Analysing the Impact of Car Features on Price and Profitability

Final Project-3

Project Description:

The automotive industry has been rapidly evolving over the past few decades, with a growing focus on fuel efficiency, environmental sustainability, and technological innovation. With increasing competition among manufacturers and a changing consumer landscape, it has become more important than ever to understand the factors that drive consumer demand for cars. In recent years, there has been a growing trend towards electric and hybrid vehicles and increased interest in alternative fuel sources such as hydrogen and natural gas. At the same time, traditional gasoline-powered cars remain dominant in the market, with varying fuel types and grades available to consumers.
For the given dataset, as a Data Analyst, the client has asked how can a car manufacturer optimize pricing and product development decisions to maximize profitability while meeting consumer demand?
This problem could be approached by analyzing the relationship between a car's features, market category, and pricing, and identifying which features and categories are most popular among consumers and most profitable for the manufacturer. By using data analysis techniques such as regression analysis and market segmentation, the manufacturer could develop a pricing strategy that balances consumer demand with profitability, and identify which product features to focus on in future product development efforts. This could help the manufacturer improve its competitiveness in the market and increase its profitability over time.

Dataset Description:

The dataset contains information on various car models and their specifications, and is titled "Car Features and MSRP". It was collected and made available on Kaggle by Cooper Union; a private college located in New York City.

Here is a brief overview of the dataset:

- **Number of observations:** 11,915
- **Number of variables:** 16
- **File type:** CSV (Comma Separated Values)

The variables in the dataset are:

- **Make:** the make or brand of the car
- **Model:** the specific model of the car
- **Year:** the year the car was released
- **Engine Fuel Type**: the type of fuel used by the car (gasoline, diesel, etc.)
- **Engine HP:** the horsepower of the car's engine
- **Engine Cylinders:** the number of cylinders in the car's engine

- **Transmission Type**: the type of transmission (automatic or manual)
- **Driven_Wheels:** the type of wheels driven by the car (front, rear, all)
- **Number of Doors:** the number of doors the car has
- **Market Category:** the market category the car belongs to (Luxury, Performance, etc.)
- **Vehicle Size:** the size of the car
- **Vehicle Style:** the style of the car (Sedan, Coupe, etc.)
- **Highway MPG:** the estimated miles per gallon the car gets on the highway
- **City MPG:** the estimated miles per gallon the car gets in the city
- **Popularity:** a ranking of the popularity of the car (based on the number of times it has been viewed on Edmunds.com)
- **MSRP:** the manufacturer's suggested retail price of the car

This dataset could be useful for a variety of data analysis tasks, such as:

- Exploring trends in car features and pricing over time
- Comparing the fuel efficiency of different types of cars
- Investigating the relationship between a car's features and its popularity
- Predicting the price of a car based on its features and market category

Tech-Stack Used:

- **Microsoft Excel 2021** — A spreadsheet editor software used mainly by professionals to enter data in table format, perform computations, plot graphs etc. Here Microsoft Excel is used to filter data and plot graphs to get insights about the movies.
- **Microsoft Word:** A word processing application for preparing report.

Data Pre-Processing: Before diving into the analysis of the given dataset, it is important to perform thorough data cleaning to ensure accurate and reliable results.

Handling Duplicate Values

- Found duplicate rows on analysis. Except the first instance, dropped all other duplicate rows. Handling Null values. After removing duplicate values, the dataset contained 11199 rows.

Handling NULL Values

- For Null values in **Engine Fuel Type** column, we found that there were only one such Car model. For Suzuki Make and Verona model in year 2004 was missing therefore we replaced it "regular unleaded" because for Suzuki Make and Verona Model year in 2005 and 2006 all was "regular unleaded".
- For Null values in **Engine HP** column, we searched the value of **Engine HP** by searching for rows with same **Make**, **Model** and **Year**. If found then replaced null value with the mode of **Engine HP** value of all the matching rows. For rest of the rows which were unaffected by above process, we found that they were all **Electric Cars**. So searched for the Car model's **Engine HP (Motor's Power)** in the website **evcompare.io** and replaced null values with its correct values.
- For Null values in **Market Category** column,
  — Separated the Categories into different columns.
  — Searched for **Market Categories** of all rows of same Car **Make** and **Model** in non-null

values
— Found mode of the **Market Categories** and replaced null value with the mode.

- For Null values in **Market Category** column which were not affected by the previous process, we left it with N/A so in total we had 3250 market categories. out of these 3188 were regular unleaded, 5 electrics, 9 flex fuel, 2 natural gas, 31 premium unleaded (recommended), 15 premiums unleaded (required) Engine fuel type. Replaced all with Crossover because the mode of crossover was highest.
- Column **Transmission Type** had some (12) rows with column value '**UNKNOWN**', we searched for the information online and replaced the value '**UNKNOWN**' with correct ones. (All were automatic).

