Impact of Car Features on Price and Profitability

Project Description:

The objective of this project is to assist a car manufacturer in optimizing pricing and product development decisions to maximize profitability while meeting consumer demand.

The project aims to analyse the relationship between a car's features, market category, and pricing to identify the most popular and profitable factors that drive consumer demand.

By conducting data analysis and visualization, we aim to provide valuable insights to optimize decision-making for the car manufacturer and meet consumer demand effectively.

Goals:

1. **Understand Car Model Popularity:** Explore the dataset to determine how the popularity of car models varies across different market categories. Identify which car models are most preferred by consumers within specific market segments.

2. **Analyse Engine Power and Price Relationship:** Investigate the relationship between a car's engine power and its price. Determine whether higher engine power translates to higher pricing and understand the dynamics between these two factors.

3. **Identify Key Factors in Price Determination:** Identify the car features that significantly influence a car's price. Determine which features are most important in determining the price of a car and their impact on consumer demand.

4. **Evaluate Average Prices by Manufacturer:** Assess how the average price of cars varies across different manufacturers. Identify which manufacturers have the highest and lowest average MSRPs and examine how this varies by body style.

5. **Study price variation with Transmission Type:** Investigate how different transmission types affect the Manufacturer's Suggested Retail Price (MSRP) of cars and how this relationship varies by body style. Analyse the impact of transmission types on pricing and understand the preferences of consumers in different body styles.

6. **Examine Fuel Efficiency across Body Styles and Model Years:** Analyse the dataset to determine how the fuel efficiency of cars varies across different body styles and model years. Identify trends and patterns in fuel efficiency to understand the preferences and demands of consumers for different types of vehicles.

7. **Explore Horsepower, MPG, and Price Variation by Brands:** Investigate the variations in horsepower, MPG (Miles Per Gallon), and price across different car brands. Analyse how these factors differ among various brands and understand the relationships between horsepower, MPG, and price.

The findings will help optimize the manufacturer's offerings, enhance competitiveness, and align their strategies with the demands of the market and target consumers.

Dataset Description:

The "**Car Features and MSRP**" dataset contains information on various car models and their specifications.

The raw dataset consists of 11,915 observations and 16 variables. The dataset was obtained in CSV format and from Kaggle, made available by Cooper Union, New York.

Variables in the raw dataset include:

1. **Make**: The brand or make of the car.
2. **Model**: The specific model of the car.
3. **Year**: The year the car was released.
4. **Engine Fuel Type**: The type of fuel used by the car (e.g., gasoline, diesel).
5. **Engine HP**: The horsepower of the car's engine.
6. **Engine Cylinders**: The number of cylinders in the car's engine.
7. **Transmission Type**: The type of transmission (automatic or manual).
8. **Driven Wheels**: The type of wheels driven by the car (e.g., front, rear, all).
9. **Number of Doors**: The number of doors the car has.
10. **Market Category**: The market category the car belongs to (e.g., Luxury, Performance).
11. **Vehicle Size**: The size of the car.
12. **Vehicle Style**: The style of the car (e.g., Sedan, Coupe).
13. **Highway MPG**: The estimated miles per gallon the car gets on the highway.
14. **City MPG**: The estimated miles per gallon the car gets in the city.
15. **Popularity**: A ranking of the popularity of the car based on views on Edmunds.com.
16. **MSRP**: The manufacturer's suggested retail price of the car.

Data Cleaning and Preprocessing:

During the project, several data cleaning and preprocessing steps were performed to ensure the accuracy and quality of the dataset. These steps included:

1. **Handling Missing Values:** The missing values in the dataset were identified and removed.
2. **Data Formatting:** The data was reviewed for consistent formatting across columns-
    - **Model** column had numerical and textual entries; the format was changed to **Text**.
    - **MSRP** column format was changed from General to Accounting to represent the prices in US Dollars.
3. **Removing Duplicates:** 715 duplicate records were identified and removed from the dataset to avoid any duplication biases during analysis. This helped ensure that each observation represented a unique entity or case.

Assumptions:

Throughout the project, certain assumptions were made based on available information and domain knowledge. These assumptions include:

1. **Data Accuracy:** It was assumed that the provided dataset was accurate and representative of the car industry. However, limitations in data collection or reporting processes could exist, which might impact the analysis.

2. **Data Completeness:** It was assumed that the dataset provided was reasonably complete, and missing values were handled appropriately. However, it is acknowledged that there might be missing data points that could affect the overall analysis and results.

3. **Relevance of Features:** It was assumed that the selected features in the dataset (e.g., horsepower, MPG, price, transmission type) were relevant and appropriate for addressing the research questions and goals of the project. Other potentially important factors not included in the dataset were not considered.

4. **Data Relevance:** The data was last updated in 2017, so it's assumed that it may not reflect current patterns and trends.

These assumptions were made to guide the analysis and provide insights. It is important to acknowledge these assumptions and their potential impact on the results and interpretations.

Tech-Stack Used:

For this project, I utilized both **Microsoft Excel** and **Tableau Desktop** as part of the data analysis and visualization process. Here's an explanation of why I used each tool:

- **Microsoft Excel**: Excel is a powerful spreadsheet software that provides various data manipulation and analysis capabilities.
  - I used Excel for tasks such as data cleaning, data formatting, handling missing values, and performing basic calculations and statistical analysis like regression, correlation analysis.
  - Excel's functionality, including formulas, filters, pivot tables, and conditional formatting, allowed me to efficiently preprocess and explore the data, identify patterns, and perform basic statistical analysis.
- **Tableau Desktop:** Tableau is a data visualization and business intelligence software that enables interactive and visually appealing data representations.
  - I employed Tableau to create charts, graphs, and interactive dashboards to visualize and present the analysed data.
  - Tableau's drag-and-drop interface, robust visualization options, and interactivity features like filters, allowed me to create insightful and engaging visualizations, making it easier to communicate complex information and uncover meaningful insights from the data.

This combination facilitated a comprehensive analysis of the dataset, making it easier to understand patterns, trends, and relationships within the data, and ultimately enabled me to derive valuable insights to address the research questions and goals of the project.

Approach:

Analysis

Here's my approach to the different tasks:

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

- I inserted a pivot chart for the cleaned dataset after pre-processing. In the "Create PivotTable" dialog box, I ensured that the correct data range was selected and chose to place the pivot table on a new worksheet.
- In the pivot table field list, I dragged the "**Market Category**" field to the "Rows" section.
- I then dragged the "**Model**" field to the "Values" section, which displayed the count of car models by default.
- Then I added the "**Popularity**" field to the "Values" section as well, using summary function average.
- Finally, I sorted the pivot table in largest to smallest order, in terms of popularity.

