Final Paper

Alice Mee Seon Chung and Ningyin Xu

I. Introduction

Since China has been one of the key players of world economy, many studies have examined the economic development and income level of China. Income level is directly related to the economic power so it is important to know how it will change in the future at various levels like local, community, family and individual level. In this project, our goal is to study what factors affect the level of individual income in China. It is commonly believed that the socioeconomic backgrounds and characteristics of residential community have large impact on the income level of individual. To extend existing findings, we proposed the research question on what factors impact personal income and how they impact.

Our main data source is China Family Panel Studies (CFPS), the first large-scale academically oriented longitudinal survey data in China. We used data for 2 periods, with 33,018 observations in 2010 and 37,147 observations in 2014. After cleaning the missing variables, selecting the interested variables and merging two data by personal id, we ended up with 12,237 observations and 202 variables. We focused on the important variables associated with personal income: age, gender, years of education and other variables on community characteristics: the number of convenience stores, the number of primary schools, the number of hospitals, water source, community population, travel time to nearest town, travel time to county seat, agricultural GDP, non-agricultural GDP and income per capita.

Using these variables, we can measure averaged personal income by each province and further inspect the relationship of it with individual and community characteristics. First, we examined the overall features of these variables to get a general sense of the data. Then, we ran a multivariate regression with the variables and examine the prediction of model.

II. Graphical Methods

To answer our research question, the selection of data visualization methods is significant and the most important part of our research to deliver meaningful information of the data. As we have province identifier in the dataset, we were able to look at personal income by each province. Based on the characteristics of our data, we used two different types of graph: geospatial and statistical. Technically, we mainly used the graphic package "ggplot2" in r and plot function in package "lmtest".

Since we also focused on what factors have impact on the income level, we need to investigate the relationship between the dependent variable, income, and the independent variables. We examined the interested variables considering their specific characteristics and type, then chose the effective graphical method to show its distribution and its relationship with personal income.

First, our main dependent variable is personal income so we started to explore the relationship between the average personal income and GDP in each province.

Province

In the first plot, "Mean of Personal Income vs. GDP", we used scatter plot with "geom_point" to show the relationship of two variables, average personal income and GDP of each province in China. Here, we intended to provide the overall trend and relationship of these two variables without giving any information on which dot is corresponding to which province in China. We used color to indicate year and also used title, x-axis, y-axis, legend, and captions to convey the purpose and information to interpret the plot. From this plot we can see that mean of personal income and GDP has positive relationship in each province and both variables increased in 2014.

From the plot above, one might be curious about exactly how much mean personal income increased from 2010 to 2014 across China. We used geospatial visualization to show this using China map. Since R studio does not have data to draw China map with provincial geospatial data, longitude and latitude, we referred to the *ggplot2-china-map* in github repository (https://github.com/xiaohk/ggplot2-china-map) to draw each province in the map.

For the second plot, "Difference of Mean Personal Income between 2014 and 2010 in China", we used "geom_polygon" and "geom_path" to draw the whole China map and set the fill in "geom_polygon" and used "scale_fill_gradient" to show the difference scale with color in ggplot2. We used color scale to indicate the level of difference, so the darker color means larger difference in mean personal income. To make a consistent color palette, we selected the turquoise color and also used title and captions to deliver the additional information like, grey area means there are no available income information in our data. From this plot we can see the overall distribution of difference in mean personal income and perceive which area has the highest difference.

Further, in the third and fourth plot, "Difference of Mean Personal Income between 2014 and 2010 in East

China" and "Difference of Mean Personal Income between 2014 and 2010 in Midwest China", we divided Chinese provinces into two groups, east and midwest, to see closely which area has become richer than the other in four years. We used subset of provinces in China and manually set the scale to be the same across the three geospatial plots to make it easy to compare. Using same color and scale bound, we can observe that east region in China has become richer than midwest region in China in the past four years.

Next, we moved to the distribution of interested variables that we stated in the first section and the relationship between those and personal income. These variables are also categorized as individual level and community level. We first looked at the variables on the individual level and then moved to the variables related with the characteristics of residential communities. Here, we used statistical visualization method using "ggplot2" and applied specific graphical functions considering the type of variable. For every variable, we first looked at the distribution and then examined the relationship between the variable and personal income. For basic setting of the graphs, we used consistent color palette used in the previous plots and the same colors to indicate the same year as well.

Individual Level

Age

For the distribution of age, we used "geom_line" with stat setting "density". We can easily see how the distribution looks like and compare data across years using colors. For the relationship, we used "geom_smooth" and spanned it on the top of the "geom_point" scatter plot. From the plot, we can see that until age 30, personal income seems to increase with age, and after that, there is a decreasing trend.

Age-squared

For "age-squares", a variable that economists claim to have impact on personal earnings, we used the same "ggplot2" method as for the age variable since the characteristics of these two variables are similar. To see the relationship, we chose to draw "geom_smooth" with method "lm" and plotted it above the scatter plot using "geom_point". The relationship between age-squared and personal income seems to be negative from the plot, which could be because the "lm" method we use.

