Aviation Islands

Tarun Malarmaran

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

November 28, 2021

**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 Where other published and unpublished source materials have been used, these have been acknowledged.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**: 6374

**Student Name**: Tarun Malarmaran

**Date of Submission**: November 28, 2021

**Signature**:

**Abstract**

The single airport island nations’ data were gathered from various source including World Bank’s Dataset. This project aims at cleaning this ambiguous data with a lot of missing values using appropriate imputation methods wherever necessary and form a subset for various aviation island nations the data for a period of thirteen years.

To find correlation of the datapoints generated by these island’s data for each year to discover various attribute’s dependency and find corelation between islands based on a specific attribute to draw hypothesis based on certain pattern of growth. Applying various data visualising techniques to answer specific queries on finding similarities and inter-relationships such as area, population etc of different islands.

**Contents**

1 Introduction 1

1.1 Problem Statement 1

1.2 Motivation 1

1.3 Dataset Initial Analysis 1

2 Background Research 3

2.1 Literature Survey 3

2.2 Conclusions and Hypothesis 4

3 Data Wrangling 5

3.1 Obtain and Understand Data 5

3.2 Structuring and Cleaning 5

3.3 Enriching, Validate and Publish 5

4 Data Pre-processing 7

4.1 Imputation 7

4.1.1 Changing the Object types 7

4.1.2 Dealing with Missing Values 7

4.2 Dimension Reduction using PCA 9

4.2.1 2 - Components Analysis 9

4.2.2 2 - Components with Visual Loadings 10

4.2.3 PCA with variance explained 10

5 Exploratory Data Analysis 12

5.1 Hypothesis based on Corelations 12

5.2 Cluster Analysis 13

5.2.1 Hierarchical Clusters 14

5.2.2 Gaussian Mixture Model 16

6 Visualisation based on Analysis 18

6.1 GDP Trends 18

6.2 Flights Frequencies Trends 19

Conclusions 20

References 21

A Installing packages 22

B Code Listings 4

C Professional Issues 4

# Introduction

In the modern-day world, data is generated in large quantity every passing second. Every organisation relies on data to gain insights on their progress and act accordingly. In terms of the business needs, an efficient data analysis is very essential to study the data and make decisions. The reports which are generated and presented using the concept of Data Visualisation holds key in this process. Sometimes visualisation can be used by industries to trick the customers to get things in their favour, a classic example is a three-dimensional pie chart. So, it’s essential that a Data Analyst uses the best practises to Visualise the data and how it need to represented to the executives and general public.

Taking the above points into consideration, the task in this project is to Analyse the dataset of different island nations across the world. The data recoded in this dataset are about the economic statistics of the all the islands with their population data, geographic data and aviation statistics such as number of flights bound and via various modes of transportation. In order to proceed, the key concepts required are Exploratory Data Analysis and Data Visualisation.

## Problem Statement

The problem arises when I try to incorporate some meaningful insights from this dataset. The various statistical data are vaguely recorded in various sheets in the Microsoft Excel file which makes nearly impossible to make sense. To resolve this situation, various data cleaning and imputation methods have been incorporated in this project to perform the analysis.

## Motivation

The aim of this project it to understand the relationship between the island countries and it’s economic, geographic, population aviation data trends and draw hypothesis from the data based on the pattern of growth.

## Dataset Initial Analysis

This subsection outlines the details of the data available in the initial raw dataset, stating the attributes available for the different countries. The details are outlined based on each sheet in the dataset.

**Sheet**: The table in this sheet of the dataset seems to have the data of 40 island countries, with the attributes being about United nations and non-United Nations developing states. About population and year, it was recorded, area, number of arrivals from in different years.

**Sheet 1**: This sheet contains the data of traffic between Mauritius and Rodrigues which appears to be in a wide format but the column attributes are unclear, and around 70% of the cells have missing values.

**Sheet 2**: This sheet contains the individual data for Mauritius island, the attributes are almost same as the ones recorded in sheet 5, but the data here is in wide format.

**Sheet 3**: This sheet contains the individual data for Seychelles island, the attributes are almost same as the ones recorded in sheet 5, but the data here is in wide format.

**Sheet 4**: This sheet appears to have 2 separate tables, the wide table on the top appears to have some data about the rest of the world, Mauritius and Seychelles. The tall format table below seems to have the expenditure comparison among the three islands: Antigua and Barbuda, Mauritius and Seychelles.

**Sheet 5**: This sheet contains the data of 27 different islands recorded over 13 years of time span, where for each year the population, area, GDP, number of incoming flights, hotel rooms, visitor’s average expenditure, number of day visits, number of arrivals from different modes of transportation, number of arrivals from different counties, and certain inbound/outbound tourism expenditure attributes.

**Sheet 6**: This contains the data from Dutch territory, the attributes are almost same as the ones recorded in sheet 5, but the data here is in wide format.

From the initial analysis, it clear that the data is not proper structure in order to analyse it using visualisation plots. Before I could perform any analysis, there was a need to make a decision on what I am trying to explain. As my motivation was to observe the data trends, I will discuss on the decisions made on the dataset in section 3 Data Wrangling.

# Background Research

In this section I have discussed the observations and takeaways from the background survey carried out for the procedures in this project. This guides us in to the areas and concepts which I will be looking into in this project.

## Literature Survey

The authors [1] were assessing the contribution of different means of transport expenditures in the economic development of island country Mauritius. The concept which they had used is on-line time series on the data collected back from the year 1950 for Mauritius. The models in their study were Vector Autoregressive (VAR) and differenced Vector Autoregressive (DVAR) statistical model with feedbacks, which was used to present the relationship between the different modes of transportation and economic development individually. The interesting observation in their study was the revenues generated by the road transport and port harbors had a major influence on the island nation’s economic growth when compared to the revenue generated by the airport.

The authors [2] had conducted research on how tourism development in Small Island Developing States (SIDS) can potentially improve the economic state if the island country and considering Zanzibar, a state in Republic of Tanzania for the case study. In this study, the authors [2] have surveyed the data of Zanzibar’s international arrivals and the different modes of accommodation available in the state for the tourists. The final outcome of their study is that if the government had invested more with respect to the tourism of the country, Zanzibar could have improved their economic growth which would have taken them to a better state.

The authors [3] had conducted similar research as the authors [2], on Tourism as a tool to improve economy in a poor country as Comoro Island. It is observed that, the main reason of incoming passengers to Comoro is mainly because if tourism. And the arrivals being mostly from certain African countries and France. So, the conclusion they draw at the end is that, if the government of Comoro establish a partnership with the above-mentioned countries, then these partner countries can help Comoro sustain small-scale poverty reduction projects, so that Comoro can focus of scaling up the tourism resorts to improve the tourist experience and attract more tourists, ultimately leading to improve the country’s economy.

The authors [4] study about the Economic impact of World Bank’s investment in Aviation sector in Pacific Island countries. The world bank over the years had granted and supported the nations in Pacific islands towards projects related to aviation and air transportation. They analyse the effect on GDP by the different factors with respect to revenues from airports and efficient use of resources towards the improvement of aviation sector. They conclude by stating, the support from World Bank plays a vital role in the globalisation of the economy and boost the employment opportunities this leads to in that process.

