Visualising a multidimensional car-dataset

Steven Raaijmakers, 10804242

September 2018

Abstract
What region makes the most powerful cars? And is there a relation between the weight of a car and its economy? In this report we are finding an answer to these questions by visualising a multidimensional cars data-set into a Parallel Coordinates Graph (PCG).

1 Introduction

In this research we are using a multidimensional data-set containing information about the following details of 391 different cars:

1. Model
2. Miles per gallon (MPG)
3. Cylinders
4. Horsepower
5. Weight (in kg)
6. Year
7. Origin

The cars are manufactured in the US, Japan or Europe between 1970 and 1982.

1.1 Research questions

Within this data-set we want to find an answer to the following research questions:

1. Which region manufactures the most powerful cars?
2. What is the relation between the weight of a car and its economy?

To answer these research questions we must analyse the cars data-set. This can be done in numerous ways, one of them being visualisation. When a multidimensional data-set is being visualised one must choose an efficient and accurate way to represent the data.

2 Method

Visualising the data-set must give us an answer to the research question. To do so we must first define the type of data of each dimension. The *model* and *origin* dimensions contain nominal data. The remaining five dimensions (MPG, cylinders, horsepower, weight, year) contain quantitative data. The nominal data can be visualised in seven ways, the quantitative can be visualised in three ways [1]: position, size, texture. Position being the most accurate way to visualise (quantitative) data [2]. Thus our goal is to visualise all dimensions by position.

Therefore will visualise the data-set with a Parallel Coordinates Graph (PCG). This gives us the ability to represent each dimension by a position. This is straightforward for the dimensions containing quantitative data. However, for the nominal dimensions (origin and model) we must first map the nominal values to quantitative values in order to visualise them on a 1D-axis. The *model* dimension contains 301 unique values which means we must create an axis with 301 values. all of them representing another model, making the PCG harder to visually analyse. Besides the fact that the *model* dimension is not relevant to our research, we thus choose not visualise this dimension.

The *origin* dimension contains three unique values, which means we could create an axis with three different values. However we choose to not assign *origin* to its own axis but to visualise it by colour, making the result aesthetically more pleasing and more efficient to analyse. The remaining five dimension (MPG, cylinders, horsepower, weight, year) will have their own axis in the PCG, making the visualisation accurate [2].

We will create the PCG using a Python library called Plotly [1]. Since the PCG created by Plotly gives us the opportunity to interactively change the visualised range on each axis we do not have to give a specific order to the dimensions. By changing the range on one axis we will be able to have a more clear vision of one dimension and its relation to all the other dimensions, instead of just the relation between one dimension and its adjacent dimensions. This interactive-ability help finding an answer to our research questions.

3 Results

The result of our research is an interactive PCG (see cars.html), which is displayed in figure 1.

In our PCG we can see five different axes which all represent one of the dimensions containing quantitative data. One line in the PCG represents one specific car model and the line-colour represents the car origin:

- US: red
- Europe: blue

[1] Plotly: https://plot.ly/

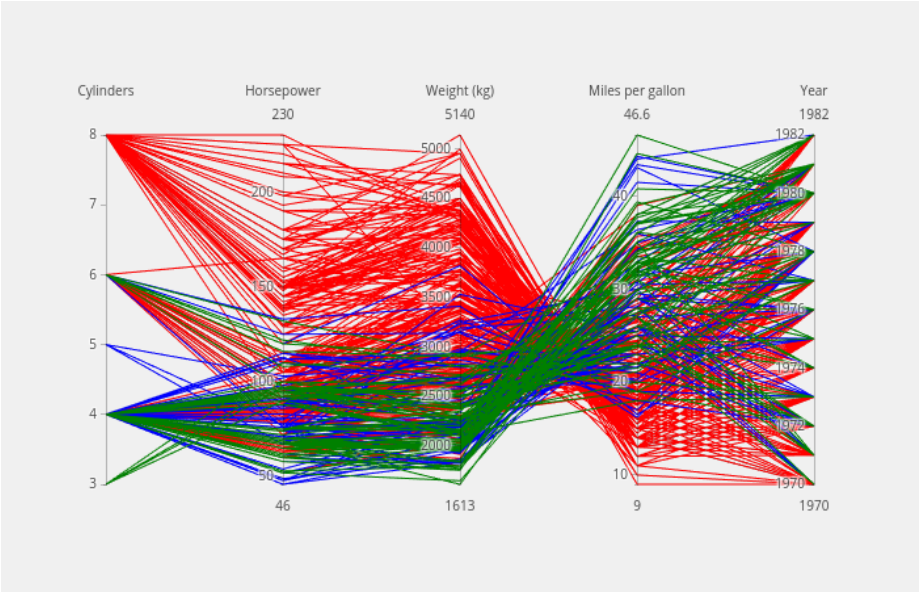

Figure 1: PCG

- Japan: green

Using the PCG we can draw some conclusions of the data-set at a glance. For example we can see that the data-set does not contain cars made with 7 cylinders. As well as the fact that all the 8 cylinder cars are originated in the US.

