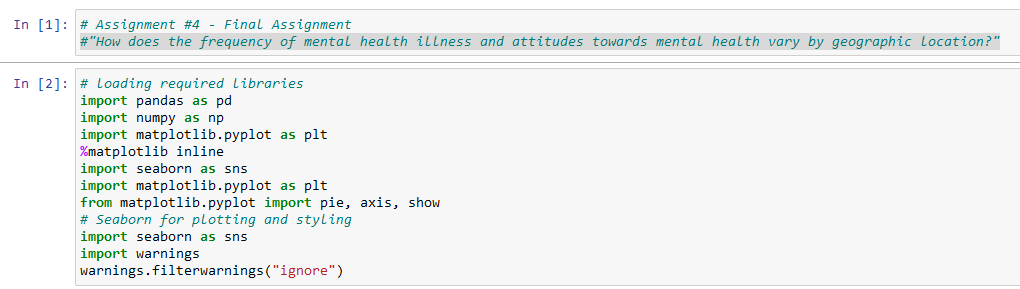
**Assignment #4 - Final Assignment\_120201\_DATA\_Swati\_Pal\_100845961**

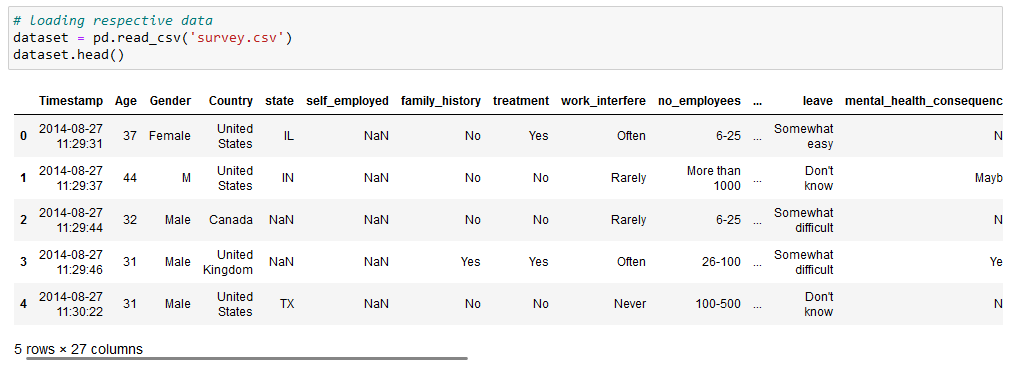
**Problem Definition:** How does the frequency of mental health illness and attitudes towards mental health vary by geographic location?

**Solution:** We have dataset survey.csv which has multiple columns containing Age, Gender, Country, state, mental\_health\_consequences, etc. This column will help us to determine our solution

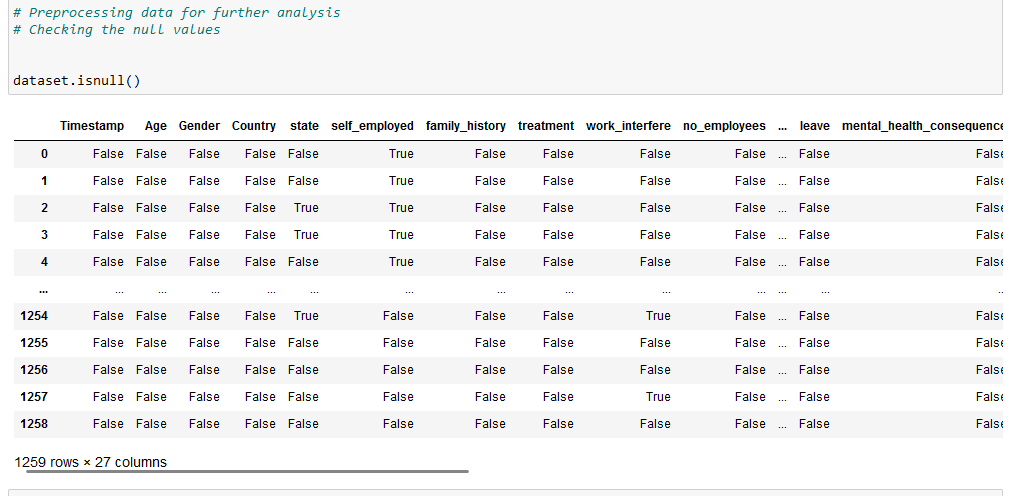
**Step 1:** Loading respective libraries required for further analysis AND plotting of graph.

