

Aviation Islands

Prajwal Sridhar

Submitted for the Degree of Master of Science in
Machine Learning



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

August 31, 2022

Declaration

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 Where other published and unpublished source materials have been used, these have been acknowledged.

Word Count: 13935

Student Name: Prajwal Sridhar

Date of Submission: 31/08/2022

Signature: Prajwal Sridhar

Abstract

There has been some interest in understanding the geographical features of nations that have just single airports and these nations have been termed small island states. Hence the project name 'Aviation Islands'. This project aims to understand the relationships of several single airport island nations. There have been several instances when the geographical features such as economic growth, population, wellbeing, etc, of large states or islands have been analyzed. However, less attention has been paid to the small islands and their geographical features. Our main goal is to first collect the data, perform exploratory data analysis of the dataset, visualize the data with the help of different visualization techniques and also find out if there are any similarities and interrelationships among various features. Finally, draw possible hypotheses from the data, especially if small island nations have a common pattern of growth. Apart from just the demographics, we also try to understand the majority of such nations belong to which continent. A detailed description of the dataset and all the plots explaining the relationships have also been provided. The data collected for each nation has been collected manually and the tool used for visualization is ggplot2 in R. The entire dataset has been collected from open-source websites. A detailed description of the missing values and how to deal with them has also been explained. One can also extend this project by developing a user interface and API to visualize and analyze this data set. This project can also be integrated with other social data sets for analysis.

Table of Contents

s1 Introduction	1
1.1 Importance of Aviation in Small Islands.....	2
2 Data Description	3
2.1 Missing Data	5
2.1.1 Missing Data Handling Methods	6
2.1.2 Method of Imputation.....	7
2.1.3 Check for Missing values	10
2.1.4 What are Outliers?.....	13
2.1.5 Recommendations for Detecting and Handling the Outliers?	14
3 Analyzing the Data	20
3.1 What are the States and Number of States in each Continent?	21
3.2 What is the population and how is the population growth in each state?	23
3.3 What is the correlation between the numerical variables in our data?	25
3.3.1 What is the ANOVA test and How do we do it in R?	26
3.4 How is the population density for each state?	27
3.4.1 Is the mean distribution of the population density of all the states the same?	27
3.5 How are infant mortality and life expectancy related?	28
3.5.1 Is the mean distribution of life expectancy of all the continents the same?	29
3.5.2 Which are the states that have both low and high life expectancy and which continent do they belong to?	30
3.5.3 Is the mean distribution of the life expectancy in the 3 continents the same?	31
3.5.4 What is the distribution of the infant mortality rate per continent?	31
3.5.5 Is the mean distribution of the infant mortality rate same for all the continents?	33
3.6 How is Life Expectancy related to Educational Attainment?	34
3.7 How are educational attainment related to other variables?	36
3.8 How are the Aircraft Movements related to other variables?	40

3.9	Which state has the highest economy and how is economy related to other variables?.....	44
3.9.1	Which is the state that has the least economic and which continent does it belong to?.....	46
3.9.2	How important are the aircraft movements for these small islands?	47
3.10	What is the official language of each state?	48
3.10.1	What is the language spoken in the state that has the highest population?	49
4	Conclusion	50
4.1	Challenges Faced	51
4.2	Steps to run the R file	51
5	References	53

1 Introduction

The number of sovereign entities that are regarded as minor has increased throughout the processes of decolonization and nationalism that have characterized this century. The term "little islands" is used to describe many of these (Hindmarsh J., 1996, p.36.) These islands only have one airport, thus for the purposes of this essay, we will refer to them as being little. There has also been curiosity about the demography of these islands due to their rapid expansion (Hindmarsh J., 1996, p.36.).

According to (Briguglio, 1995), because of their diminutive size, isolation, and susceptibility to natural calamities, many small island republics have disadvantages because of their small size, insularity, remoteness, and proneness to natural disasters. This may have an impact on all these states' economic health too, impacting their GNP per capita (Briguglio, 1995, pp.1615-1632).

The fundamental query, however, is whether small states are adversely affected by their size. There are numerous justifications for why they do (Easterly & Kraay, 2000). As stated in the article by (Easterly & Kraay, 2000) small states are those with small populations, which hinders them from engaging in a variety of activities, which has an impact on their development and economic standing. Many small states suffer as a result of their poor locations because they may be found in areas that are more vulnerable to hurricanes and other natural catastrophes (Easterly & Kraay, 2000).

When we consider the economic situation of such small countries with few activities, tourism becomes a very important factor that can help them in their economy and in a variety of other ways. However, tourism can be improved when connectivity to these countries improves, whether by land, sea, or air and in order to improve any mode of transportation, transportation capital must also be improved. Making regions more accessible and appealing has a significant impact on economic growth. Air travel is the most common of the three modes of transportation. This is due to the fact that it is faster, more efficient, and more dependable than the other two options.

The paper is organized as follows: in the introduction, information regarding small island states and why they are called small is provided. A number of Caribbean islands are facing a wide range of difficulties. Following that, we would focus more on air transport. Subsection 1.1 will discuss the significance of aviation in small island states. Section 2 will provide a detailed description of the data gathered, as well as a list of states with only one airport and a description of the features or attributes gathered for each state. Section 2.2 will go over the various types of missing data, what data is missing in our data, and how to deal with it. In the same subsection, a general understanding of the various approaches to dealing with missing data would be provided. This section will go over the concept of imputation. Following that, in Section 3, we will look at data visualization. Finally, we'll ask some data-related questions to see if there are any patterns or similarities. The fourth section of the paper would be devoted to data statistical analysis. This would include a discussion of the research methods used as well as the analysis results.

1.1 Importance of Aviation in Small Islands

All of these small island states face similar challenges, such as small but growing populations, limited supplies, and reliance on international trade. Other challenges that states face when working to implement sustainable development goals include their vulnerability to the effects of climate and natural disasters (Aviation, 2018). Furthermore, when faced with massive disruptions, such as natural disasters, aviation aided as the most effective means of delivering humanitarian aid (Clark, 2018).

Without a doubt, aviation plays a significant role in a country's economy. Aviation connects the world. More importantly, air transport improves connectivity for island nations and expands job opportunities. According to the blog (Gill, 2014), Modern air transport supports over 58 million jobs and \$2.2 trillion in global GDP. While aviation aids in the development of business, trade, and tourism services, it also serves a much more important role in some areas, namely connectivity (Gill, 2014).

Tourism is a significant source of foreign revenue for islands because the majority of visitors arrive by air. It can also be argued that tourism revenue is more sustainable than other sources of income. With the expansion of tourism opportunities comes an increase in passengers and the jobs they support (Gill, 2014). Air traffic in small island states is expected to increase at a 5.4 percent annual rate over the next 20 years (Gill, 2014). The tourist connection helps to connect residents in social and business ways. It also helps to promote tourism in the area (Gill, 2014).

Since we've discussed how dependent such islands are on the aviation industry, you're probably wondering what happened to these islands during the pandemic, when the global aviation sector suffered a major bump in the road. But from this article (Macola, 2020) it is clear that all of these small island states suffered heavily as the number of visitors decreased for about two years (Macola, 2020). Now, most of the states have slowly started air operations and are recovering whilst coming back to normal.

While aviation has many advantages, it also has some drawbacks. It contributes about 2% of global CO₂ emissions. However, the industry has been working to reduce these emissions (Gill, 2014).

The preceding discussion shows that air transport or the aviation industry makes a significant difference, particularly for these small island states. Even if it is only one airport, it makes a significant contribution to the country's development.

2 Data Description

The dataset was compiled by hand. There are 33 island nations from the data that has been collected. This dataset contains ten features, each of which is described in detail below. The characteristics are as follows: Landmass, Economy, Population, Life Expectancy, Infant Mortality, Educational Attainment, Aircraft Movements, Official Language, and Continent are some factors to consider.

Everyone may be wondering why these features were chosen in particular. These characteristics are the most important aspects of a country. These characteristics comprise the country's geography and demographics. Except for aircraft movements and educational attainment, all data has been gathered from Wikipedia (Wikipedia, 2022). Aircraft Movement data has been collected from (The World Bank Data Group, 2021) whereas Educational attainment data has been collected from (The World Bank Data Group, 2021).

The description of each feature is as follows:

- Landmass: Everyone understands what landmass is; it is simply the size of the state in square kilometers.
- Economic: A state's economy includes all activities such as production, consumption, and trade of goods and services in a given area. We used the gross national product per capita in US dollars (GNP PPP) for each state in this dataset.
- Population: We considered the population over the last 50 years, namely the population in 1972 and the population today.
- Life Expectancy: I have considered the total life expectancy of the state. However, one can also collect the life expectancy for males and females separately according to gender. Life expectancy is nothing but the average period that a person may expect to live.
- Infant Mortality: The infant mortality rate is defined as the number of children that die under one year of age per 1000 live births in a given year. Just like life expectancy, the data for both males and females can be collected separately. However, I have collected the total infant mortality rate of the states.
- Educational Attainment: I considered the percentage of university graduates (at least a bachelor's degree) or the equivalent of the total population aged 25 and up. Educational attainment is related to a country's population's skills and competencies (The World Bank Data Group, 2021). It also reflects the educational system's capacity at the corresponding level of education.
- Aircraft Movements Data: Airports and air traffic control systems are critical to household and government activities. This data

represents the total scheduled traffic carried by a country's registered air carriers. Domestic and international aircraft passengers registered in the country are among those carried (The World Bank Data Group, 2021).

