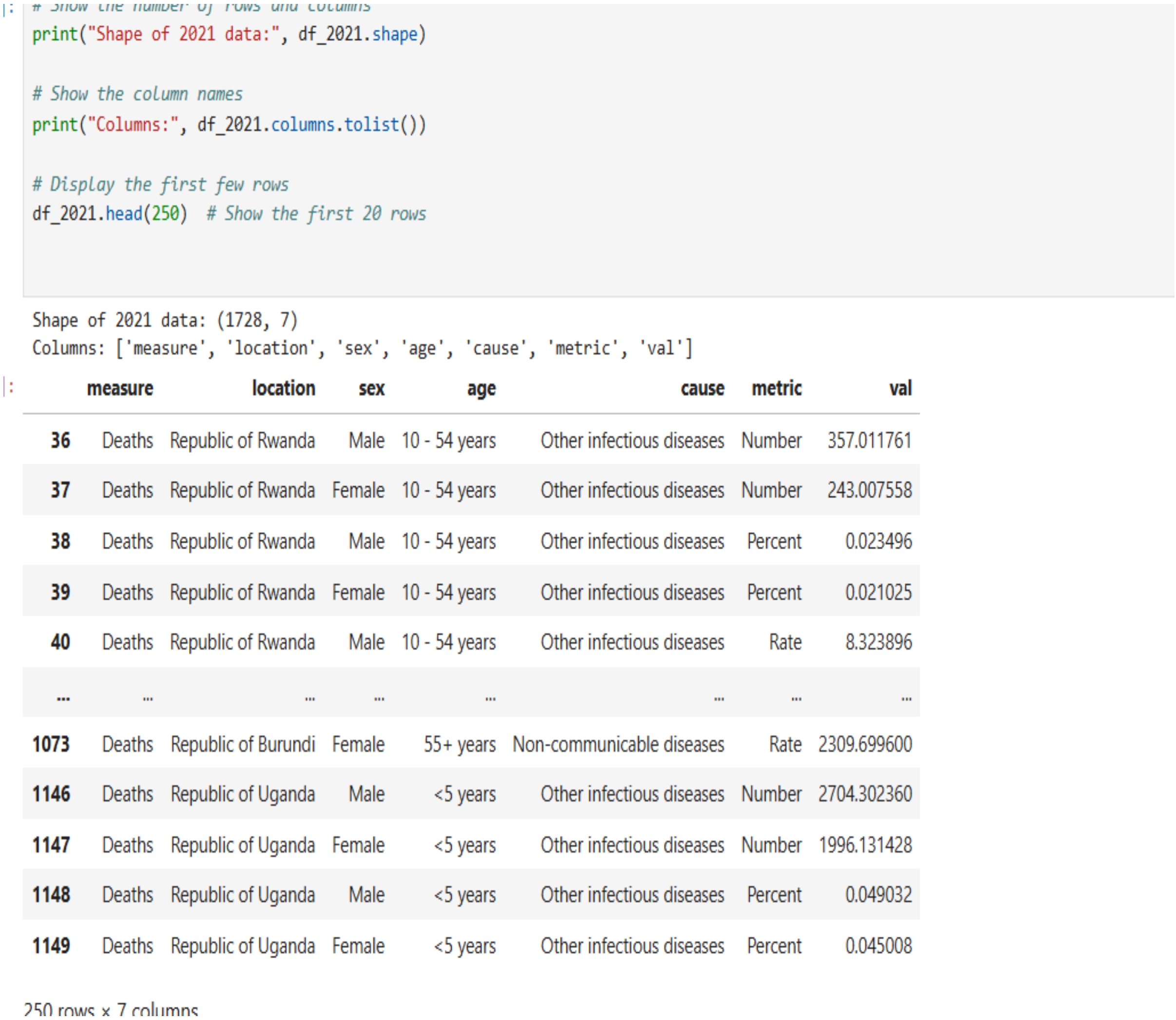
1. **Overview of the Cleaned Dataset**



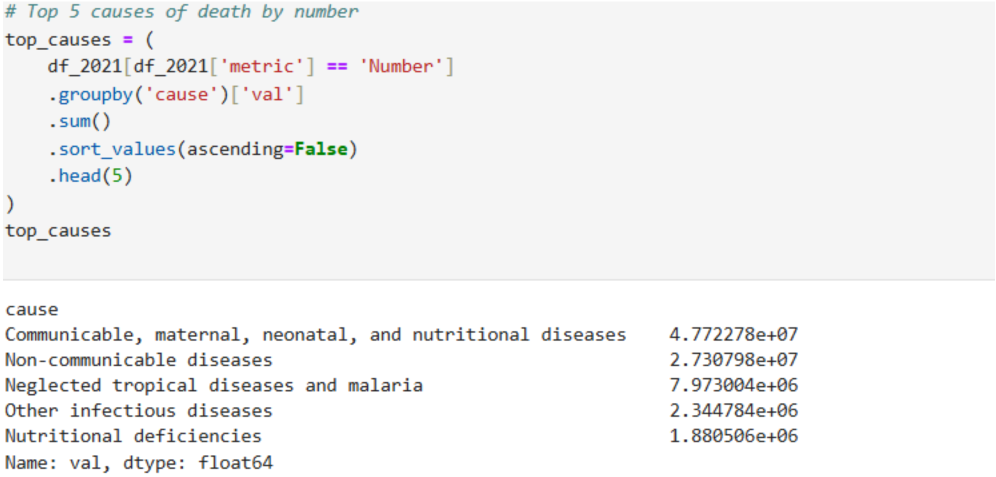
```
# Show the number of rows and columns
print("Shape of 2021 data:", df_2021.shape)

# Show the column names
print("Columns:", df_2021.columns.tolist())

# Display the first few rows
df_2021.head(250)  # Show the first 20 rows
```

```
Shape of 2021 data: (1728, 7)
Columns: ['measure', 'location', 'sex', 'age', 'cause', 'metric', 'val']
```

	measure	location	sex	age	cause	metric	val
36	Deaths	Republic of Rwanda	Male	10 - 54 years	Other infectious diseases	Number	357.011761
37	Deaths	Republic of Rwanda	Female	10 - 54 years	Other infectious diseases	Number	243.007558
38	Deaths	Republic of Rwanda	Male	10 - 54 years	Other infectious diseases	Percent	0.023496
39	Deaths	Republic of Rwanda	Female	10 - 54 years	Other infectious diseases	Percent	0.021025
40	Deaths	Republic of Rwanda	Male	10 - 54 years	Other infectious diseases	Rate	8.323896
...	...	...	...	...	...	...	...
1073	Deaths	Republic of Burundi	Female	55+ years	Non-communicable diseases	Rate	2309.699600
1146	Deaths	Republic of Uganda	Male	<5 years	Other infectious diseases	Number	2704.302360
1147	Deaths	Republic of Uganda	Female	<5 years	Other infectious diseases	Number	1996.131428
1148	Deaths	Republic of Uganda	Male	<5 years	Other infectious diseases	Percent	0.049032
1149	Deaths	Republic of Uganda	Female	<5 years	Other infectious diseases	Percent	0.045008

