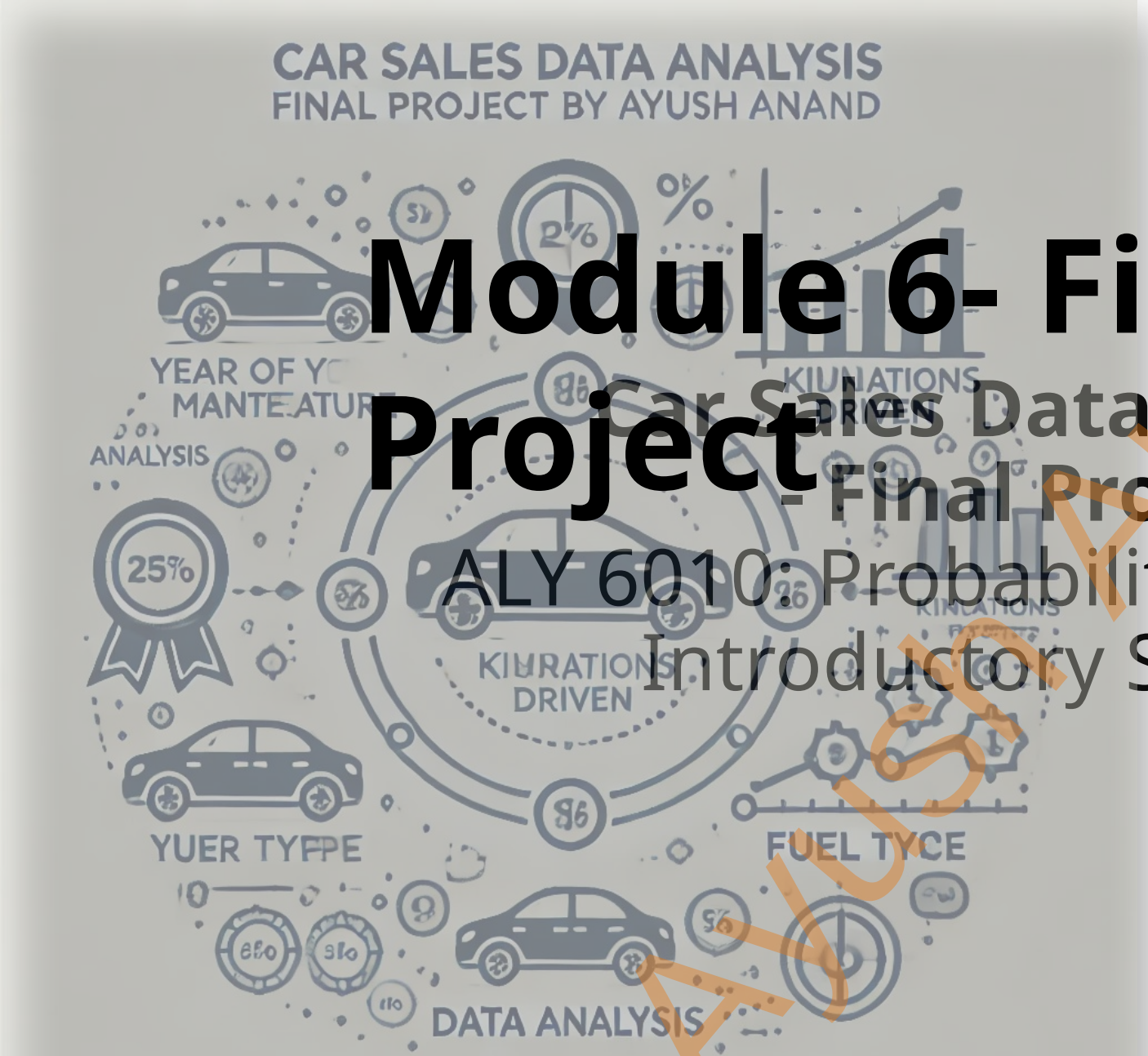


Northeastern University
College of Professional Studies

Module 6- Final Project

Car Sales Data Analysis - Final Project

ALY 6010: Probability Theory and Introductory Statistics

Table of Content

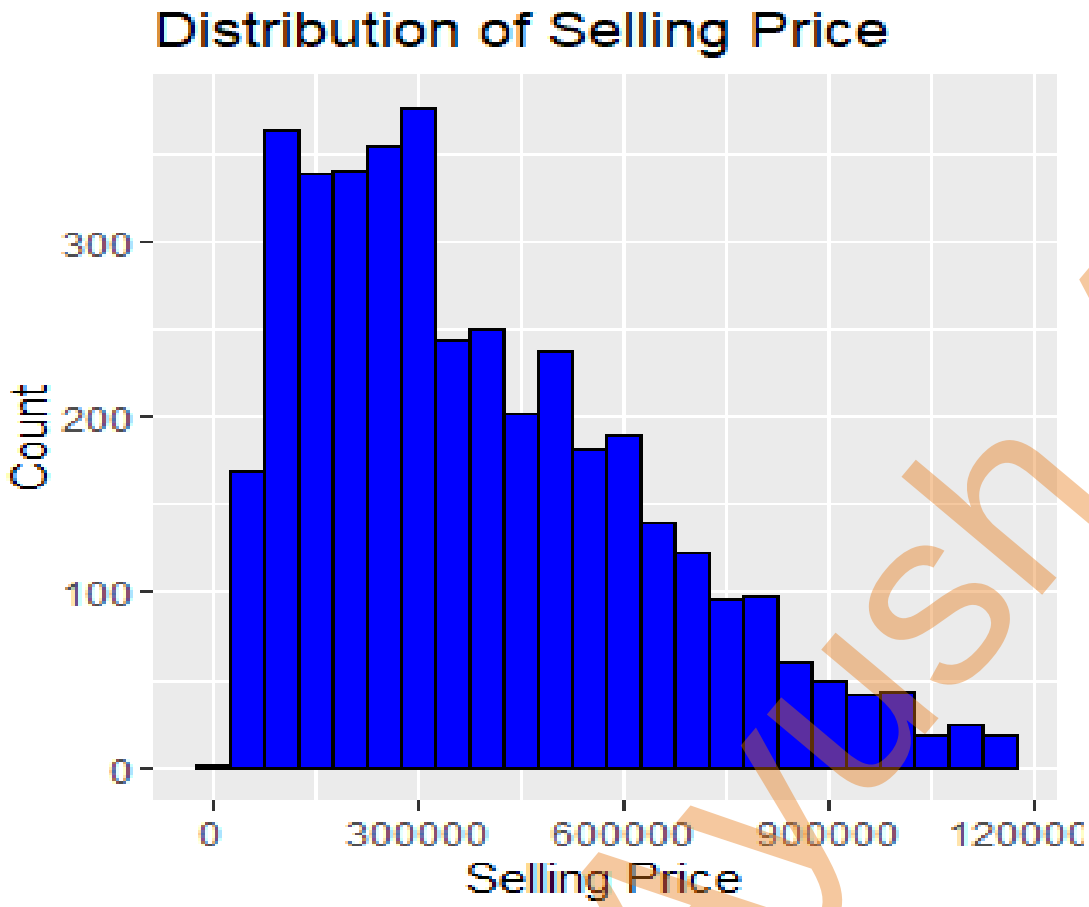
About the Dataset

This report provides a comprehensive analysis of car sales data, with a focus on understanding the relationships between **selling price** and various factors such as the **manufacturing year**, **kilometers driven**, and **fuel type**. Linear regression and hypothesis testing were employed to investigate these relationships, with a final evaluation of model accuracy and robustness through error metrics and cross-validation.

