Simulation and Data Visualisation Assignment 2

Carlota Vazquez Gonzalez
K1890616

ACADEMIC HONESTY INTEGRITY

Students at King's are part of an academic community that values trust, fairness and respect and actively encourages students to act with honesty and integrity. It is a College policy that students take responsibility for their work and comply with the university's standards and requirements. Online proctoring / invigilation will not be used for our online assessments. By submitting their answers students will be confirming that the work submitted is completely their own. Misconduct regulations remain in place during this period and students can familiarise themselves with the procedures on the College website. ***I agree to abide by the expectations as to my conduct, as described in the academic honesty and integrity statement.***

1 PART 1: ANALYTICS

1.1 Exploratory Research Questions Proposed

Q1 - Analyse the development of virus mutations over time. Are there detectable seasonal trends? For this question, the focus will be on the dataset provided by the United Kingdom. From December 2019 until October 2020. Sub-questions for this section could be, which variants have mutated the most frequently and is there a particular time of the year this peak has occurred.
Q2 - Analyse and compare the impact of each variant of the virus. Are the variants following a trend? News articles and scientists are expressing that the virus could become more contagious while the deaths will not be as prominent [1, 6]. For this question, the dataset will focus on US population numbers only. Within the question, queries like how each age group has been affected could be answered.
Q3 - Analyse the development of the vaccine campaigns. Has each dose been received at the same pace? It is interesting to see if time and the effects of the virus on people have made UK citizens more prone to getting the vaccine. Moreover, the question of are people more reluctant to get the newer doses like the third and booster vaccines [9, 10].

1.2 Data Types and Datasets

Extensive research of datasets was carried out for all the research questions. For this section the most appropriate data found will be introduced and assessed for each question, other less appropriate datasets can be found on the Miro board with their own assessments of appropriateness. The link can also be seen in the first part of the appendix.

1.2.1 Nextstrain Dataset

For the exploratory research question 1 the data required would be quantitative discrete data of the number of mutations of each variant over a period of time, in the UK. Moreover, the frequency would also help answer the question and showcase patterns throughout time. The temporal field is necessary to visualise seasonal patterns. The most appropriate dataset found was the dataset provided of Nextstrain. The time of each sample is given, and all the variants are given with all their mutations. The frequency of the mutations per variant is collected over the time period specified. Furthermore, the connections between each mutation are given to see if certain mutations are connected to certain variants in a specific way. Nevertheless, this dataset does not contain data beyond the date of October 2020, therefore the seasonal trends that could be detected can not be validated. However, the data within the year provided can still be analysed for smaller periods of time, for example months and weeks.

1.2.2 CDC Dataset

Research question 2 requires quantitative discrete data on the number of cases and the number of deaths for each of the variants of the Covid-19 virus. In addition, further breakdown of those numbers can be collected to see which specific groups have been impacted the most. For this question most datasets found for the UK only provided the cases and deaths, however, the specific variant was not mentioned. Nevertheless, the CDC dataset broke the timeline into periods for each variant [5]. Making the possibility of comparing cases and deaths in each period possible. The periods are defined with the most prominent variant and a further breakdown of the age groups affected is provided. Even though the exact number of each case is not measured since variants can continue to affect people outside of their period there were no more appropriate datasets that could break the deaths and cases in these groups. Therefore, out of all the datasets found for this section the CDC Dataset was the most appropriate to answer the proposed exploratory question.

1.2.3 UK Government Dataset

To be able to see the development of quantitative discrete data of the number of received vaccines throughout each dose has to be seen against time. This is to see if certain doses were more enthusiastically taken than others. Additionally, a separation of the data is necessary to compare regions of the UK and see if there are any areas that are more reluctant to a specific dose. To answer the exploratory research question the most appropriate dataset can be seen on the UK Government website [14]. In-depth data of the dates each dose was supplied and the regions with their own breakdown is provided. Both time and quantity of vaccines are given. However, the dataset groups both the third dose and the booster which in some cases it is not the same, as some people have had 4th doses as boosters [2]. It would be important to separate these numbers to analyse thoroughly if the newer doses are less popular.

