SI 618 Project 1 Report

Using PySpark to Explore the educational expenditure

Sijun Tao

October 23, 2022

1 Motivation

Education spending is an investment that can promote economic growth, boost productivity, support individual and societal growth, and lessen social inequality. This expenditure includes spending on educational institutions like colleges, universities, and other public and private organizations that provide or support educational services. One of the important decisions that governments, businesses, students, and their families make is what percentage of overall financial resources should be allocated to education.

To explore how the education spending can really affect individual and societal growth, I want to explore if there is any relationship between education expenditure and some social indicators like education level, employment rate and unemployment rate. There might be interesting inner relationship between education expenditure and these indicators, which may helpful with the financial arrangements and scheduling of future governments.

In this project, I surrounded and explored the following three instructional questions:

- Does the amount of instruction expenditure effectively affect the education level?
  - Does the proportion of instruction expenditure make an influence on education level?
  - Is the higher education level related to the higher instruction expenditure?
- Does the amount of instruction expenditure affect infect the living standard of the local people (indicted by indictor “unemployment rate”)?
  - Does the proportion of instruction expenditure make an influence on unemployment rate?
  - Is the lower unemployment rate related to the higher instruction expenditure?
- Does the higher educational expenditure promote the employment rate?
  - Does a higher instruction expenditure ratio lead to higher employment rate in the following year?
  - Is the higher support service expenditure related to the higher employment rate?

2 Data Sources

In this project, I used two datasets from Kaggle. The first one is the US Educational Finances dataset, while the other one is USA unemployment & educational level datasets. Both datasets are in csv format.

2.1 US Educational Finances

The url link of this dataset is https://www.kaggle.com/datasets/noriuk/us-educational-finances.

The dataset contains revenues and expenditures for all U.S. states from 1992-2016. It contains 12 variables and 1,275 records. The dataset includes information on:

- state and the year
- total expenditure
- instruction expenditure
- support service expenditure

2.2 USA unemployment & educational level

The url link of this dataset is https://www.kaggle.com/datasets/valbauman/student-engagement-online-learning-supplement. The dataset consists of the unemployed and employed information and education level of adults in the USA by county from 2000-2020. The education level dataset has 3,283 records and 48 variables, while unemployment dataset has 3,275 records and 93 variables. The whole dataset includes information on:

- state and county name
- civilian labor force
- employed number
- unemployed number and unemployment rate
- Less than a high school diploma rate
- High school diploma only rate
- Some college or associate's degree rate
- Bachelor's degree or higher rate

3 Data manipulation

The overall workflow of this project is shown as figure 1. The tools used in each step is stated. For the data manipulation part, the work mainly focused on step 1-3. The code for data manipulation is in ***sijuntao_si618_project1_data_preprocess.py***.

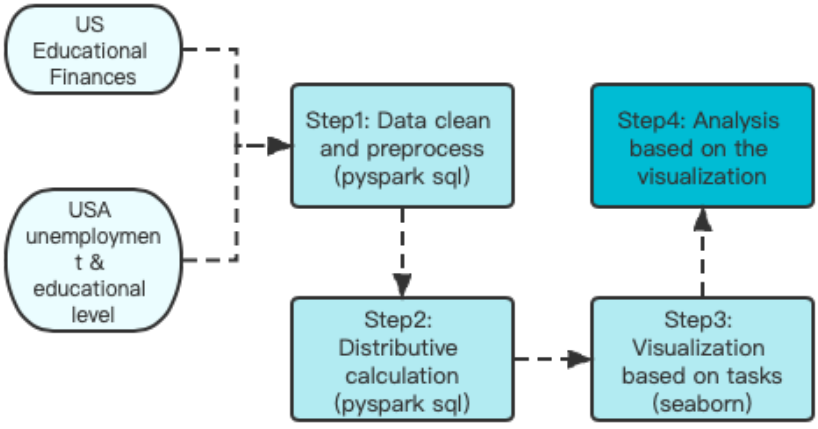

Figure 1: Project1 workflow.

3.1 Step 1: data cleaning and preprocessing

Firstly, for the datasets described before I selected the necessary variables that are required for the three questions. Then, I dropped the incomplete records. The US educational finances dataset has the information of year for each record, which is the key for further joining, so I did not do further preprocessing. For the unemployment and education level datasets, since the variables are the information in each year and did not contain the information of year itself, I added one more column to specify the year for the required information.

