Feeding the Mind: How Diet Composition Relates to Population Mental Health

Zahra Makki (zahram)
Susan Hatem (susanhat)

---

1. Motivation

This project was inspired by a discussion between Dr. Andrew Huberman and Dr. Diego Bohórquez on the Huberman Lab podcast. In the episode titled "The Science of Your Gut Sense & the Gut-Brain Axis," Dr. Bohórquez delved into the psychology of the gut microbiome and its intricate relationship with diet. He recounted a compelling anecdote about a woman who, prior to undergoing weight loss surgery, found the thought of runny egg yolks repulsive. Following the surgery, however, she developed a strong craving for them, to the extent that she would eagerly finish every bite. This behavioral shift reflects deeper physiological mechanisms. As noted in the supporting literature, "satiety, food preference, and even mood behaviors are a few of the functions modulated by gut chemosensation" (1), underscoring the extensive influence of dietary composition on both physical and mental health outcomes. Essentially, the foods we consume are transformed into neural signals that affect virtually every aspect of our functioning.

This raises a compelling question: to what extent might the mental health outcomes of a population be influenced by its overall dietary patterns? In this project, we aim to explore this relationship at the global level by combining data from the Global Dietary Database (GDD) and the Global Mental Health Disorders dataset. Prior studies have shown links between dietary patterns and mental health outcomes. For example, Jacka et al. (2010) found that traditional Western diets were associated with lower rates of depression and anxiety. More recent reviews, such as Marx et al. (2021), explored the biological mechanisms that underlie these relationships, including the role of gut microbiota and systemic inflammation. However, few analyses have examined these patterns at the global level across countries and years, which is the aim of this project. In this report, we seek to identify broad trends and generate new insights into the diet-mental health connection by creating composite dietary scores and linking them to national mental health prevalence data.

2. Data Sources

Global Dietary Database (GDD)
The first data source used in this project is the Global Dietary Database (GDD). This dataset was obtained in CSV format (GDD Dec2020 Survey Metadata 1611c.csv) from the Global Dietary Database website. The GDD provides detailed dietary survey data from countries around the world, including information about the availability of various dietary factors. The original size was around 25,776 records for each country-year-food group combination, but after filtering for National-level surveys and years between 1990-2017 (inclusive), the final dataset contained 16,288 records.
Data location: https://www.globaldietarydatabase.org/
Format: CSV
Variables of interest:

- ISO3: 3-letter country code
- Year: Year of data collection
- Available dietary factors: A list of dietary components reported in each survey
- Sex: Gender of the survey population

Global Mental Health Disorders

The second data source used in this project is the Global Mental Health Disorders dataset. This dataset was obtained from a public Kaggle repository: https://www.kaggle.com/datasets/thedevastator/global-mental-health-disorders. It contains national-level prevalence estimates for multiple mental health disorders reported as percentages of the population. This dataset spans from 1990-2019, and after cleaning to remove rows with missing country codes and aggregate rows (e.g., OWID_WRL), the final dataset contained 1,130,580 records.

Data location: https://www.kaggle.com/datasets/thedevastator/global-mental-health-disorders

Format: CSV

Variables of interest:

- ISO3: 3-letter country code
- Year: Year of prevalence estimate
- Prevalence rates for seven mental health disorders:
    - schizophrenia
    - bipolar
    - eating_disorder
    - anxiety
    - drug_use
    - depression
    - alcohol_use

3. Data Manipulation Methods

This section will explain how we prepped and processed the data, the exploration that led to the decisions behind omitting certain columns, engineered a set of features to provide a basis for our correlational analysis, and merged the data sources for analysis.

Libraries Used

All data manipulation and analysis were performed using Python. The following libraries were used throughout the analysis:

Data Analysis Libraries:

- **pandas**: Used for reading, creating, and manipulating DataFrames. This library provided the core functionality for data loading, transformation, filtering, and exploratory analysis.
- **numpy:** Used for numerical computing and array operations. This library provided support for handling multi-dimensional arrays, performing mathematical computations, and enabling fast element-wise operations that complemented the functionality of pandas.

Visualization Libraries:

- **missingno** Used to visualize missing values within the data. The missingno correlation heatmap was particularly useful for identifying patterns of correlated missingness across variables.
- **matplotlib**: Used for generating general-purpose plots and visualizations throughout the analysis.
- **seaborn**: Used for creating heatmaps to visualize both correlation structures and missing data patterns within the datasets.

All code was documented and executed within Deepnote notebooks to ensure transparency, reproducibility, and full version control of the analytical workflow.

