eda-lab-mentalhealthdataset

November 13, 2024

Name- Sandeep Patro(202411093)

Importing Libraries

[34]:
```
import pandas as pd
import numpy as np
```

[35]:
```
url='https://raw.githubusercontent.com/Sandeep-git1/E-D-A/refs/heads/main/
↪mental_health_dataset.csv'
df=pd.read_csv(url)
```

[36]:
```
df.head()
```

[36]:

	id	Name	Gender	Age	City	Working Professional or Student	Profession	Academic Pressure	Work Pressure	CGPA	Study Satisfaction	Job Satisfaction	Sleep Duration	Dietary Habits	Degree	Have you ever had suicidal thoughts ?	Work/Study Hours
0	0	Aaradhya	Female	49.0	Ludhiana	Working Professional	Chef	NaN	5.0	NaN	NaN	2.0	More than 8 hours	Healthy	BHM	No	1.0
1	1	Vivan	Male	26.0	Varanasi	Working Professional	Teacher	NaN	4.0	NaN	NaN	3.0	Less than 5 hours	Unhealthy	LLB	Yes	7.0
2	2	Yuvraj	Male	33.0	Visakhapatnam	Student	NaN	5.0	NaN	8.97	2.0	NaN	5-6 hours	Healthy	B.Pharm	Yes	3.0
3	3	Yuvraj	Male	22.0	Mumbai	Working Professional	Teacher	NaN	5.0	NaN	NaN	1.0	Less than 5 hours	Moderate			
4	4	Rhea	Female	30.0	Kanpur	Working Professional	Business Analyst	NaN	1.0	NaN	NaN	1.0	5-6 hours	Unhealthy			

```
3      BBA                                    Yes              10.0
4      BBA                                    Yes               9.0

   Financial Stress Family History of Mental Illness  Depression
0               2.0                                No           0
1               3.0                                No           1
2               1.0                                No           1
3               1.0                               Yes           1
4               4.0                               Yes           0
```

[37]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 140700 entries, 0 to 140699
Data columns (total 20 columns):
 #   Column                                 Non-Null Count   Dtype
---  ------                                 --------------   -----
 0   id                                     140700 non-null  int64
 1   Name                                   140700 non-null  object
 2   Gender                                 140700 non-null  object
 3   Age                                    140700 non-null  float64
 4   City                                   140700 non-null  object
 5   Working Professional or Student        140700 non-null  object
 6   Profession                             104070 non-null  object
 7   Academic Pressure                      27897 non-null   float64
 8   Work Pressure                          112782 non-null  float64
 9   CGPA                                   27898 non-null   float64
 10  Study Satisfaction                     27897 non-null   float64
 11  Job Satisfaction                       112790 non-null  float64
 12  Sleep Duration                         140700 non-null  object
 13  Dietary Habits                         140696 non-null  object
 14  Degree                                 140698 non-null  object
 15  Have you ever had suicidal thoughts ?  140700 non-null  object
 16  Work/Study Hours                       140700 non-null  float64
 17  Financial Stress                       140696 non-null  float64
 18  Family History of Mental Illness       140700 non-null  object
 19  Depression                             140700 non-null  int64
dtypes: float64(8), int64(2), object(10)
memory usage: 21.5+ MB
```

[38]:
```
df.describe()
```

[38]:
```
                 id            Age  Academic Pressure  Work Pressure  \
count  140700.000000  140700.000000       27897.000000  112782.000000
mean    70349.500000      40.388621           3.142273       2.998998
std     40616.735775      12.384099           1.380457       1.405771
min         0.000000      18.000000           1.000000       1.000000
```

```
25%     35174.750000      29.000000           2.000000       2.000000
50%     70349.500000      42.000000           3.000000       3.000000
75%    105524.250000      51.000000           4.000000       4.000000
max    140699.000000      60.000000           5.000000       5.000000

                CGPA  Study Satisfaction  Job Satisfaction  Work/Study Hours  \
count  27898.000000        27897.000000     112790.000000     140700.000000
mean       7.658636            2.944940          2.974404          6.252679
std        1.464466            1.360197          1.416078          3.853615
min        5.030000            1.000000          1.000000          0.000000
25%        6.290000            2.000000          2.000000          3.000000
50%        7.770000            3.000000          3.000000          6.000000
75%        8.920000            4.000000          4.000000         10.000000
max       10.000000            5.000000          5.000000         12.000000

