Understanding relationships between geography, tourism and economy for Small Island Developing States with single airports

Shivani Nikam

Submitted for the Degree of Master of Science in

Data Science and Analytics

Department of Computer Science
Royal Holloway University of London
Egham, Surrey TW20 0EX, UK

September 3, 2019

Declaration

This report has been prepared on the basis of my own work. Where other published and unpublished source materials have been used, these have been acknowledged.

Word Count: 9151

Student Name: Shivani Nikam

Date of Submission: 03/09/2019

Signature:

1 Acknowledgement

I'd like to take this opportunity to thank my supervisor Dr Hugh Shanahan for all the guidance and support he has provided me with in the last few months. He always made time for all my queries and pushed me towards completing the project. I am grateful to the faculty that has imparted so much useful knowledge which allowed me to take up this project. At the end I'd like to thank my parents for giving me the opportunity to study in such a prestigious University. Thank you all.

Contents

2 Abstract

Small Island Developing States are a group of islands littered across the world which deal with similar problems and issues in their development and growth. There is great similarity in the development of SIDS which are close to each other geographically[1,2]. Tourism has been by far the biggest economic contributor for these countries and countries that have developed their tourism associated infrastructure have seen tremendous growth relative to the other SIDS[3]. The SIDS harbor exotic natural beauty as well as flora and fauna which makes them perfect for all types of tourists from nearby countries as well as enthusiasts from across the world. There has been great interest recently in understanding the similarities and differences in the economic development of Small Island Developing States (SIDS) with single airports[3,4]. This project uses a data-driven approach to discover patterns of similarities between the economic and tourism profiles of SIDS. I have used unsupervised machine learning to discover clusters of SIDS within a dataset containing information about tourist arrivals and economic conditions of 27 different SIDS from the year 200 to 2011. A key finding was that often geographically close countries show similar patterns across these criteria. I have shown correlation between tourist arrivals from various developed countries in the world and how they correlate to the economy of the SIDS. This work can serve as a foundation for further data-driven approaches towards issues being faced by SIDS. Furthermore, this work can also help in SIDS being able to identify the tourists who are contributing to their tourism-based revenue the most and take actions to further attract them.

3 Introduction

Small Island developing states (SIDS) are a group of small island nations littered across the world that share similar challenges such as small but growing population, limited resources, limited connectivity to the rest of the world, and susceptibility to natural disaster. Many of these countries also rely heavily on global trade. SIDS also have limited availability of human, institutional and financial resources to manage and use natural resources on a sustainable basis and face ever increasing demographic and economic pressures on existing natural resources and ecosystems. Their remoteness and the lack of opportunities for scaling up the economy stunts their growth and often leads to periods of stagnation. Blessed with natural and exotic beauty, these islands often turn to tourism as the source of revenue and development. Many of these islands have a single airport and relatively low other means of connectivity to the world which makes it harder for the tourism industry to really bloom[1].

Fig 1. Geographical locations of various SIDS around the world[5].

According to the World Tourism Organization (2000) report and estimates, the travel and tourism industry would experience an increase in the number of international tourists to a tune of 1602 million by the year 2020. With rising worldwide tourism, many of these small island nations are hoping to take advantage and grow economically and many have already shifted away from traditional international trade and export to the service industry as the primary source of revenue[3]. Tourism development does not only enhance foreign exchange earnings, but also provides job opportunities, invigorates tourism industry growth, and thereby stimulates overall economic growth. The tourism sector shares a special link with the agriculture sector as the agriculture industry can supply much-needed produce to the tourism sector to be served in the hotels and restaurants. Strengthening the links between the two industries is perceived to be a useful way benefit the local economies[6,7].

The magnitude of the economic benefits that accrue from tourism also depend heavily on the degree of good governance in SIDS. Lack of investment, in general and specifically in tourism facilities and attractions, corruption and parochialism, lack of institutional accountability and failure to plan and implement policies have dampened the potential for tourism to benefit the small islands[4]. Because of the unique challenges that they face and the need to

ensure sustainability, island states need to pay particular attention to carrying capacities, community involvement, the dynamic political environment and special interest activities[8].

Tourism sector growth is more apparent in the Caribbean and Mediterranean SIDSs, when compared with the African and Asia-Pacific regions. For instance, Cyprus and Malta over the years have experienced advancement in tourism, which manifested in the aggregated foreign exchange earnings and enormous percentage share of tourism in the GDP[3]. The sector has been regarded as the most significant source of foreign currency earnings. Data from the World Bank (2017) report show a commensurate increase in international tourist arrival and receipts. Similarly, tourism is the single largest earner of foreign exchange in 16 of the 30 countries in the Caribbean. In 2006, Caribbean Tourism Organization members hosted 22.2 million overnight international arrivals, who spent, in total $US 21 billion. This means destinations with less than 1% of the world's population attract approximately 3% of worldwide tourism. Caribbean SIDS also outperform the their Pacific / Indian counterparts on a range of socioeconomic and demographic variables including GDP per capita, life expectancy, and infant mortality[9,10]. Three underlying reasons for these differences are given: the geographic proximity to major global markets; early post-war development of international tourism; and a longer and more intense period of colonization that led to the early establishment of basic infrastructure and market institutions[11,12]. On the other side of things, the level of development and its contribution to economic growth/development has been argued to be more uneven and particularly much lower in the SIDS in the Asia-Pacific, Oceanic and African regions. six of these states, such as Kiribati, Maldives, Solomon Islands, Samoa, Tuvalu, and Vanuatu, have been categorized as the least developed states[3].

4 Background

4.1 The dataset

The dataset I have used for this project was provided to me by my supervisor which can be downloaded from the following link:

https://drive.google.com/drive/folders/1ZKHZ84zOOm3drPk1B-3dhEUCGUAYH9dV?usp=sharing

It consists of 314 observations of 67 variables for 27 different Island nations with a single airport. The data observed was collected from the year 2000 to the year 2011. The aim of this project was to utilize this dataset in order to find relationships and similarities between these Island nations. The list of countries in the dataset has been provided below along with a list of variables.

4.1.1 Countries observed

```
##  [1] "Mauritius"                 "Seychelles"
##  [3] "Antigua and Barbuda"       "Grenada"
##  [5] "Bahrain"                   "Barbados"
##  [7] "Bermuda"                   "Cape Verde"
##  [9] "Comoros"                   "Dominica"
```

```
## [11] "Kiribati"                    "Maldives"
## [13] "Malta"                       "Marshall Islands"
## [15] "Micronesia"                  "Samoa"
## [17] "Sao Tome and Principe"       "Saint Kitts and Nevis"
## [19] "Saint Lucia"                 "Cayman Islands"
## [21] "Saint Vincent & Grenadines" "Tonga"
## [23] "Tuvalu"                      "Palau"
## [25] "Singapore"                   "Trinidad and Tobago"
## [27] "Solomon Islands"
```

4.1.2 Variables observed

The variables consist of basic information such as land area, population, GDP as well as fairly granular information such as tourist arrival counts from various regions of the world.

```
##  [1] "country"      "year"         "pop"          "areakm2"
##  [5] "gdpnom"       "flights...WB" "hotels"       "hotrooms"
##  [9] "expend"       "receipt"      "ovnarriv"     "dayvisit"
## [13] "crusvis"      "arrpleas"     "arrbus"       "arrair"
## [17] "arrwat"       "emptour"      "carrycap"     "exptour"
## [21] "expbus"       "travx"        "travpas"      "arrafr"
## [25] "arram"        "arreap"       "arreur"       "arrme"
## [29] "arrse"        "arroth"       "arrausl"      "arrbel"
## [33] "araus"        "arrfra"       "arrger"       "arrchn"
## [37] "arrit"        "arrnet"       "arrswz"       "arrind"
## [41] "arruae"       "arrreu"       "arrsey"       "arrspa"
## [45] "arrsin"       "arrsaf"       "arrrus"       "arrswe"
## [49] "arruk"        "arrmad"       "ocrooms"      "ocbed"
## [53] "avstay"       "avcap"        "intrxgdp"     "ottexpdg"
## [57] "tbalgdp"      "topen"        "tcov"         "intxexg"
## [61] "intxexs"      "intxexal"     "intxcac"      "oteximg"
## [65] "otxims"       "otximal"      "otxcad"
```

4.1.3 Key observations

- The data is littered with missing values.
- Certain countries may have to be excluded as they might be outliers in terms of very high GDP due to excellent economic growth.
  - The high GDP also leads to high overall development in various sectors of the industry.
  - An example is Singapore.
- A large proportion of data is missing for years before 2007.

4.2 Data Cleaning

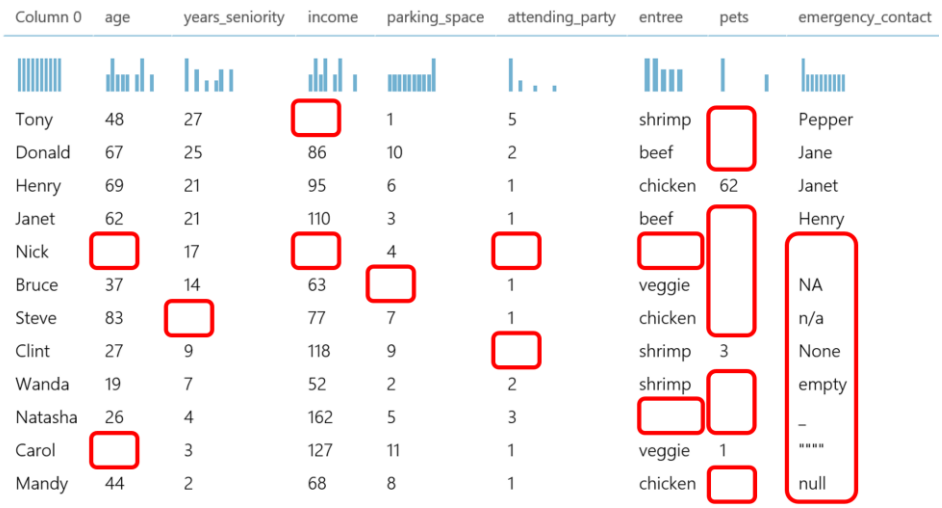
4.2.1 Missing Values

Real world data sets are often littered with missing data. An example of this is given in Fig. 2. This can be due to a variety of reasons as I will discuss below. Having missing data in our

data sets can often mislead us in terms of the patterns we capture and quite often hide some patterns from us. This means that handling missing data is paramount for building reliable and robust machine learning models as well performing accurate data analysis[13].

4.2.1.1 Reasons for missing values

- Data Extraction: It is possible that there are problems with extraction process. Errors at data extraction stage are typically easy to find and can be corrected easily as well. In such cases, we should double-check for correct data with the data warehouse. Some hashing procedures can also be used to make sure data extraction is correct[14].
- Data collection: These errors occur at time of data collection and are harder to correct. They can be categorized in four types[14].
- Missing completely at random: This is a case when the probability of missing variable is same for all observations. For example: respondents of data collection process decide that they will declare their earning after tossing a fair coin. If a head occurs, respondent declares his / her earnings & vice versa. Here each observation has equal chance of missing value[14].
- Missing at random: This is a case when variable is missing at random and missing ratio varies for different values / level of other input variables. For example: We are collecting data for age and female has higher missing value compare to male[14].



Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			–
Carol		3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

Fig 2. Examples of missing value in the dataset[15].

- Missing that depends on unobserved predictors: This is a case when the missing values are not random and are related to the unobserved input variable. For example: In a medical study, if a diagnostic study causes discomfort, then there is higher chance of drop out from the study. This missing value is not at random unless we have included "discomfort" as an input variable for all patients[14].
- Missing that depends on the missing value itself: This is a case when the probability of missing value is directly correlated with missing value itself. For example: People with higher or lower income are likely to provide non-response to their earning[14].

4.2.1.2 Treating Missing Values

4.2.1.2.1 Deletion

- In list wise deletion a case is dropped from an analysis because it has a missing value in at least one of the specified variables. The analysis is only run on cases which have a complete set of data. This works well only when we have a large dataset with relatively low proportion of missing data[15].

List wise deletion

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

Pair wise deletion

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

Fig 3. Example of list-wise and pair-wise deletion of missing data[15].

- Pairwise deletion occurs when the statistical procedure uses cases that contain some missing data. The procedure cannot include a variable when it has a missing value, but it can still use the case when analyzing other variables with non-missing values. A case may contain 3 variables: VAR1, VAR2, and VAR3. A case may have a missing value for VAR1, but this does not prevent some statistical procedures from using the same case to analyze variables VAR2 and VAR3. This can be very computationally expensive[13].

- Deletion methods are used when the nature of missing data is "Missing completely at random" else nonrandom missing values can bias the model output[15].

4.2.1.2.2 Mean/ Mode/ Median Imputation

Imputation is a method to fill in the missing values with estimated ones. Summary statistic imputation is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable[16].

4.2.1.2.3 Predictive Modelling

Prediction model is one of the sophisticated methods for handling missing data. Here, we create a predictive model to estimate values that will substitute the missing data by predicting it based on the rest of the variables. In this case, we divide our data set into two sets: One set with no missing values for the variable and another one with missing values. First data set becomes the training data set of the model while the second data set with missing values is test data set and the variable with missing values is treated as target variable. Next, we create a model to predict target variable based on other attributes of the

training data set and populate missing values of test data set. We can use regression, ANOVA, Logistic regression and various modeling technique to perform this[14]. There are 2 drawbacks for this approach:

- The model estimated values are usually more well-behaved than the true values.
- If there are no relationships with attributes in the data set and the attribute with missing values, then the model will not be precise for estimating missing values.

4.2.1.2.4 KNN Imputation

KNN Imputation is an example of using predictive modelling to impute missing values. In this method of imputation, the missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing. The similarity of two attributes is determined using a distance function. It is also known to have certain advantage & disadvantages[13].

4.2.1.2.4.1 Advantages:

- k-nearest neighbor can predict both qualitative & quantitative attributes
- Creation of predictive model for each attribute with missing data is not required
- Attributes with multiple missing values can be easily treated
- Correlation structure of the data is taken into consideration

4.2.1.2.4.2 Disadvantage:

- KNN algorithm is very time-consuming in analyzing large database. It searches through all the dataset looking for the most similar instances.
- Choice of k-value is very critical. Higher value of k would include attributes which are significantly different from what we need whereas lower value of k implies missing out of significant attributes[15].

4.2.2 Outliers

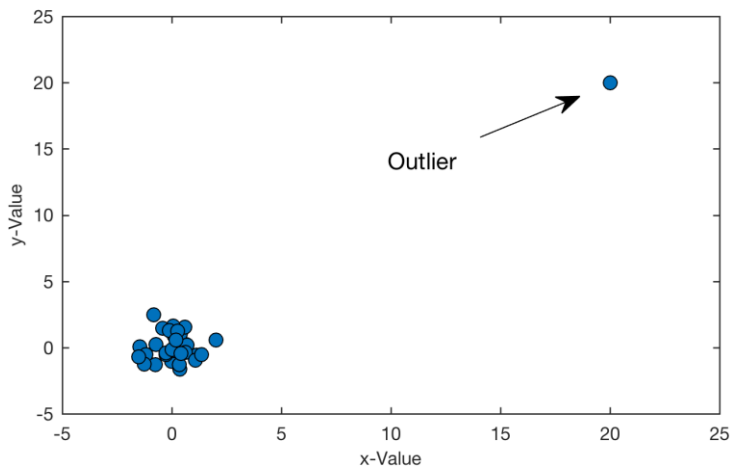


Fig 4. Example of a bivariate outlier[15].

Simply speaking, Outlier is an observation that appears far away and diverges from an overall pattern in a sample. For example, if we do customer profiling and find out that the average annual income of customers is USD 0.3 million. But there are two customers having annual income of USD 6 and USD 4.2 million. These two customers annual income is much higher than rest of the population. These two observations will be Outliers[17].

4.2.2.1 Impact of outliers on data mining

Without Outlier	With Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7,300
Mean = 5.45	Mean = 30.00
Median = 5.00	Median = 5.50
Mode = 5.00	Mode = 5.00
Standard Deviation = 1.04	Standard Deviation = 85.03

Fig 5. Effect of outliers on the summary statistics of a dataset[15].

Outliers can drastically change the results of the data analysis and statistical modeling. There are numerous unfavorable impacts of outliers in the data set:

- They increase the error variance and reduces the power of statistical tests.

- If the outliers are non-randomly distributed, they can decrease normality of the dataset.
- They can bias or influence estimates that may be of substantive interest.
- They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions. [15,17]

4.2.2.2 Outlier Detection

Most commonly used method to detect outliers is visualization. We use various visualization methods, like Boxplot, Histogram, Scatter Plot. There are also various thumb rules to detect outliers. Some of them are:

- Any value, which is beyond the range of -1.5 x IQR to 1.5 x IQR can be considered an outlier.
- Any value which out of range of 5th and 95th percentile can be considered as outlier.
- Data points, three or more standard deviation away from mean are considered outliers.
- Bi variate and multivariate outliers are typically measured using either an index of influence or leverage, or distance. [15,17]

4.2.2.3 Treating Outliers

Most of the ways to deal with outliers are similar to the methods of missing values like deleting observations, transforming them, binning them, treat them as a separate group, imputing values and other statistical methods. Here, we will discuss the common techniques used to deal with outliers:

- Deleting observations: We delete outlier values if it is due to data entry error, data processing error or outlier observations are very small in numbers. We can also use trimming at both ends to remove outliers.

- Transforming and binning values: Transforming variables can also eliminate outliers. Natural log of a value reduces the variation caused by extreme values. Binning is also a form of variable transformation. Decision Tree algorithm allows to deal with outliers well due to binning of variable. We can also use the process of assigning weights to different observations.

