NAME : NIKHIL ANAND
EMAIL : NIKHIL3313@GMAIL.COM

TASK 7 : IMPACT OF CAR FEATURES (FINAL PROJECT)

TECH STACK USED : PYTHON

- The purpose of this project is to analyze the impact of car features on price and profitability in the automotive industry. By exploring a dataset containing information on various car models and their specifications, we aim to provide insights that can help car manufacturers optimize their pricing and product development decisions while meeting consumer demand. The dataset was collected and made available by Cooper Union and contains variables such as make, model, year, engine specifications, transmission type, market category, size, style, fuel efficiency, popularity, and manufacturer's suggested retail price (MSRP).

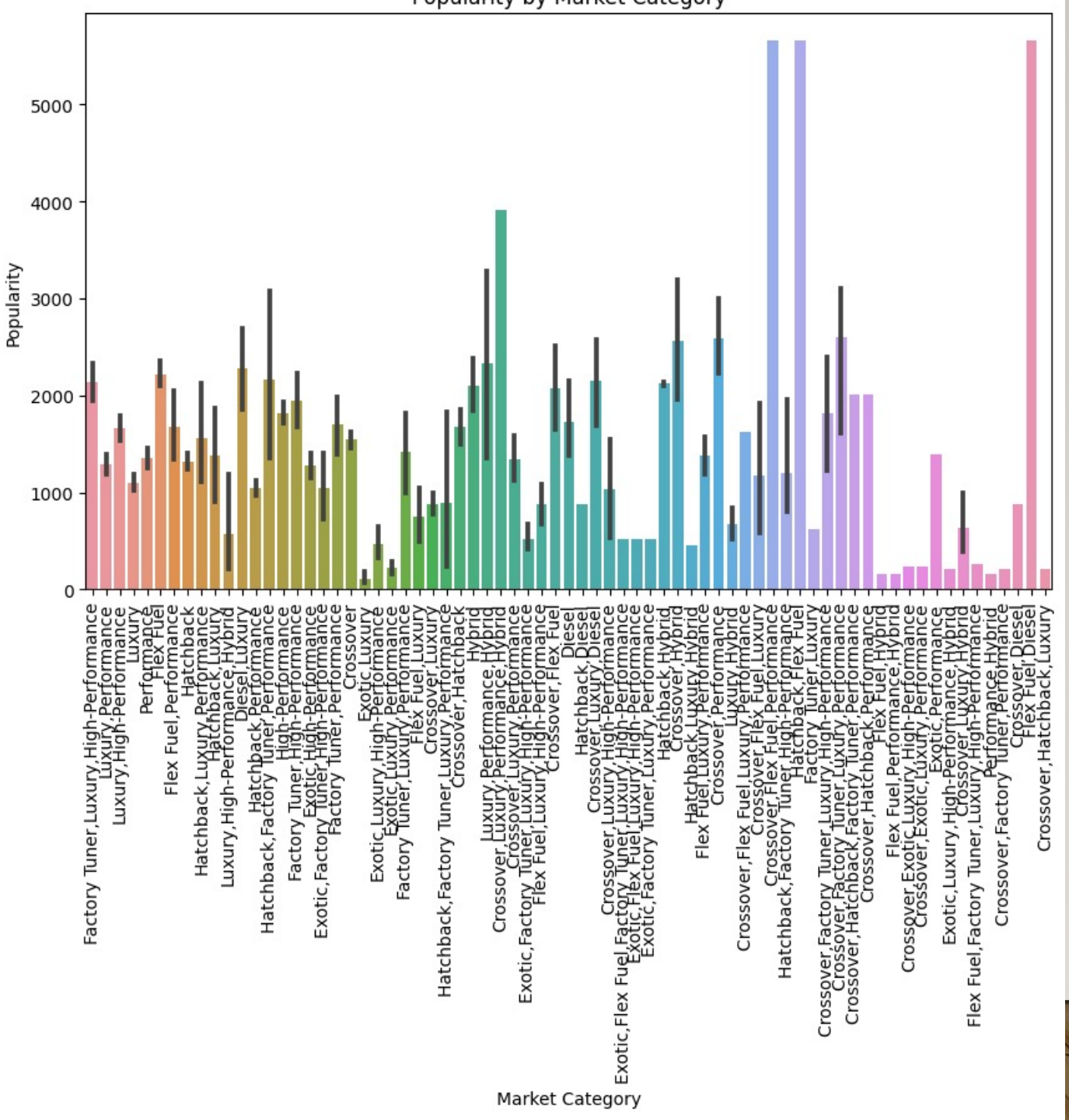
Business and Data Analytics

→ **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.

Market Category	Popularity
Crossover	1545.263063
Crossover,Diesel	873.000000
Crossover,Exotic,Luxury,High-Performance	238.000000
Crossover,Exotic,Luxury,Performance	238.000000
Crossover,Factory Tuner,Luxury,High-Performance	1823.461538
...	...
Luxury,Hybrid	673.634615
Luxury,Performance	1292.615156
Luxury,Performance,Hybrid	2333.181818
Performance	1348.873544
Performance,Hybrid	155.000000

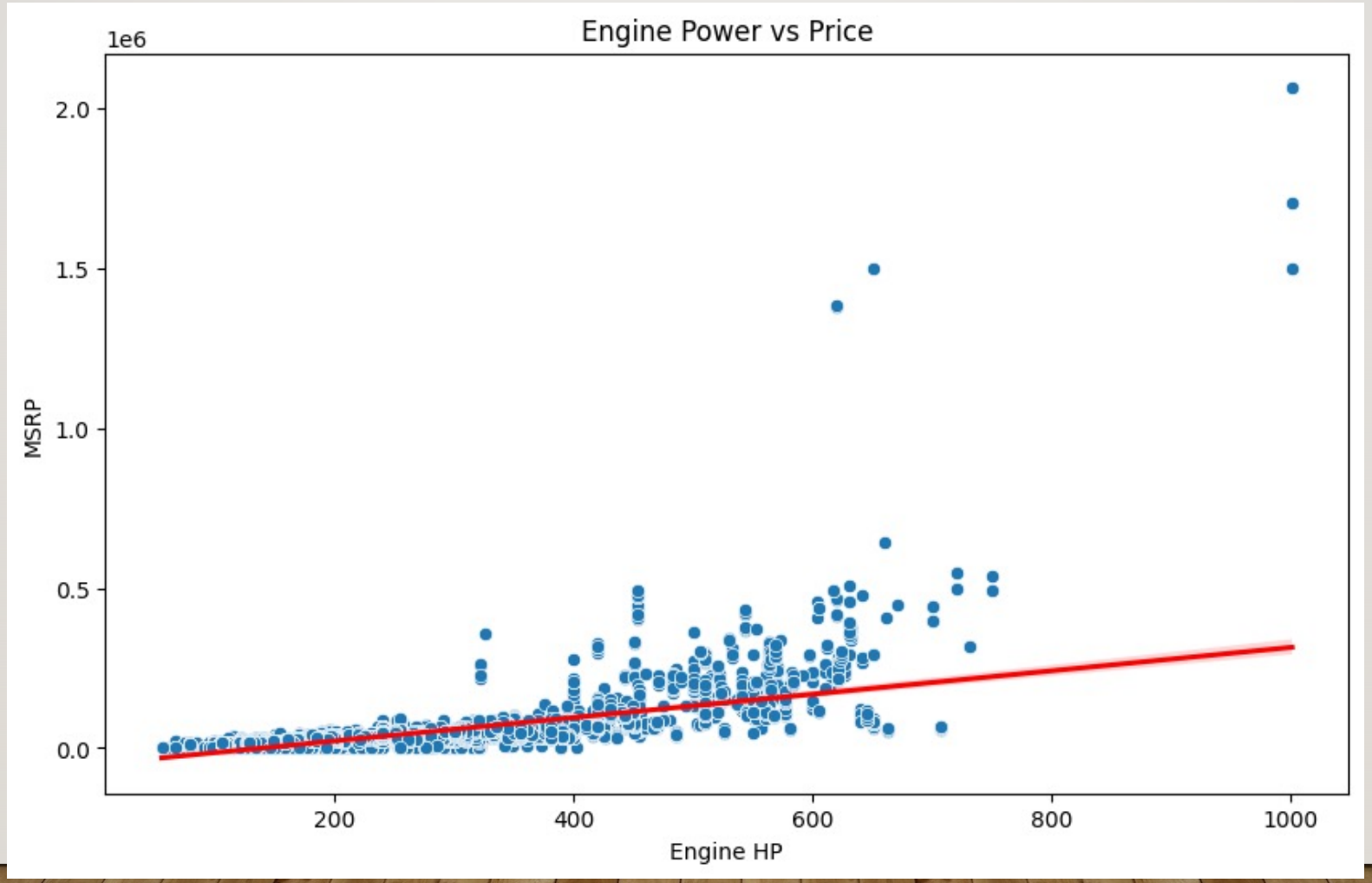
[71 rows x 1 columns]



