# Load and Inspect the Data

```
pwd

'C:\\Users\\Administrator'

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")

df=pd.read_csv('C:/Users/Administrator/Documents/
student_depression_dataset1.csv',header=0)

df.head()
```

```
   id  Gender  Age           City Profession  Academic Pressure  CGPA
\
0   2    Male   33  Visakhapatnam    Student                  5  8.97

1   8  Female   24      Bangalore    Student                  2  5.90

2  26    Male   31       Srinagar    Student                  3  7.03

3  30  Female   28       Varanasi    Student                  3  5.59

4  32  Female   25         Jaipur    Student                  4  8.13


   Study Satisfaction       Sleep Duration Dietary Habits   Degree  \
0                   2          '5-6 hours'        Healthy  B.Pharm
1                   5          '5-6 hours'       Moderate      BSc
2                   5  'Less than 5 hours'        Healthy       BA
3                   2          '7-8 hours'       Moderate      BCA
4                   3          '5-6 hours'       Moderate   M.Tech

  Have you ever had suicidal thoughts ?  Work/Study Hours Financial
Stress  \
0                                    Yes                3
1
1                                     No                3
2
2                                     No                9
1
3                                    Yes                4
5
4                                    Yes                1
```

```
1

  Family History of Mental Illness  Depression
0                               No           1
1                              Yes           0
2                              Yes           0
3                              Yes           1
4                               No           0
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27901 entries, 0 to 27900
Data columns (total 16 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   id                                     27901 non-null  int64
 1   Gender                                 27901 non-null  object
 2   Age                                    27901 non-null  int64
 3   City                                   27901 non-null  object
 4   Profession                             27901 non-null  object
 5   Academic Pressure                      27901 non-null  int64
 6   CGPA                                   27901 non-null  float64
 7   Study Satisfaction                     27901 non-null  int64
 8   Sleep Duration                         27901 non-null  object
 9   Dietary Habits                         27901 non-null  object
 10  Degree                                 27901 non-null  object
 11  Have you ever had suicidal thoughts ?  27901 non-null  object
 12  Work/Study Hours                       27901 non-null  int64
 13  Financial Stress                       27901 non-null  object
 14  Family History of Mental Illness       27901 non-null  object
 15  Depression                             27901 non-null  int64
dtypes: float64(1), int64(6), object(9)
memory usage: 3.4+ MB
```

```
df.describe()
```

```
                  id           Age  Academic Pressure          CGPA  \
count   27901.000000  27901.000000       27901.000000  27901.000000
mean    70442.149421     25.822300           3.141214      7.656104
std     40641.175216      4.905687           1.381465      1.470707
min         2.000000     18.000000           0.000000      0.000000
25%     35039.000000     21.000000           2.000000      6.290000
50%     70684.000000     25.000000           3.000000      7.770000
75%    105818.000000     30.000000           4.000000      8.920000
max    140699.000000     59.000000           5.000000     10.000000

       Study Satisfaction  Work/Study Hours    Depression
count        27901.000000      27901.000000  27901.000000
mean             2.943837          7.156984      0.585499
```

```
std              1.361148           3.707642      0.492645
min              0.000000           0.000000      0.000000
25%              2.000000           4.000000      0.000000
50%              3.000000           8.000000      1.000000
75%              4.000000          10.000000      1.000000
max              5.000000          12.000000      1.000000
```

# Data Cleaning

a. Check for Missing Values

```
df.isnull().sum()

id                                       0
Gender                                   0
Age                                      0
City                                     0
Profession                               0
Academic Pressure                        0
CGPA                                     0
Study Satisfaction                       0
Sleep Duration                           0
Dietary Habits                           0
Degree                                   0
Have you ever had suicidal thoughts ?    0
Work/Study Hours                         0
Financial Stress                         0
Family History of Mental Illness         0
Depression                               0
dtype: int64
```

b. Clean Column Names

```
df.columns = df.columns.str.strip().str.replace(" ", "_").str.lower()

df.columns

Index(['id', 'gender', 'age', 'city', 'profession',
'academic_pressure',
       'cgpa', 'study_satisfaction', 'sleep_duration',
'dietary_habits',
       'degree', 'have_you_ever_had_suicidal_thoughts_?',
'work/study_hours',
       'financial_stress', 'family_history_of_mental_illness',
'depression'],
      dtype='object')
```

c. Categorical Data Normalization

```
df['city'].value_counts()
```

```
city
Kalyan                  1570
Srinagar                1372
Hyderabad               1340
Vasai-Virar             1290
Lucknow                 1155
Thane                   1139
Ludhiana                1111
Agra                    1094
Surat                   1078
Kolkata                 1066
Jaipur                  1036
Patna                   1007
Visakhapatnam            969
Pune                     968
Ahmedabad                951
Bhopal                   934
Chennai                  885
Meerut                   825
Rajkot                   816
Delhi                    768
Bangalore                767
Ghaziabad                745
Mumbai                   699
Vadodara                 694
Varanasi                 685
Nagpur                   651
Indore                   643
Kanpur                   609
Nashik                   547
Faridabad                461
Saanvi                     2
Bhavna                     2
City                       2
Harsha                     2
Kibara                     1
Nandini                    1
Nalini                     1
Mihir                      1
Nalyan                     1
M.Com                      1
ME                         1
Rashi                      1
Gaurav                     1
Reyansh                    1
Harsh                      1
```

```
Vaanya                           1
Mira                             1
'Less than 5 Kalyan'             1
3                                1
'Less Delhi'                     1
M.Tech                           1
Khaziabad                        1
Name: count, dtype: int64
```

```
df['gender'].value_counts()
```

```
gender
Male      15547
Female    12354
Name: count, dtype: int64
```

```
df['degree'].value_counts()
```

```
degree
'Class 12'    6080
B.Ed          1867
B.Com         1506
B.Arch        1478
BCA           1433
MSc           1190
B.Tech        1152
MCA           1044
M.Tech        1022
BHM            925
BSc            888
M.Ed           821
B.Pharm        810
M.Com          734
MBBS           696
BBA            696
LLB            671
BE             613
BA             600
M.Pharm        582
MD             572
MBA            562
MA             544
PhD            522
LLM            482
MHM            191
ME             185
Others          35
Name: count, dtype: int64
```

