Data Visualisation

Amelia Handley [40326169]

[1] Edinburgh Napier University
40326169@live.napier.ac.uk

Abstract. A dataset containing 10000 measurements was explored finding relationships between the various attributes. The dataset recorded participants who were asked to complete a set of three tasks in five different locations. These participants scores for each task were noted and were timed. The participants ages were recorded and were split into four age groups: under 16, young adult, middle-aged and older-aged.

It was found that in the dataset there were cases of outliers - data without the normal range - within two relationships: the participants age against the participant score for part A and the participants age band against the participants score for part B. In addition, there were five relationships in the attributes from the dataset. One of these relationships was further optimised to make it easier to understand.

1 Outliers in the Dataset

1.1 Participants Age vs Participant Score for Part B

Throughout the exploration of the dataset it was found that were outliers found within the relationship between the participants and the participant score for test part B (see Fig. 1). The graph chosen to display this relationship was a scatter plot. This is because each data entry is plotted on the chart as a point, making it easy to spot trends in the data and any possible outliers [3]. Outliers are out with the normal and are abnormal instances which are out with the normal range of data [2] which are a result from numerous factors including participant error responses or incorrect data entries [7]

The law of proximity, a Gestalt principle, states that the human brain makes connections based on visual perception [8]. Therefore, the graph highlights the relationship between participants that are 0-100 years old as well as being between 0-100% score for part B. Whilst the outliers follow a similar trend to the rest of the graph they are also separate from the general trend. It shows ages between 200 and 210 years old and a score of 153-163% - which is out with the possible data range of 0-100%.

1.2 Participants Age Band vs Participants Score for Part A

There was another instance of an outlier found within the relationship of the participants age band and the score for part A of the test (see Fig. 2). The chart used to show this relationship was a box plot. This is because it displays the relationship, using a “box” between the two attributes as well as highlighting instances outliers, which appear as

points above and below the “whiskers” [6]. Reviewing the graph, the attributes are split by gender into male and female. Initially, it shows that both males and females have a similar trend in data. However, examining the females scores, specifically the “Under 16” age band, shows that there are negative results for part A. Therefore, these can be considered outliers as it is not within the data range for the score for part A, for 0-100%.

2 Relationships found in Dataset

2.1 Participants Age vs Participant Score for Part B

The participants age, when compared against the participant score, showed a significant relationship. Using a scatter plot showed that there was a positive correlation between the attributes [5] (see Fig. 1). Meaning, that as the participants age increases, the participant score for part B also increases. A scatter plot was an effective visualisation as both attributes being compared were continuous [3] and as the attributes showed a strong correlation to one another [5]. However, the graph includes outliers. To alter the graph to not show the outliers would either mean removing the data which had the outliers in them or by reducing the x and y axis to not show the outliers [10].

2.2 Participants Age Band vs Participants Score for Part A

Examining the participants age band and the participants score for part A showed a positive relationship between the attributes (see Fig. 2). To visualise this relationship, a box plot was used as box plots can be used to show the average score of the participants [6]. The graph shows that Young Adults (16-34 yr.), Middle-aged adults (35-64 yr.) and Older Adults (65yrs+) all displayed higher mean scores of around 65-76%. However, those under 16 years old had a mean of around 20% of the part A score. Thus, meaning that as the age of the participant increases, as does the score for part A.

2.3 Completion Time for each Participant vs their Age Band

The completion time to complete all three of the tasks was recorded for each of the participants. This was compared against the age band of the participants. However, to see the relationship, the attributes were split by the Gender (see Fig. 3). To show this relationship a box plot was used. This is because box plots display the average of the participants [6]. For the males there is a slight increasing trend between the attributes as the completion time increases, so does the age band (or age) of the participants. Whilst for females there is a negative correlation between the completion time and the age band. Thus, meaning the older the participant, the less time it takes to complete all three of the activities.

