A New Light on Depression:

A study of the effect of daylight hours in winter on self-perceptions of depression

Authors: Vikram Vaddamani, Alexis Parker, Peter King

Abstract

This study explores the relationship between winter daylight and self-reported depressive symptoms, using data from the 2022 Behavioral Risk Factor Surveillance (BRFSS) along with a compiled dataset on daylight exposure. A Depression Index (DI) was developed to quantify self-assessed mental well-being, allowing for an analysis of how variations in daylight correlate with reported depressive symptoms across different regions of the United States.

Results indicate a statistically significant yet weak negative correlation, suggesting that longer daylight hours are associated with a slight decrease in reported depressive symptoms. While these findings do not establish a direct causal relationship, they contribute to the broader discussion on environmental influences on mental health. This study highlights the value of incorporating measurable physical factors into psychological research and reinforces the need to further explore environmental effects on mental well-being.

Motivation

Humans, which we consider to be biological organisms consisting of interconnected colonies of roughly 30 trillion communicative cells and another 38 trillion bacteria, are extremely complex entities (Sender, et. al.). The human brain, with its roughly 86 billion neurons and 100 trillion neural connections, is in the modern era generally regarded as responsible for producing a wide range of thought, feeling, mood, and observed behavior, all of which in turn are also extremely multidimensional and complex (Caruso).

In Western civilization, modern psychological methods to investigate, categorize and explain the inner workings of the mind and associated range of human behavior date back primarily to the 19th century with the rise of experimental psychology in Europe, although there is evidence that globally, humans were thinking about the human mind in our earliest known civilizations (Hergenhahn). Investigative methods have varied widely over recorded history, reflective of the wide variety of underlying theories of mechanism of the mind, and constrained by the technologies available to gather information and conduct scientific inquiry.

In the last two centuries, one dominant thread of inquiry relied primarily on a dialogue between a trained professional (a psychologist) and a patient. Analysis of such dialogues culminated in the Diagnostic and Statistical Manual of Mental Disorders (DSM), which at present sets the standard for professional evaluation of mental health conditions. It is primarily through this dialogue, coupled with observations of manifest behavior (e.g., stuttering, eye contact avoidance) that mental disorders are professionally diagnosed in the present day.

In recent decades, however, there has been a growing trend to incorporate advanced technologies such as electroencephalogram (EEG) recordings and genetic profiling into the mix of observable human characteristics. Unlike the interpretation of dialogue and displayed behavior, which relies on the subjective evaluation of a professional, such "physical world" information is (perhaps erroneously) regarded as more objective, concrete, and unbiased. It is in this vein of a more modern, biologically-based approach to the investigation of human behavior -- at the intersection of the physical world with the abstract world of thought, consciousness, and the mind -- that we introduce our current inquiry.

Our aim in this study was to investigate the influence of daylight hours during winter on an individual's self-perceptions of depression. Daylight (i.e., sunlight) is a "physical world" variable that can influence human mood (an "abstract world" variable) through its effect on melatonin levels in the human eye. Melatonin is widely regarded as responsible for regulating the circadian rhythm and the sleep-wake cycle (Khullar). Disruptions in the sleep-wake cycle have been linked to alterations in mood, including depression (Walker, et. al.).

Our initial belief was that reduced periods of daylight in winter, as seen in the northern latitudes in the northern hemisphere, would show correlation with an increase in self-perceptions of depression among individuals living in the northern latitudes. We did not propose to prove a chain of causation in this study -- we merely sought to investigate whether a correlation exists between daylight hours in winter and feelings of depression as reported by individuals themselves (as opposed to a clinical diagnosis). Our underlying goal was to demonstrate that it is possible to quantify and measure a relationship between a "physical world" variable (daylight) and an "abstract world" variable (a "Depression Index" that we will develop, based on self-perception).

Our hope was that, in demonstrating a connection between the "physical world" of, for example, photons, and the "abstract world" of thought and feeling, we would contribute to the body of evidence that suggests an urgent need to transform the discipline of psychology once again for our present time. The method of doctor-patient dialogue that informs the DSM, and at present provides the foundation for professional classification and diagnosis of mental health disorders, is limited in effectiveness due to its reliance on centuries-old concepts and technology. It is outdated and desperately in need of an overhaul. This small demonstration should help reinforce the idea that humans are complex biological and physical organisms -- that their thoughts and feelings are deeply intertwined with physical and biological factors. Our ultimate goal is not merely academic: we believe that in order to improve classification, diagnoses, and treatment of mental health disorders, the ossified thinking around psychology that persists in the present day needs to give way to a new paradigm grounded in more objective measures.

Literature Review

In a study investigating correlation between latitude and the annual prevalence of Major Depressive Episode (MDE) in Canada, Patten, et. al., found that, "In models including latitude as a continuous variable, a statistically significant association was observed, with prevalence increasing with increasing latitude. This association persisted after adjustment for a set of known risk factors. The latitude gradient was modest in magnitude, a 1% to 2% increase in the prevalence odds of MDE per degree of latitude was observed."

