Lab 5: KNN

Date: 20.03.2024

Created by: Preksha Shah | 2348446

Dataset: Burden disease from each mental-illness

Import necessary libraries:

```
#Import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import confusion_matrix, classification_report
```

Display basic information about the dataset:

```
#Load the dataset of your choice into your Python environment.
data = pd.read_csv("/content/2- burden-disease-from-each-mental-illness(1).csv")
```

```
# Print the number of samples (rows) and features (columns) in the dataset
print("Number of samples:", data.shape[0])
print("Number of features:", data.shape[1])
```

```
Number of samples: 6840
Number of features: 8
```

```
# Print data types of each feature
print("\nData types of features:")
print(data.dtypes)
```

```
Data types of features:
Entity                                                                          object
Code                                                                            object
Year                                                                             int64
DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Depressive disorders    float64
DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Schizophrenia           float64
DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Bipolar disorder        float64
DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Eating disorders        float64
DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Anxiety disorders       float64
dtype: object
```

```
# Print the first few rows of the dataset
print("\nFirst few rows of the dataset:")
print(data.head())
```

```
First few rows of the dataset:
        Entity Code  Year  \
0  Afghanistan  AFG  1990
1  Afghanistan  AFG  1991
2  Afghanistan  AFG  1992
3  Afghanistan  AFG  1993
4  Afghanistan  AFG  1994

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Depressive disorders  \
0                                          895.22565
1                                          893.88434
2                                          892.34973
3                                          891.51587
4                                          891.39160

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Schizophrenia  \
0                                          138.24825
1                                          137.76122
2                                          137.08030
3                                          136.48602
4                                          136.18323

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Bipolar disorder  \
0                                          147.64412
1                                          147.56696
```

```
2                                        147.13086
3                                        146.78812
4                                        146.58481

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Eating disorders  \
0                                        26.471115
1                                        25.548681
2                                        24.637949
3                                        23.863169
4                                        23.189074

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Anxiety disorders
0                                        440.33000
1                                        439.47202
2                                        437.60718
3                                        436.69104
4                                        436.76800
```

Univariate Analysis

For Numerical Variables:

```
#Calculate basic descriptive statistics
numerical_variables = ['DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Depressive disorders',
                       'DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Schizophrenia',
                       'DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Bipolar disorder',
                       'DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Eating disorders',
                       'DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Anxiety disorders']
print("\nBasic Descriptive Statistics for Numerical Variables:")
print(data[numerical_variables].describe())
```

