Report: The distribution of coronavirus in China and whether the spread of the virus is related to the population and GDP of provinces

Report Prepared By:
Yixue Wang

February 27, 2020

Background

Coronaviruses (CoV) are a large family of viruses that cause illness ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV) [1]. Coronaviruses are zoonotic, meaning they are transmitted between animals and people. This report will focus on a novel coronavirus that has not been previously identified in humans, Coronavirus Disease 2019, first discovered in Wuhan, Hubei Province, China.

Problem Description

For most people, images are always sharper and more intuitive than data. It's easier to catch the point of the problem from the image. This paper will be presented the distribution of coronavirus in China in the form of geographic images and whether the spread of the virus is related to the population and GDP of provinces.

Objectives and Metrics

The purpose of the project is to determine the correlation between the number of people diagnosed with coronavirus in China, the population of each province, and GDPpc.

The indicators we can use to confirm the correlation are the number of diagnosed coronavirus in 2019, the geographic latitude and longitude of each province in China, population data, and GDP data.

Source of the data

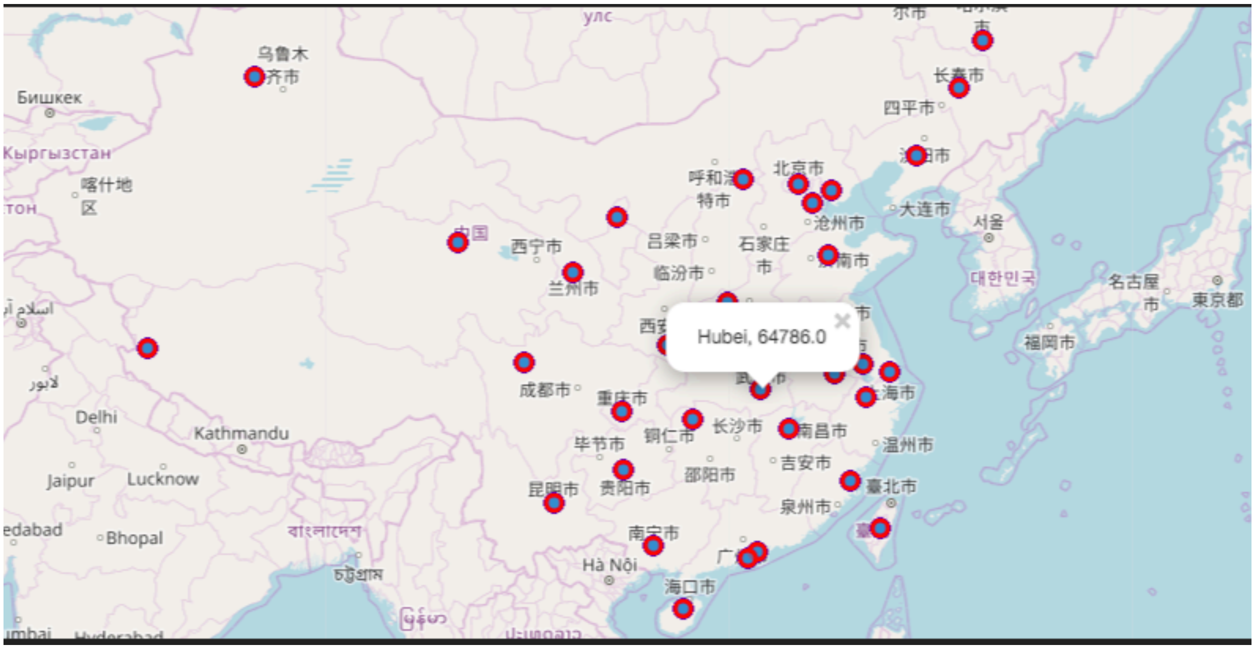
So far, the coronavirus is still spreading, and people have different understandings about where the virus came from and how to define and treat it. Because medical data is only accessible to insiders in any country, it is not easy to get first-hand data.

