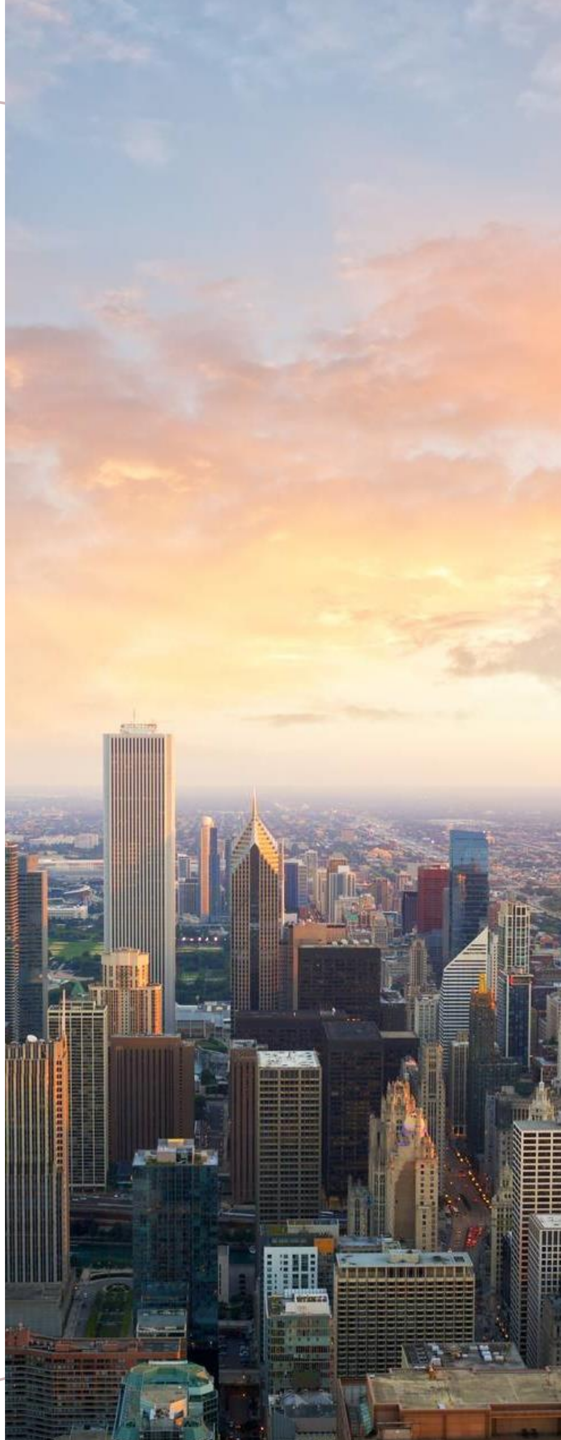
OESON DATA ANALYTICS TRAINING AND INTERNSHIP PROJECT 3:

LOAN DEFAULT ANALYSIS; EDA AND SIMPLE REGRESSION MODEL USING PYTHON LIBRARIES

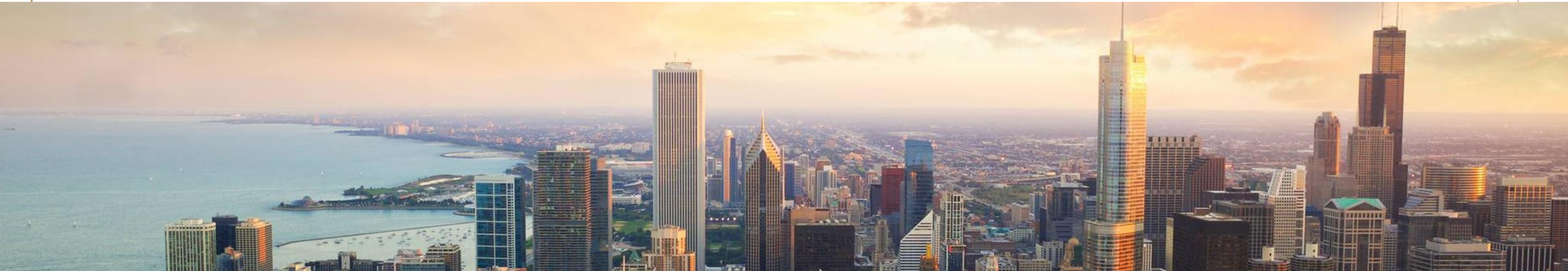
AGENDA

- INTRODUCTION TO DATASET
- DATA PREPERATION AND CLEANING
- DATA ANALYSIS AND VISUALISATION
- REGRESSION AND STATISTICAL ANALYSIS
- INSIGHTS



PROJECT:

- THIS EDA PROJECT AIMS TO IDENTIFY PATTERNS WHICH INDICATE IF A CLIENT HAS DIFFICULTY IN PAYING THEIR LOAN INSTALLMENTS WHICH MAY BE USED FOR TAKING ACTIONS BY THE BANK 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 BY THE BANK. IDENTIFICATION OF SUCH APPLICANTS USING EDA IS THE AIM OF THIS PROJECT. (SOURCE: OESON)



DATA PREPARATION AND CLEANING

- Data cleaning involves several key steps, starting with data visualization to identify patterns, missing values, and duplicates. This process includes converting data types as needed, removing columns with excessive missing values, and handling missing data by imputing values, such as using the median of the dataset.

- Before and after the cleaning the data frame can be inspected for head and tail values, statistical descriptions, shape. Data was preprocessed and was been prepared for further analysis and visualisation.

DATA VISUALISATION

- MATPLOTLIB AND SEABORN FOR VISUALIZATION:

- TO IDENTIFY PATTERNS IN THE DATA, THE MATPLOTLIB AND SEABORN LIBRARIES HAVE BEEN UTILIZED. FOLLOWING THE DATA CLEANING PROCESS, THE DATASET IS PREPARED FOR PATTERN RECOGNITION, REGRESSION ANALYSIS, AND FURTHER MACHINE LEARNING APPLICATIONS

