Exercise 6.1
Laura DeCesare

Data Source

For this project, I will use a Kaggle data set entitled Mental Health Depression Disorder Data. It provides information on the prevalence of depression and other mental health disorders across the world. It looks at depression rates by age, level of education, and gender.

According to the contributor, this data comes from international health organizations and health data sources, including the World Health Organization (WHO) and Global Health Data Exchange (GHDx). The data underwent validation before publication.

I selected this data set because I have a background in mental health and an interest in using data to design, deliver, and assess mental health services.

Data Profile

1. Data Cleaning

The majority of worksheets in this Excel document feature information from 1990 and later, but the gender prevalence and suicide tabs include centuries-old population data. Filtering by year, I was able to remove those rows.

I also noticed that many rows labeled 2018 and 2019 were missing data. By applying another filter to view only those years, I saw that there was no suicide rate or depressive disorder data for those years. I eliminated those years as well.

Finally, I removed rows for

- Africa,
- Anguilla
- Aruba
- Asia
- Bonaire Sint Eustatius and Saba
- British Virgin Islands
- Cayman Isa
- Channel Islands
- Cook Islands
- Curacao

- Europe
- Faeroe Islands
- Falkland Islands
- French Guiana
- French Polynesia
- Gibraltar
- Guadeloupe
- Hong Kong
- Isle of Man
- Latin America
- Liechtenstein
- Macao
- Martinique
- Mayotte
- Monaco
- Montserrat
- Nauru
- New Caledonia
- Niue
- Palau
- Reunion
- Saint Barthlemy
- Saint Helena
- Saint Kitts and Nevis
- French Saint Martin
- Saint Pierre and Miquelon
- San Marino
- Dutch Sint Maarten
- Tokelau
- Turks and Caicos Islands
- Tuvalu
- Vatican
- Wallis and Futuna
- Western Sahara

I also moved several categorical records, including “High-income” and “High SDI” to another sheet, because this data is potentially valuable but would throw errors into a geographic analysis. However, I ended up re-integrating those values into my final spreadsheet.

I then integrated the data using VLOOKUP to consolidate multiple worksheets into one table I could work with in Python.

2.Understanding the Data

The dataset provides information for world countries and regions between 1990 and 2017. There are 6,468 unique records in total. Each record represents one year in a particular country or region (e.g. “Argentina, 2017” or “Australia, 2009.”

A brief descriptive analysis in Jupyter revealed the following:

- On average, 3.5% of a country’s population struggles with depression.
- Only anxiety is more common on average at 4%.
- Average rates of depression seem to increase by age, peaking at a mean of 6.14% in the 70+ age group.
- Mean depression prevalence is significantly higher in females than males (4.16% vs. 2.81%)

4. Data Limitations

Since we lack a detailed picture of collection methods, we can’t be sure of the data’s accuracy and completeness. Its origin from government sources is promising, however.

The primary limitation is the unavailability of post-2017 data. Global mental health has changed significantly since then, but that information has not been collected and aggregated in this way.

Questions to Explore

- Which countries and regions have the highest and lowest depression rates
- How have depression rates changed over time?
- What ages are most likely to experience depression, and how has that changed over the years?
- Which other mental illnesses have the strongest correlation with depression?
- When and where are suicide rates the highest, and do those peaks correlate with high rates of depression?