2020

Narrative Visualisation Project Report

POURIA EBRAHIMNEZHAD
STUDENT ID: 30035678
TUTOR NAME: MOHAMMAD HAQQANI

Contents

1. Introduction

As my data exploration project suggested I analysed and explored any existing relationships between the spread of COVID-19 rate in different countries and various health, economic or environmental attributes. I aimed to answer mainly below questions.

1. Is there any relationship between the spread of COVID-19 and the median age of the countries impacted?
2. Is there any relationship between the spread of COVID-19 and the population density of the countries impacted?
3. Is there any relationship between the spread of COVID-19 and different health and governmental factors (i.e. smoking, life expectancy, Government spending on health)?

In the process I discovered some interesting trends and almost positive relationships in most of these factors with the spread of COVID-19 however some seemingly related more to the spread of the disease than others.

Reflecting on the underlying relationships between the factors, I discovered that age played a bigger role than I initially thought and, in a way, it influenced the relationship between these factors and spread of disease. For example, countries with higher life expectancy tend to show higher rate in the spread of COVID-19 as well and these countries not surprisingly have a higher median age. And this underlying relationship with age seemed to be consistent at least with respect to the data I sourced and explored. Meaning the less of a positive relationship the factor seemed to show with the spread of COVID-19 the less it had a relationship with age.

After careful consideration of the objective in this part of our course and discussions in tutorials, I decided that this could be a good story line amongst many observations from my EDA and could allow me to not only use visualisation to demonstrate this relationship but also to demonstrate the work I have done in the EDA section and allow the user to visually observe this narrative for themselves. As for the audience I decided that my audience would be the Data Science students as my findings where more geared towards having a basic understanding of the statistical relationships. Also decided that the title of my narrative visualisation would be "***Spread of COVID-19 and the undeniable influence of age"***

2. Design

For reaching the optimal narrative visualisation I had various ideas in mind reflecting back on the material that was discussed in the course and looking at many designs throughout the course of the project, the 5 sheet design methodology allowed for a good summarising of all the ideas and helped in reaching one final design. I also discussed and visualisation and the alternative designs with a few friends and colleagues and get feedback before considering the final implementation.

Brainstorm sheet

For brainstorming I considered all the ways in which I can show the relationship between the factors in my study and the spread of the disease also considered different ideas and interactivity levels the user could have, the mix included trend charts, scatter plot sorted bar charts and individual country comparison line chart, bubble charts and race chart categorised into the amount of interactivity it provides.

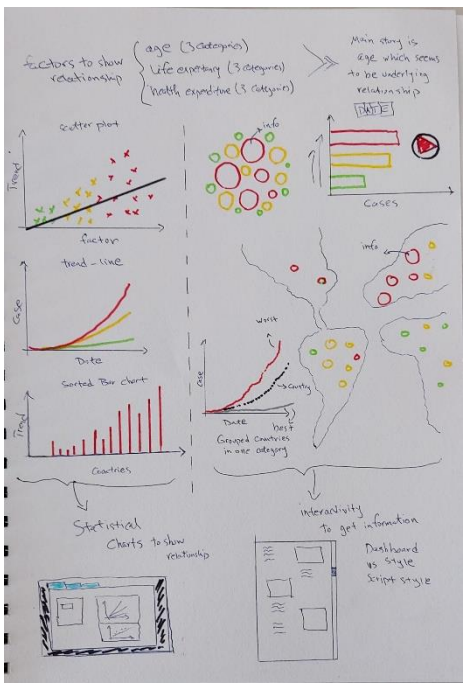


Figure 1 - Brainstorm sheet

Initial design 1

In this design, I chose a 2 tab page where the in the first tab the user gets to see the most important factor relationship between age and the COVID-19 spread and show some narration which discusses the study findings and how different factors influence the relationship this tab will set the stage for the user and also allow them to see the most important factor. In the second tab the idea is to show bar charts for the calculated Rate of the disease for each country sorted by any of the studied factors which would allow the user to see how the sorting shows the relative importance of relationship. Also, they get to choose a country and view its average

number of cases per day compared to the worst- and best-case country in that category.

The advantage of this design is that it allows the user to see the relationship between factors and spread of covid-19 and compare different countries Rates in their categories with each other.

