Car Data Analysis

By Krishna Rao Ramesh

Table of contents

1 Problem Statement

To explore the dataset provided by performing Exploratory Data Analysis (EDA), Data cleaning, and Feature Engineering to discern actionable insights for the company. This will involve filling missing values, replacing outliers, and generating a number of visualizations for understanding each variable, which can be further used to provide recommendations for improving the company's functioning.

2 Solution Statement

A variety of packages from R would be leveraged to carry out data cleaning, feature engineering, and EDA visualizations. The following steps will be taken;

- **Pre-Cleaning Visualization** to get a glimpse of the distribution of data and inconsistencies such as extreme values, formatting, missing values, etc.
- **Data cleaning** to get rid of all the aforementioned data inconsistencies.
- **Feature Engineering** to create new variables for deriving additional insight.
- **Exploratory Data Analysis (EDA)** to identify KPI's, patterns, and trends in data. Visualizations will point out interesting occurences for each feature present in the dataset.

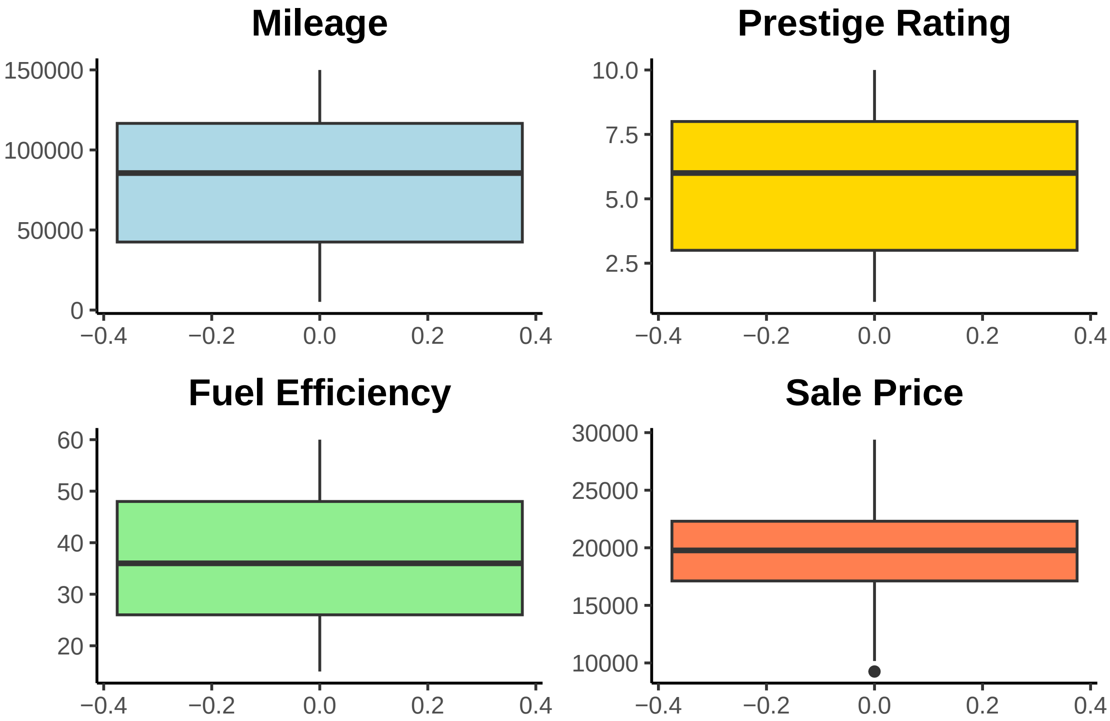
3 Data Preprocessing & Exploratory Analysis

3.1 Section A: Visualizing data pre-cleaning to understand type of distribution

Before cleaning our data, let us explore unclean data to understand its distribution and identify insufficiencies that can deter effective analysis. Each of the 4 numerical columns contain **50 missing values each**.

