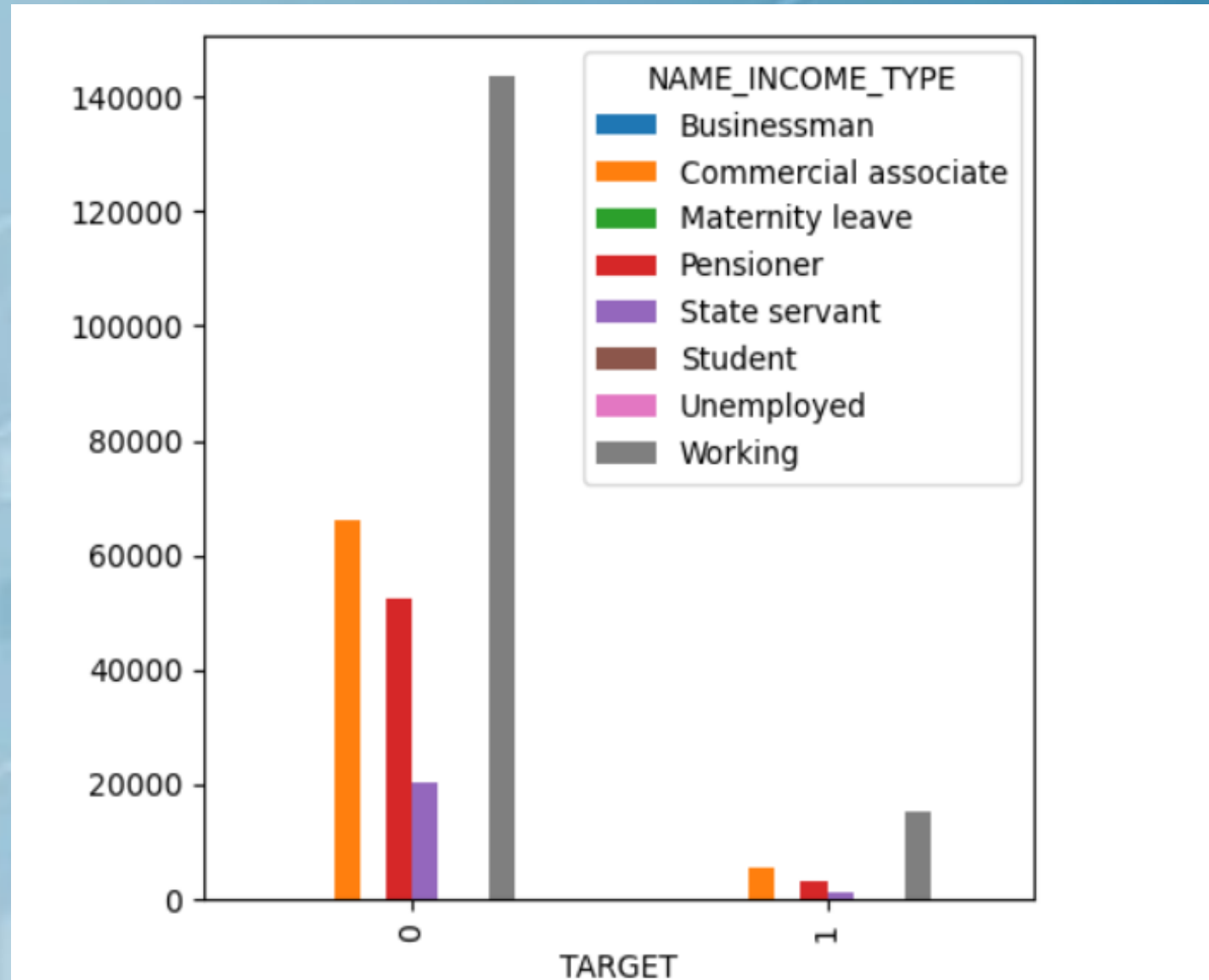
➤ Insight- Females Take More Loans Than Others



- Insight- Most Of The People Come Alone For Applying Loan

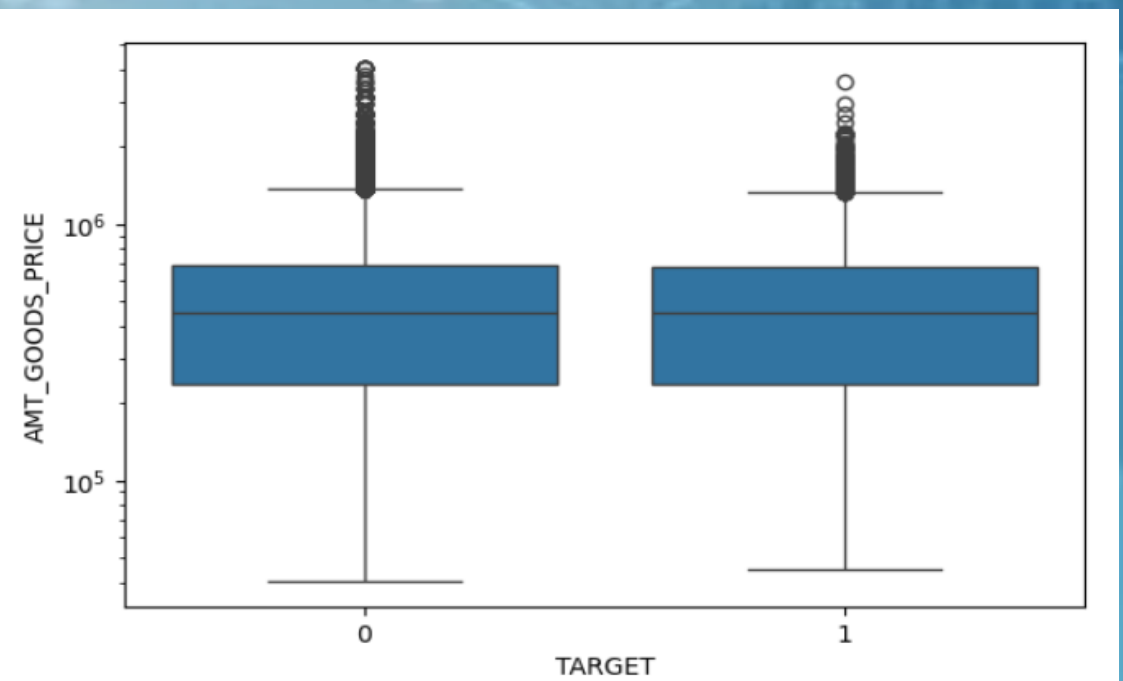
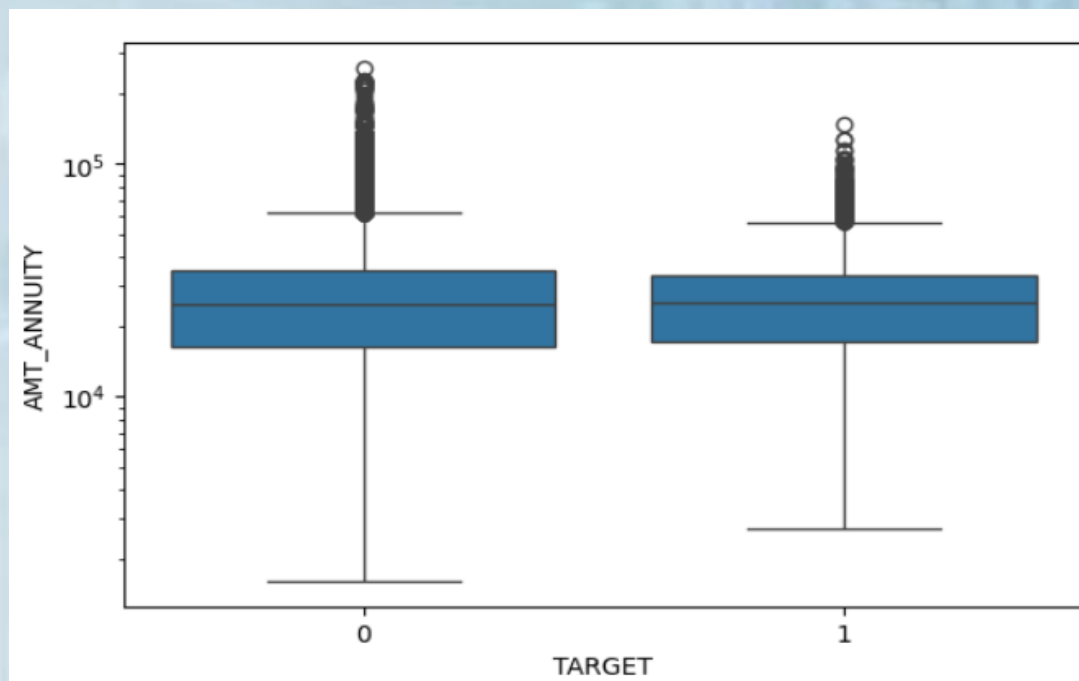
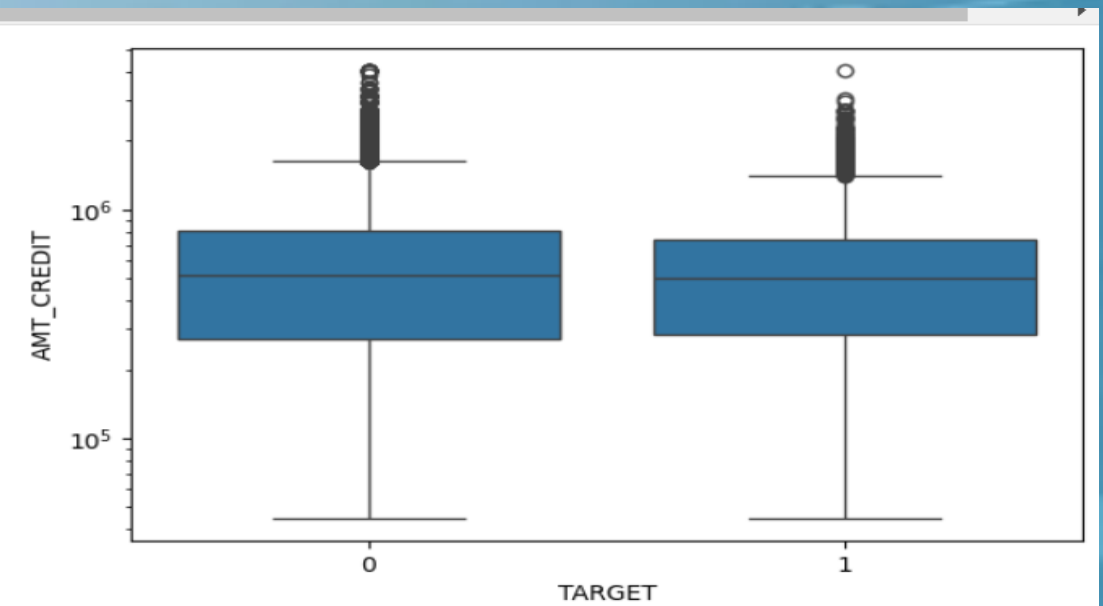
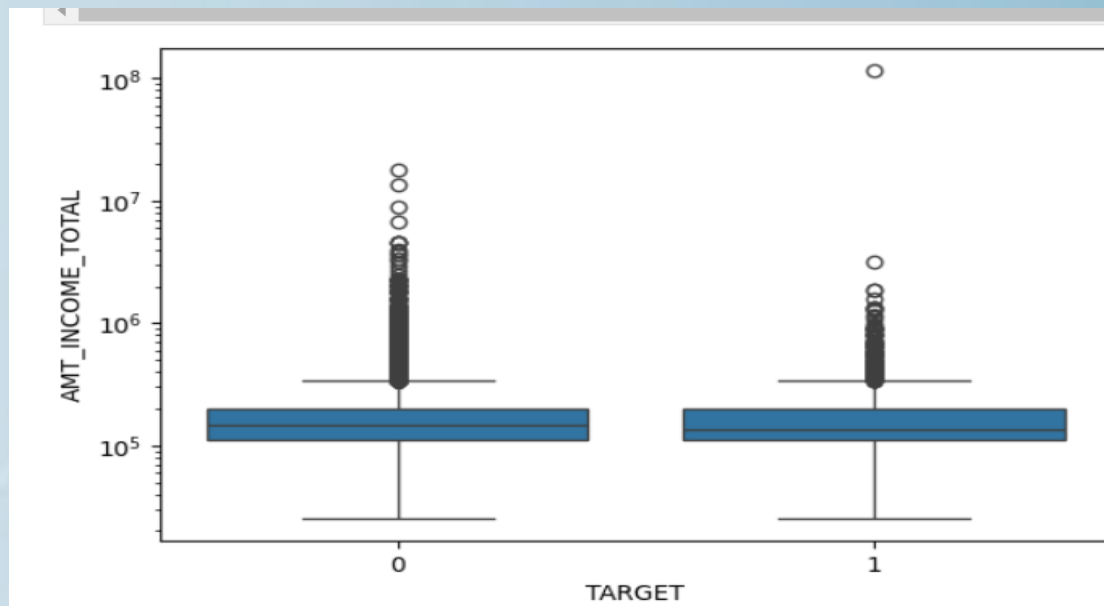


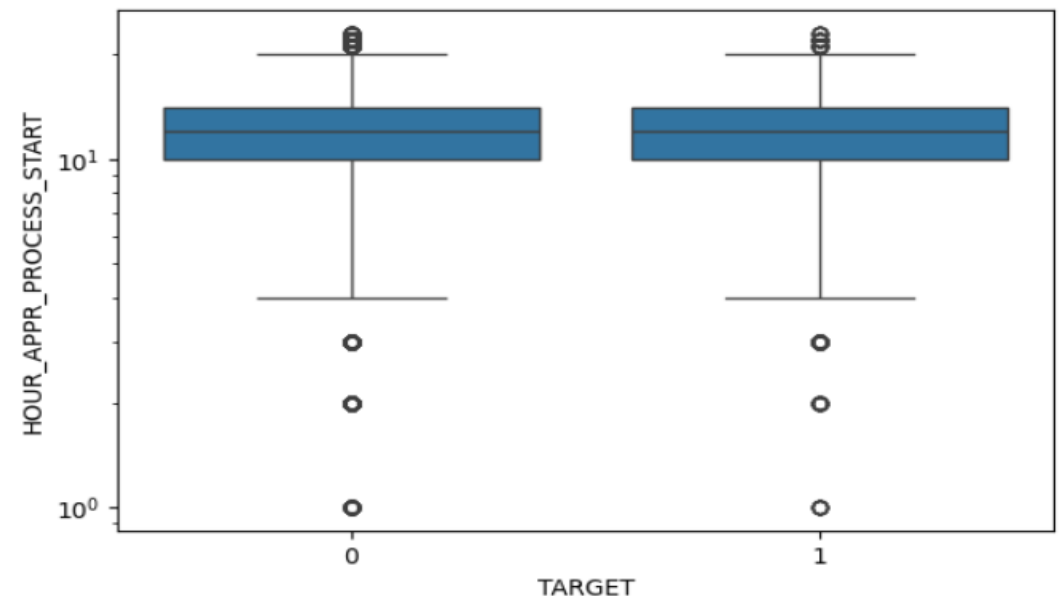
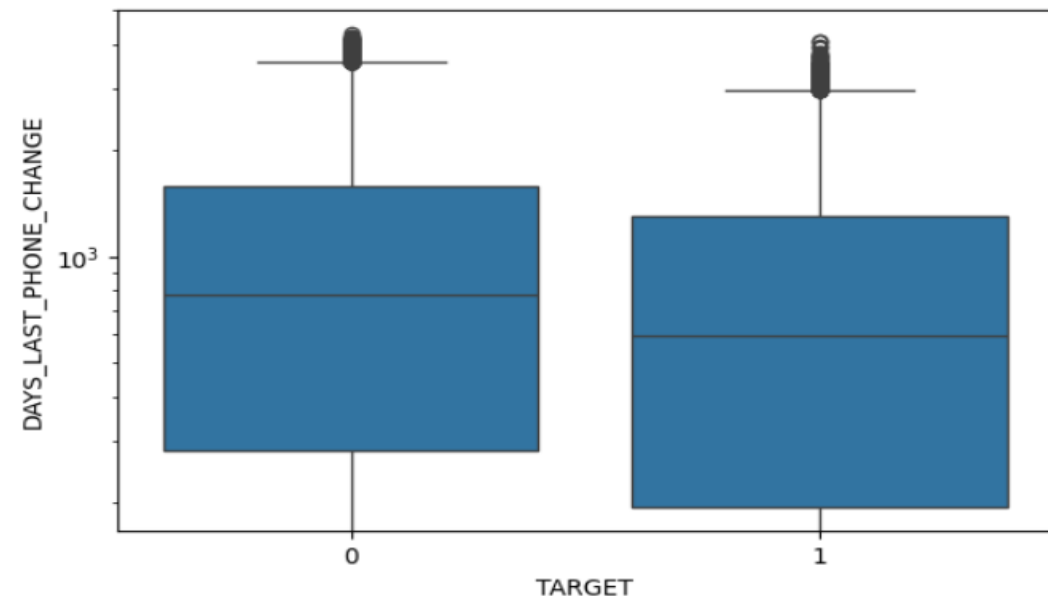
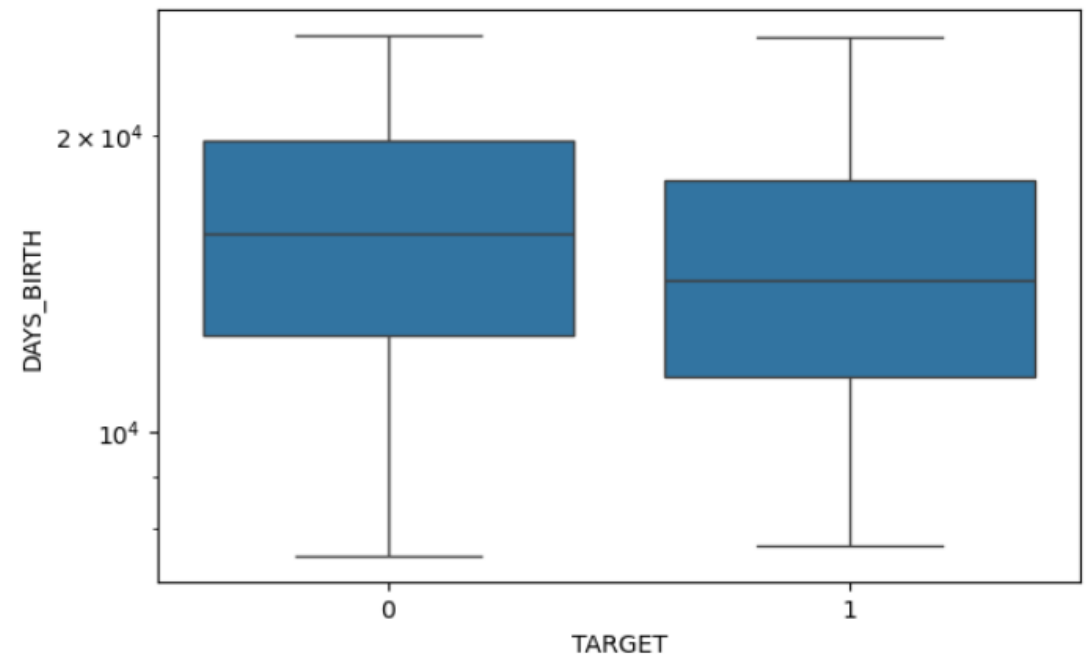
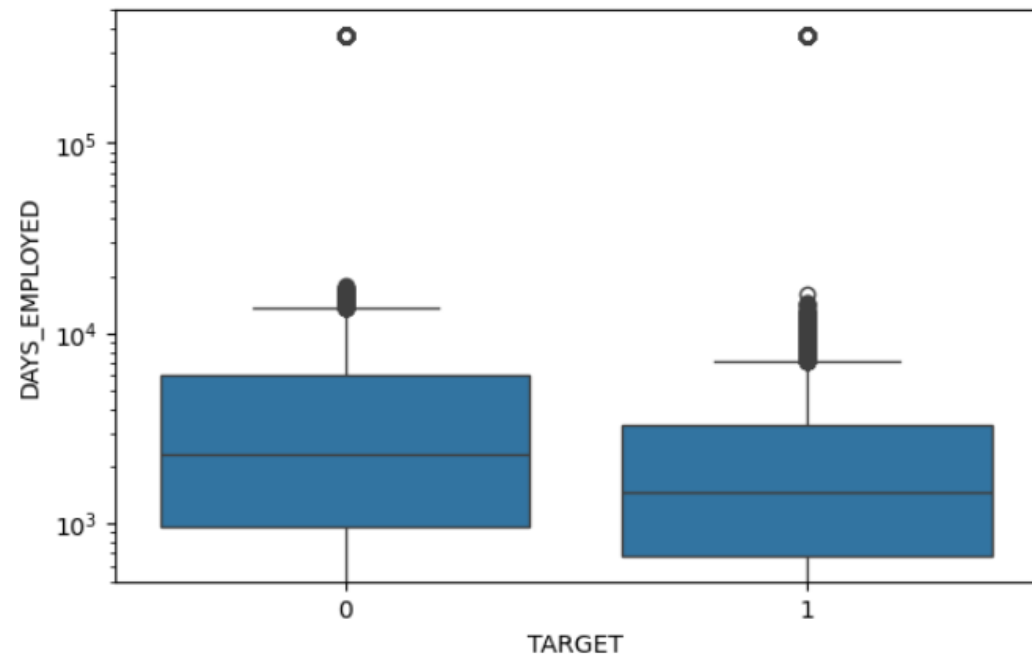
➤ Insight-working People Take More Loans Than Others.