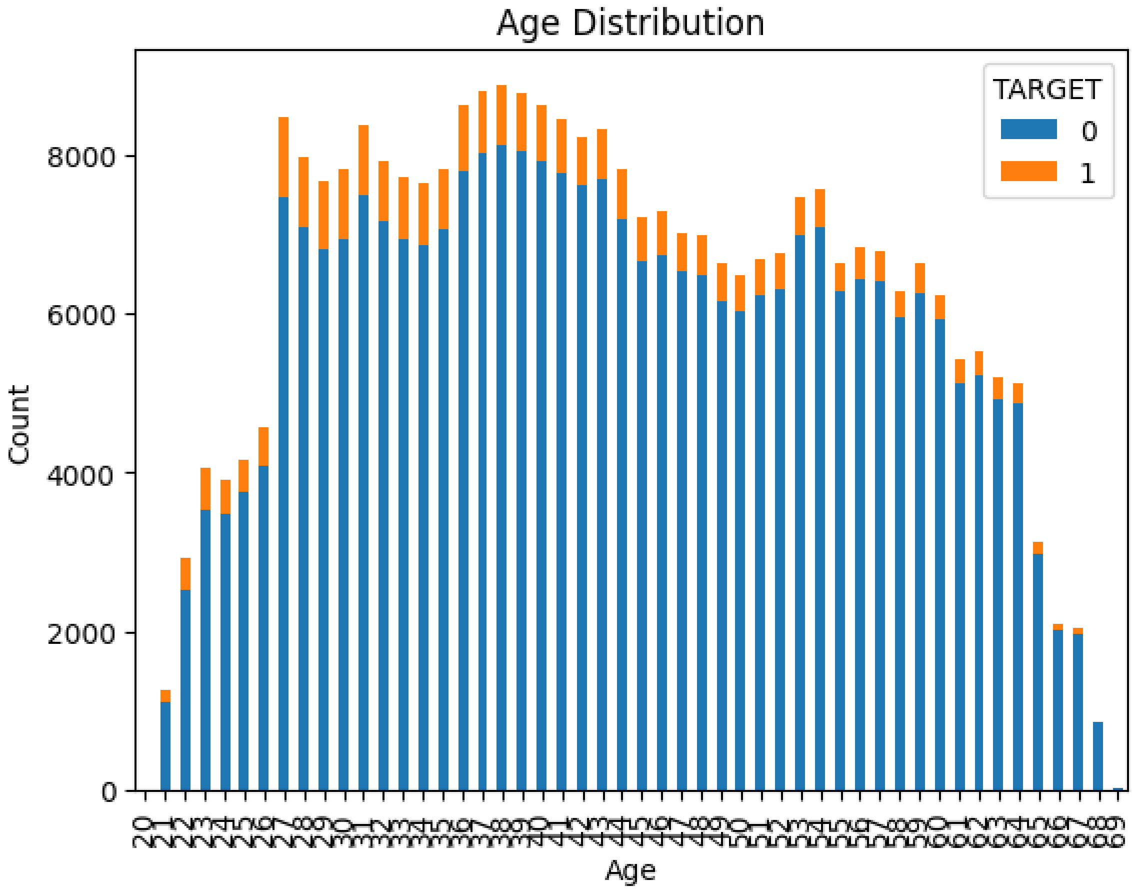
GLOSSARY AND INSIGHTS FOR DATA ANALYSIS

- TARGET. 0 OR 1: 1 FOR HAVING DIFFICULTIES PAYING THE LOAN, 0 FOR HAVING NO DIFFICULTIES

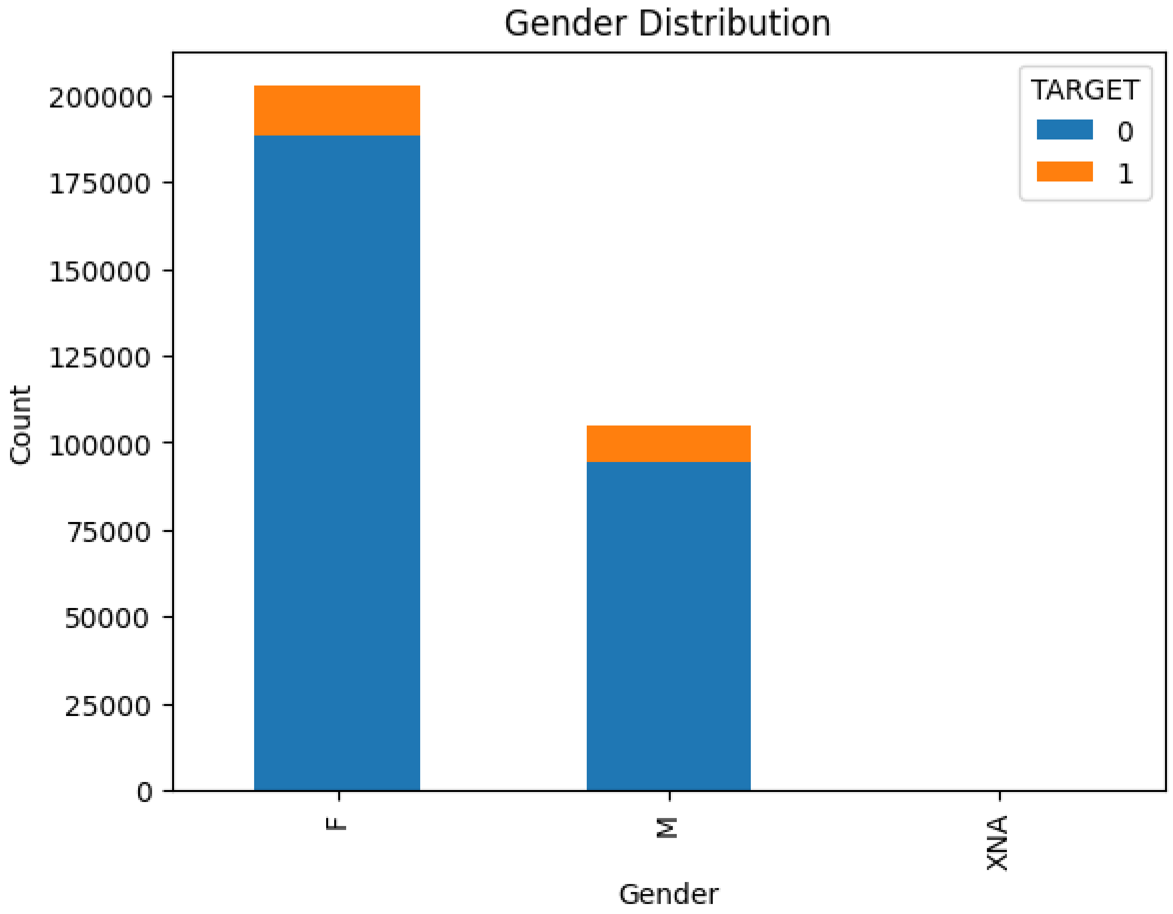
- IN THE FOLLOWING PART THE FOLLOWING WILL BE USED: BOXPLOT, HISTOGRAMS AND BARCHARTS

- - IT WILL BE EXAMINED THE RELATIONSHIPS BETWEEN LOAN DEFAULT AND GENDER, AGE, EDUCATION LEVEL, OCCUPATION, HOUSING TYPE, CREDIT AMOUNT, FAMILY STATUS, CONTRACT TYPE AND CAR OWNERSHIP, A REGRESSION ANALYSIS WILL ALSO BE DONE. A CORRELATION MATRIX WILL BE SHOWN.