Market Category	Number of car models	Average of Popularity
Hatchback,Flex Fuel	7	5657.0
Flex Fuel,Diesel	16	5657.0
Crossover,Flex Fuel,Performance	6	5657.0
Crossover,Luxury,Performance,Hybrid	2	3916.0
Crossover,Factory Tuner,Luxury,Performance	5	2607.4
Crossover,Performance	69	2586.0
Crossover,Hybrid	42	2563.4
Diesel,Luxury	47	2416.1
Luxury,Performance,Hybrid	11	2333.2
Flex Fuel	855	2225.7
Crossover,Luxury,Diesel	33	2195.8
Hatchback,Factory Tuner,Performance	21	2173.7
Factory Tuner,Luxury,High-Performance	215	2133.4
Hybrid	121	2116.6
Hatchback,Hybrid	64	2111.2
Crossover,Flex Fuel	64	2073.8
Crossover,Hatchback,Factory Tuner,Performance	6	2009.0
Crossover,Hatchback,Performance	6	2009.0

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

- I selected the pivot chart and going to the "Insert" tab, I clicked on "Combo Chart" from "Recommended Charts" option.
- I chose a clustered column chart for the "Avg. Popularity" data and a line chart for the "No. of Models" data.
- I assigned the "No. of Models" which was the line chart, on the secondary axis.
- I customized the chart by adding axis labels, a chart title, and adjusting colours and formatting options as needed.
- To enhance readability, I applied any necessary scaling or formatting to the axes and data labels.

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.

- I selected the columns “**Engine HP**” and “**MSRP**”.
- Going to the "Insert" tab, I clicked on "Scatter" to insert a scatter chart.
- In the scatter chart, the **Engine HP** was automatically assigned to the x-axis and the price to the y-axis.
- I reviewed the initial scatter chart to assess the distribution and relationship between engine power and price.
- To add a trendline, I right clicked on a data point in the scatter chart and selected "Add Trendline."
- In the "Format Trendline" pane that appeared, I chose the linear trendline.
- I customized the trendline by adjusting options like line colour, style, and thickness to make it visually distinguishable.
- I ensured that the trendline adequately represented the relationship between engine power and price, capturing any noticeable patterns or correlations.
- Additionally, I added axis labels, a chart title, and any other necessary formatting to enhance the clarity and presentation of the scatter chart.

Engine H	MSRP
335	$ 46,135
300	$ 40,650
300	$ 36,350
230	$ 29,450
230	$ 34,500
230	$ 31,200
300	$ 44,100
300	$ 39,300
230	$ 36,900
230	$ 37,200
300	$ 39,600
230	$ 31,500
300	$ 44,400
320	$ 48,250
320	$ 43,550

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.

- I selected the columns that included all the numerical variables, including **“Year”**, **“Engine HP”**, **“Engine Cylinders”**, **“Number of Doors”**, **“highway MPG”**, **“city mpg”**, **“Popularity”** and **“MSRP”.**
- I performed regression analysis using Excel's built-in regression analysis tool, the "Data Analysis" add-in.
- The regression analysis provided coefficient values for each predictor variable, indicating their relationship with the car's price.
- I reviewed the coefficient and P-values to identify the variables with the strongest relationships, focusing on those with the highest absolute coefficient values and P-values <0.05.
- The variable **“Year”** had P-value >0.05 so it was not relevant, so I performed the analysis again without the **“Year”** column.
- All the rest of the variables had P-values <0.05 and hence were important for determining the car price.
- To create a bar chart to visualize the relative importance of the variables, I selected the predictor variables and their corresponding coefficient values.
- Going to the "Insert" tab, I chose the clustered column chart.
- I formatted the chart by adding axis labels, a chart title, and adjusting the colours, font sizes, and other visual elements as needed.
- I ensured that the bar chart accurately represented the coefficient values, with longer bars indicating stronger relationships with the car's price.
- Finally, I reviewed the bar chart to assess the relative importance of the variables in influencing the car's price.

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

Manufacturer	Average of MSRP
Plymouth	$ 3,296.87
Oldsmobile	$ 12,843.80
Suzuki	$ 18,021.05
Pontiac	$ 19,800.04
Scion	$ 19,932.50
Mazda	$ 20,106.56
Mitsubishi	$ 21,316.35
FIAT	$ 22,206.02
Subaru	$ 24,240.67
Dodge	$ 24,857.05
Hyundai	$ 24,926.26
Kia	$ 25,318.75
Honda	$ 26,608.88
Chrysler	$ 26,722.96
Saab	$ 27,879.81

- I inserted a pivot chart for the cleaned dataset after pre-processing. In the "Create PivotTable" dialog box, I ensured that the correct data range was selected and chose to place the pivot table on a new worksheet.
- In the pivot table field list, I dragged the "**Make**" field to the "Rows" section.
- I then dragged the "**MSRP**" field to the "Values" section and changed the aggregation function to "Average" for the Price values.
- I sorted the pivot table to arrange the data in smallest to largest order of price.

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

- I selected the pivot table and going to the "Insert" tab, I chose the horizontal bar chart.
- I ensured that the manufacturers were represented on the category axis and the average prices on the value axis.
- I customized the chart by adding axis labels, a chart title, and adjusting colours and formatting options as needed.
- I reviewed the chart to ensure it effectively visualized the relationship between manufacturer and average price.

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.

Engine Cylinders	Fuel Efficiency
0	103.00
3	38.67
4	31.50
5	26.07
6	24.00
8	20.18
10	20.00
12	17.74
16	14.00

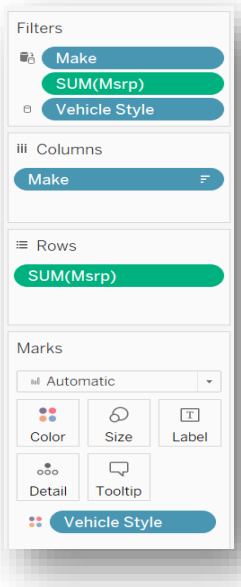
- I created a pivot table with **“Engine Cylinders”** in ‘Rows’ and average of **“highway MPG”** in ‘Values’.
- Going to the "Insert" tab, I clicked on "Scatter" to insert a scatter plot.
- In the scatter plot, the number of cylinders was automatically assigned to the x-axis and highway MPG to the y-axis.
- To add a trendline, I right clicked on any data point in the scatter plot and selected "Add Trendline" and added a linear trendline.
- I customized the trendline by adjusting options like line colour, style, and thickness to make it visually distinguishable.
- Finally, I reviewed the scatter plot and trendline to visually assess the significance of the relationship.

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

- To find the correlation between the **“Engine Cylinders”** and **“highway MPG”** variables, I used the “Data Analysis” add-in.
- I used conditional formatting to assign colour scales to the correlation table and get a heat map of the data.

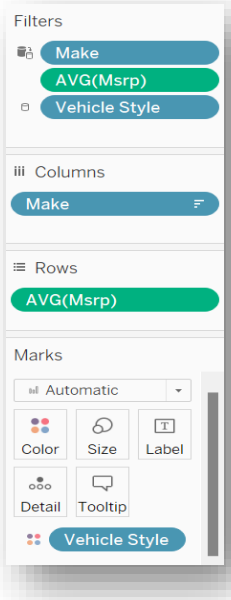
Building the dashboard

I made the dashboard on Tableau Desktop using following approach:



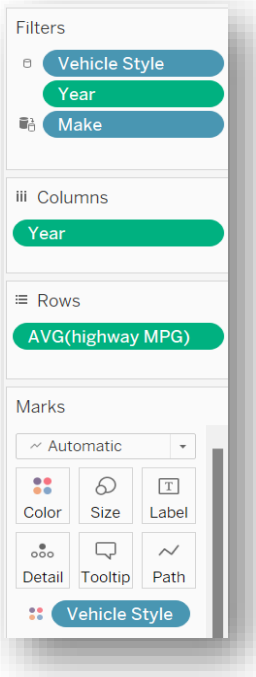
Task 1: How does the distribution of car prices vary by brand and body style?

