SOCIAL CONNECTION x PREVELANCE OF DEPRESSIVE & ANXIETY DISORDERS

Global prevalence of Depressive and Anxiety Disorders was evaluated in relation to Augmented HDI, government investment, cultural attitudes, coping mechanisms, and World Happiness Index, utilizing **Python** in **Jupyter**. Techniques such as **pair plots**, **correlation coefficient heatmaps**, **spatial** and **time-series analysis**, **linear regressions**, and **k-means clustering** were applied.

This case study details the analytical process undertaken and the skills developed throughout.
For immediate access to the findings and scripts, please click the links below:

TABLEAU STORY | GITHUB | STATIC VISUAL

DATASET LIMITATIONS:

Before delving into the analysis, it is essential to acknowledge the inherent limitations of the data used:

- **Reported prevalence of Depressive and Anxiety Disorders is derived solely from medically diagnosed cases:**
  - This dataset does not encompass the broader spectrum of individuals experiencing symptoms of depression and anxiety who have not received a formal diagnosis.
  - This limitation is significant as it directly affects the comprehensiveness of the data, influenced by factors such as accessibility to medical services and the presence of healthcare professionals trained in diagnosing these conditions.
- **Subjective nature of diagnosing Depressive and Anxiety Disorders:**
  - There is no universal diagnosis standard, thus these are based on subjective interpretations of symptoms described by patients, filtered through medical professionals.
- **Datasets were compiled from various sources, each originally produced for different purposes:**
  - The differing intentions and purpose across the datasets makes it challenging to fully standardize the data. Ideally, the data collection would have been designed specifically for this analysis to ensure standardization and relevance.

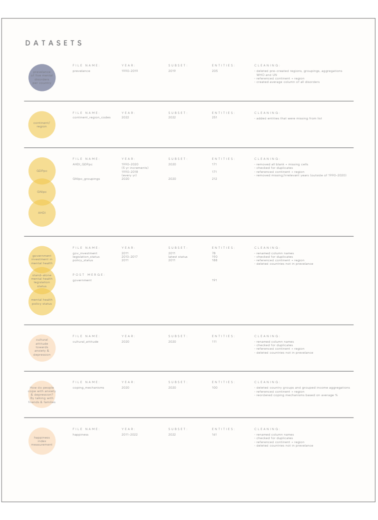
DATASETS:

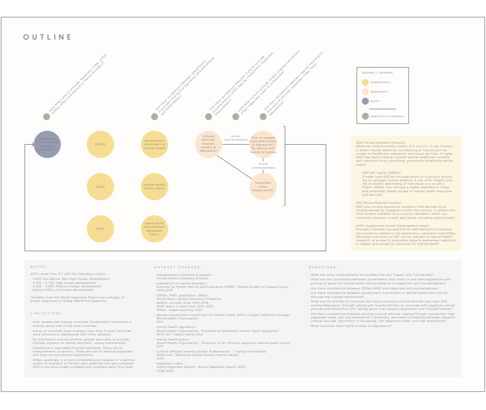
This project analyzed a compilation of the following datasets:

- Prevalence of Depressive & Anxiety Disorders per country
- continents / regions
- GDP per capita (GDPpc), GNI per capita (GNIpc), Augmented HDI (AHDI)
- government investment in mental health
- stand-alone mental health legislation status
- mental health policy status
- cultural attitude towards anxiety & depression
- coping mechanisms
- World Happiness Report

Collection methods and units of variables were reviewed carefully to understand the premise and limitations of each.

Click below for more details on the compiled datasets and their sources, and the outline and questions proposed for this analysis.





DATA PREPARATION:

The initial cleaning of the datasets was performed in Excel, while simultaneously imported into Jupyter, where each column and unique countries were counted, ensuring there was overlap between dataframes, and missing or duplicated data within each dataframe were addressed.