The authors [5] study the impact of COVID-19 on aviation industry. They have analysed the data using the online-machine learning concept of weekly time series on the seats available on flights and using moving average concept of differencing and seasonal effect by Seasonal Autoregressive Integrated Moving Average (SARIMA). The data they have considered is form the John Hopkins University’s database. Their models are used to observe the regional statistics of number of cases and the number of flight seats booked based on the effect of pandemic and restrictions. As a result, they had provided the trends in the data and have quantified the disruption caused because on the pandemic on the aviation sector.

## Conclusion and Hypothesis

After the through study from the background research with the help of the literatures, I have arrived at a conclusion that all the authors have tried to show the relationship between the nation’s aviation industry and GDP. The point to note to that for a nation, it’s very important to have their tourism industry a success. This refers to all the ports of entry to the country and the level of services in airports, the lodging accommodation.

All the above-mentioned factors are the essential areas a government needs to look into and invest in, as the tourism industry contributes up to around 20% of a country’s economy.

# Data Wrangling

In most of the cases in real world, the data which I get to work on are not in meaningful formats. The data are mostly very noisy. So, for this purpose I perform Data Wrangling. Data Wrangling is a process of exploring data and transforming it step by step in order to perform analysis and obtain useful insights.

The ultimate goal of this process is to make data useful, that is to model the data using various tools and programs to gain valid insights using various visualisation techniques and to detect and rectify the various data quality issues.

## Obtain and Understand Data

The data was collected from an aviation outlet as the source and was initially in a Microsoft excel format. The first job was to look into the format. In this case the format of the data was pretty unclear at first.

The data recorded were segregated across different sheets in the excel file, there wasn’t a proper structure because of which I could have made an initial visual analysis. So, it was better to look through all the sheets clearly and make a decision on which one of those to retain to form the base data-frame.

It was clear that in the “sheet 5” the data was recorded for 27 of the island nations and for a span of 13 years. The other sheets did not have much data by which I could have made them as the base data-frame, but I have referred them in the future steps as and where required.

## Structuring and Cleaning

In this sub-section, I form a proper structure for the data-frame. In the previous section I had observed that “sheet 5” had the most useful data to perform the necessary tasks required in this project. So, diving deep into this data-frame from “sheet 5”, the data had many unwanted texts which had to be removed in order to make it useful.

The first task in this process is assign a data-frame header which behaves as the column name or attributes of the datapoints. The second task to follow is to remove the rows which contain the unwanted texts and the various completely empty rows, this is done using the pandas function dropna(). This function takes the argument, the axis along which I need to perform the task (in this case it’s the axis 0, the rows) and how I need to remove the Not a Number cells either its all of them are NaN or if just even one of them is NaN, (in this case all of them are Not a Number) as I will be dealing with the rows with just one or a few NaN using imputation.

## Enriching, Validate and Publish

In this subsection, I am looking at the data at hand after structuring it and looking at the ratio of missing values in the dataset. I was able to observe that nearly half of the island countries do not have enough data recorded in order to apply imputation and consider those for further analysis.

Now I am making the decision on removing the country or the column in order to progress further. The column which are almost completely empty are “hotrooms”, “hotel” and “expenditure”. These columns are removed because they don’t have enough data to proceed with imputation. The countries which are removed are “Grenada”, “Barbados”, “Bermuda”, “Cosmoros”, “Domnica”, “Kiribati”, “Fed Micro”, “St. Kits and Nevis”, “St. Lucia”, “Cayman Islands”, “St. Vincent & Grenadines”, “Tuvalu”, “Palau”, “Marshall Islands”, and “Cape Verde”. It’s a difficult decision to eliminate nearly half of the countries in the dataset, but this had to be done as keeping these countries data would and imputation straight way the whole column would have lead data ambiguity and incorrect conclusions.

After finalising on the data by removing the rows and columns, I performed a descriptive analysis for getting the statistics of the features that I had in the dataset. The function describe() analyses both the numeric and object types to give a summary statistics as show in Fig.3.3.1.

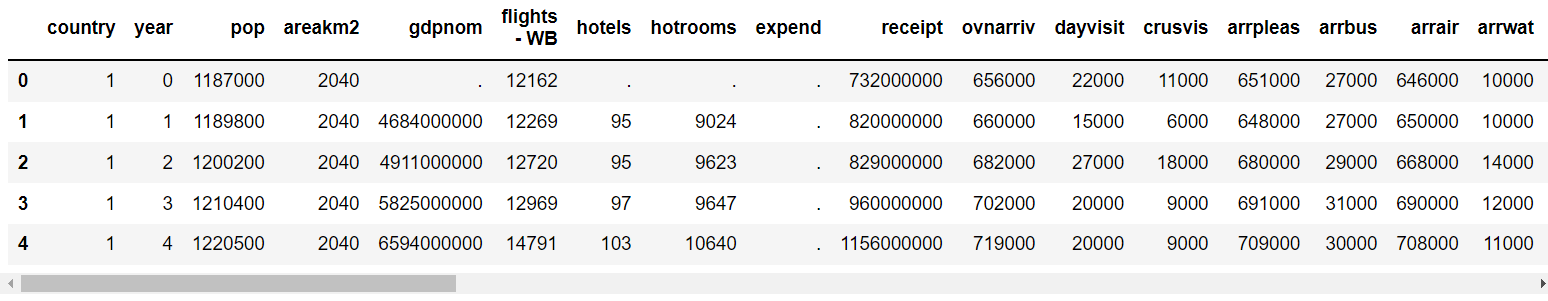


Fig.3.3.1 Summary Statistics before imputation.

The country codes new to be replaced with their actual name and this is done to finalise the dataset before processing for data imputation in the next section.

# Data Pre-processing

The data available in the real world tend to be inconsistent, missing values and noisy as the snippets or subsets are collected from various sources and of various types: structured, unstructured and on-line feed inputs (or time series).

The Data Pre-processing steps which I am applying in this section includes Imputation and Dimension reduction, and it is explained as follows.

## Imputation

The process of imputation which basically means assign the values where missing, that is I am finding a way for assigning or substituting values in the dataset where the values are missing mainly because the following 3 reasons:

**Missing at Random** (MAR): this happens when a feature or attribute which has a missing value, depends on the kind of attribute but not entirely on the missing values itself.

**Missing Completely at Random** (MCAR): this happens when a feature or attribute which has a missing value, does not depend on the other values in that attribute nor the missing values itself.

**Not Missing at Random** (NMAR): this happens when a feature or attribute which has a missing value, depends completely on the missing value.

In this regard of dealing with missing values, there are various strategies available: deletion methods, imputation methods and creating dummy variables.

### Changing the object types

The first thing I need to do in order to start imputation is have all the column data in numeric type. In this regard, I have replaced the NaN values to empty string and then change the whole column values is “pop” attribute to numeric using apply(). This converts the values to 64-bit floating-point values and then in turn to 64-bit integer values as it is meant to be for population attribute (pop). This makes the job easier to being the imputation process.

