This chapter will discuss a basic univariate analysis of the survey results, in which we look at the summary statistics and make by-eye comparisons between groups withing the whole population.

In the chapter 3 we will delve deeper into the data, sub-setting for various groups and performing comparisons between and within these groups to ascertain whether different demographic and population factors have an effect on responses.

Ninja Warrior - Part 1

The first part of the survey consisted of showing the respondents three differently scaled bar plots representing how many times four obstacles were used throughout 10 seasons of the television show American Ninja Warrior. The three presented visualisations all showed the same raw data, but each was produced with a different y-axis scaling, in order to assess whether changing the scale in these ways affects viewer interpretation of interpreting both differences and exact bar heights. Each respondent was asked four questions; two free form answer and two multiple choice.

The use of free form answers did result in occasional non-valid answers, such as statements along the lines of “Don’t know” when a number was required. Most people also opted to write a number between 0 and 1 for the fourth question when a percentage was required.

Approximately many times would you say the ‘Salmon Ladder’ was used?

This question, the first of the survey, asked participants to type the how many times Salmon Ladder was used, based on the bar plot. The ‘correct’ answer, or rather the true height of the corresponding bar, was 42. There were three invalid answers in these responses; one for the R versions of the survey and two for the Python versions. The invalid response in the R survey was ‘41/42’, which we will take as 41.5, and the invalid Python responses were 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 we see both were entered for the logarithmically scaled plot made in Python.

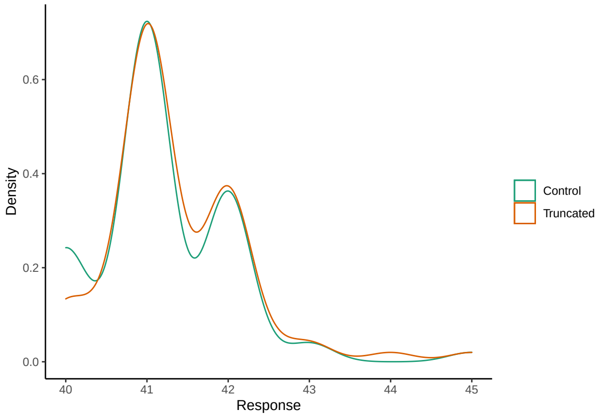
The default logarithmic scaling in Python uses standard form notation, which perhaps these two participants were less exposed to and not familiar with. Alongside these responses, the Python logarithmic scaling also elicited two answers of ‘10^15’ and ‘10^9’, which again could point towards the respondents being less familiar with this notation in addition to not correctly gauging the distancing between points. They have tried to use standard form to answer the question, but have misinterpreted that these values are many magnitudes larger than the other points on the scale. A lack of familiarity with standard form may have led the respondents to not consider the actual value of these numbers, and rather attempt to interpolate based on the indices, given the scale shows 10^0 and 10^1. Adding to this is the spacing between the two values already on the scale. These extreme responses, alongside others of 1000 and 100, potentially again due the the lack of standard form knowledge, significantly altered the statistics for the total population. These are statistically counted as outliers, but will still be important as they again provide important insights into the reactions of the respondents.

The sets of responses regarding the control and truncated plots have means 41.21 and 41.35 respectively and both have median 41; as compared to the default scaling, the truncated scale didn't much alter the respondents' perception of the bar's height. The mean of the responses for the logarithmically-scaled plot, on the other hand, was magnitudes higher at 1.493e+13, although with a median of 35; lower than the median response of the control and truncated plots responses. The variance in the responses for the logarithmic plot is also very 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 $[30, 40]$, which sits below the true value, displaying that over 50% of the observations in the total population actually underestimate the value. The control and truncated plots have small, and again very similar, variances with each having variance 0.752 and 0.753 respectively, depicting that the most of the observations lie fairly close to the respective means.

