OCD Patients EDA and ML Analysis Project

1. Importing Libraries

In [1]:
```
# Basic Libraries for Data Manipulation
import pandas as pd
import numpy as np

# Library for summary report
from summarytools import dfSummary

# Data Visualization
import matplotlib.pyplot as plt
import seaborn as sns

# Machine Learning  # type: ignore
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score
```

2. Load Data

In [2]:
```
# Load the dataset
data = pd.read_csv(r"E:\PROJECTS\Unified-Mentor_Projects\3-OCD Patient Analysis\OCD Patients Analysis-Python\OCD Pati
```

In [3]:
```
# Display the first few rows
data.head()
```

Out[3]:

	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
0	1018	32	Female	African	Single	Some College	2016-07-15	203	MDD	No	Harm-related	Checking
1	2406	69	Male	African	Divorced	Some College	2017-04-28	180	NaN	Yes	Harm-related	Washing
2	1188	57	Male	Hispanic	Divorced	College Degree	2018-02-02	173	MDD	No	Contamination	Checking
3	6200	27	Female	Hispanic	Married	College Degree	2014-08-25	126	PTSD	Yes	Symmetry	Washing
4	5824	56	Female	Hispanic	Married	High School	2022-02-20	168	PTSD	Yes	Hoarding	Ordering

3. Data Cleaning and Preprocessing

Basic Data Info

In [4]:
```
# Check data types and missing values
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Patient ID                      1500 non-null   int64
 1   Age                             1500 non-null   int64
 2   Gender                          1500 non-null   object
 3   Ethnicity                       1500 non-null   object
 4   Marital Status                  1500 non-null   object
 5   Education Level                 1500 non-null   object
 6   OCD Diagnosis Date              1500 non-null   object
 7   Duration of Symptoms (months)   1500 non-null   int64
 8   Previous Diagnoses              1252 non-null   object
 9   Family History of OCD           1500 non-null   object
 10  Obsession Type                  1500 non-null   object
 11  Compulsion Type                 1500 non-null   object
 12  Y-BOCS Score (Obsessions)       1500 non-null   int64
 13  Y-BOCS Score (Compulsions)      1500 non-null   int64
 14  Depression Diagnosis            1500 non-null   object
 15  Anxiety Diagnosis               1500 non-null   object
 16  Medications                     1114 non-null   object
dtypes: int64(5), object(12)
memory usage: 199.3+ KB
```

In [5]:
```
# Statistical summary
data.describe()
```

Out[5]:

	Patient ID	Age	Duration of Symptoms (months)	Y-BOCS Score (Obsessions)	Y-BOCS Score (Compulsions)
count	1500.000000	1500.000000	1500.000000	1500.000000	1500.00000
mean	5541.254000	46.781333	121.745333	20.048000	19.62600
std	2562.389469	16.830321	67.404610	11.823884	11.78287
min	1017.000000	18.000000	6.000000	0.000000	0.00000
25%	3338.000000	32.000000	64.000000	10.000000	9.00000
50%	5539.500000	47.000000	121.000000	20.000000	20.00000
75%	7745.500000	61.000000	178.000000	31.000000	29.00000
max	9995.000000	75.000000	240.000000	40.000000	40.00000

In [6]:
```
# Get the detailed summary of the DataFrame
dfSummary(data)
```

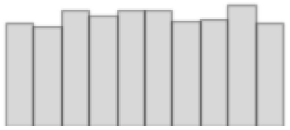
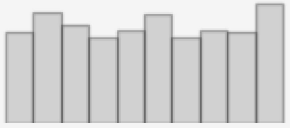
Out[6]:

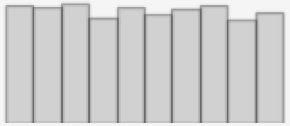
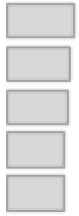
Data Frame Summary

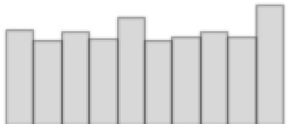
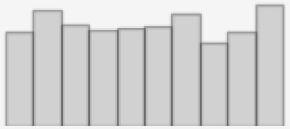
data

Dimensions: 1,500 x 17

Duplicates: 0

No	Variable	Stats / Values	Freqs / (% of Valid)	Graph	Missing
1	**Patient ID** [int64]	Mean (sd) : 5541.3 (2562.4) min < med < max: 1017.0 < 5539.5 < 9995.0 IQR (CV) : 4407.5 (2.2)	1,393 distinct values		0 (0.0%)
2	**Age** [int64]	Mean (sd) : 46.8 (16.8) min < med < max: 18.0 < 47.0 < 75.0 IQR (CV) : 29.0 (2.8)	58 distinct values		0 (0.0%)
3	**Gender** [object]	1. Male 2. Female	753 (50.2%) 747 (49.8%)		0 (0.0%)
4	**Ethnicity** [object]	1. Caucasian 2. Hispanic 3. Asian 4. African	398 (26.5%) 392 (26.1%) 386 (25.7%) 324 (21.6%)		0 (0.0%)
5	**Marital Status** [object]	1. Single 2. Married 3. Divorced	511 (34.1%) 507 (33.8%) 482 (32.1%)		0 (0.0%)
6	**Education Level** [object]	1. Some College 2. Graduate Degree 3. College Degree 4. High School	394 (26.3%) 376 (25.1%) 366 (24.4%) 364 (24.3%)		0 (0.0%)

No	Variable	Stats / Values	Freqs / (% of Valid)	Graph	Missing
7	**OCD Diagnosis Date** [object]	1. 2017-06-21 2. 2019-04-29 3. 2014-02-06 4. 2016-03-16 5. 2022-09-28 6. 2021-08-13 7. 2014-01-01 8. 2016-09-02 9. 2021-09-30 10. 2017-02-04 11. other	4 (0.3%) 3 (0.2%) 3 (0.2%) 3 (0.2%) 3 (0.2%) 3 (0.2%) 3 (0.2%) 3 (0.2%) 3 (0.2%) 3 (0.2%) 1,469 (97.9%)		0 (0.0%)
8	**Duration of Symptoms (months)** [int64]	Mean (sd) : 121.7 (67.4) min < med < max: 6.0 < 121.0 < 240.0 IQR (CV) : 114.0 (1.8)	235 distinct values		0 (0.0%)
9	**Previous Diagnoses** [object]	1. MDD 2. Panic Disorder 3. GAD 4. PTSD 5. nan	345 (23.0%) 313 (20.9%) 298 (19.9%) 296 (19.7%) 248 (16.5%)		248 (16.5%)
10	**Family History of OCD** [object]	1. Yes 2. No	760 (50.7%) 740 (49.3%)		0 (0.0%)
11	**Obsession Type** [object]	1. Harm-related 2. Contamination 3. Religious 4. Symmetry 5. Hoarding	333 (22.2%) 306 (20.4%) 303 (20.2%) 280 (18.7%) 278 (18.5%)		0 (0.0%)

No	Variable	Stats / Values	Freqs / (% of Valid)	Graph	Missing
12	**Compulsion Type** [object]	1. Washing 2. Counting 3. Checking 4. Praying 5. Ordering	321 (21.4%) 316 (21.1%) 292 (19.5%) 286 (19.1%) 285 (19.0%)		0 (0.0%)
13	**Y-BOCS Score (Obsessions)** [int64]	Mean (sd) : 20.0 (11.8) min < med < max: 0.0 < 20.0 < 40.0 IQR (CV) : 21.0 (1.7)	41 distinct values		0 (0.0%)
14	**Y-BOCS Score (Compulsions)** [int64]	Mean (sd) : 19.6 (11.8) min < med < max: 0.0 < 20.0 < 40.0 IQR (CV) : 20.0 (1.7)	41 distinct values		0 (0.0%)
15	**Depression Diagnosis** [object]	1. Yes 2. No	772 (51.5%) 728 (48.5%)		0 (0.0%)
16	**Anxiety Diagnosis** [object]	1. Yes 2. No	751 (50.1%) 749 (49.9%)		0 (0.0%)
17	**Medications** [object]	1. nan 2. Benzodiazepine 3. SNRI 4. SSRI	386 (25.7%) 386 (25.7%) 376 (25.1%) 352 (23.5%)		386 (25.7%)

In [7]:
```
dfSummary(data, is_collapsible= True)
```

Out[7]:

Show Summary - data

In [8]:
```
# Check for missing values
missing_values = data.isnull().sum()
```

```
missing_values[missing_values > 0]
```

Out[8]:
```
Previous Diagnoses    248
Medications           386
dtype: int64
```

Handling Missing Values

In [9]:
```
# Fill missing values in 'Previous Diagnoses' and 'Medications' "df.method({col: value}, inplace=True)"
data.fillna({'Previous Diagnoses': 'No Previous Diagnosis'}, inplace=True)
data.fillna({'Medications': 'No Medication'}, inplace=True)
```

In [10]:
```
# Convert 'OCD Diagnosis Date' to datetime
data['OCD Diagnosis Date'] = pd.to_datetime(data['OCD Diagnosis Date'], errors= 'coerce')
```

