Clustering the Caribbean 🌴

STUDENT ID | SA 01D01

Methodology of Data Collection

Task: Cluster the countries of the Caribbean based on Population, Land Size, Gross Domestic Product (GDP), and Crime Rate.

`Our research methodology defined "*The Caribbean*" specifically for our study, focusing on the thirteen sovereign island nations: *Antigua and Barbuda, Bahamas, Barbados, Cuba, Dominica, Dominican Republic, Grenada, Haiti, Jamaica, Saint Kitts and Nevis, St Lucia, St Vincent and the Grenadines, and Trinidad and Tobago.* We intentionally excluded data from Central American countries like Belize & Guyana that borders the Caribbean Sea, and dependent territories of the Caribbean such as Anguilla, Martinque, & Puerto Rico.

The decision to focus on sovereign states was based on the premise that these nations have unique economic, political, and social structures, which could significantly influence the metrics under study. The data collection process was comprehensive within these nations, involving information gathering from diverse sources like government databases, international organizations, and academic research. This data was then meticulously analyzed and cross-verified to ensure its accuracy and reliability.

The 2023 population data for the mentioned Caribbean countries were sourced from the International Monetary Fund (IMF) website, a trusted and reliable source for current economic and financial data. This lends significant credibility to our research. For consistency and accuracy, this data was cross verified with data in the Britannica Encyclopedia, an authoritative and reliable source established in 1768 with over 250 years of longevity. The consistency between these sources further validates the accuracy of our data. This also holds true for the data the researcher collected on the Land Size as well as the Gross Domestic Product (GDP) -with the exception for Cuba as no GDP data could be found for this country on the IMF website.

Whereas it speaks to the GDP data, the researcher contemplated using The Real GDP for it provides a more accurate measure of economic performance by accounting for differences in price levels between countries in the Caribbean. The formula used:

$$\text{Real GDP} = \frac{Nominal\ GDP}{1 + Inflation\ Rate}$$

where the Nominal GDP was in billions of US Dollars Annually

and the Inflation Rate was the Average Percentage Rate of Inflation represented as a decimal.

In the quest for finding GDP data on Cuba, the researcher turned to a trusted source: The Economist Intelligence Unit. This organization, with a rich history spanning 60 years, specializes in providing valuable insights to companies operating across international borders. As a member of The Economist Group, it has a reputation for delivering reliable information. This speaks to the credibility of this source. The researcher utilized data from the 2023 Cuba Report, a comprehensive study conducted by the Unit, to gather the necessary information for the economic analysis.
The researcher extracted the nominal GDP values in the report as well as the inflation rate and calculated the Real GDP of the country using the above-stated formula to come by the value for the Real GDP for Cuba.

Our research utilized crime rate data from the World Population Review website, a source frequently cited in Caribbean newspapers and articles, thus underscoring its credibility. However, we found that crime rates were only available for four of the report's Caribbean countries, indicating a need for further data collection and analysis for the remaining eight countries.

In-depth research into the specific crime rates for these remaining countries proved inconclusive, as local reports primarily referenced Homicide rates, which alone are insufficient to explain the composite metric of a country's overall Crime Rate. To address this data gap, we considered using a *proxy variable* to represent the Crime Rate within these remaining countries. We turned to NUMBEO, a website known for providing accurate and up-to-date information about various socio-economic factors across cities and countries worldwide. NUMBEO's credibility is highlighted by its frequent references in international newspapers and magazines like *BBC, Time, The Week, Forbes, The Economist, & Business Insider*.

NUMBEO's Crime Index, derived from visitor surveys, offers insights into perceptions of crime levels, safety, and concerns about specific crimes. This index is updated continuously, using data up to 36 months old, and is presented on a scale from 0 to 100. While this index provides a comparative tool for assessing safety across different locations, it's important to note that it's based on user-contributed data and perceptions, which may differ from official statistics. The methodology for Crime Rate data collection used on the NUMBEO site was found to be similar to that of the World Population Review website. Furthermore, the data obtained from NUMBEO showed a high correlation with the data from the World Population Review website, further attesting to the reliability of both sites. Therefore, we decided to use these values for the remaining countries in our study. This approach allowed us to maintain the integrity and comprehensiveness of our research.

