Mads Corfixen, 2016 4715
mcorfi16@student.aau.dk

Data Visualisation Exam Paper
Visualisation of Pokémon!

Aalborg University
Data Science, 3rd Semester

Source: [1]

Table of Contents

1 Introduction

The interactive visualisations shown throughout the report can be found on https://madscorfixen.github.io/, where this report can also be downloaded. Furthermore, a link can also be found there that leads to the source code for the visualisations.

1.1 Description of Used Data

The static dataset utilised is of the table type, containing 800 items pertaining to different Pokémon species (Pikachu, Bulbasaur, etc.), and 23 attributes which describe the different characteristics of each species. The dataset originates from [2]. The 6 attributes, which are used in this report, are shown below, along with an explanation of them and their type. Finally, an example of how the items and attributes of the dataset are structured can be seen in Table 1.

Type1 & Type2
The attributes Type1 and Type2 describe the typing of a Pokémon. Each species of Pokémon is defined to have a typing consisting of either one or two *elements*. These elements are Normal, Fire, Fighting, Water, Flying, Grass, Poison, Electric, Ground, Psychic, Rock, Ice, Bug, Dragon, Ghost, Dark, Steel, and Fairy [3]. If a Pokémon only has a single element as its typing, e.g. Pikachu, which is Electric, the entry of Type2 for that Pokémon is empty. Pokémon with only a single element are called mono-typed, and Pokémon with two elements are called dual-typed. Both of these attributes are nominal and qualitative as they are labels with no ordering.

Base Total
Not all Pokémon are equally strong. Each species is defined by their *statistic*, which assigns a numerical value describing their Hit Points, Attack, Defense, Special Attack, Special Defense, and Speed respectively [4]. The sum of these individual statistics is called the *base total* of that species of Pokémon. This attribute is quantitative with values in $I_{BT} = [180;\ 780]$. It is of the type interval as it is possible to order and calculate differences between its values, but no zero value exists. The ordering direction is sequential.

Capture Rate
Not every Pokémon is equally easy to capture. This attribute describes the rate with which each Pokémon is caught, with lower values indicating more difficult catches. With values in the interval $I_{CR} = [3;\ 255]$, the type of this attribute is also interval and quantitative, and its ordering is sequential.

Is Legendary
Certain species of Pokémon are extraordinarily stronger than all others. Such Pokémon are called legendary, and are usually used as the mascot for each Pokémon game. The type of this attribute is dichotomous qualitative as it describes a boolean value; either a Pokémon is legendary, and is assigned the value of 1, or it is not, and is assigned the value of 0.

Base Egg Steps
The last aspect is Pokémon breeding. In the Pokémon games, every non-legendary Pokémon is capable of laying an egg, which is incubated by taking steps, i.e. walking, in the game. This attribute describes the exact number of steps needed for each Pokémon species' egg to hatch. With values in the interval $I_{BES} = [1,280;\ 30,720]$, the type of this attribute is also interval and quantitative, and its ordering is sequential.

Table 1: An example of how the items and attributes of the dataset are structured.

Species	Type1	Type2	Base Total	Capture Rate	Is Legendary	Base Egg Steps
Pikachu	Electric	None	320	190	0	2,560
Jigglypuff	Normal	Fairy	270	170	0	2,560
Mew	Psychic	None	600	45	1	30,720

1.2 Description of Tasks

In this section, some tasks that can be solved using the aforementioned data will be discussed. Furthermore, the relevance of these tasks will be elaborated upon.

The Pokémon genre is widely popular all around the world, with the mobile app, Pokémon GO having been downloaded a total of 1 billion times since its release in July, 2016 until March, 2019 [5]. Furthermore, the latest release of the games in the main series, Pokémon Sword and Shield, sold 19.02 million copies in the world in the period from its release on November 15, 2019 to November 5, 2020 [6].

Because of this huge popularity, it makes sense to attempt to compare different Pokémons' rarity to each other, such that the players can compare between themselves who has the most rare Pokémon. To accomplish this, a task would be to discover whether any trends exist for the typings of Pokémon. Specifically, are some typings more common than others, and are mono-typed Pokémon more prevalent than dual-typed? This could help players identify how rare a certain Pokémon's typing actually is as well as compare it to other, equally uncommon typings. This task can be abstracted to *exploring distributions* and *identifying outliers*.

Another consideration in a Pokémon's rarity is the difficulty of obtaining, or catching, it in the games. Different factors could come into play here, and an interesting task would be to explore the correlation between a Pokémon's Capture Rate and its other statistics, such as Base Total and Legendary status. This task can be abstracted to *identifying correlations*.

Finally, some Pokémon are deemed as being Legendary by the games' developer, Nintendo. To determine if there are any similarities between the attributes of Legendary Pokémon, it is relevant to compare the attributes of Legendary Pokémon to those of non-Legendary Pokémon. This task can be abstracted to *discovering (dis)similarities*.

On the basis of these tasks, and the attributes described in Section 1.1, a problem statement can be formulated, summarising the desired goals that the visualisations should be able to provide an answer for.

1.3 Problem Statement

How can interactive visualisation idioms be constructed to efficiently solve the following tasks?

1. Explore distributions and identify outliers.
    - Determine most common mono- and dual-typing of Pokémon
2. Identify correlations.
    - Identify whether a correlation exists between the Base Total and other attributes of Pokémon.
3. Discover (dis)similarities
    - Discover if being Legendary has an effect on a Pokémon's attributes

2 Analysis of Idioms

In this section, the appearance of the program itself, consisting of a part of the GitHub Pages site referenced in Section 1 as well as the different visualisation idioms used herein, will be analysed using the theoretical knowledge obtained in the Data Visualisation course at Aalborg University [7]. For some of the idioms, an alternative was also considered and analysed. These alternatives are found immediately after the idiom which they would replace.

2.1 The Program

The visualisation program itself is a part of the GitHub Pages site as shown in Figure 1, which will from now on be referenced as the *Main Window*.

