Impact of Car features on price and profitability

Project Description:

The client seeks insights into how a car manufacturer can strategically price its vehicles and develop new products to meet consumer demands effectively. Specifically, they want to understand the relationship between car features, market category, pricing and profitability.

Data collection and Understanding: obtain the dataset and explore it to understand its structure, features and datatypes. Identify the target variable and distribution of each feature to gain insights into the data.

Data cleaning and exploration: cleaning dataset by handling missing values, removing duplicates, and addressing any inconsistencies. Conduct EDA to understand distribution of variables, identify patterns and detect outliers.

Statistical Analysis: Conduct statistical tests to assess the significance of relationships between variables and MSRP. use measures such as correlation coefficients to quantify relationships and identify significant factors.

Product Development Recommendations: Based on the analysis results, recommendations will be made regarding which features to prioritize in future product development efforts. This will help the manufacturer align their offerings with consumer preferences and market trends.

Interpretation and Reporting: interpret the findings from EDA and model results to identify factors influencing popularity and MSRP. Provide actionable insights and recommendation to the Car company for making decisions about car features to drive profitability.

Tech Stack use:

Jupiter notebook- its is versatile tool that is widely used for data science, machine learning, scientific computing, and education. It provides an intuitive and interactive environment for exploring data, prototype code and communicating results.

Tableau- Business Intelligence tool to visualize and analyse data to create interactive and insightful visualizations.

Github

https://github.com/radhikanagaraj/Data-analyst-projects/tree/Python-car-feature-analysis

Load the data into the notebook and **Identify missing data and deal with it appropriately**
First step is to remove duplicates, there were **715 duplicates** found and removed
check for count of missing values in each column of the dataframe.
The Columns **Engine fuel type, Engine Hp, Engine Cylinders and Number of doors** have 3, 69, 30, 6 missing values missing values respectively.

```
df=pd.read_csv('/kaggle/input/car-data/car_data.csv') #load the dataset into the notebook
df.drop_duplicates(inplace=True) # remove duplicate entries
print(df.isnull().sum()) # find the count of null values in each column
```

```
Make                  0
Model                 0
Year                  0
Engine Fuel Type      3
Engine HP            69
Engine Cylinders     30
Transmission Type     0
Driven_Wheels         0
Number of Doors       6
Market Category    3376
Vehicle Size          0
Vehicle Style         0
highway MPG           0
city mpg              0
Popularity            0
MSRP                  0
dtype: int64
```

Market Category has 3376 N/A values and it is replaced by most common category for each make type by using Mode method

```
#Market category has around 3000 null values and it is replaced by most common category for
#mode of Market category for each Make
m=df.groupby('Make')['Market Category'].transform(lambda x: x.mode().iloc[0])
#print(m)

#filling the missing values by the above value
df.fillna({'Market Category':m},inplace=True)
```

we check for **Number of doors** with corresponding column for ‘Make’ of the vehicles. by filtering we found out the missing values are for ‘Model S’ and ‘FF’ and by checking the value for ’No of Doors’ for other rows of the particular model we replace the same for same kind of vehicles.

Engine Hp has 69 missing values, we check for Engine HPs of particular Make and for the electric vehicles the HP is replaced by 0 and for the rest of the empty values we use mean method to replace them.

```
def hp(a):

    if a['Engine Fuel Type']=='electric':
        return 0
    else:
        return a['Engine HP']


def fi(a):

    if a['Model']=='Model S':
        return 4
    elif a['Model']=='FF':
        return 2
    else:
        return a['Number of Doors']

df.fillna({'Engine HP':df.apply(hp,axis=1),'Number of Doors':df.apply(fi,axis=1),'Engine Fue

#e=df.groupby('Model')['Engine HP'].mean() would still give correct ans but inorder to fill
e=df.groupby('Model')['Engine HP'].transform('mean')
df['Engine HP'].fillna(e,inplace=True)
print(df.isnull().sum()) # check for count of null values after impuation in  each columns
```

Engine fuel type has 3 empty values for SUZUKI VERNOA 2004 MODEL by filtering and checking we can replace the value with regular unleaded as rest of the rows have same type.

Engine cylinders have 30 missing values and their corresponding Fuel type are electric for which we can replace it with 0 as electric vehicles have no cylinders and for the rest of the missing values the corresponding make and model is mazda RX7 and RX8 which do not have cylinders hence the missing values are all replaced with 0

```
ıxis=1),'Engine Fuel Type':'regular unleaded','Engine Cylinders':0},inplace=True)
```

Column **Transmission Type** has few UNKNOWN values and we use mode method to replace them.

