Data Collection and Preprocessing Phase

Date	8 July 2024
Team ID	SWTID1720075414
Project Title	Panic Disorder Detection
Maximum Marks	6 Marks

Data Exploration and Preprocessing

Data Overview

Dataset Name: panic_disorder_dataset_training
Dimensions: 100000 rows and 17 columns
Descriptive Statistics:

	Participant ID	Age	Panic Disorder Diagnosis
count	100000.000000	100000.000000	100000.00000
mean	50000.500000	41.454300	0.04285
std	28867.657797	13.839204	0.20252
min	1.000000	18.000000	0.00000
25%	25000.750000	29.000000	0.00000
50%	50000.500000	41.000000	0.00000
75%	75000.250000	53.000000	0.00000
max	100000.000000	65.000000	1.00000

Dataset Name: panic_disorder_dataset_testing
Dimensions: 20000 rows and 17 columns
Descriptive Statistics:

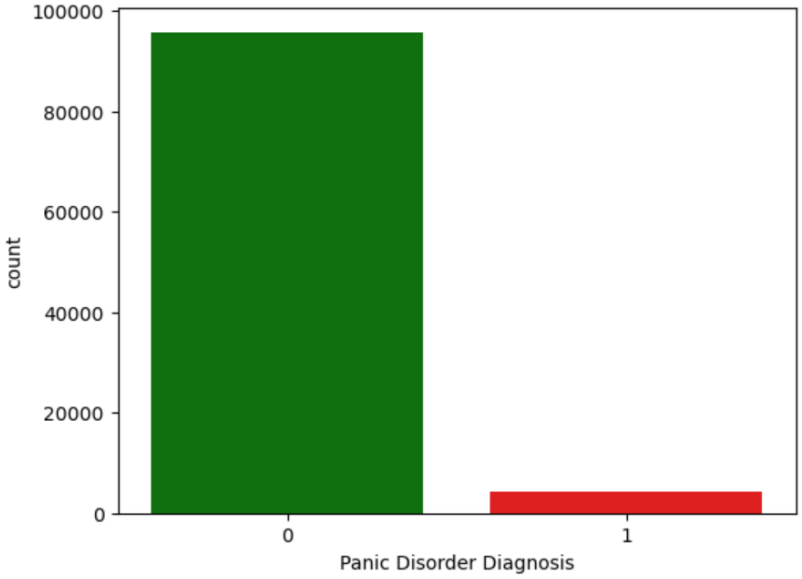
	Participant ID	Age	Panic Disorder Diagnosis
count	20000.000000	20000.000000	20000.000000
mean	10000.500000	41.489250	0.042050
std	5773.647028	13.887773	0.200708
min	1.000000	18.000000	0.000000
25%	5000.750000	29.000000	0.000000
50%	10000.500000	42.000000	0.000000
75%	15000.250000	54.000000	0.000000
max	20000.000000	65.000000	1.000000

Univariate Analysis

Panic Disorder Diagnosis

```
sns.countplot(data=train_invt,x='Panic Disorder Diagnosis',palette=['green','red'])
```

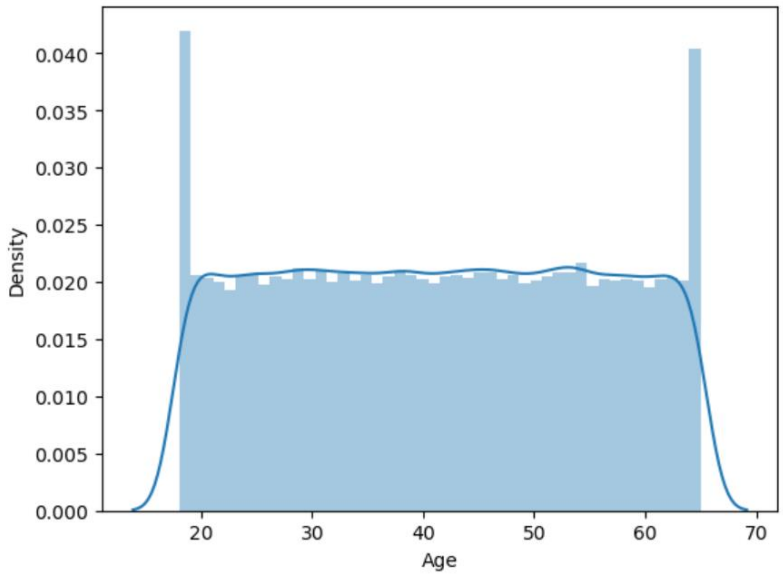
```
<Axes: xlabel='Panic Disorder Diagnosis', ylabel='count'>
```



