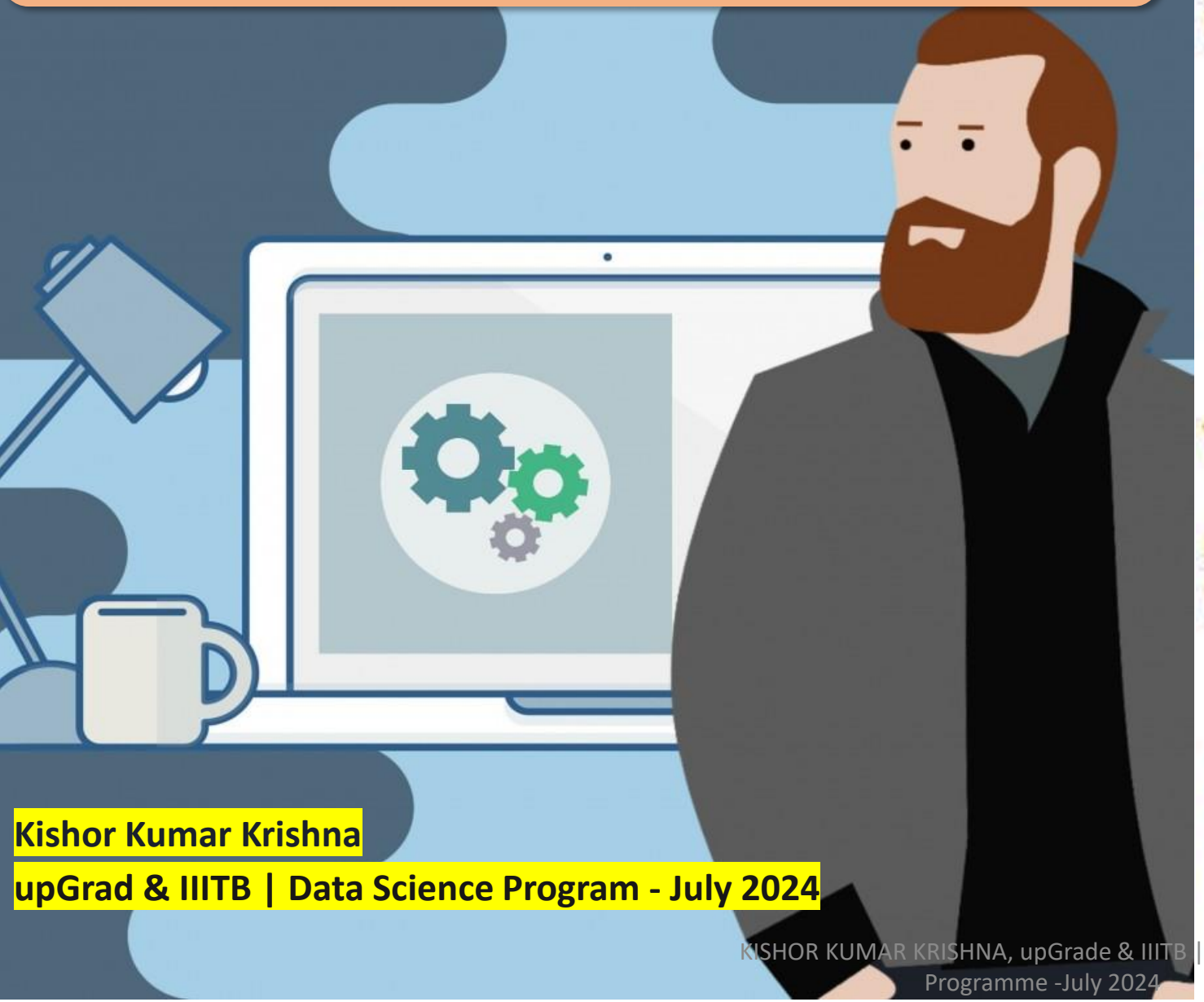
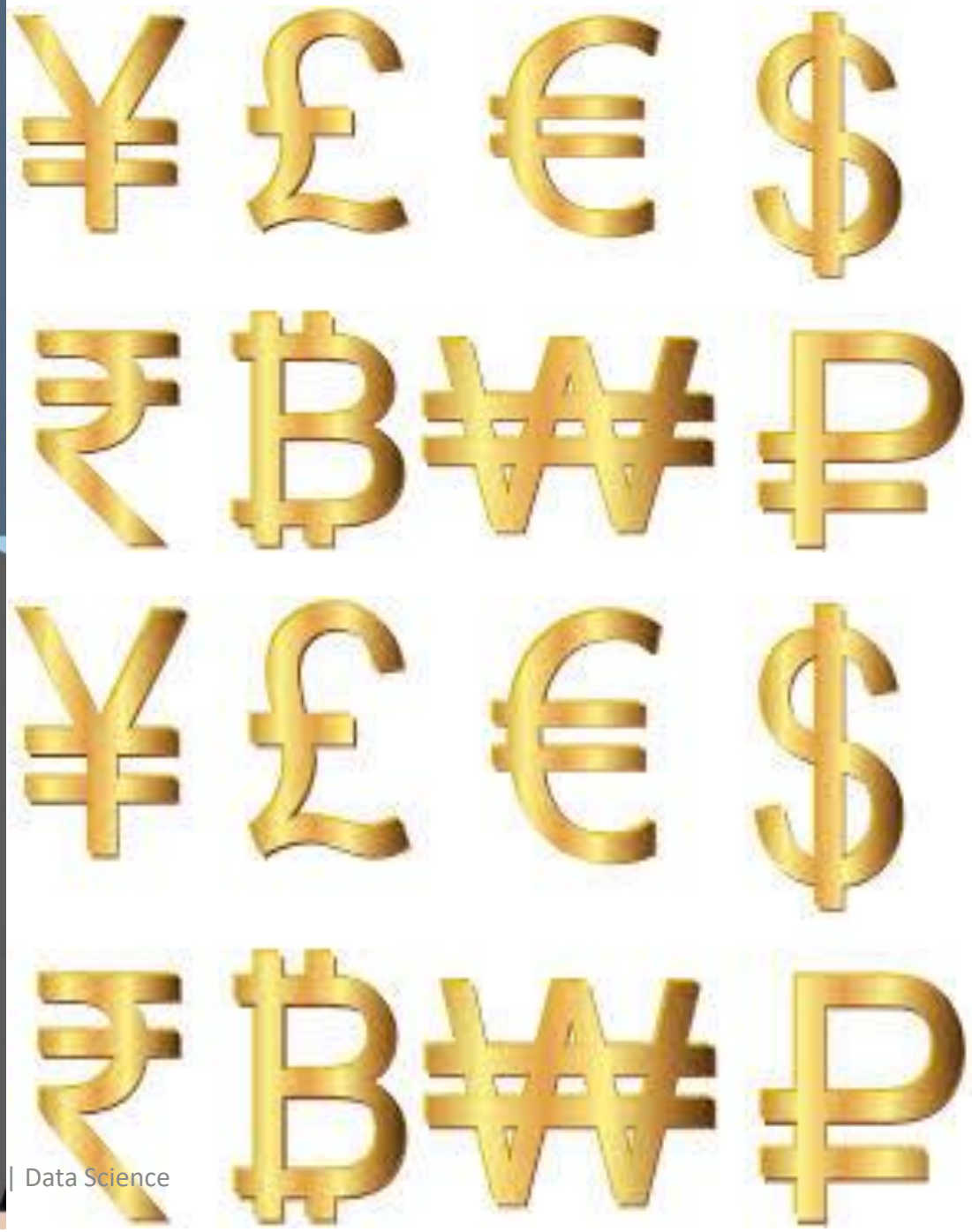
Credit EDA Assignment



Kishor Kumar Krishna

upGrad & IIITB | Data Science Program - July 2024



Introduction

1

Apply EDA in a real business scenario

2

Develop understanding of risk analytics in banking and financial services

3

Use data to minimize the risk of losing money while lending to customers

Business Understanding

- Loan companies struggle with insufficient or non-existent credit history of applicants
- Some consumers exploit this by becoming defaulters
- Use EDA to analyze data patterns to ensure capable applicants are not rejected

Business Objectives

- Identify patterns indicating difficulty in paying installments
- Actions: Deny loan, reduce loan amount, lend at higher interest rates to risky applicants
- Understand driving factors behind loan default for portfolio and risk assessment

Data Understanding and Data-Cleaning.

'application_data.csv': Information of the client at the time of application

- **Checking number of feature and their datatype in data frame and Data Understanding and Data-Cleaning**

1. Application data set contains 121 feature, 1 target variable and 307511 rows
2. We have a total of 49 columns with more than 40% null values and insignificant to our analysis
3. All the columns except AMT_ANNUITY, AMT_GOODS_PRICE, CNT_FAM_MEMBERS, OCCUPATION_TYPE and NAME_TYPE_SUITE seems insignificant
4. Fetch all indicator FLAG columns
5. On further analysis, we can observe there are columns with values of 0/1 or N/Y
6. Here other than FLAG_OWN_CAR and FLAG_OWN_REALTY all other columns seems insignificant further for analysis

Data Understanding and Data-Cleaning.

'application_data.csv': Information of the client at the time of application

- Here other than FLAG_OWN_CAR and FLAG_OWN_REALTY all other columns seems insignificant further for analysis. Hence dropping all other columns from flag_col except FLAG_OWN_CAR and FLAG_OWN_REALTY.

- **Handling Nulls - Filling in appropriate values for analysis**

- **Identifying and handling Outliers**

Data Understanding and Data-Cleaning.

previous application data set':

Information of the client at the time of application

- **previous application data set** contains 37 features and 1670214 rows (out of which contain 15 float64 feature, 6 int64 features, and 16 feature are object data type)

Data Understanding and Data-Cleaning.

previous application data set':

Information of the client at the time of application

Observations

There are 16 features in prev application data frame that have missing values

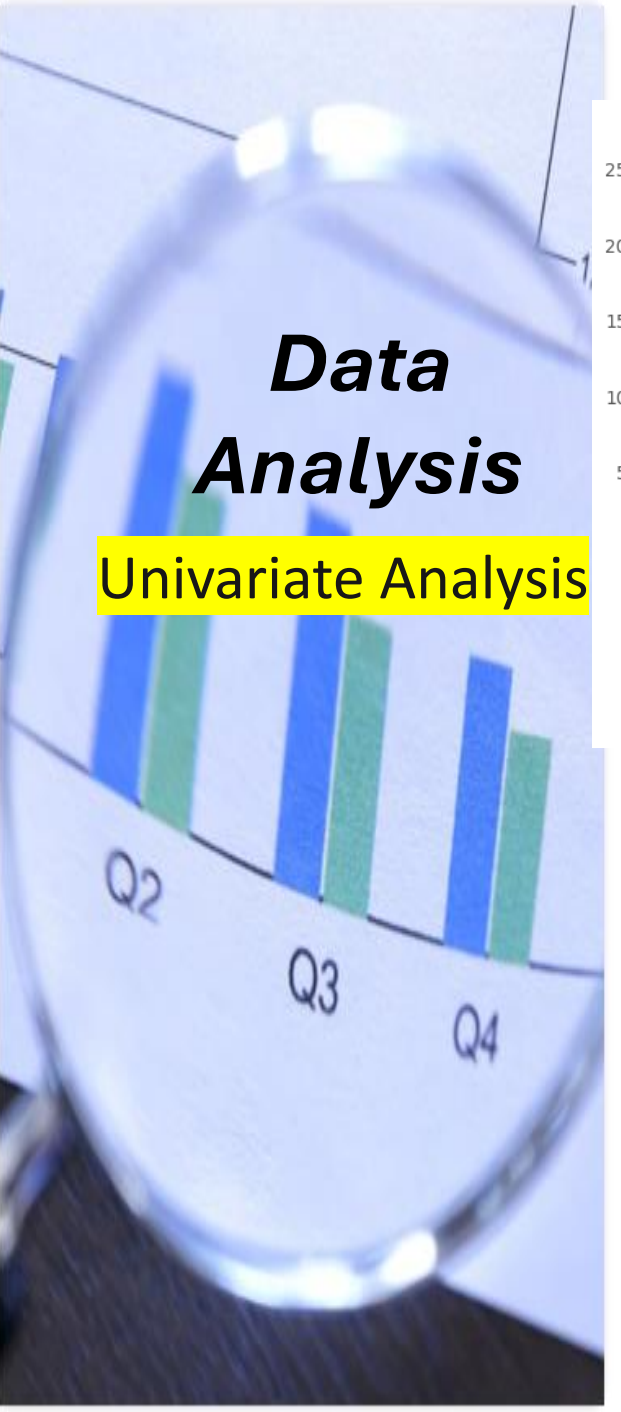
A significant number of features have missing values, with some exceeding 50% missing data. Permanently dropping the features (RATE_INTEREST_PRIMARY ,RATE_INTEREST_PRIVILEGED) as 99% data is missing.