In [11]:
```
# Check for any remaining missing values
data.isnull().sum()
```

Out[11]:
```
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: int64
```

In [12]:
```
# Verfiy the data
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Patient ID                      1500 non-null   int64
 1   Age                             1500 non-null   int64
 2   Gender                          1500 non-null   object
 3   Ethnicity                       1500 non-null   object
 4   Marital Status                  1500 non-null   object
 5   Education Level                 1500 non-null   object
 6   OCD Diagnosis Date              1500 non-null   datetime64[ns]
 7   Duration of Symptoms (months)   1500 non-null   int64
 8   Previous Diagnoses              1500 non-null   object
 9   Family History of OCD           1500 non-null   object
 10  Obsession Type                  1500 non-null   object
 11  Compulsion Type                 1500 non-null   object
 12  Y-BOCS Score (Obsessions)       1500 non-null   int64
 13  Y-BOCS Score (Compulsions)      1500 non-null   int64
 14  Depression Diagnosis            1500 non-null   object
 15  Anxiety Diagnosis               1500 non-null   object
 16  Medications                     1500 non-null   object
dtypes: datetime64[ns](1), int64(5), object(11)
memory usage: 199.3+ KB
```

In [13]:
```
# Recheck the detailed summary of the DataFrame
dfSummary(data, is_collapsible= True)
```

Out[13]:

Show Summary - data

Encoding Categorical Variables

In [14]:
```
# List of categorical columns to encode
categorical_cols = ['Gender', 'Ethnicity', 'Marital Status', 'Education Level',
                    'Previous Diagnoses', 'Family History of OCD', 'Obsession Type',
                    'Compulsion Type', 'Depression Diagnosis', 'Anxiety Diagnosis', 'Medications']


# Use LabelEncoder for simplicity
```

```
le = LabelEncoder()
for col in categorical_cols:
    data[col] = le.fit_transform(data[col])
```

In [15]:
```
# Verify the changes
data.head()
```

Out[15]:

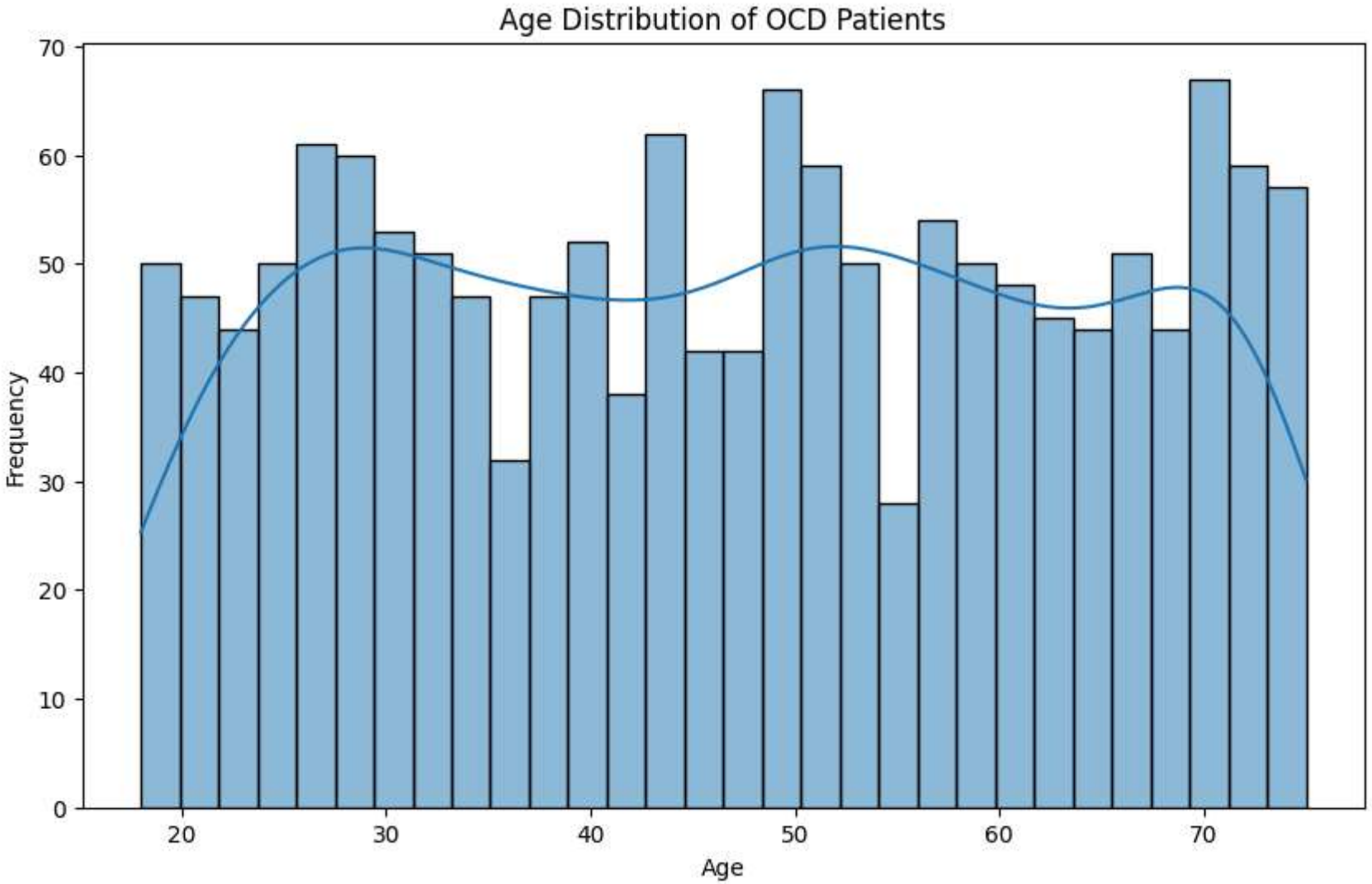
	Patient ID	Age	Gender	Ethnicity	Marital Status	Education Level	OCD Diagnosis Date	Duration of Symptoms (months)	Previous Diagnoses	Family History of OCD	Obsession Type	Compulsion Type	(Ob
0	1018	32	0	0	2	3	2016-07-15	203	1	0	1	0	
1	2406	69	1	0	0	3	2017-04-28	180	2	1	1	4	
2	1188	57	1	3	0	0	2018-02-02	173	1	0	0	0	
3	6200	27	0	3	1	0	2014-08-25	126	3	1	4	4	
4	5824	56	0	3	1	2	2022-02-20	168	3	1	2	2	