Age

```
sns.distplot(train_invt.Age)
```

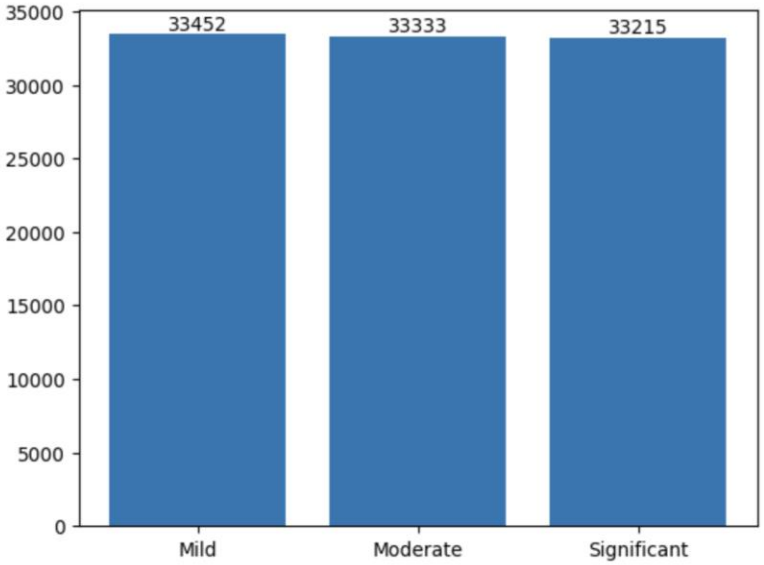
```
<Axes: xlabel='Age', ylabel='Density'>
```



Impact on Life

```
y_axis = train_invt['Impact on Life'].value_counts()
x_axis = y_axis.index.values
bars = plt.bar(x_axis,y_axis)
plt.bar_label(bars)
```

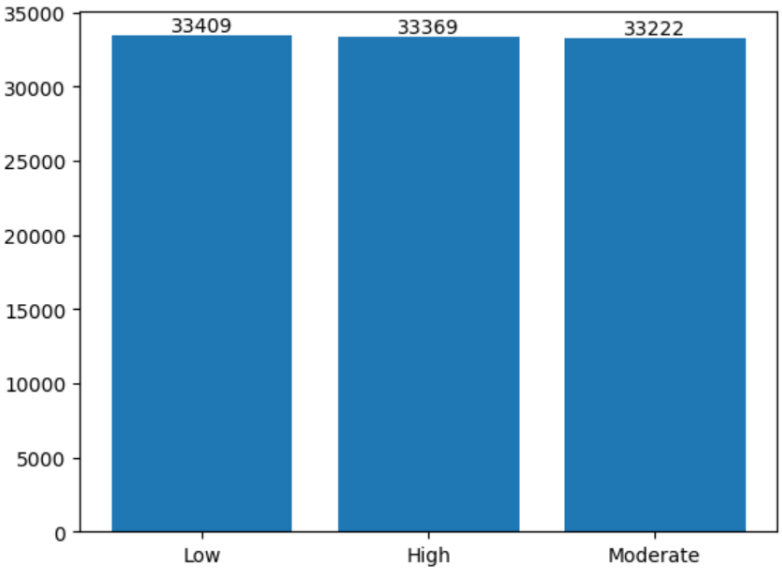
```
[Text(0, 0, '33452'), Text(0, 0, '33333'), Text(0, 0, '33215')]
```



Current Stressors

```
y_axis = train_invt['Current Stressors'].value_counts()
x_axis = y_axis.index.values
bars = plt.bar(x_axis,y_axis)
plt.bar_label(bars)
```

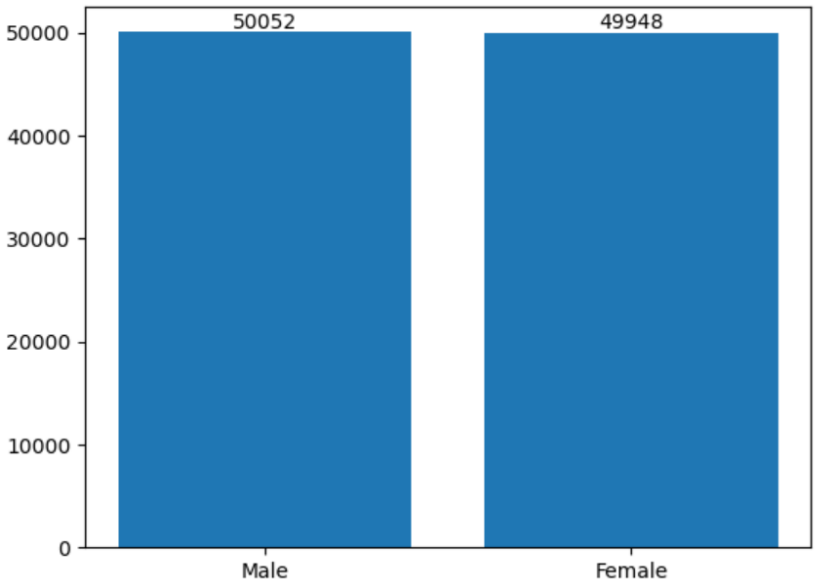
```
[Text(0, 0, '33409'), Text(0, 0, '33369'), Text(0, 0, '33222')]
```