- I imported the car dataset into Tableau by connecting to the Excel file.
- I dragged the **"Make"** field to the Columns shelf and the **"MSRP"** field to the Rows and selected the Measures to Sum.
- Then, I dragged the **"Vehicle Style"** field to the 'Color' shelf to represent the body styles of the cars.
- I customized the chart by adding labels, adjusting colours, and formatting options to enhance its visual appearance and clarity.
- To make the chart interactive, I added filters for **"Vehicle Style"** and **"MSRP"** allowing users to select specific body styles and dynamically update the chart.



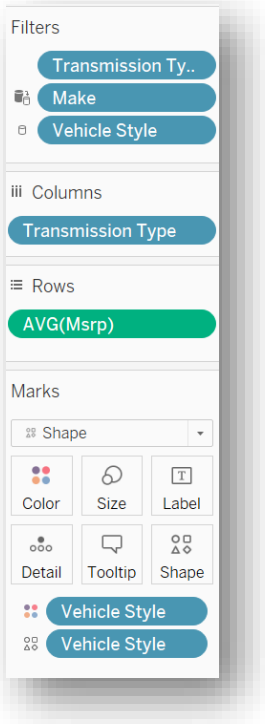
Task 2: Which car brands have the highest and lowest average MSRPs, and how does this vary by body style?

- I created a stacked column chart just for this just like in Task 1 except the **"MSRP"** measure was selected to be Average instead of Sum.
- I customized the chart and added filter for **"Vehicle Style"** and **"MSRP".**



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

- To create a scatter plot chart, I dragged the **"MSRP"** field to the Rows shelf, with Measure selected as Average and the **"Transmission Type"** field to the Columns shelf.
- I further customized the scatter plot by dragging the **"Vehicle Style"** field to the Color and Shape shelf to differentiate symbols for each body style.
- I formatted the scatter plot chart by adding axis labels, a chart title, and adjusting colours and formatting options to enhance clarity and visual appeal.
- To make the chart interactive, I utilized Tableau's filtering capabilities, and I added filters for body styles, allowing users to dynamically explore the data and observe the effects on MSRP.

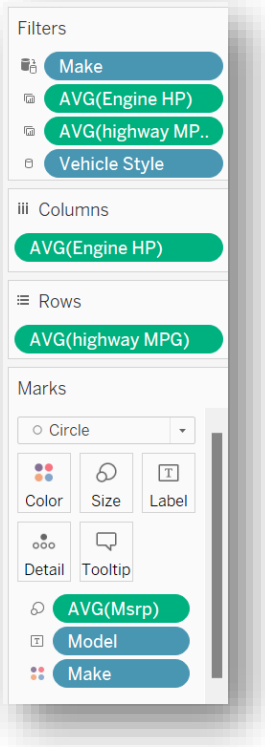


Task 4: How does the fuel efficiency of cars vary across different body styles and model years?

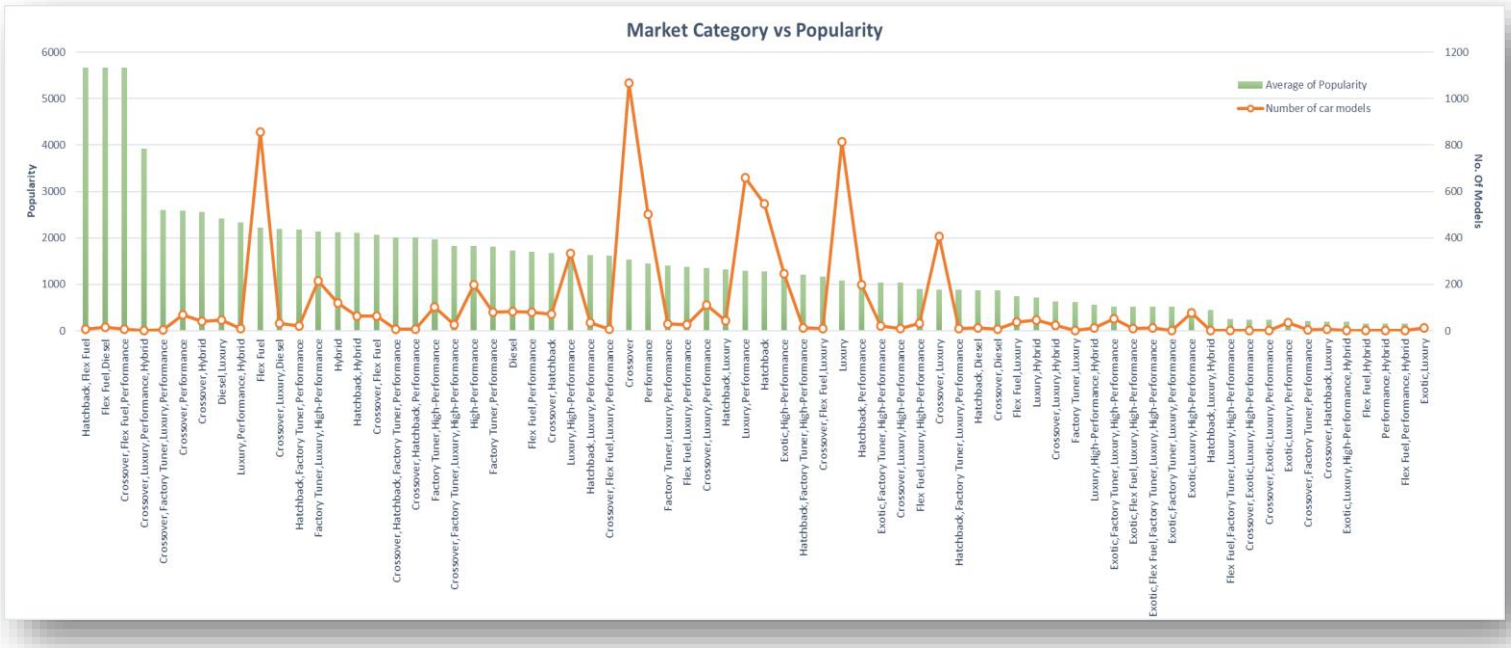
- To create a line chart, I dragged the **" Year"** field to the Columns shelf and the **"highway MPG"** field to the Rows shelf to show Average.
- I further customized the line chart by dragging the **"Vehicle Style"** field to the Color or Label shelf to differentiate the lines for each body style.
- To make the chart interactive, I added filters for specific body styles and years.

