Society Score Prediction

Exploratory Data Analysis (EDA)

In [1]:
```
# Import all library
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.impute import SimpleImputer
from sklearn.pipeline import Pipeline
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import LabelEncoder
```

Insights

Importing libraries is the first step in setting up the environment for data analysis and modeling. Libraries like Pandas, NumPy, Matplotlib, Seaborn, and scikit-learn are commonly used for various data-related tasks.

Data Loading and Overview

In [2]:
```
pd.set_option("Display.max_column",100)
df = pd.read_csv('Energy_and_Water_Data_Disclosure_for_Local_Law_84_2017__Data_for_Calendar_Year_2016_.csv')
```

Insights

Adjusting display options can improve the readability of DataFrame outputs, especially when dealing with datasets with many columns. And Loading data from external sources like CSV files is a common task in data analysis.Pandas provides the 'read_csv()' function for reading data from CSV files into DataFrames.

In [3]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11746 entries, 0 to 11745
Data columns (total 60 columns):
 #   Column                                                Non-Null Count  Dtype
---  ------                                                --------------  -----
 0   Order                                                 11746 non-null  int64
 1   Property Id                                           11746 non-null  int64
 2   Property Name                                         11746 non-null  object
 3   Parent Property Id                                    11746 non-null  object
 4   Parent Property Name                                  11746 non-null  object
 5   BBL - 10 digits                                       11735 non-null  object
 6   NYC Borough, Block and Lot (BBL) self-reported        11746 non-null  object
 7   NYC Building Identification Number (BIN)              11746 non-null  object
 8   Address 1 (self-reported)                             11746 non-null  object
 9   Address 2                                             11746 non-null  object
 10  Postal Code                                           11746 non-null  object
 11  Street Number                                         11622 non-null  object
 12  Street Name                                           11624 non-null  object
 13  Borough                                               11628 non-null  object
 14  DOF Gross Floor Area                                  11628 non-null  float64
 15  Primary Property Type - Self Selected                 11746 non-null  object
 16  List of All Property Use Types at Property            11746 non-null  object
 17  Largest Property Use Type                             11746 non-null  object
 18  Largest Property Use Type - Gross Floor Area (ft²)    11746 non-null  object
 19  2nd Largest Property Use Type                         11746 non-null  object
 20  2nd Largest Property Use - Gross Floor Area (ft²)     11746 non-null  object
 21  3rd Largest Property Use Type                         11746 non-null  object
 22  3rd Largest Property Use Type - Gross Floor Area (ft²)  11746 non-null  object
 23  Year Built                                            11746 non-null  int64
 24  Number of Buildings - Self-reported                   11746 non-null  int64
 25  Occupancy                                             11746 non-null  int64
 26  Metered Areas (Energy)                                11746 non-null  object
 27  Metered Areas  (Water)                                11746 non-null  object
 28  ENERGY STAR Score                                     11746 non-null  object
 29  Site EUI (kBtu/ft²)                                   11746 non-null  object
 30  Weather Normalized Site EUI (kBtu/ft²)                11746 non-null  object
 31  Weather Normalized Site Electricity Intensity (kWh/ft²)  11746 non-null  object
 32  Weather Normalized Site Natural Gas Intensity (therms/ft²)  11746 non-null  object
```

