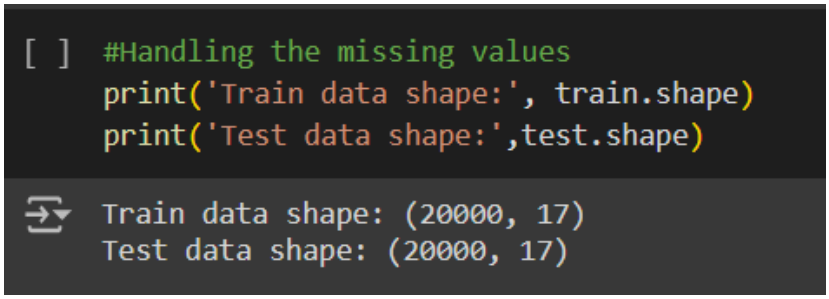
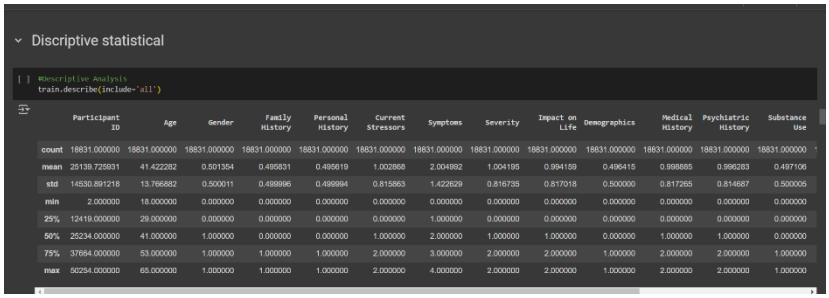
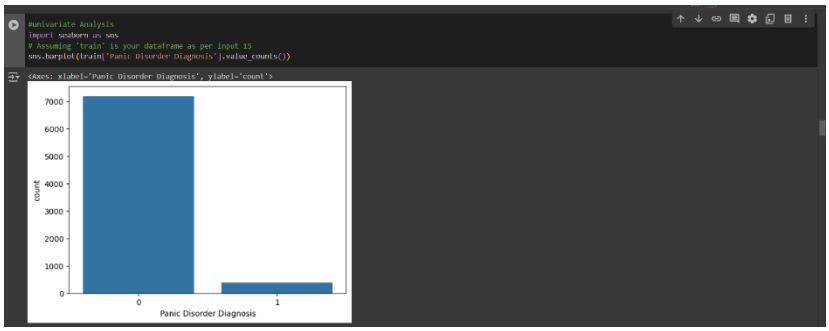
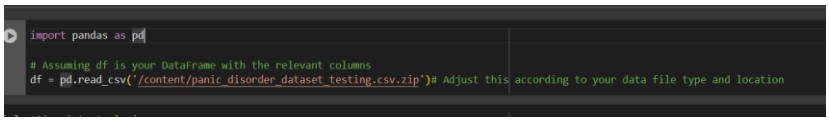
Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	740301
Project Title	Panic Disorder Detection
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

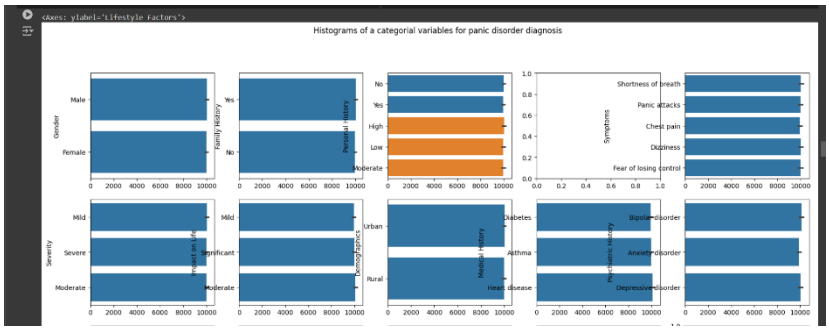
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

