GNG 5300K: Introduction to Data Visualization & Analytics

Deliverable 4: The Story

Title: Car Price Data Analysis

University of Ottawa

Faculty of Engineering

Guided by:

Dr. Andrew Sowinski

Presented by:

Group 1

Yash Dalvi 300257910

Shivani Deshmukh 300251019

Vikram Khanzode 300198886

Rishi Kumar Pandey 300219699

Date: December 12th, 2022

1. Introduction

The dataset we chose for the study includes several characteristics essential in the vehicle manufacturing and sales business. We intended to forecast the sales of several car manufacturers using a data-centric approach. As we know, sales analysis aids the business in determining the current size of the niche it has chosen, as well as the business of its competitors and how customers react to change. The chosen dataset includes KPIs that are important for the vehicle manufacturing business. Based on the correlation between the available variables and the various qualities in the data, our goal was to conduct an automobile sales analysis.

The data was collected via Kaggle, which we cleaned by removing missing values and duplicate values and carried out basic data cleaning using a python library called Pandas. The dataset had sixteen columns or features used to determine their impact on the Manufacturer Suggested Retail Price (MSRP). While investigating, we found that the null hypothesis for the currently investigated dataset was ($P > 0.05$), which indicates that there is no linear association between the distinctive features of an automobile and its MSRP. The alternate hypothesis was ($P < 0.05$) which says that there is a linear relationship between unique features of cars and cars' MSRP. We have used aggregate statistics and regression analysis to analyze the data and extract insights that will further be analyzed for the dashboarding and visualization.

For this study, we have used Microsoft's PowerBI: an interactive data visualization application used for obtaining and developing critical business insights. Dashboards are important for tracking a business's success and quickly reviewing all its key performance indicators (KPIs). Additionally, at last, we carried out some exploratory data analysis over the data set using python packages called Matplotlib and Seaborn.

2. Dataset

2.1 INTRODUCTION AND DATASET

The selected dataset contains KPIs (key performance indicators) which are vital in any automobile manufacturing industry. Using the available variables, we intend to perform a car sales analysis based on the correlation with different attributes in the data. Sales analysis helps the company identify the current market for the selected niche, analyze the competitor's business and how the customers behave/respond to the change [1]. Currently, the automobile sector is at its peak level of saturation, and most manufacturers rely on state of art innovations that distinguish themselves from their competitors [2], [3]. We want to use a data-centric approach to predict the sales of different makes of cars.

We have amassed the dataset from Kaggle and the information in the dataset is scrapped from Twitter and Edmunds.com (an American website for automotive inventory and information which also includes expert automobile reviews). The dataset has sixteen columns and approximately twelve thousand rows such that each member can perform their analysis on a particular make of the car. The key factors that might help us better correlate the car sales are Engine HP, Engine Cylinders, and MSRP (Manufacturer Suggested Retail Price), yet we are not restricted to the factors mentioned and will try to incorporate all the information available at our disposal.

2.2. DATA INTEGRITY

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11914 entries, 0 to 11913
Data columns (total 16 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Make               11914 non-null  object
 1   Model              11914 non-null  object
 2   Year               11914 non-null  int64
 3   Engine Fuel Type   11911 non-null  object
 4   Engine HP          11845 non-null  float64
 5   Engine Cylinders   11884 non-null  float64
 6   Transmission Type  11914 non-null  object
 7   Driven_Wheels      11914 non-null  object
 8   Number of Doors    11908 non-null  float64
 9   Market Category    8172 non-null   object
 10  Vehicle Size       11914 non-null  object
 11  Vehicle Style      11914 non-null  object
 12  highway MPG        11914 non-null  int64
 13  city mpg           11914 non-null  int64
 14  Popularity         11914 non-null  int64
 15  MSRP               11914 non-null  int64
dtypes: float64(3), int64(5), object(8)
memory usage: 1.5+ MB
```

Figure 1 Data Information

	Year	Engine HP	Engine Cylinders	Number of Doors	highway MPG	city mpg	Popularity	MSRP
count	11914.000000	11845.00000	11884.000000	11908.000000	11914.000000	11914.000000	11914.000000	1.191400e+04
mean	2010.384338	249.38607	5.628829	3.436093	26.637485	19.733255	1554.911197	4.059474e+04
std	7.579740	109.19187	1.780559	0.881315	8.863001	8.987798	1441.855347	6.010910e+04
min	1990.000000	55.00000	0.000000	2.000000	12.000000	7.000000	2.000000	2.000000e+03
25%	2007.000000	170.00000	4.000000	2.000000	22.000000	16.000000	549.000000	2.100000e+04
50%	2015.000000	227.00000	6.000000	4.000000	26.000000	18.000000	1385.000000	2.999500e+04
75%	2016.000000	300.00000	6.000000	4.000000	30.000000	22.000000	2009.000000	4.223125e+04
max	2017.000000	1001.00000	16.000000	4.000000	354.000000	137.000000	5657.000000	2.065902e+06

Figure 2 Data Description

We have used Kaggle (a public database source). This data is collected by web scrapping done on Edmunds.com and Twitter. Our current dataset can be compared with another price prediction model [4], which shows a close resemblance to the entities used in our data and thus confirms the validity of the data. Figure 1 and 2 can also be referred to confirm the data's validity.

2.3. FUTURE DATA ANALYSIS AND VISUALIZATION

In this project, we will focus on correlating different traits in data and will try to identify trends that make the car more sellable. This may include, however, is not restricted to, effects of the car features on the price, brand value, and much more. We will also try to remove the null values and may try to replace them. Based on our analysis so far, we are thinking about consolidating the mileage and the year of the cars and what effect it will have on the price of the car. According to our understanding, we may use regression analysis to recognize trends. We are hoping to experiment with using diverse kinds of visualization plots and which will best suit our data, and later we can make decisions based on our observation.

2.4. Changes made into Deliverable 1

The term "automotive industry" refers to all businesses and endeavors concerned with producing automobiles, including most of its parts, such as bodywork and engines. The invention of the gasoline engine in the 1860s and '70s, primarily in France and Germany, laid the foundation for the automobile sector, even if steam-powered road vehicles were manufactured early. The World War 1 and 2 brought a massive technological enhancement throughout the world. As a result, assembly lines were built to assemble the automobiles by reputable manufacturers like Ford and General Motors. This led to the birth of mass manufacturing of vehicles. Since then, the development of cars has continued, and thanks to Industry 4.0, we can now experience a digital world, where automobile manufacturing is not restricted to mechanical concepts [15]. The utilization of data has a substantial impact on production; it boosts productivity across the entire process, which benefits both the producers and the customers. [16].

Some biases like selection bias and omitted variable bias can be seen in the initial raw dataset during the data collection phase, where the selection of data is not represented by all the people in the world and number of passengers in the car would have been a powerful addition to the dataset,

respectively. Exclusion bias may be observed during the data preprocessing stage since the data contain several null values that will be eliminated, which might have a major impact on the data.

During the exploratory data analysis (EDA) process, we will use various calculations like, COUNT, AVERAGE, SUM etc. to normalize the data. We will use box plots to identify outliers and remove them.

3. Data Cleaning

We have carried out data cleaning and preparation using Python, Jupyter Notebooks IDE for data analysis. The Python packages we have utilized are Pandas for data cleaning and preparation as well as Matplotlib and Seaborn for data visualization. The Jupyter Notebook file named "GNG5300_Cars_Cleaning+EDA.ipynb" included in the attachments uses 'data.csv' as its input and gives out 'AllCarsData.csv" as the cleaned dataset as its output. The data cleaning was executed by dropping any duplicate rows as well as any rows containing a null value(s). Then columns were renamed to more appropriate names or to correct spelling errors. The feature "Market Category" had too many null values, hence, we replaced the null values in that column with "Miscellaneous." Additionally, some features were converted from one data type to another, e.g., Year was made a string type, mileages, and engine HP were converted into a float as well as the number of doors and number of engine cylinders were converted to integer datatype, etc. Finally, some aggregate statistics for the numerical features were calculated as shown in Table 1.

	Engine HP	Engine Cylinders	Number of Doors	Highway Mileage (mpg)	City Mileage (mpg)	Popularity Rating	MSRP
count	11084.000000	11084.000000	11084.000000	11084.000000	11084.000000	11084.000000	1.108400e+04
mean	253.593919	5.688109	3.451822	26.247925	19.303861	1558.142367	4.194166e+04
std	110.184744	1.766399	0.873961	6.804170	6.606635	1444.116034	6.175385e+04
min	55.000000	0.000000	2.000000	12.000000	7.000000	2.000000	2.000000e+03
25%	172.000000	4.000000	2.000000	22.000000	16.000000	549.000000	2.159000e+04
50%	240.000000	6.000000	4.000000	25.000000	18.000000	1385.000000	3.061000e+04
75%	303.000000	6.000000	4.000000	30.000000	22.000000	2009.000000	4.303125e+04
max	1001.000000	16.000000	4.000000	111.000000	137.000000	5657.000000	2.065902e+06

Table 1: Aggregate Statistics of the Numerical Features of the Dataset After Cleaning

There were some additional data cleaning steps taken regarding the "Market Category" column (feature) in this deliverable that was not executed in the previous one. One of the largest challenges we were facing in the last deliverable was that the "Market Category" feature of the dataset had various comma-separated attributes which gave rise to a complicated classification of the market category of the various cars and therefore made it difficult to divide all the cars into set categories. In Figure 3 we can see a snippet of the "Market Category" column. One thing we need to note is that this column had a lot of null values, to begin with, which had been very easily changed into being tagged as "miscellaneous."

Market Category
Factory Tuner,Luxury,High-Performance
Luxury,Performance
Luxury,High-Performance
Luxury,Performance
Luxury
...
Crossover,Hatchback,Luxury
Crossover,Hatchback,Luxury

Figure 3: Market Category column uncleaned version.

The simplest way to solve the above problem was to only consider the first category in the case of multiple comma-separated category names. For example, “Crossover, Hatchback, Luxury” becomes simply “Crossover” assuming that it is the car’s main characteristic because it was listed first. Figure 4 illustrates the python list comprehension method that helped achieve this. Additionally, Figure 5 shows the “Market Category” after being cleaned.