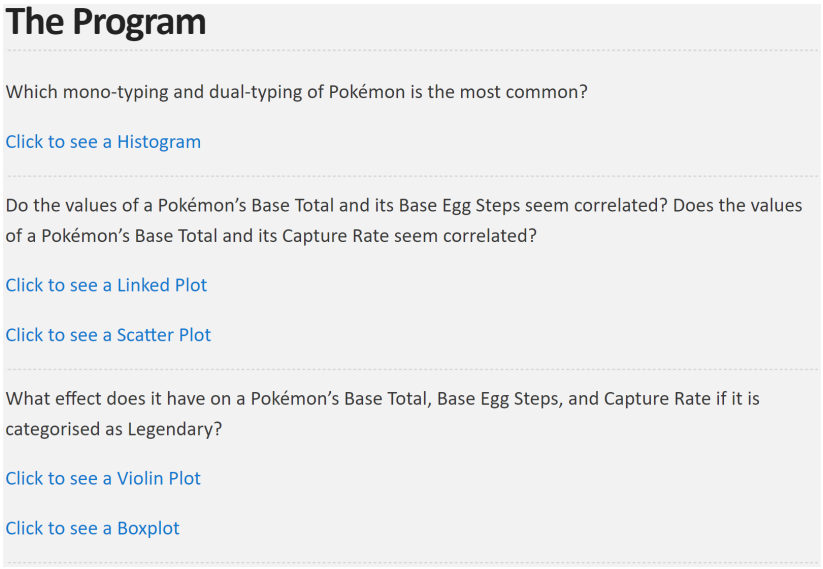

Figure 1: The part of the GitHub Pages that contains the actual program. Referenced in the body text as the Main Window.

The first thing one might notice in the Main Window is the lines of blue text. These typically signify that the text affords clicking, which is also the case here. Furthermore, these lines of blue text contain feed-forwarding of what will happen if the user clicks them in the form of the actual text, i.e. 'Click to see [...]'.

Other than the blue text, the Main Window is partitioned into three blocks using horizontal dashed lines as constraints. Each of these blocks contains further feed-forwarding in the form of one or more questions. The combination of these constraints and feed-forwarding is meant to communicate which questions each visualisation can be used to answer to avoid slips from the user, i.e. to help the user align their goal and actions. In other words, to ensure that they click the correct piece of blue text that maps to the visualisation that can answer their question.

2.2 The Visualisations

2.2.1 The Plotly Toolbar

A built-in functionality in all visualisations created using **plotly** in Python is the Plotly Toolbar, shown in Figure 2, hereafter simply called the *Toolbar*.

Figure 2: The Plotly toolbar.

The Toolbar utilises signifiers in the form of small pictures and text upon hovering to feed-forward what each tool does. Furthermore, it uses highlighting of each button to indicate what state the Toolbar is currently in. From left to right, the functionality of each button is:

- The Camera, an iconic signifier, affords the saving of the plot on your computer. It feed-forwards that pressing it will *capture* the visualisation, like a camera captures a picture.
- The Magnifying Glass, an iconic signifier, affords the zooming in on the visualisation. It imitates the function of a real-life magnifying glass.
- The Four Arrows, an indexical signifier, affords panning around the visualisation. It directly points in four two-dimensional directions, feed-forwarding exactly what it allows someone to do; to pan up, down, to the right, or the left.

- The Dotted Box, a symbolic signifier, affords the selection of an area on the visualisation. As a convention, the Dotted Box usually is a selection tool.
- The Lasso, a symbolic signifier, affords the lasso selection of an area on the visualisation. As a convention, the Lasso usually signifies this action.
- The Plus and Minus, a symbolic signifier, respectively affords the zooming in and out on the visualisation. It feed-forwards that clicking will either increase (+) or decrease (-) the zoom-level.
- The Box and Arrows, a symbolic signifier, affords re-scaling the visualisation to the standard zoom-level. Seeing the box as the visualisations, the arrows are stretching it outwards, towards it original form.
- The House, a symbolic signifier, resets the axes of the visualisation, returning to the original view. It feed-forwards that clicking it will bring you 'home' to the original view.
- The Dotted Corner, an indexical signifier, toggles the spike lines of the visualisation on or off. The spike lines look exactly like the depicted signifier.
- The One and Two Left-Arrows, an indexical signifier, controls how to view the date upon hovering over it. Selecting the One Left-Arrow, shows all information at the cursor, while the Two Left-Arrow shows the primary axis value at the cursor and the secondary axis value on the secondary axis.
- The Plotly Icon, a symbolic signifier, directs the user to the Plotly website. Functions like a watermark.

In the following analyses, the Toolbar will not be shown, however take note that it is actually present in all the visualisation idioms.

2.2.2 The Histogram

Clicking the text 'Click to see a Histogram' opens a new tab in web browser containing the idiom depicted in Figure 3, hereafter called the *Histogram*. The target of the Histogram is to answer the first task of identifying the most common Pokémon typings. In abstract terms, it affords the exploration of the distribution and the identification of outliers.

