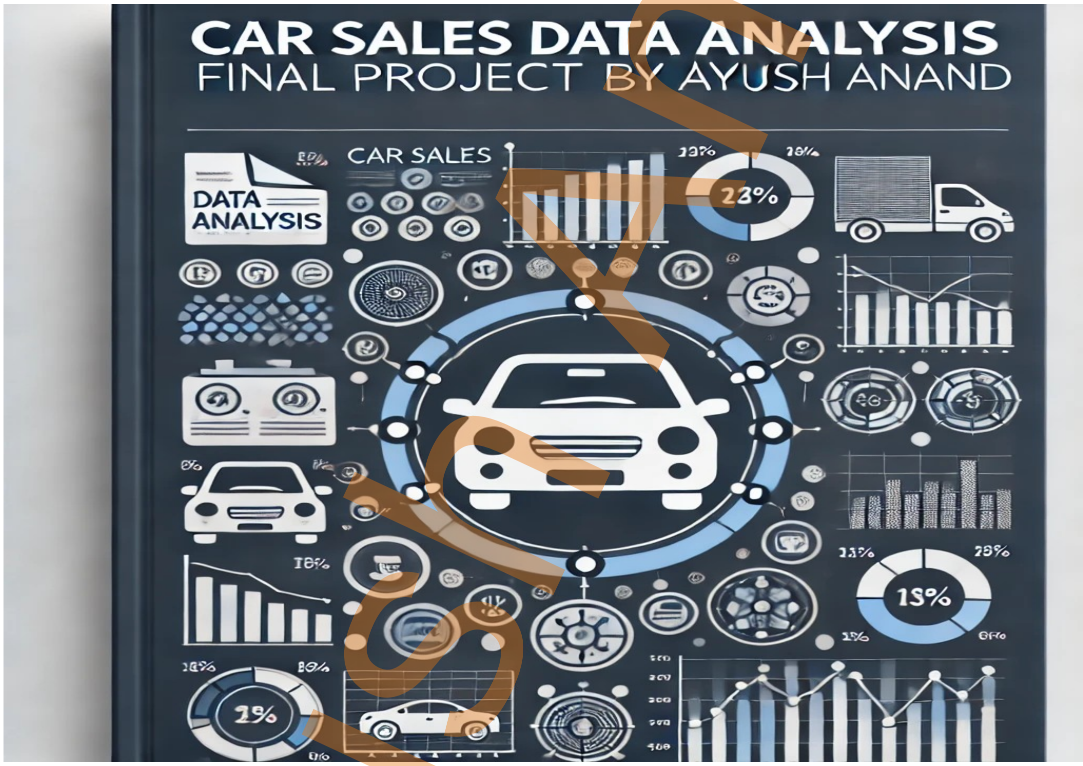
Final project

Ayush Anand



NORTHEASTERN UNIVERSITY: COLLEGE OF PROFESSIONAL STUDIES
ALY 6010.71820: PROBABILITY THEORY AND INTRODUCTORY STATISTICS
PROFESSOR XYZ
OCTOBER 31ST, 2024

Car Sales Data Analysis Report

Overview

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.

Dataset Summary

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:

Summary Statistics

The following table provides an overview of the key numerical variables in the dataset:

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Year	1992	2011	2014	2013	2016	2020
Selling Price (₹)	20,000	200,000	345,000	393,306	550,000	1,165,000
Kilometers Driven	1	35,000	60,000	63,031	90,000	172,000

Additional Notes:

- The categorical variables include fuel type, seller type, transmission, and owner.
- Data types were verified using the str(car_data) function, ensuring that each variable was appropriately assigned for analysis.

Data Structure

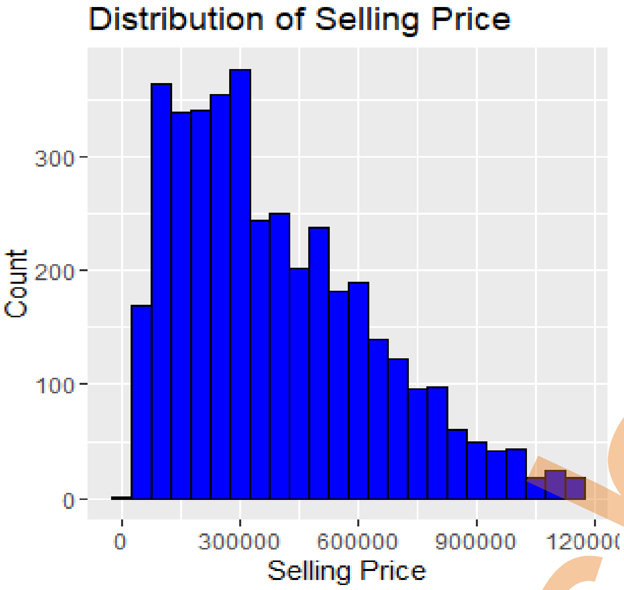
Using str(car_data), we confirmed that the data types for each variable are correctly assigned:

- name, fuel, seller_type, transmission, and owner are **character**.
- year, selling_price, and km_driven are **integer**.

