# A REVIEW OF THE BOREALIAN CONSULATE NETWORK

Laura Mann[1], Devan Scholefield[1], Bhavika Sewpal[1], John Tsang[1]

**Abstract**

The aim of this project was to provide the Borealian Foreign Office (BFO) with valuable insights about their daily operations based on the case and program information available in the PIMENTO data set. In order to efficiently analyse the data set and extract valuable observations from it we split the project into two phases. The first phase involved a number of transformations and assumptions and focused mainly on understanding the data set through data visualisations. These assumptions were necessary to provide reasonable definitions of the variables in the data set while still preserving the integrity of the data. This was also necessary to understand the relationships between variables. We then tidied the data due to the presence of missing values, inconsistent instances, and outliers. We were able to extract figures that summarised the data set. These findings were then illustrated using different kinds of graphical representations. The second phase of the project focused exclusively on clustering in an attempt to group missions by similarity. To do so, six more compact data sets were derived from the original massive data set, based on several criteria described in the clustering section. Then, three different kinds of clustering algorithms were performed on each data set, namely PAM, hierarchical clustering and K-means clustering. This resulted in 18 clustering results which were then fed into the final clustering algorithm that used the DBSCAN technique. The DBSCAN algorithm then grouped all the missions into five clusters. Unfortunately, we were not able to precisely deduce the cause of these differences. This report ended with five recommendations on how the BFO can improve the Borealian consular network.

[1]Department of Mathematics and Statistics, University of Ottawa, Ottawa
**Email**: wtsan094@uottawa.ca

## Contents

## 1. Introduction

The Borealian Foreign Office (BFO) is responsible for operating all Borealian consulates worldwide. Each mission employee in this consulate network offers services to Borealian citizens abroad and soon-to-be immigrants/visitors to Borealia. The BFO collects information about its mission employees in the PIMENTO database. This database contains daily details on the number of cases these employees work on and the amount of time they spend on cases and various programs.

The BFO seeks an in-depth understanding of the PIMENTO database to make Borealian missions as efficient and helpful as possible. Therefore, this project reviews the Borealian consulate network, focuses on analysing the consulate network's human resource allocation, and achieves these goals in two phases. Phase 1 aims to provide the BFO with an initial understanding of the PIMENTO database, how this database represents the Borealian consular network and prepare data sets for the subsequent phase. Phase 2 uses clustering analysis to uncover similar missions in the network. In the end, based on evidence established in these two phases, this project makes recommendations to the BFO about possible ways to improve the Borealian consulate network for more efficient and helpful missions.

The scope of this mission was to analyse human resource allocation and similar missions between 2021 and 2026. In order to reach the full breadth of this scope we divided the project into multiple stages. Starting with understanding the makeup of the PIMENTO database and the data contained within it. We created schema, such as in Figure 1, to assist in our understanding. This in turn supported our future explorations into the data using data visualisation techniques and clustering algorithms.

We aimed to provide a greater depth of understanding of the data set through visualisations. These visualisations looked at the progression of cases over time, compared the time spent on cases vs. programs, and looked at the distributions of time spent on the various types of cases and programs.

We then tested clustering methods to look at grouping similar missions. We used hierarchical Clustering, k means, PAM clustering, and finally DBSCAN to look at multiple ways of clustering similar missions together. The goal of this was to find unique and insightful ways in which missions could be grouped in order to find potential operating efficiencies.

This report ends with five recommendations concerning how the BFO can improve operations of Borealian consulate worldwide through rotational programs for employees, review committees, restructuring the consulate network and expanding information recorded in the PIMENTO data set.

## General Statistics about the BFO

In Table 1 some general statistics about the BFO are given. The number of unique employees is given as the number of unique employees for the full time span of 2021 - 2026. For example, if an employee worked in 2021 and 2022, but then retired, they are still counted as a unique employee. For unique case and programs types "other" was counted as one unique type. "Other" could contain a multitude of other types, however it is generally accepted that none of these sub-types would be as significant as the standalone types.

**Table 1.** General Statistics about the BFO

| | |
|---|---|
| Number of Unique Missions | 231 |
| Number of Unique Employees | 1966 |
| Unique Case Types Including "Other" | 15 |
| Unique Program Types Including "Other" | 14 |

This report comes with an Appendix for technical detials and a GitHub repository containing `R` programs this project uses. The link to the repository: `github.com/john-tsang/MAT4376G-Project`.

## 2. Objectives and Scope

This project aims to use the PIMENTO database from the BFO to provide the office with an overview of the Borealian consulate network worldwide. Analysis in this project focuses on improving the human resource allocation among these consulates.

### Phase 1: Data Understanding and Data Visualisation

The purpose of phase 1 is to understand the Borealian consulate network from the PIMENTO database, provide a narrative and visualisation of the human resource allocation among all Borealian consulates in the world and prepare data sets for the next phase.

### Phase 2: Similarities among Missions

The project also aims to improve human resource allocation among Borealian consulates globally by identifying similarities among missions. Understanding the current human resource allocation contributes to the more efficient use of human resources in the Borealian consulate network.

In the end, this project makes some recommendations to improve operations of Borealian consulates worldwide.

### Project Scope

This project uses techniques described in section 3 to analyse human resource allocation and similar missions with the PIMENTO database spanning from 2021 to 2026.

## 3. Methodology

This section outlines the methods used to analyse the PIMENTO database in phases one and two of this project. Specifically, the first phase develops a general understanding of the database through preparing data sets for the second phase. The second phase uses clustering to identify similar missions.

### Phase 1: Data Understanding and Data Visualisation

Phase 1 of this project develops a general understanding of the Borealian consular network represented by the PIMENTO database through a three-step approach.

**Step 1: Identifying conceptual entities and establishing schema of the PIMENTO database**

Identifying all conceptual entities in the Borealian consular network represented by the PIMENTO database aims to facilitate an abstract understanding of the network and what aspects the data set represents the network.

Because the PIMENTO database does not have descriptions for each column other than its name, this project aims to develop a definition for each column (schema). These definitions facilitate the use of the database in this project and future uses and database maintenance by the BFO.

**Step 2: Validating and Cleaning Data**

Because of the large size of the raw PIMENTO database (1,300,526 instances (or rows) in total), it is inevitable for the database to contain problematic instances, including missing values, illogical and inconsistent instances and outliers. Resolving these instances furthers the understanding of the network and provides data sets usable for analysis in the remainder of this project.

**Step 3: Exploratory Data Analysis**

A descriptive analysis of the PIMENTO database provides a basic understanding of the Borealian consular network. This initial diagnosis of the network from this exploratory analysis identifies the strengths and weaknesses of the network and facilitates the analysis in Phase 2.

### Phase 2: Similarities among Missions

Through clustering analysis, this phase aims to uncover groups of similar missions. These groups allow for the investigation of more efficient human resource allocation in the Borealian consular network.

## 4. The PIMENTO Database

Each instance (row) of the PIMENTO database from the BFO is a daily work record from each mission employee in each Borealian consulate worldwide. Mission employees of each Borealian consulate in each geographical region perform two types of consular services: (1) program and (2) case:

**Consular Program:** Non-routine consular services to promote Borealian interests

**Consular Case:** Routine consular services provided to Borealian citizens abroad, soon-to-be immigrants and visitors to Borealia.

Accordingly, there are two work record database tables in the PIMENTO database, each corresponding to the work record of a type of consular service. The remainder of this document denotes **Program** as the data set containing mission employees' work records for consular programs, while **Case** as the data set containing mission employees' work records for consular cases.

### 4.1 Conceptual Entities

The PIMENTO database consists of four conceptual entities:

- geographical regions
- missions
- mission employees
- tasks

**Geographical Regions** The BFO divided the world into 14 different regions. Each region has a distinct name in the database. This classification applies to both the Program and the Case data sets (Table 2).

**Table 2.** Geographical Regions in the PIMENTO Database (Both Program and Case Data Sets)

| **America** | |
|---|---|
| United States | Caribbean |
| Central America and Mexico | South America |
| **Europe** | |
| Europe (West) | Europe (East) |
| **Africa** | |
| Africa (North) | Africa (South) |
| **Middle East** | |
| Middle East | |
| **Asia** | |
| Asia (Central) | Asia (South) |
| Asia (Southeast) | Asia (East) |
| Asia (Oceania) | |

**Missions** The BFO further divided each of the 14 geographical regions into different sub-regions where each mission employee carried out their assigned missions. The BFO office records the start date (the day, the month and the year) of each mission in the database and names the mission of each mission employee by the assigned sub-region.

The Program and the Case data sets do not have the same set of missions. Specifically, missions “Asmara" and “Antananarivo" in geographical region Africa (South) exist only in the Program data set. Overall, the Program data set has 229 distinct missions, while the Case data set has 221 missions from 2021 to 2026, with 219 in common. The following 10 missions exist only in the Case data set.

| | |
|---|---|
| **Africa (South):** | Ammertuma, Yamai, Asbi, Usumbura, eGoli and Chach |
| **Europe (East):** | Tartu and Wilna |
| **South America:** | Atarillo |
| **Asia (South East):** | Puranupakorn |

**Mission Employees** The BFO hires mission employees to carry out missions in different sub-regions of the 14 geographical regions. The PIMENTO database uniquely identifies each employee by an identification number (ID).

Throughout the period from 2021 to 2026, 1,679 distinct employees worked for consular programs while 1,967 mission employees worked for consular cases. Every mission employee either works on cases only or works on both cases and programs.

**Tasks** Each mission consists of different types of tasks. In the Program data set, each mission involves at most 14 different types of tasks. In the Case data set, each mission involves at most 15 different kinds of tasks. The Case data set records both the count for each type of task and the employee’s time on each type of task. However, the Program data set only records the time the employee spent on each type of task. Table 3 lists all types of tasks in the Program and the Case data sets.

**Relationships Among Conceptual Objects** The PIMENTO database shows the following relationships among geographical regions, missions, mission employees and tasks.

- Each geographical region has more than one mission, more than one task and more than one mission employee.
- Each mission corresponds to multiple tasks and multiple mission employees and corresponds to only one geographical region.
- Each mission employee works for more than one mission and more than one task. However, each mission employee corresponds to one geographical region only.

**Table 3.** Enumeration of All Types of Tasks in the Program and the Case Data Set

| Program Data Set | Case Data Set |
|---|---|
| Arrest | Commerce & Trade |
| Assistance Communications | Development |
| Child Abduction/ Custody | Emergency |
| Citizenship | Immigration |
| Death | Informatics |
| Disasters | Liaison |
| Evacuation | Other |
| Family Distress | Political & Economic Interests |
| Financial Assistance / Transfers | Police |
| Immigration | Program Management |
| Legal / Notary | Program Services |
| Other | Public Communications |
| Passport | Training |
| Registration | Visitor Management |
| Service | |

Figure 1 shows an entity-relationship model[1] summarizing

[1] An entity-relationship model represents the design of a database and describes relationships between conceptual entities

the relationships among conceptual entities in the PIMENTO database for both the Program and the Case data sets.

