Visualizing Car Cost-Effective: a Comparative Analysis of Fuel Consumption and Price for 2017 Vehicles in Australia

Yaoyu Liu*

The Faculty of Engineering, The of university of Sydney, Sydney, Australia, 2008

*yliu0430@uni.sydney.edu.au

Abstract. In today's big data era, the management and analysis of data is by far the most important. And the graphical visualisation of data can fully reflect the correlation between data. The dataset studied in this paper is the price of cars and fuel consumption in Australia in 2017. The cost-effectiveness of the car is measured by studying the city fuel consumption, highway fuel consumption, combined fuel consumption and car price. In this paper, using two software, tableau and Visual-paradigm, 3 visualisation charts are created by 3 the symbolic representation. The final visualisation can clarify the relationship between the data and finally conclude that 2017 Volkswagen Beetle is the most cost effective vehicle.

Keywords: visualisation, data, cost-effective, tableau

1 Introduction

The data used in this article comes from the Kaggle [1][2]. This original dataset has over 100,000 rows and 20 columns. The aim of this thesis is to examine the **cost effective price** of family type cars in Australia in 2017. The authors filtered the dataset for the final study of this paper by selecting 2017, compact cars, front-wheel drive and automatic. The specific dataset has 7 rows and 7 columns as shown on figure 1.

A	B	C	D	E	F	G
year	make	model	city_mpg	highway_m	combined_	price(K)
2017	Audi	Q3	20	28	23	35
2017	Ford	Focus FWD	26	36	29	29
2017	Honda	Civic 2Dr	30	39	34	21
2017	Kia	Forte Koup	25	32	27	30
2017	Mercedes-	CLA250	24	36	29	40
2017	Volkswage	Beetle	24	33	28	24

Fig. 1. The specific dataset

It is important to analyze the dataset to understand each features before visualization.

E. P. H. Lau et al. (eds.), *Proceedings of the 2024 3rd International Conference on Information Economy, Data Modelling and Cloud Computing (ICIDC 2024)*, Advances in Computer Science Research 114,
https://doi.org/10.2991/978-94-6463-504-1_4

- The first column shows the year in which the car was produced. For this dataset, it is harmonized to 2017.
- The elements in the second column represent the brand of the car. There are a total of 6 different brands in the dataset.
- The third column shows the model number of each sample. It represents the specific model of each car.
- The characteristics in column 4 represent the car's gas mileage in the city. MPG stands for 'miles per gallon', which stands for the number of miles per gallon in the city.
- Column 5 shows the car's fuel consumption at highway speeds. This stands for the number of miles the car gets on a gallon of fuel on the highway.
- Column 6 represents the combined fuel consumption of the car. It represents the distance per gallon that can be travelled in a combined environment.
- The last column represents the price of each car model in 2017. Price is an important factor that affects car sales [3]. Here the authors have standardized the units to K, implying that 3K would represent A$3000.

The data was collected in two ways. Firstly, fuel consumption data for each vehicle model was taken from the Vehicle Certification Agency (VCA), the executive agency of the UK Department for Transport [4]. The dataset cleansed and consolidated vehicle fuel consumption and emissions data from 2000 to 2017. Secondly, the price per car was taken from over 16,000 car listing records from various online platforms in Australia.

2 Stakeholder

The objective of this paper is to study the cost effectiveness of automobiles. And the dataset of this paper is about the fuel consumption as well as the price of the car, so the stakeholders of this dataset are vehicle manufacturers, vehicle dealers, and consumers who have the need to buy a car.

- **Vehicle manufacturers**: Manufacturers need to understand the final position of the vehicles they produce in the marketplace to analyze whether they need to make enhancements or changes. For example, if the current vehicle has high fuel consumption and is very expensive, then the corresponding product is not as competitive and needs to be improved.
- **Vehicle dealers**: Dealers need to compare the fuel consumption and price of competing products to make appropriate sales strategies. Such as the same fuel consumption, the price of the competitor product is lower, then the dealer should also reduce the price.
- **Consumers**: Fuel consumption and price are two important criteria for purchasing a vehicle in the eyes of consumers [3]. So consumers will also compare the data in the dataset to decide the final result.