4. Exploratory Data Analysis

Demographic Analysis

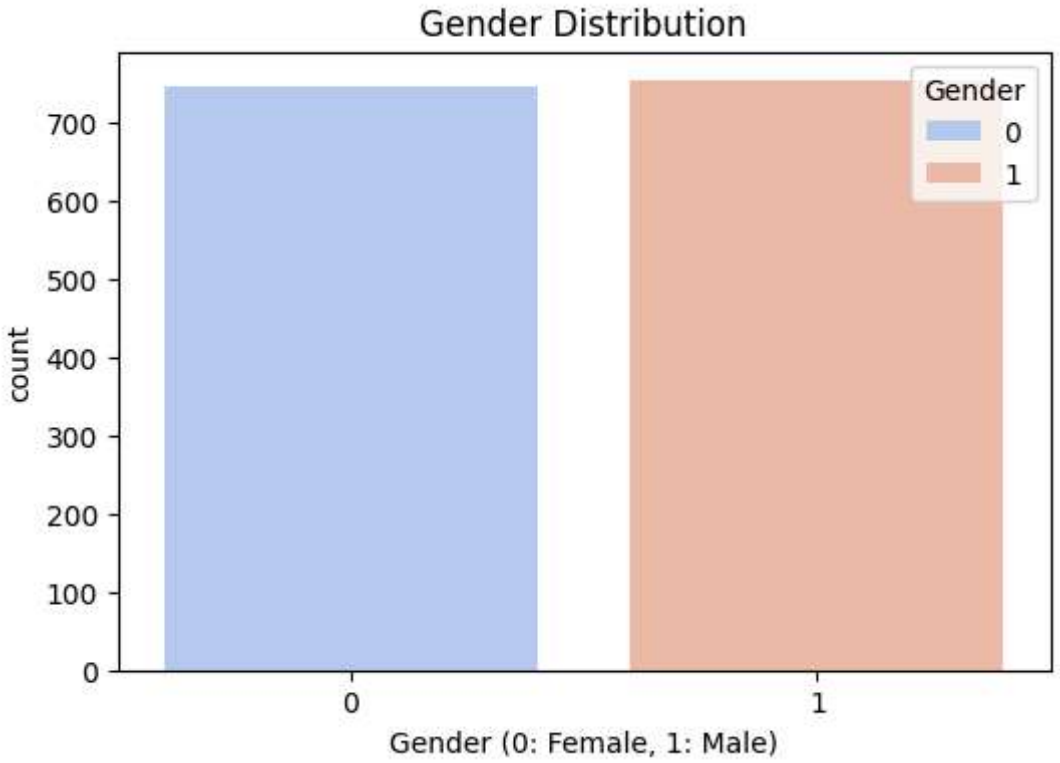
1. Age Distribution

In [16]:
```
plt.figure(figsize = (10, 6))
sns.histplot(data['Age'], bins=30, kde=True)
plt.title('Age Distribution of OCD Patients')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```



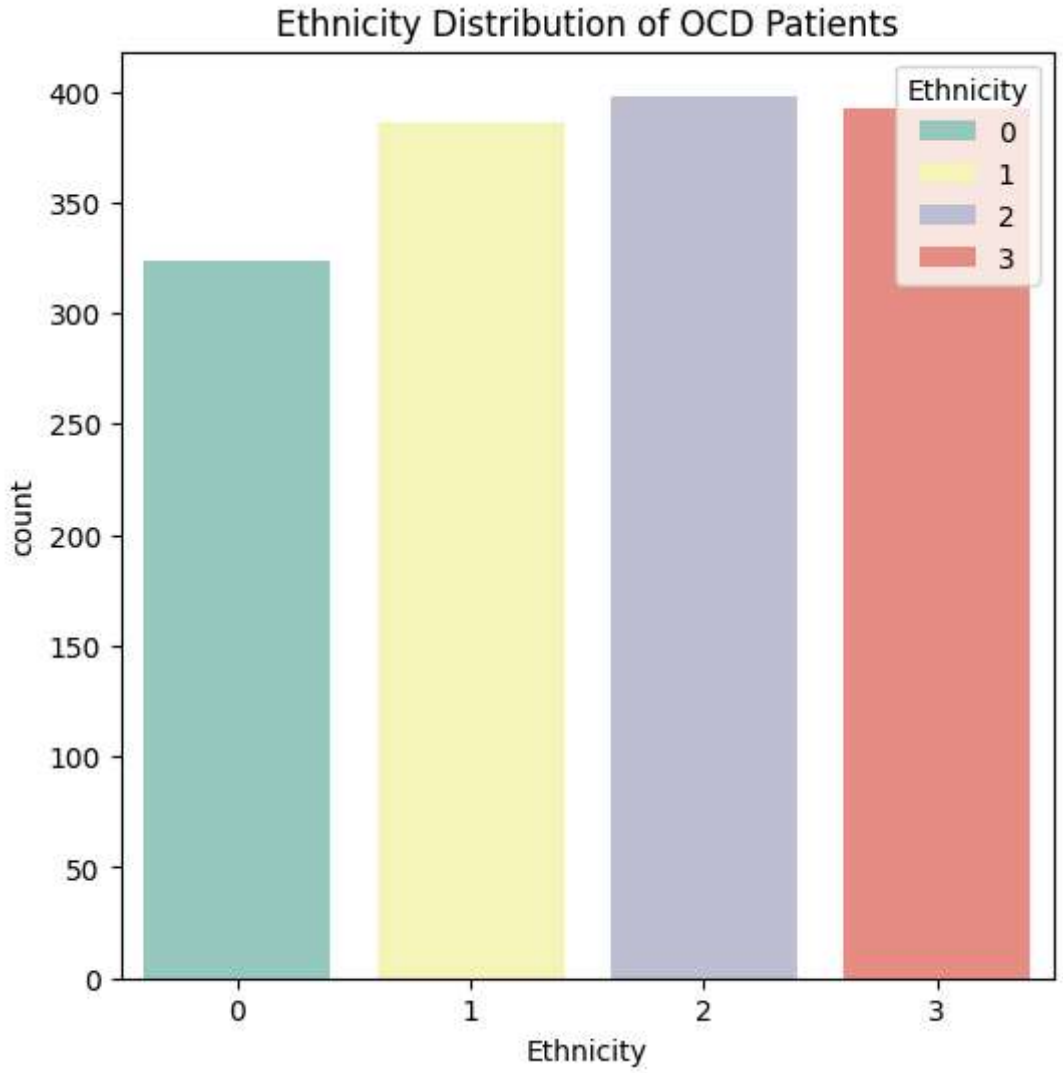
2. Gender Distribution

In [17]:
```
plt.figure(figsize=(6, 4))
sns.countplot(x='Gender', data=data, hue='Gender', palette='coolwarm', legend=True)
plt.title('Gender Distribution')
plt.xlabel('Gender (0: Female, 1: Male)')
plt.ylabel('count')
plt.show()
```



3. Ethnicity Distribution

In [18]:
```
plt.figure(figsize=(6, 6))
sns.countplot(x='Ethnicity', data=data, hue='Ethnicity', palette='Set3', legend=True)
plt.title('Ethnicity Distribution of OCD Patients')
plt.show()
```

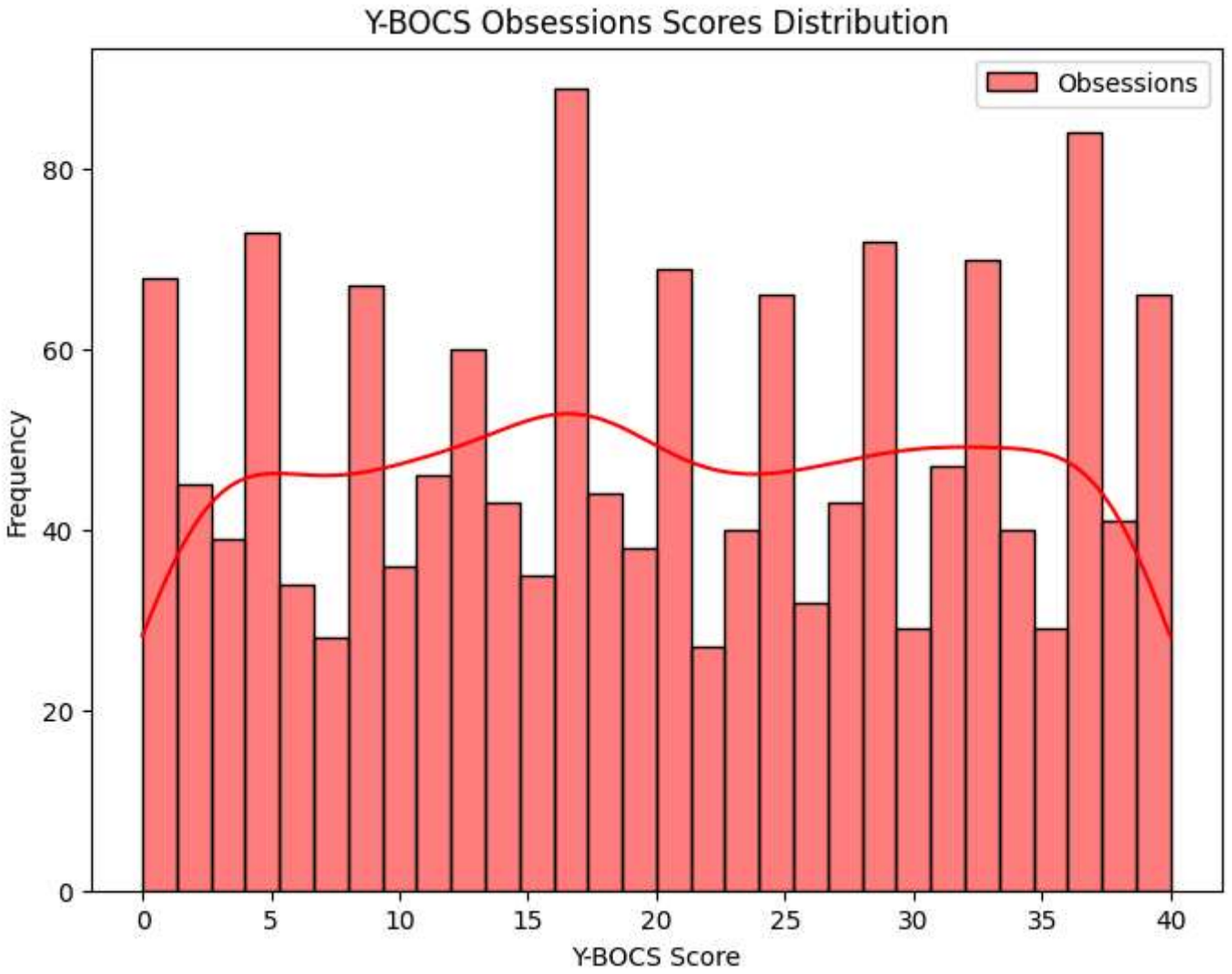


Clinical Data Analysis

1. Y-BOCS Scores Distribution for Obsessions

In [19]:
```
plt.figure(figsize=(8, 6))
