BANK LOAN CASE STUDY

-MRUNALI PETAKAR

DESCRIPTION

The project aims to analyze the loan application dataset to identify patterns and factors that influence loan default. The goal is to use Exploratory Data Analysis (EDA) to understand how customer attributes and loan attributes impact the likelihood of default, so that the company can make informed decisions and reduce the risk of default. The dataset contains information about loan applications, including customers with payment difficulties and all other cases. The project will focus on identifying missing data, outliers, data imbalance, and performing univariate, segmented univariate, and bivariate analysis to gain insights into the driving factors of loan default. The top correlations for different scenarios will also be identified.

SCOPE

The project will focus on analyzing the loan application dataset provided, which contains information about loan applicants and their loan applications. The dataset will be analyzed to identify patterns and trends that can help predict the likelihood of loan default.

TECH - STACK USED

- Python 3.11
- Libraries: Pandas, Matplotlib, Seaborn, NumPy
- Google Drive
- PowerPoint

here I am giving the link to Excel sheet & github link

APPROACH

Data Preparation and Cleaning

Removing any irrelevant or redundant data to ensure that the remaining variables are relevant to the analysis.

Descriptive Statistics and Visualization

to detect and identify outliers.

Bivariate and Multivariate Analysis

to identify any interactions or non-linear relationships between variables.

Model Building and Evaluation

to evaluate the performance of the classification models.

Insights and Recommendations

to create visualizations and present the findings and recommendations in a clear and concise manner.

UNDERSTANDING THA DATA

a. previous_application.csv: Contains information about previous loan applications.

b. application_data.csv: Provides details about the current loan applications.

c. columns_description.csv: Describes the columns present in the other datasets, explaining what each column represents.

Approach-

- Loaded three datasets: **application_data**, **previous_application**, and **columns_description**.
- Examined the structure and content of the datasets

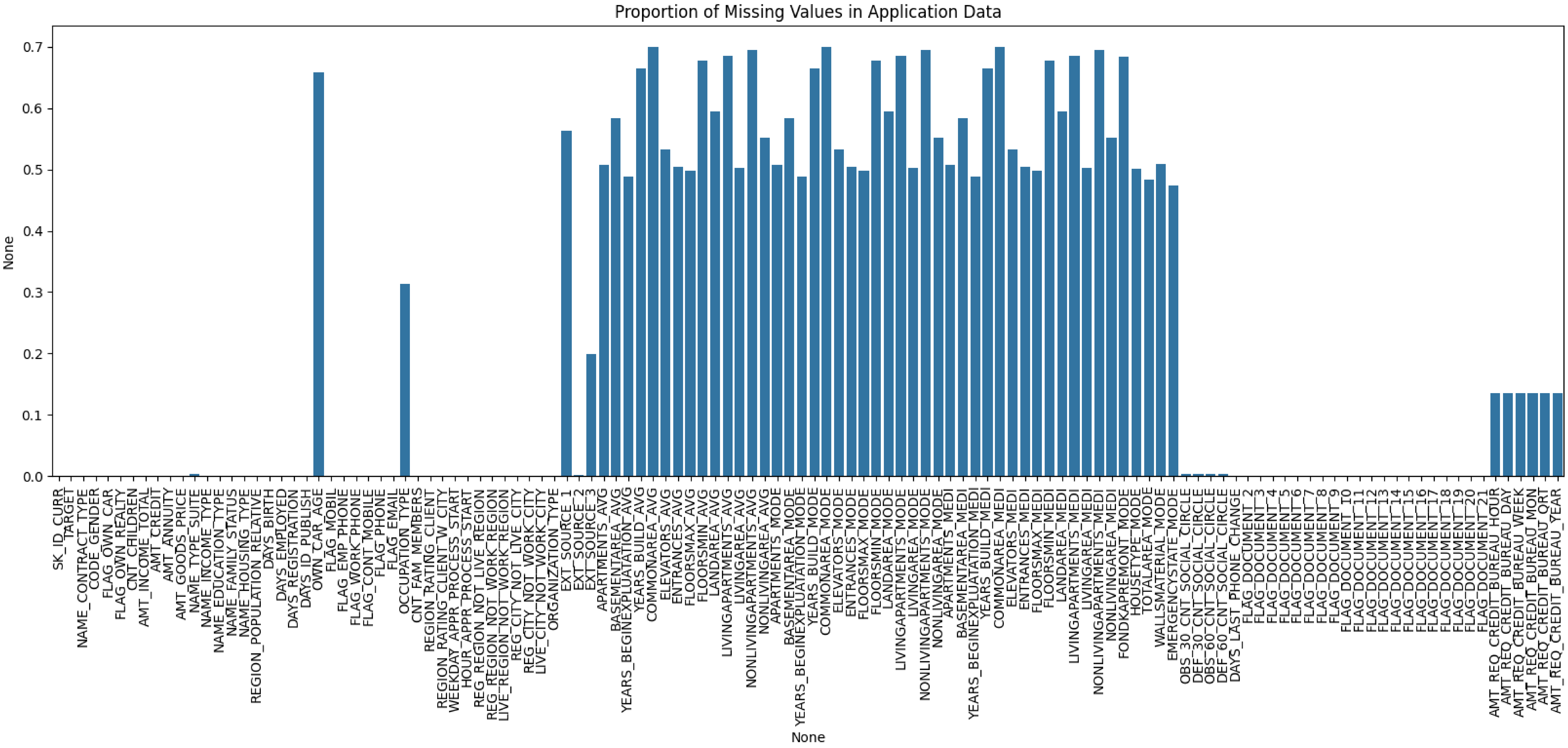
1] DATA CLEANING

1. Identified missing values in application_data.
2. Visualized the proportion of missing values using a bar plot.
3. Handled missing values:
   - Imputed missing numerical values with the mean.
   - Imputed missing categorical values with the mode.

columns to drop -

HOUSETYPE_MODE
WALLSMATERIAL_MODE
BASEMENTAREA_MEDI
FLOORSMIN_MEDI
LIVINGAREA_AVG
ELEVATORS_AVG
LANDAREA_AVG
LIVINGAPARTMENTS_AVG
LIVINGAREA_MODE
ELEVATORS_MODE
LANDAREA_MODE
LIVINGAPARTMENTS_MODE
LIVINGAREA_MEDI
ELEVATORS_MEDI
LANDAREA_MEDI
LIVINGAPARTMENTS_MEDI
ENTRANCES_AVG
NONLIVINGAREA_AVG
OWN_CAR_AGE
FONDKAPREMONT_MODE
ENTRANCES_MODE
NONLIVINGAREA_MODE
YEARS_BUILD_AVG
NONLIVINGAPARTMENTS_AVG
ENTRANCES_MEDI
NONLIVINGAREA_MEDI
YEARS_BUILD_MODE
NONLIVINGAPARTMENTS_MODE
APARTMENTS_AVG
EXT_SOURCE_1
YEARS_BUILD_MEDI
NONLIVINGAPARTMENTS_MEDI
APARTMENTS_MODE
BASEMENTAREA_AVG
FLOORSMIN_AVG
COMMONAREA_AVG
APARTMENTS_MEDI
BASEMENTAREA_MODE
FLOORSMIN_MODE
COMMONAREA_MODE
COMMONAREA_MEDI

