Task 6.1: Sourcing Open Data Sydney Storer

Data Source -
https://www.kaggle.com/datasets/programmerrdai/mental-health-dataset?select=mental-and-substance-use-as-share-of-disease.csv

This dataset contains information about mental illnesses and the burden of mental illness in different countries around the world.

Data Collection Method:
It was unclear how this data was collected based on the Kaggle page, however with further research, I believe the data is based on the Global Burden of Disease Study. (based on exploration of an external website https://vizhub.healthdata.org/gbd-results/)

Data Contents:

Variable	Description
Entity (name changed to Country)	a categorical variable representing the country
Code (column dropped)	Abbreviation of the country
Year	Continuous variable showing the year the row of data is from (1990-2019)
Prevalence - Schizophrenia - Sex: Both - Age: Age-standardized (Percent)	A continuous variable showing the prevalence of schizophrenia disorders (percent of population)
Prevalence - Bipolar disorder - Sex: Both - Age: Age-standardized (Percent)	A continuous variable showing the prevalence of bipolar related disorders (percent of population)
Prevalence - Eating disorders - Sex: Both - Age: Age-standardized (Percent)	A continuous variable showing the prevalence of eating disorders (percent of population)
Prevalence - Anxiety disorders - Sex: Both - Age: Age-standardized (Percent)	A continuous variable showing the prevalence of anxiety disorders (percent of population)
Prevalence - Drug use disorders - Sex: Both - Age: Age-standardized (Percent)	A continuous variable showing the prevalence of drug use disorders (percent of population)
Prevalence - Depressive disorders - Sex: Both - Age: Age-standardized (Percent)	A continuous variable showing the prevalence of depressive disorders (percent of population)
Prevalence - Alcohol use disorders - Sex: Both - Age: Age-standardized (Percent)	A continuous variable showing the prevalence of alcohol use disorders (percent of population)
DALYs (Disability-Adjusted Life Years) - Mental	A continuous variable showing the percentage of

disorders - Sex: Both - Age: All Ages (Percent)	total potential healthy life that is lost due to premature death or disability

Reason for Choosing:
I chose this dataset because I am passionate about mental health and find it fascinating to see the percentages of populations in different countries dealing with mental illnesses. I look forward to discovering possible correlations between variables and forecasting global mental health diagnoses rates.

Data Profile -

Data Cleaning:
There were no missing values or duplicates. I dropped the Code column because it was unnecessary to the analysis. I changed the name of the Entity column to “Country” for ease of understanding. I also shortened all of the other columns by deleting “Sex: Both - Age: Age-standardized” from the prevalence columns and “Sex: Both - Age: All Ages” from the DALYs column, because they are constant among all of the variables and can be assumed to be true for the rest of the analysis.

Descriptive Statistics:

	Year	Prevalence - Schizophrenia (Percent)	Prevalence - Bipolar disorder (Percent)	Prevalence - Eating disorders (Percent)	Prevalence - Anxiety disorders (Percent)	Prevalence - Drug use disorders (Percent)	Prevalence - Depressive disorders (Percent)	Prevalence - Alcohol use disorders (Percent)	DALYs - Mental disorders (Percent)
count	6840.000000	6840.000000	6840.000000	6840.000000	6840.000000	6840.000000	6840.000000	6840.000000	6840.000000
mean	2004.500000	0.281167	0.673891	0.211062	4.327525	0.746708	3.950449	1.578807	4.818062
std	8.656074	0.047561	0.258594	0.152559	1.177961	0.463026	0.921021	0.934655	2.294029
min	1990.000000	0.191621	0.189344	0.045425	1.974823	0.225471	1.640902	0.319900	0.215647
25%	1997.000000	0.255468	0.539791	0.099857	3.567064	0.423502	3.258977	0.732826	3.006507
50%	2004.500000	0.287456	0.591893	0.154143	4.094443	0.646050	3.904117	1.460045	4.679177
75%	2012.000000	0.304760	0.897248	0.276891	4.797286	0.890013	4.550505	2.261262	6.387488
max	2019.000000	0.506018	1.676204	1.136541	9.015948	3.699504	7.688213	4.698694	13.761517

Data Limitations:
Some people may not be comfortable reporting mental health diagnoses, regardless of anonymity. This is due to the stigma around mental health and individuals possibly being ashamed/embarrassed of their diagnoses.
Additionally, some people may not be aware of their mental health symptoms and believe life is normal as they experience it.

Ethical Considerations:
This dataset has a CC0: Public Domain license. Although there are no copyright laws surrounding the use of this dataset, there is no clear source of this dataset mentioned in the Kaggle description, so there is a lack of transparency where this data is sourced/collected from. For this reason, it is important to note that bias could've been introduced into the data at any point from collection to being shared.

Key Questions:

- How has the prevalence of different mental health disorders changed over time?
- How will the prevalence of mental health disorders in a country change as time goes on?
- How will the prevalence of substance use disorders change as time goes on?
- How do mental health diagnoses affect DALYs (Disability-Affected Life Years)?
- How do substance use diagnoses affect DALYs?
- How does the prevalence of disease vary amongst different regions of the world?
- How does the prevalence of substance-use disorders vary amongst regions of the world?