An Emprirical Study of Data Visualisation

An investigation into the theory behind data visualisation

Katie Murphy - 1632254

MMORS Final Year Dissertation

Cardiff School of Mathematics

Abtract

A study into the theory behind data visualisation, looking into different ways in which data can be interpreted visually to portray key information accurately in an elegant and concise manner using the libraries 'ggplot2' from R and 'matplotlib' from Python. The investigation will focus on aspects of visualisations that may either deliberately or accidentally mislead the observer, such as inapropriate axis scalings and labeling, alteration of aspect ratios, and the use of colour. A survey is administered to gather quantitative information and interviews with experienced programmers are held to gain an understanding on the opinions of the codes and visualisation packages themselves.

Acknowlegements

I would like to thank my supervisor, Dr. Vince Knight, for his invaluable support and encouragement throughout the last year. It has been a difficult year for many reasons, and Vince has always been very understanding and supportive, and has helped me greatly in managing to get this dissertation to this point. I could not have managed without his continued support and guidance.

I also deeply appreciate his patience in teaching me and assisting in my own learning of Git and VS Code, and for always being willing to answer my many questions.

I would additionally like to thank Vince's children and pets for their lively contributions to meetings.

Finally, I would like to thank everyone who participated in the study, in both the survey and the interviews. Thank you to the interviewees for their time and very informative contributions to the study; Vince Knight, Geraint Palmer, Nikoleta E. Glynatsi, Andreas Artemiou, Henry Wilde and Owen Jones.

This report was written and compiled using R Markdown (Allaire et al. 2020) with pdflatex and version control using Git and GitHub (Chacon and Straub 2014).

Contents

1 Literature Review **5**

1.1 Introduction . . . 5
1.2 Good Visualisation Practice . . . 6
1.2.1 Axis Scaling and Aspect Ratios . . . 6
1.2.2 Colour . . . 7
1.2.3 Bar Charts . . . 9
1.2.4 Visualisation Taxonomy . . . 9
1.3 Studies in Visualisation . . . 10

2 Data collection **13**

2.1 Background on survey design . . . 13
2.2 Specific goals of survey tool for this study . . . 16
2.3 Survey Design . . . 17
2.4 Creating the Visualisations . . . 19
2.4.1 The Data . . . 19
2.4.2 The Bar Plots . . . 20
2.4.3 The Line Plots . . . 21
2.5 The Survey . . . 21
2.5.1 Demographic Questions . . . 21
2.5.2 American Ninja Warrior - Part 1 . . . 22
2.5.3 American Ninja Warrior - Part 2 . . . 23
2.5.4 American Ninja Warrior - Part 3 . . . 24
2.5.5 Sales - Part 1 . . . 26
2.5.6 Sales - Part 2 . . . 26

3 Univariate Analysis **27**

3.1 American Ninja Warrior - Part 1 . . . 27
3.1.1 Effect of Y-Axis Truncation . . . 28
3.1.2 Effect of Logarithmic Scaling . . . 33
3.1.3 Differences Between Question 2 and 3 Responses . . . 38

3.2 American Ninja Warrior - Part 2 . . . 38
3.2.1 How many times would you say 'Floating Steps' were used? . . . 42
3.2.2 Comparison questions on aesthetics and ease of interpretation . . . 45
3.3 American Ninja Warrior - Part 3 . . . 45
3.4 Sales - Part 1 . . . 52
3.4.1 How much would you say sales of each company increased between January and December? [Company A] . . . 53
3.4.2 How much would you say sales of each company increased between January and December? [Company B] . . . 53
3.4.3 How large would you say the drop in sales between April and July of Company A is? 54
3.5 Sales - Part 2 . . . 54
3.5.1 Based on the above graph, how large would you say the difference is between the number of sales Company C makes and the number of sales Company D makes? . 54
3.6 Conclusion . . . 54

4 Appendices **57**
4.1 Appendix 1 - The Survey . . . 57
4.1.1 R . . . 57
4.1.2 Python . . . 95
4.2 Apendix 2 - R Code for Visualisations . . . 124
4.3 Apendix 3 - Python Code for Visualisations . . . 138
4.3.1 . . . 153
4.4 Appendix 4 - Univariate Analysis . . . 153
4.5 Ninja Warrior - Part 1 . . . 154
4.5.1 "Approximately many times would you say the 'Salmon Ladder' was used?" . . . 154
4.5.2 Approximately how much more than 'Log Grip' would you say 'Salmon Ladder' was was used? . . . 160
4.5.3 'Approximately how much more than 'Quintuple Steps' would you say 'Salmon Ladder' was used?' . . . 164
4.6 Ninja Warrior - Part 2 . . . 168
4.6.1 "How large would you say the difference between 'Jumping spider' and 'Salmon Ladder' is?" . . . 168
4.6.2 How large would you say the difference between 'Log Grip' and 'Floating Steps' is? 172
4.6.3 How many times would you say 'Floating Steps' were used? . . . 176
4.7 Ninja Warrior - Part 3 . . . 181
4.7.1 How many times would you say 'Floating Steps' were used in the Finals (Regional/City) round? . . . 181
4.7.2 How many times would you say 'Log Grip' was used in the Finals (Regional/City) round . . . 184
4.7.3 Please select the statement you feel applies to the bar chart above. . . . 187

4.7.4 Which obstacle do you think was used MORE in Finals (Regional/City) rounds, 'Log Grip' or 'Floating Steps'? . . . 188
4.7.5 Which bar chart do you feel is easiest to read and interpret? . . . 189
4.7.6 Which colour scheme do you find most aesthetically pleasing? . . . 191
4.7.7 Do you feel that one of the colour schemes makes it easier to read and interpret? If so, please select which one. . . . 192
4.8 Sales - Part 1 . . . 193
4.8.1 How much would you say sales of each company increased between January and December? [Company B] . . . 198
4.8.2 How large would you say the drop in sales between April and July of Company A is? 203
4.9 Sales - Part 2 . . . 207
4.9.1 Based on the above graph, how large would you say the difference is between the number of sales Company C makes and the number of sales Company D makes? . . . 207
4.9.2 . . . 210
4.10 Appendix 5 - Interview Transcripts . . . 210

5 Conclusion **243**

6 Future Work **245**

7 References **247**

Chapter 1

Literature Review

1.1 Introduction

Data visualisation is a method of conveying data in an easily digestible manner through graphics. It is an important aspect of data presentation and allows key information to be quickly identified by the observer. Very many subject areas rely on such visualisations to relay messages that may get lost or have less impact when presented as written word or raw numbers.

The main objective of visualisation is to create figures that display the data accurately in an aesthetic manner, giving non-misleading messages in a format that is pleasing to the eye. A good visualisation strikes a balance between aesthetics and information, where the aesthetic features are designed such that they 'enhance the message of the visualisation' (Wilke 2019).

An incorrect balance of aesthetics to information can lead to figures that are misleading, confusing, or unengaging. Wilke discusses the way in which, for example, a research scientist with limited design experience may produce a visualisation displaying the data in an informative manner, but fail to draw immediate attention to the desired message, and on the other hand, someone with a main interest in the aesthetic design of a visualisation could create a figure that is very pleasing to the eye, but create a misleading visualisation in the process.

This literature review will discuss a range of publications discussing various aspects of data visualisation with a focus on how poor or uninformed visualisation design can produce misleading figures, as well as how such visualisations may be abused to deliberately deceive the observer. Starting with publications discussing general good visualisation practice, the discussion will then lead on to look at studies investigating the implementation of different visualisation practices, from which inspiration will be drawn to design the study for this paper.

1.2 Good Visualisation Practice

The book *'Fundamentals of Data Visualization'* (Wilke 2019) is renowned as *'an excellent reference about producing and understanding static figures, figures'* (see Bebeau 2019) and described as being *'suitable to be used as a reference manual'* (see Hwang 2020). Thus, this book provides a good basis to understanding the principles behind data visualisation, and how to create effective, informative and aesthetic figures.

In the book, Wilke discusses a variety of topics under the data visualisation umbrella, from relatively simple but important and often overlooked ideas such as deciding on coordinate systems, axis scales and colouring, to how to visualise distributions, trends and geospatial data. This literature review will focus on the areas being investigated in the 'Empirical Study of Data Visualisation' survey; namely coordinate systems, axis scaling, colouring for bar charts, alongside stacked and grouped bar charts, as well as axis scaling and formats for time series plots.

1.2.1 Axis Scaling and Aspect Ratios

In discussing coordinate systems and axis scaling, Wilke highlights that, prior to deciding on a coordinate system, it is important to consider the form the data will take, and where the data will be positioned, as well as how many dimensions this data takes. The example used is a classic two-dimensional scatter plot, in which each data value is represented by a point positioned in a distinct location on the 2d plane, and thus two scales are required to define where this location falls, traditionally with a linear scale and horizontal x-axis with the y-axis perpendicular to this.

Alternative coordinate systems can include the perpendicular model with non-linear axes, or circular or 'curved' models such as polar coordinates in addition to flipped axes, where the dependent variable is represented by the x-axis as opposed to the y.

'An Empirical Study of Data Visualisation' will be mainly analysing perception of categorical bar charts, for which the 'locations' are the category as defined by the x-axis and the bar height. It will be discussed how the perception of these locations could be altered to stray from 'good practice', and how these alterations may mislead the observer.

In discussion of the linear, two-dimensional cartesian system, the author describes the various formulations that this system can take, in terms of variables with the same or different units. For example, if two variables with different units are represented perpendicular to one another on a cartesian system, the designer has the freedom to stretch or compress the data in a way to best represent the data and, as Wilke states, 'maintain a valid visualisation of the data'.

Another point of interest mentioned by Wilke here is that the ratio of x to y-axis should be such that 'important differences in position are noticeable'. This is regarded as good practice by Wilke, but could potentially be exploited as discussed by Few (2016); the aspect ratio can be manipulated to make differences appear larger depending on the story that the creator wishes to sell. On the other hand, Wilke does state that it is 'important differences' that should be noticeable, and so may relate to differences that are already significant and crucial to see, and which may be minimised by an inappropriate aspect ratio.

For example, consider a company facing a drop in profits from one time step to another. An aspect ratio minimising the height in comparison to the width can allow this difference to appear smaller. On the other side of this coin, a company may have marginal profit gain between two time steps, and can abuse principles of perception to lengthen the y-axis as compared to the x, potentially making the difference seem larger.

This will be considered when writing the survey, as the perceived differences in position will be tested when changing features such as y-axis scaling or aspect ratio. A standard practice as laid out by Wilke is that, for two variables with the same unit, the aspect ratio should ensure that the space between ticks for each variable are equal in size, ie. such that the grid lines (real or imagined), form regular squares. This is to ensure that the sizes of tick spacing represent the same values in the same way, as it could be misleading to show two equal numerical differences with different visual spacing. This is regarded as less important, however, for variables in differing units, as the tick spacings represent different values for each variable. Thus, one has less freedom with re-scaling plots while still ensuring an accurate representation of the data when working with variables in the same units.

The plots in this study will show categorical data, and thus have character variables on the x-axis, with numerical values on the y-axis, and so the effect of altering the aspect ratio on two same-unit variables will not be investigated, but this could be an interesting topic for future investigation.

After this, Wilke goes on to discuss logarithmic scaling, which will be investigated in this study alongside axis truncation. Conversely to what will be investigated in this study, he talks about both logarithmic scaling and log-transformed data whereas this study will consider logarithmic scaling alone. He describes how this is a preferable format when dealing with ratios, and explains that this is a result of the fact that the product of linear numbers is analogous to the sum of the logarithms when using a logarithmic scale. Additionally, data containing a large variation in magnitudes is also stated to benefit from logarithmic scaling.

The book also explains the differences between plotting the original data on a logarithmic scale, and log-transformed data on a linear scale. In terms of mathematics, these are analogous, but Wilke states that plotting the original data on a log-transformed scale is favoured as this shows the true data values as opposed and allows easier reader interpretation of values. This will be considered and the original data on a logarithmic scale will be investigated. A study with narrower scope on only logarithmic scaling and log-transformed data, or even just a study of axis scaling, may allow to test both log-transformations and log-scalings, however the wide scope of this study means that study topics have to be reduced.

1.2.2 Colour

A very important aspect of visualisation is the use of colour. Colour is a very useful means of showing features of the data such as groupings and gradients of values, as well as to highlight key values. It could be said that colour works as a third dimension to the visualisation, showing another dimension of information on top of that shown by the position or size of the data points. It is important that the use of colour is carefully thought out, and not just applied with aesthetics in mind. Aesthetics are, or course, an important factor in colouring a visualisation, as 'pretty' colouring allows the visualisations to be eye-catching and memorable, which can be beneficial to, for example, brands or pharmaceutical companies giving regular data presentations as very aesthetic plots are more likely to be remembered.

Once again, Wilke has a good explanation and examples of each type of colour usage. There can be defined to be two types of colour scale, or palette; qualitative or quantitative. Qualitative colour scales do not have a logical order of colouring and are used for data for which the ordering of either values of groupings is inconsequential, such as for much categorical data. The latter, quantitative colour scales, provide gradients of colour and can consist of either a gradual scale moving through two or more colours, or a single colour with varying saturation. This type of scale provides a continuous colouring and can show grouping tendencies in continuous data, and is also often used in maps. As mentioned by Wilke, continuous graduated colour scales have the ability to show the degree to which two values are similar. There are several examples of both types of colour scale, and those discussed in this study will involve the defaults for both R and Python, and two colourblind friendly palettes; viridis and greyscale.

It is important to consider that the visualisations may be viewed by people with colourblindness in order to ensure they are accessible to any viewer. Using certain colour scales may be problematic for someone who is colourblind as they may find it difficult to distinguish between certain colours, and thus lose the impact and story told by the third dimension of the visualisation. Shaffer (2016) has a good explanation of how visualisation can be modified to accommodate those with the condition. Using colours such as red and green, or a traffic light scaling is not ideal, as these can be harder to distinguish. However, red and green is a very common and useful combination to use, as it can give a highly intuitive story of 'good' and 'bad' values, or positive and negative. Shaffer describes that this can be worked around by adding arrows, icons or annotations to distinguish values. Another workaround is to, for example, use very light green and very dark red as a saturation as opposed to hue comparison. The viridis palette has been specifically formulated to allow easier perception for people with colourblindness.

Colour can very easily be misused, however, and a common misuse, described by Wilke, is to colour each individual bar in a simple bar chart. This colouring reveals no additional information about the data and labeling is much preferred as the added colour here draws attention away from the message of the data and reduces the efficiency of information transmission. However, colour is useful for stacked or grouped bar charts to distinguish between groupings, with each bar in a group relating to a different category, with the labels representing the overall group.

Additionally, a good way of using colour is to highlight values of interest. For example, one technique is to use a greyscale palette for the majority of the figure, with a small selection of bars or points highlighted for fast communication that these values are most important to the message. This could easily be abused, however, to highlight the required message while potentially hiding or lowering the significance of values that could contradict the message being portrayed.

In regard to legends, Wilke refers to *'redundant coding'* of legends, which is the principal of using colour as an aesthetic tool to 'enhance' the message of the visualisation as opposed to using this as a primary tool for relaying information. Firstly, as mentioned prior, Wilke discusses how using colour as an identifier can be problematic for people with colourblindness if the colours are poorly selected, and shows how a given colour scheme would look for people with varying forms of colourblindness. The example plots the well-known Iris data set on a scatter plot, separating the species by colour. The colours are poorly chosen, with the colours for two overlapping species becoming almost indistinguishable. Solutions to this are laid out to be changes of colour, or changes of point shape, where the change of point shape provides a fourth perceptual dimension. For line plots this can be seen as dashed or dotted lines.

The author then explains the principle of 'direct labeling', that is, plotting without a legend and rather labeling the objects in the plot itself. This can reduce the amount of information the observer must take in and potentially improve ease of interpretation. Once again, this would be a topic to be investigated further in another study.

1.2.3 Bar Charts

When discussing good practice for bar charts, Wilke discusses many aspects of visualisation including axis alignment and bar ordering, as well as discussion on stacked and grouped bars. The axis alignment, ie vertical vs horizontal bars, is dependent on the data being visualised. Wilke uses the example of bars with labels that may become either difficult to read or unaesthetic when shown on the vertical chart, but appear clearer on the horizontal.

In terms of bar ordering, Wilke discusses that it is important aesthetically to order bars from largest to smallest, given there is no pre-specified ordering in the data, such as age ranges. Bar plots will be discussed in more detail in the past study reviews.

1.2.4 Visualisation Taxonomy

The paper Shneiderman (1996) provides a 'task by data type' basis for creating visualisations, summarising this with the 'Visual Information Seeking Mantra'; 'Overview first, zoom and filter, then details-on-demand'. This mantra provides a starting point when thinking about creating a visualisation, and relates to the different messages encoded in a visualisation; the viewer must first be able to gain a good overview of the whole data when taking a glance at the plot, but then discern more detail by paying closer attention, as per the 'zoom and filter' principle. The third principle in the mantra is useful to consider when creating interactive visuals; the user is able to, for example, obtain further tables of values and information-based visuals which, and as described in Taylor (2014), are 'less visual, and more text-heavy'. The mantra allows the designer to focus on not making visuals too busy whilst also encoding the necessary information.

Based on this mantra, Shneiderman suggests a 'task by data type taxonomy', which involves cross referencing 7 data types with 7 tasks, for which he doesn't provide a diagram but the description envisions what seems to be a 7-by-7 table of tasks against data types. The idea is to discuss these alongside each other to draw meaningful conclusions as to how best to produce the visual representation.

1.3 Studies in Visualisation

There is a large amount of research and literature surrounding the topic of misleading visualisations, looking into how various techniques can either deliberately or unintentionally deceive an observer in the message of the data. Results from some of these papers will be replicated, as well as used to form hypotheses which this survey will investigate. A large amount of the literature exploring misleading tactics in data visualisation focuses mainly on bar plots and line plots for categorical and time series data, and so this is what the study and literature review will focus on.

The 2020 paper "The Deceptive Potential of Common Design Tactics Used in Data Visualizations" (Lauer and O'Brien 2020), as the title suggests, explores how using different design tactics may mislead the person seeing the visualisation. Similarly to "An Empirical Study of Data Visualisation", the Claire and O'Brian paper uses a survey to explore how deceptive visualisation techniques can be employed as well as their impact on perception of the data. The survey discussed in this paper presents the participant with four plots; a bar plot, a line plot, a pie chart and a bubble plot. Additionally to changing aesthetic features of the plots themselves, the study investigates the use of exaggerated, leading titles, for example one control plot has the title " Home Sales Show Increase From 2015 - 2016", which is altered to"Huge Increase in Home Sales From 2015 – 2016¡'. The control plots consist of using a y-axis scaling beginning at 0 for the bar and line plots, a standard pie chart, and a bubble plot with proportionally sized bubbles, all alongside the non-exaggerated titles. The altered plots involve truncating the y-scale for the bar and line plots, making the pie chart in 3D, and arbitrarily altering the sizes of the bubbles on the bubble plot. The altered plots are referred to as the"deceptive" plots. The survey used sets of plots as crossed between deceptive aesthetics and deceptive titles; two had control aesthetics, one with the control title and one for the exaggerated title, and two had deceptive aesthetics with one having the exaggerated titling.

With regard to truncated axes, Claire and O'Brian asked participants to subjectively judge the difference between two data points using a 6 point scale ranging from "a little" to "a lot". For both the bar plot and line plot it was found the the use of a truncated scale increases the perceived difference between the data points. The use of a truncated scale is also discussed by Yang et al. (2021), whereby 5 empirical studies were performed in order to assess the effect of altering the scale in this way. The first of the 5 studies once again assessed how large the difference between data points is perceived to be in the truncated plot as compared to a control, again using a subjective scale from "Not at all different" to "Extremely different" on a 7 point scale. This scale differed, however, in the way that a midpoint label of "Moderately different" was provided. The 7 point scale may be preferable to the 6 point scale as the 7 point has a defined midpoint at 4, whereas the 6 point does not. This study once again concludes that the differences in data points tended to be perceived as larger than for the control plot. Alongside these studies, a 2014 blog post (Parikh 2014) discusses axis truncation and its effect on perceived data point difference for bar plots alongside other aesthetic features. The first example shows how truncating the y-axis of a bar plot can over-exaggerate differences in the heights of the bars, perhaps leading to incorrect observations regarding comparisons of values within the data.

The paper Hlawatsch et al. (2013) performs a similar study, but instead investigates the use of ‘stack-scale’, or ‘stacked’ bar charts and logarithmic scaling. The aim of the study was to explore whether stack-scale bar charts are an effective way to visualise large value data, which is less relevant to since the Ninja Warrior and Sales data are relatively low-valued data compared to the paper, but nevertheless provides a framework for exploring the use of logarithmic scaling and stacked bars in a respondent study. Participants were shown three plots; a control with a linear scale, a bar plot using a stack-scale, and one with logarithmic scaling. The questions asked determined how the different scaling affected accuracy in reading individual values, interpreting differences in values and determining which time-step exhibits the largest difference in values. Motulsky (2009) additionally discusses the use of a logarithmic axis in bar plots, explaining how it is impossible for a zero value to be displayed on this axis, and thus the bar start points are arbitrary and produce an inaccurate representation of the bar height with relation to the true value. To quote the paper, *"Don’t create bar graphs using a logarithmic axis if your goal is to honestly show the data"*. It can be observed that the logarithmic scale makes the perceived difference appear smaller than in the control.

As well as scaling, another aspect of visualisation design that could potentially mislead the observer is bar width and aspect ratios. When adding a visualisation into a publication, re-sizing the visualisation to fit a specific gap may include altering the aspect ratio, in turn affecting the length to width ratio of the bars in a bar plot. As explored by Steven Few in a 2016 article for the *’Visual Business Intelligence Newsletter’* (Few 2016), altering this ratio can affect viewer perception in the way of a narrower and taller image distorting bars to appear longer, and vice versa, meaning that perceived differences between bar heights may be affected.

Part 2 of the survey will be based around investigating this idea, alongside how the reading of exact values is affected. The second section of the survey tests whether altering aspect ratio of plots affects interpretation. The purpose of this is to mirror what my occur when visualisations are published, and may be resized to fit the section of the page they sit on. As in (Few 2016), it will be hypothesised that an aspect ratio that effectively narrows the bars may cause overestimation in values, and vice versa, using a ratio that widens bars could lead to underestimation. In the paper, the author discusses how increasing the widths of bars could distract from the bar height as well as take up excessive space on a page. It is also mentioned that wider bars may be “aesthetically displeasing”. This section of the survey will test both how bar width alters perceived difference between bars as well as opinions on the aesthetics. The method in the paper also involves altering spaces between bars, including bar plots with spaces at 50% of the bar widths and then reducing the width of the space by a third. Conversely to this, width of spaces between bars will not be considered, only the effective widths of the bars themselves. The author concludes that a length-to-width ratio of 10:1 appears to suffer from perceptual imbalance, but increasing this such that the bars become narrower and longer does not appear to have as much of an impact; the ratio can be increased relatively far with out causing much perceptual imbalance.

An article from the University of Stuttgart (Huynh 2017) gives an overview of many types of bar chart, including stacked and grouped bars. The author remarks that grouped bar charts may make the comparison of bars in the same category more difficult, while the stacked bar chart sacrifices ease of comparison of values in the bars for increased spacial efficiency. A 2018 work from the journal of *'Visual Informatics'* (Indratmo et al. 2018) also provides a discussion on the use of various forms of stacked and grouped bar charts and their efficacy. The paper notes how a classical stacked bar chart can be useful for overall comparisons as the height of the bar represents the value of the item, with the different attributes depicted as a segmentation of this single bar into different colours. When discussing grouped bar charts it is mentioned that stacked bar charts may be less useful when performing attribute comparisons, in other words comparisons between different categories on the same bar, as a result of the bar segments being non-aligned. This results in comparison taking the form of length judgment as opposed to position judgment. Cleveland and McGill in their 1984 article in the *'Journal of the American Statistical Association'* (Cleveland and McGill 1984) discuss how judgments based on length are likely to be less accurate than those based on position. A grouped bar chart is a way to allow for easy comparison between individual categories, but is discussed to be less effective in overall comparison.

Chapter 2

Data collection

2.1 Background on survey design

As explained by Wiley-Interscience (2004), a survey is a means of obtaining quantitative information regarding opinions and experiences of the respondents in order to explore the views of the target population as a whole. In this book, a survey is noted as a “systematic” method of collecting data, where the author states that the word “systematic” is deliberately used in order to separate surveys from other methods of information collection. “systematic” is defined by the Collins English Dictionary as something that *"is done according to a fixed plan, in a thorough and efficient way"* (Collins n.d.), and this reflects the manner in which surveys are created in accordance with a given system, where methods for distribution, implementation and analysis are defined under a pre-determined structure. The survey will be delivered to potential respondents in the target population, who will then be asked to complete a series of standardised questions, or questions for which the question ordering and wording is identical for every respondent, unless different formats are to be used to research purposes. It is once again discussed by Wiley-Interscience (2004) that standardised questioning was not always the norm; most interviewers would more likely have a list of objectives, and each interviewer would formulate and word questions based around these. It was discovered that question wording can have a drastic effect on respondents’ answers.

Whether or not the survey is ‘thorough’ and ‘efficient’ depends heavily on the survey structure and design. Designing an effective, systematic survey involves balancing efficiency with completeness, creating a survey that can obtain as much information as possible whilst not boring or fatiguing participants, which can lead to non-response and measurement errors due to participants skipping questions or selecting answers at random. A well-designed systematic survey has the capacity to yield large amounts of both qualitative and quantitative information regarding the research topic while minimising these errors.

There exist a variety of methods for delivering a survey, such as self-completed questionnaires and interviewer-administered interviews. Depending on the aims of the study, there will be advantages and disadvantages to each method. There may also be times when a combined approach is helpful in gathering the necessary information. The first method of surveying, a questionnaire, may consist of either physical paper forms that are mailed or handed out to people within the target population, or in an online format. As discussed by Brace (2004), this form of surveying constitutes a method of indirect communication between the respondent and researcher, in effect a non-verbal conversation in which the respondent is replying to the researcher's questions. The non-face-to-face aspect of this method can be beneficial in terms of anonymity; an anonymous respondent is more likely to be honest in their answers than a respondent for whom the identity is known. As a result, an anonymous questionnaire can mitigate errors that may be caused by respondents fearing judgment of their answers. It is also possible to administer a large number of these questionnaires in a short period of time since they are self-administered, and thus constraints such as the number of interviewers or time taken to administer the survey has less effect on the amount of information obtained.

There are, however negatives to this questionnaire method. In his book, Brace discusses the way in which question wording must be very carefully thought about when using this method of indirect conversation, for reasons such as there being no way to correct participant misunderstanding of questions. Additionally, the fact that the researcher and participant never come into contact may allow the researcher to write questions without considering the human nature of the participants; it is easy to become absorbed in attempting to gather information and fall into forgetting that long-winded or complicated questions may bore or confuse respondents, leading to poorer quality responses. Similarly including too many questions in the questionnaire may lead to response errors for the same reasons. It is then crucial to be as clear and concise as possible in question wording, leaving little room for interpretation. This type of survey is also a very static medium; it does not allow for much expansion on participants' answers, with reasoning behind answers unknown unless specifically requested, which again could add to respondent fatigue and affect quality of response.

We can attempt to implement some dynamic discussion into a questionnaire in the form of 'open-ended questions', mentioned above as specifically requesting reasoning behind answers. A questionnaire is composed of two types of questions; closed-ended questions, for which the respondent selects their answer from a given set of potential responses, and open-ended questions, in which the participants are able to write their answers in a free-form format. Closed-ended questions are very good for obtaining quantitative data that may be easily categorised and counted, which is useful for gathering empirical evidence in order to form objective conclusions regarding the sample population.

Open-ended questions are generally used where more expansion may be required in addition to the closed-form answer, or if using a closed-form question would limit the answer range. The Leibniz Institute for the Social Sciences (Züll 2016) provides guidance on open-ended questions, in which the occasions for using open-ended questions are outlined as:

- "knowledge measurement"; with with multiple choice, respondents would have a chance of guessing the correct answer, and thus this would be a sub-optimal way to measure raw knowledge

- "Unknown range of possible answers"; multiple choice may be limiting for certain questions, and may cause the researcher to miss important information

- "Avoidance of excessively long lists of response options"; if there is a known range of answers, but this range is very large, it may overwhelm respondents to see all of these as options
- "Avoidance of directive questions"; certain questions may have options based on the researcher's own opinions, and thus have the potential to direct the participant in a certain direction, and may not reflect the participants' true views. This links to "unknown range of answers" in that the researcher may incorrectly assume the potential range of answers and thus the given options may not cover the respondents' true opinions.
- "Cognitive pretesting", which covers instances such as ensuring the question was understood correctly.

To summarise, open-ended questions are useful when either there is not enough information to set a standardised range of potential responses or if more information is needed after a closed-ended response.

A method of surveying that is, by design, more dynamic is an interview. An interview may be structured, semi-structured or structured and each of these have a different set of features that distinguish them from one another. Structured interviews, as by the name, are rigid in nature and comprise of a vocal conversation in which the interviewer has a specific set of questions from which the discussion does not deviate. The slightly less rigid semi-structured interview is similar, but slight deviation from the plan is allowed in order to explore new avenues and ideas that might not be found with a structured interview, but the interviewer will still have a set of specific questions for which to obtain responses. For the most flexible of the three, the unstructured interview, the interviewer will tend to follow a loose plan of what they wish to explore rather than a strict question schedule, with the discussion led by the respondent's answers.

Phone calls and other forms of interview-based survey allow the interviewer to form a personal connection with the survey participant, which can be especially helpful for a company's image if the interviewer is particularly professional or charismatic. Additionally, while the interviewer will still be limited to asking the pre-set questions, the format of such a survey can be considered semi-structured and with much more room for interpretation. This can lend itself to gaining additional insights that may not have otherwise been gathered from a more closed-form paper or online survey. Additionally, the more open format can negate any error as a result of participants misinterpreting questions due to the interviewer's ability to immediately clarify on any misunderstandings. This type of survey also provides an instant response, which is beneficial if there is only a short time frame available in which to gather information.

However, there are also shortfalls to an interview-based survey method. For instance, although a charismatic interviewer can positively impact the image of whoever is conducting the survey, this could also lead to biases, such as the respondent answering in a way they feel will please the interviewer. Additionally, the image of the organisation could potentially be tainted if the interviewer appears rude or unprofessional, alongside potentially providing bias in the opposite direction. As well as this, telephone surveys are likely to be interpreted as a telemarketing scheme, and thus potentially have a negative impact on the number of willing respondents. The reduced anonymity of this type of survey may also create bias in the way of participants avoiding making statements that could be deemed socially unacceptable, or that they feel they may be judged for, and therefore may not provide answers accurate to their true line of thought.

The UK Household Longitudinal Study ("Understanding Society - the Uk Household Longitudinal Study," n.d.) is an ongoing study and an example of implementation of a combined use of the above mentioned surveying methods. Initially, in 'wave 1' of the study, a sample of 40,000 households in the UK were selected to be surveyed on a yearly basis. The survey involves all members of each selected household, overall comprising of around 100,000 individuals, and asks them a range of questions regarding areas such as family life, income, employment and health. The study consists of a self-administered youth paper questionnaire given to respondents ages 10-15, and an interview for those aged 16 and up. This split in age demographic allows some questions to be omitted from the youth survey, such as those about income and employment, and some to be added such as about pocket money habits and 'future intentions', as the website states. Giving the youth respondents a paper questionnaire may help obtain more useful or relevant answers, as the respondent may be more comfortable with this than being interviewed by an adult. The youth questionnaire is also shorter, which could perhaps just be a result of many questions not being relevant to this demographic, or it could be a conscious decision, but either way this with help to ensure the young respondent doesn't lose interest and potentially incur bias in their answers due to either rushing to finish the survey or not paying attention. The adult survey also includes a section specific to 16-21 year olds. The surveys contain a standardised set of core questions asked each year alongside a set asked every other year. The reasoning behind this is given to be that this study has a very large scope, asking about many aspects of each respondents' life, and so it becomes inefficient and counterproductive to include all questions every year since, as mentioned previously, the longer a survey is, the more likely a respondent is to get bored or mentally fatigued. The fact that the adult survey is administered in an interview also means that there may be limits on the amount of time the survey can take, as interviewers may have to get through a certain number of respondents in a day, additionally to the interviewer potentially also becoming fatigued. If the interviewer is fatigued, their tone and how they hold themselves may change, and potentially cause a subconscious bias in how the respondent answers the questions.

2.2 Specific goals of survey tool for this study

While visualisations can be a very useful tool for understanding data, they also have the potential to be highly misleading. This section of the study will explore how modifying certain aesthetic features of visualisations can impact perception and interpretation of data, and how these modifications can be exploited in order to mislead the observer. Misleading visualisations may be created in an effort to deliberately influence the viewers' perceptions, or accidentally as a result of poor practice and knowledge surrounding data visualisation. In either case, visualisations have the ability to communicate different messages and stories depending on how they present the data to the observer.

The specific aim of the survey is to test whether altering y-axis scaling, bar width, bar grouping method and colouring will have an impact on single data value interpretation and subjective interpretation of differences in data values.

2.3 Survey Design

The survey design will be inspired by the papers discussed in the previous literature review, all of which investigate how different aesthetic and design choices have the potential to mislead the observer or alter perception.

Following this, questions included in part 1 the survey will focus on gauging whether altering the y-scale to be truncated or logarithmic has an effect on user perception of difference in data point values, for both bar and line plots. The respondents will be asked to gauge both individual values and differences in values, with the former providing an open answer box in which the may type their answer to allow for maximum freedom and obtain their true observation, unimpeded by the bias of having a specific set of numbers to pick from when their true observation may lie outside this range. The question for gauging difference perception follows Lauer and O'Brien (2020) and Yang et al. (2021) in using a numbered scale with numbers representing a range from not much difference up to a large difference. The Yang et al. (2021) method of a 7-point scale was employed here. From these papers, it is hypothesised that the truncated scale will cause respondents to overestimate differences between data values, and the logarithmic scale will be hypothesised to result in underestimation.

Additionally, stacked bar charts will be investigated, showing a comparison between using the stacking method as opposed to a grouped bar plot. Based on reviewing the literature, part 3 of the survey will include questions with the objective of testing standard stacked against grouped bar charts, alongside questions relating to the colour palettes used in depicting the different groups. We aim to test which colour palette is preferred in terms of aesthetics as well as ease of interpretation and reading.

The last two parts of the survey, noted henceforth as 'Sales - part 1' and 'Sales - part 2', explore the different y-axis scalings with respect to line plots, but for these, as opposed to the bar plots, the default was a truncated axis. The three plots investigated will consist of line plots relating to time series data for two fictitious companies. One will display each of the two lines on separate plots with the default axis, one will show both on the same plot with the default axis, and finally one with both on the same plot but with a zeroed axis. It is hypothesised that a difference in value for two time points will be perceived as smaller fopr the zeroed axis, and larger for the separated plots.

As discussed in Peytchev and Peytcheva (2017), too long a survey can result in higher measurement error due to factors such as waning interest or mental fatigue of respondents, resulting in careless responding and non-response. This is also further explored in Brower (2018), whereby a study is carried out to determine causes of careless responding, and specifically looks at questionnaire length and participant disinterest. The study performed in this work provides evidence that longer survey length can have a detrimental affect on careless responding; a long survey may make participants more likely to respond carelessly, and this must be considered when designing an effective and efficient survey. An additional conclusion states that participant interest in the survey content could have an effect, but also that evidence is less supported for this claim. There is significant enough evidence, however, to say that this should also be considered when designing the survey.

The Peytchev and Peytcheva (2017) paper explains that a 'split survey' design, where each respondent is only asked to answer a selection of questions from the whole set, is effective in reducing error while gathering large amount of information, however this will not be employed here. The reasoning for this is that there will already be a set of 12 different surveys being sent, and creating further splits could potentially lead to much too small sample sizes and thus inconclusive results. Additionally to this, the paper investigates how placement of questions in the survey can affect responses, concluding that questions asked later in the survey are more susceptible to bias, which tracks with the conclusion of survey length being a cause of careless responding; the longer a participant is taking a survey for, the more likely they are to start being careless with responding.

Due to this, the survey was designed to last in the range of approximately 15-20 minutes, as suggested in Revilla and Ochoa (2017). One paper (Crawford, Couper, and Lamias 2001) explores the pecieved burden of a survey on the participant, and performs a study whereby respondents were assigned a questionnaire, but given one of two different time estimates, for which the true length of the survey lay between. It was found that more people started the survey with the lower estimated completion time, but more also dropped out. However, the time at which respondents dropped out did not significantly differ in the two groups. In order to obtain maximum response, it is wise to as accurately as possible disclose the true survey length, and even slightly over-estimate in the disclosure.

With regard to the interest factor, the survey was designed with engaging respondents. The topic of the majority of the survey was chosen to be data relating to the television show *American Ninja Warrior*, as this could be subjectively viewed as a 'more interesting' topic than seemingly meaningless numbers. The survey was administered to a test subject, who commented that they found this topic interesting, with the additional comment that perhaps some pictures of the Ninja Warrior obstacles would be nice, however was not employed. The survey also took this respondent about 20 minutes to complete.

Although the content of the surveys for this study is not likely to be controversial or highly personal, anonymity is still important as the participants could otherwise potentially feel pressure to give a 'correct' answer, given the mathematical nature of the questions. As mentioned prior, anonymity here means that this pressure is potentially reduced and thus the relevant measurement bias may be mitigated. Additionally to the more technical visualisation questions, respondents were asked a series of demographic questions such as age, degree subject (if applicable), and whether they are colourblind or have any disorders that my affect visual processing. Additionally, three Likert scaled questions relating to well they would rate their spatial, observational and numerical skills. The Yang et al. (2021) paper, which explores the truncation effect of barplots, looks at graph literacy and its relation to perception, and hypothesises that those undertaking quantitative subjects at PhD level would be less impacted by the truncation effect as compared to humanities PhD students. It was found that the truncation effect did impact both groups, but those in quantitative fields had their perception marginally less affected. Thus the degree subject question was included to explore if this has an effect here. In relation to the visual processing and colorblindness questions, these are again included to test whether they have any significant impact on perception, as it may be important to consider these factors when creating visualisations to ensure they are accessible to all, and the study will examine the potential impact of such disorders.

The set will consist of two groups of surveys, which will be identical up to the visualisation package used. Particularly, one group will contain visualisations made with R's ggplot2, the next with matplotlib from Python. These surveys will be distributed to the general public by sharing links on social media platforms such as Facebook. The reasoning behind creating two separate surveys in different languages is to ascertain whether the language used influences the interpretation. Within the groups there are 6 surveys, with each altering the order of visualisations shown in part 1 to assess the perception of each plot type without reference or comparison to another, and the same with part 2. in Part 3, each of the 6 used one of 3 colour palettes as the main colour, and another as a comparitor to test which the preferred colour palette is and which respondents find easier to read and interpret. Note however that, while both languages were intended to be as close to default as possible, the ggplot visualisations were made such that the theme `theme_classic` was applied, as this is mirrors the Python format in terms of the absence of grid lines.

2.4 Creating the Visualisations

See appendix 1 for a pdf of the survey, containing the finished visuals, and appendices 2 and 3 for code. The R visualisations were created using R version 4.0.2 (R Core Team 2017) using ggplot2 version 3.3.3 (Wickham 2009). The Python visualisations were made using Python version 3.7.4 (Van Rossum and Drake Jr 1995) with pyplot from matplotlib version 3.3.3 (Hunter 2007).

2.4.1 The Data

The visualisations for the survey were created with inspiration from the papers discussed above. The bar plots were created using a data set regarding the history of obstacles used over 10 seasons of *'American Ninja Warrior'* (LAESSIG, n.d.). Each row of the data represents a single instance of an obstacle being used, and each instance has variables as specified in table [?].

This data was manipulated in R to produce a data frame containing the count of the number of times each obstacle was used over the course of the whole ten seasons. For the stacked and grouped bar plots, a data frame was produced, once again in R, containing columns 'obstacle' and 'stage', where 'obstacle' is a vector containing the name of each obstacle repeated the number of times it was used, and 'stage'

Table 2.1: Table with explanation of variables

Variable Name	Explanation
season	Season in which instance occured
location	Location of use
round_stage	Stage of competition in which instance occured
obstacle_name	Name of the obstacle
obstacle_order	Order in which the obstacle was placed in the course

similarly contains the names of all the stages of the competition, with each repeated the number of times it appeared. For example, Salmon Ladder was used 41 times, and thus is also repeated this many times, and there are 41 entries in the 'stage' vector corresponding to this. For the python version, the frequency tables were created manually.

The data for the time series plots was taken from the data set `BJsales` in the base R package `datasets` (R Core Team 2017). This data consists of a single vector of values with 150 entries, where each entry corresponds to a measurement taken at some arbitrary time point. Four subsets were taken from this data such that a start index was selected, and then this entry and the 11 following consecutive entries were extracted. The vectors were put into a data frame with the time steps set as months, giving a year of sales data for four fictional companies. This again was used to manually create a data frame in Python. To select the starting index, several seeds were tested for random selection, and four seeds were selected that would create plots to best test the hypotheses.

2.4.2 The Bar Plots

As explained before, the bar plots for part 1 were made such that one uses the default axis scaling, one uses a truncated axis, and one uses a logarithmically-scaled axis. It is worth noting that in R attempting to truncate the bar plot itself does not work; the bar must start at the zero tick mark otherwise the bars do not show up. To get around this issue, the data itself was truncated before applying to a bar plot with the tick labels then altered to fit the truncation, using intervals of 10 as in the default plot. Python, on the other had, will perform the truncation without this issue and defaults to steps of 2.5, which could affect the reading of values. For the logarithmically scaled plots, R by default starts at 1 and uses a non-standard form notation with tick labels of 1, 3, 20, 30. Python does use standard form and has labels 0, 10^0 and 10^1, starting at zero. The Python scale starting at zero was before mentioned as potentially misrepresenting the data. The height gauging of the R plot could maybe be impacted by the scale starting at 1. The default for the Python control plot scaling was more granular than the R, with steps on 5 as opposed to 10. The control scales for both languages have a range [0, 40], and [20, 40] for the truncated plots. There were 4 bars corresponding to 4 of the most used obstacles, arranged in descending order.

The next part plays with the aspect ratio of the plots. In order to keep this accurate, the plots were saved within the code as opposed to saving from the viewing window. The default aspect ratio for the ggplot is

1/1 for height to width, and using pyplot.gca() and comparing to the default we see that the default for Python using this method is 0.1. For the ‘wide’ plot, the aspect ratios are halved to 0.5/1 and 0.05, respectively. For the narrow, the aspect ratios were doubled to 2/1 and 0.2. Note that the aspect ratios include the entire plotting area, including labels and titles. These plots contained 7 bars as opposed to the 4, but were still arranged in decending order.

The plots in the third part of the survey were the stacked and grouped plots. The three colour schemes were the package default, a greyscale, and the colurblind-friendly Viridis palette (Garnier 2018). The obstacles here were the same 4 as displayed in part 1, but with the added colours for the competition rounds. The default axis ratios here mean that the R plots appear taller in comparison to their width than the Python plots, due to the legends.

2.4.3 The Line Plots

The plots for part 1 of this show the false sales data in the form of time series line plots, where the x-axis displays the months and y-axis shows number of sales. In the R version, the x-axis displays the 12 months in words, whereas the x-axis of Python version numbers the months and plots them in intervals of 2 months. This was an unintentional error on the part of the designer, however could be used to draw conclusions regarding how the two systems differ; monthly ticks in words or bi-monthly numbers. The plots in sales- part 2 were created very similarly, just with two different start indices.

2.5 The Survey

This section will discuss the specific survey questions and explain the differences in plot ordering and colour schemes between survey versions. Google forms was chosen as the medium for delivering the survey, as it is a free service and provides easy way to send out survey links and automatically compiles responses in a Google sheet along with time stamps, which can be exported to csv for analysis. To randomly assign each participant a survey, a javascript code was created to link to a landing page, which redirected the participant randomly to one of the 12 surveys. As time progressed it was possible to see how many respondents were taking each survey, and it was possibly to alter the Javascript accordingly to ensure each survey had an approximately even number of respondents. The survey was set such that each page contained a single question with a set of related sub-questions and only the plots relevant to these sub-questions, to prevent participants scrolling through the survey and seeing other figures which may alter their perception. This can also be used to analyse the effect of seeing other plots on perception of the plots following.

2.5.1 Demographic Questions

As discussed, the questions below are used to assess whether these factors have an impact on graph literacy and graph perception.

- Please enter your age (Open)

- If you are a university student or past university graduate please specify your area of study. (Drop down box: Science, Technology, Engineering, Maths, Arts, Social Sciences, Humanities, Business, N/A, Other (please specify))
- How strongly do you agree with each of the following statements? (Linear scale with 1 - 5, 1=strongly disagree, 5=strongly agree)
- - I have good spatial awareness skills
- - I have good observational skills
- - I have good numerical skills
- Are you colourblind? (Checkbox: Yes, No, Prefer not to answer)
- Do you have any disorders that may affect visual processing? (this could be a general visual processing disorder or dyslexia, dyscalculia, ADHD etc) ((Checkbox: Yes, No, Prefer not to answer))

2.5.2 American Ninja Warrior - Part 1

The questions regarding each of the three bar plots were as follows:

- Approximately many times would you say the 'Salmon Ladder' was used? (Open)
- Approximately how much more than 'Log Grip' would you say 'Salmon Ladder' was used? (1-7 scale)
- Approximately how much more than 'Quintuple Steps' would you say 'Salmon Ladder' was used? (1-7 scale)
- In your opinion, approximately how many times would you say 'Log Grip' was used, as a percentage of the number of times 'Salmon Ladder' was used? (Open)

Here, the two questions with the difference rating scale are used to assess whether having the bars next to each other vs on opposite ends of the plot has an effect on the difference in rating when comparing the responses for each of the plots. The use of the word 'more' in these questions could perhaps be considered leading, as it indicates to the respondent the direction of the difference. The wording of part 2, specifying that distance is being compared but without specifying a direction, may have been better here, and next time would be used. However, this is unlikely to have has too large of an effect, as it is easy to see that the 'Salmon Ladder' was used most out of the four obstacles.

Question 4 here was perhaps too wordy and/or confusing, and asks effectively the same question as the two previous. If this survey were to be applied again, this would be omitted as it adds unnecessary respondent fatigue. This is especially the case since the responses to this question mean it has been omitted from analysis, and so added to respondent fatigue without any additional information gain.

Table 2.2: Order of plots in part 1

	Q1	Q2	Q3
V1	Control	Log	Truncated
V2	Control	Truncated	Log
V3	Log	Control	Truncated
V4	Log	Truncated	Control
V5	Truncated	Control	Log
V6	Truncated	Log	Control

The table below shows all the permutations of the three plot types, and which questionnaire version they appear in.

The table shows that, for example, in version 1, the control plot was shown in question 1, the log-scaled in question 2 and the truncated in question 3.

2.5.3 American Ninja Warrior - Part 2

The questions regarding each of the three bar plots were as follows:

- How large would you say the difference between 'Jumping spider' and 'Salmon Ladder' is? (1-7 scale)
- How large would you say the difference between 'Log Grip' and 'Floating Steps' is? (1-7 scale)
- How many times would you say 'Floating Steps' were used? (Open)

In hindsight, the value judgment question should perhaps have used the same phrasing as part 1. Removing the word 'approximately' from the value judgment question could have an adverse affect on responses by comparison to part one in the way of perhaps making respondents feel they have to give a more 'accurate' and less subjective response than part 1.

Similar to part 1, the below table gives all permutations of the three plot types.

Questions regarding comparisons between the plots were then administered as follows, while showing respondents all of the three plots on a single page.

- Which of the three bar charts do you find most aesthetically pleasing? (Multiple choice with options "A", "B" or "C")

Table 2.3: Order of plots in part 2

	Q1	Q2	Q3
V1	Default	Narrow	Wide
V2	Default	Wide	Narrow
V3	Narrow	Default	Wide
V4	Narrow	Wide	Default
V5	Wide	Default	Narrow
V6	Wide	Narrow	Default

- Which bar chart do you feel is easiest to read and interpret? (Multiple choice with options "A", "B" or "C")
- Which bar chart do you find hardest to read and interpret? (Multiple choice with options "A", "B" or "C")

2.5.4 American Ninja Warrior - Part 3

This part explored the differences in perception for stacked and grouped bar charts, alongside colour preferences. This part had 4 questions, with the first two asking about the stacked and grouped bar plots, with either the stacked first or grouped first.

The first two sub-questions are given below.

- How many times would you say 'Floating Steps' were used in the Finals (Regional/City) rounds? (Open)
- How many times would you say 'Log Grip' was used in the Finals (Regional/City) rounds? (Open)

The next question is "Please select the statement you feel applies to the bar chart above." and consists of a multiple choice answer with the following options:

- 'Log Grip' was used MORE in Finals (Regional/City) rounds than in Qualifying (Regional/City) rounds.
- 'Log Grip' was used Less in Finals (Regional/City) rounds than in Qualifying (Regional/City) rounds.
- 'Log Grip' was used an EQUAL number of times in Finals (Regional/City) rounds and Qualifying (Regional/City) rounds.")

This is followed by another multiple choice question, given as Which obstacle do you think was used MORE in Finals (Regional/City) rounds, 'Log Grip' or 'Floating Steps'?, with the following options:

Table 2.4: Colour palette pairings used in each question

Version	Main colours	Comparitor
V1	Viridis	Default
V2	Default	Viridis
V3	Default	Greyscale
V4	Greyscale	Default
V5	Viridis	Greyscale
V6	Greyscale	Viridis

- ’Log Grip’
- ’Floating Steps’
- They were used the same amount of times

After answering these questions for both plot types, the respondents were shown both on the same page and asked to select which of the two they found easier to read and interpret, and were then shown the stacked bar plot in two different colour palettes; the one used for the questions so far and a comparitor, with the questions below.

For the stacked vs grouped comparison:

- Which bar chart do you feel is easiest to read and interpret? (Multiple choice with options "A", "B", "C")

For the colours comparison:

- Which colour scheme do you find most aesthetically pleasing? (Multiple choice with options "A", "B", "C")
- Do you feel that one of the colour schemes makes it easier to read and interpret the data than the other? If so, please select which one. (Multiple choice with options "No", "Yes, A is easier", "Yes, B is easier")

For this part, survey versions 1, 2 and 4 showed the stacked bars first, followed by the grouped, and versions 3, 5 and 6 displayed the grouped first. It is shown in the below table which colour schemes were used in each survey.

Table 2.5: Order of plots in part 3

	Q1	Q2	Q3
V1	Separated	Truncated	Zeroed
V2	Separated	Zeroed	Truncated
V3	Truncated	Separated	Zeroed
V4	Truncated	Zeroed	Separated
V5	Zeroed	Separated	Truncated
V6	Zeroed	Truncated	Separated

2.5.5 Sales - Part 1

The respondents then moved onto part 1 of the sales section of the survey, in which they are asked to once again give subjective opinions regarding the y-axis scaling, but this time relating to time series line plots.

Once again, the same set of questions is asked for each plot which consist of, firstly, a two-row multiple choice grid, with each row relating to one of the companies. Respondents were asked the question “How much would you say sales of each company increased between January and December?” and were to give a response on the 7-point scale. Again, this could be seen as leading due to the word “increased”, and in hindsight this could be altered to read “changed”, with the response options then potentially on a scale with the centre value (ie 4 on a 7 point scale) corresponding to no change, and the numbers on either side representing a positive or negative change. This would also perhaps be implemented in the questions regarding comparisons of bar heights.

The ordering of the plots for each version number are given below.

The second question was “How large would you say the drop in sales between April and July of Company A is?”, which once again was rated based on the 7-point scale.

2.5.6 Sales - Part 2

The final part of the survey showed zeroed and truncated plots once again, for two different fictitious companies, this time with the intention of gaining an overall view. For each of the two, each respondent was asked a single 7-point scale rating question; “Based on the above graph, how large would you say the difference is between the number of sales Company C makes and the number of sales Company D makes?”.

Chapter 3

Univariate Analysis

This chapter will discuss basic univariate analysis of the survey results, including summary statistics and univariate testing for the whole population as well as the subsetting for the programming language used and degree type. Additionally, subsets will be created considering only the first plot shown for each question, drawing comparisons between responses for these plots themselves without influence of the others. The analysis will be performed in R version R version 4.0.2 (R Core Team 2017).

In terms of testing, Shapiro-Wilk tests will be applied with the `shapiro.test()` function to gauge whether the data sets can be considered normally distributed and thus whether parametric T-Tests are suitable for either one-sample or paired comparisons, for the Shapiro-Wilk test, the alternative hypothesis is that the data is not normally distributed. Failing he normality condition, a symmetry test will be administered via the `symmetry.test()` function from the package `lawstat` (Gastwirth et al. 2020), and providing there is insufficient evidence to reject the null hypothesis that the data is symmetric, a Mann-Whitney-Wilcoxon (MWW) test will be used. If there is sufficient evidence that data proves neither symmetric nor normally distributed, sign tests will be applied. MWW will also be used for two sample testing where perhaps a sign test would be most appropriate, but cannot be used as the samples are of different sizes.

The sample sizes are 70, 38 and 32 for the whole population, R subgroup and Python subgroup, respectively before removing NA of invalid values. The sample means and medians will be notated as $\bar{x}$ and $\tilde{x}$, respectively.

See appendix 4 for all statistical testing results and p-values.

3.1 American Ninja Warrior - Part 1

This part of the survey assess the effect of truncated and logarithmic scaling on bar plots perception and interpretation.

The final question in part 1 of the survey, *'In your opinion, approximately how many times would you say 'Log Grip' was used, as a percentage of the number of times 'Salmon Ladder' was used?'* will not be considered as it is similar to the previous questions, and responses ranged in form, between percentages and decimals, and it can not just be assumed that all the decimals can be converted to percentages; for example a value of 0.5 could be the decimal value for 50%, or the respondent could have meant this as 0.5%.

3.1.1 Effect of Y-Axis Truncation

In general, truncating the y-axis had less of an effect than anticipated. In question 1, *"Approximately many times would you say the 'Salmon Ladder' was used?"*, for which the true value was 41, the distribution of responses for the truncated plot ($\bar{x} = 41.35$) as compared to that of the control plot responses ($\bar{x} = 41.21$) shows a small difference, with the mean perceived value of the bar being slightly higher for the truncated plot. The median for both of these is 41, showing that both distributions are centered around the true value of 41. The control and truncated plots have contextually fairly small variances of 0.752 and 0.753 respectively, depicting both that there is limited variation in the responses and most of the observations lie fairly close to the respective means. The variances are also quite similar, showing that the distributions appear fairly similar, as emphasised by observing figure 3.1 below.

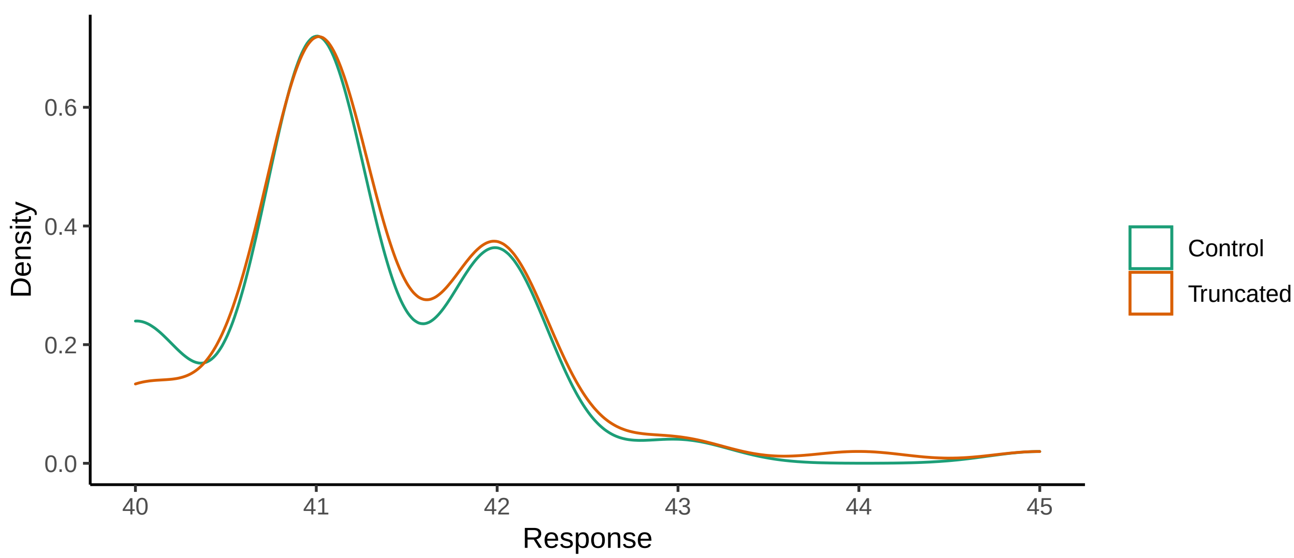


Figure 3.1: Density plot showing distributions of responses regarding the control and truncated plots for the question 1

Performing a dependent-samples sign test comparing these two sets of responses confirms that there is no significant difference ($p = 0.1877$) in the response distributions. However, the one sample sign tests show that there is not sufficient evidence to suggest the control plot responses differ from the true value of 41 ($p = 0.1214$), but there is evidence to accept the hypothesis that the truncated plot responses differ from the true value ($p = 0.0026$). This shows that, while there is insufficient evidence from sign testing to suggest a statistically difference in the responses for the two plots, the location of the truncated plot responses may be slightly further from the true value than the control, and it is confirmed by a one sided sign test with an alternative hypothesis that the true median of truncated responses is greater than 41 ($p = 0.0002$). This gives evidence that the truncated plot results in a slight overestimation in reading of the bar height as compared to the true value of 41. Note that in the responses for the control plot for question 1, there was a response of "41/41", which was taken to be 41.5.