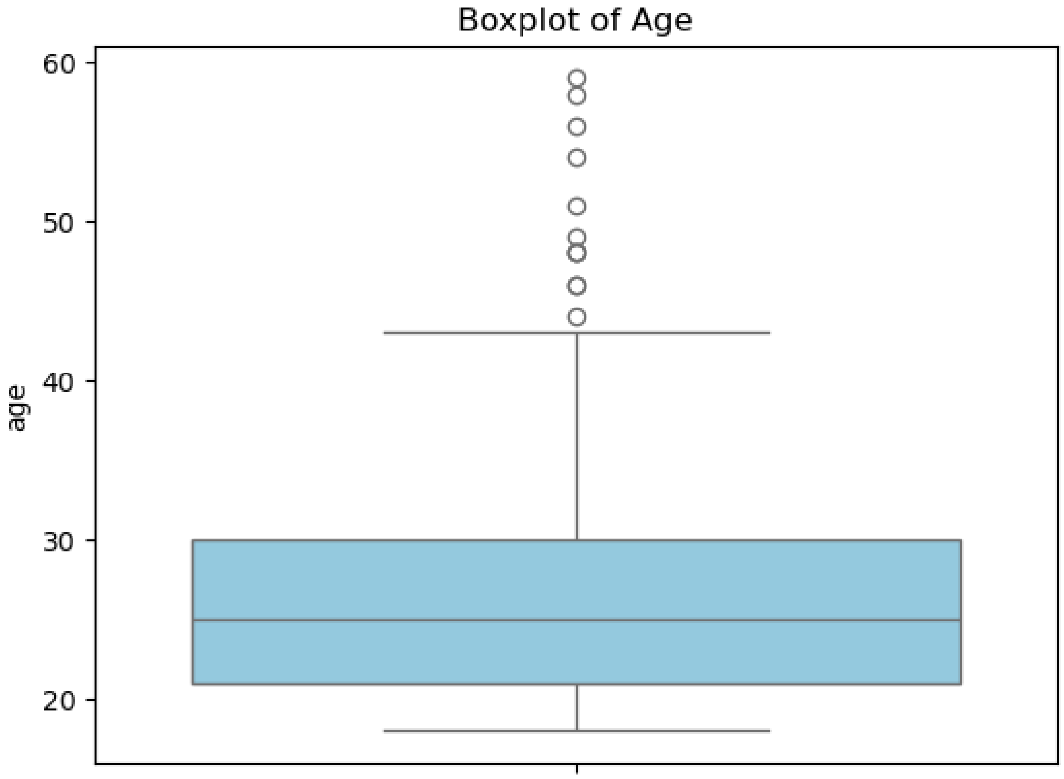
# Exploratory Data Analysis (EDA)

a. Outliers detection

```
print(df['age'].describe())
```

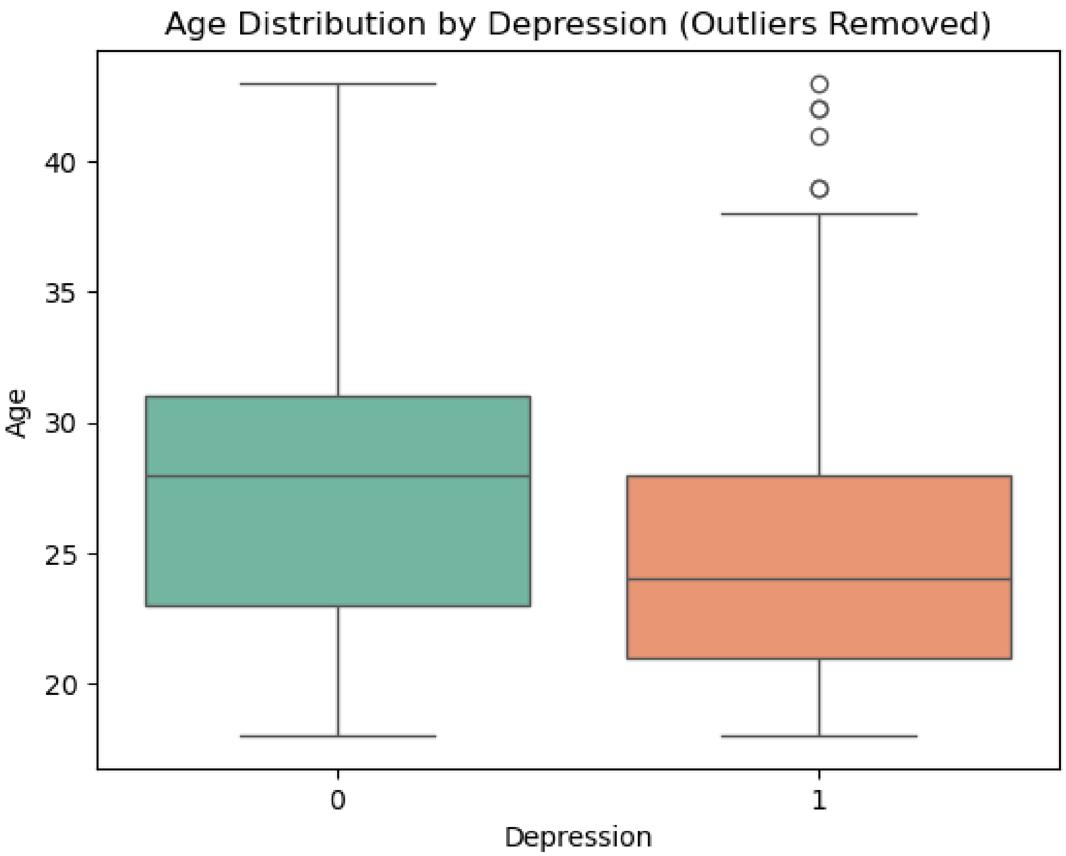
```
count    27901.000000
mean        25.822300
std          4.905687
min         18.000000
25%         21.000000
50%         25.000000
75%         30.000000
max         59.000000
Name: age, dtype: float64
```

```
sns.boxplot(y=df['age'], color='skyblue')
plt.title("Boxplot of Age")
plt.show()
```



```
Q1 = df['age'].quantile(0.25)
Q3 = df['age'].quantile(0.75)
IQR = Q3 - Q1

lower_bound = Q1 - 1.5 * IQR
```

```
upper_bound = Q3 + 1.5 * IQR

df_no_outliers = df[(df['age'] >= lower_bound) & (df['age'] <=
upper_bound)]

sns.boxplot(data=df_no_outliers, x='depression', y='age',
palette='Set2')
plt.title('Age Distribution by Depression (Outliers Removed)')
plt.xlabel('Depression')
plt.ylabel('Age')
plt.show()
```



