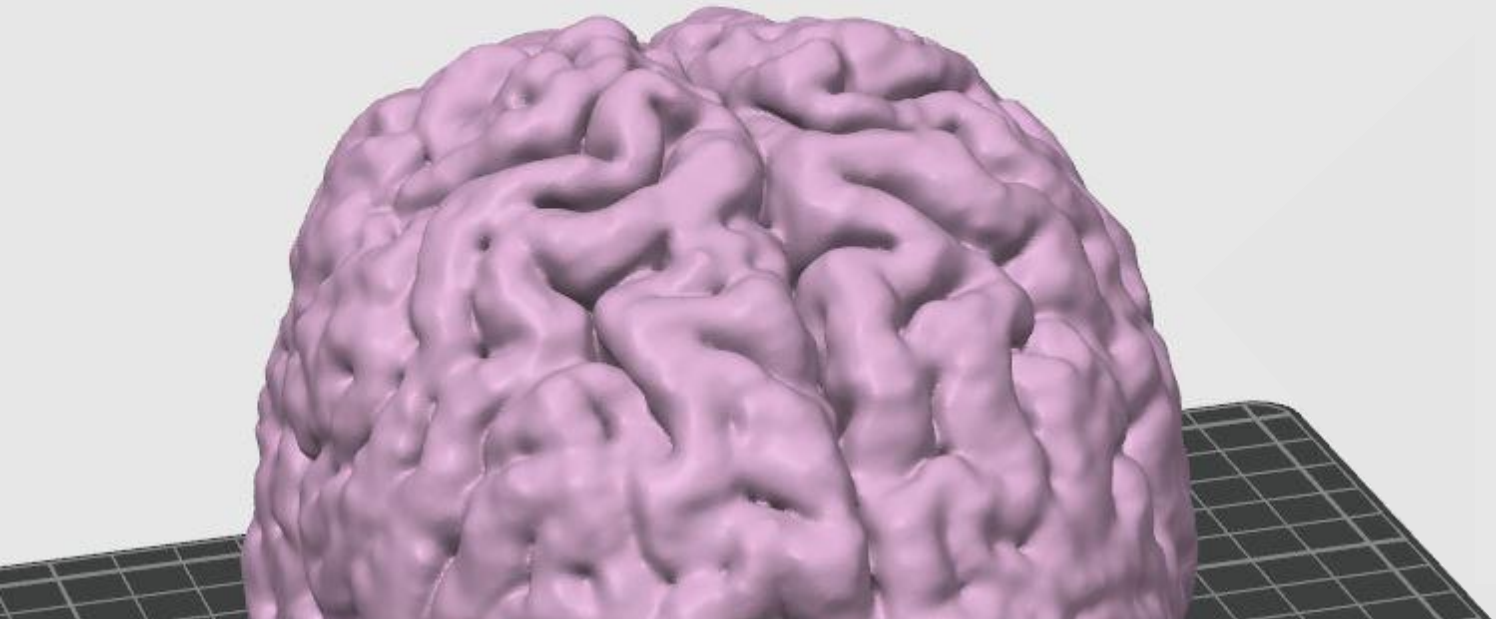
The Global Scope of Mental Health Disorders

Analysis Case Study

Marta Majer: Data Analyst
Project Time: 01.10.2024 - 30.10.2024



Project Overview: Motivation, Objectives, and Scope

🧠 Motivation

In this analysis, I delve into the global landscape of mental health disorder rates, focusing on the varying prevalence across different countries. I explore the correlation between the frequency of depression and the overall burden and severity it imposes on individuals' lives, as measured by the DALYs metric.

🎯 Objective

I aim to highlight regional similarities by:

- Visualizing mental health disorder rates on a map.
- Analyzing trends over time from 1990 to 2019.
- Grouping regions into clusters based on the rates of depression and bipolar disorder, given that bipolar disorder is often misdiagnosed as depression.

📊 Scope

This analysis utilizes the *Our World in Data* dataset and focuses on:

- Depression and bipolar disorder rates globally.
- Temporal trends from 1990 to 2019.
- Clustering regions by prevalence and examining the rates of other mental health disorders within those clusters to provide a clearer picture of geographical differences.

Technical Methodology and Resources

Analytical Tools, Skills, and Techniques:

Data Sourcing and Preparation for Exploratory Data Analysis & Geospatial Analysis.

Supervised Machine Learning (Regression): preparing and split the data into training and test sets, and run linear regression to analyze model performance.