data cleaning



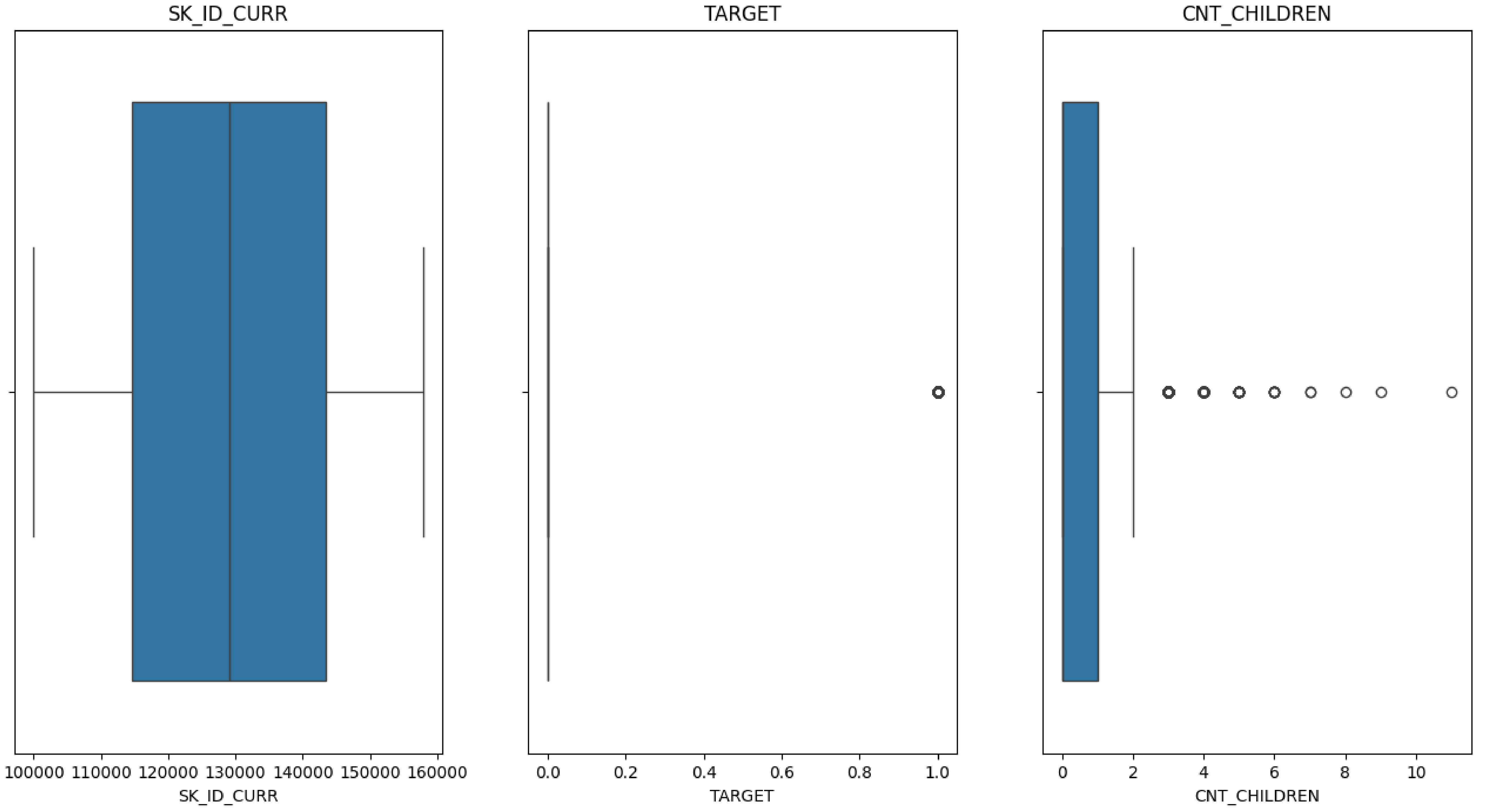
```
# Identify numerical columns
numerical_columns = application_data.select_dtypes(include=['float64', 'int64']).columns

# Replace missing values with the mean for numerical columns
application_data[numerical_columns] =
application_data[numerical_columns].fillna(application_data[numerical_columns].mean())

# Identify categorical columns
categorical_columns = application_data.select_dtypes(include=['object']).columns

# Replace missing values with the median for categorical columns
application_data[categorical_columns] =
application_data[categorical_columns].fillna(application_data[categorical_columns].mode().il
oc[0])
```

TASK 2]

Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.

- Outliers can only be identified on Numeric variables.

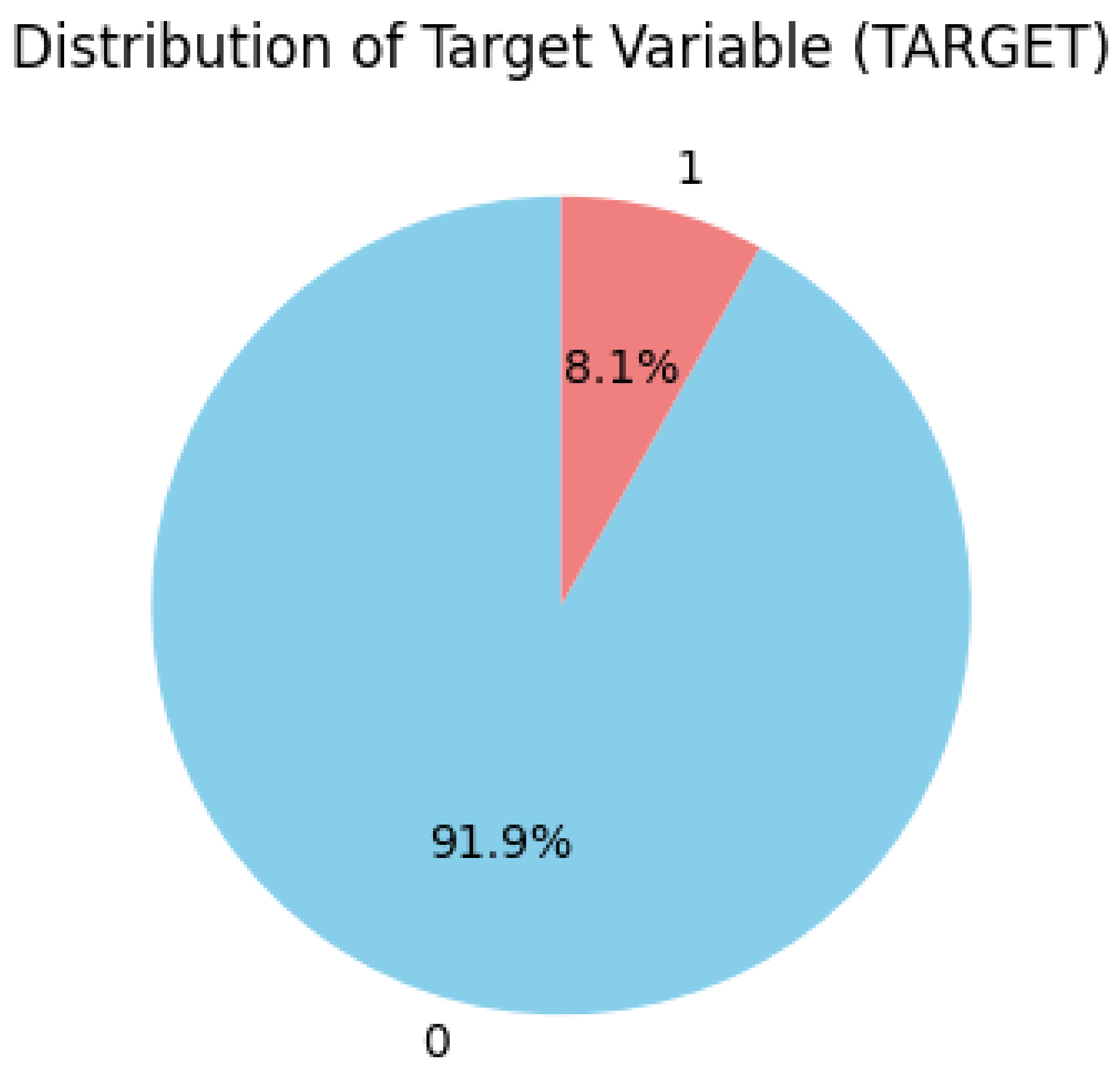
Identified outliers in numerical columns of application_data using box plots shown as below

