ANALYZING THE IMPACT OF CAR FEATURES ON PRICE AND PROFITABILITY

MIRRA. G

PROJECT DESCRIPTION:

This project aims to build an interactive Excel dashboard to provide insights into various aspects of the automotive market using a comprehensive dataset of car models. The analysis involves several key tasks: examining how car model popularity varies across market categories, exploring the relationship between engine power and price through scatter plots, and identifying the most influential car features on pricing using regression analysis. Additionally, the project investigates the average price variations across different manufacturers, the relationship between fuel efficiency and engine cylinders, and the distribution of car prices by brand and body style. Interactive elements like filters and slicers will enhance the user experience, allowing for dynamic exploration of the data. The final dashboard will feature various visualizations, including combo charts, scatter plots, bar charts, and bubble charts, to effectively communicate the insights and support data-driven decision-making for automotive stakeholders.

TECH STACK USED: Microsoft Excel, Python (Pandas, Statsmodels & MatplotLib)

WHY EXCEL?

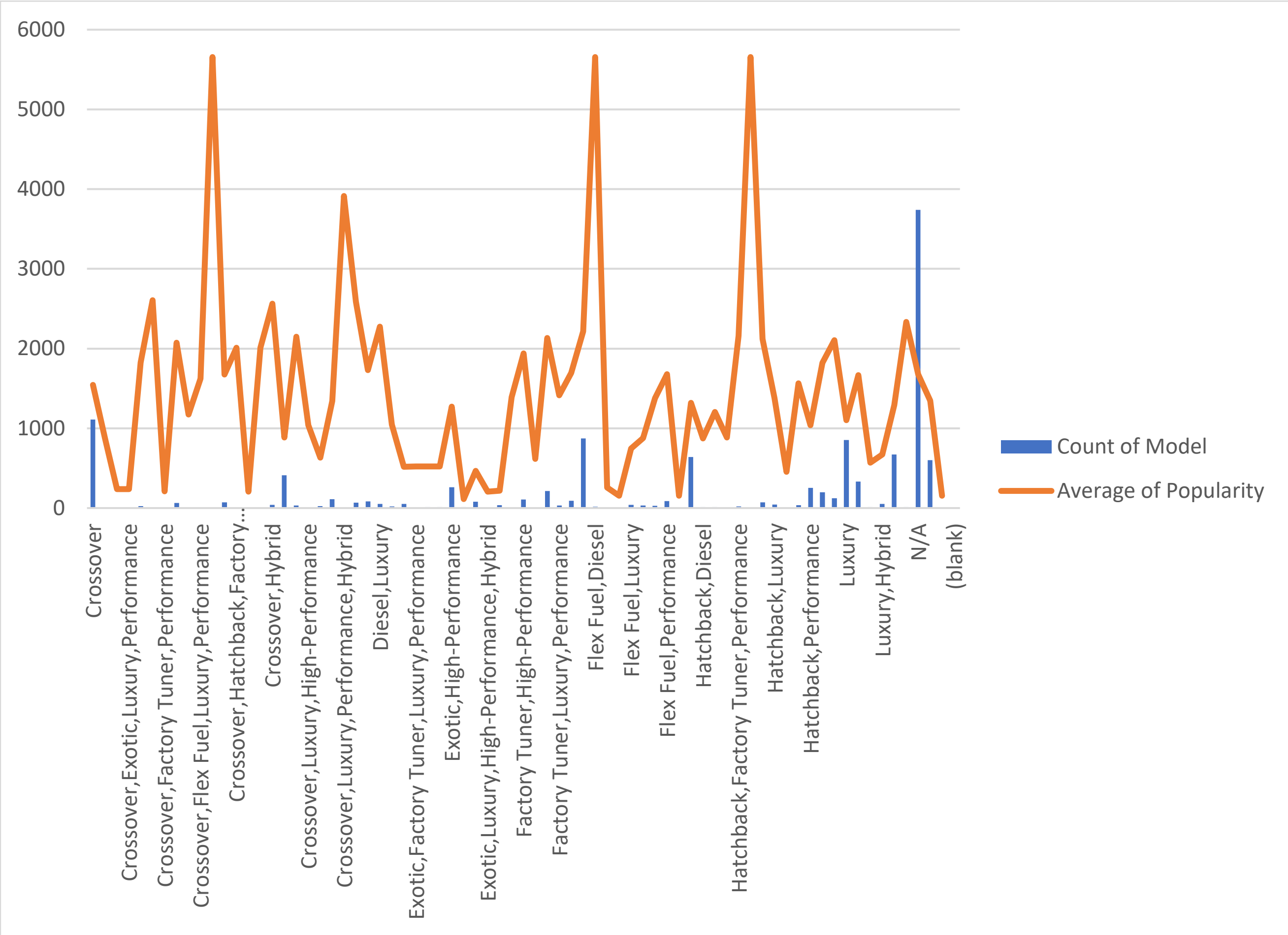
Using Excel and Python (pandas and statsmodels) is an excellent choice for this automotive market analysis project due to their complementary strengths and capabilities. Excel is highly user-friendly, widely accessible, and excels in creating interactive dashboards with features like filters, slicers, and various chart types, making it ideal for visualizing data insights in a dynamic and engaging manner. Python, on the other hand, offers powerful libraries such as pandas and statsmodels, which are essential for handling large datasets, performing complex data manipulations, and conducting robust statistical analyses. Pandas efficiently manages data cleaning, transformation, and exploratory analysis, while statsmodels provides advanced statistical modelling and regression analysis capabilities. Combining Excel's interactive and visualization prowess with Python's data processing and analytical power ensures a comprehensive, insightful, and user-friendly approach to understanding the automotive market dynamics.

