Car Resale Price prediction

SAT 5165 - Intro to Big Data Analytics

Bhavani Chalamalla
Haranadh Ravi

Contents

1. INTRODUCTION

Goal:

- To accurately predict the resale value of used cars by analyzing historical sales data using data science and machine learning techniques.

Significance:

- This project aims to revolutionize the used car market by providing accurate price predictions, benefiting both sellers and buyers.
- Enables more transparent and fair pricing, helping consumers make better-informed purchasing decisions.

Approach Overview:

- Data preprocessing, exploratory analysis, and applying regression models.

Tools & Techniques:

- Programming Language: Python
- Libraries: Pandas for data manipulation, Matplotlib and Seaborn for data visualization, scikit-learn for machine learning.
- Machine Learning Techniques: Regression algorithms (like Linear Regression, Random Forest,& Gradient Boosting).

2. Data Overview

	Brand	Price	Body	Mileage	EngineV	Engine Type	Registration	Year	Model
0	BMW	4200.0	sedan	277	2.0	Petrol	yes	1991	320
1	Mercedes-Benz	7900.0	van	427	2.9	Diesel	yes	1999	Sprinter 212
2	Mercedes-Benz	13300.0	sedan	358	5.0	Gas	yes	2003	S 500
3	Audi	23000.0	crossover	240	4.2	Petrol	yes	2007	Q7
4	Toyota	18300.0	crossover	120	2.0	Petrol	yes	2011	Rav 4

4345 total observations , 9 Columns (Features)

- 5 categorical variables
- 3 continuous variables.
- 1 response variable

Categorical variables

- Brand
- Body
- Engine Type

Continuous variables

- Mileage
- EngineV
- Year

Response Variable

- Price

3. Data Preprocessing

- We have missing values in 2 columns “Price” and “EngineV”. We have removed all null values as seen in Figure 2.

```
df.isna().sum()
```

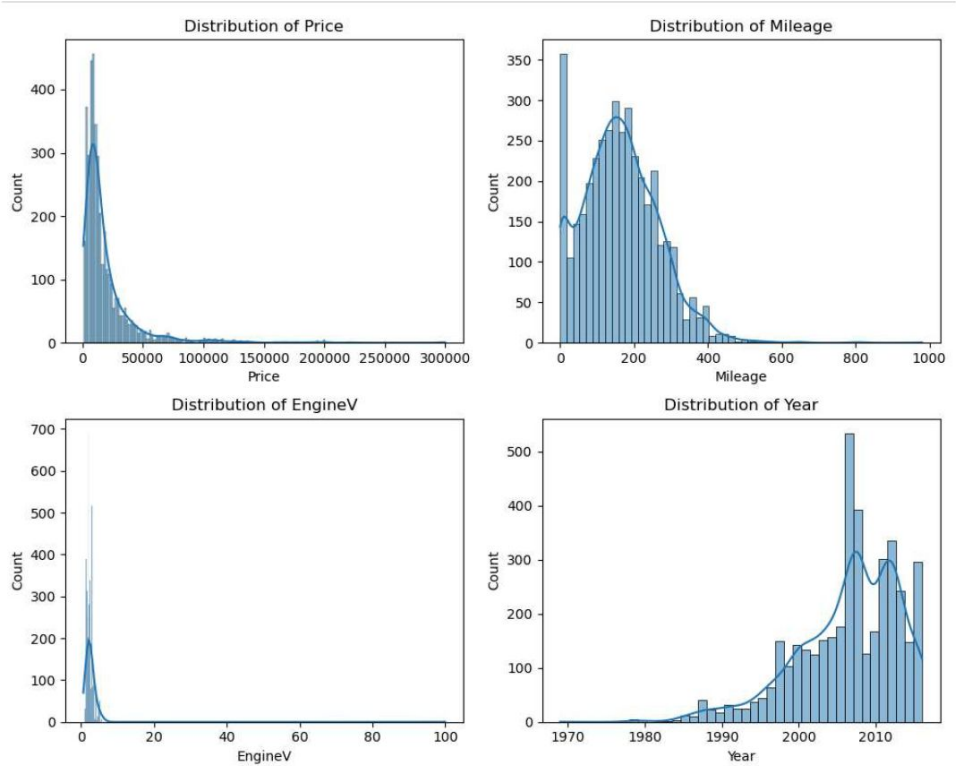
```
Brand             0
Price           172
Body              0
Mileage           0
EngineV         150
Engine Type       0
Registration      0
Year              0
dtype: int64
```