AGE AND GENDER DISTRIBUTION

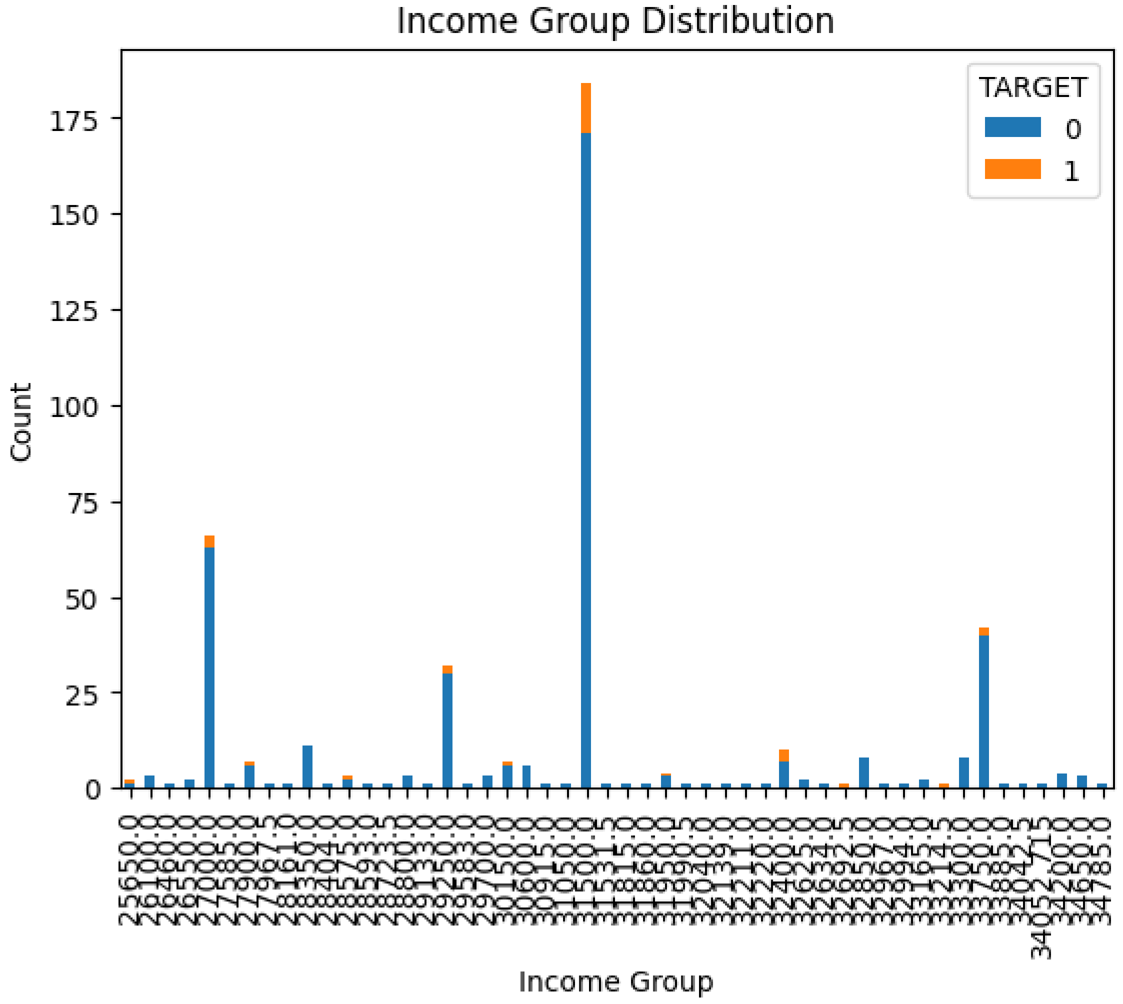


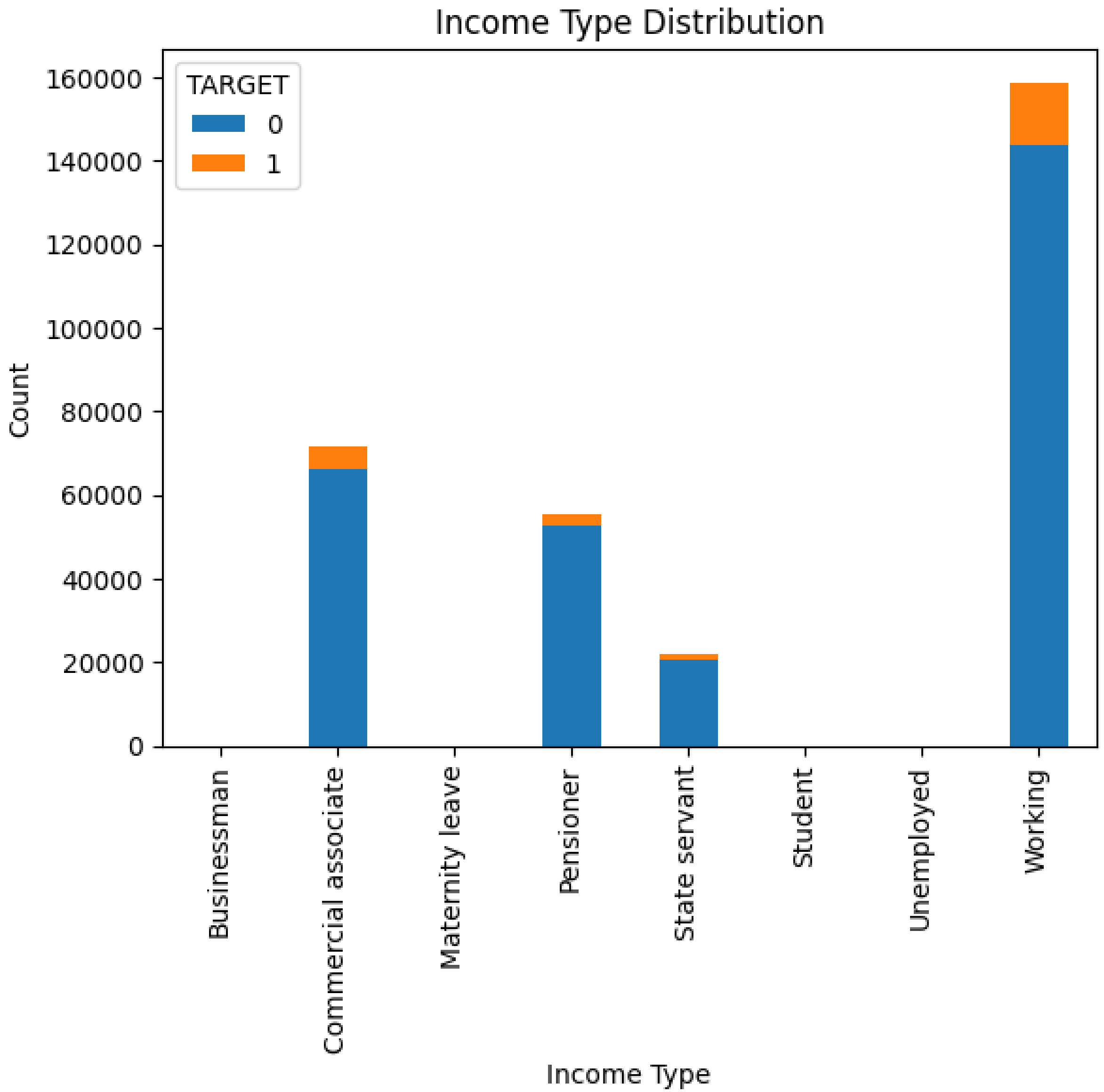
Age Distribution



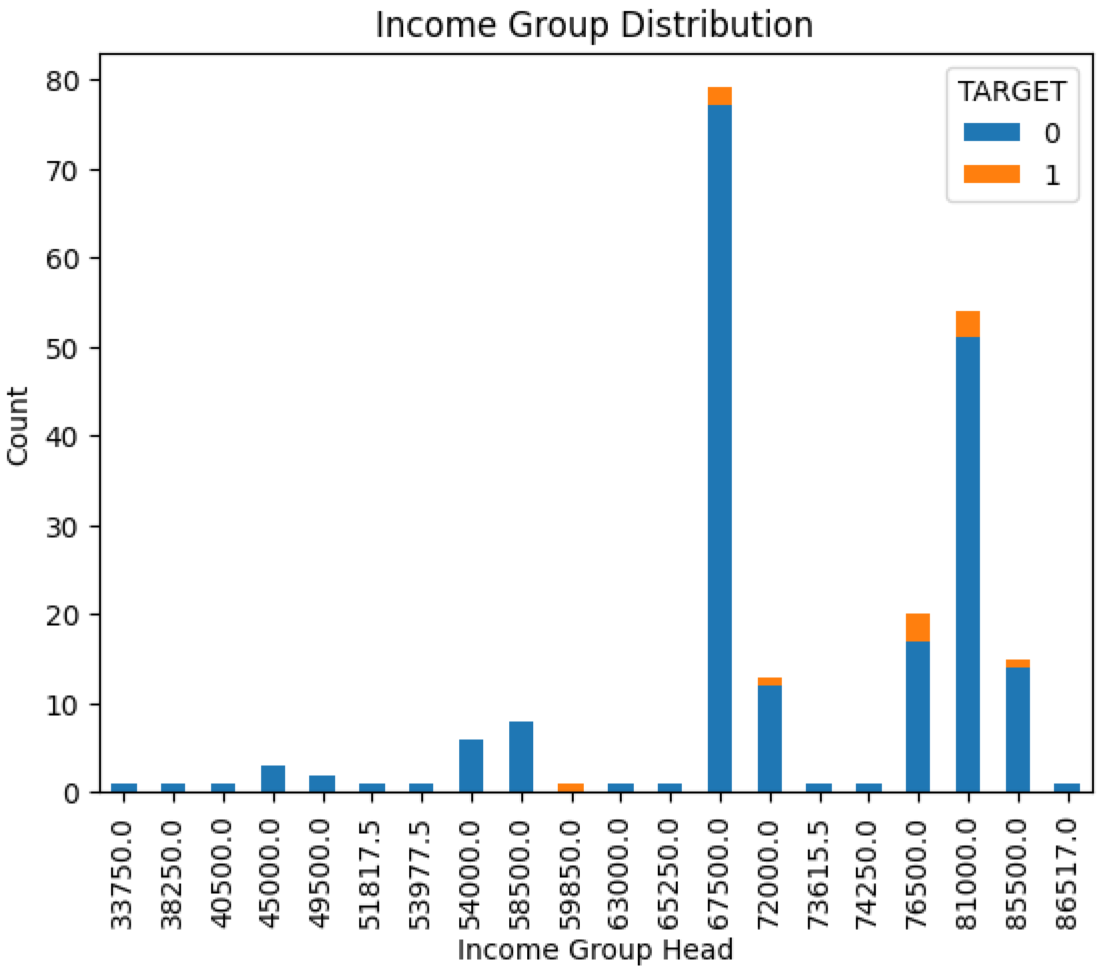
Gender Distribution

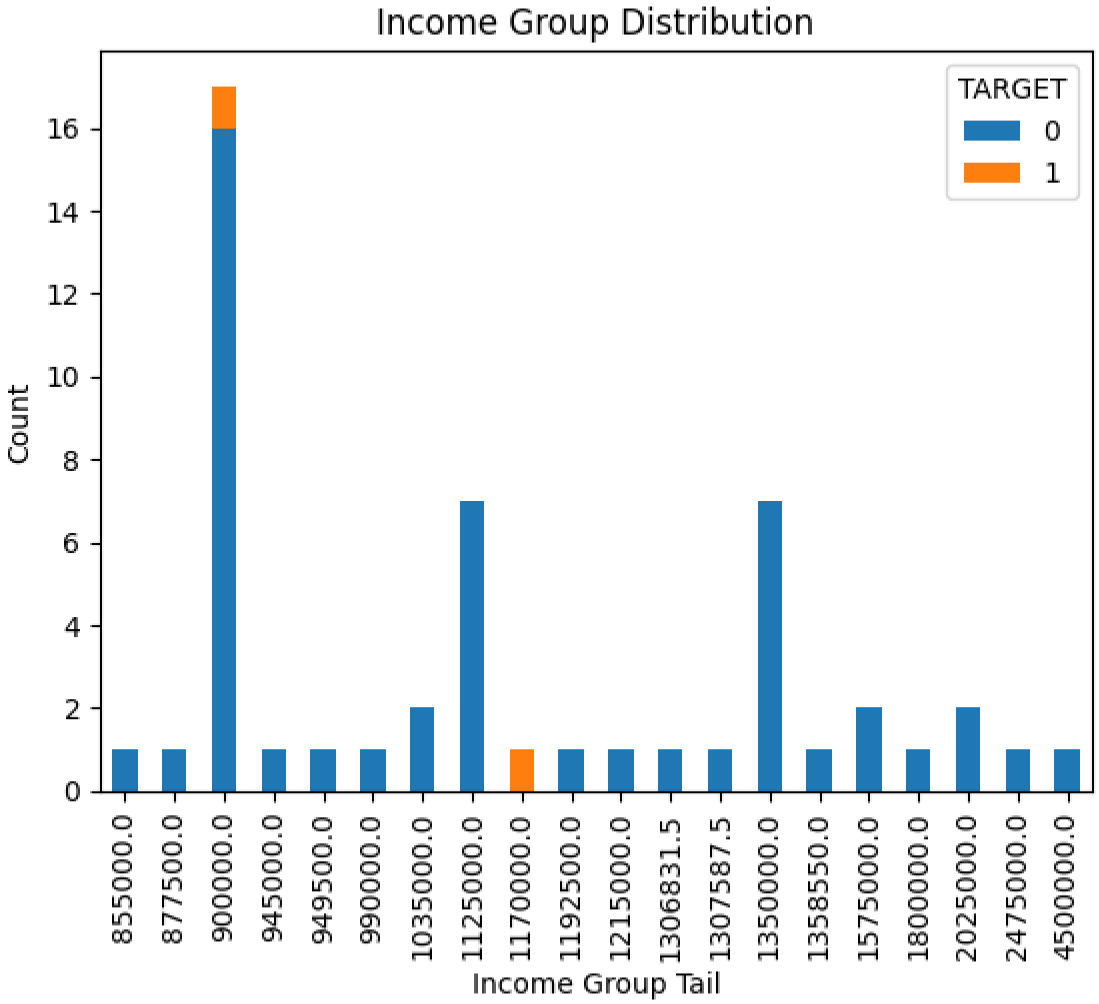
INCOME TYPE AND GROUP



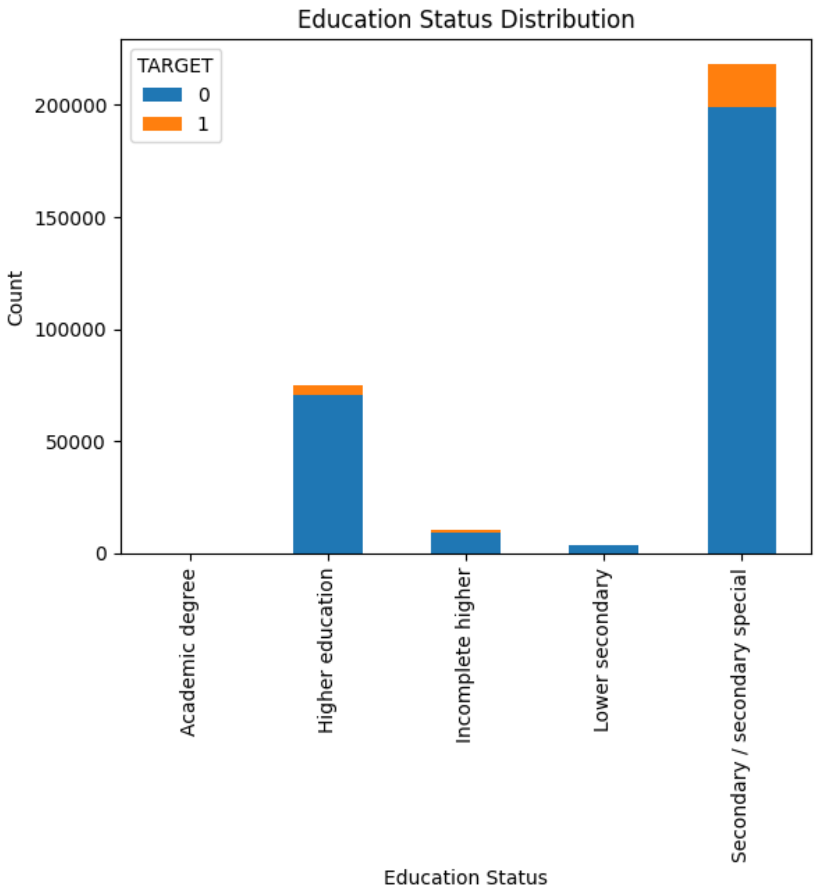