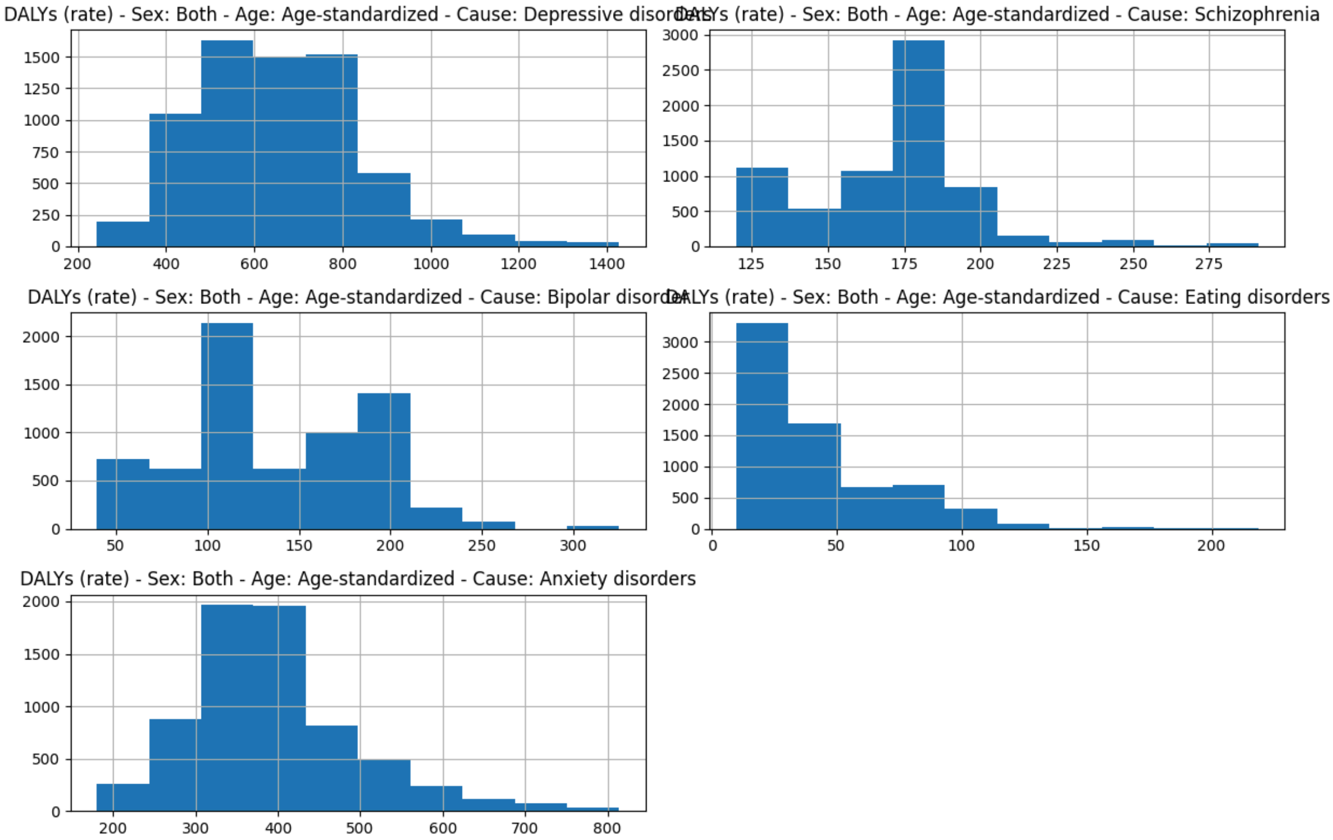
```
Basic Descriptive Statistics for Numerical Variables:
       DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Depressive disorders  \
count                                        6840.000000
mean                                          652.215475
std                                           183.643326
min                                           243.097840
25%                                           506.857413
50%                                           640.099150
75%                                           765.842910
max                                          1427.423600

       DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Schizophrenia  \
count                                        6840.000000
mean                                          171.090876
std                                            26.234514
min                                           119.913380
25%                                           155.950035
50%                                           175.115100
75%                                           183.999005
max                                           291.100100

       DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Bipolar disorder  \
count                                        6840.000000
mean                                          137.930619
std                                            51.197175
min                                            39.438133
25%                                           112.140244
50%                                           124.228445
75%                                           184.438120
max                                           325.152800

       DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Eating disorders  \
count                                        6840.000000
mean                                           42.392972
std                                            29.394380
min                                             9.671199
25%                                            20.837689
50%                                            31.430651
75%                                            55.850353
max                                           218.704390

       DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Anxiety disorders
count                                        6840.000000
mean                                          392.942475
std                                           100.820728
min                                           180.049640
25%                                           327.652407
50%                                           376.317940
75%                                           438.437842
max                                           814.302300
```

Inference:

- The dataset spans from the year 1990 to 2019.
- The mean DALYs rates vary across different causes of disorders, with depressive disorders having the highest mean rate.
- There is considerable variability in DALYs rates across different causes, as indicated by the standard deviations.
- The distribution of DALYs rates for each cause can be further explored through visualization techniques such as histograms and box plots.

