Introduction

The Indian automobile market is evolving rapidly, fuelled by urbanization, economic growth, and diverse customer preferences across major cities like Bangalore, Surat, Delhi, Hyderabad, and Mumbai. Each region presents unique opportunities, ranging from budget-conscious first-time buyers to affluent customers seeking luxury vehicles. To thrive in this dynamic landscape, Car Diggers employs a data-driven approach to revolutionize the dealership experience.

By analyzing **regional trends** and customer behaviour, we curate an inventory that aligns with market demands, ensuring the right mix of vehicles for every buyer segment. Our advanced pricing models and market-backed inventory strategies enable competitive pricing while maintaining profitability. Additionally, targeted marketing efforts allow us to engage effectively with diverse buyer personas, from cost-sensitive customers to premium car enthusiasts. Through this comprehensive strategy, Car Diggers aims to deliver a seamless and tailored experience, establishing itself as a forward-thinking and trusted dealership across India.

Car Diggers, a new-age car dealership, aims to revolutionize the customer experience by leveraging historical and current data and machine learning models. By analyzing car specifications and pricing patterns, Car Diggers provides actionable insights for accurate price estimation, customer segmentation, and targeted marketing. Our business strategy is built on four key customer clusters, enabling us to tailor offerings for budget-friendly, mid-range, high-performance, and luxury car buyers across different regions. With a focus on innovation and customer-centric solutions, Car Diggers is poised to set a new benchmark for dealerships in India.

Scope in Car Dealership Market:

India is rapidly emerging as a significant automobile market, with the used car segment playing a crucial role in its economic structure and cultural landscape. A recent trend indicates a growing preference for pre-owned vehicles, driven by factors such as affordability, lower depreciation rates, and the availability of certified used cars. With more than 4 million used cars sold annually, this segment highlights the rising demand for cost-effective mobility solutions across the country. To address this demand, this project leverages a dataset encompassing various factors related to used cars, including specifications, condition, mileage, and ownership history. By analyzing these attributes, the project aims to develop a reliable pricing prediction model that can help buyers and sellers make informed decisions about a vehicle's market value, supporting the growth of India's car market.

Problem Statement:

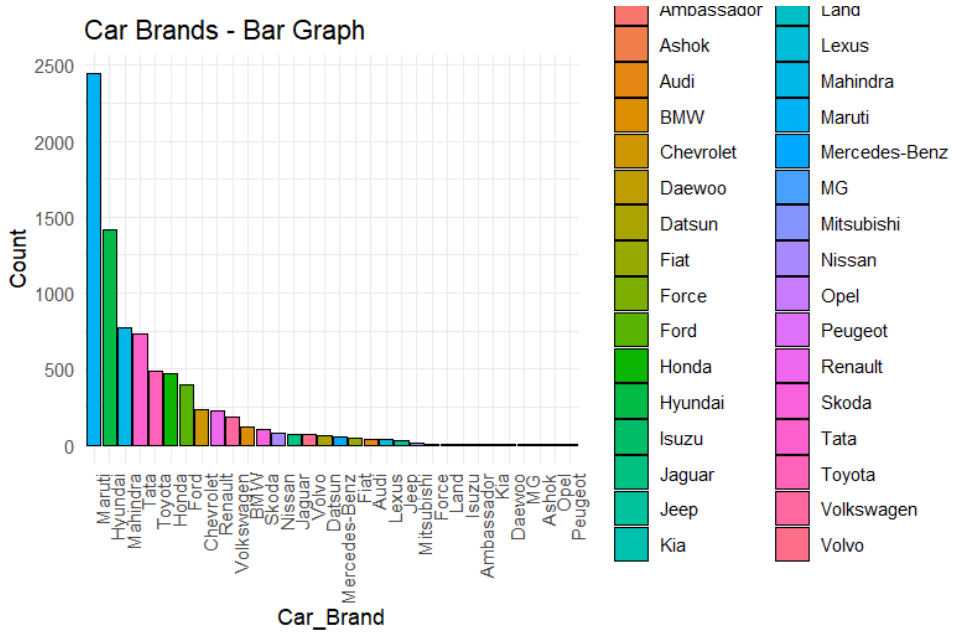
This project aims to address the challenges faced by potential buyers or sellers of used and brand-new automobiles in estimating the price of that particular vehicle, with the help of the multiple variables from this dataset we intend to create models that can accurately estimate prices of cars, help in customer segmentation, and optimized inventory which can then be used to support informed decision-making for managing a car dealership and helps the customer in buying a car on their personal preferences.

Data Overview:

This dataset our company has consists of 8130 observations approximately and the vehicle data is recorded over the past 5 years i.e., 2018-2023. There are various variables in the dataset which are used further in to train the models and predict the selling prices of cars. The variables and their types are as follows:

Columns	***Types***
Name	Categorical
Fuel	Categorical
Seller type	Categorical
Transmission	Categorical
Owner	Categorical
Torque	Categorical
Mileage	Numerical
Engine	Numerical
Max Power	Numerical
Seats	Numerical
Selling price	Numerical

Data Cleaning & Exploration:



Here the data of various cars and the companies of which those cars belong to is shown and from the bar graph we can observe that the there are multiple car brands in the dataset but there are few companies which have multiple and a greater number of models in the dataset, there are also high-end luxury cars but in smaller number. This kind of data helps in catering to large number of customers and diversify the clientele.

<u>Cleaning blank and N/A values:</u>

In the cleaning process, the names of the car companies are converted into ordinal numbers for referencing convenience and further training of the model

```
# A tibble:
  name
  <chr>
1 Maruti …
2 Skoda R…
3 Honda C…
4 Hyundai…
5 Maruti …
5 Hyundai…
```

name
0
1
2
3
4
5
6
7
8

There are multiple columns with blank spaces and there are also “N/A” values which are not useful and are in fact an obstacle in creating a model, Here the **blank and “N/A” values** are **replaced** with

```
> colSums(blank_or_na)
        name         year selling_price    km_driven         fuel  seller_type transmission
        1371            0             0            0            0            0            0
       owner      mileage       engine    max_power       torque        seats
           0          221          221          215          222          221
```

mean and median. But first, in order to replace them with arithmetic measures we converted the blank values them to “N/A”. The following are the blank and N/A values combined in each column:

After further cleaning of the data, we converted the blank values to N/A:

```
> colSums(blank_or_na) #final is.na values
         name          year selling_price     km_driven          fuel   seller_type  transmission
         1371             0             0             0             0             0             0
        owner       mileage        engine     max_power        torque         seats
            0           221           221           215           222           221
```

From the above, it can be seen that all the blank values are converted into N/A.

The values which have N/A are further changed to have the mean and median values:

Replacing null values with mean & median:

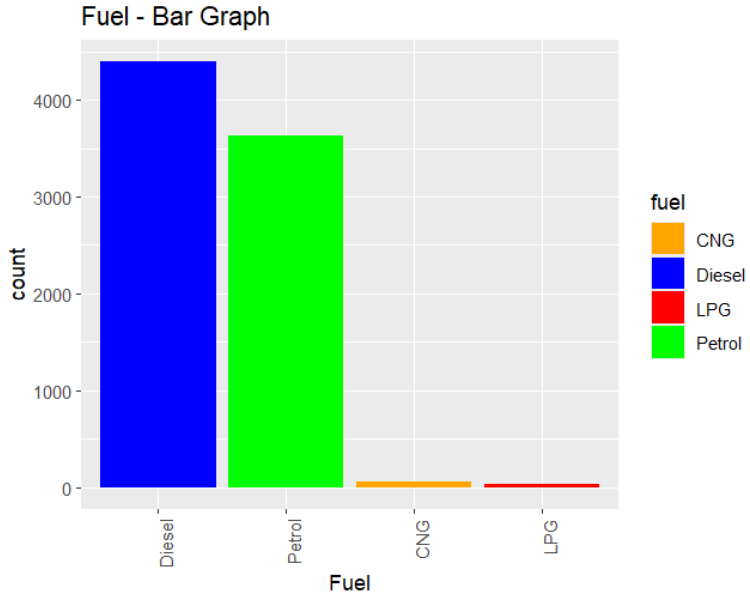
The columns: **Mileage, Engine, Max Power** are changed to have the values of **Mean.** The **Seats** column is changed to have the **Median.** Here we wanted to have the central value at which most customers are willing to consider their desired car to have and replaced the N/A values with those particular values.

mileage	engine	max_power	seats
23.40000	1248.000	74.00000	5
21.14000	1498.000	103.52000	5
17.70000	1497.000	78.00000	5
23.00000	1396.000	90.00000	5
16.10000	1298.000	88.20000	5
20.14000	1197.000	81.86000	5
17.30000	1061.000	57.50000	5
16.10000	796.000	37.00000	4
23.59000	1364.000	67.10000	5
20.00000	1399.000	68.10000	5
19.01000	1461.000	108.45000	5
17.30000	993.000	60.00000	5

After this step the data has become **ready** for training models and getting effective insights which helps in making business decisions.

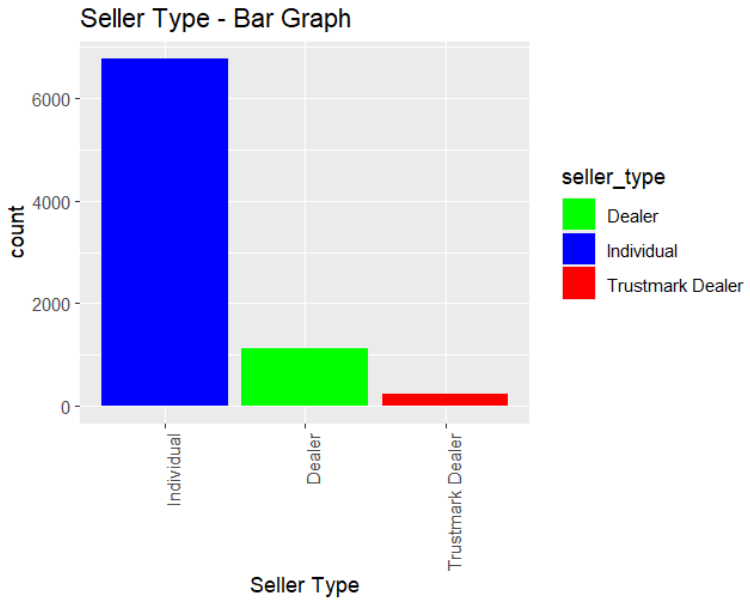
Data Exploration:

FUEL TYPE:



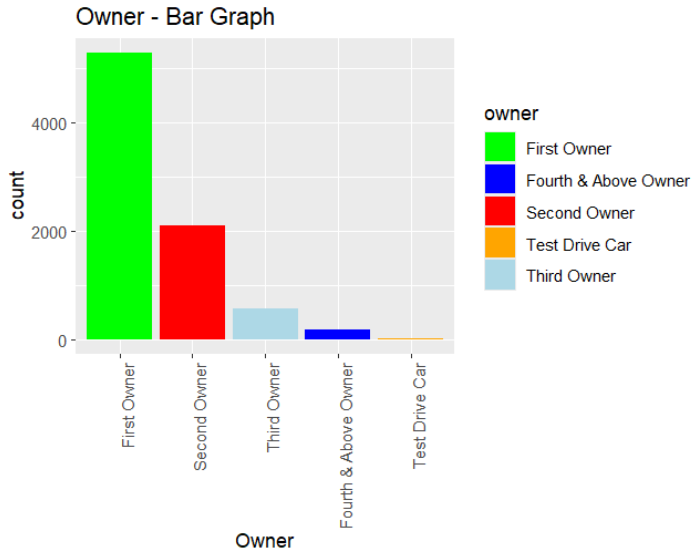
This graph enables to understand that the vehicle's fuel and their count in the dataset, from this dataset we can understand that the demand for diesel vehicles is more in the market, followed by petrol vehicles. This helps in understanding what kind of vehicles to have more in the inventory and increase that particular type of inventory based on requirements for the customers and trends in the market. CNG & LPG are in circulation in the market but do not have much demand because of various reasons like availability, higher initial purchase costs compared to petrol and diesel and also studies show that the vehicles with CNG & LPG are not much preferred for personal uses as much as they are used for commercial purposes.

