**Analyzing the Impact of Car Features on Price and Profitability**

The aim of this project is to gain the insight about how the certain features are impacting the sells and profit of company. The data we get from the source is as follows –

* **Project\_7\_Dataset.xlsx**

Certain Information about the dataset.

**Features in the Columns:**

* Make- The company that has made the Vehicle.
* Model - Model of the vehicle.
* Year - Manufactured year.
* Engine Fuel Type – Fuel type on which Vehicle runs on.
* Engine HP - Power of the Engine
* Engine Cylinders – Number of Cylinders.
* Transmission Type – Gear Transmission System.
* Driven Wheels - Drive Wheel System of the car.
* Number of Doors - Number of the doors in the car
* Market Category – Market Segment Car is for.
* Vehicle Size - Vehicle Size.
* Vehicle Style - Type of Vehicle
* highway MPG - Average of Vehicle on Highway
* city mpg - Average of Vehicle on City
* Popularity - Popularity Scores from the Consumers/People.
* MSRP - Maximum Selling Price in Retail from Manufacture.

The data has 16 Columns and 11914 rows: (11914, 16) 11

**Files Use for Analysis**

To perform the analysis, I have used Python. Python allows us to visualize the data with different Libraries and also cleaning of the data is made very simple. The Jupyter Notebook I use is named **car\_feature.ipynb.**

**Python Version Use – 3.11**

**Libraires Used –**

* **Pandas**
* **Numpy**
* **MatplotLib**
* **Seaborn**

**Utilities.py –**

I have made utilites.py to store the function that I used in project performs various task. That include visualization, feature selection, data cleaning.

Functions that I have used in the data.

**Clean data –** To handle missing Value and formatting column data types.

**Bar Line Graph –** To create the bar plots of average msrp and average popularity score.

**Scatter plots –** To generate the Regplot and correlation score and R2 values.

**DATA CLEANING**

Overall, data cleaning is an essential prerequisite for any data analysis or machine learning project. It lays the foundation for accurate, reliable, and actionable insights that drive decision-making and drive success in various domains

So here are the steps that I took too Clean the data.

**Cleaning Missing Values**

Like any other data we need to first take care of the missing Value of the data.

The data has many missing Values.

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 3742

Vehicle Size 0

Vehicle Style 0

highway MPG 0

city mpg 0

Popularity 0

MSRP 0

To I dropped the Missing Values in the Columns.

**Correct Formatting of the Data**

The Feature Model has some most of the data in string format. However around 3000 records are in Date time format which during visualization would cause problems. Therefore, I have changed those dates to Unknown Model.

To perform the data cleaning, I used the Function stored in the Utilities.py

**Analysis**

The features are of two main types –

* **Categorical**
* **Numerical**

So, to perform the EDA I begin with first segregating both the columns in to two different groups. Cat col and Num Col.

One extra step I took as I Feel I would create the correlation matrix between Categorical Column as well during further analysis. So, I created columns based on that Cat Col Corr which has Msrp along with the categorical columns.

**Count Plots**

Visualize the Count Plots of the Categorical Columns presented in the data. It would help me understand the mode or the popular features of the car based on their counts.

**Aim** *– The count plot of the categorical features will allow us to understand what features most cars have in the market. The counting of each value of feature will help us understand which car features are in huge supply in the market.*

**Make**

The Columns talk about the Brand of the car or the Brand Manufacturer will be the most appropriate term to use here.

TOP Seller and there Make:

Make

Chevrolet 614

Volkswagen 581

Ford 492

Cadillac 397

Mercedes-Benz 352

BMW 334

Infiniti 330

Audi 328

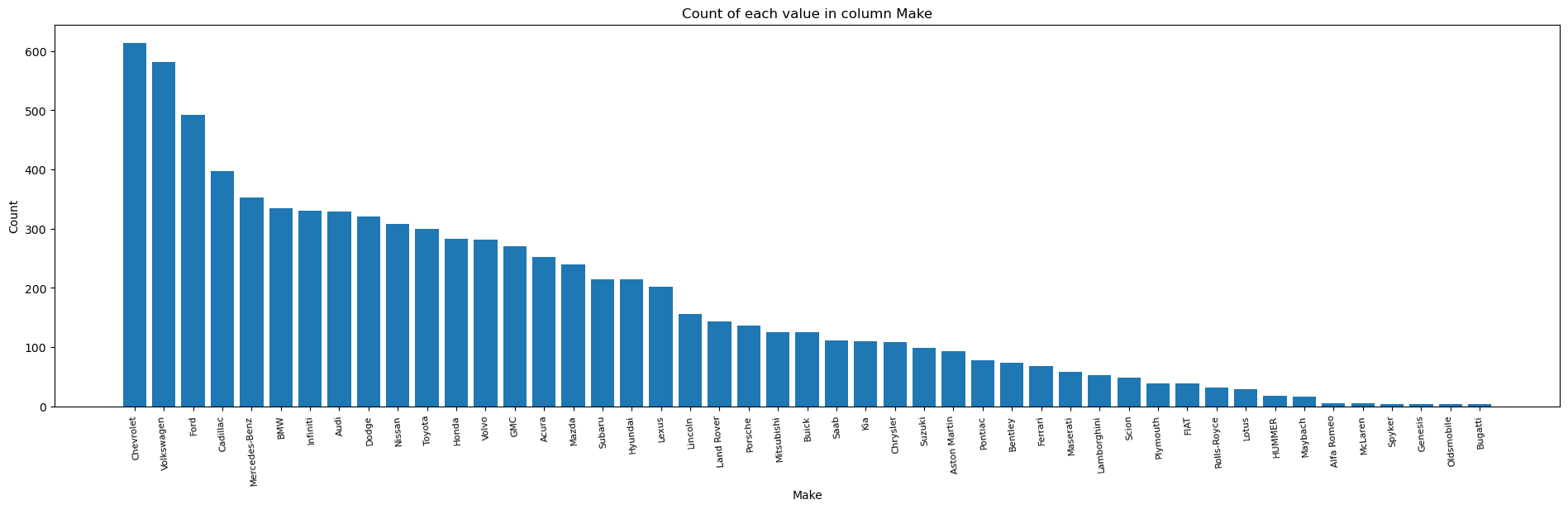
Dodge 320

Nissan 308

Toyota 300

Honda 282

Volvo 281



**Model**

The columns tells about the models of the car.