- Imputing: Like imputation of missing values, we can also impute outliers. We can use mean, median, mode imputation methods. Before imputing values, we should analyze if it is natural outlier or artificial. If it is artificial, we can go with imputing values. We can also use statistical model to predict values of outlier observation and after that we can impute it with predicted values.

- Treat separately: If there are significant number of outliers, we should treat them separately in the statistical model. One of the approach is to treat both groups as two different groups and build individual model for both groups and then combine the output.[15,17]

4.3 Unsupervised Learning

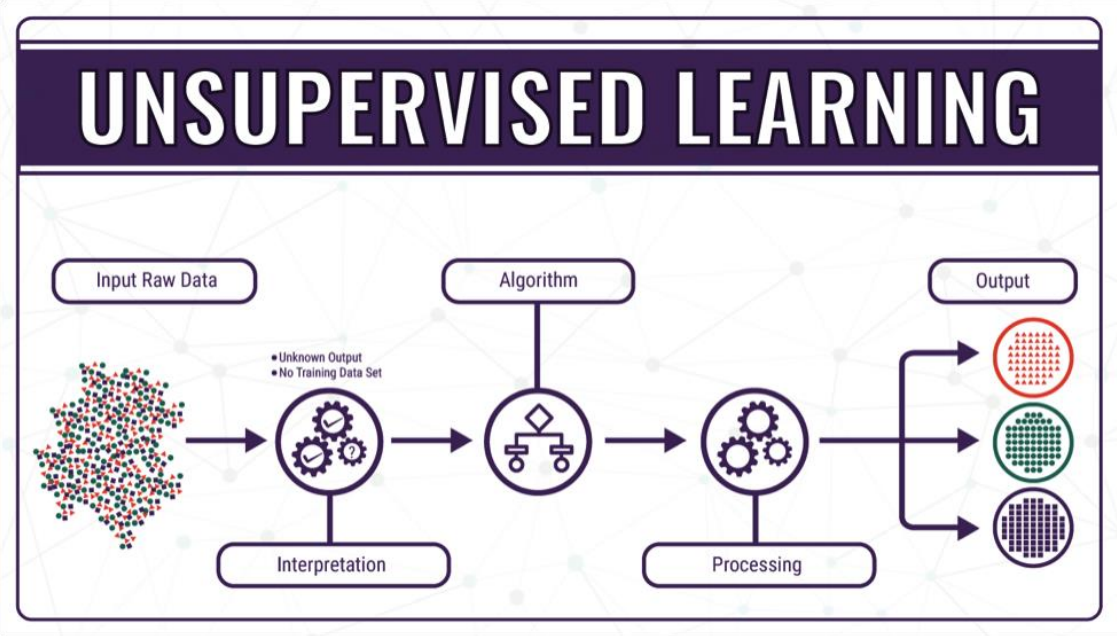


Fig 6. Workflow of unsupervised learning approaches[18].

In this branch of machine learning, the machine is not provided with labelled data. Instead, it is left alone to discover patterns and associations in the data on its own, using mathematical and statistical heuristics. These are then used to categorize the data points thus labeling them[19]. There are two primary types of unsupervised learning algorithms:

- Clustering: Clustering is used to discover inherent categories of data points in the data. The aim to generate clusters of data points that are as similar as possible within a cluster and as different as possible across different clusters.
- Association: Association is used to discover rules that allow the machine to describe large portions of the overall dataset[19].

In this project I primarily use clustering techniques in order to group single airport Island countries into groups of similar countries. This is an example of unsupervised learning as the dataset is not labelled and we are only interested in discovering groups of countries sharing similar patterns in the variables. Below I have described some of the most commonly used clustering algorithms.

4.4 Hierarchical Clustering

Hierarchical clustering is a clustering technique that builds a hierarchy of clusters based on their relative similarities. What we obtain from hierarchical clustering is a set of nested clusters organized as a tree. Each node except the terminal nodes of the tree represents a cluster formed by the union of its child nodes. The child nodes represent sub clusters within

each cluster(*Cluster Analysis: Basic Concepts and Algorithms*, n.d.). Below I describe the algorithm for agglomerative hierarchical clustering:

- The algorithm starts with representing each data point as a unique cluster.
- Proximity of each cluster to all other unique clusters is calculated.
- Based on the similarity of these clusters, the two most similar clusters are merged into one.
- The second and third steps are repeated iteratively until all data points are a part of one big cluster.
- Each time two clusters are merged, the distance between them is recorded. This forms the basis of visualizing the hierarchical clustering as a dendrogram.
  - A dendrogram is a tree-like diagram that records the sequence of merges.
  - Whenever two clusters are merged, they are joined in the dendrogram and the height of the join is determined by the distance between the clusters.
  - More the distance of the vertical lines in the dendrogram, more the distance between those clusters.
- Once the dendrogram is prepared, a distance threshold is chosen to cut the dendrogram into a number of clusters[19].

4.4.1 Defining proximity between clusters

The key step for the hierarchical clustering algorithm is the calculation of proximity of each cluster to all other unique clusters. Before this proximity between clusters comes into play, distance between each data point has to be calculated. This distance can be calculated in a variety of different ways[19,20]. Listed below are two of the more commonly used distance metrics for a & b representing two data points:

- Euclidean distance

$$dist(a,b) = \sqrt{\sum_i (a_i - b_i)^2}$$

- Manhattan Distance

$$dist(a,b) = \sum_i |a_i - b_i|$$

- Once the pairwise distance between all points are known, the two points closest together are merged to form a cluster. When it comes to merging two different clusters together, a linkage criteria has to be used[19,20]. Below I define some of the commonly used linkage criteria for A & B representing two clusters of observations:

- Complete Linkage clustering

$$dist(A,B) = max(d(a.b): a \in A, b \in B)$$

- Single linkage clustering

$$dist(A,B) = min(d(a.b): a \in A, b \in B)$$

- Average linkage clustering

$$dist(A,B) = \frac{1}{|A|.|B|}\sum_{a\in A}\sum_{b\in B} d(a,b)$$

- Wards method
  - Very similar to Average linkage clustering but takes sum of **Squared** distances.

$$dist(A,B) = \frac{1}{|A|.|B|}\sum_{a\in A}\sum_{b\in B} d(a,b)^2$$

- Centroid linkage clustering
  - For c_A and c_B as cluster centroids for cluster A and B respectively.

$$dist(A,B) = || c_A - c_B ||$$

4.4.2 Advantages

- Hierarchical Clustering is inherently very informative as it provides a structured hierarchy of clusters.
- It is completely unsupervised as we do not have provide the number of clusters.
- It is much easier to determine the optimal number of clusters in the data by looking at the dendrogram.[19,20]

4.4.3 Disadvantages

- It is very expensive in terms of time and computational resources for large data sets.

- Initial seeds have a strong impact on the final results.
- The order of the data has an impact on the final results.
  - Many distance metrics use the magnitude of the data point coordinates to calculate the distance.
  - Distances between higher order columns will have a greater impact than distances between lower order columns.
- The algorithm is very sensitive to outliers.
  - Distances between outliers and any other point will be larger than other pairwise distances.[19,20]

4.5 K-Means Clustering

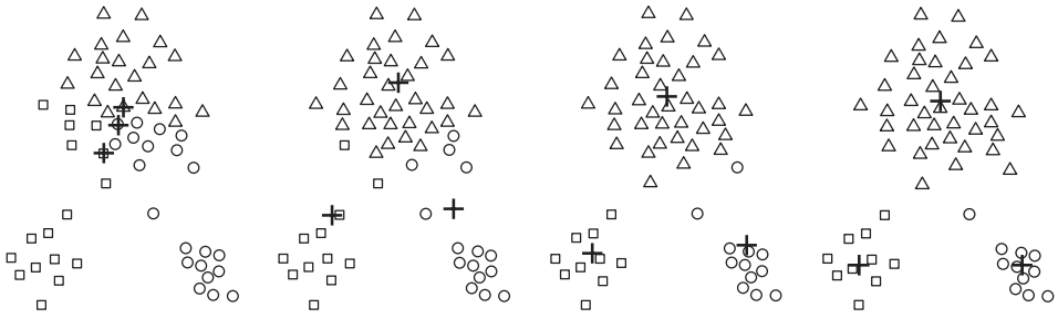
The K-means clustering algorithm groups the data points into a predefined number of non-overlapping clusters such that the sum of the squared distance between the data points and the corresponding cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum[19,20]. The algorithm behind K-means clustering is described below:

- The algorithm is initiated by predefining the number of clusters K to be formed.
- K random data points are arbitrarily chosen without replacement as cluster centroids.
- For each of the remaining data points the distance to each of the K centroids is calculated.

  - The distance metric can be any of the distance metrics defined above.
- Each data point is assigned to one of the K clusters based on their proximity to the cluster centroid.
- Cluster centroids are recalculated based all the data points in a cluster as follows.

$$c_i = \frac{1}{|\ S_i\ |} \sum_{x \in S_i} x_i$$

- The last three steps are iterated until there is no change of centroid.[19,20]



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

Fig 7. Example results of the iterative K-means clustering algorithm[20].

4.5.1 Objective Function

For the objective function, K-means uses the **sum of the squared error (SSE)**, which means the sum of the distance of each datapoint to the closest cluster centroid.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} d\ ist(c_i, x)^2$$

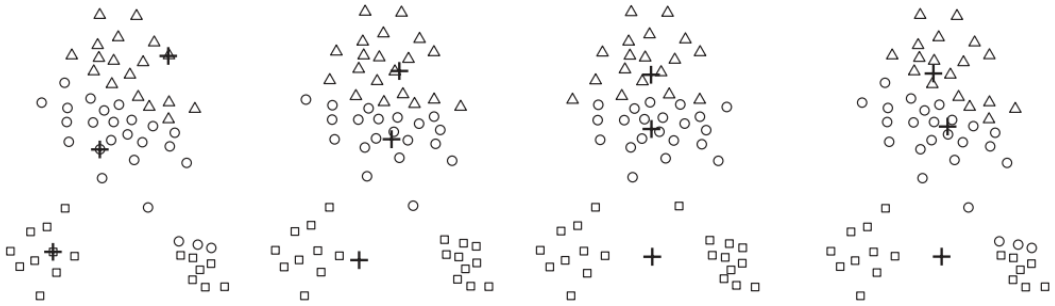
Here the distance can be measured using a variety of different distance metrics such as Euclidean, Manhattan, or cosine distances[20].

4.5.2 Advantages

- The K-means algorithm is very intuitive and works well in most cases, but it does have some limitations:
- With many variables, K-Means may be computationally faster than hierarchical clustering as distances only need to be calculated against the cluster centroids.
- K-Means may produce tighter clusters than hierarchical clustering.
- An instance can change cluster (move to another cluster) when the centroids are recomputed.[19,20]

4.5.3 Disadvantages

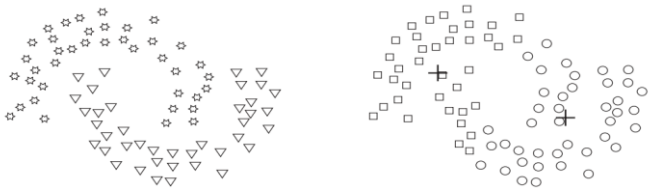
- The final result of the K-means clustering is heavily dependent on the initial arbitrary choice of cluster centroids. This means that clustering results in two different experiments may differ significantly.



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

Fig 8. Example results with sub-optimal initial centroid choice for K-means Clustering[20].

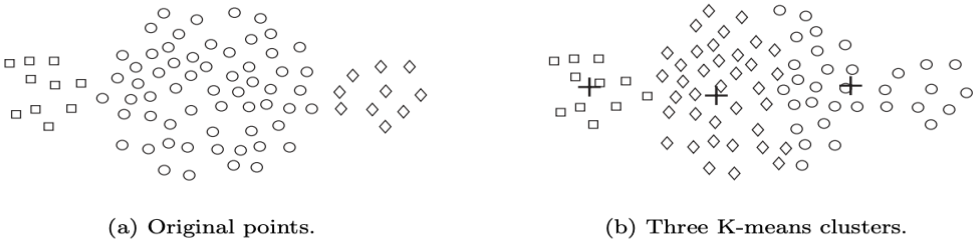
- K-means algorithm tries to fit into spherical clusters. If the dataset has clusters with a complicated geometry, K-means clustering may lead to inaccurate results.



(a) Original points. (b) Two K-means clusters.

Fig 9. Example results with non-spherical clusters for K-means Clustering. The algorithm still tries to fit to spherical-shaped clusters[20].

- It requires us to pre-define the number of clusters meaning that it is semi-supervised.
- Since it tries to reduce the intra-cluster variation, it favors bigger clusters against smaller ones.[19,20]



(a) Original points. (b) Three K-means clusters.

Fig 10. Example results with uneven sized clusters for K-means Clustering. The algorithm favors 2 big and 1 small clusters as compared to 1 big and 2 small ones[20].

4.6 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique used to reduce the dimensions of large data sets or of data sets with multiple highly correlated dimensions. The aim is to transform the dataset into a lower dimension one without losing too much information. Smaller data sets are much easier to explore and analyze which makes it easy to barter lost information[21]. PCA is achieved using the following steps:

- Standardization: PCA is very sensitive to the variances of initial variables. Variables with larger ranges will dominate over those with small ranges. Thus, it becomes imperative to standardize the dataset variables as that gets all the variables on the same scale. Mathematically a variable X with a given mean U and standard deviation s.d is standardized as follows.

$$z = \frac{Xi - U}{s.d}$$

The distribution of z has a mean approximately equal to 0 and a standard deviation approximately equal to 1. This ensures comparable scales for all dataset variables.

- Co variance Matrix Calculation: This step is aimed at understanding the relationships between the variables of the dataset. It does so by understanding how the variables of the input data set are varying from the mean with respect to each other. This helps in identifying correlated variable which are therefore containing redundant information. The co variance matrix is a p × p symmetric matrix (where p is the number of dimensions) that has as entries the co variances associated with all possible pairs of the initial variables. In the main diagonal (Top left to bottom right) we have the variances of each initial variable and the entries of the co variance matrix are symmetric with respect to the main diagonal, which means that the upper and the lower triangular portions are equal. The key piece of information here is actually the sign of the co variance between two variables. A positive co variance indicates that the two variables increase or decrease together (directly correlated) while a negative co variance indicate inverse correlation.

- Compute the eigenvectors and eigenvalues of the co variance matrix to identify the principal components: Every eigenvector has an eigenvalue. And their number is equal to the number of dimensions of the data. For example, for a 3-dimensional data set, there are 3 variables, therefore there are 3 eigenvectors with 3 corresponding eigenvalues. The eigenvectors of the Co variance matrix are actually the directions of the axes where there is the most variance (most information) and that we call Principal Components. And eigenvalues are simply the coefficients attached to eigenvectors, which give the amount of variance carried in each Principal Component. Principal components are the new vectors that form the feature set of our reduced dataset post-PCA. These are constructed as linear combinations of the initial variables in an orthogonal manner. Most of the information contained in the original variables is forced or compressed into the first principal component then the maximum remaining information in the second and so on. By ranking your eigenvectors in order of their

eigenvalues, highest to lowest, you get the principal components in order of significance. This results in a scree plot as shown below. In this plot the percentage of variance explained by each principal component is plotted against the number of components. This plot always displays a downward curve and the 'elbow' of the plot where the Y-values seem to level off represent the optimal number of components.

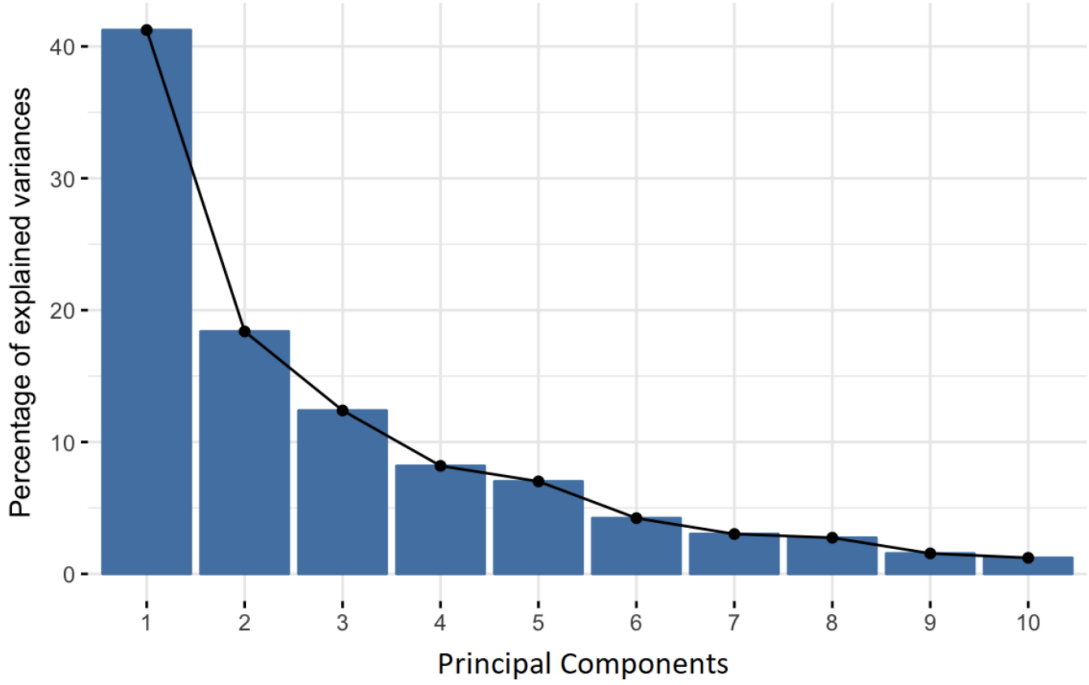


Fig 11. Example scree plot obtained as number of principal components increase[22].

$$FinalDataset = FeatureVector^T * StandardizedOriginalDataset^T$$

- Forming the Feature Vector: Choose whether to keep all these components or discard those of lesser significance (of low eigenvalues), and form with the remaining ones a matrix of vectors that we call Feature vector.
- Recasting the original data into PCA axes: Use the feature vector formed using the eigenvectors of the co variance matrix, to reorient the data from the original axes to the ones represented by the principal components (hence the name Principal Components Analysis). This can be done by multiplying the transpose of the original data set by the transpose of the feature vector.[21]

5 Exploratory data analysis

5.1 Missing value analysis

One of the first things I observed with this dataset was that it was plagued with missing data. A reason for this could that certain countries didn't start capturing certain variables until recently. I have calculated the missing proportion in each variable and plotted the distribution below.