For the unemployment dataset, I selected the records in 1990, 2000, and 2015 for further joining with other datasets. I extracted the subtables for each year and then concatenated them together, formatting the original dataset into a new processed dataset with only 7 variables. For the education level dataset, I selected the records from 1992 to 2016 for further joining. Since it has similar format with the original data of unemployment dataset, I also extracted and concatenated to produce a new processed dataset for education level dataset with 8 variables.

All the manipulations in step 1 were accomplished by spark SQL.

3.2 Step 2: distributive calculation

For each main question I mentioned in “Motivation” section, I created two sub-questions. Using the preprocessed data obtained before, I generated six csv files for each sub-question.

The first two sub-questions which belong to question 1 required to join the expenditure data and the education level data together. But before joining, some extra manipulations should be conducted to get the same key for joining. For expenditure data, I use data in 1992 to simulate the data in 1990 for further joining. For education level data, since the state name are all abbreviations, I added a column with the full name by merging it with an extra dataset containing information of the state name and the corresponding abbreviations. After that, I grouped the data by state name and year to get the average education level for each state for each year. Finally, the two processed datasets were joined together by keys state and year. The first sub-question extracted state, year, instruction expenditure ratio, less than high school ratio, equal to high school diploma ratio, equal to college diploma ratio, greater than college diploma ratio from the joined dataset for further visualization, while the second sub-question extracted state, year, instruction expenditure amount, high education ratio.

The next two sub-questions which belong to question 2 required to join the expenditure data and the unemployment data together. Similar as above, before joining, two datasets were manipulated to fulfill the condition for future joining. Besides the manipulation mentioned above, I filtered both datasets to get data from 2000 to 2016 and then joined them together. The first sub-question extracted state, year, instruction expenditure ratio, unemployment ratio from the joined dataset for further visualization, while the second sub-question extracted state, year, instruction expenditure amount, and unemployment ratio.

The final two sub-questions which belong to question 3 required to join the expenditure data and the unemployment data together. Different from the joined dataset obtained in question 2, this time I used the support service expenditure, its corresponding ratio, and the employment rate. Except these, all the others are the same. The first sub-question extracted state, year, support service expenditure ratio, employment ratio from the joined dataset for further visualization, while the second sub-question extracted state, year, support service expenditure amount, and employment ratio.

3.3 Step 3: visualization based on the tasks

After generating the csv files using RDDs, I wrote a Python notebook to load the files and do plotting. The visualizations were created to demonstrate the relationship between various educational expenditure and education level, unemployment rate, employment rate. For each sub-question, one plot was created.

4 Analysis and Visualization

The csv files used in this part were generated using spark SQL in the above section. For each sub-question, one plot and the corresponding analysis are given below. The code for visualization is in ***sijuntao_si618_project1_data_visualization.ipynb.***