ANALYSIS TASKS:

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

Row Labels	Count of Model	Average of Popularity
Crossover	1110	1545.263063
Crossover,Diesel	7	873
Crossover,Exotic,Luxury,High-Performance	1	238
Crossover,Exotic,Luxury,Performance	1	238
Crossover,Factory Tuner,Luxury,High-Performance	26	1823.461538
Crossover,Factory Tuner,Luxury,Performance	5	2607.4
Crossover,Factory Tuner,Performance	4	210
Crossover,Flex Fuel	64	2073.75
Crossover,Flex Fuel,Luxury	10	1173.2
Crossover,Flex Fuel,Luxury,Performance	6	1624
Crossover,Flex Fuel,Performance	6	5657
Crossover,Hatchback	72	1675.694444
Crossover,Hatchback,Factory Tuner,Performance	6	2009
Crossover,Hatchback,Luxury	7	204
Crossover,Hatchback,Performance	6	2009
Crossover,Hybrid	42	2563.380952
Crossover,Luxury	410	884.5487805
Crossover,Luxury,Diesel	34	2149.411765
Crossover,Luxury,High-Performance	9	1037.222222
Crossover,Luxury,Hybrid	24	630.9166667
Crossover,Luxury,Performance	113	1344.849558
Crossover,Luxury,Performance,Hybrid	2	3916
Crossover,Performance	69	2585.956522
Diesel	84	1730.904762
Diesel,Luxury	51	2275
Exotic,Factory Tuner,High-Performance	21	1046.380952
Exotic,Factory Tuner,Luxury,High-Performance	52	517.5384615
Exotic,Factory Tuner,Luxury,Performance	3	520
Exotic,Flex Fuel,Factory Tuner,Luxury,High-Performance	13	520
Exotic,Flex Fuel,Luxury,High-Performance	11	520
Exotic,High-Performance	261	1271.333333
Exotic,Luxury	12	112.6666667
Exotic,Luxury,High-Performance	79	467.0759494
Exotic,Luxury,High-Performance,Hybrid	1	204
Exotic,Luxury,Performance	36	217.0277778
Exotic,Performance	10	1391
Factory Tuner,High-Performance	106	1941.415094
Factory Tuner,Luxury	2	617
Factory Tuner,Luxury,High-Performance	215	2133.367442
Factory Tuner,Luxury,Performance	31	1413.419355
Factory Tuner,Performance	92	1695.695652
Flex Fuel	872	2217.302752
Flex Fuel,Diesel	16	5657
Flex Fuel,Factory Tuner,Luxury,High-Performance	1	258
Flex Fuel,Hybrid	2	155
Flex Fuel,Luxury	39	746.5384615
Flex Fuel,Luxury,High-Performance	33	878.9090909
Flex Fuel,Luxury,Performance	28	1380.071429
Flex Fuel,Performance	87	1680.471264
Flex Fuel,Performance,Hybrid	2	155
Hatchback	641	1318.865835
Hatchback,Diesel	14	873
Hatchback,Factory Tuner,High-Performance	13	1205.153846
Hatchback,Factory Tuner,Luxury,Performance	9	886.8888889
Hatchback,Factory Tuner,Performance	22	2159.045455
Hatchback,Flex Fuel	7	5657
Hatchback,Hybrid	72	2121.25
Hatchback,Luxury	46	1379.5
Hatchback,Luxury,Hybrid	3	454
Hatchback,Luxury,Performance	38	1566.131579
Hatchback,Performance	252	1039.646825
High-Performance	199	1821.447236
Hybrid	123	2105.569106
Luxury	855	1102.65731
Luxury,High-Performance	334	1668.017964
Luxury,High-Performance,Hybrid	12	568.8333333
Luxury,Hybrid	52	673.6346154
Luxury,Performance	673	1292.615156
Luxury,Performance,Hybrid	11	2333.181818
N/A	3742	1676.889364
Performance	601	1348.873544
Performance,Hybrid	1	155
(blank)		
Grand Total	**11914**	**1554.911197**