Gender

```
y_axis = train_invt['Gender'].value_counts()
x_axis = y_axis.index.values
bars = plt.bar(x_axis,y_axis)
plt.bar_label(bars)
```

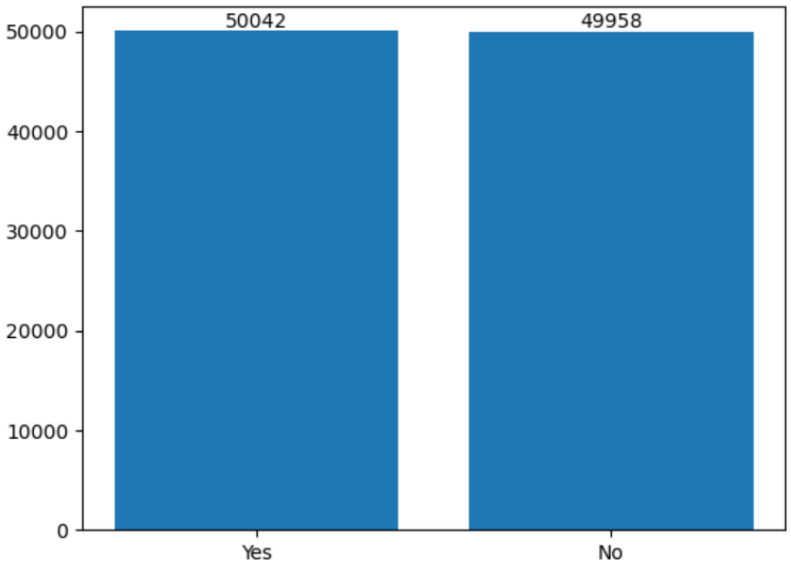
```
[Text(0, 0, '50052'), Text(0, 0, '49948')]
```



Family History

```
y_axis = train_invt['Family History'].value_counts()
x_axis = y_axis.index.values
bars = plt.bar(x_axis,y_axis)
plt.bar_label(bars)
```

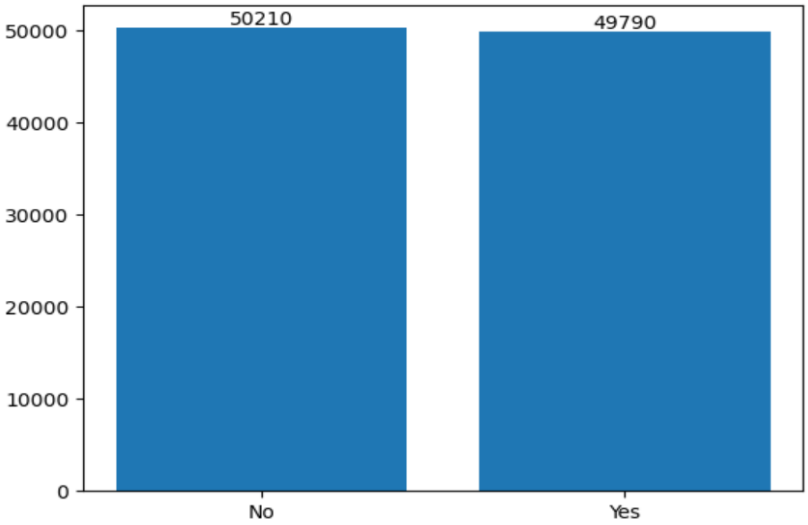
```
[Text(0, 0, '50042'), Text(0, 0, '49958')]
```



Personal History

```
y_axis = train_invt['Personal History'].value_counts()
x_axis = y_axis.index.values
bars = plt.bar(x_axis,y_axis)
plt.bar_label(bars)
```