In question 2, *'Approximately how much more than 'Log Grip' would you say 'Salmon Ladder' was was used?'*, the set of responses for the truncated plot ($\bar{x} = 5.87$, $\tilde{x} = 6$) is considered significantly different by a dependent-samples sign test from the control plot responses ($\bar{x} = 5.36$, $\tilde{x} = 5$). By eye, the average values do not seem too different between the two plot types, although the p-value of the sign test ($p = 0.00019$) shows that there is in fact a statistically significant difference. The perceived difference for the truncated plot being rated higher on average than for the control plot provides evidence to accept the hypothesis that using a truncated scale can cause differences in bar height to appear larger, once again this is confirmed by a one-sided sign test ($p = 9.554e - 05$), with the alternative hypothesis that the true median of truncated responses is greater than that of the control responses.

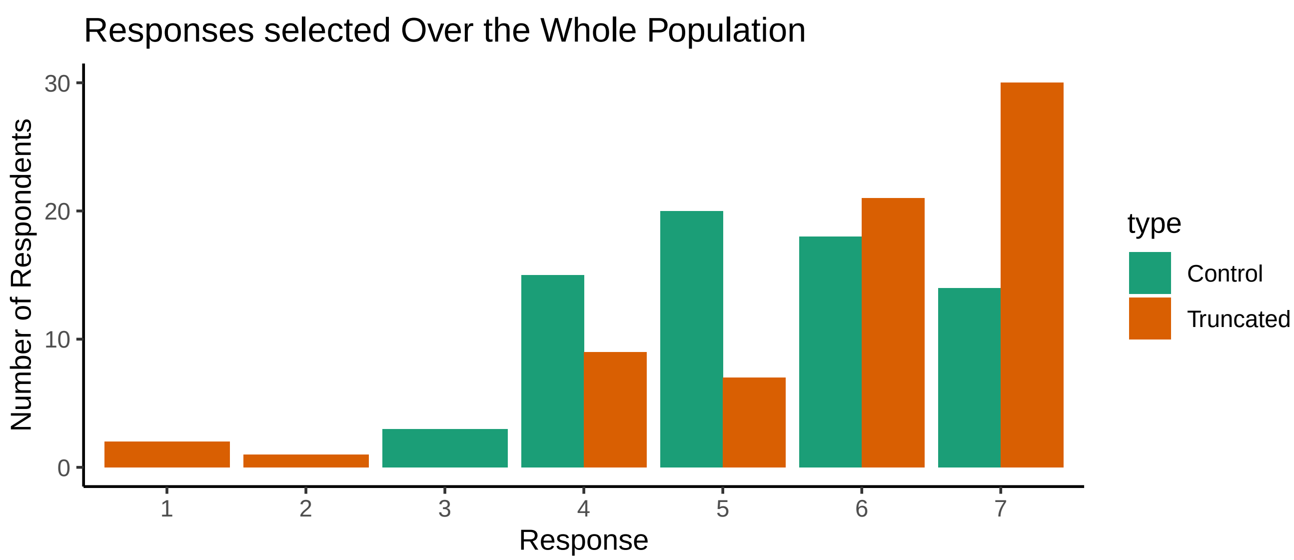


Figure 3.2: Bar plot showing distributions of responses regarding the control and truncated plots for question 2

The spread for the truncated and control plot responses are slightly skewed to the right, depicting that the subjective view on the difference between the bar heights was that it was in general on the larger side. Looking at the bar heights, for the responses of 4 and 5 the control plot bars are higher, and vice versa for the truncated plot response bars. This again emphasises the evidence to support the hypothesis that truncation leads to larger perceived difference.

Question 3 of part 1, *'Approximately how much more than 'Quintuple Steps' would you say 'Salmon*

Ladder' was used?', asks a similar question to question 2, but asks respondents to judge the difference for bars on opposite ends of the plot as opposed to next to each. Again, the by eye comparison shows not a massive difference between distributions of responses for the control ($\bar{x} = 3.12$, $\tilde{x} = 3$) and truncated ($\bar{x} = 3.12$, $\tilde{x} = 3$) plots, although the sign test shows that the there is evidence to suggest that the truncated plot responses are in fact on average greater than for the control plot ($p = 4.624e - 06$). figure 3.3 shows the distribution of responses.

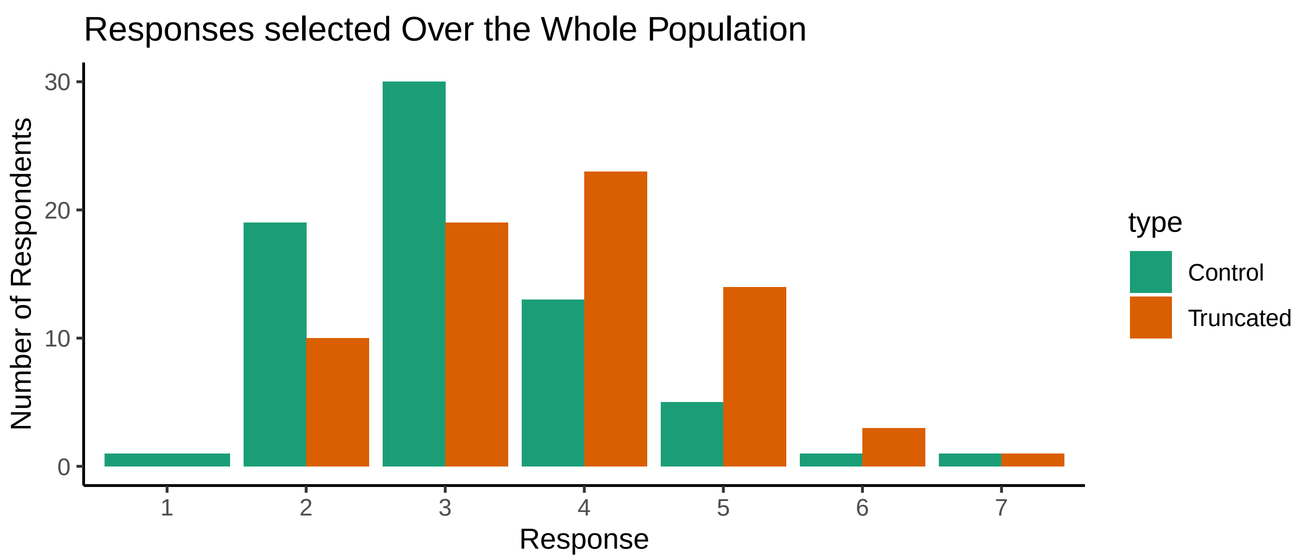


Figure 3.3: Bar plot showing distributions of responses regarding the control and truncated plots for question 3

The response distributions, conversely to question 2, now seem skewed more to the left. However there is a similarity in the way that for the lower ratings of 2 and 3, the control plot response bars dominate, and for the responses of 4 and 5 the opposite is true.

Overall, it seems that the use of truncation has a small but statistically significant effect on perception of height difference between bars, with respondents tending to judge the difference as slightly larger than for the control plot, although this effect is smaller than initially anticipated, and larger for bars that are further apart. In terms of reading values from bars, the truncation did not have a statistically significant effect when comparing the two distributions, however in one sample testing the truncated plot responses did differ significantly from the true value.

When considering the language subgroups, note that there is a discrepancy here between languages in terms of the axis tick breaks and labeling, with the R plot being incremented in steps of 10 for both the control and truncated plots and the Python being more granular in steps of 5 for the control and steps of 2.5 for the truncated.

Consider question 1. Comparing the two language subgroups for the truncated plot, the distributions for both the R ($\bar{x} = 41.56$, $\tilde{x} = 41$) and Python ($\bar{x} = 41.01$, $\tilde{x} = 41$) responses to question 1 appear similar in location to those of both each other and the whole population ($\bar{x} = 41.35$, $\tilde{x} = 41$).

Comparisons via MWW testing show that the responses related to the control plot differ statistically significantly between the two language cohorts ($p = 0.00012$), and similar for the truncated plot responses ($p = 0.02163$), where the tests were performed comparing first the R and Python responses for the control plot, and then for the truncated.

A sign test shows sufficient evidence that the R subgroup responses relating to the truncated plot differ from the true value ($p = 0.0004$), whereas there is insufficient evidence when applying a MWW test to the Python responses ($p = 0.718$). Similarly, the R subgroup's responses in relation to the control plot statistically significantly differ from the true value ($p = 7.629e - 05$), but the Python subgroup's do not ($p = 0.1185$). This could potentially be a result of the less granulated R plot scaling, due to the reduced precision.

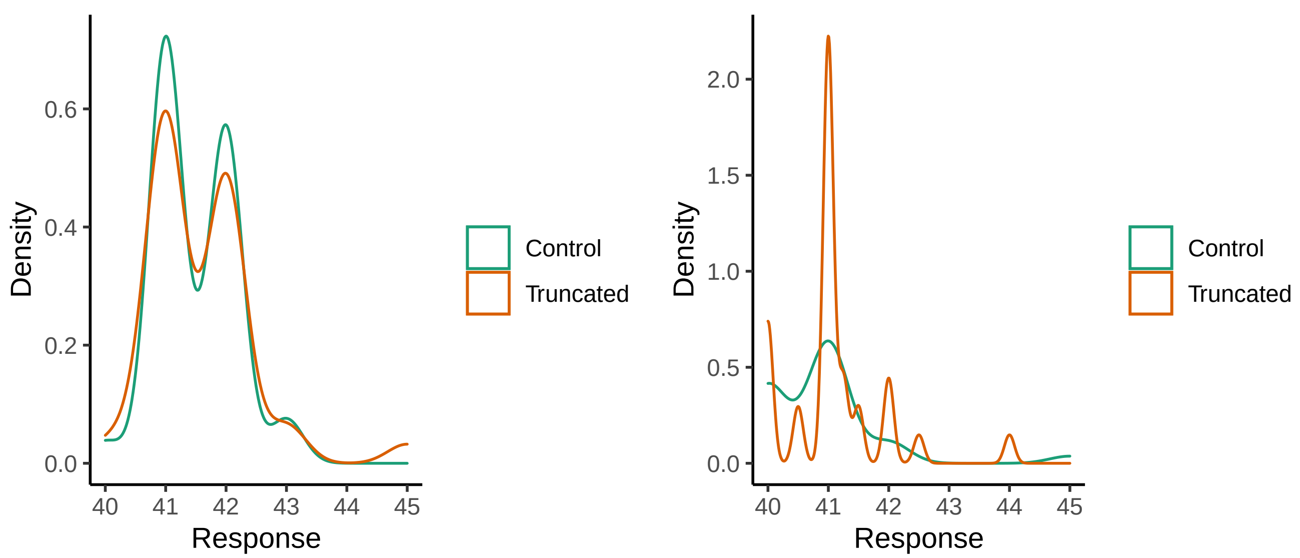


Figure 3.4: Density plot showing distributions of responses regarding the control and truncated plots for the question 1

The distributions for the control and truncated plot responses for the R subgroup are fairly similar to the whole population, although the peaks for the logarithmic plot responses are marginally lower. The distribution of the truncated plots is unexpected from lokking at the numbers, and more 'chaotic'. This shows potentially more variation in the responses.

For question 2 it is similarly seen that the language used does not have a statistically significant impact on the response for the truncated plot, with means 5.500 and 5.187, and medians 6 and 5 respectively for R and Python for the control plot, and means 5.98 and 5.84 both with median 6 for the truncated. Comparative testing with MWW gives $p = 0.2199$ for the control plot and 0.9105 for the truncated. Thus, the scale granulation or any other differing aspect of the plots does not seem to have a significant effect. See figure 3.4 for the distributions.

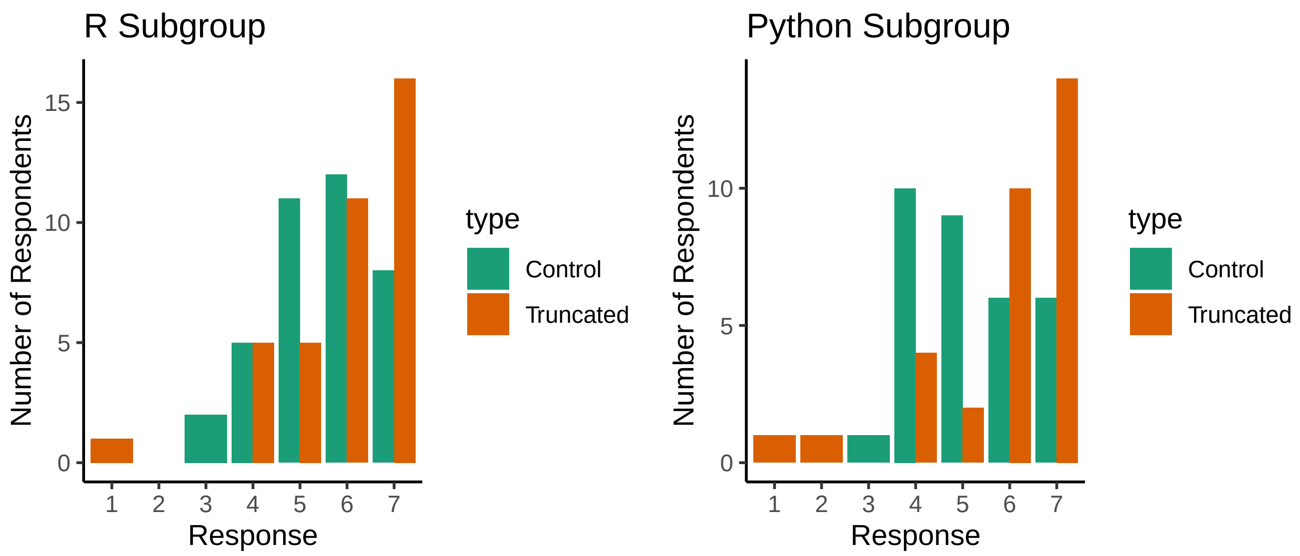


Figure 3.5: Bar plot showing distributions of responses regarding the control and truncated plots for question 2, for the R and Python subgroups

For question 3, it is again seen that the responses in relation to the R version of truncated plot ($\bar{x} = 3.76$, $\tilde{x} = 4$) do not differ significantly to those related to the Python version ($\bar{x} = 3.78$, $\tilde{x} = 4$), with a two sample MWW p-value of 0.9708. Similarly the control plot, there is little difference between the R ($\bar{x} = 3.342$, $\tilde{x} = 3$) and the Python ($\bar{x} = 2.87$, $\tilde{x} = 3$) versions of the plot, again with am MWW p-value 0f 0.1465.

Figure 3.4 shows both distributions, with the R appearing more positively skewed and the python looking fairly symmetric for both plot types, which was also found when performing symmetry tests. For the Python it can also easily be seen that the bars for the truncated plot responses seems 'shifted' to the right slightly as compared to the control.

Now considering subsetting for the respondents that saw the truncated plot first out of the three. Note that 25 saw the control plot first and 23 saw the truncated plot first.

The distribution of responses for the truncated plot in question 1 shows a slightly higher mean (41.696) and median (41.25) than for the whole population, but a MWW test shows that the difference is not significant ($p = 0.1379$). Similarly for questions 2 and 3, performing tests on the truncated plot for respondents who saw this first as compared to the truncated plot responses for the whole population result in p-values of 0.2614 and 0.3145, providing evidence that the plot order doesn't have much of an impact on perception for the truncated plot.

The conclusions appear to be consistent with results from the Yang et al. (2021) paper, in which the

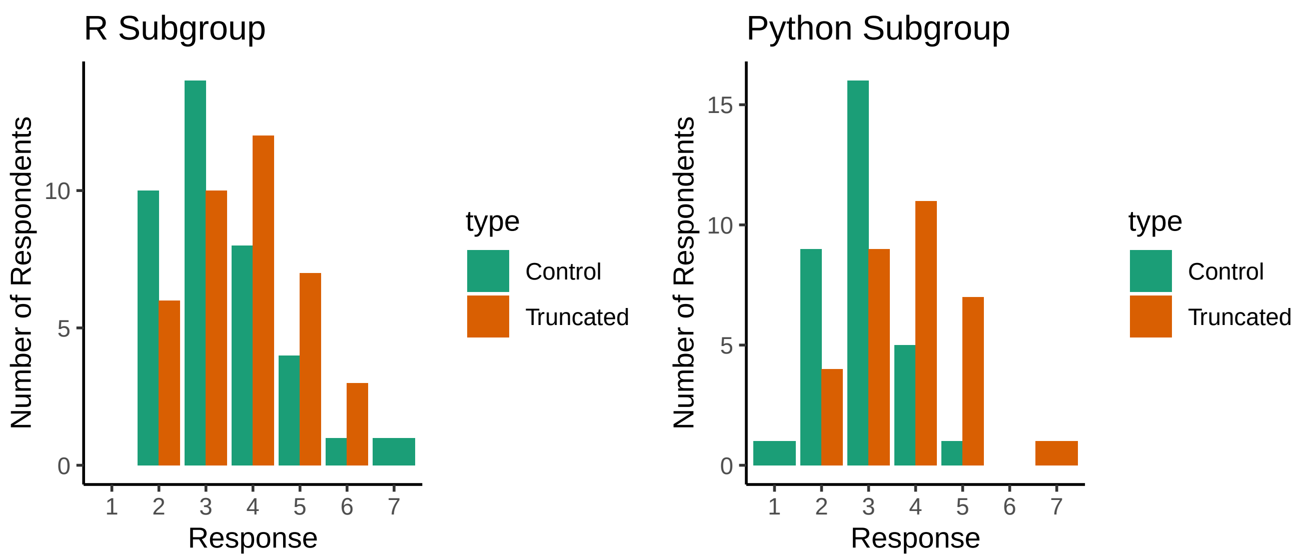


Figure 3.6: Bar plot showing distributions of responses regarding the control and truncated plots for question 3, for the R and Python subgroups

researchers, similar to this survey, showed participants a series of control bar plots alongside those with a truncated axis, and concluded that the difference in values for the truncated axis were perceived to be larger than those of the control plots.

3.1.2 Effect of Logarithmic Scaling

Within the logarithmic responses, there were two invalid responses, given as 'Don't know' and 'Next to none.'. These will be considered as 'NA' responses and discounted from the quantitative analysis, however they do provide useful qualitative insights into how the respondents reacted to the plots, particularly as both were entered for the logarithmically scaled plot made in Python.

The mean of the responses for the logarithmically-scaled plot, on the other hand, was magnitudes higher than the true value at 1.493e+13, although with a median of 35; lower than the median response of the control and truncated plots responses. The high magnitude is the result of two answers of '10ˆ15' and '10ˆ9', both again for the python version of the plot.

The default logarithmic scaling in Python uses standard form notation, which perhaps the two participants who entered the high magnitude answers were less exposed to and not as familiar with. Looking at the degree subjects for these respondents, it is observed that they study Social Sciences and Psychology, respectively. This could add to the idea that they are less familiar with this notation as it is more commonly used in mathematical and physical science disciplines. One of the respondents also rated their numerical skills at 1/5, showing they feel that numerical skill is not their specialty. The other rated their numeric skills at 4/5, showing that even with a good self-perceived level of numerical skill, standard form could be considered misleading.

This should perhaps be considered when designing visualisations; the creator of the visualisations may find the logarithmic scale or standard form more effective in showing the data, but they should consider the target audience. Are the audience going to be familiar with this? If, for example, visualisations are being published in a paper targeted at academics in a subject likely to use such scalings often and understand them, this may be a good way to depict the data. However, using this in something such as an advertising campaign could mislead the public, causing them to either over or under estimate values. As previously discussed, however, this is often done deliberately in order to push the message the creator wishes to sell.

The variance in the responses for the logarithmic plot is also high, with value 1.492×10^{28}, showing that a large amount of the observations differ from the very high mean, and considering this alongside the lower median may point towards many of the respondents either giving an accurate response or even underestimating. Furthering this point, the IQR for the logarithmic responses is the interval $[30, 40.5]$, which sits below the true value, displaying that over 50% of the observations in the total population actually underestimate the value.

The distribution of responses in the R subgroup also shows on average a slight underestimation ($\bar{x} = 39.73$, $\tilde{x} = 35$) and, as expected, vast overestimation for the Python version ($\bar{x} = 39.73$, $\tilde{x} = 35$). This shows that, with a linearly notated logarithmic scale, the scale may cause underestimation, but this is counteracted by using a standard form notation.

It can be considered to follow the convention of values that have value outside the range $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$, where $Q1$ and $Q3$ are the first and third quartiles, which here would be the range $[14.25, 60.75]$ and results in a sample size of 59. Consider now the response distribution for the logarithmically-scaled plot, after removing these responses, for which figure 3.7 gives the density plot. Both plots show the response distribution of the outlier-removed set of responses, with the providing a comparison with the distribution of responses relating to the control plot.

The Python default of standard form notation appears to have confused certain respondents, who are perhaps not as used to seeing this notation, and there was a large range in the responses along with one person not even entering a number, but rather stating that they "Don't know", and another stating they believed the value was "Next to none". The "Next to none" entry is subjective, but could potentially be be assumed as a value close to 0, once again maybe as a result of standard form being less well known to this respondent.

The distribution of responses for question 2 is displayed in figure 3.8.

The spread of logarithmic plot values is fairly wide, with at least one response for each option, and the control is the same as stated before. The plot depicts how there is a wide spread of values, with some respondents having very different subjective views of the size of the difference to others. On average, the subjective perceived difference in bar heights was significantly lower for the logarithmic plot responses ($\bar{x} = 3.67$, $\tilde{x} = 3.5$) than for the control ($\bar{x} = 3.35$, $\tilde{x} = 5$). This is evidenced by a one-sided sign test with the alternative hypothesis that the logarithmic plot responses are on average lower than the control plot

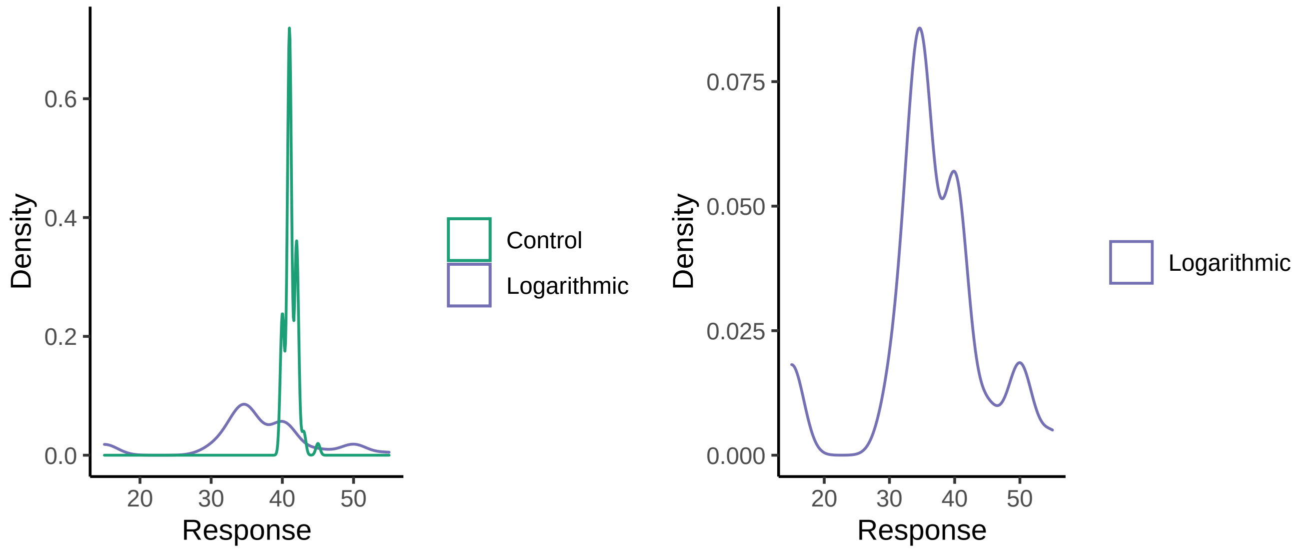


Figure 3.7: Density plot showing distributions of responses regarding the control and logarithmic scaled plot, after removing values of greater or equal to 1000

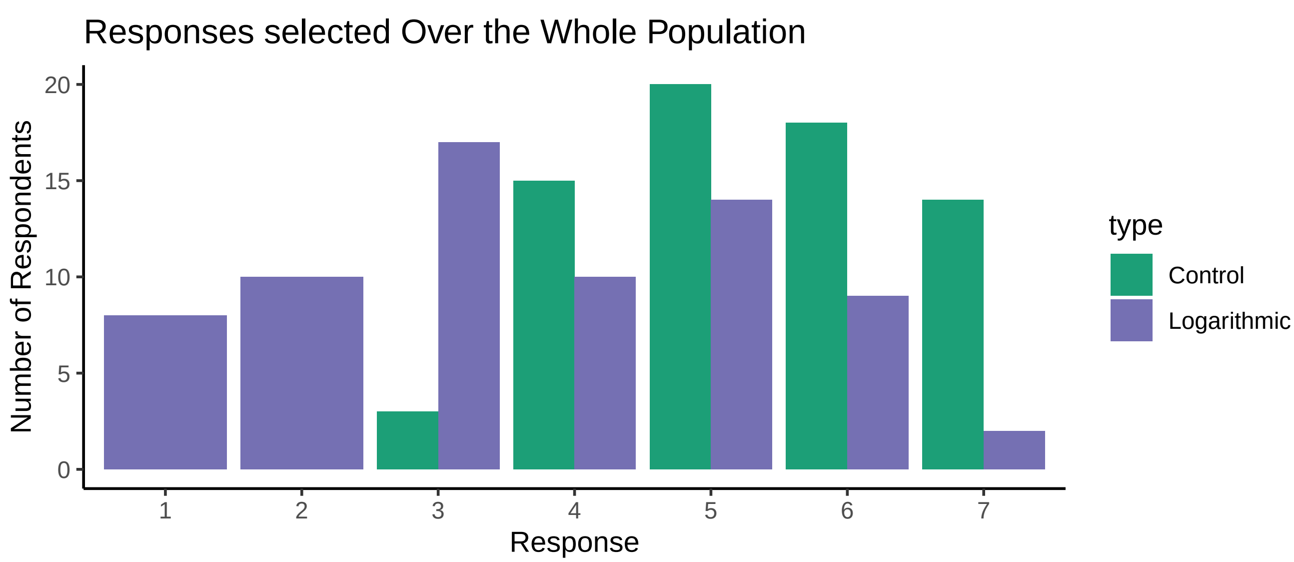


Figure 3.8: Bar plot showing distributions of responses regarding the control and logarithmic plots for question 2

responses.

There is evidence to show that the difference between the R and Python versions of the logarithmic plot is significant ($p = 0.00096$, $\bar{x}_R = 4.263$, $\bar{x}_{Py} = 2.969$). The distributions for the two language subsets are shown in figure 3.9.

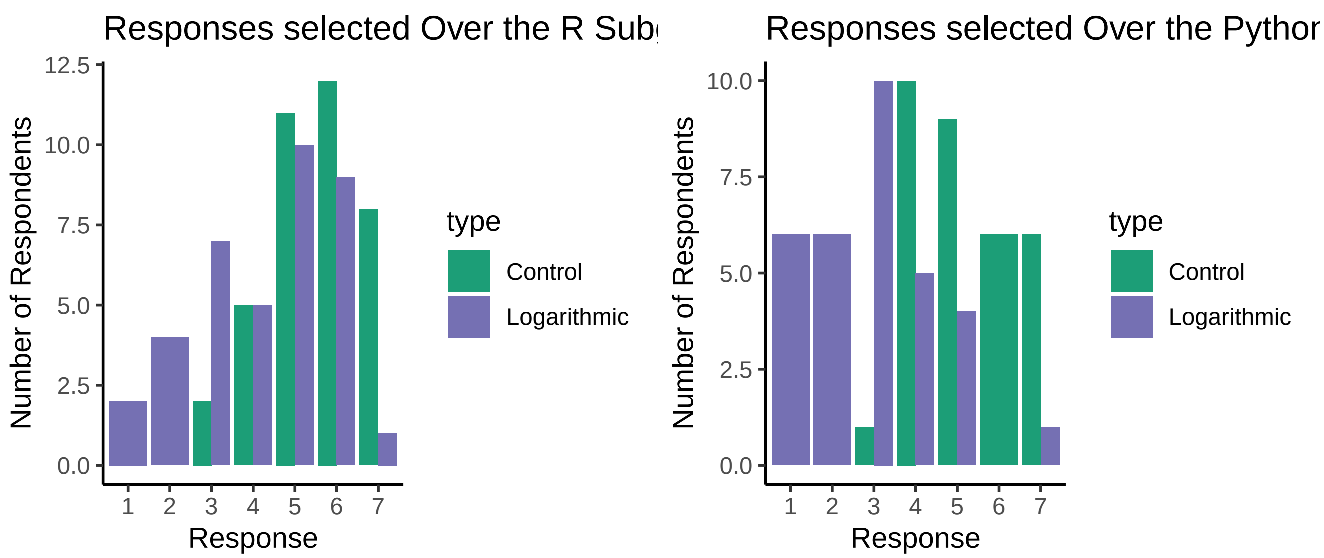


Figure 3.9: Bar plots showing distributions of responses regarding the control and logarithmic plots for the question 2, separated by language

In regard to question 3, see again the below figure for the plotted distributions.

The responses for the logarithmically scaled plot are skewed towards the lower end of the scale, similar to the control and truncated responses, and there does not appear to be much difference between distributions of the two populations. Looking at the numbers, however, the averages for the logarithmic plot ($\bar{x} = 2.22$, $\tilde{x} = 2$) seem lower than that of the control plot ($\bar{x} = 3.77$, $\tilde{x} = 4$). Indeed, a one sided MWW test comparing the logarithmic and control plot responses elicits a p-value of $1.317e - 06$, showing evidence that the logarithmic scale resulted in lower rating in difference of bar height.

Figure 3.10 shows the distributions for R and Python subgroups.

The distributions of the logarithmic plot responses for the R ($\bar{x} = 2.5$, $\tilde{x} = 2$) and Python ($\bar{x} = 1.9$, $\tilde{x} = 2$) subgroups appear fairly similar, with the same median albeit with the mean for the R subgroup being slightly higher. The plots to appear to show the R subgroup responses being slightly positively skewed and the Python responses more centered around 3. A two sample, one sided MWW test provides sufficient evidence that the R responses appear in average greater than the Python ($p = 0.03689$).

Looking at the responses from the respondents who saw the logarithmic plot first of the three, the average responses from this group for question 1 ($\bar{x} = 40$, $\tilde{x} = 40$) were closer to the true value of 41 than for the

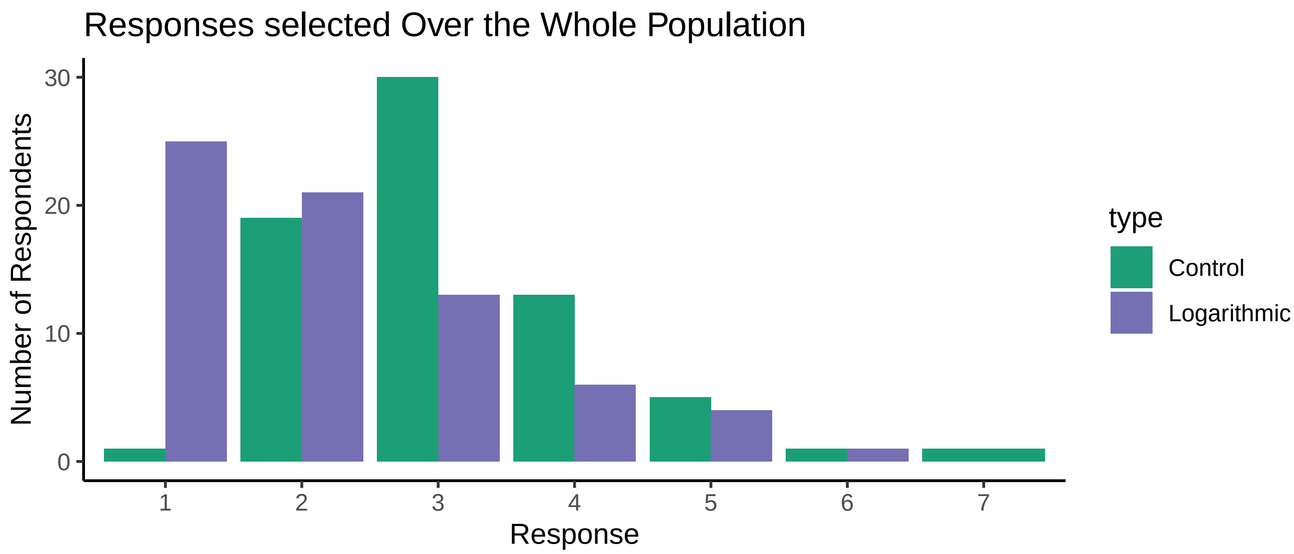


Figure 3.10: Bar plots showing distributions of responses regarding the control and logarithmic plots for the question 3

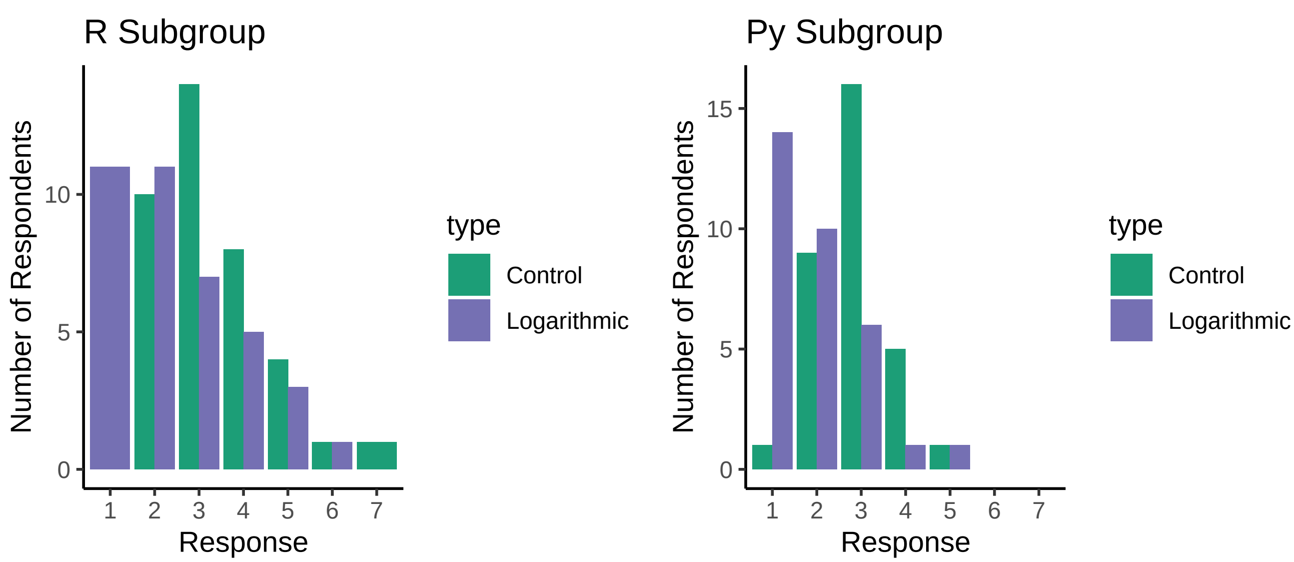


Figure 3.11: Bar plots showing distributions of responses regarding the control and logarithmic plots for the question 2, separated by language

whole population ($\bar{x} = 36.277$, $\tilde{x} = 35$), although the former still differs significantly from the true value ($p = 6.104e - 05$), and there is not significant evidence to state that the two distributions differ ($p = 0.1705$). Comparing the response statistics for the whole population and for those who saw the logarithmic plot first, the log first group perhaps show the bar height difference being perceived slightly higher than for the whole population ($\bar{x}_{overall} = 3.67$, $\bar{x}_{logfirst} = 4.13$), however a two-sample MWW test gives an insignificant p-value of 0.2614 when comparing them. Similarly, the difference between the responses for the whole population and for those who saw the logarithmic plot first for question 3 is also statistically insignificant, with means of 3.08 and 2.68 for and a p-value of 0.1889.

3.1.3 Differences Between Question 2 and 3 Responses

Now take $\bar{x}_{control} - \bar{x}_{truncated}$ and $\bar{x}_{control} - \bar{x}_{logarithmic}$ for each of questions 2 and 3, which is shown in figure 3.11.

Table 3.1: Table showing difference in the percieved difference for the logarithmic-scaled and truncated plots as compared to the control, for questions 2 and 3

	Con - Trnc	Con - Log
Q2	-0.5142857	1.685714
Q3	-0.6428571	0.900000

This again shows that the responses for the truncated plot were in general rated higher than the control plot responses, and also that the effect was more significant for the bars on opposite ends of the plot as compared to the bars next to each other. The opposite is true for the logarithmic plot responses; on average they were rated lower than the control plot, but this was greatly more significant for the bars next to each other, as opposed to the truncated plot. Figure 3.11 shows this visually.

On average, truncating the scale had a similar effect for both questions, albeit with slightly more effect for when comparing ‘Salmon Ladder’ with ‘Quintuple Steps’ as opposed to ‘Log Grip’. For the logarithmically scaled plots, however, the re-scaling appears to have had a significantly greater effect when considering the bars directly next to each other, with respondents on average judging the difference in bar height to be greater by 1.68 on the 7-point scale, whereas this is 0.9 for the bars further apart. It can be concluded from this that truncating the scale had more of an impact when bars were on opposite ends of the plot as opposed to next to each other, and the way round for the bars close to each other; the logarithmic scaling had more of an impact.

3.2 American Ninja Warrior - Part 2

This part of the survey assessed whether different aspect ratios would have an impact on perception of bar height differences as well as reading of true values. This part will be analysed question by question.

Question 1 asked *’How large would you say the difference between ’Jumping spider’ and ’Salmon Ladder’*

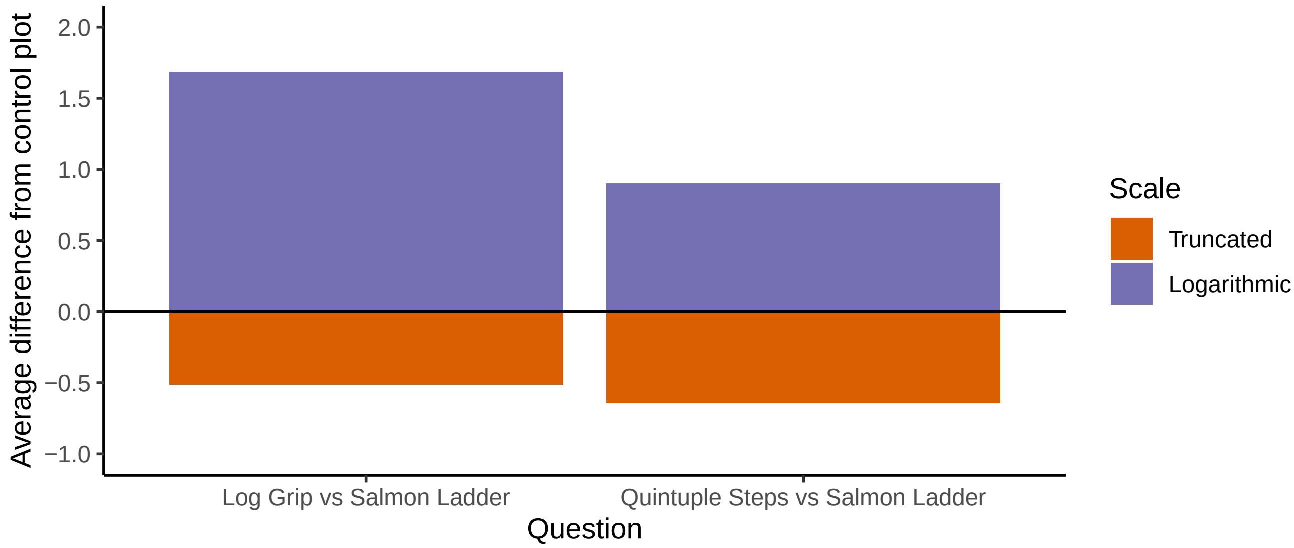


Figure 3.12: Bar plot giving a visual representation of the table

is?'. This question once again uses the 7-point scale to gain a subjective view on the degree to which respondents felt the heights between the two bars corresponding to 'Jumping Spider' and 'Salmon Ladder' differed for three bar plots of 7 obstacles, where 'Salmon Ladder' is furthest to the left, and 'Jumping Spider' furthest to the right.

Looking at the means and medians here, it doesn't seem like there is that much of a difference in perception of the differences between the three aspect ratios, as displayed in table[?].

Table 3.2: Table showing means and medians

	Default	Narrow	Wide
Mean	5.914	6.129	5.357
Median	6.000	6.000	6.000

Note that 'narrow' is defined as the plot with the aspect ratio of smaller width to greater height, and vice versa for the 'wide' plot. The means show marginal differences, whereby the default plot mean is the middle-valued mean of the three, with the mean perceived difference for the wide plot being slightly smaller than this and the mean perceived difference for the narrow plot is slightly larger. This result, although at first glance marginal, follows the hypothesis that the wide plot would cause differences to be perceived as smaller and narrow bars to cause differences to be perceived to be greater.

Now looking at figure[?], showing the three distributions. There isn't an immediately obvious difference in distributions, but on closer inspection it can be seen that the orange "Wide" bars dominate over the three for the range [2, 5], and the purple "Narrow" dominated for the response of 7, following the above analysis of summary statistics. There was a fairly strong consensus that in general that a rating of 6 was applicable to all three plots.

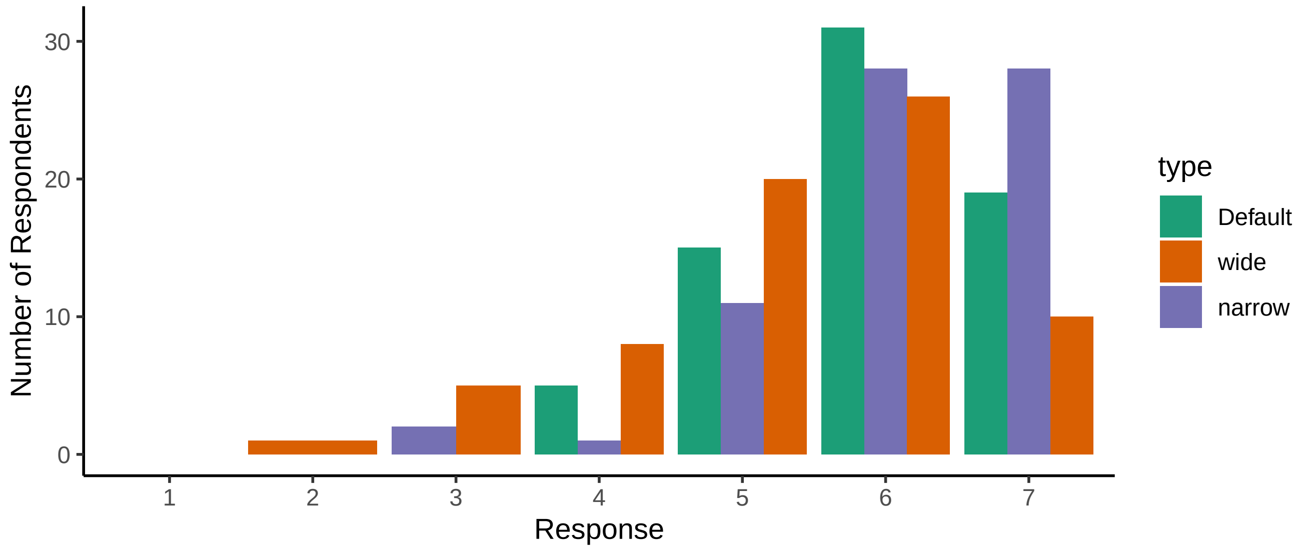


Figure 3.13: Bar plots showing distributions of responses regarding the three plots

Running a one-sided MWW test to compare the responses for default plot to the narrow plot, it is confirmed that there is evidence to suggest that using a ‘narrow’ aspect ratio causes the perceived difference to be greater ($p = 0.0468$). Then applying a one-sided sign test to compare the default to the wide plot, the perceived difference is shown to be smaller ($p = 6.457e - 06$).

Question 2 then went on to ask *’How large would you say the difference between ’Log Grip’ and ’Floating Steps’ is?’*. Similar to part 1, there are two questions for gauging differences between bars, for which one asks about bars far away from each other, and one about bars next to each other. In the case of this section, the first question contained bars on opposite ends of the x-axis, and this question asks about two bars that sit adjacent to one another.

The analysis results here show that altering the axis ratio appears to have even less of an effect than in the first question, with the means of the responses for the default and wide plots being identical at 3.057, with the mean of the narrow plot responses only 0.157 greater at3.214. The median for all three is 3, and the IQRs are all $[2, 7]$. The variances, however, do differ from one another, with values 1.301, 0.866 and 1.214 for the default, wide and narrow bars, respectively. The distribution of values are shown in figure[?]. The results of two-sided MWW tests show that neither aspect ratio appears to have a significant effect on the rating of the perceived difference ($p = 0.2446$ and $p = 0.5688$).

At least 50% of respondents placed the difference in the range $[2, 4]$ for all three plots, showing that they believed the difference was small to moderate, and this didn’t change depending on the plot type, and thus for the bars further apart from each other, changing the aspect ratio does not appear to make much of a difference. The overall distributions are shown in the figure 3.14.

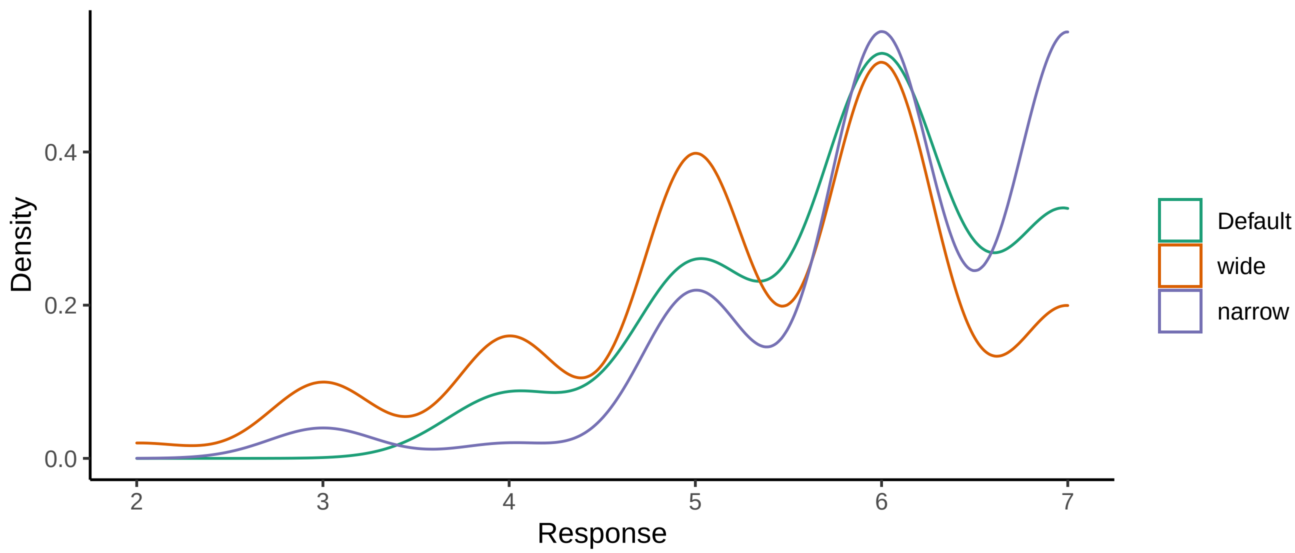


Figure 3.14: Density plot showing distributions of responses regarding the three plots

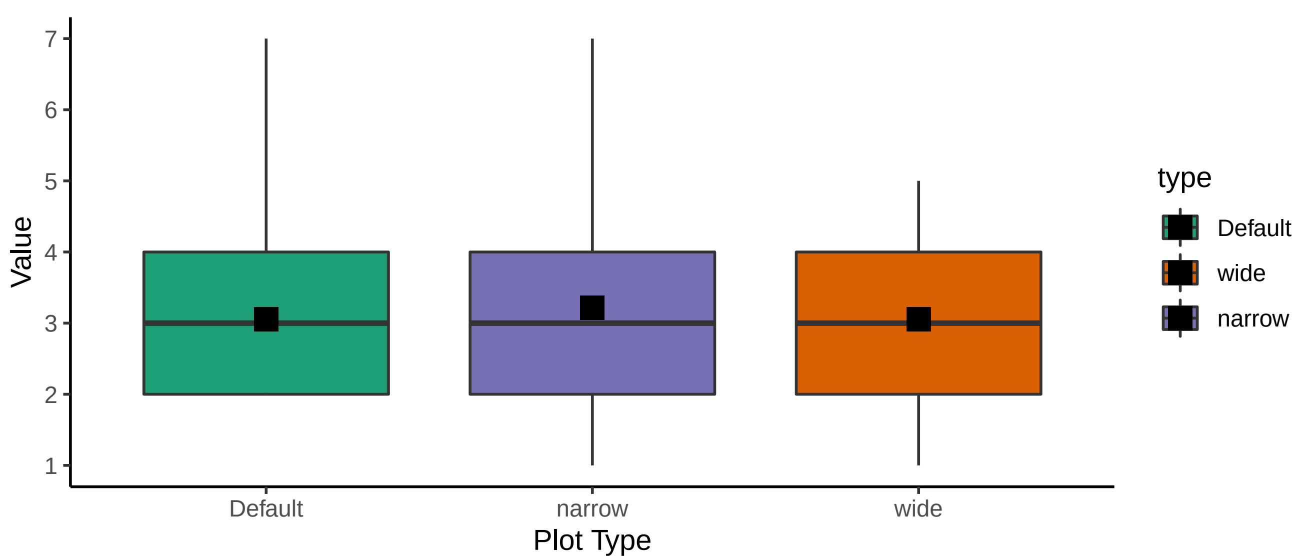


Figure 3.15: Box plots showing distributions of responses regarding the three plots

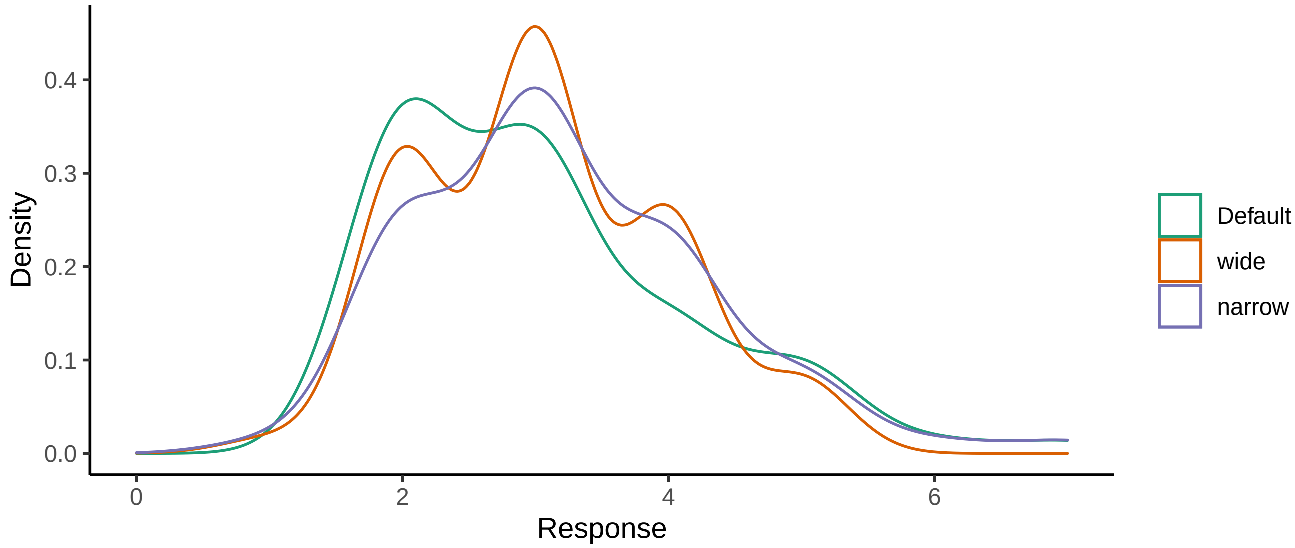


Figure 3.16: Density plots showing distributions of responses regarding the three plots

All three distributions are very similar, and almost appear to form bell curve shaped distributions, albeit with some irregularities and very slight negative skew.

As in part 1, the two height difference perception questions will be compared, calculating $\bar{x}_{default} - \bar{x}_{narrow}$ and $\bar{x}_{default} - \bar{x}_{wide}$.

Table 3.3: Table showing difference in the percieved difference for plots with narrow and wide bars as compared to the default, for questions 1 and 2

	Def - Narrow	Def - Wide
Q1	-0.2142857	0.5571429
Q2	-0.1571429	0.0000000

As before, the figure below gives a visual representation.

Both by eye comparisons of values and statistical testing show that the language used has negligible effect on the perceived difference, as does the order in which the plots were shown. See tables 51 - 61 in appendix 4 for more details.

3.2.1 How many times would you say 'Floating Steps' were used?

This is again similar to question 1 of part 1, where participants were asked to state what they believed to be the height of the bar for 'Salmon Ladder', however this time the third bar from the axis is chosen. This is to ascertain whether the distance of the bar from the axis may have an effect alongside any potential perceived distortion of values. Note that the true value was 28.

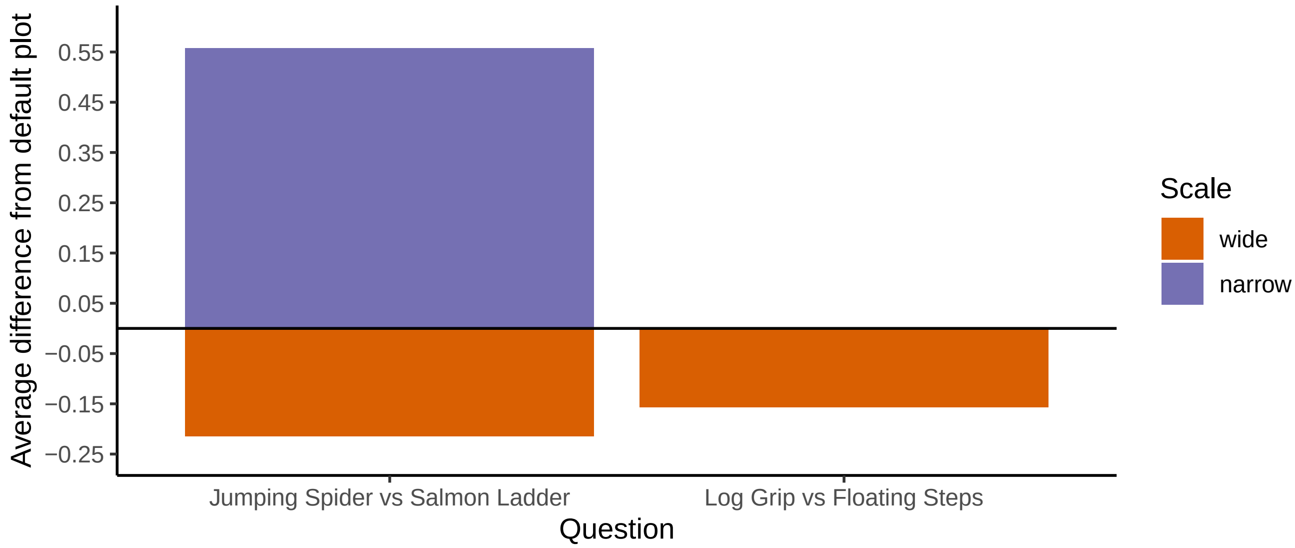


Figure 3.17: Bar plot giving a visual representation of the table

The means of each of the three sets of responses were very close to the true value, at 27.97, 28.04 and 27.39, respectively for the default, wide and narrow, and the medians are exactly equal to the true value. Based on the means and medians it appears that, once again, altering the aspect ratio had minimal, if any, effect on interpretation of the data value. The value for the default plot also appears to be closer to the true value than the control plot in part 1, question 1.

Looking at the box plots, there are very small ranges in the values, signifying that there was a large consensus between respondents in terms of what they perceived the height to be. It can also be seen that there are three outliers below the box plot for the narrow plot responses, and two above for the default plot responses. There is very little overlap between the boxes, and it appears again that there altering the aspect ratio of the bar plot has little to no impact on reading the height of the bar. Additionally, there was less agreement between respondents for the wide plot than for the other two, although this doesn't seem to be too significant.

The distributions for the default and narrow plot responses are similar, both seeming to be fairly centred on the mean with a steep decrease in density on either side of the mean to shallow tails within the range $[25, 30]$. The responses for the wide plot appear to be more spread with lower density function values, with a slight negative skew.

After removing the outliers the medians have stayed the same, and the mean has obviously decreased for the default and increased for the narrow, however, these means are all still fairly similar to each other and at a first glance prior to testing it again seems that changing the aspect ratio, at least to the degree tested here, is inconsequential to interpretation of the actual value. As expected as well, the variances for the outlier-removed sets have decreased.

