Term: Fall 2023 **Subject:** Data Science, Analytics and Engineering **Number:** 598

Course Title: Statistics for Data Analysts (DSE 598)

FINAL PROJECT:

HYPOTHESIS TESTING

ON

MENTAL HEALTH CARE DATASET

Submitted To:

Prof. Rong Pan

Submitted By:

Name: Dhruva Patel	Student ID: 1230583512
Name: Jay Jayani	Student ID: 1226668926
Name: Siddharth Ranjan	Student ID: 1229564910
Name: Kalyani Tidke	Student ID: 1230605781

I. INTRODUCTION

Mental health is an important component of overall well-being, affecting individuals, families, and societies as a whole. There has been a growing recognition in recent years of the importance of understanding and addressing mental health issues. This report delves into a large dataset on mental health, with the goal of shedding light on various aspects such as prevalence, risk factors, and potential interventions.

The United States Census Bureau, in collaboration with several federal agencies, launched the Household Pulse Survey to collect data on the societal and economic effects of the COVID-19 pandemic on American households. This survey aims to provide insights into the pandemic's impact on medical conditions after covid-19 pandemic effect. The survey data could provide crucial insights to policymakers, researchers, and the public, allowing them to make informed decisions and take appropriate actions to address the issues and challenges arising from the pandemic's widespread impact.

1.1 PROBLEM DESCRIPTION

The primary objective of this analysis is multifaceted, aiming to unravel intricate patterns and discern disparities in the utilization of mental health care services among diverse demographic groups within the United States. The significance of this investigation lies in its potential to furnish invaluable insights into the dynamics of mental health care usage, especially considering the nuanced influence of demographic factors. The inclusion of a timeline within the dataset is instrumental in conducting a temporal analysis.

By examining trends over specific time periods, this analysis can identify evolving patterns in mental health care utilization. For instance, it may uncover whether there has been a notable increase in therapy usage among a particular age group in recent years or if prescription medication utilization has changed over the last decade. By delving into these patterns over various time periods, this analysis seeks to offer critical information that can inform health policy formulation and enhance the delivery of mental health services.

1.2 PROBLEM CONTEXT

In today's mental health landscape, the need to analyze patterns in mental health care utilization stems from increased societal awareness of mental health, ongoing efforts to reduce stigma, and the dynamic interplay of demographic factors influencing help-seeking behaviors. Understanding these patterns is critical for effective resource allocation, policymaking, and the design of responsive and inclusive mental health services as treatment modalities evolve and external factors impact mental health. This analysis serves as a guidepost for navigating the complex terrain of mental health care, illuminating trends that can inform targeted interventions and foster a more equitable and resilient mental health care system capable of addressing individuals' diverse needs over time.

1.3 ORIGINAL DATA DESCRIPTION

The initial dataset comprises 10,404 rows and 15 features or attributes. It has been subdivided based on four unique indicators, with each indicator having 2,601 data entries.

- Indicator: The specific metrics or measures that have been recorded for analysis are most likely represented by this attribute. These may include various social and economic factors influenced by the COVID-19 pandemic, such as took prescription medicine, took prescription medicine and/or receiving counseling or therapy, received counseling or therapy, needed counseling or therapy but did not get it.
- Group: This attribute classifies the data into specific groups or segments, which may include various demographics, household types, specific age groups, education, race, and so on.
- State: The states in the United States for which data is collected are denoted by this attribute. It enables a regional analysis of the pandemic's effects.
- Subgroup: This attribute refers to a subcategory or subset of the previously mentioned group attribute, providing more information about the specific population being studied. It essentially provides specific and detailed information to the group variable assigned in the previous attribute. For example, if the group attribute has the value "by age," the sub group attributes specify the age range, such as 0-15, 16-30, and so on.
- Phase: This attribute can represent various stages or phases of the pandemic or the associated response, allowing for the tracking of changes over time.Time Period: This attribute specifies the time periods in which the data was collected or is relevant. It aids in comprehending the indicators' temporal trends and changes.
- Time Period Label: A descriptive label or identifier for the time provides the detailed range of intervals with specific date when the record was recorded. It helps to gain details about the same indicator label over different days or intervals.
- Time Period start date: This attribute indicates the start date of the survey.
- Time Period End date: This attribute indicates the end date of the survey.
- Value: The actual numerical data for the various indicators is contained in this attribute, providing quantitative insights into the effects of the pandemic on various aspects of society and the economy.
- LowCI: This attribute may represent the data's lower bound of a confidence interval.
- HighCI: This attribute may represent the upper bound of the data's confidence interval.
- Confidence Interval: This attribute provides additional information about the data's statistical uncertainty, indicating the range within which the true value is expected to lie with a specified level of confidence. The width of the confidence interval influences the level of uncertainty in the data, with a wider interval implying greater uncertainty and a narrower interval signifying more confidence in the estimated values.

- Quartile Range: This attribute may provide information about the data distribution's quartile range. Understanding the quartile range is crucial for detecting potential outliers or disparities within the dataset, thus aiding in a more nuanced analysis of the data distribution.
- Suppression Flag: Indicates whether specific data points have been suppressed or withheld due to privacy concerns or insufficient sample sizes, ensuring data confidentiality and integrity.

1.4 DATA RECOLLECTION

The addition of regional categorization to the original dataset by dividing U.S. states into five zones—Southwest, Southeast, Midwest, West, and Northeast—significantly enhances the dataset's analytical depth. This regional classification introduces a valuable dimension for comparison, effectively reducing the number of categorical values associated with each state. The rationale behind this division likely stems from the recognition that geographical proximity and shared socio-cultural factors within each zone may influence mental health care patterns.[1]

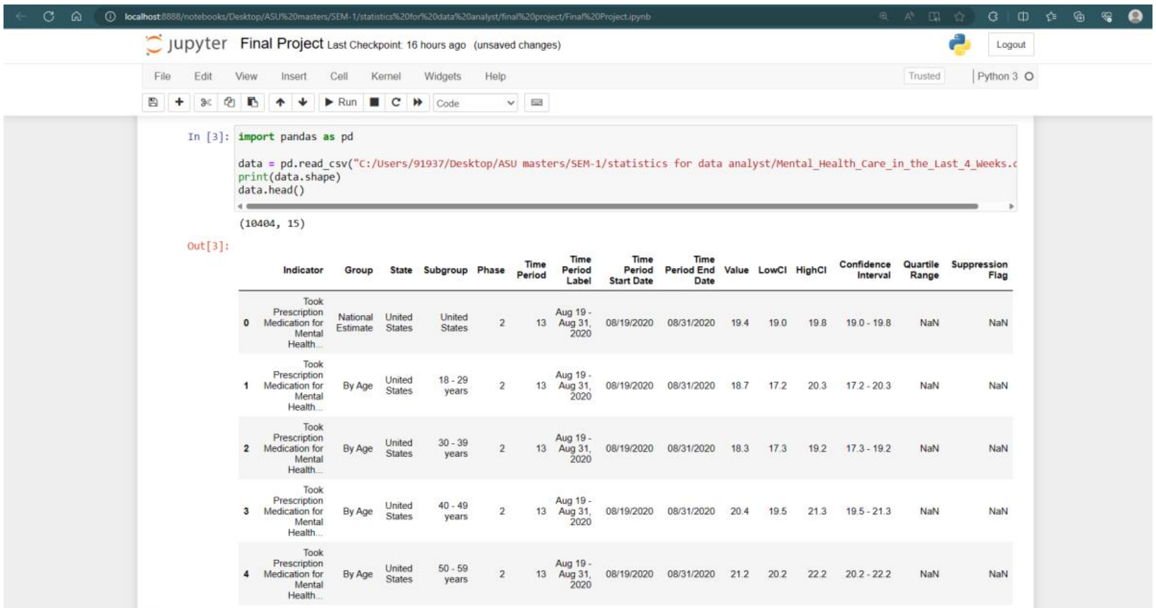
The addition of “Number of Clinics” to each data point with state grouping represents a valuable addition to the dataset. This addition recognizes the critical role that mental health clinic availability plays in shaping the landscape of mental health care utilization within each state. The number of clinics can be a critical determinant of the accessibility and availability of mental health services, influencing individuals' willingness to seek treatment. This improvement allows for a more thorough analysis, allowing researchers and policymakers to investigate the relationship between the prevalence of mental health care utilization, demographic factors, and the local infrastructure of mental health clinics. By including this field, the dataset becomes a more robust tool for investigating the intricate interplay of geographical, demographic, and healthcare infrastructure factors in the United States allowing to understanding mental health care patterns across different states.[1]

After observing similarities and patterns in the data we decided to combine the "Group" fields "By Sex" and "By Gender" into a single category. This amalgamation entails renaming the "Group" field "By Sex" and renaming the sub-group entries "Female," "Male," and "Transgender." This consolidation streamlines the dataset, reducing redundancy and allowing for a more focused analysis while maintaining a nuanced understanding of sex and gender dynamics in mental health care utilization. The dataset becomes more user-friendly and interpretable by creating a unified framework, allowing us to draw insights from a cohesive representation of sex and gender-related patterns in mental health care, thereby improving the dataset's overall clarity and utility.

1.5 DATA CLEANING

The initial dataset comprises 10,404 rows and 15 features or attributes. Null values are initially present in fields such as "Value," "LowCI," "HighCI," "Confidence Range," "Quartile Range," and "Suppression Flag." Because this information is derived from real-time surveys, arbitrary

approximation of missing values is avoided to avoid bias. "Value," "LowCI," "HighCI," and "Confidence Range" each have 490 NaN values, while "Quartile Range" has 3672 and "Suppression Flag" has 10382. Because of the large number of missing values, it was decided to remove the "Quartile Range" and "Suppression Flag" columns from the dataset.
We observed that the "Time Period" field in the dataset exhibited varying date formats, posing challenges during visualization. To address this inconsistency and ensure uniformity, a decision was made to standardize the date formats across the dataset. Consequently, the dates in the "Time Period" field have been modified to adhere to a common format, YY-MM-DD



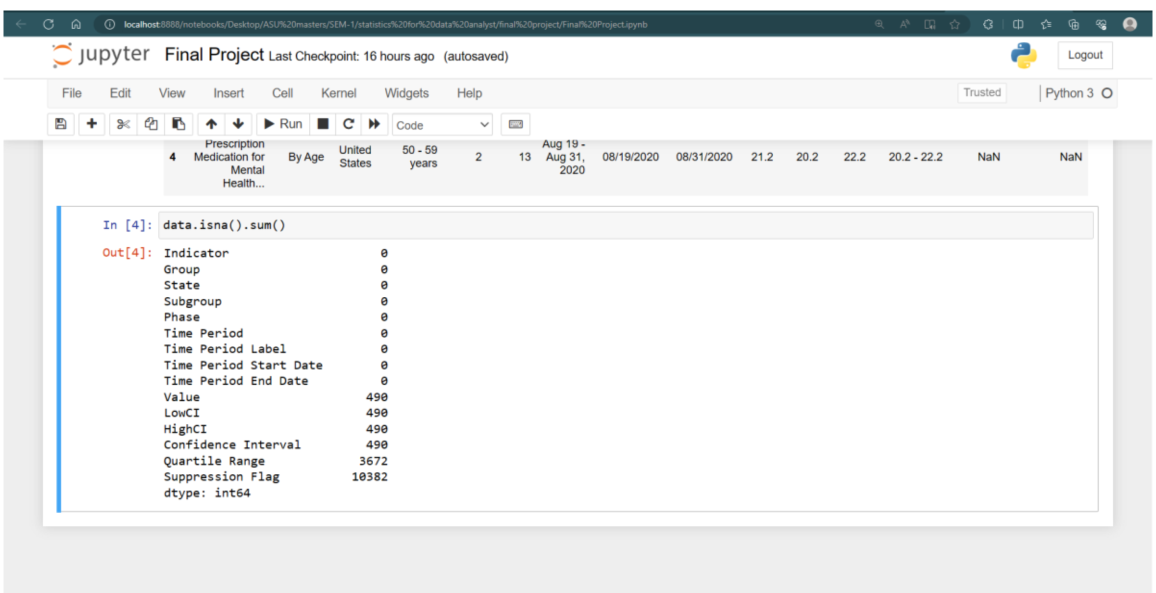

```
import pandas as pd

data = pd.read_csv("C:/Users/91937/Desktop/ASU masters/SEM-1/statistics for data analyst/Mental_Health_Care_in_the_Last_4_Weeks.c
print(data.shape)
data.head()
```

```
(10404, 15)
```

