Exercise 1

Cars Dataset

Lucas Piper - 14571773
Scientific Visualization and Virtual Reality
University of Amsterdam
14571773@uva.nl

ABSTRACT

In order to extract useful information from a multi-dimensional data set, it's often needed to visually represent the data. This enables a better understanding of the data as a whole and allows for improved clarity in the data analysis. In this work, a visualization application is proposed to create a visual representation of all data points within a data set of cars. Besides visualizing all data points of the data set, the visualization should also enable the representation of as much variables as possible in a clear and insightful manner.

KEYWORDS

Scientific visualization; Data encoding; Visual mapping; Visual attributes; Multidimensional data

1 INTRODUCTION

Information visualization and, more specifically, scientific visualization have been continuously shown to be a notorious tools for data analysis [Anscombe(1973)]. Moreover, as big data becomes more common and the complexity of data itself increases, scientific visualization also grows into a more challenging task [Mason(2019)].

A crucial step in the visual mapping process is the choice of visual attributes for the represented variables. This choice is highly dependent on the type of data of the variable. There are four distinct types of data: nominal, ordered, quantity-interval and quantity-ratio [Belleman(2022)]; and several possible visual attributes, according to Bertin's "Levels of Organization" [Bertin(1974)] and Mackinlay's "Ranking of perceptual tasks" [Mackinlay(2019)].

Besides the assignment of visual attributes, the clarity of multi-dimensional data representation is also dependent on the visualization tools used and the capabilities in simultaneously displaying multiple data dimensions. A useful framework to develop a visualization application for multi-dimensional data is Python with it's comprehensive libraries like Matplotlib [Hunter(2007)] and SeaBorn [Waskom(2021)] which allow the representation of data with up to six dimensions (and possibly more) [Sarkar(2018)].

Thus, in scientific visualization it's necessary to balance the need to visualize multiple dimensions at the same time and the clarity of what is meant to be visualized, raising the question *given a data set, how to represent multi-dimensional data in order to represent as many dimensions as possible and, simultaneously, maximize clarity?*

2 METHODS

To answer the aforementioned research question, the following approach was followed with the `cars.csv` data set that was provided:

(i) Assigning variables to data types
(ii) Determining variable representation priority order
(iii) Assigning each variable to the best available visual attribute (according to the variable data type and a visual attribute accuracy ordering)
(iv) Visualization creation

Firstly, each variable was assigned a data type. Numeric variables like `mpg`, `horsepower`, `weight` and `cylinders` were considered to be quantity-ratios, since it's possible to pinpoint a non-arbitrary fixed zero. On the other hand, `year` is a quantity-interval, given that that the location of the zero is arbitrary. For non-numeric variables like `origin` and `model`, it was assigned the nominal data type, given the lack of intrinsic order among the values and the inability to perform other operations besides = and ≠ between two values.

In the second step, the goal was to answer the question *what message should the visualization tell?* by establish an order of priority of the data items to represent. This was accomplished by analysing several scatter plots of the data (e.g. Figure 1) and searching for insightful relations between variables. In this stage the `model` variable was discarded, since the aim was to look for trends among the data with a higher level of abstraction. Afterwards, the priority ordering of Table 1 was determined, prioritizing quantity-ratios like `mpg`, `horsepower` and `weight` that show clear trends in the aforementioned pair-wise scatter plots. The following variables in the ordering are `cylinders`, `origin` and `year`, respectively, given the decreasing tendency to display interesting trends in the scatter plots. The `model` variable was left in the last position of the ordering for the same reason as before.

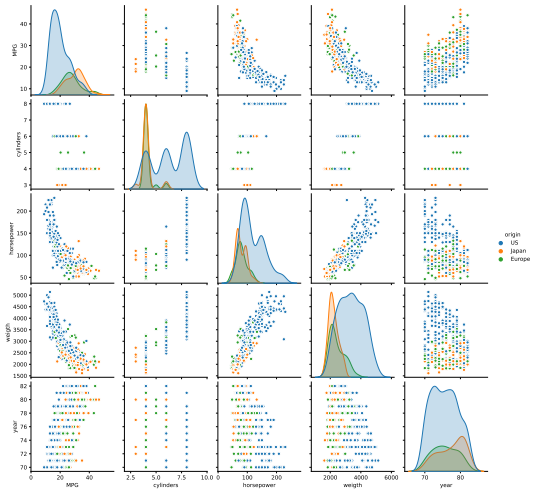

Figure 1: Pair-wise scatter plot of data set variables

Column	Data Type	Priority Ordering	Visual Attribute
mpg	quantity-ratio	1^{st}	Position x
horsepower	quantity-ratio	2^{nd}	Position y
weight	quantity-ratio	3^{rd}	Position z
cylinders	quantity-ratio	4^{th}	Area
origin	nominal	5^{th}	Shape
year	quantity-interval	6^{th}	Value
model	nominal	7^{th}	—