1.3 Correlation

All the datasets contain the common field of time as a row or column of the dataset, and so all of them can be correlated in that manner. However, several problems appear with this, first the Nextstrain dataset finishes in October 2020 and so no further data is given leading to missing values. Secondly, the CDC dataset does not have a continuous period of data, for example, between the periods of Winter and the Delta variant there is a large gap in time. This could leave gaps in the correlation. Another problem is how the time is measured in the datasets, for instance, the dataset of the CDC dataset separates the data into months where the peak values are calculated whilst both Nextstrain and the UK government datasets have daily data entries. In addition, the UK government dataset is not completely up to date the latest 11 to 7 days are missing. Finally,

the only directly proportional datasets for the question are the ones answering questions 2 and 3, both of them have cases and deaths which can be seen following a similar pattern even though they are from different countries.

2 Part 2: Design and Discussion

For all visualisations there were a few concepts in common that all of them contained. These concepts can be separated into 3 categories.

- **Colour:** The use of a range of hues and different saturations was always taken into account to make the colours as accessible as possible. In the implementation, the colour scheme chosen was even set through a colourblindness simulator to be able to confirm the accessibility.
- **Size:** The visualisations were also all designed with the concept of size in mind to be readable for anyone. Most final implementations of the concepts designed would contain voice features for accessibility and the text could be changed to make it larger.
- **Aesthetics:** Finally, all designs were designed to be minimalist and not over-saturate the screen with information. The designs given were to be able to share with users of all levels of context to be able to understand. All the designs would incorporate a feature where the user can get more information or display an in-depth analysis of the visualisation to explore the information further.

2.1 Question 1

For the first proposed research question, the first concept was to create a timeline using line graphs where the cumulative number of the mutations is visualised to see where there are more or fewer mutations over time. Each strand would represent a new variant and the smaller strands are the clades. However, after a few alternative designs, the final design was conceptualised as a 3D connected bar chart, as seen in Figure A.1 in the appendix. The three dimensions would be time, the clade and the frequency of mutations. The interaction of the visualisation would let the user rotate and zoom into the graph to see specific sections in detail. Not only that but in certain angles, only two of the numeric channels would be visible and therefore the data could be seen with a particular focus in mind. Additionally, the user could hover their mouse and the exact coordinates of the data where the mouse is placed would be displayed, those coordinates would describe the data of the three dimensions [4]. Moreover, the design would incorporate animations of the development over time for the user to see the progress. From the research carried this visualisation is not common as it can be described as a combination of a 3D dendrogram and a 3D bar chart [7, 8, 11, 12]. The scale used in this visualisation would follow a similar scale to the one provided by the Nextstrain dataset. Where the time would be separated into days, the clades would define the different variants and the frequency could be given in percentages. The idea to create this particular design instead of following other more common visualisations was to clearly see the independent variants and their development over time. Compared to 2D line graphs the visualisation would provide further insight into the possible seasonal trends.

2.2 Question 2

The idea of a multi-leveled pie chart was thought when it comes to the challenge of visualising research question 2. This design would contain large interactions where the user would be able to click through the wedges and specifically see how each variant affected the population. A particular implementation would be done to make the design more accessible when it comes to the colours. When a specific wedge of the pie chart was selected and opened, the rest of the wedges would change colour to grey scale to decrease the output and let the user focus more on that particular information. The user would be able to open a new layer by clicking on the specific wedge and then close the new area by clicking on the center of the pie chart. As seen in Figure A.2 in the appendix, the design would have a custom-ability to it where the user could decide the order of the layers of information and could delete certain aspects they might not be as interested in. The only data that would be fixed would be level 0, the center of the pie, with the separations of all the variants of the Covid-19 virus. Moreover, on the screen when the user clicks on a specific wedge and this opens the specific breakdown of the numbers would be displayed on the screen. This way further information is provided and the design offers more transparency. In this prototype design, the user would be able to open two or more wedges and their respective layers at the same time. Thus, would let the user analyse and compare the exact data for each of the levels and variants.