## 4.2 Database Schema

This section describes the logical structure (schema or metadata) of the Program and the Case data sets in the PIMENTO database. The main differences between the schemata of the Program and the Case data sets are twofold.

(1) The Program and the Case data sets have different types of tasks.

(2) The Program data set does not have columns to count each task while the Case data set has such columns.

### 4.2.1 The Program Data Set

Each row in the Program data set contains 20 columns representing an instance of a mission employee's participation in a mission (i.e. a work record) in a day. This data set characterizes an instance of a mission employee's participation in a mission by

(1) the identification number of the mission employee,

(2) the geographical region,

(3) the English names of the mission

(4) the number of hours the employee spent on 14 different types of tasks and

(5) the start date of the mission employee's participation.

Table 4 summarizes the definitions of each column in the Program data set.

### 4.2.2 The Case Data Set

Each row of the Case data set contains 35 columns representing an instance of a mission employee's participation in a mission in a day. This data set characterises an instance of a mission employee's participation in a mission by

(1) the identification number of the mission employee,

(2) the geographical region,

(3) the English names of the mission,

(4) the number of hours the employee spent on 15 different types of tasks,

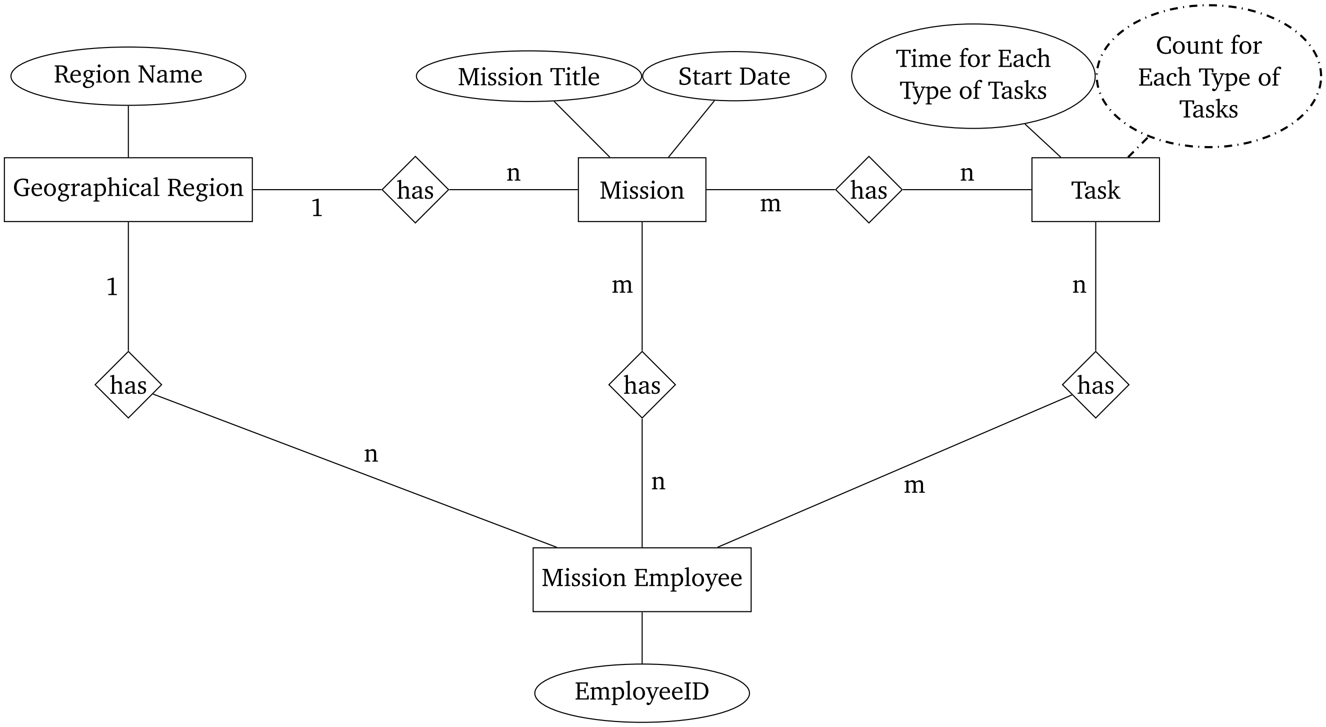
(5) the number tasks completed for each type of tasks and

(6) the start date of the mission employee's participation.

Table 4 summarizes the definitions of each column in the Case data set.

in a database. The Wikipedia page of Entity-relationship model (https://en.wikipedia.org/wiki/Entity%E2%80%93relationship_model) explains the model in detail.

**Figure 1.** Entity Relationship Diagram of the Program and the Case Data Sets



Notes:

- Each rectangle represents one of the four conceptual entities. Each oval represents an attribute of the corresponding conceptual object. Each diamond refers to the relationship between two entities.
- The number or variables (*m* or *n*) corresponding each line joining two entities (rectangles) with a relationship (diamond) signifies the type of the relationship between two entities. For instance, each geographical region has *n* missions while each mission corresponds to only 1 geographical region.
- The dot-dashed attribute of task (count for each type of tasks) exists only in the Case data set. Other entities and attributes in both the Program and the Case data sets are the same.

**Table 4.** Summary of Definitions, Data Types and Ranges of Every Column in the Program Data Set

**(a) Mission Instance Record Identifiers**

| Column | Definition | Data Types | Min. | Max. |
|---|---|---|---|---|
| GeoRegionNameE | English name of the geographical region | String | - | - |
| MissionTitleE | English name of the mission | | - | - |
| EmployeeCode | Identification code of the corresponding mission employee | Integer | 135 | 19260 |

**(b) Mission Instance Start Date (Integers)**

| Column | Definition | Min. | Max. |
|---|---|---|---|
| Month | Month of the start date | 1 | 12 |
| Day | Day of the start date | 1 | 31 |
| Year | Year of the start date | 2021 | 2026 |

**(c) Minutes Spent on Program Management for Each Instance (Integers)**

| Column | Definition | Min. | Max. |
|---|---|---|---|
| Emergency | Number of minutes spent on emergency services | 0 | 5150 |
| Program_Mgmt | Number of minutes spent on program management | 0 | 9420 |
| Visit_Mgmt | Number of minutes spent on visitor management | 0 | 6600 |
| Training | Number of minutes spent on training | 0 | 4890 |
| Liaison | Number of minutes spent on liaison | 0 | 2400 |
| Informatics | Number of minutes spent on informatics | 0 | 2000 |
| Program_Services | Number of minutes spent on program services | 0 | 9600 |

**(d) Minutes Spent on Different Types of Program Tasks (Integers)**

| Column | Definition | Min. | Max. |
|---|---|---|---|
| Pol_Econ | Number of minutes spent on promoting the political and economic interest | 0 | 2160 |
| Comm_Trade | Number of minutes spent on promoting international commerce and trade | 0 | 3000 |
| Development | Number of minutes spent on international development | 0 | 2400 |
| Police | Number of minutes spent on communicating with local police | 0 | 960 |
| Immigration | Number of minutes spent on promoting immigration to Borealia | 0 | 4500 |
| Public_Comms | Number of minutes spent on public communications with local stakeholders | 0 | 13801 |
| Other | Number of minutes spent on other services | 0 | 14400 |

**Table 5.** Summary of Definitions, Data Types and Ranges of Every Column in the Case Data Set

**(a) Mission Instance Record Identifiers**

| Column | Definition | Data Types | Min. | Max. |
|---|---|---|---|---|
| GeoRegionNameE | English name of the geographical region | String | - | - |
| MissionTitleE | English name of the mission | | - | - |
| EmployeeID | Identification code of the corresponding mission employee | Integer | 135 | 19,260 |

**(b) Mission Instance Start Date (Integers)**

| Column | Definition | Min. | Max. |
|---|---|---|---|
| Month | Month of the start date | 1 | 12 |
| Day | Day of the start date | 1 | 31 |
| Year | Year of the start date | 2021 | 2026 |

**(c) Counts and Minutes for Each Type of Tasks for Each Instance (Integers)**

| Column | Definition | Min. | Max. |
|---|---|---|---|
| Disasters | Number of new disasters-related tasks opened | 0 | 240 |
| Disasters Time | Total Number of minutes spent on disasters-related tasks | 0 | 13080 |
| Death | Number of new death-related tasks opened | 0 | 60 |
| Death Time | Total Number of minutes spent on death-related tasks | 0 | 6500 |
| Assistance Communications | Number of new assistance communications-related tasks opened | 0 | 1719 |
| Assistance Communications Time | Total Number of minutes spent on assistance communications | 0 | 26580 |
| Legal/Notary | Number of new Legal/Notary-related tasks opened | 0 | 66150 |
| Legal/Notary Time | Total Number of minutes on legal assistance and notary services | 0 | 3600 |
| Evacuation | Number of new evacuation opened | 0 | 240 |
| Evacuation Time | Total Number of minutes spent on evacuation | 0 | 7200 |
| Child Abduction/ Custody | Number of new child abduction/custody cases opened | 0 | 220 |
| Child Abduction/ Custody Time | Total Number of minutes spent on child abduction/ custody cases | 0 | 2490 |
| Family Distress | Number of new tasks related to family distress | 0 | 5560 |
| Family Distress Time | Total Number of minutes spent on tasks related to family distress | 0 | 2000 |
| Registration | Number of new tasks related to registration | 0 | 532 |
| Registration Time | Total Number of minutes spent on registration-related tasks | 0 | 14265 |
| Immigration | Number of new tasks related to immigration | 0 | 2,230 |
| Immigration Time | Total Number of minutes spent on immigration-related tasks | 0 | 9105 |
| Arrest | Number of new tasks related to arrest | 0 | 960 |
| Arrest Time | Total Number of minutes spent on arrest-related tasks | 0 | 6000 |
| Citizenship | Number of new tasks related to citizenship | 0 | 450 |
| Citizenship Time | Total Number of minutes spent on citizenship-related tasks | 0 | 4590 |
| Passport | Number of new tasks related to passport | 0 | 2,020 |
| Passport Time | Total Number of minutes spent on passport-related tasks | 0 | 28800 |
| Financial Assistance/ Transfers | Number of new tasks related to financial assistance/transfers | 0 | 248 |
| Financial Assistance/ Transfers Time | Total Number of minutes on financial assistance/transfers | 0 | 3600 |
| Service | Number of new tasks related to service | 0 | 7730 |
| Service Time | Total Number of minutes spent on service-related tasks | 0 | 28800 |

## 5. Data Preparation

This section discusses the process to validate and clean the Program and the Case data sets. Specifically, this project checks if the data types, ranges and consistency of all variables in these two data sets are logical. Then this project cleans these two data sets by identifying and resolving all observations with missing values and outliers.