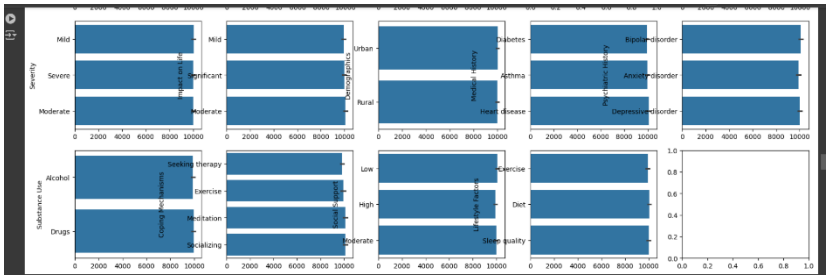
Section	Description
Data Overview	**#Structure of the data: -**  **#Descriptive Statistical:** Descriptive analysis is to study the basic features of data with the statistical process. Here pandas have a worthy function called describe. With this describe function we can understand the unique, top and frequent values of categorical features. And we can find mean, std, min, max and percentile values of continuous features.

	
Univariate Analysis	**Univariate Analysis:** In simple words, univariate analysis is understanding the data with single feature. Here we have displayed two different graphs such as Histplot and countplot. Seaborn package provides a wonderful function histplot. With the help of histplot, we can find the distribution of the feature. To make multiple graphs in a single plot, we use subplot. First let's check if the data is balanced or not. 
Bivariate Analysis	**#Bivariate Analysis:-** 

```
#Bivariate Analysis
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

fig,axes=plt.subplots(3,5,figsize=(20,10))
fig.suptitle("Histograms of a categorial variables for panic disorder diagnosis")
sns.barplot(df['Gender'],ax=axes[0,0])
sns.barplot(df['Family History'],ax=axes[0,1])
sns.barplot(df['Personal History'],ax=axes[0,2])
sns.barplot(df['Current Stressors'],ax=axes[0,3])
sns.barplot(df['Symptoms'],ax=axes[0,4])
sns.barplot(df['Severity'],ax=axes[1,0])
sns.barplot(df['Impact on Life'],ax=axes[1,1])
sns.barplot(df['Demographics'],ax=axes[1,2])
sns.barplot(df['Medical History'],ax=axes[1,3])
sns.barplot(df['Psychiatric History'],ax=axes[1,4])
sns.barplot(df['Substance Use'],ax=axes[2,0])
sns.barplot(df['Coping Mechanisms'],ax=axes[2,1])
sns.barplot(df['Social Support'],ax=axes[2,2])
sns.barplot(df['Lifestyle Factors'],ax=axes[2,4])
```





From the plot we came to know,

- Both the genders are diagnosed equally with panic disorder.
- The current stressors of the subjects are mostly high with a sleep deprived lifestyle.
- Panic disorder plays an important role in one's life and is severely affected to most of the subjects.
- The symptoms of the panic disorder are mainly 5 out of which Panic attacks are mostly observed.
- The social support provided for these subjects is also low and the coping mechanisms include seeking therapy by large number of the affected.

Visual Analysis	Visual analysis is the process of using visual representations, such as charts, plots, and graphs, to explore and understand data. It is a way to quickly identify patterns, trends, and outliers in the data, which can help to gain insights and make informed decisions. #In this no plots are available
Outliers and Anomalies	-

Data Preprocessing Code Screenshots

Loading Data	#Loading the data
	`#Read the dataset`
	`train = pd.read_csv("/content/panic_disorder_dataset_training.csv")`
	`train.head()`
	`test = pd.read_csv("/content/panic_disorder_dataset_testing.csv.zip")`
	`test.head()`

Handling Missing Data	`#Handling the missing values`
	`print('Train data shape:', train.shape)`
	`print('Test data shape:',test.shape)`
	`Train data shape: (20000, 17)`
	`Test data shape: (20000, 17)`
	`train.info()`

	 For checking the null values, . isnull() function is used. To sum those null values we use . sum() function. From the below image we found that there are no null values present in our dataset. So we can skip handling the missing values step.
Data Transformation	-
Feature Engineering	-
Save Processed Data	-