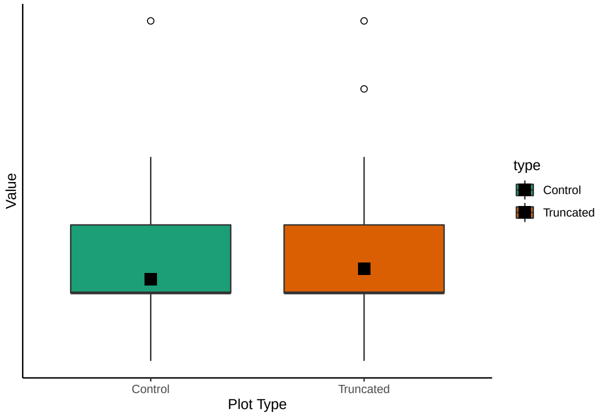
At a first glance, the means of 41.21 and 41.35 do not appear to differ significantly from the true value of 42. We can determine if this is statistically the case through univariate testing. First running Shapiro-Wilk and symmetry tests, we see that the data is not considered normally distributed or symmetric, violating required conditions for both a t-test and a Mann-Whitney-Wilcoxon test. To test this data, then, we use sign tests, with the null hypothesis that the true median of the data is equal to 42. For both of the sets of responses we obtain p-values $<< 0.05$ with 95% confidence intervals of $[41, 41]$ and $[41, 41.25]$, respectively, signifying that this difference is in fact statistically significant and we can thus infer that both the control and truncated plots resulted in a slight underestimation of the bar height. Additionally to the sign test, despite the non-normality of the data, we can also get an understanding of the results using the Central Limit Theorem by taking a series of samples from a normal distribution of the same mean and variance of our data, for which we calculate the means and then run a t test on the set of means, with the null hypothesis that the mean is equal to 42. Here we take 100 samples of size 100. Once again we achieve p-values $<< 0.05$ with 95/ confidence intervals lying just below the true value and within $[41, 42]$, reflecting the previous results.The underestimation may be as a result of the scales going up to only 40; the respondents may have seen the bar being slightly above the 40 mark, and taken this as 41 rather than the true value of 42. Here this doesn't have a significant impact, being only a single unit different, but say the scale was in £ billions for example, then an underestimation of one is significant. It may then be advised to specify the exact numerical values of the bars alongside the bars themselves in order for the interpretation to be accurate. This again could be used to either deliberately or accidentally mislead consumers, such as into believing a rival company may be doing worse than they actually are.

Compared to the control scaled plot and based on the information discussed to far, it appears that truncating the y-scale has marginally improved the respondents' ability to judge the height of the bar. This makes sense as truncating the scale results in less of a spread of values on the same sized scale, and so wider gap between each value and thus potentially allows a more accurate judging of values. There is a discrepancy here between languages in terms of the axis labeling, with the R plot being incremented in steps of 10 for both the control and truncated plots and the Python in steps of 5 for the control and steps of 2.5 for the truncated. In the multivariate analysis we will subset based on language, university degree and more to compare between these and assess whether this numbering made a difference, alongside other factor, such as the impact of standard form on the Python responses.

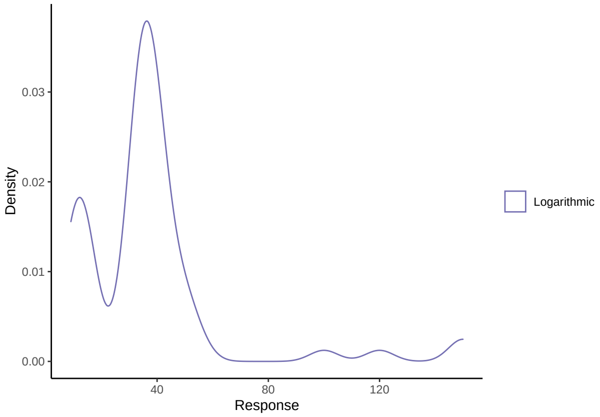
Now we consider the shape of the distributions of the responses for the control and truncated plots, with the response set for logarithmically-scaled plot excluded from this, since the high magnitude values will stretch out the x-axis and not allow us to properly see the distributions.



As expected from the above analysis the two distributions are very similar and we can see that they are, in fact, very much non-normal and asymmetrical. Similarly, the box plot of the response sets (below) displays the similarity between them, where the boxes look almost identical aside from the black square depicting the mean for the truncated plot being slightly higher than for the control.



