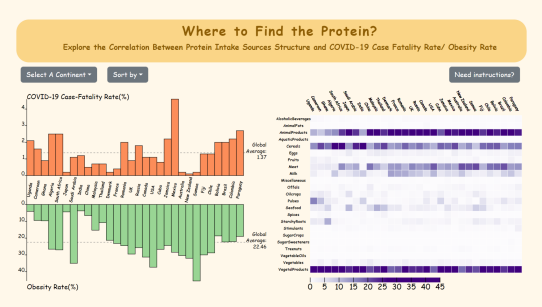
Title

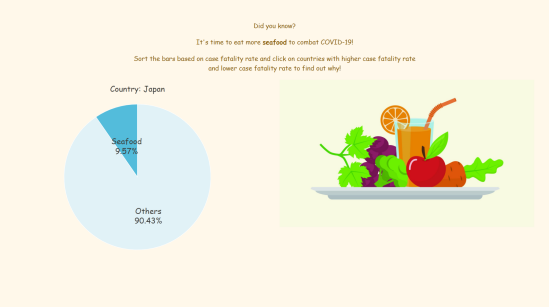
Where to Find the Protein? Explore the Correlation Between Protein Intake Sources Structure and COVID-19 Case Fatality Rate/ Obesity Rate

Members

Ivana Li (yl8060); Christina Yu (ty2233)

Overview





Diet and health have always been closely related. When measuring a person's health, obesity, a common and costly chronic disease often affected by diets, is an important indicator closely related to the potential incidence of various diseases. Moreover, in recent years, people's lives have been greatly affected by COVID-19. Many medical interventions were focused on the development of vaccines. Despite that, non-pharmaceutical interventions gradually come to the stage as people start to be aware of and regulate their health status. Nonetheless, not sufficient attention was put on how people can adjust their diets to prevent diseases, especially infectious diseases. In order to maintain good health from chronic disease and improve the immune system against the COVID-19 virus, we proposed a visualization system to help people better understand the correlation between two pairs of different variables, specifically between protein intake structure and COVID-19 case fatality rate as well as obesity rate.

Data

We visualized a dataset consisting of 170 records from 170 countries, with each record consisting of certain attributes: 23 of them represent the weights of different sources of protein intake, such as the weights of protein intake from animal products, aquatic products, and so on; And these weights sum up to 100%. One of the attributes is the obesity rate of the country; one of the attributes represents the COVID-19 case fatality rate within that country.

We preprocessed raw dataset from kaggle[1] using excel to clean up invalid values and extract useful information and retain 30 records from 30 countries from different continents (6 in total except Antarctica) in order to maintain the diversity of the dataset.

We check the coronavirus resource center webpage from John Hopkins University[2] and found that the original death rate from the kaggle dataset was not that accurate so we replaced the death rate indicator to case fatality rate indicator which is more accurate. The case fatality rate was computed by the number of deaths divided by the number of confirmed cases so it can also manifest one's immunity. We can merge these information together because people's food intake structure doesn't change a lot within two years so we can link the protein sources intake

[1] https://www.kaggle.com/datasets/mariaren/covid19-healthy-diet-dataset
[2] https://coronavirus.jhu.edu/data/mortality

data to the most up-to-date case fatality rate. The original data also includes attributes such as undernourished rate, COVID-19 confirmed cases rate, etc. Since these attributes are not related to our main goal of visualization, in order to preserve clarity and conciseness in the visualizations, these attributes were discarded as well. We also add information such as income levels from the World Bank dataset but this information does not contribute to our final design.

Goals and Tasks

Our goal is to uncover the correlation between two pairs of different variables, specifically between protein intake structure and COVID-19 case fatality rate as well as obesity rate.