```
 33  Weather Normalized Source EUI (kBtu/ft²)                  11746 non-null  object
 34  Fuel Oil #1 Use (kBtu)                                    11746 non-null  object
 35  Fuel Oil #2 Use (kBtu)                                    11746 non-null  object
 36  Fuel Oil #4 Use (kBtu)                                    11746 non-null  object
 37  Fuel Oil #5 & 6 Use (kBtu)                                11746 non-null  object
 38  Diesel #2 Use (kBtu)                                      11746 non-null  object
 39  District Steam Use (kBtu)                                 11746 non-null  object
 40  Natural Gas Use (kBtu)                                    11746 non-null  object
 41  Weather Normalized Site Natural Gas Use (therms)          11746 non-null  object
 42  Electricity Use - Grid Purchase (kBtu)                    11746 non-null  object
 43  Weather Normalized Site Electricity (kWh)                 11746 non-null  object
 44  Total GHG Emissions (Metric Tons CO2e)                    11746 non-null  object
 45  Direct GHG Emissions (Metric Tons CO2e)                   11746 non-null  object
 46  Indirect GHG Emissions (Metric Tons CO2e)                 11746 non-null  object
 47  Property GFA - Self-Reported (ft²)                        11746 non-null  int64
 48  Water Use (All Water Sources) (kgal)                      11746 non-null  object
 49  Water Intensity (All Water Sources) (gal/ft²)             11746 non-null  object
 50  Source EUI (kBtu/ft²)                                     11746 non-null  object
 51  Release Date                                              11746 non-null  object
 52  Water Required?                                           11628 non-null  object
 53  DOF Benchmarking Submission Status                        11716 non-null  object
 54  Latitude                                                  9483 non-null   float64
 55  Longitude                                                 9483 non-null   float64
 56  Community Board                                           9483 non-null   float64
 57  Council District                                          9483 non-null   float64
 58  Census Tract                                              9483 non-null   float64
 59  NTA                                                       9483 non-null   object
dtypes: float64(6), int64(6), object(48)
memory usage: 5.4+ MB
```

In [4]:
```
df.columns
```

Out[4]:
```
Index(['Order', 'Property Id', 'Property Name', 'Parent Property Id',
       'Parent Property Name', 'BBL - 10 digits',
       'NYC Borough, Block and Lot (BBL) self-reported',
       'NYC Building Identification Number (BIN)', 'Address 1 (self-reported)',
       'Address 2', 'Postal Code', 'Street Number', 'Street Name', 'Borough',
       'DOF Gross Floor Area', 'Primary Property Type - Self Selected',
       'List of All Property Use Types at Property',
       'Largest Property Use Type',
       'Largest Property Use Type - Gross Floor Area (ft²)',
       '2nd Largest Property Use Type',
       '2nd Largest Property Use - Gross Floor Area (ft²)',
       '3rd Largest Property Use Type',
       '3rd Largest Property Use Type - Gross Floor Area (ft²)', 'Year Built',
       'Number of Buildings - Self-reported', 'Occupancy',
       'Metered Areas (Energy)', 'Metered Areas  (Water)', 'ENERGY STAR Score',
       'Site EUI (kBtu/ft²)', 'Weather Normalized Site EUI (kBtu/ft²)',
       'Weather Normalized Site Electricity Intensity (kWh/ft²)',
       'Weather Normalized Site Natural Gas Intensity (therms/ft²)',
       'Weather Normalized Source EUI (kBtu/ft²)', 'Fuel Oil #1 Use (kBtu)',
       'Fuel Oil #2 Use (kBtu)', 'Fuel Oil #4 Use (kBtu)',
       'Fuel Oil #5 & 6 Use (kBtu)', 'Diesel #2 Use (kBtu)',
       'District Steam Use (kBtu)', 'Natural Gas Use (kBtu)',
       'Weather Normalized Site Natural Gas Use (therms)',
```

```
'Electricity Use - Grid Purchase (kBtu)',
'Weather Normalized Site Electricity (kWh)',
'Total GHG Emissions (Metric Tons CO2e)',
'Direct GHG Emissions (Metric Tons CO2e)',
'Indirect GHG Emissions (Metric Tons CO2e)',
'Property GFA - Self-Reported (ft²)',
'Water Use (All Water Sources) (kgal)',
'Water Intensity (All Water Sources) (gal/ft²)',
'Source EUI (kBtu/ft²)', 'Release Date', 'Water Required?',
'DOF Benchmarking Submission Status', 'Latitude', 'Longitude',
'Community Board', 'Council District', 'Census Tract', 'NTA'],
dtype='object')
```

In [5]:
```
df.head()
```

Out[5]:

	Order	Property Id	Property Name	Parent Property Id	Parent Property Name	BBL - 10 digits	NYC Borough, Block and Lot (BBL) self-reported	NYC Building Identification Number (BIN)	Address 1 (self-reported)	Address 2	Postal Code	Street Number	Street Name	Borou
0	1	13286	201/205	13286	201/205	1013160001	1013160001	1037549	201/205 East 42nd st.	Not Available	10017	675	3 AVENUE	Manhatt
1	2	28400	NYP Columbia (West Campus)	28400	NYP Columbia (West Campus)	1021380040	1-02138-0040	1084198; 1084387;1084385; 1084386; 1084388; 10...	622 168th Street	Not Available	10032	180	FT WASHINGTON AVENUE	Manhatt
2	3	4778226	MSCHoNY North	28400	NYP Columbia (West Campus)	1021380030	1-02138-0030	1063380	3975 Broadway	Not Available	10032	3975	BROADWAY	Manhatt
3	4	4778267	Herbert Irving Pavilion & Millstein Hospital	28400	NYP Columbia (West Campus)	1021390001	1-02139-0001	1087281; 1076746	161 Fort Washington Ave	177 Fort Washington Ave	10032	161	FT WASHINGTON AVENUE	Manhatt

4	5	4778288	Neuro Institute	28400	NYP Columbia (West Campus)	1021390085	1-02139-0085	1063403	710 West 168th Street	Not Available	10032	193	FT WASHINGTON AVENUE	Manhatta

Insights

The info() method gives a concise summary of the DataFrame, including the number of non-null entries and data types of each column. Accessing columns attribute provides the list of column names. head() method displays the first few rows of the DataFrame, which helps understand its structure and contents.

Data Cleaning

In [6]:
```
print(df.isnull().sum())
```