### Dealing with Missing Values

The imputation of missing values in this dataset are dealt differently from each column and by using different imputation for each country. There are different imputation methods based on the type of missing values, whether they are Missing at Random or Missing Completely at Random or Not Missing at Random.

In this dataset there are only Missing at Random or Missing Completely at Random values, this is where I can use imputation methods. If any dataset contains missing values which Not Missing at Random, in this case I cannot using imputation instead I should get back to the source of the data to acquire more information.

**Imputation by Most Common value**: If an attribute containing a missing value is continuous, I can replace the missing values with the mean of the attribute. If an attribute containing a missing value is discrete, I can replace the missing values with the most frequent value. This is one of the most simple and faster way to impute the missing values.

**Imputation by regression**: In this method, I replace the missing values by predicting values obtained by solving regression equations. This is very good as it utilises observed information in a very informative way.

To start with the imputation on this dataset,

* “pop”: this attribute denotes the population of the island countries for various years. I applied imputation country by country, and for the island ‘Seychelles’ and ‘Antigua and Barbuda’ I use the mean value and for the other islands except for Mauritius (no missing data) I used maximum value in that column for that particular country, this is because for these islands only the thirteenth year’s values were missing and the population values were continuous and increasing.
* “gdpnom”: this attribute denotes the nominal gross domestic product of the islands for various years. I applied imputation for just ‘Mauritius’ as only it’s first year’s value was missing and the series was continuous and increasing.
* “flights – WB”: this attribute denotes the world bank international departures of flag carrier. I applied imputation for ‘Mauritius’, ‘Maldives’, ‘Samoa’, ‘Sao Tome & Principle’ and ‘Tonga’ with their most frequent attribute values (median) as the missing data are in discrete series.
* “receipt”: this attribute denotes the world bank receipts for inward tourism. I applied for all the islands with each of their maximum values respectively, this is because in all the countries only the thirteenth year’s values were missing which were continuous and increasing.
* “ovnarriv”, “arrpleas” and “arrair”: these attributes denote the world bank overnight arrivals, the air arrivals and arrivals by sea planes respectively. I applied for all the islands with each of their most frequent attribute values (median) as the missing data are in discrete series.
* “tcov”, “intxexg”, “intxexs”, “intxexal”, “intxcac”, “oteximg”, “otxims”, “otximal” and “otxcad”: these attributes are world bank expenditure values calculated by the tourism expenditure, imports and exports. All of these values are continuous and hence imputation method used in these cases is mean.

By following the above steps, I have filled all the missing values in the dataset imputation methods which were efficient enough for each island country case by case. Now that the imputation is done, the data is ready to be analysed and gain insights, corelation and make note of the patterns.

## Principal Components Analysis (PCA)

The purpose of pre-processing is making the data ready to get better insights and understand it’s characteristics efficiently. In this process I come across a usual problem, which is the “Curse of Dimensionality”. The term Curse of Dimensionality is to cite the problems arising because of having many attributes as this increases the dimension of the data and makes analysis difficult. In order to deal with this, I performed a Principal Components Analysis.

Principal Components Analysis is linear feature extraction method to reduce the dimension of the dataset by creating entirely new features by combing the existing ones. The aim is to represent the data points in a lower dimensional space by applying linear transformations which maximise the variance.

The new features in the dataset are called the principal components, which are orthogonally corelated and ordered by the amount of information based on their original values. Principal components analysis is an unsupervised learning method which is used to discover patterns in data without referring to the prior knowledge on how the samples in a feature originated. One of most useful benefits of Principal Components Analysis is the lower dimensional data components can be projected on to 2-dimensional or 3-dimensional visualisations. Following are the visualisations obtained after Principal Components Analysis for 2 and 3 components.

### 2 – Components Analysis

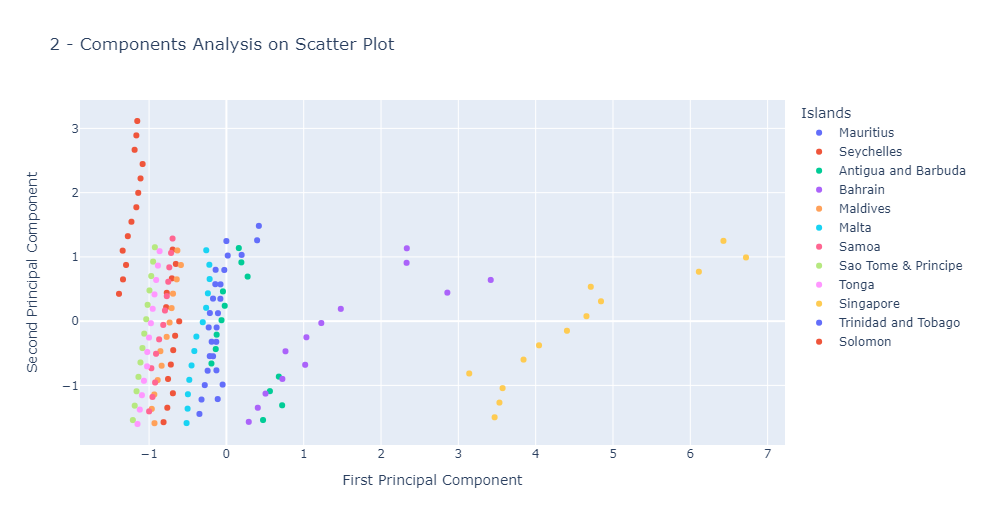
In this subsection, I am looking into the basic Principal Components Analysis plot for 2 components. Here, I reduce the data of 5 dimensions to 2 dimensions. In order to visualise, I use the help of plotly to plot a scatter plot which shows how the islands are distributed.

Fig.4.2.2 Two-Components Analysis.

As depicted in Fig.4.2.1, I was able to observe that the islands which are very different in terms of their data in this case are Solomon and Singapore. Singapore seems to be separated from the rest of the pack by a very huge margin, this may be because of how popular Singapore is with respect to its number of international visitors, the country’s Gross Domestic Product and so on which I will be getting into more depth in the next section.