Out[3]:

	Indicator	Group	State	Subgroup	Phase	Time Period	Time Period Label	Time Period Start Date	Time Period End Date	Value	LowCI	HighCI	Confidence Interval	Quartile Range	Suppression Flag
0	Took Prescription Medication for Mental Health...	National Estimate	United States	United States	2	13	Aug 19 - Aug 31, 2020	08/19/2020	08/31/2020	19.4	19.0	19.8	19.0 - 19.8	NaN	NaN
1	Took Prescription Medication for Mental Health...	By Age	United States	18 - 29 years	2	13	Aug 19 - Aug 31, 2020	08/19/2020	08/31/2020	18.7	17.2	20.3	17.2 - 20.3	NaN	NaN
2	Took Prescription Medication for Mental Health...	By Age	United States	30 - 39 years	2	13	Aug 19 - Aug 31, 2020	08/19/2020	08/31/2020	18.3	17.3	19.2	17.3 - 19.2	NaN	NaN
3	Took Prescription Medication for Mental Health...	By Age	United States	40 - 49 years	2	13	Aug 19 - Aug 31, 2020	08/19/2020	08/31/2020	20.4	19.5	21.3	19.5 - 21.3	NaN	NaN
4	Took Prescription Medication for Mental Health...	By Age	United States	50 - 59 years	2	13	Aug 19 - Aug 31, 2020	08/19/2020	08/31/2020	21.2	20.2	22.2	20.2 - 22.2	NaN	NaN

Fig 1: Original Dataset




```
data.isna().sum()
```

Out[4]:

```
Indicator                    0
Group                        0
State                        0
Subgroup                     0
Phase                        0
Time Period                  0
Time Period Label            0
Time Period Start Date       0
Time Period End Date         0
Value                      490
LowCI                      490
HighCI                     490
Confidence Interval        490
Quartile Range            3672
Suppression Flag         10382
dtype: int64
```

Fig 2: NaN values in original dataset

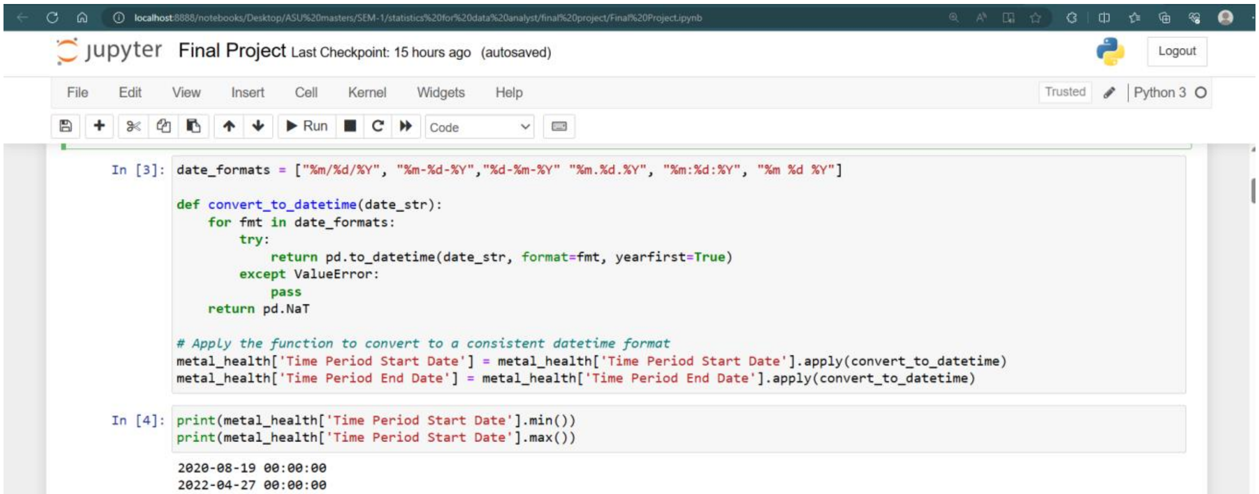


Fig 3: Time Period Column Formatting

1.6 HYPOTHESIS OF INTEREST

1. The first potential hypothesis for investigation involves assessing the availability of medical treatment based on the race of individuals. The United States, being a country characterized by a diverse range of populations, exhibits variations in racial demographics across regions. It is conceivable that certain areas have a majority population of a specific race, while others have a minority representation. In this context, it becomes crucial to examine whether there are disparities in the accessibility of medical services, specifically mental health care, across different racial groups. This hypothesis addresses the important concern of ensuring equitable availability of medical services to all individuals, regardless of their racial background, and underscores the need for a thorough examination of potential disparities in access to mental health care based on race.

2. The second hypothesis posits that younger and middle-aged individuals experienced more pronounced mental health challenges during the pandemic compared to older individuals. This hypothesis is grounded in the potential influence of various factors, such as heightened stress stemming from academic or professional obligations and the impact of social isolation imposed by pandemic-related restrictions. The younger and middle-aged populations may be particularly vulnerable to the disruptions caused by the pandemic, given the significance of career development, educational pursuits, and the social connections that are integral to their life stages. To rigorously test this hypothesis, a detailed analysis of the dataset is imperative. By examining variables related to mental health indicators, treatment modalities, and demographic factors across different age groups, we aim to provide empirical evidence either supporting or refuting the hypothesis

3. The third hypothesis proposes that states in the country's southern and central regions have more access to mental health service utilization, over other parts of the country. This

hypothesis suggests that factors such as a lack of access to mental health professionals or a misunderstanding of the benefits of counselling may contribute to the observed pattern. A comprehensive analysis of the dataset is required to rigorously examine and validate this hypothesis. The analysis seeks to uncover patterns that support or refute the hypothesis by scrutinizing variables related to mental health service utilization, treatment preferences, and regional demographics.

1.7 IMPORTANCE OF THE SOLUTION

1. The hypothesis relating to availability of medical treatment based on the race of individuals in the United States is a crucial hypothesis, given the nation's diverse demographic landscape. This analysis is pivotal for assessing the equitable distribution of healthcare resources and identifying potential disparities in access to medical services among various racial groups. Understanding the intersectionality of race and healthcare access is essential for informing targeted interventions, shaping healthcare policies, and fostering health equity. By addressing this hypothesis, policymakers can work towards eliminating racial disparities, ensuring that all individuals have equal access to medical treatment, and ultimately enhancing the overall inclusivity and effectiveness of the healthcare system.

2. The hypothesis that younger and middle-aged people faced more severe mental health challenges during the pandemic is significant because it has the potential to shape targeted mental health interventions, inform resource allocation strategies, and guide policy formulation. This hypothesis, if validated, highlights the need for tailored support mechanisms that address the specific stressors and vulnerabilities of younger and middle-aged populations. It has implications for preventive measures, educational campaigns, and stigma reduction efforts, all of which contribute to a more nuanced and effective approach to mental health care. Recognizing age disparities during pandemics is critical for improving public health preparedness and understanding the long-term societal impact of mental health challenges, ultimately fostering a more resilient and empathetic society.

3. The third hypothesis, asserting that states in the southern and central regions of the country exhibit higher mental health service utilization, is of paramount importance as it addresses potential disparities in access to mental health resources. The outcome of this analysis holds the potential to inform targeted strategies, ensuring equitable access to mental health care across diverse geographic areas and advancing the goal of a more inclusive and responsive mental health system nationwide.

II DATA ANALYSIS AND VISUALIZATION

Following the initial project proposal meeting, the decision has been made to focus exclusively on two indicators from the available set of four. The selected indicators are "Received Counseling or Therapy, Last 4 Weeks" and "Needed Counseling or Therapy But Did Not Get It, Last 4 Weeks." The clean dataset in consideration has 5202 data points and 11 attributes.

All subsequent hypothesis testing and analyses will be centered around these chosen indicators. This focused approach streamlines the investigation, allowing for a more in-depth examination of specific aspects related to the utilization of counseling or therapy services in the last four weeks. By focusing in on these key indicators, the analysis aims to draw targeted insights and conclusions that align directly with the project's objectives and research questions.

The dataset is initially divided into two sections: one that includes all group fields except the state, and the other that is specific to the 'By State' group. The 'By State' group stands out because it contains over 30 unique classifiers, necessitating a unique approach to analysis.

2.1 Analyzing all group vales except 'By State'

The focus of the investigation is primarily on the time period and the corresponding value measure for each group category, with a detailed analysis performed separately for each indicator within each group category. To assess the variations in the value measure across all indicators within the specified time frame. This analysis aims to scrutinize how each indicator evolves over time, providing insights into trends, fluctuations, and potential patterns.