- Official Language: Some states have multiple official languages. However, for this dataset, I considered the language spoken by the majority of the population in the respective states to be their official language.
- Continent: This feature, as the name implies, refers to the continent to which the states belong. This would also provide us with a clear idea of how the features on the continent interact with one another.

The dataset table is shown in Figure 1. Some values in this dataset are missing because they were unable to be gathered from any source and are, therefore, absent. Information about missing values and possible solutions are described in the following section.

States	Economic	Landmass	Pop in 197	Pop in 202	Life Expect	Infant Mor	Education	Aircraft M	Official La	Continent
Aruba	2.86E+09	180	59800	107657	76.29	12.698	N/A	274280	Dutch	South Ame
Barbados	4.16E+09	430	241000	288017	79.19	11.4	1.1	N/A	English	North Ame
Tuvalu	70000000	26	5786	12080	66.16	30.8	N/A	N/A	Tuvaluan	Oceania
Singapore	4.92E+11	728.6	2152000	5453600	83.9	0.8	33	7884373	English	Asia
Saint Hele	47768924	121.7	5919	6127	78.6	16.98	N/A	N/A	English	Africa
Kuwait	2.48E+11	17818	870000	3068155	79.13	7.43	11.1	1823594	Arabic	Asia
Guyana	1.21E+10	214969	721949	794128	67	25.12	N/A	17990	English	South Ame
Niue	24938000	261.5	4724	1648	73.6	0	N/A	N/A	Niue	Oceania
Malta	2.04E+10	316	302600	525285	82.6	6.1	21.8	549319	Maltese	Europe
Mayotte	N/A	374	40200	286145	60.6	65.98	N/A	N/A	French	Africa
Macau	2.52E+10	118	243900	666932	84.55	4.35	N/A	544411	Chinese	Asia
Cote d'Ivo	1.40E+11	322462	5416000	28713423	62.26	55.67	N/A	322841	French	Africa
Dominica	7.73E+08	750	72000	74027	77.4	0	5	N/A	English	North Ame
Jersey	672089	119.5	69329	103267	78.48	3.94	44	679000	English	Europe
Uganda	1.03E+11	241038	9930000	46205893	68.96	30.45	1.7	6159	English	Africa
Mauritius	2.83E+10	2040	828000	1308222	74.86	12.08	N/A	407291	English	Africa
Samoa	1.28E+09	2831	146647	206179	75.19	18	3.9	11957	Samoan	Oceania
Nauru	2.24E+11	21	6798	10960	59.7	23.8	N/A	9356	Nauruan	Oceania
Isle of Ma	86482	572	57166	86381	80.6	3.9	N/A	N/A	English	Europe
Saint Lucia	2.24E+09	617	104192	167122	78.95	11.99	N/A	N/A	English	North Ame
Gibraltar	2.04E+09	6.8	29000	33671	79.93	4.39	N/A	N/A	English	Europe
Gambia	5.41E+09	11300	492424	2173999	65.1	60.2	2	109542	English	Africa
Faroe Islan	54833	1393	48300	48865	79.85	5.4	N/A	N/A	Faroese	Europe
Bhutan	8.09E+09	38394	316822	867775	71.31	27.04	10.2	48825	Tsangla	Asia
Bermuda	5525000	53.2	52976	61819	81.87	0	31.1	N/A	English	North Ame
Qatar	2.54E+11	11571	130505	2508182	79.81	6.62	19	10640789	Arabic	Asia
Luxembou	4.52E+10	2856	347000	650364	82.98	3.25	N/A	753114	French	Europe
Papua Nev	3.8E+10	462840	2918000	9593498	69.43	33.59	N/A	1501149	Tok Pisin	Oceania
The Solom	1.84E+09	28896	172550	702694	76.2	14.41	N/A	138661	English	Oceania
Suriname	8.8E+09	163820	372000	632638	72.42	30.25	9.2	89027	Dutch	South Ame
Mongolia	3.59E+10	1564000	1301400	3296866	68.63	13.4	23.7	143860	Mongolian	Asia
Hong Kong	4.68E+11	1114	3936630	7276588	83.61	2.55	N/A	5878548	Chinese	Asia
Albania	3.85E+10	28748	2243126	3095344	79.47	10.82	12.9	124714	Albanian	Europe

Figure 1: My Dataset that has been collected before cleaning

2.1 Missing Data

Missing data are a typical issue with all types of data, but they are particularly prevalent in survey-based research investigations. There are various approaches to addressing missing data (Heijden, et al., 2006). The results of analyses based on such data can be significantly impacted by how these missing values are handled.

In a large-scale investigation, the incidence of missing data on one or more variables that are significant for the sample size has turned into the norm rather than the exception (Cheema, 2014). A study with a lot of variables and only a few missing values might significantly reduce the overall effective sample size (Cheema, 2014). For instance, listwise deletion can reduce the effective sample size to just $0.9^{10} \times 400 = 140$ for a data set with 400 observations and 10 variables and 10% of the data missing from each variable individually (Cheema, 2014). Just to note, the Listwise deletion that I have mentioned here is a method to deal with the missing data and will be explained further in the discussion.

You must be asking how the data could be missing at this point. The lack of data in surveys can be attributed to a number of factors. Sometimes it may be as a result of respondents purposefully ignoring some questions. Other times, a respondent can really forget to respond to a particular question. Another cause can be that the respondent was asked a question that was incorrect or that they were unable to provide an adequate response. The values in my example are missing since I was unable to locate data that was appropriate for the variable when I acquired the data from open-source websites. We will examine missing data mechanisms such as MCAR, MAR, and MNAR before discussing how missing data might be handled (Cheema, 2014).

- MCAR:- Missing entirely at random is known as MCAR. The data are said to be fully missing at random if the probability of their absence is the same in all cases (S., 2018). In other words, MCAR refers to data that is randomly completely missing without any discernible pattern. This indicates that the reasons why some data are absent are unimportant to the data (S., 2018). Many of the blatant information losses may go unaccounted for (S., 2018). Additionally, data loss due to unlucky circumstances may occur (S., 2018). This strategy is acceptable provided the target population is still accurately represented despite the sample size being reduced due to missing data (Cheema, 2014). For the available data, MCAR frequently fails (S., 2018). Let's take an example where a variable Y has no relationship to any of the predictors in the dataset and missing data on Y does not have any relationship to its value (Cheema, 2014).
- MAR:- Missing at random is abbreviated as MAR. When a variable's risk of missing data is independent of its own value but may be connected to the values of other variables in the data set, the data are MAR (Cheema, 2014). In other words, when data are

missing exclusively within data sub-samples and not at random. Compared to MCAR, MAR is a substantially larger class. Overall, MAR is more realistic and inclusive (S., 2018). For instance, when X is controlled for, the MAR theory predicts that the missing data for Y may rely on X but not on Y (Cheema, 2014). My situation is an example of MAR because data is only missing within data sub-samples and not randomly.

- **NMAR:-** Not Missing At Random is referred to as NMAR. When the probability of missing data for a variable depends on the value of the variable itself, this is the case. Another way to put it is that NMAR can be used when there is a clear pattern in the way data is missing. Salary is one instance of NMAR (Cheema, 2014). As opposed to individuals with lower wages, people with high salaries have a tendency to withhold their salaries. As a result, the likelihood of a missing pay value depends on the salary itself (Cheema, 2014).

There is no need to model the missing data method as part of the estimating process when the data is either MCAR or MAR, which means that any analysis approach can be applied to the resulting data set as if it were complete (Cheema, 2014). However, the missing data technique must be carefully represented as part of the estimating process when the data is NMAR (Cheema, 2014).

The question of what happens if we use these tactics wrongly may now be raised in light of the explanation of these strategies. The missing data process is not accurately modelled when NMAR data are wrongly handled as MCAR or MAR, and parameter estimates will not be accurate (Cheema, 2014). Similar to how erroneous management of MCAR and MAR data as NMAR results in the researcher adding needless complexity to the handling of missing data (Cheema, 2014). Last but not least, using MAR data as MCAR indicates that the researcher is oversimplifying how to handle missing data (Cheema, 2014). We will then examine how to deal with the missing data.

2.1.1 Missing Data Handling Methods

- **Listwise Deletion**
This is the most common and simplest method for dealing with missing data. As defined by (Cheema, 2014) this method simply discards observations that have missing values for one or more variables of interest (Cheema, 2014). For this reason, it is also known as the entire case method (Cheema, 2014). Convenience is one advantage of listwise deletion (S., 2018). If the data is MCAR, it creates standard errors and significance levels that are appropriate for the data subset (S., 2018). However, listwise deletion has the disadvantage of being inefficient, especially when the number of variables is large (S., 2018). If the data is not MCAR, this technique

can drastically distort mean and correlation estimates (S., 2018). In some cases, listwise deletion can produce more accurate predictions than even the most complex procedures (S., 2018).