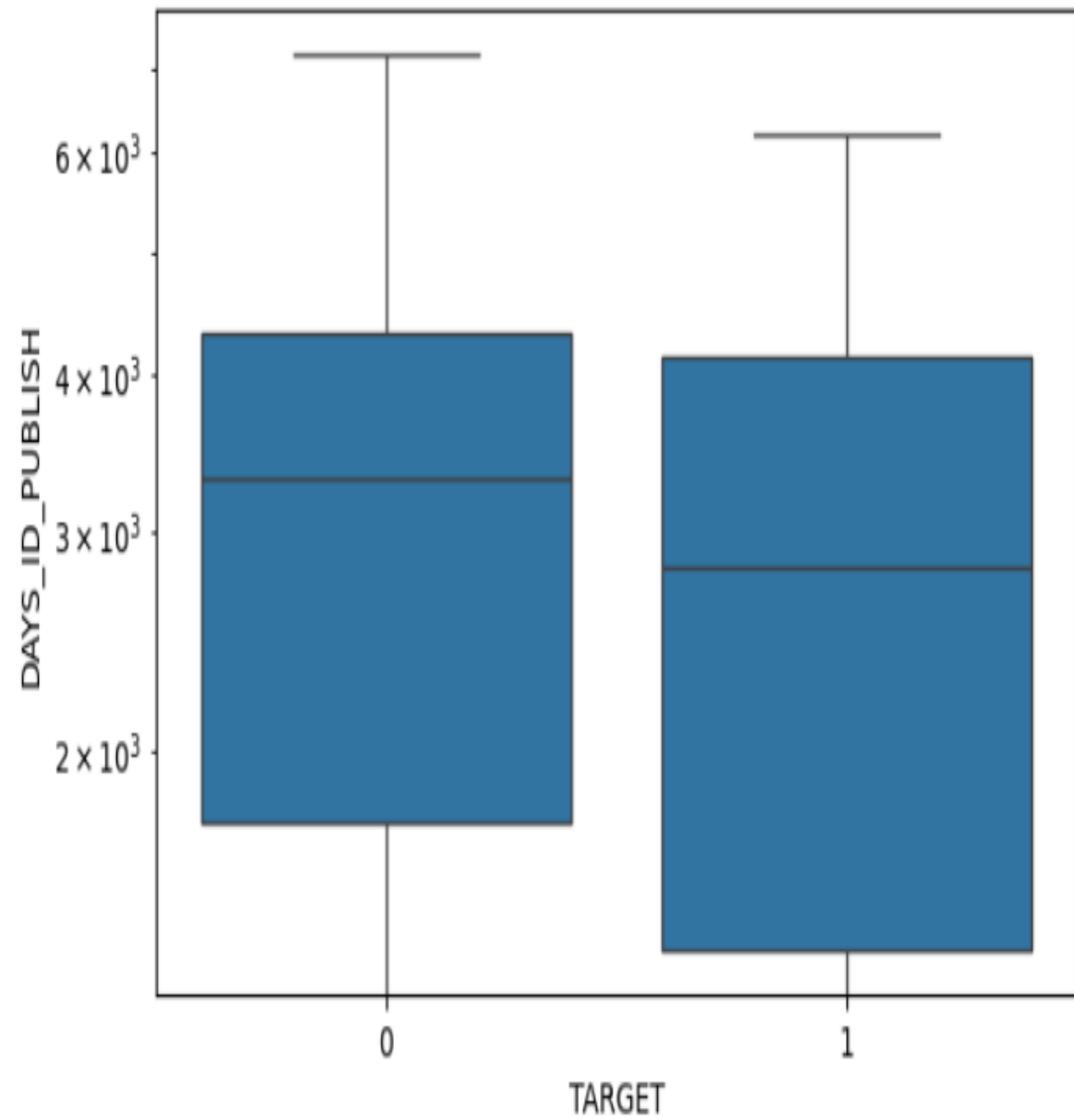
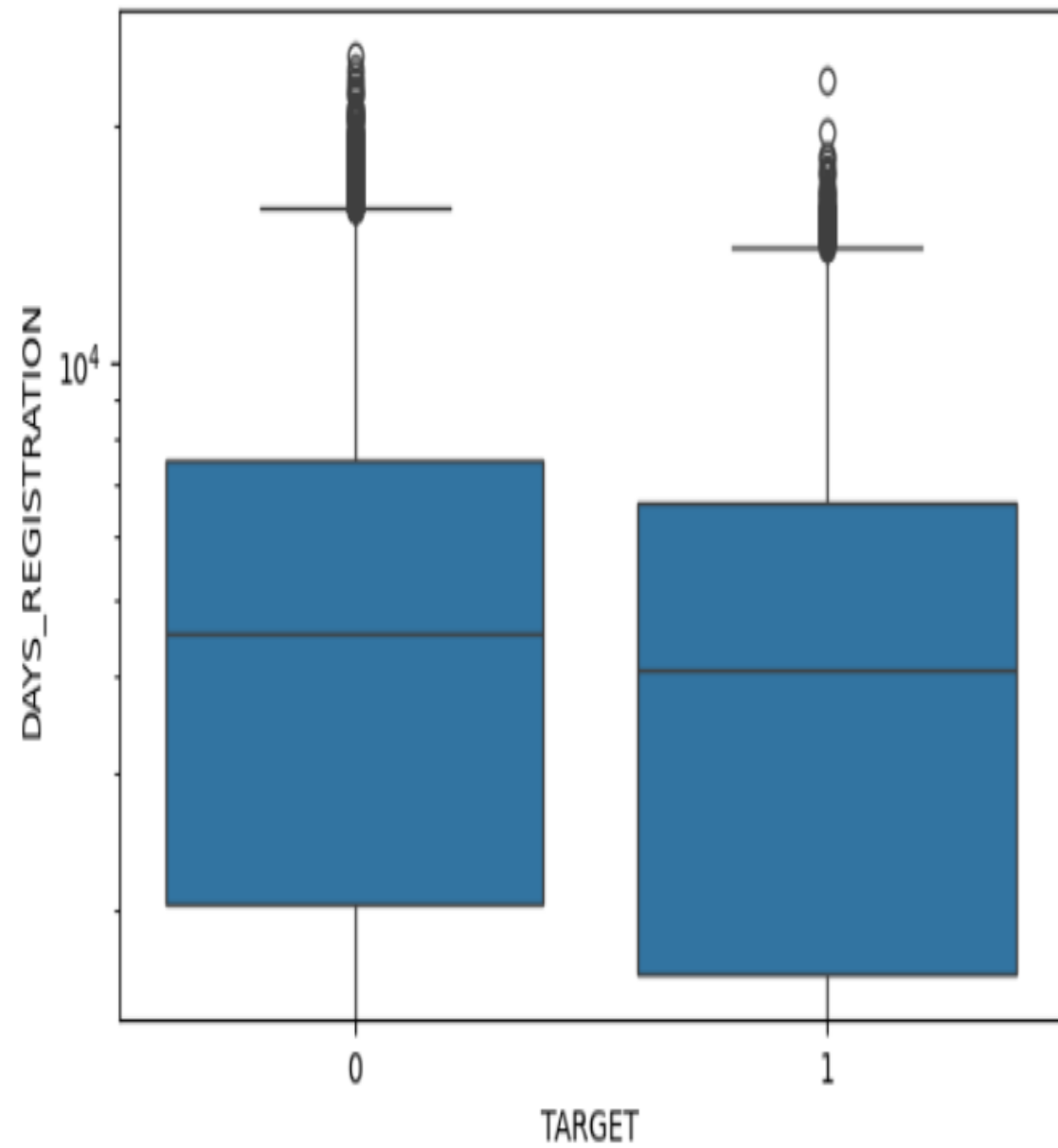




Bi-variate Categorical- continuous Plot



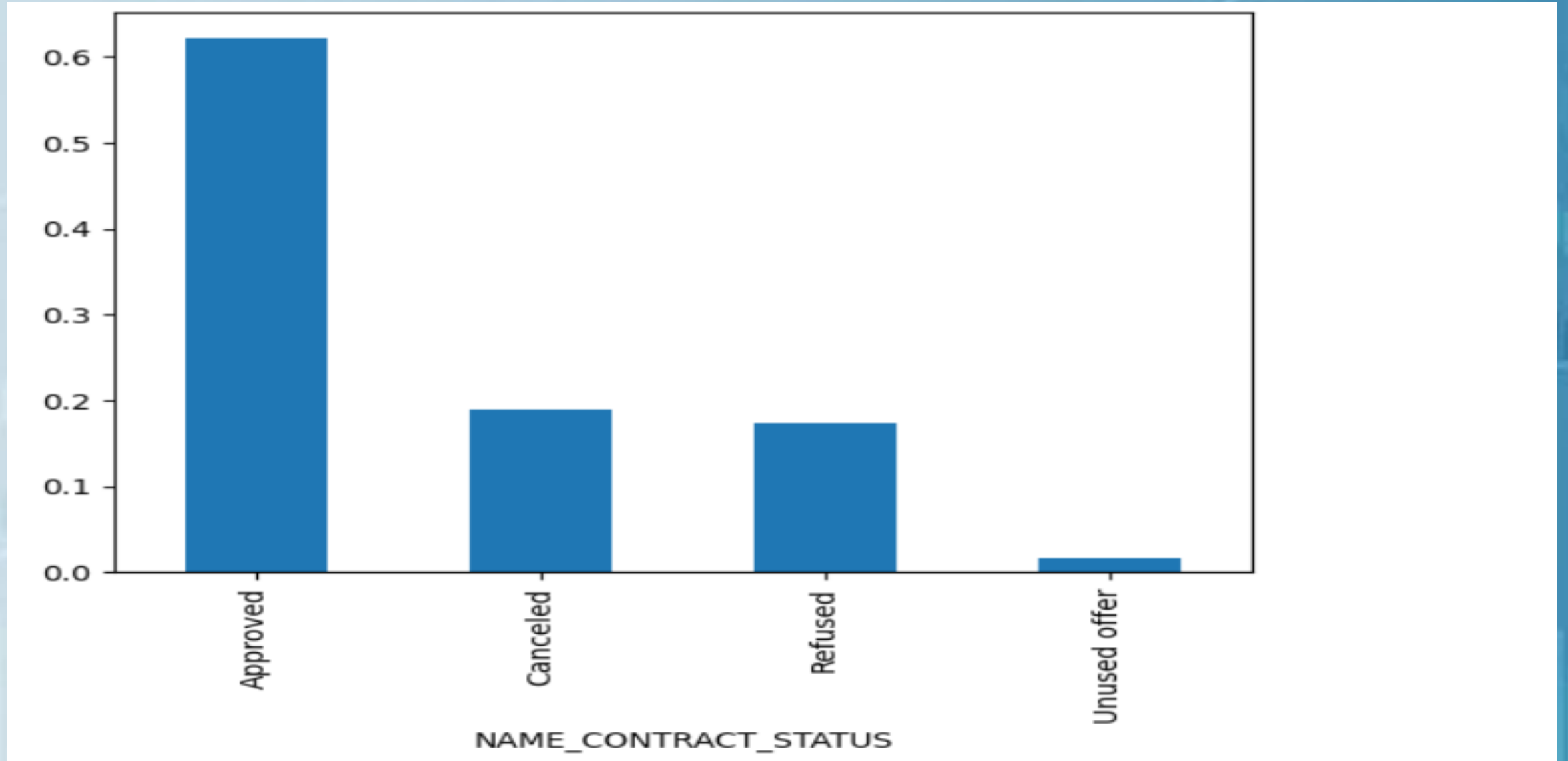


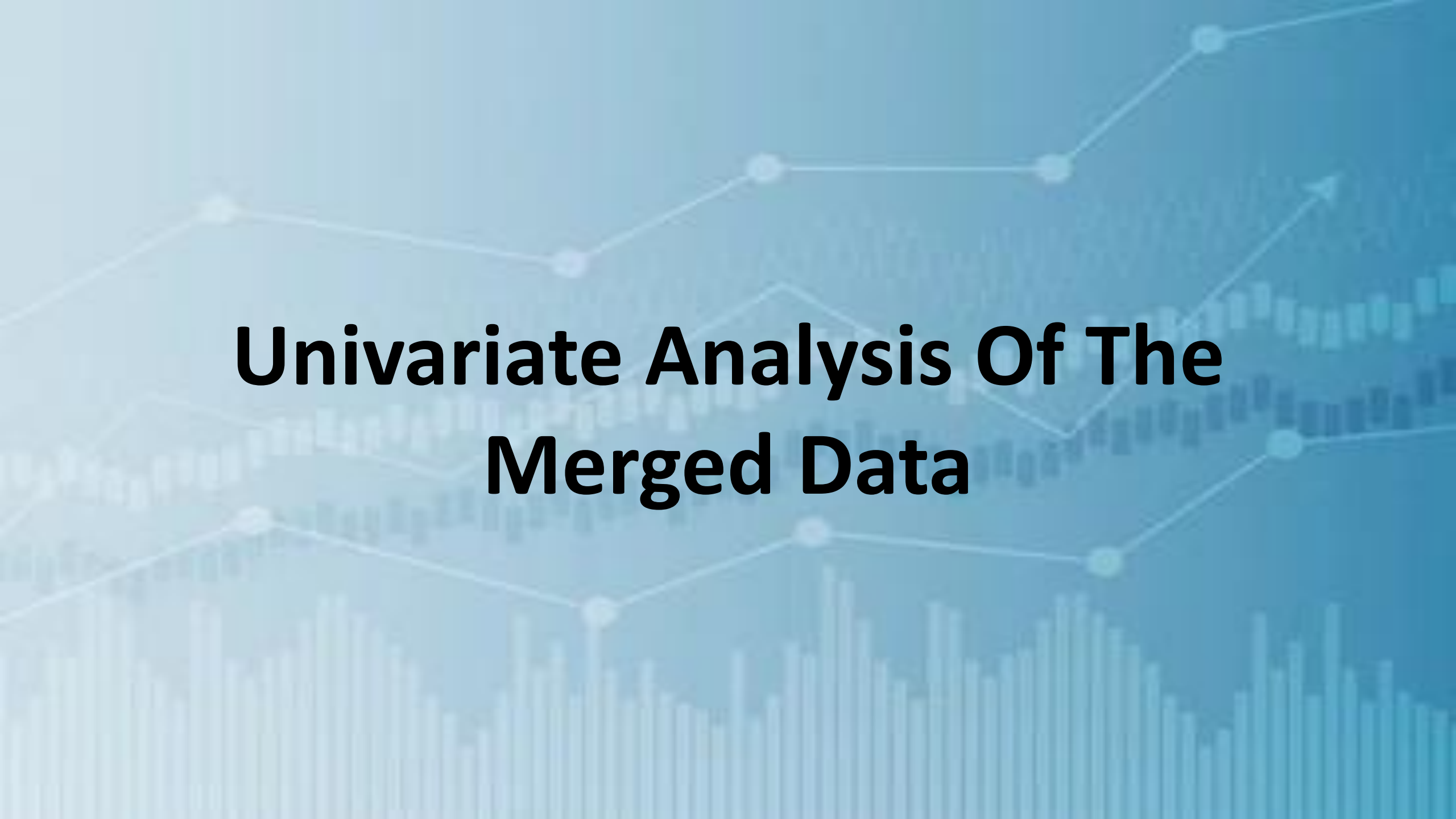


Insights from bivariate continuous plots-

- There Exists More Clients Who Changed They're Their Enrollment Details After 4000 Days Of Approval Of Loan.
- For Few Not Default Clients, Time Taken To Publish Id's Are Greater Than Default Clients.
- The Application Process Start Hours Taken For Default And Not Default Cases Are Analogous.
- In Non-Default Cases, People Keep Their Phone Numbers For Greater Time Than Default Cases.
- People With Greater Number Of Days Born Count Are Less Defaulters.
- Amt_goods Price Contains More Outliers In Non-default Case Than Default Case.
- In Default Cases, Most Of The Clients Amount Annuity Tends To Be Greater Than Median I.E.. 25000.
- Whose Credit Amount Is Greater Than 50000 Are Less Defaulters .
- People With Higher No Of Employment Days Are Less Defaulters.
- Maximum Defaulters Are Having Less Total Income.