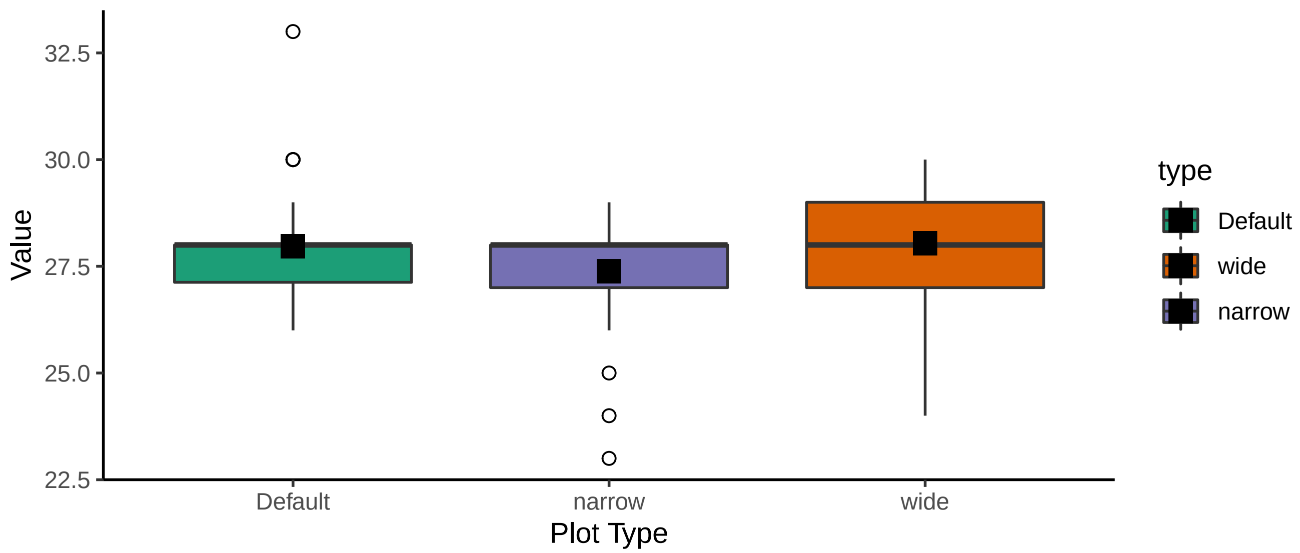


Figure 3.18: Box plots showing distributions of responses regarding the three plots

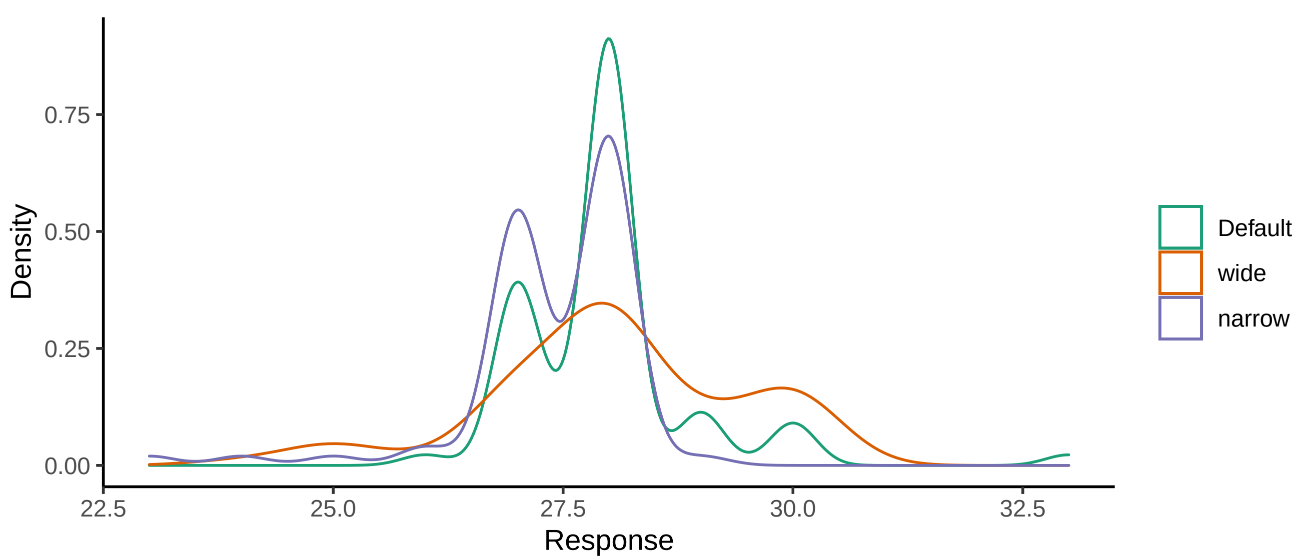


Figure 3.19: Density plots showing distributions of responses regarding the three plots

However, statistical tests do actually show that while the default responses did not differ significantly from the true value of 28 ($p = 0.5667$), the responses for the narrow plot did ($p = 2.0955e - 09$), but the wide didn't ($p = 0.5067$).

Changing the language and plot order was once again inconsequential here.

3.2.2 Comparison questions on aesthetics and ease of interpretation

The last set of questions in part 2 show respondents all three of the bar plots presented in this section and ask them to select which they find most aesthetically pleasing, and which they find easiest and hardest to interpret. Table[?] gives the number of respondents that selected each plot for each of the three questions.

Table 3.4: Numbers of responses for each option

	Default	Narrow	Wide
Most aesthetically pleasing?	37	14	18
Easiest to read and interpret?	36	15	19
Hardest to read and interpret?	20	20	30

For the first question, relating to how aesthetically pleasing respondents found each plot, just over half of the respondents chose the default aspect ratio as the most aesthetically pleasing, with 37 out of the 69 who responded selecting this.

Similarly, 37 out of the 70 that responded to the second question found the plot with the default aspect ratio easiest to read and interpret. Perhaps the people that preferred this aspect ratio aesthetically did so because they found it easiest to interpret. Investigating this, 27 respondents who chose the default for question 1 also chose this for question 2.

The plot judged hardest to read and interpret by the most respondents was the one with the wide bars, with 30 selecting this and 20 selecting each of the other two. While a significant number chose the default and narrow bars, the slightly higher amount selecting the plot with wide bars matches the previously stated hypothesis formulated from following the Stephen Few paper, which discusses that an ratio of greater width to length could suffer from perceptual imbalance. While this imbalance isn't seen in the numbers from the previous questions, the result here does give some indication that the aspect ratio producing wide bars may impact on ease of interpretation.

3.3 American Ninja Warrior - Part 3

The third and final part of the questions about the American Ninja Warrior data discusses stacked bars and colour schemes. The questions asked in this part are used to decipher how data with multiple categories may be best represented in a bar plot. The plots presented use the same bars as in part 1, but this time the number of times each obstacle was used in each stage of the competition for each bar is highlighted. Each participant was shown both a stacked and a grouped bar plot in one of three colour schemes; the default for the language, viridis, and greyscale. For three versions of the survey, the stacked bars were shown first, and for the other three versions the first shown was the grouped bars. The final question of this part also asked respondents to compare two colour schemes, and through the 6 surveys there are comparisons of every colour scheme against every other colour scheme.

The question *"How many times would you say 'Floating Steps' were used in the Finals (Regional/City) round?"* is the first here, and is regarding the reading of a numerical value off the axis. In this question respondents were asked about 'Floating Steps', which is the bar third along from the y-axis. The question asks respondents to view the bar plot, where the bars will either be grouped of stacked, and decipher how many times this obstacle was used in the specified round of the competition. The true value for this was 11. The hypothesis for this question is that the respondents will more accurately gauge the value for the grouped bar than the stacked, which as seen below appears to be the case.

The mean for the values estimated by respondents using the stacked bars is 14.32, a fair bit larger than the true value of 11, and the mean estimated value for the grouped bars was closer to the true value, at 11.8. The IQR for the grouped bars is also smaller than for the stacked, and comprises of the range $[11, 12]$, insinuating that the estimated values tended to be fairly accurate but with some respondents perhaps slightly overestimating. The IQR for the stacked bars on the other hand covers the interval $[10, 14]$, which does contain the true value, but shows a tendency for both over and underestimation of respondents. Additionally to this, there is a large variance in the responses to this question, at 54.8 compared to the variance of 13.1 for the responses regarding the grouped bar plots. This adds to the picture that there was much less agreement between respondents, with many straying away from the mean of 14.3. It is seen however that the median for both the stacked and grouped bars is 11, showing that the higher mean of the stacked bars may be a result of an influential value at the upper end of the distribution, and that many observations do actually sit around 11. The fact that many values actually sit around 11 could be contributing to the higher variance, as variance is simply the sum of the squared distances from the mean, and so will be elevated if there are many values that sit some distance away from the mean. The higher mean could be reflected in the maximum of the stacked responses being 35, although the maximum of the grouped responses is 40, so there may be more than one influential point in the stacked responses. Outliers can be checked for by looking at the box plots for this data.

It can in fact be seen that the box for the grouped responses is short and centered around 11. The box for the stacked responses shows many high valued outliers that could be causing the mean to be higher, although the IQR is still a fair bit larger than that of the responses for the grouped bars. The mean for this also sits above the IQR, and thus the outliers may be having a significant influence. Now the outliers will be removed, assuming, from the box plot, that outliers are any values above or equal to 25 for the stacked responses and above or equal to 20 for the grouped.

Removing the outliers as specified by the box plot, the mean of the stacked responses is now just above 11, and actually closer to the true value than the mean of the other set of responses, and the median has

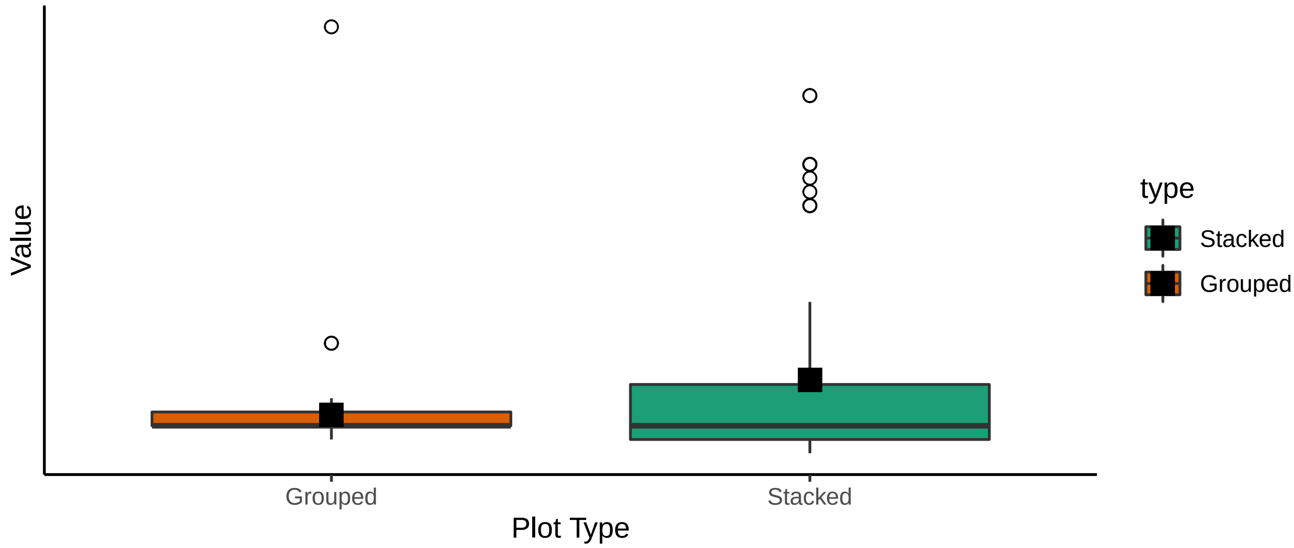


Figure 3.20: Box plots showing distributions of responses regarding the two plots

decreased to 10. From this one could infer that there is no difference between each type of bar plot in terms of gauging the size of the bars. However, there are 12 outliers in the stacked responses, which leads to the idea that these are not in fact all outliers and may be valid responses that just sit on the upper end of the distribution. However, it seems the cause of the high values could be respondents taking the whole height of the bar, which has an actual height of 28, rather than the section of interest. Many of the potentially influential values fall around the range $[25, 30]$, with all but 2 of the 12 potential outliers sitting in this interval, with the remaining two both being 35. Looking below at the summary statistics for only the values picked up as outliers, there is a mean of 29.83, which is higher than the true value of 28, and interestingly goes against the analysis from part 1, question 2 whereby respondents were asked to judge the height of this bar and on average underestimated. The fact that so many participants misinterpreted this plot and signify that stacked bar plots may not be the best way to present data to general public, as there may be the potential to misread the height of the whole bar as the size of the top category.

As a result of this, this set of 12 values will be discounted from the analysis, and thus come to the conclusion that, for the respondents that appear to have judged the height of the correct section, there was little to no impact when using stacked vs grouped bar charts, and most of the difference comes from misinterpretation of the plot itself, as opposed to a poorer judgment of size.

To see if either of these values are significantly far from the true value, tests are once again run. A sign test on the stacked bar plot responses gives a high p-value of 0.5258, showing that for the stacked bar plot responses (after removing the values as priorly specified), the participant estimated values do not differ significantly from the true value. For the grouped bar plot the obtained p-value is $0.009 < 0.05$, and thus these responses are statistically significantly different from the true value.

The next question, *'How many times would you say 'Log Grip' was used in the Finals (Regional/City) round?'*, is similar the above, but for the next bar to the right. The purpose of this question was to test the same hypothesis as the previous question, and also to lead into the following question, where respondents were asked to compare the 'Floating Steps' and 'Log Grip'. Additionally, the bar in the previous question had only two categories, of which the respondents were asked to judge the size of the category on the top of the bar in the stacked plot, whereas the bar for 'Log Grip' has 5 categories, of which the category of interest sits above 4. The true value of this was 9.

Similarly to the previous question, the mean response for the stacked bar plots are higher than that of the grouped, and the mean of the stacked also slightly overestimates the value. Once again however, a selection of respondents appeared to judge the full height of the bar rather than the category as asked.

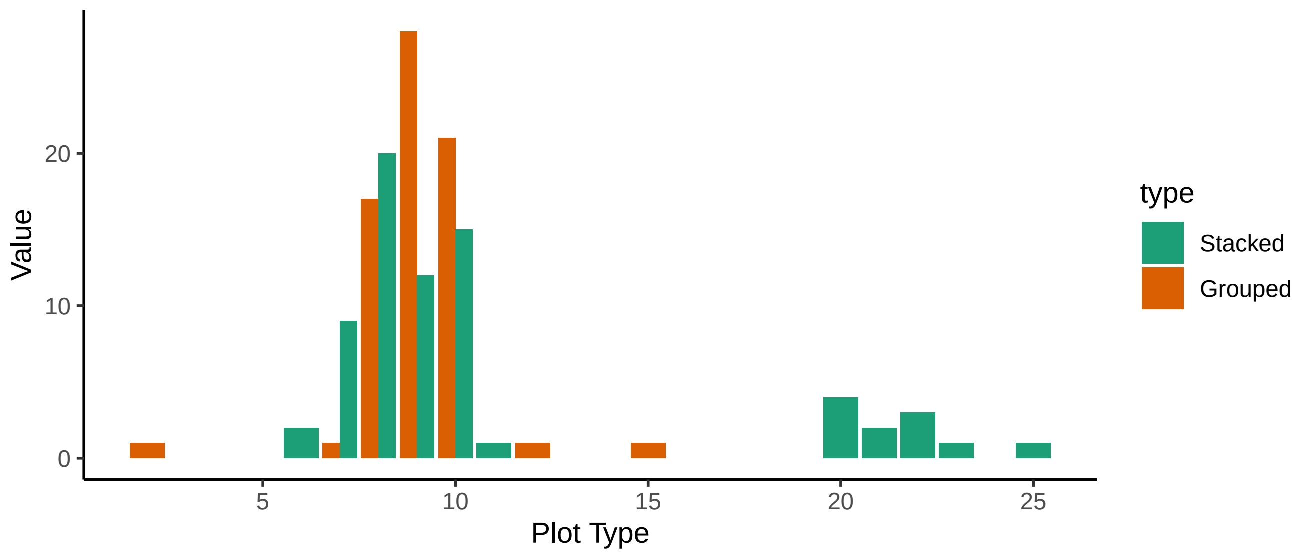


Figure 3.21: Bar plots showing distributions of responses regarding the three plots

Indeed, the distributions of values for each of the two response sets appear to be almost identical.

After removing the outlying values, there tended to be a slight underestimation in the value for the stacked bar plot, however this is approximately 0.46 away from the true value, and unlikely to be significant.

Once again the response sets are non-normally distributed and asymmetric, and so sign tests are applicable. The response set for the stacked bar plots produces a p-value of around 0.04, which shows a statistically significant difference in the responses from the true value of 9 at the 0.05 level of significance. However, this would easily become insignificant by slightly lowering the significance level to, say, 0.035. The p-value for the grouped bar responses, however, is » 0.05, as expected given that the median of the data sits at the true value.

The respondents were then asked to *'Please select the statement you feel applies to the bar chart above.'*.

This question asked respondents to judge whether log grip was used more, less, or an equal amount in the Finals (Regional/City) and Qualifying(Regional/City) rounds. This was to see how well differences between sizes of categories are judged when relating to the same variable, and are in the same bar. The results for this are given in the table below.

The table shows overwhelmingly that significantly more people accurately judged that the two values were the same for the grouped bars than for the stacked bars. This was the hypothesised result, and has presented to an even greater extent than previously anticipated. All but 7 of the respondents who responded to this question correctly judged from the grouped bars that the obstacle was used an equal number of times in each of the two rounds, whereas the responses for the grouped bar seemed fairly well split between the three options. It may be interesting in the multivariate analysis section to compare responses depending on whether respondents were shown the stacked or grouped bars first.

Perhaps a reason for the incorrect judging with the stacked is that the human brain works best when dealing with comparison in position than with length, by the '10 elementary tasks' idea put forward by Cleveland and McGill (1984), since comparing two height next to each other is a comparison in position as opposed to length, whereas the stacked bars are a length comparison.

Respondents were then asked *'Which obstacle do you think was used MORE in Finals (Regional/City) rounds, 'Log Grip' or 'Floating Steps'?'*Similar to the previous question, this asks for a comparison between the size of two categories, but this time about how many times two different obstacles were used in the round Finals (Regional/City), where these two obstacles are those discussed at the start of this part of the survey.

This was a potentially poorly formulated question, as the respondents had already been asked to specify how many times each of these obstacle was used in this round and respondents mostly judged this accurately with regard to both plots, but this could have been impacted by the previous questions. However, this does follow from the results from the past questions showing that respondents mostly accurately judged the values correctly, aside from those who instead judged the height of the whole bar.

The aim of the question *'Which bar chart do you feel is easiest to read and interpret?'* was to assess the perceived ease of interpretation of both bar plots. This is to gain an understanding in how data may best be presented in an easily understandable, easily readable manner. This is an important factor in visualisation, as a main aim in creating visuals is to provide an aid for the viewer to simply and quickly see the message. The opposite may be beneficial in certain applications however; based on the misreadings in the question regarding judging the number of times 'Log Grip' was used in the specific round, viewers of the visualisations could be easily mislead by incorrectly interpreting the plot. The people being shown the plot in, for example, an advert, may only take a fleeting look and not go beyond to analyse the plot to see accurate differences between values, and thus it is important to produce a plot that gives the easiest interpretation.

The large majority of participants found the grouped bar chart easier to read and interpret, as predicted.

Table 3.5: Number of respondents finding each of the two charts easier to read and interpret

Var1	Freq
Grouped	59
Stacked	11

The questions *'Which bar chart do you feel is easiest to read and interpret?'* and the one following *'Do you feel that one of the colour schemes makes it easier to read and interpret? If so, please select which one.'* are asked with the purpose of assessing the colour scheme that gives the greatest aesthetic pleasure, or effectively which colour palette the respondents feel is subjectively the ‘prettiest’ or ‘nicest’. It is important to note here that aesthetics and readability do not always go hand-in-hand; a plot that is made to look very aesthetically pleasing may sacrifice readability, and vice versa. For each of the two languages, six pairings of three different colour palettes were created, whereby the first colour was the one displayed for the main questions, and the second used only for the comparison questions. As previously discussed, the three colour schemes considered are viridis, greyscale, and each language’s default plotting colour palette. The colour palette pairings are outlined below, where each set of two colours is assigned a ‘Pairing ID’ from A to F.

Table 3.6: Colour pairings

Pairing ID	Main Palette	Secondary Pallette
A	Viridis	Default
B	Default	Viridis
C	Default	Greyscale
D	Greyscale	Default
E	Viridis	Greyscale
F	Greyscale	Viridis

Table 3.7: Easiest colour scheme to read and interpret

	A	B
Set A	7	6
Set B	6	6
Set C	9	1
Set D	3	9
Set E	11	0
Set F	1	11

This table shows that when it came to the default/viridis pairings, displayed in the first two rows, the respondents tended to have no preference overall. Comparing this to the bottom two rows, in which

Table 3.8: Easiest to read and interpret colour scheme, for R

	A	B
Set A	4	4
Set B	2	4
Set C	4	1
Set D	2	5
Set E	5	0
Set F	1	6

Table 3.9: Easiest to read and interpret colour scheme, for Python

	A	B
Set A	3	2
Set B	4	2
Set C	5	5
Set D	1	4
Set E	6	0
Set F	5	5

viridis is put against greyscale, only 1 respondent out of the 23, a proportion of 0.04, found the grey more aesthetically pleasing, as hypothesised. When considering greyscale/default, there was still a majority preferring the non-greyscale palette, but a higher proportion preferred this as compared to the viridis/greyscale, with 4 out of the 22, or a proportion of 0.18, preferring the grey. Overall, 35 preferred viridis, 30 the default, and 5 the greyscale.

As anticipated, the two more-colourful palettes are preferred aesthetically over the grey, and the viridis was preferred over the default.

Complementing the aesthetic preferences, the second question assesses the colour preference with regard to readability and ease of interpretation. As mentioned before, this will be used to test both the colour palette preference itself alongside whether this preference matches up with aesthetic preference.

Interestingly here, the top two rows appear to give slightly opposing results; the respondents who were presented with viridis for the main questions and the default as a secondary palette stated that they found either viridis easier to interpret or had no preference, whereas those presented with the default first and viridis second tended to find the default easier. This could perhaps be a result of the respondents becoming used to their primary colour scheme.

Once again looking at the comparisons with the greyscale, there were some respondents that found this easier to read, but the majority chose the alternative, whether this is viridis or the default.

Table 3.10: Aesthetic preference of colour schemes

	A	B	None
Set A	7	3	3
Set B	11	0	11
Set C	9	1	0
Set D	2	10	0
Set E	11	0	11
Set F	2	9	1

Table 3.11: Aesthetic preference of colour schemes, for R

	A	B	None
Set A	5	3	5
Set B	5	0	1
Set C	4	1	0
Set D	1	6	0
Set E	5	0	0
Set F	2	4	1

Table 3.12: Aesthetic preference of colour schemes, for Python

	A	B	None
Set A	2	3	2
Set B	6	0	0
Set C	5	0	0
Set D	1	4	0
Set E	6	0	0
Set F	0	5	0

The results seem fairly similar for the R and Python responses, showing that the default colourings for each language elicit a similar level of ease of interpretation.

The sample of respondents with colour blindness was too small to test this analysis.

3.4 Sales - Part 1

Now consider the sales part of the survey. In this section data was taken from a the `BJsales` data set in R, which is a time series data set containing 150 observations. This data set constitutes a single vector of values with no specified timings, and the visualisation data was formed by taking subsets of size 12 this and setting a month between each point to give a year of fictional sales data.

3.4.1 How much would you say sales of each company increased between January and December? [Company A]

This question was included for the purpose of testing whether, again, axis scaling impacts the perceived differences between values, but this time with time series line plots as opposed to bar plots. Respondents were asked to assess how much the sales of company A increased over the course of the year, or in other words to look at and compare each end of the line.

The plot for which the respondents, on average, found the difference to be smallest was the zeroed, followed by the truncated, and then the separated, with means of 1.371, 2.414 and 3.043 respectively. These differences are found to be statistically significant, as outlined in table[?].

Table 3.13: Table of p-values for this question

Alternatve Hypothesis	P-value
Truncated > Zeroed	8.870681966755e-14
Truncated < Separated	0.00654175643803223
Separated > Zeroed	3.48079934270661e-13

The differences between languages and plot ordering were shown to be inconsequential (see table [?])

3.4.2 How much would you say sales of each company increased between January and December? [Company B]

The zeroed was once again perceived to have the smallest difference ($\bar{x} = 1.371$), but this time with the separated in the middle ($\bar{x} = 4.1304$) and truncated with the largest difference ($\bar{x} = 4.1304$). See appendix 4, table [?] for p-values. The p-values show sufficient evidence that the truncated responses were on average greater than the zeroed, as were the responses for the separated plots. However, the difference between the ratings for the truncated and separated plot responses was inconsequential, along with the language comparisons and plot order.

Table 3.14: Table of p-values for this question

Hypothesis	P-value
Truncated > Zeroed	8.95254768631571e-23
Truncated not equal to Separated	0.2162
Separated not equal to Zeroed	12.46327564235365e-23

3.4.3 How large would you say the drop in sales between April and July of Company A is?

The means for this question appear very significantly different by eye, once again with the zeroed plot eliciting the lowest average rating ($\bar{x} = 1.371429$), followed by the truncated ($\bar{x} = 2.814286$) and then the separated ($\bar{x} = 4.028571$). The p-values confirm the significance of the differences between all three variables.

Table 3.15: Table of p-values for this question

Hypothesis	P-value
Truncated not equal to Zeroed	1.03832498155043e-11
Truncated not equal to Separated	0.00012743463393642
Separated not equal to Zeroed	1.1261341031207e-16

3.5 Sales - Part 2

3.5.1 Based on the above graph, how large would you say the difference is between the number of sales Company C makes and the number of sales Company D makes?

The final question of the survey compares just two plots, for which the difference in the ratings is shown to be significant, with the mean for the truncated plot ratings at $\bar{x} = 4.271$ and for the zeroed $\bar{x} = 2.7$ and a one-sided p-value of $p = 4.44089209850063e - 15$ showing the difference in the truncated was on average rated as larger than for the zeroed.

3.6 Conclusion

From this analysis, it can be concluded that altering axis scales in the way of truncating the axis or converting to a logarithmic scaling may have an effect on interpretation of differences in values, for both bar and line plots. The axis truncation has the effect of increasing the perceived difference in value and the logarithmic does the opposite. In general, from both literature, it is advised against to truncate the axis of a bar plot and this study confirms that it does in fact have an effect on interpretation. A logarithmic scaling may be ill-advised where it will distort the perceived size of the difference in point value, such as for the bar chart used here, however as discussed before from literature could be useful for other purposes, such as data that differs greatly in orders of magnitude. The labeling of this may also need to be considered, since the standard form labeling here confused some respondents.

Altering the aspect ratios had less of an effect, however there was a marginal effect of the wide plot making the difference in bar height appear smaller, and vice versa for the narrow, with the language not making a huge difference. This means that, while some consideration should be given to aspect ratios when re-scaling plots, it shouldn't have too much of an effect on interpretation.

It was found that grouped bar charts lead to a higher accuracy in interpretation of data values by that a stacked chart, and that the judgement of size difference is also more accurate for the grouped, along with ease of interpretation. Based on this as well as the literature, it appears grouped bars are mostly preferable to stacked.

Chapter 4

Appendices

4.1 Appendix 1 - The Survey

The following shows a pdf paper format of the first versions of the surveys, for R and Python.

4.1.1 R

An Empirical Study of Data Visualisation

Thank you for considering taking part in this study, it will take around 25 minutes to complete.

Information for participants:
This study, 'An Empirical Study of Data Visualisation' aims to explore the use and implementation of data visualisation as a tool for understanding data. The survey will ask you subjective questions regarding a series of data visualisations to assess how various factors may impact interpretation of the underlying data.

The survey involves reading charts and giving brief interpretations of them, in the form of multiple choice, ranking or a giving a single number. It is NOT designed to be treated as a test; all answers should be fully subjective and there are no 'correct' answers.

This survey may also investigate the impact of different visualisations on people with disorders in which visual and/or numerical processing may be inhibited. The purpose of this would be to investigate how to create visualisations more accessible to people with such disorders. However, if you feel distressed or overwhelmed by the style of questions presented, please do not hesitate to exit the survey at any point.

The data presented is not expected to involve controversial or distressing information;
- Section 1 will contain data relating to 'American Ninja Warrior'
- Section 2 will involve sales data for some fictional companies.

Important:
1) Participation is voluntary and you may exit the survey at any time. If you decide to cease participation your responses will not be recorded.

2) Post submission, your responses will be saved anonymously in a google sheet <https://www.google.co.uk/forms/about/> and used for analysis, which may be published.

3) If you have any questions please do not hesitate to contact me at murphyka1@cardiff.ac.uk.

*Required

1. If you have read and fully understood the above information and wish to voluntarily participate in the survey, please select 'I agree' below. *

Tick all that apply.

- [ ] I agree

(Optional)

Demographic Information

2. Please enter your age

____

3. If you are a university student or past university graduate please specify your area of study.

Mark only one oval.

- ( ) Science
- ( ) Technology
- ( ) Engineering
- ( ) Maths
- ( ) Arts
- ( ) Social Sciences
- ( ) Humanities
- ( ) Business
- ( ) N/A
- ( ) Other: ______________________________

How strongly do you agree with each of the three following statements?

4. I have good spatial awareness skills

Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	( )	( )	( )	( )	( )	Strongly Agree

5. I have good observational skills

Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	⬭	⬭	⬭	⬭	⬭	Strongly Agree

6. I have good numerical skills

Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	⬭	⬭	⬭	⬭	⬭	Strongly Agree

7. Are you colourblind?

Mark only one oval.

⬭ Yes

⬭ No

⬭ Prefer not to answer

⬭ Other: ____________________

8. Do you have any disorders that may affect visual processing? (this could be a general visual processing disorder or dyslexia, dyscalculia, ADHD etc)

Mark only one oval.

⬭ Yes

⬭ No

⬭ Prefer not to answer

⬭ Other: ____________________

American Ninja Warrior

The first sets of questions will show you a series of bar charts presenting information regarding how many times some obstacles were used over the course of 10 seasons of the TV show 'American Ninja Warrior'.

For each question you will be asked to look at a bar chart and answer the corresponding questions, where the height of each bar shows how many times the obstacle was used.

Please note that the answers to these questions are meant to be subjective, and that there are no correct answers.

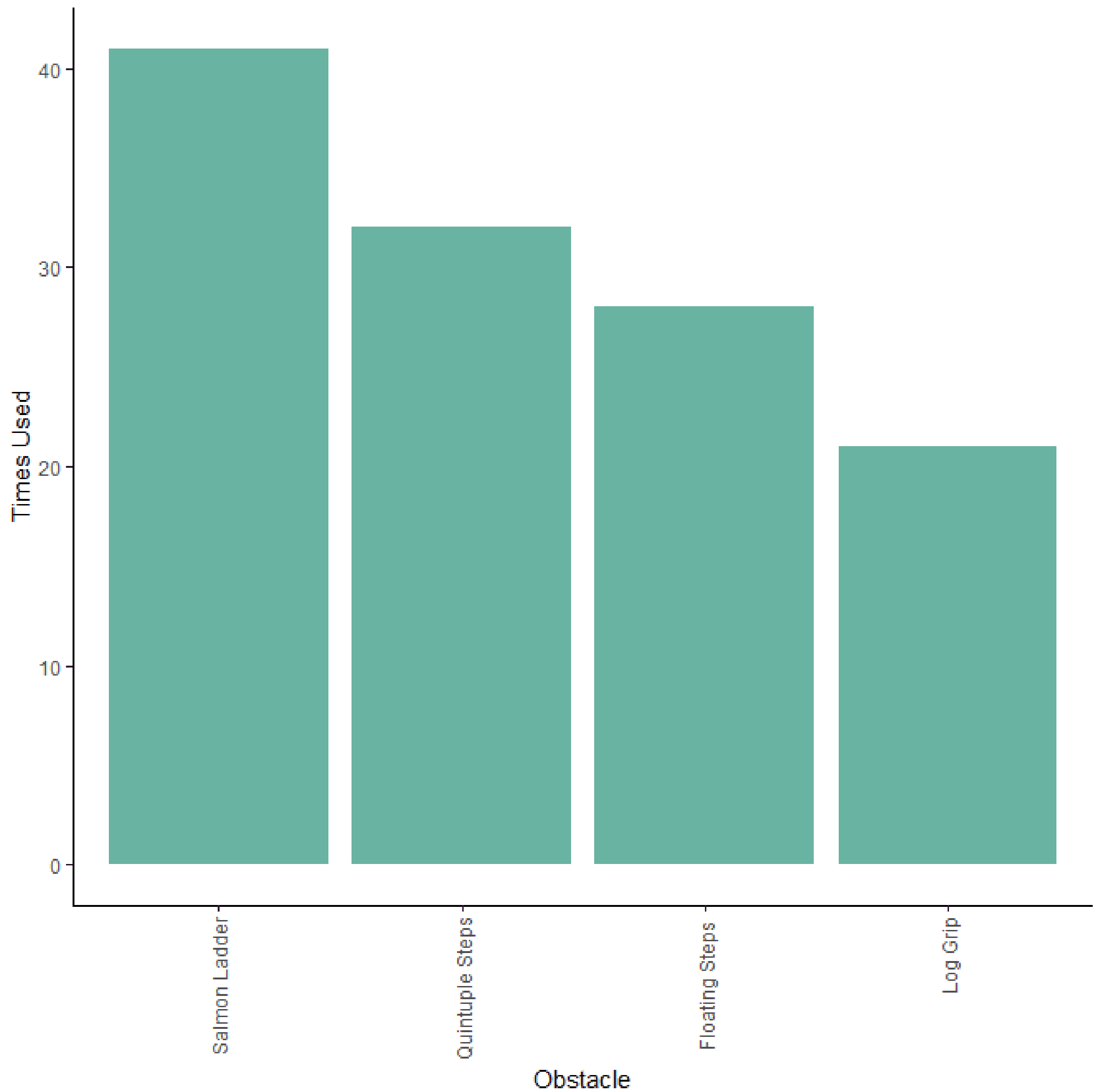
American Ninja Warrior - Part 1

The following bar charts present data for 4 of the most frequently used obstacles. Please look at the following charts and answer the corresponding questions.

American Ninja Warrior - Part 1, Question 1

Please refer to bar chart A, below.

Bar chart A



In relation to bar chart A:

9. Approximately many times would you say the 'Salmon Ladder' was used?

___

10. Approximately how much more than 'Log Grip' would you say 'Salmon Ladder' was used?

Mark only one oval.

	1	2	3	4	5	6	7	
Not much more	⬭	⬭	⬭	⬭	⬭	⬭	⬭	A lot more

11. Approximately how much more than 'Quintuple Steps' would you say 'Salmon Ladder' was used?

Mark only one oval.

	1	2	3	4	5	6	7	
Not much more	⬭	⬭	⬭	⬭	⬭	⬭	⬭	A lot more

12. In your opinion, approximately how many times would you say 'Log Grip' was used, as a percentage of the number of times 'Salmon Ladder' was used?

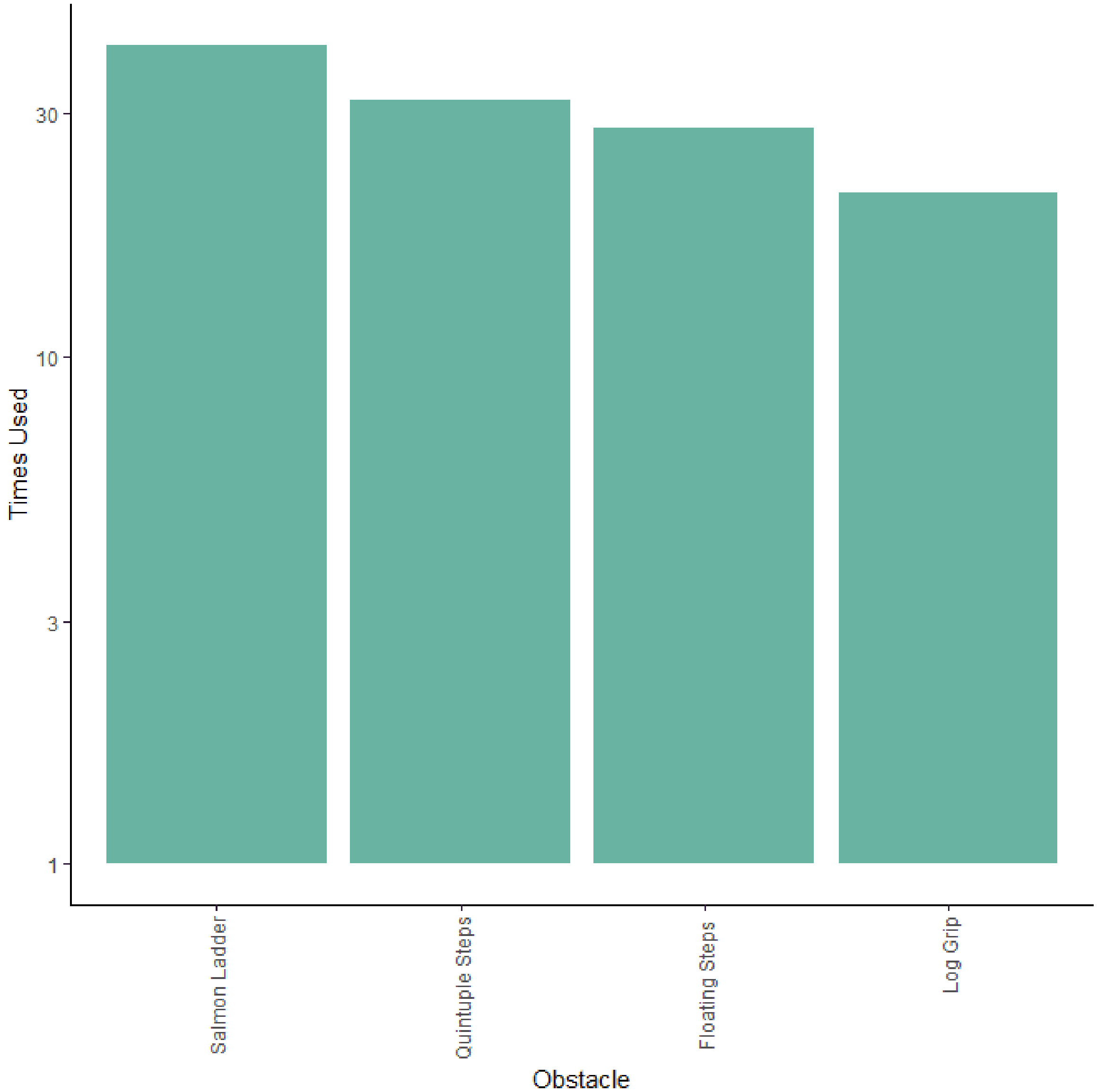
For example, if you think 'Log Grip' was used half as much as 'Salmon Ladder', enter 50%. If you think 'Log Grip' was used a quarter as much as 'Salmon Ladder', enter 25%, and so on.

____

American Ninja Warrior - Part 1, Question 2

Now consider the below bar chart, B.

Bar chart B



In relation to bar chart B:

13. Approximately many times would you say the 'Salmon Ladder' was used?

14. Approximately how much more than 'Log Grip' would you say 'Salmon Ladder' was used?

Mark only one oval.

	1	2	3	4	5	6	7	
Not much more	◯	◯	◯	◯	◯	◯	◯	A lot more

15. Approximately how much more than 'Quintuple Steps' would you say 'Salmon Ladder' was used?

Mark only one oval.

	1	2	3	4	5	6	7	
Not much more	◯	◯	◯	◯	◯	◯	◯	A lot more

16. In your opinion, approximately how many times would you say 'Log Grip' was used, as a percentage of the number of times 'Salmon Ladder' was used?

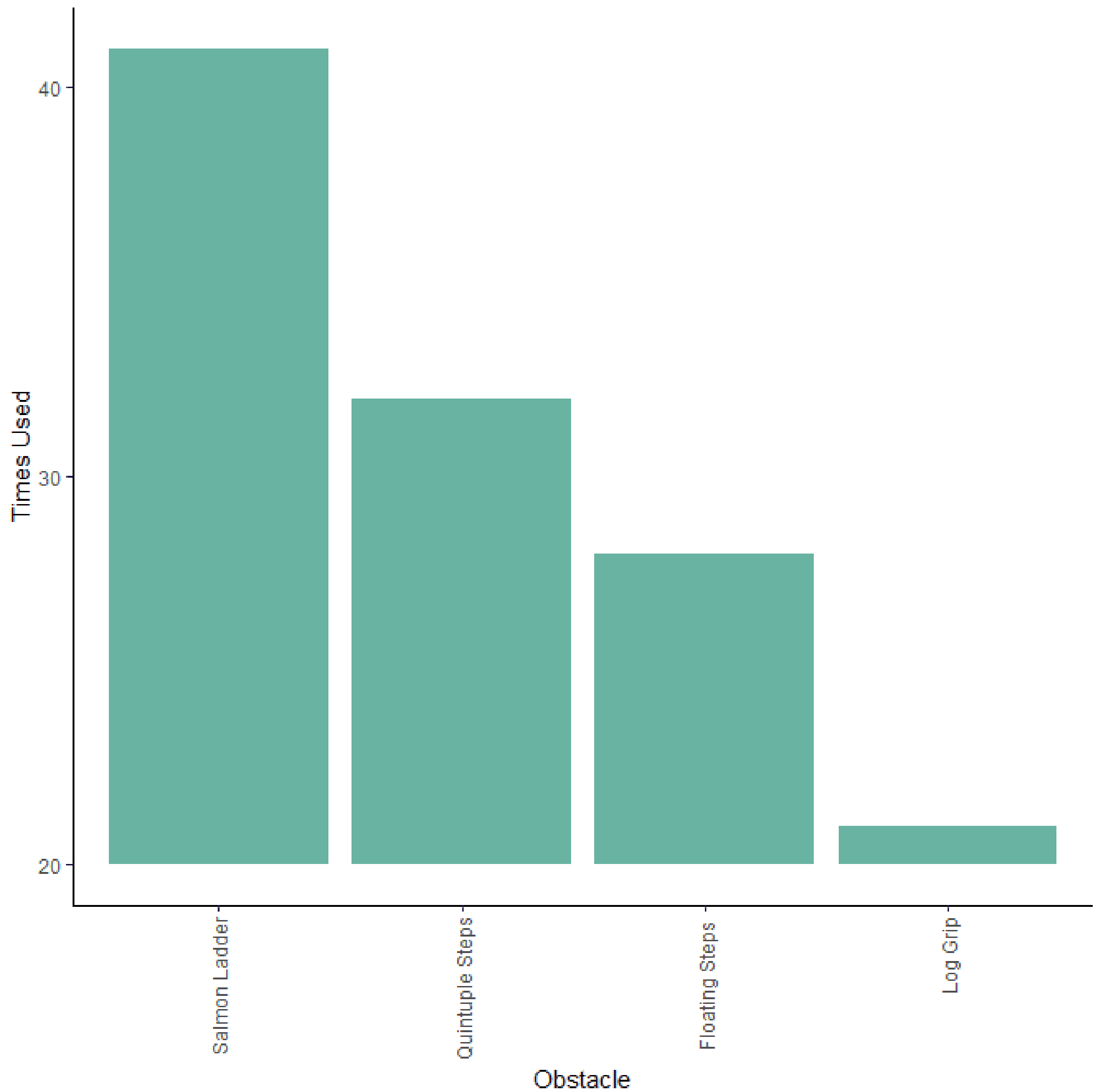
For example, if you think 'Log Grip' was used half as much as 'Salmon Ladder', enter 50%. If you think 'Log Grip' was used a quarter as much as 'Salmon Ladder', enter 25%, and so on.

____

American Ninja Warrior - Part 1, Question 3

Now consider bar chart C, below.

Bar chart C



In relation to bar chart C:

17. Approximately many times would you say the 'Salmon Ladder' was used?

____

18. Approximately how much more than 'Log Grip' would you say 'Salmon Ladder' was used?

Mark only one oval.

	1	2	3	4	5	6	7	
Not much more	⬭	⬭	⬭	⬭	⬭	⬭	⬭	A lot more

19. Approximately how much more than 'Quintuple Steps' would you say 'Salmon Ladder' was used?

Mark only one oval.

	1	2	3	4	5	6	7	
Not much more	⬭	⬭	⬭	⬭	⬭	⬭	⬭	A lot more

20. In your opinion, approximately how many times would you say 'Log Grip' was used, as a percentage of the number of times 'Salmon Ladder' was used?

For example, if you think 'Log Grip' was used half as much as 'Salmon Ladder', enter 50%. If you think 'Log Grip' was used a quarter as much as 'Salmon Ladder', enter 25%, and so on.

____

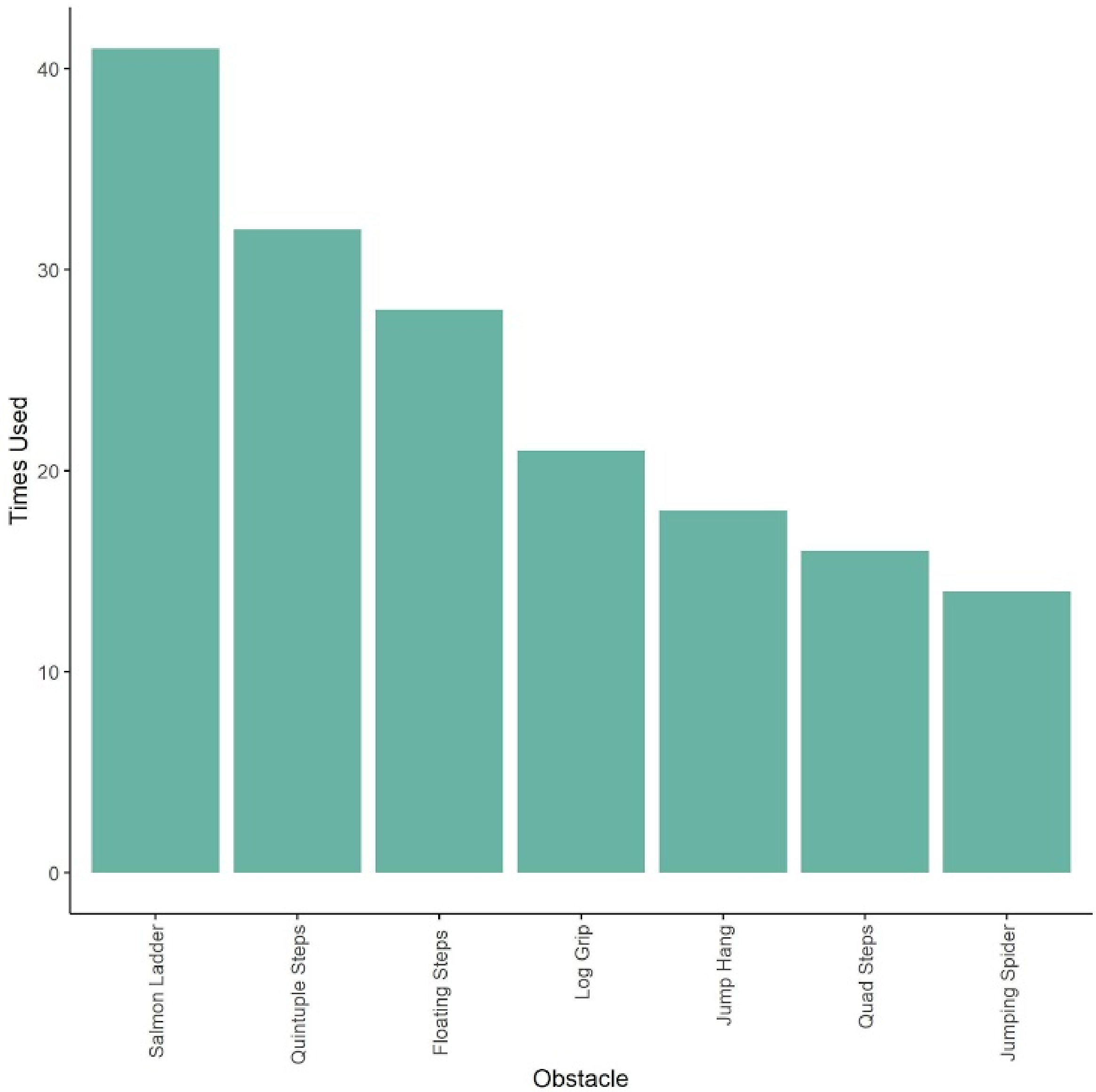
American Ninja Warrior - Part 2

You will now see bar charts regarding 7 of the most frequently used obstacles. Once again, for each question you will be asked to consider the bar chart and answer the corresponding questions. Again note that the answers to these questions are meant to be subjective, and that there are no correct answers.

American Ninja Warrior - Part 2, Question 1

Please see bar chart A, below.

Bar Chart A



In relation to bar chart A:

21. How large would you say the difference between 'Jumping spider' and 'Salmon Ladder' is?

Mark only one oval.

	1	2	3	4	5	6	7	
Very Small	⬭	⬭	⬭	⬭	⬭	⬭	⬭	Very Large

22. How large would you say the difference between 'Log Grip' and 'Floating Steps' is?

Mark only one oval.

	1	2	3	4	5	6	7	
Very Small	◯	◯	◯	◯	◯	◯	◯	Very Large

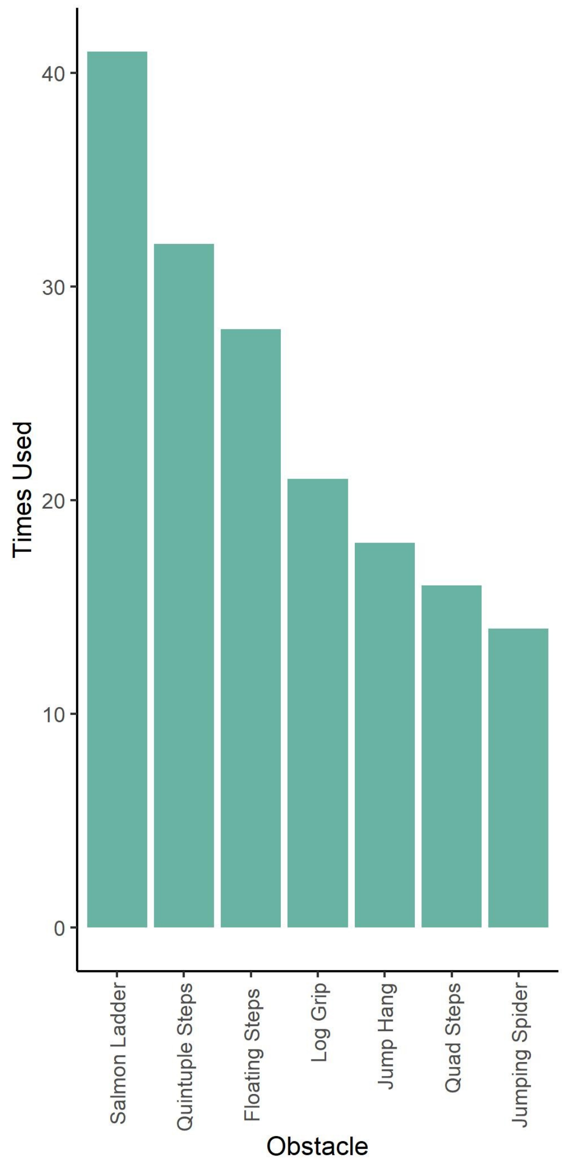
23. How many times would you say 'Floating Steps' were used?

____

American Ninja Warrior - Part 2, Question 2

Now please see bar chart B.

Bar Chart B



In relation to bar chart B:

24. How large would you say the difference between 'Jumping spider' and 'Salmon Ladder' is?

Mark only one oval.

	1	2	3	4	5	6	7	
Very Small	⬭	⬭	⬭	⬭	⬭	⬭	⬭	Very Large

25. How large would you say the difference between 'Log Grip' and 'Floating Steps' is?

Mark only one oval.

	1	2	3	4	5	6	7	
Very Small	◯	◯	◯	◯	◯	◯	◯	Very Large

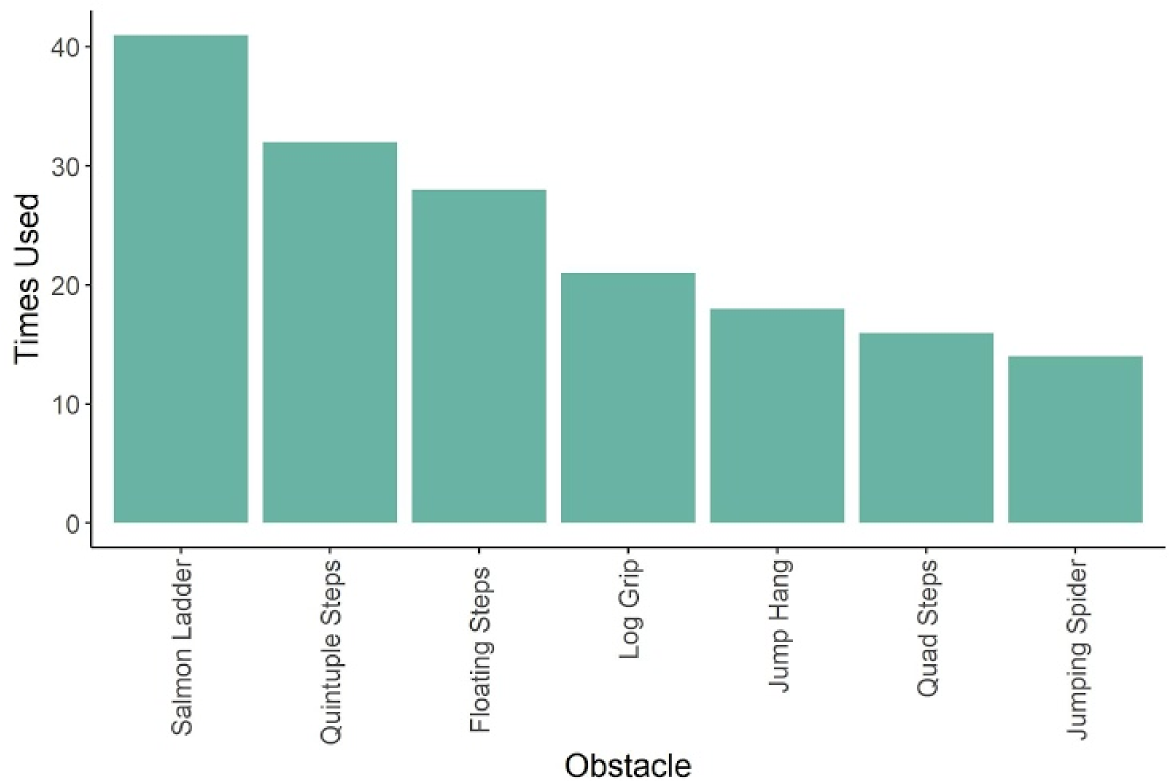
26. How many times would you say 'Floating Steps' were used?

____

American Ninja Warrior - Part 2, Question 3

Consider bar chart C, below.

Bar Chart C



In relation to bar chart C:

27. How large would you say the difference between 'Jumping spider' and 'Salmon Ladder' is?

Mark only one oval.

	1	2	3	4	5	6	7	
Very Small	◯	◯	◯	◯	◯	◯	◯	Very Large

28. How large would you say the difference between 'Log Grip' and 'Floating Steps' is?

Mark only one oval.

	1	2	3	4	5	6	7	
Very Small	⬭	⬭	⬭	⬭	⬭	⬭	⬭	Very Large

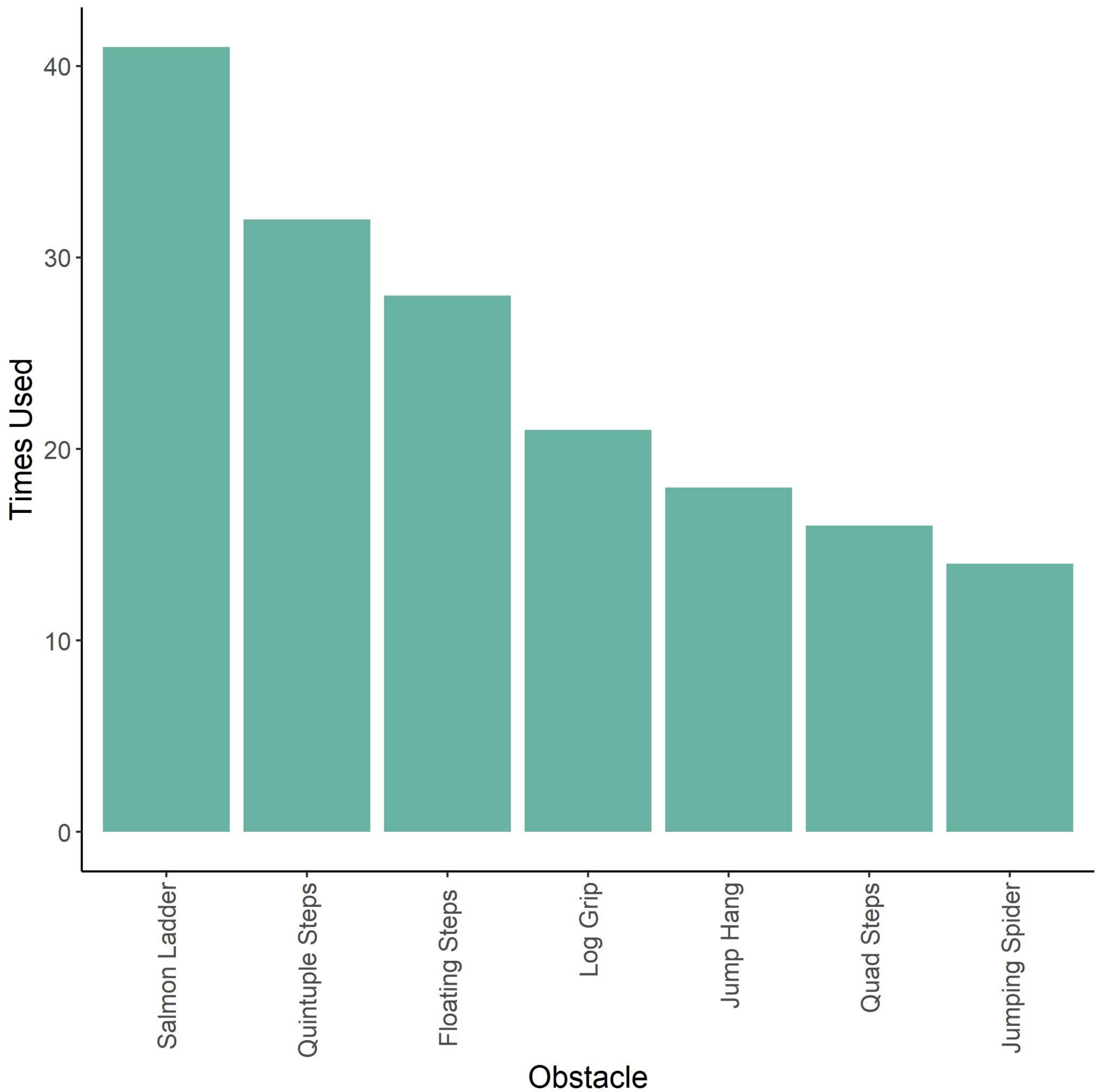
29. How many times would you say 'Floating Steps' were used?