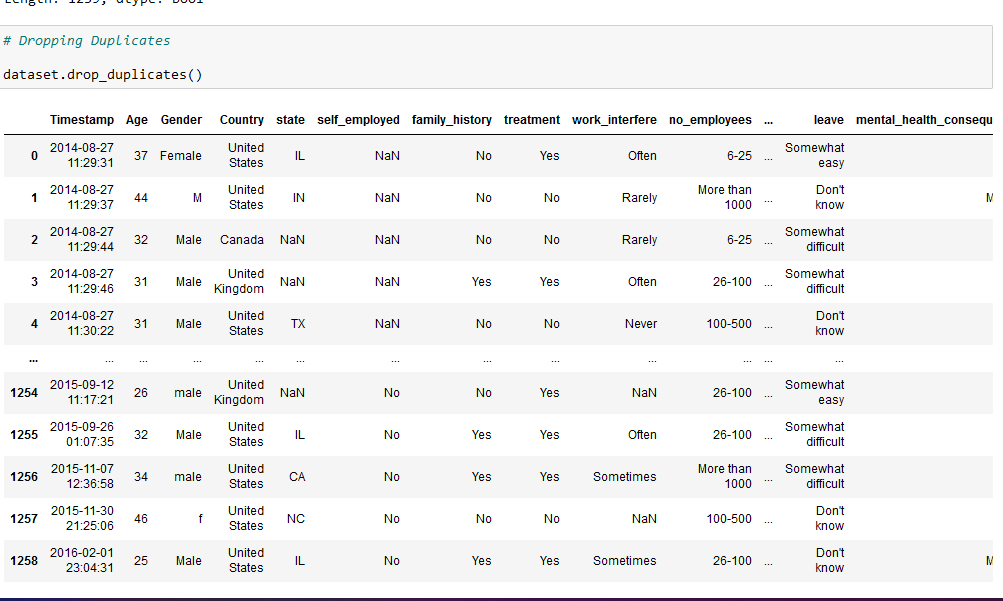


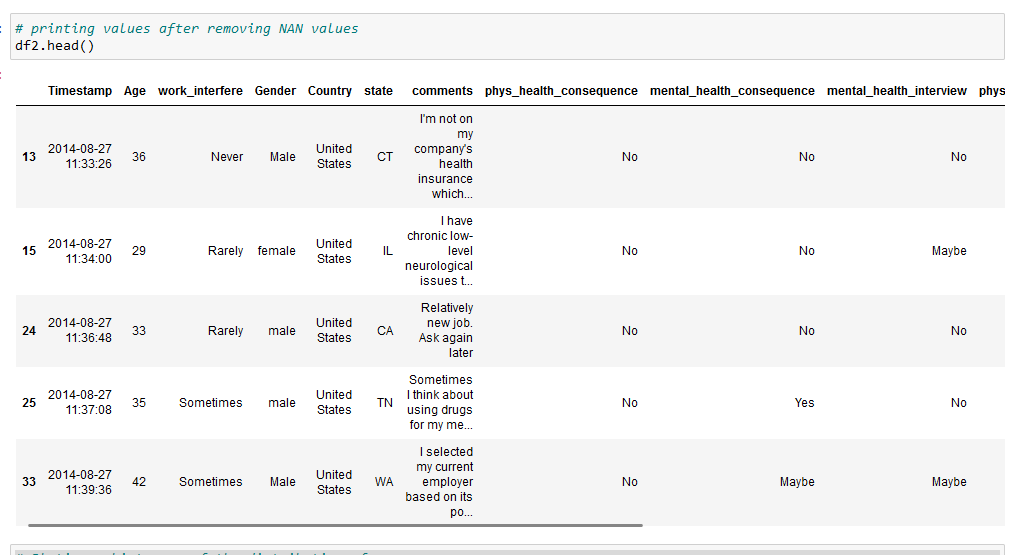
**Step 2:** Loading Survey.csv into our jupyter notebook to work on further data.