We assume our user to be a professional nutritionist called Diana. She wants to have a comprehensive understanding of the correlation between protein intake structure and health, to develop an optimal protein intake structure for people in different regions. When Diana visits the webpage, she will see two views, a symmetric bar chart and a heat map. In the heat map, the horizontal axis represents the countries and the vertical axis represents the sources of protein intake. We used a purple sequential colormap to fill the cells. As the legend shows, the greater the percentage of protein intake among all sources, the darker the color of the cell in which it is located. In a symmetric bar chart, there is one horizontal axis and two vertical axes. The horizontal axis represents countries, one of the vertical axes represents obesity rates and the other vertical axis represents COVID-19 case fatality rates. Diana can choose any of the vertical axes to sort on, and she can select a specific continent to focus on the countries in it. In this way, she can also explore relationships between obesity rates and COVID-19 case fatality rates. She can also find out the pattern of people consuming more seafood have lower case fatality rates by clicking on the bar chart to see shifts in percentage of seafood in the pie chart below.

Besides, the two views above are interlinked. When Diana hovers over a cell on the heat map, a tooltip that contains the exact value of the cell will pop out beside the cell. At the same time, the entire vertical column of the country corresponding to that cell will be filled with an orange sequential colormap. It allows Diana to get a clear picture of the protein intake structure in the same country. And when Diana hovers over a bar on the symmetrical bar chart, both the bar and the corresponding upper(lower) bar change to another color, so as to distinguish the information from that of the other countries and vice versa. Meanwhile, a tooltip will pop out with information on the country's name, its concrete obesity rate and its COVID-19 case fatality rate.

In the visualization language, point mark and length channel help users identify the country or specific protein sources. Color hue was used in the bar chart to distinguish indicators and color saturation was used to show protein sources magnitude change. Angle and color hue channels in pie charts help users to realize the importance of seafood in the finding. Interactions include mouse hovering and clicking to help users to find information by linking multiple views.

Visualization

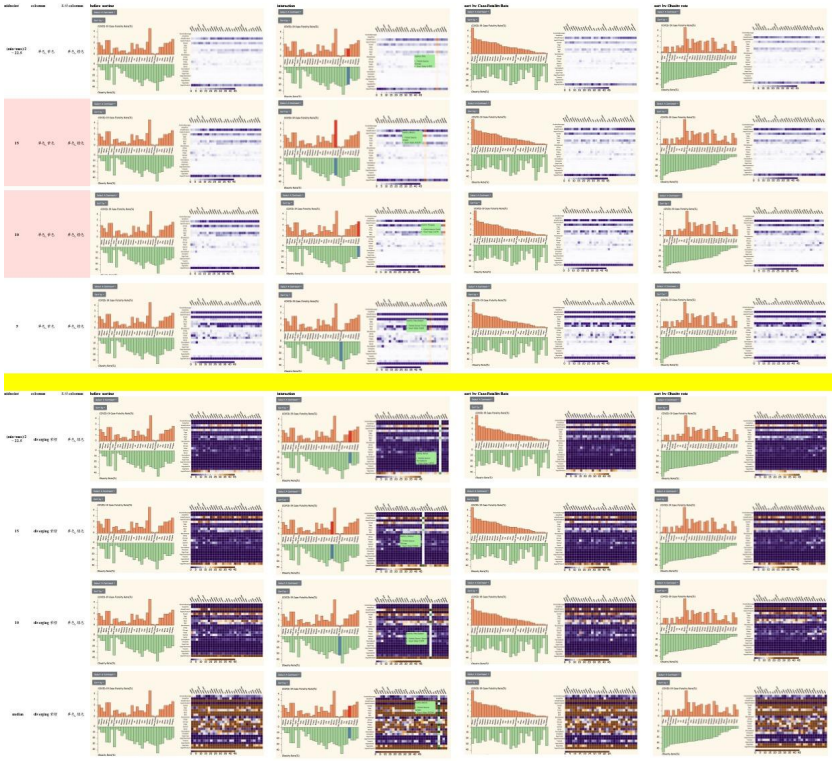
There are 3 views, a symmetric bar chart, heatmap, and pie chart.