____

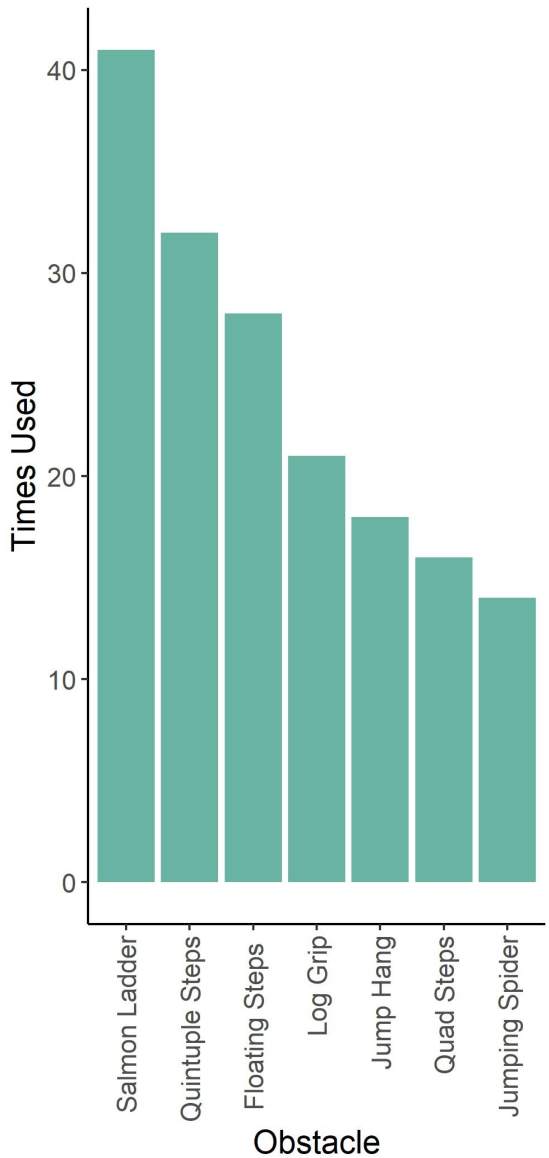
American Ninja Warrior - Part 2, Question 4

You will now see the bar charts A, B and C from part 2 again. Please answer the corresponding questions.

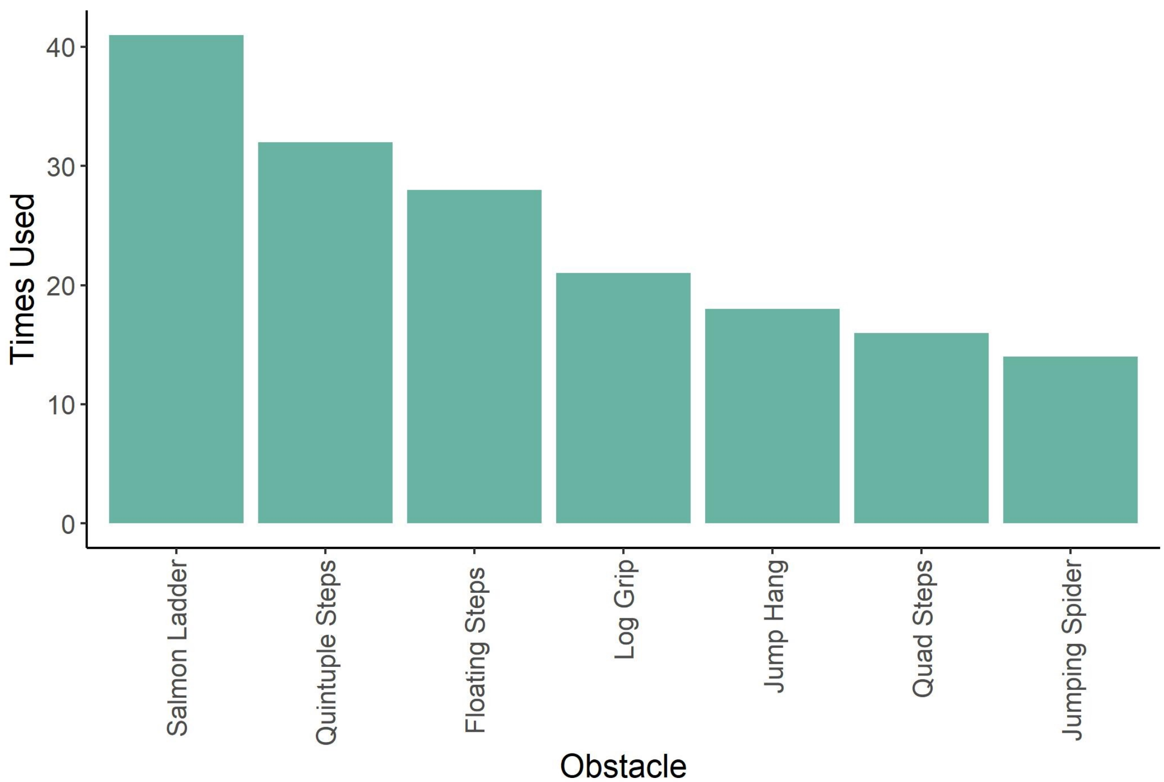
Bar chart A


40
30
20
10
0
Times Used
Salmon Ladder
Quintuple Steps
Floating Steps
Log Grip
Jump Hang
Quad Steps
Jumping Spider
Obstacle

Bar chart B


Times Used
40
30
20
10
0
Salmon Ladder
Quintuple Steps
Floating Steps
Log Grip
Jump Hang
Quad Steps
Jumping Spider
Obstacle

Bar chart C



30. Which of the three bar charts do you find most aesthetically pleasing?

Mark only one oval.

- A
- B
- C

31. Which bar chart do you feel is easiest to read and interpret?

This includes ease in reading labels, ease of reading the value of each bar, and ease in seeing the relative differences in values of the bars.

Mark only one oval.

- A
- B
- C

32. Which bar chart do you find hardest to read and interpret?

Mark only one oval.

- A
- B
- C

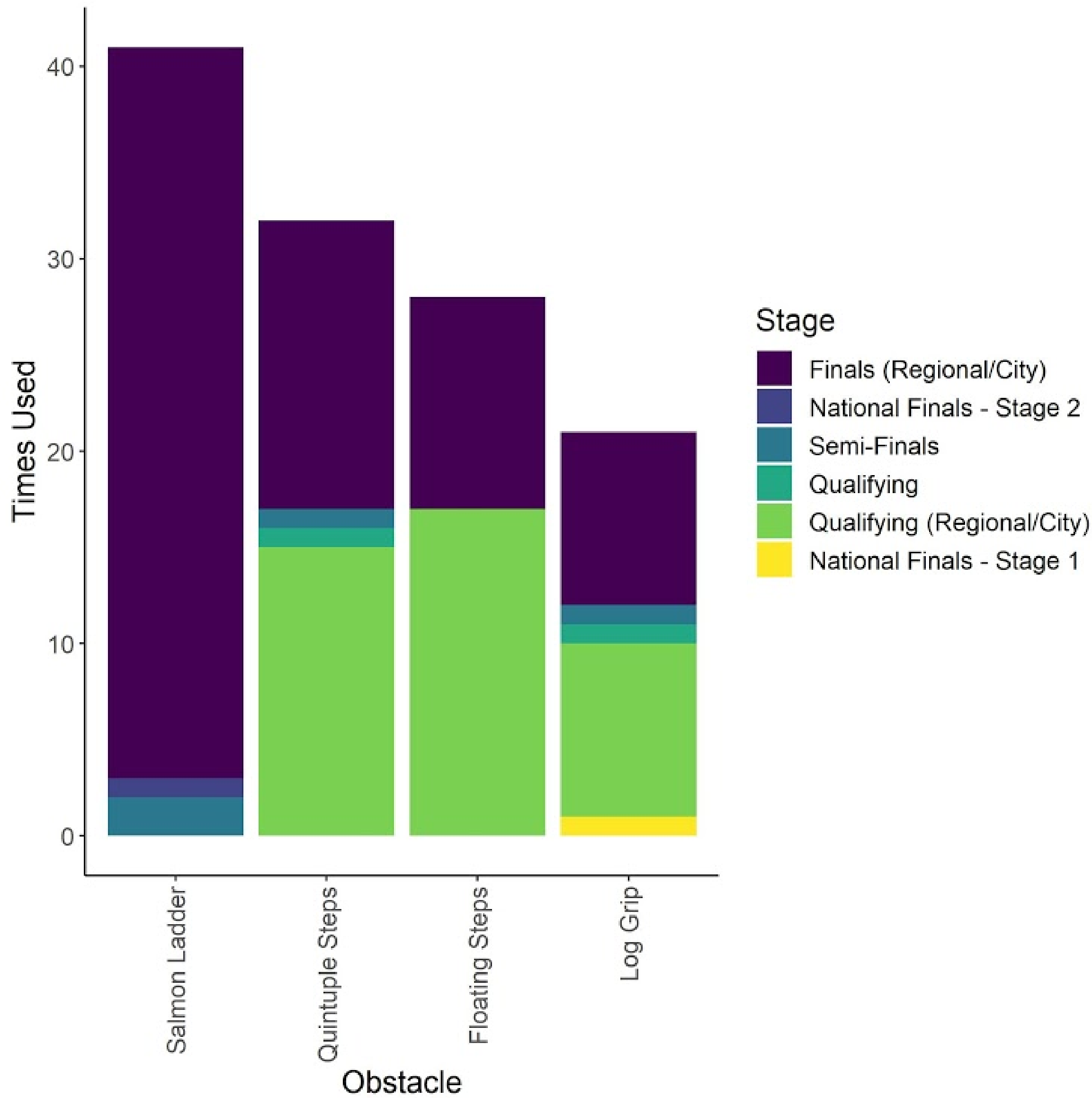
American Ninja Warrior - Part 3

The below plots show the same data as part 1, but are additionally coloured to show how many times each obstacle was used in each stage of the competition.

American Ninja Warrior - Part 3, Question 1

Please answer the questions below, for bar chart A.

Bar chart A



In relation to bar chart A:

33. How many times would you say 'Floating Steps' were used in the Finals (Regional/City) rounds?

34. How many times would you say 'Log Grip' was used in the Finals (Regional/City) rounds?

35. Please select the statement you feel applies to the bar chart above.

Mark only one oval.

- 'Log Grip' was used MORE in Finals (Regional/City) rounds than in Qualifying (Regional/City) rounds.
- 'Log Grip' was used LESS in Finals (Regional/City) rounds than in Qualifying (Regional/City) rounds.
- 'Log Grip' was used an EQUAL number of times in Finals (Regional/City) rounds and Qualifying (Regional/City) rounds.

36. Which obstacle do you think was used MORE in Finals (Regional/City) rounds, 'Log Grip' or 'Floating Steps'?

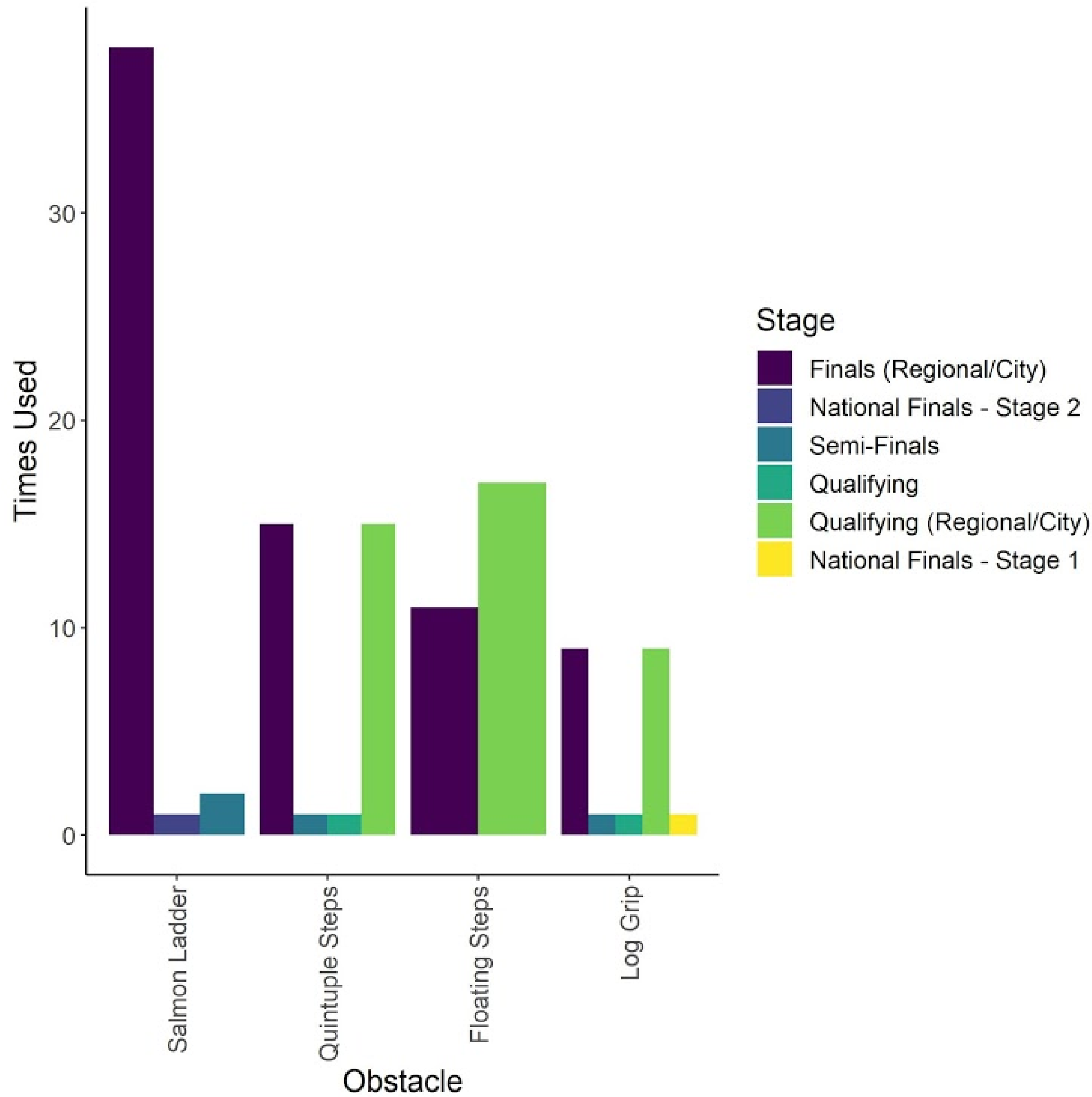
Mark only one oval.

- 'Log Grip'
- 'Floating Steps'
- They were used the same amount of times

American Ninja Warrior - Part 3, Question 2

Now please answer the questions again, for bar chart B.

Bar chart B



In relation to bar chart B:

37. How many times would you say 'Floating Steps' were used in the Finals (Regional/City) rounds?

38. How many times would you say 'Log Grip' was used in the Finals (Regional/City) rounds?

39. Please select the statement you feel applies to the bar chart above.

Mark only one oval.

- 'Log Grip' was used MORE in Finals (Regional/City) than in Qualifying (Regional/City).
- 'Log Grip' was used LESS in Finals (Regional/City) than in Qualifying (Regional/City).
- 'Log Grip' was used an EQUAL number of times in Finals (Regional/City) and Qualifying (Regional/City).

40. Which obstacle do you think was used MORE in Finals (Regional/City) rounds, 'Log Grip' or 'Floating Steps'?

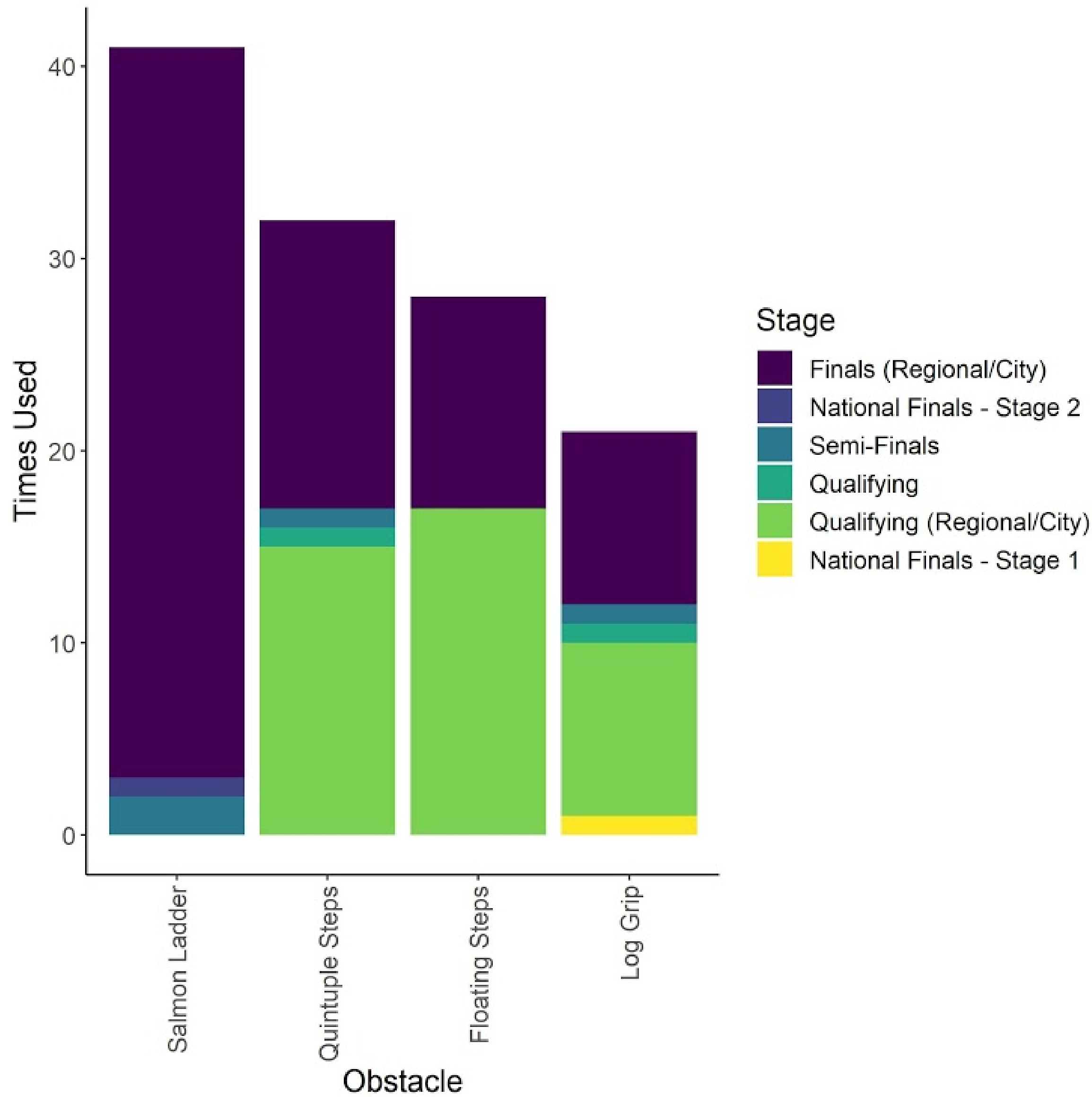
Mark only one oval.

- 'Log Grip'
- 'Floating Steps'
- They were used the same amount of times

American Ninja Warrior - Part 3, Question 3

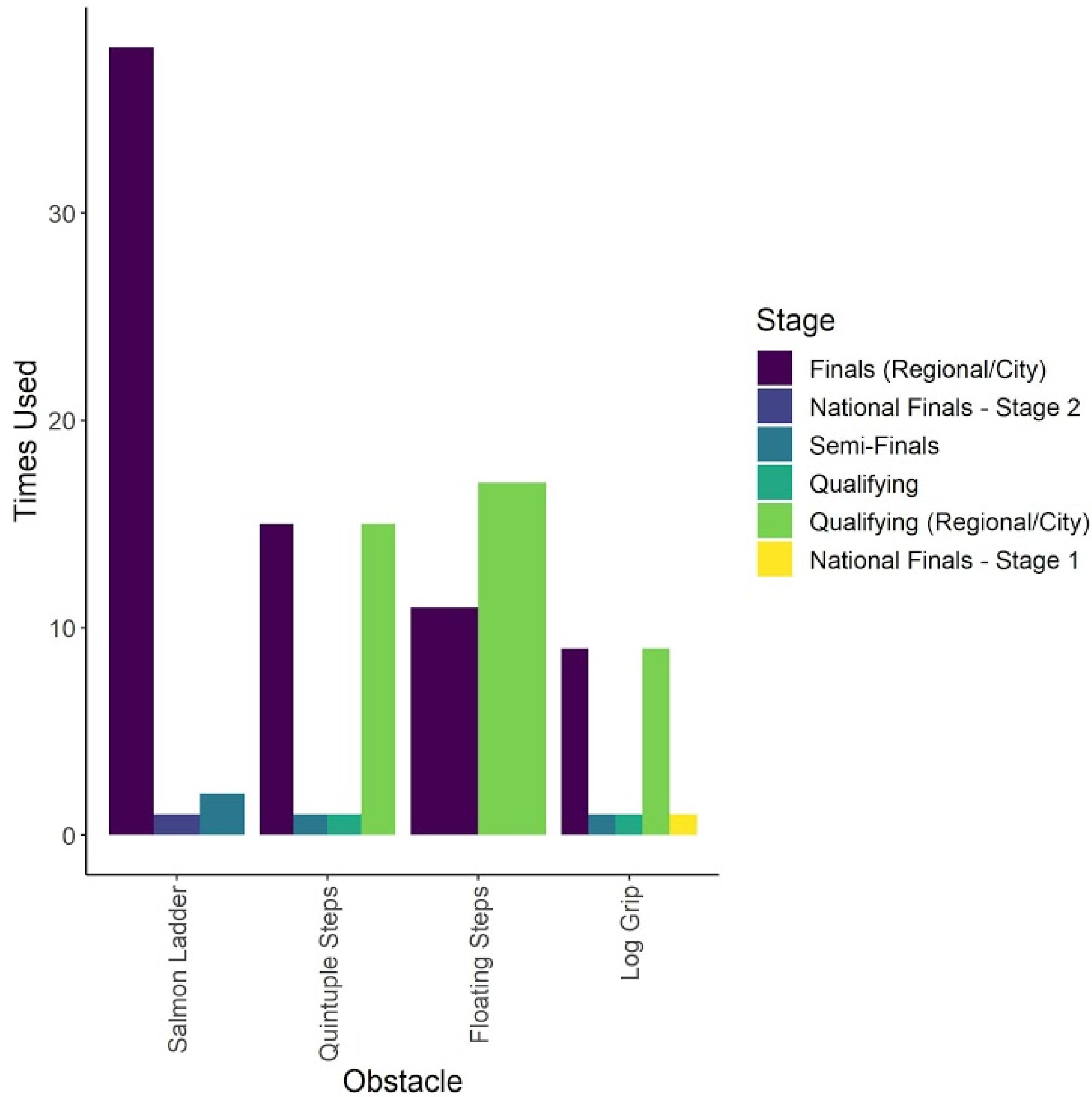
You will now see the two bar charts again. Please answer the question below.

Bar Chart A



Times Used
0
10
20
30
40
Salmon Ladder
Quintuple Steps
Floating Steps
Log Grip
Obstacle
Stage
Finals (Regional/City)
National Finals - Stage 2
Semi-Finals
Qualifying
Qualifying (Regional/City)
National Finals - Stage 1

Bar chart B



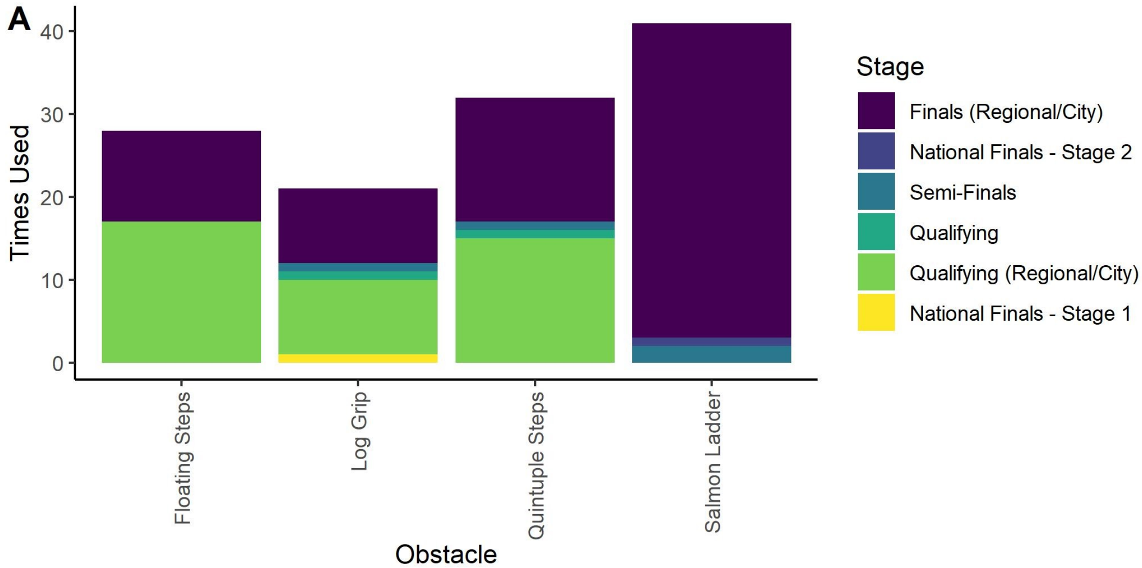
41. Which bar chart do you feel is easiest to read and interpret?

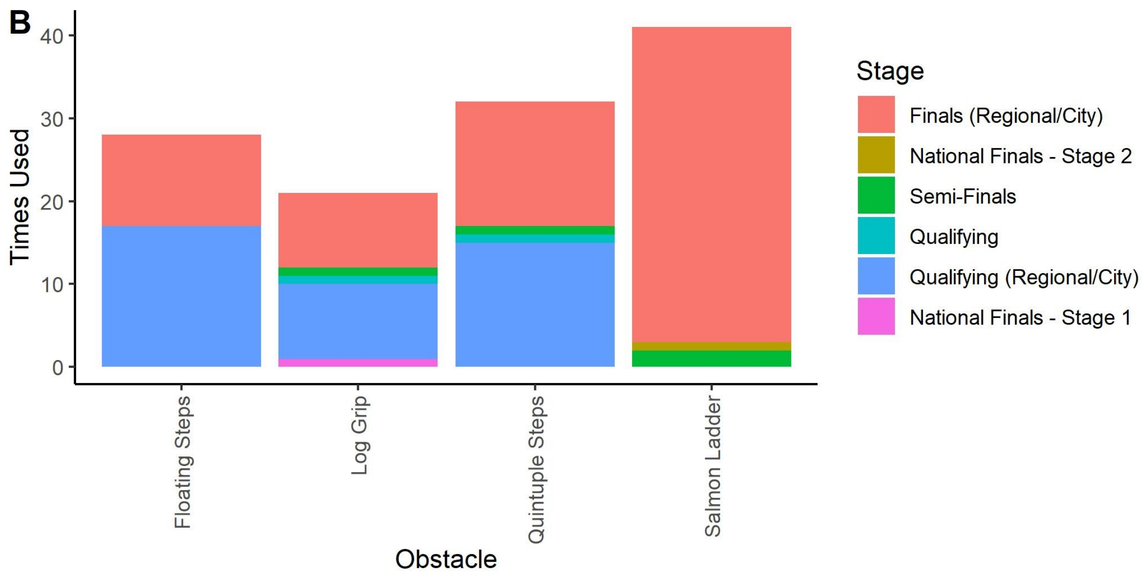
Mark only one oval.

- A
- B

American Ninja Warrior - Part 3, Question 4

Below compares the color scheme of the charts above, colour scheme A with another colour scheme, B.





42. Which colour scheme do you find most aesthetically pleasing?

Mark only one oval.

- ◯ A
- ◯ B

43. Do you feel that one of the colour schemes makes it easier to read and interpret the data than the other? If so, please select which one.

Mark only one oval.

- No
- Yes, A is easier
- Yes, B is easier

Sales

Thank you for completing the first section of this survey. The second, and final, section is shorter and will show you a series of line graphs relating to monthly sales data from four companies over the course of a year.

Once again, please refer to the following charts and answer the questions.

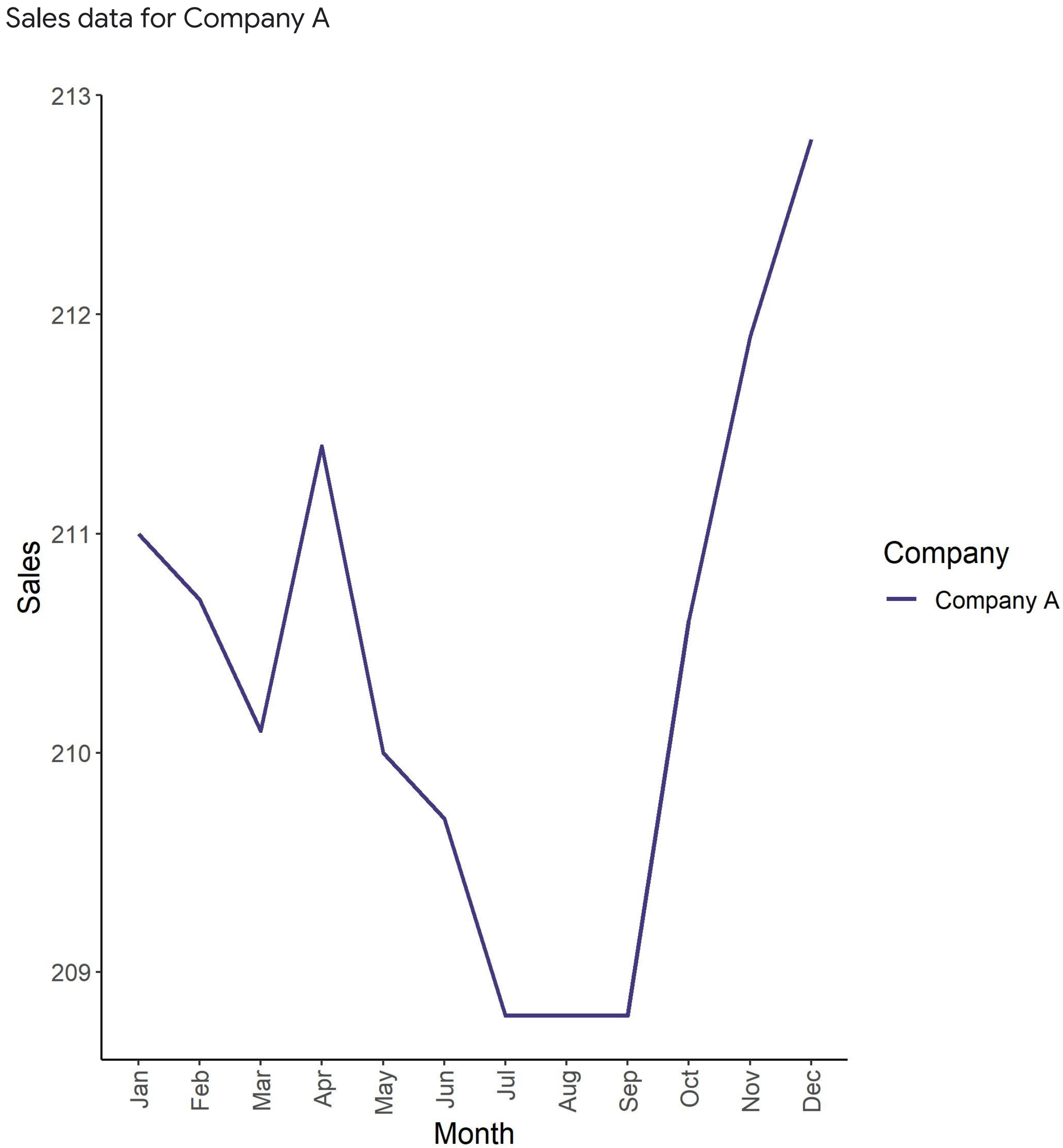
Please note that the answers to these questions are meant to be subjective, and that there are no correct answers.

Sales - Part 1

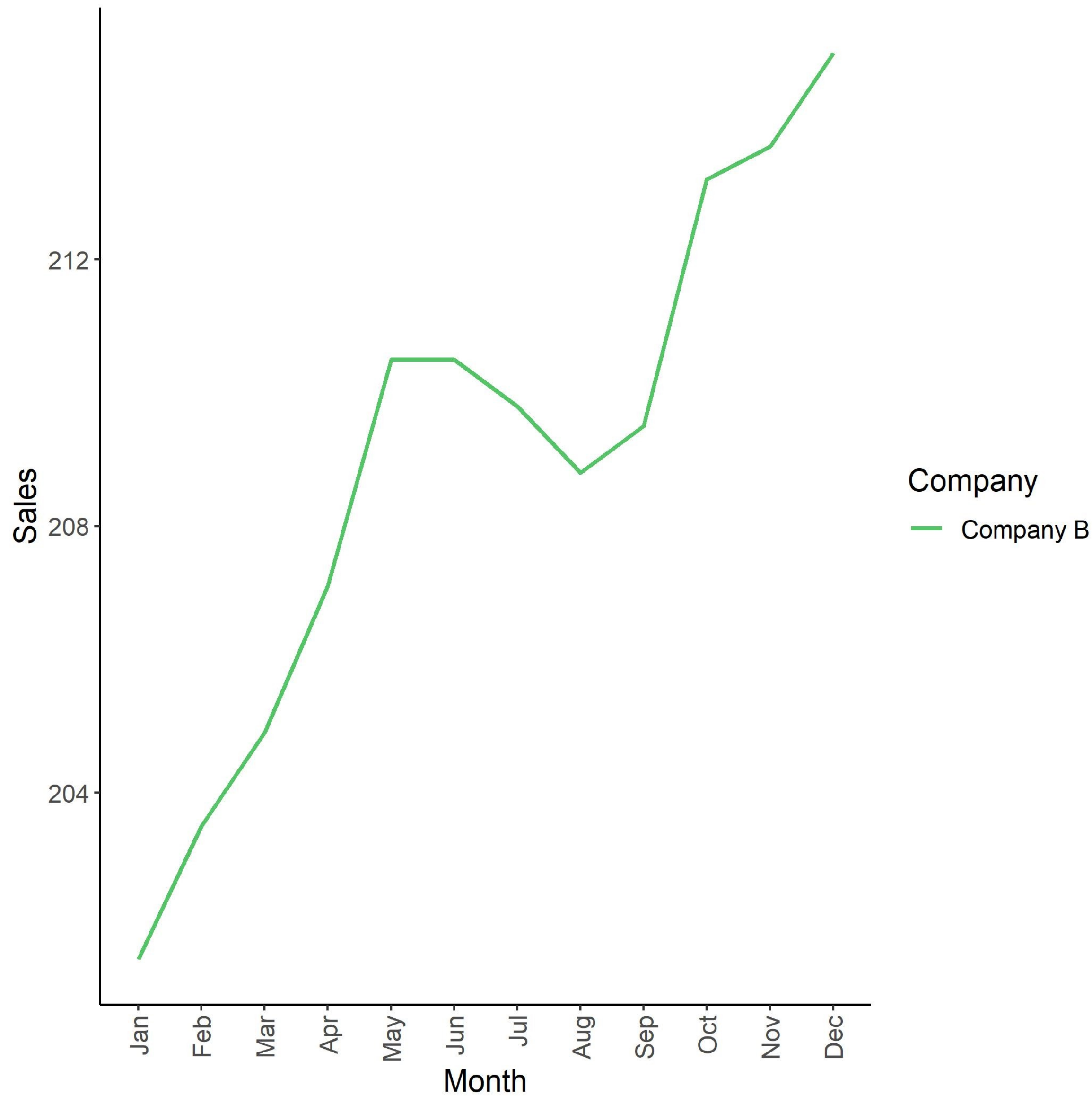
The following section will present graphs showing monthly sales data of two competing companies over the course of a year, and ask you questions relating to the graphs.

Sales - Part 1, Question 1

Please refer to the two charts below


Sales data for Company A
213
212
211
210
209
Sales
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
Company
Company A

Sales data for Company B



44. How much would you say sales of each company increased between January and December?

Mark only one oval per row.

	1 (A little)	2	3	4	5	6	7 (A lot)
Company A	◯	◯	◯	◯	◯	◯	◯
Company B	◯	◯	◯	◯	◯	◯	◯

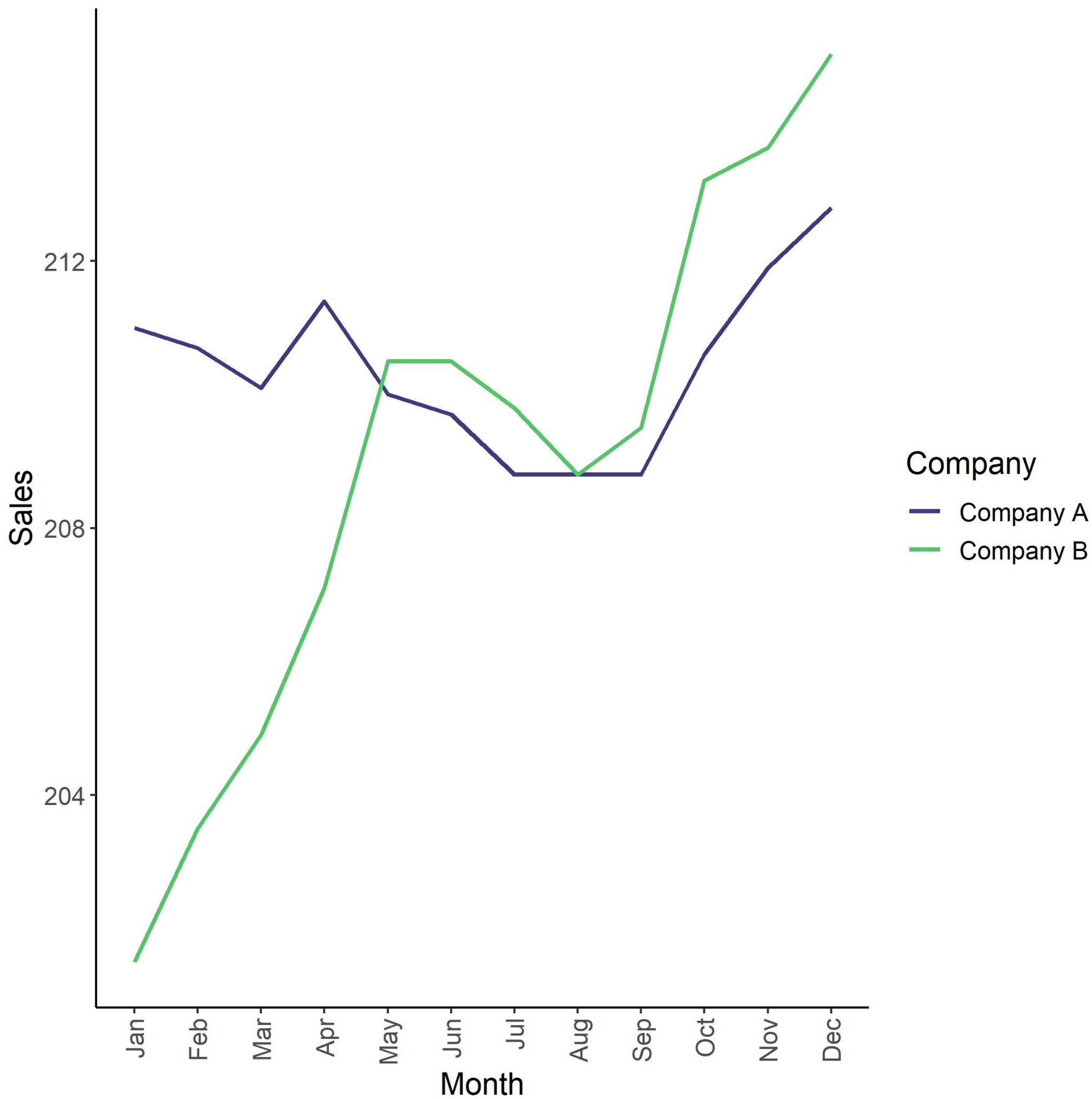
45. How large would you say the drop in sales between April and July of Company A is?

Mark only one oval.

	1	2	3	4	5	6	7	
A little	⬭	⬭	⬭	⬭	⬭	⬭	⬭	A lot

Sales - Part 1, Question 2

Consider the chart below



46. How much would you say sales of each company increased between January and December?

Mark only one oval per row.

	1 (A little)	2	3	4	5	6	7 (A lot)
Company A	⬭	⬭	⬭	⬭	⬭	⬭	⬭
Company B	⬭	⬭	⬭	⬭	⬭	⬭	⬭

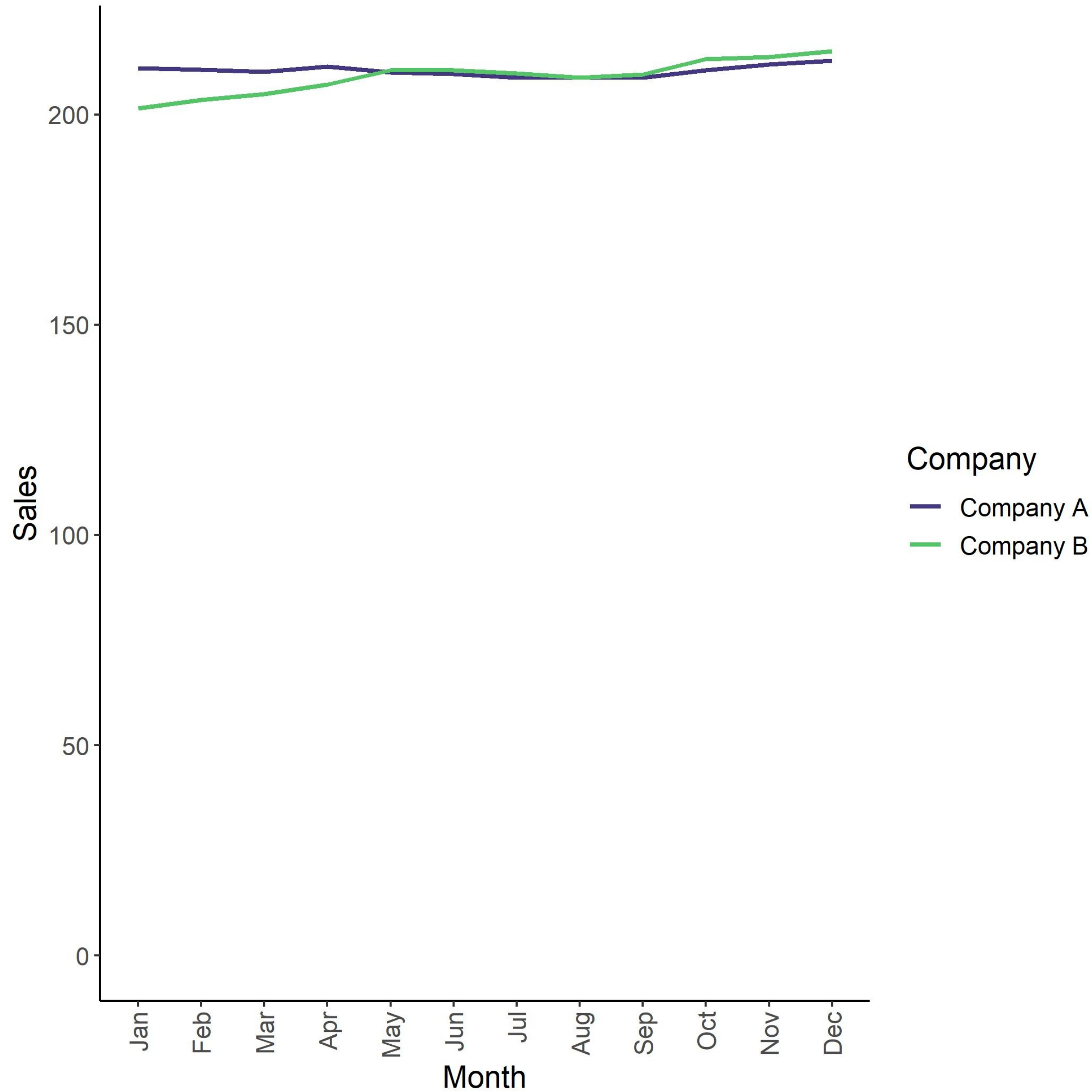
47. How large would you say the drop in sales between April and July of Company A is?

Mark only one oval.

	1	2	3	4	5	6	7	
A little	⬭	⬭	⬭	⬭	⬭	⬭	⬭	A lot

Sales - Part 1, Question 3

Consider the chart below



48. How much would you say sales of each company increased between January and December?

Mark only one oval per row.

	1 (A little)	2	3	4	5	6	7 (A lot)
Company A							
Company B							

49. How large would you say the drop in sales between April and July of Company A is?

Mark only one oval.

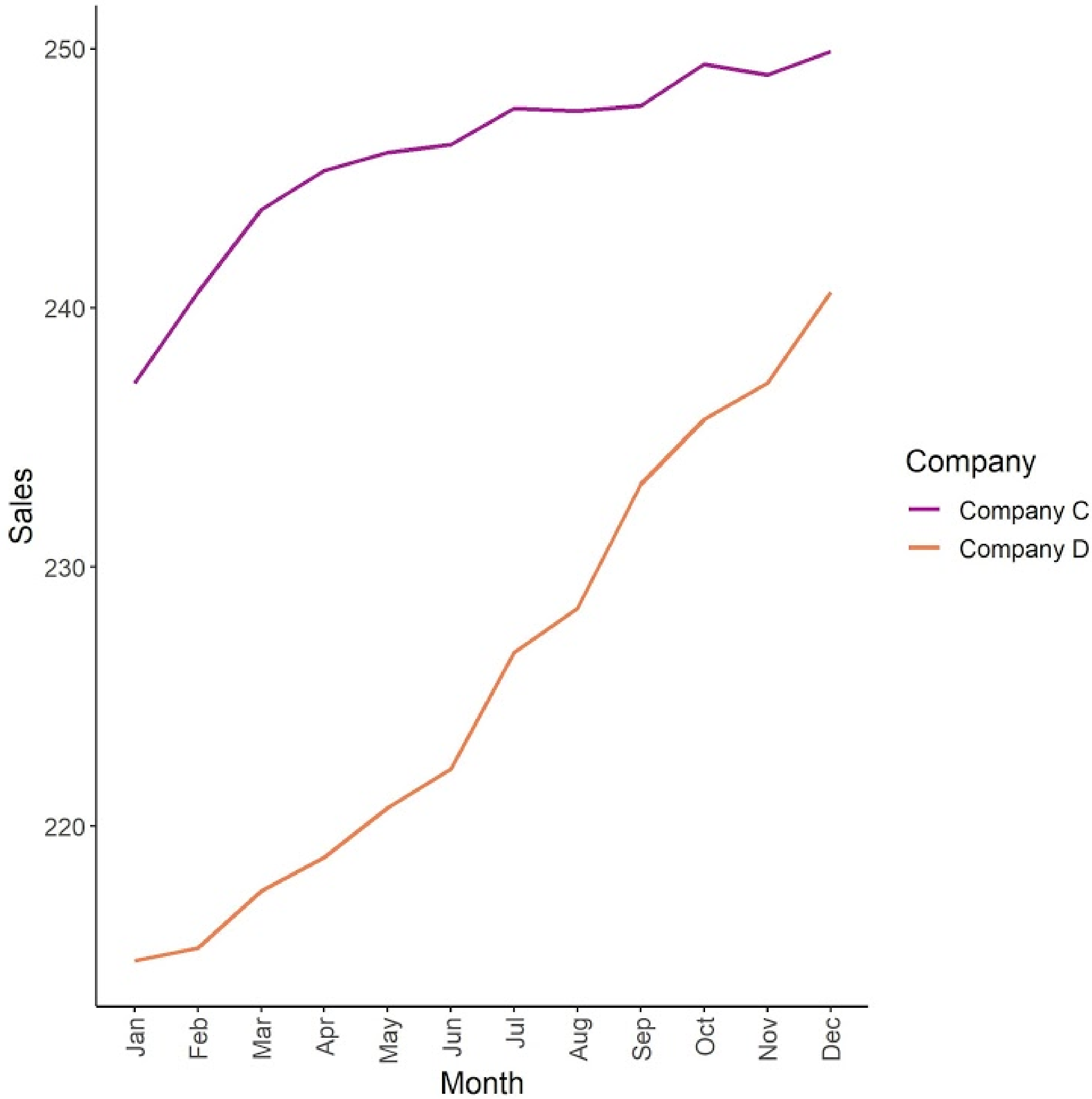
	1	2	3	4	5	6	7	
A little	⬭	⬭	⬭	⬭	⬭	⬭	⬭	A lot

Sales - Part 2

The graphs below show the sales data for two more companies, Company C and Company D.

Sales - Part 2, Question 1

Consider the chart below



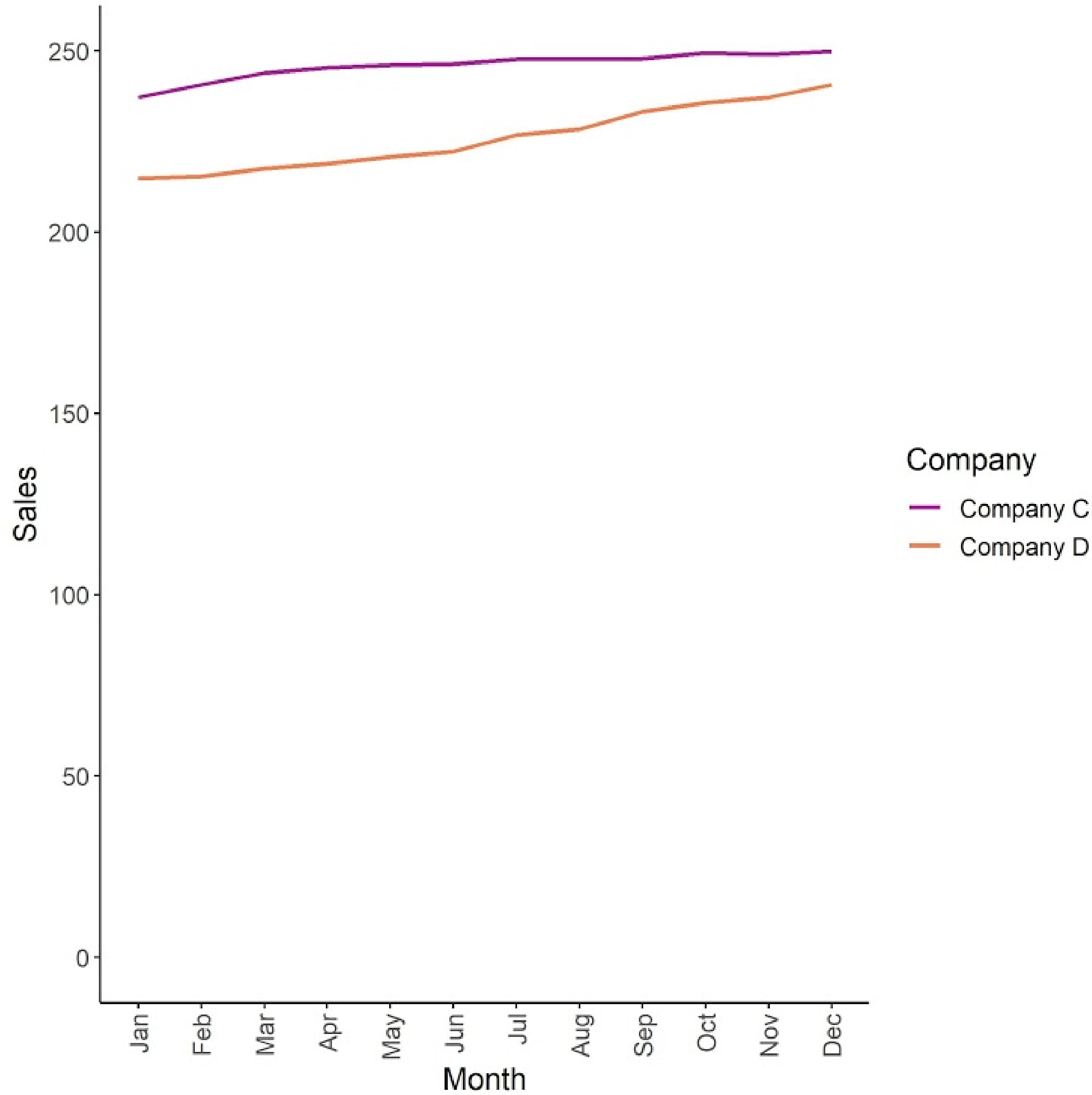
50. Based on the above graph, how large would you say the difference is between the number of sales Company C makes and the number of sales Company D makes?

Mark only one oval.

	1	2	3	4	5	6	7	
A little	◯	◯	◯	◯	◯	◯	◯	A lot

Sales - Part 2, Question 2

Consider the chart below



51. Based on the above graph, how large would you say the difference is between the number of sales Company C makes and the number of sales Company D makes?

Mark only one oval.

	1	2	3	4	5	6	7	
A little	⬭	⬭	⬭	⬭	⬭	⬭	⬭	A lot

Thank you for completing this survey.

Reminder - Your responses to this survey will remain anonymous.

If you would like to know more about the use of data visualisation and how different techniques may impact interpretation, please see the links below.
- https://searchbusinessanalytics.techtarget.com/definition/data-visualization
- https://datajournalism.com/read/handbook/one/understanding-data/using-data-visualization-to-find-insights-in-data
- https://heap.io/blog/data-stories/how-to-lie-with-data-visualization
- https://www.idashboards.com/blog/2019/02/27/when-data-visualizations-mislead-and-how-to-prevent-it/

If you are interested in how colour blindness can affect reading of visualisations and colour blind friendly palettes to use for data visualisation, please see the links below:
- https://blog.datawrapper.de/colorblindness-part3/
- https://www.youtube.com/watch?v=xAoljeRJ3lU

Once again, if you have any questions please do not hesitate to contact me at murphyka1@cardiff.ac.uk.

4.1.2 Python

An Empirical Study of Data Visualisation

Thank you for considering taking part in this study, it will take around 25 minutes to complete.

Information for participants:
This study, 'An Empirical Study of Data Visualisation' aims to explore the use and implementation of data visualisation as a tool for understanding data. The survey will ask you subjective questions regarding a series of data visualisations to assess how various factors may impact interpretation of the underlying data.

The survey involves reading charts and giving brief interpretations of them, in the form of multiple choice, ranking or a giving a single number. It is NOT designed to be treated as a test; all answers should be fully subjective and there are no 'correct' answers.

This survey may also investigate the impact of different visualisations on people with disorders in which visual and/or numerical processing may be inhibited. The purpose of this would be to investigate how to create visualisations more accessible to people with such disorders. However, if you feel distressed or overwhelmed by the style of questions presented, please do not hesitate to exit the survey at any point.

The data presented is not expected to involve controversial or distressing information;
- Section 1 will contain data relating to 'American Ninja Warrior'
- Section 2 will involve sales data for some fictional companies.

Important:
1) Participation is voluntary and you may exit the survey at any time. If you decide to cease participation your responses will not be recorded.

2) Post submission, your responses will be saved anonymously in a google sheet <https://www.google.co.uk/forms/about/> and used for analysis, which may be published.

3) If you have any questions please do not hesitate to contact me at murphyka1@cardiff.ac.uk.

*Required

1. If you have read and fully understood the above information and wish to voluntarily participate in the survey, please select 'I agree' below. *

Tick all that apply.

☐ I agree

Demographic Information

(Optional)

2. Please enter your age

____

3. If you are a university student or past university graduate please specify your area of study.

Mark only one oval.

- ( ) Science
- ( ) Technology
- ( ) Engineering
- ( ) Maths
- ( ) Arts
- ( ) Social Sciences
- ( ) Humanities
- ( ) Business
- ( ) N/A
- ( ) Other: ____________________

How strongly do you agree with each of the three following statements?

4. I have good spatial awareness skills

Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	( )	( )	( )	( )	( )	Strongly Agree

5. I have good observational skills

Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

6. I have good numerical skills

Mark only one oval.

	1	2	3	4	5	
Strongly Disagree	◯	◯	◯	◯	◯	Strongly Agree

7. Are you colourblind?

Mark only one oval.

