

loan-sanction-prediction

October 11, 2023

1 Creating a model for Loan Sanction ammount Prediction

data from <https://www.kaggle.com/datasets/boss0ayush/loan-sanction-amount-prediction-data/>

Importing all the Library Needed

```
[208]: import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.model_selection import GridSearchCV, train_test_split
from sklearn.tree import DecisionTreeRegressor
from sklearn.metrics import mean_absolute_error, \
    mean_absolute_percentage_error, mean_squared_error, r2_score
from sklearn.preprocessing import LabelEncoder
from scipy import stats
import math

import gdown
```

Loading the data to colab

you can change this block of code depending on how your loading your data, I'am using Google drive

```
[209]: def load_data(link):
    # Extract file ID from the provided link
    file_id = link.split('/')[2]

    # Generate direct download link
    csv_link = f'https://drive.google.com/uc?id={file_id}'

    # Download the CSV file
    gdown.download(csv_link, f'{file_id}.csv', quiet=False)

    # Read the CSV file
    df = pd.read_csv(f'{file_id}.csv', encoding='utf-8')

    return df
```

```
# Use the provided links
test_link = 'https://drive.google.com/file/d/1WVW4P32kTIB6RNbufikoBzdT5K7DkEo_/
↳view?usp=sharing'

train_link = 'https://drive.google.com/file/d/1WSui92w1LQ-ZLybpXoQ8_wpMV0tPe90E/
↳view?usp=sharing'

# Load the data
test = load_data(test_link)

train = load_data(train_link)
```

Downloading...

From: https://drive.google.com/uc?id=1WVW4P32kTIB6RNbufikoBzdT5K7DkEo_

To: /content/1WVW4P32kTIB6RNbufikoBzdT5K7DkEo_.csv

100%| | 2.87M/2.87M [00:00<00:00, 232MB/s]

Downloading...

From: https://drive.google.com/uc?id=1WSui92w1LQ-ZLybpXoQ8_wpMV0tPe90E

To: /content/1WSui92w1LQ-ZLybpXoQ8_wpMV0tPe90E.csv

100%| | 4.48M/4.48M [00:00<00:00, 232MB/s]

[210]: train.head()

```
[210]:   Customer ID      Name Gender  Age  Income (USD) Income Stability \
0      C-36995  Frederica Shealy    F   56      1933.05             Low
1      C-33999  America Calderone    M   32      4952.91             Low
2        C-3770    Rosetta Verne    F   65        988.19             High
3      C-26480      Zoe Chitty    F   65           NaN             High
4      C-23459    Afton Venema    F   31      2614.77             Low
```

```
   Profession      Type of Employment      Location  Loan Amount Request (USD) \
0    Working              Sales staff  Semi-Urban             72809.58
1    Working                NaN  Semi-Urban             46837.47
2  Pensioner                NaN  Semi-Urban             45593.04
3  Pensioner                NaN    Rural             80057.92
4    Working  High skill tech staff  Semi-Urban             113858.89
```

```
...  Credit Score  No. of Defaults  Has Active Credit Card  Property ID \
0  ...      809.44                0                NaN             746
1  ...      780.40                0            Unpossessed             608
2  ...      833.15                0            Unpossessed             546
3  ...      832.70                1            Unpossessed             890
4  ...      745.55                1              Active             715
```

```
   Property Age  Property Type  Property Location  Co-Applicant \
0      1933.05                4              Rural                1
```

1	4952.91	2	Rural	1
2	988.19	2	Urban	0
3	NaN	2	Semi-Urban	1
4	2614.77	4	Semi-Urban	1

	Property Price	Loan Sanction Amount (USD)
0	119933.46	54607.18
1	54791.00	37469.98
2	72440.58	36474.43
3	121441.51	56040.54
4	208567.91	74008.28

[5 rows x 24 columns]

2 Data Preprocessing Part 1

```
[211]: #Check the number of unique value from all of the object datatype
train.select_dtypes(include='object').nunique()
```

```
[211]: Customer ID          30000
Name          30000
Gender         2
Income Stability  2
Profession     8
Type of Employment 18
Location        3
Expense Type 1   2
Expense Type 2   2
Has Active Credit Card  3
Property Location  3
dtype: int64
```

```
[212]: # Check the amount of missing value
check_missing = train.isnull().sum() * 100 / train.shape[0]
check_missing[check_missing > 0].sort_values(ascending=False)
```

```
[212]: Type of Employment      24.233333
Property Age      16.166667
Income (USD)      15.253333
Dependents        8.310000
Credit Score      5.676667
Income Stability   5.610000
Has Active Credit Card  5.220000
Property Location   1.186667
Loan Sanction Amount (USD)  1.133333
Current Loan Expenses (USD)  0.573333
```

Gender 0.176667
dtype: float64

```
[213]: def preprocess_data(df):  
        # Remove identifier columns  
        df.drop(columns=['Customer ID', 'Name'], inplace=True)  
  
        # Segment Type of Employment into smaller unique values  
        df['Type of Employment'] = df['Type of Employment'].  
        ↪ apply(segment_employment_type)  
  
        # Handle missing values  
        df.fillna({  
            'Property Age': df['Property Age'].median(),  
            'Income (USD)': df['Income (USD)'].median(),  
            'Dependents': df['Dependents'].median(),  
            'Credit Score': df['Credit Score'].median(),  
            'Loan Sanction Amount (USD)': df['Loan Sanction Amount (USD)'].median(),  
            'Current Loan Expenses (USD)': df['Current Loan Expenses (USD)'].  
        ↪ median()  
        }, inplace=True)  
  
        df.dropna(subset=['Income Stability', 'Has Active Credit Card', 'Property_  
        ↪ Location', 'Gender'], inplace=True)  
  
        return df.head(10)
```

3 Segment Type of Employment into smaller unique value

```
[214]: train['Type of Employment'].unique()
```

```
[214]: array(['Sales staff', nan, 'High skill tech staff', 'Secretaries',  
            'Laborers', 'Managers', 'Cooking staff', 'Core staff', 'Drivers',  
            'Realty agents', 'Security staff', 'Accountants',  
            'Private service staff', 'Waiters/barmen staff', 'Medicine staff',  
            'Cleaning staff', 'Low-skill Laborers', 'HR staff', 'IT staff'],  
          dtype=object)
```

```
[215]: def segment_employment_type(value):  
        if pd.isna(value):  
            return 'Unknown'  
        elif 'Sales' in value or 'Realty' in value:  
            return 'Sales/Realty'  
        elif 'Tech' in value or 'IT' in value:
```

```

    return 'Tech/IT'
elif 'Secretaries' in value or 'HR' in value:
    return 'Secretaries/HR'
elif 'Laborers' in value or 'Low-skill Laborers' in value:
    return 'Laborers'
elif 'Managers' in value:
    return 'Managers'
elif 'Cooking' in value or 'Waiters/barmen' in value:
    return 'Hospitality'
else:
    return 'Other'

```

```

[216]: # Apply the function to create a new column
train['Type of Employment'] = train['Type of Employment'].
    ↪ apply(segment_employment_type)

```

```

[217]: preprocess_data(train)

```

```

[217]:
  Gender  Age  Income (USD)  Income Stability  Profession \
1      M   32    4952.910             Low      Working
2      F   65     988.190             High    Pensioner
3      F   65    2222.435             High    Pensioner
4      F   31    2614.770             Low      Working
5      F   60    1234.920             Low  State servant
6      M   43    2361.560             Low      Working
7      F   45    2222.435             Low  State servant
8      F   38    1296.070             Low      Working
9      M   18    1546.170             Low      Working
10     M   18    2416.860             Low  State servant

  Type of Employment  Location  Loan Amount Request (USD) \
1                Other  Semi-Urban          46837.47
2                Other  Semi-Urban          45593.04
3                Other    Rural          80057.92
4                Other  Semi-Urban         113858.89
5    Secretaries/HR    Rural          34434.72
6            Laborers  Semi-Urban         152561.34
7            Managers  Semi-Urban         240311.77
8                Other    Rural          35141.99
9            Laborers    Rural          42091.29
10             Other  Semi-Urban          25765.72

  Current Loan Expenses (USD)  Expense Type 1  ...  Credit Score \
1                495.81                N  ...        780.40
2                171.95                N  ...        833.15
3                298.54                N  ...        832.70
4                491.41                N  ...        745.55

```

5	181.48	N ...	684.12
6	697.67	Y ...	637.29
7	807.64	N ...	812.26
8	155.95	N ...	705.29
9	500.20	N ...	613.24
10	140.02	N ...	652.41

	No. of Defaults	Has Active Credit Card	Property ID	Property Age \
1	0	Unpossessed	608	4952.91
2	0	Unpossessed	546	988.19
3	1	Unpossessed	890	2223.25
4	1	Active	715	2614.77
5	1	Inactive	491	1234.92
6	0	Unpossessed	227	2361.56
7	0	Active	314	2223.25
8	1	Active	241	1296.07
9	0	Unpossessed	883	1546.17
10	0	Active	325	2416.86

	Property Type	Property Location	Co-Applicant	Property Price \
1	2	Rural	1	54791.00
2	2	Urban	0	72440.58
3	2	Semi-Urban	1	121441.51
4	4	Semi-Urban	1	208567.91
5	2	Rural	1	43146.82
6	1	Semi-Urban	1	221050.80
7	2	Urban	1	401040.70
8	4	Rural	1	54903.44
9	2	Urban	1	67993.43
10	2	Rural	1	32423.71

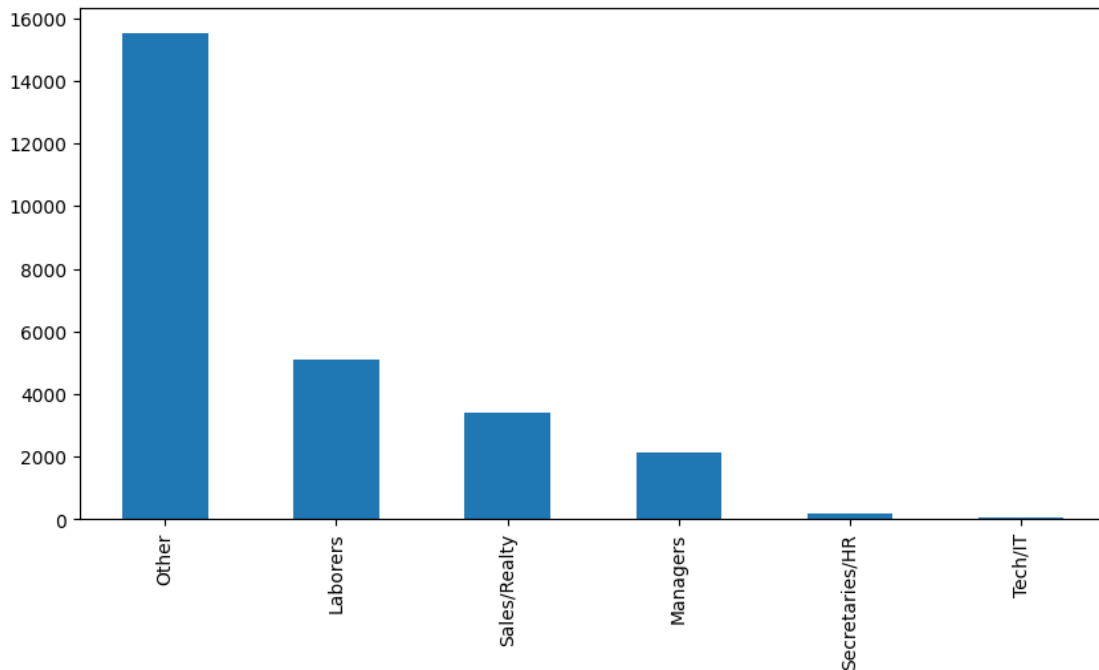
	Loan Sanction Amount (USD)
1	37469.98
2	36474.43
3	56040.54
4	74008.28
5	22382.57
6	0.00
7	168218.24
8	22842.29
9	0.00
10	16747.72

[10 rows x 22 columns]

```
[218]: # We visualize the column Type of Employment how many times it repeates
plt.figure(figsize=(10,5))
```

```
train['Type of Employment'].value_counts().plot(kind='bar')
```

[218]: <Axes: >



```
[219]: # Loop over each column in the DataFrame where dtype is 'object'
for col in train.select_dtypes(include=['object']).columns:

    # Print the column name and the unique values
    print(f"{col}: {train[col].unique()}")
```

Gender: ['M' 'F']

Income Stability: ['Low' 'High']

Profession: ['Working' 'Pensioner' 'State servant' 'Commercial associate' 'Unemployed']

Type of Employment: ['Other' 'Secretaries/HR' 'Laborers' 'Managers' 'Sales/Realty' 'Tech/IT']

Location: ['Semi-Urban' 'Rural' 'Urban']

Expense Type 1: ['N' 'Y']

Expense Type 2: ['Y' 'N']