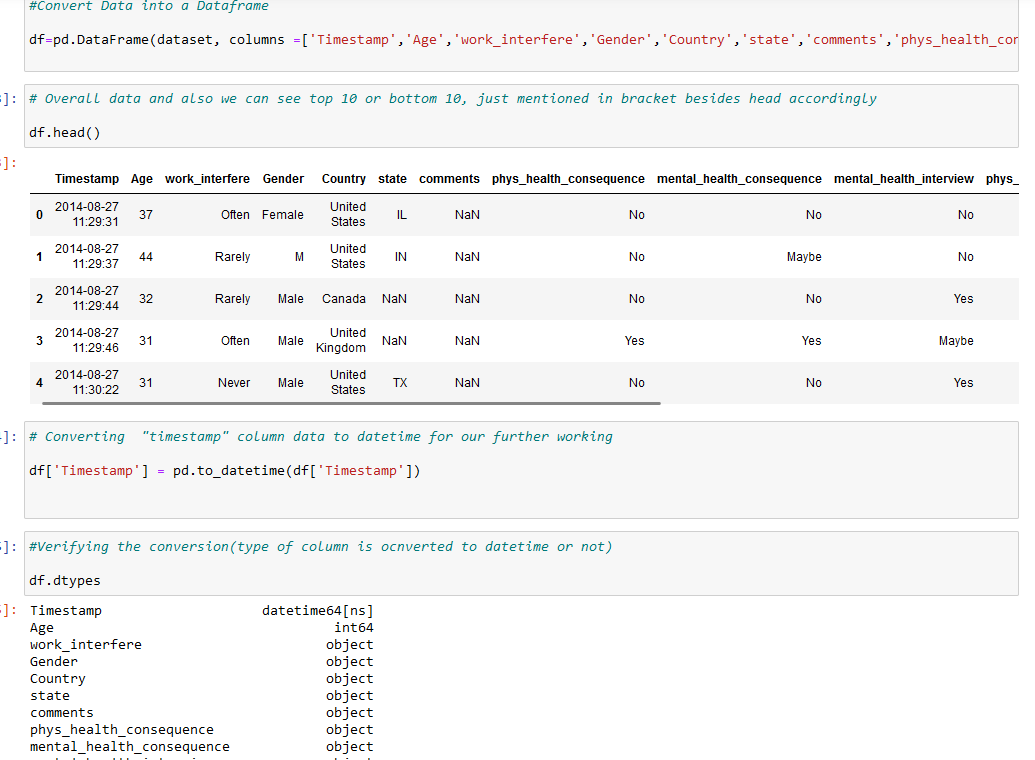


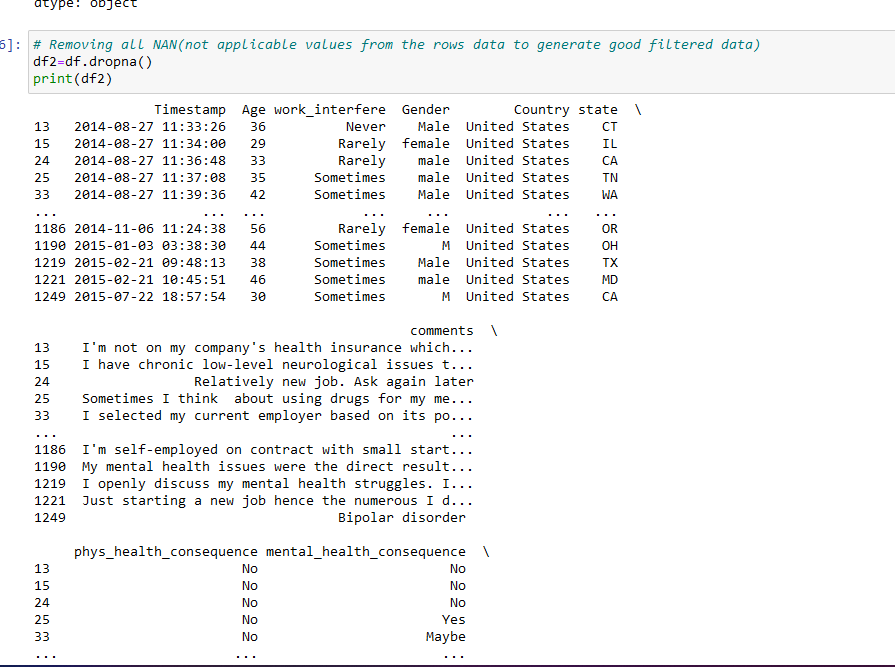
**Step 3:** Massaging the data, which is basically pre-processing the data to removing null and duplicate values, etc. And also, transformation of Data according to our need.





