Univariate Analysis

It is the simplest form of data analysis where we examine each variable individually. The purpose of this analysis is to understand the distribution and characteristics of each feature in the dataset. It includes statistical summaries and visualizations such as histograms, box plots, and bar charts.

Features of Univariate Data Analysis:

- Central Tendency: Measures like mean, median, and mode.
- Dispersion: Measures like range, variance, and standard deviation.
- Shape: Skewness and kurtosis of the distribution.
- Visualization: Histograms for numerical data, and bar plots for categorical data.

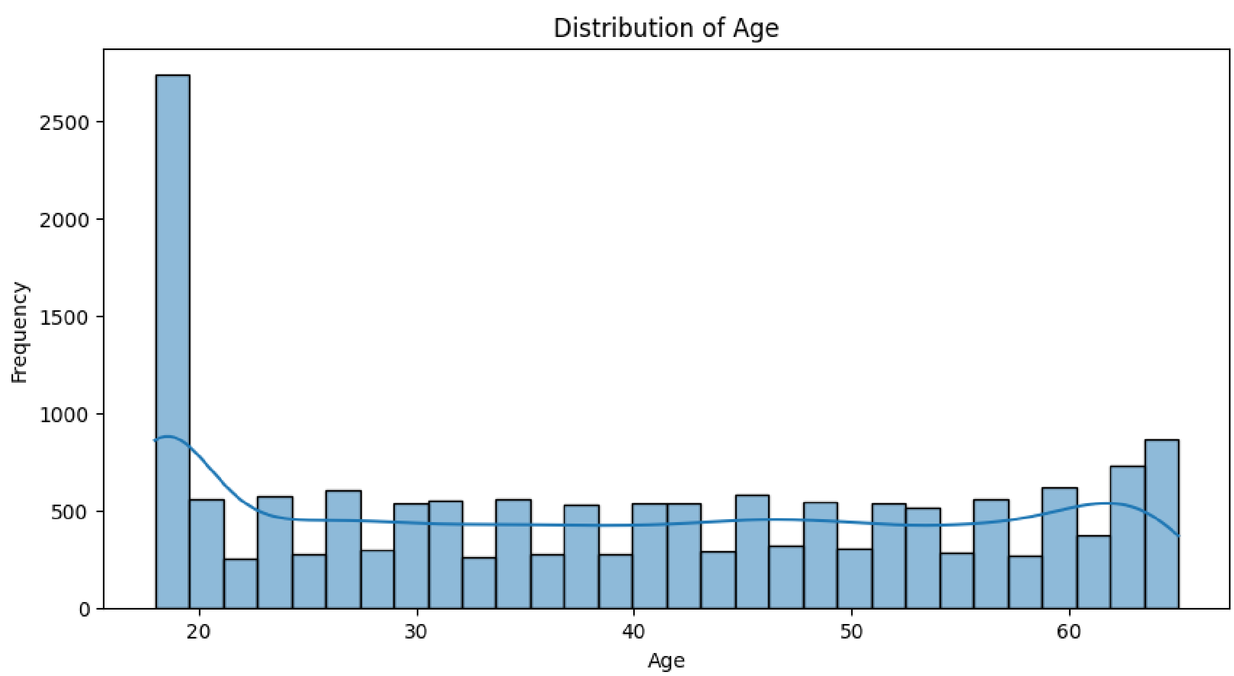
```
# importing libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import *

# Load the data
file_path = 'processed_loan_data.csv'
data = pd.read_csv(file_path)

# Display the first few rows of the data
print("First few rows of the dataset:")
print(data.head())
```

```
First few rows of the dataset:
  Customer ID               Name Gender  Age  Income (USD) Income
Stability  \
0     C-36995  Frederica Shealy      F   56       1933.05
Low
1     C-23855  Nathalie Olivier      M   43       2361.56
Low
2     C-24944    Barbie Goetsch      M   18       1546.17
Low
3     C-40801      Laree Staton      M   18       2416.86
Low
4     C-30073        Brinda Vaz      F   48        777.25
Low

      Profession Type of Employment    Location  Loan Amount Request
(USD)  \
0        Working        Sales staff  Semi-Urban
72809.58
1        Working           Laborers  Semi-Urban
```

```
152561.34
2        Working           Laborers       Rural
42091.29
3  State servant         Core staff  Semi-Urban
25765.72
4        Working           Laborers  Semi-Urban
96080.60

   ...  Credit Score No. of Defaults Has Active Credit Card  Property
ID  \
0  ...        809.44               0                 Active
746
1  ...        637.29               0            Unpossessed
227
2  ...        613.24               0            Unpossessed
883
3  ...        652.41               0                 Active
325
4  ...        764.11               0                 Active
678

   Property Age  Property Type Property Location  Co-Applicant  \
0       1933.05              4             Rural             1
1       2361.56              1        Semi-Urban             1
2       1546.17              2             Urban             1
3       2416.86              2             Rural             1
4        777.25              1        Semi-Urban             1

   Property Price  Loan Sanction Amount (USD)
0       119933.46                    54607.18
1       221050.80                        0.00
2        67993.43                        0.00
3        32423.71                    16747.72
4       146073.26                    67256.42

[5 rows x 24 columns]

# Univariate Analysis for Numerical Features

numerical_columns = data.select_dtypes(include=['float64',
'int64']).columns

print("\nSummary Statistics for Numerical Features:")
for column in numerical_columns:
    print(f"\nStatistics for {column}:")
    print(f"Mean: {data[column].mean()}")
    print(f"Median: {data[column].median()}")

    # Measures of Dispersion
    print(f"Range: {data[column].max() - data[column].min()}")
```

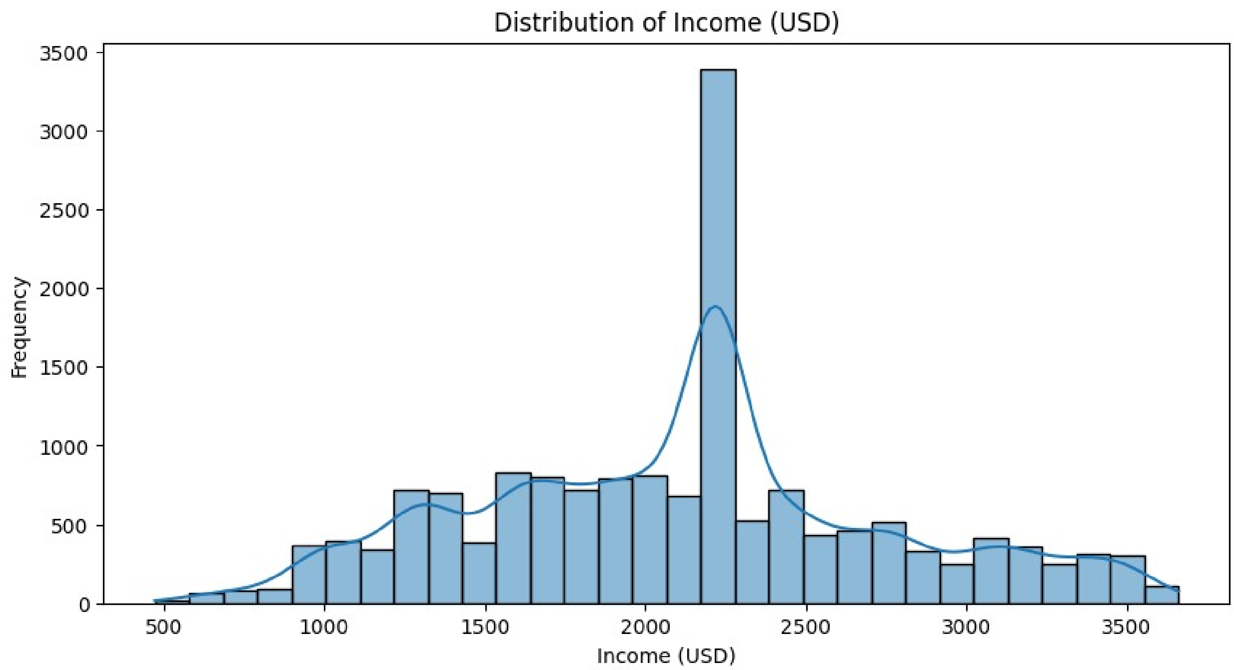
```
    print(f"Variance: {data[column].var()}")
    print(f"Standard Deviation: {data[column].std()}")

    # Measures of Shape
    print(f"Skewness: {data[column].skew()}")
    print(f"Kurtosis: {data[column].kurtosis()}")

    # Histogram
    plt.figure(figsize=(10, 5))
    sns.histplot(data[column], kde=True, bins=30)
    plt.title(f'Distribution of {column}')
    plt.xlabel(column)
    plt.ylabel('Frequency')
    plt.show()
```

```
Summary Statistics for Numerical Features:

Statistics for Age:
Mean: 39.14283055159149
Median: 39.0
Range: 47
Variance: 246.25596947569096
Standard Deviation: 15.692545028633532
Skewness: 0.10005717290869368
Kurtosis: -1.3460309166884092
```

