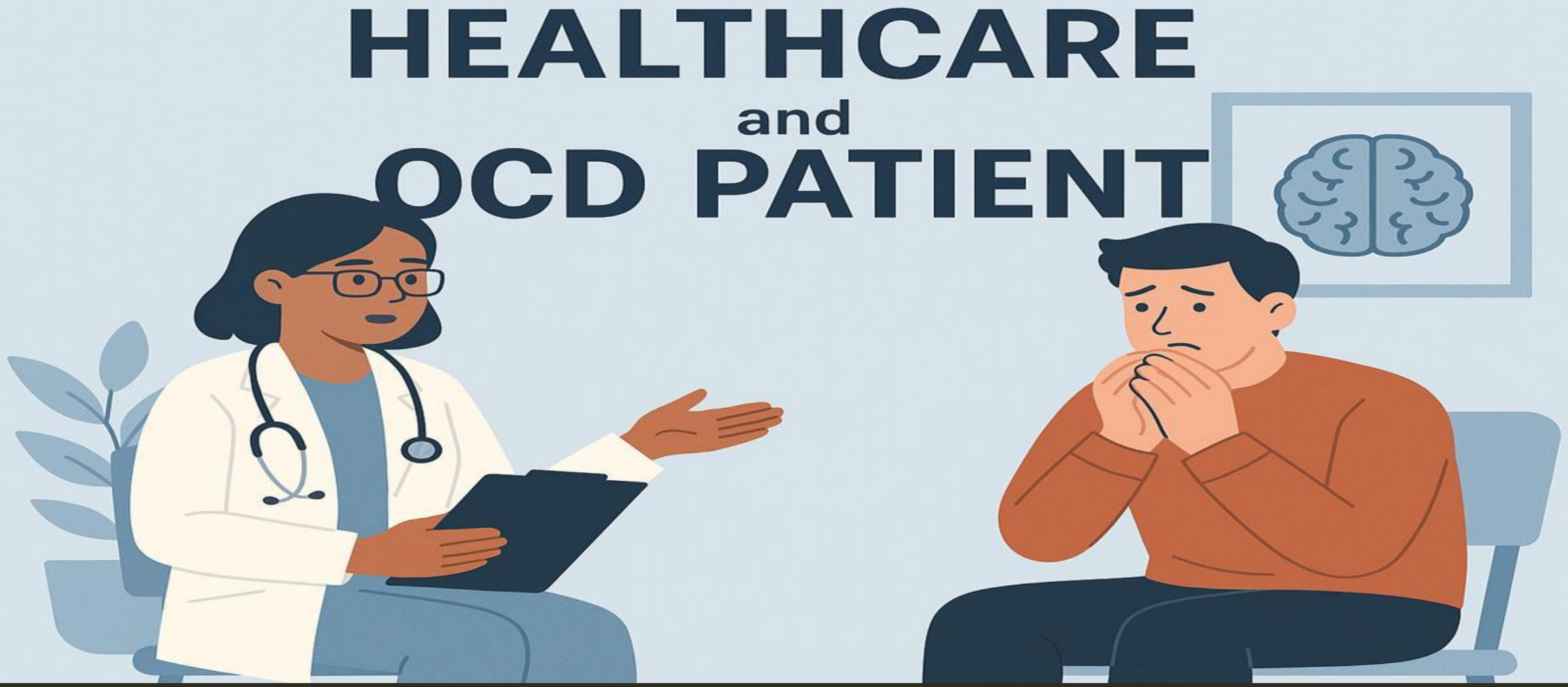


OCD PATIENT DATASET: DEMOGRAPHICS & CLINICAL DATA

DEEPAK SEN
INTERN – HEALTHCARE ANALYST | UNIFIED MENTOR

INTRODUCTION

Obsessive-Compulsive Disorder (OCD) is a mental health condition that affects how people think and behave, often leading to distressing thoughts (obsessions) and repetitive actions (compulsions). The impact of OCD goes beyond mental health—it influences social life, relationships, and daily routines.

In the evolving landscape of healthcare, data plays a vital role in understanding patient profiles and treatment outcomes. This analysis explores a dataset of OCD patients, focusing on their demographics and clinical history.

By analyzing this data, we aim to uncover hidden patterns, detect trends, and provide insights that could support better decision-making in mental health care. This study reflects how data analytics can help healthcare professionals deliver more personalized and effective interventions.

OBJECTIVE

The primary goal of this analysis is to explore and understand the **demographic and clinical characteristics** of patients diagnosed with Obsessive-Compulsive Disorder (OCD). Using data-driven techniques, the objective is to identify meaningful patterns that can support improved mental health research and care planning.

Specific Objectives:

- To analyze the distribution of patients based on **gender, age, and marital status**
- To examine **clinical factors** such as **diagnosis type**, **duration of illness**, and **treatment outcomes**
- To detect any **hidden relationships** between demographic and clinical variables
- To identify potential **trends or outliers** that could guide further investigation
- To support **evidence-based insights** for healthcare professionals using visualizations and statistical tools

DATASET OVERVIEW

The dataset contains detailed information about patients diagnosed with **Obsessive-Compulsive Disorder (OCD)**, capturing both **demographic** and **clinical** variables. After initial data cleaning (removing duplicates and handling missing values), the dataset is structured and ready for analysis.

Key Details:

- **Total Records (Rows): [1393]**
- **Total Features (Columns): [17]**
- **Patient-Level Data: Each row represents one unique patient**
- **Data Type: Structured CSV (Comma-Separated Values)**

Main Features in the Dataset:

- Patient ID - Unique identifier for each patient
- Gender - Categorical variable (Male/Female)
- Age - Numeric value (in years)
- Ethnicity – ethnic background of the patient
- Marital Status – marital background (single, married, divorced, etc.)
- Education Level – highest education level achieved
- OCD Diagnosis Date – date of diagnosis
- Duration of Symptoms (months) – length of illness
- Previous Diagnoses – other mental health conditions
- Family History of OCD – yes or no

- Obsession Type – type of obsession (e.g., symmetry, contamination)
- Compulsion Type – type of compulsion (e.g., checking, washing)
- Y-BOCS Score(obsessions) - score measuring obsessive severity
- Y-BOCS Score(compulsions) - core measuring compulsive severity
- Depression Diagnosis – presence of depression (yes/no)
- Anxiety Diagnosis – presence of anxiety (yes/no)
- Medications – type of psychiatric medication used

IMPORTING NECESSARY LIBRARIES AND LOADING DATASET

```
# importing Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import random
```

```
# Loadind the Dataset into dataframe
data = pd.read_csv('/content/OCD Patient Dataset_ Demographics & Clinical Data.csv')
```

```
data.head()
```

	Patient ID	Age	Gender	Ethnicity	Marital Status	Education Level	OCD Diagnosis Date	Duration of Symptoms (months)	Previous Diagnoses	Family History of OCD	Obsession Type	Compulsion Type	Y-BOCS Score (Obsessions)	Y-BOCS Score (Compulsions)	Y-BOCS Score (Compulsions)	Depression Diagnosis	Anxiety Diagnosis	Medications
0	1018	32	Female	African	Single	Some College	15-07-2016	203	MDD	No	Harm-related	Checking	17	10	10	Yes	Yes	SNRI
1	2406	69	Male	African	Divorced	Some College	28-04-2017	180	NaN	Yes	Harm-related	Washing	21	25	25	Yes	Yes	SSRI
2	1188	57	Male	Hispanic	Divorced	College Degree	02-02-2018	173	MDD	No	Contamination	Checking	3	4	4	No	No	Benzodiazepine
3	6200	27	Female	Hispanic	Married	College Degree	25-08-2014	126	PTSD	Yes	Symmetry	Washing	14	28	28	Yes	Yes	SSRI
4	5824	56	Female	Hispanic	Married	High School	20-02-2022	168	PTSD	Yes	Hoarding	Ordering	39	18	18	No	No	NaN