```
Order                                                  0
Property Id                                            0
Property Name                                          0
Parent Property Id                                     0
Parent Property Name                                   0
BBL - 10 digits                                       11
NYC Borough, Block and Lot (BBL) self-reported         0
NYC Building Identification Number (BIN)               0
Address 1 (self-reported)                              0
Address 2                                              0
Postal Code                                            0
Street Number                                        124
Street Name                                          122
Borough                                              118
DOF Gross Floor Area                                 118
Primary Property Type - Self Selected                  0
List of All Property Use Types at Property             0
Largest Property Use Type                              0
Largest Property Use Type - Gross Floor Area (ft²)     0
2nd Largest Property Use Type                          0
```

```
2nd Largest Property Use - Gross Floor Area (ft²)                0
3rd Largest Property Use Type                                    0
3rd Largest Property Use Type - Gross Floor Area (ft²)           0
Year Built                                                       0
Number of Buildings - Self-reported                              0
Occupancy                                                        0
Metered Areas (Energy)                                           0
Metered Areas  (Water)                                           0
ENERGY STAR Score                                                0
Site EUI (kBtu/ft²)                                              0
Weather Normalized Site EUI (kBtu/ft²)                           0
Weather Normalized Site Electricity Intensity (kWh/ft²)          0
Weather Normalized Site Natural Gas Intensity (therms/ft²)       0
Weather Normalized Source EUI (kBtu/ft²)                         0
Fuel Oil #1 Use (kBtu)                                           0
Fuel Oil #2 Use (kBtu)                                           0
Fuel Oil #4 Use (kBtu)                                           0
Fuel Oil #5 & 6 Use (kBtu)                                       0
Diesel #2 Use (kBtu)                                             0
District Steam Use (kBtu)                                        0
Natural Gas Use (kBtu)                                           0
Weather Normalized Site Natural Gas Use (therms)                 0
Electricity Use - Grid Purchase (kBtu)                           0
Weather Normalized Site Electricity (kWh)                        0
Total GHG Emissions (Metric Tons CO2e)                           0
Direct GHG Emissions (Metric Tons CO2e)                          0
Indirect GHG Emissions (Metric Tons CO2e)                        0
Indirect GHG Emissions (Metric Tons CO2e)                        0
Property GFA - Self-Reported (ft²)                               0
Water Use (All Water Sources) (kgal)                             0
Water Intensity (All Water Sources) (gal/ft²)                    0
Source EUI (kBtu/ft²)                                            0
Release Date                                                     0
Water Required?                                                118
DOF Benchmarking Submission Status                              30
Latitude                                                      2263
Longitude                                                     2263
Community Board                                               2263
Council District                                              2263
Census Tract                                                  2263
NTA                                                           2263
dtype: int64
```

Insights

Checking for missing values is an essential step in data preprocessing to ensure data quality and reliability.

In [7]:
```
# Drop rows with missing values for the 'ENERGY STAR Score' column
df.dropna(subset=['ENERGY STAR Score'], inplace=True)
print(df.dtypes)
```

