Part1 Dataset

1. Dataset with number of confirmed, cured and death cases every in different provinces across China
   https://github.com/BlankerL/DXY-COVID-19-Data/blob/master/csv/DXYArea.csv
2. Population of China's provinces： http://data.stats.gov.cn/
3. The number of passengers from Wuhan to each province from 1.1 to 2.3
   http://qianxi.baidu.com/
4. China map data, accurate to the provincial level
   https://github.com/echarts-maps/echarts-china-provinces-pypkg

Part 2: Library

I use **pyEcharts** for visualization.

Part 3: Library

1. Six COVID-19 related questions

(1) China's epidemic prevention and control are nearing completion, and I am interested in the severity of the epidemic in different regions. So, the first question is how to visualize the cumulative confirmed cases in different provinces.

(2) Besides the cumulative confirmed cases, I also want to analyze the trend of COVID-19 in China. We know that the epidemic first broke out in Hubei and then spread to other regions. So, this time, I focus on not only the overall trend in China, but also the trends in and outside Hubei. The second question is how to visualize the everyday existing confirmed cases in China, in and outside Hubei.

(3) In Q1, we get the absolute number of confirmed cases in different provinces. However, we also want to show the difference among different provinces more intuitively. Visualizing the proportion may be a good idea. Also, I am interested in the severity of the epidemic in different geographic regions in China. The third question is (a) for different geographical areas' confirmed cases, how to visualize their proportions in China's total confirmed cases, and (b) for different provinces' confirmed cases, how to visualize their proportions in one geographical area's total confirmed cases.

(4) When assessing the risk of an area in an epidemic, the number of confirmed cases is a crucial indicator, reflecting the local epidemic's scale. The growth rate of confirmed cases reflects the development speed of the epidemic. Combining these two indicators, the degree of danger of the regional epidemic can be assessed from two aspects: importance

and urgency. The fourth question is how to analyze the regional epidemic risk in terms of importance and urgency.

(5) The correlation analysis part. I want to analyze whether the spread and cure of the epidemic are related to social factors, such as population and transportation. The fifth question is how to visualization the relationship between population, transportation, and confirmed cases.

(6) I want a more comprehensive assessment of the current situation and try to use a comprehensive evaluation method to establish a regional risk index. The sixth question is for each province, how to establish a regional risk index and visualize it.

2. Analyze the information from the data

Overall, in my dataset, there are 8 variational concepts (components). They are
province name,
confirmed cases,
cured cases,
death cases,
date,
population,
number of passengers,
province boundaries.

The first dataset is a **table** that consists of a series of **items** and 5 **attributes**. The 5 attributes are province name, confirmed cases, cured cases, death cases, and date. Among them, the type of province name is **categorical**, and the types of the other 4 attributes are **quantitative**.

The second dataset is a **table** that consists of 34 **items** and 2 **attributes**. The 2 attributes are province name and population. The type of population is **quantitative**.

The third dataset is a **table** that consists of a series of **items** and 3 **attributes**. The 3 attributes are province name, number of passengers and date. The types of the number of passengers and date are **quantitative**.

The fourth dataset is a **geometry** which consists of a series of **items** and **positions**.

3. what data analysis I need to do

(1) For the **table** dataset, I store them in a CSV file and use pandas to read it into a data frame. For each province, I select the item with the most recent update time and use the “confirmed cases” attribute as the cumulative confirmed cases.

For the **geometry** dataset, when I use pyEcharts to draw the map, I set parameter maptype = ‘china’ to load China map data accurate to the provincial level.

(2) I use the same dataset as Q1. Existing confirmed cases are computed by confirmed cases – cured cases – death cases. For each updateDate, I aggregate the existing cases in all provinces to get the existing case numbers in China. Cases outside Hubei = cases in China – cases in Hubei.

(3) I use the same dataset as Q1 and select the item with the most recent update time. I aggregate the confirmed cases in different provinces by their geographical areas. Then, for each geographical area’s confirmed cases, I compute its proportion in China’s total confirmed cases, and for each province’s confirmed cases, I compute its proportions in one geographical area’s total confirmed cases.

(4) After answering Q2, I find non-Hubei confirmed cases peaks on around 2.11. I select the confirmed cases and compute the weekly growth rate on 2.11 for each province.