Fortunately, the Chinese government has adopted a more open and transparent approach to the coronavirus data. The Ministry of Health of each province publishes the latest confirmed diagnosis daily on the official website.

Besides using Foursquare location data to define the provinces' locations, this time I also collected research data from sina.com [2], an authoritative website in mainland China. The data includes the names of provinces, dates, number of confirmed patients, number of cures, and number of deaths. In addition, in order to study whether the spread of the virus is related to the population and GDP of each province, I also searched and downloaded the latitude and longitude of each province, the population and GDP in 2018 on google.ca

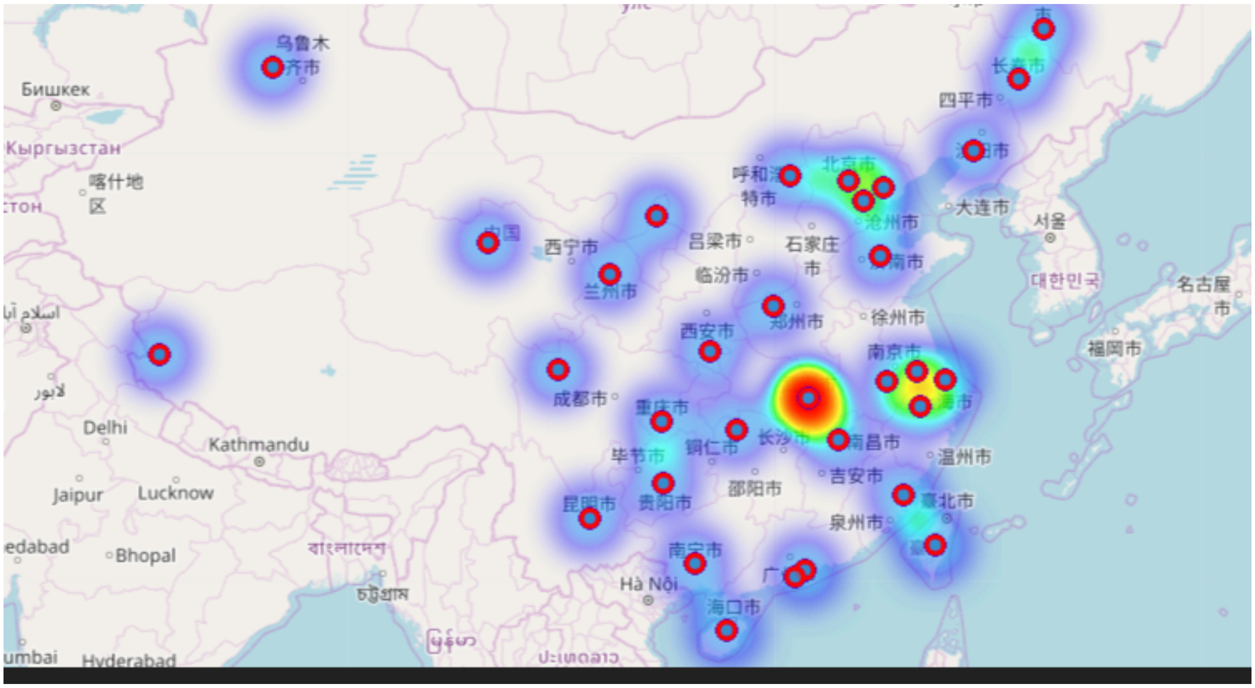
Methodology and data analysis

First, I used markers on the map to visually show the distribution of the epidemic in China. All provinces with confirmed cases will be displayed on the map, and when the user clicks on a marker, the map will show the name of the corresponding province and the number of confirmed people. The data I used in the following picture is the epidemic data on February 24, 2020.

