Choosing the suitable chart to present comparison about multiple-variable data

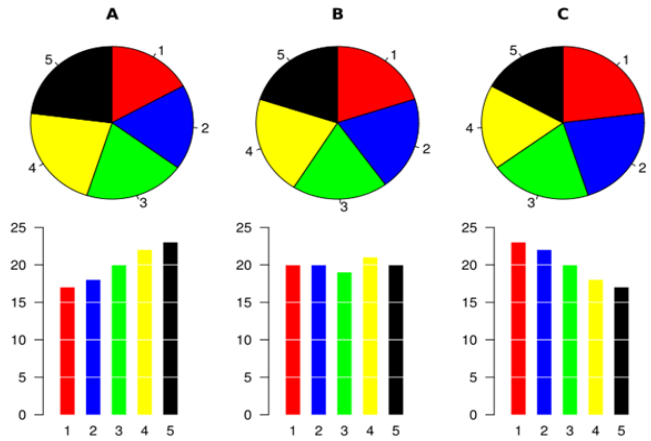
Student Name(s) WEI LAN , ZENG CHEN
Student Number(s)15211145 , 15211282

1 Proposed Hypothesis (What Question Are You Asking?)

Our proposed hypothesis: when changing different type of charts in multiple-variable data visualization, there will have the different accuracy and speed of answering the related questions.

The question we want to ask is what kind of chart can be best choice to show the multiple variables in terms of the effectiveness of the proposed multi-variables visualization technique. Multiple variables showed in a chart is very popular and it could use one chart to show majority information. People can make a decision on choosing the best one by comparing multiple variables values. Whether the user could understand and find the information show to them, the choice of the chart is very important.

Who doesn't love pies or donuts, right? Not in data visualization, though. These charts are among the most frequently used and also misused charts. The one on the right is a good example of a terrible, useless pie chart - too many components, very similar values. A pie chart typically represents numbers in percentages, used to visualize a part to whole relationship or a composition. Pie charts are not meant to compare individual sections to each other or to represent exact values (you should use a bar chart for that). When possible, avoid pie charts and donuts. The human mind thinks linearly. When it comes to angles and areas, most of us can't judge them well. Source: Oracle.com[5]



Map charts are good for giving your numbers a geographical context to quickly spot best and worst performing areas, trends, and outliers. If you have any kind of location data like coordinates, country names, state names or abbreviations, or addresses, you can plot related data on a map. Maps won't be very good for comparing exact values, because map charts are usually color scaled and humans are quite bad at distinguishing shades of colors. Sometimes it's better to use overlay bubbles or numbers if you need to convey exact numbers or enable comparison. A good example would be website visitors by country, state, or city, or product sales by state, region or city. But, don't use maps for absolutely everything that has a geographical dimension. Today, almost any data has a geographical dimension, but it doesn't mean that you should display it on a map [5].

From the research above, it is clear to find that if we do not use the suitable chart to show multiple variable in the chart, we could not get the information from the chart accurately. So choosing the proper chart to show the multiple variable data is important. To solve this problem, we decided to use some charts to show the Multi-variables data , Moreover, the question we asked is whether the feedback getting from the different specific group of audiences about the performance of Multi-variables visualization technique will be significantly different. For the multiple variables domain of cars, it might be interested for analyse the performance of the

car. Which would give some feedback to people who interested in buying cars. For example, different people like different models of car, for the same car model, they might be consider the weight or the MPG (mile per gallon) of the car. What's more, the price of the car might be also considered. So using the chart to show multiple variables would give some information to the users at the same time. Meanwhile, showing these variables properly on the chart is very significant.

Additionally, we want to get some information about for the Multi-variables chart, what would the user focus on and what information they want to know from the chart, but it does not show well on chart, we want to try our best to create chart for Multi-variables data to more intuitive and readable. Based on the response time, the score which the respondents give to each chart and the accuracy of the respondents' answers which asked in questionnaire, we could get the basic idea about what kind of chart is better to performance the multiple variables.