Mersch, et. al., conducted a comprehensive review of the literature regarding the impact of latitude on Seasonal Affective Disorder (SAD), a condition distinct from but related to depression. They concluded that, "Over all prevalence studies, the correlation between prevalence and latitude was not significant. A significant positive correlation was found between prevalence and latitude in North America. For Europe there was a trend in the same direction."

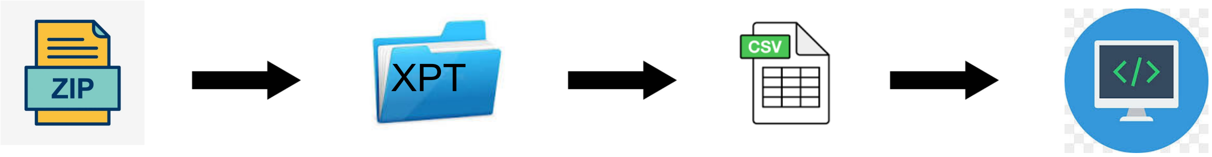
Data Sources

We used two datasets for this project: 1) the Behavioural Risk Factor Surveillance System (BRFSS) 2022, and 2) the Daylight dataset.

The BRFSS is an annual health-related survey conducted by the Centers for Disease Control and Prevention (CDC) to collect self-reported data from adults across all 50 U.S. states, the District of Columbia, and U.S. territories[1]. As the largest continuously conducted health survey in the United States, BRFSS tracks health behaviors, chronic conditions, and healthcare access to inform policy decisions. The 2022 dataset includes responses on lifestyle habits, self-reported depression, and healthcare interactions.

The second dataset (Daylight) provides a rough measure of the amount of daylight an individual had available to them during the winter season in 2022. This dataset is itself a combination of three separate data sources.

Acquisition of the BRFSS:
The 2022 BRFSS dataset was obtained from the CDC BRFSS Data Portal, where publicly available health survey data is released annually[2]. The dataset was downloaded as a ZIP file containing a SAS Transport (.XPT) file, a format commonly used for storing and sharing large survey datasets, including public health data. While XPT files can be read directly in Python, we converted the dataset to CSV format using the *pyreadstat* library for convenience and accessibility, ensuring a more seamless workflow for preprocessing and exploratory analysis.



Acquisition of Daylight data:
To assemble the Daylight dataset, we first combined data from two *plain text* files that we downloaded from a publicly available website[3]. These gave us a list of the 50 U.S. states (plus Puerto Rico and D.C.) along with their capital cities and the latitude and longitude of each capital city. We chose to focus on capital cities in part because we did not have specific geographic location data for each individual surveyed in the BRFSS, and we wanted to choose areas that would typically be population centers. Given more time, we could have obtained a third dataset showing population distribution/density for each state and used this to obtain a more realistic view of daylight available to each individual.

[1] Centers for Disease Control and Prevention. "Behavioral Risk Factor Surveillance System (BRFSS)." Available at: https://www.cdc.gov/brfss/

[2] Centers for Disease Control and Prevention. "Behavioral Risk Factor Surveillance System (BRFSS) 2022 Data." Available at: https://www.cdc.gov/brfss/annual_data/annual_2022.html in SAS transport format, https://www.cdc.gov/brfss/annual_data/2022/files/LLCP2022XPT.zip

[3] These two plain text files can be downloaded directly using the following URLs:
- https://people.sc.fsu.edu/~jburkardt/datasets/states/state_capitals_name.txt
- https://people.sc.fsu.edu/~jburkardt/datasets/states/state_capitals_ll.txt

We then used the U.S. Naval Observatory's (USNO) API to request sunrise and sunset times for each capital city (using the lat/long) on the Winter Solstice in 2022 (i.e., 21 Dec 2022)[4]. The API returns information in JavaScript Object Notation (JSON) format. Instead of focusing on the Winter Solstice as a single data point, we could have gathered sunrise/sunset data for an entire year and then selected a 3-month window as the "winter season." However, this more cumbersome process would have ultimately resulted in a very similar independent variable -- and one with less variation than a variable based on the Winter Solstice alone.

Data Manipulation Methods

A note about <u>File Management</u>:

- We developed our original code in a Google Colab environment.
- We used a shared folder from a team member's Google drive to store all notebooks and data files as we developed them.
- Prior to submitting this report, we changed the default filesystem path for all notebooks so that our code would work in the Jupyter environment provided via Coursera (as part of the University of Michigan's Master of Applied Data Science (MADS) program).