Exploratory Data Analysis (EDA)

The EDA aimed to understand the distribution of key variables and identify potential patterns.

1. **Distribution of Selling Price:**

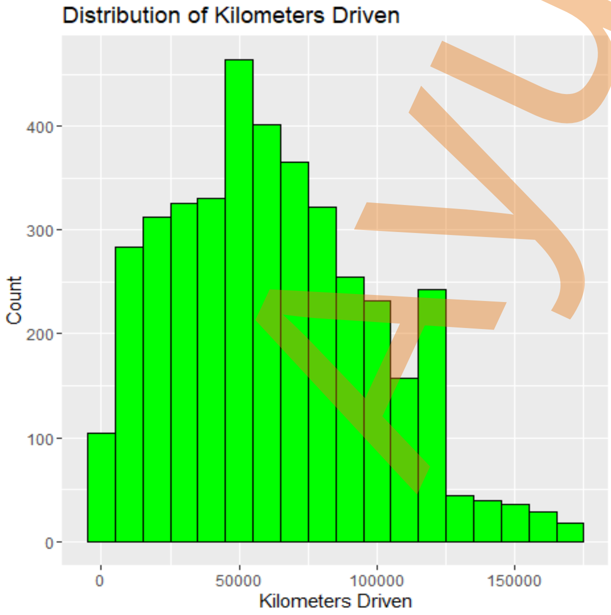


A histogram was plotted to understand the spread and frequency of car prices.

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.

2. **Distribution of Kilometers Driven**:



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.

