**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

***Exploratory Data Analysis on Mental health care***

Submitted by:

Parikshit Kaushal

Registration No: 12323036

Programme and Section: K23FK

Course Code: INT375

Under the Guidance of

**Dr. Karan Bajaj**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that **Parikshit Kaushal** bearing Registration no **12323036** has completed **INT375** project titled, **“Exploratory Data Analysis on Mental health Dataset”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Karan Bajaj**

**Designation of the Supervisor**

**Assistant professor**

**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab.

Date: 12-04-2025

**DECLARATION**

I, **Parikshit Kaushal**, student of **B.Tech** under CSE Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12-04-2025 Signature

Registration No: **12323036** **Parikshit kaushal**

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to **Lovely Professional University** for providing me with the opportunity to work on this project titled **"Exploratory Data Analysis on Mental health Dataset"**

I would like to thank my respected faculty mentor, **Dr. Karan Bajaj**, for his valuable guidance, support, and encouragement throughout the course of this project. His insightful feedback and suggestions have played a crucial role in the successful completion of this work.

I am also thankful to the university's faculty and technical staff for creating a supportive and resourceful learning environment. Their teachings and assistance helped me apply theoretical knowledge to practical implementation.

Working on this Python project has not only enhanced my technical skills but also deepened my understanding of real-world data analysis and its applications.

**Reg No.: 12323036**  
**2nd Year, B.Tech (CSE)**  
**Lovely Professional University**

**TABLE OF CONTENT**

1. Introduction

2. Source of dataset

3. EDA process

4. Analysis on dataset (for each analysis)

1. Introduction
2. General Description
3. Specific Requirements, functions and formulas
4. Analysis results
5. Visualization

5. Conclusion

6. Future scope

7. References

**INTRODUCTION**

In recent years, the analysis of crime data has become increasingly important for understanding patterns, trends, and potential causes of criminal activity. With the growing availability of public datasets and powerful data analysis tools like Python, it is now possible to extract meaningful insights that can help authorities, researchers, and policymakers make informed decisions.

This project titled **"Exploratory Data Analysis on Mental health Dataset using Python"** focuses on analyzing a comprehensive dataset containing mental health . The dataset includes Mental health care

Data of different states and regions.

Using libraries such as **Pandas, NumPy, Matplotlib, and Seaborn**, this project aims to explore the data, clean and preprocess it, identify trends and patterns, and visually represent the distribution of different types of crimes over time. Additionally, efforts are made to predict future crime trends using non-machine learning approaches such as moving averages.

The ultimate goal of this analysis is to draw meaningful conclusions about the nature of crimes, identify mental health areas or agencies, and provide a foundation for further study or decision-making based on data-driven insights.

**Source of Dataset:** <https://catalog.data.gov/dataset/mental-health-care-in-the-last-4-weeks>

**EDA PROCESS**

The Exploratory Data Analysis (EDA) process was conducted using Python libraries such as **Pandas**, **NumPy**, **Matplotlib**, **Seaborn**, and **SciPy**. The primary objective was to explore the structure, distribution, and underlying trends in the mental health care dataset and extract meaningful insights that could assist in future research or predictive modeling.

### **Data Understanding**

* The dataset was loaded and its structure was examined using df.info(), df.head(), and df.shape.
* Key features such as Time Period, Group, Subgroup, Indicator, Value, LowCI, and HighCI were analyzed to understand their data types and significance.
* Temporal aspects were highlighted by extracting Year from the Time Period Start Date column for trend-based insights.

### **2. Data Cleaning**

* Missing values were identified using df.isnull().sum() and handled by imputation or removal depending on context.
* Duplicate entries were checked using df.duplicated().sum() and dropped to ensure data integrity.
* Date formats were standardized using pd.to\_datetime(), and categorical columns were cleaned for consistency in naming.

### **3. Summary Statistics**

* Descriptive statistics were generated using df.describe() for numerical features such as Value, LowCI, and HighCI.
* This helped summarize central tendencies, dispersion, and distribution shape, offering a quick numeric profile of the dataset.

### **4. Univariate Analysis**

* Bar plots and pie charts were used to understand the frequency and percentage distribution of mental health records by Group and Subgroup.
* This step identified which populations (e.g., by age or demographic) reported the highest and lowest levels of mental health care engagement.

### **5. Bivariate and Multivariate Analysis**

* A correlation heatmap was created to explore relationships between numerical variables such as Value, LowCI, and HighCI.
* Line plots were used to visualize how mental health indicators change across years, grouped by Indicator and Group.
* Subgroup-level and group-level comparisons revealed which demographic or regional segments required more mental health care over time.

### **6. Trend Analysis**

* Time series plots were developed to observe mental health care trends year by year.
* Smoothing techniques like a 3-period **Moving Average** were applied to distinguish long-term trends from short-term fluctuations.

### **7.Forecasting**

* Future trends were estimated using basic moving average forecasting, offering a preliminary prediction of how mental health care needs might evolve.
* The insights can assist health agencies in preparing for potential changes in resource allocation and intervention strategies.

### **8. Visualization**

* Visualization libraries like **Seaborn** and **Matplotlib** were used to create clear and insightful graphs:  
  → Line plots for trends,  
  → Boxplots for outlier detection,  
  → Bar charts for distribution,  
  → Pie charts for proportion analysis,  
  → Heatmaps for correlation exploration.
* Custom labels, legends, and color palettes were applied to ensure the visualizations were informative and easy to interpret.

