STAT 479 Project Milestone 4: Final Report

Group 8: Yuan Cao, Shikun Liu, Xiangyu Wang, Zijin Wang, Runze You

1. Introduction

China is the country with the largest population in the world. While the distribution of population is always playing a vital role in many regional social development fields such as climate and economy, it is also affected by these fields. Because of that, we are planning to implement a series of visualization tools to convert the relationship into easily understood images. We hope our visualization will give a clear thought of the reasons and the patterns behind some social phenomena to researchers who are interested. Our project goal is to exploit the relationship between regional population difference and some economic indicators in China, and find how they impact regional development. If we can conclude and quantify some relationships between population distribution and some social development features, our goal is achieved.

2. Steps and Methods

Our statistical data are collected from the *National Bureau of Statistics of China*. We extract population data and several socioeconomic indicators related to population, such as regional GDP, electricity consumption and consumer price index. We then transform the data into tidy format with each row to be a province and each column is related to an economic attribute. Given that our data is annual from 2001 to 2019, we include an extra column to mark which year the data belongs to exactly.

We obtain our spatial data with geometric features of provinces on *DataV platform*. Then, we translate provinces names for integrating statistics in visualization and use *Mapshaper* website converting format to TopoJSON, in order to apply Vega-lite for interactivity.

After data curation, we develop interactive choropleth population map and time series plot to visualize the variables of provinces in China in the period 2001-2019. Then we implement factor analysis to find and visualize some latent variables, which evaluate the development of a province in China. Besides, to further reveal the relationship among provinces with similar properties, we conduct hierarchical clustering. Details about our analysis are demonstrated below.

3. Interactive Choropleth Map

In order to perform some exploratory analysis about population and economic indexes we chose, a time series plot for changes of corresponding variable and an interactive choropleth map are established. The economic index data of Taiwan, Hong Kong and Macau are not included for now, so they are not shown. To avoid screenshots showing interaction taking up too much space, details about this map and some findings are contained in our *Observable notebook*.

4. Factor Analysis

Factor analysis mainly deals with the problem which is not clear or precise from the data. Here, we want to get some latent index from the data, which evaluate the development for a province or city of China. In fact, there is no kind of column in the data which could describe the development of that region directly. However, instinctively, the data of population, agriculture and so on, consist of the information of the development of the represented region.

Furthermore, in the past few years, Chinese government always used GDP, which is "product" variable in our data, as the only index to evaluate a province or city's developing situation. Our goal here is to give some more scientific indexes to help government evaluate the development of selected region.

Hence, we use factor analysis (FA) to derive that kind of indexes, called factor in FA, which is a kind of transformation of the original data. We visualize the consistence of factors among each variables, which are population, agriculture and so on, which called loadings in FA.

The left following plot shows the correlation between variables of Beijing with the right plot shows that two factors seems to be enough.