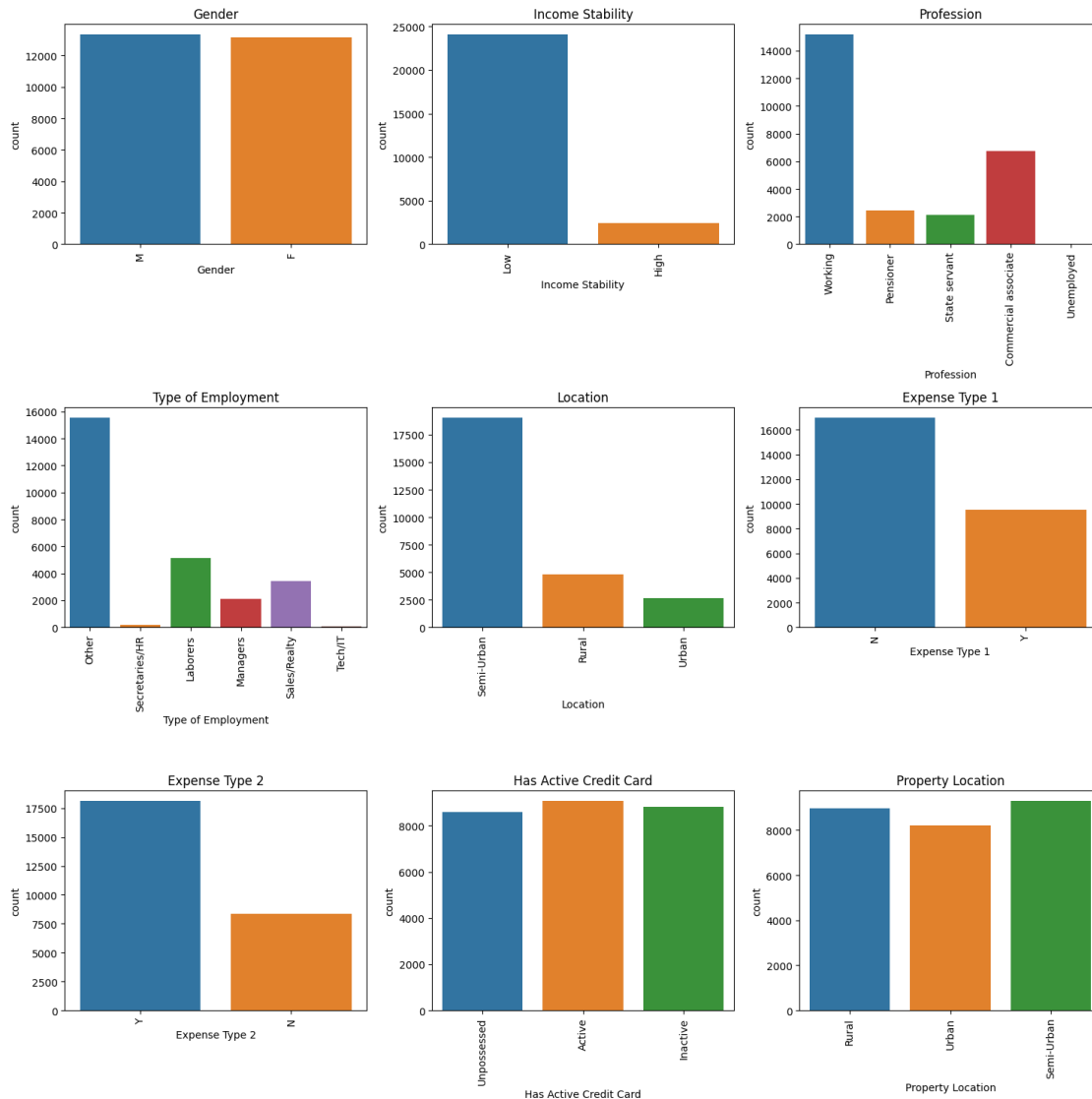
Has Active Credit Card: ['Unpossessed' 'Active' 'Inactive']

Property Location: ['Rural' 'Urban' 'Semi-Urban']

4 Exploratory Data Analysis

```
[220]: def explore_data(df):  
    # Get the names of all columns with data type 'object' (categorical columns)  
    cat_vars = df.select_dtypes(include='object').columns.tolist()  
  
    # Create a figure with subplots  
    num_cols = len(cat_vars)  
    num_rows = (num_cols + 2) // 3  
    fig, axs = plt.subplots(nrows=num_rows, ncols=3, figsize=(15, 5*num_rows))  
    axs = axs.flatten()  
  
    # Create a countplot for the top 6 values of each categorical variable using  
    ↪ Seaborn  
    for i, var in enumerate(cat_vars):  
        top_values = df[var].value_counts().nlargest(6).index  
        filtered_tf = df[df[var].isin(top_values)]  
        sns.countplot(x=var, data=filtered_tf, ax=axs[i])  
        axs[i].set_title(var)  
        axs[i].tick_params(axis='x', rotation=90)  
  
    # Remove any extra empty subplots if needed  
    if num_cols < len(axs):  
        for i in range(num_cols, len(axs)):  
            fig.delaxes(axs[i])  
  
    # Adjust spacing between subplots  
    fig.tight_layout()  
  
    # Show plot  
    plt.show()
```

```
[221]: explore_data(train)
```

```
[222]: def eplore_2(df):
    # Get the names of all columns with data type 'int' or 'float' except 'cltv'
    # and 'marital_status'
    num_vars = df.select_dtypes(include=['int', 'float']).columns.tolist()
    exclude_vars = ['Loan Sanction Amount (USD)']
    num_vars = [var for var in num_vars if var not in exclude_vars]

    # Create a figure with subplots
    num_cols = len(num_vars)
    num_rows = (num_cols + 2) // 3
    fig, axs = plt.subplots(nrows=num_rows, ncols=3, figsize=(15, 5*num_rows))
    axs = axs.flatten()
```

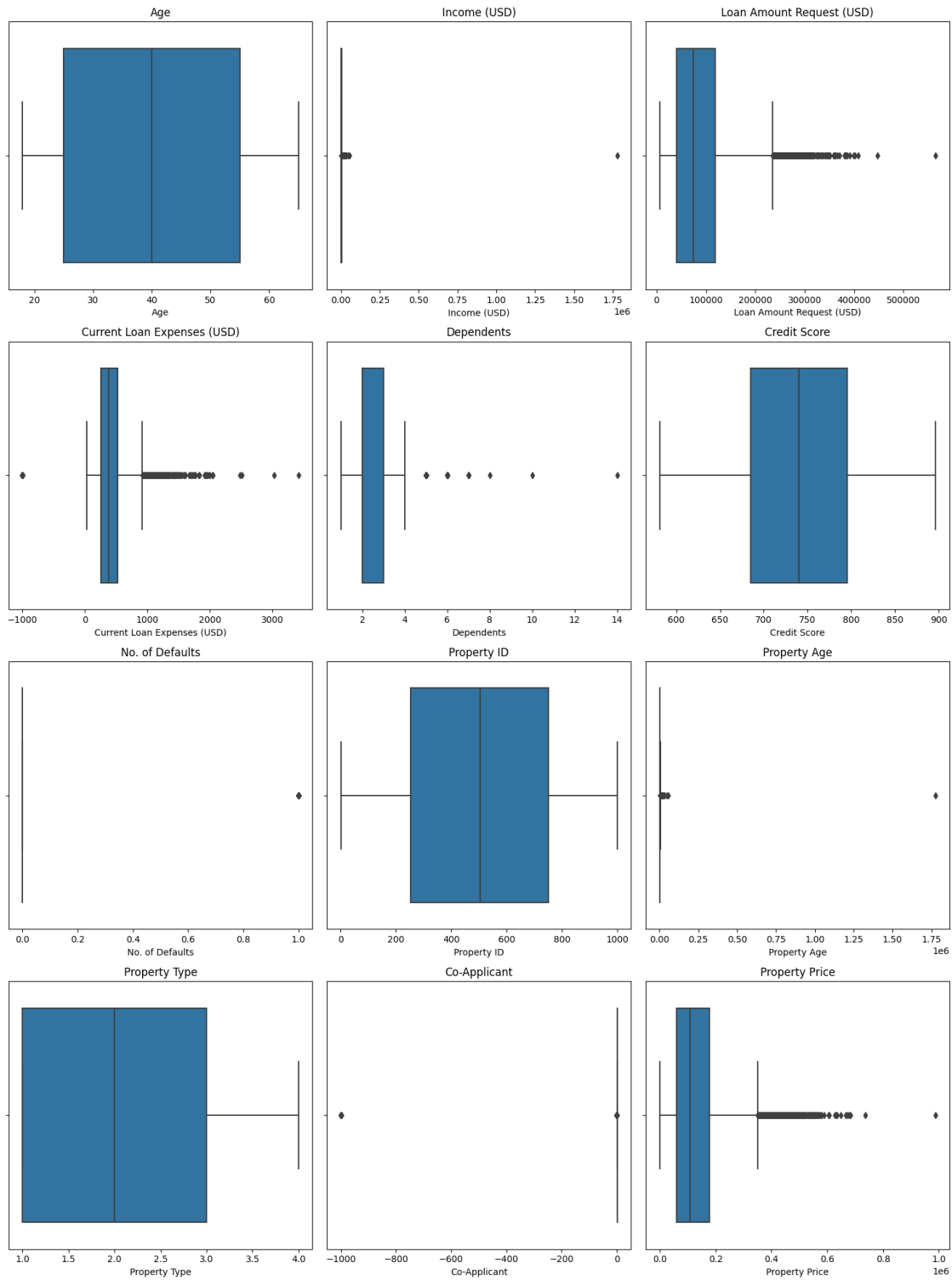
```
# Create a box plot for each numerical variable using Seaborn
for i, var in enumerate(num_vars):
    sns.boxplot(x=df[var], ax=axes[i])
    axes[i].set_title(var)

# Remove any extra empty subplots if needed
if num_cols < len(axes):
    for i in range(num_cols, len(axes)):
        fig.delaxes(axes[i])

# Adjust spacing between subplots
fig.tight_layout()

# Show plot
plt.show()
```

```
[223]: explore_2(train)
```



```
[224]: def exprole_3(df):
```

```

# Get the names of all columns with data type 'int' or 'float' except
↳ 'marital_status' and 'cltv'
int_vars = df.select_dtypes(include=['int', 'float']).columns.tolist()
exclude_vars = ['Loan Sanction Amount (USD)']
int_vars = [var for var in int_vars if var not in exclude_vars]

# Create a figure with subplots
num_cols = len(int_vars)
num_rows = (num_cols + 2) // 3 # To make sure there are enough rows for the
↳ subplots
fig, axs = plt.subplots(nrows=num_rows, ncols=3, figsize=(15, 5*num_rows))
axs = axs.flatten()

# Create a histogram for each integer variable
for i, var in enumerate(int_vars):
    df[var].plot.hist(ax=axs[i], bins=20) # You can adjust the number of
↳ bins as needed
    axs[i].set_title(var)

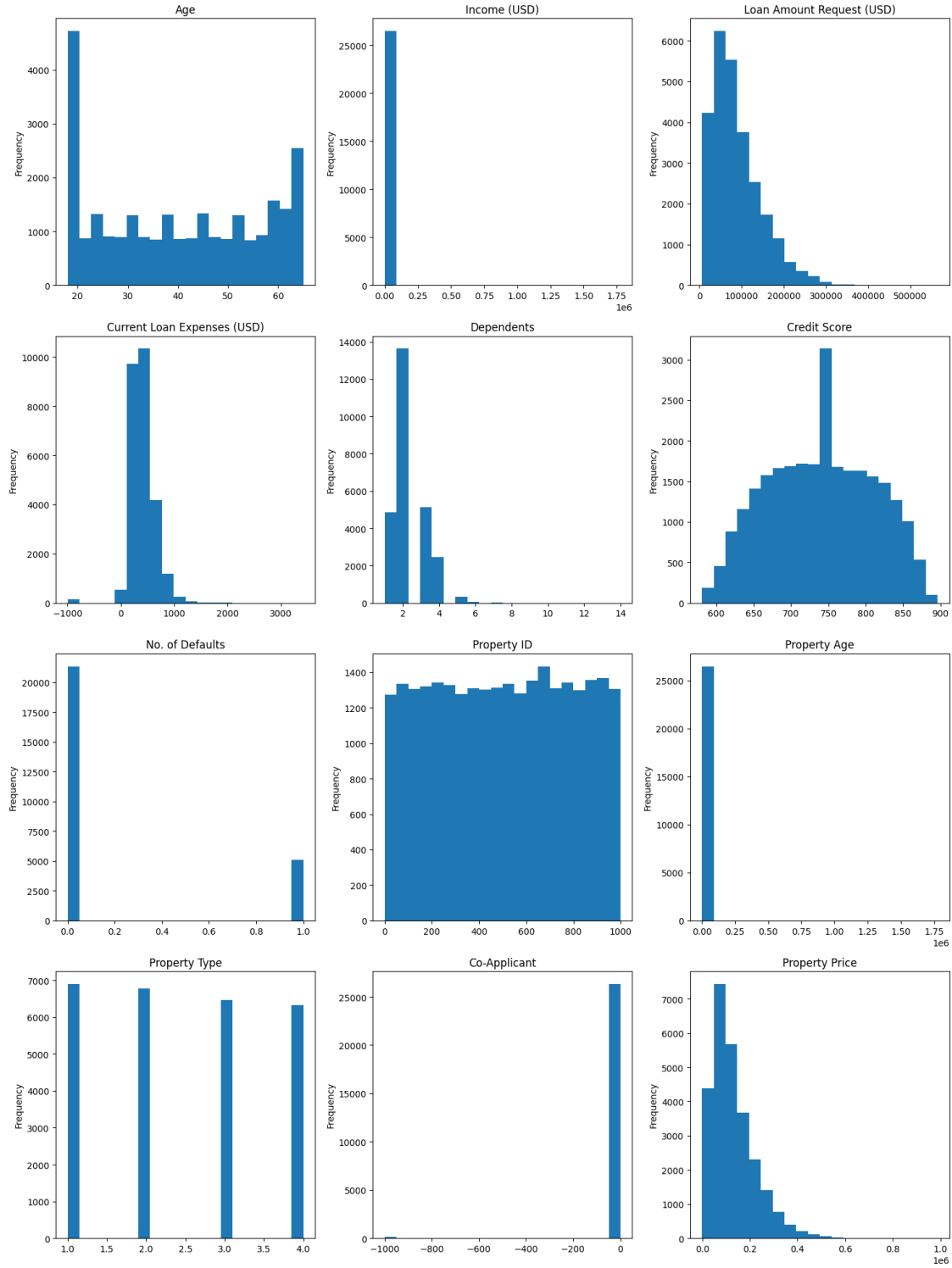
# Remove any extra empty subplots if needed
if num_cols < len(axs):
    for i in range(num_cols, len(axs)):
        fig.delaxes(axs[i])

# Adjust spacing between subplots
fig.tight_layout()

# Show plot
plt.show()

