Group: Team LS2
Title: Implications of Diets and Measures on COVID-19
Members: Maharshi Thakkar, Nick Edington, Jeet Raol, and Bharathwaj Reddy Yanala

Introduction

With the current COVID-19 pandemic having a huge impact on everyone's life, it is important to look at related data in as many ways as possible. We will not only be observing general trends of the virus, such as rates of infection and deaths in different places, but also how diets in different countries may affect these rates. It is a well-known fact that a person's diet greatly affects their health, and we want to see if the diet of the general population of a country can have different results on how much they are affected by COVID. Another thing that we will be examining is the effect of measures taken by governments to slow down the spread of COVID. These measures include social distancing, movement restrictions, public health measures, social and economic measures, and lockdowns.

Data Overview

Case and Death Data

This data set is compiled by the Johns Hopkins University Center for Systems Science and Engineering found here:
https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases
It is time series data on the aggregate number of confirmed cases of covid-19 starting January 22nd 2020 and going to the present. It includes numeric data for the count of confirmed cases for each day, by region, as well as categorical data for country/region and province/state (if applicable). It also includes geographic data corresponding to the aforementioned categorical data. We will also be using another data set from the same source, containing similarly presented data on deaths caused by the virus.

Diet Data

This dataset on Kaggle is put together by Maria Ren and the data for different food group supply quantities, nutrition values, obesity, and undernourished percentages are obtained from the Food and Agriculture Organization of the United Nations.
https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset/data?select=Food_Supply_Quantity_kg_Data.csv

Each row represents a country's food intake in kg as a percentage, and the columns contain the country name, the percentage intake for the different food types, and the percentages of the population that is obese and undernourished.

Measure Data

This data set is taken from ACAPS. It covers government measures taken to mitigate the impact of the pandemic found here:
https://www.acaps.org/covid19-government-measures-dataset
It contains categorical data for country, region, admin level describing where the measure was implemented, date information of when the measures were implemented, and 4 categorical variables describing the measure. Additionally it provides date and 4 more categorical variables describing the source of the information and when it was entered into the data set and an ordinal variable for the id of the measure.

Population Data

This data set comes from world meter which derived their data from the United Nations, Department of Economic and Social Affairs. This was then transcribed into CSV format by a user on Kaggle.com. The data shows statistics for total population, population density and other growth related variables for various countries. The data was found here: https://www.kaggle.com/tanuprabhu/population-by-country-2020/data
It contains a categorical variable as well as ten numeric variables for the following features: Population in 2020, yearly change as a percentage, net change, density as population per square kilometer, land area by square kilometer, net migrants, fertility rate, median age, urban population percentage, and world share of population in percentage.

Exploratory Analysis

Distributions of government measures

The focus while exploring the measure data set was to investigate transformations that would lend to joining the data with the death and/or confirmed cases data sets. I first looked for the distributions of the measures as shown in appendix A figure 1. With that information I decided on a few factors to clean up the data: remove empty/NA values, ignore phase outs, reduce implementations of measures to the first instance per country. I recreated my distribution graph in appendix A figure 2 and colored by category. Additionally, I created a tree map, grouped by

category in appendix A figure 3 and a box plot for implementations of the measures over time in appendix A figure 4. I attempted to find a methodology for reducing the measures to simplify the graphs, however, I found such reduction difficult without losing information I felt my audience would be interested in. This cleaning did however contribute to the joining of the data with the death data. Lastly, I examined the distributions of country population densities to be able to create population density classes as shown in appendix a figure 5.

Diet and its Implications on Covid

As the world scurries to find the cure for the novel coronavirus, we as scholars attempt to find who the coronavirus really attacks. It's no surprise that the virus will take down many people as it's been upgraded to a pandemic. In this time, it is a known factor that like many viruses the ones that are the most impacted are the elderly and those with pre-existing conditions. However, we do not want to end the investigation on that front. While those that are the most susceptible are those individuals, we dig deep into the data to see what other demographics this disease perhaps attacks. The idea of this research is to understand if diet has in any way some sort of an implication on people's interactions with covid. Whether that be in catching the virus, recovering, or perhaps passing away due to it. Figure 6 appendix A has a corrplot which shows positive density towards obesity and it's interaction with Covid. This is where the focus started to shift towards understanding this correlation.

In the search for this, we have enough data to also try to understand patterns of virus with different countries and their population. While this is not the focus for our research, it's supplementary to understanding geographically if certain countries fared better than others. In figure 7 & 8 a geographic map is created that also takes into consideration population. It would only make sense that these two figures would counter one another. For the most part, either the person survived, or they passed away from the virus. However, it is still interesting to see how the virus wreaked havoc geographically. This also gives us an indication in terms of which countries really struggled with this. This is done by looking at the size of the text. This is important as it gives another dimension in learning more about the virus and what way it can be handled. We have four data sets for the diet. They focus around fat & protein consumption, and food supply in quantity and kilocalories. All four datasets behave in a very similar fashion. Therefore we will focus on the fat consumption primarily just for better understanding.

Explanatory Analysis

Hypothesis: The time between the first Covid-19 related death in a country, and the initial implementation of various government measures will show a relation to the number of deaths those countries experience as a result of Covid-19.

The intended audience for this visualization are those interested in taking a more detailed and nuanced look at how various government measures affect death rates. The intention of using the included data, is to provide as responsible as possible context to these evaluations while trying to not overwhelm the audience.

To prepare the death data set I reduced it to be organized by country so it could be joined easily with the other data. I identified that some of the death counts were split by provinces of a country. I chose to ignore the province data but this required me to calculate totals for Canada and China as they did not have nationwide totals already available. After this I joined the death data with the population data based on country.

At this point I pivoted the data such that each day's death count had its own row for each country. This resulted in each row containing the death total for a given day of a given county along with the population and geographic for that country. This first allowed me to calculate the date of the first death for a country by letting me find the first row grouped by country that had a non-zero death count. Additionally, I was able to create a column for the average deaths per data by population since the first death in a country. This was done by taking the current death totals for each country and dividing that by population and then the number of days since the country's first death. Lastly I added a column to show to group countries with similar population densities. Once these summarizations were made I reduced the data to only have one row per country.

As described in the exploratory analysis for the measure data set, the measure data had been reduced to specify the first implementation of each measure for each country available. This allowed me to join the population and death data with the measures data. I was then able to calculate lag by comparing the first death of each country to the date of the implementation of each measure.

With the data frame created from these transformations, I exported it to tableau and began creating an interactive visualization. The main visualization I used mapped lag against deaths per day for each country on a scatter plot. Color was used to specify the measure and point size was used to encode population density class. I finally added a trendline and allowed it to be filtered by population density class and measure.

To supplement this data I added a few extra charts to my dashboard. I decided to retain my distribution graphs for measures and population density classes as they could act as self evident filters as well as give appropriate context to the provided data.

Lastly I added a geographic plot to give my audience situational awareness of where the data was distributed on the globe.

The resulting visualization provides the ability to examine many facets of the data. Trends for individual measures can be viewed as shown in appendix B figure 1. More detail can be gained by isolating those measures for individual population density classes as in appendix B figure 2. Comparisons can be made between measures as shown in appendix B figure 3. Lastly detailed data about the countries' individual responses to covid via selecting the country in the geographical plot and/or highlighting the data points in the main graph as shown in appendix B figure 4.

Hypothesis: The overall health and diet of a country may have an impact on the recovery rates of COVID-19 patients. Different percentages of food intake will affect the recovery rate based on the type of food.