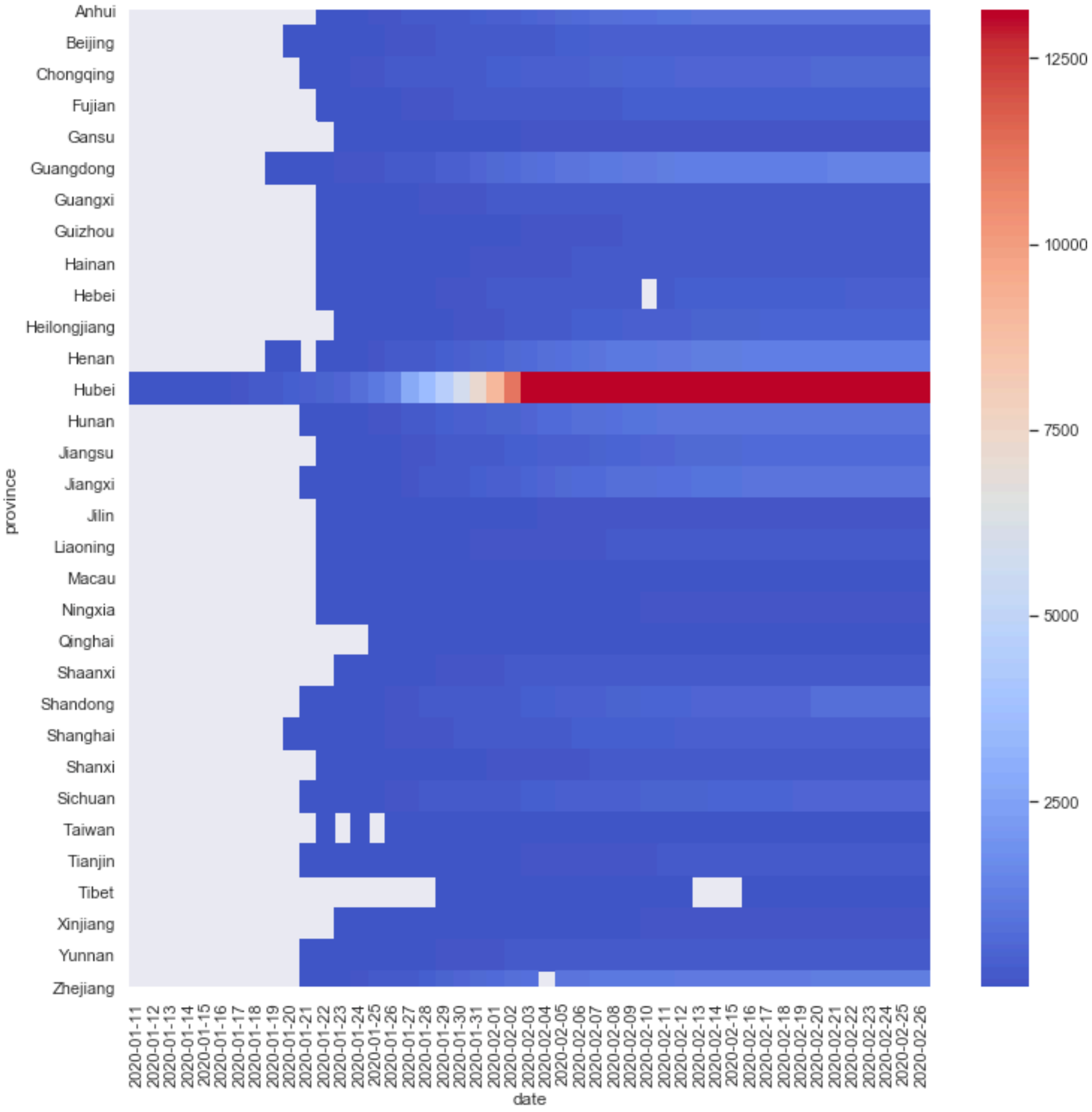


In addition, I also calculated the number of venues within 5,000 meters of each province via Foursquare. Foursquare is a platform that can help users find the venues around that location by entering the name of that location. It contains ordinary user version and programmer version, which can realize many commercial applications.

