**Temporal changes and spatial distribution of educational level for the US**

Group member: Xiangmei Zhang; Xiangzhen Li

## Introduction

As an important indicator for comprehensive national strength, educational level information is collected at each Censuses of Population. Educational level is not only the embodiment of the advanced nature, but also an important influencing factor of social stability. The temporal changes and spatial distribution of educational level are useful reference for investment, school district selection, and employment. However, it is not easy for a citizen to dig up this information and have a visual perception. Even the government website might provide raw dataset of native historical education level. Citizen without basic statistical training can only get limited information from raw data that government’s website provided. Besides, most of the raw data is huge and messy, it is nearly impossible for an untrained citizen to get information from. In this study, we aim to provide a convenient visualization tool to help people obtain and analysis the educational level for the US. The raw data is from the Census Bureau of US. Data technologies are needed to clean the data into a long tidy version, and check the completeness of the data.

The main deliverable products of this study will be an R package called “Edulevel”, and a corresponding shiny app for people who are not familiar of using R package. The package and the shiny app contains 25 years of education level of every county in the US. First, the package has basic function of searching a database, which could gather information we want. For example, the package and the app can provide the educational level of the specific time and county, if the county name and year is specified. What’s more, the package can provide more detailed information after we define the requirement. The second function of this product is to provide simple graphical tools to analyze the educational level spatially from year to year. This package can help the users to identify the temporal changes and spatial distribution of educational level for an area of interest. Besides, a user could not only have plot comparing the educational level along with year of a specific location, but also have an intuitive map with gradient color which indicate the different educational level. The map could be a US map which has distinguished colored state, or a state map with gradient color indicate different counties.

## Data clean

* 1. Data description

The data is collected from the website of US government open source database (<https://www.data.gov/)>. The dataset is an excel file that contains 3283 rows which indicate 3283 counties in US and 47 columns which are variables. Table 2.1 shows a part of the raw data. In the variables, seven of them are attributes of counties, 40 of them are educational level variables. The seven attributes are FIPS code, State, Area name (county), 2003 Rural-urban Continuum Code and so on. The FIPS code is unique for each county, so it could be kept helping users to select county in R. The “State” and “Area name” is kept helping users to select the specific area they are interested in. Since other four variables are not unique for each county and also not well known by users, they will not be used in the package or shiny app. The 40 variables contain four educational level which is “Less than a high school diploma”, “High school diploma only”, “Some college (1-3 years)”, “Four years of college or higher” and percentage of adults of these four levels. There are five data point for every educational level from 1970 to 2015 by every ten or five years. The 40 educational level variables will all be used and they will be classified as two types: count and percentage. Value of each variable means in the selected county at a specific time point, the number of people or the percentage of adults in a specific education level.

Table 2.1 Part of raw data



* 1. Data clean

The raw data might have problems for R to select information even it looks like clean. There might be missing data and wrong data points. The hybrid column name is combination of educational level and year which is not straightforward for R to select either educational level or time point. The word description for the same educational level is not exactly the same. Besides, the format of the dataset is not convenient for users. So, several steps are needed to make the dataset tidy before the application stage.

First step is to revise the dataset into a tidy format. R Function “gather” is used to gather columns 8 which is first column of educational level variables to the last column into new column “values”, and previous column name which indicate the educational variables into new column “education”. Other seven attributes columns do not change. We also use R function “separate” to separate cells under Column “education” into Column “level” and “year”. For example, first row of Column “education” is “Less than a high school diploma, 1970”; it is separated into two columns: Column “level” which has value of “Less than a high school diploma” and Column “year” which has value of “1970”. In the same time, four columns of not useful attributes are deleted using R Function “select”. Then, we add a Column “type” with R Function “mutate” to distinguish if the value is a count of people or percentage of adults with an educational level. Thus, we already have an initial format of dataset. However, there are still several simple changes needed. R Function “parse\_number” is used to drop the non-numeric characters before or after the first number in Column “FIPS”. R Function “trimws” is used to remove the whitespace in Column “year”. Table 2.2 is the first six rows of dataset with initial format.