Global Dietary Database (GDD) Processing

Initial Import and National-Level Filtering

The Global Dietary Database (GDD) was first imported into a pandas DataFrame (gdd_df). The dataset includes a column for survey representativeness, with four levels: National, Local, Subnational, and MISSING. Since our mental health

dataset contains national-level prevalence rates, we restricted the GDD dataset to only national-level surveys conducted between 1990 and 2017 to ensure consistency across datasets.

Missing Data Assessment
We used both Missingno correlation heatmaps and Seaborn missingness heatmaps to explore patterns of missing data. The Missingno heatmap revealed strong correlated missingness between certain columns such as Youngest age and Oldest age, while other variables exhibited more independent missingness patterns. Seaborn heatmaps further highlighted data sparsity in columns such as FoodEx2 harmonization status, Sample size, and Available dietary factors. We also computed missing value counts and percentages for each column, which guided our decision to omit columns with excessive missingness while retaining variables necessary for robust analysis.

Age and Sex Filtering
We explored the possibility of filtering by both age and sex to match the granularity of the mental health dataset. While the GDD includes age ranges, we found that many surveys lacked complete or consistent age information. Following guidelines from the Global Burden of Disease (GBD) project, we restricted our analysis to adult populations (age 18+), which aligned with the scope of the mental health outcomes under study (6). Although the dataset included sex-specific estimates, over 40% of surveys lacked gender-specific data. As neither dataset could support sufficiently granular gender-specific analysis, we excluded the sex variable from further analysis and conducted all analyses at the population level.

Dietary Factor Extraction
The Year column in the original dataset is stored as a float. The value was converted to integer datatypes after handling non-numeric entries. The Available dietary factors column contained pipe-delimited strings, which were split into sets of individual dietary factors. These were then exploded into one row per dietary factor, with leading and trailing whitespace removed. Rows with missing dietary factor values or duplicates were dropped to prevent redundancy.

Global Mental Health Dataset Processing

Initial Import and National-Level Filtering
The Global Mental Health Disorders dataset was read into a pandas DataFrame (mental_health_df). The Year column was cleaned and converted to integers. Rows with missing or invalid year values were dropped. To ensure consistent merging with the GDD data, we renamed the Code column to ISO3, which is the shared country identifier. We removed rows with missing ISO3 values, as well as rows where ISO3 was set to ‘OWID_WRL’, which represents global aggregate data not tied to any specific country.

Column Renaming
For ease of analysis, several columns were renamed to more concise variable names. This facilitated both readability and code implementation during modeling.

Missing Data Assessment
As with the GDD dataset, we visualized missingness using both Missingno and Seaborn heatmaps. The missingno heatmap showed perfect correlations in missingness between several mental health outcomes (anxiety, depression, drug_use, and alcohol_use), indicating that when one variable was missing, the others were missing as well. In contrast, eating disorders exhibited less correlated missingness, suggesting it was more independently reported across country-year pairs. Seaborn heatmaps provided a row-wise visualization, further highlighting contiguous blocks of missingness for certain disorders. These visualizations informed our decision to handle missingness on a per-outcome basis, rather than using a complete-case approach that would have resulted in excessive data loss.

Data Sparsity

Exploratory analysis showed that applying a global complete-case filter would reduce the dataset from 102,780 observations to only 5,460, representing significant information loss. To maximize the usable data, we filtered missing values separately for each disorder during correlational analyses, allowing each mental health outcome to retain the maximum available data for its specific model.

Value Conversions

The mental health dataset contained a large amount of variation in the values populated for the mental health outcome percentages. For example, the schizophrenia outcome column could contain values such as 0.33 and 24539.72. For correlational analysis, we wanted to make the data comparable on the same scale of 0 - 1, representing percentage values and matching the scale that the engineered features were scored on. To achieve this, we did the following: (1) Drop all numbers that we could not infer any percentage values from. In this case, any number greater than 100; (2) Any number on a scale of 1 - 100 was divided by 100; and (3) Any number between 0 and 1 was unchanged.

4. Feature Engineering

In order to analyze how diets may be correlated to mental health outcomes, the following set of features were created: Plant Based Score, Animal Based Score, Processed Food Score, and Diversity Score. The focus is dietary consistency, and these scores allow for some categorization to give a basis of comparison to various mental health outcomes to give context to interesting patterns. Before categorizing each food item for comparison, all vitamins included were removed. This is due to vitamins not fitting any of the new features and having generally sparse appearance in the dataset.