One of the recent studies on chart perception is by Namee and Zubiaga [16] who proved that bar chart is the right chart to present one variable data distributions for non-expert users. Keim [17] summarized five visualization techniques used to display multiple variables: (1) standard 2D/3D displays, for example, bar charts and x-y plots; (2) Geometrically transformed displays; (3) Icon-based displays, such as star icons; (4) Dense pixel displays, such as circle segments and recursive pattern; (5) Stacked displays. Furthermore, the literatures does not have some studies on how well the viewers understand or interpret different charts which show comparison about multiple-variable data.

2 Experimental Method

2.1 Overview

We decided to make a survey, in order to make more people to participate in the experiment to support our hypothesis and it focus on verifying what kind of charts to display multiple variables can be more intuitive and readable, especially, use some data charts to show some data. In the survey, some different type charts will be displayed on a questionnaire.

By inviting some University College Dublin students (who study in different subjects), using survey [1] can be developed in less time (compared to other data-collection methods) and it cost-effective, but cost depends on survey mode it can be administered remotely via online, email or social apps(Facebook, slack, tweeter etc.) Conducted remotely can reduce or prevent geographical dependence and capable of collecting data from a large number of respondents, for some questions can be asked about a subject, giving extensive flexibility in data analysis. With survey software, advanced statistical techniques can be utilized to analyze survey data to determine

validity, reliability, and statistical significance, including the ability to analyze multiple variables, a broad range of data can be collected.

In this experiment, the following variables are included in our experiment:
1. Independent variables: 4 different chart types: stacked histograms, Lines chart, small multiples, Radar chart; 3 different types of questions (total 5 questions): (1) The highest or lowest values and comparison: the questions were of the following two types: (i) "which car is the best value?", (ii) " which car had the worse value? ". (2) For each chart type, two versions of statement type were shown to participants: one that was true and one that was false. Meanwhile, there was an option for impossible to tell. For example, the true statement " Car A is the better than B " and the false statement "Car B is better than A ". (3) Meanwhile, for each chart, viewers will scale the level of readability of the chart.

2. Dependent variable: measuring the accuracy and speed of answering questions

3. Controlled variables: capability of using computer; gender; age; the same colour for all charts

4. Confounding variables: To be get the most accurate output as much as possible, we take into account confounding variables, which can influence measurement of independent variable. In our case the confounding variables considered are: 1) previous experience; 2) learning effect; 3) colour blindness.
In order to begin to answer our questions, there were few basic information about studying background and some pictures that identified what's number under the colour pictures. We need to avoid some participants who study statistics or any related with data subjects, which would help us to avoid previous experience. Generally, we also had some colourful pictures to test the participants colorblindness. Meanwhile, Task was presented in random order to control for learning effect.

The combination of chart types (4) and the questions (4) amounted to a total 16 different tasks. In other words, the participants should answer the same questions (4) for each chart type. So, the experimental condition is to use the different charts in data visualization and the visualizations will be randomly shown for each participant. And then the participant will be presented a set of questions to answer with his opinion regarding with the charts. Comparing them to the ground truth and checking the accuracy and calculating the time to compute the speed of answering questions in each different chart will analyze the answers.

To acquire available data, respondents' opinions were surveyed through the use of a questionnaire of survey involving 100 respondents from different kinds of majors were conducted in UCD. Survey can be developed in less time and administered remotely via online, email or social apps. Meanwhile, the questions can be asked about a subject, giving extensive flexibility in data analysis by survey. Overall, this survey experiment investigates and analyses what kind of chart is good for Multi-variables improve the intuitive and readability of Information Visualisation. We will use our dataset and we will use within group for experiment approach. Because it does not require large number of subjects and less participants are needed.

