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 Where other published and unpublished source materials have been used, these have been acknowledged.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

**Student Name**: Tarun Malarmaran

**Date of Submission**: November 26, 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 Abstract 2

1.3 Dataset Initial Analysis 2

2 Background Research 3

2.1 Literature Survey 2

2.2 Conclusions and Hypothesis 2

3 Data Wrangling 4

3.1 Obtain and Understand Data 1

3.2 Structuring Data 1

3.3 Cleaning the Data 1

3.4 Enriching Data 1

3.5 Validate and Publish 1

4 Data Pro-processing 3

4.1 Imputation 1

4.1.1 Changing the Object types 1

4.1.2 Dealing with Missing Values 1

4.2 Dimension Reduction using PCA 2

4.2.1 2 - Components Analysis 1

4.2.2 2 - Components with Visual Loadings 1

4.2.3 PCA with variance explained 1

5 Exploratory Data Analysis 4

5.1 Hypothesis based on Corelations 1

5.2 Cluster Analysis 1

5.2.1 Hierarchical Clusters 2

5.2.2 Gaussian Mixture Model 2

6 Visualisation based on Analysis 3

6.1 GDP Trends 1

6.2 Flights Frequencies Trends 1

Conclusions 4

References 4

A Installing packages 4

B Code Listings 4

C Professional Issues 4

# Introduction

The dataset is about the 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 area and aviation statistics such as number of flights bound and via various modes of transportation.

The problem arises when we 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.

## Problem Statement

Here goes the project problem description.

## Abstract

Explaining the abstract of the project.

## Dataset Initial Analysis

**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.

# Background Research

The dataset is about the 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 area and aviation statistics such as number of flights bound and via various modes of transportation.

## Literature Survey

Literature Survey required for carrying out the project.

### Sub-subsection

This is an example of a sub-subsection. The heading should be in style ‘Heading 3’.

### Sub-subsection

This is another example of a sub-subsection

## Conclusion and Hypothesis

This is an example of a subsection.

# 

# Data Wrangling

In most of the cases in real world, the data which we get to work on are not in meaningful formats. The data are mostly very noisy. So, for this purpose we 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

Literature Survey required for carrying out the project.

### Sub-subsection

This is an example of a sub-subsection. The heading should be in style ‘Heading 3’.

### Sub-subsection

This is another example of a sub-subsection

## Structuring Data

This subsection to explain about how I structure the data in this project.

## Cleaning the Data

This subsection to explain about the various tasks done in order to clean the dataset.

## Enriching Data

This subsection to say my opinion about whether any more data enriching can potentially benefit the analysis.

## Validate and Publish

This subsection covers the report on finalising the data after wrangling before proceeding with pre-processing.

# 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 we are applying in this section includes Imputation and Dimension reduction, and it is explained as follows.

## Imputation

The process of imputation which basically means assign, that is we are 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

Stating about any potential changes to the object types as we being the imputation process.

### Dealing with Missing Values

This sub-subsection describes the whole imputation process with respect to each country of each column.

## Principal Components Analysis (PCA)

This subsection to explain about how insights very drawn by performing a Principal Component Analysis.

### 2 – Components Analysis

Stating about any potential changes to the object types as we being the imputation process.

### 2 – Components with Visual Loadings

This sub-subsection describes the PCA visualisation with the visual loadings direction in which the features are deviating.

### PCA with variance explained

This sub-subsection describes the PCA variance across different number of components.

# Exploratory Data Analysis

Description of the exploratory analysis process.

## Hypothesis and Imputation

Stating and explaining the steps followed in pre-processing and imputation.

### Changing the object types

Stating about any potential changes to the object types as we being the imputation process.

### Imputation of Missing Values

This sub-subsection describes the whole imputation process with respect to each country of each column.

## Principal Components Analysis (PCA)

This subsection to explain about how insights very drawn by performing a Principal Component Analysis.

### 2 – Components Analysis

Stating about any potential changes to the object types as we being the imputation process.

### 2 – Components with Visual Loadings

This sub-subsection describes the PCA visualisation with the visual loadings direction in which the features are deviating.

PCA with variance

# 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.

## GDP Trends

This sub-section details findings regarding the evolution of GDP for the islands across the 13 years’ timeline, it shows the growth pattern of all the island countries.

Raise in population means, the country is capable of putting more towards increasing its economy. So, to argue the correctness of this point, below is the log-log plot visualisation for population v gdpnom, with the help of trendline we can observe that as the receipt increases, the gdpnorm of a country increase. The purpose of using a log-log scale is to have a better visualisation of this correlation between receipt and gdpnom. and explaining the steps followed in pre-processing and imputation.

## Flights Frequencies Trends

This sub-section to explain about the various patterns observed from the data with respect to the trends in flight frequencies.

# 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**

The list of sources that you have used, which should have been referenced in the body of the report.

1. Plotly: Scientific and Interactive open-source data visualisation application integrated with python 3. <https://plotly.com/>

# A Installing packages

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

# B Installing packages

Details on how the code developed for this project can be accessed on the web.

# C Professional Issues

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