The features 'AMT_GOODS_PRICE', 'RATE_DOWN_PAYMENT', 'RATE_INTEREST_PRIMARY', 'RATE_INTEREST_PRIVILEGED', 'NAME_TYPE_SUITE', 'AMT_ANNUITY' have relatively high missing values, indicating potential issues with data collection or recording for those specific variables.

The 'DAYS_FIRST_DRAWING', 'DAYS_FIRST_DUE', 'DAYS_LAST_DUE_1ST_VERSION', 'DAYS_LAST_DUE', 'DAYS_TERMINATION' features related to loan terms and repayment schedules also have substantial missing values, suggesting either incomplete or inconsistent data entry.

Dropping rows containing missing values for the features (AMT_CREDIT and PRODUCT_COMBINATION) for vary low % of missing data. Dropping entices would not cause impact the analysis as percentage of missing value is very low (-2%)

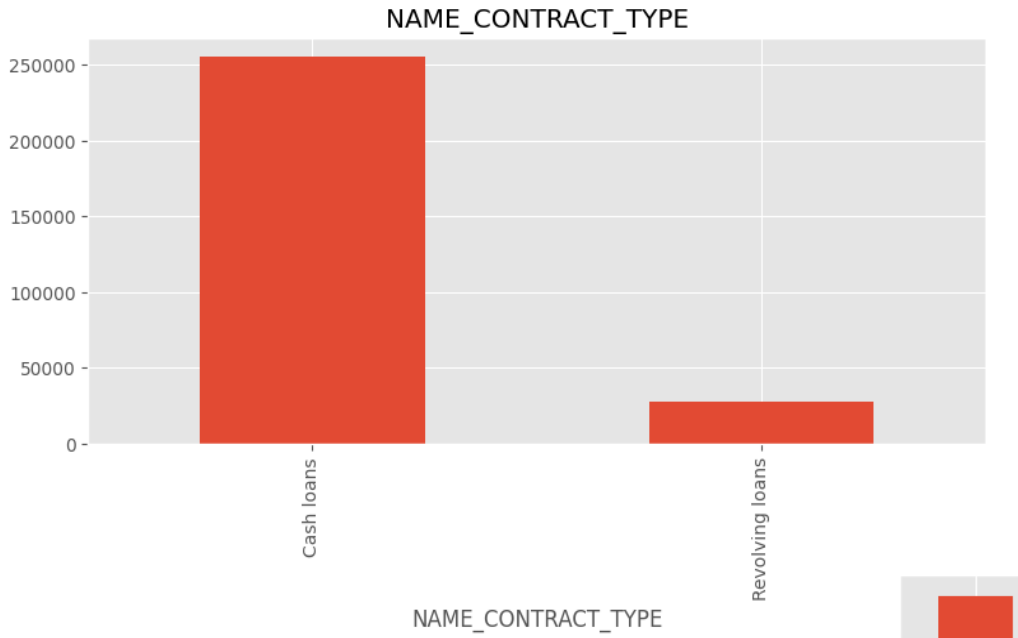
The presence of missing values can pose challenges for analysis and modelling. Depending on the extent of missing data and the nature of the analysis, techniques like imputation or removal of rows/columns with missing data may be necessary.