```
Order                                                    int64
Property Id                                              int64
Property Name                                           object
Parent Property Id                                      object
Parent Property Name                                    object
BBL - 10 digits                                         object
NYC Borough, Block and Lot (BBL) self-reported          object
NYC Building Identification Number (BIN)                object
Address 1 (self-reported)                               object
Address 2                                               object
Postal Code                                             object
Street Number                                           object
Street Name                                             object
Borough                                                 object
DOF Gross Floor Area                                   float64
Primary Property Type - Self Selected                   object
List of All Property Use Types at Property              object
Largest Property Use Type                               object
Largest Property Use Type - Gross Floor Area (ft²)      object
2nd Largest Property Use Type                           object
2nd Largest Property Use - Gross Floor Area (ft²)       object
3rd Largest Property Use Type                           object
3rd Largest Property Use Type - Gross Floor Area (ft²)  object
Year Built                                               int64
Number of Buildings - Self-reported                      int64
Occupancy                                                int64
Metered Areas (Energy)                                  object
Metered Areas  (Water)                                  object
ENERGY STAR Score                                       object
Site EUI (kBtu/ft²)                                     object
Weather Normalized Site EUI (kBtu/ft²)                  object
Weather Normalized Site Electricity Intensity (kWh/ft²) object
Weather Normalized Site Natural Gas Intensity (therms/ft²) object
Weather Normalized Source EUI (kBtu/ft²)                object
Fuel Oil #1 Use (kBtu)                                  object
Fuel Oil #2 Use (kBtu)                                  object
Fuel Oil #4 Use (kBtu)                                  object
Fuel Oil #5 & 6 Use (kBtu)                              object
Diesel #2 Use (kBtu)                                    object
District Steam Use (kBtu)                               object
Natural Gas Use (kBtu)                                  object
Weather Normalized Site Natural Gas Use (therms)        object
Electricity Use - Grid Purchase (kBtu)                  object
Weather Normalized Site Electricity (kWh)               object
Total GHG Emissions (Metric Tons CO2e)                  object
Direct GHG Emissions (Metric Tons CO2e)                 object
Indirect GHG Emissions (Metric Tons CO2e)               object
Property GFA - Self-Reported (ft²)                       int64
Water Use (All Water Sources) (kgal)                    object
Water Intensity (All Water Sources) (gal/ft²)           object
```

In [7]:
```
# Drop rows with missing values for the 'ENERGY STAR Score' column
df.dropna(subset=['ENERGY STAR Score'], inplace=True)
print(df.dtypes)
```

