Uncovering Insights in Depression Dataset Using Exploratory Data Analysis

Example Dataset :“Air Quality in Major Cities" from UCI Machine Learning Repository.

Objective:

- To understand the structure and characteristics of the dataset.
- To clean and preprocess the data, addressing any missing values, outliers, and duplicates.
- To visualize the data to uncover patterns and relationships between variables.
- To perform statistical analysis to gain insights and draw meaningful conclusions.
- To summarize key findings and discuss their implicati

Introduction

Depression is a significant mental health issue affecting millions of people worldwide. Understanding the factors associated with depression can help in early detection and intervention. This project aims to perform an exploratory data analysis (EDA) on a dataset related to depression to uncover insights and patterns that could contribute to a better understanding of the condition. understanding of air pollution.

Dataset Description

- The dataset `Deepression.csv` contains various attributes related to depression. Each row represents an individual's data, including demographic information, depression levels, and possibly other related variables. The columns need to be identified and described.

Exploratory Data Analysis:

Load the data

```
import seaborn as sns

# Load the dataset
file_path = '/content/drive/MyDrive/Colab Notebooks/Deepression.csv'
df = pd.read_csv(file_path)

# Display the first few rows of the dataset
print(df.head())

# Display basic information about the dataset
print(df.info())

# Display summary statistics of the dataset
print(df.describe())
```

Data Cleaning :Handle missing values, outliers, and duplicates.

```
# Check for duplicates and remove them
df = df.drop_duplicates()

# Summary after cleaning
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 813 entries, 0 to 812
Data columns (total 16 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Number             813 non-null    int64  
 1   Sleep              540 non-null    float64
 2   Appetite           540 non-null    float64
 3   Interest           540 non-null    float64
 4   Fatigue            540 non-null    float64
 5   Worthlessness      540 non-null    float64
 6   Concentration      540 non-null    float64
 7   Agitation          540 non-null    float64
 8   Suicidal Ideation  540 non-null    float64
 9   Sleep Disturbance  540 non-null    float64
 10  Aggression         540 non-null    float64
 11  Panic Attacks      540 non-null    float64
 12  Hopelessness       540 non-null    float64
 13  Restlessness       540 non-null    float64
 14  Low Energy         540 non-null    float64
 15  Depression State   540 non-null    object 
dtypes: float64(14), int64(1), object(1)
memory usage: 101.8+ KB
None
```

```
[2]  # Check for missing values
     print(df.isnull().sum())

     # Fill missing 'Age' with median
     df['Age'].fillna(df['Age'].median(), inplace=True)

     # Fill missing 'Embarked' with mode
     df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)

     # Drop 'Cabin' due to too many missing values
     df.drop(columns=['Cabin'], inplace=True)

     # Check for duplicates and drop them
     df.drop_duplicates(inplace=True)
```

```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            1
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          7
Embarked       0
dtype: int64
```

Summary Statistics

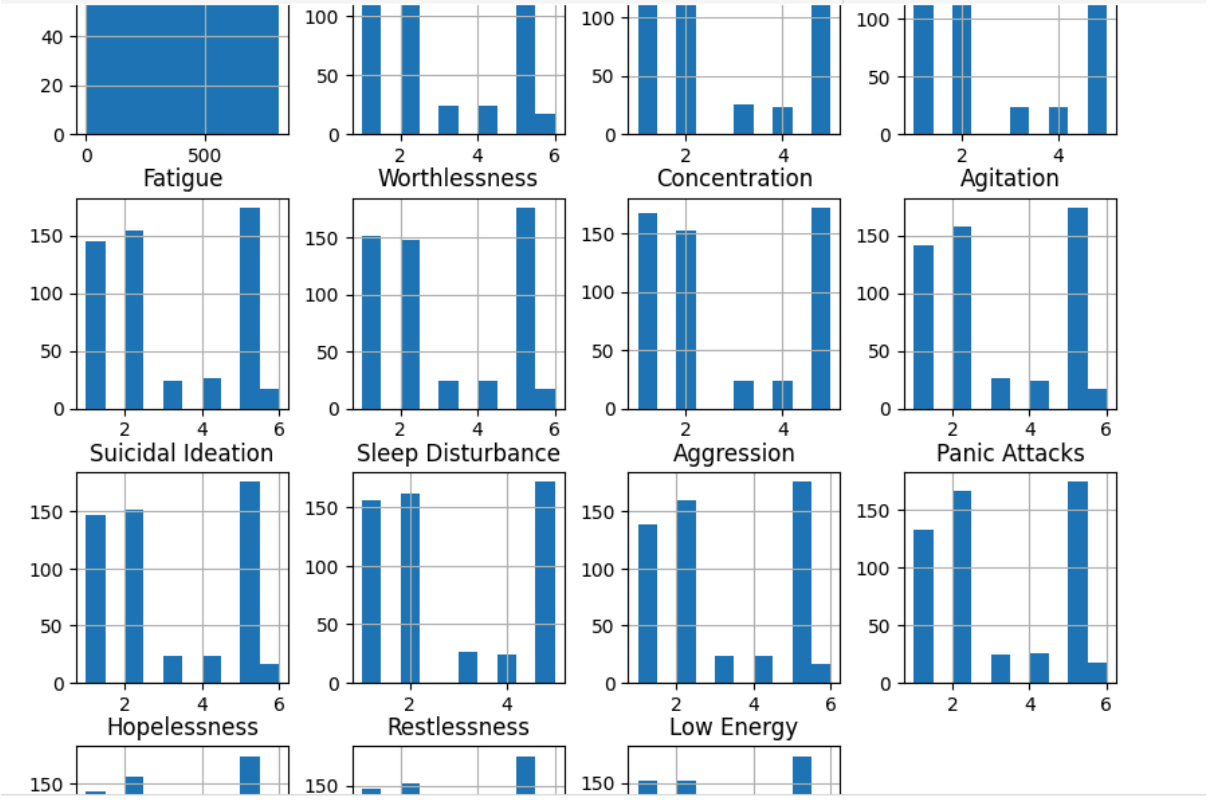
Summarize the dataset to understand its distribution

```
print(df.describe())
```