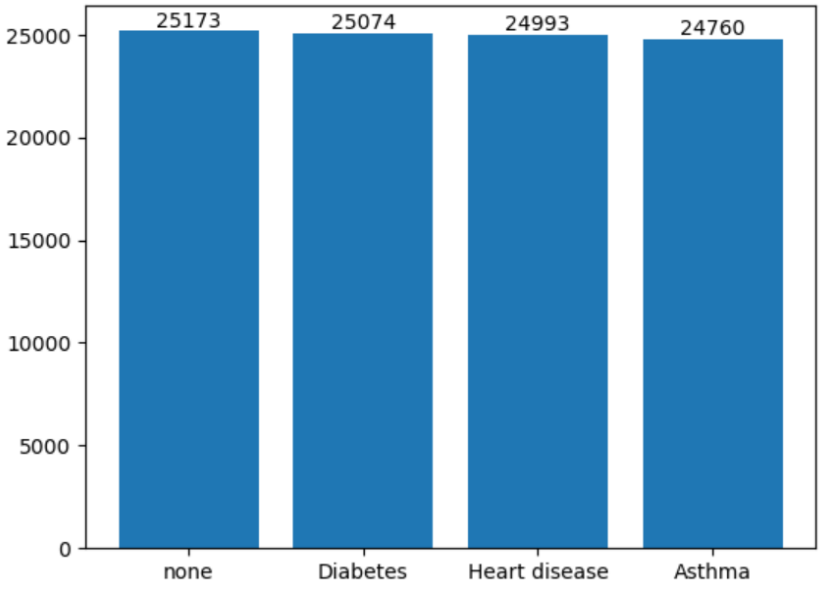
```
[Text(0, 0, '50210'), Text(0, 0, '49790')]
```



Medical History

```
y_axis = train_invt['Medical History'].value_counts()
x_axis = y_axis.index.values
bars = plt.bar(x_axis,y_axis)
plt.bar_label(bars)
```

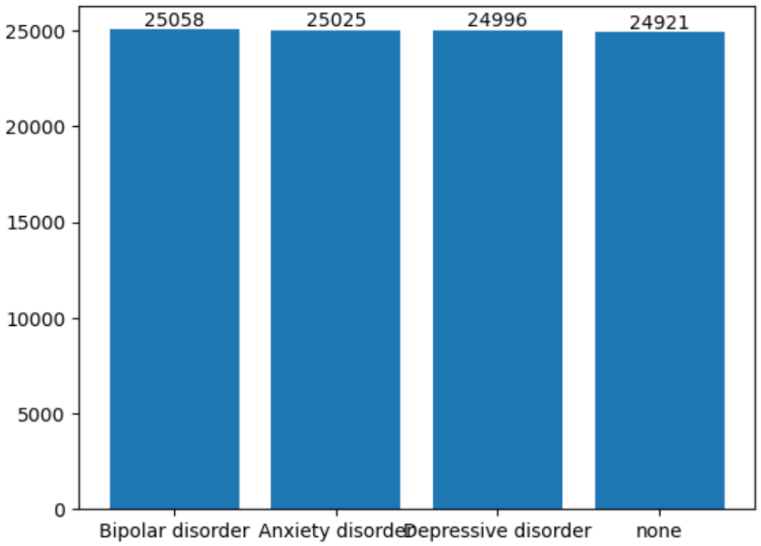
```
[Text(0, 0, '25173'),
 Text(0, 0, '25074'),
 Text(0, 0, '24993'),
 Text(0, 0, '24760')]
```



Psychiatric History

```
y_axis = train_invt['Psychiatric History'].value_counts()
x_axis = y_axis.index.values
bars = plt.bar(x_axis,y_axis)
plt.bar_label(bars)
```

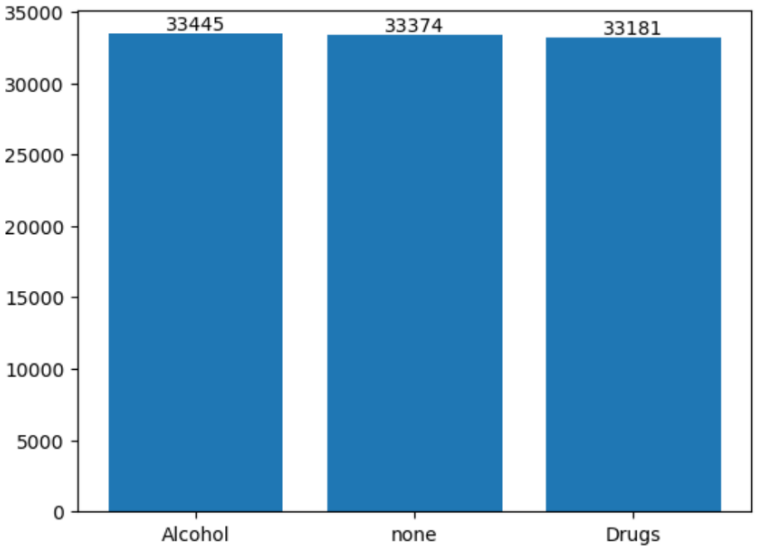
```
[Text(0, 0, '25058'),
 Text(0, 0, '25025'),
 Text(0, 0, '24996'),
 Text(0, 0, '24921')]
```



Substance Use

```
y_axis = train_invt['Substance Use'].value_counts()
x_axis = y_axis.index.values
bars = plt.bar(x_axis,y_axis)
plt.bar_label(bars)
```