Business and Data Analytics

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

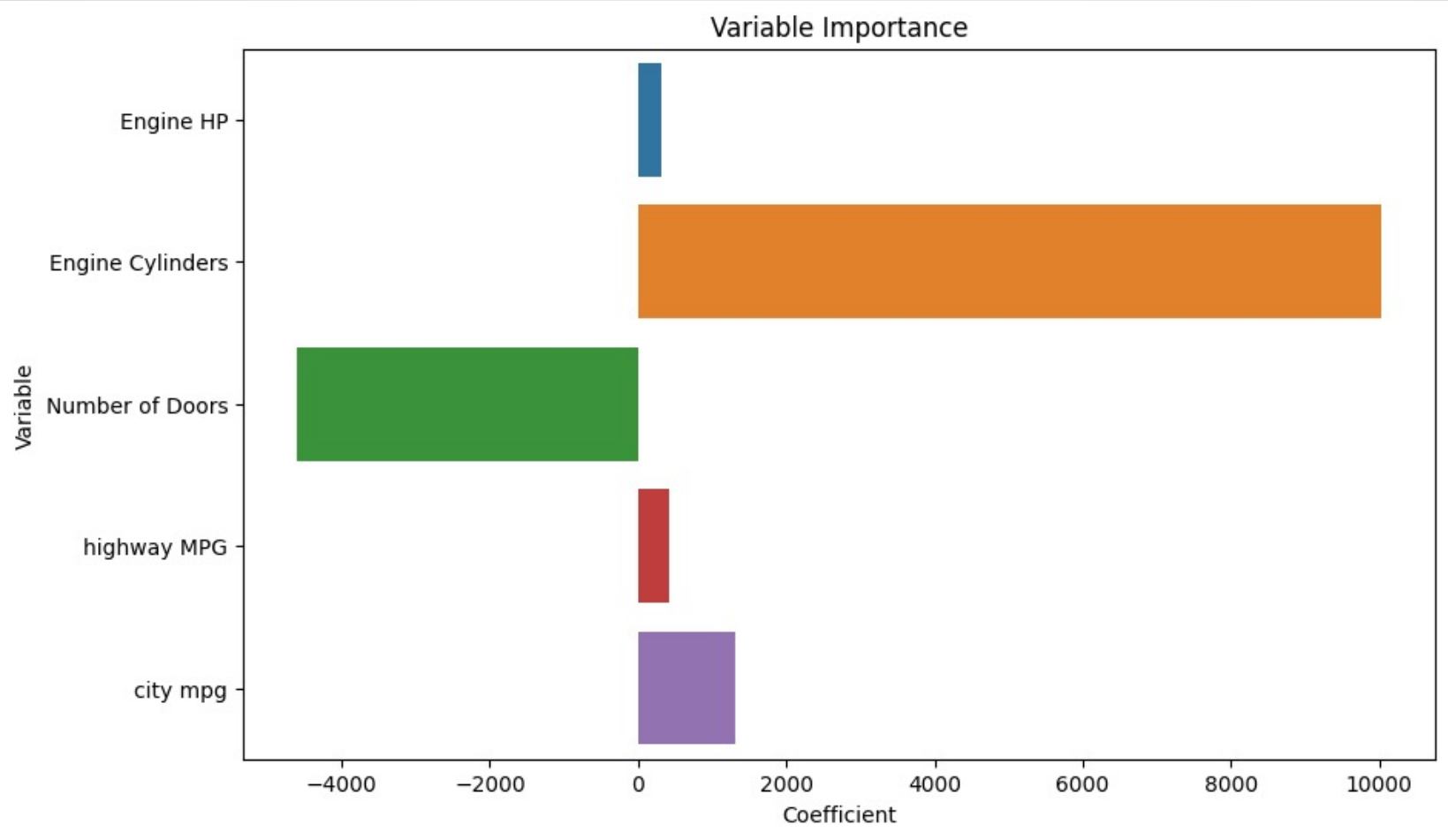
Task 2: Create a scatter chart that plots engine power on the x-axis and price on the y-axis. Add a trendline to the chart to visualize the relationship between these variables.



Business and Data Analytics

→ **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.