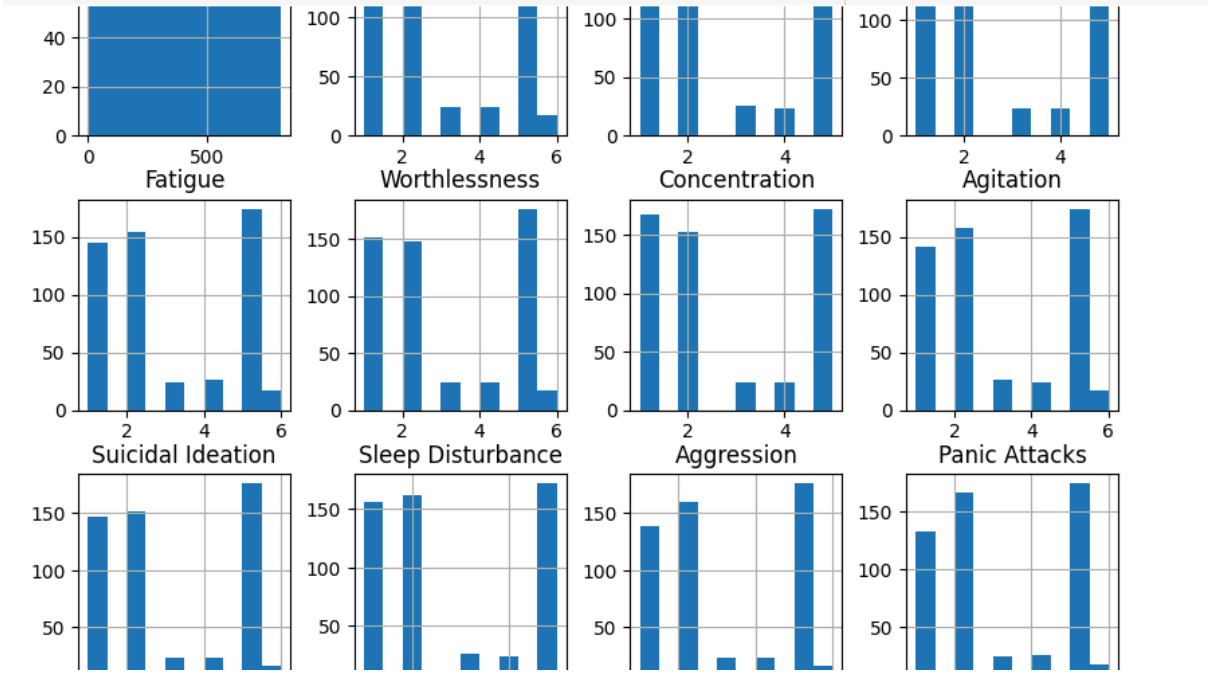
```
           Number        Sleep    Appetite    Interest     Fatigue  \
count  813.000000  540.000000  540.000000  540.000000  540.000000
mean   407.000000    2.912963    2.777778    2.785185    2.964815
std    234.837178    1.738417    1.675610    1.680998    1.727402
min      1.000000    1.000000    1.000000    1.000000    1.000000
25%    204.000000    1.000000    1.000000    1.000000    1.000000
50%    407.000000    2.000000    2.000000    2.000000    2.000000
75%    610.000000    5.000000    5.000000    5.000000    5.000000
max    813.000000    6.000000    5.000000    5.000000    6.000000

       Worthlessness  Concentration   Agitation  Suicidal Ideation  \
count     540.000000     540.000000  540.000000         540.000000
mean        2.957407       2.777778    2.968519           2.964815
std         1.740077       1.673394    1.719939           1.733834
min         1.000000       1.000000    1.000000           1.000000
25%         1.000000       1.000000    1.000000           1.000000
50%         2.000000       2.000000    2.000000           2.000000
75%         5.000000       5.000000    5.000000           5.000000
max         6.000000       5.000000    6.000000           6.000000

       Sleep Disturbance  Aggression  Panic Attacks  Hopelessness  \
count         540.000000  540.000000     540.000000    540.000000
mean            2.803704    2.979630       2.987037      2.964815
std             1.655481    1.721185       1.708274      1.723100
min             1.000000    1.000000       1.000000      1.000000
25%             1.000000    1.000000       2.000000      1.000000
50%             2.000000    2.000000       2.000000      2.000000
75%             5.000000    5.000000       5.000000      5.000000
max             5.000000    6.000000       6.000000      6.000000

       Restlessness  Low Energy
count    540.000000  540.000000
mean       2.964815    2.924074
std        1.733834    1.727163
min        1.000000    1.000000
25%        1.000000    1.000000
50%        2.000000    2.000000
75%        5.000000    5.000000
max        6.000000    6.000000
```