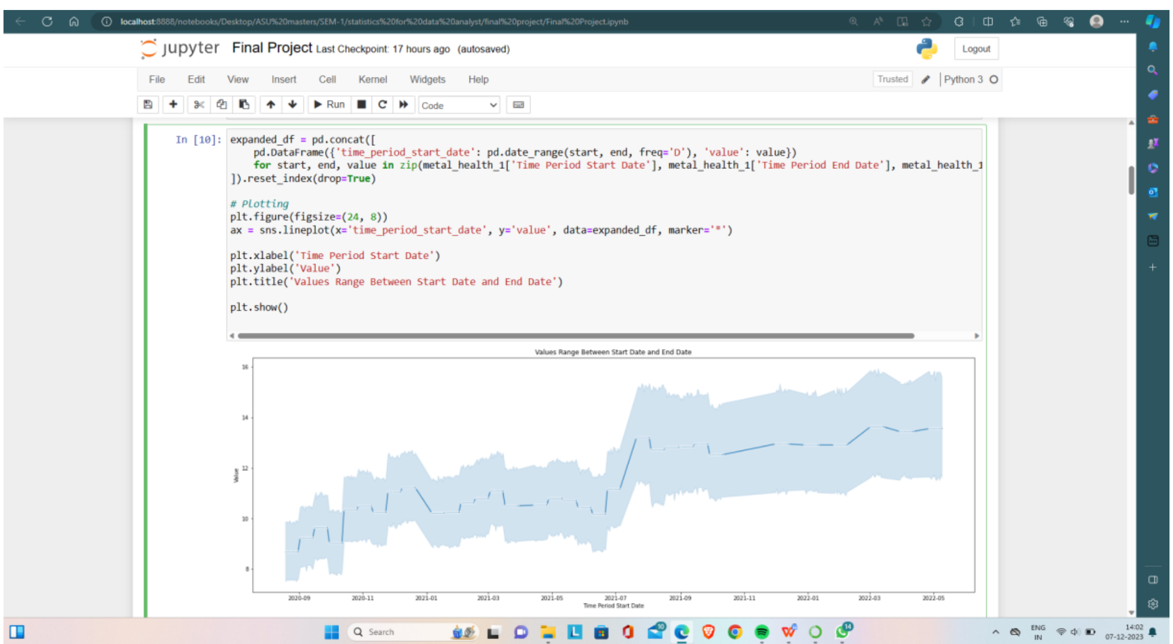
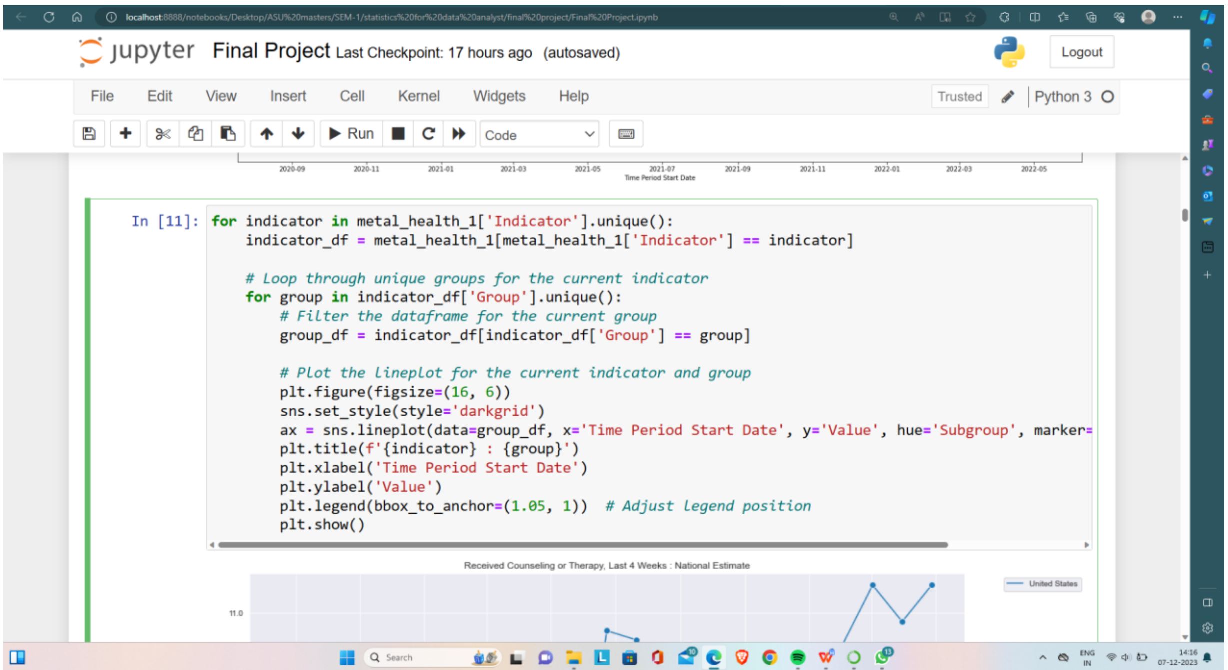


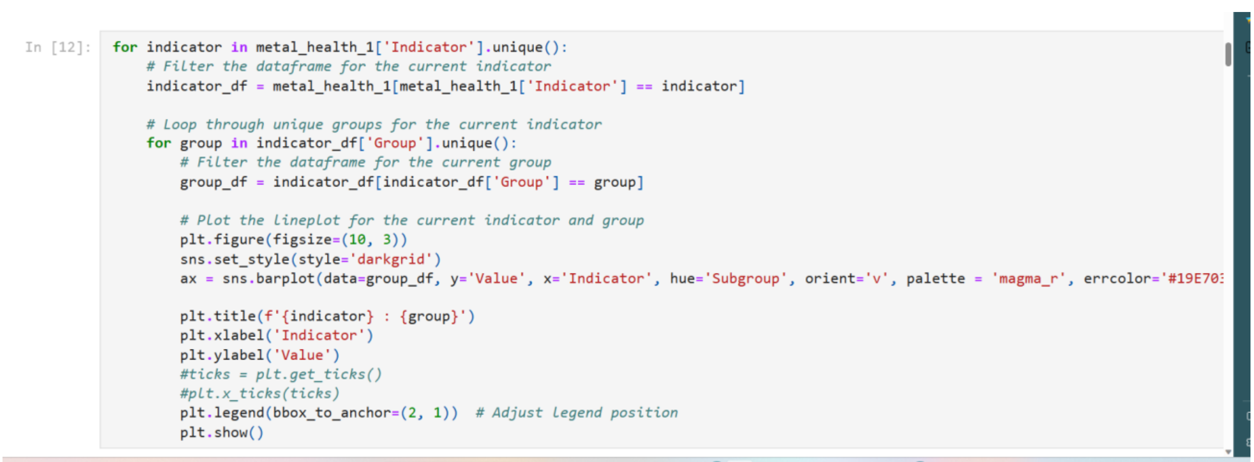
Fig 4: Line plot for value vs Time period

For each selected indicator, "Received Counseling or Therapy, Last 4 Weeks" and "Needed Counseling or Therapy But Did Not Get It, Last 4 Weeks," a line plot has been employed to assess the evolution of the value measure over the specified time period, spanning from mid-2020 to 2022. This time frame is deemed particularly pertinent as it encapsulates a crucial phase of the COVID-19 pandemic. The utilization of line plots offers a visual representation of trends and fluctuations, enabling a clear depiction of how each group's value has changed over time.



```
In [11]: for indicator in metal_health_1['Indicator'].unique():
             indicator_df = metal_health_1[metal_health_1['Indicator'] == indicator]

             # Loop through unique groups for the current indicator
             for group in indicator_df['Group'].unique():
                 # Filter the dataframe for the current group
                 group_df = indicator_df[indicator_df['Group'] == group]

                 # Plot the lineplot for the current indicator and group
                 plt.figure(figsize=(16, 6))
                 sns.set_style(style='darkgrid')
                 ax = sns.lineplot(data=group_df, x='Time Period Start Date', y='Value', hue='Subgroup', marker=
                 plt.title(f'{indicator} : {group}')
                 plt.xlabel('Time Period Start Date')
                 plt.ylabel('Value')
                 plt.legend(bbox_to_anchor=(1.05, 1))  # Adjust legend position
                 plt.show()
```

Fig 5: Code For line plot



```
In [12]: for indicator in metal_health_1['Indicator'].unique():
             # Filter the dataframe for the current indicator
             indicator_df = metal_health_1[metal_health_1['Indicator'] == indicator]

             # Loop through unique groups for the current indicator
             for group in indicator_df['Group'].unique():
                 # Filter the dataframe for the current group
                 group_df = indicator_df[indicator_df['Group'] == group]

                 # Plot the lineplot for the current indicator and group
                 plt.figure(figsize=(10, 3))
                 sns.set_style(style='darkgrid')
                 ax = sns.barplot(data=group_df, y='Value', x='Indicator', hue='Subgroup', orient='v', palette = 'magma_r', errcolor='#19E70

                 plt.title(f'{indicator} : {group}')
                 plt.xlabel('Indicator')
                 plt.ylabel('Value')
                 #ticks = plt.get_ticks()
                 #plt.x_ticks(ticks)
                 plt.legend(bbox_to_anchor=(2, 1))  # Adjust legend position
                 plt.show()
```

Fig 6: Code for Histogram plot

2.1.1 Received Counseling or Therapy for group By Age

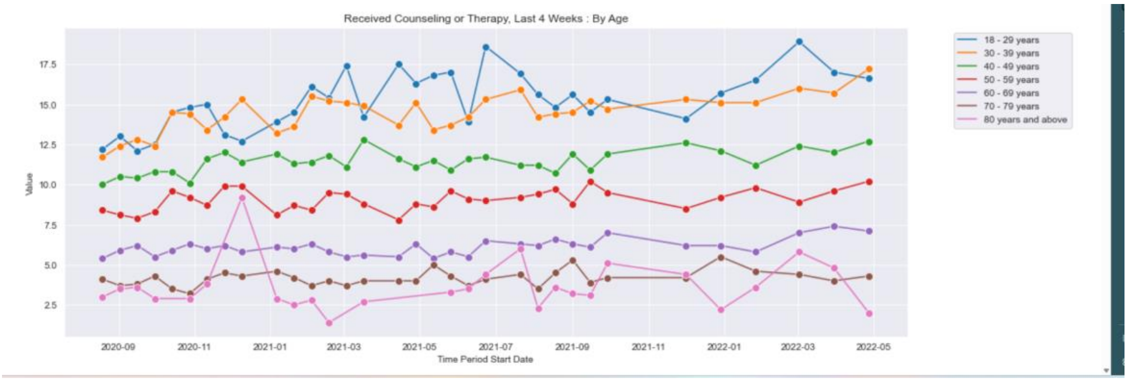


Fig 7: Value over Time Period for Group Age

Fig 8: Value for different categories in Age

Inference: Throughout the time span spanning from 2020 to 2022, individuals in the younger and middle age brackets consistently exhibit higher value measures compared to their older counterparts. Figure 8 further illustrates that the histogram values for the younger and middle-aged demographics are notably elevated. Consequently, for hypothesis testing, the age groups of 18-29 years, 30-39 years, and 40-49 years have been singled out for analysis.

2.1.2 Received Counseling or Therapy for group By Sex

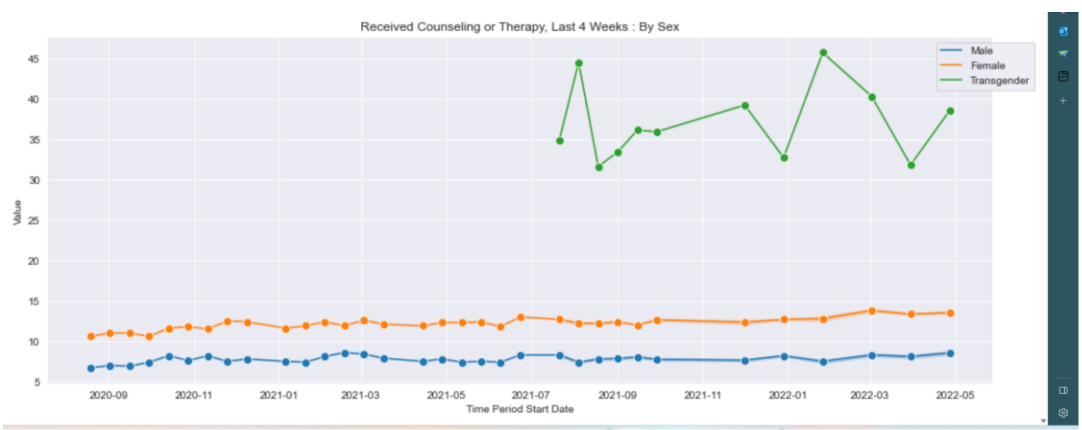


Fig 9: Value over Time Period for Group Sex

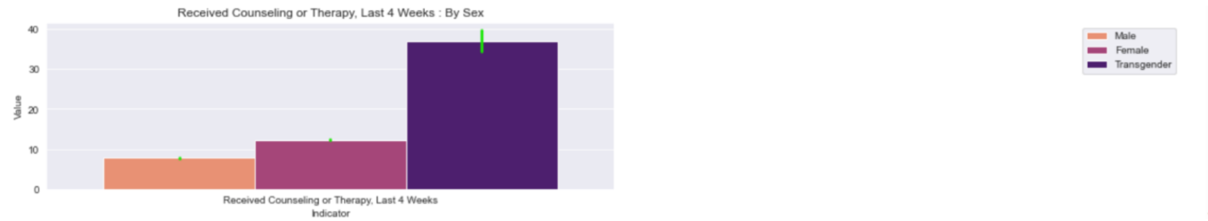


Fig 10: Value for different categories in Sex

Inference: The values for males and females demonstrate a relatively constant trend throughout the period from 2020 to 2022. In contrast, transgenders exhibit higher values, which can be attributed to the relatively smaller number of records for this group. It's noteworthy that the values for transgenders are observed more prominently in the later phase of the time period, suggesting a potential shift or emerging trend in mental health service utilization within this demographic. The concentration of higher values for transgenders may indicate evolving patterns in seeking and receiving counseling or therapy services, warranting further investigation into the factors influencing this observed change.