(5) The dataset I get contains the number of passengers from Wuhan to each province from 1.1 to 2.3. So, I decide to analyze the relationship between population, transportation and confirmed cases on 2.3. I compute the number of passengers from Wuhan to each province per day.

(6) To establish a regional risk index, I consider four indicators: the number of confirmed cases, growth rate, population, and the number of passengers. Before integrating the indicators, I normalize them and then use PCA to extract the first principal component and rank according to its value.

4. the visualization design process

(1) marks: areas
channels: spatial region for province name (categorical attribute), color with a color map for confirmed cases (ordered attribute)
visual variables: geospatial location, color with a cool to warm color map
For different provinces, I arrange them on China map based on its geospatial location, and I map the cumulative confirmed cases to the color variable via a color map.

(2) marks: lines
channels: horizontal and vertical positions, color
visual variables: 2D plane, color
I map date (ordered attribute) to the horizontal postion, existing confirmed cases (ordered attribute) to the vertical postion, and China, Hubei, non-Hubei to color (categorical attribute).

(3) marks: area
channels: central angle and color
visual variables: size and color
I map the propotion (ordered attribute) to size (central angle) and geographic regions or

provinces (categorical attribute) to color.

(4) marks: points
channels: horizontal and vertical positions, color
visual variables: 2D plane, color
I map current confirmed cases (ordered attribute) to the horizontal position, weekly growth rate (ordered attribute) to the vertical position, and provinces (categorical attribute) to color.

(5) marks: points
channels: horizontal and vertical positions, color, area
visual variables: 2D plane, color, size
I map annual permanent population (ordered attribute) to the horizontal position, passengers from Wuhan per day to the vertical position, provinces (categorical attribute) to color, current confirmed cases (ordered attribute) to the circle size.

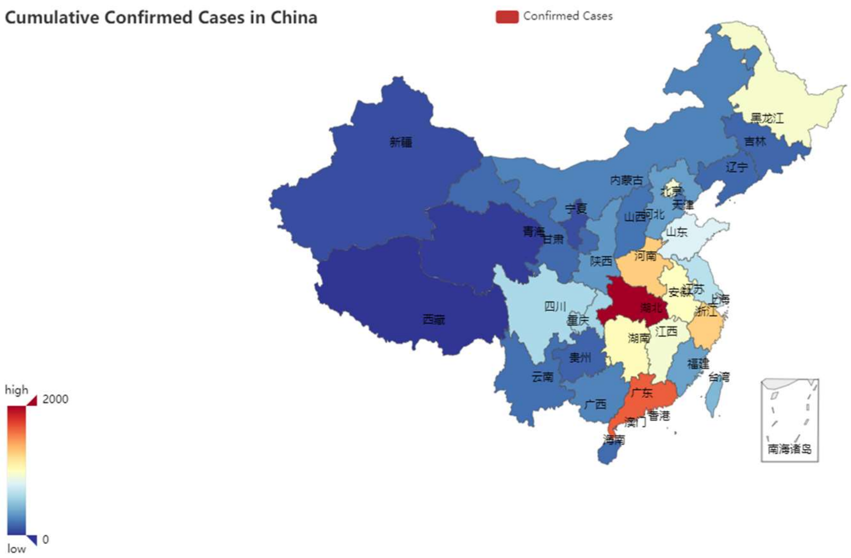
(6) marks: areas
channels: polar coordinate position, color, area
visual variables: 2D plane, color, area
I map province (categorical attribute) to the polar coordinate position and color, province risk index (ordered attribute) to the area (all the sectors' central angle is the same, and only the radius shows the data size).

5. Justify the choice of visualization

(1) For different provinces, I arrange them on China map based on its geospatial location. It is selective, allowing us to isolate a specific province immediately.

The visual variable color with a cool to warm color map can serve as the visual variable value. It is associative, allowing us to isolate the provinces that have similar cumulative confirmed cases immediately. It is ordered so we can have an intuitive comparison between different provinces.