Now consider the response distribution for the logarithmically-scaled plot, after removing the two responses of “10ˆ15” and “10ˆ9”, as well as the values at 1000.



We see that after removing the outlying values, taken as any value above 1000, the distribution is positively skewed and centres around 40 with a tail up to 150. The summary statistics for the outlier removed set shows a mean of 37.45, with a range of $[9, 150]$ and median of 35. The IQR shows that 50% of these values lie in the range $[22.5, 40]$, meaning we appear to potentially have a greater underestimation than for the other two plots, also shown by the mean.

Finally consider the three distributions together.

The distribution looks remarkably different to the other two, with much less height and a wider spread.

Overall, it appears that the use of the truncated scale had little impact on judging the height of an individual bar as compared to a control with no scale alterations, and if anything marginally improved the respondents' ability to judge the height. It has also been observed that the scale by default ending at 40 while the bar height is at 42 leads to some underestimation in the height for both the control and truncated bar plots.

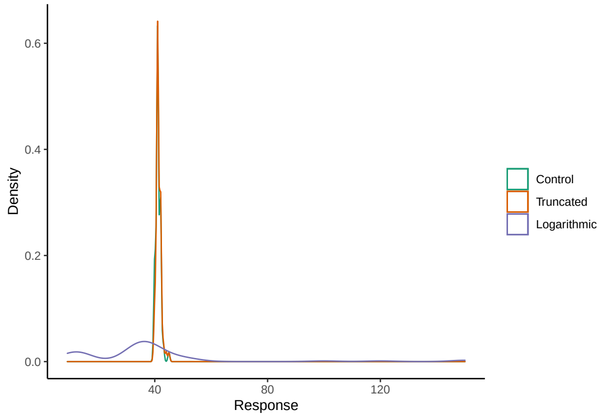


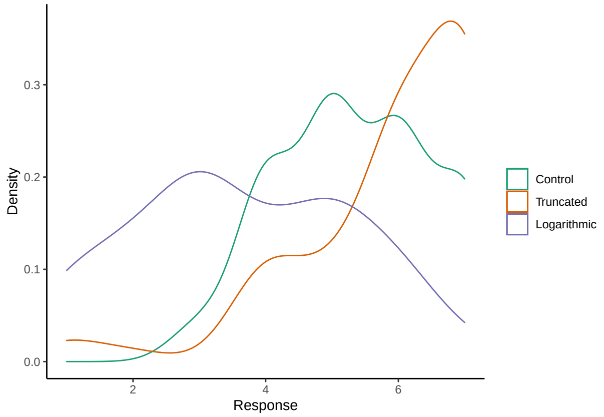
Figure 1: Density plot showing distributions of responses regarding all three plots, after removing values greater or equal to 1000 from the responses fore the logarithmic plot

Now considering the logarithmic-scale 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 very 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 very 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.

This should 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.

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

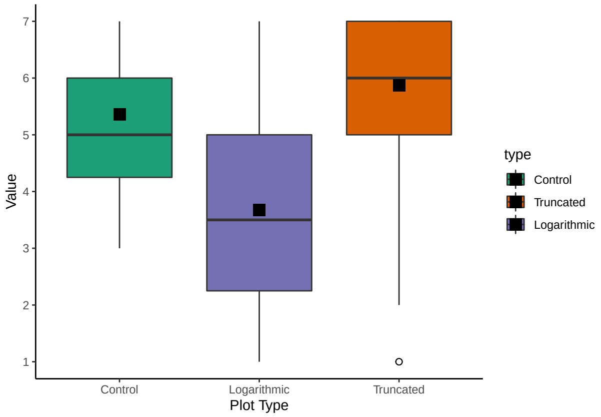
Now we consider the results from the second question, in which the participants were asked to respond on a scale from 1-7. The density plot below depicts the distribution of results for each of the three plot types. We see that the distributions appear to once again be non-normal.



The distribution for the logarithmic plot values has a fairly wide, flat curve, showing that the subjective view appeared to vary a fair amount from respondent to respondent. The distribution for the truncated plot seems very skewed to the right, depicting that the subjective view on the difference between the bar heights was that the difference was on the larger side.