       Financial Stress     Depression
count     140696.000000  140700.000000
mean           2.988983       0.181713
std            1.413633       0.385609
min            1.000000       0.000000
25%            2.000000       0.000000
50%            3.000000       0.000000
75%            4.000000       0.000000
max            5.000000       1.000000
```

[39]:
```
df.isnull().sum()
```

[39]:
```
id                                          0
Name                                        0
Gender                                      0
Age                                         0
City                                        0
Working Professional or Student             0
Profession                              36630
Academic Pressure                      112803
Work Pressure                           27918
CGPA                                   112802
Study Satisfaction                     112803
Job Satisfaction                        27910
Sleep Duration                              0
Dietary Habits                              4
Degree                                      2
Have you ever had suicidal thoughts ?       0
Work/Study Hours                            0
Financial Stress                            4
Family History of Mental Illness            0
Depression                                  0
```

```
dtype: int64
```

[40]:
```
df.isnull().sum() / df.shape[0] * 100
```

[40]:
```
id                                        0.000000
Name                                      0.000000
Gender                                    0.000000
Age                                       0.000000
City                                      0.000000
Working Professional or Student           0.000000
Profession                               26.034115
Academic Pressure                        80.172708
Work Pressure                            19.842217
CGPA                                     80.171997
Study Satisfaction                       80.172708
Job Satisfaction                         19.836532
Sleep Duration                            0.000000
Dietary Habits                            0.002843
Degree                                    0.001421
Have you ever had suicidal thoughts ?     0.000000
Work/Study Hours                          0.000000
Financial Stress                          0.002843
Family History of Mental Illness          0.000000
Depression                                0.000000
dtype: float64
```

DATA CLEANING

[41]:
```
del df['Academic Pressure']
del df['Study Satisfaction']
del df['CGPA']

df.isnull().sum() / df.shape[0] * 100
```

[41]:
```
id                                        0.000000
Name                                      0.000000
Gender                                    0.000000
Age                                       0.000000
City                                      0.000000
Working Professional or Student           0.000000
Profession                               26.034115
Work Pressure                            19.842217
Job Satisfaction                         19.836532
Sleep Duration                            0.000000
Dietary Habits                            0.002843
Degree                                    0.001421
Have you ever had suicidal thoughts ?     0.000000
```

```
Work/Study Hours                          0.000000
Financial Stress                          0.002843
Family History of Mental Illness          0.000000
Depression                                0.000000
dtype: float64
```

[42]:
```
df['Financial Stress']=df['Financial Stress'].fillna(df['Financial Stress'].
 ↪mean())
```

[43]:
```
df['Degree']  = df['Degree'].fillna (df['Degree'].mode()[0])
```

[44]:
```
df['Dietary Habits'] = df['Dietary Habits'].fillna (df['Dietary Habits'].
 ↪mode()[0])
```

[45]:
```
categorical_variables = df.select_dtypes(include=['object']).columns
categorical_variables
```

[45]:
```
Index(['Name', 'Gender', 'City', 'Working Professional or Student',
       'Profession', 'Sleep Duration', 'Dietary Habits', 'Degree',
       'Have you ever had suicidal thoughts ?',
       'Family History of Mental Illness'],
      dtype='object')
```

[52]:
```
numerical_variables = df.select_dtypes(include=['int64', 'float64']).columns
numerical_variables = numerical_variables.difference(['Depression'])
numerical_variables
```

[52]:
```
Index(['Age', 'Financial Stress', 'Job Satisfaction', 'Work Pressure',
       'Work/Study Hours', 'id'],
      dtype='object')
```

[53]:
```
for var in categorical_variables:
  df[var] = df[var].fillna(df[var].mode()[0])
```

[57]:
```
for var in numerical_variables:
  df[var] = df[var].fillna(df[var].median())
```

[59]:
```
df.isnull().sum() / df.shape[0] * 100
```

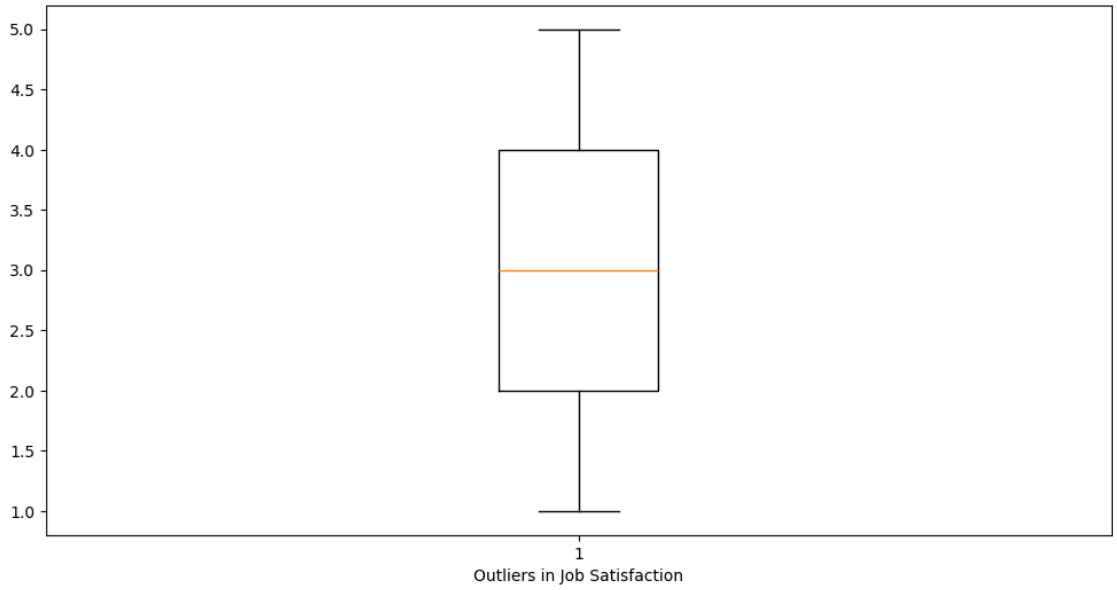
[59]:
```
id                                        0.0
Name                                      0.0
Gender                                    0.0
Age                                       0.0
City                                      0.0
Working Professional or Student           0.0
Profession                                0.0
Work Pressure                             0.0
```