Top Count of Model:

Model

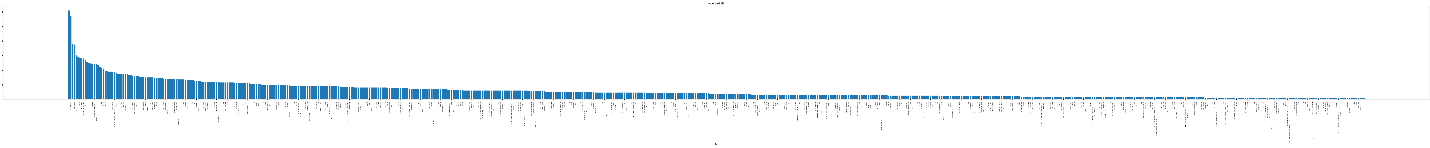
F-150 122

Silverado 1500 114

GTI 76

Beetle 75

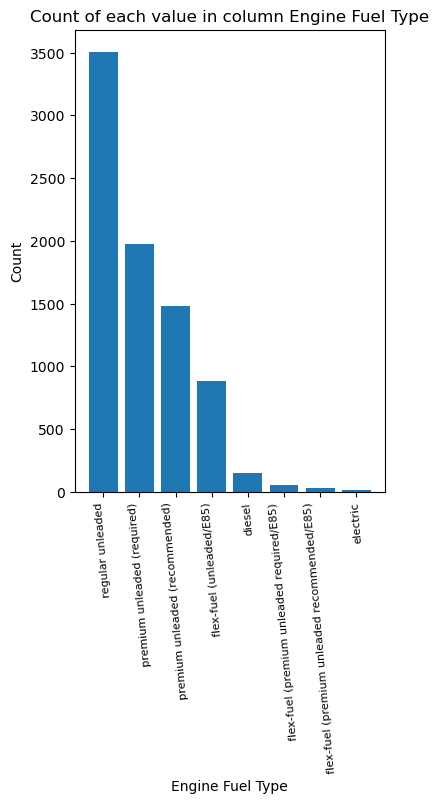
Corvette 60



Since value are too large in this column the plot is bit messy. However, one can check the top 8 car model in above table.

**Engine Fuel Type**

Engine Fuel Type of the car. Tell us about the fuel the car needs to run on roads.



Engine Fuel Type

regular unleaded 3507

premium unleaded (required) 1973

premium unleaded (recommended) 1477

flex-fuel (unleaded/E85) 881

diesel 153

flex-fuel (premium unleaded required/E85) 54

flex-fuel (premium unleaded recommended/E85) 26

electric

**Transmission Type**

The Gear Shifting system of the car.

Transmission Type

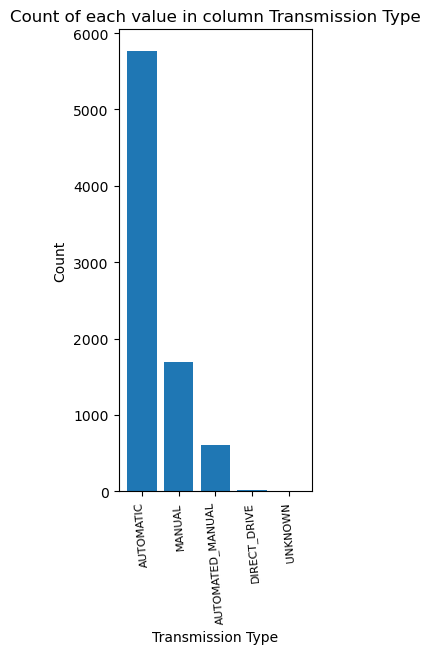
AUTOMATIC 5761

MANUAL 1695

AUTOMATED\_MANUAL 610

DIRECT\_DRIVE 15

UNKNOWN 3



**Driven Wheels**

Driven\_Wheels:

Driven\_Wheels

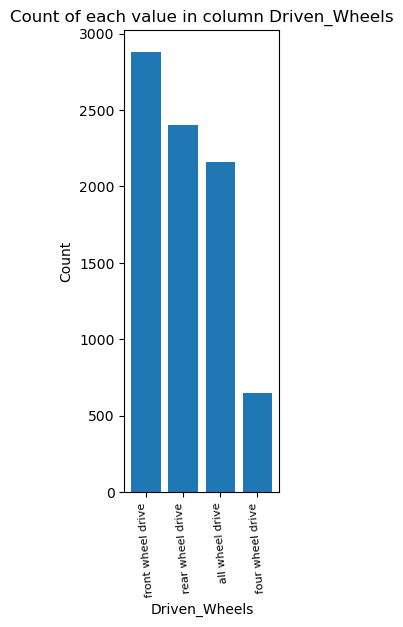
front wheel drive 2879

rear wheel drive 2401

all wheel drive 2158

four wheel drive 646

Name: count, dtype: int64

****

Market Category

Values count of each categorical column Market Category:

Market Category

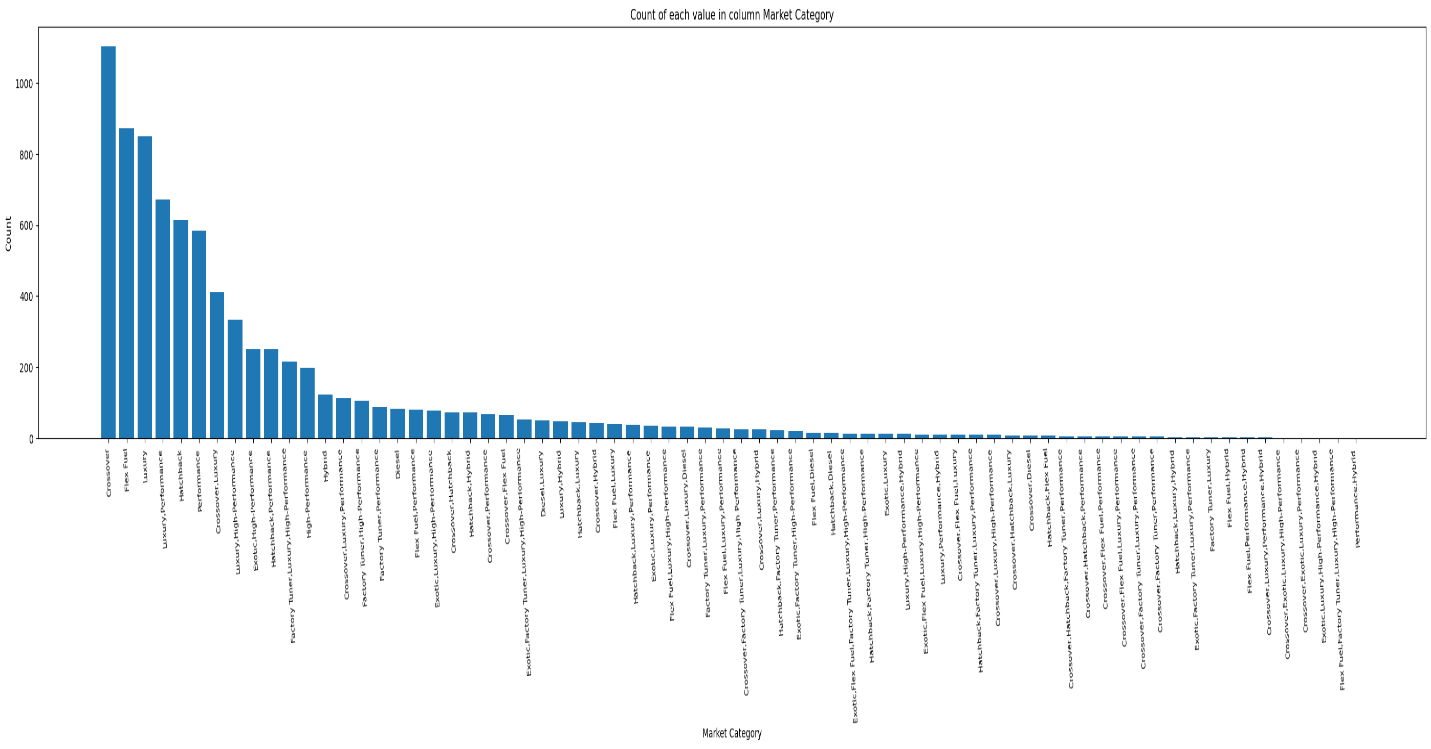
Crossover 1103

Flex Fuel 872

Luxury 851

Luxury,Performance 673

Hatchback 614



**Vehicle Size**

Values count of each categorical column Vehicle Size:

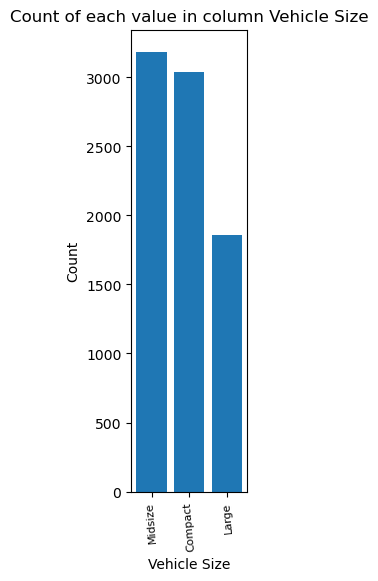
Vehicle Size

Midsize 3187

Compact 3039

Large 1858

Name: count, dtype: int64

****

**Vehicle Style**

Values count of each categorical column Vehicle Style:

Vehicle Style

4dr SUV 2093

Sedan 1909

Coupe 939

4dr Hatchback 678

Convertible 670

2dr Hatchback 503

Wagon 357

Crew Cab Pickup 344

Extended Cab Pickup 191

Regular Cab Pickup 176

Passenger Van 100

Passenger Minivan 65

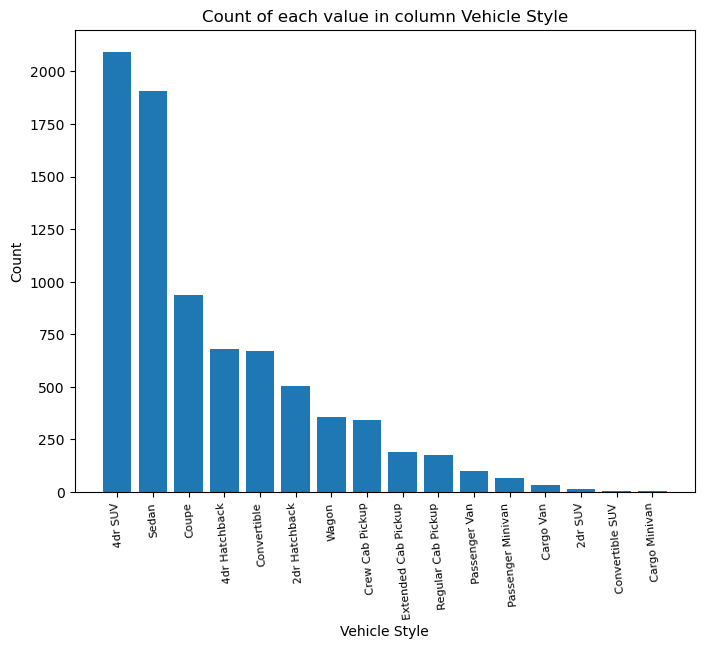
Cargo Van 34

2dr SUV 14

Convertible SUV 6

Cargo Minivan 5

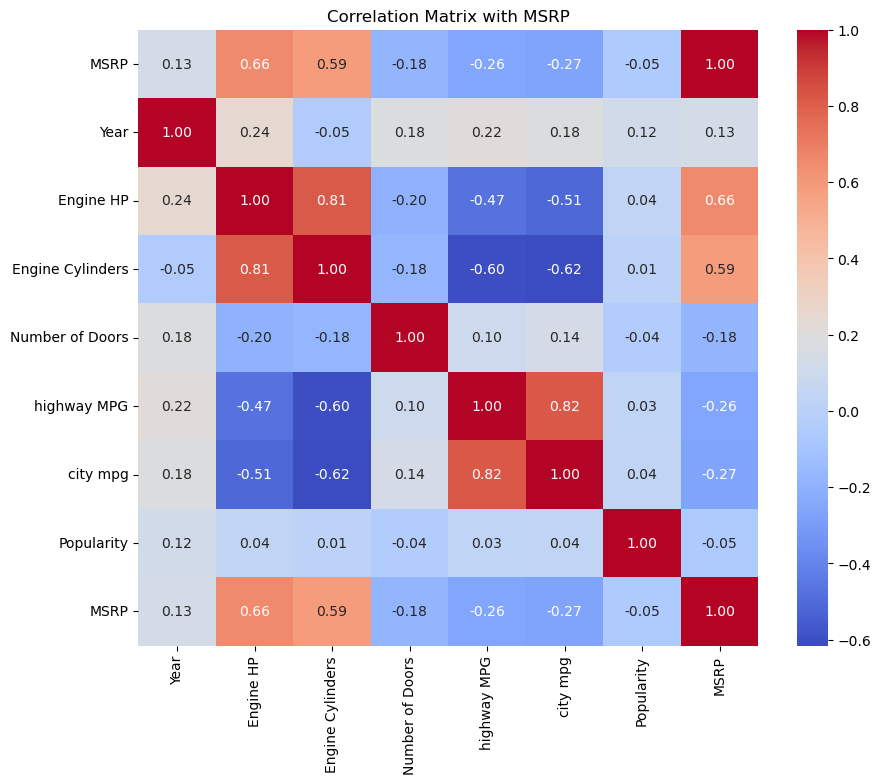
Name: count, dtype: int64

****

**Findings**

* The Brands such as Chevrolet, Ford, Volkswagen are Dominating the Market with more number of cars availability. Also, Nissan Toyota, Hinda the Asian brand of the car have good produced more cars.
* The world does want to transfer to the green fuels. However electric cars sales are still very few. Regular Unleaded still dominates the markets and are available more than others.
* Automatics cars have good count. One can say people are more shifting to automatic system than manual gear system. Maybe the convenience is the point behind such a move.
* Driven Wheels front wheel cars are produced in large number than the four-wheel drive. Less cars have four-wheel drive. The reason maybe it is the special feature and not all cars and series could include such feature.
* The market has seen more cross over category along with the flex fuel, luxury, luxury performance and Hatch back. But most of the cars are in hatchback category. Could be the demand of hatch back have grown over years therefore more companies are launching hatch backs.
* Now in vehicle size Mid-size cars are produced more than the others. Reason could be they are better for the city.
* Four door SUV are produced more than another category. The style is Sedan, Coupe and 4dr Hatchback.

**CORRELATION MATRIX WITH FEATURE NUMERICAL COLUMNS**

**Aim –** The aim is to find which features better correlates with the MSRP and Popularity. The feature that correlates with MSRP can help us optimize and gain good market price. Finding the feature Correlating well with the popularity score will allows us to optimize the car feature that can become instant hit among the audience. ****

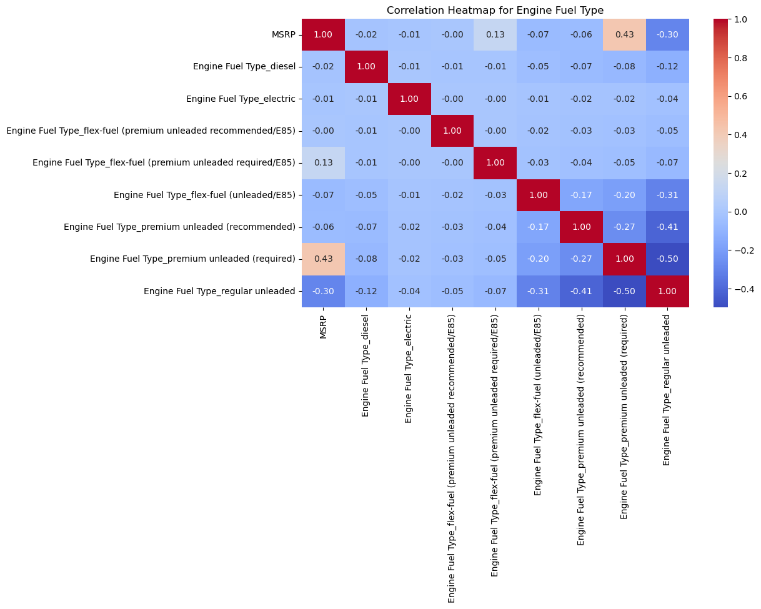
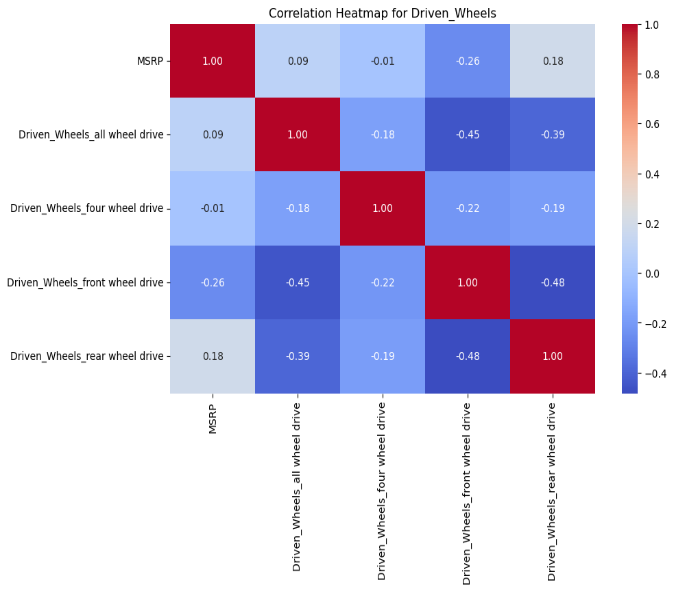
**FINDINGS**

**MSRP –** MSRPhas good positive correlation with the features such as Engine Cylinders and Engine HP. So, if we can sell good horse power to audience we can sell it on the high price. Also, they have normal positive linear relationship with the Popularity index.

**Popularity Scores –** The correlation with another feature is not strong. Besides, number of doors and MSRP the correlation is negative**.** So not much could be said about the which feature can help us to gain more popularity. So, if the door increases of the car the popularity drops among the masses. Also, the higher price range can decrease the popularity among the consumers.

**MSRP VS OTHER CATEGORICAL FEATURES**

I have used some of the categorical features that have less unique values to check the correlations. The correlation can help us understand which features are better for makers to avail high price from the markets.