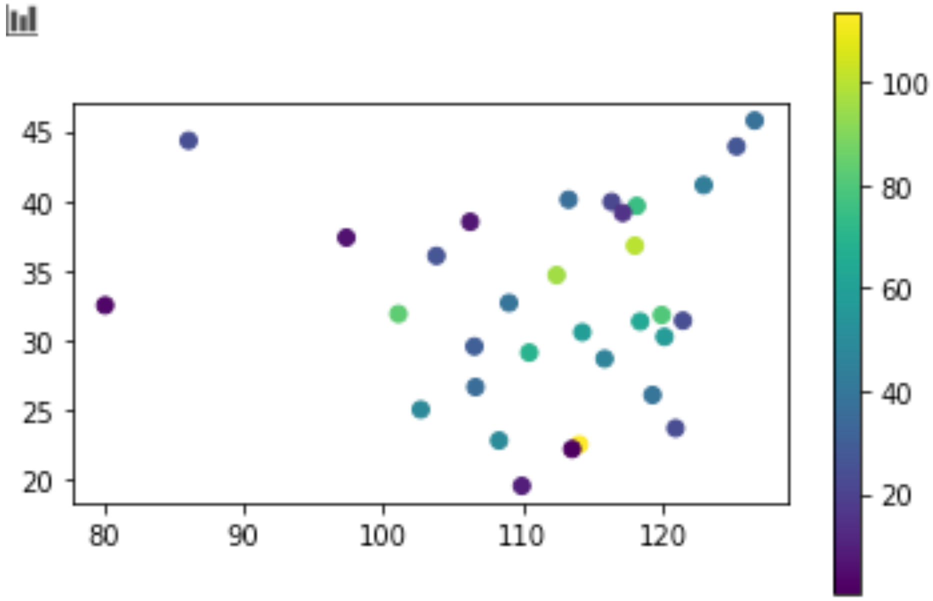
Markers on the map can show the distribution of data very well, however, it cannot let users discern the difference in degree at a glance. Hence, I thought of using heatmap to show the change in the amount of data. In this example I used the plugins function under folium. In the figure below, I still use the data of February 24, 2020. The difference is that the heatmap shows different colors according to the number of people diagnosed, so that users can immediately know the severity of infection in different areas.



The time dimension will then be taken into account. The following heatmap uses the seaborn package. We use province as the ordinate, time as the abscissa, and value is the number of confirmed patients. The bar on the right shows the degree of the number of people, which is indicated by blue to red.