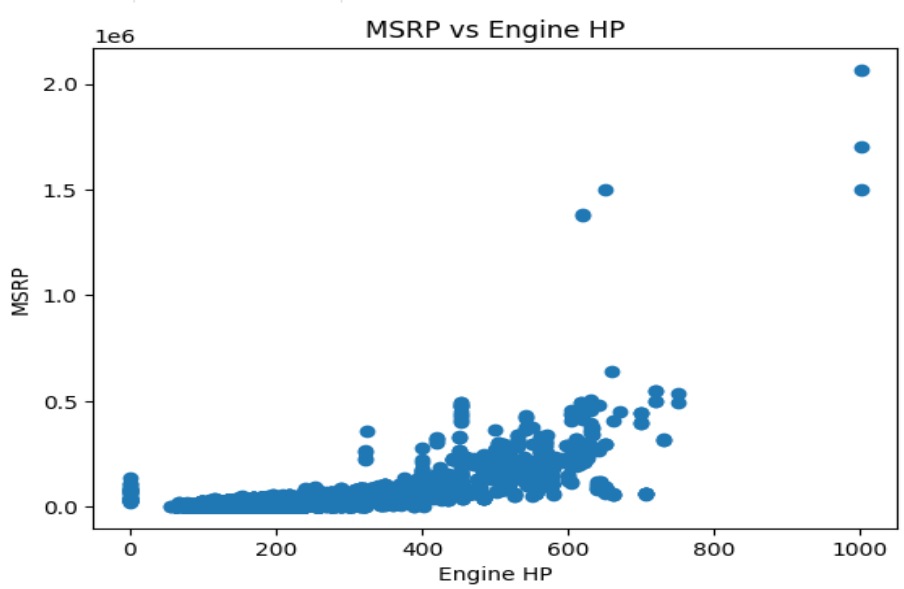
```
#check for count of unique entries in each column to identify any incorrect entries
for i in df.columns:
    print(df[i].value_counts())
#Find mode of transmission type for each Make type
    b=df.groupby('Make')['Transmission Type'].transform(lambda x: x.mode().iloc[0])
#replace "Unknown" with mode of each Make type
df.loc[df['Transmission Type'] == 'UNKNOWN', 'Transmission Type'] = b
#print(df['Transmission Type'].value_counts())
```

Task 1 : What is the relationship between a car's engine power and its price?

The relationship between MSRP and Engine HP is linear in positive direction they have correlation coefficient of **0.65**, higher the Engine HP higher the price. As pricing depends on Engine HP, manufacturer can focus on putting engines with lower HP to target budget conscious consumers, whereas higher HP engines can be placed into high end vehicles attracting luxury consumers.

```
#relationship between a car's engine power and its price
import matplotlib.pyplot as plt
print( df[['Engine HP', 'MSRP']].corr()) #Correlation coefficient of Engine HP and MSRP
plt.scatter(df['Engine HP'],df['MSRP']) #scatter plot to vizualize correlation

# Add labels and title
plt.xlabel('Engine HP')
plt.ylabel('MSRP')
plt.title('MSRP vs Engine HP')

# Show plot
plt.show()
```

```
           Engine HP      MSRP
Engine HP   1.000000  0.650203
MSRP        0.650203  1.000000
```

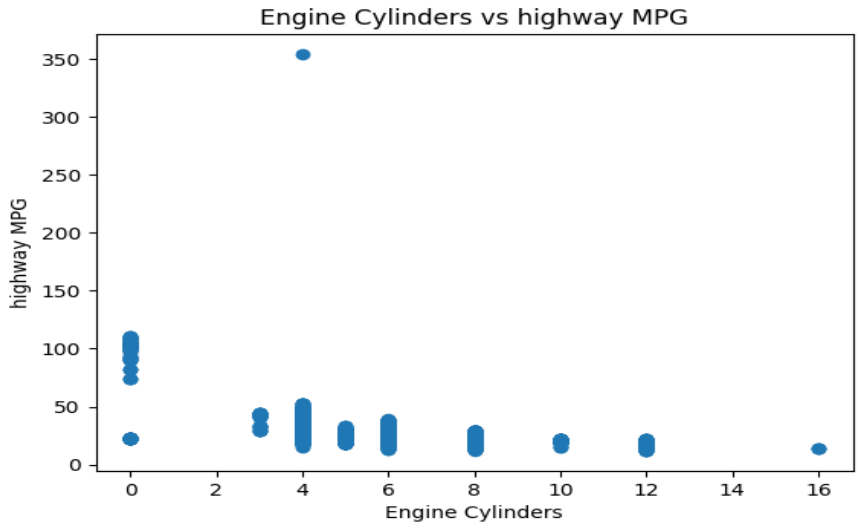


Task 2: Relationship between Fuel efficiency and the number of cylinders

The relationship between no of cylinders and Highway MPG can be seen clearly with increasing no of cylinders the mpg does not really show any increase, in-fact 0 cylinders have the highest MPG means electric cars serves the highest speed on highway. The correlation coefficient is **-0.61034** they are negatively correlated. No of cylinders can affect the car's performance which is crucial for any buyer, consumer need high MPG hence decreasing the no of cylinders can lower price as seen earlier but giving the highest MPG which consumers prefer.

```
print( df[['Engine Cylinders','highway MPG']].corr())
plt.scatter(df['Engine Cylinders'],df['highway MPG'])

plt.xlabel('Engine Cylinders')
plt.ylabel('highway MPG')
plt.title('Engine Cylinders vs highway MPG')
```

```
                  Engine Cylinders  highway MPG
Engine Cylinders          1.000000    -0.610338
highway MPG              -0.610338     1.000000
```