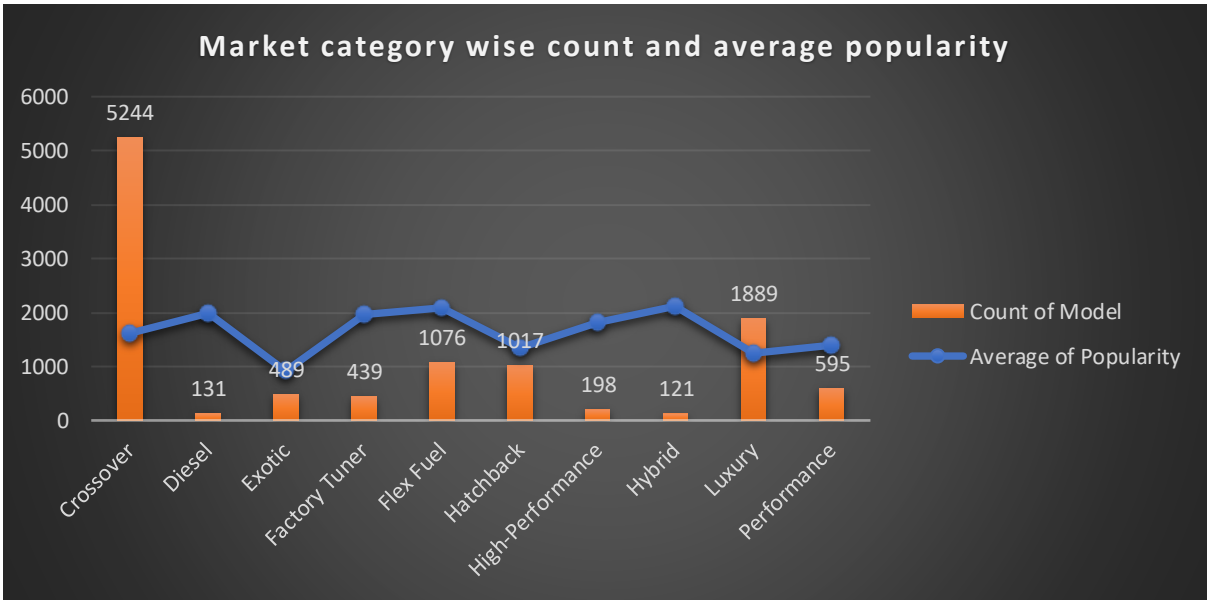
Tasks: Analysis

Insight Required: How does the popularity of a car model vary across different market categories?

- **Task 1.A:** Create a pivot table that shows the number of car models in each market category and their corresponding popularity scores.

Row Labels	Average of Popularity	Count of Model
Crossover	1615	5244
Diesel	1977	131
Exotic	915	489
Factory Tuner	1963	439
Flex Fuel	2086	1076
Hatchback	1340	1017
High-Performance	1823	198
Hybrid	2117	121
Luxury	1250	1889
Performance	1396	595
Grand Total	**1558**	**11199**

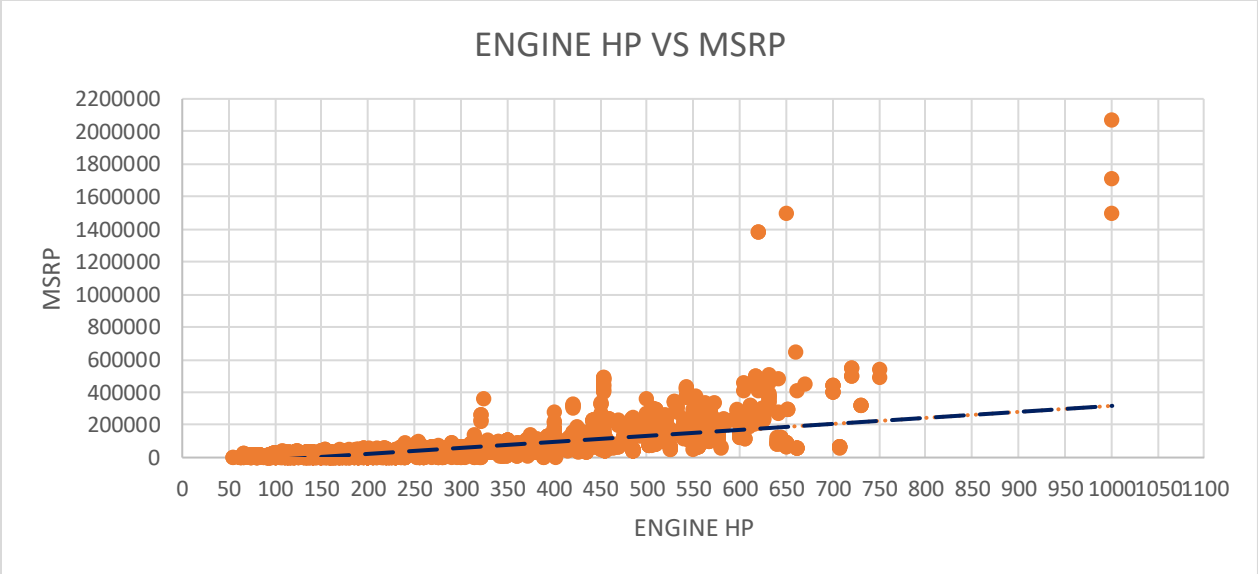
- **Task 1.B:** Create a combo chart that visualizes the relationship between market category and popularity.



- We can observe that the average popularity of cars based on their Market Category mainly ranges from 1200 to 2117 with the exception of Exotic, Luxury cars being the lowest popular and Hybrid, Flex Fuel cars being the most popular.
- The dataset has comparatively higher number of Crossover followed by Luxury cars.

Insight Required: What is the relationship between a car's engine power and its price?

- **Task 2:** Create a scatter chart that plots engine power on the x-axis and price on the y-axis. Add a trendline to the chart to visualize the relationship between these variables.