### 5.1 Data Validation

Data validation ensures that the data is in an appropriate format for further processing.

#### 5.1.1 Data Type and Range Check

Confirming all columns have logical data types with reasonable ranges is the first step to ensuring the integrity of data sets. In the Program data set, each column represents either a name, a date or an amount of time and hence should fulfill the following constraints.

- **Name**: each value in the column must be a string.
- **Date**: each value in the column corresponding to date must be an integer.
  - If the column represents a year, its value must be between 2021 to 2026.
  - If the column represent a month, its values must be between 1 to 12.
  - If the column represents a day, its values must be between 1 to the number of days of the corresponding month in the year.
- **Amount of time**: each value representing an amount of time has to be any number greater than or equal to 0.

The Case data set contains all types of columns of the Program data set, in addition to a count of each type of task. The count of each type of task must be a whole number. Every value in these two data sets conforms to data types and ranges (minima and maxima) of the in Tables 4 and 5.

#### 5.1.2 Consistency Check

Checking if values in all columns are logical is also important to establish the integrity of data sets. Specifically, this project checks the following conditions.

(1) In both the Program and the Case data set, the total number of minutes all mission employees spent on a mission must be greater than 0. An instance with a total of zero minutes on all tasks is not reasonable.

**Violations in the Program Data Set:**
About 58.7 percent (367,970 of 627,298) of the instances have zero total number of minutes spent on all tasks. These instances come from all geographical regions and 216 of 221 missions from 2021 to 2026.

**Resolution to Violations in the Program Data Set:**
We drop all these observations in the Program data set for clustering analysis because these observations cannot provide information for our analysis. However, employee information from these observation retains.

**Violations in the Case Data Set:**
About 22.5 percent (151,325 of 673,228) of the instances have zero total minutes spent on all tasks. However, 300 of these instances have a total count of tasks greater than 0 and hence provide some information for analysis.

**Resolution to Violations in the Case Data Set:**
We drop all observations with both zero total number of minutes spent on all tasks and zero total counts (22.4 percent, 151,025 of 673,228 observations. However, employee information from these observation retains.)

(2) Among the remaining 522,203 instances in the Case data set, we have to check the following two conditions.

(i) if the total number of tasks of the instance is greater than 0, the total time spent should also be greater than 0.

**Violations:** About 25.9 per cent (135,329) instances have zero total numbers of tasks but a positive total time spent on all tasks.

(ii) the number of new cases opened every day for each mission employee has to be reasonable.

We regard an employee opening more than 1,400 new tasks a day as unreasonable because this high number of new tasks is equivalent to starting more than 58 new tasks an hour in a day. Table 6 summarizes employees with these characteristics. Each of these ten rows of Table 6 corresponds to only one instance in the Case data set. Given this small number of illogical instances, we decided to remove these instances

**Table 6.** Summary of Unreasonable Number of New Cases a Mission Employee Starting in a Day

| Employee ID | Date | Mission | # New Cases |
|---:|---|---|---:|
| 1641 | 2/26/2026 | Saragossa | 1536 |
| 8889 | 3/27/2025 | Chuqi Yapu | 1652 |
| 867 | 9/9/2021 | Vienna | 1723 |
| 8889 | 5/31/2021 | Chuqi Yapu | 1765 |
| 13677 | 12/1/2025 | Everglades | 2031 |
| 357 | 6/12/2023 | Mont Blanc | 2245 |
| 6999 | 11/30/2025 | Schduagert | 3480 |
| 1965 | 5/31/2023 | ePitoli | 6011 |
| 357 | 4/7/2024 | Mont Blanc | 7752 |
| 2091 | 5/28/2025 | Kenitra | 66186 |

from the Case data set.

These ten illogical instances should be data entry errors. For example:

- Employee 2091 in Kenitra, Africa (North) opened 66,186 new tasks on May 28th, 2025, 66,150 of which coming from legal/notary services,
- Employee 357 in Mont Blanc, Europe (West) opened 7,752 new tasks on April 7th, 2024, 7,730 of which coming from services and
- Employee 1965 in ePitoli, Africa (South) opened 6,011 new tasks on May 31st, 2023, 5,560 of which coming from family distress.

A mission employee opening such a large volume of cases in a day is far from being reasonable. Therefore, we believe these ten entries in Table 6 have data entry errors.

### 5.2 Data Cleaning

Data cleaning consists of removing all NA values[2] in the datasets. This process led to the removal of 2 rows (0.0003%) from the Cases dataset and 453 rows (0.072%) from the Programs dataset. We also added a new date column in both datasets that was a combination of the already existing “Day,” “Month,” and “Year” columns for producing better visualizations.

In order to produce better visualizations certain high values were also removed from manipulated datasets. Specifically, this project excludes all instances in the PIMENTO dataset with more than 900 total minutes. Instances with more than 900 total minutes means that a mission employee Worked more than 15 hours a day. As alluded to in the violations section certain days had unreasonable levels of new cases opened. These unreasonable levels translated to visual outliers that drastically skewed spaghetti plots and therefore provided useless visual results. Therefore in order to get a better sense of how the number of cases evolved over time we decided to remove any day that had over 1400 new cases opened that day.

[2] A value of NA refers to no value. In other words, if a cell in a table is NA, the cell is empty.

## 6. Exploratory Analysis

Through examining the distribution and fluctuations in

(1) the number of unique employees per mission in the Program and the Case data sets,

(2) the average number of cases opened per day,

(3) the number of cases opened per day per mission and

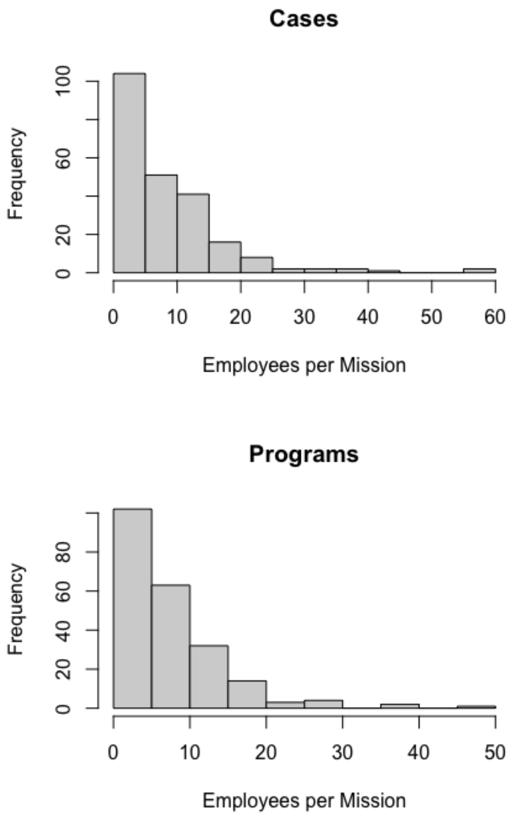
(4) the average amount of time spent per mission per day on cases and programs,

this project provides a general picture of the Borealian consulate network and identifies that mission employees, on average, spent more time on cases than programs during the 2021-2026 period. This project also identifies six stylised facts to highlight this network's key characteristics, strengths and weaknesses.

### 6.1 Exploratory Data Visualisations

**The Number of Unique Employees for Cases and Programs**
We first started by looking at the distribution of the number of unique employees for each mission split by cases and programs (Figure 2).

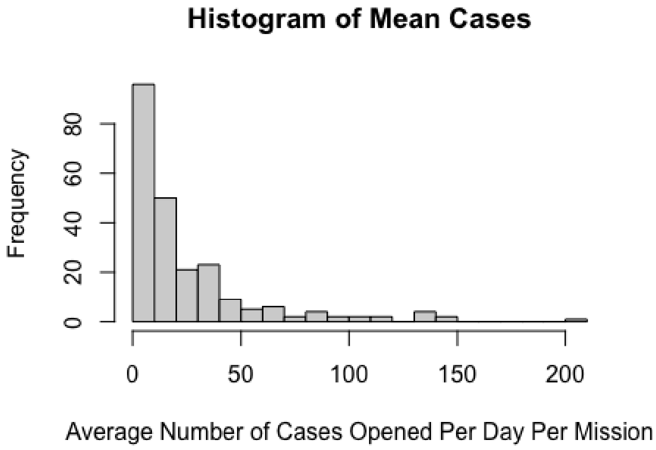
**Figure 2.** Distribution of the Number of Unique Employees per Mission, by Consular Case and Program, 2021-2026



These distributions are relatively similar, which makes sense given that for the most part an employee working on a mission would deal with both cases and programs. In rare cases this is not true, hence the distributions are not exactly the same, but in general this is true.

**The Average Number of Cases Opened per Day per Mission**
The next thing we looked at, based on these distributions, was the average number of cases opened per day per mission (Figure 3). In theory this distribution should be very similar to the number of employees per mission.

**Figure 3.** Distribution of the Average Number of Cases Opened per Day per Mission, 2021-2026

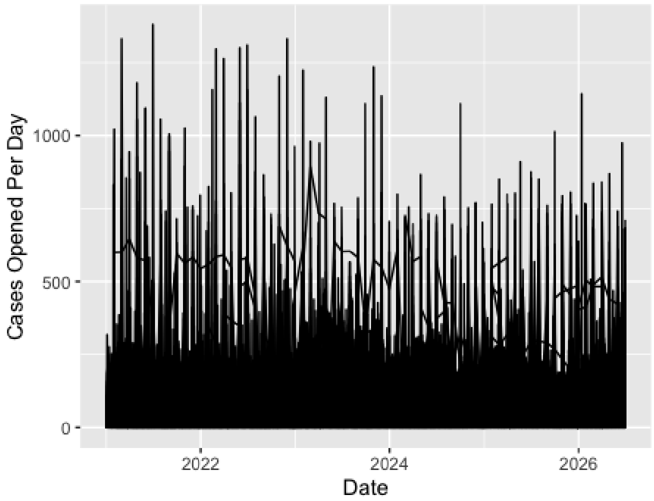


Indeed, this distribution is very similar to that of the number of employees per mission. In theory this indicates that there should be a correlation between number of employees and number of cases opened, however more analysis will be necessary to prove this link.

**The Number of Cases Opened per Day per Mission**
Cases opened per day are also logically not uniform. We looked at the number of cases opened per day per mission to see if there were identifiable trends (Figure 4).