An initial look at the table of summary statistics reveal means of 5.375, 3.671 and 5.871 respectively for the control, log and truncated plots, meaning that for the 'baseline' control plot, participants on average judged the difference to be moderately significant, with the perceived difference being smaller for the log plot and marginally larger for the truncated plot. This appears to be consistent with results from [[[CITE chrome-extension://cbnaodkpfinfiipjblikofhlhlcickei/src/pdfviewer/web/viewer.html?file=file:///C:/Users/Katie/Downloads/YangVargasRestrepoStanleyMarsh%20(2020).pdf]]], in which the 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. However, the average perceived difference here is fairly small, much smaller than initially hypothesised, so tests will be needed to decipher whether this is significant. The logarithmic plot causing the average perceived difference to be smaller follows the hypothesis from prior to running the survey.

The box plot shows these results for each plot type.

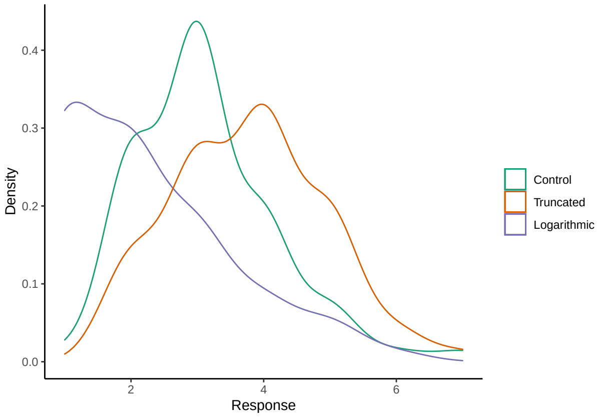


We see that the interquartile range for the control plot is smallest of the three at 1.75, followed by the truncated plot at 2, and then the log plot at 2.75. This depicts that overall, there was more of a consensus in the subjective perception of the difference for the control plot than the other two, and less agreement between participants for the logarithmic scale.

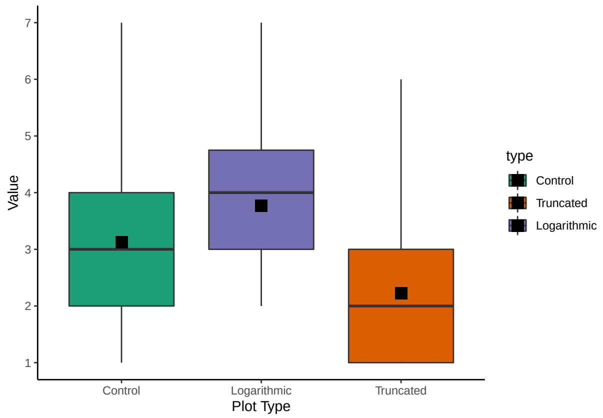
The black squares represent the means here, and we can see that for the control and truncated boxes, the mean is higher than the median, perhaps signifying a positive skew, with a slightly negative skew for the logarithmic plot.

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

This is a similar question to the one prior, but the purpose was to see if there was a difference in perceived difference for bars next to each other vs bars on opposite sides of the plot.



The distributions here appear more regular than for the previous question, with the truncated plot appearing almost normal, with the control plot skewed slightly more to the left, and then the logarithmic skewed once again more to the left. The means for these three, in the same order, are 3.771, 3.129 and 2.229, respectively, with medians of 4, 3 and 2. The ranges for these are all similar, with the truncated plot responses sitting in $[2, 7]$, the control responses in $[1, 7]$, and the logarithmic in $[1, 6]$, with variances of 1.309, 1.157 and 1.599. All of this points towards the respondents perceiving the differences in the truncated plot as lower than the control plot, and the logarithmic plot higher. This is again confirmed in the below box plots.



