# Car Selling Price Prediction Summary

# LLM used: GPT-4.1

# What was done:

## Data Cleaning:

Converted ‘max\_power’ to numeric.

Filled missing values using median and mode based on categorical groups.

Clustered rare categories in ‘owner’ and ‘fuel’ columns for better feature representation.

*LLM was used to ask if it was best to drop the missing value rows or fill them.*

## Exploratory Data Analysis (EDA):

Visualized distributions of key features like ‘selling\_price’, ‘km\_driven’, ‘mileage’, ‘engine’, and ‘max\_power’.

Analyzed relationships between ‘selling\_price’ and categorical features like ‘fuel’, ‘transmission’, and ‘owner’.

*LLM was used to ask the best way to visualize the data*

## Feature Engineering:

Extracted ‘car\_age’ from the ‘year’ column.

Applied log transformations to ‘selling\_price’ and ‘km\_driven’ to reduce skewness.

Scaled numerical features using ‘StandardScaler’.

One-hot encoded categorical features for model compatibility.

*LLM suggested to extract the ‘car\_age’ from ‘year’ and apply log transformation to ‘selling\_price’ and ‘km\_driven’. Log transformation were not applied to other features because they were not heavily skewed with minimum outliers. One-hot encoding was used instead of label encoding for linear regression to perform better.*

## Modeling:

Split the data into training and testing sets (80/20 split).

Trained three models: **Linear Regression**, **Random Forest**, and **XGBoost**.

Evaluated models using **RMSE** (for measuring the average magnitude of the error between the predicted and actual values) and **R²** (for measuring how well the model explains the in the target variable) metrics.

# What Was Found in the Data:

* 1. Selling prices are heavily skewed, with most cars priced below 10 lakhs.

Features like ‘max\_power’ and ‘engine’ have strong positive correlations with ‘selling\_price’.

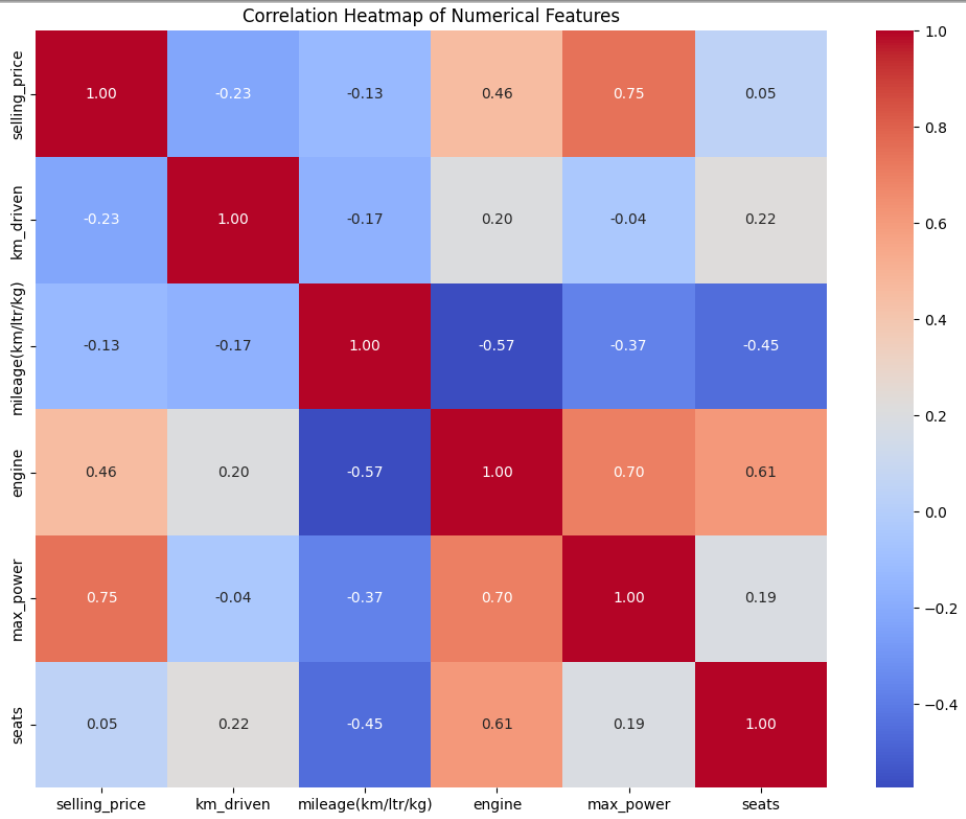
Mileage and ‘km\_driven’ show weaker negative correlations with ‘selling\_price’.

Diesel and petrol cars dominate the dataset, while automatic transmission cars tend to have higher prices.

# Graph Interpretations:

*LLM was used to give interpretations of the data obtained from EDA step.*

## Correlation Heatmap:



Selling price has a strong positive correlation with max power (0.75) and engine capacity (0.46).