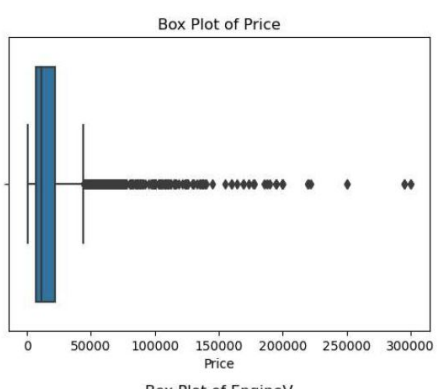
```
df=df.dropna()
```

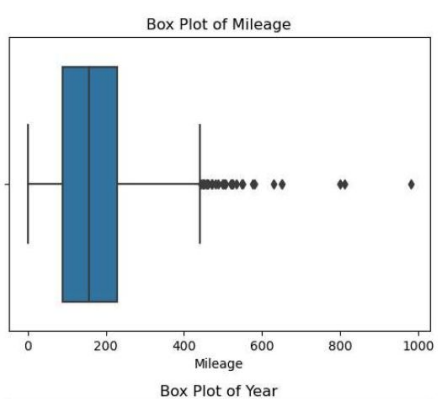
```
df.isna().sum()
```

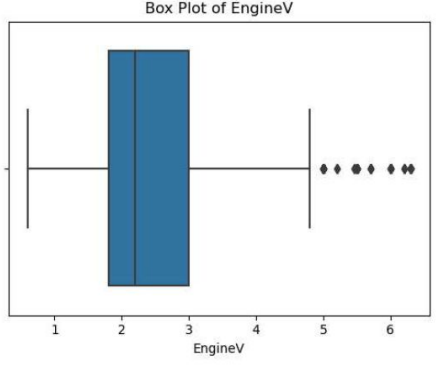
```
Brand             0
Price             0
Body              0
Mileage           0
EngineV           0
Engine Type       0
Registration      0
Year              0
dtype: int64
```