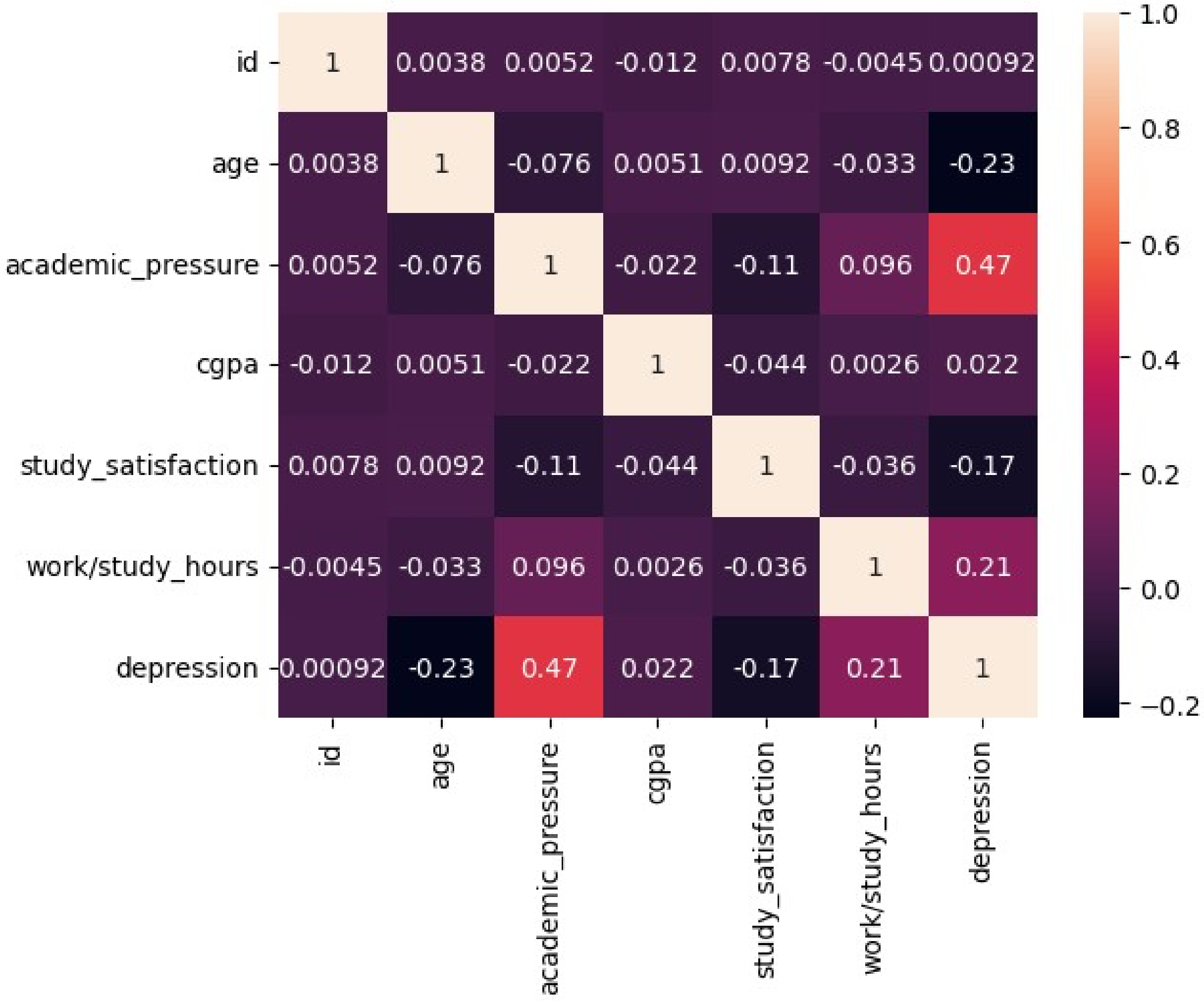
b.Corelation Analysis

```
df.describe()

                   id           age  academic_pressure          cgpa  \
count    27901.000000  27901.000000       27901.000000  27901.000000
mean     70442.149421     25.822300           3.141214      7.656104
std      40641.175216      4.905687           1.381465      1.470707
min          2.000000     18.000000           0.000000      0.000000
25%      35039.000000     21.000000           2.000000      6.290000
50%      70684.000000     25.000000           3.000000      7.770000
75%     105818.000000     30.000000           4.000000      8.920000
```

```
max    140699.000000     59.000000           5.000000     10.000000

       study_satisfaction  work/study_hours    depression
count        27901.000000      27901.000000  27901.000000
mean             2.943837          7.156984      0.585499
std              1.361148          3.707642      0.492645
min              0.000000          0.000000      0.000000
25%              2.000000          4.000000      0.000000
50%              3.000000          8.000000      1.000000
75%              4.000000         10.000000      1.000000
max              5.000000         12.000000      1.000000
```

```
sns.heatmap(df.corr(numeric_only=True), annot=True)
```