The disadvantage is that there is a lot of interactivity and the user must move between the two tabs in order to see the whole story and sorting of the bar charts may not as clearly visualise the strengths of the relationships. Also, the complexity of the implementation has to be taken into account here as disadvantage given the time constraints of the project

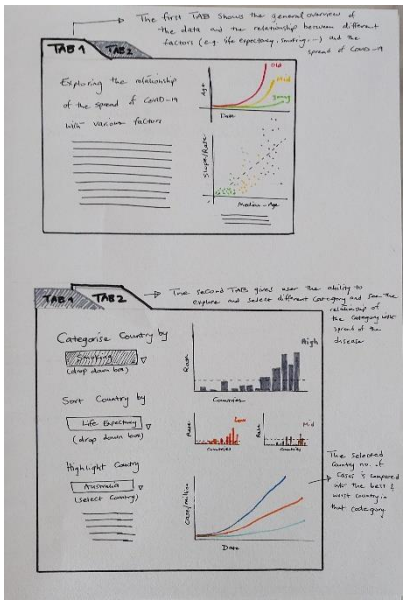


Figure 2 - Initial design 1

Initial design 2

The second design revolves around the idea of having a news article style with the different factors relationships explained by the trend chart or scatter plots where required and the narrative taking the user through the discovery of the relationships and discussing their relative importance and underlying relation to age, in contract to design 1 this design would give limited interactivity to the user and rather guide the user by providing a sequential story telling narrative to walk through from top to down on the page.

The advantage of this design could be that it is very easy for the user to navigate as they don't have any options and depending on the quality of graphs generated it could provide a clear visual demonstration of the message in each section and for each factor and allow more flexibility for the narration to be done effectively.

The disadvantage could be that the user doesn't have much interactivity and given the topic is COVID-19 if they wish to see their own country information and compare the factors that have been studied and presented they won't have this option and could just step through the story as they are confined to do. Implementation could also be challenging depending on the tool used and visualisations.

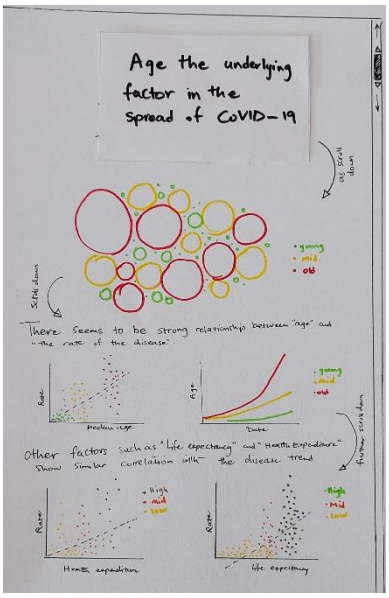


Figure 3 - Initial design 2

Initial design 3

Here the main idea of the design was based around an idea to give the user a holistic and concise way of viewing the narrative of the story and to present to them at first glance a bubble chart for each factor analysed with the size of each bubble proportional to the relative importance of the relationship with the Spread of COVID-19 and then when the user clicks on the bubbles a second window will open which shows them the relationship graph including the scatter plot and a line chart of the average number of cases grouped by the attribute.

There would also be a small hidden narrative which the user could click on to see the narration related to that section. This was the user only works on one page and everything is there for them to view with their choice and exploring the narrative of the story while setting the stage with the first initial bubble chart

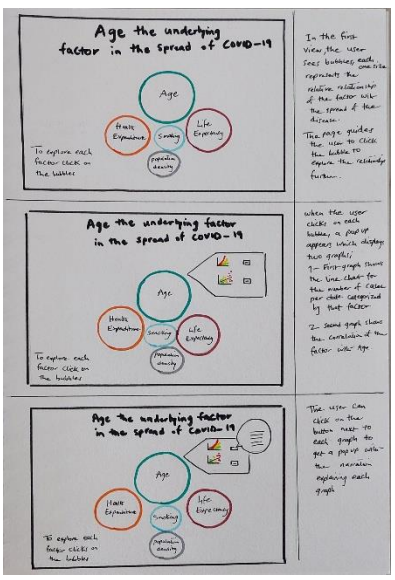


Figure 4 - Initial design 3

The advantage of this design is that the user is given relatively some part of the message in a concise view at the beginning and then the supporting evidence is presented to them by clicking on the factors.

