# Data Analytics Coursework Report

Jonathan Binns, 40311703@live.napier.ac.uk

[1] Edinburgh Napier University, Edinburgh EH10 5DT, Scotland

**Abstract.** This report shows the five relationships and two outliers found while analyzing the supplied data. The relationships range from correlation between part B scores and age to differences in median completion time across different age bands and genders. Both outliers and relationships are shown in Figures 1 through 6 all of which show the trends in the simplest way possible with explanation as to why I think each graph is the most optimized and is the best way to explain a given trend.

## 1 Outliers

### 1.1 Ages Over 200 Years Old

The first set of outliers found are people whose age is greater than 200. These outliers are shown in Figure 6, being compared with Part B Score. The graph shows a collection of points in the top right corner far away from the trend. If you look closely at the x axis all of the ages of these points are at either 200 or above. I consider these to be outliers as the oldest woman in the world is currently 116 years old[1] so anyone who's age appears to be greater than that is probably a mistake, or should be holding the record instead. I would remove this type of outlier simply by ignoring it, this is because there is more than enough data to show a strong correlation without needing this data.

### 1.2 Part B Scores Over 100

The second set of outliers found were part B scores that were over 100%. These are also visualized on Figure 6. If you look in the top right corner again, this class of outlier is shown as all of the data points are over 150 on the Y axis. I think that these data points are outliers as each part of this test is scored as a percentage and it is impossible for a percentage score to be more than 100%. I would also remove this class of outlier by ignoring the data points as they correspond with the people whose ages are over 200, so even if this error was fixed the data would still class as an outlier and should not be analysed. I used a scatterplot to show these outliers as it spaces them out nicely from the rest of the data, making them easy to spot. My use of Gestalt's law of proximity[2] means that the viewer can see that while the outliers are one group they are not close to the group of valid data points, meaning that they won't be per-

ceived as part of the valid group and instead will be perceived as their own group of outliers.

# 2 Relationships Found

## 2.1 The Difference in Part A Score Based on Age Band

In part A age band U (people under the age of 16) scored considerably less than any other age group. I visualised this trend in Figure 1 using a box and whisker plot. This graph shows the box for age band U being significantly lower than any of the other age bands. The graph is ordered by the median value, the line in the middle of each box, to further show the difference between age groups and making the trend easier to understand. This visualization uses height to show a difference in proximity, meaning that because of Gestalt's laws[2] a viewer will perceive the age band U to be different from other age bands. This type of graph also uses size as a retinal variable to show the range of data for each age band. As humans can perceive differences in size[3] we can make more detailed comparisons between the different age bands. Finally, one possible chart could be median part A score against age band I feel like the loss of information of doing that was not worth the ease of readability of something like a bar chart.

## 2.2 As Age Increases So Does Part B Score

The second relationship found is that as peoples ages increase so does their part B score. This can be seen in Figure 2 as nearly every black dot is in a neat diagonal line from the bottom left to the top right. This trend is visualised as a scatterplot as it was the simplest method to show both the trend and the few anomalies that exist in the data. The trend is so strongly correlated, adding a line of best fit would have obscured the individual data points. Also adding any other additional information such as gender or location with a second retinal variable, for example colour or shape would have had the same effect.

## 2.3 Differences in Median Completion Time for Each Age Band in Each Gender

For women, the lower the age band that they are in the longer the median completion time of all parts of the test. For men, as their age band decreases there is a slight decrease in median completion time. I visualized this trend by using two bar charts placed next to each other, one per gender. A bar chart was used to show this data as length has a natural order[4] which will help people to understand the trends as effectively as possible. I first found this trend as a box and whisker plot however, focusing on median completion time was the best thing to do as the trend was more subtle than in Figure 1 and the additional information added by a box and whisker plot would have made the graph more cluttered without adding relevant information.