2.4 Participants Score for Test C vs the Participants Age

Examining the dataset highlighted a significant between the participants score for test C and the age. At first, it appeared that there was no significant relationship that could be examined between the attributes. This is until the participants score was split by the

five locations (see Fig. 4). In locations B, C and E it showed that there was no change in the score as the age increased, remaining 48%, 41% and 37% respectively. However, in location A it illustrates a decrease in the test score for part c as the age increases, whilst location D showed an increasing trend in the score as the age increased. To illustrate this relationship a bar chart was used. This is because the age band is a discrete attribute whilst test score for part c is a continuous attribute whilst also following Gestalts Principle of proximity [8] - as the bars for each location are together - making it a simplistic and effective visualisation.

2.5 Completion Time of All Three Tests vs Participant Score for Part B

Completion Time to complete all three of the tests was compared against the participant score for part B (see Fig. 5). This was separated by gender. A scatter plot chart was used to visualize this relationship as both the attributes are continuous of data. Whilst the males displayed no correlation, the females displayed a negative correlation. Additionally, due to the law of Gestalt Law of Continuation, which states that the human eye follows a sequence of shapes to determine a relationship between the attributes [8]. Meaning, scatter plot is an effective visualisation to display this relationship. Therefore, it can be concluded that for the female participants, as the completion time increased, the participant score for part B decreased whilst for the males, the completion time does not affect performance.

3 Optimised Visualisation

For the optimised visualisation the relationship between the Participants Age Band against the Participants Score for Part A was selected. Initially a box plot was used to display this relationship (see Fig. 2). Whilst the box plot is an effective visualisation choice for showing the average score of the participants for each age group and to highlight outliers, showing the relationship as a bar chart is more efficient (see Fig. 6).

The gestalt principle, the law of continuation, asserts that the human eye can determine relationships by following lines or a sequence of shapes [8]. Therefore, by using a bar chart it makes the chart simpler to read as the reader can now see the trend - that as the age of the participant increases, so does the participants score. Additionally, since a bar chart cannot show a negative number, there is no need to alter the dataset to remove the outliers [6]. This also means that there is no need to split the graph by gender. It can also be said that the law of proximity shows that each of the bars are an individual part of the attribute, in this case the age band, as there is white space between the bars [8].

Additionally, retinal variables, proposed by Jacques Bertin, are an important to consider when developing a graph [9]. For instance, the size of the bars on the bar chart highlights how significant the difference is between each of the age bands. From the chart it can be seen that the under 16 category, when compared with the young adult, middle-aged and older-adult are all nearly three times the number than those under 16. Additionally, by having different colours, another retinal variable, for each of the bars it indicates that each of the bars are for different age bands.

Appendices

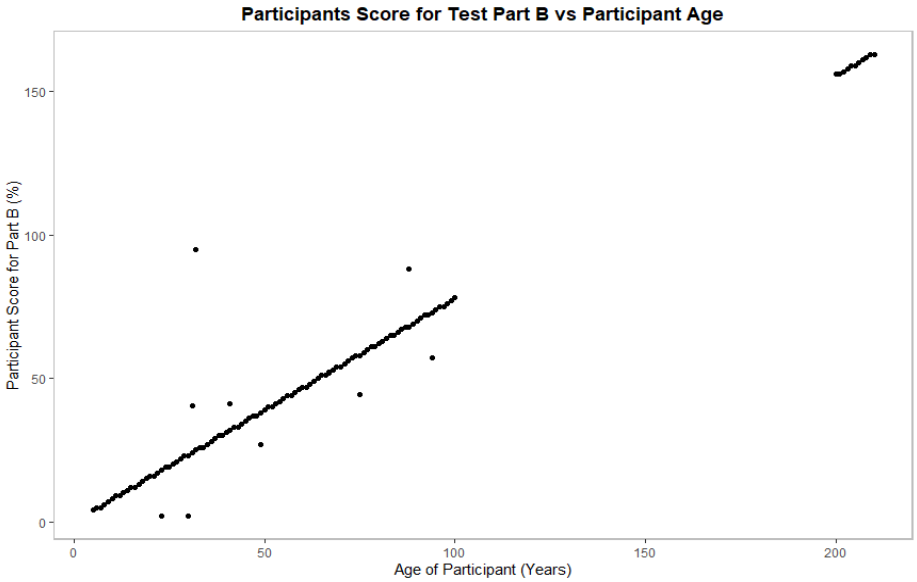

Fig. 1. Participant Score for Test B vs Age of Participant.