3 Data Types

Researching data types is an important step in data visualization. It allows the researcher to go for the right type of graph to avoid misleading or misinterpreting the data [5]. Since no examples of applicable visualizations were found at the source [1] [2], this paper starts with a basic visualization (Figure 2) to facilitate the user to understand the data type in the early stage of the research.

- **Years** can be considered as interval data because the differences between years are uniform and meaningful[6] . However, there is no natural order to the years themselves and it only represent a specific identification of a period of time. Year is generally used as a scalar and is represented in this dataset as the year in which the car was produced. So in Figure2 Years is used as dimensions of attribute.
- **Car make** and **model** belong to nominal data. They only serve as a classification of different cars without any order [6]. And their order is irrelevant [6]. This dataset is a dataset containing car makes and models, so they are represented as tensor dimensions of attribute in Figure 2.
- The fuel consumption of a car, including **city** fuel consumption, **highway** fuel consumption and **combined** fuel consumption is ratio data. Similarly, the **price** of a car is also ratio data. This is because they are numerical data and have specific values [6].
- They can be compared with each other, for example, the price of an Audi Q3 is twice the price of a Volkswagen Beetle. They are both Integer data and appear as quality of attribute in figure 2.

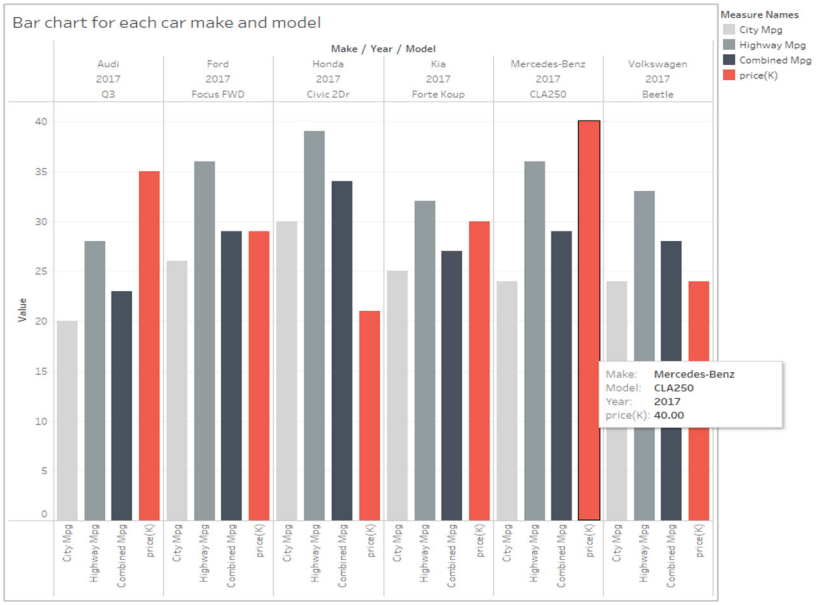

Fig. 2. Basic visualization

4 Question

The KGI in this paper is to study the cost effective of different brands of vehicles. KPI is applied to the fuel consumption of the car and the price of the car which are ratio data.

The following questions can be asked around the topic, visualization and dataset.

1. Which make and model of car can be observed to have the highest and lowest price in figure 2?

   Answer:

   From the top right corner of figure 2, people can notice that the red color represents the price, so the **Mercedes-Benz CLA250** is the car with the **highest** price, and **Honda Civic 2Dr** has the **lowest** price.

2. Which car in figure 2 has the highest fuel consumption?

   Answer:

   By observing the figure 2, it's simple to understand that the **Honda Civic 2Dr** has the **highest fuel consumption** in both city, highway, and combined fuel consumption.

3. Is the car with the lowest price the car with the highest cost effective price?

   Answer:

   No, The Honda Civic 2Dr is the cheapest car, but it also has the highest fuel consumption.

