



PRESENTATION ON CREDIT EDA ASSIGNMENT

:-AMAN TIKALE

PROBLEM STATEMENT

- This assignment aims to give you an idea of applying EDA in a real business scenario.
- In this assignment, apart from applying the techniques that you have learnt in the eda module.
- And to also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

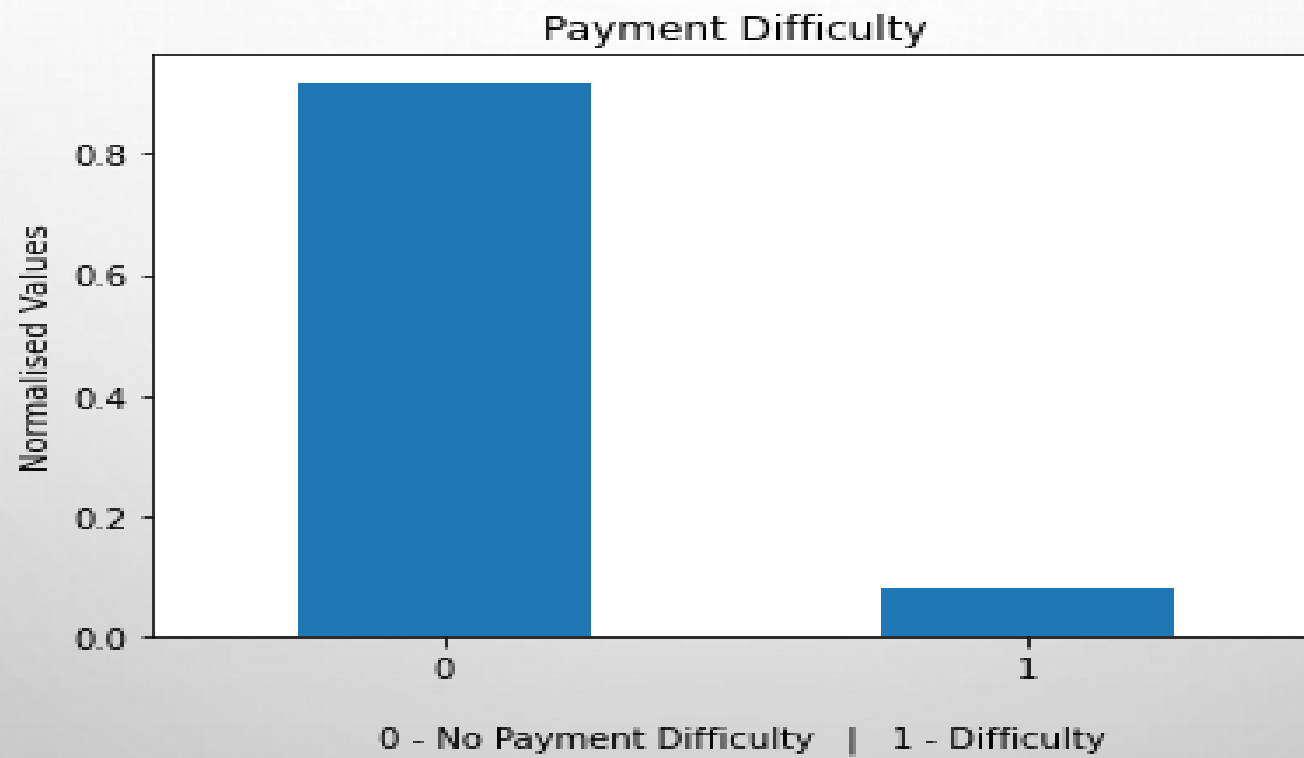
PURPOSE

- THIS CASE STUDY AIMS TO IDENTIFY PATTERNS WHICH INDICATE IF A CLIENT HAS DIFFICULTY PAYING THEIR INSTALMENTS 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.

OVERALL APPROACH

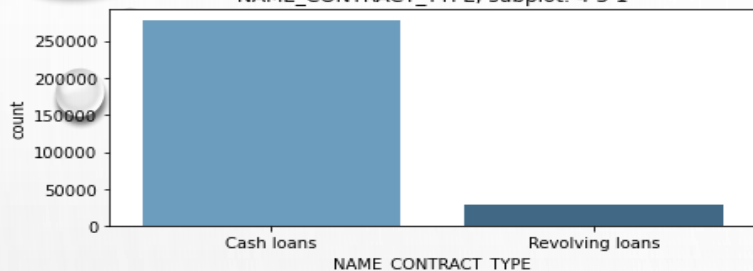
- LOADING AND INSPECTING DATA
- DATA UNDERSTANDING AND SOURCING
- DATA CLEANING AND IMPUTATION
- CHECKING DATA QUALITY AND BINNING
- CHECKING DATA IMBALANCE, UNIVARIATE, SEGMENTED UNIVARIATE ANALYSIS
- BIVARIATE ANALYSIS AND CORRELATION
- MERGING TWO DATA FRAMES APPLICATION DATA AND PREVIOUS APPLICATION
- INTERFERENCE/INSIGHTS RECOMMENDATIONS AND RISKS

UNIVARIATE ANALYSIS



OBJECT TYPE VARIABLES AND THEIR VALUES

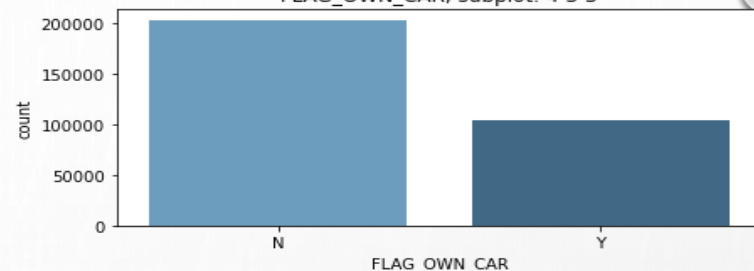
NAME_CONTRACT_TYPE, subplot: 4 3 1



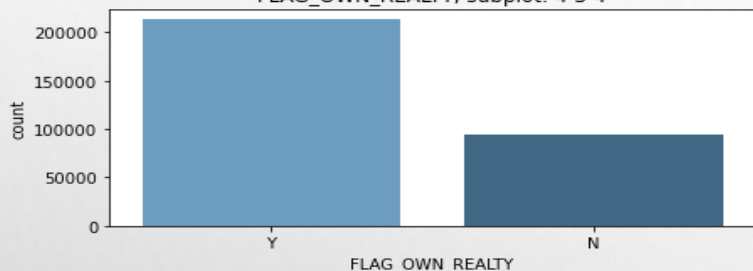
CODE_GENDER, subplot: 4 3 2



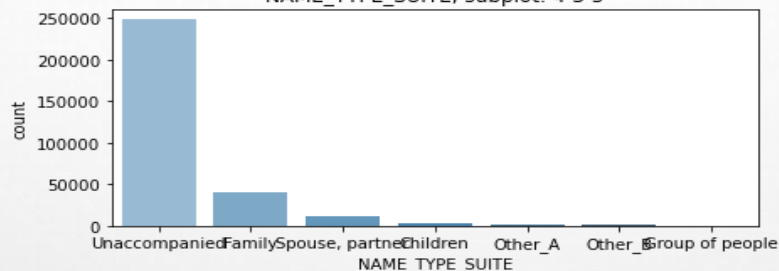
FLAG_OWN_CAR, subplot: 4 3 3



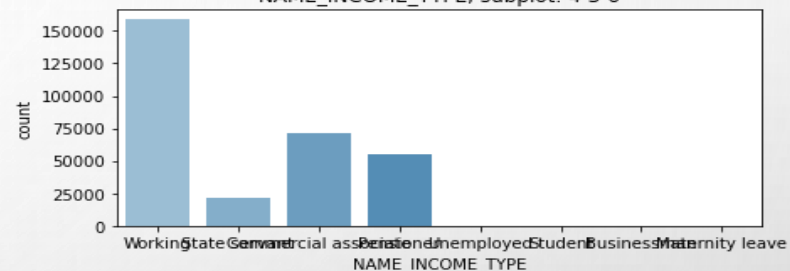
FLAG_OWN_REALTY, subplot: 4 3 4



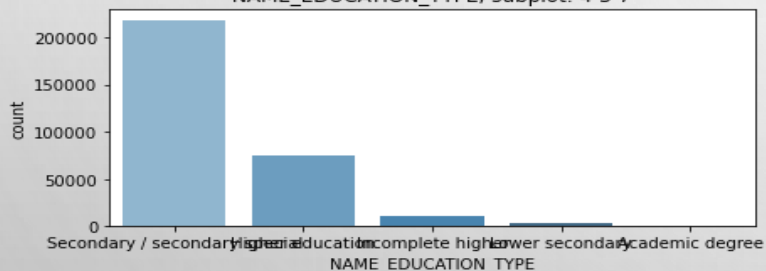
NAME_TYPE_SUITE, subplot: 4 3 5



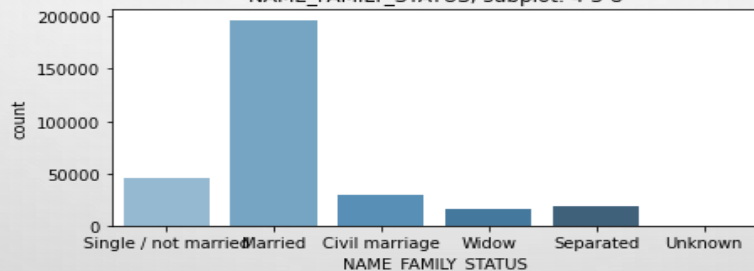
NAME_INCOME_TYPE, subplot: 4 3 6



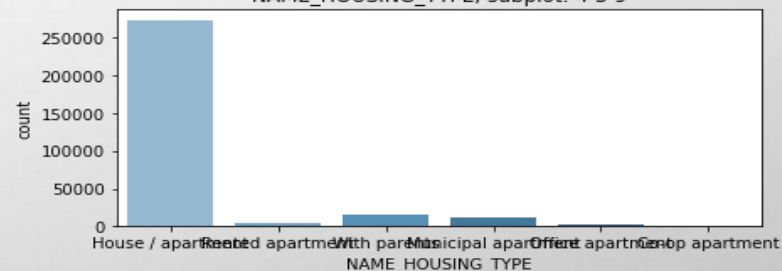
NAME_EDUCATION_TYPE, subplot: 4 3 7



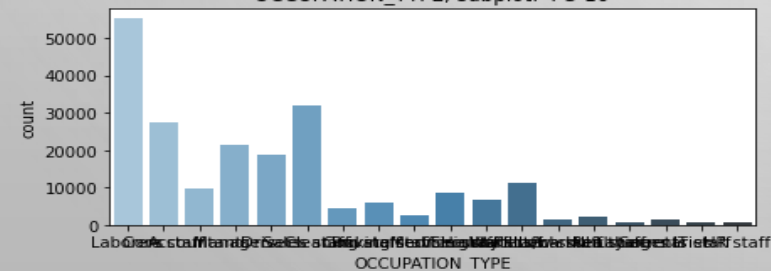
NAME_FAMILY_STATUS, subplot: 4 3 8



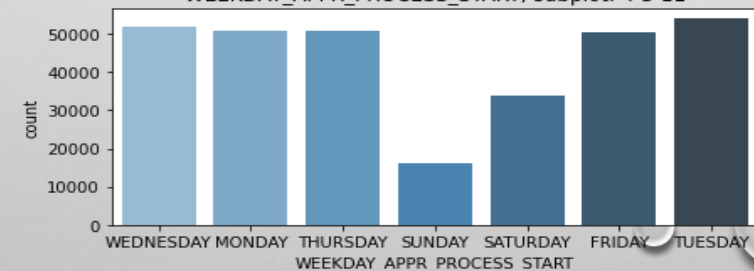
NAME_HOUSING_TYPE, subplot: 4 3 9



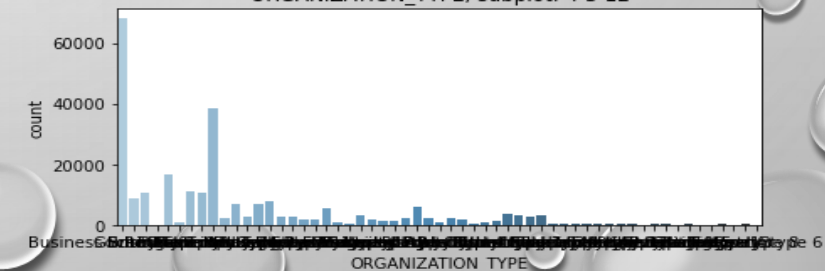
OCCUPATION_TYPE, subplot: 4 3 10



WEEKDAY_APPR_PROCESS_START, subplot: 4 3 11

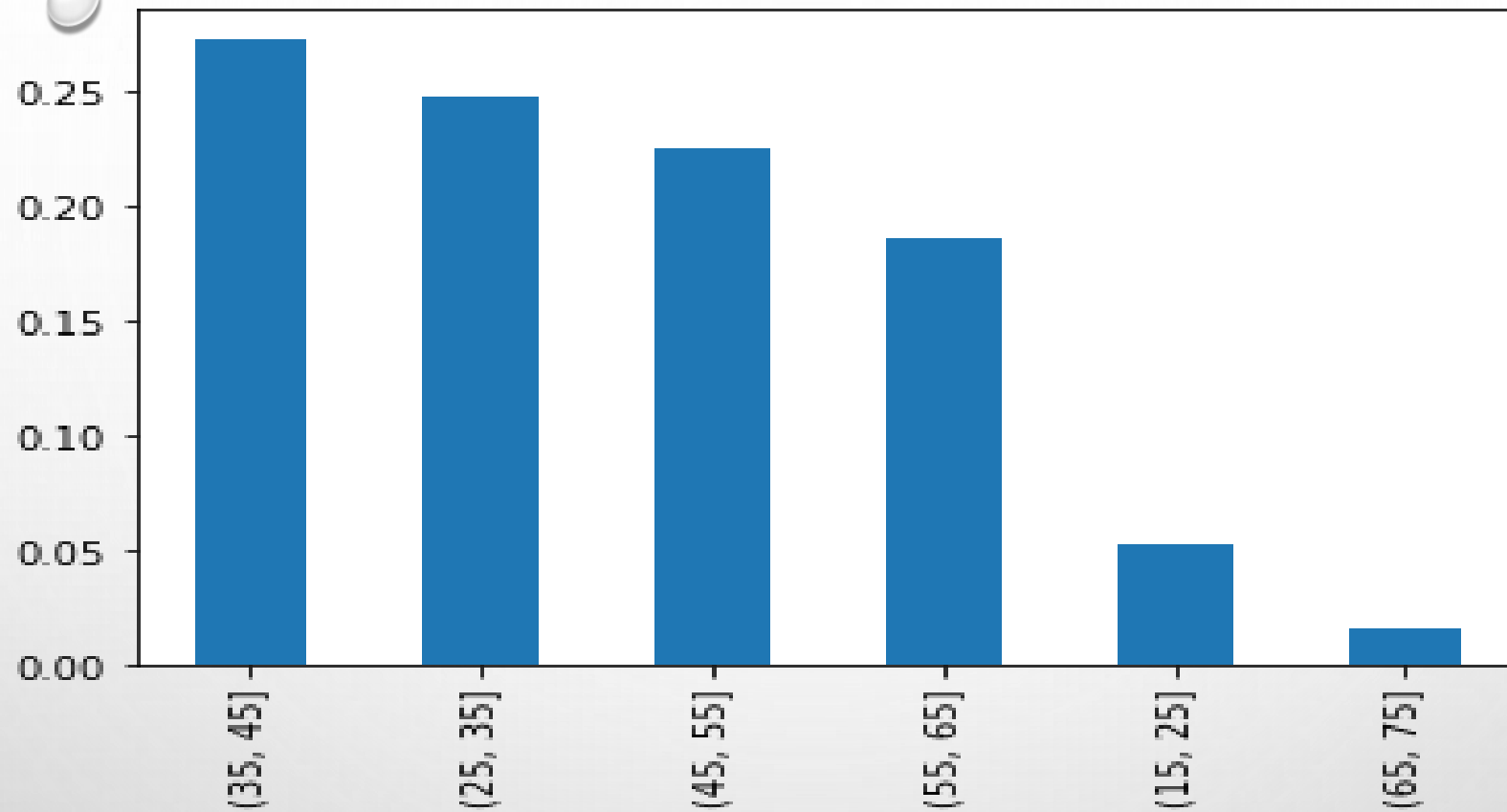


ORGANIZATION_TYPE, subplot: 4 3 12

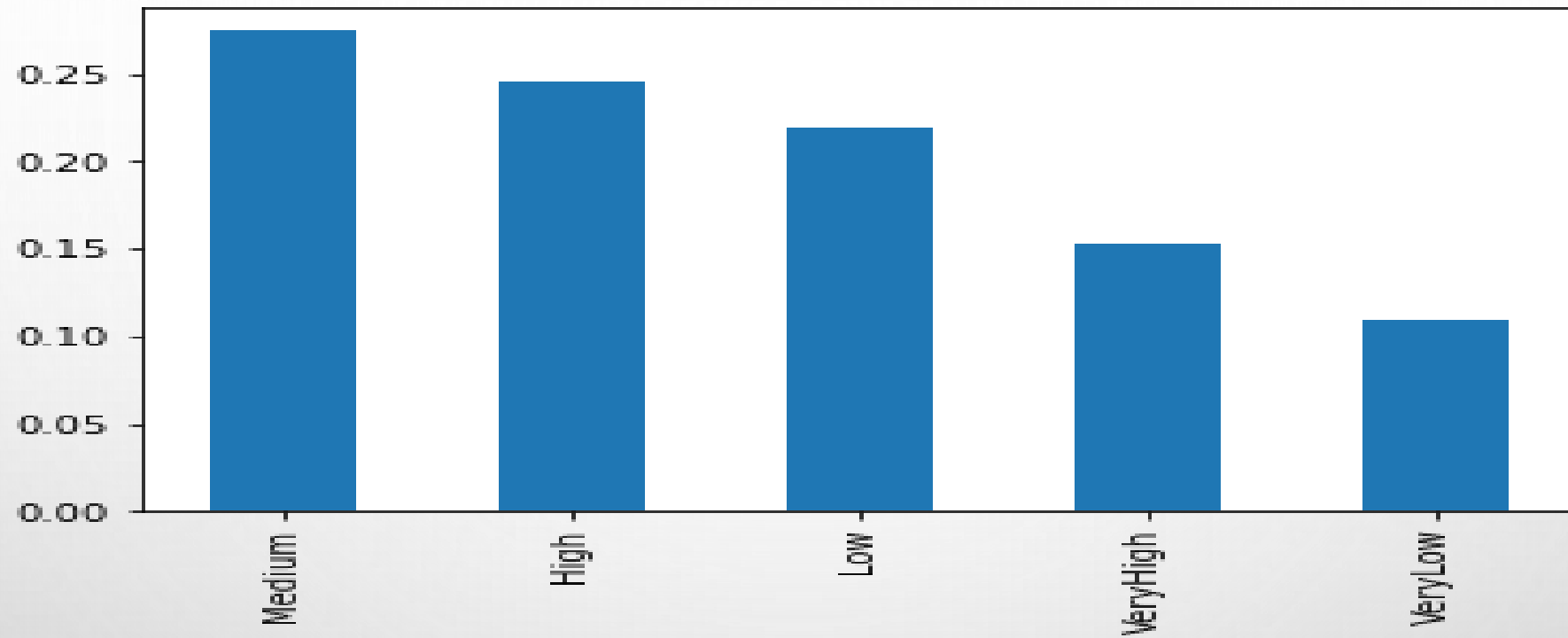


NOTABLE POINTS

- CASH LOANS OFFERED ARE MORE THAN REVOLVING LOANS, AT 90%.
- 65% FEMALES HAVE TAKEN LOANS IN COMPARISON TO 34% MALE. THIS IS VERY INTERESTING AND NEEDS TO BE STUDIED FURTHER.
- 65% APPLICANT DON'T OWN CARS.
- 69% APPLICANTS OWN LIVING QUARTERS.
- 81% APPLICANTS CAME ACCOMPANIED FOR LOAN APPLICATION WHILE MOST APPLICANTS ARE WORKING CLASS, 18% ARE PENSIONERS.
- 71% HAVE SECONDARY EDUCATION.
- 63% ARE MARRIED.
- 31% HAVE NOT MENTIONED THEIR OCCUPATION TYPE.