### 2 – Components with Visual Loadings

In this subsection, I am plotted the same plot as in the previous subsection, but with the visual loading depicting the direction of variance. This is calculated as

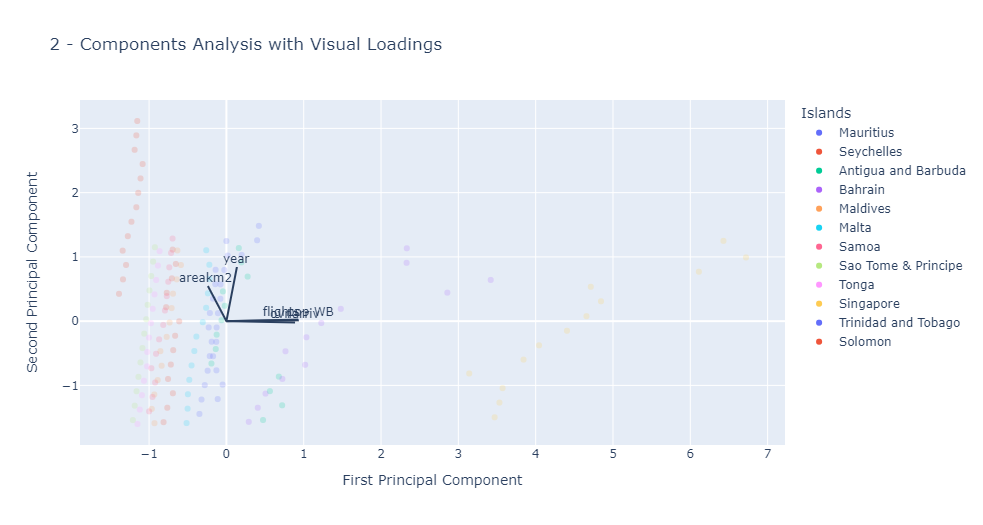
The direction here is the covariances between the initial values of attributes and the scaled down principal components.

Fig.4.2.2 Two-Components Analysis with Visual Loading.

From the Fig.4.2.2, I was able observe that, as the features “areakm2” and “ovnarrive” are very closely related. There may be a pattern emerging there between those two features, which I explained in detail in the next section.

### PCA with variance explained

In this subsection, I have observed an interesting concept of the explained variances. This is to analyse how the PCA fit explains the variance of the attributes with increasing number of components.

From Fig.4.2.3(a) below, I observed that the variance increases with the increase in number of components. This makes it clear, with a high variance explained I was able to incur more variability in the data which can be very helpful in building a better training model to get the best performance.

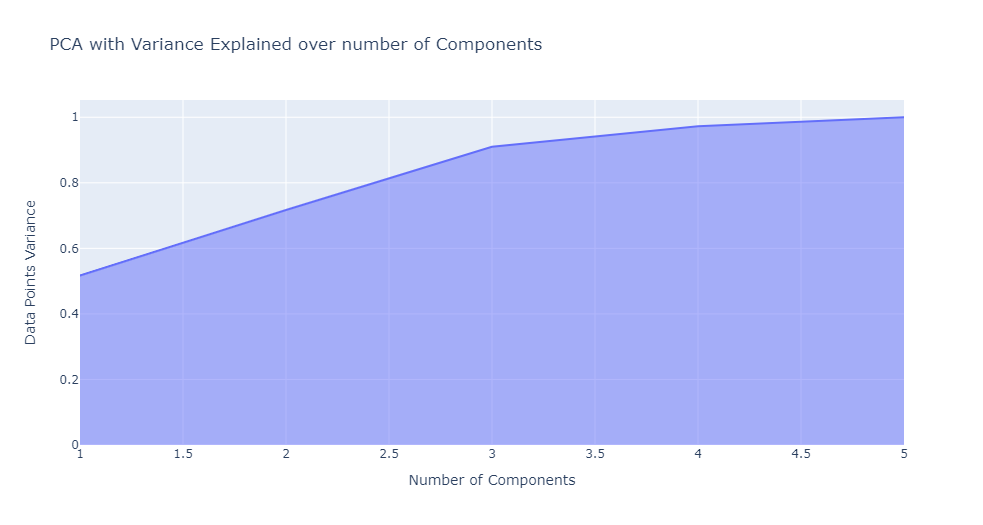
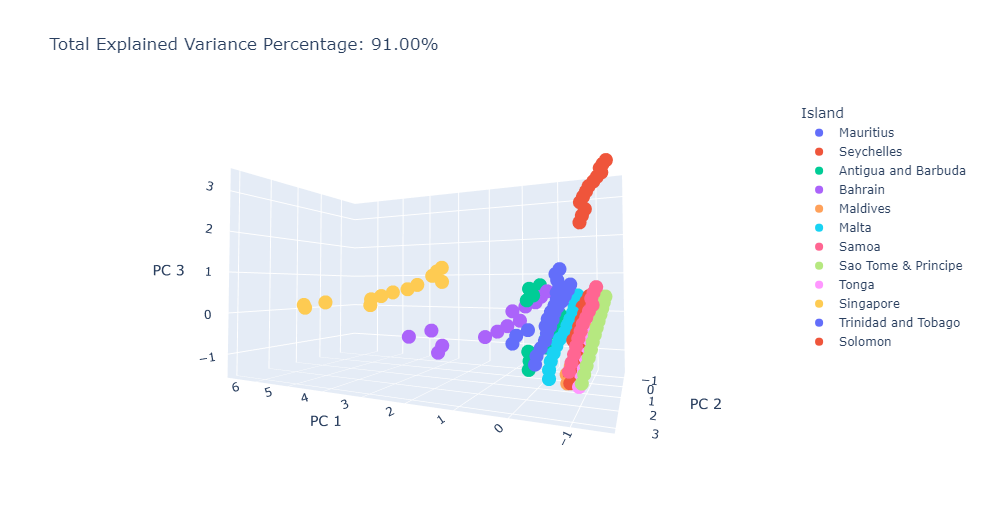
The Fig.4.2.3(b) above plot is a three-dimensional view of the PCA scatter plot for 3 components, which states the total explained variance of 91% for the given dataset.

Fig.4.2.3(b) 3D PCA plot with Total Explained Variance.

Fig.4.2.3(a) PCA with Explained Variance over number of Components.

# Exploratory Data Analysis

The term Exploratory data analysis refers to analysing the dataset to summarize their main characteristics using machine learning algorithms and different visualisation tools and techniques.

The data when first looked at after data wrangling, cleaning and imputation, may be overwhelming to asses and draw insights. To ease onto this process, I first begin with finding the corelations between the attributes in the first subsection. And I proceed with this to gain further insights by performing a multivariate cluster analysis.

## Hypothesis based on Correlations

In this subsection, I am studying about the correlations of the attributes in this dataset. The term correlation is the statistical measure of the level of dependance between two variables. Since in this case I was trying to find the correlation between all the attributes, this can be represented as a correlation matrix.

