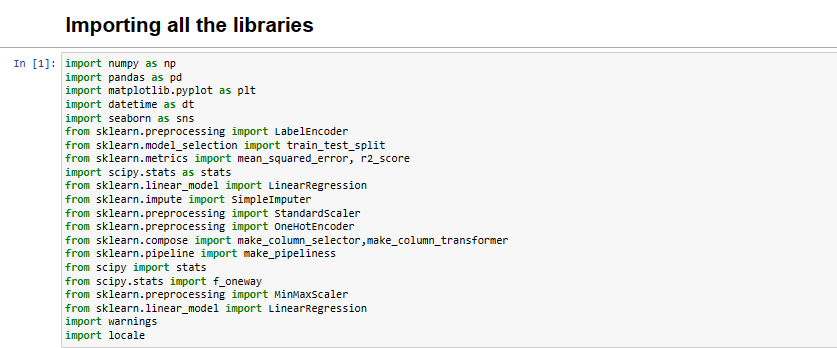
**BHARATINTERN TASK#1 – HOUSE PRICE PREDICTION**

**TASK#1 – HOUSE PRICE PREDICTION USING SIMPLE LINEAR REGRESSION(SLR)**

**Step1-** Importing All The required libraries of Python in Jupyter Notebook.

****

# Step2- Understanding the Dataset

**Features Included in the Dataset:** The dataset consists of various features or columns, each providing specific information about property sales. Common features may include property type (e.g., house, apartment), location details (suburb, postcode), property size (number of bedrooms, bathrooms), land size, historical sales prices, and potentially other attributes like proximity to amenities, school ratings, or economic indicators.

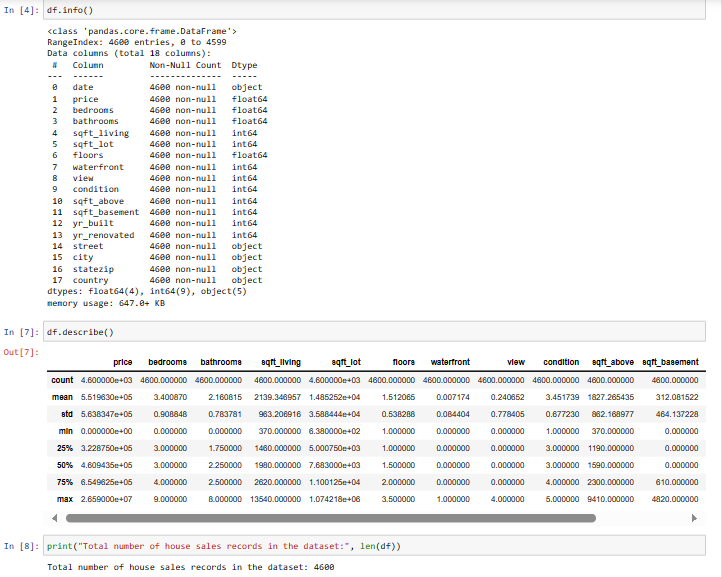
**Range of Values for Each Feature:** To gain insights into the dataset's characteristics, we will examine the range of values for each feature. This involves calculating the minimum, maximum, mean, and standard deviation for numeric features and listing unique categories for categorical features. This step helps identify any potential outliers or data inconsistencies.

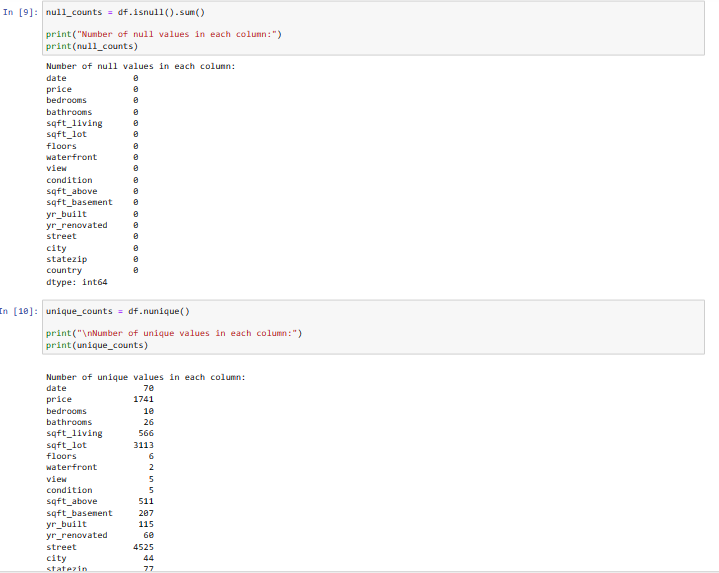
**Handling Missing Values and Outliers:** During our analysis, I also pay close attention to missing values and outliers. Missing data can be problematic, as it may affect the quality of our predictions. I will decide on an appropriate strategy for handling missing values, such as imputation or removal. Additionally, identifying and addressing outliers is essential, as they can distort our models' predictions. Outliers may be handled through techniques like trimming or transformation.