**ANALYSIS RESULTS**

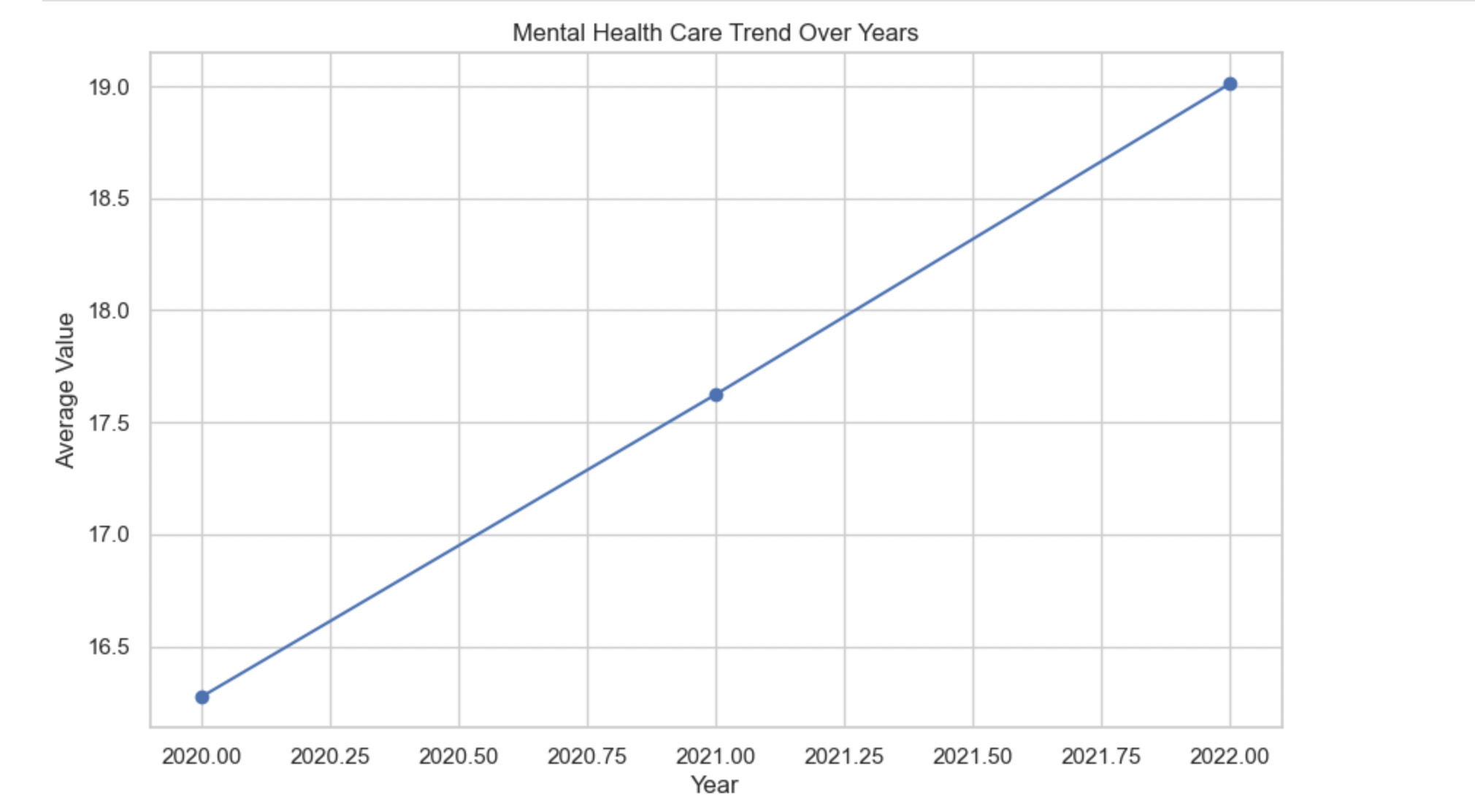
**OBJECTIVES ACHIEVED:**

