CAR PRICE PREDICTION MODEL REPORT:

1.Introduction:

This report details the process of building and evaluating a linear regression model to predict car selling prices based on the given dataset. This analysis covers data understanding, exploratory data analysis(EDA),data processing, model development, evaluation and interpretation.

2. Data Understanding and Exploration

The CarPricePrediction.csv dataset has 4340 rows and 8 columns, which present information on various car features and selling prices.

Key Observations:

No Missing Values: The data was clean with no missing values, making the data preparation process easy.

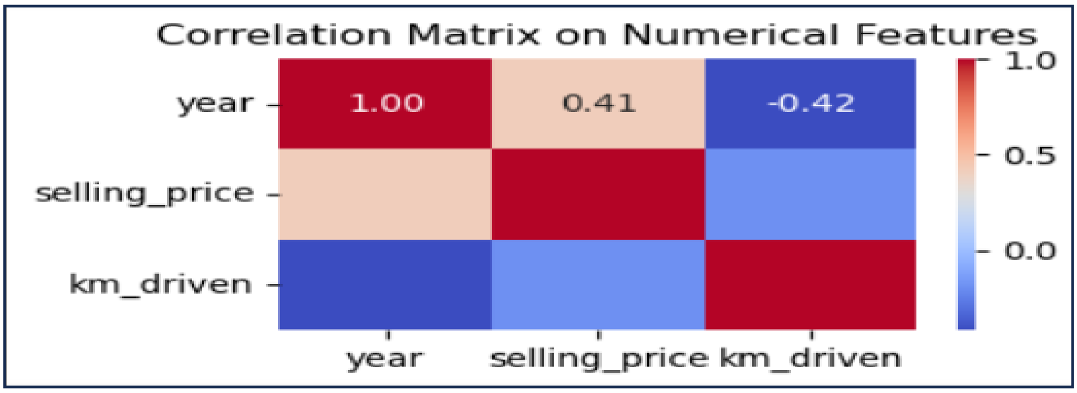
Data Types: Attributes like year, selling_price, and km_driven were numeric, while name, fuel, seller_type, transmission, and owner were categorical.

Correlation Analysis:

year had a moderate positive correlation (0.41) with selling_price, reflecting the fact that newer cars are more expensive.

km_driven had a weak negative correlation (-0.19) with selling_price, reflecting that higher mileage is linked to lower price.

Inverse relationship (-0.42) was found between year and km_driven, i.e., older vehicles tend to have higher mileage.



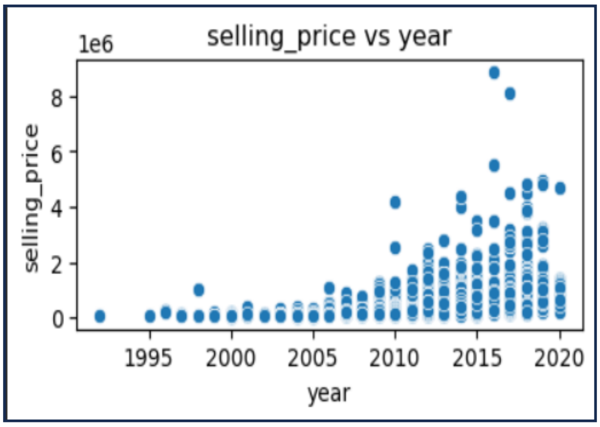
Data Distributions:

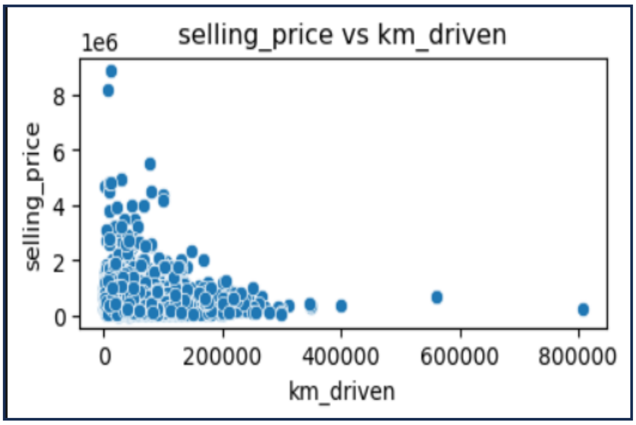
distributions of selling_price and km_driven were right-skewed, reflecting the dominance of lower-priced, lower-mileage vehicles with some high-value and high-mileage outliers.

The distribution of year exhibited a dominance of vehicles from older years.

Feature vs. Price Visualizations (EDA):

Numerical Features: Scatter plots validated the positive trend in year vs. selling_price, and a negative trend in km_driven vs. selling_price.





