CREDIT EDA CASE STUDY

BY ANUKUL MOREY

INDEX

PROBLEM STATEMENT

The objective of this analysis is to examine loan application data to uncover patterns that can help predict loan defaults. The aim is to differentiate between applicants who are likely to repay their loans and those who pose a higher risk of default. This ensures that financially reliable borrowers are approved while minimizing potential financial losses for the company. The study primarily leverages Exploratory Data Analysis (EDA) to identify key factors influencing loan default.

LOADING DATASET

application_data.csv - Contains applicant details.
(307511, 122)

previous_application.csv - Records of past loan
applications.
(1670214, 37)

DATA CLEANING

```
Columns with missing values:
                           Missing Values   Percentage
commonarea_medi                   214865    69.872297
commonarea_avg                    214865    69.872297
commonarea_mode                   214865    69.872297
nonlivingapartments_medi          213514    69.432963
nonlivingapartments_mode          213514    69.432963
...                                  ...          ...
ext_source_2                         660     0.214626
amt_goods_price                      278     0.090403
amt_annuity                           12     0.003902
cnt_fam_members                        2     0.000650
days_last_phone_change                 1     0.000325

[67 rows x 2 columns]
```
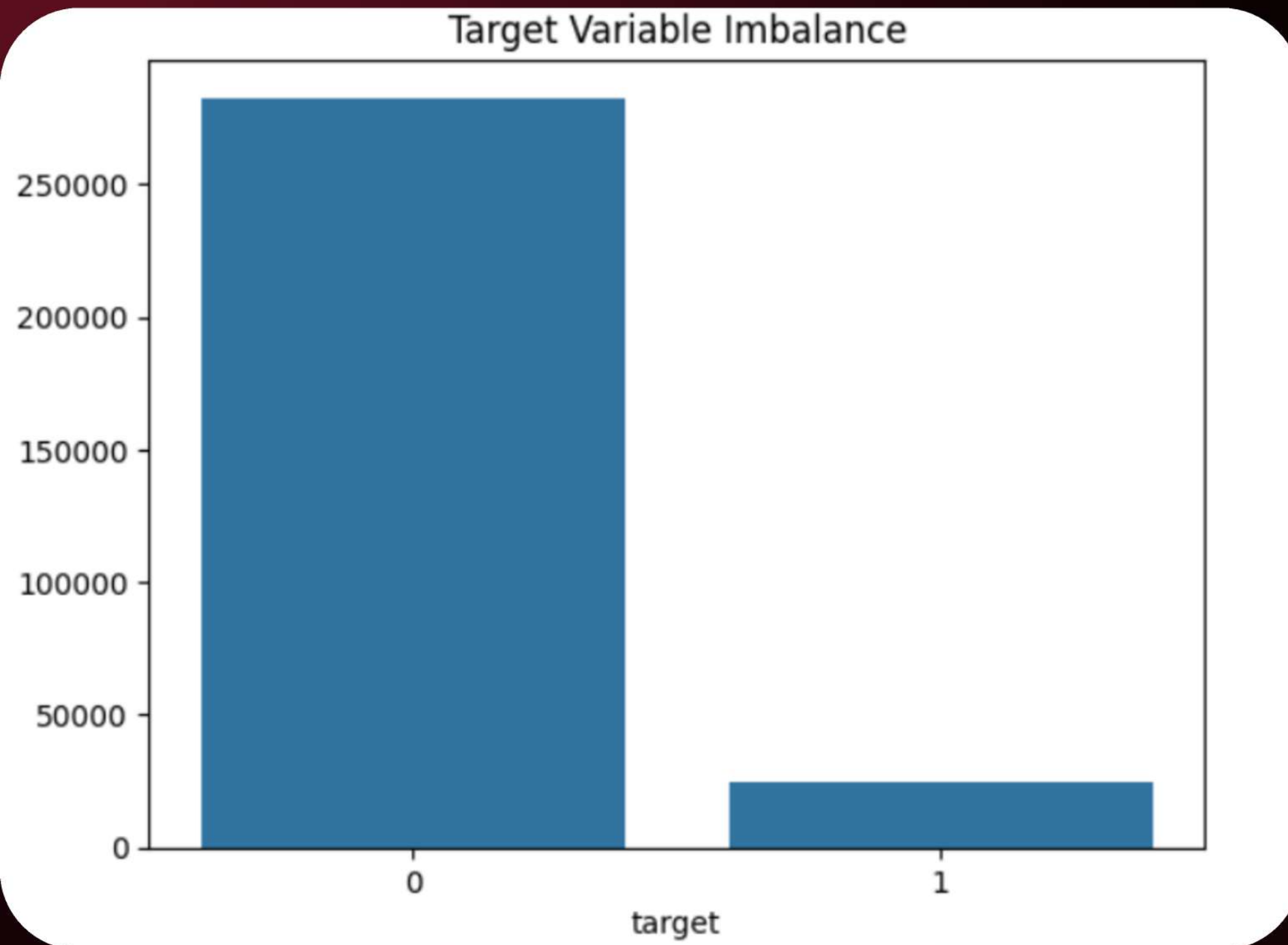
- Handling Missing Values:
    a. Removed columns with excessive missing data.
    b. Imputed missing numerical values with median.
    c. Categorical values replaced using mode.