SELLER TYPE:



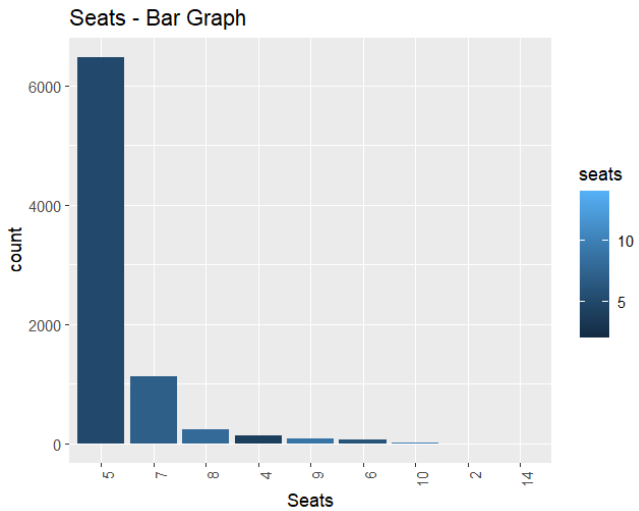
This particular variable tells us about the **type of seller** the car has been purchased previously from and here we can see that, many people in this dataset are interested in purchasing from an individual rather than from a dealer and trusted dealers i.e., government certified agencies. One reason for this could be is that a dealer has many vehicles in their inventory and there are many costs incurred by a person going to buy such car like markups, costs on customization and accessories replacement. When it comes to an individual the person who is buying the car will immediate transfer of ownership and there won't be any sales pressure on the individual and they can get to know the history behind the vehicle.

OWNER TYPE:



Here the graph represents the data of what type of pre-owned cars are customers willing to go for more, there is a clear majority for the first owners and rest all are at the lower end in comparison. There is a definitive reason for this because an individual who interested in buying a car will go for to the first owner because of reasons like vehicle performance, transparent records and documentation, lower chances of wear and tear of the vehicle. It is worth noting that there is still a market for second third and fourth owned cars but it is not a substantial market to cater but there is no harm in having such vehicles in our inventory.

SEATS:

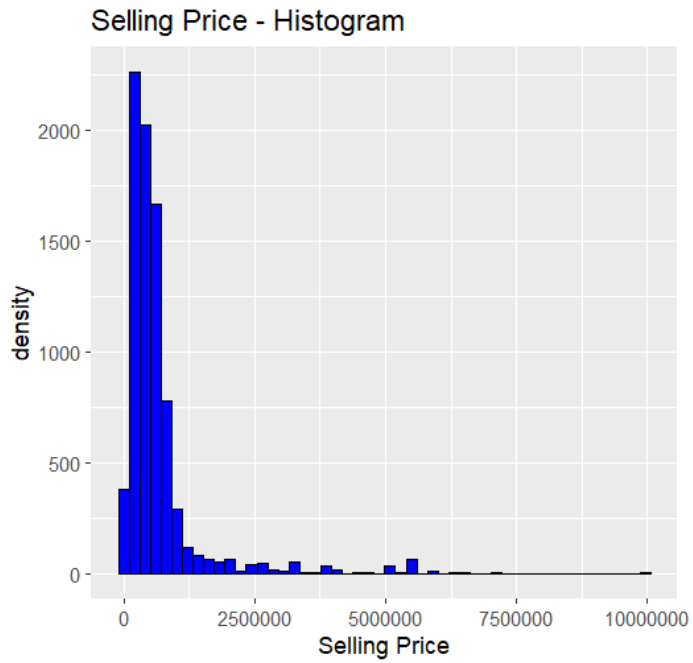


The requirement of number of seats in a vehicle are also graphed above, the graph shows a highly skewed distribution, with 5-seaters dominating and other seating capacities progressively having fewer vehicles. The bar for 5 seats is the tallest, indicating that most vehicles in the dataset are 5-seaters. This category significantly outnumbers all others, with a count exceeding 6,000. The inventory should be more focused on the 5-seater vehicles since there is a requirement in the market.

These are the insights from the graphs and plots for the above-mentioned variables, for the following variables **we created dummies** to split the data for easy understanding and interpretation.

```
> car.df.cat=dummy_cols(car.df,select_columns = c('owner','transmission','fuel','seller_type'), remove_selected_columns = TRUE)
```

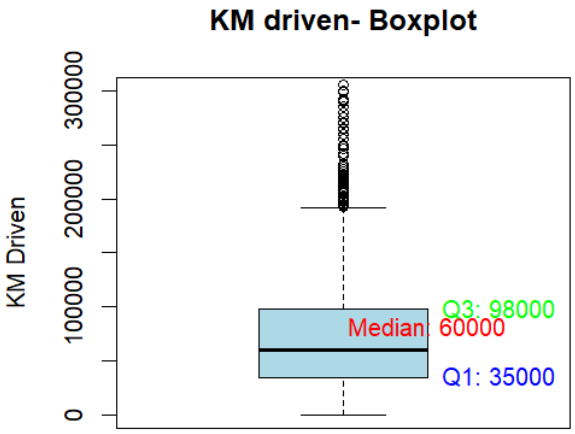
SELLER TYPE:



This graph is highly right skewed, here most vehicle selling prices are concentrated towards the lower end of the price spectrum. A few vehicles have very high selling prices, which stretch the distribution to the right. As the selling price increases beyond 2,500,000, the density decreases significantly. Prices above 5,000,000 and nearing 10,000,000 are very rare. The concentration of selling prices in the lower range indicates that most vehicles in the dataset are **relatively affordable and likely aimed at budget-conscious buyers.** From this graph alone we can come up with a vague price segmentation for the customers:

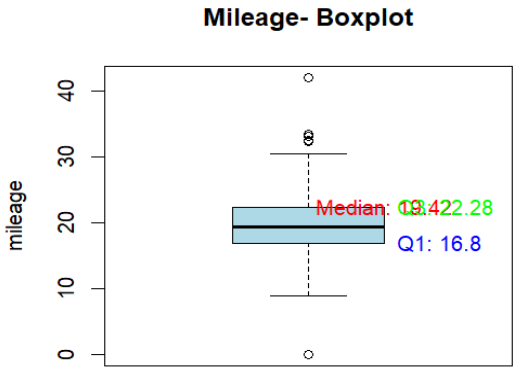
- **Economy Segment**: Vehicles below 2,500,000 dominate the market.
- **Mid-Range Segment**: Vehicles priced between 2,500,000 and 5,000,000 have moderate representation.
- **Premium Segment**: Vehicles above 5,000,000 cater to a very small, high-income customer base.

KILOMETERS_DRIVEN:



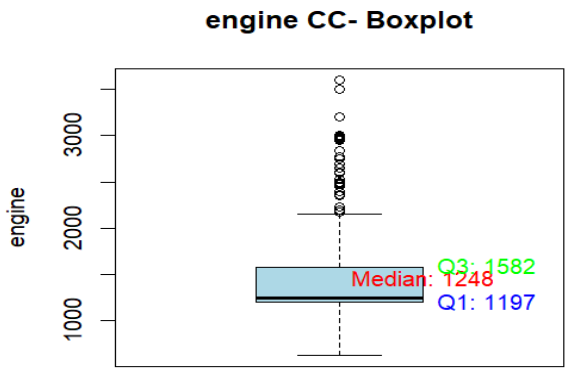
From this kilometre driven boxplot, we can see that the median is 60,000 km, half the vehicles have been driven less than 60,000 km, and half more. First Quartile is 35,000 km: 25% of the vehicles have been driven less than 35,000 km. Third Quartile is 98,000 km: 75% of the vehicles have been driven less than 98,000 km. We can interpret that the vehicles with lower mileage (below 60,000 km) are likely to attract buyers looking for newer, less-used cars. High-mileage vehicles, especially outliers, may require additional scrutiny for maintenance history and durability. The business should maintain a balanced inventory with a focus on mid-range mileage vehicles (35,000–98,000 km) since this fall within the interquartile range and are likely the most common and marketable.

MILEAGE OF THE VEHICLE:



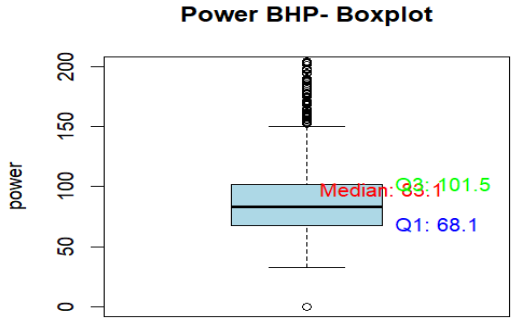
From this boxplot we can infer that the median mileage of 19.4 km/l suggests that most vehicles in the dataset offer good fuel efficiency, which is an appealing factor for buyers prioritizing cost savings on fuel. Vehicles with mileage close to the median and within the interquartile range (16.8 to 24.22 km/l) are likely the most marketable, balancing performance and efficiency. Mileage is a key determinant of vehicle pricing. Vehicles offering higher mileage can command premium prices, while lower-mileage vehicles might require discounts or incentives to attract buyers.