CONCISE SUMMARY OF THE DATASET

```
data.shape
```

```
(1393, 17)
```

```
data.columns
```

```
Index(['Patient ID', 'Age', 'Gender', 'Ethnicity', 'Marital Status',
       'Education Level', 'OCD Diagnosis Date',
       'Duration of Symptoms (months)', 'Previous Diagnoses',
       'Family History of OCD', 'Obsession Type', 'Compulsion Type',
       'Y-BOCS Score (Obsessions)', 'Y-BOCS Score (Compulsions)',
       'Depression Diagnosis', 'Anxiety Diagnosis', 'Medications'],
      dtype='object')
```

```
data.index
```

```
RangeIndex(start=0, stop=1393, step=1)
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1393 entries, 0 to 1392
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Patient ID                      1393 non-null   int64
 1   Age                             1393 non-null   int64
 2   Gender                          1393 non-null   object
 3   Ethnicity                       1393 non-null   object
 4   Marital Status                  1393 non-null   object
 5   Education Level                 1393 non-null   object
 6   OCD Diagnosis Date              1393 non-null   object
 7   Duration of Symptoms (months)   1393 non-null   int64
 8   Previous Diagnoses              1169 non-null   object
 9   Family History of OCD           1393 non-null   object
 10  Obsession Type                  1393 non-null   object
 11  Compulsion Type                 1393 non-null   object
 12  Y-BOCS Score (Obsessions)       1393 non-null   int64
 13  Y-BOCS Score (Compulsions)      1393 non-null   int64
 14  Depression Diagnosis            1393 non-null   object
 15  Anxiety Diagnosis               1393 non-null   object
 16  Medications                     1037 non-null   object
dtypes: int64(5), object(12)
memory usage: 185.1+ KB
```

```
 # This syntex gives data of entire row if in row has duplicate value
data[data.duplicated()].sum()
```

	0
Patient ID	0
Age	0
Gender	0
Ethnicity	0
Marital Status	0
Education Level	0
OCD Diagnosis Date	0
Duration of Symptoms (months)	0
Previous Diagnoses	0
Family History of OCD	0
Obsession Type	0
Compulsion Type	0
Y-BOCS Score (Obsessions)	0
Y-BOCS Score (Compulsions)	0
Depression Diagnosis	0
Anxiety Diagnosis	0
Medications	0

dtype: object

```
data.isnull().sum() # Checking Null Values in Dataset
```

	0
Patient ID	0
Age	0
Gender	0
Ethnicity	0
Marital Status	0
Education Level	0
OCD Diagnosis Date	0
Duration of Symptoms (months)	0
Previous Diagnoses	224
Family History of OCD	0
Obsession Type	0
Compulsion Type	0
Y-BOCS Score (Obsessions)	0
Y-BOCS Score (Compulsions)	0
Depression Diagnosis	0
Anxiety Diagnosis	0
Medications	356

dtype: int64

```
data.describe()
```

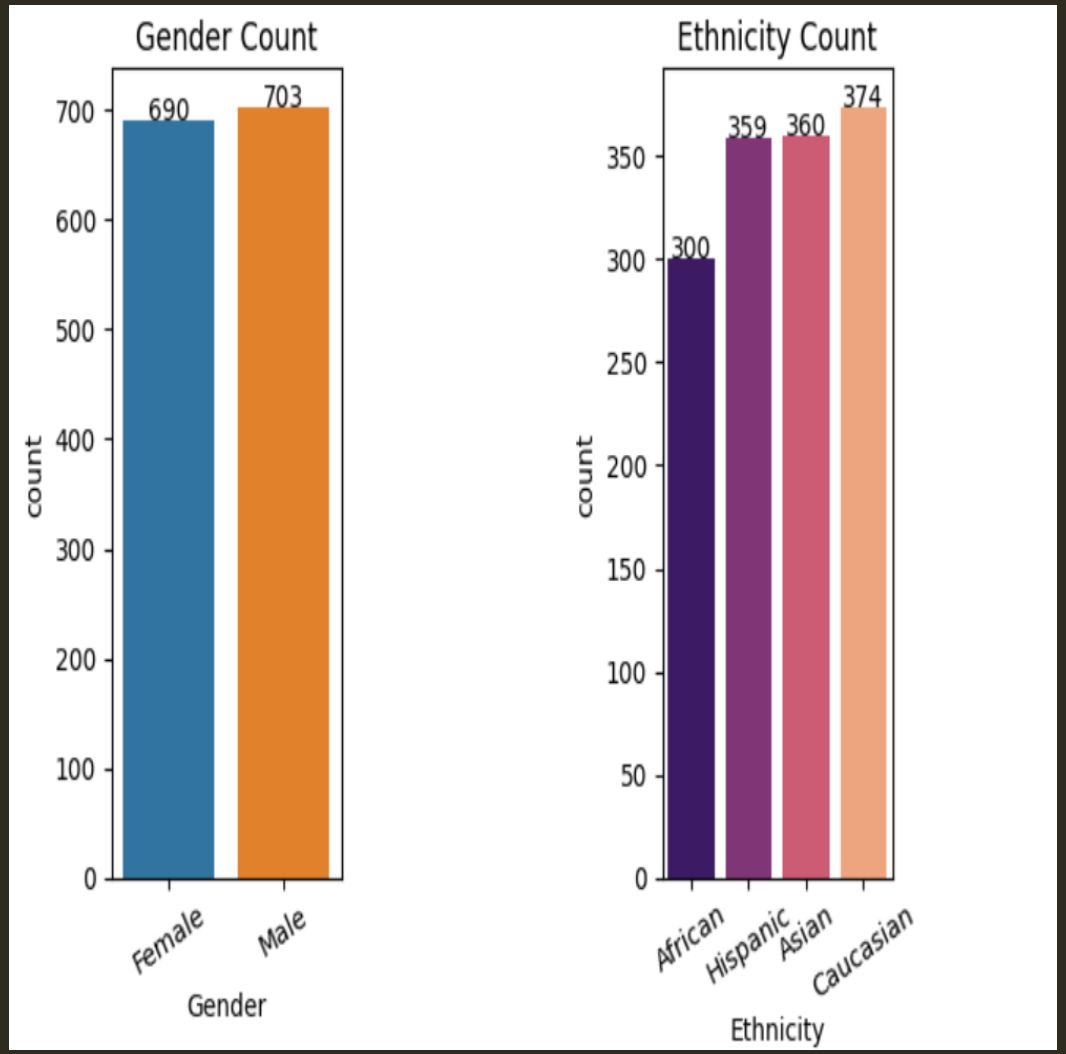
cell output actions

	Patient ID	Age	Duration of Symptoms (months)	Y-BOCS Score (Obsessions)	Y-BOCS Score (Compulsions)
	00000	1393.000000	1393.000000	1393.000000	1393.000000
mean	5528.895908	46.986360	122.244078	20.145729	19.720029
std	2574.800758	16.832951	66.894308	11.880131	11.759589
min	1017.000000	18.000000	6.000000	0.000000	0.000000
25%	3324.000000	32.000000	65.000000	10.000000	9.000000
50%	5476.000000	48.000000	122.000000	20.000000	20.000000
75%	7764.000000	61.000000	178.000000	31.000000	30.000000
max	9995.000000	75.000000	240.000000	40.000000	40.000000