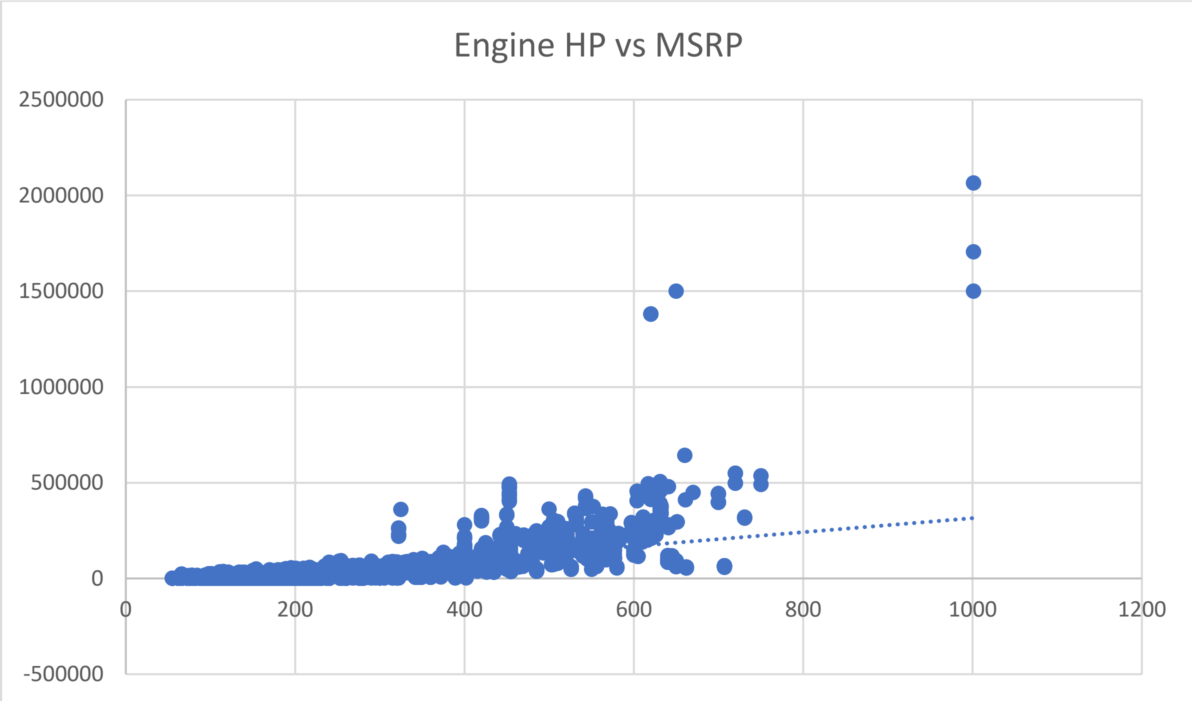
1. **B. Create a combo chart that visualizes the relationship between market category and popularity.**