ENGINE:



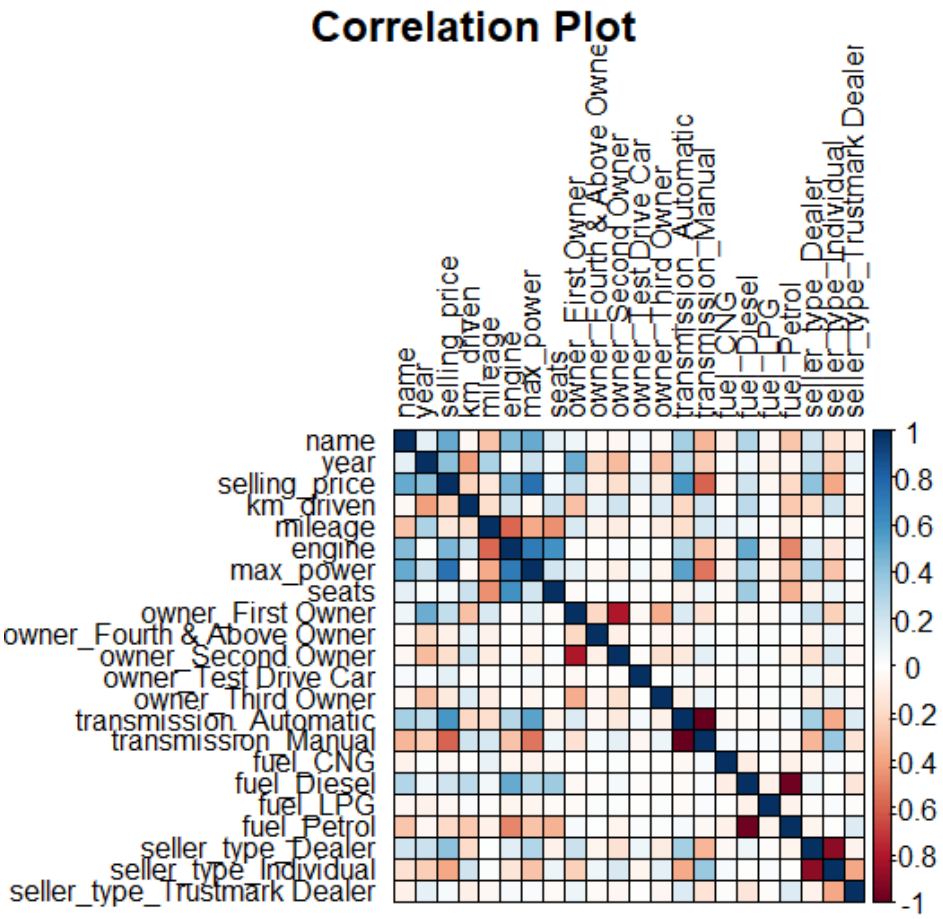
From this plot we can understand that the median of the engine requirement in this dataset is 1248 CC, first quartile: 1197 CC, third quartile: 1582 CC The median engine capacity of 1248 CC suggests that most vehicles in the dataset fall within the compact car segment, which is suitable for city driving, fuel efficiency, and affordability. These vehicles are likely popular among budget-conscious buyers or those prioritizing fuel efficiency. Ideal for urban markets with heavy traffic or parking constraints. Large-engine vehicles should be marketed strategically, highlighting unique features such as power, luxury, or off-road capabilities, to differentiate them from the mainstream market.

POWER BHP:



From this above plot we can see that the median, first and second quartile are 83.1, 68.1, 101.5 respectively. The clustering around 68.1–101.5 BHP indicates a preference for vehicles that balance power and efficiency, making them ideal for urban and moderate highway use. Vehicles with low power ratings might represent hybrid, electric, or ultra-compact models and can be marketed for eco-conscious buyers. Buyers looking for vehicles above 150 BHP likely prioritize performance or luxury, indicating a need for differentiation in features and branding. Diversifying offerings with a few high-performance models could cater to niche markets and boost brand perception.

CORRELATON OF THE VARIABLES IN THE DATASET:



From the correlation plot we can understand that the variables in **blue are highly correlated and variables in red are less correlated** with the other variables. This plot explains the correlation between the selling price which is one of the most significant factors for the vehicle and what are the factors which can and will affect the price of the car.

Strong Positive Correlations:

- **Selling Price vs. Engine, Power, and Year:** Selling price has a strong positive correlation with **engine size**, **max power**, and **year**
- Newer vehicles i.e. recently manufactured vehicles are associated with higher selling prices.
- Vehicles with larger engines and higher horsepower command higher selling prices, likely due to their superior performance or luxury status.

Strong Negative Correlations:

- **Selling Price vs. KM Driven**: Selling price is negatively correlated with kilometres driven. Vehicles with higher mileage tend to have lower selling prices due to wear and tear.
- **Year vs. KM Driven**: Newer vehicles tend to have fewer kilometres driven, reflecting limited use and better condition.

Insights:

- Ensure vehicles from first owners are priced higher than those with multiple owners, reflecting their better condition and perceived value.

- Include vehicles with automatic transmission, newer models (recent years), and diesel engines in your inventory, as these features correlate with higher resale prices.
- Customers are likely willing to pay more for diesel engines over petrol, emphasizing long-term fuel savings.
- Suggest regular maintenance and limited mileage to vehicle owners planning to sell in the future.
- Encourage potential sellers to use Trustmark Dealers to maximize their resale price.

Model Building Linear Regression:

To develop the linear regression model, we approached the task in two ways: without normalization and with normalization of the feature variables. The goal is to compare the performance of both approaches and evaluate their effectiveness in predicting car prices.

Dataset Preparation:

The dataset consists of 8,129 observations and 13 variables. To train and validate the model, the data was split into two subsets:

- Training Set (60%): Contains 4,876 observations used to train the model.
- Validation Set (40%): Contains 3253 observations used to evaluate the model's performance.

This split ensures the model has sufficient data to learn from while reserving a significant portion for unbiased evaluation.

Evaluation Metrics:

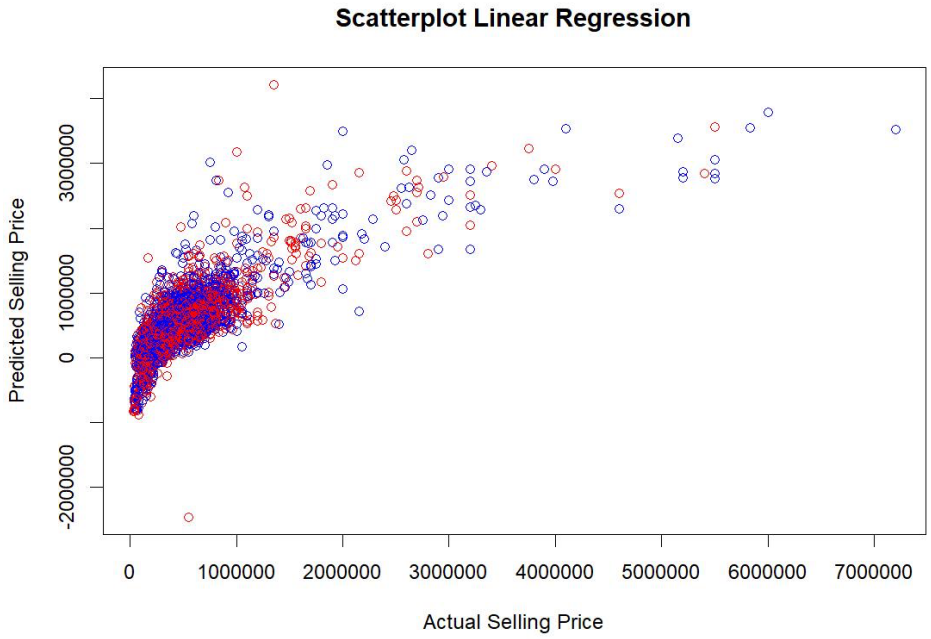
The model's performance was assessed using the following metrics:

1. Root Mean Squared Error (RMSE): Measures the average deviation of predicted values from the actual values, giving higher weight to larger errors.
2. Mean Absolute Error (MAE): Calculates the average of absolute differences between predicted and actual values, providing a straightforward interpretation of prediction accuracy.
3. R-Squared (R^2): Indicates the proportion of variance in the dependent variable explained by the independent variables, measuring the model's overall fit.

Steps in Model Building:

Without Normalization:

- The raw data was used directly without scaling the feature variables.
- The model was trained on the training dataset and evaluated on the validation set.



Scatterplot Analysis for Linear Regression (Without Normalization)

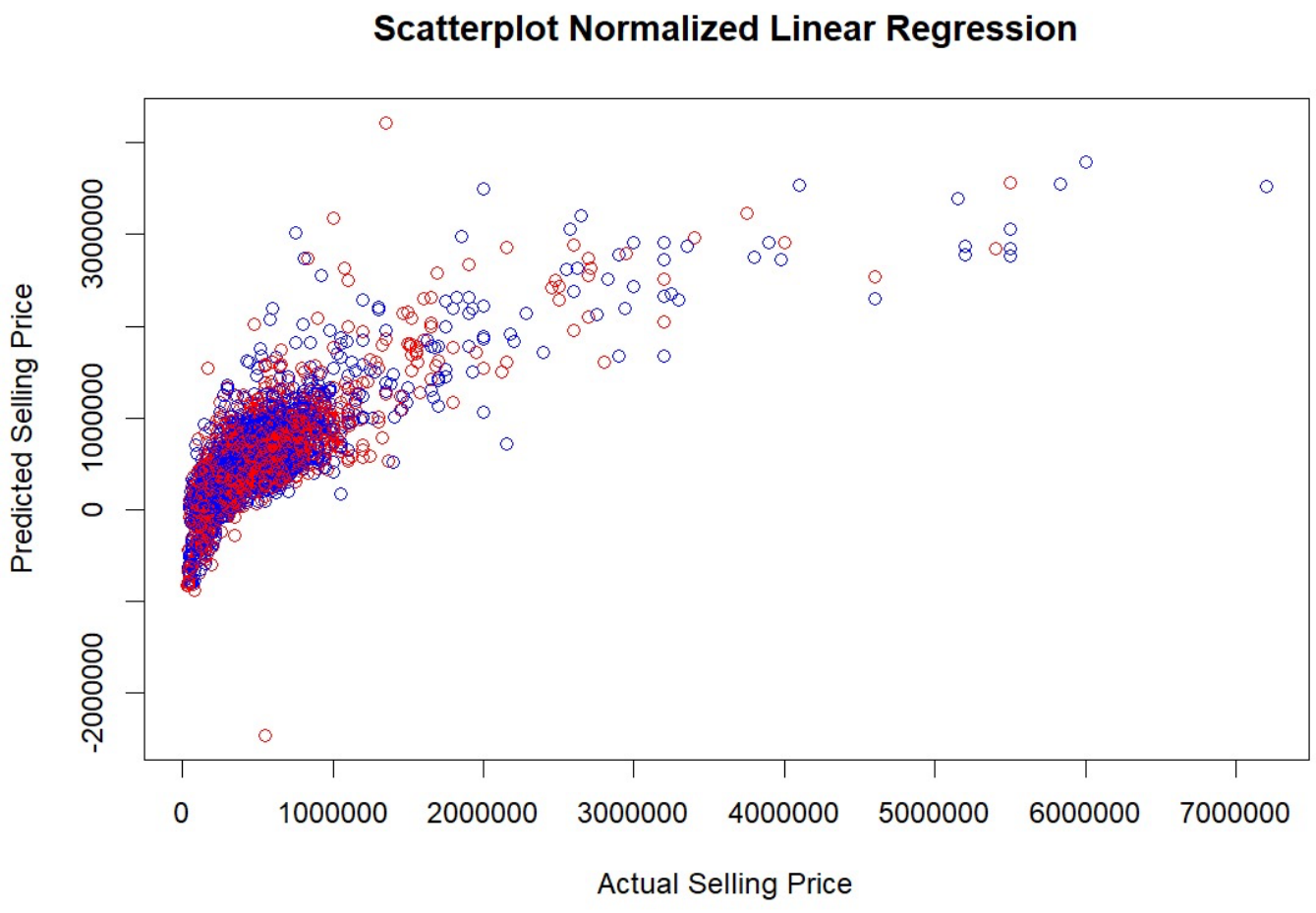
The scatterplot shown above depicts the relationship between the predicted selling prices and the actual selling prices for the linear regression model built without normalization. Each point on the graph represents an observation in the validation dataset.