2.2 Data collection

We will use Google form to collect the questionnaire. Google form is a powerful and free way to collect the numbers of questionnaires in a short time. The questionnaire, which builds in Google form, contains some different type of charts; some of them are objective question for respondents to give score to each chart to show multiple variables. According to the score, which the respondents give us, we could find which chart is better to show the multiple variables. Meanwhile, we set some questions, which relative to the chart, according to the accuracy of the answers, we can find which chart is better. For the response time, the time they used to answer each questions for each chart would be recorded by software automatically. According to their response time for the chart to read the data, the intuitive of the chart could be calculated. What's more, some relative questions would be asked to respondents. Based on the accuracy of the answers, we can get whether the chart is clear to show the data.

We would collect data about the participants' performance from the experiments in the following ways:

1. Accuracy to evaluate how well participants answer questions matches the ground truth.
2. Speed of answering the questions from the first visualization to the last data visualization by measuring time taken for their responses in each chart type.
3. The scale which respondents give us to rank which one is better to show the multiple data.

To compute the accuracy, we initially depend on majority voting that has been chosen by most participants. The final accuracy values we reported was that the total correct answers by match ground truth. For the scale, which respondents give us to consider, which chart is better on visual effect. The response time could show whether the chart is easy to get the data. Different charts have their own advantages and disadvantages. Combining these three conditions could give us the basic thought about which chart is better to show the multiple variables.

Objective measurements are based on how well the participants' performance, such as task completion time, percentage of task completed and errors percentage and so on. On the other hand, subjective measurements are to measure how participants feel or say they actually experience, such as, ease of use and user satisfaction and so on.

In our experiment, using questions in the section 3 data visualization could identify how well the participants performed and using the scale information about chart types in section 3 to see how easily they used different chart types . So we used the mixed objective measurements and subjective measurements.

From data we collected, we could get the information about how people perceive and understand the information through use the different chart types. We could notice that changing different chart types increased or decreased their speed of

answering questions and accuracy. Meanwhile, we can see from the confusion matrix that which chart types had the lowest error made by participants.

According to analyze the data collected, this experiment helps in providing our proposed hypothesis by recording speed of answering questions and the accuracy that the questions answered correctly.

2.3 Selected subjects

Subjects: are the people in the researcher's experiment - usually quantitative research. (Example: in a medical experiment the control group of 10 subjects did not receive the medicine, while the experimental group of 10 subjects received the medicine.) Subjects are a term used more in science. Subjects are generally a more passive term (Example: Ten subjects were given the behavior therapy) [3].

A representative sample is a small quantity of something that accurately reflects the larger entity. An example is when a small number of people accurately reflect the members of an entire population [4].

I will use 20 of students in UCD to be the subject (10 male and 10 female), and select them by different ages and different majors, the data visualization will show to different people and different platform, they could give the answers with the different ideas. Firstly I will non-random select the students from different majors and ages, then randomly select the responses of 20 students for analyzing. Choose them because they are representative, in the same professional may be bound by some experience, it would not good and not accurate for the conclusion, choose different subjects who has different background would get better result for analyzing. The rest of the students as the respondents support the analysis.

2.4 Data analysis

First, we use some pictures to identify the colorblindness because we use some different color in charts. According these pictures, we will choose the questionnaires, which have the right answers to these questions. Moreover, we will not choose the people who major is relative to data, like computer science , statistics and so on. Then we will use analysis of variance; the analysis of variance can be used as an exploratory tool to explain observations. "Classical ANOVA for balanced data does three things at once: As exploratory data analysis, an ANOVA is an organization of an additive data decomposition, and its sums of squares indicate the variance of each component of the decomposition (or, equivalently, each set of terms of a linear model). Comparisons of mean squares, along with an F-test allow testing of a nested sequence of models. Closely related to the ANOVA is a linear model fit with coefficient estimates and standard errors [11]." In short, ANOVA is a statistical tool

used in several ways to develop and confirm an explanation for the observed data. Additionally, it is computationally elegant and relatively robust against violations of its assumptions. ANOVA provides industrial strength (multiple sample comparison) statistical analysis. It has been adapted to the analysis of a variety of experimental designs. As a result: ANOVA "has long enjoyed the status of being the most used (some would say abused) statistical technique in psychological research [12]". ANOVA "is probably the most useful technique in the field of statistical inference [13]". ANOVA is difficult to teach, particularly for complex experiments, with split-plot designs being notorious [11]. In some cases the proper application of the method is best determined by problem pattern recognition followed by the consultation of a classic authoritative test [11].