Plant-Based and Animal-Based Scores

Plant-Based and Animal-Based scores are meant to measure how many plant based items and animal based items are featured in a countries diet, respectively. To measure the feature, two sets were created based on available dietary data:

- **Plant-based food features:** Fruits, Non-Starchy Vegetables, Beans and Legumes, Nuts and Seeds, Fruit Juice, Dietary Fiber, Whole Grains, Plant Omega-3 Fat, Plant Protein.
    - ❖ **Plant Based Score:** Plant Count / Total Number of Plant-Based Factors
- **Animal-based food features:** Unprocessed Red Meats, Total Seafoods, Seafood Omega-3 Fat, Dietary Cholesterol, Total Milk, Cheese, Yogurt, Eggs, Whole Fat Milk, Reduced Fat Milk, Total Processed Meats, Dairy Protein, Animal Protein, Total Animal Protein.
    - ❖ **Animal Based Score:** Animal Count / Total Number of Animal-Based Factors

This score was initially going to be a composite score on a scale of -1 to 1, with -1 being animal based and 1 being plant based. Separating these features out allows us to measure the impact of these specific factors on a diet independent of one another. The new scale was modified to a 0 to 1 scale to match the scales of the remaining features.

Processed Diet Score

Processed food consumption scores range on a scale of 0 - 1, with 0 representing a completely processed diet and 1 representing a completely unprocessed diet. The following feature sets were created and the score was then computed using the following formula.

- **Processed food features:** Added Sugars, Sugar-Sweetened Beverages, Fruit Juice, Refined Grains, Unprocessed Red Meats, Total Processed Meats, Total Animal Protein, Total Energy, Saturated Fat, Trans Fatty Acid, Dietary Cholesterol, Total Carbohydrates, Dairy Protein, Whole Fat Milk, Reduced Fat Milk, Cheese, Glycemic Index, Glycemic Load, Other Starchy Vegetables, Potatoes, Monounsaturated Fat, Animal Protein.
- **Unprocessed food features:** Fruits, Non-Starchy Vegetables, Beans and Legumes, Nuts and Seeds, Whole Grains, Plant Protein, Plant Omega-3 Fat, Seafood Omega-3 Fat, Dietary Fiber, Dietary Sodium, Total Seafoods, Yogurt, Eggs, Total Protein, Coffee, Tea.
- ❖ **Processed Diet Score:** Unprocessed Count / (Unprocessed Count + Processed Count)

Diversity Score
Finally, a dietary diversity score was calculated. The set was a union of the Processed and Unprocessed food features. Following a similar pattern as the other features, the score was calculated using:

- **Diversity Score:** Unique Factors / Total Factors

This follows the same 0 - 1 scale as the other features, with a highly diverse diet being close to 1.

5. Model Selection

Our primary goal was to explore potential relationships between dietary patterns and national-level mental health outcomes, not to predict or classify outcomes per se. Therefore, we focused on correlation-based exploratory analysis rather than fitting more complex statistical models.

We initially considered implementing logistic regression models to predict the presence or absence of prevalence rates for various mental health outcomes based on dietary patterns. However, we decided against this approach because our outcome variables were continuous prevalence rates (proportions), not binary outcomes, making logistic regression a less natural fit. Our primary goal was to explore general patterns of association across countries and dietary patterns, rather than to produce predictive models or infer causal effects.

Instead, we selected a correlation-based analysis strategy. We computed Pearson correlation coefficients between engineered dietary scores and the various mental health outcome prevalence rates. This allowed us to quantify the strength of the associations without imposing a specific model structure.

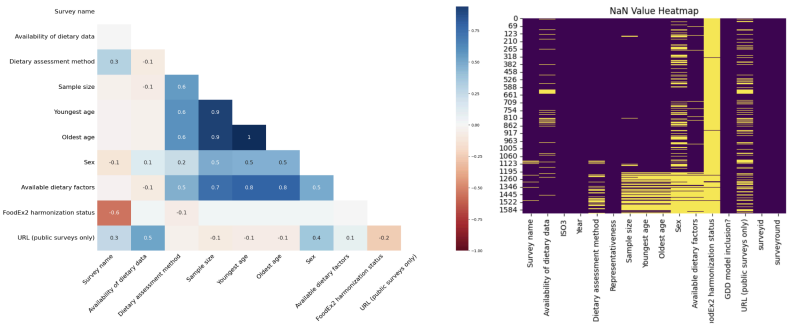
For our correlation analysis, we implemented two visualization techniques, correlation matrices and scatter plots. Using the plot_outcome_correlations helper function, we generated correlation heatmaps between dietary features and mental health outcomes. We also used the plot_scatter_with_correlation helper function to create scatter plots for each feature-outcome pair.