From the scatterplot, it is evident that the predicted values closely align with the actual values for most data points, indicating that the model performs well in capturing the general pricing trends. While there is some deviation for higher selling prices, the overall alignment demonstrates that the model provides a reasonably accurate prediction without requiring normalization. This suggests that the model's coefficients and relationships between variables are robust even without scaling the input features.

This analysis serves as a foundation for comparing the results with the normalized dataset to evaluate whether normalization further enhances the accuracy or interpretability of the model.

With Normalization:

- The feature variables were normalized to have a mean of 0 and a standard deviation of 1.
- This ensures that variables with different scales (e.g., mileage and engine size) do not disproportionately influence the model.



Scatterplot Analysis for Linear Regression (With Normalization)

The scatterplot above illustrates the relationship between the predicted selling prices and actual selling prices for the linear regression model built with normalized feature variables. By normalizing the data, all features were scaled to have a mean of 0 and a standard deviation of 1, ensuring that variables with different units and ranges (e.g., mileage and engine size) contributed equally to the model.

In this plot, the predicted values closely follow the actual values for most observations, similar to the non-normalized model. However, normalization improved the model's handling of variables with large ranges, resulting in a more consistent spread of predictions, particularly for higher selling prices. The alignment of predicted and actual values demonstrates that normalization enhances the model's ability to generalize across the dataset, especially in scenarios where features have varying magnitudes.

Comparison of Normalized and Non-Normalized Linear Regression Models

From the scatterplots of both the normalized and non-normalized linear regression models, we observe that there is no significant difference in the alignment between the predicted selling prices and actual selling prices. Both models demonstrate a similar pattern, where most predictions closely follow the actual values, with minor deviations observed at higher selling prices.

The lack of a significant difference suggests that the data, in its raw form, was already structured in a way that allowed the model to capture the relationships between the features and the target variable effectively. Normalization primarily ensures fairness among features with varying scales, but in this case, the results indicate that both approaches provide comparable predictive performance.

Thus, while normalization offers theoretical benefits in certain scenarios (e.g., models sensitive to feature scaling), its impact on this dataset and linear regression model appears minimal. Evaluation metrics like RMSE, MAE, and R^2 further confirm the similarity in performance between the two approaches. This insight can help guide the choice of preprocessing steps for future models.

Error Metrics Analysis: Normalized vs. Non-Normalized Linear Regression

```
> print(rmse_df)
                      Algorithm Train_RMSE Valid_RMSE Train_MAE Valid_MAE  R2.Train  R2.Valid      Run_Time
1             Linear Regression   437254.4   452062.4  274829.6  278446.2 0.7086599 0.6809505 1.134209 secs
2 Linear Regression normalized   437254.4   452062.4  274829.6  278446.2 0.7086599 0.6809505 0.836720 secs
```

The table above compares the performance metrics for the linear regression model built using normalized and non-normalized data. The following observations can be made:

Training and Validation RMSE:

Both models have identical Root Mean Squared Error (RMSE) values for the training data (437254.4) and validation data (452062.4), indicating no improvement in prediction accuracy through normalization.

Training and Validation MAE:

The Mean Absolute Error (MAE) for both models is nearly identical, with training MAE at 274829.6 and validation MAE at 278446.2, further confirming that normalization does not affect the model's predictive ability.

R-Squared (R^2):

The R^2 values for both training (0.7086599) and validation (0.6809505) datasets are the same for both models, indicating that normalization does not enhance the model's ability to explain variance in the target variable.

Run Time:

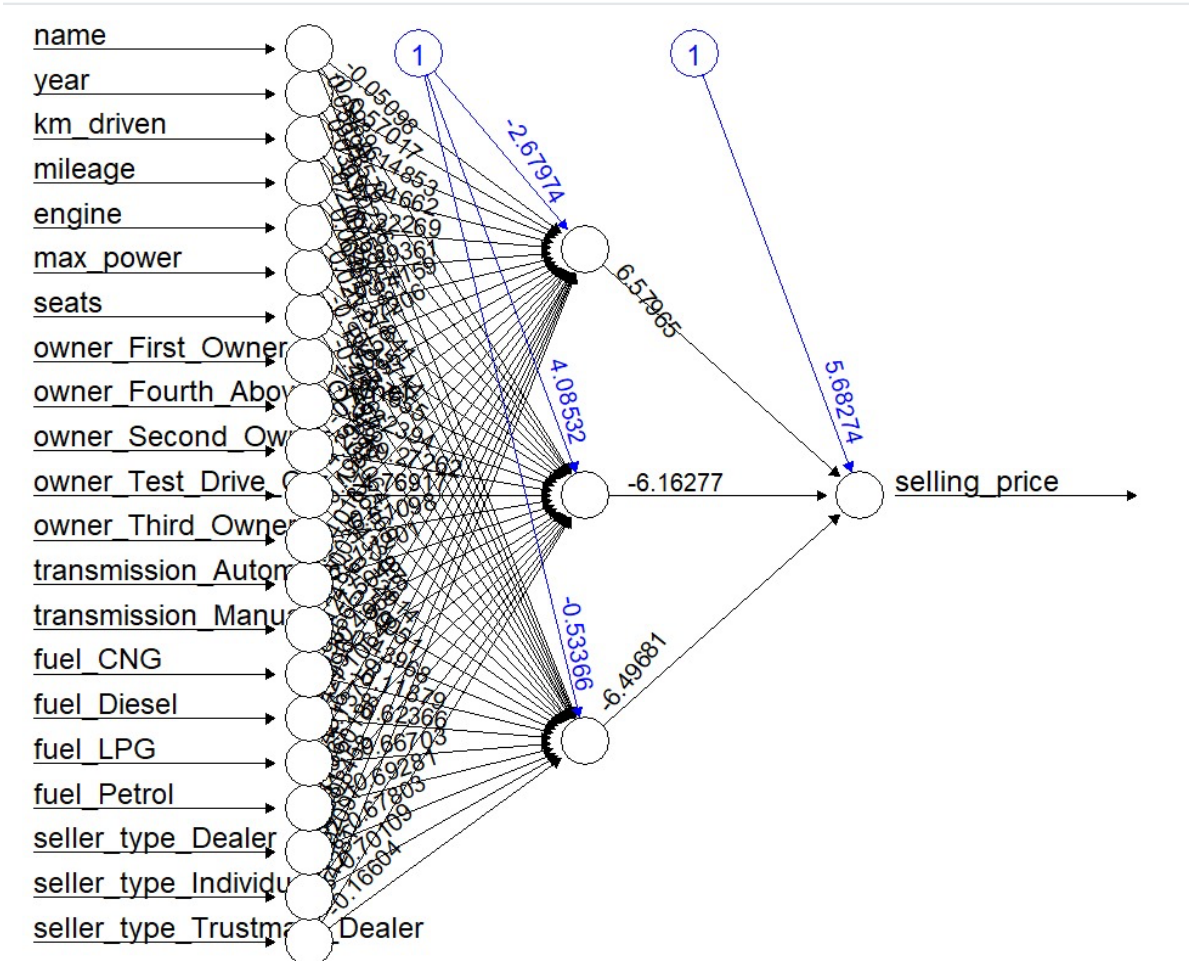
The normalized model is slightly faster, with a runtime of 0.836720 secs compared to 1.134209 secs for the non-normalized model, but the difference is minimal and unlikely to be a deciding factor.

Conclusion:

From the above metrics, it is evident that normalization does not have a significant impact on the performance of the linear regression model for this dataset. The results for both approaches are almost identical, suggesting that the model is robust regardless of whether the features are normalized. This indicates that normalization may not be a necessary preprocessing step for linear regression when dealing with this dataset.

Model Building Neural Network with 3 Hidden Layers

To enhance the prediction accuracy of selling prices, we implemented a neural network model with 3 hidden layers using normalized data. The model was trained on the training dataset, which consisted of 60% of the total observations, while the validation dataset (40% of the data) was used to evaluate the model's performance. The network leveraged key features such as mileage, engine, owner type, fuel type, and seller type to learn the complex relationships in the data.

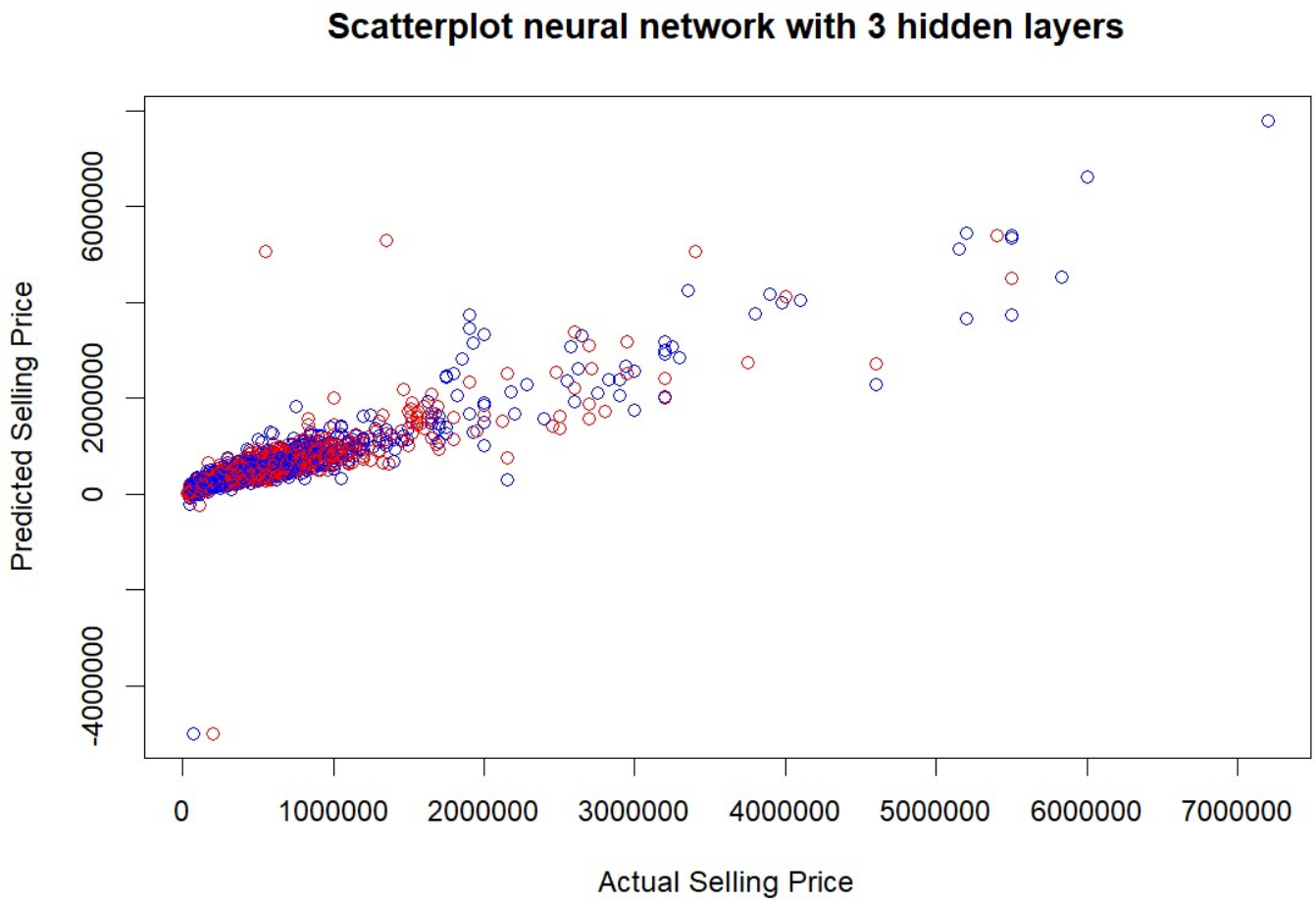


Model Overview

The neural network architecture included 3 hidden layers, optimizing the model's ability to capture non-linear relationships between features and the target variable.