2.1.3 Received Counseling or Therapy for group By Race

Inference: Over the time span from 2020 to 2022, non-Hispanic individuals of other races consistently exhibit higher values, indicating a sustained trend in mental health service utilization. In contrast, non-Hispanic Asians show a comparatively narrower range of values, suggesting a potentially less fluctuating pattern. Non-Hispanic black, white, and Hispanic individuals present a mixed line plot of values, indicating variability and potential shifts in mental health service utilization within these groups

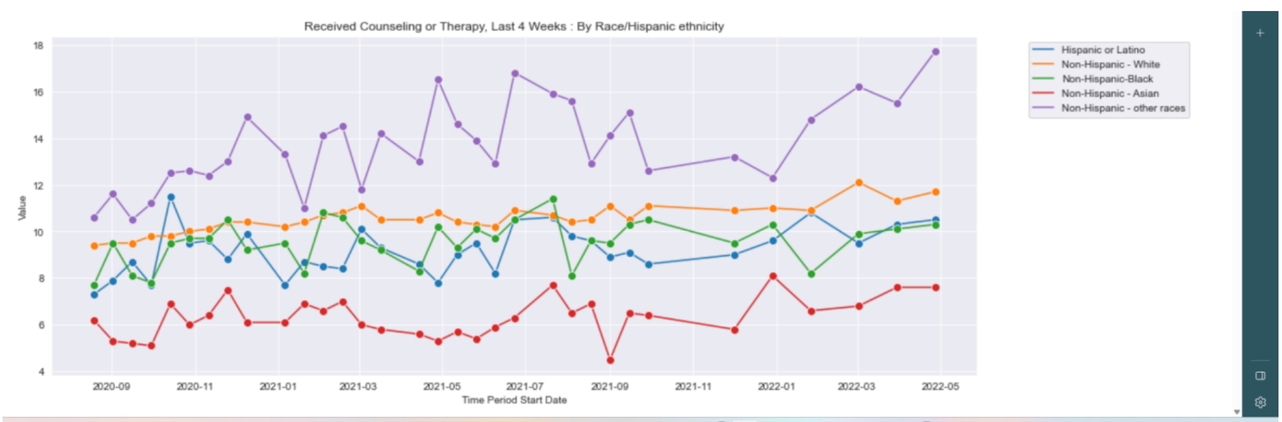


Fig 11: Value over Time Period for Group Race

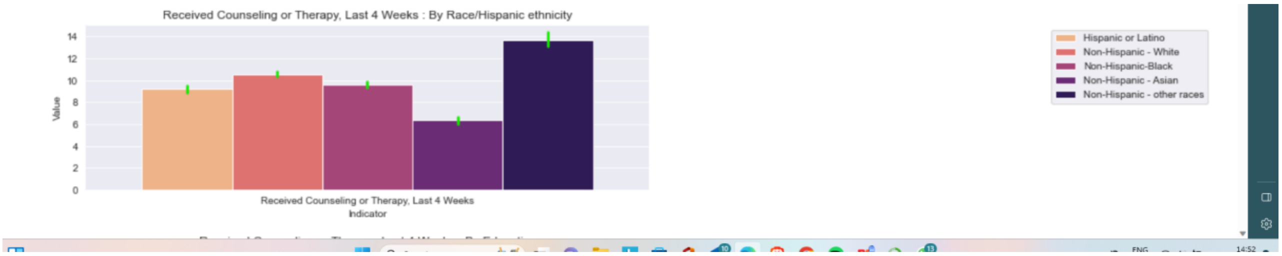


Fig 12: Value for different categories in Race

2.1.4 Received Counseling or Therapy for group By education

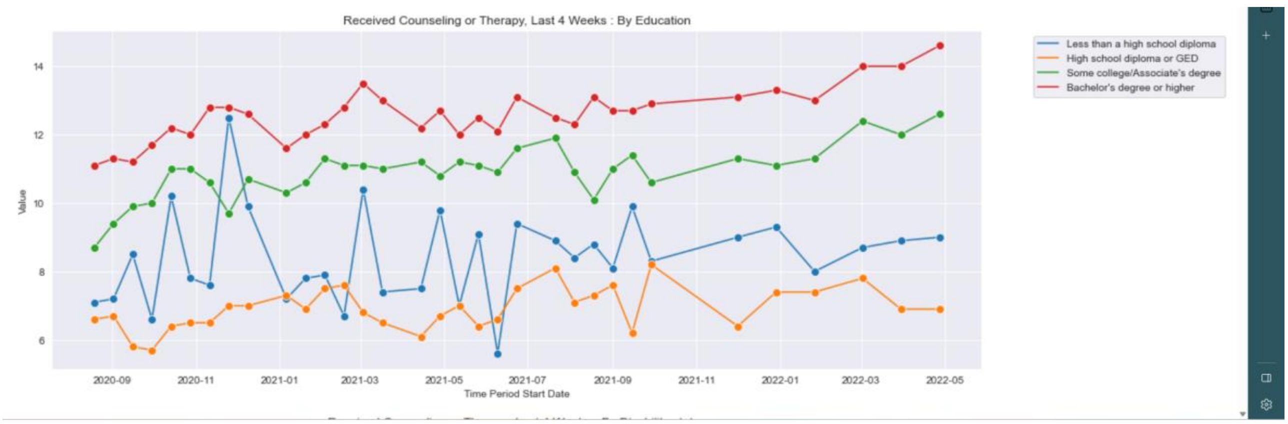


Fig 13: Value over Time Period for Group Education

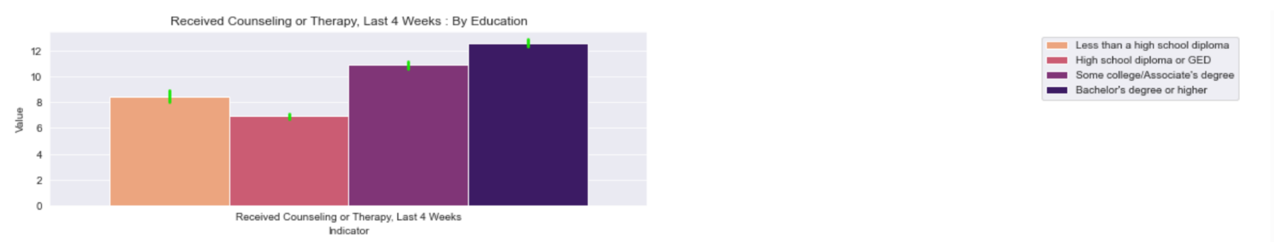


Fig 14: Value for different categories in Education

Inference: The data reveals a distinct pattern where individuals with higher education levels, such as college, bachelor, or associate degree holders, consistently exhibit a broader range of values over the entire period from 2020 to 2022. In contrast, those who have completed high school or have a diploma or lower education level show a comparatively narrower range of values. This implies that individuals with higher educational attainment are more likely to seek and receive mental health treatment, as reflected in the higher values. The observed trend underscores a potential correlation between education level and mental health service utilization, suggesting that those with more education may be more proactive in seeking and accessing mental health support

