Final Report
Global COVID Data through the Lens of the World Happiness Report

Hasmitha Chatla
Elizabeth Kerrigan
Sanjana Kulshrestha
Aditya Lanjewar
Endri Rama

Introduction

Our group will be investigating whether there has been any correlation between COVID-19 data and the following attributes: perception of corruption, social support, freedom of choice, and health life expectancy from the 2019 World Happiness Report dataset.

The dataset Impact of Covid 19 Pandemic on the Global Economy was found using Kaggle and taken from The Impact of Covid-19 Pandemic on the Global Economy: Emphasis on Poverty Alleviation and Economic Growth by Prince Asare Vitenu-Sackey who obtained the data from Our World In Data. The dataset consists of 8 variables and 50,419 observations. The data comes from 204 countries which serve as our categorical variables. The other variables used are date, total cases, stringency index, population, GDP per capita, and human development index, all of which are numeric, except for date which will be used for time series analysis as ordinal data. The dates range from December 31, 2019 to October 19, 2020, though some countries do not have data for the entirety of that time.

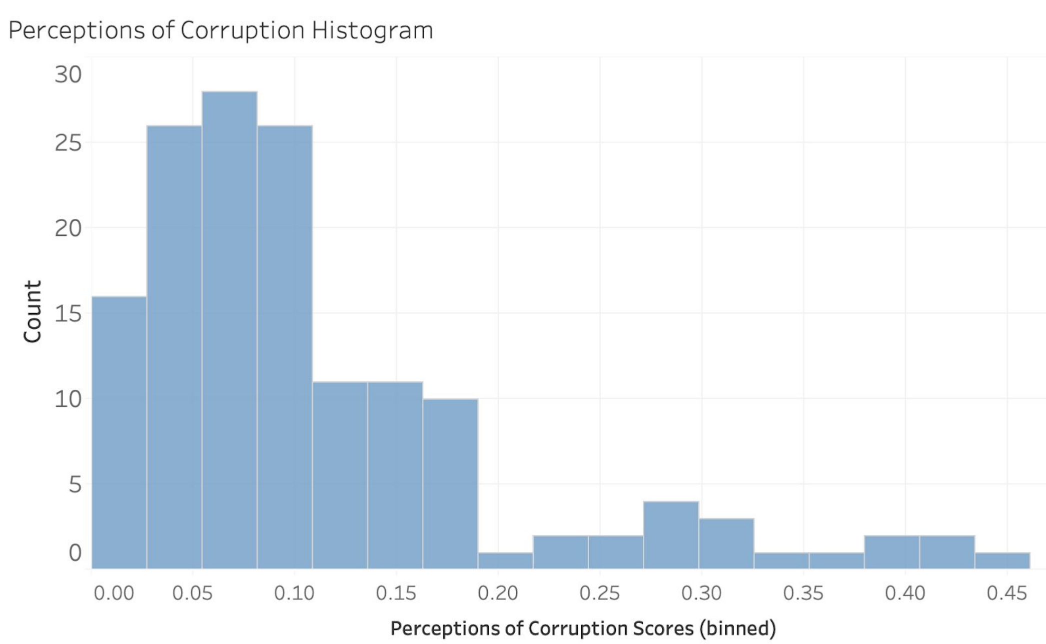
The 2019 World Happiness Report uses 156 countries to link 6 key factors to happiness. We will use some of these factors to see if there is any connection between them and a country or region's COVID response. These attributes include social support, perception of corruption, healthy life expectancy, and freedom of choice. Healthy life expectancy is constructed from the World Health Organization's Global Health Observatory data repository. The social support value for each country is the national average of binary responses to the Gallup World Poll (GWP) "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?" The freedom to make life choices is the national average of binary responses to the GWP question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?" Perceptions of corruption are the average of binary answers to two GWP questions: "Is corruption widespread throughout the government or not?" and "Is corruption widespread within businesses or not?" We will examine how these four features connect to and possibly impact national pandemic responses using the COVID-19 data.

Exploratory Analysis

We began by first exploring our two datasets separately before we looked at how the two interacted. By first analyzing the distributions within the COVID dataset, we were able to gain an understanding of how attributes like population and case number interacted. We have also included other areas such as corruption, freedom, social support, and life expectancy. By incorporating the Happiness dataset we could examine whether some of these countries' less tangible attributes could have affected their COVID-19 responses in the first 10 months of 2020.

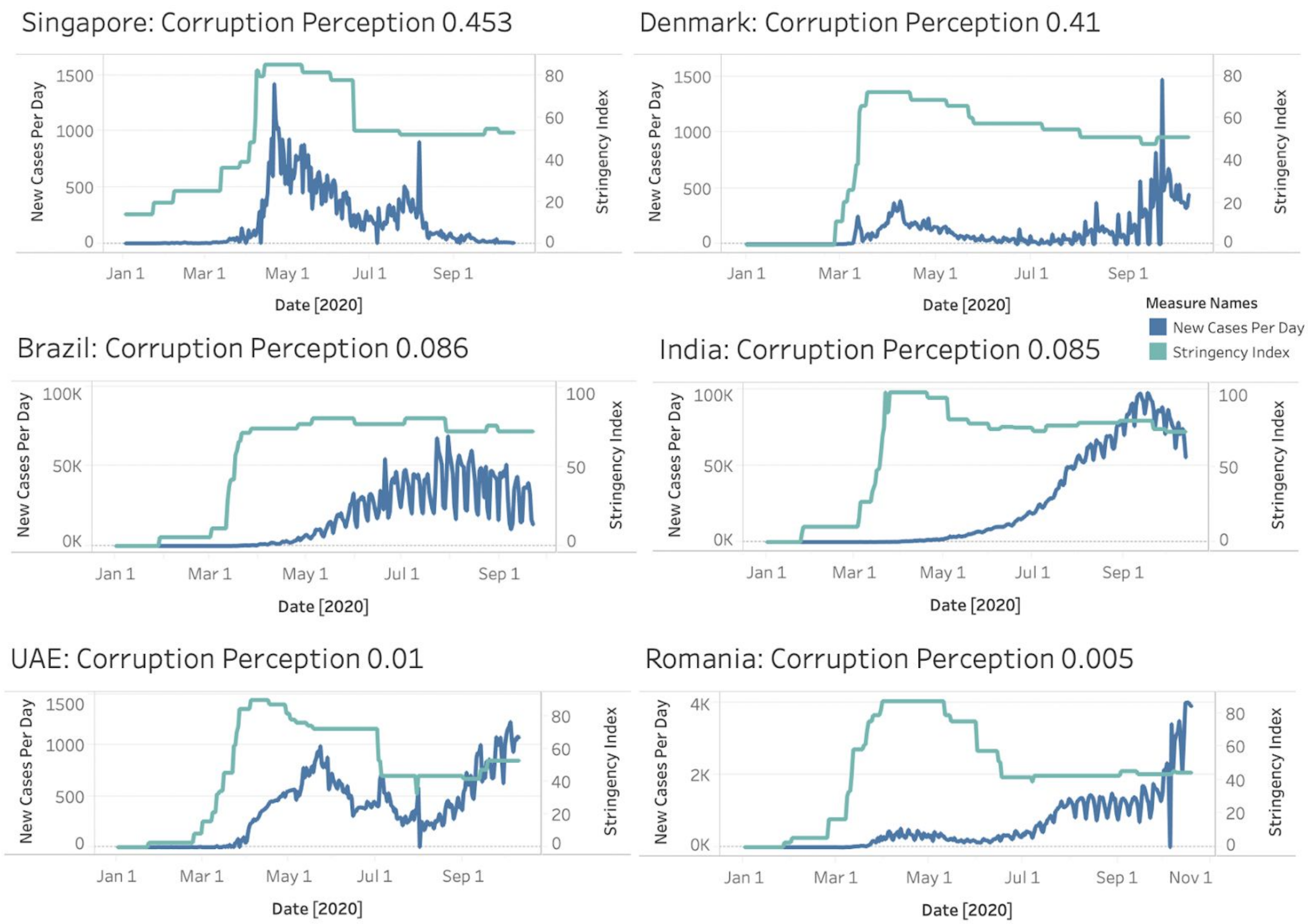
Examples of these preliminary exploratory visualizations are included in the appendices. These were primarily done using Tableau because it's a little quicker to just give us an idea of what was happening within the datasets and what would be interesting to pursue.

Here is one example of how we approached exploratory analysis. I've included this histogram to help highlight that most countries' trust in government is quite low, and most countries believe that there is corruption within their government. The "perception of corruption" score is taken from a Gallup survey of citizens of that country, not based on certain metrics pertaining to laws. The median value is only 0.086 meaning only about 8.6 percent of those surveyed do not think their government is corrupt. While we do have some countries on the higher end of the spectrum, there are fewer and even the highest country's trust is less than half of those surveyed. This histogram showed us that we needed to approach perceptions of corruption in distinct brackets and use the median rather than the mean to determine the middle since the distribution was so skewed.