sns.histplot(data['Y-BOCS Score (Obsessions)'], bins=30, kde=True, color='red', label='Obsessions')
plt.title('Y-BOCS Obsessions Scores Distribution')
plt.xlabel('Y-BOCS Score')
```

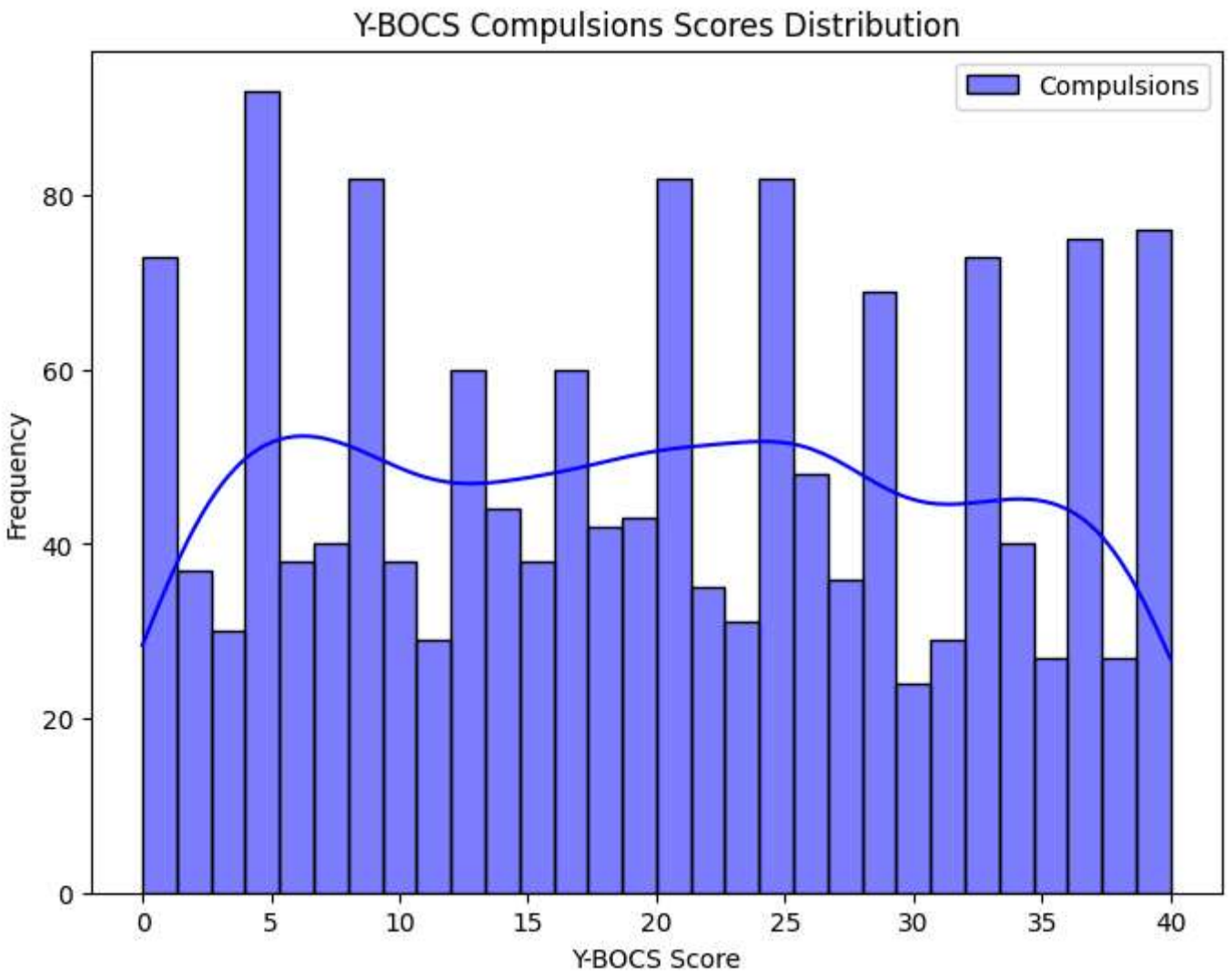
```
plt.ylabel('Frequency')
plt.legend()
plt.show()
```



2. Y-BOCS Scores Distribution for Compulsions

In [20]:
```
plt.figure(figsize=(8, 6))
sns.histplot(data['Y-BOCS Score (Compulsions)'], bins=30, kde=True, color='blue', label='Compulsions')
plt.title('Y-BOCS Compulsions Scores Distribution')
```

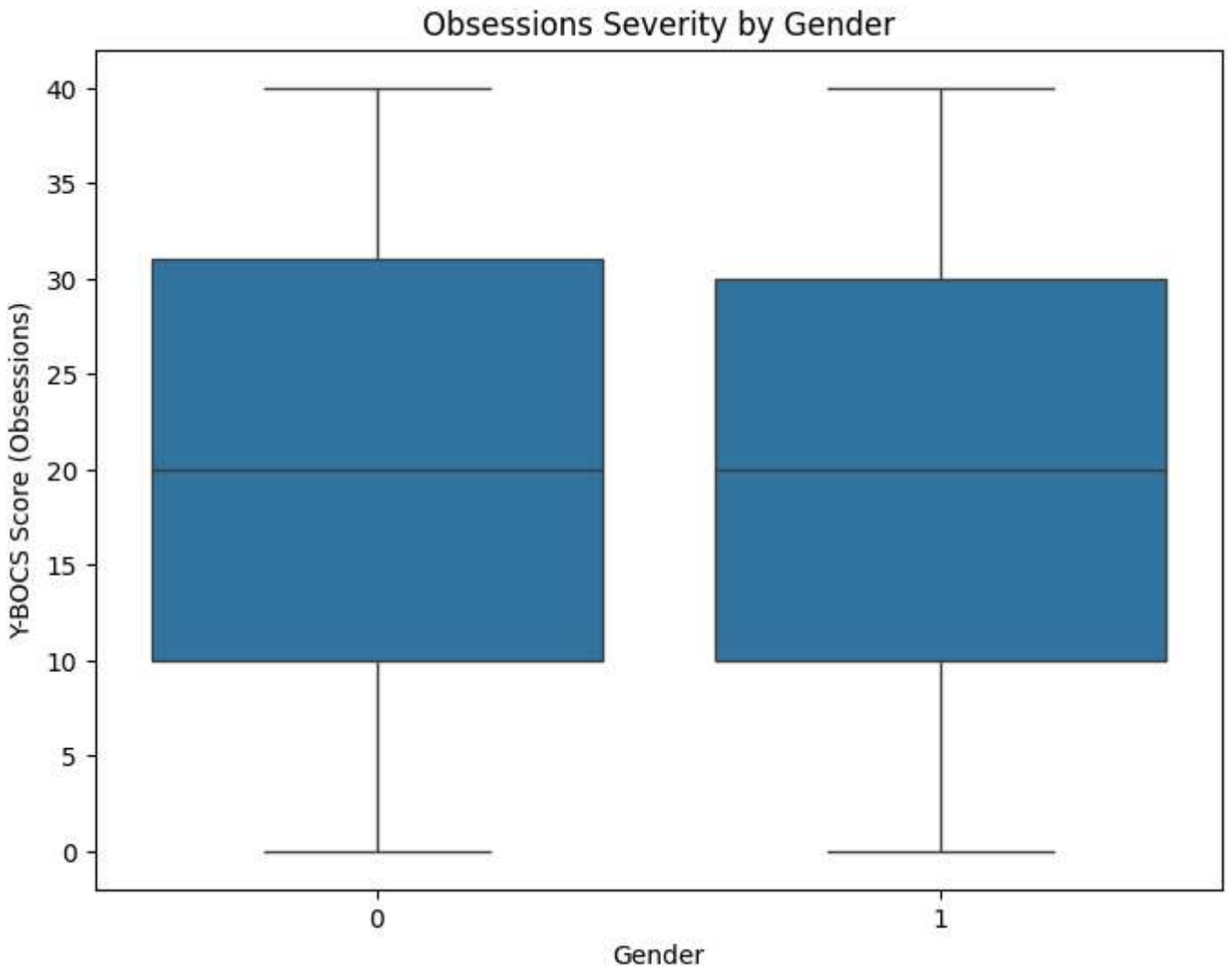
```
plt.xlabel('Y-BOCS Score')
plt.ylabel('Frequency')
plt.legend()
plt.show()
```



3. Boxplot of Y-BOCS Scores Distribution by Gender

In [21]:
```
plt.figure(figsize=(8, 6))
sns.boxplot(x='Gender', y='Y-BOCS Score (Obsessions)', data=data)
```

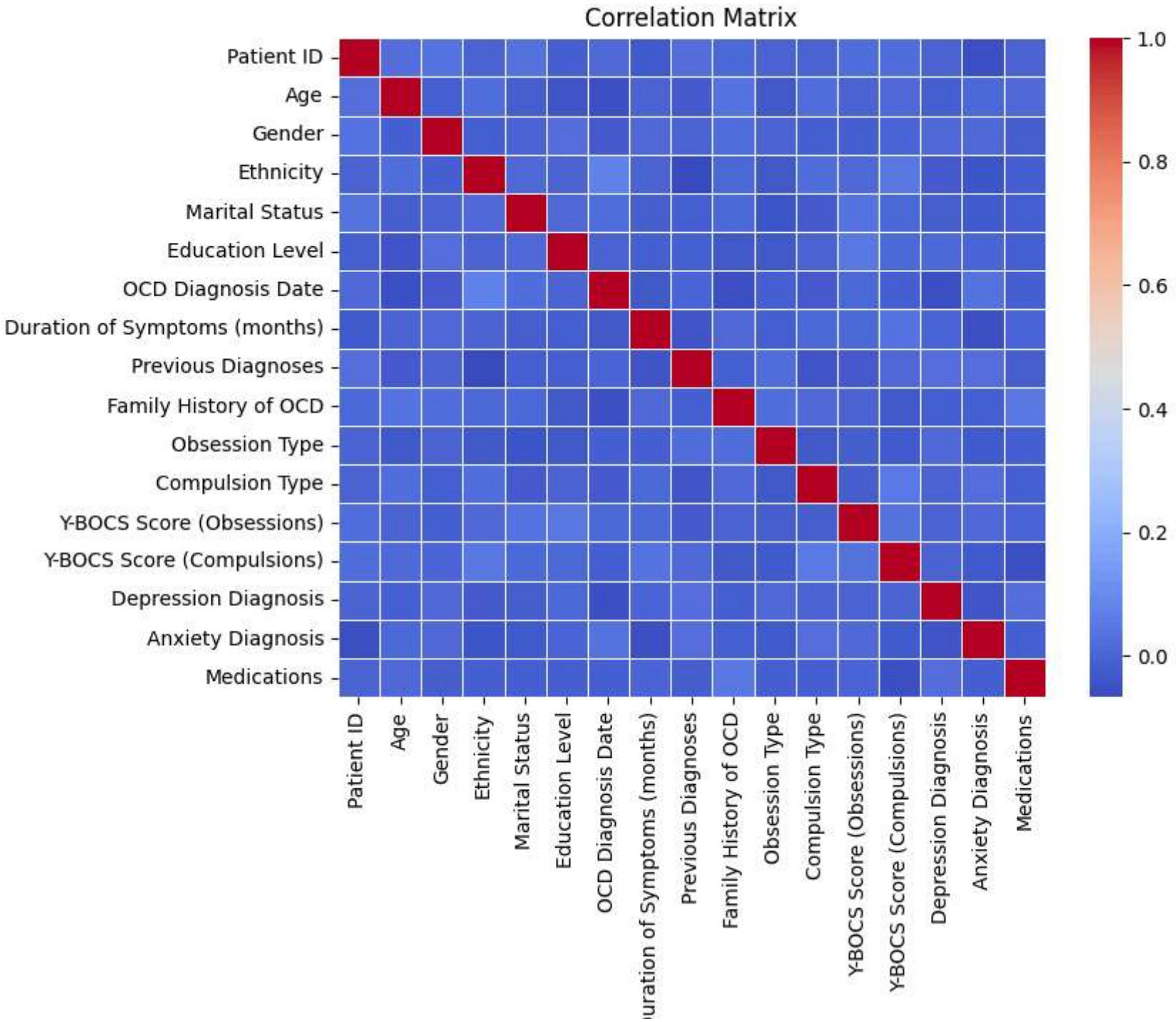
```
plt.title('Obsessions Severity by Gender')
plt.show()
```



Relationships between Demographics & Clinical Data

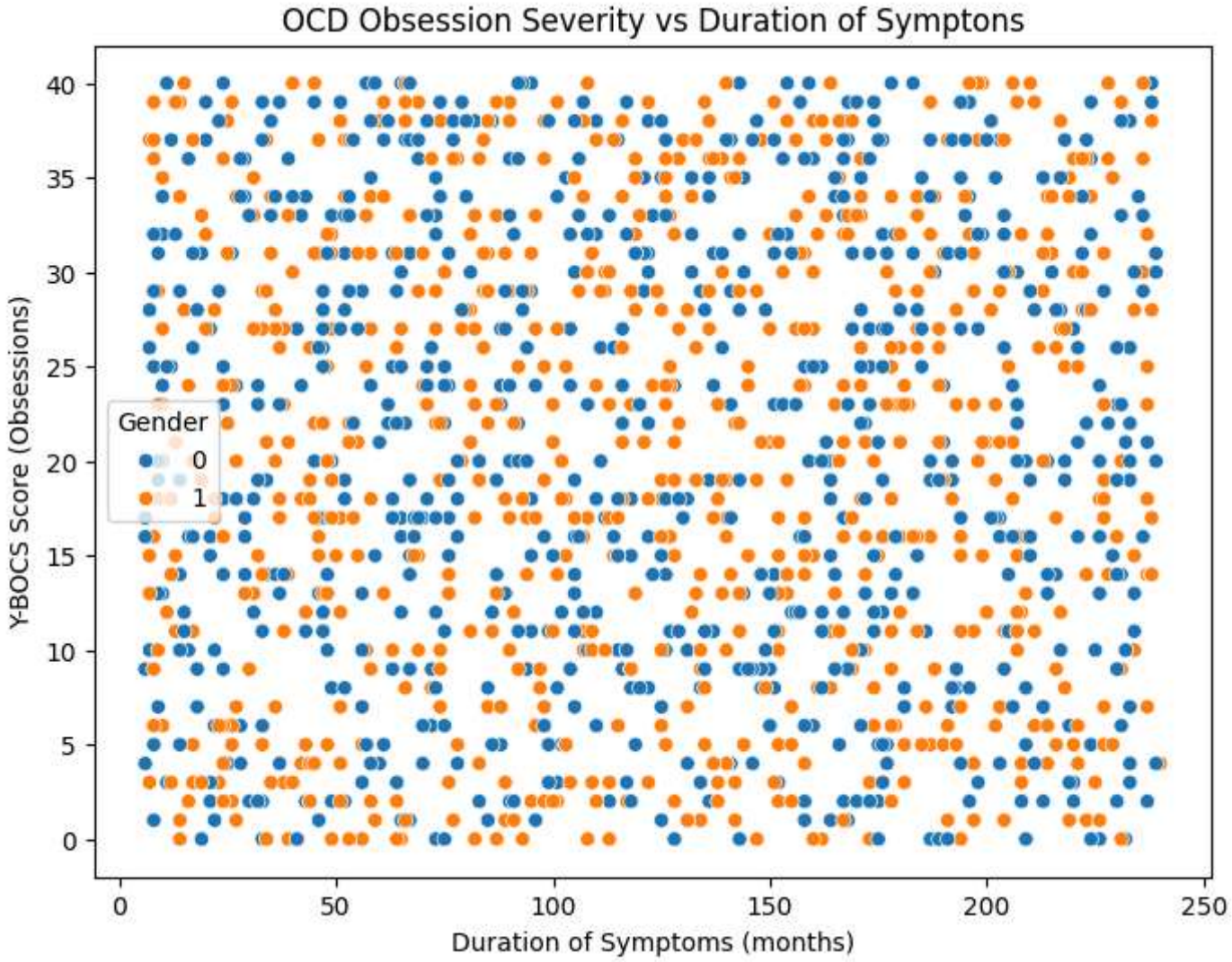