The BRFSS dataset:

<u>Data Preprocessing</u> for the BRFSS dataset:

- Converted the XPT file to CSV format using Python to enable data manipulation in Pandas.
- Performed an initial examination of the dataset to identify missing values, redundant columns, and categorical variables requiring transformation.
- Dropped irrelevant columns unrelated to the analysis to streamline the dataset.
- Standardized categorical variables by replacing coded numeric responses (e.g., *1* = *Yes*, *2* = *No*) with meaningful labels.
- Filtered out responses categorized as “Don’t know,” “Refused,” or blank values to ensure only valid responses were retained.
- Checked for missing data across key variables and evaluated multiple imputation strategies before finalizing an approach[5].

<u>File management</u> for the BRFSS dataset:

- Stored the processed CSV file in a shared Google Drive, specifically in the “data” folder, ensuring accessibility and ease of use for the entire team.

<u>Exploring</u> the BRFSS dataset:

- Analyzed survey responses to understand the distribution of depression indicators across states.
- Examined the distribution of DI scores by state, identifying variations and potential outliers.[6]
- Generated visualizations to illustrate regional differences in DI scores, including a choropleth map and a boxplot.

[4] A description of how to use the USNO API can be found at https://aa.usno.navy.mil/data/api
Reference the enclosed Jupyter Notebook ("2-K-Fetch_daylight_data.ipynb") for the exact code used to query the API.

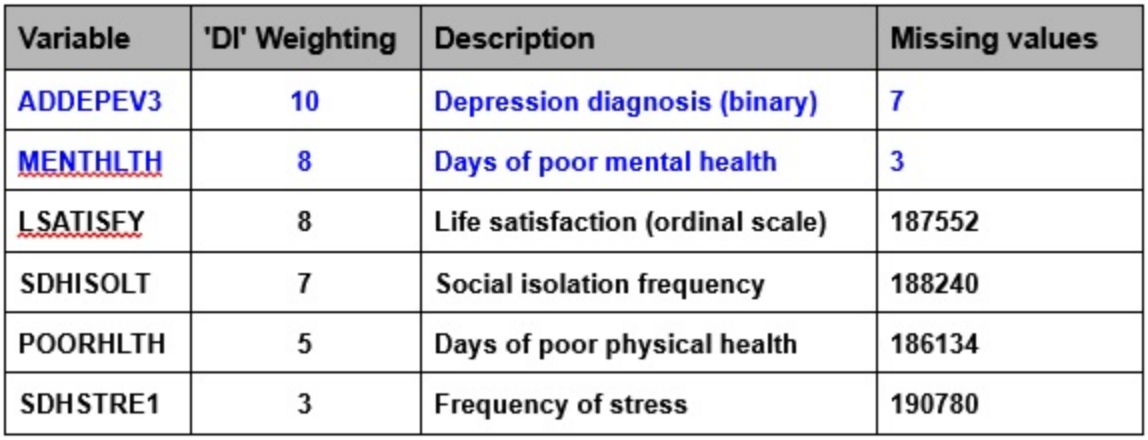
[5] See "3-A-BRFSS_EDA.ipynb" for missing data preprocessing steps.

[6] See “6-A-Depression_Index_Analysis.ipynb” for DI score distributions.

- Identified state-level trends in self-reported depression indicators, highlighting geographic disparities.

Data Transformation: Developing the Depression Index (DI) for the BRFSS dataset

- Selected key BRFSS variables related to emotional well-being to construct the DI.
- Preprocessed the dataset by filtering out incomplete or inconsistent responses before applying the DI calculation.
- Used statistical summaries and visualizations to confirm reliability of selected variables.
- A weighted scoring method was applied to ensure a balanced representation of depression indicators.
- Applied a min-max scaler to standardize the range of all columns, then applied DI Weighting, as shown in the table below.
- The sum of all weighted columns forms the Depression Index.



Variable	'DI' Weighting	Description	Missing values
ADDEPEV3	10	Depression diagnosis (binary)	7
MENTHLTH	8	Days of poor mental health	3
LSATISFY	8	Life satisfaction (ordinal scale)	187552
SDHISOLT	7	Social isolation frequency	188240
POORHLTH	5	Days of poor physical health	186134
SDHSTRE1	3	Frequency of stress	190780