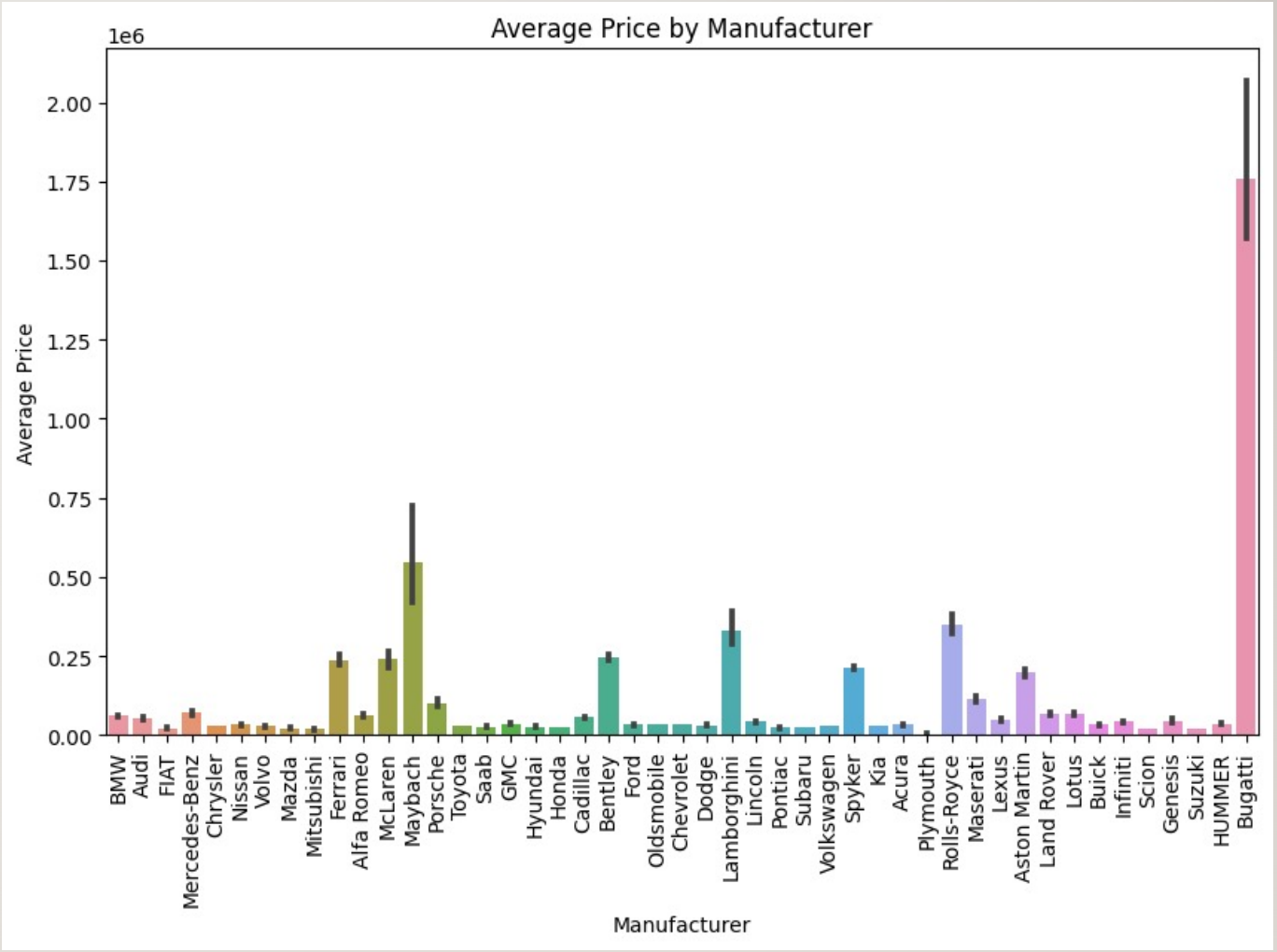
Business and Data Analytics

→ **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.

Make	MSRP
Acura	3.488759e+04
Alfa Romeo	6.160000e+04
Aston Martin	1.979104e+05
Audi	5.345211e+04
BMW	6.154676e+04
Bentley	2.471693e+05
Bugatti	1.757224e+06
Buick	3.377040e+04
Cadillac	5.623132e+04
Chevrolet	3.583328e+04
Chrysler	2.997887e+04
Dodge	3.099538e+04
FIAT	2.237066e+04
Ferrari	2.373838e+05
Ford	3.324537e+04
GMC	3.738575e+04
Genesis	4.661667e+04
HUMMER	3.646441e+04
Honda	2.695689e+04
Hyundai	2.698596e+04
Infiniti	4.239421e+04
Kia	3.014931e+04
Lamborghini	3.315673e+05
Land Rover	6.782322e+04
Lexus	4.754907e+04
Lincoln	4.249437e+04
Lotus	6.918828e+04
Maserati	1.142077e+05
Maybach	5.462219e+05
Mazda	2.298335e+04
McLaren	2.398050e+05
Mercedes-Benz	7.153781e+04
Mitsubishi	2.022188e+04
Nissan	3.290842e+04
Oldsmobile	3.486800e+04
Plymouth	4.076821e+03
Pontiac	2.472813e+04
Porsche	1.016224e+05
Rolls-Royce	3.511306e+05
Saab	2.741350e+04
Scion	2.039594e+04
Spyker	2.133233e+05
Subaru	2.640722e+04
Suzuki	2.115305e+04
Toyota	3.104422e+04
Volkswagen	2.990867e+04
Volvo	2.854116e+04

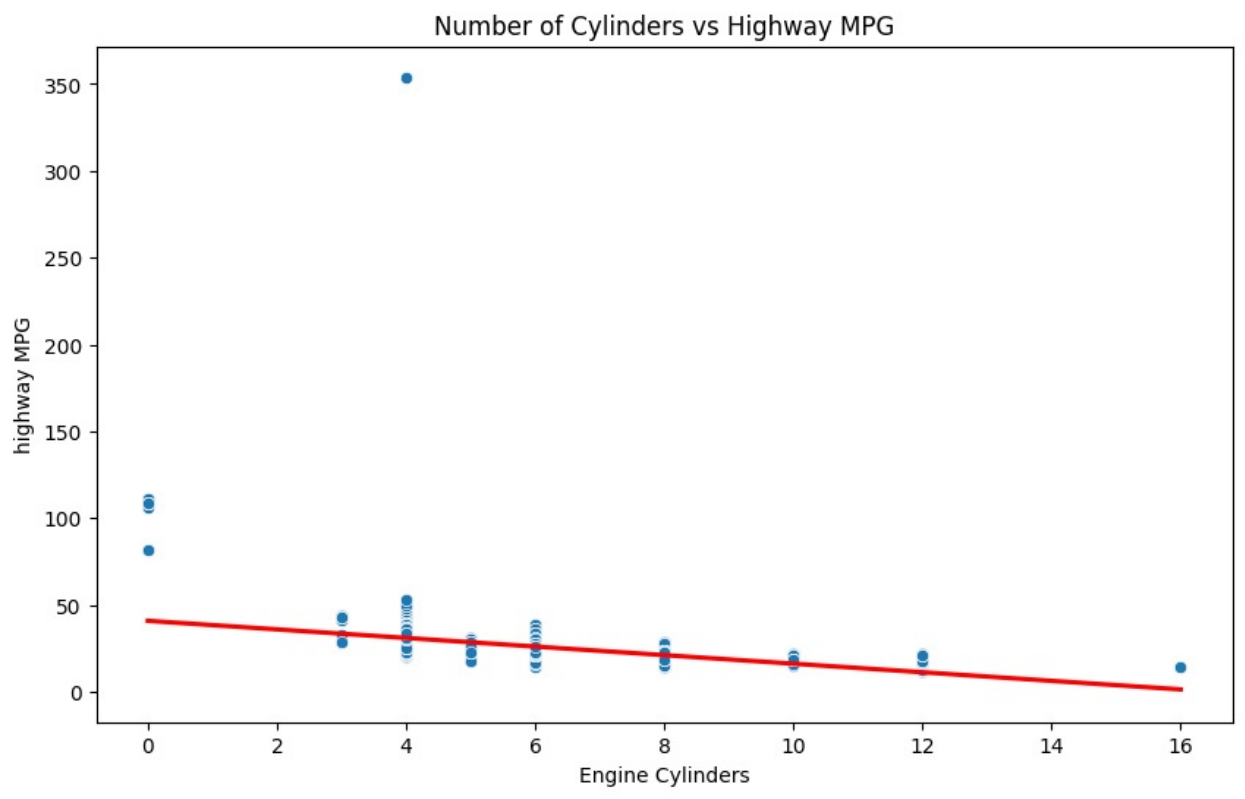


Business and Data Analytics

→ **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.

