STAT 479 Project Milestone 3: Initial Report

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

1. Introduction

Our project goal is to exploit what impact on earth the distribution of population have in the provinces of China. While the distribution of population is always playing a vital role in many social development fields such as climate and economy. Because of that, we are planning to implement a series of visualization tools to convert the distribution of population associated with social development features 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. If we can conclude and quantify some relationship 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 map and implement factor analysis to find and visualize some latent variables, which evaluate the development of a province in China. Besides, to fur-

ther reveal the relationship among provinces with similar properties, we conduct hierarchical clustering. Details about our analysis are demonstrated below.

3. Interactive Choropleth Map

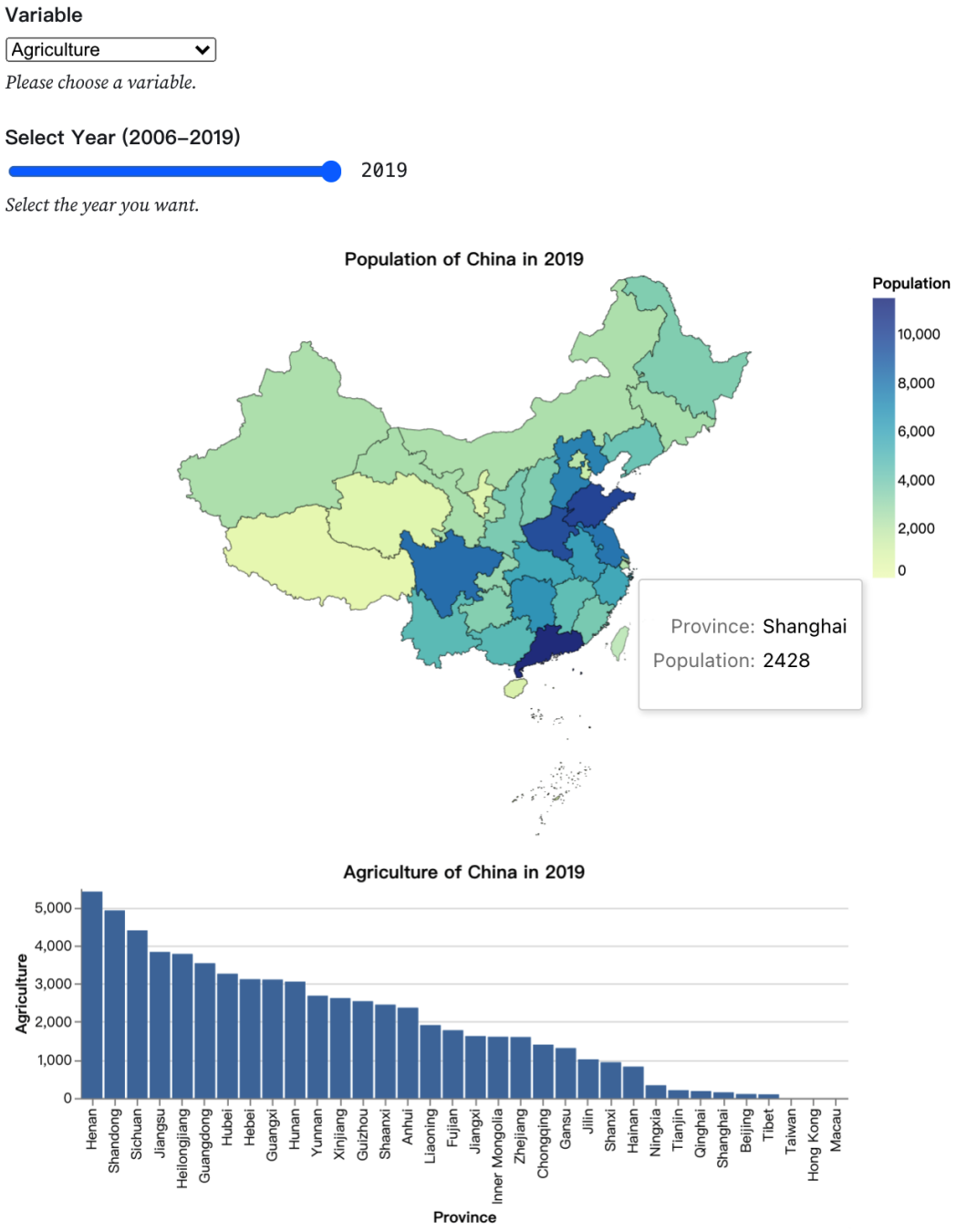

Figure 1: Choropleth Map

In order to perform some exploratory analysis about population and economic indexes we chose, an interactive choropleth map (see Figure 1) was established. The economic index data of Taiwan, Hong Kong and Macau are not included for now, so they are not shown in Figure 1. There are three parts in this visualization, top is the input part for dynamic query, including "Variable" and "Select Year", middle is our choropleth map showing population of each province, and the bottom is a sorted bar plot for corresponding variable.

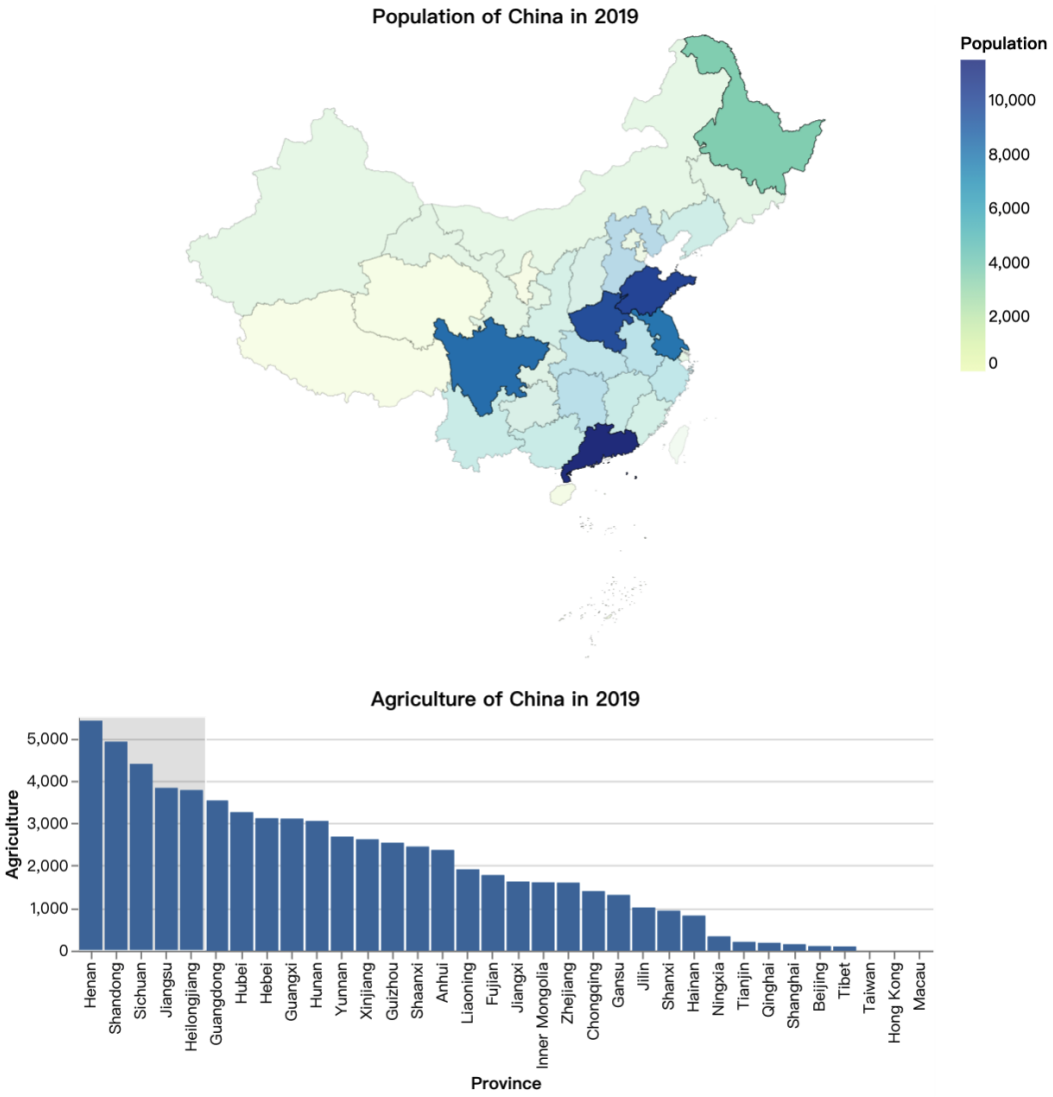

Figure 2: Dynamic Linking (query part omitted)