For different provinces, I arrange them on China map based on its geospatial location, and I map the cumulative confirmed cases to the color variable via a color map. Then I fill each geospatial region by the color. Here is the visualization:



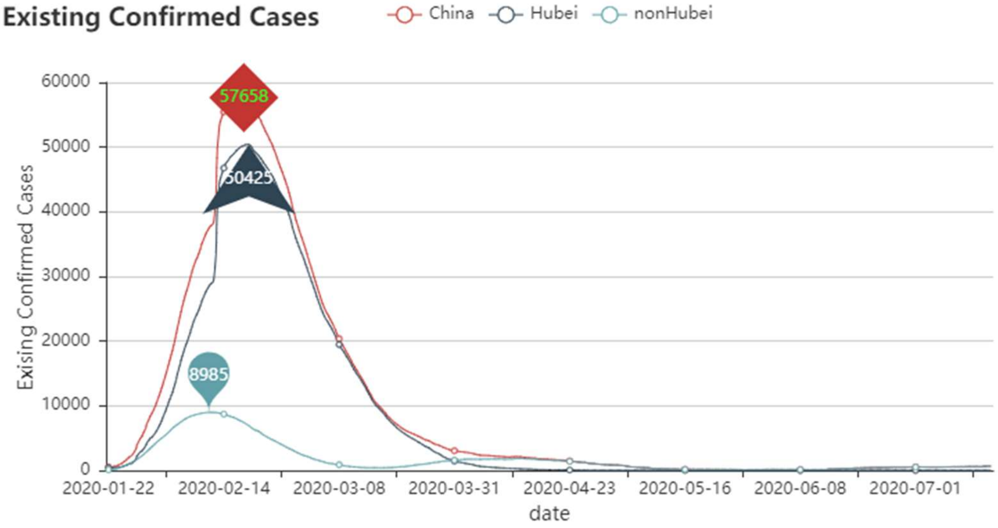
(2) The horizontal position is selective, so we can select a date to see the existing confirmed cases on that day. And since the vertical position and color are associative, we can even group the cases in China, Hubei, and outside Hubei with the same date (mapped to the horizontal position variable) since these items are differentiated by the number of cases and regions.

The vertical position is selective, so we select one existing confirmed case number to see what dates the cases are above the number.

The horizontal position is ordered, so it can be used to encode the dates.

The vertical position is ordered and quantitative so that it can encode the number of cases. We can compare them, and a numerical ratio can immediately express the visual distance.

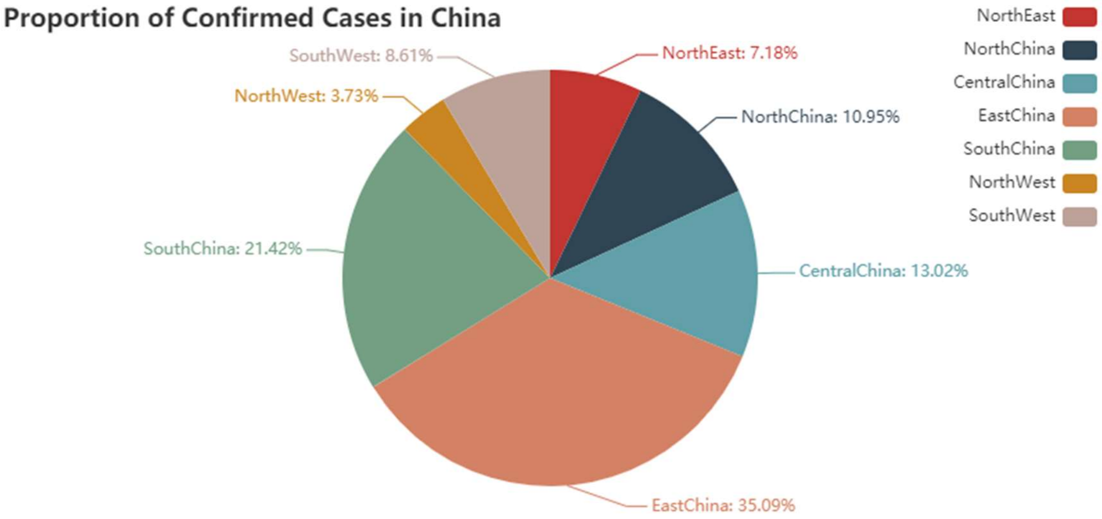
We use the line chart to visualize the everyday existing confirmed cases in China, in and outside Hubei. Here is the visualization:

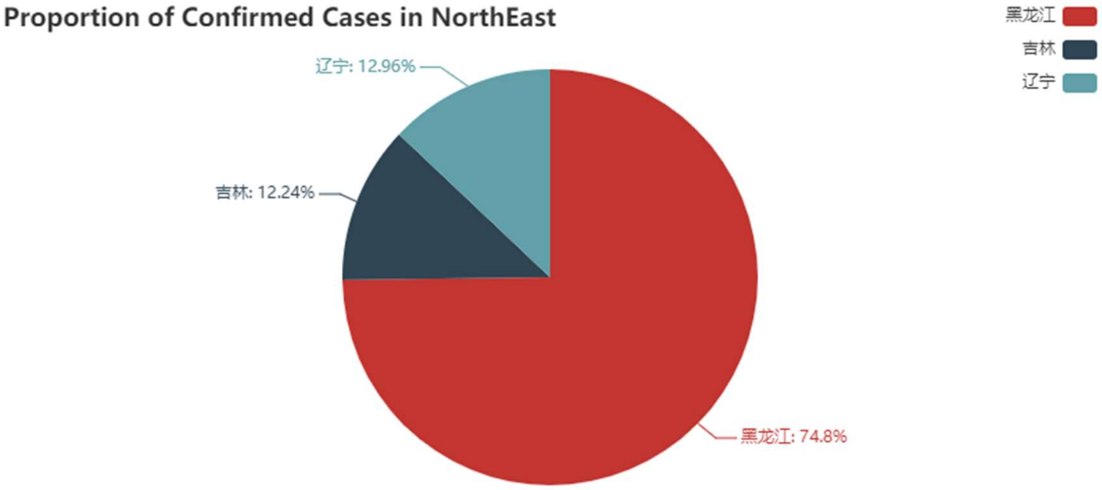


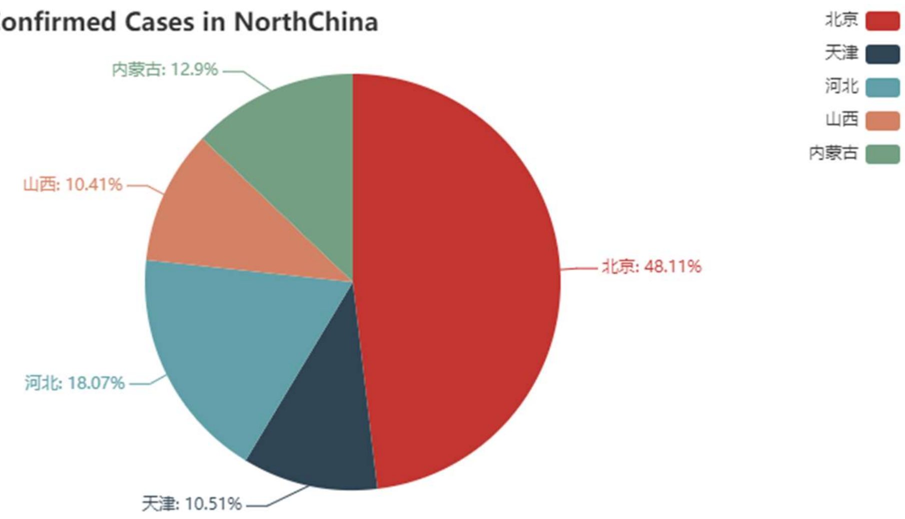
(3) The central angle is ordered and quantitative, so that it can encode the propotions. We can compare them, and a numerical ratio can immediately express the visual distance.

The color is is selective, allowing us to isolate a specific geographic region or province immediately.