Before looking at the diets, I wanted to check the impact of the health of a country on the recovery and death rates of patients. To examine this, I created multiple scatterplots, plotting the percent of COVID-19 patients who recovered or died with the percentages of the country's population that is obese or undernourished as displayed in Appendix B Figure 5. It was important to show the percentages instead of the actual number because the populations of countries vary greatly so the total numbers may be misleading. On top of these plots, I added regression lines to visualize possible correlations.

Analyzing the plots and regression lines does show one trend, and that is between the COVID-19 recovery percentage compared to the percentage of the population that is undernourished. It seems like as the undernourished population percentage goes up, the recovery percentage goes down. This could make sense because countries with more people that do not consume what they may need to function properly will not be able to fight off a virus as easily as others. There is not too much of a correlation seen in the other plots.

Now moving onto the actual diets and food groups, I plotted the percentage intake in kg for each food group as seen in Appendix B Figure 6. I added jitter to separate the points and used a diverging color scheme for the points to separate what I considered were low and high recovery percentages based on the data, instead of splitting at 50%. I used green to represent the high end of the recovery percentages and red for the low end.

Upon initially examining the visualization, there is just a clutter of red and green points with no type of distinction that can be made. But looking closely, specifically at the "Cereals - Excluding Beer" and "Starchy Roots" categories, it is possible to see that

countries who tend to consume more of those types of foods, have had a lower percentage of people recover from COVID. Then looking at the “Vegetable” column, there seem to be a little more green points seen towards the high end of the intake area. There are red points as well, but they are not very saturated, indicating a just below normal recovery percentage.

After first making some rough geographic graphs, and performing a correlation plot, the story began to unfold on its own. The geographic graphs gave us enough insight to understand what countries were intriguing to look at. From there, a heat map in figure 7 was created along with a bubble chart in figure 9. The heat map was created with the exclusion of any null data points as this would not be helpful in our case. Each block of course incorporates the measure of obesity of each country. Because of its organization, it allows us to understand death rates better by the different countries. We can see which countries really struggled against Covid, and see how much population they have. From the heat map and the bubble chart we see that while Brazil struggled with death rates much like the United States, they had better recovery rates statistically compared to the US. The United States, in fact, really struggled with it’s recovery rates along with Russia when adjusting for population. In the middle of the bubble chart we see smaller countries, mainly from Europe that struggled with Covid, and some of their death rates can be seen on the heat map. The heat map also gives an interesting perspective to look at, to see if the percentage of obesity is indicating a country will struggle with death rates. While it’s not a huge display, there does seem to be a higher death rate amongst the United States, the UK, and Brazil, all of which have higher rates of obesity.

In figure 10, we take it one step further and put the data for death and recovery rates on the same bar plot graph. We also include in this case however obesity and population. Similar to the other graphs, this one is filtered as well to leave out any null data. This graph is amazing because it has the capability to show all the different factors in one place. From this we are able to see that while Oman and Kuwait have similar populations according to the width of the each bar plot, and have similar death rates, the recovery rate is not the same. When we factor in obesity we see in fact that the Oman has a higher recovery rate and also lower percentage of population with obesity. While there are not enough countries for us to compare this type of result, this is a start.

To further study this, a contour plot in figure 8 we created to understand death rates and obesity better. We see in the bottom left corner that the lower the obesity the lower the death rates. In fact all the points on the plot are all clustered together. However, as obesity goes up, so does the death rate and randomness of the scatter in

the plot. This tells us that there is much more left to the story and the contour plot is on the right track and perhaps so are the rest of the graphs. We are able to piece different graphs together to tell one unifying story and that is always the approach when studying data science. We cannot for certain say there is statistically significant information to prove this, however, graphically we have some evidence to claim there may be a relationship between obesity and the death rates of covid patients.

Exploratory data analysis on Covid19-Dataset

Firstly, we have pivoted each date column from 01-01-2020 to 08-08-2020 into one column which is Date and removed dates columns full of null values from the dataset. I have used tableau to visualize the geographical plot showing covid-19 cases of each country with latitude generated in rows and longitude in columns and Date in Page with covid-19 cases in size and detail by country. Plotted line graph using tableau showing total number of covid19 cases in rows with date in columns to compare country wise covid-19 cases I have plotted bar graph in tableau with country and region in and covid-19 cases in columns which shows that highest number of covid-19 cases are use with around 50 million cases followed by brazil with around 30 million cases and India with 20 million cases approximately. I have also plotted similar plots for covid19 deaths by plotting bar graphs to compare the number of deaths caused due to covid-19 by each country wise.

Conclusion

Studying the actions taken by individuals and governments of various nations, is helpful to understand what practices mitigate the spread of a global epidemic such as Covid-19. From the plots of geographical plot and bar graphs on Covid-19 confirmed cases show that the highest number of cases are in the USA with around 5 million followed by Brazil with around 3 million cases and India with around 2 million cases. Alternatively, from the plots of geographical and bar graphs on Covid-19 deaths shows that the highest number of deaths are in the USA with around 0.16 million followed by Brazil and the UK.

While strictly studying diet, it is hard to come to precise conclusions that the overall diet of a country may affect their COVID recovery rates given all of the factors that contribute to it. However, with the data that we had, we found there is a slight decrease in the recovery rate for countries who consume the most cereals (excluding beer) and starchy roots in their diets.

Diving deeper into diet, the most common trend we see, is that there is some correlation between obesity and Covid. While it may not be a powerful one, there is enough evidence to suggest there is some correlation. More research into this would be beneficial and beyond the scope of this course. We have also created and studied many different graphs that help us in pointing to this evidence. Some geographical plots were created as well to see if there are any other patterns in the graphs that we could see. It was not very clear, however we do see certain countries struggle throughout all the data diving. Unfortunately, the United States, and Brazil seem to be the biggest culprits of not dealing with Covid comparatively.

On the other hand, there is a lot of insight to be gained seeing how each government responded to the virus and how quickly. While plotting deaths against lag summarized the data, it is unclear how efficient of a method it is for evaluating effectiveness as some of the sample sizes were rather small. With the visualizations provided here we have just scratched the surface but hopefully recognized some practices that are helpful.