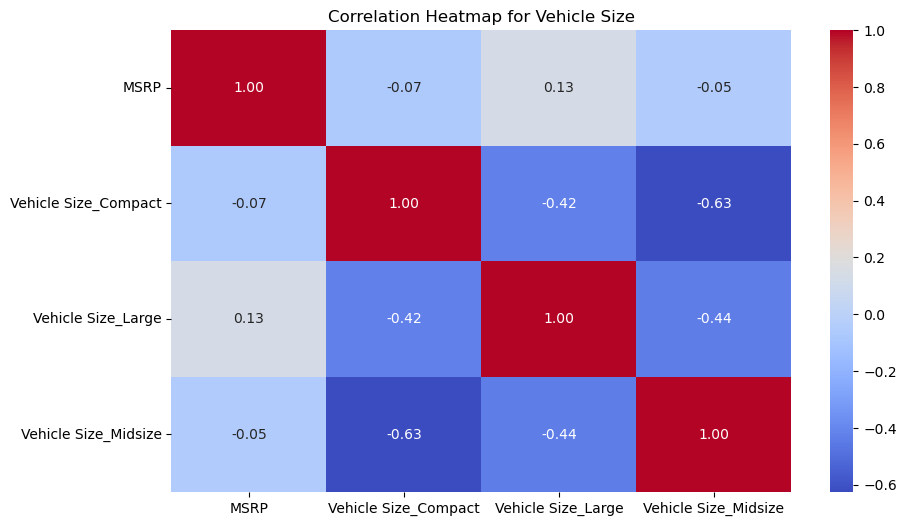
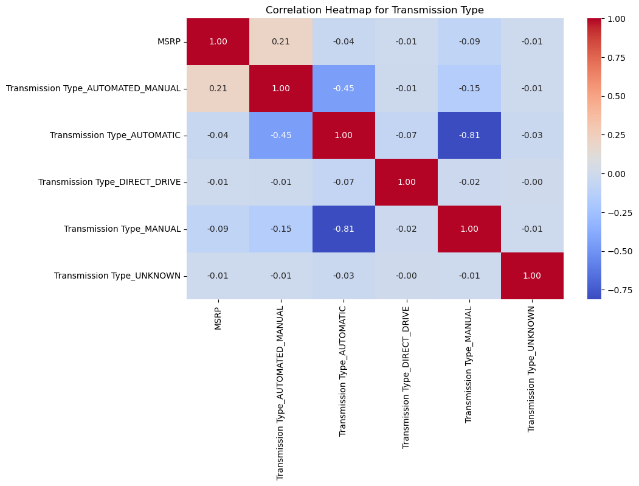


**Driven Wheel**

Driven Wheels front wheel drive have negative correlation with the MSRP. However, not strong but it still negatives correlation. When it comes to rear wheel drive, we can ask for money from the market.

**ENGINE FUEL TYPE**

The cars with the fuel type Premium Unleaded (recommended) are expensive and can help us gain more money from the markets.



**Transmission Type**

Only Automated Manual had a high correlation with MSRP. Its good as most of the vehicle in the market available have the automated system. The company can use this feature in the car to avail the higher price for the car.

**Vehicle Size**

The Large Vehicle Size allows companies to sell the car at high market price. However, compact and Midsize are negatively correlated although not so strong.

**Key Findings –**

* Engine HP can help us get more price from the market.
* The rear wheel drive is the feature which can help us avail a good price from market.
* Automated car is way to go if we need to charge more money from the market.
* Vehicle drive has not much higher correlation in any category. But still focusing on large vehicle size can help us charge more money from the market.

**MSRP VS POPULARITY**

**FOR**

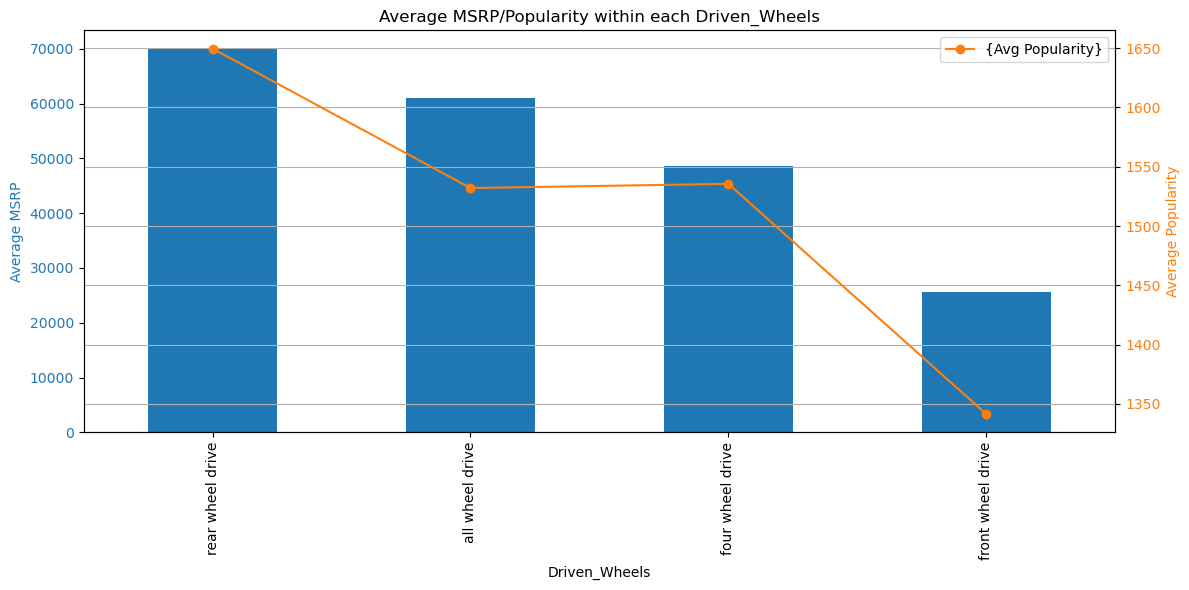
**THE CATEGORICAL COLUMNS**

Since we have plot the correlation heatmap with other categorical I would like to check which feature values does well with MSRP and Popularity. The features that have correlated more with MSRP if they have good popularity score we can be assured of them to include in our car.

