**DataSci 200 Final Project:** **Our GitHub Repository**
**Henry Gardner, Patrick Yim**
**Original Dataset Acknowledgement:** US Open Data Portal, data.gov

# U.S Household Mental Health & COVID-19

## Overview:

This dataset contains information on mental health care received by households in the United States during the Covid-19 pandemic in the last four weeks. The data is valuable for tracking and measuring mental health needs nationwide and making comparisons between regions based on available support.

To understand the viability and usability of the data, it is important to discuss where the data comes from at a slightly deeper level. The U.S. Census Bureau, in collaboration with five federal agencies, conducted a study called the “Household Pulse Survey” to assess social and economic impacts of the Covid-19 pandemic on US households. The survey was launched to help provide a more accurate and timely estimation around a variety of topics such as consumer spending, employment status, and in this case, mental health/wellness. The survey was conducted online via email and text messaging, with a random selection of housing units linked to email addresses and cell phone numbers. The Census Bureau matched estimates of the population by age, gender, race, and other demographic information and the data has met the NCHS Data Presentation Standards for Proportion, meaning the data has been analyzed and presented in a consistent and reliable manner in accordance with the National Center for Health Statistics.

To gain meaningful insights from the data, it is crucial to understand each column or variable in the dataset. The "Indicator" column specifies whether a percentage or an absolute number is being measured, while the "Group" and "Subgroup" columns provide additional details about the surveyed population for each indicator. The "Phase" and "Time Period" columns indicate when each measurement was taken, whether during a specific phase or over a particular timespan. Other columns such as "Value," "LowCI," and "HighCI" show quartile ranges for each measurement, such as the number of individuals who reported feeling lonely rarely. The "Suppression Flag" column identifies cases where a value has been suppressed if it falls below a certain benchmark, enabling more accurate estimates without the need to sort through suppressed values manually during analysis. Additionally, columns such as "Time Period Start Date" and "Time Period End Date" provide information on the exact dates used for measurements taken over different periods[1], which was useful for conducting time-series analyses over extended periods in research. The ability to look at certain demographic information such as ethnicity and age also were very helpful and informative when performing our analysis.

The following is a description of each column with their respective types:

Indicator: The indicator being measured. (String)
Group: The demographic group being surveyed. (String)
State: The state or territory where the survey was conducted. (String)
Subgroup: The specific demographic being measured. (String)
Phase: The phase of the collection process. (String)

[1] Devastator, The. “U.S. Household Mental Health & Covid-19.” *Kaggle*, 21 Jan. 2023, https://www.kaggle.com/datasets/thedevastator/u-s-household-mental-health-covid-19.

Time Period: The time period of the survey. (String)
Time Period Label: The label of the time period of the survey. (String)
Time Period Start Date: The start date of the time period of the survey. (Date)
Time Period End Date: The end date of the time period of the survey. (Date)
Value: The value of the indicator being measured. (Numeric)
LowCI: The lower confidence interval of the value. (Numeric)
HighCI: The higher confidence interval of the value. (Numeric)
Quartile Range: The quartile range of the value. (String)
Suppression Flag: The flag indicating whether the result has been withheld from public release or not. (Boolean)

## Assumptions:

Throughout the entirety of this research, we made one key assumption that the "Value" attribute corresponded to an overall mental health score, so to speak. The higher the value, the more the person experienced symptoms of anxiety/depression/loneliness. This follows the description of the dataset. We also assumed that the data was accurate and gathered ethically as it came from the US Government, sampling across the country.

We hypothesize that the different demographics had a variety of mental health scores, with the most variability between age and gender/sex. We also presume that location and time was a contributing factor to higher/lower scores.

## Background:

The subsequent analysis was completed in Python with specific multi-level structuring using NumPy, Pandas, Matplotlib, and seaborn. A variety of visualization techniques were used, some of which are defined below, as a reader might be unfamiliar:

- Boxplot: introductory analysis that shows a data's distribution. It allows for understanding if the data is skewed, what outliers might exist, and displays an interquartile range with the median value[2].
- Heat map: shows the magnitude of what's being described in a 2-dimensional color structure, where variation in color represents intensity or variability[3].
- Violin Plot: a combination of a boxplot and a kernel density plot, outlining the median value, the interquartile range, any outliers, and additionally shows the entire distribution of the data with peaks, unlike a normal boxplot[4].

