Tableau Data Analysis

Part 1: Introduction

The automotive industry is an ever-changing one, utilizing extensive research and development in order to create the best products for customers, while also bringing in greater profits towards the company and minimizing environmental impact at the same time. The 1970s and the 1980s was a prolific time for this particular industry. Electronic Ignition Systems, Fuel Injection, Anti-Lock Braking Systems, Turbocharging and All-Wheel Drive Transmissions are just a few examples of automotive technological advancements from those aforementioned eras. With that in mind, as a researcher, one might want to look back and see how cars have improved in terms of general performance and fuel efficiency throughout those years. That's why I have chosen this dataset, which is the "Car Information Dataset" from Kaggle. In the following part of this paper, I will do a deep dive onto the dataset.

Part 2: Dataset explanation

For this paper, I have chosen the "Car Information Dataset" dataset from Kaggle. There are, by default, 9 features in this dataset, namely;

- Name, which is a unique identifier for each automobile, based on the make and model of the respective automobile.
- MPG, which is fuel efficiency, measured in miles per gallon.
- Cylinders, which is the number of cylinders in the engine.
- Displacement, which is the size or capacity of the engine, measured in cubic inches.
- Horsepower, which is the commonly used unit of power for cars.
- Weight, which is the weight of the automobile in kilograms.
- Acceleration, which is the capability of the automobile to increase its speed from 0 to 100 kilometers per hour, measured in seconds. Lesser values mean better performance.
- Model Year, which is the last 2 digits of the year that the automobile was made in.
- Origin, which is the country or region of origin of the automobile.

It appears that, with there being a great bias in the number of models of American cars, it is safe to assume that this dataset represents cars that were sold in the American market from 1970 to 1982.

Part 3: Data readiness (Data cleaning process)

I found this particular dataset to be quite clean, with it only having 6 nulls throughout the entire dataset, all being in the “Horsepower” field. However, I personally did not like how the data was structured. For example, all the names here were lowercase in the .csv file, so I had to use excel to make it use proper case, first using an Excel function, and then manually tweaking some model names (e.g. “Sw” to “SW”, which stands for “Station Wagon”). In this process, I also discovered that there were also some naming inconsistencies within the .csv file, such as “Chevrolet” being referred to using 3 names; “Chevrolet”, “Chevy” and “Chevroelt”, which was most likely a typographical error. Some models also had inconsistent naming, e.g. “Datsun 300ZX and Datsun 300-ZX”. After I had fixed the naming schemes, I also manually changed every “Europe” value, which was the only continent, as opposed to the other names in “Origin”, which consists of countries, into respective European countries where the respective automobile originated. (England, Sweden, France, Germany, Italy). The next steps I did was with Tableau Prep, in which I did these steps;

1. Changed the role of “Origin” to “Country/Region”
2. Changed “Model_Year” to “Date”, and then only showing the year aspect of the date since there is no other information given regarding the date.
3. Wrote a command to create a new field which replaces the 6 null values to 100HP. I chose 100 because most of the nulls were American cars with fairly big displacements. I then deleted the original horsepower field.
4. Wrote commands to split the “Name” attribute to “Make” and “Model”, in order to add an extra attribute to gain insight from (Make).
5. Finalized the flow with exporting it into a “.hyper” Tableau file extension.
6. Ran the flow.

After running the flow, the data preparation/cleaning part of this paper is done.

Part 4: Questions

Before doing the visualization part of this paper, I prepared 10 research questions in order to base the visualization on, which are:

1. **Performance Over Time**
    a. Are there any possible implications of relationships between these variables based on how the graphs changed?
    b. Which years had the maximum and minimum values for each variable?
    c. Analyzing how different amounts of engine cylinders will affect these graphs.
    d. How do these graphs differ over different weight groups with respect to the unfiltered graph? (1500-2500 (Light), 2500-3500 (Midsize) and 3500-5000 (Heavy))
2. **Fuel Efficiency vs Horsepower and Displacement**
    a. Does the horsepower and displacement of an automobile affect its fuel efficiency?
    b. Does the weight of a car have a relationship with the other factors in this graph?
    c. Are there any trends we can infer in this graph from an automobile's origins?

