**Capstone Project – 1**

**Predicting Property Prices in a Specific Location Using Machine Learning**



**Project Summary Report**

Submitted by:-

Aviral Singh

e-mail id – [avvi1001@gmail.com](mailto:avvi1001@gmail.com)

Table of Contents

1.Introduction to the Project

2.Objectives of the Project

3.Flow Chart of operations

4. Python Codes

5. Screenshot of the Outputs

6. Report on EDA (include pictures of the graphs)

7. Learning Outcomes

8. Conclusion

9. Citations – Books and Websites used for research.

Introduction to the Project

Real estate is one of the most dynamic and high-value sectors, where accurate property price prediction plays a crucial role for buyers, sellers, investors, and policymakers. Property prices are influenced by a wide range of factors such as location, property size, number of rooms, age of the building, proximity to essential services, and prevailing market trends. Traditional methods of price estimation often rely on limited data or expert intuition, which can lead to subjective and inconsistent results. With the availability of large datasets and advancements in data science, **Machine Learning (ML)** provides a powerful alternative for delivering more precise, data-driven property valuations.

This project focuses on predicting property prices in a specific location using machine learning techniques. The aim is to build a predictive model that can analyze historical property data, identify key features that influence prices, and provide accurate price estimates for new properties. The workflow involves several critical steps: data collection, pre-processing, exploratory data analysis (EDA), feature engineering, model selection, training, and evaluation. By comparing multiple machine learning algorithms—such as Linear Regression, Decision Trees, Random Forests, and Gradient Boosting—we seek to identify the most effective approach for this problem.

Beyond predicting property prices, the project also provides insights into the most significant factors driving price variations in the chosen location. These insights can help potential buyers assess fair market values, assist sellers in setting competitive prices, and guide investors in making informed decisions. In addition, the predictive framework developed here can be extended to other regions or scaled with larger datasets, making it a valuable tool in the broader real estate industry.

Objectives of the Project

1.To collect and clean real estate data from a specific location.

2.How can ordinal and nominal columns be handled separately in property price prediction using the metadata sheet provided to identify which columns are ordinal or nominal?

3.How can scaling, PCA, and fillna() techniques be used in property price prediction to handle missing data and improve the accuracy of the model?

4.To perform exploratory data analysis (EDA) on the collected data to identify key variables that influence property prices.

5.What is the proper encoding technique to be used for ordinal and nominal variables in property price prediction, based on the requirements of the model?

6.To develop a machine learning model that can predict property prices based on the selected variables.

7.To evaluate the performance of the model and compare it with other machine learning algorithms.

8.To present the findings and insights from the project in a clear and concise manner

Flow Chart of Operations

Conclusion

Model’s Performance Comparison

Exploratory Data Analysis

Model Creation and Evaluation

Data Pre-processing Part-2

Data Pre-Processing Part 1

Data collection

Data Exploration

Python Codes

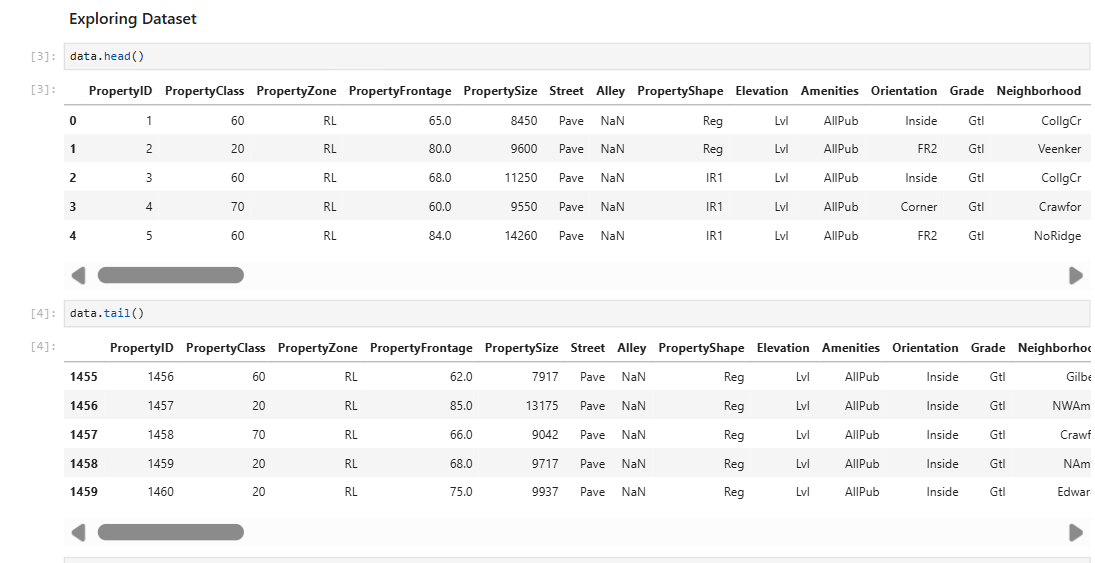
1. Importing Necessary libraries:

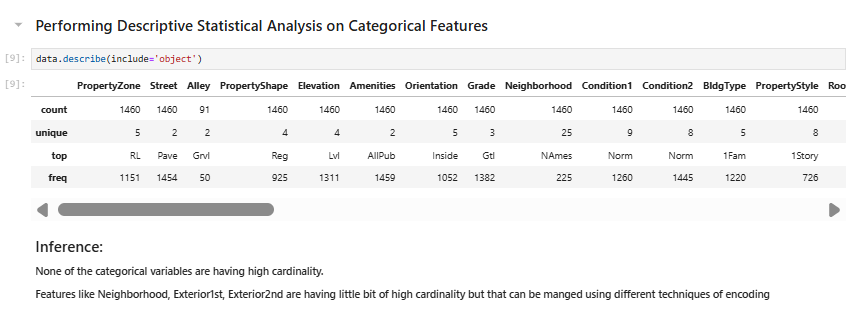
First we imported all the necessary libraries:

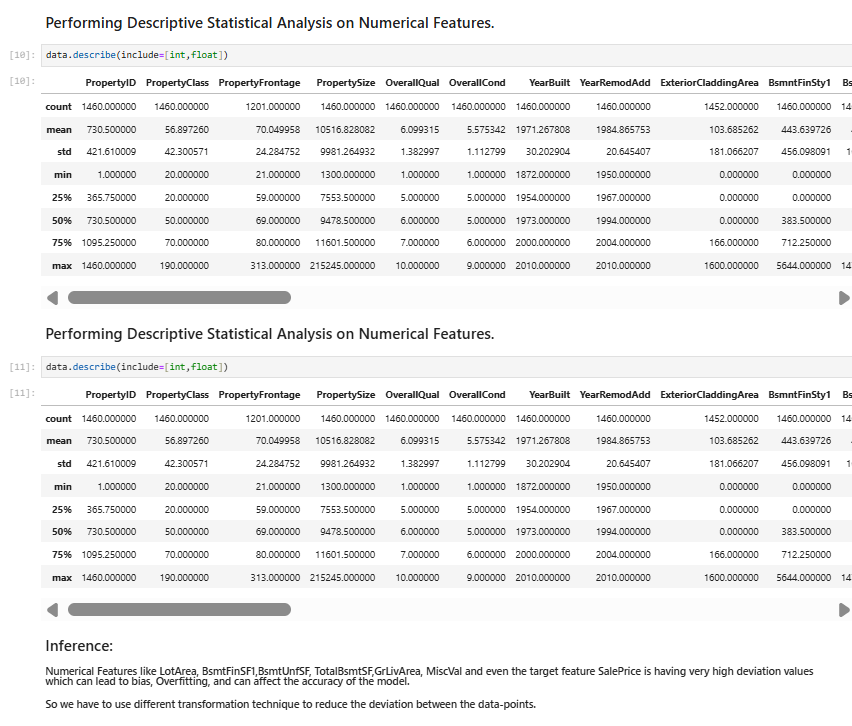


1. Data Collection:



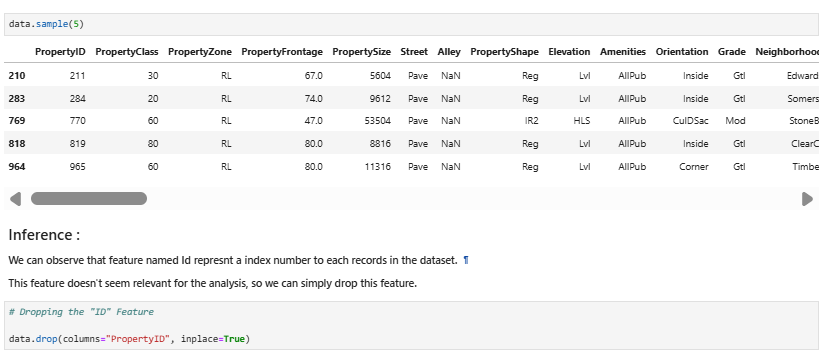
1. Exploring Dataset:



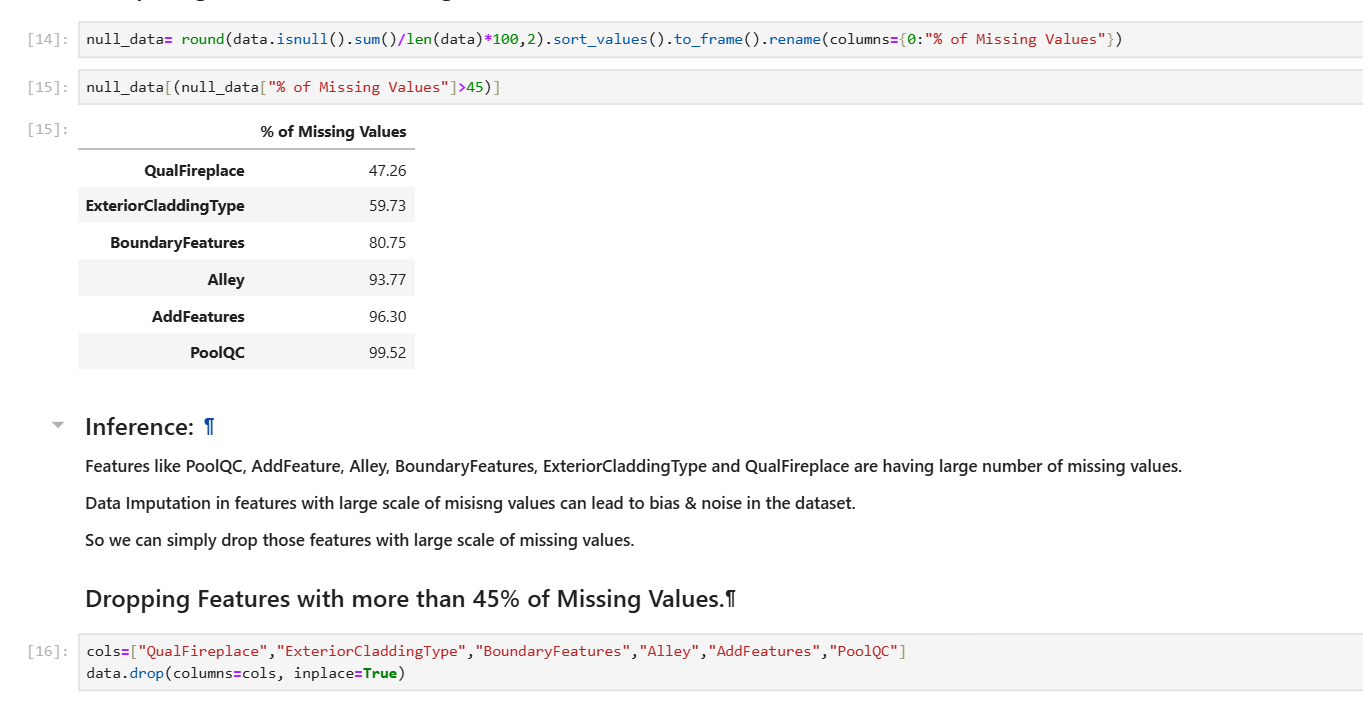


1. Data Pre-Processing Part-1:-

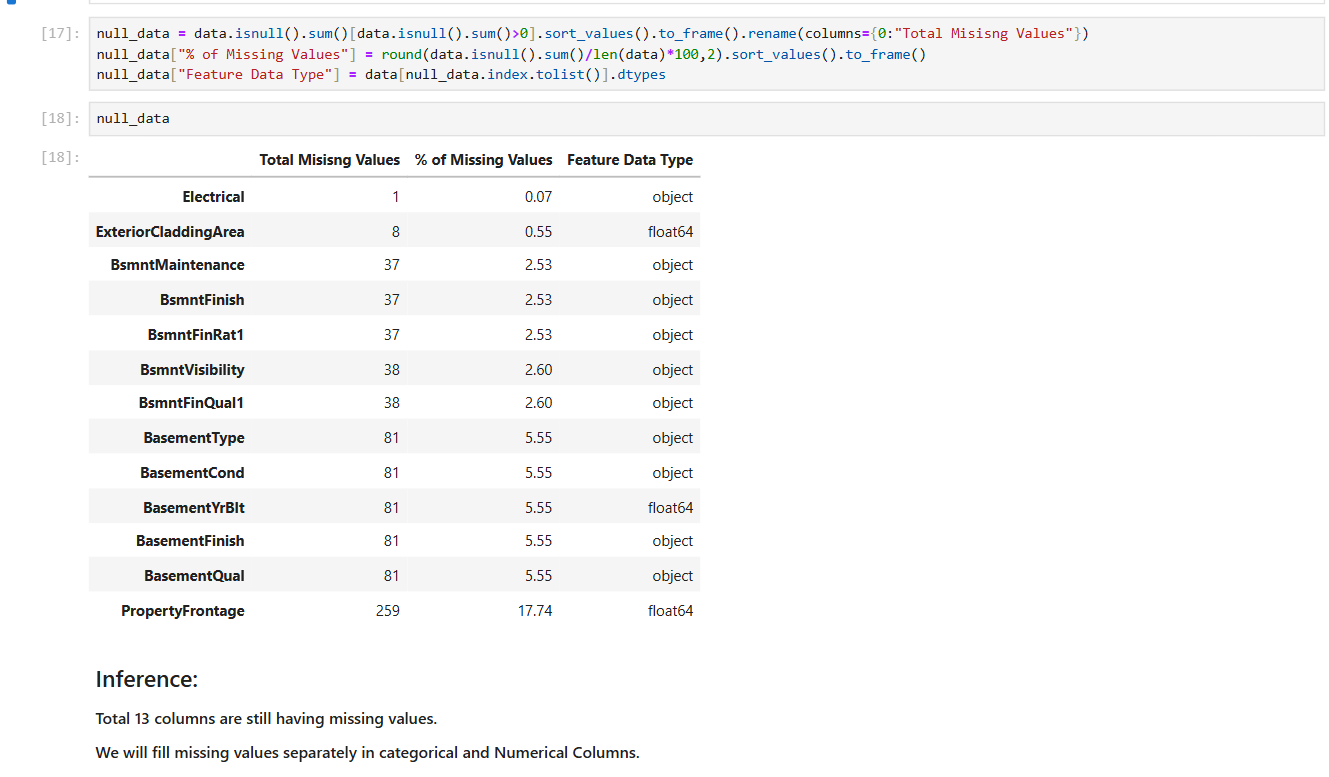
* Removing Unnecesary ID feature:



* Computing Features with Missing Values More than 45%:



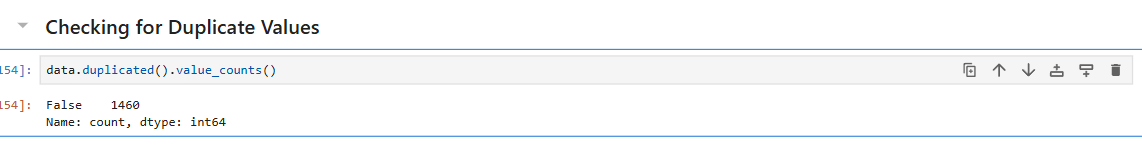
* Computing Total Missing Values and % of Missing Values:



* Filling Missing Values:

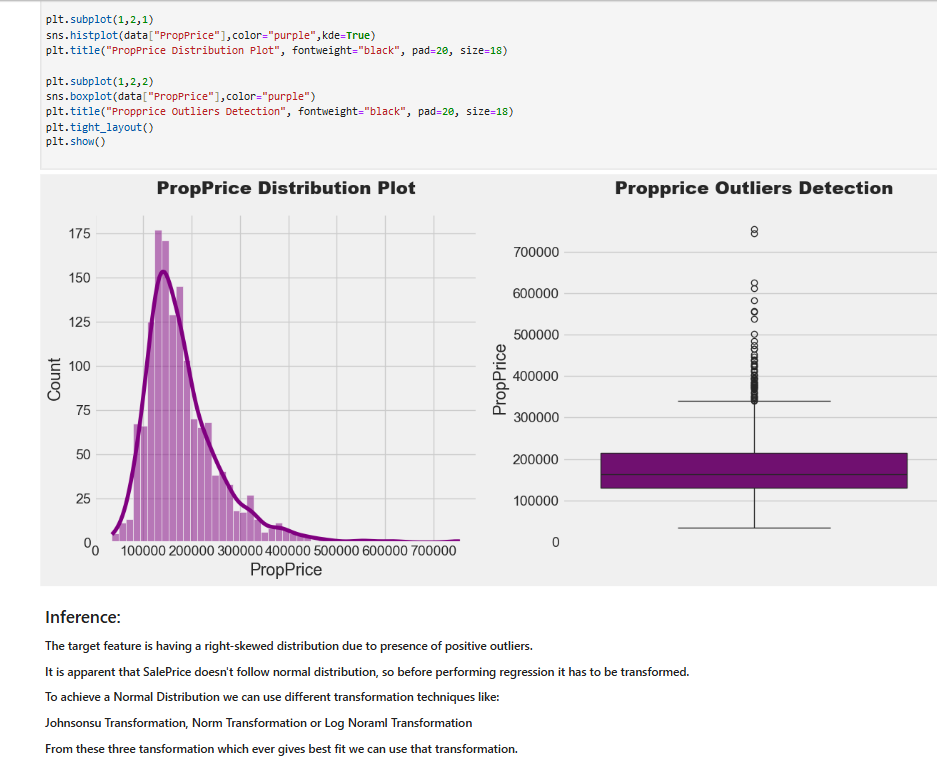


* Checking For Duplicate Values:



Exploratory Data Analysis

* Analysing and Visualising Target Variable:



### Visualizing Distribution of Continous Numerical Features

### 

### 

### 

### Visualizing the Skewness of Continous Numerical Features.

### 

### 

### Visualizing the Correlation of Continous Numerical Features

### Visualizing Categorical Features w.r.t SalePrice.

### 

### 

### 

### 

### 

### 

### Visualizing Discrete Numerical Features w.r.t Average "SalePrice".

### 

## **Feature Engineering**

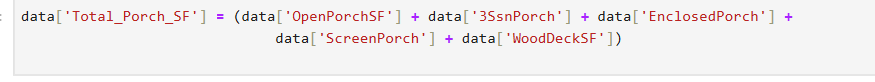
### 1.Creating Two New Features "RenovationStatus" and "AgeAtSale" of the House.

### 

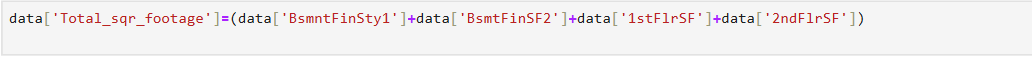
### 2. Creating a New Feature using all the columns storing "Bathroom Values."