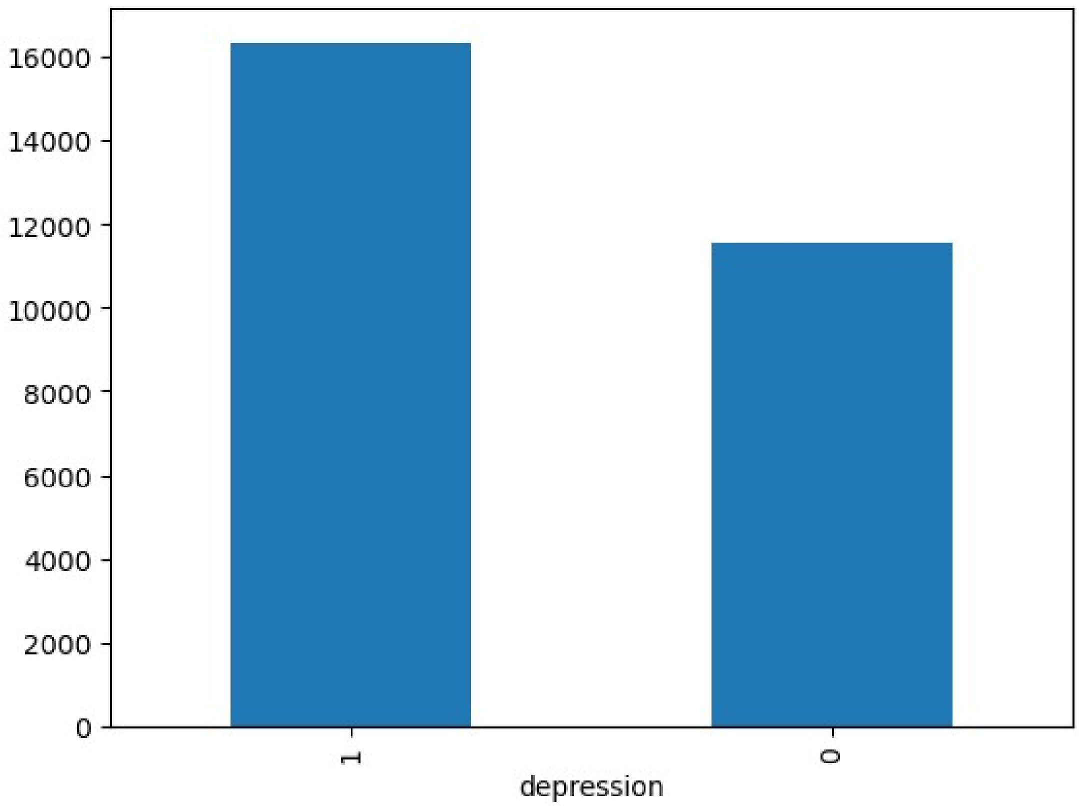
```
<Axes: >
```



c. univariate analysis

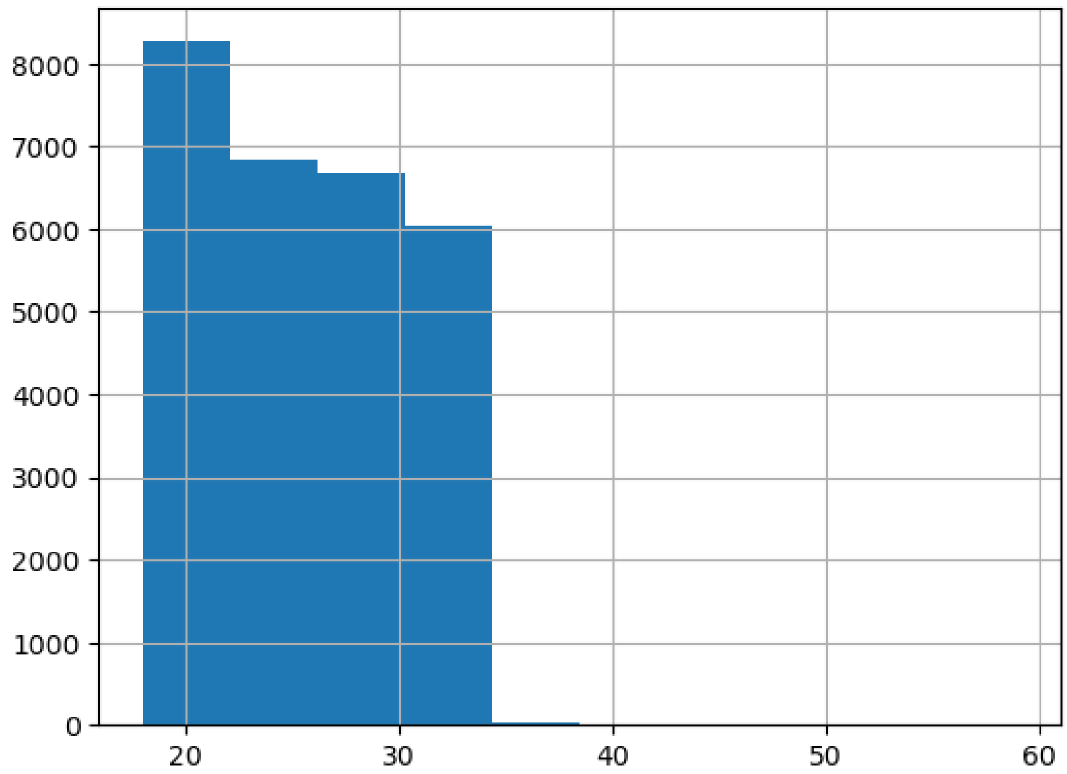
```
df['depression'].value_counts().plot(kind='bar')
```

```
<Axes: xlabel='depression'>
```



```
df['age'].hist(bins=10)
```