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

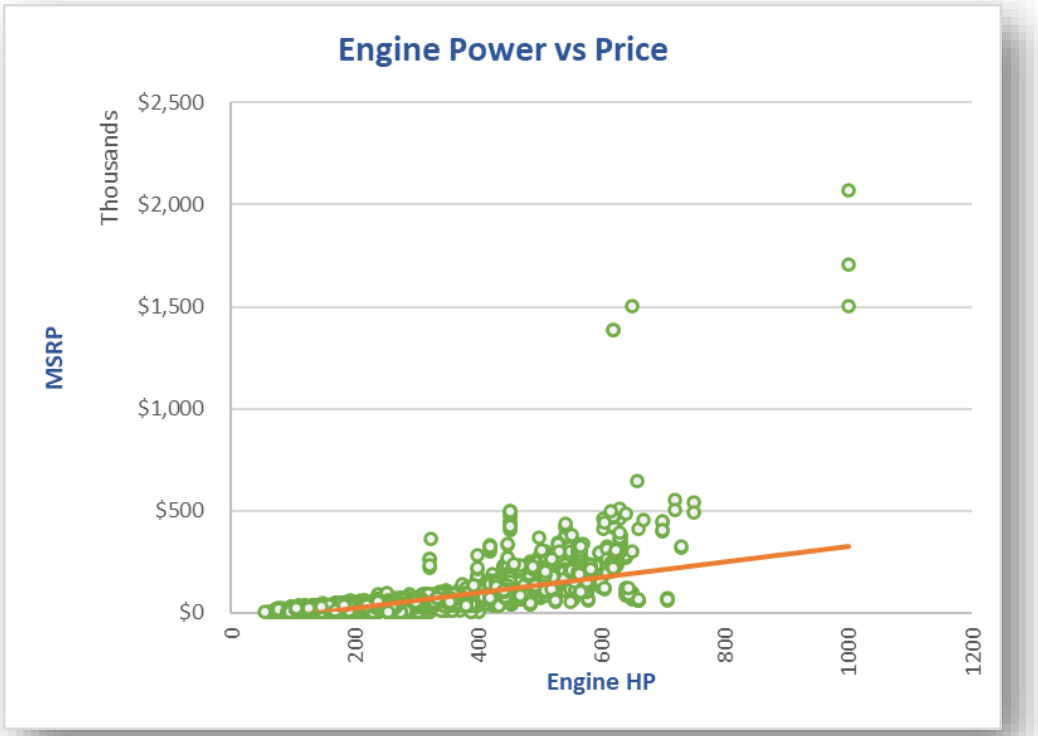
- To create a bubble chart, I dragged the **"Engine HP"** field to the Columns shelf, the **"highway MPG"** field to the Rows shelf, and the **"MSRP"** field to the Size shelf. I selected the measure Average for all these fields.
- I further customized the bubble chart by dragging the **"Make"** field to the Color shelf to assign different colours to each brand bubble.
- Each bubble represented a car model, with the size indicating the average price and the placement indicating the horsepower and MPG.
- To make the chart more informative, I labelled the bubbles with the car model name by dragging the **“Model”** field to the Label shelf, for better identification and analysis.



Insights:



- **'Hatchback**, **Flex Fuel', 'Flex Fuel**, **Diesel'**, **'Crossover, Flex Fuel, Performance**', seems to be the most popular market category with average popularity of 5657, but has very low number of cars.
- **'Exotic**, **Luxury'** seems to be the least popular market category of cars with average popularity of 112.7.
- Maximum car models are of the **'Crossover'** category (1068) followed by **'Flex Fuel'** (855) and **'Luxury'** (815), although they don't have high popularity.



- The relationship between a car's engine horsepower and its price is sort of linear, meaning as the engine power increases, the price also increases.

- We can see a couple of high-priced cars for engine horsepower ~650, almost comparable to the prices those of 1000HP engine. This could be because of the brand or model of those cars.

Regression Statistics	
Multiple R	0.683695001
R Square	0.467438855
Adjusted R Square	0.467095836
Standard Error	45339.44159
Observations	10876

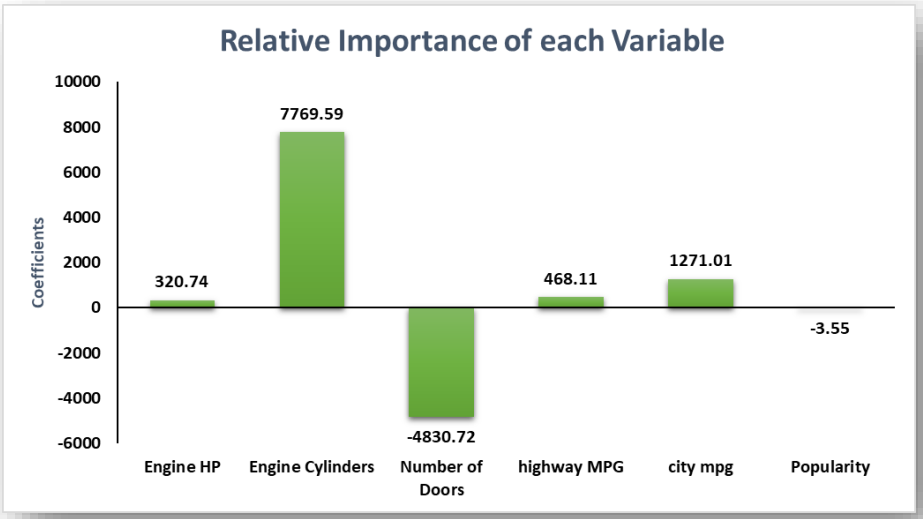
	Coefficients	Standard Error	t Stat	P-value
Intercept	-180578.4755	167107.1982	-1.080614584	0.279892571
Year	41.4588481	83.85885655	0.494388426	0.621041879
Engine HP	318.3130351	8.147429257	39.06913765	0
Engine Cylinders	7855.106106	498.7677809	15.74902471	2.84334E-55
Number of Doors	-4916.030854	533.3694306	-9.216934026	3.62013E-20
highway MPG	459.4906806	111.6614473	4.115034255	3.89979E-05
city mpg	1266.141773	126.9638752	9.972456898	2.53385E-23
Popularity	-3.566266022	0.30370743	-11.74243916	1.19906E-31

- The R square or the 'Coefficient of Determination' shows that 46.7% of the variance in the dependent variable (MSRP) can be explained by the independent variables.
- To understand which independent variables, contribute towards the dependent variable, we check for the P-value. If the individual P-values are >0.05 (our confidence interval) then they are not relevant. Here **Years** is irrelevant.