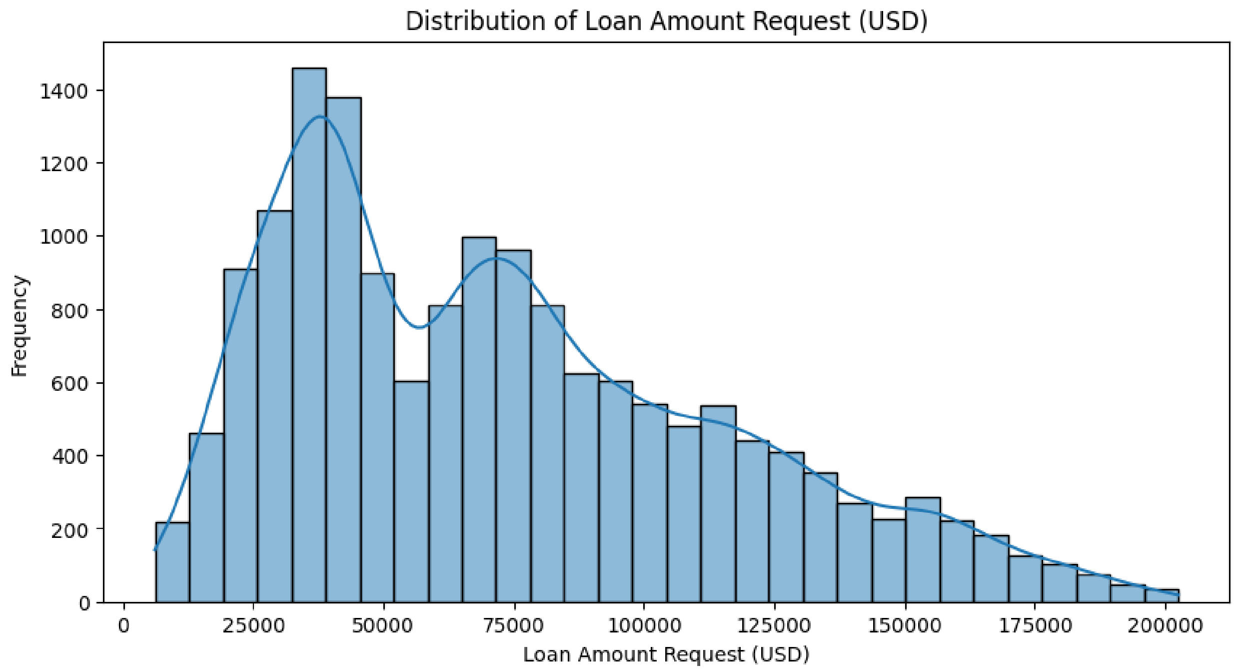



```
Statistics for Income (USD):
```

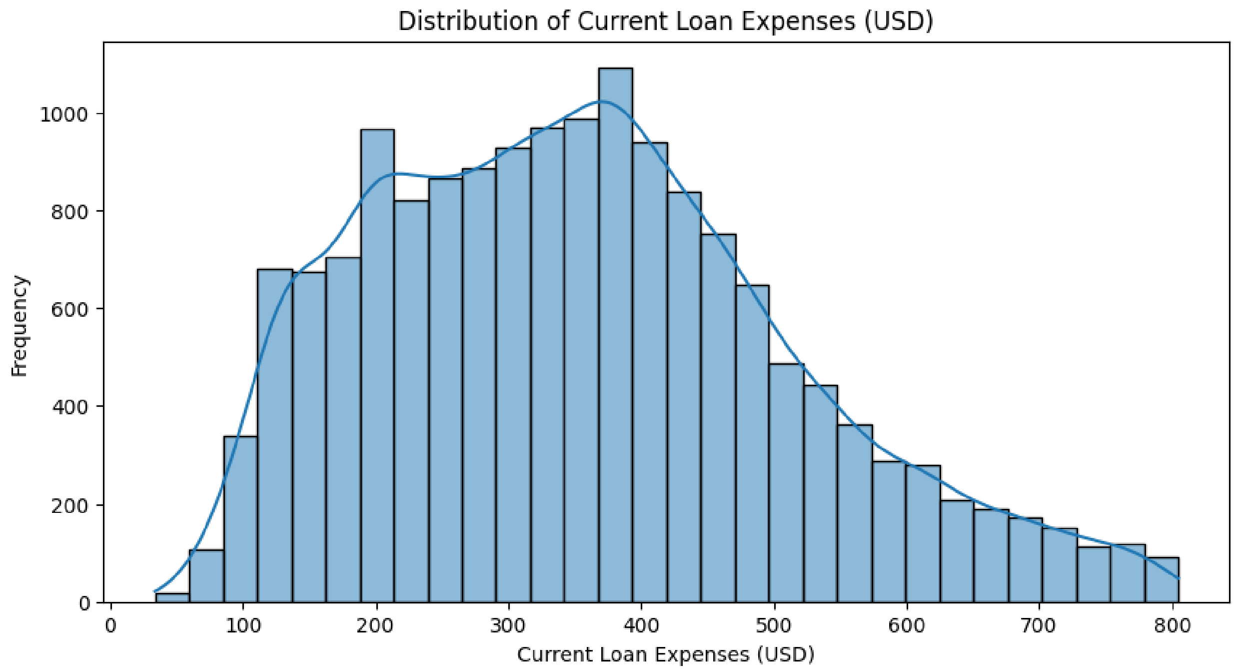
```
Mean: 2106.1819553887203
Median: 2222.435
Range: 3188.03
Variance: 412714.03950779355
Standard Deviation: 642.4282368543537
Skewness: 0.16199362526661434
Kurtosis: -0.3350356751557846
```



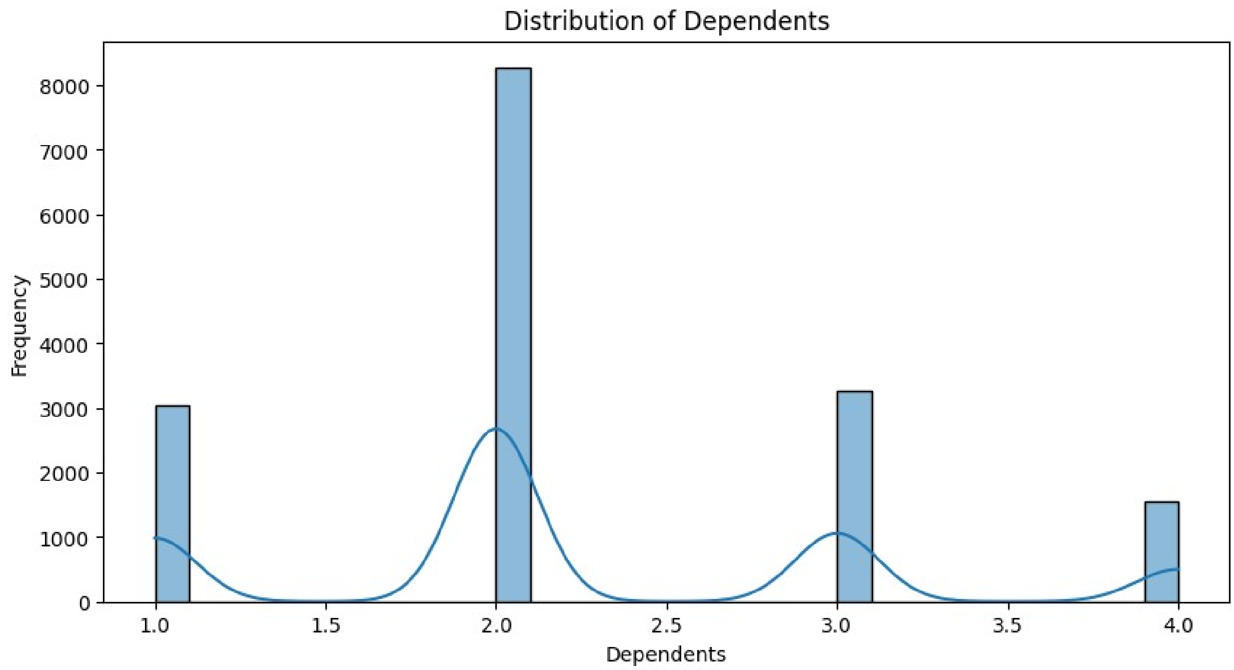

```
Statistics for Loan Amount Request (USD):
Mean: 73492.24677421356
Median: 67002.87
Range: 196404.06
Variance: 1797192539.0875115
Standard Deviation: 42393.30771581184
Skewness: 0.7012425386536644
Kurtosis: -0.2902449301530061
```