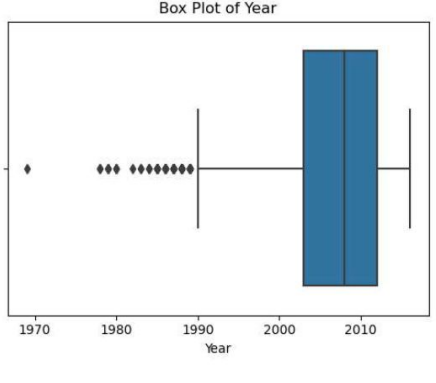
Distributions of Continuous Data

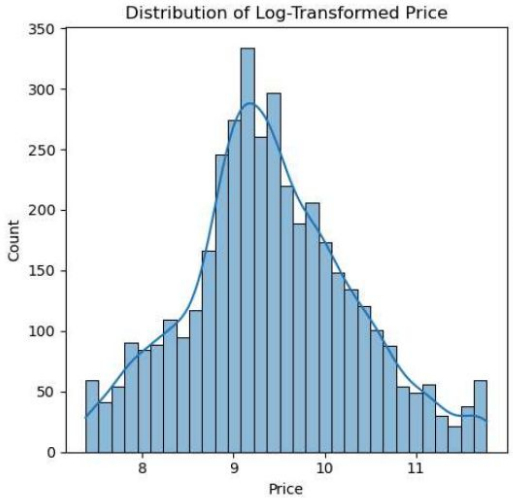


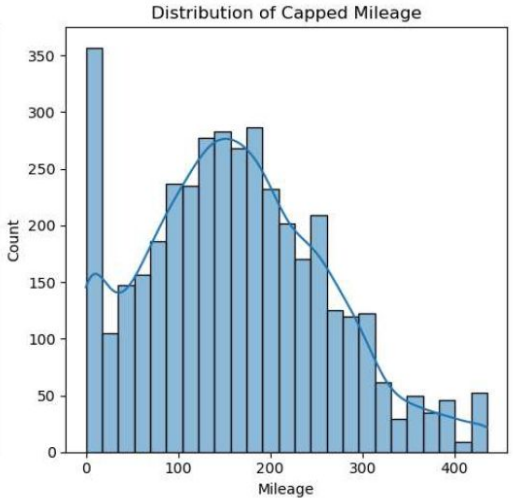


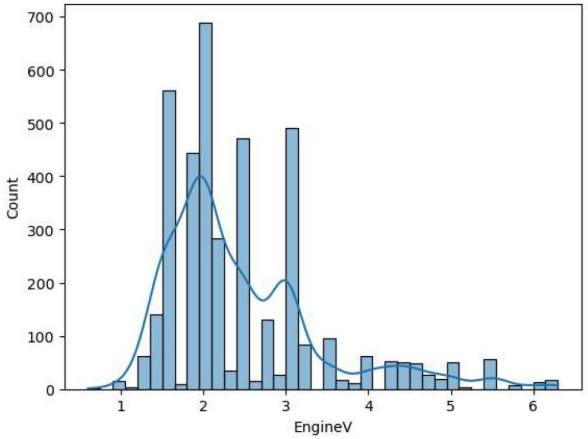


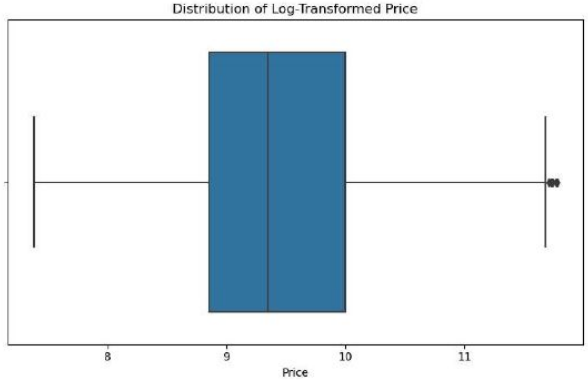


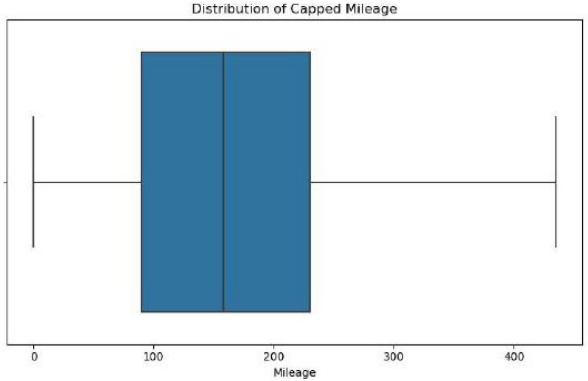








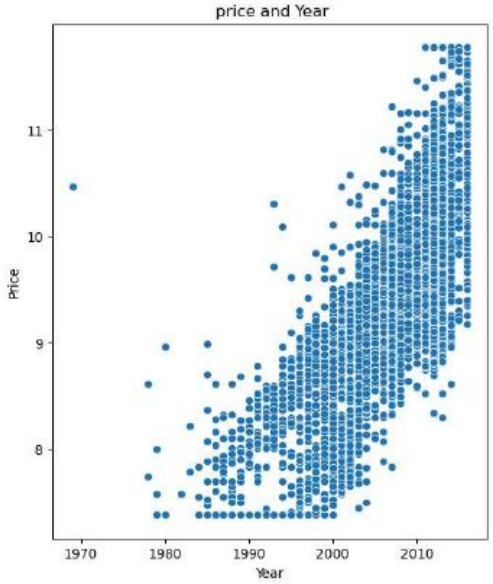


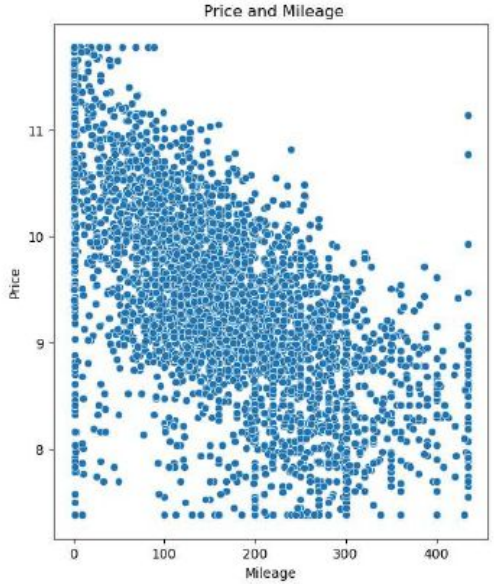


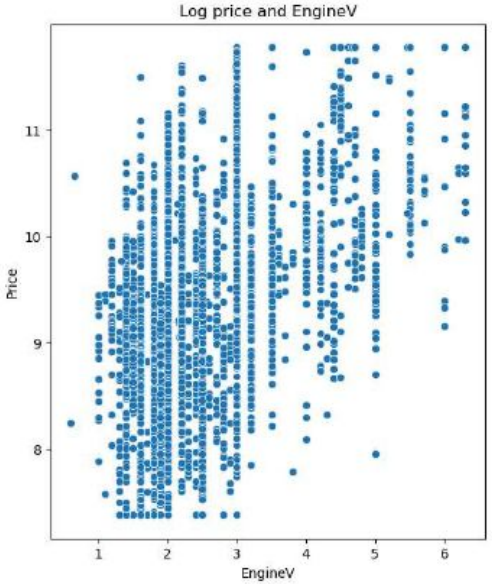