```

```
[225]: exprole_3(train)
```



```
[226]: def exprole_4(df):
        # Specify the maximum number of categories to show individually
```

```

max_categories = 5

# Filter categorical columns with 'object' data type
cat_cols = [col for col in df.columns if col != 'y' and df[col].dtype ==
↳ 'object']

# Create a figure with subplots
num_cols = len(cat_cols)
num_rows = (num_cols + 2) // 3
fig, axs = plt.subplots(nrows=num_rows, ncols=3, figsize=(20, 5*num_rows))

# Flatten the axs array for easier indexing
axs = axs.flatten()

# Create a pie chart for each categorical column
for i, col in enumerate(cat_cols):
    if i < len(axs): # Ensure we don't exceed the number of subplots
        # Count the number of occurrences for each category
        cat_counts = df[col].value_counts()

        # Group categories beyond the top max_categories as 'Other'
        if len(cat_counts) > max_categories:
            cat_counts_top = cat_counts[:max_categories]
            cat_counts_other = pd.Series(cat_counts[max_categories:].sum(),
↳ index=['Other'])
            cat_counts = cat_counts_top.append(cat_counts_other)

        # Create a pie chart
        axs[i].pie(cat_counts, labels=cat_counts.index, autopct='%1.1f%%',
↳ startangle=90)
        axs[i].set_title(f'{col} Distribution')

# Remove any extra empty subplots if needed
if num_cols < len(axs):
    for i in range(num_cols, len(axs)):
        fig.delaxes(axs[i])

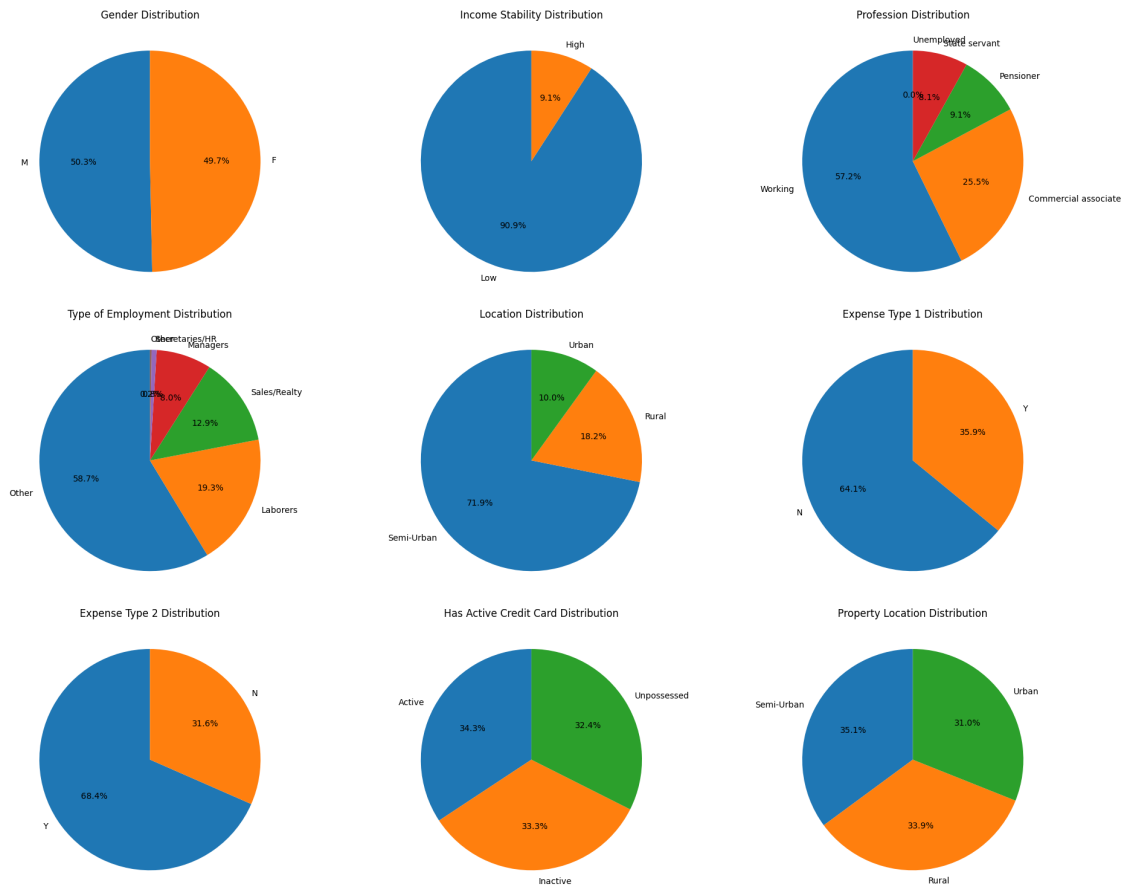
# Adjust spacing between subplots
fig.tight_layout()

# Show plot
plt.show()

```

```
[227]: exprole_4(train)
```

The `series.append` method is deprecated and will be removed from pandas in a future version. Use `pandas.concat` instead.



5 Data Preprocessing Part 2

```
[228]: # Check the amount of missing value
check_missing = train.isnull().sum() * 100 / train.shape[0]
check_missing[check_missing > 0].sort_values(ascending=False)
```

```
[228]: Series([], dtype: float64)
```

6 Label Encoding for Object Datatypes Encode categorical columns

```
[229]: ## Label Encoding for Object Datatypes Encode categorical columns
def encoder_object(df):
    # Loop over each column in the DataFrame where dtype is 'object'
    for col in df.select_dtypes(include=['object']).columns:
        label_encoder = LabelEncoder()
```

```
# Fit the encoder to the unique values in the column
label_encoder.fit(df[col].unique())

# Transform the column using the encoder
df[col] = label_encoder.transform(df[col])

# Print the column name and the unique encoded values
print(f"{col}: {df[col].unique()}")
```

```
[230]: # Print the column name and the unique encoded values
encoder_object(train)
```

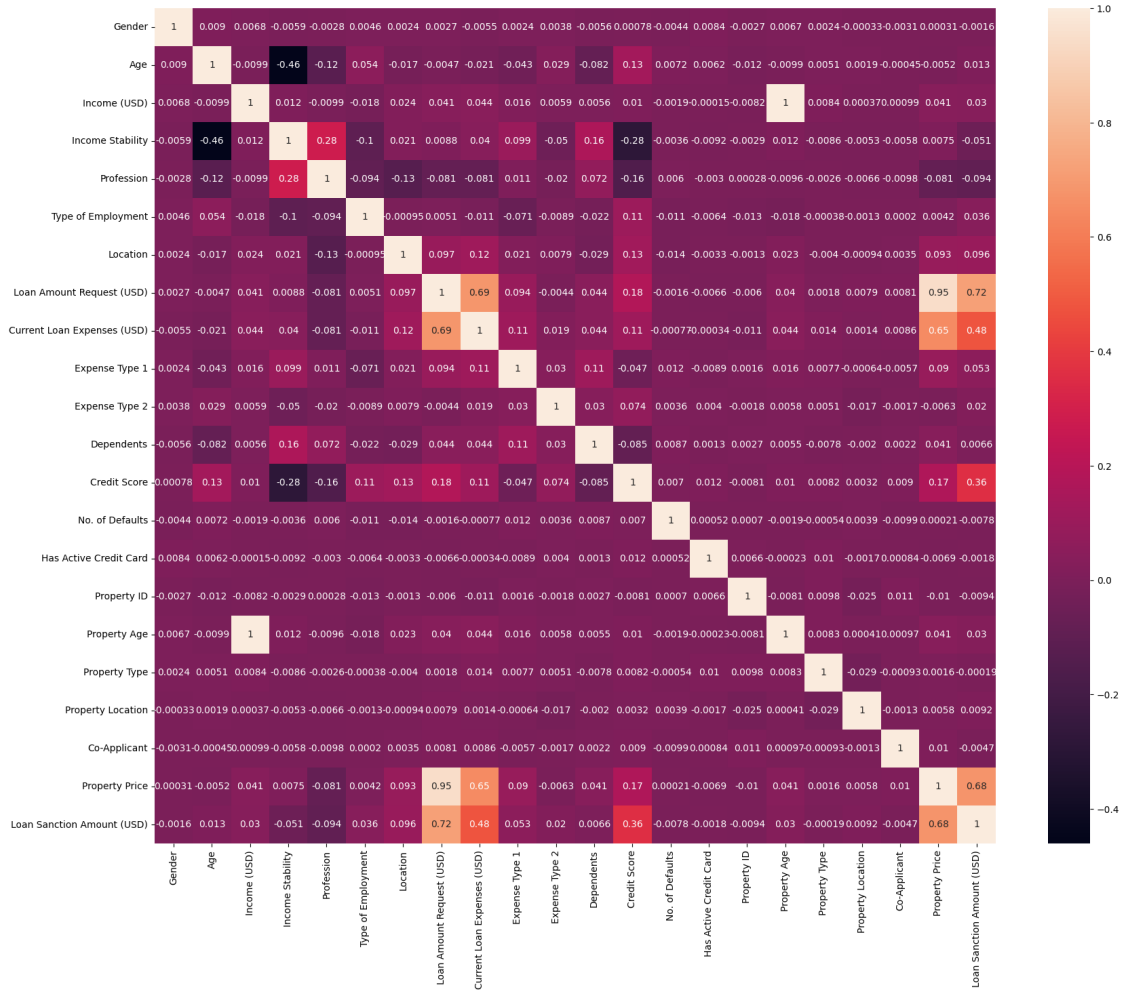
```
Gender: [1 0]
Income Stability: [1 0]
Profession: [4 1 2 0 3]
Type of Employment: [2 4 0 1 3 5]
Location: [1 0 2]
Expense Type 1: [0 1]
Expense Type 2: [1 0]
Has Active Credit Card: [2 0 1]
Property Location: [0 2 1]
```

```
[231]: train.shape
```

```
[231]: (26474, 22)
```

```
[232]: # Correlation Heatmap
plt.figure(figsize=(20, 16))
sns.heatmap(train.corr(), fmt='.2g', annot=True)
```

```
[232]: <Axes: >
```

7 Remove Outlier from Train Data using Z-Score

```
[233]: def remove_outliers(df, columns_to_remove_outliers):
# Remove outliers using Z-score
z_scores = np.abs(stats.zscore(df[columns_to_remove_outliers]))
threshold = 3
outlier_indices = np.where(z_scores > threshold)[0]
df = df.drop(df.index[outlier_indices])
return df.head(10)
```

```
[234]: selected_columns = ['Income (USD)', 'Loan Amount Request (USD)', 'Current Loan_
↳Expenses (USD)',
                           'Dependents', 'Property Age', 'Co-Applicant', 'Property_
↳Price']
remove_outliers(train, selected_columns)
```

[234]:

	Gender	Age	Income (USD)	Income Stability	Profession \
1	1	32	4952.910	1	4
2	0	65	988.190	0	1
3	0	65	2222.435	0	1
4	0	31	2614.770	1	4
5	0	60	1234.920	1	2
6	1	43	2361.560	1	4
7	0	45	2222.435	1	2
8	0	38	1296.070	1	4
9	1	18	1546.170	1	4
10	1	18	2416.860	1	2

	Type of Employment	Location	Loan Amount Request (USD) \
1		2	1 46837.47
2		2	1 45593.04
3		2	0 80057.92
4		2	1 113858.89
5		4	0 34434.72
6		0	1 152561.34
7		1	1 240311.77
8		2	0 35141.99
9		0	0 42091.29
10		2	1 25765.72

	Current Loan Expenses (USD)	Expense Type 1 ...	Credit Score \
1	495.81	0 ...	780.40
2	171.95	0 ...	833.15
3	298.54	0 ...	832.70
4	491.41	0 ...	745.55
5	181.48	0 ...	684.12
6	697.67	1 ...	637.29
7	807.64	0 ...	812.26
8	155.95	0 ...	705.29
9	500.20	0 ...	613.24
10	140.02	0 ...	652.41

	No. of Defaults	Has Active Credit Card	Property ID	Property Age \
1	0	2	608	4952.91
2	0	2	546	988.19
3	1	2	890	2223.25
4	1	0	715	2614.77
5	1	1	491	1234.92
6	0	2	227	2361.56
7	0	0	314	2223.25
8	1	0	241	1296.07
9	0	2	883	1546.17
10	0	0	325	2416.86

	Property Type	Property Location	Co-Applicant	Property Price \
1	2	0	1	54791.00
2	2	2	0	72440.58
3	2	1	1	121441.51
4	4	1	1	208567.91
5	2	0	1	43146.82
6	1	1	1	221050.80
7	2	2	1	401040.70
8	4	0	1	54903.44
9	2	2	1	67993.43
10	2	0	1	32423.71

	Loan Sanction Amount (USD)
1	37469.98
2	36474.43
3	56040.54
4	74008.28
5	22382.57
6	0.00
7	168218.24
8	22842.29
9	0.00
10	16747.72

[10 rows x 22 columns]

8 Train Test Split

9 Build the decision Tree

```
[235]: def train_decision_tree(df):
        X = df.drop('Loan Sanction Amount (USD)', axis=1)
        y = df['Loan Sanction Amount (USD)']

        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
        random_state=0)

        # Create a DecisionTreeRegressor object
        dtree = DecisionTreeRegressor()

        # Define the hyperparameters to tune and their values
        param_grid = {
            'max_depth': [2, 4, 6, 8],
            'min_samples_split': [2, 4, 6, 8],
            'min_samples_leaf': [1, 2, 3, 4],
```


[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

keep the past behaviour, explicitly set `max_features=1.0`.

`max_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0`.

`max_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0`.

`max_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0`.

`max_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0`.

`max_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0`.

`max_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0`.

{'max_depth': 8, 'max_features': 'auto', 'min_samples_leaf': 3, 'min_samples_split': 8, 'random_state': 0}

10 Evaluate the decision tree

```
[237]: def evaluate_decision_tree(model, X_test, y_test):
    y_pred = model.predict(X_test)
    mae = mean_absolute_error(y_test, y_pred)
    mape = mean_absolute_percentage_error(y_test, y_pred)
    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)
    rmse = math.sqrt(mse)

    print('MAE is {}'.format(mae))
    print('MAPE is {}'.format(mape))
    print('MSE is {}'.format(mse))
    print('R2 score is {}'.format(r2))
    print('RMSE score is {}'.format(rmse))
```

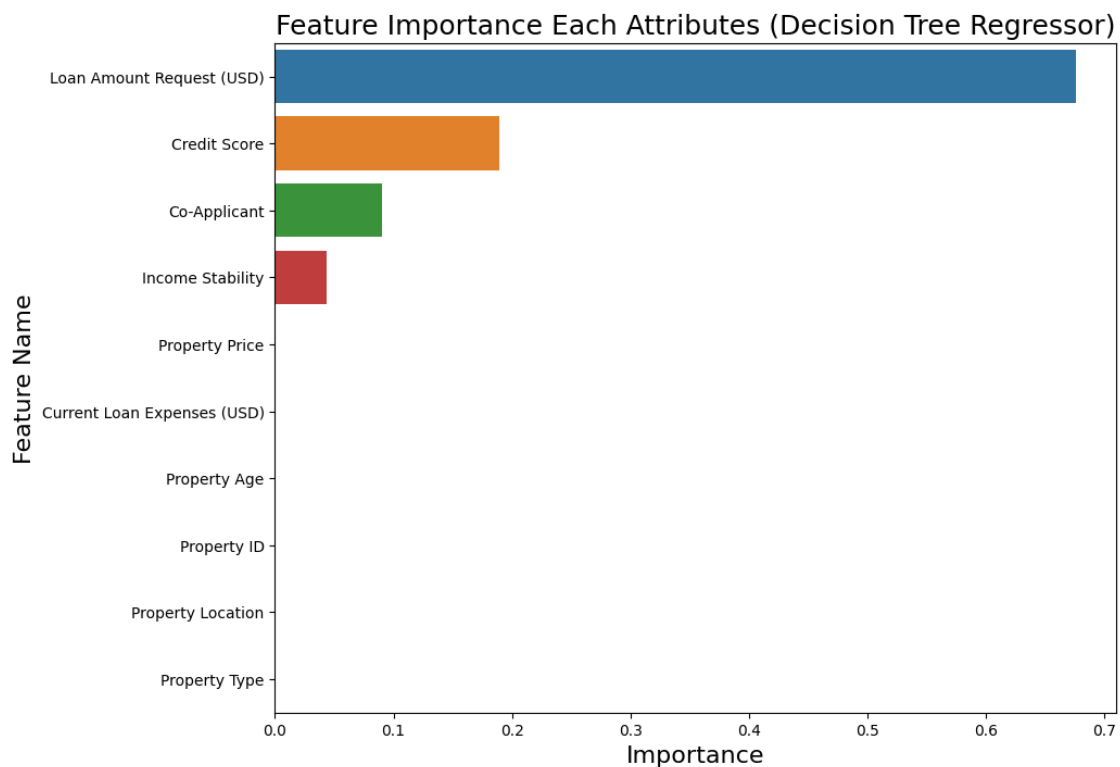
```
[238]: print("Decision Tree Model:")
    evaluate_decision_tree(dtrees, X_test, y_test)
```

```
Decision Tree Model:
MAE is 12651.787227844097
MAPE is 2.3472735407383396e+19
MSE is 591410680.1464931
R2 score is 0.7426083343470093
RMSE score is 24318.936657397113
```

10.1 Plot The Feature Importance

```
[239]: imp_df = pd.DataFrame({
    "Feature Name": X_train.columns,
    "Importance": dtree.feature_importances_
})
fi = imp_df.sort_values(by="Importance", ascending=False)

fi2 = fi.head(10)
plt.figure(figsize=(10,8))
sns.barplot(data=fi2, x='Importance', y='Feature Name')
plt.title('Feature Importance Each Attributes (Decision Tree Regressor)',
          ↪fontsize=18)
plt.xlabel ('Importance', fontsize=16)
plt.ylabel ('Feature Name', fontsize=16)
plt.show()
```



```
[240]: !pip install shap
```

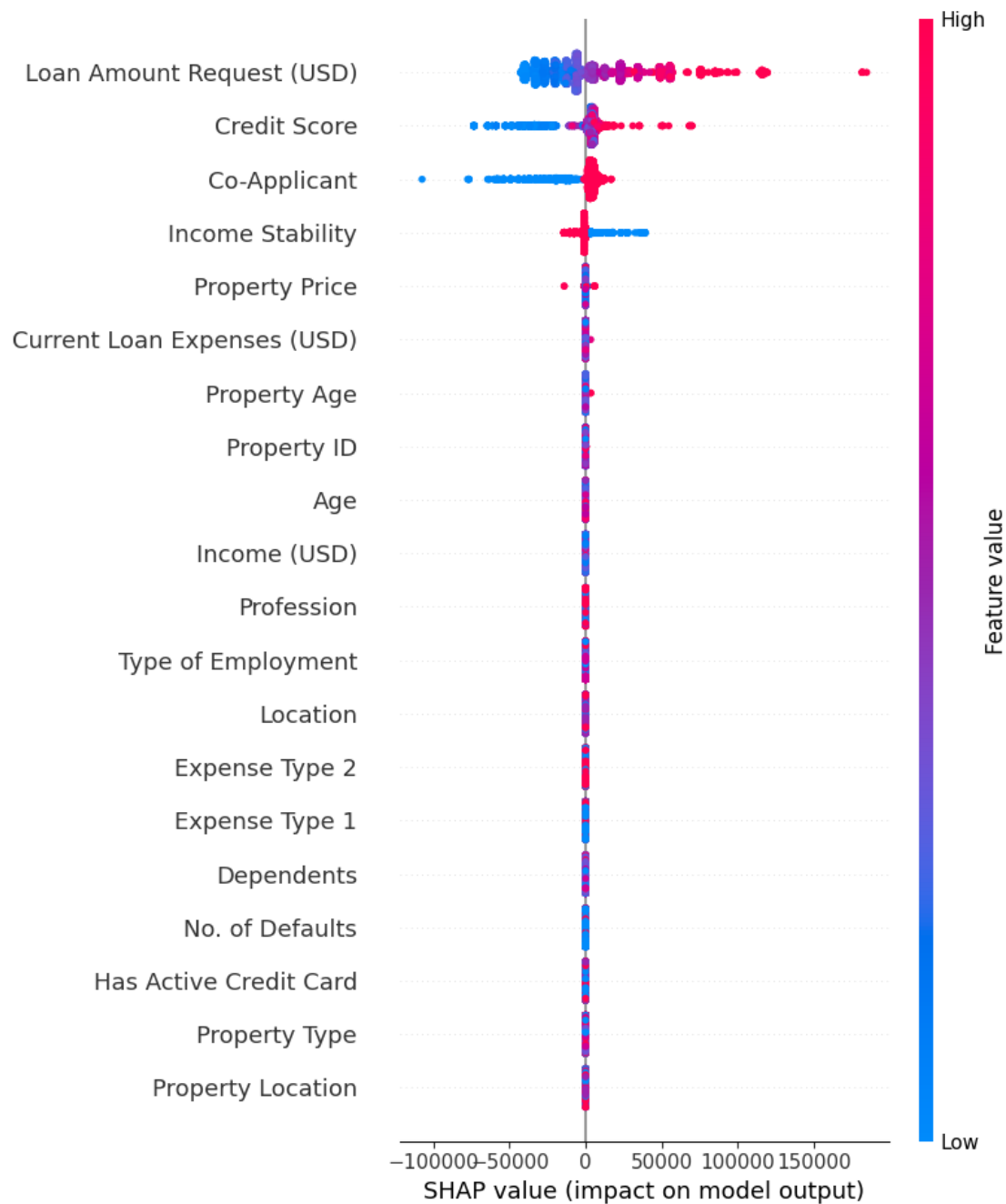
Requirement already satisfied: shap in /usr/local/lib/python3.10/dist-packages (0.42.1)

Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (from shap) (1.23.5)

