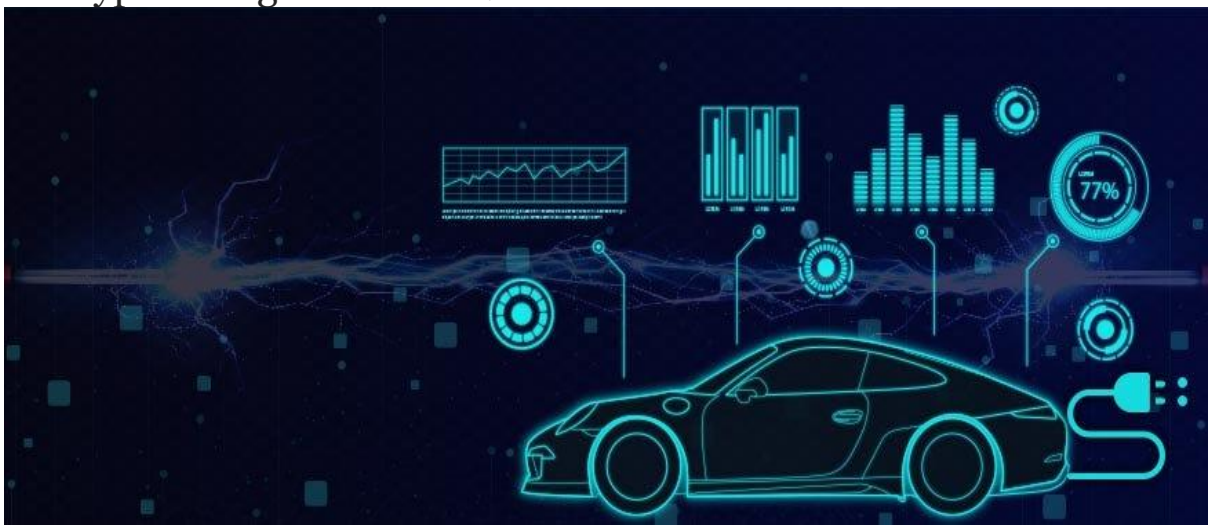
Analyzing the Impact of Car Features on Price and Profitability

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.



This Project is about analyzing patterns in the various Car features to understand the factors that drive consumer demands for Cars. These insights will be helpful for Car manufacturers, dealers and

other stakeholders to determine the important factors to boost sales of Cars.

Contents

Tech Stack Used

1. **Python 3.9.6** — Programming language used for Data Pre-processing.
2. **Jupyter Notebook 6.5.2 —** Interactive platform to write and execute codes in various programming languages (in this case Python).
3. **Microsoft Excel 2021** — A spreadsheet editor software used mainly by professionals to enter data in table format, perform computations, plot graphs etc.

4. **Tableau Public 2021.2** — A visualization tool to represent data in graphs and plots. Mainly used to create Dashboard

Credit: Google Images

Dataset Overview

Source of Data: https://drive.google.com/drive/folders/1lSDUg0iy71tJ5rzs1QWb9UeDq_7IjAYw?usp=sharing

The dataset provides details about the current loan applications like the type of contract, annuity amount, credit amount etc.

- The Dataset details are:
  - Number of Data-Points: **11,914**
  - Number of Features: **16**
  - Column Details:

  1. Make: Manufacturer of the Car
  2. Model: Model name of the Car
  3. Year: Year of launch of the Car
  4. Engine Fuel Type: Type of Fuel that the Car uses
  5. Engine HP: Horsepower of the Car
  6. Engine Cylinders: Number of cylinders in the Car's engine
  7. Transmission Type: Transmission type of the Car
  8. Driven_Wheels: Which wheels does the engine

transfer power to
9. **Number of Doors**: Number of doors in the Car
10. **Market Category**: Market categories the Car can be classified into
11. **Vehicle Size**: Size category of the Car
12. **Vehicle Style**: Style category of the Car
13. **highway MPG**: Mileage of the Car in highways
14. city **mpg**: Mileage of the Car in cities
15. **Popularity**: Popularity score of the Car
16. **MSRP**: Price of the Car

Data Pre-Processing

Handling Duplicate Values

- Found duplicate rows on analysis. Except the first instance, dropped all other duplicate rows.

Handling Null Values

- 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 **Engine Cylinders** column where **Engine Fuel Type** is **electric,** we replaced them with **0** as on analysis we found that **electric** Cars have **0 Engine Cylinders** which is logical.

- For Null values in **Engine Cylinders** column where **Engine Fuel Type** is other than **electric,** we found that there were only two such Car models. So searched for the information online and replaced them with the correct values.

- For Null values in **Number of Doors** column, we found that there were only two such Car models. So searched for the information online and replaced them with the correct values.

- For Null values in **Engine Fuel Type** column, we found that there were only one such Car model. So searched for the information online and replaced them with the 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 dataframe.
  — 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 performed **Multinomial Logistic Regression** to get

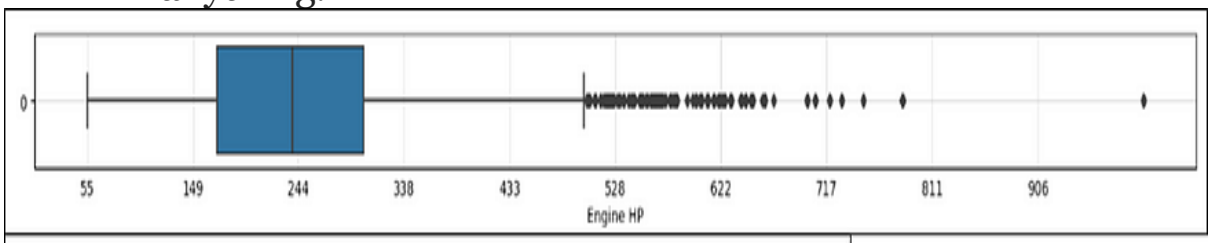
the **top 4 Market Categories** for the rows with null values of **Market Category**.

Handling Errors

- Column **Transmission Type** had some rows with column value '**UNKNOWN**', we searched for the information online and replaced the value '**UNKNOWN**' with correct ones.

Handling Outliers

- For the Outliers in **Engine HP** column, we checked **Market Category** column for all rows with Null values of **Engine HP**. All cars are either **Exotic** or **High-Performance** or **Luxury** vehicles. So didn't change anything.