I observed that a majority of variables have more than 50% of the values missing. However, the missing pattern is not completely random. On visual inspection I was able to see that the data for many variables is systematically missing for certain country & year combinations. There are some variables where the missing proportion is higher than 75%. For many of these variables, I noticed that the missing data was predominantly found for years 2006 and below. In order to analyze the effect of the year and the country on the missing proportion of various variables I visualized these features using a heat map. A couple of examples of this analysis have been shown below.

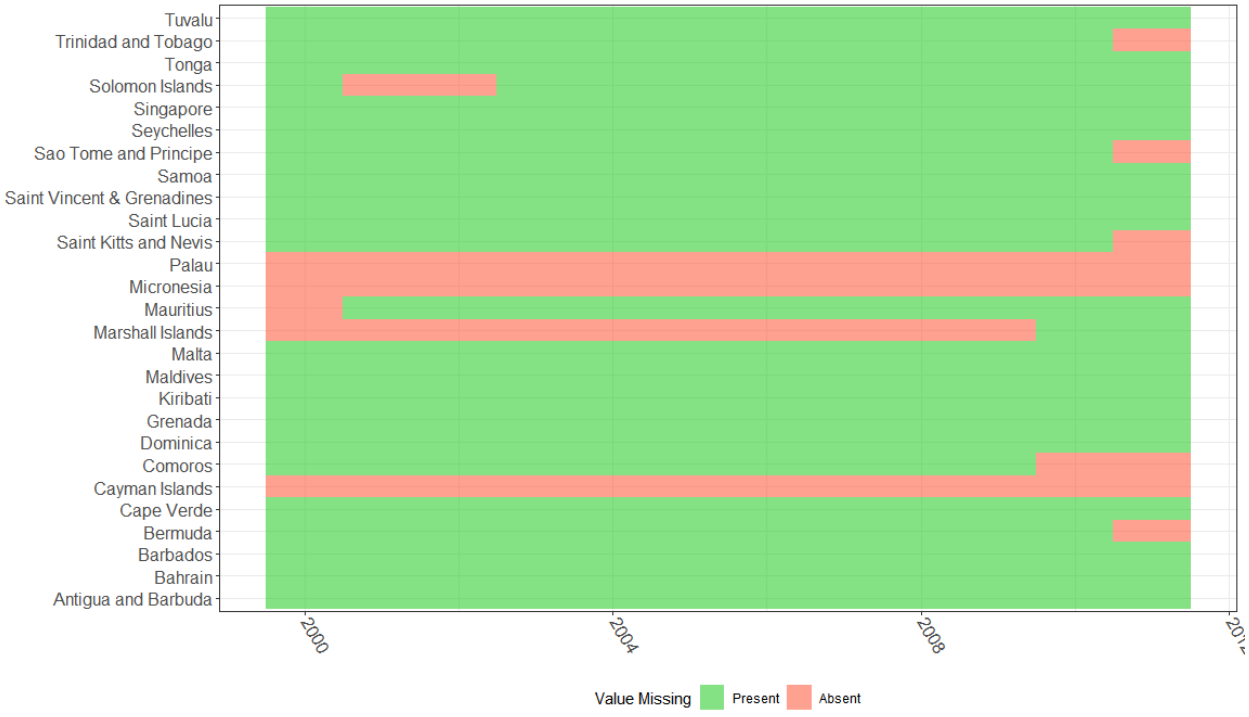


Fig 12. Present/absent values by year and country for the variable 'gdpnom'.

As we see here, the countries 'Tuvalu', 'Cayman Islands', 'Micronesia', 'Palau', and 'Marshall Islands' have more than 80% of the data missing for either 'gdpnom' or 'receipt'. Both variables are critical for this analysis and imputing them in presence of such a high proportion of missing values would be very inaccurate. For the further analysis these countries will be dropped.

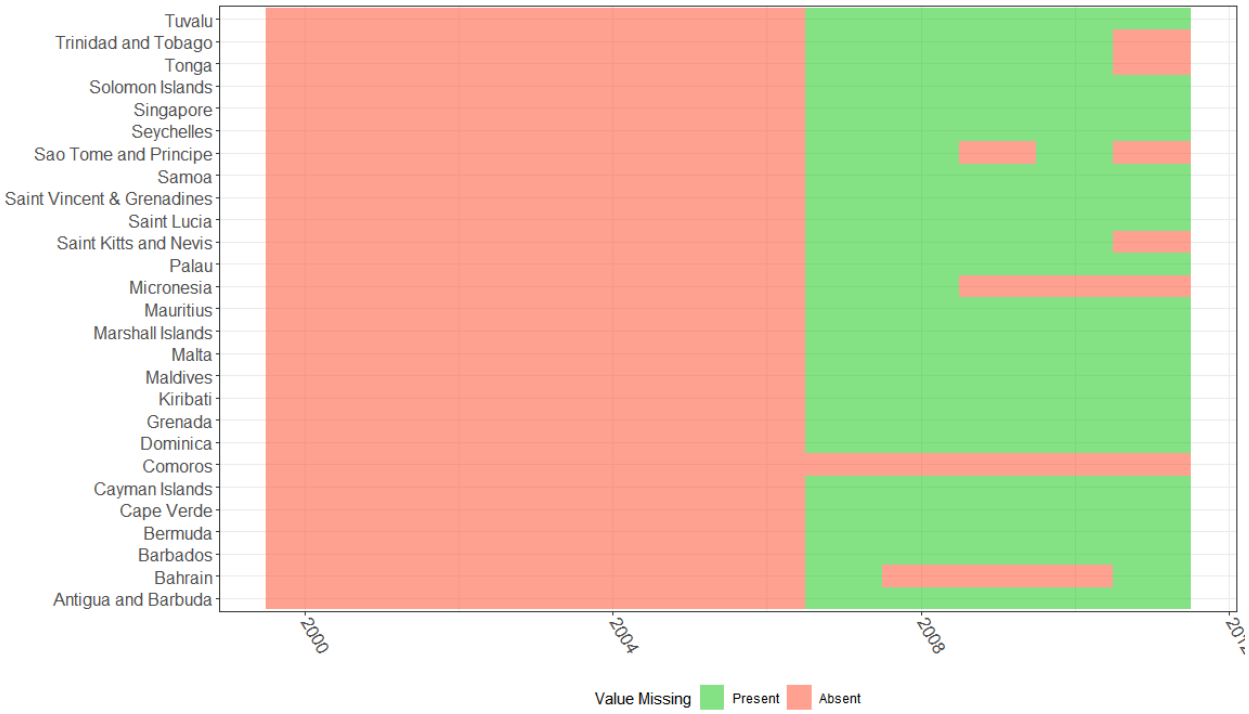


Fig 13. Present/absent values by year and country for the variable 'arram'.

As seen above for the variable 'arram' which means arrivals of tourists from America, the data is systematically missing for all countries before 2007. Such heat maps were created for every variable in the dataset to analyze the patterns of missing data. What I found was that for certain variables the analysis could be done for all years by removing data for certain countries. On the other hand, for most variables what needed to be done was to remove all the data until the year 2007 for all countries to analyze them properly. Thus, subsets of the dataset were created for various clustering experiments I have performed to group the countries together.

Similarly, another key finding on visual inspection was that certain variables were often missing together. All of this meant that I had to decide on a set of core variables which were present throughout the dataset for all years and countries for a primary clustering experiment. For subsequent clustering experiments I kept adding sets of variables that shared the same missing proportion pattern and filtering the countries where the shared missing proportion pattern was found.

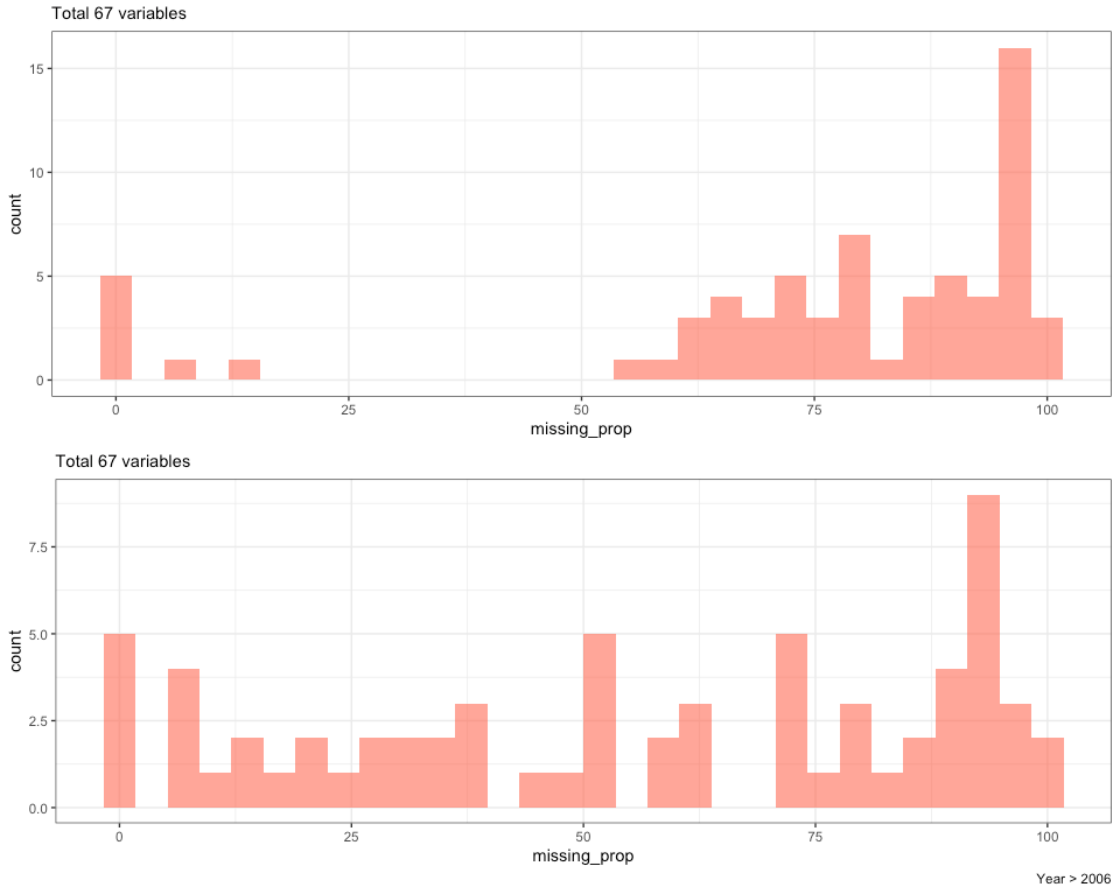


Fig 14. Distribution of missing values in the complete dataset (top) and the dataset after the year 2006 (bottom).

5.2 Similarities in Missing values

In order to figure out which variables were found missing together to generate complete subsets of the data, I performed a hierarchical clustering of the missing data distributions of the variables. This was only performed on the data beyond 2006 as most of the data was missing for most of the variables up until 2006 and thus couldn't be imputed reliably. After doing this all variables with more than 75% data missing were removed from the dataset along with any rows that had more than 80% of the variables missing. This removed 23 variables out of the 63 total variables. This dataset now comprised of 21 unique countries.

Shown below is the dendrogram for the hierarchical clustering of the missing data patterns for each variable in the dataset. The way this plot can be interpreted is the lower the height of the join between two variables the more similar their missing data pattern. This means for example that the variables 'arram' and 'arreur' (right most on the x axis) share very similar missing data pattern.

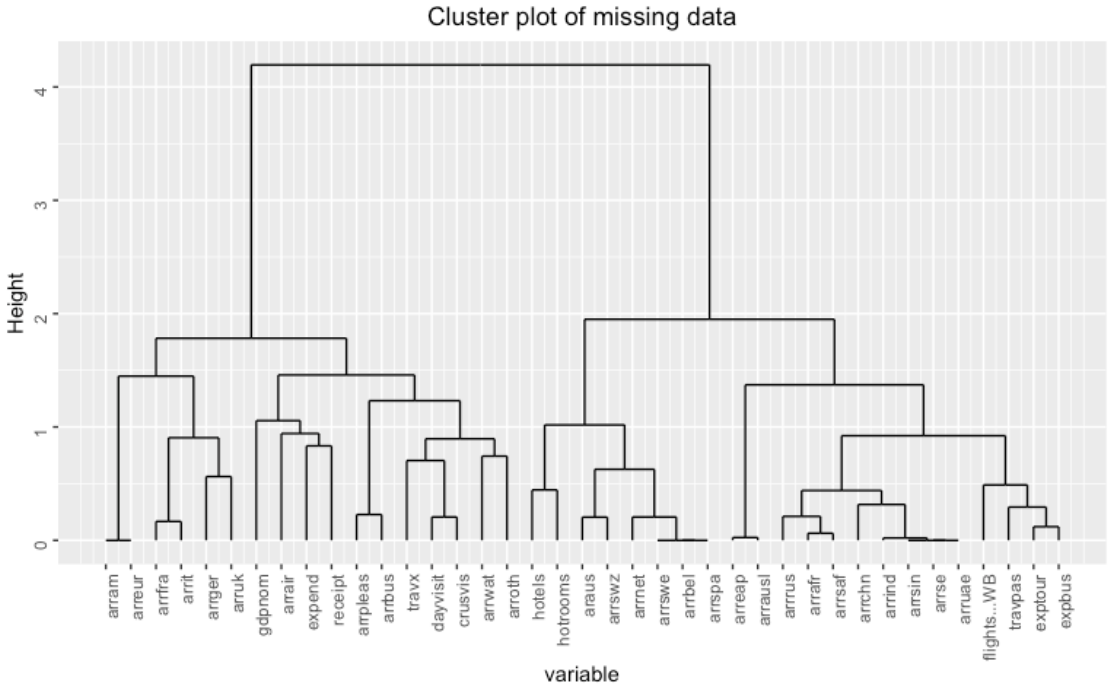


Fig 15. Dendrogram representing the clustered missing data distributions of various variables in the dataset.

Based on this clustering the variable sets I perform clustering experiments on are as follows:

- ‘arram’ and ‘arreur’
- ‘arrswe’, ‘arrbel’, ‘arrspa’
- ‘arrfra’ and ‘arrit’

6 Clustering experiments

For clustering experiments, I have used the Hybrid K-means clustering algorithm in which the initial cluster centers for K-means clustering are calculated using hierarchical clustering.

6.1 Clustering Experiment 1

The primary clustering experiment was done with a set of core variables that were present most of the times in the dataset. These variables represented the basic information associated with the island nations. The following variables were used for this analysis as they were both core to this project and had low missing data proportions.

```
## [1] "country"  "year"     "pop"      "areakm2"  "gdpnom"   "receipt"
## [7] "ovnarriv"
```

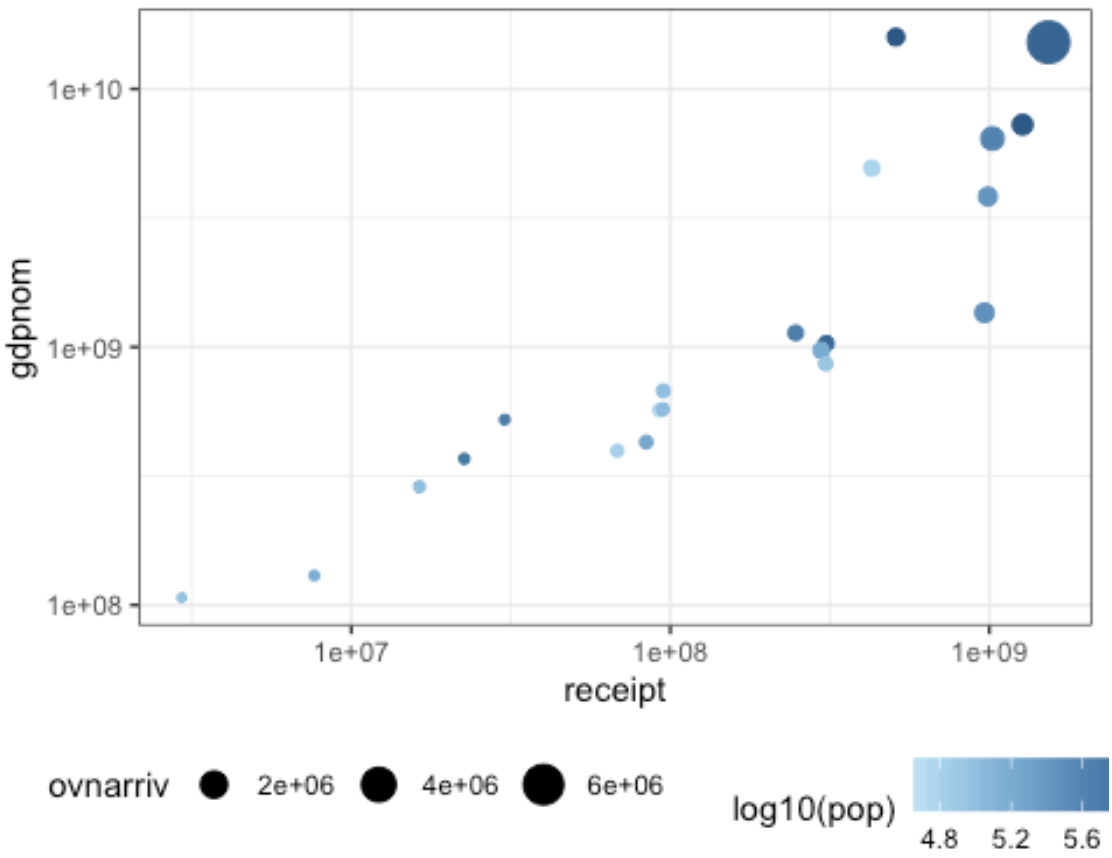


Fig 16. Scatter plot showing the relationships between the core variables.

