Analyzing the Impact of Car Features on Price and Profitability

Arcot Navya Sai (navyasaipatnaik@gmail.com) +918074851394

Project Description:

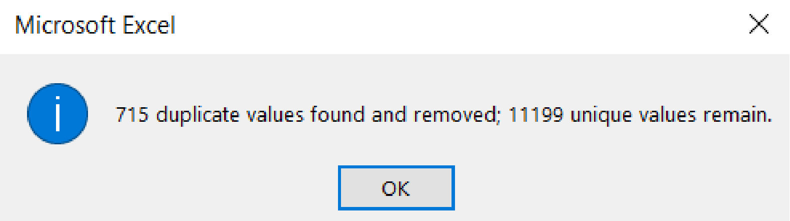
This Project Focuses on automotive industry, how the Prices of Cars change based on fuel efficiency, environmental sustainability, and technological innovation. We are going to conduct analysis to know How can a car manufacturer optimize pricing and product development decisions to maximize profitability while meeting consumer demand?

We will be analyzing the relationship between a car's features, market category, and pricing, and identifying which features and categories are most popular among consumers and most profitable for the manufacturer. By using data analysis techniques such as regression analysis and market segmentation, the manufacturer could develop a pricing strategy that balances consumer demand with profitability, and identify which product features to focus on in future product development efforts. This could help the manufacturer improve its competitiveness in the market and increase its profitability over time.

The Data Source Used in the project is Car_data.csv, Which contains Columns like 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, Popularity, MSRP.

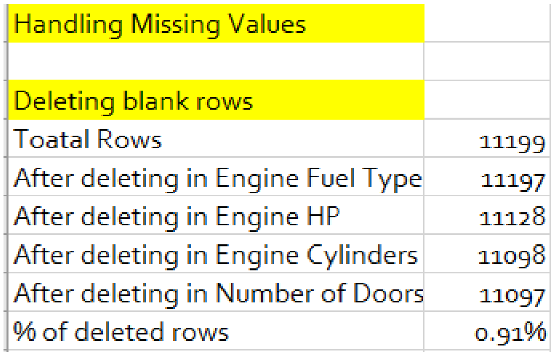
Before Performing any analysis on the data we will be performing the Data Cleaning and preprocessing like Outlier Analysis on the data.

The duplicates were first removed



715 duplicate values found and removed; 11199 unique values remain.

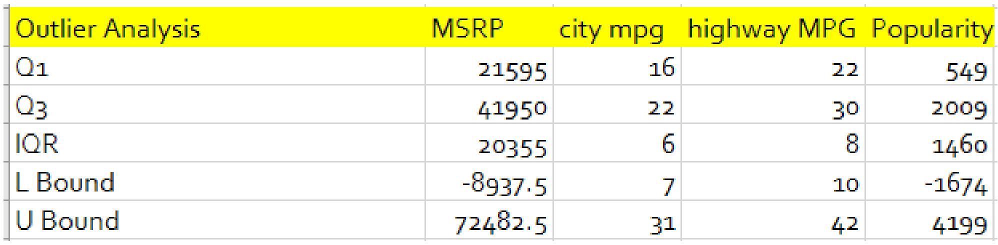
Then We Handled the missing values there were very less empty cells so we directly removed them



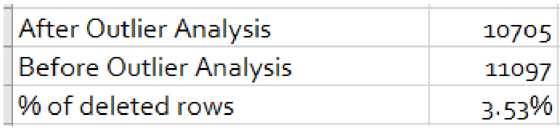
Handling Missing Values	
Deleting blank rows	
Toatal Rows	11199
After deleting in Engine Fuel Type	11197
After deleting in Engine HP	11128
After deleting in Engine Cylinders	11098
After deleting in Number of Doors	11097
% of deleted rows	0.91%

Later we continued with the outlier analysis on the clean data, there are many numerical columns in the data, we performed the outlier analysis only on 4 columns namely highway MPG, city mpg, Popularity, MSRP. We cannot remove all the outlier data as we will be losing many rows so we only removed the rows where all the four columns were outliers by using formula.

=AND(OR([@MSRP] <'Data Cleaning'!$B$15,[@MSRP]>'Data Cleaning'!$B$19), OR([@[city mpg]]<'Data Cleaning'!$C$15,[@[city mpg]]>'Data Cleaning'!$C$19),OR([@[highway MPG]]<'Data Cleaning'!$D$15,[@[highway MPG]]>'Data Cleaning'!$D$19),OR([@Popularity]<'Data Cleaning'!$E$15,[@Popularity]>'Data Cleaning'!$E$19))



Outlier Analysis	MSRP	city mpg	highway MPG	Popularity
Q1	21595	16	22	549
Q3	41950	22	30	2009
IQR	20355	6	8	1460
L Bound	-8937.5	7	10	-1674
U Bound	72482.5	31	42	4199

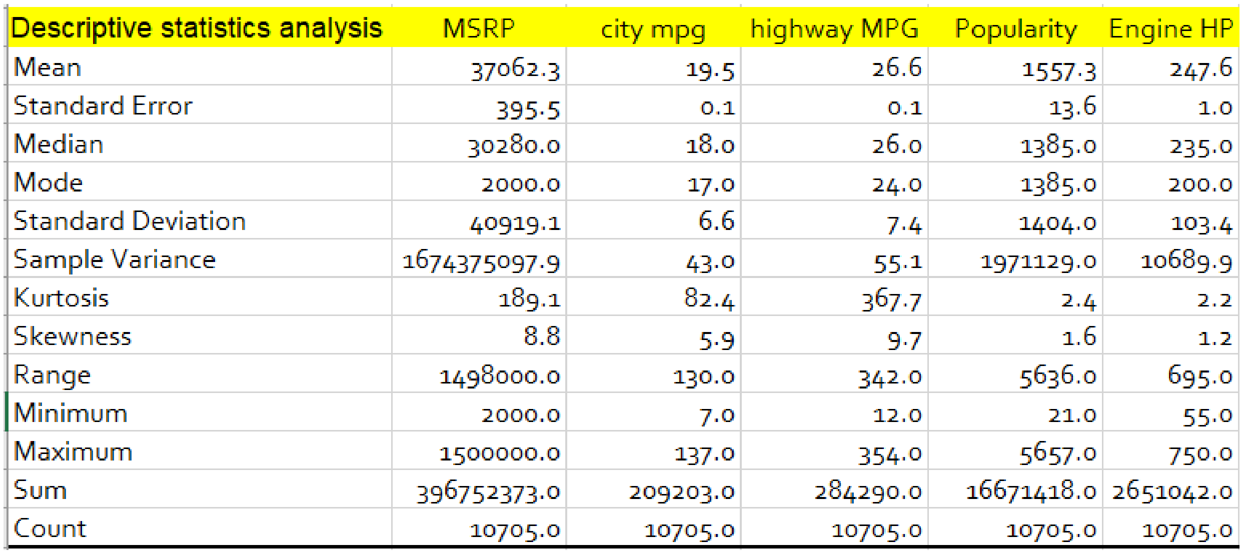


After Outlier Analysis	10705
Before Outlier Analysis	11097
% of deleted rows	3.53%

We then continue the Analysis assuming we now have the clean data with very less outliers.

Approach:

To Understand the data better we are going to do the descriptive Analysis on the few numerical columns using the Data Analysis tool pack with Excel.



Descriptive statistics analysis	MSRP	city mpg	highway MPG	Popularity	Engine HP
Mean	37062.3	19.5	26.6	1557.3	247.6
Standard Error	395.5	0.1	0.1	13.6	1.0
Median	30280.0	18.0	26.0	1385.0	235.0
Mode	2000.0	17.0	24.0	1385.0	200.0
Standard Deviation	40919.1	6.6	7.4	1404.0	103.4
Sample Variance	1674375097.9	43.0	55.1	1971129.0	10689.9
Kurtosis	189.1	82.4	367.7	2.4	2.2
Skewness	8.8	5.9	9.7	1.6	1.2
Range	1498000.0	130.0	342.0	5636.0	695.0
Minimum	2000.0	7.0	12.0	21.0	55.0
Maximum	1500000.0	137.0	354.0	5657.0	750.0
Sum	396752373.0	209203.0	284290.0	16671418.0	2651042.0
Count	10705.0	10705.0	10705.0	10705.0	10705.0

There is more Standard deviation in the MSRP because of Luxury cars like Bugatti we can take only those data out and perform separate analysis or even remove the data, but I am going to keep the data as we have comparatively small data.

We have used visualization like multiple types of Charts in the analysis to get the insights. Which we will be seeing in detail through the report.

We have also used Multiple Linear Regression analysis to get the coefficients of regression equation and understand what columns are actually influencing the car price.