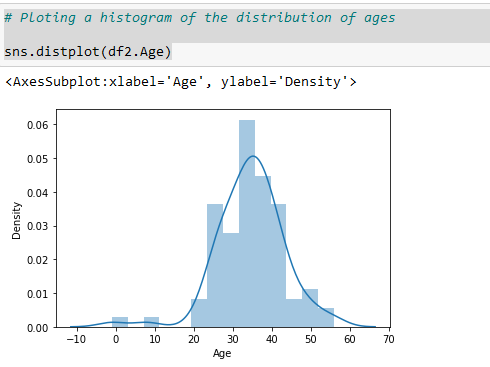




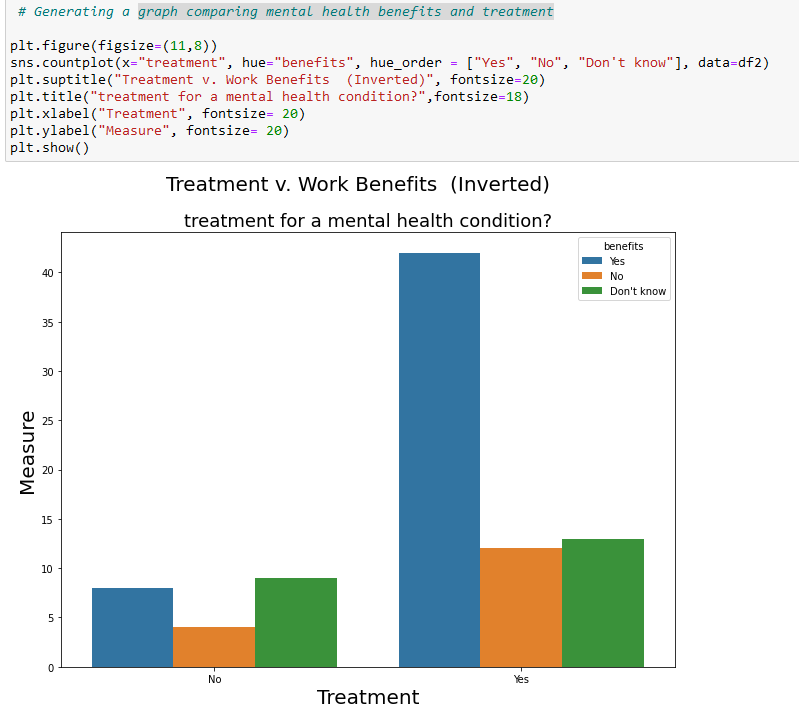


**Step 4:** We want to check frequency of mental health illness and attitude varied by geographic location. So, for this we will draw multiple graphs, which will helps us to determine our relevant solution.

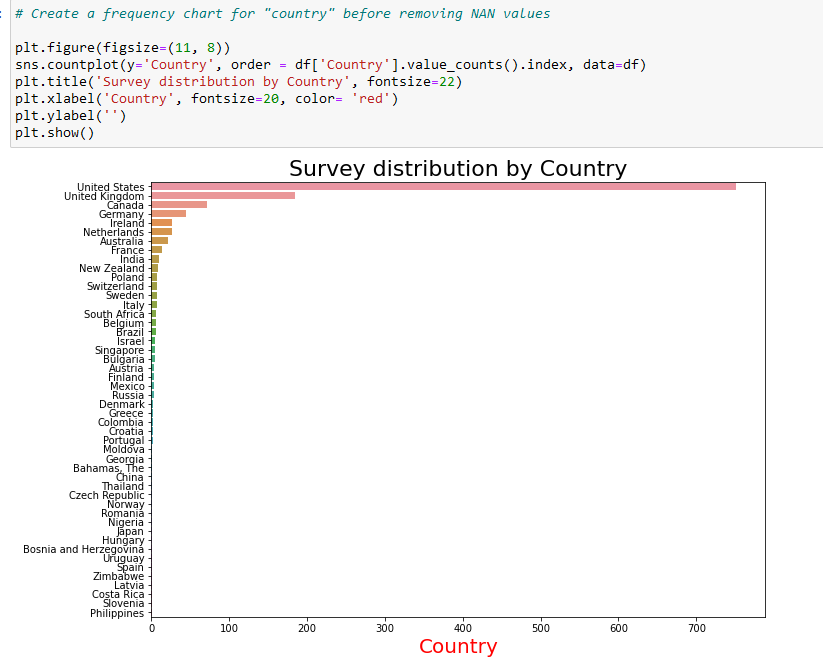
**Graph\_1:** Mental health Age distribution according to our data-set