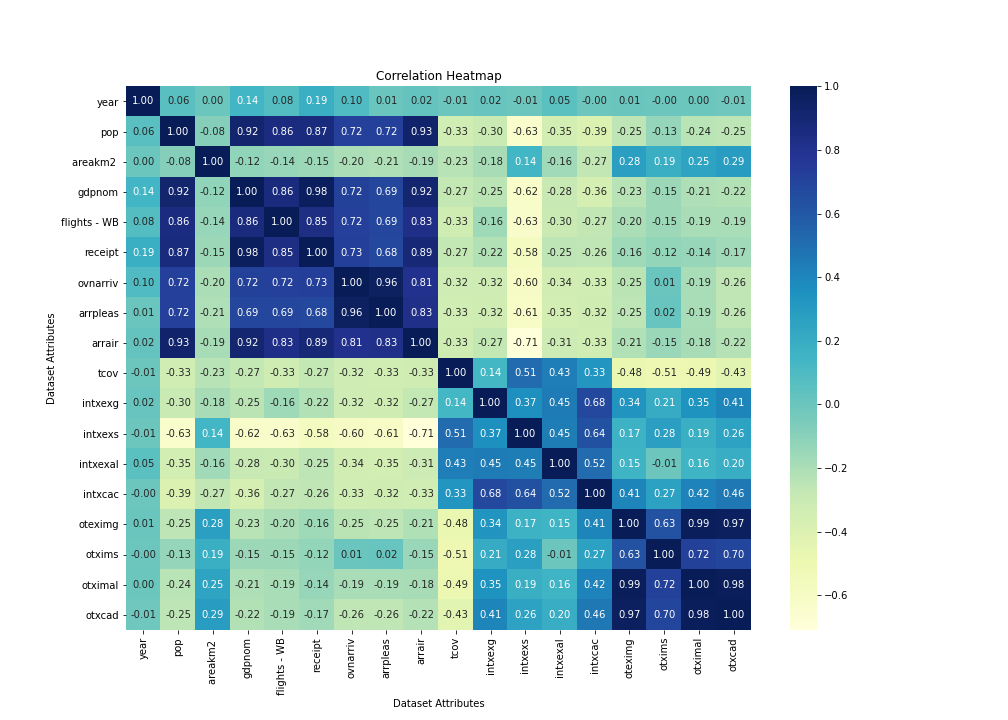
In a correlation matrix, the values of the first dimension (dependent attribute) are the rows and the values of the second dimension (depending attribute) are the columns.

Fig.5.1 Correlation Heatmap of the Attributes in the Dataset.

The correlation is calculated as per the Pearson correlation coefficient theory which gives the measure of how two variables differ with respect to one another, this is explained by the formula below.

Here, and are the mean values of *x* and *y* respectively. And *xi* and *yi* are the values at each point of calculation in X and Y.

In this project, I am illustrating the correlations between attributes using a correlation heatmap. The color in each cell of the heatmap is the strength of affinity between the attributes assisted by the colourbar. Fig.5.1 shows the heatmap correlation for the dataset.

From the correlation heatmap I was able notice some very strong affinities. The most closely related attributes are “gdpnom” and “receipt”, this tells us that based on the incoming tourism receipts is directly proportional to a country’s GDP. The attributes denoting the population, arrivals by air, flights traffic, overnight arrivals and arrivals by sea planes, play a very important role in forming a country’s GDP.

Similarly, the international tourism receipts of a country depend on the number of flights bound and the number of overnight arrivals. So basically, a country’s tourism index plays a huge role in generating the revenue and building the economy.

## Multivariate Cluster Analysis

Multivariate analysis is the study of experiments in which more than one features or attributes are taken in to consideration on every experiment to discover their relationships among all the other variables.

One of the categories of this type of analysis is cluster analysis. Cluster analysis is considered as statistical method for analysing data by categorising them into groups or here called as clusters based on how closely thy are related to each other. I use clustering to find datapoints which are similar and find patterns associated with these groups of data.

Clustering is an unsupervised machine learning algorithm. One would not know the exact number of clusters in the dataset beforehand. The clusters are formed by calculating the distances between the data points. There are two types of measurements which I can take into consideration, the Intra-cluster distance: which calculates the distance between any datapoint within a cluster. If there is a strong affinity, this value should be small which is more homogenous. And the Inter-cluster distance: which calculates the distance between any datapoint in different clusters. If the clusters have greater affinity, this value should be very large which is more heterogenous.

### Hierarchical Clusters

Hierarchical clustering as the name suggests, creates a nested tree structure of the clusters by amalgamating or bifurcating the datapoints. This hierarchy is illustrated as a tree structure called as a dendrogram.

**Agglomerative clustering**: This type of hierarchical clustering algorithm builds a tree in a bottom-up approach. At first, each data points are assigned their own clusters, and based on the different types of merge strategy, these clusters containing single datapoints are merged, this processed is recursively done until it arrives at the top of the tree creating a root cluster.

The different linkage methods used to form the merge strategies are:

* **Single linkage**: This method minimises the closest pair of clusters.
* **Average linkage**: This method minimises the average distances between all the pair of clusters.
* **Complete or Maximum linkage**: This method minimises the maximum distances between all the pair of observations.
* **Ward or Minimum linkage**: This method minimises the sum of squared distances between all the pair of observations.

In Fig.5.2.1(b), I have plotted the complete linkage dendrogram for this dataset considering all the datapoints with the country names being the labels. Here I observed that, nothing is clear to us, but I was able to notice the datapoints for the countries like Solomon and Singapore are all grouped together from the very beginning (small branches down the tree).

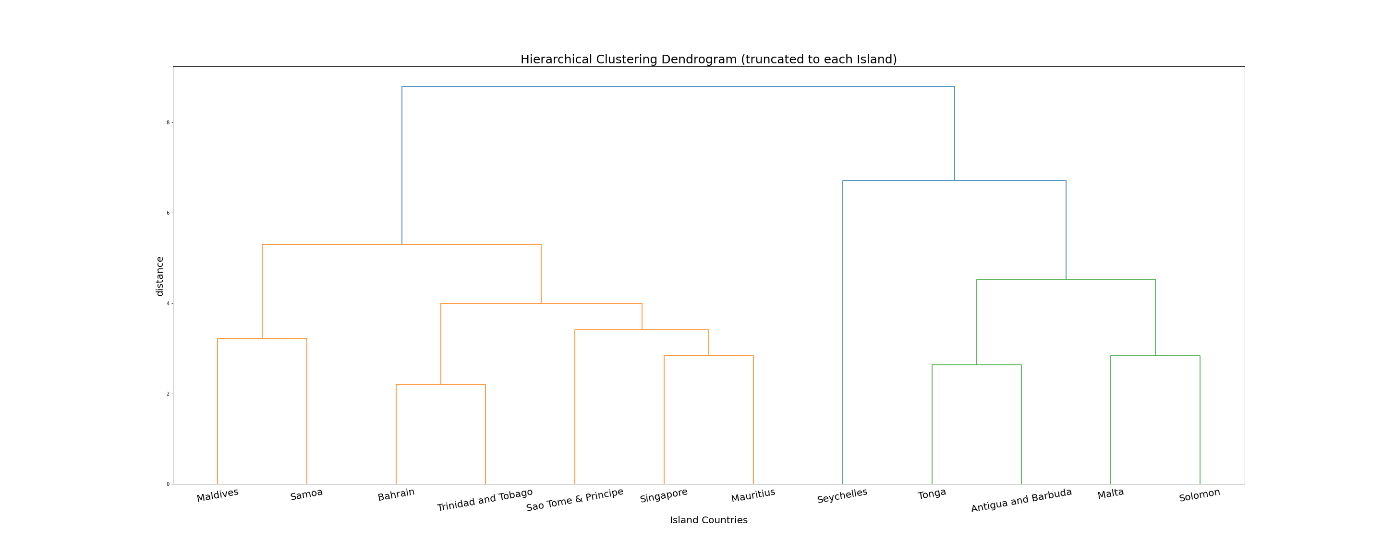
But, on interesting thing to note is Bahrain is clustering with Singapore and also some of its datapoints shows similarity with other islands, this mean that for some of the years the data of Bahrain shows changes vigorously in affinity to other island countries in those years.