Additional figures describing indicator distributions are presented in the appendix for ease in reading.

It is also important to note that the 'Subgroup' attribute has a section labeled 'By Presence of Symptoms of Anxiety/Depression,' unfortunately this is in the same category as the age, race, gender, etc. values so

[2] McDonald, Andy. "Creating Boxplots of Well Log Data Using Matplotlib in Python." *Medium*, Towards Data Science, 15 July 2022, https://towardsdatascience.com/creating-boxplots-of-well-log-data-using-matplotlib-in-python-34c3816e73f4.

[3] "What Are Heat Maps? A Guide to Heatmaps & How to Use Them." *What Are Heat Maps? A Guide to Heatmaps & How to Use Them*, https://www.hotjar.com/heatmaps/.

[4] Lewinson, Eryk. "Violin Plots Explained." *Medium*, Towards Data Science, 31 Jan. 2022, https://towardsdatascience.com/violin-plots-explained-fb1d115e023d.

we cannot filter by both, for example, cannot filter by age and those that had symptoms of anxiety/depression.

To see how the values in the dataset played out in the different indicators, we made a table (as seen in figure 1) displaying the average value per indicator.

Figure 1

```
Indicators and their average values:
Indicator
Received Counseling or Therapy, Last 4 Weeks                                          9.641863
Needed Counseling or Therapy But Did Not Get It, Last 4 Weeks                        10.777516
Took Prescription Medication for Mental Health, Last 4 Weeks                         21.111859
Took Prescription Medication for Mental Health And/Or Received Counseling or Therapy, Last 4 Weeks    24.409081
Name: Value, dtype: float64
```

Figure 1 shows the average value per indicator.

These results were very similar to the median value (less subject to outlier manipulation), which means that these averages reflect a general understanding of mental health stability per indicator. Notice how those that took prescription medication had higher values, indicating they had more mental health problems, which makes sense as they were diagnosed with something prevalent enough to be prescribed medication. We seek to use this base understanding to identify relationships between US households using this data.

### Questions To Be Explored:

Key-Questions:

- How does the distribution of mental health indicators change over time?
- What is the mental health impact of COVID based on: age, location, and sex/gender?

These key-questions were generated after some initial analysis to see variability within the “Group” and “Subgroup Attributes.” We discovered that there was notable variability over time and between sex/genders. And curiosity in state-wide differences prompted the inspiration for location analysis.

### Cleaning and Filtering:

Before any analysis took place, we focused on understanding the structure of the data and what was required to be changed for more accurate results. Since the data met the NCHS Data Presentation Standards for Proportion, there was little that ended up needing fixing. However, when understanding the null values dispersed throughout the dataset, it was discovered that some columns needed attention. As shown in figure 2, most columns had no null values, but the ‘Suppression Flag’ and ‘Quartile Range’ attributes had ~99.895% and ~30.721% of their values null, respectively. Due to nearly every value being null in the ‘Suppression Flag’ attribute, we dropped it as it provided no insight answering the questions posed. Additionally, we ended up calculating the quartile ranges manually through box and violin plots, which did not require the need for this attribute. Therefore, all unexpected null values were dealt with.

Figure 2

```
Percent of null values per column:
index                      0.000000
Indicator                  0.000000
Group                      0.000000
State                      0.000000
Subgroup                   0.000000
Phase                      0.000000
Time Period                0.000000
Time Period Label          0.000000
Time Period Start Date     0.000000
Time Period End Date       0.000000
Value                      2.298851
LowCI                      2.298851
HighCI                     2.298851
Confidence Interval        2.298851
Quartile Range            30.721003
Suppression Flag          99.895507
dtype: float64
```

Figure 2 shows the percentage of null values per column in the dataset.

Further filtering was needed to answer the specific sub-questions, which are discussed in their respective sections.

## Question #1:

**How does the distribution of mental health indicators change over time?**

To answer this question, we created a boxplot with time periods marked on the y-axis and indicator values depicted on the x-axis. The box represents the interquartile range (IQR), which is the middle 50% of all data being shown. Minor data formatting and cleaning was required to answer this sub question. The "Time Period Start Date" and "Time Period End Date" were converted to a datetime format using the pandas "to_datetime" function, allowing for proper formatting of the datetime index before usage within our analysis.