Appendix A: Exploratory Data Analysis

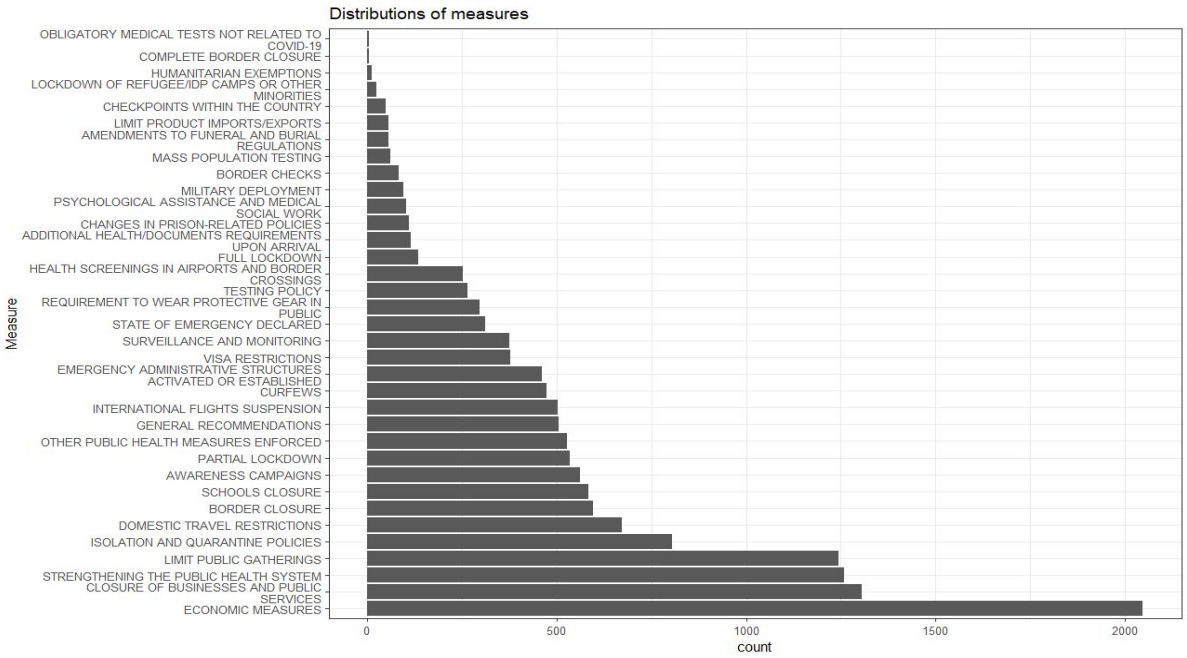


Figure 1: Distributions of Covid-19 mitigation measures

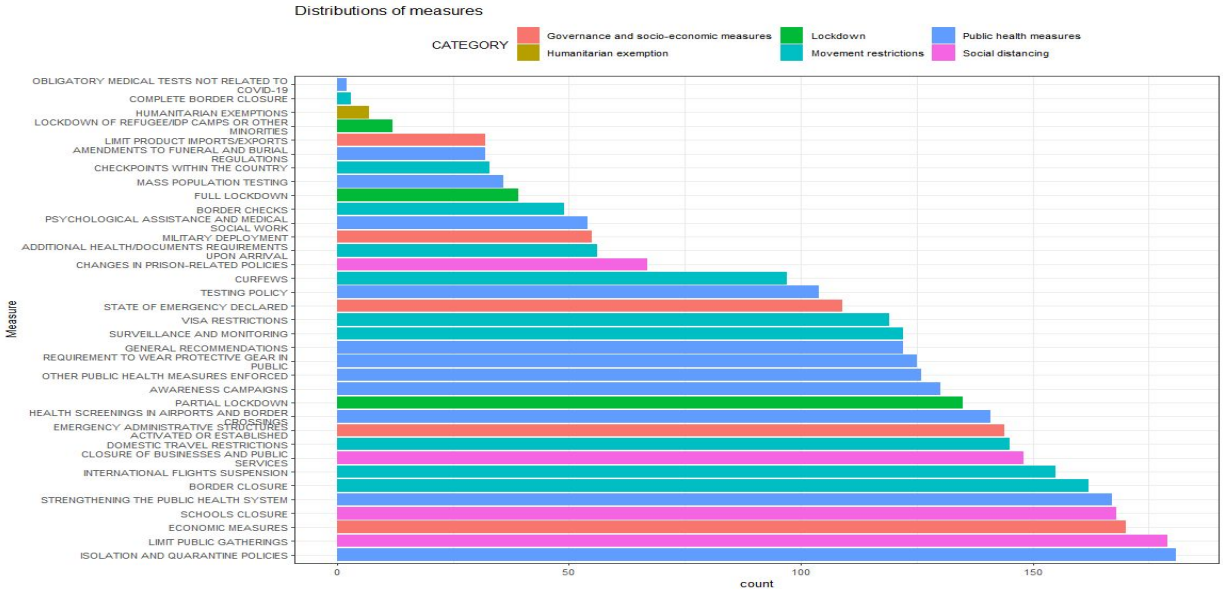


Figure 2: Cleaned distributions of measures by category.

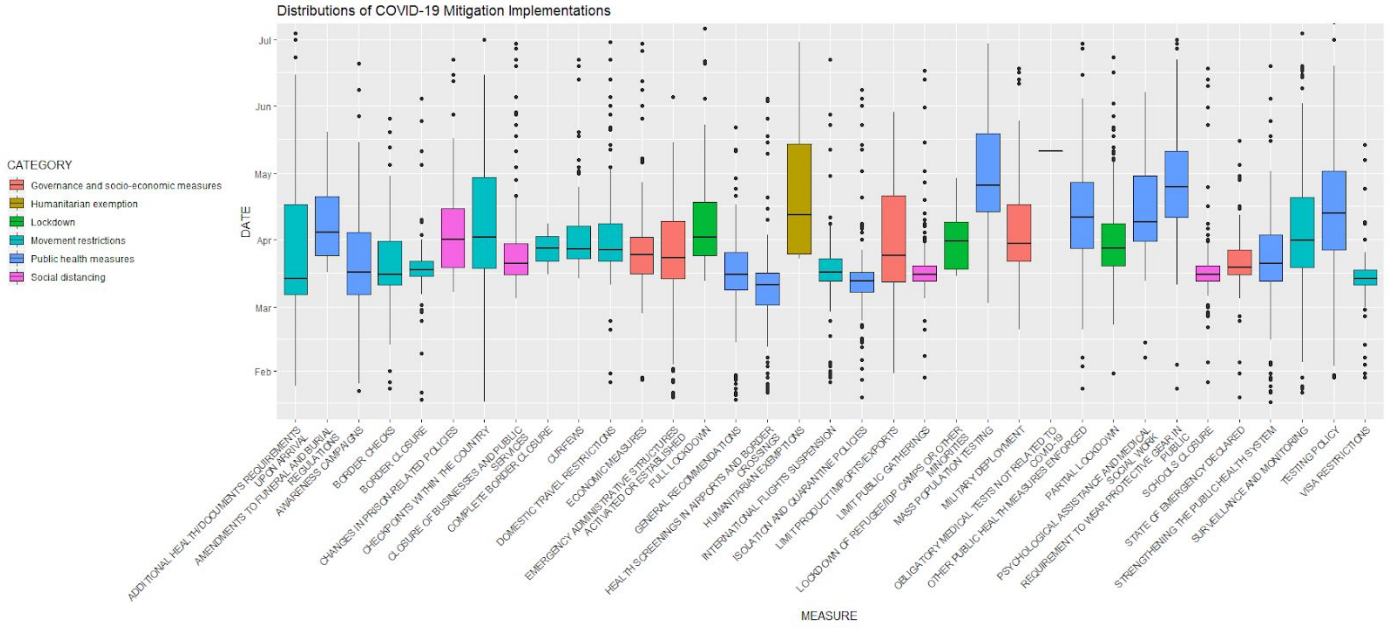


Figure 3: Distributions of measures over time.