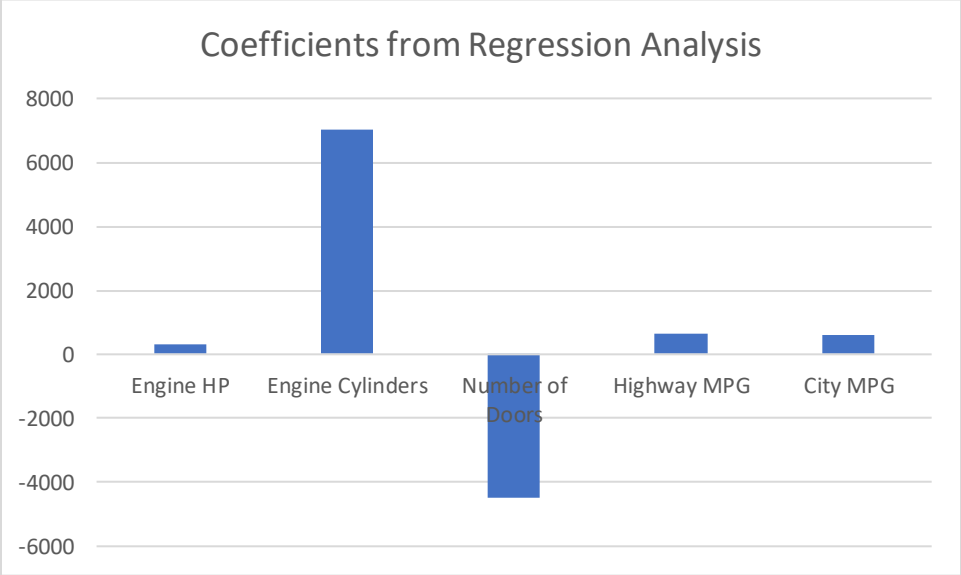


- We can observe that the relationship is **positive** as the trendline has **positive** slope. This is logical as higher **Engine HP** requires more complex level of **design** and **engineering** and more expensive sub-parts.

Insight Required: Which car features are most important in determining a car's price?

- **Task 3:** Use regression analysis to identify the variables that have the strongest relationship with a car's price. Then create a bar chart that shows the coefficient values for each variable to visualize their relative importance.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.6765832							
R Square	0.4577649							
Adjusted R Square	0.4575226							
Standard Error	45322.448							
Observation	11199							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	5	1.94101E+13	3.882E+12	1889.86714	0			
Residual	11193	2.29918E+13	2054124257					
Total	11198	4.24019E+13						
	Coefficient	*Standard Error*	*t-Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-91053.9	3598.181332	-25.3055347	1.94E-137	-98106.971	-84000.83	-98106.971	-84000.834
Engine HP	312.93191	6.306453686	49.6208998	0	300.5701476	325.2937	300.570148	325.293665
Engine Cylinders	7025.5619	456.0544051	15.4050959	5.2843E-53	6131.61498	7919.509	6131.61498	7919.50873
Number of Doors	-4470.942	496.9279569	-8.99716313	2.6891E-19	-5445.00812	-3496.876	-5445.0081	-3496.8757
Highway MPG	658.7235	107.3018055	6.13897869	8.5867E-10	448.3930792	869.0539	448.393079	869.053916
City MPG	603.95506	101.5859276	5.94526302	2.8425E-09	404.8287667	803.0814	404.828767	803.081351



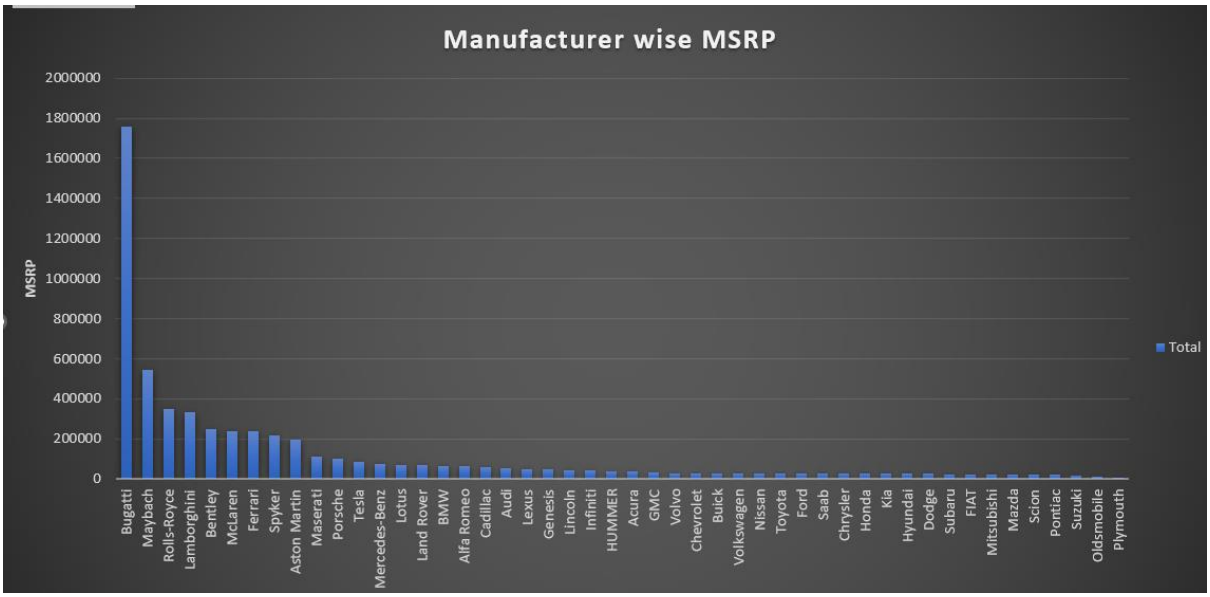
- We can observe that the highest coefficient value is that of **Engine Cylinders.** This shows that the Engine Cylinders is very important relationship with Car's price.