There were some missing values remaining in the variables 'pop', 'gdpnom' and 'receipt'. These values were imputed using K-nearest neighbors. The hyper parameter 'K' for the KNN algorithm was chosen as 2 empirically.

The dataset was then grouped by country and all variables were summarized as mean. When I looked at the distribution of the variables in the dataset, I found that all 4 variables had a highly skewed distribution with most of values lying on the lower intervals with a few large values skewing the distribution. To deal with this skewness I performed a log10 transformation on the entire dataset. To bring all variables to the same scale I standardized the numerical variables. Following this, two separate methods were used to calculate the optimal number of clusters. The elbow plot suggested the number of optimal clusters to be 2 while the silhouette plot suggested the same. After experimenting with the K value and inspecting the clustering results I decided to cluster this dataset into 2,3 & 4 clusters using the hybrid K-means clustering technique.

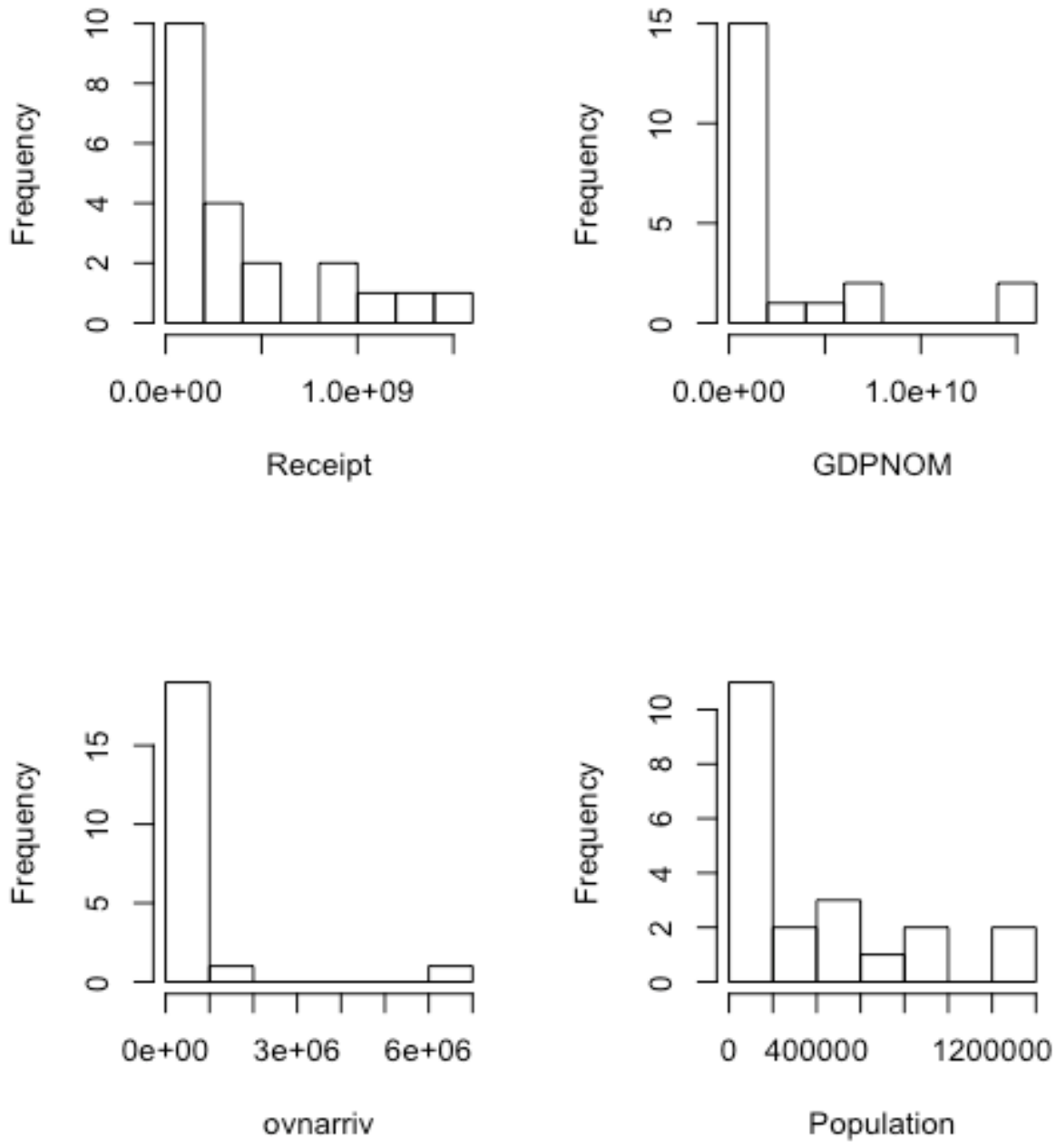


Fig 17. Distribution of the core variables before any transformation.

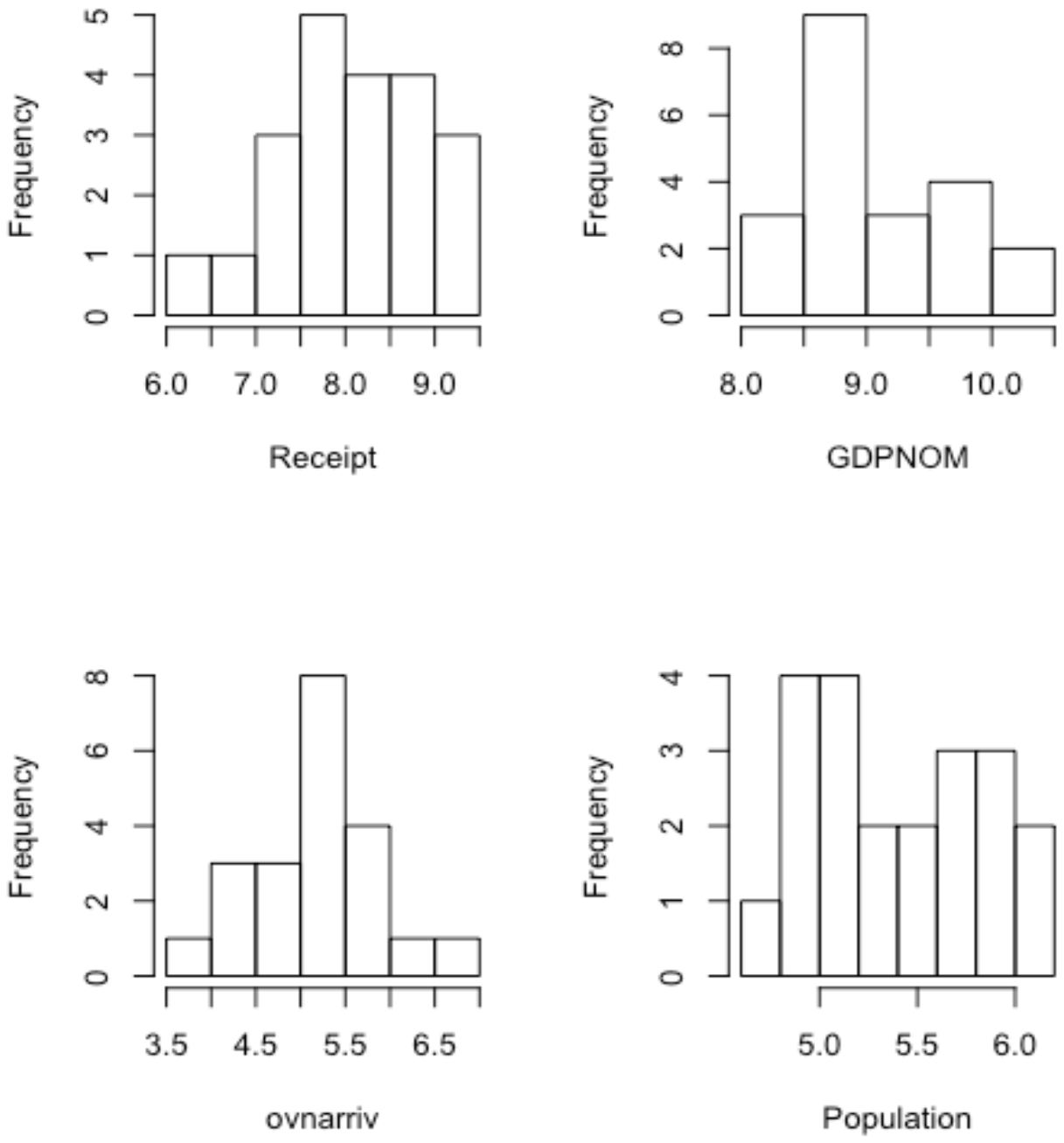


Fig 18. Distribution of the core variables after log10 transformation.

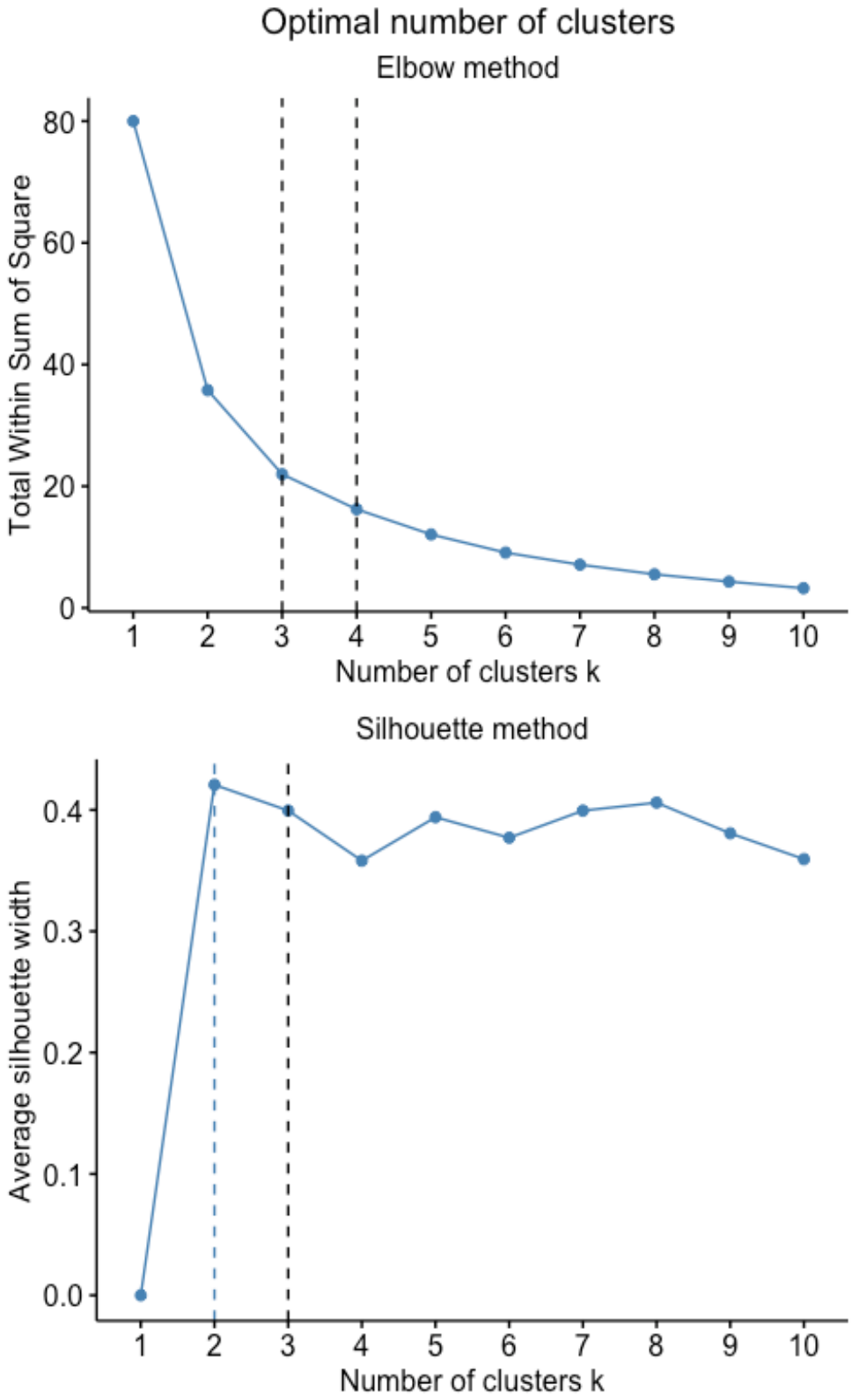


Fig 19. Determination of the optimal number of clusters using the elbow (top) and the silhouette (bottom) methods for clustering experiment 1.

6.1.1 2 Clusters

Table 1. *Cluster profiles for clustering experiment 1 with 2 clusters*

cluster	receipt	gdpnom	ovnarriv	pop
1	7.57e+08	5.80e+09	1.14e+06	5.94e+05
2	7.46e+07	4.48e+08	6.60e+04	1.88e+05

As seen above the clustering experiment seems to have separated the dataset into high value and low value clusters. The complete dataset with the cluster memberships can be seen below.

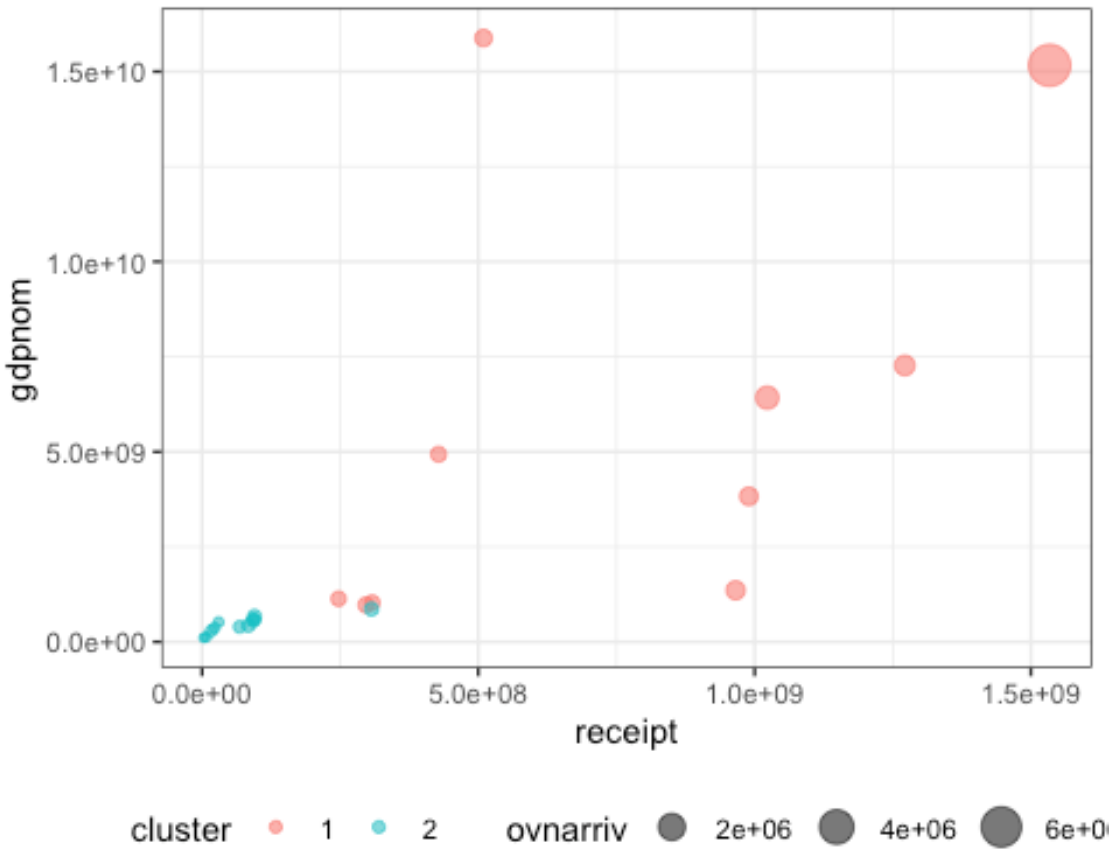


Fig 20. Relationship between the core variables and the clusters for clustering experiment 1 with 2 clusters

Table 2. *Cluster memberships obtained from clustering experiment 1 with 2 clusters*

cluster	countries
1	Antigua and Barbuda, Bahrain, Barbados, Bermuda, Cape Verde, Maldives, Malta, Mauritius, Saint Lucia, Trinidad and Tobago
2	Comoros, Dominica, Grenada, Kiribati, Saint Kitts and Nevis, Saint Vincent & Grenadines, Samoa, Sao Tome and Principe, Seychelles, Solomon Islands, Tonga

What we see is that countries with high GDP such as 'Bahrain', 'Trinidad and Tobago', 'Mauritius' and 'Malta' lie in the first cluster. In general, the first cluster of countries has higher revenue from tourism, higher number of tourist arrivals and higher population than the second cluster. An exception to this seems to be Seychelles which also has a high GDP and income from tourists but has relatively tourist arrival and population.

To visualize the clusters, I used the function 'fviz_cluster' from the factoextra package in R. This function uses PCA to reduce the dimensions of the dataset to 2 and then plots a scatter plot in those 2 dimensions for each point. Below we can clearly see how the clusters are separated. Doing PCA often leads to information loss and as such this visual is only for the purpose of visualization.