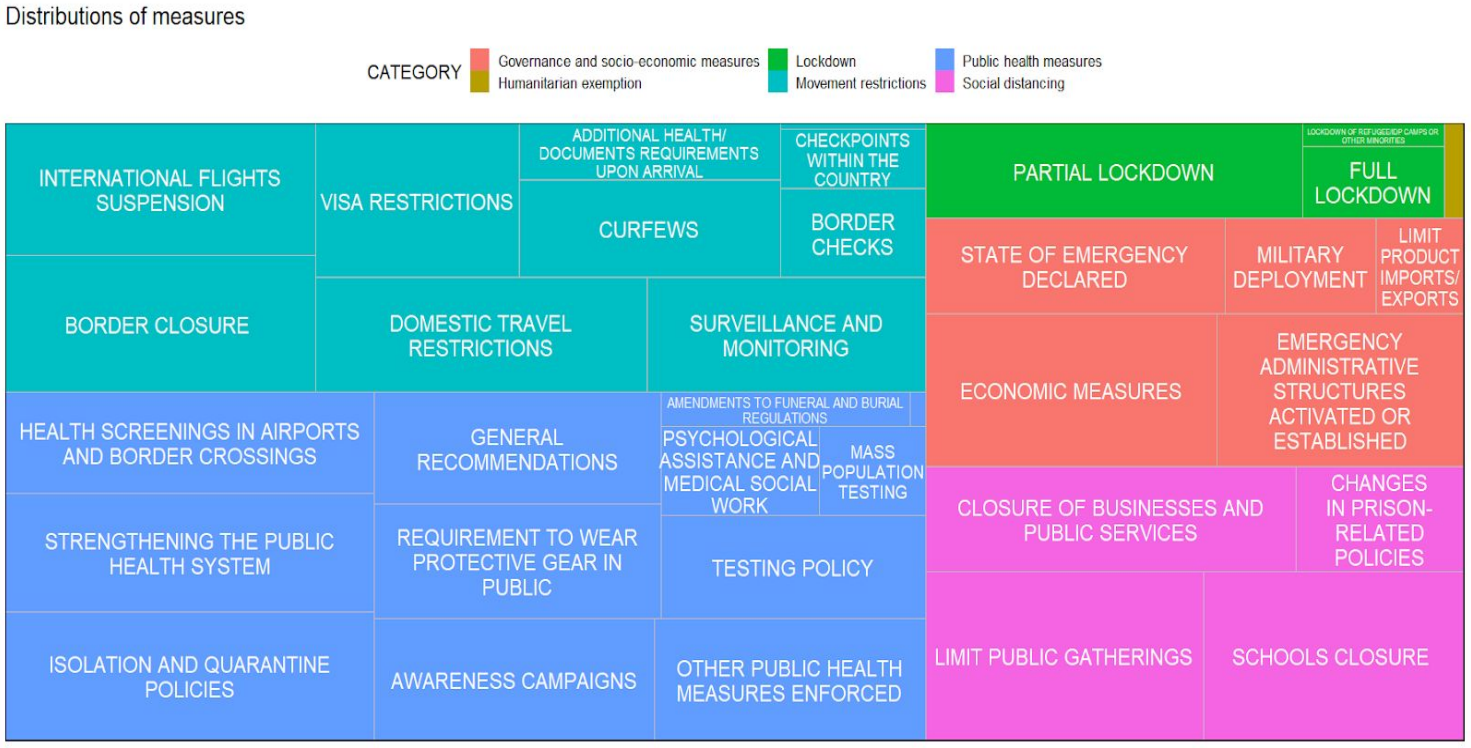


Figure 4: Treemap of measures by category.

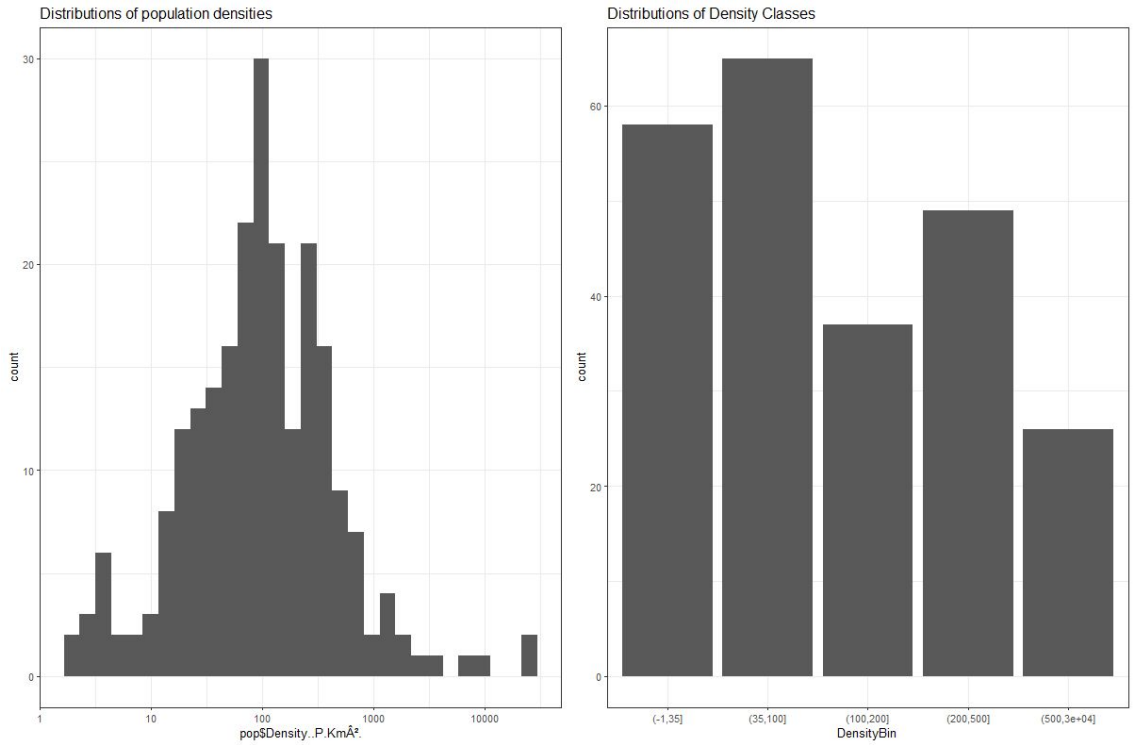


Figure 5: Density distributions and density class distributions.

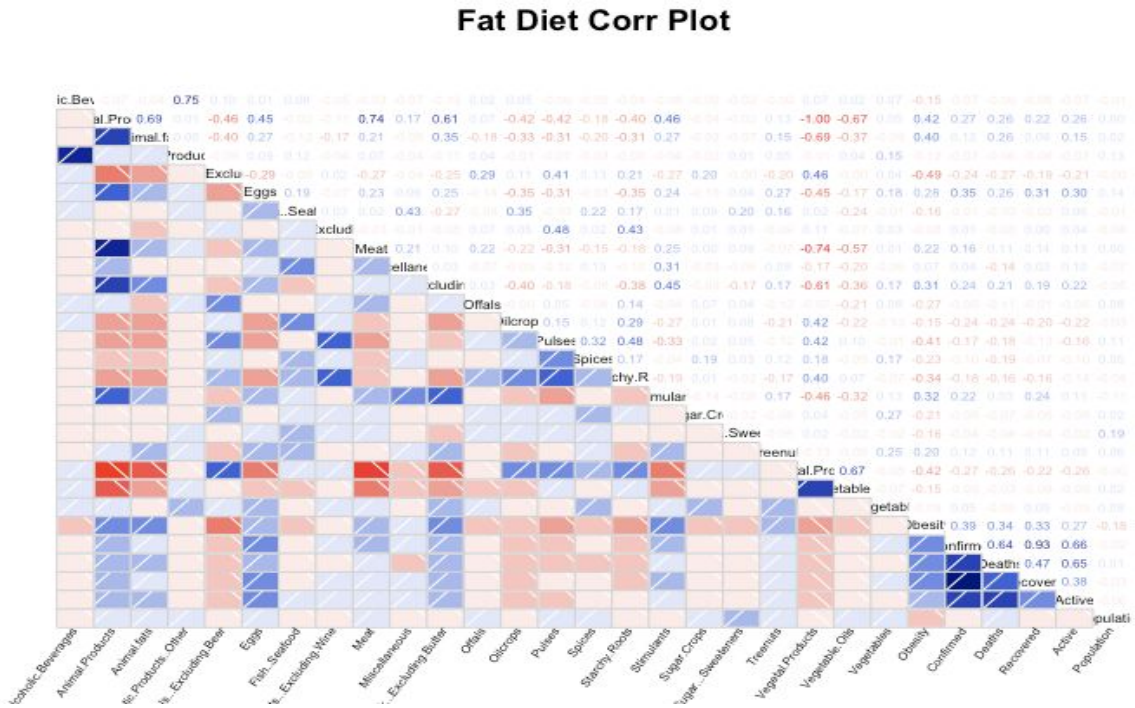


Figure 6: Correlation plot for Covid

Recovery Rates by Country

Map based on Longitude (generated) and Latitude (generated). Color shows details about Recovered (bin). Size shows sum of Population. The marks are labeled by Country.