2.3 Question 3

When it comes to the third proposed exploratory question and its visualisation the emphasis of the design was to make it highly interactive and have a more fun design. There are multiple designs that incorporate representative visuals to the graphs, and it is thought that this type of design is not only understood by more people but also eye-catching [3, 13]. For this particular, design the same hue of blue would be used but different saturations and values would be implemented. This monochromatic colour scale would in the end create a clear grey gradient for the users who have any kind of colour blindness. The colour of the region would change colour as soon as one person has had the vaccine, though it is more common to show a specific range the question is to see if there are any areas that did not get the vaccine due to a specific reason. This will visualise more drastically the areas that are more reluctant to the third and fourth dose of the vaccine. Not only that, but the change in colour would demonstrate the regions of the UK that are more enthusiastic about the newer doses and further analysis of other external factors could be carried out. An animation is provided in the Miro board following the link given in Figure A.3 in the appendix. The bar charts inside the syringes could represent the percentages of the population that have had the specific dose. Hence, that particular bar chart would only be visible once the vaccine dose is accessible based on the timeline. The data and the exact numbers could be seen with further interaction, such as, by clicking on a region or hovering over the bar charts.

References

[1] D. J. Dawson. Debunking the idea viruses always evolve to become less virulent, 2022.

[2] A. Finnis. Who is eligible for the fourth covid jab as spring booster is rolled out, 2022.

[3] A. Glivinska. The 25 best data visualizations of 2020 [examples], 2021.

[4] Y. Holtz. 3d — the r graph gallery, 2018.

[5] A. D. Iuliano, J. M. Brunkard, T. K. Boehmer, E. Peterson, S. Adjei, A. M. Binder, S. Cobb, P. Graff, P. Hidalgo, and M. J. e. a. Panaggio. Trends in disease severity and health care utilization during the early omicron variant period compared with previous sars-cov-2 high transmission periods — united states, december 2020–january 2022, 2022.

[6] D. Kopecki. The next covid variant will be more contagious than omicron, but the question is whether it will be more deadly, who says, 2022.

[7] N. LLC. 3d bar charts, 2021.

[8] W. MOSZCZYŃSKI. Dendrogram and clustering 3d - the data science library, 2019.

[9] J. A. Murugesu. Vaccine hesitancy: Uk booster campaign must reach out to pregnant women and people in ethnic minority groups — new scientist, 2021.

[10] J. Tapper. One in 13 covid double-jabbed in uk reluctant to get booster, study finds, 2021.
[11] U. Unknown. 3d bar charts — anychart gallery.
[12] U. Unknown. Beautiful dendrogram visualizations in r: 5+ must known methods - unsupervised machine learning - easy guides - wiki - sthda.
[13] U. Unknown. Visual capitalist.
[14] U. Unknown. Uk summary, 2022.

Appendix A

The project was also visually developed on a Miro Board where further research and further analysis can be seen.

The link to the Miro:

https://miro.com/app/board/uXjVO8p8Vv8=/?share$_l ink_i d$ = 555391113366

The implementation can be seen in Codepen:

https://codepen.io/carlotavagon/pen/gOoZOoB

or on Github:

https://github.com/CarlotaVagon/SDV

Miro link to the other specific assessed datasets:

https://miro.com/app/board/uXjVO8p8Vv8=/?moveToWidget=3458764523437583754cot=14

Links to all the datasets:

Nextstrain:

https://nextstrain.org/groups/ecdc/ncov/united-kingdom?gmin=280l=scatterp=full

UK Government:

https://coronavirus.data.gov.uk/details/vaccinations

CDC dataset:

https://www.cdc.gov/mmwr/volumes/71/wr/mm7104e4.htm?s$_c id = mm$7104$e4_w$

The link to the D3 example used in the implementation:

$https://codepen.io/danbrellis/pen/aEMGMp$

A.1 Figures of the Prototypes for Part 2: Design and Discussion

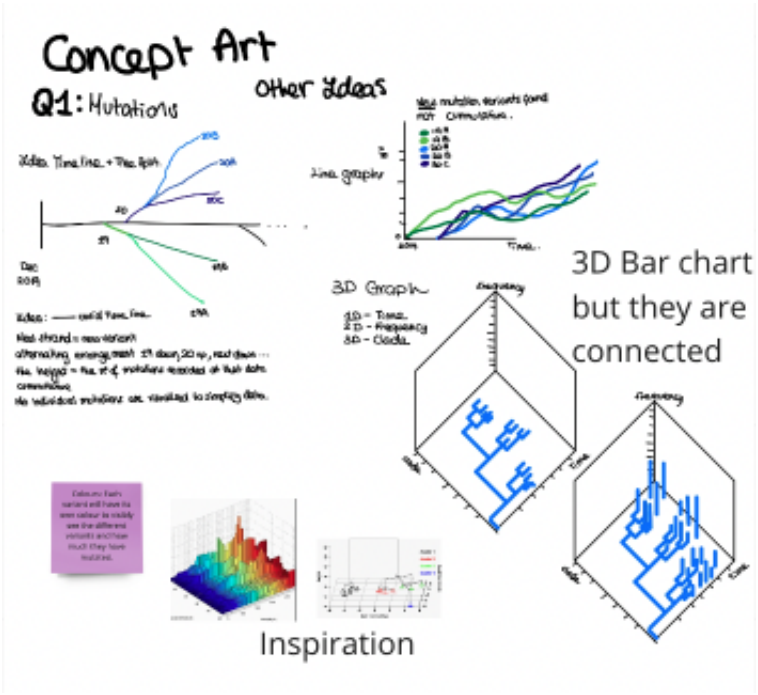

Figure A.1: Concept art of the exploratory research question one proposed. Design can be seen in detail following this link: https://miro.com/app/board/uXjVO8p8Vv8=/?moveToWidget=3458764523375416789cot=14

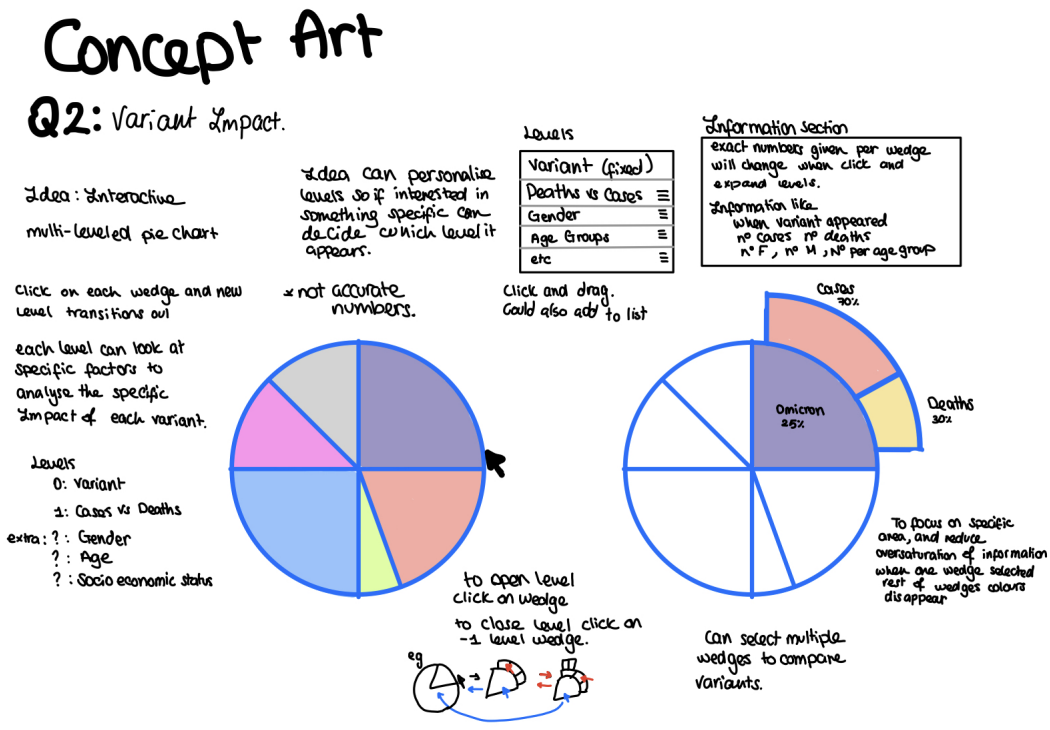

Figure A.2: Concept art of the exploratory research question two proposed. Design can be seen in detail following the link: https://miro.com/app/board/uXjVO8p8Vv8=/?moveToWidget=3458764523375416748cot=14