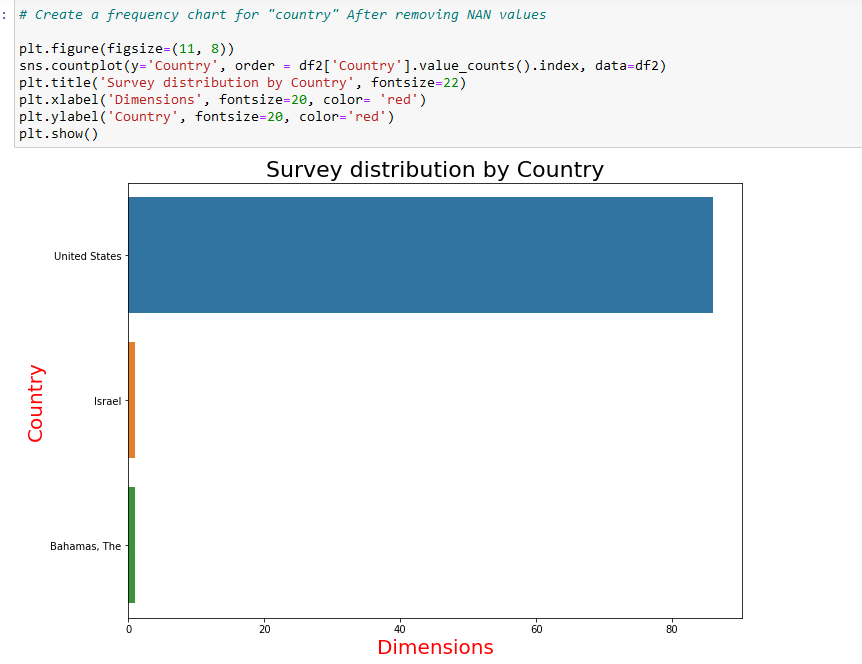
**Graph-2:** Comparing mental health benefits and treatment.



**Graph-3:** Survey Distribution by Country before removing NAN values



**Graph-4:** Survey Distribution by Country After removing NAN values



Now, giving some variable name to some specific data, it will help us to plot our graph easily and efficiently.

For example, Attitudes= mental\_health\_consequence(this column is given in dataset)

Average\_Attitude = ['Maybe']

Positive\_Attitude = ["No"]

Negative\_Attitude = ["Yes"]

Work\_interfere\_illness= work\_interfere

Average\_interfere\_illness= ["Sometimes"]

Rare\_interfere\_illness= ['Rarely']

Positive\_interfere\_illness = ["Never"]

Negative\_interfere\_illness = ["Often"]

Zone= State

West\_zone = ["WA", "OR", "CA", "NV", "ID", "MT", "WY", "UT", "AZ", "NM", "CO"]

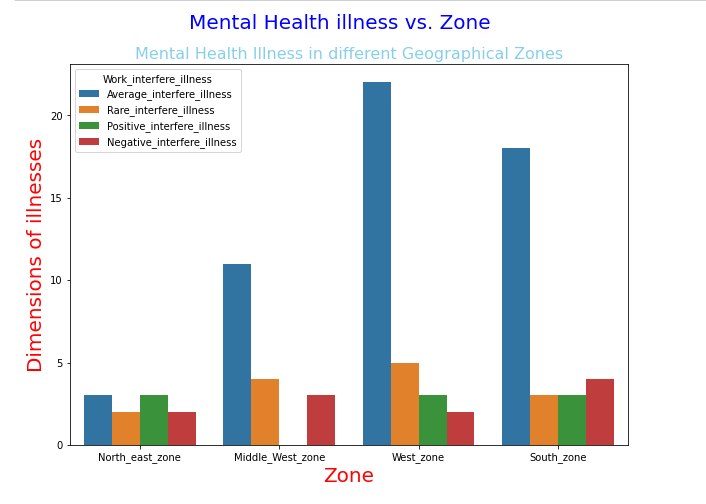
Middle\_West\_zone = ["ND", "SD", "NE", "KS", "MN", "IA", "MO", "WI", "IL", "IN", "OH", "MI"]

North\_east\_zone = ["ME", "NH", "VT", "MA", "CT", "RI", "NY", "PA", "NJ"]

South\_zone = ["MD", "DE", "DC", "WV", "VA", "NC","SC", "GA", "FL", "KY", "TN", "AL", "MS", "AR", "LA", "OK", "TX"]

Note: We did, all this to understand data more accurately, as we know, there are multiple state, so plotting graph is not easy, that is why we have accumulated different states in different zones/regions respectively.