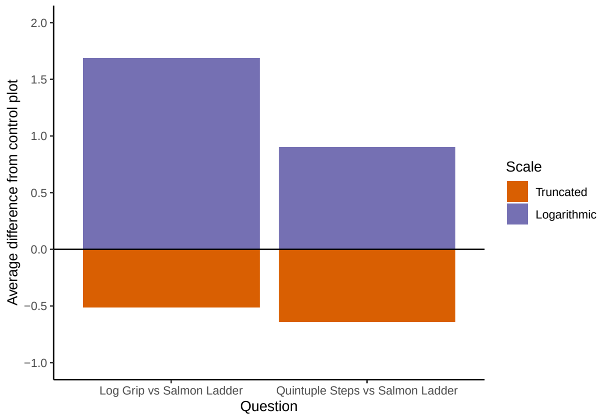


Figure 2: Bar plot giving a visual representation of the table

This is similar to the conclusions from the previous question.

Now, we will look at the means for both of the two questions regarding bar height difference perception, and compare the mean responses for the logarithmic and truncated plots with to the mean of the control plot responses. This will be in the form of taking the difference between the mean control response and mean of the truncated or logarithmic responses, respectively. This will give an idea of how much of a impact each scaling makes has on perception as compared to the control plot, depending on if the bars are next to each other or further apart. The table below gives these differences, alongside a bar plot giving a visual representation of these.

Table 1: Table showing difference in the percieved difference for the logarithmic-scaled and truncated plots as compared to the control, for the two above questions

	Con - Trnc	Con - Log
Log Grip	-0.5142857	1.685714
Quintuple Steps	-0.6428571	0.900000

On average we see 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. We can conclude 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.

Ninja Warrior - Part 2

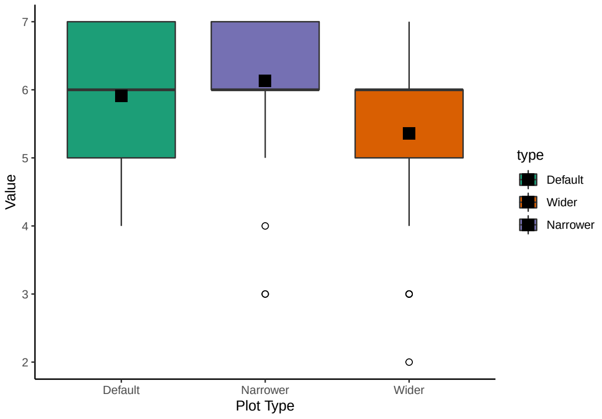
The second section of the survey tests whether altering aspect ratio of plots affects interpretation. The purpose of this is to mirror what my occur when visualisations are published, and may be resized to fit the section of the page they sit on. As in [[[CITE: http://perceptualedge.com/articles/visual_business_intelligence/bar_widths.pdf]]], it was hypothesised prior to the survey that an aspect ratio that 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 tests 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, we will not be considering different width of spaces between bars, but only the 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. In our version of this investigation, we have three bar plots of 7 obstacles, each with a different aspect ratio. The control, or default, plot is given as the plot for which the aspect ratio has not been altered, the plot with narrow bars has a doubled aspect ratio, and the plot wide bars a halved aspect ratio.

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

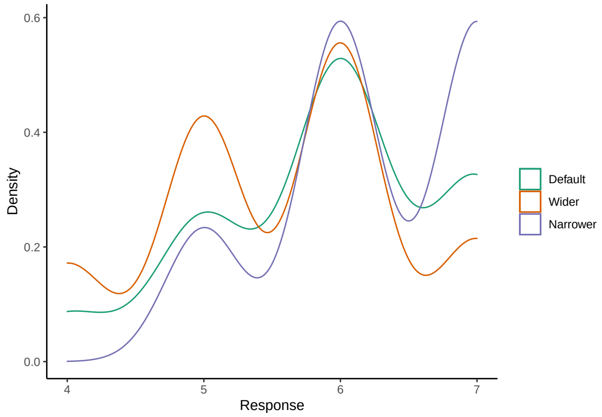
Looking at the means and medians here, it doesn't seem like there is that much of a difference in perception of the differences between the three aspect ratios. With means of 5.914, 5.357 and 6.129 for the default narrower and wider plots respectively, where 'narrower' is defined as the plot with the aspect ratio of smaller width to greater height, and vice versa for the 'wider' plot, and all have a median of 6. The means show marginal differences, whereby the default plot mean is in the middle-valued mean, with the mean perceived difference for the wider plot being slightly smaller than this and the mean perceived difference for the narrower plot is slightly larger. This result, although marginal, follows the hypothesis that the wider plot would cause differences to be perceived as smaller and narrower bars to cause differences to be perceived to be greater.