Table 2.2 The first six rows of initial format



To have a general knowledge of the dataset. We use R Function “table” to look at the educational level. It turns out the names of the same educational level are not consistent. R Function “group\_by” and “tally” could make the same point clearer. For example, in 1970 and 1980 year, the highest educational level is “Four years of college or higher”, the highest educational level in 2000 is “Bachelor's degree or higher”. These two descriptions should be classified as the same educational level, however. To matching the names of the same educational level, we use R Function “gsub”. Function “gsub” is used to select key word and replace the description of variables in Column “level” with the same appropriate names. After several attempts to try to describe all the levels with regular expressions, the educational levels are combined as four levels: “Bachelor's degree or higher”, “Some college (1-3 years)”, “High school diploma only”, “Less than a high school diploma”.

On the other hand, the correctness and completeness of the data is needed to be checked. R Function “spread” is used to spread the “value” by “type”, which means spread the value of type “count” and “percentage” into two columns. For one specific FIPS code (county) and a year, the percentage for each level is the corresponding count divided by the total count. To check if the percentage variable is calculated correctly, the R Function “mutate” is used to add a Column “new\_percentage” which is calculated from variable “count”. Check if the variable “new\_percentage” is the same with the variable “percentage” and select these rows that are not the same with function “filter”. It turns out there are seven rows with percentage calculated incorrectly. Use R Function “mutate” to correct the variable “percentage”. Last, R Function “gather” is used to get back to tidy version of dataset (as shown in Table 2.3). Until this step, the dataset is tidy and correct to conduct further analyze.

Table 2.3 First six rows of tidy data



## R package “Edulevel”

The main deliverable product of this project is an R package and a shiny app. The package is used for people who have experience on R and grasp the skills to apply simple R functions. The package called “Edulevel”, which contains one dataset and four functions. The dataset is the tidy data names “tidy” from Section 2 (showed in Table 2.3 as an example) which contains information of four educational level for adults (more than 25 years old) in US from 1970 to 2015. The dataset is stored using “LazyData”. Function "showcounty" will return a list of county names for a given state abbreviation, which will be used in other functions, and it also provide reference for users when using other functions. Function "lineEd" will plot a line chart ("count" or "percent") of all (four levels) education levels for a given county. Function "mapstate" will give a map of the "count" or "percent" data on the scale of states in US for a given education level. Function "mapcounty" will give a map of the "count" or "percent" data on the scale of counties for a given state and a given education level.

When using function “showcounty”, the user need to specify the two-letter abbreviated state name that is interested in. Since the state name is character, it has to be quoted in double quotes. For example, if we are interested in Iowa, the command and output is:

*> showcounty("IA")*

*> [1] Adair County Adams County Allamakee County Appanoose County*

*[5] Audubon County Benton County Black Hawk County Boone County*

*[9] Bremer County Buchanan County Buena Vista County Butler County*

*[13] Calhoun County Carroll County Cass County Cedar County*

*[17] Cerro Gordo County Cherokee County Chickasaw County Clarke County*

*…*

*[97] "Woodbury County" "Worth County" "Wright County"*

*99 Levels:*

Function “lineEd” is used when one is interested in one specific county and want to analyze the trend educational level in the past years. The input of the function is the abbreviated state name and county name. The state name and county name should be quoted in double quotes. The county name has to be exactly the same with the county name in the dataset “tidy”. If the user is not clear with the spell of county name, function “showcounty” could help. The output of the function is a plot with year as the x-axis and value (count or percentage) as the y-axis. The data is group by four educational level with four different color and facet by value type (count or percentage). For example, to find the educational level in past years of Story county in Iowa, we could use following command:

*> lineEd("IA", "Story County")*

If the input is “Story” instead of “Story County”, R will give an error message.

Figure 3.1 is the output of the above command. We could see from the plot that the people with Bachelor’s degree or higher (red line) is increasing rapidly over the years. The percentage of adults with a Bachelor’s degree or higher is increased by about 20% percentage in 35 years. The people who has been to college is also increasing. Meanwhile, the people who has high school diploma or lower level is decreasing. The figure tells us the education level in Story county is stepping into a good direction. It seems that without considering the education progress in US, Story county is a pretty good place to stay if parents are emphasis on children’s education.