Unsupervised Machine Learning (Clustering): preparing data for cluster analysis; elbow technique to find optimal cluster numbers; the k-means algorithm, and attach cluster results to the DataFrame while visualizing and calculating descriptive statistics with groupby() function.

Time Series Analysis: sourcing relevant time-series data via an API, subset the data for historical insights, visualize the data and decompose its structure, conduct the Dickey-Fuller test for stationarity, and perform differencing to stationarize non-stationary data.

Storytelling in Tableau.

Resources: Access Jupyter notebooks in the GitHub repository.

Analysis Strategy and Approach

When analyzing mental health data, numbers reveal profound realities about human well-being. My investigation employed both supervised and unsupervised machine learning techniques to understand: first, the burden of depression, and second, the complex relationship between depression and bipolar disorder diagnoses.

The investigation progressed through three strategic phases:

1. **Foundation and Hypothesis:** I began by examining depression's societal impact through its relationship with Disability-Adjusted Life Years (DALYs). This initial analysis would set the stage for a deeper investigation into diagnostic patterns between mental health conditions.

2. **Two-Stage Analytical Approach:** I implemented a dual machine learning strategy

Understanding DALYs: Key Metric Overview

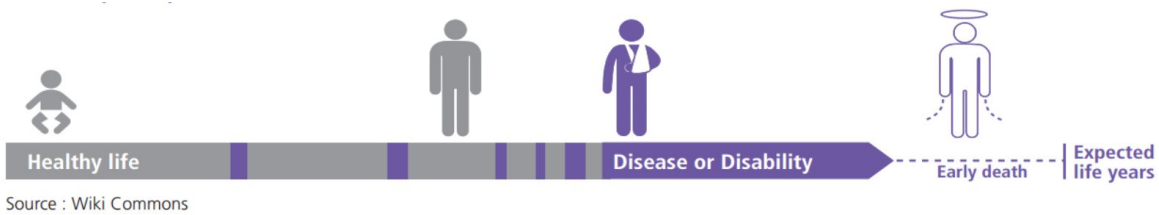
But first… What is DALYs?

Disability

Adjusted

Life

Years



Source : Wiki Commons

It measures the total impact of health issues by combining two factors:

- Years of Life Lost (YLL): The years of life cut short by early death.
- Years Lived with Disability (YLD): The years spent living with a health condition, adjusted for its severity.

Together, DALYs offer a way to understand the overall impact of diseases or conditions on a population. This measure helps identify top health priorities, compare how different conditions affect people's lives, and guide decisions on where public health resources are most needed.

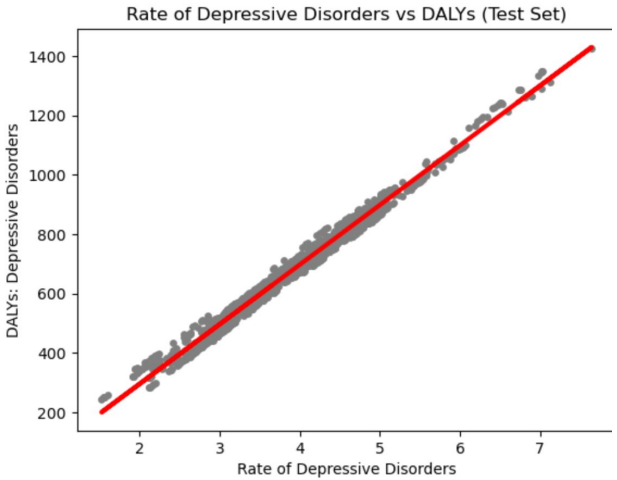
Stage 1: Depression Impact Analysis through Linear Regression

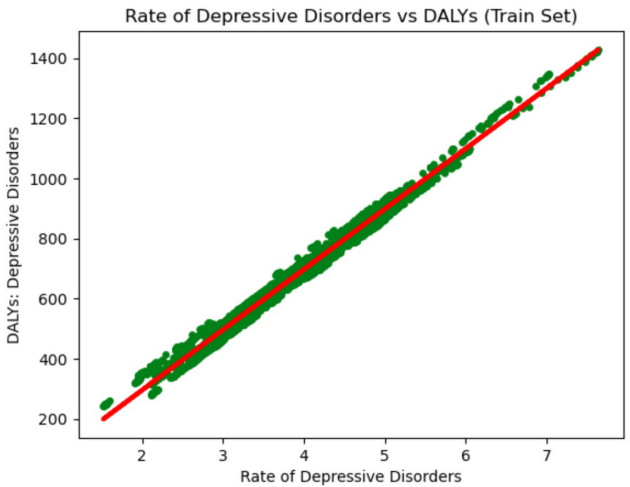
Linear Regression Analysis of Depression's Impact