Storing the answer to the question for each chart, we want to compare the several matched groups and we can assume the differences between any two options are equal; we treat them as interval data. We use analysis of variance (ANOVA) to analyse the data. Analysis of variance is a collection of statistical models used to analyze the differences among group means and their associated procedures (such as "variation" among and between groups). This analysis can help us to understand the each chart ability to reflect the information under different condition [18]. We record time consumption when answer the same question in different charts. This analysis relate to the time consumption, which can show the result that different chart are appropriately, apply to solving different question. Selecting each respondent's scores to each chart can demonstrate that the standard for visualization are different, according to different scores they give us we can get which chart is better for showing Multi-variables.

2.5 Practical setup

For this experiment, a survey will be conducted in online method by questionnaire (Google form) with 100 respondents from different kinds of subjects in UCD. The visualizations will be displayed on screen through the questionnaire. Before the actual study, we will introduce the types of chart. After the introduction, the respondents will have the opportunity to familiarize themselves with the charts and to ask questions concerning the concept and interactions by E-mail, Facebook or other apps. In addition to measuring task completion time for the answers, we took notes on the participants' approaches to the tasks and what problems they encountered.

The step of this experiment [2]:

1. Formulate the survey keeping in mind your overall substantive and analytical needs: I want to have an experiment about what kind of data use what kind of chart will more intuitive for reading, for example, if we want to know the trend of the whole data, use bar chart or use line chart is better to read.

2. Determine specifically what mode of collecting the data will be used: the survey online can collect the survey by many ways like email or by apps(like Facebook and Skype).

3. Determine an appropriate sampling plan: I choose students in UCD because they are different kinds of people, they majored in different subjects and some of them from different countries and at different ages, they can be treated as random sample.

4. Develop the questionnaire: A questionnaire was created by use some data to draw some charts for respondents to choose which kind of charts is more intuitive and whether they have some suggestions about use other charts would be more intuitive they can contact us.

5. Execute the survey in the field: it would have some missing data; I decide to use the majority answers to fill the missing data.

6. Edit and process the data: edit the data into form, it is better for analyze data

7. Analyze the data: Based on the result of the questionnaire to analyze the better way to show multiple variable data.

8. Conclusion: Based on the result to make conclusion what kind of data is better to use what kind of chart is more intuitive and readable.

9. Result: Develop our findings and conclusions to write up a summary of what has been found and how to create chart for multiple variable data would be more intuitive and readable.

3 Data Visualizations

The sample questionnaire which used for this experiment:

Dear respondents,

Hello, we are students who major in computer science, as we know, data has an important role in our life, different charts would focus on different parts of data, we want to have an experiment on what kind of chart to show the multiple domains data would be more intuitive and more readable, through this survey, your answers are helpful for us to have a basic thought for the data. Your response will only be used for survey. If you have some questions and suggestions please contact us on lan.wei@ucdconnect.ie or chen.zeng@ucdconnect.ie Thank you for your time.

Basic information:

Gender: male female

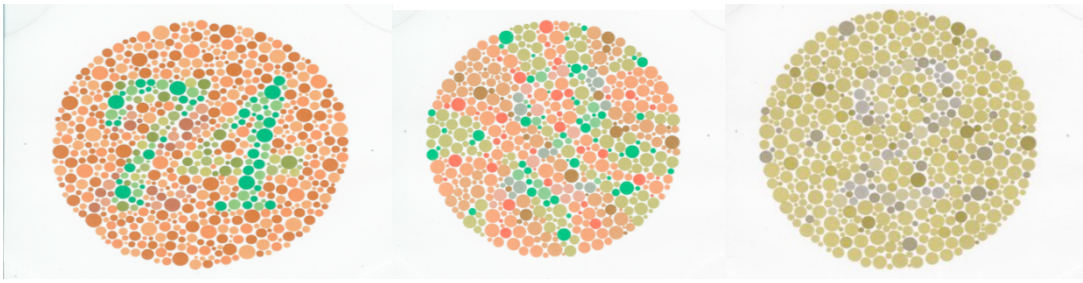
Age: _____

Major:_____

Where are you from:_____

Pictures [15] that could identify the colorblindness:

Question: What number did you read from the picture?