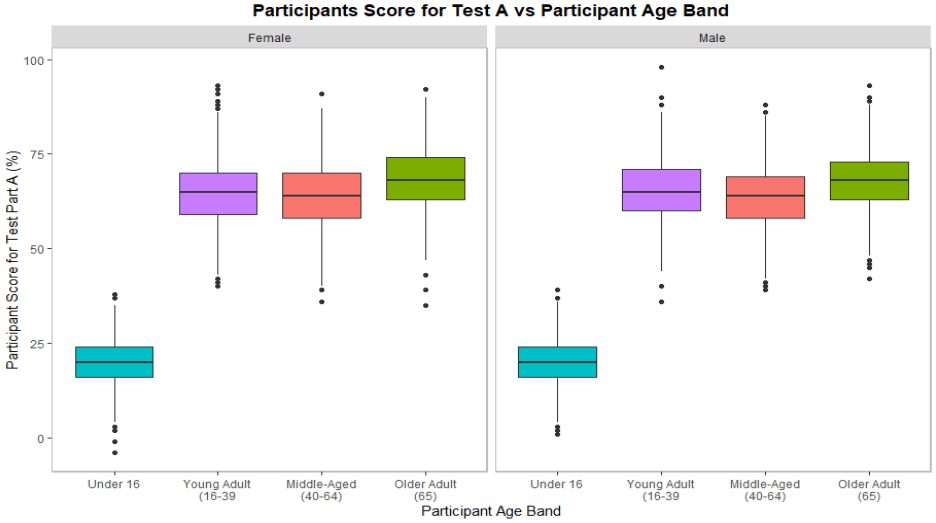

Fig. 2. Participant Score for Test A vs Participant Age Band separated by Gender.

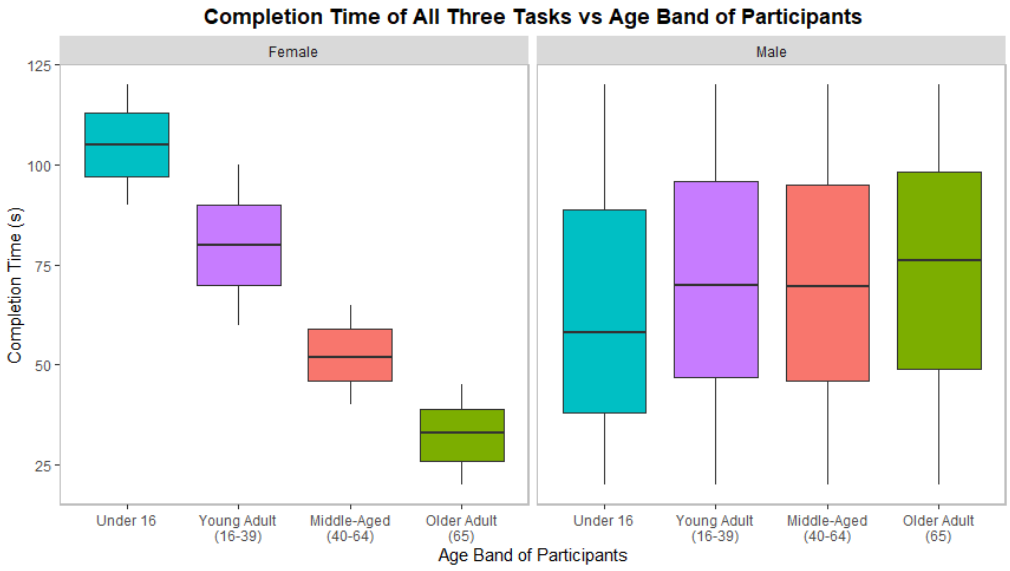

Fig. 3. Completion Time for each Participant vs their Age Band, separated by Gender.