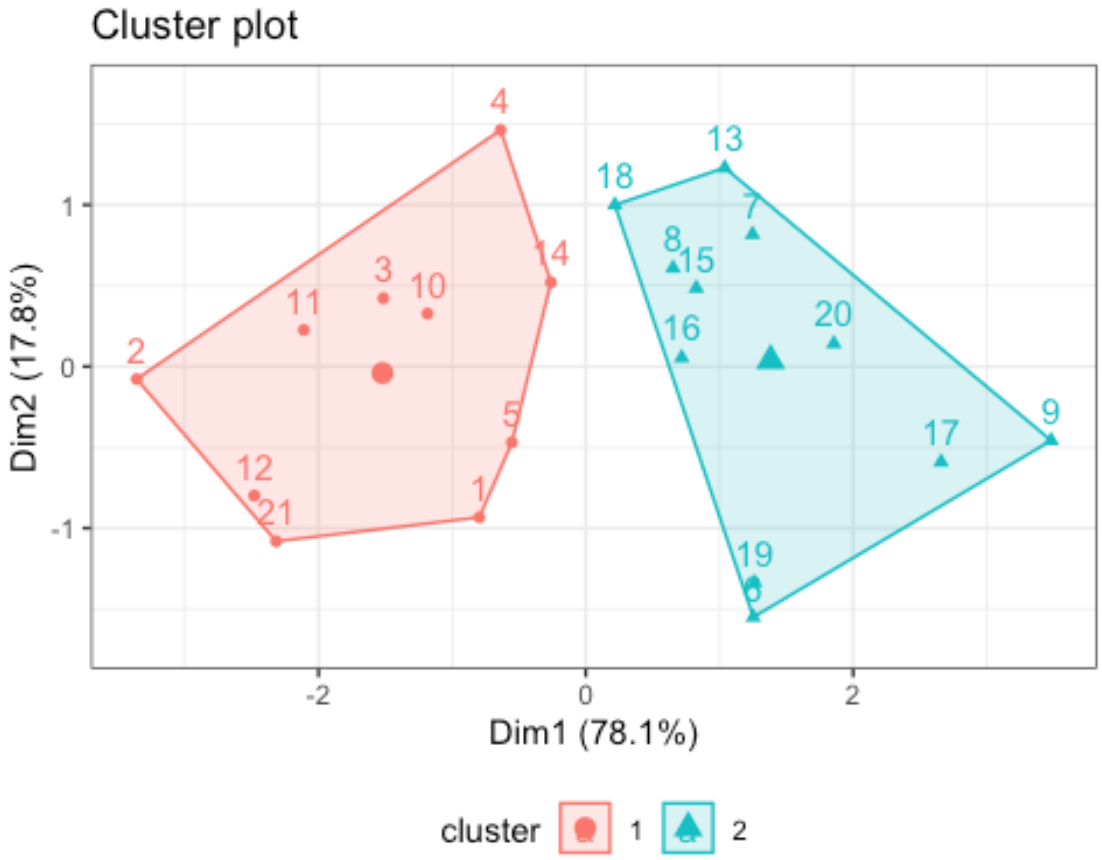


Fig 21. Cluster separation achieved on two principal components of the dataset in clustering experiment 1 with 2 clusters.

6.1.2 3 Clusters

Table 3. *Cluster profiles for clustering experiment 1 with 3 clusters*

cluster	pop	receipt	gdpnom	ovnarriv
1	7.14e+05	8.56e+08	6.51e+09	1.35e+06
2	1.03e+05	1.83e+08	1.18e+09	1.49e+05
3	2.94e+05	1.60e+07	2.83e+08	1.81e+04

As seen above the clustering experiment seems to have separated the dataset into high, moderate and low value clusters. The complete dataset with the cluster memberships can be seen below. Here, cluster 3 clearly separates out countries where the tourism sector is not performing as well as the other two clusters as both the variables 'ovnarriv', and 'receipt' have the lowest values across all 3 clusters. Cluster 1 on the other hand is clearly dominating in terms of tourism and economic health. Cluster 2 boasts moderate values across the board with an exception of Bermuda which is a curious case. Bermuda performs well in GDP and tourism terms, even outperforming some of the cluster 1 countries. Bermuda does this despite having an order of magnitude less population.

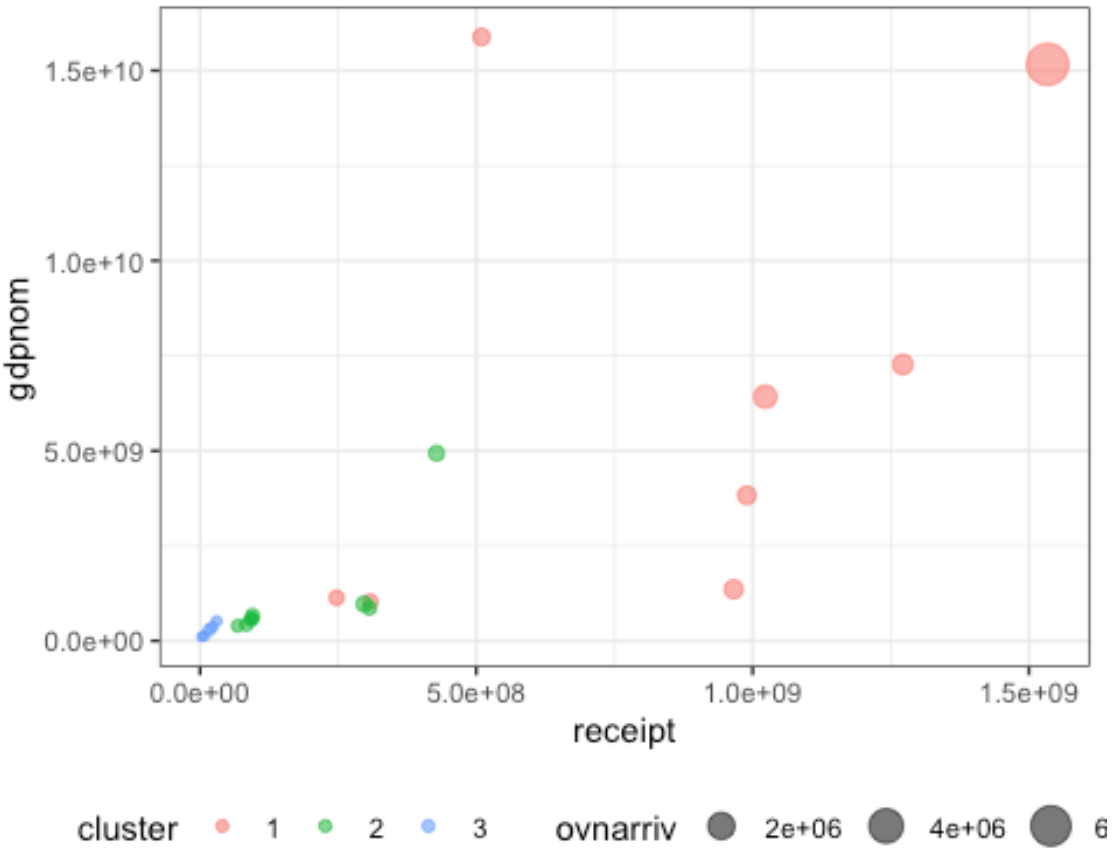


Fig 22. Relationship between the core variables and the clusters for clustering experiment 1 with 3 clusters

Table 4. *Cluster memberships obtained from clustering experiment 1 with 3 clusters*

cluster	countries
1	Antigua and Barbuda, Bahrain, Barbados, Cape Verde, Maldives, Malta, Mauritius, Trinidad and Tobago
2	Bermuda, Dominica, Grenada, Saint Kitts and Nevis, Saint Lucia, Saint Vincent & Grenadines, Samoa, Seychelles
3	Comoros, Kiribati, Sao Tome and Principe, Solomon Islands, Tonga

To visualize the clusters, I used the function 'fviz_cluster' from the factoextra package in R. This function uses PCA to reduce the dimensions of the dataset to 2 and then plots a scatter plot in those 2 dimensions for each point. Below we can clearly see how the clusters are separated. Doing PCA often leads to information loss and as such this visual is only for the purpose of visualization.

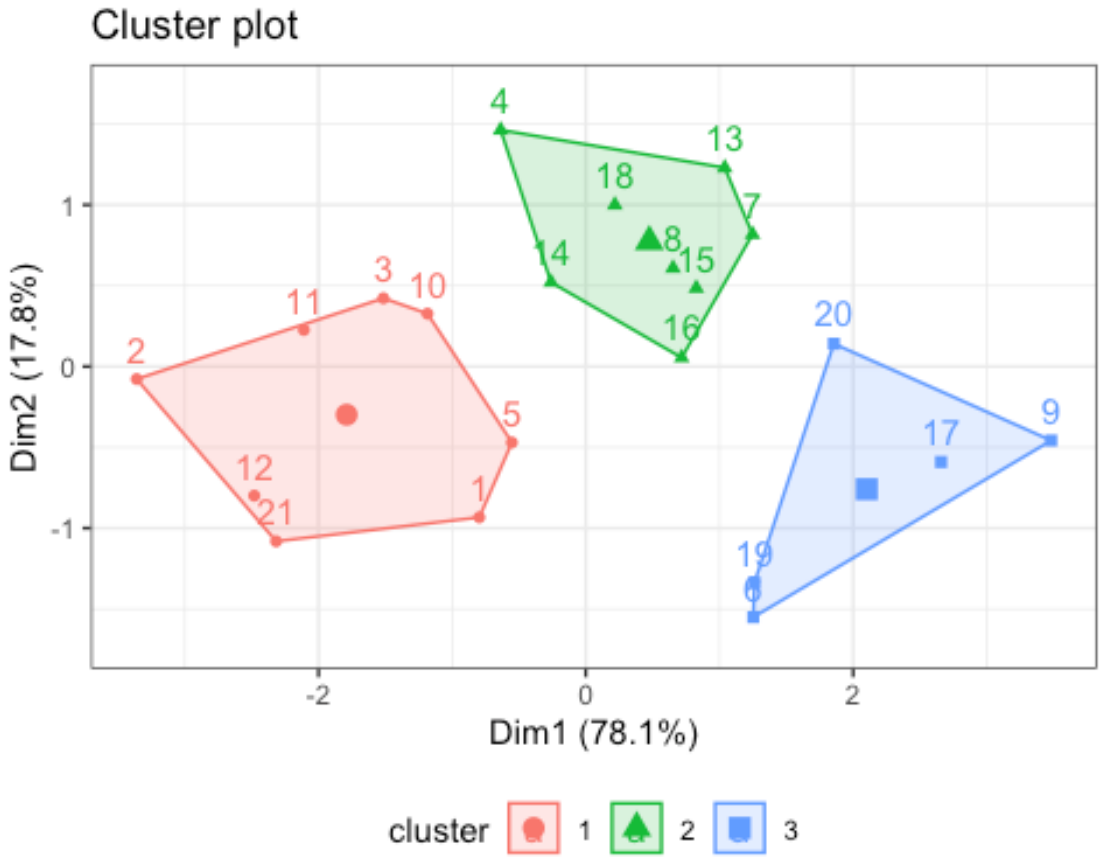


Fig 23. Cluster separation achieved on two principal components of the dataset in clustering experiment 1 with 3 clusters.

6.1.3 4 Clusters

Table 5. *Cluster profiles obtained from clustering experiment 1 with 4 clusters*

cluster	ovnarriv	pop	receipt	gdpnom
1	4.05e+05	4.69e+05	6.28e+08	1.84e+09
2	2.30e+06	9.59e+05	1.08e+09	1.12e+10
3	1.49e+05	1.03e+05	1.83e+08	1.18e+09
4	1.81e+04	2.94e+05	1.60e+07	2.83e+08

As seen above the clustering experiment seems to have separated the dataset into high, moderate and low value clusters. The complete dataset with the cluster memberships can be seen below.

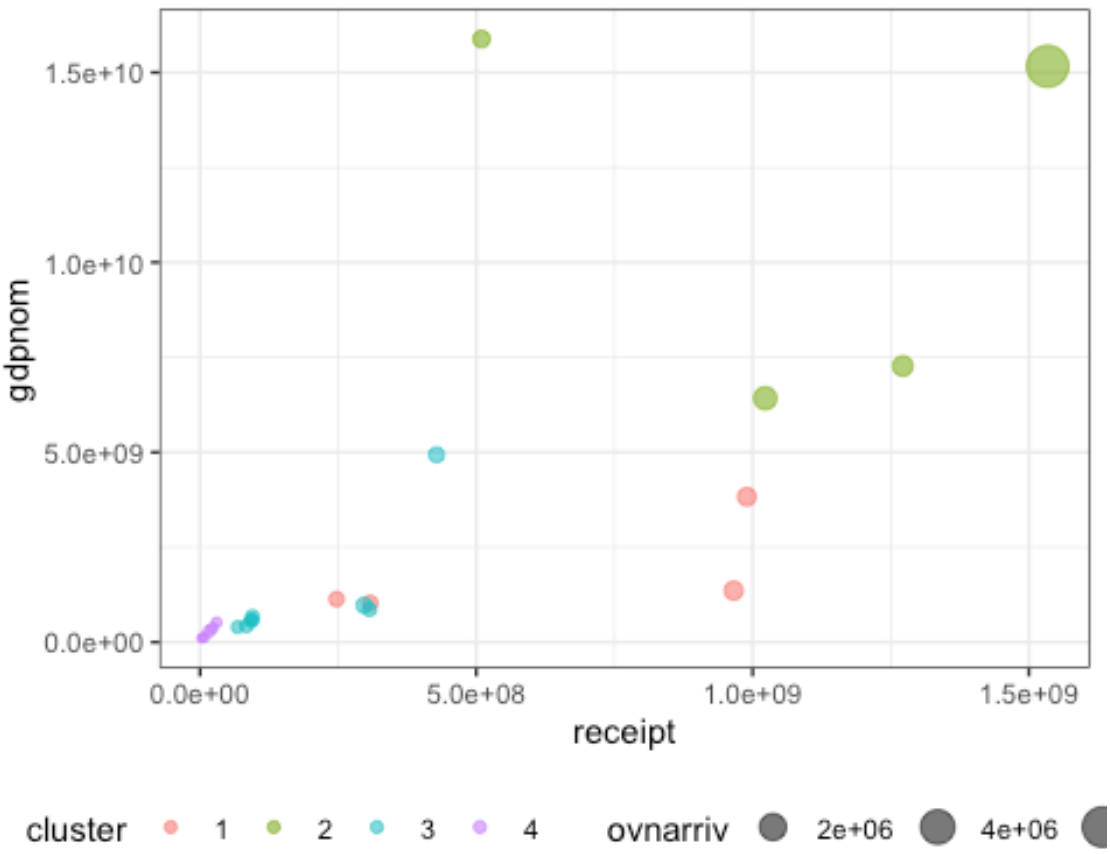


Fig 24. Relationship between the core variables and the clusters for clustering experiment 1 with 4 clusters

Table 6. *Cluster memberships obtained from clustering experiment 1 with 4 clusters*

cluster	countries
1	Antigua and Barbuda, Barbados, Cape Verde, Maldives
2	Bahrain, Malta, Mauritius, Trinidad and Tobago
3	Bermuda, Dominica, Grenada, Saint Kitts and Nevis, Saint Lucia, Saint Vincent & Grenadines, Samoa, Seychelles
4	Comoros, Kiribati, Sao Tome and Principe, Solomon Islands, Tonga

As before, two clusters clearly contain the countries doing the best and the worst in terms of tourism and economy namely cluster 2 and 4 respectively. Cluster 1 contains the moderately performing countries while cluster 3 contains low-to-moderate performing countries with an exception of Bermuda.

To visualize the clusters, I used the function 'fviz_cluster' from the factoextra package in R. This function uses PCA to reduce the dimensions of the dataset to 2 and then plots a scatter plot in those 2 dimensions for each point. Below we can clearly see how the clusters are

separated. Doing PCA often leads to information loss and as such this visual is only for the purpose of visualization.

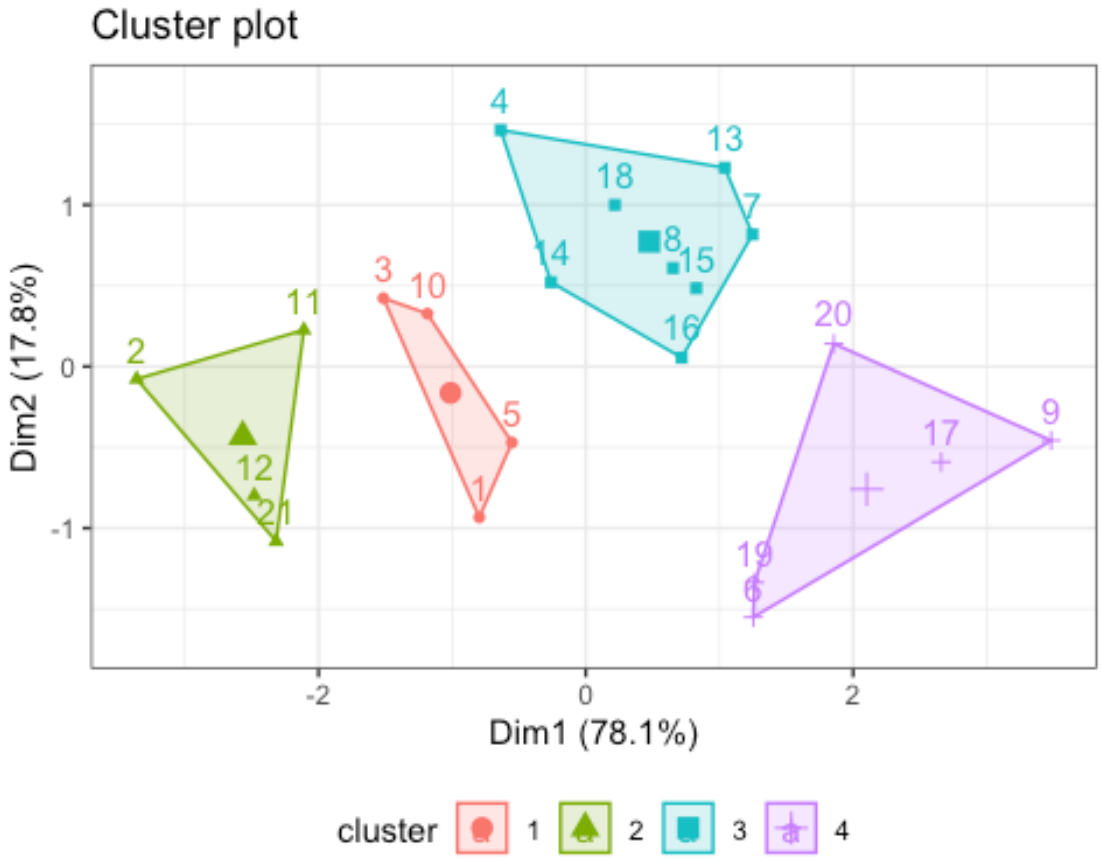


Fig 25. Cluster separation achieved on two principal components of the dataset in clustering experiment 1 with 4 clusters.