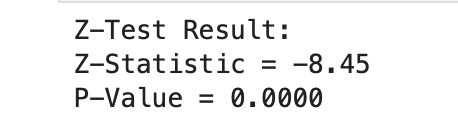
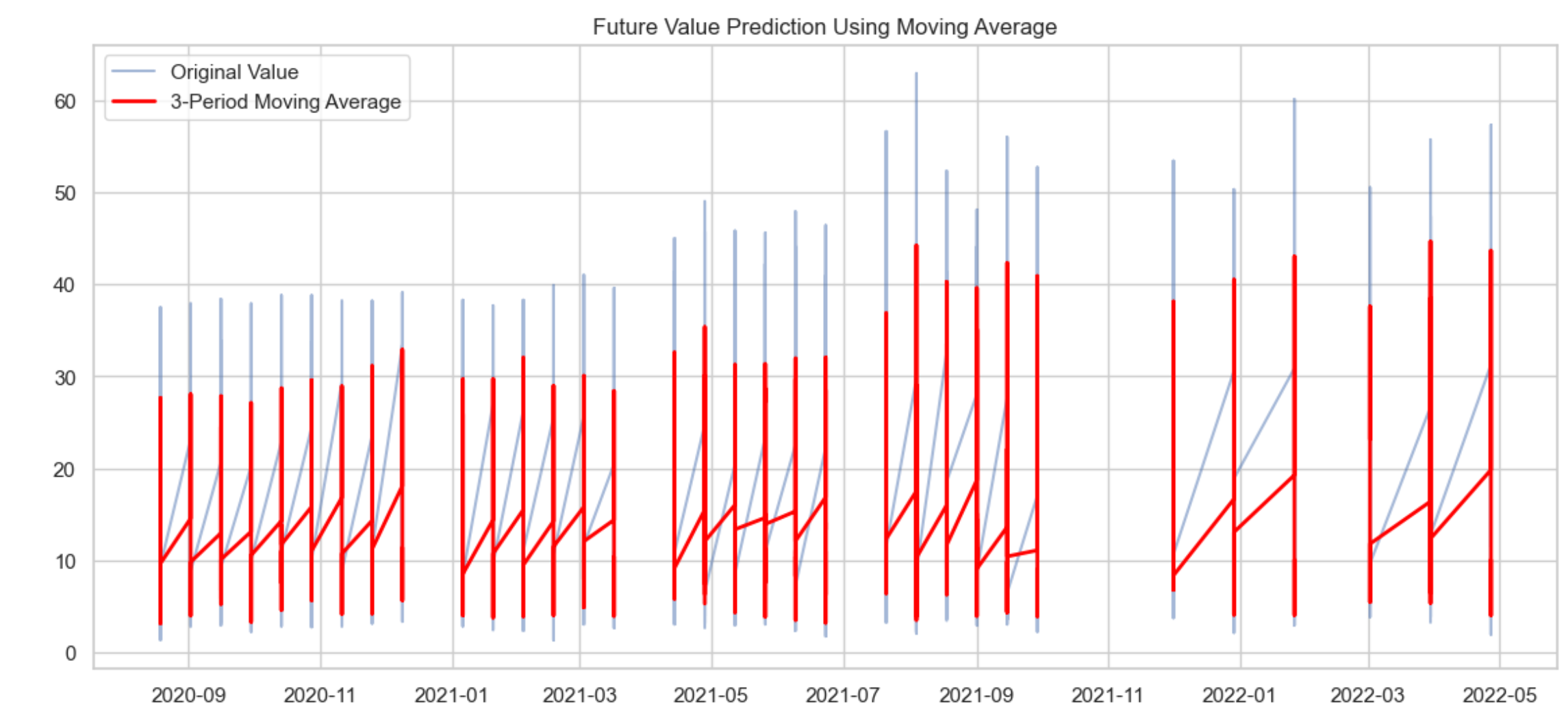
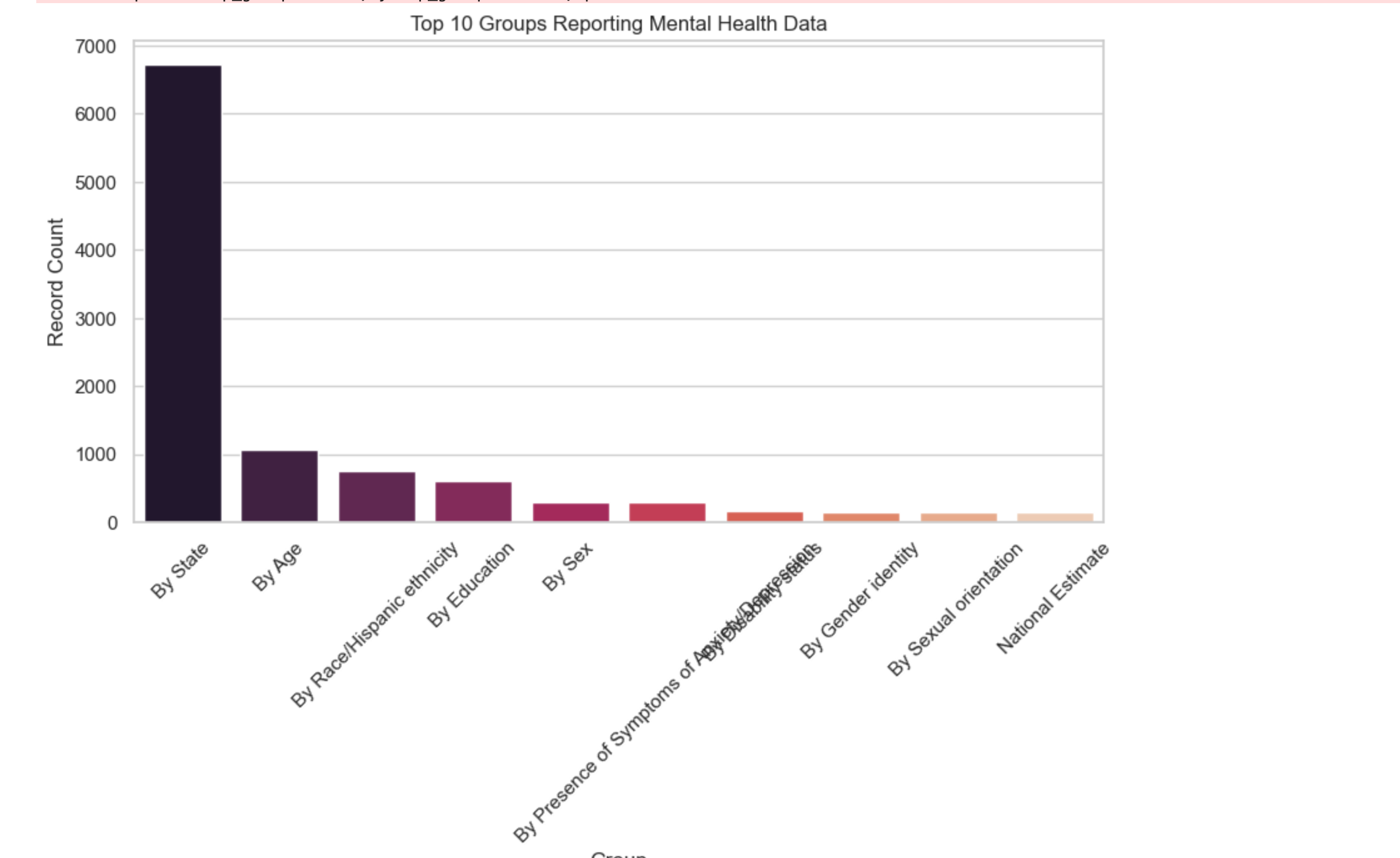
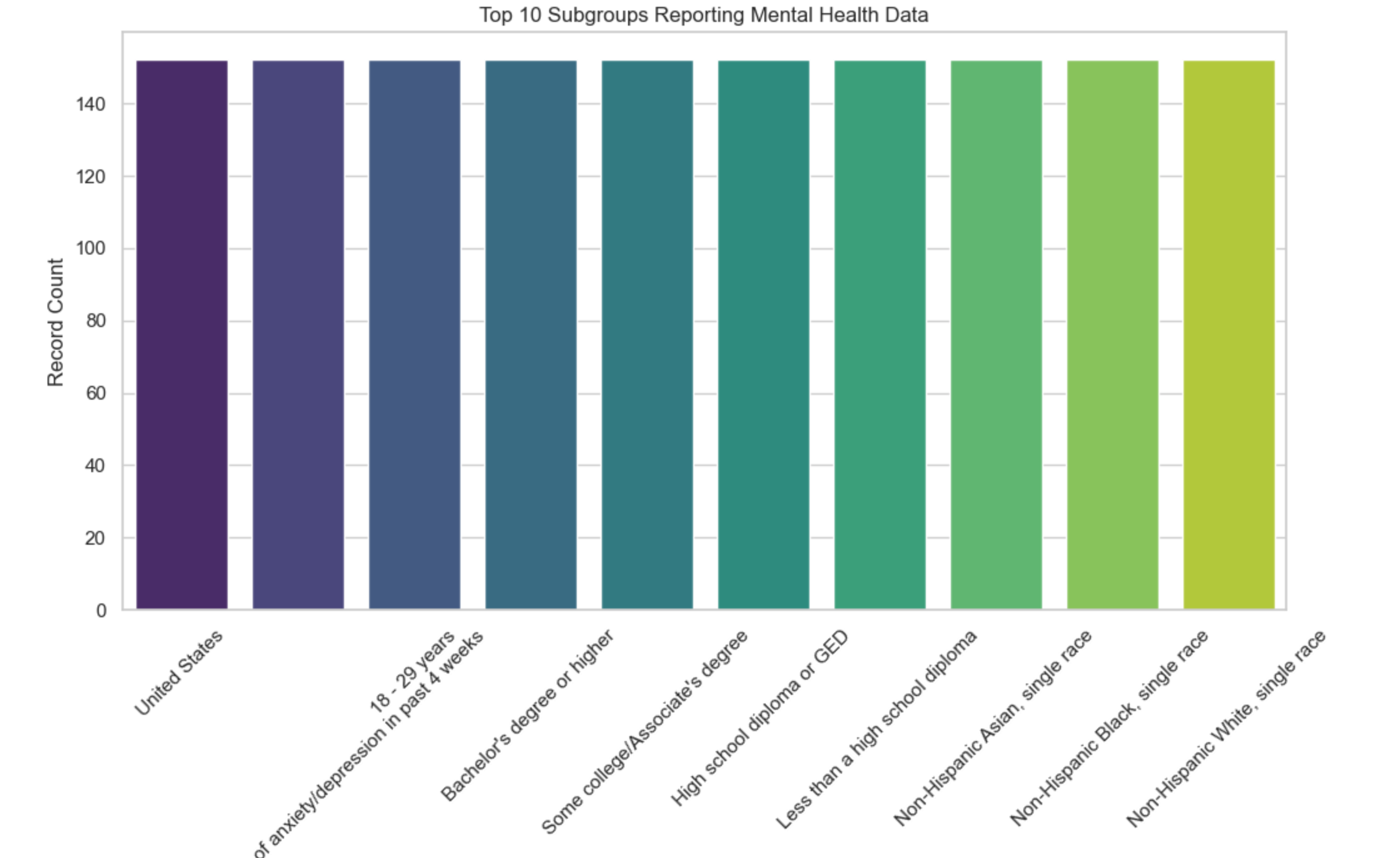
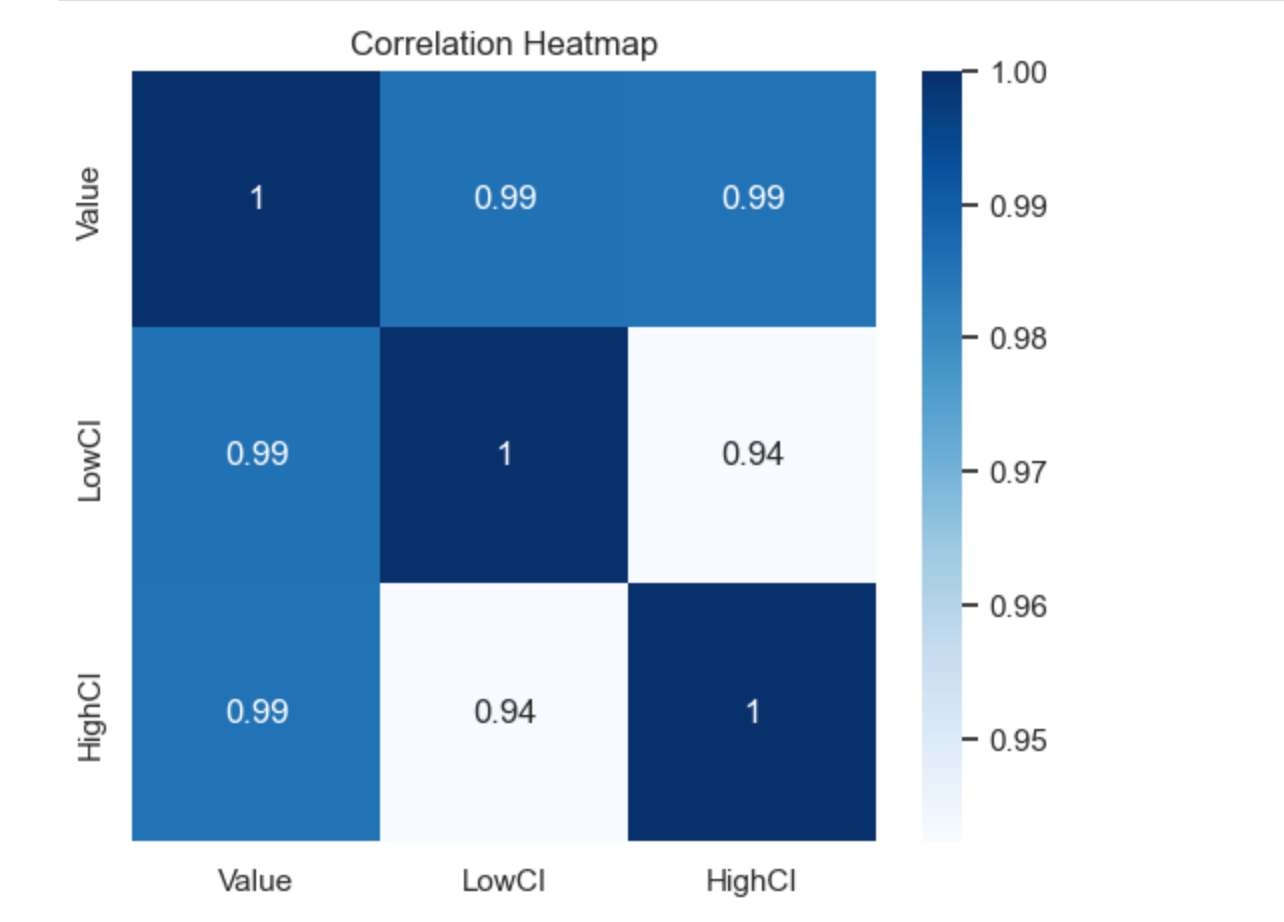
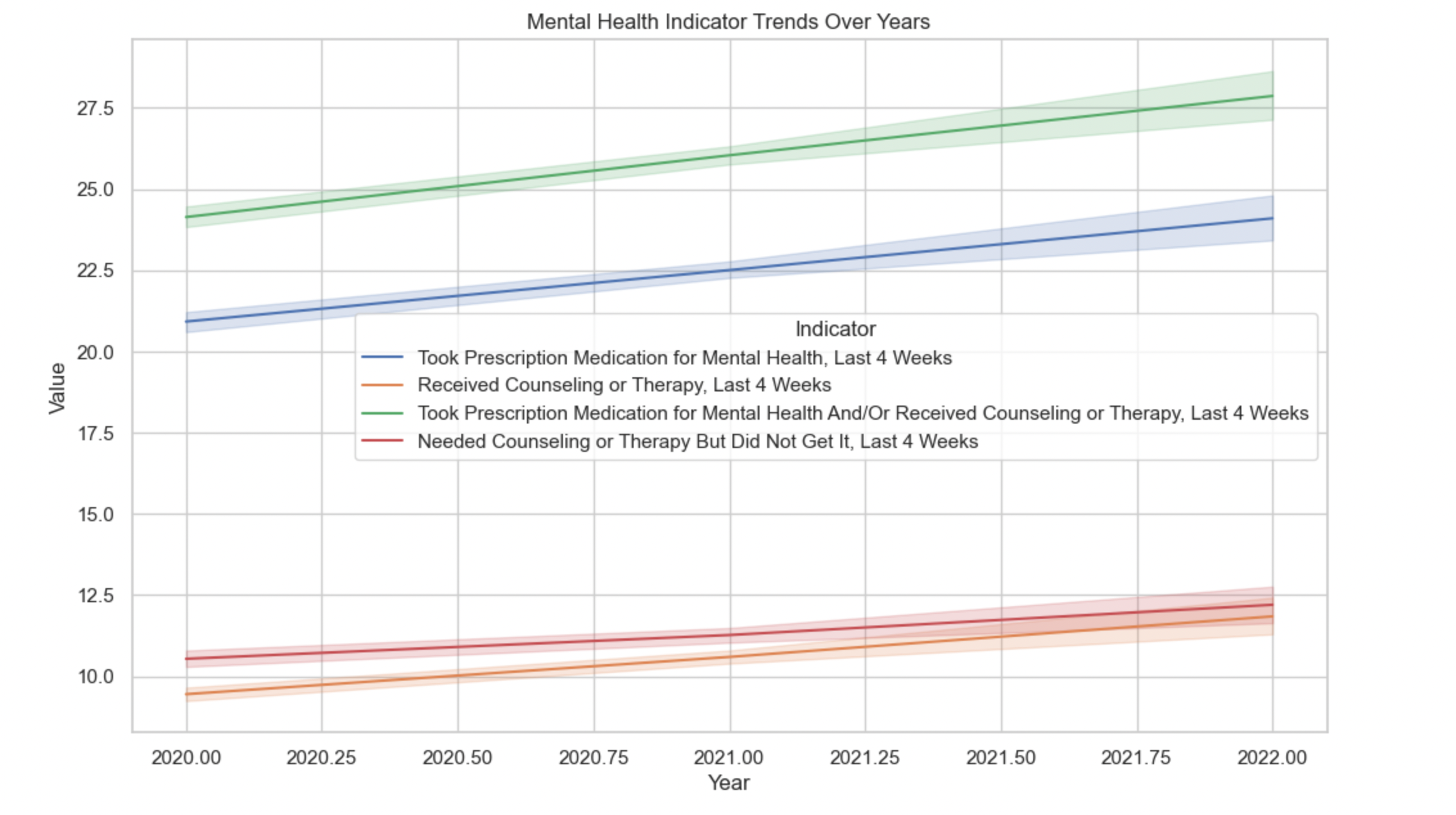
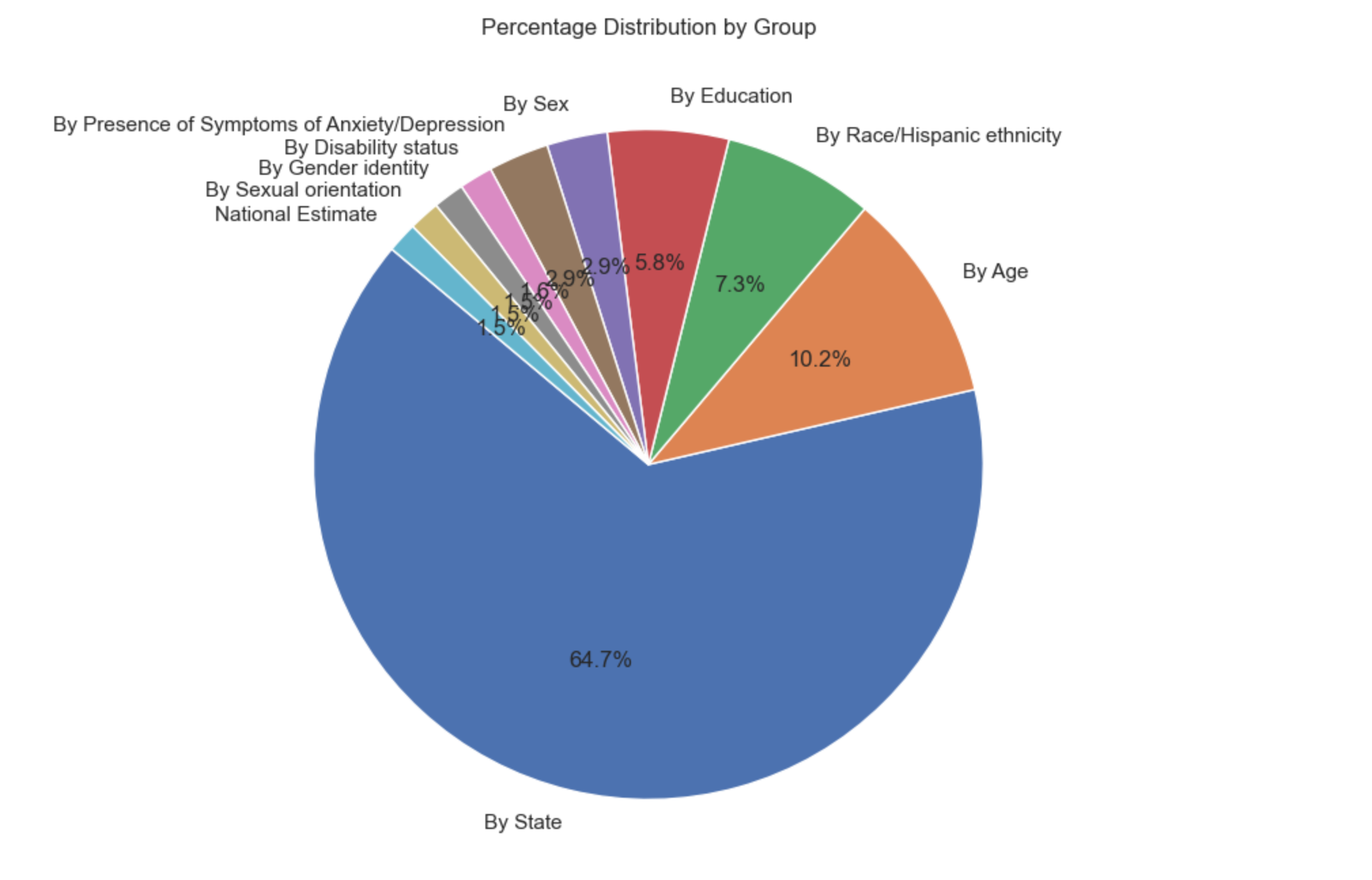
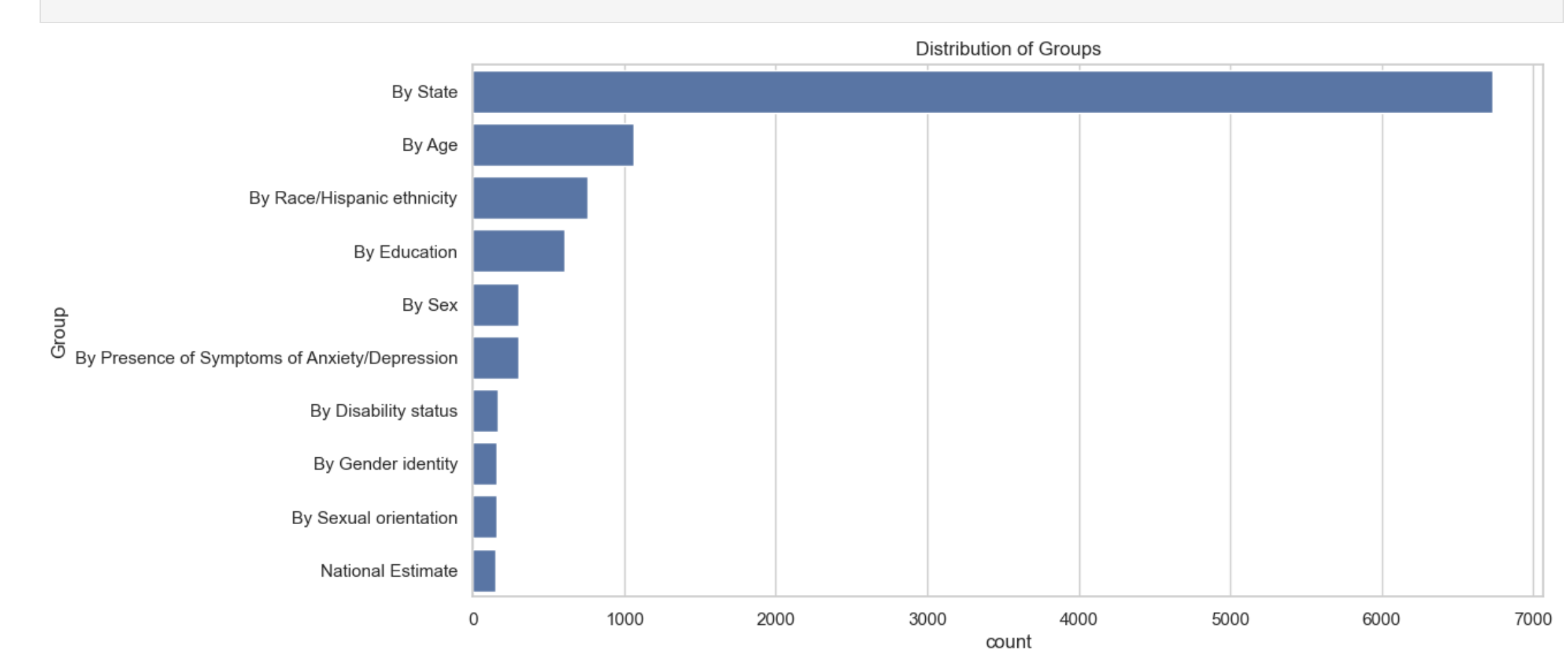
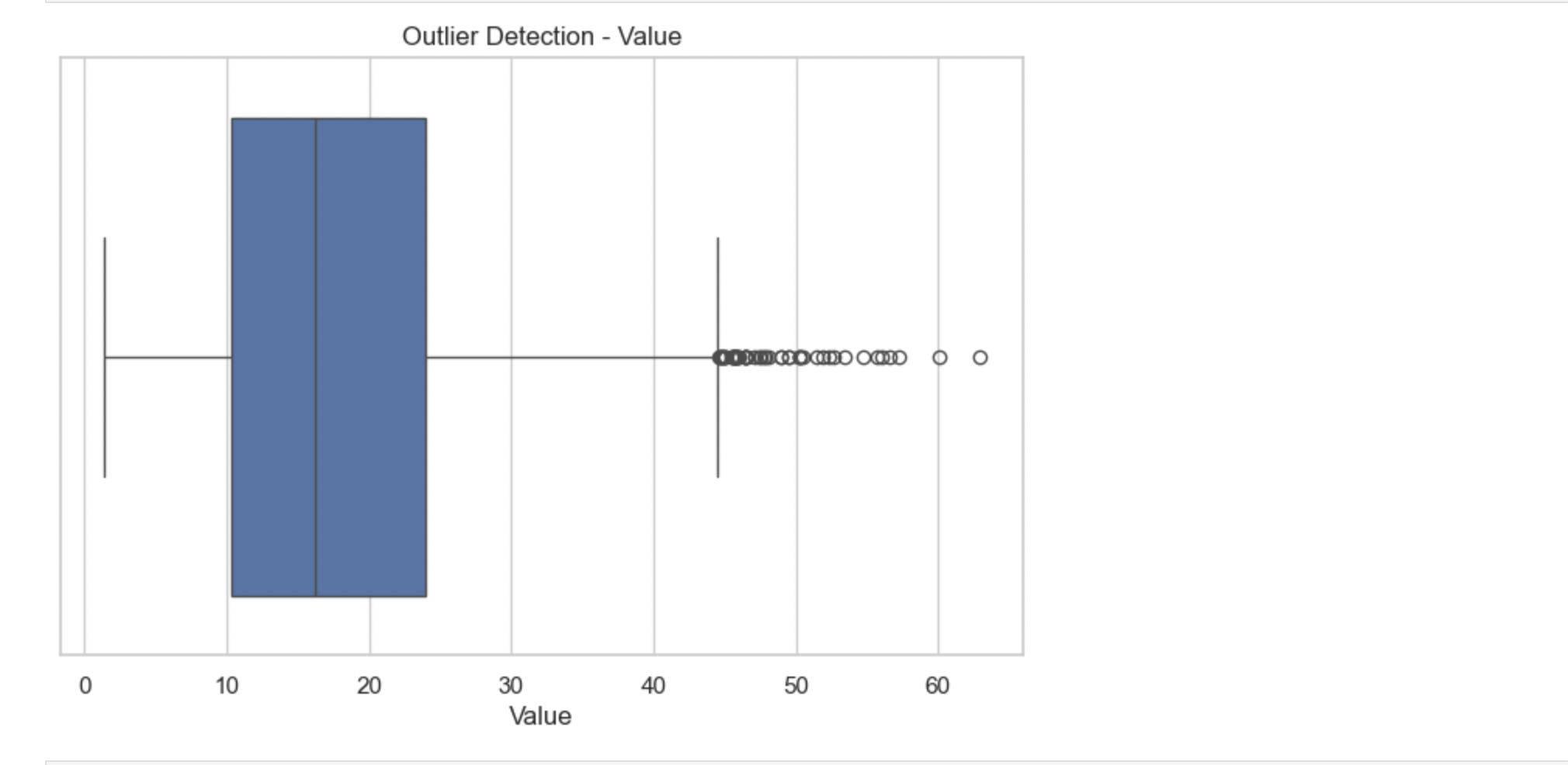
**Exploratory data analysis**

1. **Mental health trend over years**
2. **Outlier Detection**
3. **Bar plot showing distribution of mental health care**
4. **Pie Chart showing percentage distribution of Mental health**
5. **Type of mental health trend over year**
6. **Correlation Heatmap Between Numeric Columns**
7. **Top 10 Subgroups Reporting Mental Health Data**
8. **Top 10 Groups Reporting Mental Health Data**
9. **Future Value Prediction Using Moving Average**
10. **Z-Test: Value Before and After 2021**

Data Head:  
 Group \  
0 Took Prescription Medication for Mental Health... National Estimate   
1 Took Prescription Medication for Mental Health... By Age   
2 Took Prescription Medication for Mental Health... By Age   
3 Took Prescription Medication for Mental Health... By Age   
4 Took Prescription Medication for Mental Health... By Age   
  
 State Subgroup Phase Time Period Time Period Label \  
0 United States United States 2 13 Aug 19 - Aug 31, 2020   
1 United States 18 - 29 years 2 13 Aug 19 - Aug 31, 2020   
2 United States 30 - 39 years 2 13 Aug 19 - Aug 31, 2020   
3 United States 40 - 49 years 2 13 Aug 19 - Aug 31, 2020   
4 United States 50 - 59 years 2 13 Aug 19 - Aug 31, 2020   
  
 Time Period Start Date Time Period End Date Value LowCI HighCI \  
0 08/19/2020 08/31/2020 19.4 19.0 19.8   
1 08/19/2020 08/31/2020 18.7 17.2 20.3   
2 08/19/2020 08/31/2020 18.3 17.3 19.2   
3 08/19/2020 08/31/2020 20.4 19.5 21.3   
4 08/19/2020 08/31/2020 21.2 20.2 22.2   
  
 Confidence Interval Quartile Range Suppression Flag   
0 19.0 - 19.8 NaN NaN   
1 17.2 - 20.3 NaN NaN   
2 17.3 - 19.2 NaN NaN   
3 19.5 - 21.3 NaN NaN   
4 20.2 - 22.2 NaN NaN   
  