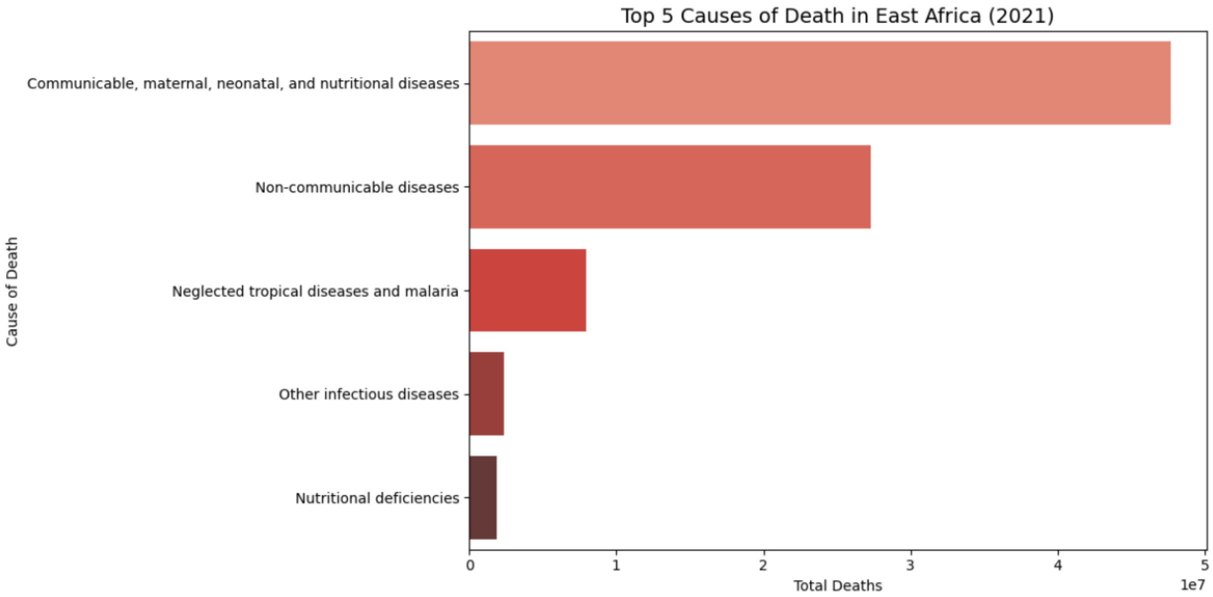
250 rows × 7 columns

2. **Top 5 Causes of Death Chart**



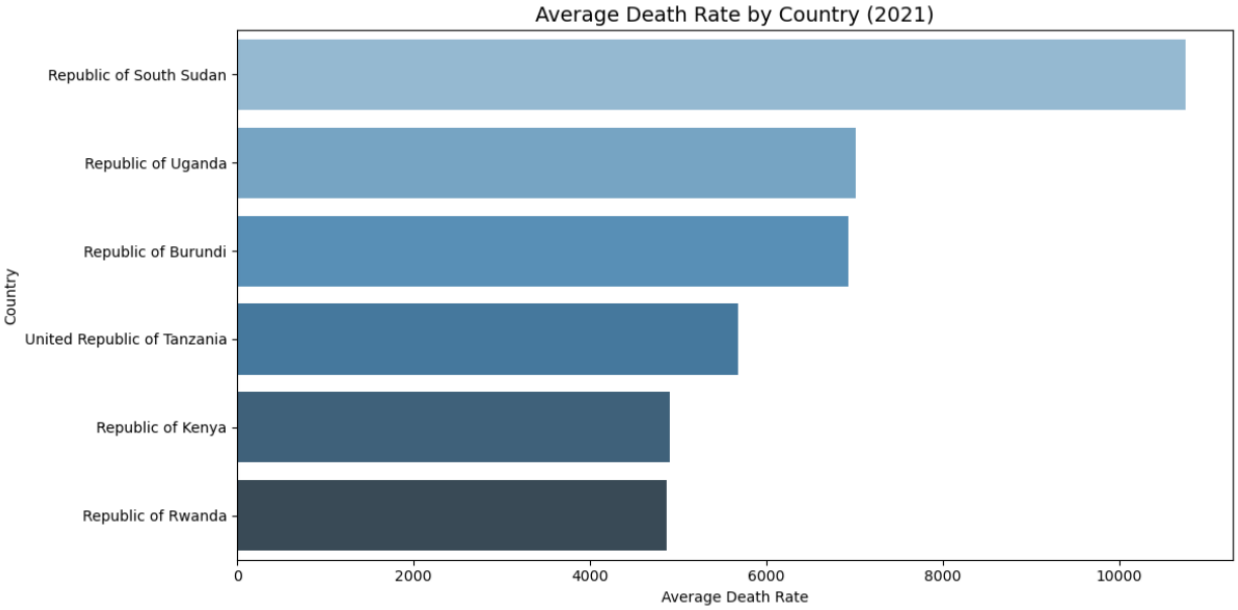
```
# Top 5 causes of death by number
top_causes = (
    df_2021[df_2021['metric'] == 'Number']
    .groupby('cause')['val']
    .sum()
    .sort_values(ascending=False)
    .head(5)
)
top_causes
```

```
cause
Communicable, maternal, neonatal, and nutritional diseases    4.772278e+07
Non-communicable diseases                                     2.730798e+07
Neglected tropical diseases and malaria                       7.973004e+06
Other infectious diseases                                     2.344784e+06
Nutritional deficiencies                                      1.880506e+06
Name: val, dtype: float64
```