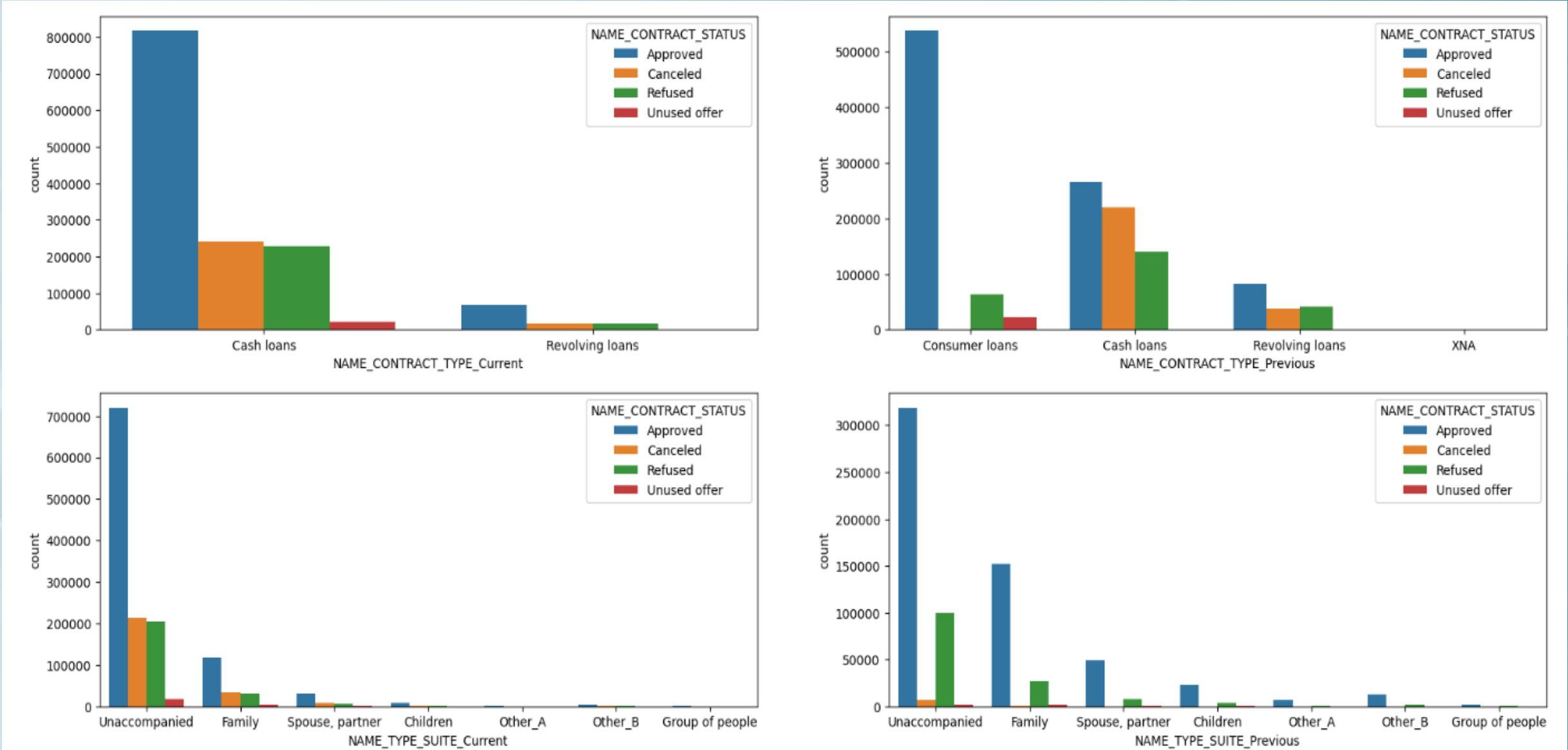
Previous_application_df NAME_CONTRACT_STATUS

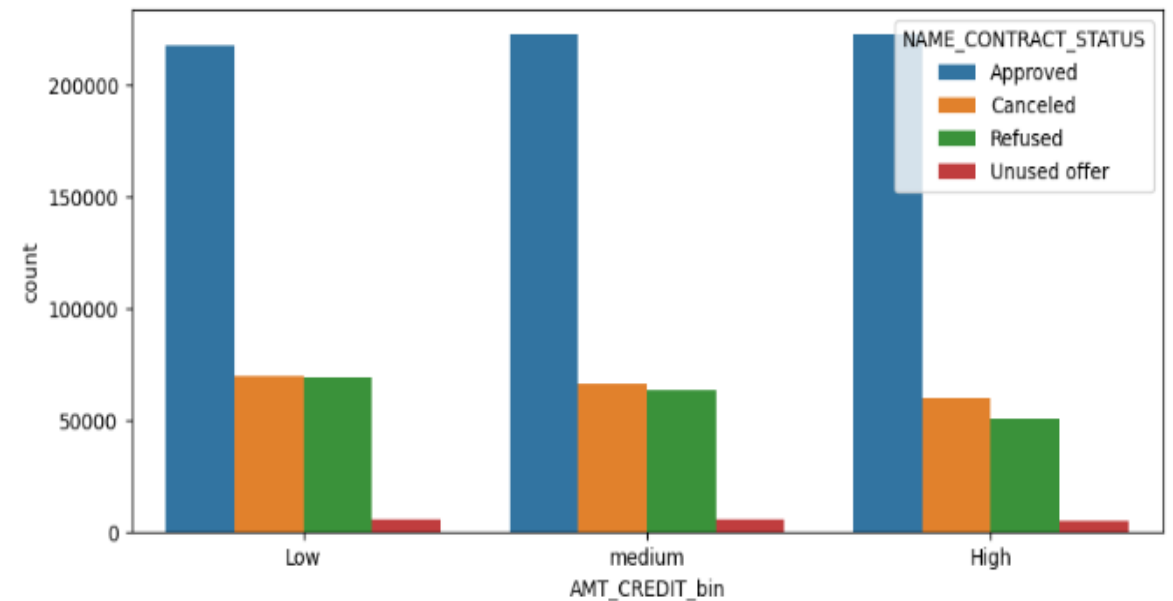
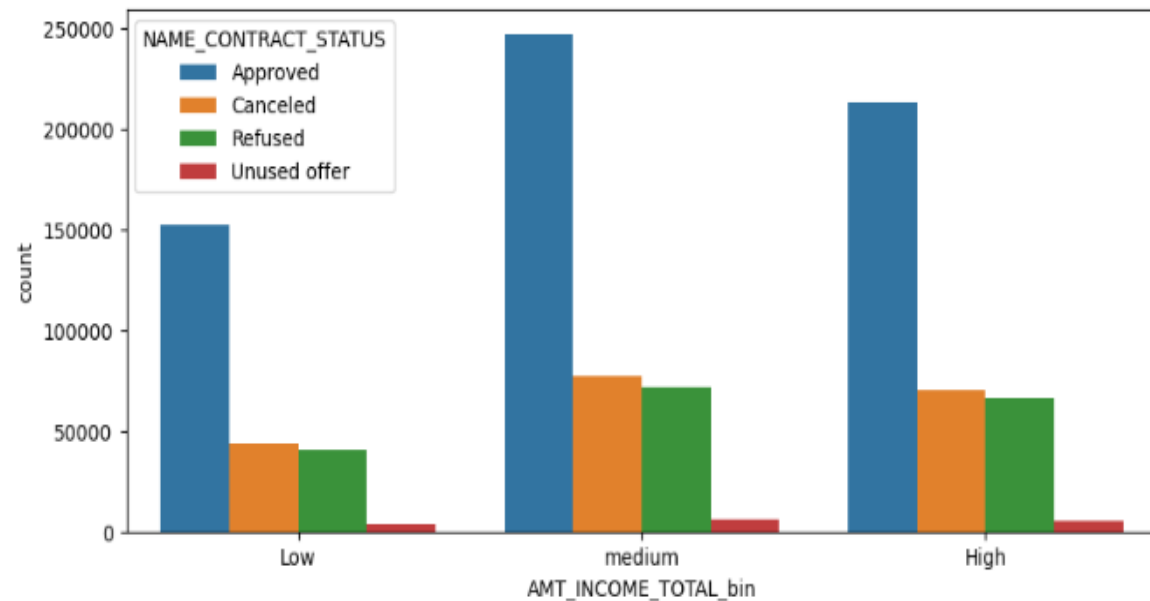
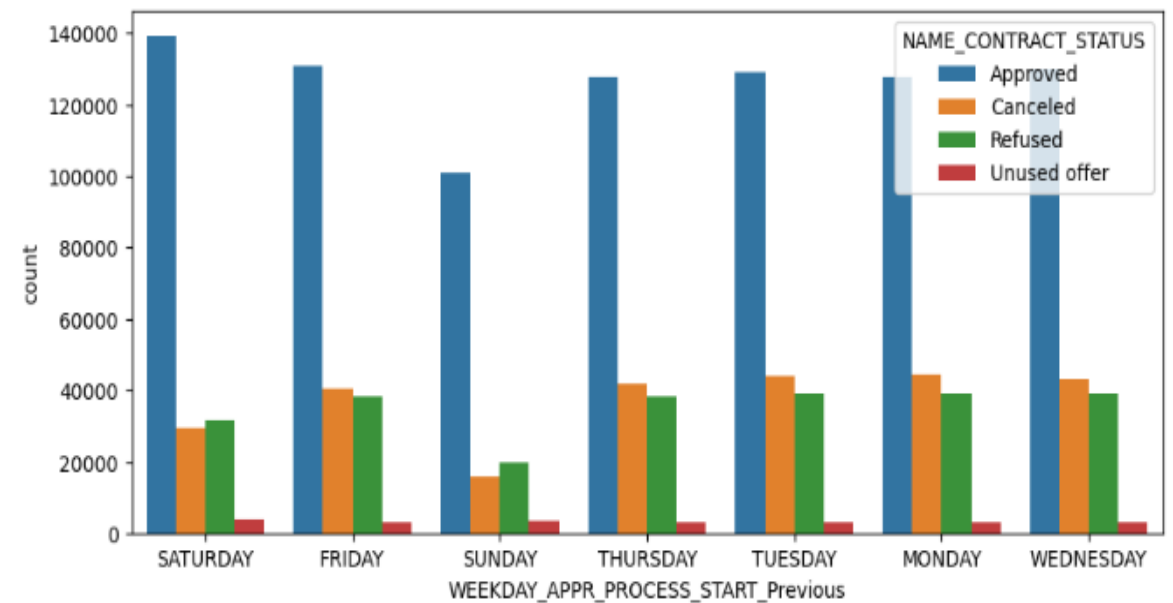
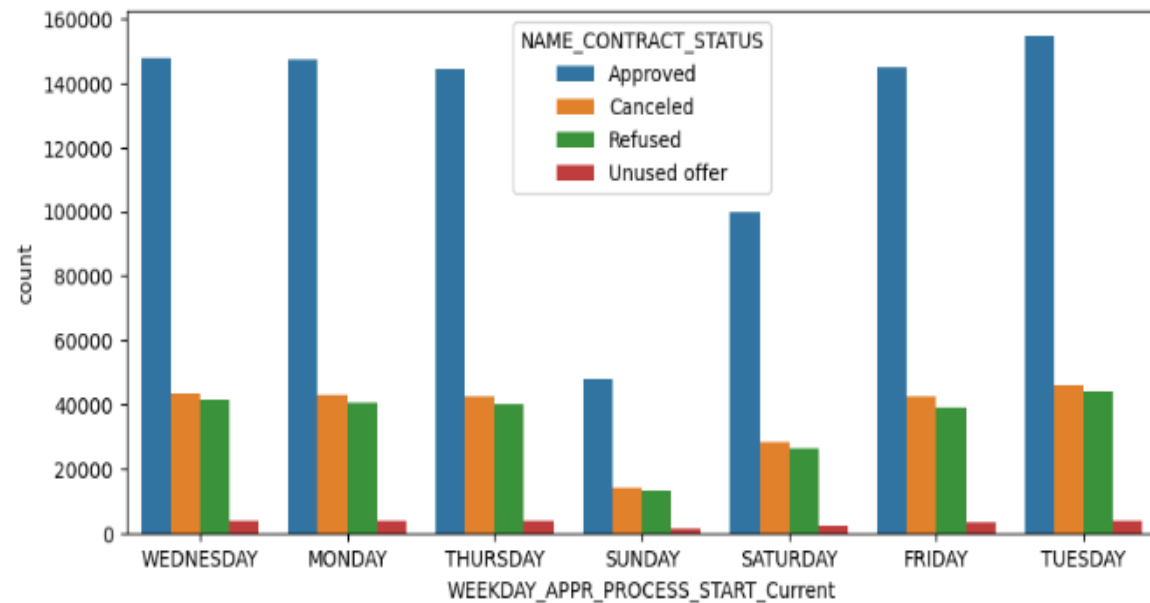