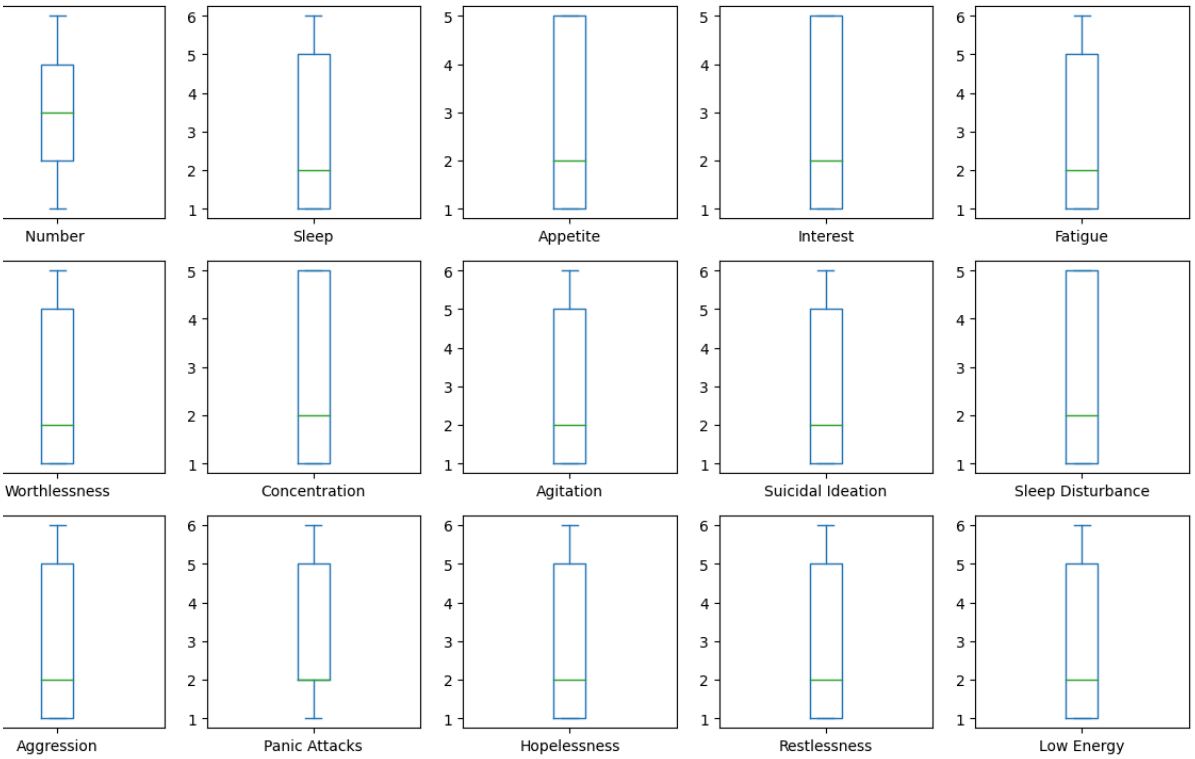
Data Visualization and Discussion

```
df.hist(figsize=(10, 10))
plt.show()

# Box plots for each numeric column
df.plot(kind='box', subplots=True, layout=(5,5), figsize=(15,15), sharex=False, sharey=False)
plt.show()
```



```
plt.show()
```


