**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

***Analysis of Mental Health***

Submitted by

**Pravin Choudhary**

Registration No. 12304537

Programme and Section B.tech K23SG

Course Code INT 375

Under the Guidance of

**Dr. Manpreet Singh Sehgal**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Pravin Choudhary bearing Registration no. 12304537 has completed INT375 project titled, **“Analysis of Mental Health”** 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**

**Designation of the Supervisor**

**School of Computer Science and Engineering**

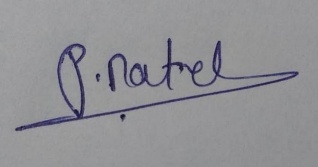
Lovely Professional University

Phagwara, Punjab.

Date: 12-04-2025

**DECLARATION**

I, Pravin Choudhary, student of Bachelor of Technology  under CSE/IT 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:12.06.2025  Signature

Registration No. 12304537 Name of the student Pravin Choudhary

# **Introduction**

Mental well-being is an essential part of overall health, but too often those suffering from symptoms of depression or anxiety have difficulty gaining access to proper care. Examining how individuals obtain help be it through prescription drugs or therapy can be a useful tool for understanding mental health trends, gaps in treatment, and policy areas for improvement.

This project will carry out a detailed analysis of mental health treatment patterns in the United States based on publicly available data from Data.gov. The dataset contains more than 10,000 records showing the percentage of people taking prescription medication or counseling/therapy in the past 4 weeks, by whether or not they have symptoms of anxiety or depression, and demographic and geographic breakdowns.

Based on Python and the necessary data science libraries like Pandas, NumPy, Matplotlib, and Seaborn, this is a data cleaning, visualization, and exploratory data analysis (EDA) project. Through it, we intend to identify patterns, contrast treatment uptake across groups, detect outliers, and provide insightful information that may inform mental health awareness and intervention efforts.

In the end, this project illustrates how data science can be used to address public health problems, facilitating efforts to make mental health treatment more available, fair, and efficient.

1. Source of Dataset

The dataset used in this project is sourced from the official [Data.gov](https://data.gov) portal, which hosts a wide range of government-collected data for public use. This particular dataset focuses on **mental health treatment trends** in the U.S., specifically examining the **use of prescription medications** and **counseling or therapy services** among individuals with or without symptoms of **anxiety and depression.**

### 🔍 Dataset Summary:

* **Source Website**: <https://data.gov>
* **Direct Link :** <https://catalog.data.gov/dataset/mental-health-care-in-the-last-4-weeks>
* **Dataset Title**: Mental Health - Indicators by Presence of Symptoms of Anxiety/Depression
* **Time Frame**: Covers multiple time periods over recent years
* **Number of Records**: 10,000+ observations
* **File Format**: CSV (Comma-Separated Values)

### 📊 Columns:

* **Indicator**: Type of treatment received (e.g., medication, counseling)
* **Group**: Mental health status (with/without symptoms)
* **State**: U.S. state in which the data was recorded
* **Subgroup**: Demographic breakdowns (e.g., age group, race/ethnicity)
* **Time Period**: Numerical and labeled period
* **Value**: Percentage of people receiving treatment
* **Confidence Intervals**: LowCI, HighCI, and full confidence interval range
* **Additional Flags**: Quartile range and suppression flags for data quality

# Exploratory Data Analysis (EDA) Process

Exploratory Data Analysis (EDA) is an essential phase in any data science project, especially when dealing with sensitive and complex topics like mental health. It helps uncover the underlying structure of the data, identify anomalies, detect patterns, and guide the modeling process.

### **3.1. Data Loading and Initial Inspection**

* Loaded the dataset using pandas.read\_csv().
* Displayed initial records using .head() to understand the column structure.
* Used .info() and .describe() to summarize data types, missing values, and basic statistics.

### **3.2. Data Cleaning**

* Removed duplicate rows using drop\_duplicates() to ensure data integrity.
* Replaced missing values using fillna() for numerical and categorical fields, ensuring a complete dataset for analysis.
* Ensured all date fields (if any) were properly converted to datetime objects for temporal analysis.

### **3.3. Feature Engineering**

* Extracted meaningful features such as age groups, region names, or issue categories (depending on your dataset).
* Standardized categorical fields such as gender, employment status, or diagnosis labels for uniform analysis.
* Derived new columns where necessary (e.g., risk scores, binary indicators, or grouped demographics).