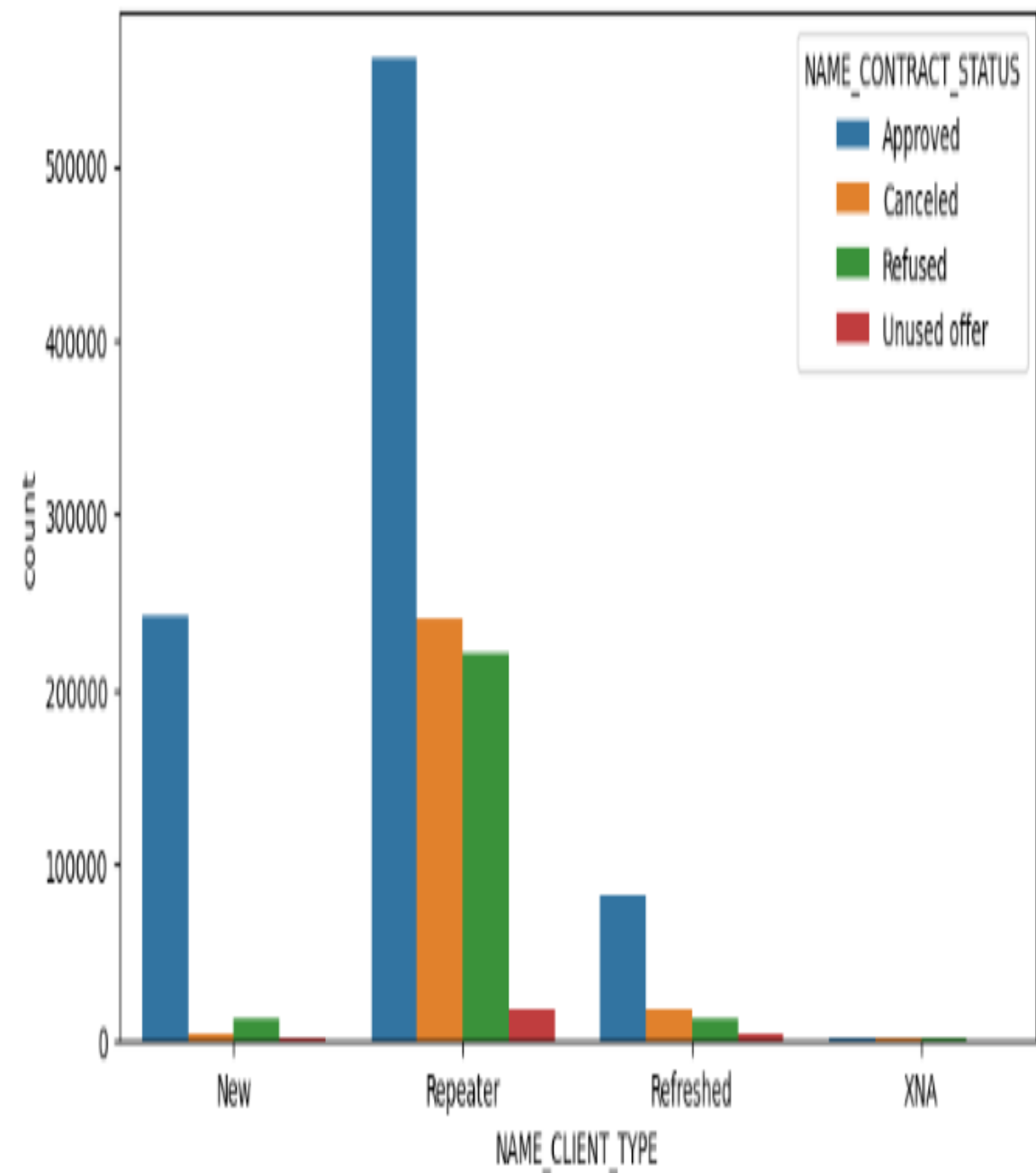
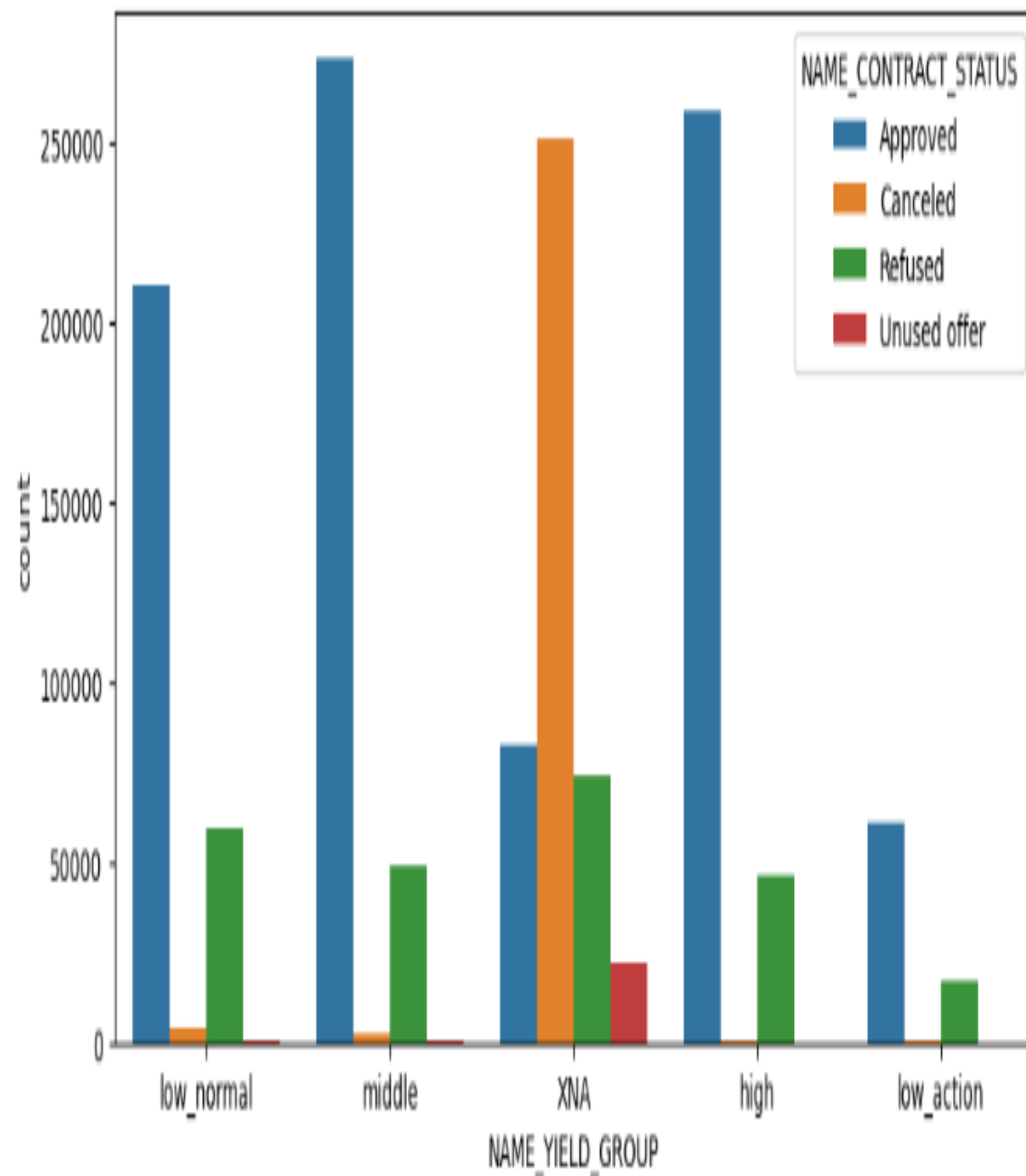
The background of the slide features a light blue gradient with faint, stylized financial data visualizations. These include several line graphs with circular markers at data points and a bar chart with numerous vertical bars of varying heights. The overall aesthetic is clean and professional, typical of a business or academic presentation.

Univariate Analysis Of The Merged Data

Univariate Categorical Analysis Of Merged Data



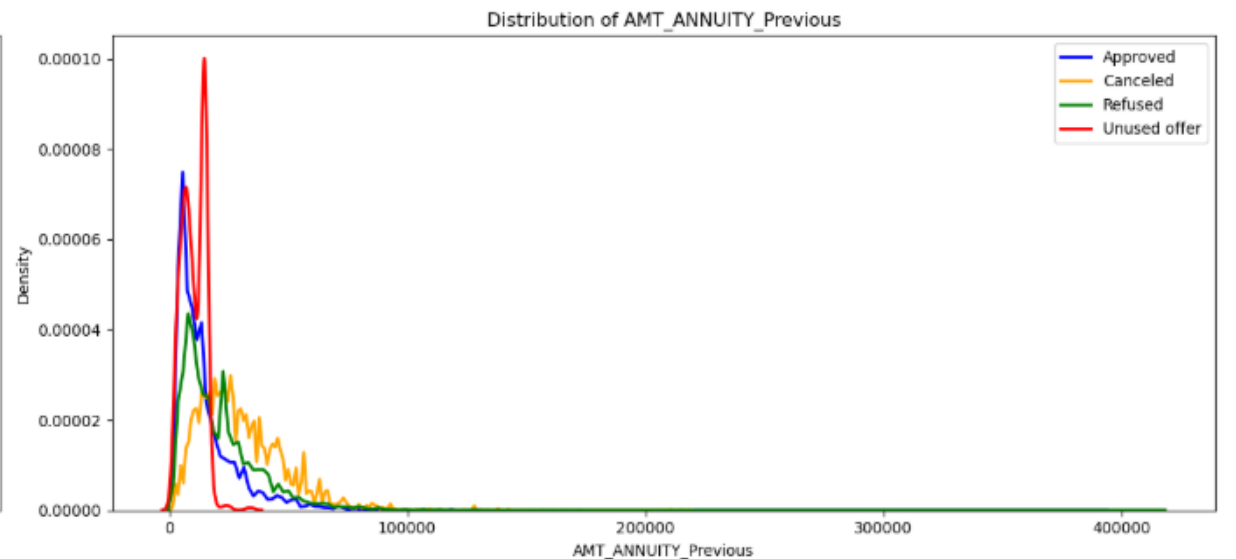
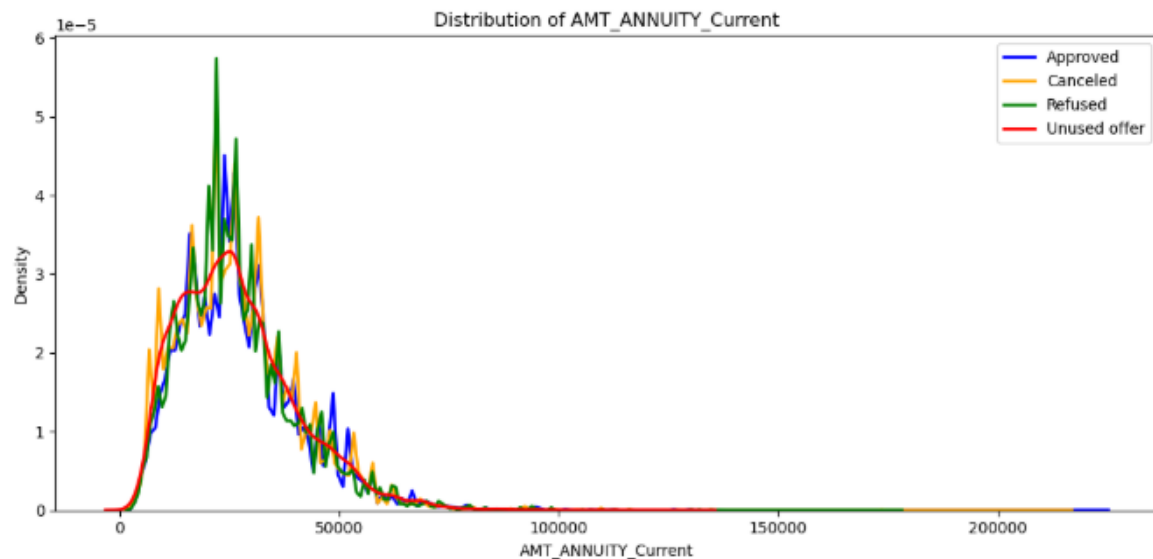
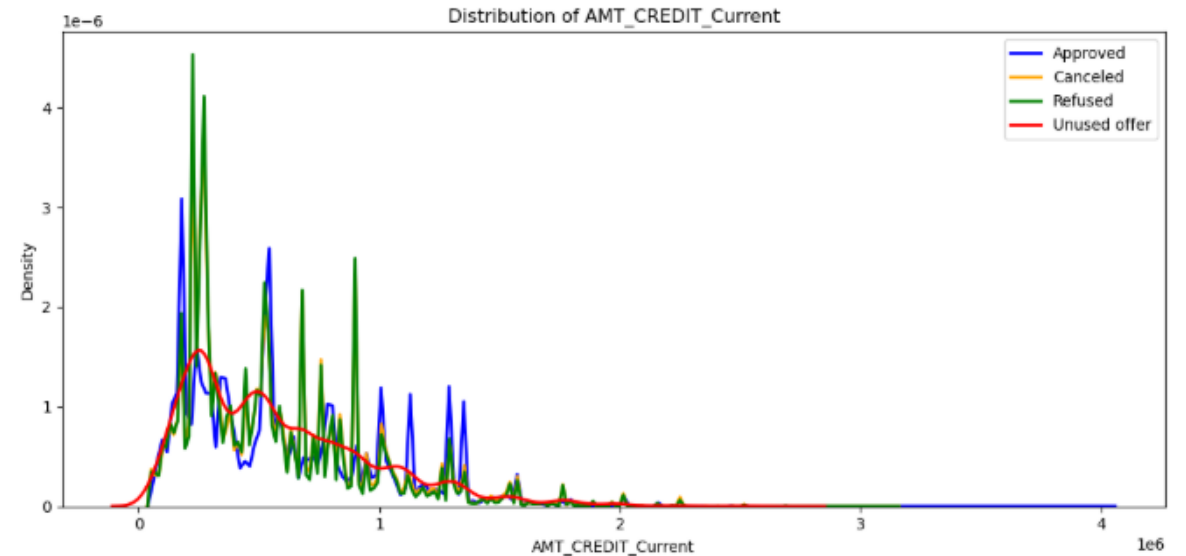
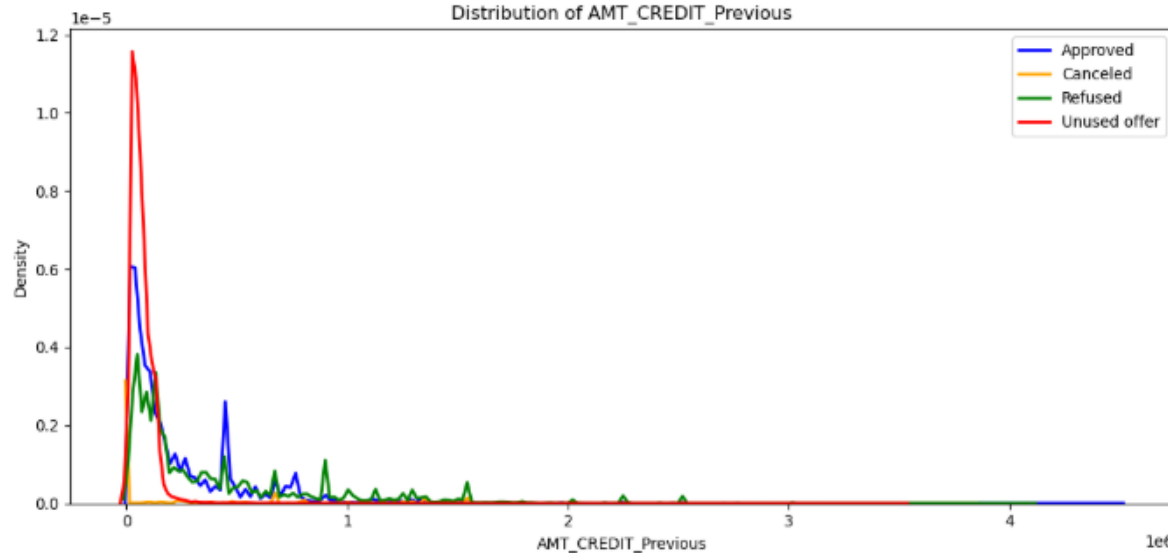


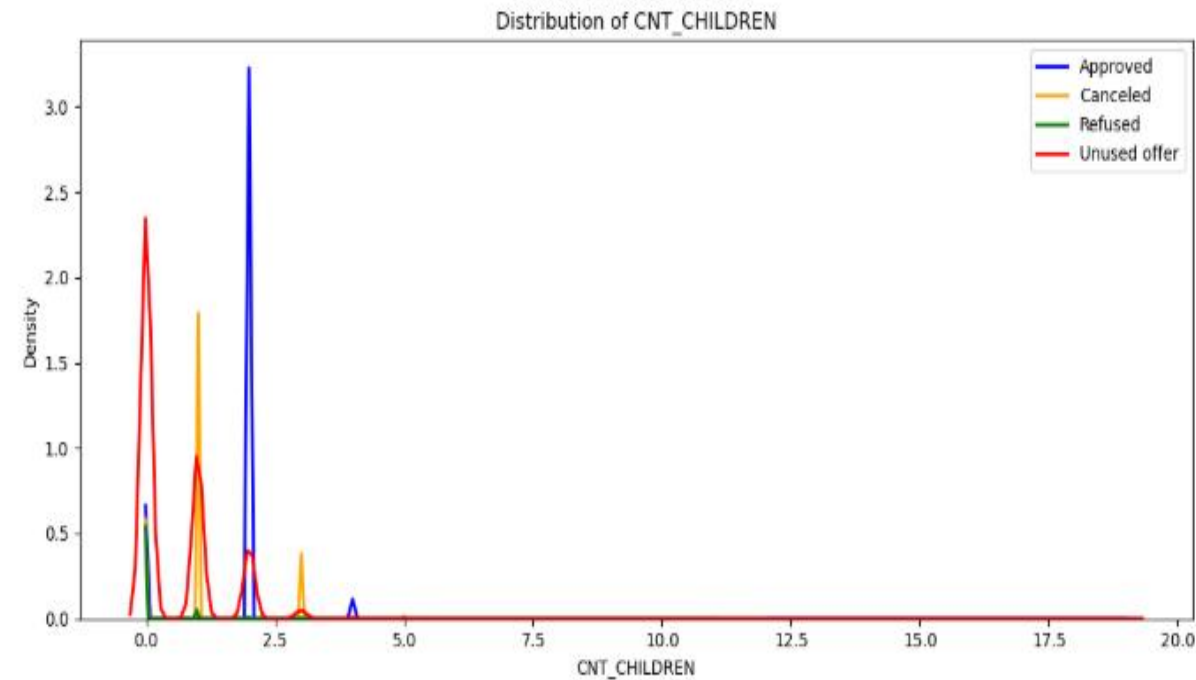
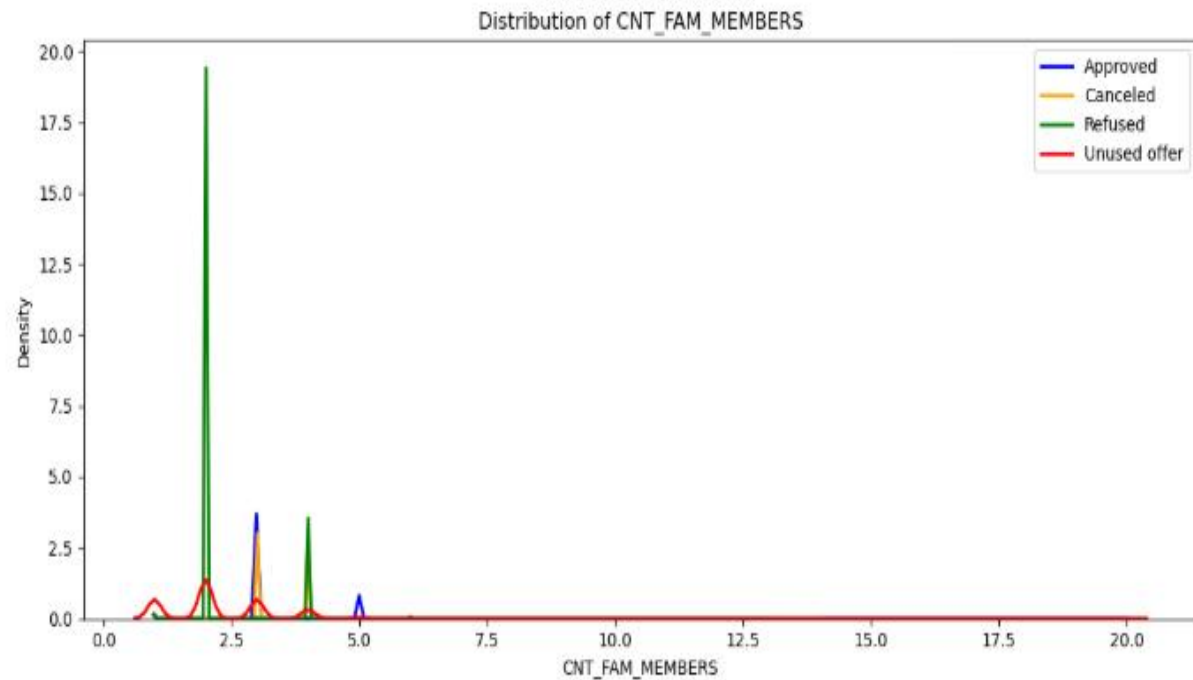
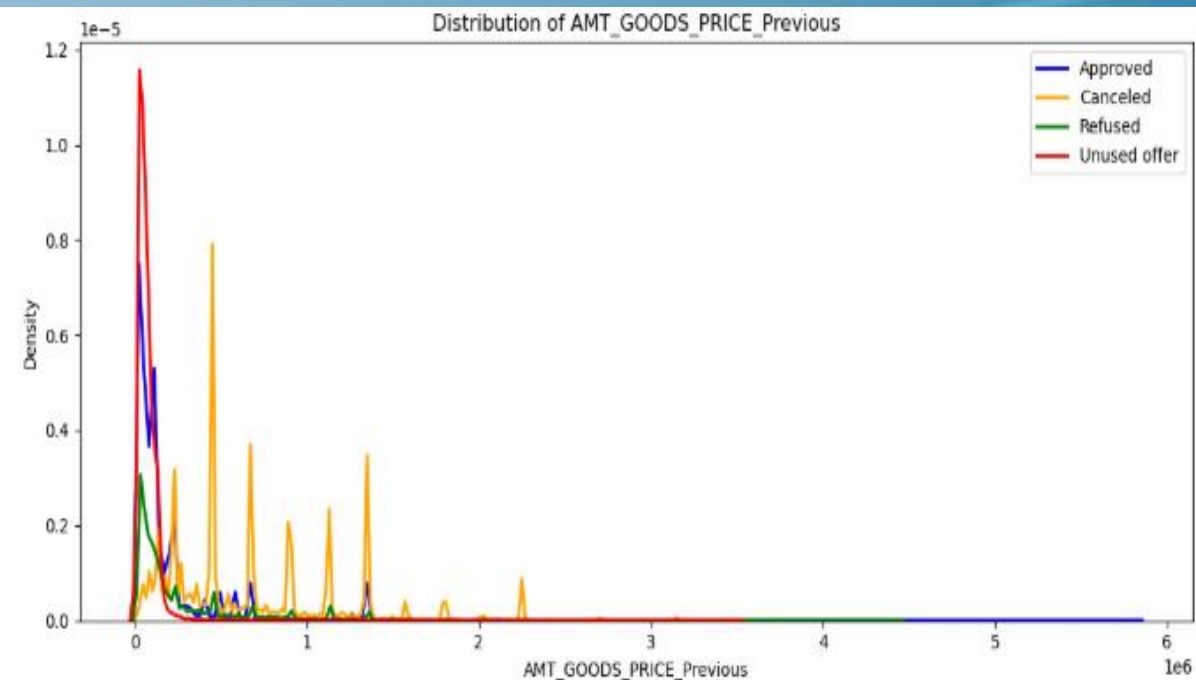
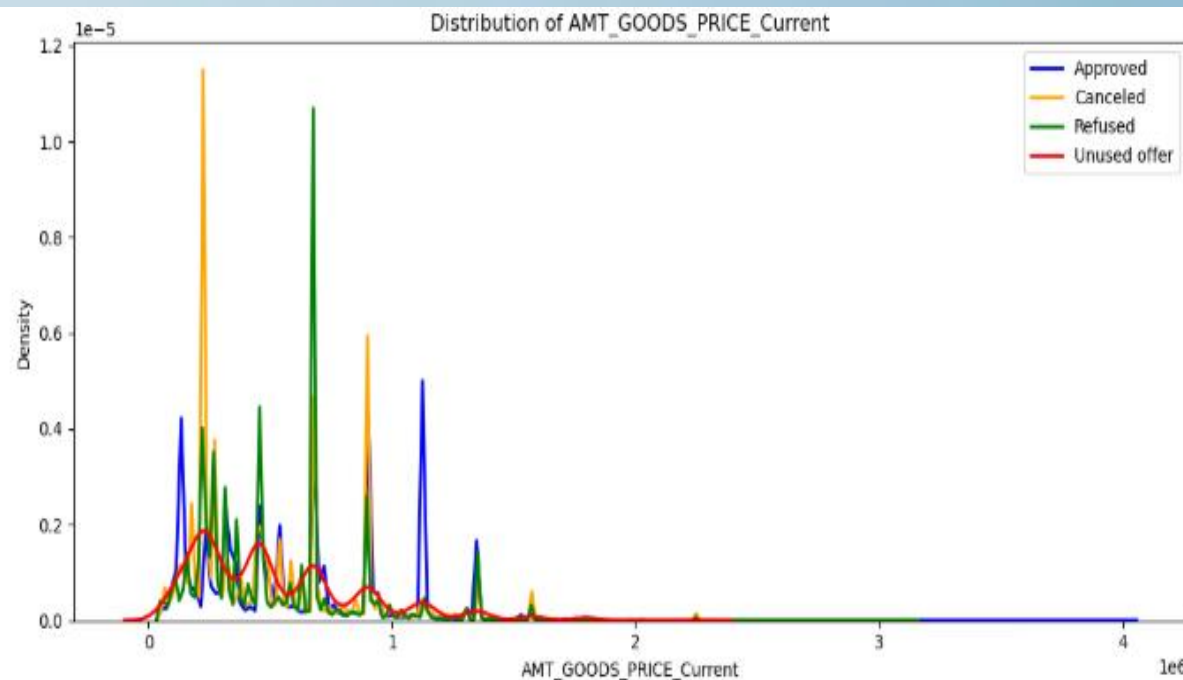


Insights From Univariate Categorical Analysis Of Merged Data-

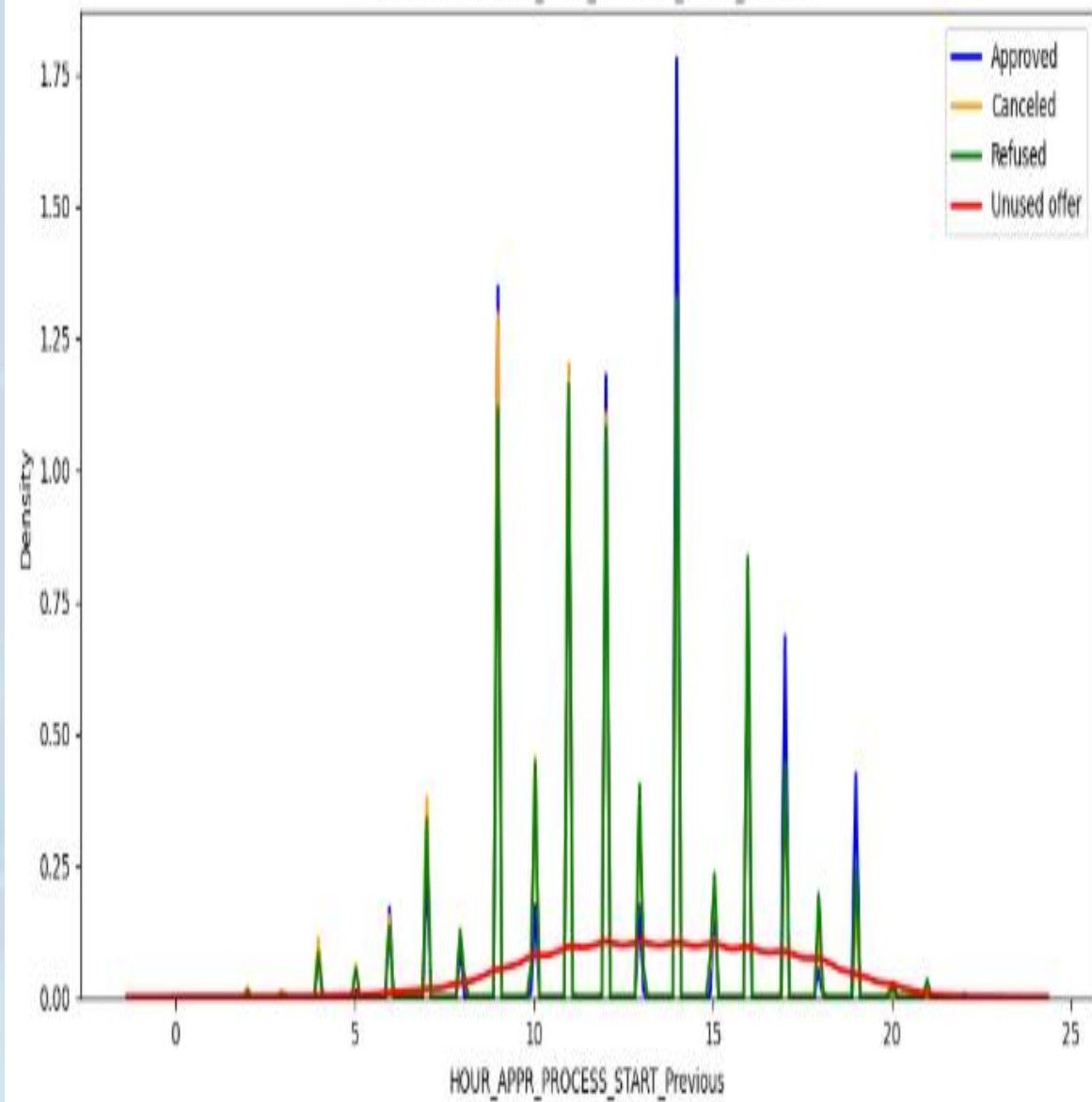
- Repeater Are Having Highest Number Of Approved Loans.
- Middle Name_yield_group Are Having Highest Approval.
- Amt_credit_bin Values Does Not Affect Loan Approvals.
- For Medium Category In Amt_income_total_bin The Approval Is Highest .
- In Previous Application_df Saturday Has The Highest Approval Rate.
- But In Current Application_df It Is Tuesday.
- Unaccompanied Has The Highest Number In Both (In Name_contract_type_previous And Name_contract_type_current).
- Bank Is Only Giving Two Types Of Loans Currently -Cash And Revolving Loans.
- Bank Was Providing Cash, Revolving And Consumer Loans Previously.
- Previously Number Of Consumer Loans Were Highest And Now Cash Loans Is Highest In Number.

Univariate Numerical Analysis Of Merged Data

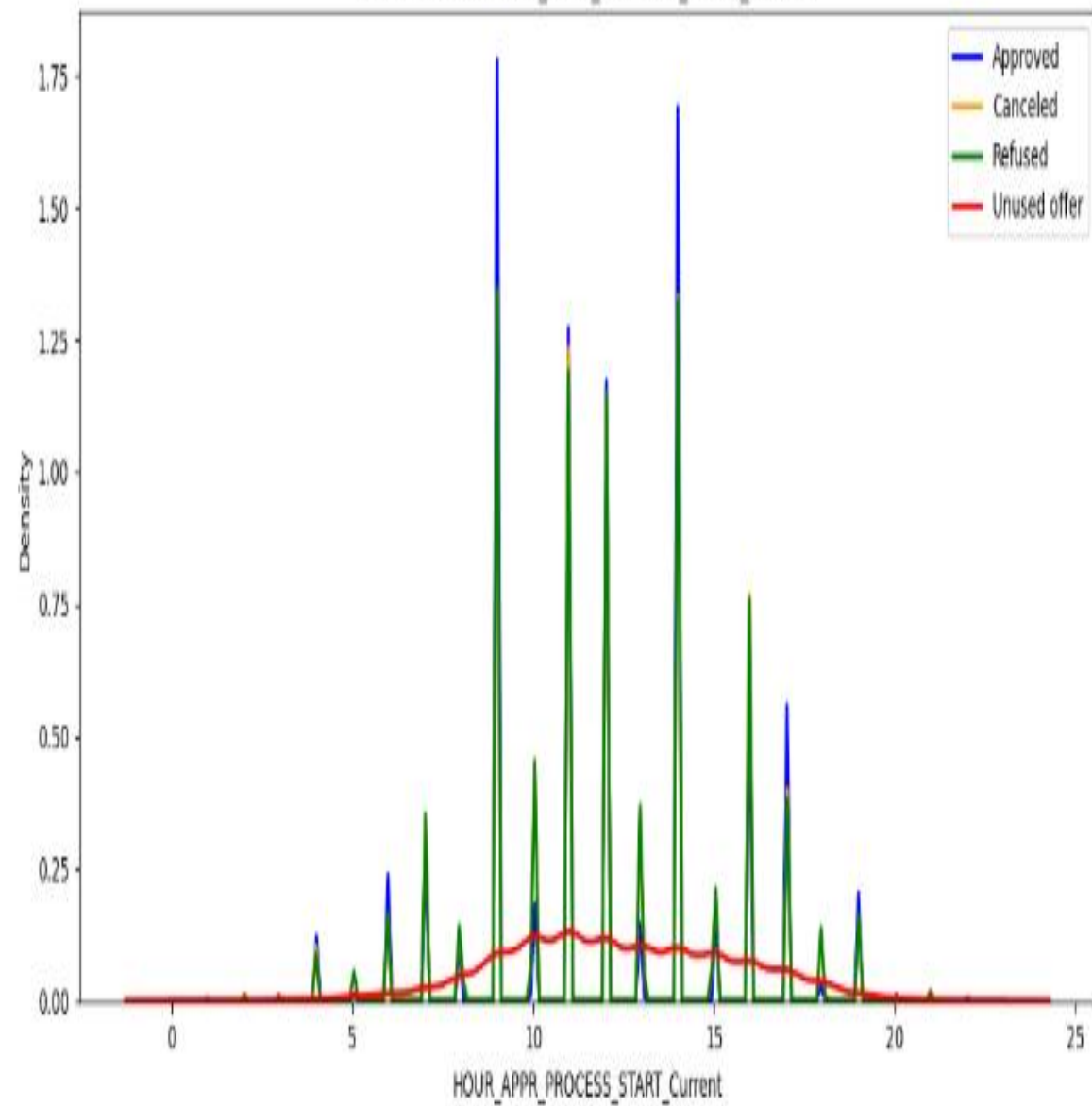




Distribution of HOUR_APPR_PROCESS_START_Previous



Distribution of HOUR_APPR_PROCESS_START_Current



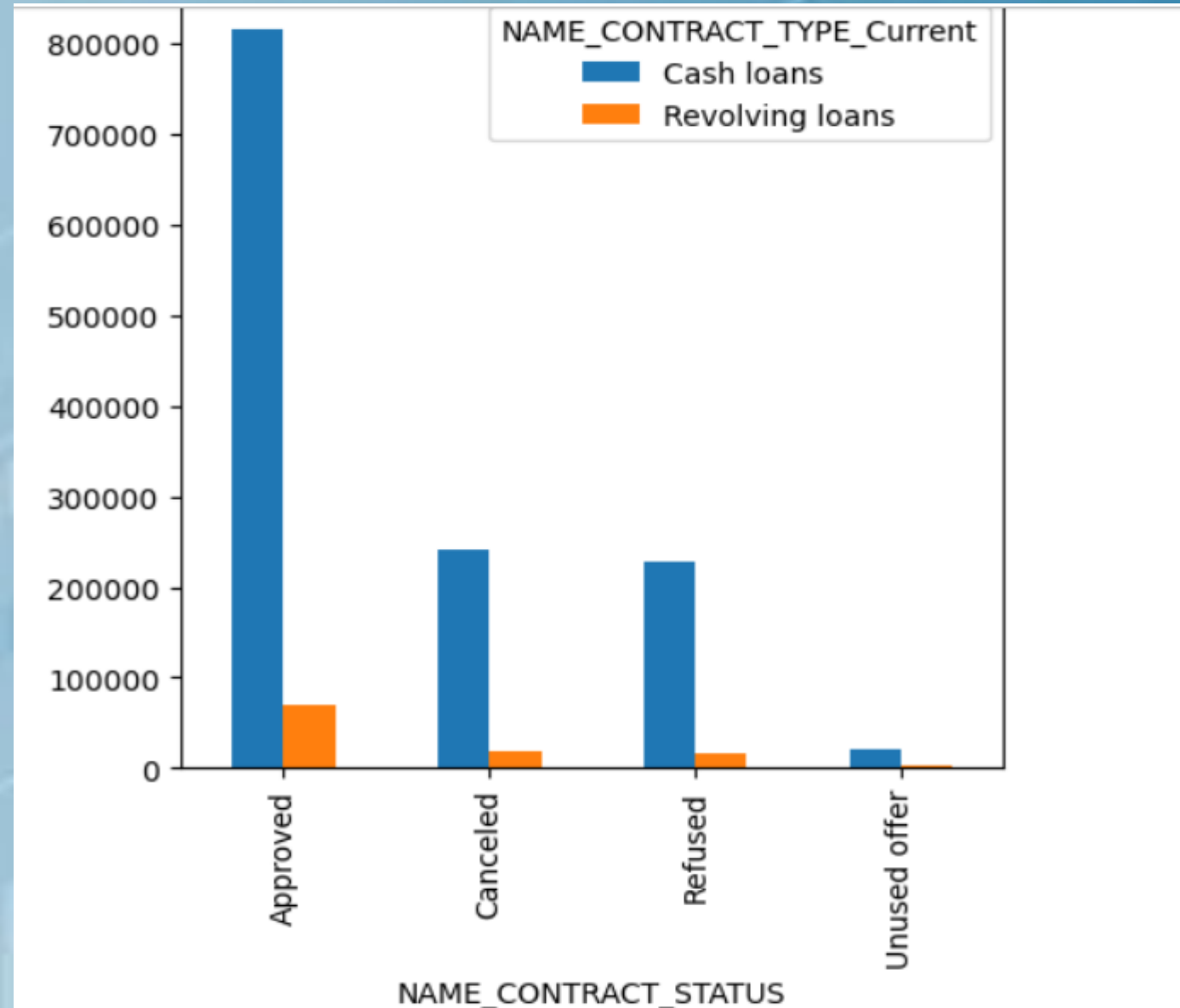
Insights From Univariate Numerical Plots-

- For Both Current And Previous Data High Number Of Applications Are Filed In 9 AM To 2 PM, So Busiest Hours For Bank Are During 9 AM To 2 PM.
- It Can Be Seen That Nuclear Family Take More Loans.
- Bank Had High Unused Offers Previously But Refused Is High Incase Of Amt_goods_price Currently.
- Bank Had High Unused Offers Previously And Cancelled/Refused Offers Are Similar For Amt_annuity Currently.
- Bank Had High Unused Offers Previously And High Number Of Refused Offers For Amt_credit Currently.

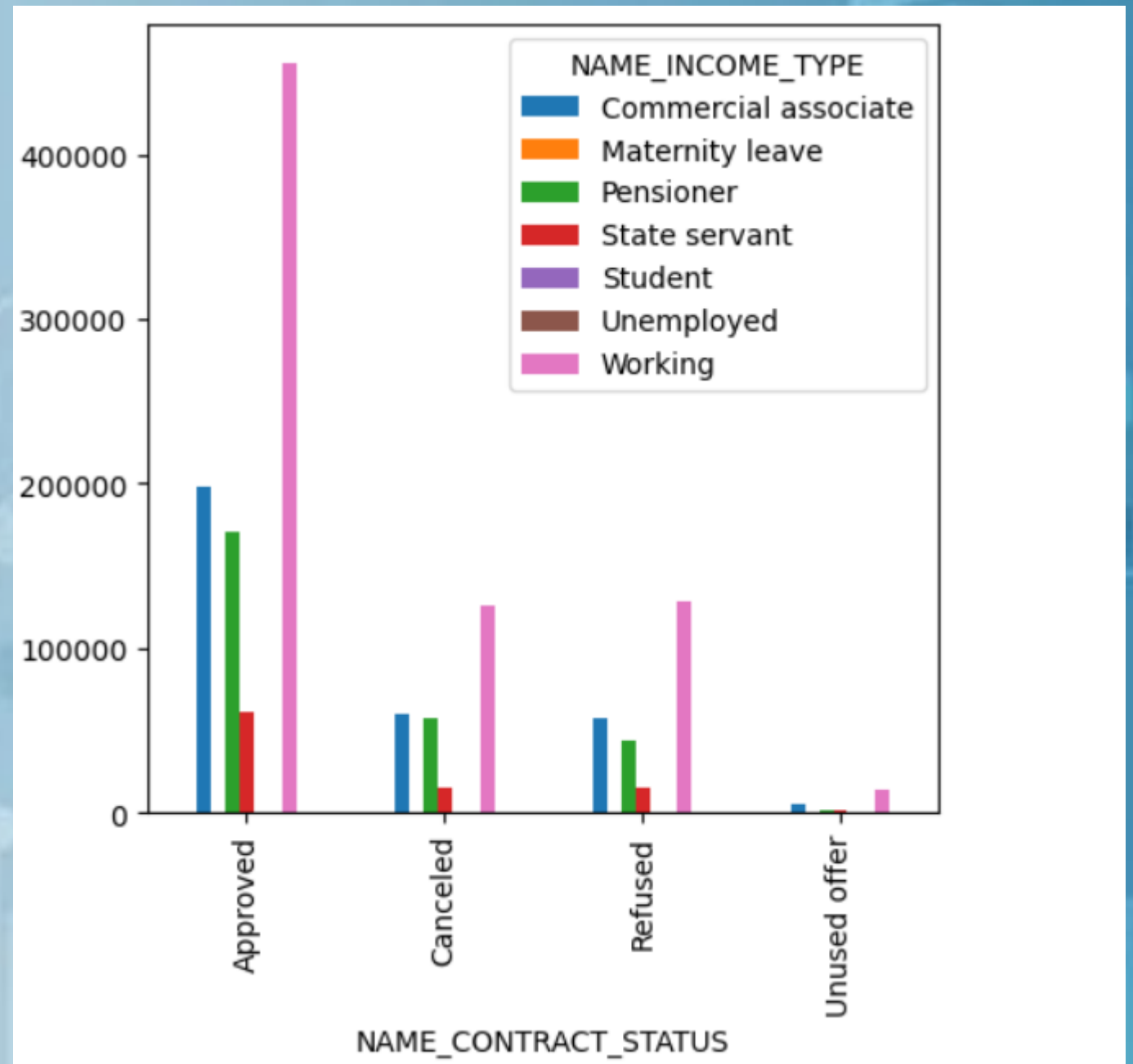
The background of the slide features a light blue gradient with faint, stylized financial data visualizations. These include several line graphs with circular markers at data points and a bar chart with numerous vertical bars of varying heights. The overall aesthetic is clean and professional, typical of a business or academic presentation.

Bi-variate Analysis Of Merged Data

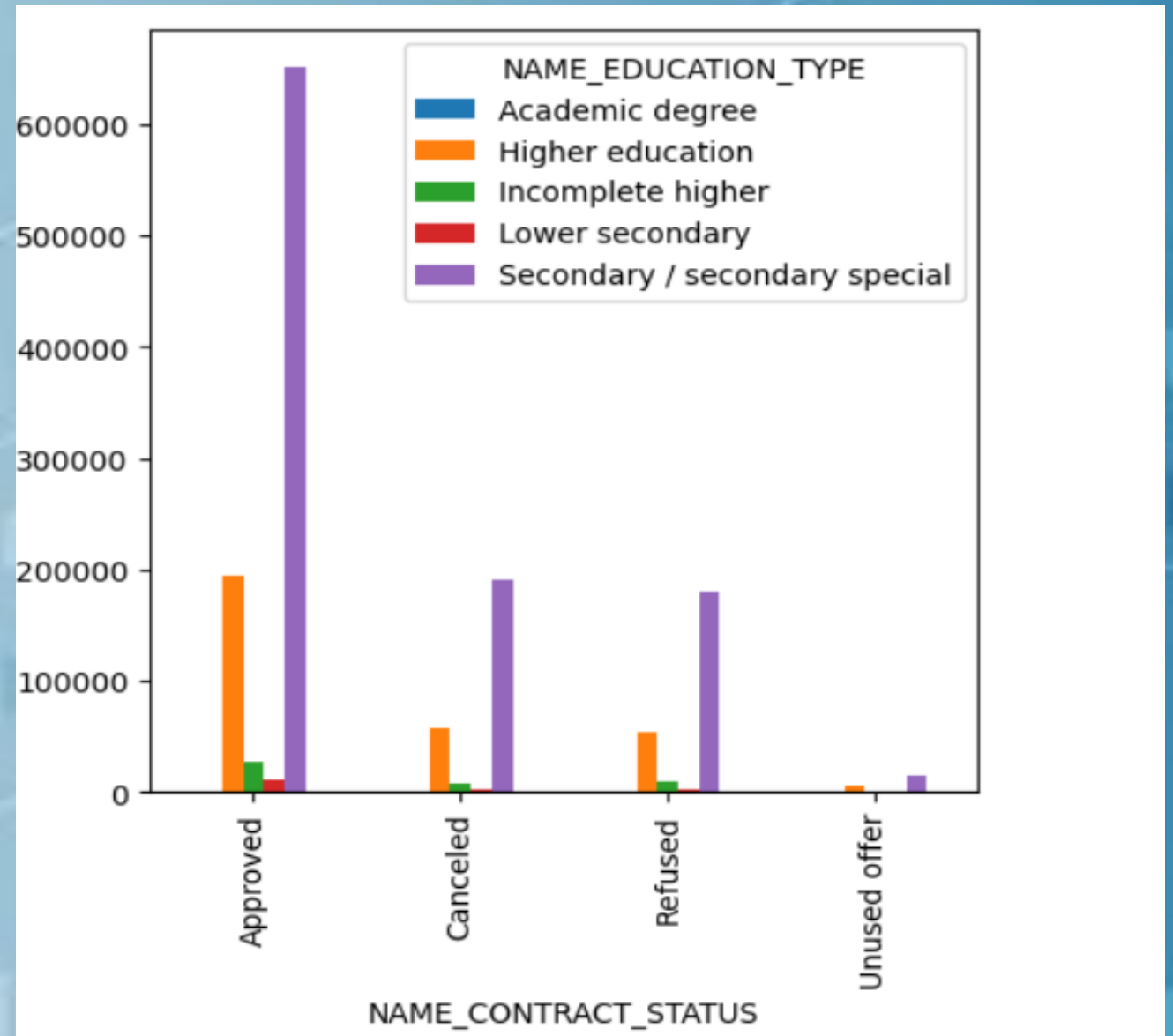
- Insight- Cash Loans Have The Highest Number Of Approved Loans



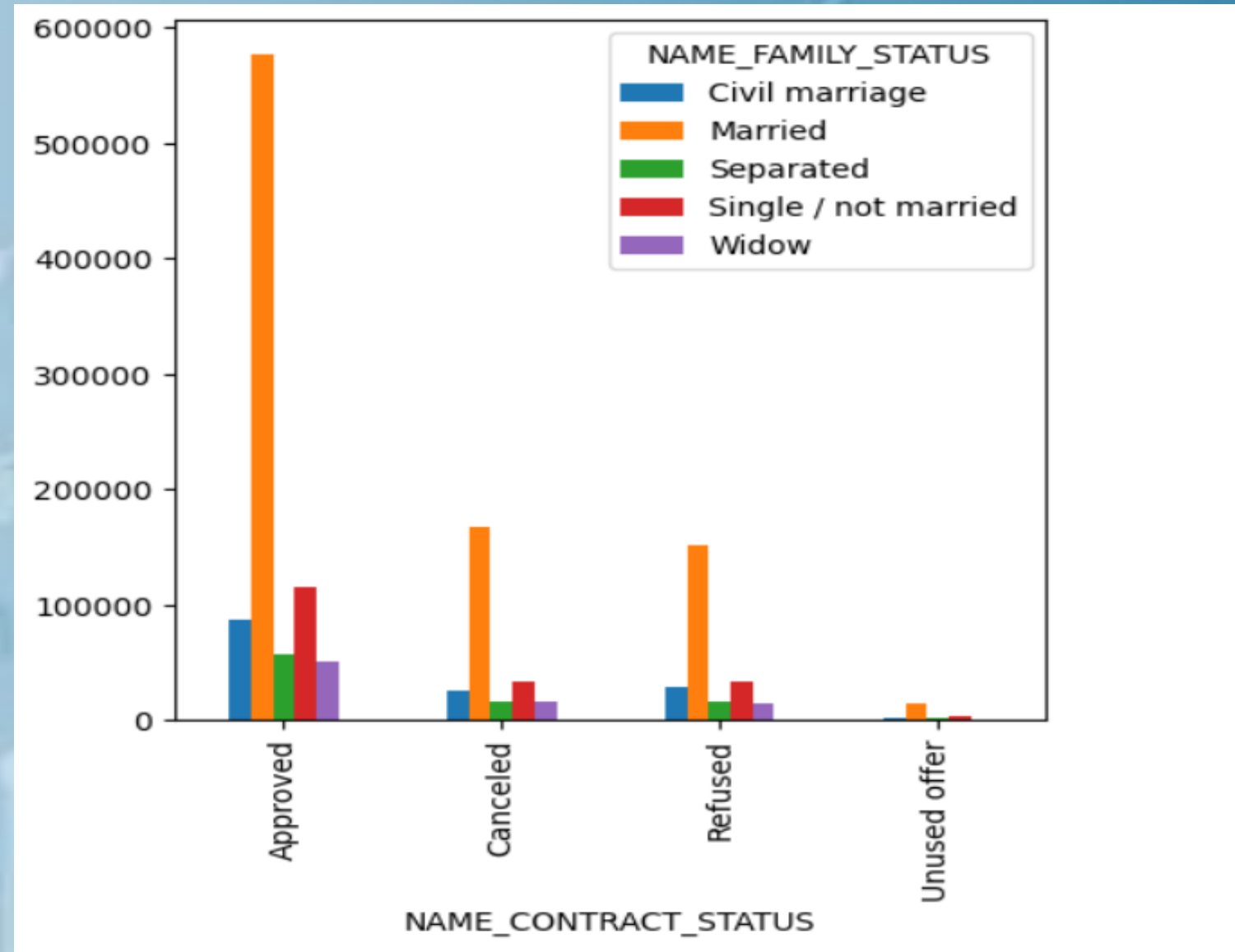
- Insight- Working Applicant Have Highest Number Of Approvals.



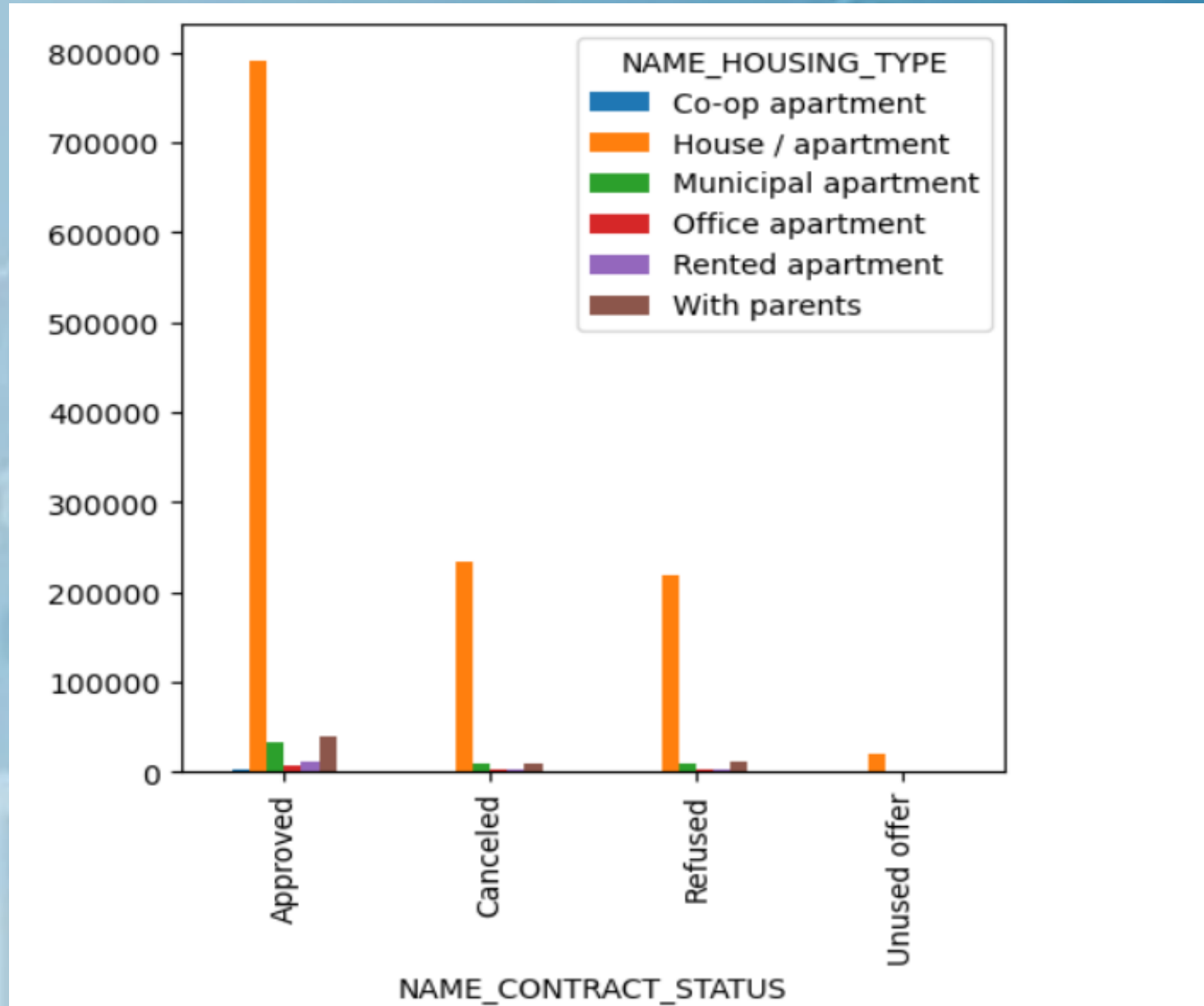
➤ Insight- Secondary/Secondary Special Educated Applicant Have Highest Approval Of Loan.



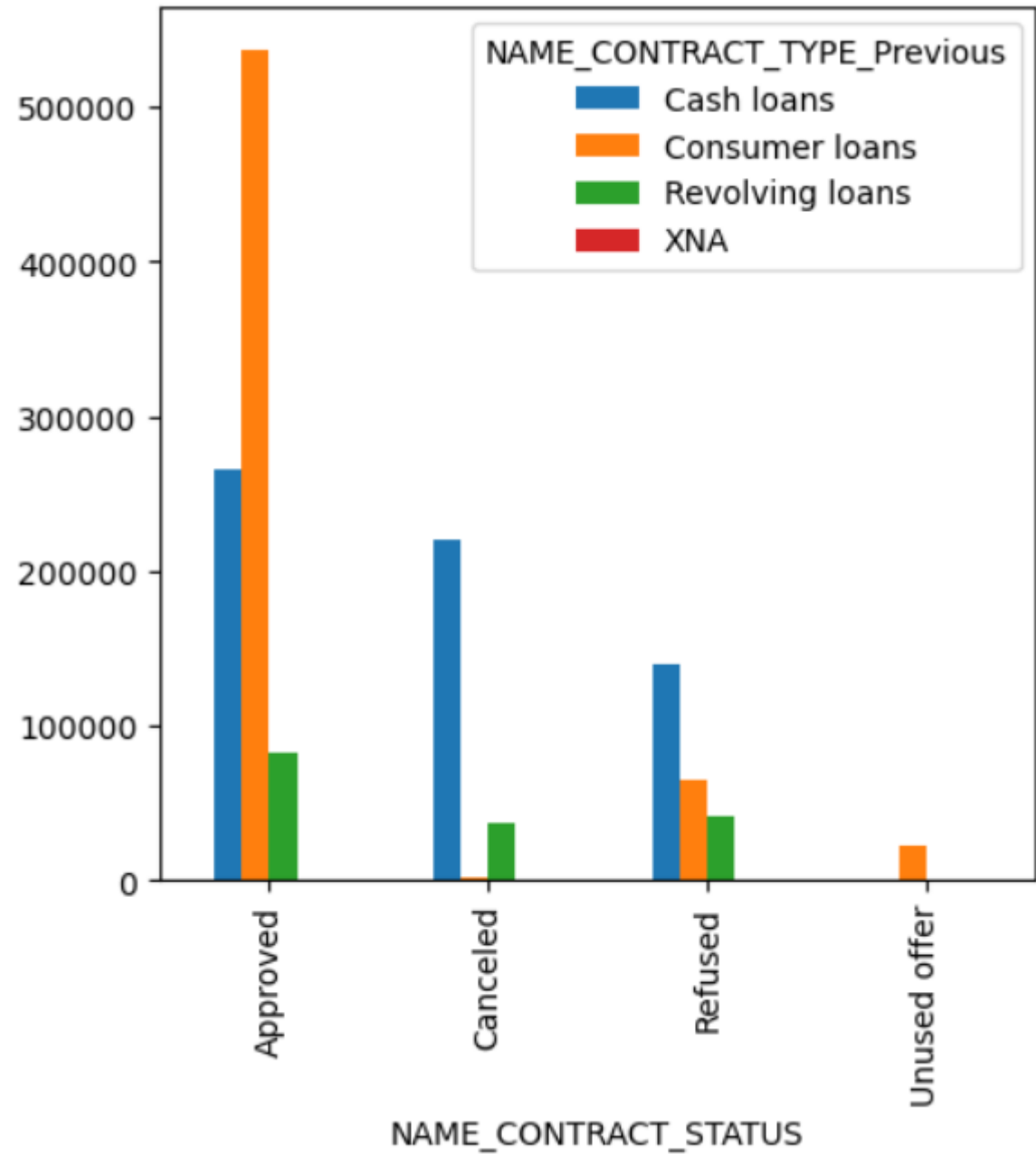
- Insight- Married Applicant Has Highest Number Of Approvals



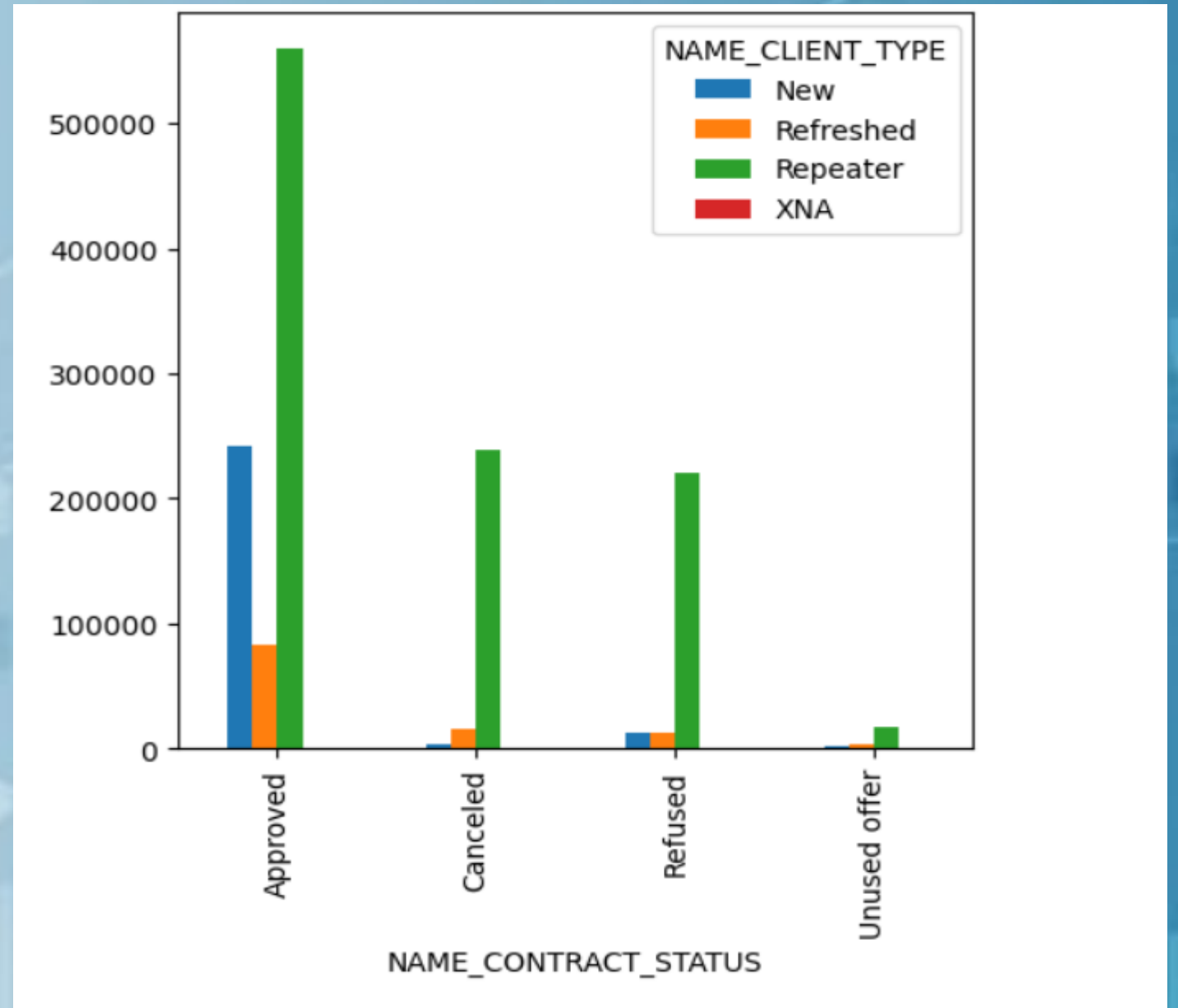
- Insight- House/Apartment Owner Having Highest Number Of Approvals.



- Insight- Consumer Loans Have Highest Number Of Approval.

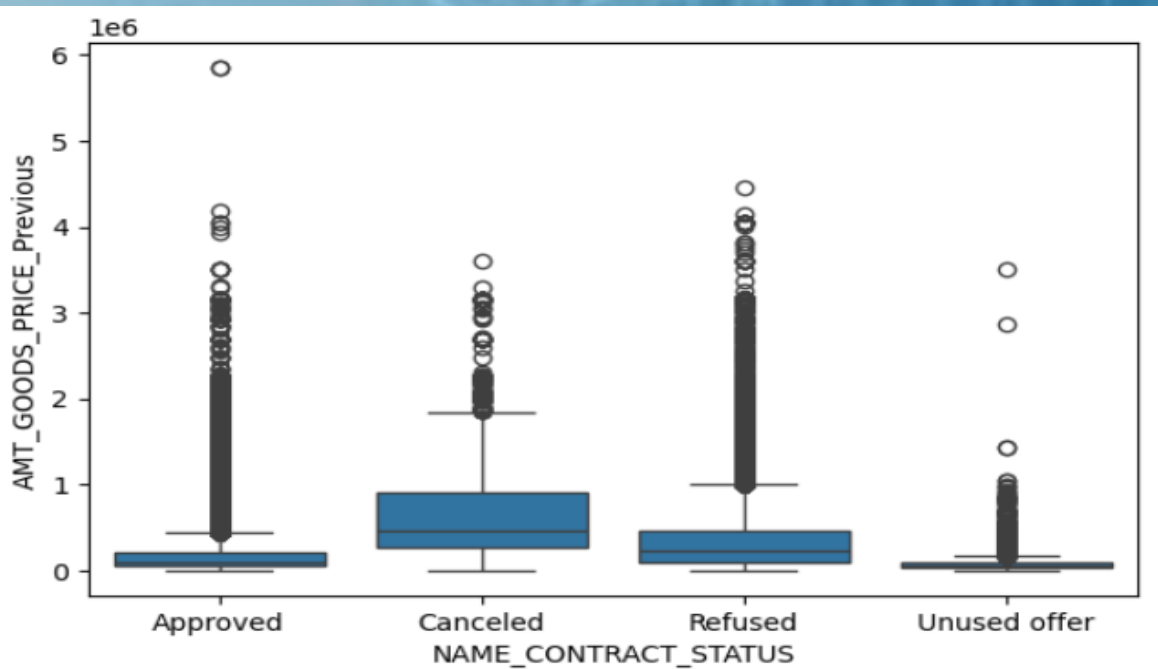
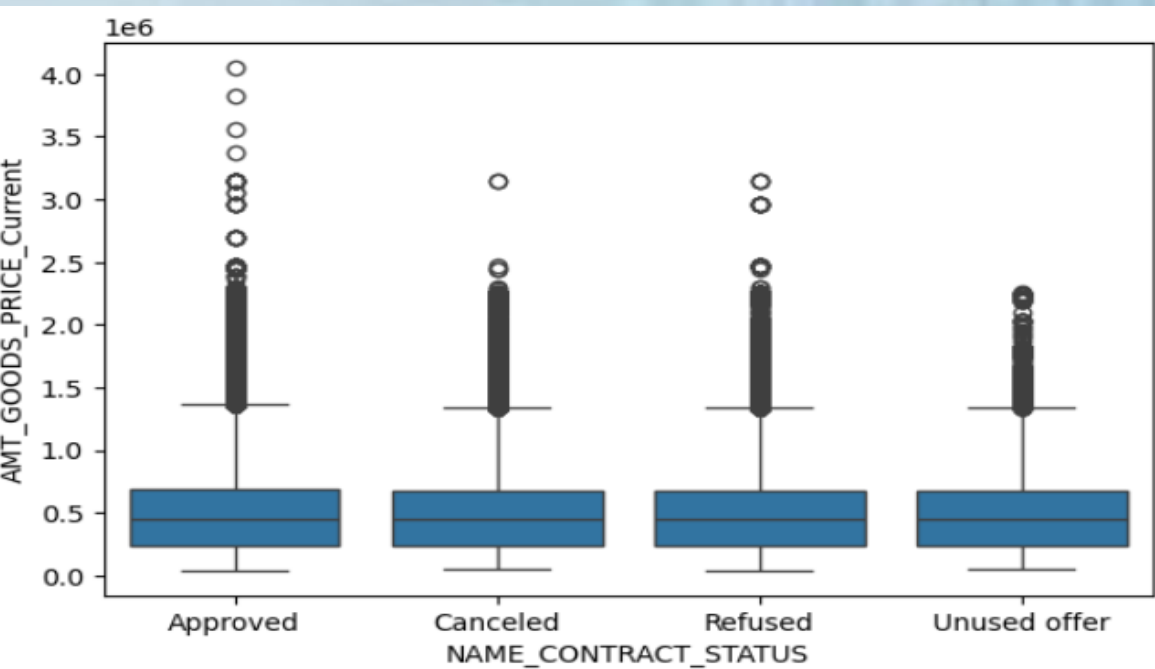
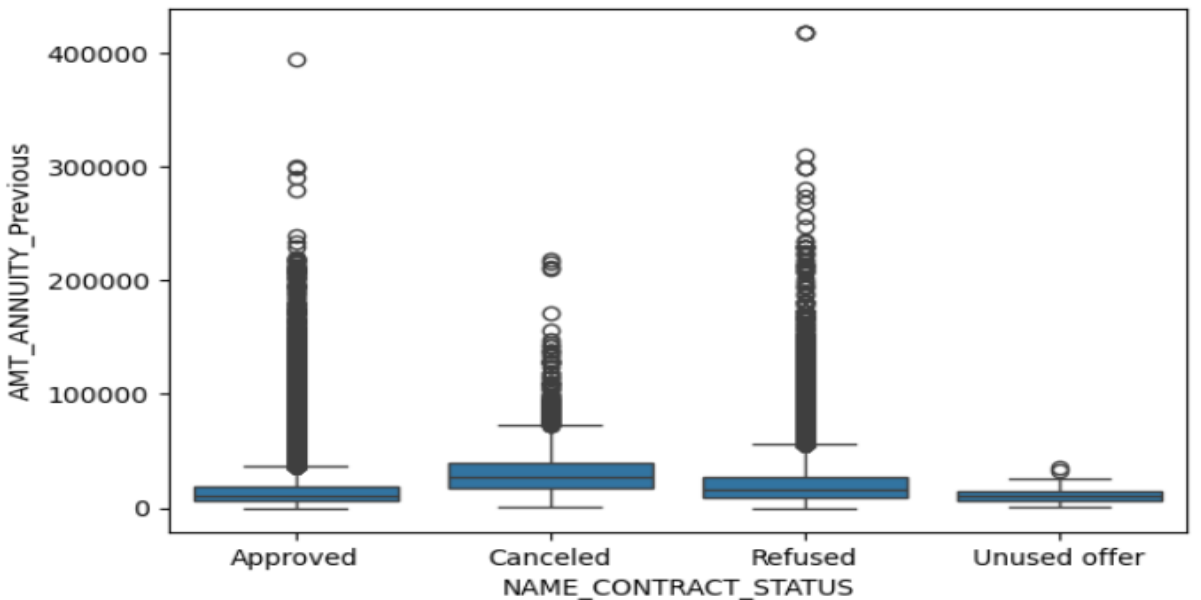
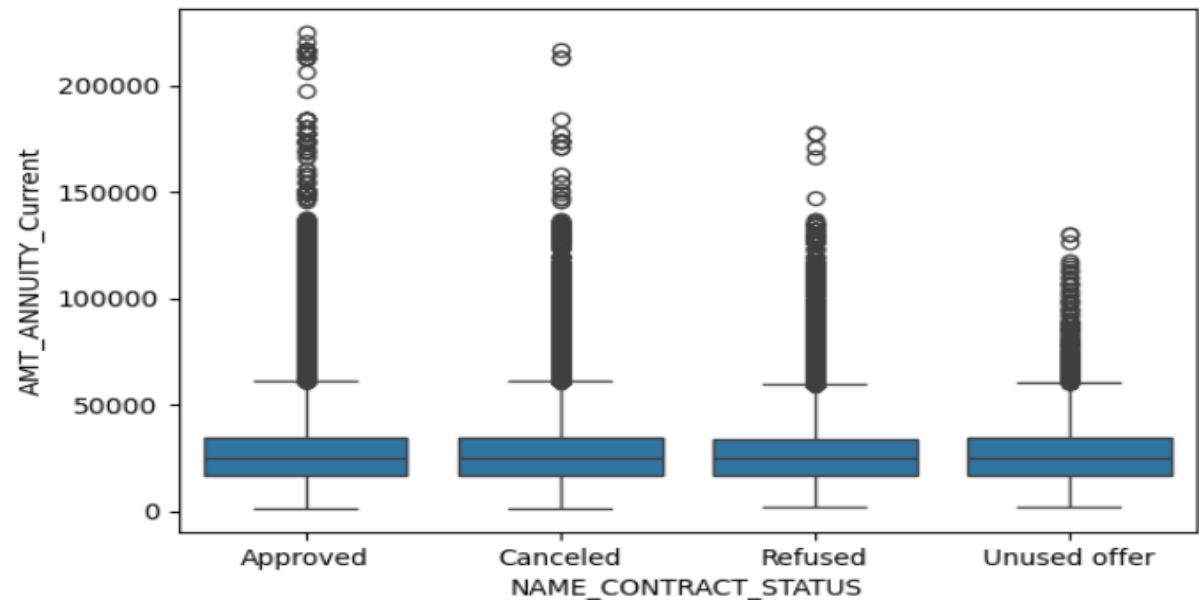