As we have the data which is having both numerical and categorical columns there were challenges during the regression analysis and also during plotting the scatter plot and bubble plot. To make the analysis possible we have created columns converting the categorical data to numerical. We assigned the numbers for each category to make analysis easier to perform.

Tech-Stack Used:

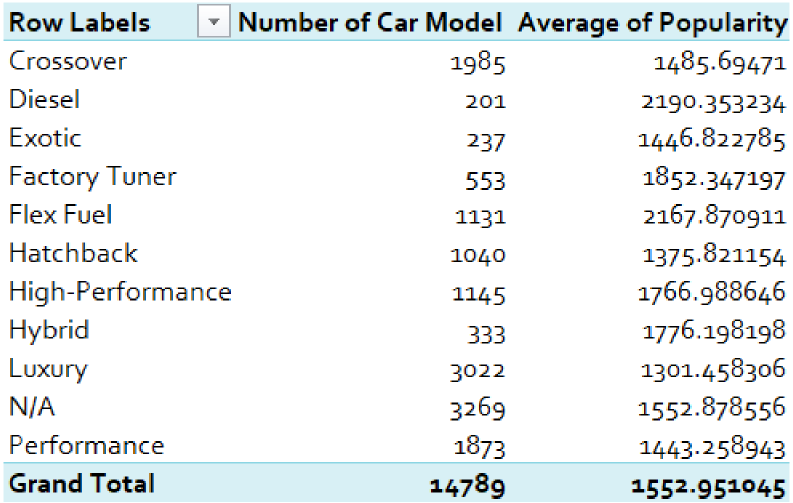
We have used

1. Microsoft Excel Professional Plus 2016.
2. Microsoft Word Professional Plus 2016.
3. Data Analysis Tool Pack from Excel.
4. Pivot Tables and Pivot Charts from Excel.

Insight:

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.

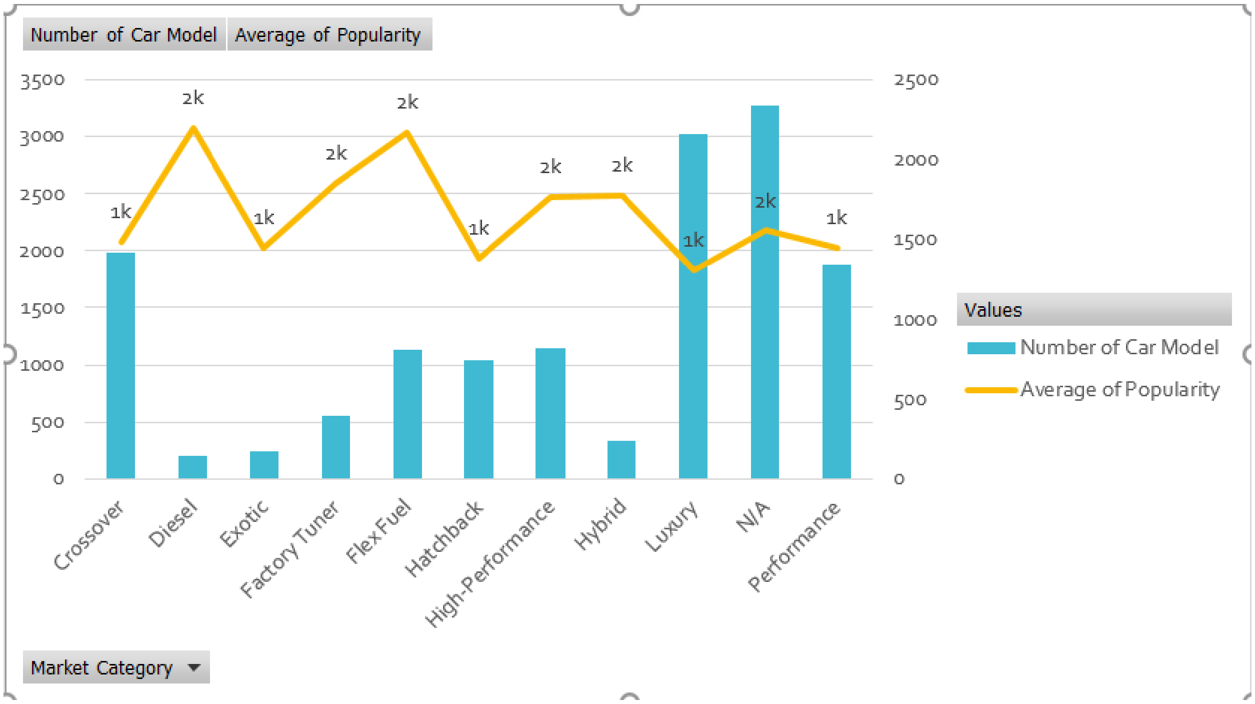


Row Labels	Number of Car Model	Average of Popularity
Crossover	1985	1485.69471
Diesel	201	2190.353234
Exotic	237	1446.822785
Factory Tuner	553	1852.347197
Flex Fuel	1131	2167.870911
Hatchback	1040	1375.821154
High-Performance	1145	1766.988646
Hybrid	333	1776.198198
Luxury	3022	1301.458306
N/A	3269	1552.878556
Performance	1873	1443.258943
Grand Total	**14789**	**1552.951045**

Insights:

1. Luxury Car models are more in number where N/A is not known so we are not considering them.
2. Diesel Car Models are less number than all the other models, but we can see the Average popularity is more.
3. Luxury has very less average popularity

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



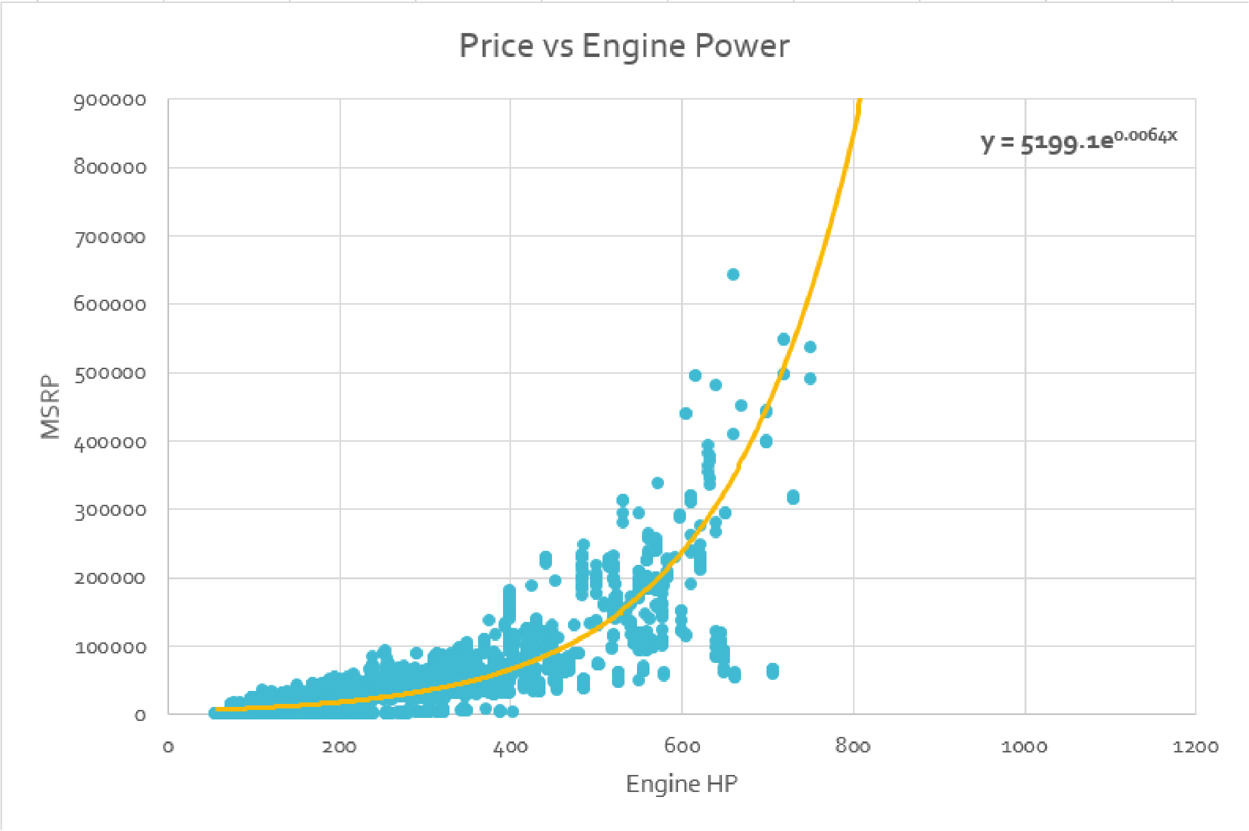
As we can see from the chart though Diesel Cars are less in number, it has more average popularity.