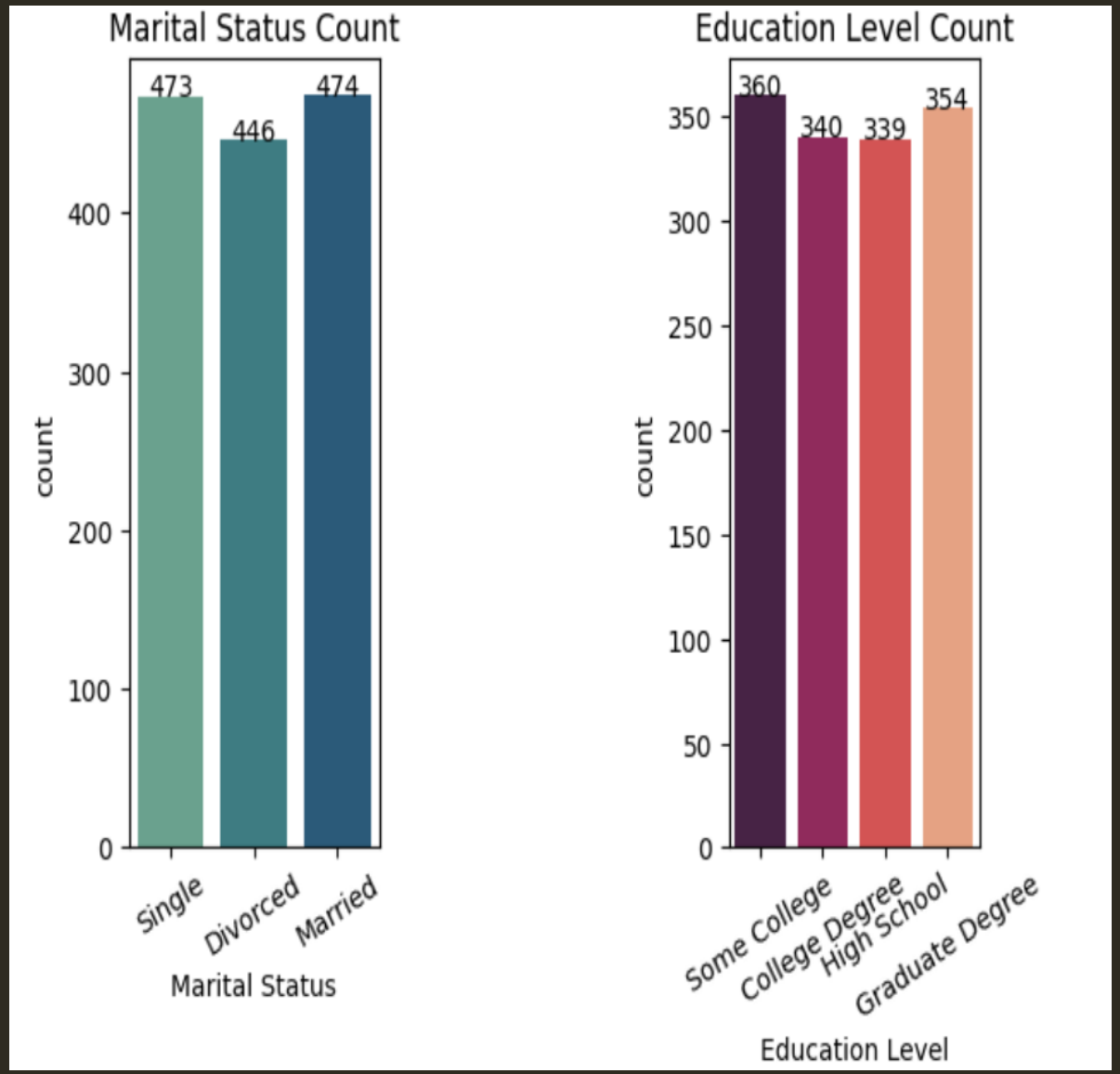
DATA VISUALIZATION

USING COUNTPLOT TO CHECKING CATEGORY COUNT

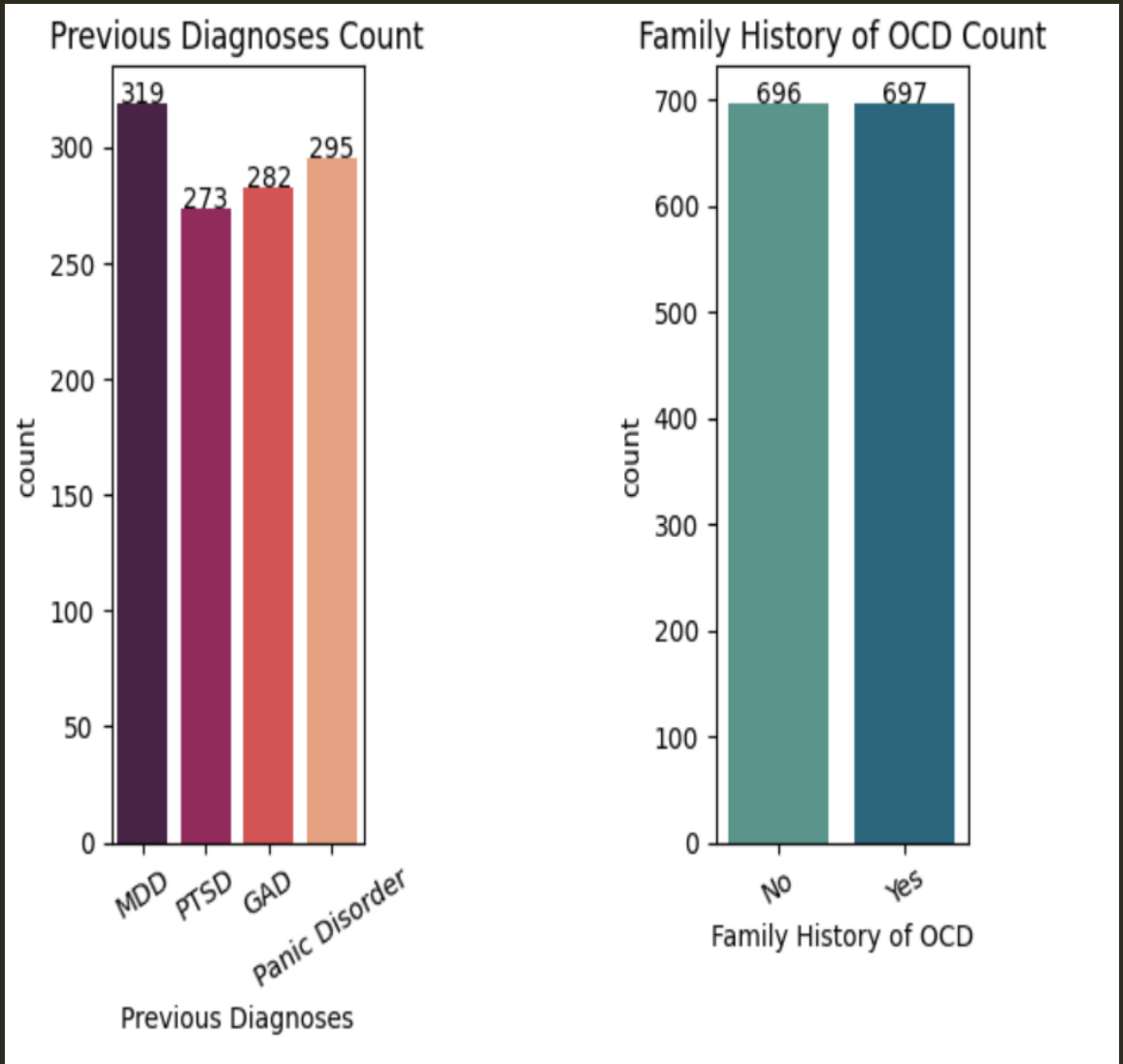
```
#countplot are use to checking category count
plt.subplot(1, 3, 1)
df = sns.countplot(data = data, x = 'Gender', hue = 'Gender')
for p in df.patches:
    df.text(p.get_x() + p.get_width()/2, p.get_height() + 0.3, int(p.get_height()), ha='center')
plt.xticks(rotation = 30);
plt.title('Gender Count')

plt.subplot(1, 3, 3)
df1 = sns.countplot(data = data, x = 'Ethnicity', palette = 'magma')
for p in df1.patches:
    df1.text(p.get_x() + p.get_width()/2, p.get_height() + 0.3, int(p.get_height()), ha='center')
plt.xticks(rotation = 30);
plt.title('Ethnicity Count')
```



```
plt.subplot(1, 3, 1)
df3 = sns.countplot(data= data, x = 'Marital Status', palette= 'crest')
for p in df3.patches:
    df3.text(p.get_x() + p.get_width()/2, p.get_height() + 0.3, int(p.get_height()), ha='center')
plt.title('Marital Status Count')
plt.xticks(rotation = 30)

plt.subplot(1, 3, 3)
df4 = sns.countplot(data = data, x = 'Education Level', palette='rocket')
for p in df4.patches:
    df4.text(p.get_x() + p.get_width()/2, p.get_height() + 0.3, int(p.get_height()), ha='center')
plt.title('Education Level Count')
plt.xticks(rotation = 30)
```