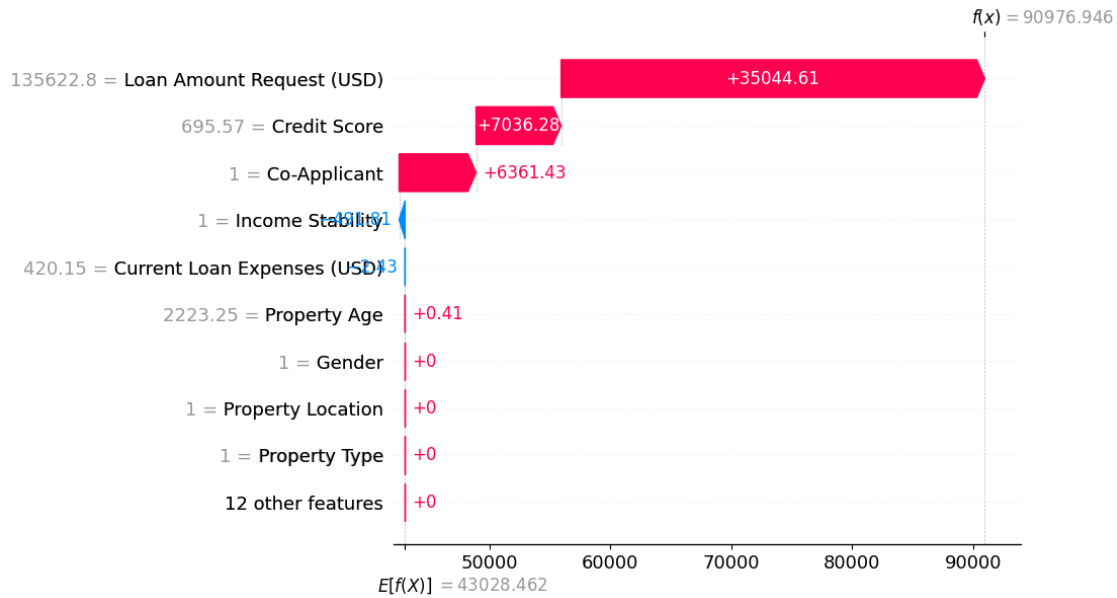
Requirement already satisfied: scipy in /usr/local/lib/python3.10/dist-packages (from shap) (1.11.3)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.10/dist-packages (from shap) (1.2.2)
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (from shap) (1.5.3)
Requirement already satisfied: tqdm>=4.27.0 in /usr/local/lib/python3.10/dist-packages (from shap) (4.66.1)
Requirement already satisfied: packaging>20.9 in /usr/local/lib/python3.10/dist-packages (from shap) (23.2)
Requirement already satisfied: slicer==0.0.7 in /usr/local/lib/python3.10/dist-packages (from shap) (0.0.7)
Requirement already satisfied: numba in /usr/local/lib/python3.10/dist-packages (from shap) (0.56.4)
Requirement already satisfied: cloudpickle in /usr/local/lib/python3.10/dist-packages (from shap) (2.2.1)
Requirement already satisfied: llvmlite<0.40,>=0.39.0dev0 in /usr/local/lib/python3.10/dist-packages (from numba->shap) (0.39.1)
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-packages (from numba->shap) (67.7.2)
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/python3.10/dist-packages (from pandas->shap) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas->shap) (2023.3.post1)
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.10/dist-packages (from scikit-learn->shap) (1.3.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn->shap) (3.2.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.1->pandas->shap) (1.16.0)

```
[241]: import shap
explainer = shap.TreeExplainer(dtree)
shap_values = explainer.shap_values(X_test)
shap.summary_plot(shap_values, X_test)
```

No data for colormapping provided via 'c'. Parameters 'vmin', 'vmax' will be ignored



```
[242]: explainer = shap.Explainer(dtree, X_test)
shap_values = explainer(X_test, check_additivity=False)
shap.plots.waterfall(shap_values[0])
```



11 Hyperparameter Tuning

```
[243]: from sklearn.ensemble import RandomForestRegressor
def tune_random_forest(X_train, y_train):
    # Create a Random Forest Regressor object
    rf = RandomForestRegressor(max_features=1.0) # Set max_features explicitly

    # Define the hyperparameter grid
    param_grid = {
        'max_depth': [3, 5, 7, 9],
        'min_samples_split': [2, 5, 10],
        'min_samples_leaf': [1, 2, 4],
        'max_features': ['auto', 'sqrt'],
        'random_state': [0, 42]
    }

    # Create a GridSearchCV object
    grid_search = GridSearchCV(rf, param_grid, cv=5, scoring='r2')

    # Fit the GridSearchCV object to the training data
    grid_search.fit(X_train, y_train)

    # Print the best hyperparameters
    best_params = grid_search.best_params_
    print("Best hyperparameters: ", best_params)
```

```
# Return the best model with tuned hyperparameters
best_rf_model = grid_search.best_estimator_
return best_rf_model
```

```
[244]: best_rf_model = tune_random_forest(X_train, y_train)
best_rf_model.fit(X_train, y_train)
```

`max_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0` or remove this parameter as it is also the default value for RandomForestRegressors and ExtraTreesRegressors.

`max_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0` or remove this parameter as it is also the default value for RandomForestRegressors and ExtraTreesRegressors.

`max_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0` or remove this parameter as it is also the default value for RandomForestRegressors and ExtraTreesRegressors.

`max_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0` or remove this parameter as it is also the default value for RandomForestRegressors and ExtraTreesRegressors.

`max_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0` or remove this parameter as it is also the default value for RandomForestRegressors and ExtraTreesRegressors.

`max_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0` or remove this parameter as it is also the default value for RandomForestRegressors and ExtraTreesRegressors.

`max_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0` or remove this parameter as it is also the default value for RandomForestRegressors and ExtraTreesRegressors.

`max_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0` or remove this parameter as it is also the default value for RandomForestRegressors and ExtraTreesRegressors.

`max_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0` or remove this parameter as it is also the default value for RandomForestRegressors and ExtraTreesRegressors.

`max_features='auto'` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set `max_features=1.0` or remove this parameter as it is also the default value for RandomForestRegressors and ExtraTreesRegressors.

``max_features='auto'`` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set ``max_features=1.0`` or remove this parameter as it is also the default value for `RandomForestRegressors` and `ExtraTreesRegressors`.

``max_features='auto'`` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set ``max_features=1.0`` or remove this parameter as it is also the default value for `RandomForestRegressors` and `ExtraTreesRegressors`.

``max_features='auto'`` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set ``max_features=1.0`` or remove this parameter as it is also the default value for `RandomForestRegressors` and `ExtraTreesRegressors`.

Best hyperparameters: `{'max_depth': 9, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 10, 'random_state': 0}`

``max_features='auto'`` has been deprecated in 1.1 and will be removed in 1.3. To keep the past behaviour, explicitly set ``max_features=1.0`` or remove this parameter as it is also the default value for `RandomForestRegressors` and `ExtraTreesRegressors`.

```
[244]: RandomForestRegressor(max_depth=9, max_features='auto', min_samples_leaf=2,
                             min_samples_split=10, random_state=0)
```

12 Model Evaluation

```
[245]: def evaluate_model(model, X_test, y_test):
        # Make predictions
        y_pred = model.predict(X_test)

        # Calculate evaluation metrics
        mae = mean_absolute_error(y_test, y_pred)
        mape = mean_absolute_percentage_error(y_test, y_pred)
        mse = mean_squared_error(y_test, y_pred)
        r2 = r2_score(y_test, y_pred)
        rmse = math.sqrt(mse)

        # Print the evaluation metrics
        print('MAE is {}'.format(mae))
        print('MAPE is {}'.format(mape))
        print('MSE is {}'.format(mse))
        print('R2 score is {}'.format(r2))
        print('RMSE score is {}'.format(rmse))
```

```
[246]: print("Random Forest Model:")
        evaluate_model(best_rf_model, X_test, y_test)
```

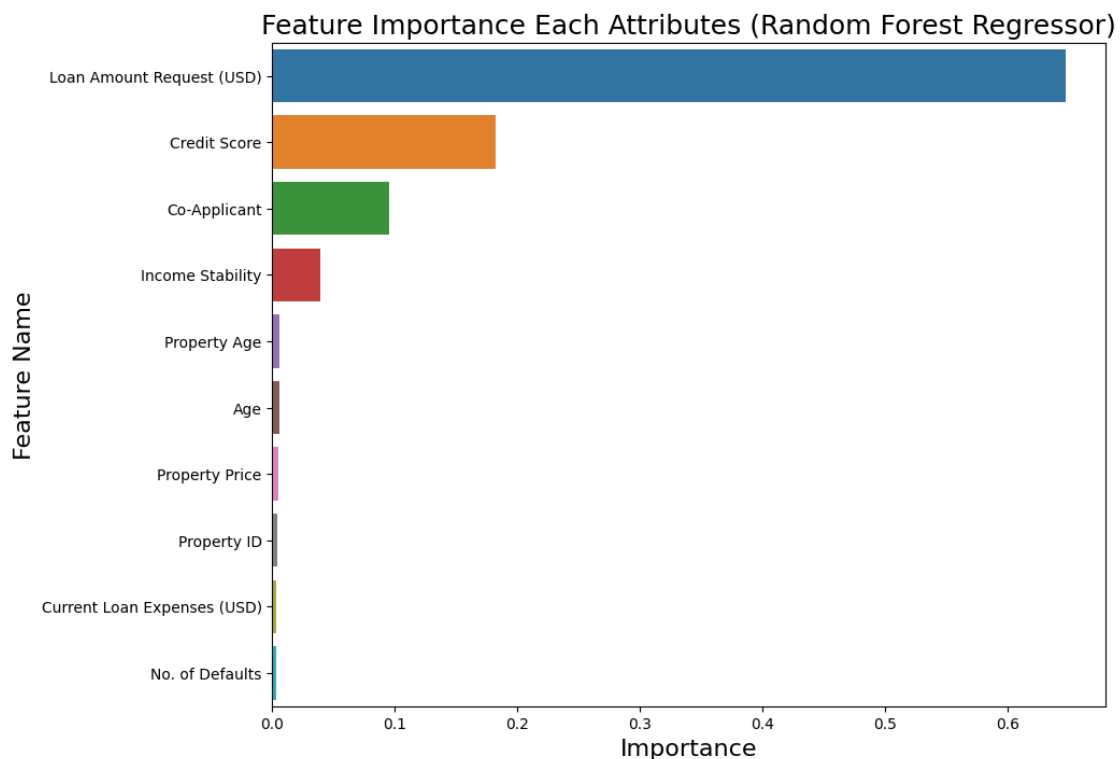
Random Forest Model:

MAE is 11690.229893138958
MAPE is 2.1939785754512978e+19
MSE is 556197583.8459972
R2 score is 0.757933653645164
RMSE score is 23583.841583719928

13 Feature Importance Analysis

```
[247]: def plot_feature_importance(model, X_train):  
    imp_df = pd.DataFrame({  
        "Feature Name": X_train.columns,  
        "Importance": model.feature_importances_  
    })  
    fi = imp_df.sort_values(by="Importance", ascending=False)  
    fi2 = fi.head(10)  
  
    plt.figure(figsize=(10, 8))  
    sns.barplot(data=fi2, x='Importance', y='Feature Name')  
    plt.title('Feature Importance Each Attributes (Random Forest Regressor)',  
             ↪fontsize=18)  
    plt.xlabel('Importance', fontsize=16)  
    plt.ylabel('Feature Name', fontsize=16)  
    plt.show()
```

```
[248]: plot_feature_importance(best_rf_model, X_train)
```



14 SHAP Analysis

```
[249]: def shap_analysis(model, X_test):  
        # SHAP Tree Explainer  
        explainer = shap.TreeExplainer(model)  
        shap_values = explainer.shap_values(X_test)  
  
        # Summary Plot  
        print ('Summary Plot')  
        shap.summary_plot(shap_values, X_test)  
  
        # Waterfall Plot for a single prediction (you can customize the index)  
        print('\n Waterfall Plot for a single prediction ')  
        explainer = shap.Explainer(model, X_test, check_additivity=False)  
        shap_values = explainer(X_test, check_additivity=False)  
        shap.plots.waterfall(shap_values[0])
```

```
[250]: shap_analysis(best_rf_model, X_train)
```

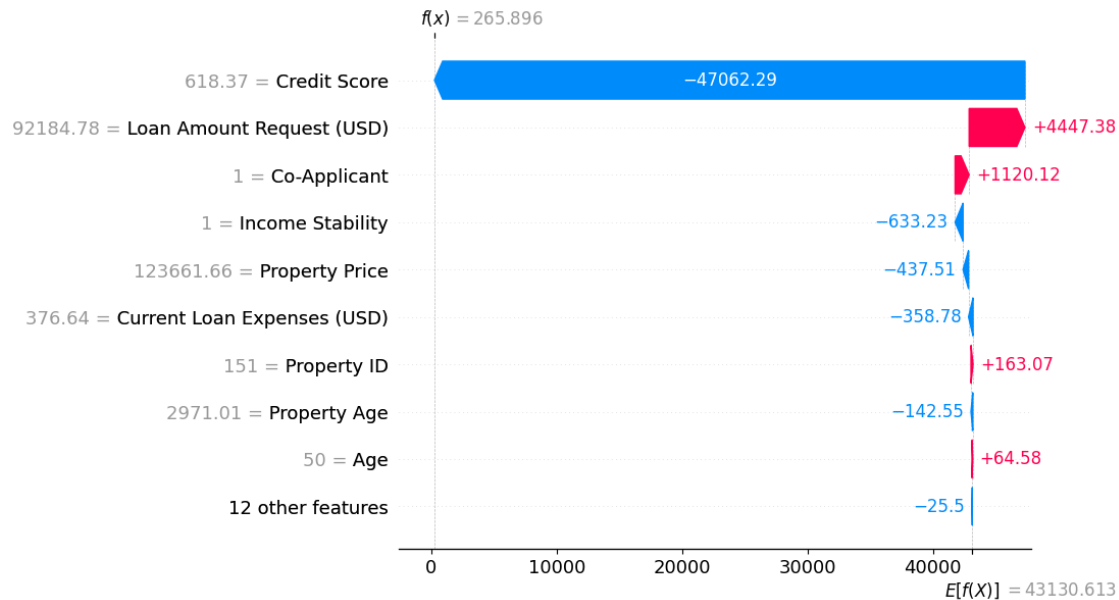
Summary Plot

No data for colormapping provided via 'c'. Parameters 'vmin', 'vmax' will be ignored



Waterfall Plot for a single prediction

99%|=====| 21056/21179 [02:33<00:00]