6. Results

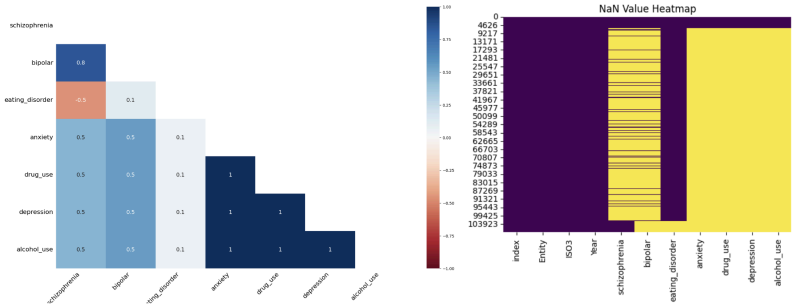
NaN Exploration
As a result of the NaN analysis on both the gdd_df and the mental_health_df DataFrames, the high sparsity of the datasets led to a modified approach to the correlation analysis.

In the Global Dietary Database (GDD), the Missingno correlation heatmap revealed highly correlated missingness between the columns Youngest age and Oldest age (correlation of 0.9 to 1.0). There was also moderate correlation between Available dietary factors and variables such as Sample size, Youngest age, Oldest age, and Sex. The Seaborn NaN heatmap showed that Youngest age, Oldest age, and Available dietary factors contained substantial missing values across country-year pairs. Other variables, such as Sample size and Sex, exhibited more scattered missingness.

Based on these patterns, we decided to retain only the columns necessary for constructing our dietary scores. We also limited our analysis to adult populations (age 18+), following the Global Burden of Disease (GBD) framework. This decision was made because the diagnosis and reporting of many mental health disorders differ between children and adults, with most disorders typically being diagnosed and reported in adult populations. Focusing on adults allows us to ensure greater consistency and comparability across countries in our merged dataset. However, the Mental Health Disorders dataset did not provide age-specific breakdowns, so this required us to assume that the dataset primarily reflected prevalence in adult populations.



In the Mental Health Disorders dataset, the Missingno correlation heatmap showed how the variables anxiety, depression, drug_use, and alcohol_use exhibited perfectly correlated missingness (correlation = 1.0), meaning that when one was missing, all were missing for that country-year pair. This was in contrast to eating_disorder, which displayed more independent reporting as it had lower correlations with other outcomes. The Seaborn NaN heatmap confirmed this: large contiguous blocks of missing data were present for anxiety, depression, drug_use, and alcohol_use, while eating_disorder, bipolar, and schizophrenia had a more variable distribution of missing values.



These findings encouraged us to avoid excessive data loss by observing the relationships between these disorders differently. We decided to create and analyze three different datasets to maximize the available data for each analysis. First, we created a clean Eating Disorder (ED) dataset by dropping rows with missing values for eating_disorder and keeping only the ISO3, Year, and eating_disorder columns. This allowed us to explore the relationship between dietary scores and eating disorder prevalence without being affected by missing data in other columns. Second, for schizophrenia, we created a balanced subset by sorting the schizophrenia column and selecting both the first 5,000 and the last 5,000 rows (lowest and highest prevalence values). This ensures that our schizophrenia analysis was not biased toward a particular portion of the distribution and allows us to capture variation across the range of reported prevalence. Finally, we constructed a third dataset focused on all the other mental health outcomes (bipolar, anxiety, drug_use, depression, and alcohol_use) by excluding schizophrenia and eating_disorder and selecting the first 5,000 rows. This provided a clean and manageable subset for exploring correlations between these five outcomes and our engineered dietary scores. By structuring the data in this way, we were able to perform focused analyses on each outcome of interest while minimizing the impact of the strong patterns of correlated missingness present in the full dataset.

Correlation Matrices

To explore the relationships between our engineered dietary scores and mental health outcomes, we computed Pearson correlation matrices between each mental health outcome and the four primary dietary features: plant-based score, animal-based score, processed diet score, and dietary diversity score.

In the full population dataset (diet_mental_health_df), dietary diversity score showed positive correlations with anxiety (r = 0.28), eating disorders (r = 0.25), and depression (r = 0.12). In contrast, processed diet score showed small negative correlations with eating disorders (r = -0.18), anxiety (r = -0.13), and depression (r = -0.066). Correlations between processed diet score and other outcomes were generally weak or negligible. The plant-based score also showed weak correlations across most outcomes, but it exhibited small positive associations with eating disorders (r = 0.24), anxiety (r =