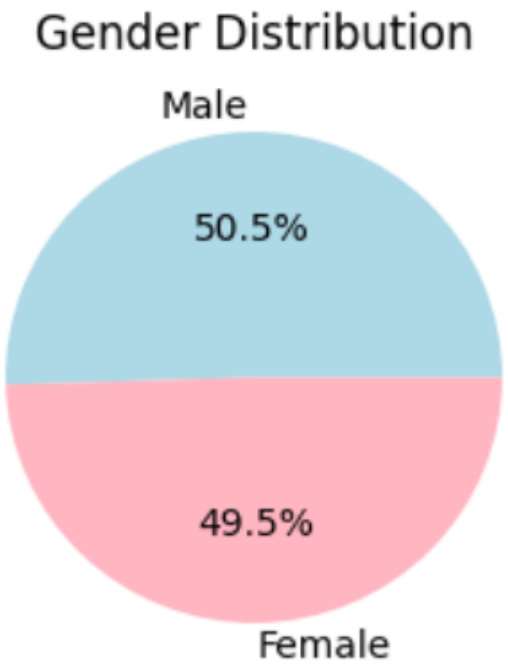
```
plt.subplot(1, 3, 1)
df5 = sns.countplot(data = data, x = 'Previous Diagnoses', palette='rocket')
for p in df5.patches:
    df5.text(p.get_x() + p.get_width()/2, p.get_height() + 0.3, int(p.get_height()), ha='center')
plt.title('Previous Diagnoses Count')
plt.xticks(rotation = 30)

plt.subplot(1, 3, 3)
df6 = sns.countplot(data= data, x = 'Family History of OCD', palette= 'crest')
for p in df6.patches:
    df6.text(p.get_x() + p.get_width()/2, p.get_height() + 0.3, int(p.get_height()), ha='center')
plt.title('Family History of OCD Count')
plt.xticks(rotation = 30)
```



GENDER DISTRIBUTION

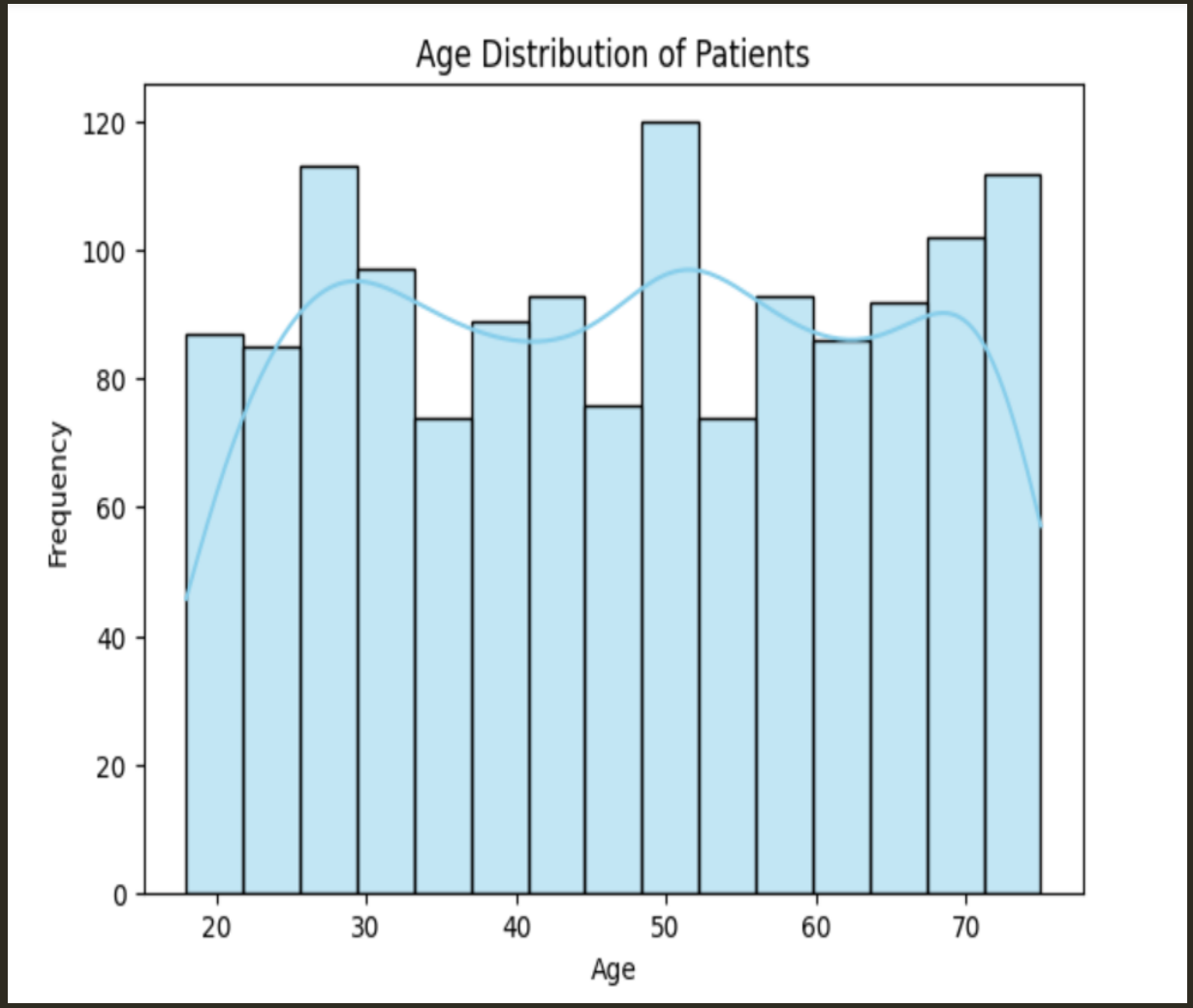
(PIE CHART USE FOR UNDERSTANDING PROPORTION OF CATEGORIES)

```
#pie chart use for understanding  proportion of categories
plt.figure(figsize=(5,3))
data["Gender"].value_counts().plot.pie(autopct="%1.1f%%", colors=["lightblue", "lightpink"])
plt.title("Gender Distribution")
plt.ylabel("")
plt.show()
```