2.1.5 Needed Counseling or Therapy But Did Not Get It, for group By Age

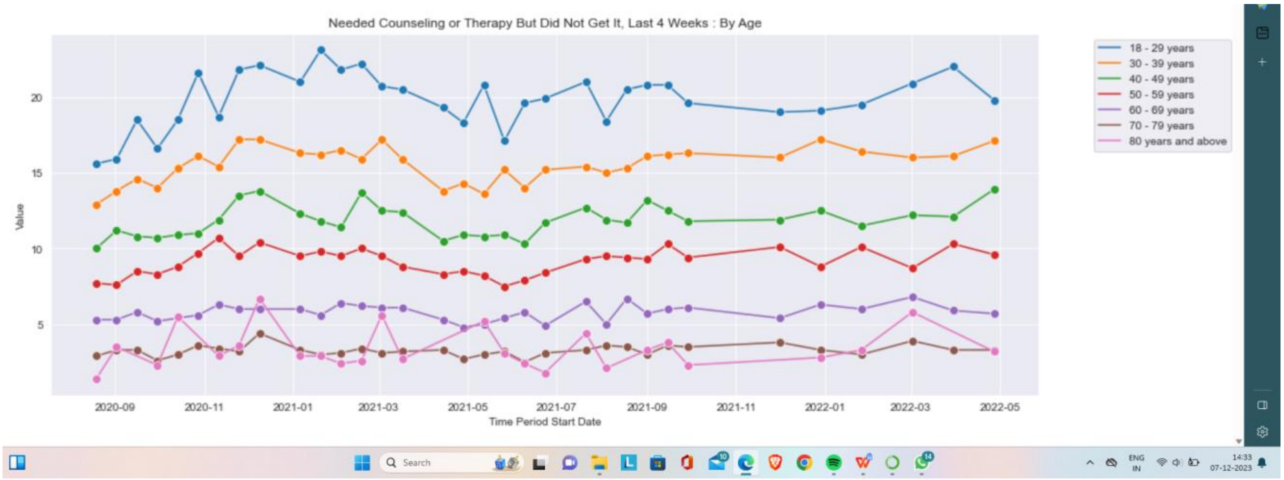


Fig 15: Value over Time Period for Group Age

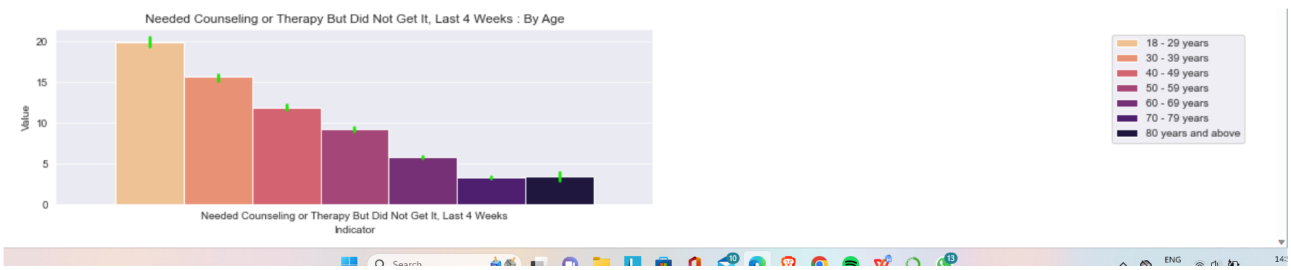


Fig 16: Value for different categories in Age

2.1.6 Needed Counseling or Therapy But Did Not Get It for group By Sex

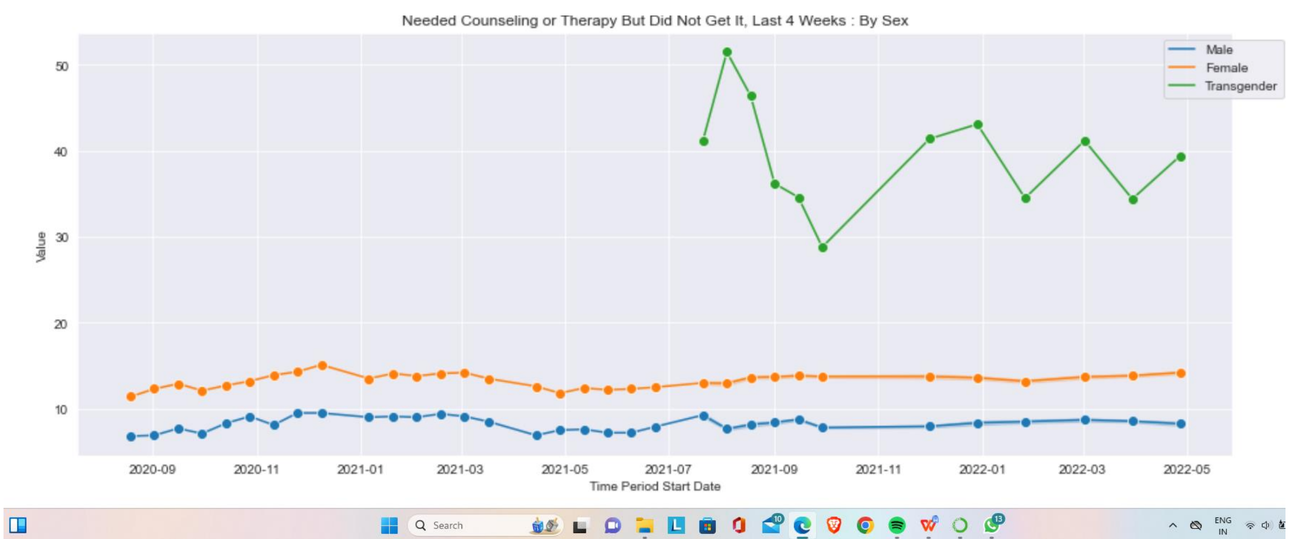


Fig 17: Value over Time Period for Group Sex

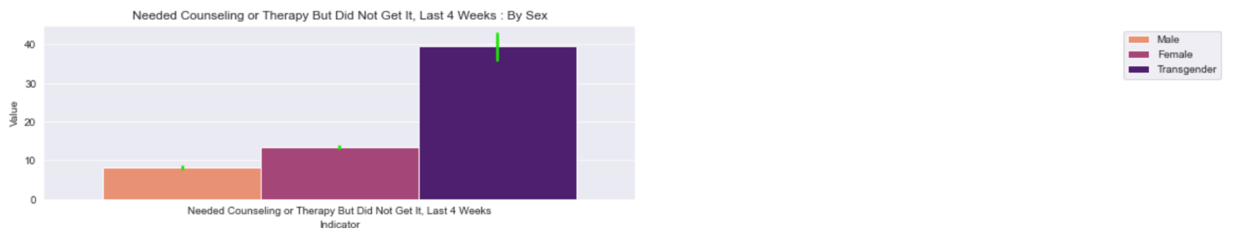


Fig 18: Value for different categories in Sex

2.1.7 Needed Counseling or Therapy But Did Not Get It for group By Race

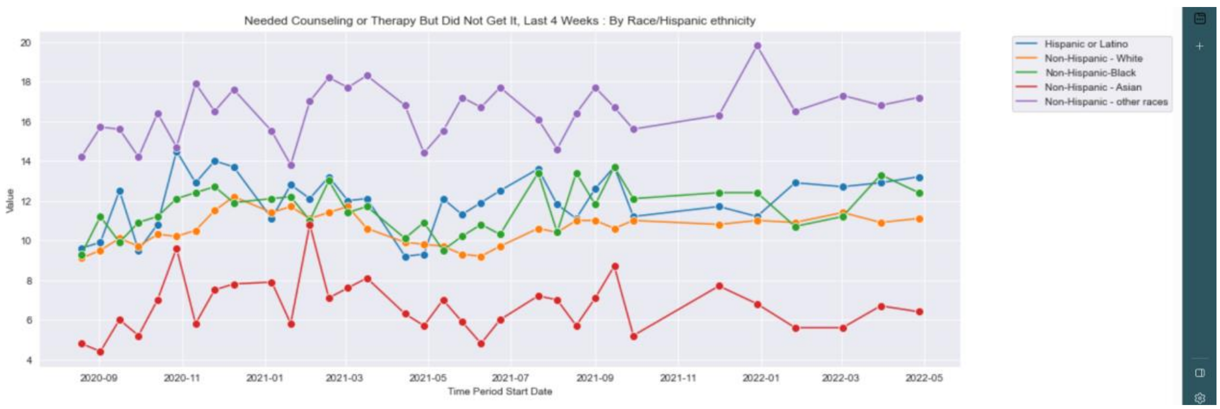


Fig 19: Value over Time Period for Group Race

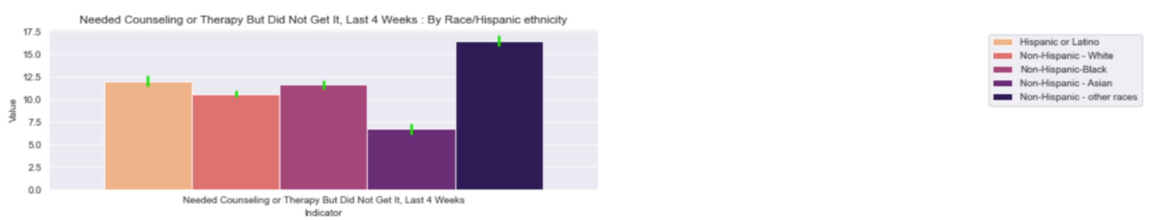


Fig 20: Value for different categories in Race

2.1.8 Needed Counseling or Therapy But Did Not Get It for group By education

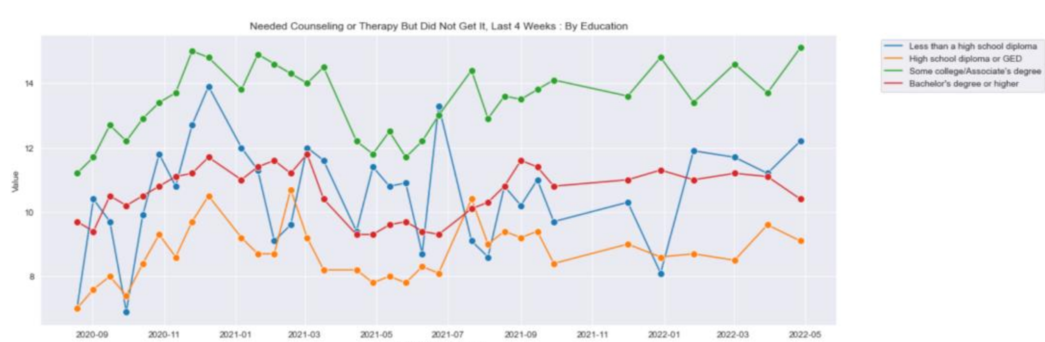


Fig 21: Value over Time Period for Group Education

Fig 22: Value for different categories in Education

2.2 Analyzing Group value ‘By State’

The latter part of the visualization is dedicated to the group with values categorized by state, with a notable division of all states into five distinct zones based on their geographic locations. These zones, namely Northeast, West, Midwest, Southwest, and Southeast, serve as a geographical framework for organizing and analyzing the data. The segmentation into zones is informed by the details presented in [1], providing a clear and systematic approach to examining mental health service utilization patterns across different regions of the country. This regional stratification enhances the precision of the analysis, allowing for a more targeted exploration of trends and variations in mental health indicators within specific geographic contexts.