- As the datasets were sourced from different organizations, I faced a challenge of them having varied classifications and naming conventions for countries, regions, country groupings, and GDP groupings. To address this, a standardized list of countries, regions, country grouping names, and GDI groupings with standardized spelling and values was created, ensuring all overlapping countries were included.
- Datasets like GDP, AHDI, government investment, policy and legislation statuses were missing data for some countries. For a more complete and robust analysis, additional research was performed to fill in any gaps manually for each country.
- Some datasets included both countries and territories. To maintain consistency, territories were removed across all datasets.
- After the completion of the standardized list, classifications across all datasets were harmonized using VLOOKUP.
- Unnecessary or irrelevant columns, non-overlapping countries across datasets, and data outside the timeline scope were removed. All datasets were checked for duplicates.
- In datasets such as 'legislation_status', which covered a span of five years, some entries were missing data for the most recent year. To ensure its completeness, a 'last observation carried forward' imputation was utilized, substituting the missing values with the most recent available data for each country.

This preprocessing to increase data consistency and accuracy ensured subsequent clean merging and analysis in Python and prevented any issues that could have arisen from mismatched or misclassified data.

Once imported into Jupyter:

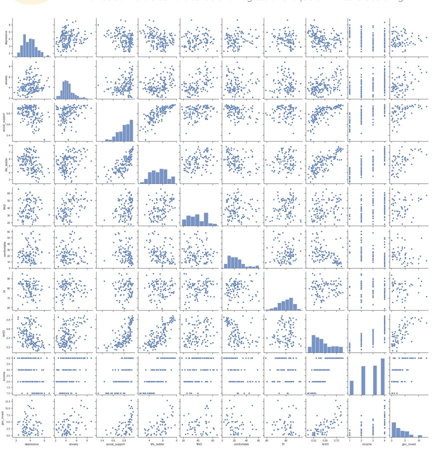
- Data types were reviewed and standardized
- All relevant columns were merged into one dataframe
- Categorical data ('yes', 'no'; GDP groupings) was encoded and altered from object to float64 in preparation for pair plot and correlation coefficient heatmap analyses
- Column names were altered for improved readability

DATA EXPLORATION:

Below reviews the tools that were used for data exploration. Click on images for a detailed view and view of scripts.

PAIR PLOT

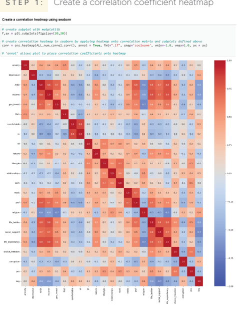
The exploratory data analysis was initiated by selecting a subset of variables from the original dataframe and creating a pair plot using Seaborn. This visualization facilitated a comprehensive look of the relationships between variable sets, allowing for an identification of discernible patterns. The insights gained from this broad overview informed decisions on which sets of variables were suitable for regression analysis or k-means clustering.



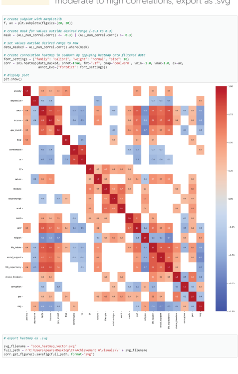
HEATMAP

Correlations between each set of variables were assessed by creating a correlation coefficient heatmap using Seaborn.

STEP 1: Create a correlation coefficient heatmap



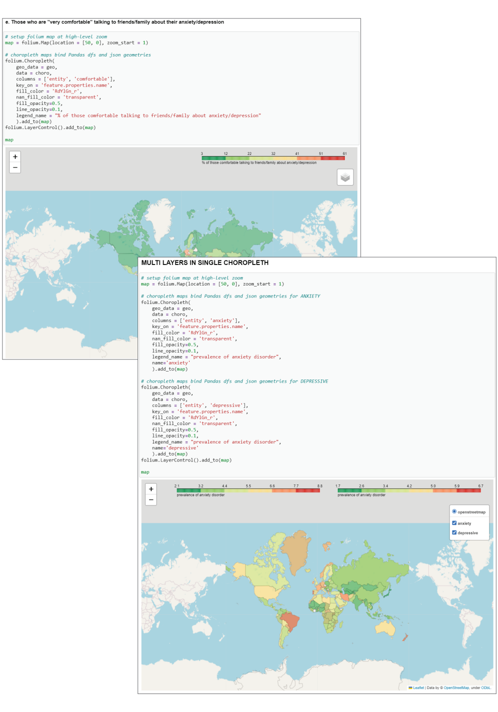
STEP 2: Mask low correlation coefficients to highlight moderate to high correlations; export as .svg