### 2.4 Locations A and D Location Part C Score Changes with Age

In location A, as age increases so does part C score, whereas in location D as age increases part C score decreases. No such trends exist in any other locations with the two factors having no correlation. These relationships can be seen in Figure 4. These trends are shown as scatterplots for each location being placed next to each other in the same graph. In order to show two differing trends, I used juxtaposition. By placing smaller graphs next to each other in order to help users see the different patterns between the objects[5]. Alongside this, each location is given its own colour, this helps to differentiate all the locations and allow people to see the trends more clearly. Again no line of best fit was added as the data is so strongly correlated it would not add any useful information and in the case of locations B, C and E it would have slightly misrepresented the data by making it seem as if part C scores stay the same no matter age rather than age having no impact on score.

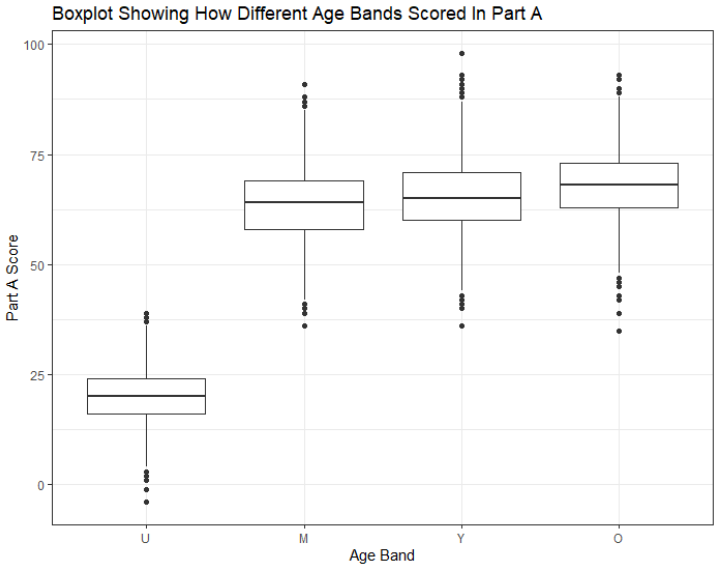
### 2.5 Median Part B Score Decreases as Completion Time Grows for Women

For women, completion time gets longer as the median part B score decreases, however for men there is no correlation. This is shown in Figure 5 and is visualised as a line graph because a scatterplot was very crowded. Displaying the data as a line graph of the median values for each age allowed me to remove a large part of the data while still being able to show the overall trend. The line graph still isn't a perfect method of displaying the data as for men as it shows a slight rise in median part B score where in fact the data resembled something close to locations B, C and E from Figure 4 showing no real difference or correlation between the two variables. I decided to show the trend over two graphs as when the data is plotted as a scatterplot on one graph no amount of retinal variables would make the graph free from clutter.