- **Pairwise Deletion**

The distinction between it and listwise deletion is that only cases with missing data on variables involved in a statistical technique are removed (Cheema, 2014). The accessible case method (Cheema, 2014) is another name for it. On all observed data, the approach estimates the means and covariances (S., 2018). Both listwise and pairwise methods provide the same result. This technique yields a consistent estimate of the means and correlation for MCAR (S., 2018). This strategy has drawbacks as well. First and foremost, if the data are not MCAR, the estimations may be biased (S., 2018). Furthermore, pairwise deletion necessitates numerical data with a roughly normal distribution, when in fact, we frequently have variables of mixed types (S., 2018). If the correlations between the variables are modest and the MCAR assumption is plausible, so this strategy is recommended. Otherwise, this approach is not advised (S., 2018).

These are the two approaches for dealing with data. Imputation is also a method that could be used to deal with missing values. The following section will explain what imputation is and how it might be useful. The many types of imputation will be discussed in the same section.

2.1.2 Method of Imputation

This section will begin by introducing imputation and the essential idea underlying this strategy. According to (Heijden, et al., 2006), in the conventional statistical paradigm, the findings drawn in any study should not be dependent on the sample. If the study is repeated with a different sample, identical results must be obtained. The results are not dependent on the sample's specified set of individuals. This means that any subject in a randomly selected sample can be replaced by a new subject drawn at random from the same source population as the original subject without affecting the conclusions. This is known as the replacement principle, and imputation techniques are based on it as well (Heijden, et al., 2006). There are different imputation methods that are used and can be explained as follows :

- **Mean Imputation:-**

This is a straightforward procedure that involves replacing missing data on a variable with the mean of the missing data. Because the effect of other variables is not partial out of the mean used to replace missing data, it is also known as marginal mean imputation. It is one of many strategies that rely on substituting missing data on a

variable with a measure of that variable's central tendency. This method is also known as median imputation or mode imputation, depending on how the central tendency is measured. This strategy reduced the standard error of the mean, which raises the possibility of rejecting the null hypothesis when it should not be rejected. Just to clarify, a comprehensive description of hypothesis testing will be provided in the latter section of this paper, as will the definition of the null hypothesis. There is a concern with minimizing the standard error for imputation based on other metrics of central tendency (Cheema, 2014).

- **Regression Imputation:-**
This strategy entails regressing the missing data variable on all other factors in the data set using instances with complete information for those variables. As a result, given the values of other variables, this approach allows for the computation of anticipated values of the variable with missing data. As a result, this approach is also known as conditional mean imputation. This technique, however, does not properly describe the natural variance in missing data, resulting in skewed standard errors of parameter estimations (Cheema, 2014).
- **Expectation-Maximization Imputation:-**
This is a mathematical approach for determining one or more statistical models that maximize the observed probability distribution given observed data. This approach consists of two phases, the first of which assigns initial values to the missing data. The expectations created with those starting values are maximized in the second stage. This technique is continued until the imputed values converge according to the requirements. This approach yields parameter estimates with unbiased standard errors (Cheema, 2014).
- **Multiple Imputation:-**
As explained by (S, et al., 2015) if correct estimates and p-values have to be obtained then we have to take into account the imprecision caused by the fact that the distribution of the variables with missing values is estimated. This can be done by creating not a single imputed dataset, but multiple imputed datasets (S, et al., 2015).
A multiple Imputation is a sophisticated approach for simulating natural variance in missing data by repeatedly imputing missing data and therefore obtaining numerous complete data sets. Averaging is used to aggregate the sets of estimates provided by these full sets into a single set of estimates. Because this technique mimics the inherent variance in missing data, the standard errors are unbiased (Cheema, 2014).

- **Hot Deck Imputation:-**

It is widely used in social science research, however, it is less established than other missing data imputation approaches. As defined by (Cheema, 2014), this strategy includes matching the missing data on a variable with other cases in the dataset that contain complete information on many other essential variables. There are other versions of this procedure, but the one that allows for natural variability in missing data entails picking a pool of all instances, known as the donor pool, that is identical to the case of missing data, and then selecting one at random from that pool. This randomly selected data is then utilized to replace the missing value on the case with partial data. Another version of this strategy involves replacing the nearest donor neighbor instead of picking one donor from a pool. This technique overlooks missing data variability. Hot deck imputation would require huge sample sizes in studies with a large number of variables in order for cases to be matched. Now then how does one pick the donor that is most appropriate? To find an eligible donor (Cheema, 2014), the receiver is matched with comparable instances based on all potential criteria, not only those mentioned in the analytic technique. This approach has two further variations: weighted sequential hot deck and weighted random hot deck. A weighted sequential hot deck is used to prevent the problem of matching the same donor with a high number of receivers (Cheema, 2014). This is accomplished by limiting the number of times a donor may be picked (Cheema, 2014). Unlike the weighted random hot deck, which does not limit the number of times a donor may be chosen, the donors are chosen at random from the donor pool. This approach can also be used with other imputation methods, such as the multiple imputations described above, in which the results of various imputed data sets are merged to provide average parameter estimations (Cheema, 2014).

- **Zero Imputation:-**

As defined by (Cheema, 2014) this technique, as the name implies, simply replaces missing values on a variable with zeros. This method's simplicity is outweighed by its limited use. This strategy, however, makes sense only in particular situations, such as when dealing with missing accomplishment scores, when a missing value can be fairly expected to occur because the responder did not know the proper response (Cheema, 2014). However, when additional factors contribute to the incidence of missing data, this technique may provide biased parameter estimations (Cheema, 2014).

- **Single Random Imputation:-**
According to (Cheema, 2014) this approach might be thought of as a hybrid of regression and multiple imputation. For scenarios with complete information, it entails regressing the variable with missing values on all other variables, supplementing the resultant projected values with random drawings, and then utilizing the augmented values to replace the missing data. Because the dataset after imputation is deemed full, the resultant standard errors are typically underestimated by their population equivalents (Cheema, 2014).

From all the discussions we have seen below that our data contains outliers and in section 3 we saw that our data in some of the features are skewed to either side, hence considering these 2 factors in my mind I decided to impute the missing values using the median imputation as it is more preferred when the data contains outlier or skewness.

2.1.3 Check for Missing values

After all of the previous explanations concerning missing data, missing data methods, and missing data management, one needs to know how to check for missing values, which is what this part will cover.

Once you get a dataset, the first step is to determine which columns contain missing values and how many of them there are. I read the dataset using pandas. `IsNull()` is a function that, when used with the `sum()` function, returns the total number of missing values in each column (R, 2021). A snapshot of the number of missing values in each column of my dataset is shown below.

```
missing_values = data.isnull().sum()
```

```
missing_values
```

States	0
Economic	1
Landmass	0
Pop in 1972	0
Pop in 2022	0
Life Expectancy	0
Infant Mortality	0
Educational Attainment	18
Aircraft Movements	11
Official Language	0
Continent	0

dtype: int64

Figure 2: Depicts the number of missing values in each column

Although we know the total number of missing values in each column, we also need to know the percentage of missing data (R, 2021). The picture below depicts the percentage of missing values in my dataset.

```
miss_value_percent = 100*data.isnull().sum() / len(data)
```

```
miss_value_percent
States                0.000000
Economic              3.030303
Landmass              0.000000
Pop in 1972           0.000000
Pop in 2022           0.000000
Life Expectancy       0.000000
Infant Mortality      0.000000
Educational Attainment 54.545455
Aircraft Movements    33.333333
Official Language     0.000000
Continent             0.000000
dtype: float64
```

Figure 3: Percentage of missing values to the total values

According to the accompanying graphic, 54 percent of the Educational Attainment column is missing, which is a considerable number. Aircraft Movements has around 34% of data missing, whereas Economic has just 3%. This is the quantitative analysis I have for missing data.

Until now we have seen the quantitative analysis of missing data, now let's look at the qualitative analysis. There are different maps, charts, and matrices, (R, 2021) using which we can visualize, analyze missing data and also understand the distribution of missing data. By analyzing how they are distributed, we can conclude what category they fall into which is either MCAR, MAR, or NMAR (R, 2021). We can also find the correlation between the columns containing the missing values and the target column (R, 2021).

Let us start by making a bar chart for non-null values. We can do this by using the bar() function (R, 2021).