4. Which graphs are nominal data, ordinal data, interval data, and ratio data applicable to? And for which graphs is the data set of this paper applicable?

   Answer:

   This paper examines the cost effective of a car, so the main thing that needs to be visualized is the fuel consumption as well as the price of the car. Since these are ratio data, this paper mainly applies to scatter plots, bubble plots, area plots. These charts help the observer to understand the proportional relationship and trend of fuel consumption among different cars.

   In visualization, nominal data is usually used in bar charts, pie charts, stacked bar charts, which helps the user to understand the distribution between different categories [6].

   Ordinal data is usually used in bar charts, line charts, radar charts, which can understand the sequential relationships between different categories [6].

   Interval data used in histograms, box plots, line plots,, which is easy to understand the distribution, trends and changes in the data [6].

5 Common Mistake

Figure 2 is a basic visualization showing all the data. However, the current dataset is not sorted by values because of the large number of features. Therefore, due to visual limitations, people are prone to make mistakes or draw too extreme conclusions when analyzing the data.

- **Large data**: there are a total of 24 values and 4 colors in figure 2. Observers may have difficulty analyzing the relationship between different models because of the large amount of data.
- **Misunderstanding correlations**: Without visualization, people may incorrectly scale the relationship between individual features. For example, they may assume that price is inversely proportional to fuel efficiency, and that a car with lower fuel consumption is more expensive like the Honda Civic 2Dr. In fact, this is not always the case, and there are many other factors that affect the price, including the engine, space, vehicle system, etc. [7].
- **Misplaced Focus**: people have limited vision and may focus too much on one feature and ignore other important information. For example, they may focus on combined fuel consumption and ignore city and highway fuel consumption.
- **Lack of intuitive understanding**: Suitable data visualization can help people understand the data more intuitively [8]. However, figure 2 is just a generic visualization that does not show the correlation between vehicle data, and it may be more difficult for people to understand what is represented behind the data. For example, it remains difficult for customers to understand the cost effective of each car brand.

6 Symbolic Representations

In order to improve the visualization, it is very important to use graphical symbolic representation first. It also makes the data more analyzable.

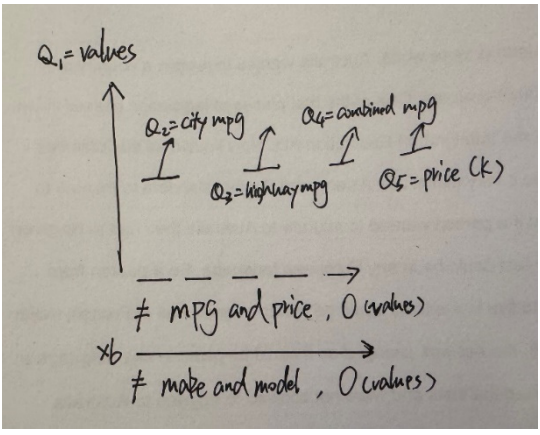

Fig. 3: Semiology of bar chart

Figure 3 shows the symbolic representations of figure2. Although semiology of graphics cannot show specific data, it can help observers to more readily discover relationships between data by mapping the data onto different symbols and visual variables.

First, the vertical coordinates indicate the values for the different quantities. Set values as a quantities perception(Q1). Second, the horizontal coordinate is represented by the 6 different car make and model. It shows the mpg and price of each car make, repeated 6 times. Finally, each car make in figure2 shows 4 different data. So in figure 3 these 4 data are drawn as a separate quantities perception. these 4 quantities are city_mpg(Q2), highway_mpg(Q3), combined_mpg(Q4) and price_K(Q5).

From the analysis of figure3 above it is easy to see that the figure2 visualization contains all the data. But it's still just a basic visualization with lots of elements in it. As can be seen from figure3, there are total 5 quantities perceptions and is repeated 6 times. This results this visualization has too many variables and cannot express the association between each of them.