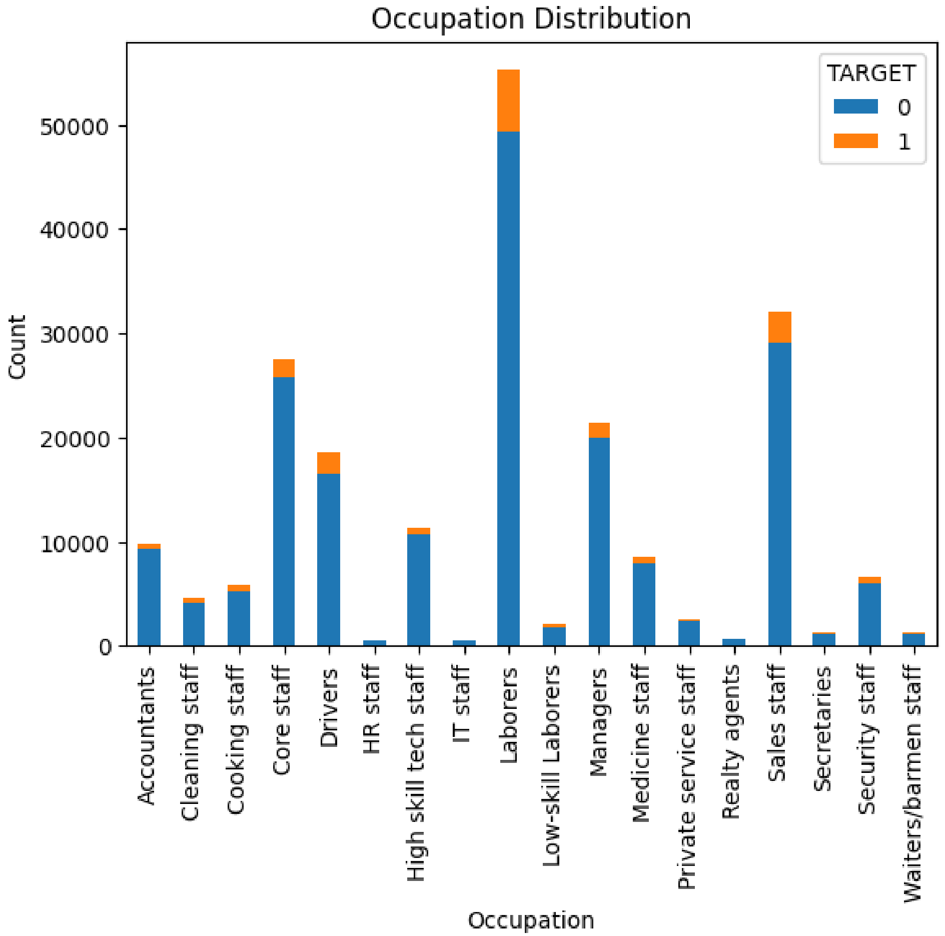
INCOME

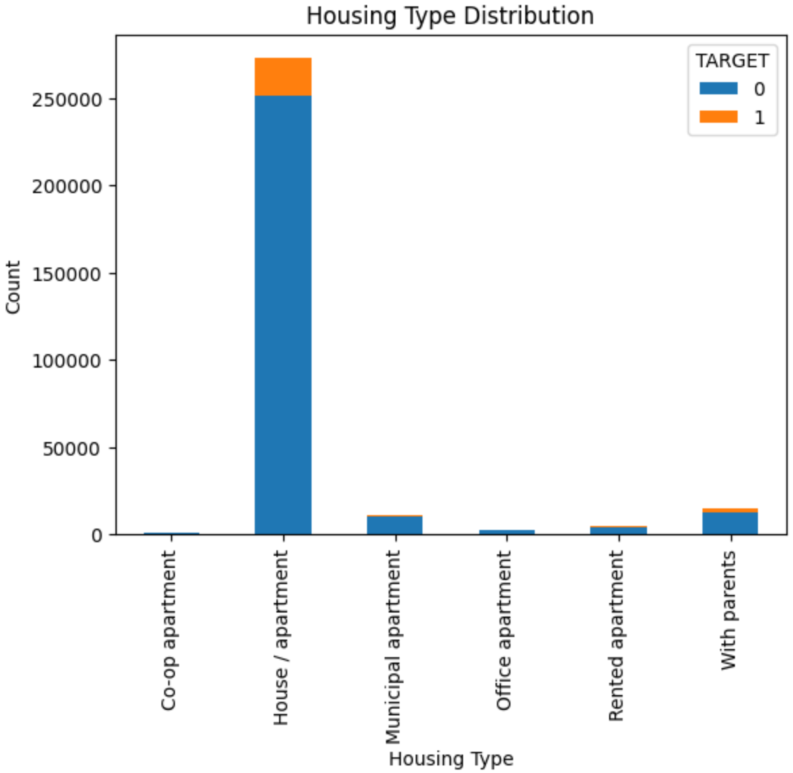

Income Group Distribution
Count
Income Group Head
TARGET
0
1
0
10
20
30
40
50
60
70
80
33750.0
38250.0
40500.0
45000.0
49500.0
51817.5
53977.5
54000.0
58500.0
59850.0
63000.0
65250.0
67500.0
72000.0
73615.5
74250.0
76500.0
81000.0
85500.0
86517.0