```
#extracting only the first category name for all the listed categories in a row
df['Market Category'] = [x.split(',',1)[0] for x in df['Market Category']]
```

Figure 4: List comprehension in python used to clean the feature.

Market Category
Factory Tuner
Luxury
Luxury
Luxury
Luxury
...
Crossover
Crossover

Figure 5: Market Category column cleaned version.

Subsequently, some rudimentary analysis was also carried out on the newly cleaned version of the "Market Category" feature as shown in figure 6.

```
In [55]: #listing out all the values for market category and their counts
         marketcategoryvalues = df['Market Category'].value_counts()
         marketcategoryvalues

Out[55]: Miscellaneous       3353
         Crossover           1986
         Luxury              1878
         Flex Fuel           1056
         Hatchback            955
         Performance          504
         Exotic               470
         Factory Tuner        432
         High-Performance     198
         Diesel               131
         Hybrid               121
         Name: Market Category, dtype: int64
```

Figure 6: Python script for listing out the breakdown of the Market Category by their respective counts.

4. Exploratory Data Analysis

EDA is an important part of data analysis as it can give a clear picture of what to expect from the data, what can be done with the data set available etc. and helps in decision-making. With the data set available, we have used statistical and visualization tools to analyze the trends in the data set. Some of the visualizations and statistics snapshots have been attached below:

Count of Transmission Type	Column Labels					
Row Labels	AUTOMATED_MANUAL	AUTOMATIC	DIRECT_DRIVE	MANUAL	UNKNOWN	Grand Total
Acura	21	169		56		246
1992				1		1
1993				6		6
1994		1		5		6
1995		2		4		6
1997		2				2
1998		1				1
1999		1		6		7
2000				7		7
2001		6		7		13
2002		2				2
2003		3		2		5
2004		3		4		7
2005		3		4		7
2006		2		3		5
2010		3				3
2011		6				6
2012		19		2		21
2013		14		2		16
2014		15		2		17
2015	2	25		1		28
2016	10	34				44
2017	9	27				36
Alfa Romeo	5					5
2015	3					3
2016	2					2
Aston Martin	22	38		31		91
2003		1		3		4
2004		1				1
2005	2					2

Table 2: Basic Statistics of the Dataset

In this, we have created a pivot chart to understand how many cars of different transmission types were designed in different years. This can help us determine the future sale pattern and which is the most preferred transmission type in recent years.

	MSRP
Mean	41901.11895
Standard Error	586.0003561
Median	30600
Mode	2000
Standard Deviation	61730.62484
Sample Variance	3810670043
Kurtosis	258.4416379
Skewness	11.59963665
Range	2063902
Minimum	2000
Maximum	2065902
Sum	464976717
Count	11097

Table 3: Pivot Table of the Dataset

To summarize the MSRP into a smaller data set and get an insight into how much people are willing to spend on cars we have generated a statistical analysis of MSRP. This table gives an average of how much people spend on buying a car. The total sum of the cars purchased to date from the dealers and the number of cars sold.