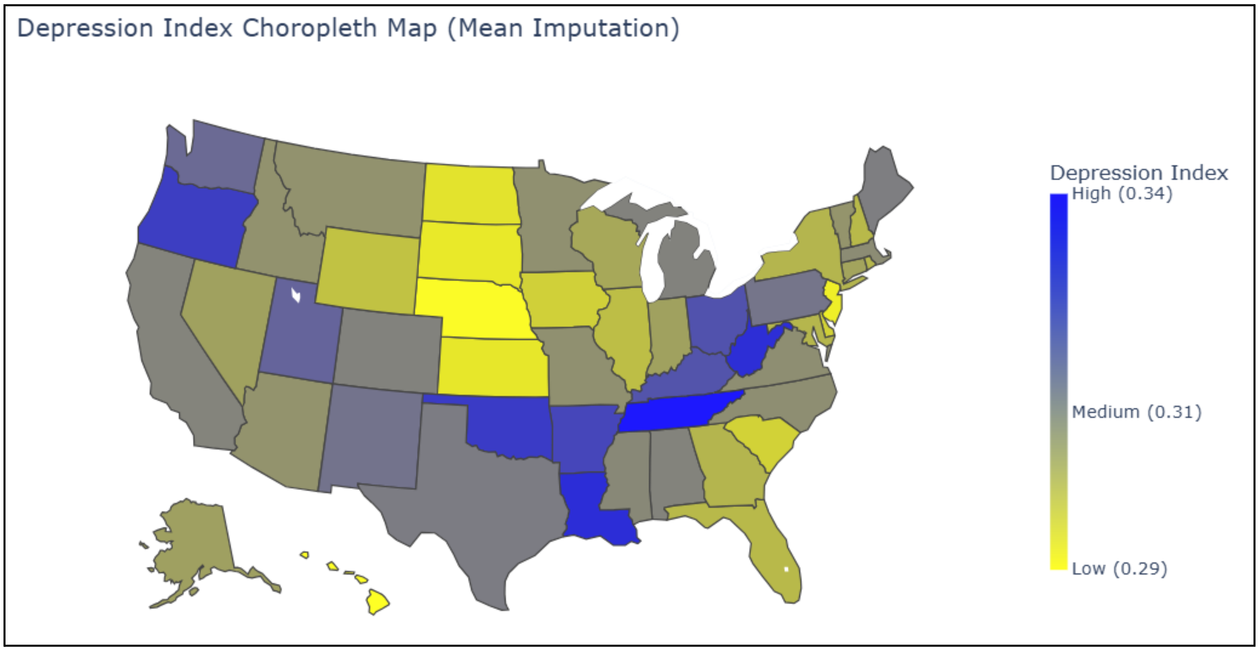
Table 1: Depression Index

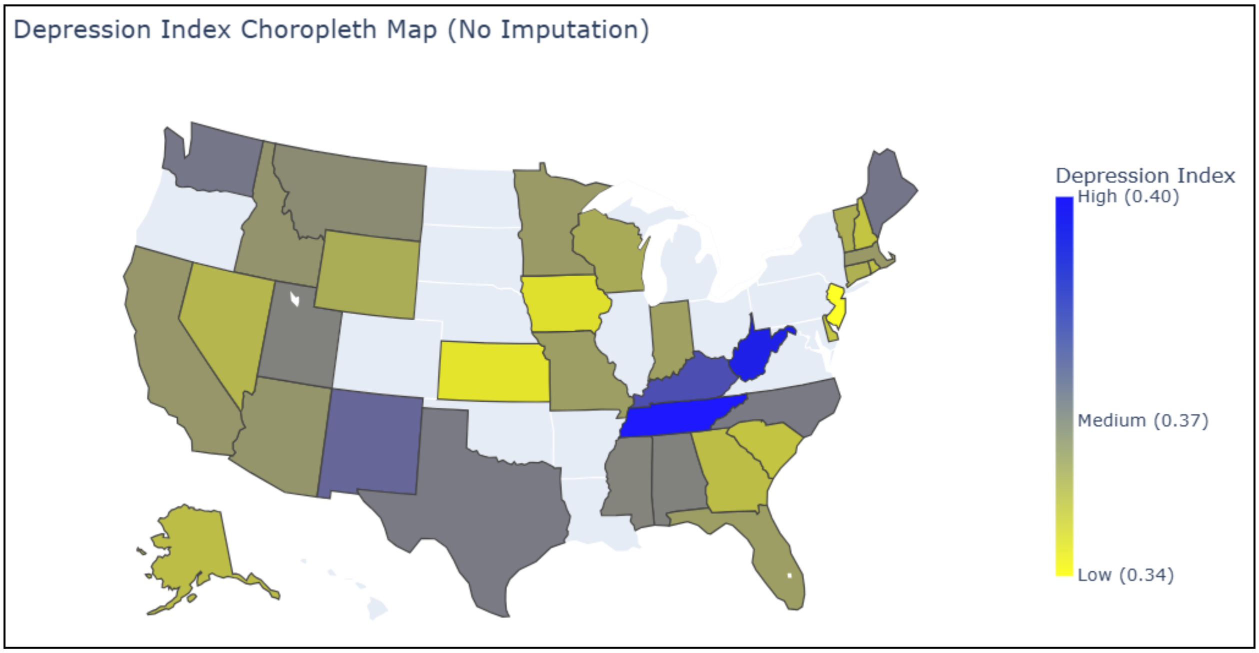
Imputing Missing Data for the BRFSS dataset:

- Four imputation strategies were considered: none, mean, median, and zero imputation.
- Compared how each method affected the distribution of DI at a state level.
- Found that mean and median imputation introduced smoothing effects, while zero imputation to underestimation.
- Determined that excluding imputation (none) preserved the most accurate state-level variations.