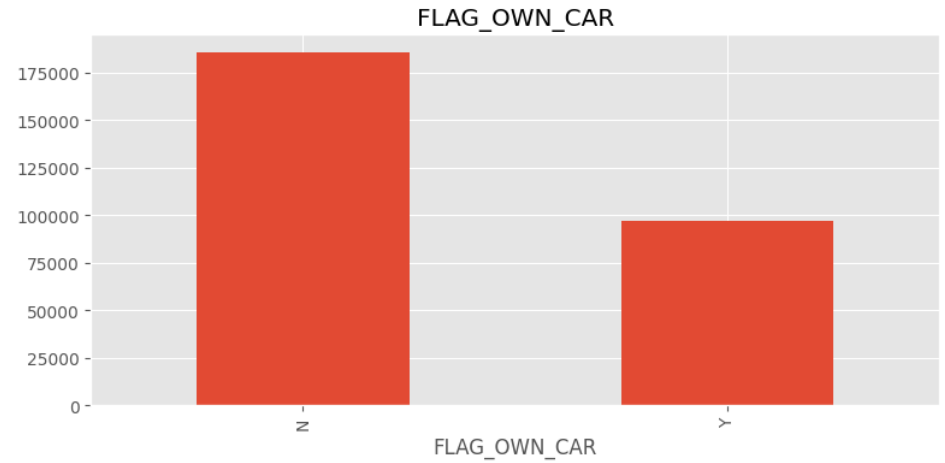


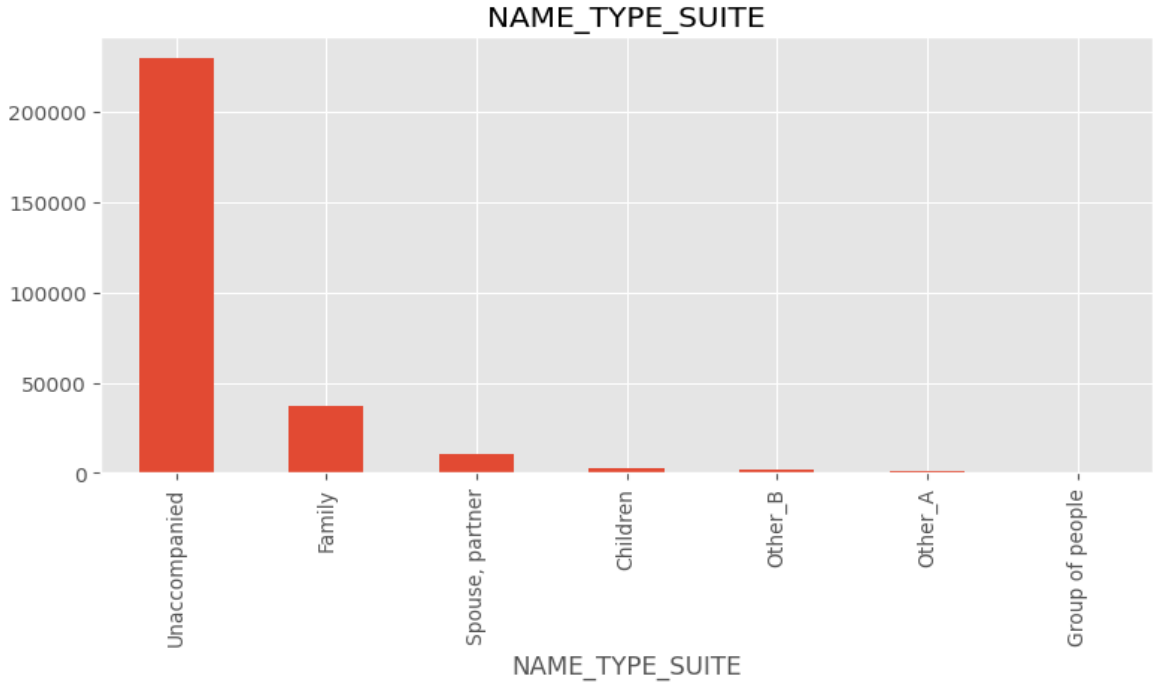
Data Analysis

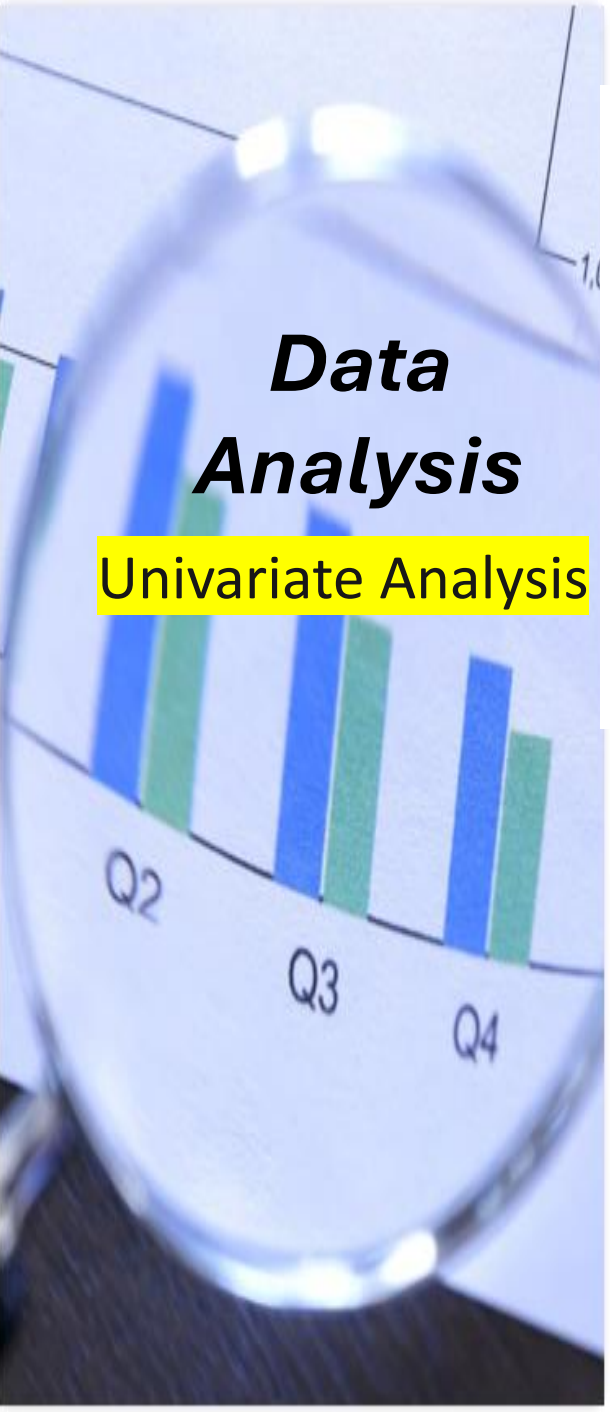
Univariate Analysis

Categorical variables





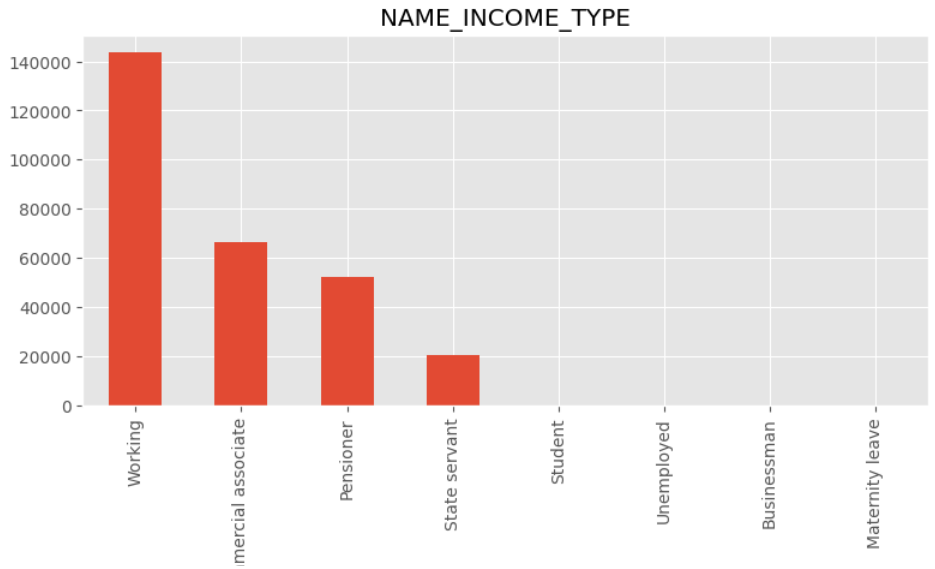


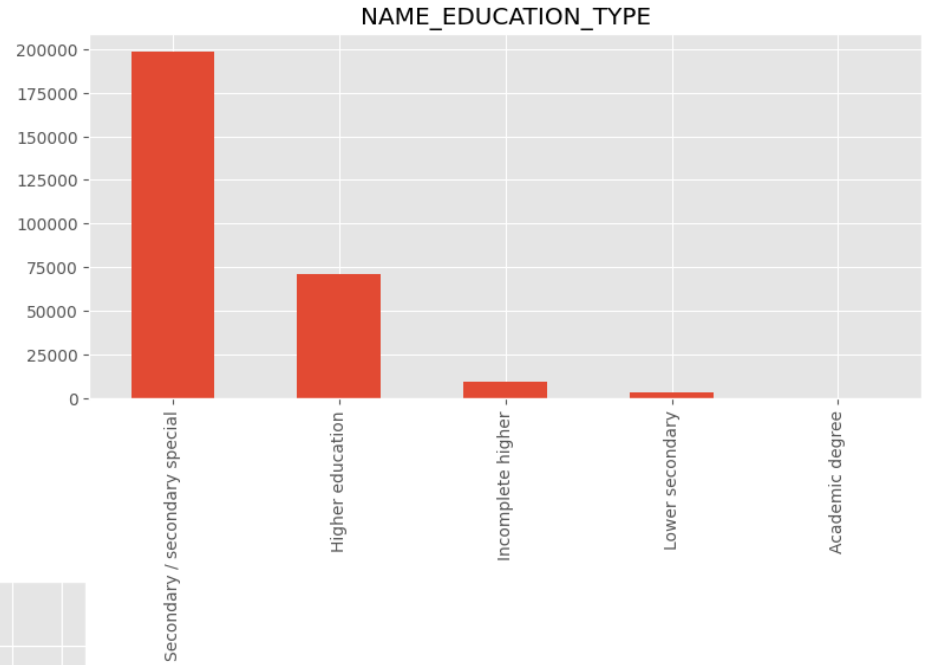


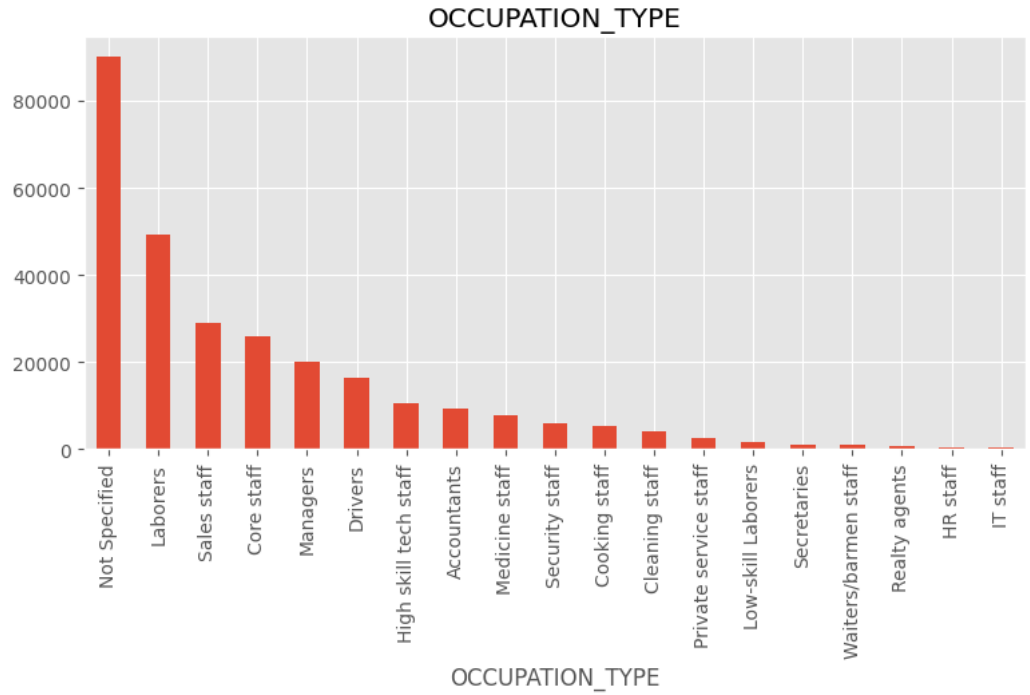
Data Analysis

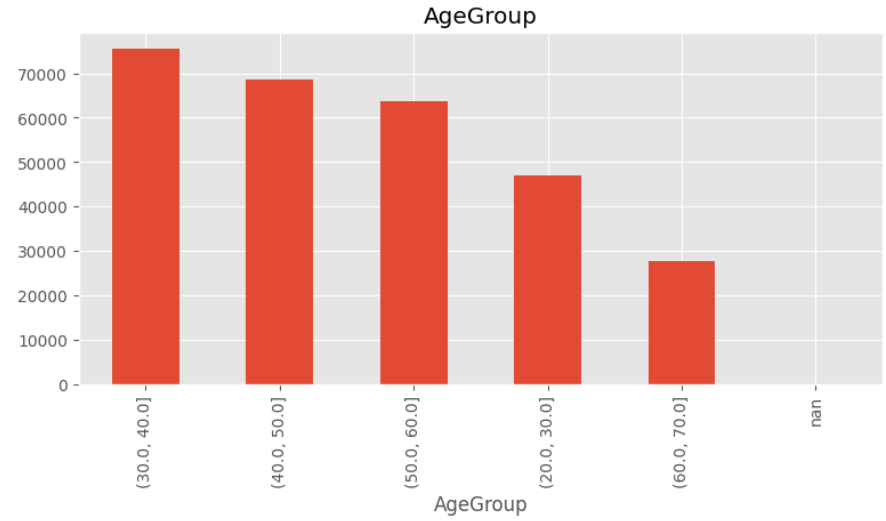
Univariate Analysis

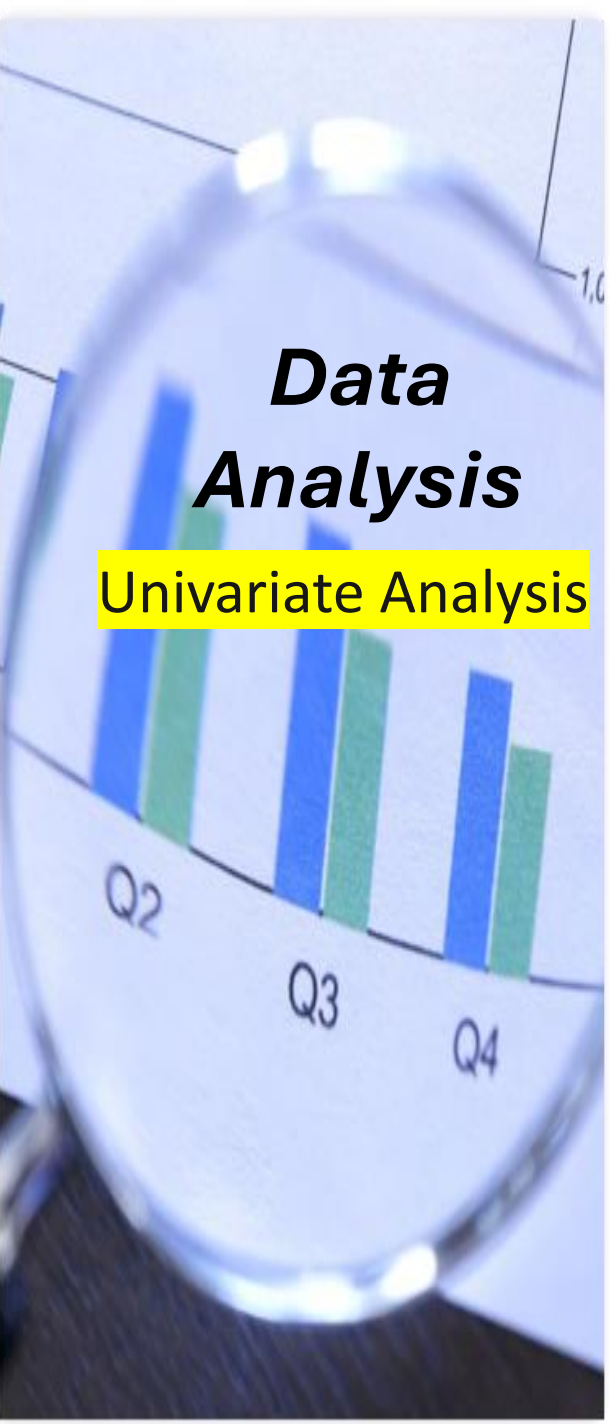
Categorical variables





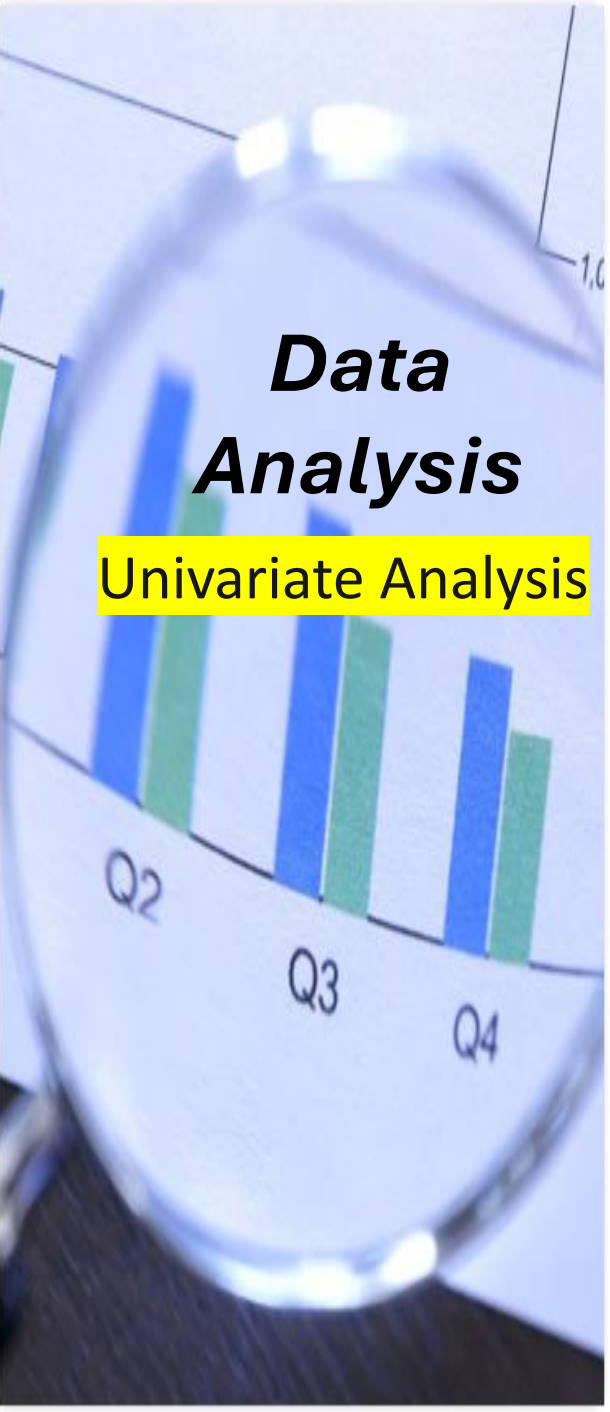






Inferences from Univariate analysis of Categorical variables Applicants making payments on time.