```
Order                                                 int64
Property Id                                           int64
Property Name                                        object
Parent Property Id                                   object
Parent Property Name                                 object
BBL - 10 digits                                      object
NYC Borough, Block and Lot (BBL) self-reported       object
NYC Building Identification Number (BIN)             object
Address 1 (self-reported)                            object
Address 2                                            object
Postal Code                                          object
Street Number                                        object
Street Name                                          object
Borough                                              object
DOF Gross Floor Area                                float64
Primary Property Type - Self Selected                object
List of All Property Use Types at Property           object
Largest Property Use Type                            object
Largest Property Use Type - Gross Floor Area (ft²)   object
2nd Largest Property Use Type                        object
2nd Largest Property Use - Gross Floor Area (ft²)    object
3rd Largest Property Use Type                        object
Source EUI (kBtu/ft²)                                object
Release Date                                         object
Water Required?                                      object
DOF Benchmarking Submission Status                   object
Latitude                                            float64
Longitude                                           float64
Community Board                                     float64
Council District                                    float64
Census Tract                                        float64
NTA                                                  object
dtype: object
```

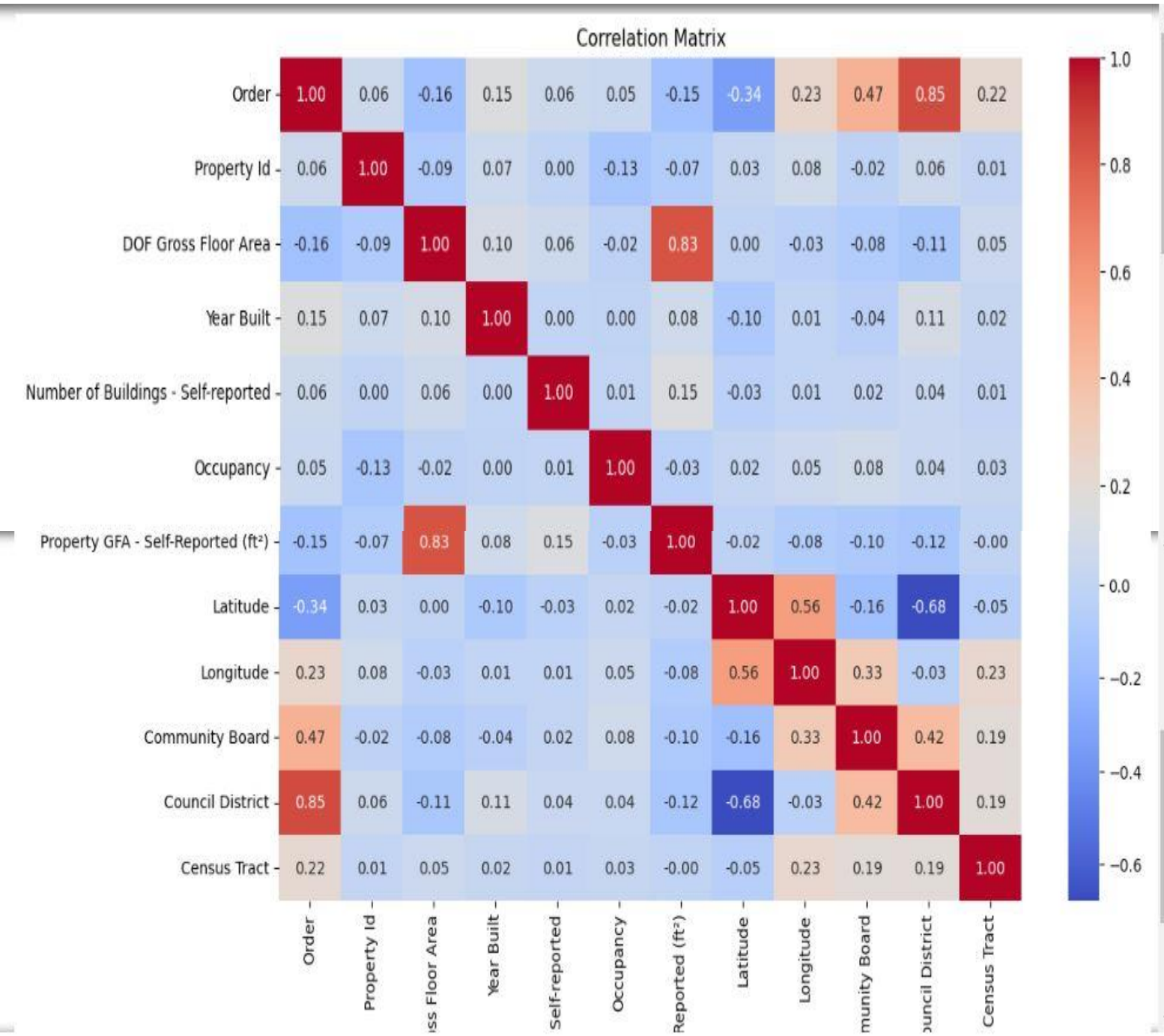
Insights

Dropping rows with missing values in specific columns may be necessary if those columns are critical for analysis. And Checking data types after preprocessing helps ensure that columns have the correct data types for further analysis.

Data Visualization and Analysis

```
# Exclude non-numeric columns from correlation calculation
numeric_columns = df.select_dtypes(include=[np.number])
corr_matrix = numeric_columns.corr()

plt.figure(figsize=(12, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix')
plt.show()
```



DOF Gr

Number of Buildings -

Property GFA - Self-

Com

C

Insights

Exploring correlations between variables helps identify relationships and dependencies within the dataset. And Visualizing correlations through heatmaps provides a quick overview of the data's interrelationships.

Preprocessing and Model Training

```
# Drop rows with missing values for the 'ENERGY STAR Score' column
df.dropna(subset=['ENERGY STAR Score'], inplace=True)

# Convert categorical variables to numeric using label encoding
label_encoder = LabelEncoder()
for column in df.select_dtypes(include=['object']).columns:
    df[column] = label_encoder.fit_transform(df[column].astype(str))
```

Insights

Handling missing values and encoding categorical variables are essential preprocessing steps in preparing data for machine learning models. And Label encoding converts categorical variables into numerical representations, which can be understood by machine learning algorithms.