**Aim –** To find the categorical features that have good popularity score along with the MSRP.

**DRIVEN WHEELS**

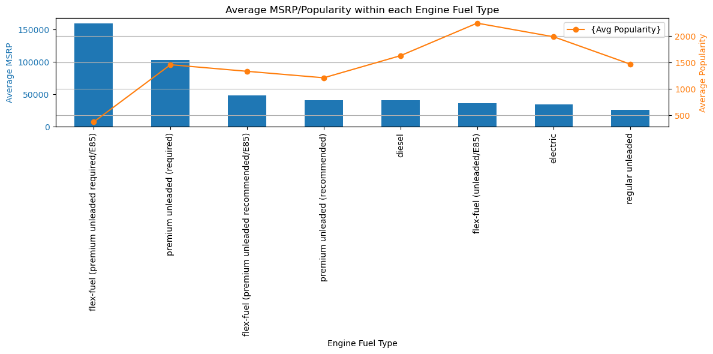
In above correlation plots we find out that rear wheel drive is best is we want to charge more price from the markets. And the Popularity of the rear wheel drive among the consumers is also great we can move forward. Includig rear wheel drive will surely help us gain more money from market and also demand from consumers.

****

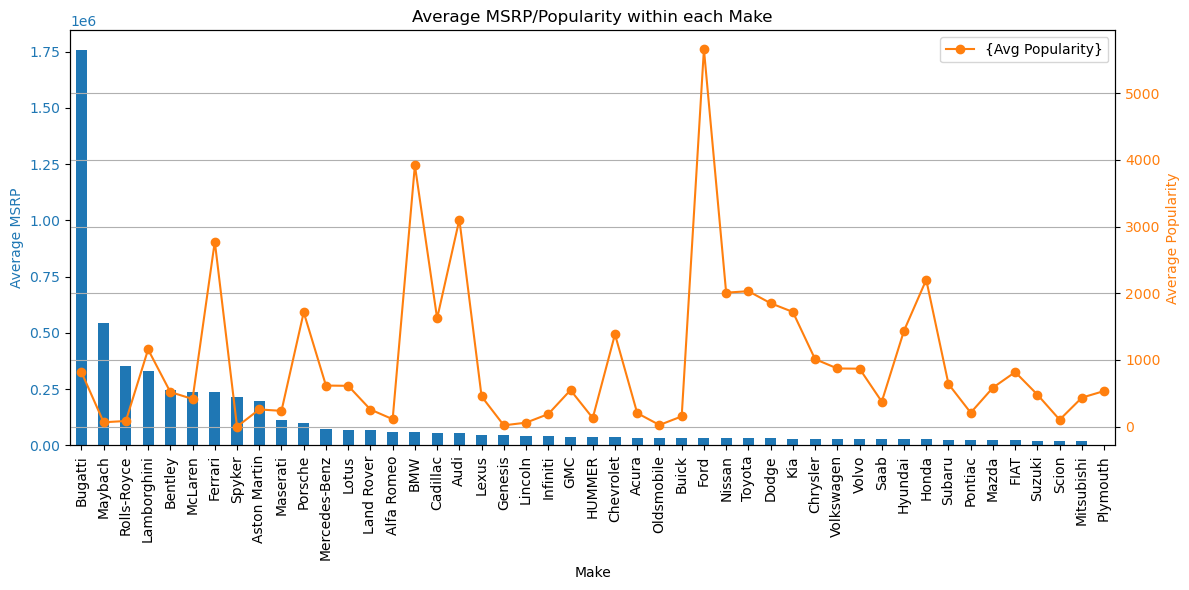
**ENGINE FUEL TYPE**

The correlation plots show us that using the car with premium unleaded fuel allows us to charge more money from the market. But in the Market the popularity for the flex fuel is more and electric too than the Premium Unleaded (required) fuel types. However, MSRP is not good than the Premium Unleaded (required).

The regular unleaded gives the high MSRP from but the popularity from the people is very low. Using such feature may help us gain more price from market. But could be risky as not many people like the feature as shown by popularity score. Using the premium Unleaded (required) is best choice we can do. It has good popularity among the consumers we can use the feature to ask more from the market.

****

**MAKE**

****

**High MSRP cars in MAKE**

If the car has tag like Bugatti, Maybach, Lamborghini, McLaren they can charge more price for the car in the market. Since they are the premium sports car company the popularity of the car is justified. In such premium sport car section brand Ferrari is most popular out of all. Although they rank very low in MSRP compared to Bugatti.

**Popular cars in MAKE**

Audi and BMW have good popularity score and also charge good MSRP if we exclude the Top 15 brands from the data. This category is doing pretty good in terms MSRP and popularity score.

The American Brand Ford is the most popular among all cars we have in the data. However, MSRP is low compare to others. It has 27 ranks in terms of Average MSRP. We can see if the cars of ford that use rear wheel drive, automatic system, and large car size can help improve on the MSRP.

Ferrari is the most balanced brand that has good Average MSRP and Popularity among the market. Apart from that Porsche can be included in that list with good average MSRP and Popularity more than the Bugatti with highest average MSRP out of all.

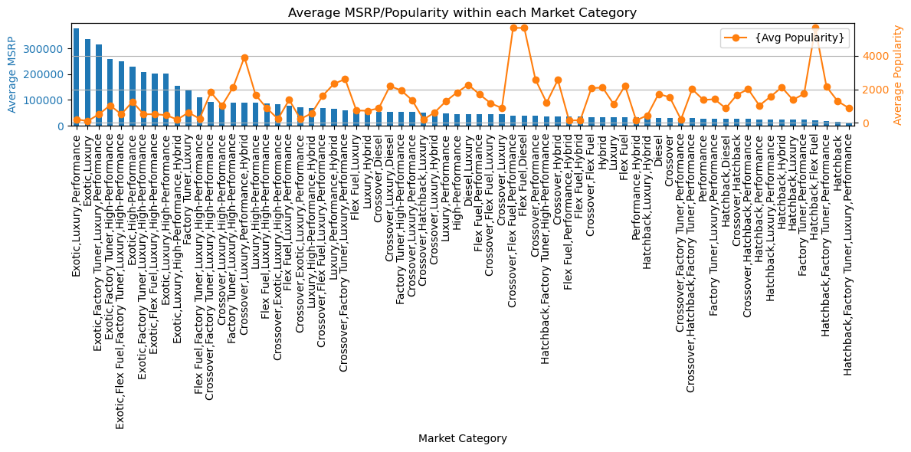
**MARKET CATEGORY**

The Market Category that offers highest MSRP are Exotic, Luxury, Performance and combination of these three categories.