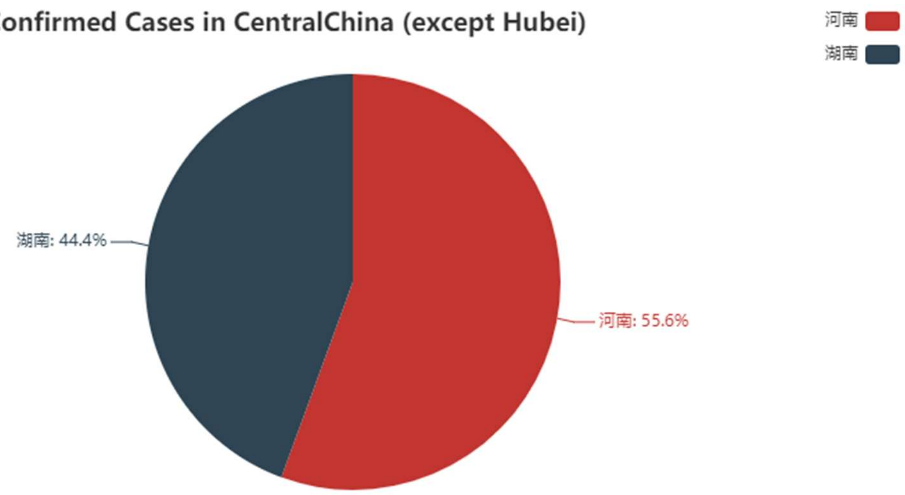
I use pie charts to visualize (a) different geographical areas' confirmed cases proportions in China's total confirmed cases, and (b) different provinces' confirmed cases proportions in one geographical area's total confirmed cases. Here is the visualization:



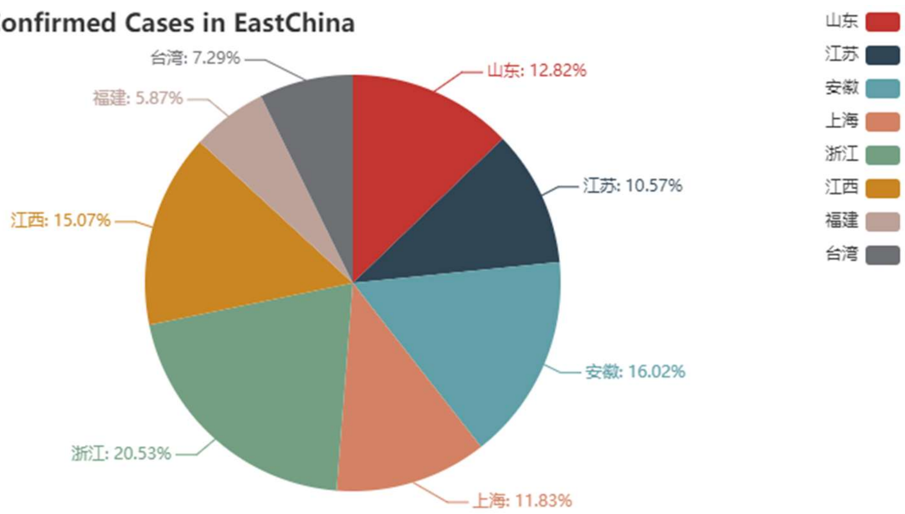




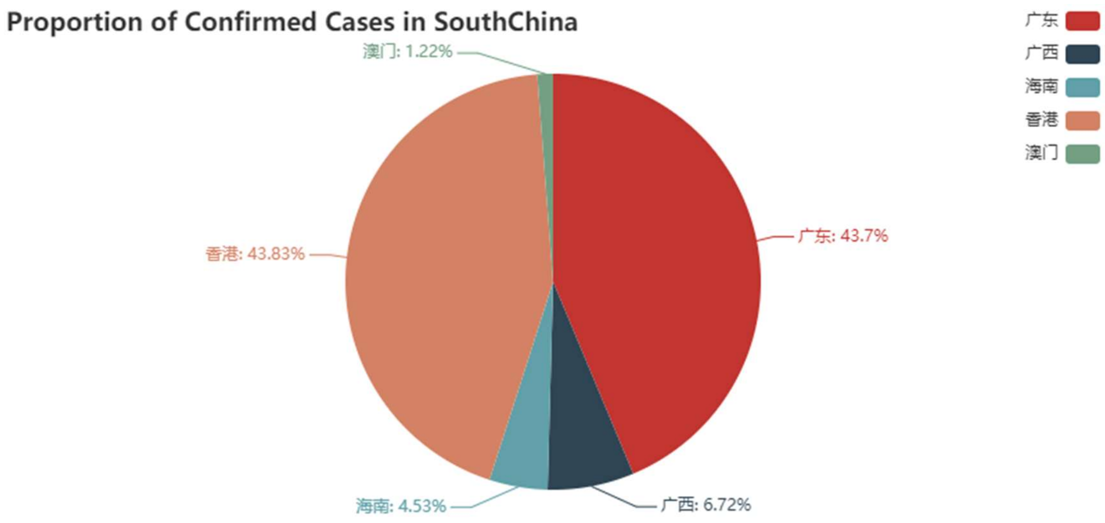
Proportion of Confirmed Cases in NorthChina

