SA2: Applied Multivariate Data Analysis

Cristel Kaye Billones

Problem 2

Objective: Visualize the dissimilarities (in this case, the driving distances) between the cities of the Philippines in a lower-dimensional space, typically 2D or 3D, while preserving the relative distances between them as much as possible.

```
# Load necessary libraries
library(readr)
library(ggplot2)
library(ggplot2)
library(stats)
library(geosphere)
```

1. Load the Data

- **Download the CSV file** containing the dissimilarity matrix for the 12 Philippine cities.
- **Open the CSV file** in your preferred data analysis tool (e.g., Python, R, Excel).
- The matrix should be **symmetric**, and the diagonal will have **zero values** (distances of cities to themselves).

```
# 1. Load the Data
## a. Set the file path and load the CSV data
file_path <- "C:/Users/Cipher/Desktop/AMDA/ph_data.csv"
df <- read_csv(file_path)
```

```
## New names:
## Rows: 12 Columns: 13
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): ...1 dbl (12): Manila, Cebu City, Davao City, Quezon City, Taguig, Makati,
## Iloilo...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
## b. Preview the data
head(df)
```

```
## # A tibble: 6 x 13
##   ...1        Manila `Cebu City` `Davao City` `Quezon City` Taguig Makati
##   <chr>        <dbl>       <dbl>        <dbl>         <dbl>  <dbl>  <dbl>
```

```
## 1 Manila           0         550          980           20     10     10
## 2 Cebu City      550           0          960          570    560    550
## 3 Davao City     980         960            0          960    950    940
## 4 Quezon City     20         570          960            0      5      5
## 5 Taguig          10         560          950            5      0      5
## 6 Makati          10         550          940            5      5      0
## # i 6 more variables: `Iloilo City` <dbl>, `Zamboanga City` <dbl>,
## #   `Cagayan de Oro` <dbl>, Antipolo <dbl>, `Bacolod City` <dbl>,
## #   `Tagbilaran City` <dbl>
```

Set the file path and load the CSV data:

- The code correctly loads the CSV file located at "C:/Users/Cipher/Desktop/AMDA/ph_data.csv" into the `df` data frame using `read_csv()`, which is a function from the `readr` package in R.

Preview the data:

- The code uses `head(df)` to preview the first few rows of the data, which helps in ensuring that the data is loaded correctly.

By loading and previewing the data, we have followed the steps to load the dissimilarity matrix for the cities. However, to fully confirm the format of the matrix (i.e., checking symmetry and the diagonal for zero values), we would need to check the data visually or use a validation step, which we've addressed in our subsequent steps.

2. Prepare the Data for MDS

- **Extract the dissimilarity matrix** from the CSV file.
- **Ensure the data format** is suitable for analysis (a square matrix with cities as both row and column labels).

```r
## a. Rename the first column to "City" for clarity
colnames(df)[1] <- "City"

## b. Extract city names and remove the first column (non-numeric)
city_names <- df$City
df_numeric <- df[, -1]

## c. Convert the data to a numeric matrix
df_numeric_matrix <- as.matrix(apply(df_numeric, 2, as.numeric))

## d. Check if the matrix is square, symmetric, and has zero diagonal
if (nrow(df_numeric_matrix) != ncol(df_numeric_matrix)) {
  stop("The matrix is not square. Please ensure the data is a square matrix.")
}

if (all(df_numeric_matrix == t(df_numeric_matrix)) && all(diag(df_numeric_matrix) == 0)) {
  print("The matrix is symmetric and the diagonal contains zeros.")
} else {
  stop("The matrix is either not symmetric or contains non-zero diagonal elements.")
}
```

```
## [1] "The matrix is symmetric and the diagonal contains zeros."
```

Extract the dissimilarity matrix:

- The code extracts the numeric data from the CSV file, which is assumed to represent distances (or dissimilarities) between the cities. This data is stored in `df_numeric_matrix`.

