Data Wrangling with Python

Data wrangling is the process of transforming raw data into a more structured format. The process includes collecting, processing, analyzing, and tidying the raw data so that it can be easily read and analyzed. We can use the common library in python, that is "pandas".

```
# Before you load your dataset, Its a advised to import the libraries you will be using.
#If you get moduleNotFoundError run pip install pandas numpy matplotlib seaborn to install the modules

import pandas as pd
 #for data manipulation and analysis. Useful for cleaning, filtering, and transforming datasets.
import numpy as np
# for scientific computing in Python. It provides support for large, multi-dimensional arrays and matric
import matplotlib.pyplot as plt
#2D plotting library that produces static, animated, and interactive visualizations in Python.
import seaborn as sns
#tatistical data visualization library based on matplotlib. It provides a high-level interface for creat
```

```
#Use Pandas to load your Dataset

# loading and reading dataset
ds_path = r"D:\APHRC\Projects\Mental Health\data\MH.csv"
df = pd.read_csv(ds_path)
df.head()
```

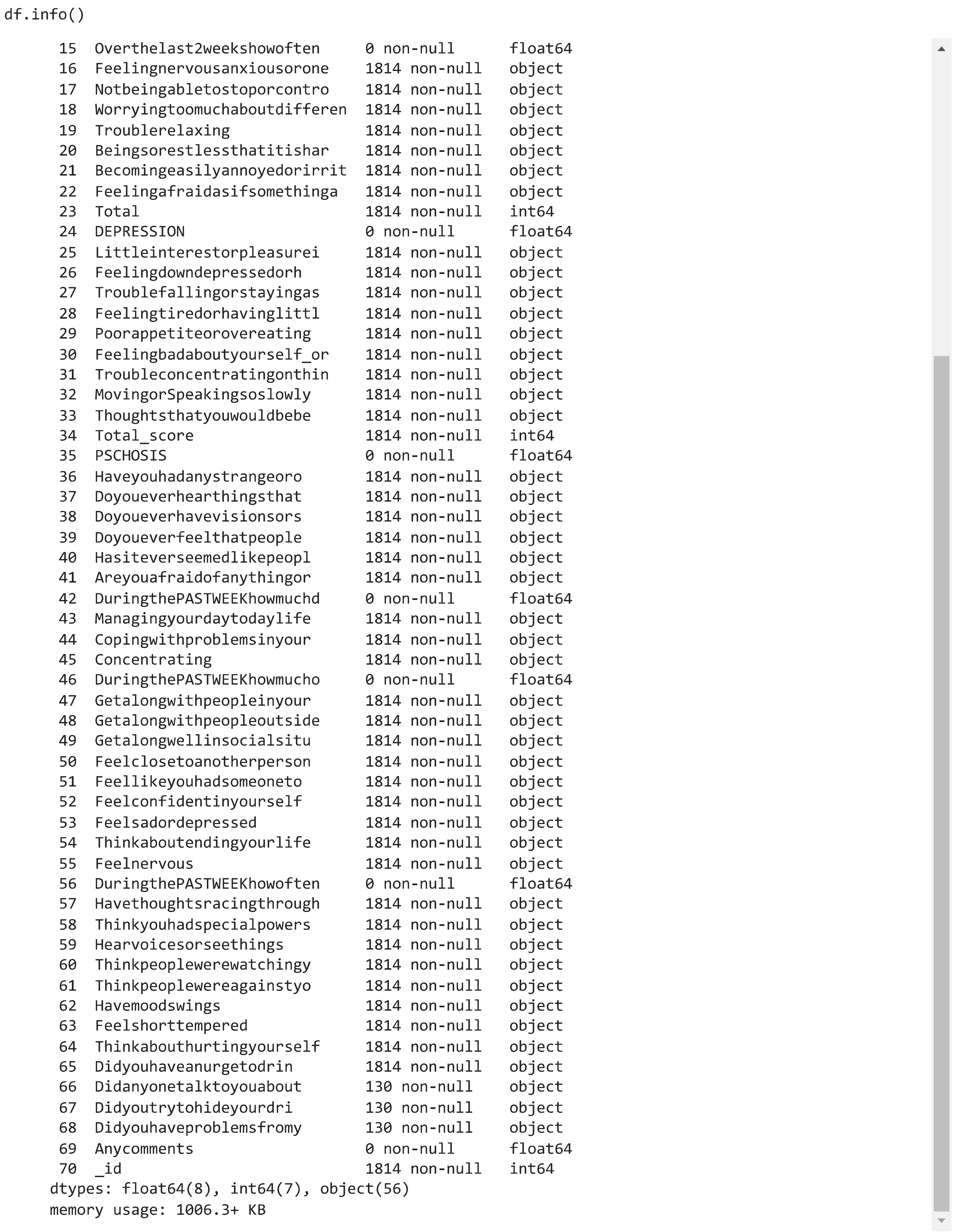
	Unnamed: 0	start	end	_submission_time	FA_Code	StudyID	_index	locationid	v
0	1	2023-06-29	2023-06-29	2023-06-30	ANE	12737	1735	I41607189001	NAI
1	2	2023-09-13	2023-09-13	2023-09-14	GYH	14487	7027	I31501165003	N
2	3	2023-09-28	2023-09-28	2023-09-28	BGO	12027	9019	M20901008002	
3	4	2023-07-27	2023-07-27	2023-07-27	MGA	5802	3667	I31201305001	
4	5	2023-08-14	2023-08-14	2023-08-21	MLF	7449	5101	I10503309001	E

5 rows × 71 columns

```
# shape of the data
df.shape
```

```
(1814, 71)
```

```
#data information
```



```
df.info()
```

```
 15  Overthelast2weekshowoften     0 non-null      float64
 16  Feelingnervousanxiousorone    1814 non-null   object
 17  Notbeingabletostoporcontro    1814 non-null   object
 18  Worryingtoomuchaboutdifferen  1814 non-null   object
 19  Troublerelaxing               1814 non-null   object
 20  Beingsorestlessthatitishar    1814 non-null   object
 21  Becomingeasilyannoyedorirrit  1814 non-null   object
 22  Feelingafraidasifsomethinga   1814 non-null   object
 23  Total                         1814 non-null   int64
 24  DEPRESSION                    0 non-null      float64
 25  Littleinterestorpleasurei     1814 non-null   object
 26  Feelingdowndepressedorh       1814 non-null   object
 27  Troublefallingorstayingas     1814 non-null   object
 28  Feelingtiredorhavinglittl     1814 non-null   object
 29  Poorappetiteorovereating      1814 non-null   object
 30  Feelingbadaboutyourself_or    1814 non-null   object
 31  Troubleconcentratingonthin    1814 non-null   object
 32  MovingorSpeakingsoslowly      1814 non-null   object
 33  Thoughtsthatyouwouldbebe      1814 non-null   object
 34  Total_score                   1814 non-null   int64
 35  PSCHOSIS                      0 non-null      float64
 36  Haveyouhadanystrangeoro       1814 non-null   object
 37  Doyoueverhearthingsthat       1814 non-null   object
 38  Doyoueverhavevisionsors       1814 non-null   object
 39  Doyoueverfeelthatpeople       1814 non-null   object
 40  Hasiteverseemedlikepeopl      1814 non-null   object
 41  Areyouafraidofanythingor      1814 non-null   object
 42  DuringthePASTWEEKhowmuchd     0 non-null      float64
 43  Managingyourdaytodaylife      1814 non-null   object
 44  Copingwithproblemsinyour      1814 non-null   object
 45  Concentrating                 1814 non-null   object
 46  DuringthePASTWEEKhowmucho     0 non-null      float64
 47  Getalongwithpeopleinyour      1814 non-null   object
 48  Getalongwithpeopleoutside     1814 non-null   object
 49  Getalongwellinsocialsitu      1814 non-null   object
 50  Feelclosetoanotherperson      1814 non-null   object
 51  Feellikeyouhadsomeoneto       1814 non-null   object
 52  Feelconfidentinyourself       1814 non-null   object
 53  Feelsadordepressed            1814 non-null   object
 54  Thinkaboutendingyourlife      1814 non-null   object
 55  Feelnervous                   1814 non-null   object
 56  DuringthePASTWEEKhowoften     0 non-null      float64
 57  Havethoughtsracingthrough     1814 non-null   object
 58  Thinkyouhadspecialpowers      1814 non-null   object
 59  Hearvoicesorseethings         1814 non-null   object
 60  Thinkpeoplewerewatchingy      1814 non-null   object
 61  Thinkpeoplewereagainstyo      1814 non-null   object
 62  Havemoodswings                1814 non-null   object
 63  Feelshorttempered             1814 non-null   object
 64  Thinkabouthurtingyourself     1814 non-null   object
 65  Didyouhaveanurgetodrin        1814 non-null   object
 66  Didanyonetalktoyouabout       130 non-null    object
 67  Didyoutrytohideyourdri        130 non-null    object
 68  Didyouhaveproblemsfromy       130 non-null    object
 69  Anycomments                   0 non-null      float64
 70  _id                           1814 non-null   int64
dtypes: float64(8), int64(7), object(56)
memory usage: 1006.3+ KB
```

```
# describing the data
df.describe()
```

	Unnamed: 0	StudyID	_index	age	ANXIETY	Overthelast2weekshow
count	1814.000000	1814.000000	1814.000000	1814.000000	0.0	
mean	907.500000	8431.299338	4580.059537	52.060088	NaN	
std	523.801012	4788.627705	2670.713830	14.922866	NaN	
min	1.000000	37.000000	7.000000	19.000000	NaN	
25%	454.250000	4770.250000	2221.250000	42.000000	NaN	
50%	907.500000	7998.500000	4493.000000	50.000000	NaN	
75%	1360.750000	12746.750000	6891.750000	61.000000	NaN	
max	1814.000000	17013.000000	9302.000000	105.000000	NaN	

The DataFrame “df” is statistically summarized by the code df.describe(), which gives the count, mean, standard deviation, minimum, and quartiles for each numerical column. The dataset’s central tendencies and spread are briefly summarized.

```
#column to list

df.columns.tolist()
```

```
      'IndividualId',
      'ANXIETY',
      'Overthelast2weekshowoften',
      'Feelingnervousanxiousorone',
      'Notbeingabletostoporcontro',
      'Worryingtoomuchaboutdifferen',
      'Troublerelaxing',
      'Beingsorestlessthatitishar',
      'Becomingeasilyannoyedorirrit',
      'Feelingafraidasifsomethinga',
      'Total',
      'DEPRESSION',
      'Littleinterestorpleasurei',
      'Feelingdowndepressedorh',
      'Troublefallingorstayingas',
      'Feelingtiredorhavinglittl',
      'Poorappetiteorovereating',
 'Managingyourdaytodaylife',
 'Copingwithproblemsinyour',
 'Concentrating',
 'DuringthePASTWEEKhowmucho',
 'Getalongwithpeopleinyour',
 'Getalongwithpeopleoutside',
 'Getalongwellinsocialsitu',
 'Feelclosetoanotherperson',
 'Feellikeyouhadsomeoneto',
 'Feelconfidentinyourself',
 'Feelsadordepressed',
 'Thinkaboutendingyourlife',
 'Feelnervous',
 'DuringthePASTWEEKhowoften',
 'Havethoughtsracingthrough',
 'Thinkyouhadspecialpowers',
 'Hearvoicesorseethings',
 'Thinkpeoplewerewatchingy',
 'Thinkpeoplewereagainstyo',
 'Havemoodswings',
 'Feelshorttempered',
 'Thinkabouthurtingyourself',
 'Didyouhaveanurgetodrin',
 'Didanyonetalktoyouabout',
 'Didyoutrytohideyourdri',
 'Didyouhaveproblemsfromy',
 'Anycomments',
 ' id']
```

Cleaning your Dataset

1. Dealing with Null Values
2. Duplicates
3. changing datatypes
4. Deleting irrerevant columns

```
#Checking for null values

# Check for null values in each column
null_columns = df.columns[df.isnull().any()]

# Create a DataFrame with only columns containing null values
df_null_columns = df[null_columns]

# Calculate the total null values in each column
total_null_values_per_column = df_null_columns.isnull().sum()

# Display the result
print("Columns with null values and their total null values:")
print(total_null_values_per_column)
```

```
Columns with null values and their total null values:
residence                      75
ANXIETY                      1814
Overthelast2weekshowoften    1814
DEPRESSION                   1814
PSCHOSIS                     1814
DuringthePASTWEEKhowmuchd    1814
```

```
DuringthePASTWEEKhowmucho    1814
DuringthePASTWEEKhowoften    1814
Didanyonetalktoyouabout      1684
Didyoutrytohideyourdri       1684
Didyouhaveproblemsfromy      1684
Anycomments                  1814
dtype: int64
```

Go back and explore your missing data so that you know how to handle them.

Handling Missing Data

1. Dropping Missing Values:

Method: Use dropna() method in pandas.

Pros: Simple and quick. Useful when the missing data is random and removing those rows doesn't significantly affect the analysis.

Cons: May lead to loss of information, especially if the missing data is not entirely random.

2. Imputation:

Method: Fill in missing values with a specific value (e.g., mean, median, or mode) or use more advanced imputation methods.

Pros: Retains more data compared to dropping. Can be suitable for datasets with systematic missingness.

Cons: Imputed values may introduce bias, and the choice of imputation method is critical.

3. Forward or Backward Fill:

Method: Propagate the last valid observation forward or use the next valid observation to fill gaps.

Pros: Simple and suitable for time-series data.

Cons: The method may not be suitable for all types of data, and it assumes a certain temporal pattern.

4. Interpolation:

Method: Use methods like linear interpolation to estimate missing values based on surrounding values.

```
#In our case we can drop columns with 1814 missing values as they are insignificant

# Find columns with 1814 null values
columns_to_drop = df.columns[df.isnull().sum() == 1814]

# Drop the columns
df_dropped = df.drop(columns=columns_to_drop)

# Display the resulting DataFrame
print("DataFrame after dropping columns:")
print(df_dropped.columns)

df_dropped.shape
```

```
DataFrame after dropping columns:
Index(['Unnamed: 0', 'start', 'end', '_submission_time', 'FA_Code', 'StudyID',
       '_index', 'locationid', 'villagenam', 'residence', 'HHHead_id',
            'gender', 'age', 'IndividualId', 'Feelingnervousanxiousorone',
            'Notbeingabletostoporcontro', 'Worryingtoomuchaboutdifferen',
            'Troublerelaxing', 'Beingsorestlessthatitishar',
            'Becomingeasilyannoyedorirrit', 'Feelingafraidasifsomethinga', 'Total',
            'Littleinterestorpleasurei', 'Feelingdowndepressedorh',
            'Troublefallingorstayingas', 'Feelingtiredorhavinglittl',
            'Poorappetiteorovereating', 'Feelingbadaboutyourself_or',
            'Troubleconcentratingonthin', 'MovingorSpeakingsoslowly',
            'Thoughtsthatyouwouldbebe', 'Total_score', 'Haveyouhadanystrangeoro',
            'Doyoueverhearthingsthat', 'Doyoueverhavevisionsors',
            'Doyoueverfeelthatpeople', 'Hasiteverseemedlikepeopl',
            'Areyouafraidofanythingor', 'Managingyourdaytodaylife',
            'Copingwithproblemsinyour', 'Concentrating', 'Getalongwithpeopleinyour',
            'Getalongwithpeopleoutside', 'Getalongwellinsocialsitu',
            'Feelclosetoanotherperson', 'Feellikeyouhadsomeoneto',
            'Feelconfidentinyourself', 'Feelsadordepressed',
            'Thinkaboutendingyourlife', 'Feelnervous', 'Havethoughtsracingthrough',
            'Thinkyouhadspecialpowers', 'Hearvoicesorseethings',
            'Thinkpeoplewerewatchingy', 'Thinkpeoplewereagainstyo',
            'Havemoodswings', 'Feelshorttempered', 'Thinkabouthurtingyourself',
            'Didyouhaveanurgetodrin', 'Didanyonetalktoyouabout',
            'Didyoutrytohideyourdri', 'Didyouhaveproblemsfromy', '_id'],
           dtype='object')
     (1814, 63)
```

```
#For residence, 75 people didnt answer whether they are from Rural or Urban areas. If we explore the dat

# Check for null values in the residence column
null_residence_indices = df_dropped[df_dropped['residence'].isnull()].index

# Iterate over the rows with null residence values
for index in null_residence_indices:
    # Check if the village name starts with "BULUBANDI"
    if df_dropped.loc[index, 'villagenam'].startswith("BULUBANDI"):
        # Fill missing value with the corresponding village name
        df_dropped.loc[index, 'residence'] = df_dropped.loc[index, 'villagenam']

# Display the resulting DataFrame
print("DataFrame after filling missing values in the residence column:")
df_dropped
```

```
DataFrame after filling missing values in the residence column:
```

	Unnamed: 0	start	end	_submission_time	FA_Code	StudyID	_index	locationid
0	1	2023-06-29	2023-06-29	2023-06-30	ANE	12737	1735	I41607189001
1	2	2023-09-13	2023-09-13	2023-09-14	GYH	14487	7027	I31501165003
2	3	2023-09-28	2023-09-28	2023-09-28	BGO	12027	9019	M20901008002
3	4	2023-07-27	2023-07-27	2023-07-27	MGA	5802	3667	I31201305001
4	5	2023-08-14	2023-08-14	2023-08-21	MLF	7449	5101	I10503309001
...	...	...	...	...	...	...	...	...
1809	1810	2023-07-04	2023-07-05	2023-07-05	NAS	15401	2209	I21002157001
1810	1811	2023-08-11	2023-08-11	2023-08-11	BGO	7240	4246	I10503125001
1811	1812	2023-06-26	2023-06-26	2023-06-30	NFL	16808	1718	M20704062004
1812	1813	2023-09-18	2023-09-18	2023-09-18	NPR	3517	7336	I31401882003
1813	1814	2023-08-22	2023-08-22	2023-08-22	KAR	15133	5371	I10404015001

1814 rows × 63 columns

```
df.shape
```

```
(1814, 71)
```

```
#This dataset contains mental health info for Anxiety, depression and Pschosis. For the purposes of this
 #List of columns to drop
columns_to_drop = ['Unnamed: 0', '_index','Haveyouhadanystrangeoro',
       'Doyoueverhearthingsthat', 'Doyoueverhavevisionsors',
       'Doyoueverfeelthatpeople', 'Hasiteverseemedlikepeopl',
       'Areyouafraidofanythingor', 'Managingyourdaytodaylife',
       'Copingwithproblemsinyour', 'Concentrating', 'Getalongwithpeopleinyour',
       'Getalongwithpeopleoutside', 'Getalongwellinsocialsitu',
       'Feelclosetoanotherperson', 'Feellikeyouhadsomeoneto',
       'Feelconfidentinyourself', 'Feelsadordepressed',
       'Thinkaboutendingyourlife', 'Feelnervous', 'Havethoughtsracingthrough',
       'Thinkyouhadspecialpowers', 'Hearvoicesorseethings',
       'Thinkpeoplewerewatchingy', 'Thinkpeoplewereagainstyo',
       'Havemoodswings', 'Feelshorttempered', 'Thinkabouthurtingyourself',
       'Didyouhaveanurgetodrin', 'Didanyonetalktoyouabout',
       'Didyoutrytohideyourdri', 'Didyouhaveproblemsfromy', '_id']

# Drop the specified columns
df_dropped1 = df_dropped.drop(columns=columns_to_drop, axis=1)

df_dropped1.shape
```

```
(1814, 30)
```

```
#Rename the _submission_time column to Submission date
df_dropped1 = df_dropped1.rename(columns={'_submission_time': 'Submission date'})
df_dropped1.dtypes
```

```
start                           object
end                             object
Submission date                 object
FA_Code                         object
StudyID                          int64
locationid                      object
villagenam                      object
residence                       object
HHHead_id                       object
gender                          object
age                              int64
IndividualId                    object
Feelingnervousanxiousorone      object
Notbeingabletostoporcontro      object
Worryingtoomuchaboutdifferen    object
Troublerelaxing                 object
Beingsorestlessthatitishar      object
Becomingeasilyannoyedorirrit    object
Feelingafraidasifsomethinga     object
Total                            int64
Littleinterestorpleasurei       object
Feelingdowndepressedorh         object
Troublefallingorstayingas       object
Feelingtiredorhavinglittl       object
Poorappetiteorovereating        object
Feelingbadaboutyourself_or      object
Troubleconcentratingonthin      object
MovingorSpeakingsoslowly        object
Thoughtsthatyouwouldbebe        object
```

```
     Total_score                       int64
     dtype: object
```

```
#Save our dataframe as Cleaned_df

clean_df = df_dropped1

clean_df.head(5)
```