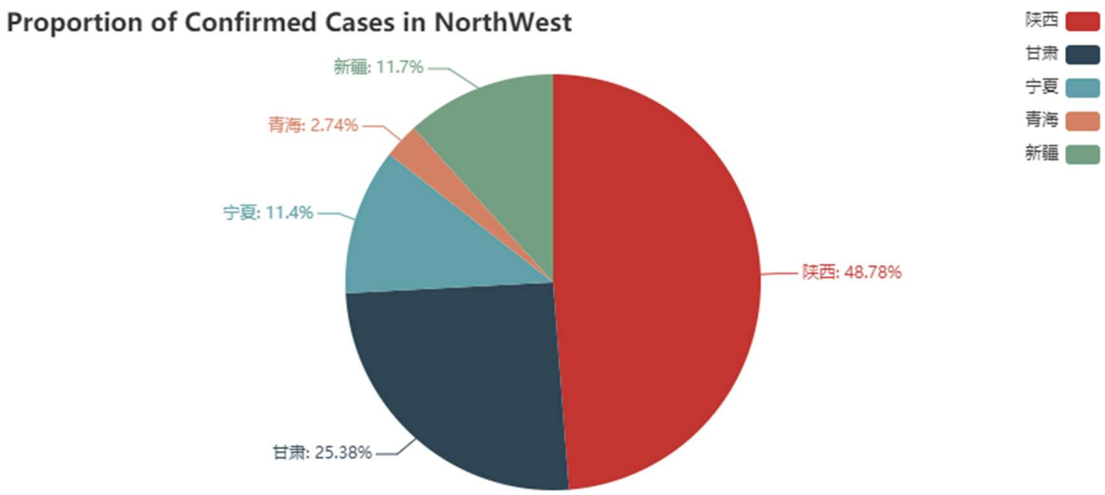


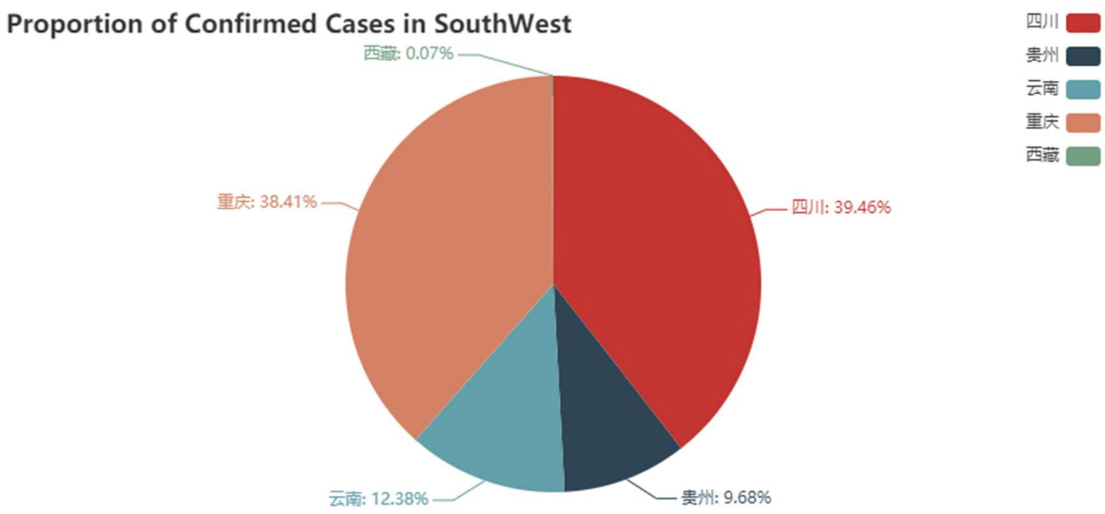
Proportion of Confirmed Cases in CentralChina (except Hubei)