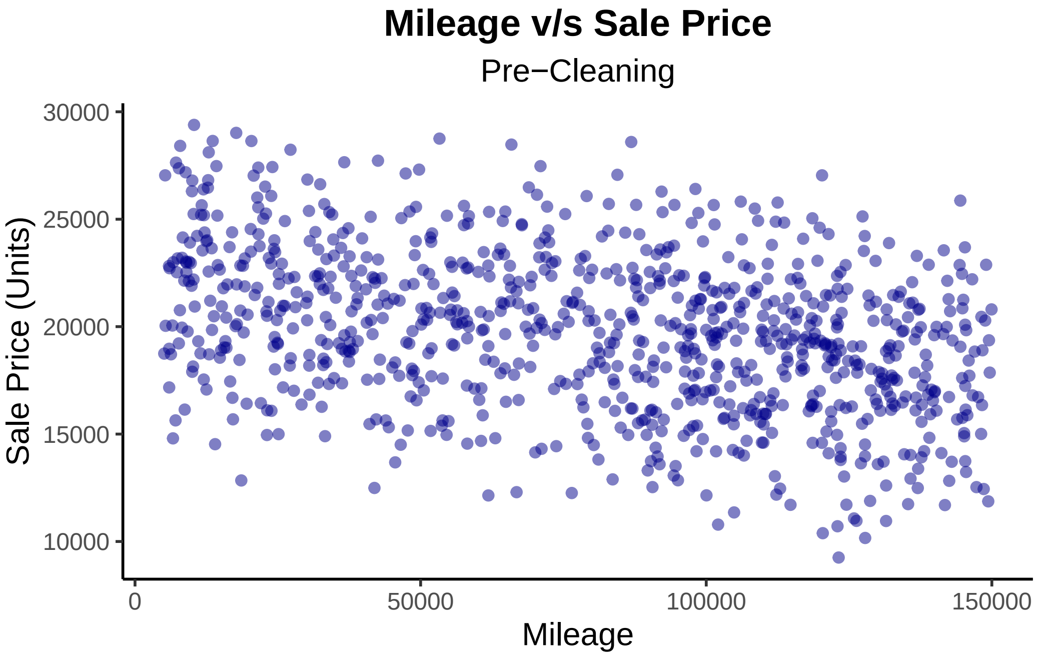
3.1.1 Boxplots for numerical columns

The boxplots represent the data distribution for each of the 4 numerical columns. The variables have moderate variability in data with no pronounced spread, except for an outlier in the "Sale Price" feature. This might denote an unusually low priced car.



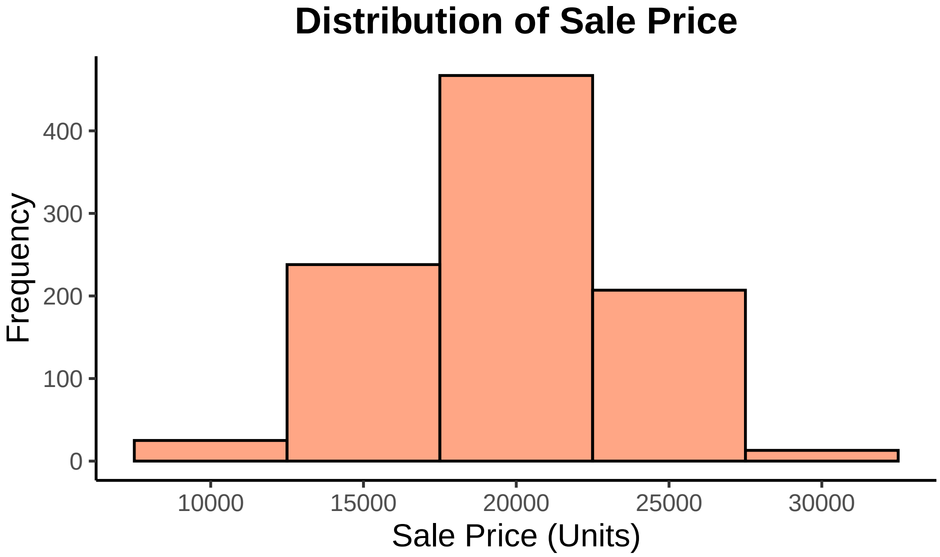
3.1.2 Scatter plot - Mileage v/s SalePrice