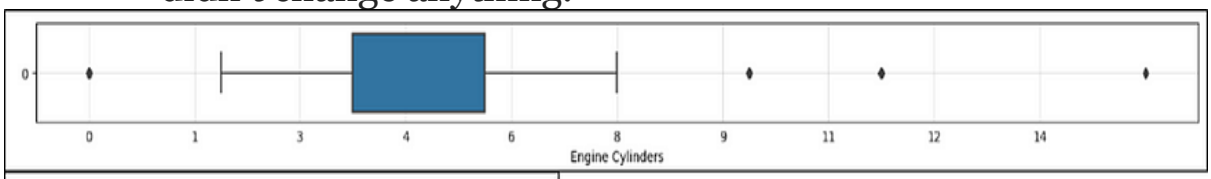

```
df[df['Engine HP']>520]['Market Category'].value_counts()
```

Market Category	
Exotic,High-Performance	117
Factory Tuner,Luxury,High-Performance	73
Factory Tuner,High-Performance	41
Luxury,High-Performance	33
Exotic,Factory Tuner,Luxury,High-Performance	29
Exotic,Luxury,High-Performance	21
Crossover,Factory Tuner,Luxury,High-Performance	16
Exotic,Luxury,Performance	13
Exotic,Flex Fuel,Factory Tuner,Luxury,High-Performance	13
Exotic,Factory Tuner,High-Performance	11
Exotic,Flex Fuel,Luxury,High-Performance	11
Exotic,Performance	10
Exotic,Luxury	6
High-Performance	6
Crossover,Luxury,High-Performance	2
Factory Tuner,Luxury	2
Exotic,Luxury,High-Performance,Hybrid	1
Flex Fuel,Factory Tuner,Luxury,High-Performance	1

Name: count, dtype: int64

Dealing with Outliers in **Engine HP** column

- For the Outliers in **Engine Cylinders** column, we checked **Engine Fuel Type** for rows with **0 Engine Cylinders** which are all are **electric** which is logical. For **Engine Cylinders** greater than **8**, we checked the **Market Category** and all cars were either **Exotic** or **High-Performance** or **Luxury**. So didn't change anything.



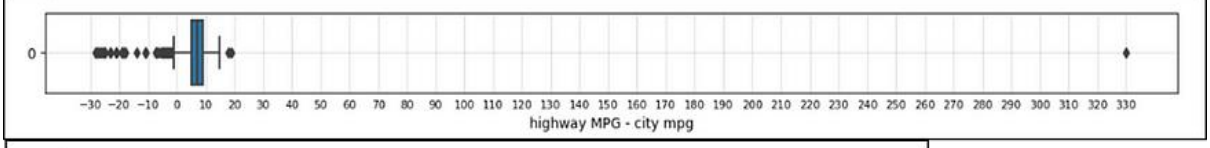

```
1 df[df['Engine Cylinders']<1]['Engine Fuel Type'].unique()

array(['electric'], dtype=object)

1 df[df['Engine Cylinders']>8]['Market Category'].value_counts()

Market Category
Exotic,High-Performance                                141
Exotic,Luxury,Performance                               25
Luxury,High-Performance                                 25
Exotic,Luxury,High-Performance                          19
Exotic,Factory Tuner,High-Performance                   19
Factory Tuner,Luxury,High-Performance                   13
Exotic,Flex Fuel,Factory Tuner,Luxury,High-Performance  13
Exotic,Flex Fuel,Luxury,High-Performance                11
Exotic,Luxury                                           10
Luxury,Performance                                      10
Exotic,Factory Tuner,Luxury,High-Performance             5
Luxury                                                   3
Factory Tuner,Luxury,Performance                         1
Crossover,Luxury,Diesel                                  1
Name: count, dtype: int64
```

Dealing with Outliers in **Engine Cylinders** column

- For the Outliers in **highway MPG** and **city mpg** columns, we plotted a box plot of difference values between the two columns. We observed that there are few negative value of differences. Generally Mileage in Highway is more than that of City. So we replaced the **highway MPG** of such rows with the corresponding **city mpg**. Also, there were few rows where **highway MPG** is a lot more than **city mpg**. On further investigation, we found that these are not correct values. So we searched online for the

correct Highway and City mileages and replaced them with the correct values.

```
figure,axs = plt.subplots(1,1, figsize=(15,1))
sns.boxplot(df['highway MPG']-df['city mpg'], orient='h', ax=axs)
axs.set_xticks(np.arange(-30, 331, 10))
axs.tick_params(axis='x', pad=5, length=0, labelsize=8)
axs.set_xlabel('highway MPG - city mpg')
axs.grid(axis='y', alpha=0.4)
axs.grid(axis='x', alpha=0.4)
axs.set_axisbelow(True)
plt.show()
```




```
df['highway MPG'].loc[df[d<0].index]=df['city mpg'].loc[df[d<0].index]
```

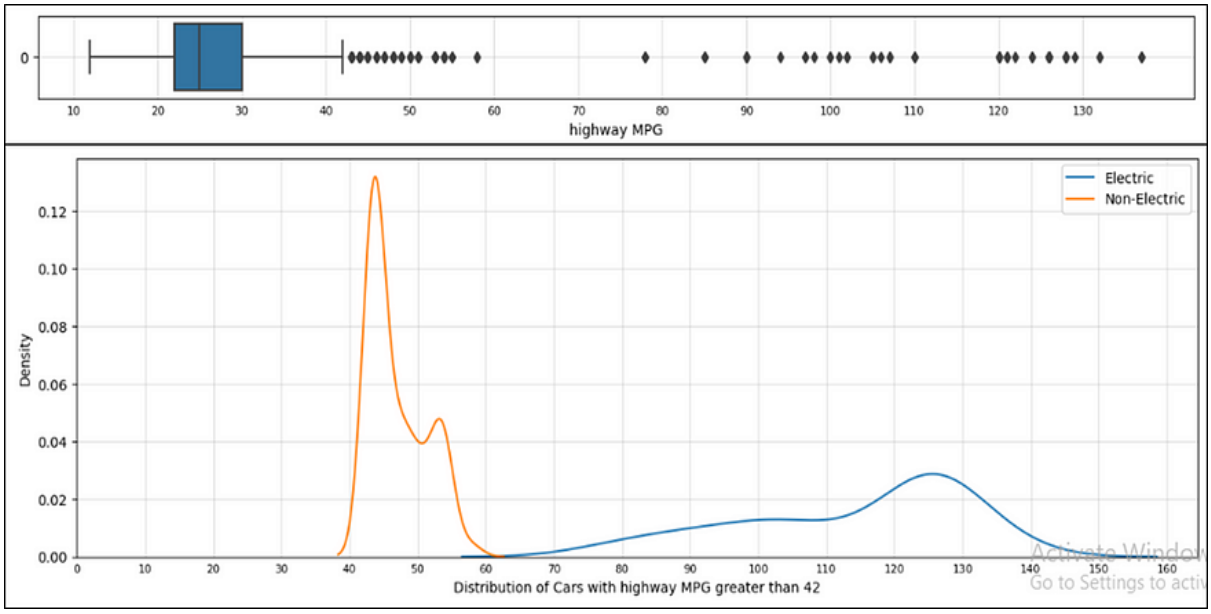
```
d=df['highway MPG']-df['city mpg']
df[d>15]
```