Fig.5.2.1(a) Truncated Hierarchical Clustering Dendrogram.

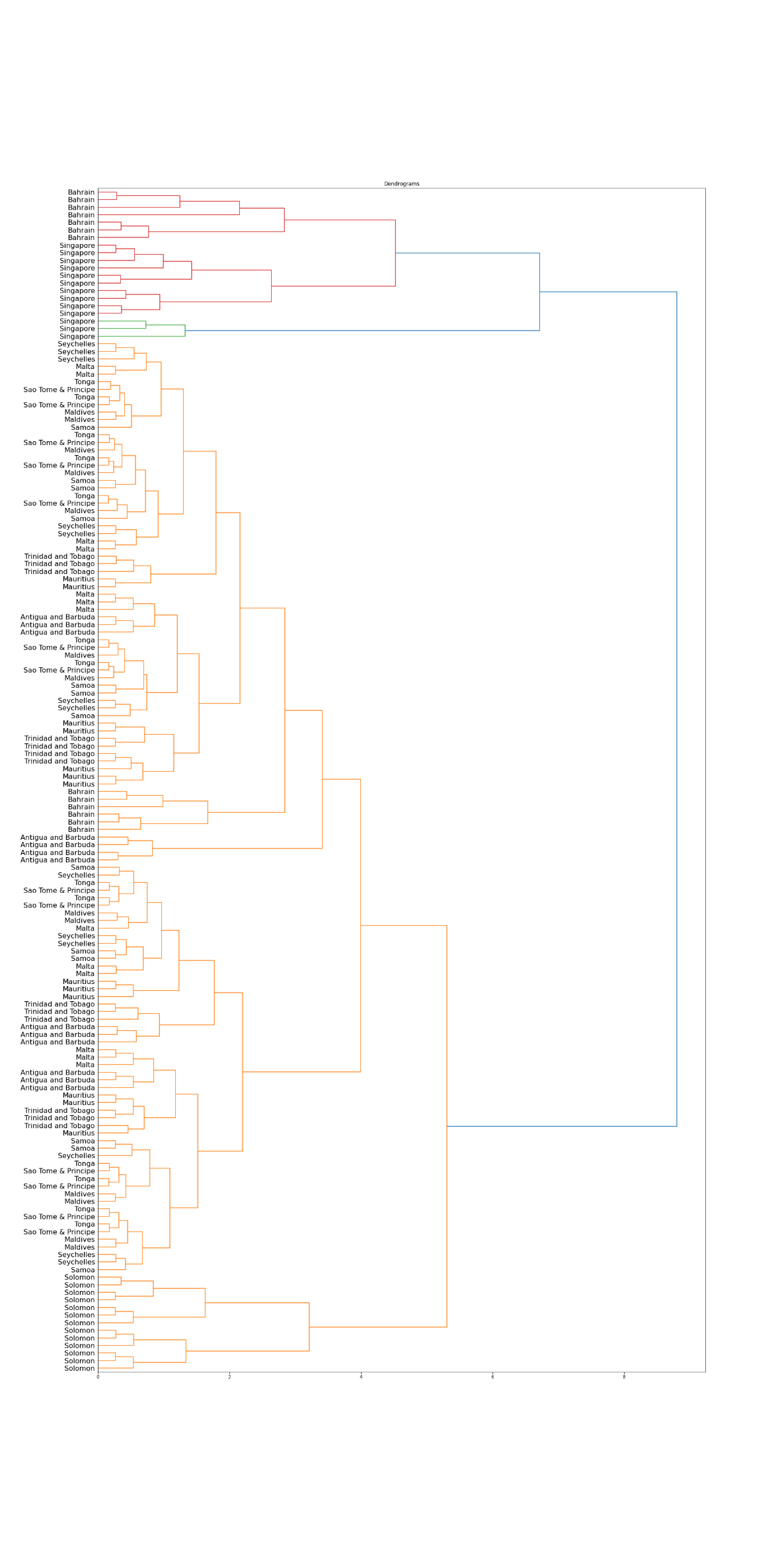


Fig.5.2.1(b) Hierarchical Clustering Dendrogram.

In Fig.5.2.1(a), I have cut down the tree to just 12 leaves stating, representing the 12 islands I have considered in the dataset. Using complete linkage, I have plotted the most occurring country in the cluster at the leaf levels and finding the clusters between them.

Now, from my earlier observation, in which I had noticed Mauritius and Seychelles having closely related datapoints, this argument can be support using this dendrogram. I can see these two islands are clustered first in the right most branch of the dendrogram.

### Gaussian Mixture Model

Gaussian mixture model is a probabilistic clustering algorithm to represent a distribution of subpopulations enclosed by the complete population. As Gaussian mixture model is an unsupervised algorithm, the model doesn’t have any prior knowledge on any datapoint’s label.

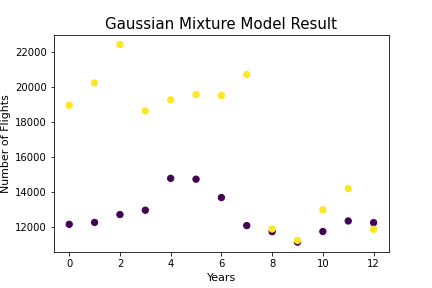
In a Gaussian mixture model if the number of clusters k is already known, the method used to estimate the mixture model’s parameters is expectation maximisation algorithm. In this method the model is trained with the help of maximum likelihood estimation techniques, which is used to maximise the likelihood or probability of the observed data against the model parameters.

Fig.5.2.2 Gaussian Mixture Model Result

In Fig.5.2.2, I am using the gaussian mixture model to show the relation between the islands countries Mauritius and Seychelles. I have illustrated the pattern of changes to the number of flights traffic to and from the islands with the increasing time in years.

I observed that, the gaussian mixture model with two components one for Mauritius and the other for Seychelles. I observed the model’s prediction of clusters for this set, and the cluster 1 seems to have a raise in flights traffic between years 4 to 6, then the number seems to take a dip (lowest of all the observations in year 9.

With respect to the predicted cluster 2, I observed the values are chaining very rapidly, but can notice a low traffic period between year 8 to year 12, there must be other factors due to which this has happened, I have discussed more about this in the next section.

# Visualisation based on Analysis

The dataset is ready to be applying pre-processing steps and impute the missing values. After this is done, the Dimensionality reduction is done using the Principal Components Analysis to get a better understanding of the data presented to us.

**Note**: The plots in this section are interactive. The islands can be selected by double clicking the selected countries on the Jupyter Notebook outputs.