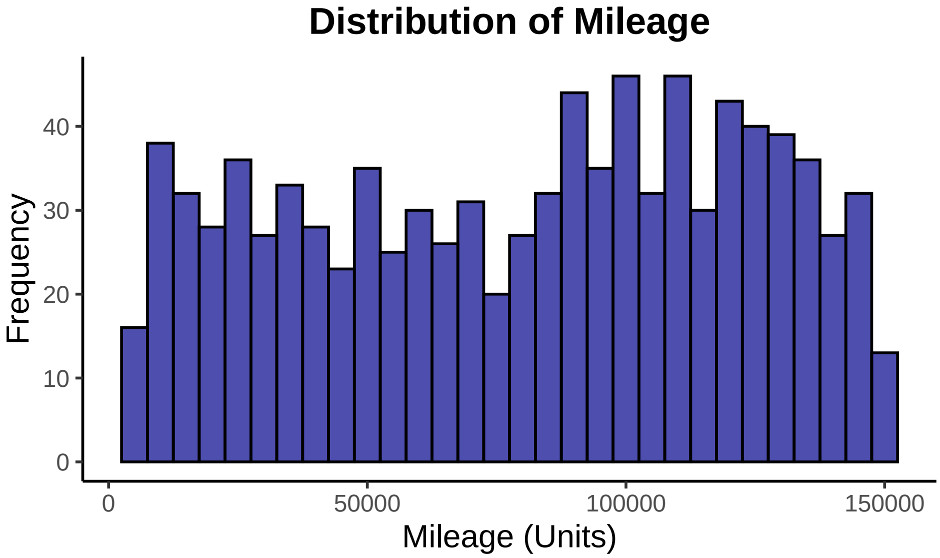
The plot reveals a downward trend wherein the price of cars seems to be decreasing with an increase in mileage. The presence of extreme points, especially at the top of the plot, can potentially be outliers.



3.1.3 Histograms for SalePrice and Mileage

The Sale Prices of cars are normally distributed. Different Cars tend to have different sets of mileage, with the maximum concentration around the 85,000 to 135,000 mark.





3.2 Section B: Data Cleaning

a. Cleaned the ‘Makemodel’ column and split it into Manufacturer and Model. Special characters were removed from the column and regular expression was used to replace different combinations of incorrect spellings (like “aud” or “au” to “Audi”)

b. Imputed missing values using **MICE (Multivariate Imputation by Chained Equations)**. MICE was chosen since it takes relationships or correlations between variables into account before imputing missing values. This can lead to better analysis.

c. Removing incorrect data entries and removing duplicates, such as negative values. Years out of range (2000-2024) are also removed from dataset.

d. Removing Outliers by calculating the **Mahalanobis distance** between datapoints. Similar to MICE, Mahalanobis also takes the relationship between variables into account before judging outliers. This can be useful since a car with an unusually high mileage count could be deemed an outlier unless the age of the car is also taken into account.

Table 1: Cleaned Car Data (First 10 Rows)

Manufacturer	Model	YearOfManufacture	SalePrice	Mileage	FuelEfficiency	PrestigeRating
Chevrolet	Impala	2014	22120.79	94307	48	7
Toyota	Highlander	2015	16835.13	138829	45	2
Audi	A4	2003	14895.14	20118	32	1
Toyota	RAV4	2010	25042.48	118584	54	8
Ford	Ex plorer	2016	17717.65	97745	46	1
Tesla	Model S	2022	22979.35	57355	43	5
BMW	X5	2002	22479.38	32354	47	7
Audi	Q5	2011	15134.18	92194	31	2
Tesla	Model X	2010	21825.79	65807	52	6
Chevrolet	Corvette	2010	21148.99	23367	29	8

3.3 Section C: Feature Engineering

Newer columns derived from primary data are added to dataset to extract additional insight.

(I) **Car_age**: To identify age of car.

(II) **FuelEfficiency_Category**: Grouping fuel efficiencies of vehicles into different classes of Low, medium and high.

(III) **Price_Category**: Grouping SalePrice of vehicles into different classes.

(IV) **Mileage_Category**: Grouping Mileage of vehicles into different classes.

(V) **PricePerMile**: A column to indicate money spent for each mile driven.

(VI) **EfficiencyPrice_Ratio**: To measure the cost-effectiveness of a car's fuel efficiency.

Table 2: Feature Engineered Columns (First 10 Rows)

Car_Age	FuelEfficiency_Category	Price_Category	Mileage_Category	PricePerMile	EfficiencyPrice_Ratio
11	High	High	High	0.23	0.47
10	High	Medium	High	0.12	0.63
22	Medium	Medium	Low	0.74	0.46
15	High	High	High	0.21	0.47
9	High	Medium	High	0.18	0.61
3	Medium	High	Medium	0.40	0.37
23	High	High	Medium	0.69	0.44
14	Medium	Medium	High	0.16	0.43
15	High	Medium	Medium	0.33	0.54
15	Medium	Medium	Low	0.91	0.21