However, Popularity scores are low among for them. Moreover, Crossover, Luxury, high Performance Hybrid have the well balance between Popularity and Popularity Scores.

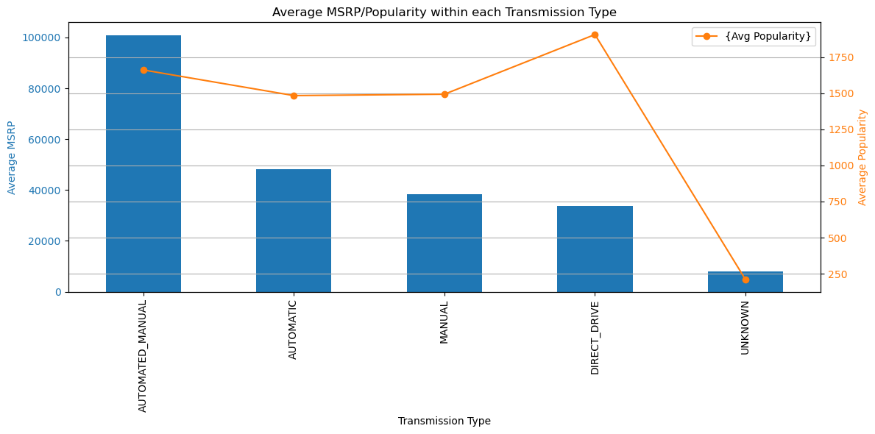
The popular category with average MSRP is Flex Fuel-Diesel-Performance, and Flex Fuel Diesel.

The Hatchback-Flex fuel have high popularity but we can’t charge high amount for it in market.

****

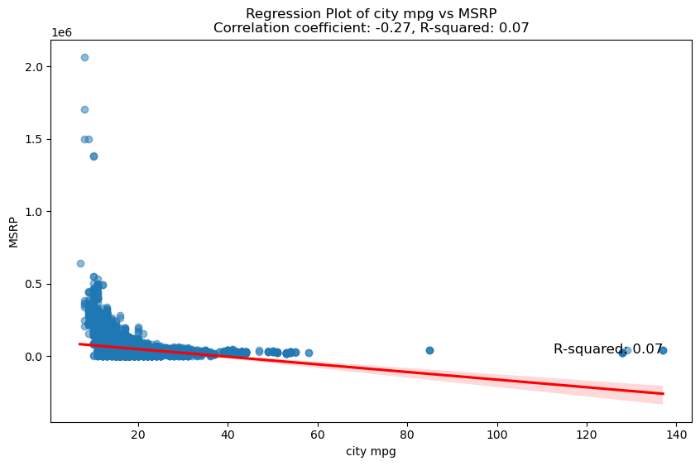
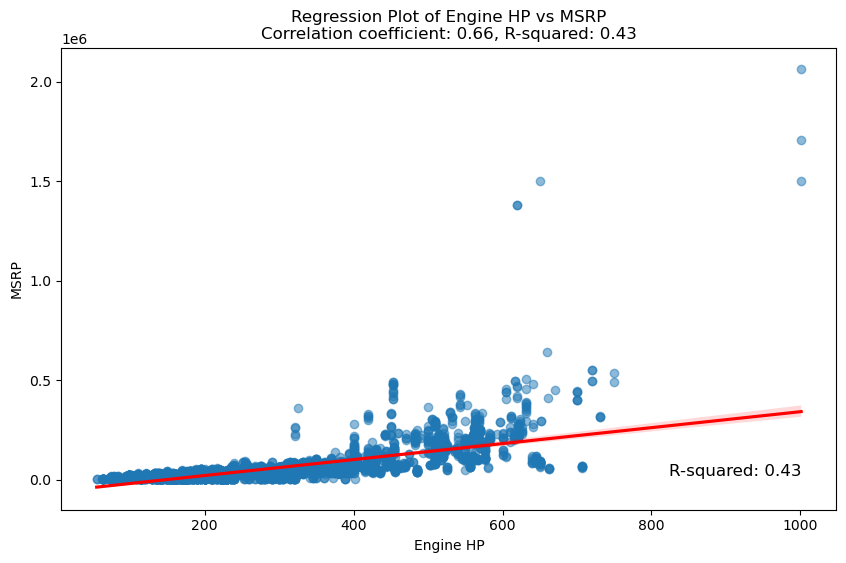
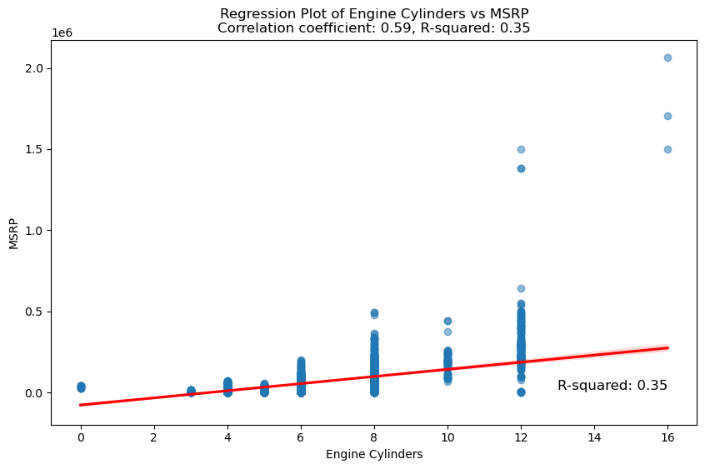
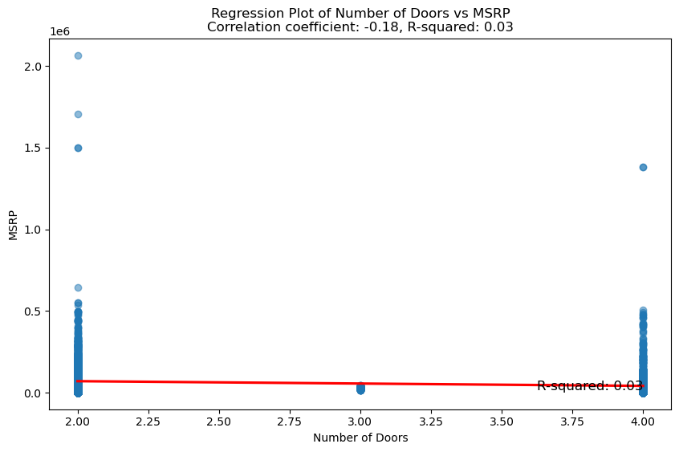
**TRANSMISSION TYPE**