Visualizing the Effects of Data Imputation:

- To enhance comparability across states, the DI range was rescaled from 0-10 to 0-1.
- Across all imputation methods, the DI ranged from 0.06 to 0.40, with varying degrees of spread.
    - **No Imputation (0.34 - 0.40)**: The smallest range, as only states with existing data are included, preserving natural variability but excluding missing values.
    - **Zero Imputation (0.06 - 0.35)**: The broadest range, replacing missing values with zero, significantly lowers the minimum, artificially increasing the spread.
    - **Median Imputation (0.26 - 0.33)**: A moderate range, as the median smooths data but still allows for some natural variation.
    - **Mean Imputation (0.29 - 0.34)**: The narrowest range, replacing missing values with the mean, reduces overall variability by minimizing the impact of missing values.





Impact of Data Imputation: **Adjusting the Depression Index**

- We observed that for all data imputation methods, the states where no data was originally available showed significant bias, as can be clearly seen in the above visualizations[7]
- Identified that imputed values distorted DI distributions, particularly in states with significant data gaps, introducing artificial smoothing effects.

[7] Reference the enclosed Jupyter Notebook for additional data imputation methods and visualizations

- Noted inconsistencies in DI rankings, where imputation altered state-level comparisons.
- *After discovering these biases, we scaled down the original DI from six variables to two, removing those most impacted by missing data.*
- Finalized DI calculations using only original, unaltered responses, ensuring the dataset accurately reflected reported depression indicators.

The Daylight dataset:

Data Preprocessing for the Daylight dataset:

For the two plain text files:

- Split each plain text file on the newline character ('\n') to create two Python lists.
- Used a regular expression to capture the values of interest from each string in the lists
- Converted each list to a Pandas Series object and assembled the Series objects into a DataFrame.

For the JSON data returned by the USNO API:

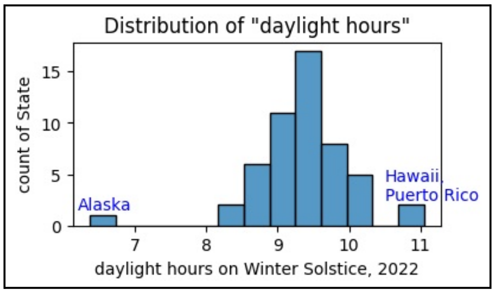
- Used the json standard module to load the text data into a Python dictionary.
- Using a "for loop", extracted the desired values for each state/territory (primarily sunrise and sunset times) into a series of Python lists.
- As part of the extraction, converted raw string data into the Pandas "datetime" data type to make it easier to work with.
- Converted the lists into Series objects and assembled them into a DataFrame.
- Took the difference between sunset and sunrise time to be the amount of daylight available, as in: *daylight = sunset - sunrise*

File management for the Daylight dataset:

- Wrote the DataFrame constructed from the USNO API to disk to make it easier to modularize our Jupyter Notebooks and facilitate collaboration for our team.

Exploring the Daylight dataset:

- Raw values for daylight ranged from 6.4 hours (seen in Alaska) to 11.0 hours (in Puerto Rico).
- Due to their geographic locations, most states had values between ~9-10 hours.
- Interquartile range: ~9-9.7 hrs
- One data instance (for Washington, D.C.) was duplicative and resulted in erroneous data, so we dropped it from the set.

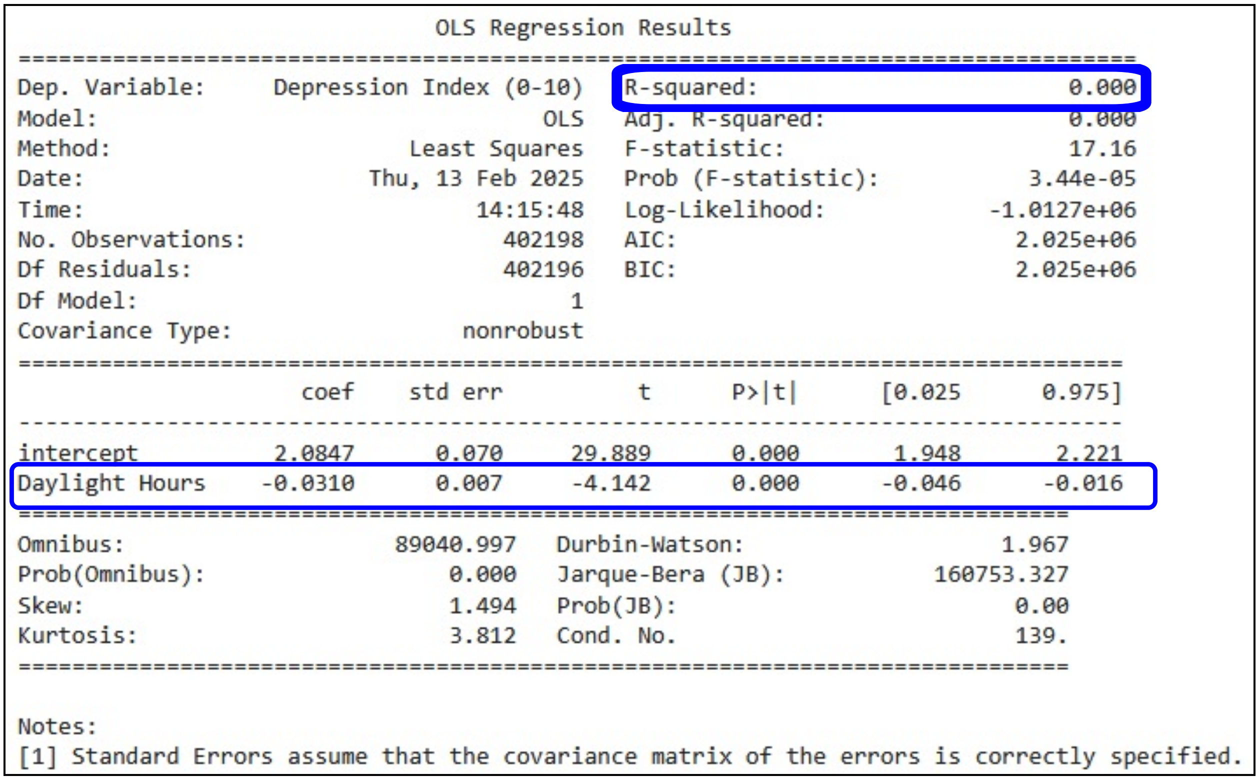


Combining the BRFSS dataset with the Daylight dataset:

- Merged the BRFSS dataset now containing the Depression Index column and the daylight hours column from the Daylight dataset into one called analysis_2022.csv. This merging was done using an inner join based on the "State" column.
- Applied Z-score function to both DI and DH to create DI_Z and DH_Z columns.
- Our primary goal was to see if there was a significant correlation between the amount of daylight hours available to an individual and that individual's self-reported feelings of depression. Merging the two datasets allowed us to perform this analysis.

Analysis

We ran an ordinary least squares (OLS) regression to test the null hypothesis that the amount of daylight available to an individual is uncorrelated with depression. If the null hypothesis were true, the regression coefficient for Daylight Hours would be zero.



```
                            OLS Regression Results
==================================================================================
Dep. Variable:     Depression Index (0-10)   R-squared:                      0.000
Model:                                 OLS   Adj. R-squared:                 0.000
Method:                      Least Squares   F-statistic:                    17.16
Date:                     Thu, 13 Feb 2025   Prob (F-statistic):          3.44e-05
Time:                             14:15:48   Log-Likelihood:           -1.0127e+06
No. Observations:                   402198   AIC:                        2.025e+06
Df Residuals:                       402196   BIC:                        2.025e+06
Df Model:                                1
Covariance Type:                 nonrobust
=================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
intercept          2.0847      0.070     29.889      0.000       1.948       2.221
Daylight Hours    -0.0310      0.007     -4.142      0.000      -0.046      -0.016
==============================================================================
Omnibus:                    89040.997   Durbin-Watson:                   1.967
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           160753.327
Skew:                           1.494   Prob(JB):                         0.00
Kurtosis:                       3.812   Cond. No.                         139.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

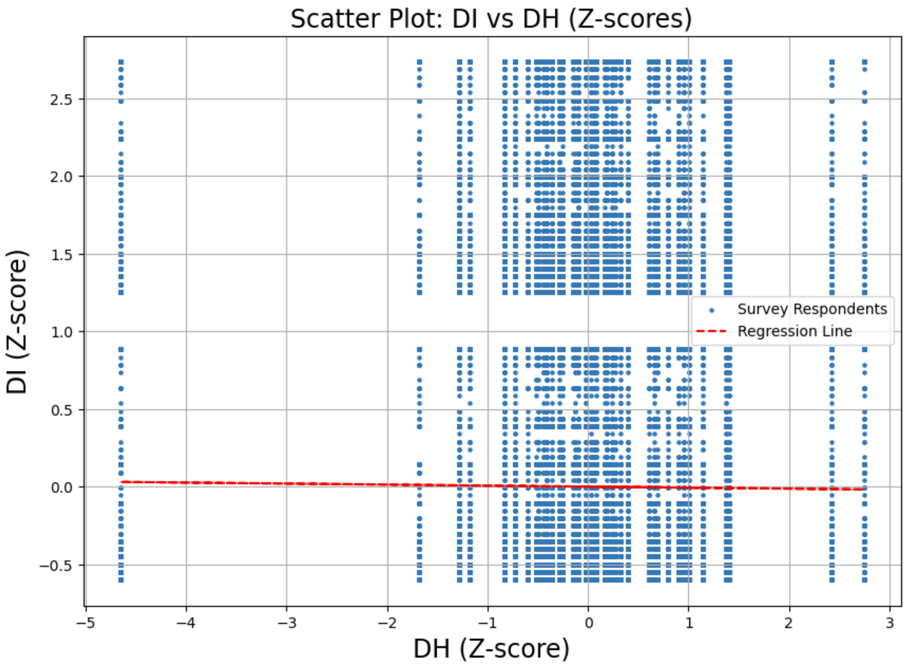
Table 2: Ordinary Least Squares (OLS) Regression Summary (via statsmodels)

From the OLS regression summary (Table 2), we see that the t-statistic for the Daylight Hours coefficient is -4.142, with a p-value of 0.000. A t-statistic less than -2 (or greater than +2) with a p-value less than 0.05 is evidence of a statistically significant correlation between the variables we used for this analysis; therefore, we can reject the null hypothesis that the coefficient is zero. We also see that the estimated coefficient for Daylight Hours is -0.0310, indicating a very weak negative correlation between the dependent and independent variables. Our Depression Index is on a 0-10 scale, so we can interpret these results to mean that, for every additional hour of daylight available to a person, the Depression Index tends to decrease very slightly: on average, by about 0.03 for every additional hour of daylight. We can also say with high confidence that this coefficient is between -0.046 and -0.016 (based on a 95% confidence interval). We also calculated the Pearson correlation coefficient for our two variables to be -0.0065, indicating a very weak negative correlation.[8] Finally, an R-squared value of 0.000 indicates that Daylight Hours explains essentially none of the variation seen in the Depression Index. These results are consistent with other findings in the literature.

[8] The fixed range of Pearson's correlation coefficient (between [-1, +1]) allows us to get an idea of the *degree* to which the two variables are correlated. A value near +1 or -1 indicates a strong correlation, while a value near zero indicates a weak correlation.

What can we infer, practically speaking, from these statistical results? Our analysis is based on observational data and we did not control for other potentially confounding variables, so we cannot infer that lack of daylight plays a causal role in depression, even if that role would be extremely small. Daylight is likely correlated to some degree with weather and temperature, for example, and perhaps with other variables we did not consider, so we really cannot say anything about its possible causal effect. The visualizations below reinforce the idea that the variation we see in the Depression Index (between different states and even between different individuals within a given state) is likely due to factors we did not consider in this analysis.

Visualizations

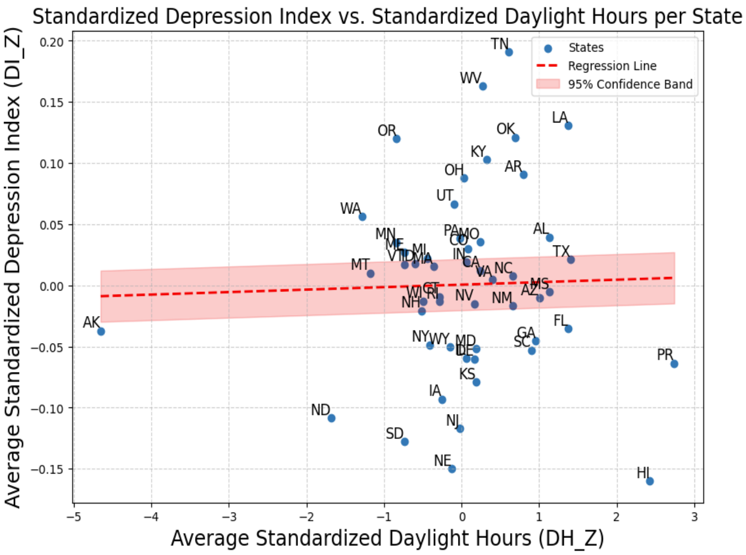


Scatter Plot: DI vs DH (Z-scores)

- Shows the relationship between the Z-scores of the DI and DH variables with a regression line
- Regression line shows that there is very weak negative correlation between the two variables.
- The stacked distribution demonstrates significant variance in DI across all values of DH.
- This plot helps to visualize the OLS regression analysis, showing how the variables are weakly correlated with significant variance.

Scatter Plot: Mean Standardized DI vs Mean Standardized DH for each State

- Takes the mean of the DI_Z and DH_Z for each state and shows the relationship between them with a banded regression line.
- Supports the choropleth map and previous analysis of weak correlation
  - AK and FL have similar DI_Z values
  - Mean DI_Z varies significantly (from ~0.15 to - 0.15) near mean DH_Z
- The 95% confidence band (shaded red) shows the area where most regression lines would fall, given different sets of sample data.



Conclusions