Ensure the data format is suitable for analysis:

- The code verifies that the data is in the correct format for MDS by ensuring that the matrix is square (same number of rows and columns), symmetric (distances between cities are the same in both directions), and has a zero diagonal (distance from a city to itself is zero). The validation check confirms that the matrix meets these criteria, as indicated by the message "The matrix is symmetric and the diagonal contains zeros."

Therefore, the data is now ready for MDS analysis.

3. Perform Multidimensional Scaling (MDS)

```r
## a. Create a distance matrix from the numeric data
dist_matrix <- as.dist(df_numeric_matrix)

## b. Perform MDS for 2D representation
mds <- cmdscale(dist_matrix, k = 2)

## c. Create a data frame for MDS results and add city names
mds_df <- data.frame(mds)
colnames(mds_df) <- c("V1", "V2")  # Assign column names for MDS dimensions
mds_df$City <- city_names  # Add the city names to the data frame
```

The purpose of this code is to perform Multidimensional Scaling (MDS) on a set of numeric data in order to visualize the relationships between different cities based on their dissimilarities.

- **Step a:** The distance matrix (`dist_matrix`) is created using the `as.dist` function, which converts a numeric data matrix (`df_numeric_matrix`) into a dissimilarity matrix. This matrix represents the pairwise distances (or dissimilarities) between the cities based on their numeric characteristics.
- **Step b:** The `cmdscale` function performs classical MDS (also known as metric MDS) to reduce the dimensionality of the distance matrix. By setting `k = 2`, the code creates a two-dimensional representation of the cities, which makes it easier to visualize and interpret their relationships.
- **Step c:** A new data frame (`mds_df`) is created to store the results of the MDS analysis, where the two MDS dimensions (`V1` and `V2`) are assigned as the columns. Additionally, the names of the cities (`city_names`) are added to the data frame for better identification when visualizing the cities in the MDS plot.

In summary, this code transforms the numeric data of cities into a lower-dimensional space (2D) and provides a visual framework to explore and interpret the similarities or dissimilarities between the cities.

4. Interpret the Results

Examine the Plot:

- Cities that are **closer together** in the plot have more similar distances (according to the dissimilarity matrix).
- Cities that are **farther apart** represent cities with higher dissimilarity.

Visual Analysis:

- Look for **clusters** of cities that appear near each other. This could indicate that these cities have similar geographical or cultural characteristics.