<AxesSubplot:>

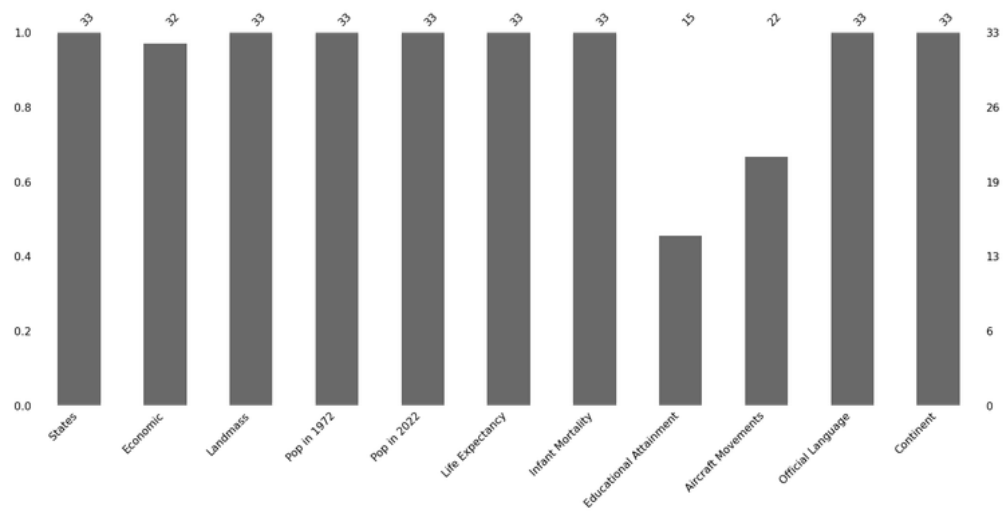


Figure 4: Bar-Graph representing the total number of non-null values

Now next let us plot the matrix visualization. Matrix visualization helps us to know how the missing data is distributed through the data, we can also know if they are localized or evenly spread or if there is any pattern and there could be many such questions (R, 2021).

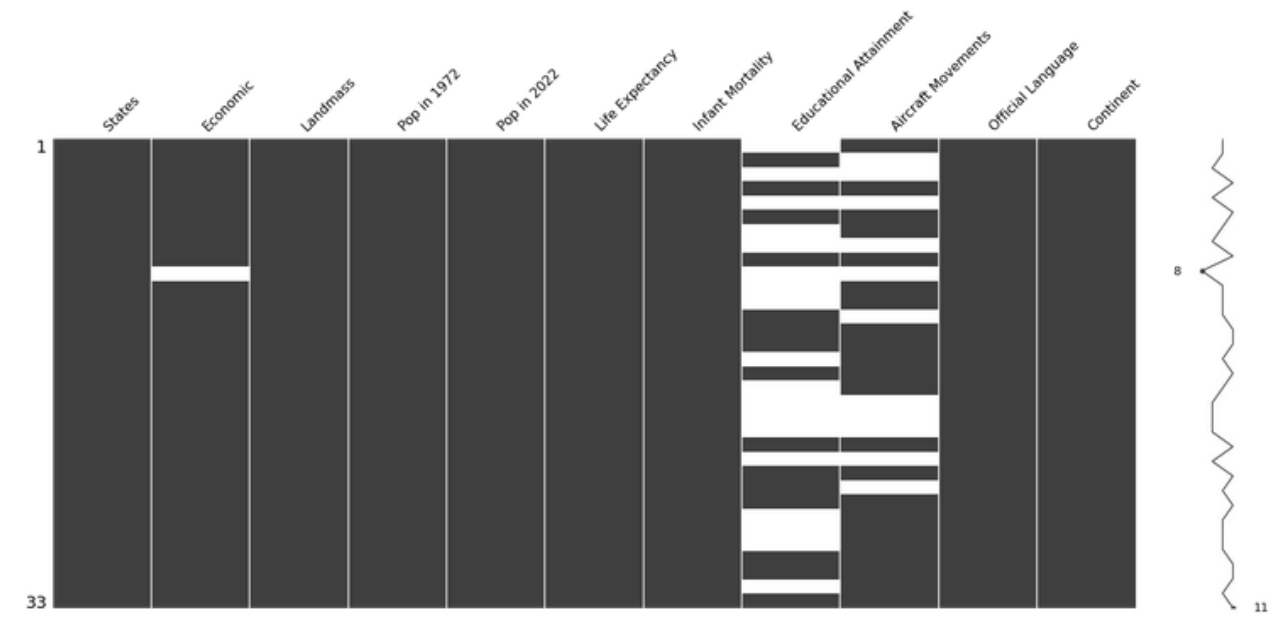


Figure 5: Matrix Visualization to understand the distribution of missing data

In the above plot, you can see blank lines for each of the missing data (R, 2021). We can notice that the 'Economic' columns have only one random missing data, which follows no pattern. This was probably lost during data collection and hence can be termed as missing completely at random (R, 2021). However, the 'Educational Attainment' and 'Aircraft Movements' columns could possibly be MAR (R, 2021). But we would like to ensure the same and for, that we would at the correlation between them. To have a look at the correlation we use the heatmap. Below is an image for the same.

<AxesSubplot:>

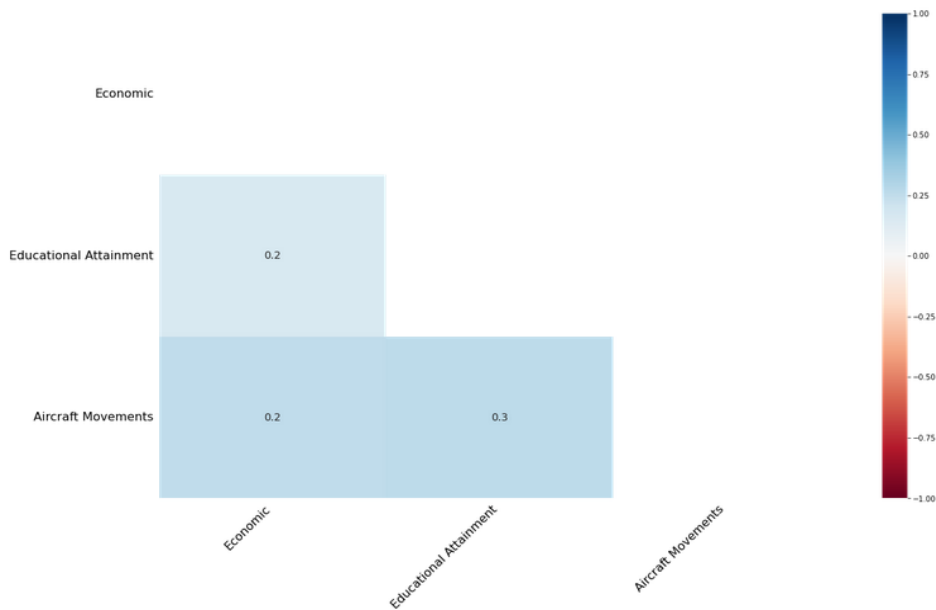


Figure 6: Correlation between the missing data columns

From figure 6 it is clear that there is no correlation between the missing data on educational attainment and aircraft movements data. So the missing data of these columns can be classified as MAR or missing at random.

2.1.4 What are Outliers?

Before we begin talking about what outliers are, I would like to talk about why I have raised this question. So this question has been raised in order to check for values that are on either extreme and we will know which columns these outliers are present in and the reason for these outliers could be anything which will be discussed further in this section.

To begin, what exactly is an outlier? Outliers are values in a dataset that are out of the ordinary. Outliers can skew statistical analysis and perhaps contradict assumptions (Frost, 2021). Outliers must be handled with caution since all analysts

will be confronted with them and required to make conclusions about them. Given the potential for difficulties, everyone may believe it is preferable to delete them from the data. However, this should not always be the case. Only for certain reasons should outliers be removed (Frost, 2021). As a result, it is also critical to understand how outliers might emerge and if they will occur again as a typical part of the process or research area. Outliers also increase the variability in the data, lowering statistical power. Another factor to consider is that eliminating the outliers may lead the results to become statistically significant (Frost, 2021). Outliers are caused by the following factors:

- **Data Entry and Measurement Errors:-** Errors may occur during measurement and data input. Typos during data entry can result in strange values (Frost, 2021). These sorts of problems are very simple to recognize and identify. If you establish that an outlier value represents a mistake, you should repair it as soon as feasible (Frost, 2021). This can range from correcting a typo to recalculating the worth of an object or person. If neither of these options is feasible, it is preferable to remove the data point since you know it contains an inaccurate number (Frost, 2021).
- **Sampling Problems:-** Samples are used in inferential statistics to draw inferences about a given population. A population should be well defined before a random sample is drawn from it (Frost, 2021). However, the research may inadvertently get an object or a person who is not from the intended group. There are several possibilities. One explanation may be that you accidentally acquire an item that is not in your target group and has some unexpected qualities (Frost, 2021).
- **Natural Variation:-** Every data distribution has a range of values. Extreme values can exist, but they are less likely (Frost, 2021). It is rather evident that you could acquire odd values if your sample size is large enough (Frost, 2021). However, there is a potential that extreme values will appear in smaller datasets as well. As a result, the process you are examining may naturally yield values. However, nothing is incorrect with these data points. They are out of the ordinary, but they are a normal element of data distribution (Frost, 2021).

2.1.5 Recommendations for Detecting and Handling the Outliers?

Following an understanding of what an outlier is, we will now focus on recommendations for dealing with outliers. There are different criteria for identifying outliers, including visual inspection and analytic procedures (PhD & MD, 2016).

A box plot is one such way, and it is the best method that analysts often employ to discover outliers. Any figure that is less than $Q1 - 1.5 \times IQR$ (the difference between the upper and lower quartiles) or greater than $Q3 + 1.5 \times IQR$ has been deemed an outlier (PhD & MD, 2016).