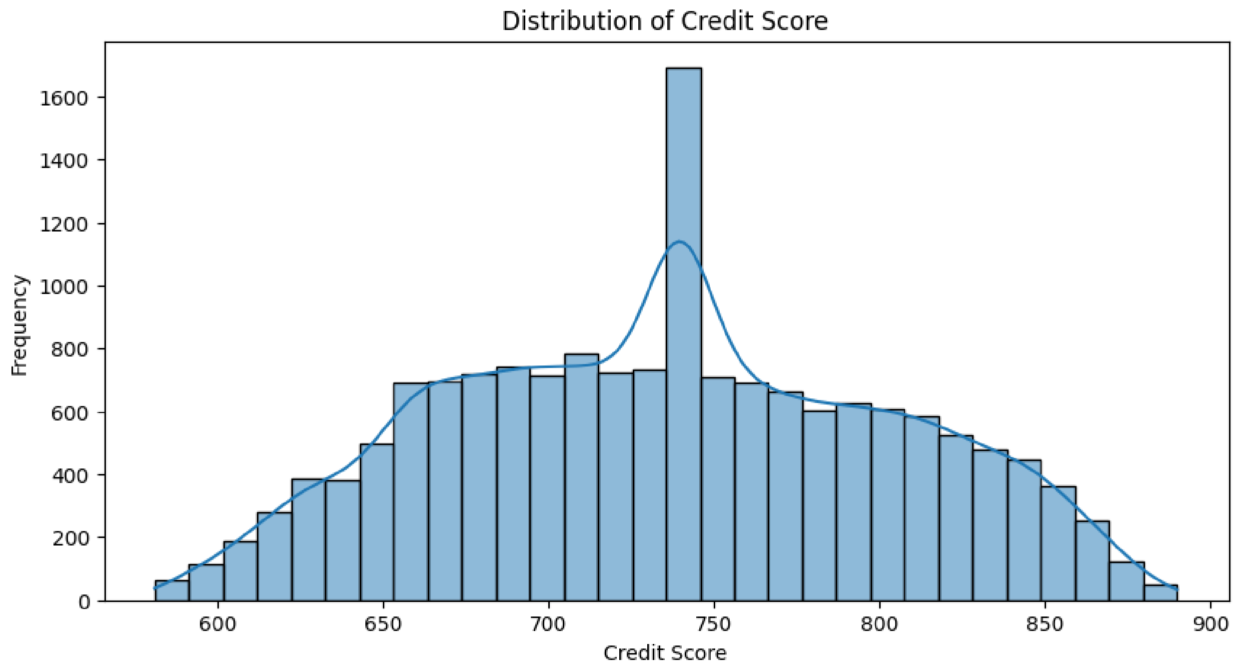

```
Statistics for Current Loan Expenses (USD):
Mean: 355.22707917106163
Median: 344.89
Range: 771.14
Variance: 24991.802391811056
Standard Deviation: 158.0879577697525
Skewness: 0.4818399929065286
Kurtosis: -0.25796560682328096
```



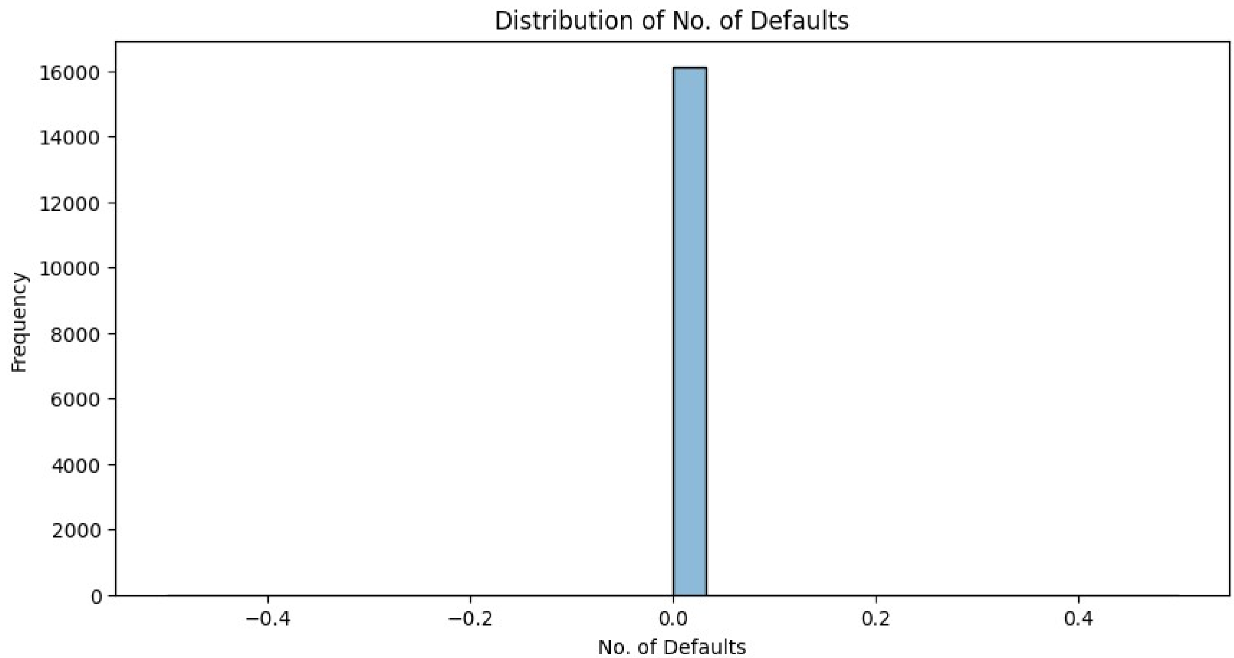

```
Statistics for Dependents:
Mean: 2.2053111621269466
Median: 2.0
Range: 3.0
Variance: 0.7319201640929469
Standard Deviation: 0.8555233276147103
Skewness: 0.5116059326060639
Kurtosis: -0.24980917823408477
```



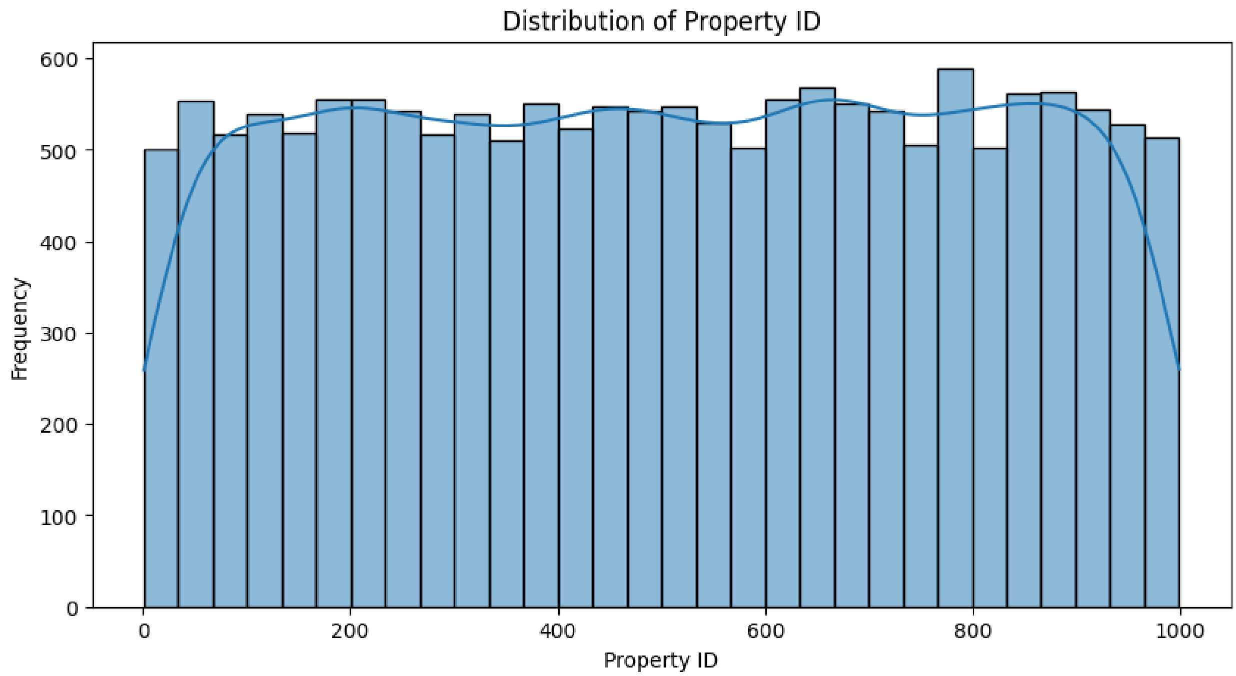

```
Statistics for Credit Score:
Mean: 736.0751504622448
Median: 739.82
Range: 309.16999999999996
Variance: 4522.564718550518
Standard Deviation: 67.25001649479736
Skewness: 0.03977036992610049
Kurtosis: -0.7673695991514542
```