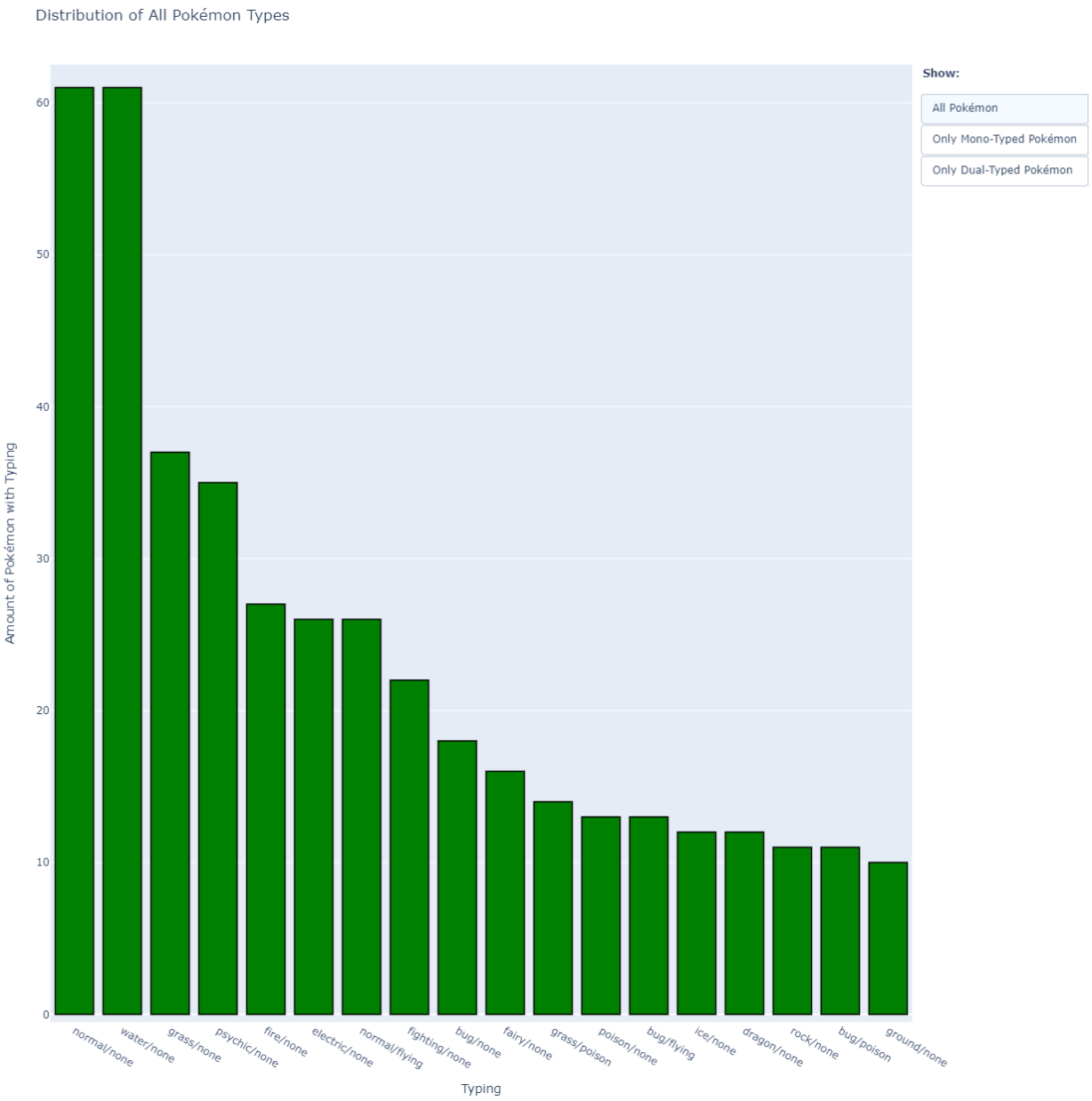

Figure 3: The histogram idiom used in the visualisation program. Referenced as the Histogram in the body text. It utilises height on a common scale to identify common Pokémon typings.

Before the Histogram was created, a new, single attribute to identify a Pokémon's typing was derived, called *Typing*. This was done in order to easily include all typings, both mono and dual, on the primary axis of the Histogram. On the secondary axis, the quantitative sequential derived value *Amount of Pokémon with Typing* is measured. The primary axis is zoomed in as standard in order to make visible the different categories - as the task is locating either the most common or the least common typings, this does not hide relevant information from the user.

The marks of this idiom are the vertical bars, with the height of these bars on a common scale being used as a channel to encode the amount of items belonging to each Typing category. As there is only one attribute of import in this idiom, the Amount, the channel with the greatest magnitude available was used. In addition, the bars are shown on a common scale to maximise ease of judgement between their heights. Furthermore, the bars are encoded to show the highest bars first and the lowest last to increase the popout, thus decreasing the time it takes to locate the most common Typing.

In order to more easily answer each sub-question of the first task, i.e. identifying the most common mono-typing and the most common dual-typing, the Histogram contains three buttons in the upper right corner to

change which items are visible on the plot. These buttons are contained in boxes and highlight when the cursor hovers over them, both of which are signifiers that they afford clicking. Furthermore, the buttons contain feed-forward mechanics in the sense that they are read as 'Show All Pokémon', 'Show Only Mono-Typed Pokémon', and 'Show Only Dual-Typed Pokémon'. Finally, the buttons also function to feed-back the status of the visualisation as the currently chosen option is highlighted with a light-blue colour as seen for the 'All Pokémon' button in Figure 3.

2.2.3 The Linked Scatter Plot

Clicking the text 'Click to see a Linked Plot' opens the idiom depicted in Figure 4, hereafter called the *Linked Plot*. The target of the Linked Plot is to answer the second task of identifying if any correlations exist between the Base Total and other attributes of Pokémon.