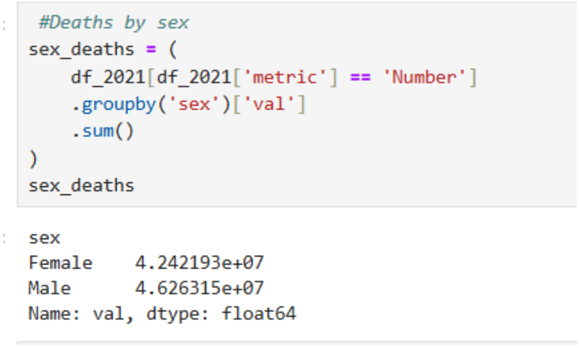


3.

4. **Death Rate by Country Chart**

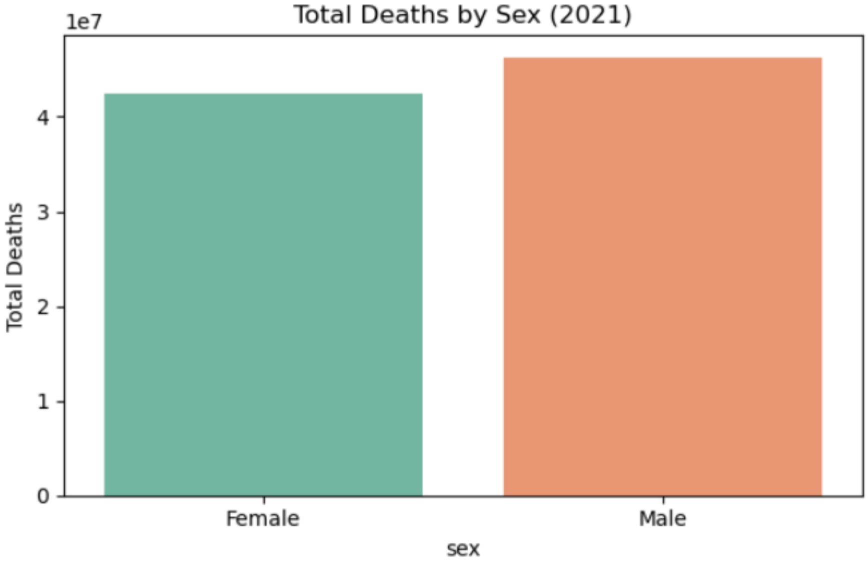


5. **Deaths by Sex**

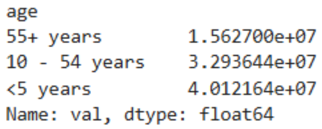


```
#Deaths by sex
sex_deaths = (
    df_2021[df_2021['metric'] == 'Number']
    .groupby('sex')['val']
    .sum()
)
sex_deaths
```