Interesting takeaways from this subquestion is first and foremost, understanding the seasonality of the data. There is a variance in the quartiles between months, including the values within January and February. Additionally, there is a gap in the dataset for the time period between December 22nd and January 1st, likely stemming from the lack of data due to the Christmas holidays. From the boxplot, we can also see the median of mental health indicator values generally increasing over time, demonstrating a decline in mental health over the measured timeframe; however, the range of values also increased, as shown by the longer whiskers in some of the later years. Some outliers within the data are present, which are shown as dots in figure 3. Potential reasons for these outliers include potential severe cases of mental health, outside events, or even data entry errors that may have skewed the data

We decided to create a boxplot analysis to further investigate the distribution of mental health indicator values over a period of time. As seen in figure 3, the median value of the mental health indicator value is around 15, with the majority of the observations falling between 10 and 25. There most definitely is a slight variance of the mental health indicator values over the period of 6 months - while the median values are relatively stable, the distribution of mental health indicator values does not appear to be symmetrical, with a longer tail towards higher values. Overall, however, the boxplot analysis suggests that while there may be outliers and seasonality, the majority of the mental health indicator values remain stable over time.

To visualize this data in another format, we created a time series chart to display the trend of mental health indicators over the period of 6 months. As shown below in figure 4, the mental health indicator shows an overall increasing trend, suggesting that mental health concerns have increased in prevalence over the pandemic period, likely due to increased feelings of isolation. This trend mimics and is consistent with themes often found in the news and articles of various sources that discusses the rise of COVID depression and mental health problems associated with complete isolation from social interactions[5]. As shown in the time series, there is a leveling off and slight decrease across 2021, which may signify the increased aptitude to cope with mental health issues in the beginning of the year.

One key action item we took within the time series chart is because we found the missing values within the end of 2020 through the Boxplot chart, we made sure to fill in the gaps. Within a time series, having a gap would reduce the viability of the overall visual representation, so we made sure to leverage forward

[5] Hayward, Ed, Eryk. "COVID-19's toll on mental health." Boston College Article, April 2020, https://www.bc.edu/bc-web/bcnews/campus-community/faculty/anxiety-and-stress-spike-during-pandemic.html

filling to fill in the missing values with the last observed value. Since the timeframe that was missing wasn't a large portion of the overall analysis, this method was viable and produced good, clean results.

Figure 3

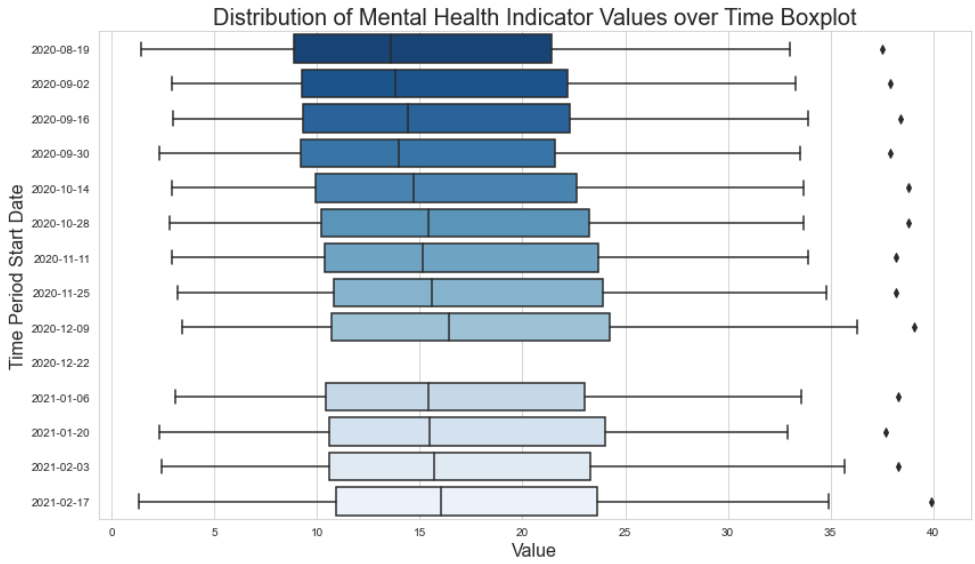



Figure 3 shows the boxplots over time to show the mental health indicator distribution.