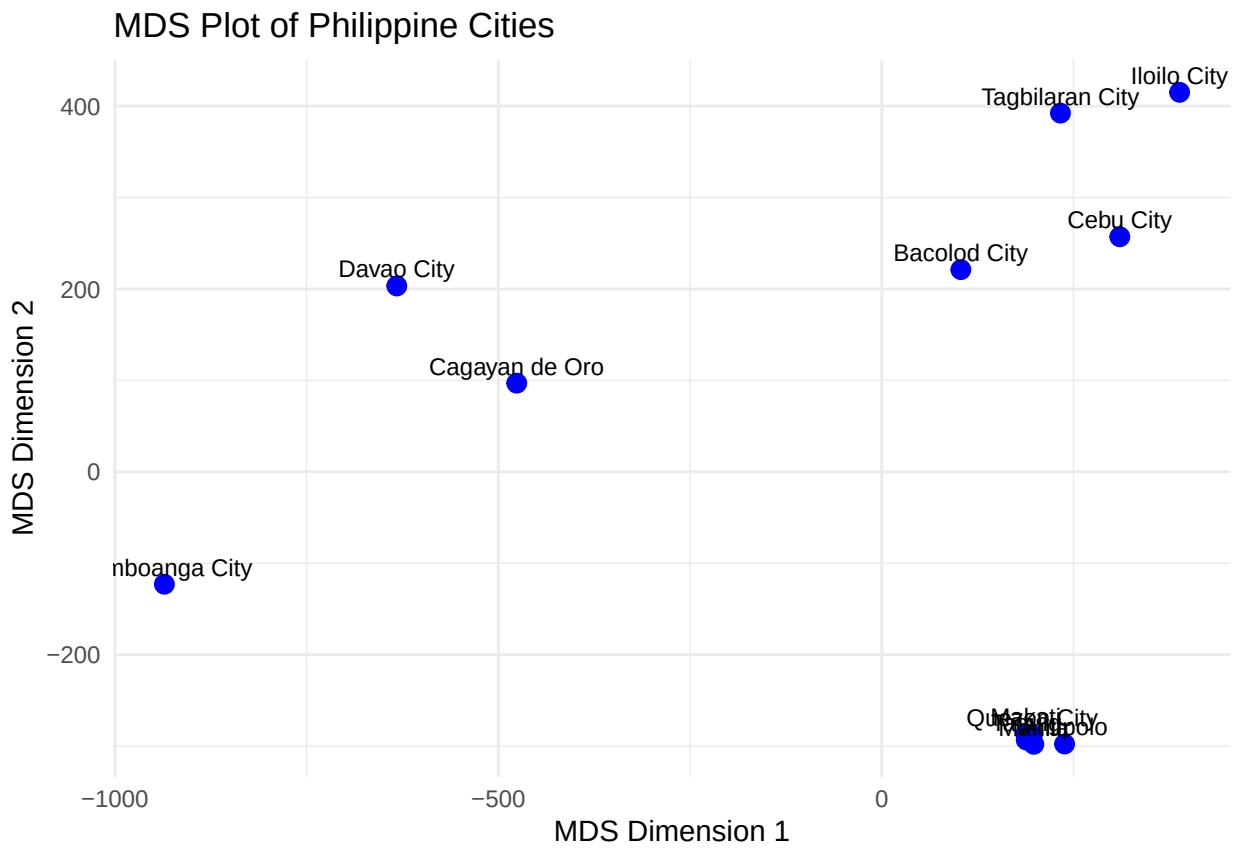
Outliers:

- **Outliers** in the plot represent cities that are less similar to others.