3.4 Section D: EDA Visualizations post-cleaning

3.4.1 Understanding the correlation between different numerical features

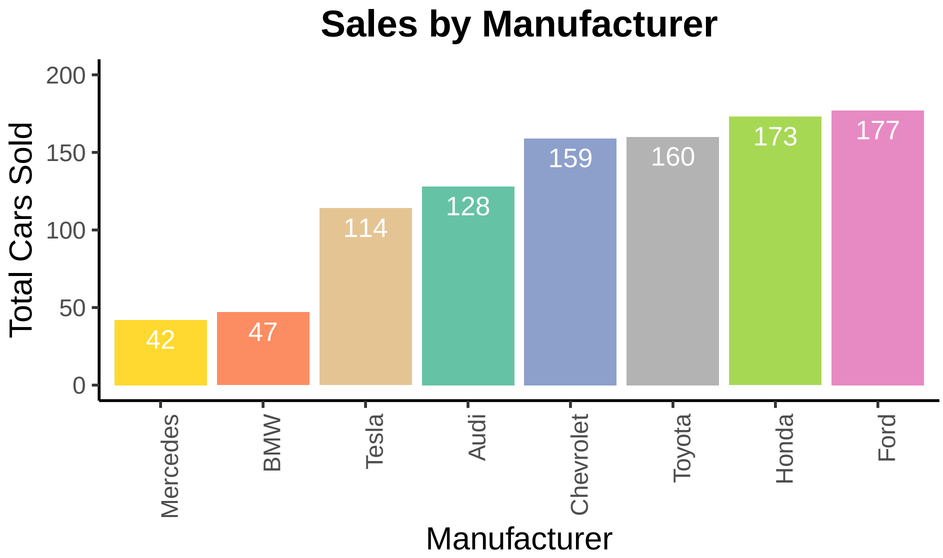
Sale Price is moderately correlated with Fuel Efficiency and Prestige Rating. None of other correlations are significant.

Table 3: Correlation Matrix of Numeric Variables

	SalePrice	Mileage	FuelEfficiency	PrestigeRating
SalePrice	1.00	-0.37	0.54	0.63
Mileage	-0.37	1.00	0.01	0.01
FuelEfficiency	0.54	0.01	1.00	0.03
PrestigeRating	0.63	0.01	0.03	1.00

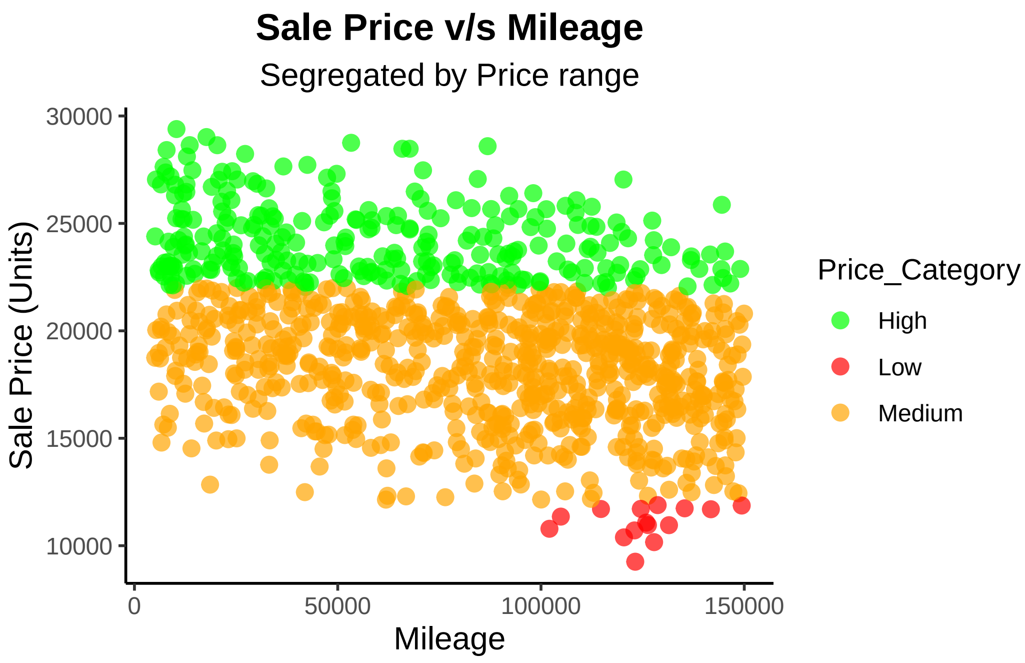
3.4.2 Bar Chart of Total Sales by Manufacturer

The bar chart shows the total number of cars sold by each manufacturer. Ford is leading the race with 177 cars followed by Honda with 173 sold.