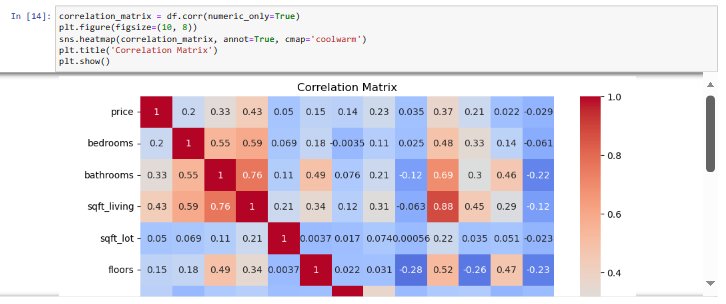
# Step3- Analysing Features in the Dataset

# Each feature plays a significant role in influencing property prices in the real estate markets of Sydney and Melbourne. By thoroughly examining these features, I aim to uncover valuable insights that will inform my predictive modeling efforts. I tried to explore the relationships between features, identify key variables that impact property prices, and assess the overall data quality. This analysis will serve as a foundation for data preprocessing and modeling steps, ensuring that I build robust and accurate predictive models for property prices.

# 

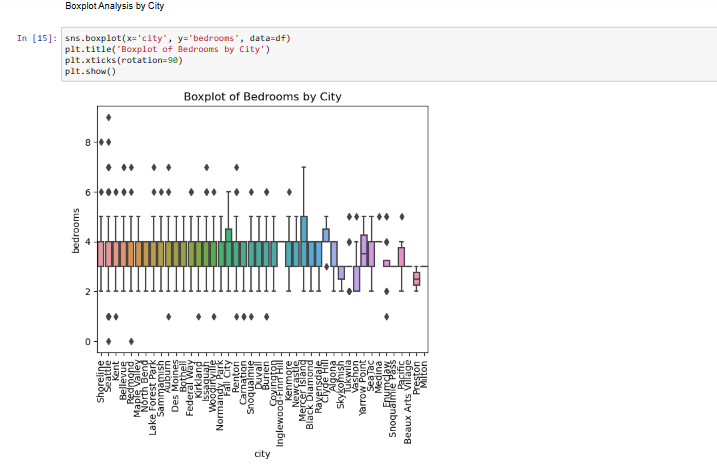






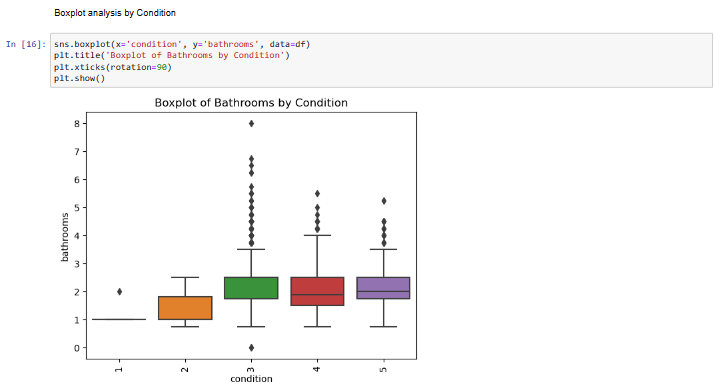
#### **Step4-** Boxplot Analysis by City

To gain insights into the distribution of the number of bedrooms in different cities, I created a boxplot. The boxplot provides a visual representation of the data's central tendency and spread. I plotted the number of bedrooms on the y-axis and the city on the x-axis. Each box represents a city's distribution of bedroom counts.