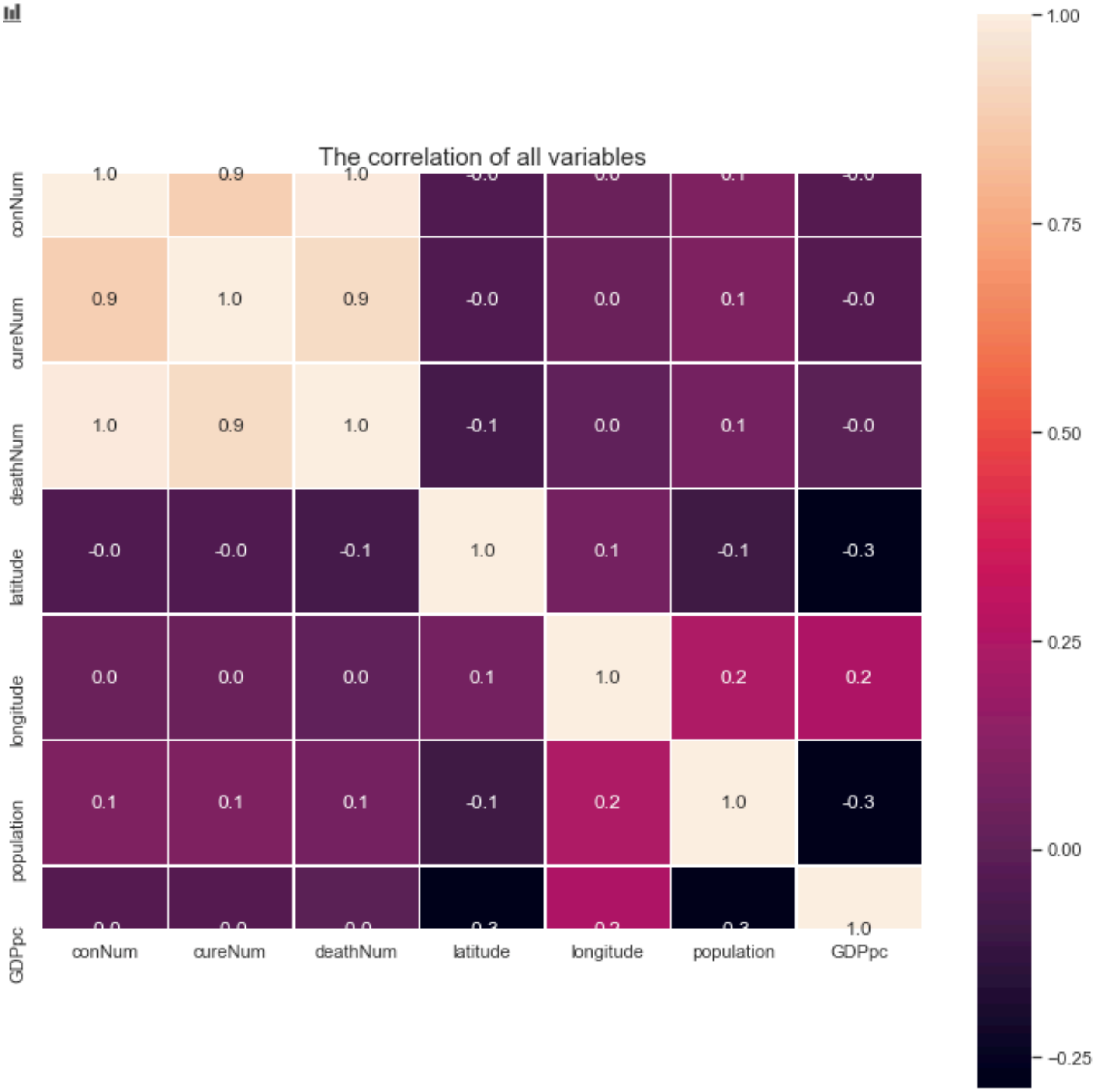


Next, I will show the population and GDP of each affected province in the form of a Choropleth chart. Here I used matplotlib.pyplot. It is a collection of command style functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc[3].



Choropleth Map of each infected province

Finally, in order to determine whether the population and GDP are related to the number of patients diagnosed. I plot the heatmap on all numerical variables to get an overview of relationships between them. In the figure below, a larger number indicates a higher correlation. The maximum value is 1.



Results Discussion and recommendations

Through the observation of heatmaps, we can find that in the past two months, the high incidence of coronavirus in China in 2019 was mainly concentrated around Hubei and Jiangsu. From the number of confirmed cases, the number of people in Hubei province is much higher than in other provinces, and there is a surge in the number of people in late January and early February. From the perspective of correlation, the number of confirmed patients, the number of cured patients and the number of deaths have a very high positive correlation. However, from the analysis of current data, it is difficult to see the impact of population and GDP on virus transmission.

My recommendation is that in order to further refine this study, broader and more accurate data should be included. For example, I can expand the research to include data from other countries. It can also expand accuracy on this research, such as collecting data for each city, otherwise, include longer time frames and add more variables.

Conclusion

This project uses many heatmaps to display research results. From the analysis of the data, we can conclude that the high and frequent areas of coronavirus in China in January and February 2020 were mainly concentrated in Hubei Province, and the number of confirmed patients increased significantly at the end of January and early February. In addition, the impact of population and GDP on the number of diagnoses cannot be seen from the currently limited data. Hope that further research can find more details to help us increase our understanding of this coronavirus transmission.

1. https://www.who.int/health-topics/coronavirus
2. https://news.sina.cn/zt_d/yiqing0121?vt=4&cid=257615&node_id=257616
3. https://matplotlib.org/tutorials/introductory/pyplot.html