6.2 Clustering Experiment 2

In this experiment I used the variables 'arram' and 'arreur' for the clustering analysis. These variables represent tourist arrivals from America and Europe respectively.

On analysis I noticed that these variables are jointly missing for Bahrain, and Comoros for 3 years and for Tonga and Sao Tome and Principe for 1 year. Because of this I removed those observations from the data. The distributions of these two variables were skewed and I corrected that using log transformation. Following this I clustered the dataset into 3 clusters (based on elbow and silhouette method along with experimentation).

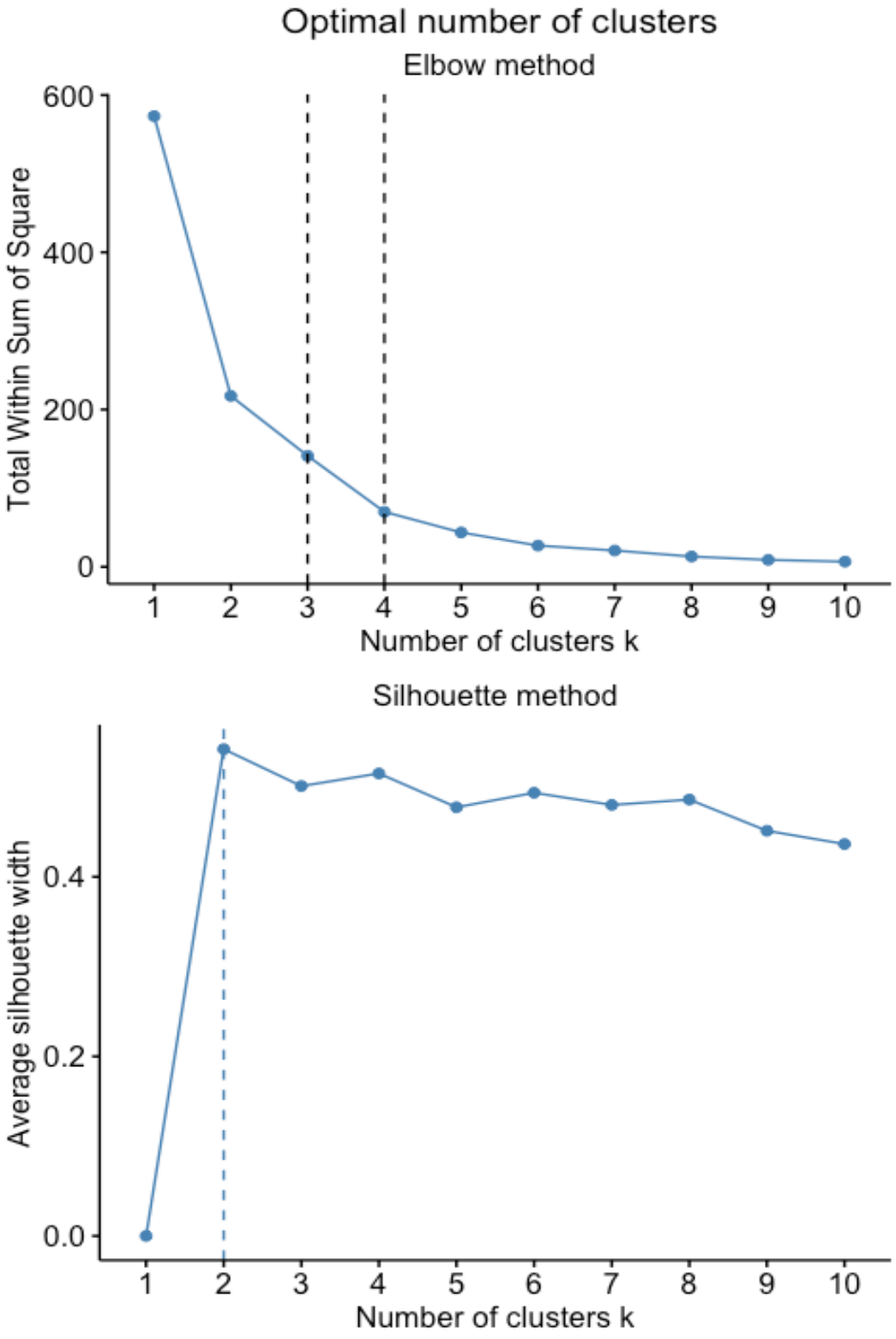


Fig 26. Determination of the optimal number of clusters using the elbow (top) and the silhouette (bottom) methods for clustering experiment 2.

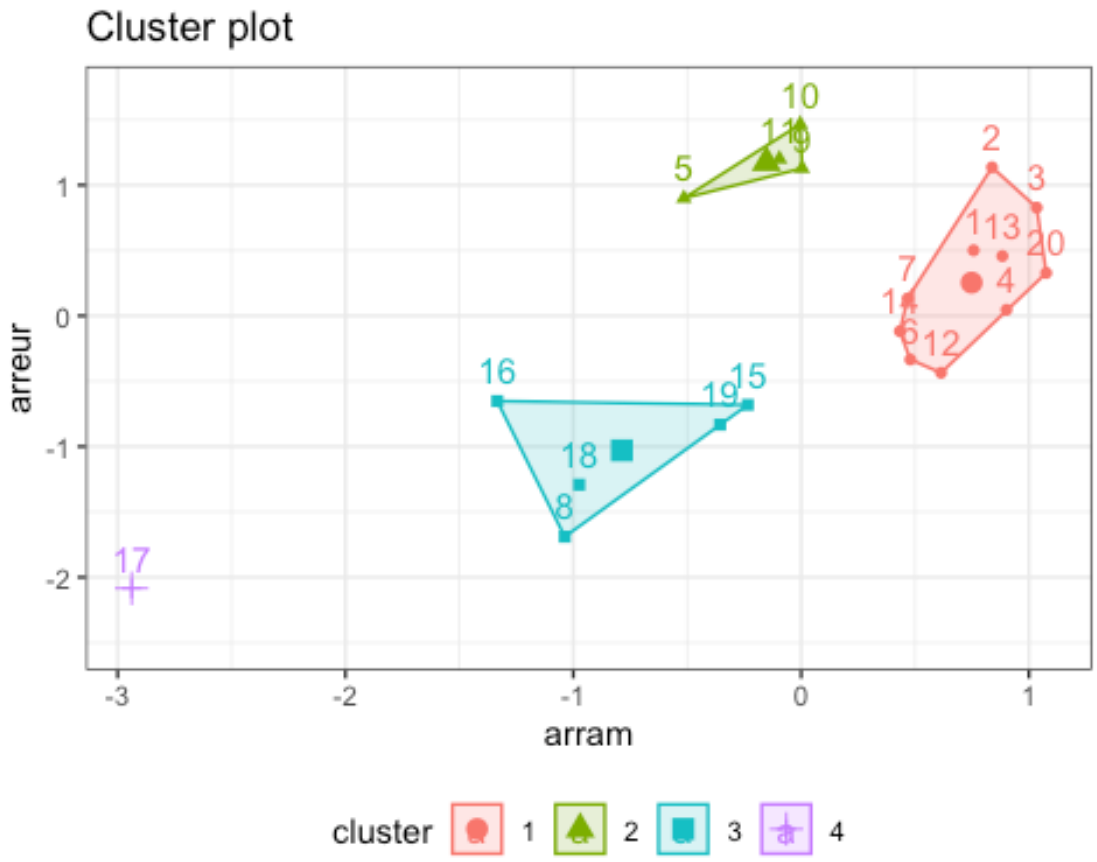


Fig 27. Cluster separation achieved on the two variables of the dataset in clustering experiment 2.

Table 7. *Cluster profiles obtained from clustering experiment 2*

cluster	pop	gdpnom	arram	arreur	receipt	ovnarriv
1	4.16e+05	5.99e+09	1.68e+05	1.10e+05	5.03e+08	9.44e+05
2	6.23e+05	5.33e+09	1.28e+04	6.42e+05	1.22e+09	8.21e+05
3	2.14e+05	3.77e+08	3.49e+03	2.92e+03	3.92e+07	4.04e+04
4	8.72e+04	9.75e+08	4.40e+00	1.31e+02	3.77e+08	1.69e+05

Table 8. *Cluster memberships obtained from clustering experiment 2*

cluster	countries
1	Antigua and Barbuda, Bahrain, Barbados, Bermuda, Dominica, Grenada, Saint Kitts and Nevis, Saint Lucia, Saint Vincent & Grenadines, Trinidad and Tobago
2	Cape Verde, Maldives, Malta, Mauritius
3	Kiribati, Samoa, Sao Tome and Principe, Solomon Islands, Tonga
4	Seychelles

Above we can clearly see the distinction between countries where tourists from America and Europe pay lots of visits and where they do not. Cluster 1 consists of countries with high tourist volume from the American tourists. Cluster 2 on the other hand consists of countries with high tourist volumes from Europe. Cluster 3 belongs to countries who have low to very low tourist volume from these regions. Lastly cluster 4 consists of only a single country 'Seychelles' which has extremely low number of tourists from the American and European region. This country is also off the coast of Africa and is relatively less known. Next I compared these two clusters across the core variables to see what effect of increased tourism from these two regions has on those.

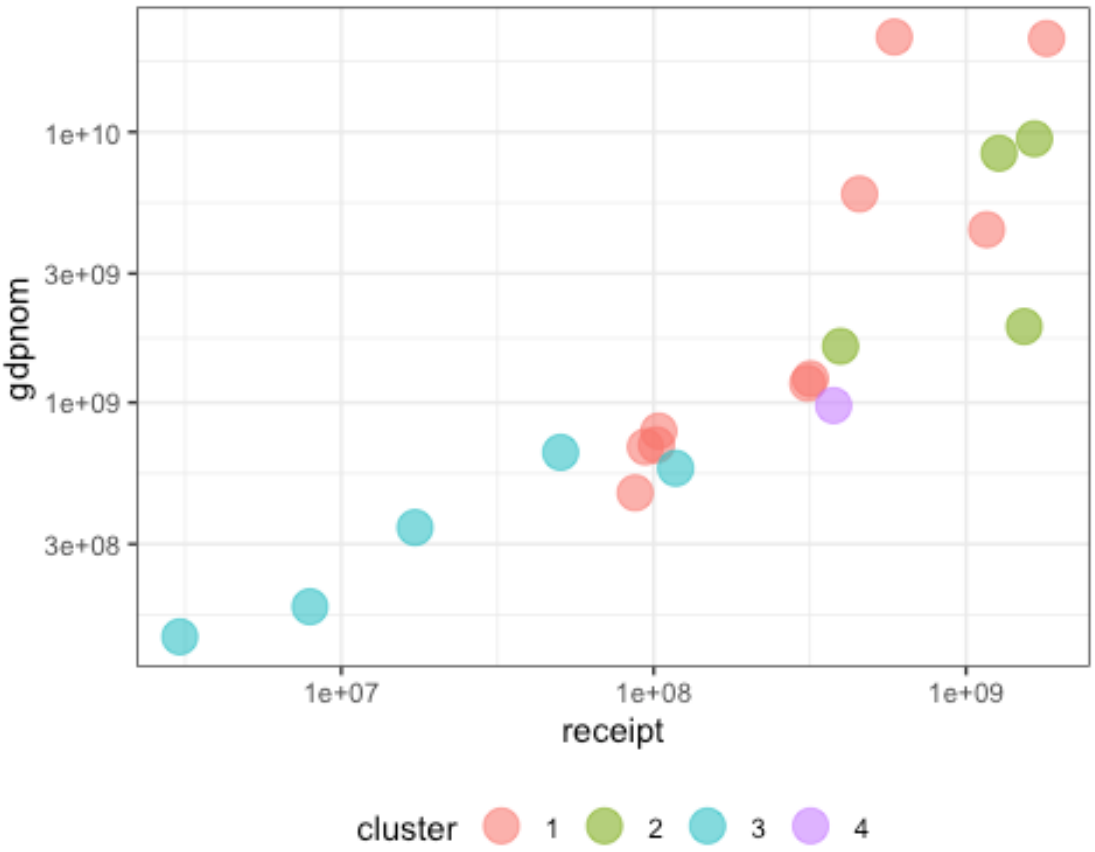


Fig 28. Relationship between the core variables and the clusters for clustering experiment 2.

Quite clearly, we see that cluster 1 countries are doing really well both tourism and economy wise as compared to cluster 2 countries. Interestingly 'Seychelles' the sole cluster 3 country with extremely low tourist counts from these regions seems to be fairly okay.

6.3 Clustering Experiment 3

In this experiment I used the variables 'arrbel', 'arrswe', and 'arrspa' for the clustering analysis. These variables represent tourist arrivals from Belgium, Sweden, and Spain respectively.

On analysis I noticed that these variables are jointly missing for many of the countries. Because of this I removed those observations from the data. The distributions of these two variables were skewed and I corrected that using log transformation. Following this I clustered the dataset into 2 clusters (based on elbow and silhouette method along with experimentation).

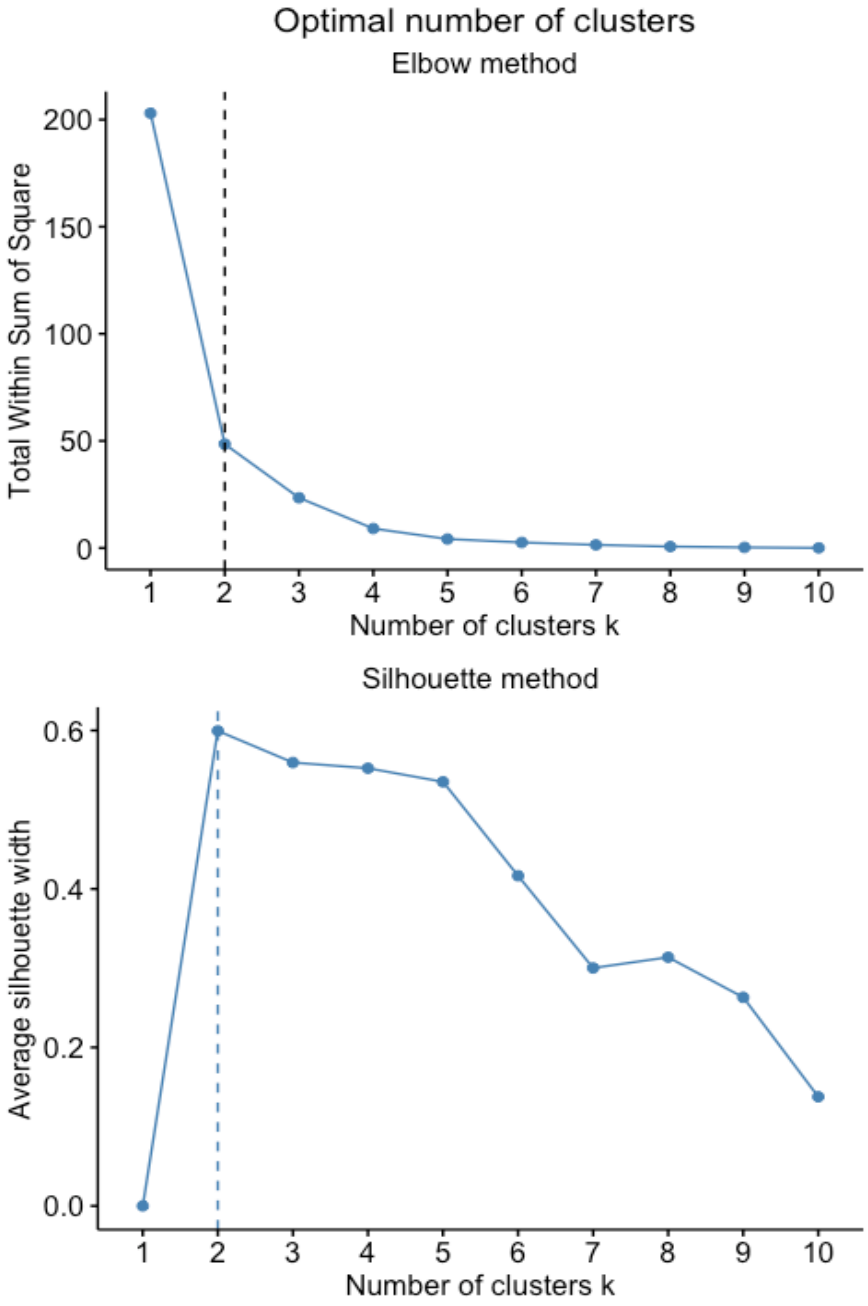


Fig 29. Determination of the optimal number of clusters using the elbow (top) and the silhouette (bottom) methods for clustering experiment 3.