Gender

To show the distribution of gender, we used "geom_bar" with position as dodge since this is a categorical variable. It is useful to use bar chart to convey information and show the distribution when the variable is categorical. For the relationship, we used dodged "geom_boxplot" to compare the relationship between the gender and personal income. From the plot, we can compare between genders, and males seem to have a higher level of personal income than female in both years.

Education Years

We used “geom_line” with stat setting as density since education years is a continuous numeric variable. The overall distribution looks similar within two years, but the fluctuation of density in 2010 is larger than 2014. In the part of showing the relationship with income, we chose to draw “geom_smooth” with method as “lm” and span the lines above the scatter plot using “geom_point”. From the plot, we can see that as education years get higher, personal income also gets higher. So we can observe the positive relationship between two variables and the increasing trend is steeper in 2014 than 2010.

Residential Community Level

The Number of Convenience Stores in Communities

For the distribution, we used “geom_line” with stat setting as density since the number of convenience stores is a discrete numeric variable. The overall distribution looks a lot similar within the two years and this variable has large variance. For the relationship, we chose to draw “geom_smooth” with model as “lm”, linear model, and spanned it above the scatter plot using “geom_point”. From the plot, we can see that as the number of convenience stores gets larger, personal income also gets larger. And this relationship is stronger in 2010 than 2014 from the plot. However, it seems that for 2014, there are a large amount of data points falling near 0 number of convenience store, this could be an issue with the data.

The Number of Primary Schools in Communities

To see the distribution, we used dodged “geom_bar” because the number of primary schools is a discrete variable with small range and the difference of count between those numbers is large. In this case, using bar chart is more suitable to show the distribution. For showing the relationship, we used “geom_boxplot” with “dodge” position to compare the relationship between the two variables and resized the box width to indicate the group size. From the plot, there is a positive relationship in 2010, while we can not tell there is any strong relationship in 2014.

The Number of Hospitals in Communities

For the distribution, we used “geom_line” with stat setting “density” since the number of hospitals is discrete. The distribution is similar in both 2010 and 2014. For the relationship, we used “geom_boxplot” with “dodge” position to compare the relationship between the two variables, because there is a small number of different numbers of hospitals. From the plot, overall we can see there is a positive relationship between the number of hospitals and personal income in both years.

Types of Water Sources in Communities

We used dodged “geom_bar” since types of water sources are categorical. For the same reason, we again used “geom_boxplot” with “dodge” position to compare the relationship between these two variables and resized the box width to indicate the size of group as we did with the number of primary schools. From the plot, well spring water and tap water seem to have relationship with personal income in 2010 and tap water and other sources seem to have positive relationship with personal income in 2014.

Population in Communities

We used “geom_line” with stat setting as density. The overall distributions look alike within years, although there is slightly higher density in smaller population in 2014. This variable also has large variance. We chose to use “geom_smooth” with method as “lm” over the scatter plot to show relationship of two variables. From the plot, we can see that as the population gets larger, personal income also gets larger. So we can see the positive relationship between two variables in general.

Travel Time to Nearest Town in Communities

For the distribution, travel time to nearest town is a discrete variable so we used “geom_line” with stat setting as density. In 2010, the density is condensed around 1 to 2 hours. For the relationship, we chose to draw “geom_smooth” with method as “lm” and spanned it above the scatter plot using geom_point. From plot, we can see there is negative relationship between two variables and 2014 has higher upper bound on travel time than in 2010.

Travel Time to County Seat in Communities

For presenting the distribution, travel time to county seat is a discrete variable, so we decided to use “geom_line” with stat setting as density to show the distribution. The density is centered at less than 5 hours in both years. For the relationship, we used “geom_smooth” with method as “lm” and spanned it above the scatter plot using geom_point as we did with previous variable. We can see there is negative relationship between two variables in 2010 while the line in 2014 seems like a constant in the plot.

Agricultural GDP in Communities

For the distribution, we again used “geom_line” with stat setting as density to present the distribution. We can see how the distribution looks like and compare the data with two years by colors and this variable also has large variance. To present the relationship of two variables, we used “geom_smooth” with the scatter plot created by “geom_point” since agricultural GDP in communities is discrete and hard to draw the relationship as line. From the plot, we can see several curves around GDP under 2500 so it is hard to tell if the relationship is explicitly positive or negative.

Non-Agricultural GDP in Communities

The type of non-agricultural GDP is discrete so we used “geom_line” with stat setting as density to show the distribution. We used color to distinguish two years and from the plot, we can see that the distribution of non-agricultural GDP in 2014 has large variance than in 2010. We draw “geom_quantile” with method as “rqss” and with four-quantiles interval to see the relationship. From the plot we could see, as the non-agricultural GDP gets larger, personal income also gets larger in all four quantiles and across the two years. Thus, there is a positive relationship between two variables.

Income per capita in Communities

As income per capita is discrete, we used “geom_line” with stat setting as density to show the distribution of this variable. In the plot, income per capita in 2010 shows higher density under 2500. Income per capita in 2014 has higher density when it is larger than approximately 5500. To draw the relationship, we used “geom_smooth” and scatter plot using “geom_point” in this part. We can observe that income per capita and personal income has positive relationship in general.