Our analysis demonstrated a statistically significant, though extremely weak, negative correlation between the amount of daylight available to an individual and an individual's feelings of depression, on average. If we consider how complex individuals are -- recall our description of the billions of neurons and trillions of neural connections in the average human brain -- these findings make perfect sense. Human behavior, driven by our thoughts and feelings, is so multidimensional that it is unlikely to be correlated significantly with any single variable, or even by a handful of variables. We humans are incredibly complex and resilient systems. The authors find it encouraging, however, that we were able to demonstrate a link -- albeit a small one -- between an aspect of the physical world and the abstract world of human feelings.

On the other hand, one could easily interpret these results as showing that there really is no link between daylight and depression: the magnitude of the correlation is basically zero, even though we found its small deviation from zero to be "statistically significant." Perhaps humans are simply too complex and resilient for access to sunlight to matter, or perhaps any small difference that sunlight makes is simply dwarfed by other more important factors in a person's wellbeing? Or perhaps there is too much randomness in the Depression Index itself? The "number of bad mental health days" is a fairly subjective measurement. Even a psychological diagnosis of depression involves human judgement. Perhaps what is needed is a more objective measure of depression? Or perhaps we need to ditch the idea of depression altogether and find more objective measurements to inform our response variable? With modern wearable devices and machine learning techniques, such ideas are now in the realm of the possible.

Next Steps

A review of recent studies in depression that use non-traditional measurements, such as wearable devices, will help our team to identify modern "physical world" variables that are more likely to influence depression than what we studied here. We can use machine learning techniques to identify patterns in measurements such as EEG recordings, for example, that will allow us to construct more objective dependent and independent variables.

Bias and Ethical Considerations

The BRFSS relied on phone survey data, and as such excluded a significant portion of the population that likely suffers from depression: the homeless. It is also reasonable to speculate that people who feel depressed are less likely to answer a call from an unknown caller, so our Depression Index may be an underestimate. This bias is uniformly applied to all states, however, so it should not skew results for the investigation of correlation between depression and daylight.

Final Thoughts

Regardless of the end results, we hope that the approach we used here shows a viable alternative to the more subjective doctor-patient interview style of investigation that still dominates today. If nothing else, our results demonstrate that psychological studies would benefit from more objective measurements, on both the inputs (daylight) and outputs (Depression Index) of the human behavioral system.

Statement of Work

All team members contributed significantly to the problem formulation phase and throughout the project. Collaboration was typically very good, but we could have improved by being more responsive at times. We also need a fully-shared file system. Individual focus items include:

- Alexis Parker: All Jupyter notebooks starting with 'A'
    - BRFSS dataset
        - Data acquisition, exploration, and manipulation, and visualization
        - Data validation & data imputation techniques
        - Depression Index development and PHQ-9 research
        - Choropleth Maps
    - Report sections: Abstract, all BRFSS sections
- Pete King: All Jupyter notebooks starting with 'K'
    - Daylight dataset
        - Data acquisition, exploration, manipulation, and visualization
    - Merging datasets
        - Depression Index & data imputation (error checking)
        - OLS regression analysis
    - Report sections: Motivation, Analysis, Conclusions, all Daylight sections
- Vikram Vaddamani: All Jupyter notebooks starting with 'V'
    - Merging datasets
        - Depression Index creation and PHQ-9 research
        - Scatter plot & regression line visualizations
    - Report sections: Visualizations, combining datasets, additional graphics, SOW

Works Cited

Caruso, C. *A New Field of Neuroscience Aims to Map Connections in the Brain*. Harvard Medical School News and Research, 19 Jan. 2023, https://hms.harvard.edu/news/new-field-neuroscience-aims-map-connections-brain.

Hergenhahn, B. *An Introduction to the History of Psychology.* Cengage Learning, Henley, 2009, https://faculty.cengage.com/works/9780357797716.

Khullar, A. *The Role of Melatonin in the Circadian Rhythm Sleep-Wake Cycle*. Psychiatric Times, vol. 29, no. 7, 9 Jul. 2012, https://www.psychiatrictimes.com/view/role-melatonin-circadian-rhythm-sleep-wake-cycle.

Mersch, et. al. *Seasonal affective disorder and latitude: a review of the literature*. Journal of Affective Disorders, vol. 53, no. 1, 1 Apr. 1999, https://doi.org/10.1016/S0165-0327(98)00097-4.

Patten, et. al. *Major Depression Prevalence Increases with Latitude in Canada*. The Canadian Journal of Psychiatry, vol. 62, no. 1, 11 Oct. 2016, https://doi.org/10.1177/0706743716673323.

Sender, R., et al. *Revised Estimates for the Number of Human and Bacteria Cells in the Body*. PLOS Biology, 19 Aug. 2016, https://doi.org/10.1371/journal.pbio.1002533.

Walker II, et. al. *Circadian rhythm disruption and mental health*. Translational Psychiatry, vol. 10, no. 28, 23 Jan. 2020, https://doi.org/10.1038/s41398-020-0694-0.