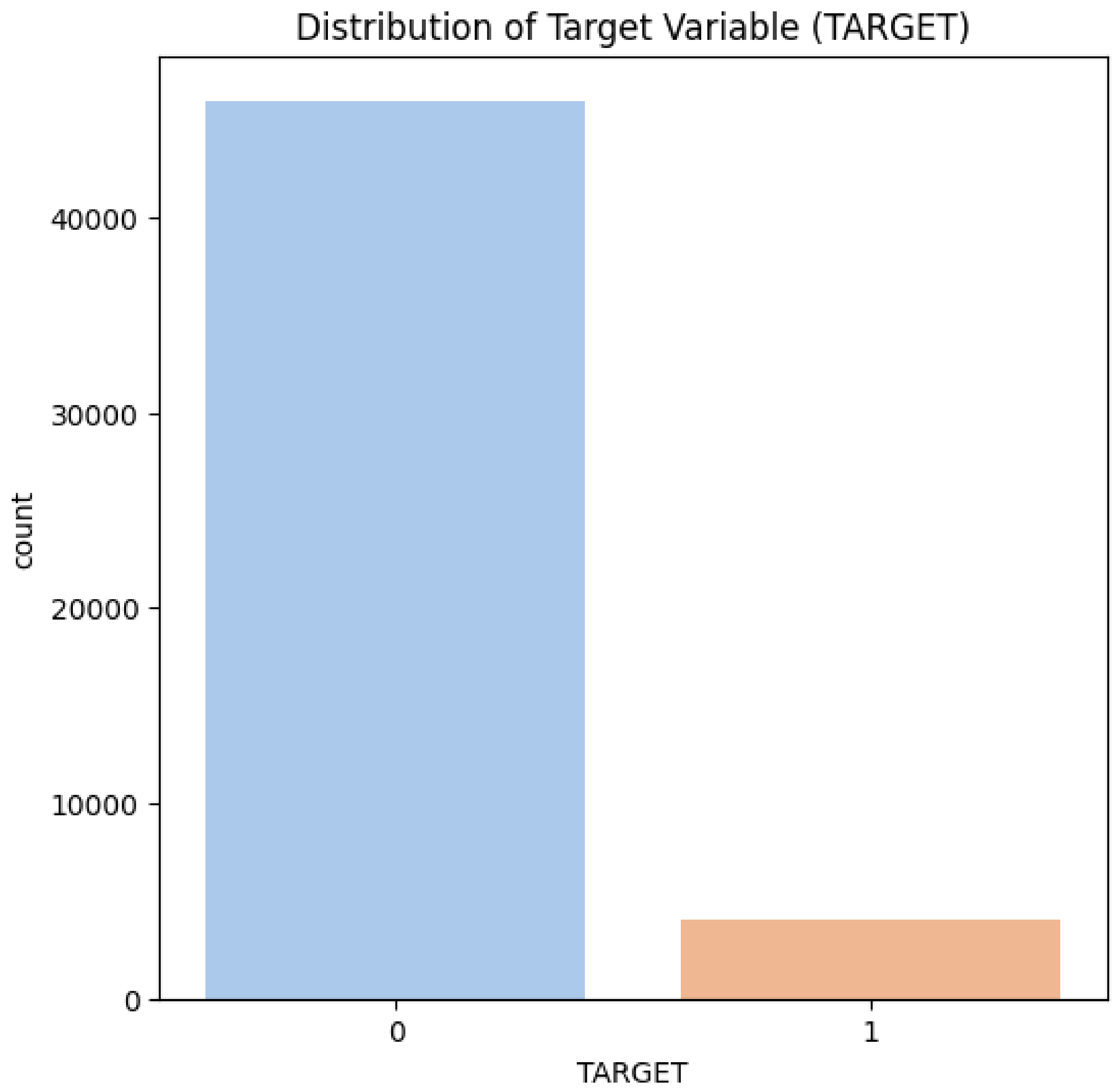


TASK C] Analyze Data Imbalance

- Determining if there is data imbalance in the loan application dataset and calculating the ratio of data imbalance Visualized the distribution of the target variable (TARGET) using a pie chart and bar chart.

```
# Assuming 'TARGET' is your target variable column
target_counts = application_data['TARGET'].value_counts()

# Pie Chart
plt.figure(figsize=(4,4))
plt.pie(target_counts, labels=target_counts.index, autopct='%1.1f%%', startangle=90, colors=['skyblue', 'lightcoral'])
plt.title('Distribution of Target Variable (TARGET)')
plt.show()

# Bar Chart
plt.figure(figsize=(6, 6))
sns.countplot(x='TARGET', data=application_data, palette='pastel')
plt.title('Distribution of Target Variable (TARGET)')
plt.show()
```


Distribution of Target Variable (TARGET)
1
8.1%
91.9%
0


Distribution of Target Variable (TARGET)
count
40000
30000
20000
10000
0
0
1
TARGET

TASK D] Perform Univariate, Segmented Univariate, and Bivariate Analysis:

Univariate Analysis:

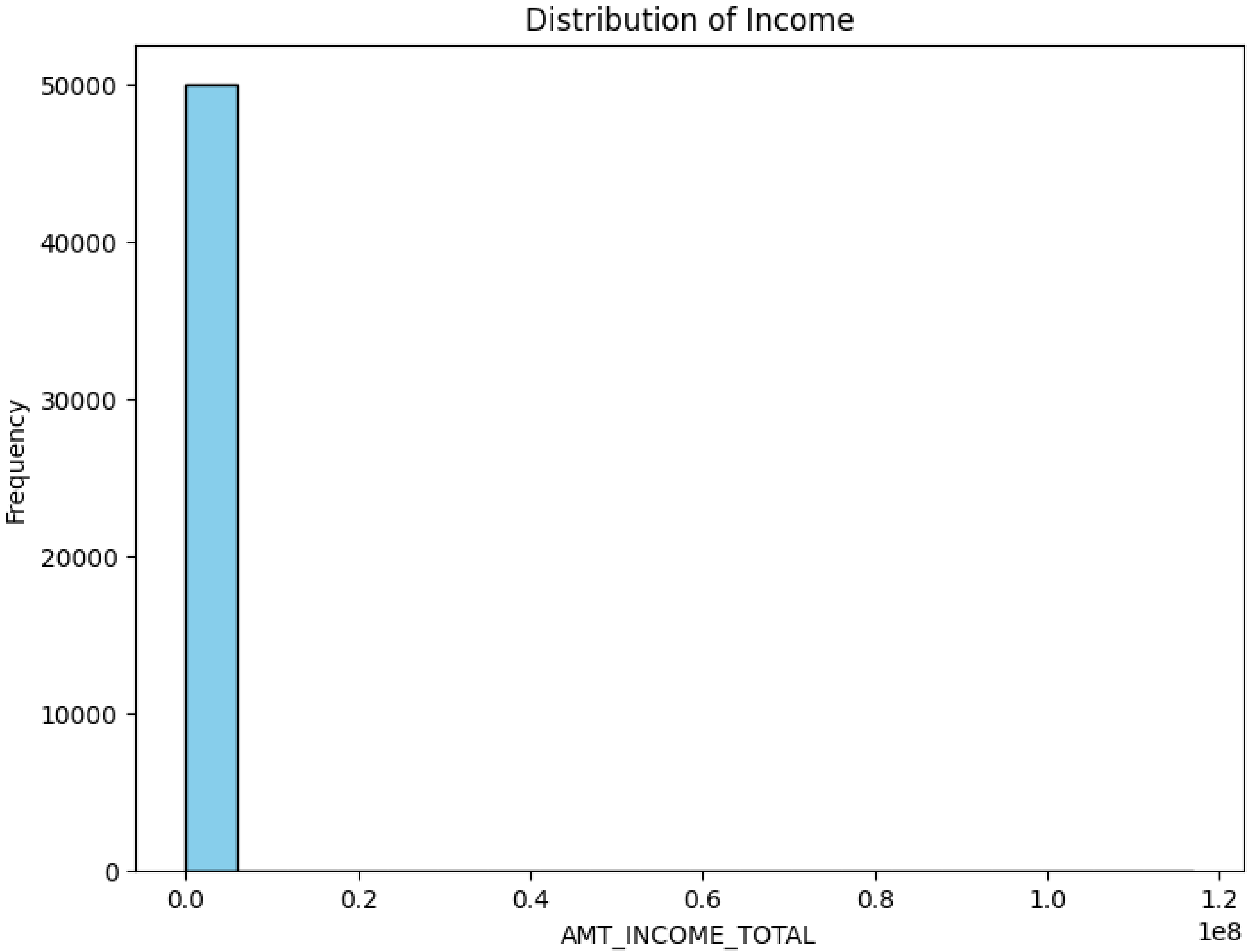
Plotted histograms for numerical variables.
Plotted count plots for categorical variables.
Segmented Univariate Analysis:

Utilized grouped bar charts for categorical variables across scenarios.
Bivariate Analysis:

Created scatter plots for numerical-target variable relationships.
Utilized box plots for analyzing the income distribution by education and target.

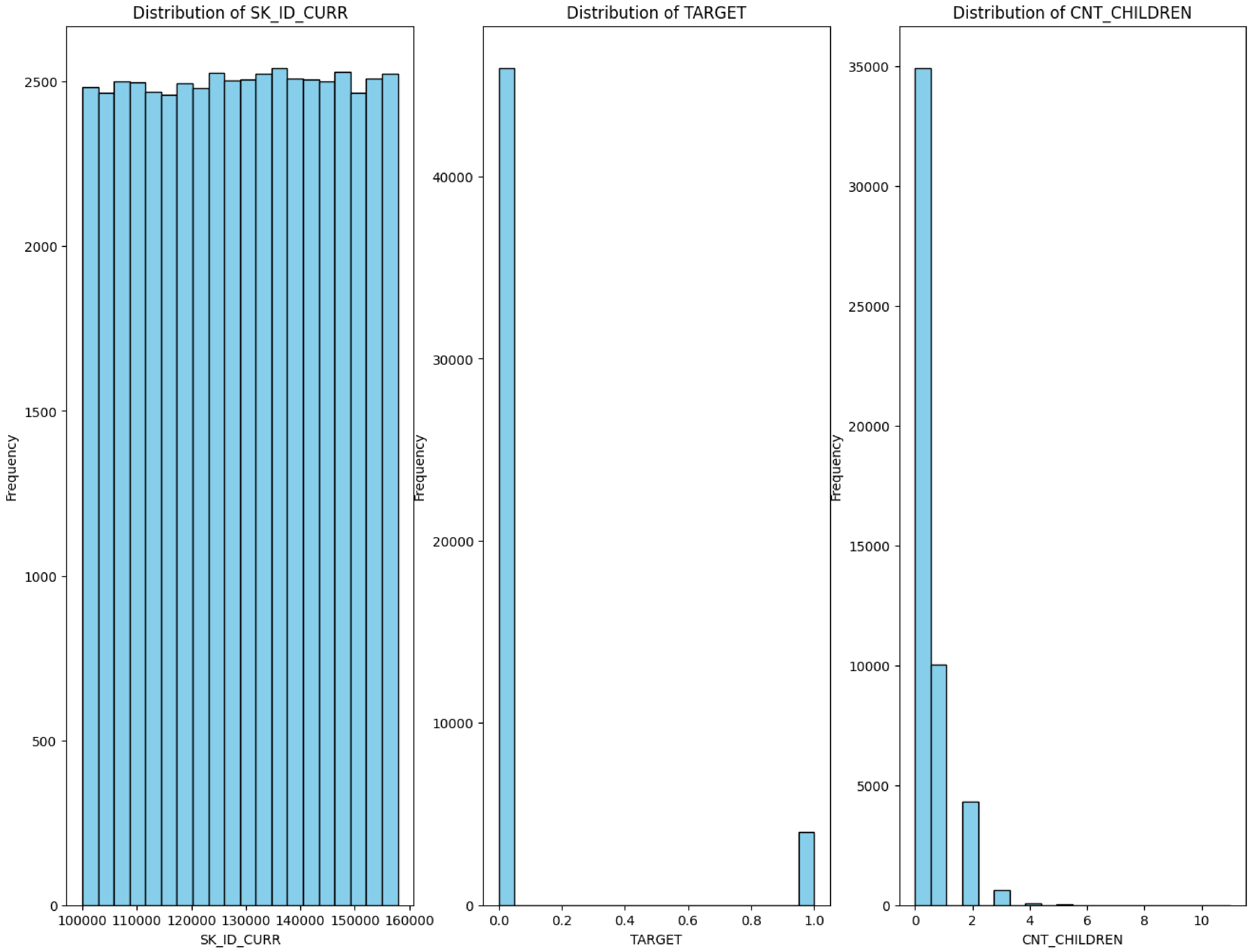
Performing univ# Example for a numerical variable 'income'

```
plt.figure(figsize=(8, 6))
plt.hist(application_data['AMT_INCOME_TOTAL'], bins=20, color='skyblue', edgecolor='black')
plt.title('Distribution of Income')
plt.xlabel('AMT_INCOME_TOTAL')
plt.ylabel('Frequency')
plt.show()
```