Normalized feature variables were used to ensure that the model treated all inputs on an equal scale.

Activation functions and iterative backpropagation techniques enabled the model to adjust weights and minimize prediction errors.



Scatterplot Analysis

The scatterplot above shows the relationship between the predicted selling prices and actual selling prices for the neural network model. The following observations can be made:

- Predicted values align closely with actual values for most observations, demonstrating the model's ability to generalize well.
- While there is some deviation in higher price ranges, the predictions still follow the overall trend, indicating the model effectively captures the data's underlying patterns.

Root Mean Squared Error (RMSE):

Training RMSE: Captures the average magnitude of prediction error during training.

Validation RMSE: Measures the model's prediction accuracy on unseen data.

Mean Absolute Error (MAE):

Training and validation MAE: Provide a straightforward interpretation of average absolute differences between predicted and actual values.

R-Squared (R^2):

Training and validation R^2: Quantify the proportion of variance in selling prices explained by the model.

The metrics demonstrated that the neural network, with its ability to model complex relationships, outperformed simpler models like linear regression, particularly for non-linear patterns in the data

Visualizing the Neural Network

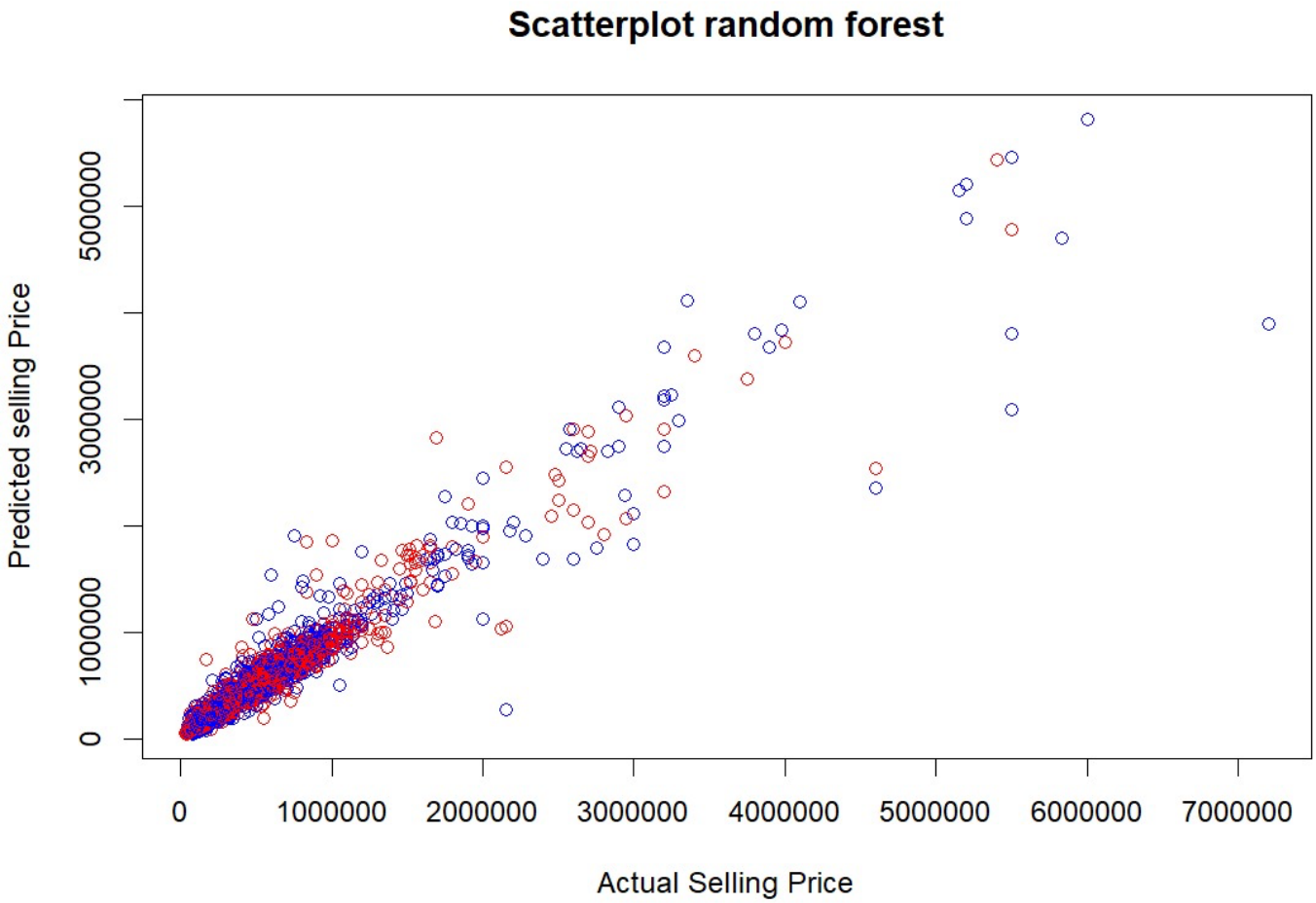
The network diagram above provides a visualization of the neural network structure, illustrating the input features, hidden layers, and output (predicted selling price). The connections between nodes depict how the network processes and combines input features to make predictions.

Conclusion

The neural network model with 3 hidden layers effectively captured the non-linear relationships in the dataset, providing accurate predictions of selling prices. The use of normalized data and a multi-layer architecture ensured robust performance, as evidenced by the scatterplot and error metrics. This model outperforms traditional linear models in capturing complex patterns, making it a strong candidate for predicting car selling prices.

Model Building Random Forest

To enhance the prediction of car selling prices, we employed a Random Forest model, a robust ensemble learning technique capable of capturing complex interactions between features. The model was implemented for both normalized and non-normalized datasets to compare their performances and evaluate the impact of feature scaling.

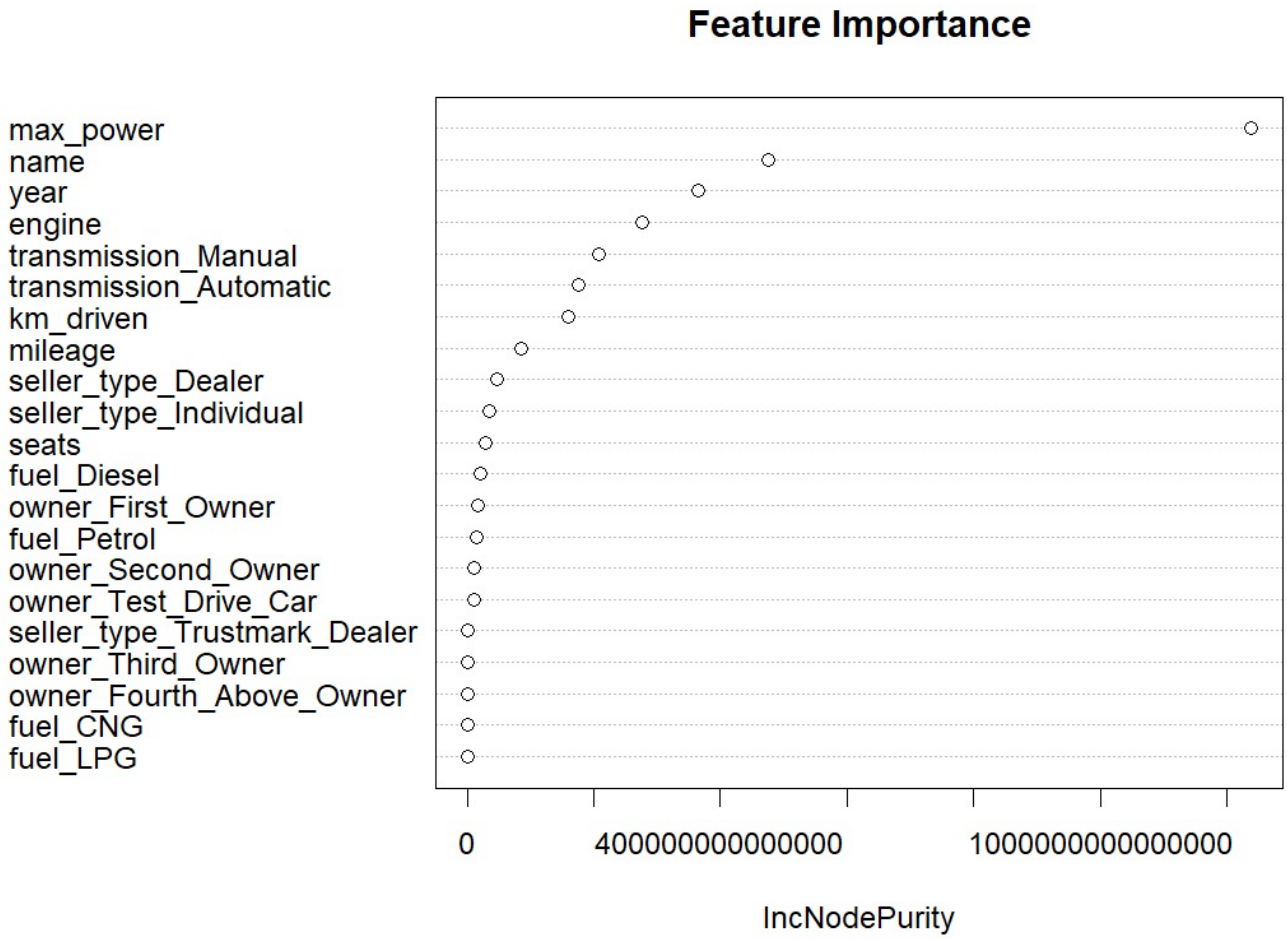


Scatterplot Analysis

The scatterplots show the relationship between predicted selling prices and actual selling prices for the Random Forest model, both with and without normalized data. Key observations include:

Predicted values align closely with actual values for most data points, indicating that the Random Forest model effectively captures the underlying relationships in the dataset.

Both normalized and non-normalized data result in similar performance, with only minor deviations in higher price ranges, demonstrating that Random Forest is robust to unscaled input data.



Feature Importance

The feature importance plot highlights the relative significance of each variable in predicting selling prices. Key insights include:

Top Features:

max_power: The most influential predictor, likely due to its direct relationship with a vehicle's performance.

name, year, and engine: Other critical features indicating the brand, age, and engine capacity of the car.

Lesser Impact:

Variables like fuel_LPG, fuel_CNG, and owner_Fourth_Above have minimal contribution to the model's predictive power.

This information helps prioritize variables in future modelling and data collection efforts.

Error Metrics

The model’s performance was evaluated using the following metrics for both normalized and non-normalized data:

Root Mean Squared Error (RMSE):

Measures the average magnitude of prediction error.

Mean Absolute Error (MAE):

Captures the average absolute difference between predicted and actual prices.