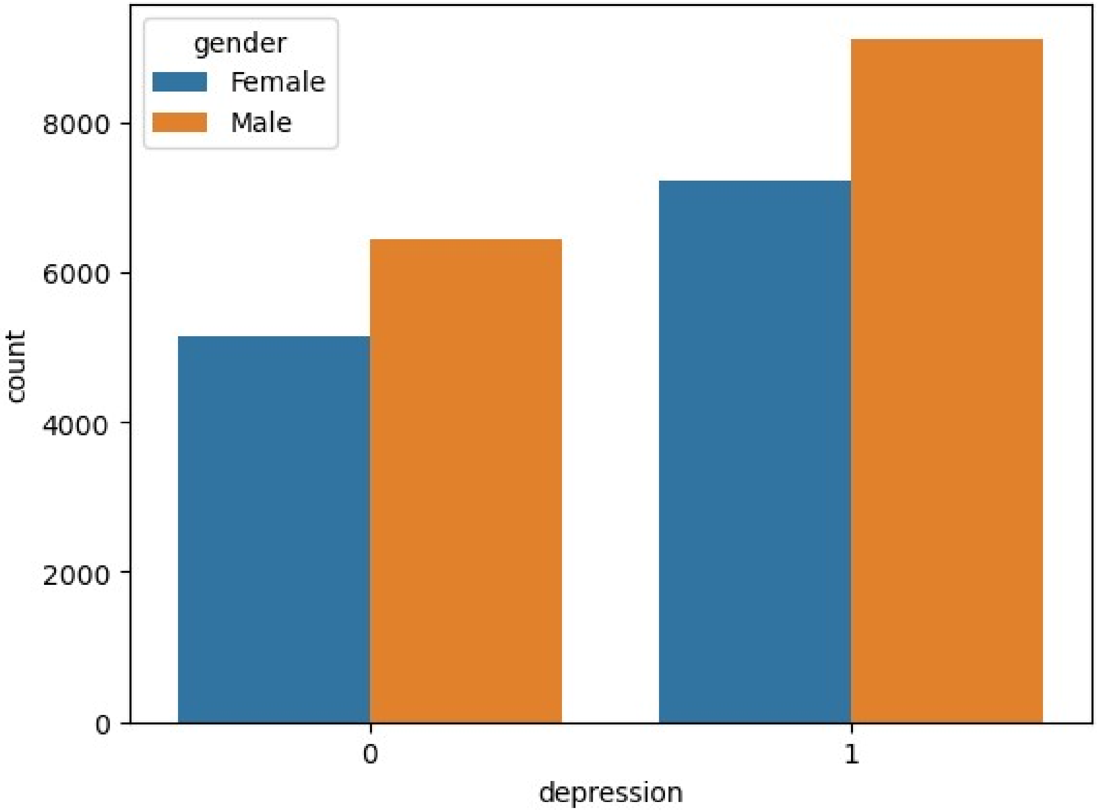
```
<Axes: >
```



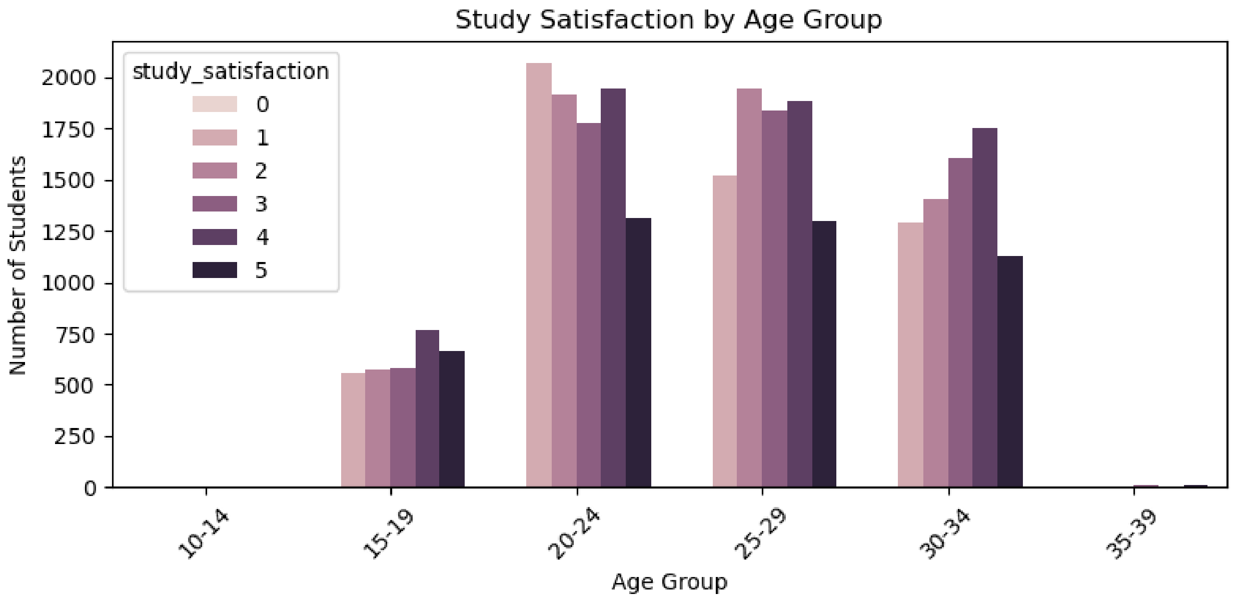
d. Bivariate analysis

```
sns.countplot(x='depression', hue='gender', data=df)
```

