|  |
| --- |
| Daylit room looking at ocean |
|  |
| |  |  | | --- | --- | |  | **Project Work Summary** |   Deba |

**Property Price Prediction Capstone Project**

**By Debasis Baidya**

TABLE OF CONTENTS

**Content Pg No.**

1. **Introduction to the Project 02**

2. **Objectives of the Report 02**

3. **Flow Chart of Operations 03**

4. **Python Codes, Outputs, Screenshots | Data Loading 04**

5. **Outputs, Screenshots | Data Cleaning & Categorization 05**

6. **Outputs, Screenshots | Handling Missing Values 06**

7. **Outputs, Screenshots | Outliers Detection 07**

8. **Outputs, Screenshots | Outliers Visualization & Handling 08**

9. **Identification & Encoding of Ordinal & Nominal Columns 09**

10. **Model Training, Feature Scaling | PCA 10**

11. **Model Performance, Saving | Actual vs Predictions Diff. 11**

12. **Improving Model using XGBoost | Actual vs Predictions 12**

13. **Learning Outcomes | Conclusion | Web References 13**

**1. Introduction to the Project:**

This project aims to analyze an existing dataset, "Property\_data.csv," to uncover patterns and understand what factors influence property prices. I began by cleaning the data to correct errors, handle missing values, and prepare it for analysis.

With the data ready, I performed exploratory data analysis (EDA) to identify key features, like location, size, and age, that may impact property prices. Based on these findings, I created new features to improve the analysis.

Following the EDA, I tested various machine learning models, including linear regression and decision trees, to determine which one best predicts property prices. Once I chose the best model, I trained and evaluated it to ensure reliable performance. Metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) helped assess the model's accuracy.

In the end, this project produced a model capable of predicting property prices based on the given data, providing valuable insights into which factors most influence property values. This analysis can serve as a guide for understanding real estate trends and making informed pricing decisions.

**2. Objectives of the Report:**

1. **Data Cleaning and Preprocessing:** To prepare the property dataset for analysis by addressing missing values and duplicates, ensuring data quality and consistency for reliable results.
2. **Data Categorization:** To separate the dataset into categorical and numerical variables, enabling a tailored approach to data analysis and visualization for more accurate insights.
3. **Exploratory Data Analysis (EDA):** To examine key features of the property data, including trends, distributions, and relationships, with the goal of understanding factors that impact property prices. This includes identifying outliers and assessing their potential influence.
4. **Feature Engineering:** To enhance predictive analysis by creating new features derived from existing data, leveraging domain knowledge and statistical techniques to improve model accuracy.
5. **Model Selection and Evaluation:** To test various machine learning algorithms (e.g., linear regression, decision trees) and identify the best model for predicting property prices. Model performance is evaluated through metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).
6. **Data Visualization for Insights:** To use graphical methods to make complex data patterns more understandable, aiding in the interpretation of property features and their influence on price.
7. **Deployment of Predictive Model:** To deploy a final, optimized model that can accurately predict property prices for new data entries, providing practical value for real estate analysis and decision-making.

**3. Flow Chart of Operations:**

The project follows this sequence:

1. Data Loading

Import libraries (`pandas`, `numpy`, `seaborn`, `matplotlib`) and load the property dataset ("Property\_data.csv") into a DataFrame for initial review.

2. Data Cleaning and Preprocessing

Inspect the dataset for missing values and duplicates.

Handle missing values by filling numerical columns with the mean and categorical columns with the most frequent category (mode).

Address unique column issues by filling specific values with median or mode, ensuring data integrity.

3. Data Categorization

Separate data into categorical and numerical columns for more focused analysis and processing.

4. Exploratory Data Analysis (EDA)

Calculate descriptive statistics to understand data trends and identify potential outliers.

Use the Interquartile Range (IQR) method to detect and analyze outliers in numerical columns.

5. Feature Engineering

Create additional features based on identified trends and relationships to enhance the model’s predictive power.

6. Data Visualization

Generate box plots and other visualizations to identify outliers, understand distributions, and visualize relationships within the dataset.

7. Model Selection

Test multiple machine learning algorithms (e.g., linear regression, decision trees) to determine the best model for predicting property prices.

8. Model Training and Testing

Train the selected model on a subset of the data and test it on the remaining data to evaluate its accuracy and performance.