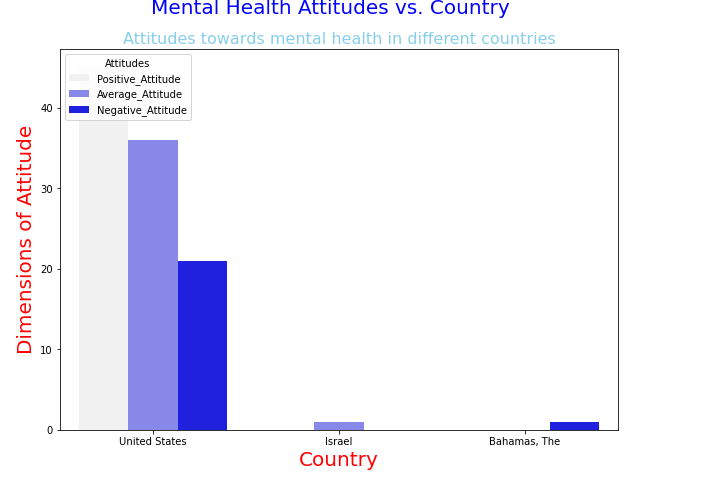
**Graph-5:** Mental Health Illnesses Vs. Zone/region

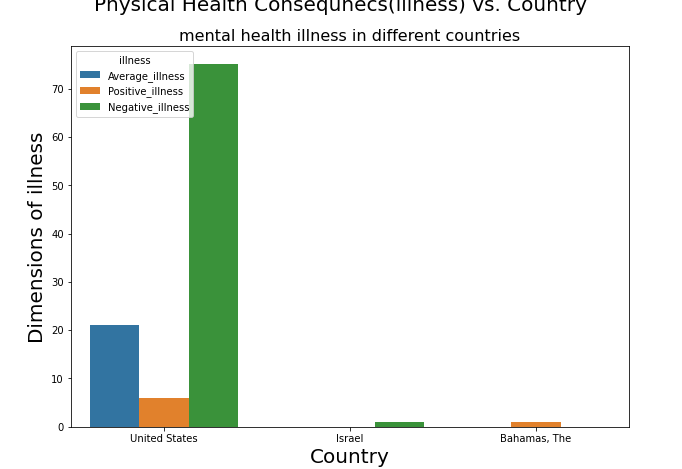


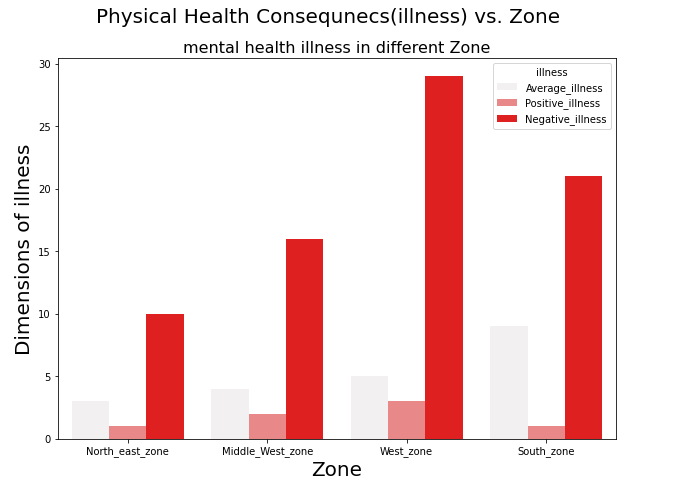
**Graph-6:** Mental health attitudes respective to different zones.



**Graph-7:** Mental Health Attitude and illnesses respective to different countries/zones after NAN data removal.







In Conclusion, We have plotted all the graphs according to countries and geographical zones(states) respective to mental health attitude and illnesses and found that "**West** zone" is the one which has high mental illnesses and mental health attitude problem. And "**United States**" in a country which has high cases of mental health illnesses and mental health attitude problem.

**Difficulties during whole assignment:** Data was not accurate, so removal of NAN values and searching and learning of Lambda function to accumulate different states values in different zones. Also, I have chosen variable df as it is the basic way to define variable in python and I wanted to keep it simple. And regarding other graph variables, I have chosen name according to the name recognizes in X and Y axis. So, that other person does not have any difficulties in understanding data and graph.

**GITHUB:**