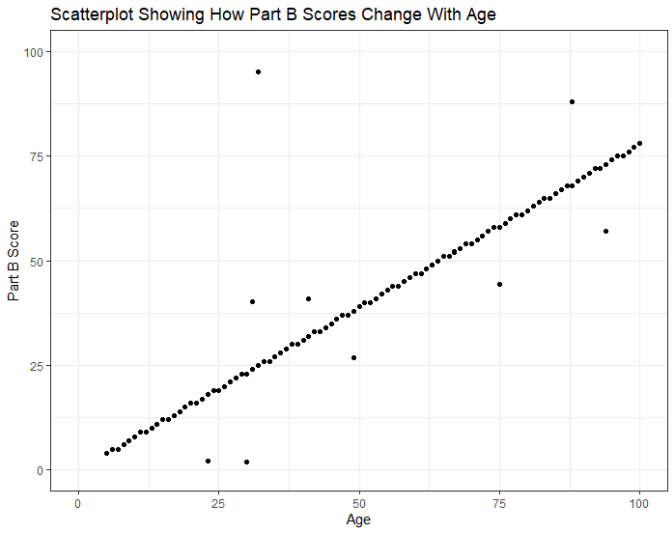
## 3 Most interesting relationship

The relationship that is the most interesting is how part B score increases alongside age. This is shown in Figure 2 as a scatterplot. This graph is the most optimised it can be. It's a very simple relationship and has very close correlation. Even when you include outliers on both axis, which is shown in Figure 6, the relationship is still shown perfectly. Adding any additional retinal variables such as colour or shape would have obscured the relationship and made the trend harder to see, and especially in this case adding additional retinal variables would not have made recognizing the trend and easier or faster. Supporting the research conducted by William A. Kealy and Chitra Subramanian[6] which challenged the widely held notion that retinal variables help people to understand visual information. I think my optimization in Figure 2 works as it shows both the data and the trend in the simplest possible way, allowing viewers to see the trend and certain exceptions without needing any excess visual information. As the graph is displayed in black and white it allows colour blind people to understand the graph to its fullest without any further help or modification to the graph.