3. **Performance by Origin and Make (Bar Charts)**
    a. Analyzing performance by origin (Which countries have the best performance in terms of Average MPG, Average Horsepower, and Average Acceleration)
    b. Which makes have the best performance according to the aforementioned criteria?
    c. Which makes have the least well performing automobiles in each aspect?

Part 5: Technical details of the used visualization methods

I have chosen three visualization methods to display the data in this dataset, which include:

1. 1 visualization which includes multiple line graphs detailing how the average MPG, acceleration, horsepower and engine displacement of automobiles changed over the time range that this data has provided.
2. Another visualization which includes two "3-Dimensional" scatter plots. In these graphs, I aimed to highlight the relationship of how the fuel efficiency of an automobile is impacted by its performance, in terms of horsepower and displacement. For extra information, I had color coded MPG into the circular points within the graphs, showing red for worse fuel efficiency and green for better fuel efficiency. The size of the dots also represents the weight of the automobile. Bigger dots mean heavier cars.
3. The last visualization is a compilation of bar charts sorted by the country of origin and make of automobile, which highlights how performance differs between makes and countries in terms of MPG, horsepower and acceleration. Color coding is also added to differentiate between countries.
4. I also added relevant filters to all of these visualizations to help the interactivity aspect of my visualization.

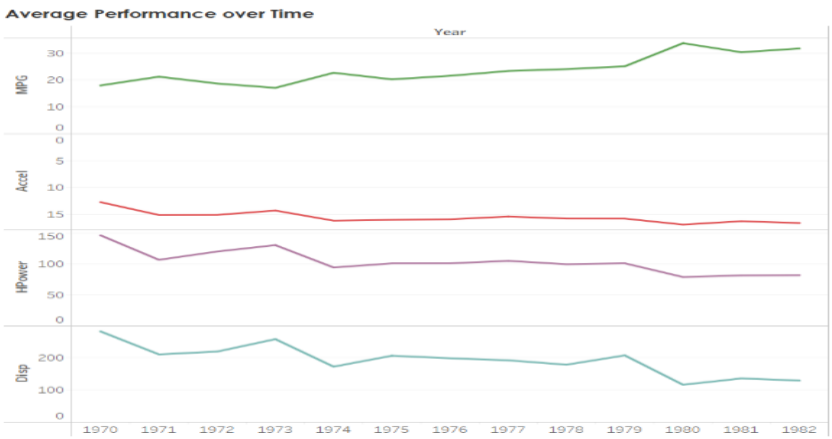
These visualization methods, I believe, are sufficient to cover all the research questions I had made earlier.

Part 6: Justification of the Visualization Methods

I had chosen these three methods because:

1. As for the amount, I tried my best to not crowd the dashboard with too many visualizations, so I only chose three which, in my opinion, represent the most relevant aspects of the data well. I experimented beforehand with up to seven, and filtered those to get the best three.
2. For time-related analysis, I believe that a line chart is the best kind of chart to highlight how something changes within a given time period. Visually, it is a fairly simple and non-gimmicky chart, which means that most people will, at a glance, understand what is happening in those charts. This allows for effective viewing and eases understanding in general.
3. I believe that a scatter plot would be perfect for highlighting the relationship between 2 variables within the data. Scatter plots also allow for color and size coding to include more information in the graphs. The built-in trend line feature also helps a lot in analyzing findings within the data.
4. Bar charts enable us to view data which are split using multiple categories well, in this case being countries of origin and car makes. I had also color coded each country as distinct colors to ease the understanding of the data from a user perspective.