	Make	Model	Year	Engine Fuel Type	Engine HP	Engine Cylinders	Transmission Type	Driven_Wheels	Number of Doors	Market Category	Vehicle Size	Vehicle Style	highway MPG	city mpg
1119	Audi	A6	2017	premium unleaded (recommended)	252	4	AUTOMATED_MANUAL	front wheel drive	4	Luxury	Midsize	Sedan	354	24
3157	Chevrolet	Cruze	2015	diesel	148	4	AUTOMATIC	front wheel drive	4	Diesel	Midsize	Sedan	46	27
4257	Lotus	Evora 400	2017	premium unleaded (required)	400	6	MANUAL	rear wheel drive	2	Exotic,High-Performance	Compact	Coupe	39	21

```
#https://www.cars.com/research/audi-a6-2017/
df['highway MPG'].loc[1119]=35

#https://media.chevrolet.com/media/us/en/chevrolet/vehicles/cruze/2015.html
df['highway MPG'].loc[3157]=40

#https://www.fueleconomy.gov/feg/bymodel/2017_Lotus_Evora.shtml
df['highway MPG'].loc[4257]=24
df['city mpg'].loc[4257]=16
```

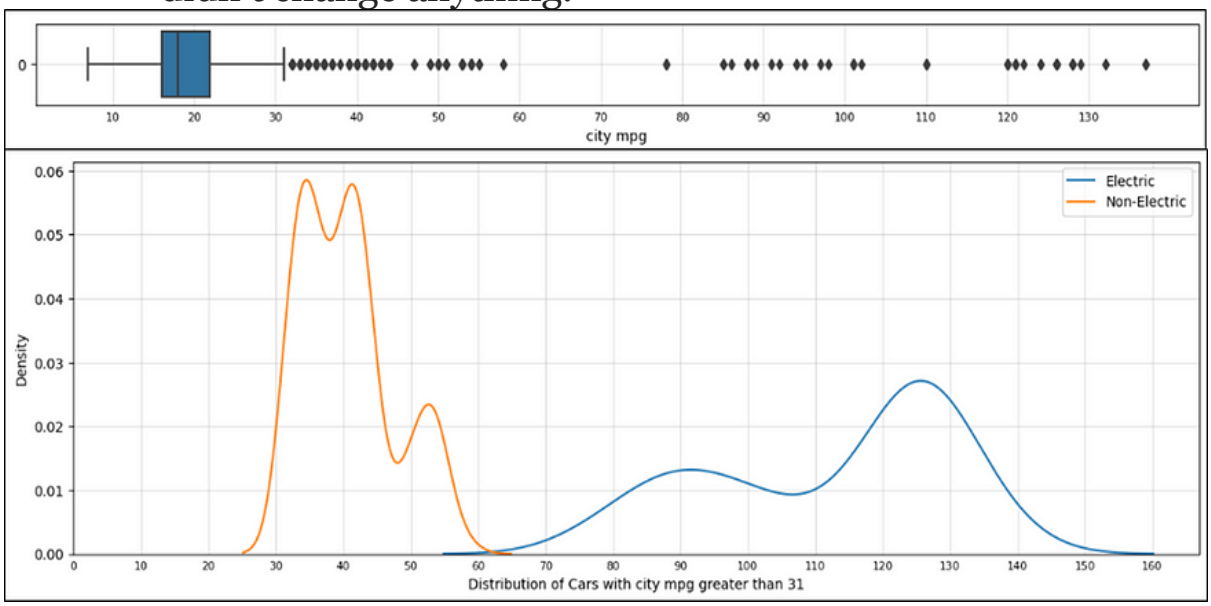

Dealing with Outliers in **highway MPG** and **city mpg** column

- For the Outliers in **highway MPG** column, we plotted box plot and as well as pdf. Considering **42** as the threshold, we observed that a large percentage of vehicles with very high mileage are **electric** vehicles which is very logical. So didn't change anything.



Dealing with Outliers in **highway MPG** column

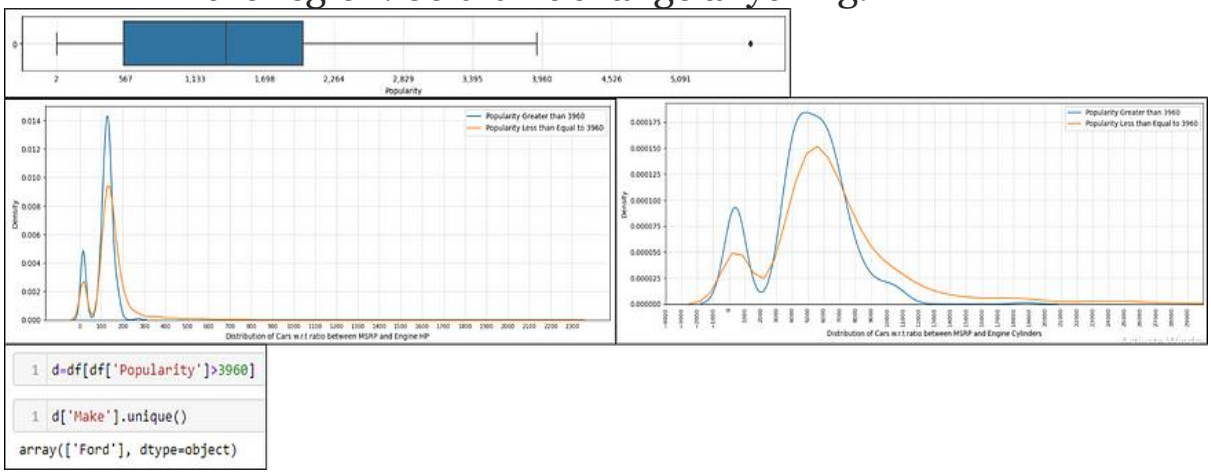
- For the Outliers in **city mpg** column, we plotted box plot and as well as pdf. Considering **31** as the threshold, we observed that a large percentage of vehicles with very high mileage are **electric** vehicles which is very logical. So didn't change anything.