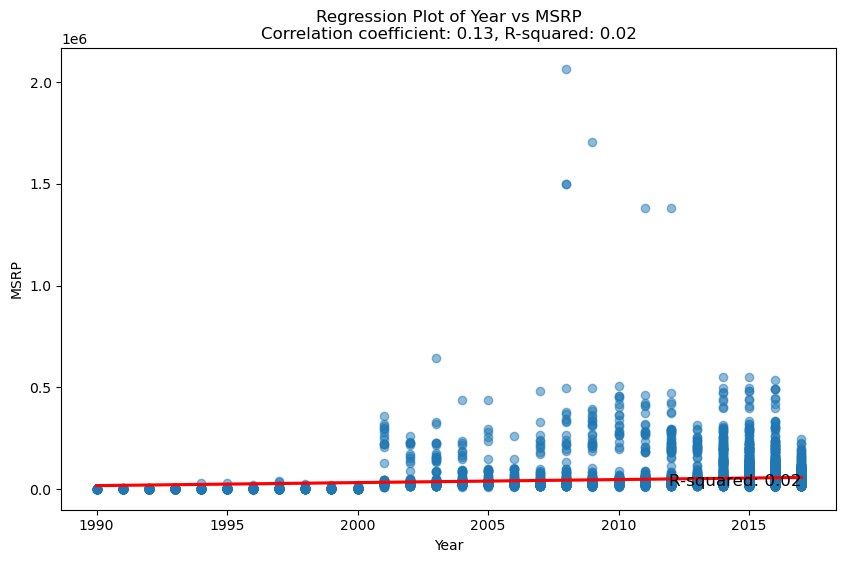
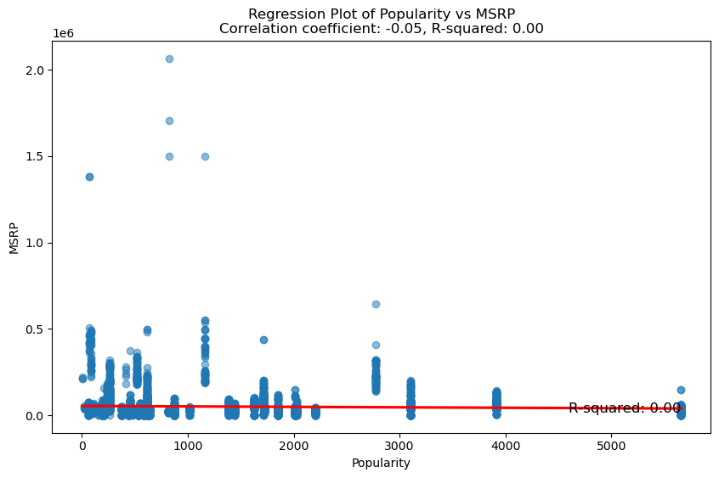
Automated Manual is the feature that correlates more with the MSRP. Also, it has good popularity score. The feature can guarantee good demand from people along with good price from the market. The Direct Drive is the Popular category but have less MSRP.

****

**SCATTER PLOT FOR NUMERICAL**

**DATA TYPES**





So the Popularity and MSRP have no correlation among each other. Which is fine, and could happen the car category is vast and they tend to different type of consumers needs.

**YEARS**

However, features such as Year show positive correlation and can explain 2 % of MSRP pattern. Logically too the new car will sell at a higher price in the markets. Only exceptions are vintage collectible cars.

**HORSE POWER**

**AND**

**NUMBER OF CYLINDERS**

The features correlate strongly with the MSRP. The horse power can explain 35 % of the variance in MSRP and cylinder 34 % of the variance in MSRP, respectively. So, higher the Horse Power or number of cylinders more money we can get from the markets.

**Conclusions**

* The Brands such as Chevrolet, Ford, Volkswagen are Dominating the Market with more number of cars availability. Also, Nissan Toyota, Hinda the Asian brand of the car have good produced more cars.
* The world does want to transfer to the green fuels. However electric cars sales are still very few. Regular Unleaded still dominates the markets and are available more than others.
* Automatics cars have good count. One can say people are more shifting to automatic system than manual gear system. Maybe the convenience is the point behind such a move.
* Driven Wheels front wheel cars are produced in large number than the four-wheel drive. Less cars have four-wheel drive. The reason maybe it is the special feature and not all cars and series could include such feature.
* The market has seen more cross over category along with the flex fuel, luxury, luxury performance and Hatch back. But most of the cars are in hatchback category. Could be the demand of hatch back have grown over years therefore more companies are launching hatch backs.
* Now in vehicle size Mid-size cars are produced more than the others. Reason could be they are better for the city.
* Four door SUV are produced more than another category. The style is Sedan, Coupe and 4dr Hatchback.

**Features that impact higher market Value**

* Using the premium Unleaded (required) is best choice we can do. It has good popularity among the consumers we can use the feature to charge more price for the car from the market.
* If the car has tag like Bugatti, Maybach, Lamborghini, McLaren they can charge more price for the car in the market. Since they are the premium sports car company the low popularity of the car is justified as they suit to certain section of society. However, even in such premium sport car section brand Ferrari is most popular out of all. Audi and BMW have good popularity score and also charge good MSRP if we exclude the Top 15 brands from the data. This category is doing pretty good in terms MSRP and popularity score.
* Crossover-Luxury-high-Performance Hybrid have the well balance between MSRP and Popularity Scores. So, we can go with these features. The popular category with average MSRP is Flex Fuel-Diesel-Performance, and Flex Fuel Diesel. So, we can look at these features for our car to avail good price along with great demand from end consumer.
* Automated Manual is the feature that correlates more with the MSRP. Also, it has good popularity score. The feature can guarantee good demand from people along with good price from the market.
* New car will sell at a higher price in the markets. Only exceptions are vintage collectible cars. So, car should be latest if we want to charge high price for it in the market
* Higher the Horse Power or number of cylinders more money we can get from the markets.