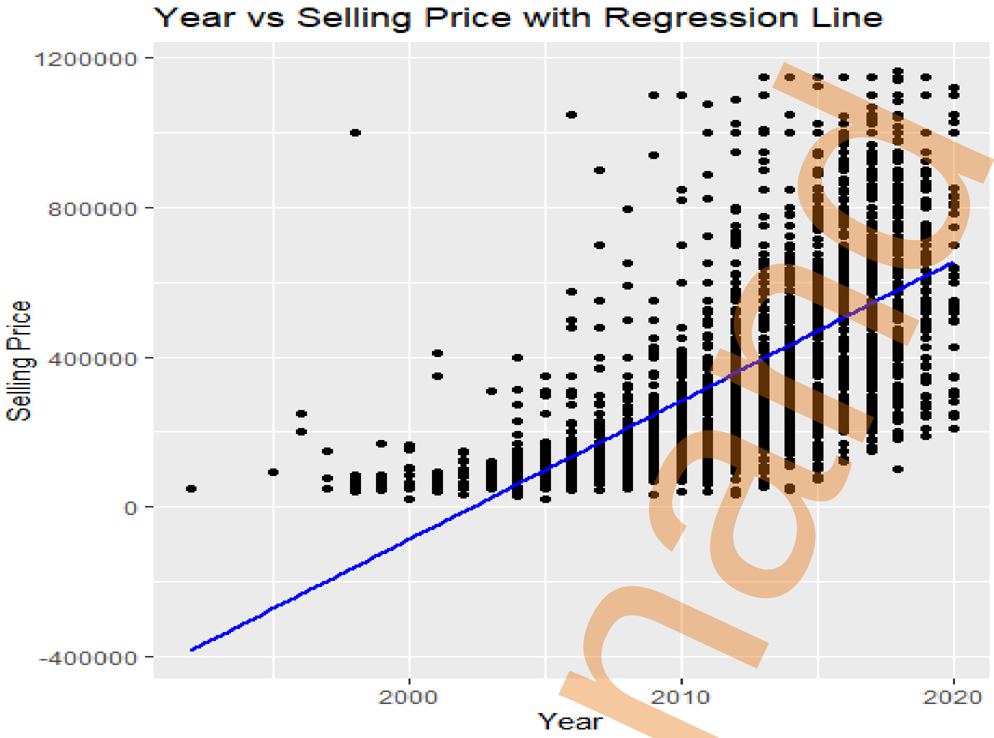
Key Questions Explored

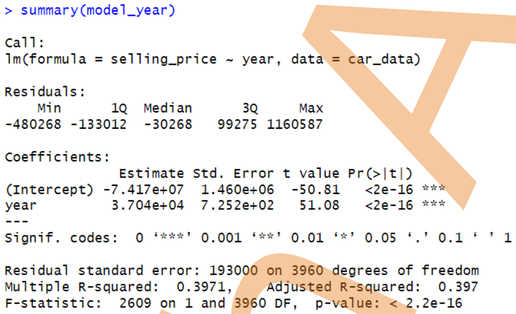
1. **Does the year of manufacture significantly affect the car's selling price?**
   - **Why This Question?** Year of manufacture often plays a key role in determining car value, as newer models tend to have updated features and improved reliability. It is important to understand whether the age of the car is a strong determinant of its price.
2. **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.
3. **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.

Regression Analysis

1. Regression Analysis of Year vs. Selling Price

To determine if the year of manufacture significantly affects selling price, a linear regression model was fitted with selling price as the dependent variable and year as the independent variable.





```
> summary(model_year)

Call:
lm(formula = selling_price ~ year, data = car_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-480268 -133012  -30268   99275 1160587 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.417e+07  1.460e+06  -50.81   <2e-16 ***
year         3.704e+04  7.252e+02   51.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 193000 on 3960 degrees of freedom
Multiple R-squared:  0.3971,	Adjusted R-squared:  0.397 
F-statistic:  2609 on 1 and 3960 DF,  p-value: < 2.2e-16
```

Model Summary:

Statistic	Value
Intercept	- 74,170,000
Year Coefficient	37,040
R-squared	0.3971
Adjusted R-squared	0.397
F-statistic	2,609
p-value	< 2.2e-16

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.

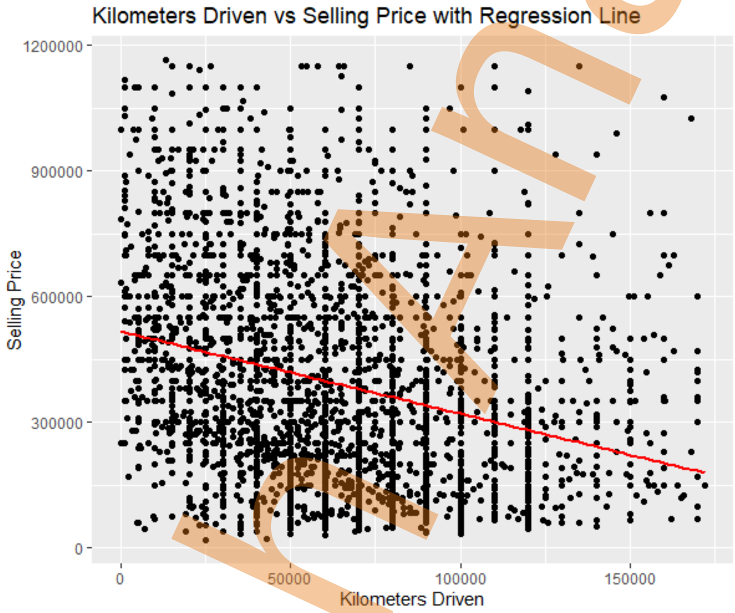
Hypothesis Test

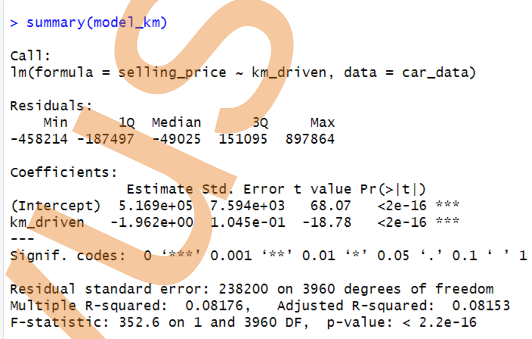
- **Null Hypothesis (H0)**: The year of manufacture has no significant positive impact on the selling price.
- **Alternative Hypothesis (H1)**: The year of manufacture has a significant positive impact on the selling price.