Mileage shows a negative correlation with engine capacity (-0.57) and max power (-0.37).

Kilometres driven has a weak negative correlation with selling price (-0.23).

## Selling Price vs Engine: A graph of a price chart AI-generated content may be incorrect.

Cars with larger engine capacities tend to have higher selling prices.

Outliers with high prices exist even for smaller engines.

## Selling Price vs Mileage: A graph of blue dots AI-generated content may be incorrect.

Cars with mileage around 15–25 km/ltr/kg dominate the dataset.

Higher mileage cars generally have lower selling prices, with exceptions.

## Selling Price vs Kilometres Driven: A graph of a price AI-generated content may be incorrect.

Cars with fewer kilometres driven tend to have higher selling prices.

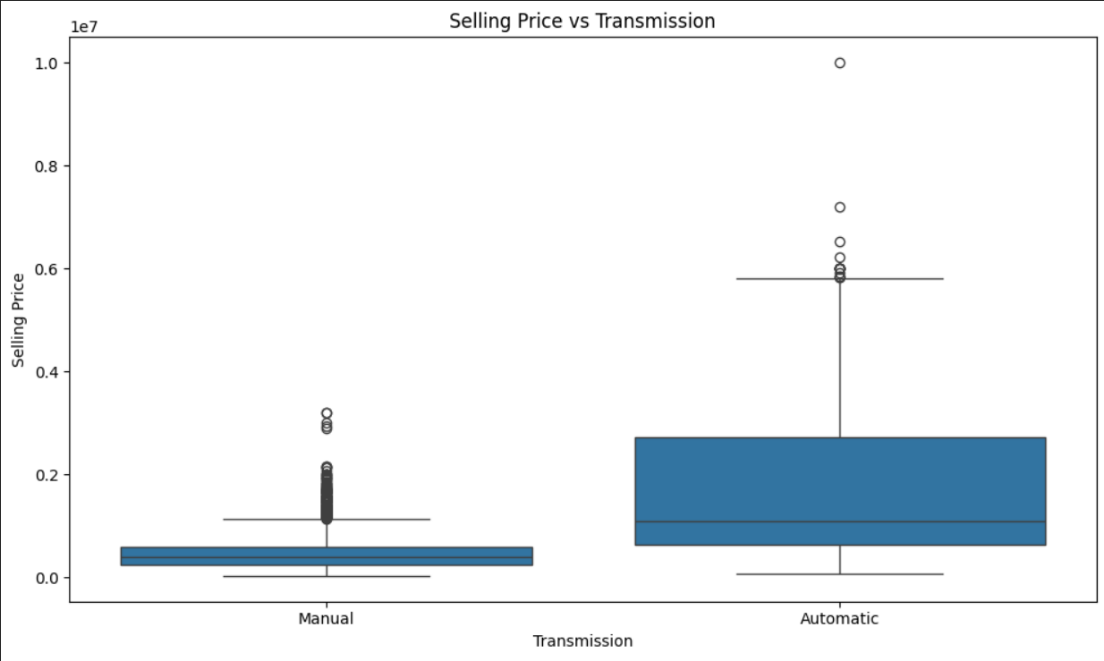
There is a steep decline in price as kilometres driven increase.

## Selling Price vs Owner: A diagram of a price comparison AI-generated content may be incorrect.

First-owner cars have higher selling prices compared to second-owner or third-owner cars.