Regression Statistics	
Multiple R	0.683686242
R Square	0.467426877
Adjusted R Square	0.467132882
Standard Error	45337.86563
Observations	10876

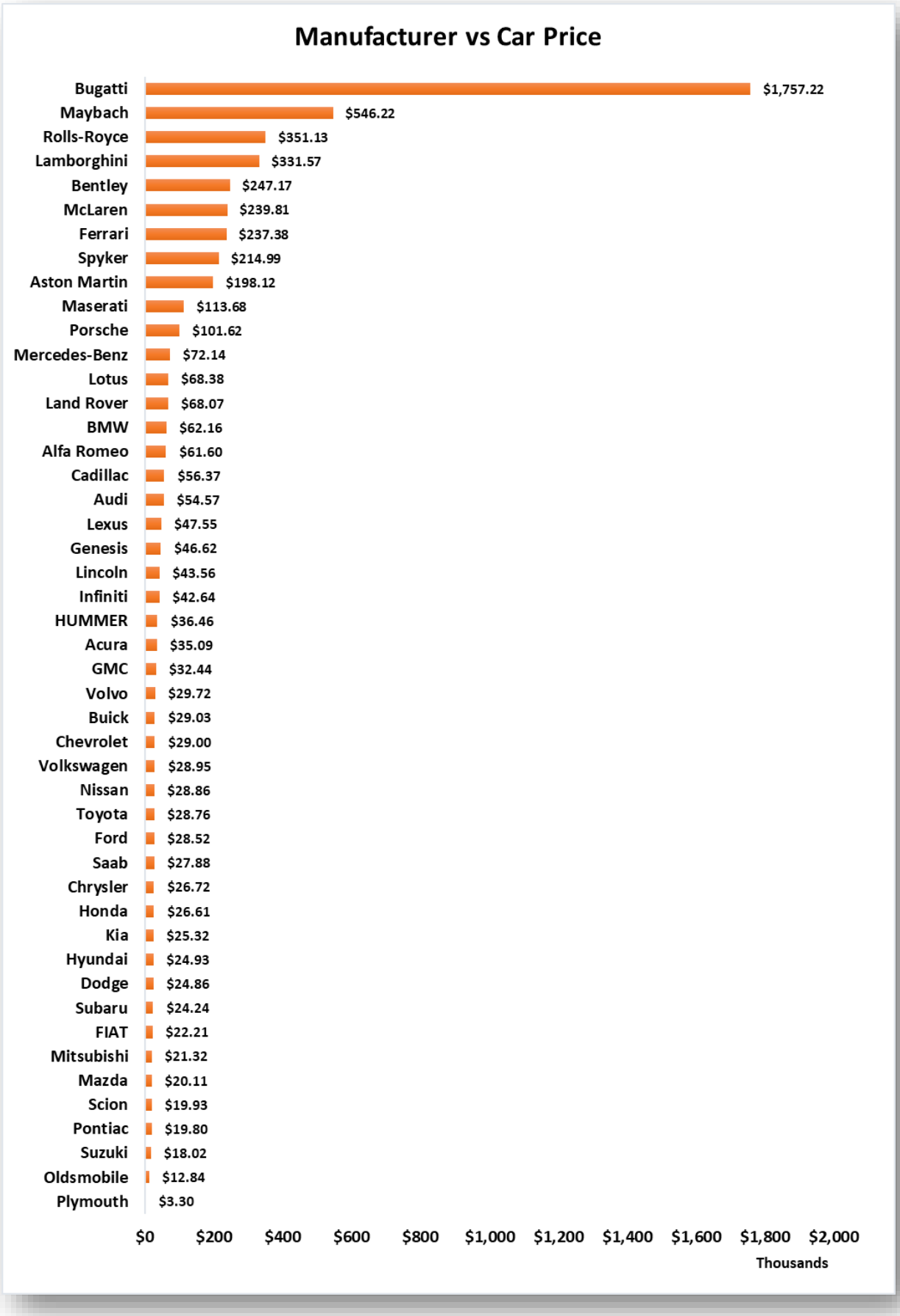
	Coefficients	Standard Error	t Stat	P-value
Intercept	-97985.72612	3952.59783	-24.79020896	5.0859E-132
Engine HP	320.7391231	6.503579949	49.31731841	0
Engine Cylinders	7769.592712	467.7992571	16.60881798	3.40646E-61
Number of Doors	-4830.720147	504.667177	-9.572091008	1.27474E-21
highway MPG	468.1064255	110.2893022	4.244350234	2.21048E-05
city mpg	1271.008424	126.5772955	10.04136183	1.27099E-23
Popularity	-3.548166438	0.301482319	-11.76906974	8.78234E-32

- Adjusted R square value increased once we removed the Years variable from analysis because adjusted R square value depends on the number of independent variables. As the number of independent variable increases, adjusted R square value decreases.
- The P-values for all the rest of the columns are <0.05 meaning our dependent variable (**MSRP**) have strong relationships with these variables.

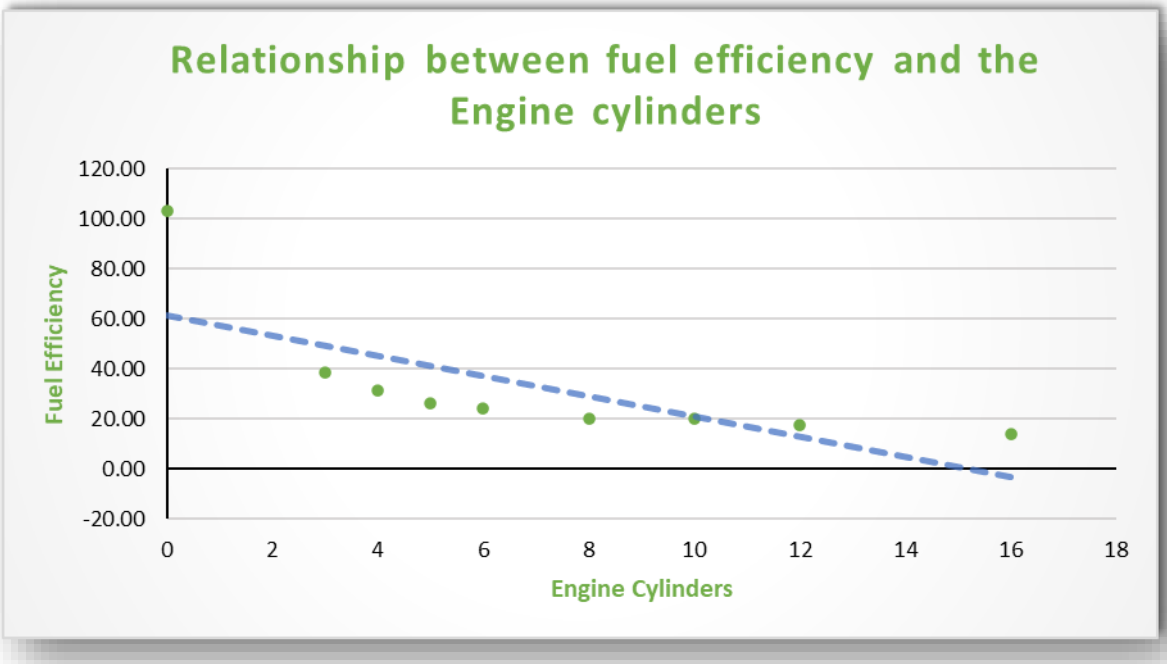


- **Engine Cylinders** seem to have a strong positive relationship with **MSRP** with coefficient of 7770, suggesting that as number of Engine Cylinders increase, we can observe an increase in car price.

- We can see a strong negative relationship between the **Number of Doors** in the car and its price (-4830) suggesting that with increase in number of car doors, price may decrease significantly.
- There is least relationship between popularity and the car price (-3.55).