Data visualization:

There are 3 different questions for each chart:

1). Which car model is the best one? 2)Which car model is the worst one?

Note: A good car has the lowest MPG, the lowest price and the lowest weight. A bad car has the contract values.

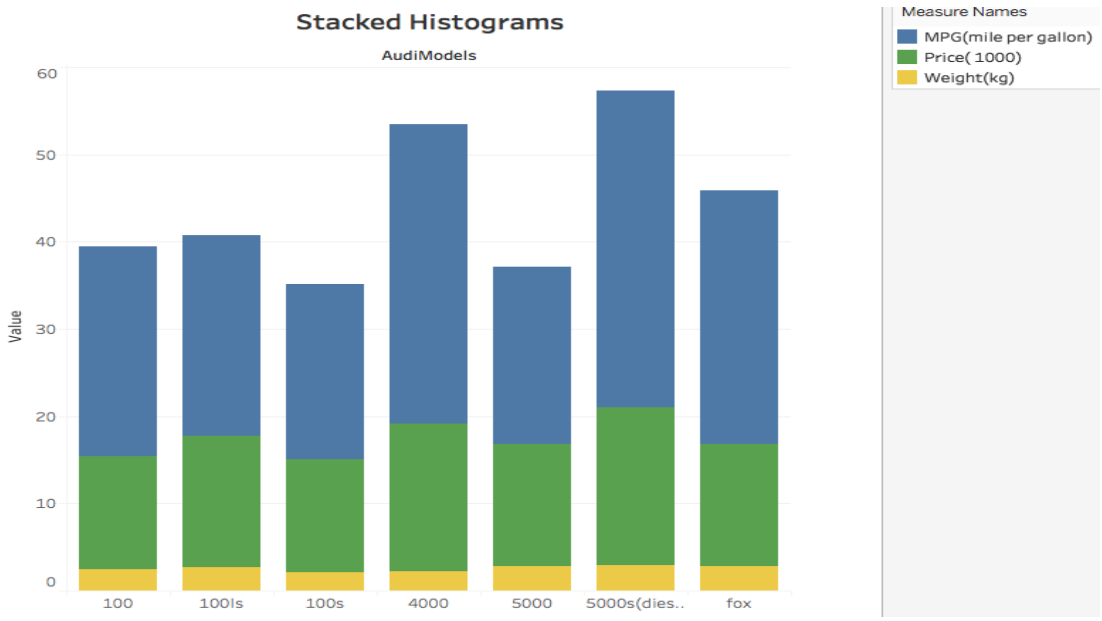
3). Please answer the questions below, and give the score to each chart according to the scale.

1	poor
2	fair
3	good
4	very good
5	excellent

There are also 2 statements were shown in the below each chart.

Images that we used for our experiment under the experimental conditions that change the different chart types to answer the 2 questions and 2 statements.

Visualization 1:



Question 1: In stacked histograms, which car model is the best one?

100	1001s	100s	4000	5000	5000s(diesel)	fox	hard to tell

Question 2: In stacked histograms, Which car model is the worst one?

100	1001s	100s	4000	5000	5000s(diesel)	fox	hard to tell

Statement 1: Model 5000 is better than Model fox.