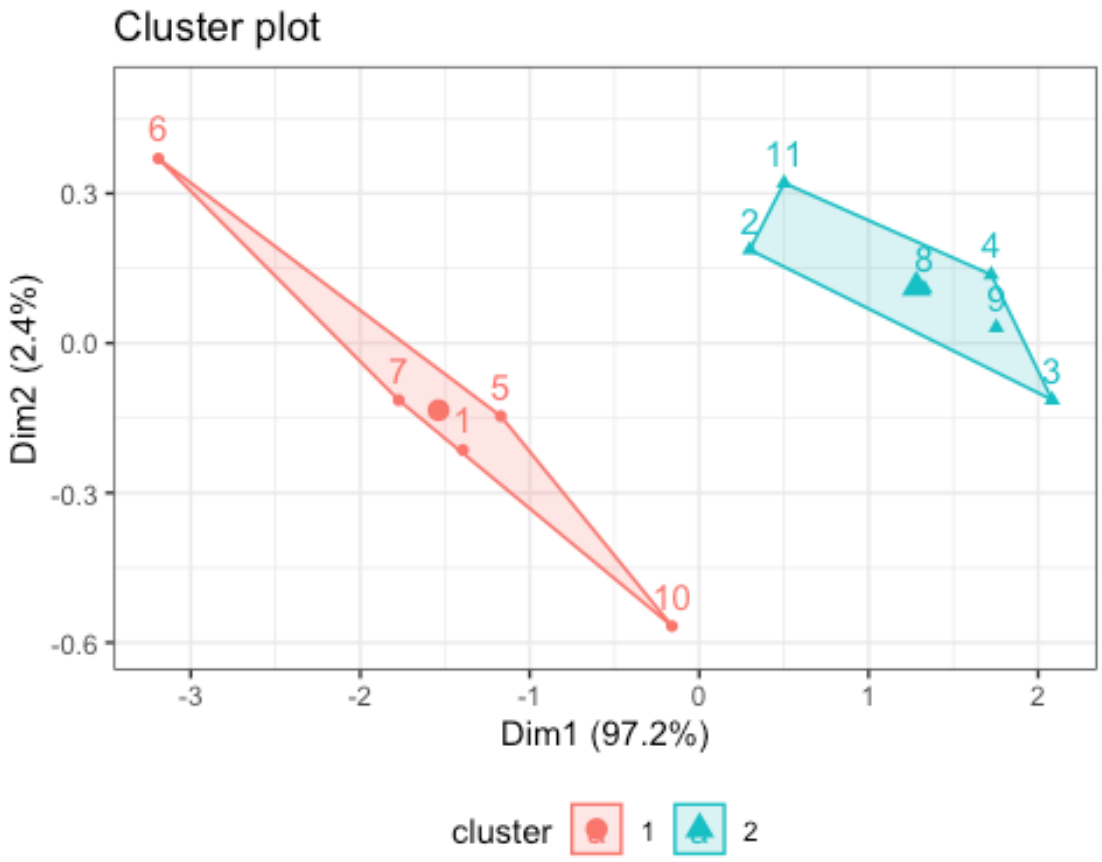


Fig 30. Cluster separation achieved on two principal components of the dataset in clustering experiment 3.

Table 9. *Cluster profiles obtained from clustering experiment 3*

cluster	pop	gdpnom	arrswe	arrbel	arrspa	receipt	ovnarriv
1	6.43e+05	8.57e+09	9.81e+03	1.04e+04	1.63e+04	1.33e+09	2.08e+06
2	3.43e+05	4.99e+09	8.25e+02	3.12e+02	4.27e+02	3.92e+08	2.58e+05

Table 10. *Cluster memberships obtained from clustering experiment 3*

cluster	countries
1	Bahrain, Maldives, Malta, Mauritius, Seychelles
2	Barbados, Dominica, Grenada, Saint Lucia, Saint Vincent & Grenadines, Trinidad and Tobago

Above we can clearly see the distinction between countries where tourists from these countries pay lots of visits and where they do not. Interestingly all the 6 countries in cluster 2 (the cluster with low levels of tourism from these countries) 4 belong to the Caribbean

region. This makes sense as Caribbean is geographically distant from these European countries relative to the countries in cluster 1 which are in the middle east, Africa, or the Mediterranean. Next, I compared these two clusters across the core variables to see what effect of increased tourism from these two regions has on those.

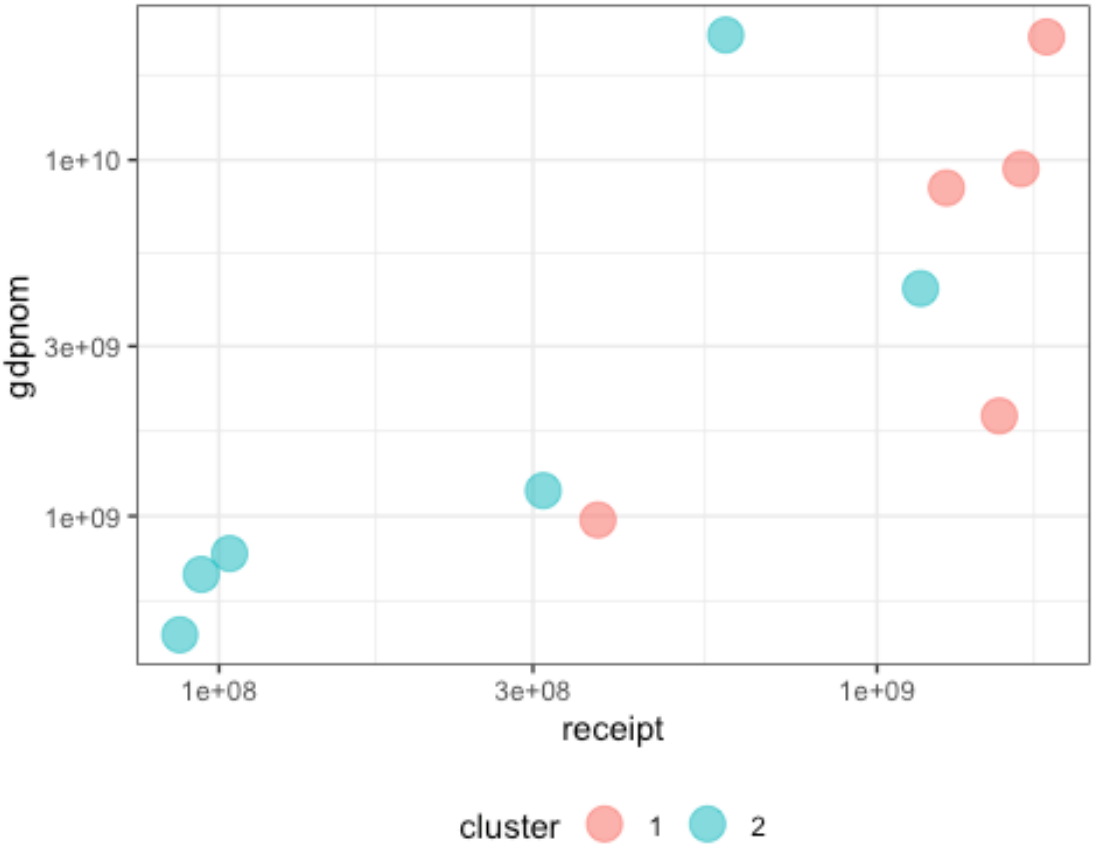


Fig 31. Relationship between the core variables and the clusters for clustering experiment 3.

For this experiment the effect on the core variables is not very clear. Further analysis will need to be done in order to draw any conclusions. Having said that, the top 4 countries on the basis of revenue generated by tourism do belong to cluster 1.

6.4 Clustering experiment 4

In this experiment I used the variables 'arrausl' for the clustering analysis. These variables represent tourist arrivals from Australia.

On analysis I noticed that these variables are jointly missing for a few countries. Because of this I removed those observations from the data. The distributions of these two variables were skewed and I corrected that using log transformation. Following this I clustered the dataset into 3 clusters (based on elbow and silhouette method along with experimentation).

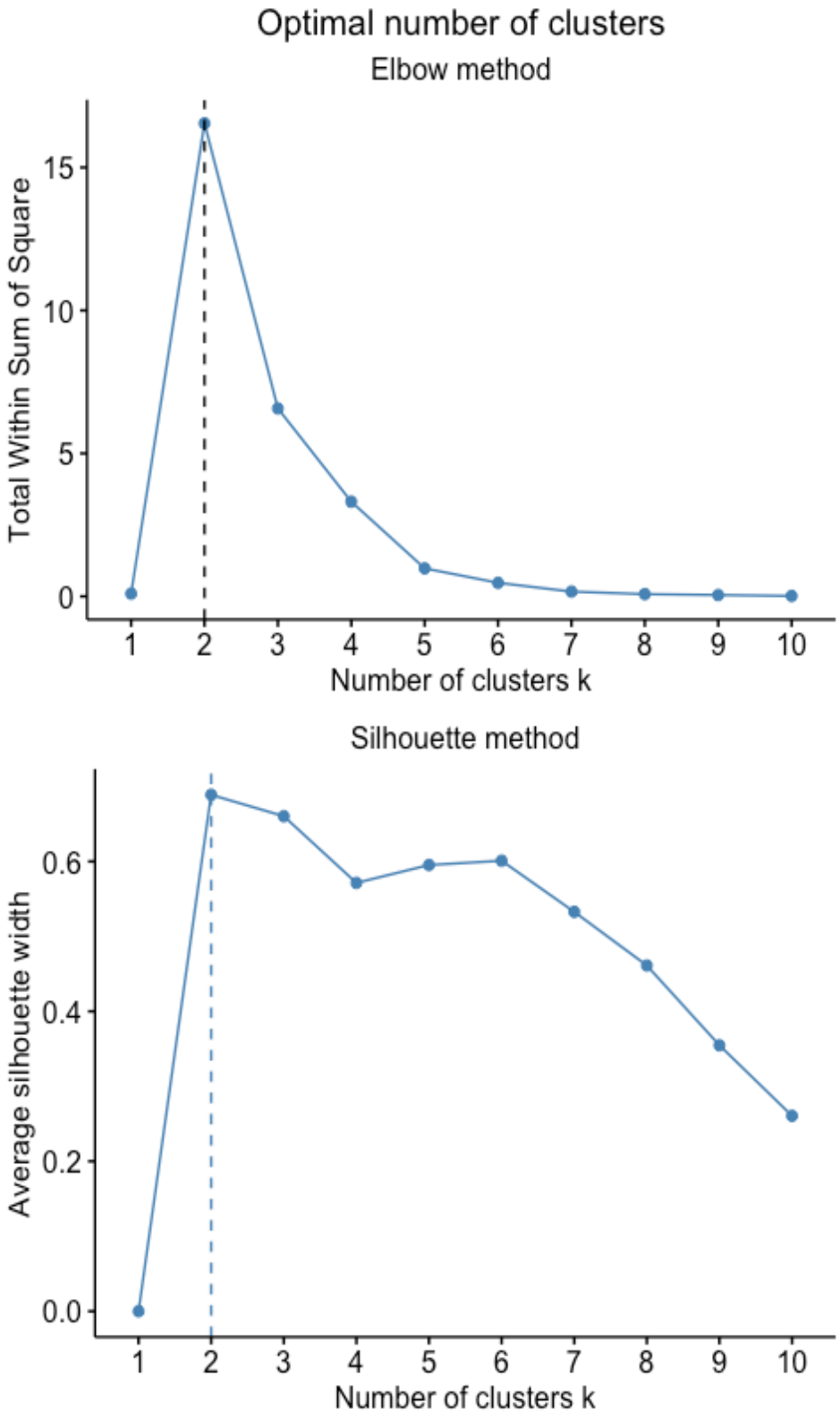


Fig 32. Determination of the optimal number of clusters using the elbow (top) and the silhouette (bottom) methods for clustering experiment 1.

Table 11. *Cluster profiles obtained from clustering experiment 4*

cluster	pop	gdpnom	arrausl	receipt	ovnarriv
1	5.42e+05	5.64e+09	1.47e+04	9.07e+08	1.38e+06
2	2.93e+05	5.12e+09	6.55e+02	2.69e+08	1.75e+05

Table 12. *Cluster memberships obtained from clustering experiment 4*

cluster	countries
1	Bahrain, Barbados, Maldives, Mauritius, Samoa, Solomon Islands, Tonga
2	Bermuda, Dominica, Grenada, Kiribati, Seychelles, Trinidad and Tobago

The aim of this clustering experiment was to see if Oceanic SIDS would do better than other SIDS when looking at tourist volume from a country much closer to them as compared to the others. Cluster 1 is the high value cluster and sure enough contains oceanic countries like Samoa, Tonga, and Solomon Islands. Interestingly Kiribati fell in the low value cluster 2 in this experiment. This goes to show that Kiribati doesn't get a lot of tourists in general and that they may need to invest in developing this sector of their economy. Below I map these clusters against the core variables.

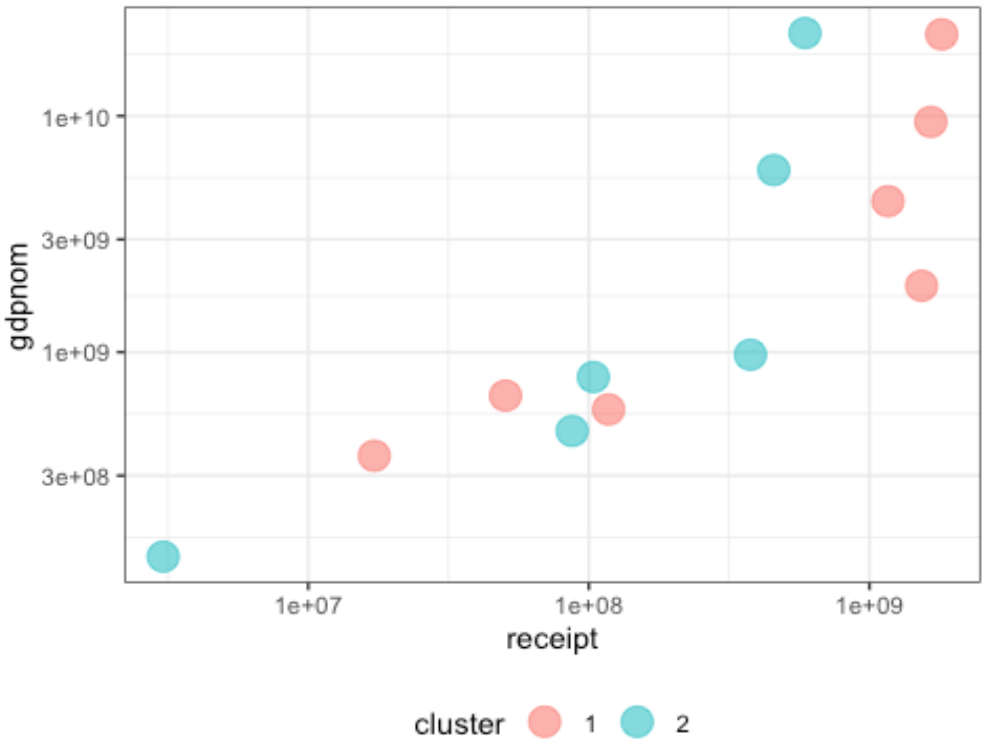


Fig 33. Relationship between the core variables and the clusters for clustering experiment 4.

6.5 Clustering experiment 5

In this experiment I used the variables 'arrfra' and 'arrit' for the clustering analysis. These variables represent tourist arrivals from France and Italy respectively. On analysis I noticed that these variables are jointly missing for a few countries. Because of this I removed those observations from the data. The distributions of these two variables were skewed and I corrected that using log transformation. Following this I clustered the dataset into 3 clusters (based on elbow and silhouette method along with experimentation).

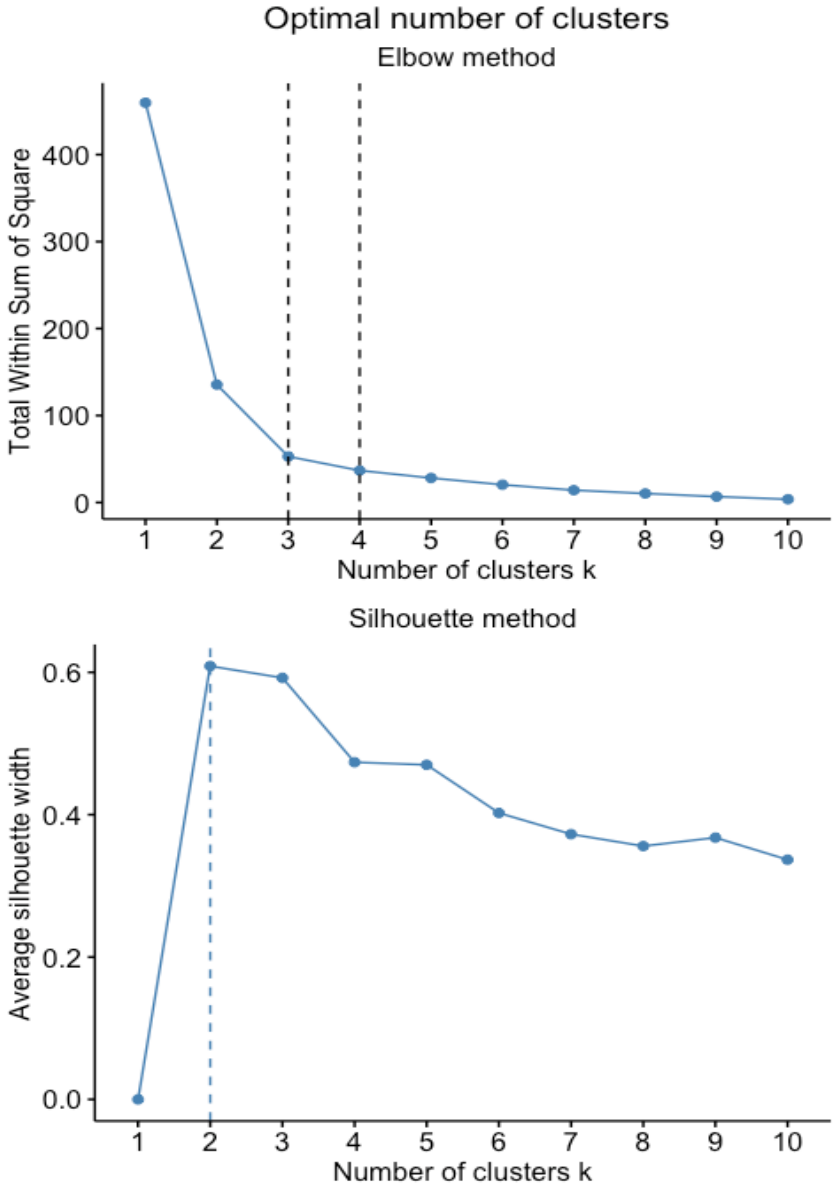


Fig 34. Determination of the optimal number of clusters using the elbow (top) and the silhouette (bottom) methods for clustering experiment 1.