While carrying out data visualization of our dataset using box plots and histograms for several numerical features, certain outliers were detected which were obviously born out of errors. One such example was the boxplot for Highway Mileage (mpg) feature.

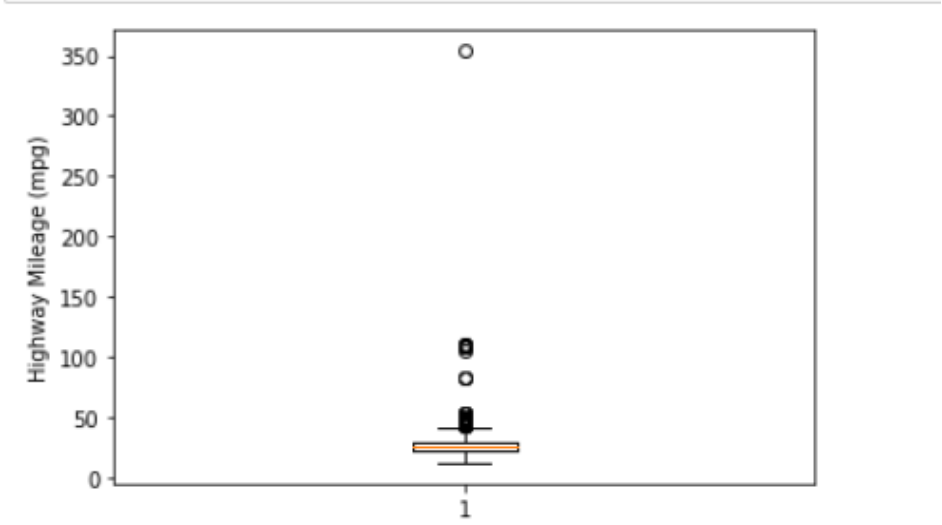

Figure 7: Boxplot of Highway Mileage feature with an outlier at 354.0 mpg (about 1 cent per mile).

There is only one outlier which shows its City Mileage as 24 miles per gallon but its Highway Mileage as 354 miles per gallon which is obviously an error. Hence, this one tuple for located and removed from the dataset to give rise to a much cleaner and more accurate boxplot as bellow:

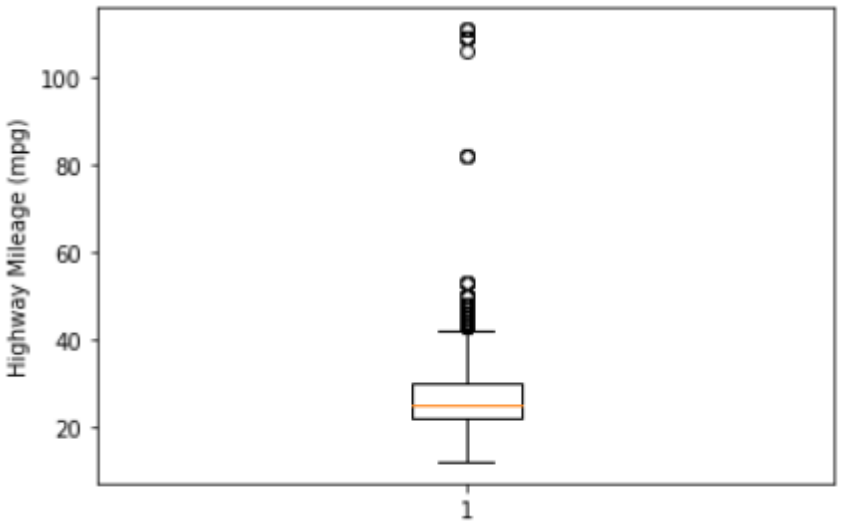

Figure 8: Boxplot of Highway Mileage feature after removing the outlier.

Distribution of many numerical features was evident by plotting histograms as follows:

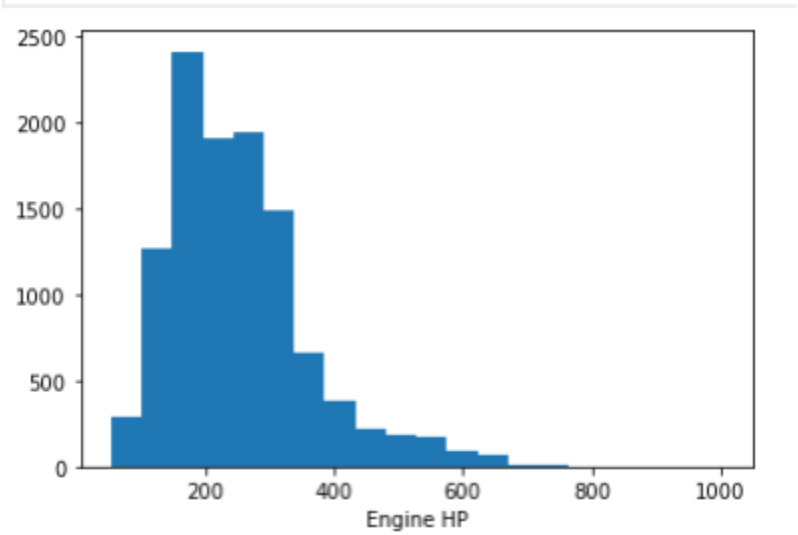

Figure 9: Histogram showing frequency distribution of Engine HP values.

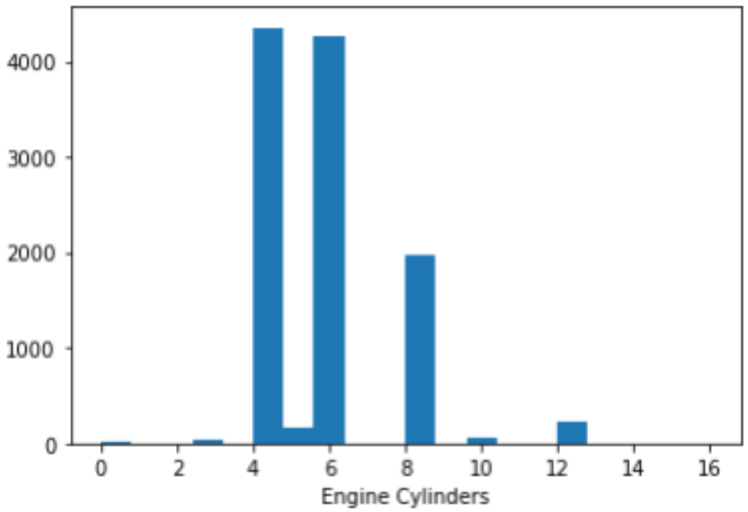

Figure 10: Histogram showing frequency distribution of number of Engine Cylinders.

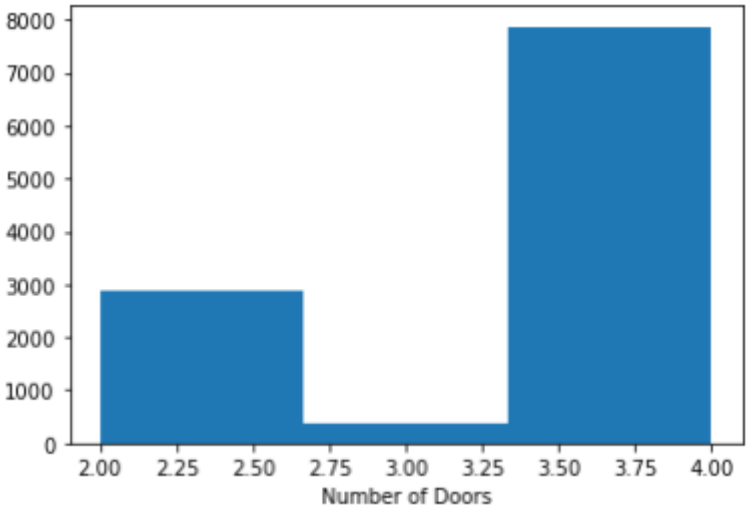

Figure 11: Histogram showing frequency distribution of Number of Doors.

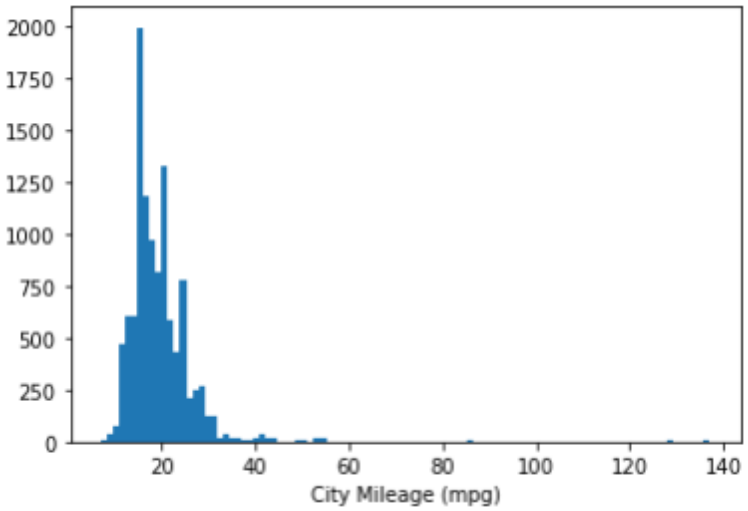

Figure 12: Histogram showing frequency distribution of City Mileage (mpg).

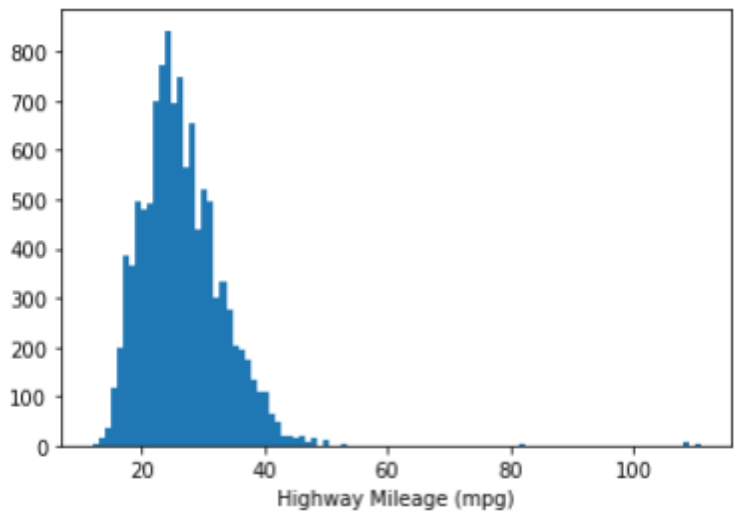

Figure 13: Histogram showing frequency distribution of Highway Mileage (mpg).

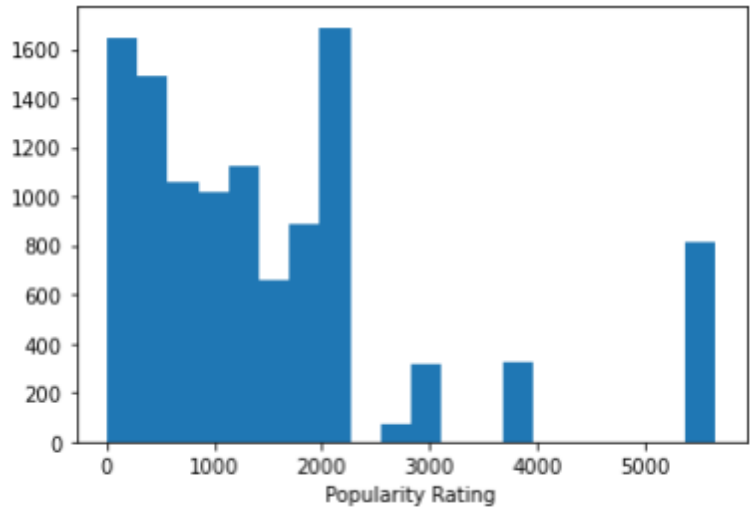

Figure 14: Histogram showing frequency distribution of Popularity Ratings.

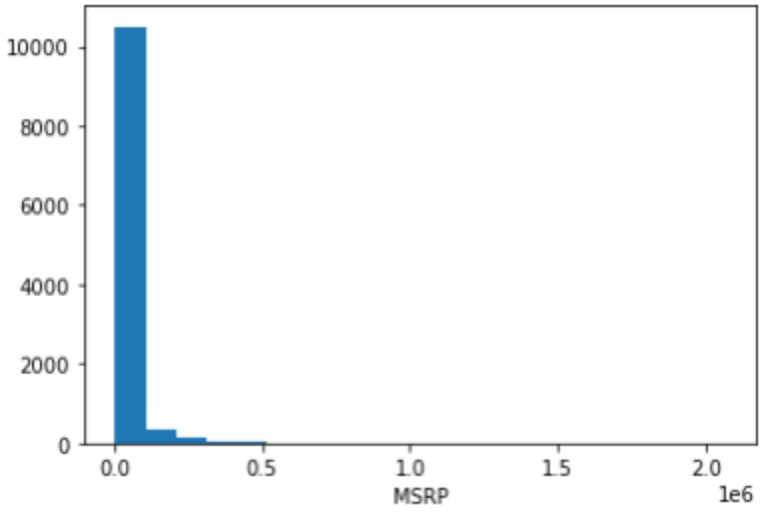

Figure 15: Histogram showing frequency distribution of MSRP values.

Distributions of the categorical features were determined by plotting pie charts for each. These features include Make, Transmission Type, Driving Wheels, Fuel Type, Vehicle Size, among others.

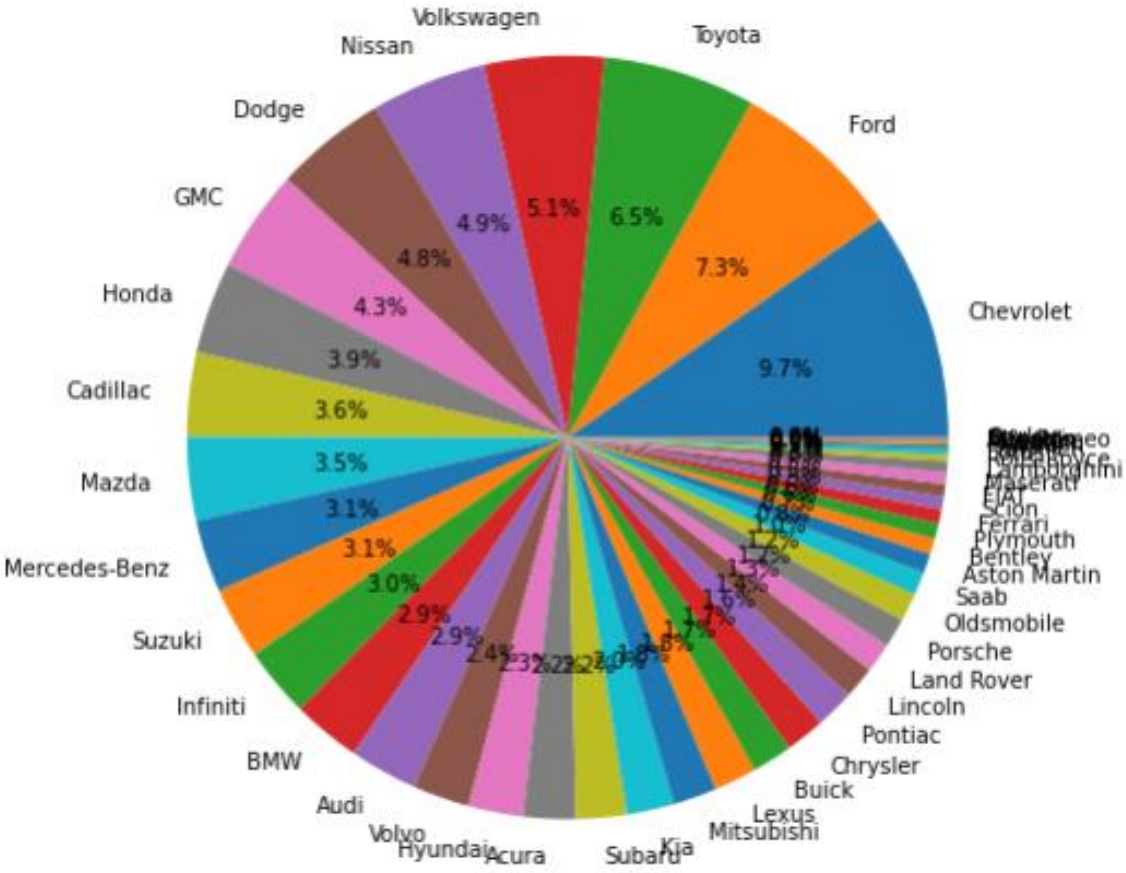

Figure 16: Pie Chart showing the distribution of various Makes.

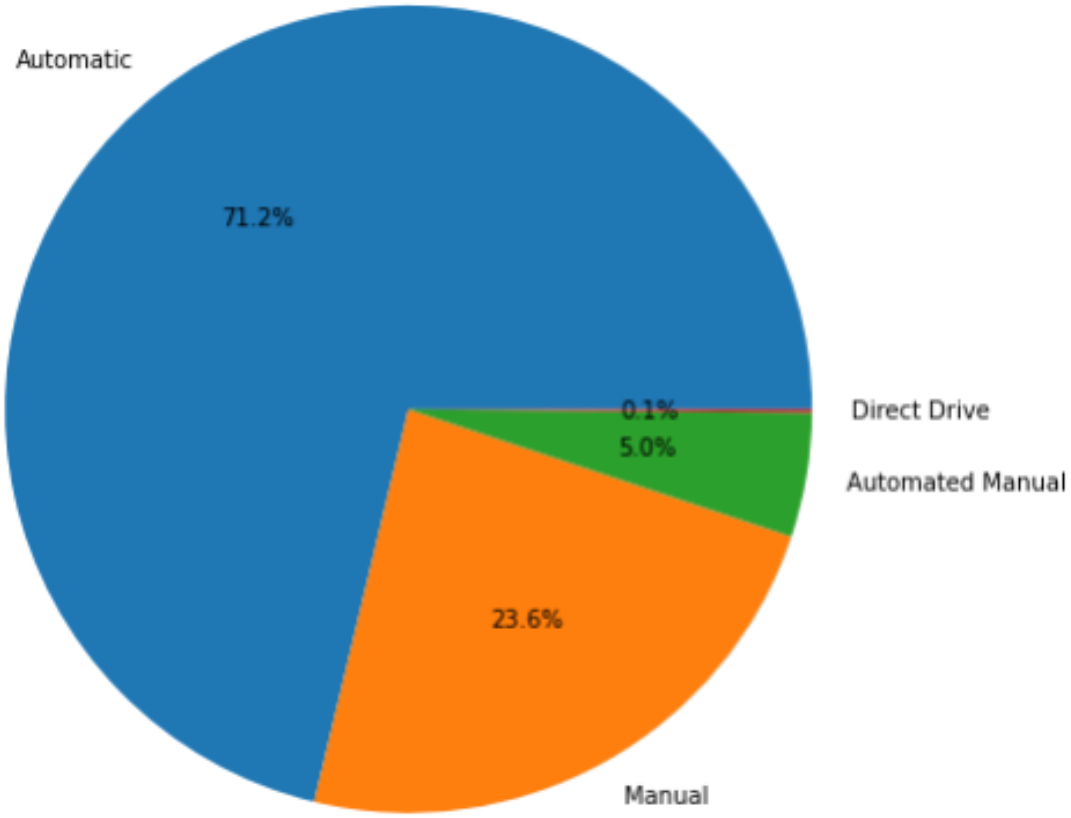

Figure 17: Pie Chart showing distribution of various Transmission Types.

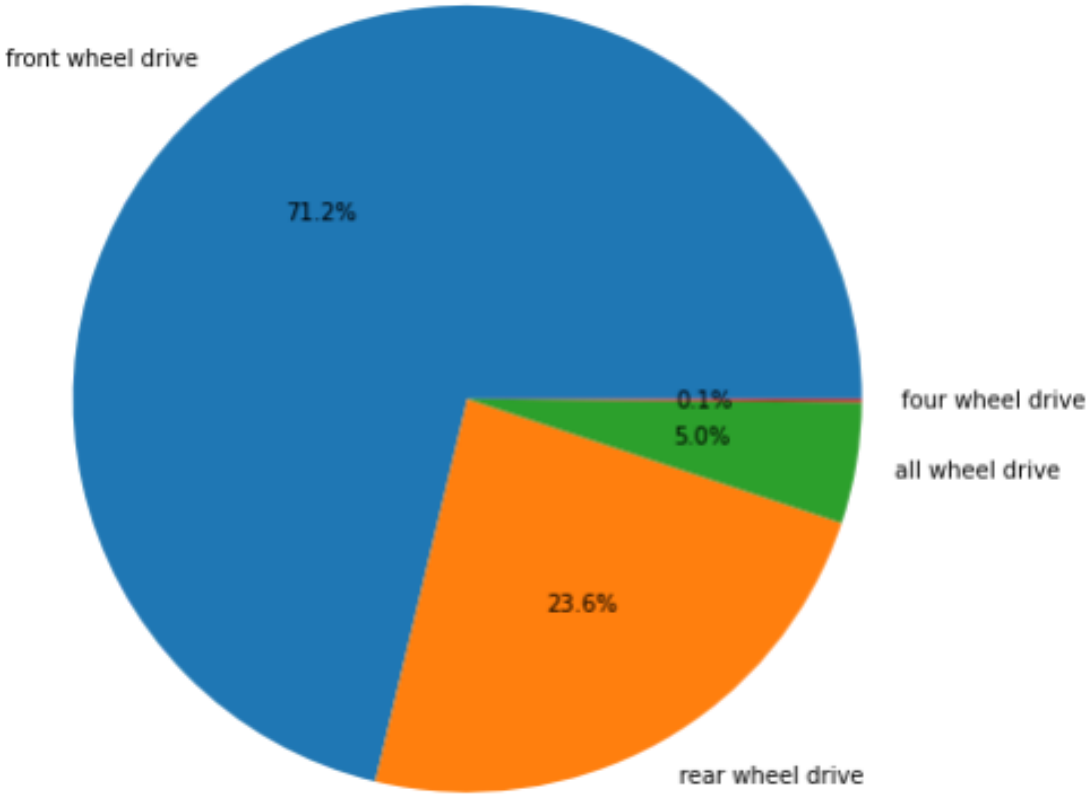

Figure 18: Pie Chart showing distribution of various Driving Wheels.

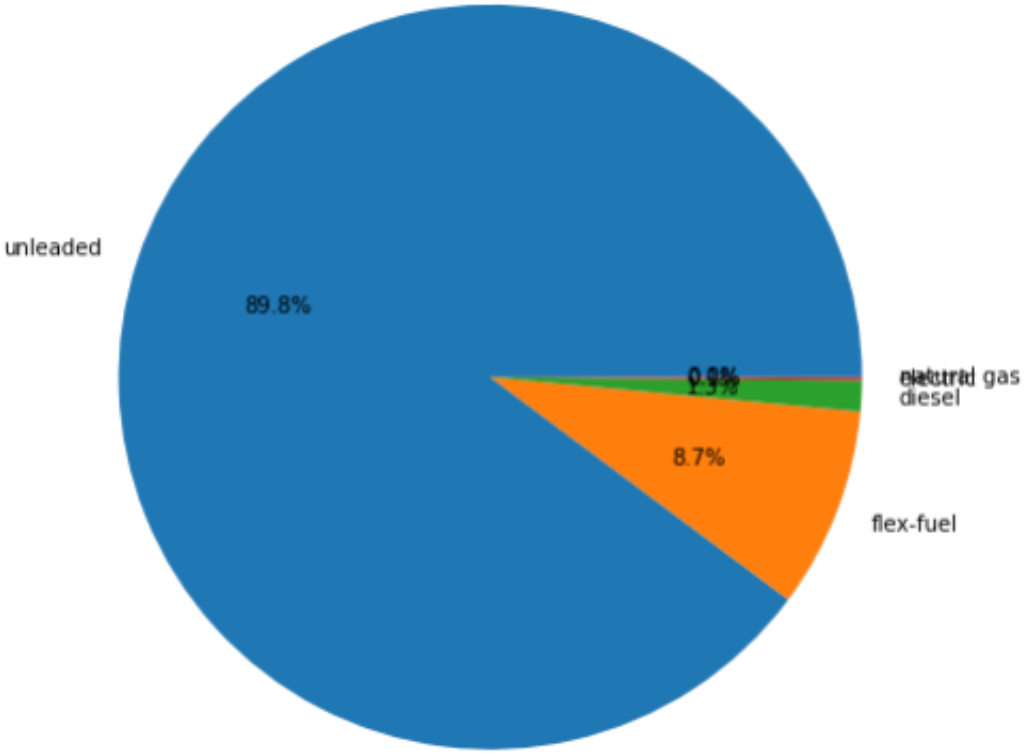

Figure 19: Pie Chart showing distribution of various Fuel Types.

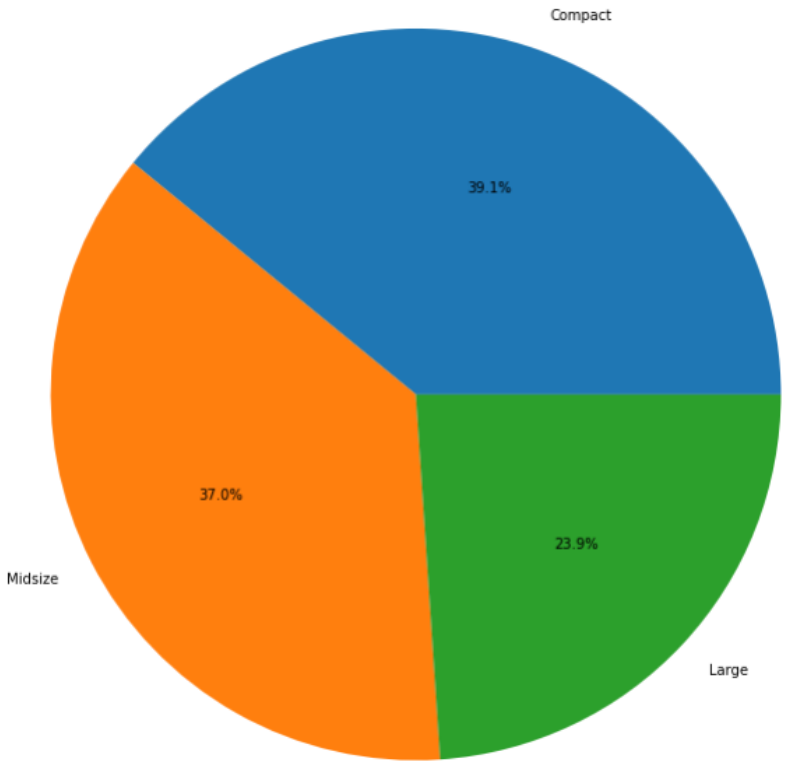

Figure 20: Pie Chart showing distribution of various Vehicle Sizes.

When it comes to Transmission Type, Fuel Type and Driving Wheels, Automatic Transmission, Unleaded Gasoline and Front-Wheel Drive respectively dominate the industry. Additionally, one can note that Chevrolet, Ford, Toyota, and Volkswagen, in that order, lead the market share and together constitute more than quarter of all brands listed in this dataset.