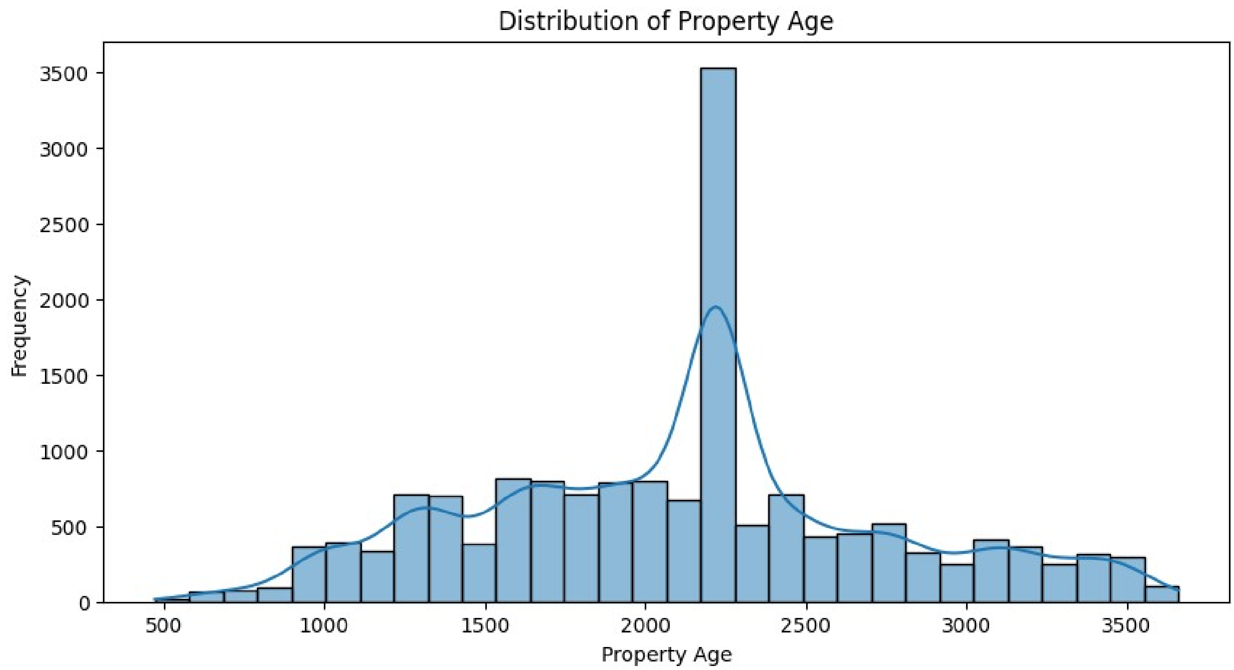

```
Statistics for No. of Defaults:
Mean: 0.0
Median: 0.0
Range: 0
Variance: 0.0
Standard Deviation: 0.0
Skewness: 0.0
Kurtosis: 0.0
```



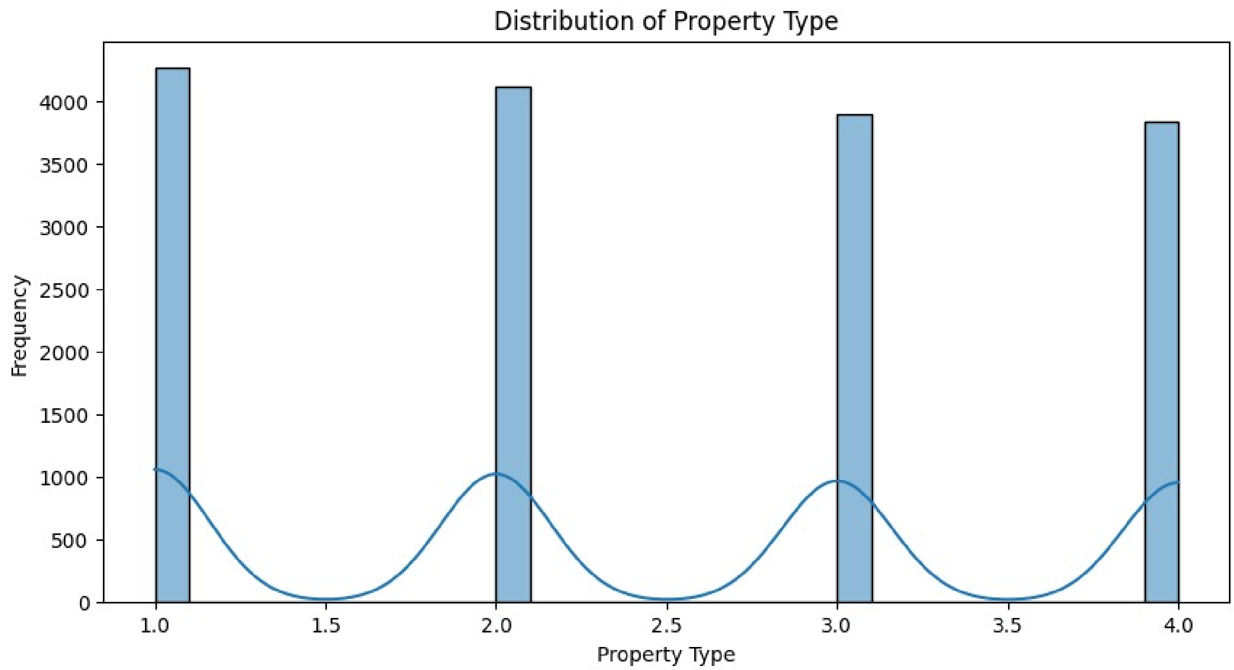

```
Statistics for Property ID:
Mean: 501.88018862071107
Median: 502.0
Range: 998
Variance: 82456.8082430024
Standard Deviation: 287.1529352853674
Skewness: -0.010724321653646328
Kurtosis: -1.2012606875450489
```



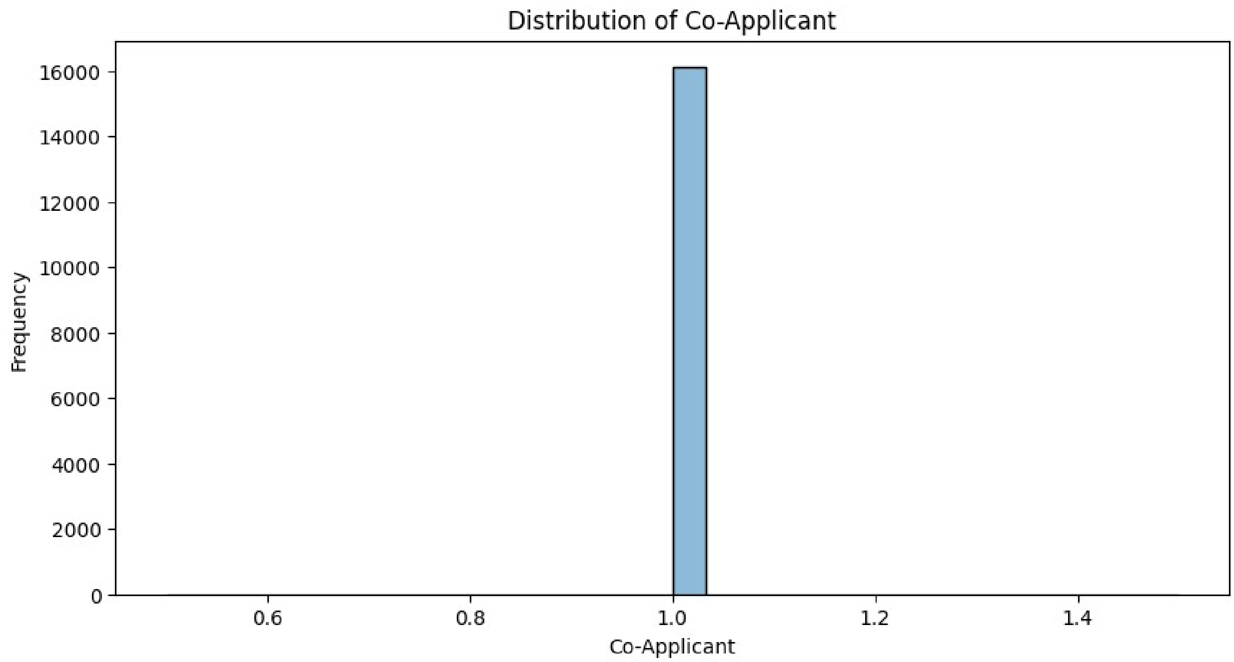

```
Statistics for Property Age:
Mean: 2107.719055035056
Median: 2223.25
Range: 3188.03
Variance: 408130.65767462825
Standard Deviation: 638.8510449820272
Skewness: 0.15656135950071126
Kurtosis: -0.3109907048635585
```



