**Introduction**

Hello everyone. My name is Sam.

Today my topic is

“The correlation of diabetes and obesity with physical inactivity in the US in 2017.”

This project, I download two CDC datasets which are talking about

“diagnosed diabetes and obesity among adults aged >= 18 years” in the US in 2017.

They include estimates for the 500 largest cities and approximately 28,000 census tracts within these cities.

I also download another dataset which is the “Prevalence of Self-Reported Physical Inactivity Among US Adults by State and Territory, 2017–2020”.

**Methods**

So first, I use API method to obtain my datasets.

You have to create your account with password.

Apply for a free token.

Last, copy your API URL.

After that, I get both datasets contain 27 columns and 29,006 rows.

And the Physical Inactivity dataset can be downloaded directly from this URL.

It contains 3 columns and 53 rows.

So after getting all datasets,

I’m going to select columns which are important.

I select data\_value(%), populationCount, stateabbr, statedesc(state name), city\_name, latitude, and longitude, total 7 columns.

From the Physical Inactivity dataset, I select state and prevalence 2 columns.

Then, change column names in order to understand them easily.

After changing column names, I remove the physical inactivity prevalence non-numeric value in its dataset.

Then, I merge two datasets by state\_abbr, populationCount, state\_name, city\_name, lat, and lon.

And merge the physical Inactivity dataset with them.

Now, I have 30,008 rows and 9 columns.

My row number increased so I have to remove duplicates.

After removing them, my rows drop to 28,505.

Next, I convert numbers from character variables into numeric variables.

And check NAs vlues, there are only 2.7% NAs values in diabetes\_percentage and obesity\_percentage columns.

Therefore, I’m going to remove them.

Last, I create a new column contain four different regions.

**Results**

**Leaflet**

From the Leaflet, the legend shows the degree of the diabetes percentage.

The red color means higher percentage of diabetes.

I see there are more orange dots in the NE and SE region from the plot of diabetes percentage.

From the second plot of obesity percentage, we can see that there are more orange dots on the east-side than on the west-side.

**Boxplots**

Now, let see the boxplot, the x-axis shows 4 regions: Northeast, Southeast, Northwest, and Southwest. On the y-axis shows the percentage of diabetes or obesity.

From the boxplot of diabetes percentage, the NE and SE region have a similar median diabetes percentage.

The NW region has the lowest median diabetes percentage. In this plot, the east-side regions’ median diabetes percentage is higher than the west-side regions.

From the boxplot of obesity percentage, the east-side regions’ median obesity percentage is also higher than the west-side regions.

Among all regions, the SE region has the highest median obesity percentage and the SW region has the lowest median obesity percentage.

**Scatter plot**

In this scatter plot, I select each state’s median of obesity and diabetes percentage with their physical inactivity prevalence.

We can see that there is a positive correlation between obesity and diabetes.

We can also see the physical inactivity prevalence by the size of dots.

I cannot tell the direct correlation of physical inactivity prevalence with the obesity or diabetes percentages.

However, it looks like there are bigger dots on the top right corner and smaller dots on the bottom left corner.

So, with higher percentage of obesity or diabetes, tend to have higher physical inactivity prevalence rate.

**Conclusion**

Question 1: How are the distribution of diabetes and obesity percentages in the US?

From the leaflet, first we can see there are more orange dots on the NE and SE regions.

From the box plot, the median of diabetes percentage looks equally high in the NE and SE regions.

The median of obesity percentage looks higher in the SE region than in the NE region.

Besides, we can also see there are higher diabetes and obesity percentages on the east-side than on the west-side.

Question 2: Is there any correlation between diabetes and obesity? If so, is there any correlation with physical inactivity prevalence?

From the scatter plot, we can see that there is a positive correlation between obesity and diabetes rates by states.

As for the correlation with physical inactivity prevalence, I don’t see a strong correlation with physical inactivity prevalence.

However, I can tell that smaller dots are on the bottom left corner and bigger dots are on the top right corner.

**Discussion**

For the discussion section, I would like to add more variables to my dataset in order to make interesting plots. For example, I want to add different years and time to check if there are any changes of diabetes and obesity percentages in each state. Adding each state’s median income level, average education level, and so on, to see whether diabetes and obesity percentages would be affected by them.