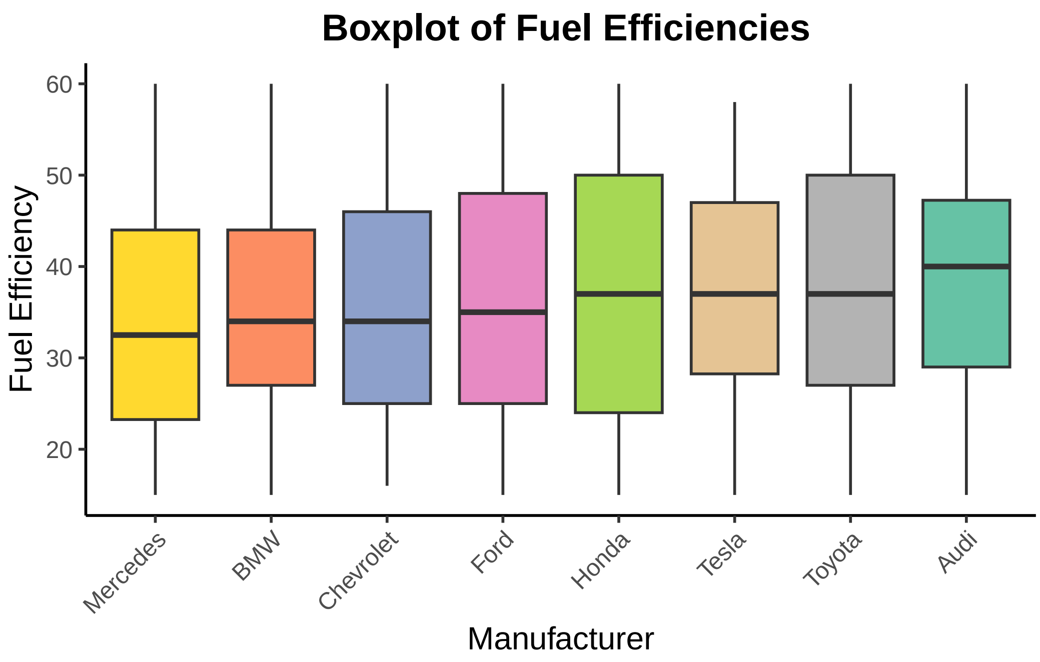
3.4.3 Scatter plot - Sale Price v/s Mileage by Price Category

The scatter plot explains the relationship between variables "SalePrice" and "Mileage", bifurcated on the basis of "PriceCategory". Cars with higher mileage tend to have lower sale prices, but some high-mileage cars continue to cost more, probably due to higher prestige rating or brand value.



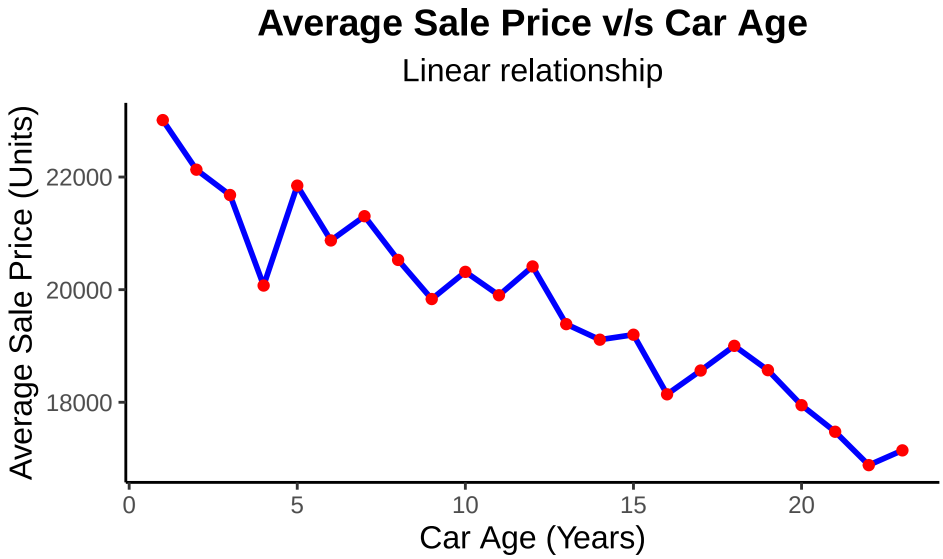
3.4.4 Range of Fuel Efficiency with respect to Manufacturer

The boxplots below compares Fuel Efficiency of cars across manufacturers. The fuel efficiencies of Audi and Toyota have a higher median value, which suggests that their cars are the most efficient when it comes to fuel consumption. Mercedes cars are on the other side of the spectrum.