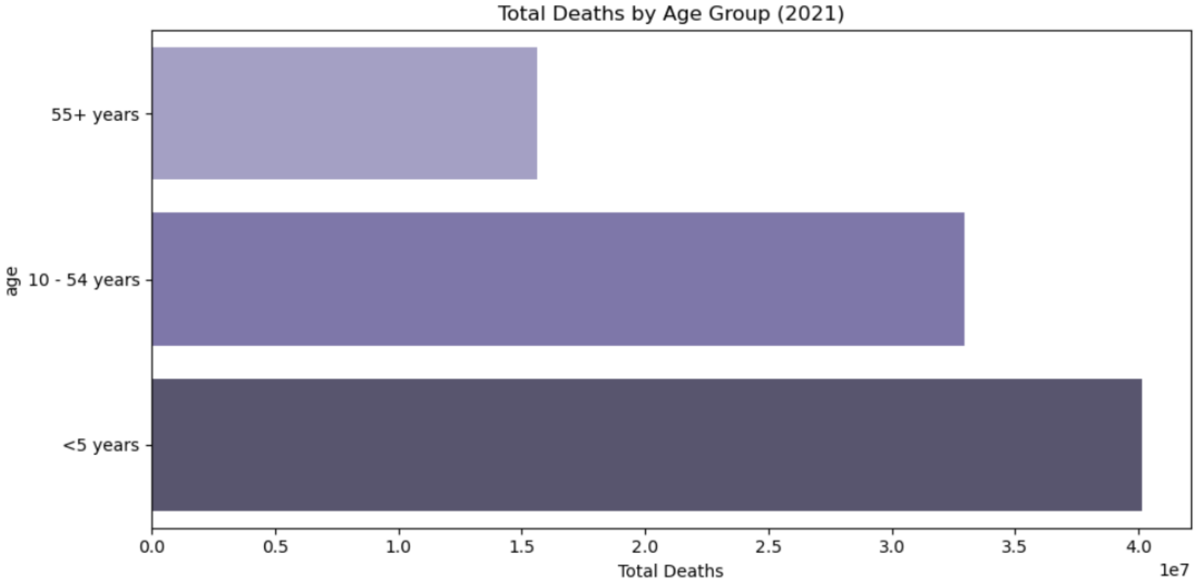
```
sex
Female    4.242193e+07
Male      4.626315e+07
Name: val, dtype: float64
```



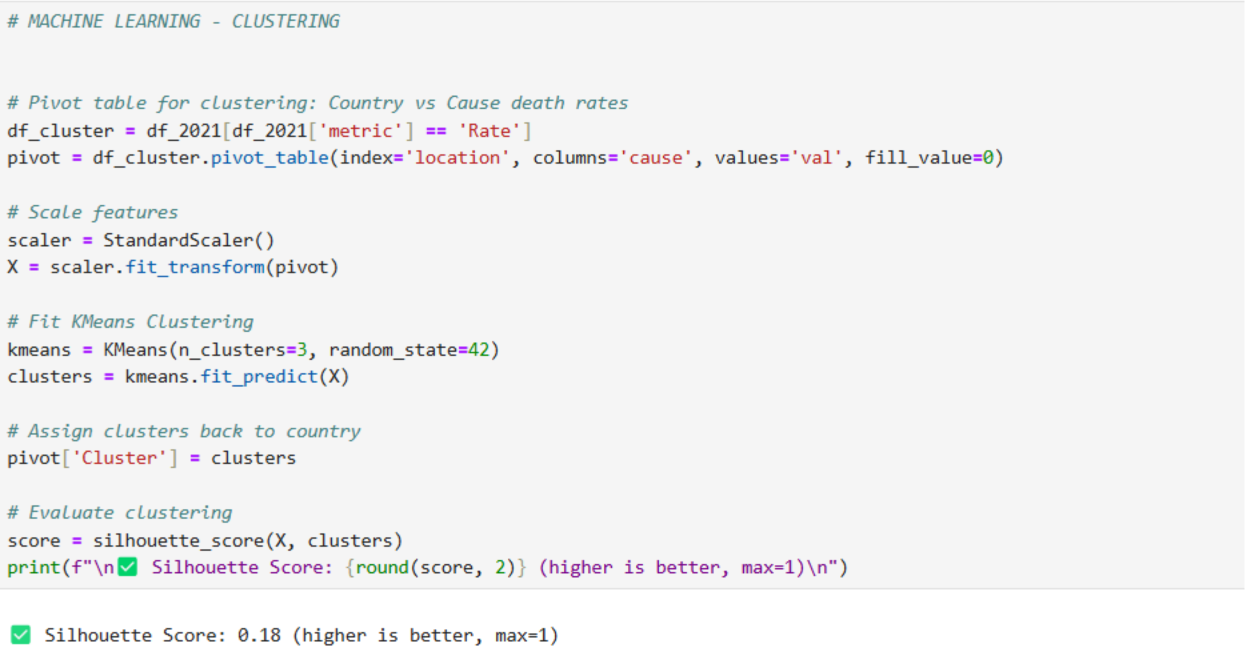
6. **Deaths by Age Group**



```
age
55+ years        1.562700e+07
10 - 54 years    3.293644e+07
<5 years         4.012164e+07
Name: val, dtype: float64
```