Dealing with Outliers in **city mpg** column

- — For the Outliers in **popularity** column, we plotted box plot and as well as pdf. Considering **3960** as the threshold,

we observed that the distribution of ratio of **MSRP** (Car Price) and **Engine HP** for cars whose popularity was above and below **3960** is almost the same.
— Also with **3960** as threshold value, we observed that the distribution of **MSRP** (Car Price) and **Engine Cylinders** for cars whose popularity was above and below **3960** was almost the same.
— Also the cars whose popularity was above **3960** are all from **Ford** which implies that **Ford** cars are very popular in the region. So didn't change anything.



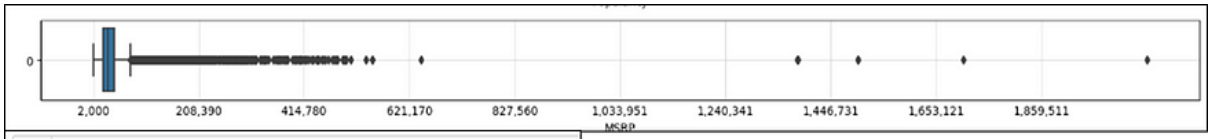

```
d=df[df['Popularity']>3960]
```

```
d['Make'].unique()
```

```
array(['Ford'], dtype=object)
```

Dealing with Outliers in **Popularity** column

- For the Outliers in **MSRP** column, we plotted the box plot. Considering **100000** as the threshold, we observed that the cars with price above **100000** are all **Exotic** or **Performance** or **Luxury** cars. So didn't change anything.




```
1 df[df['MSRP']>100000]['Market Category'].value_counts()
```

```
Market Category
Exotic,High-Performance                                    213
Luxury,High-Performance                                    105
Factory Tuner,Luxury,High-Performance                       69
Exotic,Luxury,High-Performance                              52
Exotic,Factory Tuner,Luxury,High-Performance                48
Exotic,Luxury,Performance                                   36
Exotic,Factory Tuner,High-Performance                       21
Exotic,Flex Fuel,Factory Tuner,Luxury,High-Performance      13
Flex Fuel,Luxury,High-Performance                           13
Exotic,Luxury                                               12
Crossover,Factory Tuner,Luxury,High-Performance             12
Exotic,Flex Fuel,Luxury,High-Performance                    11
High-Performance                                             6
Crossover,Luxury,Performance                                 3
Exotic,Factory Tuner,Luxury,Performance                      3
Luxury,High-Performance,Hybrid                               3
Crossover,Luxury,High-Performance                            2
Luxury                                                       2
Factory Tuner,Luxury                                         2
Factory Tuner,High-Performance                               2
Luxury,Performance                                           2
Exotic,Luxury,High-Performance,Hybrid                        1
Flex Fuel,Factory Tuner,Luxury,High-Performance              1
Name: count, dtype: int64
```

Dealing with Outliers in **MSRP** column

Analysis

Insight Required:

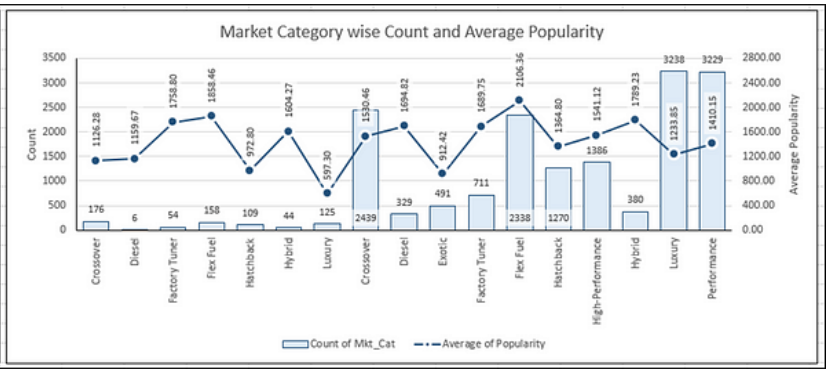
How does the popularity of a car model vary across different market categories?

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

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

Result:

Row Labels	Count of Mkt_Cat	Average of Popularity
Crossover	176	1126.28
Diesel	6	1159.67
Factory Tuner	54	1758.80
Flex Fuel	158	1858.46
Hatchback	109	972.80
Hybrid	44	1604.27
Luxury	125	597.30
Crossover	2439	1530.46
Diesel	329	1694.82
Exotic	491	912.42
Factory Tuner	711	1689.75
Flex Fuel	2338	2106.36
Hatchback	1270	1364.80
High-Performance	1386	1541.12
Hybrid	380	1789.23
Luxury	3238	1233.85
Performance	3229	1410.15



Task 1A and 1B

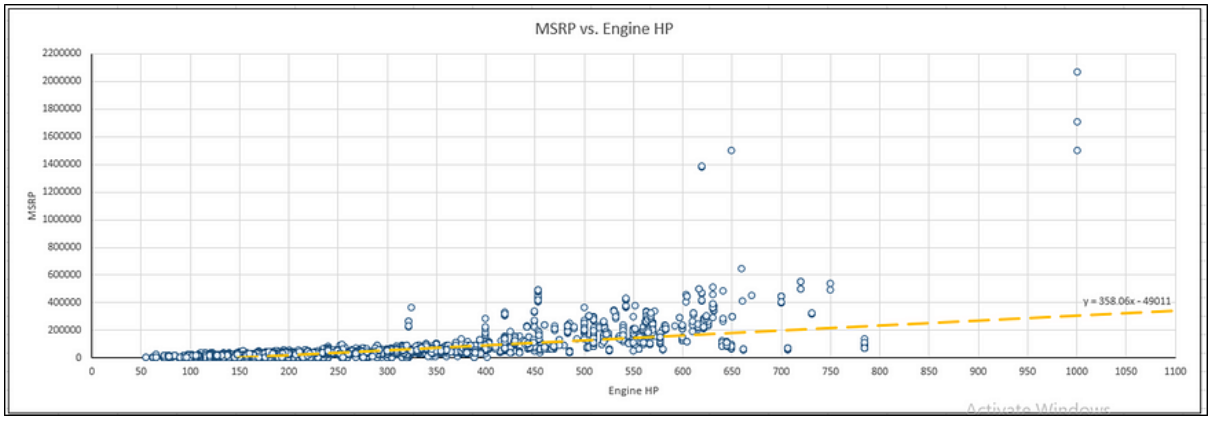
- We can observe that the average popularity of cars based on their **Market Category** is almost the same with the exception of **Luxury** cars being the **lowest** popular and **Flex Fuel** cars being the most **popular**.
- The dataset has comparatively **higher** number of **Performance** and **Luxury** cars followed by **Crossovers** and **Flex Fuel** 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.