Part 7: Discussion on analysis results



Note: Firstly, I feel that it is important to note that for this visualization in particular, I have reversed the acceleration graph, so that it is in terms of how fast a car is as opposed to the original statement which says that lower values mean that the automobile is faster.

1a. Are there any possible implications of relationships between these variables based on how the graphs changed?

I had made four graphs initially, which represented trends over time in automobile average MPG, acceleration, horsepower, and displacement in general. I noticed that the

graph for the average MPG over time looked like an inverse of the other three, which makes sense, because more cars being more powerful usually mean worse fuel efficiency. We will investigate this relationship further in the second graph.

1b. Which years had the maximum and minimum values for each variable?

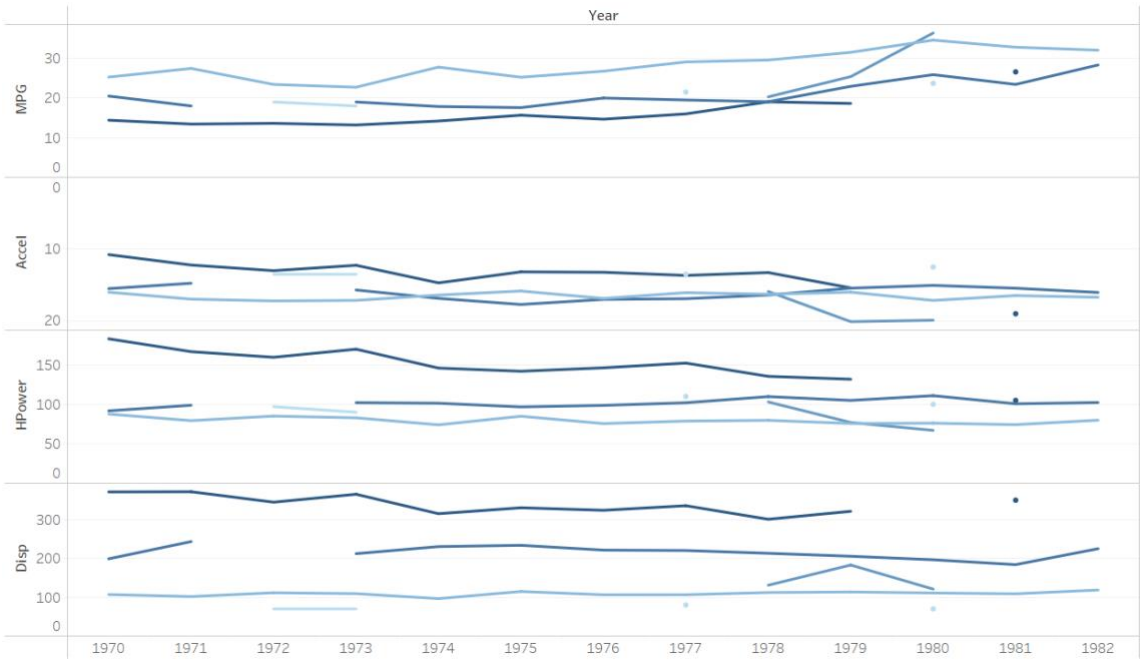
MPG: The best year, fuel efficiency wise, was 1980, with the value of 33.7 Miles per Gallon on average. The worst year, on the other hand, was 1973, with the average MPG of an automobile being 17.1.

Acceleration: The best year for acceleration was 1970, with the average acceleration being 12.75 seconds from 0 to 100 kilometers per hour, and the worst was 1980, with 16.934 seconds.

Horsepower: The year which had the most powerful cars was 1970, with 146.21 HP on average, and the year with the least powerful cars was 1980, with 79.03HP on average.

Displacement: The maximum value for engine displacement was on the year 1970, with 280.6 cubic inches on average, and the minimum value was on 1980, with 115.6 cubic inches on average.