In this graph each line indicates one mission, with each point of the line being the number of cases opened on that day. This graph indicates that there are fluctuations in the number of cases opened per day. Furthermore these fluctuations do appear to hold to some form of pattern. More analysis at a higher definition will be needed to further identify the patterns at the mission level. For example, there may be patterns that indicate that certain days of the week are generally busier than others. These questions are unfortunately beyond the scope of this project. An important conclusion from this graph is that these patterns affirm the fact that cases are not opened uniformly. Different

**Figure 4.** The Number of Cases Opened per Day per Mission, 2021-2026

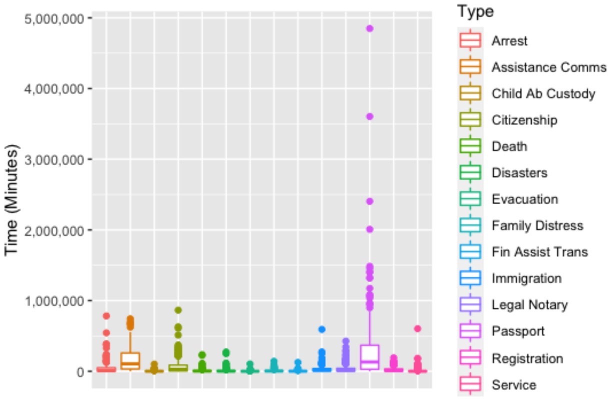


seasonal demands and geographical locations necessarily create some form of schedule of case openings.

### Overall Distributions of Time Spent on Cases and Programs

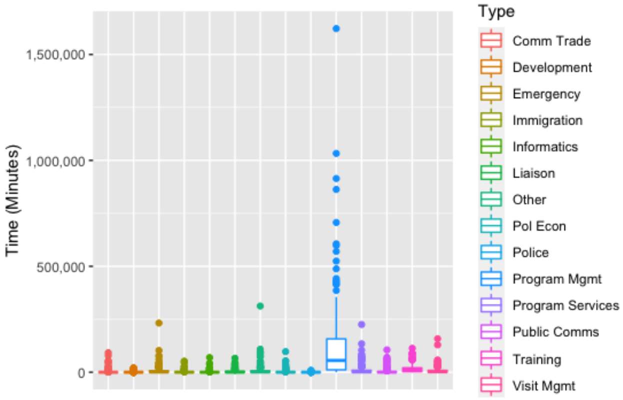
In order to further understand the aspects of the cases and programs we decided to look into the distributions of time spent on the different tasks within cases and programs (Figure 5 and Figure 6).

**Figure 5.** Distribution of Total Time Spent on Case Tasks



The two tasks that immediately stand out are passport for cases and program management for programs. While program management may seem obvious as the the task with most time spent among programs, it is interesting that among cases the most time is spent dealing with passports. This could reflect the fact that consulates quite often would deal with foreign nationals or foreign travellers and this could entail a fair amount of work with passports, thus the higher amount of time spent dealing with them.

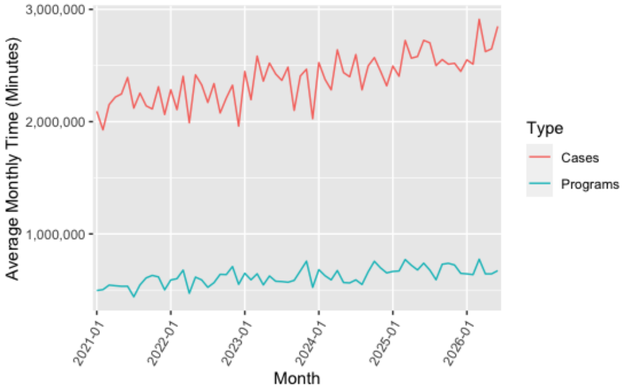
**Figure 6.** Distribution of Total Time Spent on Program Tasks



### Time Spent on Cases vs. Time Spent on Programs

To find clearer resolution on the progression of time spent on cases vs. time spent on programs we decided to look at the time spent for both by visualising the overall monthly means for cases and programs (Figure 7).

**Figure 7.** Mean Amount of Time spent per Month on Cases vs. Programs 2021-2026



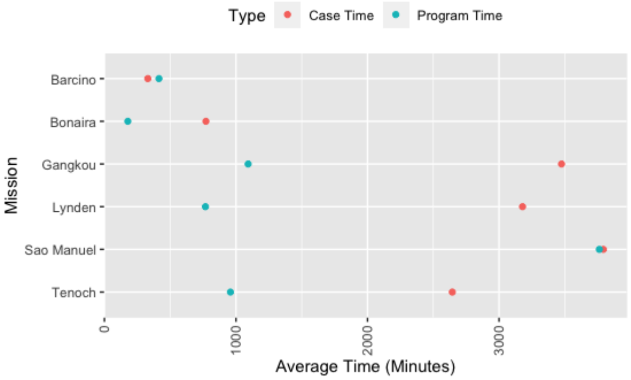
This provided much more clarity on the progression of time spent on cases and programs. The overarching theme is that missions spend, by far, more time dealing cases than running programs. While the variation in the mean monthly time spent on cases appears much higher than the variation in mean monthly time spent on programs, it is important to note that this variation primarily related to relative increases and decreases in cases, typically not the absence of cases. In contrast time spent on programs was actually more inconsistent because there are periods of time where any given mission could not be running any programs. This is likely due to the fact that programs are presumably planned events that do not necessarily occur on set intervals, whereas cases are presented to be dealt with on some form of frequent basis. This makes sense because cases are more routine tasks that a consulate would deal with on an ostensibly regular basis. In comparison programs are not given on a regular basis and probably reflect a given need at a given time for a given consulate.

Separating out cases and programs and delineating by mission allowed us to see which missions spent the highest monthly mean time on cases or programs (Appendix).

### The Average Amount of Time Spent per Mission per Month on Cases vs. Programs for Select Missions

By identifying missions that stood out from the noise in (Figure 17) and (Figure 18) we specifically looked at the difference in average monthly spent between cases and programs for these missions (Figure 8). The three missions selected from the mean amount of time spent per month per mission on cases (Figure 17) were Gangkou, Lynden, and Tenoch. The three missions selected from the mean amount of time spent per month per mission on programs (Figure 18) were Barcino, Bonaira, and San Manuel.

**Figure 8.** Average Time Spent per Month on Cases vs. Programs for Select Missions



While certain missions were selected that stood out for having a high monthly average of program time, the same missions, except Barcino, all still spent more monthly time on average on cases. In fact this sample of 6 missions is a perfect representative of the whole population as 83% of missions spend more time on average on cases than they do on programs. In practice this makes sense as a consulate is most likely dealing with cases on a regular basis, while programs are planned and not run all the time.

To further visualise general differences in average monthly time spent on cases vs. programs we decided to specifically look at the monthly average time for both cases and programs for Barcino (Figure 9) and Gangkou (Figure 10). These missions were specifically chosen because Barcino has a higher average monthly time spent on programs, while Gangkou is more reflective of most missions due to it having a much higher average of monthly time spent on cases.

Interestingly Barcino had a significant decrease in average monthly program time starting around the beginning of 2025. This may hint that program time is inconsistent. This would lend further credence to the fact that programs are not run on a regular basis, unlike cases, which a consulate is more likely to deal with on a regular basis.

**Figure 9.** Average Time Spent per Month on Cases and Programs for Barcino

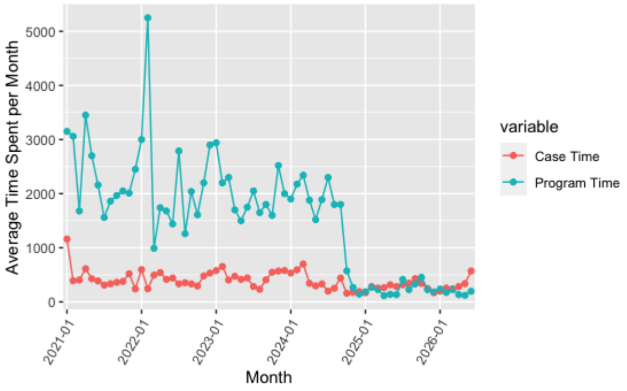
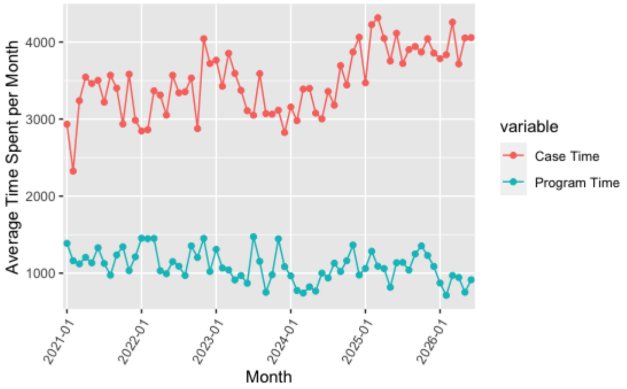


**Figure 10.** Average Time Spent per Month on Cases and Programs for Gangkou



## 6.2 Stylised Facts of the Borelian Consulate Network

After depicting a general picture of the Borealian consulate network, this project discusses six stylised facts about the Borealian consulate network. The first three stylised facts characterise trends in the number of case and program missions, the number of mission employees and the amount of time spent, while the last three facts focus on the role of mission employees in the network.

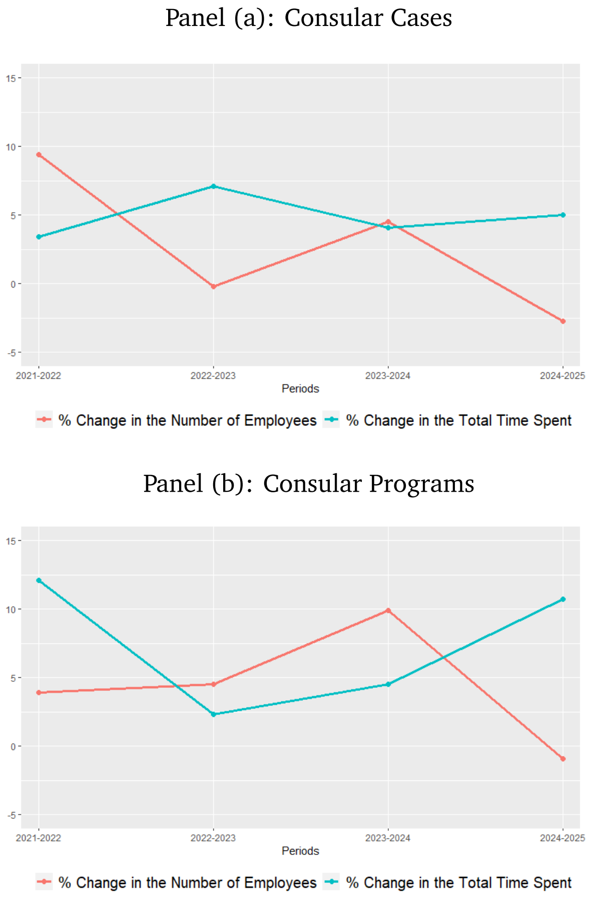
### Fact 1: Stable Number of Program and Case Missions in Each Geographical Region

Throughout the 2021-2026 period, the annual number of missions in each of the 14 geographical regions is very stable. Most of the time, the number of program and case missions did not change. The most significant change occurred between 2023 and 2024 in Africa (North), with five more missions (a 25 percent increase) for both programs and cases. Moreover, there were always fewer consular program missions than consular case missions, although there could be sub-regions with only program missions.