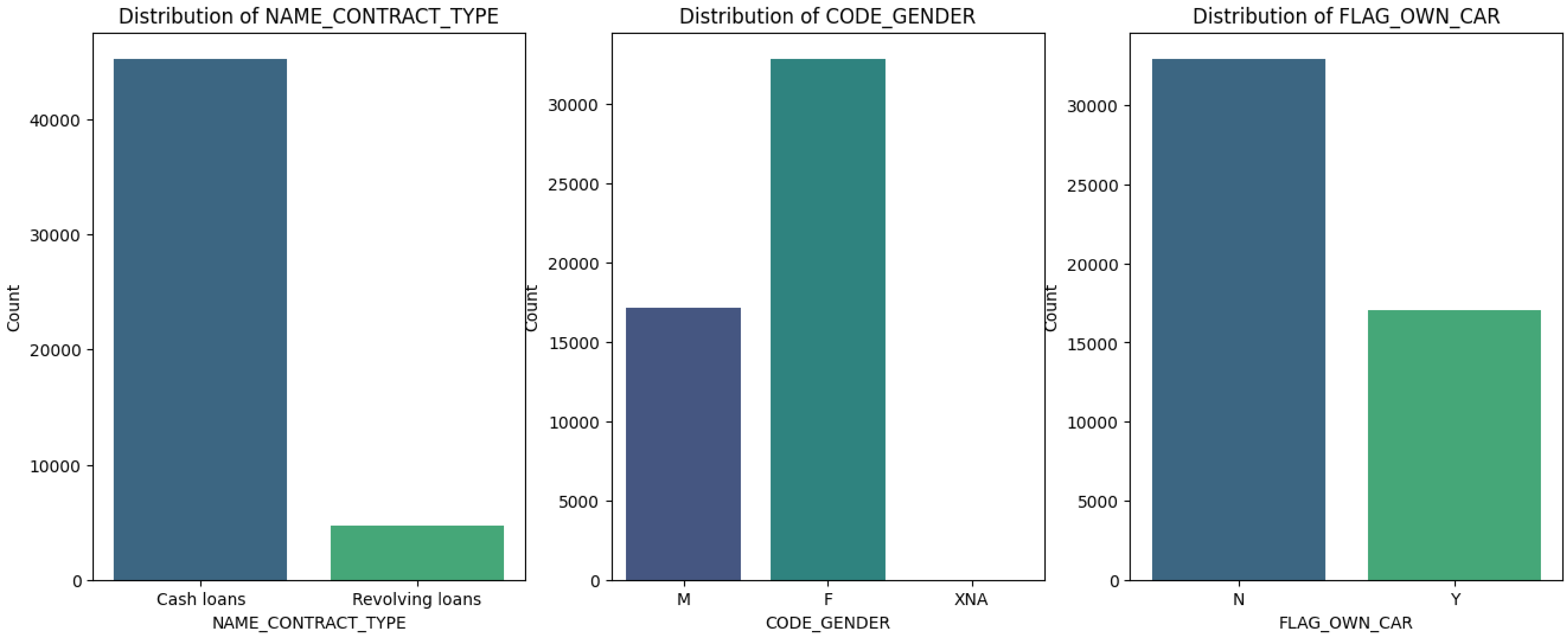
#Histograms for Numerical Variables:

```
# Selecting numerical columns
numerical_columns =
application_data.select_dtypes(include=
['int64', 'float64']).columns

# Plotting histograms for all numerical
columns
plt.figure(figsize=(16, 12))
for i, col in enumerate(numerical_columns,
1):
    plt.subplot(1, 3, i)
    plt.hist(application_data[col], bins=20,
color='skyblue', edgecolor='black')
    plt.title(f'Distribution of {col}')
    plt.xlabel(col)
    plt.ylabel('Frequency')

plt.tight_layout()
plt.show()
```



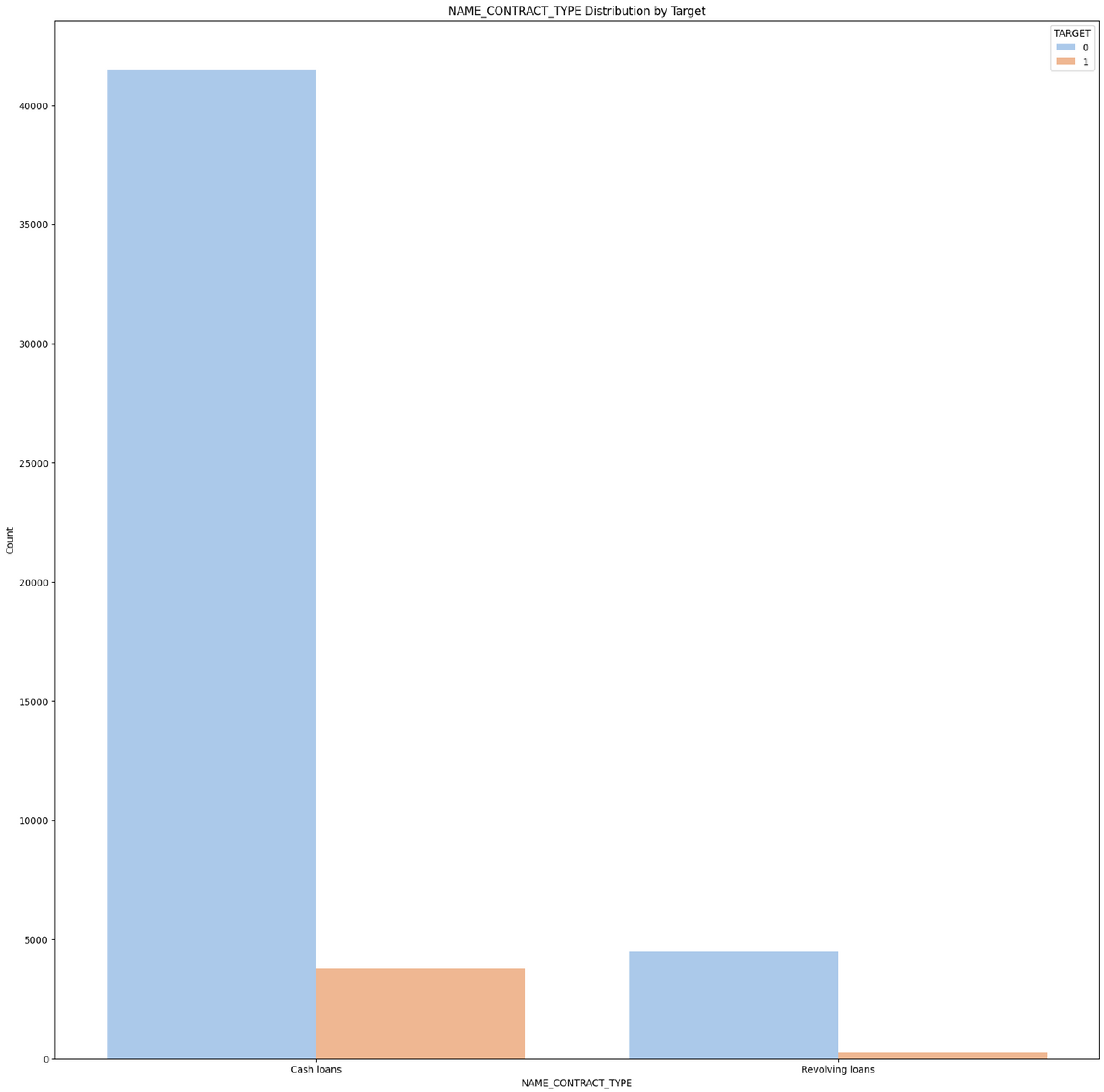
#Bar Charts for Categorical Variables:

```
import matplotlib.pyplot as plt
import seaborn as sns

# Selecting categorical columns
categorical_columns =
application_data.select_dtypes(include=['object']).columns

# Plotting count plots for all categorical columns
plt.figure(figsize=(16, 6))
for i, col in enumerate(categorical_columns, 1):
    plt.subplot(1, 3, i)
    sns.countplot(x=col, data=application_data, palette='viridis')
    plt.title(f'Distribution of {col}')
    plt.xlabel(col)
    plt.ylabel('Count')

plt.tight_layout()
plt.show()
```