True	False	impossible to tell

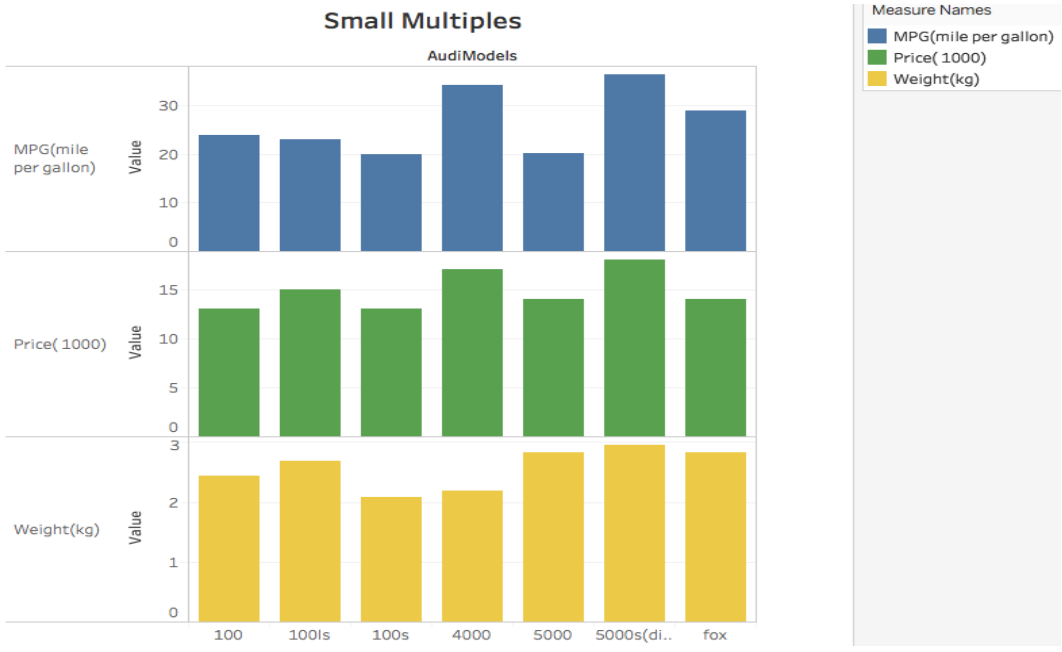
Statement2: Model 5000s is better than Model 5000.

True	False	impossible to tell

Question 3: On a scale of 1 to 5, what would you say is the ease of reading visualization 1, when the numeric values correspond to the following:

1	poor
2	fair
3	good
4	very good
5	excellent

Visualization 2:



Question 1: In small multiples chart, which car model is the best one?

100	1001s	100s	4000	5000	5000s(diesel)	fox	hard to tell

Question 2: In small multiples, Which car model is the worst one?

100	1001s	100s	4000	5000	5000s(diesel)	fox	hard to tell

Statement 1: Model 5000 is better than Model fox.

True	False	impossible to tell

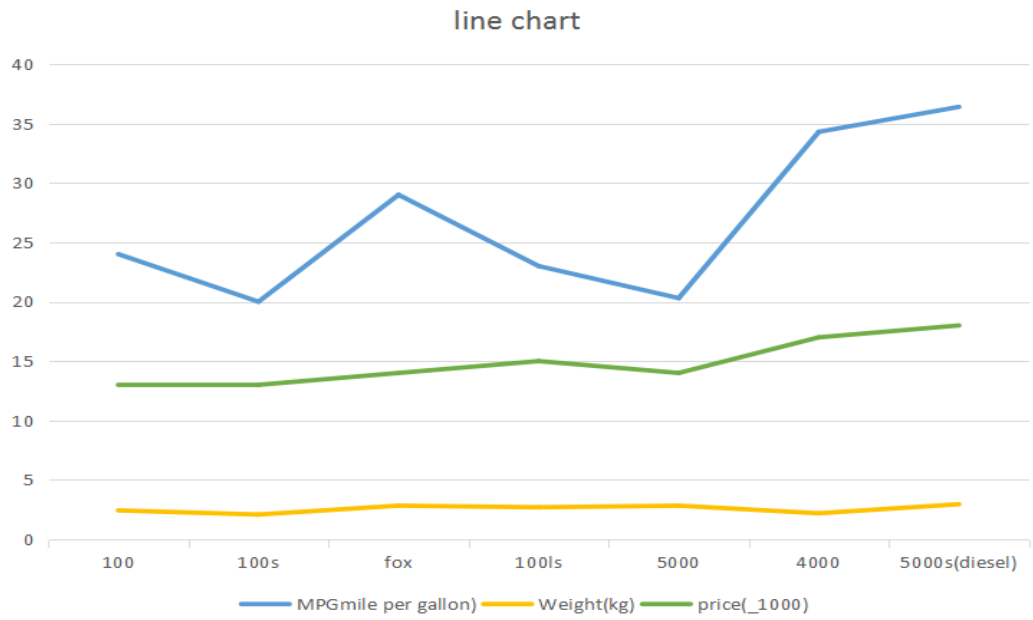
Statement2: Model 5000s is better than Model 5000.