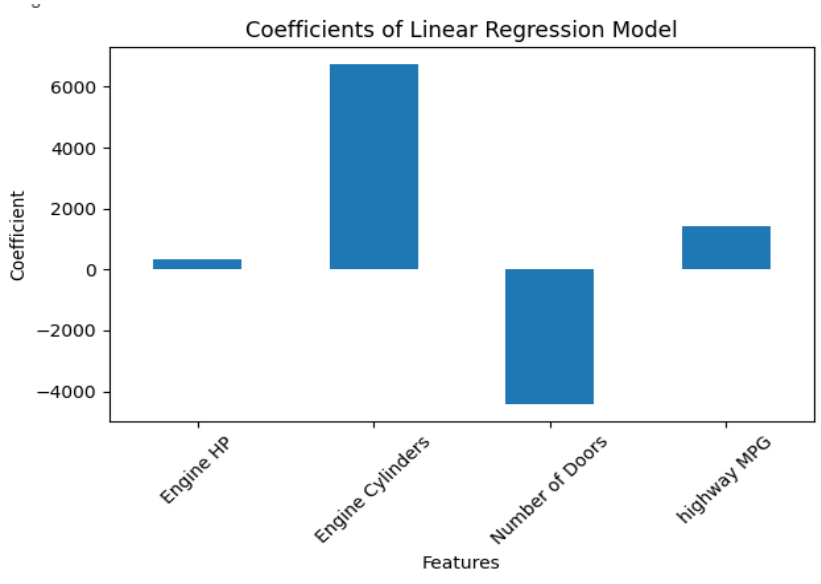


Task 3: which car features are most important in determining a cars price?

The weak relationship is with No of doors and the relation is negative, with every increase of door the MSRP decreases by 5687, this can be taken into consideration for **pricing strategy** to keep the doors for the car accordingly as per the budget.

```
#import necessary libraries to perform regression analysis
from scipy import stats
from scipy.stats import linregress
print( df[['Engine Cylinders','MSRP']].corr())
x = df['Engine Cylinders']
y = df['MSRP']
slope, intercept, r_value, p_value, std_err = linregress(x, y)


plt.figure(figsize=(8, 6))
plt.scatter(x, y, color='blue', alpha=0.5, label='Data Points')

plt.plot(x, slope * x + intercept, color='red', label='Trendline')
# Add correlation coefficient to the plot
plt.text(x.min(), y.max(), f'Correlation coefficient: {r_value:.2f}', ha='left', va='top')
plt.xlabel('Engine Cylinders')
plt.ylabel('MSRP')
plt.title('Correlation Chart with Trendline')
plt.legend()
plt.show()
```

	Coefficient
Engine HP	318.020613
Engine Cylinders	6734.674386
Number of Doors	-4423.385047
highway MPG	1417.035299



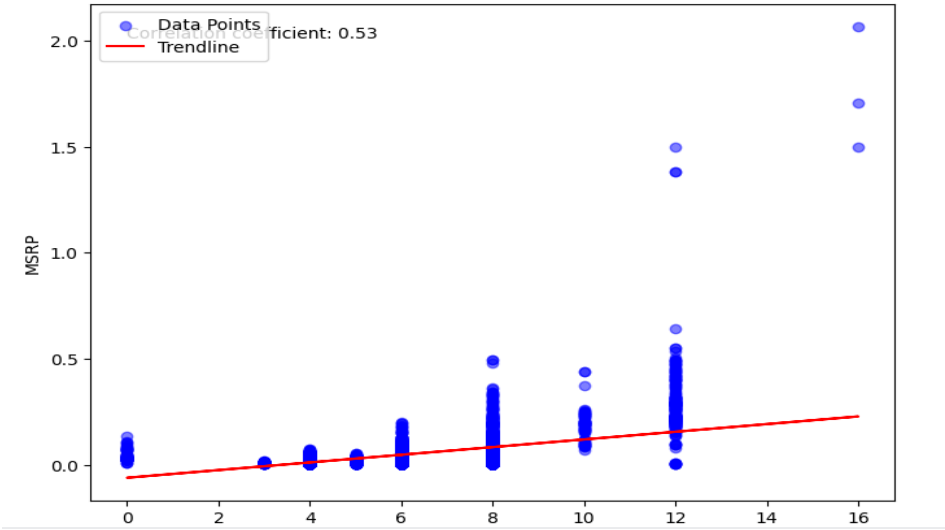
The strongest relationship of MSRP is with no of cylinders, we can observe the collinear coefficient between No of cylinders and MSRP was **0.53**

with increasing no of cylinders the price in the market sure increases. Here the trend is same as for Engine HP vs MSRP, this is because having high no of cylinders means complex internal design hence the price logically has to be high. In order to price the vehicle the manufacturer

has to think about no of cylinders that's going into the vehicle. We can see this in scatter plot as follows