Continuous Data After Transformation

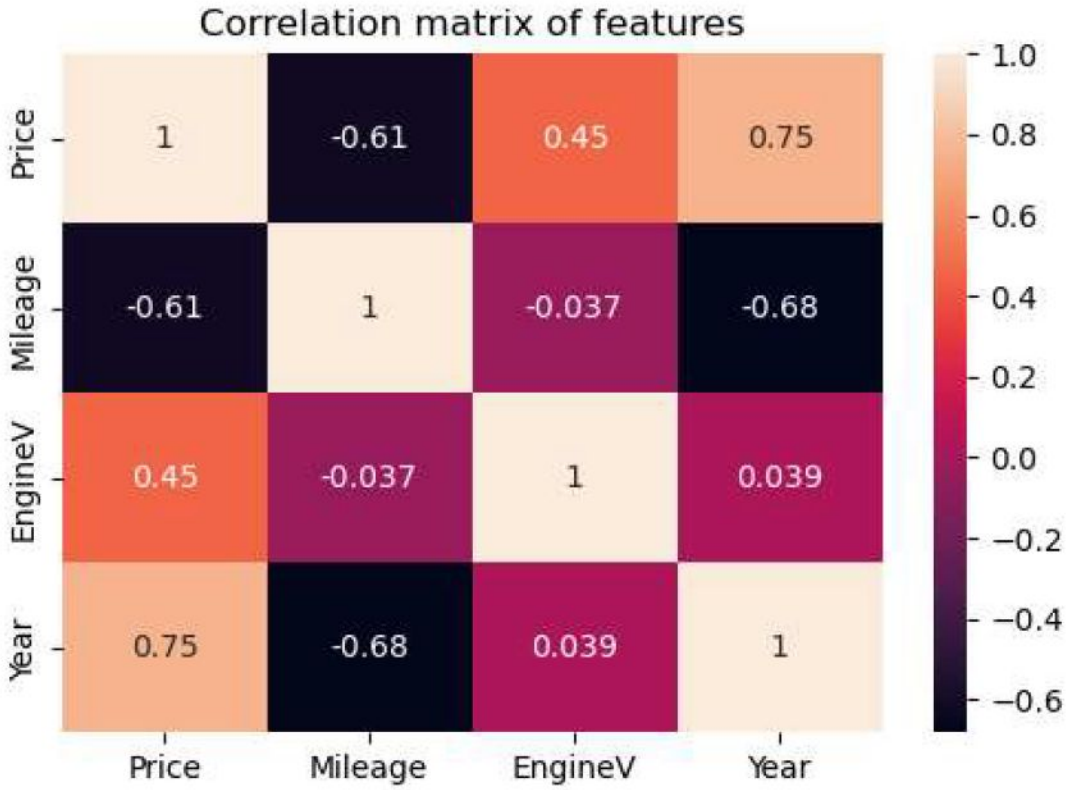
Pairs of Scatter plots after transformation

- Evidence of linear relationships suggests linear regression suitability.
- Predictable price movement with variable changes is ideal for linear regression.
- Scatter plots emphasize the need for transformations and feature engineering for better model performance.