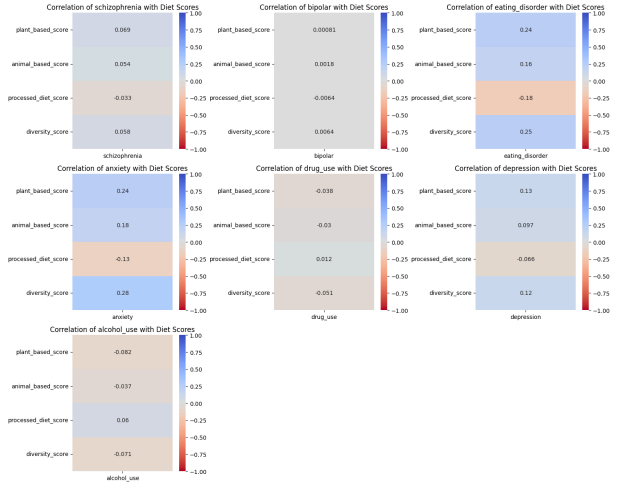
0.24), and depression (r = 0.13). The animal-based score was similar to the plant-based. For outcomes such as drug use, bipolar disorder, schizophrenia, and alcohol use, correlations with all the dietary scores were generally very small.

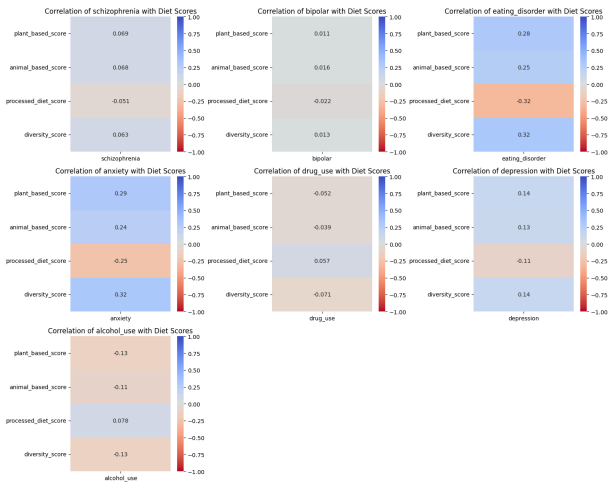
When we repeated this analysis on the adults-only subset (age 18+), the overall patterns were broadly consistent but some associations became stronger. For example, dietary diversity scores continued to show positive correlations with anxiety (r = 0.32), eating disorders (r = 0.32), and depression (r = 0.14). The negative association between processed diet score and anxiety (r = -0.25) and with eating disorders (r = -0.32) also increased. The plant-based and animal-based score again showed small positive correlations with anxiety (r = 0.29 and r = 0.24), eating disorders (r = 0.28 and r = 0.25), and depression (r = 0.14 and r = 0.13). For outcomes such as drug use, bipolar disorder, schizophrenia, and alcohol use, correlations with all dietary scores remained generally weak or near zero (absolute r < 0.15).

To further explore potential differences by outcome and data completeness, we conducted targeted analyses on three additional subsets of the data. For the eating disorder (ED) subset, the diversity score remained positively correlated with eating disorders (r = 0.30), the processed diet score showed a stronger negative correlation (r = -0.30), and the plant-based score showed a modest positive association (r = 0.25). This was consistent with the patterns observed in the full and adults-only datasets.

In the schizophrenia subset, correlations between dietary features and schizophrenia prevalence were generally weak. Diversity score showed a small positive correlation (r = 0.18), while processed diet score (r = -0.06) and plant-based score (r = 0.055) had minimal associations.

Finally, in the adults-only subset for the remaining mental health outcomes (bipolar disorder, anxiety, drug use, depression, alcohol use), correlations remained generally weak to moderate. The processed diet scores continued to show a negative association with anxiety (r = -0.17) and bipolar disorder (r = -0.12), while diversity score was positively correlated with drug use (r = 0.13) and anxiety (r = 0.10). Other correlations were small and inconsistent across outcomes.





Scatterplots

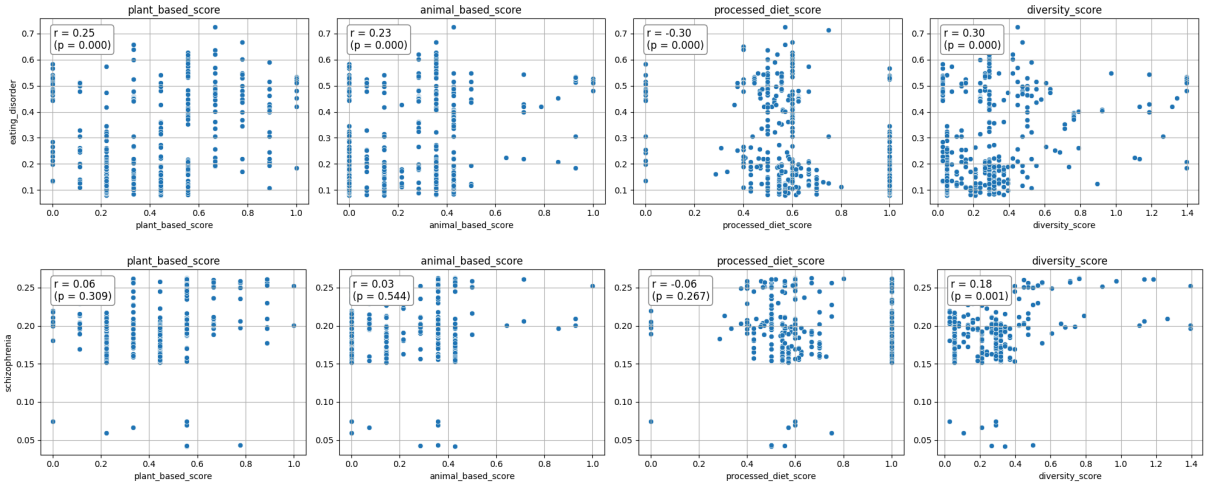
To further visualize these relationships, we created scatterplots with Pearson correlation coefficients (r) and significance values (p) to display the associations between each dietary score and each mental health outcome. In addition to the correlation matrices, these scatterplots allowed us to qualitatively assess the variability and strength of the observed associations.