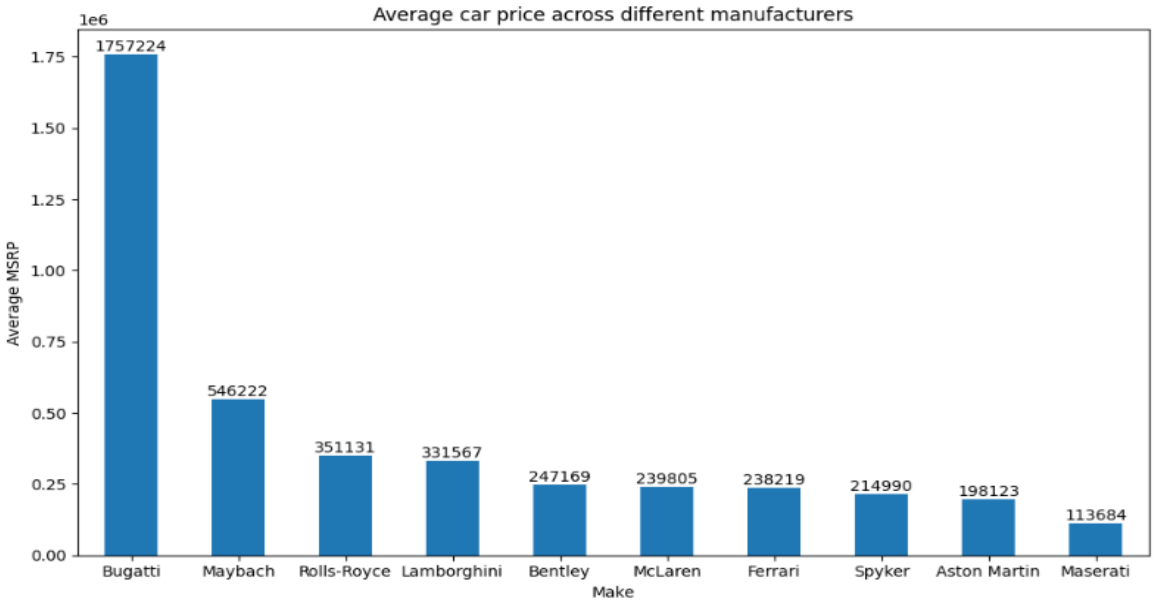
```
from scipy import stats
from scipy.stats import linregress
print( df[['Engine Cylinders','MSRP']].corr())
x = df['Engine Cylinders']
y = df['MSRP']
slope, intercept, r_value, p_value, std_err = linregress(x, y)


plt.figure(figsize=(8, 6))
plt.scatter(x, y, color='blue', alpha=0.5, label='Data Points')

plt.plot(x, slope * x + intercept, color='red', label='Trendline')
# Add correlation coefficient to the plot
plt.text(x.min(), y.max(), f'Correlation coefficient: {r_value:.2f}', ha='left', va='top')
plt.xlabel('Engine Cylinders')
plt.ylabel('MSRP')
plt.title('Correlation Chart with Trendline')
plt.legend()
plt.show()
```

```
                  Engine Cylinders      MSRP
Engine Cylinders          1.000000  0.533431
MSRP                      0.533431  1.000000
```



Task 4: How does the average price of a car vary across different manufacturers?

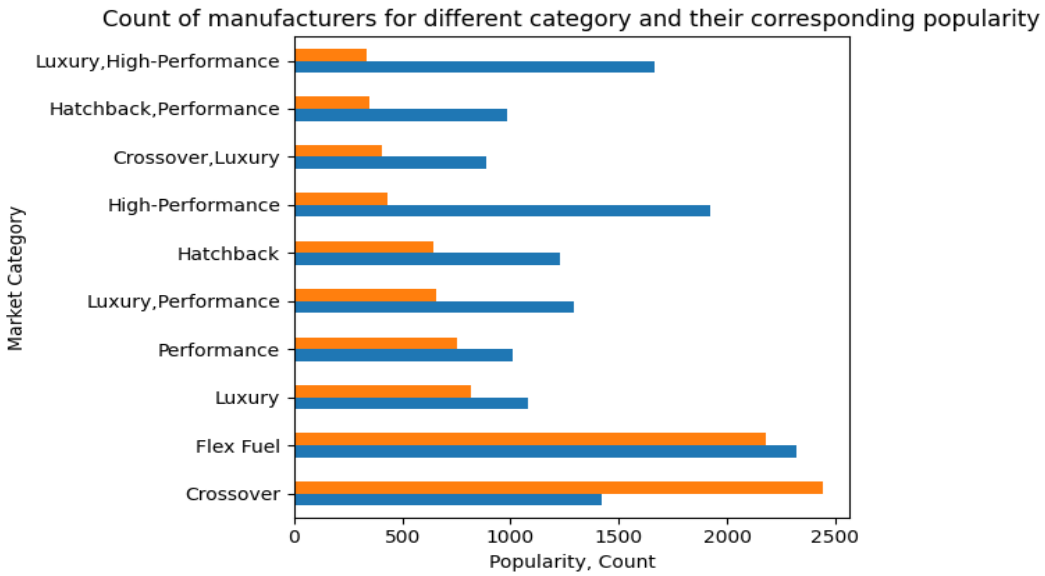
```
a=df.groupby('Make')['MSRP'].mean()

plt.figure(figsize=(10, 6))

for i, v in enumerate(a.nlargest(10)):
    plt.text(i, v, f'{v:.0f}', ha='center', va='bottom')
a.nlargest(10).plot(kind='bar', legend=None)
plt.title('Average car price across different manufacturers')
plt.xlabel('Make')
plt.ylabel('Average MSRP')
plt.xticks(rotation=0)
plt.tight_layout()
plt.show()
```



Brand Bugatti has the highest market price followed by Maybach and Rolls-Royce and these vehicles belong to luxury category, which can indicate the brand focuses only on rich consumers looking for luxury vehicles having lower customer count. New manufacturers should explore variety of market trends and consumer preferences to adjust pricing and product improvements.