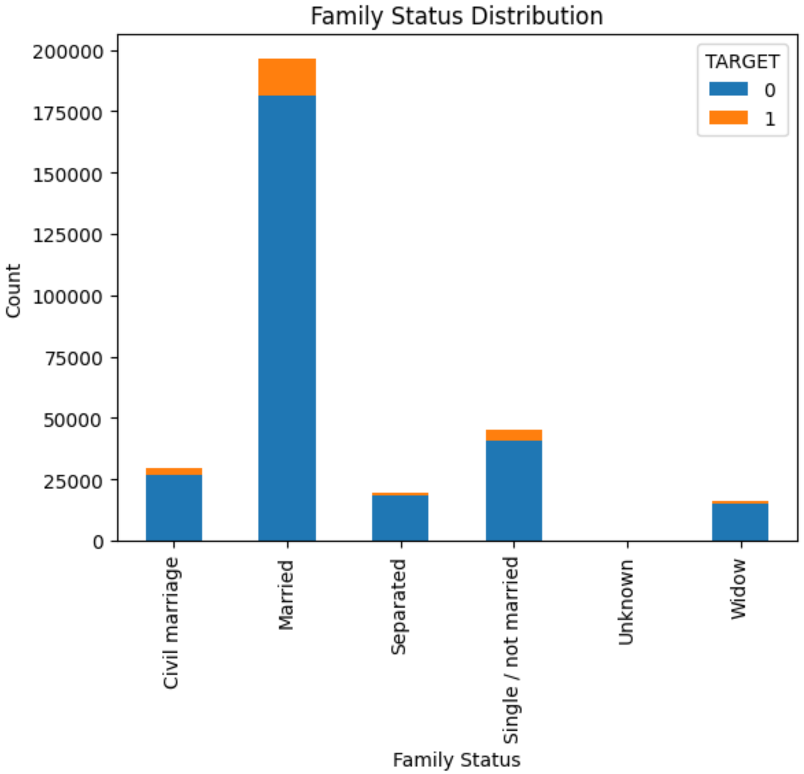

Income Group Distribution
Count
Income Group Tail
TARGET
0
1
0
2
4
6
8
10
12
14
16
855000.0
877500.0
900000.0
945000.0
949500.0
990000.0
1035000.0
1125000.0
1170000.0
1192500.0
1215000.0
1306831.5
1307587.5
1350000.0
1358550.0
1575000.0
1800000.0
2025000.0
2475000.0
4500000.0