There are several statistical strategies for spotting outliers. They are classified as parametric methods or model-free methods. The primary idea underlying parameter techniques is to compute the parameters assuming that all of the data points have the same statistical distribution. Outliers are values that have a lower likelihood of occurring from such a distribution (PhD & MD, 2016).

Following on from the discussion of analytic procedures, there are two types of analytic procedures: univariate methods and multivariate methods. One of the basic approaches for finding outliers in single samples is Grubbs' test (PhD & MD, 2016) (univariate methods). Grubbs' test, in general, detects just one outlier at a time. The method is iterative, and it ends when no more outliers can be found. Grubbs' test, on the other hand, would not work effectively for samples with fewer than or equal to 6 observations. The chi-squared test and generalized extreme studentized deviation test are two further univariate methods for detecting outliers (PhD & MD, 2016).

In terms of multivariate approaches, regression analysis is a technique. In this strategy, outliers are abnormal data in the response variable, and leverage points are abnormal observations in the predictors (PhD & MD, 2016). Although a poor leverage point might significantly affect the regression slope effect estimate (PhD & MD, 2016). Other multivariate outlier identification approaches include distance metrics to determine whether an observation is far from the center of the data distribution (PhD & MD, 2016). Mean and variance-covariance were utilized to find outliers in both univariate and multivariate approaches. However, because mean and variance are susceptible to outliers, adopting robust estimates of the distribution parameters can increase outlier identification performance (PhD & MD, 2016).

Below are some of the ways in which you could handle the outliers:

- **Keep Outliers:-** Many academics recommend including all outliers in data analysis. Outliers, for example, may be genuine extreme observations due to random variability and reflect an inherent aspect of random sampling. If this is the case, they should be stored and processed just like any other observation. When there are outliers due to data skewness, various transformations, such as the log transform, can be employed to handle the outliers. There are also robust approaches that could be utilized, where robust means less sensitivity to outliers (PhD & MD, 2016).
- **Remove Outliers:-** Removing outliers, even if they are justified, is an entirely different point of view. Outlier elimination proponents say that effective statistical analysis should focus on modeling the majority of a population. However, data with outliers removed frequently show less volatility and better match data analysis assumptions. Others may argue that eliminating data points based

on statistical analysis without a specific cause is not a good rationale. Removing observable outliers brings new issues (PhD & MD, 2016).

- Using different analysis methods:- You could also utilize statistical tests that are less affected by the presence of outliers. Using the median to compare datasets, for example, could be an option to investigate. Also, there is one more way in which we can try editing the outliers by correcting the wrong data records.

The box plots shown below were used to locate outliers and to determine which column of my dataset contained outliers.