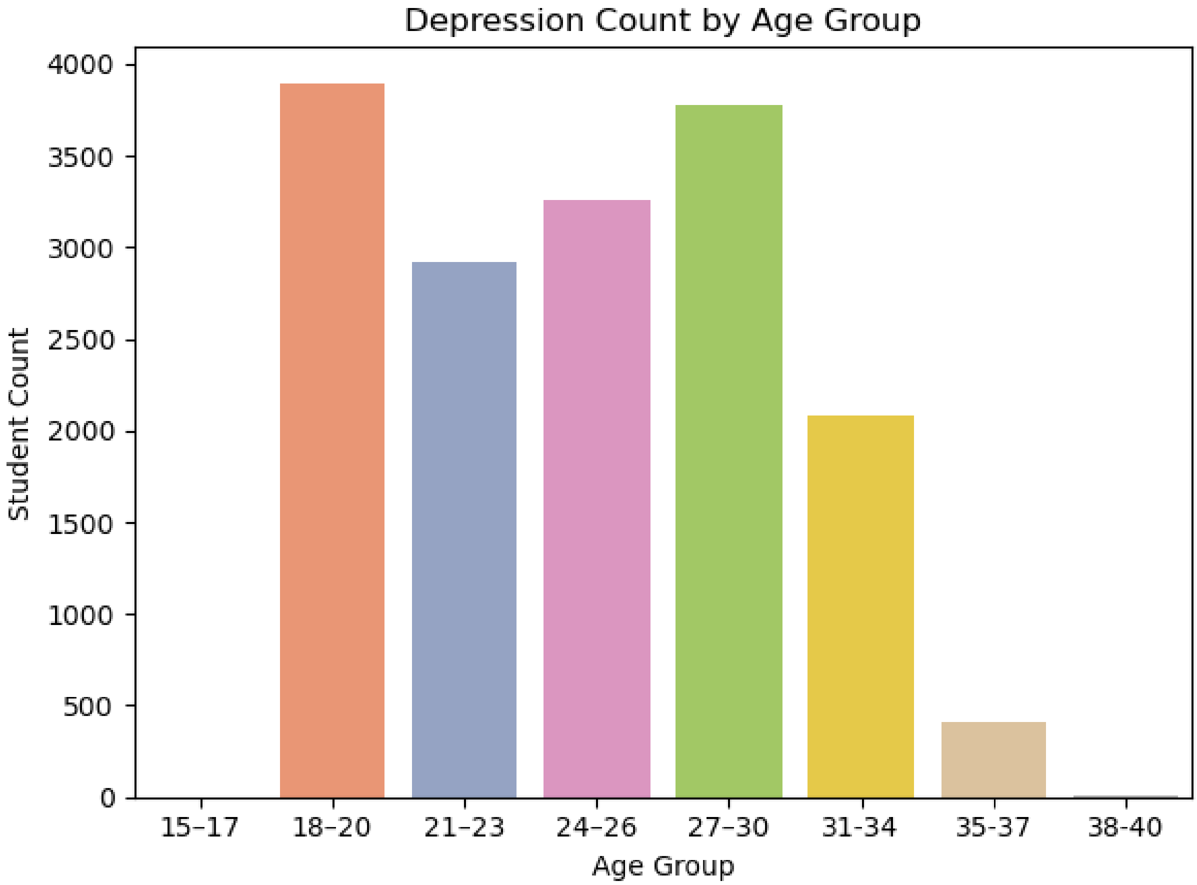
```
<Axes: xlabel='depression', ylabel='count'>
```



```
bins = list(range(10, 45, 5))
labels = [f'{i}-{i+4}' for i in bins[:-1]]

df['age_group'] = pd.cut(df['age'], bins=bins, labels=labels,
right=False)

plt.figure(figsize=(8,4))
sns.countplot(x='age_group', hue='study_satisfaction', data=df)
plt.title('Study Satisfaction by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Number of Students')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

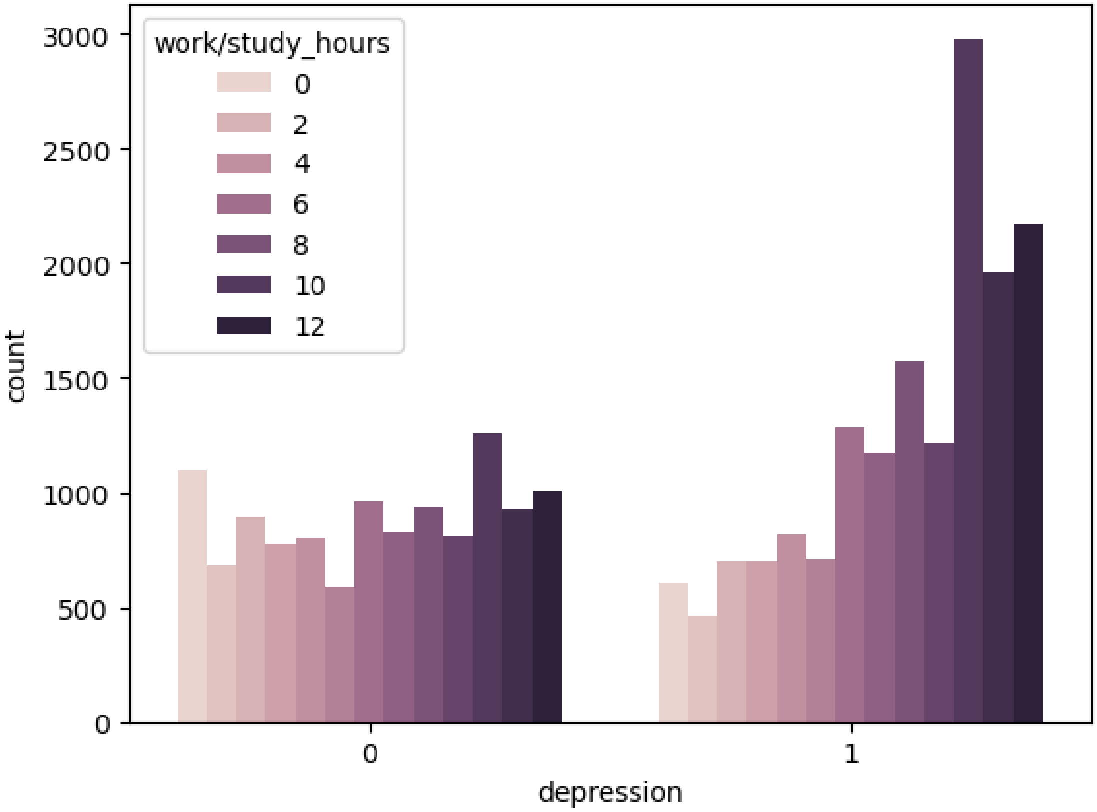


```
sns.countplot(
    x=pd.cut(df[df['depression'] == 1]['age'], bins=[14, 17, 20, 23,
26, 30, 33, 36, 40], labels=['15–17',
    '18–20', '21–23', '24–26', '27–30','31-34', '35-37', '38-40']),

    palette='Set2'
)

plt.title('Depression Count by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Student Count')
plt.tight_layout()
plt.show()
```



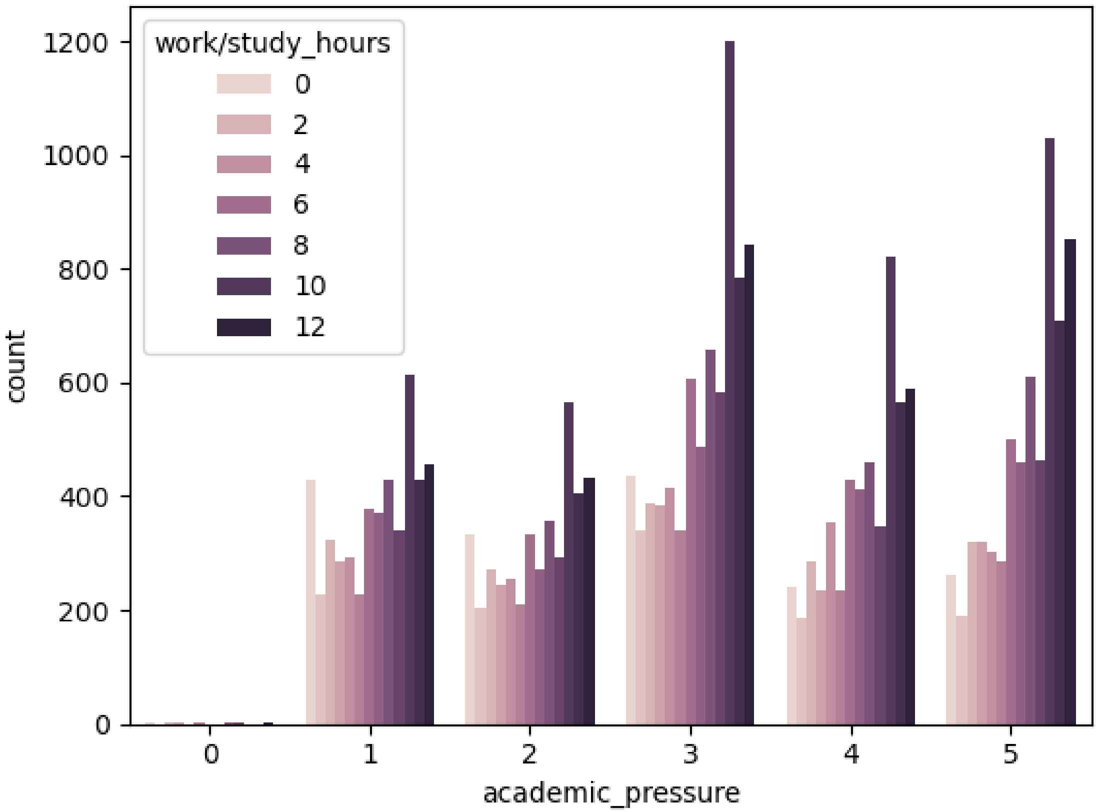
```
sns.countplot(x='depression', hue='work/study_hours', data=df)
```