**Step5-** Boxplot Analysis by Condition

To gain insights into the distribution of the number of bathrooms in different conditions, I created a boxplot. The boxplot provides a visual representation of the data's central tendency and spread. I plotted the number of bathrooms on the y-axis and the conditions on the x-axis. Each box represents a condition's distribution of bathroom counts.

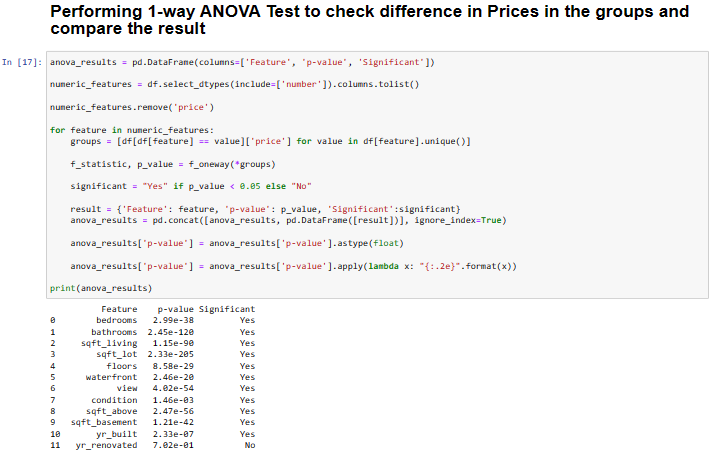


# Step6- Perform one-way ANOVA for Each Numeric Feature

In this code, one-way Analysis of Variance (ANOVA) is performed for each numeric feature in the dataset to assess their significance in explaining variations in the 'price' variable. The code follows these steps:

1. **Grouping Data:** For each numeric feature, the data is grouped based on unique values of that feature. This forms distinct groups for analysis.
2. **ANOVA Test:** One-way ANOVA is conducted on the grouped data to determine if there are statistically significant differences in 'price' among the groups.
3. **Significance Assessment:** The code checks if the p-value resulting from the ANOVA test is less than the significance level (alpha) of 0.05. If it is, the feature is marked as "Significant"; otherwise, it is marked as "Not Significant."
4. **Result Storage:** The results, including the feature name, p-value, and significance assessment, are stored in a DataFrame called 'anova\_results.'
5. **P-value Formatting:** To enhance readability, the p-values are formatted in scientific notation.
6. **Displaying Results:** The final results, including feature names, p-values, and significance assessments, are displayed.

This analysis helps identify which numeric features have a statistically significant impact on the 'price,' which can be valuable for understanding feature importance in predicting house prices.



# Step7- Visualizing Significant Features vs. Price

1. **Filtering Significant Features:**
   * I begin by filtering the dataset to select only those rows where the 'Significant' column is marked as 'Yes.' This step ensures that I focus only on features with statistical significance in relation to the 'price' variable.
2. **Extracting Feature Names:**
   * Next, I extract the names of these significant features from the filtered DataFrame. These feature names will be used to create individual plots.
3. **Subplot Configuration:**
   * I calculate the number of rows and columns needed to arrange the plots based on the number of significant features. The 'num\_cols' variable can be adjusted to control the number of columns in the layout.
4. **Creating Subplots:**
   * I create subplots within a single figure, arranging them in rows and columns. If there's only one row, the axes are reshaped accordingly.
5. **Y-Axis Label:**
   * A common y-axis label, 'Price,' is added to the left of the subplots for clarity.
6. **Creating Scatter Plots:**
   * For each significant feature, I create a scatter plot using Seaborn's sns.scatterplot. The x-axis represents the feature, and the y-axis represents the 'price.'
7. **Plot Titles and Labels:**
   * Titles and labels are set for each subplot to describe the feature being plotted and the 'Price' axis.
8. **Removing Empty Subplots:**
   * In case there are empty subplots (if the number of features is not a multiple of the number of columns), they are removed to keep the layout clean.
9. **Layout and Spacing:**
   * I make adjustments to the layout and spacing of the subplots to ensure they are well-organized and easy to read.
10. **Displaying the Plot:**
    * Finally, the plot is displayed using plt.show().

# 

# 

# Step8- Visualizing Data Distribution with Pie Charts

I utilize pie charts to visually represent the distribution of house sales by month and the distribution of houses based on their condition. The code provides a quick and intuitive way to understand the composition of data within these categories.