9. Model Evaluation

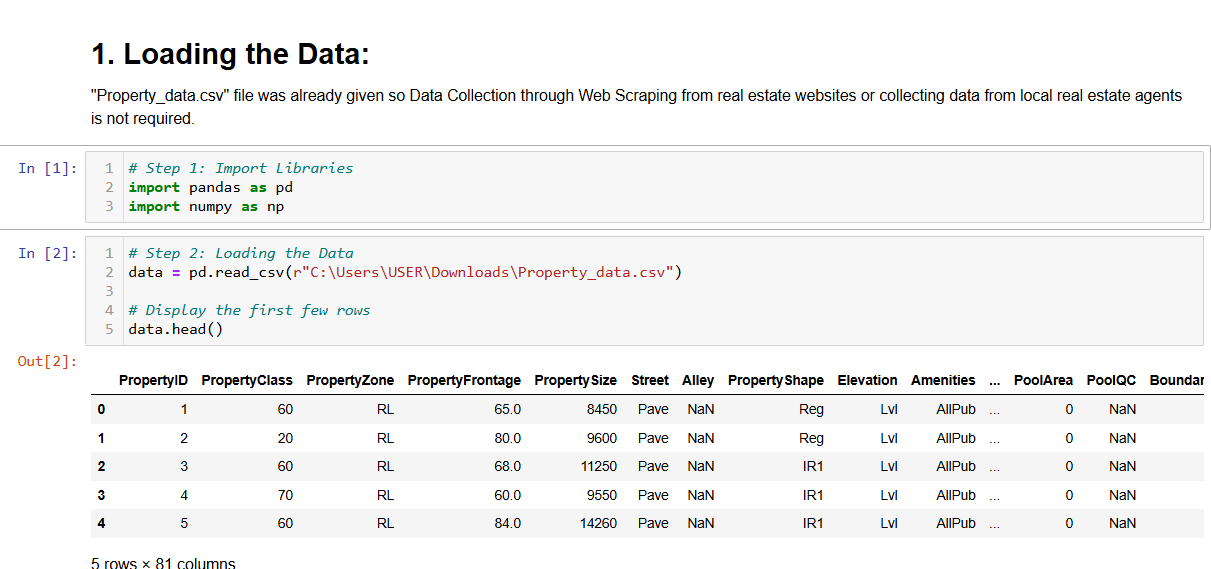
Evaluate the model’s performance using error metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to ensure reliability

10. Model Deployment

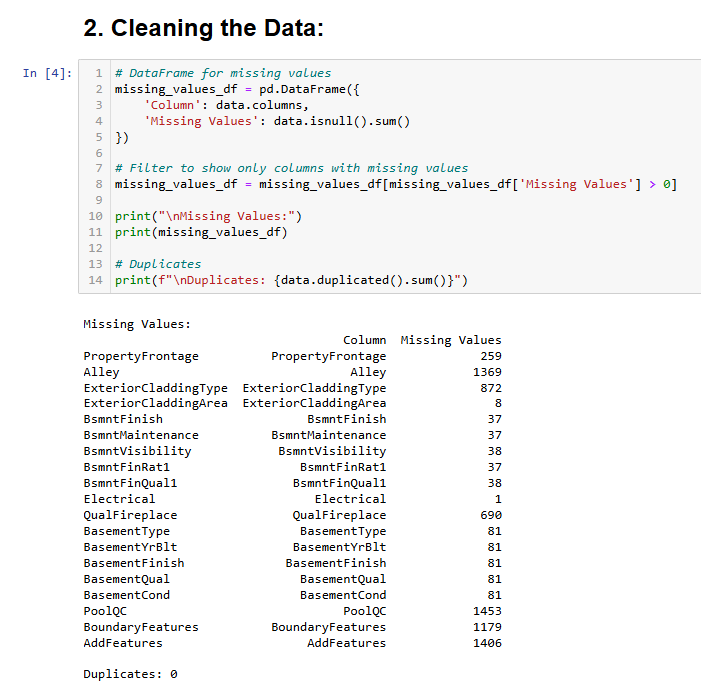
Deploy the final model to predict property prices on new data, providing a valuable tool for real estate price estimation.