EDUCATION LEVEL AND OCCUPATION



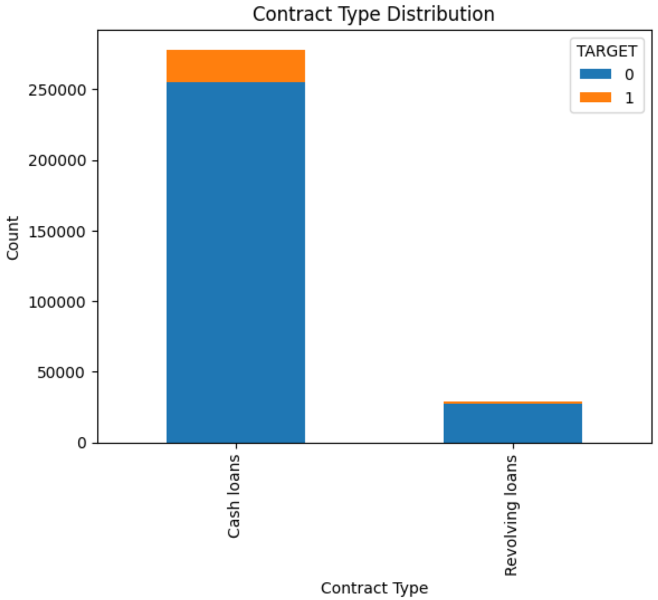


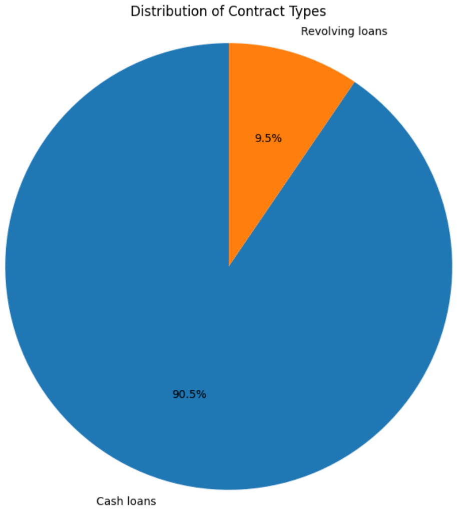


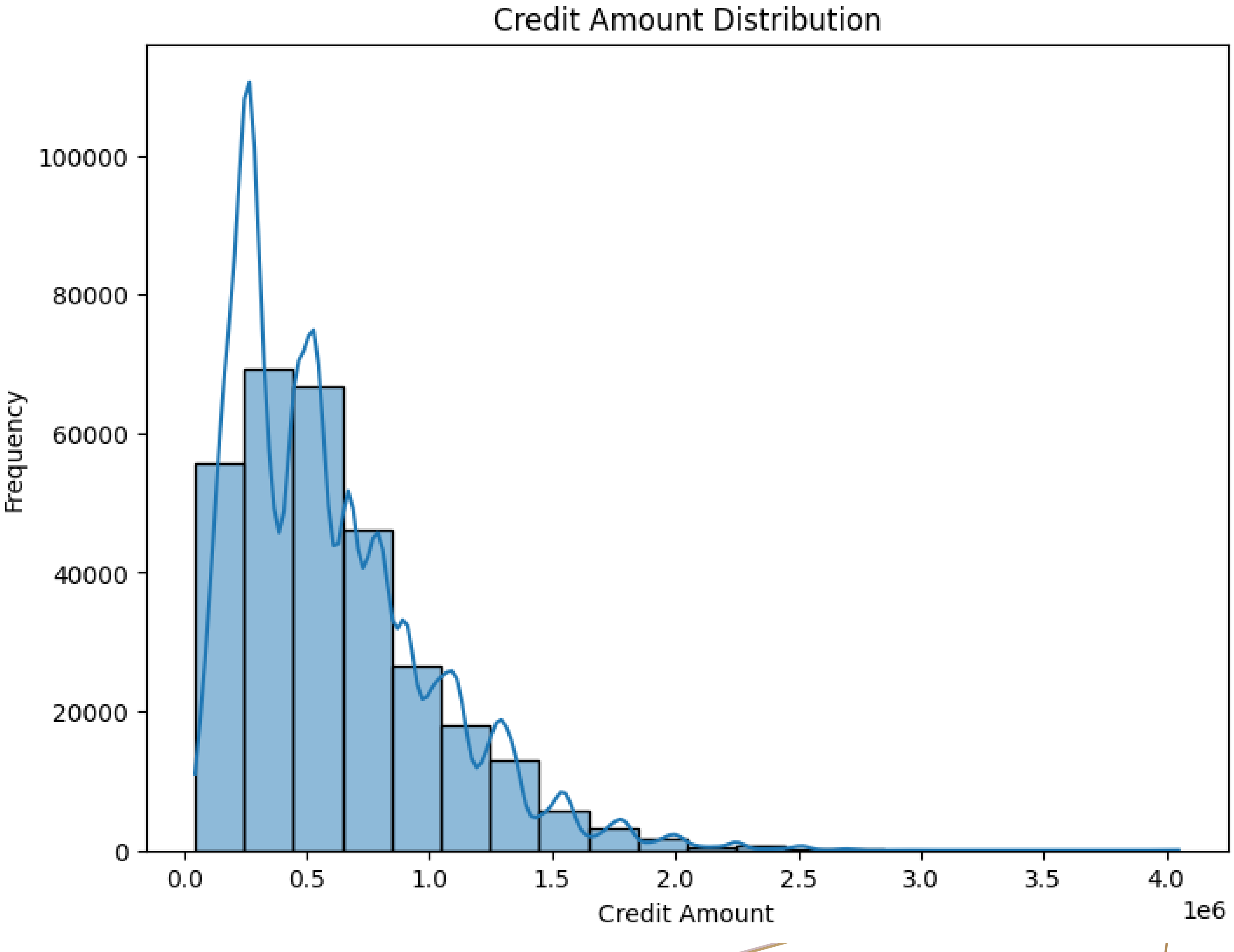


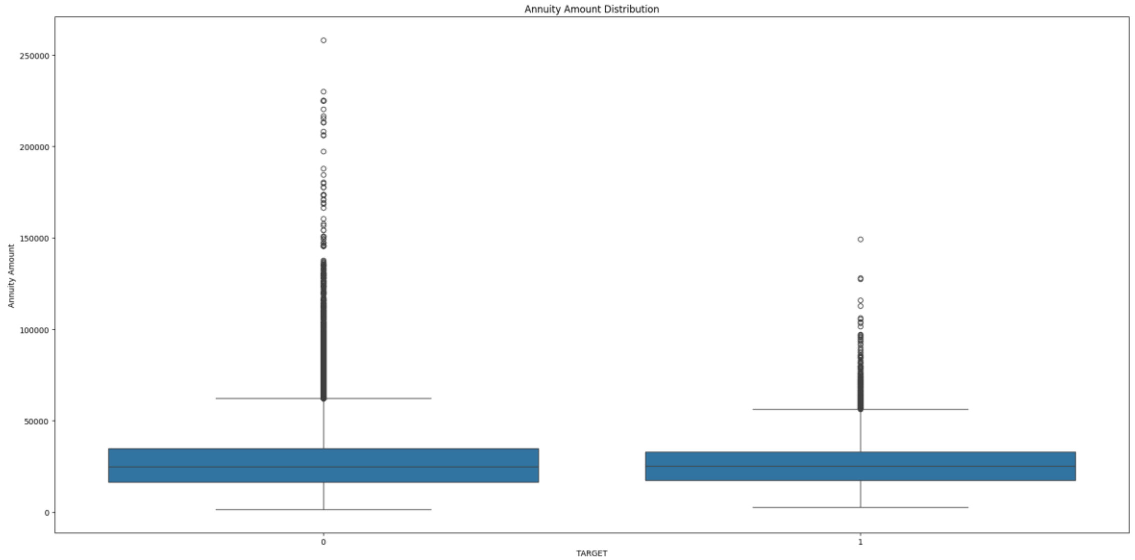
HOUSING TYPE, FAMILY STATUS.