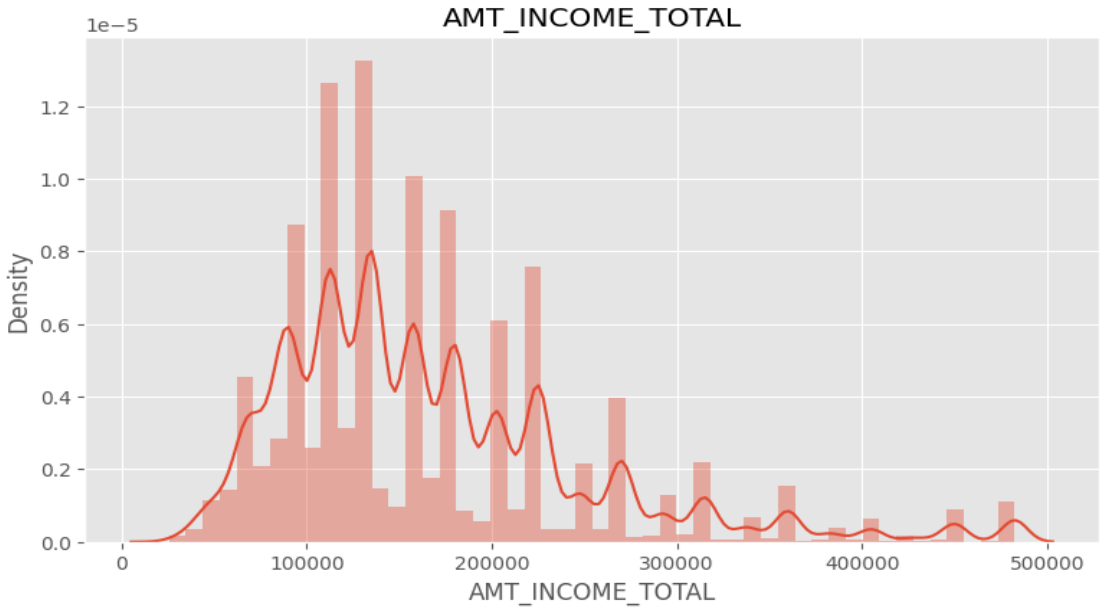
1. More than 2,50,000 applicants have applied for Cash loans and very small proportion of people have applied for Revolving loans.
2. More than 175000 loan applicants are female and slightly less than 100000 loan applicants are male and very few(4) applicants have third or unknown gender.
3. More than 175000 loan applicants doesn't own a car and little less than 100000 applicants own a car.
4. Slightly less than 200000 loan applicants own a house or a flat where as less than 100000 people doesn't own a house.
5. A large number of people i.e; almost 250000 people who applied for loans were unaccompanied. The difference between the most and the second most category i.e; family is a lot and the rest are almost negligible.
6. Top 3 categories of people who applied for loans were getting income by working, Commercial associates or were pensioners, highest being the Working class category.
7. People with Secondary or secondary special education applied the highest number of loans followed by Higher educated people.
8. Mostly married people applied for loans followed by Single people with a huge difference.
9. People having their own houses or apartments applied for the most number of loans followed by other categories with few numbers.
10. Most of the people who opted for loan, didn't mention their occupation type. Other high number of people who applied for loans are labourers, Sales Staff and Core staff.
11. Apart from Weekends, every day has almost equal distribution for loan application with Sunday being noticably least and Tuesday being the most.
12. Most of the people who applied for loans either work in Business Entity type 2 kind of organizations and the other top 2 are unknown types and Self-employed respectively.
13. People falling under the age group of 30-40, 40-50 and 50-60 are most likely to apply for loan.
14. There are relatively few people with very high income or average who have applied for loans but most of them who have applied for loans either have high income or low income.