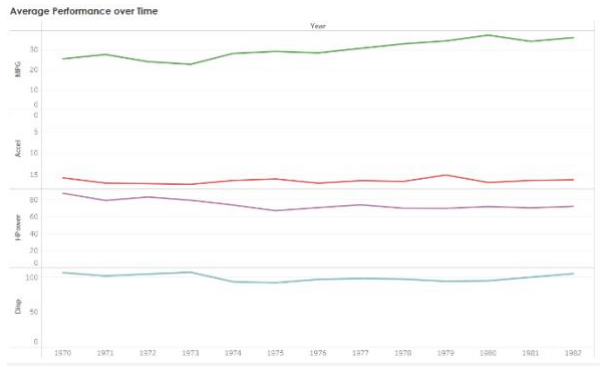
1c. Analysing how different amounts of engine cylinders will affect these graphs.

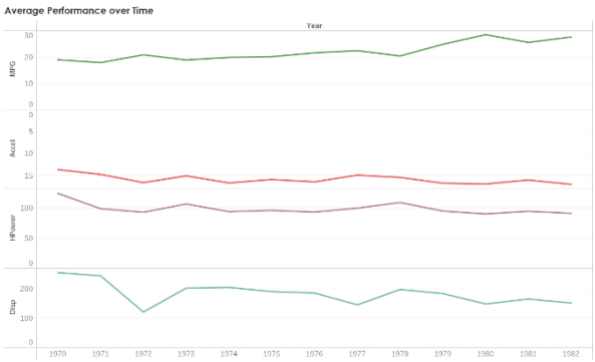


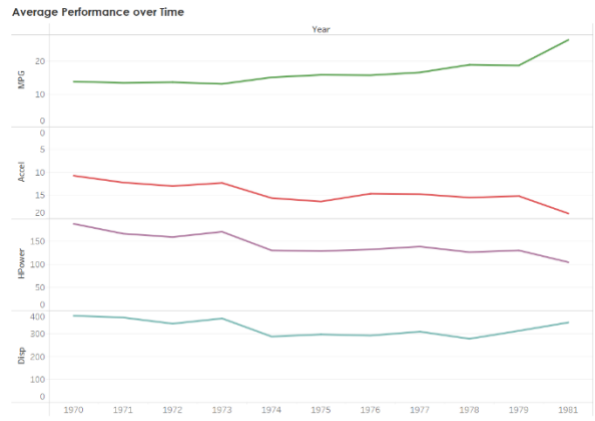
For this point, I increased the granularity of the graphs to include the number of cylinders in the graphs, color coded with the darker the shade of blue, the more cylinders there are in the automobiles. It seems that cars with certain cylinders numbers go in and out of style. We can see that for the last 3 graphs, the shade of blue seems to get darker as the graph goes more upwards, which does indicate that generally, the more cylinders an automobile

has, the more powerful and faster it is. The trend seems to be the inverse of this on the average MPG graph, which is understandable, because as stated earlier, the more powerful a car is, the worse the fuel efficiency.

1d. How do these graphs differ over different weight groups with respect to the unfiltered graph? (1500-2500 (Light), 2500-3500 (Midsize) and 3500-5000 (Heavy))