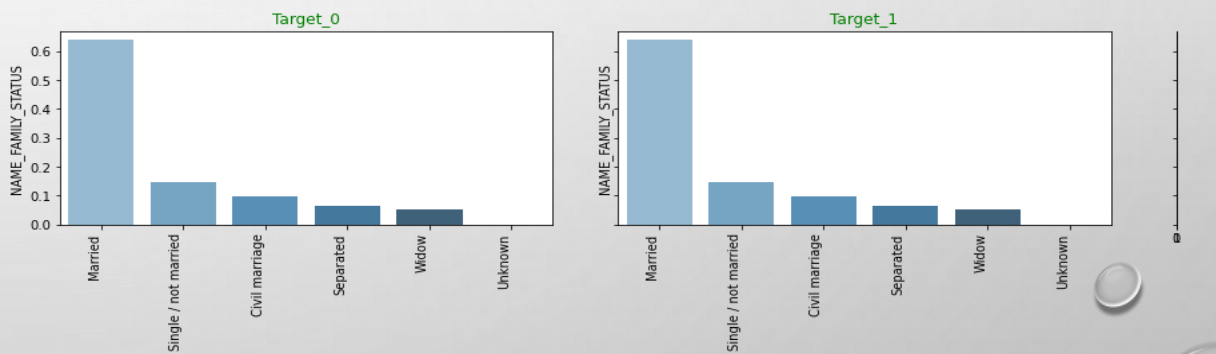
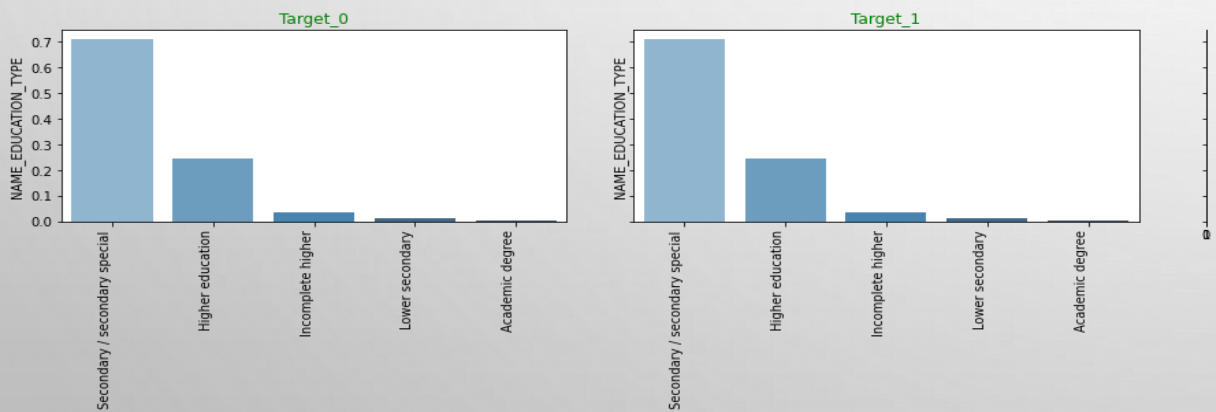
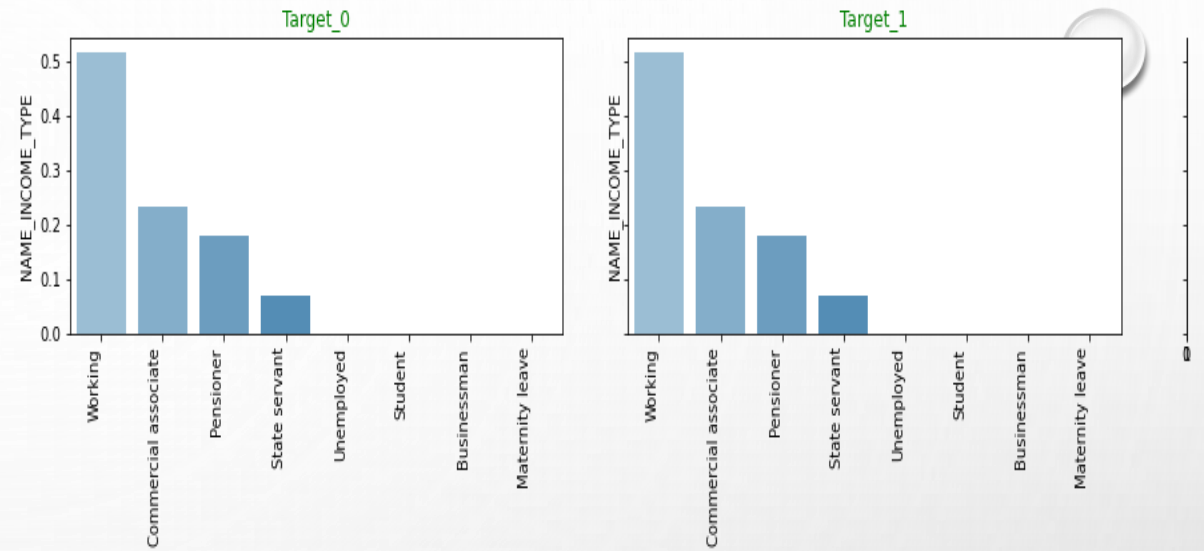
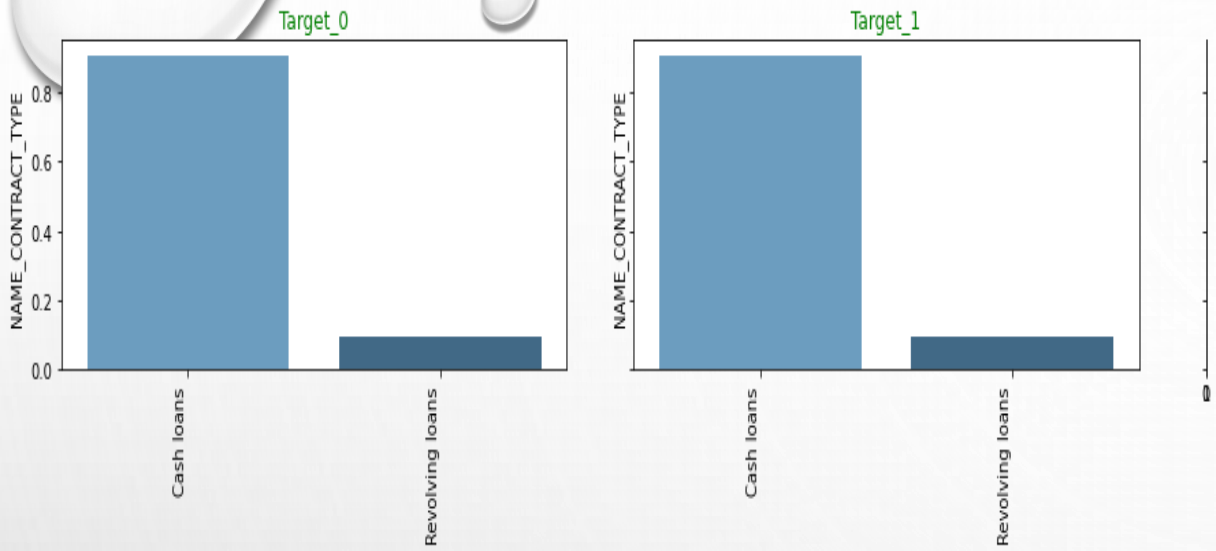


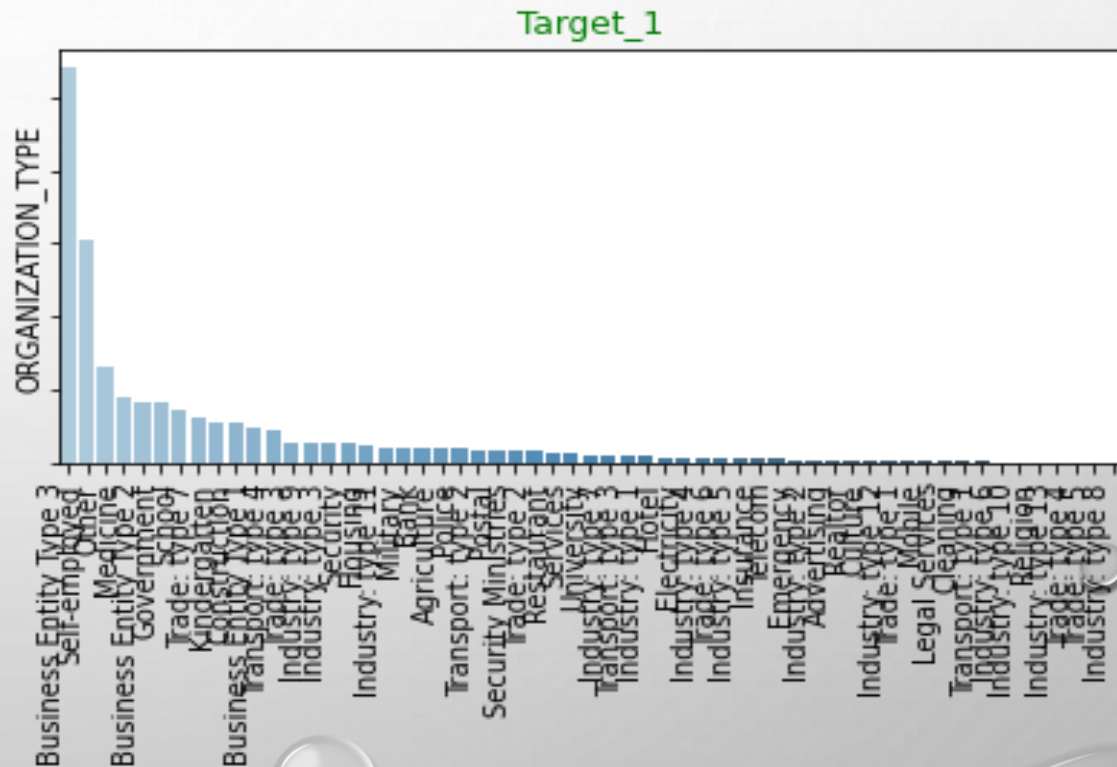
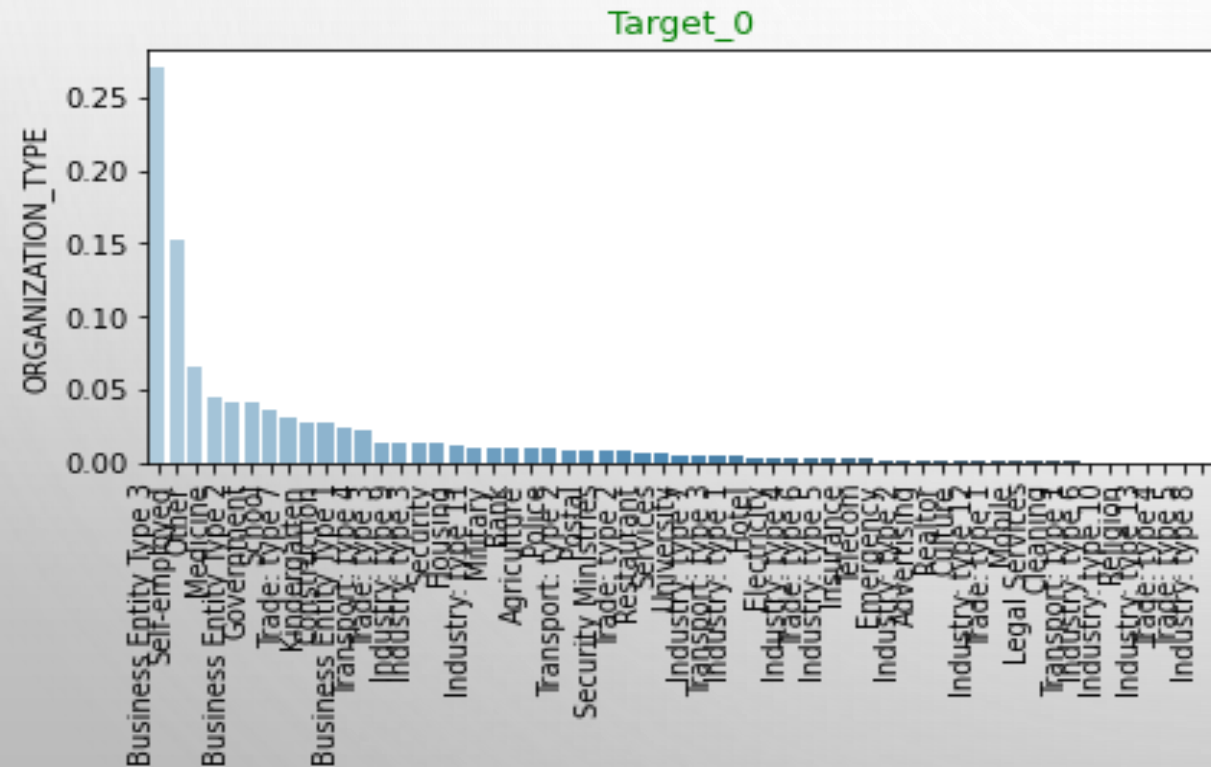
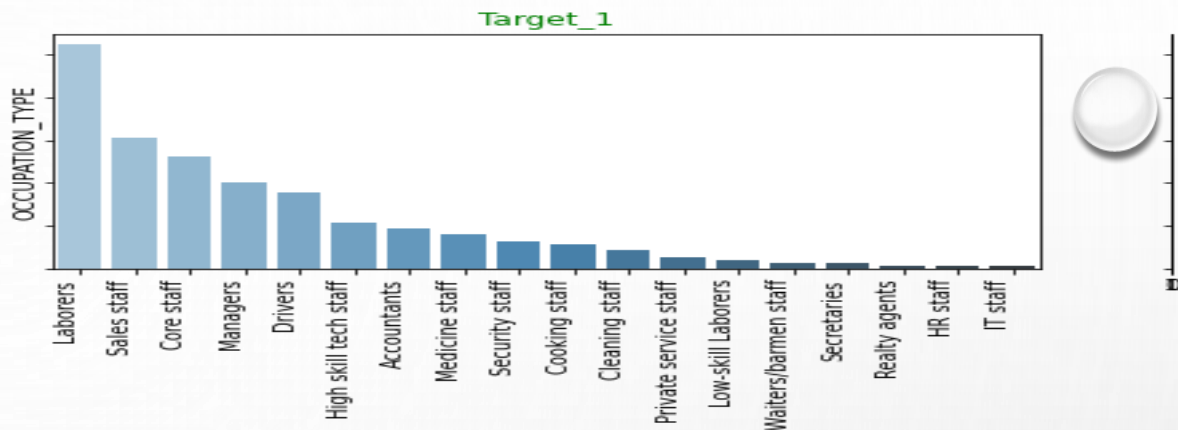
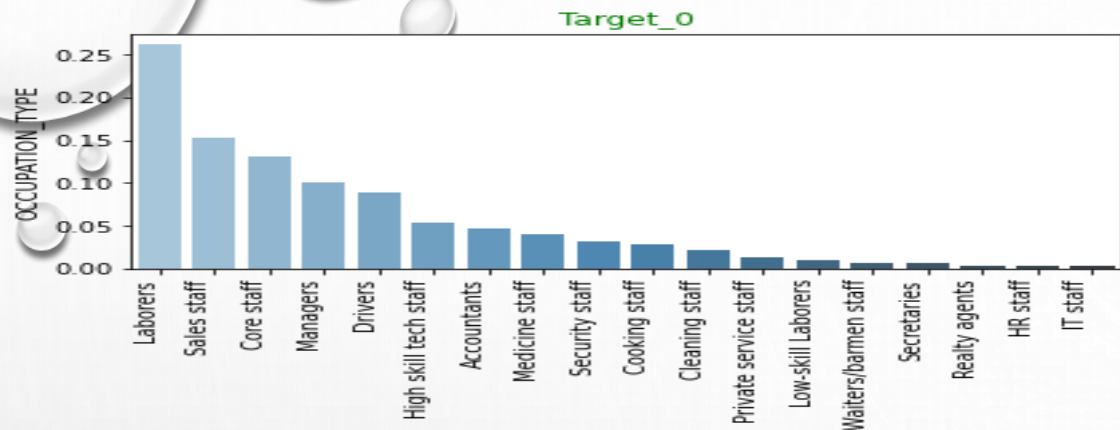
*35-45 AGE GROUP PEOPLE ARE THE LARGEST GROUP OF PEOPLE APPLYING FOR LOANS.



* INCOME GROUP HAVING MEDIUM INCOME IS LARGEST CONSUMER APPLYING FOR LOANS.

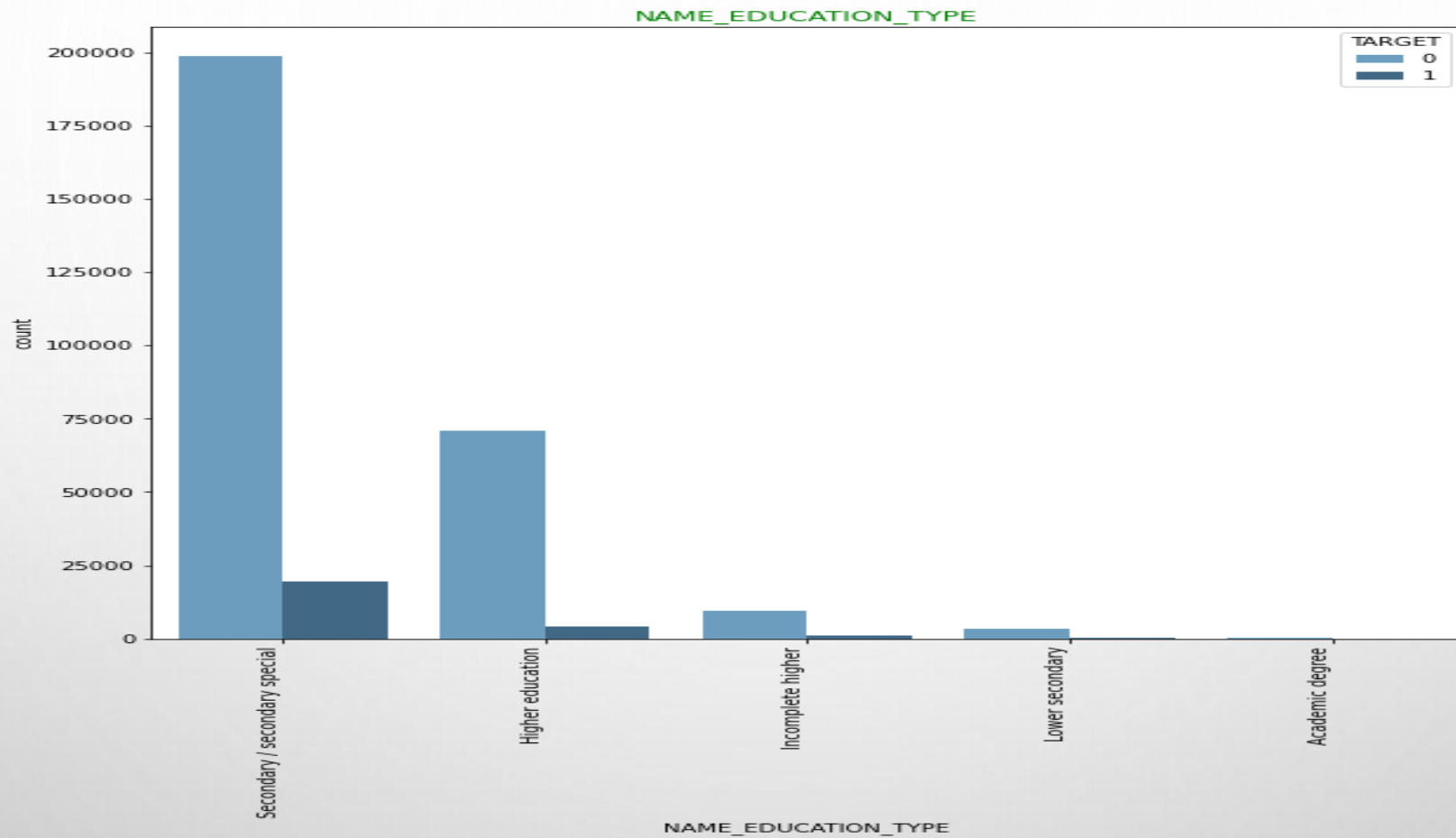
UNIVARIATE CATEGORICAL NOMINAL ANALYSIS ON BOTH DATA FRAMES





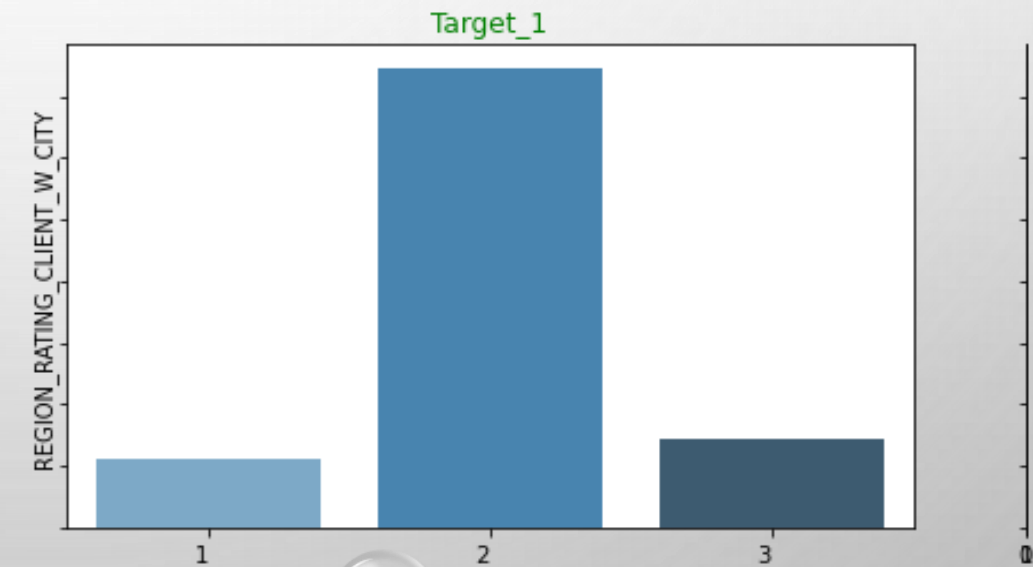
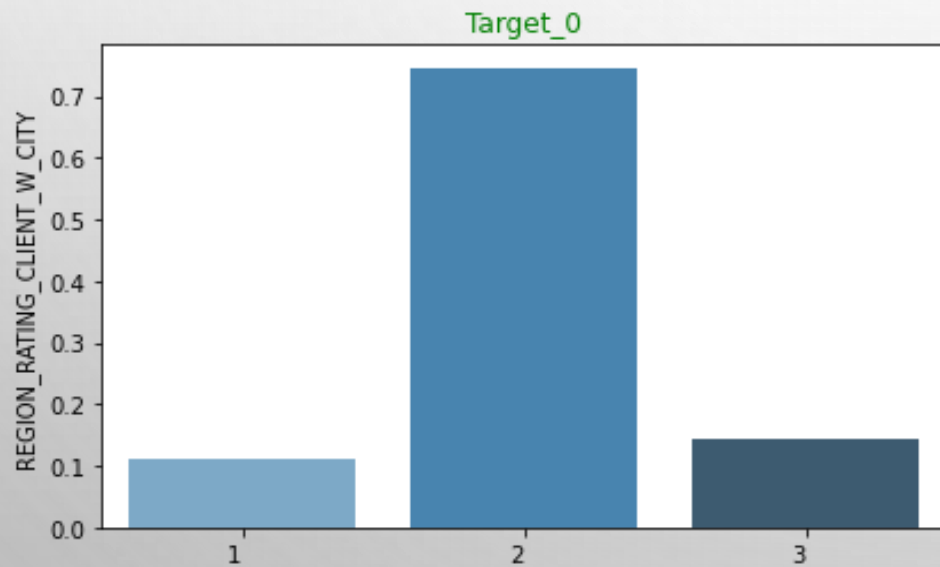
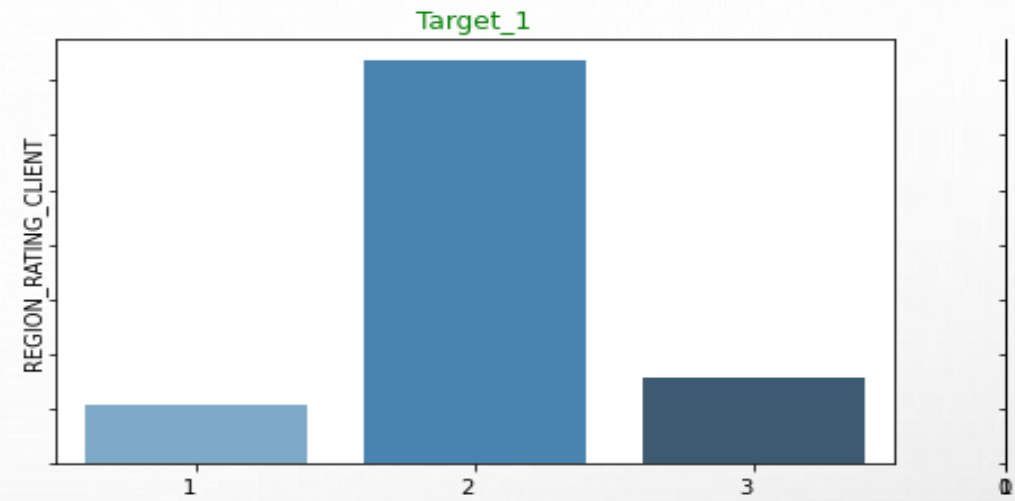
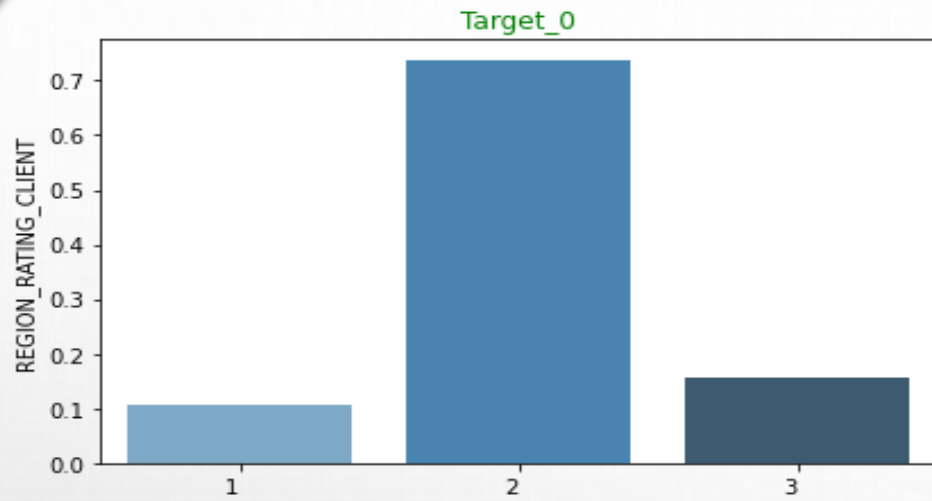
NOTABLE POINTS