## **House Sales by Month:**

* The first pie chart displays the distribution of house sales by month.
* Each slice of the pie represents a month, and its size corresponds to the proportion of house sales that occurred in that month.
* The autopct parameter adds percentage labels to each slice, indicating the percentage of sales in each month.
* Different colors are used to distinguish between the months.
* A legend is provided in the upper left corner to label the months.
* The title of the pie chart is "Distribution of House Sales by Month."

## **Houses Based on Condition:**

* The second pie chart illustrates the distribution of houses based on their condition.
* Each slice of the pie represents a specific condition category (e.g., "Good," "Excellent," etc.).
* The size of each slice reflects the proportion of houses falling into that condition category.
* autopct is used to display the percentage of houses in each condition category.
* Distinct colors are employed to differentiate between the condition categories.
* A legend is included in the upper left corner to label the condition categories.
* The title of the pie chart is "Distribution of Houses based on Condition."

# 

# 

## **Step9-** Scatter Plot of Price vs. Square Feet of Living Room

A scatter plot is created to visualize the relationship between the prices of houses and the square footage of their living rooms. The code serves to provide a visual representation of how house prices vary based on the size of the living room.

### **Scatter Plot:**

* The scatter plot is generated with 'sqft\_living' on the x-axis (representing square feet of living space) and 'price' on the y-axis (representing the price of houses).
* Each data point in the plot corresponds to a specific house, where the x-coordinate represents the size of the living room, and the y-coordinate represents the price.
* This type of plot is used to observe patterns and relationships between two continuous variables.

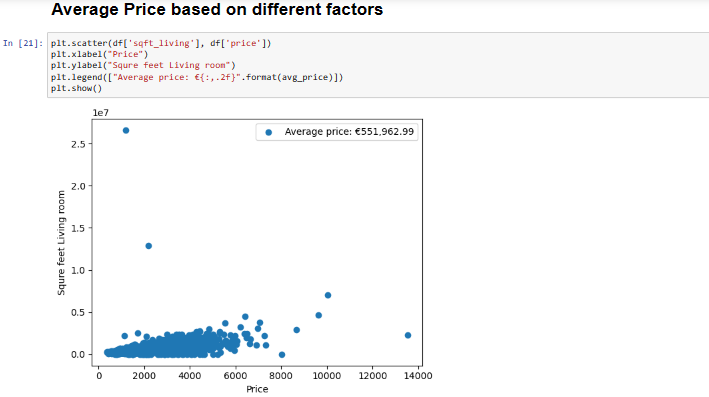
### **Legend:**

* The code adds a legend to the plot, which includes information about the average price of houses in the dataset.
* The legend displays the average price with proper formatting.

### **Axis Labels:**

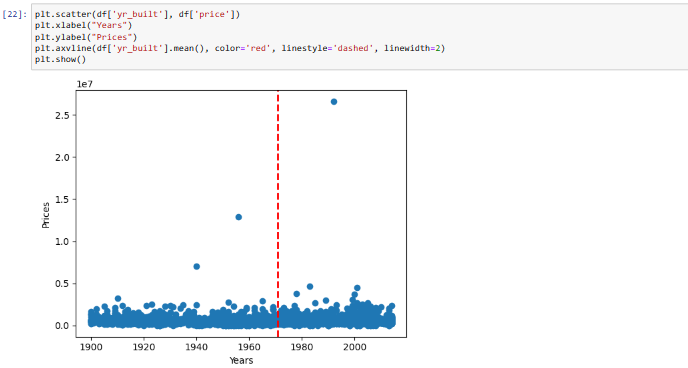
* Axis labels are provided to indicate the variables being plotted. The x-axis is labeled as "Price," and the y-axis is labeled as "Square feet Living room."