Insight Required: How does the average price of a car vary across different manufacturers?

- **Task 4.A:** Create a pivot table that shows the average price of cars for each manufacturer.

Row Labels	Average of MSRP
Bugatti	1757223.667
Maybach	546221.875
Rolls-Royce	351130.6452
Lamborghini	331567.3077
Bentley	247169.3243
McLaren	239805
Ferrari	238218.8406
Spyker	214990
Aston Martin	198123.4615
Maserati	113684.4909
Porsche	101622.3971
Tesla	85255.55556
Mercedes-Benz	72069.52786
Lotus	68377.14286
Land Rover	68067.08633
BMW	62162.55864
Alfa Romeo	61600
Cadillac	56368.26515
Audi	54574.1215
Lexus	47549.06931
Genesis	46616.66667
Lincoln	43860.825
Infiniti	42640.27134
Infiniti	42640.27134
HUMMER	36464.41176
Acura	35087.4878
GMC	32444.08506
Volvo	29724.68421
Chevrolet	29074.72576
Buick	29034.18947
Volkswagen	28978.52289
Nissan	28921.15245
Toyota	28846.5605
Ford	28511.30788
Saab	27879.80734
Chrysler	26722.96257
Honda	26655.14781
Kia	25513.75546
Hyundai	24926.26255
Dodge	24857.04537
Subaru	24240.67364
FIAT	22670.24194
Mitsubishi	21340.5625
Mazda	20416.62379
Scion	19932.5
Pontiac	19800.0442
Suzuki	18026.4152
Oldsmobile	12843.79545
Plymouth	3296.873239

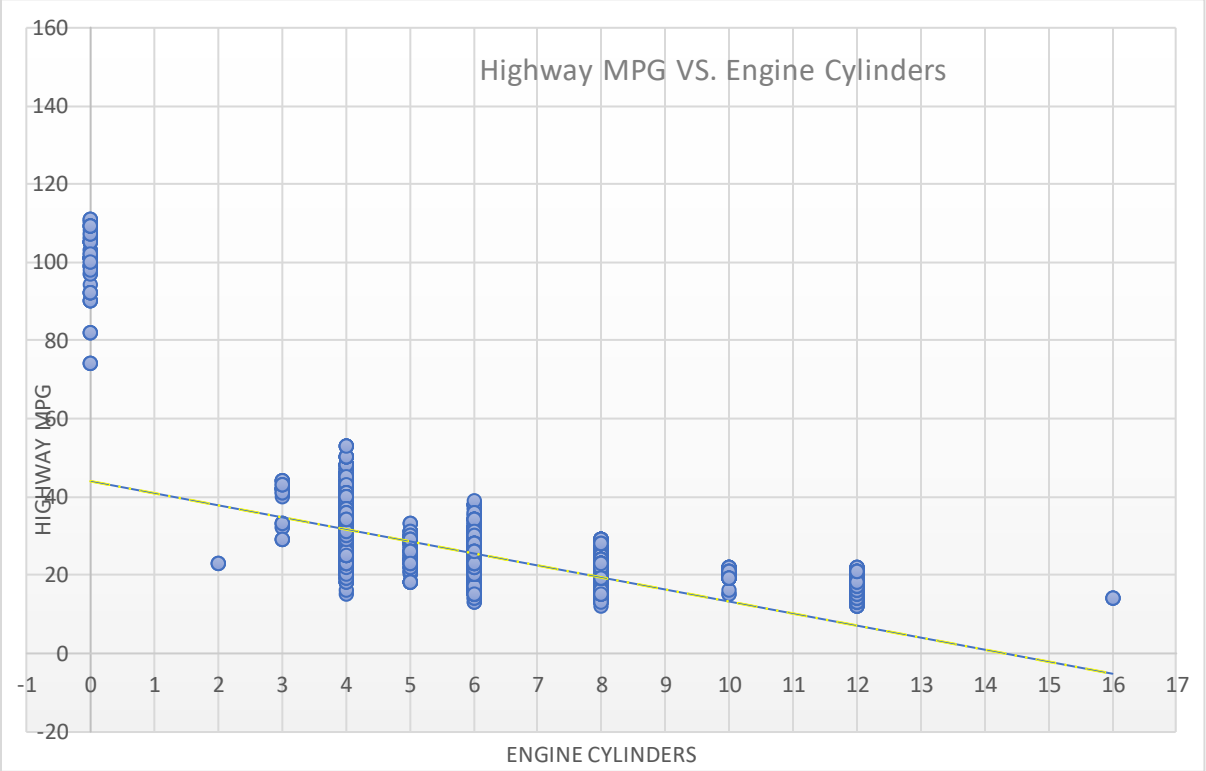
- **Task 4.B:** Create a bar chart or a horizontal stacked bar chart that visualizes the relationship between manufacturer and average price.



- We can observe that the most expensive cars are that of **Bugatti** brand followed by **Maybach**, **Rolls-Royce**, **Lamborghini** etc. All these cars' brands are **High-Performance** and **Luxury** brands.

Insight Required: What is the relationship between fuel efficiency and the number of cylinders in a car's engine?