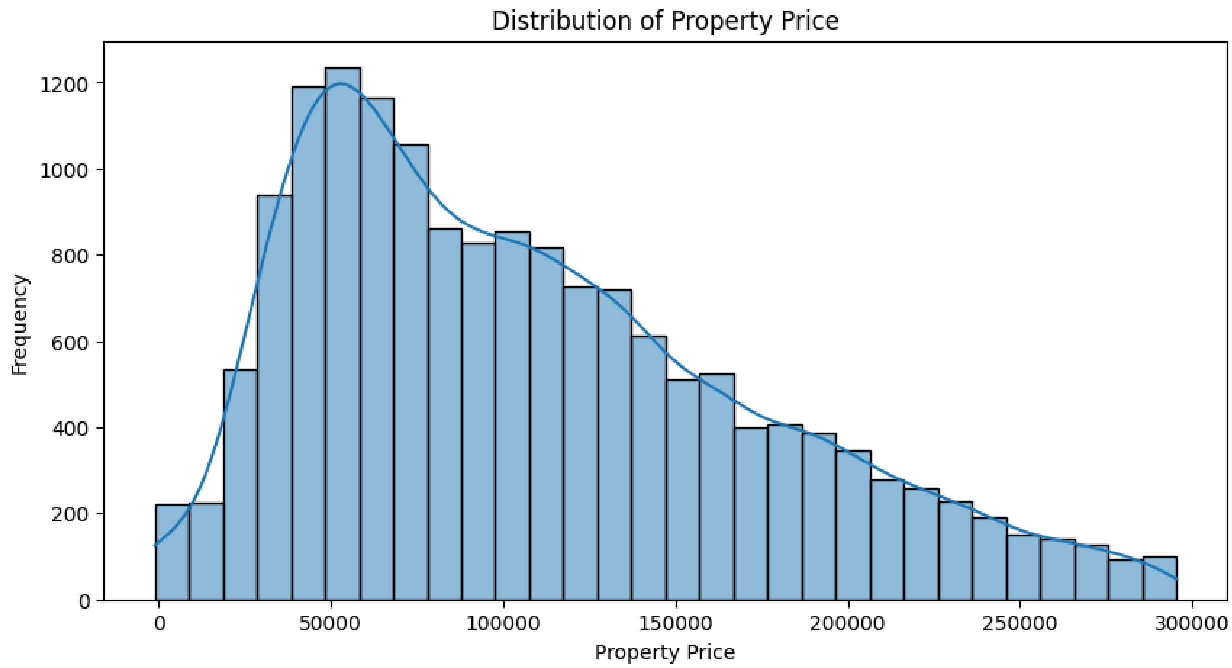

```
Statistics for Property Type:
Mean: 2.4534963082459513
Median: 2.0
Range: 3
Variance: 1.2534372972840866
Standard Deviation: 1.119570139510735
Skewness: 0.060028637117361575
Kurtosis: -1.360424209559182
```



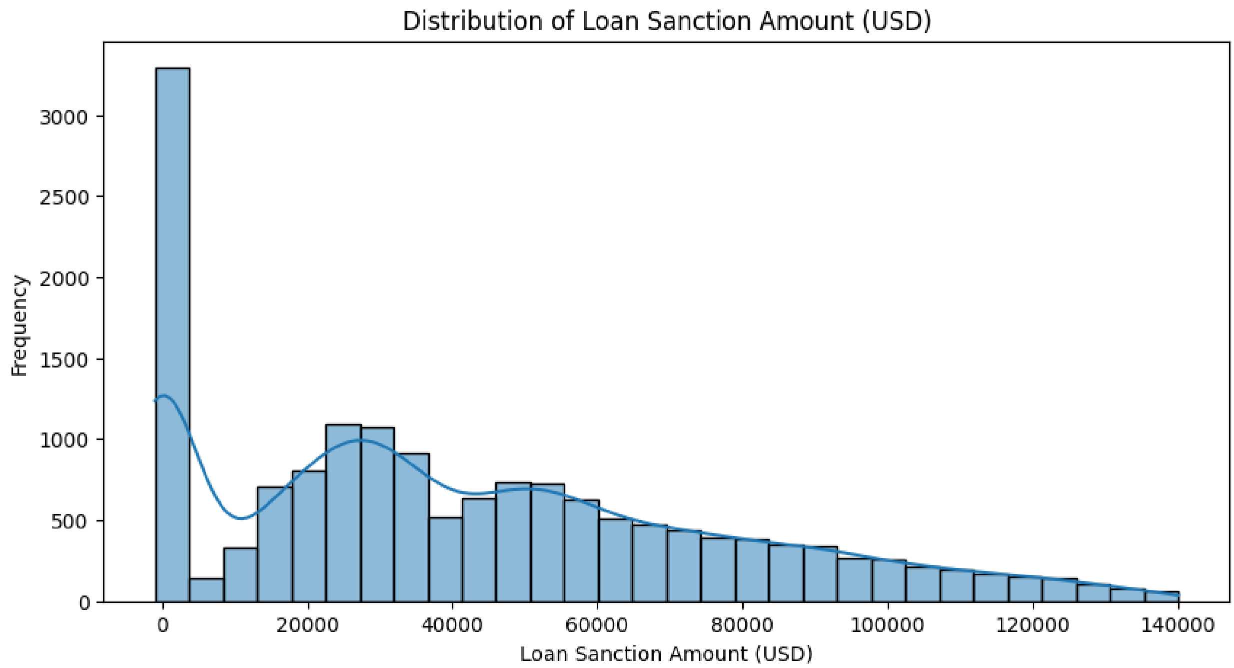

```
Statistics for Co-Applicant:
Mean: 1.0
Median: 1.0
Range: 0
Variance: 0.0
Standard Deviation: 0.0
Skewness: 0.0
Kurtosis: 0.0
```




```
Statistics for Property Price:
Mean: 107986.96156108456
Median: 95311.62
Range: 296283.39
Variance: 4270444945.8067193
Standard Deviation: 65348.64149932055
Skewness: 0.7075223482424297
Kurtosis: -0.21658929923056425
```



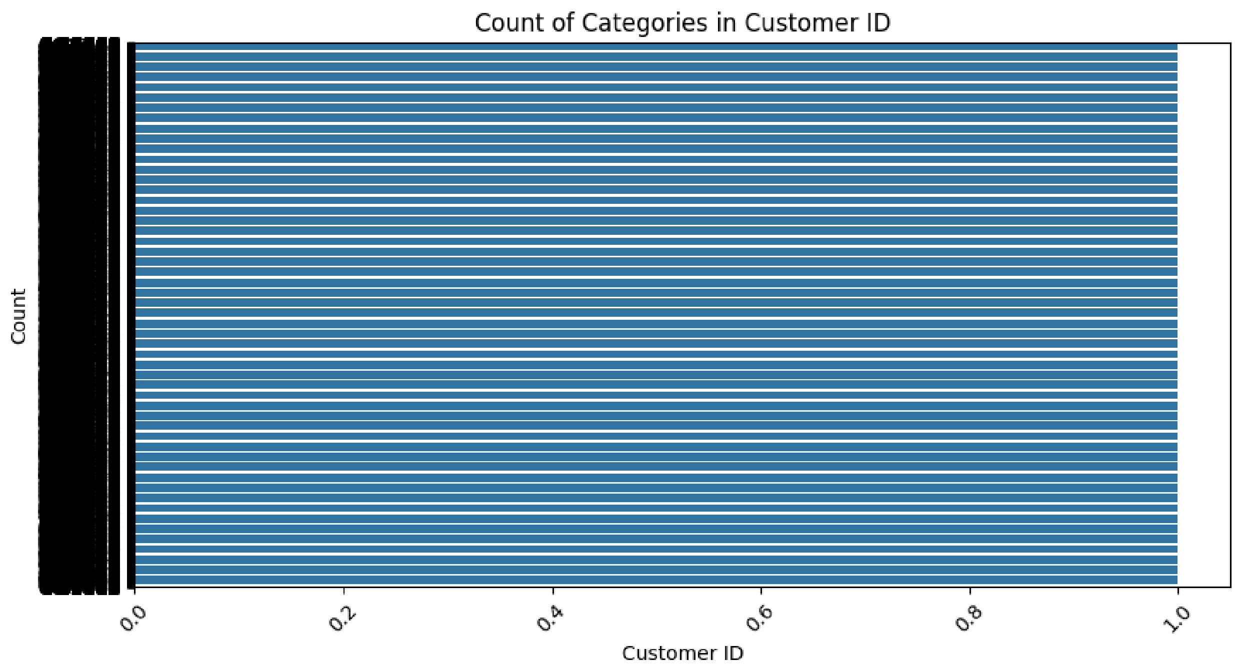

```
Statistics for Loan Sanction Amount (USD):
Mean: 42144.20699385742
Median: 35209.395000000004
Range: 141044.96
Variance: 1213051686.4161046
Standard Deviation: 34828.89154733616
Skewness: 0.6323560175985564
Kurtosis: -0.39898553108919677
```