Figure 7: Recovery Rates by Country

Death Rates by Country

Map based on Longitude (generated) and Latitude (generated). Color shows details about Deaths (bin). Size shows sum of Population. The marks are labeled by Country.

Figure 8: Death Rates by Population

Appendix B: Explanatory Data Analysis

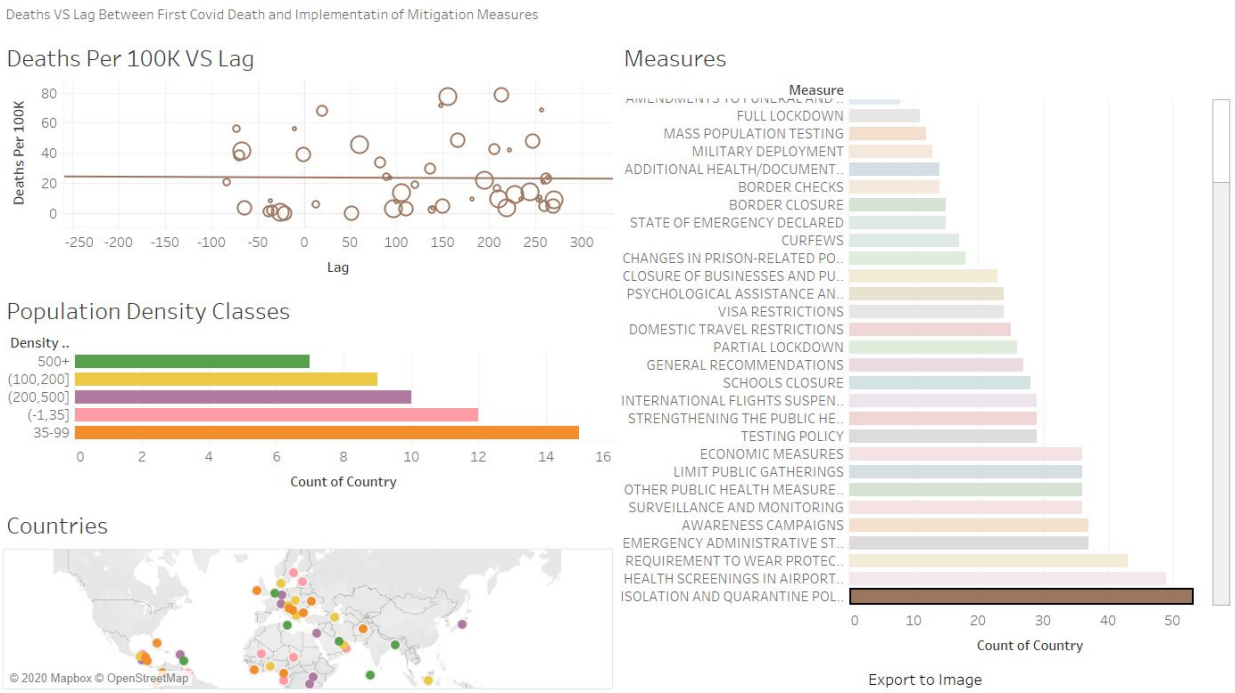


Figure 1: Examining Isolation and Quarantine Policies.

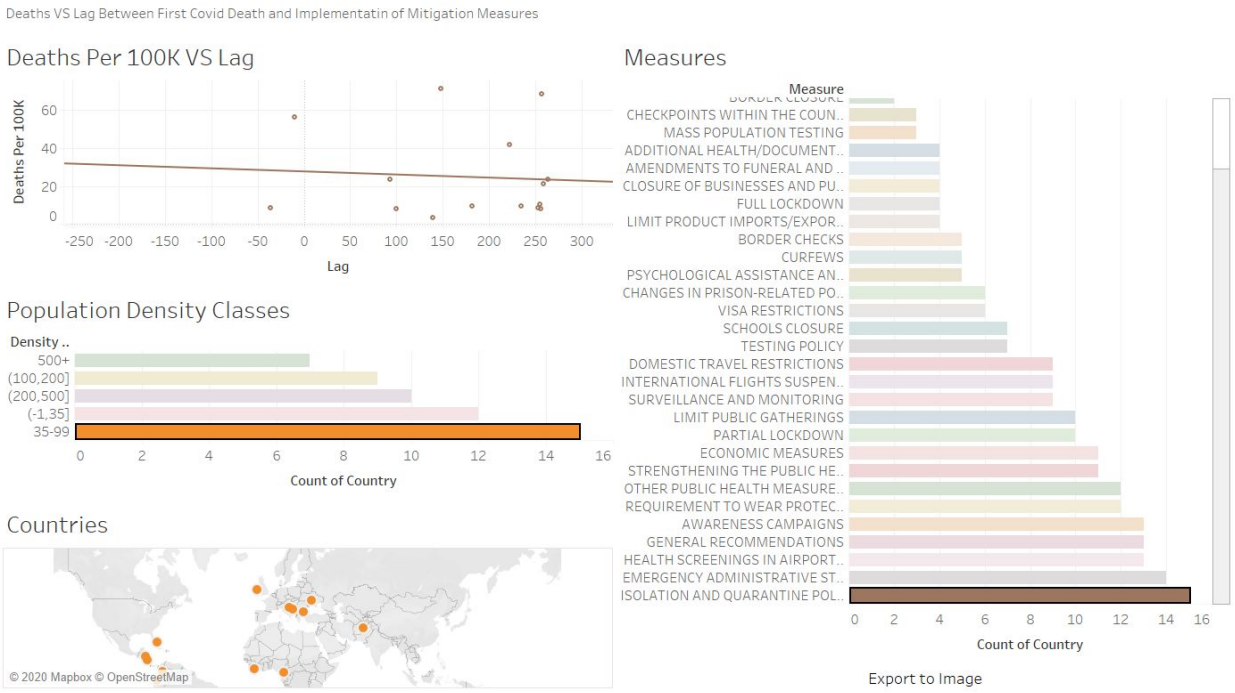


Figure 2: Examining Isolation and Quarantine Policies for Lower Population Density Countries.