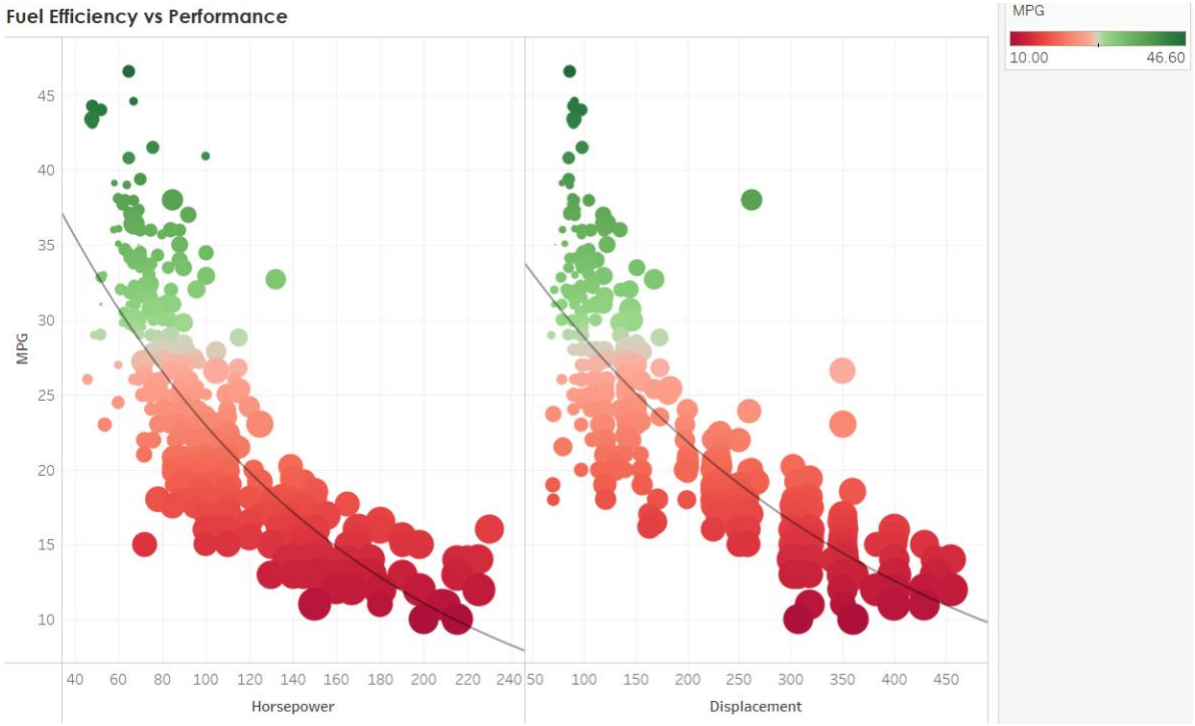






As we can see from these graphs filtered according to the weights I have specified in the question, it seems to be the case that cars in the lightweight class mirrors the trends of the unfiltered graph the best. Lighter automobiles follow the MPG and acceleration graphs of the original graph very similarly. They do however, lose the intensity of the changes that happened over time in the horsepower and displacement graphs. The midrange weight class is the most random out of all three. Albeit being quite similar at a glance, it does not have the same hills and valleys as the first graph. Heavier automobiles stopped being produced after 1981 in this dataset, but from what is there, we can infer that it has quite a similar pattern within the graphs, compared to the unfiltered graphs. From this, we can infer that midrange cars deviate the most from the general specifications of all the cars, whereas cars on the lighter and heavier sides of the data tend to follow the same trends as the unfiltered graph.

2a. Does the horsepower and displacement of an automobile affect its fuel efficiency?

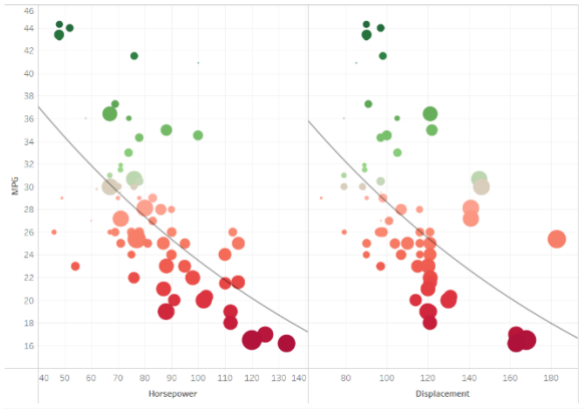


As we can see from these scatter plots, there is a clear correlation between fuel efficiency and horsepower along with the engine displacement of automobiles. The MPG, indicated by the y-axis, indicates that there is a significant drop in MPG as horsepower and displacement increases.

2b. Does the weight of a car have a relationship with the other factors in this graph?