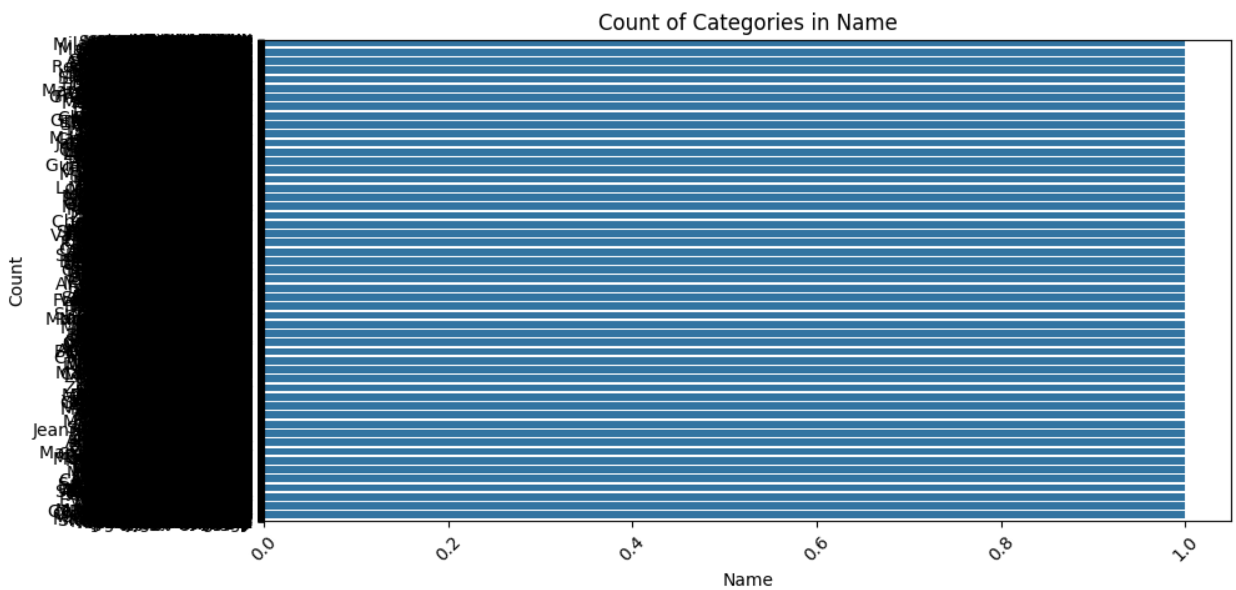

```
# Univariate Analysis for Categorical Features

categorical_columns = data.select_dtypes(include=['object']).columns

print("\nSummary Statistics for Categorical Features:")
for column in categorical_columns:
    print(f"\nStatistics for {column}:")
    print(data[column].value_counts())
    print(f"Mode: {data[column].mode()[0]}")

    # Bar Plot
    plt.figure(figsize=(10, 5))
    sns.countplot(data[column])
    plt.title(f'Count of Categories in {column}')
    plt.xlabel(column)
    plt.ylabel('Count')
    plt.xticks(rotation=45)
    plt.show()
```

```
Summary Statistics for Categorical Features:

Statistics for Customer ID:
Customer ID
C-36995    1
C-11986    1
C-35365    1
C-15737    1
C-8105     1
```

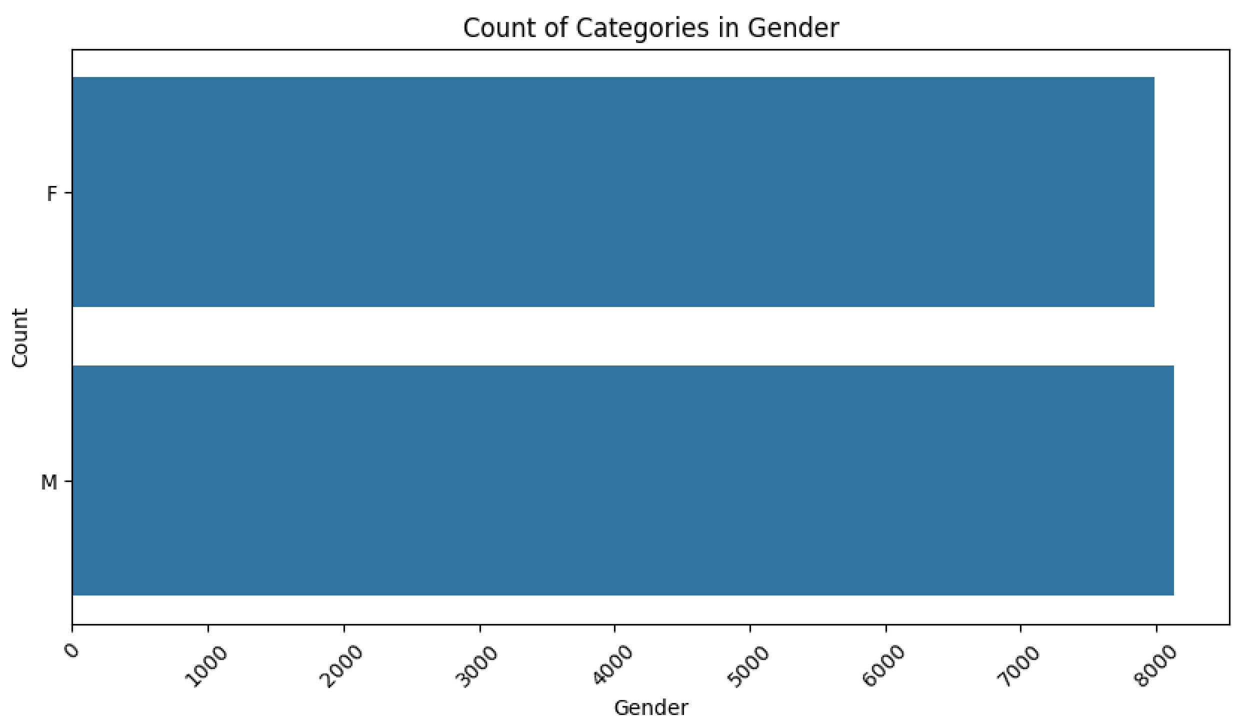
```
             ..
C-40186    1
C-5722     1
C-30014    1
C-5509     1
C-33003    1
Name: count, Length: 16117, dtype: int64
Mode: C-0
```



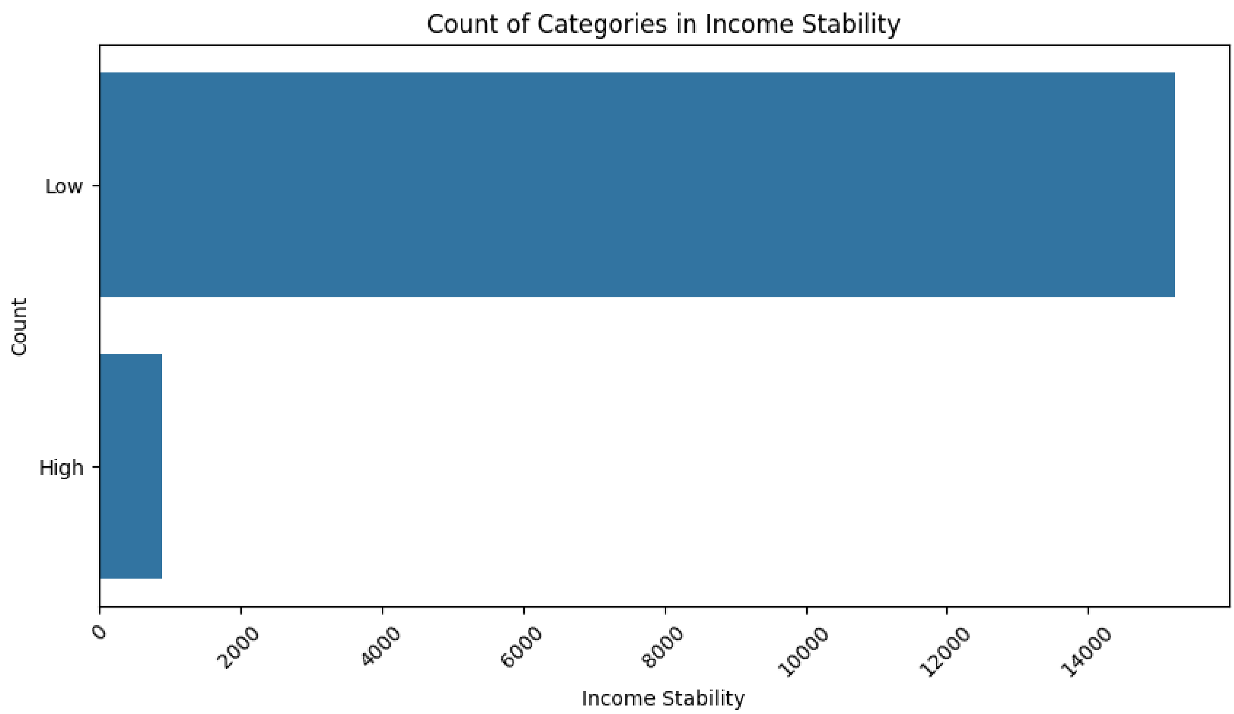

```

Statistics for Name:
Name
Frederica Shealy     1
Adina Sprankle       1
Donald Bjornson      1
Brenton Needles      1
Ted Gatto            1
                    ..
Aletha Hipps         1
Luciano Vautour      1
Elden Bumbrey        1
Leeanne Godin        1
Bridget Garibaldi    1
Name: count, Length: 16117, dtype: int64
Mode: Aaron Bingham
```