1. Correlation Heatmap

In [22]:
```
plt.figure(figsize=(8, 6))
sns.heatmap(data.corr(), annot=False, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Matrix')
plt.show()
```


Correlation Matrix
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
1.0
0.8
0.6
0.4
0.2
0.0

2. OCD Severity based on duration of Symptons

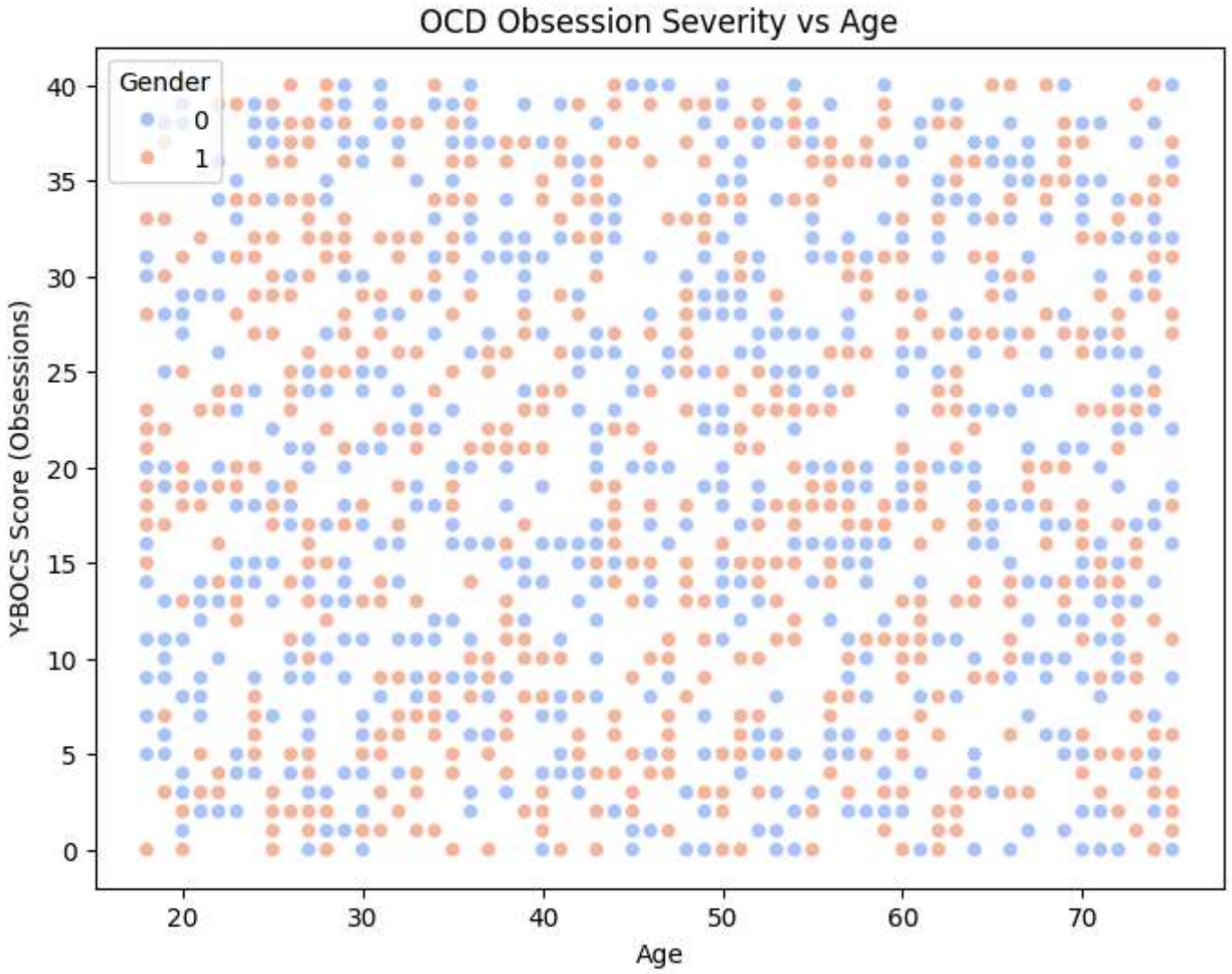
In [23]:
```
plt.figure(figsize=(8, 6))
sns.scatterplot(x='Duration of Symptoms (months)',  y='Y-BOCS Score (Obsessions)', hue='Gender', data=data)
plt.title('OCD Obsession Severity vs Duration of Symptons')
plt.show()
```

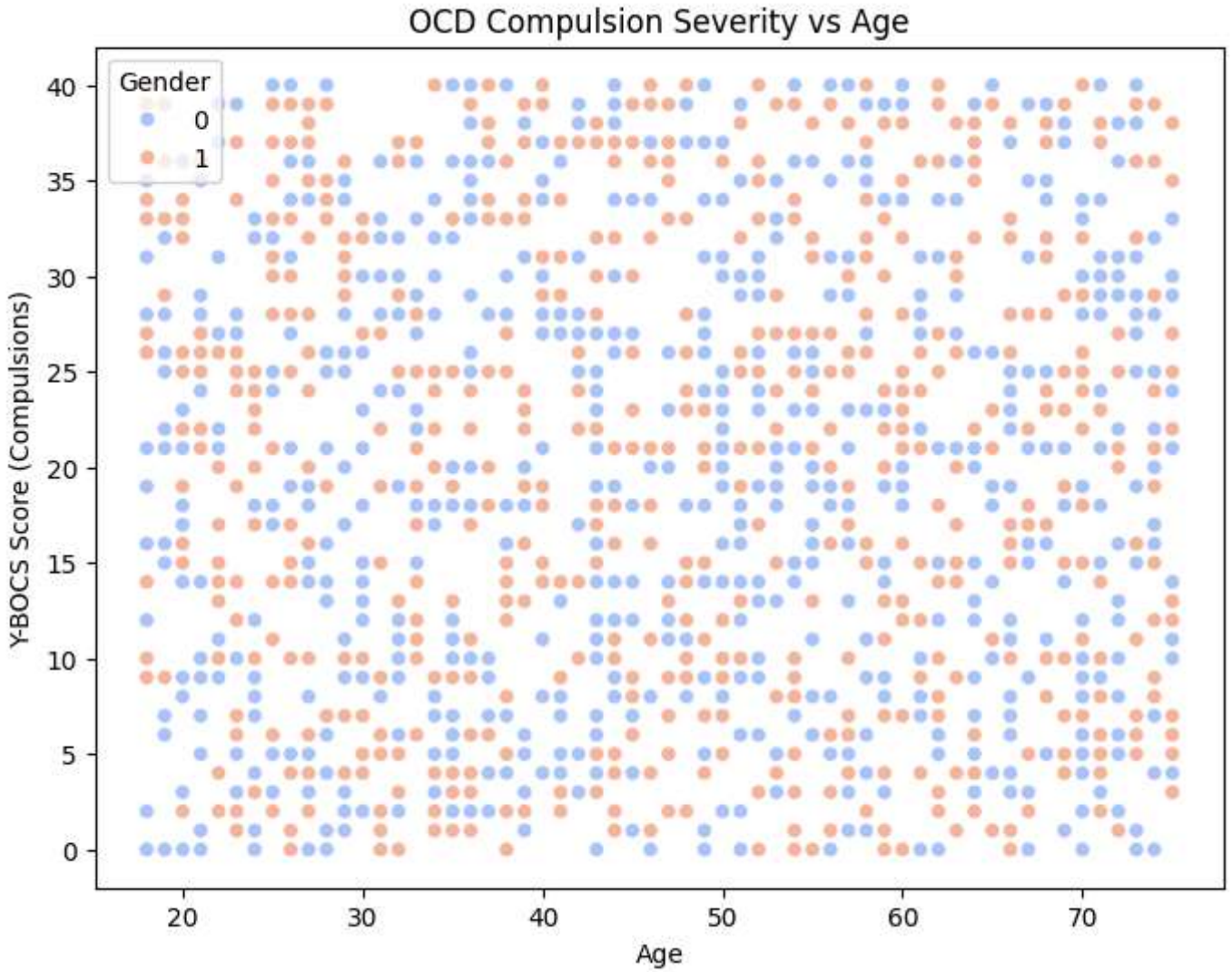


5. Data Storytelling and Insights

Age & OCD Severity

In [24]:
```
# Scatter Plot: Age vs Y-BOCS Obsessions
plt.figure(figsize=(8,6))
sns.scatterplot(x='Age', y='Y-BOCS Score (Obsessions)', data=data, hue='Gender', palette='coolwarm')
plt.title('OCD Obsession Severity vs Age')
plt.xlabel('Age')
plt.ylabel('Y-BOCS Score (Obsessions)')
plt.show()