15 Compare models based on performance metrics

```
[251]: from sklearn.metrics import mean_absolute_error, mean_squared_error,
        mean_absolute_percentage_error, r2_score
        # Compare models based on performance metrics
        def compare_models(model1_metrics, model2_metrics):
            metrics_to_compare = ['MAE', 'MAPE', 'MSE', 'R2', 'RMSE']
            for metric in metrics_to_compare:
                if model1_metrics[metric] < model2_metrics[metric]:
                    print(f"{metric}: Decision Tree model performs better")
                elif model1_metrics[metric] > model2_metrics[metric]:
                    print(f"{metric}: Random Forest model performs better")
                else:
                    print(f"{metric}: Both models perform equally")

        # Get performance metrics for Decision Tree model
        y_pred_dtree = dtree.predict(X_test)
        dtree_metrics = {
            'MAE': mean_absolute_error(y_test, y_pred_dtree),
            'MAPE': mean_absolute_percentage_error(y_test, y_pred_dtree),
            'MSE': mean_squared_error(y_test, y_pred_dtree),
            'R2': r2_score(y_test, y_pred_dtree),
            'RMSE': math.sqrt(mean_squared_error(y_test, y_pred_dtree))
        }

        # Get performance metrics for Random Forest model
```

```

y_pred_rf = best_rf_model.predict(X_test)
rf_metrics = {
    'MAE': mean_absolute_error(y_test, y_pred_rf),
    'MAPE': mean_absolute_percentage_error(y_test, y_pred_rf),
    'MSE': mean_squared_error(y_test, y_pred_rf),
    'R2': r2_score(y_test, y_pred_rf),
    'RMSE': math.sqrt(mean_squared_error(y_test, y_pred_rf))
}

# Compare models
print("\nModel Comparison:")
compare_models(dtmetrics, rf_metrics)

# Select the best model based on the comparison
if dtmetrics['MAE'] < rf_metrics['MAE']:
    best_model = dtree
else:
    best_model = best_rf_model

print("\nBest Model Selected:")
print(best_model)

```

Model Comparison:

MAE: Random Forest model performs better
 MAPE: Random Forest model performs better
 MSE: Random Forest model performs better
 R2: Decision Tree model performs better
 RMSE: Random Forest model performs better

Best Model Selected:

RandomForestRegressor(max_depth=9, max_features='auto', min_samples_leaf=2,
 min_samples_split=10, random_state=0)

16 Lets See How The Model Predicts The Train Data and by how what percentage

```

[252]: # Assuming you have 'new_data' in the same format as your training data
X_new = train.drop('Loan Sanction Amount (USD)', axis=1)
y_new_actual = train['Loan Sanction Amount (USD)']

# Use the model to make predictions on the new data
y_new_pred = best_model.predict(X_new)

# Evaluate the model on the new data
new_data_metrics = {

```

```

    'MAE': mean_absolute_error(y_new_actual, y_new_pred),
    'MAPE': mean_absolute_percentage_error(y_new_actual, y_new_pred),
    'MSE': mean_squared_error(y_new_actual, y_new_pred),
    'R2': r2_score(y_new_actual, y_new_pred),
    'RMSE': math.sqrt(mean_squared_error(y_new_actual, y_new_pred))
}

# Print the evaluation metrics for the new data
print("\nEvaluation Metrics on New Data:")
print(new_data_metrics)

```

Evaluation Metrics on New Data:

```
{'MAE': 10671.90765767297, 'MAPE': 2.047173595112394e+19, 'MSE': 461443783.13640696, 'R2': 0.7929827958328065, 'RMSE': 21481.242588277033}
```

MAE (Mean Absolute Error): MAE measures the average absolute difference between the actual values and the predicted values. It gives you an idea of how far off, on average, your predictions are from the actual values. In this case, the MAE is approximately 10,671.91.

MAPE (Mean Absolute Percentage Error): MAPE is a percentage-based error metric that measures the average percentage difference between the actual values and the predicted values. The value you've provided, 2.047173595112394e+19, is extremely high and likely indicates an issue with the model's predictions. Such a high MAPE suggests that the model's predictions are significantly off from the actual values.

MSE (Mean Squared Error): MSE measures the average squared difference between the actual values and the predicted values. It penalizes larger errors more than smaller ones. The MSE in this case is approximately 461,443,783.14.

R2 (R-squared): R-squared is a measure of how well the model explains the variance in the data. It ranges from 0 to 1, where 1 indicates a perfect fit. An R-squared value of 0.7929 suggests that the model explains about 79.29% of the variance in the data, which is relatively good.

RMSE (Root Mean Squared Error): RMSE is the square root of the MSE and is in the same units as the target variable. It provides a measure of the average magnitude of errors. The RMSE in this case is approximately 21,481.24.

In summary, while the R-squared value indicates that the model is explaining a significant portion of the variance in the data, the extremely high MAPE suggests that the model's predictions are not accurate on a percentage basis.

USING The Best Model To Predict.

First we need to prepare the data. in the formart we can use to predict the new data end evaluate the performance

#This is evaluation of the best model on new data

in preapraing the data we first transform the data into the form that will be accepted by the model

Steps taken to prepare the testing data. test.head()

Check the number of unique value from all of the object datatype
`test.select_dtypes(include='object').nunique()`

Check the amount of missing value `check_missing = test.isnull().sum() * 100 / test.shape[0]`
`check_missing[check_missing > 0].sort_values(ascending=False)`

check segment of employment type `test['Type of Employment'].unique()`

if not formatted apply this `test['Type of Employment'] = test['Type of Employment'].apply(segment_employment_type)`

Processing the new Test Data `preprocess_data(test)`

Loop over each column in the DataFrame where dtype is 'object' for col in
`test.select_dtypes(include=['object']).columns:`

Print the column name and the unique values `print(f"{col}: {test[col].unique()}")`

explore the testing data

`explore_data(test)`

`explore_2(test)`

`explore_3(test)`

`explore_4(test)`

unique encoded values `encoder_object(test)`

`test.shape`

Remove outliers using Z-Score `remove_outliers(test, selected_columns)`

[253]: `test.head()`

```
[253]:   Customer ID      Name Gender  Age  Income (USD)  Income Stability \
0    C-26247  Tandra Olszewski    F   47      3472.69             Low
1    C-35067   Jeannette Cha    F   57      1184.84             Low
2    C-34590    Keva Godfrey    F   52      1266.27             Low
3    C-16668    Elva Sackett    M   65      1369.72             High
4    C-12196   Sade Constable    F   60      1939.23             High
```

```
      Profession Type of Employment  Location \
0  Commercial associate      Managers  Semi-Urban
1              Working      Sales staff    Rural
2              Working              NaN  Semi-Urban
3          Pensioner              NaN    Rural
4          Pensioner              NaN    Urban
```

	Loan Amount Request (USD)	...	Dependents	Credit Score	No. of Defaults	\
0	137088.98	...	2.0	799.14	0	
1	104771.59	...	2.0	833.31	0	
2	176684.91	...	3.0	627.44	0	
3	97009.18	...	2.0	833.20	0	
4	109980.00	...	NaN	NaN	0	

	Has Active Credit Card	Property ID	Property Age	Property Type	\
0	Unpossessed	843	3472.69	2	
1	Unpossessed	22	1184.84	1	
2	Unpossessed	1	1266.27	1	
3	Inactive	730	1369.72	1	
4	NaN	356	1939.23	4	

	Property Location	Co-Applicant	Property Price
0	Urban	1	236644.5
1	Rural	1	142357.3
2	Urban	1	300991.24
3	Semi-Urban	0	125612.1
4	Semi-Urban	1	180908.0

[5 rows x 23 columns]

```
[254]: ##### Check the number of unique value from all of the object datatype
test.select_dtypes(include='object').nunique()
```

```
[254]: Customer ID          20000
Name          20000
Gender         2
Income Stability  2
Profession     7
Type of Employment 18
Location       3
Expense Type 1  2
Expense Type 2  2
Has Active Credit Card  3
Property Location  3
Co-Applicant    3
Property Price  19824
dtype: int64
```

```
[255]: ##### Check the amount of missing value
check_missing = test.isnull().sum() * 100 / test.shape[0]
check_missing[check_missing > 0].sort_values(ascending=False)
```

```
[255]: Type of Employment      23.445
      Dependents              5.710
      Has Active Credit Card   5.380
      Property Age             4.460
      Income Stability         4.065
      Income (USD)             3.750
      Credit Score             3.715
      Property Location        0.800
      Current Loan Expenses (USD) 0.415
      Gender                   0.155
      dtype: float64
```

```
[256]: ##### check segment of employment type
      test['Type of Employment'].unique()
```

```
[256]: array(['Managers', 'Sales staff', nan, 'Laborers', 'Core staff',
      'Medicine staff', 'Accountants', 'High skill tech staff',
      'Secretaries', 'Drivers', 'Cooking staff', 'Security staff',
      'Cleaning staff', 'Low-skill Laborers', 'Waiters/barmen staff',
      'Realty agents', 'Private service staff', 'IT staff', 'HR staff'],
      dtype=object)
```

```
[257]: ##### if segment not formatted apply this

      test['Type of Employment'] = test['Type of Employment'].
      ↪ apply(segment_employment_type)

      test['Type of Employment'].unique()
```

```
[257]: array(['Managers', 'Sales/Realty', 'Unknown', 'Laborers', 'Other',
      'Secretaries/HR', 'Hospitality', 'Tech/IT'], dtype=object)
```

```
[258]: def preprocess_data(df):
      # Remove identifier columns
      df.drop(columns=['Customer ID', 'Name'], inplace=True)

      # Segment Type of Employment into smaller unique values
      df['Type of Employment'] = df['Type of Employment'].
      ↪ apply(segment_employment_type)

      # Handle missing values
      df.fillna({
          'Property Age': df['Property Age'].median(),
          'Income (USD)': df['Income (USD)'].median(),
          'Dependents': df['Dependents'].median(),
```

```

        'Credit Score': df['Credit Score'].median(),
        'Current Loan Expenses (USD)': df['Current Loan Expenses (USD)'].
median()
    }, inplace=True)

    df.dropna(subset=['Income Stability', 'Has Active Credit Card', 'Property_L
Location', 'Gender'], inplace=True)

    return df.head(10)

```

```

[259]: ##### Processing the new Test Data
preprocess_data(test)

```

```

[259]:
Gender  Age  Income (USD)  Income Stability  Profession \
0      F   47      3472.69             Low  Commercial associate
1      F   57      1184.84             Low             Working
2      F   52      1266.27             Low             Working
3      M   65      1369.72             High            Pensioner
5      F   59      2944.81             Low             Working
6      M   43      1957.31             Low             Working
7      M   65      1403.63             High            Pensioner
8      M   64      1604.65             Low             Working
10     F   43      2037.14             Low  Commercial associate
11     M   29      2183.59             Low  Commercial associate

Type of Employment  Location  Loan Amount Request (USD) \
0      Managers  Semi-Urban      137088.98
1  Sales/Realty  Rural      104771.59
2      Other  Semi-Urban      176684.91
3      Other  Rural      97009.18
5  Sales/Realty  Semi-Urban      31465.78
6  Sales/Realty  Rural      150334.11
7      Other  Semi-Urban      121029.27
8      Other  Semi-Urban      39475.31
10     Other  Rural      123472.08
11  Laborers  Semi-Urban      53651.03

Current Loan Expenses (USD)  Expense Type 1  ...  Dependents  Credit Score \
0      396.72      N  ...      2.0      799.14
1      463.76      Y  ...      2.0      833.31
2      493.15      N  ...      3.0      627.44
3      446.15      N  ...      2.0      833.20
5      153.10      Y  ...      2.0      620.58
6      433.82      N  ...      2.0      731.37
7      348.18      N  ...      1.0      838.66
8      200.41      N  ...      1.0      589.97
10     350.99      N  ...      2.0      739.30

```

11	216.60	N ...	3.0	653.00
----	--------	-------	-----	--------

	No. of Defaults	Has Active Credit Card	Property ID	Property Age \
0	0	Unpossessed	843	3472.69
1	0	Unpossessed	22	1184.84
2	0	Unpossessed	1	1266.27
3	0	Inactive	730	1369.72
5	0	Inactive	497	2944.81
6	0	Unpossessed	206	1957.31
7	1	Active	339	1403.63
8	0	Active	465	1604.65
10	1	Active	988	2037.14
11	0	Unpossessed	307	2183.59

	Property Type	Property Location	Co-Applicant	Property Price
0	2	Urban	1	236644.5
1	1	Rural	1	142357.3
2	1	Urban	1	300991.24
3	1	Semi-Urban	0	125612.1
5	1	Semi-Urban	0	51075.31
6	4	Semi-Urban	1	232535.55
7	1	Urban	1	165652.67
8	4	Rural	1	59225.24
10	4	Rural	1	212239.85
11	1	Semi-Urban	0	79681.39