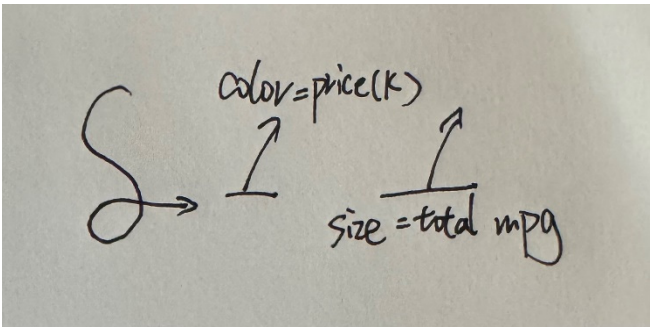

Fig. 4: Semiology of tree map

As shown in the figure above, figure4 is an improved map of figure3. figure4 is a semiology of tree map. Because there are too many variables in figure3, it is hard to see the comparisons and correlations. The tree map can represent the data by size and color, so the tree map is suitable for this dataset. In figure4, the sign on the left indicates that this is a formal diagram of a tree, neither a line nor a circle. There are a total of two variables in it. One is the color representing the price of the car, the darker the color the more expensive the car is. The other variable is size which represents the total fuel consumption, the larger the size the higher the fuel consumption.

Figure2 has more variables. With in figure4, the observer only has to focus on two variables, color and size. Tree maps are useful when there is a large amount of hierarchical data, and show the relationship in a more intuitive way.

7 Graphical Interpretation

Figure 5 is a visualization generated from semiology figure 4. The resulting visualization contains a series of rectangles, each representing a node [9]. Their size and position

indicate the data values represented by the node as well as the hierarchical relationships [9]. Each of the features in the data has a different meaning and visual variable.

- **Year**: Years is Interval data. In figure5 year is used as a separate dimensions of attribute. It is used as a label to indicate the year of each node in the graph.
- **Make** & **Model**: Car make and model are both nominal data. as dimensions of attribute in figure5. It's as two different classification labels that represent the car make and model for each node in the graph.
- **Total_mpg**: In the original dataset there is no such feature as Total_mpg. but by looking at figure2, it is not difficult to notice that there is not much variation in the city_mpg, highway_mpg and combined_mpg. In every car sample, highway_mpg is the largest and city_mpg has the smallest value. So when making the tree map, the authors added the three fuel consumption data into an overall fuel consumption, represented by Total_mpg. And in figure5, Total_mpg is represented by the size of each node. This makes it easy to understand the comparison and relationship of different car makes by the size of the rectangle.
- **Price (k)**: Price is in the dataset as ratio data. its present as a color variable in figure5. The darker the color of the node, the more expensive the car brand represented by the current node.

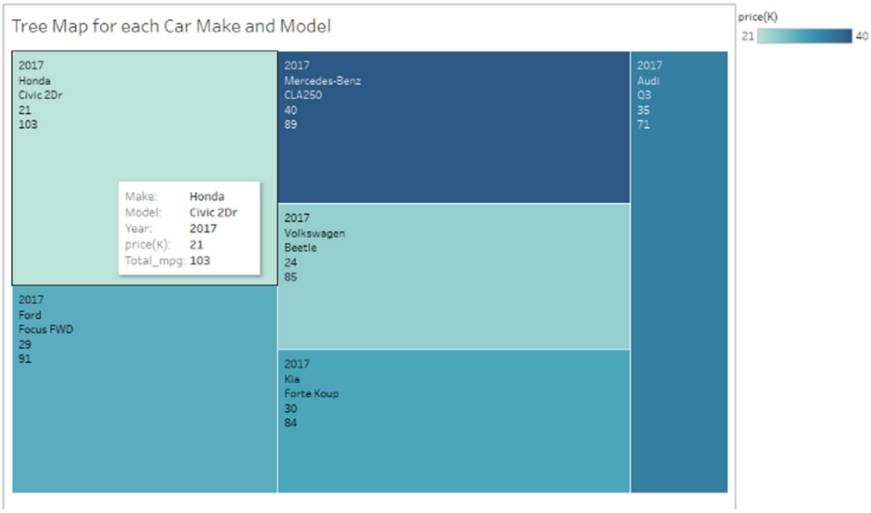

Fig. 5. Tree map visualisation

8 Implementation Methodology

Through semiology of figure4, this project was further visualized to get figure5.