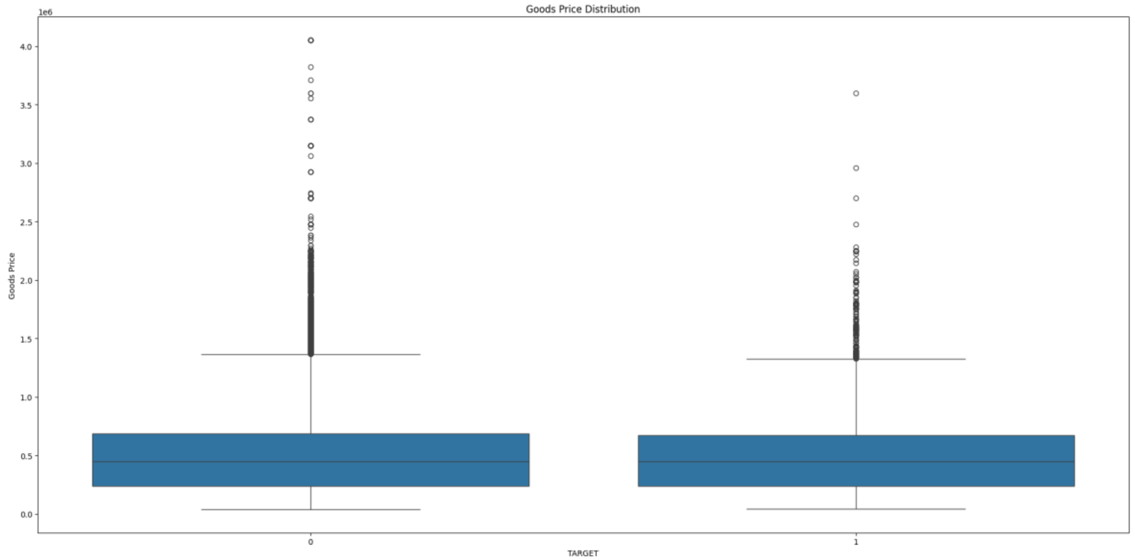
CREDIT AMOUNT, CONTRACT TYPE



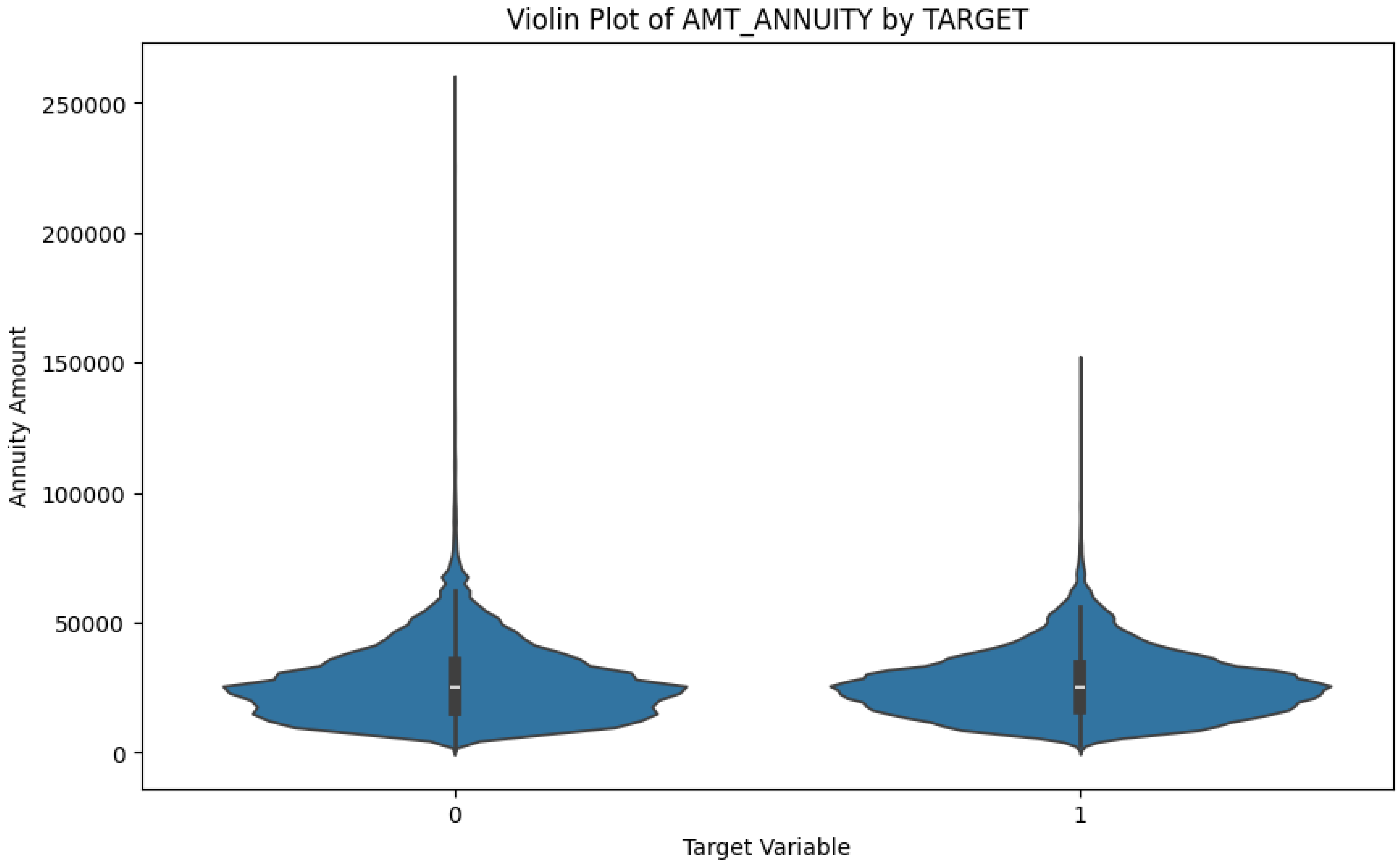


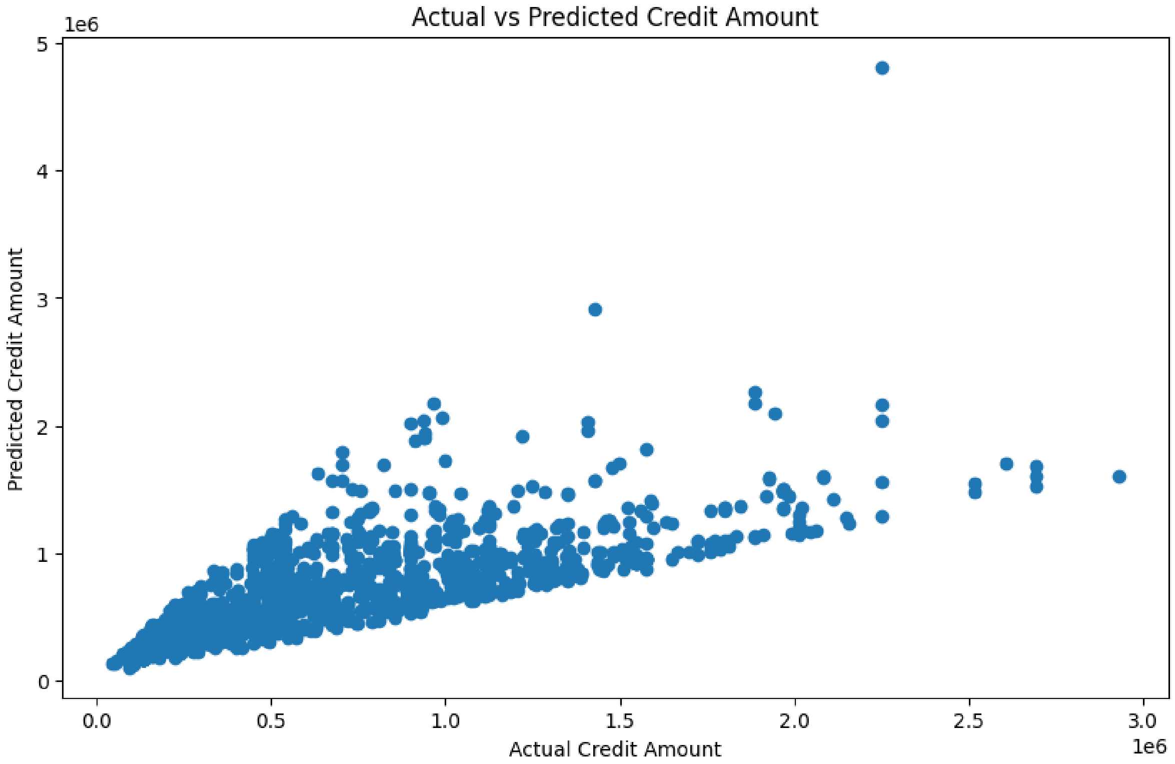






ANNUITY AMOUNT; GOODS PRICE OUTLIER DETECTION


Violin Plot of AMT_ANNUITY by TARGET
Annuity Amount
250000
200000
150000
100000
50000
0
0
1
Target Variable



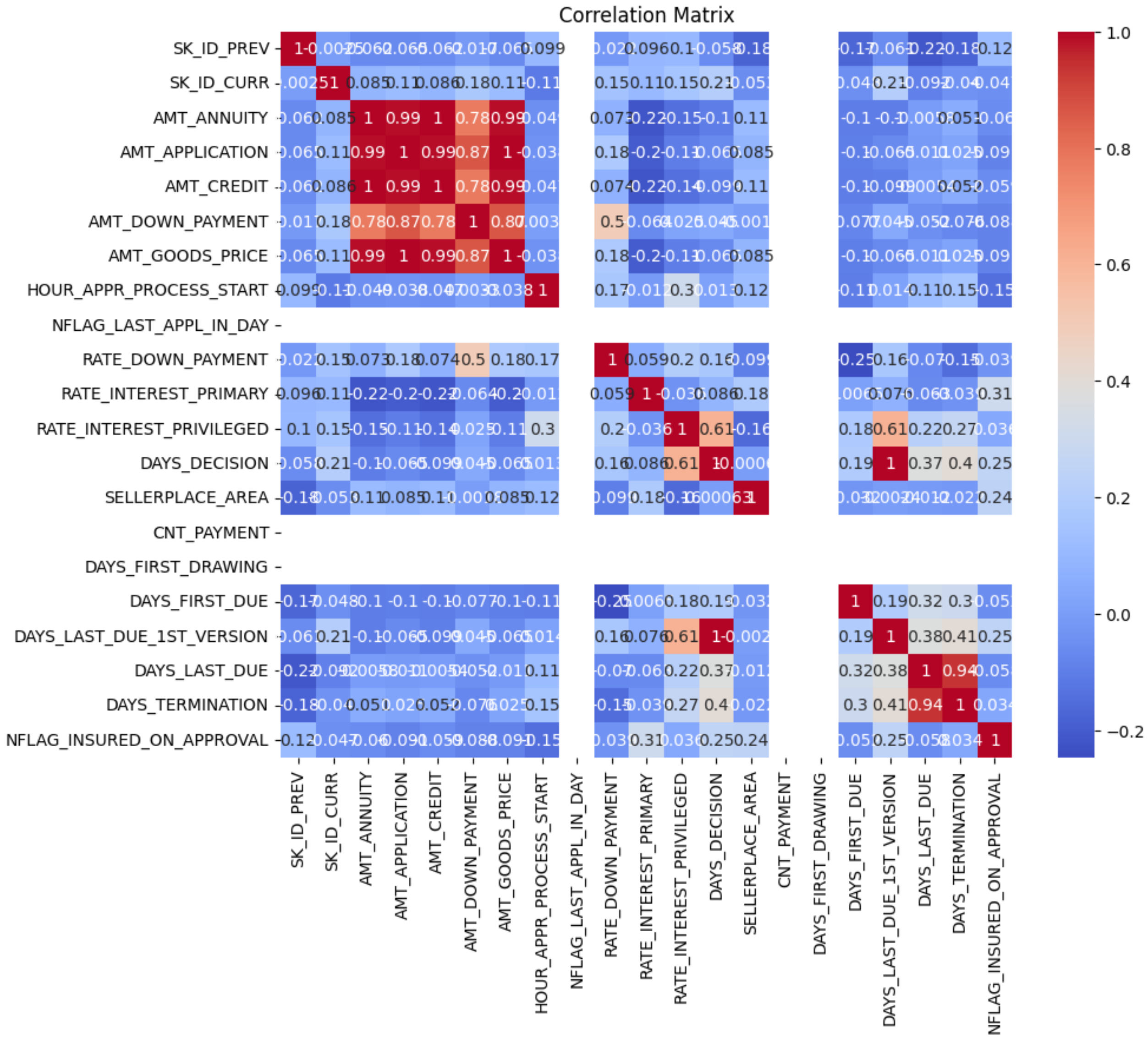
REGRESSION MODEL, ACTUAL VS PREDICTED CREDIT AMOUNT

CORRELATION MATRIX

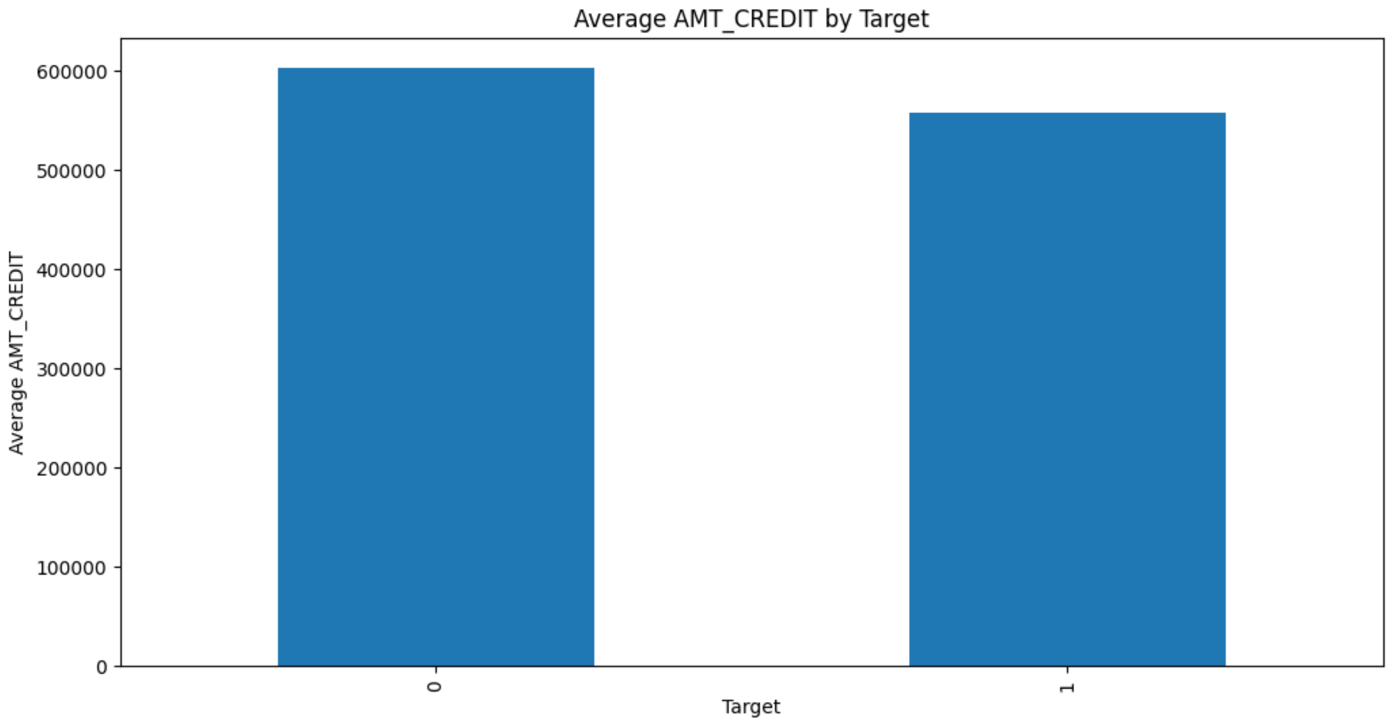
Here, some correlations between the data variables are being shown using the correlation matrix which is a built by numpy ,seaborn and matplotlib.