[10 rows x 21 columns]

```
[260]: ##### Loop over each column in the DataFrame where dtype is 'object'
for col in test.select_dtypes(include=['object']).columns:

    # Print the column name and the unique values
    print(f"{col}: {test[col].unique()}")
```

```
Gender: ['F' 'M']
Income Stability: ['Low' 'High']
Profession: ['Commercial associate' 'Working' 'Pensioner' 'State servant'
'Unemployed'
'Maternity leave' 'Student']
Type of Employment: ['Managers' 'Sales/Realty' 'Other' 'Laborers'
'Secretaries/HR' 'Tech/IT']
Location: ['Semi-Urban' 'Rural' 'Urban']
Expense Type 1: ['N' 'Y']
Expense Type 2: ['N' 'Y']
Has Active Credit Card: ['Unpossessed' 'Inactive' 'Active']
Property Location: ['Urban' 'Rural' 'Semi-Urban']
Co-Applicant: ['1' '0' '?']
```

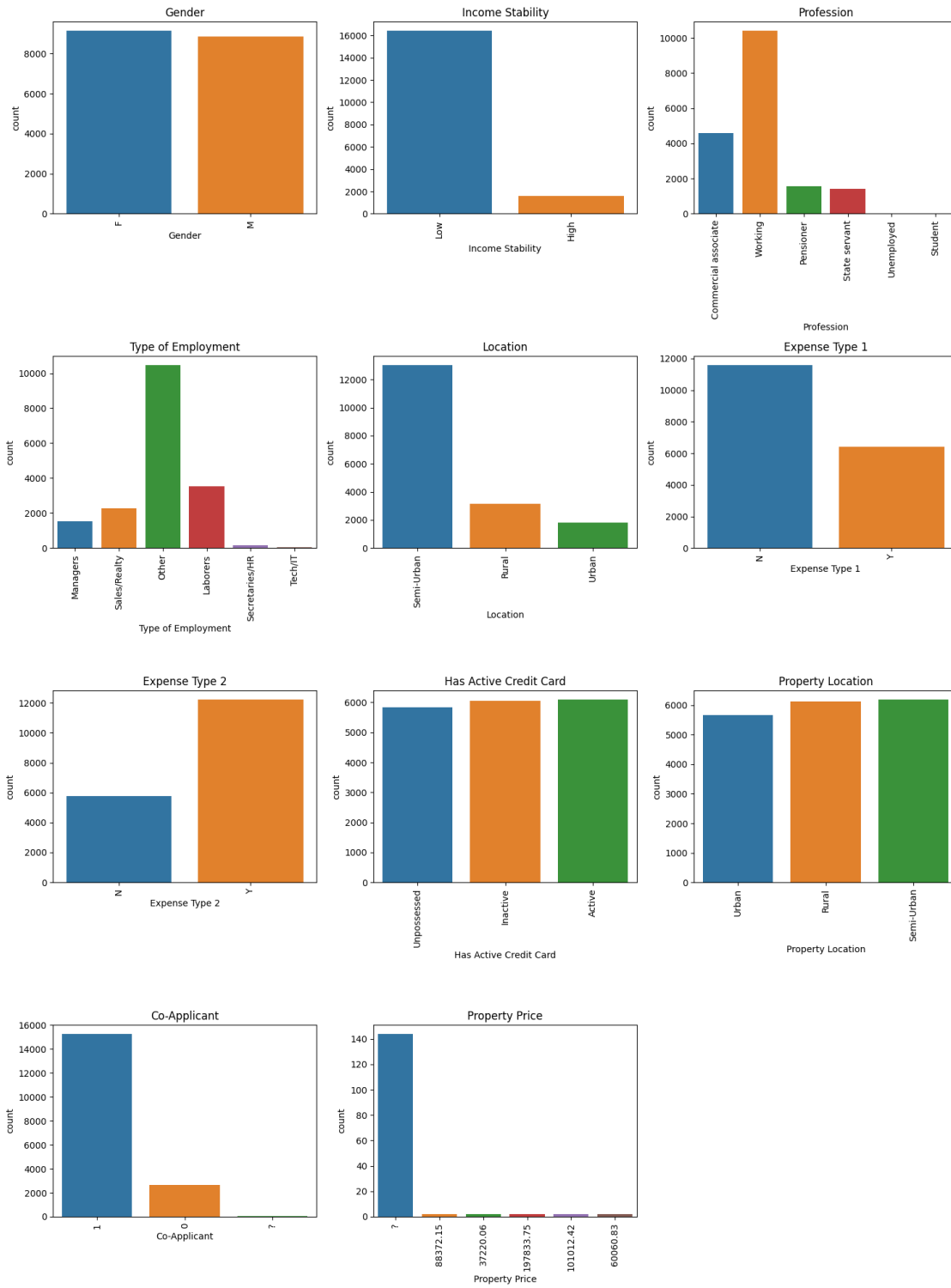
```
Property Price: ['236644.5' '142357.3' '300991.24' ... '21566.27' '120281.17'
'133425.43']
```

17 explore the testing data

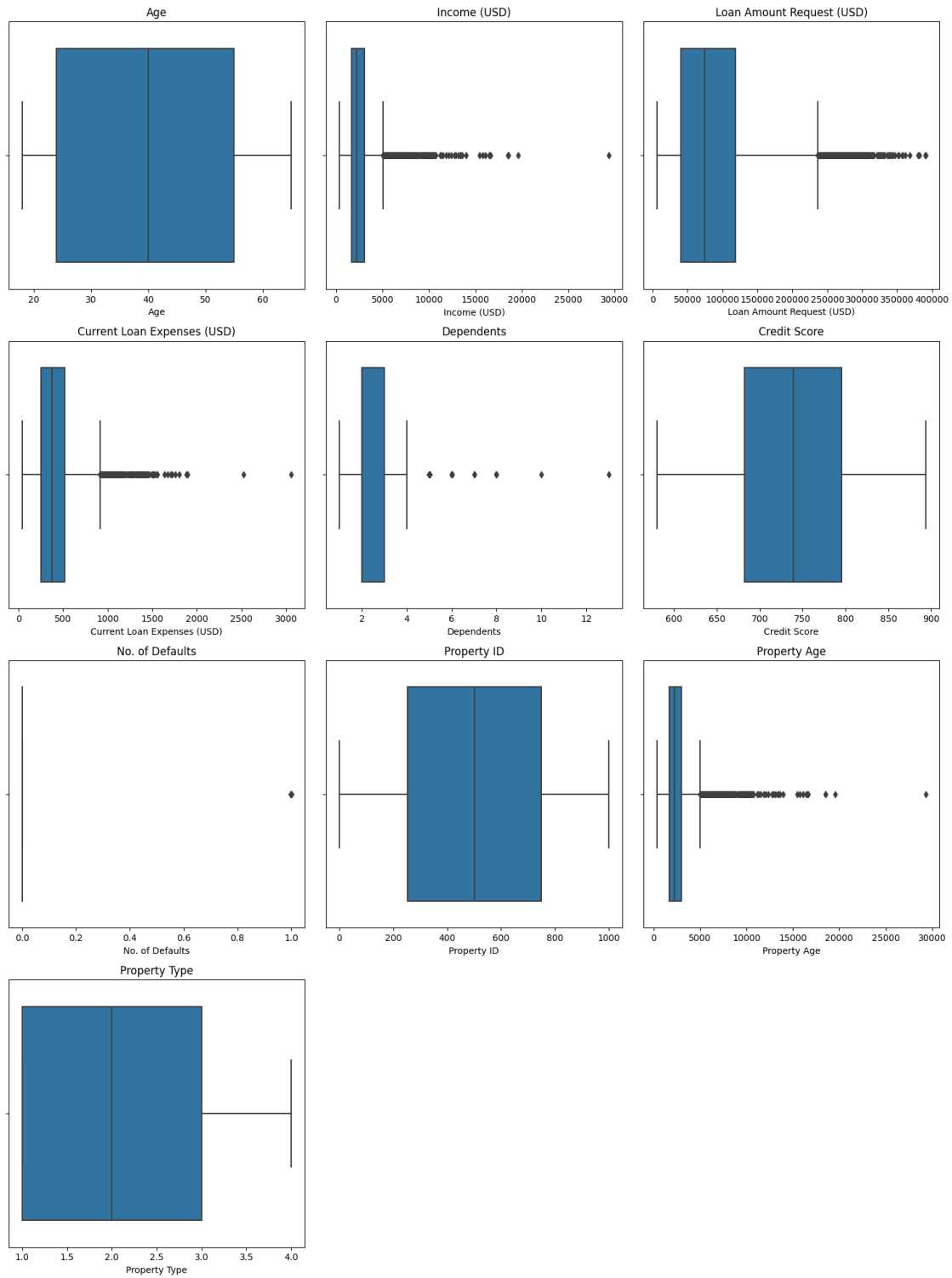
```
[261]: test.dtypes
```

```
[261]: Gender                object
Age                int64
Income (USD)       float64
Income Stability   object
Profession         object
Type of Employment object
Location           object
Loan Amount Request (USD) float64
Current Loan Expenses (USD) float64
Expense Type 1     object
Expense Type 2     object
Dependents         float64
Credit Score      float64
No. of Defaults    int64
Has Active Credit Card object
Property ID        int64
Property Age       float64
Property Type      int64
Property Location  object
Co-Applicant       object
Property Price     object
dtype: object
```

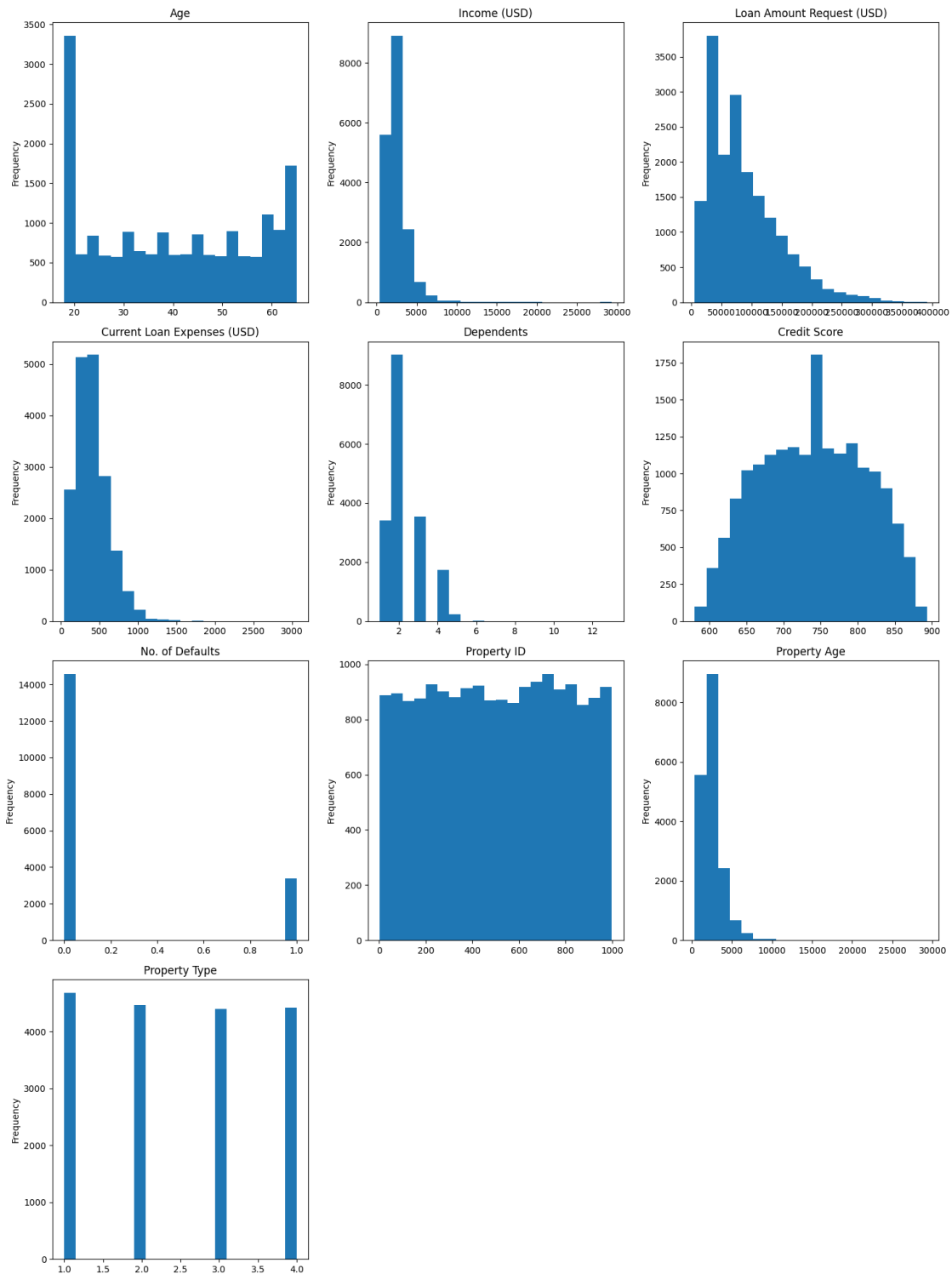
```
[262]: explore_data(test)
```



[263]: eplore_2(test)