This scatter plot allows for a quick visual assessment of how the size of the living room relates to house price



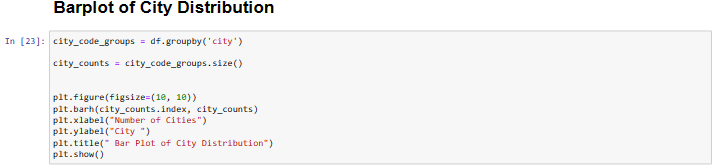
### **Step10-** Scatter Plot: Prices vs. Years

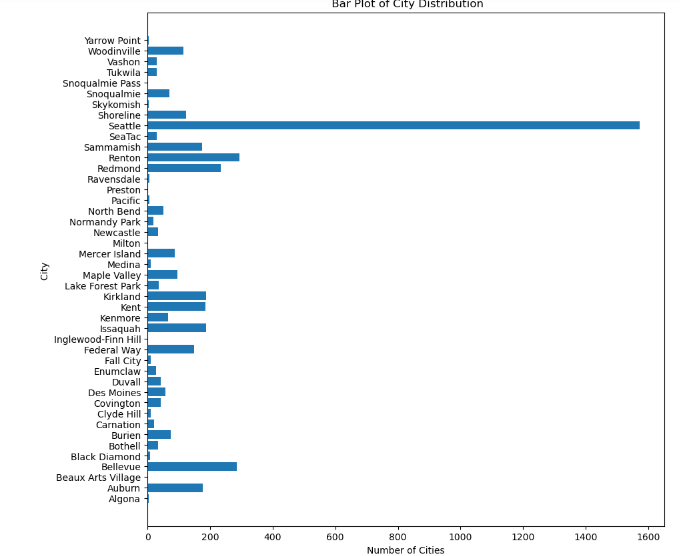
A scatter plot is created to visualize the relationship between house prices and the year they were built. Each data point on the plot represents a specific house, with the x-axis showing the construction year and the y-axis displaying the house prices. A red dashed line marks the average construction year for reference.



### **Step11-** Bar Plot: City Distribution

A bar plot is generated to visualize the distribution of cities in the dataset. Each horizontal bar represents a city, and the length of the bar corresponds to the number of occurrences of that city in the data. The x-axis displays the number of cities, while the y-axis lists the city names. This bar plot allows for a quick assessment of how data is distributed across different cities, helping to identify the prevalence of each city in the dataset.





# Step12- Feature Significance Analysis

I identified several features that do not significantly contribute to predicting home prices. These features include 'date,' 'yr\_renovated,' 'street,' 'city,' 'statezip,' and 'country.'

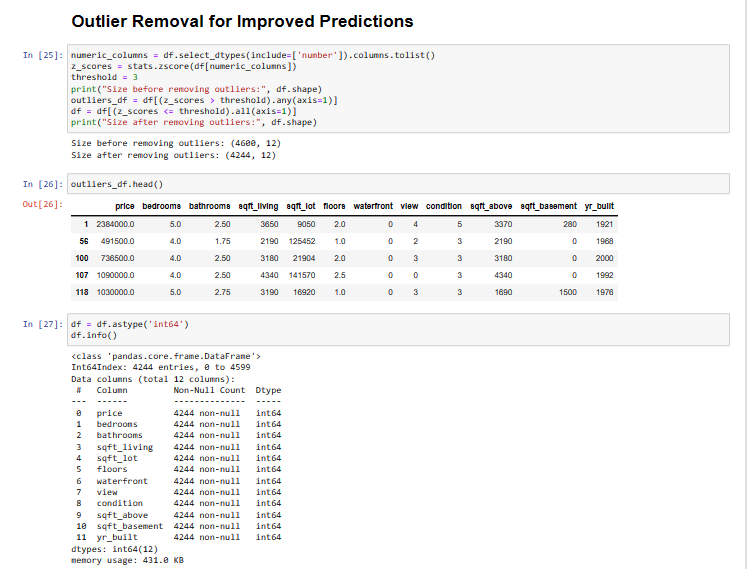
1. **Date:** The 'date' feature, which likely represents the date of property sales, may not be a strong predictor of home prices. While it could be relevant for time-series analysis or understanding market trends over time, its direct impact on individual home prices might be limited.
2. **Year of Renovation (yr\_renovated):** 'yr\_renovated' represents the year when a property was renovated. In cases where many properties have not undergone renovation, this feature may not significantly affect home prices. Additionally, it's often overshadowed by other more crucial factors, such as location and property size.
3. **Street, City, Statezip, and Country:** Categorical features like 'street,' 'city,' 'statezip,' and 'country' can indeed influence home prices, but they can be challenging to use directly in predictive models. One-hot encoding or label encoding these features may lead to a high dimensionality problem and may not always yield substantial improvements in predictive accuracy.