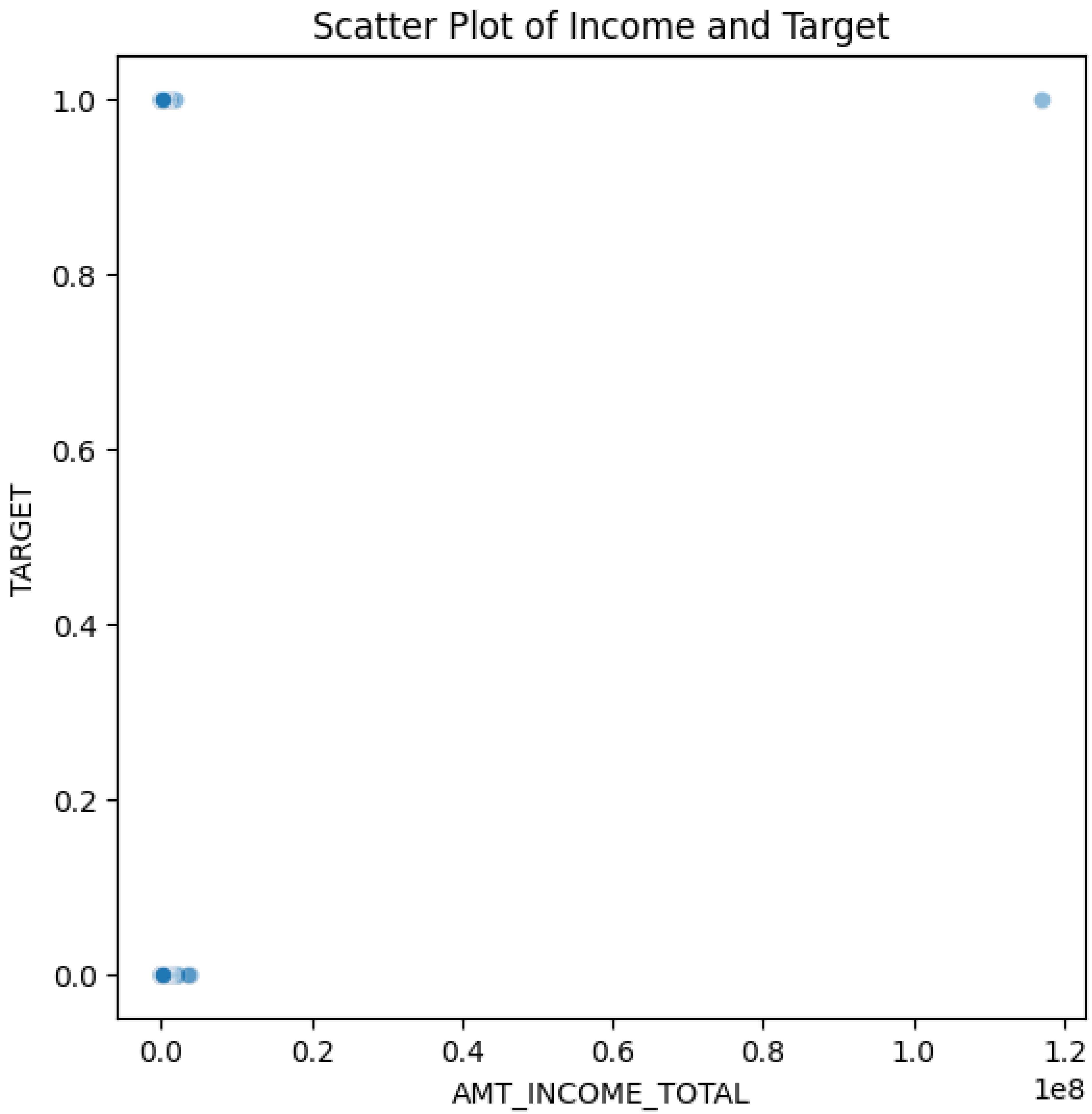


#Segmented Univariate Analysis:
#Grouped Bar Charts for Categorical Variables Across Scenarios:

```
plt.figure(figsize=(20, 20))
sns.countplot(x='NAME_CONTRACT_TYPE', hue='TARGET', data=application_data, palette='pastel')
plt.title('NAME_CONTRACT_TYPE Distribution by Target')
plt.xlabel('NAME_CONTRACT_TYPE')
plt.ylabel('Count')
plt.show()
```



```
#Bivariate Analysis:
#Scatter Plots for Numerical-Target Variable Relationships:
plt.figure(figsize=(6,6))
sns.scatterplot(x='AMT_INCOME_TOTAL', y='TARGET', data=application_data, alpha=0.5)
plt.title('Scatter Plot of Income and Target')
plt.xlabel('AMT_INCOME_TOTAL')
plt.ylabel('TARGET')
plt.show()
```