[264]: `exprole_3(test)`



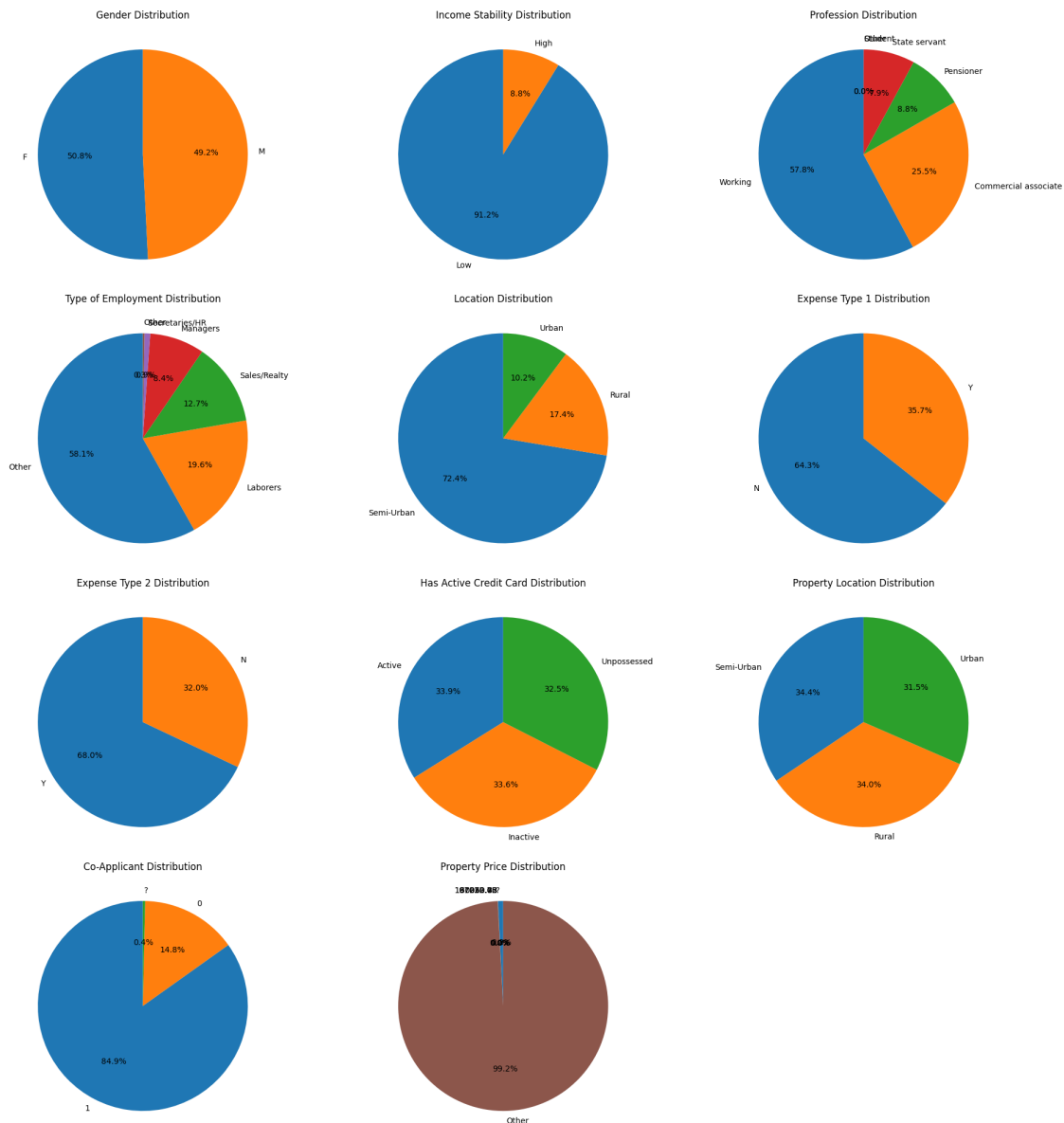
```
[265]: exprole_4(test)
```

The `series.append` method is deprecated and will be removed from pandas in a

future version. Use `pandas.concat` instead.

The `series.append` method is deprecated and will be removed from pandas in a future version. Use `pandas.concat` instead.

The `series.append` method is deprecated and will be removed from pandas in a future version. Use `pandas.concat` instead.



[266]: `#### unique encoded values`

```
encoder_object(test)
```

Gender: [0 1]

```

Income Stability: [1 0]
Profession: [0 6 2 3 5 1 4]
Type of Employment: [1 3 2 0 4 5]
Location: [1 0 2]
Expense Type 1: [0 1]
Expense Type 2: [0 1]
Has Active Credit Card: [2 1 0]
Property Location: [2 0 1]
Co-Applicant: [1 0 2]
Property Price: [7955 3439 9626 ... 7179 1781 2805]

```

```
[267]: test.shape
```

```
[267]: (17981, 21)
```

```
[268]: ##### Remove outliers using Z-Score
remove_outliers(test, selected_columns)
```

```
[268]:
```

	Gender	Age	Income (USD)	Income Stability	Profession	\
0	0	47	3472.69	1	0	
1	0	57	1184.84	1	6	
2	0	52	1266.27	1	6	
3	1	65	1369.72	0	2	
5	0	59	2944.81	1	6	
6	1	43	1957.31	1	6	
7	1	65	1403.63	0	2	
8	1	64	1604.65	1	6	
10	0	43	2037.14	1	0	
11	1	29	2183.59	1	0	

	Type of Employment	Location	Loan Amount Request (USD)	\
0		1	1	137088.98
1		3	0	104771.59
2		2	1	176684.91
3		2	0	97009.18
5		3	1	31465.78
6		3	0	150334.11
7		2	1	121029.27
8		2	1	39475.31
10		2	0	123472.08
11		0	1	53651.03

	Current Loan Expenses (USD)	Expense Type 1	...	Dependents	\
0	396.72	0	...	2.0	
1	463.76	1	...	2.0	
2	493.15	0	...	3.0	
3	446.15	0	...	2.0	

5	153.10	1 ...	2.0
6	433.82	0 ...	2.0
7	348.18	0 ...	1.0
8	200.41	0 ...	1.0
10	350.99	0 ...	2.0
11	216.60	0 ...	3.0

	Credit Score	No. of Defaults	Has Active Credit Card	Property ID \
0	799.14	0	2	843
1	833.31	0	2	22
2	627.44	0	2	1
3	833.20	0	1	730
5	620.58	0	1	497
6	731.37	0	2	206
7	838.66	1	0	339
8	589.97	0	0	465
10	739.30	1	0	988
11	653.00	0	2	307

	Property Age	Property Type	Property Location	Co-Applicant \
0	3472.69	2	2	1
1	1184.84	1	0	1
2	1266.27	1	2	1
3	1369.72	1	1	0
5	2944.81	1	1	0
6	1957.31	4	1	1
7	1403.63	1	2	1
8	1604.65	4	0	1
10	2037.14	4	0	1
11	2183.59	1	1	0

	Property Price
0	7955
1	3439
2	9626
3	2211
5	12965
6	7806
7	4874
8	13991
10	7038
11	16118

[10 rows x 21 columns]

18 Use the model to make predictions on the new data

Lets first copy the test data to new data so to avoid termparing with the data

```
[269]: new_data = test.copy()

[270]: # Assuming 'new_data' doesn't contain the 'Loan Sanction Amount (USD)' column
X_new = new_data # Replace 'new_data' with your actual new data

# Use the model to make predictions on the new data
y_new_pred = best_model.predict(X_new)

# Create a new column in 'new_data' and populate it with the predicted values
new_data['Predicted Loan Sanction Amount (USD)'] = y_new_pred

# Save the data with the predicted column to a CSV file
test_data_predicted = new_data.to_csv('test_data_with_predictions.csv',
↪index=False) # Change the filename as needed
```

19 Lets Predict Loan Sanction For an Individual

```
[272]: def predict_loan_sanction_amount():
    # Prompt the user to input values for each feature
    print("Please enter the following information:")
    print('Example First Name')
    client_name = input('Type Your Name: ')
    gender = input("Gender (1-Male/0-Female): ")
    age = int(input("Age: "))
    income_usd = float(input("Income (USD): "))
    income_stability = input("Income Stability (0 - Stable/1- Unstable): ")

    print("4 - 'Working', 1- 'Pensioner', 2-'State servant', 0 - 'Commercial_
↪associate', 3- 'Unemployed'")
    profession = input("Profession: ")
    print ("Employment : 7 -'Unknown', 3- 'Other', 5- 'Secretaries/HR', 1- 
↪'Laborers',2- 'Managers', 0- 'Hospitality',4- 'Sales/Realty' ,6- 'Tech/IT'")
    type_of_employment = input("Type of Employment: ")

    print("1 - 'Semi-Urban', 0-'Rural', 2- 'Urban'")
    location = input("Location: ")

    loan_amount_requested = float(input("Loan Amount Requested (USD): "))
    current_loan_expenses_usd = float(input("Current Loan Expenses (USD): "))

    expense_type_1 = input("Expense Type 1 (1- Yes, 0 - No): ")

    expense_type_2 = input("Expense Type 2 (1- Yes, 0 - No): ")
```

```

dependents = (input("Dependents(example: 8.9, 2.0..): "))
credit_score = float(input("Credit Score: "))
num_of_defaults = int(input("No. of Defaults: "))

has_active_credit_card = input("Has Active Credit Card (2 - 'Unpossessed', 1- 'Active', 0- 'Inactive'): ")

property_id = input("Property ID: ")
property_age = float(input("Property Age: "))

property_type = input("Property Type (1 - 4): ")

property_location = input("Property Location (0- Rural, 1- Semiurban, 2- Urban): ")

co_applicant = input("Co-Applicant (1 - Yes/0 - No): ")

property_price = float(input("Property Price: "))

# Create a dictionary with input values
input_data = {
    'Gender': [gender],
    'Age': [age],
    'Income (USD)': [income_usd],
    'Income Stability': [income_stability],
    'Profession': [profession],
    'Type of Employment': [type_of_employment],
    'Location': [location],
    'Loan Amount Request (USD)': [loan_amount_requested],
    'Current Loan Expenses (USD)': [current_loan_expenses_usd],
    'Expense Type 1': [expense_type_1],
    'Expense Type 2': [expense_type_2],
    'Dependents': [dependents],
    'Credit Score': [credit_score],
    'No. of Defaults': [num_of_defaults],
    'Has Active Credit Card': [has_active_credit_card],
    'Property ID': [property_id],
    'Property Age': [property_age],
    'Property Type': [property_type],
    'Property Location': [property_location],
    'Co-Applicant': [co_applicant],
    'Property Price': [property_price]
}

# Create a DataFrame from the input data
input_df = pd.DataFrame(input_data)

```

```

# Preprocess the input data (similar to training data preprocessing)
# ... (apply the same preprocessing steps as in your original code)

# Use the model to make predictions on the preprocessed input data
predicted_loan_sanction_amount = best_model.predict(input_df)

# Display the predicted 'Loan Sanction Amount (USD)'
print(f"\n Predicted Loan Sanction Amount (USD) for {client_name} :₹
↪{predicted_loan_sanction_amount[0]:.2f}")

```

```

[273]: # Call the function to predict 'Loan Sanction Amount (USD)' for an individual
predict_loan_sanction_amount()

```

Please enter the following information:

Example First Name

Type Your Name: Joseph

Gender (1-Male/0-Female): 1

Age: 42

Income (USD): 5000.05

Income Stability (0 - Stable/1- Unstable): 0

4 - 'Working', 1- 'Pensioner', 2-'State servant', 0 - 'Commercial associate', 3- 'Unemployed'

Profession: 4

Employment : 7 -'Unknown', 3- 'Other', 5- 'Secretaries/HR', 1- 'Laborers',2- 'Managers', 0- 'Hospitality',4- 'Sales/Realty' ,6- 'Tech/IT'

Type of Employment: 6

1 - 'Semi-Urban', 0-'Rural', 2- 'Urban'

Location: 1

Loan Amount Requested (USD): 1000000.35

Current Loan Expenses (USD): 200.32

Expense Type 1 (1- Yes, 0 - No): 1

Expense Type 2 (1- Yes, 0 - No): 1

Dependents(example: 8.9, 2.0..): 2.3

Credit Score: 568354.36

No. of Defaults: 0

Has Active Credit Card (2 - 'Unpossessed', 0- 'Active', 1- 'Inactive'): 0

Property ID: 564

Property Age: 564

Property Type (1 - 4): 2

Property Location (0- Rural, 1- Semiurban, 2- Urban): 2

Co-Applicant (1 - Yes/0 - No): 1

Property Price: 25689325.25

Predicted Loan Sanction Amount (USD) for Joseph : 302677.12