- **Task 5.A:** Create a scatter plot with the number of cylinders on the x-axis and highway MPG on the y-axis. Then create a trendline on the scatter plot to visually estimate the slope of the relationship and assess its significance.



- **Task 5.B:** Calculate the correlation coefficient between the number of cylinders and highway MPG to quantify the strength and direction of the relationship.

Correlation between Engine Cylinders and Highway MPG
-0.617805321

- We can observe that the plot between **highway MPG** and **Engine Cylinders** has a negative slope.
- The correlation coefficient is also **Negative** with a value of **-0.6178**.
- This is logical because as number of **Engine Cylinders** increases, the amount of fuel to be burnt also increasing, thus decreasing the mileage (**highway MPG**).

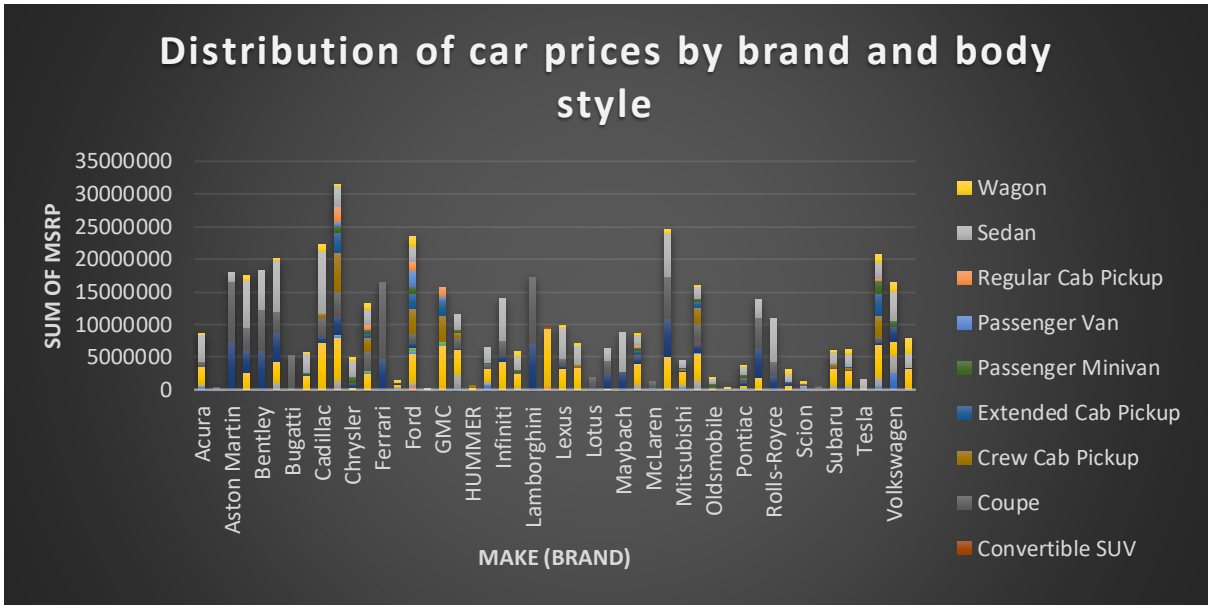
Building the Dashboard:

Now for the Next portion of the Project, we need to create the Interactive Dashboard.

Use filters and slicers to make the chart interactive. The client has requested these questions given below:

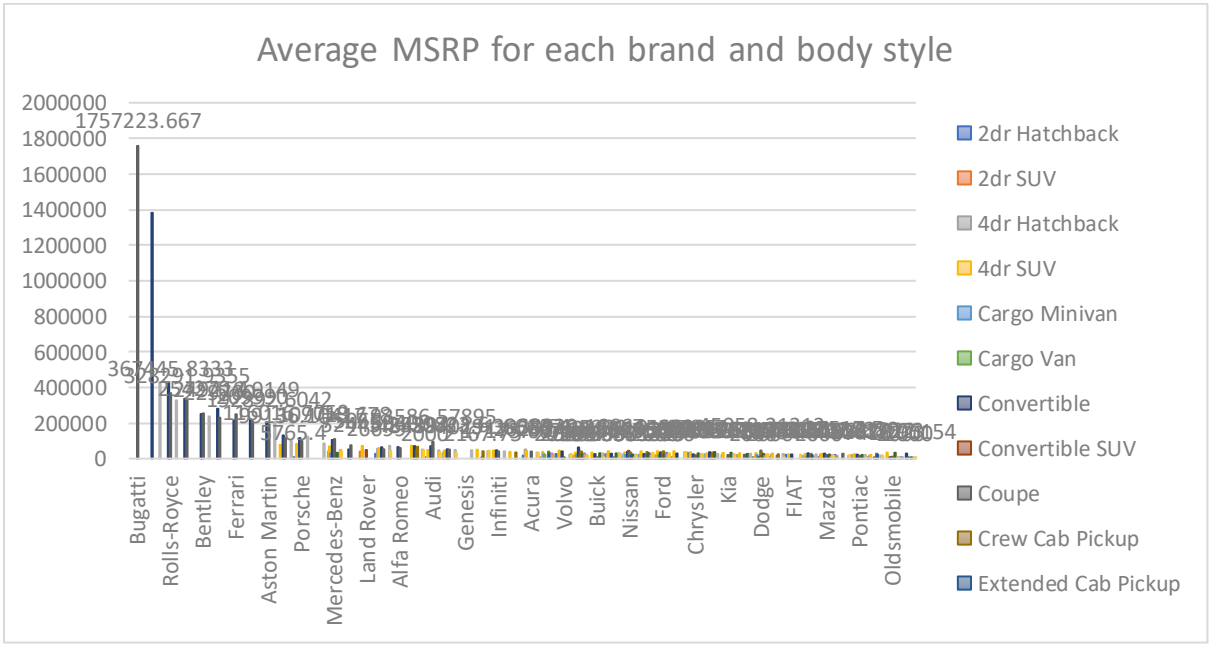
Task 1: How does the distribution of car prices vary by brand and body style?