# Scatter plot: Age vs Y-BOCS Compulsions
plt.figure(figsize=(8,6))
sns.scatterplot(x='Age', y='Y-BOCS Score (Compulsions)', data=data, hue='Gender', palette='coolwarm')
plt.title('OCD Compulsion Severity vs Age')
plt.xlabel('Age')
plt.ylabel('Y-BOCS Score (Compulsions)')
plt.show()
```


OCD Obsession Severity vs Age
Gender
0
1
Y-BOCS Score (Obsessions)
Age
0
5
10
15
20
25
30
35
40
20
30
40
50
60
70



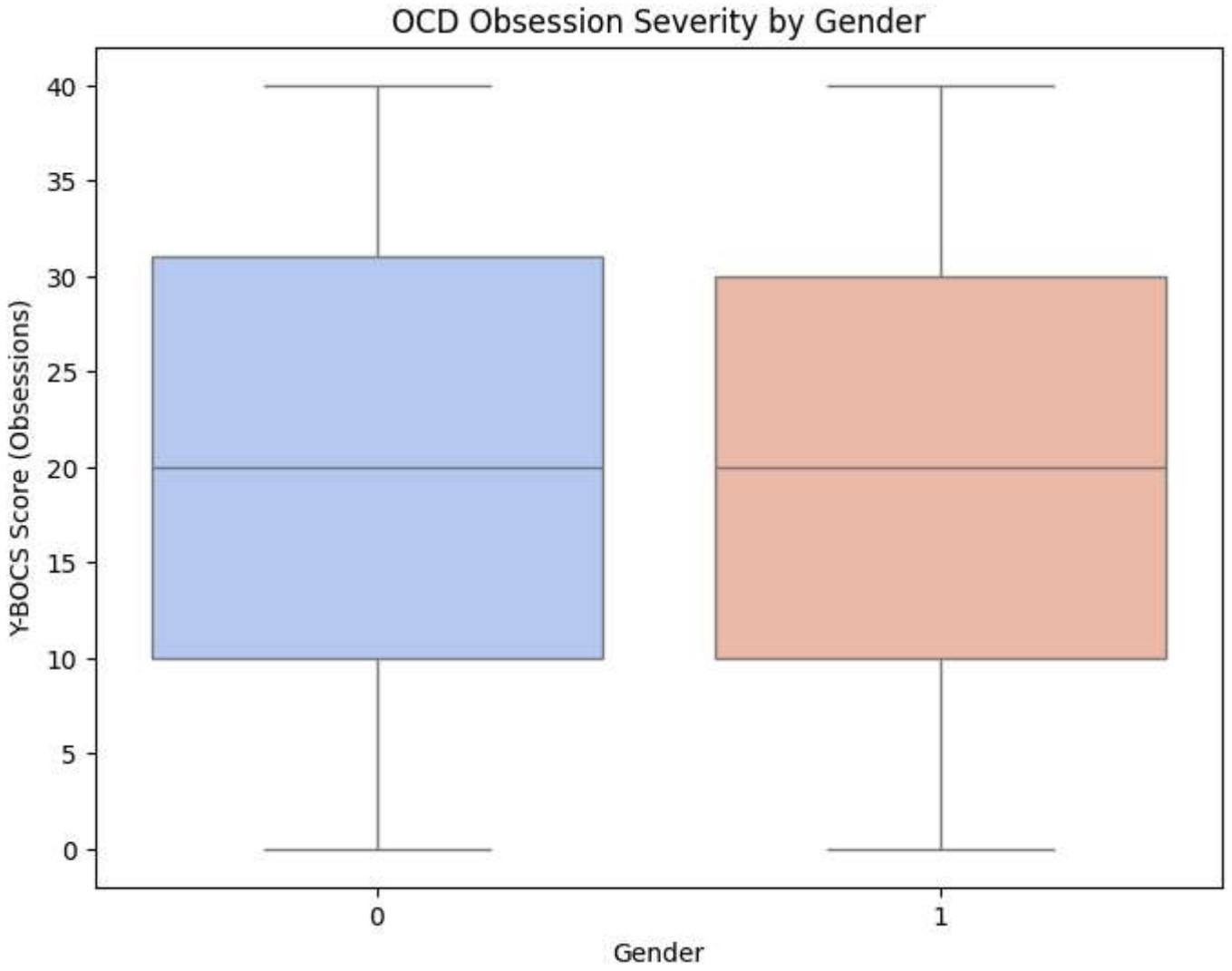
Observation: There appears to be no strong correlation between age and obsessions, age and compulsions scores, indicating that OCD severity is relatively consistent across different age groups

2. Gender Differences

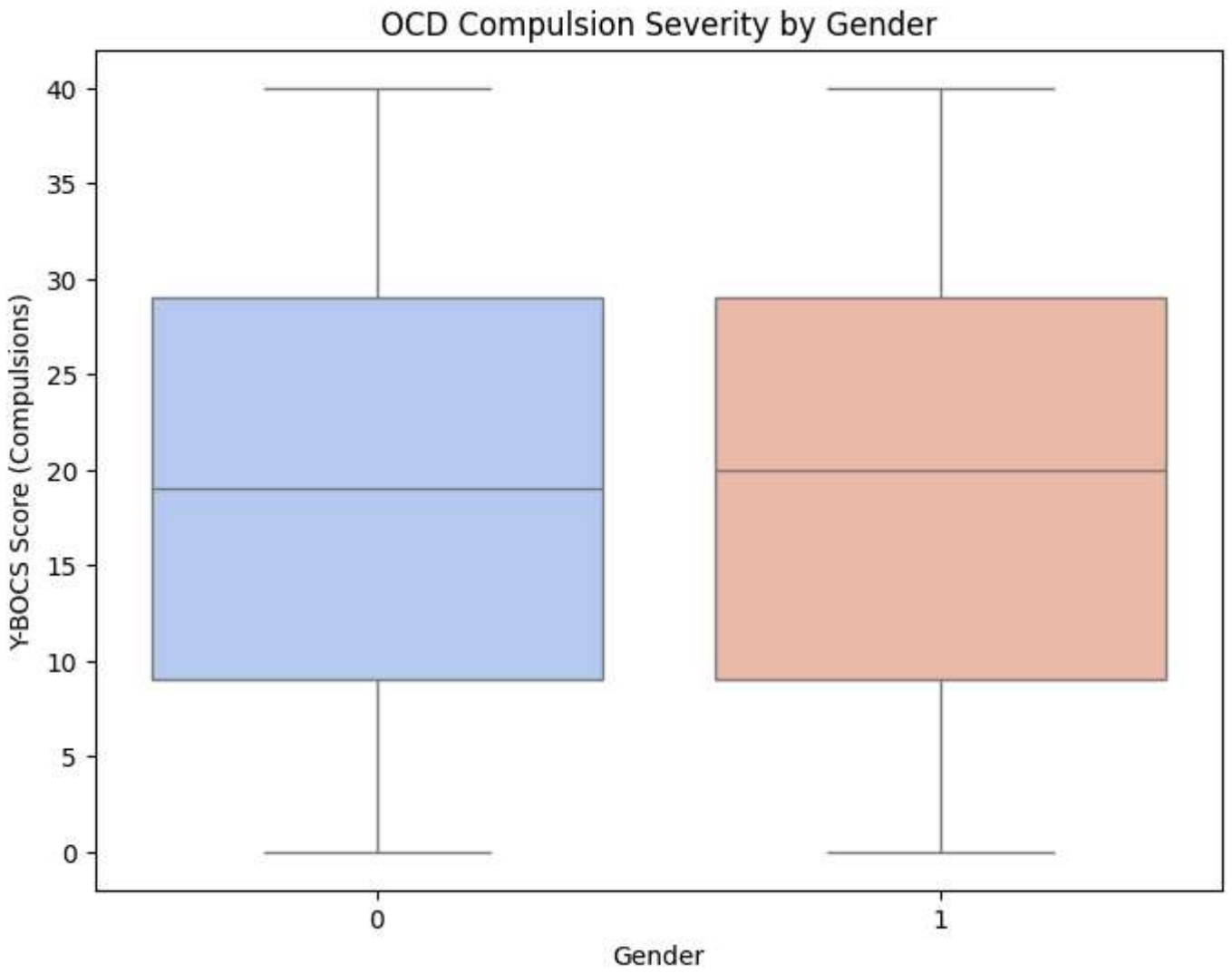
In [25]:
```
# Boxplot: Gender vs Y-BOCS Obsessions
plt.figure(figsize=(8,6))
sns.boxplot(x='Gender', y='Y-BOCS Score (Obsessions)', data=data, palette='coolwarm')
```