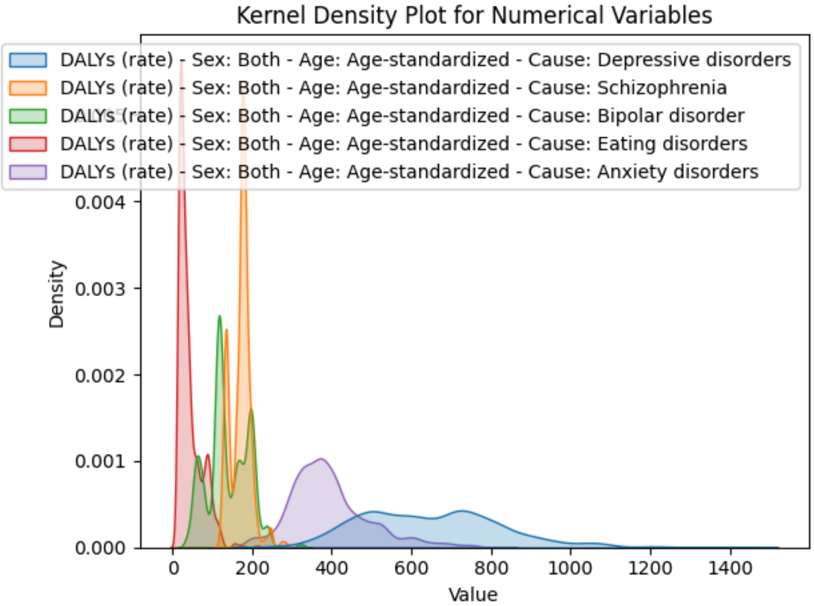
```
# Visualize the distribution using histograms
print("\nHistograms for Numerical Variables:")
data[numerical_variables].hist(figsize=(12, 8))
plt.xlabel("Values")
plt.ylabel("Frequency")
plt.title("Histograms for Numerical Variables")
plt.tight_layout()
plt.show()
```

```
Histograms for Numerical Variables:
```



Inference: Depressive Disorders is most prevelant, followed by Anxiety Disorders, Bipolor Disorders, Schizphrenia and Eating Disorder.

```
# Kernel Density Plots
sns.kdeplot(data=data[numerical_variables], fill=True)
plt.title("Kernel Density Plot for Numerical Variables")
plt.xlabel("Value")
plt.ylabel("Density")
plt.tight_layout()
plt.show()
```



Inference: Schizopherphrenia, Bipolor and Eating Idsorder have spiked density from 0 to 300, while Axiety slightly from 250 to 550, and lastly Dpressive Disorder comparatively stable from 350 to 900.

For Categorical Variables:

```
# Univariate Analysis for Categorical Variables
# Display frequency tables showing counts and percentages
categorical_variables = ['Entity', 'Code', 'Year']
for variable in categorical_variables:
    print(f"\nFrequency Table for '{variable}':")
    print(data[variable].value_counts(normalize=True))
```

```
Frequency Table for 'Entity':
Afghanistan          0.004386
Nigeria              0.004386
North America (WB)   0.004386
North Korea          0.004386
North Macedonia      0.004386
                       ...
Grenada              0.004386
Guam                 0.004386
Guatemala            0.004386
Guinea               0.004386
Zimbabwe             0.004386
Name: Entity, Length: 228, dtype: float64

Frequency Table for 'Code':
AFG    0.004878
PNG    0.004878
NIU    0.004878
PRK    0.004878
MKD    0.004878
         ...
GRL    0.004878
GRD    0.004878
GUM    0.004878
GTM    0.004878
ZWE    0.004878
Name: Code, Length: 205, dtype: float64

Frequency Table for 'Year':
1990    0.033333
1991    0.033333
2018    0.033333
2017    0.033333
2016    0.033333
2015    0.033333
2014    0.033333
2013    0.033333
2012    0.033333
2011    0.033333
2010    0.033333
2009    0.033333
2008    0.033333
2007    0.033333
2006    0.033333
```