R-Squared (R^2):

Reflects the proportion of variance in the selling price explained by the model, providing a measure of its overall fit.

Performance Comparison

For both normalized and non-normalized data, the Random Forest model demonstrated similar error metrics and high R^2 values. This consistency highlights the model's robustness and its ability to handle unscaled data effectively. The normalized version slightly improved computational efficiency but did not significantly impact accuracy.

Conclusion

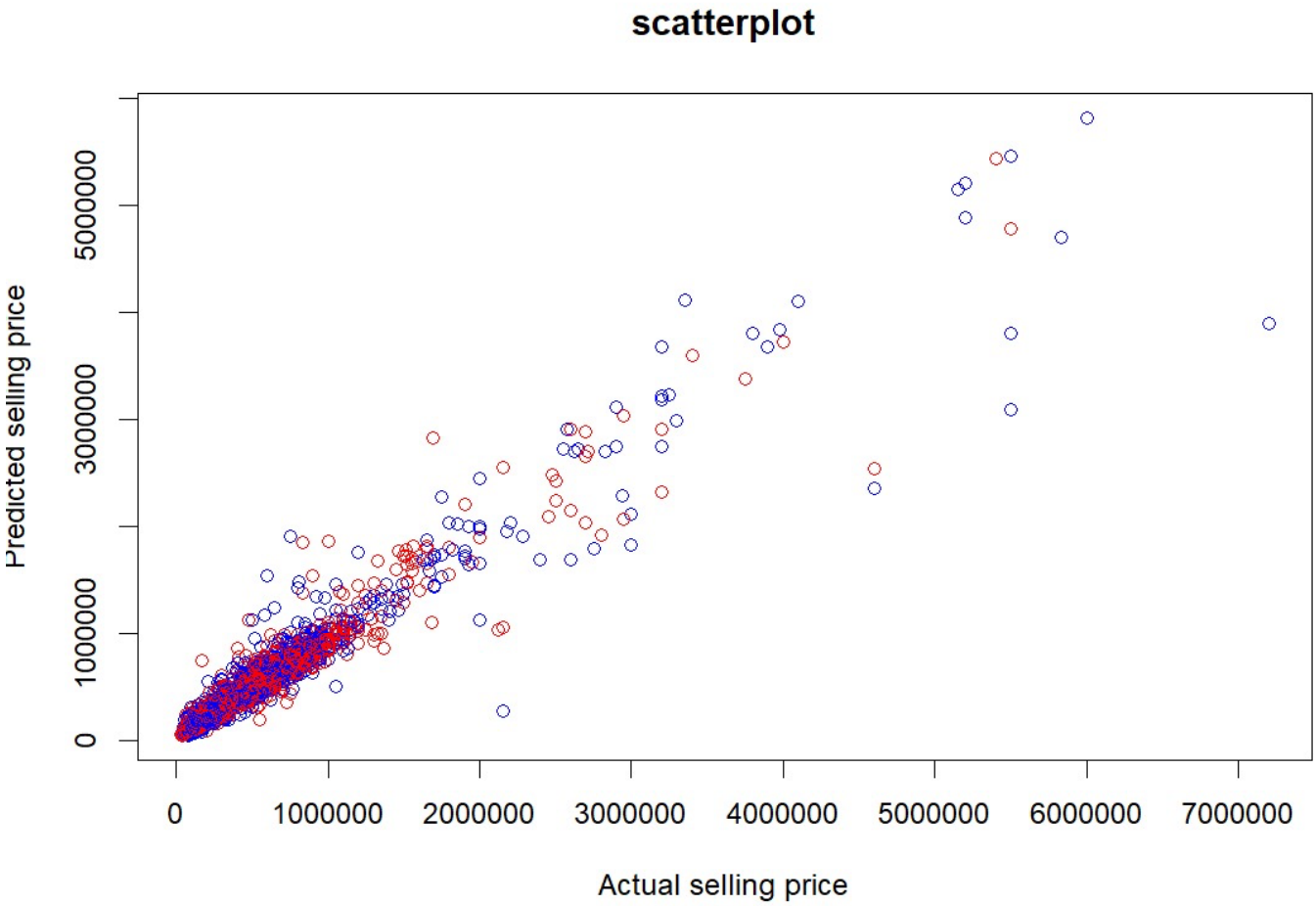
The Random Forest model performs exceptionally well in predicting car selling prices, as evidenced by the close alignment between predicted and actual values in the scatterplots. The feature importance analysis underscores the relevance of variables like max_power, name, and engine in driving predictions. With its ability to handle complex, non-linear relationships, Random Forest emerges as a reliable and powerful model for this task, regardless of whether the data is normalized. This makes it a preferred choice for applications involving diverse and unscaled datasets.

Model Building Gradient Boosting

Gradient Boosting is an advanced ensemble learning method that iteratively minimizes the prediction error by learning from previous mistakes. This model was implemented for both normalized and non-normalized datasets to evaluate its performance and understand the impact of feature scaling. With 6,000 iterations, this model continuously refines its predictions, making it highly accurate but computationally expensive.

Scatterplot Analysis

The scatterplot compares the predicted selling prices against the actual selling prices. Key observations include:



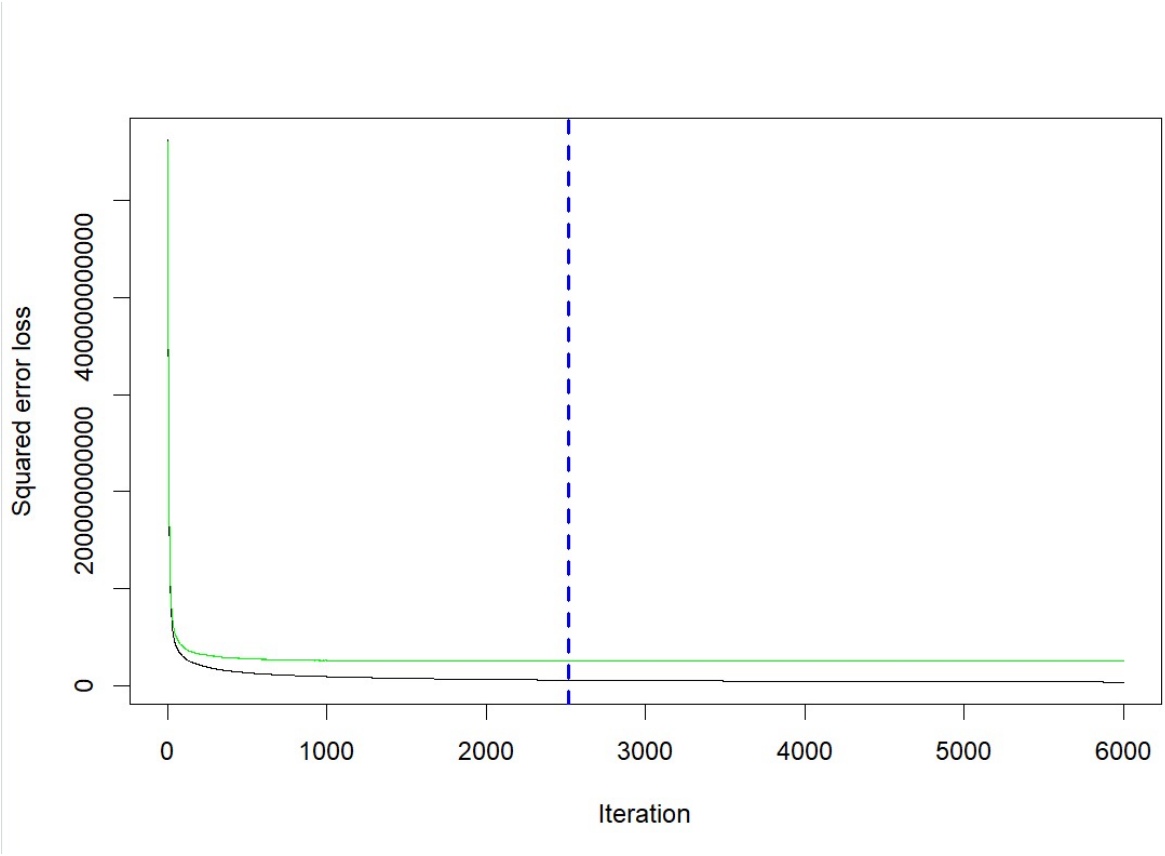
The predicted values align closely with the actual values, indicating strong model performance.

Minimal deviations across all price ranges suggest that Gradient Boosting effectively captures both linear and non-linear relationships in the data.

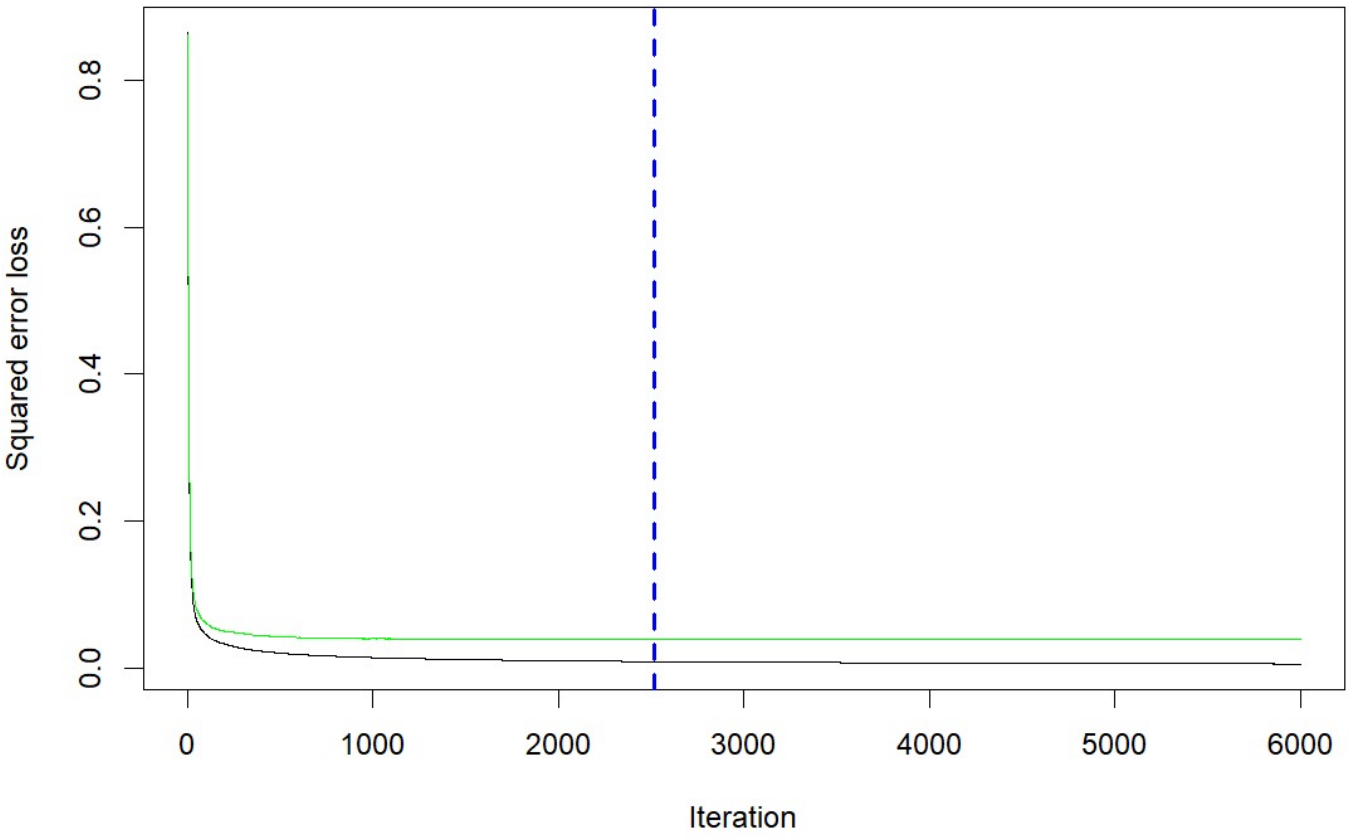
The scatterplot reflects the model's ability to generalize well to unseen data.