◯ Yes
◯ No
◯ Prefer not to answer
◯ Other: ________________

8. Do you have any disorders that may affect visual processing? (this could be a general visual processing disorder or dyslexia, dyscalculia, ADHD etc)

Mark only one oval.

◯ Yes
◯ No
◯ Prefer not to answer
◯ Other: ________________

American Ninja Warrior

The first sets of questions will show you a series of bar charts presenting information regarding how many times some obstacles were used over the course of 10 seasons of the TV show 'American Ninja Warrior'.

For each question you will be asked to look at a bar chart and answer the corresponding questions, where the height of each bar shows how many times the obstacle was used.

Please note that the answers to these questions are meant to be subjective, and that there are no correct answers.

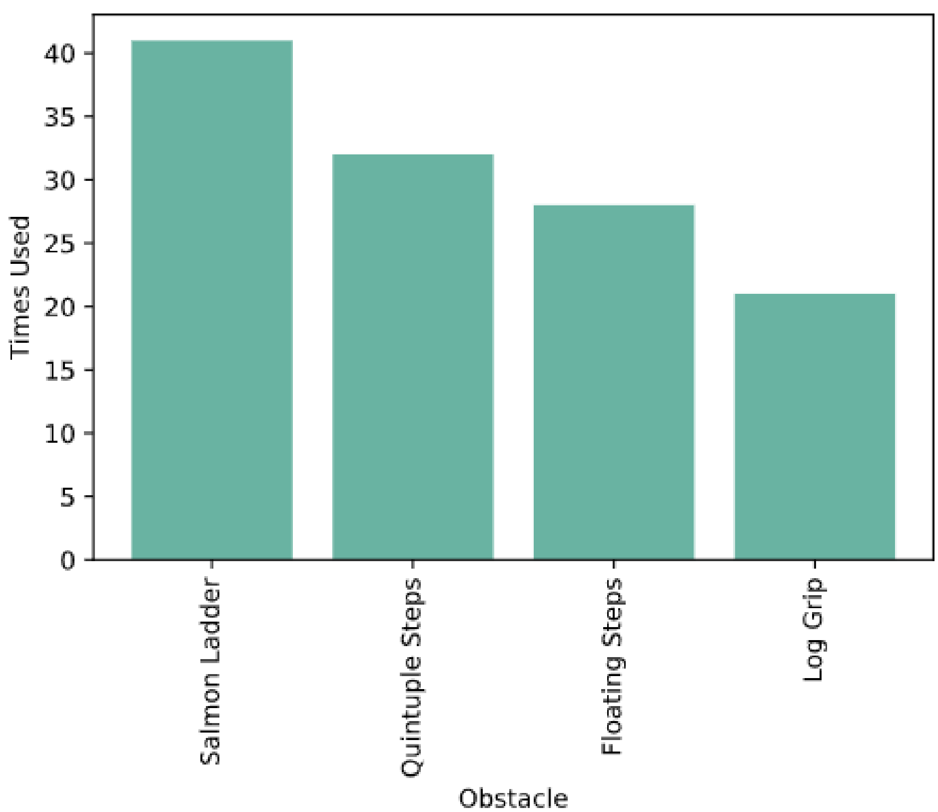
American Ninja Warrior - Part 1

The following bar charts present data for 4 of the most frequently used obstacles. Please look at the following charts and answer the corresponding questions.

American Ninja Warrior - Part 1, Question 1

Please refer to bar chart A, below.

Bar chart A



In relation to bar chart A:

9. Approximately many times would you say the 'Salmon Ladder' was used?

____

10. Approximately how much more than 'Log Grip' would you say 'Salmon Ladder' was used?

Mark only one oval.

	1	2	3	4	5	6	7	
Not much more	◯	◯	◯	◯	◯	◯	◯	A lot more

11. Approximately how much more than 'Quintuple Steps' would you say 'Salmon Ladder' was used?

Mark only one oval.

	1	2	3	4	5	6	7	
Not much more	◯	◯	◯	◯	◯	◯	◯	A lot more

12. In your opinion, approximately how many times would you say 'Log Grip' was used, as a percentage of the number of times 'Salmon Ladder' was used?

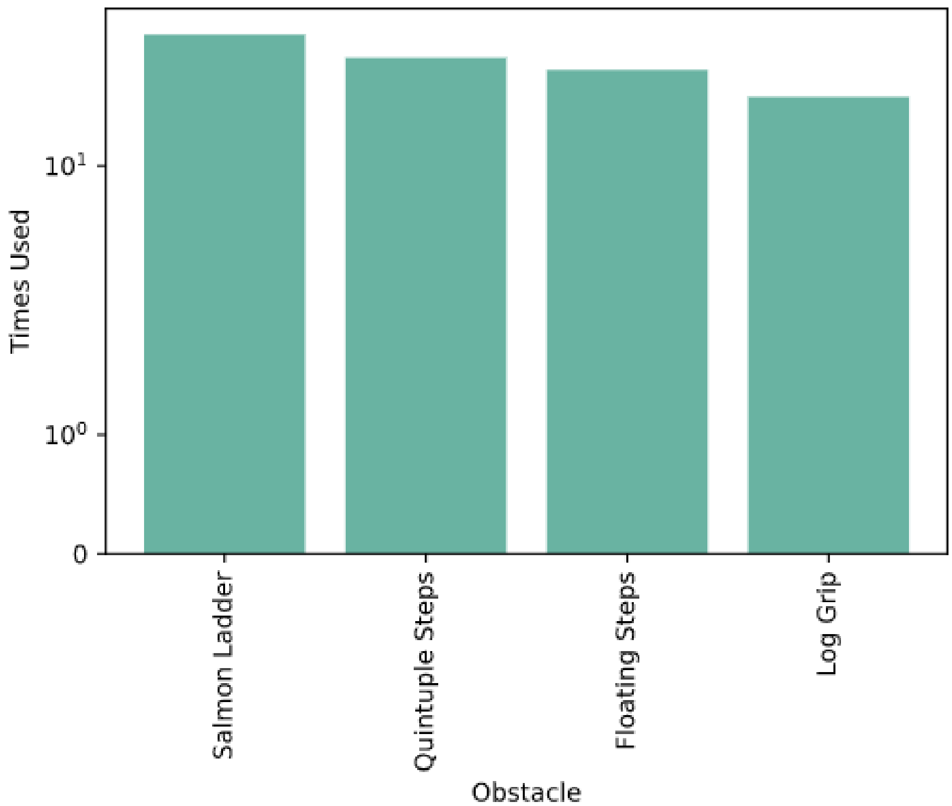
For example, if you think 'Log Grip' was used half as much as 'Salmon Ladder', enter 50%. If you think 'Log Grip' was used a quarter as much as 'Salmon Ladder', enter 25%, and so on.

____

American Ninja Warrior - Part 1, Question 2

Now consider the below bar chart, B.

Bar chart B



In relation to bar chart B:

13. Approximately many times would you say the 'Salmon Ladder' was used?

____

14. Approximately how much more than 'Log Grip' would you say 'Salmon Ladder' was used?

Mark only one oval.

	1	2	3	4	5	6	7	
Not much more	◯	◯	◯	◯	◯	◯	◯	A lot more

15. Approximately how much more than 'Quintuple Steps' would you say 'Salmon Ladder' was used?

Mark only one oval.

	1	2	3	4	5	6	7	
Not much more	◯	◯	◯	◯	◯	◯	◯	A lot more

16. In your opinion, approximately how many times would you say 'Log Grip' was used, as a percentage of the number of times 'Salmon Ladder' was used?

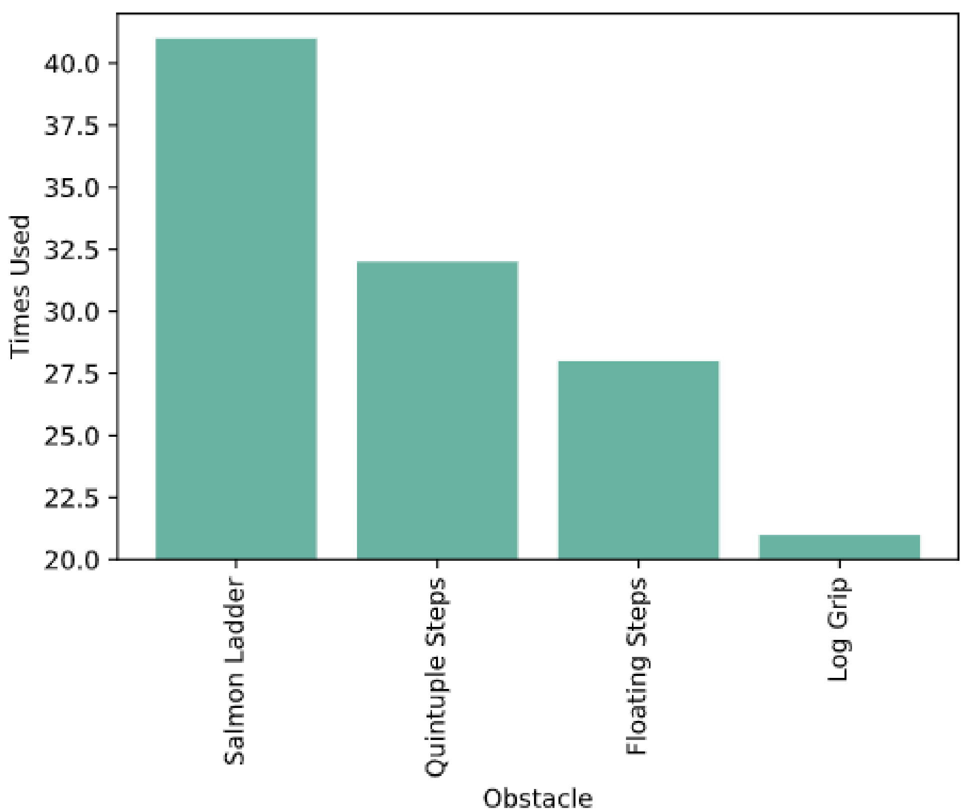
For example, if you think 'Log Grip' was used half as much as 'Salmon Ladder', enter 50%. If you think 'Log Grip' was used a quarter as much as 'Salmon Ladder', enter 25%, and so on.

____

American Ninja Warrior - Part 1, Question 3

Now consider bar chart C, below.

Bar chart C



In relation to bar chart C:

17. Approximately many times would you say the 'Salmon Ladder' was used?

____

18. Approximately how much more than 'Log Grip' would you say 'Salmon Ladder' was used?

Mark only one oval.

	1	2	3	4	5	6	7	
Not much more	◯	◯	◯	◯	◯	◯	◯	A lot more

19. Approximately how much more than 'Quintuple Steps' would you say 'Salmon Ladder' was used?

Mark only one oval.

	1	2	3	4	5	6	7	
Not much more	◯	◯	◯	◯	◯	◯	◯	A lot more

20. In your opinion, approximately how many times would you say 'Log Grip' was used, as a percentage of the number of times 'Salmon Ladder' was used?

For example, if you think 'Log Grip' was used half as much as 'Salmon Ladder', enter 50%. If you think 'Log Grip' was used a quarter as much as 'Salmon Ladder', enter 25%, and so on.

____

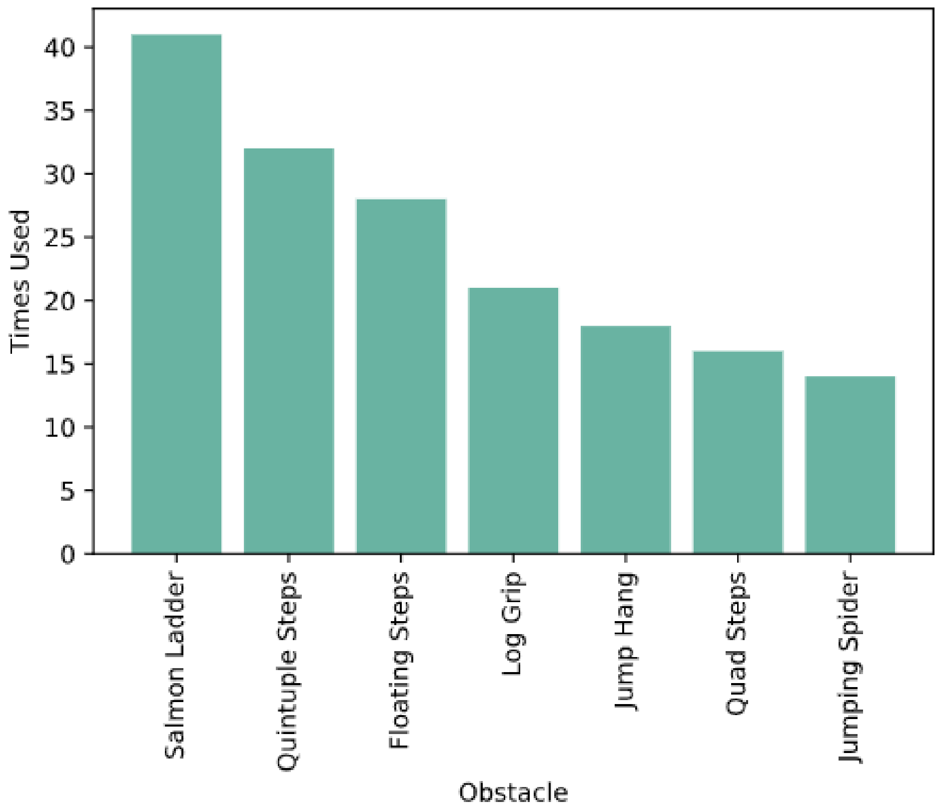
American Ninja Warrior - Part 2

You will now see bar charts regarding 7 of the most frequently used obstacles. Once again, for each question you will be asked to consider the bar chart and answer the corresponding questions. Again note that the answers to these questions are meant to be subjective, and that there are no correct answers.

American Ninja Warrior - Part 2, Question 1

Please see bar chart A, below.

Bar Chart A



In relation to bar chart A:

21. How large would you say the difference between 'Jumping spider' and 'Salmon Ladder' is?

Mark only one oval.

	1	2	3	4	5	6	7	
Very Small	◯	◯	◯	◯	◯	◯	◯	Very Large

22. How large would you say the difference between 'Log Grip' and 'Floating Steps' is?

Mark only one oval.

	1	2	3	4	5	6	7	
Very Small	◯	◯	◯	◯	◯	◯	◯	Very Large

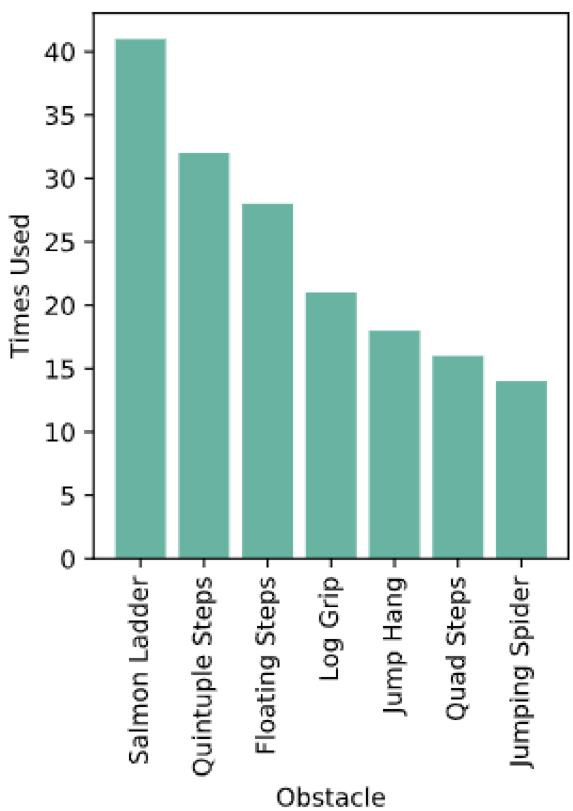
23. How many times would you say 'Floating Steps' were used?

____

Now please see bar chart B.

American Ninja Warrior - Part 2, Question 2

Bar Chart B



In relation to bar chart B:

24. How large would you say the difference between 'Jumping spider' and 'Salmon Ladder' is?

Mark only one oval.

	1	2	3	4	5	6	7	
Very Small								Very Large

25. How large would you say the difference between 'Log Grip' and 'Floating Steps' is?

Mark only one oval.

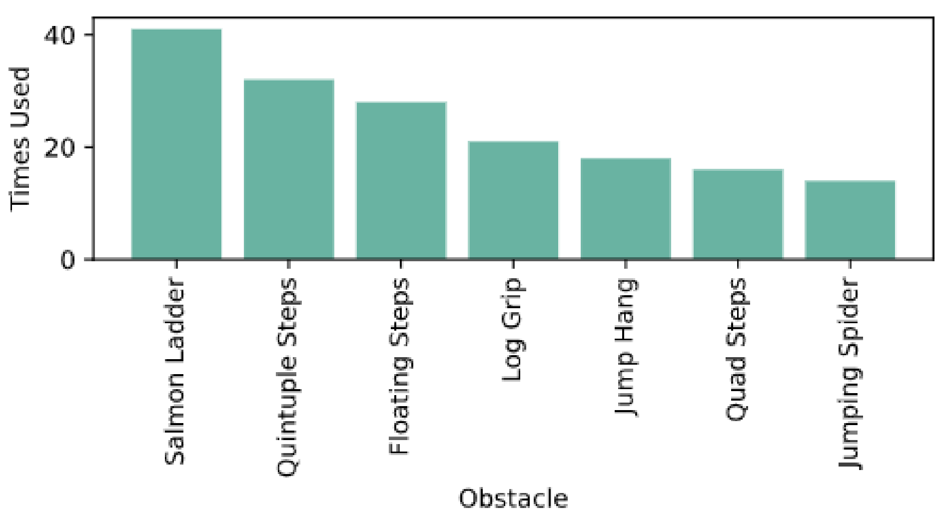
	1	2	3	4	5	6	7	
Very Small								Very Large

26. How many times would you say 'Floating Steps' were used?

American Ninja Warrior - Part 2, Question 3

Consider bar chart C, below.

Bar Chart C



In relation to bar chart C:

27. How large would you say the difference between 'Jumping spider' and 'Salmon Ladder' is?

Mark only one oval.

	1	2	3	4	5	6	7	
Very Small	⬭	⬭	⬭	⬭	⬭	⬭	⬭	Very Large

28. How large would you say the difference between 'Log Grip' and 'Floating Steps' is?

Mark only one oval.

	1	2	3	4	5	6	7	
Very Small	⬭	⬭	⬭	⬭	⬭	⬭	⬭	Very Large

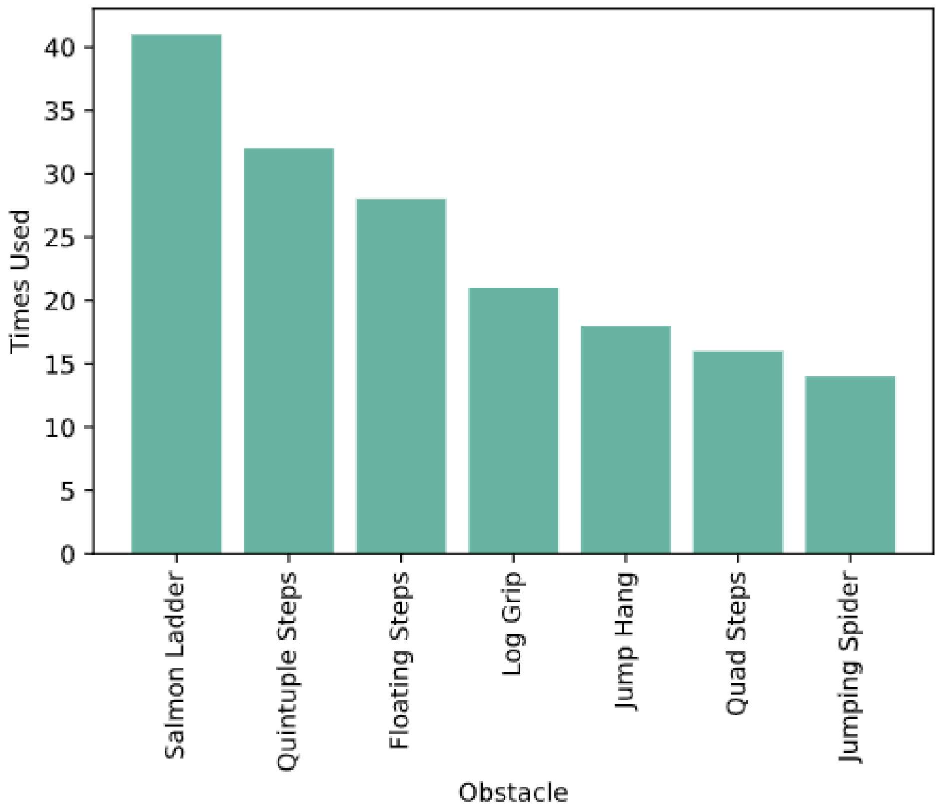
29. How many times would you say 'Floating Steps' were used?

____

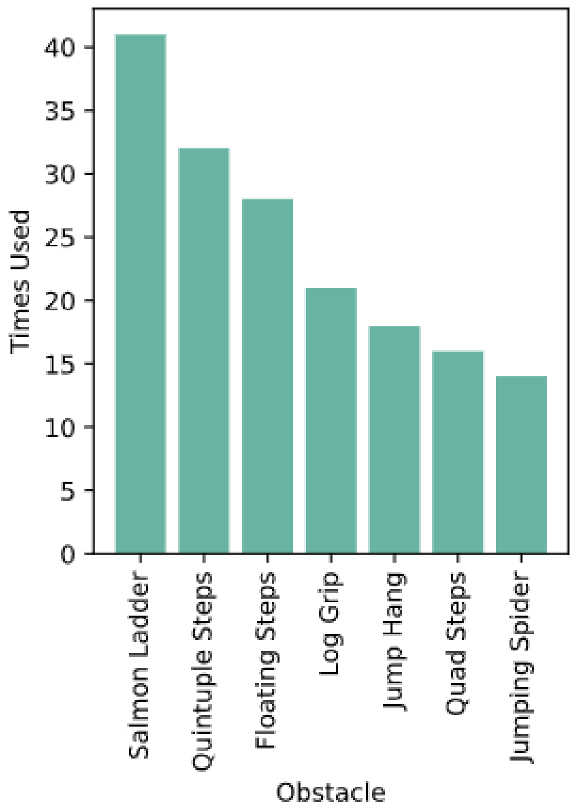
American Ninja Warrior - Part 2, Question 4

You will now see the bar charts A, B and C from part 2 again. Please answer the corresponding questions.

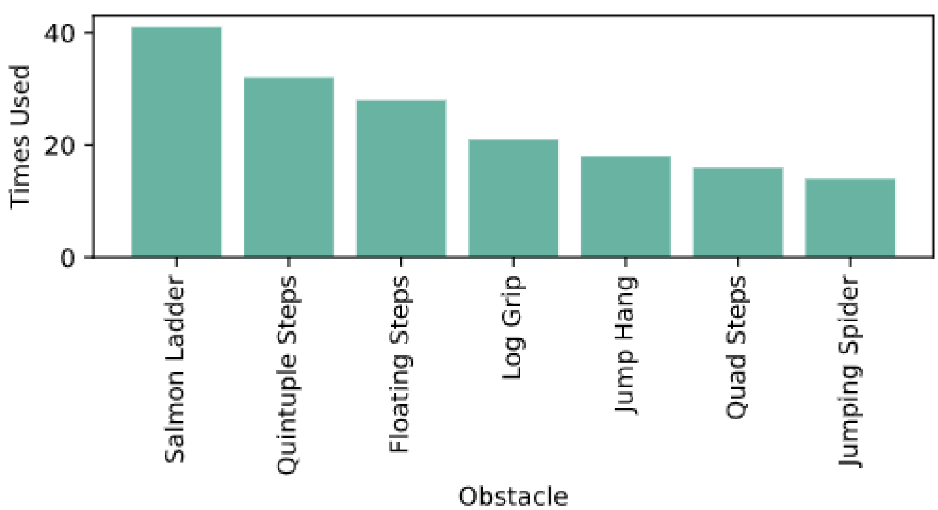
Bar chart A


Times Used
0
5
10
15
20
25
30
35
40
Salmon Ladder
Quintuple Steps
Floating Steps
Log Grip
Jump Hang
Quad Steps
Jumping Spider
Obstacle

Bar chart B



Bar chart C



30. Which of the three bar charts do you find most aesthetically pleasing?

Mark only one oval.

- A
- B
- C

31. Which bar chart do you feel is easiest to read and interpret?

This includes ease in reading labels, ease of reading the value of each bar, and ease in seeing the relative differences in values of the bars.

Mark only one oval.

- A
- B
- C

32. Which bar chart do you find hardest to read and interpret?

Mark only one oval.

- A
- B
- C

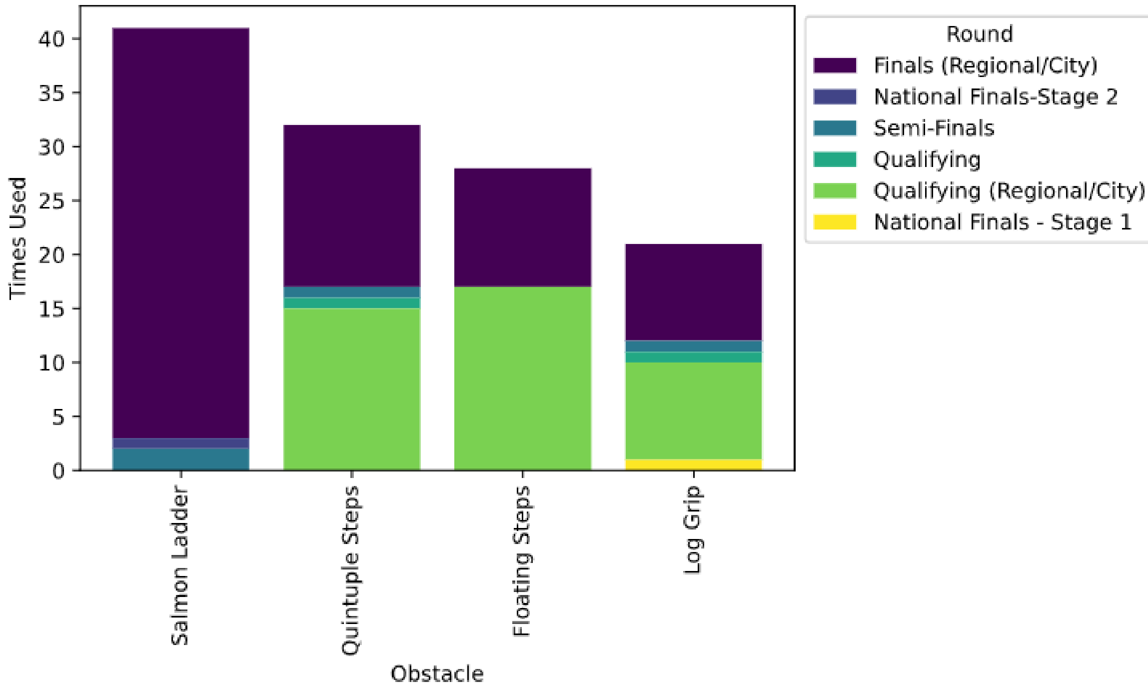
American Ninja Warrior - Part 3

The below plots show the same data as part 1, but are additionally coloured to show how many times each obstacle was used in each stage of the competition.

American Ninja Warrior - Part 3, Question 1

Please answer the questions below, for bar chart A.

Bar chart A



In relation to bar chart A:

33. How many times would you say 'Floating Steps' were used in the Finals (Regional/City) rounds?

34. How many times would you say 'Log Grip' was used in the Finals (Regional/City) rounds?

35. Please select the statement you feel applies to the bar chart above.

Mark only one oval.

- 'Log Grip' was used MORE in Finals (Regional/City) rounds than in Qualifying (Regional/City) rounds.
- 'Log Grip' was used LESS in Finals (Regional/City) rounds than in Qualifying (Regional/City) rounds.
- 'Log Grip' was used an EQUAL number of times in Finals (Regional/City) rounds and Qualifying (Regional/City) rounds.

36. Which obstacle do you think was used MORE in Finals (Regional/City) rounds, 'Log Grip' or 'Floating Steps'?

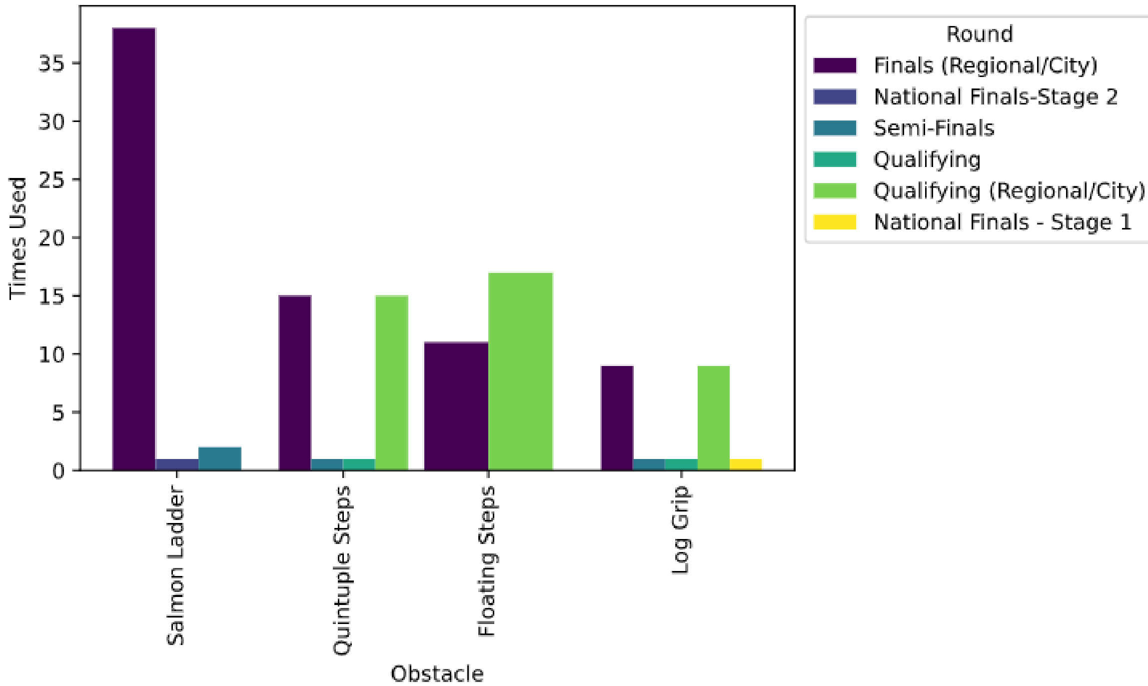
Mark only one oval.

- 'Log Grip'
- 'Floating Steps'
- They were used the same amount of times

American Ninja Warrior - Part 3, Question 2

Now please answer the questions again, for bar chart B.

Bar chart B



In relation to bar chart B:

37. How many times would you say 'Floating Steps' were used in the Finals (Regional/City) rounds?

38. How many times would you say 'Log Grip' was used in the Finals (Regional/City) rounds?

39. Please select the statement you feel applies to the bar chart above.

Mark only one oval.

- 'Log Grip' was used MORE in Finals (Regional/City) than in Qualifying (Regional/City).
- 'Log Grip' was used LESS in Finals (Regional/City) than in Qualifying (Regional/City).
- 'Log Grip' was used an EQUAL number of times in Finals (Regional/City) and Qualifying (Regional/City).

40. Which obstacle do you think was used MORE in Finals (Regional/City) rounds, 'Log Grip' or 'Floating Steps'?

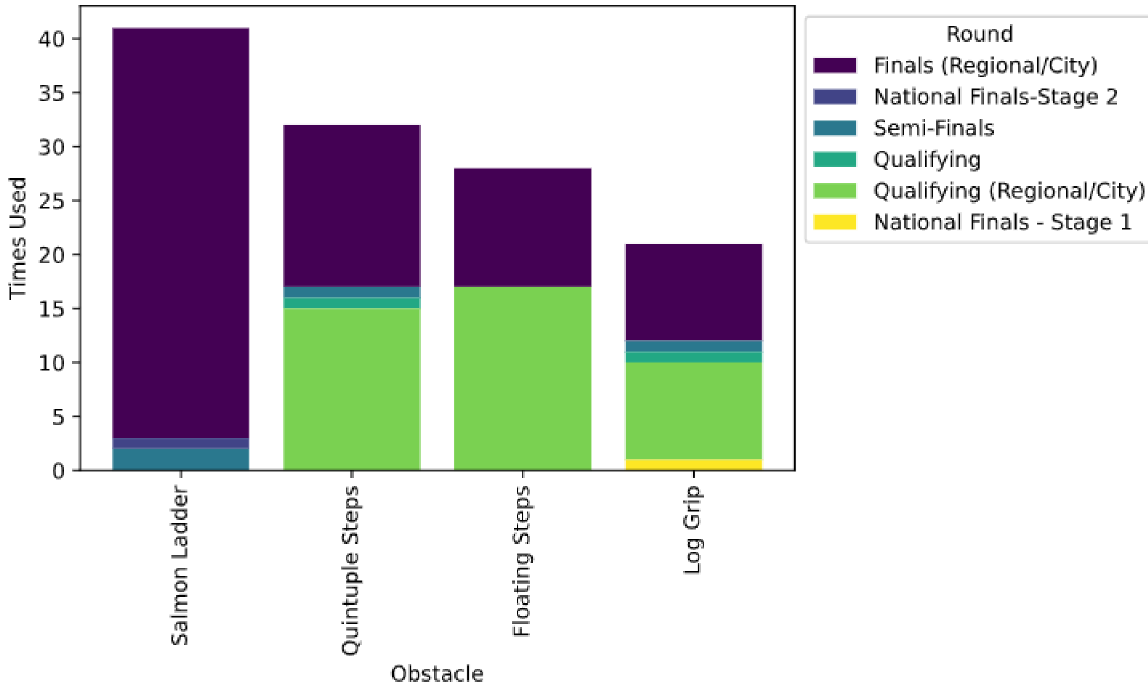
Mark only one oval.

- 'Log Grip'
- 'Floating Steps'
- They were used the same amount of times

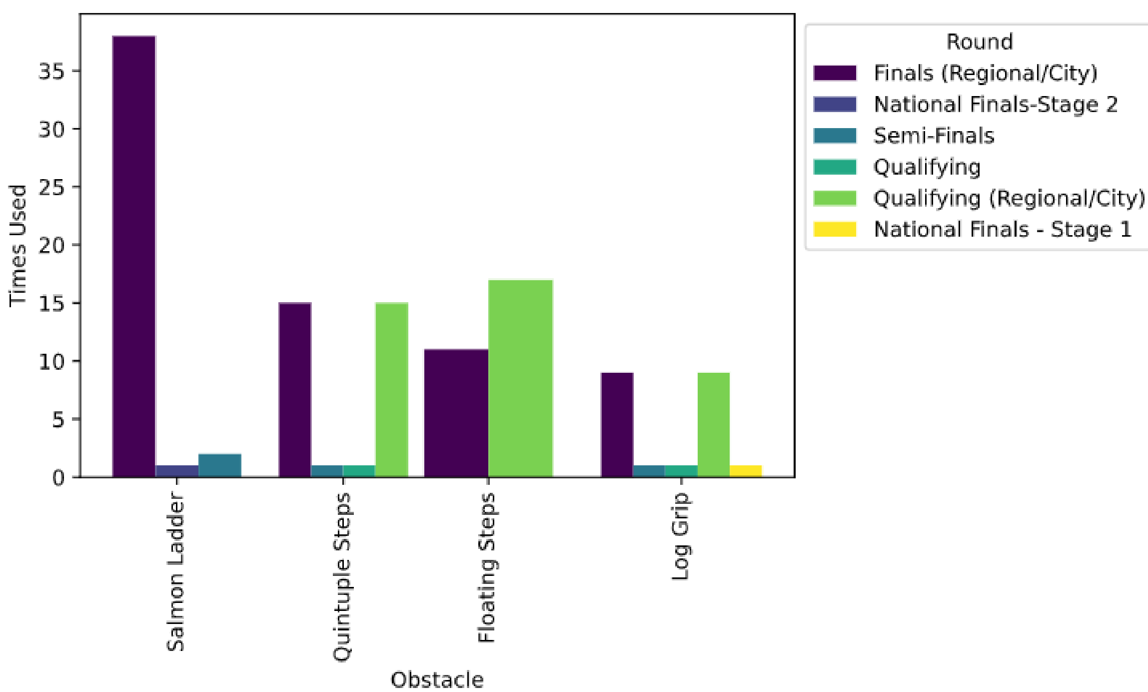
American Ninja Warrior - Part 3, Question 3

You will now see the two bar charts again. Please answer the question below.

Bar Chart A



Bar chart B



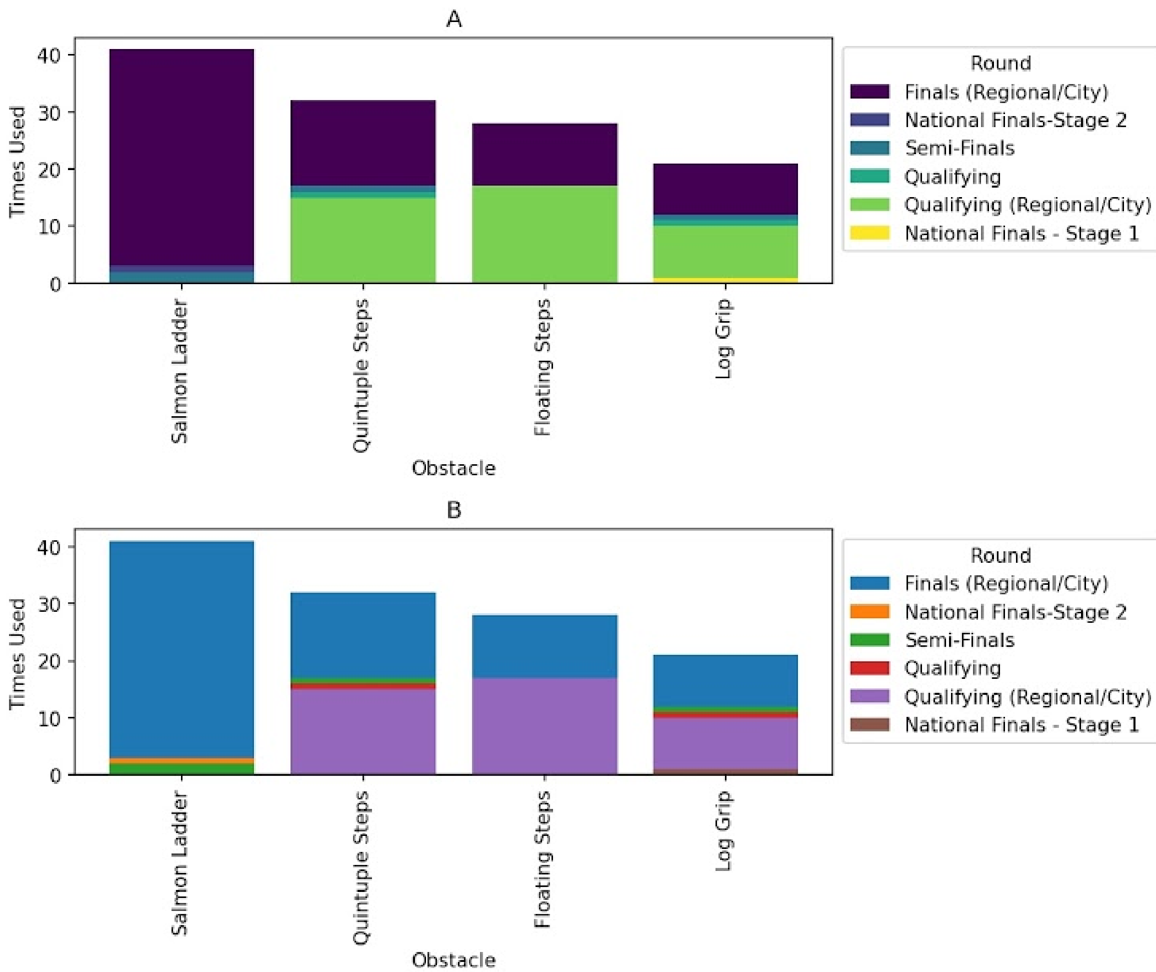
41. Which bar chart do you feel is easiest to read and interpret?

Mark only one oval.

- A
- B

American Ninja Warrior - Part 3, Question 4

Below compares the color scheme of the charts above, colour scheme A with another colour scheme, B.



42. Which colour scheme do you find most aesthetically pleasing?

Mark only one oval.

- A
- B

43. Do you feel that one of the colour schemes makes it easier to read and interpret the data than the other? If so, please select which one.

Mark only one oval.

- No
- Yes, A is easier
- Yes, B is easier

44.

Mark only one oval.

- Option 1

Sales

Thank you for completing the first section of this survey. The second, and final, section is shorter and will show you a series of line graphs relating to monthly sales data from four companies over the course of a year.

Once again, please refer to the following charts and answer the questions.

Please note that the answers to these questions are meant to be subjective, and that there are no correct answers.

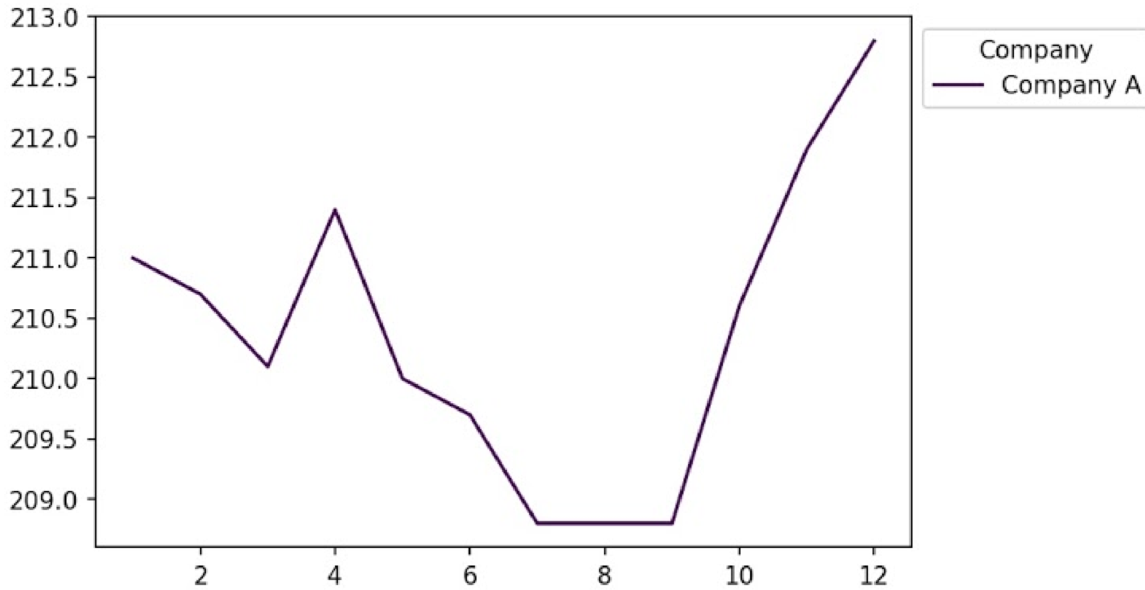
Sales - Part 1

The following section will present graphs showing monthly sales data of two competing companies over the course of a year, and ask you questions relating to the graphs.

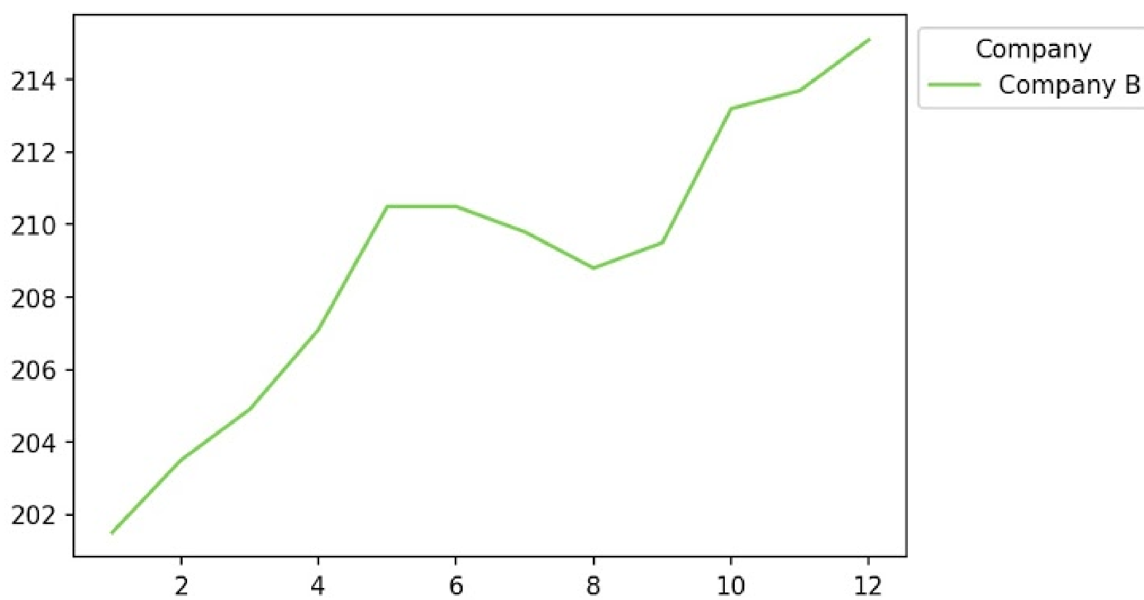
Sales - Part 1, Question 1

Please refer to the two charts below

Sales data for Company A



Sales data for Company B



45. How much would you say sales of each company increased between January and December?

Mark only one oval per row.

	1 (A little)	2	3	4	5	6	7 (A lot)
Company A	⬭	⬭	⬭	⬭	⬭	⬭	⬭
Company B	⬭	⬭	⬭	⬭	⬭	⬭	⬭

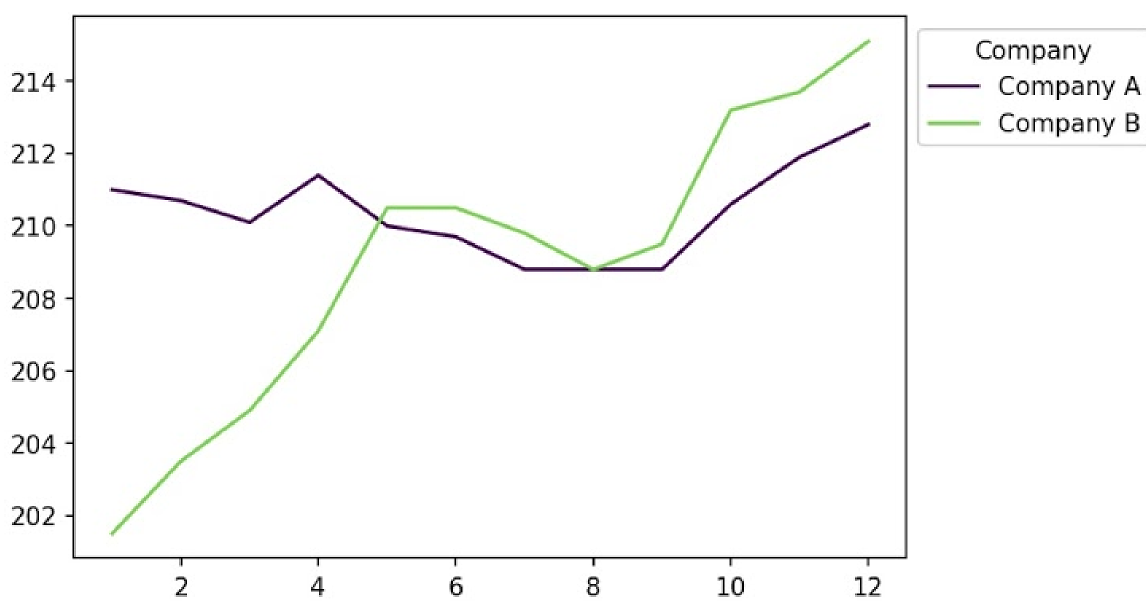
46. How large would you say the drop in sales between April and July of Company A is?

Mark only one oval.

	1	2	3	4	5	6	7	
A little	⬭	⬭	⬭	⬭	⬭	⬭	⬭	A lot

Sales - Part 1, Question 2

Consider the chart below



47. How much would you say sales of each company increased between January and December?

Mark only one oval per row.

	1 (A little)	2	3	4	5	6	7 (A lot)
Company A	◯	◯	◯	◯	◯	◯	◯
Company B	◯	◯	◯	◯	◯	◯	◯

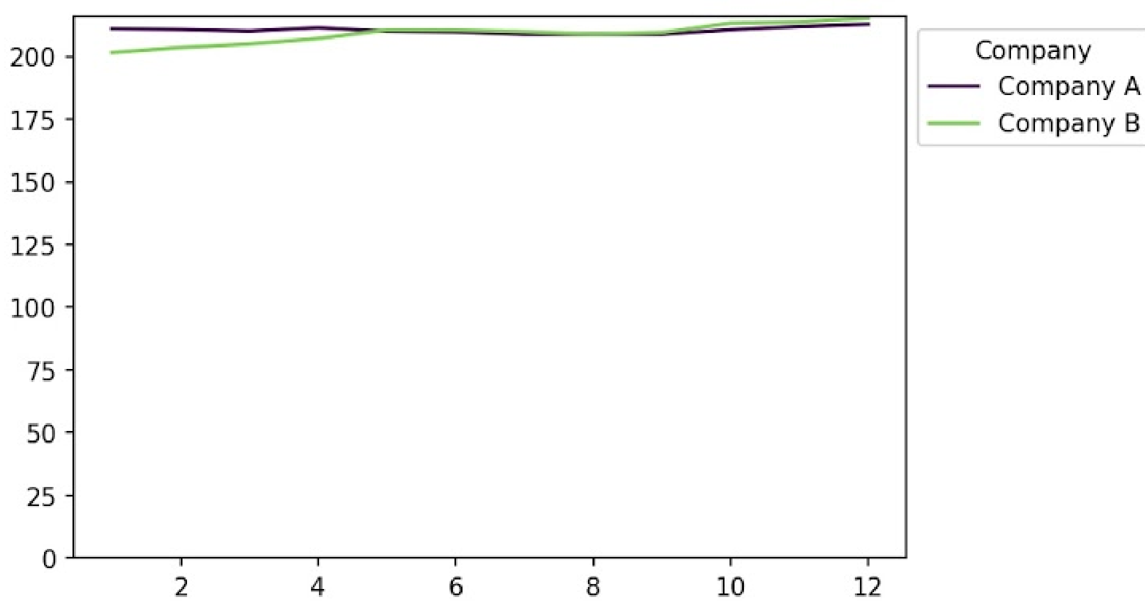
48. How large would you say the drop in sales between April and July of Company A is?

Mark only one oval.

	1	2	3	4	5	6	7	
A little	◯	◯	◯	◯	◯	◯	◯	A lot

Sales - Part 1, Question 3

Consider the chart below



49. How much would you say sales of each company increased between January and December?

Mark only one oval per row.

	1 (A little)	2	3	4	5	6	7 (A lot)
Company A	◯	◯	◯	◯	◯	◯	◯
Company B	◯	◯	◯	◯	◯	◯	◯

50. How large would you say the drop in sales between April and July of Company A is?

Mark only one oval.

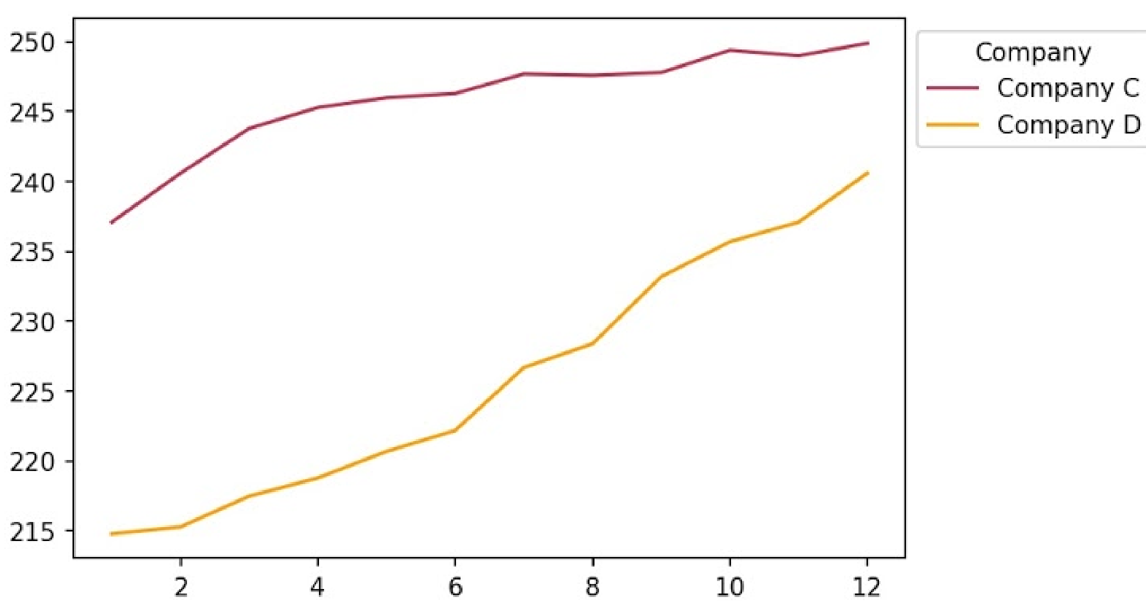
	1	2	3	4	5	6	7	
A little	◯	◯	◯	◯	◯	◯	◯	A lot

Sales - Part 2

The graphs below show the sales data for two more companies, Company C and Company D.

Sales - Part 2, Question 1

Consider the chart below



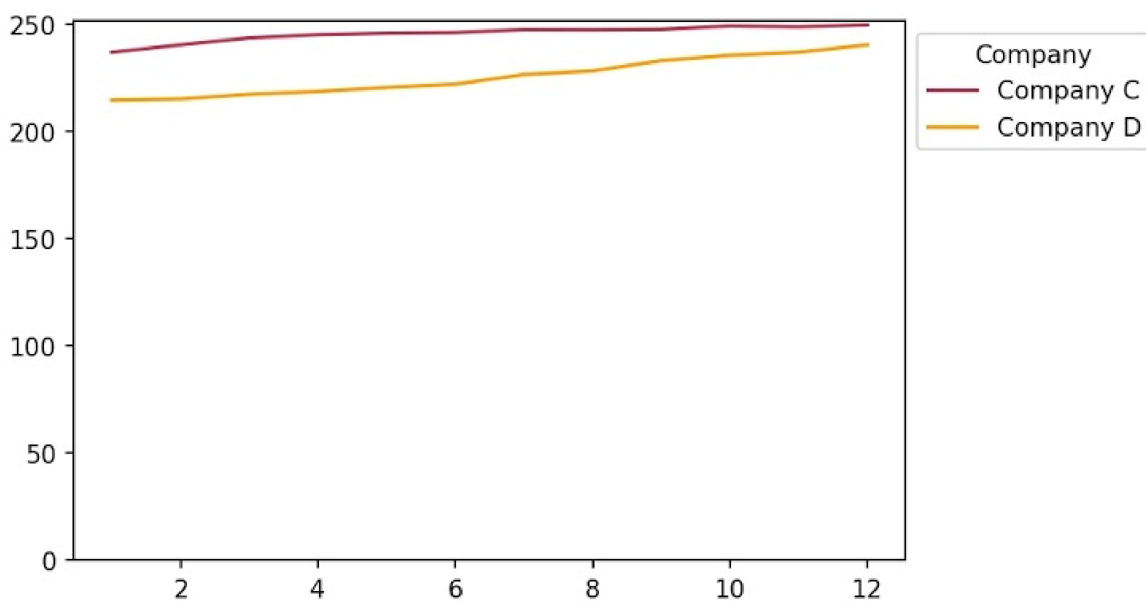
51. Based on the above graph, how large would you say the difference is between the number of sales Company C makes and the number of sales Company D makes?

Mark only one oval.

	1	2	3	4	5	6	7	
A little	◯	◯	◯	◯	◯	◯	◯	A lot

Sales - Part 2, Question 2

Consider the chart below



52. Based on the above graph, how large would you say the difference is between the number of sales Company C makes and the number of sales Company D makes?

Mark only one oval.

	1	2	3	4	5	6	7	
A little	◯	◯	◯	◯	◯	◯	◯	A lot

Reminder - Your responses to this survey will remain anonymous.

If you would like to know more about the use of data visualisation and how different techniques may impact interpretation, please see the links below.
- https://searchbusinessanalytics.techtarget.com/definition/data-visualization
- https://datajournalism.com/read/handbook/one/understanding-data/using-data-visualization-to-find-insights-in-data
- https://heap.io/blog/data-stories/how-to-lie-with-data-visualization
- https://www.idashboards.com/blog/2019/02/27/when-data-visualizations-mislead-and-how-to-prevent-it/

Thank you for completing this survey.

If you are interested in how colour blindness can affect reading of visualisations and colour blind friendly palettes to use for data visualisation, please see the links below:
- https://blog.datawrapper.de/colorblindness-part3/
- https://www.youtube.com/watch?v=xAoljeRJ3lU

Once again, if you have any questions please do not hesitate to contact me at murphyka1@cardiff.ac.uk.