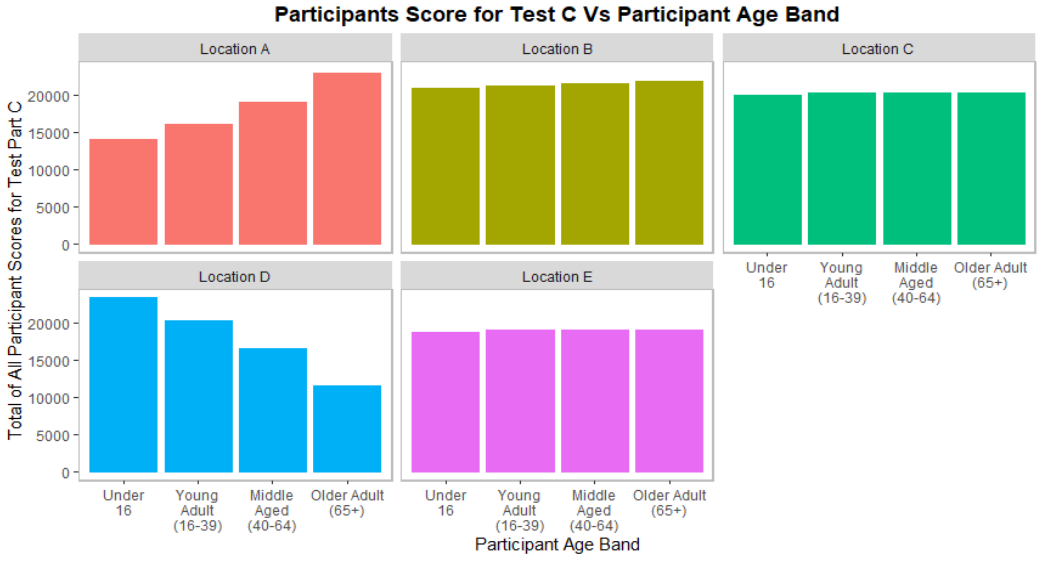

Fig. 4. Participants Score for Test C vs the Participants Age, separated by Location.