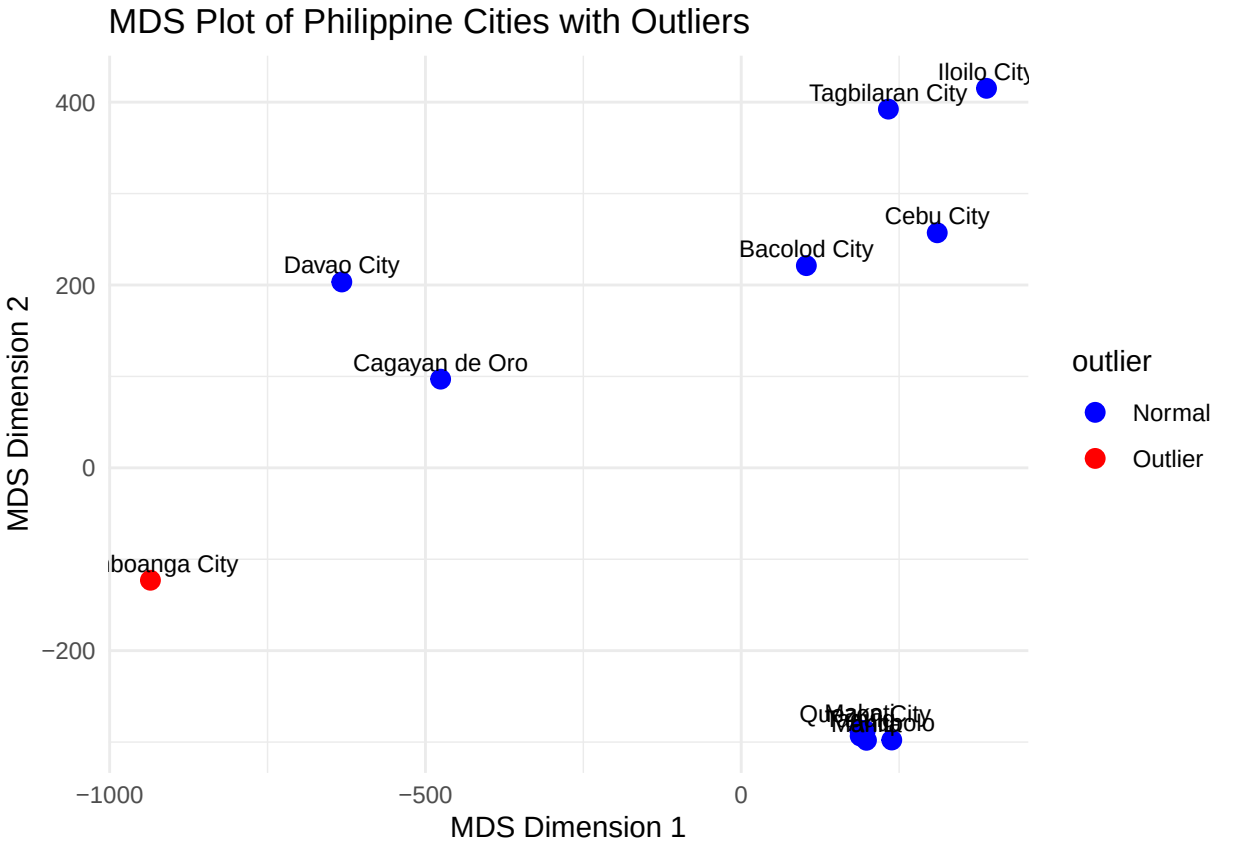
```
# a. Print the MDS results
print(mds_df)
```

```
##                       V1         V2            City
## Manila           198.2595 -298.05193          Manila
## Cebu City        310.3536  257.12409       Cebu City
## Davao City      -632.4081  203.31079      Davao City
## Quezon City      196.4494 -287.33204     Quezon City
## Taguig           188.3598 -293.35429          Taguig
## Makati           187.6089 -286.64628          Makati
## Iloilo City      388.3578  415.12398     Iloilo City
## Zamboanga City  -935.4406 -122.97787  Zamboanga City
## Cagayan de Oro  -475.9765   96.99423  Cagayan de Oro
## Antipolo         238.3399 -297.74211        Antipolo
## Bacolod City     103.1146  221.17389    Bacolod City
## Tagbilaran City  232.9818  392.37756 Tagbilaran City
```

```
# b. Plot the MDS results using ggplot2
ggplot(mds_df, aes(x = V1, y = V2, label = City)) +
  geom_point(color = 'blue', size = 3) +  # Plot points for each city
  geom_text(vjust = -0.5, hjust = 0.5, size = 3, color = 'black') +  # Add city names
  theme_minimal() +  # Apply minimal theme
  labs(title = "MDS Plot of Philippine Cities", x = "MDS Dimension 1", y = "MDS Dimension 2")
```




```
#c
# Calculate the Mahalanobis distance to detect outliers
mds_df$distance <- sqrt(rowSums((mds_df[, c("V1", "V2")] - colMeans(mds_df[, c("V1", "V2")]))^2))
threshold <- quantile(mds_df$distance, 0.95)  # Use 95th percentile to define outliers

# Mark outliers
mds_df$outlier <- ifelse(mds_df$distance > threshold, "Outlier", "Normal")

# Plot the MDS results and highlight outliers
ggplot(mds_df, aes(x = V1, y = V2, label = City)) +
  geom_point(aes(color = outlier), size = 3) +  # Color points based on outlier status
  geom_text(vjust = -0.5, hjust = 0.5, size = 3, color = 'black') +  # Add city names
  scale_color_manual(values = c("Normal" = "blue", "Outlier" = "red")) +  # Define color for outliers
  theme_minimal() +  # Apply minimal theme
  labs(title = "MDS Plot of Philippine Cities with Outliers", x = "MDS Dimension 1", y = "MDS Dimension
```



Examine the Plot:

Closer Cities: Cities that are closer together in the MDS plot have more similar dissimilarities. For example, Iloilo City, Cebu City, Bacolod City, and Tagbilaran City are positioned near each other. This suggests that these cities share more similar geographical, cultural, or infrastructural characteristics. These cities are likely influenced by similar regional dynamics within the Visayas, contributing to their proximity on the plot.