Business Problem and recommendations:

1. The car manufacturers can keep Popularity of car models in mind to increase the production of Diesel cars and Hybrid Cars.
2. Exotic Cars can be manufactured on orders or can have pre order on them.
3. High Performance cars are preferred so, their production can be increased.

<u>What is the relationship between a car's engine power and its price?</u>

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



Insights:

- We can clearly see from the chart that Engine Horse Power and Price of the car are related exponentially. So, they follow Exponential Regression.
- The equation of the regression is **$y = 5199.1e^{0.0064x}$**

Business Problem and recommendations:

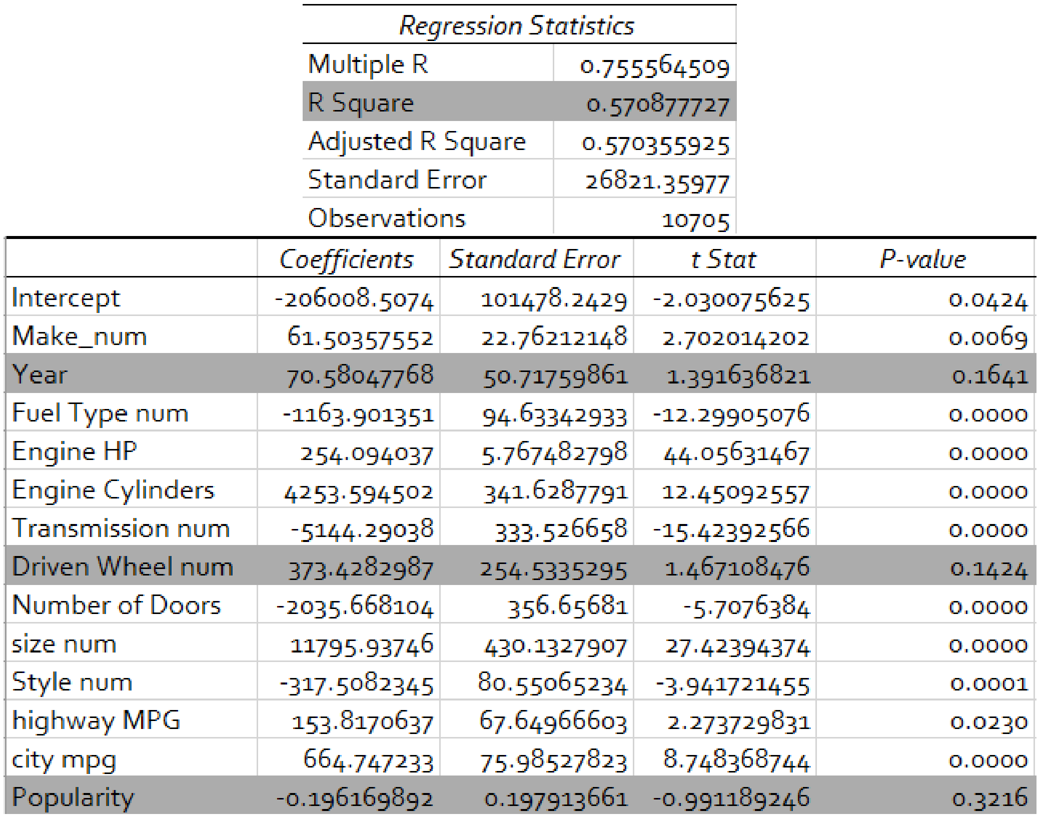
- We should find the best Engine Horse Power and Price so that there is maximum sales of the car. From Business point of view if the Horse Power is very large the cost of Car production cost will gradually increases leading to High Prices and very less people able to afford it. So we have to find the optimum point.

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.

For completing this task, we have to understand the target is the numerical continuous values so we will be using Regression models, but as we see the data has many categorical columns in it. To do the regression analysis we are first converting the categorical columns to numerical columns by simply assigning each category a number.

After which we have used the Data Analysis Toolbox to do the analysis. The below is the obtained output for the analysis.



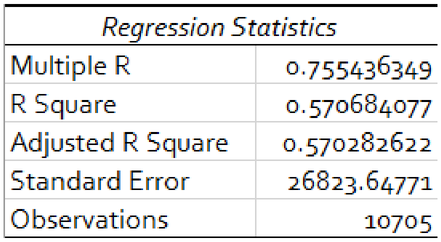
Regression Statistics	
Multiple R	0.755564509
R Square	0.570877727
Adjusted R Square	0.570355925
Standard Error	26821.35977
Observations	10705

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-206008.5074	101478.2429	-2.030075625	0.0424
Make_num	61.50357552	22.76212148	2.702014202	0.0069
Year	70.58047768	50.71759861	1.391636821	0.1641
Fuel Type num	-1163.901351	94.63342933	-12.29905076	0.0000
Engine HP	254.094037	5.767482798	44.05631467	0.0000
Engine Cylinders	4253.594502	341.6287791	12.45092557	0.0000
Transmission num	-5144.29038	333.526658	-15.42392566	0.0000
Driven Wheel num	373.4282987	254.5335295	1.467108476	0.1424
Number of Doors	-2035.668104	356.65681	-5.7076384	0.0000
size num	11795.93746	430.1327907	27.42394374	0.0000
Style num	-317.5082345	80.55065234	-3.941721455	0.0001
highway MPG	153.8170637	67.64966603	2.273729831	0.0230
city mpg	664.747233	75.98527823	8.748368744	0.0000
Popularity	-0.196169892	0.197913661	-0.991189246	0.3216

As we can see from the above Tables the P-value for three columns are highlighted.

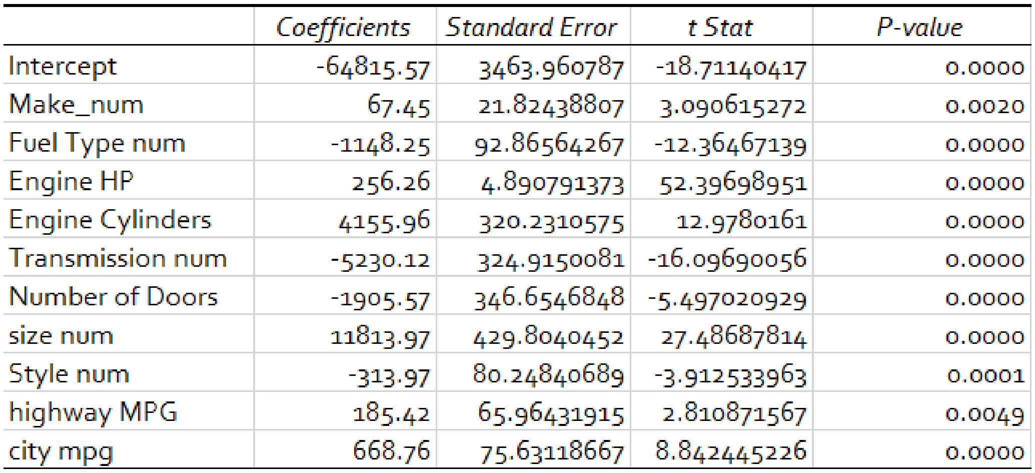
- A low P-value (< 0.05) means that the coefficient is likely not to equal zero.
- A high P-value (> 0.05) means that we cannot conclude that the explanatory variable affects the dependent variable (here, Year, driven wheel num and Popularity do not affect the target MSRP)
- A high P-value is also called an insignificant P-value

So we have removed those columns and re-performed the Regression analysis, which gave the below results.



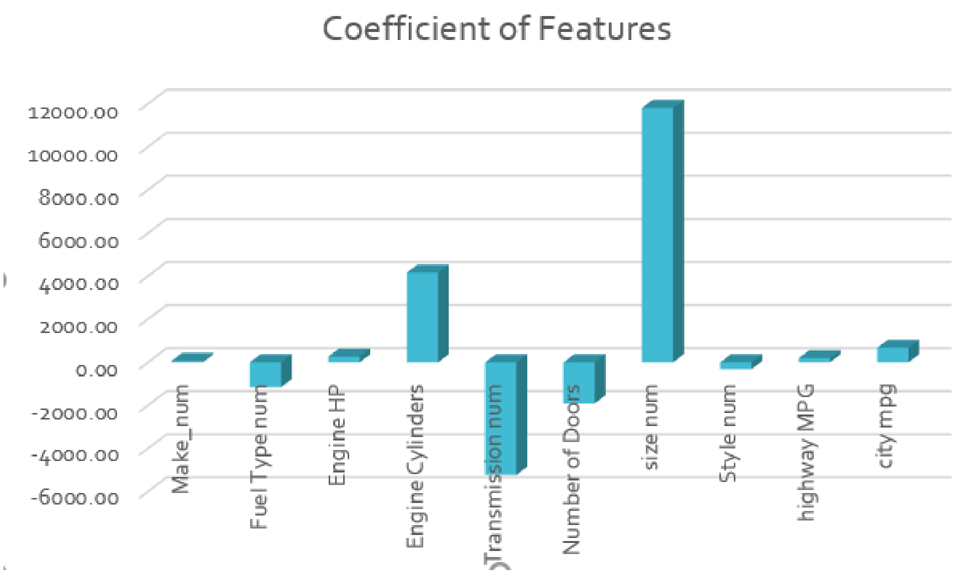
Regression Statistics	
Multiple R	0.755436349
R Square	0.570684077
Adjusted R Square	0.570282622
Standard Error	26823.64771
Observations	10705

There is not much difference in the R Square value but the p values are changed.