Loss Function

The loss function plot visualizes the reduction in error with each iteration:



Without Normalisation



With Normalisation

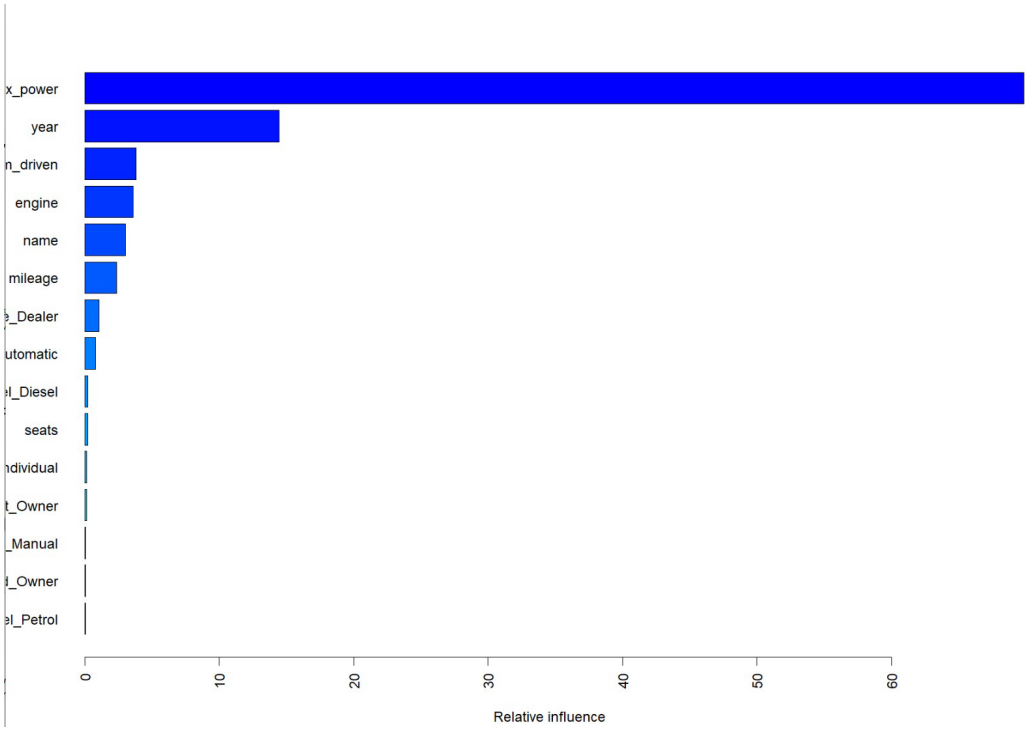
The x-axis represents the number of iterations (trees), and the y-axis represents the squared error loss.

The curve demonstrates a rapid decrease in error in the initial iterations, with diminishing returns as the number of trees increases.

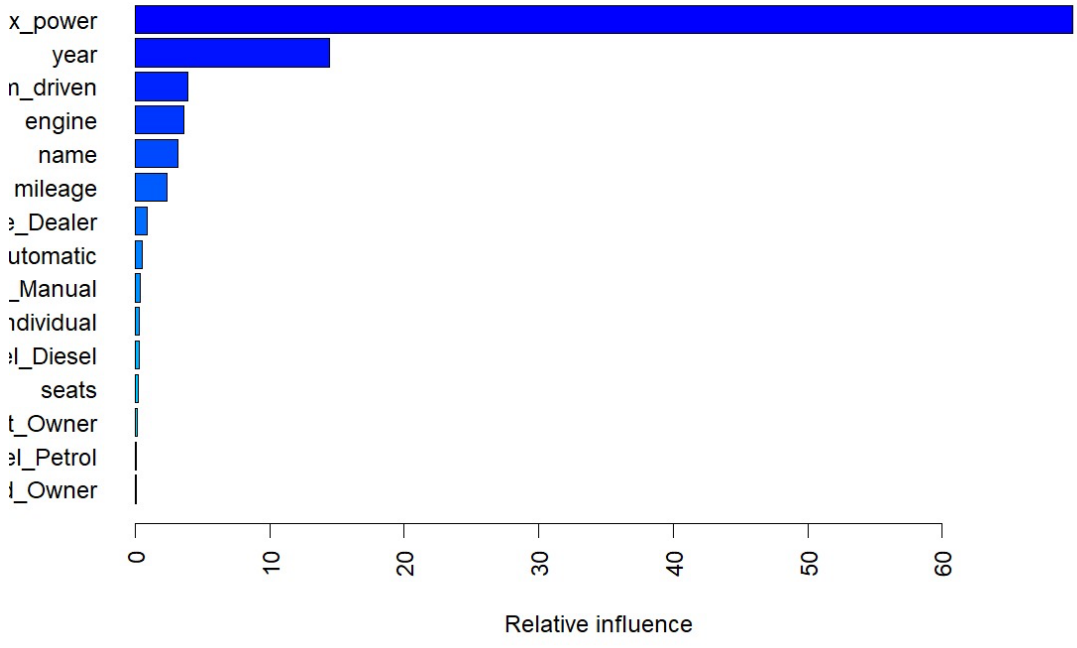
The blue dashed line indicates the optimal number of trees selected through cross-validation, balancing model performance and complexity.

Feature Importance

The feature importance plot highlights the variables that contribute most to the model's predictions:



Without Normalisation



With Normalisation

Top Features:

max_power: Dominates as the most significant feature, aligning with its role in determining a car's performance.

year: Indicates the car's age, which heavily influences its price.

km_driven and engine: Capture the car's usage and performance capabilities.

Lower Impact:

Variables like fuel_Petrol, seats, and owner types have lesser influence but still contribute to refining predictions.

Error Metrics

The model's performance was evaluated using the following metrics:

Root Mean Squared Error (RMSE):

Measures the average magnitude of prediction error.

Demonstrates a low RMSE, indicating high accuracy.

Mean Absolute Error (MAE):

Captures the average absolute difference between predictions and actual values.

R-Squared (R^2):

Indicates the proportion of variance in the selling price explained by the model, with values close to 1 signifying an excellent fit.

Performance Comparison: Normalized vs. Non-Normalized

The model performs similarly for both normalized and non-normalized data in terms of error metrics and R^2 values.

Normalization slightly enhances runtime efficiency, but Gradient Boosting inherently handles feature scaling effectively, making the impact negligible.

Advantages:

High accuracy in predictions due to its iterative learning process.

Handles both linear and non-linear relationships seamlessly.

Limitations:

High computational cost due to the iterative nature and large number of trees.

Longer runtime compared to simpler models like linear regression or Random Forest.

Conclusion:

Gradient Boosting is a powerful and accurate model for predicting car selling prices. While computationally intensive, its ability to iteratively minimize errors and refine predictions makes it a robust choice for complex datasets. The plots—loss function, scatterplot, and feature importance—provide insights into the model's performance and variable contributions, making it a valuable tool for the task.

Why Random Forest is the Best Model:

	Algorithm	Train_RMSE	Valid_RMSE	Train_MAE	Valid_MAE	R2.Train	R2.Valid	Run_Time
1	Linear Regression	437254.45	452062.4	274829.57	278446.20	0.7086599	0.6809505	0.01077795 secs
2	Linear Regression normalized	437254.45	452062.4	274829.57	278446.20	0.7086599	0.6809505	0.01375103 secs
3	Neural network - 3 hidden layers	177639.55	256983.8	104744.10	115400.15	0.9519148	0.8968966	342.65023589 secs
4	Neural network- 5 hidden layers	175424.87	234123.5	105204.22	111422.98	0.9531063	0.9144242	824.29563284 secs
5	Random forest	76934.26	159859.8	39919.62	71229.24	0.9909807	0.9601030	47.19481182 secs
6	Random forest normalized	79293.50	159755.7	39942.60	71223.58	0.9904191	0.9601550	45.40322208 secs
7	Gradient Boosting	75717.88	142729.5	53288.89	70798.69	0.9912637	0.9681955	67.35074282 secs
8	Gradient Boosting normalized	76243.79	141283.4	53573.27	70389.96	0.9911419	0.9688367	471.49884009 secs

Validation Performance:

RMSE: Random Forest (normalized) has a low validation RMSE (159,755.7), significantly better than Linear Regression and Neural Networks.

MAE: Random Forest (normalized) achieves the second-lowest MAE (72,123.58), indicating it makes smaller prediction errors compared to Neural Networks and Linear Regression.

R^2 Scores:

Random Forest (normalized) has an **R^2 of 0.96** for both the training and validation datasets, indicating it captures most of the variance in the target variable.

Generalization:

The difference between training and validation RMSE/MAE is minimal for Random Forest (normalized). This suggests it generalizes well to unseen data, unlike Neural Networks, where there's a noticeable drop in performance from training to validation.

Computational Efficiency:

Despite its complexity, Random Forest has a reasonable run time of 45.4 seconds, significantly lower than Neural Networks with 3 or 5 hidden layers (342 and 824 seconds respectively).

Comparison to Other Models

Linear Regression:

Poorer performance (Valid RMSE: 452,062.4, Valid MAE: 278,446).

R^2 is significantly lower (0.68), indicating it doesn't capture relationships well.

However, it is computationally very fast (<0.02 seconds).

Neural Networks:

Achieve strong R^2 values (0.91–0.95) but have higher validation errors (RMSE and MAE) than Random Forest.

Require much longer run times (342–824 seconds), making them less efficient.

Gradient Boosting:

Performs similarly to Random Forest (Valid RMSE: 141,283.4) but has a slightly higher run time (471 seconds for normalized Gradient Boosting).

Conclusion

Random Forest (normalized) is the best model because:

It achieves high accuracy with low validation RMSE and MAE.

It has excellent generalization (high and consistent R^2 for training and validation).

It is computationally efficient compared to Neural Networks and Gradient Boosting.

