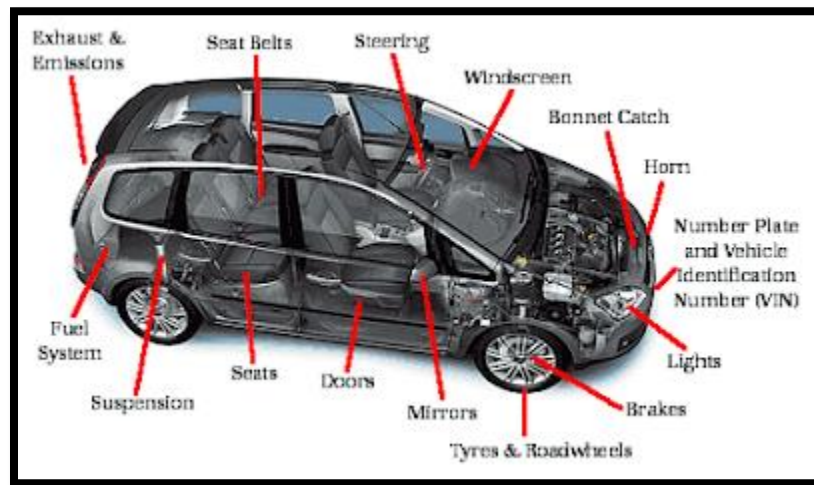


Report on the development of interactive cars data visualisation




Full name: Tuan Dung Lai

Student ID: 101467732

Unit code: COS30045 – Data Visualisation

Assessment name: Final visualisation project

I hold a copy of this assignment that can be produced if the original is lost/damaged. To the best of my belief, no part of this assignment has been copied from any other student's work or from any other source except where due acknowledgement is made in the text. No part has been written for me/us by any other person, except where such collaboration has been authorized by the lecturer concerned

Signature: 

Date: 5th Apr 2018

Contents

1. Introduction.....	2
1.1. Background and Motivation.....	2
1.2. Project Objectives	3
1.3. Project Schedule.....	3
2. Data.....	4
2.1. Data Source.....	4
2.2. Data Processing.....	6
3. Requirements	6
3.1. Must-Have Features	6
3.2. Optional Features	6
4. Visualisation Design	6
4.1. Proposal.....	6
4.2. Progress Report.....	9
4.3. Final visualization.....	10
4.3.1. Quick review of visualization by Mariano.....	10
4.3.2. Updated visualisation based on Mariano visualisation.	11
5. Validation.....	12
6. Conclusion	14
Reference	14

1. Introduction

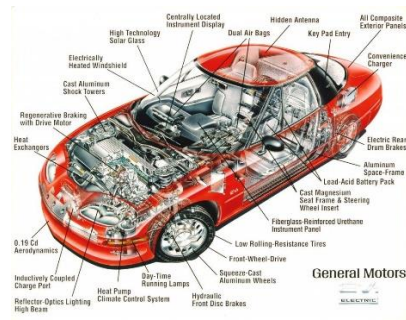
1.1. Background and Motivation

It was 10 years ago, my father bought me a mini old car model of the Peugeot 205 for my 11th birthday, he also told me a long story about the special specifications of that car compared to other modern models. The car was incredible both inside and outside, since then, I have developed a hobby which is collecting old car models researched into them.



Each car has its own story, there are more than 30 parts of a car that

Figure 1: My old cars collection



make it unique, each car brand has their own characteristics. Back in 2000, information about cars was limited, being able to know the specification about each car and compare them to other models was a difficult part. In this project, an interactive visualisation is planned and developed so that cars enthusiasts can explore the dataset and understand the differences between different cars brands as well as the correlation between specifications if they exist.

Figure 2: Inside a typical car

Another motivation is that visualizing cars data can assist buyer in the process of choosing the right car in the market. Knowing the specialization of varied brands can play a signification role in the deciding process.

1.2. Project Objectives

The primary question that could be answered by the visualisation is what types of car a brand normally makes and how the price and other factor of the cars are correlated.

The visualisation allows users to analyze the different between products of various cars companies, including prices, brands and specification. Cars enthusiasts can research into the details of parts in a car and how they are related.