```
Job Satisfaction                        0.0
Sleep Duration                          0.0
Dietary Habits                          0.0
Degree                                  0.0
Have you ever had suicidal thoughts ?   0.0
Work/Study Hours                        0.0
Financial Stress                        0.0
Family History of Mental Illness        0.0
Depression                              0.0
dtype: float64
```

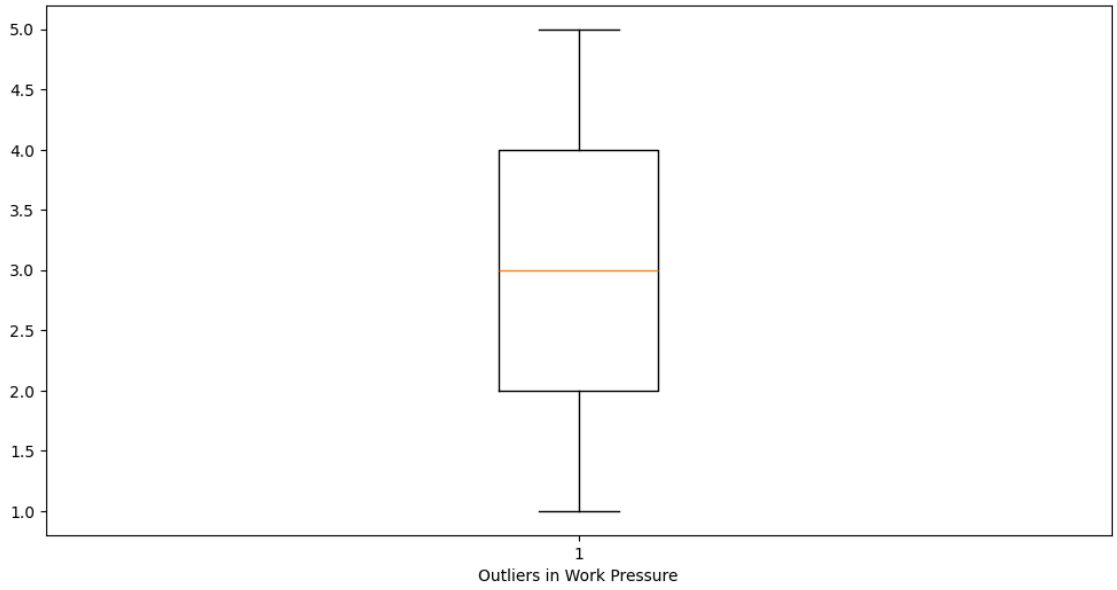
[60]:
```python
import matplotlib.pyplot as plt