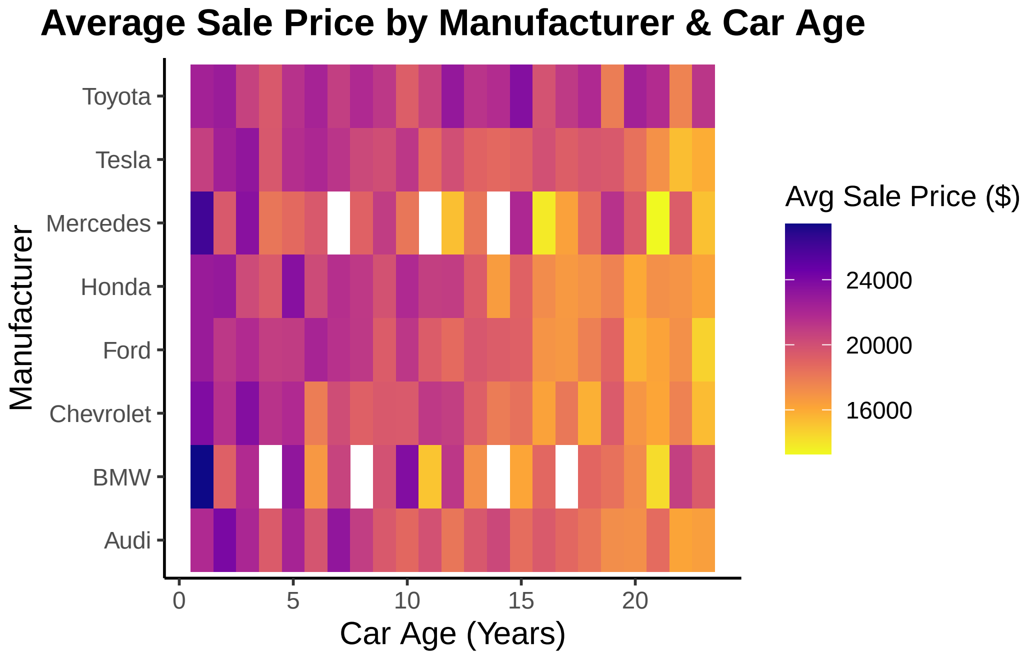
3.4.5 Average Sale Price in comparison with Car Age

The line chart indicates that "young" cars are higher priced, as expected. The sale price drops steeply after the 10-year mark, which signals drastic depreciation of commodity value.



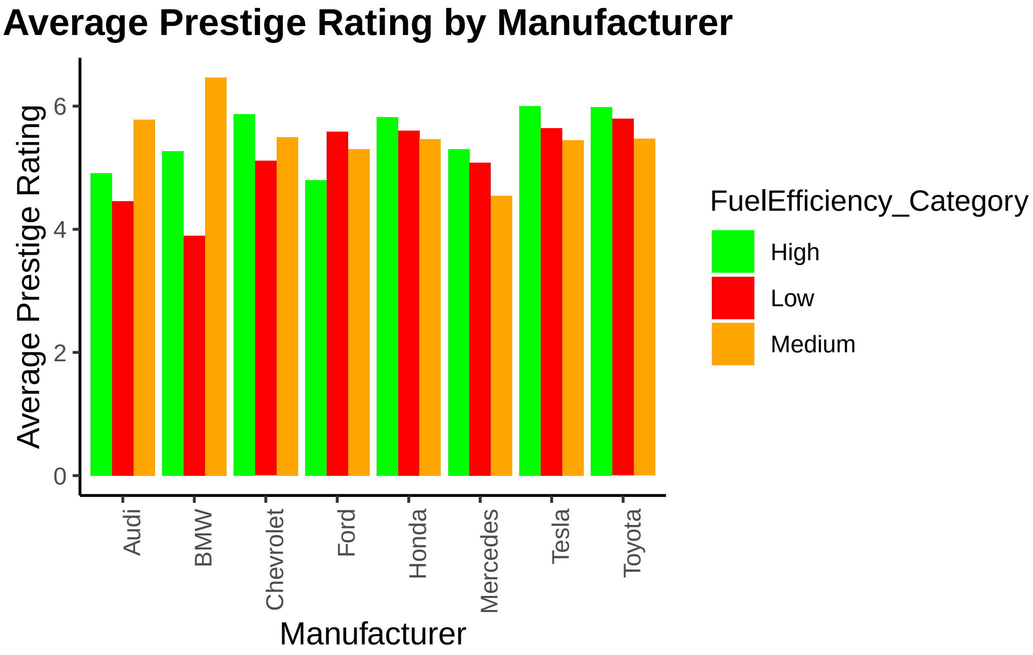
3.4.6 Heatmap - Average Sale Price by Manufacturer and Car Age