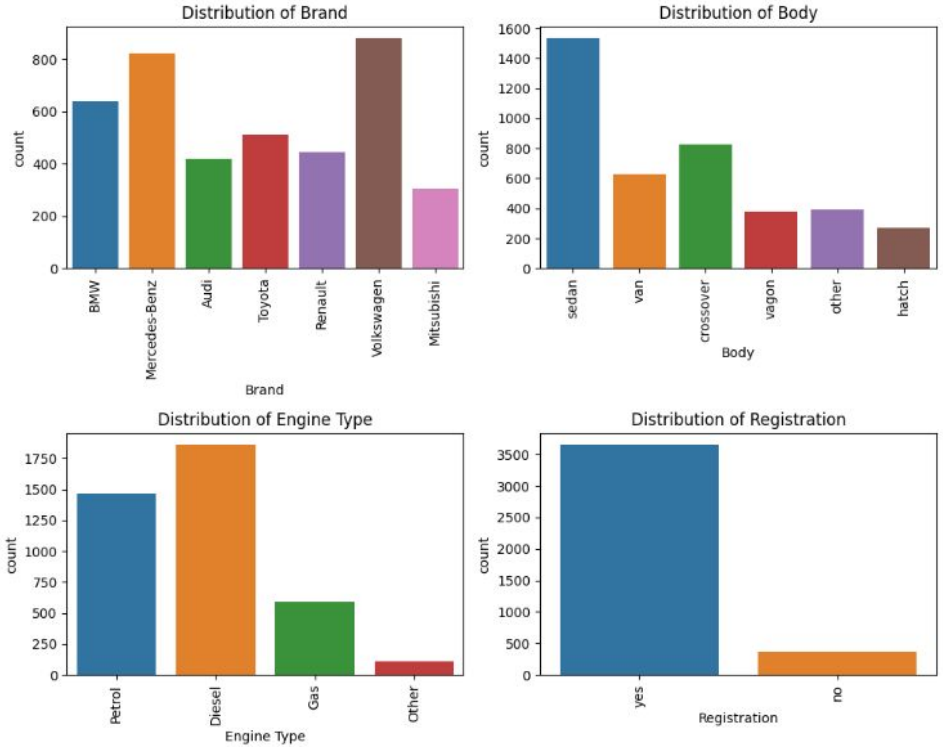






This correlation matrix gives us a snapshot of the relationships between key variables. For instance, we see a strong negative correlation of -0.61 between mileage and price, suggesting that higher mileage is associated with lower car prices. Conversely, the year of the car has a strong positive correlation of 0.75 with the price, indicating that newer cars tend to be priced higher. Engine volume (EngineV) shows a moderate positive correlation with price, at 0.45, hinting that cars with larger engines may command higher prices. These insights are pivotal in understanding which features are most influential in predicting the price of a used car.

Distributions of Categorical Data



Conversion of Categorical data to Binary

	Price	Mileage	EngineV	Year	Brand_BMW	Brand_Mercedes-Benz	Brand_Mitsubishi	Brand_Renault	Brand_Toyota	Brand_Volkswagen	Body_hatch	Bod
0	8.342840	277.0	2.0	1991	1	0	0	0	0	0	0	
1	8.974618	427.0	2.9	1999	0	1	0	0	0	0	0	
2	9.495519	358.0	5.0	2003	0	1	0	0	0	0	0	
3	10.043249	240.0	4.2	2007	0	0	0	0	0	0	0	
4	9.814656	120.0	2.0	2011	0	0	0	0	1	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	
4339	9.792556	35.0	1.6	2014	0	0	0	0	1	0	0	