Commented Code

# Importing all libraries

```
from google.colab import files

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import numpy as np

from sklearn.preprocessing import StandardScaler

from sklearn.cluster import KMeans

sns.set_theme(color_codes=True)
```

# Function that loads the data file.

```
def load_data(excel_document):

  # Loading the data

  df = pd.read_excel(excel_document)

  # Removing trailing and White spaces

  df.columns = df.columns.str.strip()

  return df
```

# Loading the Dataset

```
df = load_data('Clustering Data.xlsx')
```

# Printing the columns within the dataset

```
print(df.columns)
```

# Printing the Entire dataset to pandas Dataframe

```
df.head(14)

# Selecting the columns to use for clustering and storing them in a list called Features

features = ['Population', 'Land Size', 'Crime Rate', 'GDP']

X = df.loc[:, features].values


# Scaling the data to ensure fair distance computations

scaler = StandardScaler()

scaled_features = scaler.fit_transform(X)


# Printing the datatypes for each of the dataset columns.

print(df.dtypes)


# Checking the scaled features in the numpy array to ensure that there are fair distance computations.

scaled_features
```

## Showing the Elbow Method Graph

```
# To ignore all warnings that may show when this block of code is being runned

import warnings

warnings.filterwarnings("ignore")

# Creating a function that computes and visualizes the Elbow Method Graph

def find_optimal_k_elbow(X,max_k):

  inertias = []

  for k in range (1, max_k + 1):

    kmeans= KMeans(n_clusters=k, random_state= 42)

    kmeans.fit(X)

    inertias.append(kmeans.inertia_)

  # Plotting Elbow Graph

  plt.figure(figsize=(12,6))

  plt.plot(range(1,max_k + 1), inertias, marker = 'o')

  plt.xlabel('Number of Clusters (k)')

  plt.ylabel('Inertia')

  plt.title('Elbow Method' )

  plt.show()

# Finding the optimal k using the Elbow Method

max_k = 10

find_optimal_k_elbow(scaled_features, max_k)
```

# Showing the Silhouette Score Graph

```
# To ignore all warnings that may show when this block of code is being runned

import warnings

warnings.filterwarnings("ignore")

# Importing the sklearn's silhouette_score to compute and visualize the silhouette score

from sklearn.metrics import silhouette_score

# Creating a function that computues and visualizes the Silhouette Score

def find_optimal_k_silhouette(X , max_k):

  silhouette_scores = []

  for k in range(2 , max_k + 1):

    kmeans= KMeans(n_clusters=k, random_state= 42)

    kmeans.fit(X)

    labels = kmeans.labels_

    silhouette_scores.append(silhouette_score(X, labels))

  # Plotting Silhouette Score Graph

  plt.figure(figsize=(12,6))

  plt.plot(range(2,max_k + 1), silhouette_scores, marker = 'o')

  plt.xlabel('Number of Clusters (k)')

  plt.ylabel('Silhouette Score')

  plt.title('Silhouette Score' )

  plt.show()

# Finding the optimal k using the Silhouette Method

max_k = 10

find_optimal_k_silhouette(scaled_features, max_k)
```

## Making 5 Clusters

```
# Optimal number of clusters based on the elbow method

n_clusters = 5
```

# Performing KMeans Clustering with the optimal number of clusters as determined by the elbow method.

```
kmeans = KMeans(n_clusters = n_clusters, random_state=42)

df['label'] = kmeans.fit_predict(scaled_features)
```

# Importing the necessary packages and libraries for the cluster visualization

```
from mpl_toolkits.mplot3d import Axes3D

import matplotlib.colors as mcolors

# Creating the 3D visulization for the cluster visualization

fig = plt.figure(figsize=(10, 8))

ax = fig.add_subplot(111, projection='3d')

# Creating a colormap for the clusters

labels = df['label'].unique()

colors = plt.cm.get_cmap('tab10', len(labels))  # Using a colormap with enough colors for each of the clusters

# Creating a color map that maps each label to a unique color

color_map = {label: colors(i) for i, label in enumerate(labels)}

# Using 'Land Size' values to create sizes for the points

sizes = ((df['Land Size'] - df['Land Size'].min()) / (df['Land Size'].max() - df['Land Size'].min()) * 100) * 30  # Multiply by 30 to scale the size of the data points

# Plotting the data points

scatter = ax.scatter(df['Population'], df['GDP'], df['Crime Rate'], c=df['label'].map(color_map), s=sizes)

# Setting labels for the axes

ax.set_xlabel('Population')

ax.set_ylabel('GDP')

ax.set_zlabel('Crime Rate')

# Setting the Title

ax.set_title('Caribbean Cluster')

# Creating a legend for the clusters

handles = [plt.Line2D([],[],marker="o", ls="", color=color_map[label]) for label in labels]

legend1 = ax.legend(handles, labels, loc="upper left", title="Clusters")

ax.add_artist(legend1)

plt.show()
```

Description of Each Cluster

In the 3D scatter plot, the size of each data point is a visual metaphor for a country's 'Land Size'. Larger points represent countries with larger land areas, while smaller points represent countries with smaller land areas. This creates a visual landscape that mirrors the physical landscape of the Caribbean clusters.