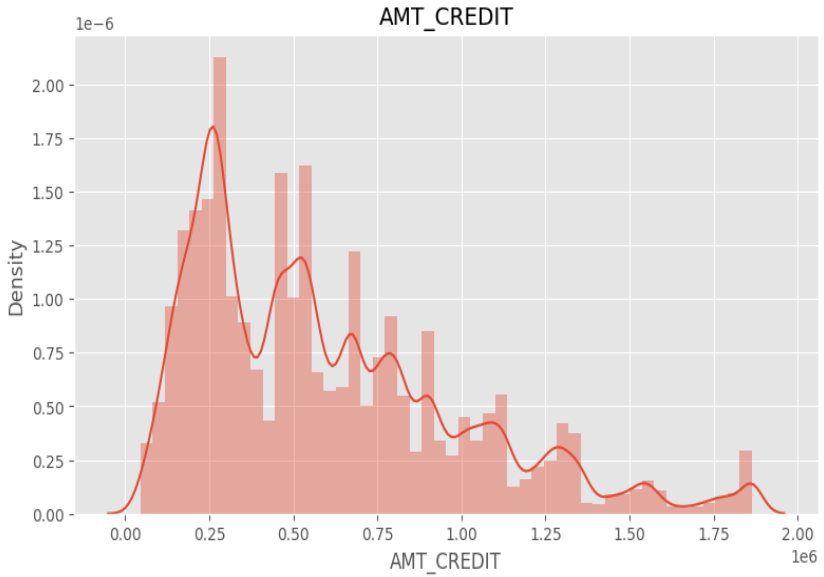


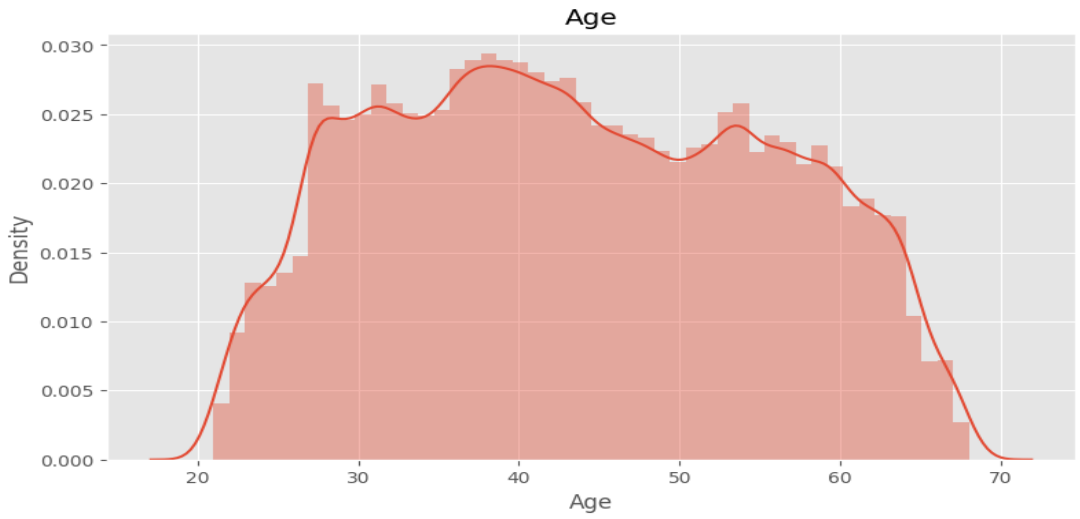
Data Analysis

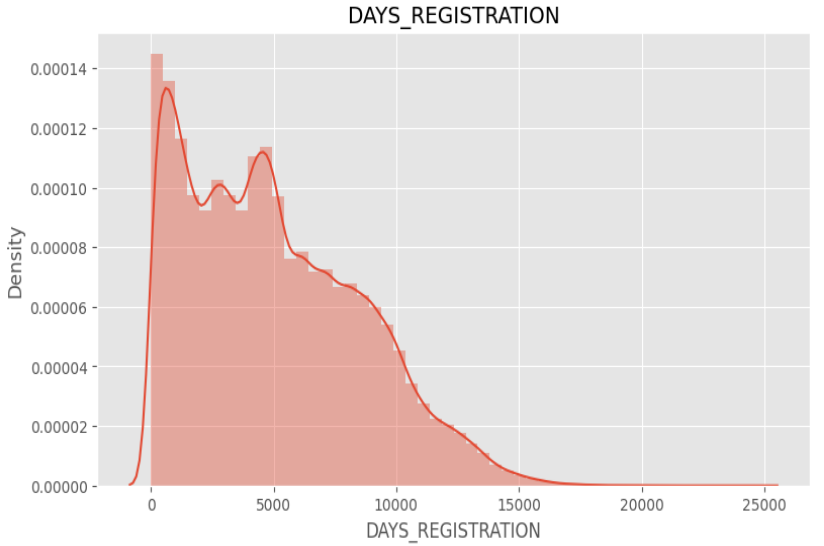
Univariate Analysis

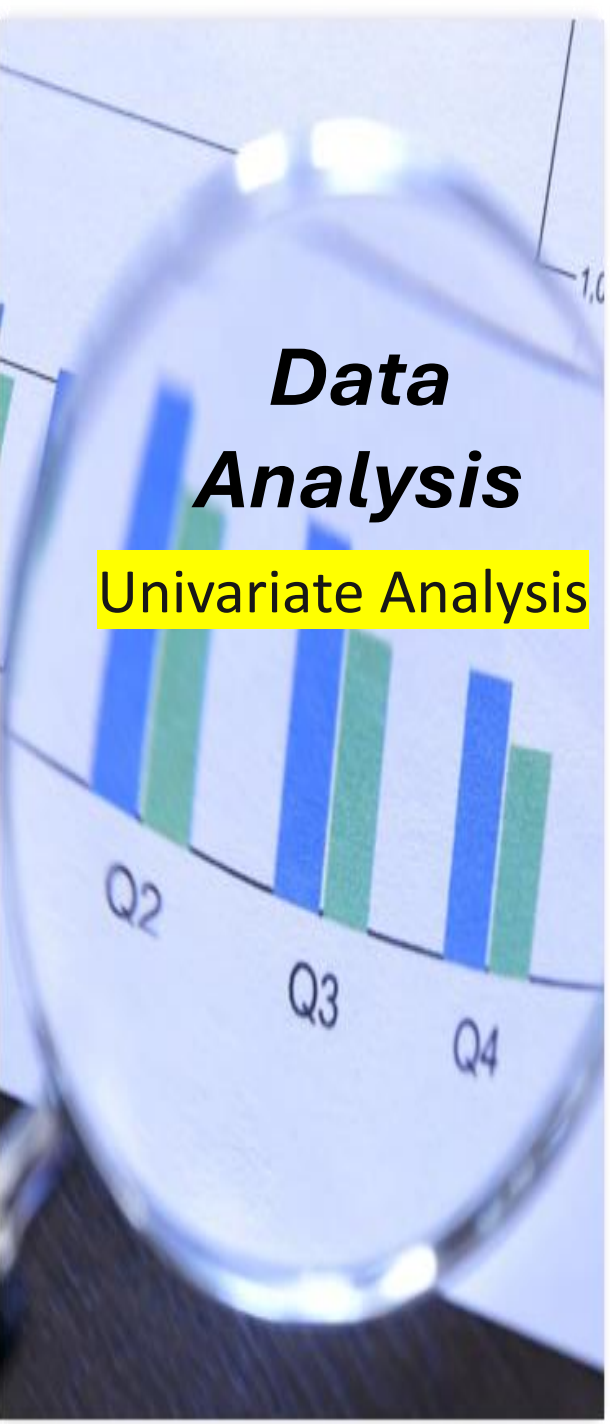
NUMERIC_COLS





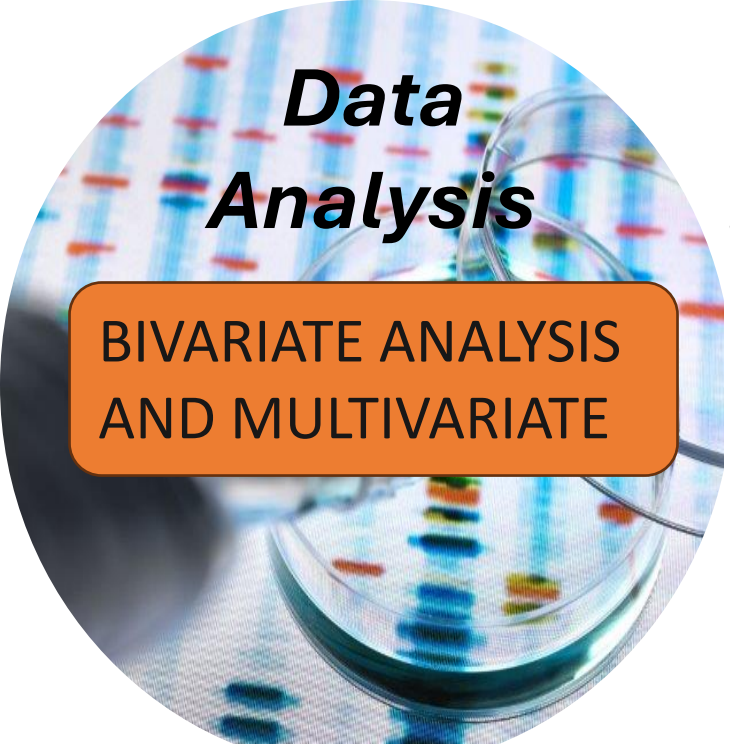


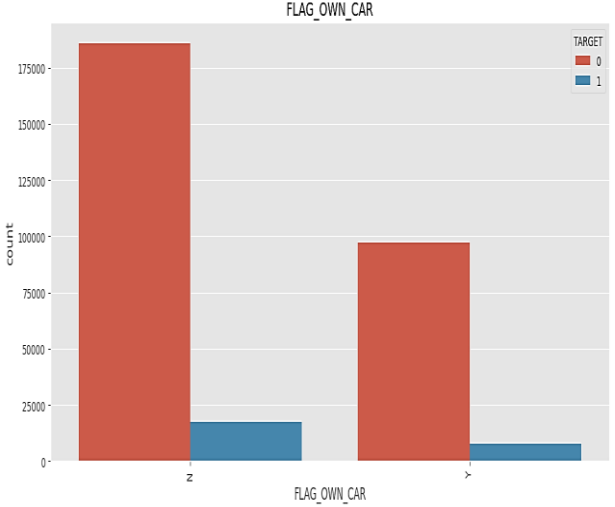


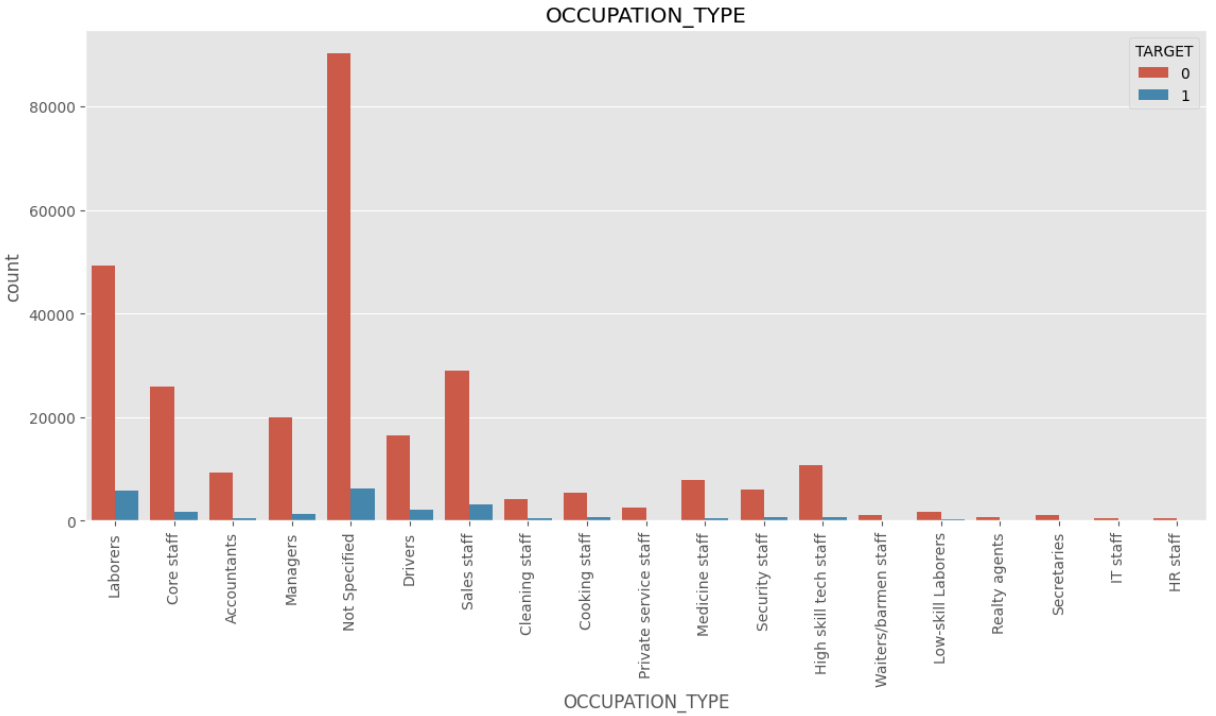


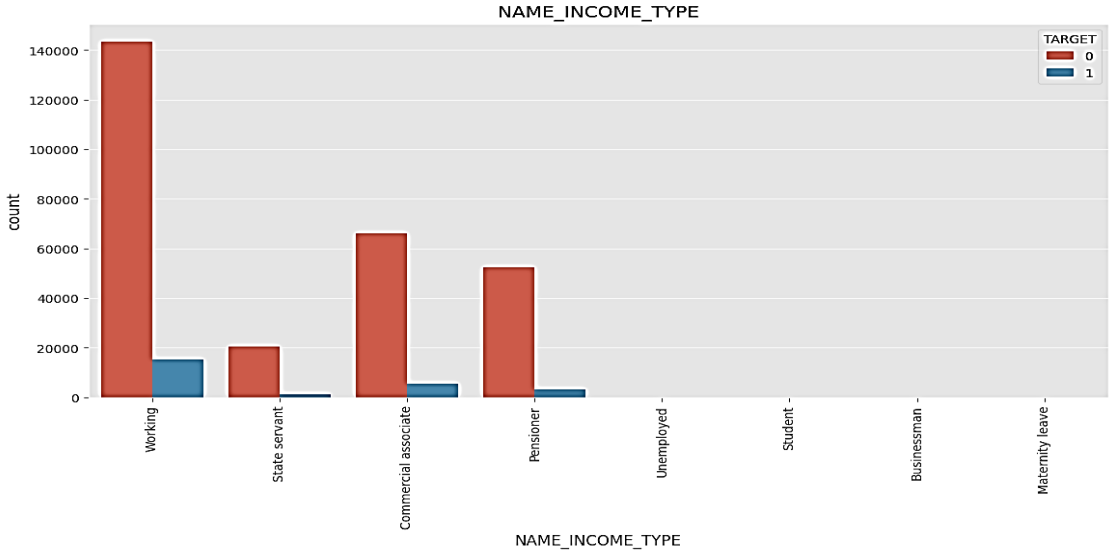
Inferences from Univariate analysis of Continuous variables of Applicants making payments on time.

1. People with 0 children are much more than people with few or more than 5 children which shows that people who are applying for loan have less people dependent financially on them which will lessen the loan payment difficulties.
2. The density of people earning between 1,00,000-2,20,000 are more likely to apply for the loans and pay them on time. There is a skew to the right which also shows people with higher salaries present as well.
3. The maximum density of loan applied by people is between 0.045x10^6-0.053x10^6 with a right skew because of the presence of 1.864000e+06. Loans with such high credits are also paid.
4. The KDE for AMT_ANNUITY almost resembles a normal distribution with a right skew because of values with more than 34749.000000 in small amounts and the maximum value 71006.500000 in it.
5. There's no pattern for AMT_GOODS_PRICE, REGION_POPULATION_RELATIVE.
6. For column DAY_EMPLOYED, value more than 12780 makes no sense because that is equivalent to 35 years. But there are many rows with values more than 12780. Hence nothing can be derived from this column either.
7. The density of applicants changing registration between 0-5000 is the most and the density of applicants who changed the identity document with which he applied between 4000-4600 is the most.
8. The applicants mostly have 2 family members.
9. REGION_RATING_CLIENT, REGION_RATING_CLIENT_W_CITY for most of the applicants is 2.
10. Most of the applicants have applied between 10 AM-12:30 PM
11. All the columns REG_REGION_NOT_LIVE_REGION, REG_REGION_NOT_WORK_REGION, LIVE_REGION_NOT_WORK_REGION, REG_CITY_NOT_LIVE_CITY, REG_CITY_NOT_WORK_CITY, LIVE_CITY_NOT_WORK_CITY which checks the address given by the applicant is matching in most of the cases here.
12. Age is almost evenly distributed with maximum density between 35-43.