## GDP Trends

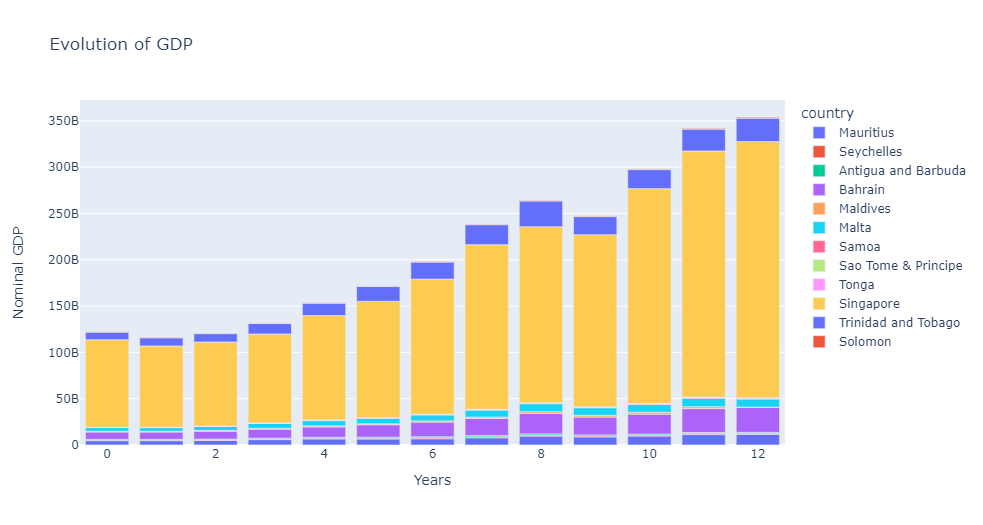
The evolution of GDP for the islands across the 13 years’ timeline, Fig.6.1(a) shows the growth pattern of all the island countries.

Fig.6.1(a) Evolution of GDP over the years.

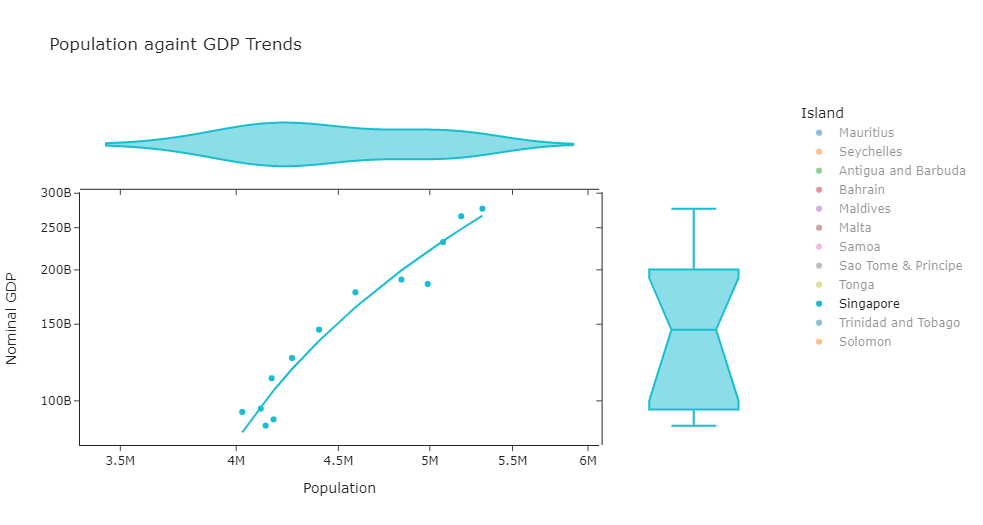
Raise in population means, the country is capable of putting more towards increasing its economy.

Fig.6.1(b) GDP Growth of Singapore.

So, to argue the correctness of this point, below is the log-log plot visualisation for population v nominal GDP, with the help of trendline I have observed that, as the receipt increases, the nominal GDP of a country increase. The purpose of using a log-log scale is to have a better visualisation of this correlation between receipt and nominal GDP and explaining the steps followed in pre-processing and imputation.

In Fig.6.1(b), I have selected the distribution of Singapore. Since Singapore is the most popular country in the dataset, from the plot I was able to note that Singapore has a uniform pattern of growth in this economy. As the population increases, the nominal GDP increases even though I was noticed a dip in GDP in year 9 from Fig.6.1(a) earlier by selecting just Singapore.

## Flights Frequencies Trends

In Fig.6.2, I observed the flight frequencies for the Islands in this dataset. The Island with the least traffic is Sao Tome & Principe with around 1.2k flights on average every year.

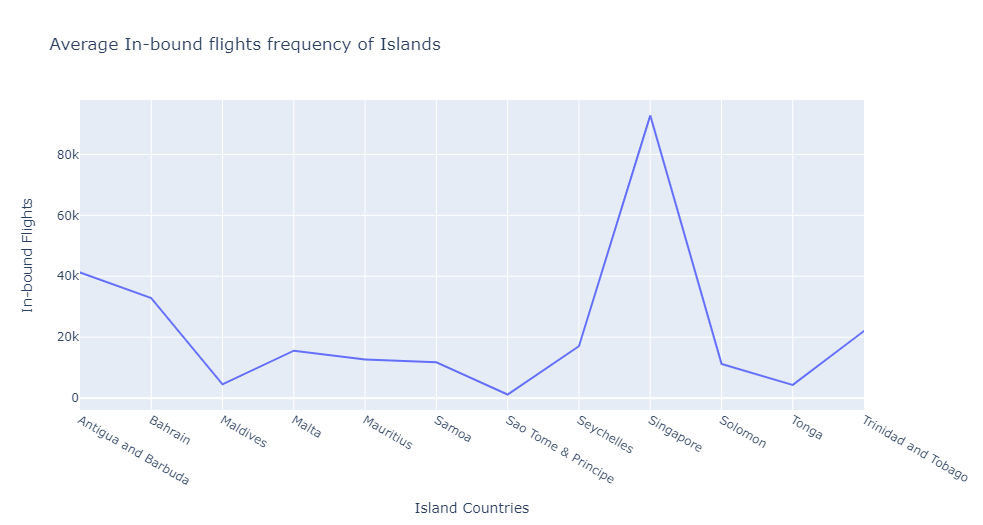
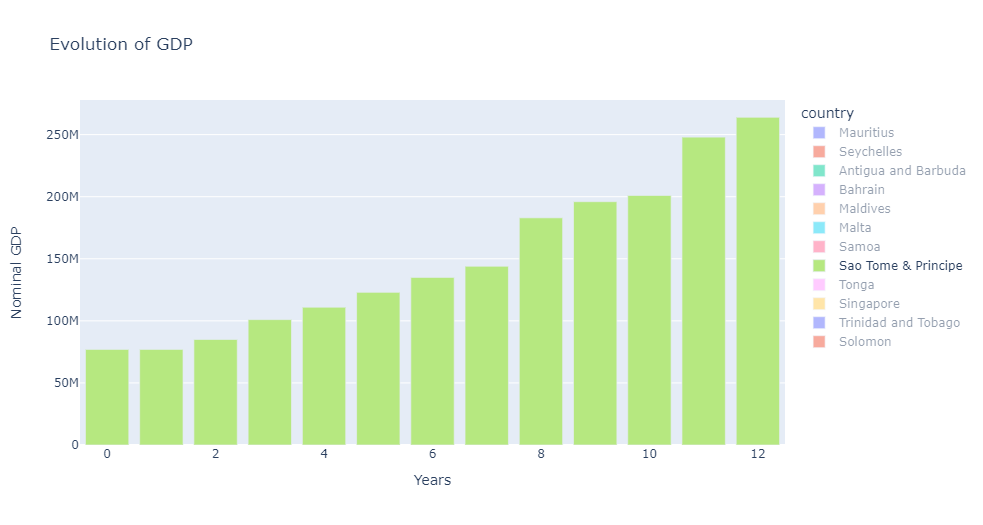
But Sao & Tome Principe’s nominal GDP has increased uniformly, which mean the island’s economy is improving with every passing year.