Data Info:  
  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 10404 entries, 0 to 10403  
Data columns (total 15 columns):  
 # Column Non-Null Count Dtype   
--- ------ -------------- -----   
 0 Indicator 10404 non-null object   
 1 Group 10404 non-null object   
 2 State 10404 non-null object   
 3 Subgroup 10404 non-null object   
 4 Phase 10404 non-null object   
 5 Time Period 10404 non-null int64   
 6 Time Period Label 10404 non-null object   
 7 Time Period Start Date 10404 non-null object   
 8 Time Period End Date 10404 non-null object   
 9 Value 9914 non-null float64  
 10 LowCI 9914 non-null float64  
 11 HighCI 9914 non-null float64  
 12 Confidence Interval 9914 non-null object   
 13 Quartile Range 6732 non-null object   
 14 Suppression Flag 22 non-null float64  
dtypes: float64(4), int64(1), object(10)  
memory usage: 1.2+ MB  
  
Descriptive Statistics:  
 Indicator Group \  
count 10404 10404   
unique 4 10   
top Took Prescription Medication for Mental Health... By State   
freq 2601 6732   
mean NaN NaN   
std NaN NaN   
min NaN NaN   
25% NaN NaN   
50% NaN NaN   
75% NaN NaN   
max NaN NaN   
  
 State Subgroup Phase Time Period \  
count 10404 10404 10404 10404.000000   
unique 52 80 8 NaN   
top United States United States 3.2 NaN   
freq 3672 152 1920 NaN   
mean NaN NaN NaN 28.134948   
std NaN NaN NaN 11.040210   
min NaN NaN NaN 1.000000   
25% NaN NaN NaN 20.000000   
50% NaN NaN NaN 29.000000   
75% NaN NaN NaN 37.000000   
max NaN NaN NaN 45.000000   
  
 Time Period Label Time Period Start Date Time Period End Date \  
count 10404 10404 10404   
unique 38 38 38   
top Apr 27 - May 9, 2022 04/27/2022 05/09/2022   
freq 320 320 320   
mean NaN NaN NaN   
std NaN NaN NaN   
min NaN NaN NaN   
25% NaN NaN NaN   
50% NaN NaN NaN   
75% NaN NaN NaN   
max NaN NaN NaN   
  
 Value LowCI HighCI Confidence Interval \  
count 9914.000000 9914.000000 9914.000000 9914   
unique NaN NaN NaN 7709   
top NaN NaN NaN 9.1 - 13.6   
freq NaN NaN NaN 8   
mean 17.450736 14.771565 20.475661 NaN   
std 8.270565 7.659396 9.052521 NaN   
min 1.400000 0.800000 2.000000 NaN   
25% 10.300000 8.000000 12.900000 NaN   
50% 16.200000 13.900000 19.200000 NaN   
75% 24.000000 20.800000 27.400000 NaN   
max 62.900000 53.200000 71.900000 NaN   
  
 Quartile Range Suppression Flag   
count 6732 22.0   
unique 500 NaN   
top 9.8-11.2 NaN   
freq 40 NaN   
mean NaN 1.0   
std NaN 0.0   
min NaN 1.0   
25% NaN 1.0   
50% NaN 1.0   
75% NaN 1.0   
max NaN 1.0

**VISUALIZATION**





**CONCLUSION**

The exploratory data analysis (EDA) of the mental health care dataset has successfully revealed critical insights into mental health trends across various subgroups and time periods. Through the use of visualizations like line plots, boxplots, pie charts, bar plots, and correlation heatmaps, we identified key patterns in how mental health care needs are distributed across different age groups, states, and indicators.

The trend analysis highlighted notable fluctuations in mental health care utilization over the years, and the outlier detection pointed to potential anomalies that warrant further investigation. The subgroup and group-level distributions provided a clear understanding of the populations that are more likely to report mental health concerns, which is essential for targeted policy-making and resource allocation.

The moving average forecasting offered a glimpse into potential future scenarios, which can be refined into full predictive models for proactive healthcare planning. Additionally, the statistical Z-test aimed to assess whether there was a significant shift in mental health reporting trends before and after 2010, though limited data prior to 2010 highlighted the need for a richer historical dataset for more accurate inference.

Overall, this EDA not only simplified the understanding of the dataset but also laid a strong foundation for advanced modeling, policy development, and future research on mental health care trends.

### **Future Scope of This EDA Analysis**

1. **Predictive Modeling for Mental Health Trends**  
Now that you’ve visualized trends and patterns, the next step would be to build predictive models (Linear Regression, ARIMA, LSTM) to forecast future mental health care usage, helping agencies and policymakers to prepare resources in advance.

**2. Identifying High-Risk Groups for Targeted Interventions**  
Using insights from subgroup and group analysis, you can identify which populations (age, gender, state) consistently report higher mental health concerns. These groups can be targeted for specialized public health campaigns.

**3. Seasonal & Event Impact Analysis**  
With the time-based trends, you can now study the effect of global events (pandemics, recessions, policy changes) on mental health care demand, helping design better response strategies for future crises.

**4. Correlation Study with External Datasets**  
You can extend this analysis by merging your mental health dataset with:

* Crime rates
* Employment statistics
* Social media sentiment
* Economic indicators  
  and check for hidden relationships using correlation, regression, or machine learning techniques.

5. **Anomaly Detection and Early Warning Systems**  
Using the outlier detection you’ve already started, you could automate anomaly detection to flag unexpected spikes or drops in mental health care usage, which could indicate social issues or reporting errors.

6. **Policy and Resource Optimization**  
Government and health organizations can use this analysis to allocate budgets, prioritize mental health programs, or optimize medical resources based on past trends and predictions.

**References**

Kaggle: <https://www.kaggle.com/code/imoore/intro-to-exploratory-data-analysis-eda-in-python>

Dataset: <https://catalog.data.gov/dataset/mental-health-care-in-the-last-4-weeks>