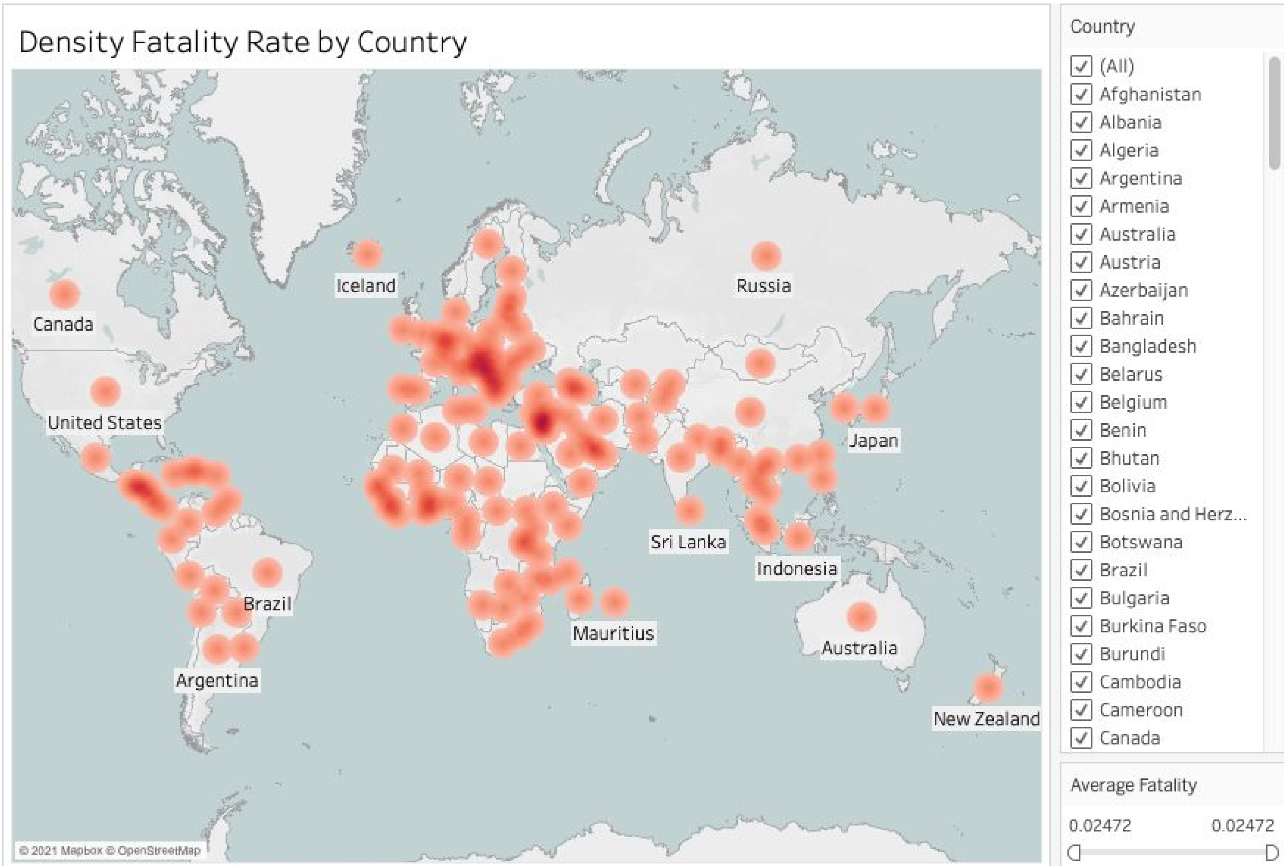
Visualizations

How Perceptions of Corruption Impact COVID-19 Response:



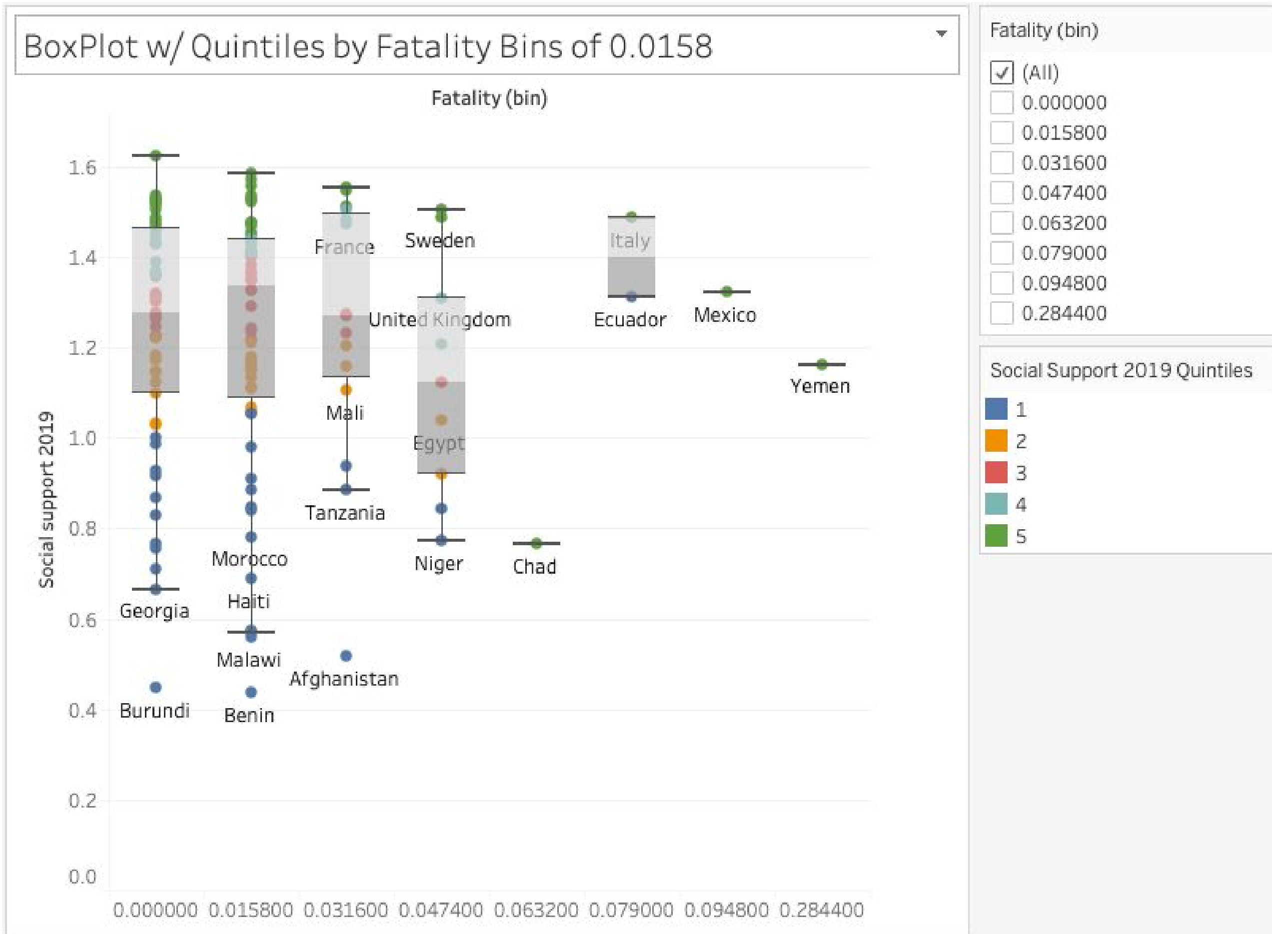
This visualization uses small multiples to show the new cases per day and stringency index for each of these six countries using time series plots from in the first 10 months of 2020. The three variables used here are date, new cases per day, and stringency index. If there was a missing value for a stringency index at a certain date, it was imputed with the value of the previous day. The stringency index typically did not change dramatically from day to day and as there were few missing values, this was a simple way to impute those values without distorting the visualization. The countries we chose did not have missing values for new cases of covid each day as the values were determined by subtracting the total cases of that day from the previous day so if the value was 0 for new cases per day, it just meant that the total cases had not increased from day to day. When we first started this process, we tried to work to create averages for different sections like highest and lowest, but this became very messy and crowded. we instead pivoted to this method of looking at a few individual countries in the same perceived corruption bracket. We started with trying to do five countries for each bracket. This was not helpful as there was too much information and some countries did not make sense to compare based on their numbers of cases. By only comparing a couple countries in each tier, we realized we could choose countries with more similar case numbers so that for each tier, the axes could be the same next to each other and help the audience see the comparison without misleading them by having different axes for each bracket. There are differences in case numbers between tiers, but we strove to make the magnitudes the same so that for each spike in cases, the stringency index still represents the response at the peak for that particular country . Once the concept of the visualization was finalized we played with the details by adjusting the colors and line thickness to what seemed most appropriate. We wanted the lines to be visible without being overpowering. Since this visualization relies on noticing trends having a very thin line to see every tiny change for each day was not necessary and distracting. The colors were chosen such that they were distinct from each other but not so different that they were harsh to the audience's eyes. By using colors of the same hue, your eye can follow them both for each graph easily. They are distinct enough that you can focus on one color or the other for all six individual countries to see the trend. The colors were also chosen to be more similar since the types of lines in the graph are distinct. The more jagged lines are the new cases per day and the "blockier" lines are the stringency indexes. This visualization uses an attribute from the Happiness dataset to take a closer look at how specific countries responded to the COVID-19 pandemic. By using the perceived corruption in a country as a measure of government trust, we could use the stringency index, the government's response to COVID-19, as a metric to see whether perceived corruption could predict how the government would respond to the pandemic.

Density of Fatality Rates by Country:



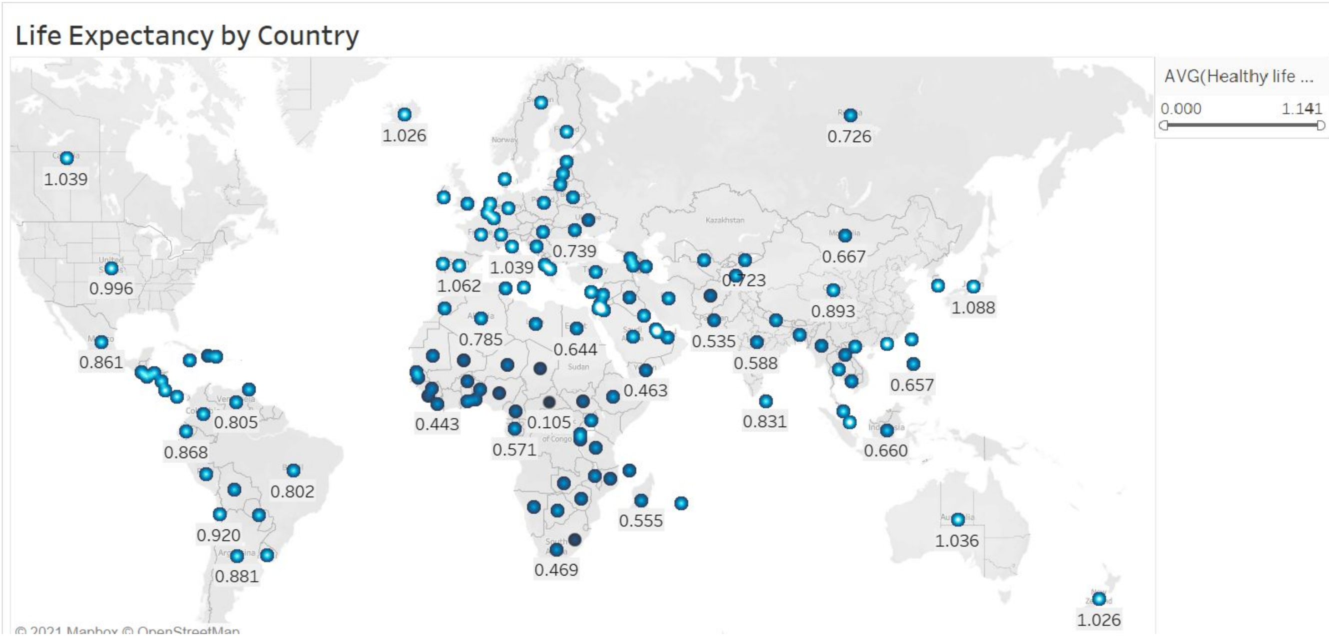
This world map shows the fatality rate for each country given the darkness of the red of the glyph. This map helped us get an idea of which countries had been hit worst by COVID-19. We can also use the Tableau features to filter out countries as well as using the average fatality rate legend to filter out directly on the average fatality rate.