Result:



Task 2

- 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. Also cars with higher **Engine HP** are mostly **Performance** cars.

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.

Result:

Final Regression Analysis

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.836939232
R Square	0.700467278
Adjusted R Square	0.700038685
Standard Error	33701.96062
Observations	11199

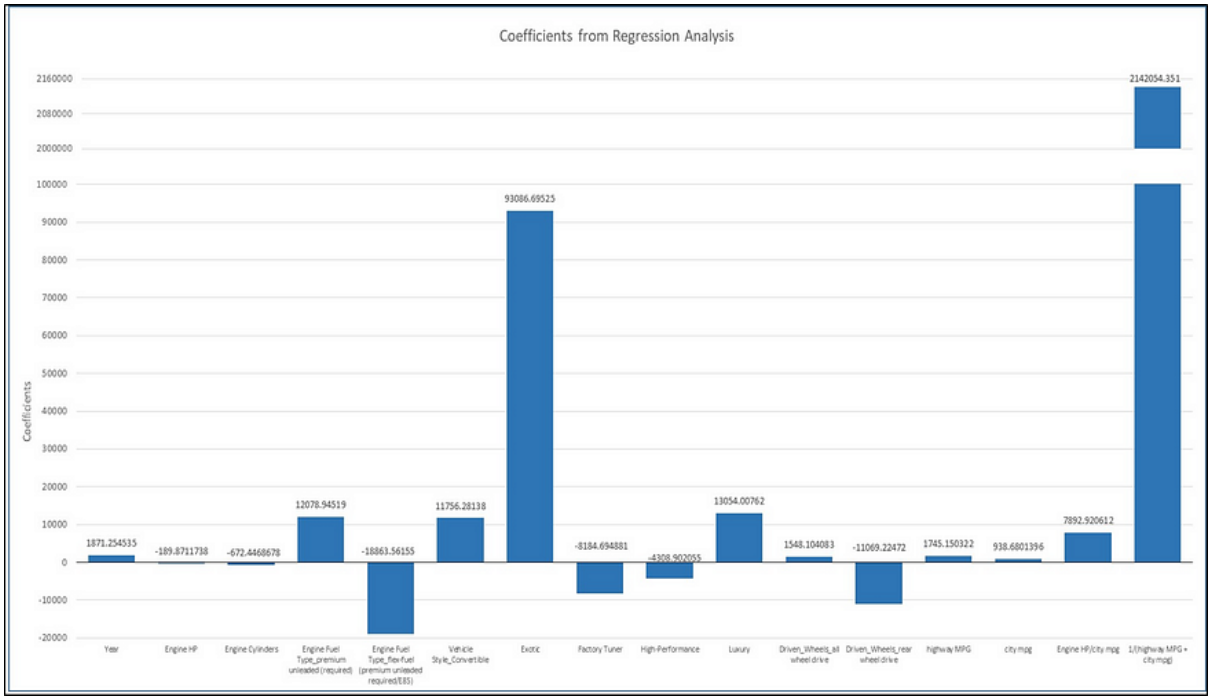
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	16	2.97012E+13	1.85632E+12	1634.342536	0
Residual	11182	1.27008E+13	1135822150		
Total	11198	4.24019E+13			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-3988458.386	145137.6152	-27.4805286	6.1866E-161	-4272953.679	-3703963.093	-4272953.679	-3703963.093
Year	1871.254535	69.62081994	26.87780088	2.9785E-154	1734.785464	2007.723606	1734.785464	2007.723606
Engine HP	-189.8711738	9.712740602	-19.54867133	1.04842E-83	-208.9098563	-170.8324912	-208.9098563	-170.8324912
Engine Cylinders	-672.4468678	427.7919571	-1.571901614	0.116001663	-1510.994463	166.1007271	-1510.994463	166.1007271
Engine Fuel Type_premium unleaded (required)	12078.94519	1138.046179	10.61375664	3.42634E-26	9848.174203	14309.71618	9848.174203	14309.71618
Engine Fuel Type_flex-fuel (premium unleaded required/E85)	-18863.56155	4866.117474	-3.876511747	0.000106571	-28402.00901	-9325.114098	-28402.00901	-9325.114098
Vehicle Style_Convertible	11756.28138	1360.151319	8.643362851	6.20069E-18	9090.145194	14422.41757	9090.145194	14422.41757
Exotic	93086.69525	2138.261713	43.53381753	0	88895.32562	97278.06488	88895.32562	97278.06488
Factory Tuner	-8184.694881	1420.939036	-5.760060546	8.63008E-09	-10969.9857	-5399.40406	-10969.9857	-5399.40406
High-Performance	-4308.902055	1467.806607	-2.935606118	0.003335744	-7186.06157	-1431.74254	-7186.06157	-1431.74254
Luxury	13054.00762	834.3417307	15.64587643	1.34197E-54	11418.55085	14689.46439	11418.55085	14689.46439
Driven_Wheels_all wheel drive	1548.104083	940.5028344	1.646038721	0.099783891	-295.4471498	3391.655315	-295.4471498	3391.655315
Driven_Wheels_rear wheel drive	-11069.22472	862.209357	-12.83820992	1.84113E-37	-12759.30694	-9379.142492	-12759.30694	-9379.142492
highway MPG	1745.150322	154.1490377	11.32118856	1.49236E-29	1442.991054	2047.309591	1442.991054	2047.309591
city mpg	938.6801396	164.5406492	5.704852536	1.194E-08	616.1514821	1261.208797	616.1514821	1261.208797
Engine HP/city mpg	7892.920612	136.7549606	57.71579016	0	7624.856798	8160.984425	7624.856798	8160.984425
1/(highway MPG + city mpg)	2142054.351	101698.2025	21.0628536	1.29369E-96	1942707.959	2341400.743	1942707.959	2341400.743

Task 3

- Using regression analysis, we found the top columns. This also include two new columns which were **Feature Engineered** (**Engine HP/city mpg**) and (**1/(highway MPG + city mpg)**).
- We can observe that the **R-Squared** score is **0.7** which can be counted as a good score.



Task 3

- We can observe that the highest coefficient value is that of **Engineered Feature, 1/(highway MPG + city mpg)**.
- This shows that the **Engineered Feature** is very important relationship with Car's price.