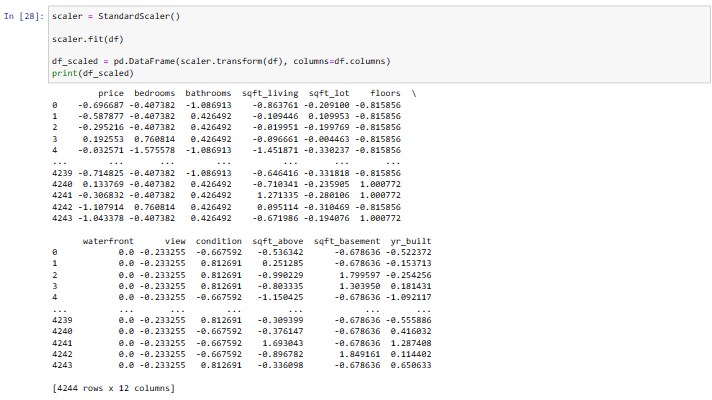
My decision to exclude these features from my predictive modeling is based on both statistical analysis and practical considerations. While every feature carries some degree of information, it's essential to focus on the most influential ones to build a concise and effective predictive model. By removing less significant features, I aim to streamline my model, reduce dimensionality, and improve its interpretability without compromising its predictive power.



**Step13-** Outlier Removal and Feature Scaling for Improved Predictions

 I employ outlier removal techniques to enhance the accuracy of my predictive model for home prices. Outliers, which are extreme data points, can distort the model's performance. Using Z-scores, I automatically identify and remove outliers in all numeric columns, ensuring that the model trains on a more representative and reliable dataset. By doing so, I aim to improve the accuracy of my predictions in the Sydney and Melbourne real estate markets.





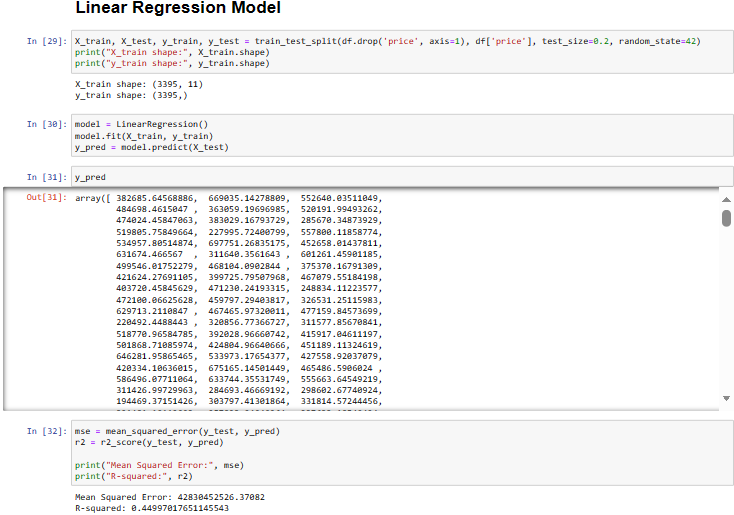
# Step14- Model Evaluation and Prediction using Model

# Models Evaluated: Linear Regression: A fundamental linear modelling approach that assumes a linear relationship between input features and the target variable.

## Evaluation Metrics**:**

For the model, I calculate two key metrics:

* **Mean Squared Error (MSE)**: Measures the average squared difference between predicted and actual values. Lower MSE indicates better predictive accuracy.
* **R-squared (R^2) Score**: Measures the proportion of variance in the target variable explained by the model. Higher R^2 scores indicate better explanatory power.



# Step15- Regression Model Visualization

I visually evaluate the performance of regression model. The scatter plots illustrate the predictions made by model against the actual target values.

## **Purpose:**

The purpose of these scatter plots is to provide a clear visual comparison of how well regression model predicts the target variable (property prices). By plotting the real values against the predicted values, I can quickly assess the models' accuracy and understand their predictive behaviour.

## **Interpretation:**

* Points in red represent the actual property prices in the dataset.
* Points marked with dots represent the predicted property prices made by model.
* The closeness of the dots to the red points indicates the accuracy of the model's predictions.
* A clear separation between the red and dotted points suggests good predictive performance.

By visually inspecting these scatter plots, we can identify patterns, trends, and potential outliers in the predictions. This visual assessment complements the quantitative evaluation using metrics such as Mean Squared Error and R-squared score.