- Dealing with Outliers:

- Data Standardization:
   a. Ensured uniform formats for numerical and categorical variables.

Target Variable Imbalance

Data Imbalance

- Observation:
  a. Data is highly imbalanced: only ~8% of the applicants defaulted on their loans.
  b. Majority class (non-defaulters) dominates, affecting statistical conclusions.
- Impact:
  a. May cause bias in predictive modeling.
  b. Requires careful consideration during analysis.

UNIVARIATE & BIVARIATE ANALYSIS
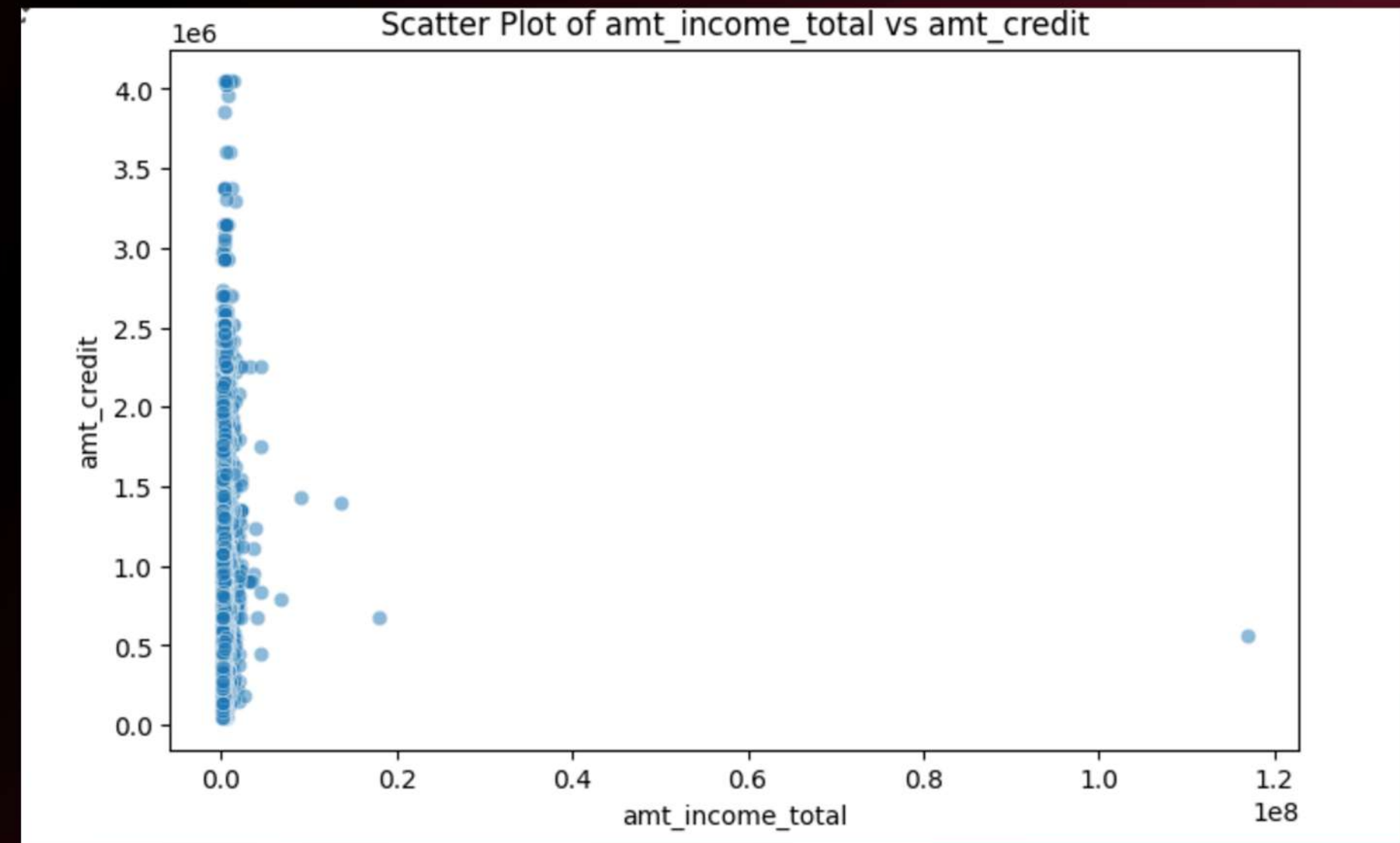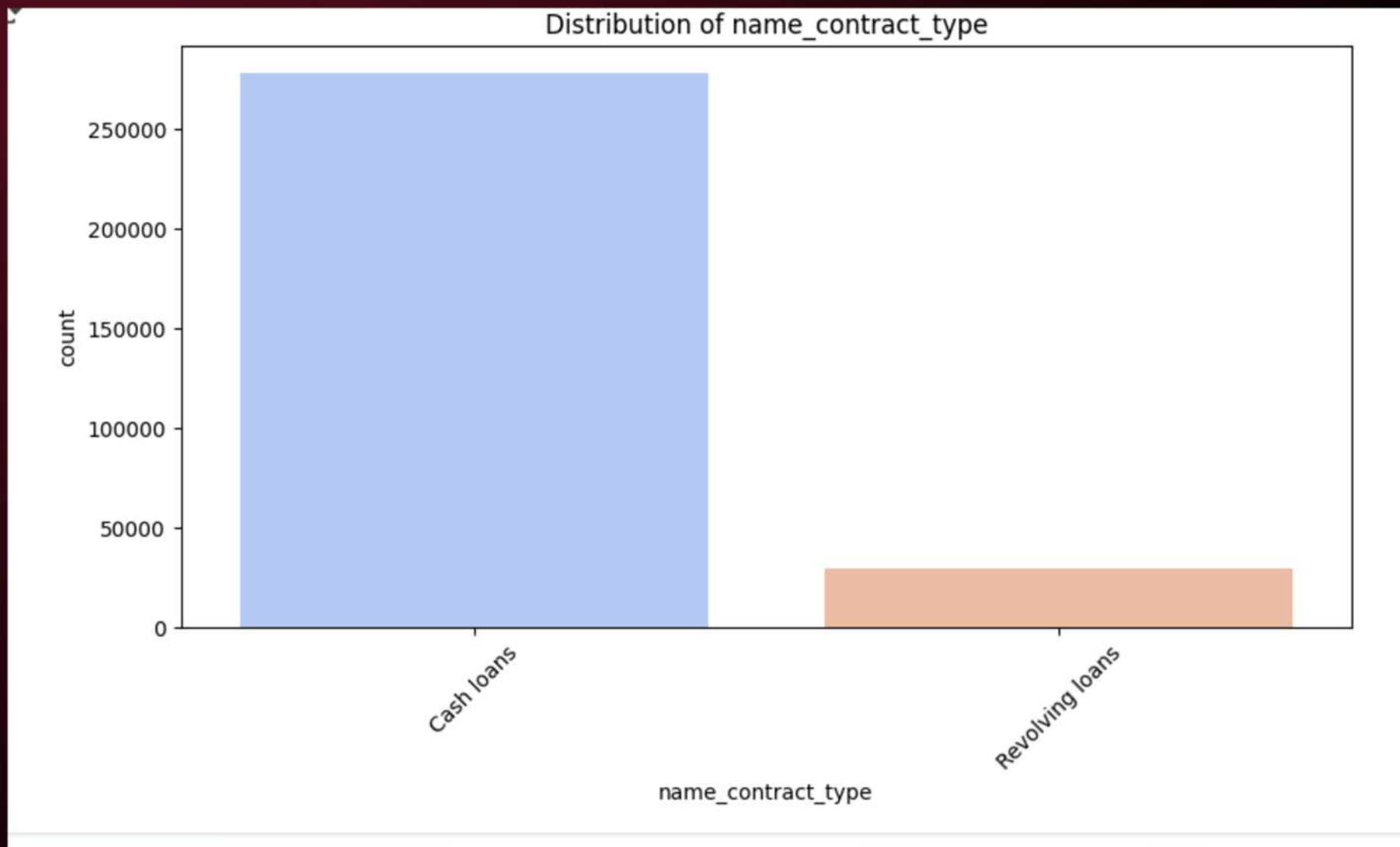
UNIVARIATE & BIVARIATE ANALYSIS