Box Plot with Quintiles by Fatality Bins of 0.0158 Distribution:



This graph was created in order to see fatality rate by bins and its correlation to social support within each box plot. The quintiles of social support have also been created to see the countries in which social support bins have fallen.

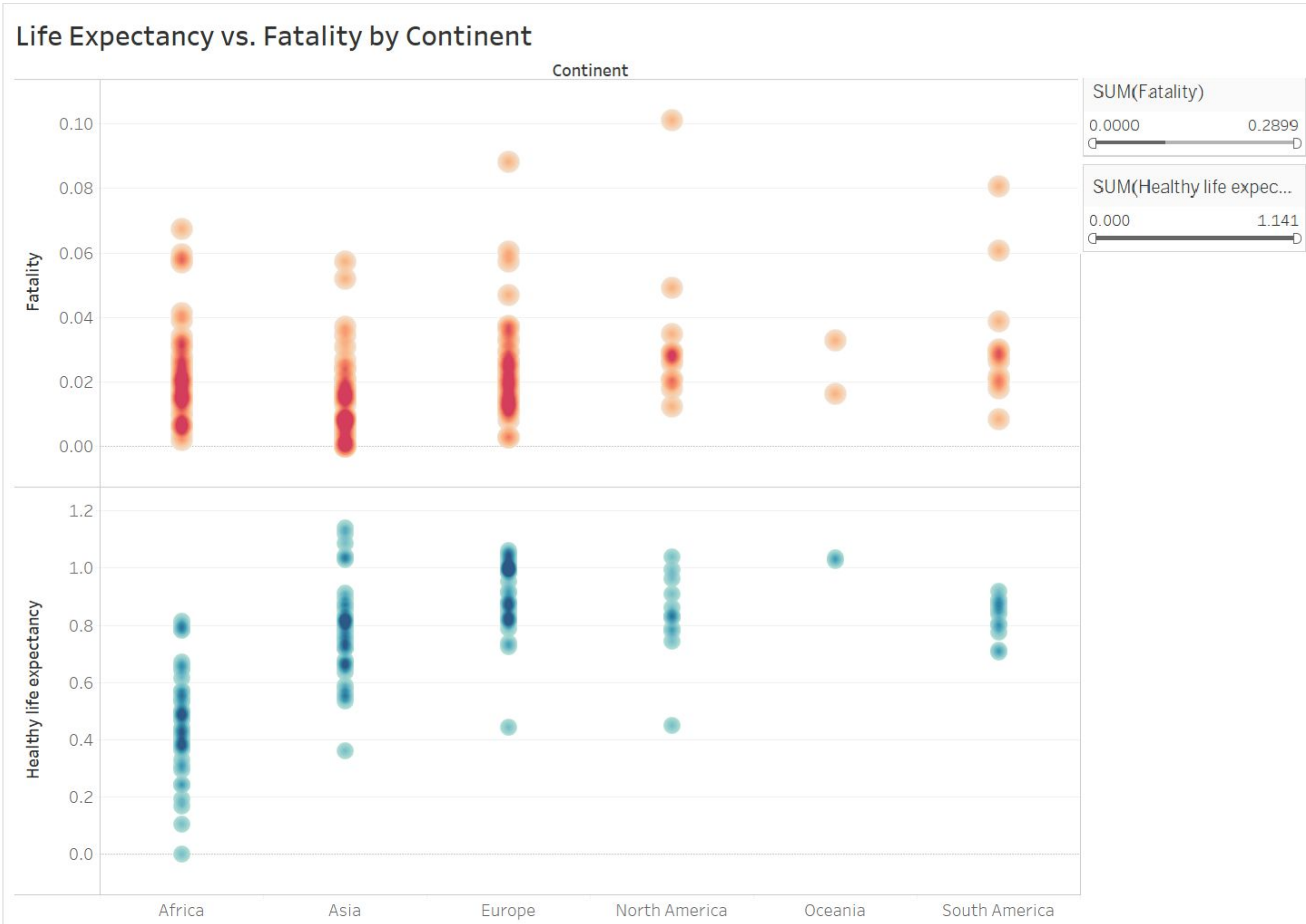
Life Expectancy by Country:



We investigated the life expectancy per country. We later compared the life expectancy by continent with fatality to see how these may be correlated to each other. We created a geographical map mixed with a density map using Tableau to show average life expectancy by country. Darker dots are lower life expectancy values and lighter dots have higher expectancy values.

Using this map, we can see how life expectancy across different countries all together. With the help of life expectancy values, we can also examine other related factors like population and fatality.

Life Expectancy vs. Fatality by Continent:



We created a scatter plot with dual axes. We used density to highlight the correlation between life expectancy and fatality. We used red color for fatality and blue for life expectancy. The darker the circles, the more countries in that fatality or life expectancy range. This visual helps us to compare different values together for specific continents.

Analysis and Discussions

How Perceptions of Corruption Impact COVID-19 Response:
We chose to see how countries reacted to their new daily cases of coronavirus by changing their stringency index. The stringency index is a measure from 0 to 100 based on how strictly a government was shut down due to COVID-19 with 0 being the most open and 100 being completely locked down. This metric was determined based on things like school closures and travel bans. The same scoring criteria was applied uniformly to all countries. The dark blue is for the new daily cases and as you can see is a bit more sporadic than the lighter blue of the stringency index. We compared pairs of countries from different regions with similar corruption scores. We look here at high, medium and low corruption scores. We tried to compare countries within the same tiers with more similar case numbers. We also plotted them on the same scale so one could more easily compare between the two countries in that perceived corruption tier. What's interesting here are the overall trends, whether countries followed the changes in case numbers by changing how locked down they were. For instance, Singapore is a great example because as their cases sharply increased, so did their stringency index. Indicating they shut down more with a higher volume of cases, and their stringency index decreases following the decrease in new cases per day. The two countries with the lowest trust in their government, the United Arab Emirates and Romania, had their stringency indexes lowered despite rising cases which were then followed by large spikes. As we were unable to apply this to all countries in the dataset, an area of future exploration would be looking at more countries and seeing if these trends follow or appear to be more isolated incidents, impacted by factors unaccounted for in this dataset.

Density of Fatality Rates by Country:
The team was exploring how fatality rate differed by country. We wanted to know how the fatality rate was distributed for country in order to find out what other fields might be useful to examine their relationship with fatality rate. A density by fatality rate geographical map was created for figuring out what other fields might correlate well with fatality rate.

Box Plot with Quintiles by Fatality Bins of 0.0158 Distribution:
This discovery provided the team with the knowledge to investigate the social support and fatality rate bins correlation. A boxplot with quintiles of social support and fatality rate bins was created in order to see the correlation between them. We wanted to investigate if a country had historically high social support it would have a lower COVID-19 fatality rate. This box plot shows all of the countries distributed within the fatality rate bins and their correlation with social support. The distribution of the quintiles within each box plot showed us that even though a country might have had high historical social support it will not necessarily lead to lower fatality rates. The box plots with the 0.0158, 0.0316, 0.0474, 0.0790 are all proof of this because what we found out was that there was no correlation between social support and fatality rate. It just showed us how the two were distributed in each box plot.

Life Expectancy by Country:
As per the map, we can analyze the higher average life expectancy value is around 1.039 and the lowest with 0.105. Darker marks show less than average life expectancy while lighter marks indicate above average life expectancy.

Life Expectancy vs. Fatality by Continent:
After analyzing the visuals, the team found both life expectancy and fatality are somehow correlated. We can see continents with higher healthy life expectancy with lower fatality rates. On the other side, continents having lower healthy life expectancy had higher fatality rates.

Conclusion

The 2019 World Happiness Report data provided us with a different way of looking at COVID-19 data by examining whether things like social support, freedom, health life expectancy and perceived government corruption were potentially related to a country's pandemic response. Since much of the happiness metrics were subjective and there are so many things that could affect COVID response, we cannot make claims about their impact but we did see some interesting connections. A possible area for future exploration would be to compare happiness data after the pandemic with prior to the pandemic and see whether COVID-19 had any effect on these Happiness Report features.

Appendices

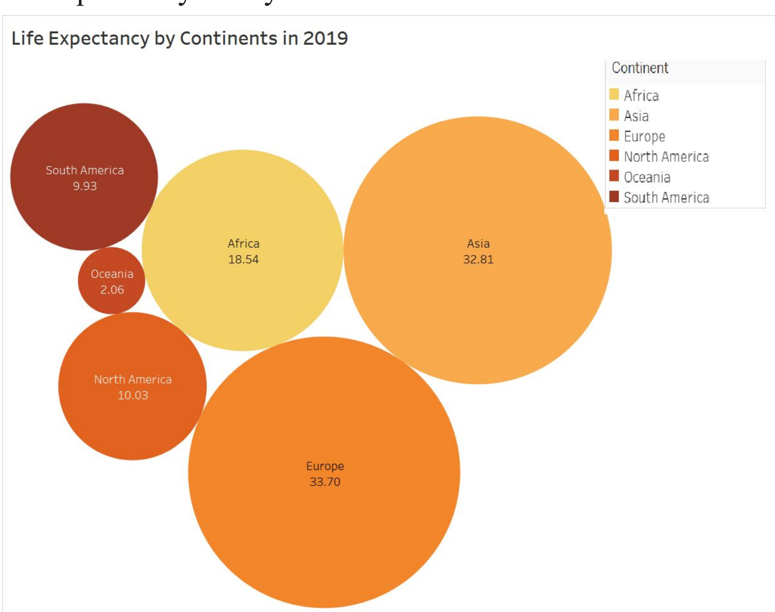
Individual Reports:

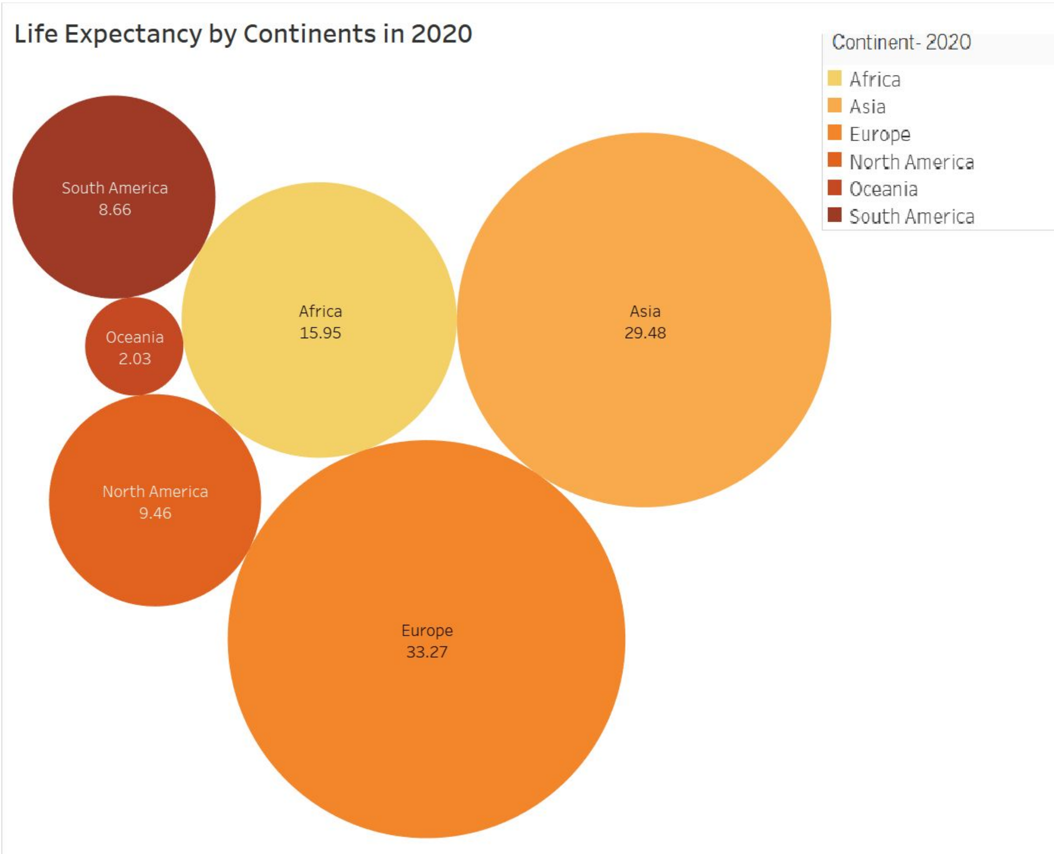
Sanjana Kulshrestha

My contribution:
After attending the group meeting, we all got different areas to cover, I got Life Expectancy.
Healthy life expectancy (HLE) is the expected number of remaining years of life spent in good health from a particular typically birth or **age** sixty-five, assuming current rates of mortality and morbidity.
So I decided to show Healthy life expectancy of all the Continents, then will be digging at a deeper level, that is, will be showing Country wise from 2019 to 2020.
In order to show the life expectancy during COVID and before COVID, I took data from both the years and compared the metrics and visuals together.

I used a packed bubbles graph to show the overall picture of life expectancy in both the years, the size of packed bubbles mixed with respected hue changing color from low to high helped in distinguishing different life expectancy values using Tableau. Mixing the packed bubble and color will help the audience to figure out which continent is having higher and lower life expectancy and easily compare the values for both the years.

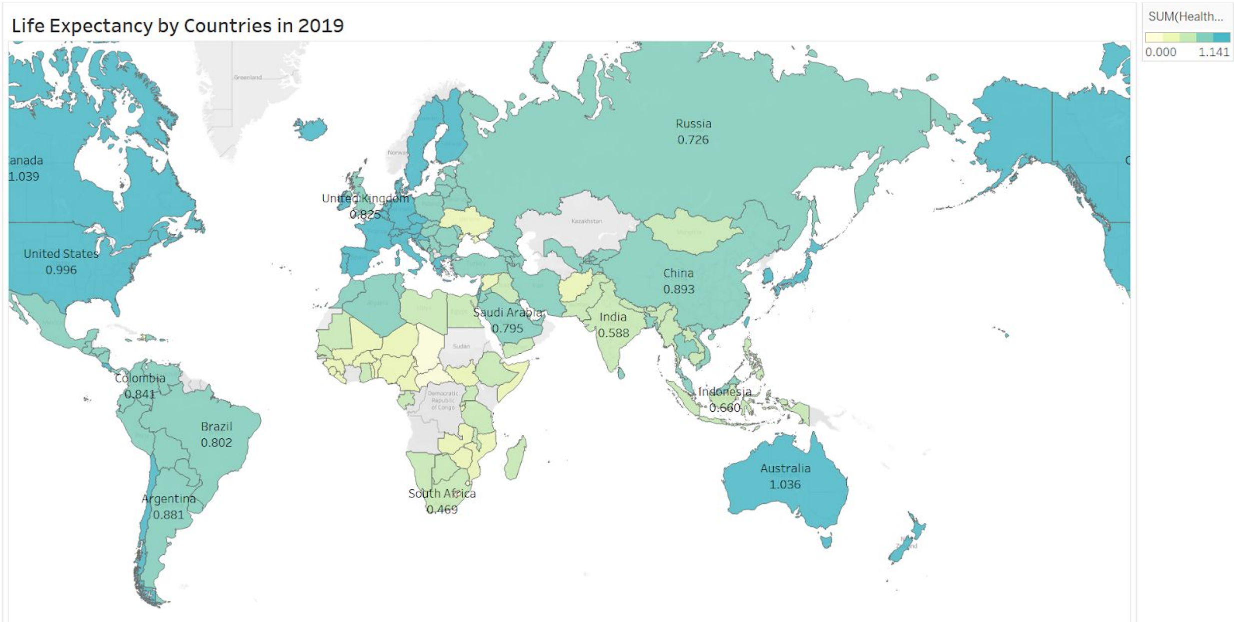
Exploratory Analysis and Visuals:

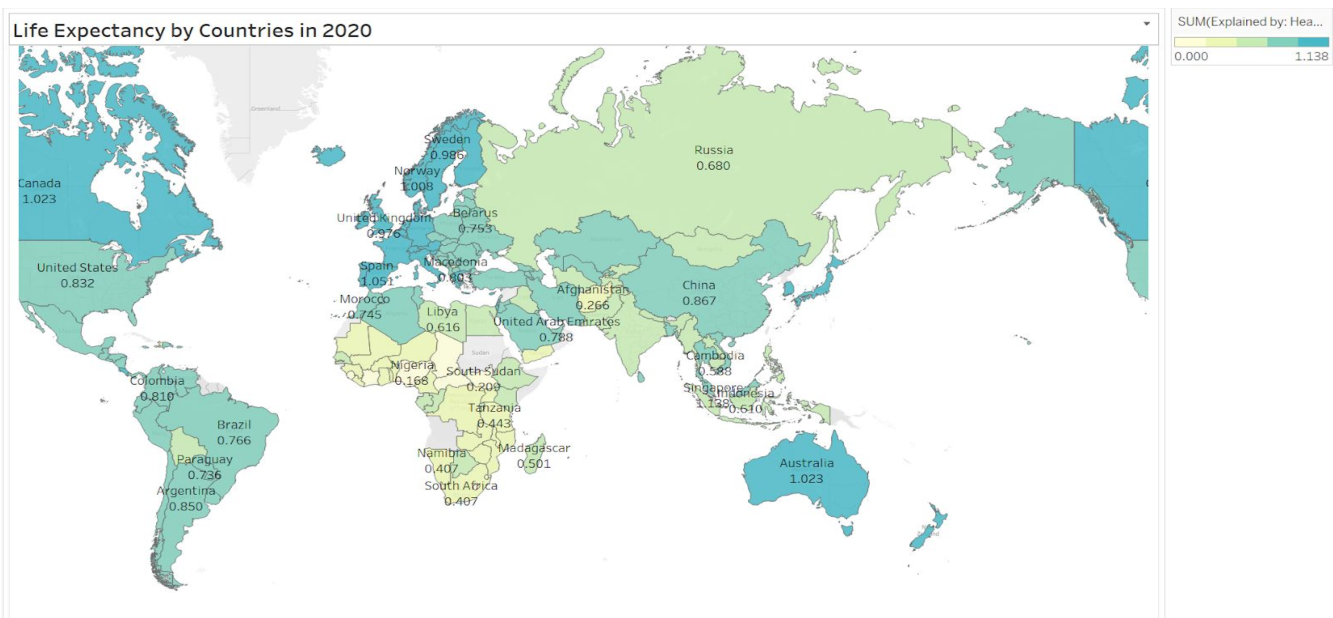




As we can assume from this map, we can assume there was a less significant impact of COVID in Healthy life expectancy on all continents except Asia. On the other side, Asia was a bit more impacted as we can see a significant decline in Healthy life expectancy from **32.81 to 29.48.** Through this visual, we cannot clearly indicate any correlation of Healthy life expectancy with other areas such as populations, fatality rate, but I can say that these continents were impacted with COVID slightly.

At the Country level:





I used choropleth map to highlight different parts of the countries and labeled a few countries in order to show the range, highest and lowest values of Healthy life expectancy value for the audience using Tableau. As per the map, the range of Healthy life expectancy from 0.000 to 1.141. Country having darker green color is representing a higher Healthy life expectancy and the lighter color is having a lower Healthy life expectancy value. In 2019, out of these sample country data, few top countries which were having higher health life expectancy like: United States, Canada, Australia, Sweden and the United Kingdom. In 2020, life expectancy for many countries got reduced, especially the United States got a major decline out of these countries.