INSIGHTS:

- *High Popularity Peaks:* Certain market categories have exceptionally high popularity scores, indicating they are more favoured by consumers despite possibly having fewer models.
- *Model Distribution:* The distribution of models is not uniform across market categories. Some categories have a dense concentration of models, while others have very few.
- *Disparity Between Popularity and Count:* The variation between the number of models and their popularity suggests that market popularity is not solely dependent on the number of available models but possibly on other factors like brand reputation, features, and market trends.

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.**



INSIGHTS:

- *Positive Correlation:* There is a general trend that as engine horsepower increases, the MSRP also increases. This indicates a positive correlation between engine HP and car price.
- *Clusters:* Most of the data points are clustered between 100 to 600 HP and between $0 to $500,000 MSRP, suggesting that the majority of cars fall within this range.
- *Trend Line:* The dotted trend line suggests a positive slope, reinforcing the idea of a positive correlation. However, the trend line is relatively flat compared to the spread of data, indicating that while there is a relationship, other factors might also be influencing the MSRP.

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.**

<u>Approach:</u>

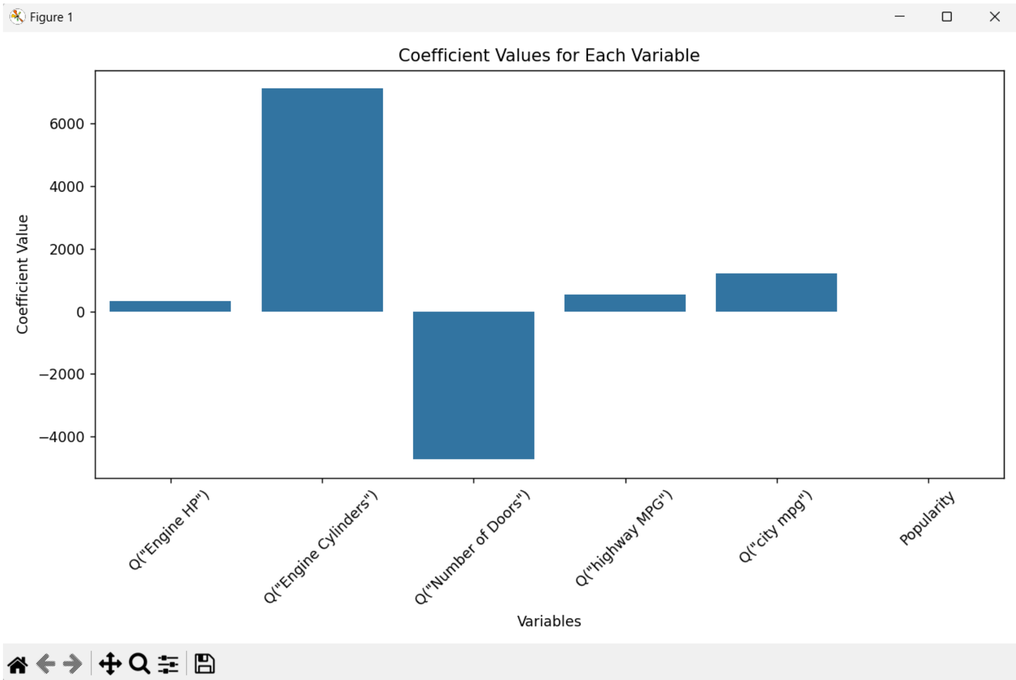
We use functions from various libraries such as Pandas, MatplotLib and Statsmodels in Python to complete this task.