Symmetric bar chart allows users to see country obesity rate and case fatality rate information. Two indicators were distinguished by orange and green. When users hover over one bar, two bars from the same countries will be highlighted with another color and exact information will be displayed. Annotation was added to show the average obesity rate and case

fatality rate. When clicking on one bar, a corresponding pie chart will be shown below to display specifically information about the amount of seafood consumed among all other sources.

Heatmap allows users to see the detailed protein sources intake among different counties. Similarly, when hovering over a cell, the column of the country will be highlighted and the exact value of that protein source will be shown.

A comparison has been made between diverging and sequential in terms of the choice of color scheme for the heatmap. In order to clarify which color scheme would represent the data more clearly, while reflecting a certain association with obesity rates and case fatality rate, we created a table. The table compared four views of the diverging and sequential color schemes for the unsorted interface, the interactive interface, sort by COVID-19 case fatality rates interface and sort by obesity rates interface by setting different values of midpoint. The screenshot below is a thumbnail of the table. After the comparison, we found out that the major problem with the diverging color scheme is that the user has to quickly understand that the left side of the middle point is from dark to light, while the right side is light to dark. It can cause confusion for the viewer when two colors appear in the same row or column at the same time, as the value of midpoint has no real meaning. As we learned in the lecture, if we need to let the audience get the information by comparing colors in our visualization, we'd better put objects with similar colors next to each other to make the comparison easier because human beings use relative judgment. The sequential color scheme is therefore clearly more aesthetically pleasing and meets the needs. In the sequential color scheme, when the midpoint is set at 12.5, the possible phenomena between the different protein intake sources and the COVID-19 case fatality rates as well as obesity rates can be presented at a clear glance.



Users can interact with these views using a sort and a select button to filter the data. They can also hover over bars and heatmap to see the link or click on bars to interact with the pie chart.

Reflection

From a general aspect, the project started following the original proposal and plan, we first built up the symmetric bar chart and the heatmap. And then we start looking for some insightful patterns in our dataset to try to tell a story based on our visualizations. We found out some highly acknowledged phenomena such as higher meat consumption leads to higher obesity rate. But we also want to find something novel with respect to COVID-19. Then we find out that countries with lower case fatality rates have a higher consumption of seafood. To help users realize this pattern, we draw a pie chart with two components, seafood and others. In this way, using angle channels which have a higher effectiveness than color saturation, users can observe the difference more obviously. We acknowledge that this pattern exists might due to a lot of reasons such as countries' income level and medical resources, so we add some texts to help users think seafood is a great source of protein as well as helping to navigate to find the pattern themselves.

In a detailed view, we add annotations to the bar chart to show the global/continent level average case fatality rate/ obesity rate. We also find out our protein source data is highly skewed so we tried several methods as described above and finally set down to choose a value of 12.5 as the middle point of the color legend which shows the best result. In the meantime, we add some bootstrap methods to make the webpage more beautiful and friendly to mobile users. Finally, there are some limitations we found along the building process, such as the protein source percentage data does not mean the absolute percentage among all the sources because some protein sources are incorporated into other protein sources. We tried to contact the dataset contributor in kaggle but found no response. Since the dataset is licensed and used by a great number of people and has authority, we assumed that the protein source dataset was normalized to make sure the sum of 23 sources is 100%. However, in order not to confuse users and keep accuracy, we add this information to our instruction part. Nevertheless, we found it important to always look into our dataset more carefully and ask for clarification as soon as possible.

Our visualization goals change a little bit. We still retain the exploratory purpose of the visualization to allow users to find information by themselves. Upon this goal, we also accomplished to convey our findings to users by creating another view and some navigation text. As for the technical goals, due to the limited aspect of the dataset and to remain clarity and conciseness, we didn’t add much more complexity to our original view design.