**4. Python Codes with Screenshots of the Outputs & Report on EDA (with pictures of the graphs):**

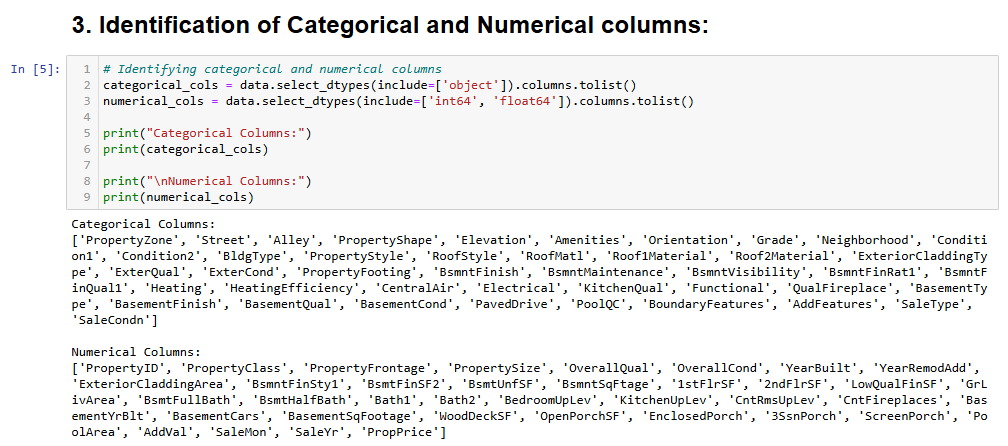
The project is implemented in Python. Here are snippets of key code sections:

**Data Loading:** Load the dataset using `pd.read\_csv()`.

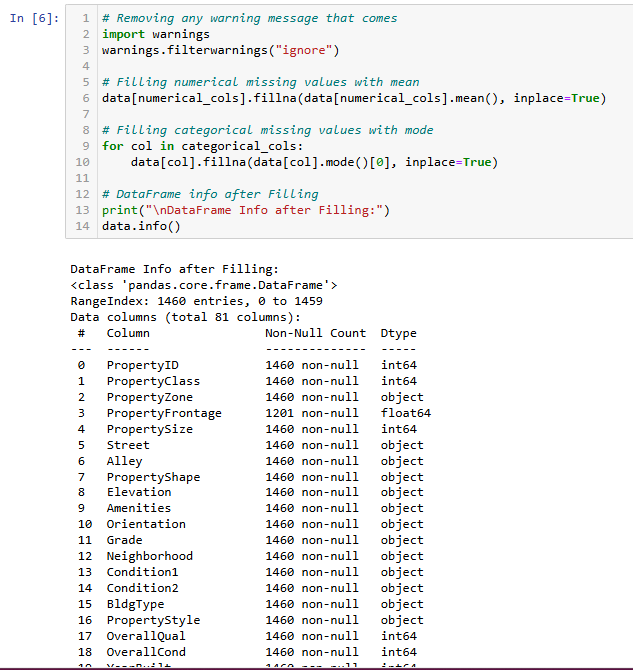
**Data Cleaning:** Identification of missing values and duplicates using DataFrame operations.



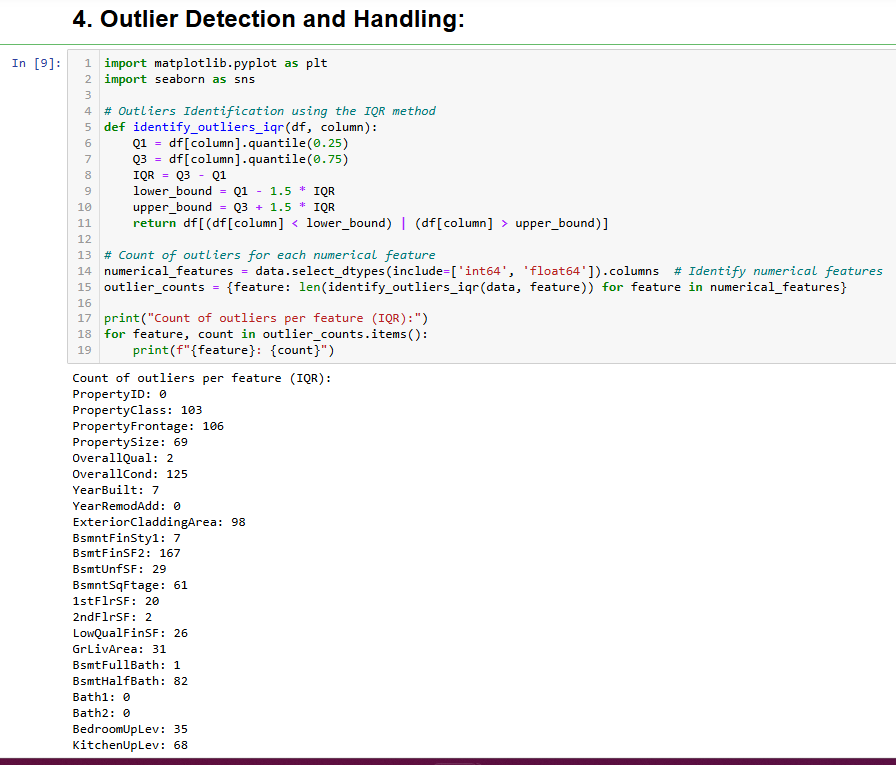
**Categorization:** Separate numerical and categorical columns.