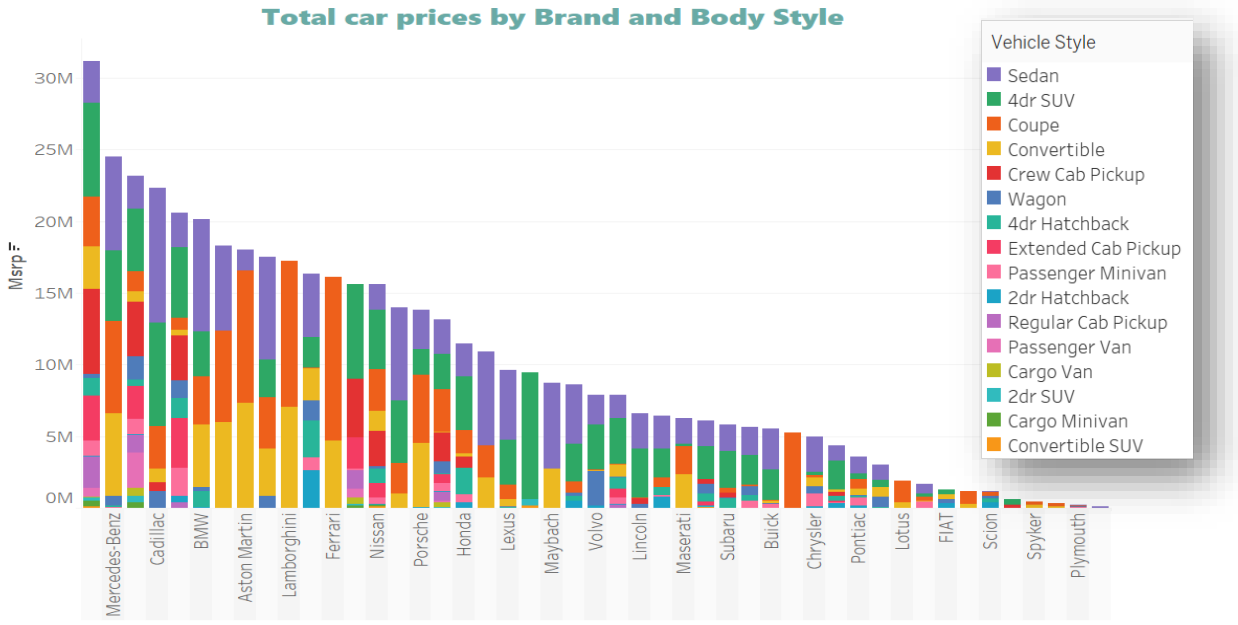
- The highest average of car price is recorded for '**Bugatti'** of around 1.7 million dollars.
- **'Plymouth**' records the lowest average price of cars.



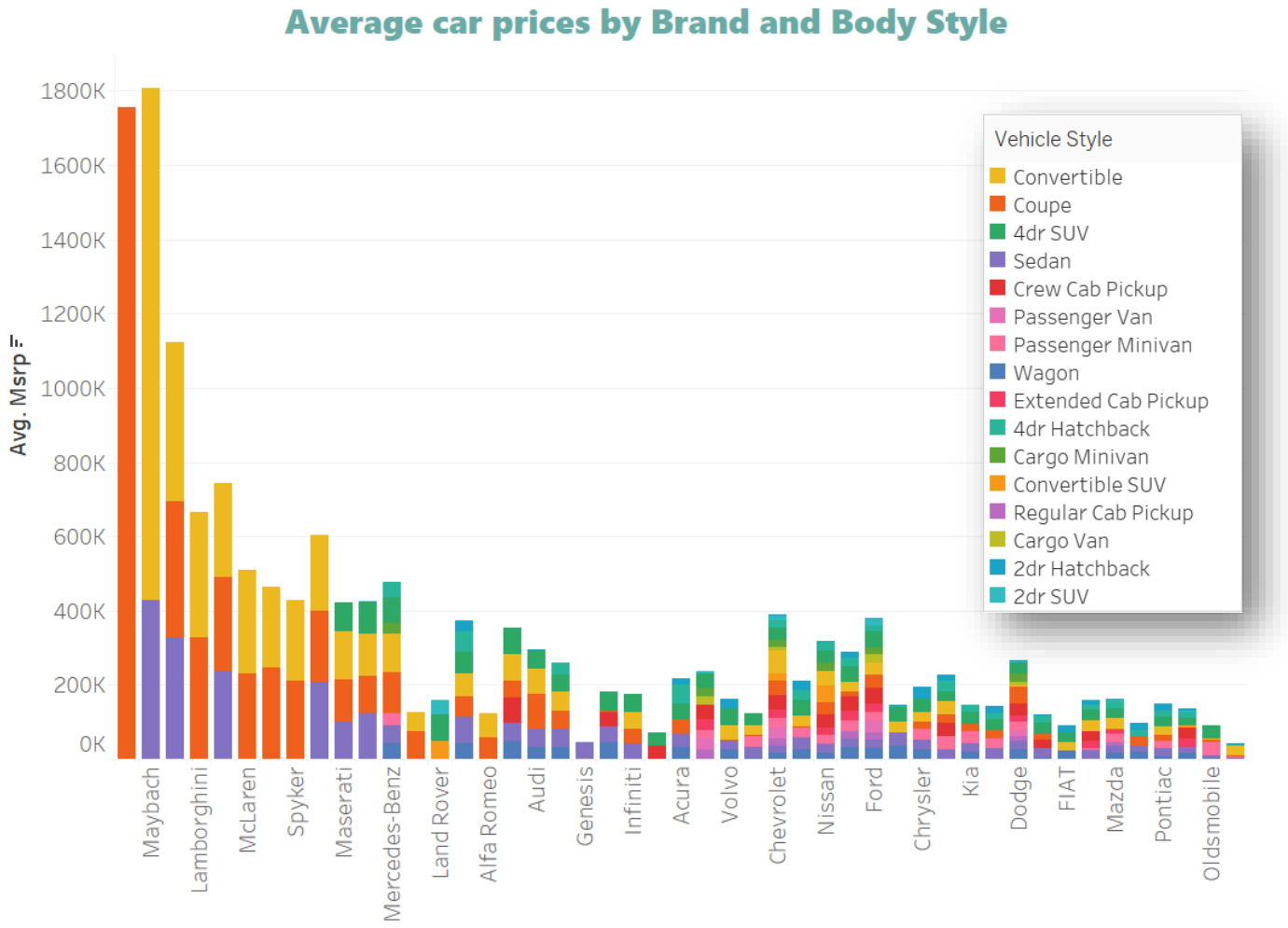
- The maximum fuel efficiency is observed for cars with 0 engine cylinders.
- The trendline shows a negative slope suggesting that the fuel efficiency decreases with increase in number of cylinders.