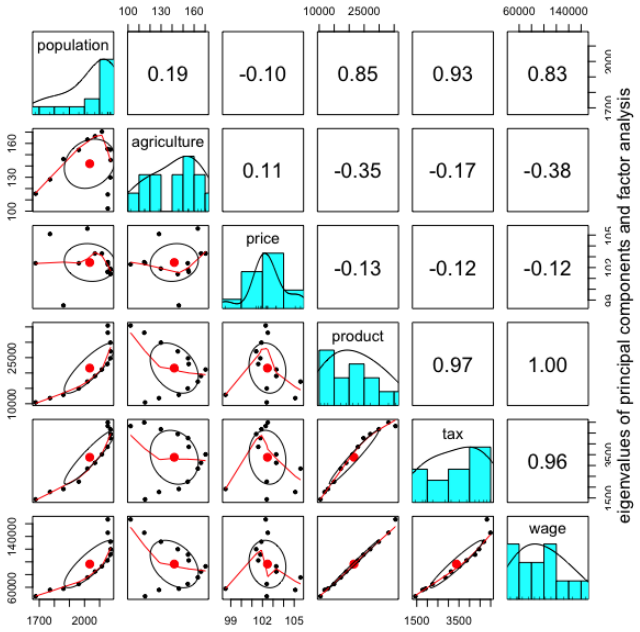

Figure 1: Co-variance Matrix plot

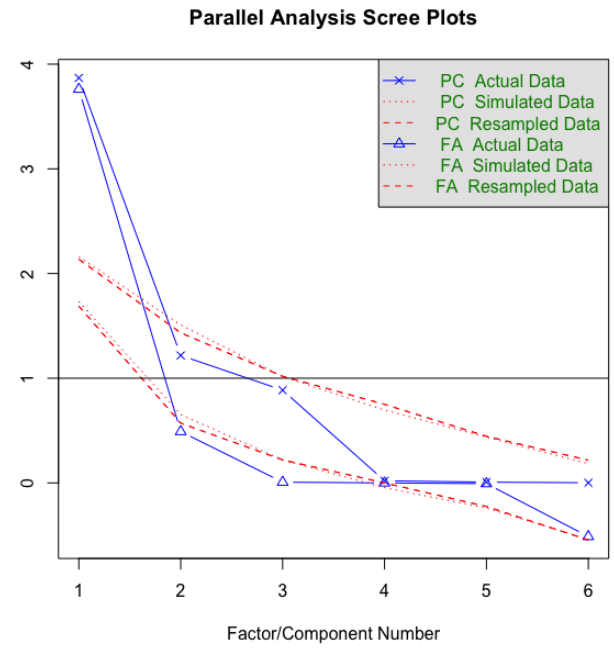

Figure 2: Needed factor numbers

We use the data after 2007, since before 2007, the collection of the data is missing a lot and not precise enough. Because of the limitation of the report, we will only show the visualization of derived factors for Beijing (Figure 3) and Guangdong (Figure 4). Besides, the reason we choose Beijing and Guangdong is that in the next cluster part, Guangdong and Beijing should be very different. Some notes of two figures for understanding are described below:

- 1 and 2 are the latent factors evaluating the development.
- In our data, price: CPI, agriculture: agriculture output, product: GDP
- Red line means negative correlation while green one means positive correlation.
- The thicker the line, the greater the influence is for this variable.

In fact, as some newspapers mentioned, for Beijing, a traditional city, Factor 1 could be called the Agricultural Support level Index and Factor 2 could be called the Consumption Index. For Guangdong, which is a newly developing province in the recent years, Factor 1 could be Light Industry Index.

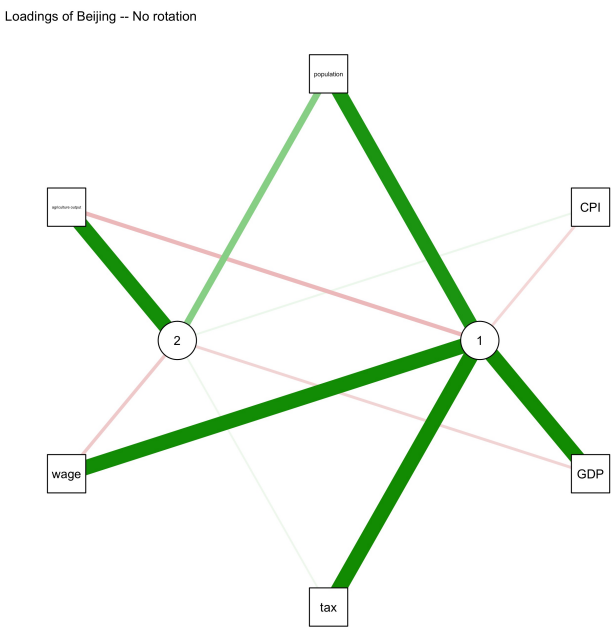

Figure 3: Two Factors in Beijing data

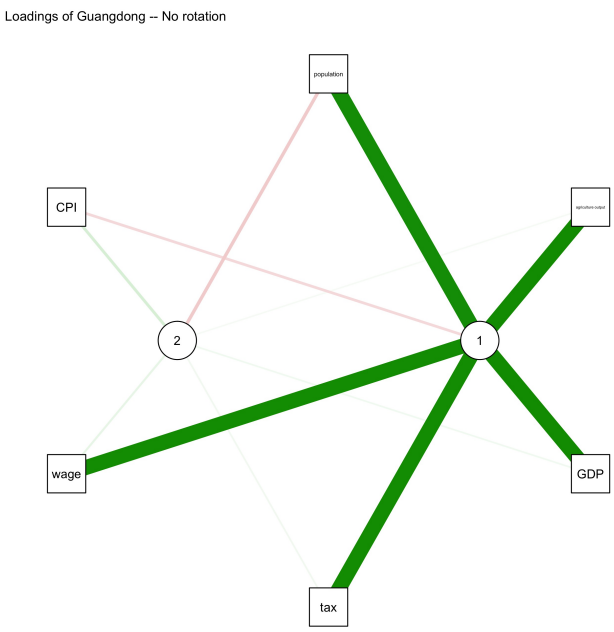

Figure 4: Two Factors in Guangdong data

In our opinion, this kind of visualization could be a useful and meaningful way to help public realize government's new complex evaluating indexes.

5. Hierarchical Clustering

From the former two steps of analysis, we realise that some provinces look pretty similar. And in order to exploit this assumption we perform a hierarchical clustering shown in Figure 5 where nearby provinces are the same cluster. Particularly, we choose the latest data we have for now which is the data of year 2019. From the visualization we can easily tell how similar two cities are across different scales which give us a solid proof of the results in the former two steps of analysis. We think such clusters will give some references between similar provinces when they make some policies as they are similar in economic conditions.