- **Hints:** Stacked column chart to show the distribution of car prices by brand and body style. Use filters and slicers to make the chart interactive. Calculate the total MSRP for each brand and body style using SUMIF or Pivot Tables.



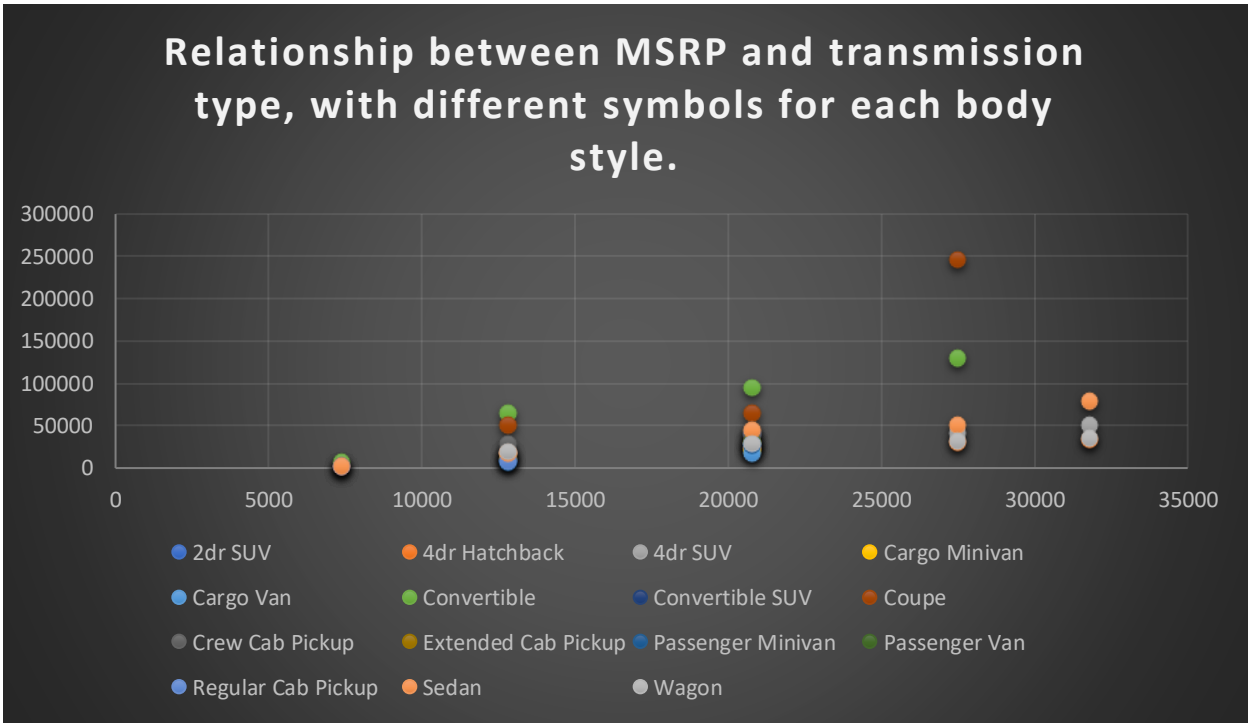
Task 2: Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?

- **Hints:** Clustered column chart to compare the average MSRPs across different car brands and body styles. Calculate the average MSRP for each brand and body style using AVERAGEIF or Pivot Tables.



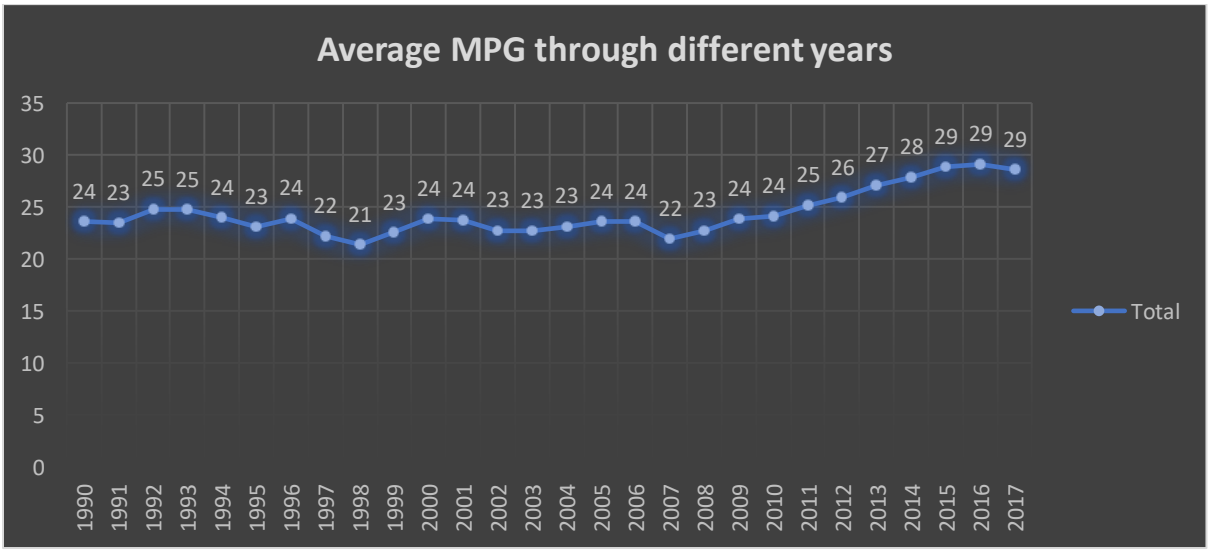
Task 3: How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?