Certain grouped statistics were also found to be insightful. Here are some examples:

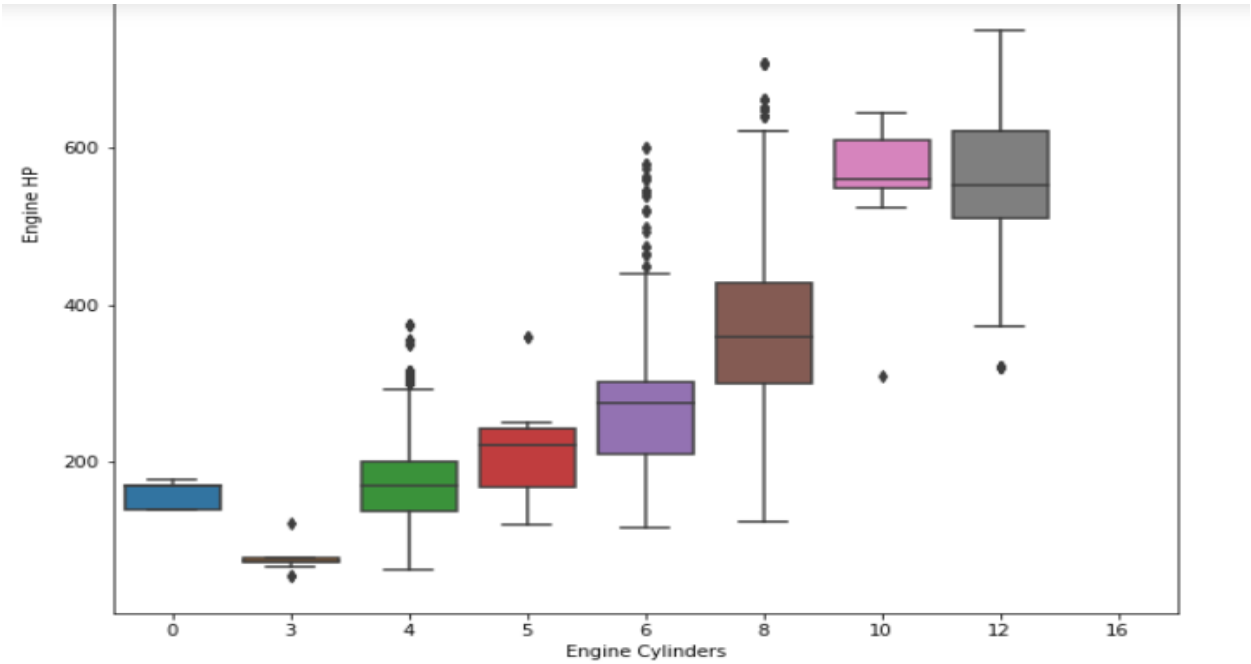

Figure 21: Boxplots showing the effect of the number of Cylinders in the Engine on its power output.

From Figure 21, one can clearly see that as the number of cylinders in the engine of a vehicle increases, the Engine HP value on an average also increases.

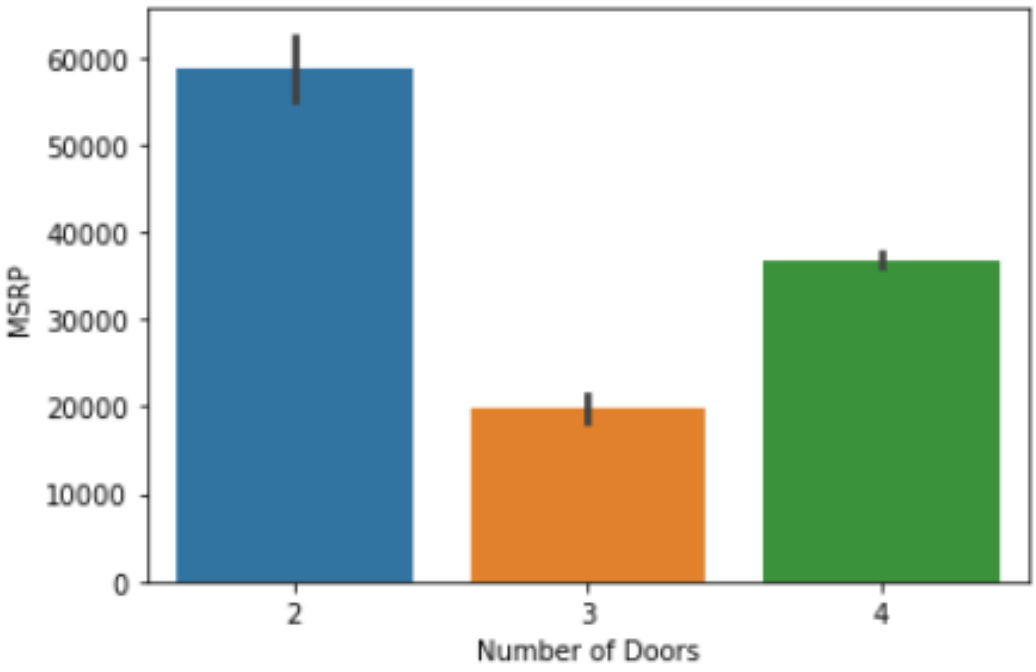

Figure 22: Boxplots showing the relationship between the number of car doors with the MSRP.

Figure 22 reminds us that most 2-door cars are the most expensive category among the three given types with an average MSRP of around 60000 followed by four-door cars with average MSRP of around 40000 and lastly, 3-door vehicles which tend to be minivans and trucks tend to be the cheapest at around 20000 as their average MSRP.

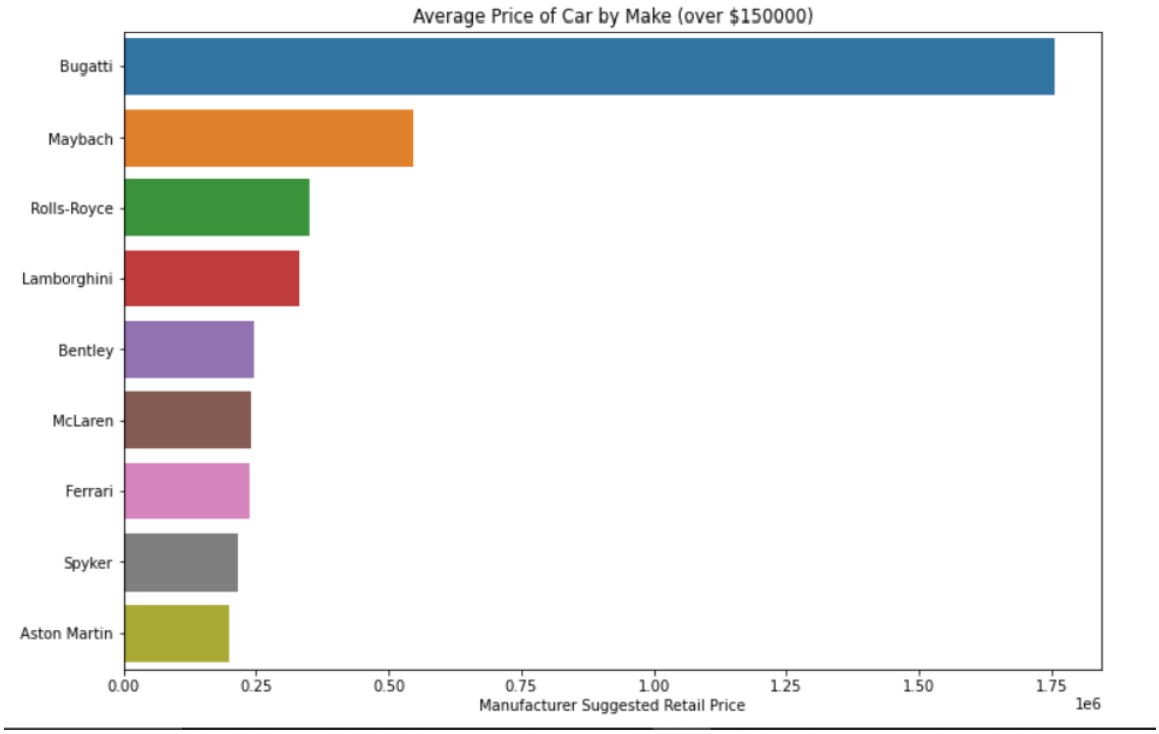

Figure 23: Average MSRP values for the topmost expensive brands.

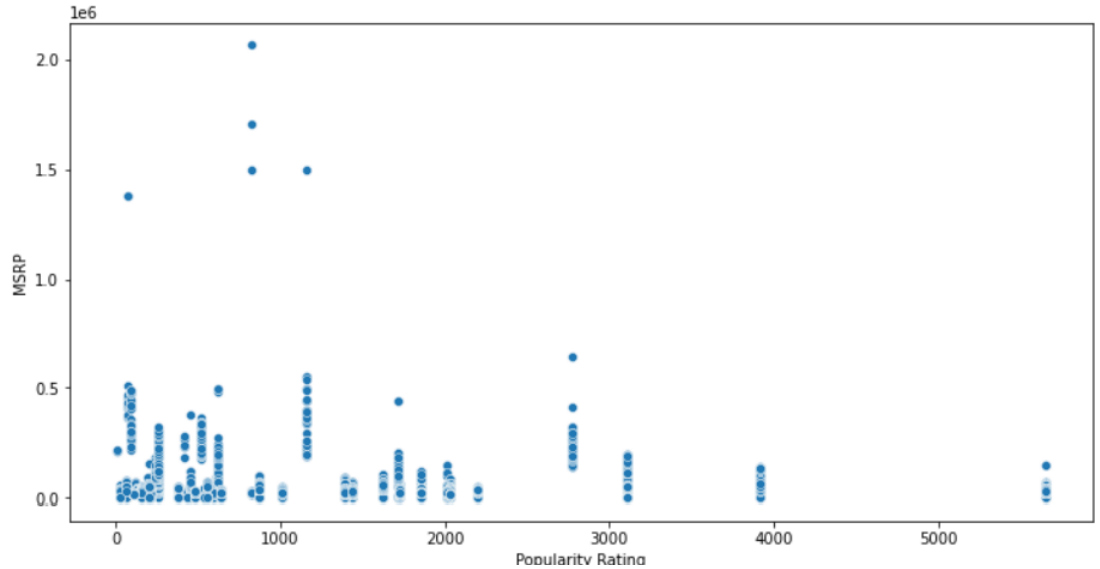

Figure 24: Scatter plot showing relationship between Popularity Rating and MSRP

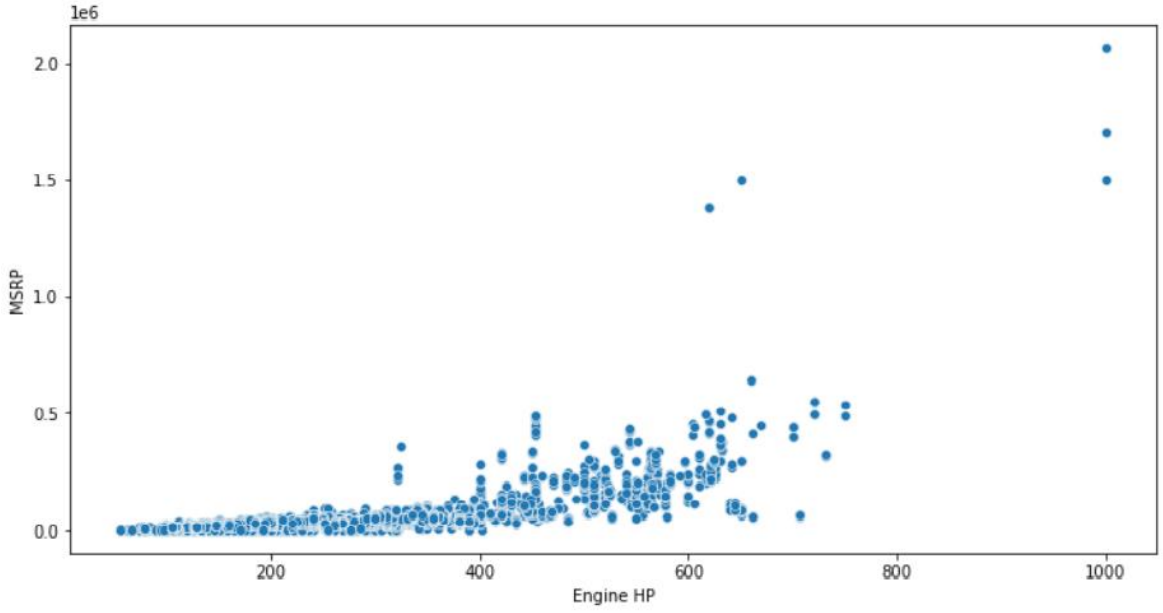

Figure 25: Scatter plot showing relationship between Engine HP and MSRP

From Figures 24 and 25, we can see that Engine HP is a better predictor of MSRP than the Popularity Rating.

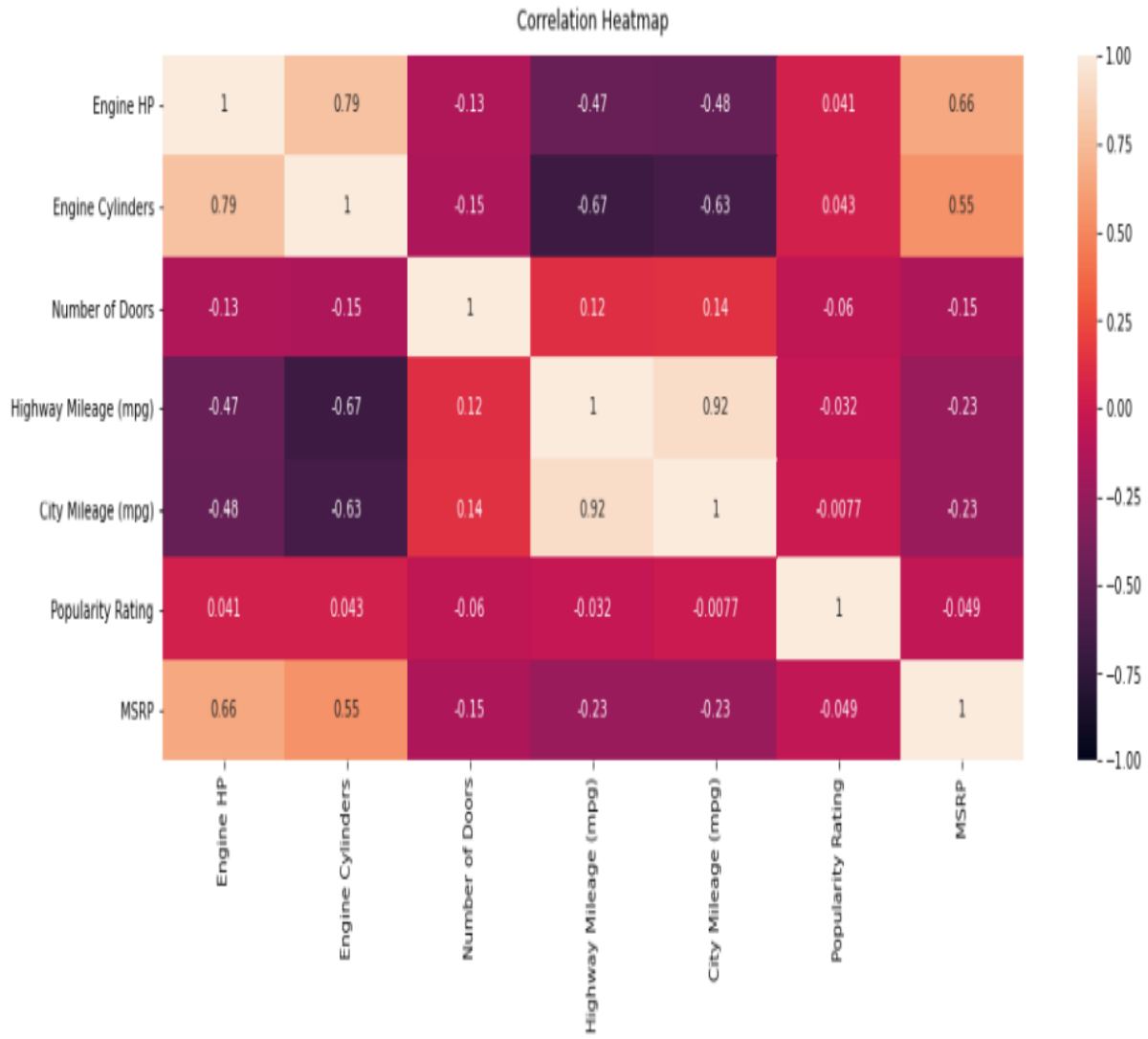

Figure 26: Correlation heatmap among all the numerical features.

One can see from the correlation heatmap illustrated in Figure 25 that Engine HP has a respectable correlation with the MSRP. This insight calls for a detailed regression analysis for the numerical features.

5. Regression Analysis

We have performed regression analysis on various features of the car to find out which feature best defines the price of the car, the results are listed below.

VEHICLE SIZE AND MSRP

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.022889
R Square	0.000524
Adjusted R	0.000434
Standard E	61717.23
Observatic	11097

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	2.22E+10	2.22E+10	5.815942	0.015898
Residual	11095	4.23E+13	3.81E+09		
Total	11096	4.23E+13			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	42983.51	738.0311	58.24079	0	41536.84	44430.19	41536.84	44430.19
Vehicle siz	-2926.74	1213.595	-2.41163	0.015898	-5305.6	-547.876	-5305.6	-547.876

Figure 27: Regression analysis Vehicle Size v/s MSRP

The regression analysis for vehicle size and MSRP was done, here multiple R value is 0.0228 and the R square is 0.0005 which shows weak linearity in the correlation. Hence, we can confirm that 0% variance in MSRP can be accounted for by the vehicle size measure. We can also see that the standard error is 61717.23, which confirms the comparison lacks precision in the regression model. The MSRP using this model can be determined by the equation

$$\mathbf{y = (-2926.74).x + 42983.51}$$

.

CITY MPG AND MSRP

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.232699
R Square	0.054149
Adjusted R	0.054064