In the PCG we can see that the cars with the most horsepower are all manufactured in the US, which answers our first research question.

The PCG gives us the option to select a range on the different axes. This helps us in lowering the information density so we can find an answer to our second question.

If we change the range on the *MPG*-axis to the lower side (9 to 17) we are able that the less economical cars are usually heavier weight (see figure 2). In figure 2 we can see also see that the cars in this specific MPG range are mostly manufactured in the US.

To strengthen this observation we can select the higher side of the MPG axis (32 to 47) showing us that the more economical cars tend to be lighter (see figure 3). In 3 we can also see that these type of cars are mostly manufactured in Japan.

4 Conclusion

Given our results we can answer the research questions:

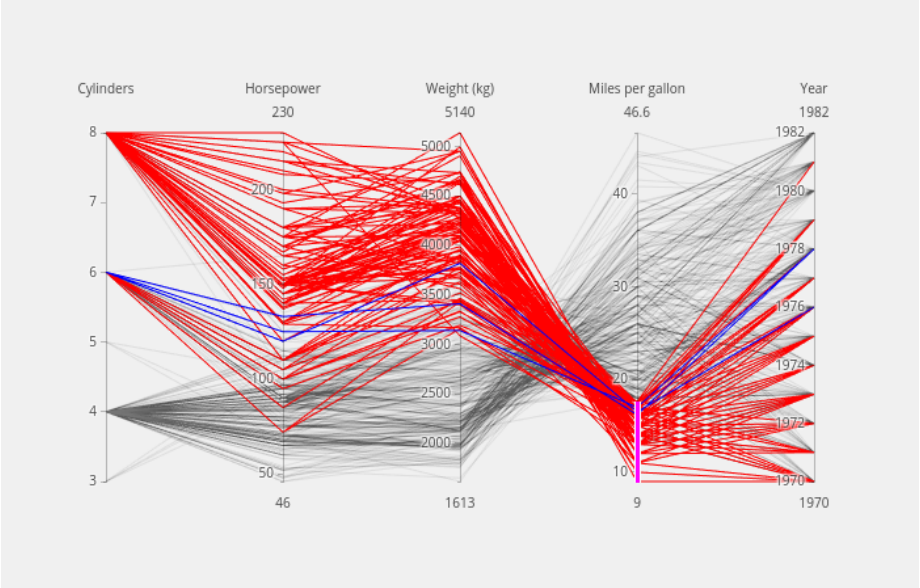

Figure 2: Cars with mpg between 9 and 17

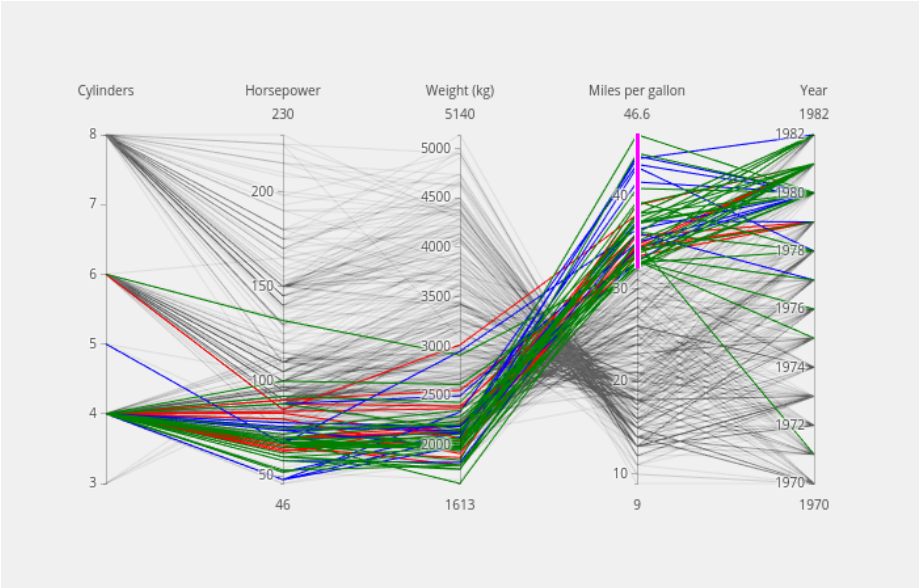

Figure 3: Cars with mpg between 32 and 47

1. **Which region manufactures the most powerful cars?**
   The most powerful cars, according to their horsepower, are manufactured in the US

2. **What is the relation between the weight of a car and its economy?**
   Heaver cars tend to be less economical than lighter cars.

5 Discussion

The PCG does not contain all the information in the data-set since we did not visualise the *model*-dimension. In future work we could make the graph more interactive by giving the viewer the ability to hover over the lines to show the corresponding *model*, as well as the exact value of the other dimensions to make it more clear to the viewer which values one line represents.

References

[1] Jacques Bertin. "Semiology of graphics: diagrams, networks, maps". In: (1983).

[2] Jock D Mackinlay. "Automatic design of graphical presentations". In: (1987).