People who wish to buy cars can have better understanding of price range and specialization of varied brands in the car market.

Some questions that can be answered by using the visualisation:

- _ Are diesel cars generally more expensive than gas cars?
- _ Is there any correlation between engine size and horsepower?
- _ Does length of cars affect the curb-weight?

1.3. Project Schedule

Week 1-4:

Review web programming, learn D3 library and visualisation theory

Week 5-7:

Choose a visualisation project based on research interests and motivations.

Choose relevant dataset, sketch visualisation ideas and analyze the dataset.

Week 8:

Start coding to build the overall structure of the visualisation.

Week 9:

Add detail improvement to the visualisation.

Week 10:

User-ability testing and evaluation, finish report.

Week 11-12:

Prepare for presentation.

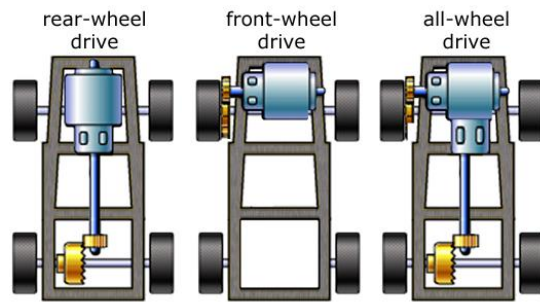
2. Data

2.1. Data Source

The dataset used in this project is called Automobile Data Set which is available for free use at UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/datasets/Automobile>). The data is a table stored in a CSV file containing 205 different cars models. The attribute characteristics are categorical, integer and real. There are 10 categorical variables (e.g. fuel type, number of cylinders) and 14 continuous variables (e.g. price, length), the risk factor (symboling) and normalized losses.

There are 26 attributes which are described in the table below:

Attribute	Description
Symboling	The degree to which the auto is riskier than its price indicates. The values are -3, -2, -1, 0, 1, 2, 3. The value -3 means it is safe and the value 3 means it is risky.
Normalized-losses	The relative average loss payment per insured vehicle year. These values are normalized, they are ranging from 65 to 256.
Make	The brands that made the cars, including alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
Fuel-type	The fuel that the cars consumed: Diesel or Gas.
Aspiration	Standard and Turbo, this refers to the amount of air the motor takes. Engines which have turbochargers run on much higher peak pressures, the air-fuel mixture is more compressed, temperature inside the cylinder are much higher.
Num-of-doors	2, 4.
Body-style	Hardtop: a rigid form of automobile roof, also those automobiles that are styled to resemble a convertible. The top may be detachable for separate storing, retractable within the vehicle itself, or permanently attached to an auto that is lacking a center side-support known as a B-pillar. wagon: automotive body-style variant of a sedan/saloon with its roof extended rearward over a shared passenger/cargo volume with access at the back via a third or fifth door (the liftgate or tailgate), instead of a trunk/boot lid.

	<p>sedan: passenger car in a three-box configuration with A, B & C-pillars and principal volumes articulated in separate compartments for engine, passenger and cargo</p> <p>hatchback: Cars that have body configuration with a rear door that swings upward to provide access to a cargo area.</p> <p>Convertible: Cars that can be driven with or without a roof in place.</p>
Drive-wheels	<p>4wd/awd, fwd, rwd</p> 
Engine-location	Front, rear
Wheel-base	The distance between the centers of the front and rear wheel, continuous from 86.6 to 120.9
Length	continuous from 141.1 to 208.1
Width	continuous from 60.3 to 72.3
Height	continuous from 47.8 to 59.8
Curb-weight	<p>the total weight of a vehicle with standard equipment, all necessary operating consumables such as motor oil, transmission oil, coolant, air conditioningrefrigerant, and sometimes a full tank of fuel, while not loaded with either passengers or cargo.</p> <p>Continuous from 1488 to 4066</p>
Engine-type	dohc, dohc, l, ohc, ohcf, ohcv, rotor.
Num-of-cylinders	eight, five, four, six, three, twelve, two.
Engine-size	continuous from 61 to 326.
Fuel-system	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
Bore	A part of piston engine, it represents the size, in terms of diameter, of the cylinder in which a piston travels. It's continuous from 2.54 to 3.94.
Stroke	continuous from 2.07 to 4.17.
Compression-ratio	continuous from 7 to 23.
Horsepower	continuous from 48 to 288.
Peak-rpm	Revolutions Per Minute, a measure of how fast a car's engine is turning at any point in time. It's continuous from 4150 to 6600.
City-mpg	continuous from 13 to 49.
Highway-mpg	continuous from 16 to 54.
Price	continuous from 5118 to 45400.