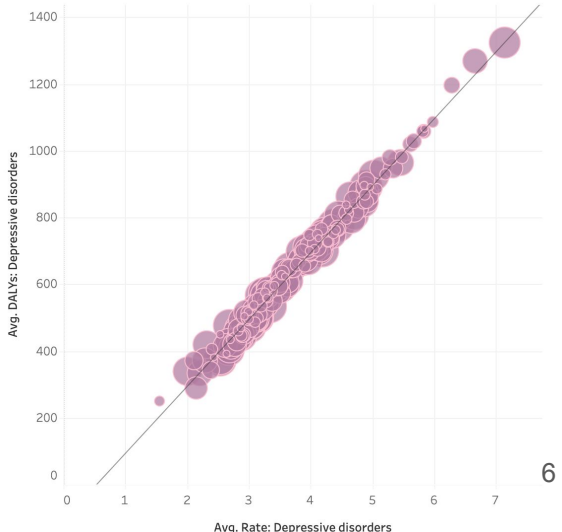
Depression's burden extends beyond prevalence rates, affecting quality of life through disability-adjusted life years (DALYs). To understand this relationship, I conducted a linear regression analysis using a supervised learning approach.

Starting with raw data, I implemented a comprehensive data cleaning process using Python's scientific stack (Pandas and NumPy). After preparing our features, I split the data into training (70%) and testing (30%) sets to ensure robust model validation. Using scikit-learn, I built a linear regression model to analyze the relationship between depression prevalence and DALYs.

Throughout the analysis, I created scatter plots using Matplotlib to visualize patterns and relationships, helping validate our hypothesis about depression severity through its correlation with disability burden. This systematic approach laid the groundwork for our subsequent cluster analysis of diagnostic patterns.





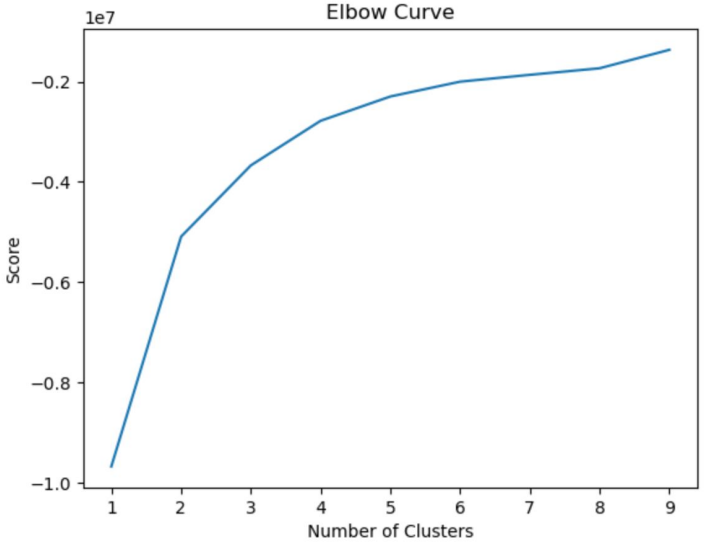


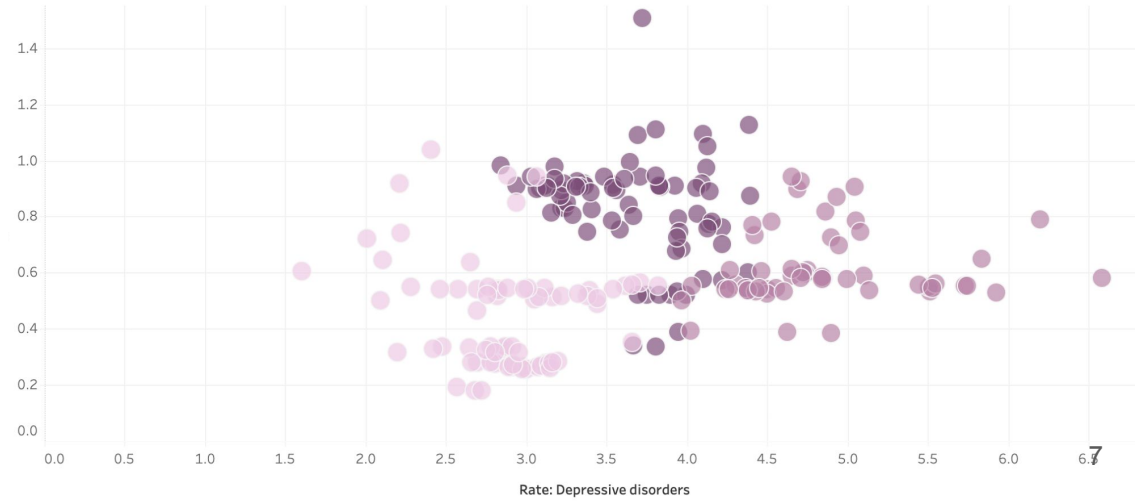
Stage 2: K-means Clustering Analysis of Mental Health Disorders