Fig.6.2 The mean Flight traffic for different islands.

# Conclusions

This section covers up on explaining all the findings, hypothesis and conclusions drawn from the results of the analysis of the Aviation Islands dataset in this project.

**References**

1. Khadaroo, A, & Seetanah, B 2007, 'Assessing the contribution of land, sea and air transport capital to the economic performance of the small island state of Mauritius', Applied Economics Letters, 14, 15, pp. 1151-1155, Business Source Complete, EBSCOhost.
2. Sharpley, R, & Ussi, M 2014, 'Tourism and Governance in Small Island Developing States (SIDS): The Case of Zanzibar', International Journal of Tourism Research, 16, 1, pp. 87-96, Business Source Complete, EBSCOhost.
3. Shaaban, I, Ramzy, Y, & Sharabassy, A 2013, 'Tourism as a Tool for Economic Development in Poor Countries: The Case of Comoro Islands', African Journal of Business & Economic Research, 8, 1, pp. 127-145, Business Source Complete, EBSCOhost.
4. Cindy Greenman1 & Javad Gorjidooz, 2017, ‘Economic Impact of World Bank Aviation Investment in Pacific Island Countries’, Journal of Economics and Public Finance ISSN 2377-1038, Vol. 3, No. 3, Business Department, Embry Riddle Aeronautical University, Prescott, AZ, USA.
5. Gianmarco Andreana, Andrea Gualini, Gianmaria Martini\*, Flavio Porta, Davide Scotti, 2021, ‘The disruptive impact of COVID-19 on air transportation: An ITS econometric analysis’, Research in Transportation Economics, Elsevier, University of Bergamo, Italy.
6. Plotly: Scientific and Interactive open-source data visualisation application integrated with python 3. <https://plotly.com/>
7. …

# A Installing packages

The project was developed using Python 3 with the help of Anaconda Jupyter Notebook. The file **aviation-islands-code.ipynb** contains the sequence of codes to run this project.

These codes in the file require the support of Python packages such as: Pandas, Numpy, Sklearn, Plotly, Matplotlib, Scipy and Seaborn. The developed code was tested on Windows 10 and Windows 11 using Python version 3.8.3. Below are the instructions to install each of the prerequisites if not already present.

Description on all the python packages used in this project and how each one of them very helpful in developing this project.

> python

>>> import pandas

>>> import numpy

>>> import sklearn

>>> import plotly

>>> import matplotlib

>>> import scipy

>>> import seaborn

>>>

By running the above commands, you can check if the python packages are already installed in the operating system, else they raise an exception. If all the requirements are satisfied, you can proceed with testing the code by following the sections 1 to 6.

Below are the descriptions of the python packages and how to install them in order to run the codes in the above-mentioned notebook.

**Pip**

It is a very important tool used for installing python packages without needing to visiting the package website and downloading from the repository links. The pip can be installed by following the procedure at [https://pip.pypa.io/en/stable/](https://pip.pypa.io/en/stable/installation/)

installation/.

**Conda**

It is an open-source package and environment manager that runs on windows and other operating systems to quickly install, run and update python packages and each of their prerequisites. Conda has its own system requirements, such as a 32/64 bit pc, windows/macOS/linux operating systems and a minimum disk space of 3GB disk space for downloading and installing it. The conda can be installed by following the procesdure at [https://docs.conda.io/projects/conda/en/](https://docs.conda.io/projects/conda/en/%0dlatest/user-guide/install/windows.html.%0d)

latest/user-guide/install/windows.html.

**NumPy**

It is a python package for the support of high-level mathematical computing in python. Some of the most useful purpose of numpy is for handling high dimensional matrices and arrays. Numpy has its own data structure which makes it easy in dealing with data for this project. You can install **numpy** as shown below.

PIP

> pip install numpy

--------------------------------------------------------

CONDA

> conda install numpy

**Pandas**

It is a very fast, powerful and easy to use python package for performing open ended data analysis and data manipulation. Pandas has its own data structure called as a data-frame which I have used a lot in this project, which make it easy to apply function on the data. You can install **pandas** as shown below.

PIP

> pip install pandas

--------------------------------------------------------

CONDA

> conda install pandas

**Matplotlib**

It is python tool for creating static and interactive visualisation plots. It also provides support for embedding plots in an external graphical user interface, it also incorporates some design features from Matrix Laboratory (MATLAB). You can install **matplotlib** as shown below.

PIP

> pip install matplotlib

--------------------------------------------------------

CONDA

> conda install -c conda-forge matplotlib

**SciPy**

It is python library for scientific and technical computing. Scipy contains modules for data optimization, linear algebra, data integration and interpolation. The data structure used is based on numpy arrays. You can install **SciPy** learn as shown below.

PIP

> pip install matplotlib

--------------------------------------------------------

CONDA

> conda install -c anaconda scipy

**sklearn**

Scikit-learn is tool in python which come with functions to perform predictive data analysis using machine learning algorithms. Scikit-learn is an easy and efficient open-source tool build with the help of **numpy**, **scipy** and **matplotlib**. You can install **Sci-kit** learn as shown below.

PIP

> pip install -U scikit-learn

--------------------------------------------------------

CONDA

> conda create -n sklearn-env -c conda-forge scikit-learn

> conda activate sklearn-env

**plotly**

Plotly is software which was developed to create online data analytic and visualisations tools. Plotly has the support to be integrated with many programming tools such as Python, R, MATLAB, Perl, Julia, Arduino and REST. You can install **plotly** learn as shown below.

PIP

> pip install plotly==5.4.0

--------------------------------------------------------

CONDA

> conda install -c plotly plotly=5.4.0

**Seaborn**

Seaborn is a data visualisation tool based on matplotlib which provides an enhanced graphical interface for creating statistical plots in python. The dependencies for Seaborn are **numpy**, **scipy**, **pandas** and **matplotlib**. You can install **seaborn** learn as shown below.

PIP

> pip install seaborn

--------------------------------------------------------

CONDA

> conda install seaborn

# B Code Listings

All the codes developed, plots and results supporting this project can be accessed on a private git repository <https://github.com/tarunm9/Aviation-Islands>.

To get access to this repository, please send an email to: [tarunm290699@gmail.com](mailto:tarunm290699@gmail.com).

--- Picture ---

# C Professional Issues

This section is to describe the professional issues faced whist developing this project and how I overcame those.

…