Proportion of Confirmed Cases in EastChina







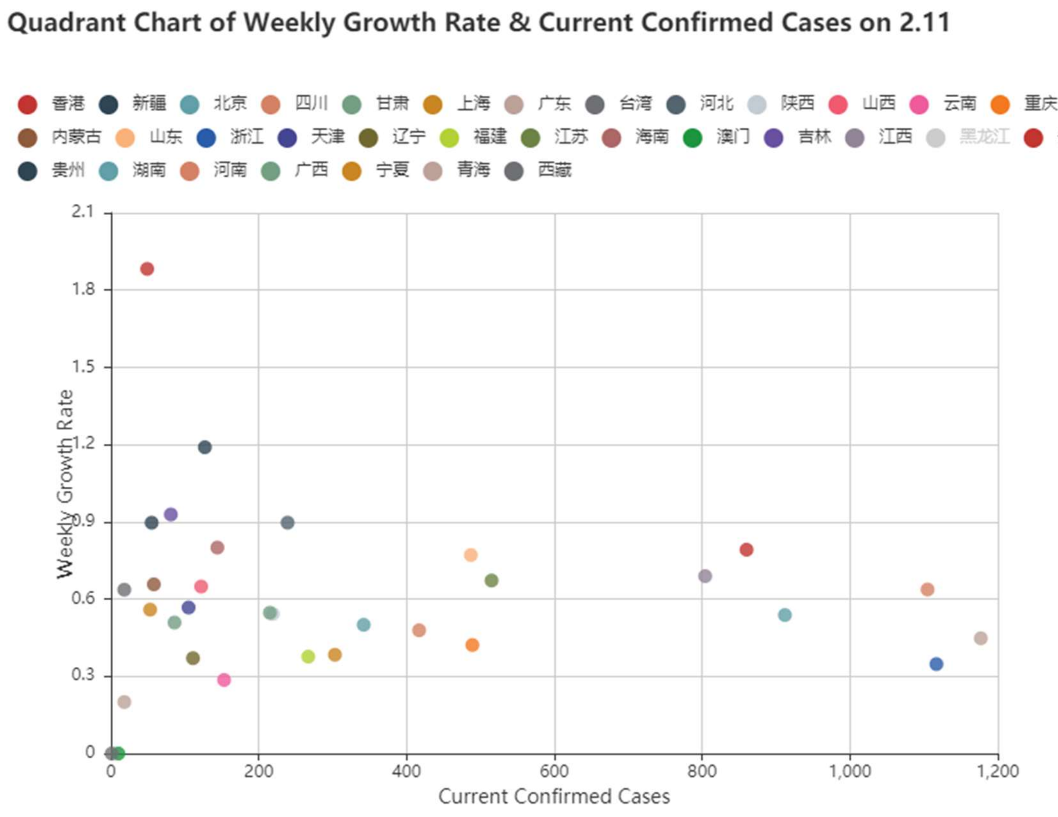
(4) The horizontal position is selective, so we can select a "current confirmed cases" range

to see some provinces' weekly growth rate. And since the vertical position and color are associative, we can even group the provinces with similar "current confirmed cases" since these items are differentiated by weekly growth rate and provinces. We can have a similar analysis of vertical position's selectivity and horizontal position and color's associativity.

The horizontal and vertical positions are both ordered and quantitative. We can compare different provinces’ confirmed cases and weekly growth rate. Besides, a numerical ratio can immediately express the visual distance.