	count	mean	std	min	25%	50%	75%	max
sk_id_curr	307511.0	278180.518577	102790.175348	100002.000000	189145.500000	278202.000000	367142.500000	4.562550e+05
target	307511.0	0.080729	0.272419	0.000000	0.000000	0.000000	0.000000	1.000000e+00
cnt_children	307511.0	0.417052	0.722121	0.000000	0.000000	0.000000	1.000000	1.900000e+01
amt_income_total	307511.0	168797.919297	237123.146279	25650.000000	112500.000000	147150.000000	202500.000000	1.170000e+08
amt_credit	307511.0	599025.999706	402490.776996	45000.000000	270000.000000	513531.000000	808650.000000	4.050000e+06
...	...	...	...	...	...	...	...	...
amt_req_credit_bureau_week	307511.0	0.029723	0.190728	0.000000	0.000000	0.000000	0.000000	8.000000e+00
amt_req_credit_bureau_mon	307511.0	0.231293	0.856810	0.000000	0.000000	0.000000	0.000000	2.700000e+01
amt_req_credit_bureau_qrt	307511.0	0.229631	0.744059	0.000000	0.000000	0.000000	0.000000	2.610000e+02
amt_req_credit_bureau_year	307511.0	1.778463	1.765523	0.000000	1.000000	1.000000	3.000000	2.500000e+01
age	307511.0	43.936973	11.956133	20.517808	34.008219	43.150685	53.923288	6.912055e+01