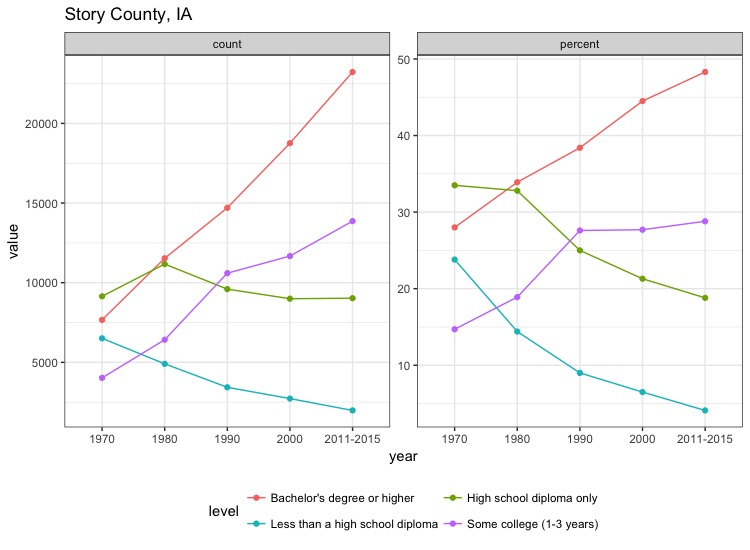


Figure 3.1 Educational level plot of Story county in Iowa

Actually, the education level of Story county is increasing rapidly. Figure 3.2 is the educational level of Liberty county in Florida. The people with high school diploma is increasing rapidly, meanwhile the people with Bachelor’s degree is increasing really slow. Also, there are other counties that people with high education level is decreasing and population that only been to high school increasing, such as Yukon-Koyukuk Census area in Alaska.

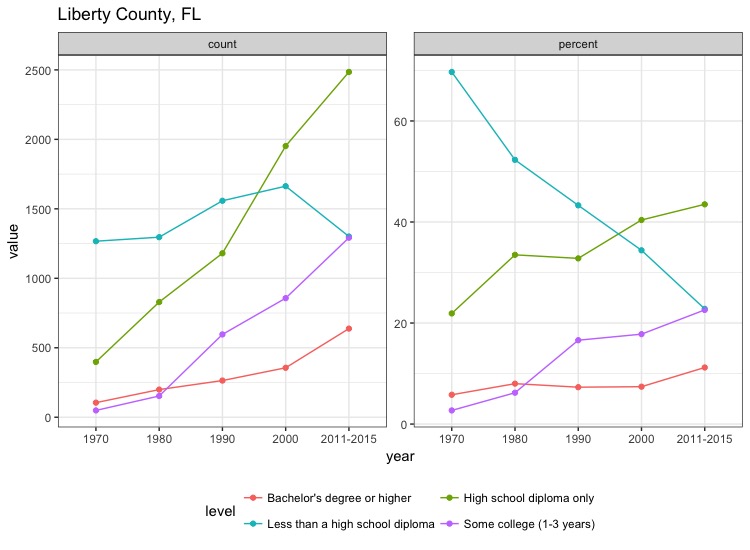


Figure 3.2 Educational level plot of Liberty county in Florida

Function “mapstate” provide a summary map of education status of US. The input parameters are “type” and “level”. Parameter “type” is either “count” or “percent”. There are four levels, with numeric number 1 to 4 to represent four educational levels from “Less than a high school diploma” to “Bachelor's degree or higher”. For example, the following command is asking for the layout of percentage of adults with Bachelor’s degree in US.

*> mapstate("percent", 4)*

Figure 3.3 is the output from R. It is obvious that the people with Bachlor’s degree is increasing. However, people with less than high school is decreasing obviously if we input command *mapstate("percent", 1)*.

