Pei-Jou Liu
7384221314
peijoul@usc.edu

Adult Diabetes in the United States

In today's world, diabetes is a very common disease and a serious health issue for many adults, especially in the United States. For the project, Adult Diabetes in the United States, the goal is to look at the real data to see where diabetes is most common and which places are affected the most. The research question is how common adult diabetes is across the United States and which places have the highest levels. To find out the answer, this assignment focused on the data from the U.S. Centers for Disease Control and Prevention and looked at both states and counties. The main purpose of this project is to clean the data, summarize it clearly, and produce the visualizations that show where adult diabetes is more common.

The data in this project come from public CDC chronic disease and surveillance datasets. In the very beginning, the get_data.py, I was going to download data from the CDC API and collect information on diabetes, obesity, and physical inactivity for many years and for every state and county. However, the API failed or returned incomplete results, so I chose to use the CSV files that I had already downloaded and placed in a folder. After combining everything, stored the cleaned data in the files which include prevalence_state_clean.csv for state measures, prevalence_county_clean.csv for county measures, and incidence_state_clean.csv for incidence. Overall, the cleaned datasets contain 810103 state records, 188456 county records, and 3143 state incidence records. In this project, only the rows are used to measure adult diabetes. The dataset has shown 29006 state diabetes records for the year 2017 and 6288 county diabetes records for the year 2020, which a large amount of data to work on.

In addition, before running the analysis code, clean_data.py is used. It reads all the raw CSV files, fixes the column names by making the lowercase and removing extra spaces and symbols, and then uses the filenames to separate out which year each file belongs to and whether it is state or county data. Files of the same type are combined into one big table, then saves the cleaned prevalence_county, prevalence_state, and incidence_state datasets into the processed folder under the data folder. After the data has been cleaned, the analysis and visualizations part can begin to work on.

For the analysis, the data are cleaned so it only keeps the rows where the CDC variable short_question_text which is Diabetes. For the state data, the rows are separate by year and calculate the average adult diabetes prevalence across all states. Because of the API issues, I only have complete state diabetes data for 2017, so the result is a single average value rather than a long trend line. For the county data, the project has focused on 2020, which is the year with consistent data that are separated by county name and measure the average adult diabetes prevalence for each county. Moreover, the 20 counties are selected with the highest values. All of these summaries are written out as CSV files in a results folder and are used to make the plots.

To produce the visualizations, the matplotlib has been used. The first plot shows the state average adult diabetes prevalence for 2017. The figure there is a single point at about 10.79 percent. Even though it is only one year, it tells an important story, on average, about one in ten adults in U.S. states has diabetes. The second plot is a horizontal bar chart of the top 20 counties by adult diabetes prevalence in 2020. The y-axis lists the county names, and the x-axis shows the average prevalence in percent. In this figure, Presidio County appears at the top with a value above 23 percent, and many of the other top counties, such as East Carroll, Dimmit, and Zavala, also have very high levels between roughly 18 and 22 percent. Many of these counties are in the South and are relatively rural. This contrast between the national average of around 11 percent and county values above 20 percent shows that diabetes is not evenly spread across the country.

In conclusion, the changes I had made from my original proposal for the project was that I wanted to combine diabetes, obesity, and lack of exercise to find the regional patterns across many years. However, with a large number of variables together while API not working out, I decided to narrow the project and focus on adult diabetes only. If I had more time, there are few things I would like to improve on this assignment. One would be to successfully pull and merge the obesity and lack of exercise data and see how strongly they are related to diabetes at the county level. Another would be to improve the data collection which I can have a better time series for both states and counties. With more years in the graph, it will be able to draw real trend lines and see whether diabetes is rising or falling in different regions. In addition, combining the diabetes data with other information, such as income, and rural or urban place, to see which factors are most closely associated with high diabetes rates. Overall, the project shows that adult diabetes is common in the United States and that some counties have especially high levels.