Variable	Description
name	The model name of the car (character)
year	The year of manufacture (integer)
selling_price	Selling price of the car in Indian Rupees (₹) (integer)
km_driven	Kilometers driven by the car (integer)
fuel	Fuel type of the car (Petrol or Diesel) (character)
seller_type	Type of seller, e.g., Individual or Dealer (character)
transmission	Transmission type of the car (e.g., Manual or Automatic) (character)
owner	Ownership status (e.g., First Owner, Second Owner) (character)

The dataset contains **3962 observations** and **8 variables** described below:

Exploratory Data Analysis (EDA)

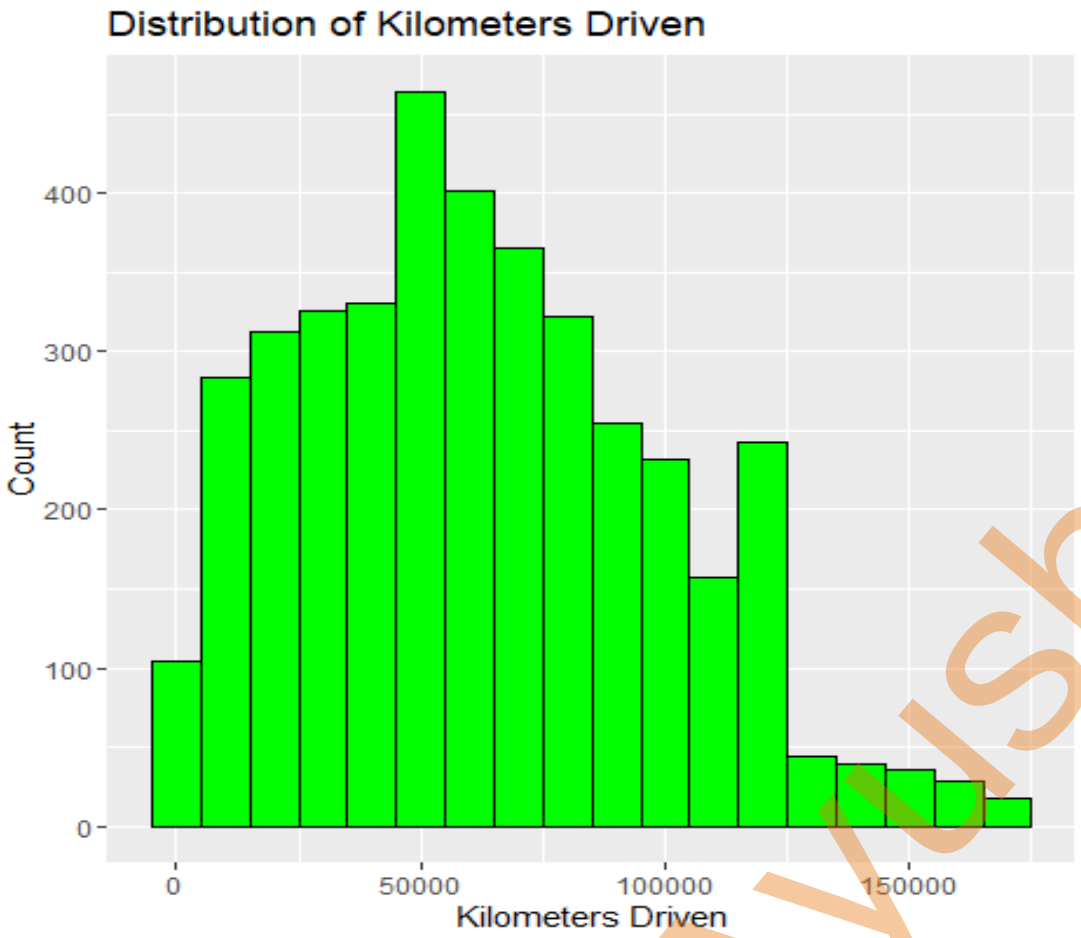


A histogram was plotted to understand the spread and frequency of car prices.

The plot uses a bin width of 50,000 units to group car prices into intervals. The bars are filled in blue with black borders.