Fig 23: Line plot for value vs time period for zones

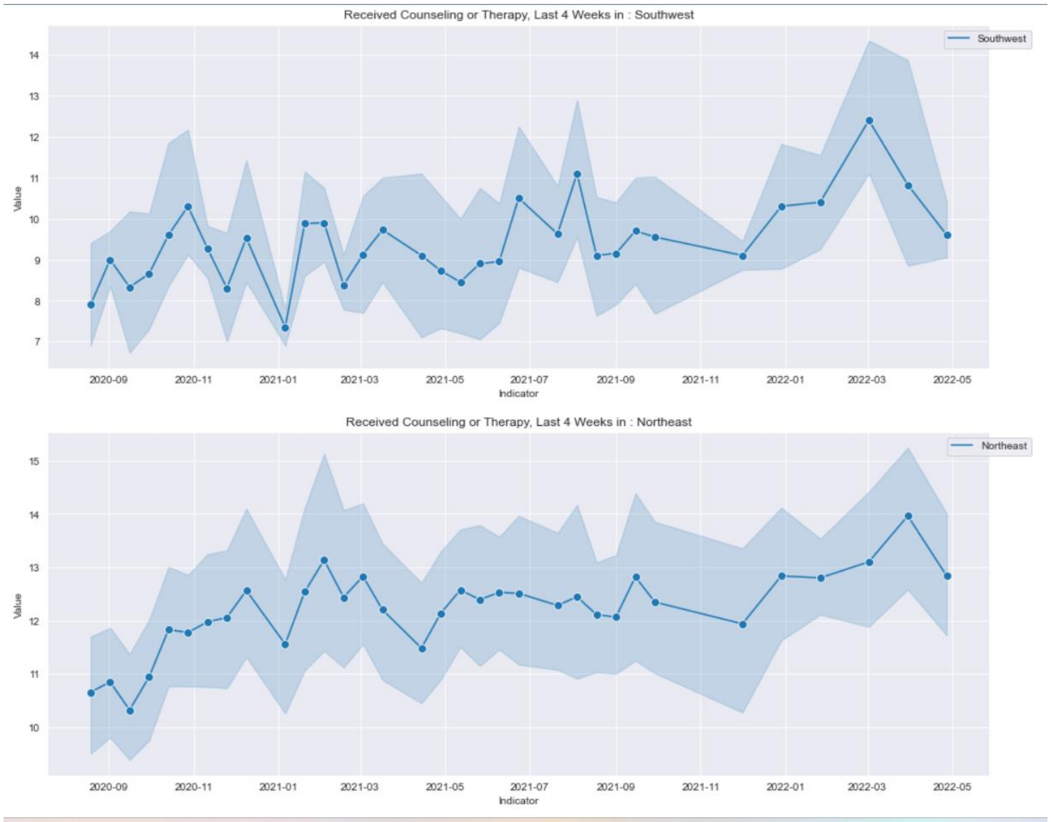


Fig 24: Line plot for value vs time period for zones

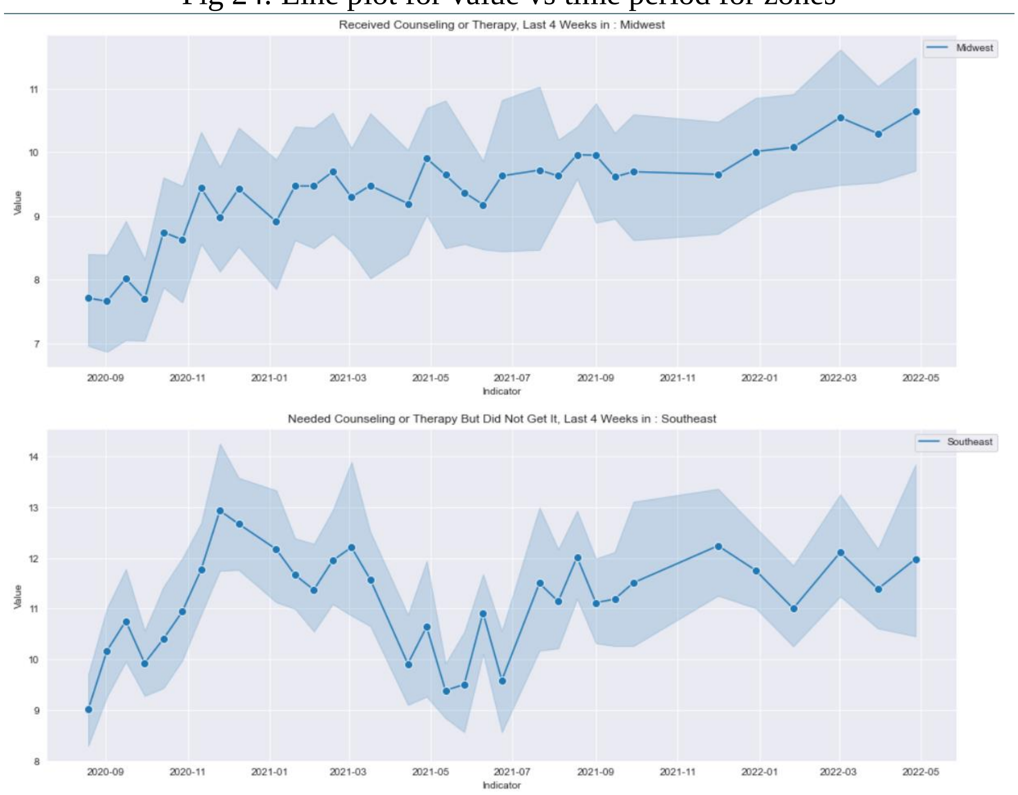


Fig 25: Line plot for value vs time period for zones

Fig 26: Line plot for value vs time period for zones

Fig 27: Line plot for value vs time period for zones

III HYPOTHESIS MODELLING

3.1 HYPOTHESIS 1

H0(Null hypothesis): There is no Significant difference in 'value' between the specified age group and the other age groups.

H1(Alternate Hypothesis): 18-29 years are more likely to receive Counseling or therapy over other age groups.

Throughout the entire time period spanning from 2020 to 2022, the analysis aims to assess whether a significant difference exists in the value measure for the specified age group (18-29 years) when compared to the remaining age groups. The null hypothesis posits that there is no noteworthy disparity in the likelihood of receiving counseling or therapy among different age groups, while the alternative hypothesis suggests a higher likelihood for individuals aged 18-29 years. The investigation seeks to determine the presence of any statistically significant variations in the value measure across these age groups during the specified time frame.

Since the distribution of the value measure is non-normal, the Mann-Whitney U test is preferred over other tests. The Mann-Whitney U test, a non-parametric test, is well-suited for comparing two independent groups when the assumption of normality is not met. By opting for this test, the analysis acknowledges and accommodates the non-normal distribution of the data, ensuring a robust and accurate assessment of the significance of differences in the value measure between the specified age group (18-29 years) and the other age groups over the entire time period from 2020 to 2022.

3.1.1 CODE:

```
# Perform Mann-Whitney U test for each combination of indicator and age group
for indicator in indicators:
    for age_group in age_groups:
        group1 = d2['Value'][(d2['Indicator'] == indicator) & (d2['Subgroup'] == age_group)]
        group2 = d2['Value'][(d2['Indicator'] == indicator) & (d2['Subgroup'] != age_group)]

        # Perform Mann-Whitney U test for independent samples
        stat, p_value = mannwhitneyu(group1, group2, alternative='two-sided')

        alpha = 0.05
        print()
        # Compare the p-value to the significance level
```

```
    if p_value < alpha:
        print(f"For the indicator '{indicator}' and age group '{age_group}'")
        print(p_value)
        print(f"Reject the null hypothesis.")
        print(f"There is a significant difference in 'value' between the specified age group and the
other age groups.")
    else:
        print(f"For the indicator '{indicator}' and age group '{age_group}'")
        print(p_value)
        print(f"fail to reject the null hypothesis.")
        print(f"There is no significant difference in 'value' between the specified age group and
the other age groups.")
```

3.1.2 OUTPUT:

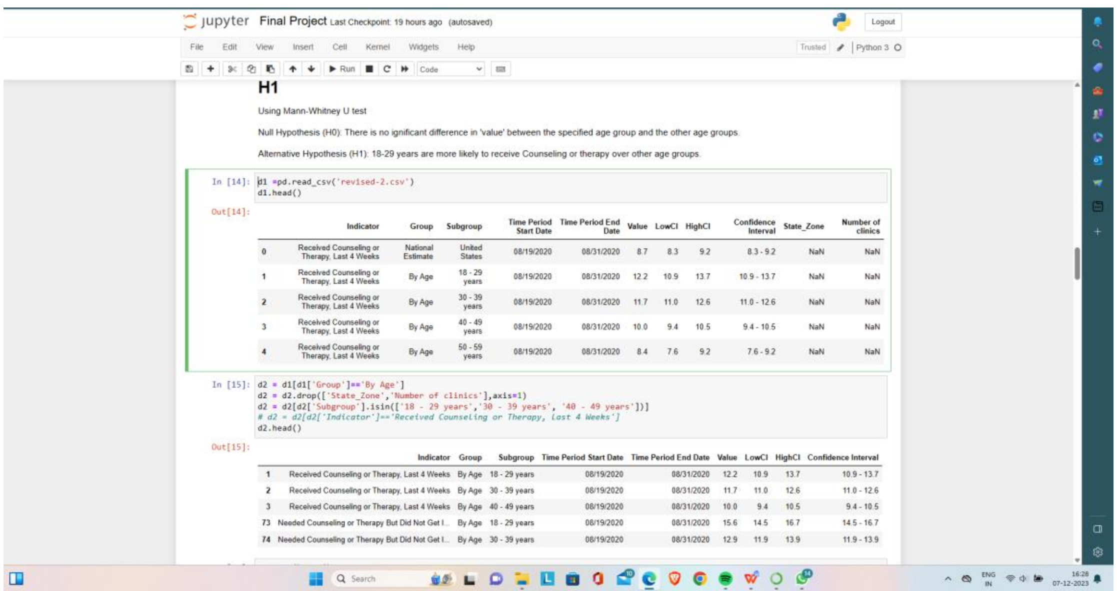


Fig 28: Hypothesis 1 testing

```
In [16]: d2.isna().sum()

Out[16]: Indicator                  0
         Group                      0
         Subgroup                   0
         Time Period Start Date     0
         Time Period End Date       0
         Value                     30
         LowCI                     30
         HighCI                    30
         Confidence Interval       30
         dtype: int64

In [17]: d2 = d2.dropna()

In [18]: import pandas as pd
         from scipy.stats import mannwhitneyu

         # Assuming 'df' is your DataFrame
         age_groups = d2['Subgroup'].unique()
         indicators = d2['Indicator'].unique()

         # Perform Mann-Whitney U test for each combination of indicator and age group
         for indicator in indicators:
             for age_group in age_groups:
                 group1 = d2['Value'][(d2['Indicator'] == indicator) & (d2['Subgroup'] == age_group)]
                 group2 = d2['Value'][(d2['Indicator'] == indicator) & (d2['Subgroup'] != age_group)]

                 # Perform Mann-Whitney U test for independent samples
                 stat, p_value = mannwhitneyu(group1, group2, alternative='two-sided')

                 # Check the significance level (commonly 0.05)
                 alpha = 0.05
                 print()
                 # Compare the p-value to the significance level
                 if p_value < alpha:
                     print(f"For the indicator '{indicator}' and age group '{age_group}'")
                     print(p_value)
                     print(f"Reject the null hypothesis.")
                     print(f"There is a significant difference in 'value' between the specified age group and the other age groups.")
                 else:
                     print(f"For the indicator '{indicator}' and age group '{age_group}'")
                     print(p_value)
                     print(f"fail to reject the null hypothesis.")
                     print(f"There is no significant difference in 'value' between the specified age group and the other age groups.")
```

Fig 29: Hypothesis 1 Mann-Whitney U test

```
File  Edit  View  Insert  Cell  Kernel  Widgets  Help          Trusted   Python 3

Run   Code

For the indicator 'Received Counseling or Therapy, Last 4 Weeks' and age group '18 - 29 years'
4.995878164591386e-07
Reject the null hypothesis.
There is a significant difference in 'value' between the specified age group and the other age groups.

For the indicator 'Received Counseling or Therapy, Last 4 Weeks' and age group '30 - 39 years'
0.0048939451244582825
Reject the null hypothesis.
There is a significant difference in 'value' between the specified age group and the other age groups.

For the indicator 'Received Counseling or Therapy, Last 4 Weeks' and age group '40 - 49 years'
4.360131559155693e-15
Reject the null hypothesis.
There is a significant difference in 'value' between the specified age group and the other age groups.

For the indicator 'Needed Counseling or Therapy But Did Not Get It, Last 4 Weeks' and age group '18 - 29 years'
9.146721863507442e-15
Reject the null hypothesis.
There is a significant difference in 'value' between the specified age group and the other age groups.

For the indicator 'Needed Counseling or Therapy But Did Not Get It, Last 4 Weeks' and age group '30 - 39 years'
0.806464263276311
fail to reject the null hypothesis.
There is no significant difference in 'value' between the specified age group and the other age groups.

For the indicator 'Needed Counseling or Therapy But Did Not Get It, Last 4 Weeks' and age group '40 - 49 years'
1.2513771476011867e-15
Reject the null hypothesis.
There is a significant difference in 'value' between the specified age group and the other age groups.
```

Fig 30: Hypothesis 1 Result

3.1.3 ANALYSIS:

1. For indicator “Received counseling or therapy” P_values for all age groups have been less than 0.05 which indicated that Reject the Null hypothesis. It means that there are significant differences between age groups in receiving the treatment.

2. For indicator “Neededcounseling or therapy But did not get it” P_values for age groups 18-29 years and 40-49 years reject the null hypothesis suggesting that there are significant differences between age groups in not receiving the treatment. P_value for 30-39 years fail to reject null hypothesis.

3. Possible challenges faced by age groups”
    - Age group 18-29 years might have experienced unique stressors during the pandemic, such as disruptions in education and early career challenges, social isolation, and uncertainty about the future, leading to a higher need or utilization of mental health services.
    - Individuals in 30-39 years of age group likely faced significant challenges related to job security, balancing remote work with family responsibilities, and managing the stress of potentially caring for young children or aging parents.
    - For 40-39 years group, the pandemic might have exacerbated stress related to job stability, health concerns, and increased family responsibilities, including the care of teenage or adult children and elderly parents.

4. Possible measures to Undergo:
    - Targeted Mental Health Interventions - Developing specific mental health programs for each age group, considering the unique challenges posed by the pandemic.[2]
    - Implementing outreach programs in schools, workplaces, and community centers to address the specific needs of these age groups.
    - Policy Making - Crafting policies that prioritize mental health resources for age groups most affected by the pandemic. Allocating funding and resources to enhance access to mental health services, especially in areas where these age groups predominantly reside.
    - Awareness and Education Campaigns - Launching awareness campaigns to reduce the stigma around seeking mental health care, tailored to the cultural and social contexts of each age group.
    - Educating the public about the importance of mental health and available resources, using age-appropriate messaging and platforms.