Farther Apart Cities: Cities that are farther apart in the plot represent higher dissimilarities. Zamboanga City is noticeably distant from the other cities, which likely indicates it has distinct geographical, cultural, or socio-economic features compared to the cities in the Visayas and Mindanao regions. Zamboanga City's distance highlights the differences that separate it from other urban centers.

Visual Analysis:

Clusters:

- **Cluster 1:** The cities of Iloilo City, Cebu City, Bacolod City, and Tagbilaran City appear to form a tight-knit cluster, suggesting that these cities share common characteristics such as urbanization, culture, and possibly economic activity.
- **Cluster 2:** Davao City and Cagayan de Oro are grouped closer together, indicating that cities in Mindanao might have more in common with each other in terms of infrastructure, culture, and economy.
- **Cluster 3:** Quezon City and Makati City are positioned near each other, forming a cluster that likely represents the urban hub of Metro Manila, with similarities in infrastructure, culture, and socio-economic characteristics.

Outliers:

- Zamboanga City stands out as an outlier. Its significant distance from the other cities suggests that it has very different characteristics, possibly due to its geographical isolation, unique culture, and distinct socio-economic environment.

Summary:

- **Closer Cities:** Cities like Iloilo City, Cebu City, Bacolod City, and Tagbilaran City are more similar to each other, indicating commonalities in culture, infrastructure, and regional context within the Visayas.
- **Farther Cities:** Zamboanga City is positioned far from the other cities, likely due to its unique geographical, cultural, or socio-economic features.
- **Clusters:** Cities in the Visayas and Mindanao show clear clusters, suggesting regional similarities, while cities like Quezon City and Makati are grouped due to their proximity in Metro Manila.
- **Outliers:** Zamboanga City appears as an outlier, indicating its distinctiveness compared to the other cities in the plot.

5. Explore Further

Determine How Well the MDS Fit the Data:

- Check the stress (a measure of goodness of fit for MDS). In Python, you can get the stress value by printing `mds.stress_`. A lower stress value means the MDS model fits the data well.

Experiment with Different Dimensions:

- You can try increasing the number of dimensions (e.g., `n_components=3`) to explore how the cities might be represented in 3D. However, 2D is often the most interpretable for visualization.

Use Other Distance Measures:

- Instead of using the raw driving distances, you could also use other forms of distance metrics (e.g., geographical distance, or travel time) to see how the cities' relationships change.