Across all datasets, the scatterplots showed weak to moderate associations between the dietary scores and mental health outcomes. None of the relationships showed strong correlations (r > 0.5) and the scatter of the data points showed high

variability. While several correlations reached statistical significance ($p < 0.05$), the corresponding r values were too small to suggest any strong relationships peresent.

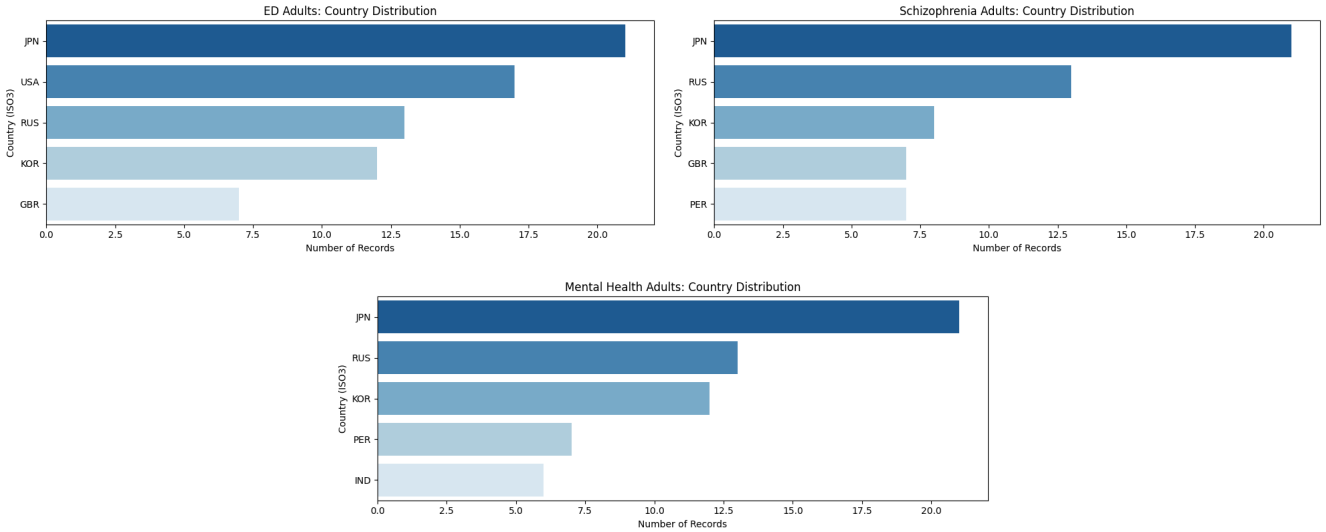
The most consistent patterns were observed for eating disorder symptoms and anxiety. They showed higher dietary diversity and plant-based scores and lower processed diet scores. Scatterplots for other outcomes, such as depression, schizophrenia, bipolar disorder, drug use, and alcohol use, showed weaker and less consistent associations.

Overall, while there were some associations between dietary diversity and mental health disorders as eating disorders and anxiety, the scatterplots highlighted high variability. This is why it is important to take caution in interpreting these results; the visual patterns supported the conclusions that dietary factors are associated with certain mental health outcomes, but the observed relationships are modest and likely influenced by many additional factors that need further study.