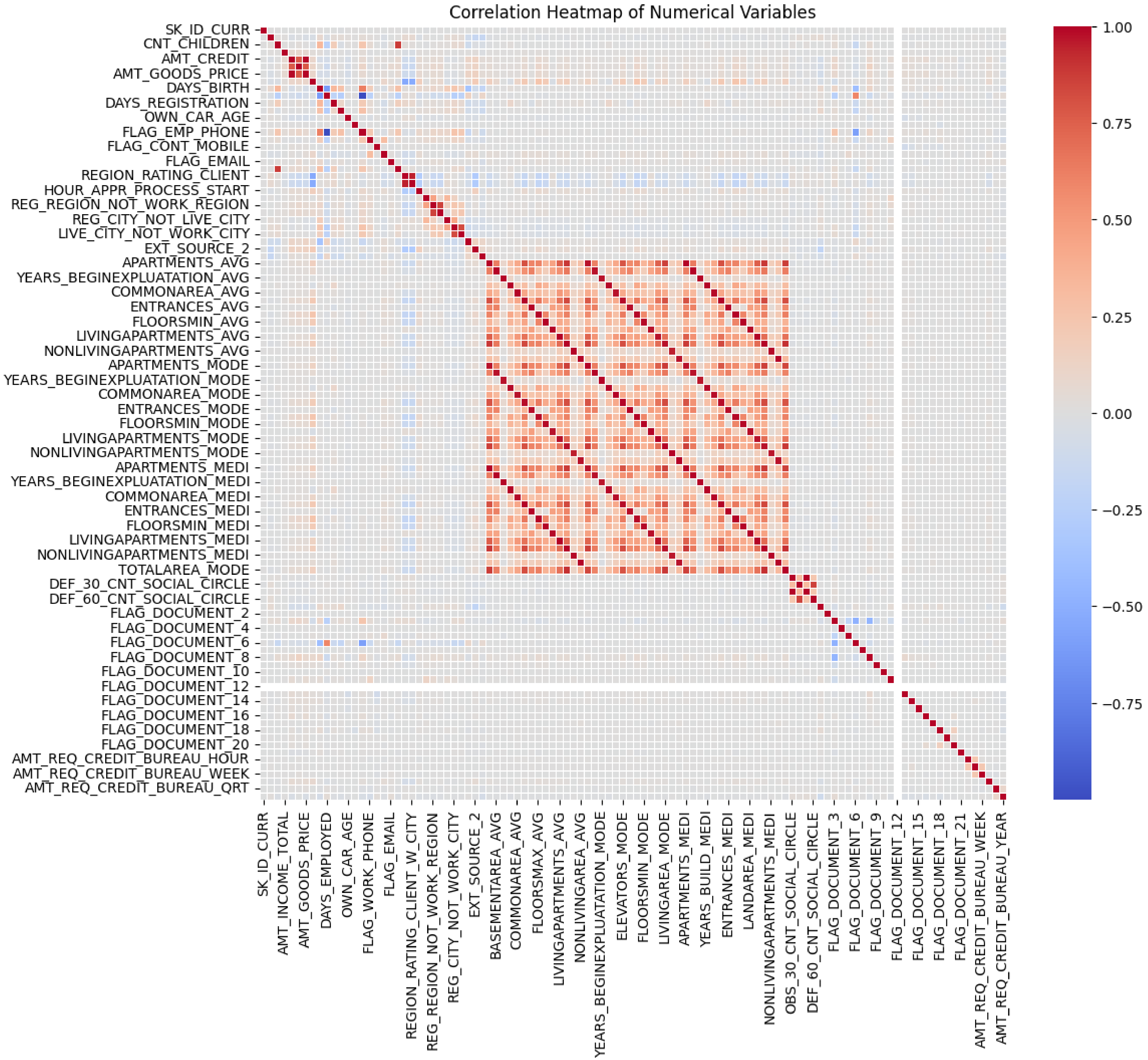


Task E]

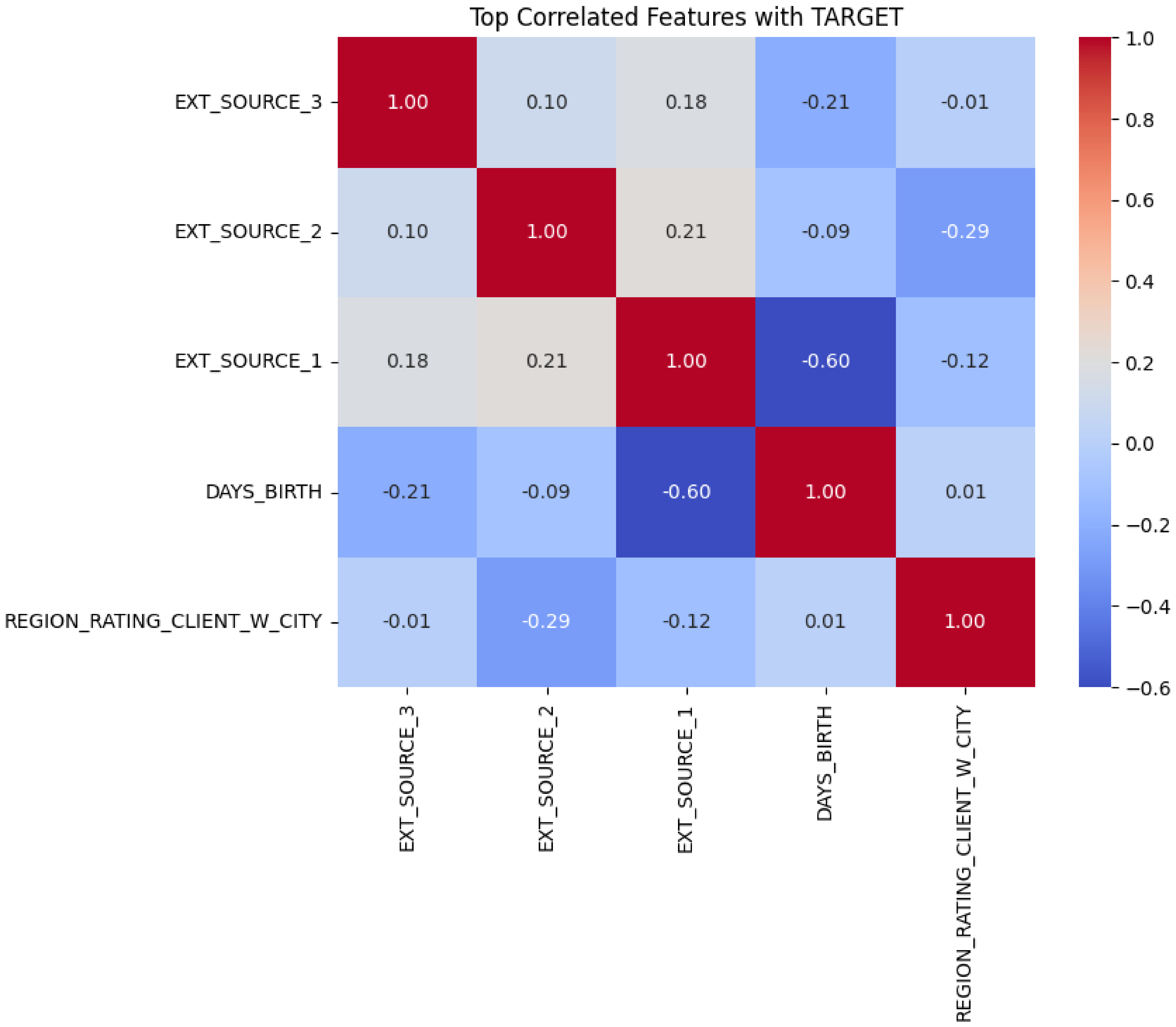
Identify Top Correlations for Different Scenarios: Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.
Created a function (analyze_correlations) to analyze and visualize correlation matrices for different segments.