2.2. Data Processing

As the nature of the dataset, all attributes are measured in standard unit. Therefore, no further data clean up is necessary.

3. Requirements

3.1. Must-Have Features

These are features without which the project is considered as failure:

- The visualisation fails to run on web browsers due to technical issues.
- The visualisation is too complicated is not easy for other people to use.
- The visualisation does not help the audience explore the dataset in a meaningful way.

3.2. Optional Features

These are extra features that are nice to have:

- There are more than one visualisation before the validate stage.
- Comparison of different existing visualisation about the same topic as chosen.

4. Visualisation Design

4.1. Proposal

The visualisation is aiming at comparing different cars companies, the design as sketched in figure 3 can let users compare multiple makes as well as visualize the correlation between 2 attributes. Users can select makes and attributes and a scatter plot will be shown. The users can select and unselect makes as well as change the attributes and a smooth transition will be applied. The data points in the scatter plot can be hover and more detail will be displayed on mouse location.

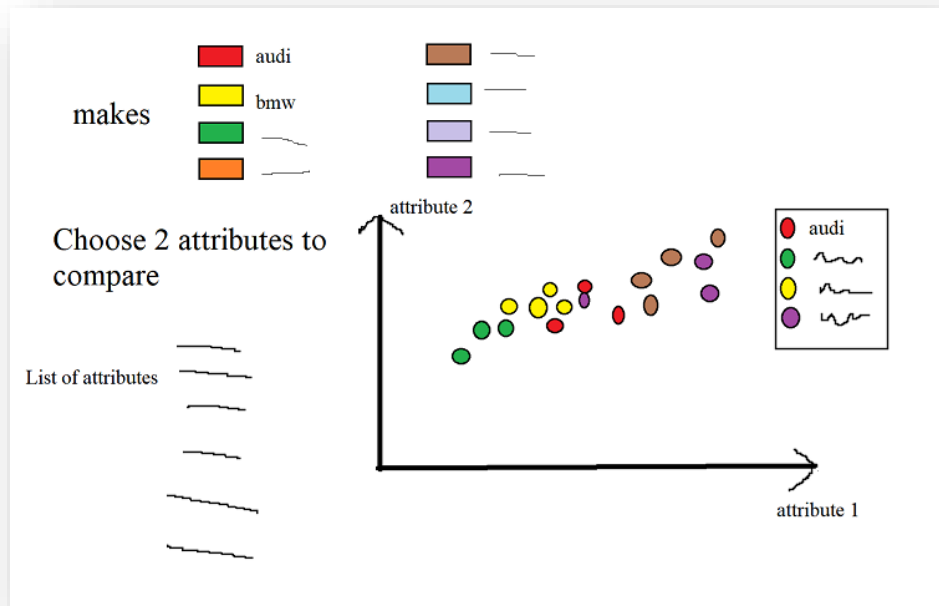


Figure 3: Visualisation idea 1

The following design allows users to explore the number of cars for each defined interval of an attributes. For example, prices are divided into different ranges and the number of cars that have that price will be displayed in a bar chart as shown in figure 4. Each bar can be hovered, and a pie chart will show the percentage of of cars according to makes.