Figure 4

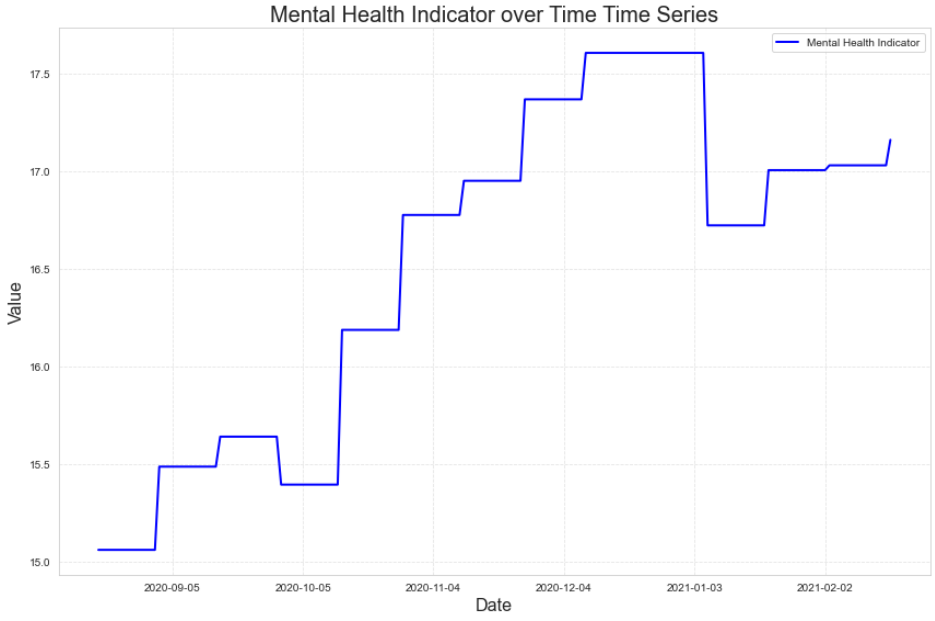



Figure 4 shows the time series plot of indicator values.

## Question #2:

**Mental Health Impact of COVID based on age?**

The data split the ages of participants into seven categories: 18 - 29, 30 - 39, 40 - 49, 50 - 59, 60 - 69, 70 - 79, and 80 years and above. To preserve confidentiality, no other information was given in the data, so all analyses deal with these 10 year subsets. From initial inspection, there were no null values in the age category so minimal cleaning/filtering was needed. As described, the indicator “Value” attribute describes the severity of the mental health, meaning that the higher the value, the more one experiences symptoms of depression/anxiety. Therefore, grouping these values by age category should give an insight into which age group experienced the most of these symptoms. To start, we grouped the data by the age categories and computed the average indicator score per age subset. We additionally used the confidence interval information to gather average estimated ranges of accuracy. The table was sorted based on the indicator values, so we could see in a top down fashion, which age groups were the most affected. The results can be seen in figure 5 below.

Figure 5

Average Value and Confidence Levels By Age:

| Subgroup | Value | LowCI | HighCI | Confidence Range |
|---|---|---|---|---|
| 18 - 29 years | 19.590385 | 17.834615 | 21.430769 | 3.596154 |
| 30 - 39 years | 19.069231 | 17.844231 | 20.355769 | 2.511538 |
| 40 - 49 years | 17.275000 | 16.188462 | 18.419231 | 2.230769 |
| 50 - 59 years | 16.223077 | 15.092308 | 17.403846 | 2.311538 |
| 60 - 69 years | 13.765385 | 12.801923 | 14.771154 | 1.969231 |
| 70 - 79 years | 11.278846 | 10.098077 | 12.561538 | 2.463462 |
| 80 years and above | 9.762500 | 7.177083 | 12.966667 | 5.789583 |

Figure 5 shows the average indicator value, average lower confidence interval, average higher confidence interval, and average confidence range per age group.

The average is often manipulated with extreme outlier values, so further analysis into the distributions of each category were done with a violin plot. In figure 6 below, notice the violin plot with 7 different distributions, one for each age subset.

Figure 6

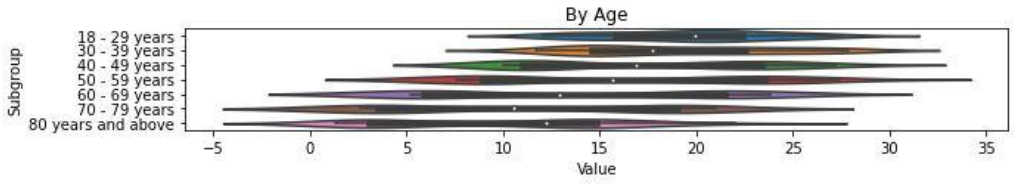



Figure 6 shows the violin plots per age group. The white dot indicates the median value and the black box represents the interquartile range.

Looking at both figures 5 and 6, it is clear that mental health problems (i.e. symptoms of anxiety/depression) decreased with age. In other words, the younger a person was, the more likely they

would experience mental health problems. In fact this stays completely true when looking at each group, as each age subset increased the indicator value lowered. However, the “80 years and above” subset had the least confident range, meaning that the true average indicator value was not as tightly confident as the other ages. However, after checking the data, each age subgroup had an equal number of units: 56, so this lack of confidence can be attributed to original data collection. It is clear that the younger people sampled experienced more mental health problems than those that were older. These results matched our original hypothesis.

## Question #3:

**Mental Health Impact of COVID based on Location?**

The most satisfactory way to answer this question from our experience was to create a bar graph detailing the variance of the mental health impact of COVID based on location. The dataset included detail around specific state level mental health impact, as shown in figure 7. The largest data cleansing and formatting came from determining what to put as our y axis. Upon thoughtful consideration on what we felt like was the best depiction of analysis to answer the question, we agreed on usage of the Mean Mental Health Indicator Value.

The bar graph revealed notable geographic disparities and impact of states towards mental health during COVID, which is quite interesting. As reflected in figure 7, the state with the lowest mean mental health indicator values (not surprisingly) is Hawaii, while the state with the highest mental health indicator value is West Virginia. To go down further down the list, Wyoming, Florida, and North Dakota all had relatively lower mental health indicator values. On the flip side of that, Rhode Island, Kentucky, and Utah had higher mental health indicator values.

Given this analysis and findings, there are several potential reasons behind why Hawaii, Florida, and Wyoming have lower mental health issues than West Virginia, Rhode Island, and Kentucky. To state the obvious from the get-go, Hawaii is a beautiful destination where individuals are carefree, nature is at the footsteps of your house, and the overall culture is one that can be deemed as supportive of mental health. Wyoming and Florida also have an abundance of outdoor activities, with a lower population likely reducing Wyoming’s mental health indication value even further because of a reduced fear of COVID infection through exposure. These findings also likely suggest that the disparities with the mean value of Mental Health Indication Scores could also result from a disparity between states with the availability and quality of mental health services.

Figure 7

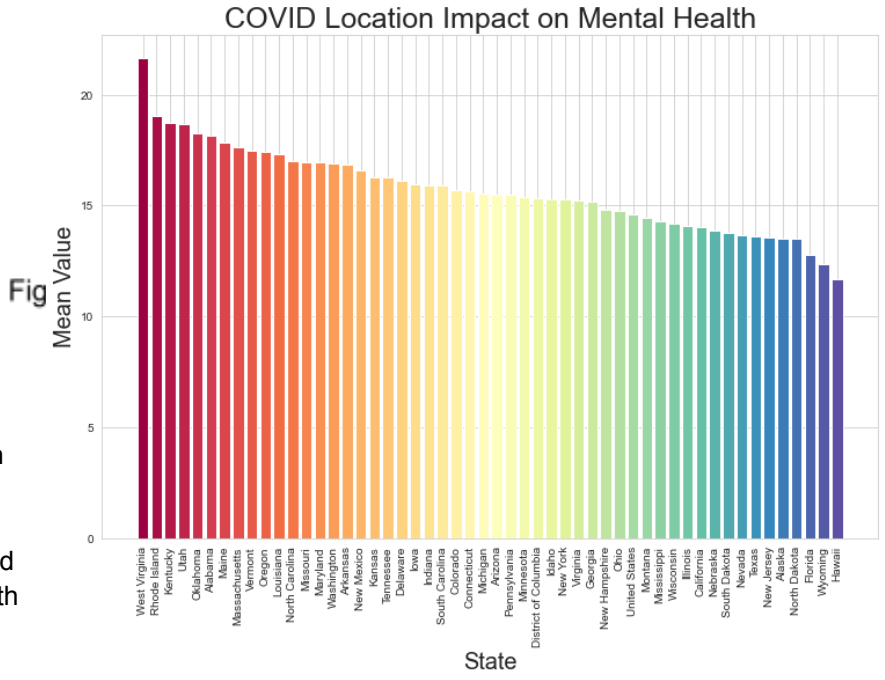



Figure 7 shows a bar graph of the average value per state, ordered by value.

## Question #4:

**Mental Health Impact of COVID based on sex/gender?**

One of the clearest subgroup variabilities came from the gender subgroup, which the data separated to “Male” and “Female” with no null values. There were 56 rows containing male data and 56 rows containing female data. Similar to the age category, we split the dataset and grouped by the gender subgroup attribute. We then computed averages for the value, lower confidence interval, higher confidence interval, and confidence range (refer to figure 8 below). We then wanted to understand the distribution and how the outliers affected these averages so we made a violin plot to further describe the distributions of gender-based variability (refer to figure 9 below).

Figure 8

Average Value and Confidence Levels By Gender:

| Subgroup | Value | LowCI | HighCI | Confidence Range |
|---|---|---|---|---|
| Female | 19.988462 | 19.298077 | 20.684615 | 1.386538 |
| Male | 12.317308 | 11.588462 | 13.071154 | 1.482692 |

Figure 8 shows the averages for the indicator value, the lower confidence interval, the higher confidence interval, and the confidence range based on gender.

Figure 9

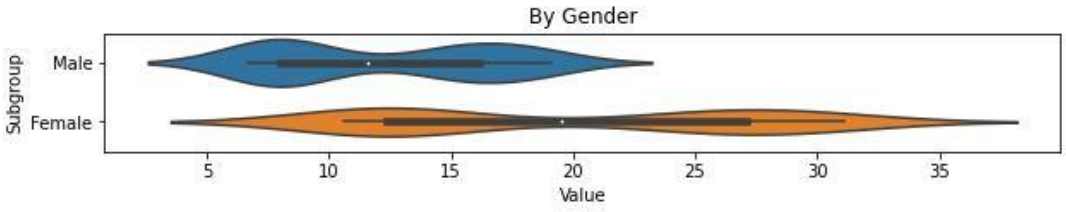



Figure 9 shows the violin plots per gender. The white dot indicates the median value and the black box represents the interquartile range.

Filtering with the ‘By Gender’ value in the ‘Group’ attribute and sub-filtering by the Indicator, it was determined that an equal number of males and females were listed under each indicator with a value of 14. Therefore, there is no way to determine a relationship between the indicator (what treatment they were getting) and the value based on the gender. In other words, with the provided data, there was no indication that more males were getting counseling or vice versa that might have contributed to a lower value. But, it is clear that Females were more affected by Covid in terms of mental health. Although we did not hypothesize that Females were more affected, we did believe that there would be a notable difference between Males and Females.

## Results and Limitations

We found that there are a variety of factors that influence Covid's impact on mental health, including demographics, location, and the length of the pandemic. From this dataset, we were able to conclude that younger aged people and females were more affected by Covid than older people and Males, respectively. We also found that location played into the overall significance of mental health problems, with the most notable being West Virginia with an average mental health severity score higher than every other state by a pretty large margin (~4 units). And, as would make sense, the score increased (on average) over time, meaning that Covid had a greater effect on people over time.

Further analysis into which indicators were more responsible for higher scores would have really aided in discovering what the problems were during the pandemic. However, every indicator had the same valid row counts for each Group/Subgroup. Therefore, we could not determine a relationship between value and treatment. Additionally, there was a subgroup called: 'By Presence of Symptoms of Anxiety/Depression.' This subgroup would have been great to use when filtering by demographics, location, etc. however this was not possible as the Subgroup attribute was split by these features. Meaning, if we filter by 'By Presence of Symptoms of Anxiety/Depression' no demographic, etc. data would show up. This was an unfortunate unavoidable limitation of the data.