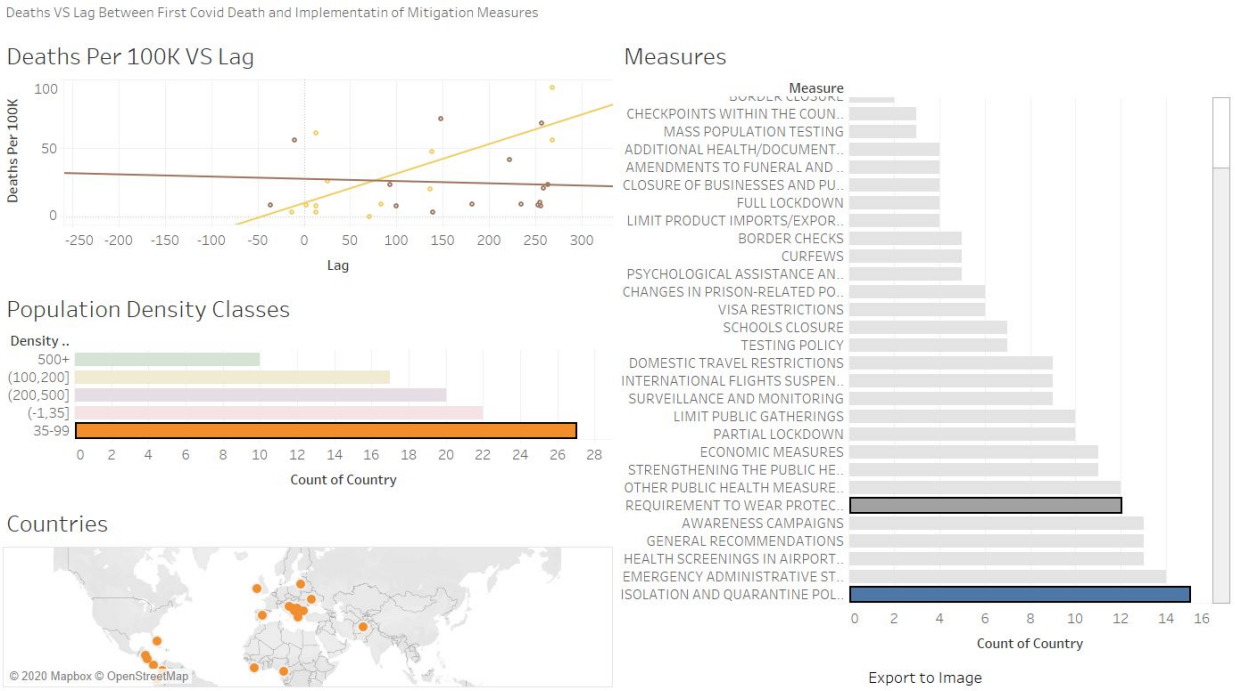


Figure 3: Comparing Isolation and Quarantine Policies to Requirements to Wear Protective Gear in Public for Lower Population Density Countries.

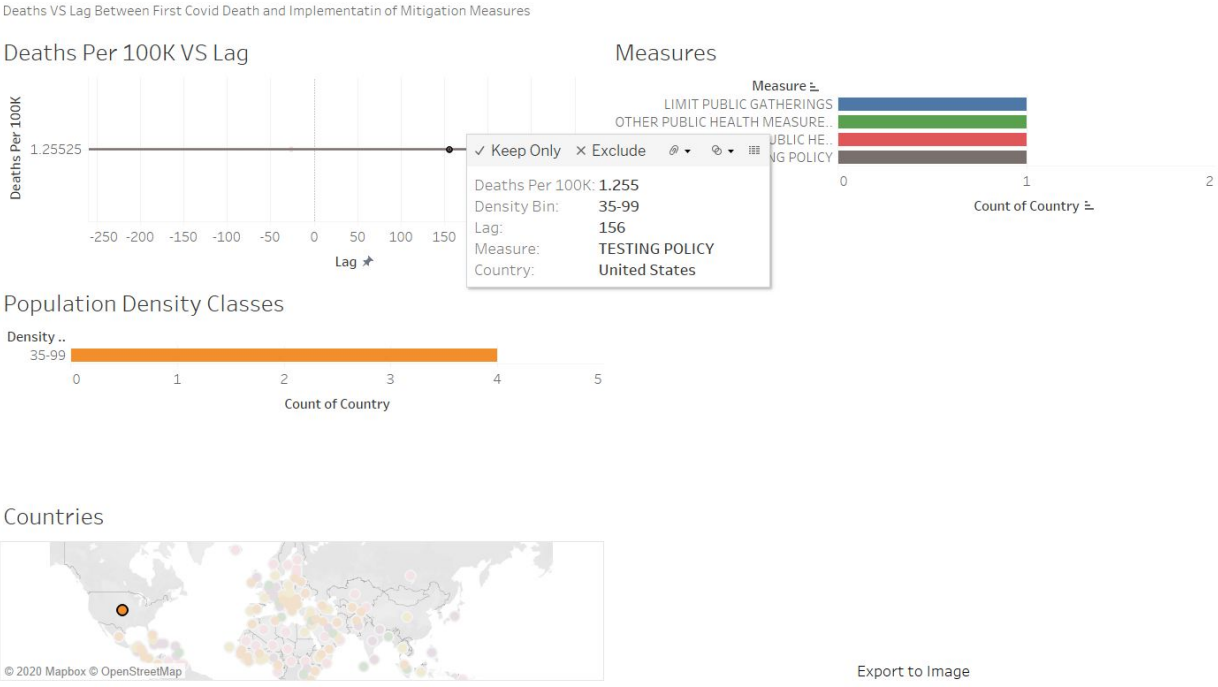


Figure 4: United States Measure Details.