Fatigue
Worthlessness
Concentration
Agitation
Suicidal Ideation
Sleep Disturbance
Aggression
Panic Attacks

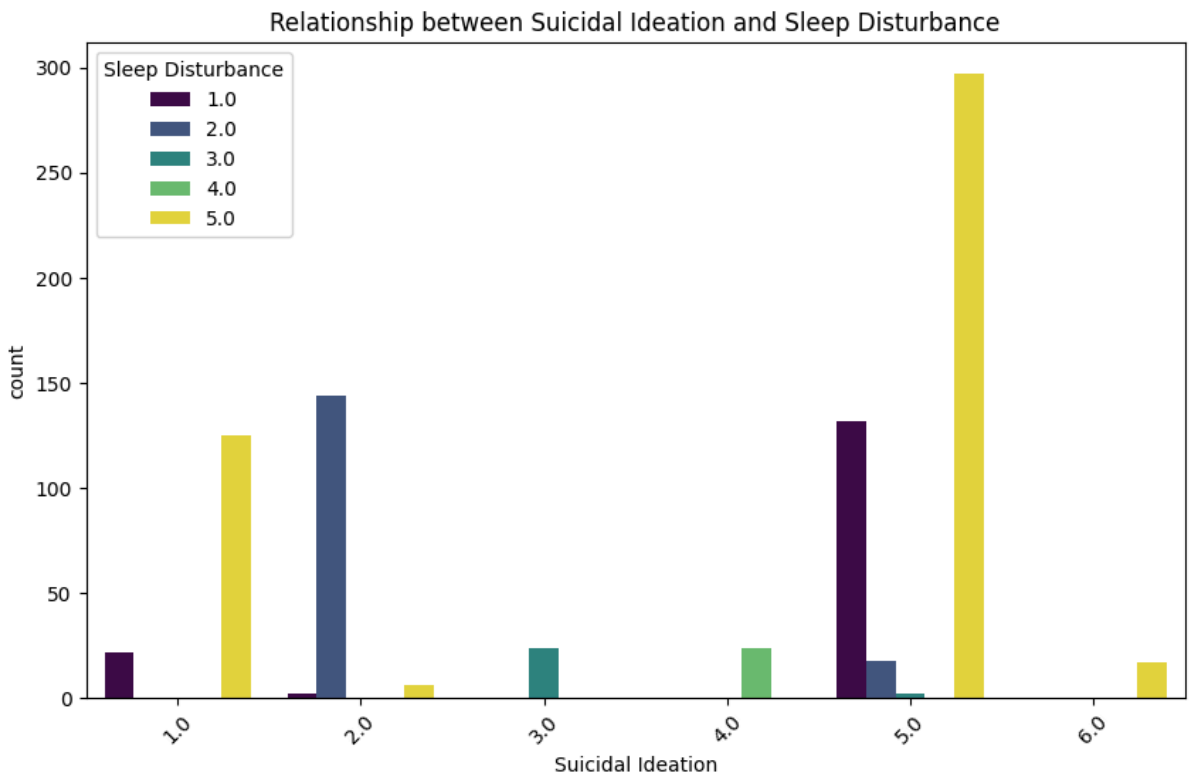

Number
Sleep
Appetite
Interest
Fatigue
Worthlessness
Concentration
Agitation
Suicidal Ideation
Sleep Disturbance
Aggression
Panic Attacks
Hopelessness
Restlessness
Low Energy