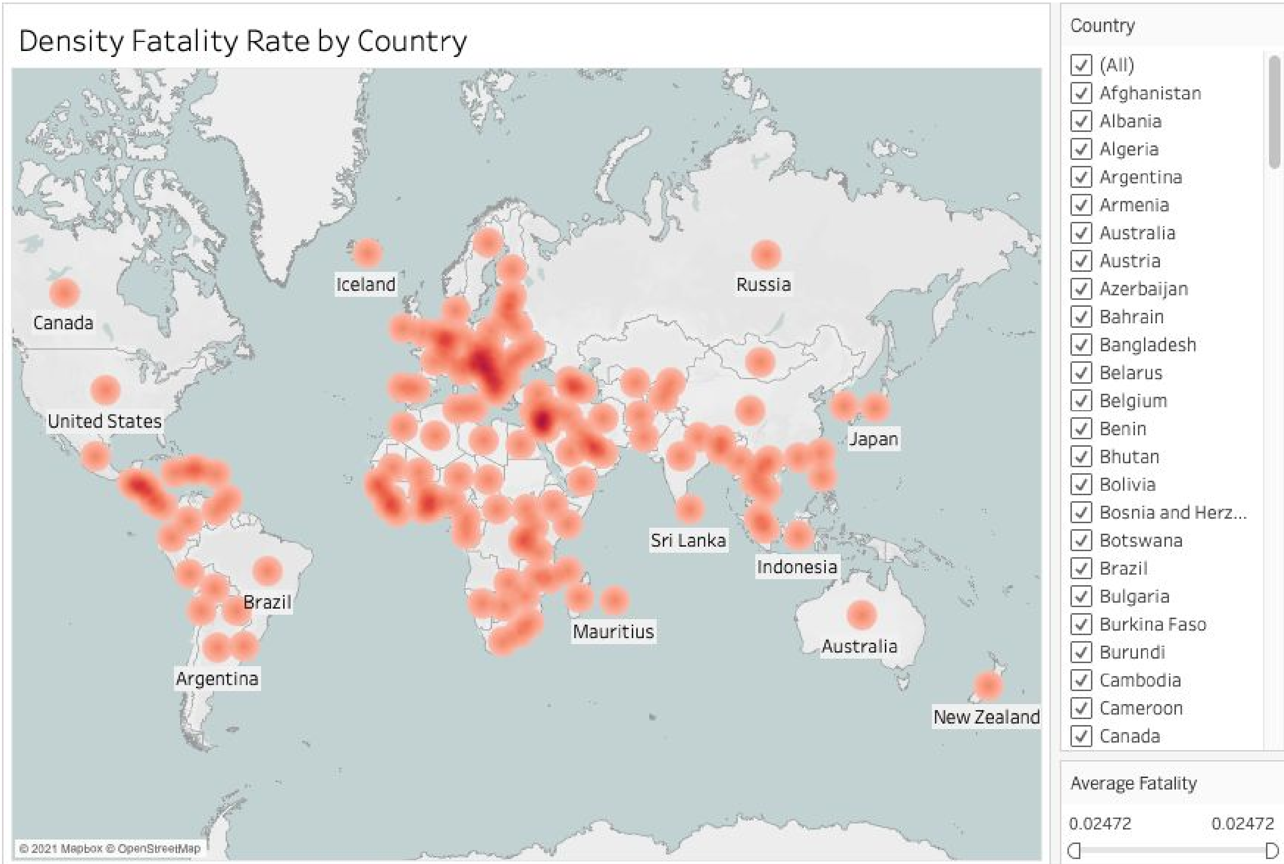
As per the map, **Canada** shows the highest Healthy life expectancy with the value of 1.039 in 2019, **which got reduced to 1.023.** One significant thing I noticed was that all the countries life expectancy got reduced in 2020, which makes me assume as per the numbers that COVID had a slight impact on most of the countries, but I cannot clearly indicate any correlation of Healthy life expectancy with other areas such as populations, fatality rate.

Reflection/ Summary:

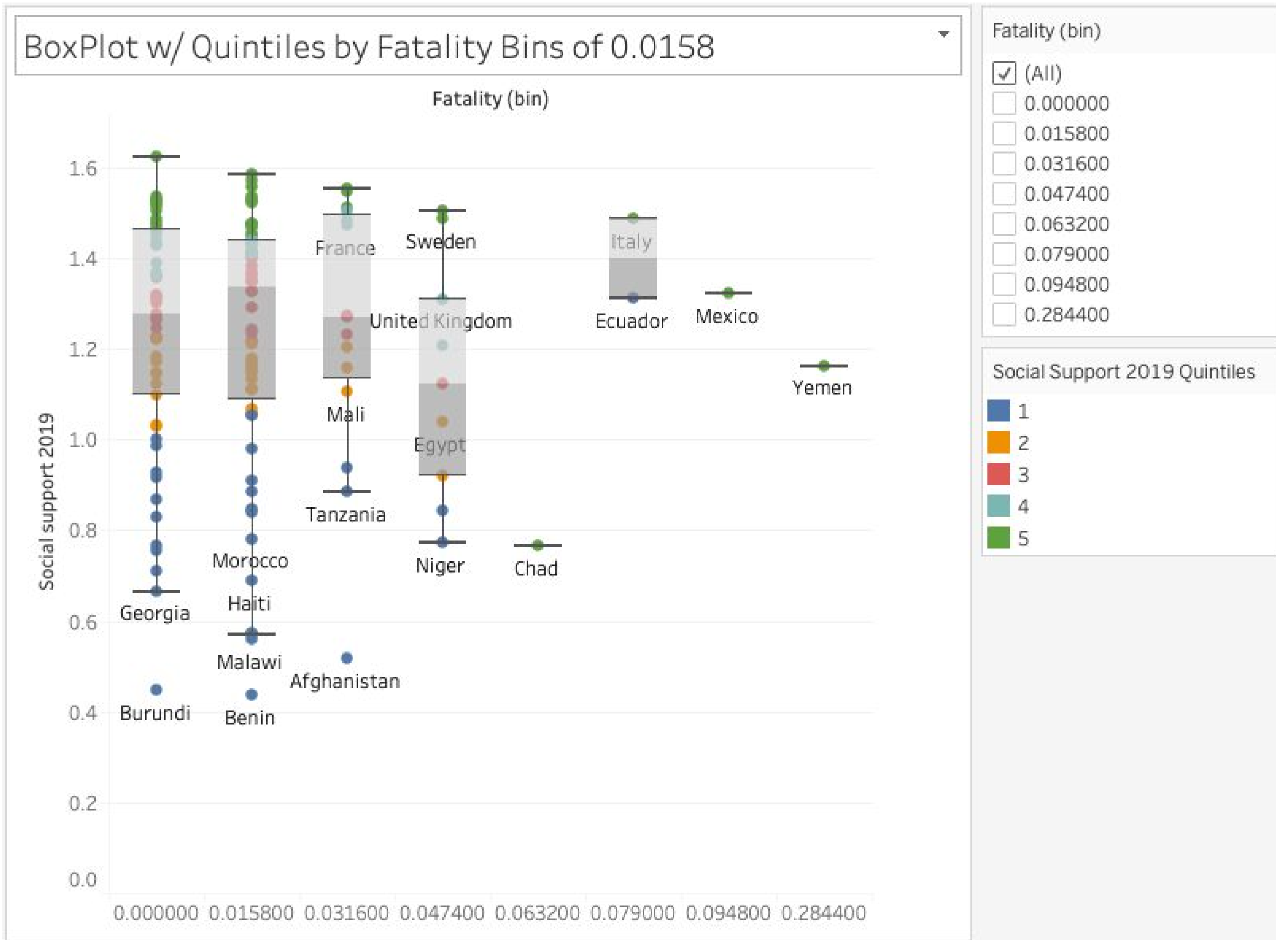
After working on the latest COVID dataset, I was able to correlate how one dataset can be connected and dependent on other datasets like fatality, generosity, COVID cases and many more.. With the help of visuals, I was able to see it properly as it was easy to understand through visuals as compared to if we just look through the numbers. With the help of this project and this course, I learned new types of graphs, which I always want to learn. I got this opportunity to work with group mates on these graphs using Tableau and R and was able to practically implement it as well. I faced few challenges while working in this project like using data of poor quality at the initial phase, then analysing what message we need to come up with and deciding what metrics to be included as per the audience perspective, instead of how I want to design any specific visuals or a dashboard.

Endri Rama

After discussing with the team, I was assigned to investigate the Density of the Fatality Rate per Country and see how this might correlate with Social Support. I created a geographical map using Tableau to investigate the Density of the Fatality Rates per country, this showed the team and I which other fields we might be able to correlate this field with and investigate the correlation, in order to discard any assumptions that might come up. The map showed the team and I how the countries were hit by Fatality Rate according to the data.



I then continued to see how the Fatality Rate might correlate with the Social Support 2019 field. Due to the Fatality Rate field being so broadly distributed, I proceeded with the method of creating bind in order to see which countries fell into which bin and see if Social Support had any correlation with this field. In order to do this I plotted the Fatality Rate bins in a box plot to see how each boxplot was distributed and then within those boxplots I added quintiles to see how Social Support was distributed. The findings were that there was zero correlation with Social Support and Fatality Rate Bins. This might not seem useful but in actuality it is as it further supports our conclusion that there was no correlation and that we cannot make assumptions correlating the two fields.

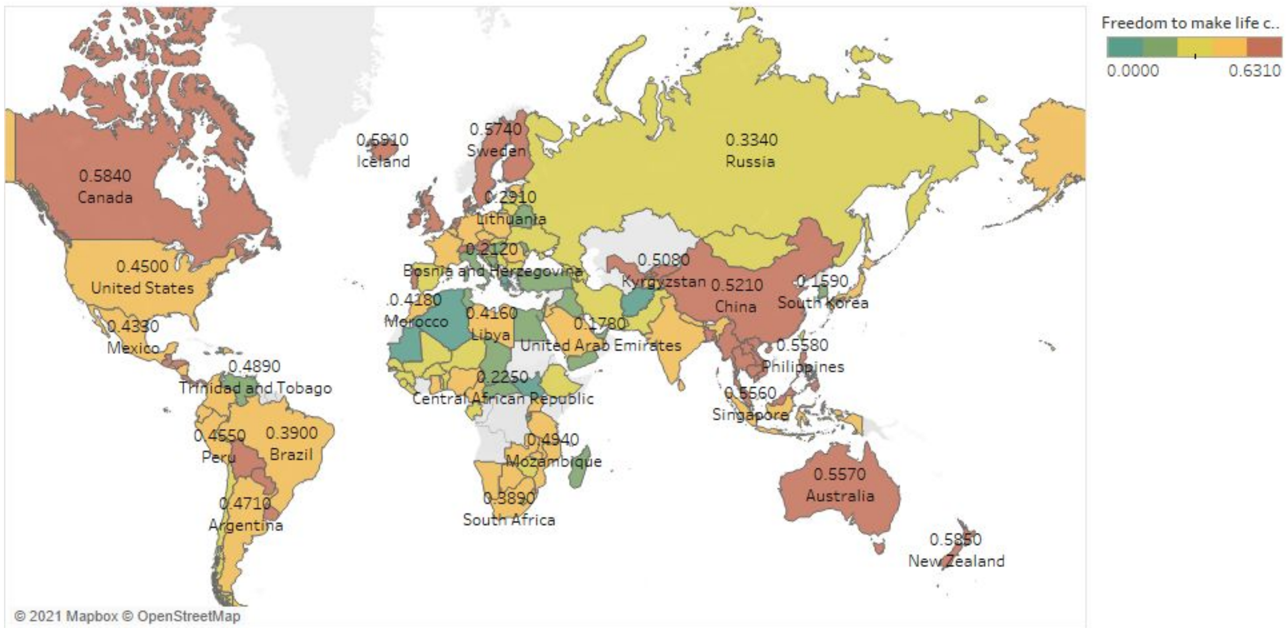


Reflection: The importance of data visualization helped us compare all of how the fields might correlate. Also learned that not all assumptions as to how a project should pan out does. Meaning we did our best to correlate the fields and find the best explanation as to how covid came to be, but as stated before we could not make any assumptions.