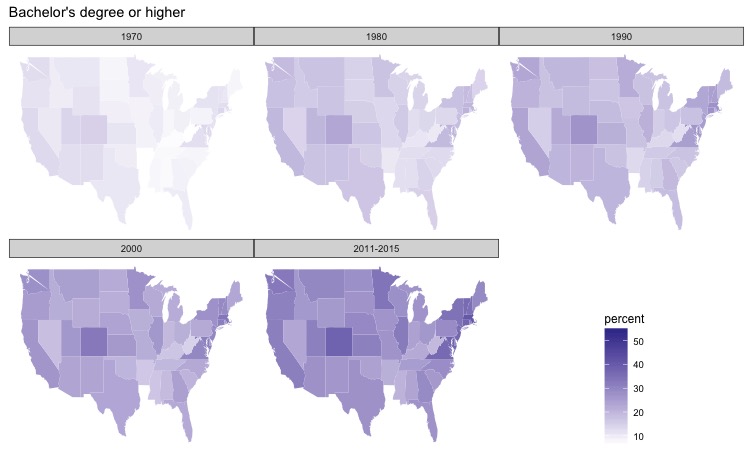


Figure 3.3 Percent of adults with Bachelor’s degree in US

Function “mapcounty” has similar output with “mapstate”, except the names are county rather than state. The input parameters are “type”, “level”, “stateshort” and “limit”. Parameters “type” and “level” have the same usage with “mapstate’. Parameter “stateshort” is the two-letter abbreviated state name. Parameter “limit” means, if a county who has a percentage greater than “limit”, the county name will be labeled; default “limit” is 28. For example, if we use following command to look for people with Bachelor's degree in Iowa, Figure 3.4 is the output. The labeled counties are adult percentage higher than 28%.

*> mapcounty("percent",levelint = 4,stateshort="IA")*

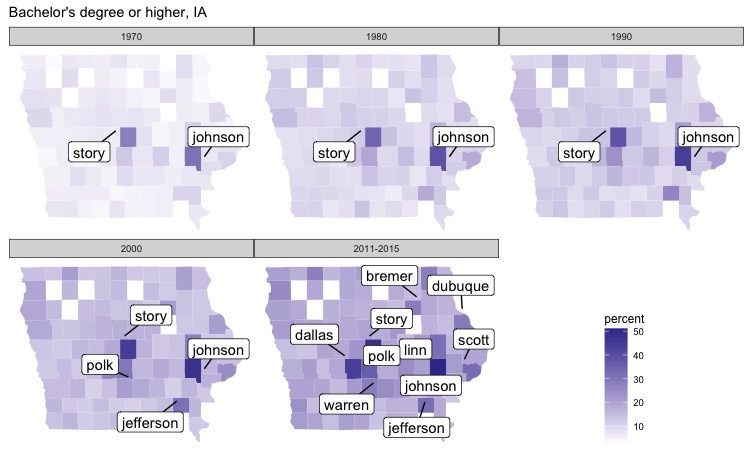


Figure 3.4 Percent of adults with Bachelor's degree in Iowa

## Shiny app (In Progress…)

For users who are not familiar with R, shiny app is a friendly tool. Also, the shiny app could provide more intuitive user interface. There are four group of widgets in shiny app total. First group is “lineEd” plot along with the text input of state name abbreviation and select box of county name. The “mapstate” widget group contains the map and radio button of “count” or “percent”. The “mapcounty” widget group contains the state map, the radio buttons of “type (count or percent)” and “level (1-4)”, select box of state name. There are another widget contains table with the selected information. This table has similar function as the dataset in R. The user could select the state, county and education level using checkbox group. There is another slider range for each type (count or percent) that is used to select the range of count or percent that the user is interested in.

## Conclusion

From this study, we apply the data technologies to conduct data clean and re-organization, draw analytical graphics, apply data onto maps, and make user interactive R package and shiny app. We have a more complete and deeper understanding of the technique we learned in class after this study. We also be more skilled in using R and R packages.

The R package “Edulevel” is more suitable for users who have experience in R. For the users who are skilled at R, they could change the source code, or do more complex analysis using raw data. For users who are not familiar with R, the shiny app is a better way for visualization, and it is easier and more interesting to use shiny app. The R package “Edulevel” can truly help people to find a good location with better education environment. What’s more, it could also show the development of technology in US and advancement of every state in education aspect.