Insight Required:

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

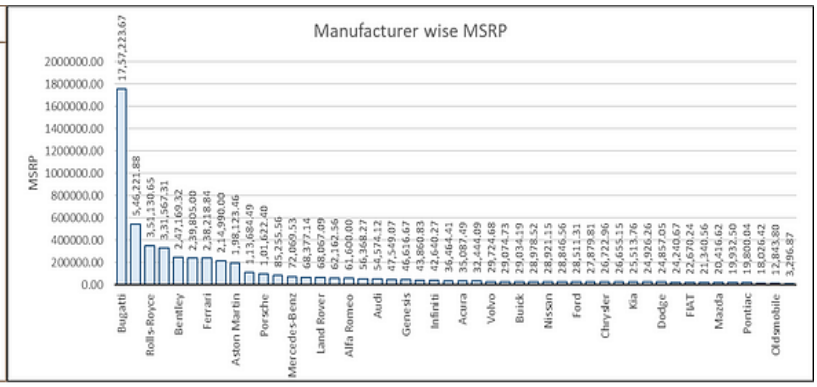
Task 4.A: Create a pivot table that shows the average price of cars for each manufacturer.

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

Result:

Row Labels	Average of MSRP
Bugatti	1757223.67
Maybach	546221.88
Rolls-Royce	351130.65
Lamborghini	331567.31
Bentley	247169.32
McLaren	239805.00
Ferrari	238218.84
Spyker	214990.00
Aston Martin	198123.46
Maserati	113684.49
Porsche	101622.40
Tesla	85255.56
Mercedes-Benz	72069.53
Lotus	68377.14
Land Rover	68067.09
BMW	62162.56
Alfa Romeo	61600.00
Cadillac	56368.27
Audi	54574.12
Lexus	47549.07
Genesis	46616.67
Lincoln	43860.83
Infiniti	42640.27
HUMMER	36464.41
Acura	35087.49
GMC	32444.09
Volvo	29724.68
Chevrolet	29074.73
Buick	29034.19
Volkswagen	28978.52
Nissan	28921.15
Toyota	28846.56
Ford	28511.31
Saab	27879.81

Row Labels	Average of MSRP
Chrysler	26722.96
Honda	26655.15
Kia	25513.76
Hyundai	24926.26
Dodge	24857.05
Subaru	24240.67
FIAT	22670.24
Mitsubishi	21340.56
Mazda	20416.62
Scion	19932.50
Pontiac	19800.04
Suzuki	18026.42
Oldsmobile	12843.80
Plymouth	3296.87



Task 4A and 4B

- 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.

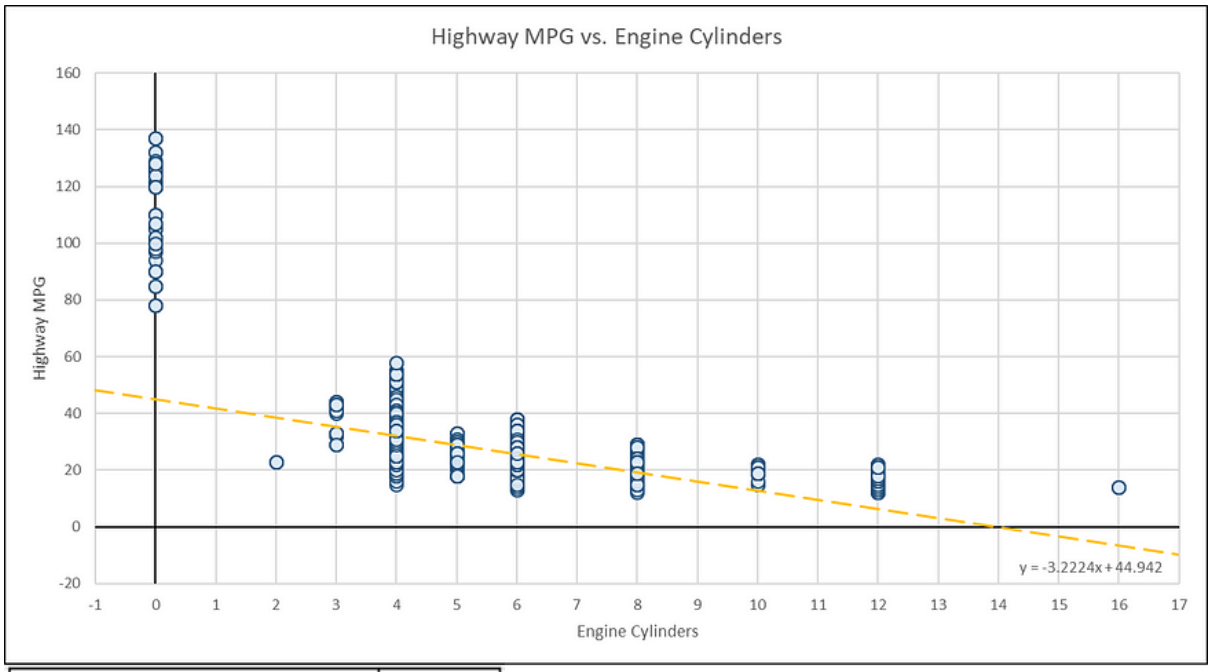
Insight Required:

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.

Result:



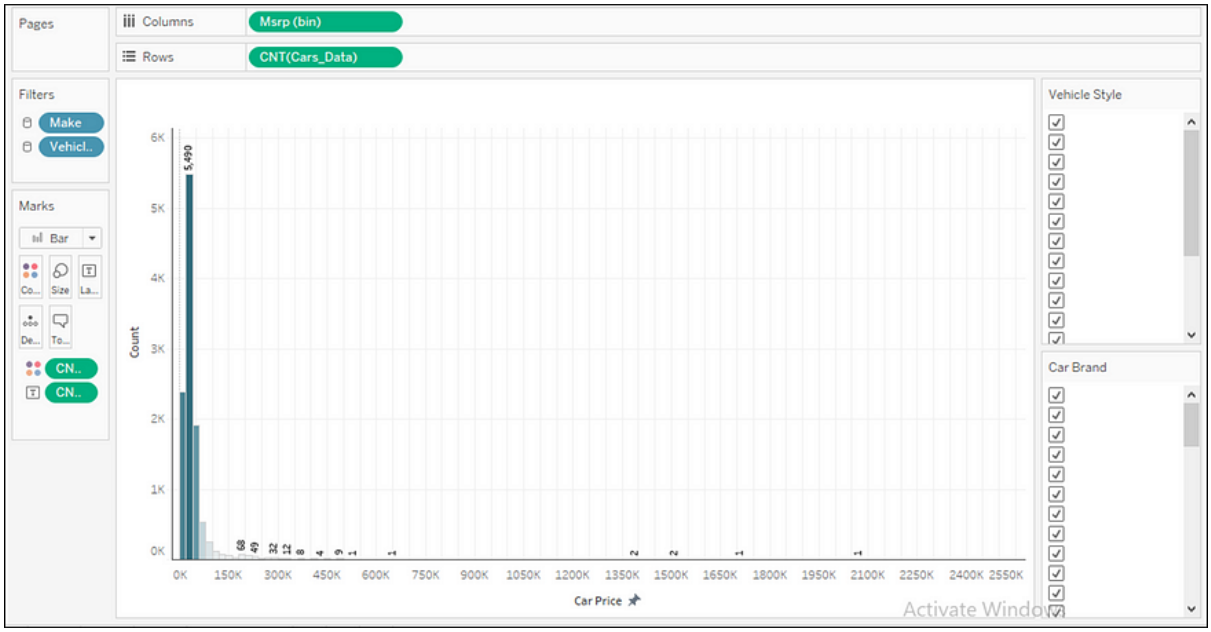
Correlation between **Engine Cylinders** and **highway MPG:**	**-0.62294**

Task 5A and 5B

- We can observe that the plot between **highway MPG** and **Engine Cylinders** has a negative slope with a value of **-3.2224**.
- The correlation coefficient is also **Negative** with a value of **-0.62294**.
- 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**).