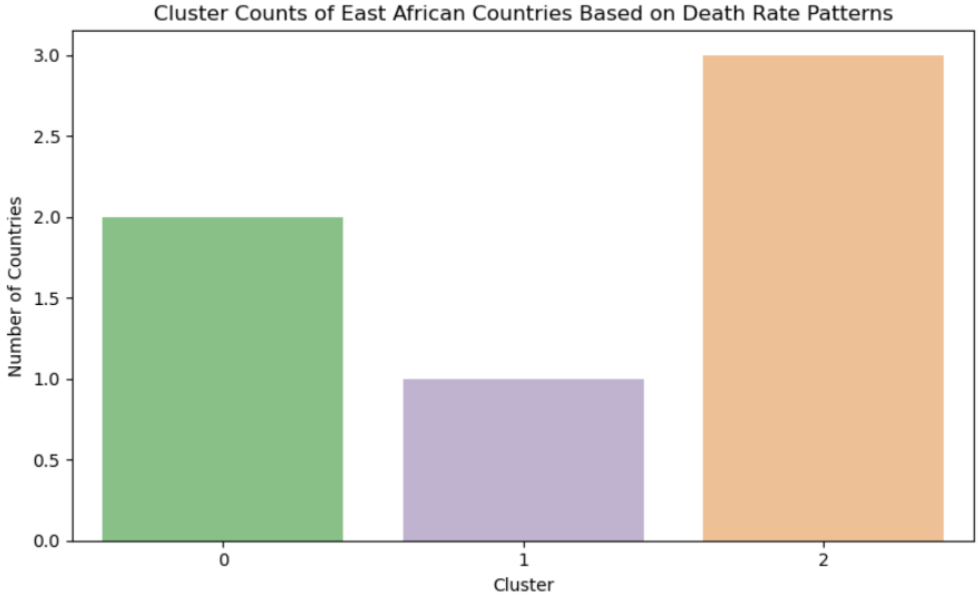


7. **Clustering Result**



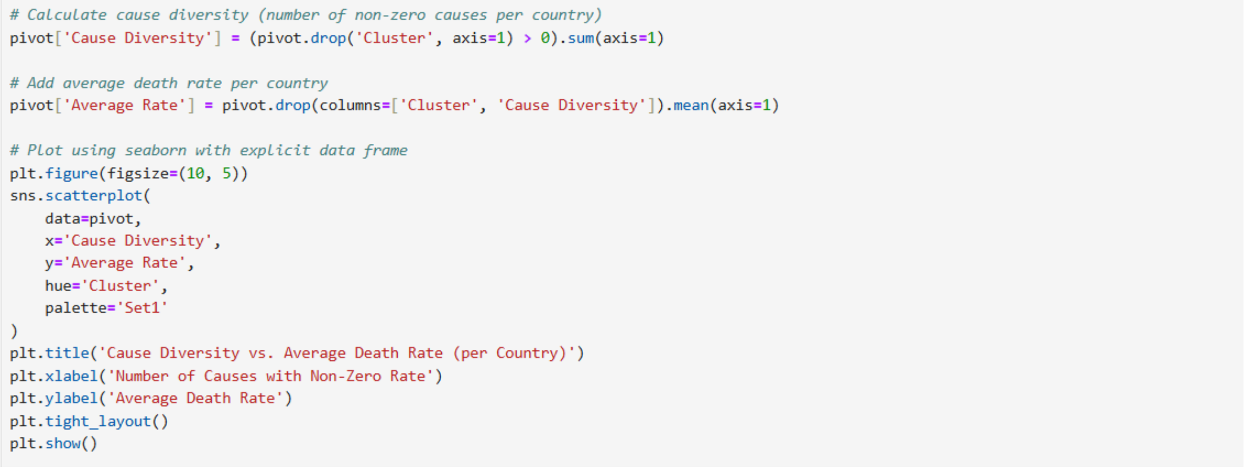
```python
# MACHINE LEARNING - CLUSTERING


# Pivot table for clustering: Country vs Cause death rates
df_cluster = df_2021[df_2021['metric'] == 'Rate']
pivot = df_cluster.pivot_table(index='location', columns='cause', values='val', fill_value=0)

# Scale features
scaler = StandardScaler()
X = scaler.fit_transform(pivot)

# Fit KMeans Clustering
kmeans = KMeans(n_clusters=3, random_state=42)
clusters = kmeans.fit_predict(X)

# Assign clusters back to country
pivot['Cluster'] = clusters

# Evaluate clustering
score = silhouette_score(X, clusters)
print(f"\n✅ Silhouette Score: {round(score, 2)} (higher is better, max=1)\n")
```

```
✅ Silhouette Score: 0.18 (higher is better, max=1)
```