Code:

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols
import matplotlib.pyplot as plt
import seaborn as sns

# Load the data
data = pd.read_excel(r'C:\Courses\TRAINITY\task 7\task 3.xlsx')

# Define the dependent variable (MSRP) and independent variables
X = data[['Engine HP', 'Engine Cylinders', 'Number of Doors', 'highway MPG',
'city mpg', 'Popularity']]
y = data['MSRP']

# Add a constant to the model (intercept)
X = sm.add_constant(X)

# Fit the regression model
model = sm.OLS(y, X).fit()

# Get the summary of the regression
summary = model.summary2()

# Extract necessary statistics
multiple_r = model.rsquared**0.5
r_square = model.rsquared
adjusted_r_square = model.rsquared_adj
standard_error = model.bse.mean()
observations = model.nobs

print("\t  Regression Statistics")
print(f"Multiple R       \t{multiple_r}")
print(f"R Square         \t{r_square}")
print(f"Adjusted R Square\t{adjusted_r_square}")
print(f"Standard Error   \t{standard_error}")
print(f"Observations     \t{observations}")

formula = 'MSRP ~ Q("Engine HP") + Q("Engine Cylinders") + Q("Number of Doors") +
Q("highway MPG") + Q("city mpg") + Popularity'

# Fit the model
model = ols(formula, data=data).fit()

# Perform ANOVA
print("\n\t\t\t\t  ANOVA Table")
anova_table = sm.stats.anova_lm(model, typ=2)
print(anova_table)

# Extract the coefficients
coefficients = model.params[1:]  # Exclude the intercept

# Create the bar chart
plt.figure(figsize=(10, 6))
sns.barplot(x=coefficients.index, y=coefficients.values)
plt.xlabel('Variables')
plt.ylabel('Coefficient Value')
plt.title('Coefficient Values for Each Variable')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Output:

```
PS C:\Courses\TRAINITY\task 7> & "C:/Program Files/Python311/python.exe" "c:/Courses/TRAINITY/task 7/task3.py"
          Regression Statistics
Multiple R                0.6853063110272515
R Square                  0.46964473993378
Adjusted R Square         0.46937525047236417
Standard Error            688.7236906928455
Observations              11815.0

                                   ANOVA Table
                              sum_sq       df            F        PR(>F)
Q("Engine HP")          5.626482e+12      1.0  2918.361271  0.000000e+00
Q("Engine Cylinders")   5.113705e+11      1.0   265.239249  5.428696e-59
Q("Number of Doors")    2.001357e+11      1.0   103.807010  2.810167e-24
Q("highway MPG")        4.967971e+10      1.0    25.768025  3.908767e-07
Q("city mpg")           1.947314e+11      1.0   101.003870  1.142813e-23
Popularity              2.692861e+11      1.0   139.674124  4.753362e-32
Residual                2.276535e+13  11808.0          NaN           NaN
```



Link: https://drive.google.com/file/d/1_jgsTzrzFRfgL7Nqftgu9Tw7ijZC5eto/view?usp=sharing