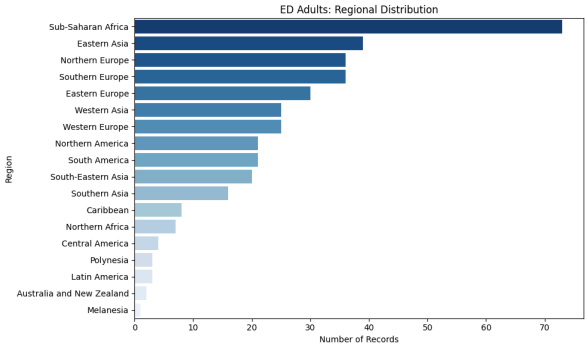
Country Distribution

The initial results highlighted the need for additional context. To better interpret the findings, we created two supplementary visualizations: one to display the distribution of countries, and another to display the distribution of regions. These visualizations were generated separately for each dataset to assess which countries and regions were most represented in the final analytic sample. To get the distribution of regions, a set that maps all ISO3 codes to the region the country belongs to was generated. The distribution of both the individual and regional distributions were plotted using a bar chart.



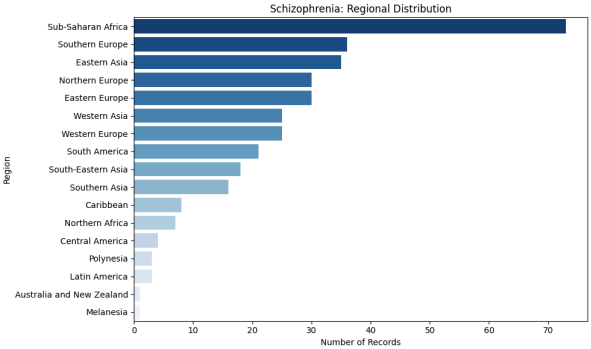
Eating Disorder Dataset Results

At the country level, Japan had the highest number of records, followed by the United States and Russia. Regionally, the top three distributions were Sub-Saharan Africa, Eastern Asia, and Northern Europe. It's important to note that regional distributions included all countries, and many Sub-Saharan African countries only had a single entry, likely contributing to the skew. This dataset was the most complete among the mental health outcomes analyzed, with approximately 95% of values present and only about 5% missing.



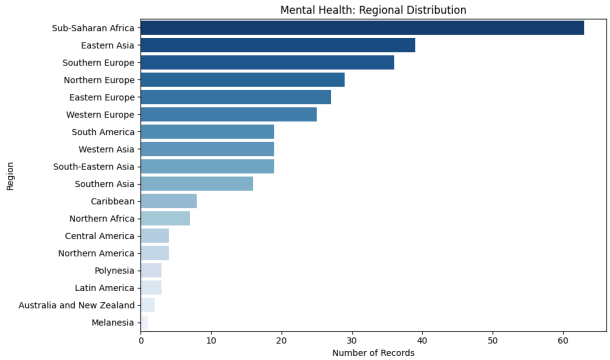
Schizophrenia Dataset Results

Japan had the highest number of records, followed by Russia and Korea. The largest regional distributions were Sub-Saharan Africa, Southern Europe, and Eastern Asia. This dataset contained a significant amount of invalid data that had to be removed.



Mental Health Dataset Results

The top three countries were Korea, Japan, and the Netherlands. Regionally, the highest distributions were in Sub-Saharan Africa, Eastern Asia, and Southern Europe. It was surprising to find that Northern America had one of the lowest representation levels in this dataset.



7. Discussion

There were no strong correlational relationships found between any of the engineered dietary features or the mental health outcomes we chose to study. This could be due to a large variety of factors. The datasets required a lot of cleaning due to inconsistent data. This loss of data could have drastically changed the trends we would be able to expose others to. To put

it simply, we cannot say anything definitive about the relationship between diet and mental health outcomes as a result of this study.

Limitations
The datasets were quite sparse for some columns, particularly the mental health outcome columns in the Mental Health data. This data loss, combined with the uneven distribution of countries represented, makes it difficult to fully assess the impact of diet. Additionally, the absence of a control group prevents us from drawing any causal conclusions. The datasets also lacked granularity. While diet may influence mental health outcomes, many other factors-- including economic, familial, and biological variables-- also contribute to their development.

Comparisons between the different mental health datasets are also challenging. Each outcome, such as Eating Disorders or Schizophrenia, has different sample sizes and varying distributions of country and regional data, which may introduce bias into the results. Furthermore, the data exhibited significant variability, and some normalization decisions may not fully capture what the data was representing.

8. Conclusion

This exploration has the potential to guide dietary choices based not only on physical, but mental health impact as well. While the data did not allow for anything conclusive the trial and error could be utilized as the basis for a more concerted effort to explore this question.

What We Would Do Different
In hindsight, we would have focused on a narrower population, such as a single country or region. This approach would have allowed for more in-depth analysis and better control over confounding variables, providing clearer insights into the relationship between diet and mental health outcomes.