	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-64815.57	3463.960787	-18.71140417	0.0000
Make_num	67.45	21.82438807	3.090615272	0.0020
Fuel Type num	-1148.25	92.86564267	-12.36467139	0.0000
Engine HP	256.26	4.890791373	52.39698951	0.0000
Engine Cylinders	4155.96	320.2310575	12.9780161	0.0000
Transmission num	-5230.12	324.9150081	-16.09690056	0.0000
Number of Doors	-1905.57	346.6546848	-5.497020929	0.0000
size num	11813.97	429.8040452	27.48687814	0.0000
Style num	-313.97	80.24840689	-3.912533963	0.0001
highway MPG	185.42	65.96431915	2.810871567	0.0049
city mpg	668.76	75.63118667	8.842445226	0.0000

The below given is the chart of the Coefficients of the column to find the MSRP of the Car.



Insights:

- Transmission Type and the fuel type have the negative correlation with the Target.
- Popularity has no effect or negligible effect on the target value which is the price of the car.
- Car price does not change much with the Year.

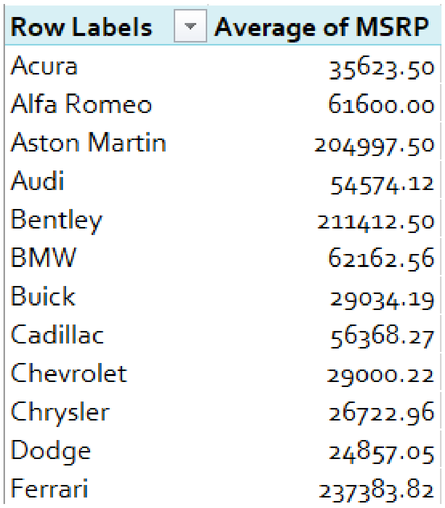
Business problem and recommendations:

- Collecting more information or doing feature engineering can increase the R Square value for the regression model.

<u>How does the average price of a car vary across different manufacturers?</u>

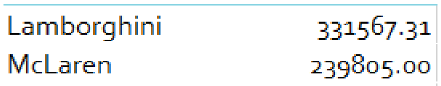
Task 4.A: Create a pivot table that shows the average price of cars for each manufacturer.

The below given is the few rows of generated pivot table with Car Brands and its Average MSRP.



Row Labels	Average of MSRP
Acura	35623.50
Alfa Romeo	61600.00
Aston Martin	204997.50
Audi	54574.12
Bentley	211412.50
BMW	62162.56
Buick	29034.19
Cadillac	56368.27
Chevrolet	29000.22
Chrysler	26722.96
Dodge	24857.05
Ferrari	237383.82

From sorting the table according the average MSRP we found that



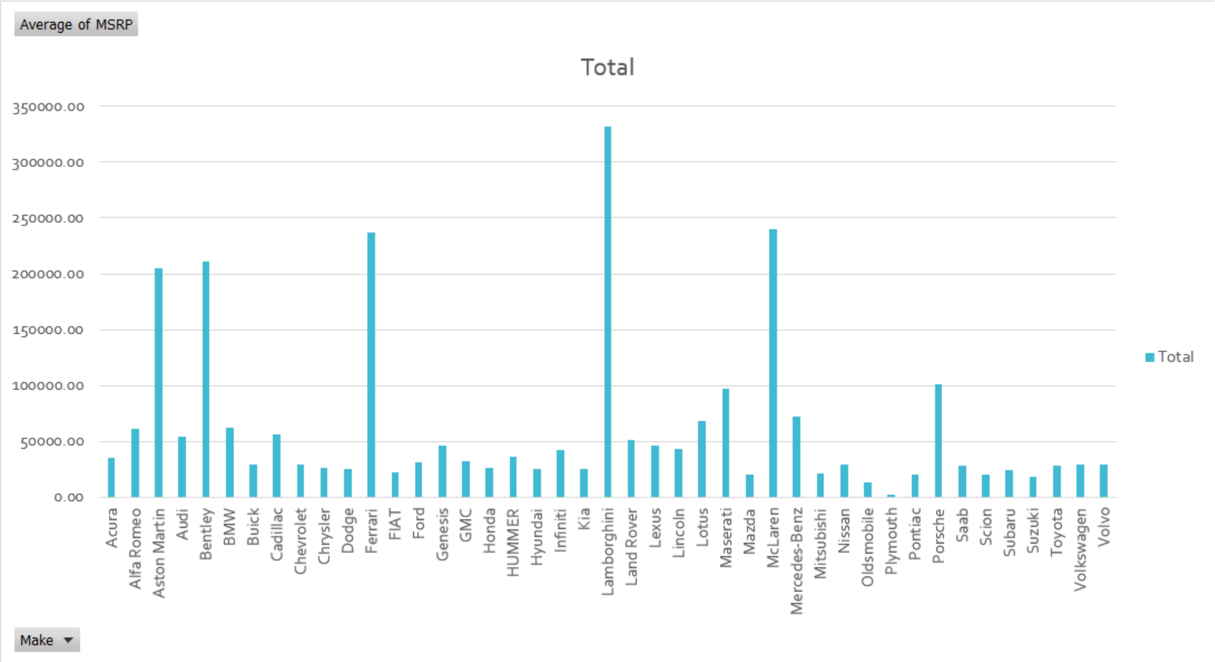
Lamborghini	331567.31
McLaren	239805.00

are having the highest average MSRP and

Oldsmobile	12843.80
Plymouth	2798.14

are having the lowest.

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



Insights:

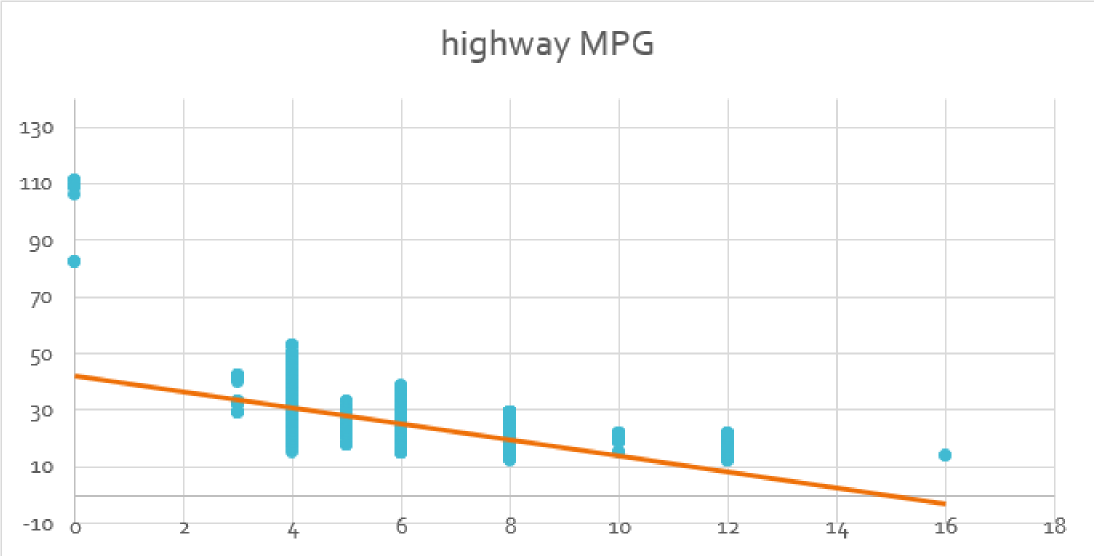
- As we can clearly see from the chart which brands have the highest price and which has the lowest.
- The weighted average is not considered any way.
- We can clearly see there is a lot of difference between the highest and the lowest price range.