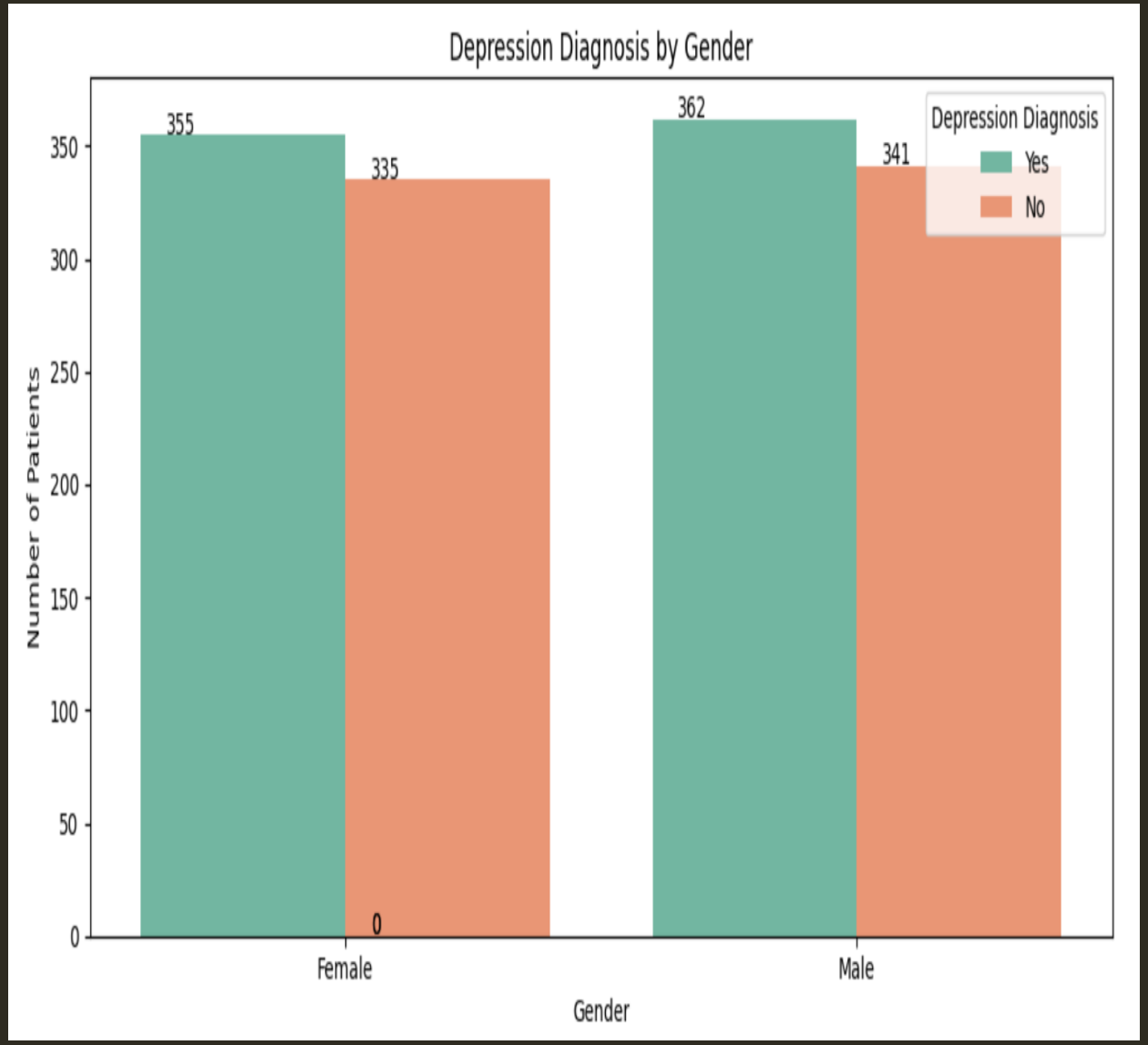
AGE DISTRIBURION OF PATIENTS

```
# Histogram checking age or any numerical distribution
plt.figure(figsize=(7,5))
sns.histplot(data["Age"], bins=15, kde = True, color="skyblue")
plt.xlabel("Age")
plt.ylabel("Frequency")
plt.title("Age Distribution of Patients")
plt.show()
```



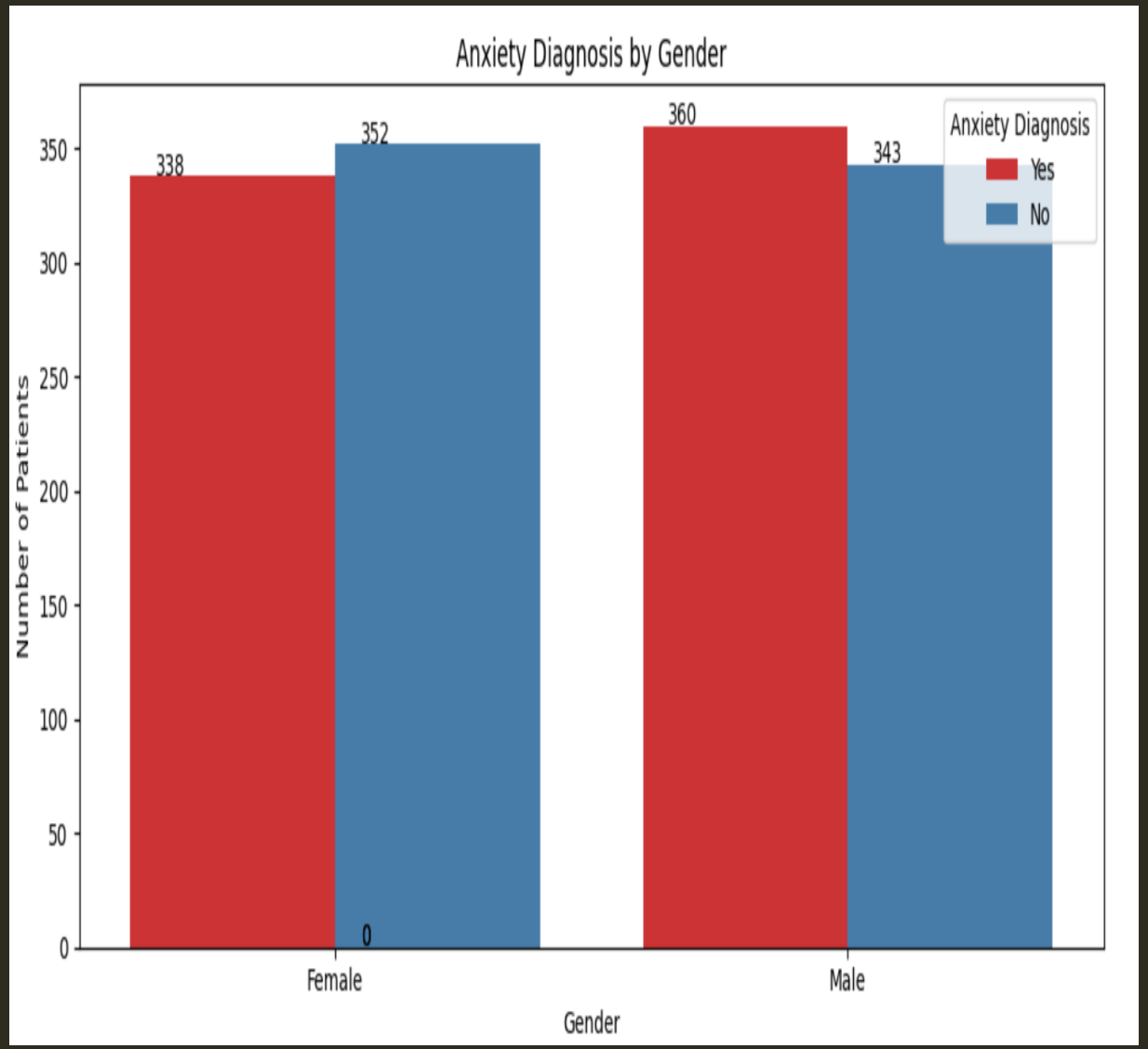
DEPRESSION DIAGNOSIS BY GENDER

```
plt.figure(figsize=(10, 5))
df7 = sns.countplot(data=data, x='Gender', hue='Depression Diagnosis', palette='Set2')
plt.title('Depression Diagnosis by Gender')
plt.xlabel('Gender')
plt.ylabel('Number of Patients')

# Add value labels
for p in df7.patches:
    df7.annotate(f'{int(p.get_height())}',
                 (p.get_x() + 0.05, p.get_height() + 0.5))

plt.legend(title='Depression Diagnosis')
plt.tight_layout()
plt.show()
```