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

To analyze the effect of **kilometers driven** on **selling price**, another linear regression was performed.




```
> summary(model_km)

Call:
lm(formula = selling_price ~ km_driven, data = car_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-458214 -187497  -49025  151095  897864 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.169e+05  7.594e+03   68.07   <2e-16 ***
km_driven   -1.962e+00  1.045e-01  -18.78   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 238200 on 3960 degrees of freedom
Multiple R-squared:  0.08176,	Adjusted R-squared:  0.08153 
F-statistic: 352.6 on 1 and 3960 DF,  p-value: < 2.2e-16
```

Model Summary:

Statistic	Value
Intercept	516,900
Kilometers Driven Coefficient	-1.962
R-squared	0.08176

Adjusted R-squared	0.08153
F-statistic	352.6
p-value	< 2.2e-16

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 (H0)**: Kilometers driven does not have a significant negative impact on the selling price.
- **Alternative Hypothesis (H1)**: 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

To assess if there is a price difference between **Diesel** and **Petrol** cars, a t-test was performed.

Hypothesis Test

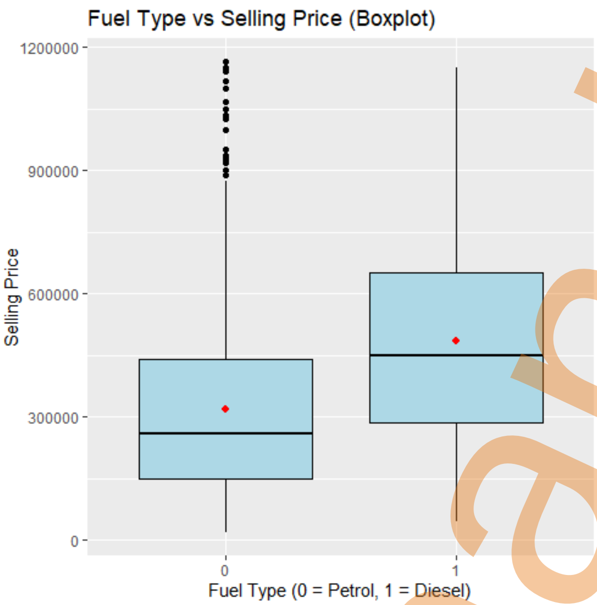
- **Null Hypothesis (H0)**: There is no significant difference in selling price between Diesel and Petrol cars.
- **Alternative Hypothesis (H1)**: Diesel cars have a significantly higher selling price than Petrol cars.

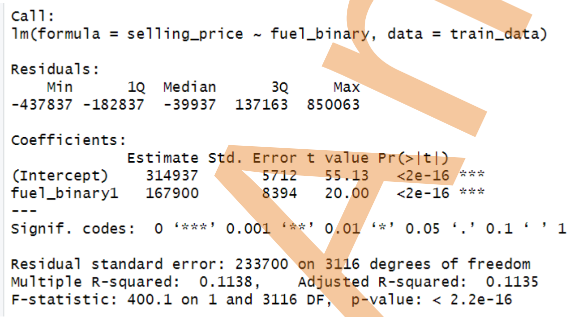
Result: The p-value is less than 0.05, allowing us to reject H0, indicating that Diesel cars have a significantly higher selling price than Petrol cars.

Fuel Type	Mean Selling Price (₹)
Petrol	314,937
Diesel	482,837

Fuel Binary Regression Model

A binary variable (fuel_binary) was created for fuel type (0 = Petrol, 1 = Diesel). A regression model was then fitted to predict selling price based on this binary variable.





```
Call:
lm(formula = selling_price ~ fuel_binary, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-437837 -182837  -39937  137163  850063 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    314937       5712   55.13   <2e-16 ***
fuel_binary1   167900       8394   20.00   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 233700 on 3116 degrees of freedom
Multiple R-squared:  0.1138,	Adjusted R-squared:  0.1135 
F-statistic: 400.1 on 1 and 3116 DF,  p-value: < 2.2e-16
```

Model Summary:

Statistic	**Value**
Intercept	**314,937**
Fuel Binary Coefficient	**167,900**
R-squared	**0.1138**
Adjusted R-squared	**0.1135**
F-statistic	**400.1**
p-value	**< 2.2e-16**

Regression Equation: The equation for this model is:

selling price=314,937+(167,900×fuel binary)

where:

- **selling price** represents the predicted price of the car.
- **fuel binary** is a binary variable where:

    - **1** indicates Diesel cars.
    - **0** indicates 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.

Model Performance

To accuracy evaluate the model's, we calculated the **Root Mean Squared Error (RMSE)** on the test dataset:

- **RMSE**: 238,834.9

This RMSE value represents the average deviation between predicted and actual selling prices, providing a measure of model error.

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.

Sample Predicted vs. Actual Selling 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 of whole report

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.

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.