Standard E	60038.75
Observatic	11097

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	2.29E+12	2.29E+12	635.1762	2.4E-136
Residual	11095	4E+13	3.6E+09		
Total	11096	4.23E+13			

	Coefficients	*tandard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	83880.31	1760.472	47.64649	0	80429.47	87331.14	80429.47	87331.14
City Mileaı	-2175.02	86.30096	-25.2027	2.4E-136	-2344.18	-2005.85	-2344.18	-2005.85

Figure 28: Regression analysis City Milage v/s MSRP

The regression analysis for city MPG (Miles per Gallon) and MSRP was done, here multiple R value is 0.2327 and the R square is 0.054 which shows weak linearity in the correlation. Hence, we can confirm that 5.41% variance in MSRP can be accounted for by the city MPG measure. We can also see that the standard error is 60038.75, which confirms the comparison lacks precision in the regression model. The MSRP using this model can be determined by the equation,

$$\mathbf{y = (-2175.02).x + 83880.31} \boldsymbol{y = -2175.02.x + 83880.31}$$

.

HIGHWAY MPG AND MSRP

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.207774
R Square	0.04317
Adjusted R	0.043084
Standard E	60386.19
Observatic	11097

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	1.83E+12	1.83E+12	500.5809	1.7E-108
Residual	11095	4.05E+13	3.65E+09		
Total	11096	4.23E+13			

	Coefficients	*standard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	86950.69	2093.518	41.53328	0	82847.02	91054.36	82847.02	91054.36
Highway M	-1714.65	76.63676	-22.3737	1.7E-108	-1864.87	-1564.42	-1864.87	-1564.42

Figure 29: Regression analysis Highway Milage v/s MSRP

The regression analysis for highway MPG (Miles per Gallon) and MSRP was performed, here multiple R value is 0.2077 and the R square is 0.054 which shows weak linearity in the correlation. Hence, we can confirm that 0% variance in MSRP can be accounted for by the highway MPG measure. We can also see that the standard error is 60386.19, which confirms the comparison lacks precision in the regression model. The MSRP using this model can be determined by the equation

$$\mathbf{y = (-1714.65).x + 86950.69} \boldsymbol{y = -1714.65.x + 86950.69}$$

POPULARITY AND MSRP

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.04849
R Square	0.002351
Adjusted R	0.002261
Standard E	61660.79
Observatic	11097

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	9.94E+10	9.94E+10	26.14912	3.21E-07
Residual	11095	4.22E+13	3.8E+09		
Total	11096	4.23E+13			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	45129.54	860.9341	52.41928	0	43441.96	46817.13	43441.96	46817.13
Popularity	-2.07322	0.40543	-5.11362	3.21E-07	-2.86793	-1.2785	-2.86793	-1.2785

Figure 30: Regression analysis Popularity rating v/s MSRP

The regression analysis for popularity and MSRP was performed, here multiple R value is 0.0484 and the R square is 0.0023 which shows weak linearity in the correlation. Hence, we can confirm that 0% variance in MSRP can be accounted for by the popularity measure. We can also see that the standard error is 61660.79, which confirms the comparison lacks precision in the regression model. The MSRP using this model can be determined by the equation,