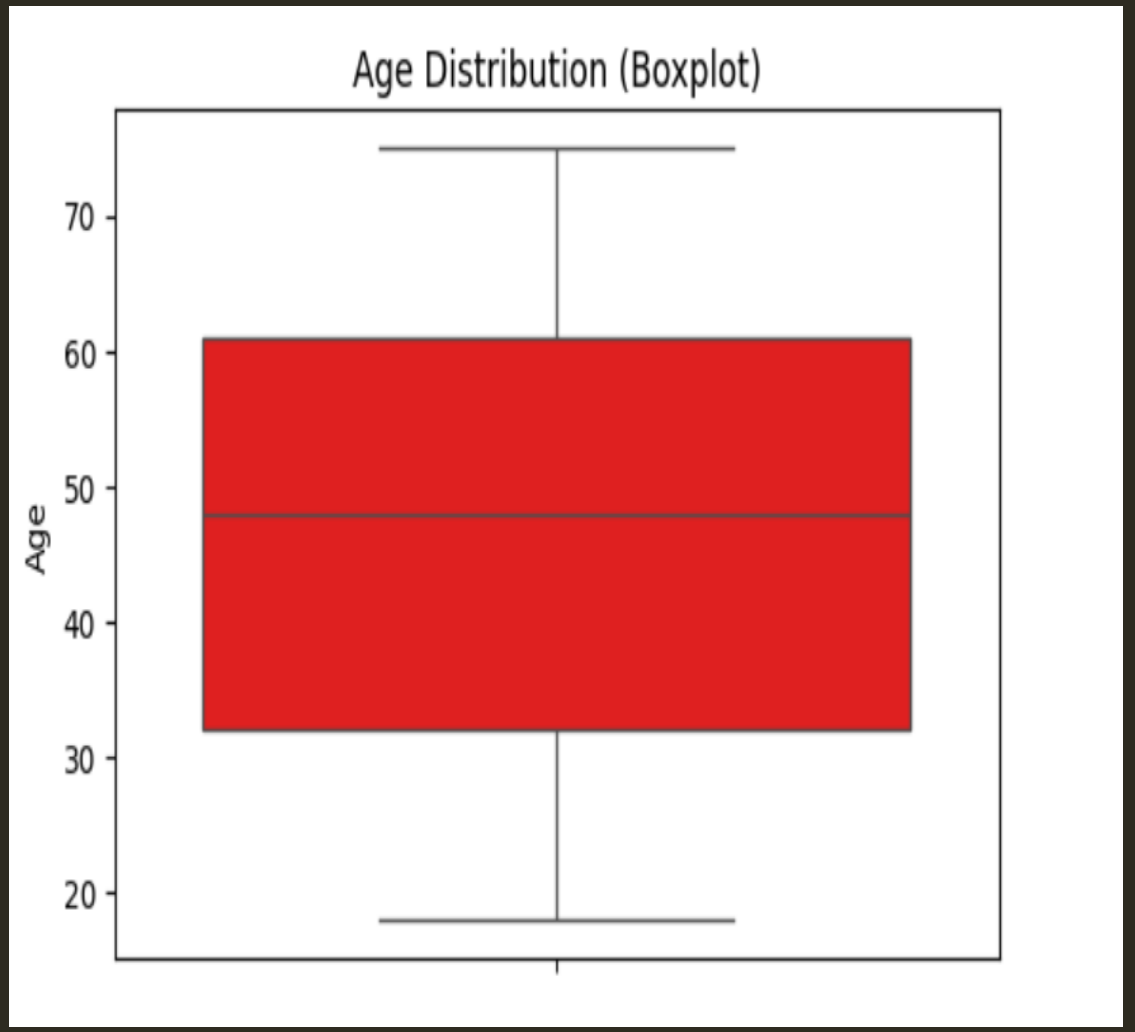
ANXIETY DIAGNOSIS BY GENDER

```
plt.figure(figsize=(10, 5))
df8 = sns.countplot(data=data, x='Gender', hue='Anxiety Diagnosis', palette='Set1')
plt.title('Anxiety Diagnosis by Gender')
plt.xlabel('Gender')
plt.ylabel('Number of Patients')

# Add value labels
for p in df8.patches:
    df8.annotate(f'{int(p.get_height())}',
                 (p.get_x() + 0.05, p.get_height() + 0.5))

plt.legend(title='Anxiety Diagnosis')
plt.tight_layout()
plt.show()
```