3.2 HYPOTHESIS –2

Null Hypothesis (H0): There is no significant difference in the distribution of value across different racial categories.

Alternative Hypothesis (H1): There is significant difference in the distribution of value across different racial categories.

The analysis conducted over the entire time period from 2020 to 2022 seeks to examine whether there is a meaningful distinction in the distribution of the value measure among various racial categories. The null hypothesis posits that there is no substantial difference, while the alternative hypothesis suggests the presence of a significant divergence in the distribution of the value measure across different racial groups. The evaluation aims to uncover and quantify any statistically noteworthy dissimilarities in the distribution of mental health service utilization among racial categories during the specified time frame.

Since the value measure is not normally distributed, the Kruskal-Wallis test is employed for hypothesis testing to assess whether there is a significant difference in the distribution of the value measure across different racial categories. Following the Kruskal-Wallis test, Tukey's Honest Significant Difference (HSD) test is utilized to conduct pairwise comparisons and calculate p-values for each pair of racial categories. This approach allows for a comprehensive examination of the specific racial groups that contribute to any observed significant differences, providing detailed insights into the nuanced patterns of mental health service utilization among different racial categories over the entire time period from 2020 to 2022. The hypothesis testing for bothe the indicator is performed separately.

3.2.1 CODE:

```
measurements = d4['Value']
sm.qqplot(measurements, line='45')
pylab.show()

result = kruskal(
    d4['Value'][d4['Subgroup'] == 'Hispanic or Latino'],
    d4['Value'][d4['Subgroup'] == 'Non-Hispanic - Asian'],
    d4['Value'][d4['Subgroup'] == 'Non-Hispanic-Black'],
    d4['Value'][d4['Subgroup'] == 'Non-Hispanic - White'],
    d4['Value'][d4['Subgroup'] == 'Non-Hispanic - other races']
)

print(f"Kruskal-Wallis H-statistic: {result.statistic}")
print(f"P-value: {result.pvalue}")
```

```
alpha = 0.05

if result.pvalue < alpha:
    print("Reject the null hypothesis. Non-Hispanic white individuals have a significantly
different distribution of receiving treatment compared to individuals of other races.")
else:
    print("Fail to reject the null hypothesis. There is no significant difference in the distribution of
receiving treatment across different race categories.")

#Tukey's Honest Significant Difference (HSD) Test
from statsmodels.stats.multicomp import pairwise_tukeyhsd

tukey_results = pairwise_tukeyhsd(d4['Value'], d4['Subgroup'])
print(tukey_results.summary())
```

3.2.2 OUTPUT FOR INDICATOR 'RECEIVED COUNSELING OR THERAPY, LAST 4 WEEKS'

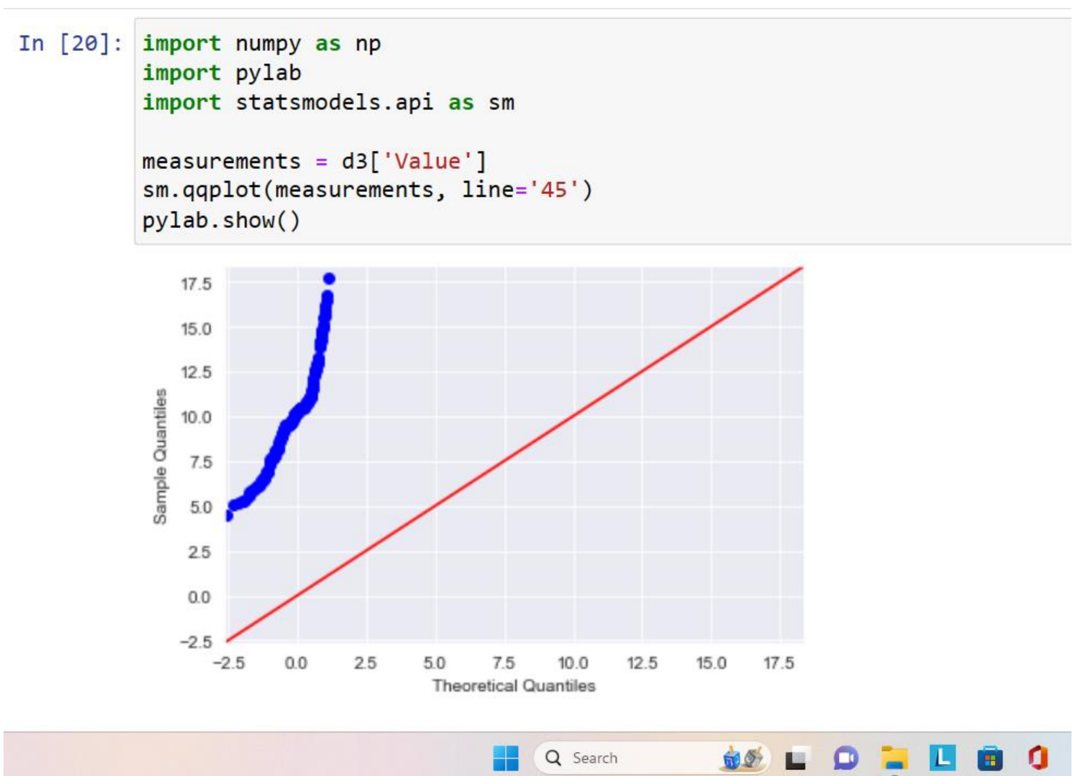


Fig 31: Normality test for Value measure when indicator is Received counselling or therapy

```
In [22]: from scipy.stats import kruskal

         # Assuming 'df' is your DataFrame
         result = kruskal(
             d3['Value'][d3['Subgroup'] == 'Hispanic or Latino'],
             d3['Value'][d3['Subgroup'] == 'Non-Hispanic - Asian'],
             d3['Value'][d3['Subgroup'] == 'Non-Hispanic-Black'],
             d3['Value'][d3['Subgroup'] == 'Non-Hispanic - White'],
             d3['Value'][d3['Subgroup'] == 'Non-Hispanic - other races']
                     )

         print(f"Kruskal-Wallis H-statistic: {result.statistic}")
         print(f"P-value: {result.pvalue}")

         # Check the significance level (commonly 0.05)
         alpha = 0.05

         # Compare the p-value to the significance level
         if result.pvalue < alpha:
             print("Reject the null hypothesis. There is significant difference in the distribution of receiving treatment across differen
         else:
             print("Fail to reject the null hypothesis. There is no significant difference in the distribution of receiving treatment acro

         Kruskal-Wallis H-statistic: 134.2178728838784
         P-value: 4.877118585685933e-28
         Reject the null hypothesis. Non-Hispanic white individuals have a significantly different distribution of receiving treatment c
         ompared to individuals of other races.
```

Fig 32: Hypothesis 2 Kruskal test

```
In [23]: from statsmodels.stats.multicomp import pairwise_tukeyhsd

         # Assuming 'df' is your DataFrame
         tukey_results = pairwise_tukeyhsd(d3['Value'], d3['Subgroup'])
         print(tukey_results.summary())

                        Multiple Comparison of Means - Tukey HSD, FWER=0.05
         ==============================================================================================
                   group1                      group2            meandiff p-adj   lower   upper  reject
         ----------------------------------------------------------------------------------------------
                 Hispanic or Latino       Non-Hispanic - Asian  -2.8848  0.001 -3.6518 -2.1179   True
                 Hispanic or Latino       Non-Hispanic - White   1.3455  0.001  0.5785  2.1124   True
                 Hispanic or Latino Non-Hispanic - other races   4.4939  0.001   3.727  5.2609   True
                 Hispanic or Latino         Non-Hispanic-Black   0.3606 0.6698 -0.4063  1.1275  False
               Non-Hispanic - Asian       Non-Hispanic - White   4.2303  0.001  3.4634  4.9972   True
               Non-Hispanic - Asian Non-Hispanic - other races   7.3788  0.001  6.6119  8.1457   True
               Non-Hispanic - Asian         Non-Hispanic-Black   3.2455  0.001  2.4785  4.0124   True
               Non-Hispanic - White Non-Hispanic - other races   3.1485  0.001  2.3816  3.9154   True
               Non-Hispanic - White         Non-Hispanic-Black  -0.9848 0.0047 -1.7518 -0.2179   True
         Non-Hispanic - other races         Non-Hispanic-Black  -4.1333  0.001 -4.9002 -3.3664   True
         ----------------------------------------------------------------------------------------------
```

Fig 33: Hypothesis 2 Turkey Pairwise Test

3.2.3 ANALYSIS:

1. H-statistic (134.2178728838784) and P-value (4.877118585685933e-28): The high H-statistic and the extremely low p-value indicate strong statistical evidence that the distributions of receiving mental health treatment among different racial groups are not the same, with a particular emphasis on the differences involving non-Hispanic white individuals.

2. Turkey pairwise test suggest that for most of the pair p_value is less and strongly supporting the rejection of null hypothesis. This means that there are differences in treatment received by people from different races.

3. Possible situations causing such results:
    - Non-Hispanic white individuals might have had different levels of access to mental health services compared to other racial groups. This could be due to factors such

as socioeconomic status, healthcare coverage, and availability of services in different communities.
- The pandemic's impact varied across racial groups, with some communities experiencing higher rates of infection, economic hardship, and stress. These disparities could influence the need and ability to seek mental health treatment.
- Cultural perceptions and stigma surrounding mental health can vary greatly between racial groups, potentially influencing the likelihood of seeking treatment.

4. Possible measures to undergo:
   - Developing policies that address the disparities in mental health care access and utilization among different racial groups. Implementing programs to increase accessibility of mental health services in underserved communities, especially those heavily impacted by the pandemic.[3]
   - Community Outreach and Education: Initiating community-based outreach programs to raise awareness about mental health issues and the importance of seeking help, with a focus on culturally sensitive approaches.
   - Enhancing Telehealth Services: Expanding telehealth services to reach populations that may have limited access to in-person mental health care. Ensuring that telehealth services are equitable and accessible to all racial groups, considering factors like internet access and digital literacy.
   - Research on Racial Disparities in Mental Health: Conducting research to further understand the root causes of racial disparities in mental health care utilization, especially during crises like the COVID-19 pandemic. Exploring the effectiveness of different mental health interventions across diverse racial groups.