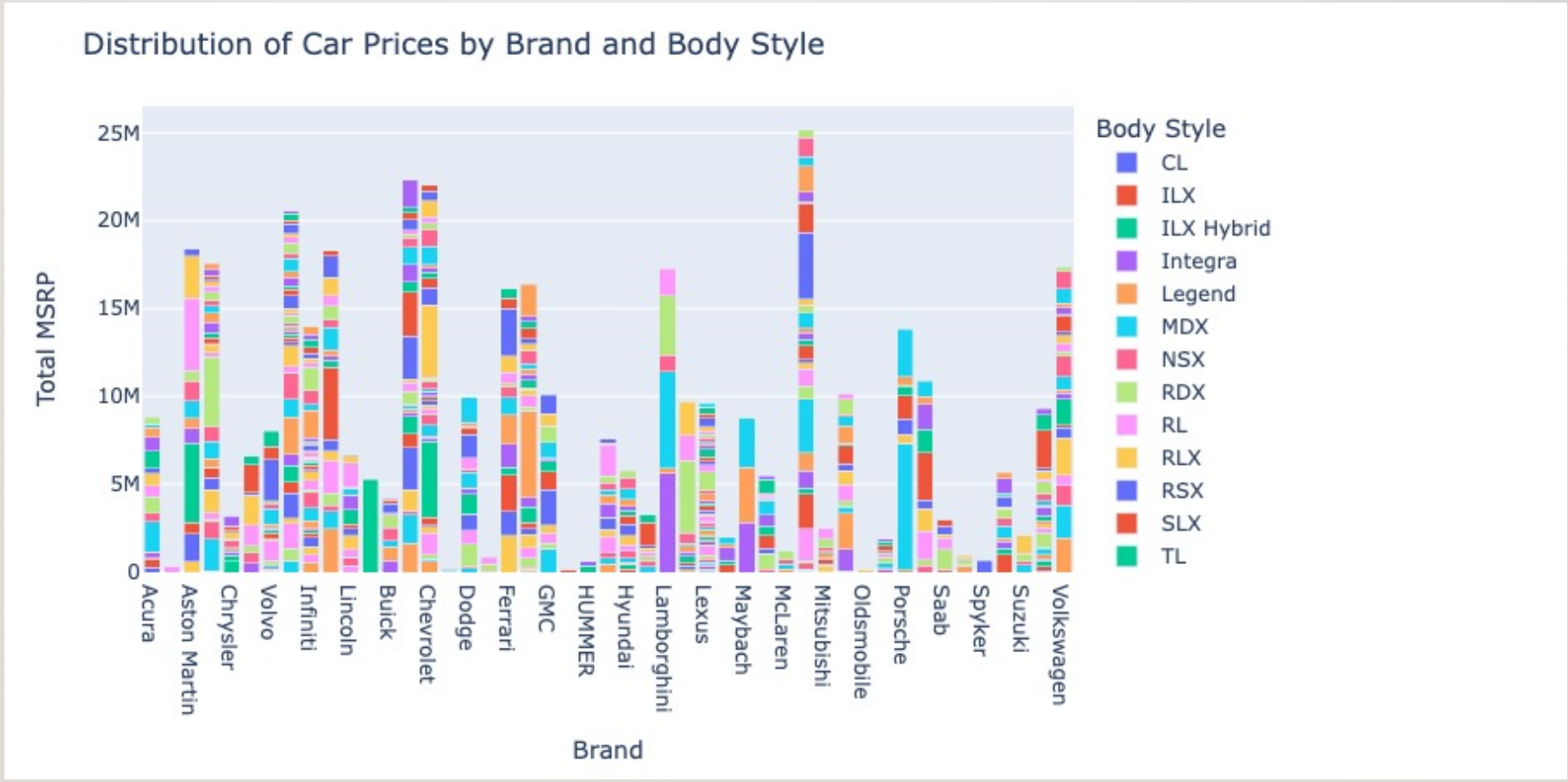


Correlation Coefficient = -0.6032365615288363

Building the Dashboard

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

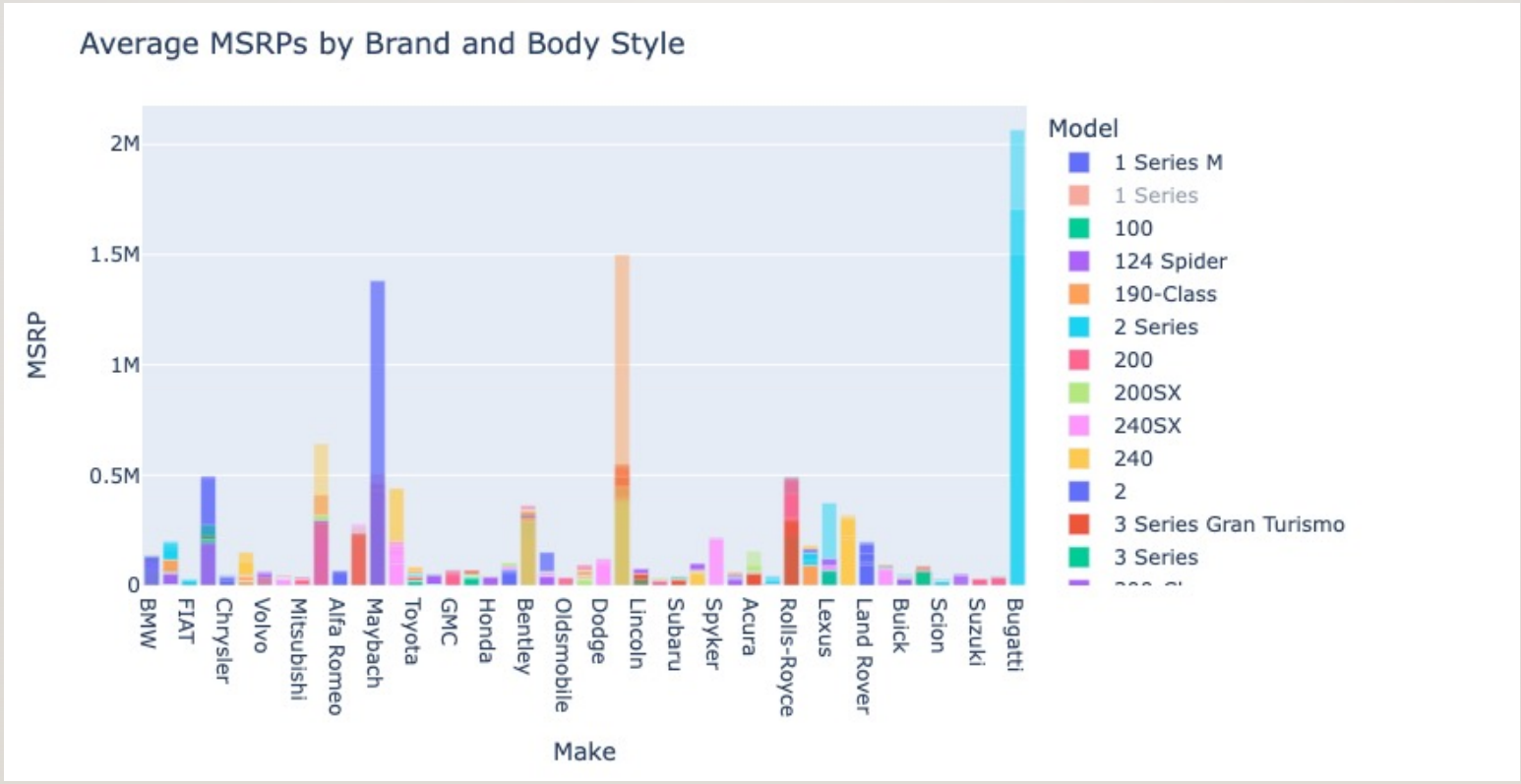
Hints: Stacked column chart to show the distribution of car prices by brand and body style. Use filters and slicers to make the chart interactive. Calculate the total MSRP for each brand and body style using SUMIF or Pivot Tables.



Building the Dashboard

→ **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.