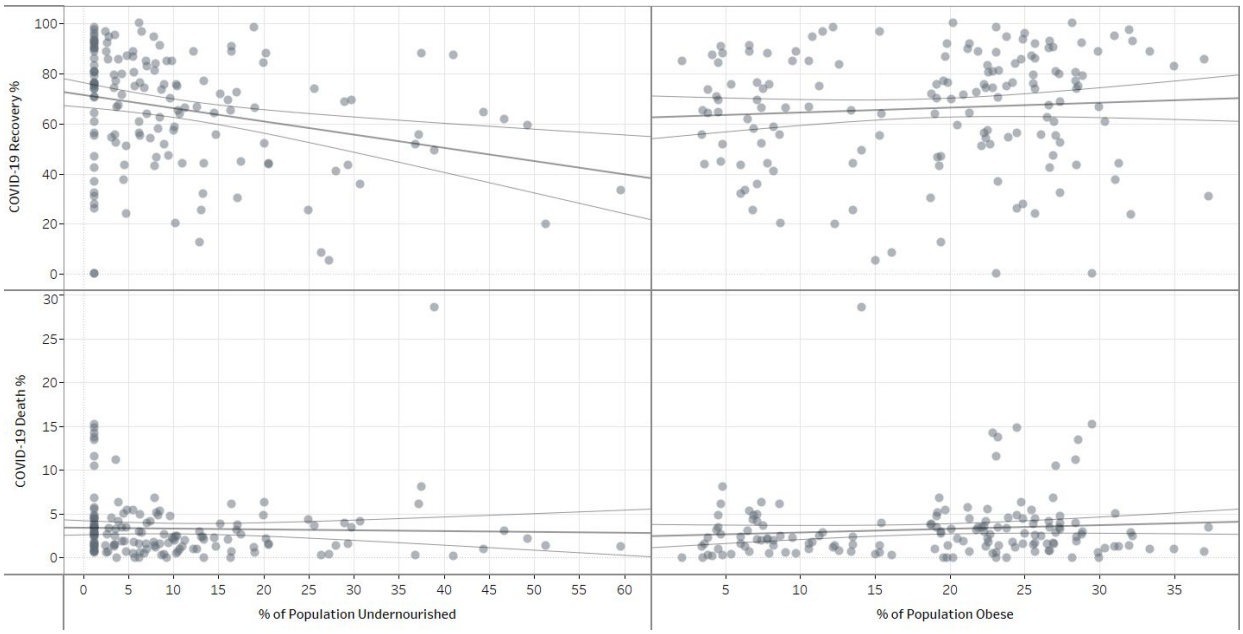


Figure 5: COVID-19 % of Deaths and Recoveries vs Health of Country's Population

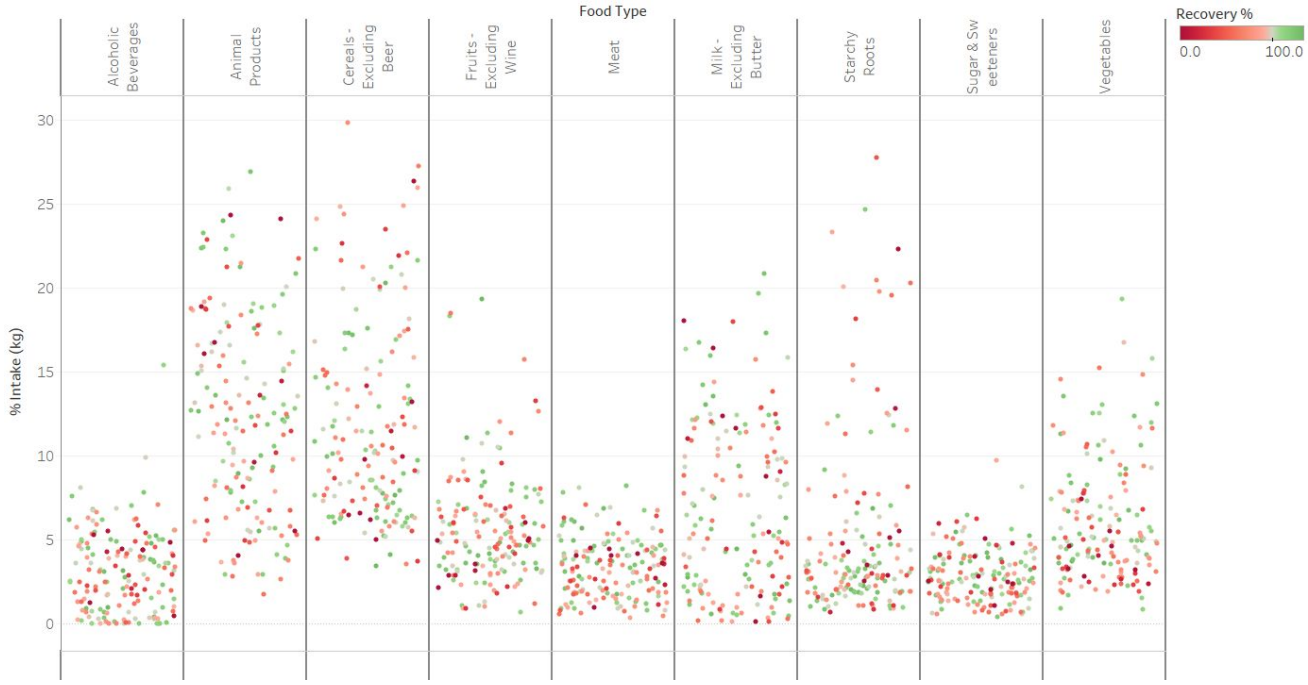


Figure 6: % of kg Food Intake with Recovery Percentage by Food Type

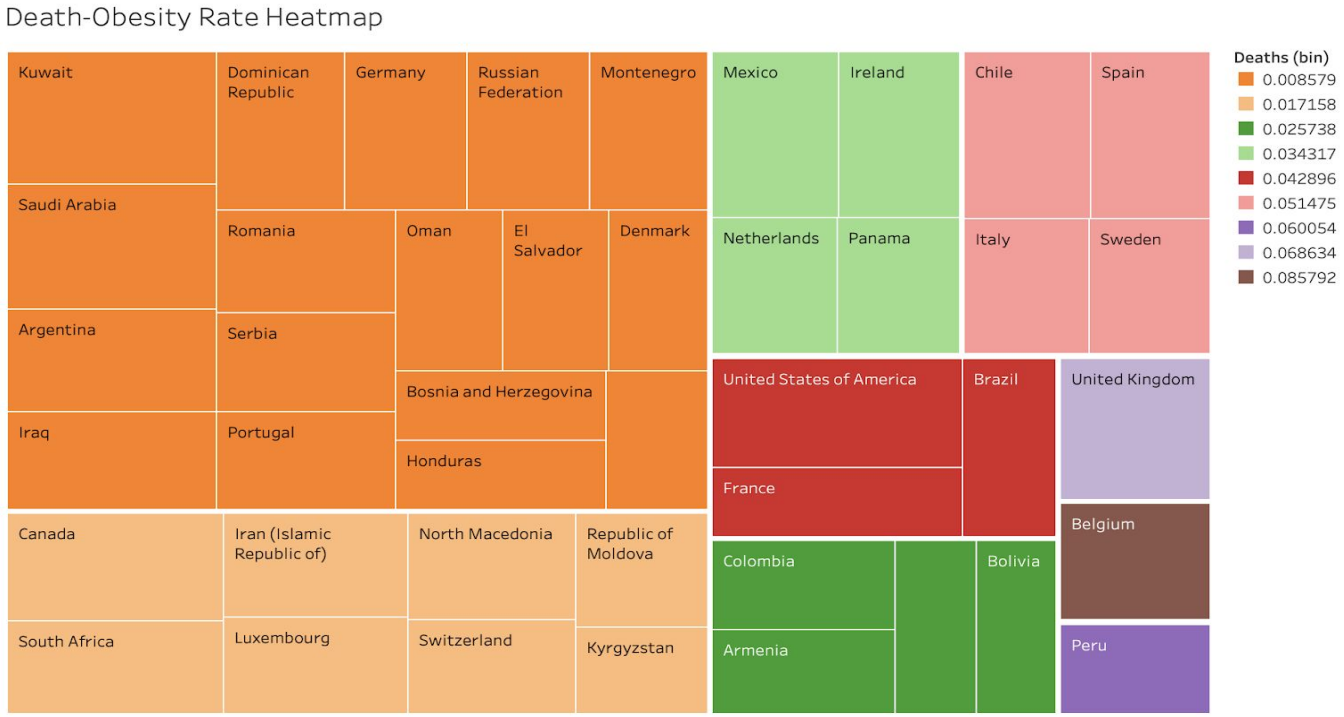


Figure 7: Heatmap Filtered

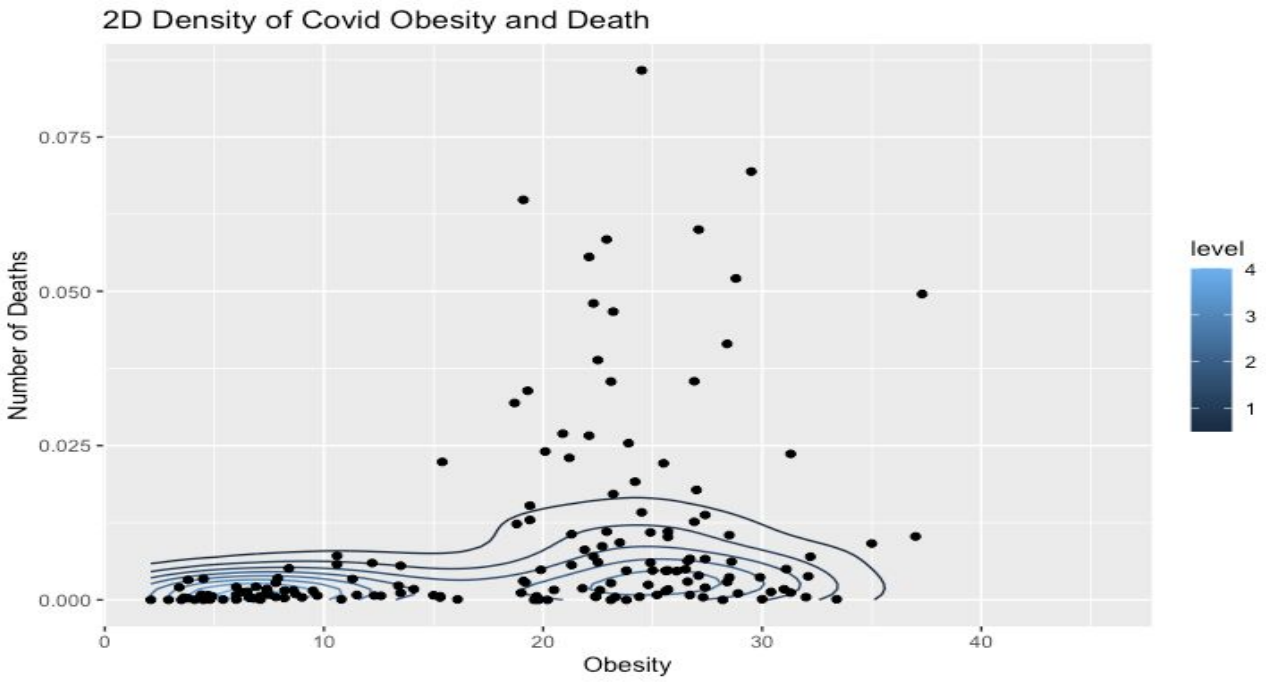


Figure 8: Death/Obesity Contour Plot

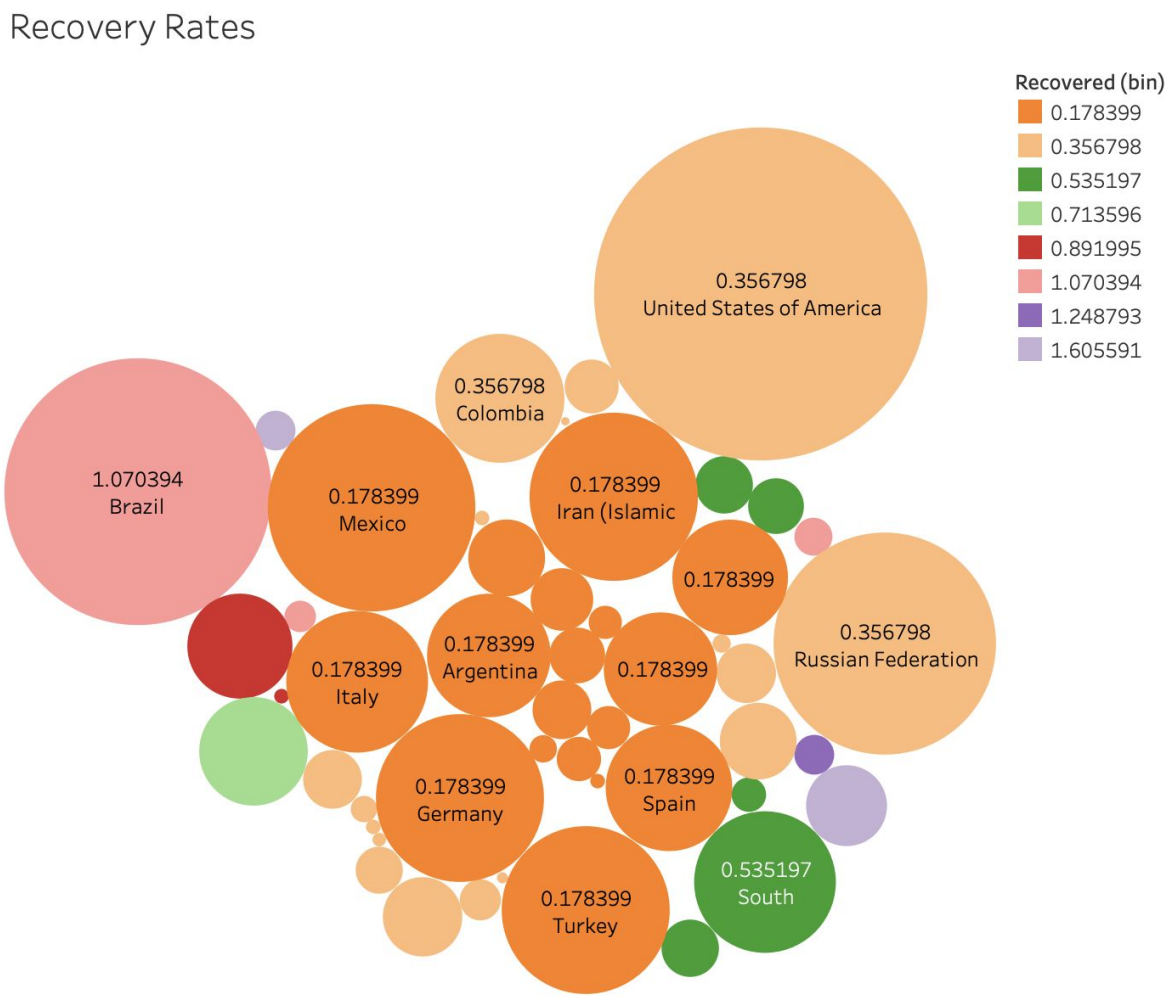
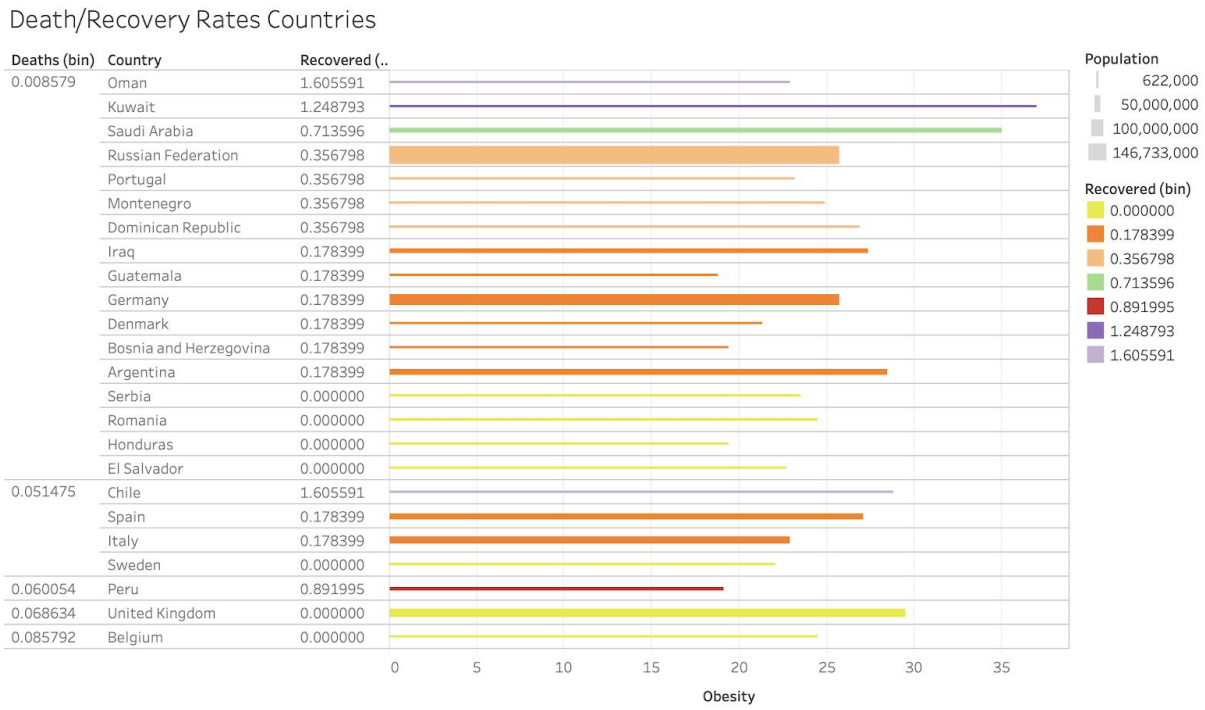


Figure 9: Bubble Chart Filtered



Sum of Obesity for each Recovered (bin) broken down by Deaths (bin) and Country. Color shows details about Recovered (bin). Size shows sum of Population. The view is filtered on Recovered (bin) and Deaths (bin). The Recovered (bin) filter keeps 10 of 10 members. The Deaths (bin) filter keeps 0.008579, 0.051475, 0.060054, 0.068634 and 0.085792.

Figure 10: Bar Plot Filtered along with population