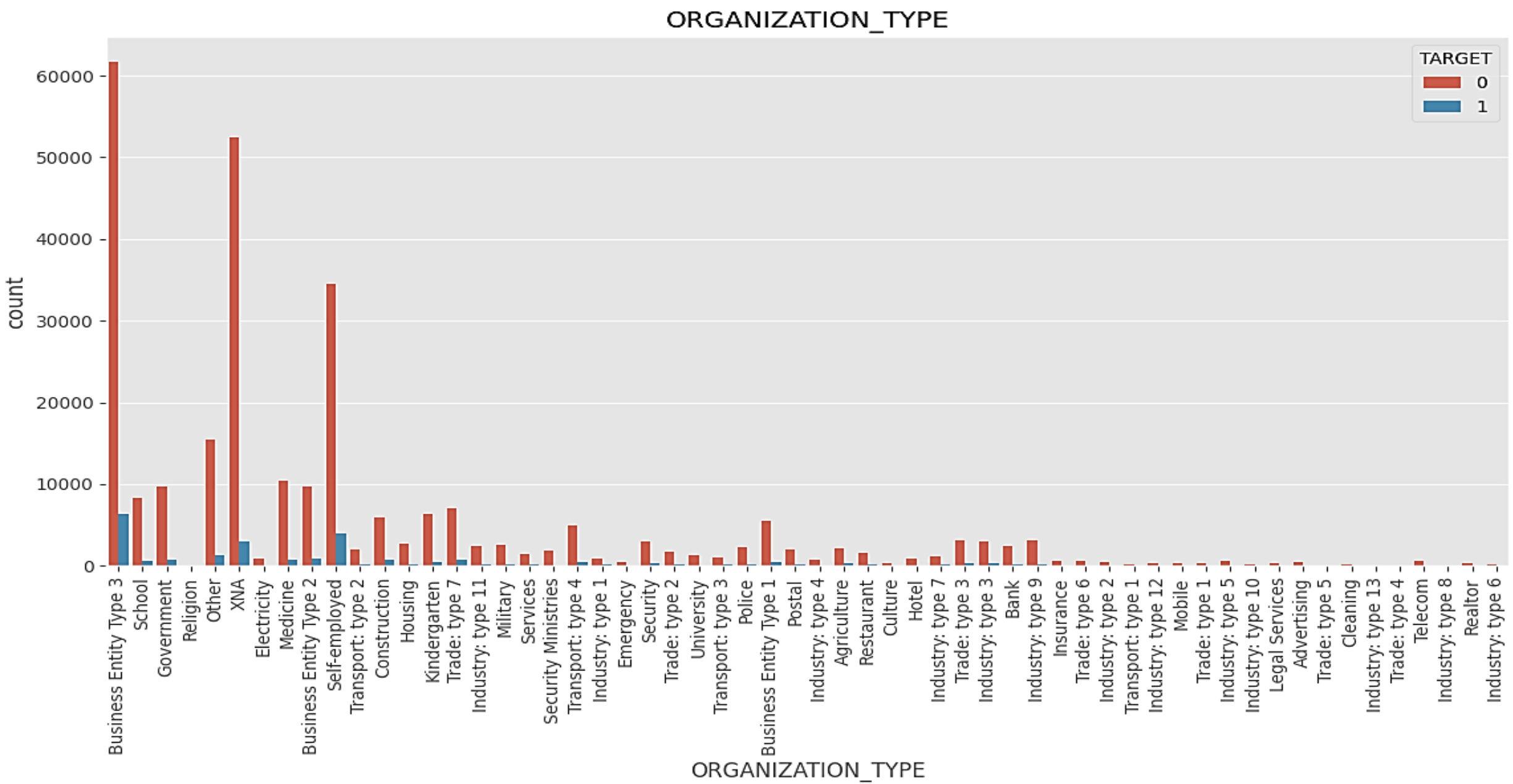


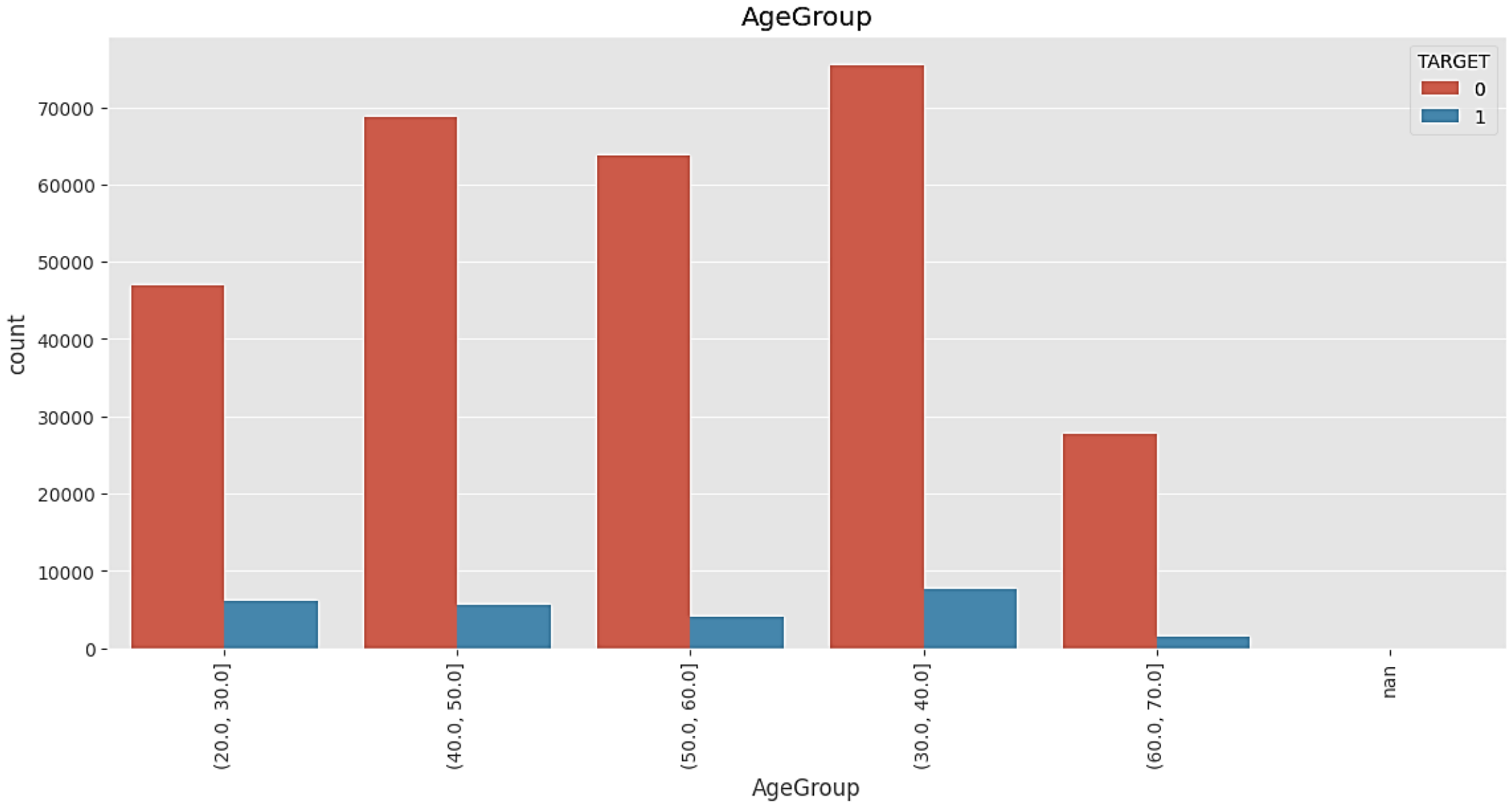


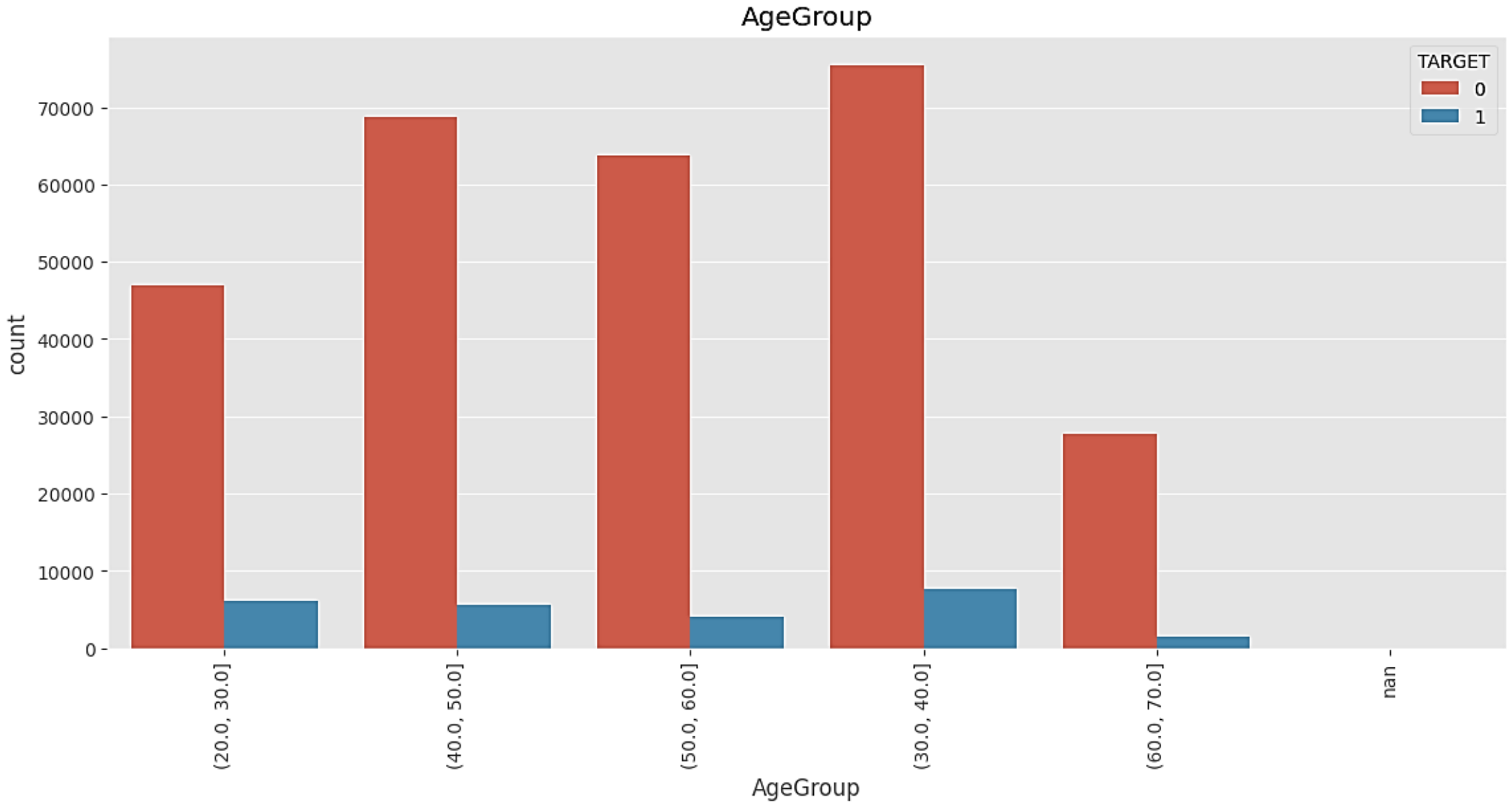


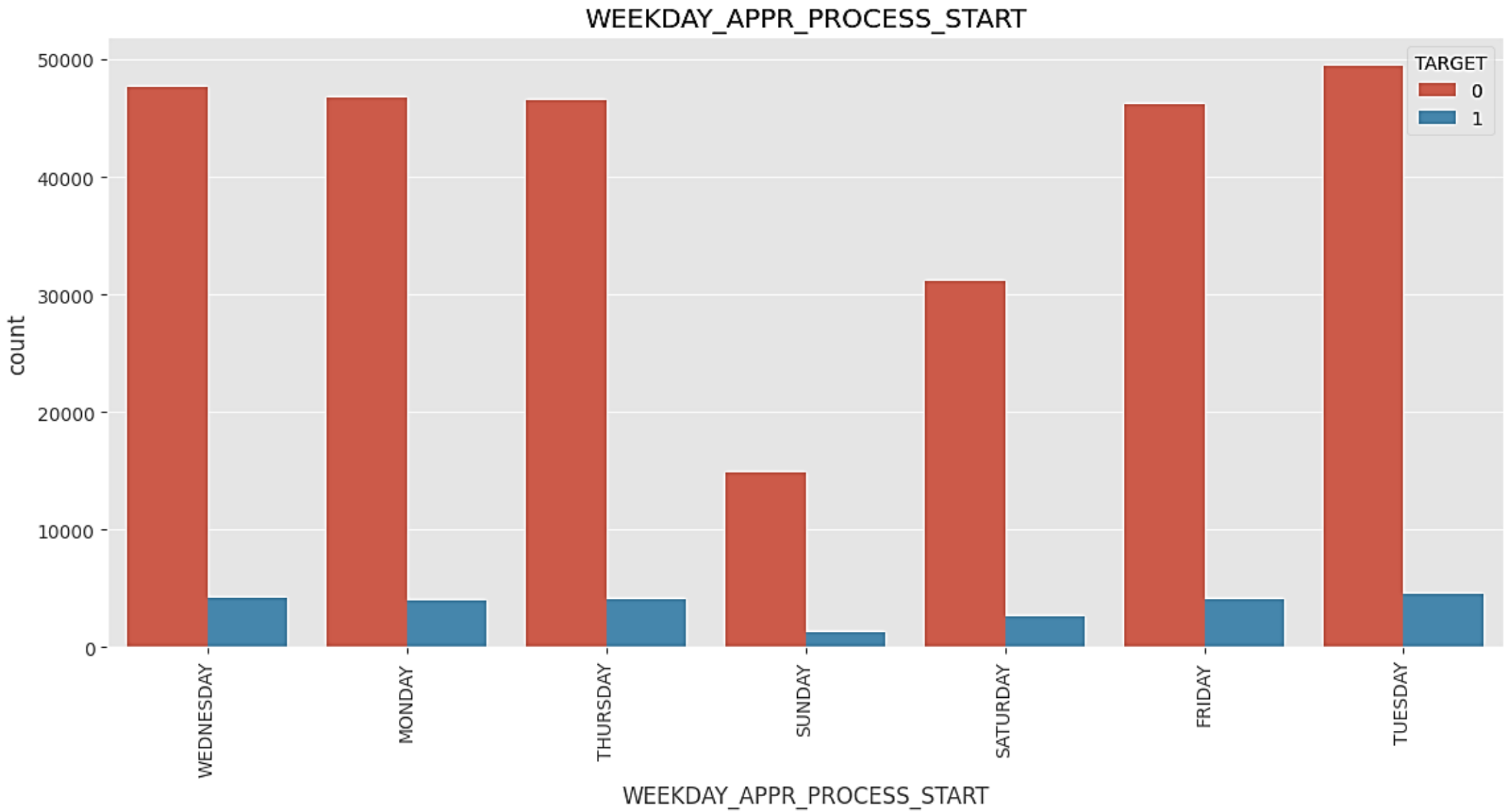


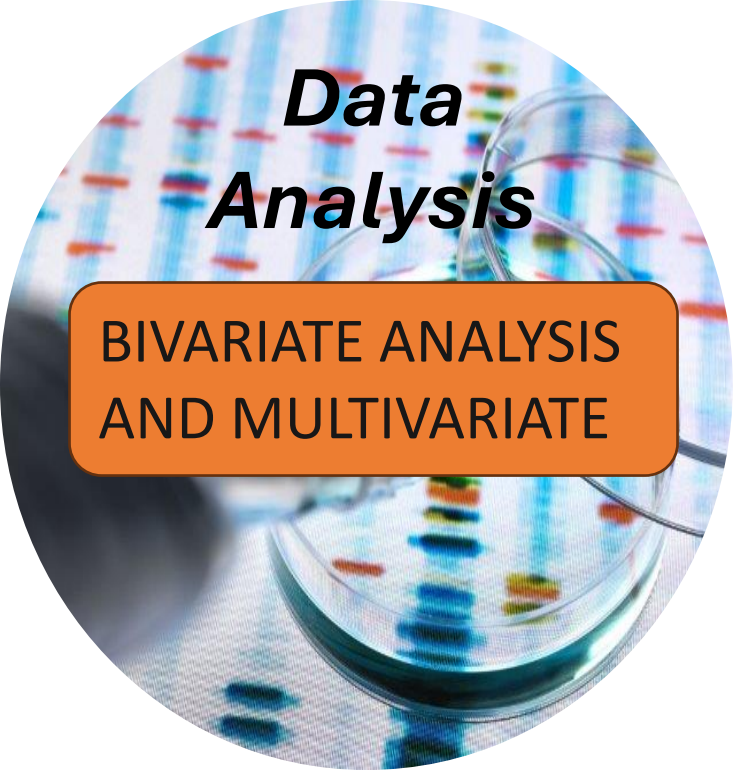
KISHOR KUMAR KRISHNA, upGrade & IIITB | Data Science Programme -July 2024


ORGANIZATION_TYPE
TARGET
0
1
count
60000
50000
40000
30000
20000
10000
0
Business Entity Type 3
School
Government
Religion
Other
XNA
Electricity
Medicine
Business Entity Type 2
Self-employed
Transport: type 2
Construction
Housing
Kindergarten
Trade: type 7
Industry: type 11
Military
Services
Security Ministries
Transport: type 4
Industry: type 1
Emergency
Security
Trade: type 2
University
Transport: type 3
Police
Business Entity Type 1
Postal
Industry: type 4
Agriculture
Restaurant
Culture
Hotel
Industry: type 7
Trade: type 3