As for interactivity, there are two ways to interact within this map. One is the dynamic query with

selection of economic indexes and slider of year, when users specify their interested variable and year, the choropleth map and bar plot will change accordingly. The other is a dynamic linking between bar plot and the map shown in Figure 2, when users brush to select some provinces, those selected provinces will be highlighted.

To sum up, this visualization gives users a tool to query socioeconomic indicators of interest over different time periods and make some basic, simple and initial judgment intuitively about how these indicators changed over time and how they related to population.

4. Factor Analysis

Factor analysis mainly deals with the problem which is not clear or precise from the dataset. Here, we want to get some latent index from the dataset, which evaluate the development for a province or city of China.

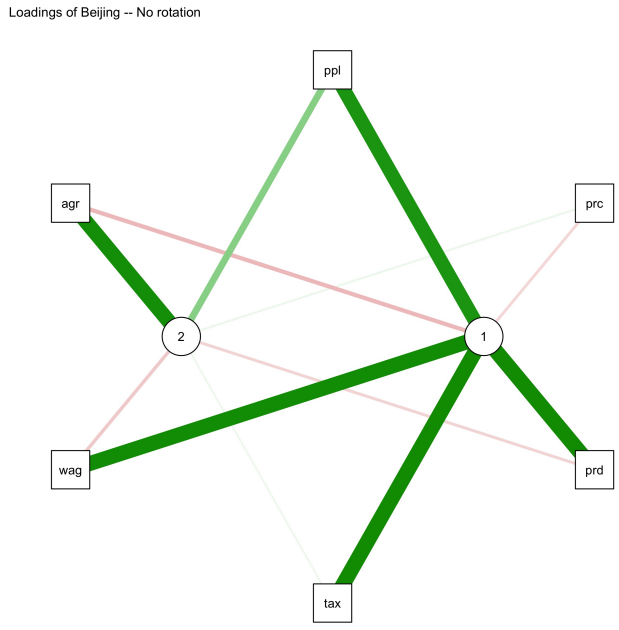

Figure 3: Two Factors in Beijing data

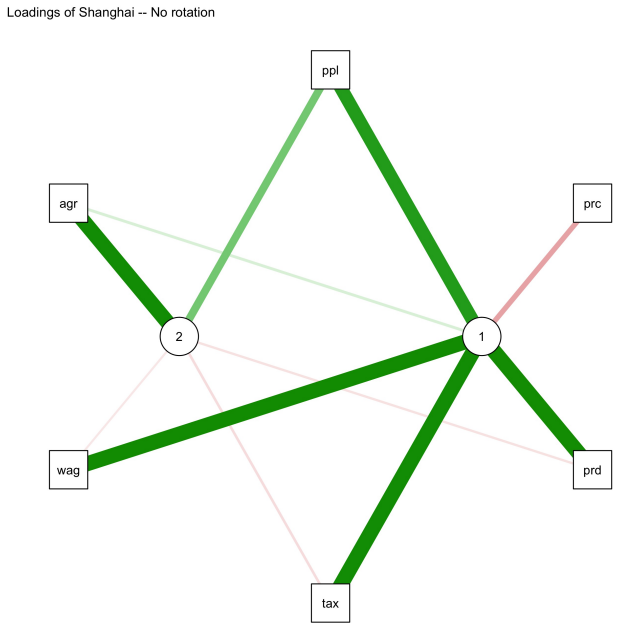

Figure 4: Two Factors in Shanghai data

In fact, there is no kind of column in the dataset 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. Hence, we use factor analysis (FA) to derive that kind of index, 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.