Business problem and recommendations:

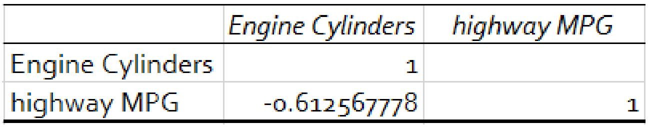
- From the average it is hard to tell recommendation we can take weighted average to better understand the data.
- The cars are either very expensive or very low priced. A range is not particularly followed.

<u>What is the relationship between fuel efficiency and the number of cylinders in a car's engine?</u>

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.



	Engine Cylinders	*highway MPG*
Engine Cylinders	1	
highway MPG	-0.612567778	1

Insights:

- We can see the correlation between Engine Cylinder and MPG is negatively correlated. Which means when one increases other will decrease.

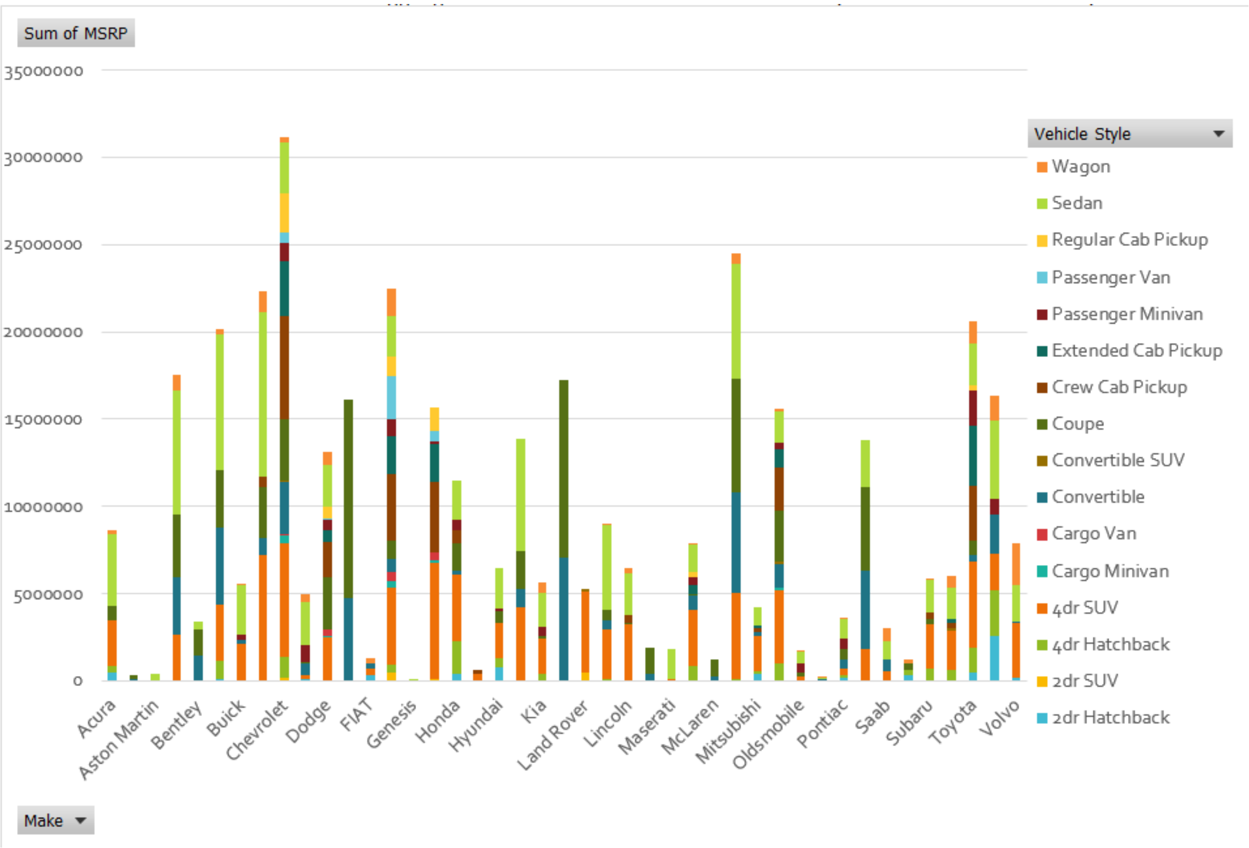
- The negative relation is clearly seen in the image above.
- The range of Correlation coefficient is between 1, -1 both included. If the correlation is 1 it is (positive) then strongly correlated and change of one variable increases the other variable. If the correlation is -1(negative) then strongly correlated and change of one variable decreases the other variable.
- If the correlation is 0 it is not correlated. If the value approaches 0 then has weak correlation.

Business Problem and recommendations:

- The optimum number of cylinders should be chosen to get maximum MPG. So the customer needs are satisfied during purchase of the car.

Building the Dashboard:

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



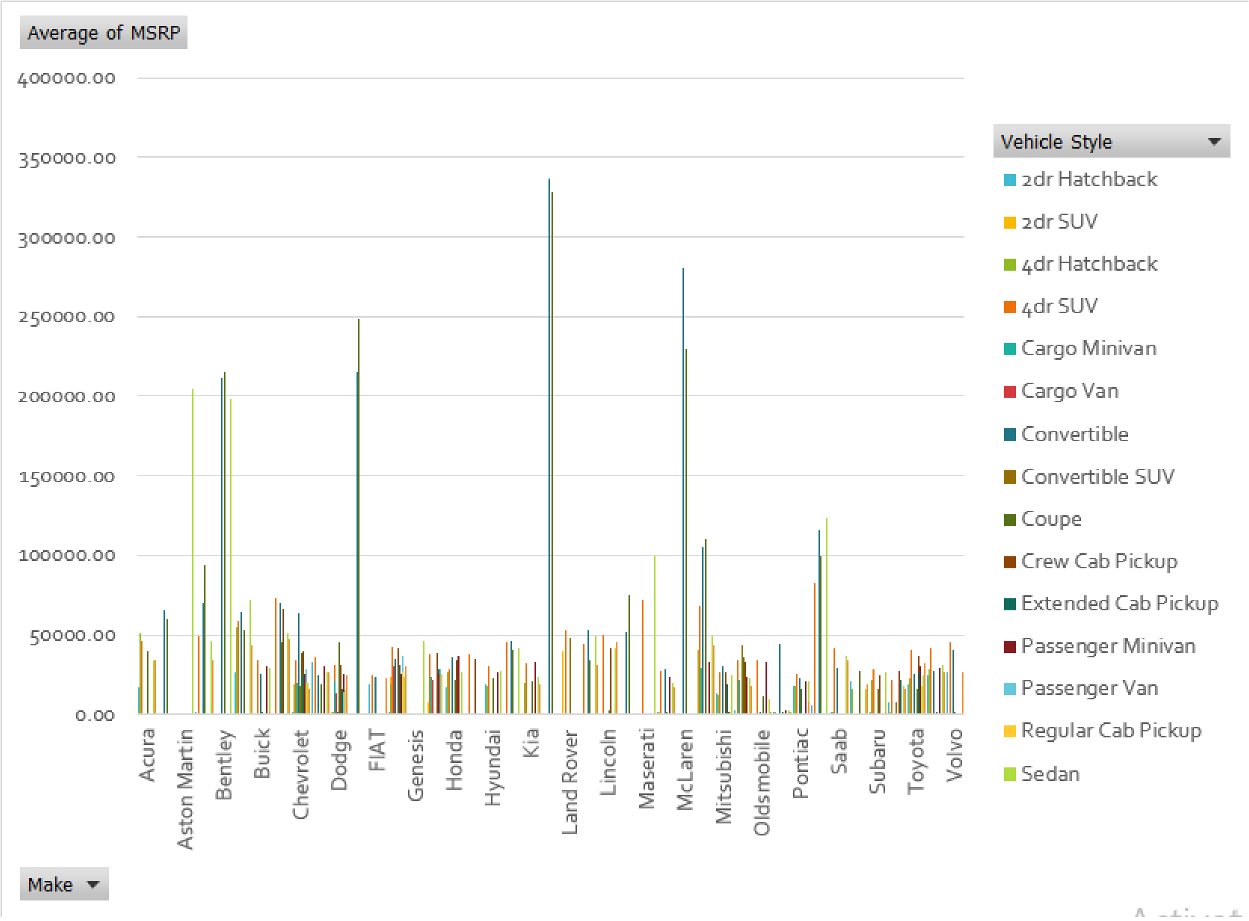
Insights:

- In few Car brands there only few or only one Model.
- In Most of the car brands many models are covered, which we can have the higher sum of MSRP.

Business problem and recommendations:

- Any car manufacturer should increase the number of models of car so there would be choice and customer preference criteria.

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



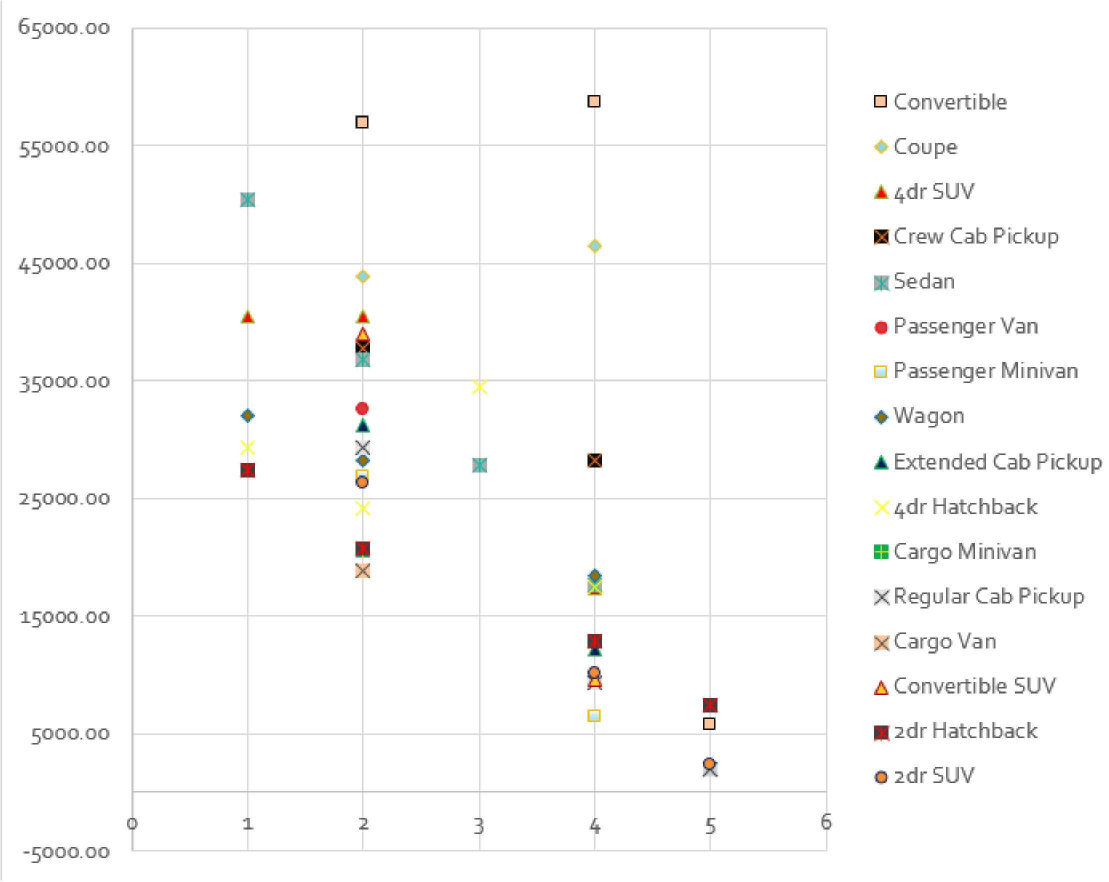
Insights:

- There is a huge difference between the Sum of MSRP of car brands according to models and their average MSRP. This is because the number of cars produced are counted in average.
- To get a better insight weighted average can be counted.
- Lamborghini is dominating other brands.

Business problem and insights:

- The Selection of brands we get the separate analysis of the model and each of its prices and the manufacturer can see what has to be improved in comparison.

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



Insights:

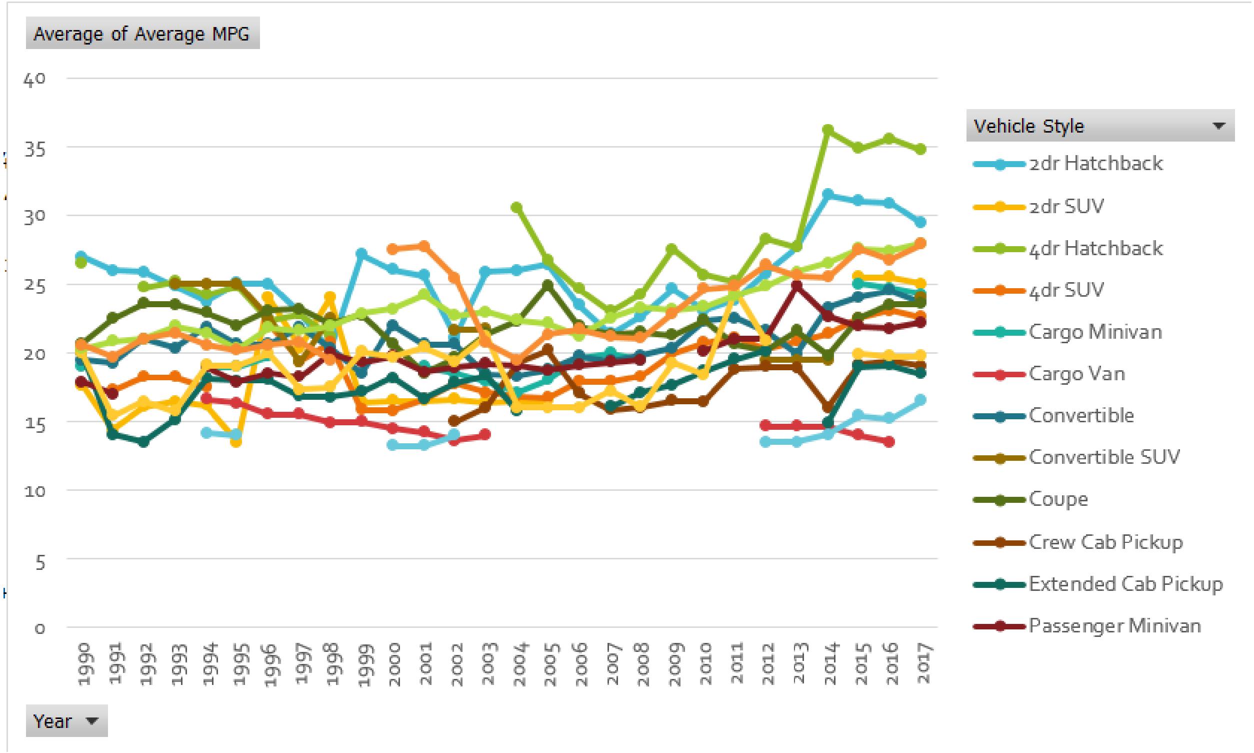
- This chart tells the relation between the Transmission type and the Body style of the vehicle.
- It was not possible to construct the scatter plot as we had category in both columns.
- To make that possible I have used pivot chat and filter after converting the columns to numerical column and selected each category and plotted each series.
- Different symbol for different body style is given.

AUTOMATED_MANUAL	1
AUTOMATIC	2
DIRECT_DRIVE	3
MANUAL	4
UNKNOWN	5

Business problem and recommendations:

- In Direct- Drive cars there are very few models.
- More number of models are there in Automatic and Manual.

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



Insights:

- There is a decrease in average price between years 2004-2010 and later is the increase in price.
- 4dr hatch back has increased more in one year.
- Few of the car models are not in production during few years.

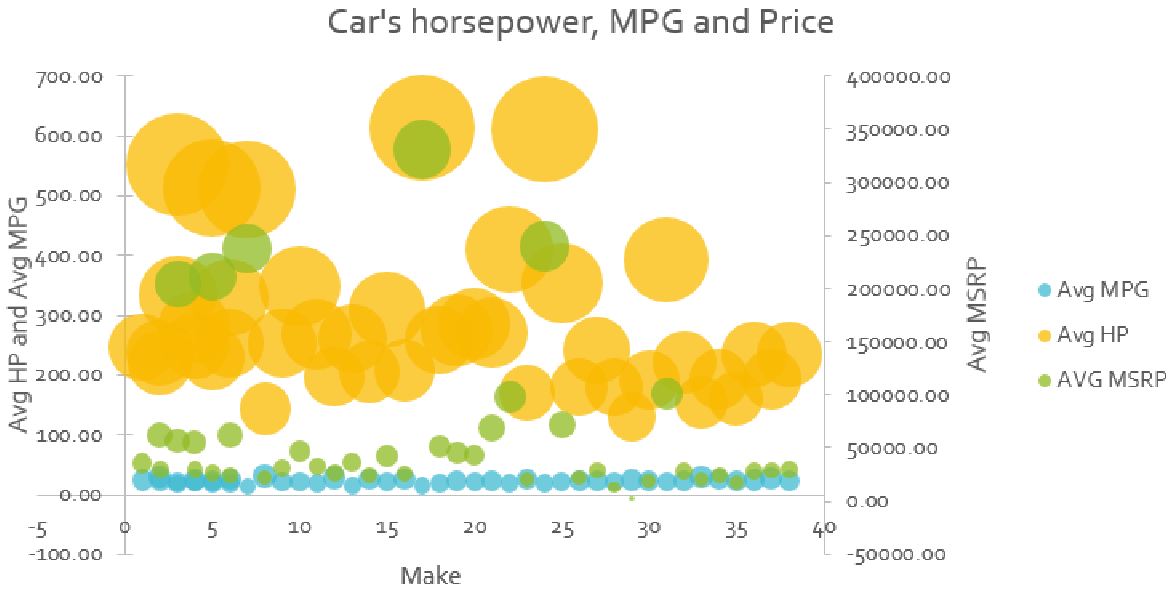
Business problems and recommendations:

- We see the discontinuity in the production or sale of the car models and there is decrease in the average price, these cars may be preferred less comparatively.
- Such models can be stopped.

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

To create this bubble chart, we have created the pivot table. Where the data we selected are Make, Make num, Engine HP (average of Engine HP), MPG (Average of MPG), MSRP (average of MSRP). We have selected these data as multiple series and created bubble chart as shown below.

In this chart in the primary y axis we have plotted Avg MPG and Avg engine HP in the secondary y axis we have plotted Avg MSRP. For the magnitude or the sizes of the bubble we have selected the same average values.



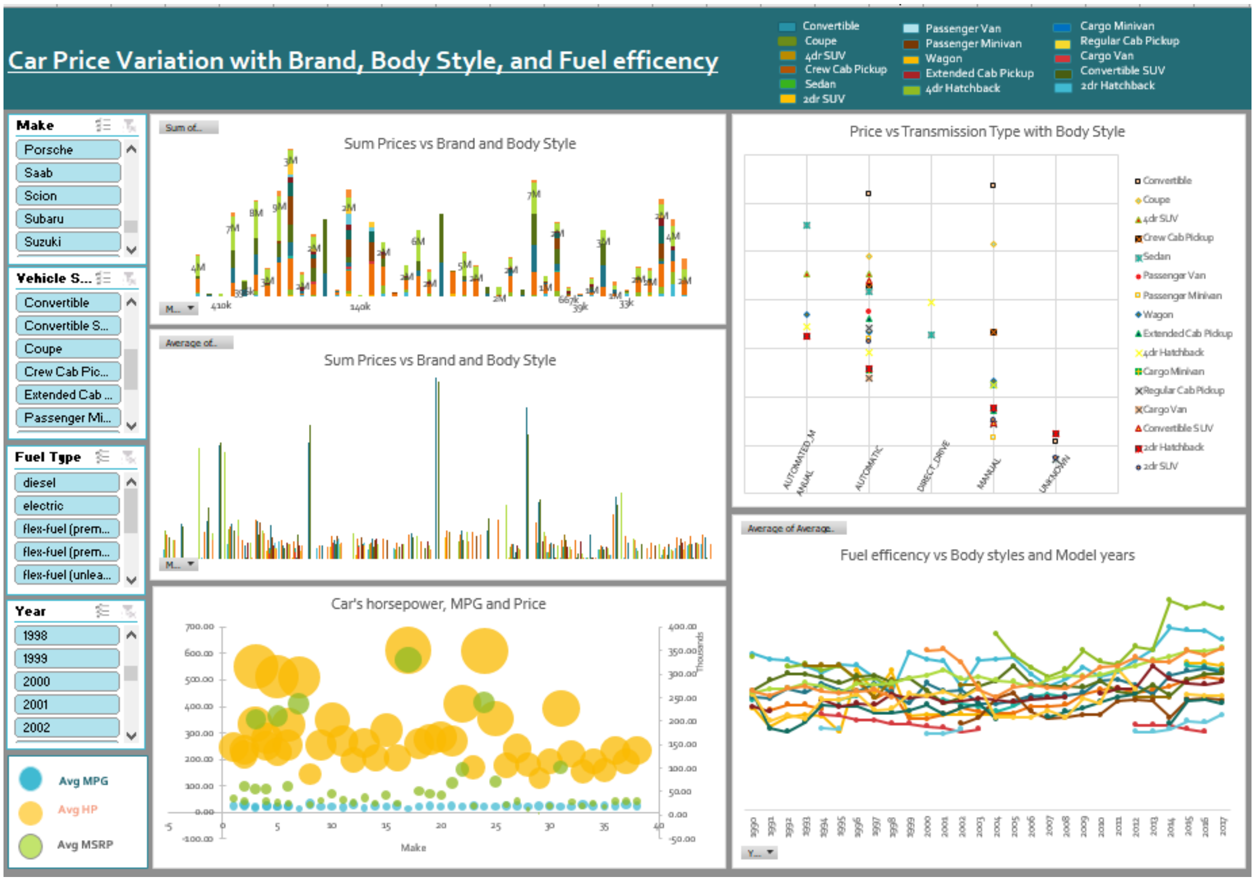
Insights:

- The size refers to the magnitude of the average, the plot is made between the value and the Make of the car.
- Colors represent the different series; I have plotted two axes for y (primary and secondary).

Steps to build the Dashboard:

1. Creating Individual charts to answer the client's questions is the first step. We have created multiple pivot tables and necessary charts from the tables.
2. Now, we have created a new sheet in the same workbook and have given appropriate heading to the dashboard.
3. We have then copied all the created charts to the new sheet and named this sheet as Dashboard, keeping this as the first sheet made it to be available easily.
4. Now inserting the appropriate filters is important. After which we will connect all these filters and slicers to all pivot tables.
5. Now we will do all the necessary changes to make the dashboard more readable.
6. Keeping all the necessary legends and axis, other axis and legends will be removed.
7. After arranging everything is an order. We will check if the dashboard is interactive enough.
8. Finally, all the color coding will be done and our dashboard is ready.

The below given is the dashboard made from the tasks we have completed.



Result:

All the visualizations like charts and dashboard, Analysis like Regression analysis, pivot tables are given above explaining the insights that we obtained during the project.

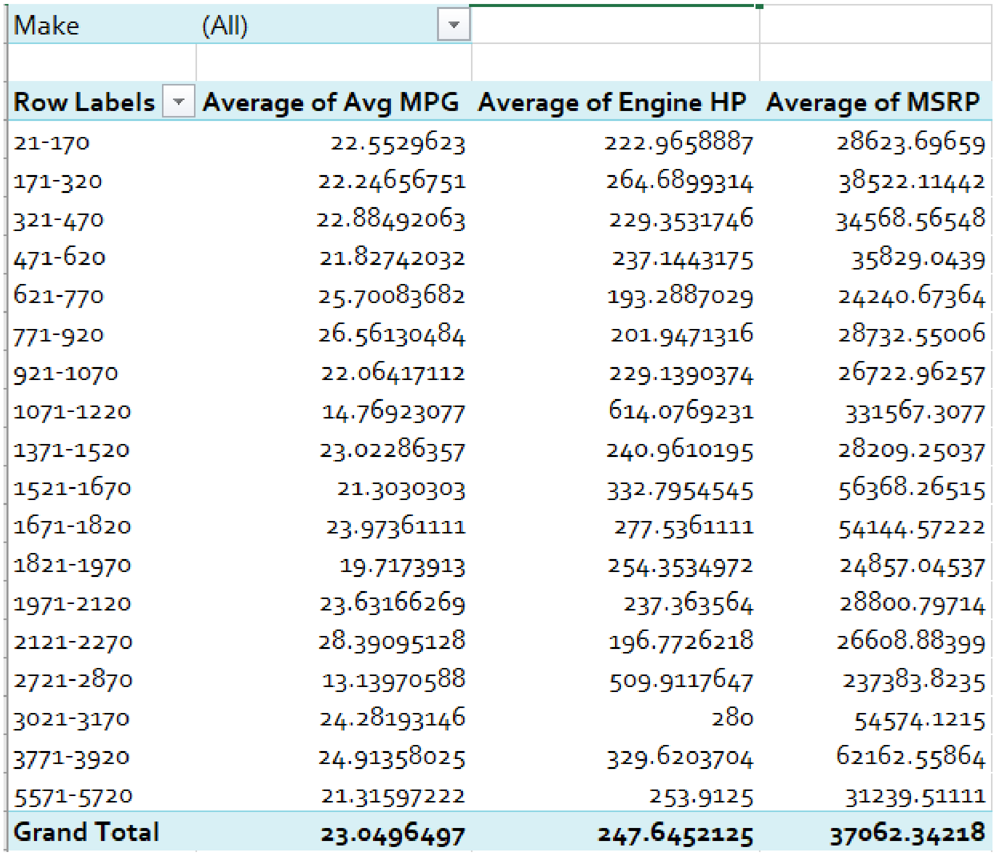
To understand the customer better, we have analyzed the popularity with other features like MPG, HP, MSRP with multiple brands. We have created a pivot table with these values and filter as the make.

It is as given below and have also created combo Chart to understand the relationship.

<u>Insights and recommendations to solve business problem:</u>

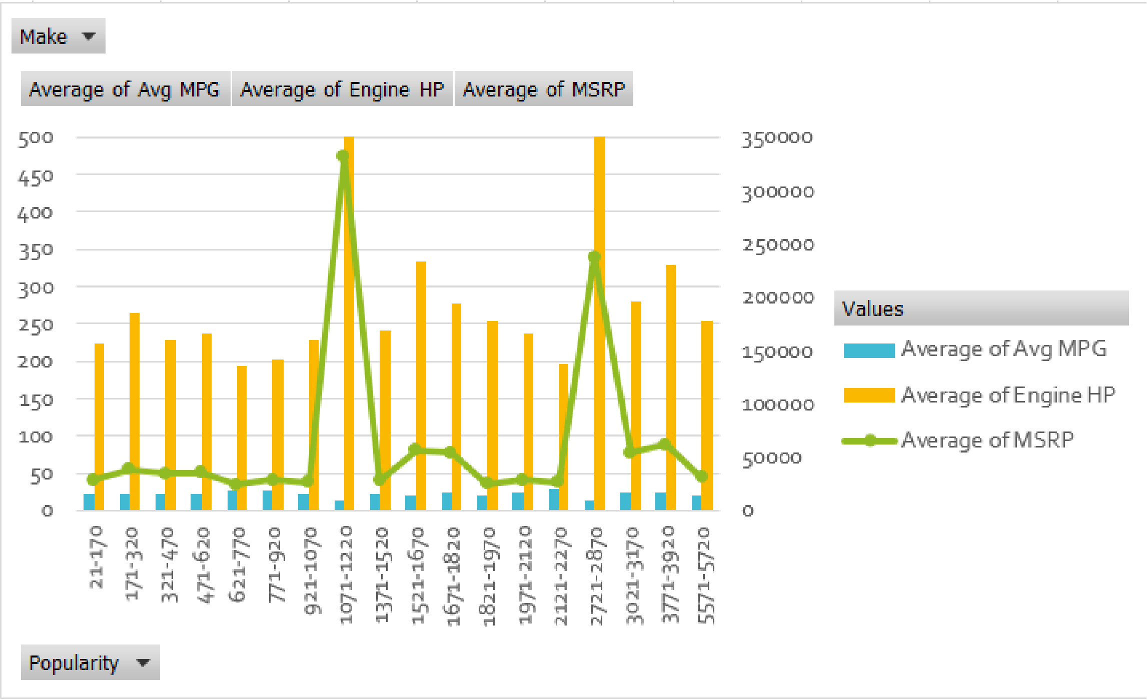
- The brands like BMW, Lamborghini have higher popularity.
- More number of popularity is for the cars having optimum number of MPG, HP, MSRP as we know more customers prefer them.
- Exotic cars can be made Pre Order.
- Cars with High performance and Diesel cars should be made in more number.

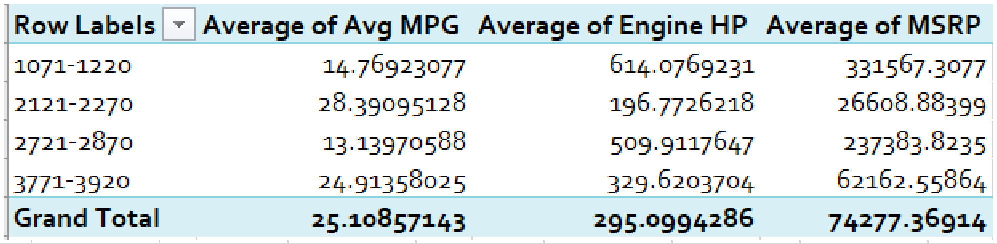
Few charts and tables are given below



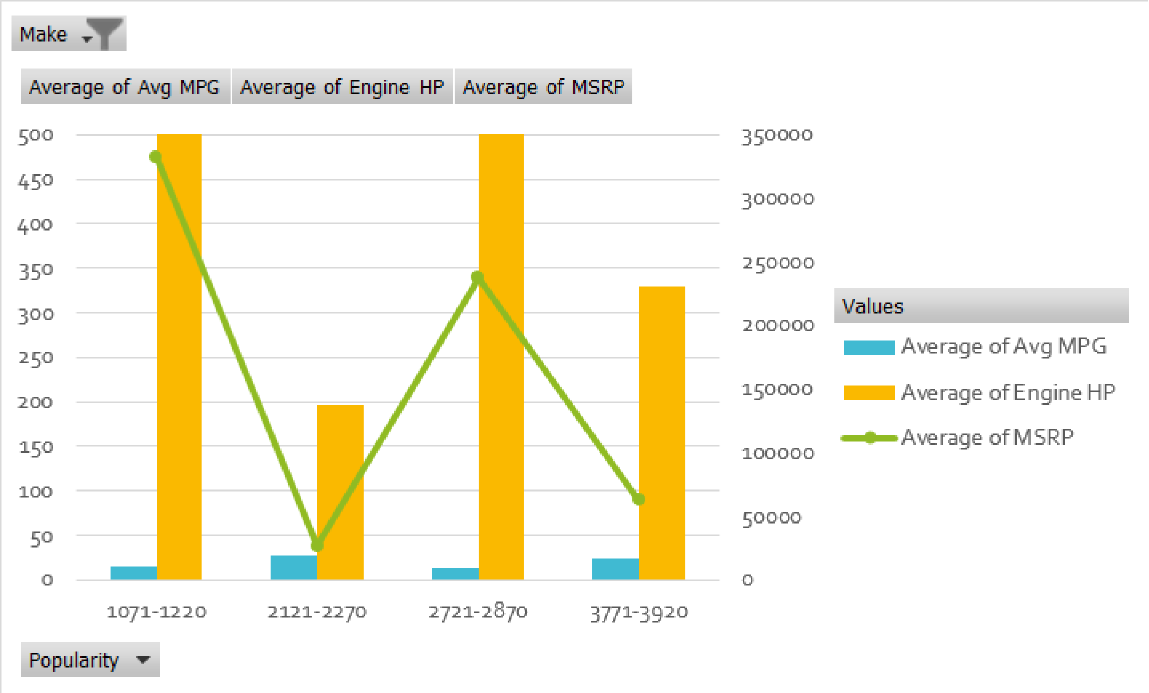
Make (All)

Row Labels	Average of Avg MPG	Average of Engine HP	Average of MSRP
21-170	22.5529623	222.9658887	28623.69659
171-320	22.24656751	264.6899314	38522.11442
321-470	22.88492063	229.3531746	34568.56548
471-620	21.82742032	237.1443175	35829.0439
621-770	25.70083682	193.2887029	24240.67364
771-920	26.56130484	201.9471316	28732.55006
921-1070	22.06417112	229.1390374	26722.96257
1071-1220	14.76923077	614.0769231	331567.3077
1371-1520	23.02286357	240.9610195	28209.25037
1521-1670	21.3030303	332.7954545	56368.26515
1671-1820	23.97361111	277.5361111	54144.57222
1821-1970	19.7173913	254.3534972	24857.04537
1971-2120	23.63166269	237.363564	28800.79714
2121-2270	28.39095128	196.7726218	26608.88399
2721-2870	13.13970588	509.9117647	237383.8235
3021-3170	24.28193146	280	54574.1215
3771-3920	24.91358025	329.6203704	62162.55864
5571-5720	21.31597222	253.9125	31239.51111
Grand Total	**23.0496497**	**247.6452125**	**37062.34218**





Row Labels	Average of Avg MPG	Average of Engine HP	Average of MSRP
1071-1220	14.76923077	614.0769231	331567.3077
2121-2270	28.39095128	196.7726218	26608.88399
2721-2870	13.13970588	509.9117647	237383.8235
3771-3920	24.91358025	329.6203704	62162.55864
Grand Total	**25.10857143**	**295.0994286**	**74277.36914**



Drive Links:

Excel link:
https://docs.google.com/spreadsheets/d/19WiDTTj4JvFYWB1xNmf5Am9SJPjeoWn8/edit?usp=sharing&ouid=106627214281497905501&rtpof=true&sd=true

Video Link:

https://drive.google.com/file/d/1dk1-DRP8cFUbQEYgGKiXgxmuf8E1F6a9/view?usp=sharing

Drive Folder:

https://drive.google.com/drive/folders/1-q2sEhS-yVkp9TtoccM2Q1a2nXYbdadN?usp=sharing