Purpose: This histogram helps to understand the distribution of car prices, identify common price points, and highlight any outliers.

Key Insights: **Selling Price Distribution**: Most cars fall within a price range of ₹200,000 to ₹550,000.



A histogram was used to visualize the distribution of kilometers driven

The plot uses a bin width of 10,000 units, with bars filled in green and bordered in black.

Purpose: This histogram helps to identify the most common mileage ranges and potential outliers in the dataset.

Key Insights:

Kilometers Driven Distribution: Majority of cars have mileage below 100,000 km.

Key Questions Explored

Does the year of manufacture significantly affect the car's selling price?

- **Why This Question?** Year of manufacture often plays a key role in determining car value, as newer models tend to have updated features and improved reliability. It is important to understand whether the age of the car is a strong determinant of its price.

Does the number of kilometers driven significantly impact the car's selling price?

- **Why This Question?** Kilometers driven is an indicator of how much a car has been used. High mileage often leads to lower value due to wear and tear. Understanding the effect of mileage helps in assessing how usage impacts car pricing.

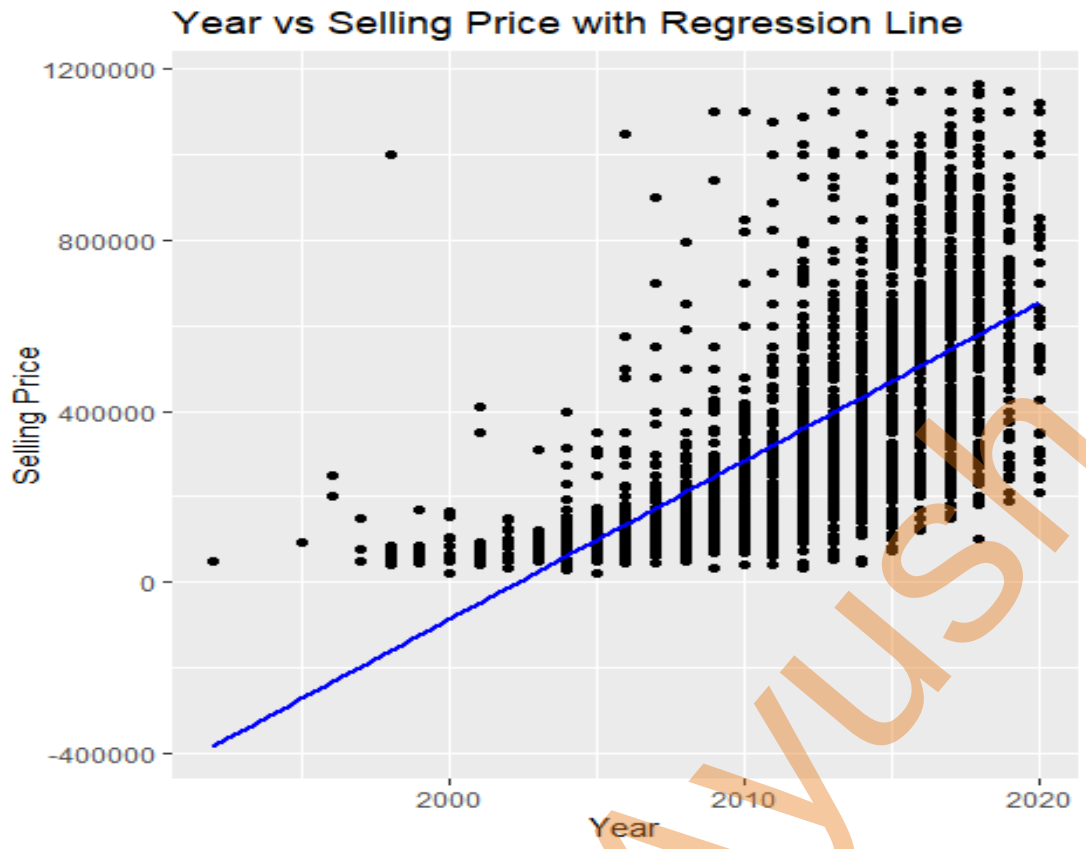
Does the fuel type (Diesel vs. Petrol) significantly affect the selling price of cars?