I have mapped size to the graph to signify heavier cars with bigger dots. It seems to be the case that in general, the heavier cars are, the bigger their power and engine displacement is.

2c. Are there any trends we can infer in this graph from an automobile's origins?

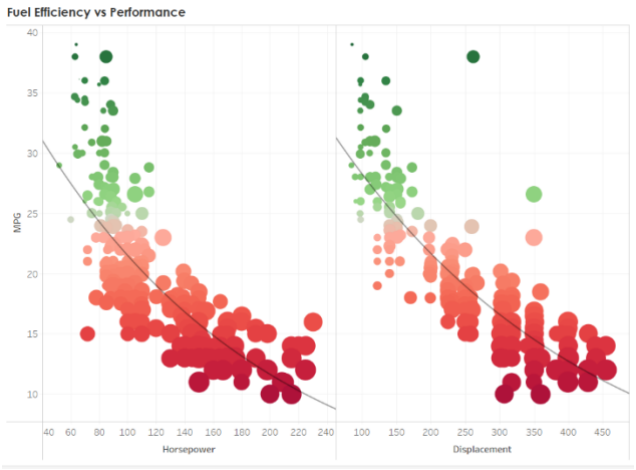


Europe

Firstly, I grouped European countries into one group, which is represented by the first picture. We can see that European cars are very random, with spots here and there which when combined, has a pretty similar pattern as the unfiltered graph.

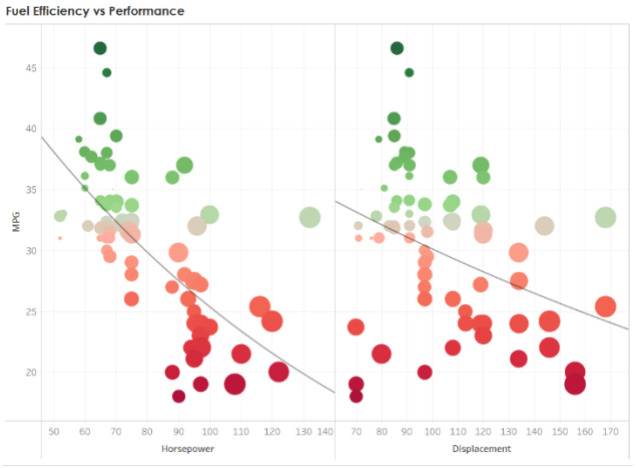
The distribution between low and high MPG cars are fairly even in European countries, and we can see that they also produce cars with varying degrees of performance.

USA



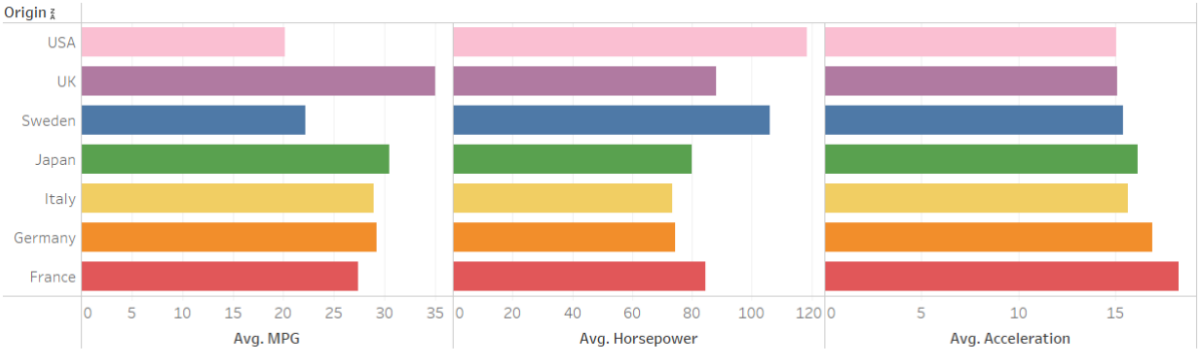
Automobiles from the USA are very well represented in this dataset, with the graph only for American cars being very identical to the unfiltered graph. However, we can see here that cars from the USA are generally heavier, with there being many big points in this filtered graph. American cars also usually have worse fuel efficiency within this time period, with there being a lot of red points in this graph.