for var in numerical_variables:
  plt.figure(figsize=(12, 6))
  plt.boxplot(df[var])
  plt.xlabel(f'Outliers in {var}')
  plt.show()
```

Outliers in Financial Stress

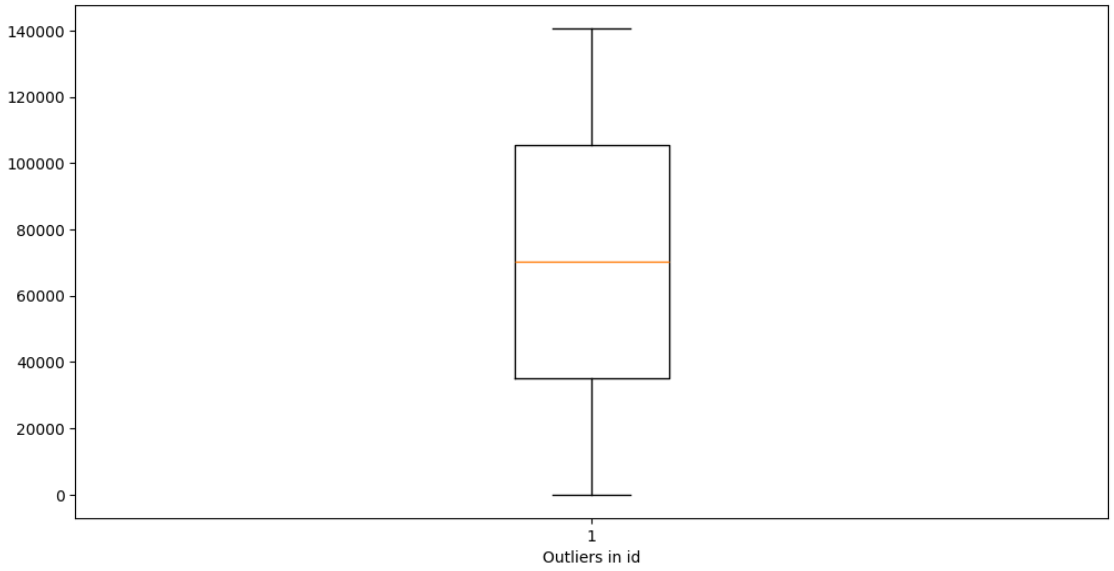


Outliers in Job Satisfaction



Outliers in Work Pressure

Outliers in Work/Study Hours



[61]:
```
for var in categorical_variables:
  print(f'{var} {len(df[var].unique())}')

categorical_variables
```

```
Name 422
Gender 2
City 98
Working Professional or Student 2
Profession 64
Sleep Duration 36
Dietary Habits 23
Degree 115
Have you ever had suicidal thoughts ? 2
Family History of Mental Illness 2
```

[61]:
```
Index(['Name', 'Gender', 'City', 'Working Professional or Student',
       'Profession', 'Sleep Duration', 'Dietary Habits', 'Degree',
       'Have you ever had suicidal thoughts ?',
       'Family History of Mental Illness'],
      dtype='object')
```

[62]:
```
df['Sleep Duration'].unique()
```

[62]:
```
array(['More than 8 hours', 'Less than 5 hours', '5-6 hours', '7-8 hours',
       'Sleep_Duration', '1-2 hours', '6-8 hours', '4-6 hours',
       '6-7 hours', '10-11 hours', '8-9 hours', '40-45 hours',
       '9-11 hours', '2-3 hours', '3-4 hours', 'Moderate', '55-66 hours',
```

```
       '4-5 hours', '9-6 hours', '1-3 hours', 'Indore', '45', '1-6 hours',
       '35-36 hours', '8 hours', 'No', '10-6 hours', 'than 5 hours',
       '49 hours', 'Unhealthy', 'Work_Study_Hours', '3-6 hours',
       '45-48 hours', '9-5', 'Pune', '9-5 hours'], dtype=object)
```

[63]:
```
to_encode_var = ['Gender', 'City','Profession', 'Dietary Habits',
                 'Have you ever had suicidal thoughts ?', 'Family History of␣
 ↪Mental Illness']