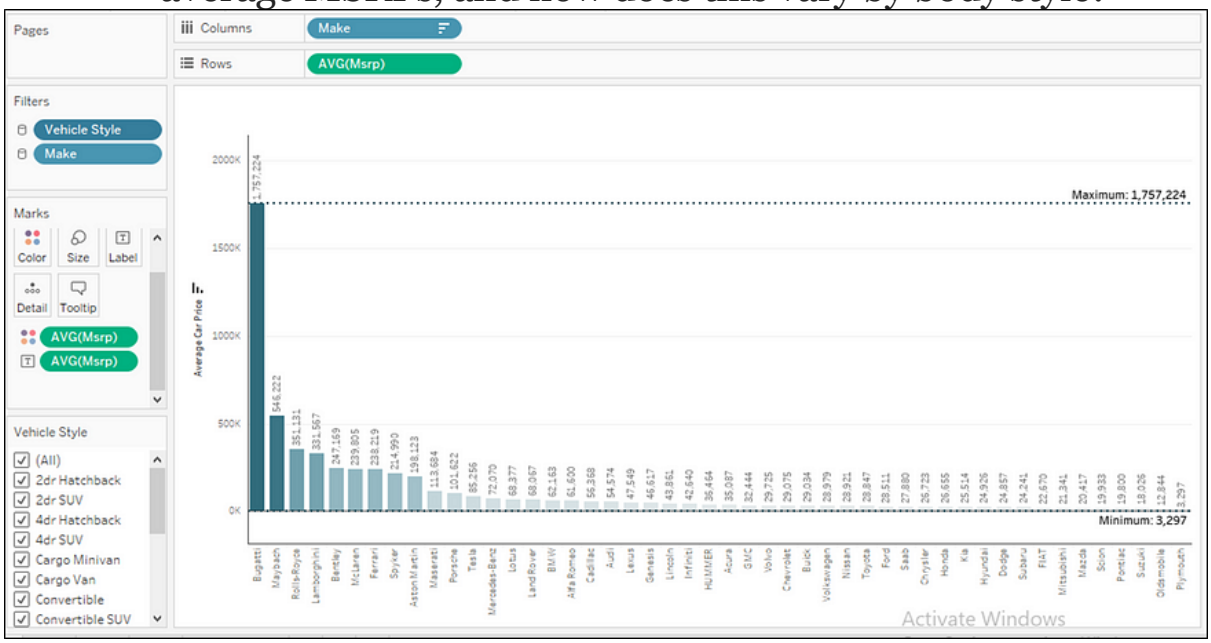
Dashboard

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



Task 1 Plot Sheet

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

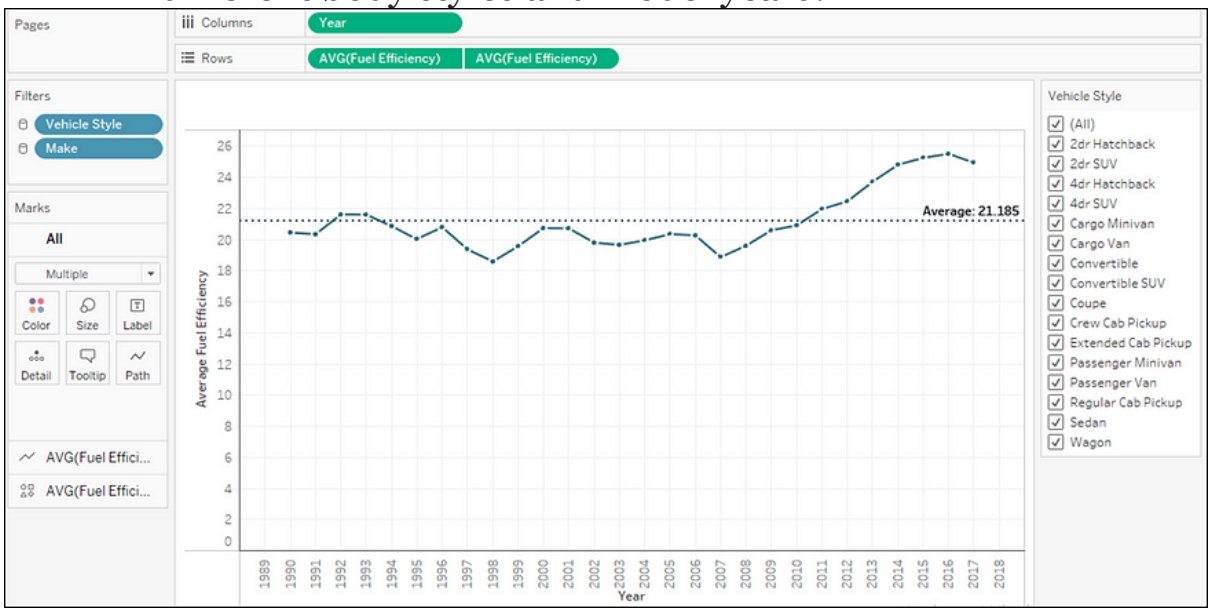


Task 2 Plot Sheet

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

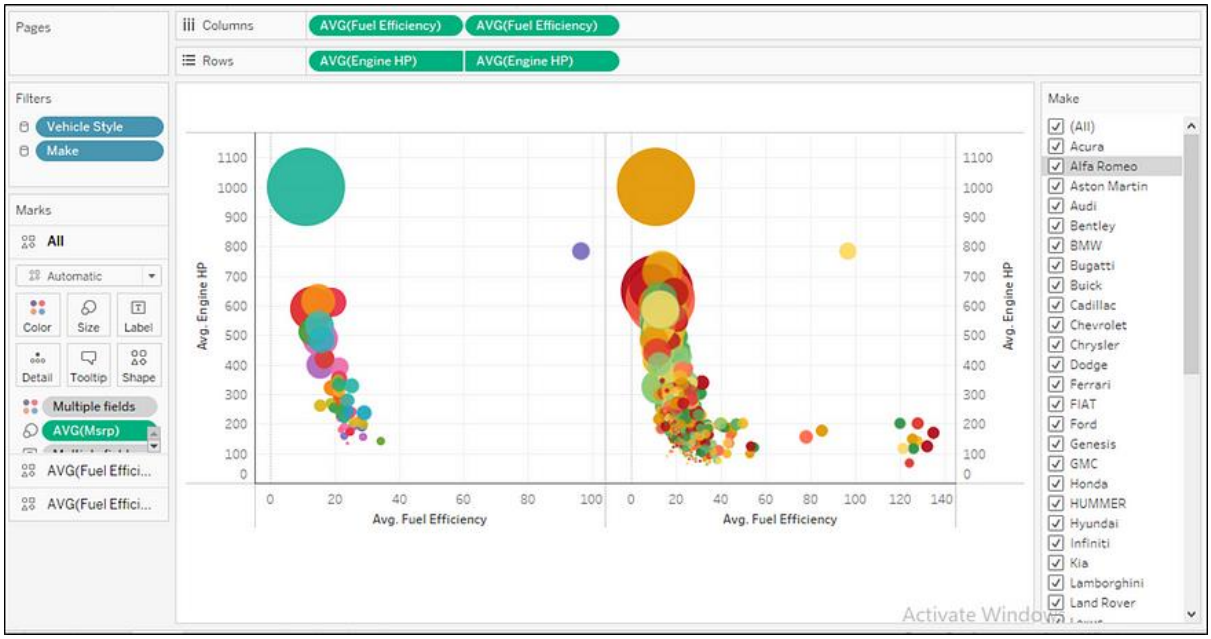
Task 3 Plot Sheet

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



Task 4 Plot Sheet

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



Task 5 Plot Sheet

- **Result:**

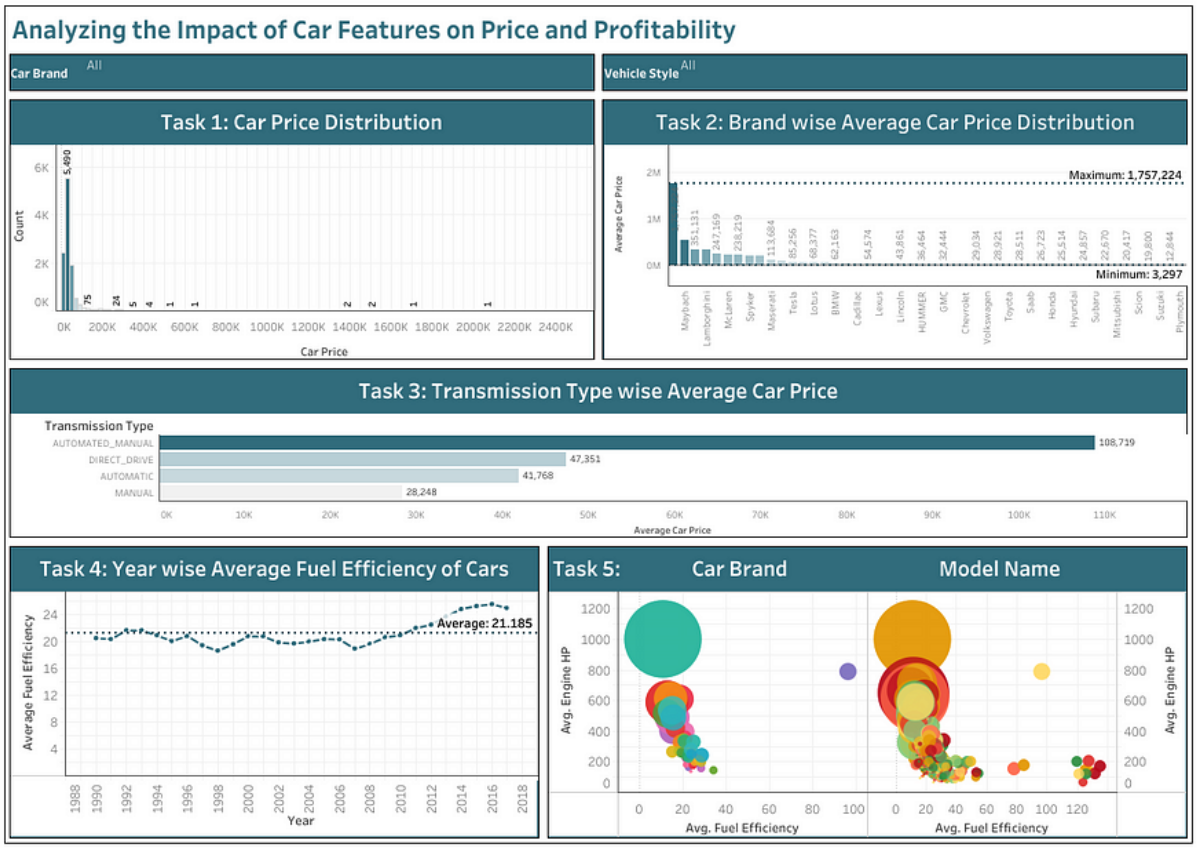


Tableau Dashboard

Conclusion

Through this project, I was able to understand the importance of **Data Analytics** in **Car Feature Analysis** as it provides valuable insights which helps in making **Data-Driven Decisions**.

In this project I was able to get insights like which features effects Car Price, relationship between Engine Cylinders and it's fuel efficiency etc. I also got experience in Data Preprocessing like Data Cleaning, handling Outliers, Feature Engineering etc. in this project which can be **communicated** to relevant stakeholders as per the requirements.