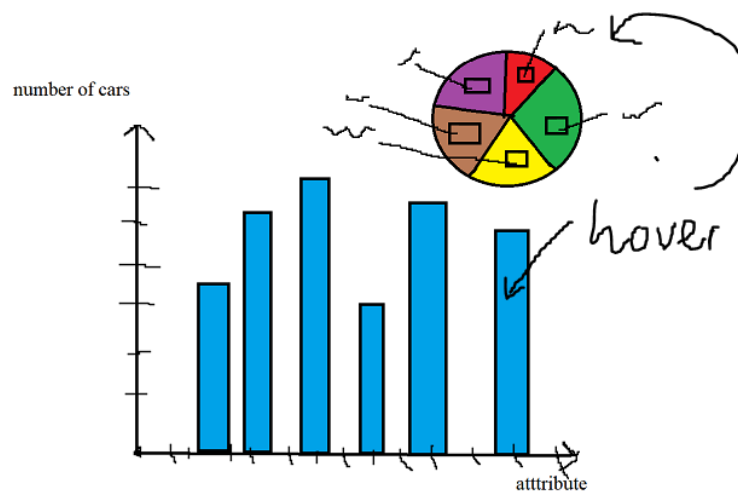


Figure 4: Bar chart and pie chart

In figure 5, a parallel coordinate plot is used, each line represents a data point and the colors represent the makes. User can select a range in an axis to highlight all the cars that have attribute values belonged to that selected range. This design allows user to visualize many attributes at the

same time as well as select the range of value to explore the correlation between variables. The colors can distinguish the makes.

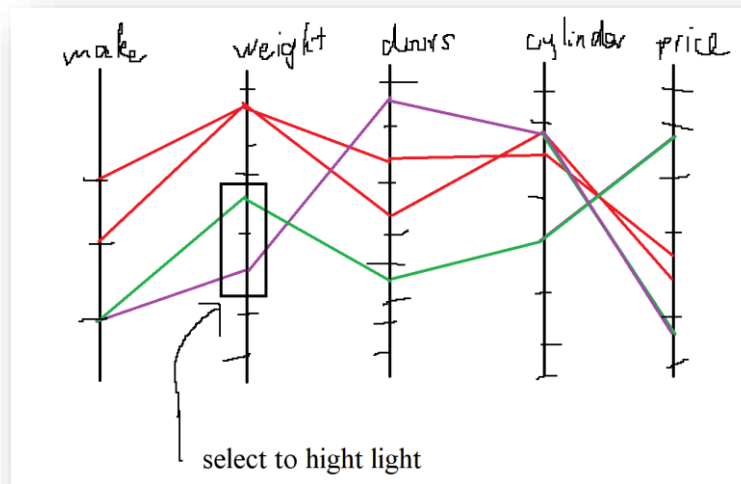


Figure 5: Parallel coordinate plot

This project aims at developing a current visualisation (Mariano Trebino, 2015). In the visualisation created by Mariano, the author pick 12 most statistically significant variables to visualize based on the result of PCA (principle components analysis, a dimensionality reduction algorithm). However, this is not a good approach to reduce the number of variables, each component of the cars has its own characteristic and is equally important. Another design of this project is to let users customize the variables that are visualized either by check box or dropdown menu.

The visualisation created by Mariano has a lot of unused space left the left and right of the screen. An effective visualisation should utilize as much space as possible and avoid unused area. The code created by Mariano is also unscalable and not flexible.