To discuss the ranges, see the box plot below.



From these box plots, it appears that the IQRs for the two plots with altered aspect ratios have very little, if any, overlap, despite the means being similar and medians being identical. The narrower plot shows a tendency for the responses to lie more towards the upper end of the scale than the wider plot, which also ranges over the upper half of the scale but between roughly 5 and 6 rather than 6 and 7. The default plot covers the entire IQR of both of the other plots, and the box plots then show that, even though the means and medians are very similar, the center bulk of the values for the narrower plot tended to be more towards the upper end of the default's IQR, whereas the central points of the wider plot sat on the lower half of this IQR. Additionally, we see that there are two outliers each for the narrower and wider plots, with values 2, 3 and 4. Perhaps excluding these from the data and re-analysing the summary statistics we will see a more marked difference.

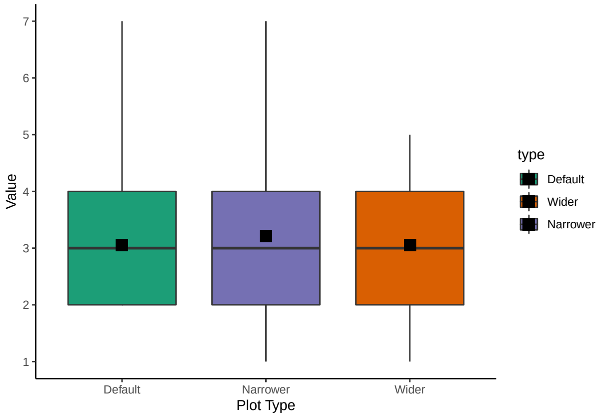
After performing the univariate analysis excluding these values, the means and medians didn't change that much, with the medians still at 6 for all the plots, and the mean response for the wider plot increasing from 5.357 to 5.543, and the mean response for the narrower plot from 6.129 to 6.257. Although as expected, the ranges and variances are lower, meaning the spread over values over the range is smaller, however this only furthers the point that altering the axis ratio appears to have minimal effect.This is again confirmed by looking at the distributions of the three plot types, which are very similar to one another. We can however see that the plot with the density for the wider plot is highest of the three for the values of 4 and 5, but is the lowest of the three for upper end of the distributions, and vice versa for the densities narrower plot. The default mostly stays in between the other two curves. For a third time, this gives way to the observation that the wider bars have a small lessening effect on gauging differences in height, and using narrower bars has a mild increasing effect on difference perception.



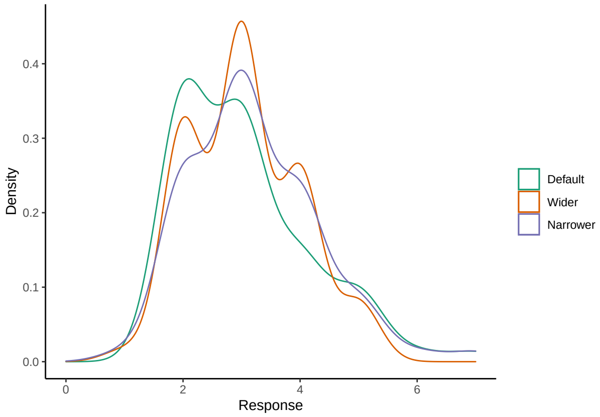
How large would you say the difference between 'Log Grip' and 'Floating Steps' is?

Similar to part 1, we have 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 wider plots being identical at 3.057, with the mean of the narrower 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, wider and narrower bars, respectively. The distribution of values are shown in the below box plots.