4.1 Instruction expenditure and education level

4.1.1 Does the proportion of instruction expenditure make an influence on education level?

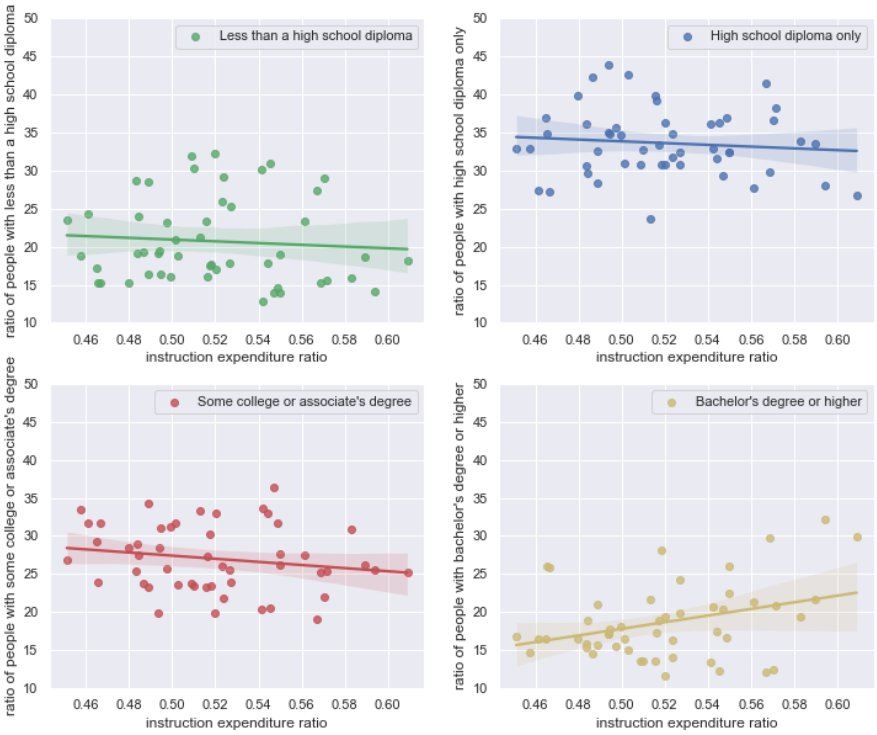

Figure 2: instruction expenditure ratio vs. education level s

I used reg plot to demonstrate the potential relationship between instruction expenditure ratio and the ratio for four different education level. The upper two are the ratio of people with less than a high school diploma and the ratio of people with high school diploma. The lower two are the ratio of people with some college or associate's degree and the ratio of people with bachelor's degree or higher. The visualization is shown in figure 2.

From figure 2, we can find that with the increasing of instruction expenditure ratio, the ratio of people with less than a high school diploma, high school diploma, some college or associate's degree decreases slightly and the ratio of people with bachelor's degree or higher increases. Since bachelor's degree or higher is highest education level among these four, we can see that the increasing of instruction expenditure could lead to higher education level in the United States.

4.1.2 Is the higher education level related to the higher instruction expenditure?

Consider some college or associate's degree and bachelor's degree or higher as high education level, I used scatter plot and reg plot to demonstrate the potential relationship between raw instruction expenditure amount and the high education level rate. Points with different color represents different years. The visualization is shown in figure 3.

From figure 3, we can find that with the increasing of instruction expenditure, the high education level rate increases as we expect. Besides, we can find that the points with darker color locate upper than points with light color, which shows that the overall high education level rate may increase by years. From these, we may give conclusion that there exists a positive relationship between instruction expenditure and education level.

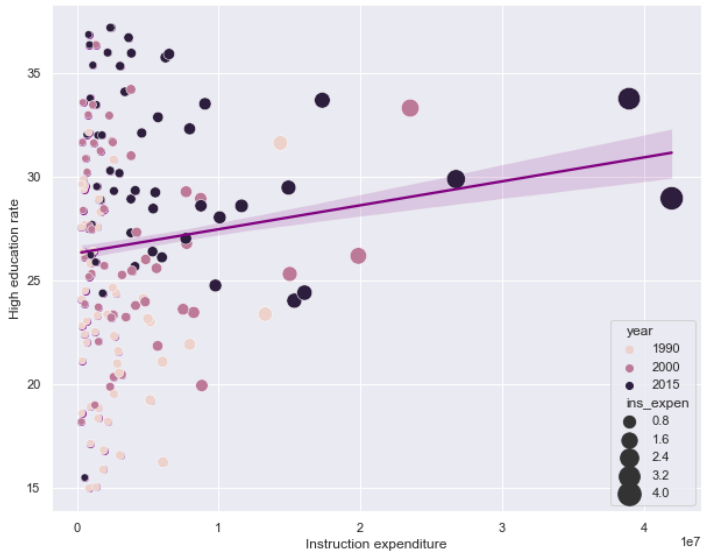

Figure 3. instruction expenditure vs. high education rate

4.2 Instruction expenditure and unemployment rate

4.2.1 Does the proportion of instruction expenditure make an influence on unemployment rate?

I used joint plot to demonstrate the potential relationship between instruction expenditure ratio and the unemployment rate in 2008. The distribution of instruction expenditure ratio and the distribution of unemployment rate are displayed on the top and left of the plot. The visualization is shown in figure 4.