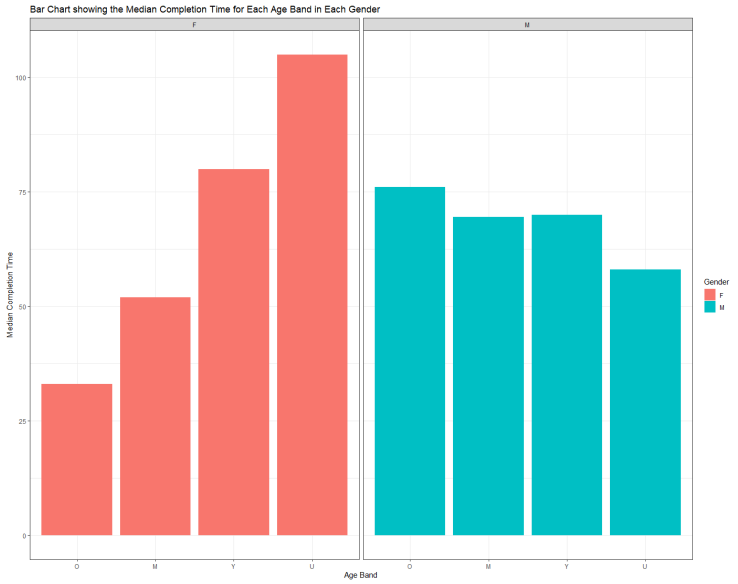
# 4 Figures



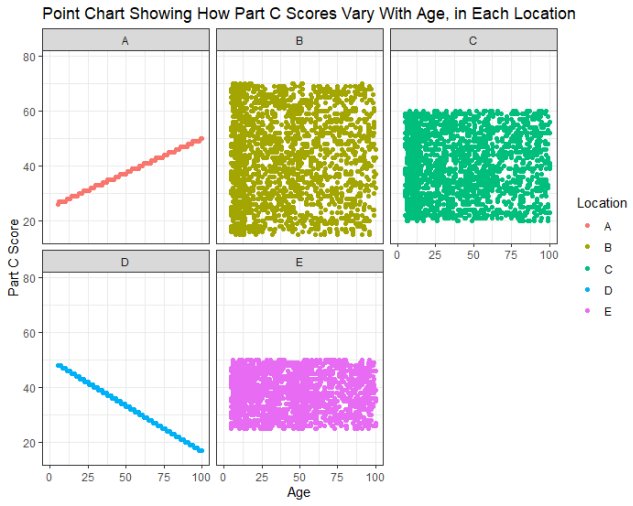

*Boxplot Showing How Different Age Bands Scored in Part A (Figure 1)*



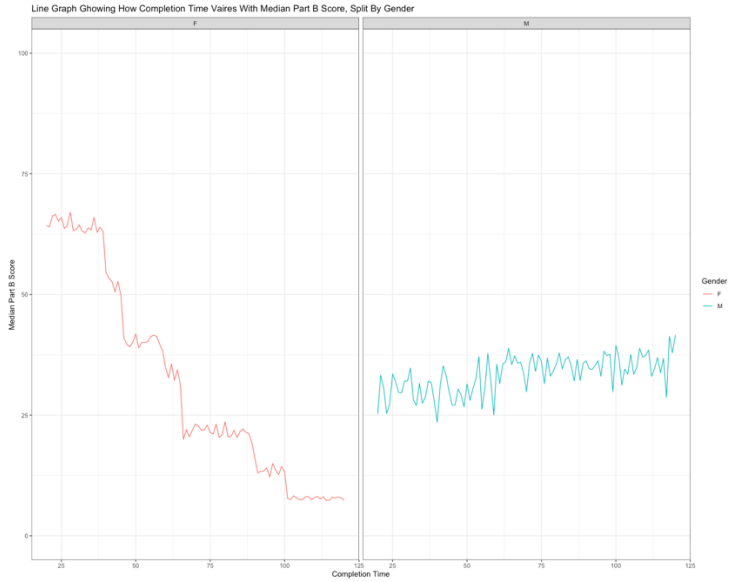

*Scatterplot Showing Age against Part B Score (Figure 2)*



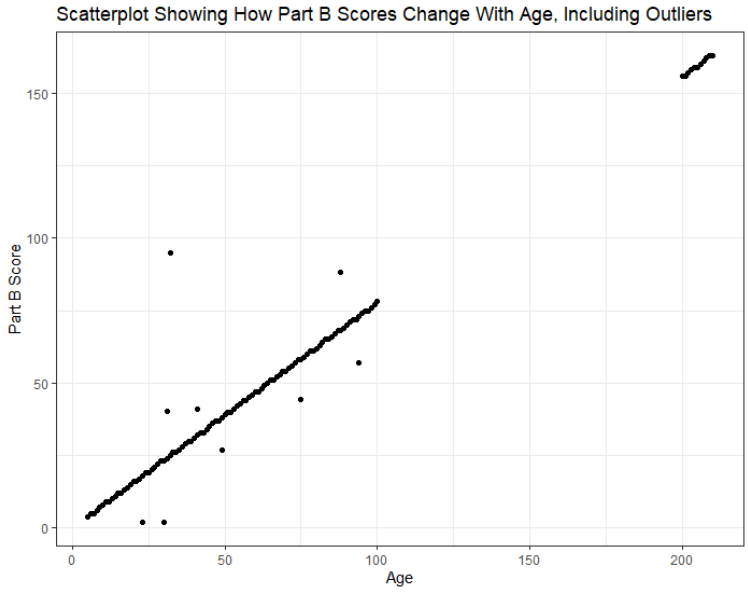

*Bar Chart Showing how Median Completion Time Changes with Age Band, Split by Gender (Figure 3)*




*Scatterplot Showing How Part C Score Varies with Age, Split by Location (Figure 4)*




*Line Chart Showing How Completion Time Varies with Median Part B Score, Split by Gender (Figure 5)*




*Scatterplot Showing How Part B Score Varies with Age Band, Including Outliers (Figure 6)*

# 5 References


1. Guinness World Records. (2019). Oldest Living Person (Female). Accessed on 23/10/19.
   https://www.guinnessworldrecords.com/world-records/67477-oldest-person-living-female

2. Gestalt Laws: Laws of Proximity and Similarity. John H. Krants & Bennett L. Shwartz. (2015). Accessed on 27/10/19.
   https://isle.hanover.edu/Ch05Object/Ch05ProxSim_evt.html

3. Data Analytics: Psychology of Vision. Dr Tomas Methven. (2019). Accessed on 27/10/19

4. Chapter 4. Choosing Appropriate Visual Encodings. Designing Data Visualisations. Julie Steele & Noah Iliinksy

5. Visual Comparison for Information Visualization. Michael Gleicher & Danielle Albers, et al. (2011). Accessed on 27/10/19.
   https://pdfs.semanticscholar.org/a5e3/cf9f7bfdbc227673a6d8f4d59112f1b5bb3a.pdf

6. The Questionable Effect of Retinal Variables on Information Displays: Implications for Problem Solving and Learning. William A. Kealy & Chitra Subramanian. (2005). Accessed on 30/10/19.
   https://libres.uncg.edu/ir/uncg/f/W_Kealy_Questionable_2005.pdf