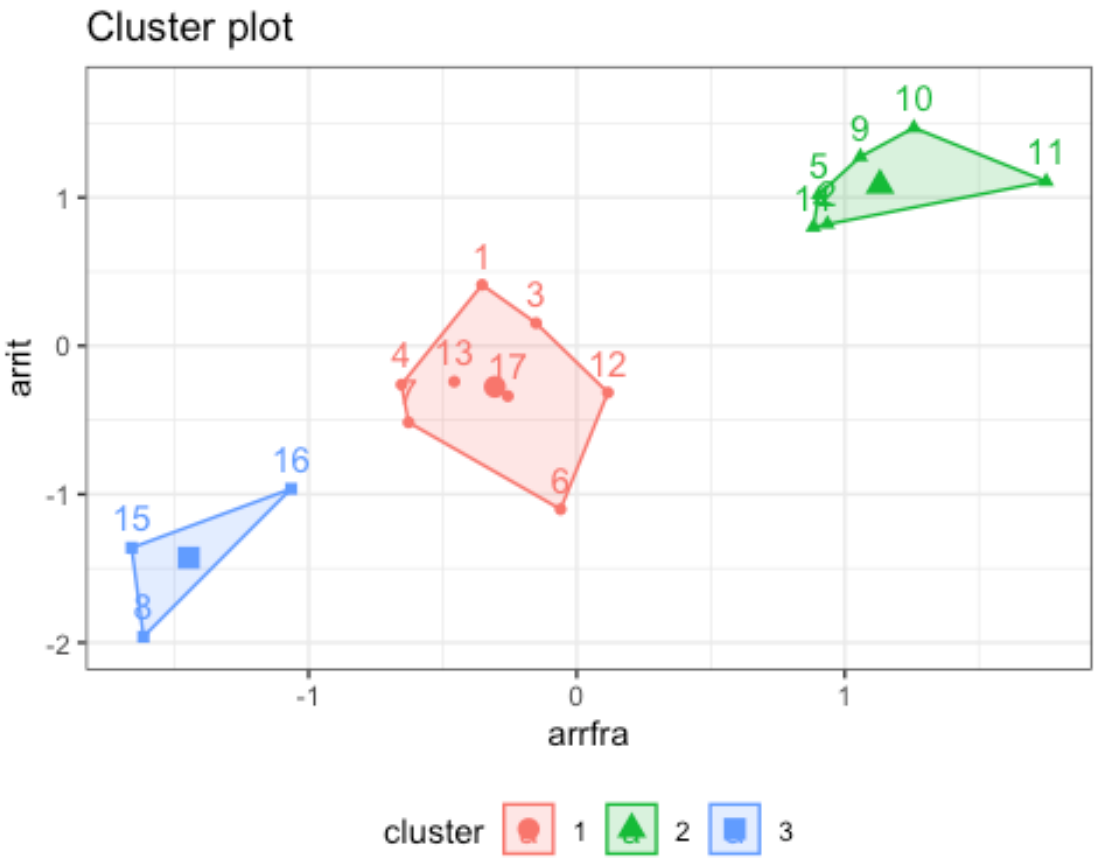


Fig 35. Cluster separation achieved on the two variables of the dataset in clustering experiment 5.

Table 13. *Cluster profiles obtained from clustering experiment 5*

cluster	pop	gdpnom	arrfra	arrit	receipt	ovnarriv
1	3.73e+05	4.63e+09	2.30e+03	2.25e+03	3.92e+08	2.56e+05
2	6.18e+05	7.41e+09	8.65e+04	7.06e+04	1.17e+09	1.79e+06
3	2.42e+05	4.08e+08	1.53e+02	8.08e+01	4.71e+07	2.29e+04

Table 14. *Cluster memberships obtained from clustering experiment 5*

cluster	countries
1	Antigua and Barbuda, Barbados, Bermuda, Dominica, Grenada, Saint Lucia, Saint Vincent & Grenadines, Trinidad and Tobago
2	Bahrain, Cape Verde, Maldives, Malta, Mauritius, Seychelles
3	Kiribati, Solomon Islands, Tonga

Above we can clearly see the distinction between countries where tourists from these countries pay lots of visits and where they do not. Interestingly all of the 11 countries in cluster 2 (the cluster with low levels of tourism from these countries) belong to the Caribbean or Oceania region. This makes sense as both Caribbean and Oceania are geographically distant from these European countries relative to the countries in cluster 1 which are in the middle east, Africa, or the Mediterranean. Next, I compared these three clusters across the core variables to see what effect of increased tourism from these two regions has on those.

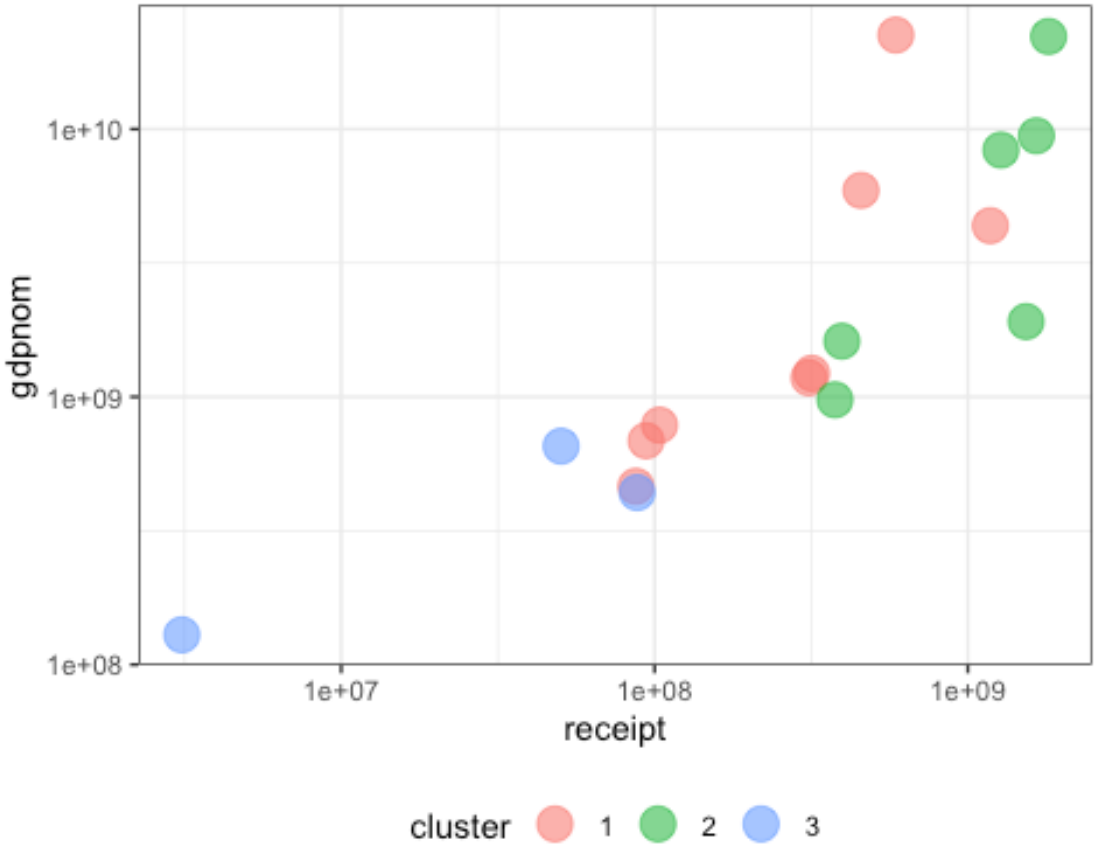


Fig 36. Relationship between the core variables and the clusters for clustering experiment 5.

7 Discussion

Overall through the various clustering experiments and multicollinearity analysis, I believe that certain patterns have emerged quite clearly. Before adding to that I'd like to clearly state that most of the conclusions I have drawn with the analysis must be taken with a grain of salt. The dataset was very messy and littered with missing data. While there might be natural reasons for that data to be missing, it still means that that affected my analysis. I had to created subsets of data on which I could run my experiments on along with imputing for missing values and thus the specific numbers I obtained may be unreliable. Having said that,

the general patterns I observed do make logical sense when combined with domain knowledge and geographical information.

As a part of the EDA, I explored the missing data patterns in the dataset and used that information to create subsets of data. I also explored the relationship between the economy of a country (represented by gdpnom) with other core variables such as population as well as tourism related variables such as tourist counts, tourism revenue, and flights data. I was able to establish quite clearly the direct relationship between GDP (logged) and Tourism revenue variable 'receipt' (logged). This is expected for these countries since their land area, population and geography is often not good for other industries and tourism often heavily dominates the economy[1,3].

In clustering experiment 1 the major idea was to establish core clusters that define how well a country is doing with respect to tourism and economy. The variables selected for this experiment reflected that. The clustering experiment was clearly able distinguish between countries doing well on that front like Bahrain and Malta and the countries that doing not so well like Kiribati and Tonga. Countries from the region Oceania all fell into the cluster with low values for core values and for good reason. They do not attract as many tourists as some of the of the other countries in the dataset due to their geographical location.

In clustering experiment 2 the arrivals from America and Europe were studied. A region that came out as a clear winner for American tourists was the Caribbean with almost all Caribbean countries falling in the cluster with high tourist volume from America. Popular tourist destinations on the other side of the world like Maldives and Mauritius fell in cluster 2 where the European tourists dominated their American counterparts. However, the lowest of the bunch was Seychelles with tourist so low that it merited its own cluster. The Caribbean countries appeal to a lot of tourists from America not only because of their attractions but also because of their proximity to the American mainland. They are easy weekend getaways with good connectivity. They also serve as popular destinations for European tourists. One reason for that may be that they have been popularized quite heavily. Interestingly out of the 5 countries in cluster 3 (the cluster with low levels of tourism from both America and Europe) 4 belong to the Oceania region. This makes sense as Oceania is geographically very distant from the American and European regions. Similarly, the other country in this cluster ('Sao Tome and Principe') is a tiny island nation near the coast of the country Gabon in the African continent. This makes sense as this is a little-known tourist destination.

In clustering experiment 3 the arrivals from Sweden, Belgium and Spain were analyzed. Again, the clustering was able to clearly separate the countries according to tourist volume and the results made geographical sense. All the countries in the low value cluster 2 were Caribbean countries while all the countries in the high value cluster 1 were relatively near Europe.

In clustering experiment 4 the arrivals from Australia was analyzed. This was done primarily to see if the geographical proximity effect also held true for oceanic countries and sure enough it did. In the high value cluster 1 oceanic countries such as Tonga, Solomon Islands and Samoa were included while the low value cluster 2 consisted of Caribbean and African countries.

The general pattern that came out from the overall analysis are that geographical proximity has a huge effect on the tourist arrival in these Island nations with a single airport. This makes sense as with a single airport the chances of getting direct connections to these islands from distant places is difficult which means that probably very few hardcore travelers visit these nations from far off places. On the other hand, people from nearby major countries like Australia, Spain, America and other countries on Europe visit the Islands nations nearby quite frequently and in large volumes. Another insight was the fact that the tourism industry plays a major role in the economic health of these countries and thus many of the low performing countries like Kiribati and Tonga should try harder to attract tourists so that they boost their economy. Another pattern that I was able to observe was that countries that share geographical proximity performed very similarly on the clustering experiments. This makes a lot of sense as more often than not the countries that share geographical proximity also share a lot of socioeconomic factors as well as similar cultures.

8 Conclusion

Through this project I was able to successfully apply unsupervised learning methods to establish patterns of similarity and differences between various SIDS on the basis of economic health, tourism, international arrivals from various countries and population. A key finding was that often geographically close countries show similar patterns across these criteria. For future work on this project a good extension could be to complete this dataset by acquiring data from external sources and augmenting it using various other data sets that may contain more demographic information like education levels, ethnic groups and professions amongst other variables.

9 Professional Issues

The dataset for this project was provided to me by my supervisor. Through some research I estimate the original source to be from the datasets released by World Bank. However, as I learnt more about report writing I realized that I should have asked my supervisor for the original source so that it could be used as a reference in this dissertation. I have only used free and open source software throughout the analysis for this project. I have also ensured that wherever I have used information from any sources to cite them. These have been included in the references section at the last of this report.

10 Self-Assessment

I think the project has been successful as I was able to obtain a good set of conclusions and patterns from my analysis. I learnt some very important transferable skills such as project management, communication, writing, scientific reports, reading literature, and managing expectations. I have learnt a great deal about unsupervised machine learning and reproducible research. Where I struggled was time management as I had a lot of trouble in

managing my academic and personal life along with this project. Hopefully, I can improve on this going forward.

11 Reproducing the analysis

In order to check the analysis done in this project, two R scripts: Data Cleaning Functions.R and Final Dissertation.R are provided with the submission of this report. Data Cleaning Function.R contains functions for cleaning of data which needs to be executed before the execution of script Final Dissertation.R. These scripts can be run directly from any directory so long as the dataset (provided as an excel spreadsheet) is in that directory. To run the scripts, R needs to be installed on your system. I would also recommend installing RStudio and using that to run the analysis.

12 References

1. Briguglio, L. *Small Island Developing States and Their Economic Vulnerabilities*. *World Development* **23**, (1995).

2. Scheyvens, R. & Momsen, J. H. Tourism and Poverty Reduction: Issues for Small Island States. *Tour. Geogr.* **10**, 22–41 (2008).

3. Roudi, S., Arasli, H. & Akadiri, S. Saint. New insights into an old issue – examining the influence of tourism on economic growth: evidence from selected small island developing states. *Curr. Issues Tour.* **22**, 1280–1300 (2019).

4. Sharpley, R. & Ussi, M. Tourism and Governance in Small Island Developing States (SIDS): The Case of Zanzibar. *Int. J. Tour. Res.* **16**, 87–96 (2014).

5. Signature Programme Development Nations United. Available at: https://cellcode.us/quotes/signature-programme-development-nations-united.html. (Accessed: 1st September 2019)

6. Telfer, D. J. & Wall, G. Linkages between Tourism and Food Production. *Ann. Tour. Res.* **23**, 635–653 (1996).

7. Ashley, C., Mitchell, J. & Mitchell, J. *Tourism and Poverty Reduction*. (Routledge, 2009). doi:10.4324/9781849774635

8. Lim, C. C. & Cooper, C. Beyond sustainability: optimising island tourism development. *Int. J. Tour. Res.* **11**, 89–103 (2009).

9. Scheyvens, R. & Momsen, J. Tourism in Small Island States: From Vulnerability to Strengths. *J. Sustain. Tour.* **16**, 491–510 (2008).

10. Croes, R. R. A paradigm shift to a new strategy for small island economies: Embracing demand side economics for value enhancement and long term economic stability. *Tour. Manag.* **27**, 453–465 (2006).

11. Parry, C. E. & Mcelroy, J. L. The Supply Determinants of Small Island Tourist Economies. **2**, (2009).

12. McElroy, J. L. & Parry, C. E. The characteristics of small island tourist economies. *Tour. Hosp. Res.* **10**, 315–328 (2010).

13. Josse, J. & Husson, F. *Handling missing values in exploratory multivariate data analysis methods*. *Journal de la Société Française de Statistique* **153**, (2012).

14. Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T. & Moons, K. G. M. Review: A gentle introduction to imputation of missing values. *J. Clin. Epidemiol.* **59**, 1087–1091 (2006).

15. A Complete Tutorial which teaches Data Exploration in detail. Available at: https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/. (Accessed: 29th May 2019)

16. Saar-Tsechansky, M. & Provost, F. Handling Missing Values when Applying Classification Models. *J. Mach. Learn. Res.* **8**, 1623–1657 (2007).

17. Aguinis, H., Gottfredson, R. K. & Joo, H. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organ.*

Res. Methods **16**, 270–301 (2013).

18. Big Data Made Simple - One source. Many perspectives. — Steemit. Available at: https://steemit.com/sciencefeed/@sciencefeed/20180213t070034427z. (Accessed: 1st September 2019)

19. Kassambara, A. *Practical guide to cluster analysis in R : unsupervised machine learning*.

20. *8 Cluster Analysis: Basic Concepts and Algorithms*.

21. Jolliffe, I. Principal Component Analysis. in *International Encyclopedia of Statistical Science* 1094–1096 (Springer Berlin Heidelberg, 2011). doi:10.1007/978-3-642-04898-2_455

22. Fusco, G. & Perez, J. Bayesian Network Clustering and Self-Organizing Maps under the Test of Indian Districts. A comparison. *Cybergeo* (2019). doi:10.4000/cybergeo.31909