## Appendix

Figure 10

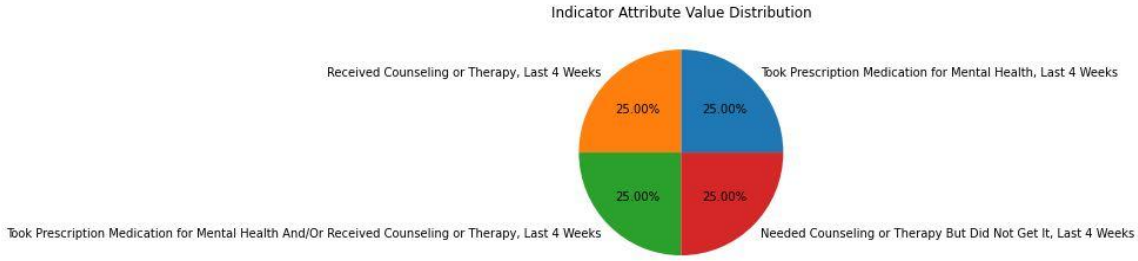
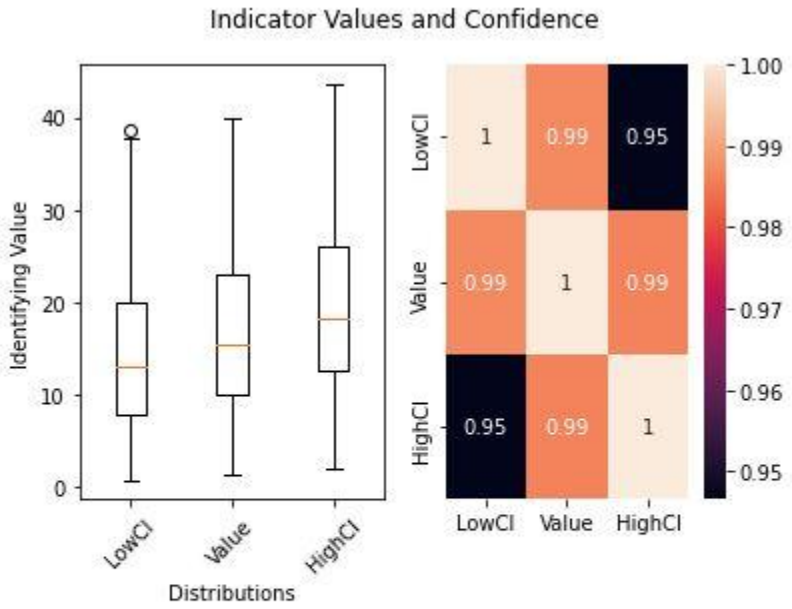



Figure 10 illustrates the indicator values over the entire dataset. Clearly, each indicator made up exactly 25% of the data.

Figure 11




```
AVERAGES:
Value     16.491791
LowCI     13.966096
HighCI    19.347086
dtype: float64
```

Figure 11 displays a boxplot and heat map over the Value, LowCI, and HighCI attributes. These illustrate the averages and distributions of each variable over the entirety of the dataset regardless of Group/Subgroup. The heat map displays the correlation between each of these listed values.

## References


“Correlation Coefficient.” *JMP*, https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-correlation/correlation-coefficient.html.

Devastator, The. “U.S. Household Mental Health & Covid-19.” *Kaggle*, 21 Jan. 2023, https://www.kaggle.com/datasets/thedevastator/u-s-household-mental-health-covid-19.

Hayward, Ed. “COVID-19’s toll on mental health.” Boston College Article, April 2020, https://www.bc.edu/bc-web/bcnews/campus-community/faculty/anxiety-and-stress-spike-during-pandemic.html

Lewinson, Eryk. “Violin Plots Explained.” *Medium*, Towards Data Science, 31 Jan. 2022, https://towardsdatascience.com/violin-plots-explained-fb1d115e023d.

McDonald, Andy. “Creating Boxplots of Well Log Data Using Matplotlib in Python.” *Medium*, Towards Data Science, 15 July 2022, https://towardsdatascience.com/creating-boxplots-of-well-log-data-using-matplotlib-in-python-34c3816e73f4.

“NumPy Documentation#.” *NumPy Documentation - NumPy v1.24 Manual*, https://numpy.org/doc/stable/index.html.

“Pandas.dataframe.groupby#.” *Pandas.DataFrame.groupby - Pandas 2.0.0 Documentation*, https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.groupby.html.

“Statistical Data Visualization#.” *Seaborn*, https://seaborn.pydata.org/.

“US Open Data Portal, Data.gov's Datasets.” *Data.world*, https://data.world/datagov-us.

“Visualization with Python.” *Matplotlib*, https://matplotlib.org/.

“What Are Heat Maps? A Guide to Heatmaps & How to Use Them.” *What Are Heat Maps? A Guide to Heatmaps & How to Use Them*, https://www.hotjar.com/heatmaps/.