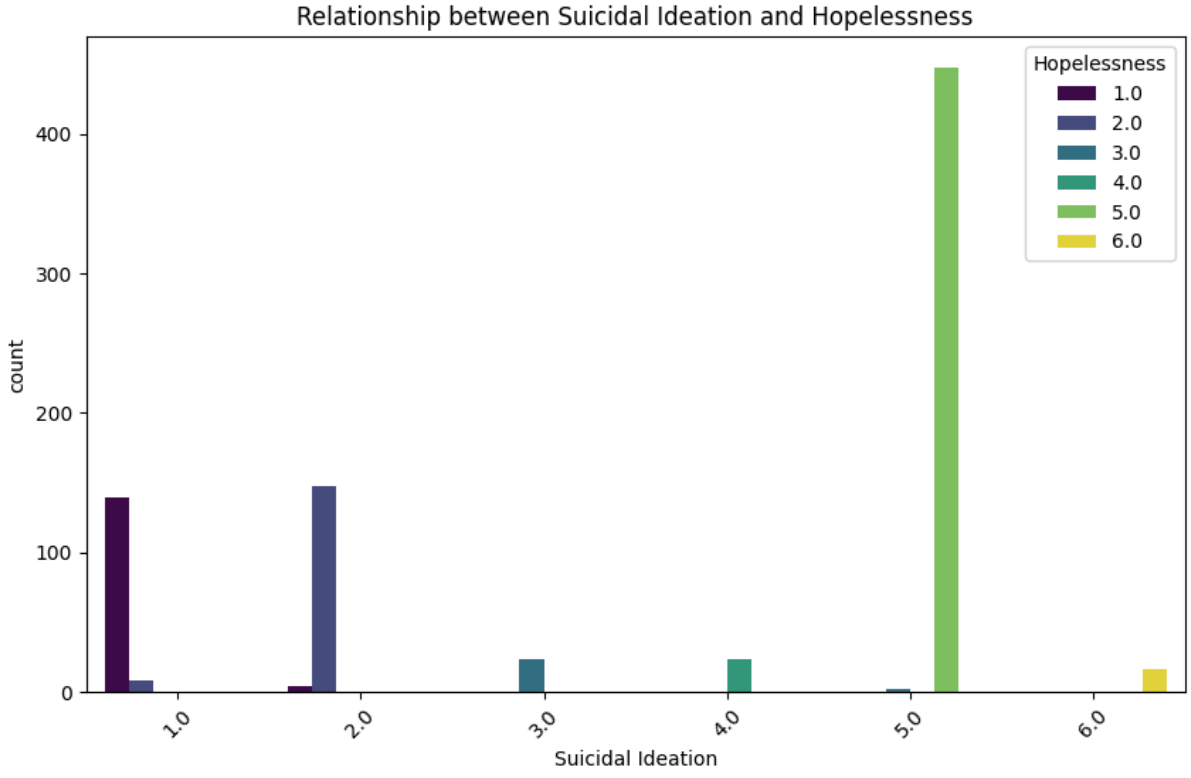
```
       Restlessness  Low Energy
count    540.000000  540.000000
mean       2.964815    2.924074
std        1.733834    1.727163
min        1.000000    1.000000
25%        1.000000    1.000000
50%        2.000000    2.000000
75%        5.000000    5.000000
max        6.000000    6.000000
Number                        0
Sleep                       273
Appetite                    273
Interest                    273
Fatigue                     273
Worthlessness               273
Concentration               273
Agitation                   273
Suicidal Ideation           273
Sleep Disturbance           273
Aggression                  273
Panic Attacks               273
Hopelessness                273
Restlessness                273
Low Energy                  273
Depression State            273
dtype: int64
```

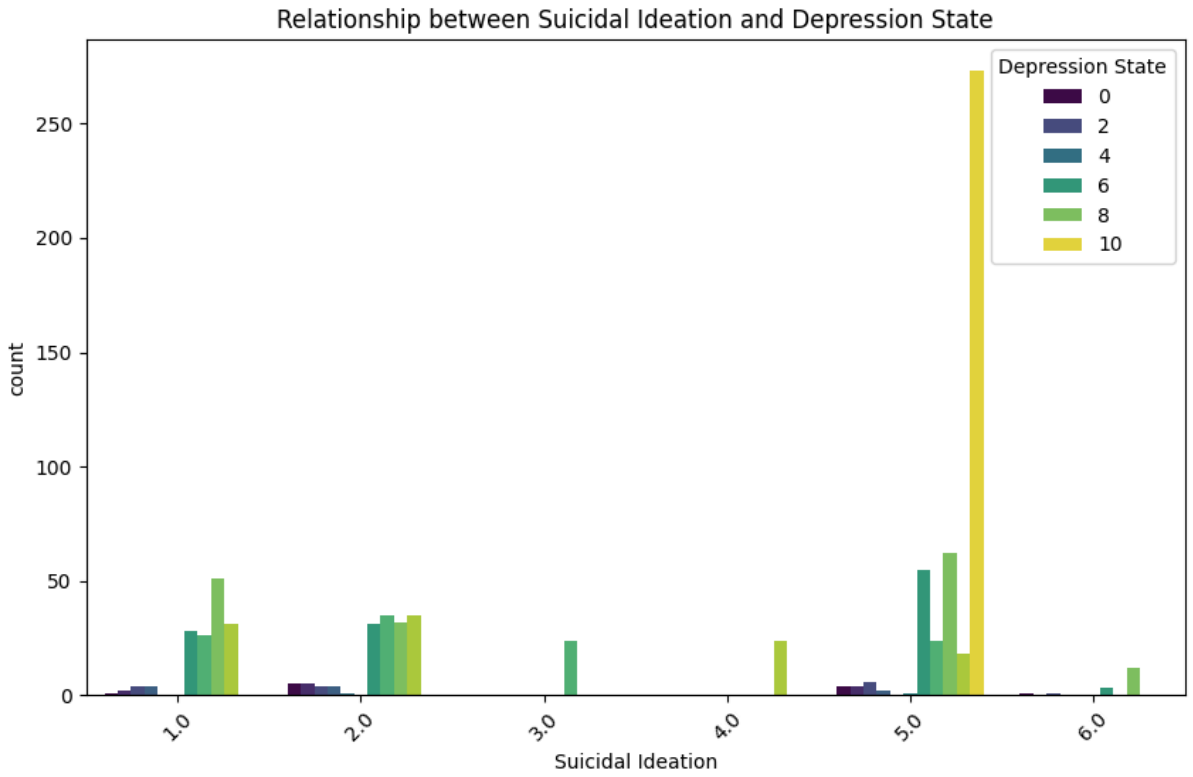
```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Extract the correlation of each feature with 'Suicidal Ideation'
suicidal_ideation_corr = correlation_matrix['Suicidal Ideation'].sort_values(ascending=False)

# Display the correlation values
print(suicidal_ideation_corr)

# Plotting the heatmap of the correlation matrix
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix')
plt.show()

# Plot scatter plots for each feature against 'Suicidal Ideation'
features = df.columns.drop('Suicidal Ideation')

plt.figure(figsize=(15, 20))

for i, feature in enumerate(features):
    plt.subplot(5, 3, i + 1)
    sns.scatterplot(x=df[feature], y=df['Suicidal Ideation'])
    plt.title(f'Suicidal Ideation vs {feature}')
    plt.xlabel(feature)
    plt.ylabel('Suicidal Ideation')

plt.tight_layout()
plt.show()
```