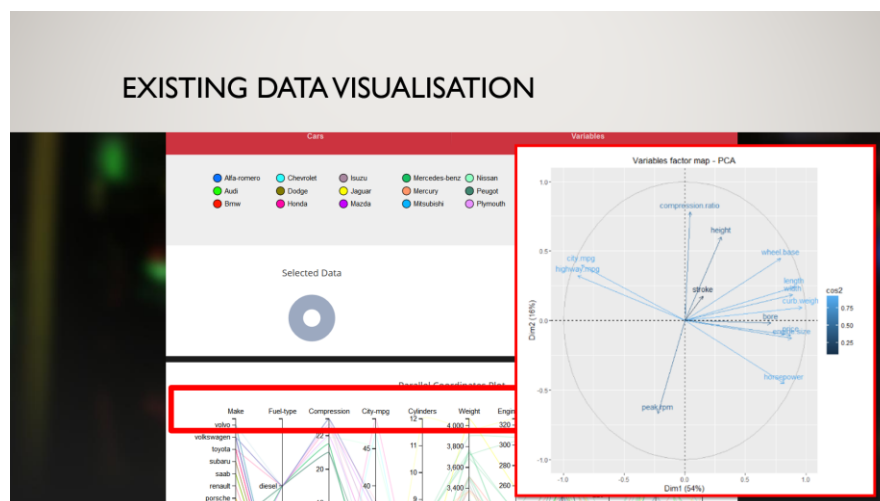


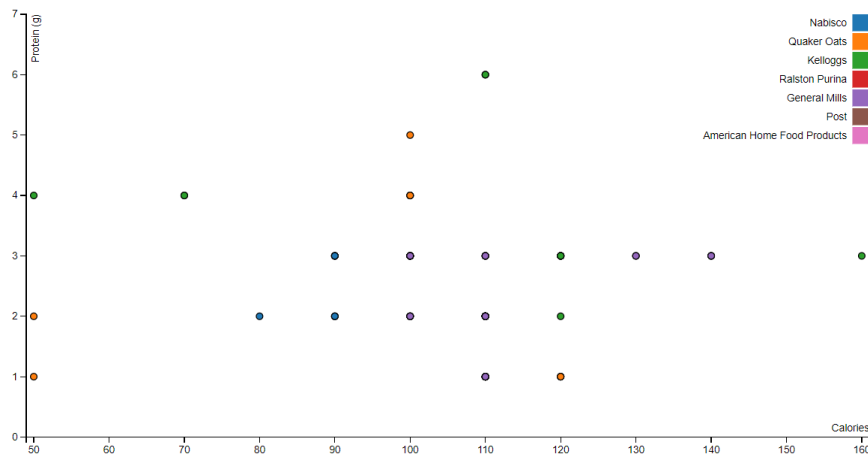
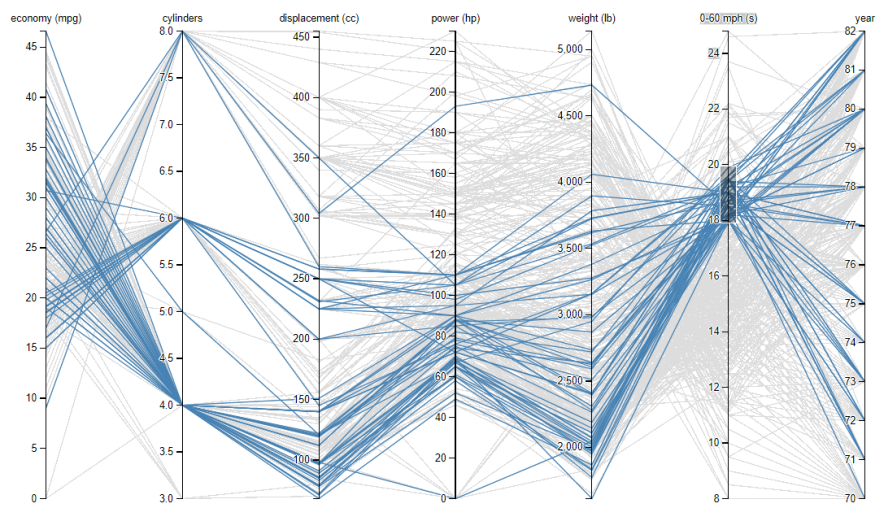
Figure 6: Existing visualisation (Mariano, 2015)

Another drawback of parallel diagram used in the visualisation in figure 6 is that users can easily explore the trends and noticeable feature of data but comparing some specific attributes is still a challenging task. This can be solved by adding a scatter plot that has customized variables. This will assist users comparing specific variables and explore hidden trend and characteristic of the data. Brush filter technique from parallel coordinate plot will be applied in the scatter plot, in other words, changing selected range from parallel diagram will change the scatter plot.

Additionally, hovering data points from scatter plot will highlight the corresponding data in parallel diagram. These 2 features are the use of gestalt principle, different visualisation is connected and linked to each other. Customized inputs option is the use of Schneider's mantra.