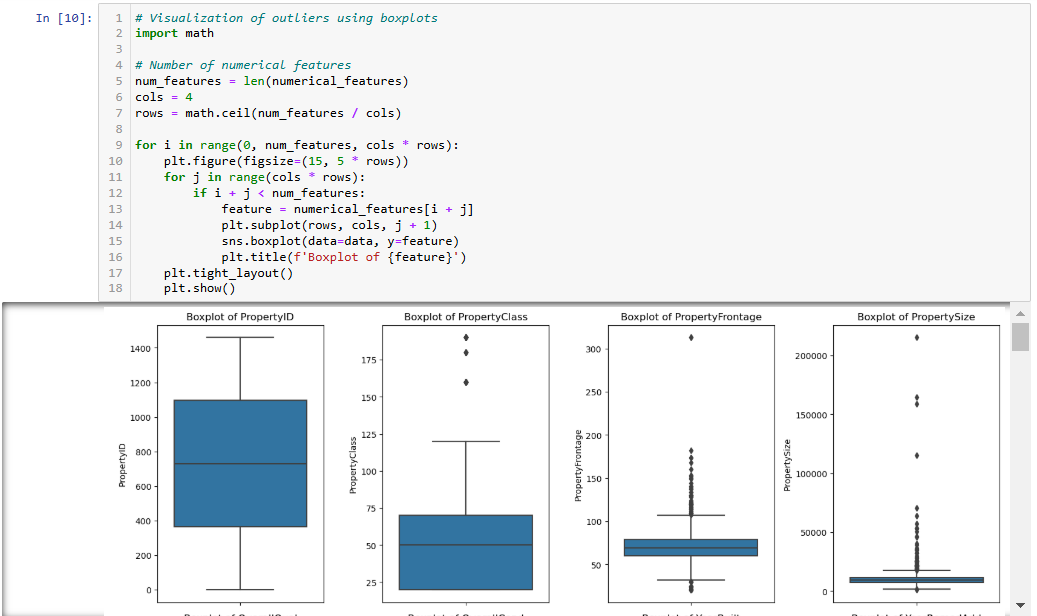
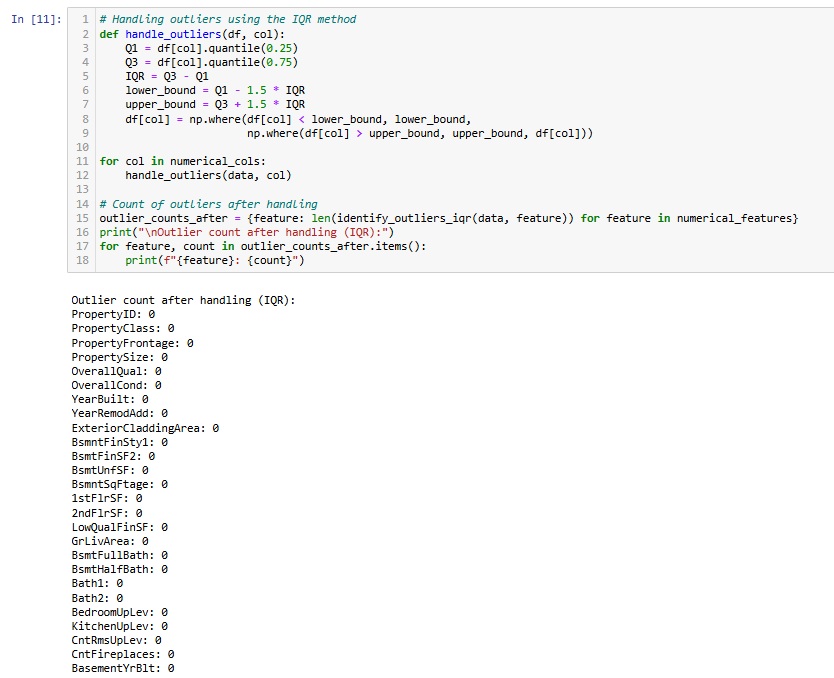