Cluster Analysis through Unsupervised Machine Learning

The high misdiagnosis rate between bipolar disorder and depression (approximately 40%) suggests these conditions frequently co-occur in population data, not necessarily because they share symptoms, but because many bipolar cases are incorrectly recorded as depression. Given this unreliability in diagnostic labels, I chose unsupervised k-means clustering (between depressive disorders and bipolar disorders) to let natural groupings emerge from the regional mental health statistics.

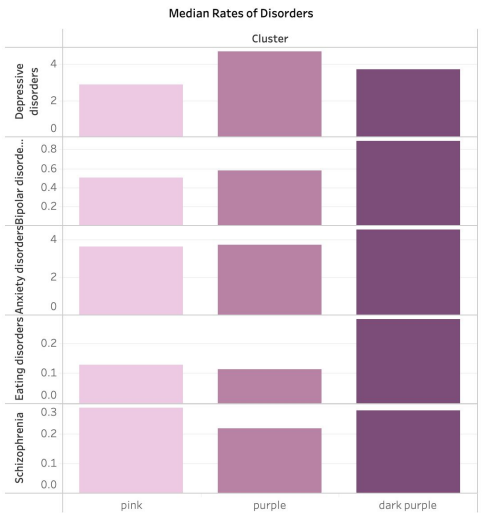
Using scikit-learn's k-means implementation, I grouped regions based on their mental health disorder rates. The elbow method helped determine the optimal number of clusters by measuring inertia (within-cluster sum of squares) across different k values. Our data pipeline utilized Pandas for preprocessing and Matplotlib/Seaborn for visualizing these cluster distributions.

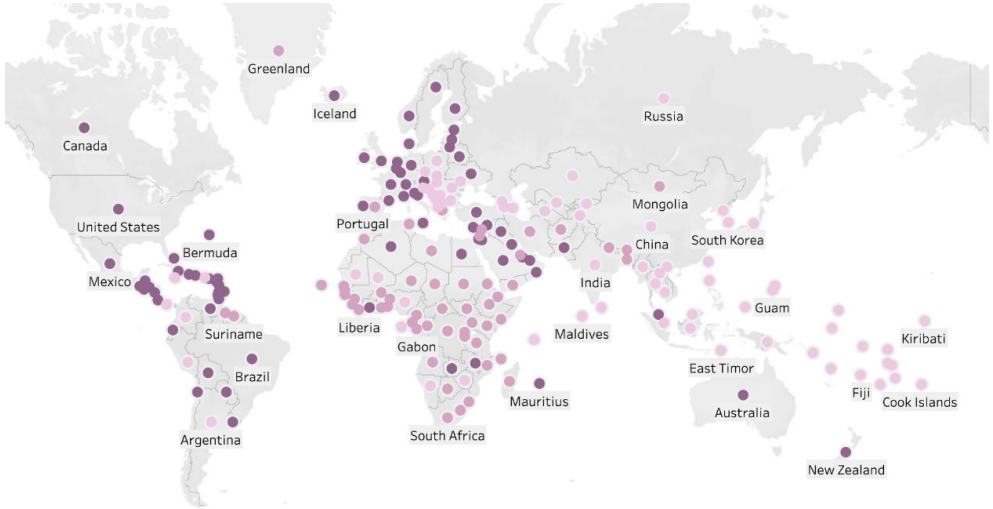




Geographical Analysis: Cluster Distribution and Patterns

After creating cluster assignments in Python, I visualized the data in Tableau to explore geographical and disorder-rate patterns across clusters. This involved mapping the clusters globally and analyzing rates of depression, anxiety, bipolar disorder, eating disorders, and schizophrenia for each group.