- **Why This Question?** Initial observations indicated a price difference between Diesel and Petrol cars. Diesel cars are often considered more fuel-efficient, which could make them more attractive to buyers. Analyzing the effect of fuel type helps understand buyer preferences and market trends.

Northeastern University
College of Professional Studies

Hypothesis testing and Regression Analysis

1. Regression Analysis of Year vs. Selling Price



Interpretation:

The positive year coefficient (37,040) suggests that each additional year increases the selling price by an average of ₹37,040.

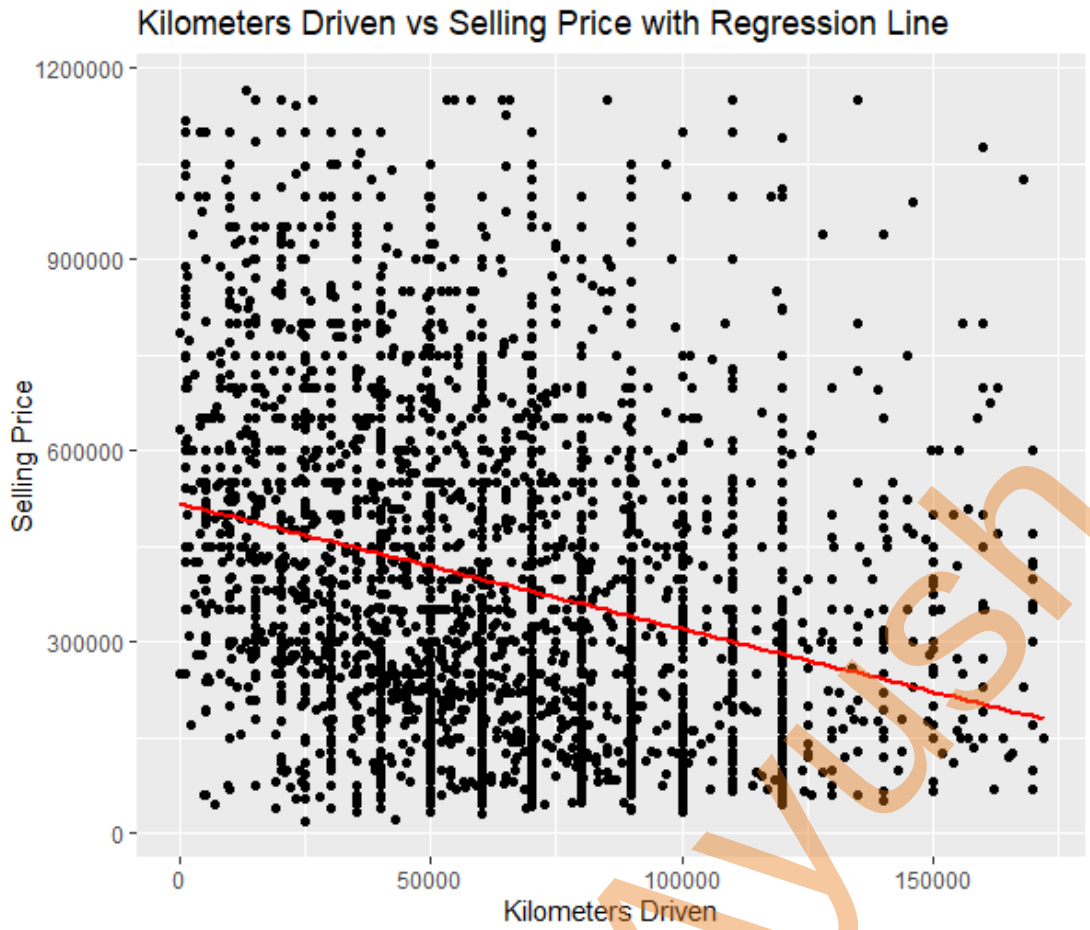
An R-squared value of 0.3971 indicates that approximately 39.71% of the variability in selling price is explained by the year of manufacture.