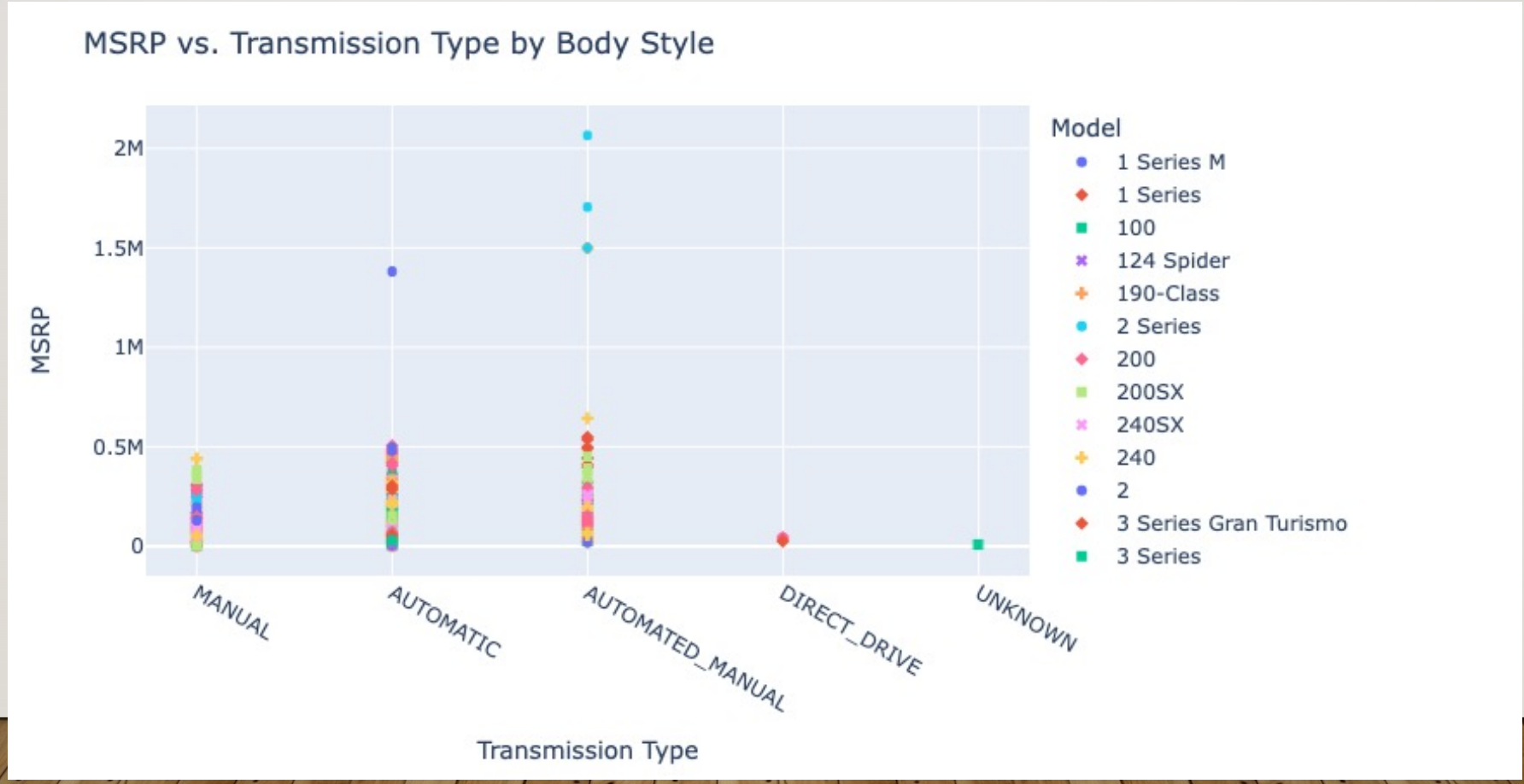


Bugatti has the highest average MSPR

Building the Dashboard

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

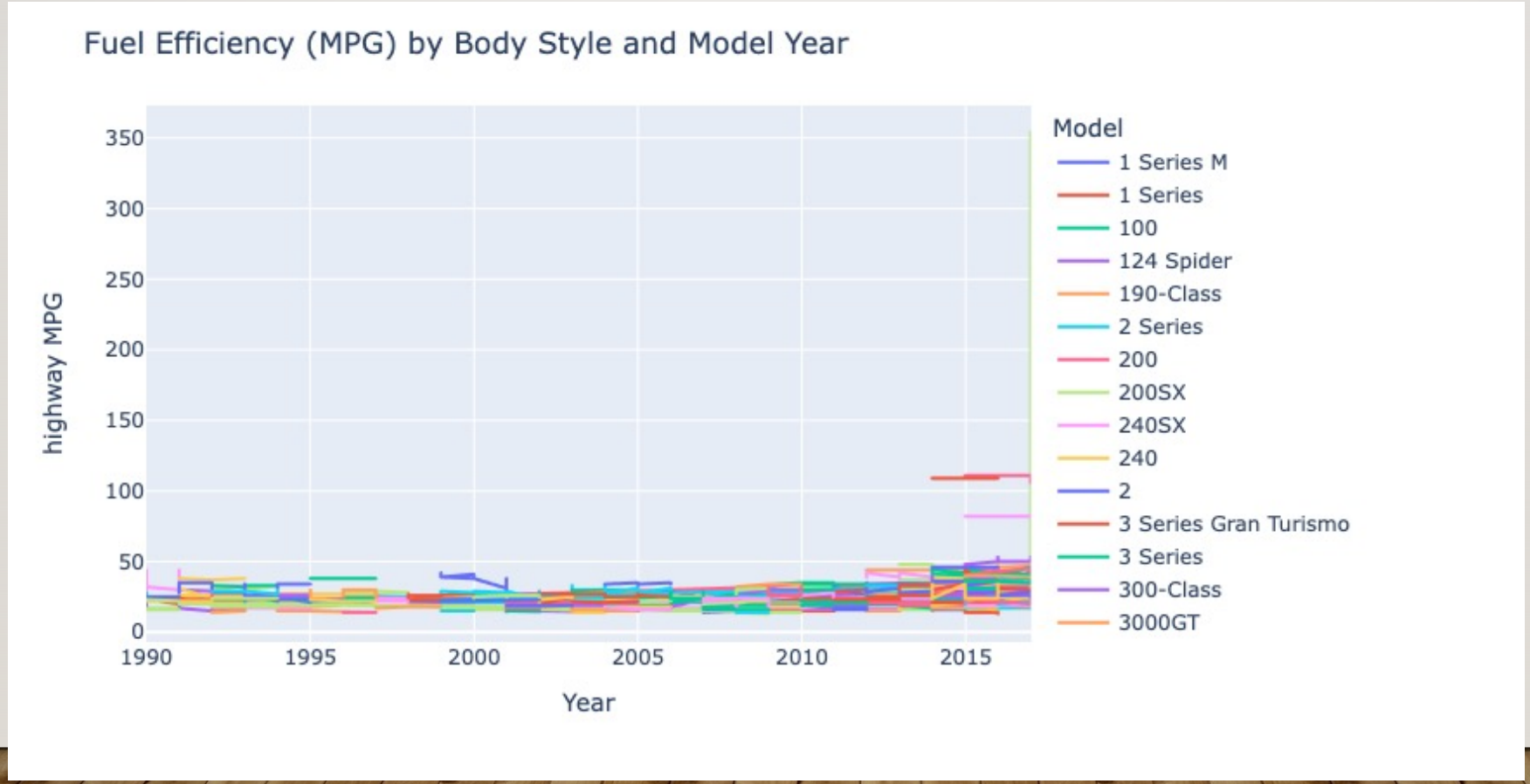
Hints: Scatter plot chart to visualize the relationship between MSRP and transmission type, with different symbols for each body style. Calculate the average MSRP for each combination of transmission type and body style using AVERAGEIFS or Pivot Tables.



Building the Dashboard

→ **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.



Building the Dashboard

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 colors 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.

