Interactive Data Visualization Final Reflection

Instructor: Prof. Ellie Frymire

Zhan Gong

The Distance that Matters:

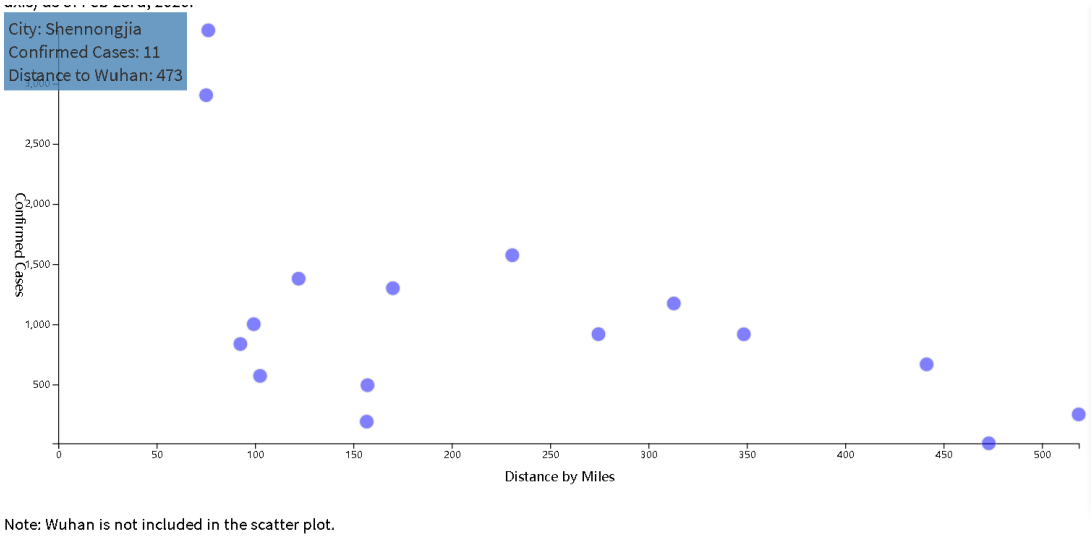
Spread of the Coronavirus Outbreak in Hubei Cities

Data Feb 23, 2020

Cities	Distance/Kilometerss From Wuhan	Confirmed Cases
Enshi	519	251
Shennongjia	473	11
Shiyan	441	667
Yichang	349	917
Xiangyang	313	1,173
Jingmen	275	918
Jingzhou	231	1,574
Suizhou	170	1,300
Tianmen	157	494
Qianjiang	157	191

Recovered: 0 Deaths: 0 Distance: 0 ConfirmedCases: 0

Figure 1



Note: Wuhan is not included in the scatter plot.

Figure 2

Abstract

My project examines the relationship between the original epicenter of coronavirus, Wuhan, and its nearby cities. Wuhan is located in Hubei Province, which is part of the Central China region. Besides Wuhan, there are 16 cities in Hubei. Together, Hubei is home to 60 million population. As of February 23rd, 2020, more than 46,607 COVID-19 cases were confirmed in Wuhan (Baidu 2020). The Chinese authorities closed off Wuhan on January 23rd; in the following week, other cities in Hubei Province also went into lockdown to control the spread of the virus. Among these cities, it is reasonable to assume that the number of confirmed coronavirus cases in one city may be directly related to its proximity to Wuhan. Cities near Wuhan may have a higher exposure to the virus than cities further away, thus having a higher number of people infected. In this project, I visualize the association between the geographic distance from each city in Hubei province to Wuhan and its total confirmed cases as of February 23rd.

My data was obtained from the website https://voice.baidu.com/act/newpneumonia/newpneumonia, which collects the most updated official coronavirus statistics in China from the National Health Commission and provincial and city governments. I found that, in general, there is a positive relationship between the distance between the city in Hubei province and Wuhan and its total confirmed cases. Cities near Wuhan had higher exposure to the virus than cities further away and thus had a higher number of infected people, though not every city fits nicely into this pattern. This finding is crucial, for it could help policymakers

in other countries create more effective quarantine policies. Instead of making a separate plan for each city, it may be more effective if neighboring cities develop a uniform standard to contain the virus.