4.2 Apendix 2 - R Code for Visualisations

```
##### Load libraries #####
library(ggplot2)
library(dplyr)
library(DAAG) # datasets
library(viridis) # colourmaps
library(pubr)
library(cowplot)
library(gridExtra)
library("httr")
library("readxl")
library("Cairo")
library(knitr)
###

##### Ninja Warrior data #####

### Load and manipulate data ###
# Data from: https://data.world/ninja/anw-obstacle-history

GET("https://query.data.world/s/ttkvg64pqsrncsfpopxl3kbxnmioyq", write_disk(tf <- tempfile(fileex
NinjaWarrior <- as.data.frame(read_excel(tf))
names(NinjaWarrior) <- c("season", "location", "stage", "name", "order")

obstacles <- function(ObstacleNumbers) {
  # Argument ObstacleNumbers specifies which obstacle index to subset for.

  # The function orders the data in decending order of how many times each obstacle
  # was used and then subsets for the required indices.

  # Outputs a table giving the name of each selected obstacle and the number of times it was used
  obst <- data.frame()
  j <- 1

  for (i in unique(NinjaWarrior$name)) {
    dat <- filter(NinjaWarrior, name == i)
    obst[j, 1] <- i
    obst[j, 2] <- dim(dat)[1]
    j <- j + 1
```

```
  }

  obst <- arrange(obst, desc(V2))
  obst <- obst[ObstacleNumbers, ]
  names(obst) <- c("name", "ntimes")
  return(obst)
}



stack_data <- function(ObstacleNumbers) {
  # Argument ObstacleNumbers specifies which obstacle index to subset for after
  # arranging in descending order.

  # Creates data set that can be used to created stacked bar plot.

  obst <- obstacles(ObstacleNumbers)

  ## List obstacles the number of times they are used##
  obstcount <- c()
  ltrs <- c()
  for (i in 1:dim(obst)[1]) {
    letter <- letters[i]
    obstcount <- append(obstcount, rep(obst[i, 1], obst[i, 2]))
    ltrs <- append(ltrs, rep(letter, obst[i, 2]))
  }

  bardat <- as.data.frame(obstcount) # Change to dataframe

  stage.count <- data.frame()
  for (i in 1:(length(ObstacleNumbers))) {
    filter_ninja <- filter(NinjaWarrior, name == obst[i, 1]) # Filters data for relevant obstacle

    freq_stages <- as.data.frame(table(filter_ninja$stage)) # Count how many times obstacle i is

    stages <- rep(freq_stages$Var1, freq_stages$Freq) # Lists each stage the number of times it c

    obs <- filter(bardat, obstcount == obst[i, 1]) # Filter for obstacle i

    stages_obs <- cbind(obs, stages)
```

```
    stage.count <- rbind(stage.count, stages_obs) # Iterate to get whole set of data
  }
  stage.count <- cbind(stage.count, ltrs)
  names(stage.count) <- c("obstacle", "stage", "order")
  return(stage.count)
}


### The two lines below are to order the bar plots in the required order as opposed to alphabetic
letters <- c("A", "B", "C", "D", "E", "F", "G")
obst_list <- c("Salmon Ladder", "Quintuple Steps", "Floating Steps", "Log Grip", "Jump Hang", "Qu


barplots_yscaling <- function(n.obst) {

  # n.obst specifies the indices of obstacles to be plotted
  # Function returns and saves bar plots with different axis scalings

  obst <- obstacles(n.obst)
  obst$name <- letters[1:length(n.obst)]

  # default scale
  dflt <- ggplot(data = obst) +
    geom_bar(aes(x = name, y = ntimes), fill = "#69b3a2", stat = "identity") +
    scale_x_discrete(labels = obst_list[1:length(n.obst)]) +
    scale_y_continuous() +
    xlab("Obstacle") +
    ylab("Times Used") +
    labs(fill = "Stage") +
    theme_classic() +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz

  # log scale
  lg <- ggplot(data = obst) +
    geom_bar(aes(x = name, y = ntimes), fill = "#69b3a2", stat = "identity") +
    xlab("Obstacle") +
    ylab("Times Used") +
    labs(fill = "Stage") +
    scale_x_discrete(labels = obst_list[1:length(n.obst)]) +
```

```
  scale_y_log10() +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz

# truncated y
obst <- obstacles(ObstacleNumbers = n.obst)
names(obst) <- c("name", "ntimes")
obst$name <- c("A", "B", "C", "D")

obstcount <- c()
for (i in 1:dim(obst)[1]) {
  obstcount <- append(
    obstcount, rep(obst[i, 1], ((obst[i, 2] - (min(obst[, 2])) + 1)))
  )
}

bardat <- as.data.frame(obstcount)



trnc <- ggplot(data = bardat) +
  geom_bar(aes(x = obstcount), fill = "#69b3a2") +
  xlab("Obstacle") +
  ylab("Times Used") +
  labs(fill = "Stage") +
  scale_x_discrete(labels = obst_list[1:length(n.obst)]) +
  scale_y_continuous(limits = c(0, 21), breaks = c(0, 10, 20), labels = c(20, 30, 40)) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz

all <- plot_grid(lg, dflt, trnc,
  labels = c("A", "B", "C"),
  ncol = 2, nrow = 2
)

ggsave("Plots/R/Ninja_Data/y-scaling/all.jpeg")

print(dflt)
ggsave("Plots/R/Ninja_Data/y-scaling/control.jpeg")
```

```
  print(lg)
  ggsave("Plots/R/Ninja_Data/y-scaling/log10.jpeg")

  print(trnc)
  ggsave("Plots/R/Ninja_Data/y-scaling/truncated.jpeg")
}
barplots_yscaling(2:5)

barplots_axisratio <- function(n.obst) {

  # n.obst specifies the indices of obstacles to be plotted
  # Function returns and saves bar plots with different aspect ratios

  obst <- obstacles(n.obst)
  obst$name <- letters[1:length(n.obst)]

  # default
  dflt <- ggplot(data = obst) +
    geom_bar(aes(x = name, y = ntimes), fill = "#69b3a2", stat = "identity") +
    scale_x_discrete(labels = ) +
    scale_y_continuous() +
    xlab("Obstacle") +
    ylab("Times Used") +
    labs(fill = "Stage") +
    theme_classic() +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz

  # narrow bars
  narrow <- ggplot(data = obst) +
    geom_bar(aes(x = name, y = ntimes), fill = "#69b3a2", stat = "identity") +
    scale_x_discrete(labels = obst_list[1:length(n.obst)]) +
    scale_y_continuous() +
    xlab("Obstacle") +
    ylab("Times Used") +
    labs(fill = "Stage") +
    theme_classic() +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz
    theme(aspect.ratio = 2 / 1)
```

```
  # wide bars
  wide <- ggplot(data = obst) +
    geom_bar(aes(x = name, y = ntimes), fill = "#69b3a2", stat = "identity") +
    scale_x_discrete(labels = obst_list[1:length(n.obst)]) +
    scale_y_continuous() +
    xlab("Obstacle") +
    ylab("Times Used") +
    labs(fill = "Stage") +
    theme_classic() +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz
    theme(aspect.ratio = 0.5 / 1)

  top <- plot_grid(narrow, dflt,
    labels = c("A", "B"),
    ncol = 2, nrow = 1
  )

  all <- plot_grid(top, wide,
    labels = c(" ", "C"),
    ncol = 1, nrow = 2
  )

  ggsave("Plots/R/Ninja_Data/bars/all.jpeg")

  print(wide)
  ggsave("Plots/R/Ninja_Data/bars/wide.jpeg")

  print(narrow)
  ggsave("Plots/R/Ninja_Data/bars/narrow.jpeg")

  print(dflt)
  ggsave("Plots/R/Ninja_Data/bars/control.jpeg")
}
barplots_axisratio(2:5)

barplots_stacked <- function(n.obst) {

  # n.obst specifies the indices of obstacles to be plotted
  # Function returns and saves stacked bar plots with different colourings
```

```
stage.count <- stack_data(n.obst)

# vir
vir <- ggplot(data = stage.count) +
  geom_bar(aes(x = ltrs, fill = stages)) +
  scale_y_continuous() +
  scale_fill_viridis(discrete = TRUE, option = "D") +
  scale_x_discrete(labels = obst_list[1:length(n.obst)]) +
  xlab("Obstacle") +
  ylab("Times Used") +
  labs(fill = "Stage") +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz

# vir2
vir2 <- ggplot(data = stage.count) +
  geom_bar(aes(x = ltrs, fill = stages)) +
  scale_y_continuous() +
  scale_fill_viridis(discrete = TRUE, option = "A") +
  scale_x_discrete(labels = obst_list[1:length(n.obst)]) +
  xlab("Obstacle") +
  ylab("Times Used") +
  labs(fill = "Stage") +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz


# grey
grey <- ggplot(data = stage.count) +
  geom_bar(aes(x = ltrs, fill = stages)) +
  scale_y_continuous() +
  scale_fill_grey() +
  scale_x_discrete(labels = obst_list[1:length(n.obst)]) +
  xlab("Obstacle") +
  ylab("Times Used") +
  labs(fill = "Stage") +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz
```

```
# fil
fil <- ggplot(data = stage.count) +
  geom_bar(aes(x = ltrs, fill = stages)) +
  scale_y_continuous() +
  scale_x_discrete(labels = obst_list[1:length(n.obst)]) +
  xlab("Obstacle") +
  ylab("Times Used") +
  labs(fill = "Stage") +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz

print(vir)
ggsave("Plots/R/Ninja_Data/colours/viridis.jpeg")

print(grey)
ggsave("Plots/R/Ninja_Data/colours/grey.jpeg")

print(fil)
ggsave("Plots/R/Ninja_Data/colours/default.jpeg")



plot_grid(vir, fil,
  labels = c("A", "B"),
  ncol = 1, nrow = 2
)
ggsave("Plots/R/Ninja_Data/colours/col_set_a.jpeg")

plot_grid(fil, vir,
  labels = c("A", "B"),
  ncol = 1, nrow = 2
)
ggsave("Plots/R/Ninja_Data/colours/col_set_b.jpeg")

plot_grid(fil, grey,
  labels = c("A", "B"),
  ncol = 1, nrow = 2
)
ggsave("Plots/R/Ninja_Data/colours/col_set_c.jpeg")
```

```
  plot_grid(grey, fil,
    labels = c("A", "B"),
    ncol = 1, nrow = 2
  )
  ggsave("Plots/R/Ninja_Data/colours/col_set_d.jpeg")

  plot_grid(vir, grey,
    labels = c("A", "B"),
    ncol = 1, nrow = 2
  )
  ggsave("Plots/R/Ninja_Data/colours/col_set_e.jpeg")

  plot_grid(grey, vir,
    labels = c("A", "B"),
    ncol = 1, nrow = 2
  )
  ggsave("Plots/R/Ninja_Data/colours/col_set_f.jpeg")
}
barplots_stacked(2:5)

barplots_grouped <- function(n.obst) {

  # n.obst specifies the indices of obstacles to be plotted
  # Function returns and saves stacked bar plots with different colourings

  stage.count <- stack_data(n.obst)

  # vir
  vir <- ggplot(data = stage.count) +
    geom_bar(aes(x = ltrs, fill = stages), position = "dodge") +
    scale_y_continuous() +
    scale_x_discrete(labels = obst_list[1:length(n.obst)]) +
    scale_fill_viridis(discrete = TRUE, option = "D") +
    xlab("Obstacle") +
    ylab("Times Used") +
    labs(fill = "Stage") +
    theme_classic() +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz
```

```
# vir2
vir2 <- ggplot(data = stage.count) +
  geom_bar(aes(x = ltrs, fill = stages), position = "dodge") +
  scale_y_continuous() +
  scale_x_discrete(labels = obst_list[1:length(n.obst)]) +
  scale_fill_viridis(discrete = TRUE, option = "A") +
  xlab("Obstacle") +
  ylab("Times Used") +
  labs(fill = "Stage") +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz


# grey
grey <- ggplot(data = stage.count) +
  geom_bar(aes(x = ltrs, fill = stages), position = "dodge") +
  scale_y_continuous() +
  scale_x_discrete(labels = obst_list[1:length(n.obst)]) +
  scale_fill_grey() +
  xlab("Obstacle") +
  ylab("Times Used") +
  labs(fill = "Stage") +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz

# def
def <- ggplot(data = stage.count) +
  geom_bar(aes(x = ltrs, fill = stages), position = "dodge") +
  scale_y_continuous() +
  scale_x_discrete(labels = obst_list[1:length(n.obst)]) +
  xlab("Obstacle") +
  ylab("Times Used") +
  labs(fill = "Stage") +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz

print(vir)
ggsave("Plots/R/Ninja_Data/colours/viridis_d.jpeg")
```

```
  print(grey)
  ggsave("Plots/R/Ninja_Data/colours/grey_d.jpeg")

  print(def)
  ggsave("Plots/R/Ninja_Data/colours/default_d.jpeg")
}
barplots_grouped(2:5)


##### Time-series #####


sales_line_plots <- function(npoints) {
  # npoints specifies number of points in each time series
  # Function returns and saves line plots with different axis scaling

  dat <- c(BJsales) # read in and concatenate data to a vector

  months_list <- c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "
  months <- as.factor(1:npoints) # Select number of months for time series

  # Selecting start point for time series and creating a data set for this and the 'npoints'-1
  # subsequent points (ie. for npoints = 12, select start point and the 11 following points)
  set.seed(seeds[1])
  start <- sample(138, 1)
  vals <- dat[start:(start + npoints - 1)]
  group <- rep("Company A", length(months))
  salesA <- data.frame(Month = months, Sales = vals, Company = group)

  set.seed(seeds[2])
  start <- sample(138, 1)
  vals <- dat[start:(start + npoints - 1)]
  group <- rep("Company B", length(months))
  salesB <- data.frame(Month = months, Sales = vals, Company = group)

  set.seed(seeds[3])
  start <- sample(138, 1)
  vals <- dat[start:(start + npoints - 1)]
  group <- rep("Company C", length(months))
```

```
salesC <- data.frame(Month = months, Sales = vals, Company = group)

set.seed(seeds[4])
start <- sample(138, 1)
vals <- dat[start:(start + npoints - 1)]
group <- rep("Company D", length(months))
salesD <- data.frame(Month = months, Sales = vals, Company = group)

sales <- rbind(salesA, salesB)
sales2 <- rbind(salesC, salesD)

# CONTROLS #
dflt_oneline_1 <- ggplot(data = salesA, aes(x = Month, y = Sales, group = Company, col = Compan
  geom_line(size = 1) +
  scale_color_manual(values = c("#453781FF")) +
  theme_classic() +
  scale_x_discrete(labels = months_list[1:npoints]) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz
ggsave("Plots/R/Sales/one_line_control_1.jpeg")

dflt_oneline_2 <- ggplot(data = salesB, aes(x = Month, y = Sales, group = Company, col = Compan
  geom_line(size = 1) +
  scale_color_manual(values = c("#55C667FF")) +
  theme_classic() +
  scale_x_discrete(labels = months_list[1:npoints]) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz
ggsave("Plots/R/Sales/one_line_control_2.jpeg")


dflt_twolines <- ggplot(data = sales, aes(x = Month, y = Sales, group = Company, col = Company)
  geom_line(size = 1) +
  scale_color_manual(values = c("#453781FF", "#55C667FF")) +
  theme_classic() +
  scale_x_discrete(labels = months_list[1:npoints]) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz
ggsave("Plots/R/Sales/two_lines_control.jpeg")


# SECOND CONTROLS #
```

```
dflt_oneline_inferno <- ggplot(data = salesC, aes(x = Month, y = Sales, group = Company, col = 
  geom_line(size = 1) +
  scale_color_manual(values = c("#a72497")) +
  theme_classic() +
  scale_x_discrete(labels = months_list[1:npoints]) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz
ggsave("Plots/R/Sales/one_line_inferno1.jpeg")

dflt_oneline_inferno <- ggplot(data = salesD, aes(x = Month, y = Sales, group = Company, col = 
  geom_line(size = 1) +
  scale_color_manual(values = c("#f07f4f")) +
  theme_classic() +
  scale_x_discrete(labels = months_list[1:npoints]) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz
ggsave("Plots/R/Sales/one_line_inferno2.jpeg")

twoline_inferno <- ggplot(data = sales2, aes(x = Month, y = Sales, group = Company, col = Compa
  geom_line(size = 1) +
  scale_color_manual(values = c("#a72497", "#f07f4f")) +
  theme_classic() +
  scale_x_discrete(labels = months_list[1:npoints]) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz
ggsave("Plots/R/Sales/twoline_inferno.jpeg")

zeroed_twoline_inferno <- ggplot(data = sales2, aes(x = Month, y = Sales, group = Company, col 
  geom_line(size = 1) +
  scale_color_manual(values = c("#a72497", "#f07f4f")) +
  theme_classic() +
  scale_x_discrete(labels = months_list[1:npoints]) +
  scale_y_continuous(limits = c(0, max(sales2$Sales))) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz
ggsave("Plots/R/Sales/zeroed_twoline_inferno.jpeg")

# ZEROED #
zeroed_oneline_1 <- ggplot(data = salesA, aes(x = Month, y = Sales, group = Company, col = Comp
  geom_line(size = 1) +
  scale_color_manual(values = c("#453781FF")) +
  theme_classic() +
  scale_x_discrete(labels = months_list[1:npoints]) +
```

```
  scale_y_continuous(limits = c(0, max(sales[1:npoints, ]$Sales))) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz
ggsave("Plots/R/Sales/zeroed_oneline_1.jpeg")

zeroed_oneline_2 <- ggplot(data = salesB, aes(x = Month, y = Sales, group = Company, col = Comp
  geom_line(size = 1) +
  scale_color_manual(values = c("#55C667FF")) +
  theme_classic() +
  scale_x_discrete(labels = months_list[1:npoints]) +
  scale_y_continuous(limits = c(0, max(sales[npoints + 1:2 * npoints, ]$Sales))) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz
ggsave("Plots/R/Sales/zeroed_oneline_2.jpeg")

zeroed_twolines <- ggplot(data = sales, aes(x = Month, y = Sales, group = Company, col = Compan
  geom_line(size = 1) +
  scale_color_manual(values = c("#453781FF", "#55C667FF")) +
  theme_classic() +
  scale_x_discrete(labels = months_list[1:npoints]) +
  scale_y_continuous(limits = c(0, max(sales$Sales))) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz
ggsave("Plots/R/Sales/zeroed_twolines.jpeg")


# ASPECT RATIO #

large <- ggplot(data = salesA, aes(x = Month, y = Sales, group = Company, col = Company)) +
  geom_line(size = 1) +
  scale_color_manual(values = c("#453781FF")) +
  theme_classic() +
  scale_x_discrete(labels = months_list[1:npoints]) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz
  theme(aspect.ratio = 2 / 1)
ggsave("Plots/R/Sales/large.jpeg")

small <- ggplot(data = salesA, aes(x = Month, y = Sales, group = Company, col = Company)) +
  geom_line(size = 1) +
  scale_color_manual(values = c("#453781FF")) +
  theme_classic() +
  scale_x_discrete(labels = months_list[1:npoints]) +
```

```
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1), text = element_text(siz
    theme(aspect.ratio = 0.3 / 1)
  ggsave("Plots/R/Sales/small.jpeg")

  return(sales2)
}



### TESTING SEEDS ###
# seeds <- sample(1000, 4)
seeds <- c(62, 905, 511, 579)
sales_line_plots(npoints = 12)
seeds
```

4.3 Apendix 3 - Python Code for Visualisations

```
# python -m pip install -U pip
# pip install matplotlib
# pip install pandas
# pip install openpyxl

import matplotlib as mat
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import openpyxl as xl
import math
from matplotlib import cm

import os

# Set working directory
os.chdir("C:/Users/Katie/OneDrive/Uni_Work_Year4/Project/Year-4-Project/Plots/Py")

# Read in and manipulate data
NinjaWarrior = pd.read_excel(
    "https://query.data.world/s/tjkd2yop2xh2x4o2j5fhatxdidcvty"
)
NinjaWarrior = NinjaWarrior.rename(
```

```
    columns={
        "Season": "season",
        "Location": "location",
        "Round/Stage": "stage",
        "Obstacle Name": "name",
        "Obstacle Order": "order",
    }
)


def obstacles(ObstacleNumbers):
  # The function orders the data in decending order of how many times each obstacle
  # was used and then subsets for the required indices.

    obst = pd.DataFrame(columns={"name", "ntimes"})

    for i in np.unique(NinjaWarrior.name):
        dat = NinjaWarrior[(NinjaWarrior.name == i)]
        new_row = {"name": i, "ntimes": len(dat["name"])}
        obst = obst.append(new_row, ignore_index=True)

    obst = obst.sort_values(by=["ntimes"], inplace=True, ascending=False)
    obst = obst.iloc[1:ObstacleNumbers]

    print(obst)

# Above function does not operate properly,
# so manually create data.
name = [
    "Salmon Ladder",
    "Quintuple Steps",
    "Floating Steps",
    "Log Grip",
    "Jump Hang",
    "Quad Steps",
    "Jumping Spider",
]
ntimes = [41, 32, 28, 21, 18, 16, 14]
```

```
FinalsRegionalCity = np.array([38, 15, 11, 9])
NationalFinalsStage2 = np.array([1, 0, 0, 0])
SemiFinals = np.array([2, 1, 0, 1])
Qualifying = np.array([0, 1, 0, 1])
QualifyingRegionalCity = np.array([0, 15, 17, 9])
NationalFinalsStage1 = np.array([0, 0, 0, 1])

rounds = np.array(
    [
        "Finals (Regional/City)",
        "National Finals-Stage 2",
        "Semi-Finals",
        "Qualifying",
        "Qualifying (Regional/City)",
        "National Finals - Stage 1",
    ]
)

obst = pd.DataFrame(data={"name": name, "ntimes": ntimes})
  # n.obst specifies the indices of obstacles to be plotted
  # Function returns and saves bar plots with different axis scalings

def barplots_yscaling(nobst):

    ##### CONTROL #####
    plt.bar(name[0:nobst], ntimes[0:nobst], color="#69b3a2")
    plt.xlabel("Obstacle")
    plt.ylabel("Times Used")
    plt.xticks(rotation=90)

    ##### LOG10 #####
    plt.bar(name[0:nobst], ntimes[0:nobst], color="#69b3a2")
    plt.yscale("symlog")
    plt.xlabel("Obstacle")
    plt.ylabel("Times Used")
    plt.xticks(rotation=90)

    ##### TRUNCATED #####
    plt.bar(name[0:nobst], ntimes[0:nobst], color="#69b3a2")
```

```
    plt.xlabel("Obstacle")
    plt.ylabel("Times Used")
    plt.xticks(rotation=90)
    plt.ylim([20, 42])


barplots_yscaling(4)


def barplots_axisratio():
  # n.obst specifies the indices of obstacles to be plotted
  # Function returns and saves bar plots with different aspect ratios

    ##### CONTROL #####
    plt.bar(name[0:nobst], ntimes[0:nobst], color="#69b3a2")
    plt.xlabel("Obstacle")
    plt.ylabel("Times Used")
    plt.xticks(rotation=90)
    plt.show()

    ##### TALL #####
    plt.bar(name[0:nobst], ntimes[0:nobst], color="#69b3a2")
    plt.xlabel("Obstacle")
    plt.ylabel("Times Used")
    plt.xticks(rotation=90)
    plt.gca().set_aspect("0.2")
    plt.show()

    ##### WIDE #####
    plt.bar(name[0:nobst], ntimes[0:nobst], color="#69b3a2")
    plt.xlabel("Obstacle")
    plt.ylabel("Times Used")
    plt.xticks(rotation=90)
    plt.gca().set_aspect("0.05")
    plt.show()


barplots_axisratio(7)
```

```
def barplots_stacked(nobst):
  # n.obst specifies the indices of obstacles to be plotted
  # Function returns and saves stacked bar plots with different colourings

    ##### CONTROL #####
    def default(title):
        plt.bar(
            name[0:nobst],
            FinalsRegionalCity,
            bottom=(
                NationalFinalsStage2
                + SemiFinals
                + Qualifying
                + QualifyingRegionalCity
                + NationalFinalsStage1
            ),
        )

        plt.bar(
            name[0:nobst],
            NationalFinalsStage2,
            bottom=(
                SemiFinals + Qualifying + QualifyingRegionalCity + NationalFinalsStage1
            ),
        )

        plt.bar(
            name[0:nobst],
            SemiFinals,
            bottom=(Qualifying + QualifyingRegionalCity + NationalFinalsStage1),
        )

        plt.bar(
            name[0:nobst],
            Qualifying,
            bottom=(QualifyingRegionalCity + NationalFinalsStage1),
        )

        plt.bar(name[0:nobst], QualifyingRegionalCity, bottom=(NationalFinalsStage1))
```

```
    plt.bar(name[0:nobst], NationalFinalsStage1)

    plt.xlabel("Obstacle")
    plt.ylabel("Times Used")
    plt.xticks(rotation=90)
    plt.title(title)
    plt.legend(rounds, loc=0, title="Round", bbox_to_anchor=(1, 1))

##### VIRIDIS #####
def viridis(title):
    viridis = cm.get_cmap("viridis", 6)

    plt.bar(
        name[0:nobst],
        FinalsRegionalCity,
        color=viridis.colors[0],
        bottom=(
            NationalFinalsStage2
            + SemiFinals
            + Qualifying
            + QualifyingRegionalCity
            + NationalFinalsStage1
        ),
    )

    plt.bar(
        name[0:nobst],
        NationalFinalsStage2,
        color=viridis.colors[1],
        bottom=(
            SemiFinals + Qualifying + QualifyingRegionalCity + NationalFinalsStage1
        ),
    )

    plt.bar(
        name[0:nobst],
        SemiFinals,
        color=viridis.colors[2],
        bottom=(Qualifying + QualifyingRegionalCity + NationalFinalsStage1),
```

```
    )

    plt.bar(
        name[0:nobst],
        Qualifying,
        color=viridis.colors[3],
        bottom=(QualifyingRegionalCity + NationalFinalsStage1),
    )

    plt.bar(
        name[0:nobst],
        QualifyingRegionalCity,
        color=viridis.colors[4],
        bottom=(NationalFinalsStage1),
    )

    plt.bar(name[0:nobst], NationalFinalsStage1, color=viridis.colors[5])

    plt.xlabel("Obstacle")
    plt.ylabel("Times Used")
    plt.xticks(rotation=90)
    plt.title(title)
    plt.legend(rounds, loc=0, title="Round", bbox_to_anchor=(1, 1))

##### GREY #####
def grey(title):
    cmap = mat.cm.get_cmap("Greys")

    plt.bar(
        name[0:nobst],
        FinalsRegionalCity,
        color=cmap(0.8),
        bottom=(
            NationalFinalsStage2
            + SemiFinals
            + Qualifying
            + QualifyingRegionalCity
            + NationalFinalsStage1
        ),
```

```
    )

    plt.bar(
        name[0:nobst],
        NationalFinalsStage2,
        color=cmap(0.7),
        bottom=(
            SemiFinals + Qualifying + QualifyingRegionalCity + NationalFinalsStage1
        ),
    )

    plt.bar(
        name[0:nobst],
        SemiFinals,
        color=cmap(0.6),
        bottom=(Qualifying + QualifyingRegionalCity + NationalFinalsStage1),
    )

    plt.bar(
        name[0:nobst],
        Qualifying,
        color=cmap(0.5),
        bottom=(QualifyingRegionalCity + NationalFinalsStage1),
    )

    plt.bar(
        name[0:nobst],
        QualifyingRegionalCity,
        color=cmap(0.4),
        bottom=(NationalFinalsStage1),
    )

    plt.bar(name[0:nobst], NationalFinalsStage1, color=cmap(0.3))

    plt.xlabel("Obstacle")
    plt.ylabel("Times Used")
    plt.xticks(rotation=90)
    plt.title(title)
    plt.legend(rounds, loc=0, title="Round", bbox_to_anchor=(1, 1))
```

```
fig = plt.figure(figsize=(7, 7))
fig.subplots_adjust(hspace=1)
plt.subplot(211)
viridis("A")
plt.subplot(212)
default("B")
plt.savefig("Ninja_Data/colours/stacked/set_a.jpg", bbox_inches="tight", dpi=150)

fig = plt.figure(figsize=(7, 7))
fig.subplots_adjust(hspace=1)
plt.subplot(211)
default("A")
plt.subplot(212)
viridis("B")
plt.savefig("Ninja_Data/colours/stacked/set_b.jpg", bbox_inches="tight", dpi=150)

fig = plt.figure(figsize=(7, 7))
fig.subplots_adjust(hspace=1)
plt.subplot(211)
default("A")
plt.subplot(212)
grey("B")
plt.savefig("Ninja_Data/colours/stacked/set_c.jpg", bbox_inches="tight", dpi=150)

fig = plt.figure(figsize=(7, 7))
fig.subplots_adjust(hspace=1)
plt.subplot(211)
grey("A")
plt.subplot(212)
default("B")
plt.savefig("Ninja_Data/colours/stacked/set_d.jpg", bbox_inches="tight", dpi=150)

fig = plt.figure(figsize=(7, 7))
fig.subplots_adjust(hspace=1)
plt.subplot(211)
viridis("A")
plt.subplot(212)
grey("B")
plt.savefig("Ninja_Data/colours/stacked/set_e.jpg", bbox_inches="tight", dpi=150)
```

```
    fig = plt.figure(figsize=(7, 7))
    fig.subplots_adjust(hspace=1)
    plt.subplot(211)
    grey("A")
    plt.subplot(212)
    viridis("B")
    plt.savefig("Ninja_Data/colours/stacked/set_f.jpg", bbox_inches="tight", dpi=150)


barplots_stacked(4)


def barplots_grouped(nobst):
  # n.obst specifies the indices of obstacles to be plotted
  # Function returns and saves stacked bar plots with different colourings

    pos = np.arange(len(name[0:nobst]))
    barwidth = np.array([0.8 / 3, 0.8 / 4, 0.8 / 2, 0.8 / 4])

    ##### CONTROL #####
    plt.bar(pos, FinalsRegionalCity, barwidth)

    plt.bar(pos + np.array([0.8 / 3, 0, 0, 0]), NationalFinalsStage2, barwidth)

    plt.bar(pos + np.array([1.6 / 3, 0.8 / 4, 0, 0.8 / 4]), SemiFinals, barwidth)

    plt.bar(pos + np.array([2.4 / 3, 1.6 / 4, 0.8 / 4, 1.6 / 4]), Qualifying, barwidth)

    plt.bar(
        pos + np.array([1.6 / 3, 2.4 / 4, 1.6 / 4, 2.4 / 4]),
        QualifyingRegionalCity,
        barwidth,
    )

    plt.bar(
        pos + np.array([1.6 / 3, 2.4 / 4, 1.6 / 4, 3.2 / 4]),
        NationalFinalsStage1,
        barwidth,
    )
```

```
plt.xlabel("Obstacle")
plt.ylabel("Times Used")
plt.xticks(ticks=np.arange(0, 4, step=1), labels=name[0:nobst], rotation=90)
plt.legend(rounds, loc=2, title="Round", bbox_to_anchor=(1, 1))
plt.show()

##### VIRIDIS #####

viridis = cm.get_cmap("viridis", 6)

plt.bar(pos, FinalsRegionalCity, barwidth, color=viridis.colors[0])

plt.bar(
    pos + np.array([0.8 / 3, 0, 0, 0]),
    NationalFinalsStage2,
    barwidth,
    color=viridis.colors[1],
)

plt.bar(
    pos + np.array([1.6 / 3, 0.8 / 4, 0, 0.8 / 4]),
    SemiFinals,
    barwidth,
    color=viridis.colors[2],
)

plt.bar(
    pos + np.array([2.4 / 3, 1.6 / 4, 0.8 / 4, 1.6 / 4]),
    Qualifying,
    barwidth,
    color=viridis.colors[3],
)

plt.bar(
    pos + np.array([1.6 / 3, 2.4 / 4, 1.6 / 4, 2.4 / 4]),
    QualifyingRegionalCity,
    barwidth,
    color=viridis.colors[4],
)
```

```
plt.bar(
    pos + np.array([1.6 / 3, 2.4 / 4, 1.6 / 4, 3.2 / 4]),
    NationalFinalsStage1,
    barwidth,
    color=viridis.colors[5],
)

plt.xlabel("Obstacle")
plt.ylabel("Times Used")
plt.xticks(ticks=np.arange(0, 4, step=1), labels=name[0:nobst], rotation=90)
plt.legend(rounds, loc=2, title="Round", bbox_to_anchor=(1, 1))
plt.show()

##### GREY #####
cmap = mat.cm.get_cmap("Greys")

plt.bar(pos, FinalsRegionalCity, barwidth, color=cmap(0.8))

plt.bar(
    pos + np.array([0.8 / 3, 0, 0, 0]),
    NationalFinalsStage2,
    barwidth,
    color=cmap(0.7),
)

plt.bar(
    pos + np.array([1.6 / 3, 0.8 / 4, 0, 0.8 / 4]),
    SemiFinals,
    barwidth,
    color=cmap(0.6),
)

plt.bar(
    pos + np.array([2.4 / 3, 1.6 / 4, 0.8 / 4, 1.6 / 4]),
    Qualifying,
    barwidth,
    color=cmap(0.5),
)
```

```
    plt.bar(
        pos + np.array([1.6 / 3, 2.4 / 4, 1.6 / 4, 2.4 / 4]),
        QualifyingRegionalCity,
        barwidth,
        color=cmap(0.4),
    )

    plt.bar(
        pos + np.array([1.6 / 3, 2.4 / 4, 1.6 / 4, 3.2 / 4]),
        NationalFinalsStage1,
        barwidth,
        color=cmap(0.3),
    )

    plt.xlabel("Obstacle")
    plt.ylabel("Times Used")
    plt.xticks(ticks=np.arange(0, 4, step=1), labels=name[0:nobst], rotation=90)
    plt.legend(rounds, loc=2, title="Round", bbox_to_anchor=(1, 1))
    plt.show()


barplots_grouped(4)

##### sales #####

month = list(range(1, 13))

# Manually create data to match data used in R plots

salesA = [
    211.0,
    210.7,
    210.1,
    211.4,
    210.0,
    209.7,
    208.8,
    208.8,
    208.8,
```

```
    210.6,
    211.9,
    212.8,
]
salesB = [
    201.5,
    203.5,
    204.9,
    207.1,
    210.5,
    210.5,
    209.8,
    208.8,
    209.5,
    213.2,
    213.7,
    215.1,
]
salesC = [
    237.1,
    240.6,
    243.8,
    245.3,
    246.0,
    246.3,
    247.7,
    247.6,
    247.8,
    249.4,
    249.0,
    249.9,
]
salesD = [
    214.8,
    215.3,
    217.5,
    218.8,
    220.7,
    222.2,
```

```
    226.7,
    228.4,
    233.2,
    235.7,
    237.1,
    240.6,
]

viridis = cm.get_cmap("viridis", 6)
#### A and B ####
###  CONTROL ###
plt.plot(month, salesA, color=viridis.colors[0])
plt.legend(["Company A"], loc=0, title="Company", bbox_to_anchor=(1, 1))
plt.savefig("Sales/A.jpg", bbox_inches="tight", dpi=150)
plt.show()
plt.plot(month, salesB, color=viridis.colors[4])
plt.legend(["Company B"], loc=0, title="Company", bbox_to_anchor=(1, 1))
plt.savefig("Sales/B.jpg", bbox_inches="tight", dpi=150)
plt.show()
plt.plot(month, salesA, color=viridis.colors[0])
plt.plot(month, salesB, color=viridis.colors[4])
plt.legend(["Company A", "Company B"], loc=0, title="Company", bbox_to_anchor=(1, 1))
plt.savefig("Sales/AB.jpg", bbox_inches="tight", dpi=150)
plt.show()
### ZEROED ###
plt.plot(month, salesA, color=viridis.colors[0])
plt.plot(month, salesB, color=viridis.colors[4])
plt.legend(["Company A", "Company B"], loc=0, title="Company", bbox_to_anchor=(1, 1))
plt.ylim(bottom=0)
plt.savefig("Sales/AB_zero.jpg", bbox_inches="tight", dpi=150)
plt.show()

#### C and D ####
### CONTROL ###
inferno = cm.get_cmap("inferno", 15)

plt.plot(month, salesC, color=inferno.colors[7])
plt.legend(["Company C"], loc=0, title="Company", bbox_to_anchor=(1, 1))
plt.savefig("Sales/C.jpg", bbox_inches="tight", dpi=150)
```

```
plt.show()
plt.plot(month, salesD, color=inferno.colors[11])
plt.legend(["Company D"], loc=0, title="Company", bbox_to_anchor=(1, 1))
plt.savefig("Sales/D.jpg", bbox_inches="tight", dpi=150)
plt.show()
plt.plot(month, salesC, color=inferno.colors[7])
plt.plot(month, salesD, color=inferno.colors[11])
plt.legend(["Company C", "Company D"], loc=0, title="Company", bbox_to_anchor=(1, 1))
plt.savefig("Sales/CD.jpg", bbox_inches="tight", dpi=150)
plt.show()
### ZEROED ###
plt.plot(month, salesC, color=inferno.colors[7])
plt.plot(month, salesD, color=inferno.colors[11])
plt.ylim(bottom=0)
plt.legend(["Company C", "Company D"], loc=0, title="Company", bbox_to_anchor=(1, 1))
plt.savefig("Sales/CD_zeroed.jpg", bbox_inches="tight", dpi=150)
plt.show()
###
```

4.3.1

4.4 Appendix 4 - Univariate Analysis

4.5 Ninja Warrior - Part 1

4.5.1 "Approximately many times would you say the 'Salmon Ladder' was used?"

Table 4.1: Summary statistics over the whole population

	Control	Truncated	Logarithmic
N	69.0000000	69.0000000	6.600000e+01
Min.	40.0000000	40.0000000	9.000000e+00
1st Qu.	41.0000000	41.0000000	3.000000e+01
Median	41.0000000	41.0000000	3.500000e+01
Mean	41.1956522	41.3442029	1.515153e+13
3rd Qu.	42.0000000	42.0000000	4.075000e+01
Max.	45.0000000	45.0000000	1.000000e+15
Var	0.7442455	0.7575394	1.515151e+28

Table 4.2: Summary statistics over the whole population after removing outliers in the logarithmic responses

	Control	Truncated	Logarithmic
N	69.0000000	69.0000000	46.00000
Min.	40.0000000	40.0000000	15.00000
1st Qu.	41.0000000	41.0000000	34.00000
Median	41.0000000	41.0000000	35.00000
Mean	41.1956522	41.3442029	36.19565
3rd Qu.	42.0000000	42.0000000	40.00000
Max.	45.0000000	45.0000000	55.00000
Var	0.7442455	0.7575394	75.98309

Table 4.3: Summary statistics relating to the control plot responses for comparison between languages

	Whole Pop	R	Python
N	69.0000000	37.0000000	32.0000000
Min.	40.0000000	40.0000000	40.0000000
1st Qu.	41.0000000	41.0000000	40.0000000
Median	41.0000000	41.0000000	41.0000000
Mean	41.1956522	41.4729730	40.8750000
3rd Qu.	42.0000000	42.0000000	41.0000000
Max.	45.0000000	43.0000000	45.0000000
Var	0.7442455	0.4159159	0.9516129

Table 4.4: Summary statistics relating to the truncated plot responses for comparison between languages

	Whole Pop	R	Python
N	69.0000000	37.0000000	32.0000000
Min.	40.0000000	40.0000000	40.0000000
1st Qu.	41.0000000	41.0000000	41.0000000
Median	41.0000000	41.0000000	41.0000000
Mean	41.3442029	41.5540541	41.1015625
3rd Qu.	42.0000000	42.0000000	41.2500000
Max.	45.0000000	45.0000000	44.0000000
Var	0.7575394	0.7747748	0.6486265

Table 4.5: Summary statistics relating to the logarithmic plot responses for comparison between languages

	Whole Pop	R	Python
N	46.00000	37.00000	10.00
Min.	15.00000	30.00000	15.00
1st Qu.	34.00000	35.00000	35.00
Median	35.00000	35.00000	35.00
Mean	36.19565	39.72973	34.60
3rd Qu.	40.00000	40.00000	40.75
Max.	55.00000	120.00000	50.00
Var	75.98309	212.70270	131.60

Table 4.6: Summary statistics of degree subgroups relating to the logarithmic plot responses

	STEM	Humanities	Social Sci	Arts	Business	NA
N	27.00000	3.00000	3.000000e+01	2.00	4.0000	1
Min.	10.00000	9.00000	1.000000e+01	33.00	10.0000	NA
1st Qu.	22.50000	21.50000	3.400000e+01	34.75	10.3750	NA
Median	35.00000	34.00000	3.850000e+01	36.50	10.7500	NA
Mean	34.25926	26.33333	3.333337e+13	36.50	16.6250	NaN
3rd Qu.	40.00000	35.00000	5.375000e+01	38.25	17.0000	NA
Max.	120.00000	36.00000	1.000000e+15	40.00	35.0000	NA
NA's	10.00000	9.00000	1.000000e+01	33.00	10.0000	1
Var	437.12251	226.33333	3.333333e+28	24.50	150.2292	NA

Table 4.7: Degree and self-rated skills for the respondents that submitted invalid or high magnitude answers

	uni	sp_aware	obs_skl	num_skl	log_1
101	Technology	4	4	3	Don't know
121	None	4	3	3	Next to none.
102	Social Sciences	5	5	4	10^15
84	psychology	3	5	1	10^9

Table 4.8: Summary statistics for the subgroups that were shown each of the three plots first. ie. here the control statistics are only for respondents who saw the control plot first etc.

	Control First	Truncated First	Logarithmic First
N	24.000000	23.000000	15.00000
Min.	40.000000	40.000000	30.00000
1st Qu.	41.000000	41.000000	35.00000
Median	41.000000	41.250000	40.00000
Mean	41.125000	41.695652	40.20000
3rd Qu.	41.000000	42.000000	42.50000
Max.	45.000000	45.000000	55.00000
Var	1.070652	1.192935	49.88571

Table 4.9: Summary statistics to compare responses for the control plot between the whole population and those shown the control plot first

	Control Overall	Control First
N	69.0000000	24.000000
Min.	40.0000000	40.000000
1st Qu.	41.0000000	41.000000
Median	41.0000000	41.000000
Mean	41.1956522	41.125000
3rd Qu.	42.0000000	41.000000
Max.	45.0000000	45.000000
Var	0.7442455	1.070652

Table 4.10: Summary statistics to compare responses for the truncated plot between the whole population and those shown the truncated plot first

	Truncated Overall	Truncated First
N	69.0000000	23.000000
Min.	40.0000000	40.000000
1st Qu.	41.0000000	41.000000
Median	41.0000000	41.250000
Mean	41.3442029	41.695652
3rd Qu.	42.0000000	42.000000
Max.	45.0000000	45.000000
Var	0.7575394	1.192935

Table 4.11: Summary statistics to compare responses for the logarithmic plot between the whole population and those shown the logarithmic plot first

	Log Overall	Log First
N	46.00000	15.00000
Min.	15.00000	30.00000
1st Qu.	34.00000	35.00000
Median	35.00000	40.00000
Mean	36.19565	40.20000
3rd Qu.	40.00000	42.50000
Max.	55.00000	55.00000
Var	75.98309	49.88571

Table 4.12: Shapiro-Wilk test results to test for normality

Variable	P-Value
control_1	5.33578754536492e-08
truncated_1	1.25452646831151e-07
logarithmic_1	0.00018516220134098
control_1_r	1.54528619638491e-05
truncated_1_r	3.97529376553e-06
logarithmic_1_r	1.00203928506718e-10
control_1_py	4.34109928466654e-07
truncated_1_py	0.000139164254994752
logarithmic_1_py	0.0692684146080762
con_first_1	0.492
trn_first_1	0.03
log_first_1	0.914

Table 4.13: Symmetry test results to test for symmetric data

Variable	P-Value
control_1	0.048
truncated_1	0
logarithmic_1	0.236
control_1_r	0
truncated_1_r	0
logarithmic_1_r	0
control_1_py	0.39
truncated_1_py	0.342
logarithmic_1_py	0.872

Table 4.14: Sign test results for data considered non-normal and asymmetric

Variable(s)	Alternative	Null Value	P-Value
control_1	two.sided	41	0.16275565745309
truncated_1	two.sided	41	0.00393317302223295
truncated_1	greater	41	0.00196658651111647
truncated_1 and control_1	two.sided	0	0.187741558998824
control_1_r	two.sided	41	0.000144958496093972
control_1_r	greater	41	7.2479248046986e-05
truncated_1_r	two.sided	41	0.000728607177733487
truncated_1_r	greater	41	0.000364303588866743
logarithmic_1_r	two.sided	41	2.2361520677805e-05
logarithmic_1_r	less	41	1.11807603389025e-05

Table 4.15: MWW test results for data considered non-normal but symmetric, and also comparisons of asymmetric data for which the two samples are of different sizes

Variable(s)	Alternative	Null Value	P-Value
logarithmic_1	two.sided	41	0.000559493277800611
logarithmic_1	less	41	0.000279746638900306
logarithmic_1 and control_1	two.sided	0	6.9825440850011e-09
logarithmic_1 and control_1	less	0	3.49127204250055e-09
control_1_py	two.sided	41	0.166730310668163
control_1_py	greater	41	0.925860973799317
truncated_1_py	two.sided	41	0.718838518267998
truncated_1_py	greater	41	0.359419259133999
logarithmic_1_py	two.sided	41	0.119901689054139
logarithmic_1_py	less	41	0.0599508445270694
control_1_r and control_1_py	two.sided	0	0.000165813028152289
control_1_r and control_1_py	greater	0	8.29065140761444e-05
truncated_1_r and truncated_1_py	two.sided	0	0.0287523780622182
truncated_1_r and truncated_1_py	greater	0	0.0143761890311091
logarithmic_1_r and logarithmic_1_py	two.sided	0	0.989387291157526
con_first_1 and log_first_1	two.sided	0	0.0327983560266542
con_first_1 and log_first_1	greater	0	0.0163991780133271
trn_first_1 and truncated_1	two.sided	0	0.120790992088065
log_first_1 and logarithmic_1	two.sided	0	0.157715607882296

4.5.2 Approximately how much more than 'Log Grip' would you say 'Salmon Ladder' was was used?

Table 4.16: Summary statistics over the whole population

	Control	Truncated	Logarithmic
N	69.000000	69.000000	69.000000
Min.	3.000000	1.000000	1.000000
1st Qu.	4.000000	5.000000	2.000000
Median	5.000000	6.000000	3.000000
Mean	5.347826	5.869565	3.637681
3rd Qu.	6.000000	7.000000	5.000000
Max.	7.000000	7.000000	7.000000
Var	1.347826	2.026854	2.705030

Table 4.17: Summary statistics relating to the control plot responses for comparison between languages

	Whole Pop	R	Python
N	69.000000	37.000000	32.000000
Min.	3.000000	3.000000	3.000000
1st Qu.	4.000000	5.000000	4.000000
Median	5.000000	6.000000	5.000000
Mean	5.347826	5.486486	5.187500
3rd Qu.	6.000000	6.000000	6.000000
Max.	7.000000	7.000000	7.000000
Var	1.347826	1.312312	1.383064

Table 4.18: Summary statistics relating to the truncated plot responses for comparison between languages

	Whole Pop	R	Python
N	69.000000	37.000000	32.000000
Min.	1.000000	1.000000	1.000000
1st Qu.	5.000000	5.000000	5.750000
Median	6.000000	6.000000	6.000000
Mean	5.869565	5.891892	5.843750
3rd Qu.	7.000000	7.000000	7.000000
Max.	7.000000	7.000000	7.000000
Var	2.026854	1.821321	2.329637

Table 4.19: Summary statistics relating to the logarithmic plot responses for comparison between languages

	Whole Pop	R	Python
N	69.000000	37.000000	32.000000
Min.	1.000000	1.000000	1.000000
1st Qu.	2.000000	3.000000	2.000000
Median	3.000000	5.000000	3.000000
Mean	3.637681	4.216216	2.968750
3rd Qu.	5.000000	5.000000	4.000000
Max.	7.000000	7.000000	7.000000
Var	2.705030	2.507508	2.160282

Table 4.20: Summary statistics for the subgroups that were shown each of the three plots first. ie. here the control statistics are only for respondents who saw the control plot first etc.

	Control First	Truncated First	Logarithmic First
N	24.000000	23.000000	22.000000
Min.	4.000000	1.000000	1.000000
1st Qu.	4.750000	5.000000	3.000000
Median	5.500000	6.000000	5.000000
Mean	5.541667	5.565217	4.136364
3rd Qu.	7.000000	7.000000	5.750000
Max.	7.000000	7.000000	6.000000
Var	1.389493	2.166008	3.075758

Table 4.21: Summary statistics to compare responses for the control plot between the whole population and those shown the control plot first

	Control Overall	Control First
N	69.000000	24.000000
Min.	3.000000	4.000000
1st Qu.	4.000000	4.750000
Median	5.000000	5.500000
Mean	5.347826	5.541667
3rd Qu.	6.000000	7.000000
Max.	7.000000	7.000000
Var	1.347826	1.389493

Table 4.22: Summary statistics to compare responses for the truncated plot between the whole population and those shown the truncated plot first

	Truncated Overall	Truncated First
N	69.000000	23.000000
Min.	1.000000	1.000000
1st Qu.	5.000000	5.000000
Median	6.000000	6.000000
Mean	5.869565	5.565217
3rd Qu.	7.000000	7.000000
Max.	7.000000	7.000000
Var	2.026854	2.166008

Table 4.23: Summary statistics to compare responses for the logarithmic plot between the whole population and those shown the logarithmic plot first

	Log Overall	Log First
N	69.000000	22.000000
Min.	1.000000	1.000000
1st Qu.	2.000000	3.000000
Median	3.000000	5.000000
Mean	3.637681	4.136364
3rd Qu.	5.000000	5.750000
Max.	7.000000	6.000000
Var	2.705030	3.075758

Table 4.24: Shapiro-Wilk test results to test for normality

Variable	P-Value
control_2	6.24342248518355e-05
truncated_2	3.88752234860782e-09
logarithmic_2	0.00262520741543162
control_2_r	0.00391427682982241
truncated_2_r	5.93996971195915e-06
logarithmic_2_r	0.0191854417681526
control_2_py	0.00285808341673906
truncated_2_py	5.97589597973905e-06
logarithmic_2_py	0.0221068279265036

Table 4.25: Symmetry test results to test for symmetric data

Variable	P-Value
control__2	0.014
truncated__2	0.446
logarithmic__2	0
control__2__r	0.016
truncated__2__r	0.562
logarithmic__2__r	0
control__2__py	0.242
truncated__2__py	0.432
logarithmic__2__py	0.88
con__first__2	0.952
trn__first__2	0.14
log__first__2	0.038

Table 4.26: Sign test results for data considered non-normal and asymmetric

Variable(s)	Alternative	Null Value	P-Value
truncated__2 and control__2	two.sided	0	0.000191076425836156
truncated__2 and control__2	greater	0	9.55382129180782e-05
logarithmic__2 and control__2	two.sided	0	2.04742334197761e-11
logarithmic__2 and control__2	less	0	1.02371167098881e-11

Table 4.27: MWW test results for data considered non-normal but symmetric, and also comparisons of asymmetric data for which the two samples are of different sizes

Variable(s)	Alternative	Null Value	P-Value
control__2__r and control__2__py	two.sided	0	0.243909916192145
truncated__2__r and truncated__2__py	two.sided	0	0.939010927446889
logarithmic__2__r and logarithmic__2__py	two.sided	0	0.00144016567559996
logarithmic__2__r and logarithmic__2__py	greater	0	0.000720082837799979
con__first__2 and control__2	two.sided	0	0.532490797237705
trn__first__2 and truncated__2	two.sided	0	0.259404475325052
log__first__2 and logarithmic__2	two.sided	0	0.182421844461329

4.5.3 'Approximately how much more than 'Quintuple Steps' would you say 'Salmon Ladder' was used?'

Table 4.28: Summary statistics over the whole population

	Control	Truncated	Logarithmic
N	69.000000	69.000000	69.000000
Min.	1.000000	2.000000	1.000000
1st Qu.	2.000000	3.000000	1.000000
Median	3.000000	4.000000	2.000000
Mean	3.144928	3.797101	2.231884
3rd Qu.	4.000000	5.000000	3.000000
Max.	7.000000	7.000000	6.000000
Var	1.155158	1.281756	1.621910

Table 4.29: Summary statistics relating to the control plot responses for comparison between languages

	Whole Pop	R	Python
N	69.000000	37.000000	32.0000000
Min.	1.000000	2.000000	1.0000000
1st Qu.	2.000000	3.000000	2.0000000
Median	3.000000	3.000000	3.0000000
Mean	3.144928	3.378378	2.8750000
3rd Qu.	4.000000	4.000000	3.0000000
Max.	7.000000	7.000000	5.0000000
Var	1.155158	1.463964	0.6935484

Table 4.30: Summary statistics relating to the truncated plot responses for comparison between languages

	Whole Pop	R	Python
N	69.000000	37.000000	32.000000
Min.	2.000000	2.000000	2.000000
1st Qu.	3.000000	3.000000	3.000000
Median	4.000000	4.000000	4.000000
Mean	3.797101	3.810811	3.781250
3rd Qu.	5.000000	5.000000	4.250000
Max.	7.000000	6.000000	7.000000
Var	1.281756	1.324324	1.273186

Table 4.31: Summary statistics relating to the logarithmic plot responses for comparison between languages

	Whole Pop	R	Python
N	69.000000	37.000000	32.000000
Min.	1.000000	1.000000	1.000000
1st Qu.	1.000000	1.000000	1.000000
Median	2.000000	2.000000	2.000000
Mean	2.231884	2.513514	1.906250
3rd Qu.	3.000000	3.000000	2.250000
Max.	6.000000	6.000000	5.000000
Var	1.621910	1.978979	1.055443

Table 4.32: Summary statistics for the subgroups that were shown each of the three plots first. ie. here the control statistics are only for respondents who saw the control plot first etc.

	Control First	Truncated First	Logarithmic First
N	24.000000	23.0000000	22.000000
Min.	2.000000	2.0000000	1.000000
1st Qu.	2.000000	3.0000000	1.250000
Median	3.000000	3.0000000	2.500000
Mean	3.125000	3.4782609	2.681818
3rd Qu.	4.000000	4.0000000	4.000000
Max.	7.000000	5.0000000	6.000000
Var	1.418478	0.9881423	2.132035

Table 4.33: Summary statistics to compare responses for the control plot between the whole population and those shown the control plot first

	Control Overall	Control First
N	69.000000	24.000000
Min.	1.000000	2.000000
1st Qu.	2.000000	2.000000
Median	3.000000	3.000000
Mean	3.144928	3.125000
3rd Qu.	4.000000	4.000000
Max.	7.000000	7.000000
Var	1.155158	1.418478

Table 4.34: Summary statistics to compare responses for the truncated plot between the whole population and those shown the truncated plot first

	Truncated Overall	Truncated First
N	69.000000	23.0000000
Min.	2.000000	2.0000000
1st Qu.	3.000000	3.0000000
Median	4.000000	3.0000000
Mean	3.797101	3.4782609
3rd Qu.	5.000000	4.0000000
Max.	7.000000	5.0000000
Var	1.281756	0.9881423

Table 4.35: Summary statistics to compare responses for the logarithmic plot between the whole population and those shown the logarithmic plot first

	Log Overall	Log First
N	69.000000	22.000000
Min.	1.000000	1.000000
1st Qu.	1.000000	1.250000
Median	2.000000	2.500000
Mean	2.231884	2.681818
3rd Qu.	3.000000	4.000000
Max.	6.000000	6.000000
Var	1.621910	2.132035

Table 4.36: Shapiro-Wilk test results to test for normality

Variable	P-Value
control_3	3.82103913813318e-06
truncated_3	0.000334440448200324
logarithmic_3	5.94758318053857e-07
control_3_r	0.000556900253104092
truncated_3_r	0.0101814150385497
logarithmic_3_r	0.00101870786334344
control_3_py	0.00187662188224314
truncated_3_py	0.0091928840428001
logarithmic_3_py	5.96901899713831e-05

Table 4.37: Symmetry test results to test for symmetric data

Variable	P-Value
control_3	0.106
truncated_3	0.106
logarithmic_3	0.096
control_3_r	0.034
truncated_3_r	0.176
logarithmic_3_r	0.02
control_3_py	0.234
truncated_3_py	0.17
logarithmic_3_py	0.5
con_first_3	0.544
trn_first_3	0.12
log_first_3	0.544

Table 4.38: Sign test results for data considered non-normal and asymmetric

Variable(s)	Alternative	Null Value	P-Value
truncated_3 and control_3	two.sided	0	9.24769992094454e-06
truncated_3 and control_3	greater	0	4.62384996047227e-06
logarithmic_3 and control_3	two.sided	0	1.21152574195093e-07
logarithmic_3 and control_3	less	0	6.05762870975467e-08

Table 4.39: MWW results for data considered non-normal but symmetric, and also comparisons of asymmetric data for which the two samples are of different sizes

Variable(s)	Alternative	Null Value	P-Value
control_3_r and control_3_py	two.sided	0	0.105883948775384
truncated_3_r and truncated_3_py	two.sided	0	0.900788473359571
logarithmic_3_r and logarithmic_3_py	two.sided	0	0.0739448816882697
logarithmic_3_r and logarithmic_3_py	greater	0	0.0369724408441348
con_first_3 and control_3	two.sided	0	0.773815433110055
trn_first_3 and truncated_3	two.sided	0	0.267815900033396
log_first_3 and logarithmic_3	two.sided	0	0.191457924192077

4.6 Ninja Warrior - Part 2

4.6.1 "How large would you say the difference between 'Jumping spider' and 'Salmon Ladder' is?"