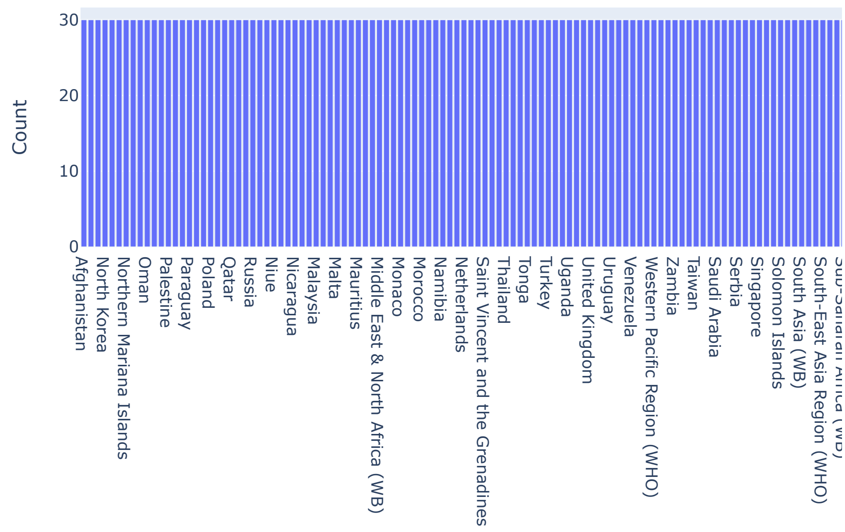
```
2005    0.033333
2004    0.033333
2003    0.033333
2002    0.033333
2001    0.033333
2000    0.033333
1999    0.033333
1998    0.033333
1997    0.033333
1996    0.033333
1995    0.033333
1994    0.033333
1993    0.033333
```

Inference:

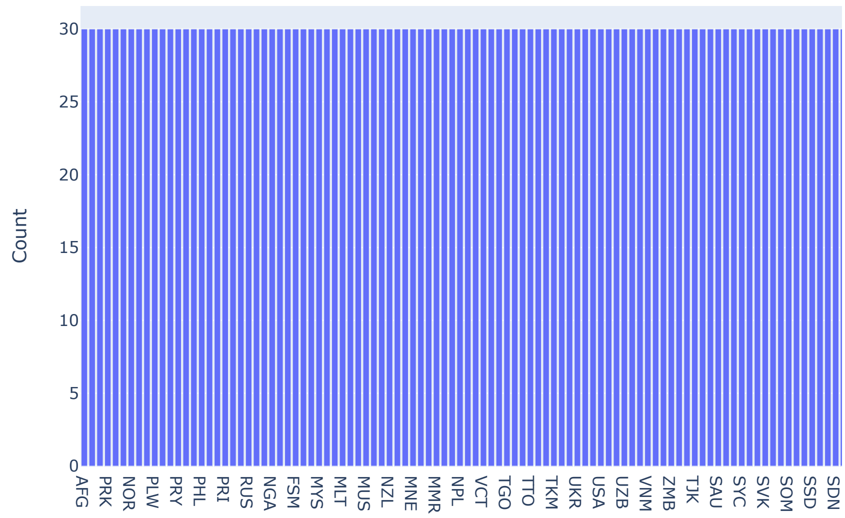
1. **Entity**: The dataset contains information about 228 different entities. The frequency of each entity is approximately 0.44%, indicating that each entity appears with roughly the same frequency in the dataset.
2. **Code**: There are 205 unique country codes in the dataset. Similar to the entity column, each country code appears with an approximate frequency of 0.49%.
3. **Year**: The dataset spans the years from 1990 to 2019. Each year from 1990 to 2019 appears with the same frequency of approximately 3.33%, indicating that the data is evenly distributed across these years.

```
# Visualize using bar plots with Plotly
for column in categorical_variables:
    fig = px.bar(data, x=data[column].value_counts().index, y=data[column].value_counts(),
                 labels={'x': column, 'y': 'Count'}, title=f'Bar Plot for {column}')
    fig.show()
```

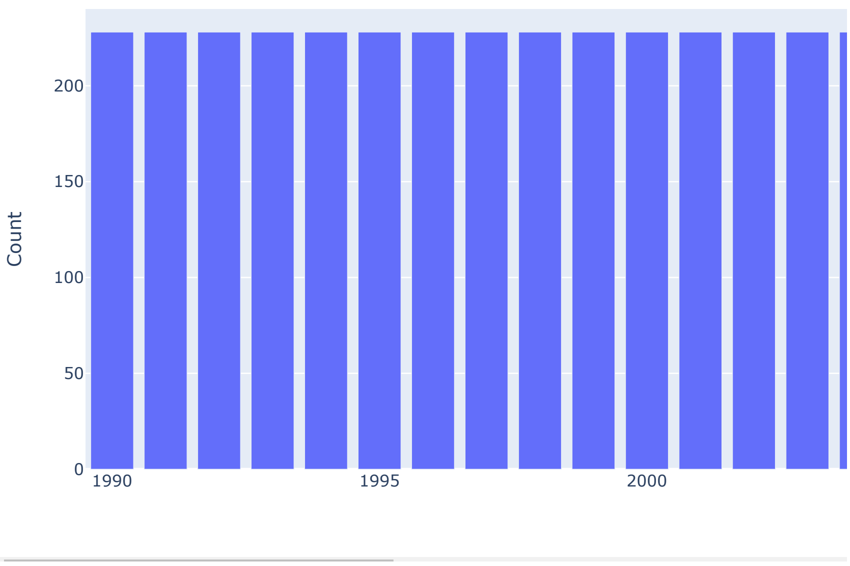
Bar Plot for Entity



Bar Plot for Code



Bar Plot for Year

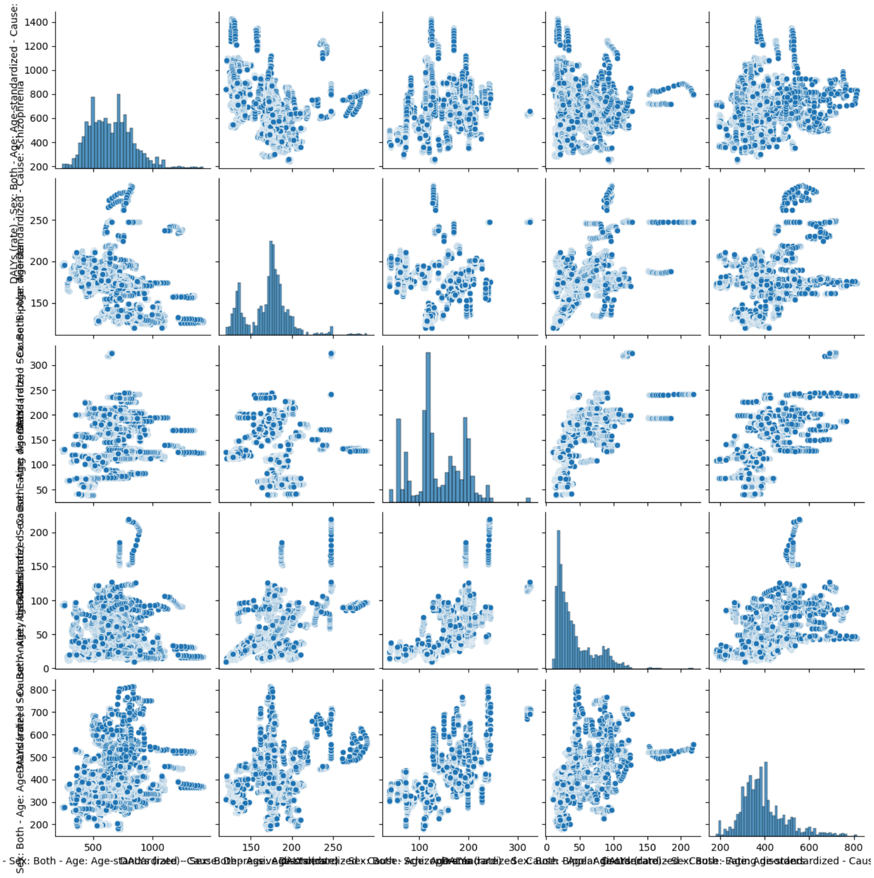


Inference:

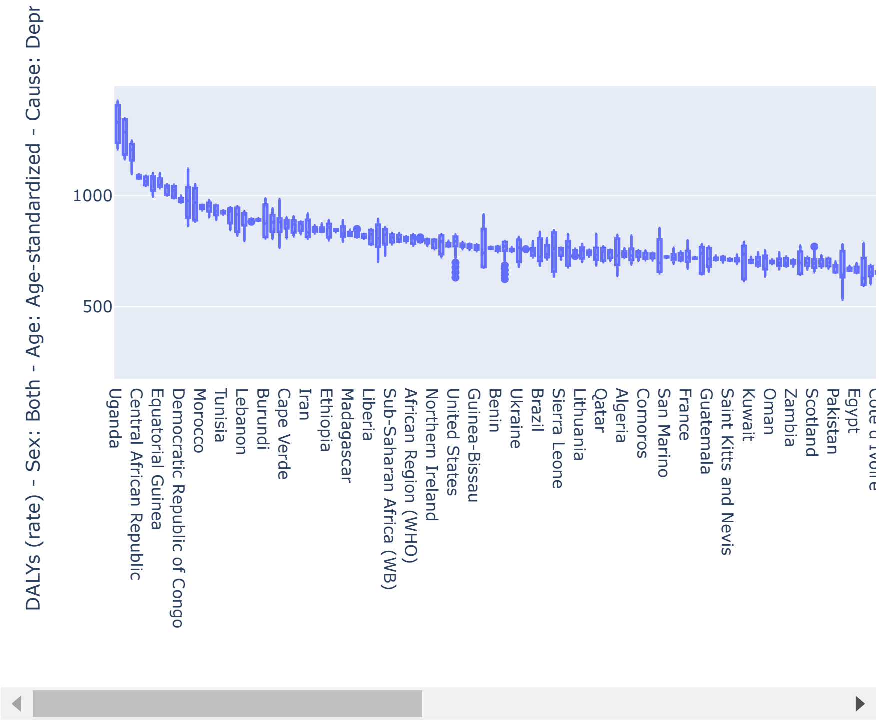
1. **Entity**: The dataset contains information about 228 different entities. The frequency of each entity is approximately 0.44%, indicating that each entity appears with roughly the same frequency in the dataset.
2. **Code**: There are 205 unique country codes in the dataset. Similar to the entity column, each country code appears with an approximate frequency of 0.49%.
3. **Year**: The dataset spans the years from 1990 to 2019. Each year from 1990 to 2019 appears with the same frequency of approximately 3.33%, indicating that the data is evenly distributed across these years.

Bivariate Analysis

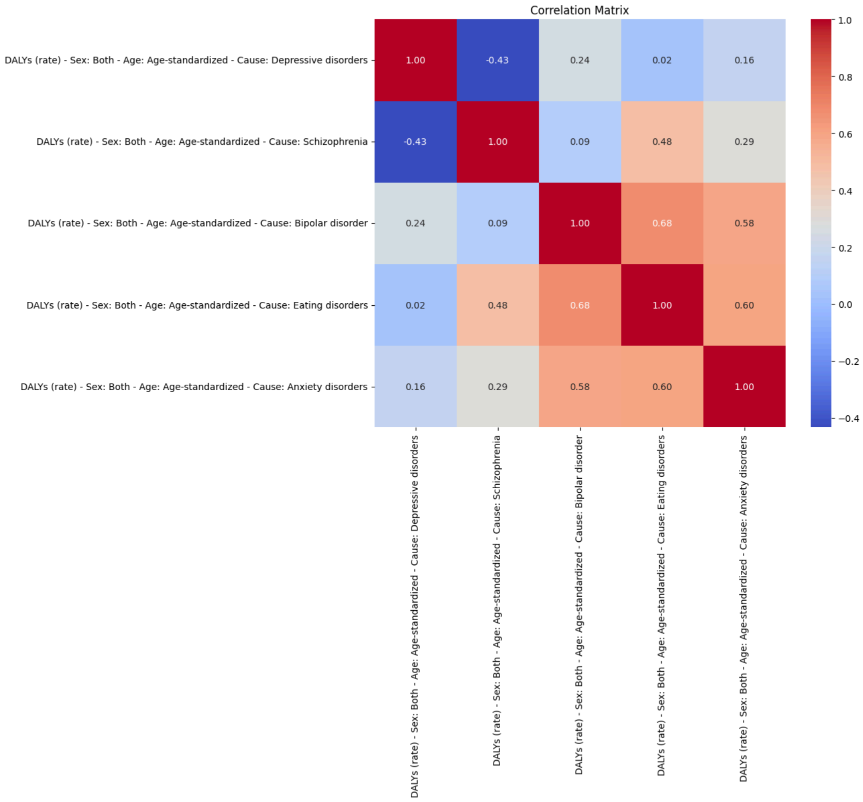
```
#Explore relationships between pairs of numerical variables using scatter plots
sns.pairplot(data[numerical_variables])
plt.show()
```



```
# Box plot for numerical variables with categorical variable 'Entity' using Plotly
fig = px.box(data, x='Entity', y='DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Depressive disorders')
fig.update_layout(xaxis={'categoryorder':'total descending'})
fig.show()
```




```
# Calculate correlation matrix
correlation_matrix = data[numerical_variables].corr()

# Heatmap for correlation matrix
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix')
plt.show()
```



Inference:

Positive Correlations:

- There is a moderate positive correlation (0.68) between Bipolar disorder and Eating disorders.
- There is a moderate positive correlation (0.6) between Bipolar disorder and Anxiety disorders.
- There is also a moderate positive correlation (0.58) between Eating disorders and Anxiety disorders.

Negative Correlations:

- There is a moderate negative correlation (-0.43) between Depressive disorders and Schizophrenia".

Drop the non-required columns / features (dependent columns)

```
# Reason: Drop columns 'Entity' and 'Code' as they are identifiers and not required for analysis
data_dropped = data.drop(['Entity', 'Code'], axis=1)
print("Data after dropping non-required columns:")
print(data_dropped.head())
```

```
Data after dropping non-required columns:
   Year  \
0  1990
1  1991
2  1992
3  1993
4  1994

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Depressive disorders  \
0                                          895.22565
1                                          893.88434
2                                          892.34973
3                                          891.51587
4                                          891.39160

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Schizophrenia  \
0                                          138.24825
1                                          137.76122
2                                          137.08030
3                                          136.48602
4                                          136.18323

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Bipolar disorder  \
0                                          147.64412
1                                          147.56696
2                                          147.13086
3                                          146.78812
4                                          146.58481

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Eating disorders  \
0                                          26.471115
1                                          25.548681
2                                          24.637949
3                                          23.863169
4                                          23.189074

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Anxiety disorders
0                                          440.33000
1                                          439.47202
2                                          437.60718
3                                          436.69104
4                                          436.76800
```

Re-arrange columns / features (if required)

```
# Reason: Move the 'Year' column to the front for better readability and understanding of temporal aspect
data_reordered = data_dropped[['Year'] + [col for col in data_dropped.columns if col != 'Year']]
print("\nData after re-arranging columns:")
print(data_reordered.head())
```

```
Data after re-arranging columns:
   Year  \
0  1990
1  1991
2  1992
3  1993
4  1994

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Depressive disorders  \
0                                          895.22565
1                                          893.88434
2                                          892.34973
3                                          891.51587
4                                          891.39160

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Schizophrenia  \
0                                          138.24825
1                                          137.76122
2                                          137.08030
3                                          136.48602
4                                          136.18323

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Bipolar disorder  \
0                                          147.64412
1                                          147.56696
2                                          147.13086
3                                          146.78812
4                                          146.58481

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Eating disorders  \
0                                          26.471115
```

```
1                                         25.548681
2                                         24.637949
3                                         23.863169
4                                         23.189074

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Anxiety disorders
0                                         440.33000
1                                         439.47202
2                                         437.60718
3                                         436.69104
4                                         436.76800
```

Separate the features (X) and target variable (y)

```
# Reason: Separate the features from the target variable
X = data_reordered.drop('DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Depressive disorders', axis=1)
y = data_reordered['DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Depressive disorders']
print("\nFeatures (X):")
print(X.head())
print("\nTarget Variable (y):")
print(y.head())
```

```
Features (X):
   Year  \
0  1990
1  1991
2  1992
3  1993
4  1994

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Schizophrenia  \
0                                         138.24825
1                                         137.76122
2                                         137.08030
3                                         136.48602
4                                         136.18323

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Bipolar disorder  \
0                                         147.64412
1                                         147.56696
2                                         147.13086
3                                         146.78812
4                                         146.58481

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Eating disorders  \
0                                         26.471115
1                                         25.548681
2                                         24.637949
3                                         23.863169
4                                         23.189074

   DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Anxiety disorders
0                                         440.33000
1                                         439.47202
2                                         437.60718
3                                         436.69104
4                                         436.76800

Target Variable (y):
0    895.22565
1    893.88434
2    892.34973
3    891.51587
4    891.39160
Name: DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Depressive disorders, dtype: float64
```

Perform Standardization

```
# Reason: Standardize the numerical features for better model performance
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
print("\nFeatures after standardization:")
print(X_scaled[:5])  # Displaying first 5 rows after standardization
```

```
Features after standardization:
[[-1.67524673 -1.25197773  0.18974115 -0.54170292  0.47005204]
 [-1.55971247 -1.27054357  0.18823393 -0.57308652  0.46154147]
 [-1.44417822 -1.29650059  0.17971525 -0.60407199  0.44304352]
 [-1.32864396 -1.31915485  0.17302026 -0.63043201  0.43395603]
 [-1.2131097  -1.33069736  0.16904885 -0.65336647  0.43471942]]
```

Split the Training and Testing Dataset

```
# Reason: Split the data into training and testing sets to evaluate model performance
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
print("\nShape of Training Features (X_train):", X_train.shape)
print("Shape of Testing Features (X_test):", X_test.shape)
print("Shape of Training Target (y_train):", y_train.shape)
print("Shape of Testing Target (y_test):", y_test.shape)
```

```
Shape of Training Features (X_train): (5472, 5)
Shape of Testing Features (X_test): (1368, 5)
Shape of Training Target (y_train): (5472,)
Shape of Testing Target (y_test): (1368,)
```

Model K-NN with different ‘K’ values and give your inference

```
k_values = [3, 5, 7, 9]
for k in k_values:
    knn = KNeighborsClassifier(n_neighbors=k)
    knn.fit(X_train, y_train)
    y_pred = knn.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    print(f"\nKNN with K={k}")
    print("Accuracy:", accuracy)
```

The target variable y contains continuous values instead of discrete class labels, then we need to convert it into a categorical variable or ensure that we are using the correct target variable.

```
# Define the thresholds for categorizing DALY rates into classes
low_threshold = data['DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Depressive disorders'].quantile(0.33)
high_threshold = data['DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Depressive disorders'].quantile(0.67)


# Categorize DALY rates into classes
data['DALYs_class'] = pd.cut(data['DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Depressive disorders'],
                             bins=[-float('inf'), low_threshold, high_threshold, float('inf')],
                             labels=['low', 'medium', 'high'])


# Separate features (X) and target variable (y)
X = data[['Entity', 'Code', 'Year']]  # Features
y = data['DALYs_class']  # Target variable


# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```