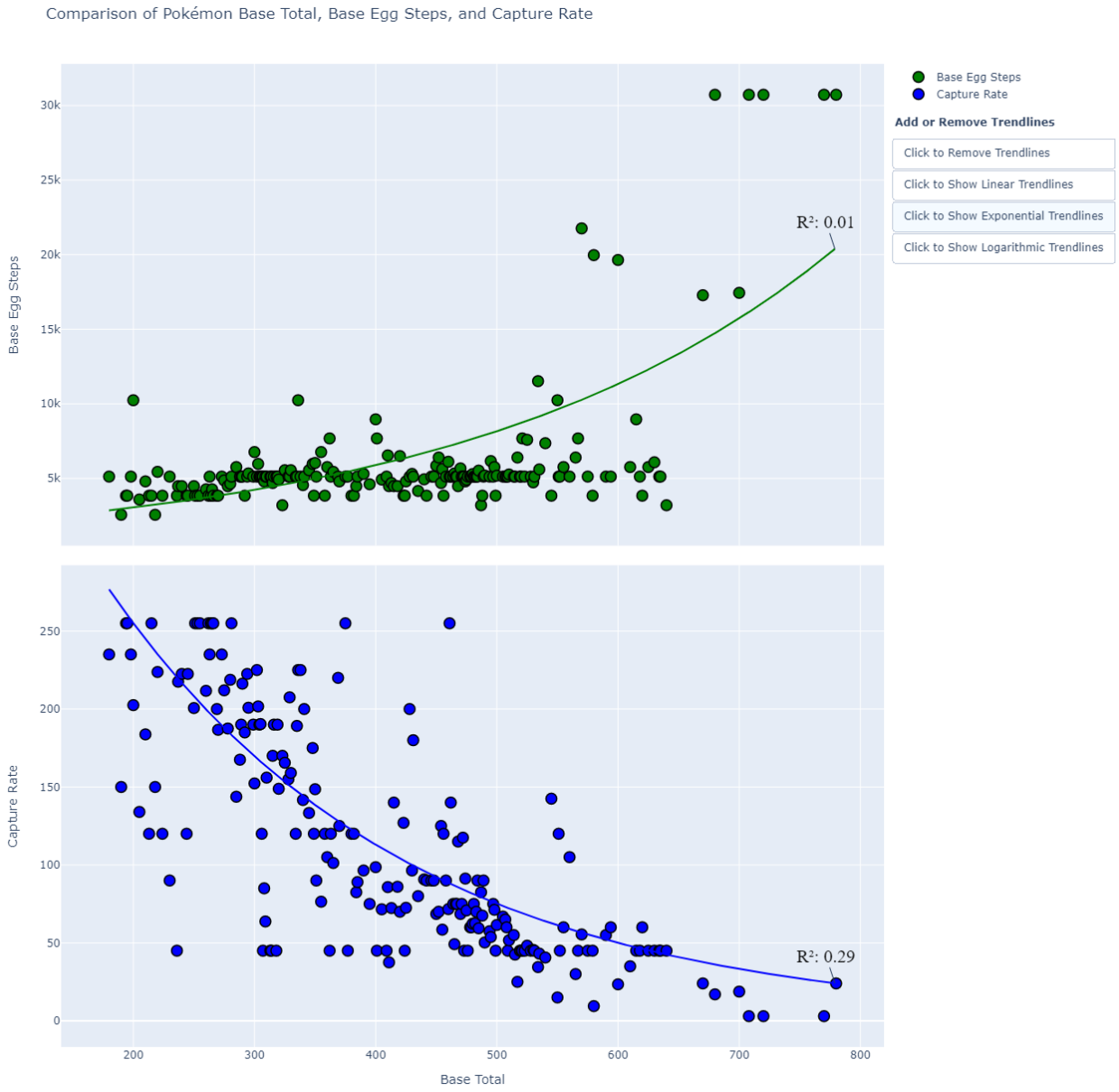

Figure 4: The linked scatter plot idiom used in the visualisation program. Referenced in the body text as the Linked Plot. It utilises position on a common scale and colour hue to identify correlations in the attributes.

The Linked Plot consists of two scatter plots on a common primary axis depicting the Base Total. On the top-most secondary axis, the Base Egg Steps are depicted, and on the bottom-most, the Capture Rate is

depicted. The marks utilised in the Linked Plot are circles, and the channel used to communicate the value on the axes is height on a common scale. In order to differentiate between the two subplots, colour hue is used as a channel for the two categories of data; one group of data points being green and the other blue. Furthermore, an annotation is added to the top-right corner that contains a legend for the subplots, further increasing the separability of the data groups. Height was used for the values of the axes as these were the most important take-away from this visualisation, and height is of a higher effectiveness than colour hue. The two scatter plots are linked in the sense that zooming on the primary axis changes the view on both plots, and clicking the buttons available produces annotations to both plots.

In order to make the task of identifying the type of correlation seen in the visualisation, four buttons were added to the right allowing the user to add either a linear, exponential, or a logarithmic trendline to both subplots, or to remove the lines again. These buttons function in a manner like that described for the Histogram, except that they produce trendlines when clicked, along with the coefficient of correlation (R^2) value of that line. For each line, colour hue and location was used to signify which group they belonged to. The effect of clicking these buttons is also feed-forwarded to the user, e.g. 'Click to Show Exponential Trendlines'.

2.2.4 Alternative - The Coloured Scatter Plot

The program also allows the use of an alternative idiom to identify correlations. Clicking the text 'Click to see a Scatter Plot' displays the idiom depicted in Figure 5, hereafter called the *Coloured Plot*.