FOLLOWING THIS IS THE CODE IN PYTHON USED FOR THE ANALYSIS

uwkr3lhbz

May 31, 2023

[29]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[30]:
```
#analysis and cleaning of data
```

[31]:
```
# Load the dataset
df = pd.read_csv('Car_data.csv')
```

[32]:
```
df.describe()
```

[32]:

	Year	Engine HP	Engine Cylinders	Number of Doors	\
count	11914.000000	11845.00000	11884.000000	11908.000000	
mean	2010.384338	249.38607	5.628829	3.436093	
std	7.579740	109.19187	1.780559	0.881315	
min	1990.000000	55.00000	0.000000	2.000000	
25%	2007.000000	170.00000	4.000000	2.000000	
50%	2015.000000	227.00000	6.000000	4.000000	
75%	2016.000000	300.00000	6.000000	4.000000	
max	2017.000000	1001.00000	16.000000	4.000000	

	highway MPG	city mpg	Popularity	MSRP
count	11914.000000	11914.000000	11914.000000	1.191400e+04
mean	26.637485	19.733255	1554.911197	4.059474e+04
std	8.863001	8.987798	1441.855347	6.010910e+04
min	12.000000	7.000000	2.000000	2.000000e+03
25%	22.000000	16.000000	549.000000	2.100000e+04
50%	26.000000	18.000000	1385.000000	2.999500e+04
75%	30.000000	22.000000	2009.000000	4.223125e+04
max	354.000000	137.000000	5657.000000	2.065902e+06

[33]:
```
df.info(verbose = True,null_counts = True)
```

```
<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
```

```
/var/folders/px/544lvycn58z65rdg800zkv9h0000gn/T/ipykernel_56880/3066153830.py:1
: FutureWarning:

null_counts is deprecated. Use show_counts instead
```

[34]:
```
#missing values
df.isnull().values.any()
```

[34]:
```
True
```

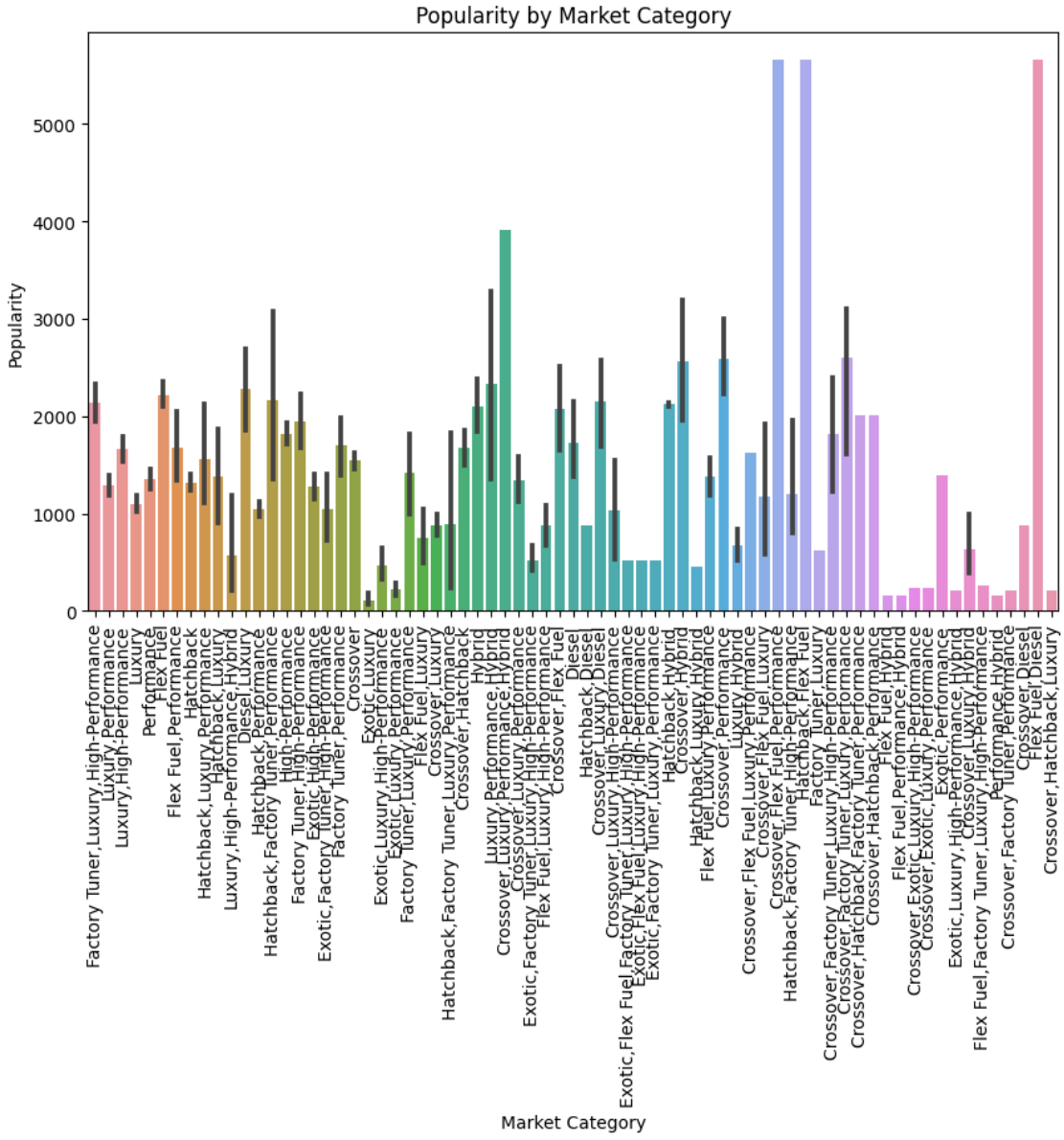
[35]:
```
#Tasks: Analysis
```

[36]:
```
# Task 1.A: Create a pivot table for market category and popularity
pivot_table = df.pivot_table(index='Market Category', values='Popularity',␣
 ↪aggfunc='mean')
print(pivot_table)
```

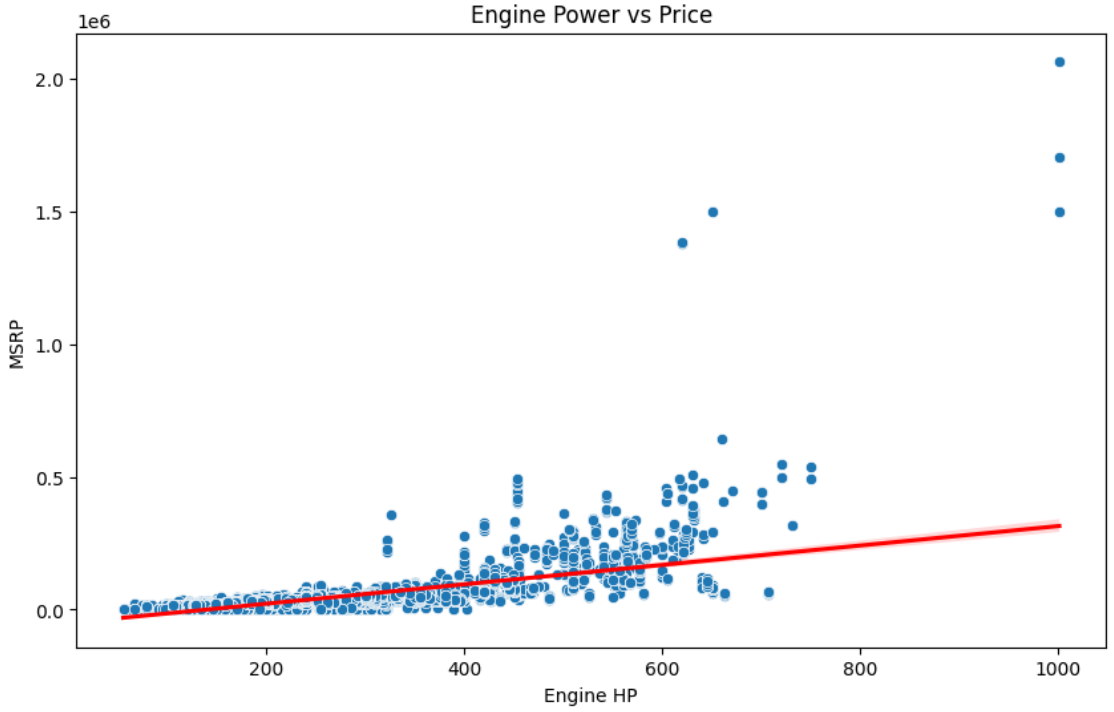
```
                                                Popularity
Market Category
Crossover                                     1545.263063
Crossover,Diesel                               873.000000
Crossover,Exotic,Luxury,High-Performance       238.000000
Crossover,Exotic,Luxury,Performance            238.000000
Crossover,Factory Tuner,Luxury,High-Performance  1823.461538
…                                                     …
Luxury,Hybrid                                  673.634615
Luxury,Performance                            1292.615156
Luxury,Performance,Hybrid                     2333.181818
```

```
Performance                                    1348.873544
Performance,Hybrid                              155.000000

[71 rows x 1 columns]
```