Japan



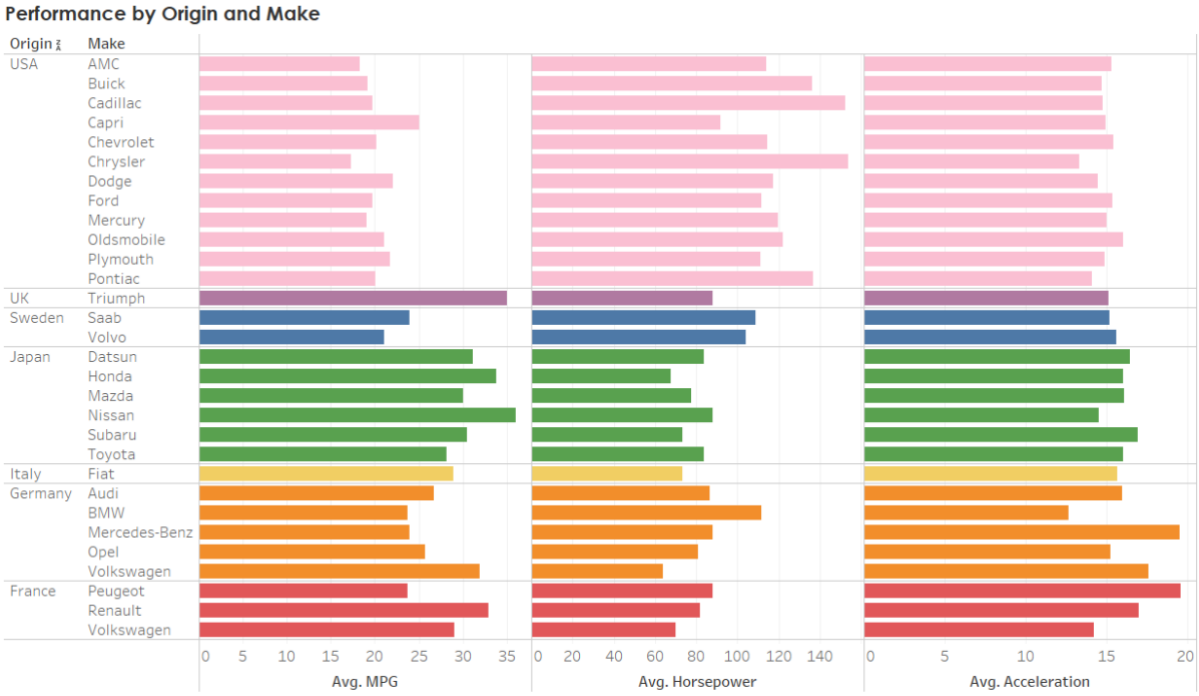
Cars from Japan are the most random out of all the regions, with there being a lot of outliers spread out. As we can see there are lots of cars with low power but disappointing fuel efficiency, and some cars with outstanding performance but decent fuel efficiency. Theres not really a trend going on here, and further research might prove useful in finding answers on why this is the case.

3a. Analyzing performance by origin (Which countries have the best performance in terms of Average MPG, Average Horsepower, and Average Acceleration)



As we can see here, cars from the UK have the best average MPG. This, however, might be biased as there is only one car from UK, which is the Triumph TR7 Coupe. Judging only from the countries with multiple data points, the title of the best average MPG goes to Japan, with 30.45 Miles per Gallon. As for horsepower, the USA wins by quite a long shot compared to other countries, with the exception of Sweden, with 118.44 HP on average. This is very reasonable, since the USA is known to make bigger engines with more power. The best average acceleration goes to the USA as well, with 15.040 seconds from 0 to 100 kilometers per hour on average.