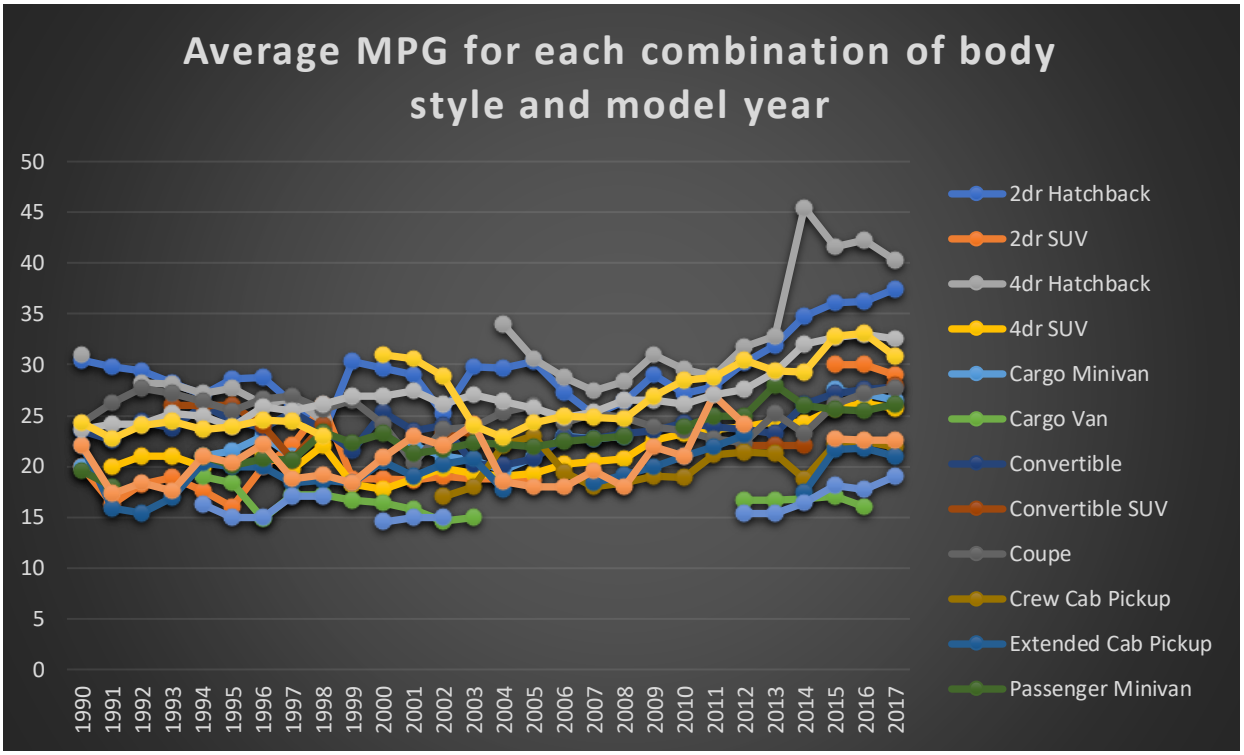
- **Hints:** Scatter plot chart to visualize the relationship between MSRP and transmission type, with different symbols for each body style. Calculate the average MSRP for each combination of transmission type and body style using AVERAGEIFS or Pivot Tables.



Task 4: How does the fuel efficiency of cars vary across different body styles and model years?

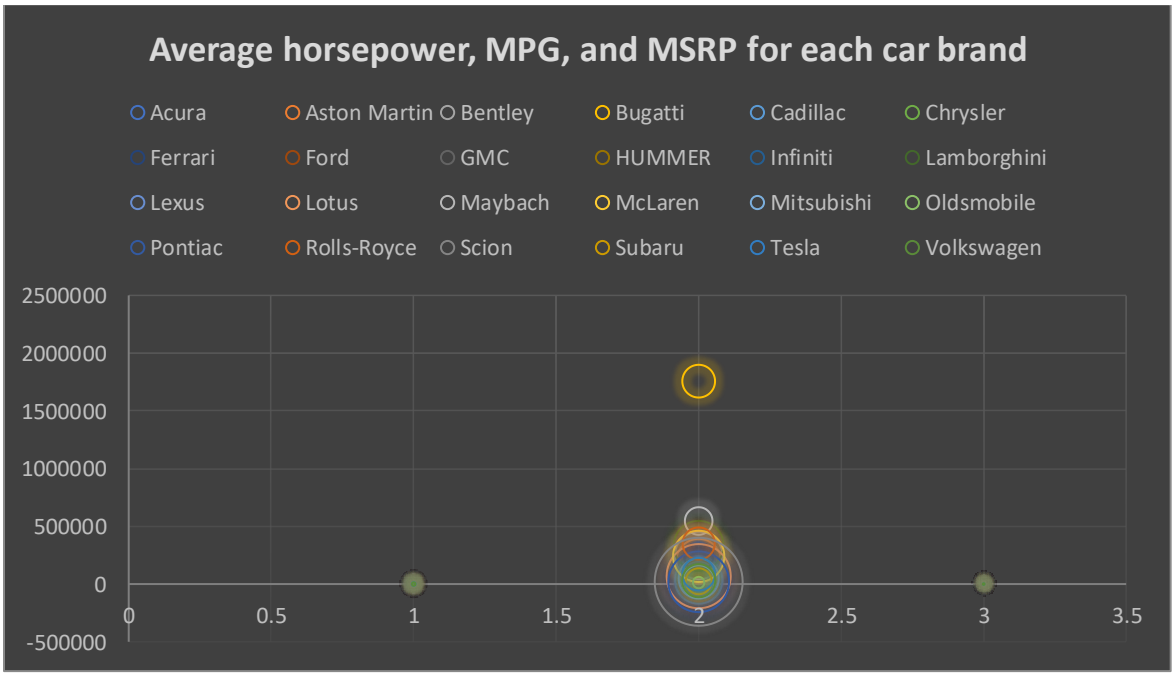
- **Hints:** Line chart to show the trend of fuel efficiency (MPG) over time for each body style. Calculate the average MPG for each combination of body style and model year using AVERAGEIFS or Pivot Tables.