We see strong correlation between amount applied and price of goods, credit amounts and goods price, down payment and goods price, amount annuity and amount applied, credit amount and annuity amount etc.

The thing we are actually looking for is the relationship between AMT_CREDIT which is the credit amount and other variables like age, gender, education, occupation, income and annuity. The matrix will show how these variables are correlated to one another.



AVG AMOUNT OF CREDIT AMONG TARGETS



OTHER ANALYSIS

T- tests and ANOVA have been performed.

T-test results for credit amount vs income, annuity, children, goods, and days employed:

Example:

- T-test for AMT_CREDIT vs AMT_INCOME_TOTAL:

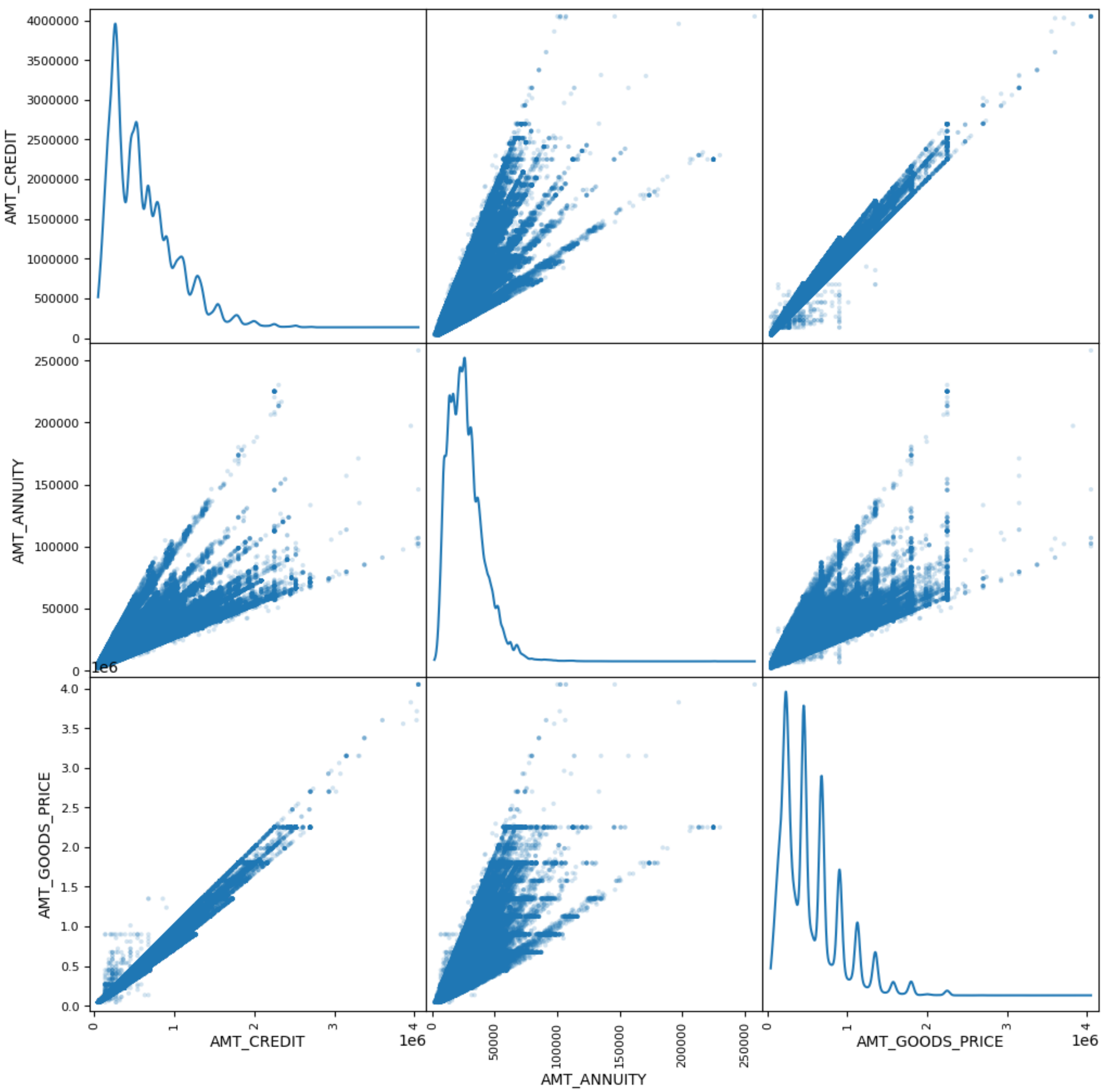
T-statistic: 2.2081011084695983

P-value: 0.027237960879677118

- T-test for AMT_CREDIT vs DAYS_EMPLOYED:

T-statistic: 24.94136608089604

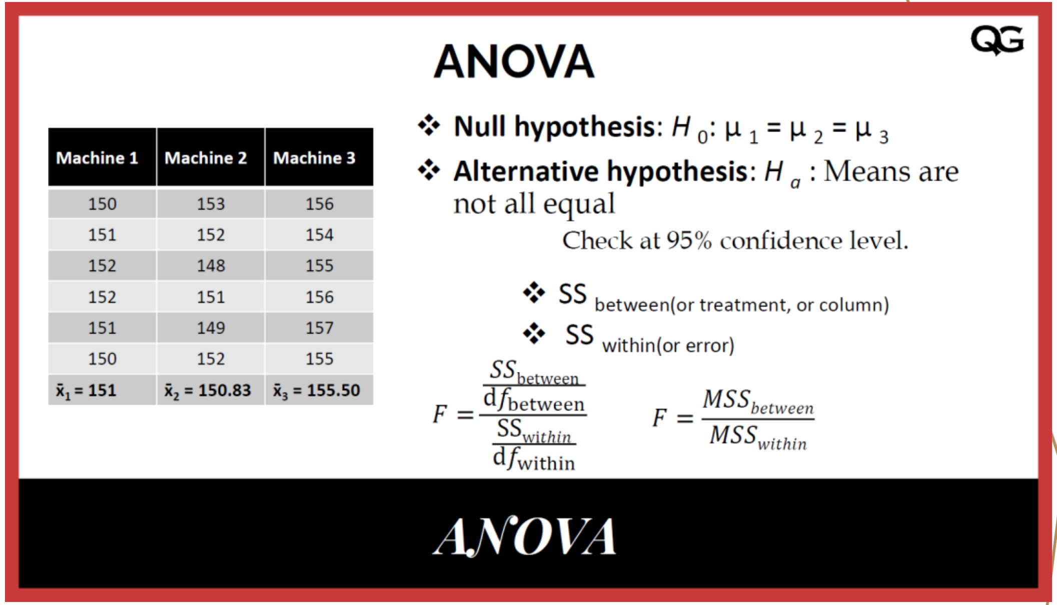
P-value: 3.6311730828848897e-137



RESULTS, INSIGHTS AND STORTELLING

In this project it has been analzed how factors such as education level, family status, owning a house or car, job, experience level and age, gender and annuity in relation to credit amount and it was compared defaulters and non defaulters. Different ykinds of visualisation techniques have been used in order to demonstrate different patterns in data, also a simple regression analysis and relevant statistical tests live t-tests and ANOVA have been done. F statistics and p values have been calculated for some cases.

1. Data cleaned, preprocessed and visualised.
2. Mostly male and the age group between 27-45 has taken or applied for a credit
3. We are seeing loan payment difficulties among high income groups, credit applicants are mostly working as a commercial associate, losts of people from labour class have applied for the credit. Mostly lawyers and sales stuff have taken credit.
4. Loans were given mostly cash form and we see from the regression analysis a strong lean to actual values from expected values.
5. From the correlation matrices, the relationships between variables can be seen.
6. About loan defaulters, the analysis has been done and it can be said that defaulters are a relative small part of the total credit takers and appliers.



QG

ANOVA

Machine 1	Machine 2	Machine 3
150	153	156
151	152	154
152	148	155
152	151	156
151	149	157
150	152	155
$\bar{x}_1 = 151$	$\bar{x}_2 = 150.83$	$\bar{x}_3 = 155.50$

- **Null hypothesis**: H_0: $\mu_1 = \mu_2 = \mu_3$
- **Alternative hypothesis**: H_a : Means are not all equal

Check at 95% confidence level.

- SS $_{\text{between(or treatment, or column)}}$
- SS $_{\text{within(or error)}}$

$$F = \frac{\frac{SS_{between}}{df_{between}}}{\frac{SS_{within}}{df_{within}}} \qquad F = \frac{MSS_{between}}{MSS_{within}}$$

ANOVA

THANK YOU

Nazim Atakan Erdogan

Data Analytics Intern at Oeson