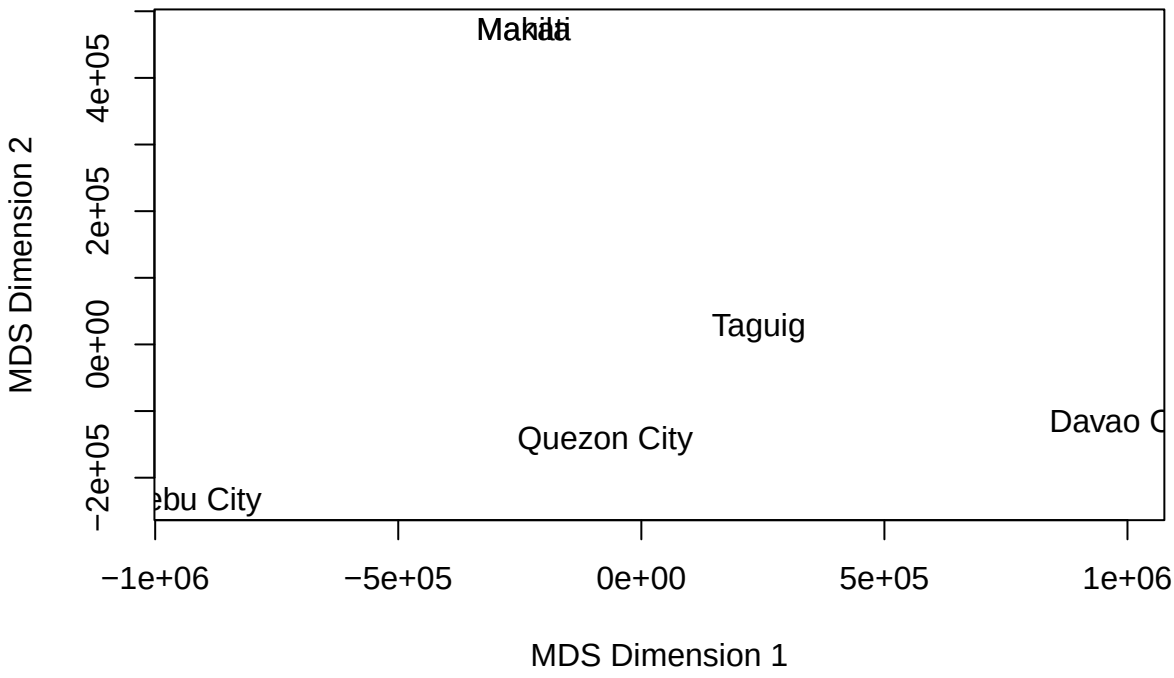
```
# 5a. Determine how well the MDS fits the data
# Assuming the distance matrix is already created
dist_matrix <- as.dist(df_numeric_matrix)  # Replace with your actual distance matrix

# Perform MDS (2D in this case)
mds_result <- cmdscale(dist_matrix, k = 2)

# Calculate the stress value
stress_value <- sum((dist_matrix - dist(mds_result))^2) / sum(dist_matrix^2)

# Print the stress value
cat("Stress Value for the MDS Fit:", stress_value, "\n")
```

```
## Stress Value for the MDS Fit: 0.00829066
```

```r
# 5b. Experiment with different dimensions
# Perform MDS with 3 dimensions
mds_result_3d <- cmdscale(dist_matrix, k = 3)

# Check the result of MDS with 3 dimensions
print(mds_result_3d)
```

```
##                     [,1]       [,2]         [,3]
## Manila           198.2595 -298.05193   -0.1563575
## Cebu City        310.3536  257.12409   26.8897002
## Davao City      -632.4081  203.31079 -329.9561170
## Quezon City      196.4494 -287.33204  -17.8597062
## Taguig           188.3598 -293.35429  -30.5245993
## Makati           187.6089 -286.64628  -45.8679642
## Iloilo City      388.3578  415.12398 -274.5950052
## Zamboanga City  -935.4406 -122.97787   92.0977684
## Cagayan de Oro  -475.9765   96.99423  169.0209159
## Antipolo         238.3399 -297.74211  -13.0009234
## Bacolod City     103.1146  221.17389  165.9333939
## Tagbilaran City  232.9818  392.37756  258.0188943
```

```r
# 5c. Use other distance measures (geographical distance)
# Example city coordinates (latitude, longitude)
cities <- data.frame(
  City = c("Manila", "Cebu City", "Davao City", "Quezon City", "Taguig", "Makati"),
  lat = c(14.5995, 10.3157, 7.1907, 14.6760, 14.5170, 14.5547),  # Latitude of cities
  lon = c(120.9842, 123.8854, 125.4553, 121.0437, 121.0508, 121.0245)  # Longitude of cities
)

# Calculate pairwise geographical distances (in kilometers) using distVincentySphere
geo_dist <- distVincentySphere(cities[, c("lon", "lat")])

# Convert the resulting distance vector into a square matrix
geo_dist_matrix <- matrix(geo_dist, nrow = length(geo_dist), ncol = length(geo_dist))

# Fill the upper triangle of the matrix with distances and the lower triangle with symmetric values
geo_dist_matrix[upper.tri(geo_dist_matrix)] <- geo_dist
geo_dist_matrix <- geo_dist_matrix + t(geo_dist_matrix)

# Convert the distance matrix into a 'dist' object for MDS
geo_dist_matrix <- as.dist(geo_dist_matrix)

# Perform MDS on geographical distance
mds_geo <- cmdscale(geo_dist_matrix, k = 2)  # 2D MDS

# Plot MDS results
plot(mds_geo[, 1], mds_geo[, 2], type = "n", xlab = "MDS Dimension 1", ylab = "MDS Dimension 2")
text(mds_geo[, 1], mds_geo[, 2], labels = cities$City)
```