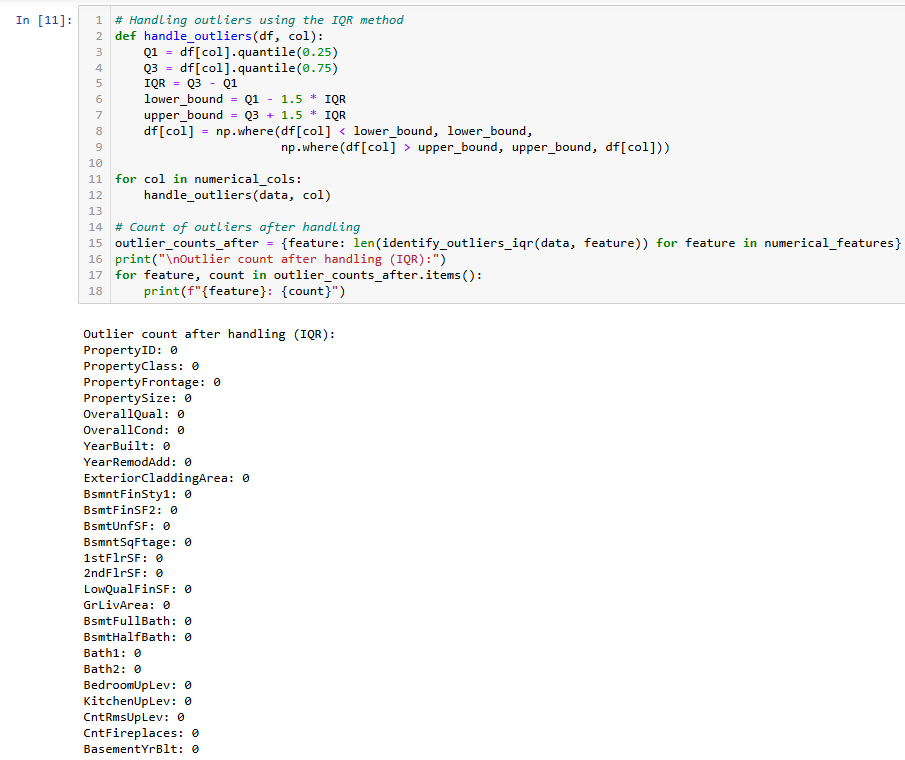
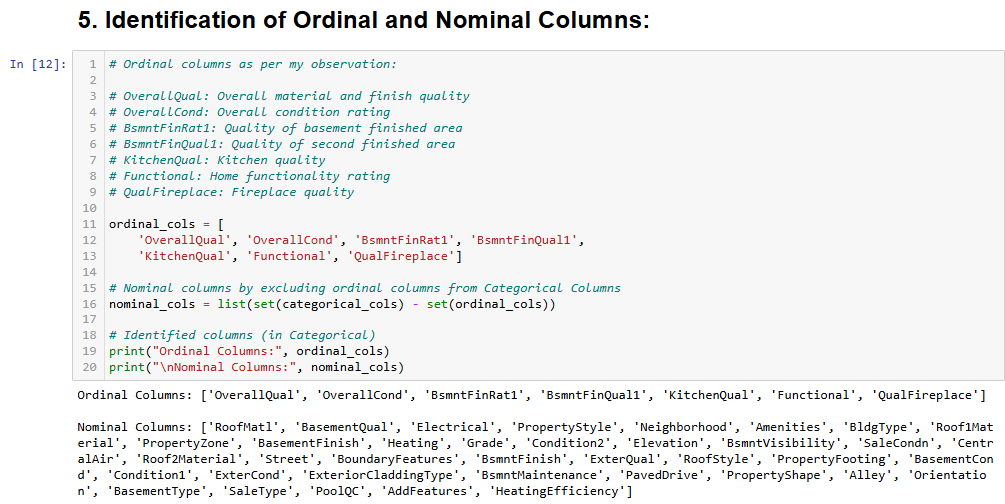
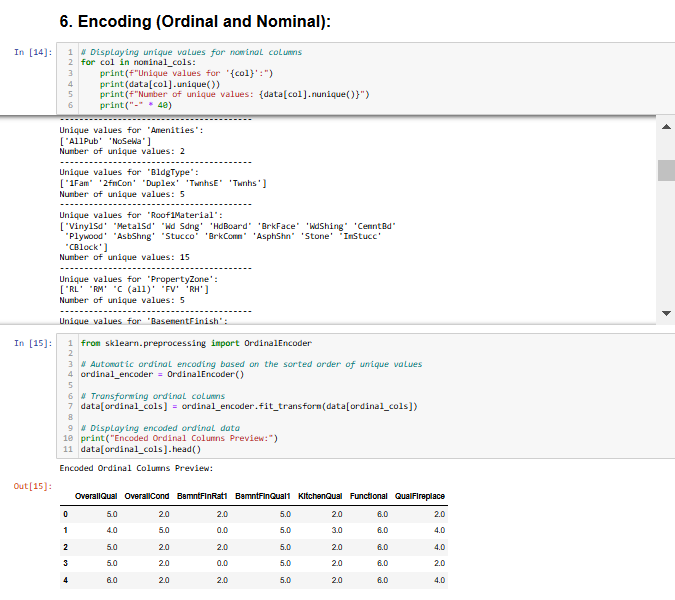
**Handling Missing Values:** Filling numerical missing values with mean & categorical missing values with mode