### Fact 2: Roughly Similar Trends in Annual Total Time Spent and Number of Employees for Cases and Programs Except 2024-2025[3]

The direction of movements in the annual total time spent and the number of mission employees for cases and programs were roughly close from 2021 to 2024 (Figure 11). However, during the 2024-2025 period, the annual total time spent increased while the number of mission employees decreased for both cases (-2.7 percent vs. 5.0 percent) and programs (-0.9 percent vs. 10.7 percent).

**Figure 11.** Annual Percentage Changes in the Number of Mission Employees and the Total Time Spent, 2021-2025, percent



Panel (a): Consular Cases

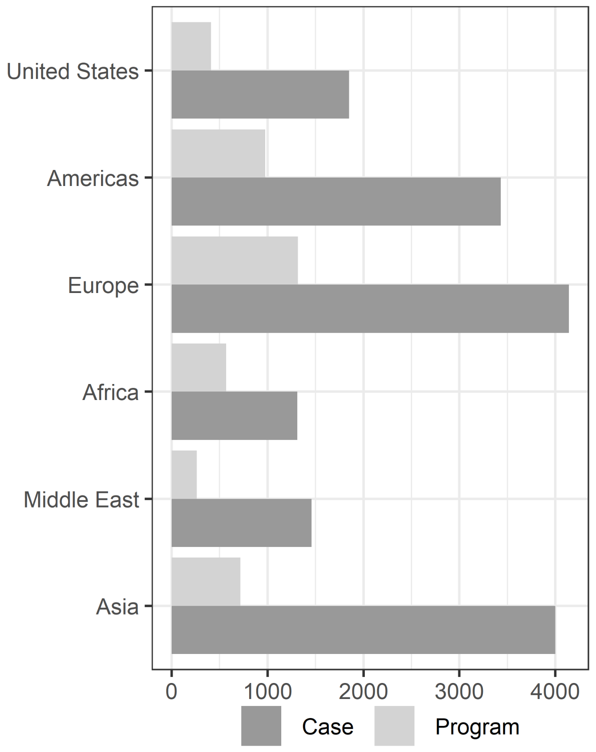
Panel (b): Consular Programs

### Fact 3: The Busiest Continent – Europe

Borelian consulates in Europe was the busiest for both consular programs and cases across all continents during the 2021-2026 period, especially the geographical region Europe (West) (Figures 12 and 13).

Followed by a gentle rise in the total amount of time from 2021 to 2025, there was a sharp decline from 2025 to 2026 (Figures 14 and 15.) This trend is consistent with the evolution of the number of mission employees over time.

**Figure 12.** Total Amount of Time Spent on Consular Programs and Cases by Continent, 2021 - 2026, in 1,000 Minutes



### Fact 4: Understaffing Issue Programs and Cases

Working long hours is detrimental to productivity. Unfortunately, it was common for a mission employee to work more than eight hours a day throughout the 2021-2025 period in the Borealian consular network (Table 7).[3] This understaffing issue was more severe for cases than programs (average annual percentages of longer-than-eight-hour workdays: 5.4 percent vs. 1.0 percent), despite the more routine nature of consular cases than consular programs. The chronic understaffing in all geographical regions except for the United States reflects potential human resource management issues for consular cases among Borealian consulates worldwide.

### Fact 5: Strong Employee Retention

Despite the fluctuating trends in the number of mission employees worldwide, from a rise during the 2023-2024 period to a decline from 2024 to 2026, the BFO had strong retention of mission employees for consular programs and cases during the 2021-2026 period. On average, around 76 percent of program employees in each geographical region stayed for the next year, while about 75 percent of case employees stayed. Table 8 shows the annual percentages of employees remained in the same geographical region, the annual retention rates. Strong employee retention provides the BFO and new employees with local knowledge, practices and experience vital for consular programs and cases.

[3] This project excludes 2026 from this discussion because there are only 6 months of data in the data set for this year.

**Table 7.** Annual Percentage of Overtime (> 8 hours/day) Workday, Cases and Programs, all Geographical Regions, 2021-2025[3]

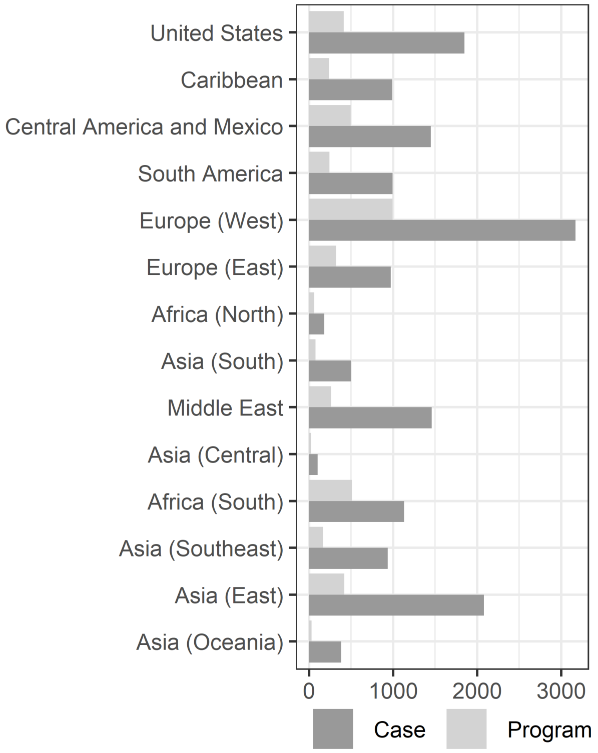
(a) Consular Cases

| Geographical Region | 21 | 22 | 23 | 24 | 25 |
|---|---|---|---|---|---|
| Africa (South) | 5.6 | 5.6 | 5.6 | 5.7 | 6.1 |
| Asia (East) | 5.3 | 5.4 | 5.4 | 5.4 | 5.7 |
| Europe (East) | 5.3 | 5.5 | 5.5 | 5.4 | 5.5 |
| Asia (South) | 5.1 | 5.1 | 5.1 | 5.1 | 5.3 |
| Africa (North) | 5.5 | 5.5 | 5.5 | 5.7 | 6.1 |
| Asia (Oceania) | 5.2 | 5.2 | 5.2 | 5.2 | 5.6 |
| South America | 7.5 | 8.1 | 8.1 | 7.5 | 6.8 |
| Middle East | 7.7 | 8.0 | 7.8 | 7.4 | 7.7 |
| Europe (West) | 7.0 | 7.3 | 7.2 | 6.8 | 7.1 |
| Caribbean | 4.8 | 4.8 | 4.7 | 4.5 | 4.6 |
| Asia (Southeast) | 4.8 | 4.7 | 4.7 | 4.6 | 5.1 |
| United States | 0 | 0 | 0 | 0 | 0 |
| Central America and Mexico | 5.1 | 5.2 | 5.2 | 5.1 | 5.1 |
| Asia (Central) | 5.6 | 5.7 | 5.7 | 5.7 | 6.1 |

(b) Consular Programs

| Geographical Region | 21 | 22 | 23 | 24 | 25 |
|---|---|---|---|---|---|
| Africa (South) | 0.9 | 0.9 | 1.0 | 1.1 | 1.1 |
| Asia (East) | 0.9 | 0.9 | 0.9 | 1.0 | 1.2 |
| Europe (East) | 1.0 | 1.1 | 1.1 | 1.3 | 1.5 |
| Asia (South) | 1.0 | 1.0 | 1.0 | 1.1 | 1.3 |
| Africa (North) | 1.1 | 1.1 | 1.1 | 1.2 | 1.2 |
| Asia (Oceania) | 0.9 | 0.9 | 1.0 | 1.1 | 1.2 |
| South America | 0.8 | 0.8 | 0.9 | 1.1 | 0.9 |
| Middle East | 1.0 | 1.0 | 1.1 | 1.3 | 1.2 |
| Europe (West) | 0.9 | 1.0 | 1.0 | 1.3 | 1.7 |
| Caribbean | 1.1 | 1.1 | 1.1 | 1.2 | 1.5 |
| Asia (Southeast) | 1.0 | 1.0 | 1.0 | 1.1 | 1.3 |
| United States | 0 | 0 | 0 | 0 | 0 |
| Central America and Mexico | 1.1 | 1.2 | 1.2 | 1.4 | 1.7 |
| Asia (Central) | 0.9 | 0.9 | 1 | 1.1 | 1.1 |

**Figure 13.** Total Amount of Time Spent on Consular Programs and Cases by Geographical Region, 2021 - 2026, in 1,000 Minutes



All 14 geographical regions have similar annual retention rates for consular programs and cases. Among these regions, on the one hand, Asia (Oceania) had the highest average annual retention rate for both consular programs and cases (85 percent and 84 percent, respectively.) On the other hand, Asia (Central) had the lowest average annual retention rates for both consular programs and cases (65 percent and 66 percent, respectively.)

### Fact 6: Abundance of Senior Mission Employees Lacking Exposure in Each Geographical Region

A senior employee in a geographical region is a mission employee participating in both consular programs and cases for six years in a roll from 2021 to 2026. Figure 16 shows the number of mission employees participating in both consular programs and this number as a percent of the number of unique employees from 2021 to 2026. South America has the highest percentage of senior employees (49 percent), followed by Europe (West) (44 percent). Europe (West) has the most senior employees (162 senior employees).

Among all senior employees in each geographical region, this project further investigates if some senior employees participated in all types of consular programs and cases. Mission employees with ID 5508 in South America and ID 8718 in the Caribbean are the only senior employees exposed to all types of program and case tasks. Their experience and exposure are valuable to the BFO for improving

**Figure 14.** Evolution of the Total Amount of Time Spent on Consular Cases by Geographical Region Over Time, 2021 - 2026, Minute

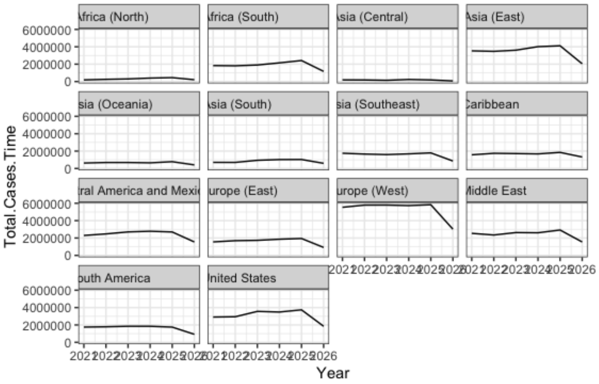
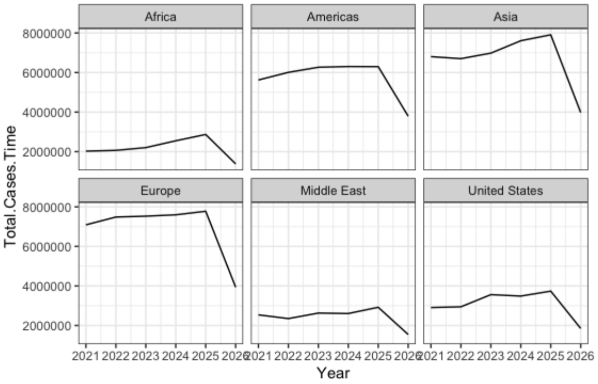


**Figure 15.** Evolution of the Total Amount of Time Spent on Consular Cases by Continent Over Time, 2021 - 2026, Minute



local operations, and the BFO should attempt to cultivate more of this type of expert to facilitate local operations.