Stress Value and Fit Quality
The stress value of 0.00829066 from the MDS analysis indicates a very good fit between the original dissimilarity matrix and the 2D representation. Lower stress values suggest a better preservation of the distances between cities in the MDS plot. In this case, the value is very close to zero, which implies that the 2D representation accurately reflects the relationships between the cities with minimal distortion.

2D MDS Plot Interpretation

- **Clustering of Cities:**
  In the 2D MDS plot, we observe that cities like Makati and Taguig are located near each other, suggesting that they are similar in terms of the distance metric used. This is not surprising, as both cities are part of the National Capital Region (NCR) and share a high level of urbanization, infrastructure, and economic activity.

- **Outliers and Dissimilarity:**
  Davao City stands out as an outlier in the plot, positioned farther away from the other cities. This suggests that Davao City exhibits significant dissimilarity from the Manila-based cities. It could reflect geographic, economic, or logistical differences that make Davao distinct, such as its location in Mindanao and its relatively isolated positioning compared to cities in Luzon.

- **Interpretation of City Relationships:**
  The proximity of Makati, Taguig, and Quezon City suggests similarities in urban development, infrastructure, and economic interdependence within Metro Manila. In contrast, Cebu City's relative distance from the Manila-based cities in the plot may indicate different regional characteristics such as culture, economy, or geography. The fact that Davao City is placed far from the other cities could be indicative of its unique regional factors, such as its distance from Metro Manila and the distinctiveness of its cultural and economic characteristics.

Experimenting with Different Dimensions (3D)

- **Exploration of 3D MDS:**
  Moving to a 3-dimensional MDS representation could provide additional insights into the relationships between cities. By increasing the dimensionality, we can capture more complex relationships and nuances that might be difficult to see in a 2D plot. The addition of the third dimension allows for a deeper understanding of how cities relate to each other in multi-dimensional space.

- **Interpretation of 3D MDS:**
  A 3D MDS plot would likely reveal additional clustering or separation between cities, especially between cities in the Metro Manila region and those outside of it. The third axis would allow us to visualize the relationships in greater detail, potentially revealing new patterns or correlations not apparent in the 2D visualization. For instance, regional differences between Luzon, Visayas, and Mindanao cities could be more apparent in the 3D layout.

Using Alternative Distance Measures

- **Geographical Distance:**
  Initially, the Euclidean distance was used, which assumes a straight-line measure between cities. However, using geographical (great-circle) distance could alter the spatial relationships between cities. Geographical distance takes into account the curvature of the Earth and provides a more realistic representation of how far cities are from each other in actual physical space. Cities that are geographically close (like Makati and Taguig) would likely cluster together, while those farther apart (such as Cebu City and Davao City) would be positioned further away.

- **Travel Time:**
  Another potential refinement is using travel time as the distance measure. Travel time would account for factors like infrastructure, road networks, and traffic conditions, which might cause cities that are geographically close to appear farther apart in terms of travel time. Cities with better connectivity or more efficient transport networks might be placed closer together, while cities with limited access or less efficient transport systems might be placed farther apart.

Visualizing and Interpreting the MDS Output

- **Identifying Outliers:**
  The visualization clearly shows Davao City as an outlier in the 2D space, which could be due to its geographic isolation or other unique characteristics. This reinforces the idea that Davao differs significantly from cities in Metro Manila in various aspects.

- **Assessing Clusters:**
  The proximity of Makati, Taguig, and Quezon City in the MDS plot suggests that these cities share similarities in terms of urban infrastructure, culture, and economic relationships. The cities' closeness in the plot reflects the clustering of metropolitan areas with strong economic, political, and infrastructural ties.