Aditya Lanjewar

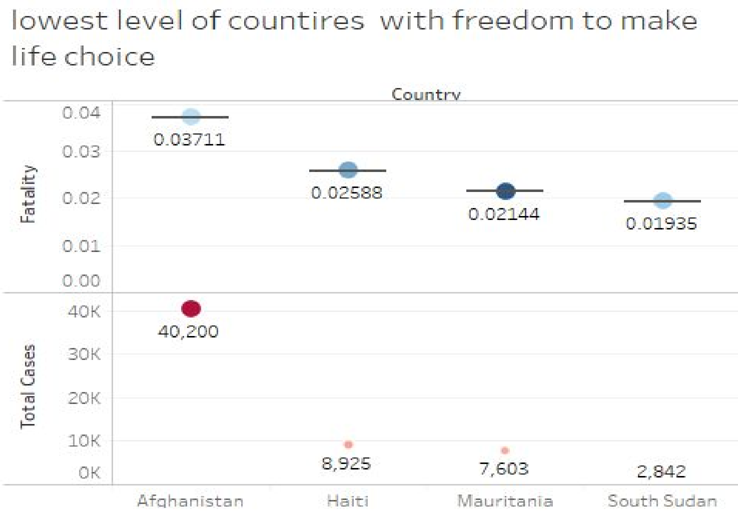
After Discussing with the team i was assigned to investigate **Freedom to make life choices** in 2019 with also using and merging the happiness dataset, I used geographical map to plot freedom to make life choices 2019 representing all countries with the saturation of color five step color variations ranging from 0.0000 to 0.6310

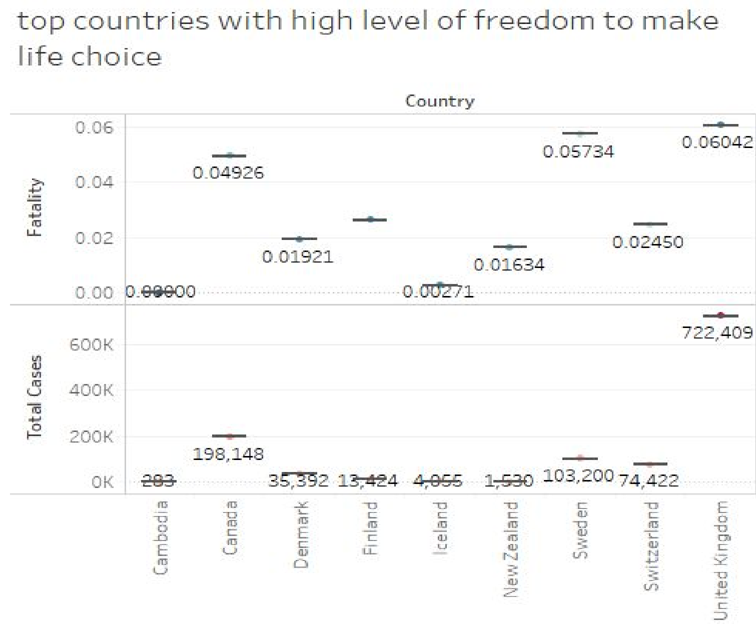
FREEDOM TO MAKE LIFE CHOICE 2019 vs COUNTRY Sheet 1



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Freedom to make life choices 2019. The marks are labeled by sum of Freedom to make life choices 2019 and Country. Details are shown for Country.

We were correlating the happiness data (freedom to make life choices) to covid 19 total cases and total deaths to find the correlation between both to find the pattern where, can we say that the countries with high level of freedom to make life choices did they affect with covid more or less as compared to the countries with the low level of freedom to make life choices.



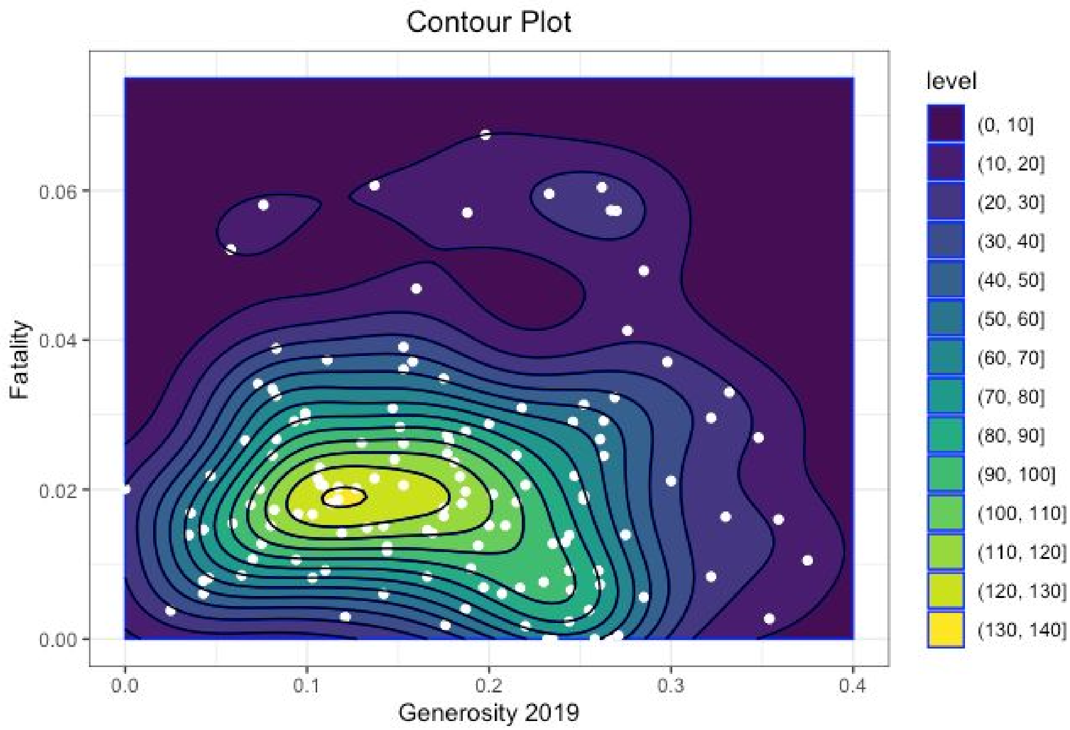


Lowest level and highest level of freedom to make life choices:
Graph representation:
Color blue: from light to dark represents (freedom to make life choices - lowest to highest)
Color Red: from light to dark represents (total cases - lowest to highest)
Size : Represents total number of cases.
As we are comparing the highest and the lowest level of freedom to make life choices, by using tableau for the same applying filter to make sure that we can get highest and the lowest countries with the freedom to make life choices so that we can correlate it with the countries which affected with covid and find out that how covid affected the countries with the highest and the lowest level of countries with the freedom to make life choices. In the graph the countries with lowest level of freedom to make life choices in the filter and adding the variables Fatality and total cases we can see the change in the graph

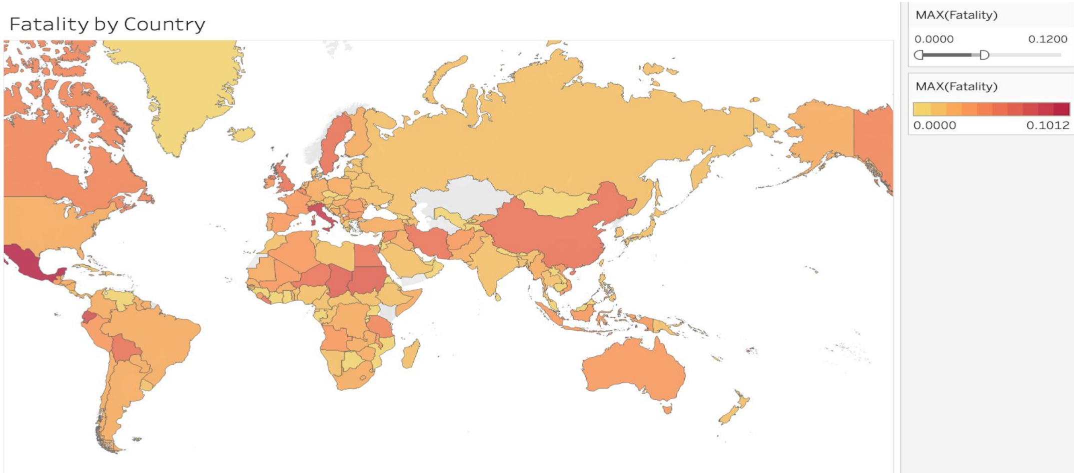
Reflection / Summary:
As going through this project by doing the visuals and playing with the data I encountered many visual aspects like reading the data and make sure that we fill up the null values in the dataset so that we won't get the false charts, as our project represents two data set covid 19 and happiness data we combined the both the dataset giving the whole new prospect see the world in terms of graphs. Just by using this data, I was able to see the covid19 impact over the world and to check the world gdp and the happiness data helped to gain more information about freedom to make life choices, fatality rate in each country, generosity of the country perception of the government and how much people trust the government in percentage, using particular visualizations we can exactly find-out how much number of people trust the government by comparing it with the population variable.