True	False	impossible to tell

Question 3: On a scale of 1 to 5, what would you say is the ease of reading visualization 2, when the numeric values correspond to the following:

1	poor
2	fair
3	good
4	very good
5	excellent

Visualization 3:



Question 1: In lines chart, which car model is the best one?

100	1001s	100s	4000	5000	5000s(diesel)	fox	hard to tell

Question 2: In lines chart, Which car model is the worst one?

100	1001s	100s	4000	5000	5000s(diesel)	fox	hard to tell

Statement 1: Model 5000 is better than model fox.

True	False	impossible to tell

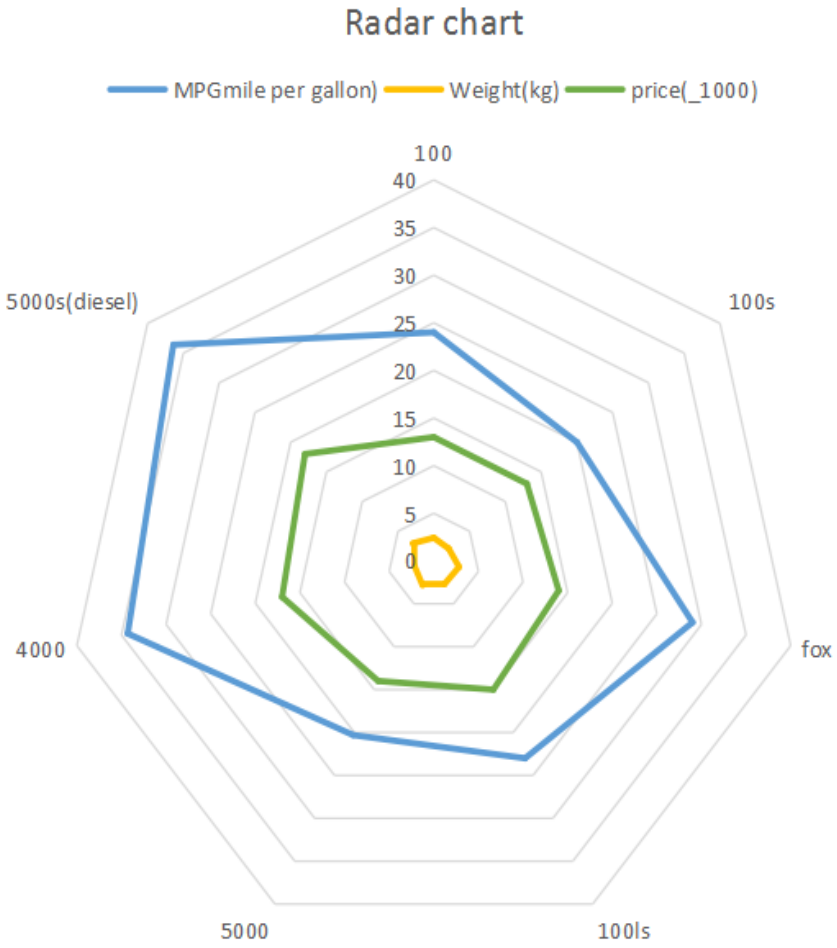
Statement2: Model 5000s is better than Model 5000.