We see that 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 below density plots



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

Once again, we can analyse the distance from the perceived bar height difference for the default aspect ratio to those of the wider and narrower ratios, for each of the two height difference questions

Table 2: Table showing difference in the percieved difference for plots with narrower and wider bars as compared to the default, for the two above questions

	Def - Narrow	Def -
Jumping Spider vs Salmon Ladder	-0.2142857	0.55
Log Grip vs Floating Steps	-0.1571429	0.00

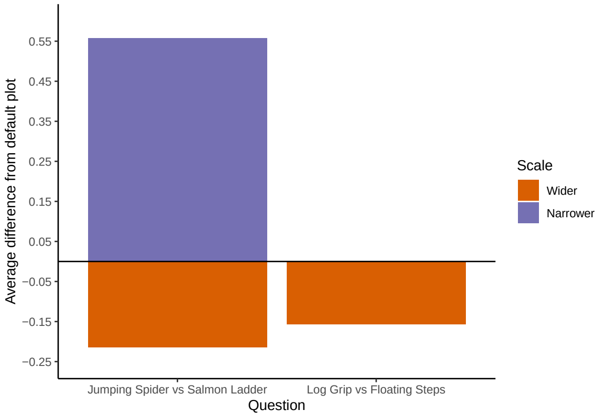
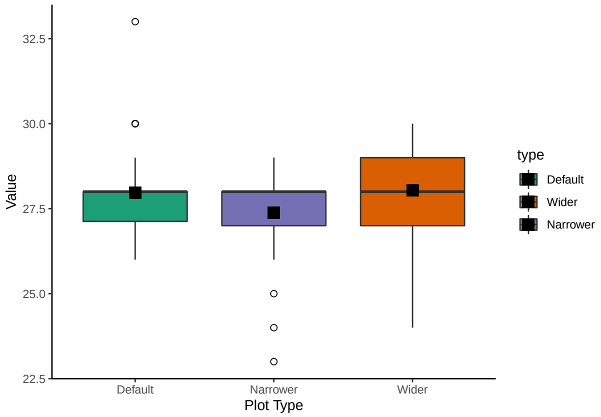


Figure 3: Bar plot giving a visual representation of the table

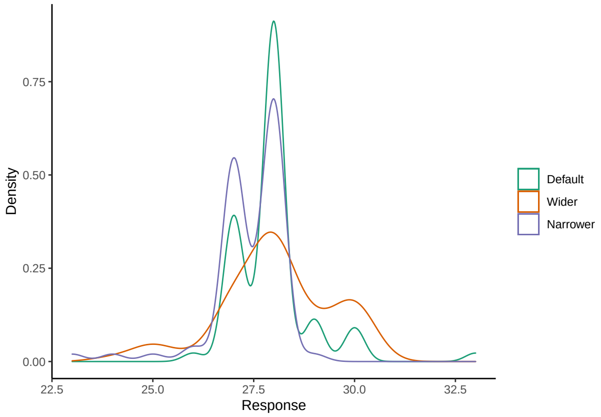
How many times would you say 'Floating Steps' were used?

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 we choose the third bar from the axis. 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.

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, wider and narrower, 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, we see 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 narrower 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 wider plot than for the other two, although this doesn't seem to be too significant.



The distributions for the default and narrower plot responses are very similar, both seeming to be fairly centred on the mean with a steep decrease in density on either side of the mean to very shallow tails within the range $[25, 30]$. The responses for the wider 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 narrower, 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.

Comparisons

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. Below a table is laid out giving the number of respondents that selected each plot for each of the three questions.

	Default	Narrower	W
Most aesthetically pleasing?	37	14	
Easiest to read and interpret?	36	15	
Hardest to read and interpret?	20	20	

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, we find that 27 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 wider bars, with 30 selecting this and 20 selecting each of the other two. While a significant number chose the default and narrower bars, the slightly higher amount selecting the plot with wider 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 we don't see this imbalance in the numbers from the previous questions, the result here does give some indication that the aspect ratio producing wider bars may impact on ease of interpretation.