$$\mathbf{y = (-2.07).x + 45128.54}$$

ENGINE HP AND MSRP

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.658983
R Square	0.434259
Adjusted R	0.434208
Standard E	46433.25
Observatic	11097

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	1.84E+13	1.84E+13	8516.445	0
Residual	11095	2.39E+13	2.16E+09		
Total	11096	4.23E+13			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	-51716.1	1106.065	-46.7568	0	-53884.2	-49548	-53884.2	-49548
Engine HP	369.2597	4.001315	92.28459	0	361.4164	377.103	361.4164	377.103

Figure 31: Regression analysis Engine v/s MSRP

Finally, the regression analysis for Engine HP and MSRP was performed, here the multiple R value is 0.6589 and the R square is 0.4342 which shows strong linearity in the correlation. Hence, we can confirm that 43.42% variance in MSRP can be accounted for by the popularity measure, which is the highest between the rest of the regression models above. We can also confirm using ANOVA table, the significance F value is 0 and hence the alternate hypothesis satisfies the condition, which confirms the comparison is linear in the regression model. The Engine HP serves as the best feature to determine the price of the car. The MSRP using this model can be determined by the equation,

$$\mathbf{y = 369.25.x + (-51716.1)}$$

6. Data Modeling & Dashboard

Power BI is a combination of software services, applications, and connections that collaborate to transform your disparate data sources into coherent, eye - catching insights that are engaging. The dataset used in Power BI may be stored in a hybrid central repository that is both cloud-based and on-premises, or it could be an Excel spreadsheet. Power BI makes it simple to connect to your databases, view the data, identify the key information, and share it with whoever you choose [6].

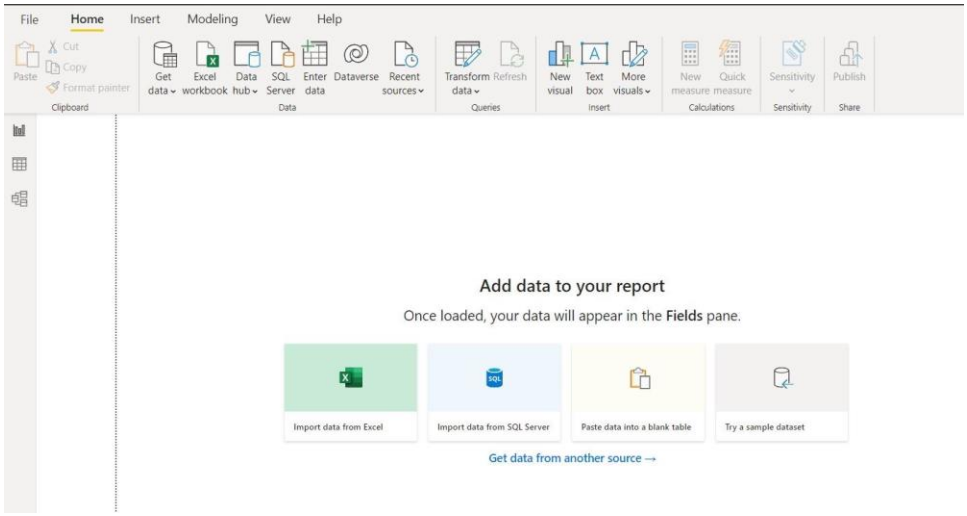

Figure 32: PowerBI interface

Power BI uses multiple tools like Microsoft teams, PowerPoint, and Excel for data visualization and analysis. The most popular data analysis program used by businesses and information professionals worldwide is Microsoft Excel. To connect the dataset in Power BI, we must click the get data ribbon on the home tab in Power BI, then we can add files with different extensions like csv, SQL query, xlsx, etc. A similar method can be followed in Excel, we must select the Power BI (Microsoft) option from the data ribbon in Excel [6].