Hasmitha Chatla

I mainly focused on this project on Generosity 2019 and finding the relationship between fatality and generosity and also created the visualizations.



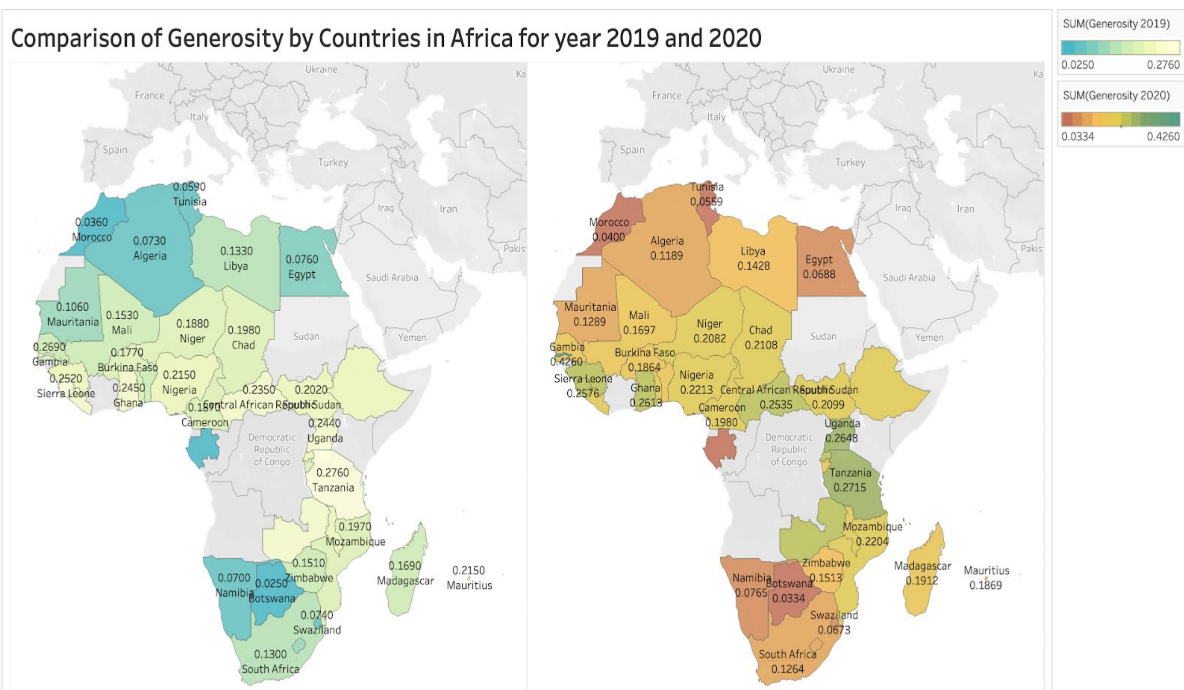
This is the correlation between Fatality and Generosity for the year 2019. From the above contour plot we can see that there are different levels of generosity. I was trying to find a pattern and compare how generosity and fatality had an impact on covid cases. For exploratory visualizations I have used geographical to find the fatality rate among the countries. There is no direct relation between geographic location and fatality and the fatality rate varies across areas.

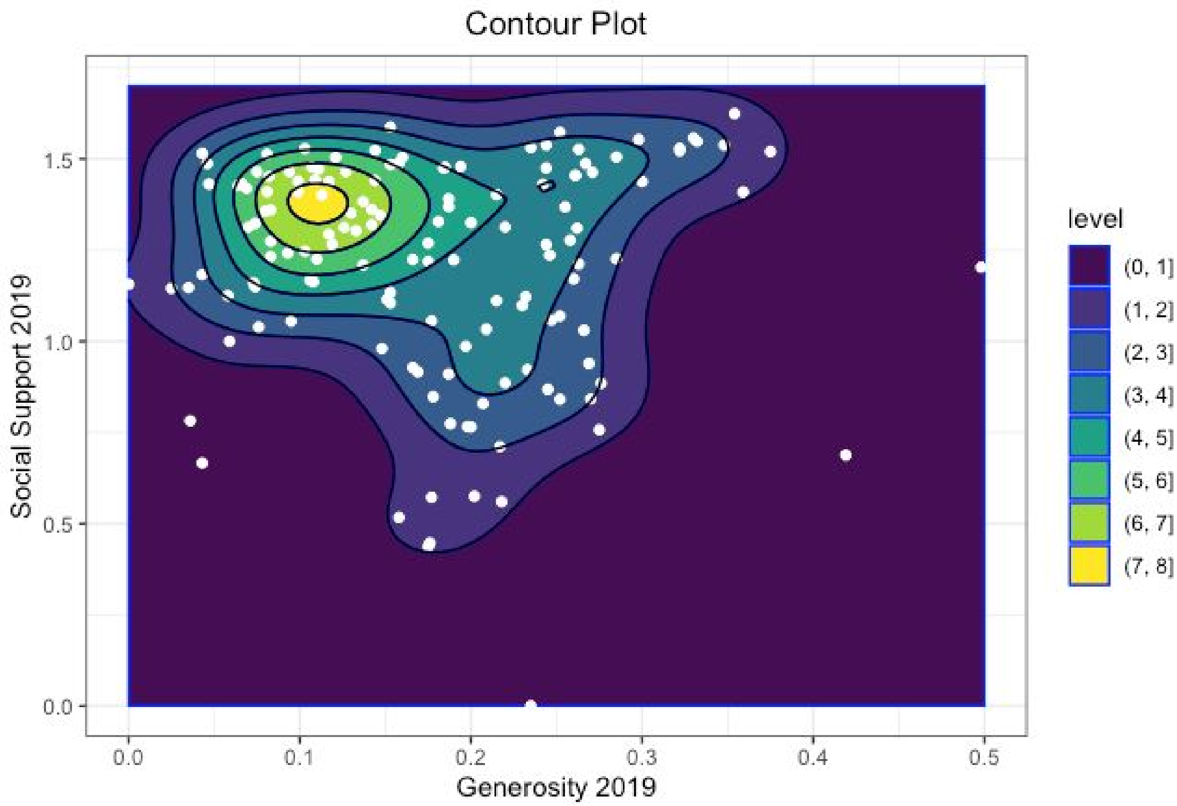


At country level

Africa

In the following map I have compared the generosity of the people by countries in Africa for the year 2019 and 2020. From the below maps we can see the range of the Generosity for both the years. The range for 2019 is from 0.0250 - 0.2760 and for the year 2020 it ranges from 0.0334 - 0.4260.





This is the correlation between Social Support and Generosity for the year 2019.The above Contour Plot shows that the countries with higher social support have a generosity around 0.1. From the below plot we can see the Generosity level 7,8 is near the Social Support 1.3 - 1.4. Through this plot we can say that the countries with higher Social Support need not have higher Generosity level.

I have learned from this project that visualization is an effective way of presenting abstract details, rather than a lengthy essay to illustrate. I have discovered a lot of techniques for applying visualization. I know that visualization is a process which requires continuous improvement in this project.

Elizabeth Kerrigan

In this project, I served as the group liaison and project coordinator. I corresponded with our professor regarding our project and submitted all assignments on behalf of the group. I created the Google drive folder where we collected all of our submissions and drafts for each step of the project. We also used this drive folder to store meeting agendas which I created for each meeting. These agendas were intended to give us a concept of what needed to be prepared before the meeting, what we would be discussing, and where I kept minutes and logged what was planned to be completed for after the meeting and our internal deadlines. I found the COVID-19 dataset and completed the initial visualizations to determine it was suitable for a project. As the project progressed, I was in charge of making sure that our project aligned with the stated information in the project milestones and project description. As the liaison, I would collect everything for submission and format it for the group by a certain deadline with group approval. I added the 2019 Happiness Report dataset to our compressed country dataset. I put together the video presentation and edited it in iMovie in order to create one cohesive presentation. I specifically created and voiced the introduction, background, and conclusion for the presentation in addition to my visualization.

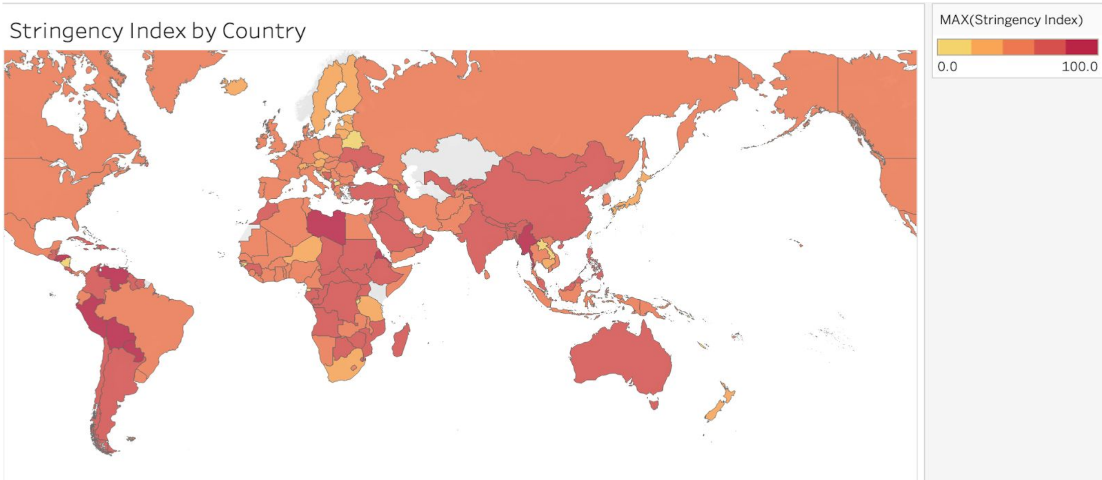
I created exploratory graphs for the COVID-19 dataset and the perceptions of corruption data from the Happiness dataset. I created the small multiples graph seen at the beginning of our visualizations portion. I also helped other group members try to find interesting visualizations with our different and combined datasets and helped make critiques based on the class concepts as well as building on feedback from our presentation.