For example, cosmopolitans like Beijing and Shanghai are similar even they have fundamental and geographical differences. While in common sense (in China), Chongqing and Sichuan are very similar in every way, the results are totally opposite. We notice that Sichuan has good hydro-power resources

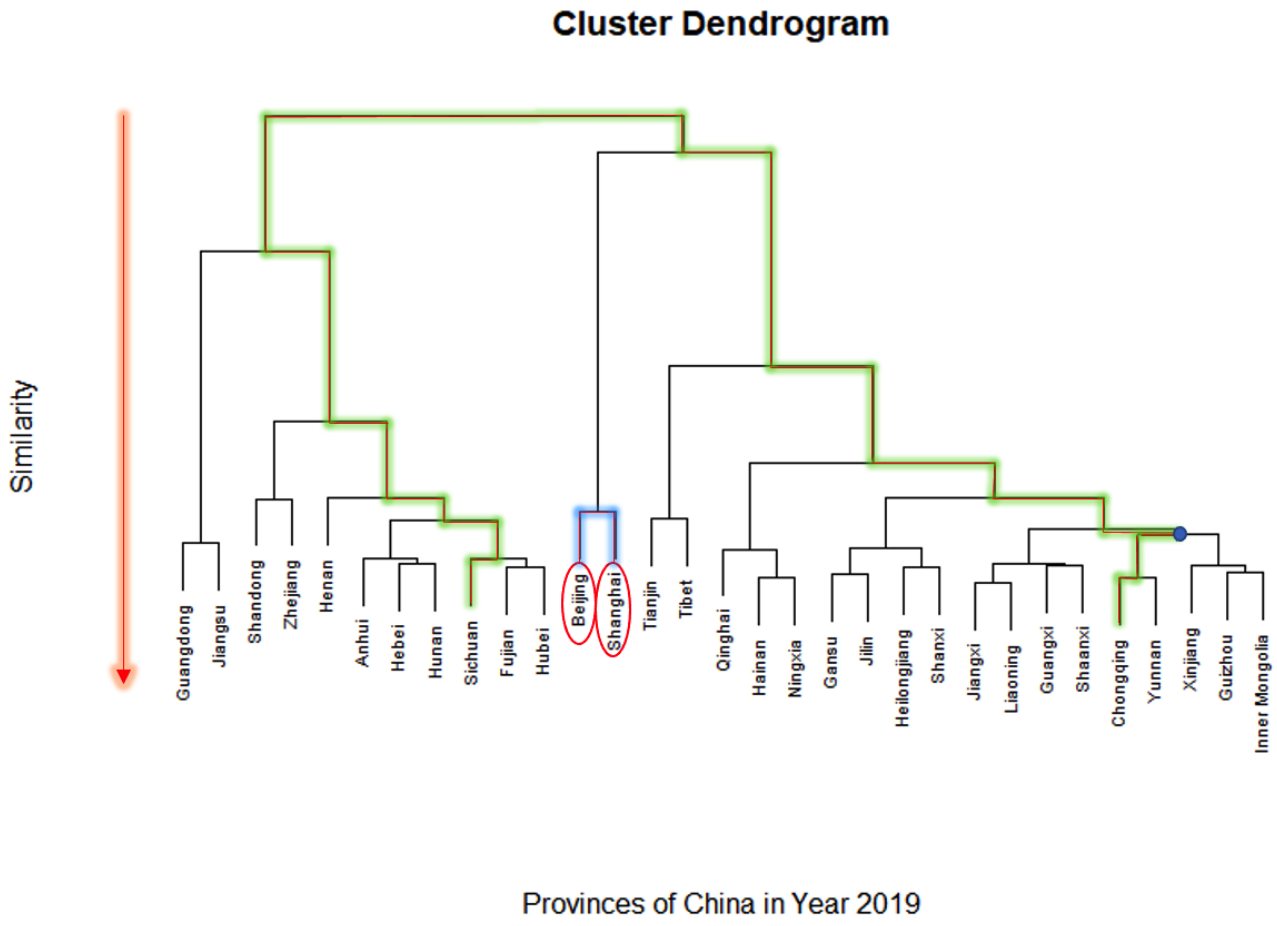

Figure 5: Hierarchical Clustering

which mainly affects the outcomes.

6. Future Prospective

There are still some drawbacks in our visualization. First, we got some comments from peer review about adding the name of province to corresponding position in the map instead of hovering to show the name, because it is difficult for people who are not familiar with China to locate certain province. However, we find it hard to add names as text to the map and hope to find a clearer way to show name of province on the map for people who are not familiar with China.

The second problem is that our factor analysis and clustering are static and only use a subset of data. We hope that these visualizations will be established with a Shiny App so that dynamic query is available.

Users can select the province and year they are interested in to check whether there are more potential relationships, or whether the cluster and latent variables of provinces change over time.