4.2. Progress Report

Screen shot of developed visualisation so far:

*Figure 7: Scatter plot (Michele Weigle, 2017)**Figure 8: Parallel Coordinates plot (Michele Weigle, 2017)*

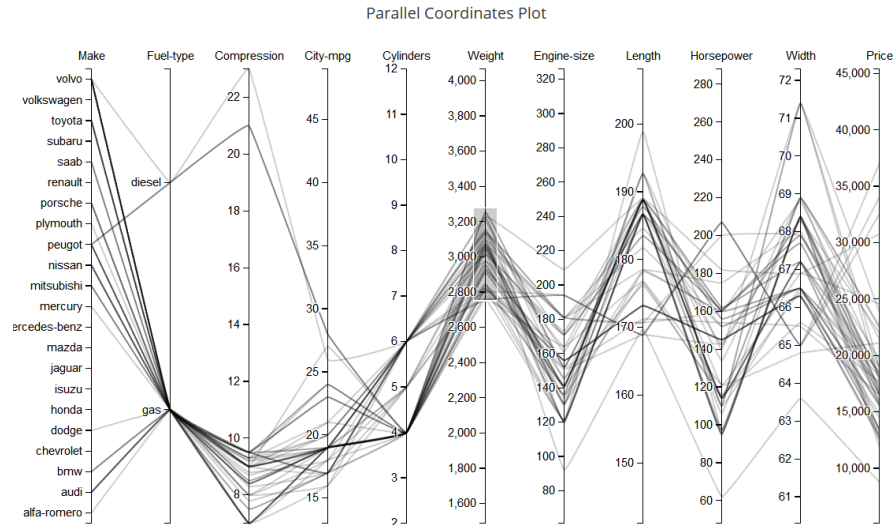


Figure 9: Current development on parallel plot based on Mariano code

4.3. Final visualization

4.3.1. Quick review of visualization by Mariano

Github link: https://github.com/mtrebi/d3_cars



Figure 10: Visualization by Mariano

The total numbers of data points are shown in the top left, it shows that there are 205 cars in the dataset, the number 12 is the total number of variables that will be visualized, the author of that visualization pick the most 12 statistically significant variables using PCA algorithm.

The next parts above the pie chart is the list of cars brand, there are 22 brands in total, each of them is represented by a unique color. The pie chart on the right demonstrate the percentage

of cars model made by each brand, each pie can be hovered by the users, when being hovered, all cars made by that brands will be highlighted in the parallel coordinate plot. Additionally, a small tooltip will appear to give the exact percentage.

The parallel coordinates plot has brush filter effect, users can pick a specific range of value for each variable, all data that are not chosen will be hidden. All data points that are chosen will be shown in the table at the bottom of the visualisation. Users can change the order of the columns in the parallel coordinates plot. Furthermore, each row from the table can be hovered, by hovering the row, the corresponding data in the parallel coordinates plot will also be highlighted.

4.3.2. Updated visualisation based on Mariano visualisation.

Web link: <https://dunglai.github.io/SwinWork/cars-visual/index.html>

Note: This visualization only run correctly in firefox.

The final visualization has accomplished the planned features:

1. Checkbox area above the parallel coordinates plot lets users add or remove variables from the diagram.
2. Scatter plot on the right utilizes the spare space and let users compare specific variables
3. Dropdown menu lets users customize the variables for scatter plot.
4. Brush filter technique is still applied when users change variables. In other words, users can highlight the range of data to visualize.
5. Table data is updated regularly upon changes from the diagram.
6. Scatter plot will change when the selected data from parallel coordinates plot is changed.
7. Data points from scatter plot can be hovered and the corresponding data will be highlighted in the other diagram.
8. Color scheme is consistent on both graphs and the table.



Figure 11: Final visualization

5. Validation

The visualisation is validated by asking different users 3 questions below:

Question 1: Are longer length cars generally more expensive than shorter length cars?

This question can be answer by using the scatter plot with customize variable, the answer is yes.

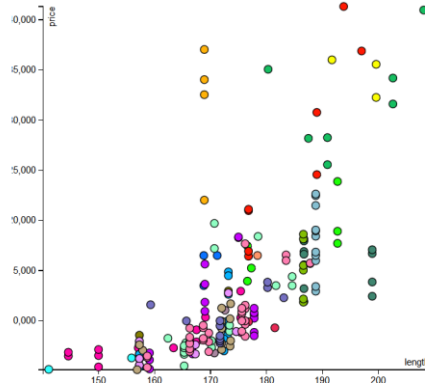


Figure 12: Relationship between length and price

Question 2: Which brand made the most expensive cars?

This question can be answer by using the pie chart or using brush filter technique in parallel coordinate plot. The answer is Mercedes-benz.