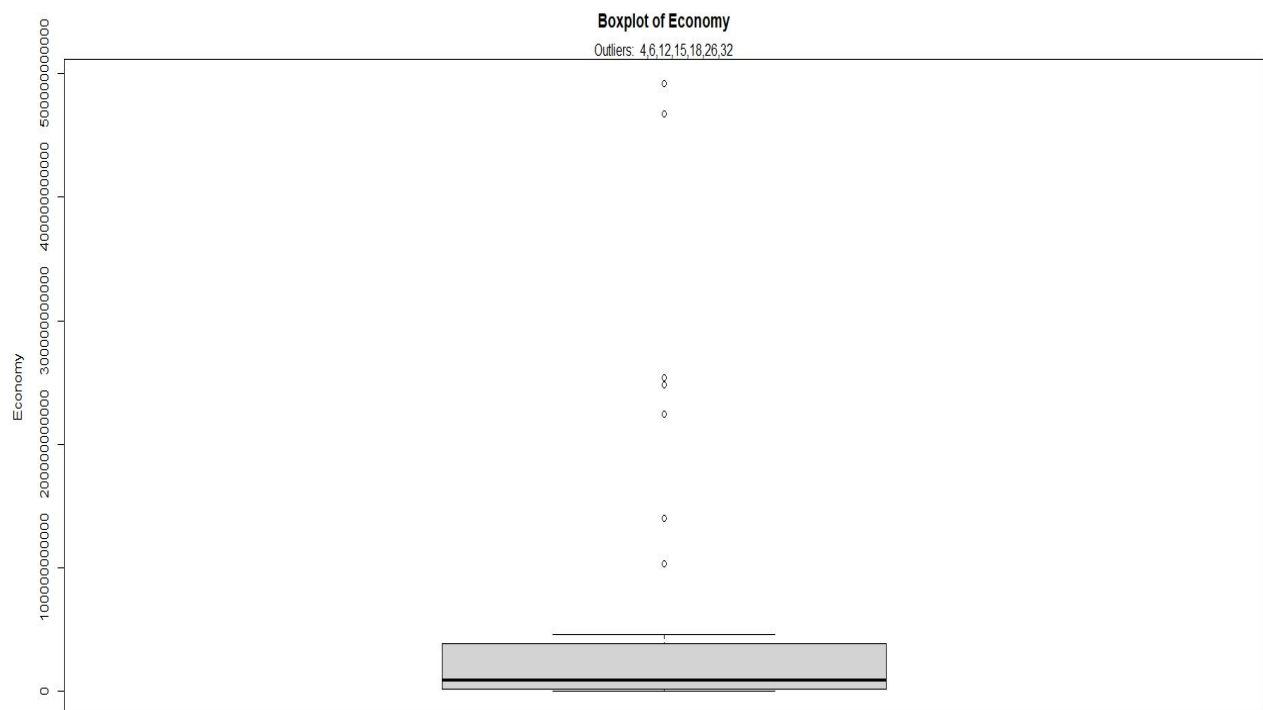


Figure 7

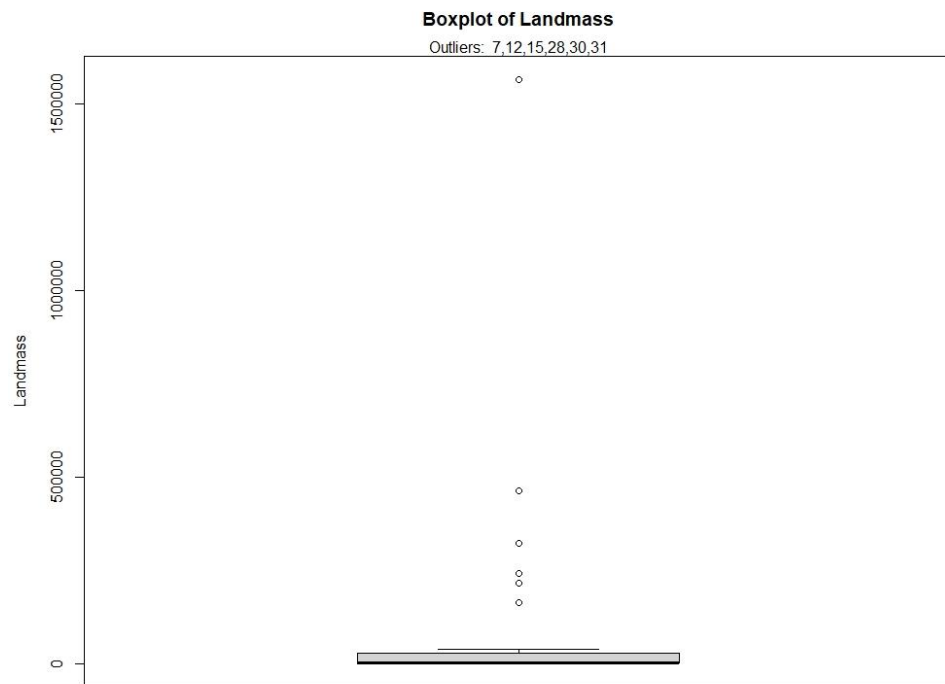


Figure 8

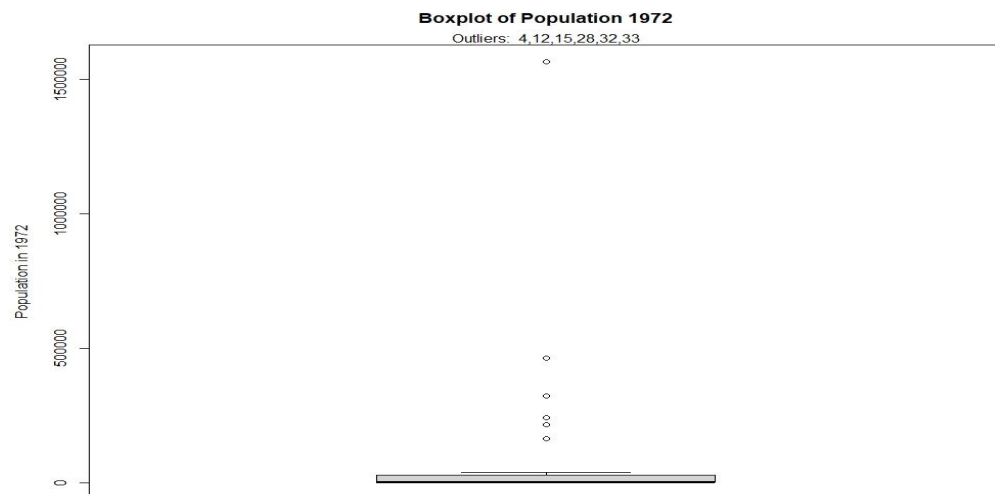


Figure 9

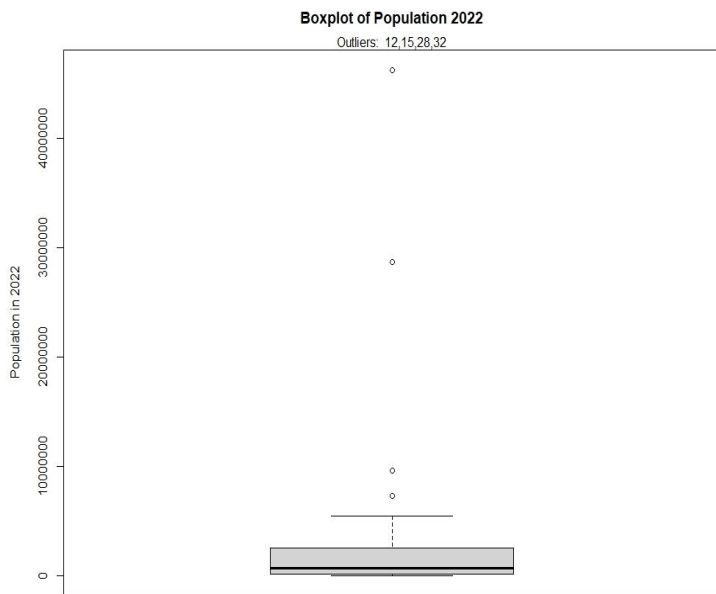


Figure 10

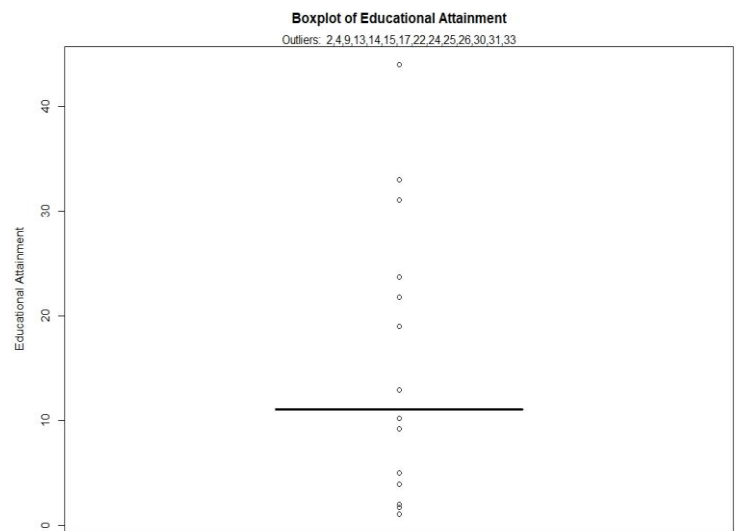


Figure 11

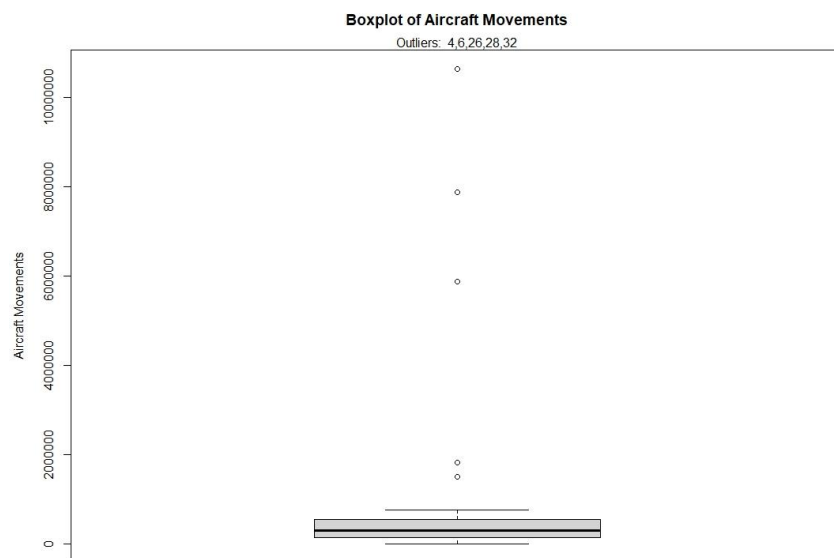


Figure 12

I used a box-plot to detect outliers in numerical columns of the dataset in all of the examples above. The box plot depicts the observations that are regarded outliers as points. Based on this criterion, we can see that Economic has roughly 7 outliers, whereas landmass and population in 1972 each have 6 observations. Population in 2022 and Aircraft Movement both have four observations. There are just two observations for the infant mortality rate. We can also see that there are no points for Life Expectancy and Educational Attainment, indicating that there are no outliers in those columns. Now that we've identified the outliers and know which column includes them, we may deal with them using any of the methods described in section 2.1.5. However, because my dataset is tiny and these outliers may be due to data entry, and those observations are a legitimate part of the population that I am investigating, I would keep those points in the dataset as they are. This is not to say that I would leave the missing values alone; I would still deal with them using one of the approaches discussed in sections 2.1, 2.1.1, and 2.1.2.

In all the above box-plots, we can there are indexes below the title of plot and those indexed indicate the states that contain those outliers.

3 Analyzing the Data

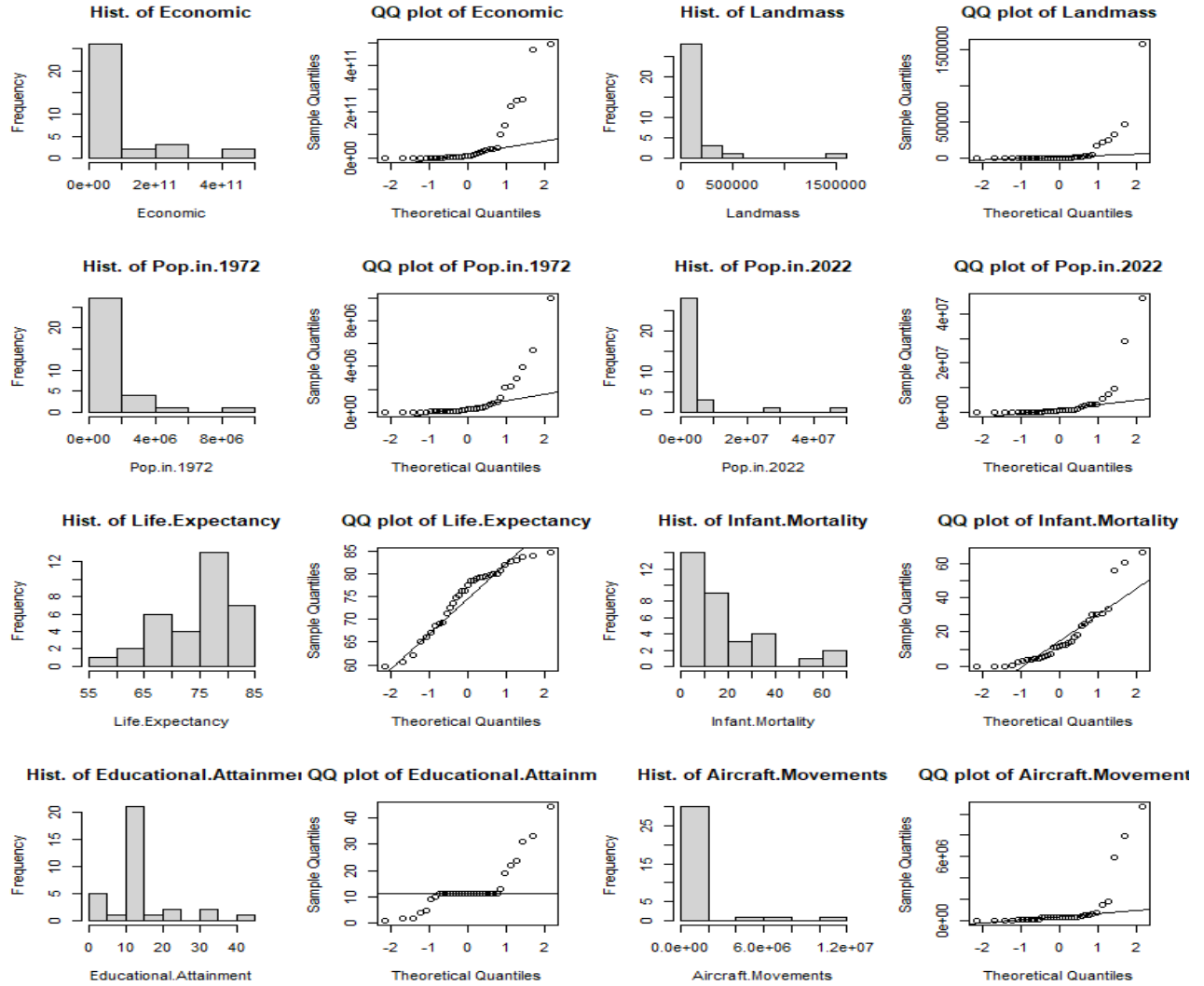


Figure 13: Distribution of Numerical Variables

Let us start by visualizing the distributions of numeric variables. There are a lot of cases where we want to know if our data follows a certain distribution or not so that we can decide whether some of the methods are suitable or not. Certain ways we can use to check the normality of a variable is by using a Histogram or a Quantile-Quantile plot.