Initially, we wrote all functions directly within the Deepnote notebook. It wasn't until later in the project that we refactored the code, moving the scoring and analysis functions into separate Python scripts and importing them into the notebook. This refactoring made additional analyses and bug fixes much faster and more efficient-- a structure that should have been implemented from the start. We also would have benefited from implementing version control early on.

At the beginning, we overlooked several inconsistencies in the data. It was only after closely examining data types that we discovered issues, such as incorrect year values and wide variations in the scales of some mental health outcome columns. Spending more time upfront to carefully review and understand the data would have made the analysis phase much smoother.

Next Steps
To gain better insight into the correlation between diet and mental health outcomes, it would be valuable to analyze more granular data. This could include factors already known to influence mental health, such as gender, economic status, family history, and many other potential contributors. Including these variables in the analysis would help identify which factors to control for in a more detailed study.

Additionally, to further investigate this relationship, one could design an experiment involving a control group and a treatment group. By monitoring how changes in diet impact mental health outcomes over time-- while controlling for other contributing factors-- it would be possible to draw stronger conclusions about potential causal links.

9. References

1. Berk, M., Williams, L. J., Jacka, F. N., O'Neil, A., Pasco, J. A., Moylan, S., ... & Maes, M. (2013). So depression is an inflammatory disease, but where does the inflammation come from? The American Journal of Psychiatry, 170(9), 891–899. https://pmc.ncbi.nlm.nih.gov/articles/PMC3846682/
2. Bohórquez, D. V., & Liddle, R. A. (2021). The gut-brain axis: Functional implications in the control of feeding. Annual Review of Neuroscience, 44, 277–295. https://doi.org/10.1146/annurev-neuro-091619-022657
3. Huberman Lab. (2023, August 14). Dr. Diego Bohórquez: The science of your gut sense & the gut-brain axis [Podcast]. Huberman Lab. http://hubermanlab.com/episode/dr-diego-bohorquez-the-science-of-your-gut-sense-the-gut-brain-axis
4. Institute for Health Metrics and Evaluation. (n.d.). Global Burden of Disease (GBD). IHME. https://www.healthdata.org/research-analysis/gbd
5. Jacka, F. N., Pasco, J. A., Mykletun, A., Williams, L. J., Hodge, A. M., O'Reilly, S. L., Nicholson, G. C., Kotowicz, M. A., & Berk, M. (2010). Association of Western and traditional diets with depression and anxiety in women. The American Journal of Psychiatry, 167(3), 305–311. https://doi.org/10.1176/appi.ajp.2009.09060881
6. Kaelberer, M. M., Buchanan, K. L., Klein, M. E., Barth, B. B., Montoya, M. M., Shen, X., & Bohórquez, D. V. (2018). A gut-brain neural circuit for nutrient sensory transduction. The Journal of Clinical Investigation, 128(2), 556–570. https://doi.org/10.1172/JCI78361
7. Marx, W., Lane, M. M., Hockey, M., Aslam, H., Berk, M., Walder, K., Borsini, A., Firth, J., Pariante, C. M., Berding, K., Cryan, J. F., Clarke, G., Craig, J. M., Su, K.-P., Mischoulon, D., Gomez-Pinilla, F., Foster, J. A., Cani, P. D., Thuret, S., Staudacher, H. M., Sánchez-Villegas, A., Arshad, H., Akbaraly, T., O'Neil, A., Segasby, T., & Jacka, F. N. (2021). Diet and depression: Exploring the biological mechanisms of action. Molecular Psychiatry, 26(1), 134–150. https://doi.org/10.1038/s41380-020-00925-x

10. Statement of Work

Susan Hatem: I was responsible for the research and analysis of the Mental Health dataset. My work included feature engineering, normalization of the dataset, and ensuring code quality and maintainability by refactoring repeated code into modular Python scripts, which were imported into the notebook environment. I also conducted the analysis of missing data and developed visualizations to assess country-level and regional distributions within the dataset. Additionally, I collaborated closely with Zahra throughout the project, providing mutual support in interpreting trends, resolving technical issues, and addressing data anomalies.

Zahra Makki: I was responsible for the initial research of the GDD/diet dataset, as well as the initial cleaning of the data and baseline exploratory analysis. This included splitting of the Available Dietary Factors, the discovery of “dirty” data in the mental health dataset, and exploring the potential use of vitamin data, which was ultimately not incorporated into the final analysis. I also worked on merging and cleaning the datasets, and conducted the correlational analysis using both correlation matrices and scatter plots with correlation coefficients. Once we decided to split the three datasets to maximize the quality and completeness of the data for each specific mental health outcome, I was responsible for splitting and cleaning these subsets to prepare them for further analysis. In addition, Susan and I collaborated through the final walk through and clean up of code, as well as the analysis and the writing of the paper.