Table 4.40: Summary statistics over the whole population

	Default	Narrow	Wide
N	69.0000000	69.0000000	69.000000
Min.	4.0000000	3.0000000	2.000000
1st Qu.	5.0000000	6.0000000	5.000000
Median	6.0000000	6.0000000	6.000000
Mean	5.8985507	6.1159420	5.333333
3rd Qu.	7.0000000	7.0000000	6.000000
Max.	7.0000000	7.0000000	7.000000
Var	0.7689685	0.8687127	1.343137

Table 4.41: Summary statistics relating to the default plot responses for comparison between languages

	Whole Pop	R	Python
N	69.0000000	37.0000000	32.0000000
Min.	4.0000000	5.0000000	4.0000000
1st Qu.	5.0000000	6.0000000	5.0000000
Median	6.0000000	6.0000000	6.0000000
Mean	5.8985507	6.0540541	5.7187500
3rd Qu.	7.0000000	7.0000000	6.0000000
Max.	7.0000000	7.0000000	7.0000000
Var	0.7689685	0.5525526	0.9828629

Table 4.42: Summary statistics relating to the narrow plot responses for comparison between languages

	Whole Pop	R	Python
N	69.0000000	37.0000000	32.000000
Min.	3.0000000	5.0000000	3.000000
1st Qu.	6.0000000	6.0000000	5.000000
Median	6.0000000	6.0000000	6.000000
Mean	6.1159420	6.3513514	5.843750
3rd Qu.	7.0000000	7.0000000	7.000000
Max.	7.0000000	7.0000000	7.000000
Var	0.8687127	0.4564565	1.232863

Table 4.43: Summary statistics relating to the wide plot responses for comparison between languages

	Whole Pop	R	Python
N	69.000000	37.0000000	32.000000
Min.	2.000000	3.0000000	2.000000
1st Qu.	5.000000	5.0000000	4.000000
Median	6.000000	6.0000000	5.000000
Mean	5.333333	5.6756757	4.937500
3rd Qu.	6.000000	6.0000000	6.000000
Max.	7.000000	7.0000000	7.000000
Var	1.343137	0.8363363	1.673387

Table 4.44: Summary statistics for the subgroups that were shown each of the three plots first. ie. here the default statistics are only for respondents who saw the default plot first etc.

	Default First	Narrow First	Wide First
N	24.0000000	22.0000000	23.0000000
Min.	4.0000000	5.0000000	3.0000000
1st Qu.	5.0000000	5.2500000	5.0000000
Median	6.0000000	6.0000000	5.0000000
Mean	5.6666667	6.0454545	5.2173913
3rd Qu.	6.0000000	7.0000000	6.0000000
Max.	7.0000000	7.0000000	7.0000000
Var	0.8405797	0.6168831	0.9960474

Table 4.45: Summary statistics to compare responses for the default plot between the whole population and those shown the default plot first

	Default Overall	Default First
N	69.0000000	24.0000000
Min.	4.0000000	4.0000000
1st Qu.	5.0000000	5.0000000
Median	6.0000000	6.0000000
Mean	5.8985507	5.6666667
3rd Qu.	7.0000000	6.0000000
Max.	7.0000000	7.0000000
Var	0.7689685	0.8405797

Table 4.46: Summary statistics to compare responses for the narrow plot between the whole population and those shown the narrow plot first

	Narrow Overall	Narrow First
N	69.0000000	22.0000000
Min.	3.0000000	5.0000000
1st Qu.	6.0000000	5.2500000
Median	6.0000000	6.0000000
Mean	6.1159420	6.0454545
3rd Qu.	7.0000000	7.0000000
Max.	7.0000000	7.0000000
Var	0.8687127	0.6168831

Table 4.47: Summary statistics to compare responses for the wide plot between the whole population and those shown the wide plot first

	Log Overall	Log First
N	69.000000	23.0000000
Min.	2.000000	3.0000000
1st Qu.	5.000000	5.0000000
Median	6.000000	5.0000000
Mean	5.333333	5.2173913
3rd Qu.	6.000000	6.0000000
Max.	7.000000	7.0000000
Var	1.343137	0.9960474

Table 4.48: Shapiro-Wilk test results to test for normality

Variable	P-Value
default_1	1.18336641940208e-06
narrow_1	2.38320342277826e-08
wide_1	3.28753750941121e-05
default_1_r	2.11400290228233e-05
narrow_1_r	2.99082479168389e-06
wide_1_r	0.000106408994479668
default_1_py	0.000646515703735704
narrow_1_py	0.000259256787785794
wide_1_py	0.0396509726470491

Table 4.49: Symmetry test results to test for symmetric data

Variable	P-Value
default_1	0.196
narrow_1	0.272
wide_1	0
default_1_r	0.594
narrow_1_r	0.006
wide_1_r	0.00800000000000001
default_1_py	0.03
narrow_1_py	0.286
wide_1_py	0.71
def_first_1	0.184
nar_first_1	0.748
wid_first_1	0.236

Table 4.50: MWW test results for data considered non-normal but symmetric, and also comparisons of asymmetric data for which the two samples are of different sizes

Variable(s)	Alternative	Null Value	P-Value
narrow_1 and default_1	two.sided	0	0.0905049464556912
wide_1 and default_1	two.sided	0	0.00377620509450692
default_1_r and default_1_py	two.sided	0	0.204411271948635
narrow_1_r and narrow_1_py	two.sided	0	0.0553113102848234
wide_1_r and wide_1_py	two.sided	0	0.00805802429702594
def_first_1 and default_1	two.sided	0	0.241886599408831
nar_first_1 and narrow_1	two.sided	0	0.506531556500665
wid_first_1 and wide_1	two.sided	0	0.554828814459765

4.6.2 How large would you say the difference between 'Log Grip' and 'Floating Steps' is?

Table 4.51: Summary statistics over the whole population

	Default	Narrow	Wide
N	69.000000	69.000000	69.0000000
Min.	2.000000	1.000000	1.0000000
1st Qu.	2.000000	2.000000	2.0000000
Median	3.000000	3.000000	3.0000000
Mean	3.072464	3.231884	3.0724638
3rd Qu.	4.000000	4.000000	4.0000000
Max.	7.000000	7.000000	5.0000000
Var	1.303495	1.210145	0.8623188

Table 4.52: Summary statistics relating to the default plot responses for comparison between languages

	Whole Pop	R	Python
N	69.000000	37.000000	32.00000
Min.	2.000000	2.000000	2.00000
1st Qu.	2.000000	2.000000	2.00000
Median	3.000000	3.000000	3.00000
Mean	3.072464	3.054054	3.09375
3rd Qu.	4.000000	4.000000	4.00000
Max.	7.000000	6.000000	7.00000
Var	1.303495	1.219219	1.44254

Table 4.53: Summary statistics relating to the narrow plot responses for comparison between languages

	Whole Pop	R	Python
N	69.000000	37.000000	32.00000
Min.	1.000000	2.000000	1.00000
1st Qu.	2.000000	3.000000	2.00000
Median	3.000000	3.000000	3.00000
Mean	3.231884	3.324324	3.12500
3rd Qu.	4.000000	4.000000	4.00000
Max.	7.000000	6.000000	7.00000
Var	1.210145	1.114114	1.33871

Table 4.54: Summary statistics relating to the wide plot responses for comparison between languages

	Whole Pop	R	Python
N	69.0000000	37.0000000	32.0000000
Min.	1.0000000	2.0000000	1.0000000
1st Qu.	2.0000000	2.0000000	2.7500000
Median	3.0000000	3.0000000	3.0000000
Mean	3.0724638	3.1081081	3.0312500
3rd Qu.	4.0000000	4.0000000	3.2500000
Max.	5.0000000	5.0000000	5.0000000
Var	0.8623188	0.9324324	0.8054435

Table 4.55: Summary statistics for the subgroups that were shown each of the three plots first. ie. here the default statistics are only for respondents who saw the default plot first etc.

	Default First	Narrow First	Wide First
N	24.00000	22.0000000	23.0000000
Min.	2.00000	2.0000000	2.0000000
1st Qu.	2.00000	2.0000000	3.0000000
Median	3.00000	3.0000000	3.0000000
Mean	3.00000	3.0000000	3.3043478
3rd Qu.	3.00000	3.0000000	4.0000000
Max.	7.00000	5.0000000	5.0000000
Var	1.73913	0.8571429	0.8577075

Table 4.56: Summary statistics to compare responses for the default plot between the whole population and those shown the default plot first

	Default Overall	Default First
N	69.000000	24.00000
Min.	2.000000	2.00000
1st Qu.	2.000000	2.00000
Median	3.000000	3.00000
Mean	3.072464	3.00000
3rd Qu.	4.000000	3.00000
Max.	7.000000	7.00000
Var	1.303495	1.73913

Table 4.57: Summary statistics to compare responses for the narrow plot between the whole population and those shown the narrow plot first

	Narrow Overall	Narrow First
N	69.000000	22.0000000
Min.	1.000000	2.0000000
1st Qu.	2.000000	2.0000000
Median	3.000000	3.0000000
Mean	3.231884	3.0000000
3rd Qu.	4.000000	3.0000000
Max.	7.000000	5.0000000
Var	1.210145	0.8571429

Table 4.58: Summary statistics to compare responses for the wide plot between the whole population and those shown the wide plot first

	Log Overall	Log First
N	69.0000000	23.0000000
Min.	1.0000000	2.0000000
1st Qu.	2.0000000	3.0000000
Median	3.0000000	3.0000000
Mean	3.0724638	3.3043478
3rd Qu.	4.0000000	4.0000000
Max.	5.0000000	5.0000000
Var	0.8623188	0.8577075

Table 4.59: Shapiro-Wilk test results to test for normality

Variable	P-Value
default_2	1.51097204656772e-07
narrow_2	2.36142627102639e-05
wide_2	1.34175431618343e-05
default_2_r	6.10374443955261e-05
narrow_2_r	0.00174605528235681
wide_2_r	0.000241526596852485
default_2_py	8.5552767804274e-05
narrow_2_py	0.000995681637986532
wide_2_py	0.00339675280251514

Table 4.60: Symmetry test results to test for symmetric data

Variable	P-Value
default_2	0.528
narrow_2	0.008
wide_2	0.394
default_2_r	0.722
narrow_2_r	0.012
wide_2_r	0.378
default_2_py	0.562
narrow_2_py	0.402
wide_2_py	0.776
def_first_2	0.986
nar_first_2	0.902
wid_first_2	0.078

Table 4.61: MWW results for data considered non-normal but symmetric, and also comparisons of asymmetric data for which the two samples are of different sizes

Variable(s)	Alternative	Null Value	P-Value
narrow_2 and default_2	two.sided	0	0.240855973238975
wide_2 and default_2	two.sided	0	0.568090611779764
default_2_r and default_2_py	two.sided	0	0.964601351759711
narrow_2_r and narrow_2_py	two.sided	0	0.36769248037497
wide_2_r and wide_2_py	two.sided	0	0.834196225296706
def_first_2 and default_2	two.sided	0	0.557858137165115
nar_first_2 and narrow_2	two.sided	0	0.380446508309098
wid_first_2 and wide_2	two.sided	0	0.29038053847017

4.6.3 How many times would you say 'Floating Steps' were used?

Table 4.62: Summary statistics over the whole population

	Default	Narrow	Wide
N	69.0000000	69.0000000	69.000000
Min.	26.0000000	23.0000000	24.000000
1st Qu.	27.0000000	27.0000000	27.000000
Median	28.0000000	28.0000000	28.000000
Mean	27.9710145	27.3768116	28.036232
3rd Qu.	28.0000000	28.0000000	29.000000
Max.	33.0000000	29.0000000	30.000000
Var	0.9917945	0.8779838	1.958227

Table 4.63: Summary statistics relating to the default plot responses for comparison between languages

	Whole Pop	R	Python
N	69.0000000	37.000000	32.0000000
Min.	26.0000000	26.000000	27.0000000
1st Qu.	27.0000000	28.000000	27.0000000
Median	28.0000000	28.000000	28.0000000
Mean	27.9710145	27.972973	27.9687500
3rd Qu.	28.0000000	28.000000	28.0000000
Max.	33.0000000	33.000000	30.0000000
Var	0.9917945	1.082583	0.9183468

Table 4.64: Summary statistics relating to the narrow plot responses for comparison between languages

	Whole Pop	R	Python
N	69.0000000	37.0000000	32.000000
Min.	23.0000000	24.0000000	23.000000
1st Qu.	27.0000000	27.0000000	27.000000
Median	28.0000000	28.0000000	27.000000
Mean	27.3768116	27.4864865	27.250000
3rd Qu.	28.0000000	28.0000000	28.000000
Max.	29.0000000	28.0000000	29.000000
Var	0.8779838	0.7012012	1.080645

Table 4.65: Summary statistics relating to the wide plot responses for comparison between languages

	Whole Pop	R	Python
N	69.000000	37.0000000	32.000000
Min.	24.000000	25.0000000	24.000000
1st Qu.	27.000000	27.0000000	27.000000
Median	28.000000	28.0000000	28.500000
Mean	28.036232	27.8108108	28.296875
3rd Qu.	29.000000	28.0000000	30.000000
Max.	30.000000	30.0000000	30.000000
Var	1.958227	0.7687688	3.271925

Table 4.66: Summary statistics for the subgroups that were shown each of the three plots first. ie. here the default statistics are only for respondents who saw the default plot first etc.

	Default First	Narrow First	Wide First
N	24.0000000	22.000000	23.000000
Min.	26.0000000	23.000000	24.000000
1st Qu.	27.8750000	27.000000	27.250000
Median	28.0000000	28.000000	28.000000
Mean	27.9791667	27.272727	27.891304
3rd Qu.	28.0000000	28.000000	28.500000
Max.	30.0000000	28.000000	30.000000
Var	0.8799819	1.445887	1.999012

Table 4.67: Summary statistics to compare responses for the default plot between the whole population and those shown the default plot first

	Default Overall	Default First
N	69.0000000	24.0000000
Min.	26.0000000	26.0000000
1st Qu.	27.0000000	27.8750000
Median	28.0000000	28.0000000
Mean	27.9710145	27.9791667
3rd Qu.	28.0000000	28.0000000
Max.	33.0000000	30.0000000
Var	0.9917945	0.8799819

Table 4.68: Summary statistics to compare responses for the narrow plot between the whole population and those shown the narrow plot first

	Narrow Overall	Narrow First
N	69.0000000	22.000000
Min.	23.0000000	23.000000
1st Qu.	27.0000000	27.000000
Median	28.0000000	28.000000
Mean	27.3768116	27.272727
3rd Qu.	28.0000000	28.000000
Max.	29.0000000	28.000000
Var	0.8779838	1.445887

Table 4.69: Summary statistics to compare responses for the wide plot between the whole population and those shown the wide plot first

	Log Overall	Log First
N	69.000000	23.000000
Min.	24.000000	24.000000
1st Qu.	27.000000	27.250000
Median	28.000000	28.000000
Mean	28.036232	27.891304
3rd Qu.	29.000000	28.500000
Max.	30.000000	30.000000
Var	1.958227	1.999012

Table 4.70: Shapiro-Wilk test results to test for normality

Variable	P-Value
default_3	9.70159388551055e-10
narrow_3	7.81013921177679e-11
wide_3	2.33999193811929e-05
default_3_r	9.47419208813669e-09
narrow_3_r	2.50020753129336e-08
wide_3_r	5.49878698354199e-05
default_3_py	4.02423544805188e-05
narrow_3_py	7.08573834534838e-07
wide_3_py	0.000292780031974141
def_first_3	0.898
nar_first_3	0.00600000000000001
wid_first_3	0.702

Table 4.71: Symmetry test results to test for symmetric data

Variable	P-Value
default_3	0.792
narrow_3	0
wide_3	0.77
default_3_r	0.838
narrow_3_r	0
wide_3_r	0.14
default_3_py	0.85
narrow_3_py	0.146
wide_3_py	0.584

Table 4.72: Sign test results for data considered non-normal and asymmetric

Variable(s)	Alternative	Null Value	P-Value
narrow_3	two.sided	28	2.09547579288481e-09
narrow_3	less	28	1.04773789644241e-09
narrow_3 and default_3	two.sided	0	0.000179991126060486
narrow_3 and default_3	less	0	8.99955630302432e-05
narrow_3_r	two.sided	28	0.0001220703125
narrow_3_r	less	28	6.103515625e-05
narrow_3_r and narrow_3_py	two.sided	0	0.147644360229049
nar_first_3 and narrow_3	two.sided	0	0.971314295834775

Table 4.73: MWW results for data considered non-normal but symmetric, and also comparisons of asymmetric data for which the two samples are of different sizes

Variable(s)	Alternative	Null Value	P-Value
default_3	two.sided	28	0.56672532578209
wide_3	two.sided	28	0.506674178654124
wide_3 and default_3	two.sided	0	0.381239105342814
default_3_r	two.sided	28	0.401160204990545
default_3_py	two.sided	28	0.928017616229503
narrow_3_py	two.sided	28	0.000113016729448749
narrow_3_py	less	28	5.65083647243744e-05
wide_3_r	two.sided	28	0.227059310644515
wide_3_py	two.sided	28	0.225993872330457
default_3_r and default_3_py	two.sided	0	0.682329364937596
wide_3_r and wide_3_py	two.sided	0	0.0872217518584691
def_first_3 and default_3	two.sided	0	0.847459410138399
wid_first_3 and wide_3	two.sided	0	0.767108139298196

4.7 Ninja Warrior - Part 3

4.7.1 How many times would you say 'Floating Steps' were used in the Finals (Regional/City) round?

Table 4.74: Summary statistics for the stacked and grouped plot responses for the whole population

	Stacked	Grouped
N	69.00000	69.00000
Min.	9.00000	10.00000
1st Qu.	10.00000	11.00000
Median	11.00000	11.00000
Mean	14.39130	11.81159
3rd Qu.	14.00000	12.00000
Max.	35.00000	40.00000
Var	55.35934	13.33163

Table 4.75: Summary statistics relating to the stacked plot responses for comparison between languages

	Whole Population	R	Python
N	69.00000	37.00000	32.00000
Min.	9.00000	9.00000	9.00000
1st Qu.	10.00000	10.00000	10.00000
Median	11.00000	10.00000	11.50000
Mean	14.39130	13.24324	15.71875
3rd Qu.	14.00000	12.00000	16.25000
Max.	35.00000	35.00000	35.00000
Var	55.35934	46.35586	64.20867

Table 4.76: Summary statistics relating to the grouped plot responses for comparison between languages

	Whole Population	R	Python
N	69.00000	37.0000000	32.00000
Min.	10.00000	10.0000000	10.00000
1st Qu.	11.00000	11.0000000	11.00000
Median	11.00000	11.0000000	11.00000
Mean	11.81159	11.2432432	12.46875
3rd Qu.	12.00000	12.0000000	12.00000
Max.	40.00000	12.0000000	40.00000
Var	13.33163	0.4114114	27.93448

Table 4.77: Shapiro-Wilk test results to test for normality

Variable	P-Value
stacked_1	1.16490097384689e-11
grouped_1	1.77605214914826e-16
stacked_1_r	1.58520962564203e-09
grouped_1_r	4.59074801464563e-06
stacked_1_py	2.72192783910971e-06
grouped_1_py	1.83295084072365e-10
stack_first_1	0
grp_first_1	0.064

Table 4.78: Shapiro-Wilk test results to test for normality

Variable	P-Value
stacked_1	0
grouped_1	0
stacked_1_r	0
grouped_1_r	0.008
stacked_1_py	0
grouped_1_py	0
stack_first_1	0
grp_first_1	0.112

Table 4.79: Sign test results for data considered non-normal and asymmetric

Variable(s)	Alternative	Null Value	P-Value
stacked_1	two.sided	11	0.608920715404828
grouped_1	two.sided	11	0.00947530427947663
grouped_1	greater	11	0.00473765213973831
grouped_1 and stacked_1	two.sided	0	0.340890941220097
stacked_1_r	two.sided	11	0.16275565745309
stacked_1_py	two.sided	11	0.571588188409806
grouped_1_r	two.sided	11	0.0490417480468759
grouped_1_r	greater	11	0.0245208740234379
grouped_1_py	two.sided	11	0.133800506591798

Table 4.80: MWW results for data considered non-normal but symmetric, and also comparisons of asymmetric data for which the two samples are of different sizes

Variable(s)	Alternative	Null Value	P-Value
stacked_1_r and stacked_1_py	two.sided	0	0.0737237477902791
grouped_1_r and grouped_1_py	two.sided	0	0.561973829695235
stack_first_1 and stacked_1	two.sided	0	0.704585465484386
group_first_1 and grouped_1	two.sided	0	0.654750665992632

4.7.2 How many times would you say 'Log Grip' was used in the Finals (Regional/City) round

Table 4.81: Summary statistics for the stacked and grouped plot responses for the whole population

	Stacked	Grouped
N	69.00000	69.000000
Min.	6.00000	2.000000
1st Qu.	8.00000	8.000000
Median	9.00000	9.000000
Mean	10.60870	9.072464
3rd Qu.	10.00000	10.000000
Max.	25.00000	15.000000
Var	24.18286	1.979966

Table 4.82: Summary statistics relating to the stacked plot responses for comparison between languages

	Whole Population	R	Python
N	69.00000	37.00000	32.00000
Min.	6.00000	6.00000	6.00000
1st Qu.	8.00000	8.00000	8.00000
Median	9.00000	9.00000	9.00000
Mean	10.60870	10.16216	11.12500
3rd Qu.	10.00000	10.00000	10.00000
Max.	25.00000	23.00000	25.00000
Var	24.18286	18.75075	30.75806

Table 4.83: Summary statistics relating to the grouped plot responses for comparison between languages

	Whole Population	R	Python
N	69.000000	37.0000000	32.000000
Min.	2.000000	7.0000000	2.000000
1st Qu.	8.000000	9.0000000	8.000000
Median	9.000000	9.0000000	9.000000
Mean	9.072464	9.0810811	9.062500
3rd Qu.	10.000000	10.0000000	10.000000
Max.	15.000000	10.0000000	15.000000
Var	1.979966	0.6321321	3.608871

Table 4.84: Shapiro-Wilk test results to test for normality

Variable	P-Value
stacked_2	3.14087188538367e-11
grouped_2	4.94337899038139e-10
stacked_2_r	9.65363700238903e-09
grouped_2_r	7.75504301811935e-05
stacked_2_py	1.34468066051548e-06
grouped_2_py	5.65036041747099e-06
stack_first_2	0
grp_first_2	0.104

Table 4.85: Shapiro-Wilk test results to test for normality

Variable	P-Value
stacked_2	0
grouped_2	0.614
stacked_2_r	0.008
grouped_2_r	0.428
stacked_2_py	0.004
grouped_2_py	0.802
stack_first_2	0.004
grp_first_2	0.0620000000000001

Table 4.86: Sign test results for data considered non-normal and asymmetric

Variable(s)	Alternative	Null Value	P-Value
stacked_2	two.sided	9	0.791366427779181
stacked_2 and stacked_2	two.sided	0	1
stacked_2_r	two.sided	9	1
stacked_2_py	two.sided	9	0.845018982887267

Table 4.87: MWW results for data considered non-normal but symmetric, and also comparisons of asymmetric data for which the two samples are of different sizes

Variable(s)	Alternative	Null Value	P-Value
grouped_2 and stacked_2	two.sided	0	0.607570902721681
grouped_2	two.sided	9	0.50198027059749
grouped_2_r	two.sided	9	0.545383797089744
grouped_2_py	two.sided	9	0.749150109631703
grouped_2_r and grouped_2_py	two.sided	0	0.833562194836486
stacked_2_r and stacked_2_py	two.sided	0	0.810854639383072
stack_first_2 and stacked_2	two.sided	0	0.714585422209954
group_first_2 and grouped_2	two.sided	0	0.268977364340368

4.7.3 Please select the statement you feel applies to the bar chart above.

Table 4.88: Table showing responses over the whole population

	Equal	Less	More
Stacked	27	30	11
Grouped	60	5	2

Table 4.89: Table showing responses over the R subgroup

	Equal	Less	More
Stacked	11	19	6
Grouped	29	4	2

Table 4.90: Table showing responses over the Python subgroup

	Equal	Less	More
Stacked	16	11	5
Grouped	31	1	31

4.7.4 Which obstacle do you think was used MORE in Finals (Regional/City) rounds, 'Log Grip' or 'Floating Steps'?

Table 4.91: Table showing responses over the whole population

	Floating Steps	Log Grip	Both the same
Stacked	55	2	12
Grouped	57	4	8

Table 4.92: Table showing responses over the R subgroup

	Floating Steps	Log Grip	Both the same
Stacked	29	8	0
Grouped	32	1	4

Table 4.93: Table showing responses over the Python subgroup

	Floating Steps	Log Grip	Both the same
Stacked	26	2	4
Grouped	25	3	4

4.7.5 Which bar chart do you feel is easiest to read and interpret?

```
                  A  B
Whole Population 32 37
R                17 20
Python           15 17
```

Table 4.94: Description of colour pairings

Colour Set	Main Palette	Secondary Pallette
A	Viridis	Default
B	Default	Viridis
C	Default	Greyscale
D	Greyscale	Default
E	Viridis	Greyscale
F	Greyscale	Viridis

Table 4.95: Table showing responses over the whole population, separated by colour set

	A	B	A Colour	B Colour
Set A	3	9	Viridis	Default
Set B	1	11	Default	Viridis
Set C	9	1	Default	Greyscale
Set D	1	11	Greyscale	Default
Set E	8	3	Viridis	Greyscale
Set F	10	2	Greyscale	Viridis

Table 4.96: Table showing responses over the R sungroup, separated by colour set

	A	B	A Colour	B Colour
Set A	2	5	Viridis	Default
Set B	6	6	Default	Viridis
Set C	4	1	Default	Greyscale
Set D	1	6	Greyscale	Default
Set E	4	1	Viridis	Greyscale
Set F	6	1	Greyscale	Viridis

Table 4.97: Table showing responses over the Python subgroup, separated by colour set

	A	B	A Colour	B Colour
Set A	1	4	Viridis	Default
Set B	1	5	Default	Viridis
Set C	5	5	Default	Greyscale
Set D	5	5	Greyscale	Default
Set E	4	2	Viridis	Greyscale
Set F	4	1	Greyscale	Viridis

4.7.6 Which colour scheme do you find most aesthetically pleasing?

Table 4.98: Table showing responses over the whole population

	A	B	A Colour	B Colour
Set A	3	9	Viridis	Default
Set B	1	11	Default	Viridis
Set C	9	1	Default	Greyscale
Set D	1	11	Greyscale	Default
Set E	8	3	Viridis	Greyscale
Set F	10	2	Greyscale	Viridis

Table 4.99: Table showing responses over the R subgroup

	A	B	A Colour	B Colour
Set A	2	5	Viridis	Default
Set B	0	6	Default	Viridis
Set C	4	1	Default	Greyscale
Set D	1	6	Greyscale	Default
Set E	4	1	Viridis	Greyscale
Set F	6	1	Greyscale	Viridis

Table 4.100: Table showing responses over the Python subgroup

	A	B	A Colour	B Colour
Set A	1	4	Viridis	Default
Set B	1	5	Default	Viridis
Set C	5	0	Default	Greyscale
Set D	0	5	Greyscale	Default
Set E	4	2	Viridis	Greyscale
Set F	4	1	Greyscale	Viridis

4.7.7 Do you feel that one of the colour schemes makes it easier to read and interpret? If so, please select which one.

Table 4.101: Table showing responses over the whole population, separated by colour set

	None	A	B	A Colour	B Colour
Set A	3	6	3	Viridis	Default
Set B	1	11	1	Default	Viridis
Set C	9	1	9	Default	Greyscale
Set D	2	10	2	Greyscale	Default
Set E	11	11	11	Viridis	Greyscale
Set F	1	2	9	Greyscale	Viridis

Table 4.102: Table showing responses over the R subgroup, separated by colour set

	None	A	B	A Colour	B Colour
Set A	0	4	3	Viridis	Default
Set B	1	5	0	Default	Viridis
Set C	0	4	1	Default	Greyscale
Set D	0	1	6	Greyscale	Default
Set E	0	5	0	Viridis	Greyscale
Set F	1	2	4	Greyscale	Viridis

Table 4.103: Table showing responses over the Python subgroup, separated by colour set

	None	A	B	A Colour	B Colour
Set A	3	2	0	Viridis	Default
Set B	0	6	0	Default	Viridis
Set C	0	5	0	Default	Greyscale
Set D	0	1	4	Greyscale	Default
Set E	0	6	0	Viridis	Greyscale
Set F	0	0	5	Greyscale	Viridis

4.8 Sales - Part 1

Table 4.104: Statistics over the whole population

	Separate	Truncated	Zeroed
Min.	1.000000	1.000000	1.000000
1st Qu.	2.000000	2.000000	1.000000
Median	3.000000	2.000000	1.000000
Mean	2.985294	2.347826	1.376812
3rd Qu.	4.000000	3.000000	2.000000
Max.	6.000000	5.000000	3.000000

Table 4.105: Summary statistics relating to the separated plot responses for comparison between languages

	Whole Pop	R	Py
Min.	1.000000	1.000000	1.000000
1st Qu.	2.000000	2.000000	2.000000
Median	3.000000	2.000000	3.000000
Mean	2.985294	2.918919	3.064516
3rd Qu.	4.000000	4.000000	4.000000
Max.	6.000000	6.000000	6.000000

Table 4.106: Summary statistics relating to the truncated plot responses for comparison between languages

	Whole Pop	R	Py
Min.	1.000000	1.000000	1.00000
1st Qu.	2.000000	2.000000	1.75000
Median	2.000000	2.000000	2.00000
Mean	2.347826	2.459459	2.21875
3rd Qu.	3.000000	3.000000	3.00000
Max.	5.000000	4.000000	5.00000

Table 4.107: Summary statistics relating to the zeroed plot responses for comparison between languages

	Whole Pop	R	Py
Min.	1.000000	1.000000	1.00
1st Qu.	1.000000	1.000000	1.00
Median	1.000000	1.000000	1.00
Mean	1.376812	1.486487	1.25
3rd Qu.	2.000000	2.000000	1.00
Max.	3.000000	3.000000	3.00

Table 4.108: Statistics for the subgroup that saw the separated plots first compared to the whole population

	Separated - Whole Population	Separated First
Min.	1.000000	1.000000
1st Qu.	2.000000	2.000000
Median	3.000000	3.000000
Mean	2.985294	3.142857
3rd Qu.	4.000000	4.000000
Max.	6.000000	6.000000

Table 4.109: Statistics for the subgroup that saw the Truncated plot first compared to the whole population

	Truncated - Whole Population	Truncated First
Min.	1.000000	1.00
1st Qu.	2.000000	1.00
Median	2.000000	2.00
Mean	2.347826	2.00
3rd Qu.	3.000000	2.25
Max.	5.000000	4.00

Table 4.110: Statistics for the subgroup that saw the zeroed plot first compared to the whole population

	Zeroed - Whole Population	Zeroed First
Min.	1.000000	1.000000
1st Qu.	1.000000	1.000000
Median	1.000000	1.000000
Mean	1.376812	1.333333
3rd Qu.	2.000000	1.250000
Max.	3.000000	3.000000

Table 4.111: Shapiro-Wilk test results to test for normality

Variable	P-Value
trn_ab_1a	6.7809735872021e-06
zero_ab_1a	9.50105229042584e-13
trn_ab_1a_r	0.000749269347280919
zero_ab_1a_r	6.72655664829562e-08
trn_ab_1a_py	0.000342433023227564
zero_ab_1a_py	6.92541983845556e-10
trn_first_ab	0.00167817869906907
zero_first_ab	3.5205920680761e-07

Table 4.112: Shapiro-Wilk test results to test for normality

Variable	P-Value
sep_ab_1a	0.922
trn_ab_1a	0
zero_ab_1a	0
sep_ab_1a	0.926
trn_ab_1a_r	0.022
zero_ab_1a_r	0
sep_ab_1a	0.916
trn_ab_1a_py	0.102
zero_ab_1a_py	0
sep_ab_1a	0.936
trn_first_ab	0.884
zero_first_ab	0

Table 4.113: Sign test results for data considered non-normal and asymmetric

Variable(s)	Alternative	Null Value	P-Value
trn_ab_1a and zero_ab_1a	two.sided	0	1.774136393351e-13
trn_ab_1a and zero_ab_1a	greater	0	8.870681966755e-14

Table 4.114: MWW results for data considered non-normal but symmetric, and also comparisons of asymmetric data for which the two samples are of different sizes

Variable(s)	Alternative	Null Value	P-Value
trn_ab_1a and sep_ab_1a	two.sided	0	0.0114489869100508
trn_ab_1a and sep_ab_1a	less	0	0.0057244934550254
sep_ab_1a and zero_ab_1a	two.sided	0	4.84412010817655e-13
sep_ab_1a and zero_ab_1a	greater	0	2.42206005408828e-13
trn_ab_1a_r and trn_ab_1a_py	two.sided	0	0.256454315230702
zero_ab_1a_r and zero_ab_1a_py	two.sided	0	0.0911448601307526
sep_ab_1a_r and sep_ab_1a_py	two.sided	0	0.507624800317214
trn_first_ab and trn_ab_1a	two.sided	0	0.159694483903836
zero_first_ab and zero_ab_1a	two.sided	0	0.868025459965406
sep_first_ab and sep_ab_1a	two.sided	0	0.648710126691048

4.8.1 How much would you say sales of each company increased between January and December? [Company B]

Table 4.115: Statistics over the whole population

	Separate	Truncated	Zeroed
Min.	1.000000	1.000000	1.000000
1st Qu.	4.000000	4.000000	1.000000
Median	4.000000	4.000000	1.000000
Mean	4.132353	4.323529	1.376812
3rd Qu.	5.000000	5.000000	2.000000
Max.	6.000000	6.000000	3.000000

Table 4.116: Summary statistics relating to the separated plot responses for comparison between languages

	Whole Pop	R	Py
Min.	1.000000	1.000000	2.0000
1st Qu.	4.000000	3.750000	4.0000
Median	4.000000	4.000000	4.0000
Mean	4.132353	4.083333	4.1875
3rd Qu.	5.000000	5.000000	4.2500
Max.	6.000000	6.000000	6.0000

Table 4.117: Summary statistics relating to the truncated plot responses for comparison between languages

	Whole Pop	R	Py
Min.	1.000000	1.000000	1.000000
1st Qu.	4.000000	4.000000	4.000000
Median	4.000000	4.000000	4.000000
Mean	4.323529	4.405405	4.225807
3rd Qu.	5.000000	5.000000	5.500000
Max.	6.000000	6.000000	6.000000

Table 4.118: Summary statistics relating to the zeroed plot responses for comparison between languages

	Whole Pop	R	Py
Min.	1.000000	1.000000	1.00
1st Qu.	1.000000	1.000000	1.00
Median	1.000000	1.000000	1.00
Mean	1.376812	1.486487	1.25
3rd Qu.	2.000000	2.000000	1.00
Max.	3.000000	3.000000	3.00

Table 4.119: Statistics over the whole population

	Separate	Truncated	Zeroed
Min.	1.0	1.000000	1.000000
1st Qu.	4.0	4.000000	1.000000
Median	4.0	4.000000	1.000000
Mean	4.3	4.405405	1.486487
3rd Qu.	5.0	5.000000	2.000000
Max.	6.0	6.000000	3.000000

Table 4.120: Shapiro-Wilk test results to test for normality

Variable	P-Value
trn__ab__1b	2.90958251317478e-06
zero__ab__1b	9.50105229042584e-13
trn__ab__1b__r	0.000176521266596845
zero__ab__1b__r	6.72655664829562e-08
trn__ab__1b__py	0.00215780969368076
zero__ab__1b__py	6.92541983845556e-10
trn__first__ab	0.000176521266596845
zero__first__ab	6.72655664829562e-08

Table 4.121: Shapiro-Wilk test results to test for normality

Variable	P-Value
sep__ab__1b	0.21
trn__ab__1b	0.028
zero__ab__1b	0
sep__ab__1b	0.238
trn__ab__1b__r	0.104
zero__ab__1b__r	0
sep__ab__1b	0.186
trn__ab__1b__py	0.24
zero__ab__1b__py	0
sep__ab__1b	0.19
trn__first__ab	0.032
zero__first__ab	0

Table 4.122: MWW results for data considered non-normal but symmetric, and also comparisons of asymmetric data for which the two samples are of different sizes

Variable(s)	Alternative	Null Value	P-Value
trn_ab_1b and zero_ab_1b	two.sided	0	4.53967704636226e-22
trn_ab_1b and zero_ab_1b	greater	0	2.26983852318113e-22
trn_ab_1b and sep_ab_1b	two.sided	0	0.216726883336969
trn_ab_1b and sep_ab_1b	greater	0	0.108363441668485
sep_ab_1b and zero_ab_1b	two.sided	0	6.33798016153557e-23
sep_ab_1b and zero_ab_1b	less	0	1
trn_ab_1b_r and trn_ab_1b_py	two.sided	0	0.543592761305094
zero_ab_1b_r and zero_ab_1b_py	two.sided	0	0.0911448601307526
sep_ab_1b_r and sep_ab_1b_py	two.sided	0	0.984369424315388
trn_first_ab and trn_ab_1b	two.sided	0	0.739783066691471
zero_first_ab and zero_ab_1b	two.sided	0	0.370578612557692
sep_first_ab and sep_ab_1b	two.sided	0	0.3983478324832

4.8.2 How large would you say the drop in sales between April and July of Company A is?

Table 4.123: Statistics over the whole population

	Separate	Truncated	Zeroed
Min.	1.000000	1.000000	1.000000
1st Qu.	3.000000	2.000000	1.000000
Median	4.000000	3.000000	1.000000
Mean	4.057971	2.826087	1.376812
3rd Qu.	5.000000	3.000000	2.000000
Max.	7.000000	7.000000	3.000000

Table 4.124: Summary statistics relating to the separated plot responses for comparison between languages

	Whole Pop	R	Py
Min.	1.000000	1.000000	1.00000
1st Qu.	3.000000	2.000000	3.00000
Median	4.000000	4.000000	4.50000
Mean	4.057971	3.864865	4.28125
3rd Qu.	5.000000	5.000000	5.00000
Max.	7.000000	7.000000	7.00000

Table 4.125: Summary statistics relating to the truncated plot responses for comparison between languages

	Whole Pop	R	Py
Min.	1.000000	1.000000	1.00
1st Qu.	2.000000	2.000000	2.00
Median	3.000000	3.000000	3.00
Mean	2.826087	2.891892	2.75
3rd Qu.	3.000000	4.000000	3.00
Max.	7.000000	7.000000	6.00

Table 4.126: Summary statistics relating to the zeroed plot responses for comparison between languages

	Whole Pop	R	Py
Min.	1.000000	1.000000	1.00
1st Qu.	1.000000	1.000000	1.00
Median	1.000000	1.000000	1.00
Mean	1.376812	1.486487	1.25
3rd Qu.	2.000000	2.000000	1.00
Max.	3.000000	3.000000	3.00

Table 4.127: Shapiro-Wilk test results to test for normality

Variable	P-Value
trn_ab_2	5.29930577781549e-05
zero_ab_2	9.50105229042584e-13
trn_ab_2_r	0.00567304991996842
zero_ab_2_r	6.72655664829562e-08
trn_ab_2_py	0.00336626681055323
zero_ab_2_py	6.92541983845556e-10
trn_first_ab	0.00175118305553834
zero_first_ab	3.5205920680761e-07

Table 4.128: Shapiro-Wilk test results to test for normality

Variable	P-Value
sep_ab_2	0.79
trn_ab_2	0.12
zero_ab_2	0
sep_ab_2	0.818
trn_ab_2_r	0.528
zero_ab_2_r	0
sep_ab_2	0.812
trn_ab_2_py	0.0700000000000001
zero_ab_2_py	0
sep_ab_2	0.798
trn_first_ab	0.294
zero_first_ab	0

Table 4.129: Sign test results for data considered non-normal and asymmetric

Variable(s)	Alternative	Null Value	P-Value
trn__ab__2 and zero__ab__2	two.sided	0	1.9068302492542e-11
trn__ab__2 and zero__ab__2	greater	0	9.53415124627099e-12

Table 4.130: MWW results for data considered non-normal but symmetric, and also comparisons of asymmetric data for which the two samples are of different sizes

Variable(s)	Alternative	Null Value	P-Value
trn__ab__2 and sep__ab__2	two.sided	0	0.000101660231966154
trn__ab__2 and sep__ab__2	less	0	5.08301159830768e-05
sep__ab__2 and zero__ab__2	two.sided	0	1.77342530165489e-16
sep__ab__2 and zero__ab__2	greater	0	8.86712650827445e-17
trn__ab__2__r and trn__ab__2__py	two.sided	0	0.925389597527489
zero__ab__2__r and zero__ab__2__py	two.sided	0	0.0911448601307526
sep__ab__2__r and sep__ab__2__py	two.sided	0	0.34145561474615
trn__first__ab and trn__ab__2	two.sided	0	0.66777647424621
zero__first__ab and zero__ab__2	two.sided	0	0.868025459965406
sep__first__ab and sep__ab__2	two.sided	0	0.934480957603288

4.9 Sales - Part 2

4.9.1 Based on the above graph, how large would you say the difference is between the number of sales Company C makes and the number of sales Company D makes?

Table 4.131: Statistics over the whole population

	Truncated	Zeroed
Min.	2.000000	1.000000
1st Qu.	4.000000	2.000000
Median	4.000000	3.000000
Mean	4.304348	2.710145
3rd Qu.	5.000000	3.000000
Max.	7.000000	5.000000

Table 4.132: Statistics for the R subgroup

	Truncated	Zeroed
Min.	2.000000	1.000000
1st Qu.	4.000000	2.000000
Median	4.000000	2.000000
Mean	4.324324	2.648649
3rd Qu.	5.000000	3.000000
Max.	7.000000	4.000000

Table 4.133: Statistics for the Python subgroup

	Truncated	Zeroed
Min.	2.00000	1.00000
1st Qu.	3.75000	2.00000
Median	4.00000	3.00000
Mean	4.28125	2.78125
3rd Qu.	5.00000	3.00000
Max.	7.00000	5.00000

Table 4.134: Statistics for the subgroup that saw the truncated plot first compared to the whole population

	Truncated - Whole Population	Truncated First
Min.	2.000000	2.000000
1st Qu.	4.000000	4.000000
Median	4.000000	4.000000
Mean	4.304348	4.166667
3rd Qu.	5.000000	5.000000
Max.	7.000000	7.000000

Table 4.135: Statistics for the subgroup that saw the zeroed plot first compared to the whole population

	Zeroed - Whole Population	Zeroed First
Min.	1.000000	1.000000
1st Qu.	2.000000	2.000000
Median	3.000000	3.000000
Mean	2.710145	3.030303
3rd Qu.	3.000000	4.000000
Max.	5.000000	5.000000

Table 4.136: Shapiro-Wilk test results to test for normality

Variable	P-Value
trn_cd	0.000379700189309497
zero_cd	2.80855241186179e-06
trn_cd_r	0.0180545424093602
zero_cd_r	9.39903013991829e-05
trn_cd_py	0.0245137395260841
zero_cd_py	0.000814074994084011
trn_first_cd	0.00524991461743558
zero_first_cd	0.00400338047851882

Table 4.137: Shapiro-Wilk test results to test for normality

Variable	P-Value
trn_cd	0.012
zero_cd	0.028
trn_cd_r	0.024
zero_cd_r	0
trn_cd_py	0.146
zero_cd_py	0.194
trn_first_cd	0.178
zero_first_cd	0.784

Table 4.138: Sign test results for data considered non-normal and asymmetric

Variable(s)	Alternative	Null Value	P-Value
trn_cd and zero_cd	two.sided	0	8.88178419700125e-15
trn_cd and zero_cd	greater	0	4.44089209850063e-15

Table 4.139: MWW results for data considered non-normal but symmetric, and also comparisons of asymmetric data for which the two samples are of different sizes

Variable(s)	Alternative	Null Value	P-Value
trn_cd_r and trn_cd_py	two.sided	0	0.920095620069404
zero_cd_r and zero_cd_py	two.sided	0	0.616940624250758
trn_first_cd and trn_cd	two.sided	0	0.581624723113106
zero_first_cd and zero_cd	two.sided	0	0.070256975666455

4.9.2

4.10 Appendix 5 - Interview Transcripts

audio_only.m4a

Andreas [00:00:05] So, um, yeah, I'll- I'll let you let the discussion here. Tell me about what I have to do.

Katie [00:00:16] Yeah, I've just got some questions to ask you.

Andreas [00:00:20] Okay.

Katie [00:00:20] So firstly, do you have a bias towards any particular language? Either R or Python.

Andreas [00:00:25] I- I prefer a much, much more familiar with R than Python.

Katie [00:00:32] Yeah, I was kind of assuming seeing as you taught me R in second year.

Andreas [00:00:38] Yeah. So you- you were in my class?

Katie [00:00:45] Yeah, I was. A few years ago.

Andreas [00:00:49] A few years ago.

[00:00:51] Yeah.

[00:00:51] But you were you were on on Placement last year, then?

Katie [00:00:57] Yes, I was- I was on placement.

Andreas [00:00:59] So it was two years ago then.

Katie [00:01:01] Yeah. So I've kind of done, well since then yeah, I've done placement, third year and now I'm in master's year. So.

Andreas [00:01:07] So you were, you were the year that they did the Zombie Dice code?

Katie [00:01:20] Um, I really can't rememeber

Andreas [00:01:23] What I think the boardgame that they did last year was Zombie Dice, that year.

Katie [00:01:31] I remember something about crabs coming up in the exam.

Andreas [00:01:35] Okay.

Katie [00:01:35] I think.

Andreas [00:01:36] Okay, yeah, yeah.

Katie [00:01:39] Yeah, yeah, that kind of get me started on R programming though and then I did it all year on placement.

Andreas [00:01:51] Okay. I'm getting I'm getting old, I really can't put the can put your face in my class. So I'm getting old.

Katie [00:02:00] You probably see a lot of people, and also I kind of make a point of not drawing attention to myself a lot of the time.

Andreas [00:02:06] Okay. Okay.

Katie [00:02:10] But um, anyway, yeah. With the code- do you have any sort of initial comments on the code. As in like readability or functionality or whatever.

Andreas [00:02:20] I don't know. I found it very, very busy. You know, I know- I know you're using the code to change some different parameters. But for me, when I do plots, visualisation and all these things, I, I do the you know, I rarely deal with them different parameters, I'm just doing a simple plot as possible that gives out the message as clear as possible.

Katie [00:02:51] Okay.

Andreas [00:02:51] Okay?

Katie [00:02:52] Okay, yeah. And sort of next, do you feel like the code- either code could be changed in any way at all? Um just in your opinion.

Andreas [00:03:10] Not really. I mean, I have it here in front of me. It depends. It can be changed. Yes, it can be changed. And in the sense that that is no reason to. I feel like sometimes there is no reason to complicate the complicated code just for visualisation purposes. Sometimes you can achieve whatever you want with a much simpler code. Yes, but at the end of the day, to what you like to do, right? I mean, that is no wrong or right, I'm happy either way. As long I'm not the one writing it, anyone who wants to do whatever they want, they can do it essentially.

Katie [00:04:02] Okay, and then sort of, based on sort of the codes and your own knowledge, how well suited do you think each language seems to be to visualisation?

Andreas [00:04:20] Now I know- I know that in R, you use that special package, the ggplot package, which theoretically is suited for visualisation. In Python, I am not sure, it doesn't seem that you used any special package. So I'm not I'm not really sure how to compare lets's say, um, the two. I will say that er, if you- if you- if you have used the, let's say the standard packages in R, I would have told you that maybe maybe not the wisest choice to do if you want to change the parameters, but using the ggplot package at least, you know, if you use something that is specifically designed for visualisation, so it should be suited for visualisation, okay?

Katie [00:05:26] You'd hope. And sort of talking about the sort of like, the two visualisation libraries I've used, sort of ggplot and matplotlib, Which do you feel would be easier for a beginner in visualisation to learn if they had sort of an equal amount of R and Python experience?

Andreas [00:05:46] I'm biased to that.

Andreas [00:05:51] Okay, this is why I asked the first question, "Do you have a bias?"

Andreas [00:05:54] I'm biased towards R, um I think, I think it's, I don't know it's very simple that ggplot functions. It's you know, it's very simple to change the parameters and do whatever you want to. So at least, you know, I'm using it, I'm using it much more frequently, I use Python, you know, I'm not that familiar with Python because I use it a couple of times and just very simple programming exercise and have to get a feeling of it. So. Yeah, I'm definitely going to go with R.

Katie [00:06:33] OK. I'm now going to show you a couple of the plots that were generated with the code. And I just want you to sort of for each, so I'm going to show you two different ones in both R and Python versions. And I would just like to know which one you feel is, sort of gives a more publication ready output. That's my emails. Yeah, so these are the first two. So these were just sort of the- I don't know what that is doing up there, that's not suppposed to be- Yeah, these are just some of the most difficult possible, sort of no changes to the axes or anything, no changes to the scaling. And just whhich do you feel is more just publication ready straight from the output?

Andreas [00:07:38] I would say the let me- this one [circles the R version].

Katie [00:07:45] Oh, thank you. OK, um do you have any sort of reasoning behind that or just-

Andreas [00:07:52] I like I like the the the fact that you don't box it, essentially.

Katie [00:08:10] Mhmm, okay, and then these two. So these ones are, so I changed it to a logarithmic scale, and the numbers down the sides are just what, so what R and Python, respectively, output as the numberings on the scales.

Andreas [00:08:29] I'm I'm not I'm not going to- I don't I don't think the logarithm it helps in any way. So in that case, I'm I'm still going to start with this one here, because it doesn't it doesn't look like the the logarithm helps.

Katie [00:08:57] OK. And just kind of as a final thing, I think you've kind of already anwered this, but um, how much freedom do you think each language allows for sort of customisation of features like, um, scale and colouring?

Andreas [00:09:16] I think I think both of them allow the freedom to do things, is that I'm just feeling more comfortable in R essentially to this than in Python because I did it much more times. It's not that I feel R, it gives you more freedom, I do believe, as I said, if you if you go to the standard package, you're not maybe you don't have as much freedom as you have in the ggplot package or it's not as easy to change things as it is in the ggplot package. But given that you are in the ggplot package, I think you have as much freedom as you want to do, to do things. And and I do believe in Python you do have the same freedom, so it's not- In terms of freedom, you can do it in both languages. The- the most, let's say, challenging thing is probably, knowing how to do it.

Katie [00:10:20] I'm kind of I'm with you on the being biased in R, which is an interesting the three people I've interviewed so far, obviously including Vince and Nikoletta, are very biased towards Python. Yeah, so it's interesting.

Andreas [00:10:37] Okay. So what do you- do you have any other questions?

Katie [00:10:41] Sorry?

Andreas [00:10:42] Do you have any other questions or can I ask something now?

Katie [00:10:46] My last question is, do you have any other comments? So, yeah.

Andreas [00:10:49] OK, no, I just wanted to ask, what exactly are you trying to understand out here in this project?Out of curiosity.

Katie [00:11:02] I have to try and put it into words. So, you know, I kind of know what I'm trying to understand, but just wording it is er- so it's just, yeah, it's just an explana- exploration of visualisation in general really. This interview is just kind of figuring out sort of opinions on sort of which languages are good for visualisation, which languages people think are good for visualisation. Which ones? So it's just kind of a discussion of sort of the pros and cons of each language sort of like subjectively and objectively. So these are kind of going to be combined, obviously, with all that research, literature reviews and things

Andreas [00:11:45] Mhmm, okay. Good, good.

Katie [00:11:50] Hopefully it should be- like based on a couple of interviews that had so far it should be interesting to discuss.

Andreas [00:11:57] you know, can, you know, send around the report after you're done send around report. I'll be more than happy to read it. OK, OK.

Katie [00:12:09] Yeah, I will do.

Andreas [00:12:10] Good. Thank you.

Katie [00:12:12] Thank you for your time.

Andreas [00:12:14] Thank you, Kate. Good luck with everything.

Geraint Interview

Katie [00:00:00] So thank you for your time today. Um. So to start do you have any sort of personal bias towards either language used?

Geraint [00:00:14] Yeah, I use Python mostly, but I'm fairly new at R.

Katie [00:00:23] Okay.

Geraint [00:00:23] I probably understand the Python a lot more.

Katie [00:00:26] OK. And do you also have any sort of initial comments on the code, sort of anything like readability, notable similarities and differences?

Geraint [00:00:46] It's obviously a big difference in how R and python like implemented, right? I've only just got to starting to know about ggplot and I quite like the way ggplot does things. It's quite nice to do things on. Yeah, but. I dunno I'm obviously seeing the python a lot easier to read because I'm more used to it. Yeah, I don't know what else you'd like me to say.

Katie [00:01:19] Oh, no, it's just, um, I don't know, just to kind of get your general opinions on it, um, so that's perfect. And sort of kind of, well, on a similar-ish note, how do you feel like either code could be changed in any way in its implementation?

Geraint [00:01:39] OK, so speaking about the Python code first. There's a few things I've noticed that maybe I wouldn't have done, but like I'm really not that fussy about any of this. For example I know you've got, I believe, as your data is all in camel case. I don't think that's normal in Python. I think we usually do the snake case, is what they call it? And there was a couple of areas where I think you you defined local variables as the same thing as a global variable.

Katie [00:02:23] OK.

Geraint [00:02:23] Which I- I think probably technically works, obviously it works you've run this code, but I would have found that- I found that difficult to read like you were, you're defining a function viridis and then you were defining a variable viridis and I was like hang on, did you just overwrite the function?

Katie [00:02:42] Yeah, OK.

Geraint [00:02:48] And then a couple of your naming things, so nobst I didn't really know what that was, I guessing I what it is, but it's just naming things in general.

Katie [00:03:00] So I think it was just, so this is just- so I did sort of like define the variables and things further up but um, I think just because I wanted to get a general impression of the code instead of kind of, I don't know. I spoke to Vince as welland we just kind of decided like yeah, leave the sort of variable definitions out of it.

Geraint [00:03:21] OK, and this is not, God this is not a criticism.

Katie [00:03:25] No, no, it's okay, I'm not taking it as one.

Geraint [00:03:26] One of the first things I noticed. I think another thing I noticed was like when you define the colours you defined with hex code, which means you get the exact hex you want, but without looking at the plot, I don't know what that is.

Katie [00:03:43] OK, I'm going to show you a couple of plots and I'm going to ask some questions about them as well, so you'll see what the colour is.

Geraint [00:03:52] Like, I know in Python you can say, like, make this blue.

Katie [00:03:55] Yeah.

Geraint [00:03:55] And it might not be the blue you want, but at least without looking into the plot, I know what it's going to look like.

Katie [00:03:59] That's true. And sort of again with just like comparing the two languages, how well suited to visualisation, do you think each seems to be, sort of based on your own knowledge and these codes?

Geraint [00:04:20] I- I don't really like saying that one is better than the other or whatever.

Katie [00:04:25] You don't really have to say that they are better or worse, just how well suited do you think each of them are?

Geraint [00:04:30] Yeah, I think I think they are, they are both really well suited, especially with the ggplot library. I think the ggplot library does things really nice. It's when you're not used to it, it feels a little unintuitive, but then you realise that you can do like some really cool things with like one variable or something and it's really neat. I really like it. Like the I don't know what stands for the aes thing where you can just tell it what your axes are, and then soon as you've told it what the axes are just throw our data at it, and it's like oh, that's so nice.

Katie [00:05:05] Yeah, just all the, sort of, the aesthetic features of the are really easy to-.

Geraint [00:05:09] Yeah.

Katie [00:05:14] And so with Python, how well do you think that sort of suited?

Geraint [00:05:21] Um good, yeah, I think. I think matplotlib is- with, with Python you tend to like have one library, that does something really good, and matplotlib does it really, really good. And there are like two ways of doing things in matplotlib, like the object oriented way and then the just plot dot way. And I think the plot dot way, I find it a bit like un-pythonic, it's very different to how you write the rest of Python. Then the object orientated way, is making plots exactly like you write the rest of Python, which makes it really natural because-

Katie [00:06:04] Yeah okay, see I'm kind of some very heavily biasd towards R, so it's interesting.

Geraint [00:06:09] And with R- it's funny, like with R as well it's the opposite, like ggplot is very object oriented. whereas the rest of R is not.

Katie [00:06:16] Yeah, that is odd actually, so I've kind of done the opposite to each language.

Geraint [00:06:24] That's just the way people do it.

Katie [00:06:30] And sort of remember the like the two libraries. So if you had a beginner to visualisation which had an equal amount of R and Python experience, which do you think would be easier to pick up, ggplot or matplotlib?

Geraint [00:06:51] Probably matplotlib. I'm only saying that because I think the way matplotlib some of the syntax of it is very similar to the rest of Python. Like, dot bar is a function which takes two things during they're similar to how you use the rest of Python, whereas with ggplot, even though that aes thing is really nice, neat adding things to it, it's not like anything else you've seen in R, so it might be really hard to get.

Katie [00:07:19] Yeah, yeah, that's a good point. It's kind of a function within a function. And so I'm going to share a couple of plots now so you can see the hex code looks like as well. And yeah, I'll just sort of- want your opinions on, like, based on these, which do you think sort of gives the most publication ready output? Sort of.

Geraint [00:07:48] OK, so out of these two that I'm seeing. Oh, yeah, yeah, the left one? The one without the top and side.

Katie [00:07:59] Yeah, okay.

Geraint [00:08:01] But I think but I do think that's just opinion. I don't think it.

Katie [00:08:06] So this is just a very subjective interview. So I can kind of go, some people thought this, some people thought this.

Geraint [00:08:13] Yeah, that's the one I like more. I don't know what publications would like more but that's what I like more.

Katie [00:08:21] Okay, and then this one was interesting as well, because this is, obviously I changed it to a logarithmic scale. And these are just kind of the default- what the different languages default if you change it to a logarithmic scale.

Geraint [00:08:38] So I, I rather the one where it's got 10 to the 0, 10 to the 1. I think that's a lot clearer about what's going on there. I suppose with the other one, well no it is, it is the actual values, it just feels really random until you- unless you knew it was a long scale, you just you've got to really look to see.