```
<Axes: xlabel='depression', ylabel='count'>
```



```
sns.countplot(x='academic_pressure', hue='work/study_hours', data=df)
```

```
<Axes: xlabel='academic_pressure', ylabel='count'>
```

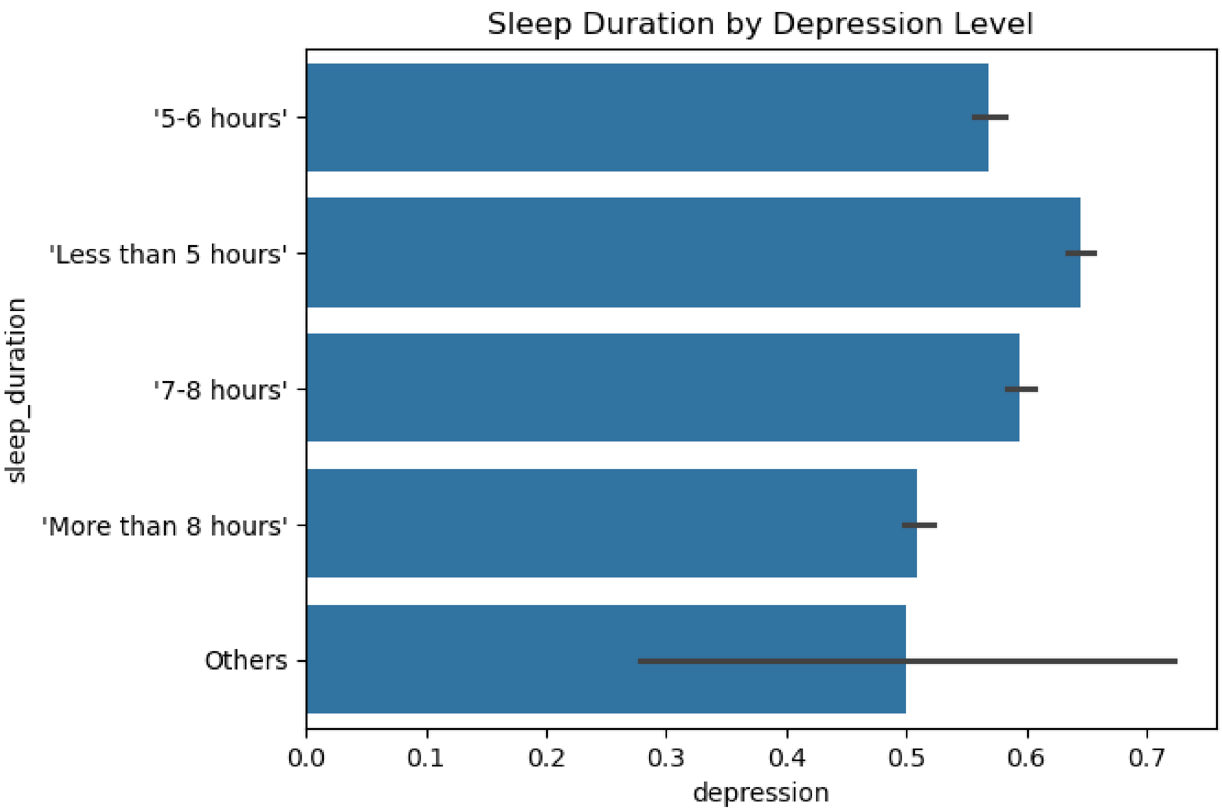


```
pd.crosstab(df['academic_pressure'], df['depression'],
normalize='index')
```

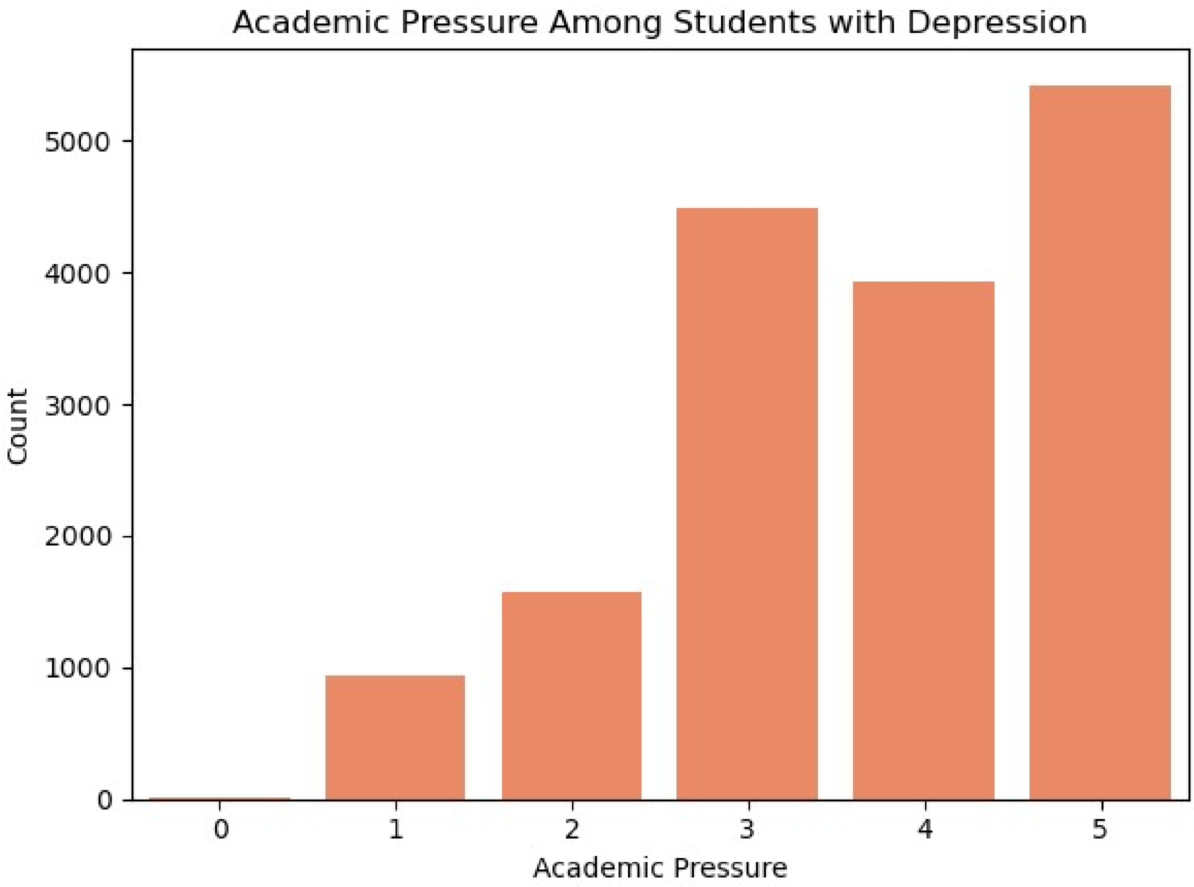
```
depression                0         1
academic_pressure
0                  0.555556  0.444444
1                  0.805874  0.194126
2                  0.625180  0.374820
3                  0.398419  0.601581
4                  0.238603  0.761397
5                  0.139136  0.860864
```

