SI 618 Project Proposal

Zhe Yang

1. Summary and motivation

This project aims to figure out the relationship between the educational expenditures and the average household income in different states in the United States. Generally speaking, people believe that higher investment in education leads to a higher education level, and thus results in higher income. Hence, if a government wants to increase people's income and living standard, it may consider increasing the educational expenditures. However, does that work for every county or state? Are there some cases where the governments invest similar amount of capital but have different results? This is a key problem to research on which would help the government review the education spending plans and do some improvement.

On the other hand, is it possible that the areas with higher income tend to spend more on education? In other words, is it possible that it can become a virtuous circle that promotes education and income growth? That is also an interesting question to investigate.

2. Description of datasets

a) The first dataset contains the educational revenues and expenditures of the elementary and high schools in different school districts of different states from 1992 to 2016. The data are available on Kaggle: https://www.kaggle.com/noriuk/us-educational-finances

b) The second dataset contains the statistics (including mean, median,

standard deviation, etc.) of US household income in different counties of different states. The data are captured in 2017 and available on Kaggle: https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations

3. Manipulation and join of datasets

First, I will join the datasets on the state level. Because both datasets have the "state" column, it will be easy to implement. Second, I will try to join these two datasets on the county level. Since they are probably different in naming different counties, I will figure out a way to group the overlapped ones together and treat them as the same counties to join them. If the second join method succeed, the project can provide a more detailed analysis on the county level additional to the state level.

4. Proposed computation tasks

a) First, I will use mrjob and spark to perform a MapReduce job on the datasets, which maps all the educational expenses and household income to different states and compute the different statistics (including the total average educational expenses, total average household income, etc.) of each state. This procedure reduces the data to the state level.

b) Second, I will use sparksql to map the obtained statistics of the educational expenses and the household incomes for each state in order to discover the relation between these two groups of statistics. I expect a positive correlation in general but also different patterns of such correlation.

c) Third, I will split the datasets by years and use mrjob or spark to construct

another MapReduce job, which maps the previous obtained statistics to different years. In this way I can build a sort of time-series model to examine if there is a virtuous circle that promotes education and income growth.

5. Visualization

One of the visualizations will be a bar graph with lines connected each bars indicating the relationship between the amount of average household income and the amount of average educational expenditures. Hopefully it can provide a rough sense of the trend of the relation. The plots may be produced with matplotlib in python.