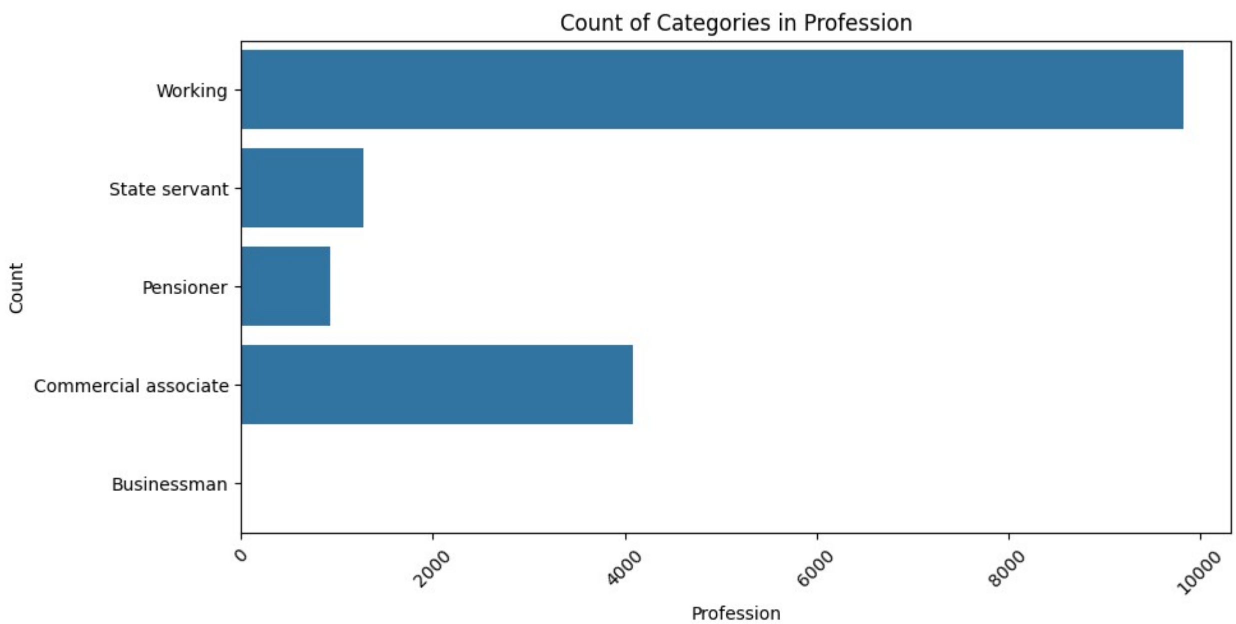

```
Statistics for Gender:
Gender
M    8134
F    7983
Name: count, dtype: int64
Mode: M
```



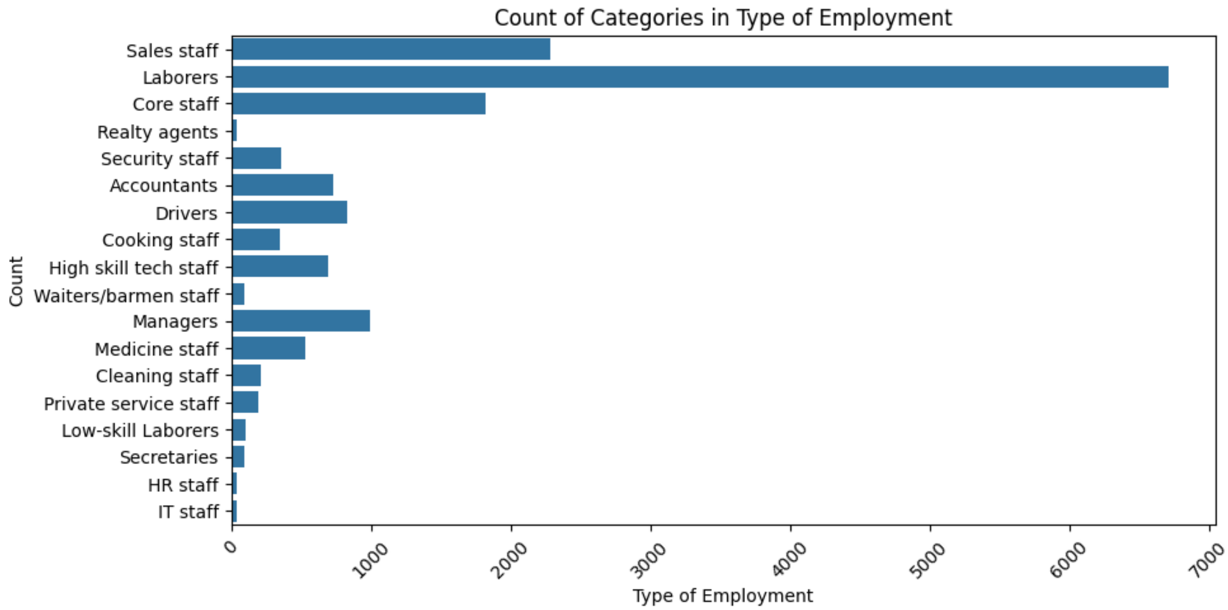
```
Statistics for Income Stability:
Income Stability
Low     15238
High      879
Name: count, dtype: int64
Mode: Low
```



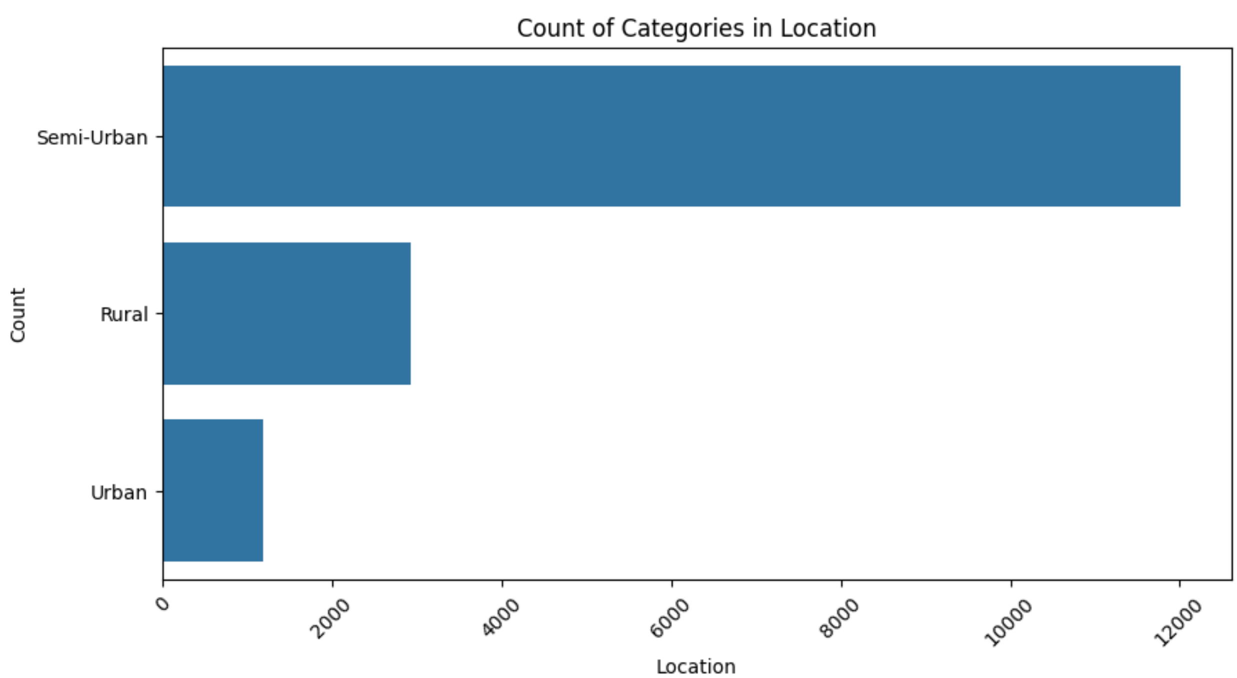

```
Statistics for Profession:
Profession
Working                 9835
Commercial associate    4079
State servant           1269
Pensioner                933
Businessman                1
Name: count, dtype: int64
Mode: Working
```