Regression

After we examined the characteristics of all interested variables, we could now have the sense that how and why we should include all these variables to see how personal income changes. With these variables, we ran the regression model with panel data and it is defined as below:

$$log(income) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 educationyear + \beta_4 numberofconveniencestore$$

$$+\beta_5 numberofprimaryschool + \beta_6 numberofhospitals + \beta_7 watersource + \beta_8 pupulation$$

$$+\beta_9 timecommittonearesttown + \beta_{10} timecommittocountyseat + \beta_{11} agriculturalGDP$$

$$+\beta_{12} Non - agriculturalGDP + \beta_{13} incomepercapita + \beta_{14} year$$

To show and validate the model, we plotted the correlation plot with selected numerical variable and coefficient plot from the model.

To validate the independence of variables, we decided to plot correlation plot. We used package corrplot to plot the correlation plot using continuous and discrete variables. We omitted the categorical variable when plotting the correlation plot since those are not suitable for testing the independence. Using the result of data frame of correlation scores within selected variables, we were able to draw correlation plot. The range is from -1 to 1 and scores closer to 1 or -1 means they are more correlated with each other. In the plot, we used the diverging color scale using two consistent colors that we kept using through the whole graphs. The lighter color indicates a less correlated relationship and the variables might be independent with each other. The correlation plot shows that our variables are independent enough to say they are not correlated so that we can further use those variables to run the model.

We used package "coefplot" to plot the coefficient of all variables. Coefficient plot is a useful graphical tool to compare the estimate of all coefficients and we can also show the estimate with confidence interval. We used 95% confidence interval in this plot. And we used consistent color palette on this plot and did not apply a variety of colors for every estimates to avoid the confusion. The graph itself can clearly deliver the information. Most of the variables except Year 2014, Tap water, and Number of hospitals in communities are distinct from 0, and some of them have a positive relationship with personal income.

III. Result and Discussion

From the work and graphs, the audience has thoroughly learned what factors, and how they impact personal income in China. The audience get a chance to examine the main dependent variable, personal income, and the factors from two aspects: geospatial and statistical. Using two approaches, we present the distribution of difference in personal income by province and the characteristics of each variables. From these visualizations, the audience can learn and understand the caveats of personal income in China. Additionally, we present the linear model using these variables. We also validate the variables using correlation plot and visualize the coefficients of model to help audience to understand. The audience could learn how the factors impact on personal income from this part.

Truthful

The visualization is truthful since we used actual survey-based data from China Family Panel Studies in 2010 and 2014, which is the first large-scale academically oriented longitudinal survey data in China. We tried to include all validate observations and cleaned the data using personal id to identify the unique observations across two years. We tried to add information using title, subtitle, and notation if we apply additional exclusion to draw graph or if there is anything needs to be clarified such as unit or color.

Functional

We believe this visualization is functional since we tried to use the visualization techniques considering the flow to understand our dataset and research question with several steps: starting from general concept to specific variables. We select the graphic types to support the characteristics of variables and to show findings effectively. We tried to arrange all the graphs in a logical way: 1) general scatter plot of personal income and GDP in China; 2) geographical map of the difference in personal income in China; 3) statistical graph to describe the distribution of each interested variable; 4) statistical graph to show the relationship between personal income and each interested variable; 5) statistical graph to validate and show the result of the linear regression model using all interested variable. The audience can follow this flow and understand the story. Moving forward, the audience keep accumulating the overall features of the variables to understand the findings. The graphs themselves are easy to understand as we added some texts and notations on the graphs.

Beautiful

We mainly used red and green as two main colors. As we tried to be consistent on using color palette in all graphs, we believe the graphs are clear, simple and very explicitly presenting the information to the audience. The main colors corresponding to each year, 2010 and 2014, so the color helps to interpret the graphs of variables. For presenting the difference in personal income growth in the maps, we used the diverging scale of color to indicate the importance. Similar diverging scale with two main colors is used in the correlation plot. The graphs are aesthetically pleasing, attractive and intriguing because of its simplicity and clarity.

Insightful

The graphs are insightful in the sense that they provide a handful of information and revealed the findings of our research question. The insight is built as the audience follows the logical chain of graphs and understand the meaning of graphs. We try to provide both “a-ha” moment and gradual process of examination of the story on research question using this logical chain. At first, scatter plot gives the curiosity of personal income growth in China from 2010 to 2014. Then, we show the personal income difference with China Map and additional regional comparison between East China and Midwest China. Next, we try to present the specific features of each variables and the relationship with personal income. From the insights that gained from previous graphs, the audience can understand and interpret the result of linear regression model in the final part.

Enlightening

Taken all together these four qualities and the findings of our research question, the visualization is enlightening to the researcher who wants to study of this dataset. Since this dataset contains lots of variables so one may feel lost when they first look at the data. We provided a way to present the information with logical chain and utilized several different graphical techniques. We applied geospatial visualization to present the income difference between two years instead of drawing a line graph. Instead of showing the result table of regression model, we visualize the result to make it easy to interpret. The story and conclusion from the visualization could provide insights to researchers or implications to the policy makers to boost personal income in future China.