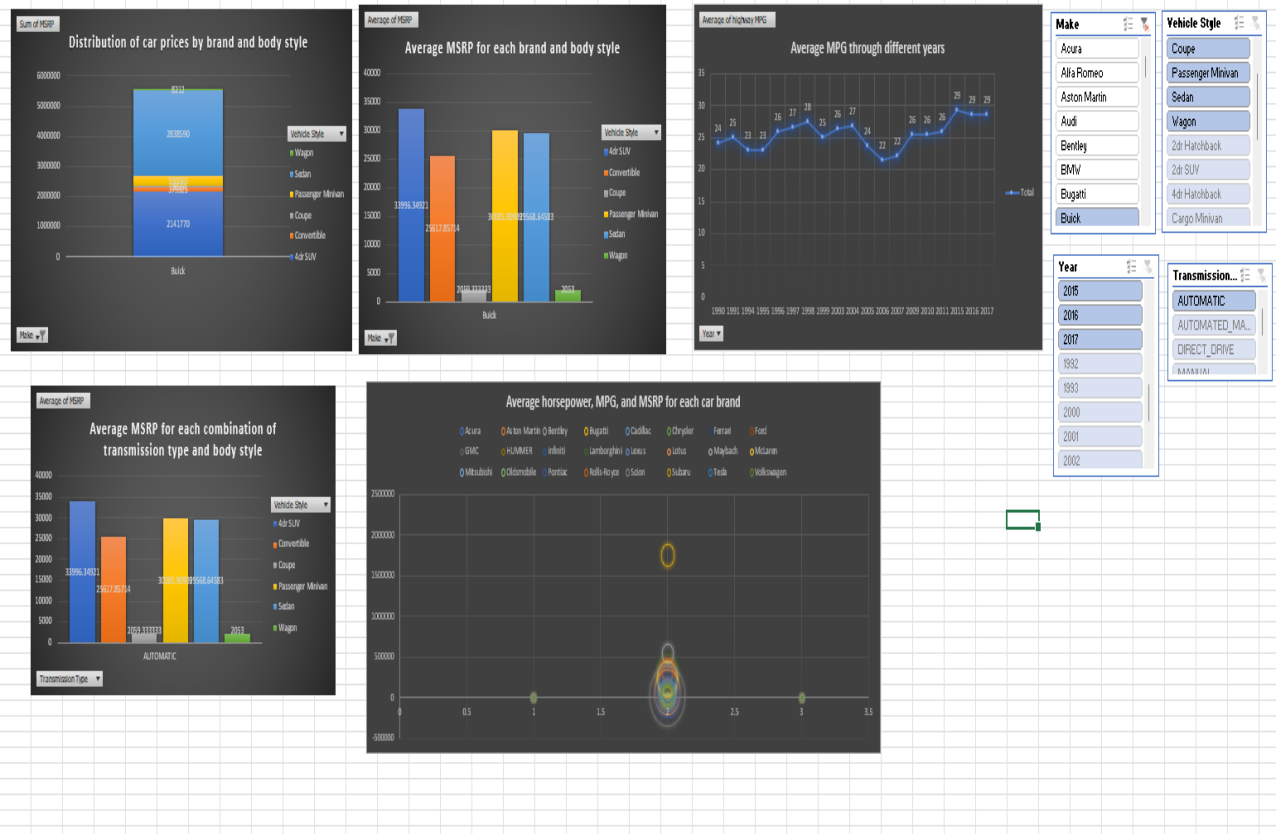


Task 5: How does the car's horsepower, MPG, and price vary across different Brands?

- **Hints:** Bubble chart to visualize the relationship between horsepower, MPG, and price across different car brands. Assign different colours to each brand and label the bubbles with the car model name. Calculate the average horsepower, MPG, and MSRP for each car brand using AVERAGEIFS or Pivot Tables.



DASHBOARD



Snap of Dashboard.

Excel Sheet LINK

https://drive.google.com/drive/folders/17LgJ-2H29QGvsXeH74xjAiFP85HPGNus?usp=sharing

Video Link: https://www.loom.com/share/996f34f10bb2457c891eda4bfd52d1a4?sid=0e4da428-8437-452a-9225-b74adcf0dee3