3.2.4 OUTPUT FOR INDICATOR ‘NEEDED COUNSELING OR THERAPY BUT DID NOT GET IT, LAST 4 WEEKS’

```
In [26]: measurements = d4['Value']
         sm.qqplot(measurements, line='45')
         pylab.show()
```

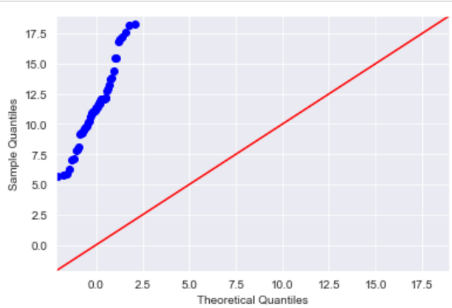


Fig 34: Normality test for Value measure when indicator is Needed counselling or therapy but did not get it

```
In [27]: result = kruskal(
    d4['Value'][d4['Subgroup'] == 'Hispanic or Latino'],
    d4['Value'][d4['Subgroup'] == 'Non-Hispanic - Asian'],
    d4['Value'][d4['Subgroup'] == 'Non-Hispanic-Black'],
    d4['Value'][d4['Subgroup'] == 'Non-Hispanic - White'],
    d4['Value'][d4['Subgroup'] == 'Non-Hispanic - other races']
             )

print(f"Kruskal-Wallis H-statistic: {result.statistic}")
print(f"P-value: {result.pvalue}")

# Check the significance level (commonly 0.05)
alpha = 0.05

# Compare the p-value to the significance level
if result.pvalue < alpha:
    print("Reject the null hypothesis. Non-Hispanic white individuals have a significantly different distribution of receiving tr
else:
    print("Fail to reject the null hypothesis. There is no significant difference in the distribution of receiving treatment acro

Kruskal-Wallis H-statistic: 36.9382622903551
P-value: 1.8548330627020784e-07
Reject the null hypothesis. Non-Hispanic white individuals have a significantly different distribution of receiving treatment c
ompared to individuals of other races.
```

Fig 35: Hypothesis 2 Kruskal test

```
In [28]: #Tukey's Honest Significant Difference (HSD) Test
from statsmodels.stats.multicomp import pairwise_tukeyhsd

# Assuming 'df' is your DataFrame
tukey_results = pairwise_tukeyhsd(d4['Value'], d4['Subgroup'])
print(tukey_results.summary())

                  Multiple Comparison of Means - Tukey HSD, FWER=0.05
=========================================================================================
         group1                    group2            meandiff p-adj   lower   upper  reject
-----------------------------------------------------------------------------------------
       Hispanic or Latino       Non-Hispanic - Asian    -4.45 0.001 -6.1805 -2.7195   True
       Hispanic or Latino       Non-Hispanic - White    -0.98 0.4992 -2.7105  0.7505  False
       Hispanic or Latino Non-Hispanic - other races     4.74 0.001  3.0095  6.4705   True
       Hispanic or Latino         Non-Hispanic-Black    -0.43   0.9 -2.1605  1.3005  False
     Non-Hispanic - Asian       Non-Hispanic - White     3.47 0.001  1.7395  5.2005   True
     Non-Hispanic - Asian Non-Hispanic - other races     9.19 0.001  7.4595 10.9205   True
     Non-Hispanic - Asian         Non-Hispanic-Black     4.02 0.001  2.2895  5.7505   True
     Non-Hispanic - White Non-Hispanic - other races     5.72 0.001  3.9895  7.4505   True
     Non-Hispanic - White         Non-Hispanic-Black     0.55 0.8906 -1.1805  2.2805  False
Non-Hispanic - other races         Non-Hispanic-Black    -5.17 0.001 -6.9005 -3.4395   True
-----------------------------------------------------------------------------------------
```

Fig 36: Hypothesis 2 Turkey Pairwise Test

3.2.5 ANALYSIS:

1. H-statistic (36.9382622903551) and P-value (1.8548330627020784e-07): These results signify a statistically significant difference in the distribution of mental health treatment across racial groups. The high H-statistic combined with the very low p-value strongly suggests that non-Hispanic white individuals' access to or utilization of mental health treatment during the COVID-19 pandemic was different compared to other racial groups.

2. Turkey pairwise test suggest that for most of the pair p_value is less and strongly supporting the rejection of null hypothesis. This means that there are differences in treatment received by people from different races.

3. Possible situations:
    - Access to Care: The pandemic may have highlighted or intensified pre-existing inequalities in access to mental health care. Non-Hispanic white individuals may have had more access to mental health resources compared to other racial groups, influenced by factors like socioeconomic status, health insurance coverage, and geographic location of services.
    - Impact of COVID-19 on Different Racial Groups: Different racial groups experienced the pandemic's stressors differently. Some communities faced higher rates of COVID-19 infection, job losses, and other socioeconomic challenges, which could have impacted their mental health and ability to seek treatment.
    - Cultural and Stigma Factors: Cultural perceptions of mental health and the stigma associated with seeking treatment vary among racial groups, potentially influencing the utilization of mental health services.

4. Possible measures to undergo:
    - Policy Development and Resource Allocation: Crafting policies that specifically address the gap in mental health care access among racial groups. Allocating resources to increase the availability of mental health services in underserved and heavily impacted communities.
    - Targeted Community Outreach: Implementing community outreach programs to raise mental health awareness, with an emphasis on cultural sensitivity and inclusivity. Developing educational materials and programs that resonate with diverse communities, potentially in multiple languages.
    - Enhancing Telehealth and Digital Solutions: Expanding telehealth services to provide equitable access to mental health care, particularly for communities that may have faced barriers to traditional in-person therapy. Ensuring digital mental health solutions are accessible and user-friendly for all racial groups, considering factors like digital literacy and internet accessibility.
    - Research on Racial Disparities and Mental Health: Conducting in-depth research to further understand the causes and implications of racial disparities in mental health treatment, especially under the unique conditions of the COVID-19 pandemic. Studying the effectiveness of mental health interventions and support systems across different racial groups to inform future healthcare strategies.

3.3 HYPOTHESIS 3

Null Hypothesis(H0): The mean values of "value" and "number of clinics" are the same for people in Zone Southeast and Midwest compared to other zones.

Alternative Hypothesis (H1): The mean values of "value" and "number of clinics" are significantly different for people in Southeast and Midwest compared to other zones, with the expectation that they are higher.

This hypothesis testing aims to evaluate whether there is a significant difference in the mean values of "value" and "number of clinics" between individuals in the Southeast and Midwest zones compared to other geographic zones. The null hypothesis assumes equality in means, while the alternative hypothesis posits a significant difference, particularly anticipating higher mean values for individuals in the Southeast and Midwest zones. The analysis explores the potential regional variations in mental health service utilization and the availability of clinics over the specified time period.

The independent t-test offers numerous benefits, particularly when applied to the comparison of mean values within the "value" column across multiple zones, where the data may not follow a normal distribution. Its robustness to deviations from normality ensures reliable outcomes, even in non-ideal distribution situations. It is designed fot comparison of two independent groups with small sample sizes, the t-test is both efficient and easily interpreted through its t-statistic and p-value.

3.3.1 CODE

```
import pandas as pd
from scipy.stats import ttest_ind
indicator = ['Received Counseling or Therapy, Last 4 Weeks']
for i in indicator:
   indicated_df = d5[d5['Indicator']==i]
   southeast_midwest_data =
indicated_df[indicated_df['State_Zone'].isin(['Southeast','Midwest'])]
   other_zones_data = indicated_df[~indicated_df['State_Zone'].isin(['Southeast','Midwest'])
      t_statistic_clinics, p_value_clinics = ttest_ind(southeast_midwest_data['Number of clinics'],
other_zones_data['Number of clinics'])
    print(f"for indicator {i}")
   print("Two-sample t-test for 'number_of_clinics': t-statistic =", t_statistic_clinics, "p-value =",
p_value_clinics)
   if p_value_clinics < 0.05:
      print("Reject Null hypothesis. There is significant evidence to suggest that people in the
Southeast, Midwest zone are more likely to receive medical treatment.")
   else:
      print("Fail to reject the null hypothesis.There is no significant evidence to suggest a
difference in the likelihood of getting medical treatment between people in the Southeast,
Midwest zone and people in other zones.")
   print()
```

3.3.2 OUTPUT:

```
In [32]: import pandas as pd
         from scipy.stats import ttest_ind

         indicator = ['Received Counseling or Therapy, Last 4 Weeks']
         # Separate data for Zone North and other zones

         for i in indicator:
             indicated_df = d5[d5['Indicator']==i]

             southeast_midwest_data = indicated_df[indicated_df['State_Zone'].isin(['Southeast','Midwest'])]
             other_zones_data = indicated_df[~indicated_df['State_Zone'].isin(['Southeast','Midwest'])]

             # Perform two-sample t-test for 'number_of_clinics'
             t_statistic_clinics, p_value_clinics = ttest_ind(southeast_midwest_data['Number of clinics'], other_zones_data['Number of cli

             # Print the resus
             print(f"for indicator {i}")
             print("Two-sample t-test for 'number_of_clinics': t-statistic =", t_statistic_clinics, "p-value =", p_value_clinics)
             if p_value_clinics < 0.05:
                 print("Reject Null hypothesis. There is significant evidence to suggest that people in the Southeast, Midwest zone are mc
             else:
                 print("Fail to reject the null hypothesis.There is no significant evidence to suggest a difference in the likelihood of g
             print()

for indicator Received Counseling or Therapy, Last 4 Weeks
Two-sample t-test for 'number_of_clinics': t-statistic = 1.717573307638159 p-value = 0.08605886639389665
Fail to reject the null hypothesis.There is no significant evidence to suggest a difference in the likelihood of getting medica
l treatment between people in the Southeast, Midwest zone and people in other zones.
```

Fig 37: Hypothesis 3 Independent T-test

3.3.3 ANALYSIS:

1. T-statistic (1.717573307638159) and P-value (0.08605886639389665): The t-statistic suggests a slight difference between the groups, but the p-value, being higher than the conventional threshold of 0.05, leads to failing to reject the null hypothesis. This implies that there is no statistically significant evidence to suggest a difference in the likelihood of receiving mental health treatment between people in the Southeast, Midwest zone, and those in other zones.

2. This test fails to reject the null hypothesis which means that the medical treatment received by the people in different zones is relatively the same. This means that the zone facilities available to people are equal for all.

3. Possible situations:
   - The lack of significant differences suggests that, during the pandemic, access to mental health treatment such as counseling or therapy was relatively equitable across these regions. This is an important finding, as it indicates that regional disparities were not a significant factor in mental health care access during this period.
   - The pandemic's impact was felt unevenly across different regions. However, this result suggests that the disparities in the healthcare system, at least in terms of mental health care accessibility, were not as pronounced as might be expected between these specific regions.

IV FUTURE EXPERIMENTATION

The dataset's intentional design, ensuring an equal number of data points for each indicator and maintaining uniformity in subgroup values, provides a structured foundation for analysis. However, acknowledging the potential for increased variability and more nuanced insights, further refinement of the dataset with real-world data points that vary across indicators and groups could enhance its representativeness. Exploring additional personal information in subsequent research, such as educational backgrounds, could yield more insightful results. More numeric columns about the individual life or place can provide a better insight in hypothesis testing. This dataset has only 'Value' attribute to measure the hypothesis. More of such independent or dependent attributes can be useful.

Introducing a hypothesis focused on individuals seeking medication or therapy based on their education level is indeed a crucial avenue for investigation. Education plays a pivotal role in shaping individual attitudes toward treatment, and such an analysis could offer valuable insights into the factors influencing mental health service utilization across diverse educational backgrounds.

V CONCLUSION

In conclusion, the hypothesis testing conducted on the mental health care dataset has provided valuable insights into various factors influencing service utilization. The analysis encompassed diverse hypotheses, including age-based patterns, racial disparities, regional variations, and the impact of educational levels on seeking medication or therapy. The use of statistical tests such as Mann-Whitney U, Kruskal-Wallis, and independent t-tests, along with Tukey's Honest Significant Difference, facilitated robust evaluations. Key findings include age groups of 18-29 years demonstrating a higher likelihood of receiving counseling or therapy, racial disparities in mental health service utilization, and no potential regional variations in the Southeast and Midwest zones. Additionally, the dataset's structured design ensures a systematic exploration of mental health indicators. Further research, incorporating real-world data and additional personal information, promises to deepen our understanding of the complex dynamics influencing mental health care patterns. These findings hold implications for policy development, resource allocation, and targeted interventions to foster a more equitable and inclusive mental health care landscape.

REFERENCES

1. 5 US Regions Map and Facts | Mappr
2. COVID-19 impact on mental health - statistics & facts | Statista
3. Mental Health | NIH COVID-19 Research