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 of Beijing (Figure 3) and Shanghai (Figure 4), and ignore the co-variance plot, which shows FA is meaningful, and the PCA plot, which conclude we need to choose two factors among data. Some notes of two figures for understanding are described below:

- 1 and 2 are the latent factors evaluating the development.
- agr = agriculture output; prc = CPI; prd = GDP; wag = wage; tax = tax.
- Red line means negative correlation while green one means positive correlation.
- The thicker the line, the greater the influence is for this variable.

5. Cluster

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 economic conditions.

6. Future Prospective

We know that the application scenarios of PCA include research on economic benefits, the level of population growth, etc. Thus, besides factor analysis, we may also want to try PCA in the next step.

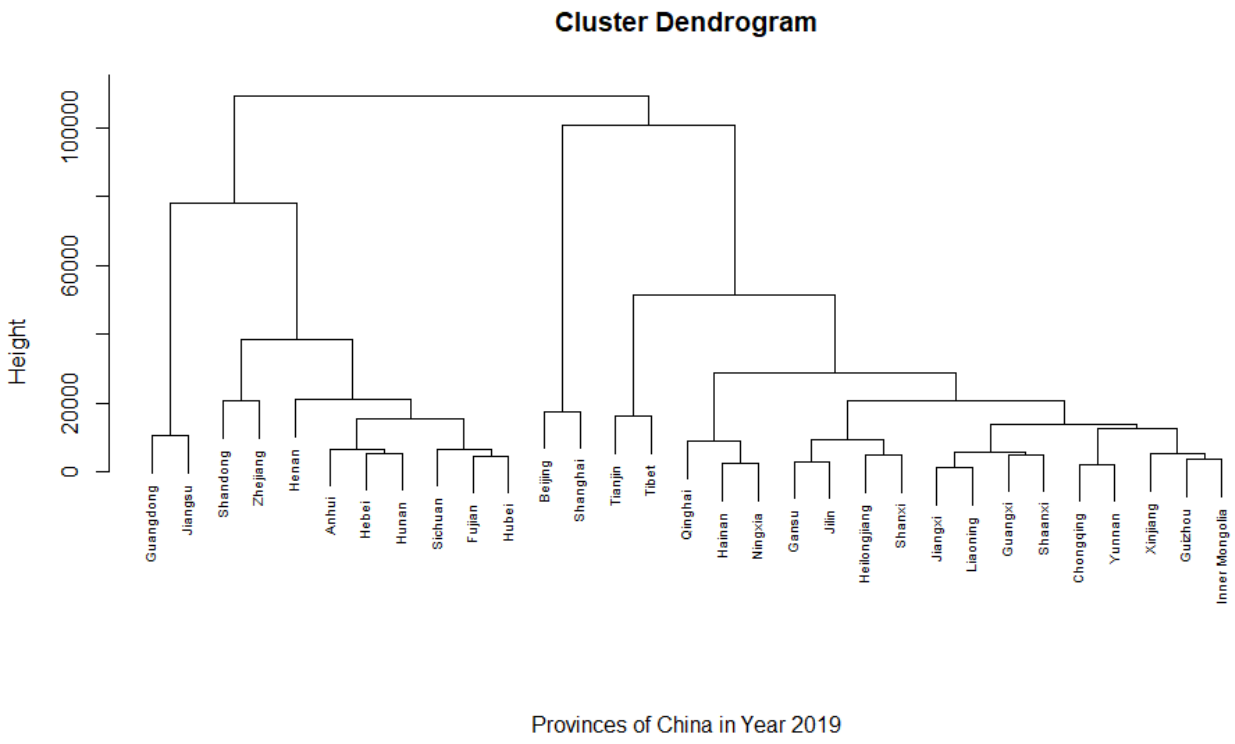

Figure 5: Hierarchical Clustering

In addition, in order to observe the trend of simultaneous changes in the population and specific indicator of each province from 2006 to 2019 more intuitively, we may consider drawing a line chart, from which we can select the province and indicator to be observed. Furthermore, we will work for detailed explanation for the result of hierarchical clustering and factor analysis such as whether similar provinces share the same latent variables and how to define these latent variables.