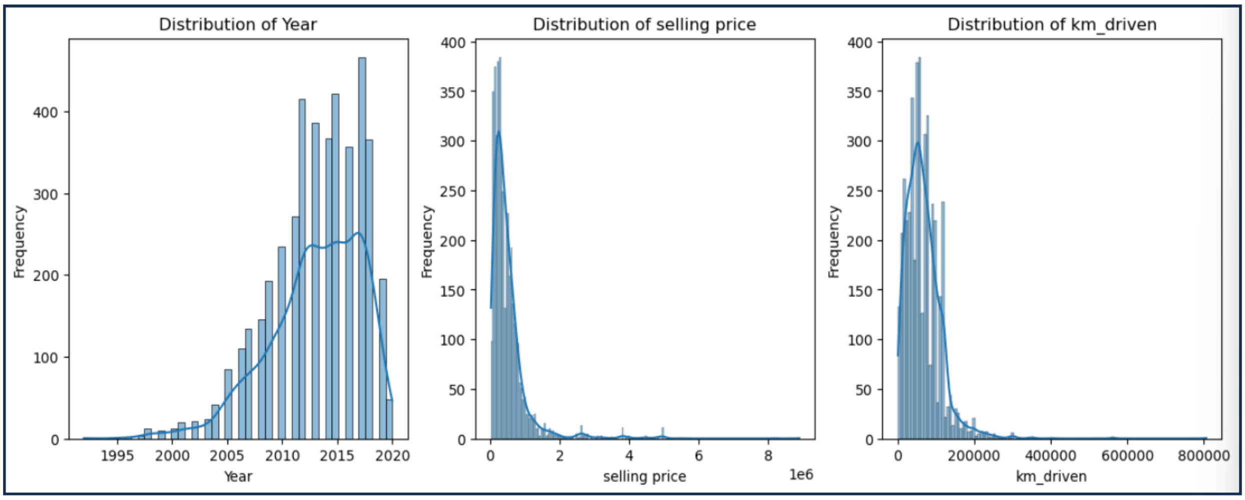
Categorical Features: Box plots showed large price variations across categories:

Diesel vehicles tended to cost more than petrol, CNG, or LPG vehicles.

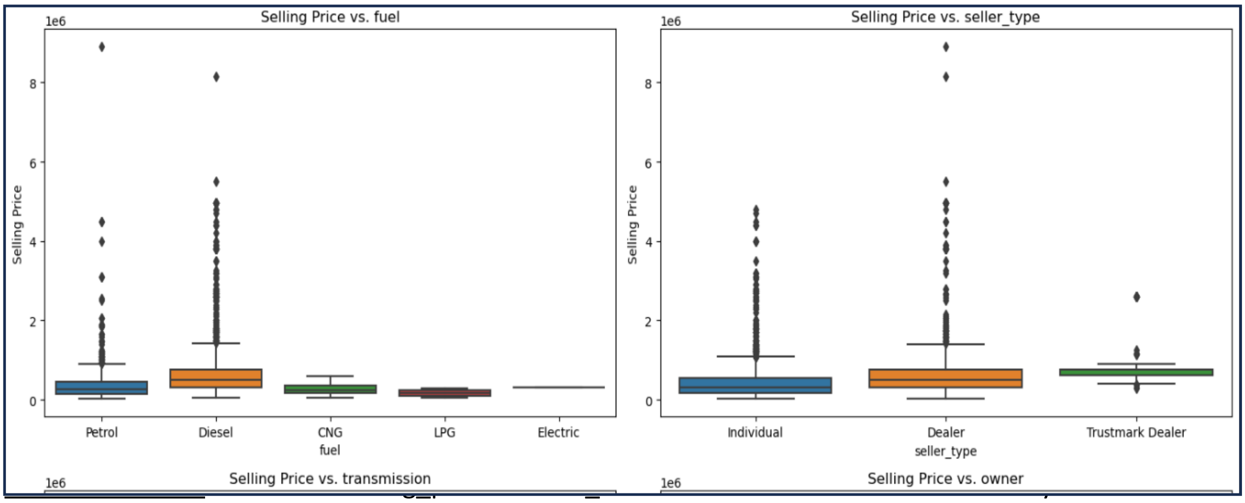
Dealers and Trustmark Dealers sold cars for a higher median price than Individual sellers.