After the data is connected to Power BI, we can initiate Data modeling. The process of describing and evaluating every type of data your company creates and gathers, as well as the connections between those data points, is known as data modelling. Data modelling is a practice in comprehending and outlining your needs, and it generates visual representation as it is deployed in any firm. Data analytics and modelling go hand in hand because you need a powerful data model to achieve the most useful analysis for business information that guides decision-making. Each franchise is forced to consider how they make a significant contribution to overarching business objectives because of the driving feature of the data model creation process. Analysis of the precise data you want is made considerably simpler when all your data is well-defined. It is straightforward to assess and identify effects when you modify procedures, pricing, or personnel since the linkages between data elements have already been established [7]. Data modeling utilizes the concept of Star schema and Snowflake schema, wherein we use a fact table which has relationships with several other dimensions tables. When the Fact table is connected to other dimension tables, it represents Star schema and when the Dimension table has its own unique branches, it represents snowflake schema.

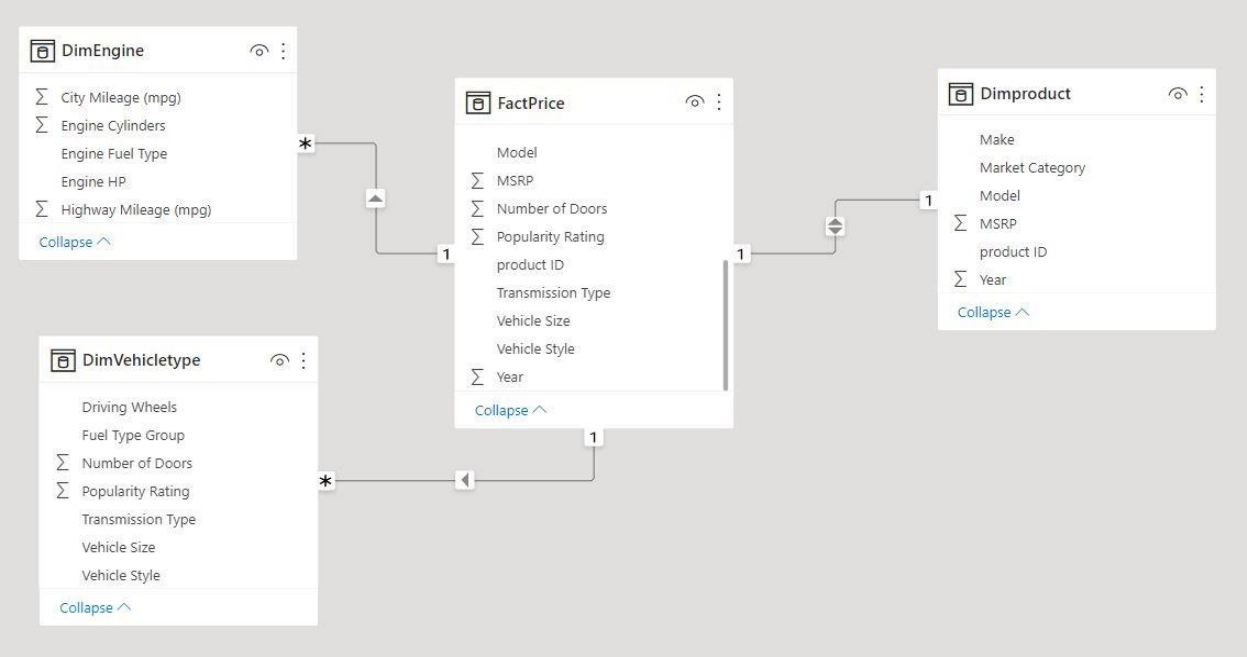

Figure 33: Data Model

The image above represents the data model for our dataset. The central chart is the fact table, and the other three charts are dimension tables. The fact table has a one-to-many relationship and filter direction with the dimension table. Our dataset did not have a product ID column; hence a product ID column was added in our cleaned dataset to set the relationships.

After the data modeling was done, we proceeded to the visualization step. Slicers were used to select multiple options from the columns so that if customers want to know the price and everything, they can select the number of doors or other things. Then a pie chart and a donut chart were used to show the relationships of the whole part. A line chart was used to show the average price by year which shows the overall shape of an entire series. Single cards were used to show a single data point like the average price of the car, etc. Treemap is used as there was a large data to represent which gives valuable information. A treemap is used instead of a bar chart as it could not provide clear visualization. A Scatter plot is used to show the relationship between two points and thus we have used it to find a correlation between the popularity rating and MSRP, and the other scatter plot was used to find a correlation between engine hp and MSRP.

7. Final Conclusions & Recommendations

In this project, we focused on comparing various data characteristics in this project and looked for patterns that would increase the car's marketability. Based on our preliminary analysis, we have linked the features or columns of the car and the impact they will have on the pricing. To evaluate the data, we first must clean the data to its finest by using all the necessary measures like data cleaning, EDA, regression, etc., to derive insights for Dashboarding and Visualization.