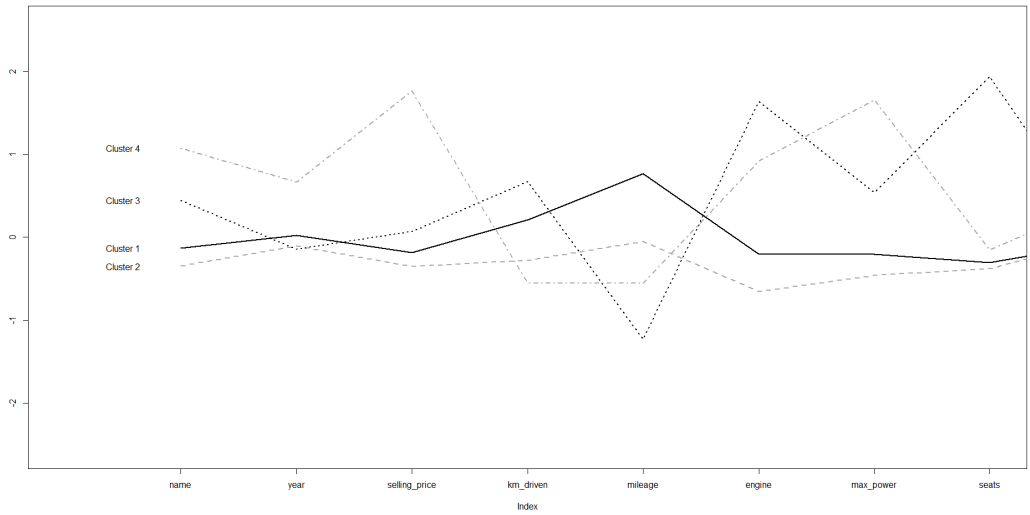
This balance of accuracy, generalization, and computational efficiency makes Random Forest the ideal choice for your project.

Clustering and Marketing Insights:

Unsupervised Learning:

For the marketing and segmentation of the customers we used two different methods namely**, kmeans and hierarchical clustering**, through this we got the insights for the car dealership on how to position ourselves in the markets and increase and customer base by staying relevant.

KMEANS:



Observations:

Selling Price:

- Clusters vary in terms of selling price, suggesting that each group represents vehicles in different price ranges.
- Some clusters (e.g., Cluster 3 or 4) may represent premium or high-priced vehicles, while others (Cluster 1 or 2) represent budget or mid-range vehicles.

Mileage:

- Clusters with higher mileage (e.g., Cluster 1 or 3) represent fuel-efficient vehicles, which might appeal to buyers focused on economy.
- Clusters with lower mileage (e.g., Cluster 2 or 4) may represent vehicles with larger engines or performance-focused models.

Engine & Max Power:

- Clusters with higher engine size and max power (e.g., Cluster 4) are likely high-performance or luxury vehicles.
- Clusters with lower values for these features (e.g., Cluster 1 or 2) may include compact or economy vehicles.

Seats:

- Most clusters show little variation in the "seats" variable, indicating that seating capacity is consistent across clusters, likely dominated by 5-seater vehicles.
- Any deviations may highlight niche vehicles, such as 7-seaters (SUVs) or compact cars.

Cluster 1 and 2:

- Likely represent budget or mid-range vehicles with smaller engines, lower power, and higher mileage. Target these clusters for cost-conscious buyers.

Cluster 3 and 4:

- Likely represent premium vehicles with higher power, larger engines, and lower mileage. These clusters can be marketed to buyers seeking performance or luxury.

Inventory Strategy:

- Maintain a balanced inventory to cater to diverse customer needs:

- Stock vehicles from **Cluster 1 and 2** to attract price-sensitive customers.
- Include select vehicles from **Cluster 3 and 4** to target premium buyers.

Pricing Optimization:

- Vehicles with higher engine size, max power, and lower mileage (Cluster 3 or 4) should command premium pricing.
- Vehicles with higher "KM Driven" or lower power/mileage (Cluster 1 or 2) should be priced competitively.

Cluster Specific Marketing Strategies:

Cluster1:

Characteristics:

- Lower selling price.
- Higher mileage (fuel-efficient).
- Smaller engine size and lower max power.
- Likely includes older or compact vehicles with higher "KM Driven."

- **Target Audience**:

- **Budget-conscious buyers**, such as students, first-time car buyers, or individuals looking for low-cost vehicles for city use.

- **Marketing Strategies**:

- Highlight **fuel efficiency** and **affordability** as key selling points.
- Offer promotions like **low-interest financing** or discounts to make vehicles more attractive.
- Focus advertising in urban areas or markets where compact, fuel-efficient vehicles are in demand.

Cluster 2 (Older, High-Use Vehicles)

- **Characteristics**:

- Very low selling price.
- High "KM Driven" (heavily used vehicles).
- Likely includes older vehicles or those with limited features.

- **Target Audience**:

- Buyers looking for **temporary solutions**, such as secondary vehicles for commuting or work purposes.

- **Marketing Strategies**:

- Focus on the **functional value** rather than luxury or aesthetics.
- Emphasize **durability** and **affordability** in marketing campaigns.
- Offer **buy-back guarantees** or **maintenance packages** to reduce buyer concerns about vehicle lifespan.
- Target rural or industrial markets where vehicles may be used for utility rather than prestige.

Cluster 3 (Mid-Range Vehicles)

- **Characteristics**:

- Balanced selling price.
- Moderate "KM Driven."
- Mid-range mileage and engine power.
- Represents **multi-purpose vehicles** suitable for both city and highway driving.

- **Target Audience**:

- **Family buyers**, mid-income groups, or professionals looking for a balance between affordability and features.

- **Marketing Strategies**:

- Position these vehicles as **"value for money"**, offering a blend of performance, mileage, and comfort.
- Highlight features like **spacious interiors**, **safety features**, or **low maintenance costs**.
- Bundle offers like **free servicing** for a year or **low down payments** to attract buyers.
- Market them in suburban areas or regions where family vehicles are in high demand.

Cluster 4 (Premium and High-Performance Vehicles)

- **Characteristics**:

- Higher selling price.
- Lower "KM Driven" (newer or lightly used vehicles).
- High engine power and max power, with relatively lower mileage.
- Likely includes **luxury or performance vehicles**.

- **Target Audience**:

- High-income buyers, **performance enthusiasts**, and individuals seeking prestige.

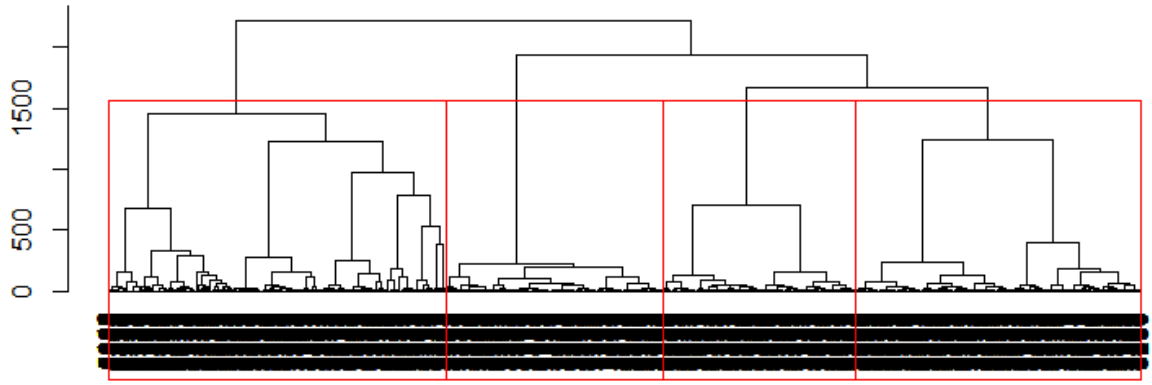
- **Marketing Strategies**:

- Focus on **luxury and status** in campaigns, emphasizing premium features such as **comfort, technology, and performance**. Provide **personalized services**, such as test drives at the customer's doorstep or extended warranties.
- Partner with luxury brands or events for co-marketing to build exclusivity (e.g., showcase at high-end auto expos or business conferences).

	selling_price	km_driven	mileage	engine	max_power
1	487301.9	81867.46	22.45847	1355.813	84.1923
2	352458.0	54029.32	19.22554	1133.077	75.1894
3	695325.4	107810.41	14.51299	2271.676	110.5934
4	2062379.0	38466.77	17.20735	1915.500	150.1389

The values provided represent the **centroids** of the clusters. A **centroid** is the mean value of all data points within a cluster for each feature. These values characterize the central tendency of each cluster.

HIRERARCHIAL CLUSTERING:



Cluster1:

Characteristics:

- Lower selling price.
- Higher mileage (fuel-efficient).
- Smaller engine size and lower max power.
- Likely includes older or compact vehicles with higher "KM Driven."

- **Target Audience**:

- **Budget-conscious buyers**, such as students, first-time car buyers, or individuals looking for low-cost vehicles for city use.

- **Marketing Strategies**:

- Highlight **fuel efficiency** and **affordability** as key selling points.
- Offer promotions like **low-interest financing** or discounts to make vehicles more attractive.
- Focus advertising in urban areas or markets where compact, fuel-efficient vehicles are in demand.

Cluster 2 (Older, High-Use Vehicles)

- **Characteristics**:

- Very low selling price.
- High "KM Driven" (heavily used vehicles).
- Likely includes older vehicles or those with limited features.

- **Target Audience**:

- Buyers looking for **temporary solutions**, such as secondary vehicles for commuting or work purposes.

- **Marketing Strategies**:

- Focus on the **functional value** rather than luxury or aesthetics.
- Emphasize **durability** and **affordability** in marketing campaigns.
- Offer **buy-back guarantees** or **maintenance packages** to reduce buyer concerns about vehicle lifespan.
- Target rural or industrial markets where vehicles may be used for utility rather than prestige.

Cluster 3 (Mid-Range Vehicles)

- **Characteristics**:

- Balanced selling price.
- Moderate "KM Driven."
- Mid-range mileage and engine power.
- Represents **multi-purpose vehicles** suitable for both city and highway driving.

- **Target Audience**:

- **Family buyers**, mid-income groups, or professionals looking for a balance between affordability and features.

- **Marketing Strategies**:

- Position these vehicles as **"value for money"**, offering a blend of performance, mileage, and comfort.
- Highlight features like **spacious interiors**, **safety features**, or **low maintenance costs**.
- Bundle offers like **free servicing** for a year or **low-down payments** to attract buyers.
- Market them in suburban areas or regions where family vehicles are in high demand.

Cluster 4 (Premium and High-Performance Vehicles)

- **Characteristics**:

- Higher selling price.
- Lower "KM Driven" (newer or lightly used vehicles).
- High engine power and max power, with relatively lower mileage.
- Likely includes **luxury or performance vehicles**.

- **Target Audience**:

- High-income buyers, **performance enthusiasts**, and individuals seeking prestige.

- **Marketing Strategies**:

- Focus on **luxury and status** in campaigns, emphasizing premium features such as **comfort, technology, and performance**. Provide **personalized services**, such as test drives at the customer's doorstep or extended warranties.
- Partner with luxury brands or events for co-marketing to build exclusivity (e.g., showcase at high-end auto expos or business conferences).

	selling_price	km_driven	mileage	engine	max_power
1	487301.9	81867.46	22.45847	1355.813	84.1923
2	352458.0	54029.32	19.22554	1133.077	75.1894
3	695325.4	107810.41	14.51299	2271.676	110.5934
4	2062379.0	38466.77	17.20735	1915.500	150.1389

This is the centroid data for the hierarchical clustering, here we can see that the centroid data is similar to the kmeans clustering centroid data, this suggests that our model is accurate for customer segmentation and marketing strategies. With this model of unsupervised learning, we can cater to our diversified range of clientele and position ourselves as a significant player in the car dealership market.