- After creating dummies the total number of columns are 19

4. Model Building

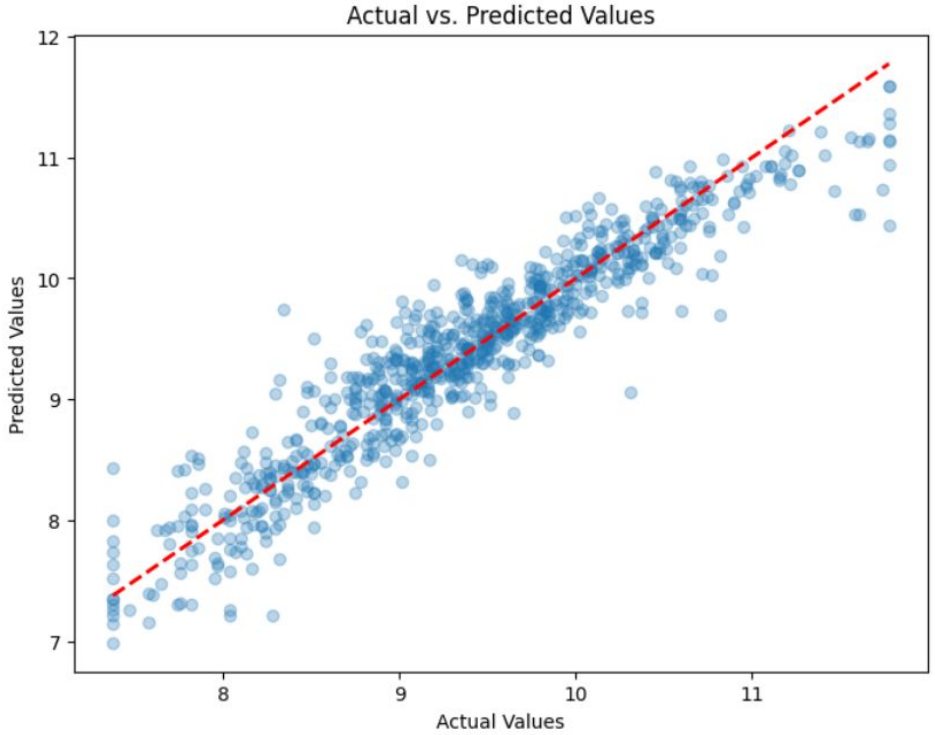
- We have split our data into 80% training and 20% testing.

Regression Models
• Linear Regression Model
• Random Forest Regressor
• Gradient Boosting Regressor

Linear Regression Model Performance

- The model generally predicts accurately, as seen by the clustering of points around the perfect prediction line.
- Deviations at higher values and scattered outliers suggest reduced accuracy in certain areas.

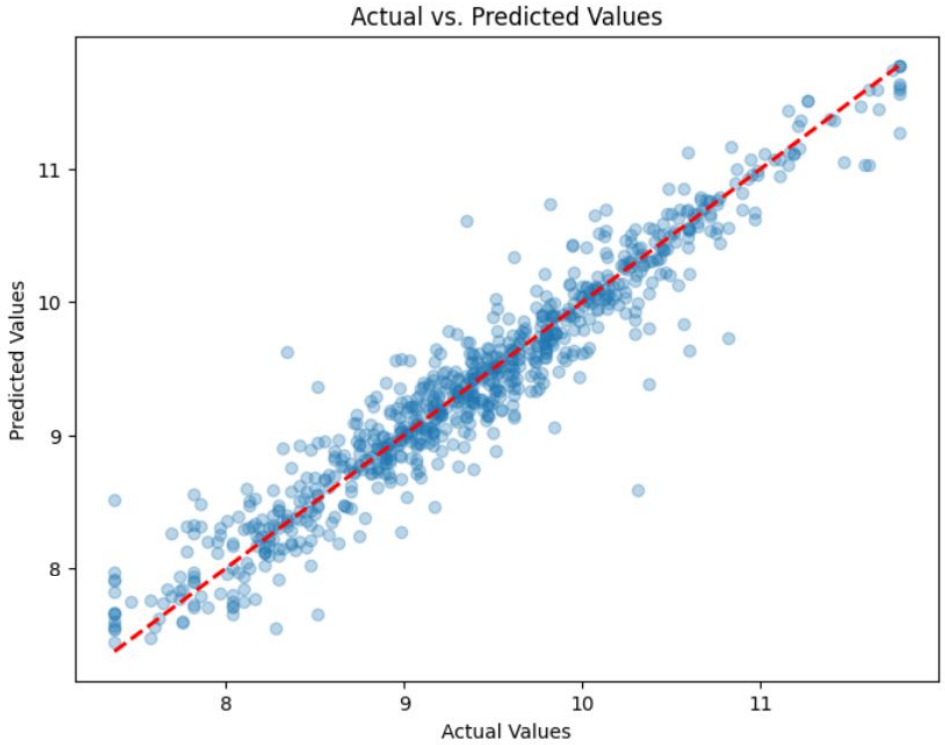
```
R_squared : 0.8834721086790O9
RMSE : 0.30661501541034153
```



Random Forest Regressor

- strong correlation between the actual and predicted values, indicating good model performance.
- There is some scatter above and below the line of perfect fit, suggesting variability in the model's predictions.