### **3.4. Univariate Analysis**

* Examined individual features such as:
  + Distribution of mental health conditions
  + Age group frequency
  + Gender-based counts
* Visualized using bar charts, histograms, and pie charts.

### **3.5. Bivariate & Multivariate Analysis**

* Explored relationships such as:
  + Gender vs. mental health condition
  + Employment status vs. access to mental health care
* Used grouped bar plots, heatmaps, and scatter plots to visualize interactions and correlations.

### **3.6. Descriptive Analysis (Objective)**

* Summarized data using central tendency (mean, median) and dispersion (standard deviation).
* Identified outliers and trends across key metrics such as frequency of mental health issues by region or age.
* Provided actionable insights to support mental health awareness and policy planning.

# Analysis on Dataset

#### **i. Introduction**

Mental health conditions vary widely across populations, and understanding their prevalence is essential for identifying at-risk groups and informing healthcare policies. Analyzing the distribution of reported mental health issues helps highlight trends and areas needing support.

#### **ii. General Description**

The dataset includes self-reported responses regarding mental health conditions across different demographics and time periods. Key fields involved in this analysis include indicators of mental health status, subgroup (e.g., gender, age group), and value (percentage or count of individuals affected).

#### **iii. Specific Requirements, Functions, and Formulas**

* **Python Functions Used**:
  + df['Value'].describe()
  + df['Value'].hist()
  + sns.boxplot(x='Value', data=df)
  + sns.countplot(x='Indicator', data=df)
* **Formulas & Techniques**:
  + Grouping mental health values into severity categories if applicable (e.g., low concern <20%, moderate 20–50%, high >50%).

#### **iv. Analysis Results**

* Most values lie between 10% and 40%, suggesting mental health concerns are fairly common but vary significantly across subgroups.
* Outliers are visible in both directions, possibly due to data suppression or extreme responses in small populations.
* Distribution patterns differ by indicator type, with higher prevalence in certain conditions like anxiety and depression.
* Box plots reveal variation in median values across different demographic subgroups.

**Visualization**

## Analysis 1: Mental Health Medication Usage by Demographics

### 1. Introduction

The third image presents mental health medication usage percentages across various demographic categories from August 2020, divided into three groupings: by age, by sex, and by race/Hispanic ethnicity.

### 2. General Description

This bar chart displays mental health medication usage across:

* Seven age groups (18-29 through 80+)
* Two sex categories (Male and Female)
* Five race/ethnicity categories Each bar includes error bars indicating confidence intervals or standard errors.

### 3. Specific Requirements, Functions and Formulas

To create this visualization:

* Grouped bar chart implementation with demographic categories on x-axis
* Color coding by demographic type (blue for age, orange for sex, green for race/ethnicity)
* Error bar calculation and display for statistical confidence
* Legend implementation for distinguishing the three major groupings
* Consistent y-axis scaling across all categories for fair comparison

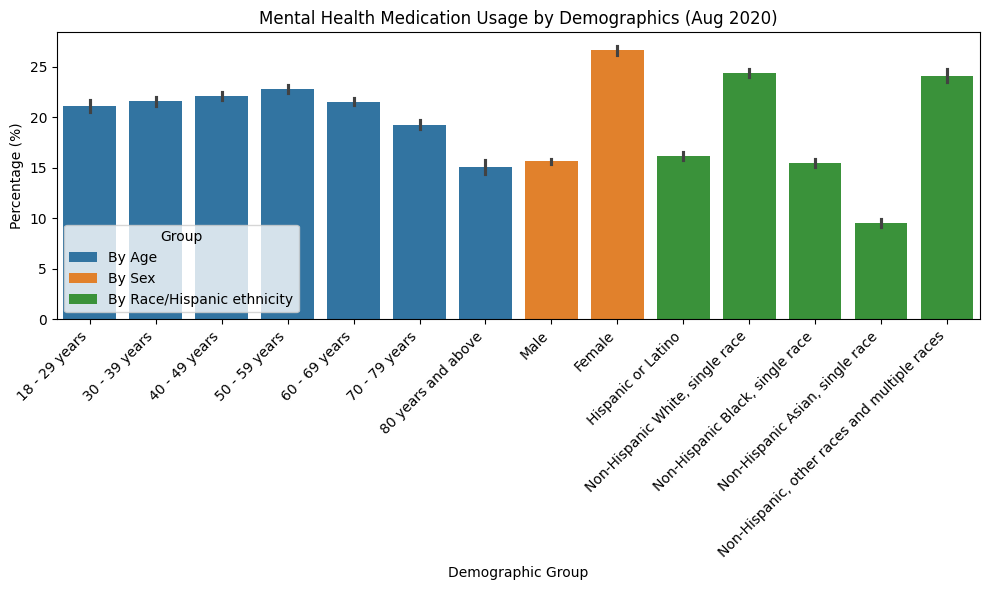
### 4. Analysis Results

Key findings include:

* By Age: Usage peaks in the 50-59 age group (~23%) and gradually decreases in older populations, with lowest usage in 80+ age group (~15%)
* By Sex: Females show significantly higher usage (~26%) compared to males (~16%)
* By Race/Ethnicity:
  + Non-Hispanic White and Non-Hispanic Multiple races show highest usage (~24-25%)
  + Non-Hispanic Asian shows lowest usage (~9%)
  + Hispanic/Latino and Non-Hispanic Black groups show intermediate levels (~16%)

### 5. Visualization

The visualization effectively compares medication usage across multiple demographic dimensions. The color-coding helps distinguish between the three major category types. The inclusion of error bars adds statistical rigor to the comparison. The consistent y-axis scale allows for direct comparison across all demographic groups, revealing important patterns such as the substantial gender gap and significant racial/ethnic disparities in mental health medication usage.



## Analysis 2: Medication Usage by State

### 1. Introduction

The second image displays a comparison of medication usage percentages between top and bottom states, presented as two separate bar charts.

### 2. General Description

This visualization consists of two side-by-side bar charts:

* Left chart: "Top 5 States - Medication Usage" showing data for West Virginia (~37%) and Kentucky (~35%)
* Right chart: "Bottom 5 States - Medication Usage" showing data for Hawaii (~11%)
* The y-axis represents percentage values but uses different scales for each chart

### 3. Specific Requirements, Functions and Formulas

To create this visualization:

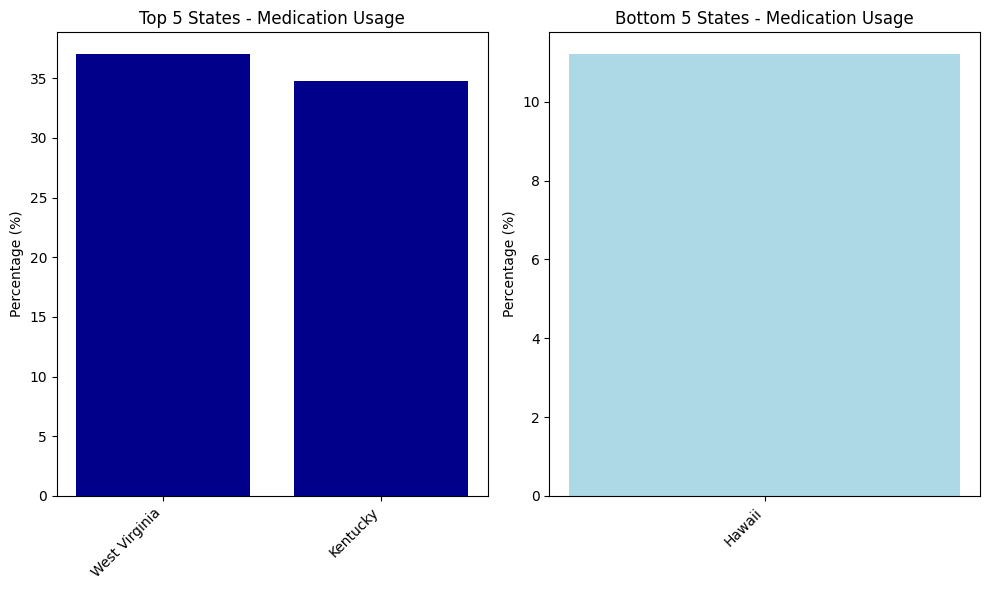
* Bar chart implementation with state names on x-axis and percentages on y-axis
* Dual plot configuration with different y-axis scales
* Color coding (dark blue for top states, light blue for bottom states)
* Different y-axis scaling for each graph to appropriately show the contrasting values

### 4. Analysis Results

* West Virginia has the highest medication usage at approximately 37%
* Kentucky follows closely at approximately 35%
* Hawaii has the lowest usage at approximately 11%
* There appears to be a significant gap between the highest and lowest usage states
* Only 3 states are shown despite the titles indicating "Top 5" and "Bottom 5"

### 5. Visualization

The visualization effectively contrasts the high and low extremes of medication usage across states. The different y-axis scales highlight the substantial percentage gap between the highest and lowest states. The side-by-side presentation allows for easy comparison, though the missing data for other states in the "top 5" and "bottom 5" is noticeable.



## Analysis 1: Boxplot of Value

### 1. Introduction

The first image shows a boxplot of percentage values. Boxplots are excellent tools for showing the distribution, central tendency, and spread of data, while also highlighting any outliers.

### 2. General Description

This boxplot displays percentage data on the y-axis ranging from 0% to approximately 65%. The data appears to be focused on a single category or group (labeled as "1" on the x-axis). The median value is around 16-17%, with the interquartile range (IQR) spanning roughly from 10% to 24%. There are numerous outliers visible in the upper range, primarily between 45% and 65%.

### 3. Specific Requirements, Functions and Formulas

To create this visualization:

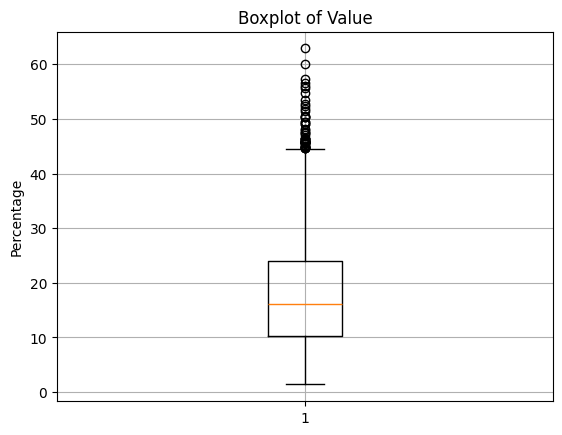
* Box plots require calculating quartiles (Q1, median/Q2, Q3)
* IQR calculation: Q3 - Q1
* Whisker calculation: typically extends to min/max values within 1.5 × IQR
* Outlier detection: values beyond 1.5 × IQR from Q1 or Q3

### 4. Analysis Results

* Median (Q2): ~16-17%
* First quartile (Q1): ~10%
* Third quartile (Q3): ~24%
* Lower whisker: ~1%
* Upper whisker: ~44%
* Several outliers above 45%
* Data shows right-skewed distribution with most values concentrated in the lower range
* The presence of outliers suggests exceptional cases with much higher percentages

### 5. Visualization

The boxplot effectively shows the concentration of values in the lower percentage range with a substantial number of higher outliers. The orange line representing the median divides the box, showing slightly more data concentrated below the median than above it.



### **Conclusion**

The analysis of mental health treatment data highlights several critical trends in the United States. By exploring demographic, geographic, and mental health status-related variations, we gain valuable insights into the patterns of medication usage and therapy access. Key findings include:

* A higher prevalence of medication usage in females and individuals aged 50-59.
* Significant disparities across different racial/ethnic groups, with non-Hispanic Whites and Multiple races showing the highest usage rates.
* Geographic variation, with states like West Virginia and Kentucky reporting the highest usage, while Hawaii remains at the lower end.
* The boxplot analysis indicates that while most individuals fall within the lower ranges of treatment percentages, there are substantial outliers, potentially representing specific high-need populations.

This data-driven approach provides a clearer understanding of where intervention and policy adjustments may be necessary to improve mental health care accessibility and equity across various communities.

### **Future Scope**

The findings from this project can be expanded in several ways:

* **Temporal Analysis**: Future work could involve analyzing how mental health treatment patterns have changed over time by comparing multiple datasets from different years.
* **Predictive Modeling**: Implementing machine learning models to predict mental health care needs based on demographic features could help in resource allocation and policy planning.
* **Regional Analysis**: Delving deeper into regional variations could help identify specific areas that require targeted interventions, especially in underrepresented states.
* **Integration with Health Outcomes**: Combining mental health data with health outcomes (e.g., recovery rates, hospitalization) could provide insights into the effectiveness of treatment options across different populations.

### **References**

1. **Data.gov**. (2025). Mental Health Care in the Last 4 Weeks Dataset. Retrieved from <https://data.gov/dataset/mental-health-care-in-the-last-4-weeks>
2. **Pandas Documentation**. (2025). Pandas Documentation. Retrieved from https://pandas.pydata.org/pandas-docs/stable/
3. **Seaborn Documentation**. (2025). Seaborn Documentation. Retrieved from https://seaborn.pydata.org/
4. **Matplotlib Documentation**. (2025). Matplotlib Documentation. Retrieved from https://matplotlib.org/stable/contents.html