For visualization, Tableau app is used in this project [10]. Tableau is a professional and efficient visualization software [10]. The steps of visualization are as follows:

- **Import data**: Import preprocessed data (as figure1) into tableau.

- **Observe the semiology of graph**: Through figure4, it is clear that the result of the current visualization is a tree map. price is used as the color variable and total_mpg is used as the size variable.
- **Calculate total_mpg**: Create a new variable Total_mpg in tableau. add city_mpg, highway_mpg and combined_mpg together to make it equal to Total_mpg using tableau's built-in operator "+".
- **Setting parameters and variables**: Set year, make and model to dimensions and labels. Then set price to the color variable and Total_mpg to the size variable.

As in figure 5, specific values for the current sample are shown in each node. These include year, car make, model, price and total_mpg. It also has a simple interactive feature that displays detailed information when the observer mouses over the node.

9 Equivalence Method

Tree map has become more and more popular in recent years, but still some users cannot accept it [9]. For example, in figure5 the comparison of the size of each sample node is not obvious, just like Kia and Volkswagen. This project tries to use a different kind of graph that shows the correlation between the data and is also acceptable to the observer.

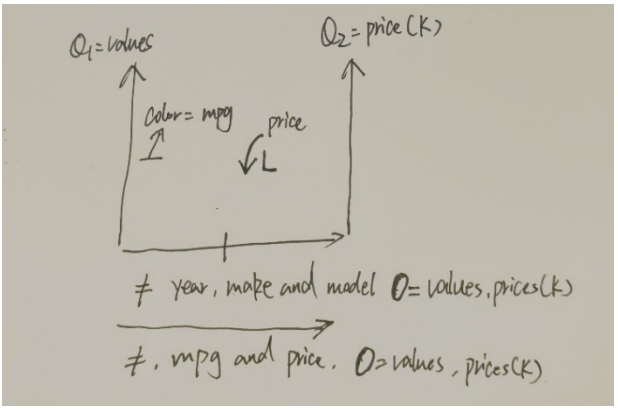

Fig. 6: Semiology of combination chart

First, the horizontal coordinates are the same as in figure3, by using year, car make and model as dimensions. Its values and price will be represented by different variables in the figure. But the **only difference** is that these variables will be superimposed on the graph.

Secondly, there are two vertical coordinates, the left one represents the value of mpg and the right one represents the price of each car.

Finally, there are a total of two variables in the graph. One is mpg, which is **similar** to figure4's Total_mpg, but its data will be generated in a different color overlay. The other is price, and this variable will be represented by a linear representation.

Figure6 is actually **equivalent** to figure4 but with a different notation. figure4 is a tree map and figure6 is a combination of a bar chart and a line graph. There are 2 variables in the tree map. One is price and the other is Total_mpg, which are color and size variables respectively. Whereas in figure 6, price is represented by a linear representation, mpg is represented by a bar chart using a stacked approach. And mpg will be composed of different colors.

10 Visualisation

For implementing the visualization of combinatorial graphs, this project uses Visual-paradigm [11]. Similarly, first import the data, then observe the semiology of graph (figure6), set the variables mpg and price, and finally complete the visualization.

Compare figure2 and figure7: in figure2 there are too many variables, only the colors are different between the variables, there are 6*4 bars in total. Whereas in figure7, there are only two variables, one is stacked mpg and one is price. and both variables one is bar one is represented by linear. In this way, figure7 can represent the relationship between the data in a clearer and more concise way.

Compare figure5 and figure7: the two diagrams are completely different but equivalent. However, when the values are similar, figure5 does not compare as well, and figure5 only shows the value of total mpg. In figure7, the highs and lows are visible even for similar values, and the total mpg is formed by stacking 3 mpg. This makes the visualization in figure7 clearer and more rigorous.