### 

#### **3.** Creatng a New Feature using all the columns related to "porch".



### 4. Creating a New Feature Using "Sqaure Footage".



### Data Pre-Processing Part-2 (Different Encoding Techniques used)

### Performing Log Transformation on Target Variable

### 

### Applying Box-Cox Transformation on Continous Numerical Features to Reduce Skewness

### 

### 

### Performing Target Encoding on Categorical Features with high Cardinality

### 

### Performing Label Encoding on other features:

### 

### Model Creation and Evaluation

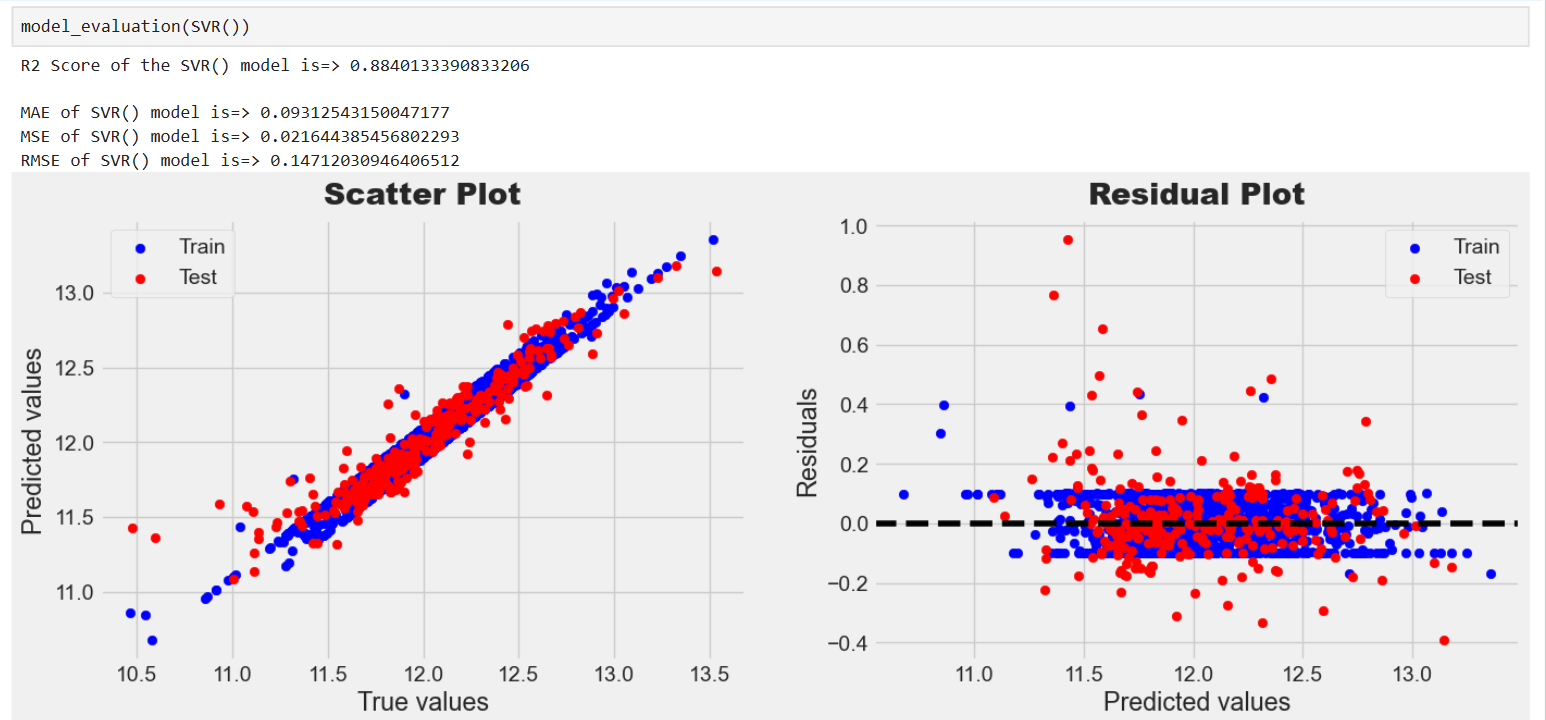
### Creating a Function to Train Model using different Regression Model

### 

### 1.Creating Linear Regression Model

### 

## 2. Creating Support vector Regressor Model.



### 3. Creating Random Forest Regressor Model

### 

### 4.Creating AdaBoost regressor Model

### 

### Creating Gradient Boosting Regressor Model

### 

### Creating XGBRegressor Model

### 

### 

### Model’s Performance Comparison

### 

### Conclusion

##### Features like 1stFlrSF,GrLivArea, 'OverallQual', 'CntRmsUpLev', 'BasementCars' and BasementSQFootage are having strong relation with the target variable.

##### The best performing model is GradientBoostRegressor with highest R2 and lowest MAE,MSE,RMSE values.

##### The second & third best performing model is LinearRegressor & XGBoostRegressor models.

##### The project developed a house price prediction model with strong performance metrics.

##### The project effectively addresses the task of house price prediction and contributes as a valuable tool in the dynamic real estate industry.