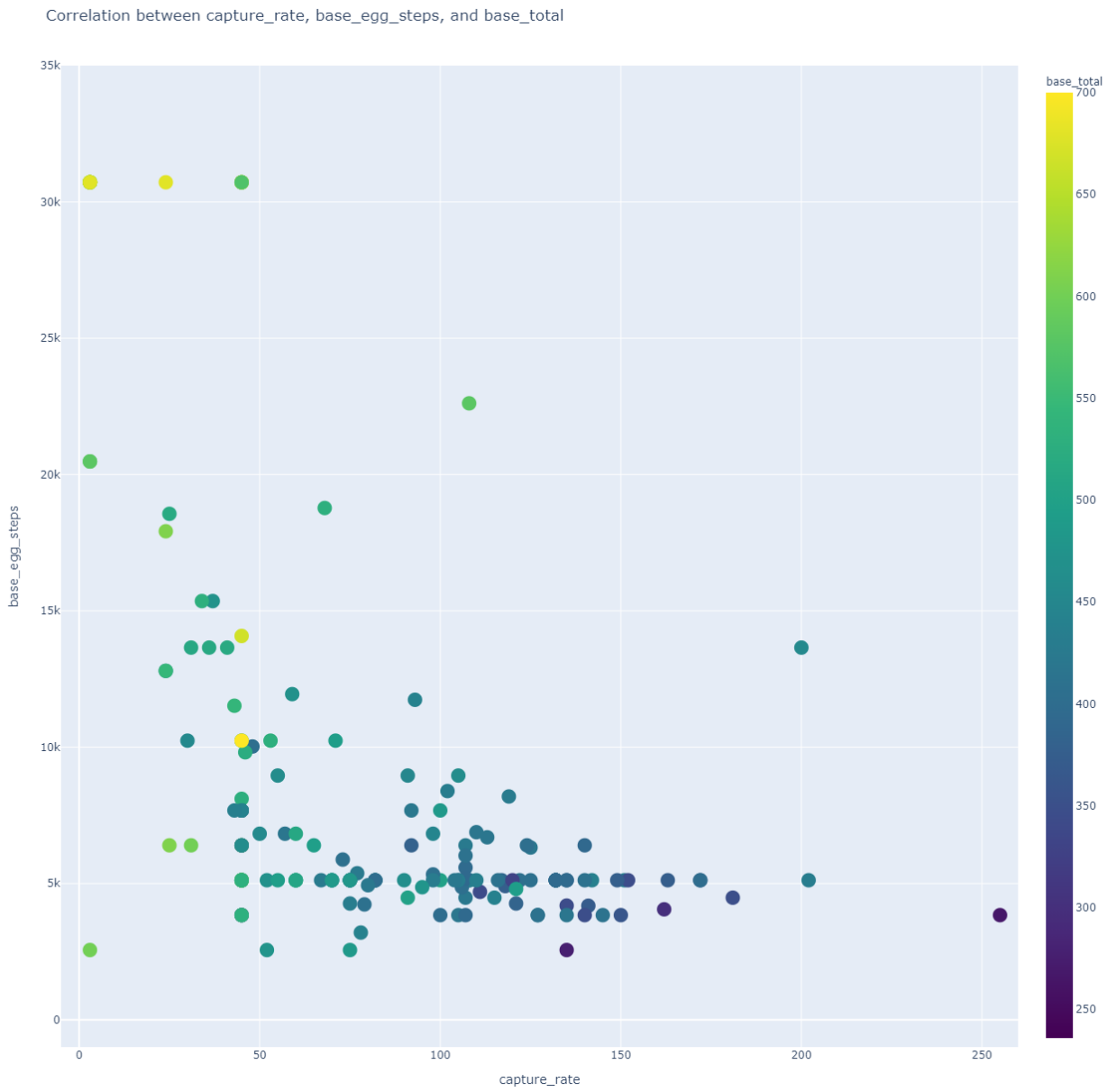

Figure 5: An alternative to the Linked Plot. Referenced in the body text as the Coloured Plot.

The Coloured Plot is a more concise way of visualising the same items as the Linked Plot. Instead of creating a subplot, it measures Capture Rate on its primary axis and Base Egg Steps on its secondary axis using circles as marks, and utilises the continuous colour scale, *Viridis*, to encode the value of Base Total, in order to avoid using three dimensions to show all the attributes in the same plot. The fact that the Viridis colour scale consists of multiple hues with equally increasing luminance is important to ensure that it is perceived as ordered and linear.

The advantages that this idiom has over the Linked Plot is that it is more concise. It visualises the same, but does not require two visualisations to do it, and neither does it require three dimensions to map three attributes to each other. Furthermore, if the information in the idiom is perceived correctly, it is easier to see the correlations. For example, it is seen that a high value on the secondary axis corresponds to a low value on the primary axis, and the colours seem to be closer to yellow than purple, and vice versa.

However, this idiom is more involved than the Linked Plot as three attributes are mapped to a 2-dimensional plane. This makes it less easy to interpret than a traditional scatter plot, leading to a higher degree of guidance required for the user to accurately read the visualisation. Another problem for the Coloured Plot is the fact that colour has a much lower effectiveness than position of a common scale, which could lead the perceiver to

believe that the attributes measured on the axes are more important than the attribute measured on the colour scale.

2.2.5 The Violin Plot

Clicking the text 'Click to see a Violin Plot' displays the idiom depicted in Figure 6, hereafter called the *Violin*. The target of the Violin is to answer the third task of discovering the effects of a Pokémon being Legendary.

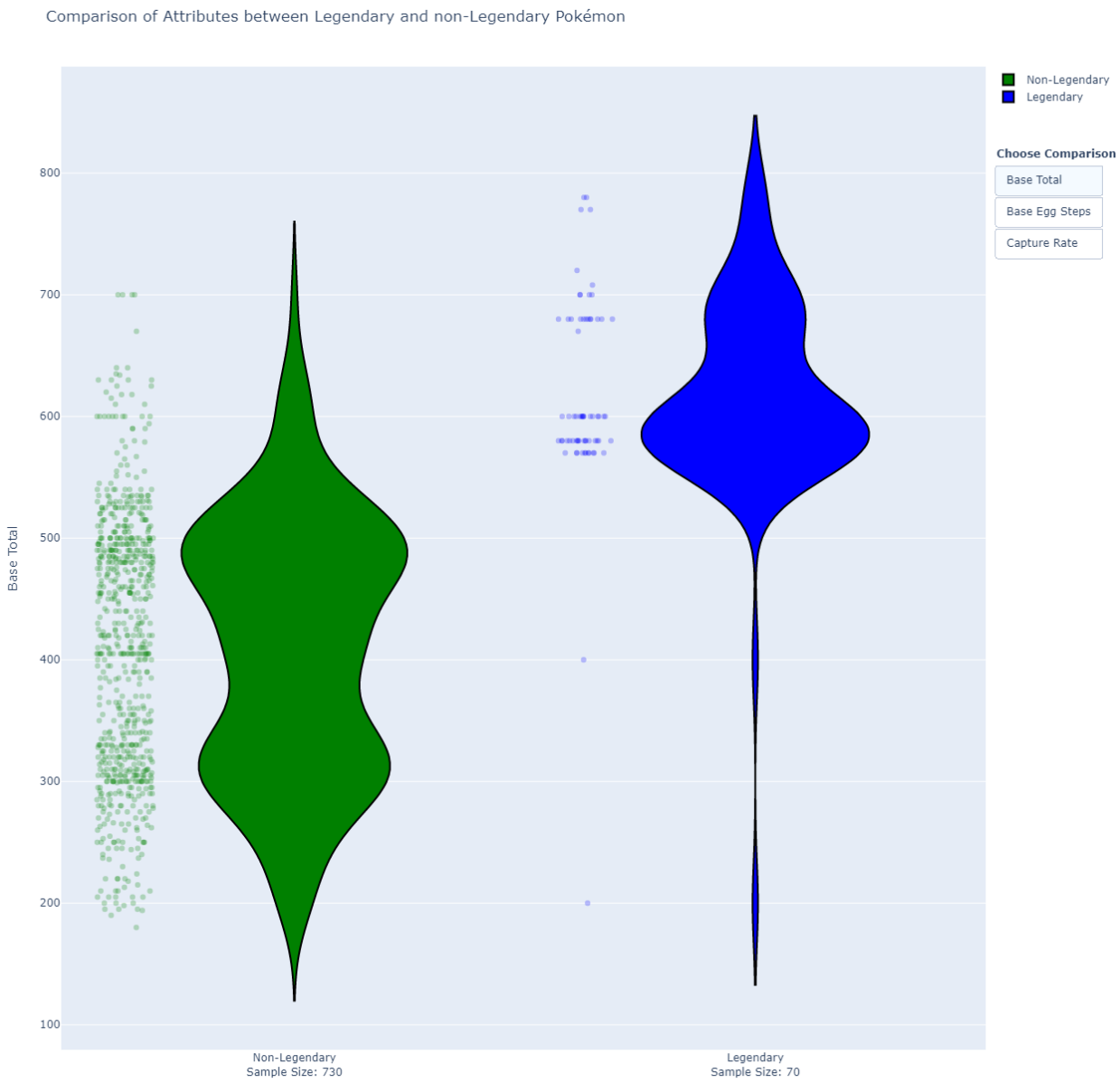

Figure 6: The violin plot idiom used in the visualisation program. Referenced in the body text as the Violin.