From the histograms and QQ plots in figure 13 we can see that Economic, Population, Infant Mortality, Aircraft Movements, and Landmass are skewed to the right. The distribution of life expectancy is skewed to the left whereas the distribution of Educational Attainment is still not known.

3.1 What are the States and Number of States in each Continent?

In the dataset, there is a categorical variable continent that contains 6 observations. What could be the distribution of the categorical variable? Now let us look at the number of states in each continent.

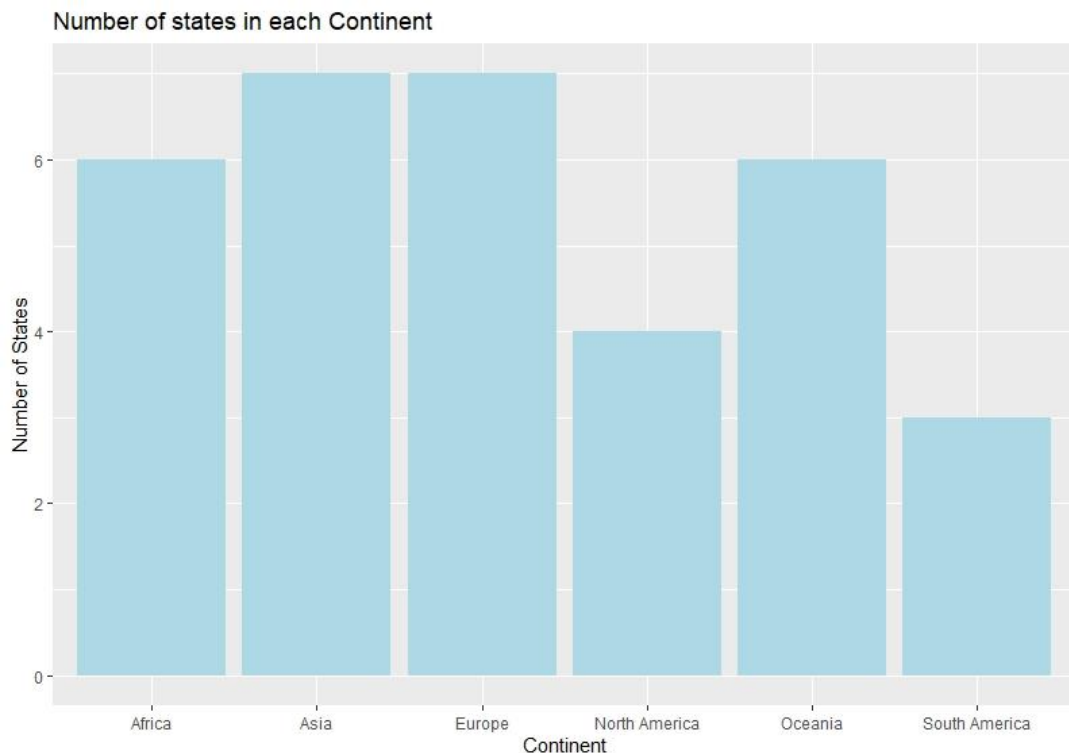


Figure 14: State count in each continent

The bar graph tells us that we have a relatively large number of small islands in Asia and Europe(7). There are six islands each in Oceania and Africa. There are 4 small islands in North America. However, we can also see that there are only 3 small islands in South America.

In the above figure, we have seen the number of states in each continent. However, do we exactly know which state belongs to which continent? We will now see a plot of the states vs continent which will give us an idea as to which state belongs to which continent.

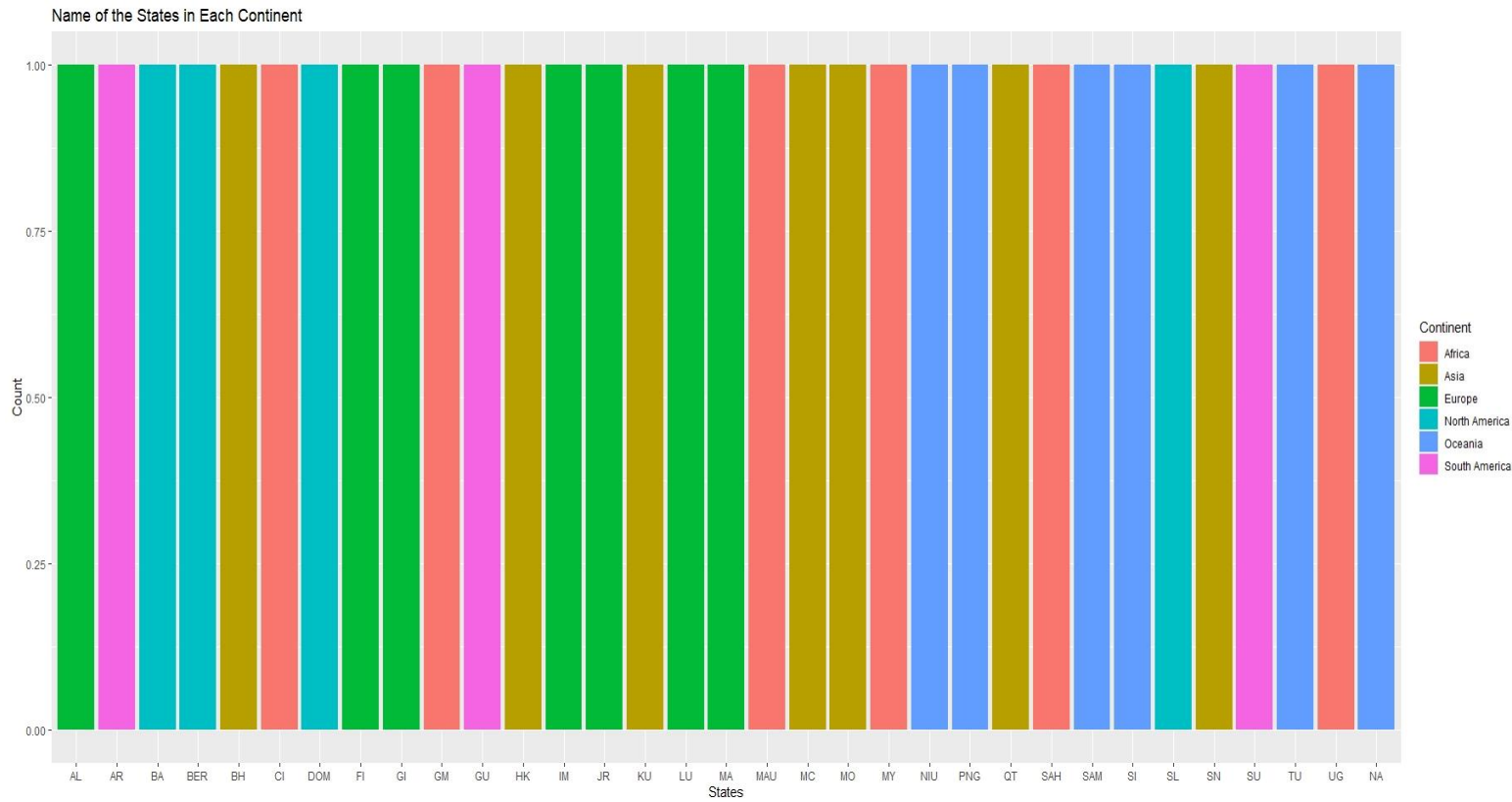


Figure 15: States in Each Continent

From figure 15 we can see that the states in South America which is represented in the color pink in the above figure are Aruba(AR), Guyana(GU), and Suriname(SU). The states in Oceania which is represented by light blue are Niue(NIU), Papua New Guinea(PNG), Samoa Islands(SAM), Solomon Islands(SI), Tuvalu(TU), and Nauru(NA). The states in North America which is represented by light green are Barbados(BA), Bermuda(BER), Dominica(DOM), Malta(MA), and St Lucia(SL). The states in Europe which are represented by dark green are Albania(AL), Faroe Islands(FI), Gibraltar(GI), Isle of Man(IL), Jersey(JR), and Luxembourg(LU). Asia which is represented by brown has the largest number of states and these are Bhutan(BH), Hong Kong(HK), Kuwait(KU), Macau(MC), Mongolia(MO), Qatar(QT), and Singapore(SN). Finally, the states in Africa which is represented by the color Orange are Cote d'Ivoire(CI), Gambia(GM), Mayotte(MY), South American Helena(SAH), and Uganda(UG).

Now that we know the number of states in each continent, we will have a look at the relationship between state and population. I would use a bar chart here to understand the population growth of each island from 1972 to 2022. Here is the bar chart that shows the relationship between the same.