Task 5 : How does the popularity of a car model vary across different market categories?

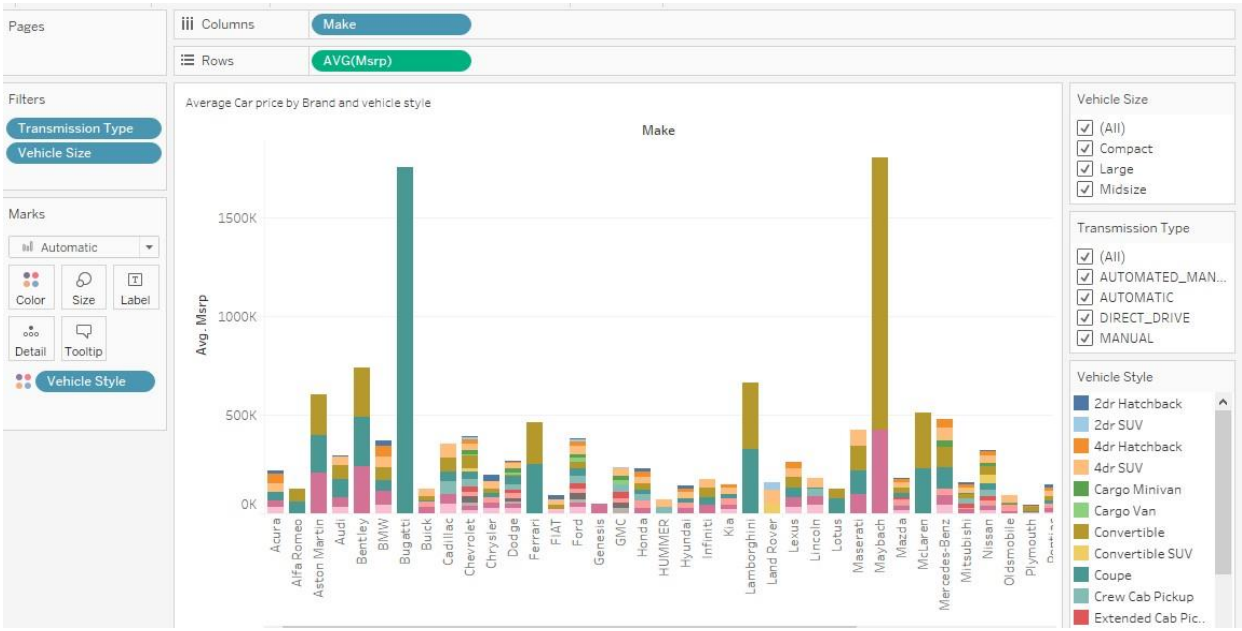
```
b=df.groupby('Market Category')['Popularity'].agg(['mean','count'])
print(b)
top10=b.nlargest(10,'count')
print(top10)
for i, v in enumerate(top10):
    if isinstance(v, (int, float)):  # Check if v is a numerical value
        plt.text(i, v, f'{v:.0f}', ha='center', va='bottom')
top10.plot(kind='barh', legend=None)
plt.title('Count of manufacturers for different category and their corresponding popularity'
plt.xlabel('Popularity, Count')
plt.ylabel('Market Category')
plt.xticks(rotation=0)
plt.tight_layout()
plt.show()
```



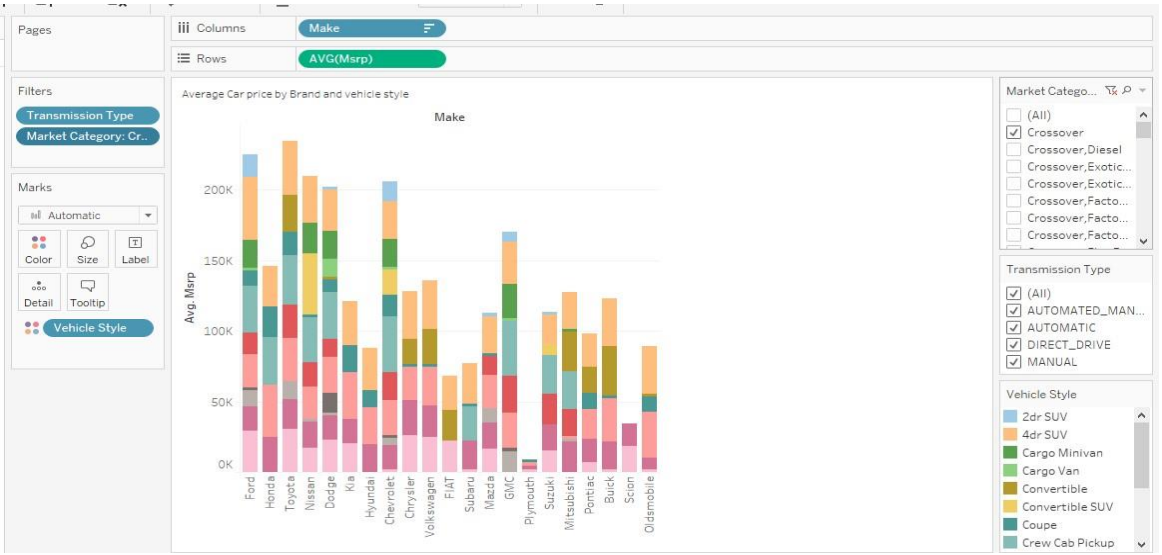
I have filtered top 10 results by highest average MSRP, crossover category has the highest MSRP followed by Flex=fuel Luxury and the popularity is high for the category Flex-Fuel, Luxury, High-performance followed by crossover. There is an inverse relation between popularity and the model count to few market category. For Luxury, High-performance the count is low but the popularity is higher indicates particular group of consumers prefer luxury vehicles while majority people prefer budget-friendly medium level cars.