categorical_variables
```

[63]:
```
Index(['Name', 'Gender', 'City', 'Working Professional or Student',
       'Profession', 'Sleep Duration', 'Dietary Habits', 'Degree',
       'Have you ever had suicidal thoughts ?',
       'Family History of Mental Illness'],
      dtype='object')
```

[64]:
```
from sklearn.preprocessing import LabelEncoder

label_encoder = LabelEncoder()

for col in to_encode_var:
  df[col] = label_encoder.fit_transform(df[col])

df.head(5)
```

[64]:

	id	Name	Gender	Age	City	Working Professional or Student	\
0	0	Aaradhya	0	49.0	50	Working Professional	
1	1	Vivan	1	26.0	93	Working Professional	
2	2	Yuvraj	1	33.0	97	Student	
3	3	Yuvraj	1	22.0	64	Working Professional	
4	4	Rhea	0	30.0	37	Working Professional	

	Profession	Work Pressure	Job Satisfaction	Sleep Duration	\
0	10	5.0	2.0	More than 8 hours	
1	55	4.0	3.0	Less than 5 hours	
2	55	3.0	3.0	5-6 hours	
3	55	5.0	1.0	Less than 5 hours	
4	9	1.0	1.0	5-6 hours	

	Dietary Habits	Degree	Have you ever had suicidal thoughts ?	\
0	7	BHM	0	
1	20	LLB	1	
2	7	B.Pharm	1	
3	15	BBA	1	
4	20	BBA	1	

```
   Work/Study Hours  Financial Stress  Family History of Mental Illness  \
0               1.0               2.0                                 0
1               7.0               3.0                                 0
2               3.0               1.0                                 0
3              10.0               1.0                                 1
4               9.0               4.0                                 1

   Depression
0           0
1           1
2           1
3           1
4           0
```

[65]:
```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
df[to_encode_var] = pd.DataFrame(scaler.fit_transform(df[to_encode_var]),␣
 ↪columns=to_encode_var)

df.head(5)
```

[65]:
```
   id      Name    Gender   Age      City Working Professional or Student  \
0   0  Aaradhya -1.106796  49.0 -0.072046            Working Professional
1   1     Vivan  0.903508  26.0  1.363522            Working Professional
2   2    Yuvraj  0.903508  33.0  1.497063                         Student
3   3    Yuvraj  0.903508  22.0  0.395348            Working Professional
4   4      Rhea -1.106796  30.0 -0.506054            Working Professional

   Profession  Work Pressure  Job Satisfaction     Sleep Duration  \
0   -1.531798            5.0               2.0  More than 8 hours
1    0.885667            4.0               3.0  Less than 5 hours
2    0.885667            3.0               3.0          5-6 hours
3    0.885667            5.0               1.0  Less than 5 hours
4   -1.585519            1.0               1.0          5-6 hours

   Dietary Habits   Degree  Have you ever had suicidal thoughts ?  \
0       -1.347199      BHM                               -0.988861
1        1.120101      LLB                                1.011265
2       -1.347199  B.Pharm                                1.011265
3        0.171140      BBA                                1.011265
4        1.120101      BBA                                1.011265

   Work/Study Hours  Financial Stress  Family History of Mental Illness  \
0               1.0               2.0                          -0.994217
1               7.0               3.0                          -0.994217
2               3.0               1.0                          -0.994217
```

```
3              10.0              1.0                          1.005816
4               9.0              4.0                          1.005816

   Depression
0           0
1           1
2           1
3           1
4           0
```

[66]:
```
to_encode_var = pd.Index(to_encode_var)

train_variables = to_encode_var.append(numerical_variables[numerical_variables !
 ↪= 'id'])
train_variables
```

[66]:
```
Index(['Gender', 'City', 'Profession', 'Dietary Habits',
       'Have you ever had suicidal thoughts ?',
       'Family History of Mental Illness', 'Age', 'Financial Stress',
       'Job Satisfaction', 'Work Pressure', 'Work/Study Hours'],
      dtype='object')
```

[67]:
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(df[train_variables],␣
 ↪df['Depression']
                                                    , test_size=.2,␣
 ↪random_state=42)
```

[68]:
```
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(random_state=42)

rf.fit(X_train, y_train)
```

[68]:
```
RandomForestClassifier(random_state=42)
```

[69]:
```
y_pred = rf.predict(X_test)
```

[70]:
```
from sklearn.metrics import accuracy_score

accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy * 100:.2f}%')
```

```
Accuracy: 92.28%
```