Overall, a strength of the Borealian consular network is its strong employee retention and hence an abundance of experienced employees. In all graphical regions except Asia (South) and Africa (North), at least 30 percent of their employees worked in the geographical region for six years in a roll (senior). This number of senior employees and strong employee retention help maintain service quality and continuous long-term tasks, especially in Europe (West). However, chronic understaffing issues can discourage and drive out senior employees. Moreover, most of these senior mission employees were specialised and lacked exposure to other types of tasks. With a stable number of program and case missions, the BFO should provide these senior employees with the opportunity to experience other tasks than those they specialise in.

**Table 8.** Annual Retention Rates of Mission Employees by Geographical Regions, 2021 - 2026, percent

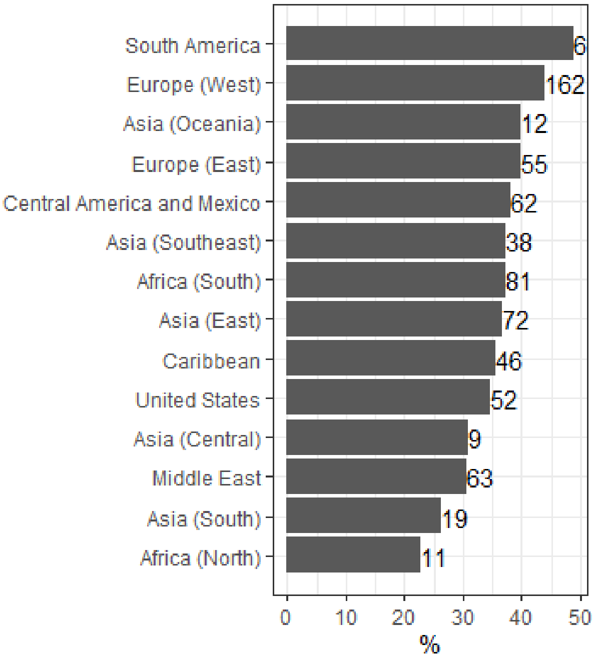
### (a) Consular Programs

| | 21 | 22 | 23 | 24 | 25 | Avg. |
|---|---|---|---|---|---|---|
| | - | - | - | - | - | |
| | 22 | 23 | 24 | 25 | 26 | |
| United States | 88 | 80 | 76 | 72 | 87 | 81 |
| Asia (South) | 84 | 80 | 74 | 77 | 73 | 78 |
| Asia (East) | 76 | 64 | 79 | 81 | 70 | 74 |
| South America | 60 | 90 | 82 | 72 | 67 | 74 |
| Middle East | 83 | 76 | 73 | 68 | 76 | 75 |
| Africa (North) | 82 | 80 | 84 | 74 | 59 | 76 |
| Europe (East) | 67 | 81 | 84 | 74 | 73 | 76 |
| Asia (Oceania) | 83 | 93 | 86 | 75 | 88 | 85 |
| Caribbean | 80 | 80 | 78 | 73 | 86 | 79 |
| Europe (West) | 77 | 76 | 76 | 80 | 74 | 76 |
| Asia (Central) | 67 | 67 | 63 | 78 | 54 | 65 |
| Central America and Mexico | 68 | 79 | 83 | 77 | 82 | 78 |
| Asia (Southeast) | 87 | 79 | 81 | 73 | 63 | 76 |
| Africa (South) | 74 | 71 | 77 | 77 | 75 | 75 |
| Annual Average | 77 | 78 | 78 | 75 | 73 | 76 |

### (b) Consular Cases

| | 21 | 22 | 23 | 24 | 25 | Avg. |
|---|---|---|---|---|---|---|
| | - | - | - | - | - | |
| | 22 | 23 | 24 | 25 | 26 | |
| United States | 87 | 80 | 70 | 69 | 86 | 79 |
| Asia (South) | 81 | 79 | 71 | 81 | 74 | 78 |
| Asia (East) | 67 | 65 | 74 | 80 | 68 | 71 |
| South America | 64 | 81 | 80 | 74 | 71 | 74 |
| Middle East | 82 | 56 | 76 | 63 | 76 | 71 |
| Africa (North) | 71 | 100 | 70 | 78 | 65 | 77 |
| Europe (East) | 72 | 79 | 79 | 72 | 71 | 74 |
| Asia (Oceania) | 73 | 93 | 94 | 70 | 88 | 84 |
| Caribbean | 75 | 78 | 79 | 73 | 83 | 78 |
| Europe (West) | 75 | 75 | 75 | 79 | 71 | 75 |
| Asia (Central) | 69 | 73 | 64 | 75 | 50 | 66 |
| Central America and Mexico | 71 | 76 | 79 | 73 | 70 | 74 |
| Asia (Southeast) | 86 | 79 | 78 | 70 | 62 | 75 |
| Africa (South) | 75 | 68 | 75 | 72 | 76 | 73 |
| Annual Average | 75 | 77 | 76 | 74 | 72 | 75 |

**Figure 16.** The Percentage and the Number of Senior Employees in Each Geographical Region, 2021-2016



**Note:** The number of senior employees in each geographical region is aside each bar. Each bar, in descending order, represents the number of senior employees as a percent of the number of unique employees from 2021 to 2026.

## 7. Clustering Analysis

This project uses clustering techniques to identify groups of similar missions.

### 7.1 Clustering Algorithms

Clustering is a type of machine learning technique to find commonalities between data elements that are otherwise uncategorised. The goal of clustering is to find distinct groups or “clusters" within a data set, based on multiple features (variables) for each individual observed value so that observations belonging to the same cluster are very similar while those belonging to different clusters are dissimilar.

### 7.2 Methodology

This project identifies groups of similar mission through the following three-step procedure. This procedure involves two layers of clustering. The first layer corresponds involves (1) PAM clustering, (2) hierarchical clustering and (3) $k$-means clustering, and the second layer uses the DBSCAN algorithm to combine the results from the first layer. Section 10.2 in the Appendix provides detailed explanation of each of these four clustering algorithms.

**Step 1: The First-level Clustering**

The first-level clustering involves three clustering algorithms (1) $k$-means clustering, (2) hierarchical clustering and (3) PAM clustering.

**$k$-means Clustering** Given a pre-specified number of cluster $k$, this algorithm partitions the entire data set into $k$ non-overlapping clusters based on a dissimilarity measure between two data points. This algorithm aims to achieve the lowest total within-cluster variation. In this case, several values of k were used for each dataset: 4,5,6,8,10 and 15.

**Hierarchical Clustering** Hierarchical clustering does not require a pre-specified number of clusters. This project This method starts with treating each data point as a cluster. Adopting a bottom-up (agglomerative) approach, this algorithm combines two closest clusters at a time based on a linkage measure. This algorithm terminates when every observation is in a single cluster. This project uses the complete linkage and cut the dendrogram to obtain 4,5,6,7,8 and 10 clusters.

**PAM clustering** PAM clustering classifies data points into clusters by partitioning around medoids[4] with a dissimilarity matrix. In this case, the dissimilarity matrix used was the gower distance. This algorithm is then ran with several values of k: 4,5,6,8,10 and 15.

Using each of the above three algorithms, this project classifies all missions based one of the following six features at a time.

- Total time spent on cases
- Total number of cases
- Total number of programs
- Monthly programs, cases, high-traffic months, and low-traffic months
- Yearly programs, cases, high-traffic years, and low-traffic years
- Summary statistics[5]

In total, there are 18 sets of clustering results.

**Step 2: Creating a Similarity Matrix of all Missions**

This project, then, creates a similarity matrix[6] based on clustering results from Step 1. Each entry in this similarity

[4] A medoid is a cluster within a data set whose sum of dissimilarities to all the objects in the cluster is minimal

[5] We thank Josh Larock and Mourad Ben Rejeb to provide us with this data set.

[6] This similarity matrix is a square and symmetric matrix with the number of rows being the number of missions. All of its diagonal entries are ones because a mission must be the same as itself and hence have the highest degree of similarity. Other entries in this matrix are strictly less than one.

matrix contains a relative frequency ranging between 0 and 1 and signifies the degree of similarity between two clusters. A higher value of an entry represents higher similarity between the two clusters.

**Step 3: The Second-level Clustering**

This project uses the DBSCAN algorithm explained below with the similarity matrix from Step 2 to classify all missions into clusters.

**DBSCAN Algorithm:** The DBSCAN algorithm starts by picking a random point x from the data set and assigns it to cluster 1. Then it counts how many points are located within a distance of $\epsilon$ from x. If this quantity is greater than or equal to minPoints(n), then x is considered as a core point and all these data points are assigned to the same cluster as x (In this case, it is cluster 1). It will then examine each member of cluster 1 and find the neighbours who are again at a distance of at most $\epsilon$ from that member. If some member of cluster 1 has n or more such neighbours, it will expand cluster 1 by putting those neighbours to the cluster. It will continue expanding cluster 1 until there are no more examples to put in it. In the latter case, it will pick another random point from the data set, not belonging to any cluster and put it to cluster 2. It will continue like this until all examples either belong to some cluster or are marked as outliers.

### 7.3 Results of Clustering Analysis

The DBSCAN algorithm in Step 3 has categorised missions into five groups. Table 9 in the Appendix shows all missions in each of the five groups. Unfortunately, this project cannot find natural cluster structures in the PIMENTO data set.

The results look ok to us because we said an indication that our results might be useful would be if the values in the similarity matrix are mostly close to 1 and close to 0. When we tested to see if that actually occurred, we looked at the percentage of values in the similarity matrix that were below 0.35 and above 0.75. That number was around 60 percent. Which is not ideal but it is good enough. When we created the clusters using DBSCAN, we see that about half of the missions are listed as outliers (or “not similar"). The other half were grouped into four clusters that are fairly evenly distributed. This is consistent with how we would expect the clustering results to look like.

## 8. Recommendations

Based on the analysis in sections 6 and 7, this project makes the following recommendations to improve operations of Borealian consulates worldwide.