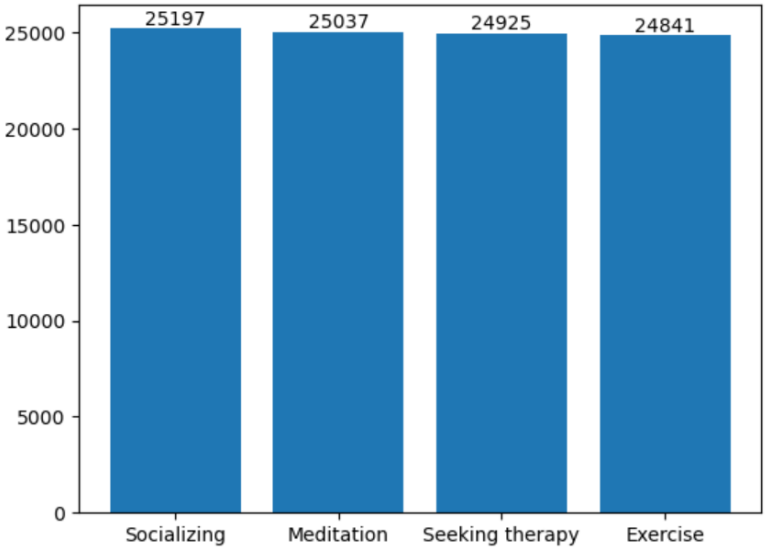
```
[Text(0, 0, '33445'), Text(0, 0, '33374'), Text(0, 0, '33181')]
```



Coping Mechanisms

```
y_axis = train_invt['Coping Mechanisms'].value_counts()
x_axis = y_axis.index.values
bars = plt.bar(x_axis,y_axis)
plt.bar_label(bars)
```

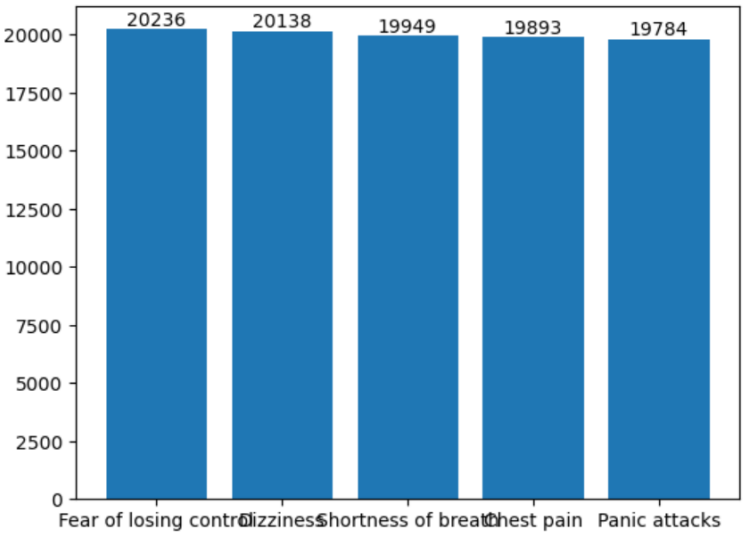
```
[Text(0, 0, '25197'),
 Text(0, 0, '25037'),
 Text(0, 0, '24925'),
 Text(0, 0, '24841')]
```



Symptoms

```
y_axis = train_invt['Symptoms'].value_counts()
x_axis = y_axis.index.values
bars = plt.bar(x_axis,y_axis)
plt.bar_label(bars)
```

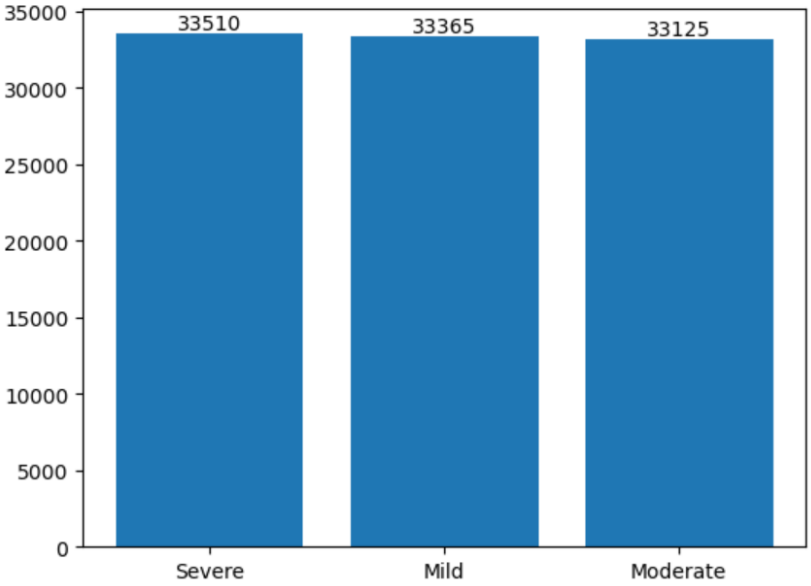
```
[Text(0, 0, '20236'),
 Text(0, 0, '20138'),
 Text(0, 0, '19949'),
 Text(0, 0, '19893'),
 Text(0, 0, '19784')]
```



Severity

```
y_axis = train_invt['Severity'].value_counts()
x_axis = y_axis.index.values
bars = plt.bar(x_axis,y_axis)
plt.bar_label(bars)
```

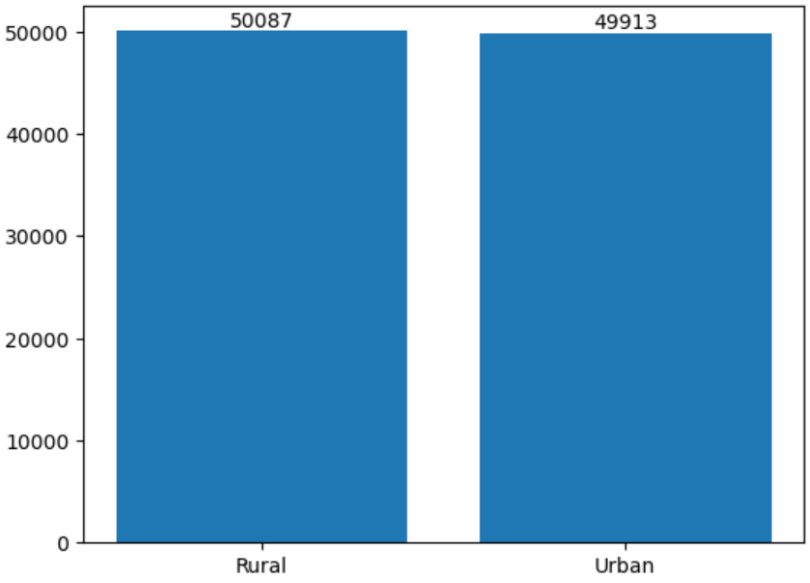
```
[Text(0, 0, '33510'), Text(0, 0, '33365'), Text(0, 0, '33125')]
```



Demographics

```
y_axis = train_invt['Demographics'].value_counts()
x_axis = y_axis.index.values
bars = plt.bar(x_axis,y_axis)
plt.bar_label(bars)
```

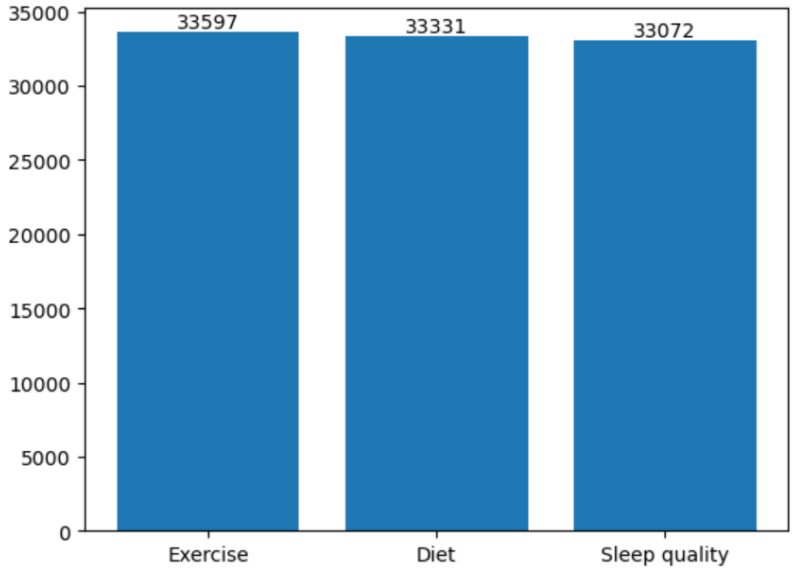
```
[Text(0, 0, '50087'), Text(0, 0, '49913')]
```



Lifestyle Factors

```
y_axis = train_invt['Lifestyle Factors'].value_counts()
x_axis = y_axis.index.values
bars = plt.bar(x_axis,y_axis)
plt.bar_label(bars)
```

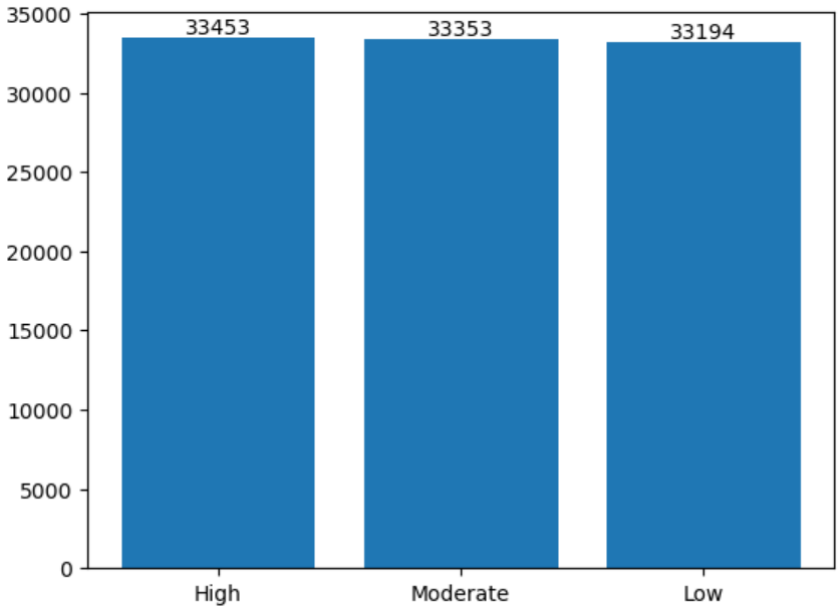
```
[Text(0, 0, '33597'), Text(0, 0, '33331'), Text(0, 0, '33072')]
```



Social Support

```
y_axis = train_invt['Social Support'].value_counts()
x_axis = y_axis.index.values
bars = plt.bar(x_axis,y_axis)
plt.bar_label(bars)
```

```
[Text(0, 0, '33453'), Text(0, 0, '33353'), Text(0, 0, '33194')]
```



Bivariate Analysis	
Multivariate Analysis	-
Outliers and Anomalies	-

Data Preprocessing Code Screenshots

Loading Data

```
[3]: train = pd.read_csv('panic_disorder_dataset_training.csv')
     train.head()
```

[3]:

	Participant ID	Age	Gender	Family History	Personal History	Current Stressors	Symptoms	Severity	Impact on Life	Demographics	Medical History	Psychiatric History	Substance Use	Coping Mechanisms	Su
0	1	38	Male	No	Yes	Moderate	Shortness of breath	Mild	Mild	Rural	Diabetes	Bipolar disorder	NaN	Socializing	
1	2	51	Male	No	No	High	Panic attacks	Mild	Mild	Urban	Asthma	Anxiety disorder	Drugs	Exercise	
2	3	32	Female	Yes	No	High	Panic attacks	Mild	Significant	Urban	Diabetes	Depressive disorder	NaN	Seeking therapy	Mod
3	4	64	Female	No	No	Moderate	Chest pain	Moderate	Moderate	Rural	Diabetes	NaN	NaN	Meditation	
4	5	31	Male	Yes	No	Moderate	Panic attacks	Mild	Moderate	Rural	Asthma	NaN	Drugs	Seeking therapy	

```
[4]: test = pd.read_csv('panic_disorder_dataset_testing.csv')
     test.head()
```

[4]:

	Participant ID	Age	Gender	Family History	Personal History	Current Stressors	Symptoms	Severity	Impact on Life	Demographics	Medical History	Psychiatric History	Substance Use	Coping Mechanisms	Su
0	1	41	Male	Yes	No	High	Shortness of breath	Mild	Mild	Urban	Diabetes	Bipolar disorder	Alcohol	Seeking therapy	
1	2	20	Female	Yes	No	Low	Shortness of breath	Mild	Significant	Urban	Asthma	Anxiety disorder	Drugs	Exercise	
2	3	32	Male	Yes	Yes	High	Panic attacks	Severe	Mild	Rural	Heart disease	Bipolar disorder	Drugs	Meditation	Mod
3	4	41	Female	Yes	Yes	Moderate	Shortness of breath	Moderate	Significant	Urban	Heart disease	Anxiety disorder	NaN	Exercise	
4	5	36	Female	Yes	No	High	Chest pain	Severe	Significant	Rural	Asthma	Depressive disorder	NaN	Seeking therapy	

Handling Missing Data

```
[21]: train["Medical History"].fillna("none", inplace=True)
      train["Medical History"].unique()
```

```
[21]: array(['Diabetes', 'Asthma', 'none', 'Heart disease'], dtype=object)
```

```
[22]: train["Psychiatric History"].fillna("none", inplace=True)
      train["Psychiatric History"].unique()
```

```
[22]: array(['Bipolar disorder', 'Anxiety disorder', 'Depressive disorder',
             'none'], dtype=object)
```

```
[23]: train["Substance Use"].fillna("none", inplace=True)
      train["Substance Use"].unique()
```

```
[23]: array(['none', 'Drugs', 'Alcohol'], dtype=object)
```

```
[24]: test["Medical History"].fillna("none", inplace=True)
      test["Medical History"].unique()
```

```
[24]: array(['Diabetes', 'Asthma', 'Heart disease', 'none'], dtype=object)
```

```
[25]: test["Psychiatric History"].fillna("none", inplace=True)
      test["Psychiatric History"].unique()
```

```
[25]: array(['Bipolar disorder', 'Anxiety disorder', 'Depressive disorder',
             'none'], dtype=object)
```

```
[26]: test["Substance Use"].fillna("none", inplace=True)
      test["Substance Use"].unique()
```

```
[26]: array(['Alcohol', 'Drugs', 'none'], dtype=object)
```