```
sns.barplot(x='depression', y='sleep_duration', data=df)
plt.title('Sleep Duration by Depression Level')
```

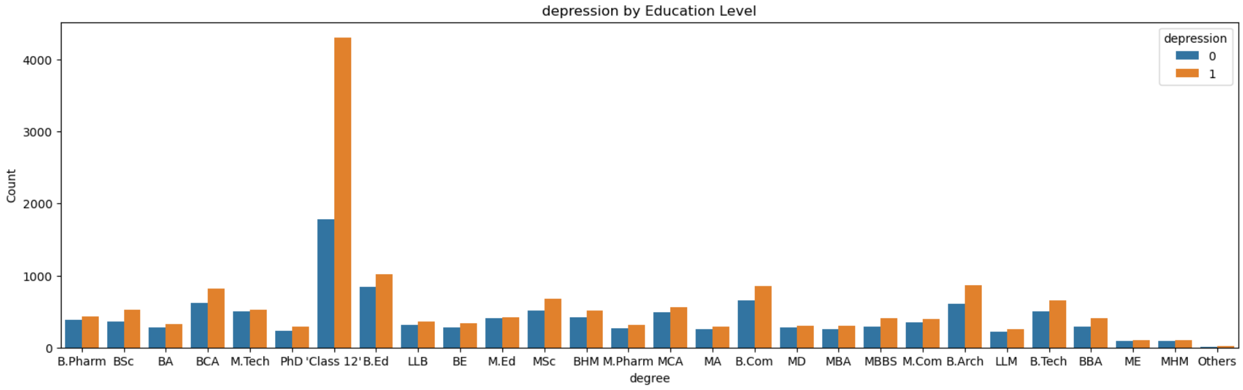
```
Text(0.5, 1.0, 'Sleep Duration by Depression Level')
```



```
 df_depressed = df[df['depression'] == 1]

sns.countplot(x='academic_pressure', data=df_depressed, color='coral')
plt.title('Academic Pressure Among Students with Depression')
plt.xlabel('Academic Pressure')
plt.ylabel('Count')
plt.tight_layout()
plt.show()
```

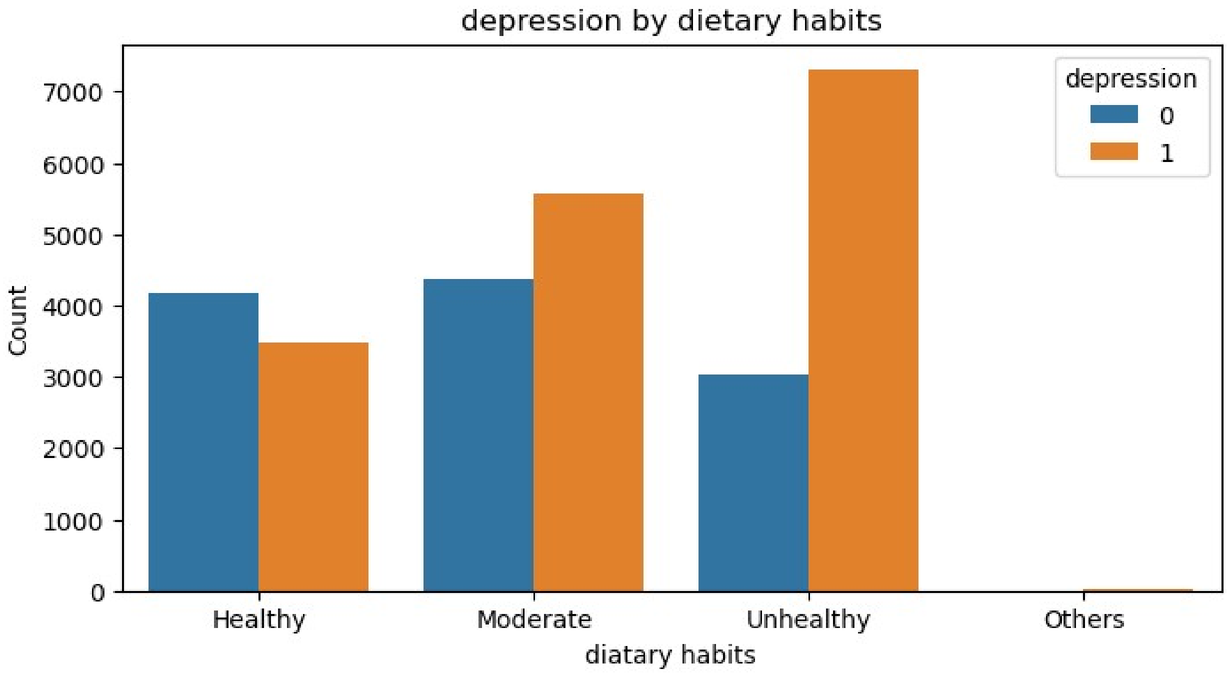


```
plt.figure(figsize=(18,5))
sns.countplot(data=df, x='degree', hue='depression')
plt.title('depression by Education Level')
plt.xlabel('degree')
plt.ylabel('Count')
plt.show()
```



```
plt.figure(figsize=(8,4))
sns.countplot(data=df, x='dietary_habits', hue='depression')
```

```
plt.title('depression by dietary habits')
plt.xlabel('diatary habits')
plt.ylabel('Count')
plt.show()
```


depression by dietary habits
depression
Count
Healthy
Moderate
Unhealthy
Others
diatary habits