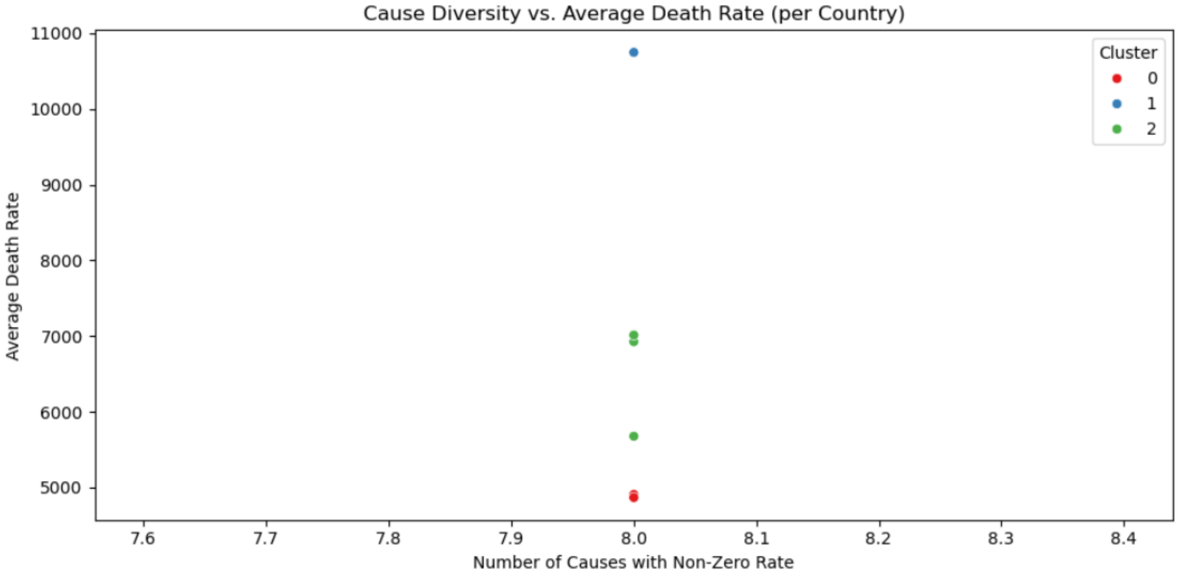


```
🌍 Country Cluster Assignments:

location
Republic of Kenya              0
Republic of Rwanda             0
Republic of South Sudan        1
Republic of Burundi            2
Republic of Uganda             2
United Republic of Tanzania    2
Name: Cluster, dtype: int32
```

8. Cause Diversity vs Avg Rate (Innovation Visual)



```python
# Calculate cause diversity (number of non-zero causes per country)
pivot['Cause Diversity'] = (pivot.drop('Cluster', axis=1) > 0).sum(axis=1)

# Add average death rate per country
pivot['Average Rate'] = pivot.drop(columns=['Cluster', 'Cause Diversity']).mean(axis=1)

# Plot using seaborn with explicit data frame
plt.figure(figsize=(10, 5))
sns.scatterplot(
    data=pivot,
    x='Cause Diversity',
    y='Average Rate',
    hue='Cluster',
    palette='Set1'
)
plt.title('Cause Diversity vs. Average Death Rate (per Country)')
plt.xlabel('Number of Causes with Non-Zero Rate')
plt.ylabel('Average Death Rate')
plt.tight_layout()
plt.show()
```


Cause Diversity vs. Average Death Rate (per Country)
Average Death Rate
Number of Causes with Non-Zero Rate
Cluster
● 0
● 1
● 2