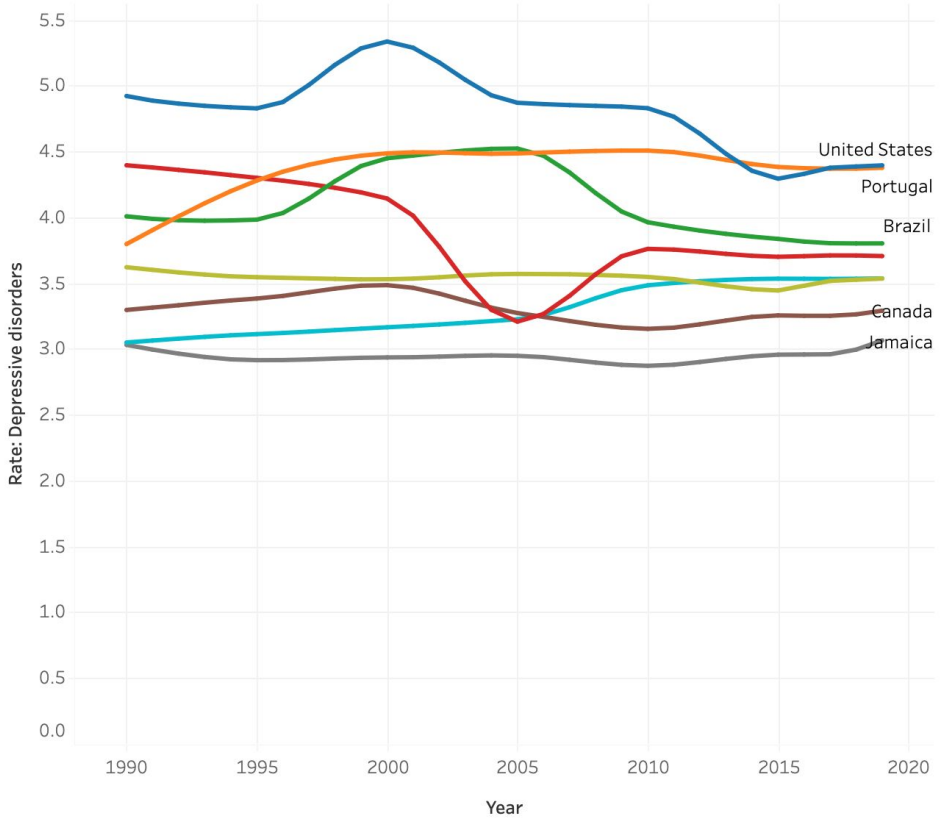




Note: Data collection inconsistencies across countries may affect the completeness of these patterns.

- Mainly high-income nations (Western Europe, Americas) with highest overall disorder rates.
- Primarily developing nations (Africa, parts of Asia) with notable depression rates but lower prevalence of other disorders.
- Mixed-income regions (Eastern Europe, Pacific islands) showing highest schizophrenia rates but moderate-to-low rates of other disorders

Temporal Analysis: Long-term Depression Rate Trends



By focusing on countries known for progressive mental health treatments, I could evaluate whether innovative approaches actually translated into improved outcomes. The long-term depression rate trends (1990-2019) revealed that only the US and Brazil showed notable decreases, despite most examined countries belonging to the dark purple cluster with typically high disorder rates.

This finding was particularly insightful because:

- It challenged assumptions about progressive policies automatically leading to reduced rates
- Highlighted the complexity of mental health outcomes - even well-resourced healthcare systems struggle to reduce prevalence
- Jamaica's consistently lower rates, despite being in the pink cluster, suggested the importance of examining cultural and systemic factors beyond just treatment approaches

The analysis demonstrated that while clustering helped identify patterns, temporal analysis was crucial for understanding policy effectiveness.

Project Conclusions: Key Findings, Challenges, and Future Directions

The project's objective was to uncover global mental health patterns. The analysis revealed distinct regional clusters and highlighted how cultural, economic, and healthcare systems influence mental health outcomes.

Key Findings:

- Dark purple cluster showed highest disorder rates, primarily in developed nations
- Purple cluster revealed notable depression rates but lower prevalence of other disorders
- Pink cluster demonstrated highest schizophrenia rates but moderate-to-low rates overall

Challenges and Limitations:

- Data gaps - only North America had 100% coverage
- Cultural biases in self-reporting and stigma affecting data accuracy
- Manual data entry errors impacting reliability

Future Improvements:

- Expand data sources to include healthcare funding and professional access metrics
- Incorporate longitudinal analysis to establish intervention effectiveness
- Refine clustering by including socioeconomic factors
- Address cultural reporting biases through improved methodology