Hypothesis Test

Null Hypothesis (H_0): $\beta_1 \leq 0$
(The year of manufacture does not have a significant positive impact on the selling price, or has no impact or a negative impact.)
Alternative Hypothesis (H_1): $\beta_1 > 0$
(The year of manufacture has a significant positive impact on the selling price.)

Result: With a p-value below 0.05, we reject the null hypothesis, confirming that there is a statistically significant positive impact of the year of manufacture on the selling price

2. Regression Analysis of Kilometers Driven vs. Selling Price



Interpretation:

The negative coefficient (-1.962) suggests that for every additional kilometer driven, the selling price decreases on average by ₹1.96.

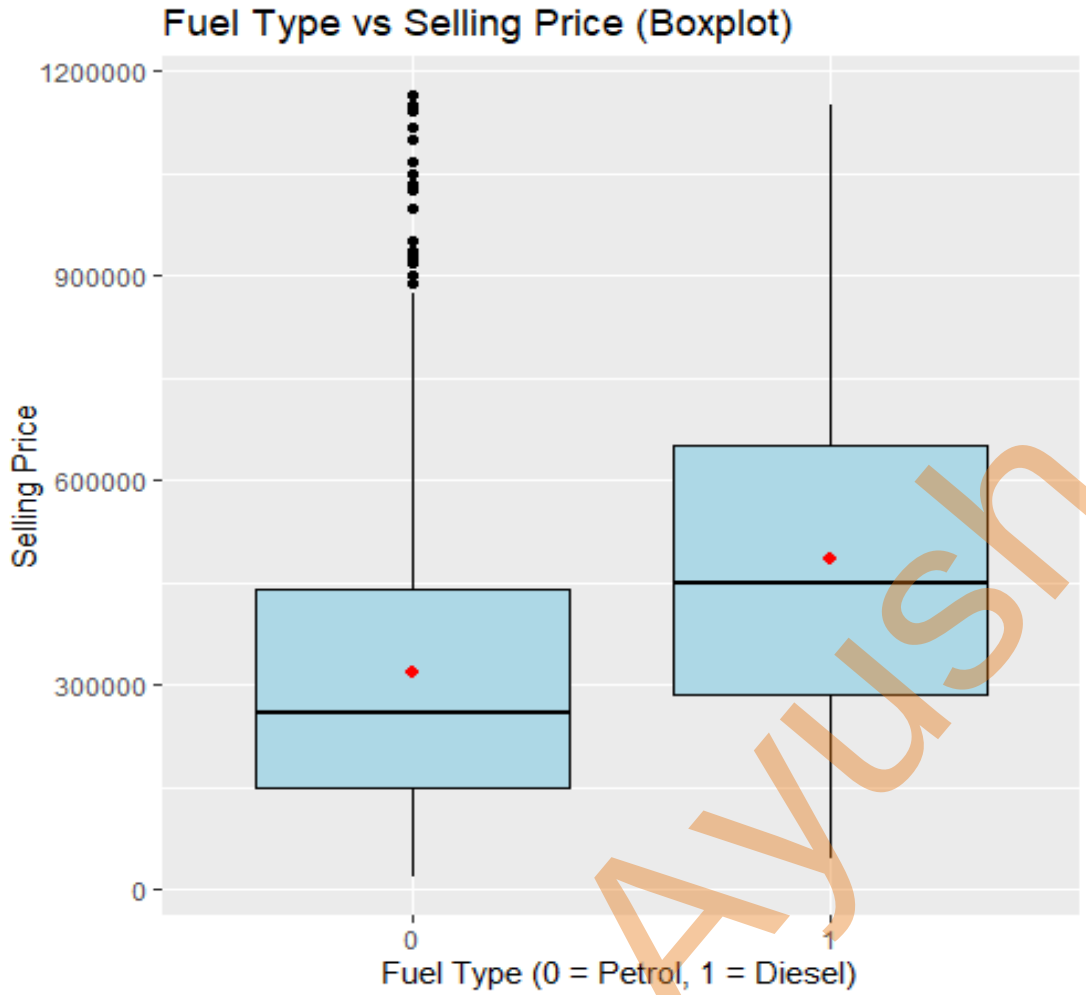
An R-squared value of 0.08176 suggests that only 8.18% of the variability in selling price is explained by kilometers driven.