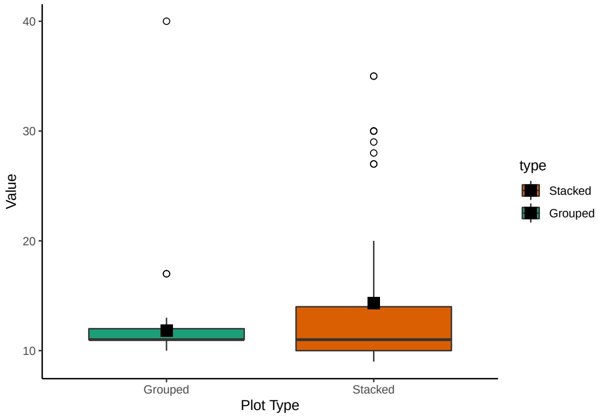
Ninja Warrior - Part 3

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 we highlight the number of times each obstacle was used in each stage of the competition for each bar. 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 we have comparisons of every colour scheme against every other colour scheme.

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

Again we start with the less subjective question regarding the reading of a numerical value off the axis. In this question we ask 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 we see 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 very 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. We do see 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. We can check for outliers by looking at the box plots for this data.



We do in fact see that the box for the grouped responses is very 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 we will remove the outliers 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.

We see that 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 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, we see that 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, we see 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, we will discount this set of 12 values 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, we once again run tests. Firstly, running Shapiro tests and symmetry tests to check for violations, we see that this data is not normal and is asymmetric. Thus, as in question 1, we run sign tests, alongside the t-tests on samples from a normal distribution with mean and variance equal to out data. We will test against a median of 11 for the sign tests, and a mean of 11 for the t-tests.

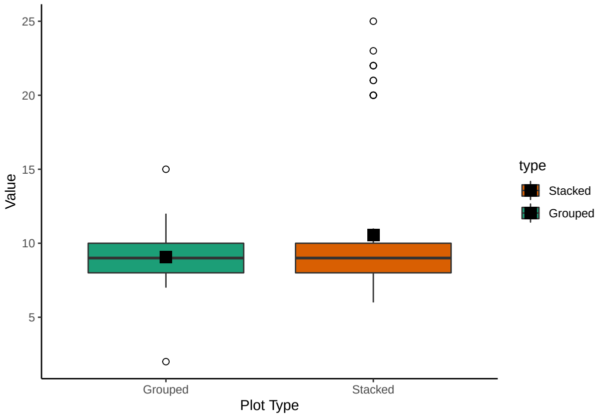
The sign test on the stacked bar plot responses gives a high p-value of 0.5258, showing that for the stacked bar plot responses (after removing the values as priorly specified), the participant estimated values do not differ significantly from the true value. For the grouped bar plot we have a p value of $0.009 < 0.05$, and thus these responses are statistically significantly different from the true value.

Running t-tests on the means, however, we see both sets of responses differ statistically significantly from the true value.

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

This question 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, we appear to see a selection of respondents judging the full height of the bar rather than the category as asked. Looking at the data, the interval for these responses seems to be $[20, 25]$, as the next response below 20 is a value of 10, seeming to separate the data into two separate subsets. This can be confirmed by a box plot.



We indeed see that the distribution of values for each of the two response sets appears to be almost identical with the exception of outliers at and above 20 for the box plot of responses for the stacked bar plot. Thus we view the sets of summary statistics for the two but with these values removed.

Here we see that 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. This can again be tested as above, where it is less clear whether the data are symmetric, so we will also run symmetry test.

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 very 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 t-tests show that the differences in the means from the true value are statistically significant, although not considering the tests we can see by eye that the means are relatively close to 9.

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

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.

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

Here was simply 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.

Var1	Freq
Grouped	59
Stacked	11

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

Which colour scheme do you find most aesthetically pleasing?

This question and the one following it 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 we created six pairings of three different colour palettes, 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.

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

	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, although this may differ between languages, which will be explored later on. Comparing this to the bottom two rows, in which we put viridis 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.

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

Complementing the aesthetic preferences, this 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.

```
 A  B None
42 20    8
```

Interestingly here, we see that the top two rows appear to give 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. 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.

Sales - Part 1

Now we move on to 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.

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. Furthermore, the zeroed plot has a small overall range, spanning $[1, 3]$ of the scale. The other two plots have range $[1, 7]$, but IQRs of $[2, 4]$ and $[2, 3]$.

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

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

Sales - Part 2

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?

Conclusions