The heatmap visualizes the average Sale price of cars concerning age and make. Toyota and Tesla cars seem to retain much of their worth through out their functioning, while Mercedes and BMW tend to quickly lose their value with time.



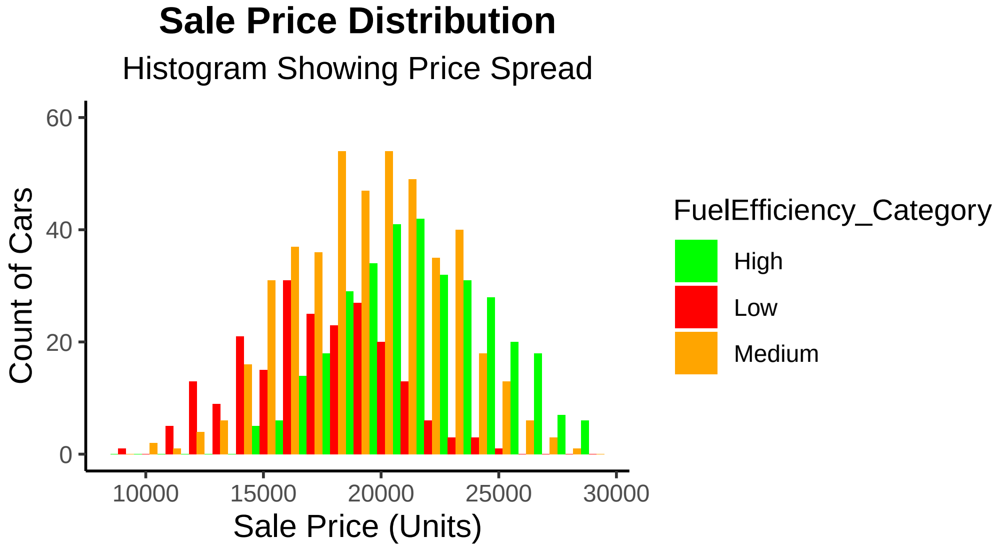
3.4.7 Bar Chart of Average Prestige Ratings with regards to Manufacturer & Fuel efficiency category

The grouped bar chart compares average prestige ratings of different manufacturers, grouped on the basis of Fuel efficiency category. Across different categories, Toyota and Tesla cars have higher prestige ratings compared to other manufacturers, while Audi and BMW have poor ratings in general.



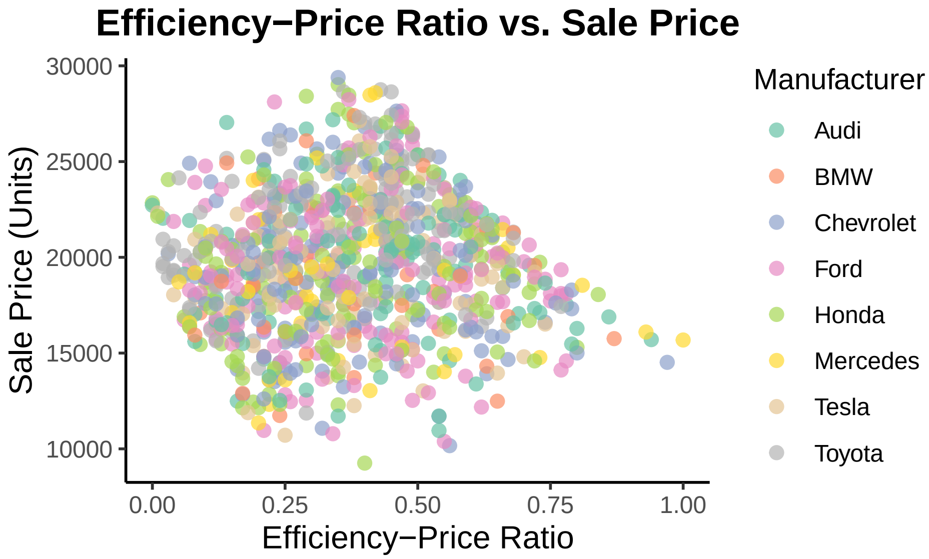
3.4.8 Histogram - Sale Price Distribution v/s Fuel Efficiency Category

This graph clearly explains that cars with higher fuel efficiency have a higher price bracket. It can be inferred that fuel efficiency drives sale prices of cars.