Data Transformation

```
scaler = StandardScaler()
x_res_train = pd.DataFrame(scaler.fit_transform(x_res_train), columns=x_train.columns)
```

```
x_res_train
```

	Age	Gender	Family History	Personal History	Current Stressors	Symptoms	Severity	Impact on Life	Demographics	Medical History	Psychiatric History
0	-0.241055	1.171061	-1.016200	0.956881	1.657461	1.453537	-1.340636	-1.382141	-0.796480	-0.275020	-0.292068
1	0.700946	1.171061	-1.016200	-1.045062	-0.826901	0.710251	-1.340636	-1.382141	1.255525	-1.214811	-1.214319
2	-0.675824	-0.853926	0.984059	-1.045062	-0.826901	0.710251	-1.340636	1.021337	1.255525	-0.275020	0.630183
3	1.642946	-0.853926	-1.016200	-1.045062	1.657461	-1.519607	-0.143681	-0.180402	-0.796480	-0.275020	1.552434
4	-0.748286	1.171061	0.984059	-1.045062	1.657461	0.710251	-1.340636	-0.180402	-0.796480	-1.214811	1.552434
...	...	...	...	...	...	...	...	...	...	...	...
191425	0.700946	-0.853926	-1.016200	0.956881	-0.826901	-0.033035	-0.143681	1.021337	-0.796480	-0.275020	-1.214319
191426	0.483561	-0.853926	0.984059	0.956881	1.657461	0.710251	-0.143681	-0.180402	-0.796480	-1.214811	-1.214319
191427	-0.313516	1.171061	0.984059	-1.045062	-0.826901	0.710251	-1.340636	1.021337	1.255525	-1.214811	0.630183
191428	0.918330	-0.853926	0.984059	-1.045062	1.657461	0.710251	1.053275	-0.180402	-0.796480	-1.214811	-0.292068
191429	-1.617825	-0.853926	-1.016200	0.956881	-0.826901	-0.033035	1.053275	1.021337	-0.796480	-0.275020	-1.214319

191430 rows × 15 columns

```
x_test = pd.DataFrame(scaler.fit_transform(x_test), columns=x_test.columns)
```

```
x_test
```

	Age	Gender	Family History	Personal History	Current Stressors	Symptoms	Severity	Impact on Life	Demographics	Medical History	Psychiatric History	S
0	-0.035230	1.00441	1.004309	-1.002704	-1.232680	1.420198	-1.219537	-1.249351	0.991933	-0.447545	-0.448969	-
1	-1.547389	-0.99561	1.004309	-1.002704	-0.006011	1.420198	-1.219537	1.199997	0.991933	-1.341829	-1.344401	
2	-0.683298	1.00441	1.004309	0.997304	-1.232680	0.715193	1.234753	-1.249351	-1.008133	0.446740	-0.448969	
3	-0.035230	-0.99561	1.004309	0.997304	1.220659	1.420198	0.007608	1.199997	0.991933	0.446740	-1.344401	
4	-0.395268	-0.99561	1.004309	-1.002704	-1.232680	-1.399823	1.234753	1.199997	-1.008133	-1.341829	0.446462	
...	...	...	...	...	...	...	...	...	...	...	...	
19995	-0.755306	-0.99561	1.004309	0.997304	-1.232680	-1.399823	0.007608	-0.024677	-1.008133	0.446740	-0.448969	
19996	-1.043336	1.00441	-0.995709	0.997304	-1.232680	0.715193	-1.219537	-0.024677	-1.008133	-1.341829	-0.448969	-
19997	-1.475381	-0.99561	-0.995709	-1.002704	-0.006011	-0.694818	0.007608	1.199997	0.991933	-1.341829	-1.344401	-
19998	-0.971328	1.00441	-0.995709	0.997304	1.220659	-0.694818	-1.219537	1.199997	-1.008133	0.446740	-1.344401	
19999	-1.403374	-0.99561	-0.995709	0.997304	-1.232680	-0.694818	1.234753	1.199997	0.991933	-0.447545	-0.448969	-

20000 rows × 15 columns

Feature Engineering

Attached the notebook in GitHub.

Save Processed Data

-