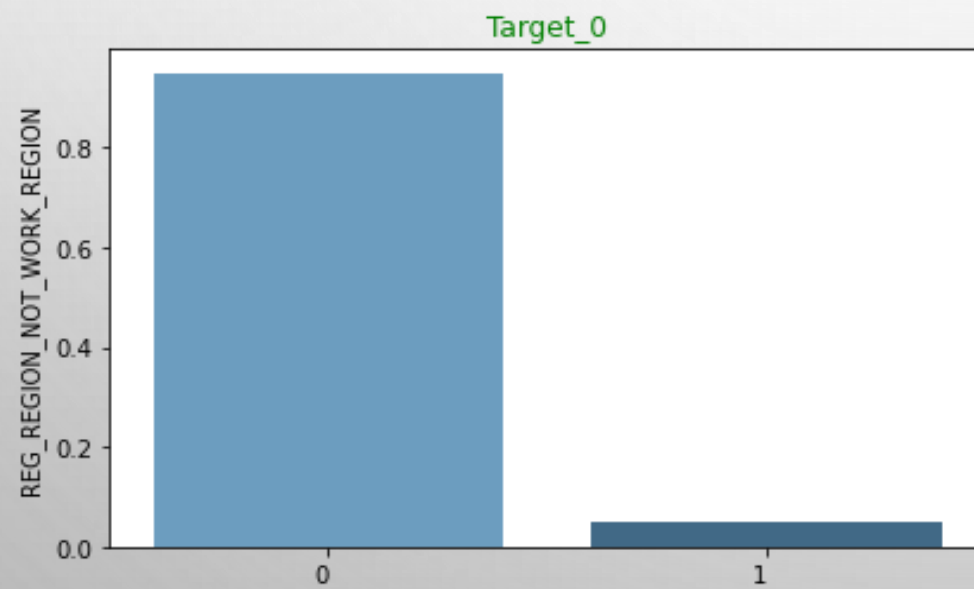
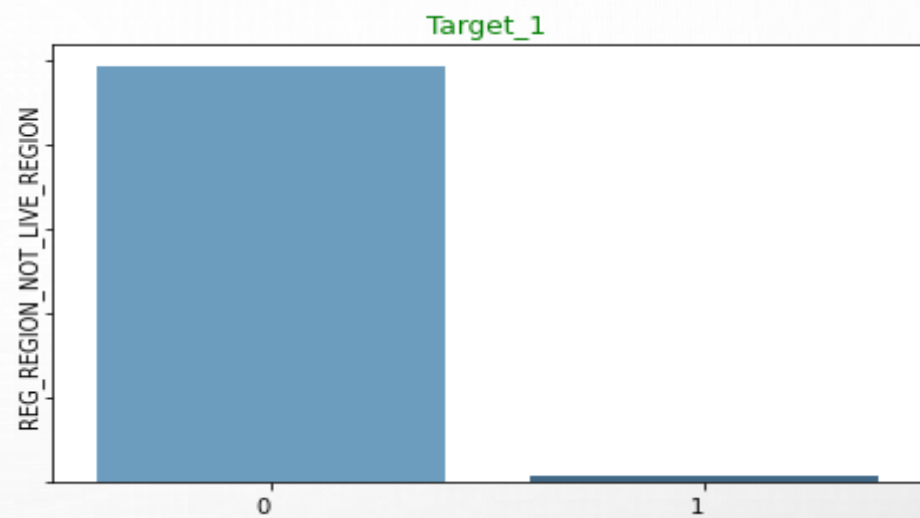
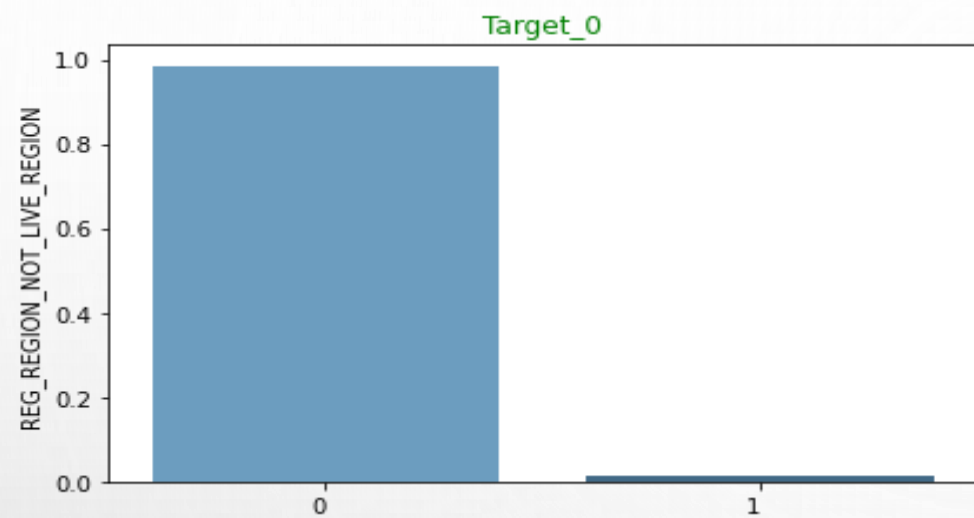
- NAME_CONTRACT TYPE- CASH LOANS ARE LARGE PART OF THE COMPANY'S PORTFOLIO. FOR TARGET 0 - 85% AND ALMOST 95% FOR TARGET-1
- NAME_TYPE_SUITE - 80-90% IN TARGET 0 AND TARGET 1 ARE APPLYING FOR LOAN UNACCOMPANIED. INDICATING, THIS IS NOT A PARAMETER THAT CAN INFLUENCE PAYMENT DEFAULT.
- NAME_INCOME_TYPE - 50% WORKING IN CASE OF TARGET 0 AND 60% IN CASE OF TARGET 1 ARE WORKING INCOME TYPES.
- NAME_EDUCATION_TYPE - IN BOTH TARGET 0 AND 1, APPLICANTS WITH SECONDARY EDUCATION HAS APPLIED FOR LOANS MORE THAN OTHERS. 90% OF DEFAULTING PAYMENTS ARE FROM APPLICANTS WITH SECONDARY INCOME. NEEDS FURTHER ANALYSIS.
- NAME_FAMILY_STATUS - MARRIED APPLICANTS - ALMOST 60% HAVE DEFAULTED ON PAYMENTS.
- NAME_HOUSING_TYPE - 85-90% IN TARGET 0 AND TARGET 1 APPLICANTS ARE STAYING IN "HOUSE/APARTMENT". INDICATING, THIS IS NOT A PARAMETER THAT CAN INFLUENCE PAYMENT DEFAULT.
- OCCUPATION_TYPE - LABOURERS, SALES STAFF, CORE STAFF, DRIVERS CONSTITUTE OF 50% OF DEFAULTERS. LABOURERS IS THE HIGHEST PERCENTAGE OF APPLICANTS TOO.
- ORGANIZATION_TYPE - BUSINESS ENTITY TYPE 3 AND SELF EMPLOYED ADD UPTO 40% DEFAULTERS. THE HIGHEST % OF LOAN TAKERS ARE ALSO THIS CATEGORY.

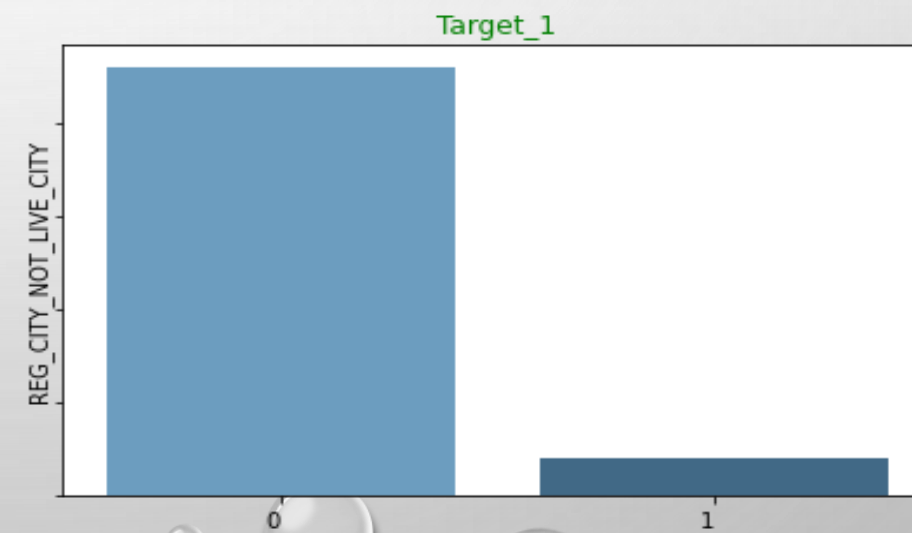
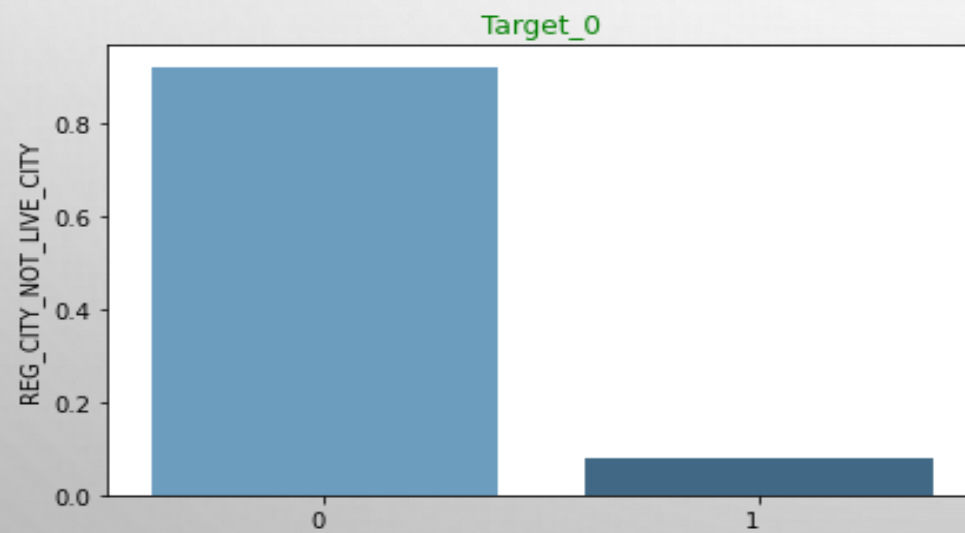
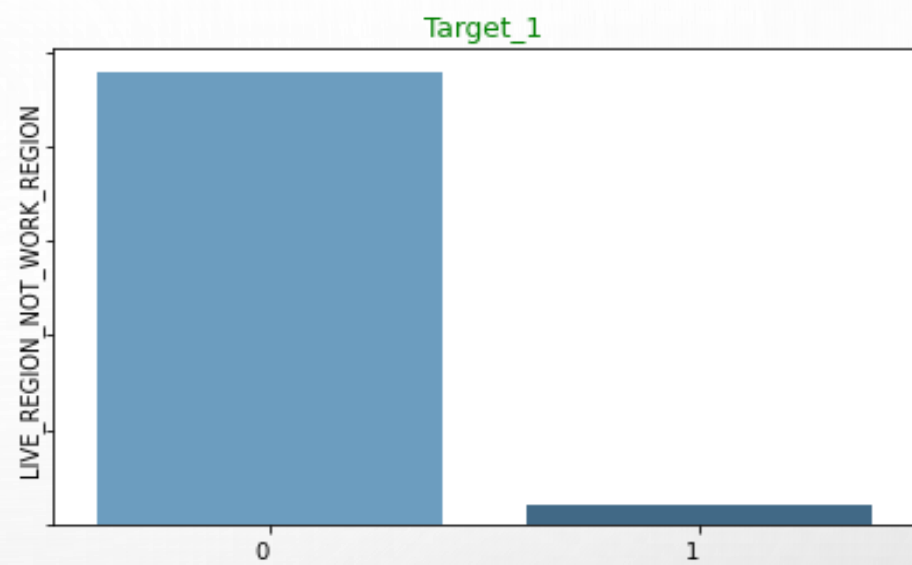
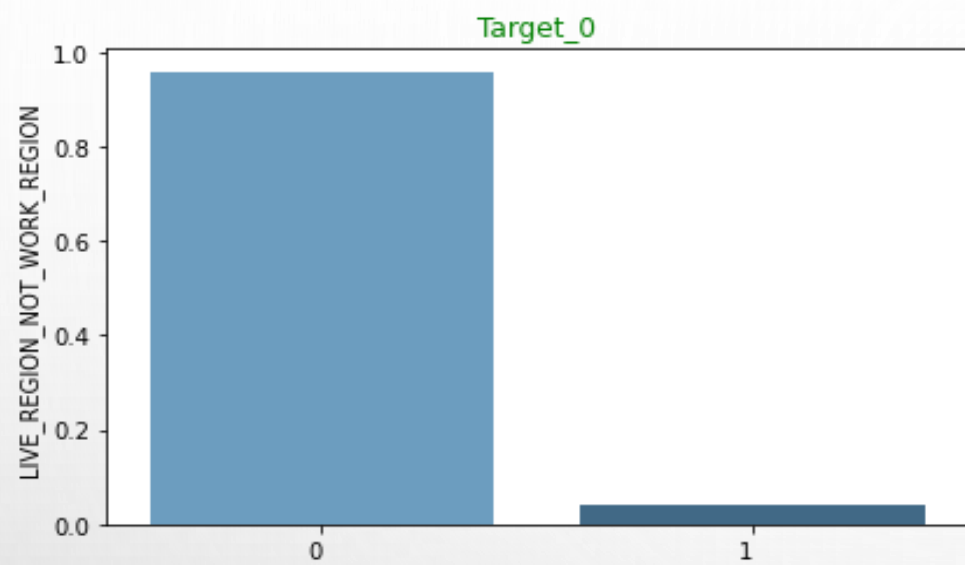


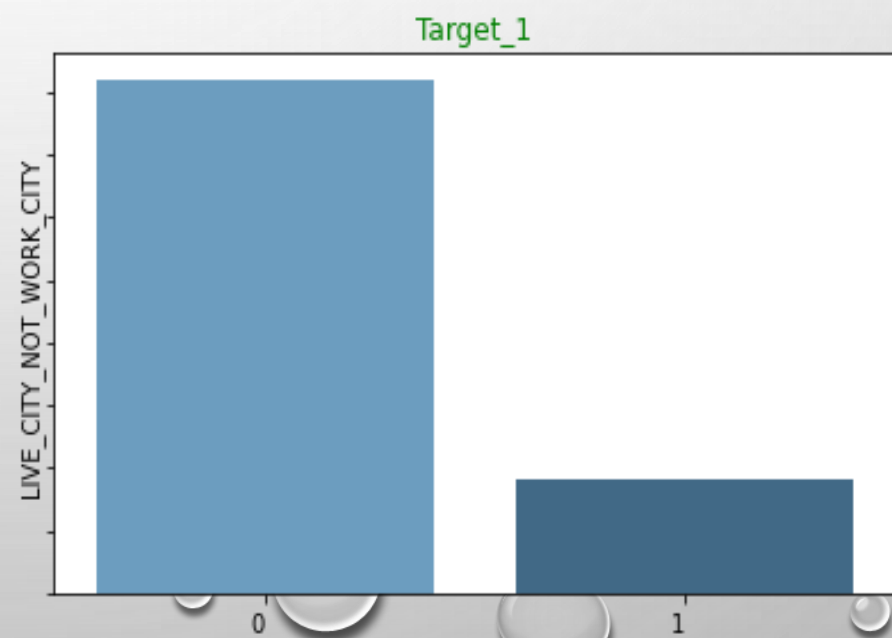
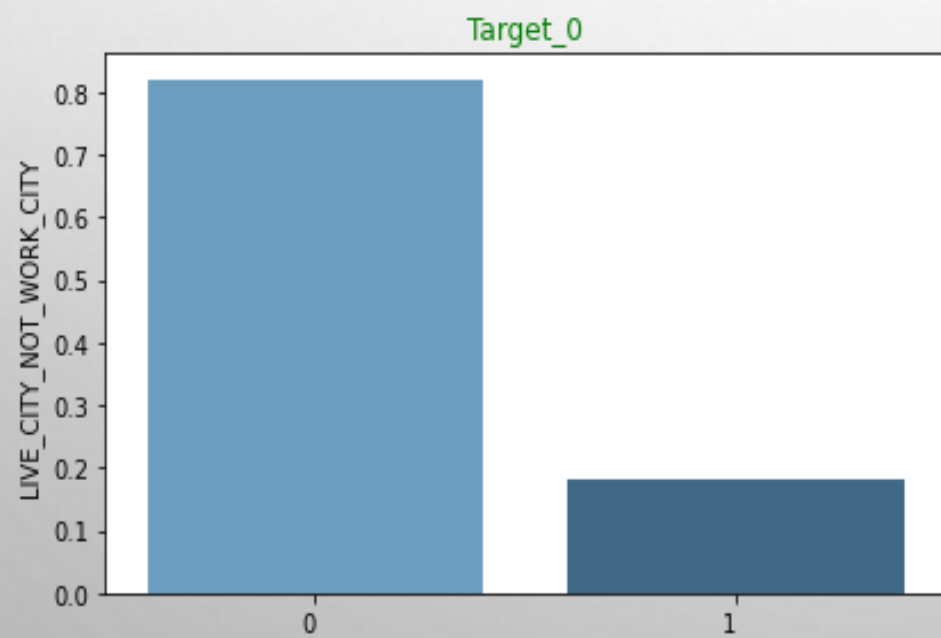
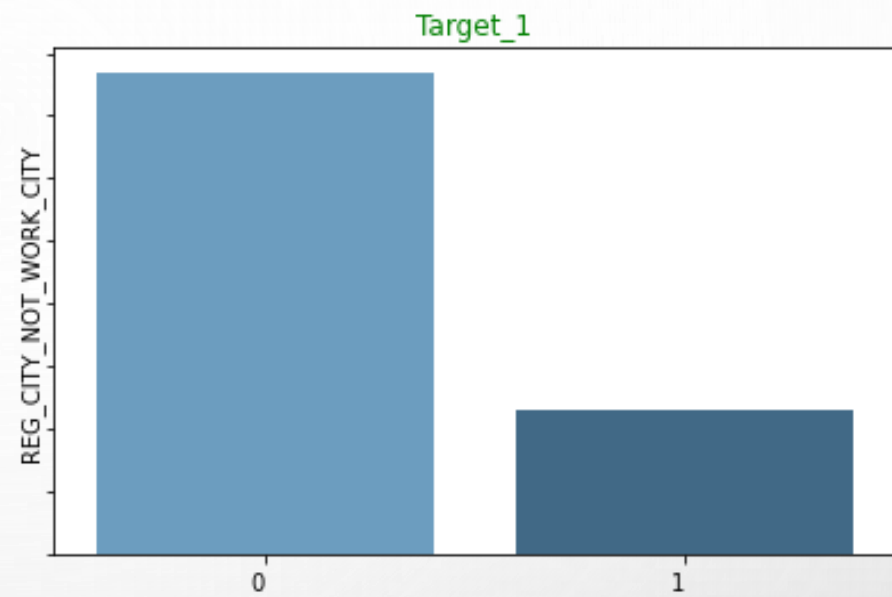
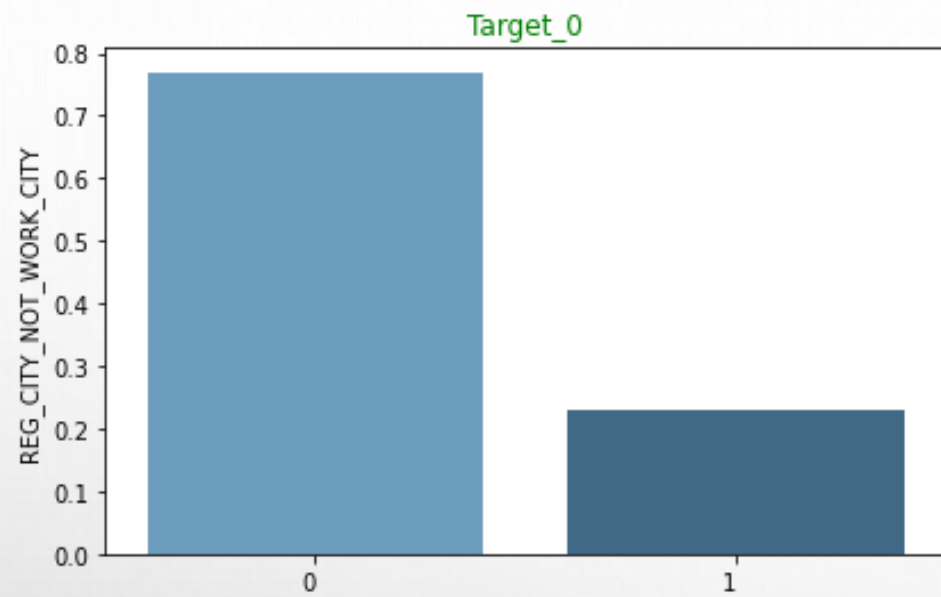
*UNBALANCED DATA, DIFFICULT FOR ANALYSIS AS SEEN ABOVE.

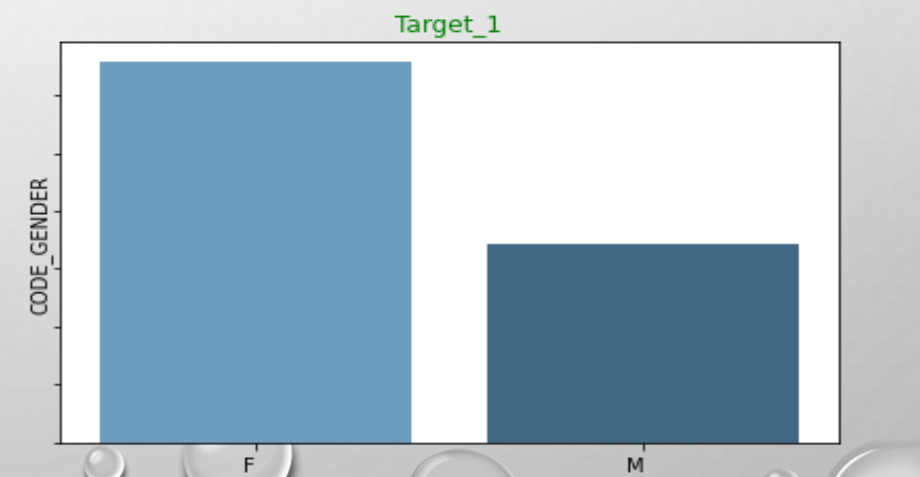
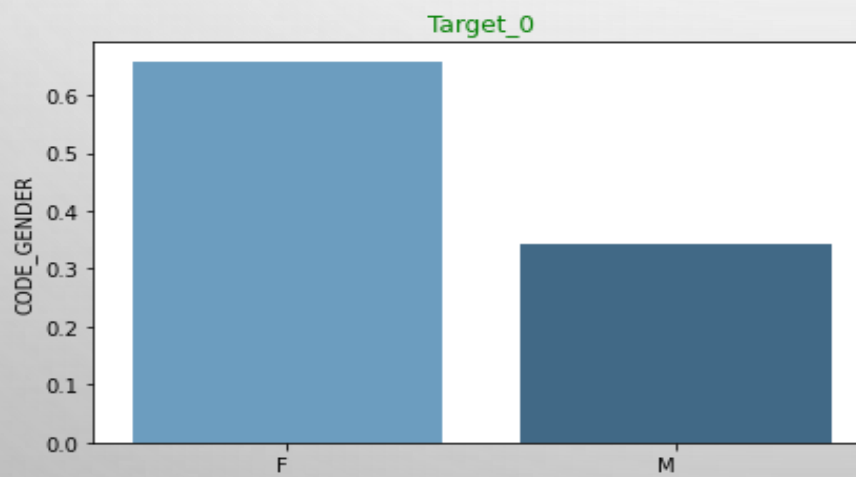
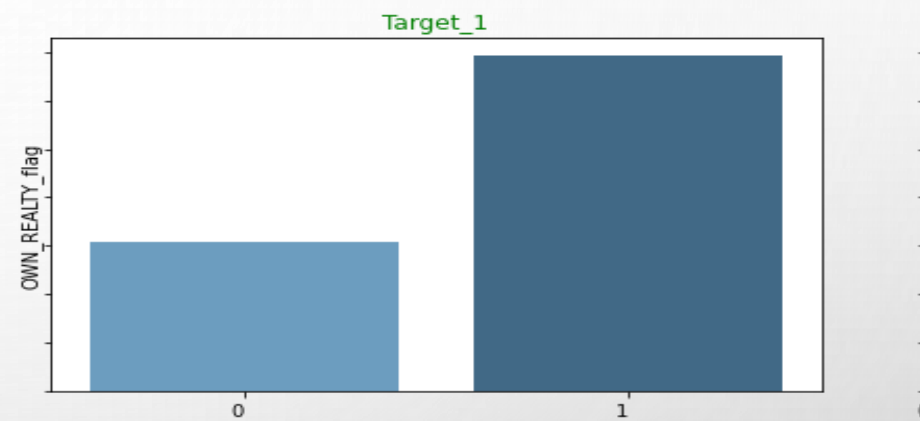
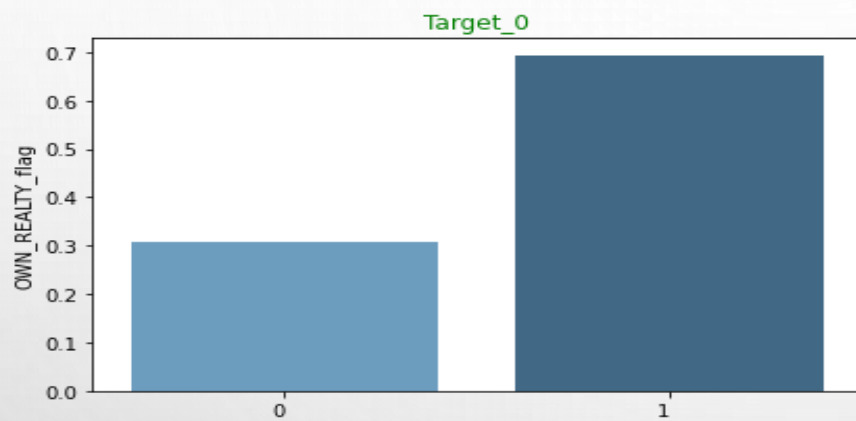
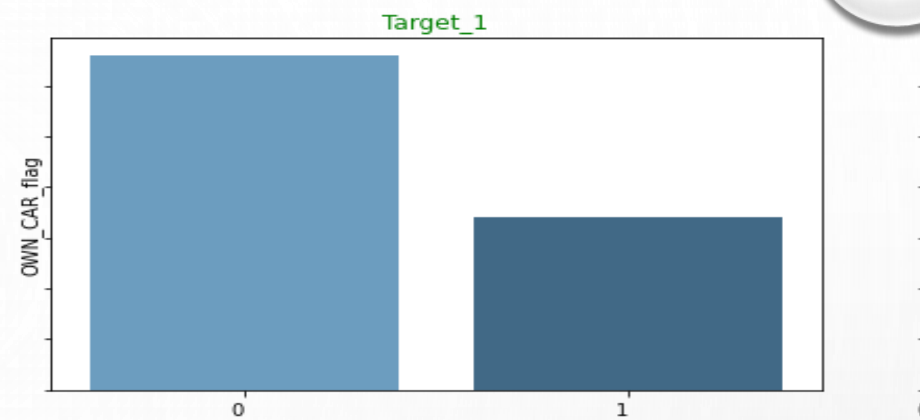
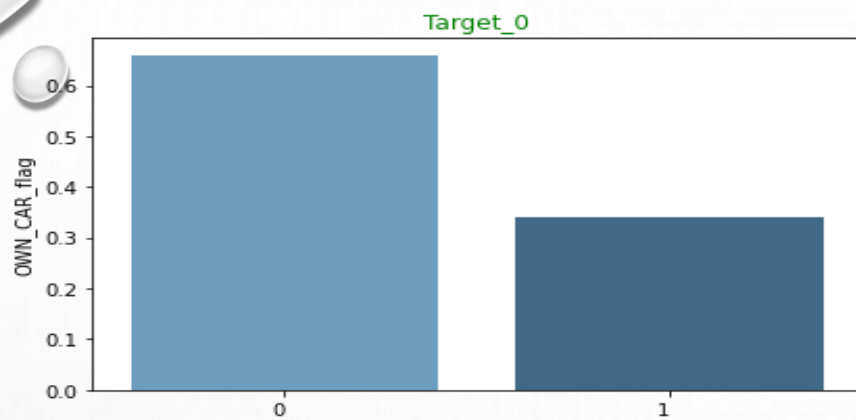
UNIVARIATE ANALYSIS ON CATEGORICAL ORDERED











NOTABLE POINTS

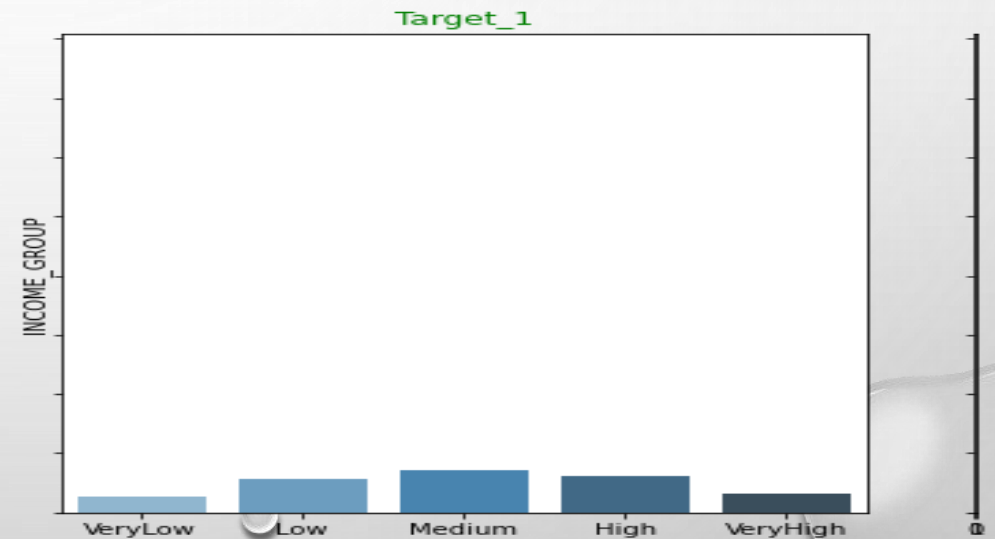
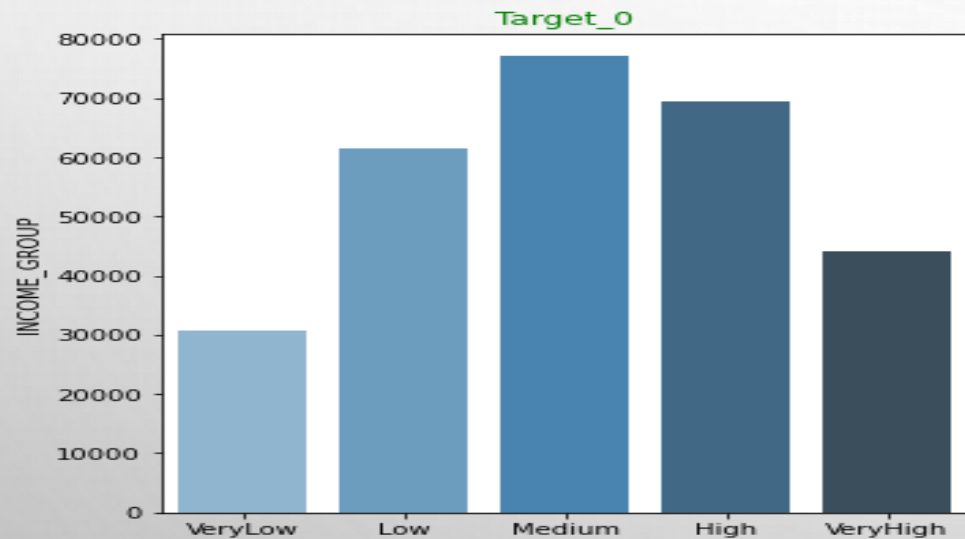
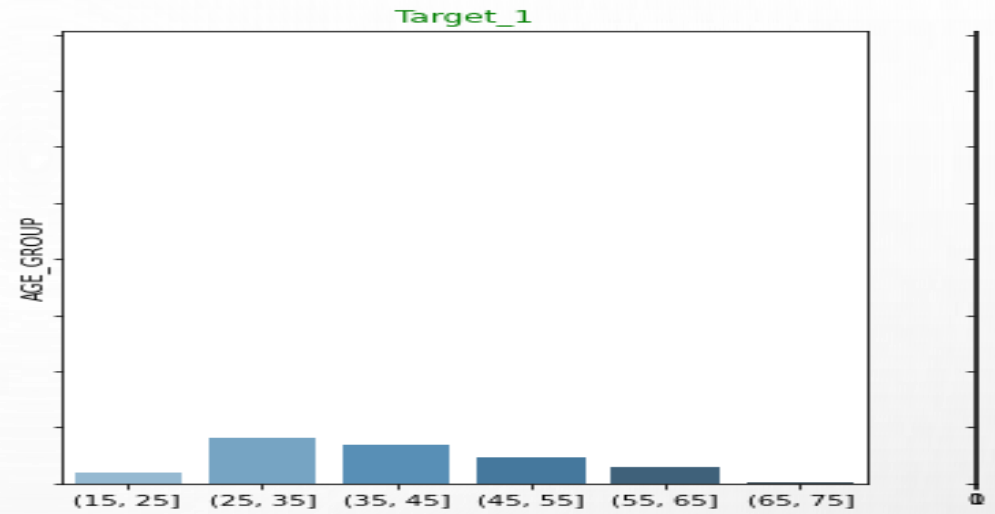
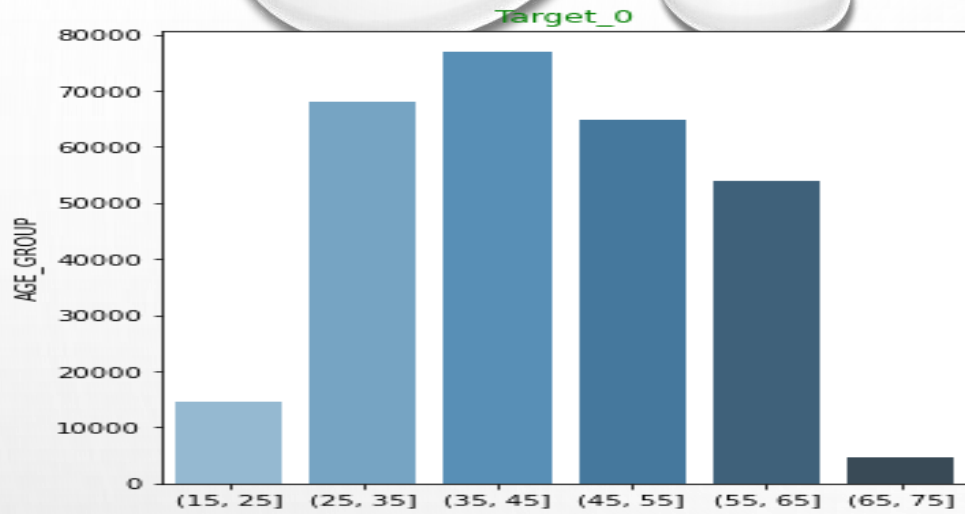
REGION_RATING_CLIENT' & 'REGION_RATING_CLIENT_W_CITY' - REGION 2 HAS THE HIGHEST % APPLICANTS BOTH IN TARGET 0 AND TARGET 1.

REG_REGION_NOT_LIVE_REGION, REGION_NOT_WORK_REGION, VE_REGION_NOT_WORK_REGION' - FOR BOTH TARGET 0 AND TARGET 1 OUT OF REGION, IE 1 IS VERY LOW AND DOES NOT SEEM TO AFFECT THE DEFAULT RATE.

REG_CITY_NOT_LIVE_CITY, REG_CITY_NOT_WORK_CITY - DEFAULT RATIO IS HIGHER FOR 1, IE DIFFERENT FROM PERMANENT ADDRESS.

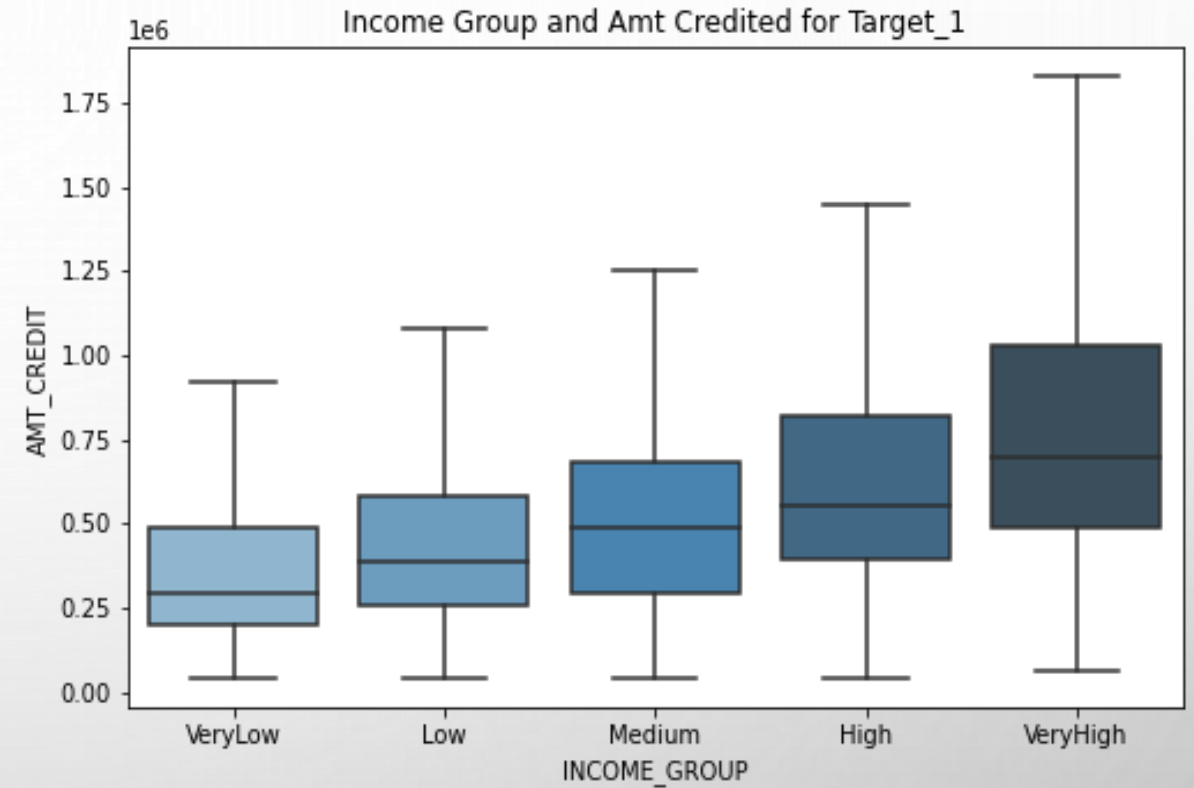
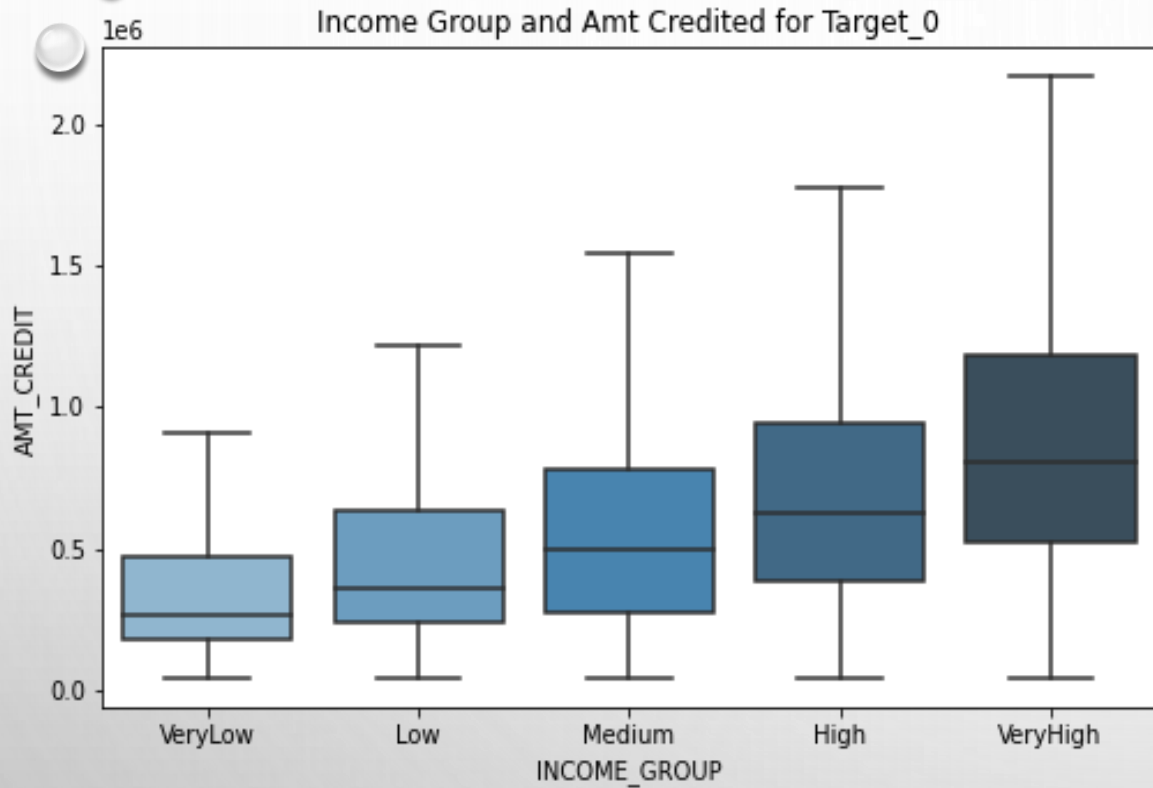
'CODE_GENDER' - RATIO OF F TO M IN TARGET 0 IS 2.3 AND F TO M IN TARGET 0 - 1.3. INDICATING THAT MEN ARE DEFAULTING MORE THAN WOMEN.

TARGETS IN INCOME AND AGE GROUP

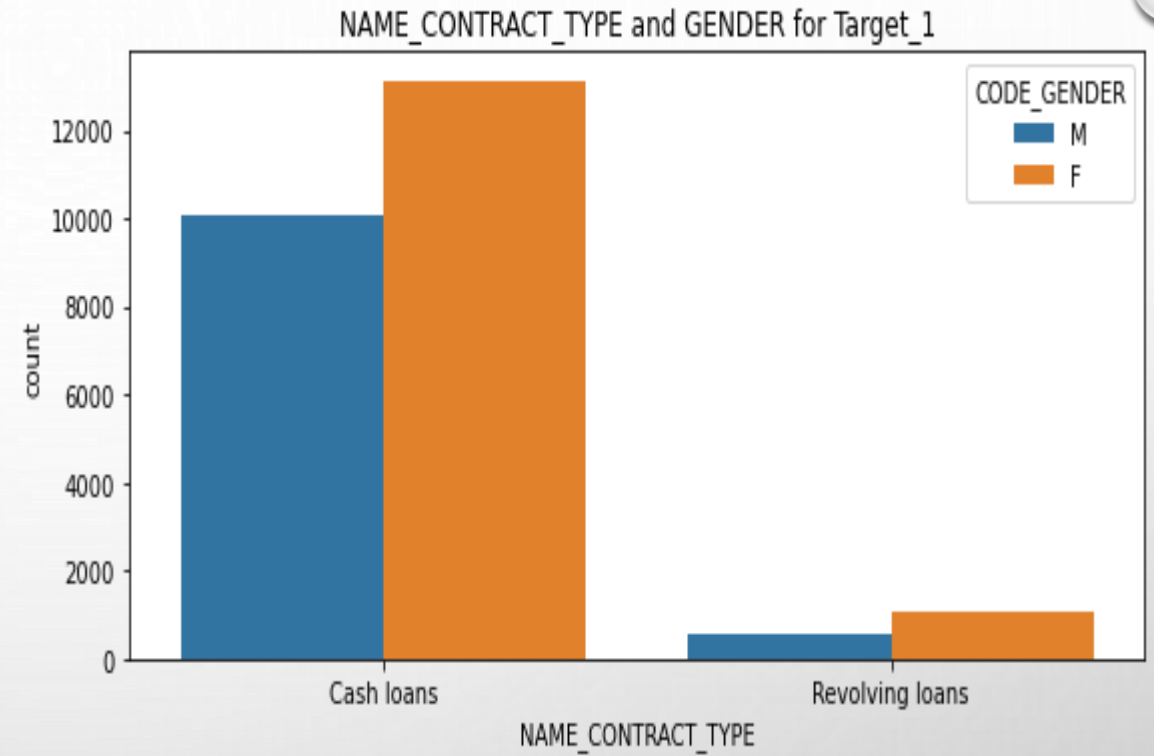
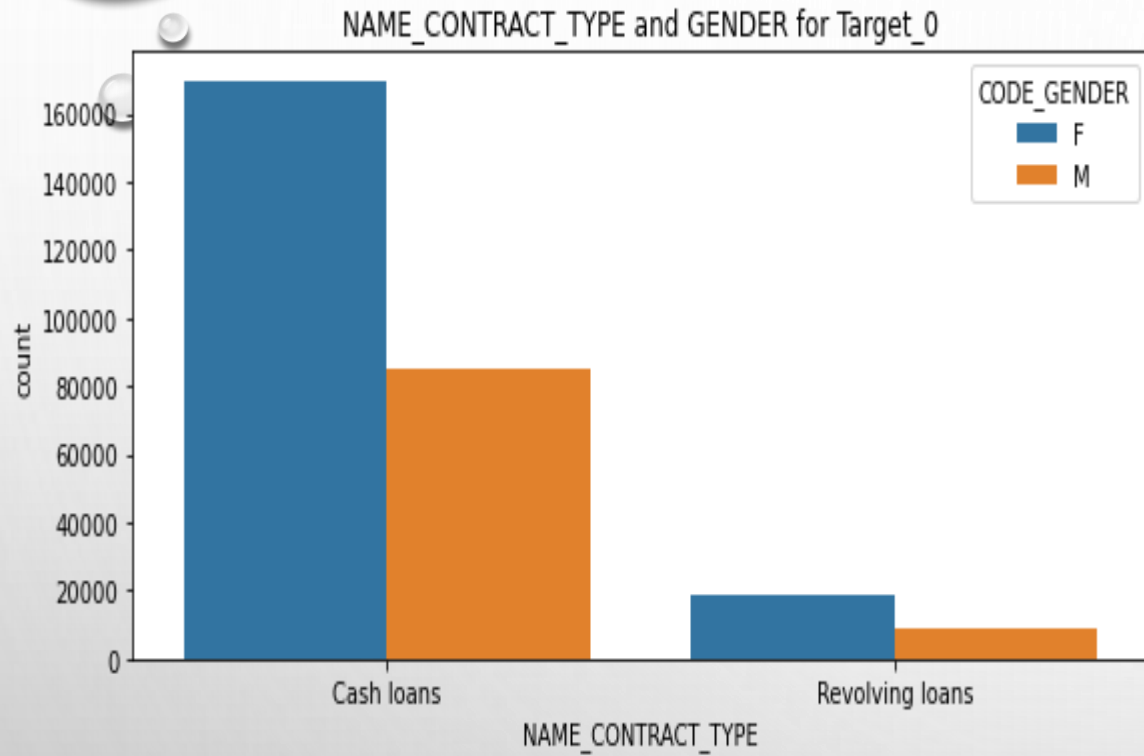


- AGE_GROPU - 35-45 ARE MORE IN TARGET_0 0. IN TARGET_1 - 25-25 HAVE HIGHER SHARE. AGE SEEMS LIKE INFLUENCING DEFAULTERS.
- 2. INCOME_GROUP - MEDIUM INCOME GROUP HAVE MORE COUNT IN TARGET_0 AND TARGET_1

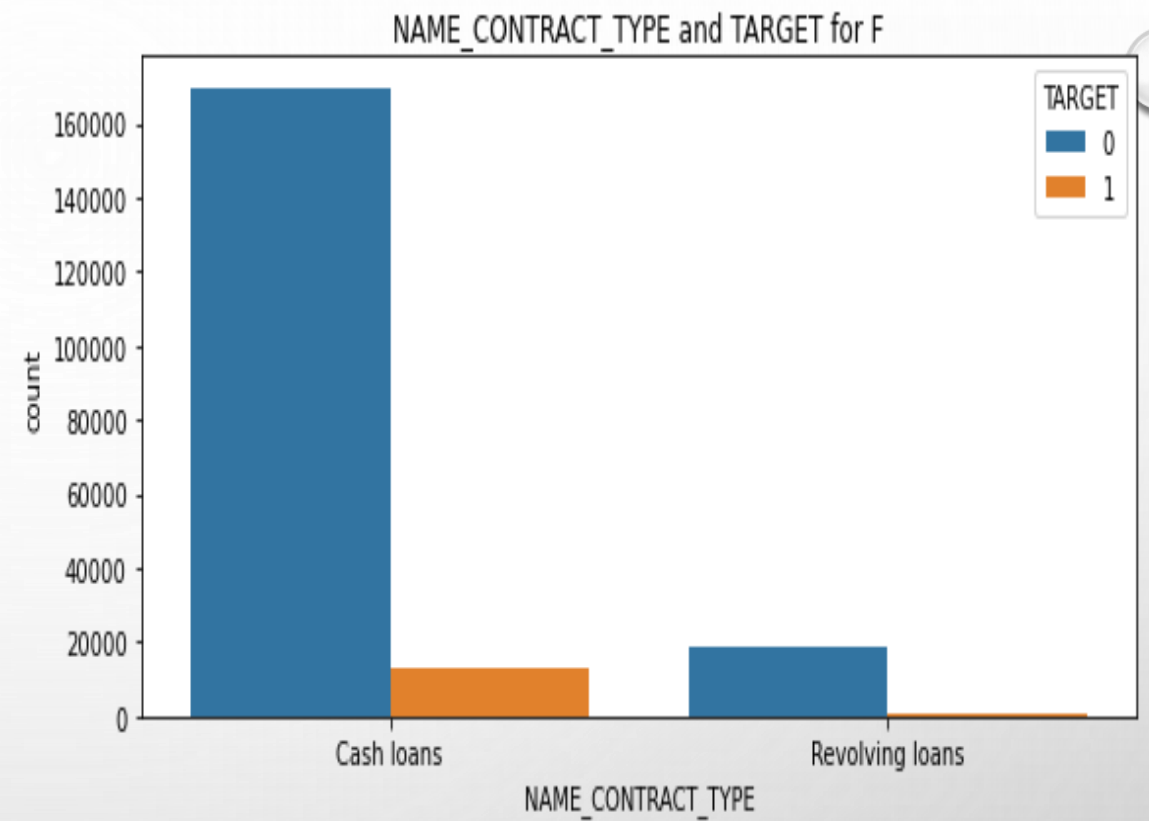
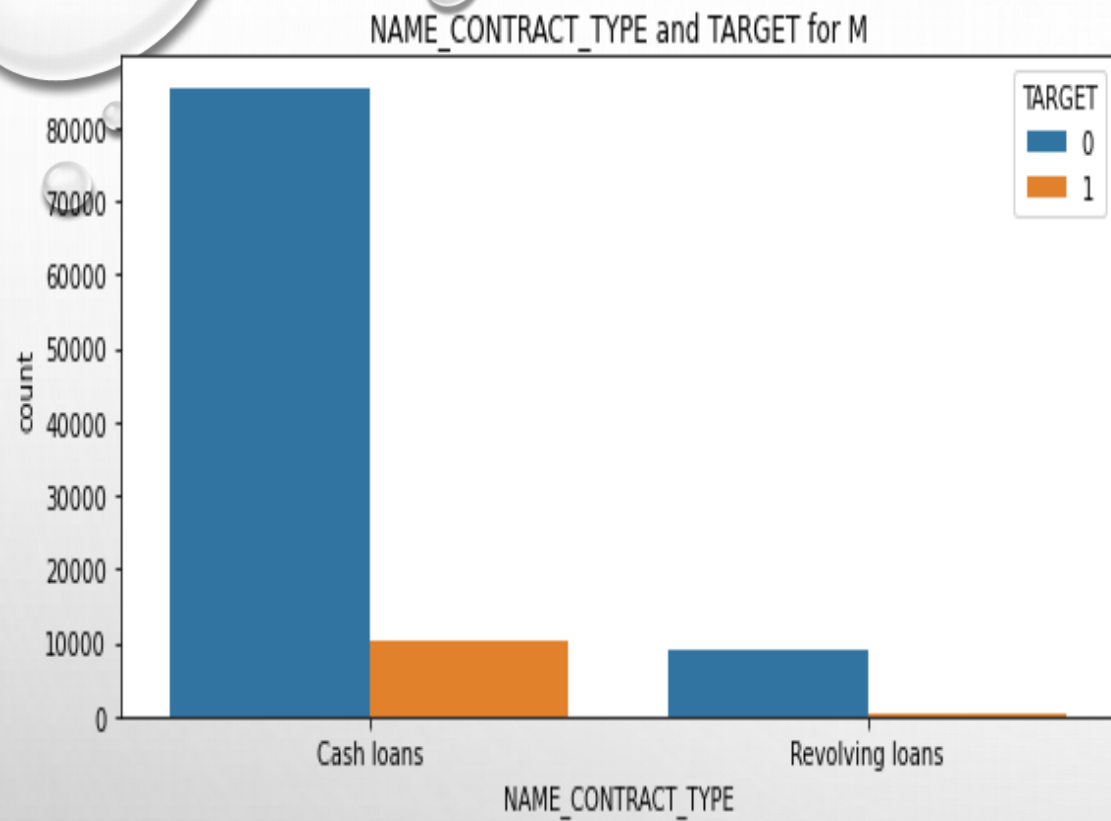
BIVARIATE ANALYSIS ON CATEGORICAL AND CONTINUOUS



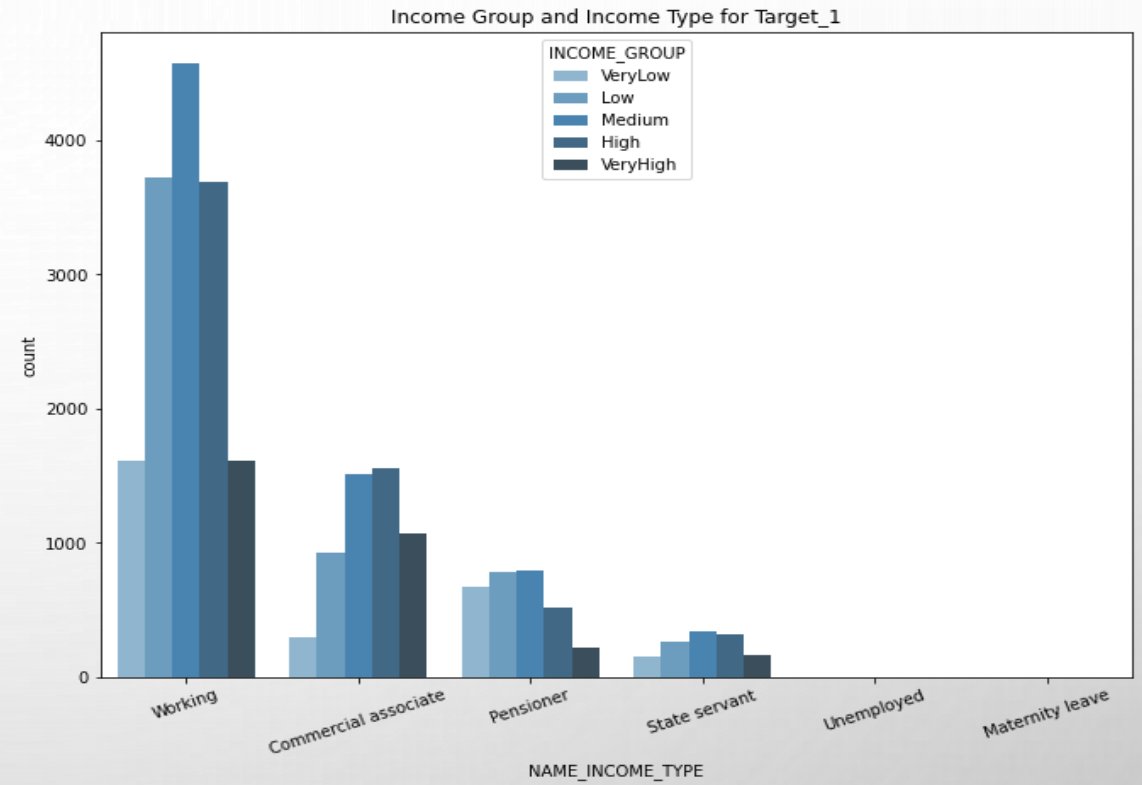
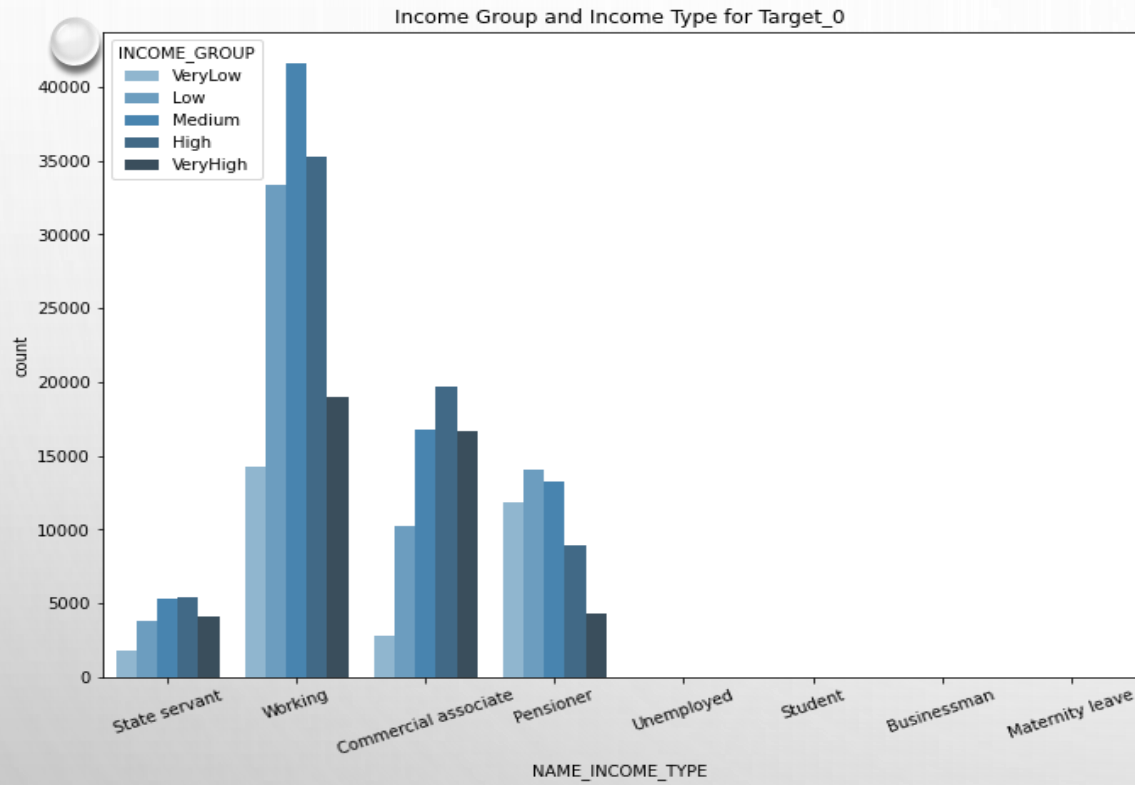
- LOOKING AT THE DATA NO INSIGHTS CAN BE DRAWN.
- APPLICANTS WITH ACADEMIC DEGREE HAVE HIGH DEFAULTERS. BUT FROM PLOT, NO OF APPLICANTS WITH ACADEMIC DEGREE IS MINIMUM.



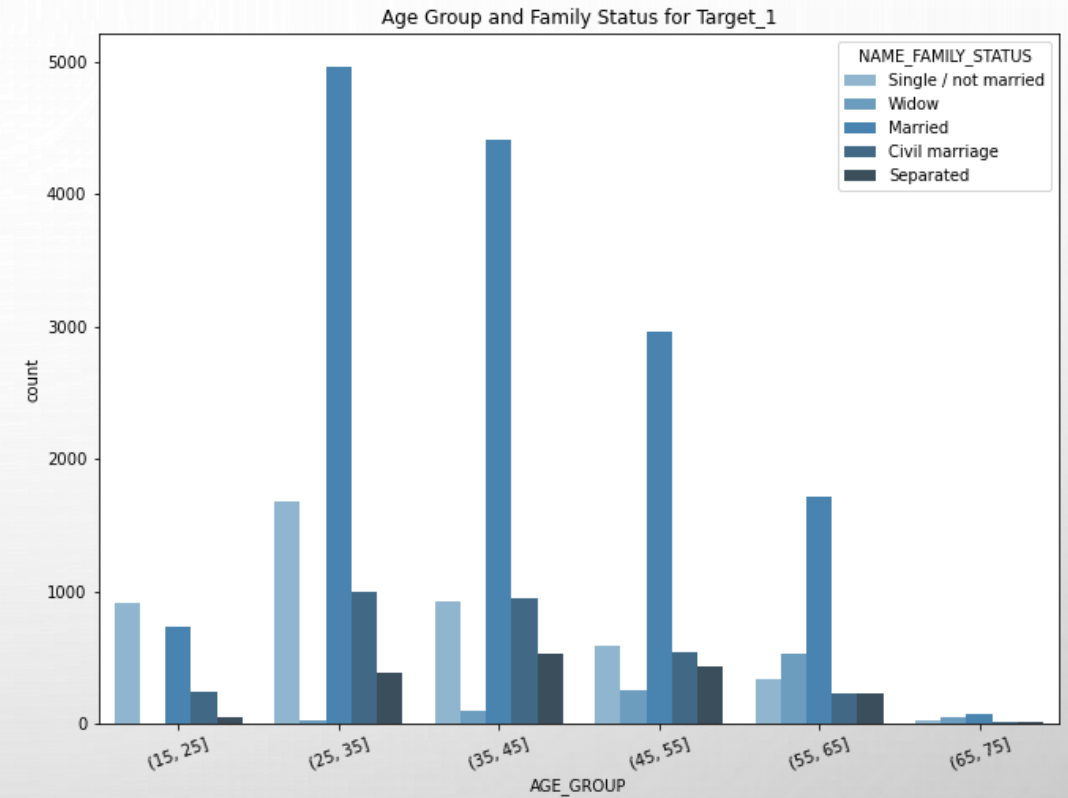
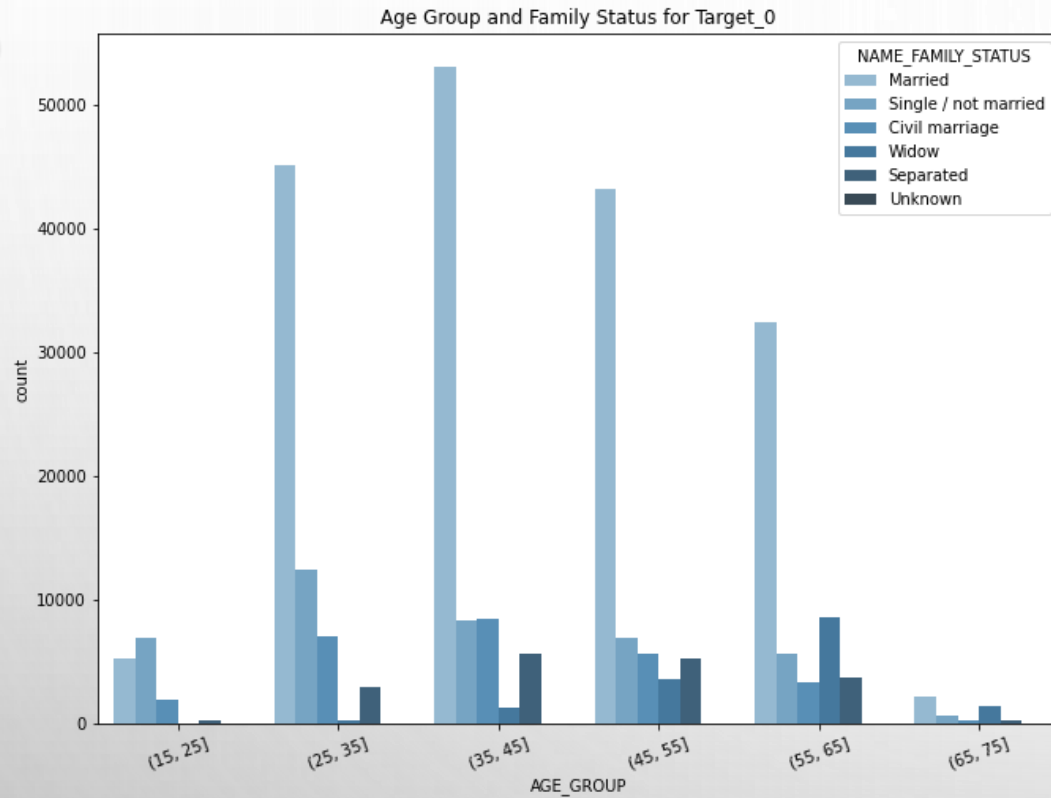
We have more females as loan applicants, although male applicants are lower, ratio of male applicants defaulters is higher.



This clearly shows more defaulters are males than female applicants .

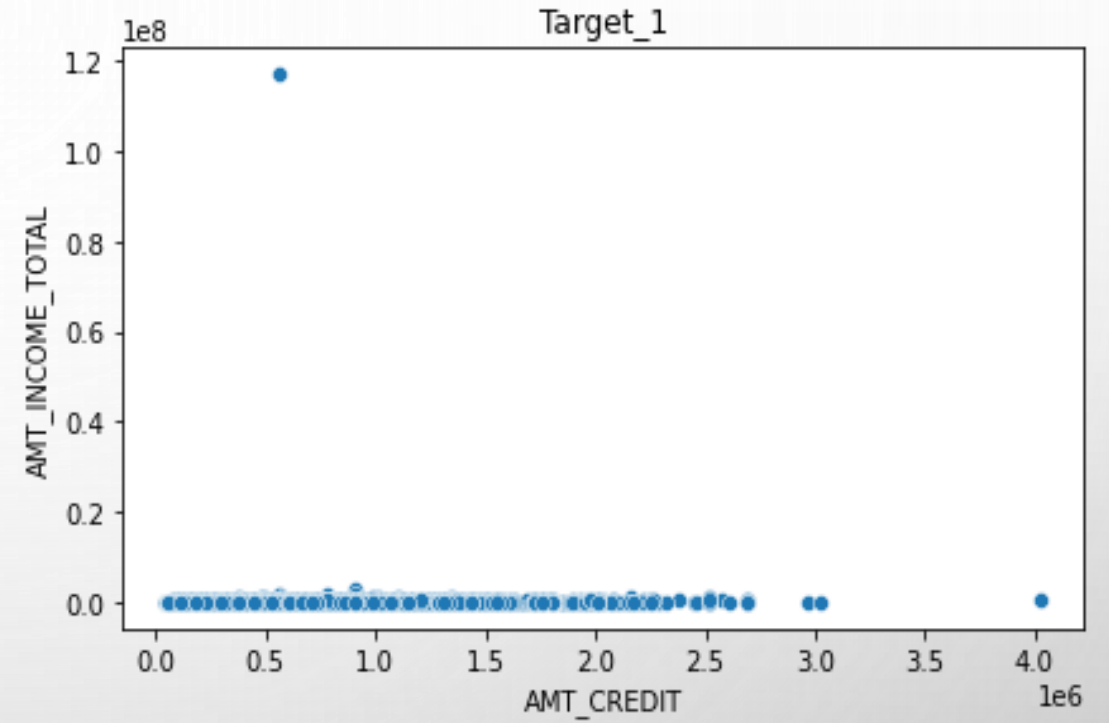
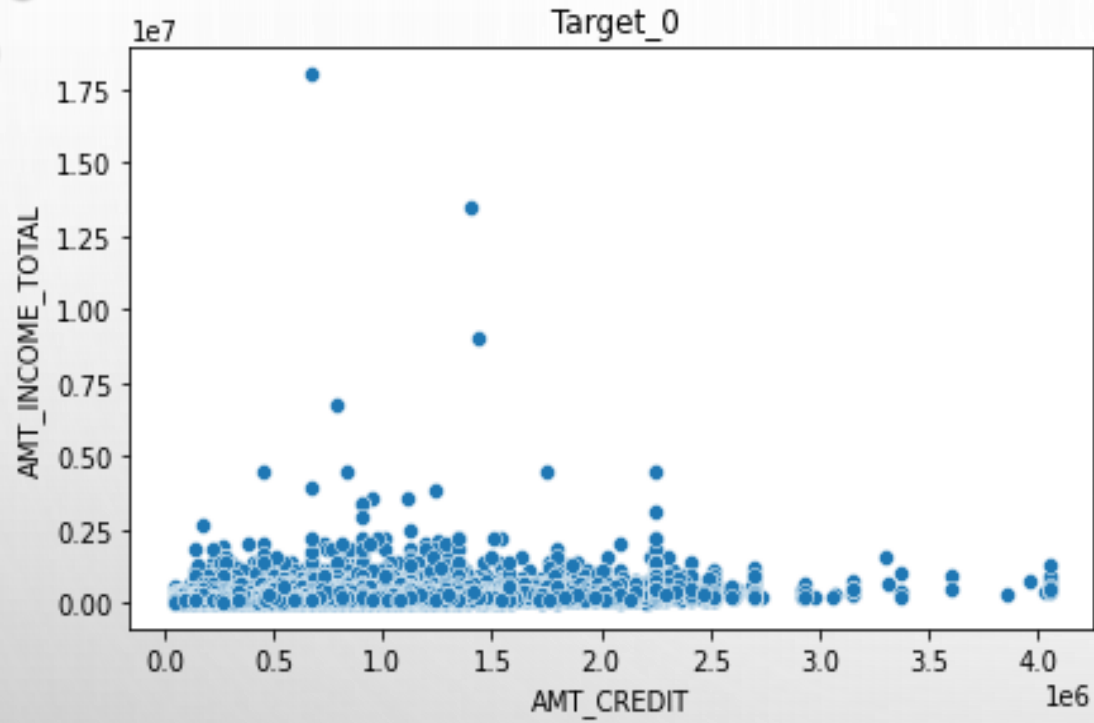


AVERAGE MEDIUM INCOME TYPE GROUP HAS ALMOST 1 IN 12 DEFAULTS WHICH IS HIGHER THAN THE AVG 1 IN 11 DEFAULTS.



In the age group 25-35 and 35-45 of married applicant have largest number of applicants with payment difficulties.

AMT_CREDIT vs AMT_INCOME_TOTAL



AMT_CREDIT AND AMT_GOOD PRICE DON'T SEEM TO BE INCREASING PROPORTIONATELY WITH AMT_INCOME FOR TARGET_1, THUS POSSIBLY LEADING TO DEFAULT.

HEATMAP CORRELATION

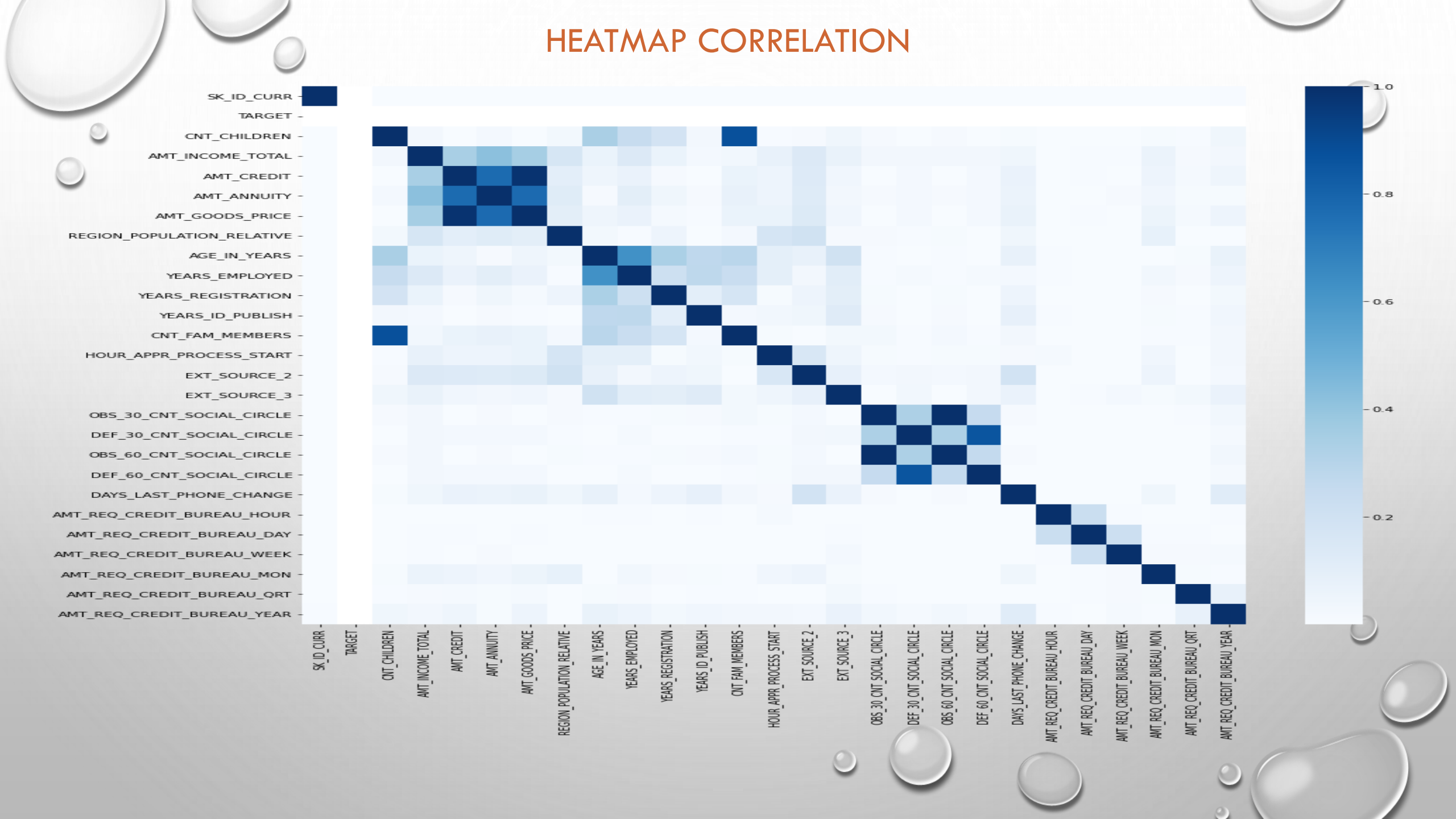
Heatmap correlation matrix showing relationships between 28 variables. The color scale ranges from 0.0 (light blue) to 1.0 (dark blue). The diagonal is dark blue, indicating perfect self-correlation. The matrix is symmetric, with the lower triangle mirroring the upper triangle. The variables are listed on the left and bottom axes.

Variables (Y-axis):

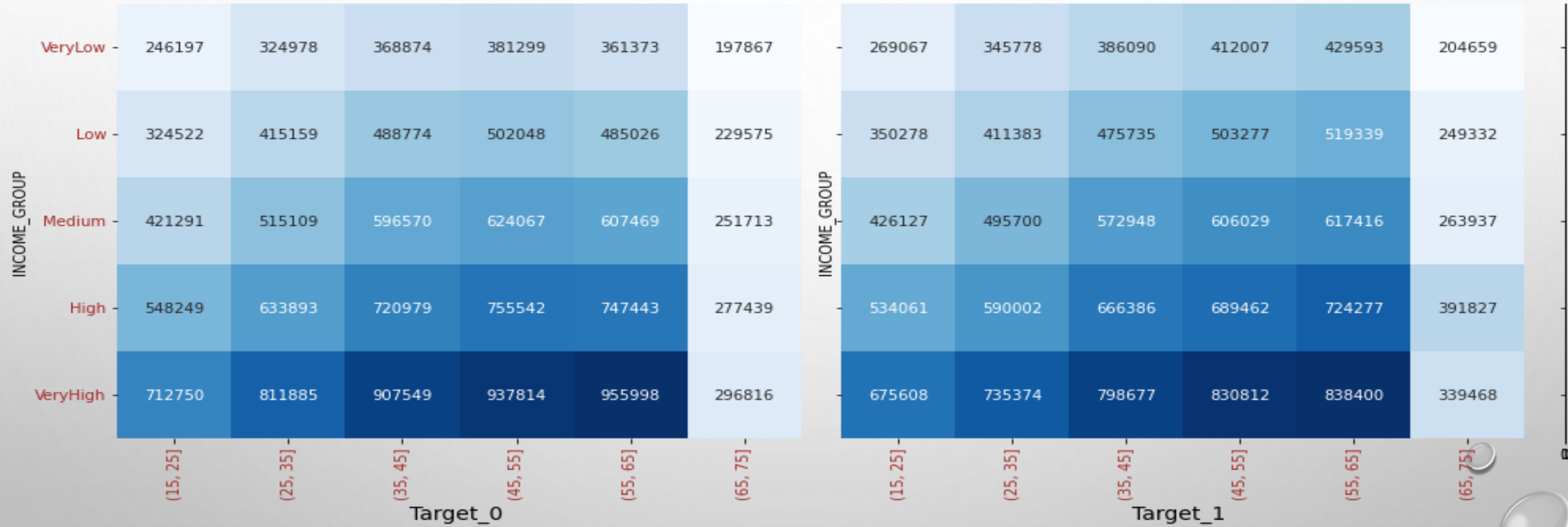
- SK_ID_CURR
- TARGET
- CNT_CHILDREN
- AMT_INCOME_TOTAL
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE
- REGION_POPULATION_RELATIVE
- AGE_IN_YEARS
- YEARS_EMPLOYED
- YEARS_REGISTRATION
- YEARS_ID_PUBLISH
- CNT_FAM_MEMBERS
- HOUR_APPR_PROCESS_START
- EXT_SOURCE_2
- EXT_SOURCE_3
- OBS_30_CNT_SOCIAL_CIRCLE
- DEF_30_CNT_SOCIAL_CIRCLE
- OBS_60_CNT_SOCIAL_CIRCLE
- DEF_60_CNT_SOCIAL_CIRCLE
- DAYS_LAST_PHONE_CHANGE
- AMT_REQ_CREDIT_BUREAU_HOUR
- AMT_REQ_CREDIT_BUREAU_DAY
- AMT_REQ_CREDIT_BUREAU_WEEK
- AMT_REQ_CREDIT_BUREAU_MON
- AMT_REQ_CREDIT_BUREAU_QRT
- AMT_REQ_CREDIT_BUREAU_YEAR

Variables (X-axis):

- SK_ID_CURR
- TARGET
- CNT_CHILDREN
- AMT_INCOME_TOTAL
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE
- REGION_POPULATION_RELATIVE
- AGE_IN_YEARS
- YEARS_EMPLOYED
- YEARS_REGISTRATION
- YEARS_ID_PUBLISH
- CNT_FAM_MEMBERS
- HOUR_APPR_PROCESS_START
- EXT_SOURCE_2
- EXT_SOURCE_3
- OBS_30_CNT_SOCIAL_CIRCLE
- DEF_30_CNT_SOCIAL_CIRCLE
- OBS_60_CNT_SOCIAL_CIRCLE
- DEF_60_CNT_SOCIAL_CIRCLE
- DAYS_LAST_PHONE_CHANGE
- AMT_REQ_CREDIT_BUREAU_HOUR
- AMT_REQ_CREDIT_BUREAU_DAY
- AMT_REQ_CREDIT_BUREAU_WEEK
- AMT_REQ_CREDIT_BUREAU_MON
- AMT_REQ_CREDIT_BUREAU_QRT
- AMT_REQ_CREDIT_BUREAU_YEAR

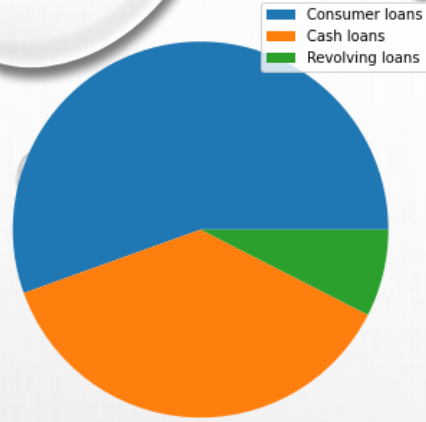


#AMT_CREDIT RELATIONSHIP ANALYSIS WITH AGE GROUP AND INCOME GROUP

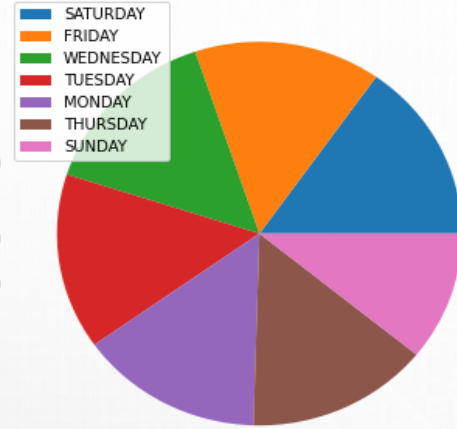


UNIVARIATE ANALYSIS ON PREVIOUS APPLICATION DATA

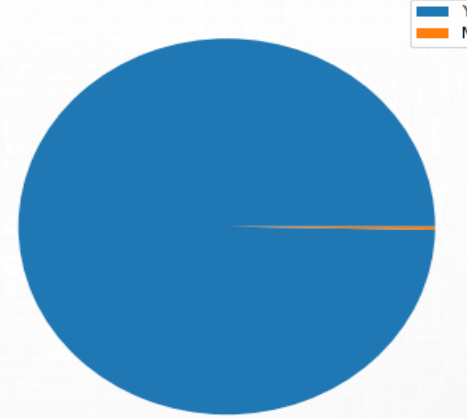
NAME_CONTRACT_TYPE



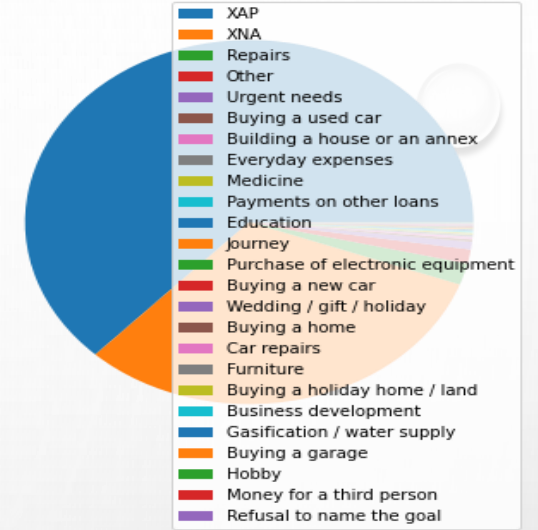
WEEKDAY_APPR_PROCESS_START



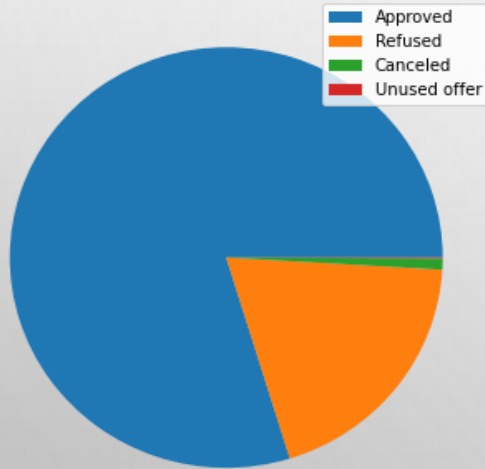
FLAG_LAST_APPL_PER_CONTRACT



NAME_CASH_LOAN_PURPOSE



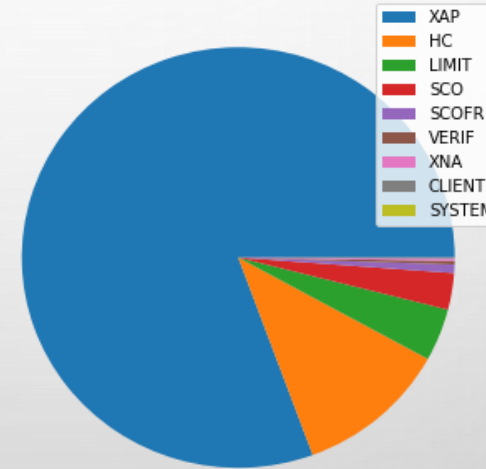
NAME_CONTRACT_STATUS



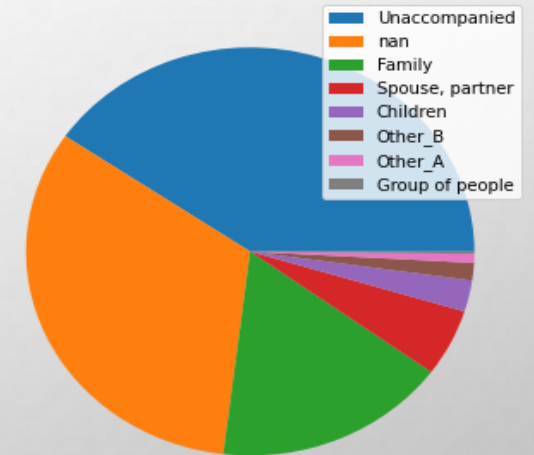
NAME_PAYMENT_TYPE



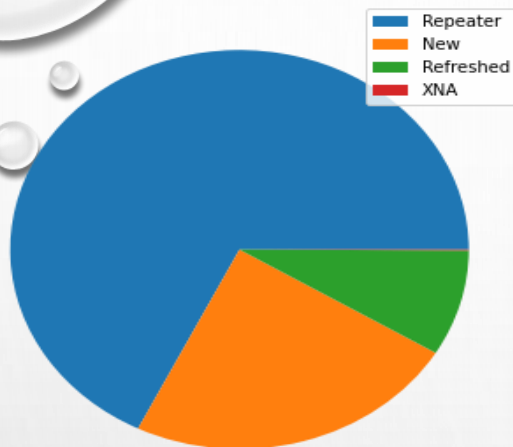
CODE_REJECT_REASON



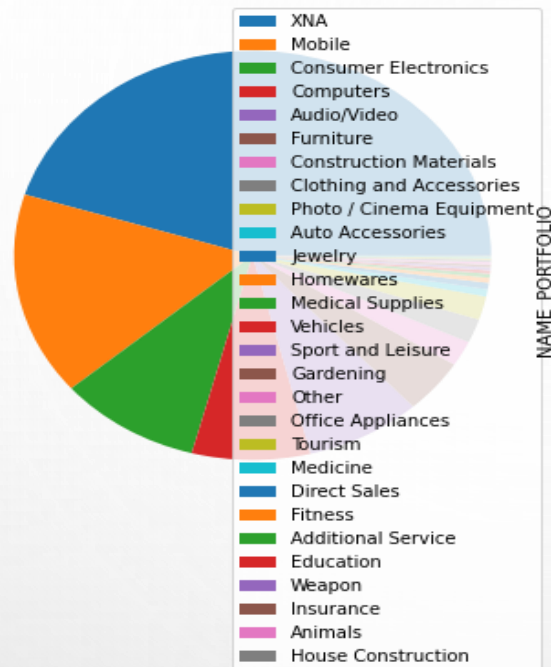
NAME_TYPE_SUITE



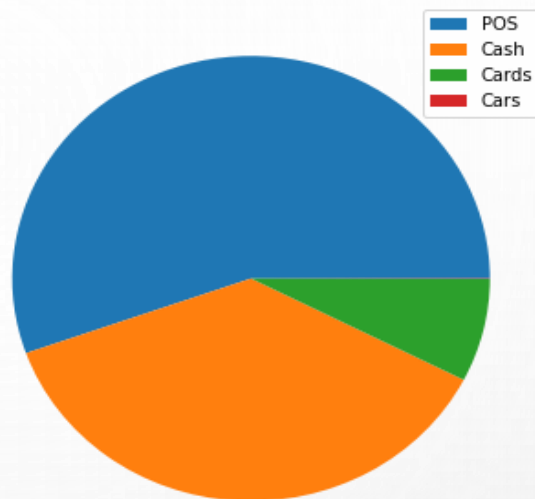
NAME_CLIENT_TYPE



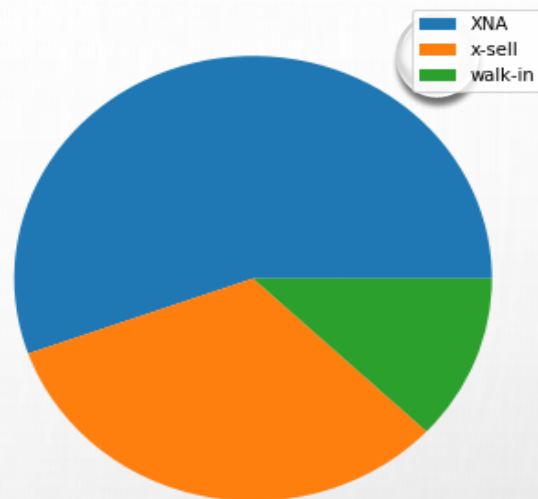
NAME_GOODS_CATEGORY



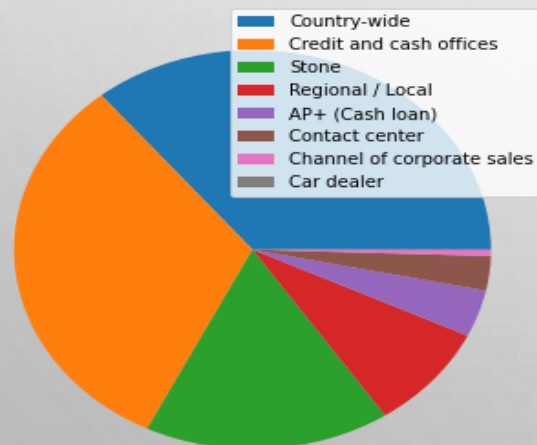
NAME_PORTFOLIO



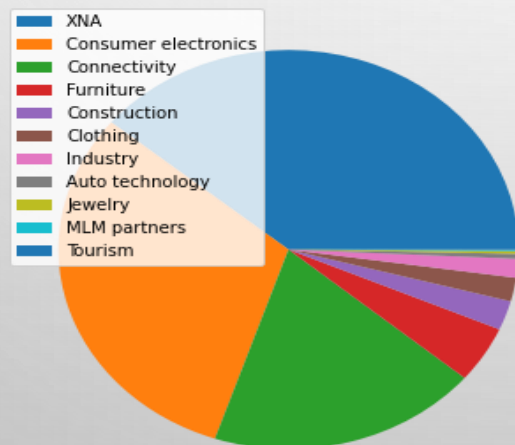
NAME_PRODUCT_TYPE



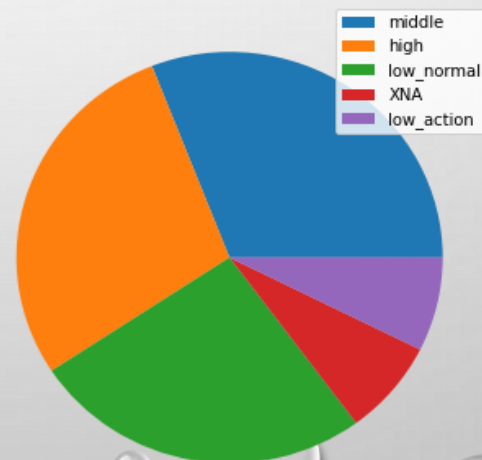
CHANNEL_TYPE



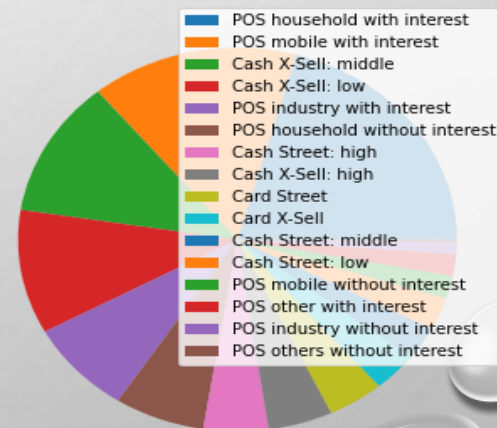
NAME_SELLER_INDUSTRY



NAME_YIELD_GROUP



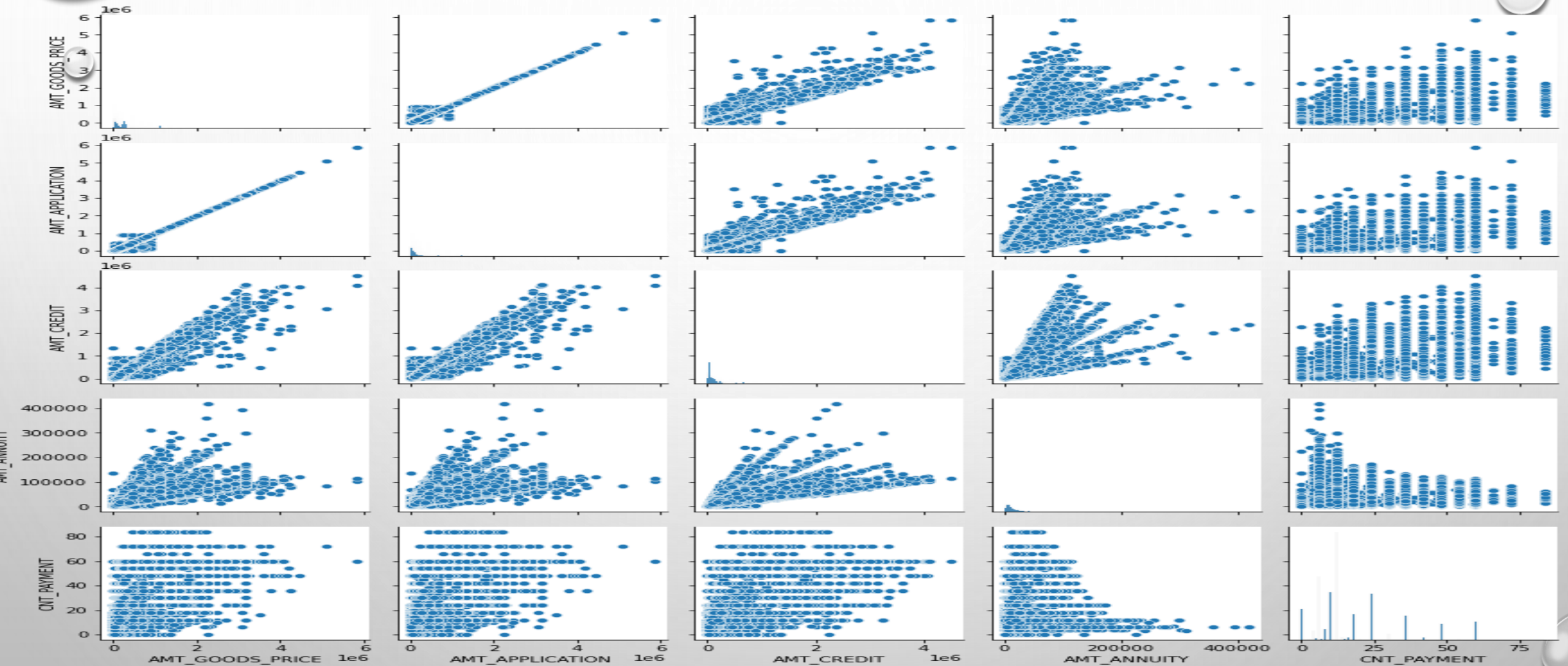
PRODUCT_COMBINATION



NOTABLE POINTS

- THIS DATA FRAME HAS A DIFFERENT TYPE OF LOAN CALLED CONSUMER LOAN, WHICH WAS NOT THERE IN APPLICATION DATA FRAME. 55% OF LOANS ARE CONSUMER LOANS. 37% CASH LOANS AND REST REVOLVING.
- APPROVED LOANS ARE 79% AND REFUSED, CANCELLED, UNUSED - REST *IMBALANCE IN DATA*.
- 67% ARE REPEATERS. NAME_CLIENT_TYPE ALSO HAS SOME NULL VALUES SHOWING AS XNA.
- 55% OF THE APPLICANTS HAVE TAKEN LOAN FOR POS PURCHASE.
- NAME SELLER INDUSTRY HAS 37% XNA VALUES, CONSUMER ELECTRONICS IS NEXT HIGHEST CATEGORY AT 30%.

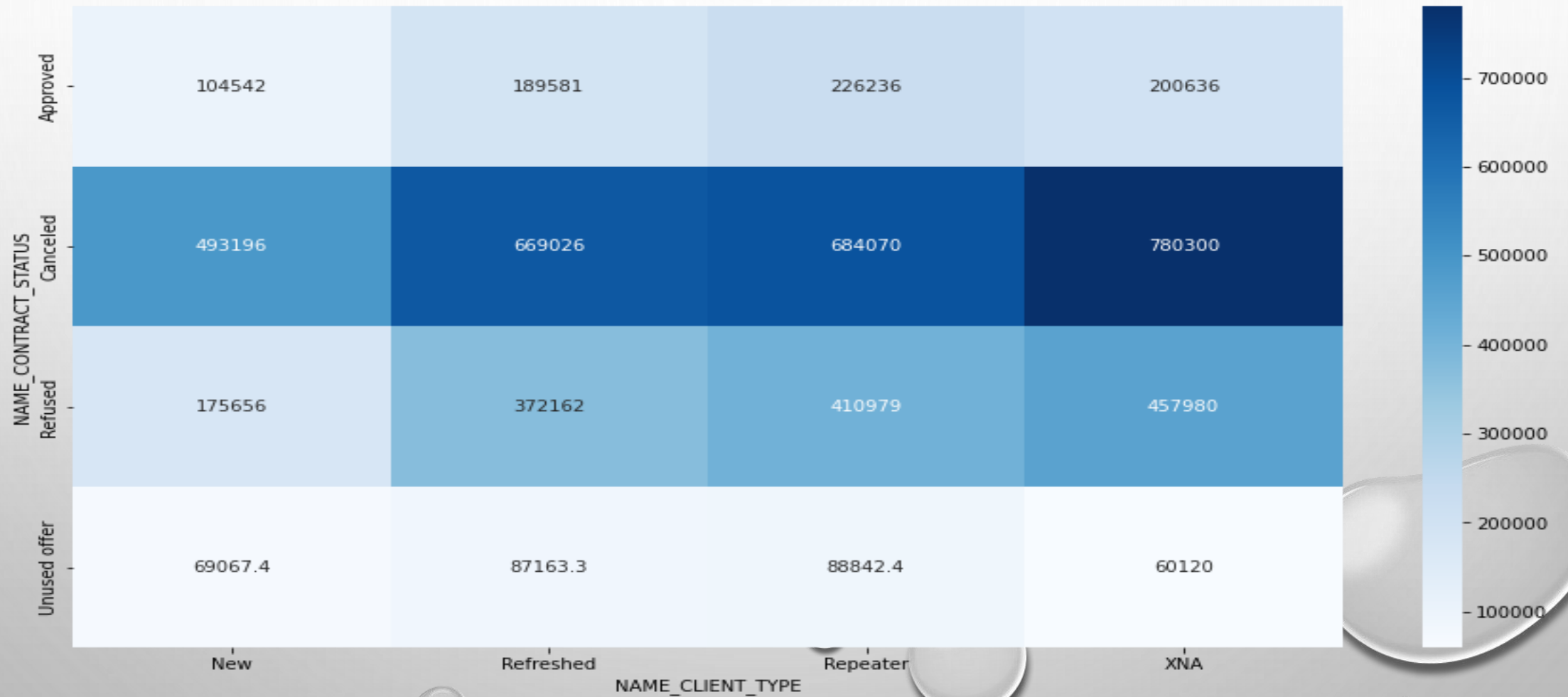
BIVARIATE ANALYSIS ON PREVIOUS APPLICATION



NOTABLE POINTS

- AMT_GOODS_PRICE, AMT_ANNUITY, AMT_APPLICATION - as expected have high correlation. Higher the value of good purchased more there will be need of loan and surely all these will correlate
- Similarly, amt_credit to AMT_GOOD_PRICE also the correlation is high
- Column cnt_payment ideally should have had a high correlation with amt_credit, ie higher credit, more the term of loan. But no such correlation can be seen.

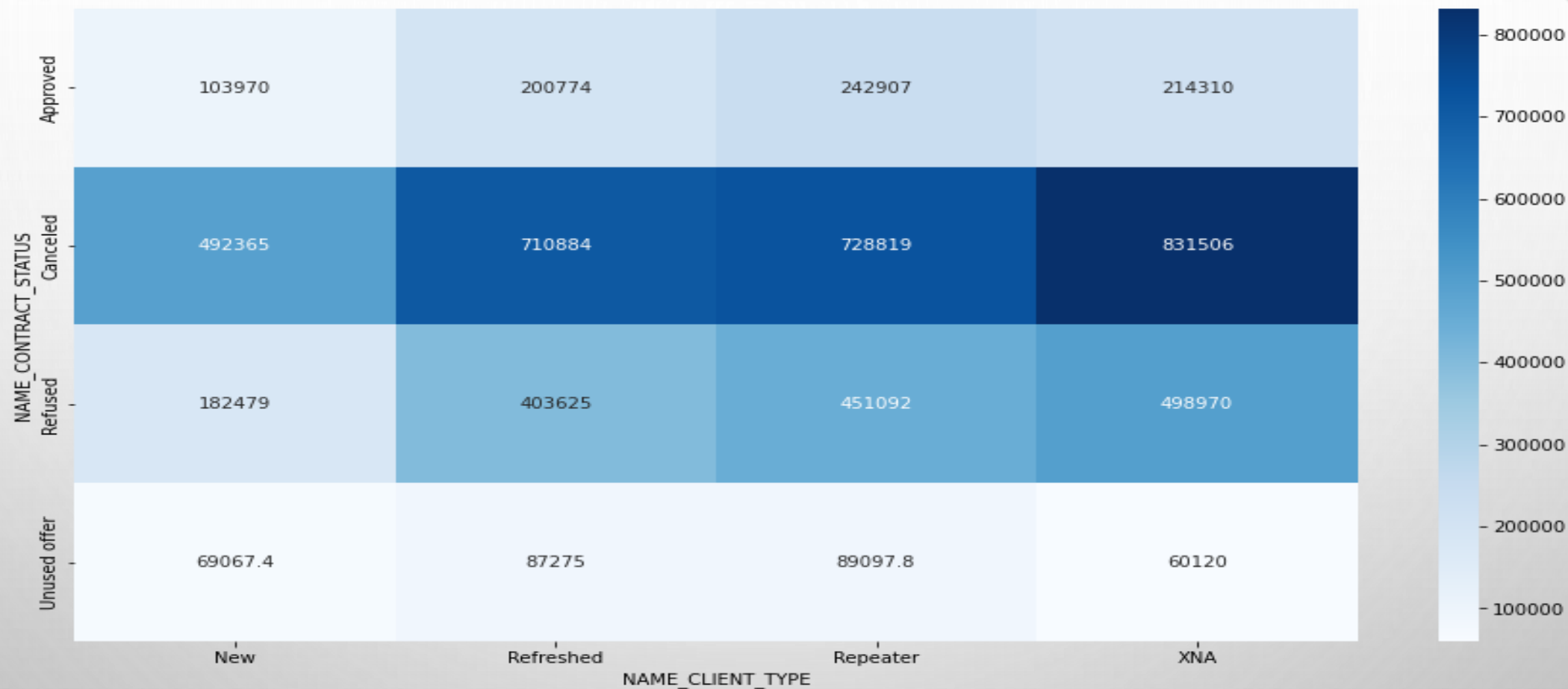
MULTIVARIATE ANALYSIS ON PREVIOUS APPLICATION



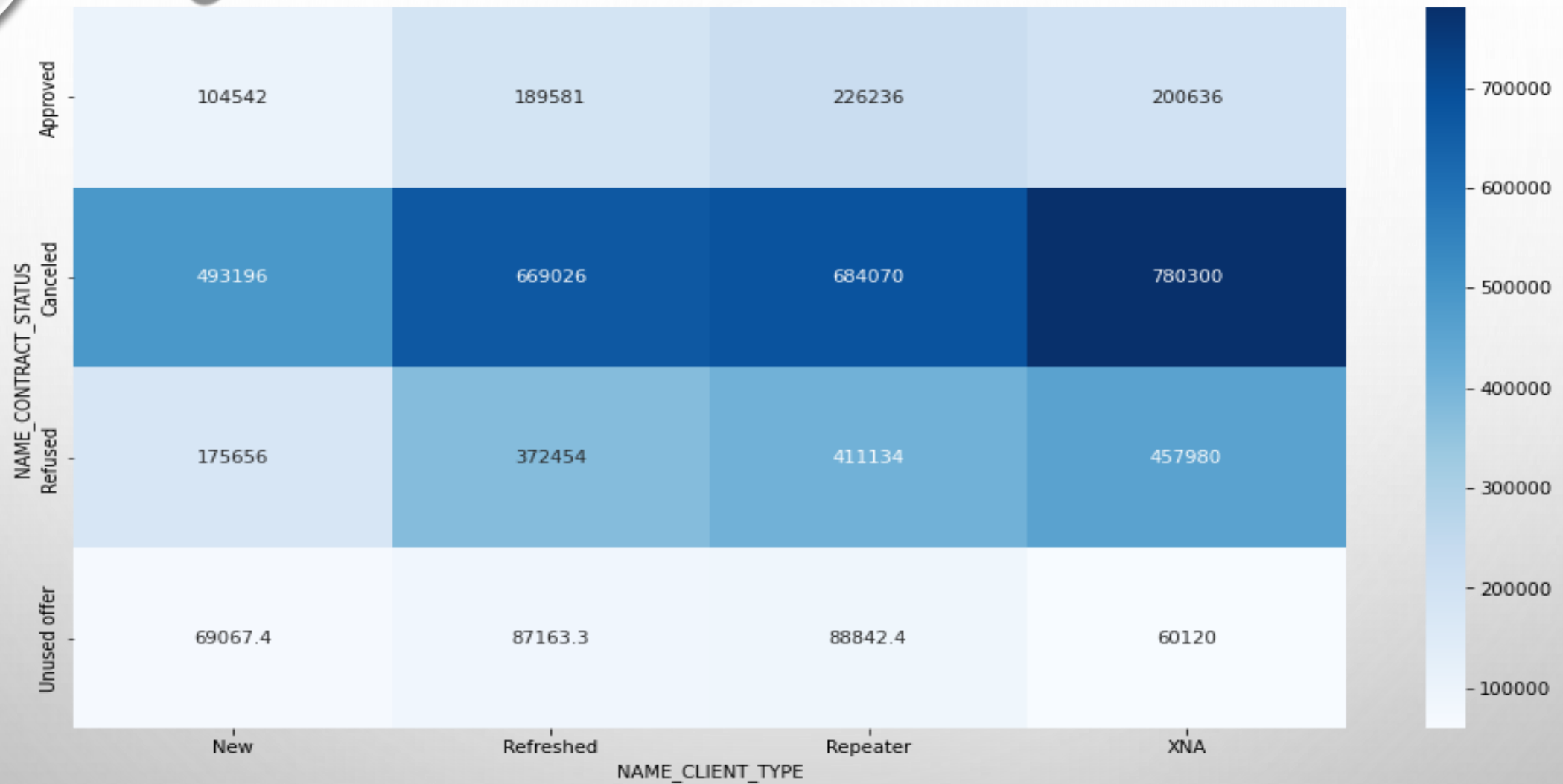
- UNUSED OFFER APPLICATION AMOUNT IS LOW.

CANCELLED APPLICATION AMOUNT IS HIGH. THE BANK MAY BE REFUSING THESE POSSIBLY AS THE DEBT LIABILITY RATIO OF CONSUMER MUST BE GOING HIGH DUE TO THE HIGH AMOUNT AND THUS CREDIT DEFAULT RISK.

REPREATER'S APPLICATION AMOUNT IS HIGHER THAN THE NEW CUSTOMERS. THIS MAY INDICATE THAT THE BANK HAS MORE CONDUCIVE POLICIES/RATE OF INTEREST ETC FOR REPEAT APPLICANTS.

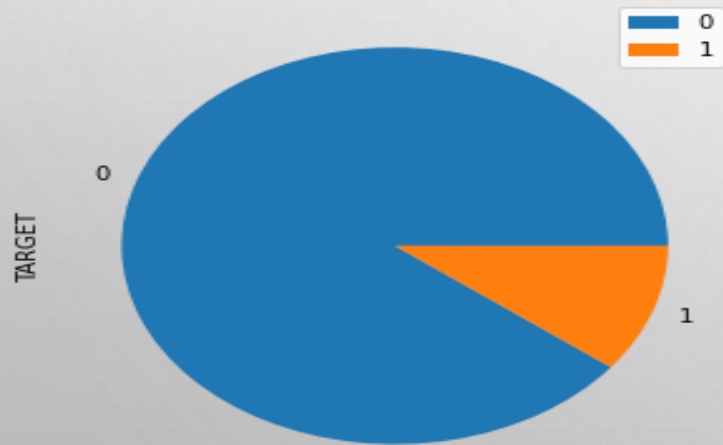
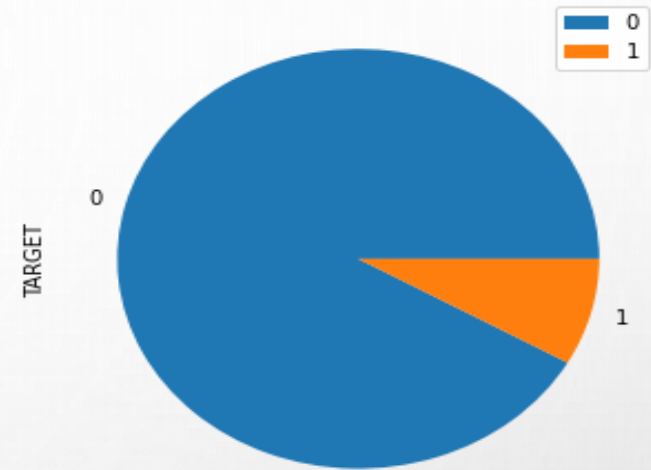
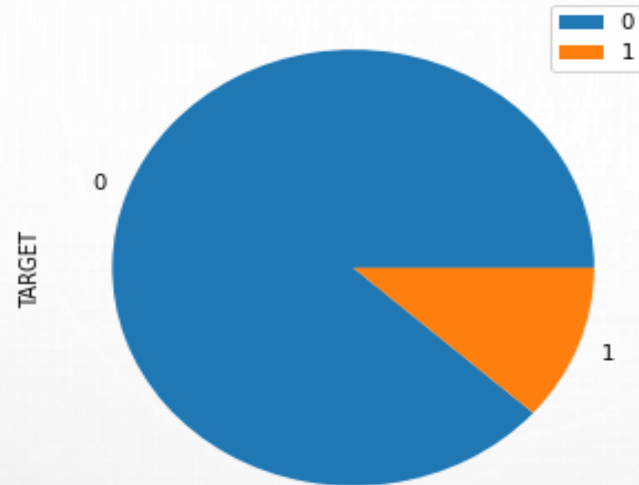


- UNUSED OFFER CREDIT AMOUNT IS LOW. THIS MAY BE THE REASON FOR CUSTOMER NOT USING IT.
UNABLE TO UNDERSTAND WHY FOR CANCELLED AND REFUSED THERE SHOULD BE ANY CREDIT AMOUNT?

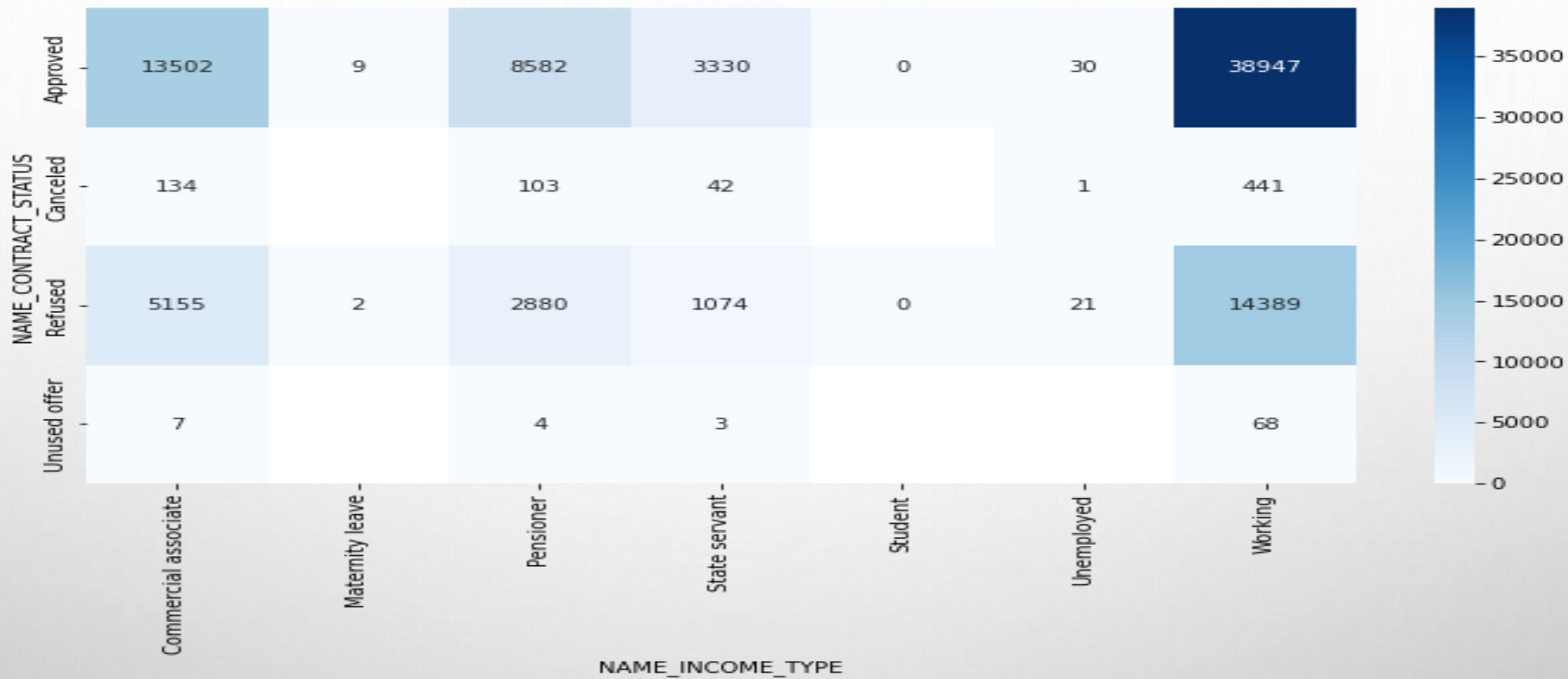


ALL CANCELLED AND REFUSED CASES ARE HAVING HIGHER VALUE OF GOODS THAN OTHER CATEGORIES.

MERGING DATA FRAMES



- 1. 7.5% OF APPROVED LOANS HAVE DEFAULTERS.
- 2. PREVIOUS APPLICATIONS WITH REFUSED, CANCELLED, UNUSED LOANS ALSO HAVE DEFAULTERS WHICH IS A MATTER OF CONCERN. THIS INDICATES THAT THE FINANCIAL COMPANY HAD REFUSED/CANCELLED PREVIOUS APPLICATION, BUT HAS APPROVED THE CURRENT AND IS FACING DEFAULT ON THESE LOANS.

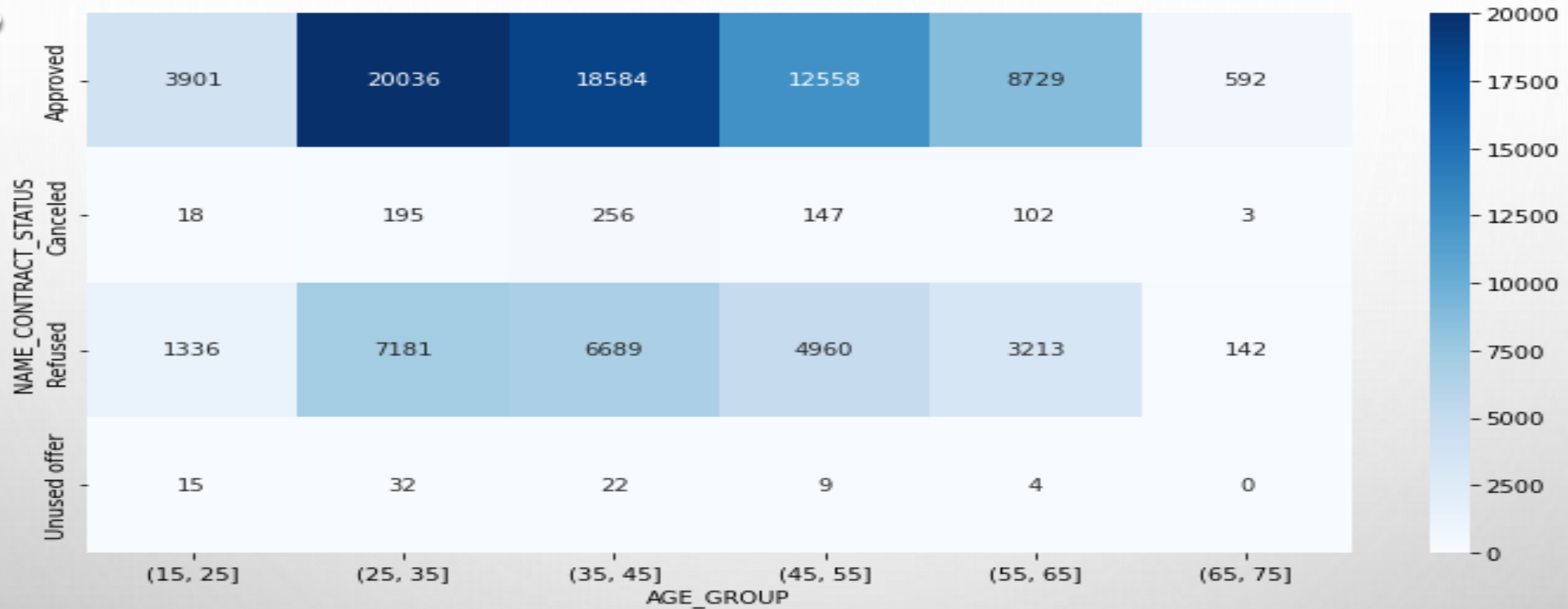


- HIGH ON ABOVE MATRIX SHOWS CORRELATION TO DEFAULTERS, TARGET_1 ARE DEFAULTERS.

WORKING APPLICANT WITH APPROVED STATUS HAVE DEFAULTED IN HIGHEST NUMBERS.

PREVIOUS APPLICATIONS WITH REFUSED, CANCELLED, UNUSED LOANS ALSO HAVE DEFAULT WHICH IS A MATTER OF CONCERN. THIS INDICATES THAT THE FINANCIAL COMPANY HAD REFUSED/CANCELLED PREVIOUS APPLICATION, BUT HAS APPROVED THE CURRENT AND IS FACING DEFAULT ON THESE LOANS.

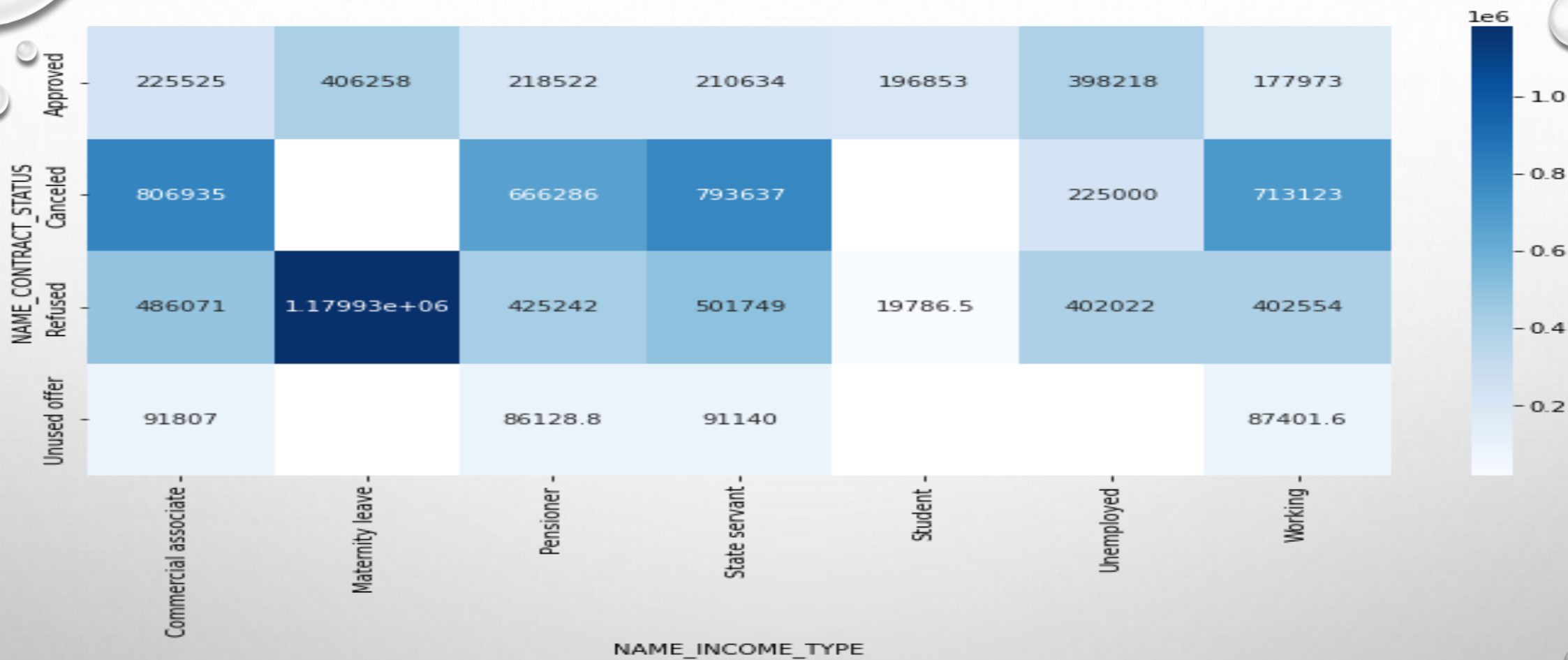
14,389 APPLICANTS OF WORKING CLASS WERE REFUSED EARLIER AND NOW HAVE DEFAULTED.



TARGET_1 ARE DEFAULTERS, HIGHER ON THE ABOVE MATRIX SHOWS CORRELATION TO DEFAULTERS.

APPROVED LOANS OF AGE GROUP 25-35 AND 35-45 HAVE HIGHER DEFAULTS.

REFUSED, CANCELLED, LOANS IN PREVIOUS APPLICATION HAVE DEFAULTED IN CURRENT.



- HIGHER CREDITS BEEN OFFERED TO UNEMPLOYED, MATERNITY LEAVE IS A NOTABLE FACTOR.

UNUSED OFFERS HAVE SMALLER CREDIT VALUES AND POSSIBLY THE REASON WHY APPLICANT IS NOT USING THEM.


DEFAULTERS IN APPROVED APPLICATIONS

- 'INCOME_GROUP' - MEDIUM INCOME
- 'AGE_GROUP' - 25-35, FOLLOWED BY 35-45
- 'NAME_INCOME_TYPE' - WORKING
- 'OCCUPATION_TYPE' - LABOURERS 31%
- 'ORGANIZATION_TYPE' - BUSINESS TYPE 3
- 'OWN_CAR_FLAG' - 31% DON'T HAVE CAR
- 'OWN_REALTY_FLAG' - 70% DON'T HAVE OWN HOME
- ABOVE ALL VARIABLES WERE ESTABLISHED IN ANALYSIS OF APPLICATION DATA FRAME LEADING TO DEFAULT. CHECKED THESE AGAINST THE APPROVED APPLICATION AND DEFAULT CASES AND IT PROVES TO BE CORRECT AND ARE HIGH DEFAULTERS.



SUMMARY/INSIGHTS OF CASE STUDY

ALL THE BELOW VARIABLES WERE ESTABLISHED IN ANALYSIS OF APPLICATION DATA FRAME AS DEFAULTERS, CHECKING THESE AGAINST THE APPROVED LOANS WHICH HAVE DEFAULTS, AND IT PROVES TO BE CORRECT:

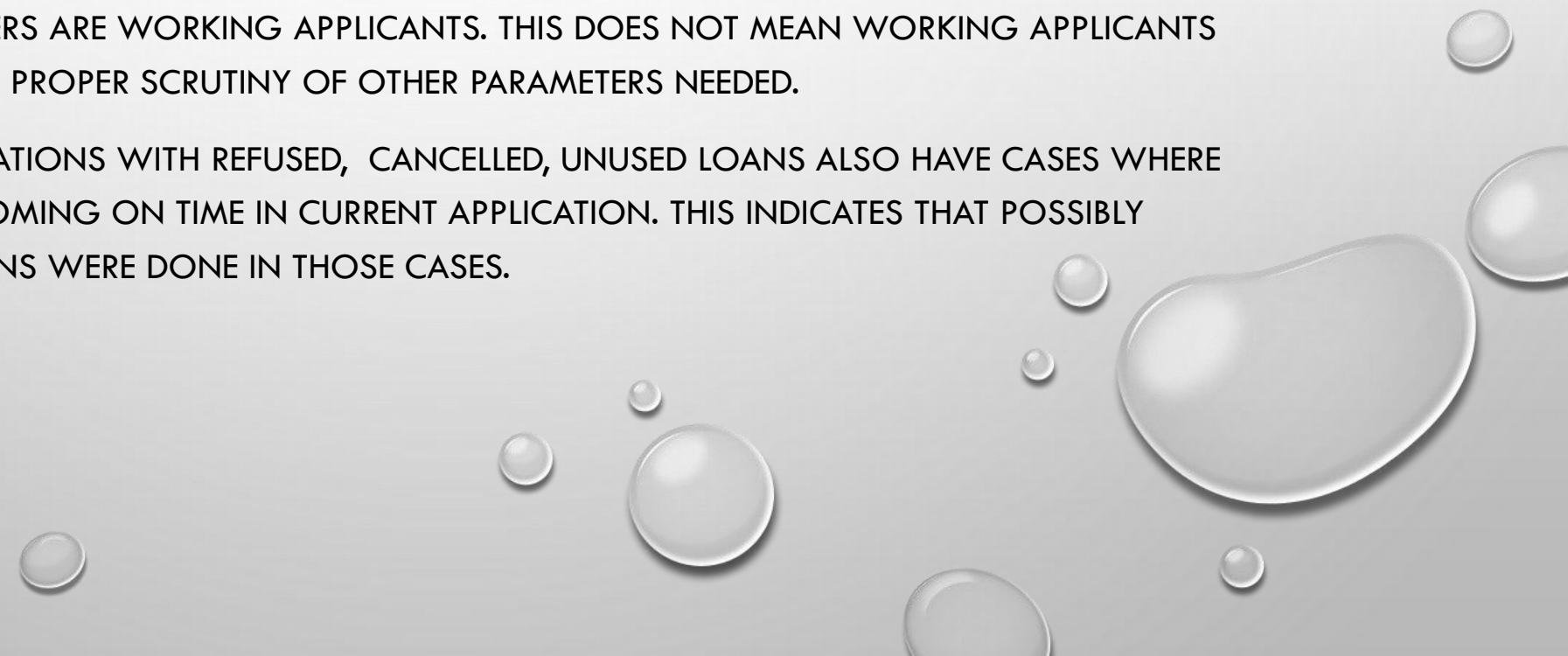
- Medium income -
 - 25-35 years old, followed by 35-45 years age group
 - Male
 - Unemployed
 - Labourers, Salesman, Drivers
 - Business type 3
 - Own House - No
- 

IMPORTANT Factors to be considered:

- Days last phone number changed.
- Lower figure points at concern.
- No of Bureau Hits in last week. Month etc.
- zero hits is good.
- Amount income not correspondingly equivalent to Good Bought
- Income low and good value high is a concern.
- Previous applications with Refused, Cancelled, Unused loans also have default which is a matter of concern.
- This indicates that the financial company had Refused/Cancelled previous application but has approved the current and is facing default on these.



CREDIBLE APPLICATIONS REFUSED

- UNUSED APPLICATIONS HAVE LOWER LOAN AMOUNT. IS THIS THE REASON FOR NO USAGE?
 - FEMALE APPLICANTS SHOULD BE GIVEN EXTRA WEIGHTAGE AS DEFAULTS ARE LESSER.
 - 60% OF DEFAULTERS ARE WORKING APPLICANTS. THIS DOES NOT MEAN WORKING APPLICANTS MUST BE REFUSED. PROPER SCRUTINY OF OTHER PARAMETERS NEEDED.
 - PREVIOUS APPLICATIONS WITH REFUSED, CANCELLED, UNUSED LOANS ALSO HAVE CASES WHERE PAYMENTS ARE COMING ON TIME IN CURRENT APPLICATION. THIS INDICATES THAT POSSIBLY WRONG DECISIONS WERE DONE IN THOSE CASES.
- 

NOTE:-VARIOUS REFERENCES INCLUDING GOOGLE AND OTHER STUDY MATERIALS.