3.4.9 Scatter Plot - Efficiency-Price Ratio v/s Sale Price

Higher Efficiency-Price Ratio signifies more cost-effective cars. This means that cars with an higher ratio have better fuel efficiency per unit of money spent.



3.4.10 Summary Statistics of KPI's with Confidence Intervals for Sale Price

Toyota cars have the highest sale value across all of their different models while also having strong fuel efficiencies and prestige ratings. Mercedes suffers from low prestige rating, which result in sub-par sale prices. The wide variation in confidence intervals for BMW and Mercedes denotes that the car prices in these brands vary more. This fluctuation in price also supports the fact that the value of these cars can be a bit unpredictable.

Table 4: Summary Statistics by Manufacturer

Manufacturer	Avg_SalePrice	CI_Lower	CI_Upper	Avg_FuelEfficiency	Avg_PrestigeRating	Avg_PricePerMile	Total_cars_sold
Ford	19330.15	18787.49	19872.81	36.64	5.21	0.51	177
Honda	19407.47	18830.63	19984.31	36.40	5.63	0.47	173
Toyota	21273.33	20769.48	21777.18	37.85	5.74	0.49	160
Chevrolet	19487.30	18897.82	20076.79	36.11	5.52	0.52	159
Audi	19646.33	19016.49	20276.17	38.32	5.23	0.47	128
Tesla	19597.92	18969.83	20226.02	37.64	5.67	0.47	114
BMW	19357.13	18281.86	20432.39	35.94	5.64	0.53	47
Mercedes	18635.21	17413.03	19857.39	35.26	4.88	0.42	42

3.4.11 Top Models by Manufacturer

Quite surprisingly, it is a Mercedes car that has outranked other cars to claim top spot in fuel efficiency. Ford Fiesta seems to be the most popular car among customers, while the Toyota Camry has the highest sale price on average.

Table 5: Top cars by manufacturer

Manufacturer	Model	Total_Sales	Avg_SalePrice	Avg_FuelEfficiency
Audi	A8	22	18865.72	35.18
BMW	X5	16	19831.72	38.31
Chevrolet	Malibu	20	19249.95	35.95
Ford	Fiesta	26	19532.05	34.35
Honda	CR V	25	20308.55	38.92
Mercedes	Benz E Class	15	18866.96	40.67
Tesla	Model S	22	19893.63	35.86
Toyota	Camry	20	21355.46	35.90

4 Key Takeaways

- Ford is the most successful manufacturer, with sales totaling 177 cars. Ford Fiesta is the most popular choice among customers when it comes to purchasing a car.
- Cars with higher mileage tend to have lower sale prices, but some high-mileage cars continue to cost more, probably due to higher prestige rating or brand value.
- The sale price of cars drops steeply after the 10-year mark, which signals drastic depreciation of commodity value.
- The fuel efficiencies of Audi and Toyota are the highest among all manufacturers.
- Cars with higher fuel efficiency will have a higher price bracket, except for some instances.

5 Recommendations

1. **Aggressive Pricing strategies ==>** Pricing can remain to be competitive for mid-range and luxury cars, but be reduced for older models of Ford to offload stock with less demand.
2. **Targeting customer segments** ==> The columns “Mileage Category**”** and”Price Category**“** can be used to segregate customers with different requirements. Tailor-made marketing campaigns can be effectively designed to attract each customer segment.
3. **Efficient Inventory Management** ==> Cars such as the Ford Fiesta or the Honda CR V in the mid-range category seems to be the fastest moving commodity. Inventory stocking should be fine-tuned with respect to demand and supply.
4. **Focus on Sustainability** ==> Promote and stock up on Eco-friendly models that offer more-than-average fuel efficiency. This can be used to fuel marketing campaigns that concentrate more on Efficiency-Price Ratio to appeal more to eco-conscious buyers.