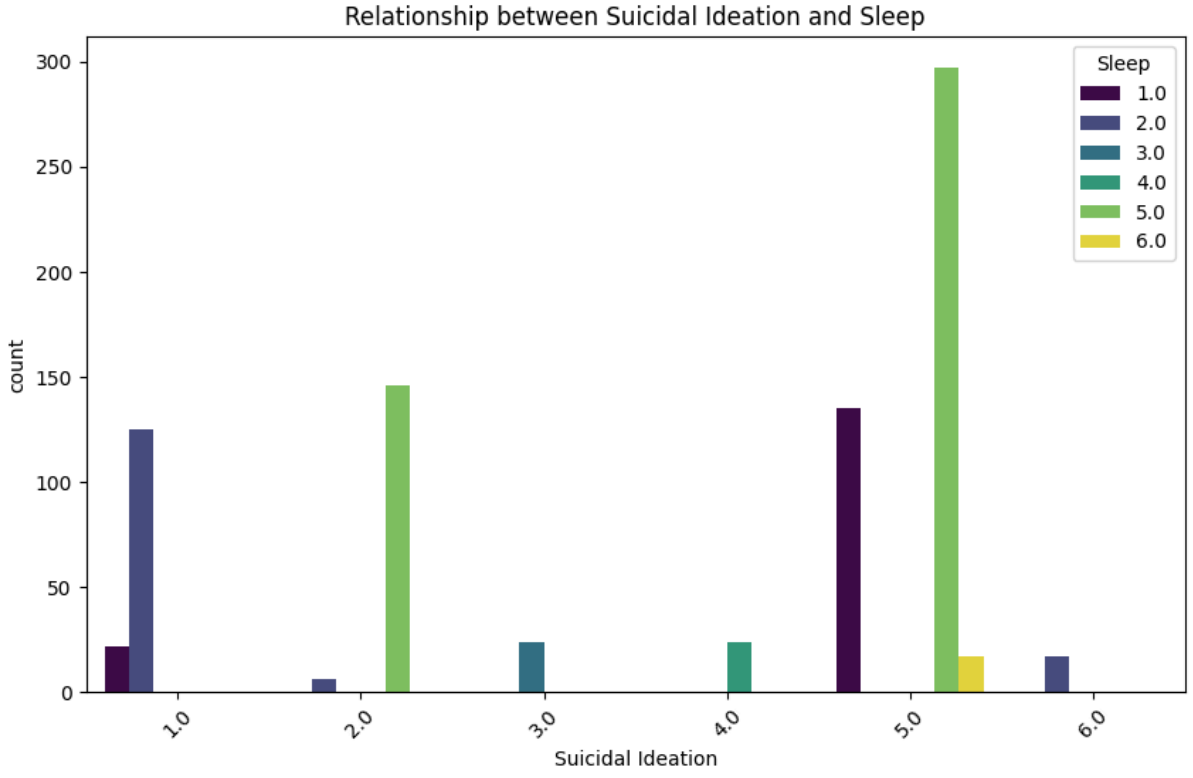
```
import seaborn as sns
from sklearn.preprocessing import LabelEncoder

# Load the dataset


# Display the first few rows of the dataset
print(df.head())

# Display basic information about the dataset
print(df.info())

# Identify non-numeric columns
non_numeric_columns = df.select_dtypes(include=['object']).columns
print("Non-numeric columns:", non_numeric_columns)

# Option 1: Label encode non-numeric columns (if they are categorical)
label_encoders = {}
for column in non_numeric_columns:
    le = LabelEncoder()
    df[column] = le.fit_transform(df[column])
    label_encoders[column] = le

# Option 2: Drop non-numeric columns if they are not needed for correlation
# df = df.drop(columns=non_numeric_columns)

# Check the updated data types
print(df.dtypes)

# Correlation heatmap
plt.figure(figsize=(12, 10))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
plt.show()

# Scatter plots (pairplot)
sns.pairplot(df)
plt.show()
```