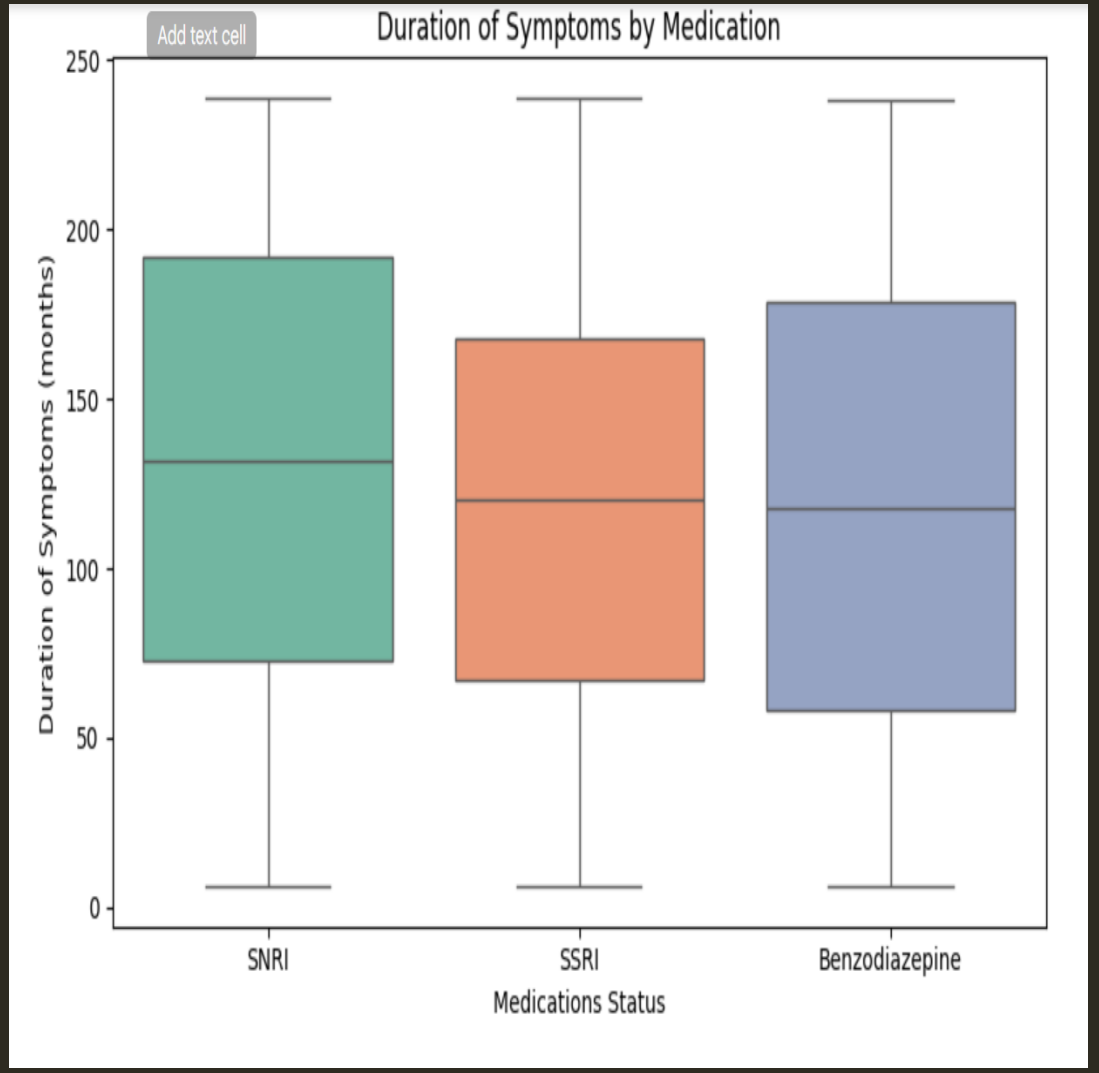


AGE DISTRIBUTION(BOXPLOT)
(DETECTING OUTLIERS)

```
#Boxplot use for detecting outliers
plt.figure(figsize=(6,4))
sns.boxplot(data=data, y="Age", color="r")
plt.ylabel("Age")
plt.title("Age Distribution (Boxplot)")
plt.show()
```

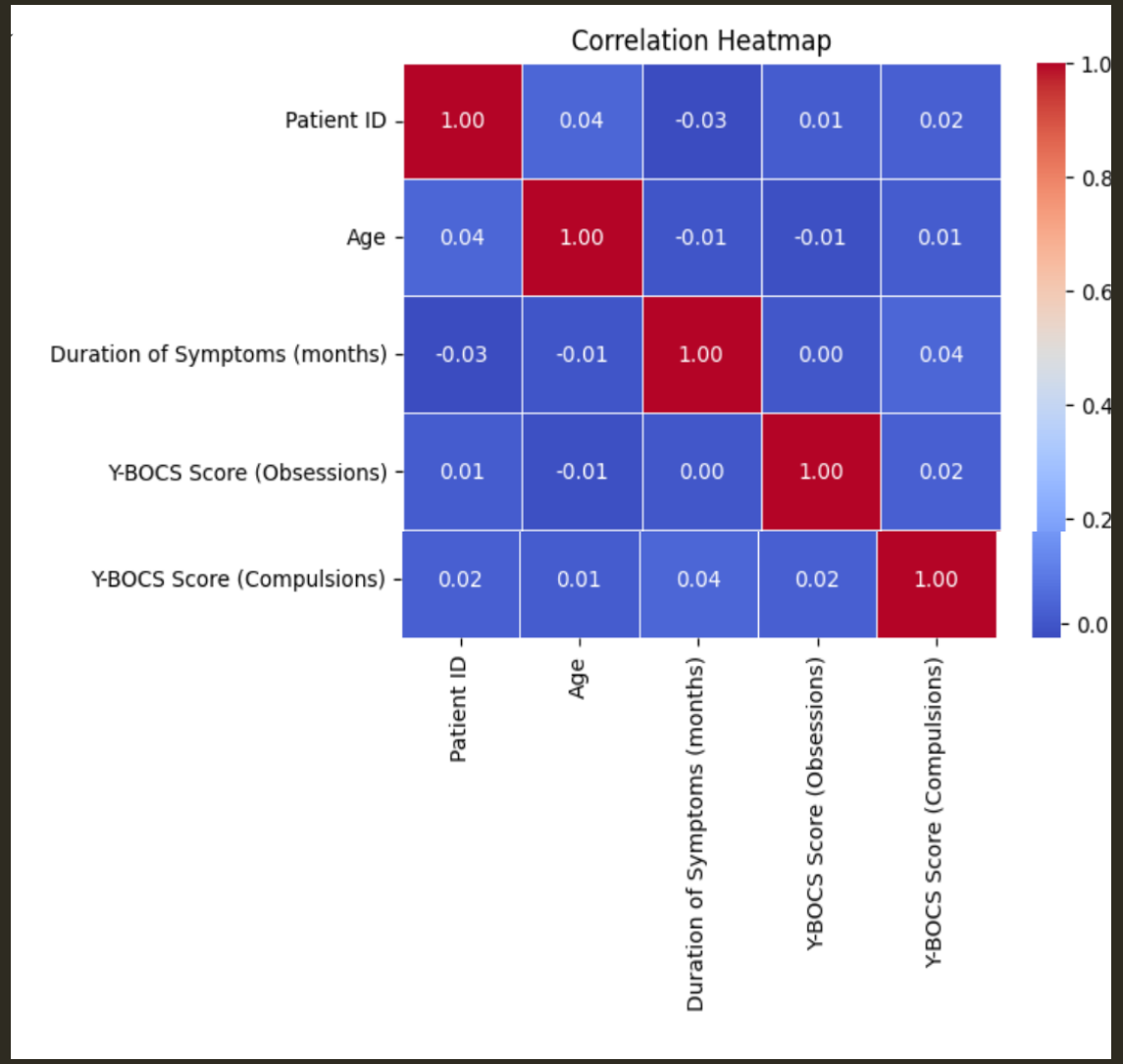


```
plt.figure(figsize=(8, 5))
sns.boxplot(data=data, x='Medications', y='Duration of Symptoms (months)', palette='Set2')
plt.title('Duration of Symptoms by Medication')
plt.xlabel('Medications Status')
plt.ylabel('Duration of Symptoms (months)')
plt.tight_layout()
plt.show()
```