The color of each data point, on the other hand, signifies the cluster to which a country belongs. Each cluster is represented by a unique color, and these are the countries that correspond to each data point:

Cluster 2 (Blue): Antigua & Barbuda, The Bahamas, Barbados, Dominica, Jamaica, Saint Lucia, & Trinidad & Tobago.

Cluster 3 (Green): Cuba.

Cluster 4 (Brown): Dominican Republic.

Cluster 0 (Grey): Grenada, Saint Kitts & Nevis & Saint Vincent & the Grenadines.

Cluster 1 (Aquamarine): Haiti.

Beginning with Cluster 1, this cluster stands out due to its high 'Land Size' and 'Population', suggesting a densely populated country. This is similar to the other outlier clusters, 3 and 4 which also exhibit high population values with larger Land Sizes, indicating densely populated countries. In these clusters, they have a higher population densities than Cluster 1, which suggests a more concentrated distribution of people.

Interestingly, despite having similar population densities, these clusters differ in their economic output. Cuba has the lowest 'GDP' value, indicating a lower economic output. When compared to the other countries, this highlights that Cuba is the only country with a large land size and population but with one of the lowest real GDP Rates. This could be due to a variety of factors such as Cuba's economic policies, level of industrialization, availability and utilization of natural resources, and the efficiency of its labor force.

Remarkably, Clusters 1 and 4 do not have any recorded 'GDP' values in the clusters which means that the 'GDP' feature does not significantly contribute to the differences between these two clusters. This suggests that Land Size, Population and Crime Rates are more influential in differentiating the countries in these clusters.

On the other end of the spectrum, we have Clusters 2 and 0, which are characterized by significantly lower population values when compared to the other clusters. However, countries like Jamaica, Trinidad & Tobago, and the Bahamas in these clusters demonstrate a higher land size, indicating sparse populations. This contrast between land size and population in those clusters creates an interesting dynamic, where larger land sizes do not necessarily equate to larger populations.

Moreover, in Clusters 2 and 0, countries with larger land sizes exhibit a positive relationship with the Real GDP Rates, suggesting stronger economies. This correlation between land size and GDP underscores the potential economic advantages of having larger land sizes, such as more resources and space for development.

Finally, when we examine the crime rates across clusters 2 & 0, we see a trend: countries with higher land sizes and higher GDP rates tend to have higher Crime Rates, despite having significantly lower population rates. This suggests that factors beyond population size, such as economic conditions and social factors, may play a significant role in influencing crime rates. This is of the exception of Cuba, which doesn't have a recorded crime rate suggesting that Land Size, Population, and GDP are more influential in differentiating this cluster.

In conclusion, larger land sizes and higher population densities are grouped together (Clusters 1, 3, and 4), but larger land size doesn't always mean a larger population (Clusters 2 and 0). GDP doesn't significantly differentiate Clusters 1 and 4, suggesting other factors like 'Land Size', 'Population', and 'Crime Rates' are more influential. Interestingly, despite its large land size and population, Cuba (Cluster 3) has a lower GDP. Lastly, countries with higher land sizes and GDP rates tend to have higher Crime Rates, despite lower population rates.

References

Britannica, E. (n.d.). In Encyclopædia Britannica. Retrieved from https://www.britannica.com

Economist Intelligence Unit. (2023, February 28). Country Report: Cuba 1st Quarter 2023. Higher Ground. Retrieved from https://higherground.mt/wp-content/uploads/2023/02/Country_Report_Cuba_1st_Quarter_2023-1.pdf

International Monetary Fund. (2023). Jamaica: Population and GDP data. Retrieved from https://www.imf.org/external/datamapper/profile/JAM

International Monetary Fund. (2023). Country Data Profile | Population and GDP data. Retrieved from https://www.imf.org/external/datamapper/profile

McVeigh, T. (2023, June 16). The Caribbean island of Trinidad grapples with a national crisis of violence. The Guardian. Retrieved from https://www.theguardian.com/global-development/2023/jun/16/the-caribbean-island-of-trinidad-grapples-with-a-national-crisis-of-violence

Numbeo. (2023.). About Numbeo.com. Retrieved from https://www.numbeo.com/common/about.jsp

Numbeo. (2023). Caribbean Crime Indices. Retrieved from https://www.numbeo.com/crime/rankings_by_country.jsp?title=2023&region=029

Wilson, M. (2023, January 5). Another bloody year…what next? The Daily Express. Retrieved from https://trinidadexpress.com/opinion/columnists/another-bloody-year-what-next/article_6ba8dc4c-8d5a-11ed-944d-c3502ebef1fe.html

World Population Review. (n.d.). Violent Crime Rates By Country 2023. Retrieved from https://worldpopulationreview.com/country-rankings/violent-crime-rates-by-country