```
# Define features (independent variables) and target variable
X = df.drop(columns=['ENERGY STAR Score'])
y = df['ENERGY STAR Score']

# Split data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize Random Forest Regressor model
rf_regressor = RandomForestRegressor(random_state=42)
```

Insights

Splitting data into training and testing sets allows for evaluating the model's performance on unseen data. And Initializing the machine learning model is a crucial step before training and evaluation.

Feature Engineering and Model Evaluation

```
# Import necessary libraries
import re

# Define the string
string_value = 'TWA:419 WEST 34'

# Extract the numerical part using regular expressions
numeric_part = re.findall(r'\d+\.*\d*', string_value)

# Convert the extracted numerical part to float
if numeric_part:
    numeric_value = float(numeric_part[0])
    print(numeric_value)
else:
    print("No numerical value found in the string.")
```

```
419.0
```

Insights

Regular expressions are powerful tools for pattern matching and string manipulation in Python. And Extracting relevant information from text data can be useful for feature engineering in machine learning tasks.

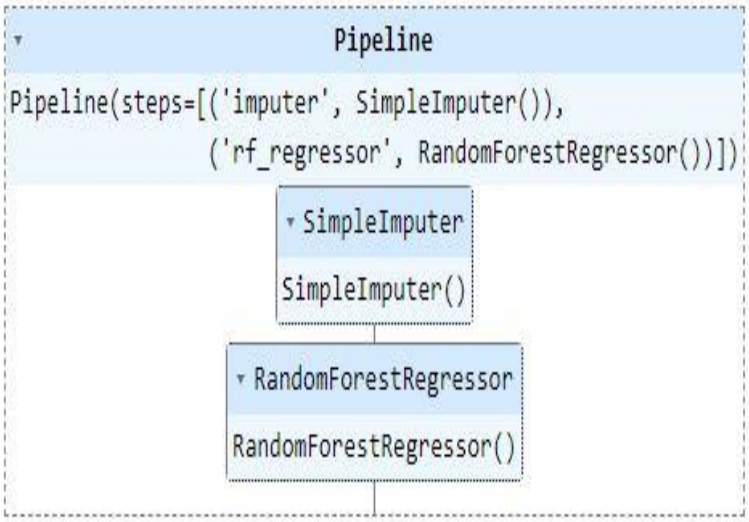
Defining and Training the Model

```
# Define the RandomForestRegressor
rf_regressor = RandomForestRegressor()

# Define the imputer to replace NaN values with the mean
imputer = SimpleImputer(strategy='mean')

# Define a pipeline to sequentially apply the imputer and then train the model
pipeline = Pipeline([
    ('imputer', imputer),
    ('rf_regressor', rf_regressor)
])

# Train the model using the pipeline
pipeline.fit(X_train, y_train)
```



Insights

Pipeline ensures uniform preprocessing and model training, enhancing workflow stability. RandomForestRegressor is robust for complex regression tasks, minimizing overfitting risks. Imputer effectively manages missing data by replacing with mean, maintaining dataset integrity. Training within pipeline simplifies the process, ensuring seamless reproducibility.

Test Set Preprocessing, Prediction, and Evaluation

```
# Preprocess the test set using the same imputer
X_test_preprocessed = imputer.transform(X_test)

# Predict on the preprocessed test set
y_pred = pipeline.predict(X_test_preprocessed)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f'Mean Squared Error: {mse}')
print(f'R-squared Score: {r2}')
```

```
Mean Squared Error: 601.6276559148937
R-squared Score: 0.4780823869858477
```

```
C:\ProgramData\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning: X does not have valid feature names, but SimpleImp
uter was fitted with feature names
  warnings.warn(
```

Insights

Consistent preprocessing techniques ensure reliable predictions on new data. And Model performance is quantitatively assessed using MSE and R2, aiding in understanding its predictive accuracy and generalization ability.

NasrinShaikh_Society_Score_ipy