62 rows × 8 columns

MULTIVARIATE ANALYSIS

CORRELATION ANALYSIS

Correlation Matrix Heatmap
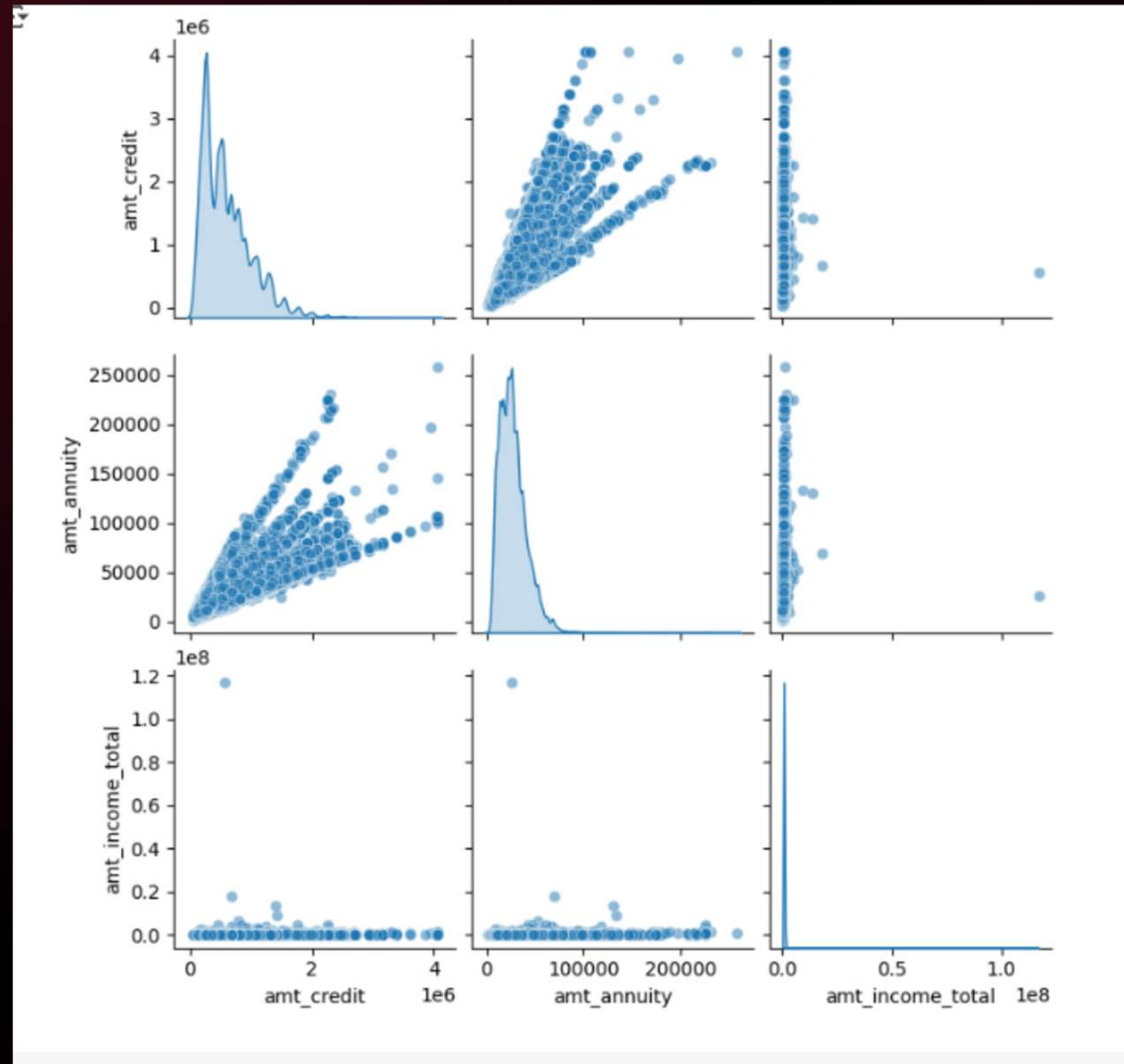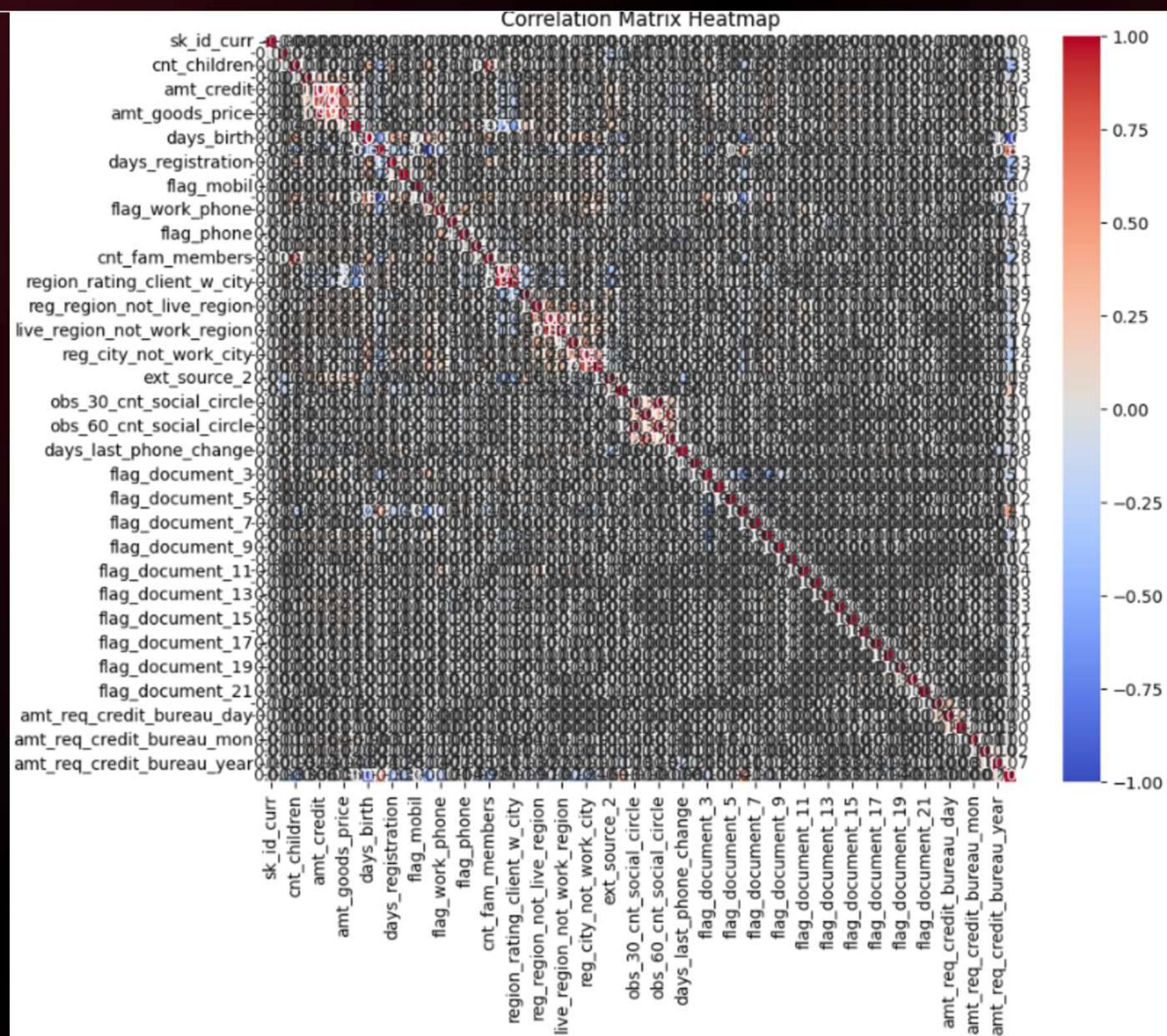
INSIGHTS

- Because of the imbalance in the data, the results may not be fully definitive. However, the analysis highlights several key factors that contribute to loan default, including:
- Loan Amount: Higher loan amounts are associated with an increased risk of default.
- Income Level: Applicants with lower income tend to have a higher likelihood of defaulting.
- Credit History: Individuals with poor or limited credit history show a greater tendency to default.
- Previous Loan Status: Applicants with previously rejected loans exhibit a higher probability of default.
- Debt-to-Income Ratio: A higher ratio indicates financial strain, making default more likely.

Thankyou

Anukul Morey

7666400518

anukulmorey987@gmail.com