True	False	impossible to tell

Question 3: On a scale of 1 to 5, what would you say is the ease of reading visualization 3, when the numeric values correspond to the following:

1	poor
2	fair
3	good
4	very good
5	excellent

Visualization 4:



Question 1: In Radar chart, which car model is the best one?

100	1001s	100s	4000	5000	5000s(diesel)	fox	hard to tell

Question 2: In Radar chart, Which car model is the worst one?

100	1001s	100s	4000	5000	5000s(diesel)	fox	hard to tell

Statement 1: Model 5000 is better than model fox.

True	False	impossible to tell

Statement2: Model 5000s is better than Model 5000.

True	False	impossible to tell

Question 3: On a scale of 1 to 5, what would you say is the ease of reading visualization 4, when the numeric values correspond to the following:

1	poor
2	fair
3	good
4	very good
5	excellent

Thank you for your cooperation and thank you for your response !

References

[1] Anon, (2017). [online] Available at:https://www.snapsurveys.com/blog/advantages-disadvantages-surveys/ [Accessed 29 Oct. 2017].

[2] Mack C. Shelley(2001).How to do a Survey (A 9-Step Process), II Fall 2001 LC Assessment Workshop

[3] Quizlet.com. (2017). Research: Participants, respondents, subjects - what is the difference Flashcards | Quizlet. [online] Available at: https://quizlet.com/72883412/research-participants-respondents-subjects-what-is-the-difference-flash-cards/ [Accessed 29 Oct. 2017].

[4] Staff, I. (2017). Representative Sample. [online] Investopedia. Available at: http://www.investopedia.com/terms/r/representative-sample.asp [Accessed 29 Oct. 2017].

[5] eazyBI. (2017). Data Visualization - How to Pick the Right Chart Type?. [online] Available at: https://eazybi.com/blog/data_visualization_and_chart_types/ [Accessed 6 Nov. 2017].

[6] Statistics Solutions. (2017). Paired Sample T-Test - Statistics Solutions. [online] Available at: http://www.statisticssolutions.com/manova-analysis-paired-sample-t-test/ [Accessed 29 Oct. 2017].

[7] Zhihu.com. (2017). Cite a Website - Cite This For Me. [online] Available at: https://www.zhihu.com/question/40903517?sort=created [Accessed 29 Oct. 2017].

[8] J. Harris. Data is useless without the skills to analyze it. Harvard Business Review, 2012.

[9] A. Beauchamp. What is data literacy?, January 2015

[10] M. Schield. Information literacy, statistical literacy and data literacy. IASSIST Quarterly, 28(2/3):6-11, 2004.

[11] Gelman, Andrew (2005). "Analysis of variance? Why it is more important than ever". The Annals of Statistics. 33: 1–53. doi:10.1214/009053604000001048.

[12] Howell, David C. (2002). Statistical methods for psychology (5th ed.). Pacific Grove, CA: Duxbury/Thomson Learning. ISBN 0-534-37770-X.

[13] Montgomery, Douglas C. (2001). Design and Analysis of Experiments (5th ed.). New York: Wiley. ISBN 978-0-471-31649-7.

[14] HSC PDHPE. (2017). Objective and Subjective Performance Measures. [online] Available at: https://www.pdhpe.net/factors-affecting-performance/how-does-the-acquisition-of-skill-affect-performance/assessment-of-skill-and-performance/objective-and-subjective-performance-measures/ [Accessed 8 Nov. 2017].

[15] Dr. Brian Mac Namee.(2017). Lecturer 6 Using Color. Information Visualization COMP40610. p32-p33.

[16] B.M. Namee, A. Zubiaga. Knowing what you don't know: Choosing the right chart to show data distributions to Non-expert users. Data Literacy Workshop, January, 2015.

[17] D. A. Keim. (2002). Information Visualization and Visual Data Mining. IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 7, NO. 1.

[18] Knowing Complex Ranking Easily: Choosing the Suitable Visualization to Show Multi-Attribute Rankings [Accessed 10 Nov. 2017]