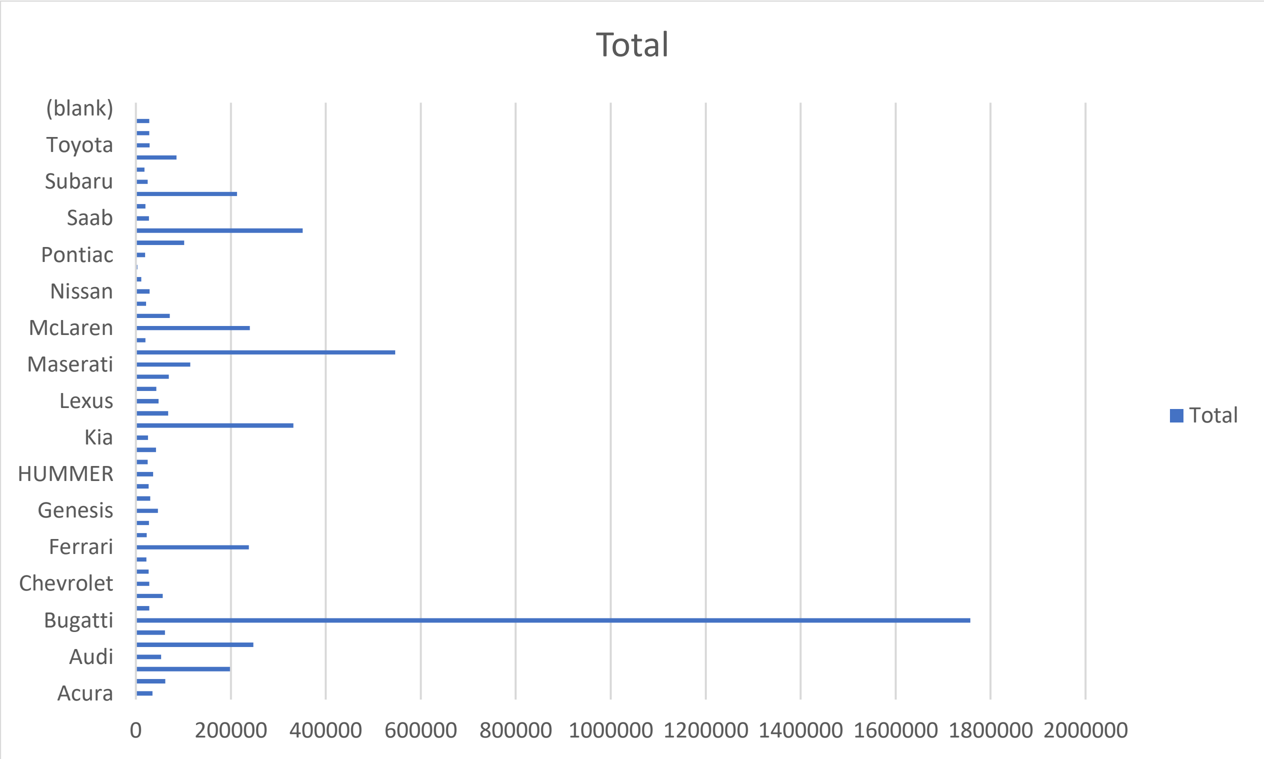
INSIGHTS:

The regression analysis indicates that Engine HP, Engine Cylinders, Number of Doors, highway MPG, city mpg, and Popularity significantly affect MSRP, with Engine HP having the largest impact ($F=2918.36$, $p<0.001$). The model explains approximately 47% of the variability in MSRP ($R^2=0.470$), suggesting a moderate fit. The ANOVA table confirms the significance of all predictors, with p-values close to zero, indicating strong evidence against the null hypothesis for each variable. A bar chart of the regression coefficients reveals that Engine HP and Engine Cylinders are the most influential variables, followed by Popularity, city mpg, Number of Doors, and highway MPG, in descending order of importance.

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

Manufacturer	Average Price
Acura	34887.5873
Alfa Romeo	61600
Aston Martin	197910.3763
Audi	53452.1128
Bentley	247169.3243
BMW	61546.76347
Bugatti	1757223.667
Buick	28206.61224
Cadillac	56231.31738
Chevrolet	28350.38557
Chrysler	26722.96257
Dodge	22390.05911
Ferrari	238218.8406
FIAT	22670.24194
Ford	27399.26674
Genesis	46616.66667
GMC	30493.29903
Honda	26674.34076
HUMMER	36464.41176
Hyundai	24597.0363
Infiniti	42394.21212
Kia	25310.17316
Lamborghini	331567.3077
Land Rover	67823.21678
Lexus	47549.06931
Lincoln	42839.82927
Lotus	69188.27586
Maserati	114207.7069
Maybach	546221.875
Mazda	20039.38298
McLaren	239805
Mercedes-Benz	71476.22946
Mitsubishi	21240.53521
Nissan	28583.4319
Oldsmobile	11542.54
Plymouth	3122.902439
Pontiac	19321.54839
Porsche	101622.3971
Rolls-Royce	351130.6452
Saab	27413.5045
Scion	19932.5
Spyker	213323.3333
Subaru	24827.50391
Suzuki	17907.20798
Tesla	85255.55556
Toyota	29030.01609
Volkswagen	28102.38072
Volvo	28541.16014
(blank)	
Grand Total	**40594.73703**