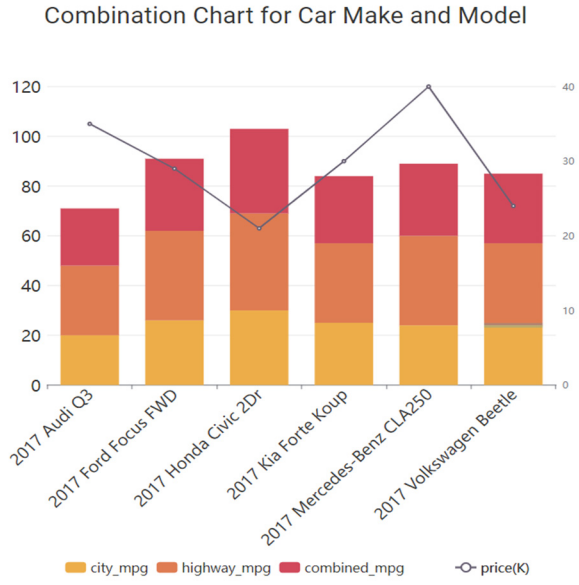

Fig. 7. The combination chart

11 Analysed

It is possible to derive from analyses the figure7:

- The 2017 Audi Q3 has the lowest fuel consumption, but is more expensive.
- The 2017 Ford Focus FWD is priced higher and gets higher fuel consumption.
- The 2017 Honda Civic 2Dr has the lowest price but the highest fuel consumption.
- The 2017 Kia Forte Koup has higher gas mileage and price.
- The 2017 Mercedes-Benz CLA250 gets better gas mileage and has the highest price tag.
- The 2017 Volkswagen Beetle has a lower price and fuel consumption and is not far from the minimum.

12 Conculsion

This paper shows 3 specific visualizations of the current dataset. Semiology of graph, bar chart, tree map and combined graph are used.

From the analysis and visualization kind of the above graph it can be concluded that in the current dataset 2017 Volkswagen Beetle is the most cost effective car product.

Acknowledgment

Thanks to Kaggle for providing data support for this project.

Thanks to Tableau and Visual-paradigm for visualization help for this project.

References

1. N. Elgiriyewithana, “Australian vehicle prices.” 27-Nov-2023.
2. A. Samoshyn, “Car Fuel & Emissions.” 03-May-2020.
3. J. M. Sallee, S. E. West, and W. Fan, “Do consumers recognize the value of fuel economy? Evidence from used car prices and gasoline price fluctuations,” J. Public Econ., vol. 135, pp. 61–73, 2016
4. “Fuel consumption & CO2 databases,” Vehicle Certification Agency, 07-Jan-2020. Available: https://www.vehicle-certification-agency.gov.uk/fuel-consumption-co2/.
5. Saquib, Nazmus. Mathematica Data Visualization. Packt Publishing, vol.1.1 pp. 21–31, 2014.
6. Velleman, P. F. and Wilkinson, L. ‘Nominal, Ordinal, Interval, and Ratio Typologies are Misleading’, The American Statistician, 47(1), pp. 65–72, 2012
7. Balce, AO, 2016, “Factors Affecting Prices In An Used Car E-Market,” İnternet Uygulamaları ve Yönetimi Dergisi, vol. 7, no. 2, pp. 5–20.
8. Ahmad, A, Leifler, O, & Sandahl, K, 2022, “Data visualisation in continuous integration and delivery: Information needs, challenges, and recommendations,” IET Software, vol. 16, no. 3, pp. 331–349.

9. Lu, L, Fan, S, Huang, M, Huang, W, & Yang, R, 2017, "Golden Rectangle Treemap," Journal of Physics: Conference Series, vol. 787, no. 1, pp. 12007
10. "Tableau: Business intelligence and analytics software," Tableau. Available: https://www.tableau.com/.
11. "Ideal modeling & diagramming tool for Agile team collaboration," Visual-paradigm.com. Available: https://www.visual-paradigm.com/.