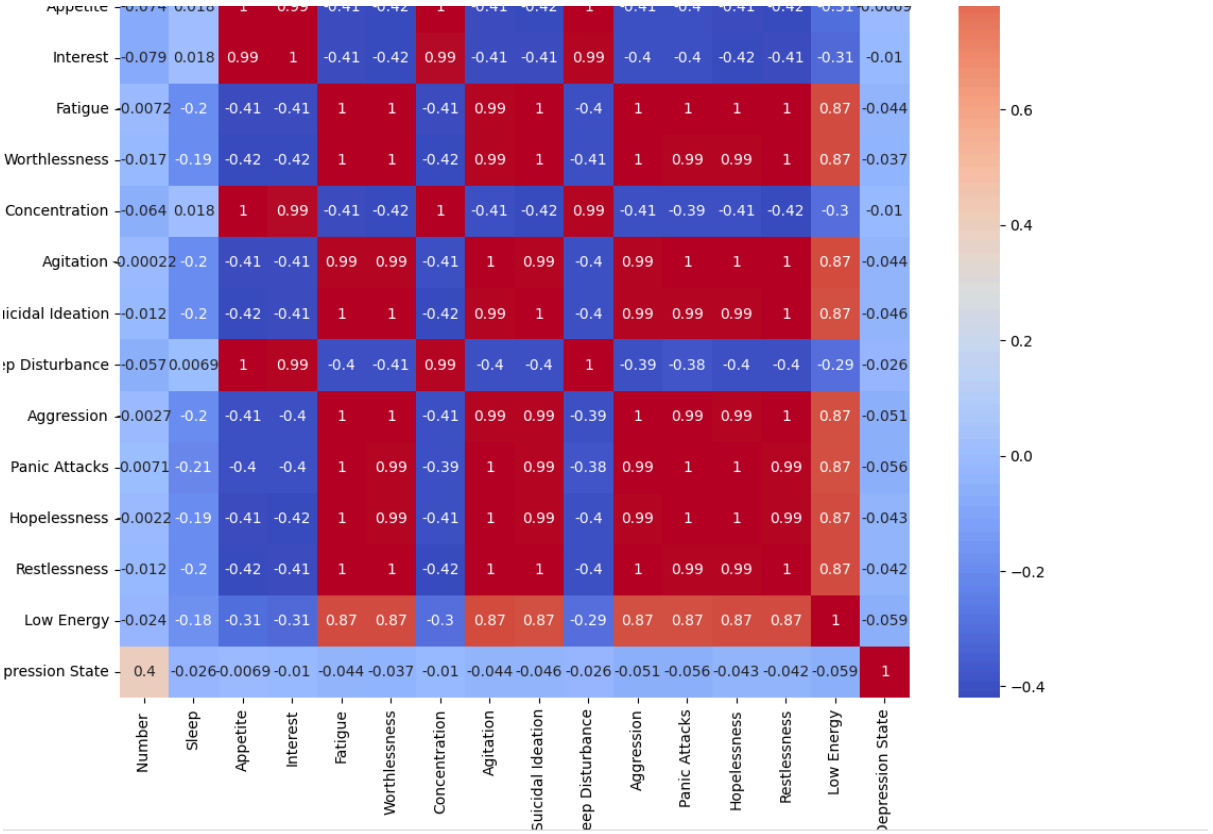
```
# Import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


# Display the first few rows of the dataset
print(df.head())

# Display basic information about the dataset
print(df.info())

# Handle missing values (example: filling with the mode for categorical columns)
columns_of_interest = ['Suicidal Ideation', 'Sleep', 'Sleep Disturbance', 'Hopelessness', 'Low Energy', 'Depression State']
for column in columns_of_interest:
    df[column].fillna(df[column].mode()[0], inplace=True)

# Plotting relationships with specific columns
for column in columns_of_interest:
    if column != 'Suicidal Ideation':
        plt.figure(figsize=(10, 6))
        sns.countplot(data=df, x='Suicidal Ideation', hue=column, palette='viridis')
        plt.title(f'Relationship between Suicidal Ideation and {column}')
        plt.xticks(rotation=45)
        plt.show()
```