The Violin consists of two categories on the primary axis, and measures Base Total on the secondary axis. The marks used in this plot are zero-dimensional points in the form of filled circles, and two-dimensional areas. The channels are height on a common scale to show the value of underlying items for the points, and area to show the distribution of the items. Furthermore, colour hue is used to differentiate between the two categories of the primary axis. Additionally, the points have been jittered, e.g. using horizontal position as a channel to allow all of the points to be shown. Finally, to indicate whether a lot of points overlap, they have been made transparent.

This idiom also includes a legend as well as three buttons that function as the buttons of the Histogram, allowing the user to change the data shown on the Violin. Furthermore, the annotations on the primary axis, which shows the title of each category, also include how large a sample size makes up the Violin, e.g. 'Sample Size: 70' for the 'Legendary' category.

2.2.6 Alternative - The Boxplot

The program furthermore allows the use of an alternative idiom to the Violin. Clicking the text 'Click to see a Boxplot' shows the idiom depicted in Figure 7, hereafter called the *Boxplot*.

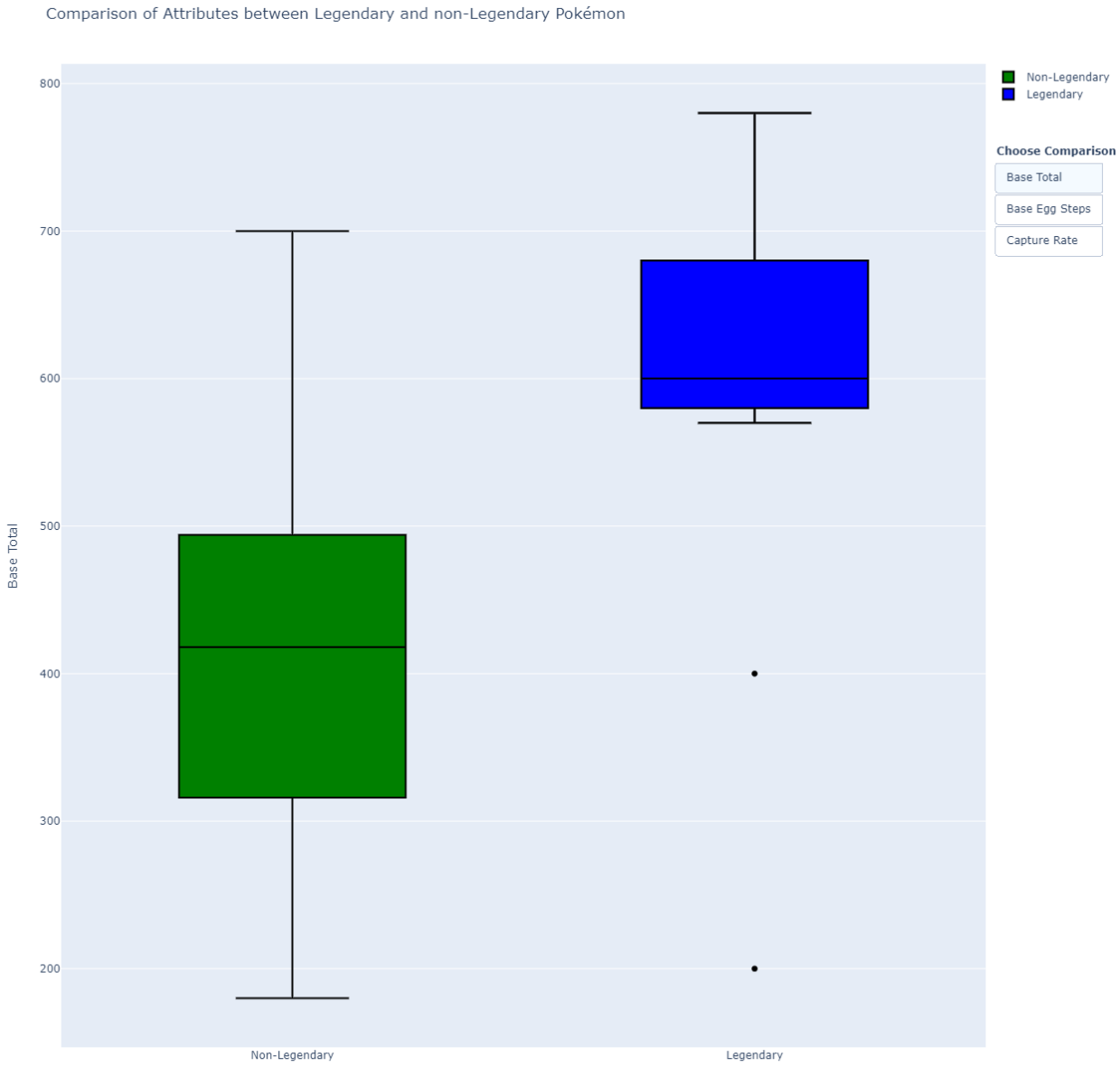

Figure 7: An alternative to the Violin. Referenced in the body text as the Boxplot.

The Boxplot consists of the same axes as the Violin, but instead uses the channel of height on a common scale to encode the distribution of the items in each category. Furthermore, it uses circular points as marks, mapped to the channel of height on a common scale, to show outliers.

The biggest advantage of the Boxplot over the Violin is that it is more widely known, and therefore a user would need less instruction on how to view this visualisation. However, the Boxplot does not capture the true sample distribution in the same manner as the Violin does. Furthermore, both idioms encode upper and lower

quartile, though it is easier to see on the Boxplot as they are marked with horizontal lines, whereas they are encoded on the Violin as the two widest parts of the area.

3 Ethical Concerns

This section will discuss several ethical concerns regarding the program itself and the visualisations contained herein. It is split into two parts; one section will discuss the data ethics of the visualisations themselves, and another will discuss the possible societal impacts of the program.

3.1 The Visualisations

This section will explore the data ethics of two of the utilised visualisations; the Linked Plot, Figure 4, and the Violin, Figure 6 as these two visualisations in particular contain elements which could be problematic.

3.1.1 The Linked Plot

The largest concern for the Linked Plot is the bias of patternicity. Viewing the standard state of the visualisation, i.e. without any added trendlines, might lead the viewer to assume that Base Egg Steps and Base Total are highly exponentially correlated. Even after adding the different trendlines, the pattern still seems obvious. For these reasons, the coefficient of correlation, which is a good measure for the actual correlation between the two attributes, was annotated on the visualisation to ensure that no untruthful correlations were derived from it. Before the R^2 value of 0.01 was added, an exponential trend seemed obvious, but seeing as only such a small part of the variance can be explained as an exponential correlation, the truth is different; there is in fact no evidence of an exponential trend.

3.1.2 The Violin

Several best practises exist for the Violin idiom in order to not propegate false information. Firstly, since the precision of the Violin increases with the sample size, the sample size should be indicated on the primary axis. Furthermore, it is important to not include impossible values on the Violin. This was a problem at first as the violin plot functionality in Plotly automatically showed negative values for both Base Egg Steps and Capture Rate, but this was solved by setting the standard view of the visualisation to begin at 0 for the secondary axis as shown in Figure 8.

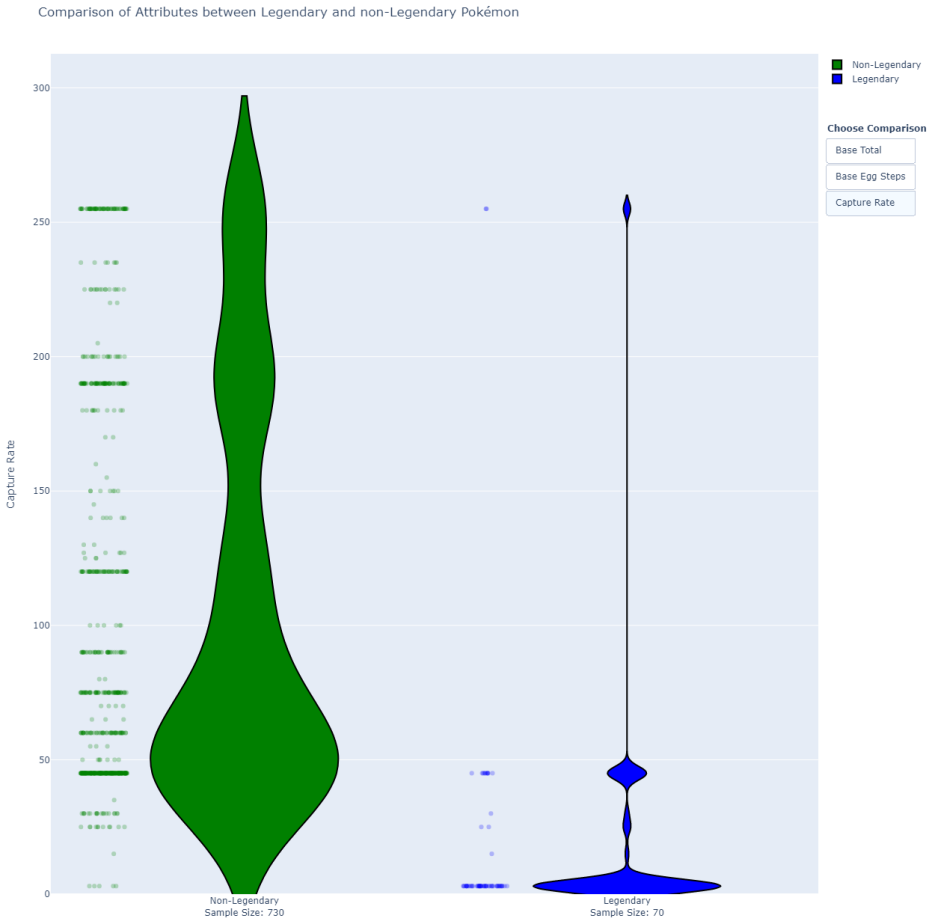

Figure 8: The standard zoom of the Violin to avoid showing impossible values.

3.2 Societal Impact

The purpose of the program is, broadly speaking, to determine what constitutes a rare and strong Pokémon. Taking this thought to the extreme, the program might show that a small subset of all Pokémon are significantly stronger than all others, leading to these few Pokémon being much more sought-after by players than others. This might sour the gaming experience for some players as the purpose of the games might then shift from enjoying oneself, the purpose of any leisure activity, to being the player with the most rare and strongest Pokémon team.

Another consideration is in regards to the main games, e.g. Pokémon Sword and Shield. In these games, the player must defeat enemy Pokémon using a team of a maximum of six of their own Pokémon. If the program developed in this report shows that six Pokémon stand above all others in terms of strength, it might lead to a homogenisation of players' Pokémon teams, meaning that all the effort that Nintendo has expended designing all the other, less powerful Pokémon, is essentially wasted.

4 Usability Test

In order to validate the effectiveness of the program in answering the tasks, a usability test was conducted. In this section, the findings of this test will be presented, and the changes that could be derived from the respondent's feed-back will be summarised. This test will focus mainly on exploring the third and forth threat to validity as shown in the black box in Figure 9 [8, ch. 4].

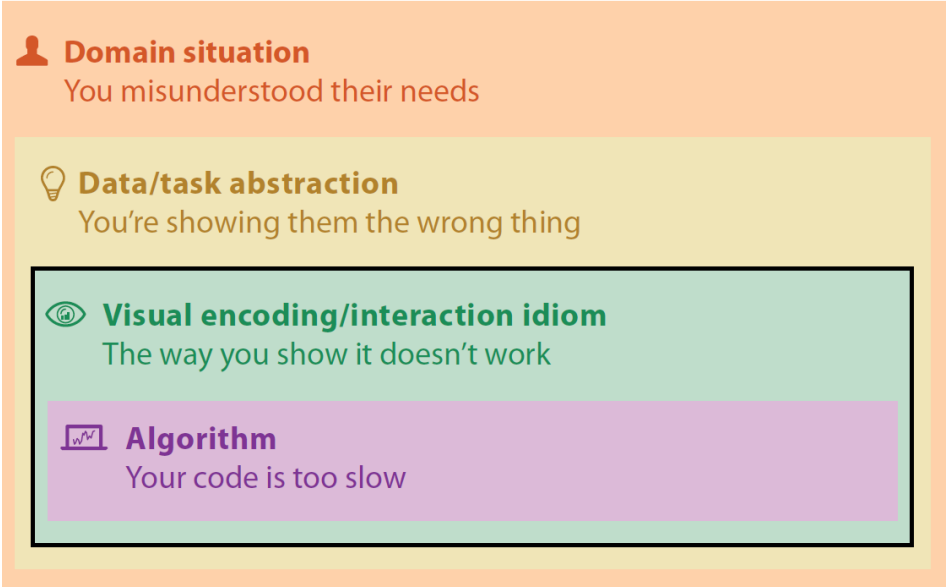

Figure 9: The four different threats to validity as explained in [8].

4.1 Tasks

The respondent was asked to solve the following tasks using the visualisations. The visualisations that the user was expected to use to answer the question are presented in parenthesis.

1. Which mono-typing of Pokémon is most common? (Figure 3)
2. Which dual-typing of Pokémon is most common? (Figure 3)
3. Does the values of a Pokémon's Base Total and its Base Egg Steps seem correlated? If so, is it mostly linear, exponential, or logarithmic? (Figure 4 or Figure 5)
4. Does the values of a Pokémon's Base Total and its Capture Rate seem correlated? If so, is it mostly linear, exponential, or logarithmic? (Figure 4 or Figure 5)
5. What effect does it have on a Pokémon's Base Total, Base Egg Steps, and Capture Rate if it is categorised as Legendary? (Figure 6 or Figure 7)

4.2 Results

In Table 2, the feedback of the usability test is seen. It is divided into different categories depending on which task the feedback was directed at. If the feedback was of a more general nature, it will be presented in the *General* category.

Table 2: An overview of the results of the usability test.

Task	**Feedback**
Tasks 1 & 2	1. Found buttons easily and understood their purpose. 2. Slight confusion due to /none. 3. Correctly solved the tasks.
Tasks 3 & 4	4. Found buttons easily and understood their purpose. 5. R^2 value does not stick out. 6. Preferred the linked plot to the coloured plot. 7. The coloured plot required a lot of hand holding to decipher. 8. Correctly solved the tasks.
Task 5	9. Liked the violin plot and found it intuitive. 10. Much preferred the violin plot to the boxplot due to the boxplot sometimes only being a horizontal line. • Note: This is due to the lower and upper quartile being equal, and not a mistake in the plot. 11. Correctly solved the task.
General	12. Momentarily unsure of how to get back to the main window. Assumed there would be a 'Back' button.

4.3 Changes Derived

The usability test supports that the alternative visualisation idioms presented above should remain as alternatives as the main visualisations were preferred by the respondent. Furthermore, the following points are seen as the main take-aways from the usability test, and need to be changed in order to improve the visualisation idioms.

- Increase pop-out of R^2 value in the Linked Plot

This has been implemented by using the channel of colour hue to colour the background of the R^2 value, and the marks were made transparent, to further increase focus on the trendline as seen in Figure 10.

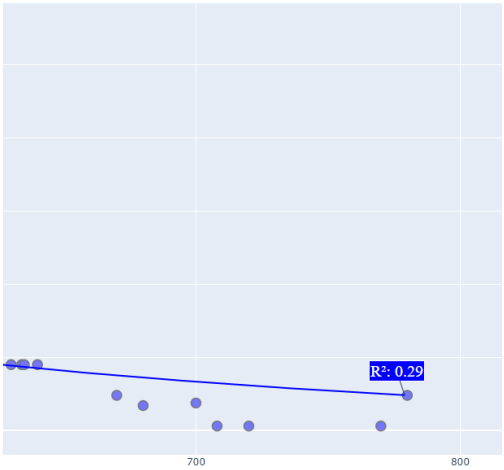

Figure 10: In order to accommodate one of the problematic points uncovered in the usability test, the R^2 value of the Linked Plot was given a background colour and the marks were made transparent.

- Remove the '/none' part of the Histogram

To accommodate solving this issue, the '/none' part was removed from the primary axis of the Histogram if the user chooses to only show mono-typed Pokémon. This change can be seen in Figure 11.

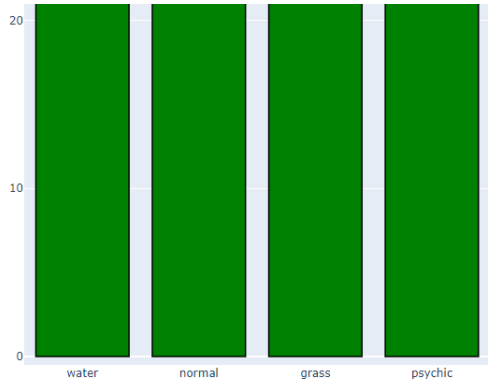

Figure 11: In order to reduce confusion, the '/none' part was removed from the Histogram if the user clicks 'Show Only Mono-Typed Pokémon'.

- Implement a button for each visualisation to return to the Main Window

Due to time constraints and the difficulty of implementing such a button as no easy collaboration exists between **plotly** and GitHub Pages, this change will not be implemented.

5 Bibliography

References

[1] 1000logos.net. (2020). "Pokemon logo," [Online]. Available: https://1000logos.net/pokemon-logo/ (visited on 01/23/2021).

[2] R. Banik. (2017). "The complete pokemon dataset," [Online]. Available: https://www.kaggle.com/rounakbanik/pokemon (visited on 11/11/2020).

[3] Bulbapedia. (2020). "Type," [Online]. Available: https://bulbapedia.bulbagarden.net/wiki/Type (visited on 12/21/2020).

[4] ——, (2020). "Statistic," [Online]. Available: https://bulbapedia.bulbagarden.net/wiki/Statistic (visited on 12/21/2020).

[5] M. Iqbal. (2020). "Pokémon go revenue and usage statistics (2020)," [Online]. Available: https://www.businessofapps.com/data/pokemon-go-statistics/ (visited on 01/23/2021).

[6] E. Swan. (2020). "Pokémon sword and shield sold over 19 million copies in one year," [Online]. Available: https://dotesports.com/pokemon/news/pokemon-sword-and-shield-sold-over-19-million-copies-in-one-year (visited on 01/23/2021).

[7] H. Knoche and C. B. Madsen. (2020). "Data visualization," [Online]. Available: https://www.moodle.aau.dk/course/view.php?id=34753 (visited on 01/25/2021).

[8] T. Munzner, *Visualization Analysis & Design*, ser. A K Peters Visualization Series. Taylor & Francis Group, 2014, ISBN: 978-1-4665-0891-0.