Hypothesis Test

Null Hypothesis (H_0): $\beta_1 \geq 0$
(The kilometers driven does not have a significant negative impact on the selling price, or has no impact or a positive impact.)
Alternative Hypothesis (H_1): $\beta_1 < 0$
(The kilometers driven has a significant negative impact on the selling price.)

Result: With a p-value below 0.05, we reject the null hypothesis, indicating a statistically significant negative impact of kilometers driven on the selling price.

3. Fuel Type Comparison: Diesel vs. Petrol Cars



Interpretation:

The **Fuel Binary Coefficient** of 167,900 indicates that Diesel cars are predicted to sell for an average of ₹167,900 more than Petrol cars.

Null Hypothesis (H_0): $\mu_1 \leq \mu_2$
(The mean selling price of Diesel cars (μ_1) is less than or equal to the mean selling price of Petrol cars (μ_2), indicating no significant difference or Diesel being lower.)

Alternative Hypothesis (H_1): $\mu_1 > \mu_2$
(The mean selling price of Diesel cars (μ_1) is significantly greater than the mean selling price of Petrol cars (μ_2).)

Model Performance Evaluation

Root Mean Squared Error (RMSE): 238,834.9

RMSE measures the average deviation between predicted and actual selling prices.

Interpretation: A lower RMSE indicates better model accuracy.

Cross-Validation

To validate model stability, a **10-fold cross-validation** was performed on a simplified model, yielding:

Cross-Validation Error: 54,649,936,551

This error indicates the model's prediction variance across different subsets of the data, assessing its robustness.

Predicted vs. Actual Prices

A comparison of actual vs. predicted values from the test set reveals how closely the model predictions align with the real selling prices:

Actual Price (₹)	Predicted Price (₹)
600,000	482,837.2
140,000	314,937.2
600,000	482,837.2
250,000	314,937.2
750,000	482,837.2
160,000	314,937.2

This comparison illustrates the model's prediction accuracy, highlighting areas for potential improvement.

Conclusion

This analysis explores car sales data to identify the factors influencing car selling prices. The dataset includes 3,962 car records with 8 features, such as year of manufacture, kilometers driven, and fuel type. Through regression analysis and hypothesis testing, it was found that newer cars, lower mileage, and Diesel fuel type significantly increase selling prices. Regression models indicate that year and fuel type are positively correlated with selling price, while higher kilometers driven negatively impacts value. Model accuracy, evaluated through RMSE and cross-validation, suggests reasonable prediction performance, with opportunities for improvement in predicting complex interactions. The analysis provides actionable insights for car dealers, such as focusing on newer Diesel vehicles for maximizing profit. Future analyses could incorporate additional attributes, such as car brand, model, and condition, to improve predictive power and better understand the determinants of car pricing.

Northeastern University
College of Professional Studies

References

Vehicle Dataset from CarDekho [Data set]. Retrieved [Date Retrieved] from Kaggle. https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho

Motorcycle Dataset [Data set]. Retrieved [Date Retrieved] from Kaggle. https://www.kaggle.com/datasets/nehalbirla/motorcycle-dataset

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). Applied Linear Statistical Models (5th ed.). McGraw-Hill Irwin.

Thank you!