From figure 4, we can find that with the increasing of instruction expenditure ratio, the unemployment rate has the tendency to decrease. For most states, the instruction expenditure ratio is around 0.52 which is more than 0.5 and the unemployment rate is 0.057. Although the fit line presents a negative

Figure 4. instruction expenditure ratio vs. unemployment rate

relationship between the two factors, it is not convincing enough since the points are too scattered to serve as evidence. Further exploration is needed to give conclusion on the possible influence of instruction expenditure ratio on unemployment rate.

4.2.2 Is the lower unemployment rate related to the higher instruction expenditure?

I used scatter plot and reg plot to demonstrate the potential relationship between raw instruction expenditure amount and the unemployment rate between 2010 and 2016. Points with different color represents different years. The points with larger instruction expenditure have larger size. The visualization is shown in figure 5.

From figure 5, we can find that with the increasing of instruction expenditure, the unemployment rate also increases. Besides, we can find that the points with darker color locate lower than points with light color, which shows that the overall unemployment rate may decrease by years. Compare with 4.2.1, we can find that the raw instruction expenditure and instruction expenditure ratio have exactly the opposite relationship with the unemployment rate, which may indicate that simply increasing the amount of instruction expenditure may not be a good idea, instead increase its ratio.

From these, we may give conclusion that there exists a positive relationship between instruction expenditure and unemployment rate.

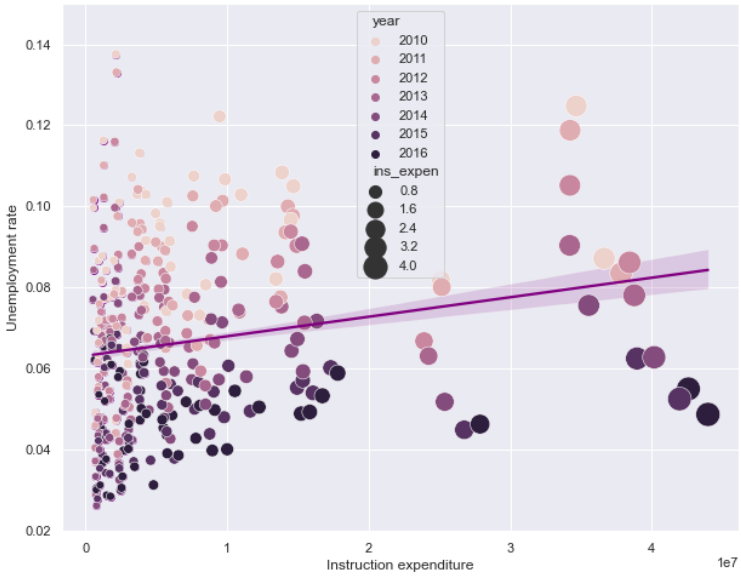

Figure 5: instruction expenditure vs. unemployment rate

4.3 Employment rate and expenditure

4.3.1 Does a higher instruction expenditure ratio lead to higher employment rate in the following year?

I used line plot to demonstrate the potential influence of instruction expenditure ratio on employment rate between 2010 and 2016. Green line represents the change of instruction expenditure ratio from 2010 to 2016, while blue line represents employment rate. The visualization is shown in figure 6.

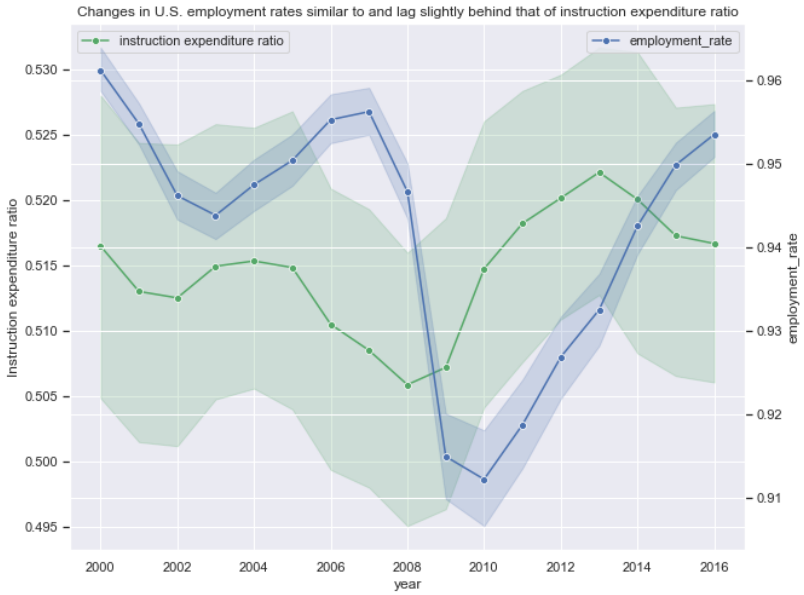

Figure 6: the change of instruction expenditure ratio and employment rate by years

From figure 6, we can find that with the changes in U.S. employment rates are similar to and lag slightly behind the changes of instruction expenditure ratio. For example, for 2000-2002, the instruction expenditure ratio decreases, then employment rate also decreases and decreases one more year from 2000 to 2003. After that, the instruction expenditure ratio increases between 2002 and 2004, then employment rate also increases and begins to increase a little bit later from 2003 to 2007. Similarly, the instruction expenditure ratio decreases after that and employment rate also decreases but is a little bit later than instruction expenditure ratio. From these we could give the analysis that the tendency of instruction expenditure ratio can affect the change of employment rate in the following years.

4.3.2 Is the higher support service expenditure related to the higher employment rate?

I used scatter plot and reg plot to demonstrate the potential relationship between raw support service expenditure amount and the employment rate between 2010 and 2016. Points with different color represents different years. The points with larger support service expenditure have larger size. The visualization is shown in figure 7.

From figure 7, we can find that with the increasing of support service expenditure, the employment rate decreases. Besides, we can find that the points with darker color locate higher than points with light color, which shows that the overall employment rate may increase by years. From these, we may give conclusion that there exists a negative relationship between raw support service expenditure and employment rate.

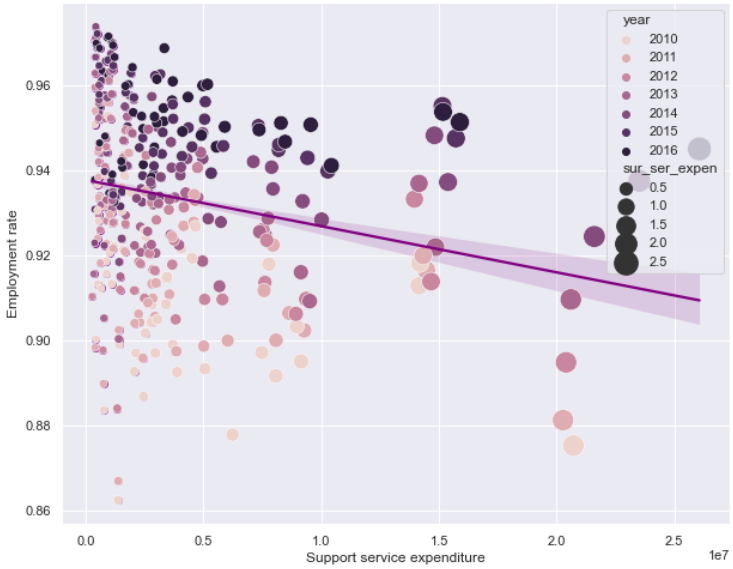

Figure 7: support service expenditure vs. employment rate

5 Challenges

In this project, I encountered several challenges. In the data cleaning and manipulation process, I found that the original column names are with ‘,’ and spaces, which makes it a bit of a hassle when using SQL to select and compute the columns. To fix it, I rename all the columns that do not meet the standard to avoid

the error. In the distributive calculation part, I found two challenges. First, the name of the state is not consistent in the datasets that need to join together. To solve this, I searched Google to find the state name and the corresponding abbreviations in dict format, then I transformed it into dataframe and saved it as csv file for future using. Second, when joining the expenditure dataset with the education level dataset, due to the fact that the expenditure dataset does not have data in 1990, I replaced the data in 1990 with data in 1992 since it is the closest data so it can represent the data in 1990 to some extent. However, it might lead to some uncertainty on my analysis. Third, when calculating the unemployment rate, I first directly use the average unemployment rate of each county as the unemployment rate of the whole state. However, I found that there may be a problem that this calculation method is inaccurate. To solve this, I changed the calculation method. I summed up the civilian labor force of the state and the unemployment number of the state then divide them. In this way, the annual unemployment rate of the state is accurate.