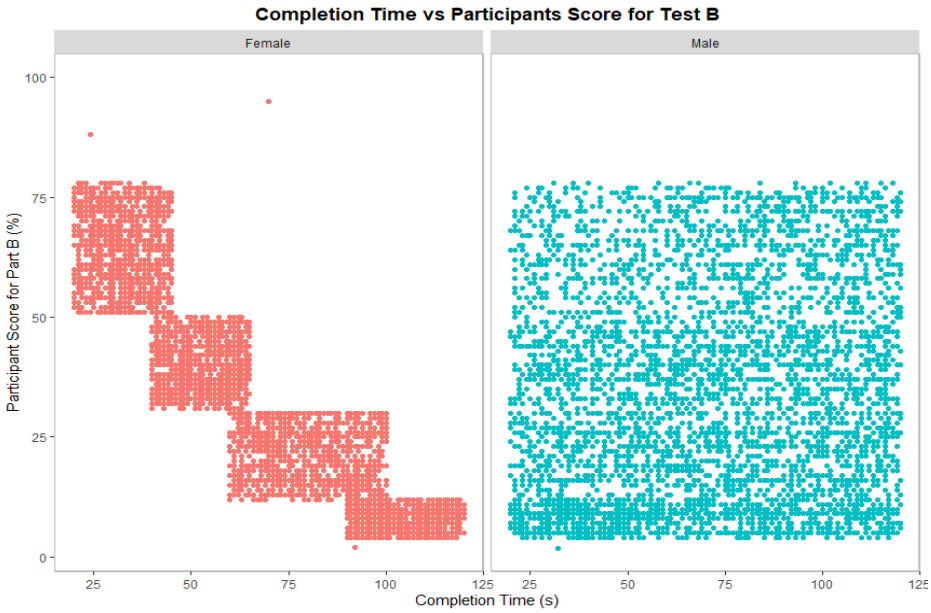

Fig. 5. Completion Time of All Three Tests vs the Participant Score for Part B (%).

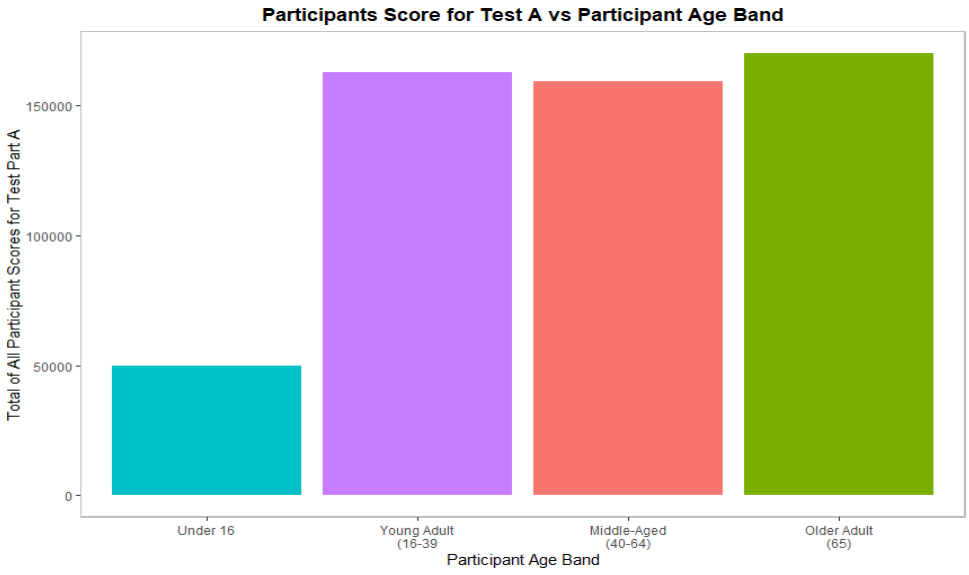

Fig. 6. Fully Visualised Bar Chart of Participant Score for Part A vs Participant Age Band (%).

References

1. Wickham, H. Getting Started with ggplot2. In: ggplot2. Use R!. Springer, Cham (2016).
2. Seeing Data, http://seeingdata.org/developing-visualisation-literacy/key-terms-in-visualisation/, last accessed 2019/10/23
3. RStudio. Data Visualisation with GGplot2, https://rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf, last access 2019/10/23
4. Engineering Statistics Handbook. What are outliers in the data?, https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm, last accessed 2019/10/23
5. Khan Academy. Scatterplots and Correlation Review, https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/introduction-to-scatterplots/a/scatterplots-and-correlation-review, last accessed 2019/10/23
6. Understanding Boxplots, https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51, last accessed 2019/11/21.
7. Kyu Kwak, S., Hae Kim, J: Statistical data preparation: management of missing values and outliers, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5548942/, last accessed, 2019/10/31
8. Soegaard, M. Laws of Proximity, Uniform Connectedness, and Continuation – Gestalt Principles, Laws of Proximity, Uniform Connectedness, and Continuation – Gestalt Principles
9. Visual Variables, https://www.axismaps.com/guide/general/visual-variables/, last accessed 2019/10/31
10. Methven, T, Data Analytics: Understanding Data, https://moodle.napier.ac.uk/mod/resource/view.php?id=1361313, last accessed 2019/10/31