Industry: type 3
Bank
Industry: type 9
Insurance
Trade: type 6
Industry: type 2
Transport: type 1
Industry: type 12
Mobile
Trade: type 1
Industry: type 5
Industry: type 10
Legal Services
Advertising
Trade: type 5
Cleaning
Industry: type 13
Trade: type 4
Telecom
Industry: type 8
Realtor
Industry: type 6
ORGANIZATION_TYPE


AgeGroup
TARGET
0
1
70000
60000
50000
40000
30000
20000
10000
0
count
(20.0, 30.0]
(40.0, 50.0]
(50.0, 60.0]
(30.0, 40.0]
(60.0, 70.0]
nan
AgeGroup


AgeGroup
TARGET
0
1
70000
60000
50000
40000
30000
20000
10000
0
count
(20.0, 30.0]
(40.0, 50.0]
(50.0, 60.0]
(30.0, 40.0]
(60.0, 70.0]
nan
AgeGroup

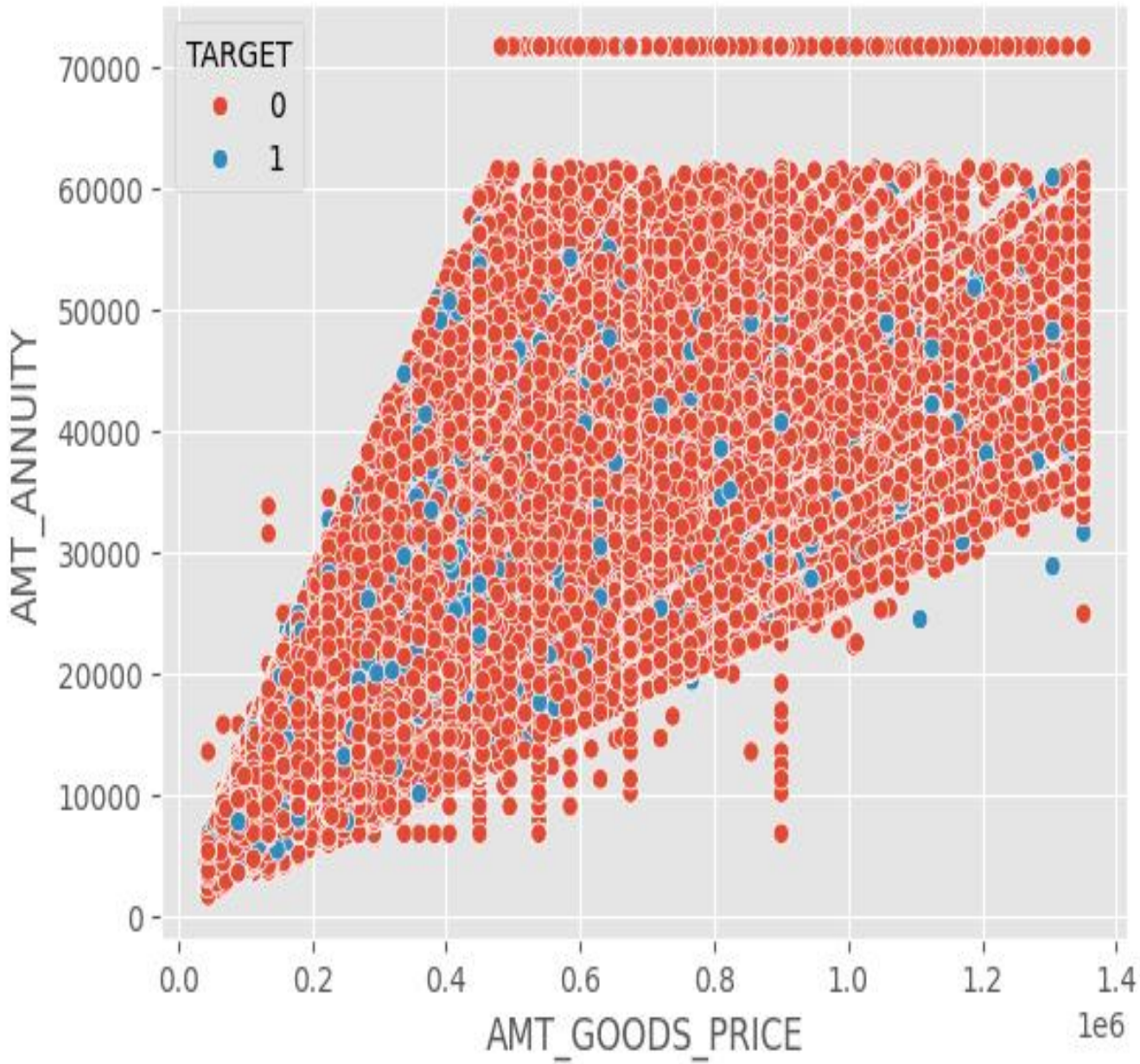

WEEKDAY_APPR_PROCESS_START
TARGET
0
1
count
50000
40000
30000
20000
10000
0
WEDNESDAY
MONDAY
THURSDAY
SUNDAY
SATURDAY
FRIDAY
TUESDAY
WEEKDAY_APPR_PROCESS_START