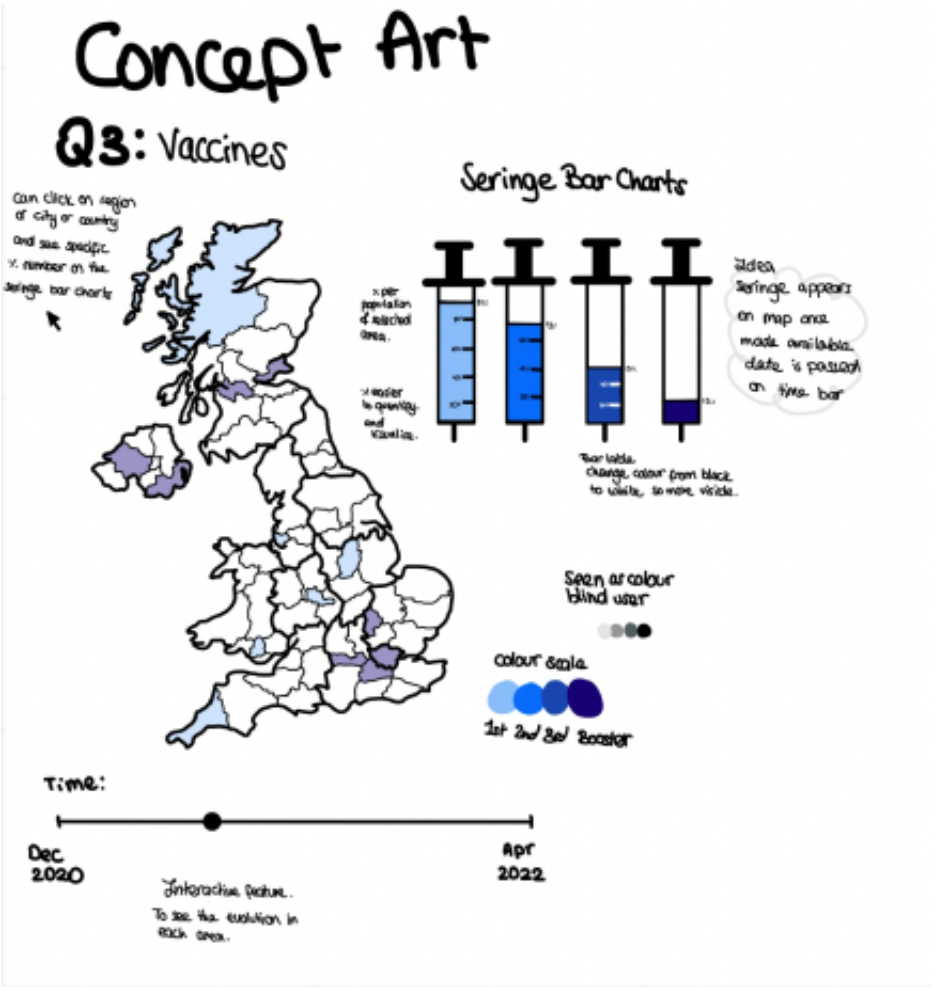

Figure A.3: Concept art of the exploratory research question three proposed. Design can be seen in detail following the link: https://miro.com/app/board/uXjVO8p8Vv8=/?moveToWidget=3458764523375416725cot=14

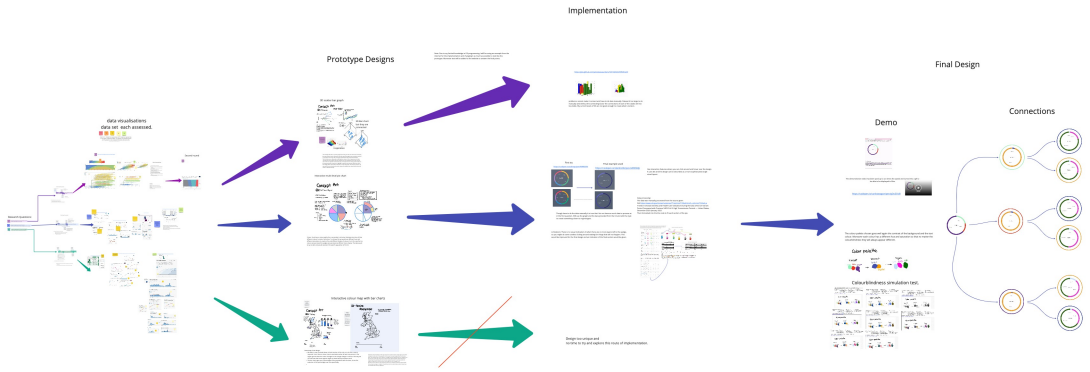

Figure A.4: The overall view of the visual representation of this assignment created on Miro Board of which the link is found on the previous section of the appendix.