Design and Challenges

I chose to use several charts to visualize my data, including the table chart, the bar chart, a regional map, and one scatter plot chart (I added the scatter plot in my final project). The explanatory variable in this study is the geographic distance (in kilometers) between cities in Hubei province and Wuhan (For Wuhan, I set the distance to be 1 km). The response variable is the number of confirmed coronavirus cases in each city as of February 23rd.

My whole project idea was based on tutorial 7, enabling the interaction between tables and bar charts. My idea was that when one clicks on the name of the city in the table, the height of different bars in the bar chart will change accordingly, displaying additional information such as the number of people that were recovered and the number of deaths in one city. I made several minor changes based on codes from tutorial 7. Specifically, I added more bars to the bar chart, as I had four categories to be visualized: recovered, deaths, distance, confirmed cases. Furthermore, I created a color scale for the table, with darker blue indicating more confirmed COVID-19 cases, while lighter blue indicates fewer cases.

To better visualize whether the city is far away or adjacent to Wuhan, I decided to add a regional map of Hubei. I started to look for a JSON file of China, but it wasn't

easy to find a file separated by provinces. Thus, I chose a world JSON file and created the Hubei province map by myself. However, after I completed the project, I found a much easier way to create a JSON file of Hubei, using the website DATAV.GeoAtlas: http://datav.aliyun.com/tools/atlas/#&lat=18.562947442888312&lng=112.32421875&zoom=3. The website contains all city and provincial level geographic data of China, which could be a useful source for any researchers who study China.

After I got the JSON file, I reviewed the codes in tutorial 5. The first problem I encountered was that if I wanted to put all charts' data in the console, I had to load the data in one file. To solve this, I used "promise all" to load the data and put the geodata into "state.geojson." The second problem was that while tutorial 5 had many dots, my data was composed of polygons (different cities). Hence, I had to write my own codes to create a color scale on my regional map.

Changes after feedback

I had a brief discussion with Prof. Frymire before I continued my final project. What I learned was that adding a scatter plot was necessary. I got the same feedback when I presented my project. In my final project, I created a scatter plot to visualize the relationship between the distance to Wuhan (X-axis) and confirmed cases (Y-axis). It should be noted that the city of Wuhan was excluded from this scatter plot. I adapted codes from tutorials 3, 4, and 6 and generated a hover, so when I clicked on each point, a box would pop up and tell me more about what this dot means: the name of the city, confirmed cases, and distance to Wuhan.

Another valuable feedback was that one of our classmates pointed out that the color scale used in the table (which was sorted by distance) was not consistent with the color scale used in my regional map (sorted by confirmed cases). I appreciated this comment a lot and fixed the color scale afterward. The color scale now indicates the frequency of the total confirmed cases in both the table and the map.

My final project was almost the same as what I expected in my prospectus, except that I wasn't planning to create a scatter plot in the beginning. This feedback taught me to think about the relationship between my data and the best visualization type that I choose to use. For example, tables are useful to present numerical data, whereas scatter plots can reveal the relationship between two different variables, and it is easy for identifying outliers. Using the scatterplot, one can quickly identify the overall positive correlation between the distance from each city in Hubei province to Wuhan and its total confirmed cases. Furthermore, another helpful thing I learned during the project is that I practiced a lot of my coding skills. After I completed the project, I'm much more familiar with D3 now.

Limitations and Future study

There are several things that I would like to change if I had more time and resources. First, I would make the table, the bar chart, and the regional map interact simultaneously, so when one clicks on any of the three charts, the other two will change accordingly (now I can only create the interaction between the bar chart and the table). Second, I chose the date of February 23rd when I started this course. It

would be better if I kept track of the COVID-19 cases and adopted a more updated dataset. If possible, I would also like to visualize the association between confirmed cases in New York State and other neighboring states. Third, it would be great to add some pointing arrows to my regional map, which directly connects the centroids of the city and Wuhan and shows the distance between the two. Fourth, my final project does not apply any inferential statistics but only presents descriptive data. Future research should use more rigorous spatial analysis methods to test the relationship between the distance and the confirmed cases. Moreover, factors such as a city's access to testing, the logged GDP per capita, and medical resources may all affect a city's total confirmed cases. Hence, future research should control these variables.

References:

1. Baidu. 2020. “实时更新：新型冠状病毒肺炎疫情地图.” Retrieved May 19, 2020 (https://voice.baidu.com/act/newpneumonia/newpneumonia?city=湖北-黄冈).