3b. Which makes from each countries have the best performance according to the aforementioned criteria?

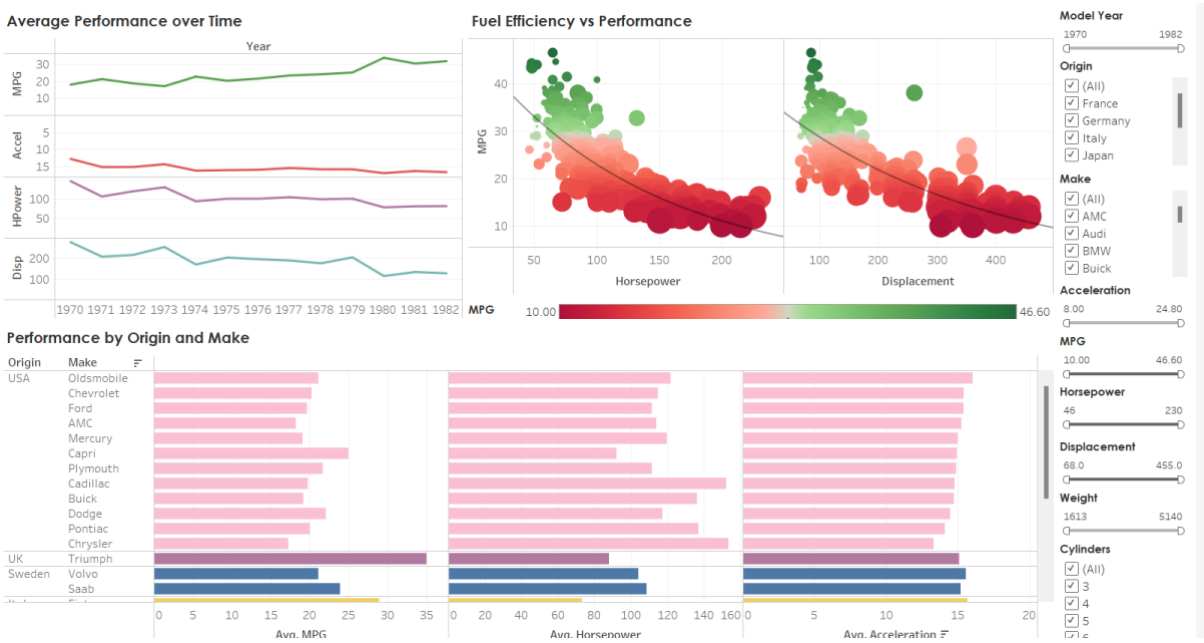


From these bar charts, the most fuel efficient makes from the US, Europe and Japan respectively are Capri, Triumph and Nissan. When it comes to horsepower, the best makes are Chrysler, BMW and Nissan. For acceleration, the best makes are Chrysler, BMW and Nissan. Overall, if we can infer a pattern from here, Nissan makes the best automobiles, with their cars on average having more power and faster acceleration, while also being fuel efficient.

3c. Which makes have the least well performing automobiles in each aspect?

On the other side of things, the least well performing automobiles from the US, Europe and Japan respectively are as follows; Fuel efficiency wise they are Chrysler, Volvo and Toyota. Horsepower wise they are Capri, Volkswagen and Honda. Acceleration wise, they are Oldsmobile, Mercedes-Benz and Subaru.

Part 8: Demonstration of Interactivity and Dashboard



I have created an interactive dashboard for people to use, in which I have displayed, in my opinion, the most insightful versions of the previously shown visualizations. I have also included some relevant sliders and filters, such as Model Years, Origin, Make, Acceleration, MPG, Horsepower, Engine Displacement, Weight and Number of Cylinders, in order to ease a user in gaining insight with certain criteria to answer their questions. This dashboard is fairly complete, equipped with relevant legends and filters, along with as much granularity as far as ease of understanding goes.