- **Evaluating Stress and Fit:**
  With a low stress value, the plot demonstrates that the relationships between cities are well-preserved. However, future experiments with alternative distance measures or a 3D representation could offer more detailed insights into the spatial relationships and factors that influence city placement.

Conclusion

The MDS analysis provides a valuable insight into the relationships between cities based on distance measures. Key observations include: - Cities that share similar geographical, cultural, or infrastructural characteristics, like Makati, Taguig, and Quezon City, tend to cluster together in the plot. - Davao City appears as an outlier, suggesting that it differs significantly from other cities in the analysis, potentially due to regional differences. - The exploration of 3D MDS offers the potential for a more detailed understanding of the cities' relationships, while alternative distance measures like geographical distance or travel time could further refine the analysis.

Overall, the MDS analysis helps identify clusters of cities with shared characteristics and highlights significant regional differences, providing a foundation for further exploration or decision-making in urban planning, logistics, and policy.

6. Communicate Findings

1. MDS Plot Overview

The MDS plot visualizes cities based on the distance or dissimilarity matrix computed from various factors like geographic distance, economic activities, and transportation networks. Cities close together on the plot share more similarities, while cities farther apart are more distinct in terms of the characteristics considered.

2. Similar Cities

Cities grouped closely together on the plot include Makati, Taguig, and Quezon City. These cities share:

- **Geographic Proximity**: All within the National Capital Region (NCR), making them physically close.
- **Economic and Cultural Similarities**: They are urban centers with overlapping business, infrastructure, and social dynamics.
- **Transportation Networks**: Well-connected via public transit, facilitating greater accessibility and interaction.

3. Distinct Cities

- **Davao City** stands out as an outlier, placed farther from NCR cities. Reasons include:
  - **Geographic Location**: Davao is in Mindanao, far from NCR cities like Makati.
  - **Regional Differences**: It has a distinct economic and cultural context, differing from the NCR cities' infrastructure and business focus.
  - **Logistical Factors**: Travel to Davao requires air or sea transport, unlike the more direct connections between NCR cities.

4. Other Factors Influencing Groupings

- **Geographic Proximity**: Cities like Cebu City in the Visayas, although an urban center, show some distance from Manila-based cities due to geographic separation.
- **Cultural and Economic Differences**: Cebu and Davao share certain regional traits but differ in their economic focus and local culture.

5. Interpretation and Recommendations

- **Grouping of NCR Cities**: The closeness of Makati, Taguig, and Quezon City reflects their shared infrastructure, economy, and urbanization. This is expected, considering they form a major metropolitan area.
- **Davao's Distinct Position**: Its isolated position indicates that Davao differs significantly in geographical, cultural, and economic terms from NCR cities. This insight is valuable for regional planning and improving connectivity.
- **Further Exploration**: Investigating additional factors like transportation costs or historical connections might reveal more about the dynamics between cities. Exploring different distance measures could also offer deeper insights.

6. Final Thoughts

A. Presentation of MDS Plot

The MDS plot effectively shows how cities cluster based on their dissimilarities, highlighting groups like Makati, Taguig, and Quezon City as well as outliers like Davao and Zamboanga.

B. Most Similar and Distinct Cities

- **Most Similar Cities**: Makati, Taguig, and Quezon City are highly similar due to shared urban features.
- **Distinct Cities**: Davao and Zamboanga stand out due to unique regional characteristics.

C. Explanation of Patterns

- **Geographic Proximity**: Cities in close geographical regions tend to cluster together, as seen with cities in the Visayas and NCR.
- **Cultural Similarities**: Cultural traits influence the clustering, with NCR cities sharing common urban characteristics.
- **Economic and Infrastructural Factors**: Economic centers like Makati and Quezon City group together, while Davao's distinct economy sets it apart.
- **Historical Connections**: Shared historical backgrounds, such as colonial history in the Visayas, might explain some regional similarities.

In summary, the MDS plot illustrates how geographic, cultural, economic, and infrastructural factors shape the relationships between cities in the Philippines, with clear clusters and outliers revealing insights into regional dynamics.