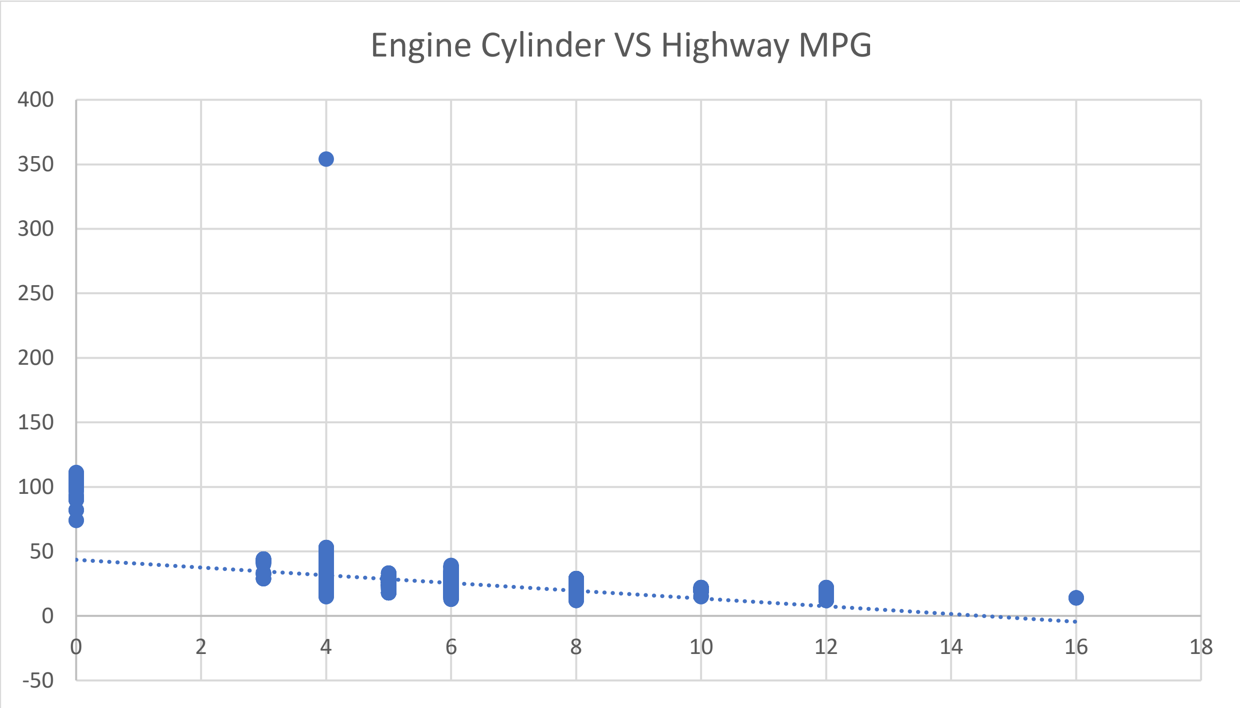
4. B. Create a bar chart or a horizontal stacked bar chart that visualizes the relationship between manufacturer and average price.



INSIGHTS:

The average price of cars varies significantly across different manufacturers. High-end manufacturers like Alfa Romeo, Aston Martin, and Bentley have much higher average prices, often exceeding $100,000, reflecting their luxury and performance-focused models. Mid-range brands like BMW, Audi, and Mercedes-Benz also show high average prices, typically in the $50,000 to $70,000 range, aligning with their premium market positioning. In contrast, mass-market manufacturers such as Chevrolet, Ford, and Toyota have lower average prices, generally under $30,000, indicating their focus on affordability and broad market appeal. This variation highlights the diversity in market positioning and target consumer segments among different car manufacturers.

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.**



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

CORRELATION COEFFICIENT	-0.62161

INSIGHTS:

The scatter plot demonstrates a negative correlation between the number of cylinders in a car's engine and its highway miles per gallon (MPG). As the number of cylinders increases, the highway MPG generally decreases, indicating that cars with more cylinders tend to have lower fuel efficiency on the highway. This trend is consistent, although there is a notable outlier with 4 cylinders achieving an exceptionally high MPG.

Link to Analysis Files:

https://drive.google.com/drive/folders/13NzaWrMnoQorREY5Yjl7krT-7YxPu2Sf?usp=sharing

RESULT:

Hence, we have completed all the analysis tasks given as a part of the Analysing the Impact of Car Features on Price and Profitability and built an interactive dashboard as per the requirements.