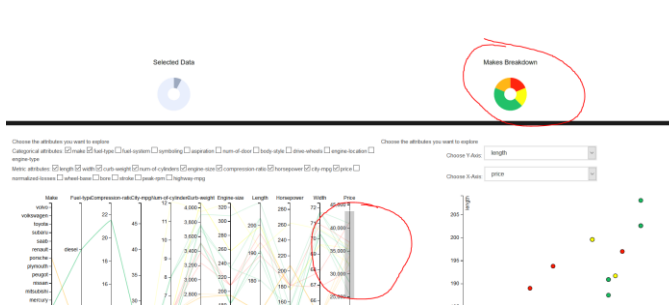


Figure 13: Using brush filter chart

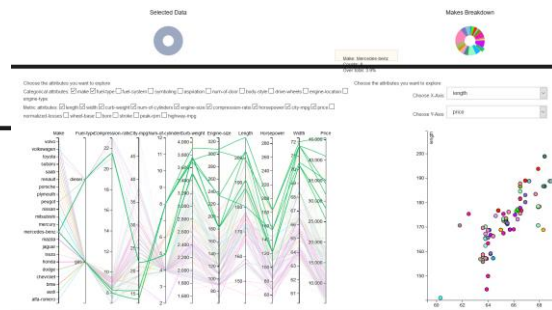


Figure 14: Using pie

Question 3: What is the peak-rpm of the most expensive car in the dataset?

This question can be answered using the table or using the checkbox to add more dimension to the parallel coordinate plot.

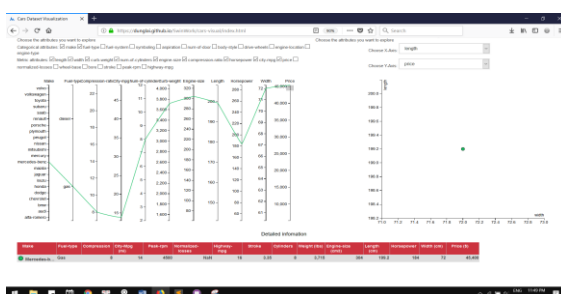


Figure 15: Using table dimension to plot

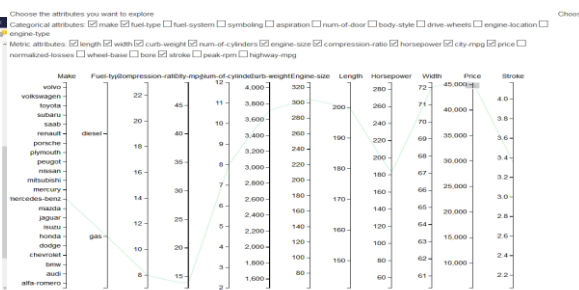


Figure 16: Adding peak-rpm

Name	Age	Question 1	Question 2	Question 3
------	-----	------------	------------	------------

Kon	12	✓	✓	✓
Albert	18	✓	✓	✓
Nguyen	25	✓	✓	✓
Tail	55	✓	✓	✓
Jane	31	✓	✓	✓

6. Conclusion

The visualization appears to be easy to use for users in different age groups. However, more testing should be conducted to collect feedback that can improve the overall design.

Through out this project, I've learnt how to criticize a visualisation, from that criticism, update and improvement are made to eliminate the issues. The design is made from logical decisions and visualisation principles. In terms of coding, the project strengthens my ability to fix, modify and extend existing code as well as using D3 library to do visualisation which will surely be valuable in my data science career.

Reference

Jeffrey C. Schlimmer (1987). Automobile Data Set, UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml/datasets/Automobile>].

Mariano Trebino (2015). Cars data visualisation [https://github.com/mtrebi/d3_cars]

Michele Weigle (2017). D3 Scatterplot example, Block [bl.ocks.org/weiglemc/6185069]

Jason Davies (2017). Parallel coordinates example, Block [bl.ocks.org/jasondavies/1341281]

Appendix A - Code developed so far in stage 2



Appendix B – Final code and working visualisation

Code: <https://github.com/DungLai/dunglai.github.io/tree/master/SwinWork/cars-visual>

Web link: <https://dunglai.github.io/SwinWork/cars-visual/index.html>

(Note: This website is tested on FireFox browser only)