STEP 3: Import .svg into Adobe Illustrator; customize for improved readability and highlight takeaways for final report

CHOROPLETH

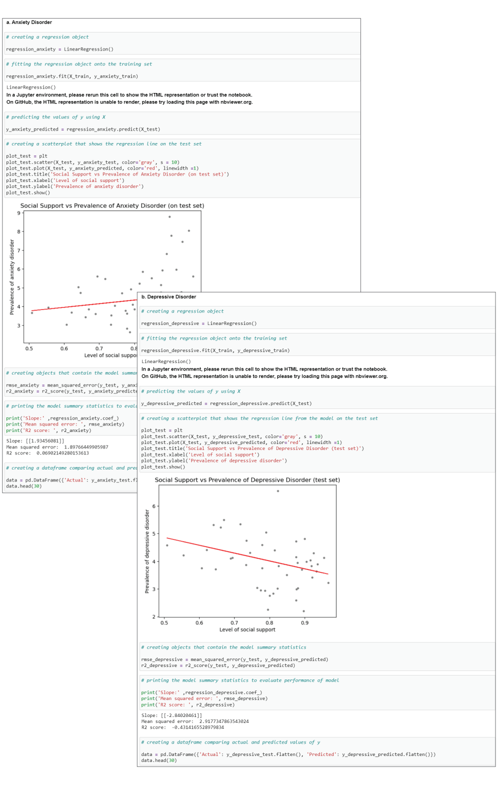
GeoJSON data was imported and choropleth maps of the prevalence of each disorder, World Happiness Indexes, and the different coping mechanisms were created using Folium.



REGRESSION ANALYSIS

Regression analysis was performed to test two hypotheses:

1. The higher the social support, the higher the prevalence of anxiety disorder.
2. The higher the social support, the lower the prevalence of depressive disorder.



ANXIETY DISORDER:

- Regression line for prevalence of anxiety disorder has a slope of 1.93, Mean Squared Error of 1.90 and an R2 score of 0.069.
- Slope of 1.93 shows a positive relationship between the level of social support and anxiety disorder.
- The Mean Squared Error is very high, meaning the distance between the regression line and the data points are large, indicating that it is not an accurate representation of the data.
- R2 score is very low, being close to 0. This indicates a poor fit. The model is not a close representation of the data. It is also clear that the relationship is not linear as there is a wider range of prevalence as the level of social support increases.

DEPRESSIVE DISORDER:

- Regression line for prevalence of depressive disorder has a slope of -2.84, Mean Squared Error of 2.92 and an R2 score of -0.43.
- Slope of -2.84 shows a negative relationship between the level of social support and depressive disorder.
- The Mean Squared Error is very high, meaning the distance between the regression line and the data points are large, indicating that it is not an accurate representation of the data.
- R2 score being a negative value, indicates a worse fit than a horizontal line. The model is not a close representation of the data.

BOTH DISORDERS:

- The limited number of data points (153; when 5,000 data points is usually desired to perform an accurate regression) could have negatively impacted the model's accuracy.
- As the respective prevalences depicted a non-linear relationship, a non-linear or curved relationship would have performed a more accurate regression.
- Both the correlation coefficient (0.3 for anxiety; -0.4 for depressive) and the scatterplots showed a lack of strong linear relationship between the respective variables, so a weak regression was expected.

PRESENTATION HIGHLIGHTS:

Correlations of anxiety and depressive disorders

Although only having weak to moderate correlations of 0.3~0.5, boxed in gray at the top highlights an interesting contrast of general patterns between anxiety and depressive disorders:

Anxiety disorder has a **positive** correlation, whereas

depressive disorder has a **negative** correlation

with high-resource variables; variables that are associated with high AHDI, income, and government investment.

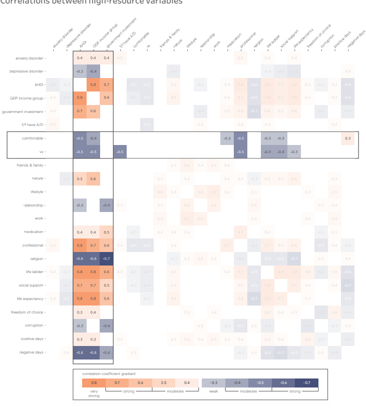
Furthermore,

countries with **higher AHDI** tend to have:
• **lower prevalence of depressive disorder** and
• **higher anxiety**, whereas

countries with **lower AHDI** tend to have:
• **lower prevalence of anxiety disorder** and
• **higher depression**

Let's take a look at these high-resource variables in more detail…

Correlations between high-resource variables



The vertical gray box in this heatmap highlights the moderate to very strong, **positive correlation** (>0.4 in orange) between AHDI, income groups, government investment, and the following high-resource variables:

• **going out to nature, taking medication, talking to a mental health professional** as coping mechanisms

• **life ladder**
(individual's score of their quality of life)

• **social support**
(score of whether or not they have friends and family they can turn to)

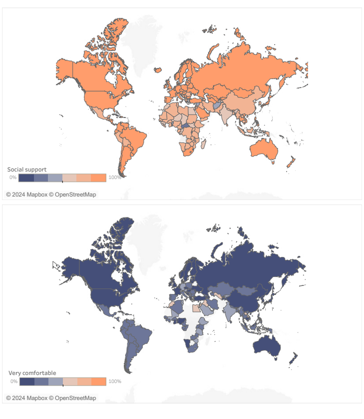
• **healthy life expectancy**

This highlights the resources available and accessible by countries with higher AHDI, income, and government investment.

Alternatively, in the horizontal gray box, the variable "comfortable", which represents the rate of people who feel *very comfortable talking to friends/family about anxiety/depression*, shows a **negative correlation** with these high-resource variables.

One would assume, with the **positive correlation** between the high-resource variables and social support, there would also be a **positive correlation** with feeling comfortable turning to their social support about anxiety and depression, but the data shows the contrary.

Social support vs feeling very comfortable



These maps show a clear contrast between the rate of **social support** and **feeling very comfortable** talking to friends and family about their own anxiety and depression.

Although the rate of **social support** is high (top), most of the world does **not** feel very comfortable talking to friends and family about their own anxiety and depression (bottom).

CHALLENGES:

Subjective nature, sample size and granularity:

- The dataset's broad scope, coupled with unaccounted societal biases, led to insightful yet non-definitive insights. A more granular approach may derive more detailed, actionable insights.
- In the case of the k-means clustering, the small sample size significantly hampered the reliability of the results, rendering them inconclusive.

Statistical analysis limitations:

Analyzing this dataset, which comprises of a mixture of subjective and partially qualitative data, posed challenges. I often felt constrained by my current analytical and statistical toolkit, underscoring the need for enhanced methods or tools better suited to handle the complexity of this data.

CONCLUSION:

There were very complicated biases present here that were difficult to account for. Self-reported and subjective diagnostic nature of mental disorders between different countries, how they are measured or surveyed, and stigmas associated with these disorders and coping mechanisms across cultures are likely to skew measurements.

One way to try to mitigate this would be through analyses divided into groupings. For example, creating cultural groups or similar biases or values, and analyzing them individually, although defining cultural groups within similar stigmas would present their own biases and challenges as well.

Due to the data's subjective nature and many levels of subjectivity in each dataset, and the challenges of standardizing datasets from multiple sources with varying collection methods, it was difficult to form strong, definitive conclusions.

One thing we can observe from this data is that stigma is strong and cross-cultural. High GDP and AHDI with the presence of high-resource variables do not curtail stigma. The data emphasizes the value and importance of social support in the face of mental health challenges. Stigma-reduction education may be a productive community implementation to build a more open and welcoming society that expands space for human experiences, validates a spectrum of mental conditions, and improves overall well-being.

SKILLS ACQUIRED:

Python Libraries:

- Pandas for data manipulation
- NumPy for numerical operations
- Matplotlib.pyplot and Seaborn for data visualization
- Folium for geographic visualization
- SciPy for scientific computing

Analyses:

- pair plot
- correlation coefficient heatmap
- geospatial analysis
- linear regression analysis
- k-means clustering
- time-series analysis and stationarying

Akiko Oguchi | Data Analyst | Information Designer | oguchiakiko@gmail.com