	Engine Cylinders	*Fuel Efficiency*
Engine Cylinders	1	-0.725773653
Fuel Efficiency	-0.725773653	1

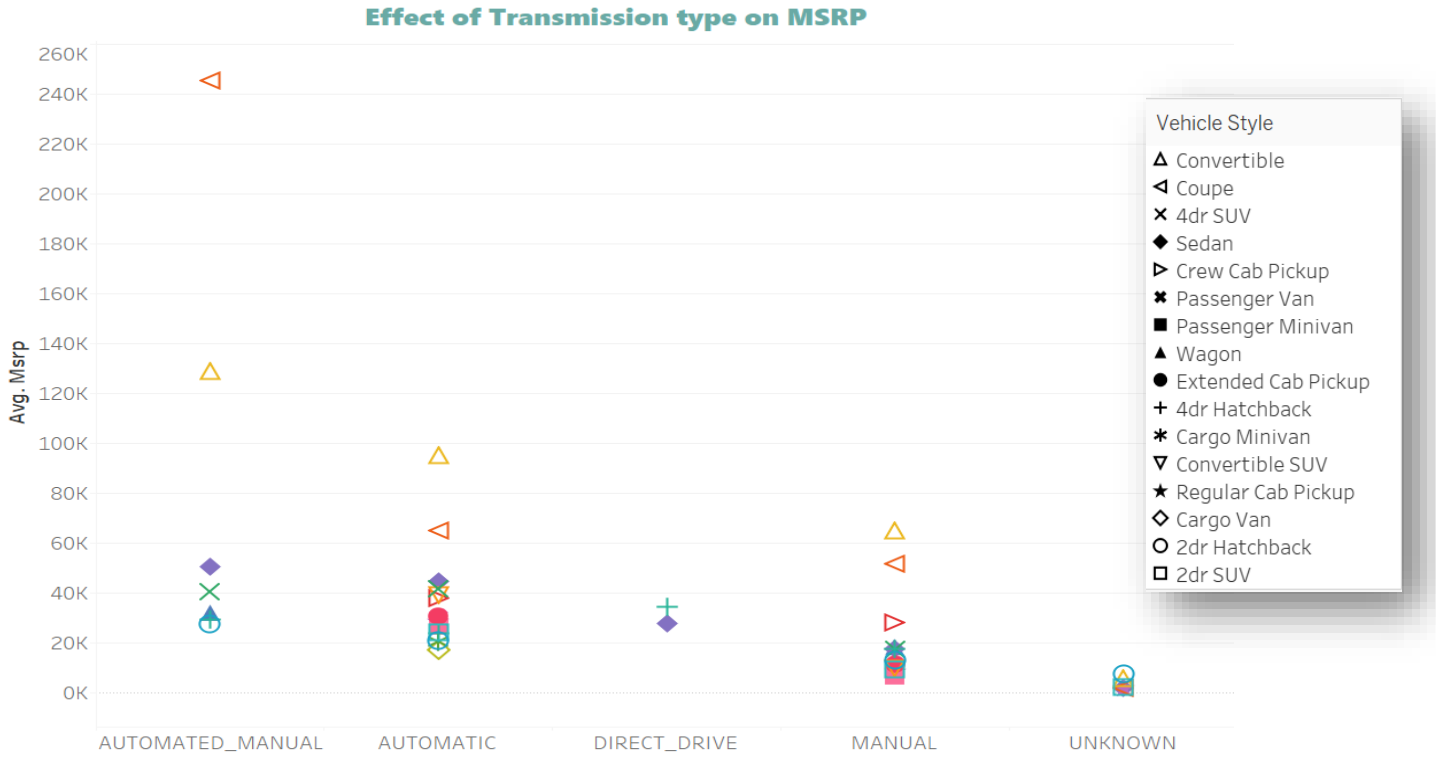
- There is a **strong negative correlation of -0.726** between number of engine cylinders and the average fuel efficiency suggesting that as number of engine cylinders increase, the fuel efficiency decreases steadily.



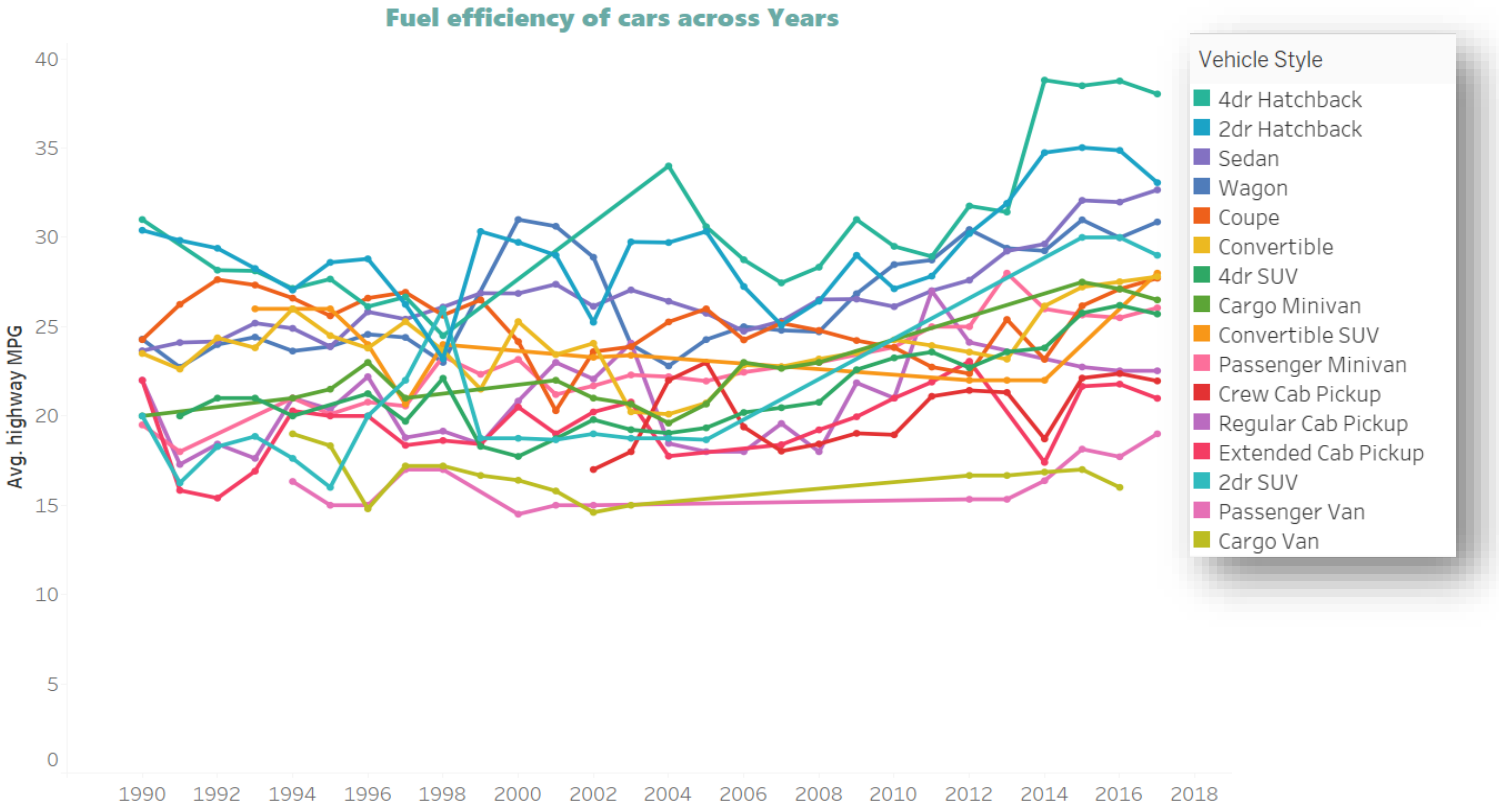
- **'Chevrolet'** has the maximum overall net MSRP (~$31M), followed by **'Mercedes-Benz'** (~$25M).
- **'Genesis-Sedan'** has the lowest net MSRP (~$140K).