Katie [00:09:03] Yeah.

Geraint [00:09:03] If if it's a large scale where it's 10 to the 0, 10 to the 1, it's really clear that that's a log scale

Katie [00:09:09] What I have found really interesting is that in a few of these interviews sort of a couple of you prefer the '10 to the' scale, but then that scale confused a lot of people in the survey as well.

[00:09:21] Oh really, okay.

Katie [00:09:24] I think said I shared it on the dissertation exchange. There's a lot of people that don't normally use logs and things. So I just, I just found that really interesting.

Geraint [00:09:34] Do you reckon, because maybe we've got- we're at different levels of how much we know maths, like your lecturer and PhD student. I don't know. I don't know.

Katie [00:09:47] I just , cause I've taken sort of university subjects and things, so be interesting to look at that, see who was-

Geraint [00:09:55] Neat, yeah, nice

Katie [00:09:58] But I think the only- another thing I've got actually is how much freedom do you think each language, or each library provides in terms of customisation features and things, I think you've like briefly been into it already.

Geraint [00:10:13] I, yeah, I, I think they both probably do just as much customisation, but I would find customising things in matplotlib a lot easier. I feel like, even if you don't know what's going on, you know how to change things because you know how to access the ticks or whatever whereas with ggplot you've got to have this new object on and then you've got to go look something up.

Katie [00:10:41] Yeah, okay, and just finally, have you got any sort of other comments about it at all, that you've thought of?

Geraint [00:10:52] No, not really. I can see- I don't know if these plots, especially the matplotlib ones I'm thinking of. Did you write those with Vince or did you write them yourself?

Katie [00:11:04] I wrote them myself, yeah.

Geraint [00:11:07] Okay, as it you were doing something there that Vince does a lot of, which I don't because I- so you've put all those in functions.

Katie [00:11:16] Okay, yeah.

Geraint [00:11:16] Vince loves a function, whereas I like functions for most things, plotting is one of the ones that I just don't see why we do it functions whereas Vince really likes putting them in functions, so it's just interesting that you used functions.

Katie [00:11:29] I default to doing that in R, so I think it's just kind of transitioned across. But I don't know if that because also obviously I don't know, I've obviously recently been taught by Vince computing in first year and game theory like last year, I could have subconsciously absorbed some of that.

Geraint [00:11:48] Yeah, he loves a function. I've been writing- well I've worked a lot with Vince and when we write code together he puts everything in a function and they're going like, does this really need to be a function like just makes it more complicated. But that's obviously subjective opinion as well.

Katie [00:12:04] Anyway, thank you for that, that was good.

Geraint [00:12:10] Cool, is there anything else you'd like from me?

Katie [00:12:12] No, that's everything. Thanks. That's great

Geraint [00:12:15] No problem. Good luck with the dissertation.

Katie [00:12:17] Thank you. Have a good rest of your day.

Geraint [00:12:19] Bye.

Katie [00:12:19] Bye.

Henry Transcript

Katie [00:00:01] OK. So um, can I just start by asking if you have any sort of personal bias towards either Python or R.

Henry [00:00:13] I would say I am a, I mainly use R, um sorry, I mainly use python, but very recently I've started using R for a project I'm working on.

Katie [00:00:26] OK.

Henry [00:00:27] So yeah. Yeah, I'd like to think- I'd consider- I'd like to think I'm unbiased, I'm not I am a python biased person.

Katie [00:00:34] I think everyone that programs probably has a bias towards some sort of language. Er, do you have any sort of initial comments on the code or like the readability, similarities, differences?

Henry [00:00:49] Uh, I mean, I mean yeah. I mean how about like what kind of comments are you looking for? I guess um, with-

Katie [00:01:01] Um, just anything that's popped into your head while you've been looking through it.

Henry [00:01:05] Um, well with the with the matplotlib stuff and using just the straight plt dot whatever, rather than creating a figure and an axis and adding stuff to the axis directly. That's one thing that I notice, I understand like there's reason for doing that because it's meant to be more readable.

Katie [00:01:23] Okay.

Henry [00:01:24] But like, erm, I wonder whether I, I personally I find that harder to read because it's harder to keep track of where things are kind of like how R code is, like R code, like there is just this sort of thing that's there and this is the figure that we're talking about. And then when you move onto the next thing, um, a new figure is created and. But yeah, otherwise, yeah, the code looks acceptable. I can imagine what the code does, which I think is the most important thing for both the R code and the and the Python code.

Katie [00:02:03] All right. And do you feel there's any way either of the codes could be changed in your opinion?

Henry [00:02:15] Yes, so, yeah, the the fig.ax matplotlib thing is something I would I would I would do because like where you're setting, like the y-scale, x-label and y-label and x-ticks, you can do all of that in one call if you do it the other way.

Katie [00:02:33] All right.

Henry [00:02:36] And, yeah the same with the R code there, like, where you have like xlab, ylab and then labs, you can just do one labs call, and just have all of them together, those kind of things. The code, I think functionally is all correct. But yeah, in terms of cleaning up, I guess that those will be my only comments.

Katie [00:02:55] Okay, that's helpful.

Henry [00:02:58] I don't know, is that what you wanted?

Katie [00:03:04] Yeah, just sort of like, just general comments of like, oh, in your opinion, how would you change this? Which you've like, answered well because I hadn't thought of doing them.

[00:03:13] Yeah, yeah no, functionally-

Katie [00:03:16] I'm not the most efficient programmer.

Henry [00:03:20] Functionally it all- it all looks looks good, but yeah.

Katie [00:03:26] All right, and sort of- so based on sort of your knowledge of programming and sort of like the two sort of Visualisation packages and these codes, how well suited do you think each seems to visualisation? For each language?

Henry [00:03:44] Well, that's a really good question, and I think I think matplotlib is great. I think it's great because you have a huge amount of control. It's kind of difficult to do complicated things very quickly in matplotlib, whereas ggplot2. You can have very, very complicated data in whatever form, and you can plot it in a relatively complex way very, very quickly. Like oh, I want to make a stacked bar chart and you just say, give me a bar chart and change the position to be stacked and it will do it, it'll do that for you, whereas in matplotlib you have to say, oh, no, this is where the bottom is now and you're setting them over the same x value, that kind of thing. So I think if what you want is to have, like, nice looking images straight away, R makes sense for that. I don't know how good ggplot2 is for doing the more complicated things. Whereas in matplotlib you just have control over every single element and it's very easy to access it, if you are familiar with how to access things, access objects in Python, matplotlib is a very, very sensible tool to have.

Katie [00:05:02] Okay, yeah, that's good. And sort of on, I guess, sort of a follow on note, which language- I'm going to show you a couple of the plots, and I just want to ask you which language you feel has a more publication ready output if you if you think either of them do. I'll just show you this, so these are just sort of the control plots that I haven't really done anything to I haven't done anything to the axes.

Henry [00:05:35] Yeah, um. They both, they both look good, like the things that the things that I look for, I guess, you know, good- good um relative size of like fonts and labels and stuff, I would argue that that's a little bit clearer on the right, which is the one made in, um, matplotlib.

Katie [00:06:03] Yes, yeah.

Henry [00:06:05] Whereas the one on the left, it looks cleaner and there's less- you know because there's fewer lines and stuff and, the uh, the size of the the labels and stuff is a little bit small compared to the size of the figure, I would say.

Katie [00:06:21] OK.

Henry [00:06:21] Um. But that's, again, as you said, you haven't changed anything, so.

Katie [00:06:30] And then just a couple more where, these ones are-. Oh I- apparently I opened the wrong image. Sorry, um, but yeah, the ones I'm going to show you now will be, I changed it to using a log scale um, and- yeah so this is just the output you get when you take those two plots I just showed you and just specify that you want a logarithmic scale.

Henry [00:07:08] Yeah.

Katie [00:07:14] Apparently I've closed both, just subconsciously, in the last one, sorry.

Henry [00:07:19] That's alright.

Henry [00:07:22] How are you finding working with Vince?

Katie [00:07:24] Er good, yeah, I mean, he's been my personal tutor since first year.

Henry [00:07:31] Oh nice.

Katie [00:07:31] Which is nice, yeah, so kind of. It's been good, we end up just talking about skating and things.

Henry [00:07:44] Cool, er, cool. And these are the these are the same plot.

Katie [00:07:50] So these are one- so it's the plots before that I've just applied a log scale to, so a specified in the code, um, I want a logarithmic scale. And again, it's just um, which do you feel is more sort of publication ready at the output?

Henry [00:08:05] So with this one, because it's a log scale, I would say the matplotlib one. One hundred percent, because the scale starts at zero.

Katie [00:08:16] Yeah.

Henry [00:08:17] And even though you can't have, like, log of whatever is never zero, having that I think is, it's kind of important because otherwise, otherwise it could be misleading and the same way the bar plots where you have a mean with a standard deviation error bar at the top can be a very misleading way of showing data, and in the same way I think not having the full scale is misleading as well. Also, I think it's clearer that this is isn't and is in the log scale because the the the Y ticks are shown as what they should be, whereas in the ggplot2 one, um, they're not. But you have like ten to the zero, ten to the one, and yeah obviously you could have the, the- what did you call them? Like the individual, the sub gradient things, whatever in the matplotlib one, you do get them sometimes, but um, yeah. If you were just looking at the plot, it's very clear that this is in the- the one of the right is in the log scale whereas the one on the left, it's not necessarily clear that it's on a log scale.

Katie [00:09:30] OK, yeah.

Henry [00:09:38] Cool.

Katie [00:09:38] And sort of finally. So which of the two sort of plotting libraries do you think would be easier for sort of beginner to visualisation to pick up if they had an equal amount of sort of R and Python experience?

Henry [00:10:01] Um, so as somebody who's picked up ggplot2, in the past, like two weeks, I would say that's easier because it is very clear like, you have your geoms, you have your stats and you just mush them together and it all- you just have lots of pieces that can all work together.

Katie [00:10:18] Yeah, okay.

Henry [00:10:18] Whereas with um, with matplotlib you have to have an understanding of what you're actually trying to do, like from a programmatic, programmatical point of view, like you have to think about this is the dimensions of my array and stuff like that whereas in ggplot you don't. So I'd say ggplot is easier to pick up.

Katie [00:10:36] Yeah, I think um speaking to someone who's also sort of, on placement ended up picking up ggplot in like a couple of weeks, yeah I kind of agree with that. And do you have sort of like, any other comments on the code that have been brought up or that you've just thought of?

Henry [00:11:01] Erm, not- not really. Yeah not really. It looks all- looks good.

Katie [00:11:10] Okay, thank you. Thank you so much for your time as well.

Henry [00:11:15] That's quite right. Don't worry about it

Katie [00:11:17] Hope you enjoyed the short interview?

Henry [00:11:22] No, it's been great. It's been really great. I'm glad I could provide some qualitative data for your work. But yeah.

Katie [00:11:30] Okay, I'll stop the recording now.

Nikoletta Transcript

Katie [00:00:46] Thank you for your time as well, um-.

Nikoletta [00:00:47] No worries, no worries.

Katie [00:00:49] I'm going to start by asking if you have any particular bias towards either language, R or Python.

Nikoletta [00:01:00] Yes, I am familiar with both languages, but I am very biased towards Python because it's the language I use mostly and I'm way more comfortable in Python than in R

Katie [00:01:13] OK, yeah. Um, and with regards to the code, um, I can- I can share them as well, if you want? Um, can share my screen, but did you have any initial sort of comments on the code? Sort of like any similarities or differences.

Nikoletta [00:01:31] Err, yeah similarities a thing, also is the way you the way that you wrote the code, right, the way that it's structured? Um, pretty much if you are more fluent in one language, you can always translate it very easily to the other one. Um, so there were commands I didn't necessarily know what they meant in R, but then I could look to the right to the Python code. It was- so,and I guess that means they're quite similar, they use similar syntax for several things. So, for example, some of the pre-recorded functions such as, you know, changing the scale and things like that. This is not exactly the same, but it's very similar. So, yes. And what was the question? I can find similarities in both, in those two languages.

Katie [00:02:25] Okay, so um. And do you feel any code could be changed in any way, in your opinion?

Nikoletta [00:02:38] Yes, for sure. Again, I'm not too familiar with, with R. So I don't know if I have any effici-. I guess if you argue to me that this is the most efficient way to do this, maybe I would have- I would agree. But if you go to bar plot one, for example, you are overwriting the plot each time, so that's an error, in my opinion. And at some point in both codes, you're referring to these things called names and ntimes, which are not defined anywhere past, even in the functions.

Katie [00:03:25] I think um, so, the point of them was I was kind of- so I defined them higher up, and I'd done sort of a load of data manipulation. But, erm, I spoke to Vince and he said just kind of just leave them out, just get the opinions on the-.

Nikoletta [00:03:38] Okay, fantastic. So my opinion is that the both codes right now, there are things that wouldn't work necessarily, right? Like if I run this it wouldn't work and there is some it seems to be some bugs as in the plots are writing themselves. So you're not returning each plot, so that seems to be- so from functionality, I think the functionality is a bit broken. Now, depending on how things are written, I think both languages, it's very good as in it's very obvious what is the colour is very obvious that you're changing the y labels and the x labels and things like that. So, yes.

Katie [00:04:25] OK.

Nikoletta [00:04:29] But the way your code is broken, the language themselves make them a bit more readable, I guess, and, yeah, yeah, I think that's, my humble opinion.

Katie [00:04:44] And sort of based on these, as well as about your own programming knowledge, how well suited do you think each language is to visualisation. Or sort of in particular matplotlib and ggplot.

Nikoletta [00:05:07] Um, so the problem with R is that the type of plotting that I've personally used in R was more pre-recorded things, pre-recorded things, sorry. So, for example, I used the stat package and it would put the arrows and things like that just with a function. But that made me feel back then that in order to manipulate the plot further, it was a hassle for me because it was a recorded thing and then I had to go on top and force things. Where with matplotlib when I plot, I do everything from scratch. So I have, yes, but that was because of my limited knowledge of R back then. And I do know that ggplot is a very powerful library in R. So how I gather when it comes to plotting both of the languages are quite powerful. And I know, for example, for plotting graphs R is fantastic where I haven't found something equivalent as good in Python. Plotting wise, both of the fantastic people will argue that one is better, but it has to do with bias, in my opinion. I would say matplotlib is fantastic, but I don't know R ggplot, so. Yes.

Katie [00:06:26] Alright, that's great. And this is going to be quite a biased question, but do you feel like ggplot or matplotlib would be easier for someone to learn if they had sort of similar R and python experience?

Katie [00:06:40] Yes, I am biased as in I find- I find Python to be more readable, also, as a person whose English is not the first language, I find Python to be more readable. So for any beginner starting something, I would say that Python, in my opinion, is slightly easier. Now, I'll say yeah plotting with the same.

Katie [00:07:03] All right.

Nikoletta [00:07:08] Yes, because I think with matplotlib it's very clear, right? You have the plt dot or figure axis as each command and you know what you're doing where in R you have these weird plusses and, weird.

Katie [00:07:20] Yeah

Nikoletta [00:07:24] So, yes, I think, like, oh, are we adding things to the things. Yes. Is not as clear to me.

Katie [00:07:34] Yeah I guess cause with ggplot you kind of, again Vince was saying you've got the whole- you have to learn the whole grammar of graphics and stuff, and whatever.

Nikoletta [00:07:42] Yes.

Katie [00:07:43] Yeah, all the pluses weird as well.

Nikoletta [00:07:44] All the pulsses are weird as well, yeah. And I guess you can't necessarily see- so the way that I'm set up, so right now I have bar plot 1 in both languages in front of me. And Fantastic, like, I understand that geom_bar is under ggplot, so it has to be part of ggplot, but apart from the fact that it's underneath it, there's no other indication that it's part of ggplot plot where with the Python code, I can see that is part of the plt library. You know, no library, but you know what I mean. Everything's after that plotting instance, and I'm like ah, OK, well.

Katie [00:08:23] OK, that's interesting, I hadn't thought of that, because I'm very biased towards R so this is interesting.

Nikoletta [00:08:29] Yes, but yes. So I think that's the thing. And also, you know, yeah that with Python.

Katie [00:08:38] Um, I'm going to show you a couple of the plots now, and it's just getting your opinion on whether you feel like either of them is more sort of publication- produces plots that are more sort of publication ready than the other. Um. So this is just to sort of just the default, well, not default. Sort of like control scale so that I have done nothing to alter scales on this one. Um, and I've tried to sort of other than altering the scales in the plots tried to sort of keep them as close to sort of like, I don't know, default as possible.

Nikoletta [00:09:21] Yes, makes sense. OK, I think a lot of people would argue that the plot on the left looks more beautiful, OK? I personally argue the I would- for my publication I would probably use the plot on the right just because the bar plots are not elevated. They're floating like they're floating on the left. I think if the left plot they were not floating and zero started from zero, I would probably go for the left because I think it looks more appealing.

Katie [00:09:52] Yes, OK.

Nikoletta [00:09:54] But I don't like the floating part.

Katie [00:09:57] Yeah, I see that actually. Um, I will show you these two as well. Oh, I've got the wrong one there, sorry.

Nikoletta [00:10:11] ~~That's okay, no worries. Someone is using Windows, what? No I'm joking.~~

Katie [00:10:20] ~~Huh?~~

Nikoletta [00:10:20] ~~Someone is using windows~~

Katie [00:10:23] ~~Yes, I've never got into the whole sort of Apple thing~~

Nikoletta [00:10:31] ~~Not even the linux thing, Vince didn't?~~

Katie [00:10:34] ~~No, he actually hasn't commented on my use of windows, yet.~~

Nikoletta [00:10:41] ~~Fair, I don't complain, I used to game a lot, I still game I used to game a lot on computer so windows are fantastic for that purpose.~~

Katie [00:10:49] ~~My friends are trying to sort of get me into a gaming at the moment being stuck at home.~~

Nikoletta [00:10:53] ~~You should, it's good, it relaxes you.~~

Katie [00:10:56] ~~And I've been playing Minecraft, which I don't think really counts as gaming.~~

Nikoletta [00:11:04] ~~It is. No, Minecraft it is counting as gaming. I'll take that, I'll take that.~~

Katie [00:11:10] ~~Okay, I'm a gamer.~~

Nikoletta [00:11:14] And said these two are different- they're the same.

Katie [00:11:18] Yeah, so they're both- they both use a logarithmic scale. Um, yeah again, I haven't done anything to. So this is just so I have the control plot to change the scale that a plot uses to logarithmic. And this is the output that it gives without any other sort of.

Nikoletta [00:11:42] OK, then the plot to the left. I think the plot to the right at least needed some sort of explanation, that is the logarithmic of, so yeah. So if these the standard that it returns I think you know for purpose of corrections, I think the left one on these in this case.

Katie [00:12:02] OK.

Nikoletta [00:12:02] But the plots are still floating.

Katie [00:12:02] Yes.

Nikoletta [00:12:02] And now they start from one right before they start from zero, and if you put those two plots next to each other, it would have been so confusing.

Katie [00:12:16] So this one did- this one particularly confuseda lot of people in the survey, I've done some initial analysis. And I think some people have just written, 'I don't know', instead of like a number, which is interesting.

Nikoletta [00:12:30] No it would make sense right, if you're not familiar with logarithmic, you're just like what am I looking at?

Katie [00:12:35] Exactly. And just kind of like a final thing. How much freedom do you think each language allows for sort of like customisation of features and sort of.

Nikoletta [00:12:52] A lot, they both do open source languages. Yes, R is open source language, so I think a lot of course, I guess you would need a more- if you want to use the features that are there, fine, you can do it as a user. But if you were meant to implement something, maybe you would need a bit to be more comfortable with coding. So maybe a beginner wouldn't be there. But given that you know, you have some knowledge, I think it's very easy for both languages.

Katie [00:13:27] OK.

Nikoletta [00:13:31] So, I wanna argue, I know a lot of people, a lot of people that I know that are not very good at coding, whatever that means. No, are not very good at software development, use R a lot and they're very happy with plotting. So I guess maybe in R without software development. You have more power over rewriting some of the functionality.

Katie [00:13:55] Mm hmm.

Nikoletta [00:13:56] I don't know why that is so because I again, I don't know that much about R. In Python if I want to change something, I feel like I am comfortable to do it, so I don't think is impossible, which is good, where if we're talking about other languages like matplotlib for like Matlab. That are not open source an issue. So both of them, I feel like I have a complete super power over it. Maybe in my head right now would say maybe R is a bit easier, maybe better given to the people I've spoken to.

Katie [00:14:31] Yeah. And have you got any sort of like final comments that haven't come up?

Nikoletta [00:14:45] Yes, sure, I guess let me just double check something. Yes, I think it's funny for some plots how so, bar plot 2, right, the Python code is way smaller than the R code, but then for bar plot, ay like, three, the opposite happens. I think that's I think that's I think that's funny. Like, some things take longer than one language and then they take less in others so, yes.

Katie [00:15:21] OK.

Nikoletta [00:15:22] I think that's funny little thing as in, alright, so it's do this in R, but then it'll take a few more lines in Python. But then I can argue that the bar plot is two lines.

Katie [00:15:35] OYeah, obviously that could also be down to sort of like my um-

Nikoletta [00:15:37] The way you write your code?

Katie [00:15:40] yeah, the way I write my code, the fact I have bias in R, and I'm not the most efficient at writing code anyway. Okay yeah, that is is interesting.

Nikoletta [00:15:52] But I yeah, I like both languages and I can't comment on the R aspect too much, but I know it's very beautiful, but I don't know, then again then that's another comment. I guess R comes out with out of the bag more beautiful plots. Beautiful, more appealing to the eye like they look better. That does not necessarily mean that is a good thing. When we're talking about plots, sometimes raw brutal visualisation is the best thing, and sometimes you're trying to make something very beautiful, that's when you start losing information.

Katie [00:16:33] I guess aesthetics can make it misleading sometimes.

Nikoletta [00:16:35] Yes. The floating right, like if you again, you put the plot in the one started from zero, the one started from one because they were both elevated. You would have thought that this started from the same point. So, yes, definitely R a bit more pleasant to the eye out of the bag, right, but if you put care in to do something beautiful matplotlib, I'm sure you can. But again, sometimes being appealing is not the best thing when you are talking about data visualisation.

Katie [00:17:03] OK, yeah, that's a really good point. And I'm not going to be able to unsee sort of like floating ggplot bars.

Nikoletta [00:17:13] How did you not? Like the first thing I was like, what? They're elevated.

Katie [00:17:16] I don't know, I mean on my placement- I'd never done any visualisation before and I had to learn ggplot. OK, this is how visualisations, this is what they look like.

Nikoletta [00:17:25] Everywhere.

Katie [00:17:32] I think that is everything.

Owen Transcript

Katie [00:00:01] OK. Yeah, OK, sorry, go ahead. You said you had a look at the code

Owen [00:00:10] Yes, so I had a quick look, but only a quick look because I wasn't entirely sure what I was supposed to be looking for, when I was looking at it so. Yes, so you you'll have to explain exactly what sort of, feedback on the code that you're after.

Katie [00:00:31] Yes, I will do. Now, er, do you have a bias towards like either R or Python, I'm kind of gathering that, you probably like R.

Owen [00:00:43] Yes! I'm an R person. I wrote a book on R, so...

Katie [00:00:51] Quite strong bias towards R then?

Owen [00:00:53] So having said that, I wrote my book some time ago. So I'm actually not enormously up to date on the whole tidy verse broohaha. I've actually been reading Hadley Wickham's book on the Tidey verse just recently, so. And the thing that's quite interesting is. The definite opinion on graphics in particular and how to present them and also write code for the graphics.

Katie [00:01:36] I haven't seen that might be interesting to look into, definitely.

Owen [00:01:40] Okay, so who I saw I mean, your code uses ggplot, I saw. So, yeah. So, I mean, that's very, very different to the way graphics was done in R when it first came out. There've been a couple of iterations actually, so the- ggplot has essentially its own programming language. Whereas standard R, it's got a bunch of printing primitives which are just that, you know, they let you they can print points, lines. Yeah, well, they can fill polygons and things-

Katie [00:02:25] I actually meant to try and use ggplot for the assignment I've just turned in and then realised, oh, I can't use that.

Owen [00:02:34] Ah, yes, for the for the next one I have, for the next one I have eased that restriction. I realised that because I think once you're in the MMORS programme, aren't you?

Katie [00:02:49] Yeah, yeah.

Owen [00:02:50] Because everyone who's in the MSc programme, the OR and applied stats do the stats module and first semester where they were introduced to ggplot, so I've realised, you know, most of the class is familiar with it, so probably I will let people start using it if they wish.

Katie [00:03:13] Yeah, that'll be cool.

Owen [00:03:13] But it's not actually necessary.

Katie [00:03:21] So, yeah, that's interesting.

Owen [00:03:25] Anyway, OK, so yes, so, so I'm biased.

Katie [00:03:29] I mean I'm also biased towards R which is- yeah, I've interviewed a lot of Python programmers so far.

Owen [00:03:37] And Vince hasn't converted you yet?

Katie [00:03:39] That's not yet no. I, um, I did complete a whole year of just doing R on placement, so I'm pretty much solid on that's my favourite language.

Owen [00:03:55] We'll have to get you onto Julia next, that's-.

Katie [00:03:59] Huh?

Owen [00:03:59] Or have you- you haven't come across Julia?

Katie [00:04:01] No, I haven't.

Owen [00:04:02] So that Julia's is designed for, er, numerical work. So it's fast, much faster than both R and Python, but still not too hard to use. So nothing like C, which is soul destroying.

Katie [00:04:23] Yeah, I have some engineer friends that have had to use C, and it looks...not great.

Owen [00:04:29] Yeah, it makes you worry about things we don't really need to worry about, or we don't want to maybe.

Katie [00:04:37] Yeah, I'll have a look into that as well.

Owen [00:04:41] In your spare time. Anyway, but meanwhile, I should probably let you get on with your interview.

Katie [00:04:49] So, I mean, the first question. Well, the next question is just do you have the initial comments on the code or like similarities and differences between them, sort of readability of them.

Owen [00:05:00] Oh, OK. Right.

Katie [00:05:04] Just like your subjective opinions on them.

Owen [00:05:05] So, you know. OK, well, in fact, I did have I actually had. One impression, strong impression I had with the code was to do with the size of the functions, and it's something that many people have noted, not- not just me, but small functions are much easier to digest. So so some of your code was written in, know, little sort of 10 line bites or chunks which are sort of then stuck together. And that, of course, is is much easier to to read than a great big, long things. So that's that's a sort of a general impression. It seems to me that your R code, those seemed to come in longer bits than the Python code. So I'm just just reminding myself, as I said it was only a quick look that I had yet. Yes, if I if I had to be honest, I say I probably found the Python code easier to read than the R code, in terms of what it was trying to do, but neither of them were particularly difficult. OK, and you've you clearly haven't documented it, so you're relying on the the names of things to convey the meaning.

Katie [00:07:53] I'd spoken to Vince about the documentation of things, and we kind of decided for this interview to just kind of leave the codes as they are just to sort of fully, sort of, compare the coding itself. And define variables further up and stuff.

Owen [00:08:11] Yeah, but I mean, with this code it wasn't- there wasn't a problem particularly so the meaning is generally fairly clear just from the sort of the names that you've chosen for the the functions of the various bits. So, yes. So that's a fairly vague response. Did you want something more quantitative or.

Katie [00:08:31] That's good. Yeah. And then the next follow on from that is what kind of do you feel either code could be changed in any way, in your opinion?

Owen [00:08:45] Well, I suppose unless I knew what was being used for, I wouldn't I wouldn't be inclined to to try, I suppose. I suppose the other general impression I got is this is a sort of code I would expect to be sitting behind the scenes with some other sort of interface for me to use. In which case, as long as it works, I don't really care too much, I suppose.

Katie [00:09:29] Oh, that's good. And then sort of, um, so in terms of how well suited for visualisation each language is. Just sort on, I don't know, based on these codes and your sort of knowledge of programming, how well suited you think each language is to visualising and visualisation.

Owen [00:09:54] Oh, I think it's a strength of R absolutely and the ggplot in particular, so I would say that's extremely well suited to visualisation. Well, ggplot is and again, talking about Python, I think, again, you have to specify what pack- or your library you're using for graphics. I must admit my understanding of Python graphics is that most people use something based on Matlab graphics, don't they met matplotlib or something?

Katie [00:10:39] Oh yeah, yeah

Owen [00:10:40] Yeah, so so that's a sort of a python import of a matlab. Graphics. Language, isn't it, Google graphics and so. Yeah, I mean, Matlab certainly has good graphics as well. I don't have enough experience, I suppose, with Python graphics, to have a really informed opinion. I think like most people, once you once you've learnt how to do graphics you are happy with, you tend to stick to the system you've learnt because learning a new system takes so long. So, for example, I mentioned Julia as a language, so I've learnt that fairly recently. And doing graphics in Julia, well, they've, they've imported matplotlib, but it's also fairly easy to use R to do plots for Julia output. And so that's what I do. And so instead of learning a new system.

Katie [00:11:56] I mean, that's definitely interesting.

Owen [00:11:58] Cling to the one I know.

Katie [00:12:00] So, yeah, it is interesting because I can sort of. So I might look into that a bit for the project now, and sort of discuss that a little bit. So thank you for, bringing that up.

Owen [00:12:15] Absolutely. Oh, yeah. No, I've tenaciously cling to the graphics, so I know it works.

Katie [00:12:28] Yeah, I, generally use ggplot for everything.

Owen [00:12:34] Yes, well, I am yes, I am learning, in fact, I've just been so I've just been writing a proposal to update the we've got a 10 credit module on our coding for the MSc students in the second semester. So. So I've just been updating the syllabus for that to specifically include the tidy verse and shiny, in fact. Which is why I've been actually reading up about these things just recently.

Katie [00:13:07] Is that the the statistical packages module? I'm taking out one.

Owen [00:13:14] Well, you yes, shiny won't appear until next year, I'm afraid. But I do think Andrei touches on ggplot a little bit. But what's there to think? Yes. So there'll be more of that hopefully next year.

Katie [00:13:33] OK. Yeah.

Owen [00:13:36] And so this was meant to think this. Because there are some new programming paradigms, too, that sort of have been introduced with the whole tidy verse, the idea of pipes in particular is, new to me as well. I think I've drifted off topic there a little bit.

[00:14:08] Well I kind of drifted along with you from so again just comparing the two well, comparing the two libraries this time, so ggplot and matplotlib. So for a beginner with sort of equal experience in both languages, which library do you feel will be easier to learn?

Owen [00:14:40] Ooh, I think actually probably ggplot. Was the. And again, as long as you're not trying to do anything too different from this sort of standard set of plots that it caters to. Which is fairly rich, I think, and then it's got, I think, a nice, logical system for for doing things. As soon as you want to do something a bit out of the ordinary, then life becomes much more difficult and my suspicion, not one that I've really tested, is that making ggplot do something a bit different is going to be harder than making matplotlib do something a bit different. So that's your a bridge to cross when you come to it. So for a beginner, that shouldn't be an issue. And I suppose I mean, another thing for a beginner, something that's very important, is getting a feel for what good plots look like and what's possible as well. And I think ggplot makes that quite easy. Well, relatively easy. Yeah.

Katie [00:16:12] So now I'm going to show you some of the plots that were made, and I'd just sort of like your opinion on which language you feel, the output is more sort of publication ready. So these are just so you can see from the codes I've tried to keep to sort of default scales, default scale numbers as much as possible and things. So I'll just share those with you. So this is the first- so these are just the control. So I haven't done any- haven't made any alterations to the scaling, or anything that's obviously on the left is ggplot, and on the right is the python. Or if you have no preference for either.

Owen [00:17:05] I have no preference. I feel like I'm at the optometrist, so I think these are both equally good.

Katie [00:17:14] OK. Then, yeah, just got the same thing, but for the so it's the logarithmic scale and again these are the numberings that are just to- come as the default if you change to a log scale.

Owen [00:17:34] Oh, interesting, okay, well, the matplotlib on the right. Yes, so that certainly makes it clearer that it's using a log scale, which I think is a good thing. That would be my preference out of those two.

Katie [00:17:57] And sort of just as a final thing, how much freedom do you feel each sort of language or package allows for sort of customisation of plots?

Owen [00:18:09] Well, I think I've probably already answered that I think matplotlib allows more freedom. Yeah, well, for the the less experienced user, I suppose that's the caveat.

Katie [00:18:24] And yeah, just any other comments that you've thought of?

Owen [00:18:30] Oh. Yes, I suppose one- Something that- an advantage that Matlab has with its graphics that Python and R don't have is interactive graphics. Yes, so matlab, you can you can you can actively tweak it and see the changes in front of you as opposed to you know going back and rewriting your code. And that that certainly has some advantages that can speed up production of graphs. But it's- It's not a deal breaker, I suppose, but.

Katie [00:19:31] Yeah, no that is interesting.

Owen [00:19:33] Yeah, I suppose that's the advantage of a commercial package over a free package.

Katie [00:19:42] Yes.

Owen [00:19:42] One of.

Katie [00:19:47] That's all my questions.

Owen [00:19:52] OK, excellent.

Katie [00:19:53] Well, thank you for your time.

Owen [00:19:56] You're very welcome. So if you have any follow up questions, let me know.

Vince Transcript

Vince [00:00:02] I'm trying to make sure not to not to forget to do that.

Katie [00:00:10] So I, so I need to try and figure out how I'm going to like introduce it and everything as well. Just like I'm going to screenshare a couple of codes with you.

Vince [00:00:23] might be like just saying, because a lot of people, a lot of people, the people involved, have a loose idea. So I might be worth saying like, hi, as you know, my name's Katie. I'm doing my final year you know, my MMATH project on visualisation. Thank you so much for your time. I don't think this should take more than, I said 15 minutes in email. I almost think we should reverse engineer it so it doesn't take more than 15 minutes. But if it does, it can. And then you can say I'm I haven't heard the this is recording thing. Oh no there it is OK, I'm going to record this. But that's just so that I have records that I can transcribe into my project. If there's anything you'd like me to redact for my project, let me know. And I'm just interested in your thoughts. I'm not interested in anything else. Just your thoughts. You know, just to clarify that academics can feel because one of the reasons we became one of the one of the explanations for which we became academics is because we're very good at tests. So any conversation with academic, they can sometimes feel like they're tested. So it always trying like say like. Yeah, this is subjective. I'm just interested in your thoughts.

Katie [00:01:43] So like in the actual survey, I repeated the phrase, this is purely subjective a lot.

Vince [00:01:49] Exactly, exactly. So just that people don't feel that they're being tested or whatever.

Katie [00:01:54] I realised after I said I'm going to screenshare, I'm actually thinking I might. Make a couple of different versions of the code, but with the data. Like reading in and stuff, and I'll just explain what each day.

Vince [00:02:16] I, I am also not sure if screensharing is the best way to do it. I think when you arrange the time and when you like, send the schedule and Zoom link and stuff, I think you should send I would send them as images so that they're not opening up. Don't send the code files because then they'll open up the code files and their editors might do something to the code files, write them in a different. So I would send specific I send pictures as high res as you can get them. So screenshots right. Of of the code.

Katie [00:02:43] I think I also realised that I need to run prettier and stuff on them. I haven't done that.

Vince [00:02:49] Cool. Yes, make sure you do that as well.

Katie [00:02:52] Would you be able to go through that in this meeting? Sure, sure, yeah. After I've had emails back from Andreas, Nikoletta and Henry, I'm meeting with the Nikoletta and Henry potentially tomorrow then Andreas on Thursday.

Vince [00:03:11] But, yeah, the thing is, as soon as you got that scheduled, I would send them beforehand, send them pictures only because like if Zoom falls over if the resolution is not good, etc., etc., and it just gives them time to take a look at it. They might not they might not have the pictures there. And they're ready in case they say, oh, I haven't looked at it. I can't find your email anymore. Right. But just have. But do send it so they have the opportunity to look at it beforehand. Right.

Katie [00:03:34] That's just. Just need to to figure out how to send the pictured because the codes are quite long.

Vince [00:03:42] Right. Right. OK, that's a good point. That's a good point. Yeah. So you can do don't make it too much, whatever it is, and see one picture file. So maybe it's only a portion of it or something like that. OK. But in a way like don't don't worry about too much, because any bias is like, for example, the fact that they haven't looked at them before or whatever. Is is exactly that right? But but I would send them this picture files, because if you open them in an editor, depending on what the editor is, they might have all sorts of stuff set up that immediately changes what it looks like so or even how it's formatted. OK, so just keep that in mind. Let's pretend we're doing this for real.

Katie [00:04:29] OK, so. Yeah, I'm Katie, this is for my project, everything.

Vince [00:04:36] Yep, great, great. When I think it might be that you transcribe this interview. Sorry for interrupting the process as well, but yeah. In that it's a dummy one, but it might be worth transcribing possibly if there's anything of value in it. But you'll just be very clear. This was with my project supervisor, right. do you know what I mean? Like who is.

Katie [00:04:56] Yeah.

Vince [00:04:57] Like who is familiar with the project. Right. So, so.

Katie [00:04:58] Yeah, and who has been involved in the whole process.

Vince [00:05:01] Exactly, exactly, exactly, exactly.

Katie [00:05:04] OK, so I will share my screen with you now.

Vince [00:05:08] Great.

Katie [00:05:11] So these are the two codes that I used for the visualisation. You can see on the left is the Python code. On the right is the R code.

Vince [00:05:20] Great. So I'm a lot more familiar with Python than I am with R or a lot more fluent in Python than I am with R so that's already one bit of bias in whatever I might suggest.

Katie [00:05:31] And it's just, obviously just looking for your thoughts, it's not it's not a test.

Vince [00:05:37] Yeah, I know.

Katie [00:05:38] Of how well you know each language.

Vince [00:05:39] Yeah, of course Thank you, thank you.

Katie [00:05:42] Um, the first-

Vince [00:05:42] So would you like my. Oh, Go ahead. Go ahead.

Katie [00:05:44] I was going to say first do you have any initial comments on the readability of the code, which obviously I'm going to run prettier and stuff on it, so.

Vince [00:05:51] Yes, yeah yeah, so it might change a little bit. So my my initial. Thing, looking at the Python one is I like that you've put your plotting code within - um in a functional way so within functions. That's good. I think maybe a docstring is my initial thing. Those are missing docstrings, you know, to kind of describe a little bit more what they're what they're doing. But other than that, I guess. But yeah, apart from that, that that all looks like standard matplotlib to me. So. So, so, yeah. Of course now taking a look at the R. And I suppose in a way, I have the the same comment that I like, that it's it's functional and maybe, they don't call the strings in R, but a leading overall bit of documentation at the start would be nice. But I understand as well that that could take up quite a fair bit of the screen, you know. I do love the assignment operator in R ,not that that's relevant to this discussion. The assignment operator, so, so much better than anything else. Apart from that. No, not no immediate comment. Okay I see on line 157 of the R there you've got, or 158 of the R there, you've got that loop going right. ~~You might, you can maybe hear Caitlin crying in the background. I don't know how how how. Well, my microphone is isolating the noise because.~~

Katie [00:07:31] ~~Ah no, I can't hear it.~~

Vince [00:07:33] ~~Cool. Hiya J! Hiya puppy! Oh do you want to come up? Sorry, this is interrupting our chat.~~

Katie [00:07:41] ~~That's okay.~~

Vince [00:07:41] ~~Yeah, that's Katie. Do you wanna say hi? Good job. Do you think she wants to see Caitlin? Yeah. Come here little girl, come here little girl. There you go. She's not very happy right now.~~

Katie [00:07:52] ~~She's so tiny.~~

Vince [00:08:04] ~~What are you doing Julian? Oh you don't want to change Caitlin's nappy? A lot of pens. Do you want a pen? Yeah, you want to take that pen to your Mom? OK, thank you Sweetheart. Yeah you can open it with your mom. OK, good boy, Riggs. Sorry, puppy. sorry, good boy. Sorry about that as well.~~

Katie [00:08:29] ~~It's a welcome- welcome interruption.~~

Vince [00:08:31] ~~The usual. The usual. Yes. My dog has slashed his paw so we've got a big cone as well so just then I had to like, help him get out. So I-.~~ Yeah, like my immediate thing like then like before when I'm like comparing them. Forgive me if I'm going kind of away from your question is I see a for loop on that, in the R code, right on line 158, but I don't see that in the Python code. Is that just because it's not there yet or ah, cool.

Katie [00:09:06] So I'm going to spend today documenting it all up. But this is just because um, so basically in the R code, you if you try and I think truncate it. The bars just don't show up because you have to have zero axis for bar plots in R.

Vince [00:09:33] So where is the equivalent of that in Python? Okay so it's the truncation there. OK, interesting.

Katie [00:09:35] Yeah, so for truncated plots in R [meant Python] you don't have to start at zero you can start anywhere, but in R-

Vince [00:09:41] Ahh yes, you you actually have to truncate the data itself.

Katie [00:09:45] Yeah.

Vince [00:09:47] Yeah, so I guess from a readability point of view, I think apart from that one line that I've looked at, I would suggest those are the only two things, whereas the thing on the the the R code, I think that truncability, truncation? That truncation is perhaps not as readable. I certainly would not have gotten that that was truncation just from a quick glance of the code. That would be my main, main comment.

Katie [00:10:12] ~~Again, I'm going to spend today fully, sort of like, documenting both of these.~~

Vince [00:10:17] ~~Sure, I think the thing it's kind of like stepping out of the interview for just half a second. Because there I was just giving you my thoughts like. It is also not a test of you, right? You know what I mean? Like like, you know, and and. For all the for all the interviewees know, this is not the code you used, this is just some code that you want to their thoughts on, does that make sense, right? So so you're saying you're saying that you're going to document it, you're going to change it. By all means, do that because you probably should anyway. From the point of view of the interview here, it's really about getting a thoughts on these two bits of code, so it doesn't matter really if thoughts are quite negative. OK?~~

Katie [00:11:05] ~~OK.~~

Vince [00:11:07] ~~Right. Because also they might not be right. Yeah. Does that make sense so don't- in the same way that a test of them, it's also not a test of you. Right. So don't don't feel the need to be, I'm saying defensive, not that you were defensive in a bad way, right? But you're not defending yourself. You're just asking for their thoughts.~~

Katie [00:11:25] ~~OK. Afterwards, I'll sort of discuss which codes to take pictures of things, um.~~

Vince [00:11:33] ~~Yeah.~~

Katie [00:11:35] Again you've kind of been through this, um, but how would you feel each code could be changed in your opinion, if there's anything you feel like you would change?

Vince [00:11:45] I think for me, the main one would just be for the Python at least, and for the R, you know, adding some leading documentation in the functionality. But my GGG, my GGG, my ggplot is certainly rusty, so I can't suggest much. But apart from that, the matplotlib looks like essentially the way to do it. I don't normally use those colour codes myself, whatever they are called, the hexes, are they hexes, are they hex code? I never use those I normally use like the RGB thing. So like. In a very subjective way, that kind of stands out to me a bit. But I am beginning to wonder if actually it's nicer, more precise to use those standard codes as opposed to like the RGB, you know, vectors. But yeah, no, I think I think that's that's all that's those are my main thoughts on the readability.

Katie [00:12:44] All right. And so sort of based on your own knowledge and sort of after looking at these codes, how would you feel each one is kind of suited to visualisation?

Vince [00:13:00] Yeah, that's an interesting question. I mean. So. Just from the fact essentially the reputation that, ggplot, gg- why can I not count to two? ggplot holds, that's the immediate thing that stands out to me. I'm like-

Katie [00:13:18] OK.

Vince [00:13:18] -you do visualisation with ggplot, you must really, you know, you must really know what you're doing. Because ggplot is kind of like the gold standard of of of visualisation. So it certainly seems appropriate, you know, that that that weird issue with the truncation. I would or I would wonder if that does mean maybe there's something else.

Katie [00:13:40] Yeah, that's interesting.

Vince [00:13:41] Because truncating some data is not, you know, that stands out to me. I'm like, oh, I don't know if there's a better way of doing it, but that would be a thing that I'd be like. I wonder if there's a better way of doing that. And then kind of in a boring way, in a boring way on the Python side, yeah, that looks perfectly appropriate. That looks like good old boring matplotlib, you know, but boring in a good way, you know what I mean? So I don't really have much to say about that at all.

Katie [00:14:13] OK.

Vince [00:14:15] One thing, you know, and this might almost be kind of like a slight bit of imposter syndrome because my matplotlib is so much better than my my ggplot plot, you know, matplotlib's the first time I go, I pick up to draw anything anything. Like there's a bit of me that's like, because it's ggplot it's probably better, you know, but but I don't know. I don't know. But that's all that's really just the reputation of ggplot. So you know, in Python. A lot of library, a lot of other visualisations, libraries come up and they kind of promised to be the ggplot of Python. Whereas in reality, matplotlib is perfectly fine. So those are my thoughts between the two of them. Sorry if that was a bit vague again.

Katie [00:15:10] No that's good. And then again, comparing the two visualisation libraries, which do you feel will be easier for a beginner to pick up if they have an equal amount of R and Python experience and do you have any reasons for that?

Vince [00:15:26] Yeah, yeah. So it's hard for me to answer that question from the code you're showing me. OK, well, let me try to answer that question from the code you showed me. But I think what I'm trying to say is I'm not sure I'm answering the code from the code, from the question and the code you're showing me. I think I'm answering it from my own knowledge. OK, so looking at the code you're showing me, you're seeing, we're seeing these additions. We're seeing these these specific grammatical calls, the ggplot library of scale, like continuous scale, y discrete, scale_x_discrete, et cetera, et cetera. Whereas the code on the left, the Python code, seems to be: create a bar plot, create an x label, create a y label and possibly modify it, modify the text. So I think I think the matplotlib plot code you're using is probably the more straightforward one for a beginner to pick up if they just need to draw something like that.

Katie [00:16:24] I can definitely see that, yeah.

Vince [00:16:25] But you are using there, you are using there the pyplot interface to matplotlib, which is kind of-there's actually, there's two ways of plotting in matplotlib, using matplotlib directly or using pyplot and pyplot is actually just meant to be matplotlib's interface to make it like Matlab's plotter. So it is meant to be relatively straightforward. So I think in a way whereas to plot with ggplot you essentially have to learn the grammar of graphics, right? That's kind of what ggplot is all based on, is the grammar of graphics. So I think my answer is twofold. If you just need to get a bar plot or a histogram or something like that, I think that matplotlib's pyplot is probably what I'd recommend to someone with equal knowledge of both. If you want to become very good at visualisation, then I think my advice is to learn, but not because it's easiest, but because by learning it, you'll not only learn syntax for visualisation, but you'll learn the graphics of visualisation as well. So I think, I think that's my answer.

[00:17:43] So kind of Python be more user friendly, but ggplot is kind of a good one to learn.

Vince [00:17:48] Python. Yeah, yeah. Python being more- the problem with matplotlib, right, is because it's got those two interfaces to matplotlib, which by the way, Python prides itself on not having. Python prides itself as being a language where there's only one way to do things. And it's unapologetically like, no, no, no, there's no confusion. There's one way to do something, OK? And that is indeed wonderful about the language. But matplotlib is a wonderful example where that doesn't apply because it's this pyplot and there's matplotlib. And and in a way, having those two options is both nice because you've got that like gateway drug, which is pipeline at the same time as someone who really, really only ever uses pyplot. The few times now where I have to use matplotlib, it's almost harder. Whereas if you just kind of say learn ggplot. You know, then you- Whilst it's more user friendly, I'm not sure it's more user helpful, if that makes sense, I think in the grand scheme of things. Yes, learning equips you equips you better than anything else.

Katie [00:18:48] Yes, that makes a lot of sense. I don't know for this one I'd probably have to show pictures of the plots, which language you feel provides a more publication ready output.

Vince [00:19:04] Yeah. Yes, I would do that. I would just I wouldn't send these pictures beforehand, though. I would just put them up on the screen. Here are some examples of visualisations created. Which one do you feel is more publication ready? Do you have any there?

Katie [00:19:23] Yeah, I'll just do the basic that control scaling for each one.

Vince [00:19:28] Yeah

Katie [00:19:42] So these plots were made kind of as close to the-

Vince [00:19:46] You're only sharing- hold on you're own sharing VScode with me, so I can't I can't see your other windows, I think.

Katie [00:19:55] Should stop sharing and then re-share?

Vince [00:19:57] If you click. Yes, stop share. Or I think if you just click the share button again, it brings you up the option.

Vince [00:20:07] Right.

Katie [00:20:09] So these were both made, sort of as close to- well as the default settings as possible with each of the languages. With minimal sort of editing.

Vince [00:20:24] Yeah, yeah, I, I, I mean, I recognise which ones which, but I can't say. I suppose the one on the right is slightly nicer that it doesn't have the box around it. OK, so out of the box ggplot is kind of slightly more publication ready, but I don't feel there's a big difference between the two there.

Katie [00:20:53] OK.

[00:20:57] I wonder if it might be nice, Katie, if you have a plot that's the same, but that is actually markedly different or maybe even a plot where you're getting different results from the data, whether where there was a difference between the Python and R questionnaire data. You could show one of the biggest difference there is because it might not be much statistically. I want to show you this one as well as this one.

Katie [00:21:23] OK, I was gonna say I have found that the biggest difference so far I found was in the log scales, because in Python, when you use- convert the y axis to log, that uses sort of 10 to the notation which I've left in because that's the default, what Python defaults to and a lot of the response and very off or people, a couple of people even just wrote, I don't know.

Vince [00:21:47] Perfect. That's really interesting. That's fantastic. Great. Show those, show those. Because even if you get both ones, the same response, which I expect is I don't really think there's much difference there. Right. I think that'll be an interesting thing. Go look. However, the data says this, whereas my discussion with what we're going to call them subject matter experts, I don't know. No one no one noticed such a difference. Right.

Katie [00:22:13] So you can see in the python one. This is what it defaults to when you change this to a y sc- to a log scale. Which confused a lot of people.

Vince [00:22:24] That's interesting, that's interesting that the one on the left confused a lot of people and not one on the right because I find one on the right more confusing there.

Katie [00:22:31] I mean that one did also sort of um-

Vince [00:22:32] But it would be interesting then to cross reference that.

Katie [00:22:35] Pardon?

Vince [00:22:39] If you see the cross-reference that with background, right?

Katie [00:22:42] Yeah, definitely.

Vince [00:22:43] Because some people literally ten to a number could mean nothing like they might be like 'what does that mean again?' You know what I mean. Which obviously sounds weird to us.

Katie [00:22:50] But because, obviously because, in the dissertation survey exchange thing that I went into there are a lot of sort of like sociology, psychology students there doing sort of tests who would potentially have different- reactions to that then sort of mathematicians and engineers that I sent the survey to.

Vince [00:23:09] Yeah absolutely.

Katie [00:23:14] Yeah, so in general, the two of them are kind of similarly publication ready?

Vince [00:23:20] Yeah, I certainly wouldn't- but I wonder if I'm asking that if I'm answering that question because I know how easy it is to make any potential minor modifications with each language that any journal, cause journals might be like, oh, we want a double boundary against every other graph, every graph or something like that. Right. So but I think we're looking at those they're essentially the same to me.

Katie [00:23:45] Yeah. I was also the thing is, I know we'd originally taken this question out, but kind of on that note, kind of a follow on could be: sort of how easily do you feel like you can modify various features of the plot? Or like, I don't know.

Vince [00:24:05] I think yeah, I'm not sure. Yeah, I feel that I could modify both of them relatively straightforwardly because I'm familiar with both, so. Yeah. And I mean, in both senses, either using the ggplot grammar where you just don't include something or change the option of something include or or, that's quite similar to the R one so I don't think I see much difference there.

Katie [00:24:29] Okay, um, do you have any other comments that haven't come up.

Vince [00:24:39] Nothing else, nothing else.

Katie [00:24:40] OK.

Vince [00:24:42] Nope, nope, no other comments from me. Yep.

Chapter 5

Conclusion

Chapter 6

Future Work

Future work on this topic would include running focus groups with survey respondents to obtain more in depth information regarding the survey questions, for example asking open ended questions to understand the reasoning behind responses and to open a discussion about the topic in general. This is in addition to the prior mentioned topics for further research. If not for time constraints, JavaScript D3 would also have been implemented and investigated, and so this could perhaps be involved in a future study and compared to R and Python. An investigation into the combined D3/Python approach could also be studied.

Chapter 7

References

Allaire, J J, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2020. "Rmarkdown: Dynamic Documents for R." https://github.com/rstudio/rmarkdown.

Bebeau, Christy M. 2019. ""Graphicacy for Numeracy: Review of Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures by Claus O. Wilke (2019)."." *Numeracy 12* 2. https://doi.org/https://doi.org/10.5038/1936-4660.12.2.18.

Brace, Ian. 2004. *Questionnaire Design: How to Plan, Structure and Write Survey Material for Effective Market Research*.

Brower, Cheyna Katherine. 2018. ""Too Long and Too Boring: The Effects of Survey Length and Interest on Careless Responding"." https://corescholar.libraries.wright.edu/etd_all/1918.

Chacon, Scott, and Ben Straub. 2014. *Pro Git*. Apress.

Cleveland, William S., and Robert McGill. 1984. "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods." *Journal of the American Statistical Association* 79 (387): 531–54. http://www.jstor.org/stable/2288400.

Collins. n.d. "Systematic." In *Collins.com Dictionary*. Accessed October 2020. https://www.collinsdictionary.com/dictionary/english/systematic.

Crawford, Scott D., Mick P. Couper, and Mark J. Lamias. 2001. "Web Surveys: Perceptions of Burden." *Social Science Computer Review* 19 (2): 146–62. https://doi.org/10.1177/089443930101900202.

Few, Steven. 2016. "Bar Widths and the Spaces in Between." Visual Business Intelligence Newsletter. Perceptual Edge. 2016. http://perceptualedge.com/library.php.

Garnier, Simon. 2018. *Viridis: Default Color Maps from 'Matplotlib'*.

https://CRAN.R-project.org/package=viridis.

Gastwirth, Joseph L., Yulia R. Gel, W. L. Wallace Hui, Vyacheslav Lyubchich, Weiwen Miao, and Kimihiro Noguchi. 2020. *Lawstat: Tools for Biostatistics, Public Policy, and Law.* https://CRAN.R-project.org/package=lawstat.

Hlawatsch, M., F. Sadlo, M. Burch, and D. Weiskopf. 2013. "Scale-Stack Bar Charts." *Computer Graphics Forum* 32 (3pt2): 181–90. https://doi.org/https://doi.org/10.1111/cgf.12105.

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

Huynh, Hai Dang. 2017. "Two-Dimensional Bar Charts." https://doi.org/http://dx.doi.org/10.18419/opus-9496.

Hwang, JP. 2020. "Create Impactful Data Visualizations with These Books." Online, towards data science. https://towardsdatascience.com/create-impactful-data-visualizations-with-these-books-ca9fbfecfde5.

Indratmo, Lee Howorko, Joyce Maria Boedianto, and Ben Daniel. 2018. "The Efficacy of Stacked Bar Charts in Supporting Single-Attribute and Overall-Attribute Comparisons." *Visual Informatics* 2 (3): 155–65. https://doi.org/https://doi.org/10.1016/j.visinf.2018.09.002.

LAESSIG, MATT. n.d. "ANW Obstacle History." https://data.world/ninja/anw-obstacle-history.

Lauer, Claire, and Shaun O'Brien. 2020. "The Deceptive Potential of Common Design Tactics Used in Data Visualizations." In *Proceedings of the 38th Acm International Conference on Design of Communication*. SIGDOC '20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3380851.3416762.

Motulsky, Harvey J. 2009. "The Use and Abuse of Logarithmic Axes." Online, GraphPad Software, inc. https://web.archive.org/web/20101123050530/http://graphpad.com/faq/file/1487logaxes.pdf.

Parikh, Ravi. 2014. "How to Lie with Data Visualization." *Heap - Data Stories.* https://heap.io/blog/data-stories/how-to-lie-with-data-visualization.

Peytchev, Andy, and Emilia Peytcheva. 2017. "Reduction of Measurement Error Due to Survey Length: Evaluation of the Split Questionnaire Design Approach." *Survey Research Methods* 11 (4): 361–68. https://doi.org/10.18148/srm/2017.v11i4.7145.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Revilla, Melanie, and Carlos Ochoa. 2017. "Ideal and Maximum Length for a Web Survey." *International Journal of Market Research* 59 (5): 557–65. https://doi.org/10.2501/IJMR-2017-039.

Shaffer, Jeffrey. 2016. "5 Tips on Designing Colorblind-Friendly Visualizations." https://www.tableau.com/about/blog/examining-data-viz-rules-dont-use-red-green-together.

Shneiderman, Ben. 1996. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations." University of Maryland.

Taylor, Twain. 2014. "How to Use the Information-Seeking Mantra in Cyber Intelligence Dashboards." https://www.recordedfuture.com/information-seeking-mantra/.

"Understanding Society - the Uk Household Longitudinal Study." n.d. https://www.understandingsociety.ac.uk/.

Van Rossum, Guido, and Fred L Drake Jr. 1995. *Python Reference Manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

Wiley-Interscience. 2004. *Survey Methodology (Wiley Series in Survey Methodology)*.

Wilke, Claus O. 2019. *Fundamentals of Data Visualization*. O'Reilly Media, Inc. https://clauswilke.com/dataviz/.

Yang, Brenda W., Camila Vargas Restrepo, Matthew L. Stanley, and Elizabeth J. Marsh. 2021. "Truncating Bar Graphs Persistently Misleads Viewers." *Journal of Applied Research in Memory and Cognition*. https://doi.org/https://doi.org/10.1016/j.jarmac.2020.10.002.

Züll, C. 2016. "Open-Ended Questions." *GESIS Survey Guidelines*. https://doi.org/10.15465/gesis-sg_en_002.