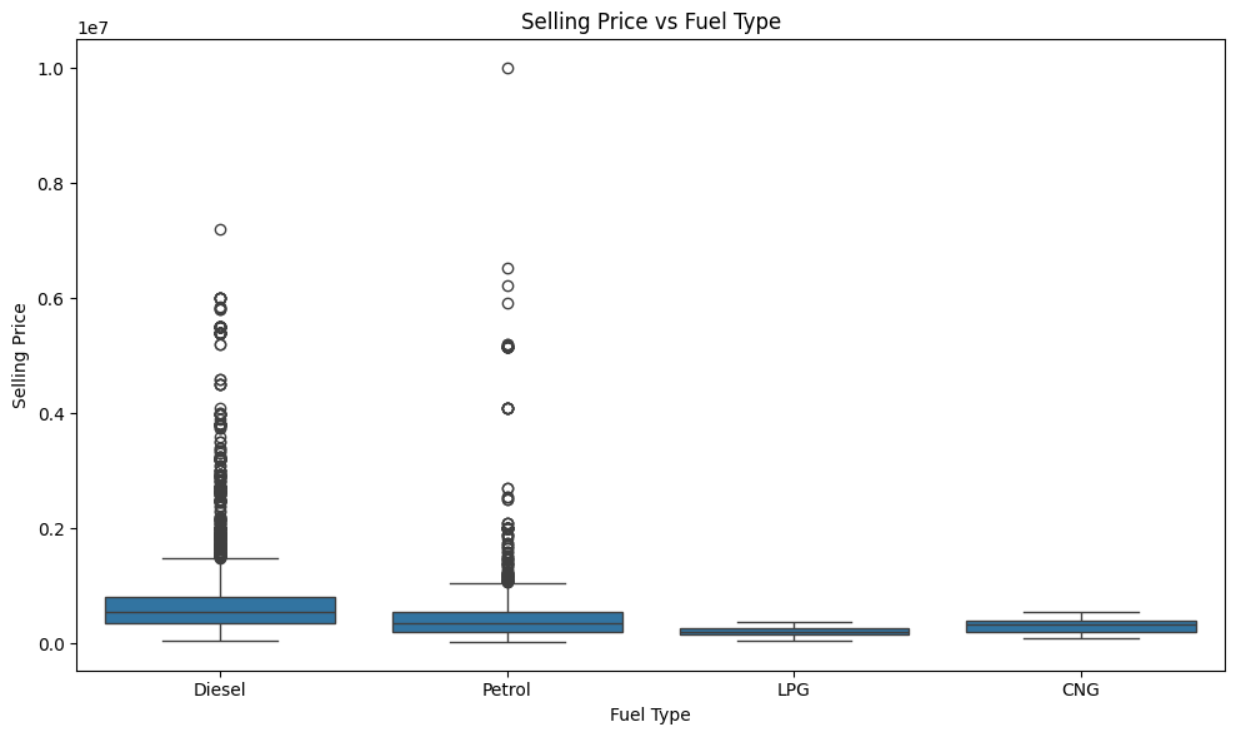
Test drive cars show the highest prices due to their near-new condition.

## Selling Price vs Transmission:



Cars with automatic transmission tend to have higher selling prices compared to manual transmission cars.

## Selling Price vs Fuel Type:



Diesel cars generally have slightly higher selling prices than petrol cars.

LPG and CNG cars have lower prices.

# How the Model Works:

* 1. Random Forest was chosen as the best model due to its ability to capture non-linear relationships and handle mixed data types.

It uses an ensemble of decision trees to predict car prices, averaging predictions to reduce overfitting and improve accuracy.

# Model Accuracy:

## **Random Forest Regression** delivered the **best performance among all models:**

**R² Score: 0.936** - This means the model explains **93.6% of the variance** in the target variable, indicating a very strong fit.

**RMSE: 0.249** - On average, the predictions are off by **approximately 0.249 units**, which is the **lowest error** among the tested models.

## **XGBoost Regression** performed similarly well:

**R² Score: 0.935** - It explains **93.5% of the variance** in the target data, just slightly below Random Forest.

**RMSE: 0.251** - The average prediction error is about **0.251 units**, very close to that of the Random Forest model.

## **Linear Regression** had the **least accurate results**:

**R² Score: 0.908** - This model explains **90.8% of the variance**, which is still good but lower than the other models.

**RMSE: 0.298** - The predictions deviate by an average of **0.298 units**, making it the **least precise** of the three.

# **Conclusion: Why Random Forest is suitable for this car dataset:**

## **Captures Non-Linear Relationships**

Car pricing is complex: The relationship between features like age, mileage, brand, and engine size is not always linear.

Random Forest can capture complex interactions between these features, such as how brand and age together affect price.

Linear Regression assumes a straight-line (linear) relationship, which can oversimplify real-world pricing patterns.

## **Handles Feature Interactions Better**

Car features interact in subtle ways. For example, a luxury brand with high mileage may be priced differently than an economy brand with the same mileage.

Random Forest automatically learns these interactions using its tree-based structure.

Linear Regression treats each feature independently, missing out on important combinations.

## **Robust to Outliers**

Car datasets often contain outliers, such as luxury or vintage vehicles with extreme values.

Random Forest is less sensitive to these outliers because it averages results across multiple trees.

Linear Regression can be heavily skewed by extreme values, reducing its reliability.

## **Better Performance on Mixed Data Types**

The dataset includes both categorical features (brand, fuel type, transmission) and numerical features (mileage, age, engine size).

Random Forest can naturally handle this mix without needing extensive preprocessing.

Linear Regression requires all inputs to be numeric and often needs encoding and normalization.

## **Performance Evidence**

2.8% improvement in R²: 0.936 vs. 0.908 (Random Forest vs. Linear Regression)

16.4% reduction in RMSE: 0.249 vs. 0.298 (lower prediction errors)

This translates to more reliable and accurate price estimates.

## **Business Impact**

93.6% model accuracy means predictions are highly reliable.

More accurate price estimates improve user trust in the system.

The Random Forest model is well-suited for deployment in car price prediction applications.