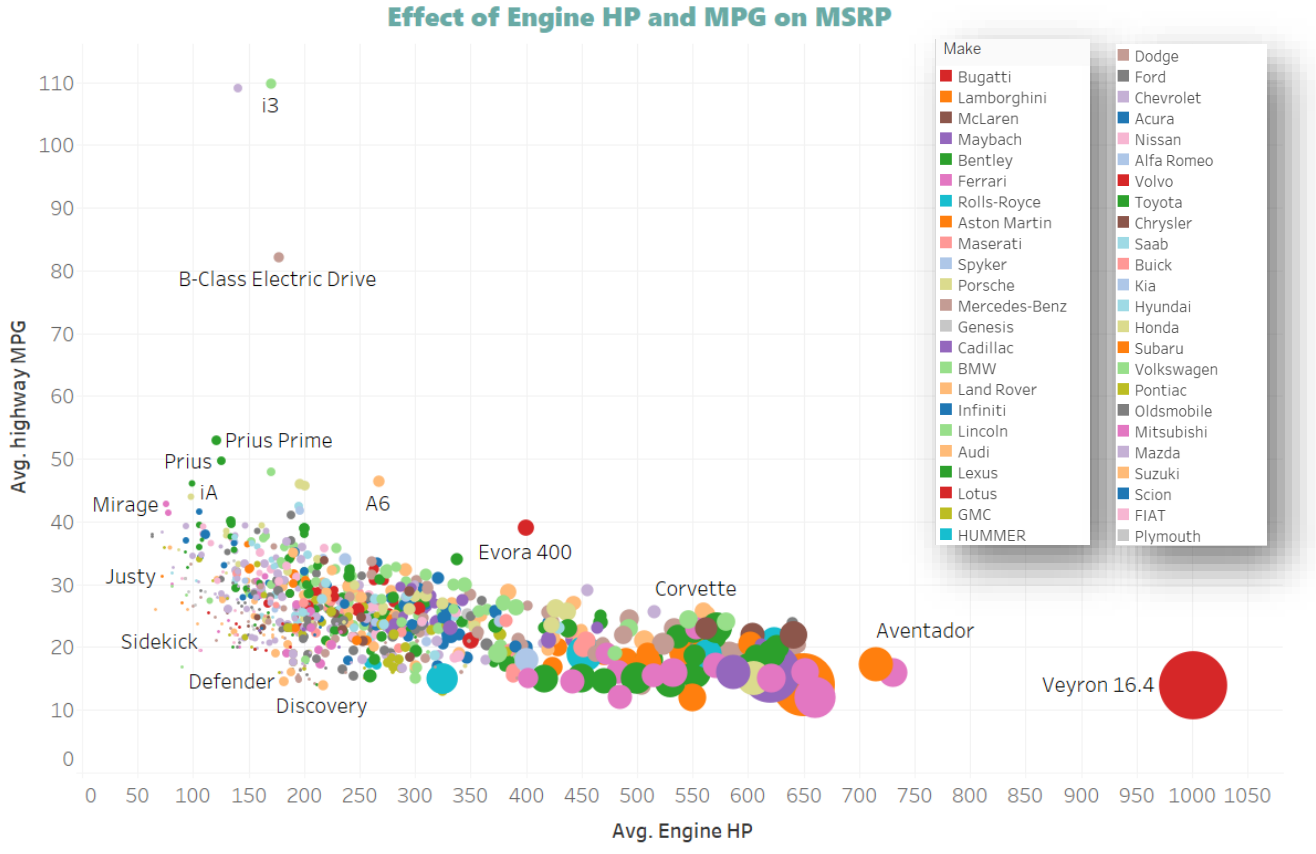
- **'Bugatti'** has the maximum average MSRP ($1.7M) for single body style **'Coupe'** whereas **'Plymouth'** has the lowest average MSRP (~$3.3K). Out of its 7 body styles as well as overall, '**Convertible'** is recorded to have the highest MSRP ($88K) compared to other body styles.



- **‘Automated Manual’** transmission type cars have the overall highest recorded average price for vehicle style **‘Coupe’** ($250K), followed by **‘Convertible’** ($129K) whereas **‘Automatic’** type has the highest average MSRP for **‘Convertible’** ($95K), same for **‘Manual’** with average MSRP of $64K. For **‘Direct’** transmission type **‘4dr Hatchback’** has the highest average MSRP ($27K).



- The maximum fuel efficiency was observed for 2014 **‘4dr Hatchback’** (38.81 MPG) whereas the lowest was recorded for 2000 **‘Passenger Van’** (14.5 MPG).
- **‘Sedan’** has been most consistent in terms of fuel efficiency throughout the years compared to other vehicle styles.



- **'Bugatti- Veyron 16.4'** has the maximum average power (1001HP) and maximum price ($1.7M), however its average fuel efficiency was one of the lowest recorded (14 MPG)
- **'BMW-i3'** has the maximum fuel efficiency (109.75 MPG) but lower power (170HP) and the average price for the same was also low ($25K).
- Average price is high for car models with high engine power whereas its low for cars with high fuel efficiency.

Results:

After the analysis, visualization of the results and drawing insights from them, we can make following conclusions:

1. Car prices increase with increase in Engine HP and number of engine cylinders, so to ensure profit, car models with more engine cylinders and power should be manufactured.
2. Since car prices are low for cars with more doors, less door models can be more profitable with high engine power.
3. Bugatti seems to be the most expensive car brand which could be because it offers high engine power, automated manual transmission type, lowest number of doors (2), maximum engine cylinders (16) and one of the lowest fuel efficiencies among the recorded brands.

So, product development decisions to maximize profitability while meeting consumer demands includes:

- Higher number of engine cylinders (>8).
- Higher Engine HP (>400 HP).
- Less number of car doors.
- Low fuel efficiency
- Automated manual transmission type models.

Some possible future directions for this project include:

- Building predictive models to forecast car prices based on various features such as horsepower, MPG, brand, and body style. This could help car manufacturers optimize pricing strategies and inform product development decisions.
- With the increasing focus on sustainability and environmental concerns, analysis of the impact of different car features, such as fuel efficiency and emissions, on consumer demand and market competitiveness can help Identify opportunities for incorporating sustainable technologies and practices into the product offerings.
- Analysis of how car features, pricing, and brand perception differ across different geographical locations can provide valuable insights for targeted marketing campaigns and regional expansion strategies.

Dashboard

Dashboard link-

https://public.tableau.com/authoring/ImpactofCarFeaturesonPriceandProfitability/CarPricesAnalysis#1

Please note that the dashboard may exceed the default window size, so it might be necessary to zoom out to view the complete dashboard.