[37]:
```
# Task 1.B: Create a combo chart to visualize market category and popularity
plt.figure(figsize=(10, 6))
sns.barplot(x='Market Category', y='Popularity', data=df)
plt.xticks(rotation=90)
plt.xlabel('Market Category')
plt.ylabel('Popularity')
plt.title('Popularity by Market Category')
plt.show()
```



[38]:
```
# Task 2: Create a scatter plot to visualize the relationship between engine␣
↪power and price
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Engine HP', y='MSRP', data=df)
plt.xlabel('Engine Power')
plt.ylabel('Price')
plt.title('Engine Power vs Price')
# Add a trendline to the scatter plot
sns.regplot(x='Engine HP', y='MSRP', data=df, scatter=False, color='red')

plt.show()
```



[39]:
```
# Task 3: Use regression analysis to identify variables with the strongest␣
↪relationship with price
from sklearn.linear_model import LinearRegression
```

[40]:
```
df.dropna(inplace=True)
```

[41]:
```
# Select relevant columns for regression analysis
X = df[['Engine HP', 'Engine Cylinders', 'Number of Doors', 'highway MPG',␣
↪'city mpg']]
```

```
y = df['MSRP']
```

[42]:
```
# Fit the linear regression model
regression_model = LinearRegression()
regression_model.fit(X, y)

# Get the coefficients and variable names
coefficients = regression_model.coef_
variable_names = X.columns
```

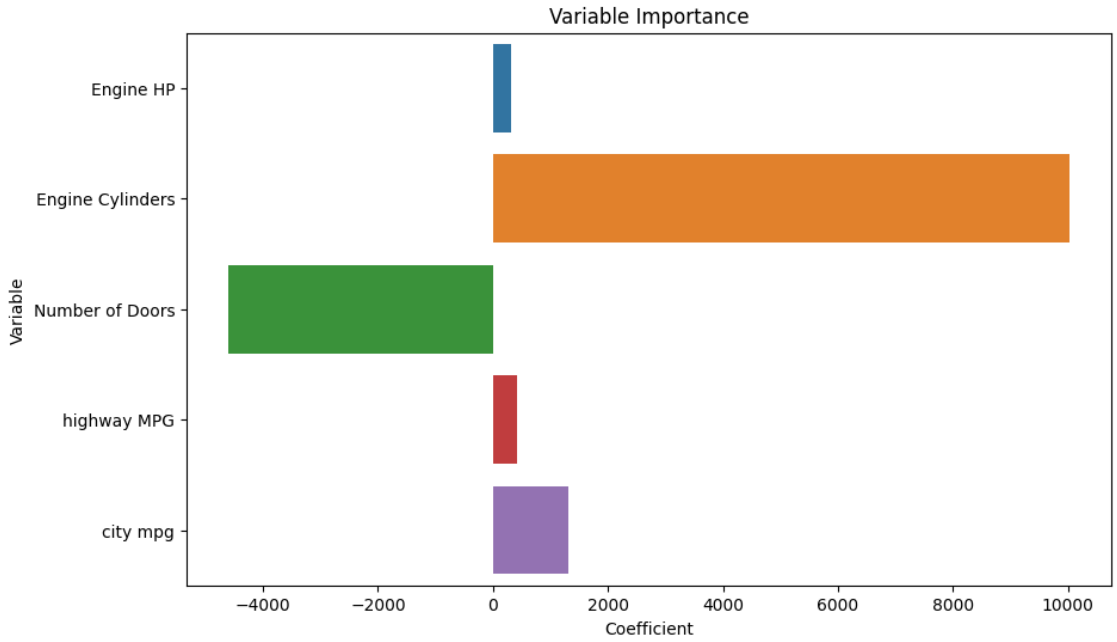
[43]:
```
from sklearn.impute import SimpleImputer

# Impute missing values with mean
imputer = SimpleImputer(strategy='mean')
X_imputed = imputer.fit_transform(X)

regression_model = LinearRegression()
regression_model.fit(X_imputed, y)
```

[43]:
```
LinearRegression()
```

[44]:
```
from sklearn.experimental import enable_hist_gradient_boosting
from sklearn.ensemble import HistGradientBoostingRegressor

# Perform regression analysis with HistGradientBoostingRegressor
regression_model = HistGradientBoostingRegressor()
regression_model.fit(X, y)
```

[44]:
```
HistGradientBoostingRegressor()
```

[45]:
```
# Create a bar chart to visualize the coefficient values
plt.figure(figsize=(10, 6))
sns.barplot(x=coefficients, y=variable_names)
plt.xlabel('Coefficient')
plt.ylabel('Variable')
plt.title('Variable Importance')
plt.show()
```



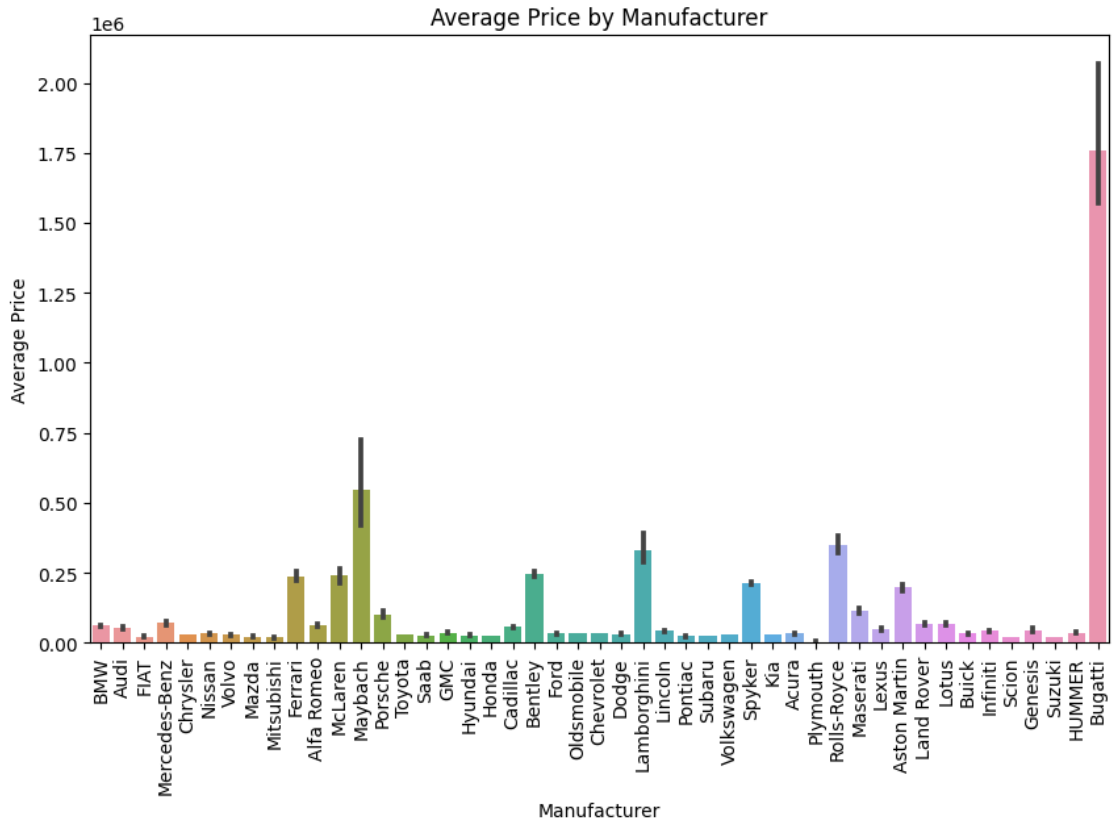
[51]:
```
# Task 4.A: Create a pivot table for average price by manufacturer
pivot_table = df.pivot_table(index='Make', values='MSRP', aggfunc='mean')
pivot_table
```