Dashboard Creation

Task1: How does the distribution of car prices vary by brand and body style?



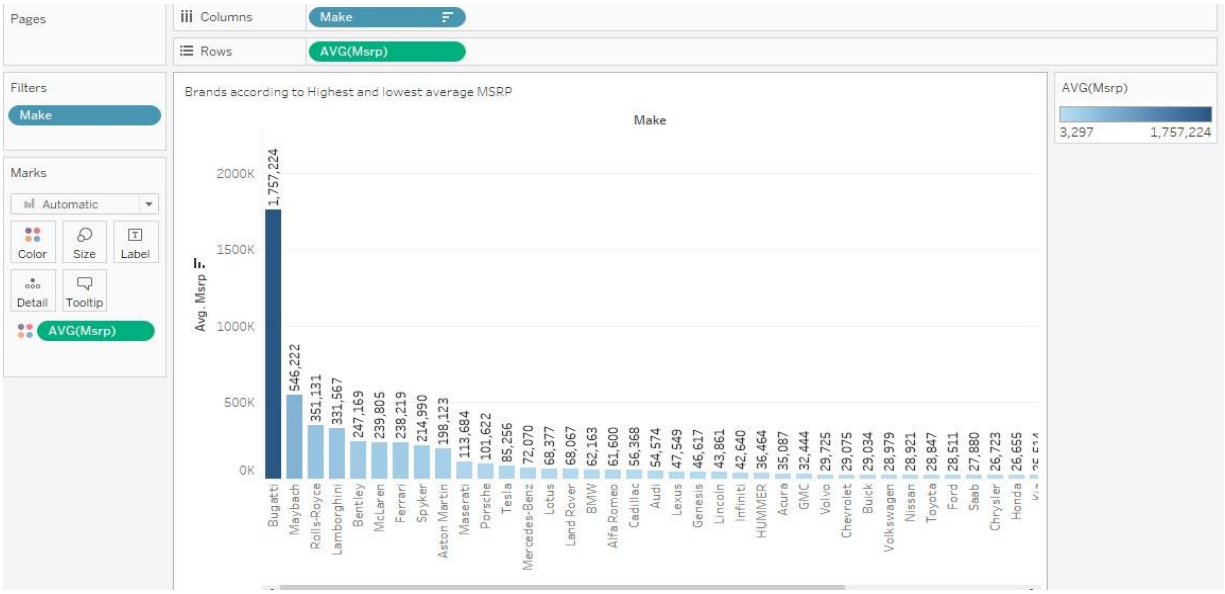
The Brand Bugatti has the highest price which is coming from the Coupe body style followed by Maybach which as two kinds of body style and convertible body style has higher price than sedan



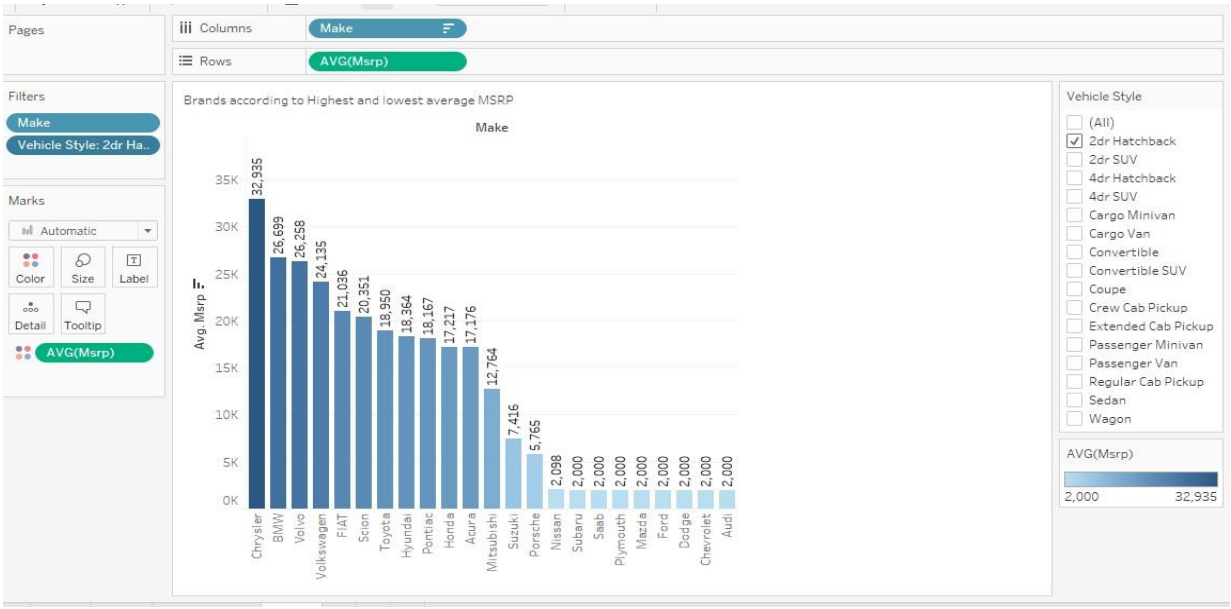
and by filtering the Market Category to a particular type the pricing varies accordingly. For category Crossover the highest price is for Toyota followed by Ford.