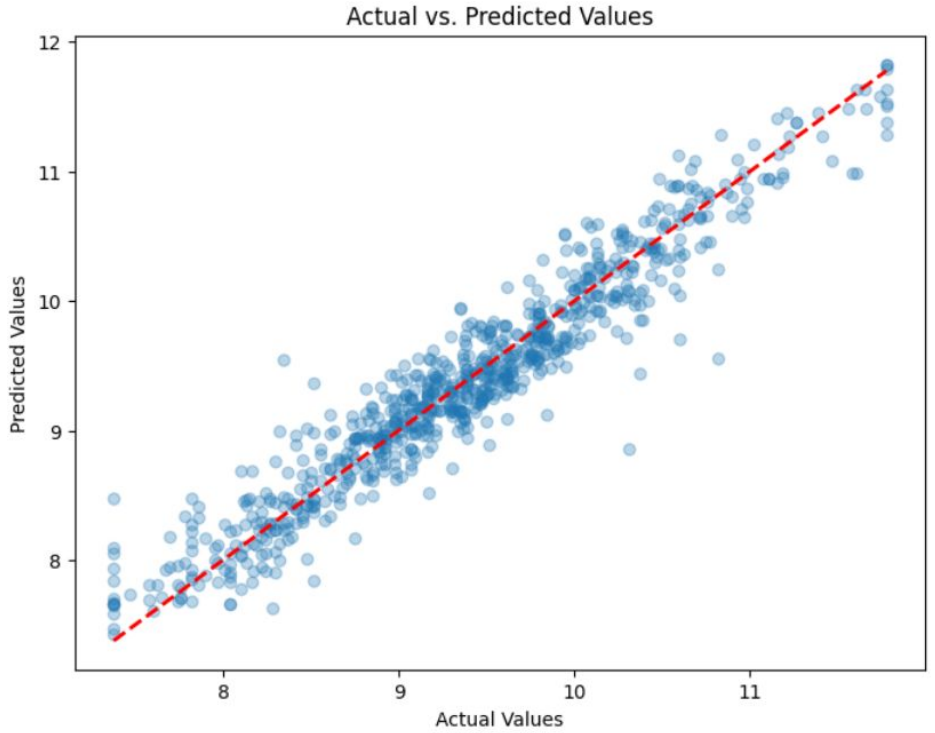
```
R_squared : 0.9187208341384878
RMSE : 0.2560755382866967
```



Gradient Boosting Regressor

- The close clustering of points around the diagonal indicates the Gradient Boosting Regressor has a high prediction accuracy.
- Despite the overall strong performance, there's a slight pattern of underpredicting at higher actual values, seen by the data points below the line.

```
R_squared : 0.9188709268388843
RMSE : 0.2558389904032532
```



5. Model Performance

Regression models	R2	RMSE
• Linear Regression Model	0.8834	0.3066
• Gradient Boosting Regressor	0.9189	0.2559
• Random Forest Regressor	0.9187	0.2560

- So as we try different Regression Algorithms and found that "**GBT Regressor Model**"is giving better accuracy compare to other models we check the predictions manually.

	Predicted Price	Actual Price	Residual	Difference%
796	11548.93	12000.00	451.07	3.76
797	18172.29	18400.00	227.71	1.24
798	28931.88	28500.00	-431.88	1.52
799	12386.44	14000.00	1613.56	11.53
800	8357.08	9500.00	1142.92	12.03

6. Conclusion

- Random Forest Regressor: R-squared of 0.9187, RMSE of 0.2560.
- Gradient Boosting Regressor: R-squared of 0.9189, RMSE of 0.2559.
- **Gradient Boosting shows a negligibly higher R-squared and a slightly lower RMSE, indicating a marginally better fit to the data.**


Thank you