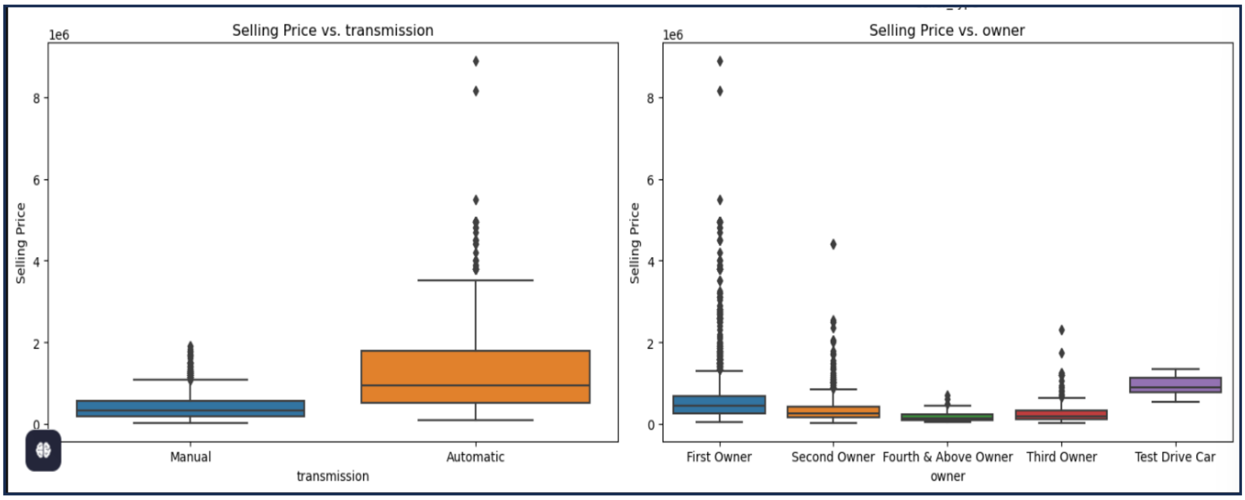
Automatic transmission vehicles tended to be more expensive than Manual vehicles.

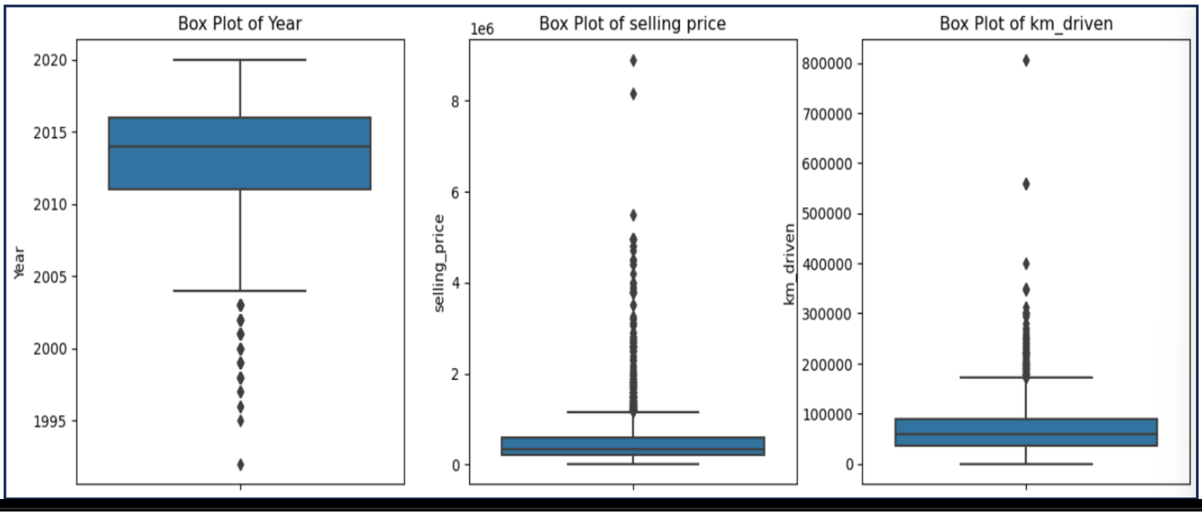
First Owner vehicles fetched the top prices, with price value dropping as the number of owners went up.



Outliers: Box plots showed that outliers existed in selling_price (luxury cars) and km_driven (extremely high mileage cars).







3.Data Preprocessing

Interquartile Range (IQR) method. This ensured the reduction of their overwhelming effect on the linear model.

Categorical Encoding: Categorical variables (fuel, seller_type, transmission, owner) were encoded into a numerical form by OneHotEncoder with drop='first'. This is to avoid multicollinearity (the "dummy variable trap") that occurs when all dummy variables are kept.

Numerical Feature Scaling: Numerical features (km_driven, year) were scaled with StandardScaler. This scales the data to a mean of 0 and standard deviation of 1 so that features with bigger numerical ranges do not disproportionately affect the model.

Data Splitting: The data after preprocessing was divided into training set and testing set with an 80:20 ratio to test the model's performance on new data

4. Model Development

A Linear Regression model was selected due to its ease of interpretation and simplicity as a baseline. The model was trained on the preprocessed training data.

5. Model Evaluation

The performance of the model was evaluated based on common regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2).

METRIC	BEFORE Cleaning	AFTER Cleaning
Mean Absolute Error	219541.48	120931.93
Mean Squared Error	181933371152.97	116189180899.29
Root mean Square error	426536	340865.34
R2 Score	0.40	0.62

Results Analysis:

The changes made (removal of outliers and correct one-hot encoding) improved the performance of the model greatly. The MAE reduced significantly, which means that the model's predictions are closer to the actual selling prices on average. The R2score improved from 0.40 to 0.62, meaning the model now accounts for 54% of car selling price variance.