CORRELATION FOR APPLICANTS WITH PAYMENT MADE ON TIME

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	REGION_POPULATION_RELATIVE	DAYS_BIRTH(Years)	DAYS_EMPLOYED (Years)	DAYS_ID_PUBLISH(Years)	REGION_RATING_CLIENT
CNT_CHILDREN	1	0.027	0.003	-0.024	-0.337	-0.245	0.029	0.023
AMT_INCOME_TOTAL	0.027	1	0.343	0.168	-0.063	-0.140	-0.023	-0.187
AMT_CREDIT	0.003	0.343	1	0.101	0.047	-0.070	0.001	-0.103
REGION_POPULATION_RELATIVE	-0.024	0.168	0.101	1	0.025	-0.007	0.001	-0.539
DAYS_BIRTH(Years)	-0.337	-0.063	0.047	0.025	1	0.626	0.271	-0.002
DAYS_EMPLOYED (Years)	-0.245	-0.140	-0.070	-0.007	0.626	1	0.277	0.038
DAYS_ID_PUBLISH(Years)	0.029	-0.023	0.001	0.001	0.271	0.277	1	0.009
REGION_RATING_CLIENT	0.023	-0.187	-0.103	-0.539	-0.002	0.038	0.009	1


Correlation Heatmap of Numerical Variables
SK_ID_CURR
CNT_CHILDREN
AMT_CREDIT
AMT_GOODS_PRICE
DAYS_BIRTH
DAYS_REGISTRATION
OWN_CAR_AGE
FLAG_EMP_PHONE
FLAG_CONT_MOBILE
FLAG_EMAIL
REGION_RATING_CLIENT
HOUR_APPR_PROCESS_START
REG_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
LIVE_CITY_NOT_WORK_CITY
EXT_SOURCE_2
APARTMENTS_AVG
YEARS_BEGINEXPLUATATION_AVG
COMMONAREA_AVG
ENTRANCES_AVG
FLOORSMIN_AVG
LIVINGAPARTMENTS_AVG
NONLIVINGAPARTMENTS_AVG
APARTMENTS_MODE
YEARS_BEGINEXPLUATATION_MODE
COMMONAREA_MODE
ENTRANCES_MODE
FLOORSMIN_MODE
LIVINGAPARTMENTS_MODE
NONLIVINGAPARTMENTS_MODE
APARTMENTS_MEDI
YEARS_BEGINEXPLUATATION_MEDI
COMMONAREA_MEDI
ENTRANCES_MEDI
FLOORSMIN_MEDI
LIVINGAPARTMENTS_MEDI
NONLIVINGAPARTMENTS_MEDI
TOTALAREA_MODE
DEF_30_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
FLAG_DOCUMENT_2
FLAG_DOCUMENT_4
FLAG_DOCUMENT_6
FLAG_DOCUMENT_8
FLAG_DOCUMENT_10
FLAG_DOCUMENT_12
FLAG_DOCUMENT_14
FLAG_DOCUMENT_16
FLAG_DOCUMENT_18
FLAG_DOCUMENT_20
AMT_REQ_CREDIT_BUREAU_HOUR
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_QRT
SK_ID_CURR
AMT_INCOME_TOTAL
AMT_GOODS_PRICE
DAYS_EMPLOYED
OWN_CAR_AGE
FLAG_WORK_PHONE
FLAG_EMAIL
REGION_RATING_CLIENT_W_CITY
REG_REGION_NOT_WORK_REGION
REG_CITY_NOT_WORK_CITY
EXT_SOURCE_2
BASEMENTAREA_AVG
COMMONAREA_AVG
FLOORSMAX_AVG
LIVINGAPARTMENTS_AVG
NONLIVINGAREA_AVG
YEARS_BEGINEXPLUATATION_MODE
ELEVATORS_MODE
FLOORSMIN_MODE
LIVINGAREA_MODE
APARTMENTS_MEDI
YEARS_BUILD_MEDI
ENTRANCES_MEDI
LANDAREA_MEDI
NONLIVINGAPARTMENTS_MEDI
OBS_30_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
FLAG_DOCUMENT_3
FLAG_DOCUMENT_6
FLAG_DOCUMENT_9
FLAG_DOCUMENT_12
FLAG_DOCUMENT_15
FLAG_DOCUMENT_18
FLAG_DOCUMENT_21
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_YEAR
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

```
application_data = pd.read_csv('application_data.csv')

def analyze_correlations(data, target_column, top_n=5):
    # Selecting only numeric columns for correlation analysis
    numeric_data = data.select_dtypes(include=[np.number])

    # Calculating correlation matrix
    correlation_matrix = numeric_data.corr()

    # Getting correlations with the target variable
    target_correlations = correlation_matrix[target_column].drop(target_column)

    # Getting the top correlated features
    top_correlations = target_correlations.abs().sort_values(ascending=False).head(top_n)

    # Plotting heatmap
    plt.figure(figsize=(8, 6))
    sns.heatmap(correlation_matrix.loc[top_correlations.index, top_correlations.index], annot=True, cmap='coolwarm', fmt=".2f")
    plt.title(f'Top Correlated Features with {target_column}')
    plt.show()

# Assuming 'TARGET' is the column indicating payment difficulties
analyze_correlations(application_data, target_column='TARGET', top_n=5)
```


Top Correlated Features with TARGET
EXT_SOURCE_3
EXT_SOURCE_2
EXT_SOURCE_1
DAYS_BIRTH
REGION_RATING_CLIENT_W_CITY
1.00
0.10
0.18
-0.21
-0.01
0.29
0.21
-0.09
-0.29
-0.60
-0.12
0.01
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Through this project, I accomplished several key objectives:

- Handling Missing Data:
- Successfully addressed missing data by imputing values based on data types.
- Achieved a dataset with minimal missing values.
- Outlier Identification:
- Identified outliers in numerical columns, providing insights into potential data quality issues.
- Data Imbalance:
- Explored the distribution of the target variable, revealing an imbalanced dataset.
- Exploratory Data Analysis:
- Conducted a comprehensive EDA, revealing patterns and trends in the data.
- Utilized various visualization techniques for effective data exploration.
- Correlation Analysis:
- Examined correlations between numerical variables in different scenarios.
- Identified top correlated features with the target variable.

Here are some key insights that I obtained from the Bank Loan Case Study project:

1. Income and Loan Requests:
    - Individuals with higher incomes are less likely to apply for loans.
    - The credit amount of a bank loan typically falls within the range of 45,000 to 1,045,000.
    - The majority of loan applications come from people between the ages of 35 and 50.
    - Those with 0 to 8 years of work experience are more likely to seek loans.
2. Homeownership and Marital Status:
    - Individuals who own homes are more likely to apply for loans.
    - Married individuals are more inclined to take out loans compared to singles or those with other marital statuses.
3. Employment Status:
    - Individuals with jobs are more likely to request loans.
    - Unaccompanied minors have requested additional loans.
4. Loan Outcomes:
    - Customers who live in low-rating areas are more likely to have loan defaults.
    - Individuals with lower incomes are more likely to default.
    - Younger individuals are more likely to default, with the trend of defaulters decreasing with age.
    - Females are less likely than males to have defaults.

- Family Size and Education:
  - Customers with more than five family members are more likely to default on their bank loans.
  - Customers with fewer educational qualifications are more likely to fail to repay their loans.
  - Clients with little work experience are more likely to have defaults.
- Loan Types and Approval Rates:
  - Consumer loans have a significantly lower rate of cancellations and the highest approval rate.
  - Loans requested for the first Selling Place Area group experienced a higher rate of cancellations.
- Previous Loan History:
  - Clients who have applied for previous loans tend to have no defaults in their current loans.
- Top Correlations with Loan Default:
  - The top factors correlated with loan default include income type, family size, children count, external source, region rating of the client, age, months employed, amount credit, amount goods price, and amount total income.

Result

- The project successfully achieved its objectives, providing valuable insights into the bank loan dataset. By addressing missing data, identifying outliers, and conducting thorough exploratory data analysis, the team gained a better understanding of the dataset's characteristics. The correlation analysis helped in identifying features strongly correlated with the target variable.

These insights are valuable for the finance company to make informed decisions about loan approvals, risk assessments, and strategies to reduce loan defaults. They highlight the importance of considering various customer attributes and loan attributes when evaluating loan applications.