Some limitations of this project were that the data set acquired for this project was just a one-table data set with already normalized tuples. This resulted in difficulty in making a data model for the PowerBI. One recommendation would be to acquire more datasets that could be joined with the existing database. This will result in a highly effective dashboard. Also, regarding the dashboard, a deeper analysis could be done with the drill-through features of PowerBI. Finally, assuming enough data is gathered, one could consider implementing various Machine Learning algorithms on the data to extract further insights into the features of an automobile that influences its MSRP.

8. Task Allocation

Yash Dalvi (300257910)

- Identifying the correct hypothesis for the given situation.
- Feature scaling and Normalization of data to achieve average value of MSRP for each fuel type and plotting a bar graph for the same.
- Encoding categorical data to ascertain the vehicle size relationship with MSRP.
- Hypothesis testing using regression analysis to reify car’s MSRP using distinctive car features.
- Data modeling using Power BI.

Shivani Deshmukh (300251019)

- Identifying the use of statistical analysis in the dataset and providing an example of the same
- Using a pivot chart to get an insight into multiple interrelated values
- Using different charts for EDA to get detailed information about the dataset
- Dashboard using Power BI.
- Data visualization part of report.

Vikram Khanzode (300198886)

- All data cleaning and preparation using Pandas.
- Data visualization using matplotlib and Seaborn.
- Introduction part of report.

- Data visualization part of report.

Rishi Kumar Pandey (300219699)

- Introduction part of report.
- Data Acquisition

Data Source

Following is the link to the dataset we collected from Kaggle.

https://www.kaggle.com/datasets/CooperUnion/cardataset

References

[1] What is price analysis and why is it important? | Minderest

[2] Market Saturation - Overview, Impact, How to Avoid (corporatefinanceinstitute.com)

[3] Auto Industry Shrinking at 'Peak Car,' Dragging Global Economy Lower (businessinsider.com)

[4] Market Saturation - Overview, Impact, How to Avoid (corporatefinanceinstitute.com)

[5] Auto Industry Shrinking at 'Peak Car,' Dragging Global Economy Lower (businessinsider.com)

[6] Car Features EDA | Kaggle

[7] https://www.kaggle.com/datasets/CooperUnion/cardataset

[8] https://www.edmunds.com/

[9] https://databox.com/why-are-dashboards-important#head1

[10] https://powerbi.microsoft.com/en-us/

[11] https://learn.microsoft.com/en-us/power-bi/create-reports/service-dashboards

[12] https://learn.microsoft.com/en-us/power-bi/collaborate-share/service-connect-power-bi-datasets-excel

[13] What is Data Modeling | Microsoft Power BI

[14] Time Series Analysis: Definition, Types & Techniques | Tableau

[15] automotive industry - Highway development | Britannica

[16] The Benefits of Data Collection for Manufacturing Companies (veryableops.com)