### 8.1 Rotational Programs for Talented and Senior Employees

As discussed in section 6, Borealian consulates have strong employee retention and an abundance of experienced mission employees lacking exposure to different tasks in all geographical regions. To develop each employee for better operations in each consulate, the BFO should organise different rotational programs targeted at its talented mission employees and its senior mission employees. The main purpose of these programs is to increase their exposure and understanding of operations in the consular network at different levels. The program for talented employees should emphasise mission task implementation, while the program for senior employees should focus more on the overall picture and the management.

**Rotational Program for Talented Employees**

This rotational program for talented employees should focus on diversifying their exposure to different programs and cases in the geographical region. With a more holistic understanding of tasks in the geographical region, their performance will improve. During the program, the BFO should also identify the best-suited tasks for each of these talents. This identification process facilitates the training of specialised employees and hence allows for more efficient use of human resources.

**Rotational Program for Senior Employees**

Senior mission employees with specialised experience in certain tasks for years are valuable assets to the consular network. As discussed in section 6, the BFO should also expand the exposure of these employees. Unlike the rotational program for talented employees, senior employees should focus more on understanding the big picture and other geographical regions. This experience fosters communication and cooperation among different missions and geographical regions and enables one geographical region to learn from another, improving operations in the network. This program also prepares these employees to take a more significant leadership role in the Borealian consular network.

Moreover, these two rotational programs provide stronger development opportunities and clearer promotion paths to mission employees. These benefits to mission employees further strengthen the BFO’s already strong employee retainment and address employees’ lack of exposure to diverse tasks.

### 8.2 Mission Advisory Committee

This project has an initial idea of recommending the BFO to start a mission advisory committee composed of senior mission employees from all geographical regions. The purpose of this committee is to regularly review mission implementation in all geographical regions and suggest ways for improvement.

### 8.3 Geographical Region Support Network

The BFO can use similar missions from Table 9 in the Appendix to develop a geographical support network. This support network can facilitate crisis management of the

Borealian consular network. Suppose there is a crisis in a geographical region or a sub-region. If there is a crisis in a geographical region, the BFO can shift mission employees from similar regions in the same cluster identified to the region in crisis for assistance.

### 8.4 Restructuring the Borealian Consulate Network: Case Hubs and Regional Hubs

One recommendation, that comes from looking at Figure 5, is case hubs. Unlike programs, cases are not necessarily mission-specific. With such a high proportion of time spent specifically on passports it may be prudent to look into having centralised case hubs. These case hubs would be highly specialised in dealing with passport related issues. This could prove more efficient as it allows consulates the ability to focus their attention on other cases and programs and not have to spend lengthy amounts of times dealing with passports and alleviate understaffing issues discussed in section 6.2.

Although this project cannot identify natural clusters, as discussed in Section 7.3, we would like to suggest the BFO to set up regional hubs to promote shared efforts for programs and cases among consulates in close proximity. Increased co-operations among missions could help establish commonalities among missions. In time, with more commonalities among missions, the BFO can then further improve resource allocations among consulates.

### 8.5 The PIMENTO Database

With this experience working with the PIMENTO database, this project would like to suggest the following ways to improve the PIMENTO database to facilitate the BFO's future operations.

#### Annual Maintenance of the Database

Section 5 identifies some instances in the PIMENTO database with missing values and inconsistencies. The BFO can avoid these problematic instances in the database by performing annual maintenance checks of the database. A consistent and reliable mission database can facilitate the BFO's ongoing mission monitoring for more efficient human resource allocation. A sudden increase in case counts can signal a future crisis so that the BFO can prepare beforehand. Unfortunately, albeit interesting, crisis prediction from the PIMENTO database is beyond the scope of this report.

#### Standardised Training For All Employees

As above, section 5 identifies some issues within the PIMENTO database. We believe that in addition to annual maintenance of the database, standardised yearly training should be introduced for all employees, with special training for new employees as part of the on-boarding process with a specific focus on the data input process. This training will be to ensure that best-practices for data input are followed in order to maximise accuracy, consistency, and precision. The standardisation of input procedures should mitigate errors such as inputting time spent on a task only when the task is completed indicating an unreasonable amount of time spent in one day.

#### Additional Information

For the PIMENTO data set to provide a more complete picture of the Borealian consulate network, we recommend the BFO to expand the current workday-centric PIMENTO data set to include another data set that links unique identification numbers of each task with the time each employee spends on each task. This task-centric data set will be a valuable addition to develop a better understanding of the nature of the Borealian consulate network.

## 9. Conclusion

In conclusion data visualisation and clustering methods have provided a deeper understanding of the PIMENTO database. We were able to identify important characteristics in missions, such as the higher average monthly time spent on cases rather than programs (Figure 7). This provided us with important conclusions. For example we surmised from this that cases occurred more frequently than programs because programs only occurred as a planned event provided by the consulate, whereas cases were dealt with on a day-to-day basis. In our clustering analysis we were able to identify similar missions through the DBSCAN algorithm, resulting in 5 different groups of missions (Table 9). Through this we were able to provide recommendations in order to increase efficiencies within the Borealian consulate network.

# 10. Appendix

This Appendix accompanying this project aims to facilitate a more detailed and technical understanding of this project.

## 10.1 Exploratory Visualization

**Figure 17.** Mean Amount of Time spent per Month per Mission on Cases, 2021-2026

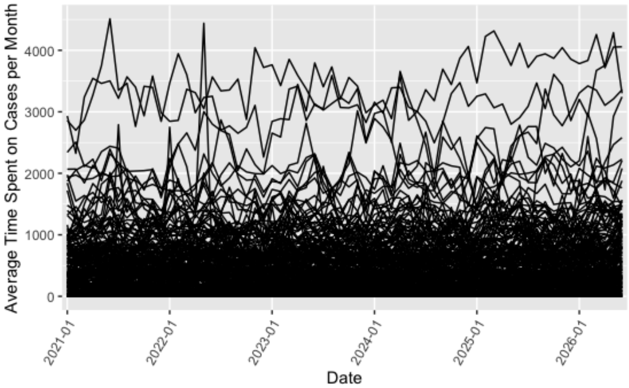
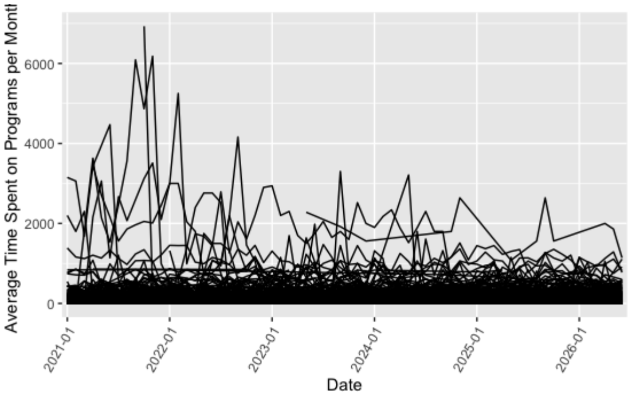


**Figure 18.** Mean Amount of Time spent per Month per Mission on Programs, 2021-2026



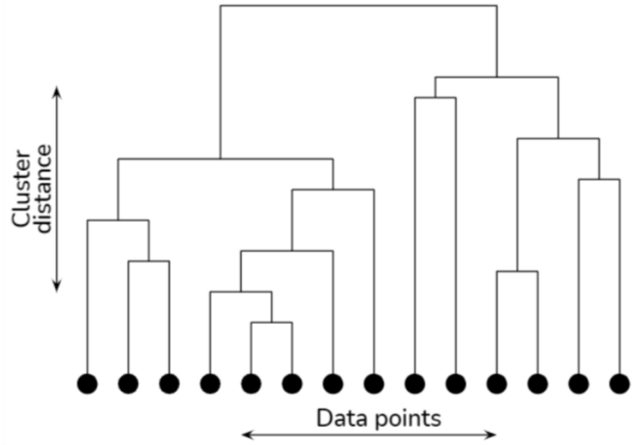
## 10.2 Clustering Analysis

This subsection aims to provide readers with a detailed explanation of each of the four clustering algorithms used in this project: (1) hierarchical clustering, (2) PAM clustering, (3) *k*-means clustering and (4) DBSCAN algorithm.

### 10.2.1 Hierarchical Clustering

The sole concept of hierarchical clustering lies in the construction and analysis of a dendrogram. A dendrogram is a tree-like structure that explains the relationship between all the data points in the system. While constructing the dendrogram, no assumptions is made about the number of clusters. Instead, these are obtained by slicing the dendrogram horizontally. All the resulting child branches formed below the horizontal cut represent the clusters. The location of the slicing is usually done by visually analysing the dendrogram. There are two clustering techniques that can be used to build the dendrogram:

**Figure 19.** Structure of a dendrogram



1. Agglomerative Clustering (bottom-up approach)
2. Divisive Clustering (top-down approach)

In agglomerative hierarchical clustering, each data point is considered as an individual cluster at first. A metric is then defined to measure the distance between all pairs of subclusters at each step and keep merging the nearest two subclusters in each step. This procedure is repeated till there is only one cluster in the system. In this type of clustering, the choice of linkage method is also important, i.e, from where the distance is computed between the clusters. There are several types of linkage methods:

- Single Linkage - is defined as the distance between two closest points in two clusters.
- Complete Linkage - is defined as the distance between two farthest points in two clusters.
- Average Linkage - is defined as the average distance between all points in the two clusters.
- Centroid Linkage - is defined as the distance between the centroids of two clusters.
- Ward's Linkage - computes the total variance between pairs of clusters and merge the ones that have the smallest variance.

In divisive hierarchical clustering, the whole data is considered as a single cluster at first. A method is then required to split this initial cluster. All resulting clusters are then split recursively until individual data have been split into singleton clusters (or the desired number of clusters). In order to determine which cluster to split, the SSE (sum of squared errors) of each cluster is computed and the one with the highest SSE is split. One way to split the cluster is to use Ward's criterion to chase for the largest reduction in the difference in the SSE criterion as a result of the split. In general, divisive clustering is more computationally expensive and has a higher complexity than agglomerative clustering and is less used. For our data set, we have used agglomerative clustering.

### 10.2.2 PAM Clustering Algorithm

To be able to use the PAM clustering algorithm, we first needed to define a notion of similarity (or dissimilarity) amongst observations. We chose the Gower distance. The Gower distance is computed as follows:

$$d(i,j) = \frac{1}{p}\sum_{f=1}^{p} d_{ij}^{(f)} \tag{1}$$

In other words, the distance between observation i and observation j is the average of all feature-specific distances (where p = the total number of features).

An important observation is that the Gower Distance will always fall in the range [0,1]. It will be 0 if two records are identical and 1 if they are opposite. The Gower distance in R will be computed by the daisy function in the cluster package. So we create a dissimilarity matrix that indicates the level of dissimilarity for each record. This is important because it puts a limitation on the data that we can use. The record that is being clusters cannot have more than one value in each column.