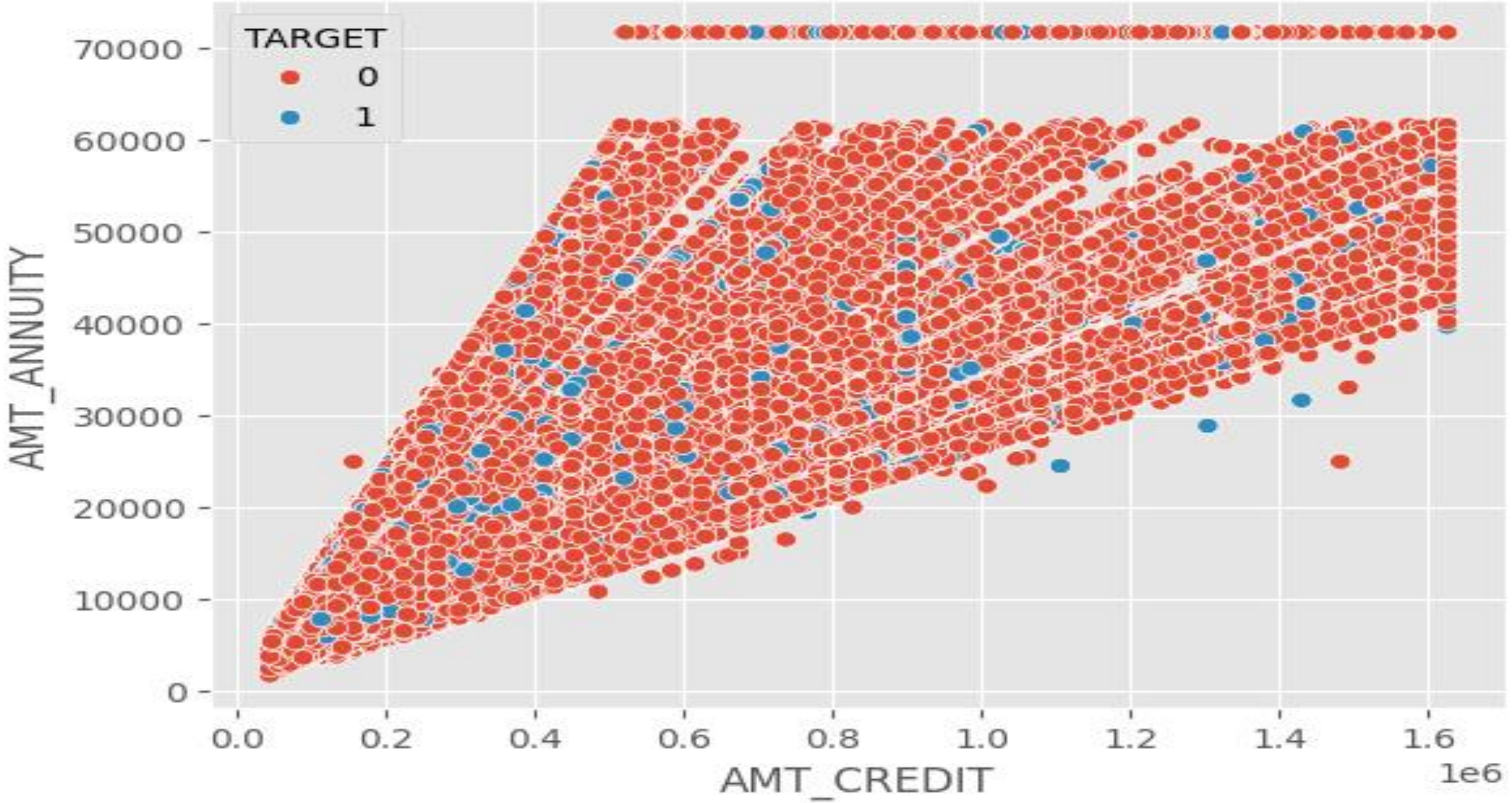


Inferences:

1. The proportion of people opting out for Cash loans and paying the amount back is more than the people opting for revolving loans.
2. The proportion of males applying for loans and having difficulties in payment is much more than females.
3. The applicants having difficulties to pay back the loans mostly come unaccompanied while applying.
4. The number of people working for income are more than any other category. But the number of people having difficulty to pay the more are also from working category people. There are very negligible amount of applicants who are unemployed, student, business man or are on maternity leave who have applied for loans or who have difficulties to pay.
5. Applicants who are Secondary or special secondary educated have applied most of the loans but are also the population facing most of the difficulties while paying the amount.
6. Married people apply for the most number of loans but tend to have difficulties in the payment as well.
7. Labourers and applicants who have not specified their occupation type have some history to face difficulties while paying back the loan and least proportion of people whi have applied for loans and have the least difficulties are IT staff, HR staff etc.
8. People belonging to organizations like Business entity type 3, unknown organization or are self- employed have applied for the most number of loans respectively and most number of people having propblems in paying back the amount is Business entity type 3, are self- employed or unknown organization respectively.
9. People in the age group 30-40, 40-50 and 50-60 have applied for the most number of loans but applicants in the age group 30-40 have difficulties to pay back the loans.
10. People with high income and low income have applied for the most number of loans(Loans applied by people with high income is much more than the applicants with low income) but people with low income and high income has almost same number of people facing difficulties to pay the loan which means the proportion of people having low income and applying for loans have most difficulties to pay the amount.


1e6
TARGET
0
1
AMT_CREDIT
AMT_GOODS_PRICE
1e6


TARGET
0
1
AMT_ANNUITY
AMT_GOODS_PRICE
1e6

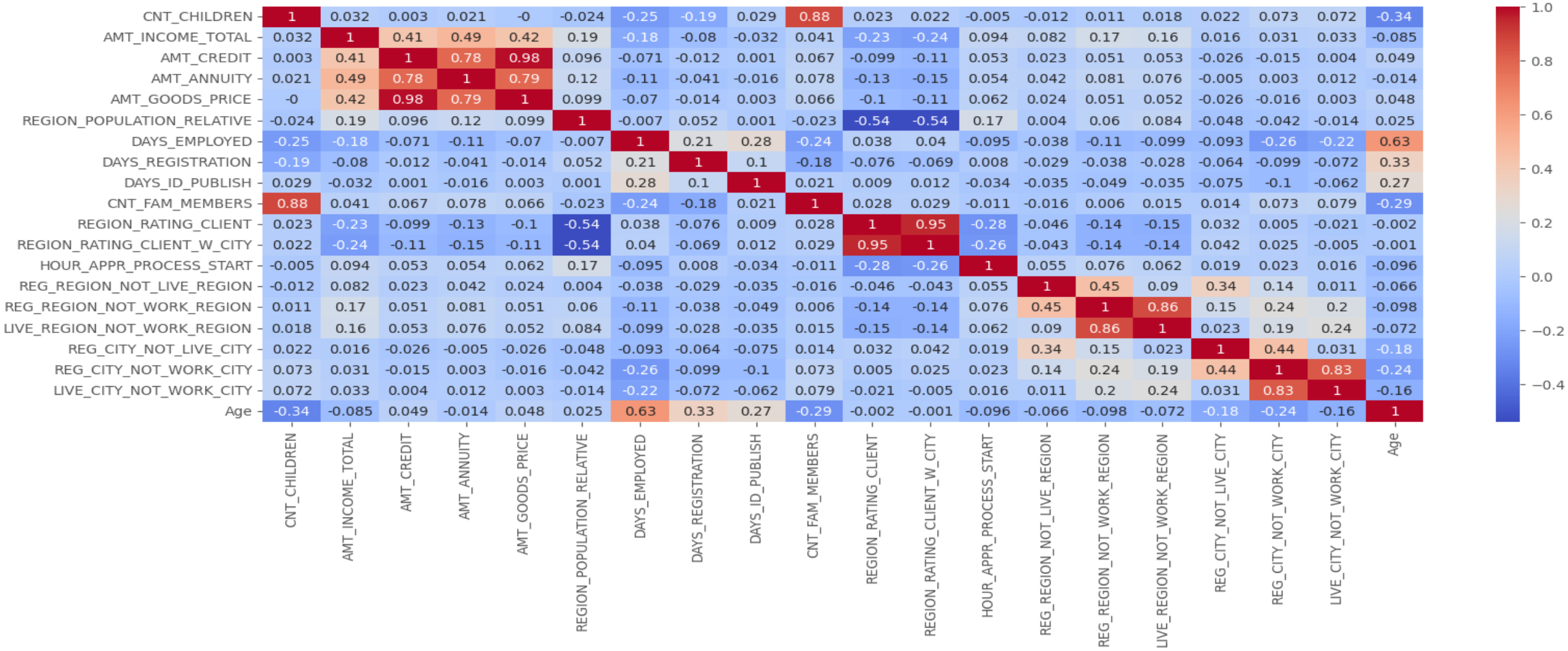

TARGET
0
1
70000
60000
50000
40000
30000
20000
10000
0
AMT_ANNUITY
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1e6
AMT_CREDIT

Inference: On saving the pair plot as an image and zooming each graph in it, following inferences are developed. Also inferences derived from the above bar graphs are also mentioned below:

1. With increasing count of children, the applicant starts facing payment problem irrespective of income, credit amount, annuity and goods price with an exception if the salary of applicant is very high and age is relatively high.
2. It becomes difficult for applicant to pay if the 'AMT_CREDIT and AMT_ANNUITY' , 'AMT_CREDIT and AMT_Goods_price' , 'AMT_ANNUITY AND AMT_GOODS_PRICE' rises together.
3. 'AMT_CREDIT and AMT_ANNUITY' , 'AMT_CREDIT and AMT_Goods_price' , 'AMT_ANNUITY AND AMT_GOODS_PRICE' have a rising relation if not completely linear relation with few exceptions.
4. Larger the DAYS_REGISTRATION shows larger chances of getting payment of loan irrespective of AMT_INCOME_TOTAL,AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE.
5. An increasing relationship is established between AMT_GOODS_PRICE and AMT_CREDIT.
6. Less family members and more days of registration leads to easy payment of loan amount. Further more CNT_FAM_MEMBERS is not related to other variable.
7. For every value of Region rating by client and REGION_RATING_CLIENT_W_CITY more DAYS_REGISTRATION indicates assured payment of loans.
8. Age doen't matter for owning a car or a house/apartment.
9. Mostly people with more age who have applied for loan are either widowed or separated.

Correlation between Target and other variables

SK_ID_CURR since calculating correlation for the id makes no sense. TARGET: Since we have divided dataframes as per target 0 and 1. Hence the variance of target will be zero which gives NaN while calculating corr.



Correlation between Target and other variables

AMT_GOODS_PRICE	AMT_CREDIT	0.99
AMT_CREDIT	AMT_GOODS_PRICE	0.99
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.95
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.95
CNT_CHILDREN	CNT_FAM_MEMBERS	0.88
CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.86
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.86
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.83
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.83
AMT_GOODS_PRICE	AMT_ANNUITY	0.79
AMT_ANNUITY	AMT_GOODS_PRICE	0.79
AMT_CREDIT	AMT_ANNUITY	0.78
AMT_ANNUITY	AMT_CREDIT	0.78
DAYS_EMPLOYED	Age	0.63
Age	DAYS_EMPLOYED	0.63
REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT	0.54
REGION_RATING_CLIENT_W_CITY	REGION_POPULATION_RELATIVE	0.54
REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	0.54
REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT_W_CITY	0.54

The top 10 correlation pairs for applicants who made their payments on time -

AMT_GOODS_PRICE AMT_CREDIT 0.99
REGION_RATING_CLIENT REGION_RATING_CLIENT_W_CITY 0.95
CNT_CHILDREN CNT_FAM_MEMBERS 0.88
LIVE_REGION_NOT_WORK_REGION * *
REG_REGION_NOT_WORK_REGION 0.86
REG_CITY_NOT_WORK_CITY LIVE_CITY_NOT_WORK_CITY 0.83
AMT_CREDIT AMT_ANNUITY 0.79
AMT_GOODS_PRICE AMT_ANNUITY 0.79
DAYS_EMPLOYED Age 0.63
REGION_RATING_CLIENT REGION_POPULATION_RELATIVE 0.54
REGION_RATING_CLIENT_W_CITY REGION_POPULATION_RELATIVE 0.54

Overall Recommendations for Loan Approval Process

Overall Recommendations for Loan Approval Process

- Efficiency Problem: The ratio of timely payers to those having difficulties is almost the same for both approved and refused loans
- Revolving Loans: The bank has lost profits by not approving revolving loans to reliable applicants
- High-Risk Applicants: Males with high income and secondary education are more likely to default
- Low-Risk Applicants: Approve more loans to applicants with high annual income, secondary or higher education, and lower credit amounts
- Reapplying Customers: Many loans applied by existing customers were previously refused