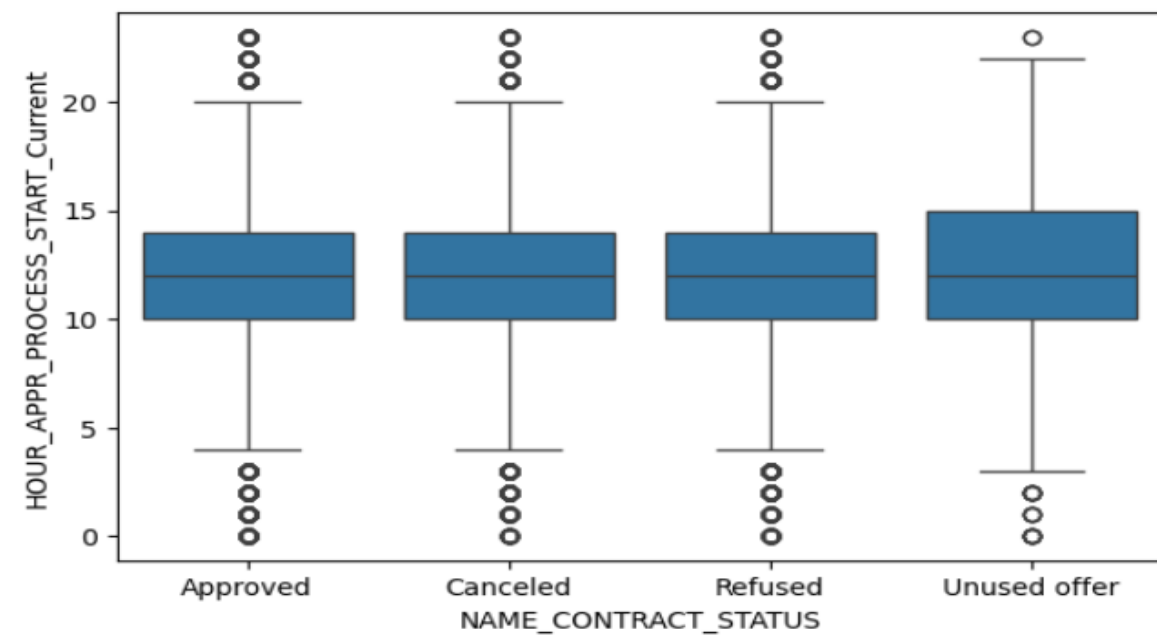
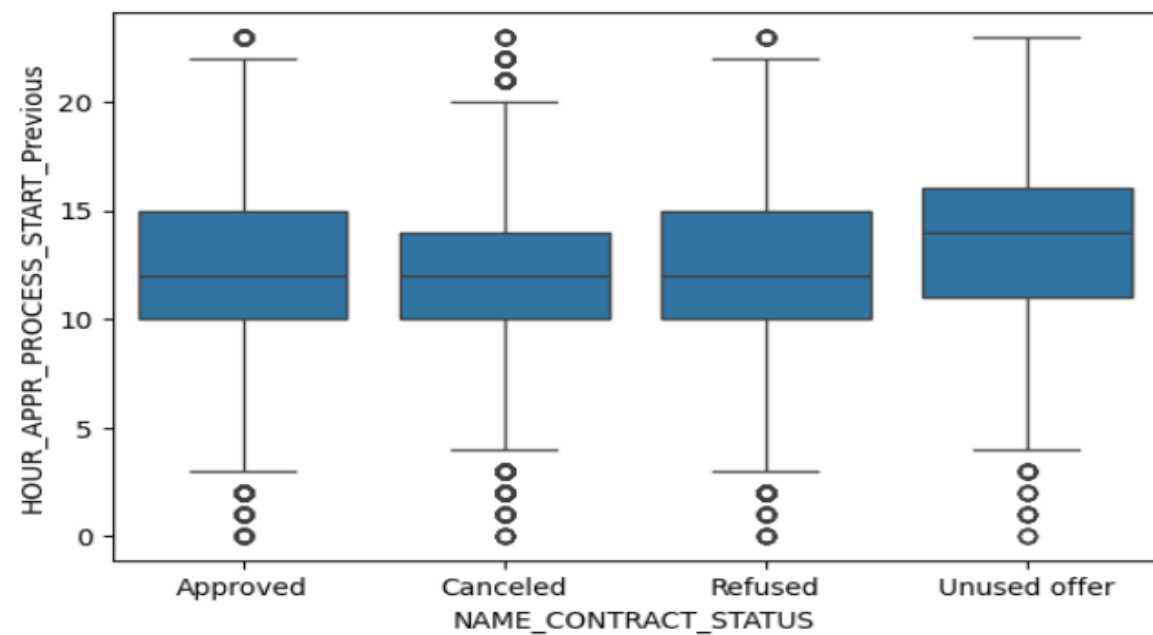
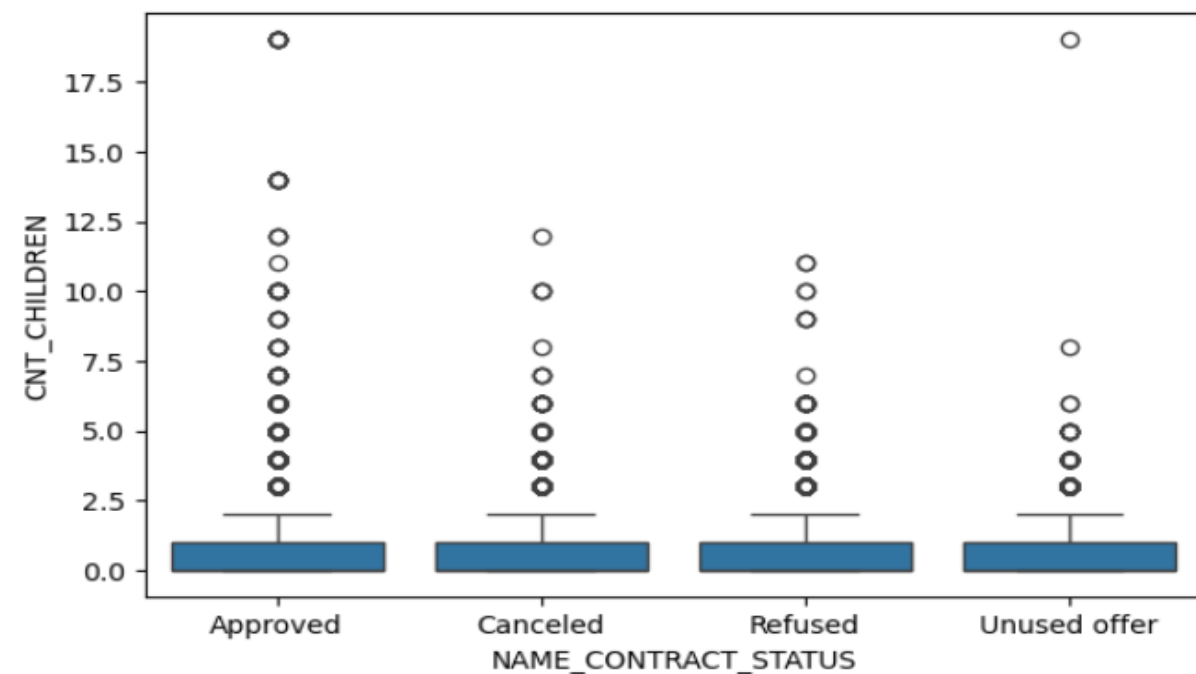
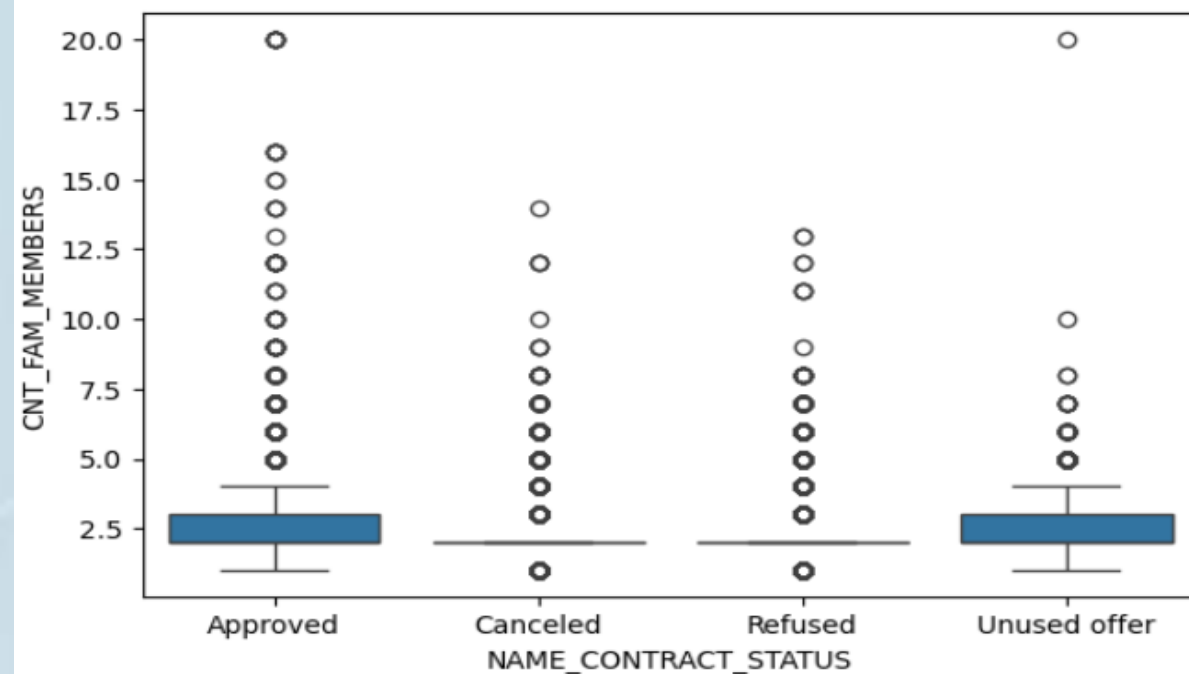
- Insight- Most Number Of Times Repeated Applications Got Approved

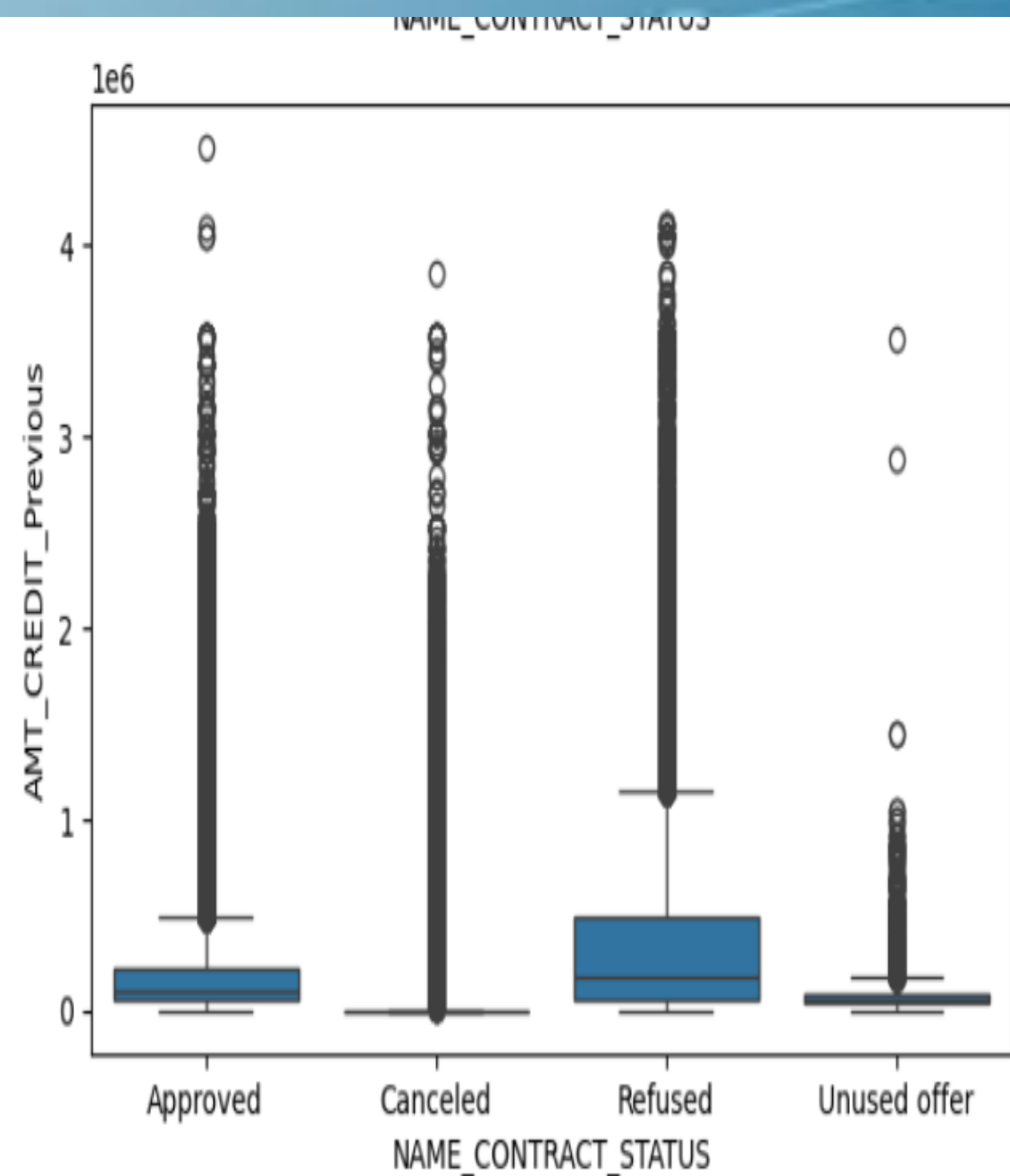
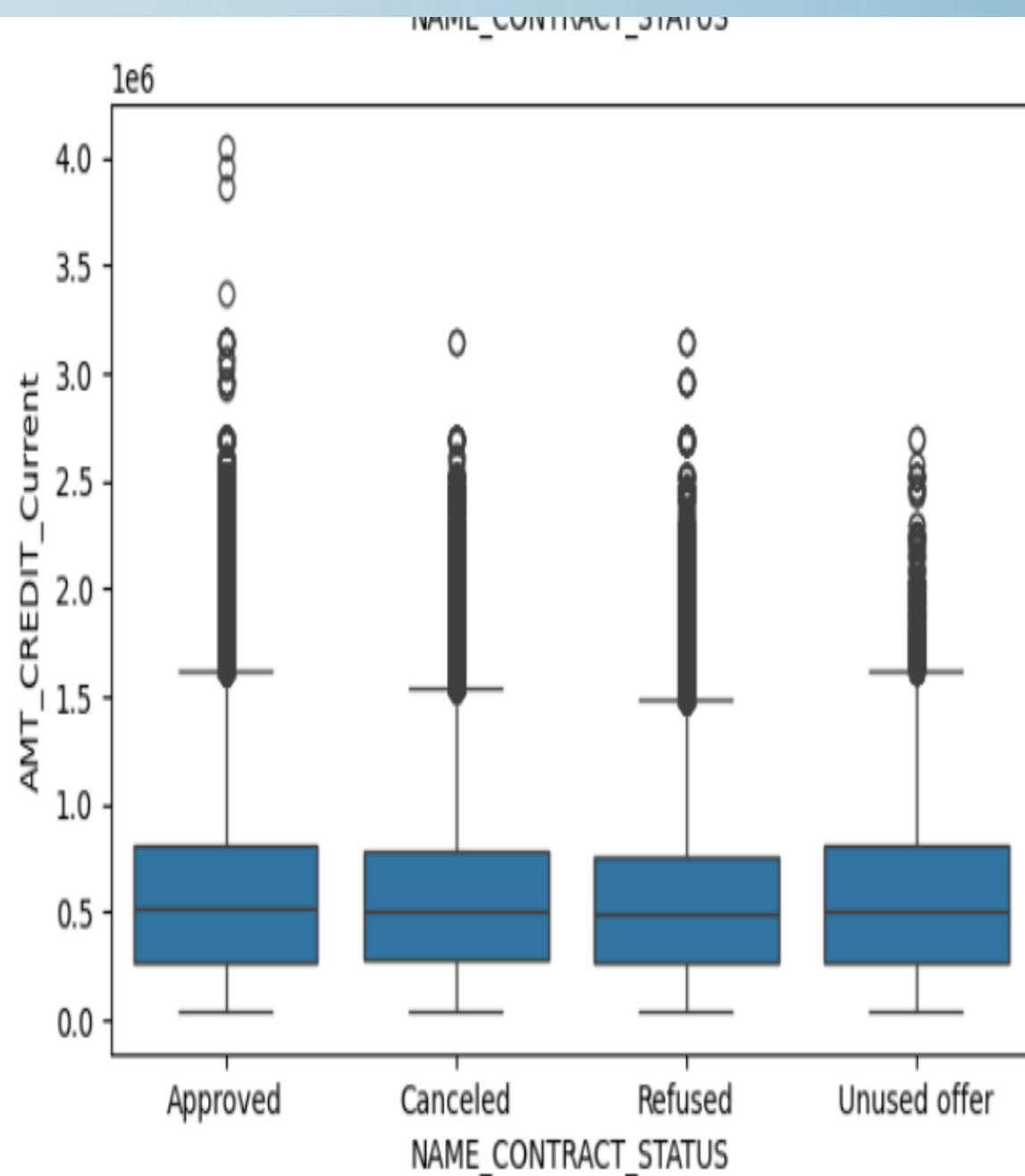


The background features a light blue gradient with several white line graphs and a bar chart. The line graphs have circular markers at data points, and one line ends in an arrow pointing towards the top right. The bar chart consists of numerous vertical bars of varying heights, creating a textured, data-like appearance.

Bi-variate categorical-Continuous analysis of merged data







Insights From Bi-variate –Continuous Of Merged Data-

- Amt_credit_current Is Similar For All 4 Cases & Amt_credit_previous Has Highest Refused Cases .
- Time Spent In Unused Offer Is Higher Than Categories So Bank Should Reduce Time Spent On Unused Offer.
- Highest Approval For Nuclear Family(2-3 People In Family) .
- Most Of The Applications Were Cancelled Or Refused Previously But Now All Four(Refused/Cancelled/Approved/Unused) Have Similar Situation For Amt_goods_price.
- Most Of The Applications Were Cancelled Or Refused Previously But Now All Four(refused/Cancelled/Approved/Unused) Have Similar Situation For Amt_annuity.

Conclusion

Target/ Focused Variable For Operation Dataset- TARGET

Target/ Focused Variable For Former Dataset- NAME_CONTRACT_STATUS

Top Major Variables The Company Should Consider Are:-

- Name_education_type
 - Amt_income_total
 - Days_birth
 - Amt_credit
 - Days_employed
 - Amt_annuity
 - Name_income_type
 - Code_gender
 - Name_housing_type
-
- The Company Should Consider The Above-mentioned Variables Before Approving Loans To Minimize The Threat Of Loss.
 - The Analysis Highlights Key Factors Influencing Loan Defaults. These Insights Can Help Financial Institutions Refine Risk Assessment Models And Optimize Loan Approval Processes.