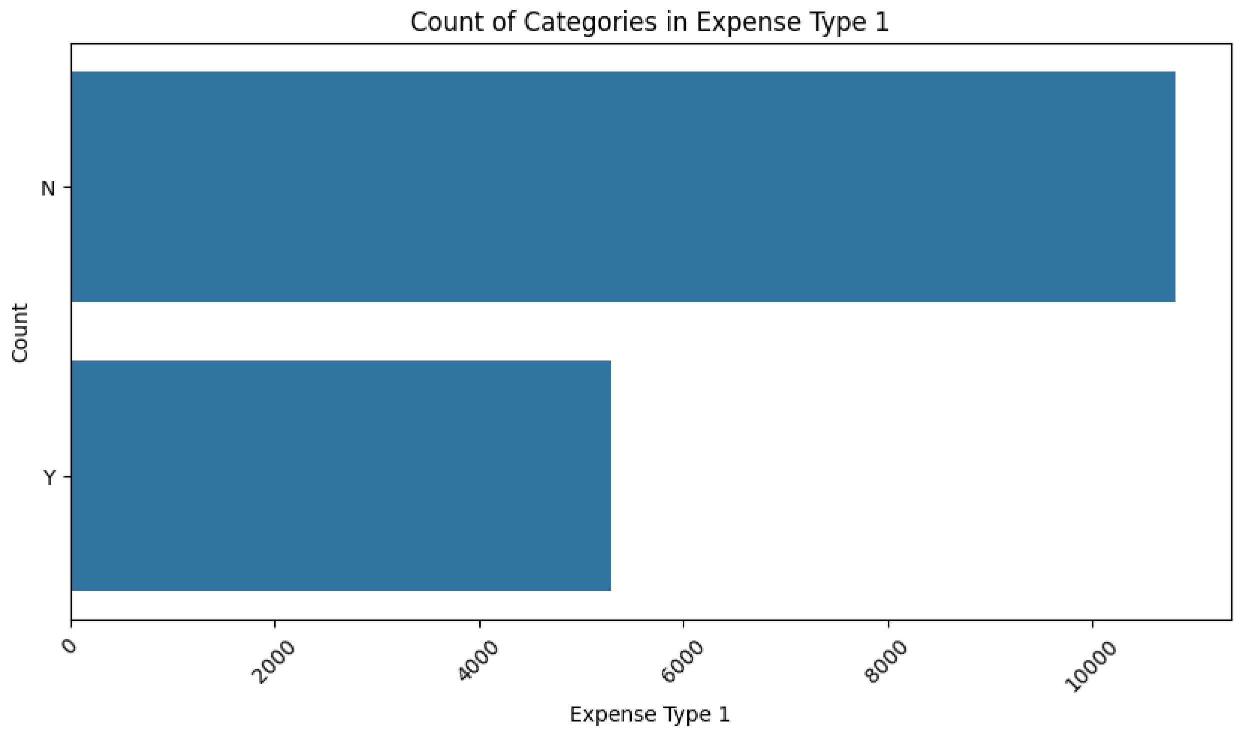

```
Statistics for Type of Employment:
Type of Employment
Laborers                 6708
Sales staff              2288
Core staff               1825
Managers                  995
Drivers                   828
Accountants               728
High skill tech staff     697
Medicine staff            530
Security staff            357
Cooking staff             349
Cleaning staff            210
Private service staff     194
Low-skill Laborers        102
Secretaries                92
Waiters/barmen staff       92
Realty agents              44
IT staff                   41
HR staff                   37
Name: count, dtype: int64
Mode: Laborers
```



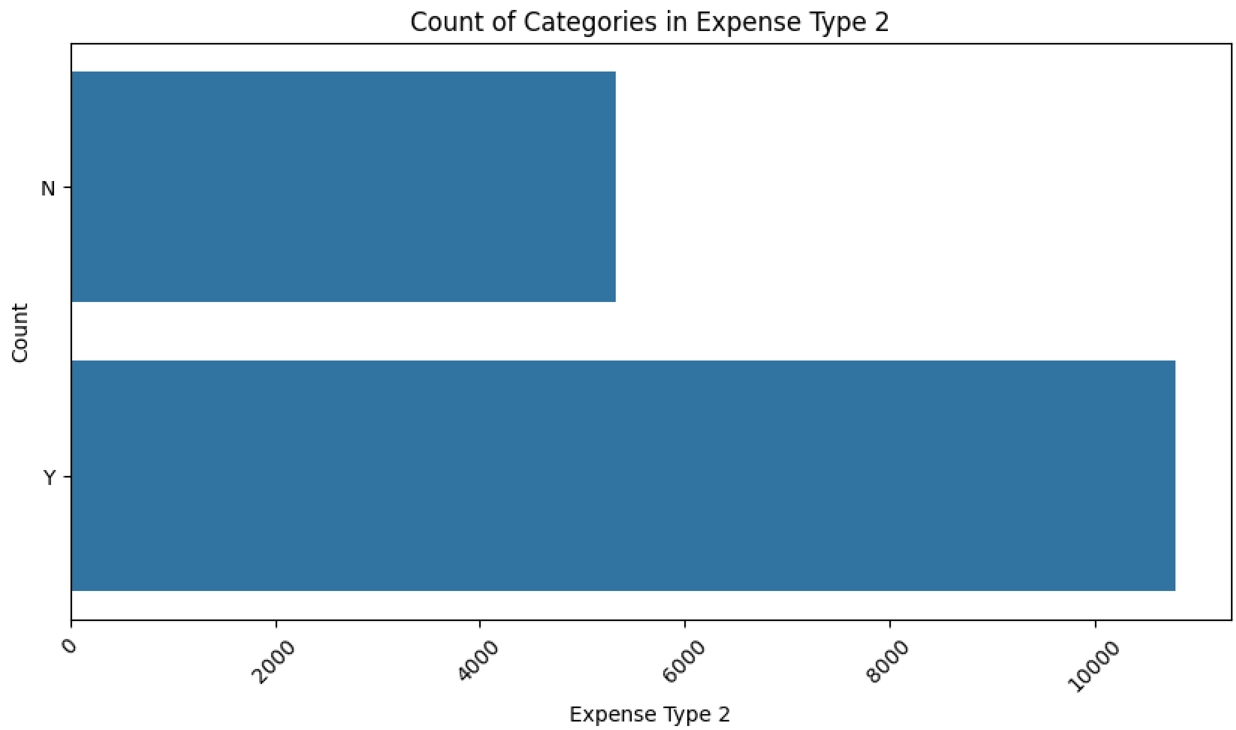

```
Statistics for Location:
Location
Semi-Urban    12009
Rural          2923
Urban          1185
Name: count, dtype: int64
Mode: Semi-Urban
```



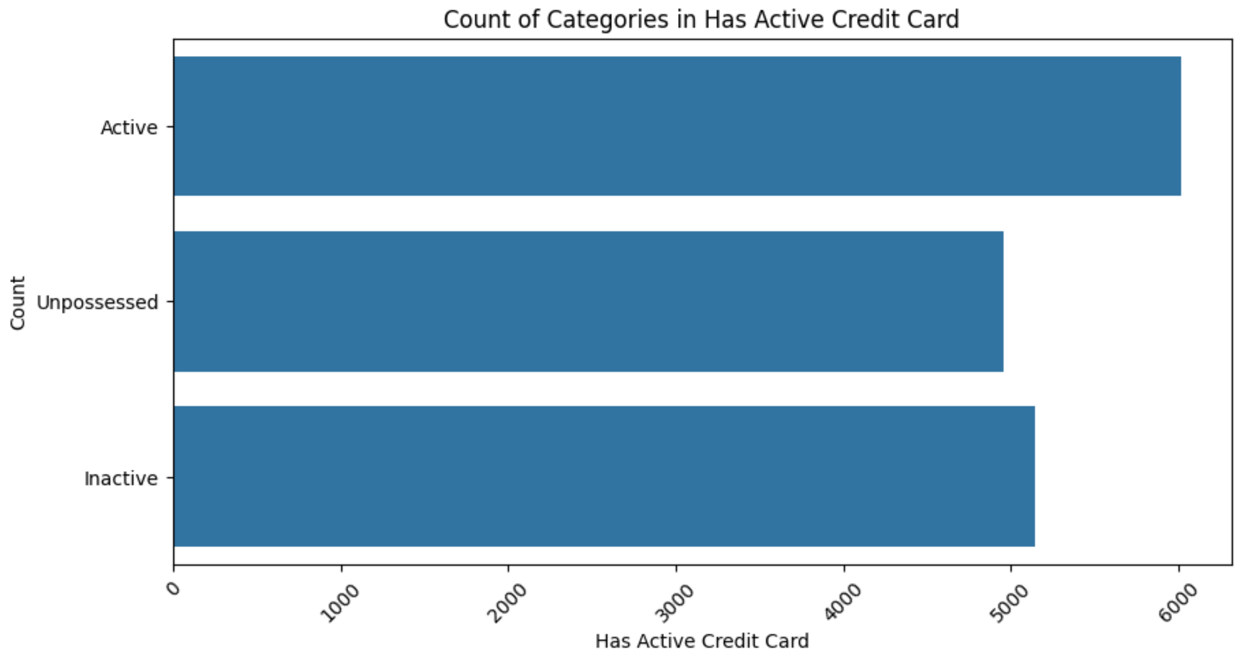
```
Statistics for Expense Type 1:
Expense Type 1
N    10816
Y     5301
Name: count, dtype: int64
Mode: N
```




```
Statistics for Expense Type 2:
Expense Type 2
Y    10787
N     5330
Name: count, dtype: int64
Mode: Y
```




```
Statistics for Has Active Credit Card:
Has Active Credit Card
Active         6015
Inactive       5145
Unpossessed    4957
Name: count, dtype: int64
Mode: Active
```




```
Statistics for Property Location:
Property Location
Semi-Urban    5790
Rural         5384
Urban         4943
Name: count, dtype: int64
Mode: Semi-Urban
```