Course: DSC 680 – Applied Data Science

Project Name: Household Income by Race in the United States

Proposal and Data Selection – Milestone I

Data sources

Just like the first project, I have used government statistics because national governments are often the only institutions with the resources to collect comprehensive social statistics, and thus publish many social statistics available. There are two datasets that I will examine. The primary data has not been generated by surveys, interviews, and experiments. It is a normal dataset that was generated from census.gov, and it is designed for understanding and solving the research problem at hand. The secondary data truly follows the definition of a real secondary data. It is also generated by census.gov. It will serve as supporting data for the project. US. Census Bureau is about the government-informed statistics on the lives of US citizens including population, economy, education, geography, and more, which is a great source to gather data.

- Data source: https://www.census.gov/search results.html?q=income+by+race+in++florida&page=1&stateGeo=none&searchtype=web &cssp=SERP&_charset_=UTF-8

- Data source: https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-detail.html

Approach

I will use Python and Power BI for this project. I am not very comfortable using Python, but I am sure that its applications are very well aligned with this type of projects. I will use NumPy to assist me with any type of linear algebra, Fourier transform, and matrices. With Pandas, the DataFrames allow me to store and manipulate tabular data in rows of observations and columns of variables, which will assist me with data wrangling. I will also use Matplotlib for my visual applications. It allows me visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, and histogram.

I will also use Power BI for the following reasons.

- Power BI offers a wide range of custom visualizations. That means visualizations made by

  developers for a specific use. Custom visuals are also available on Microsoft marketplace.

- In Power BI, I also have the option to upload and view my data in Excel. I will be able to select, filter, and slice data in a Power BI report or dashboard and put it on Excel. I will be able to open Excel and view the same data in tabular form in an Excel spreadsheet.
- Another reason why I will use Power BI is because the data visualization tool is very attractive, intuitive, and interactive. It is very easy to create and understand data through visualizations in Power BI.

Explain the Datasets

Several authors have included other American racial minorities in their analysis of racial wealth disparities. For example, Lusardi (2005) researches the differential savings behavior of both African American and Hispanic households. Hanna & Lindamood (2008) examine the decrease in stock ownership of multiple minority racial groups, including black, Hispanic, and Asian populations. Furthermore, Campbell & Kaufman (2006) perform their decomposition analysis on White, Black, Asian, Native-American, and Hispanic, using data from the 1992 Survey of Income and Program Participation, and find that group differences in the determinants of wealth, particularly indicators of socioeconomic status, have stronger effects for white populations than for racial minorities.

1. The first data is called “Average Income by State by Race”. We know that education plays a big role in income. This dataset does not reflect any educational background. This dataset will be analyzed to determine the average income of each race in each state in the United States. It will also help us understand the gap between white Americans, African-Americans, Hispanic, Asians, and Native American is particularly conspicuous, not only because these are the five largest racial groups in the U.S., but also because they are the five groups that hold the most wealth per capita and the least wealth per capita.
2. The second data is called “State Population by Race”. This dataset will be used almost the same way as Average income by state by race. The goal of this dataset is to use it to

prove that prove that these five groups are the most populated groups in the United States.

Data Analysis and Research Questions

- Does wealth inequality exist in the United States?

  Wage earnings are the single most important income source in the U.S. They account for over 70 percent of total personal income. Therefore, wage may be the key determinant of income inequality. Since the 1980s, slower economic growth, higher unemployment and reduced wage shares have been observed in the U.S. Although many studies focus on inequality of outcomes (e.g., household income or personal income or educational attainment), the outcome inequality does not adequately reflect the inequality in society since this inequality is also the result of different levels of efforts. Thus, inequality itself can be neither all good nor all bad.

- Does racial diversity play a role in the income and wealth inequality in the United States?

  Just like I mention in the data explanation, we know that education has played a big role in income and wealth inequality in America. One general analysis of the educational importance in wealth is the fact that we can clearly see the difference of educated Blacks and Hispanic and the uneducated ones when it comes to quality of life. For example, because white Americans perceive a greater threat from and exhibit more prejudice toward blacks than toward other ethnoracial groups, people living in areas

with larger black groups appear less cohesive because prejudicial attitudes toward blacks also negatively influence many measures of cohesion (Lee and Bean 2010).

- Which states that have the highest minority income?

  We know that whites and Asians fare much better than Hispanics and blacks. The gap is even wider when restricting the comparison to just middle-aged, well-educated families in each of the four groups.

Project Risk

Because the datasets that I am using are very vague, it is very easy to be off task. I mention off task because I will have to extend my datasets to analyze the years. I know that analyzing the years will be a little difficult because based on my initial research of the years they do not match accordingly to each state.

1. I will try not to miss very important features during the process.
2. I will make sure that my predictions are based on the comparisons between the years and between the states.
3. I will make sure to concentrate three major areas of the dataset - High risk states, Low risk states, and Average risk states.

References:

1. Data source: https://www.census.gov/search results.html?q=income+by+race+in++florida&page=1&stateGeo=none&searchtype=web&cssp=SERP&_charset_=UTF-8
2. Data source: https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-detail.html
3. https://www.pgpf.org/blog/2019/10/income-and-wealth-in-the-united-states-an-overview-of-data
4. https://www.pewresearch.org/fact-tank/2018/07/12/key-findings-on-the-rise-in-income-inequality-within-americas-racial-and-ethnic-groups/
5. https://datausa.io/profile/geo/united-states
6. https://www.brookings.edu/blog/the-avenue/2019/10/03/black-household-income-is-rising-across-the-united-states/
7. https://www.pewsocialtrends.org/2018/07/12/income-inequality-in-the-u-s-is-rising-most-rapidly-among-asians/
8. https://www.epi.org/blog/racial-and-ethnic-income-gaps-persist-amid-uneven-growth-in-household-incomes/
9. https://news.gallup.com/poll/4435/public-overestimates-us-black-hispanic-populations.aspx
10. https://dqydj.com/income-by-race/