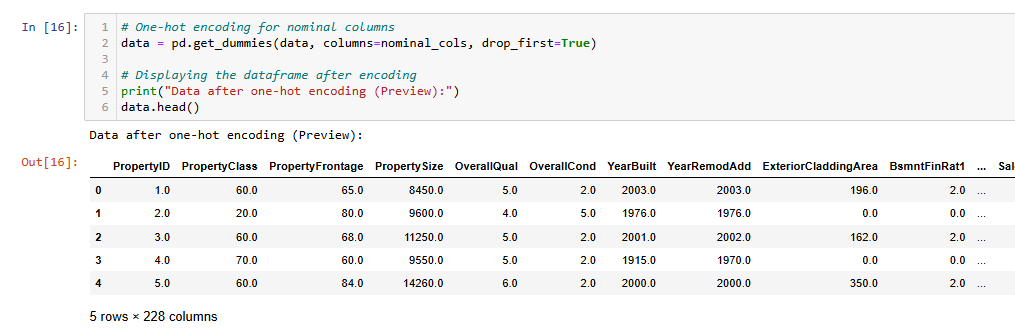


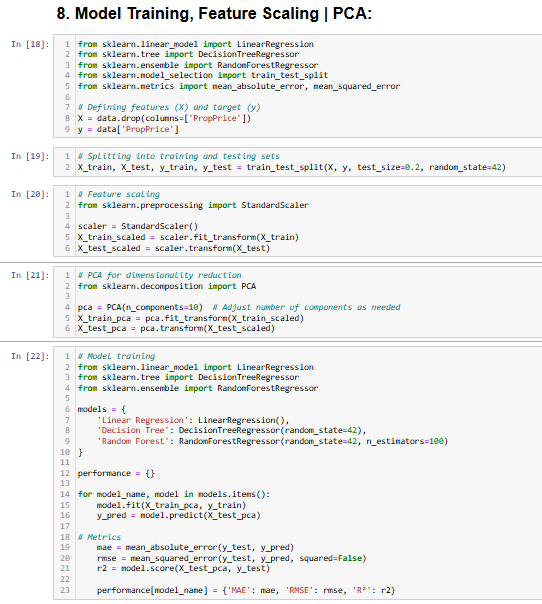
Outlier Detection: Using IQR & visualizing using Boxplots:

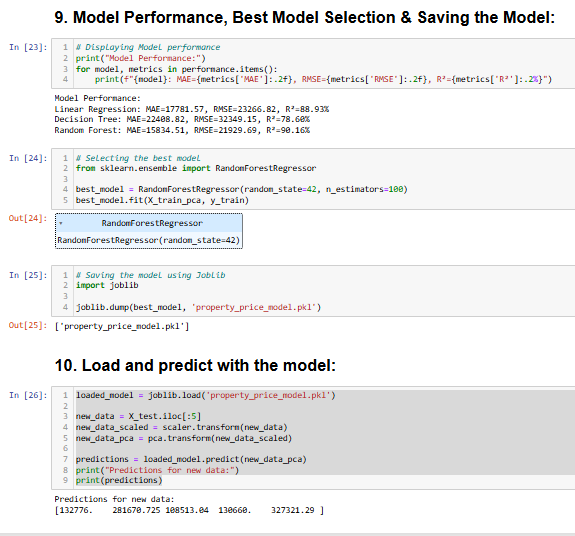


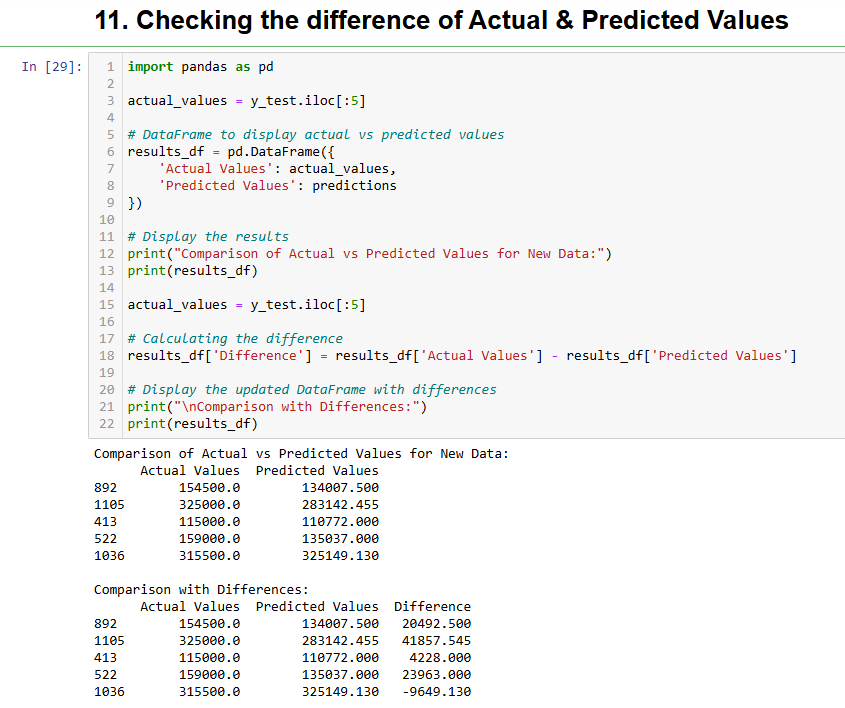
Outlier Visualization & Handling: Using Boxplots & IQR:

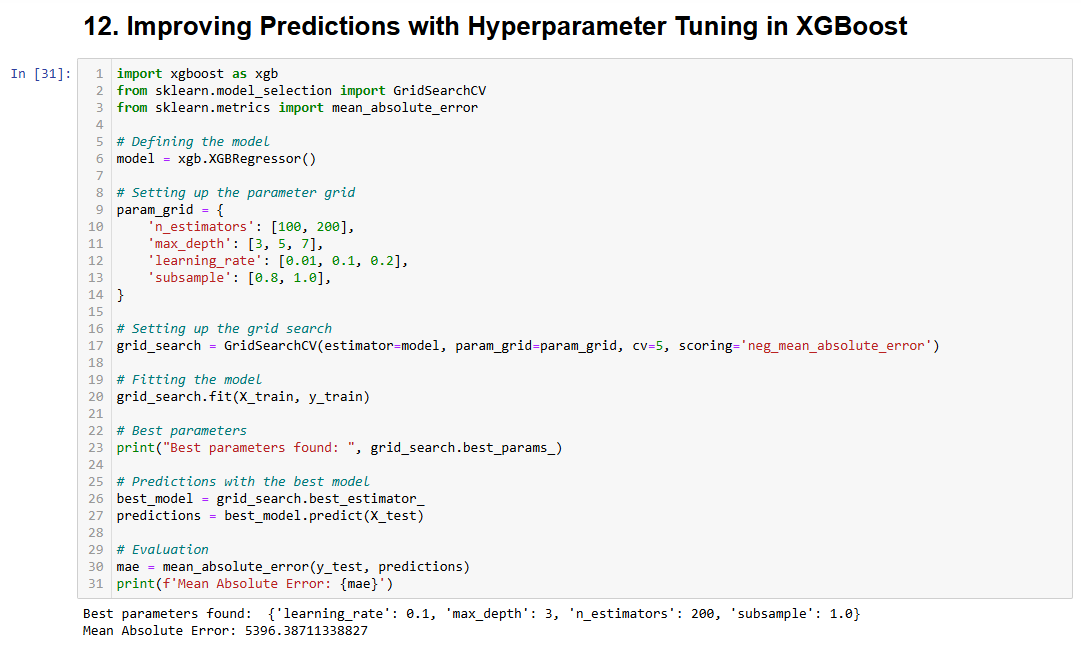
Identification & Encoding of Ordinal & Nominal Columns:

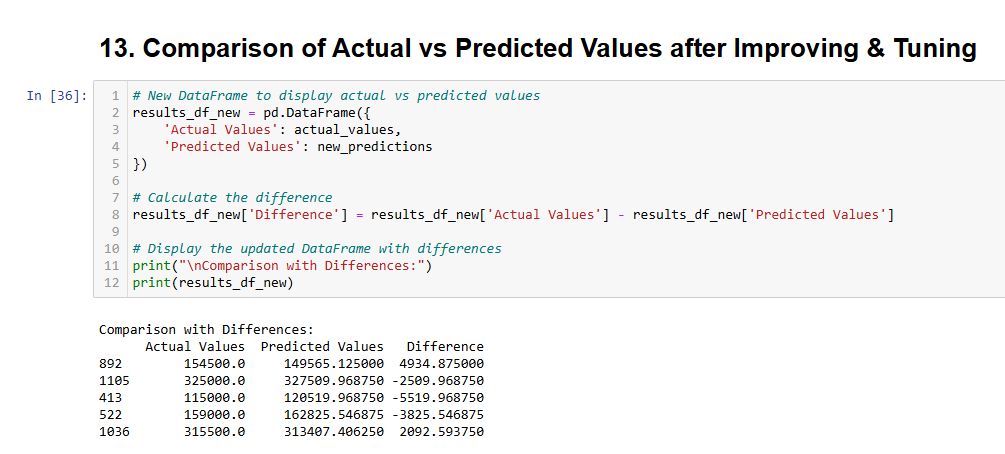


Model Training, Feature Scaling | PCA



****

****

****

**5. Learning Outcomes**

This project provided insights into:

Data preprocessing techniques, including missing value handling and data categorization.

Visualization skills for interpreting and presenting Outliers.

Analyzing real estate data, which offers practical knowledge of property data analysis.

Model Building, Performance Analysis & Choosing the best model, saving the model & making predictions

**6. Conclusion:**

This project analyzed the "Property\_data.csv" dataset to identify key factors influencing property prices. After cleaning the data and testing various machine learning models, I built a reliable prediction model.

To enhance accuracy, I applied XGBoost and fine-tuned the model, resulting in better predictions. Visualizations helped illustrate how well the model performed against actual values.

The final model serves as a practical tool for estimating property prices, benefiting real estate professionals and analysts. This work highlights the importance of thorough data preparation and advanced techniques in creating effective predictive models.

**7. Citations – Books / Websites Used for Research:**

For this project, I used a few online resources to guide my work:

* **W3Schools**: A great site for straightforward Python tutorials, W3Schools helped me brush up on Python basics and data handling, making my coding process smoother.
* **LearnPython.org**: This site offered practical exercises that reinforced Python concepts, especially useful for working with data structures in my analysis.
* **Library Documentation (Pandas, Matplotlib, Seaborn etc.)**: I relied on the official documentation for these libraries to understand specific functions and create accurate data visualizations and analyses.