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

Like seen earlier Bugatti has the highest average MSRP of 1757224 and Plymouth has the lowest average of 3297.

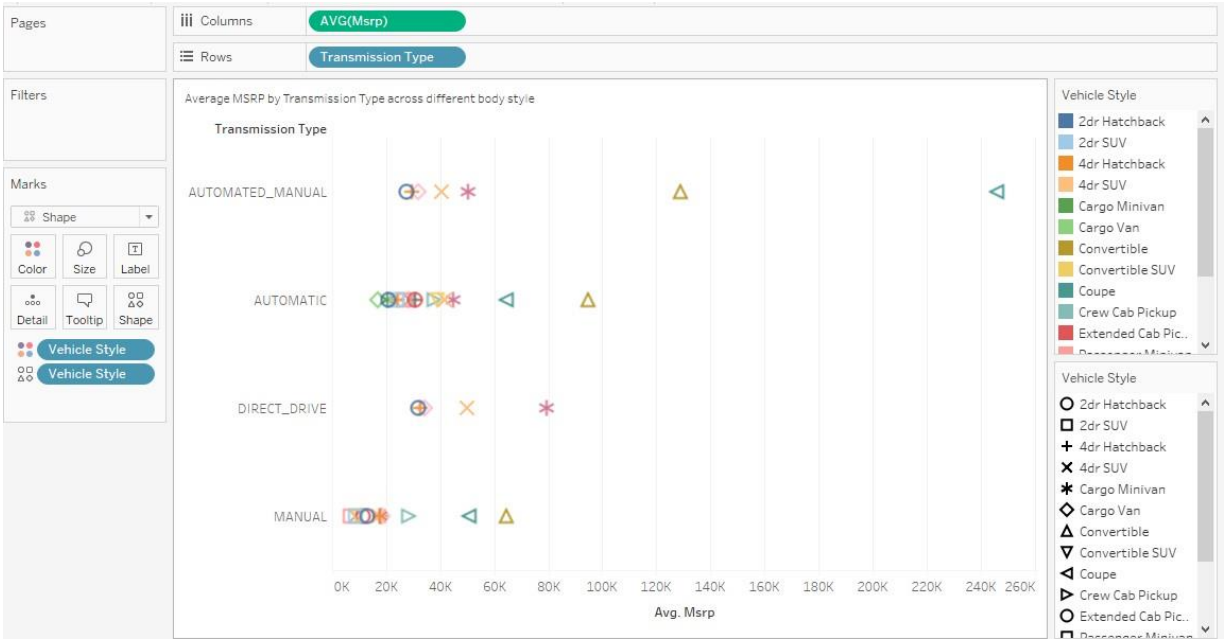


When we filter for only 2dr Hatchback, Brand crysler has the highest while audi the lowest.

