

Olivia McKellar

DATA 219

Final Project

The Impact of Chronic Illness on Mental and Physical Health

The question I'm answering is "How does chronic illness impact an individual's mental and physical health?" I am using a dataset covering participants' mental and physical health, stress levels, and lifestyle. I believe it is important to understand the impact chronic illnesses have on individuals' overall quality of life. This can be done by finding correlations between the presence of a chronic illness and things like mental illness, sleep, exercise, stress levels, and life satisfaction. I will also examine whether there's a correlation between chronic illness and education level, and employment status, to understand how a chronic illness may affect access to work or school.

The data set I used for my project is titled "Anxiety and Depression Mental Health Factors" which includes the participant's demographics and lifestyle, as well as physical and mental health information. The data lists the age, gender, education level, and employment status of the participants. The participants were asked how many hours of sleep and physical activity they engage in daily and if they have a family history of mental illness or a chronic illness. They listed if they are in therapy, on medication, using substances, or practicing meditation. It also includes a ranking system for topics like anxiety, depression, stress, self-esteem, social support, life satisfaction, and loneliness.

To begin, I started with uploading and cleaning the dataset. To do this, I uploaded and read the file containing my dataset. Then, I first checked for any null values and summed the totals for each column to understand how it impacted my data. Two columns had a large number of null values and were not directly related to my research question. So, I dropped those two columns along with a few others that were not necessary for the data analysis. Next, I needed to verify that there were no typos, and all the response types matched the category type. I was able to view the unique values of each column so that I could identify any responses that did not match. There were no typos or mismatched responses, so I moved on to looking for outliers. I used the Inter Quartile Range to identify any outliers since I expect this type of data to include extreme values. This gave me the resulting values and the corresponding column it's found in. I reviewed the outliers and did not remove any since the values are realistic numbers for the category type. It included sleep and physical activity hours, which may vary drastically depending on the individual and their lifestyle. So, I decided to keep these outliers since it is necessary for my research. I wanted to view a statistical summary of the cleaned data. I used the describe function to get results like

minimum, maximum, mean, standard deviation, and quartiles of each column. This gave me an insight into the average number of participants with chronic illnesses and the most prevalent issues.

The data structures I used to store and analyze my data are lists, heaps, and a dictionary. Lists are most efficient for storing my data since they allow for filtering and grouping of the data. I will also be able to easily perform calculations on the lists. I can easily convert the dataset into a list to be ready for analysis. Since my data is cleaned, I'm able to convert the dataset into these data structures. I sorted the dataset into two lists: those with at least one chronic illness and those without a chronic illness. I then printed the lists so I could see that both lists contain the appropriate data. I can then use heaps to quickly access the minimum and maximum values for the numerical columns in each list. This assists me in my data analysis by obtaining the highest values for comparison. I will be able to point out if the top or bottom scores belong to those with or without chronic illnesses. I will then use a dictionary to access the columns in my lists containing strings to find the mode for each category.

I began my analysis by implementing heaps to compare the minimum and maximum values for certain categories. I created a heap containing the columns ranking anxiety, depression, stress, financial stress, work stress, and loneliness. I added all of these rankings together for each participant and labeled this as a "stress score" with the highest rankings meaning the worst stress. I found the top 5 scores for those with and without chronic illnesses. Those with a chronic illness had the top scores in the low to mid-60's. Those without a chronic illness had the top scores in the mid-60's. Then, I compared the columns of sleep, physical activity, social support, self-esteem, and life satisfaction where the lowest scores are a negative experience. These results again came out very similar, with the lowest scores for those with a chronic illness ranging from 11-14 and those without a chronic illness ranging from 9-12.

Next, I calculated the averages for each column in each list to compare those with and without a chronic illness. I set up a function that would exclude the first two non-numeric columns and follow the steps to calculate the average. The averages for each column came out to be almost the same, with just a difference in decimals. I would have performed a t-test to see if the difference was significant, but the results show us it is not a significant difference. Lastly, I created a dictionary to count the values in the first two columns in each list by referencing the correct index for each one. These include the education and employment categories with string values. I was able to count each value in the columns and find the value with the highest count to calculate the mode. The printed results showed me the education level and employment status that occurred most along

with the number of occurrences. The mode for education and employment for those with chronic illnesses were high school and unemployed. The mode for those without a chronic illness was PHD and employed.

The lists were the best option for storing my data to be able to have access to all of the information in columns and rows. This allowed me to access all the values in certain columns to sum them together and calculate the average. The heaps worked best to access the minimum and maximum for the sums I had calculated. I was able to store these sums in heaps and then quickly access the top 5 and bottom 5 values. A dictionary was best for calculating the mode of columns containing strings. I was able to store the values for each column into a dictionary, and it counted the values and identified the most occurring one. The results from this analysis showed me that there is not enough evidence to state that there is a significant difference in the mental and physical health between those with and without chronic illness. However, there is evidence that leads me to believe that those with a chronic illness may face more obstacles when it comes to education and employment. Those with chronic illnesses were mostly unemployed, with the highest education level being high school.