For the clustering algorithm, we chose the PAM clustering algorithm (partitioning around medoids) because it allows us to cluster on mixed data types, i.e. categorical and continuous data unlike the k-means algorithm, which can only cluster numeric data. The PAM algorithm is more robust to noise and outliers compared to the k-means algorithm because it uses the dissimilarity matrix as compared to the mean, which can be very sensitive to outliers. However, a significant disadvantage is that both runtime and memory are quadratic for this algorithm, so it is computationally inefficient.

Lastly, we needed to choose an optimal number of clusters. This was achieved by using the silhouette coefficient. Intuitively, the silhouette coefficient for a particular observation p is defined as follows: The mean distance between the observation p and all other data points in the same cluster is called the mean intra-distance cluster and is denoted by $a$. The mean distance between the observation p and all other data points of the next nearest cluster is called the mean nearest-cluster distance and is denoted by $b$.

The silhouette coefficient ($S$) is then computed as follows:

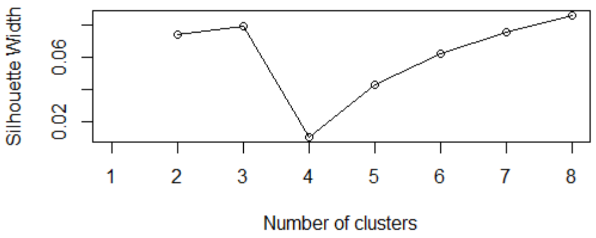
$$S = \frac{(b-a)}{max(a,b)} \tag{2}$$

For the best possible clustering, we want the mean intra-distance cluster to be as small as possible while the mean nearest-cluster distance to be as large as possible so as to have compact, well-separated clusters.

The value of the silhouette coefficient varies from -1 to 1. If the score is 1, the cluster is dense and well-separated than other clusters. A value near 0 represents overlapping clusters with samples very close to the decision boundary of the neighbouring clusters. A negative score $[-1,0]$ indicates that the samples might have got assigned to the wrong clusters.

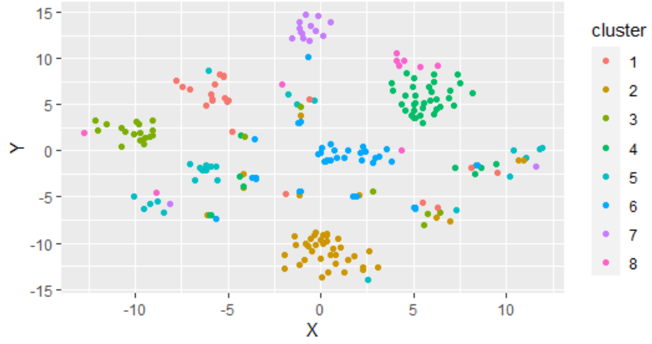
The silhouette figure for this new data set is shown below: Figure 20 shows that the best number of clusters would be either 3 or 8.

**Figure 20.** The Silhouette Figure for the New Data Set



Choosing the number of clusters as 8, we get the following result (Figure 21):

**Figure 21.** Clustering Results with 8 Clusters



### 10.2.3 K-Means Clustering Algorithm

K-means aims to partition data into k clusters by using a similarity measure, usually the euclidean distance, and the notion of centroids (the center of a cluster). Unlike hierarchical clustering, the number of clusters k needs to be predetermined by using either the silhouette or elbow method. K-means is an iterative process and works as follows: k centroids are randomly selected and the distance of all data points to the centroids are calculated. The data points are then assigned to the nearest cluster. The new centroids of each cluster are calculated by taking the mean of all data points in that particular cluster. These steps are then repeated until all points converge and the cluster centres (centroids) stop moving.

We can think of k-means as computing a distance between two points that we say corresponds to the similarity of those two points. Then, we are given a number of clusters k. Based on the distance measure, we can say we can divide up our objects into these clusters. Different values of k will give different results. So, in our case we chose several values of k and compute the clusters in each case. This will protect against the variability associated with choosing a single k.

### 10.3 DBSCAN

DBSCAN stands for density-based spatial clustering of applications with noise. It is a clustering algorithm that works on the assumption that clusters are dense regions in space separated by regions of lower density. The DBSCAN algorithm uses two parameters:

1. eps($\epsilon$) - is the radius of the circle to be created around each data point to check the density in the neighbourhood of that point. In higher dimensions, $\epsilon$ can be viewed as the radius of a hypersphere.

2. minPoints(n) - is the minimum number of points clustered together for a region to be considered dense. (The number of points required inside the circle/hypersphere for it to be considered dense).

The DBSCAN algorithm starts by picking a random point x from the data set and assigns it to cluster 1. Then it counts how many points are located within a distance of $\epsilon$ from x. If this quantity is greater than or equal to minPoints(n), then x is considered as a core point and all these data points are assigned to the same cluster as x (In this case, it is cluster 1). It will then examine each member of cluster 1 and find the neighbours who are again at a distance of at most $\epsilon$ from that member. If some member of cluster 1 has n or more such neighbours, it will expand cluster 1 by putting those neighbours to the cluster. It will continue expanding cluster 1 until there are no more examples to put in it. In the latter case, it will pick another random point from the data set, not belonging to any cluster and put it to cluster 2. It will continue like this until all examples either belong to some cluster or are marked as outliers.

There are 3 types of data points that can be observed when implementing DBSCAN.

- Core Point - is a data point that has at least minPoints (n) within a distance of $\epsilon$.

- Border Point - is a data point that has at least one core point within a distance of $\epsilon$ and lower than minPoints (n) within a distance of $\epsilon$ from it.

- Noise Point - is a data point that has no core points within a distance of $\epsilon$.

DBSCAN is very sensitive to the values of epsilon and minPoints. Therefore, it is important to make a good choice for these two parameters. For minPoints(n),as a starting point, n can be derived from the number of dimensions D in the data set, as $n \geq D + 1$. For data sets with noise, larger values are usually better and will yield more significant clusters. Hence, $n = 2D$ can be chosen, but it may even be necessary to choose larger values for very large data. As for $\epsilon$,choosing too small a value will cause a large part of the data to remain unclustered. Whereas, for a too high value, the clusters will merge and the majority of objects will be in the same cluster. The optimal value of $\epsilon$ can be decided from the K-distance graph, which we will not explain here. The DBSCAN algorithm provides several advantages over previously seen clustering algorithms such as K-means. It does not require any apriori knowledge of the number of clusters and it is very robust to outliers. With a clustering algorithm such as K-means, every observation (even an outlier) will become a part of a cluster eventually whereas DBSCAN clearly separates outliers and clusters.

## 10.4 Results of Clustering Analysis

The following table lists all missions within each of the five clusters identified in Step 3 by the DBSCAN algorithm.

**Table 9.** Missions in Each of the Five Clusters

Cluster 1

| MissionTitleE | GeoRegionNameE |
|---|---|
| Adsaloma | Africa (South) |
| Au Lac | Asia (Southeast) |
| Bigowon | Asia (Southeast) |
| Birmingham | Europe (West) |
| Borgen | Europe (West) |
| Cebu | Asia (Southeast) |
| Charles Town | Caribbean |
| Ciudad Betiz | Central America and Mexico |
| Kasim | Asia (South) |
| Keupenhavn | Europe (West) |
| Lahiri | Asia (South) |
| Mwalusa | Africa (South) |
| Rubingisa | Africa (South) |
| San Martin | Caribbean |
| Serdica | Europe (East) |
| Shedden Harbour | Caribbean |
| Soba | Africa (South) |
| Tanith | Africa (North) |
| Tarabulus | Africa (North) |

Cluster 2

| MissionTitleE | GeoRegionNameE |
|---|---|
| Airy | Middle East |
| Amibo | South America |
| Dajti | Europe (East) |
| Hamburg | Europe (West) |
| Jabuuti | Africa (South) |
| Laibach | Europe (East) |
| Malake | Europe (West) |
| Ritgo | Europe (East) |
| Rueda | Central America and Mexico |
| Wouri | Africa (South) |

Cluster 3

| MissionTitleE | GeoRegionNameE |
|---|---|
| Addasibaba | Middle East |
| Aitioch | Africa (North) |
| Al-geris | Middle East |
| Amani | Europe (East) |
| Arifwala | Asia (South) |
| Baile Atha Cliath | Europe (West) |
| Bann | Europe (West) |
| Beirite | Middle East |
| Bridetown | Caribbean |
| Bukovie | Europe (East) |
| Chaguanas | Caribbean |
| Damascia | Middle East |
| ePitoli | Africa (South) |
| Funan | Asia (Southeast) |
| Irishton | United States |
| Kenitra | Africa (North) |
| Limona | South America |
| Loong | Asia (Southeast) |
| Magra | Europe (West) |
| Malacanang | Asia (Southeast) |
| Mughalhi | Asia (South) |
| Queenstown | Caribbean |
| Randstad | Europe (West) |
| Remutaka | Asia (Oceania) |
| San Guevara | South America |
| St. Francis | United States |
| Sungai | Asia (Southeast) |
| Vienna | Europe (West) |
| Warcislaw | Europe (East) |
| Yadub | Middle East |

Cluster 4

| MissionTitleE | GeoRegionNameE |
|---|---|
| Bakoumour | Africa (South) |
| Batte | Africa (South) |
| Chari et Logone | Africa (South) |
| Dzhunushaliev | Asia (Central) |
| Erebuni | Europe (East) |
| eThekwini | Africa (South) |
| Freetown | Africa (South) |
| Fuladiwosituoke | Europe (East) |
| Hadid | Middle East |
| Huidobro | South America |
| Konaakiri | Africa (South) |
| Korsou | Caribbean |
| Livingstone | Africa (South) |
| Lugville | Europe (West) |
| Maputsu | Africa (South) |
| Martintar | Asia (Oceania) |
| Mukherjee | Asia (South) |
| Napule | Europe (West) |
| Nawaksut | Africa (North) |
| Perai | Asia (Southeast) |
| Polibeka | Europe (West) |
| Port Heywood | Africa (South) |
| Pressburg | Europe (East) |
| Qart Hadasht | South America |
| Rysel | Europe (West) |
| Saint Patrick | Europe (West) |
| San Ramon | Caribbean |
| Schduagert | Europe (West) |
| Sesotho | Africa (South) |
| St-Petersburg | Europe (East) |
| St. James | Caribbean |
| Sulmona | Europe (West) |
| Surabaya | Asia (Southeast) |
| Wubri | Africa (South) |
| Yu | Asia (East) |

Cluster 5:

| MissionTitleE | GeoRegionNameE |
|---|---|
| Angel City | United States |
| Attila | Europe (West) |
| Bank Krung Thep | Asia (Southeast) |
| Everglades | United States |
| Ludovia | Europe (West) |
| New Albany | United States |
| Reme | Europe (West) |