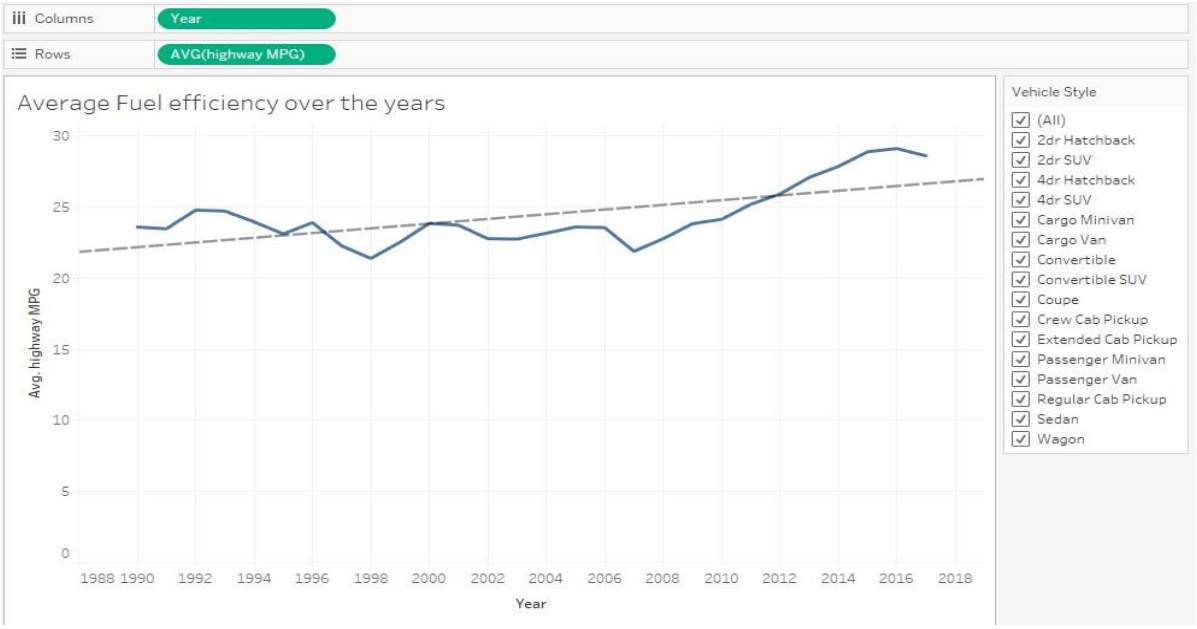


Task 3: How do the different feature such as transmission type affect the MSRP, and how does this vary by body style?

Automatic has the highest MSRP, while direct drive has the lowest, and majority of vehicle styles have automatic transmission type followed by manual, and direct drive with least vehicle styles.

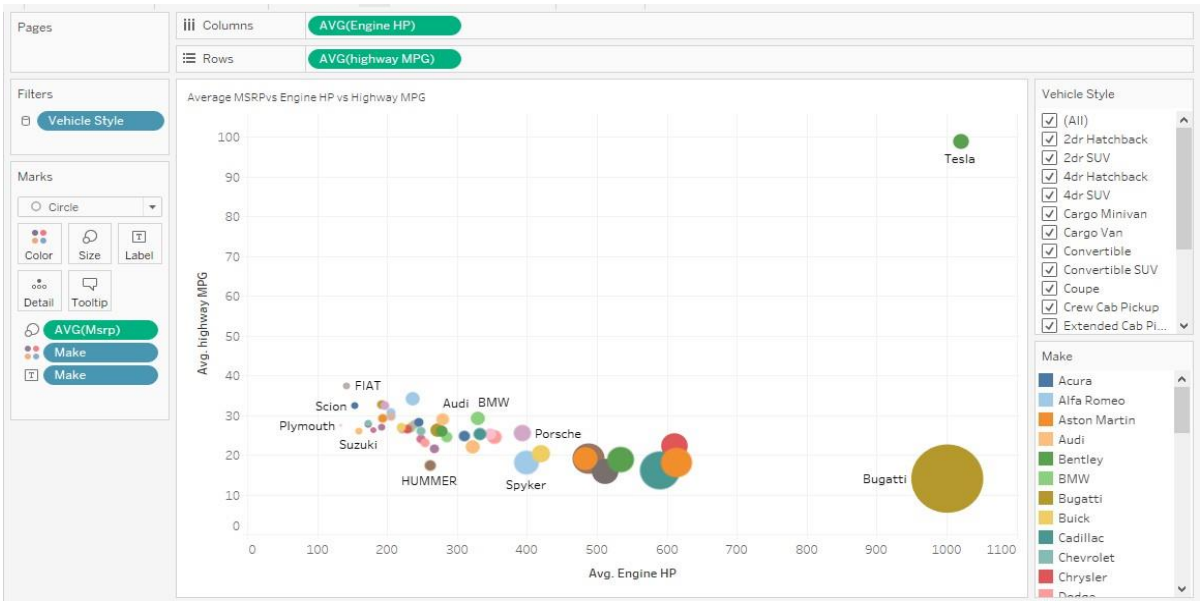


Task 4: How does the fuel efficiency of cars vary across different body styles and model years?



The Highway MPG was certain years till 2006 as been consistent with little ups and downs, after 2006 the MPG has gradually increased till 2016. With varying body styles the MPG varies.

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



The relationship between average HP an average MPG is negative correlation with increasing HP MPG decreases, but the engine HP is positively correlated with average MSRP with increasing Engine HP MSRP increases as well which we can see with the size of the bubbles.

Result:

Understanding consumer preferences: through analysing, we can gain insights into consumer preferences regarding car features, fuel efficiency, and pricing, understanding

what features are most popular among consumers can inform product development decisions and help tailor offerings to meet market demand.

Optimizing pricing strategies: by observing the relationship between car features, like Engine HP, Cylinders with MSRP the manufacturer can develop more effective pricing strategies by identifying features that drive pricing, ultimately maximizing profitability.

Product development: Insights gained from the analysis can help the manufacturer improve product development. By understanding market trends and consumer preferences, they can develop products that better align with customer needs.

Data analysis skills: project offers an opportunity to gain experience in data cleaning, imputation, feature engineering, regression analysis, and predictive modelling techniques.

Effective communication of insights: Communicating the findings of the analysis effectively is crucial. Both the manufacturer and the data analyst learn how to present complex analytical results in a clear and understandable manner, using visualizations, reports and presentations to convey key insights and recommendations.