I think the thing I learned most acutely, and obviously this is bound to happen, but sometimes even with a large dataset you're not going to have these gorgeous, obvious visualizations come. In most of my previous work, the visualization comes after so much other work has gone in, and I'm just trying to get R to work so I can submit and move along. In this class, the analysis and message was important but the core of the work was the visualization itself. This gave me such a wonderful space to really dig into the visualizations and remember that those are something that also need to be worked and iterated on as well. I loved getting more comfortable with R (it's still a little rocky) and was disappointed when I realized I needed dual axes to complete my visualizations but was unable to accomplish that in R. I would have liked to have more control over the visualizations that we get in R. For example, I could not get rid of the "[2020]" on my x axes no matter what I did and tried.

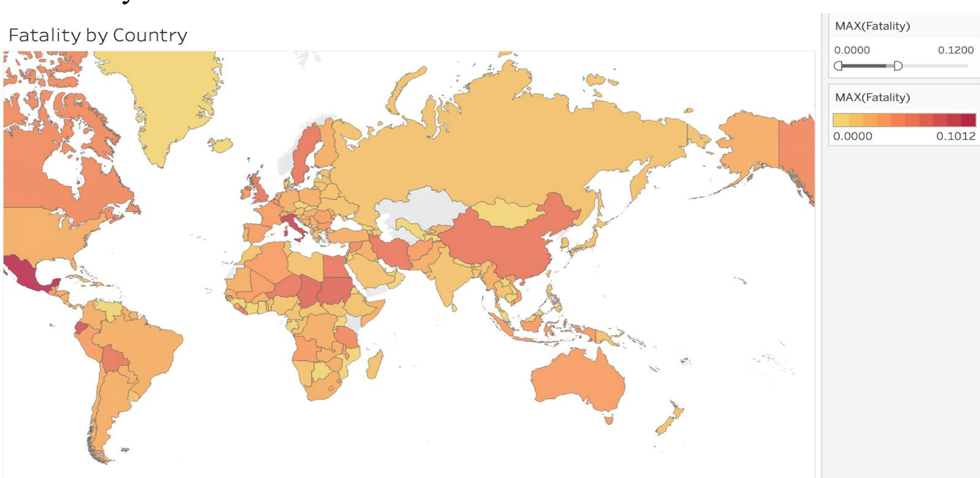
Other Exploratory Analysis

COVID Data

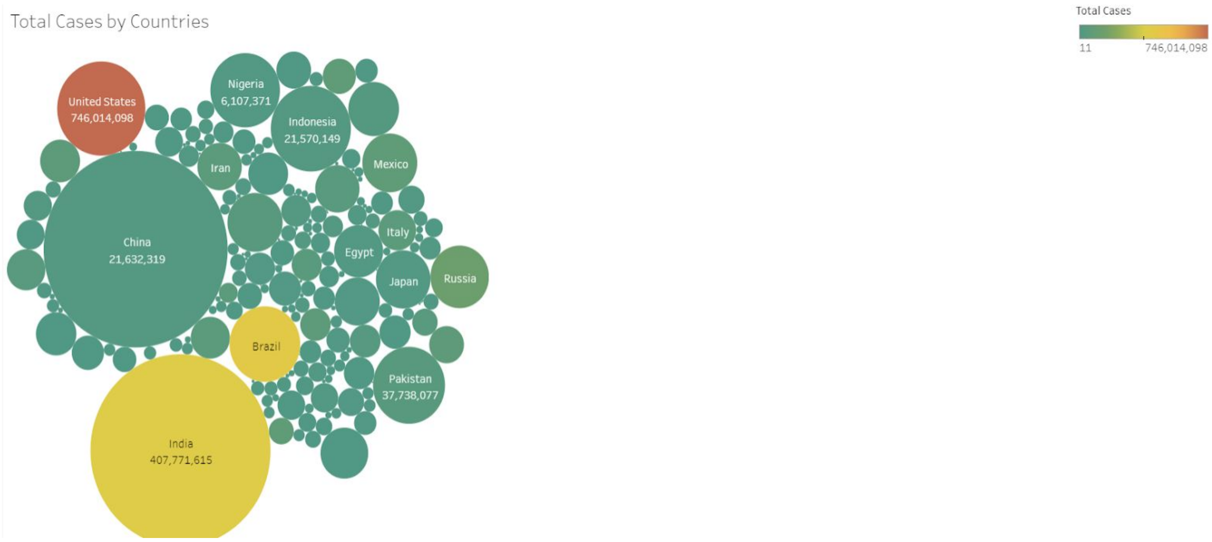
The stringency index is a measure based on 9 government response indicators such as work closures, school closures, and travel bans on a scale from 0 to 100, with 100 meaning the strictest regulations. In the visualization below, we can see that the majority of countries took a moderate approach to responding to COVID-19 while fewer had a more severe response and even fewer a relaxed approach. This visualization was based on a country's average stringency index so analyzing across time will be important to really understand each country's response.



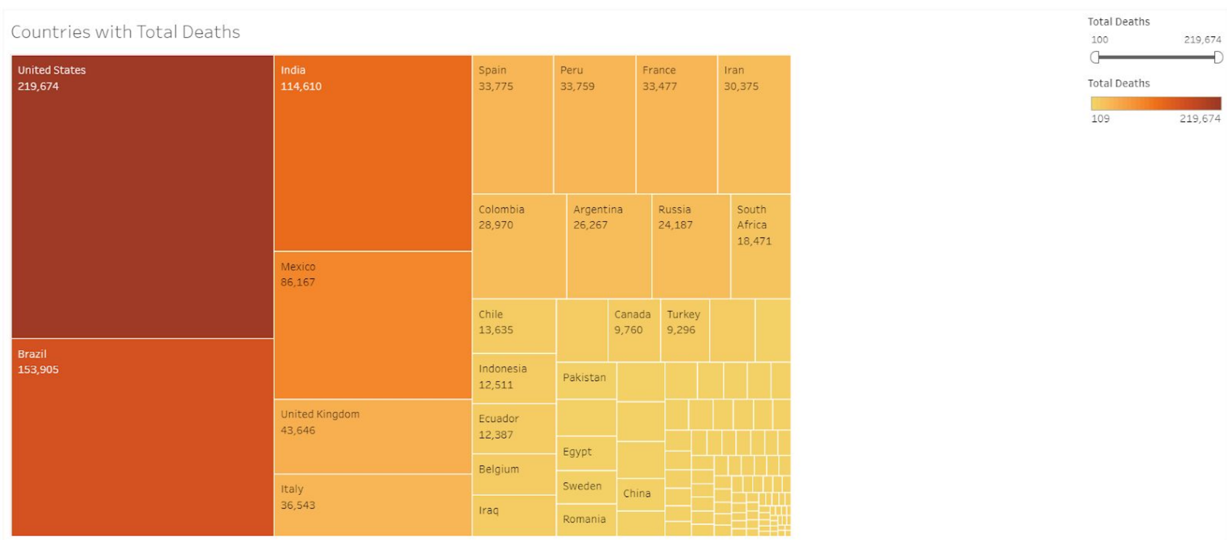
Below we look at the fatality by country. The maximum fatality rate for this visualization is 10%, meaning of every 10 cases reported in that country 1 person died. The fatality rate varies within regions and there does not immediately appear to be a strong correlation between geographic location and fatality.



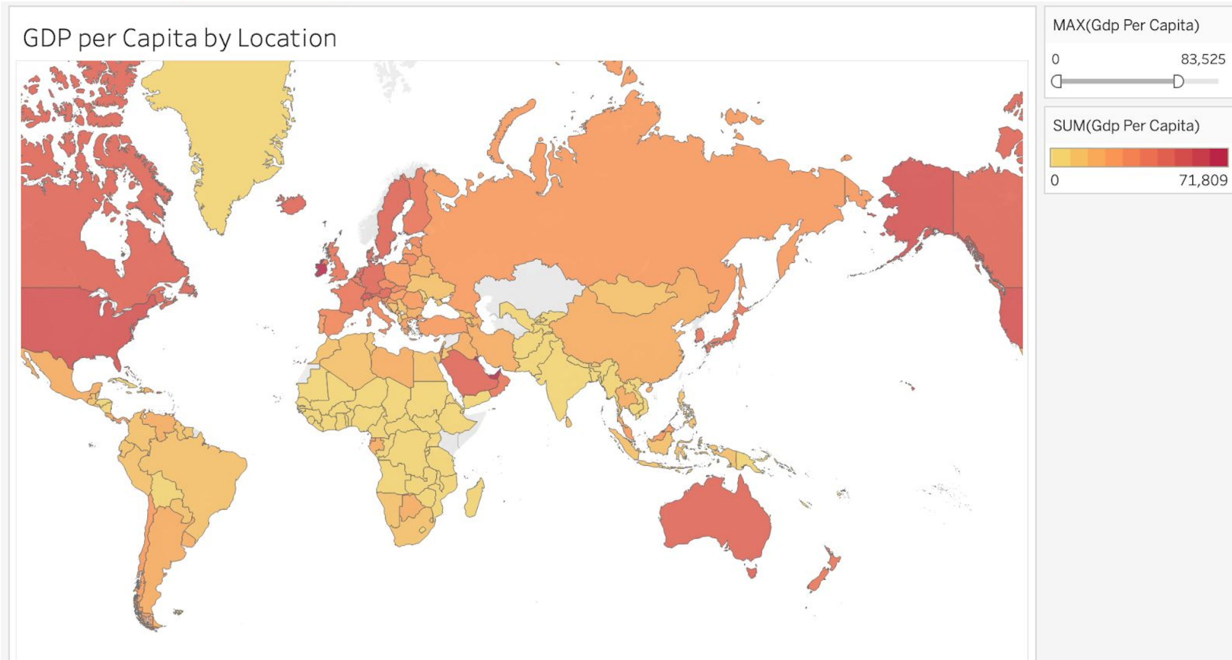
In Total Cases by Countries, we can see total cases of each country based on color with yellow and red indicating a higher number of cases and the size of the bubble reflecting the size of the population.



This visualization, Countries with Total Deaths, helped us really understand the difference in magnitude between the countries most affected. The size and color help illustrate the gravity of the number of deaths in those countries.



In the GDP per Capita visualization below, we can see that most of the wealth is concentrated in North America and Europe with rarer pockets of wealthy nations throughout the world. For example, South America and Africa are much poorer regions by comparison. Please note that the GDP per capita measure used in this dataset is an average



World Happiness Report Data

I wanted to investigate whether certain regions had stark differences in how they felt about their governments. Most regions have a similar mean and spread, concentrating around 0.1. Australia and New Zealand are particularly high, but there are only two countries in that region. Northern Europe has the greatest difference.