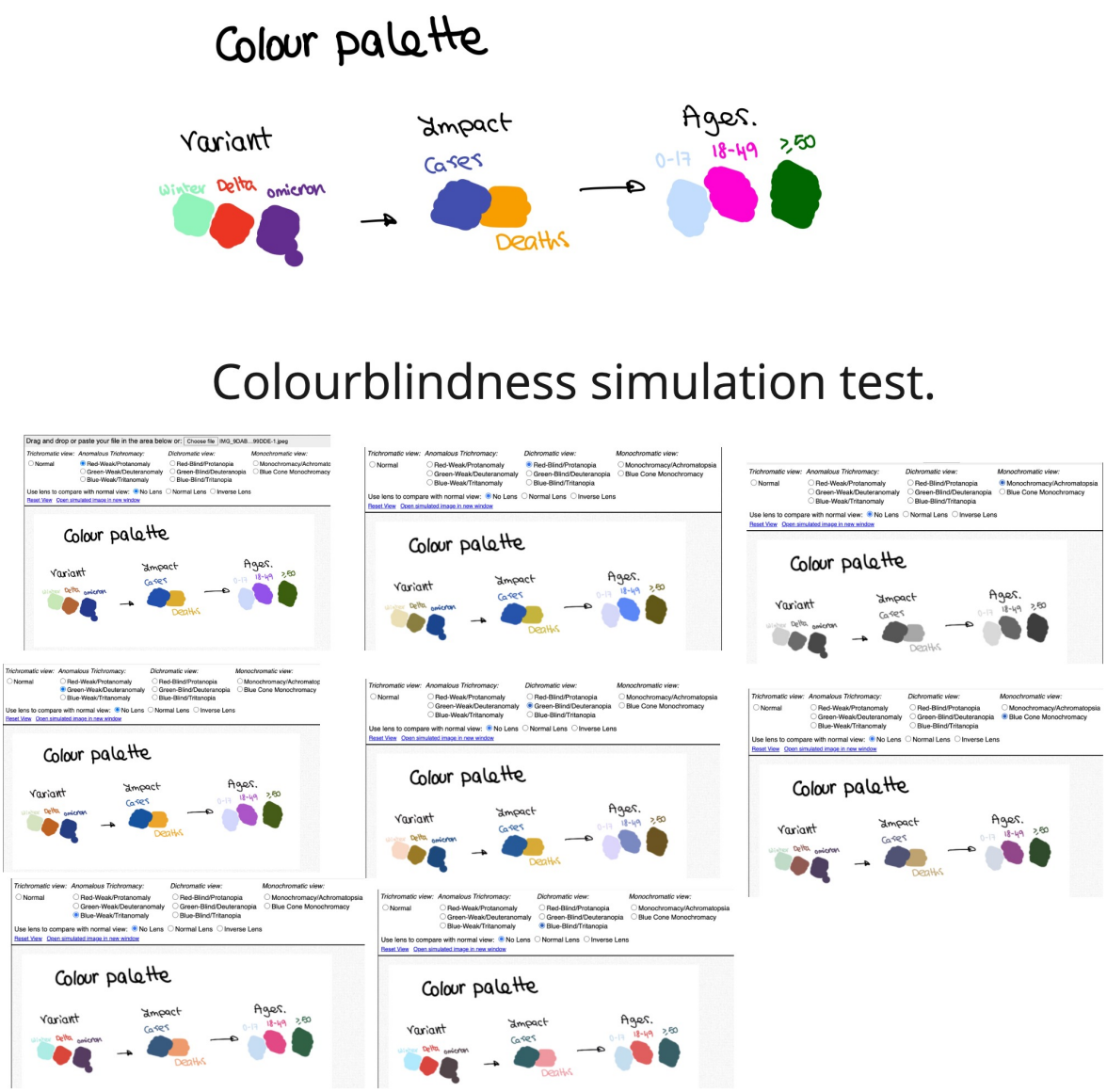

Figure A.5: The demonstration of the colour pallette that would be implemented in all prototypes and can be seen in the final visualisation. The simulation of colourblindness is used to portray its accessibility.

Figure A.6: An in-depth description of the data processing carried to apply it to the final visualisation. Link to this can be found in the previous section of the appendix.