```
plt.title('OCD Obsession Severity by Gender')
plt.xlabel('Gender')
plt.ylabel('Y-BOCS Score (Obsessions)')
plt.show()

# Boxplot: Gender vs Y-BOCS Compulsions
plt.figure(figsize=(8,6))
sns.boxplot(x='Gender', y='Y-BOCS Score (Compulsions)', data=data, palette='coolwarm')
plt.title('OCD Compulsion Severity by Gender')
plt.xlabel('Gender')
plt.ylabel('Y-BOCS Score (Compulsions)')
plt.show()
```

```
C:\Users\Kumar\AppData\Local\Temp\ipykernel_44780\3344031067.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `h
ue` and set `legend=False` for the same effect.

  sns.boxplot(x='Gender', y='Y-BOCS Score (Obsessions)', data=data, palette='coolwarm')
```




```
C:\Users\Kumar\AppData\Local\Temp\ipykernel_44780\3344031067.py:11: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `h
ue` and set `legend=False` for the same effect.

  sns.boxplot(x='Gender', y='Y-BOCS Score (Compulsions)', data=data, palette='coolwarm')
```



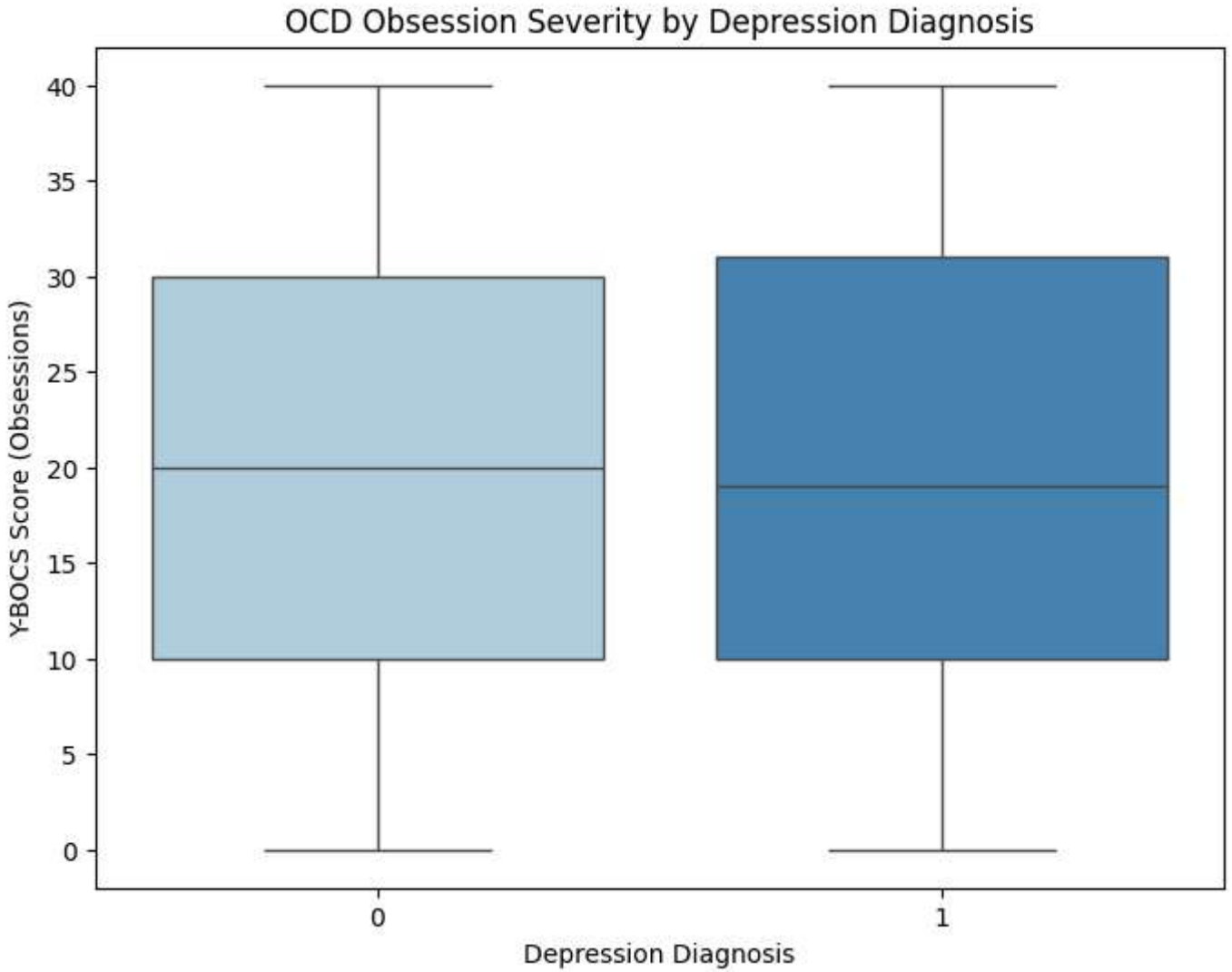
Observation: If males score significantly higher than females, this could suggest the need for gender-specific therapeutic approaches for OCD.

3. Comorbidities

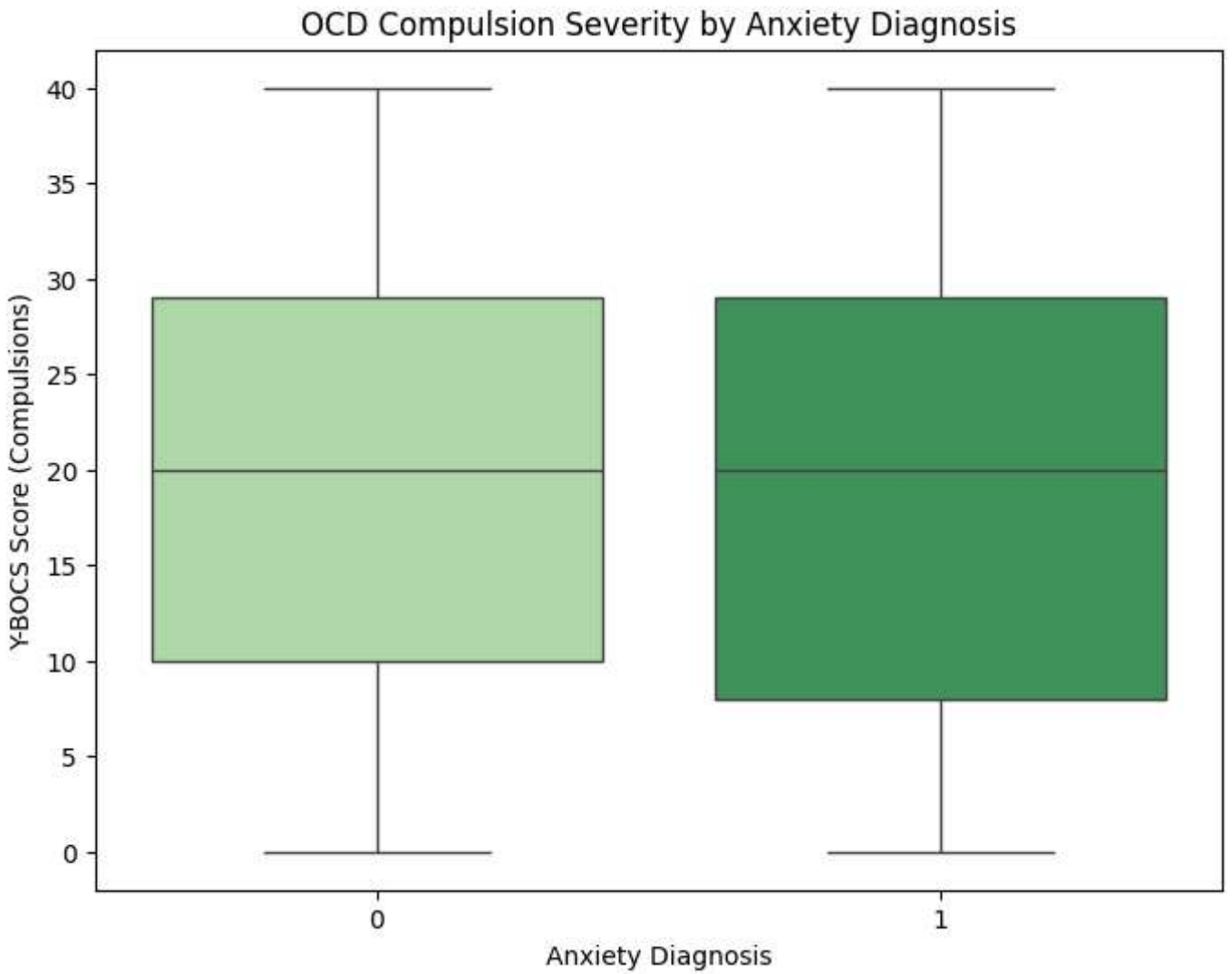
In [26]:
```
# Bar plot: Depression Diagnosis vs Y-BOCS Obsessions
plt.figure(figsize=(8,6))
sns.boxplot(x='Depression Diagnosis', y='Y-BOCS Score (Obsessions)', data=data, palette='Blues')
```