The disadvantage is that the design by its nature may not allow for full narrations to be done and could require the user to click to see the narrative in order to allow for space to be used efficiently. implementation could also provide a challenge and complicated to implement with multiple nested popups and requires advanced knowledge of JavaScript and D3 to implement.

Final design

The Final design which I aimed to implement was based around the idea to give the user the narrative visualisation of the story around the relationship of the various factors and the spread of COVID-19, as well as their correlation significance with age. At the same time giving some freedom to the user to explore their own country of choice and visually enable them to tell the relative importance of the attributes and not making the design too exploratory.

The design was influenced by having a world map in the background to allow the view of information per country and provide the user with a relative holistic comparison of the Rate of the disease with the use of proportional symbols while at the same time taking advantage of colour hue and saturation where required to differentiate the categories of the countries by the factor being viewed.

In this way I am allowing the viewer to visually compare if one colour hue is dominating in size of the symbol than another, at the same time allow them to get more detailed information when hovering above the country of their choice.

Also, the design allows to take maximum advantage of the space in one page without the need of navigation to another tab or scrolling up and down. The user can simply view the relationship graphs and the required narrative on top of the map in a floating transparent board while guiding them to choose the attributes to see the importance of each one. The legend on the map also provides more information regarding how the categories were split from the EDA results, rather than having this legend for each of the charts and making them redundant. They can also zoom in on the map where required and move the transparent panel around to get a better view if required.

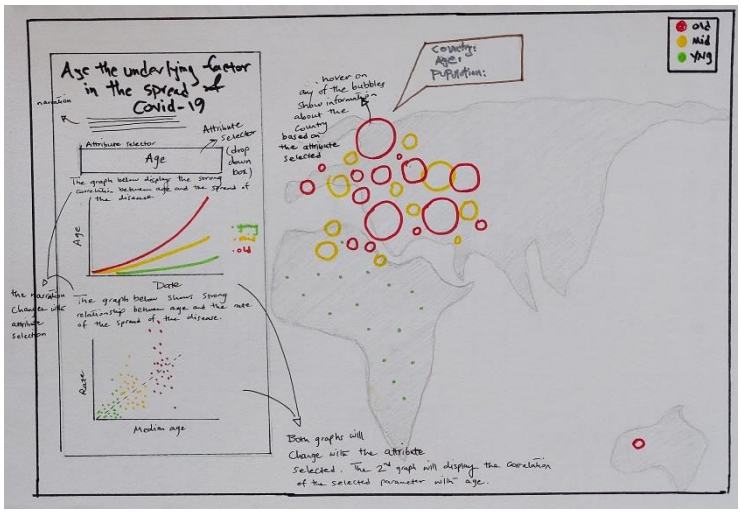


Figure 5 - Final design

3. Implementation

For the implementation I decided to use Shiny in R for two reasons one was its relative flexibility around designing the page with the leaflet map background and my familiarity with R which allowed me to deliver the project within the timeframe intended.

First off, I implemented a reshaping tool in R to allow me to extract the data in the shape and form required to allow for easier implementation of my statistical charts. Basically I needed a tabular data with each country attributes as well as the calculated rate from my EDA and I also required a data frame which calculated the average number of cases for each calendar date per countries grouped by their category for each attribute, this would allow me to create the trend plots for each factor per group.

For the development of the interface I used a base template of a shiny app (Parker, n.d.) in the shiny gallery template (2020 RStudio Inc., 2020) that allowed me to implement the idea I had in mind and modified to cater for my visualisation needs.

After reading the CSVs generated from the reshaping step in the code, the next step was basically to do some simple wrangling tasks such as fixing date column and add order to the categories for each attribute. The order would allow me to modify the plot legends appropriately.

Then I wrote some functions which would be used in the server side to generate the required plots when the appropriate selector is chosen in the UI side and passed as input to these functions, each function would set the required attributes for the

plots and update case conditions to allow the plots to carefully display the appropriate information as well as selecting a colour pallet that could effectively distinguish the categories in accordance to what we have learnt as part of this course.

The UI code would include a tab panel page within which I would use leaflet map, the hovering panel on top of the map and the panel itself would include the narrative of the story as well as the plots and the selector to allow for attribute selection, part of the narrative would be fixed in this panel and the parts which should change dynamically with respect to the attribute, using the textoutput feature in shiny.

On the server side and using the reactive feature in shiny I populate a variable for the selector used by the user, based on this value I will populate appropriate narrative as well as call the plot passing in the selector value to ensure the correct plot is being displayed in the panel.

The next section of the server code will use the renderleaflet and leafletproxy to allow me to populate the background map with markers that are colour coded to the specific pallet reflecting the chosen attribute this cool feature allows the viewer to consistently see the categories colours on the map as well as visually see the trend on the graphs. There is also matching legends which will populate indicating the category thresholds used in the study.

4. User guide

In order to run the visualisation. The COVID_App.R file and the supporting csv and other files provided with this report need to be in the same folder for the app to run properly. The app can be run from R-studio which will result in the page to load. The best view of this visualisation would be on a wide screen and aspect ratio of 16:9 which allows the panel to fit nicely to the screen and best initial view.

At first load The floating panel will have the first attribute selected as Media Age where the user would be able to see the title, the visualisation purpose and the narration for the selected attribute. Also, the two plots will load one indicating the average number of cases by date for each category of countries and the other would

be the scatter plot displaying the relationship of median age and the COVID-19 Rate coloured by each age category.

Figure 6 - View of the Shiny App

There is also a footprint in the floating panel to ensure the user knows caveats about the data including the date range, how rate has been calculated and some further information regarding the study.

They would also see the world map with the proportional circles representing the COVID-19 Rate calculated for each country. The symbols are coloured according to each country median age category with a legend where the thresholds are displayed for each category. Hovering above each country would show the user the name of the country, population, the rate of the disease as well as the median age of the country.

The panel can be moved around if the visualisation is viewed on a narrower aspect ratio to allow the user to view the entire map and also the map can be zoomed in and out for better view of countries.

The narration also guides the user to select each attribute and view its relationship with the spread of the disease and with the median age attribute.

When the user selects other attributes from the drop down the plots update as well as the narrative and the map colour coding will change to reflect that attribute category for each country, hovering over the countries would display that attribute value along with the country name, population and Rate of the spread of COVID-19.

Figure 7 - hovering over countries view

The important distinction here is that the second plot will now display the correlation between the selected attribute and the median-age highlighting how much that factor has a correlation with age. The user can view the separation between the categories on the line chart as well as the map which indicated how important that attribute relationship is with the Rate of the disease.

5. Conclusion

Overall, when comparing the relationship of these factors with the linear trend of the spread of the disease one could see some positive relationship in all cases with.

At face value it's an interesting and maybe an unexpected observation to see that governments who have spent a larger percentage of their GDP on health have shown a faster trend of spread of the COVID-19, also those with a higher life-expectancy are showing a faster rate of spread. However when looked carefully at the relationships we can see that the more a factor is correlated with age the more it seems to have a relationship with the rate of the disease and age seems to be the underlying factor here, However this is most likely not the only factor and other things such as amount of testing done in countries could greatly impact this analysis, however as for the story line to talk about this study, it could well provide a catchy story for data science students specifically and could generate some initial interest to look into the visualisation and compare the trends and relationships to reach the same conclusion.

I was expecting to find a more meaningful relationship between smoking data and its relationship to the spread of the disease which doesn't seem to be as significant here and also the population density of countries which is surprising to see. However there maybe caveats here in the data, for one I replaced the missing values of smoking data in a lot of the countries with the average of this value which could have impacted the relationship, also not all countries had most recent data for this parameter which makes the observations questionable. With regards to population density I used the data provided in data.worldbank.org (Group, 2020) which provides description for this data as people per square kilometre of land area in each country. If I could've sourced data regarding average density of the countries with respect to their cities, I would have expected different results using the same analysis I have done.

As for the shiny app and the techniques used to make the final visualisation I enjoyed the learnings and appreciate the chance I got to make this app as I think it allowed me to gain at least some understanding and exposure to a new tool which I think could be easily used to help with publishing similar projects in the future, the challenges of making the app reactive to the selections and making the charts as well the map in a way to allow for a meaningful visualisation was interesting for me to deal with and I enjoyed the project for the most part EDA and the final visualisation.

In hindsight I would have liked to be able to make the charts in D3 and to integrate with the shiny app this would have allowed me to familiarise myself with D3 as well as shiny and would possibly create better quality graphs which could have allowed for some further interactivity on the graphs which is not possible with ggplot in R. unfortunately due to my limited knowledge of JavaScript, D3 and CSS and due to the project timing constraints, I was not able to experiment with this idea.

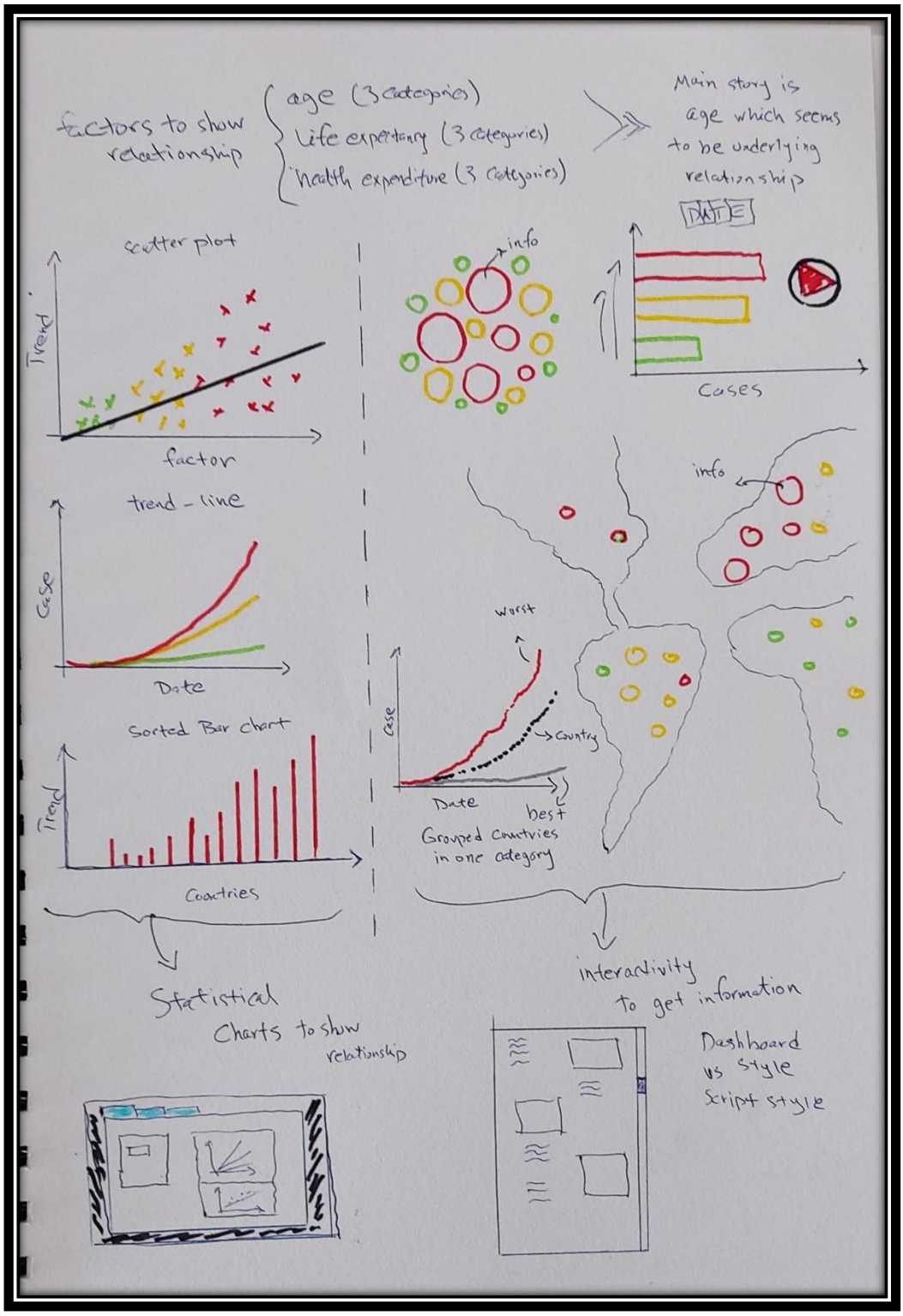
Also, I would have liked to have the chance to further study the relationship of other environmental and socioeconomic factors such as Cardiovascular disease, respiratory disease, income, education, average temperature, etc. these would have required further data sourcing, wrangling and may have led to discovering other interesting trends in the EDA part of the project and possibly a more interesting story to tell using the data.

6. Bibliography

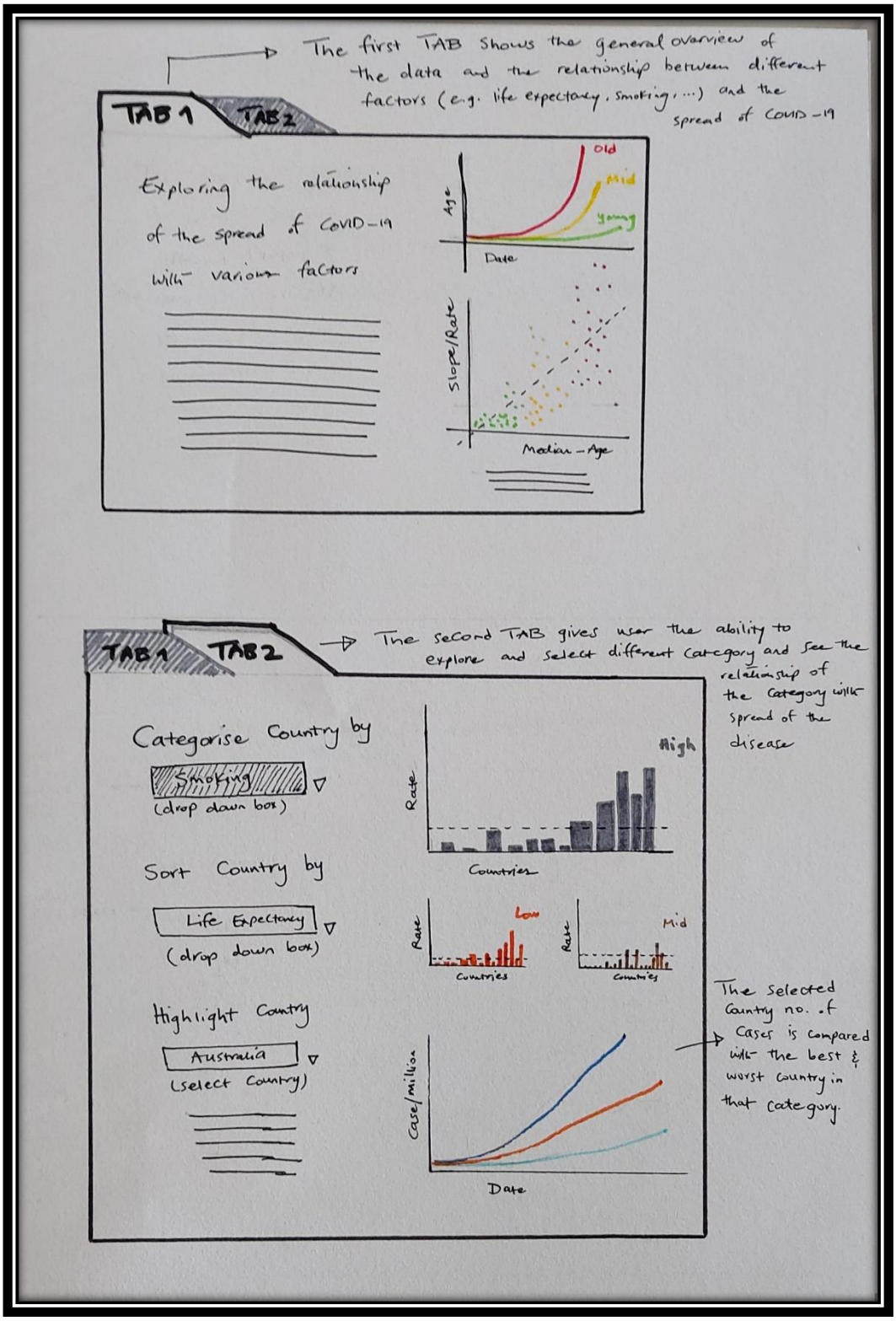
- 2020 RStudio Inc. (2020). *Shiny from R Studio*. Retrieved from https://shiny.rstudio.com/gallery/
- community, T. S. (2017, March). *scipy.org*. Retrieved from docs.scipy.org: https://docs.scipy.org/doc/scipy-0.19.0/reference/generated/scipy.stats.mstats.linregress.html
- Group, W. B. (2020). *The World Bank*. Retrieved from worldbank.org: https://www.worldbank.org/
- OCHA. (2020). *HUMANITARIAN DATA EXCHANGE*. Retrieved from data.humdata.org: https://data.humdata.org/dataset/coronavirus-covid-19-cases-data-for-china-and-the-rest-of-the-world
- Parker, E. (n.d.). *COVID-19 tracker*. Retrieved from https://shiny.rstudio.com/: https://shiny.rstudio.com/gallery/covid19-tracker.html
- Review, W. P. (2020). *World Population Review*. Retrieved from https://worldpopulationreview.com/countries/smoking-rates-by-country/
- Teggart, A. (2016, February). *The Data School*. Retrieved from https://www.thedataschool.co.uk/anuka-teggart/tipweek-calculating-z-scores-in-tableau/

7. Appendix

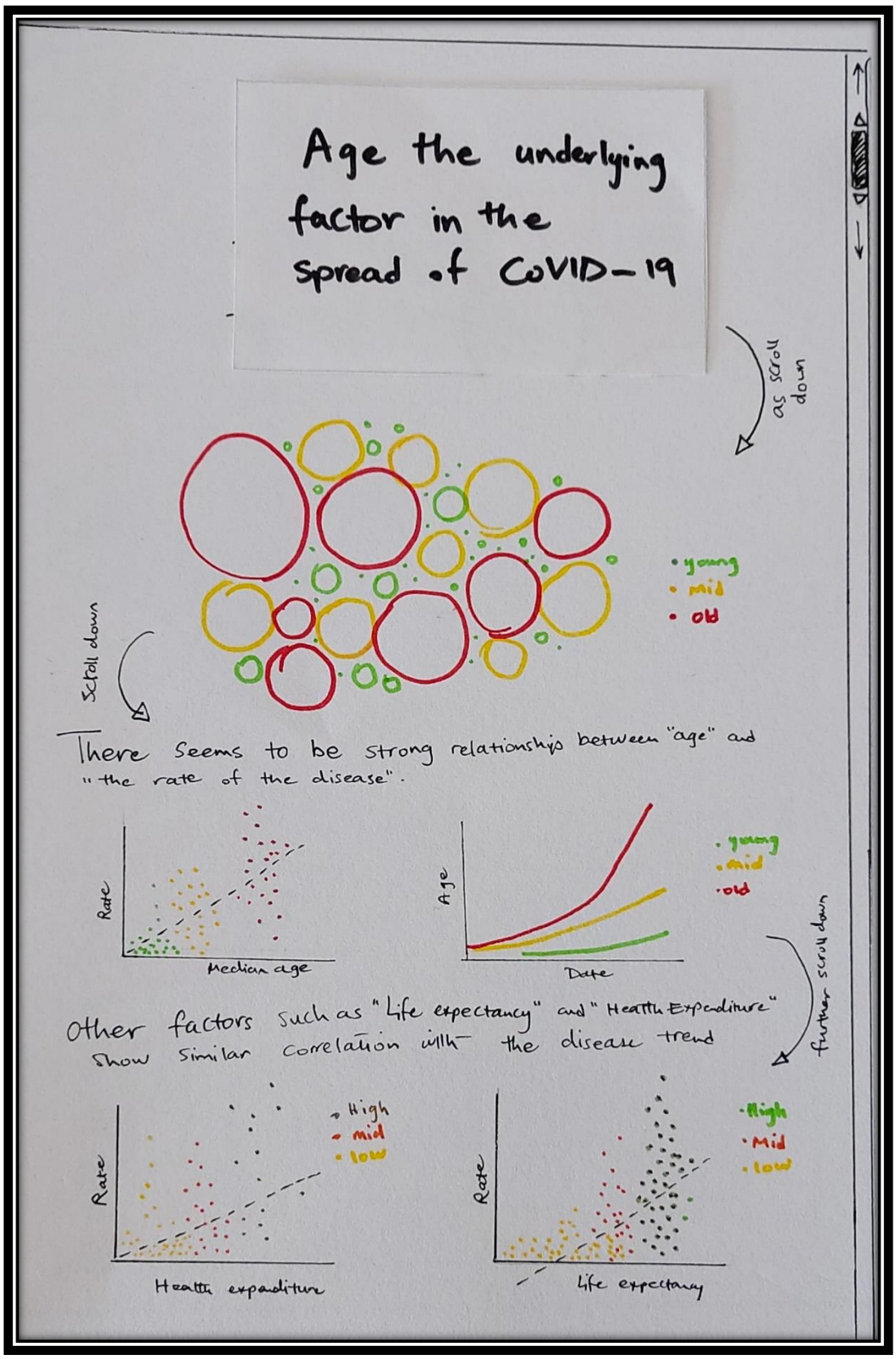
Brainstorm Sheet



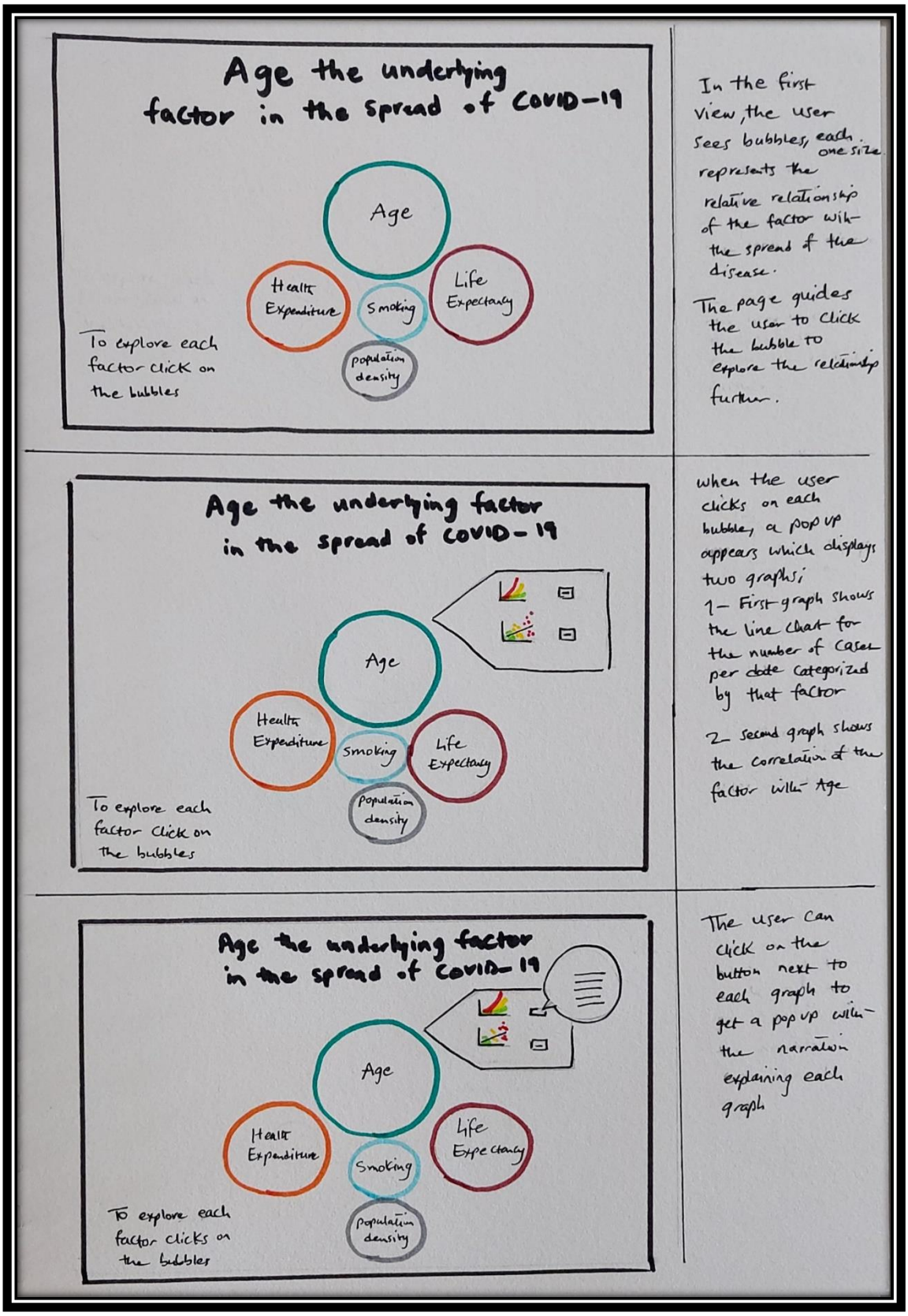
Initial Design 1



Initial Design 2



Initial Design 3



Final Design