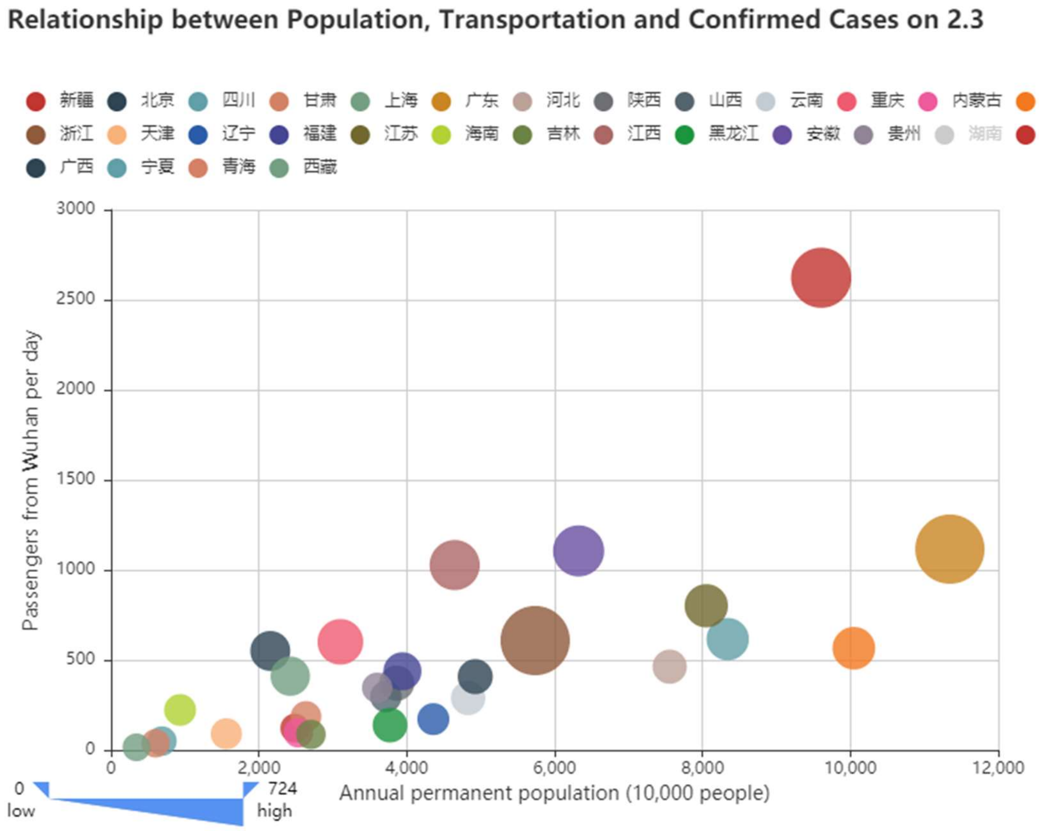
The color is selective, so we can select a specific province and see its current confirmed cases and weekly growth rate.

I use a scatter plot to visualize the regional epidemic risk in terms of importance and urgency. Here is the visualization:



(5) Besides the analysis made in (4), I want to comment on the circle size. Since the circle size is ordered and quantitative, we can use it to encode the confirmed cases. We can compare them, and a numerical ratio can immediately express the visual distance.

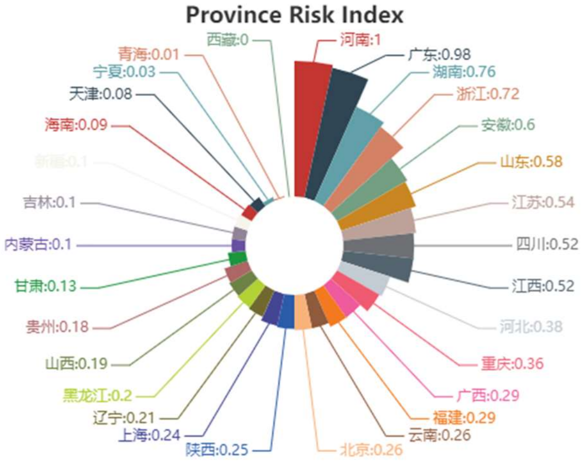
I use a scatter plot to visualize the relationship between population, transportation, and confirmed cases. Here is the visualization:


Relationship between Population, Transportation and Confirmed Cases on 2.3

(6) The polar coordinate position and color are both selective, so we can use either one to select the province that we are interested in.

The radius is ordered and quantitative so that it can encode province risk index. We can compare them among different provinces, and a numerical ratio can immediately express the visual distance.

I use a rose map to visualize the province risk index. Here is the visualization:


Province Risk Index