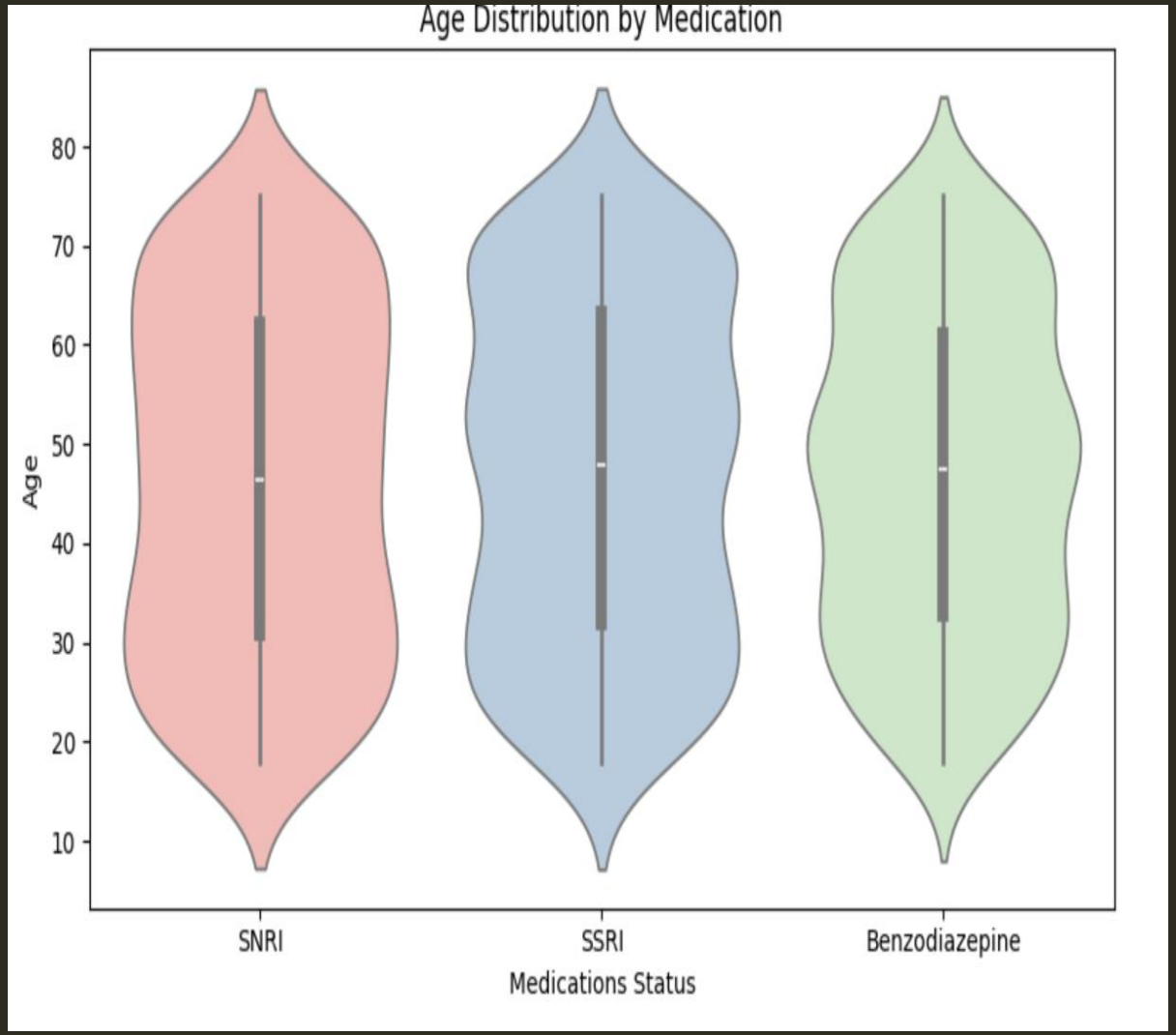
CORRELATION HEATMAP

```
# Correlation Heatmap - Finding strong relationships between variables

# Select only numerical columns for correlation
numeric_df = data.select_dtypes(include=["number"])

# Compute correlation matrix
correlation_matrix = numeric_df.corr()

# Plot heatmap
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)
plt.title("Correlation Heatmap")
plt.show()

#+1.0 - Perfect positive correlation (as one increases, the other increases)
# 0.0 - No correlation
# -1.0 -  Perfect negative correlation (as one increases, the other decreases)
```



VIOLIN PLOT – AGE DISTRIBUTION BY MEDICATIONS

```
#Violin Plot  - Understanding distribution differences
plt.figure(figsize=(8, 5))
sns.violinplot(data=data, x='Medications', y='Age', palette='Pastel1')
plt.title('Age Distribution by Medication')
plt.xlabel('Medications Status')
plt.ylabel('Age')
plt.tight_layout()
plt.show()
```



KEY INSIGHTS AND SUMMARY

1. **Age Distribution**:
- The majority of OCD patients fall between **18 to 35 years**.
- The histogram and boxplot indicate that **young adults are more affected**, with a few outliers above age 50.

2. **Gender Distribution**:
- Slightly more **male patients** than females in the dataset.
- Gender is fairly balanced, but this helps identify **target groups** for awareness and support.

3. **Education, Marital Status & Ethnicity**:
- Most patients have a background in **higher secondary or graduate education**.
- **Marital status and ethnicity** do not show strong trends, indicating OCD affects people from various social groups.

4. **Diagnosis Patterns (Depression & Anxiety):**

- A large portion of patients are also diagnosed with **Depression and Anxiety**.
- **Females** show slightly more depression cases, while anxiety is observed across genders.

5. **Family History & Previous Diagnosis:**

- Many patients have a **family history of OCD**, supporting a potential **genetic link**.
- Some have been **previously diagnosed** with other mental health conditions.

6. **Medication Patterns:**

- Patients with a **longer duration of symptoms** are more likely to be on **medication**.
- The violin and boxplots show **slight variation by age**, but duration is more strongly related to treatment.

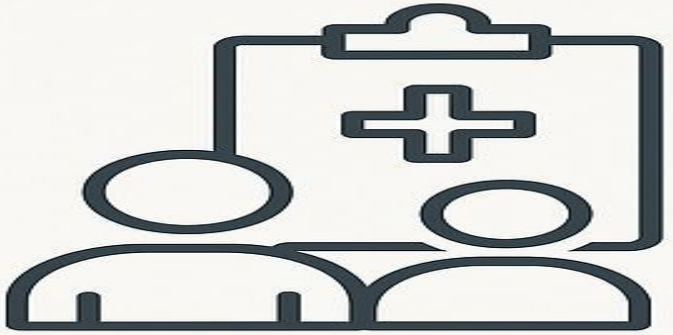
7. **Correlation Insights:**

- A **moderate correlation** exists between **obsession and compulsion scores** (Y-BOCS).
- Other numerical values (age, symptom duration) have **low correlation**, showing **individual variation** in severity and treatment.

CONCLUSION

1. The analysis of OCD patient data reveals valuable insights into their demographics, clinical patterns, and treatment behaviors.
2. Young adults, especially between 18 to 35 years, form the majority of OCD cases in the dataset.
3. A significant number of patients also suffer from comorbid conditions like depression and anxiety.
4. Family history appears to play an important role in OCD occurrence.
5. Medication is more common among patients with a longer duration of symptoms, suggesting that chronicity influences treatment plans.
6. The data shows that there is no strong correlation among most numeric features, indicating that each patient may require a personalized treatment approach.

THANK YOU



DEEPAK SEN

MAIL ID – SENDEEPAK008@GMAIL.COM

LINKEDIN - https://www.linkedin.com/in/deepak-sen-b55612226