[51]:

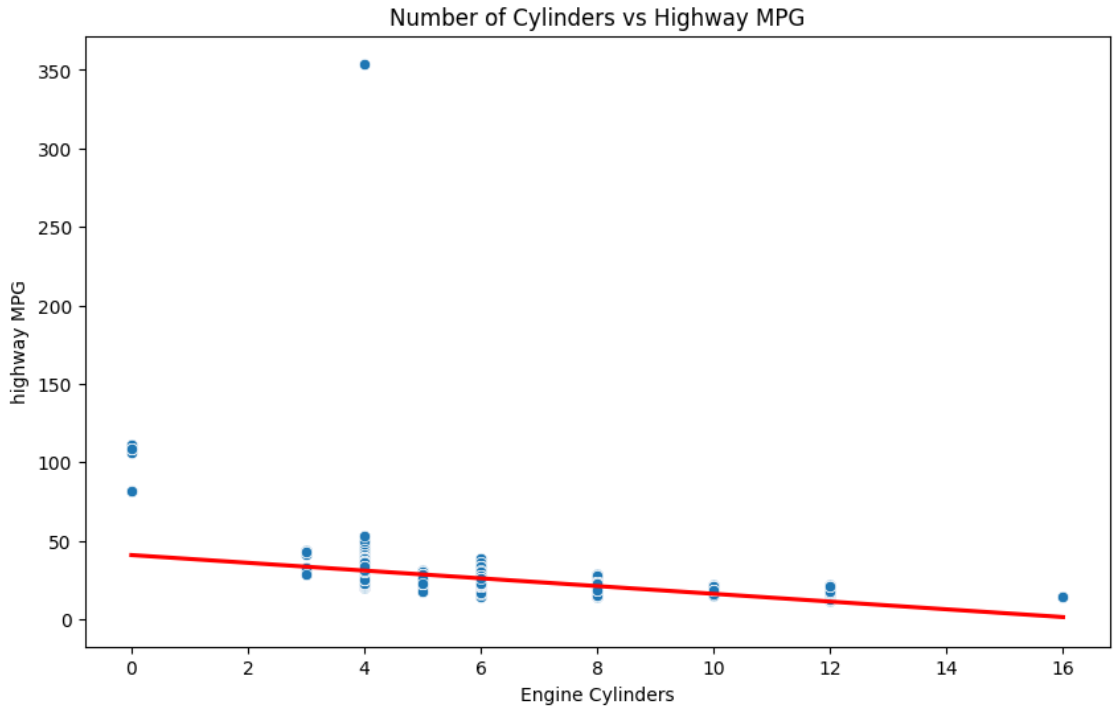
Make	MSRP
Acura	3.488759e+04
Alfa Romeo	6.160000e+04
Aston Martin	1.979104e+05
Audi	5.345211e+04
BMW	6.154676e+04
Bentley	2.471693e+05
Bugatti	1.757224e+06
Buick	3.377040e+04
Cadillac	5.623132e+04
Chevrolet	3.583328e+04
Chrysler	2.997887e+04
Dodge	3.099538e+04
FIAT	2.237066e+04
Ferrari	2.373838e+05
Ford	3.324537e+04
GMC	3.738575e+04
Genesis	4.661667e+04
HUMMER	3.646441e+04
Honda	2.695689e+04
Hyundai	2.698596e+04

```
Infiniti        4.239421e+04
Kia             3.014931e+04
Lamborghini     3.315673e+05
Land Rover      6.782322e+04
Lexus           4.754907e+04
Lincoln         4.249437e+04
Lotus           6.918828e+04
Maserati        1.142077e+05
Maybach         5.462219e+05
Mazda           2.298335e+04
McLaren         2.398050e+05
Mercedes-Benz   7.153781e+04
Mitsubishi      2.022188e+04
Nissan          3.290842e+04
Oldsmobile      3.486800e+04
Plymouth        4.076821e+03
Pontiac         2.472813e+04
Porsche         1.016224e+05
Rolls-Royce     3.511306e+05
Saab            2.741350e+04
Scion           2.039594e+04
Spyker          2.133233e+05
Subaru          2.640722e+04
Suzuki          2.115305e+04
Toyota          3.104422e+04
Volkswagen      2.990867e+04
Volvo           2.854116e+04
```

[19]:
```python
# Task 4.B: Create a bar chart to visualize average price by manufacturer
plt.figure(figsize=(10, 6))
sns.barplot(x='Make', y='MSRP', data=df)
plt.xticks(rotation=90)
plt.xlabel('Manufacturer')
plt.ylabel('Average Price')
plt.title('Average Price by Manufacturer')
plt.show()
```



[46]:
```
# Task 5.A: Create a scatter plot to visualize the relationship between number␣
 ↪of cylinders and highway MPG
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Engine Cylinders', y='highway MPG', data=df)
plt.xlabel('Number of Cylinders')
plt.ylabel('Highway MPG')
plt.title('Number of Cylinders vs Highway MPG')
# Add a trendline to the scatter plot
sns.regplot(x='Engine Cylinders', y='highway MPG', data=df, scatter=False,␣
 ↪color='red')

plt.show()
```



[21]:
```
# Task 5.B: Calculate the correlation coefficient between number of cylinders
 ↪and highway MPG
correlation = df['Engine Cylinders'].corr(df['highway MPG'])
print(f"Correlation coefficient: {correlation}")
```

```
Correlation coefficient: -0.6032365615288363
```

[22]:
```
#Building the Dashboard
```

[23]:
```
import plotly.express as px
```

[55]:
```
# Task 1: Distribution of car prices by brand and body style
fig_task1 = px.histogram(df, x='MSRP', color='Make', barmode='stack')
fig_task1.update_layout(title='Distribution of Car Prices by Brand and Body
 ↪Style')
```

[56]:
```
# Calculate the total MSRP for each brand and body style
df_total_msrp = df.groupby(['Make', 'Model'])['MSRP'].sum().reset_index()

# Create the stacked column chart
fig = px.bar(df_total_msrp, x='Make', y='MSRP', color='Model', barmode='stack')

# Customize the chart layout
fig.update_layout(
```

```
    title='Distribution of Car Prices by Brand and Body Style',
    xaxis_title='Brand',
    yaxis_title='Total MSRP',
    legend_title='Body Style'
)

# Show the chart
fig.show()
```

[50]:
```
# Task 2: Average MSRPs by brand and body style
fig_task2 = px.bar(df, x='Make', y='MSRP', color='Model', barmode='overlay')
fig_task2.update_layout(title='Average MSRPs by Brand and Body Style')
```

[26]:
```
# Task 3: MSRP vs. Transmission type by body style
fig_task3 = px.scatter(df, x='Transmission Type', y='MSRP', color='Model',␣
 ↪symbol='Model')
fig_task3.update_layout(title='MSRP vs. Transmission Type by Body Style')
```

[27]:
```
# Task 4: Fuel efficiency (MPG) by body style and model year
fig_task4 = px.line(df, x='Year', y='highway MPG', color='Model')
fig_task4.update_layout(title='Fuel Efficiency (MPG) by Body Style and Model␣
 ↪Year')
```

[28]:
```
# Task 5: Horsepower, MPG, and Price by brand
fig_task5 = px.scatter(df, x='Engine HP', y='highway MPG', size='MSRP',␣
 ↪color='Make', hover_name='Model')
fig_task5.update_layout(title='Horsepower, MPG, and Price by Brand')
```