```
plt.title('OCD Obsession Severity by Depression Diagnosis')
plt.xlabel('Depression Diagnosis')
plt.ylabel('Y-BOCS Score (Obsessions)')
plt.show()

# Bar plot: Anxiety Diagnosis vs Y-BOCS Compulsions
plt.figure(figsize=(8,6))
sns.boxplot(x='Anxiety Diagnosis', y='Y-BOCS Score (Compulsions)', data=data, palette='Greens')
plt.title('OCD Compulsion Severity by Anxiety Diagnosis')
plt.xlabel('Anxiety Diagnosis')
plt.ylabel('Y-BOCS Score (Compulsions)')
plt.show()
```

```
C:\Users\Kumar\AppData\Local\Temp\ipykernel_44780\554129878.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `h
ue` and set `legend=False` for the same effect.

  sns.boxplot(x='Depression Diagnosis', y='Y-BOCS Score (Obsessions)', data=data, palette='Blues')
```




```
C:\Users\Kumar\AppData\Local\Temp\ipykernel_44780\554129878.py:11: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `h
ue` and set `legend=False` for the same effect.

  sns.boxplot(x='Anxiety Diagnosis', y='Y-BOCS Score (Compulsions)', data=data, palette='Greens')
```

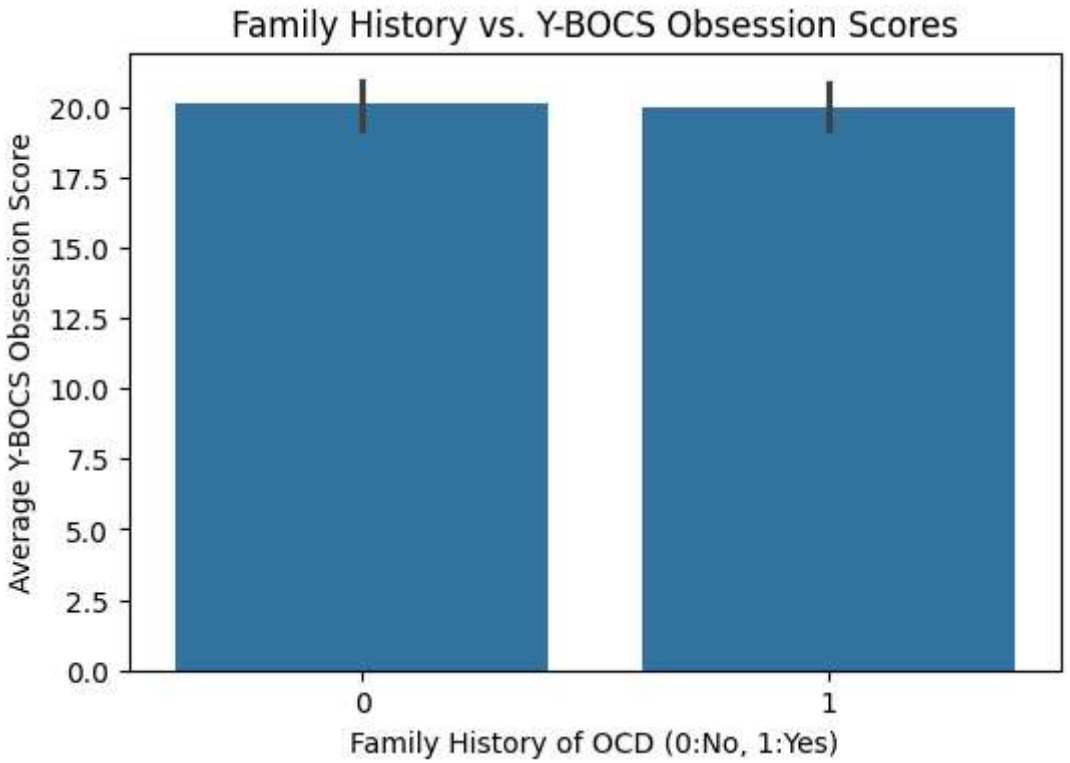


Insight: If comorbid patients (with depression or anxiety) show higher OCD scores, there is a clear case for integrated treatment plans that address both OCD and other mental health issues.

4. Family History Impact

In [27]:
```
plt.figure(figsize=(6, 4))
sns.barplot(x='Family History of OCD', y='Y-BOCS Score (Obsessions)', data=data)
plt.title('Family History vs. Y-BOCS Obsession Scores')
```

```
plt.xlabel('Family History of OCD (0:No, 1:Yes)')
plt.ylabel('Average Y-BOCS Obsession Score')
plt.show()
```



Observation: Patients with a family history of OCD tend to have higher obsession scores, indicating a possible genetic influence on OCD severity

6. Machine Learning Implementation

6.1 Preparing the Data

In [28]:
```
# Features and target variable
X = data.drop(['Patient ID', 'OCD Diagnosis Date', 'Y-BOCS Score (Obsessions)', 'Y-BOCS Score (Compulsions)'], axis=1
y = data['Y-BOCS Score (Obsessions)']

# Split into training and test sets
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Feature scaling
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.fit_transform(X_test)
```

6.2 Model Training and Evaluation

In [29]:
```
# Initialize the Random Forest Regressor
model = RandomForestRegressor(n_estimators=100, random_state=42)

# Train the model
model.fit(X_train_scaled, y_train)

# Predictions
y_pred = model.predict(X_test_scaled)

# Evaluation
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Mean Squared Error (MSE): {mse:.2f}")
print(f"R-squared (R2): {r2:.2f}")
```

```
Mean Squared Error (MSE): 154.12
R-squared (R2): -0.05
```

Observation: The model's performance metrics indicate how well it can predict OCD severity. A R-squared value closer to 1 implies a better fit

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: