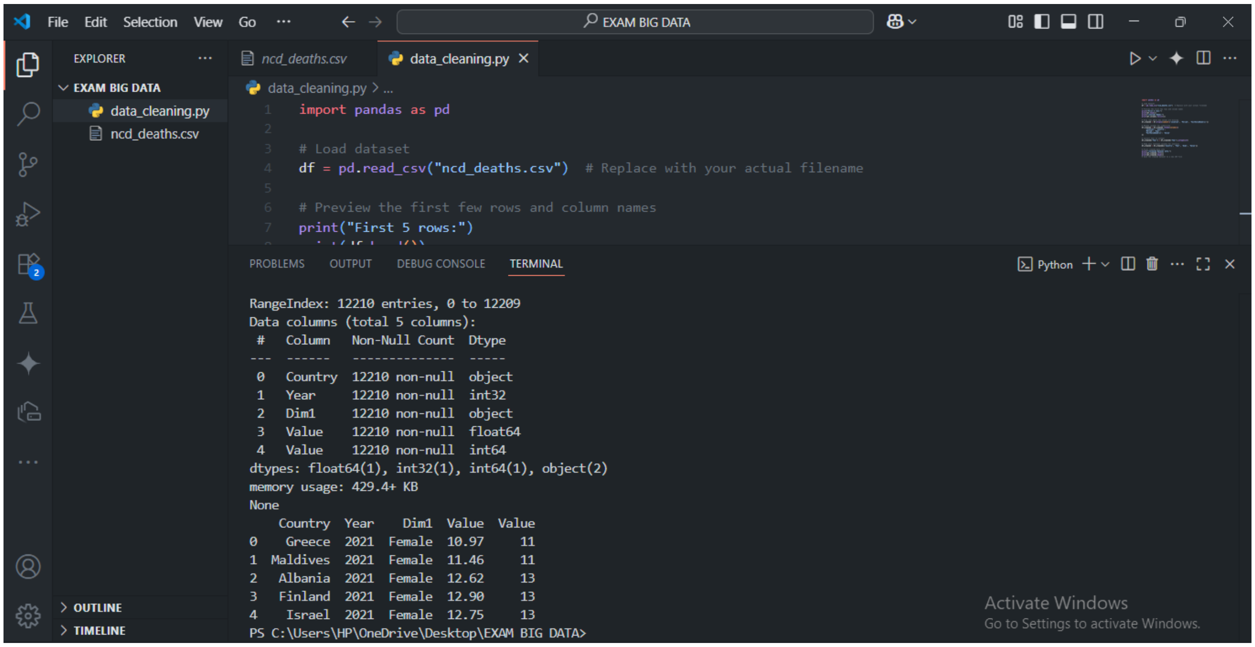
Names: MUHIRE Samuel

Id: 26092

INTRODUCTION TO BIG DATA EXAM SCREEN SHOOT TAKEN

1. Clean the Dataset ▪ Handle missing values, inconsistent formats, and outliers ▪ Apply necessary data transformations (e.g., encoding, scaling)
Data Cleaning

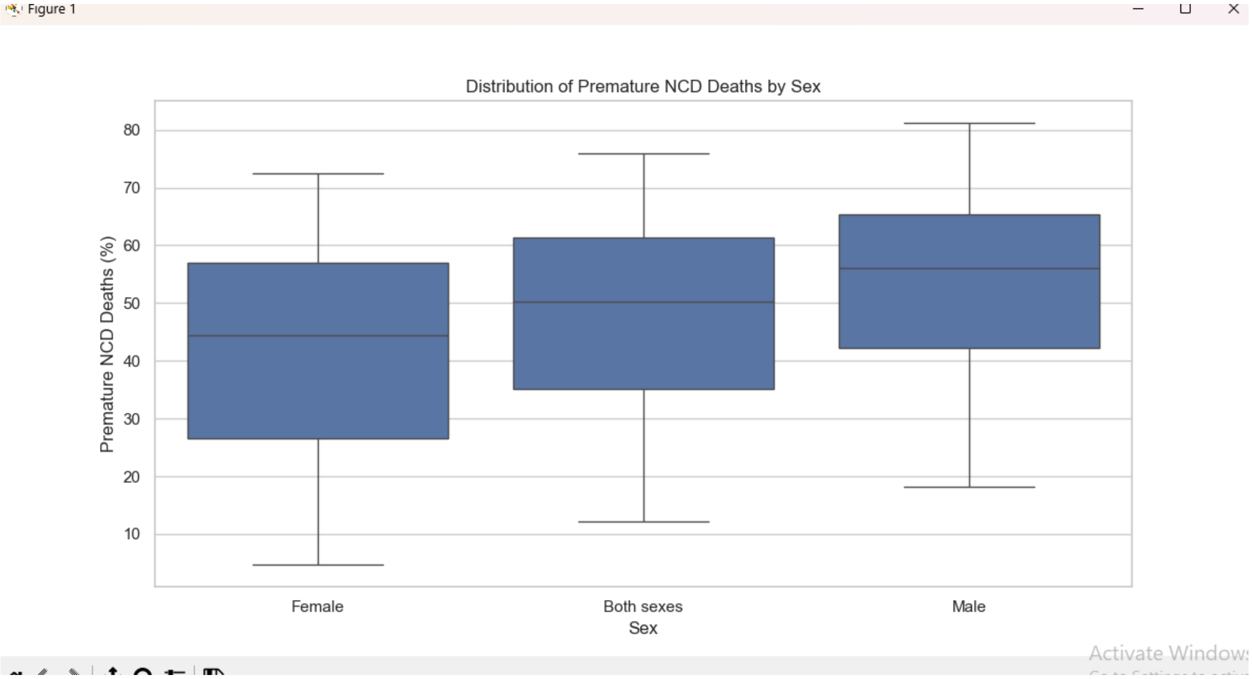
We start by loading the WHO NCD dataset and handling missing values, standardizing formats, and identifying outliers.

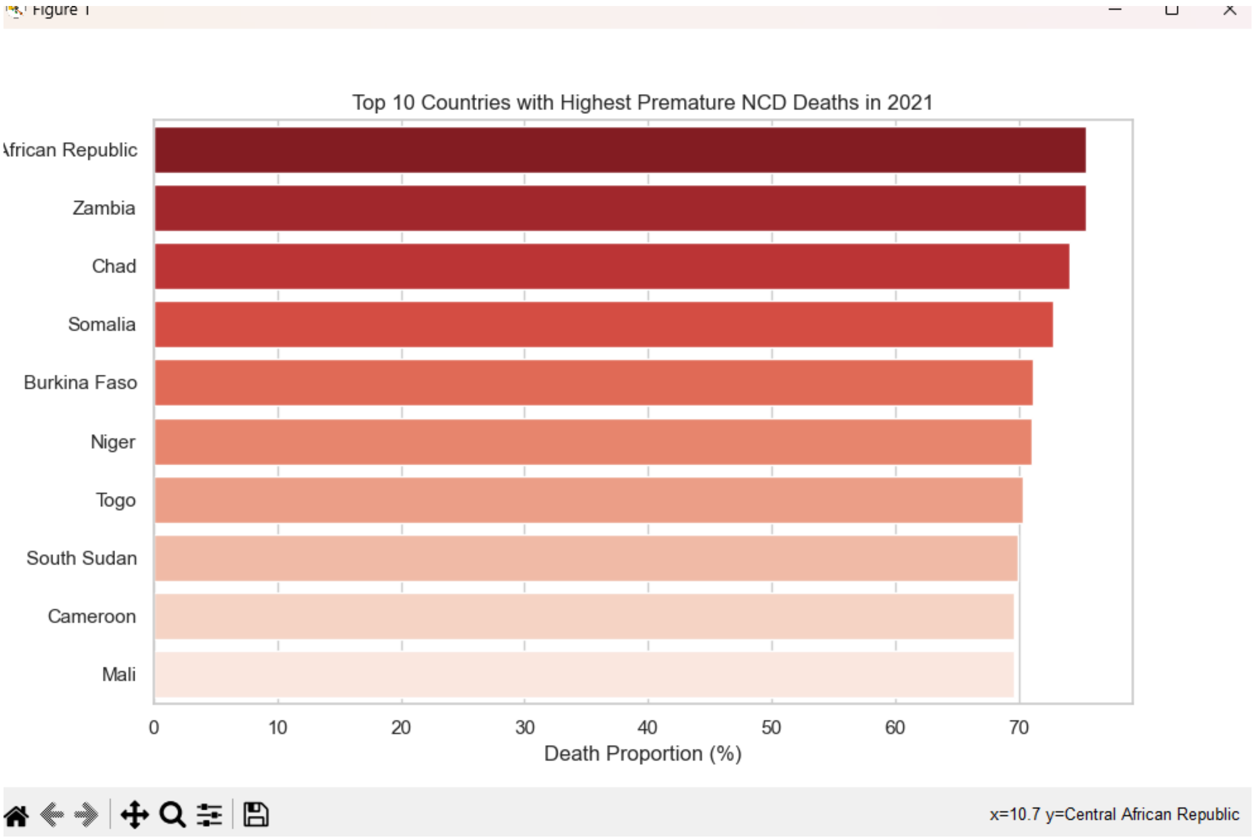


```python
import pandas as pd

# Load dataset
df = pd.read_csv("ncd_deaths.csv")  # Replace with your actual filename

# Preview the first few rows and column names
print("First 5 rows:")
```

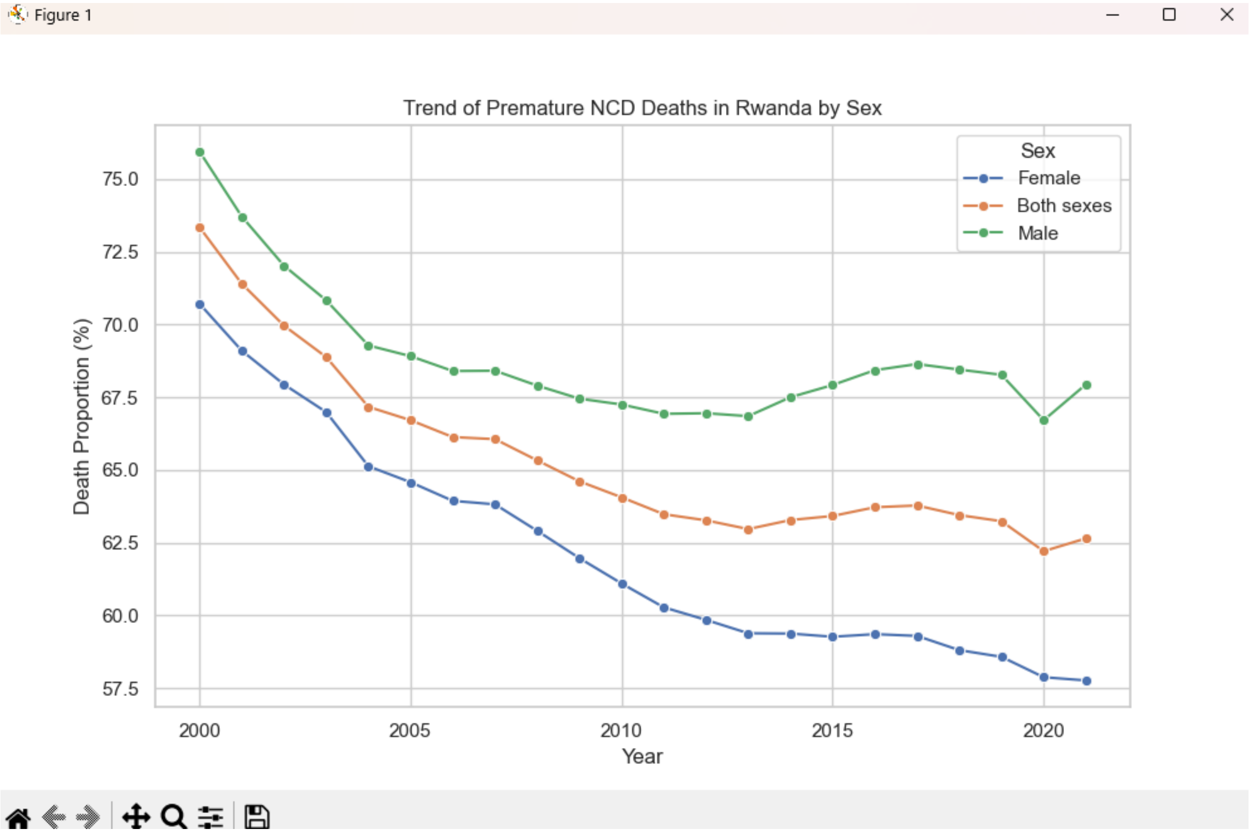
```
RangeIndex: 12210 entries, 0 to 12209
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   Country  12210 non-null  object
 1   Year     12210 non-null  int32
 2   Dim1     12210 non-null  object
 3   Value    12210 non-null  float64
 4   Value    12210 non-null  int64
dtypes: float64(1), int32(1), int64(1), object(2)
memory usage: 429.4+ KB
None
    Country  Year    Dim1  Value  Value
0    Greece  2021  Female  10.97     11
1  Maldives  2021  Female  11.46     11
2   Albania  2021  Female  12.62     13
3   Finland  2021  Female  12.90     13
4    Israel  2021  Female  12.75     13
PS C:\Users\HP\OneDrive\Desktop\EXAM BIG DATA>
```

Conduct Exploratory Data Analysis (EDA) ▪ Generate descriptive statistics ▪ Visualize distributions and relationships among variables





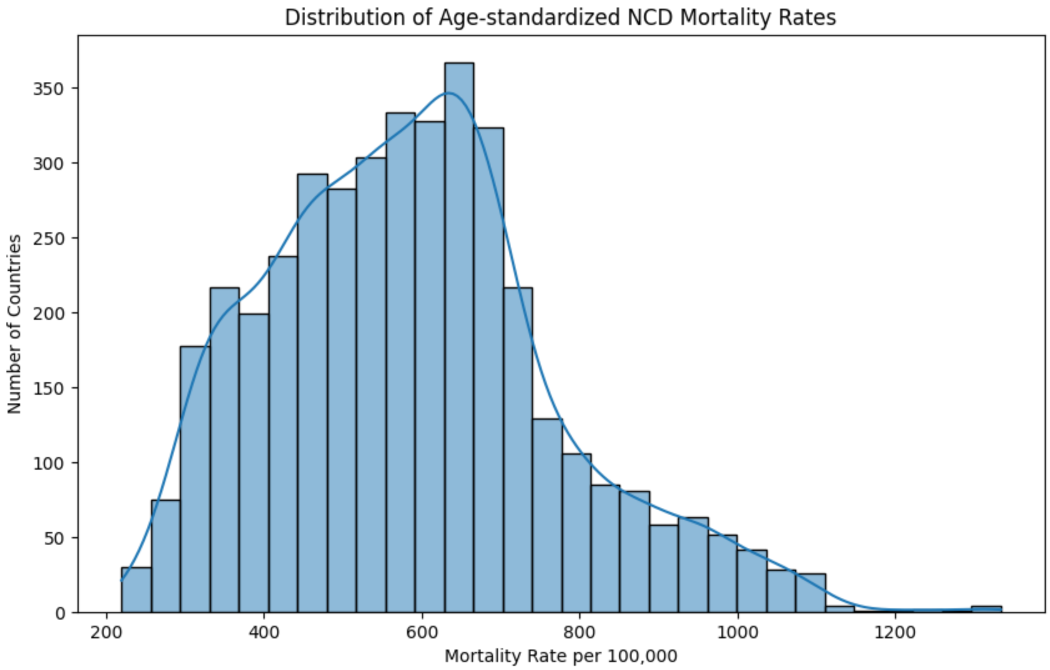
x=10.7 y=Central African Republic

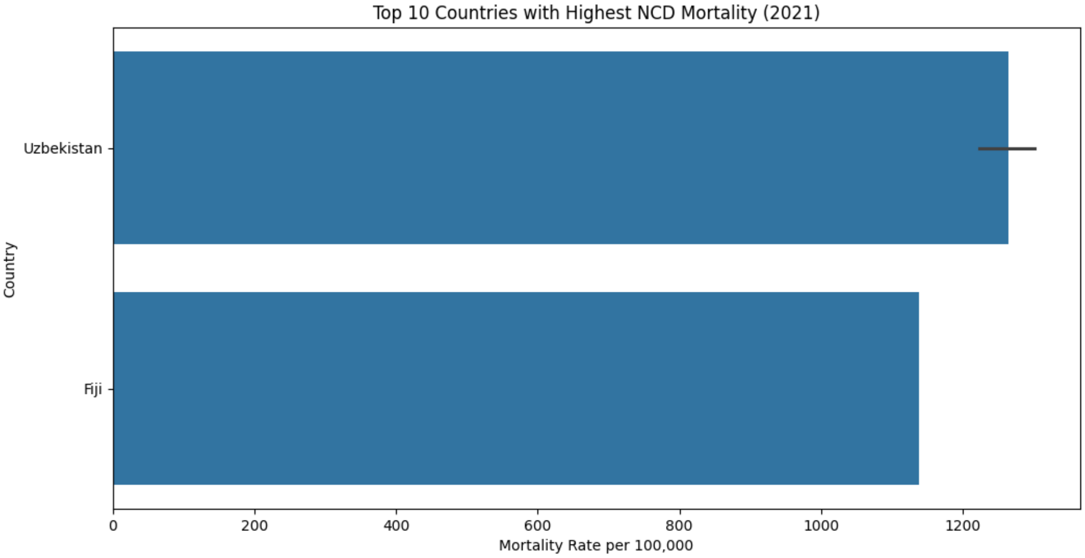


Exploratory Data Analysis (EDA)

Objectives:

- Understand distributions
- Compare countries/regions
- Highlight Rwanda



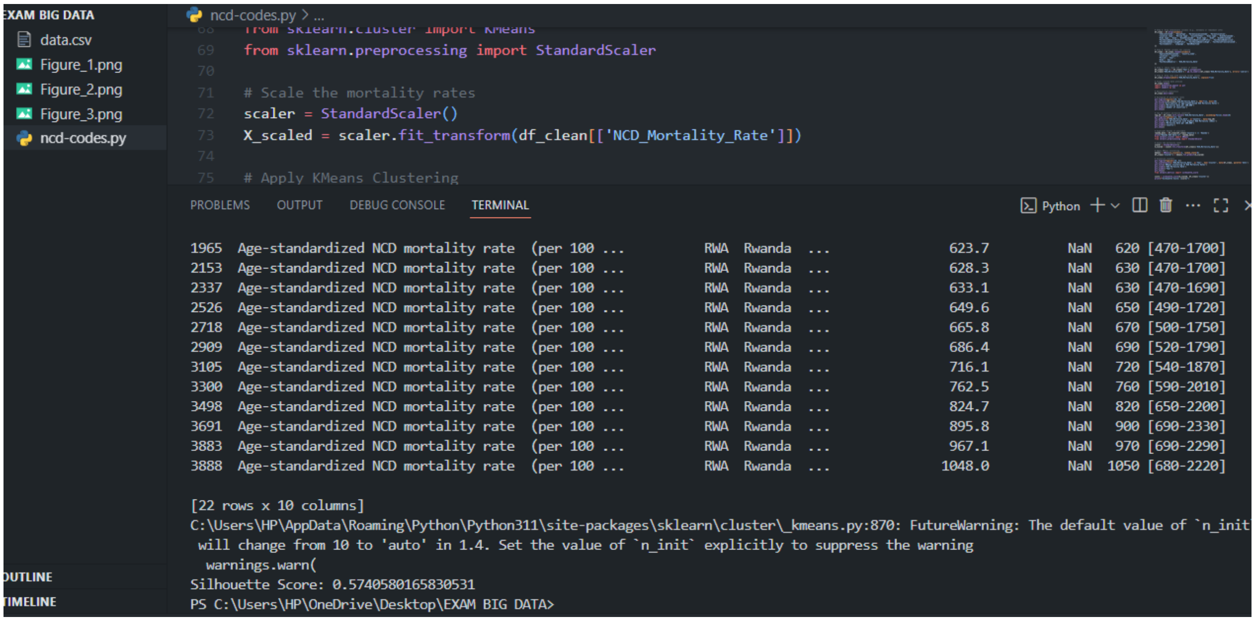


3. Apply a Machine Learning or Clustering Model

Because this is not a predictive task, we can use **Clustering (e.g., KMeans)** to group countries by mortality levels.

Evaluate the Model

For clustering, we use **silhouette score**.



```python
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# Scale the mortality rates
scaler = StandardScaler()
X_scaled = scaler.fit_transform(df_clean[['NCD_Mortality_Rate']])

# Apply KMeans Clustering
```

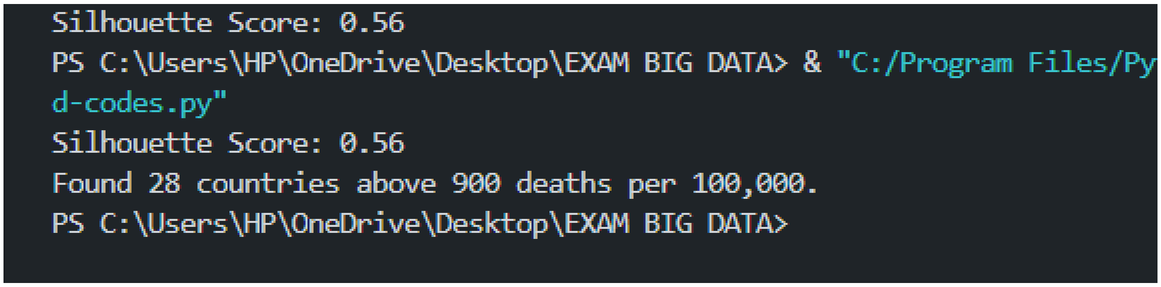
```
1965  Age-standardized NCD mortality rate  (per 100 ...        RWA  Rwanda  ...             623.7        NaN   620 [470-1700]
2153  Age-standardized NCD mortality rate  (per 100 ...        RWA  Rwanda  ...             628.3        NaN   630 [470-1700]
2337  Age-standardized NCD mortality rate  (per 100 ...        RWA  Rwanda  ...             633.1        NaN   630 [470-1690]
2526  Age-standardized NCD mortality rate  (per 100 ...        RWA  Rwanda  ...             649.6        NaN   650 [490-1720]
2718  Age-standardized NCD mortality rate  (per 100 ...        RWA  Rwanda  ...             665.8        NaN   670 [500-1750]
2909  Age-standardized NCD mortality rate  (per 100 ...        RWA  Rwanda  ...             686.4        NaN   690 [520-1790]
3105  Age-standardized NCD mortality rate  (per 100 ...        RWA  Rwanda  ...             716.1        NaN   720 [540-1870]
3300  Age-standardized NCD mortality rate  (per 100 ...        RWA  Rwanda  ...             762.5        NaN   760 [590-2010]
3498  Age-standardized NCD mortality rate  (per 100 ...        RWA  Rwanda  ...             824.7        NaN   820 [650-2200]
3691  Age-standardized NCD mortality rate  (per 100 ...        RWA  Rwanda  ...             895.8        NaN   900 [690-2330]
3883  Age-standardized NCD mortality rate  (per 100 ...        RWA  Rwanda  ...             967.1        NaN   970 [690-2290]
3888  Age-standardized NCD mortality rate  (per 100 ...        RWA  Rwanda  ...            1048.0        NaN  1050 [680-2220]

[22 rows x 10 columns]
C:\Users\HP\AppData\Roaming\Python\Python311\site-packages\sklearn\cluster\_kmeans.py:870: FutureWarning: The default value of `n_init
 will change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
Silhouette Score: 0.5740580165830531
PS C:\Users\HP\OneDrive\Desktop\EXAM BIG DATA>
```

5. Structure Your Code with Markdown & Functions

Example of reusable function:

6

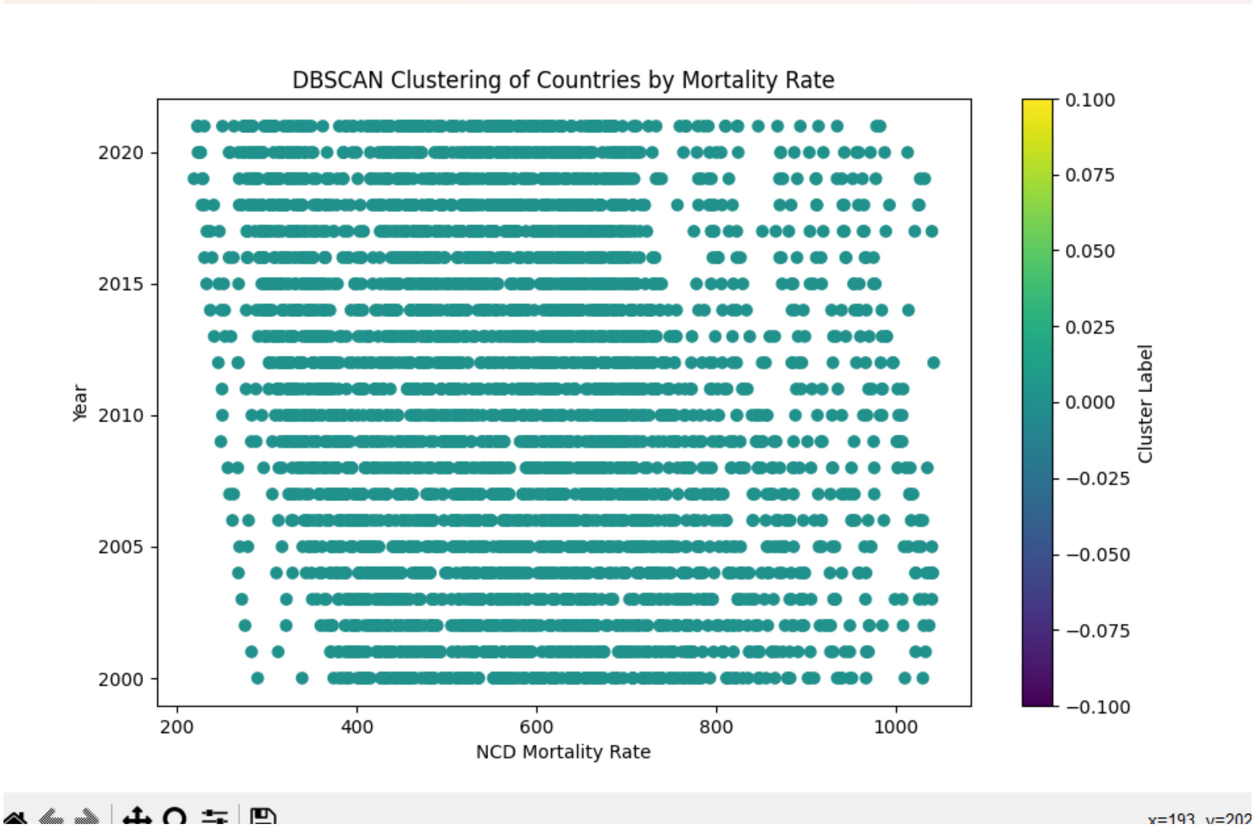
1. Custom Function: Highlight Countries Exceeding a Threshold



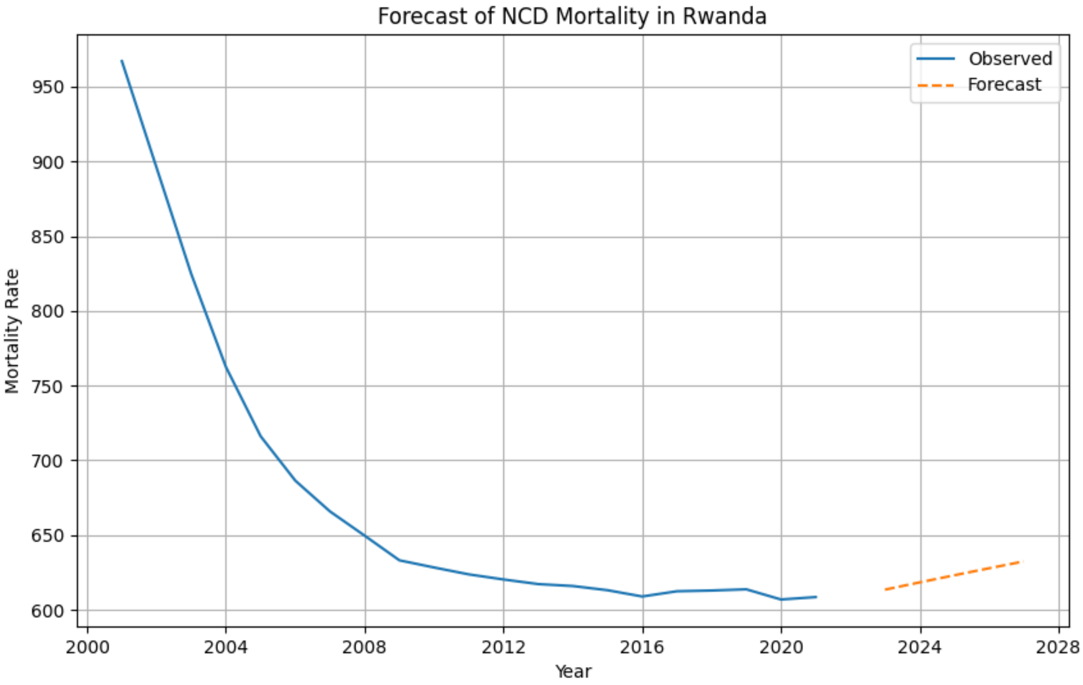
```
Silhouette Score: 0.56
PS C:\Users\HP\OneDrive\Desktop\EXAM BIG DATA> & "C:/Program Files/Py
d-codes.py"
Silhouette Score: 0.56
Found 28 countries above 900 deaths per 100,000.
PS C:\Users\HP\OneDrive\Desktop\EXAM BIG DATA>
```

DBSCAN Clustering for High-Risk Region Detection

Python



Time Series Forecasting with ARIMA for Rwanda


Forecast of NCD Mortality in Rwanda
Observed
Forecast
Mortality Rate
950
900
850
800
750
700
650
600
Year
2000
2004
2008
2012
2016
2020
2024
2028