Table 1: Properties of data set columns

Once the priority ordering was established, the visual attribute assignment was done using Bertin's "Levels of Organization" and Mackinlay's "Ranking of perceptual tasks" as a visual attribute accuracy rankings. The most overall accurate visual attribute, position, was assigned to the first three variables in the priority ordering: `mpg`, `horsepower` and `weight`. This was done, assuming the use of the x, y and z axis of a three-dimensional frame of reference. For the forth variable, i.e. `cylinders`, the chosen visual attribute was area, given that there were no positional attributes left and that the length, angle and slope attributes from Mackinlay's "Ranking of perceptual tasks" are incompatible with the representation of a point in a reference frame. Furthermore, the shape attribute was assigned to the `origin` variable, since it's a nominal variable and the value attribute is necessary for the last quantitative variable (only position, size and value can be used for quantitative types of data). Additionally, the `year` variable can take more values than the `origin` one, being, as such, clearer to represent `year` with values as opposed to shapes. Therefore, from all the columns of the data set only six were represented in the visualization.

Finally, the resulting visualization was tested by ascertaining the visual clarity of the relations and trends among the data variables in comparison to pair-wise scatter plots previously mentioned.

3 RESULTS

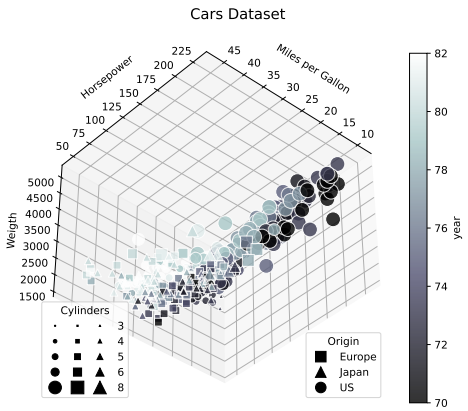

Figure 2: Visualization application

The final result is a an application developed with Python and Matplotlib's visualization tools. The `origin` variable was represented squares, triangles and circles for Europe, Japan and US, respectively. These markers have five different sizes depending on the value of the `cylinders` variable. Matplotlib's `bone` colormap was used for the value of the markers, given the lack of color saturation which allows for a better visualization of chronological evolution and the range of the colormap. The visualization application also allows for user interaction: within the 3D projection, the camera position can be adjusted, by dragging the mouse pointer.

4 DISCUSSION

As a mere static picture (e.g. Figure 2), the visualization of the data points and the trends that can be found in the data can be occasionally hard to visualize. Nevertheless, as an interactive application it's possible to change the camera position in order to achieve a better viewpoint of the frame of reference and, therefore, an improved visualization of the data, when a given relation between columns of the data set is meant to be visualized.

Beside the conceptual model for the visual representation, the representation power of the application is also limited by the capabilities of the visualization tools used. Further developments could be made that also take advantage of time or even more visual attributes. In the specific case of the data set used, a possible way to incorporate the `model` values in the visualization and not loose clarity could be through aggregation of values into categories and representation of values with a coarser granularity.

5 CONCLUSION

Despite scientific visualization being far from a trivial task, it's possible to easen the process and develop useful and insightful visualizations using the guidelines provided by Bertin's "Levels of Organization" and Mackinlay's "Ranking of perceptual tasks".

To create a visual representation of the `cars.csv` data set, for each of the variables, a data type, a priority ordering rank and a visual attribute was assigned. Moreover, in order to determine the priority ranking, for each pair of numeric variables, a scatter plot was created and trends among the data points we're identified. Finally an interactive visualization application was developed with Python and the Matplotlib library. From all the data set columns only `model` was not visually represented. This application attempts to answer the research question for the specific case of the data set at hands.

REFERENCES

[Anscombe(1973)] F. J. Anscombe. 1973. Graphs in Statistical Analysis. *The American Statistician* 27, 1 (1973), 17–21. http://www.jstor.org/stable/2682899

[Belleman(2022)] Robert Belleman. 2022. 1. Introduction. [PowerPoint Slides].

[Bertin(1974)] Jacques Bertin. 1974. *Sémiologie Graphique*. Paris: Editions Gauthier-Villars, Paris, FR.

[Hunter(2007)] J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9, 3 (2007), 90–95. https://doi.org/10.1109/MCSE.2007.55

[Mackinlay(2019)] Jock Mackinlay. 2019. Automating the Design of Graphical Presentations of Relational Information. (2019). https://research.tableau.com/sites/default/files/p110-mackinlay.pdf

[Mason(2019)] Betsy Mason. 2019. Why scientists need to be better at data visualization. (2019). https://knowablemagazine.org/article/mind/2019/science-data-visualization

[Sarkar(2018)] Dipanjan Sarkar. 2018. The Art of Effective Visualization of Multi-dimensional Data. *Medium* (2018). https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57

[Waskom(2021)] Michael L. Waskom. 2021. seaborn: statistical data visualization. *Journal of Open Source Software* 6, 60 (2021), 3021. https://doi.org/10.21105/joss.03021