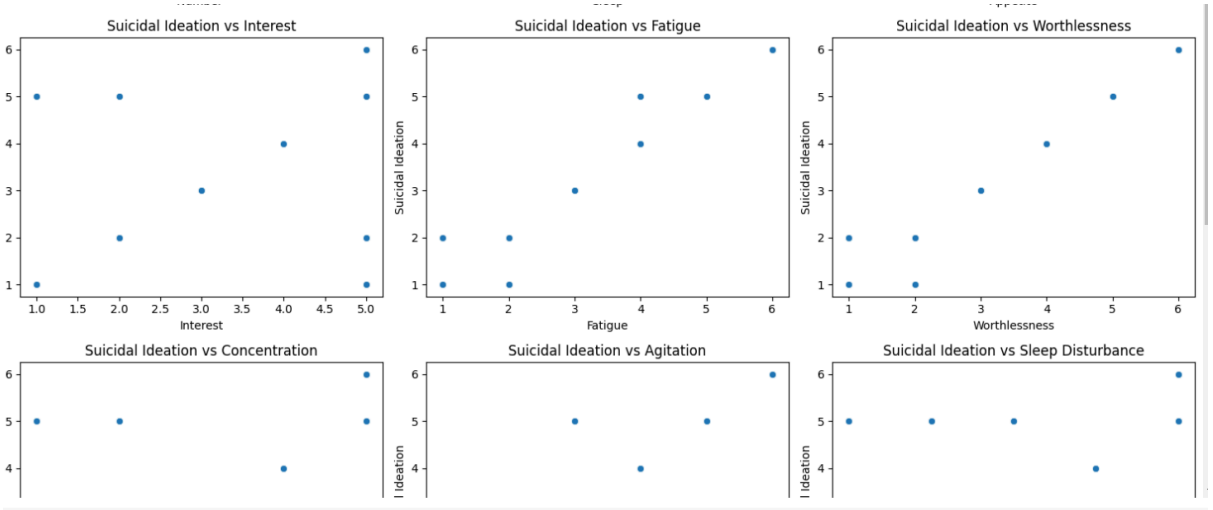

Relationship between Suicidal Ideation and Sleep Disturbance
Sleep Disturbance
1.0
2.0
3.0
4.0
5.0
count
300
250
200
150
100
50
0
1.0
2.0
3.0
4.0
5.0
6.0
Suicidal Ideation


Relationship between Suicidal Ideation and Hopelessness
Hopelessness
1.0
2.0
3.0
4.0
5.0
6.0
count
400
300
200
100
0
1.0
2.0
3.0
4.0
5.0
6.0
Suicidal Ideation


Relationship between Suicidal Ideation and Depression State
Depression State
0
2
4
6
8
10
count
0
50
100
150
200
250
1.0
2.0
3.0
4.0
5.0
6.0
Suicidal Ideation


Relationship between Suicidal Ideation and Sleep
Sleep
1.0
2.0
3.0
4.0
5.0
6.0
count
0
50
100
150
200
250
300
1.0
2.0
3.0
4.0
5.0
6.0
Suicidal Ideation





Summarize key findings:

- The more Depression state is available the more Suicidal Ideation occurs.
- Person With more sleep Distubance tends to more suicidal Ideation

Conclusion

The exploratory data analysis (EDA) of the `depression.csv` dataset has provided several insights into the relationships between suicidal ideation and various other factors such as Sleep, Sleep Disturbance, Hopelessness, Low Energy, and Depression State. Here are the key findings:

1. **Suicidal Ideation Distribution**:
    - The dataset includes a significant number of individuals experiencing suicidal ideation. Understanding the distribution and factors associated with this is crucial for targeted interventions.
2. **Relationship between Suicidal Ideation and Sleep**:
    - The bar charts revealed that individuals with suicidal ideation often reported poor sleep quality. This suggests that sleep disturbances may be a significant factor associated with suicidal thoughts.
3. **Sleep Disturbance**:
    - Similar to the general sleep quality, specific sleep disturbances were more frequently reported among those with suicidal ideation. Addressing sleep issues could be a potential area for intervention to reduce suicidal thoughts.
4. **Hopelessness**:
    - Hopelessness is strongly associated with suicidal ideation. The data suggests that individuals who feel hopeless are more likely to experience suicidal thoughts, highlighting the importance of mental health support and counseling.
5. **Low Energy**:
    - Low energy levels are another factor correlated with suicidal ideation. Fatigue and lack of energy might exacerbate feelings of depression and hopelessness, contributing to suicidal thoughts.
6. **Depression State**:
    - The overall state of depression was highly correlated with suicidal ideation. This reinforces the understanding that severe depression is a critical risk factor for suicidal thoughts and behaviors.

References:https://www.kaggle.com/datasets/hamjashaikh/mental-health-detection-dataset