**Project Description:**

With the dataset provided for car features, my job is to analyze it and provide insights that can help the car manufacturers regulate the prices for maximum profits while also ensuring that consumer demands are met. In the later pages of this report, I will apply regression analysis and market segmentation on the data which will help the car manufacturers understand what car features and categories are the most popular among the customers using which they can set their prices. The approach and the corresponding insights and results will be listed down in detail in the coming pages of this report.

**Tech-Stack Used:**

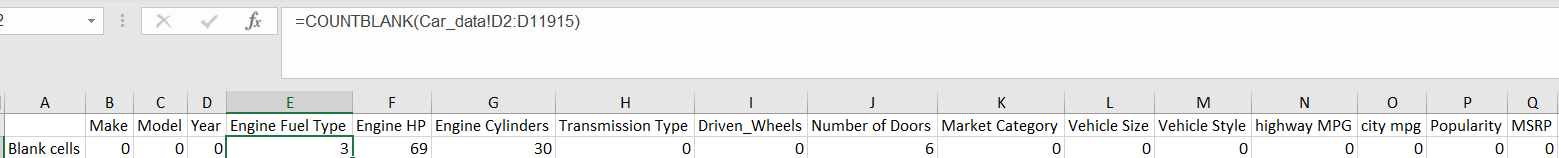
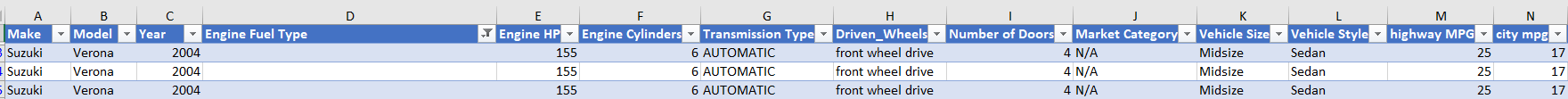
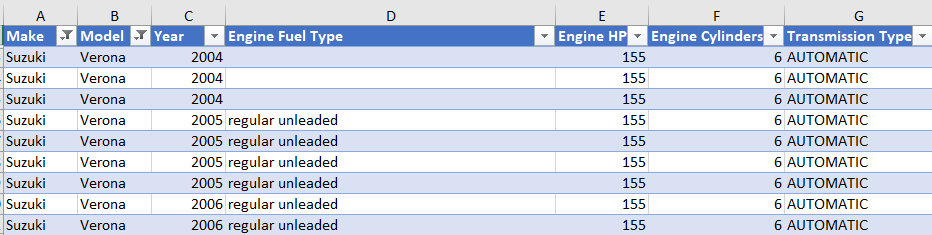
To implement this project, I have used Microsoft Excel 365 Version 2302. There are two reasons for this:

* MS Excel 365 allows a lot of features that are not available in the previous version such as there are a lot of automated charts which help me to quickly use them on my data without having to perform the calculations manually.
* This subscription is offered to me by the company I currently work in, so I am quite familiar with its ins and outs.

Kindly download the dataset from the link below and view in Excel to see all the charts and tables for the problems (Google sheets doesn’t show all the analytics done). In total there are 7 sheets in this workbook.

[**ImpactOfCarFeatures\_analyzed**](https://drive.google.com/drive/folders/1QU0AOdSG5DacoOpEnEA0JV3Up6hWj7cM?usp=sharing)

**Data cleaning:**

1. Create a table from the raw dataset given.
2. To check the number of blank cells for each variable, I have used the COUNTBLANK function.  
   
3. We have missing data for 4 variables.
4. The engine fuel type is missing for Suzuki Verona.  
   
5. From the other years’ releases, we can see that it is ‘regular unleaded’ and hasn’t changed for the two years with the same Engine HP and number of cylinders. Hence, I will fill the missing cells for the year 2004 also with regular unleaded.  
   
6. We will go one by one for each make and their corresponding model and year details and fill the HP as we did for the engine fuel type for Suzuki Verona above.
7. To fill the 30 missing cells for Engine Cylinders, I have done my research and used the data available on web (some sources given at the end of this section) for my dataset or simply comapred the data available in the dataset for the other models and makes.
8. For example, we can see that electric vehicles have 0 cylinders, so I have filled 0 for all the missing cells under engine cylinder whose engine fuel type is electric.
9. Note that the **engine hp is not consistent** and accurate therefore I had taken the average of the values throughout the selection I had gone through on the web.
10. For the cars where the market category is NA, I have kept as it is. In my analysis later, I have ignored the records associated with these.
11. **Note that after filling all the missing cells, the COUNTBLANK function in the “DataCleaning” sheet returns 0 for all the features since the missing cells are not empty anymore.**

Sources:

<https://www.automobile-catalog.com/#gsc.tab=0>

<https://www.auto123.com/en/new-cars/technical-specs/tesla/model-s/2016/base/70/>

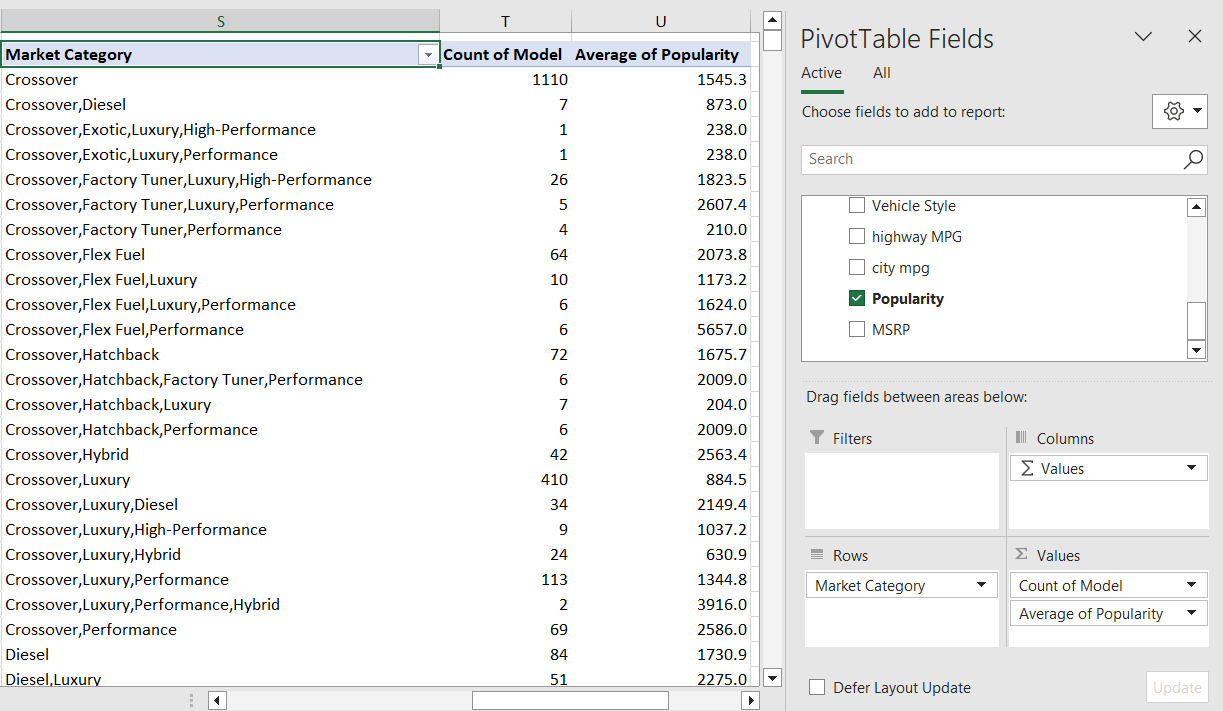
Problem Description No. 1:

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.

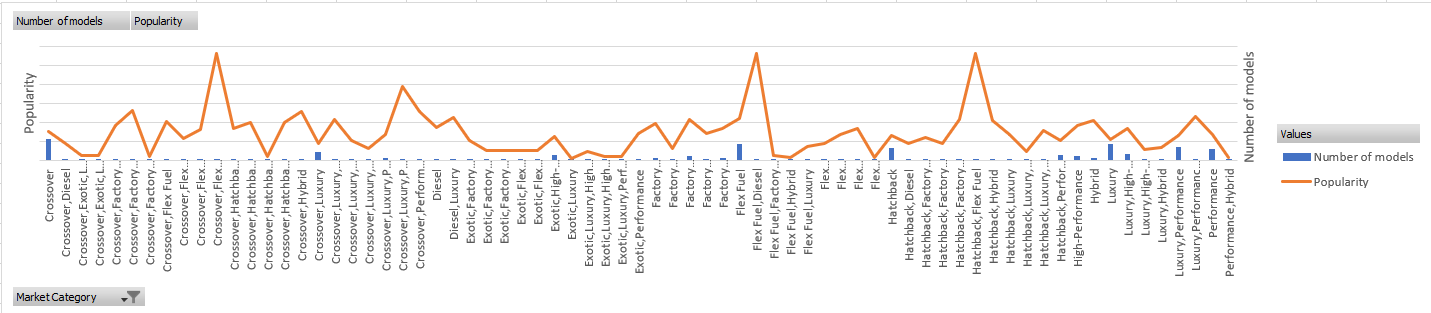
Task 1.B: Create a combo chart that visualizes the relationship between market category and popularity.

**Approach:**

1. I have created the Pivot table using the option available in excel and selected the fields shown below:  
   
2. To demonstrate the popularity score, I have used the average for each market category.

**Insights:**

1. From the chart, we see that there are 3 popular market categories with the same average popularity score of around 5657.



1. It is also evident, that for the popular market categories, the number of models is lesser as compared to the less popular market categories.
2. There is a market category which is not available. I have filtered that away in the pivot table.

Problem Description No. 2:

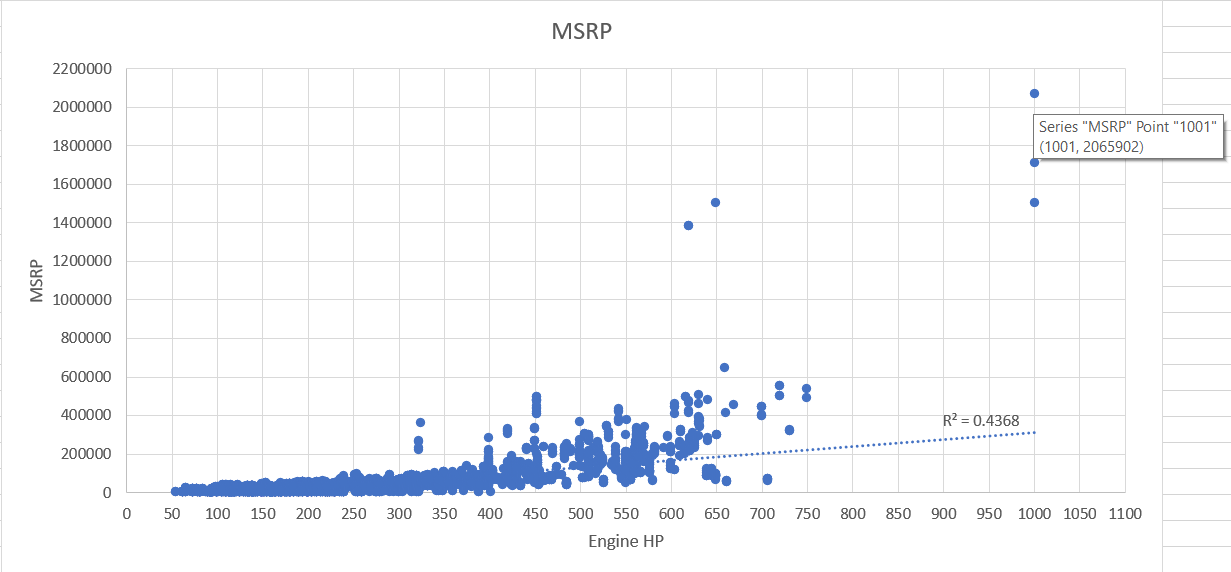
What is the relationship between a car's engine power and its price?

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.

**Approach:**

1. I have selected the engine hp and the MSRP columns and inserted the scatter plot chart.

**Insights:**

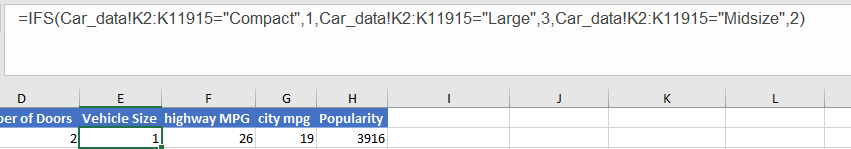
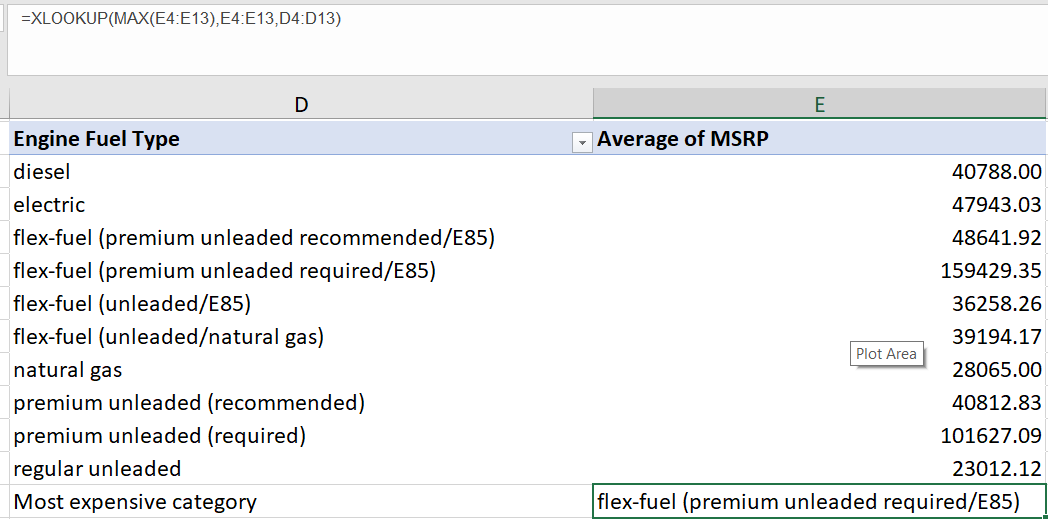
It is evident from the trendline that as the engine hp increases so does the maximum suggested retail price of the car.  
We have an R-squared value of around 0.4 which is neither too close nor too far away from 1 which means we can rely on this relationship only 50% of the times. In other words, the horsepower is a defining factor for the MSRP of a car but not the only one.  


Problem Description No. 3:

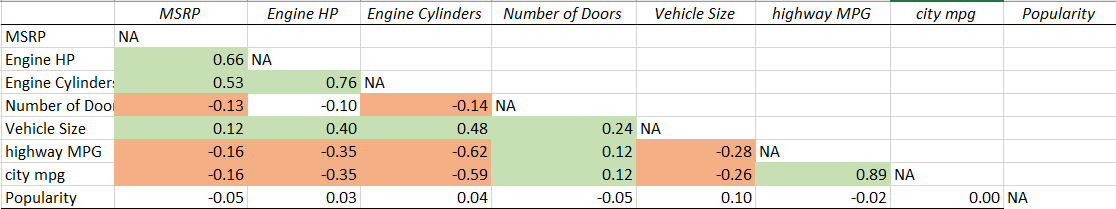
Which car features are most important in determining a car's price?

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.

**Approach:**

1. The variables which can affect a car’s price have been collected in the sheet “CorrelationWithMSRP”.
2. For the textual variable – Vehicle Size I have used 1 for Compact, 2 for Midsize and 3 for Large so that these numbers can indicate the growing size and replaced the category with this number with the help of IFS function:  
   
3. After this, I have created the correlation matrix for these variables using the “Data Analysis” add-in.
4. For analysis of one variable with the remaining, I have replaced the correlation (1.0) of the variable with itself with NA.
5. Finally, I have highlighted the top and bottom 10 correlations using the conditional formatting option.
6. For the other textual variables, I have created pivot tables in the sheet “TextualVariableRelationWithMSRP” and determined what is the most expensive category within these variables using the XLOOKUP and MAX functions.  
   

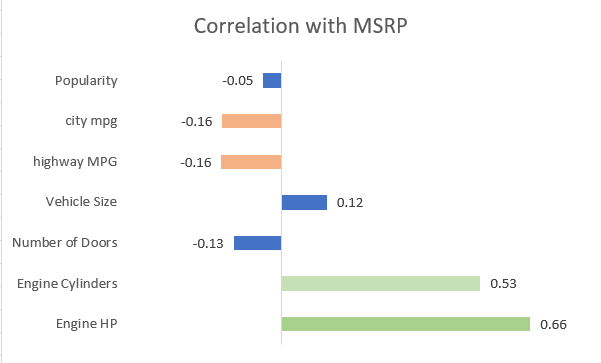
**Insights:**



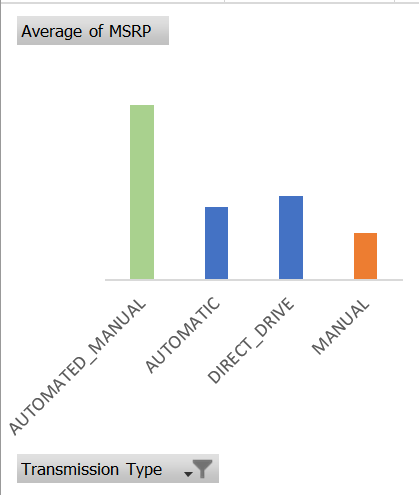
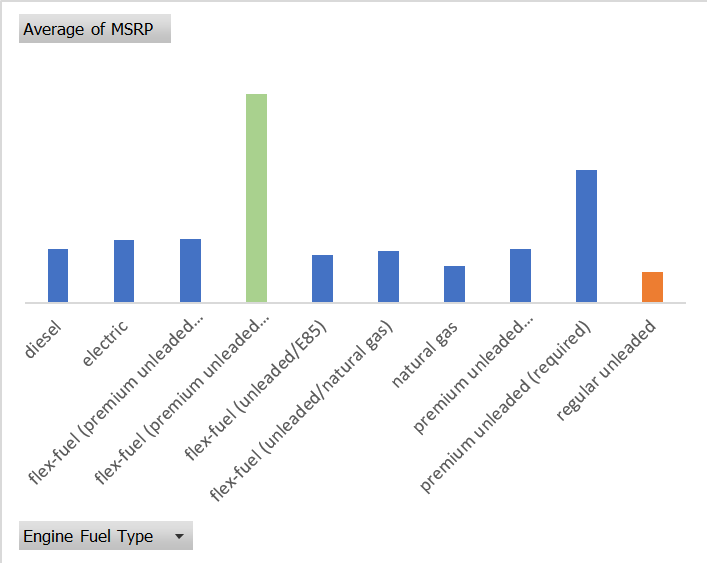
From the above correlation matrix, it is clear that the engine hp is the most significant variable which determines the MSRP of a car. Next in line is the number of engine cylinders.

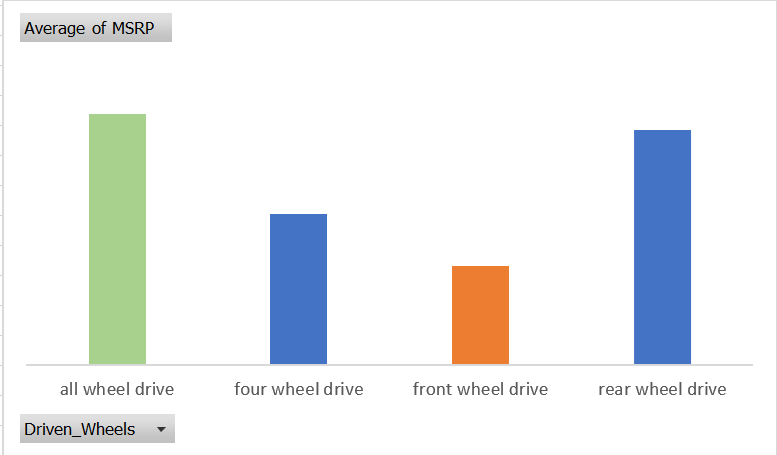
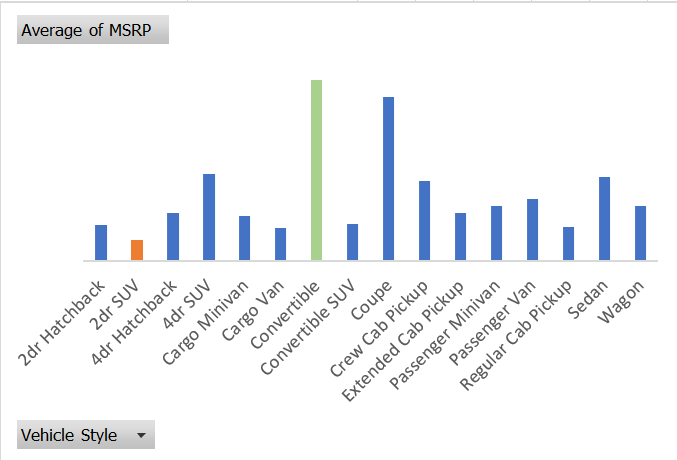
Another important thing is to note that there is a negative correlation between the popularity of the car and its MSRP which means that more popular cars have a lower MSRP whereas less popular cars have higher MSRP. **But this is not a reliable finding since the value is way less than -1.0**

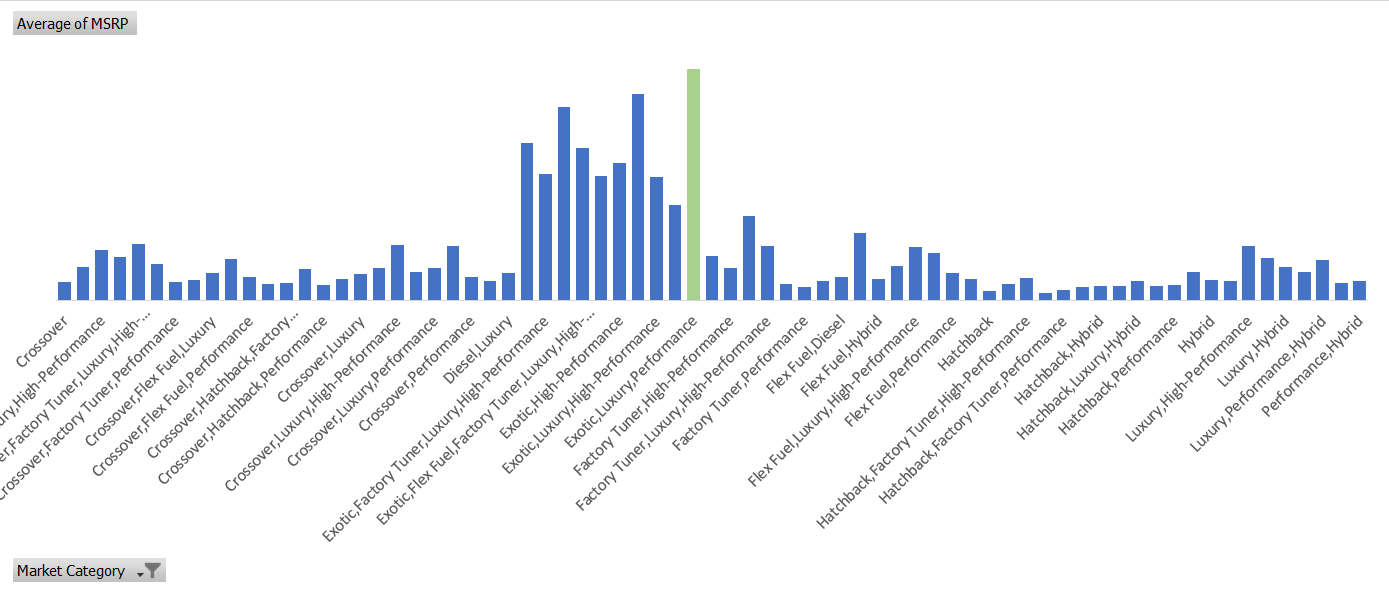
The vehicle size also while it has a positive correlation is way too far away from the value 1.0 which is why it’s not a valuable factor in determining the MSRP.

The bar chart demonstrated below further enforces the above findings.  


For the textual variables, cars with transmission type as both automated and manual, engine fuel type as flex-fuel (premium unleaded required/E85), driven on all wheels, are under the market category “Exotic,Luxury,Performance” and are of Convertible style have the maximum average MSRP.   
The results have been demonstrated below:



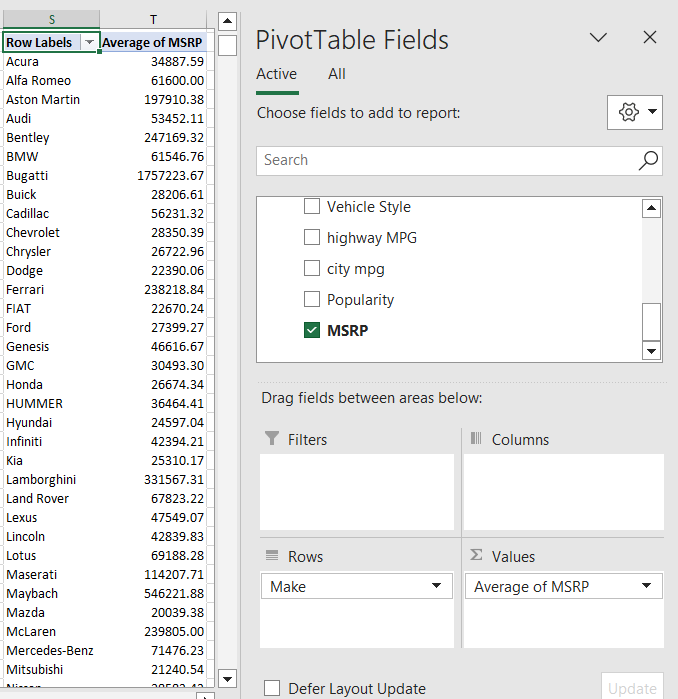
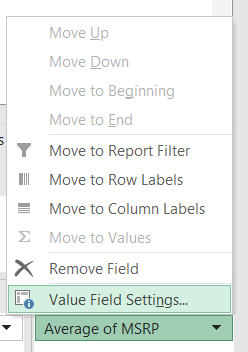
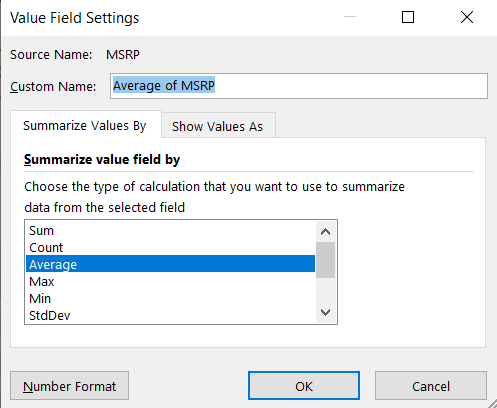
Problem Description No. 4:

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.

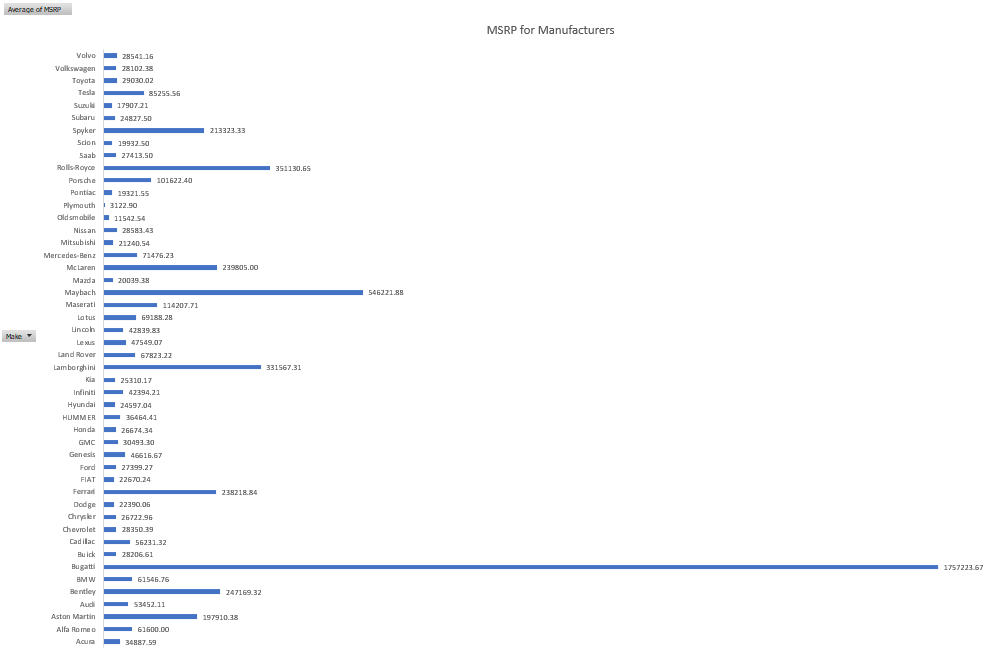
Task 4.B: Create a bar chart or a horizontal stacked bar chart that visualizes the relationship between manufacturer and average price.

**Approach:**

1. Create the pivot table and select the required fields as shown below:  
   
2. By default, the values section shows the count of the variable selected. This can be changed by going inside the “Value Field Settings” option.  
    
3. Next, select the 2D bar chart to visualize the average MSRP for each manufacturer.

**Insights:**

From the chart, it is evident that Bugatti has the highest average MSRP of around 17 lakhs for its cars and there’s no other manufacturer even close to the prices set by Bugatti.



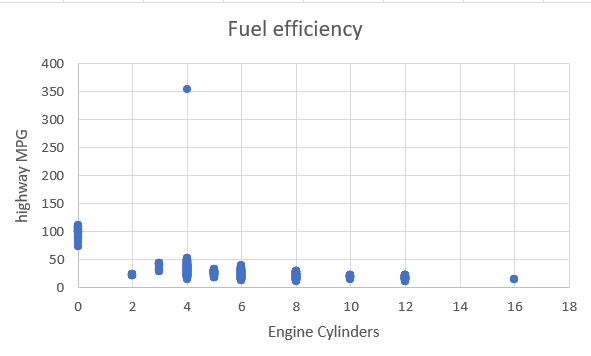
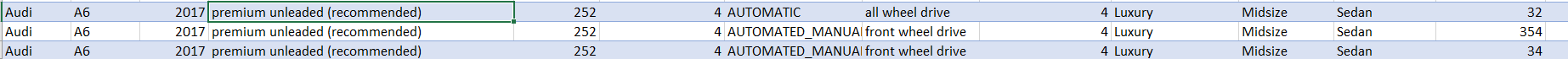
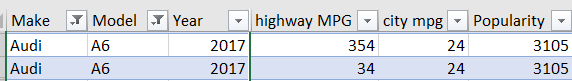
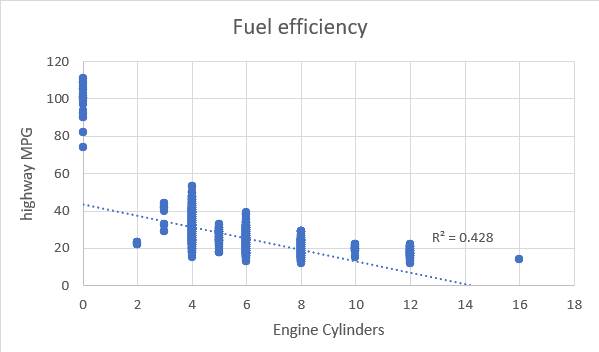
Next in line are Maybach, followed by Rolls-Royce, Lamborghini, Bentley, and McLaren.

Problem Description No. 5:

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.

**Approach:**

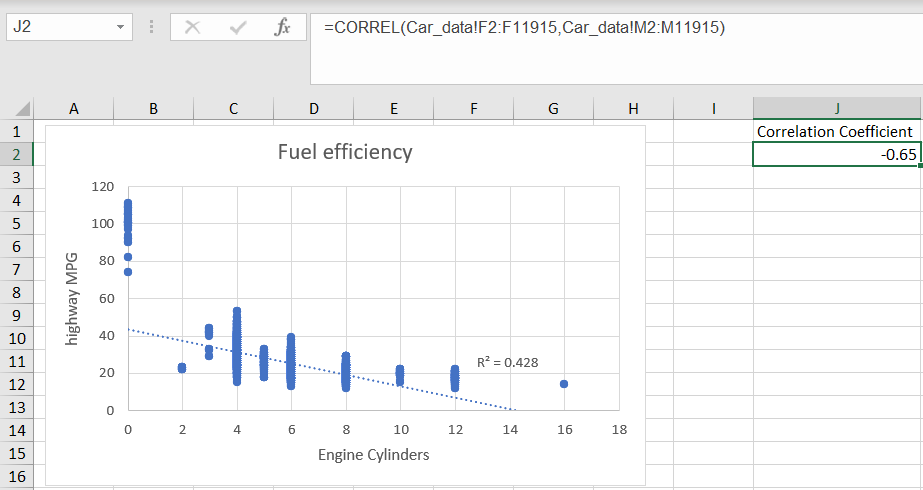
1. Select the columns “Engine Cylinders” and “highway MPG” and create the scatter plot chart in the sheet “CyclindersCountVsFuelEfficiency”.
2. From the chart created we can see that there is one MPG value (354) which is too high as compared to the rest of the MPG values.  
   
3. When I checked in Car\_data, I found that this value belongs to Audi A6. I filtered out all the Audi A6 records and found the following:  
   
4. **We can also see that the city MPG for this outlier record is only 24 which proves that 354 is a discrepancy.**  
   
5. The other records have highway MPG 32 and 34. Therefore, **I will replace 354 with the other Audi record for front wheel drive, i.e., 34.**
6. Now our scatter plot looks more reliant.  
   

**Insights:**

1. Following the trendline we can see that there is a negative slope or in other words a negative relationship between the highway MPG and number of engine cylinders. Since, the R-squared value of around 0.4 is neither too far away nor too close to 1.0, we will calculate the correlation coefficient to confirm this result.
2. **As the number of cylinders in a car increases, the highway MPG offered by it decreases.**
3. **Electric cars having 0 engine cylinders are offering the highest highway MPG values.**

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

**Approach:**

Use CORREL() excel function between the columns “Engine Cylinders” and “highway MPG”.  


**Insights:**

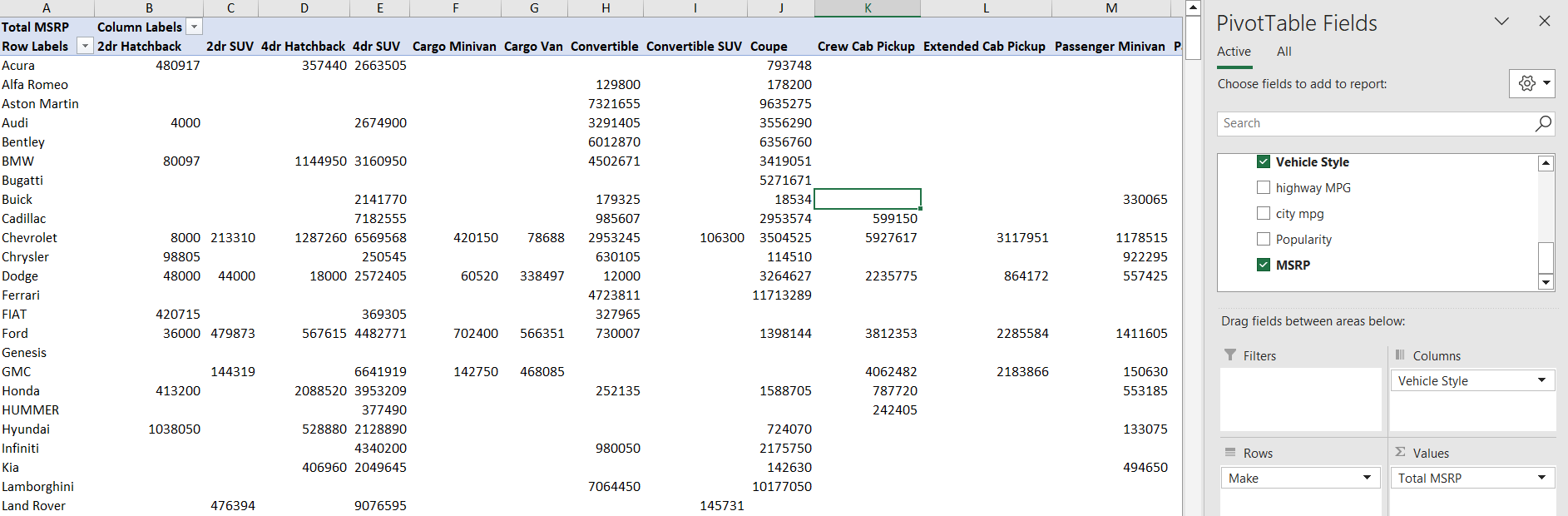
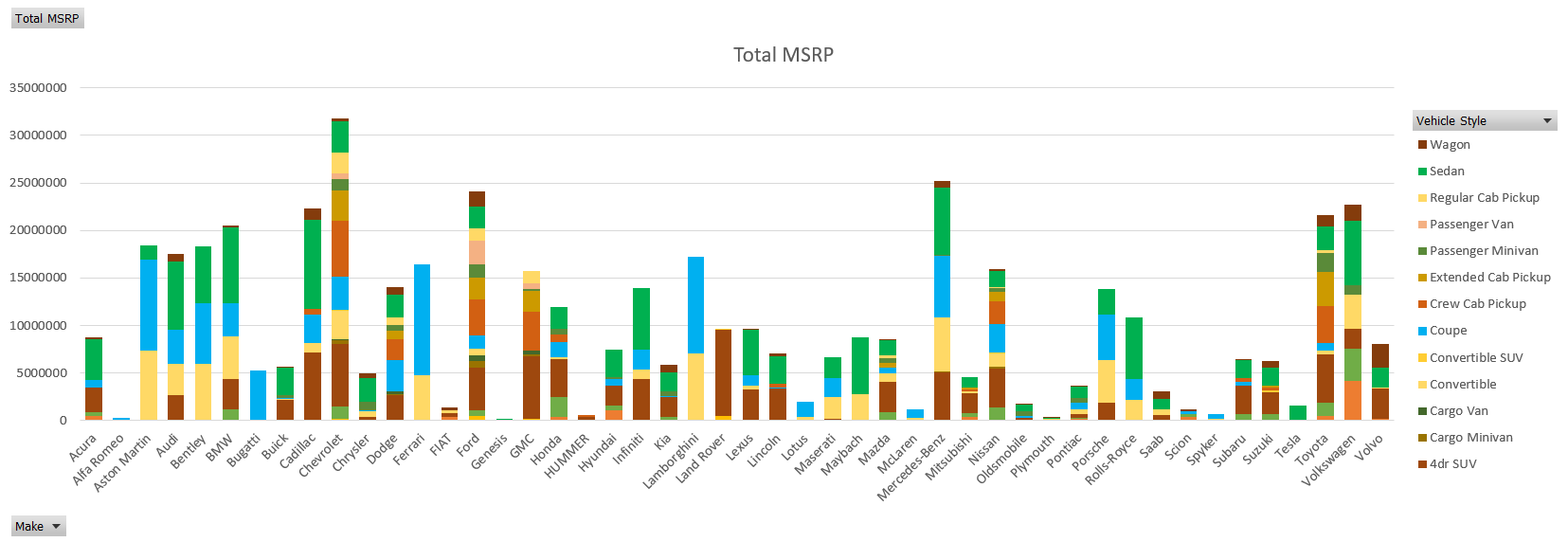
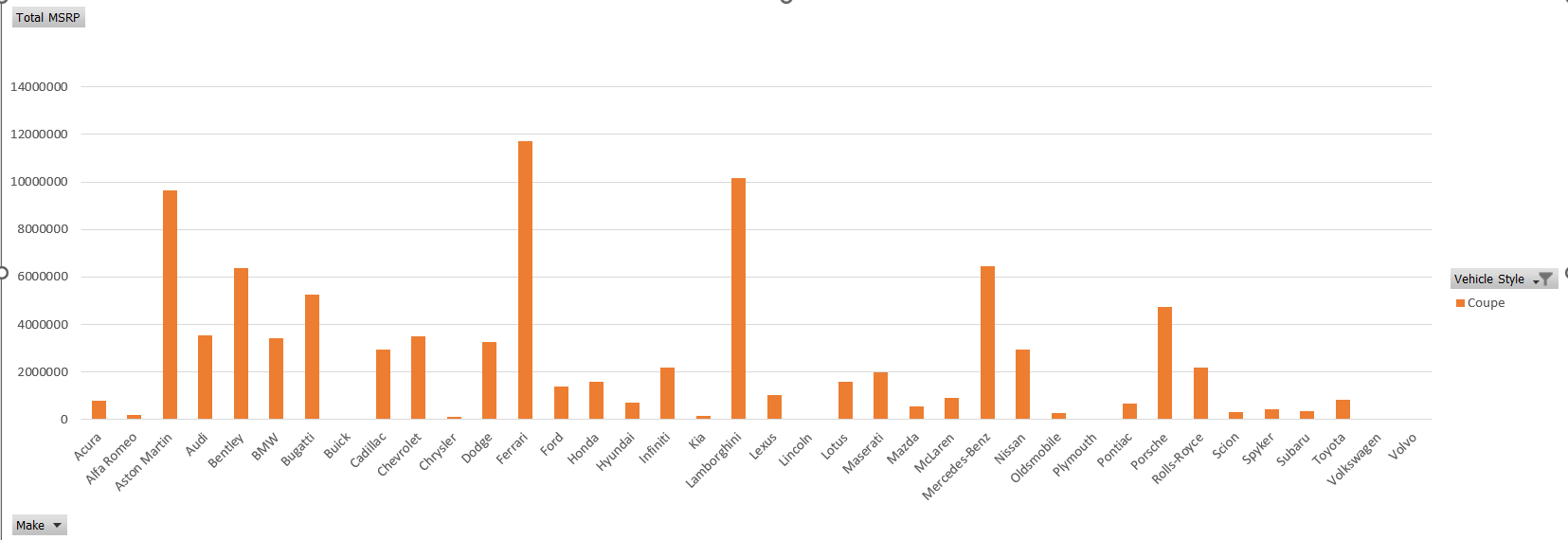
The correlation coefficient is a negative value of -0.65 which further strengthens our finding from above that there is a negative relationship between the number of engine cylinders and the highway MPG. Since the value is close to -1.0, we can easily rely on it too.

Dashboard Creation:

Problem Description No. 1:

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

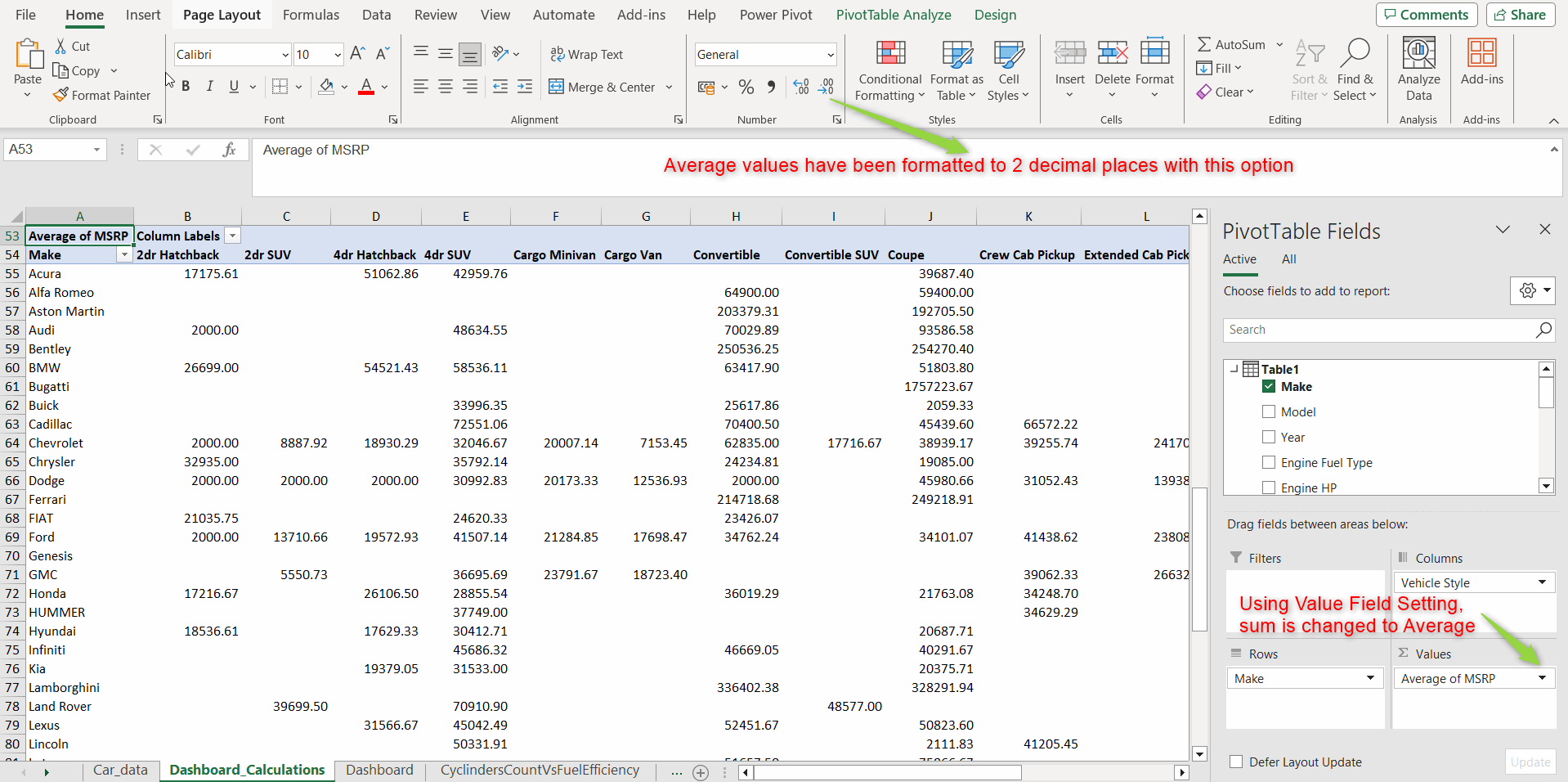
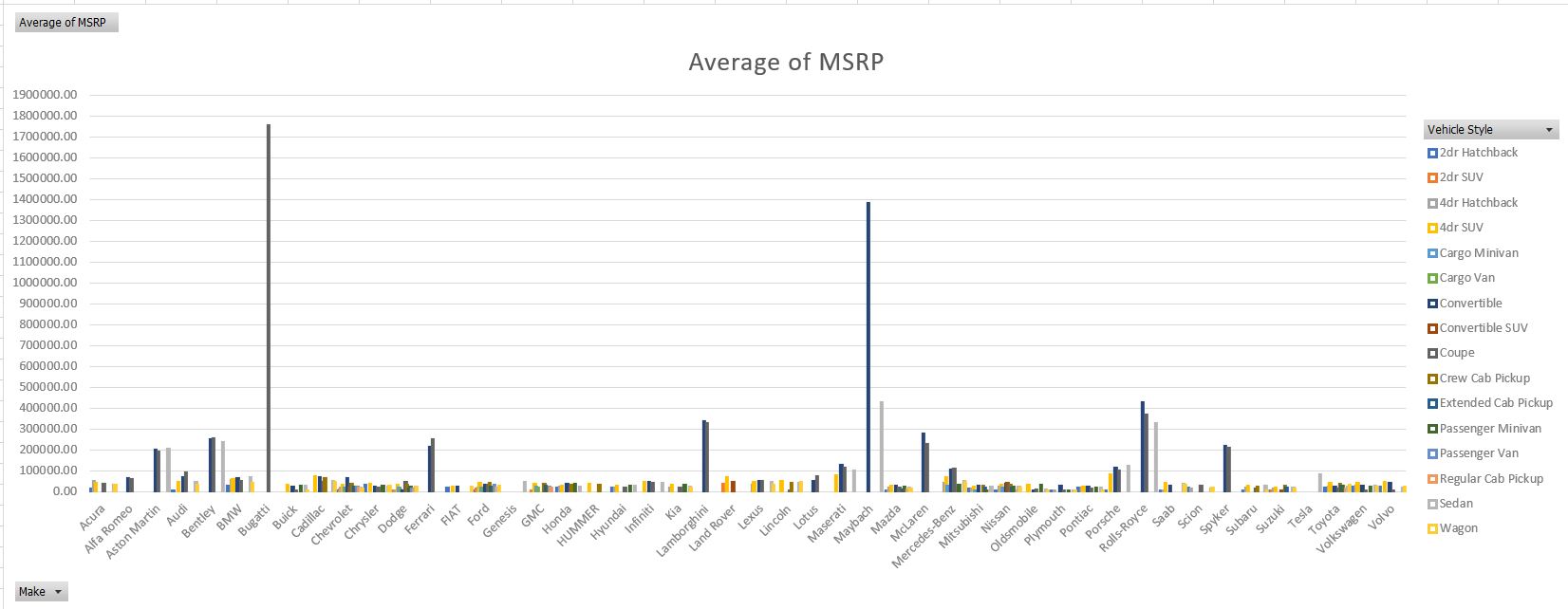
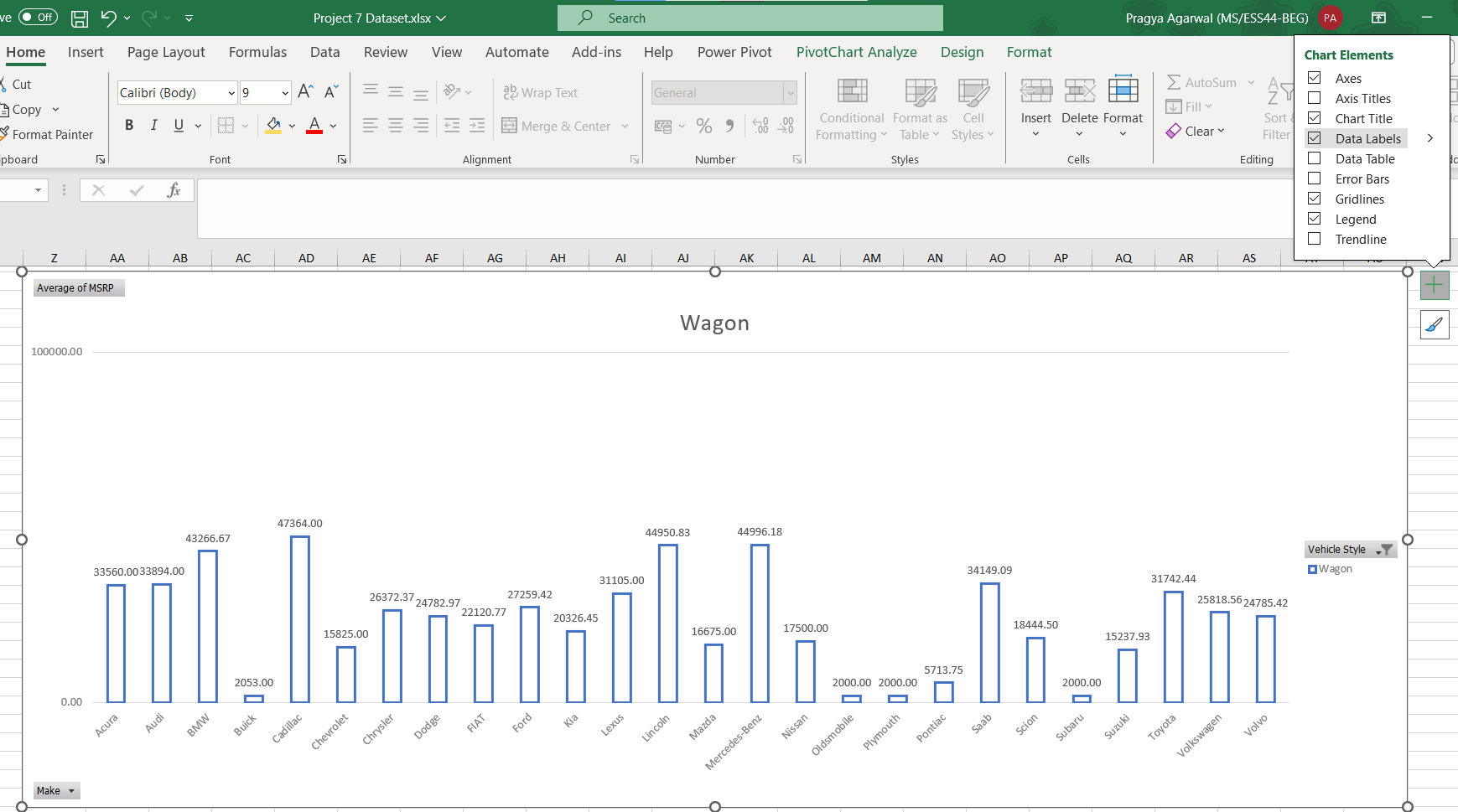
**Approach:**

1. Select the table in the “Car\_data” sheet and create a pivot table in the sheet “Dashboard\_Calculations” with the following fields:  
   
2. Select the stacked column chart for this pivot table, do some formatting and move it to the “Dashboard” sheet.  
   
3. From the first look at the chart, we can see that for almost all the brands, Sedan is the most popular style.
4. We can filter out the Make and Vehicle Style with the filters present in the chart.
5. For example, if we want to see the total MSRP for the vehicle style Coupe, I can select the filter available on the right-hand side of the chart. This is the result:  
   

Problem Description No. 2:

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

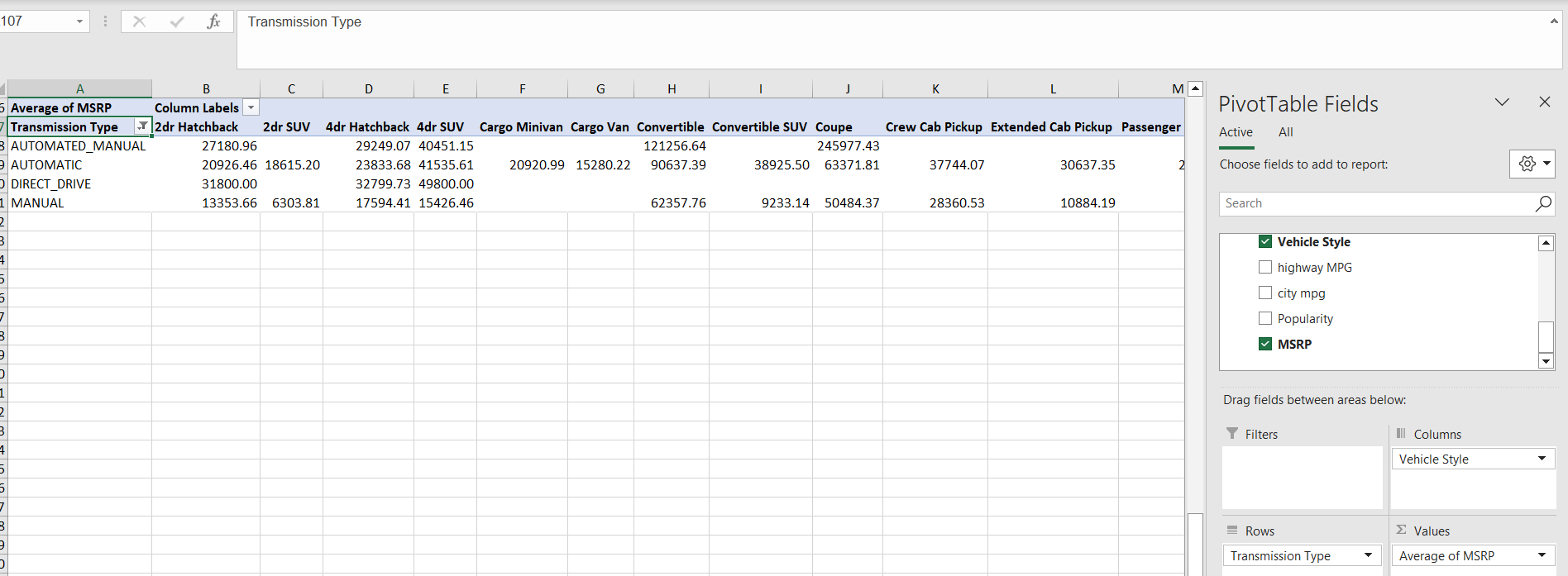
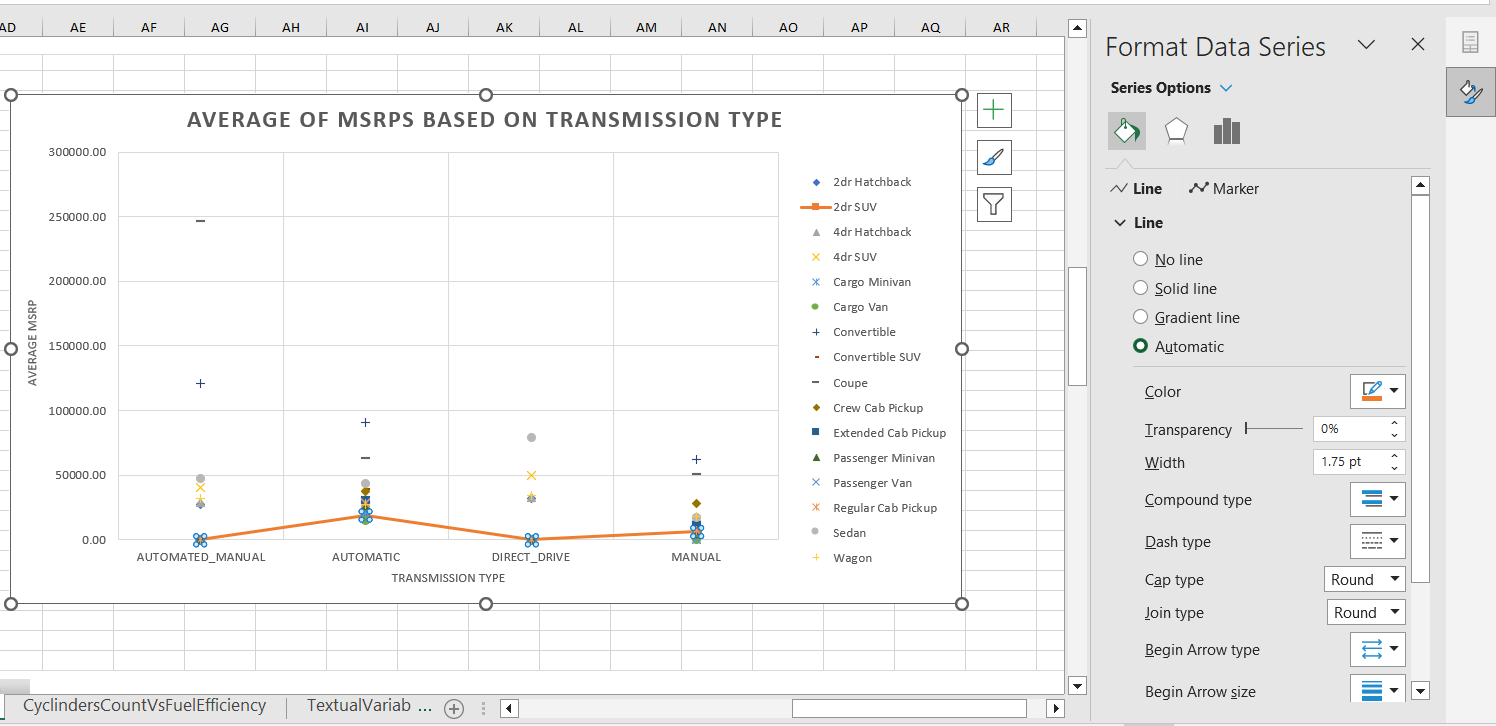
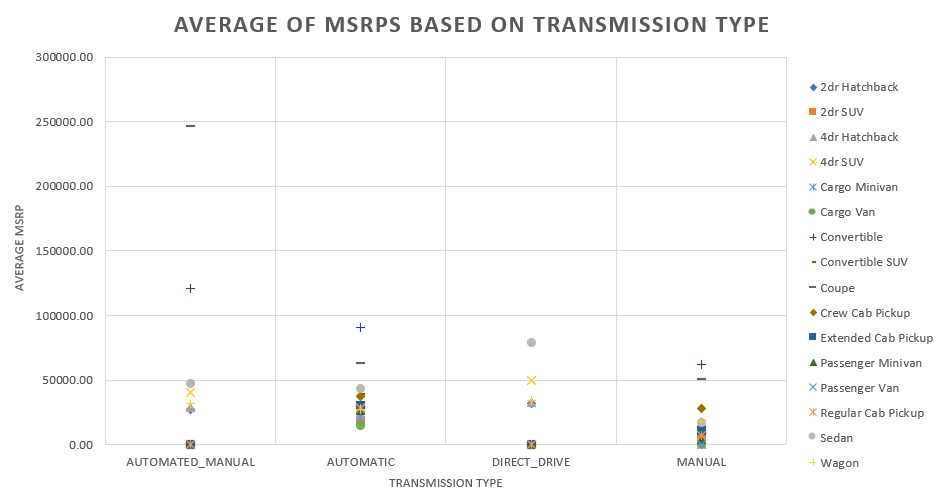
**Approach:**

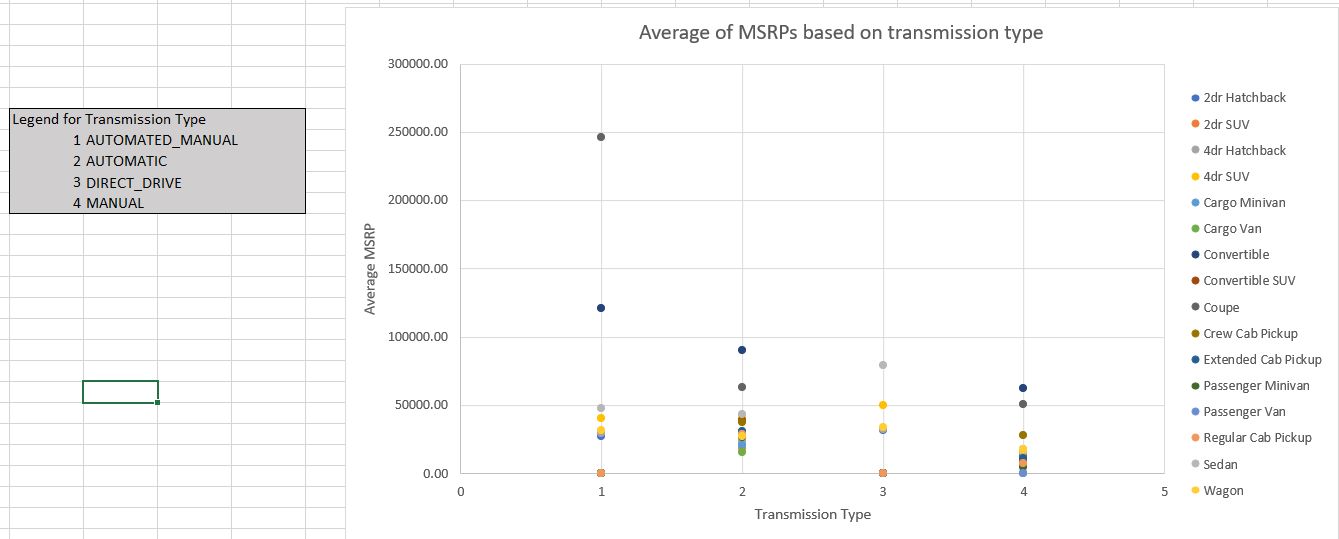
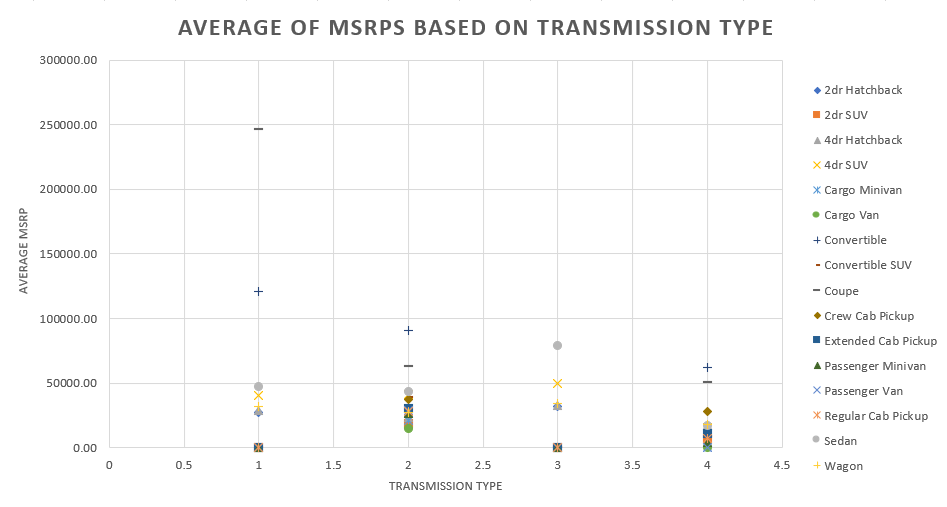
1. Select the table in the “Car\_data” sheet and create another pivot table in the “Dashboard\_Calculations” sheet with the following fields:  
   
2. Select this pivot table and create the clustered 2d column chart. I have moved the chart to the “Dashboard” sheet.  
   
3. From the first look at the chart, it is clear that Bugatti has the highest average MSRP for its cars which is not close to any other manufacturer.
4. We can filter out the Make and Vehicle Style with the filters present in the chart.
5. For example, if we want to see the total MSRP for the vehicle style Wagon, I can select the filter available on the right-hand side of the chart. This is the result:  
   
6. I have temporarily selected the Data labels option to view the average MSRPs. This can be played around with the person who wants to use the dashboard.

Problem Description No. 3:

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

**Approach:**

1. Select the table in the “Car\_data” sheet and create another pivot table in the “Dashboard\_Calculations” sheet with the following fields:  
   
2. Copy the contents of this pivot table and paste it outside it and create a table.
3. Select this newly created table and create the line with markers chart plot chart. I have moved the chart to the “Dashboard” sheet.
4. Next, remove the lines by selecting the “No line” option here since we have to create a **scatter plot** chart:  
     
   
5. The above line chart I had created since the scatter plot chart does not display textual values on the horizontal axis. If we want to create the scatter plot, we can directly create that also and provide a legend:

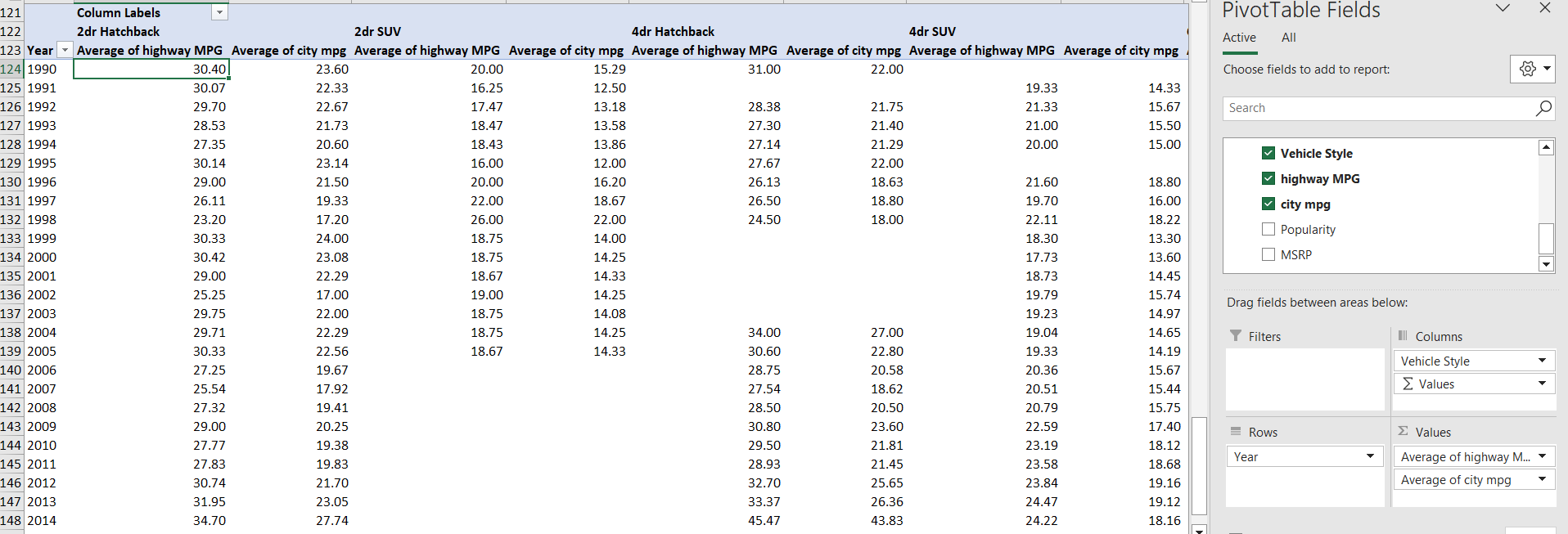
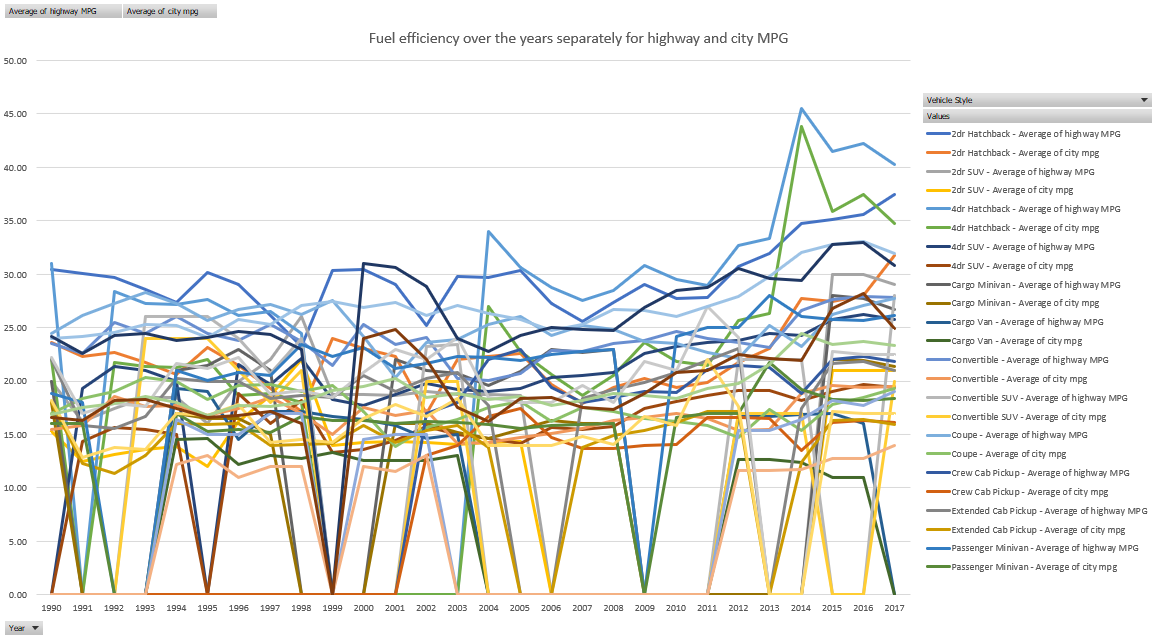
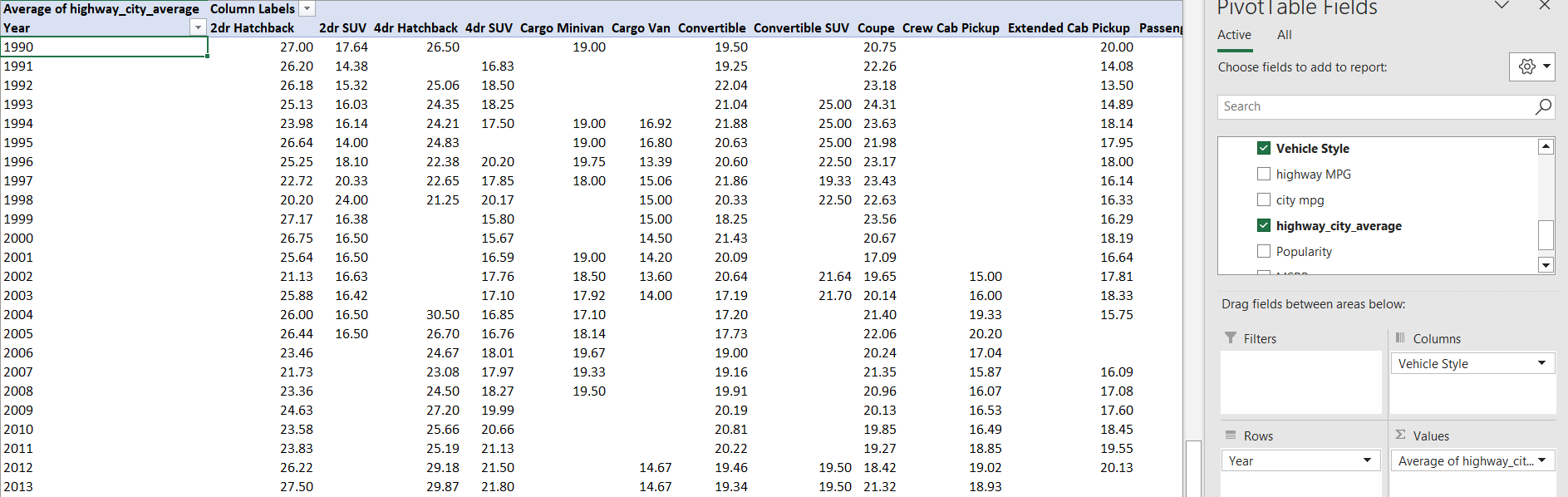
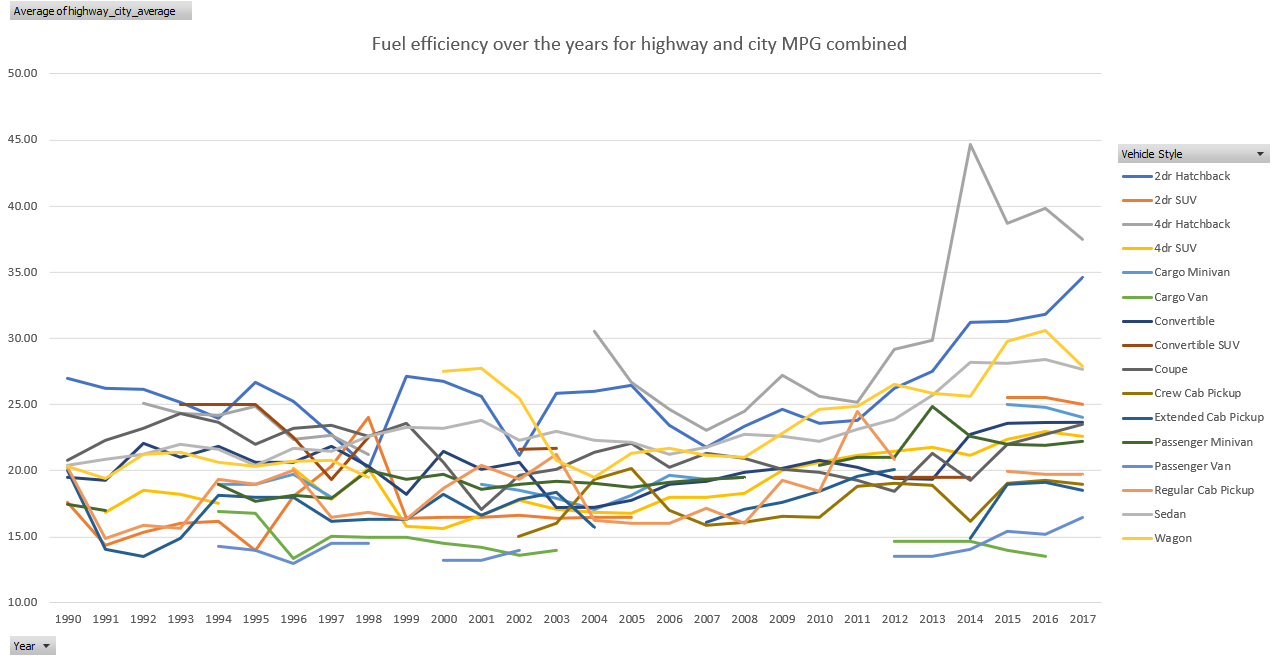
 

1. Now, if we want to see the average MSRP for 4dr SUV, we can select the filter option offered by the chart and select 4dr SUV as demonstrated below:  
   
2. It is evident that the cars having transmission type as “DIRECT\_DRIVE” and vehicle style as 4dr SUV have the highest average MSRP.

Problem Description No. 4:

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

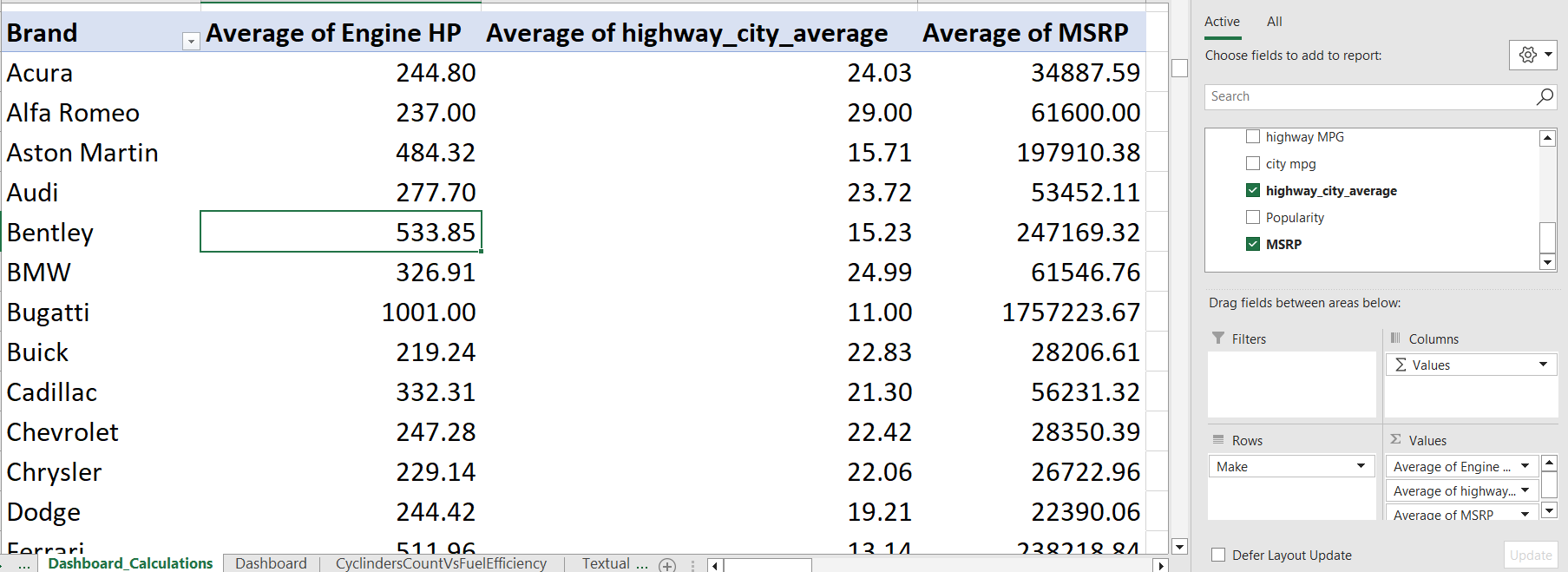
**Approach:**

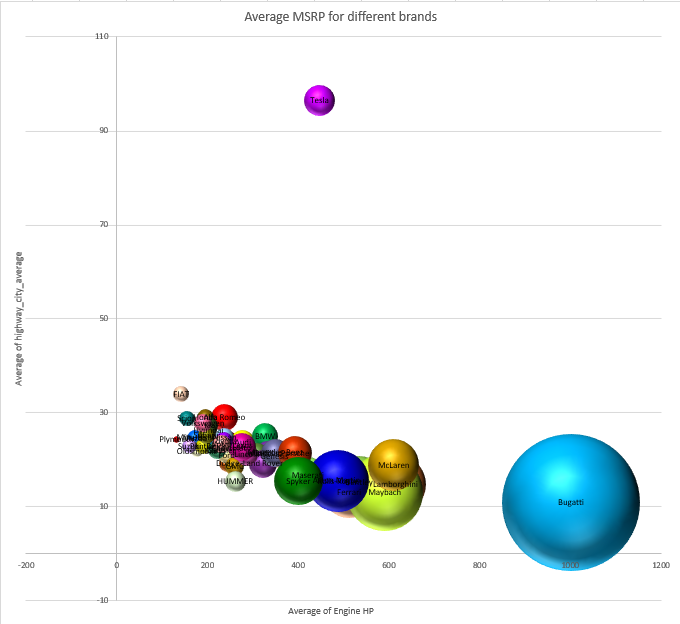
1. First, I will create a pivot table with separate average of highway and city MPG columns as shown below:  
   
2. After this, I have created the line chart and moved it to the “Dashboard” sheet.  
   
3. Here we can see that for all the body styles, the highway and city MPG averages follow the same trend over the years.
4. With the filters available, we can select the body style or the year and view the average highway and city MPG lines separately.
5. Next, I have created a column highway\_city\_average in the “Car\_data” sheet by taking the average of the highway and city MPG values with the help of AVERAGE function.
6. Then, I have created another pivot table which shows for a particular year the average of the highway\_city\_average:  
   
7. This can be interpreted as: For the year 1990 the average MPG offered by 2dr hatchback cars is 27. When we cross verify this value with the previous pivot table, we see that the average highway MPG is 30.40 and average city MPG is 23.60. If we take the average of these two values, it will come to be 27 only.
8. Finally, I have created another chart in the “Dashboard” sheet for this pivot table as well.  
   
9. Either of the two charts can be used for further analysis and deriving results. For a bigger picture analysis, the combined chart can be used and for a detailed analysis, the separated chart can be used.

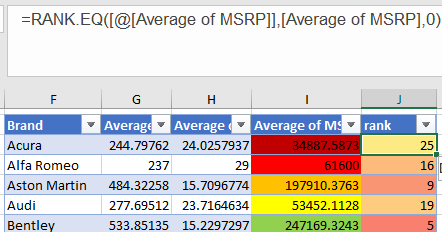
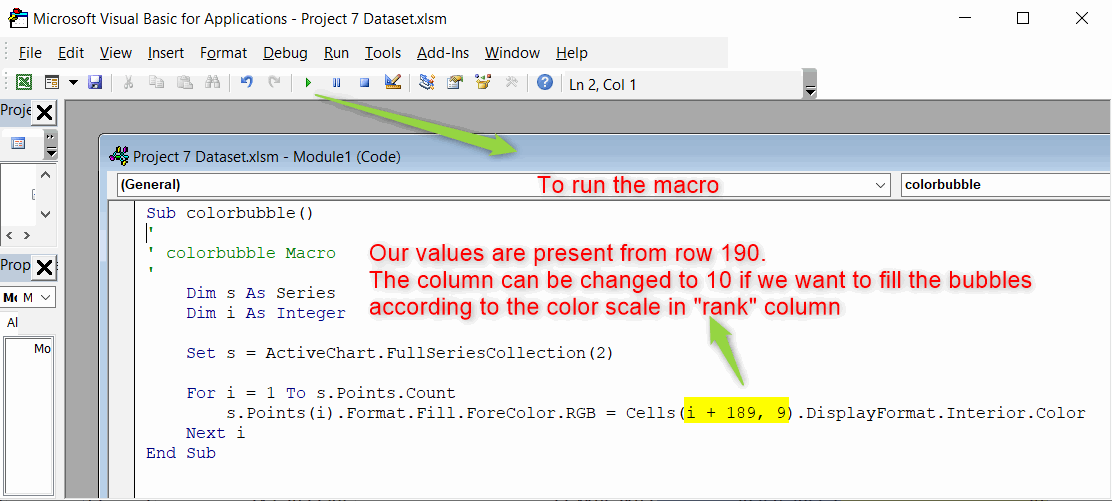
Problem Description No. 5:

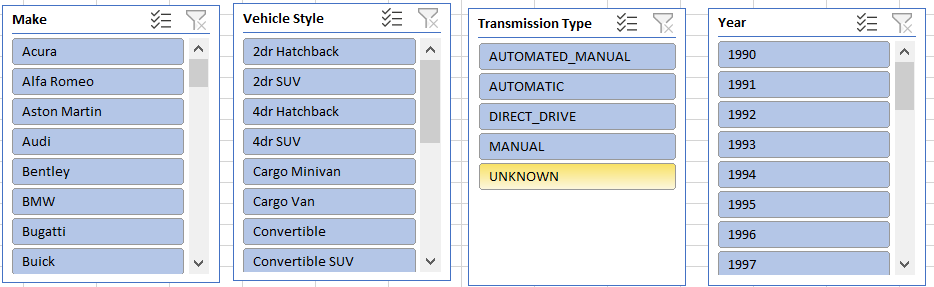
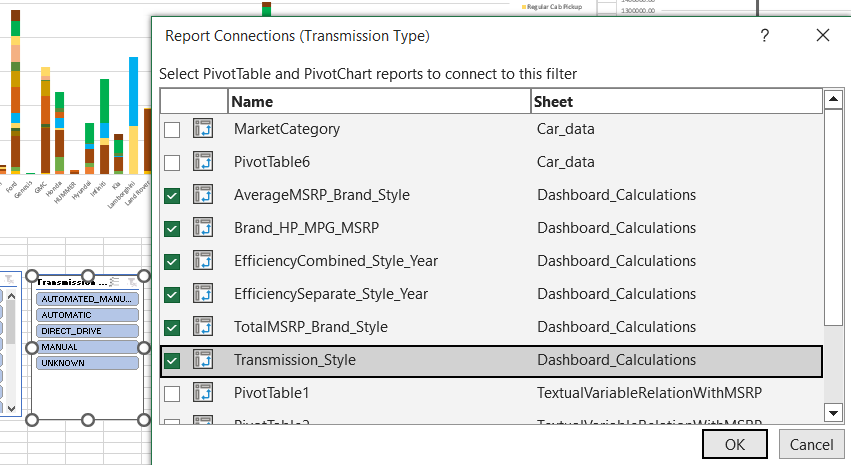
How does the car's horsepower, MPG, and price vary across different Brands?

**Approach:**

1. Create the pivot table from the “Car\_data” sheet with the following fields:  
   
2. Next, copy the contents of this table in a new range since we need to create a Bubble chart.
3. From this newly selected table insert the Bubble chart where the size of the bubble depicts the Average MSRP for each brand.



1. From the size of the bubble, we can see that Buggati has the highest value of Average MSRP.
2. For colouring the bubbles with different colours, I have used VBA and/or conditional formatting.
3. In a first method, I have manually formatted the column “Average of MSRP” with different colours.
4. In the second method, I have created another “rank” column to display the rank of the different Average MSRP values.  
   
5. Then using either the column I or the column J, I have recorded a macro “colorbubble” as shown below:  
   
6. The variable ‘i’ as it increments will fill down the bubbles with different colours which is why I have used a for loop.
7. Next, I have moved the chart to the “Dashboard” sheet.

Finally, for all the charts created in the dashboard, I have added 4 slicers and connected them to the pivot/normal tables that are used to create these charts. **I have created only 4 slicers according to the mode of analysis required by the client.**  
 

**Result:**

From the insights derived above, the car manufacturers can considerably regulate the prices based on the most popular features among the customers which will help them derive maximum profits.

Overall, I found this project particularly interesting as I did my research for some of the missing data and got to learn about the features that impact the price of a car. Creating the dashboard and playing around with it was also new to my Data Analytics learning which further increased my confidence in handling a wide range of client requirements. In terms of difficulty I found that this project was comparatively easier than the previous two, even so it helped me refine my anaytical skills owing to new learnings and critical thinking.