3.2 What is the population and how is the population growth in each state?

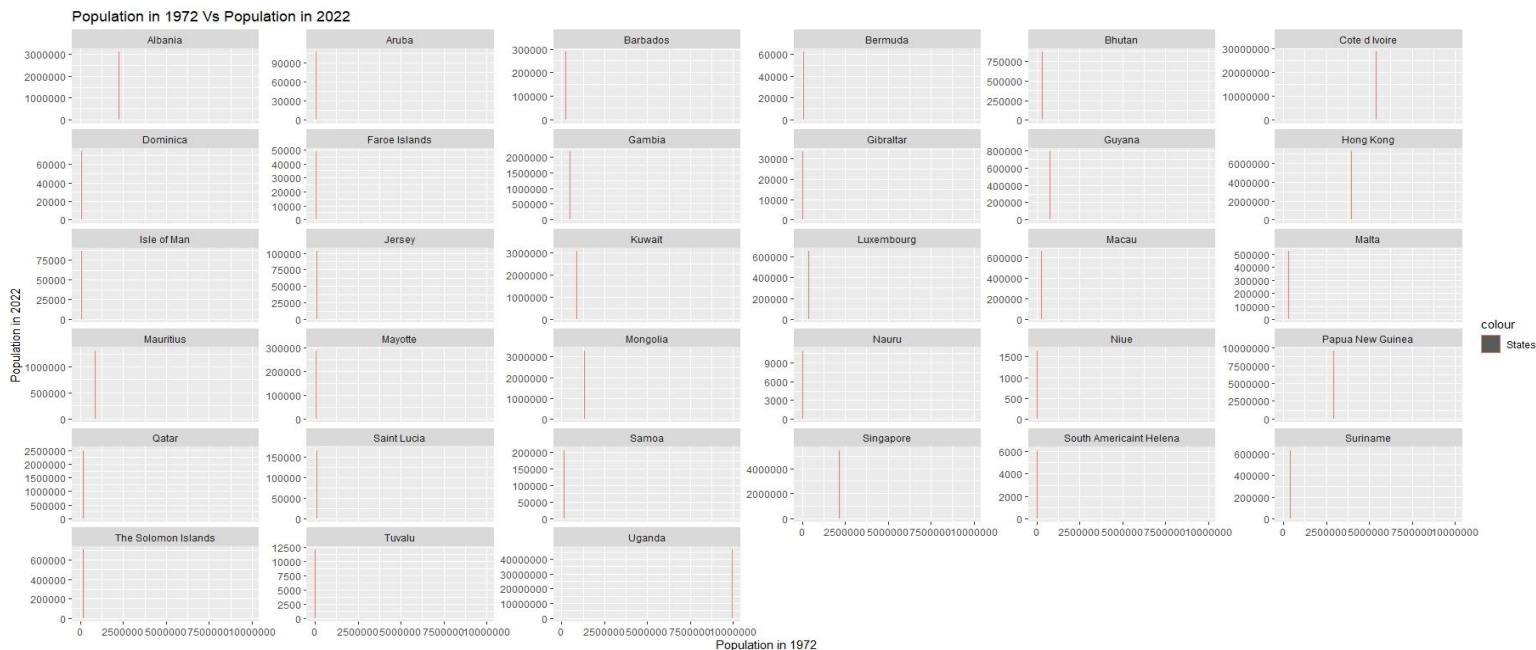


Figure 16: Population Growth in the Different States

From the plot, we can see that almost all the states have seen some increase in population from 1972 to 2022. However, there is one state that is Niue which has seen a decrease in the population from 1972 to 2022. It is also clear that Uganda is the state that has had the highest population in the early days as well as today. Uganda has seen significant growth from 10 million in 1972 to 40 million in 2022. On the other hand, Niue has seen a drop in the population from about 4000 to about 1500 in 2022. Other questions we may ask are what could be the economy of these states? What could be the factors that might be affecting these states? What could be the educational attainment in these states and what about the aircraft movements from such states where there is only 1 airport? We will be trying to answer a lot of questions about the data and visualize them further in the discussion. Also, further in the discussion population data will be based on the current population of each state.

Now that we have seen the population for different states, let us now visualize the landmass vs population for each state.



Figure 17: Landmass Vs Population for each state

From the plot, we can see that most of the states that have landmass close to zero (cannot be exactly zero) indicate that their landmass is very small, and might be within 1000 square km. Because of the scale on the x-axis, you can see that it shows zero on the above plot but it is not really zero. Further, if we have a look at the above plot, we can also see that the landmass of Uganda is 0.25 million, however, it has the largest population among all the states even though it does not have the largest landmass. We can also infer that Mongolia which has the largest landmass of over 1.5 million among all the states has a small population of 3 million. Does this mean that the Population of the state is independent of its landmass?

Before we move further, let us first analyze the relationship among the numerical variables and for this, we will be using a correlation plot which is the best way to understand the relationship between the numerical variables.

3.3 What is the correlation between the numerical variables in our data?

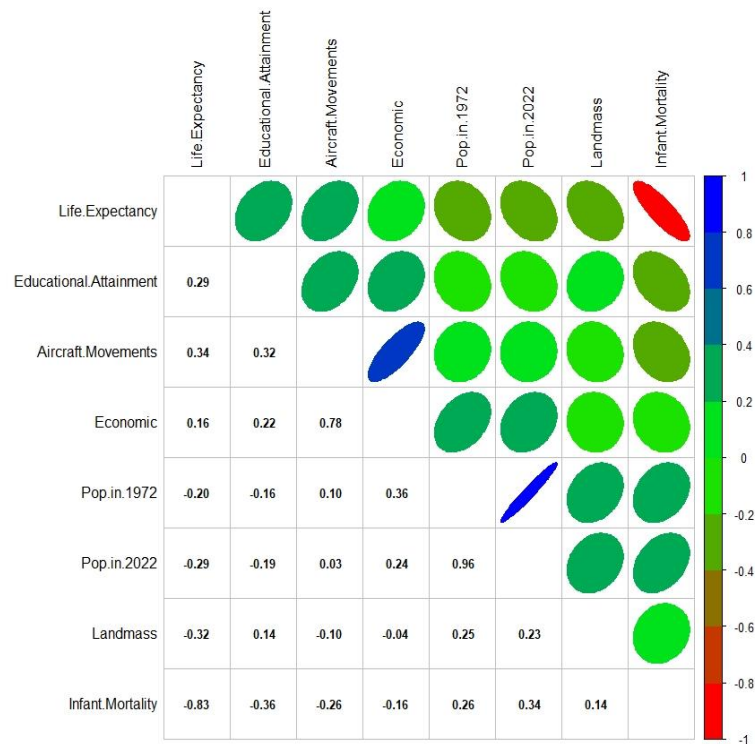


Figure 18: Correlation between the Numerical Variables

In the top right corner of figure 18, it is clear that the red and narrow shape between Infant mortality and life expectancy shows a high negative correlation whereas the blue narrow shape between economic and aircraft movements shows a high positive correlation. The light green shapes between population(in 1972 and 2022) and educational attainment, landmass and aircraft movements, landmass and economy, infant mortality rate, and economy show a small negative correlation. The dark green shapes between educational attainment and life expectancy, economy and educational attainment, population and economy, aircraft movements and educational attainment, landmass and population, infant mortality and population show a small positive correlation.

A positive correlation between economy and aircraft movements with an R-value of 0.78 means that the higher the economy the state has, the higher will be the aircraft movement, in other words, they are closer to +1 and hence are highly positively correlated. This is one of the pairs that will be really interesting for analysis. This makes sense as we have already spoken about how important aircraft movement data are for these small island nations and how much value they add. On the other hand, a negative Correlation between life expectancy and infant mortality, infant mortality, and educational attainment means that the higher the infant

mortality, the lower will be the educational attainment and life expectancy of the state. These pair would also be interesting to analyze as there are strongly negatively correlated.

3.3.1 What is the ANOVA test and How do we do it in R?

To examine our hypothesis, we will utilize the One Way ANOVA test, but first, what is the ANOVA test? ANOVA is an acronym that stands for analysis of variance. This is a statistical procedure for comparing the means of two or more groups for one dependent variable. When the investigation comprises more than two groups, this technique is required. In this scenario, the normal distribution assumption is not required. We utilize the built-in ANOVA function `aov()` in R for the ANOVA test. Also note that before performing the ANOVA test, a minimum sample of size 30 is generally advised as this will reduce the risk of making type II error and if you want to reduce the risk of making type I error then you may choose to lower the alpha level or level of significance (Ross & Willson, 2017).

Our null hypothesis from figure 17 is that landmass and population are strongly correlated with each other whereas the alternative hypothesis is that they are weakly correlated. Performing the ANOVA test with a p-value of 0.20 which is greater than 0.05 also gives us evidence that the null hypothesis is not rejected. Figure 21 also shows that landmass and population are correlated. Below is a figure that shows the summary of the ANOVA test.

```
> model <- aov(data$Landmass ~ data$Pop.in.2022, data)
> summary(model)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data\$Pop.in.2022	1	135110856257	135110856257	1.696	0.202
Residuals	31	2468945874348	79643415302		

```
>
```

Figure 19: ANOVA test between Landmass and Population

Now that we have visualized the landmass vs population per state, let us now visualize the population per square km area by continent.

3.4 How is the population density for each state?

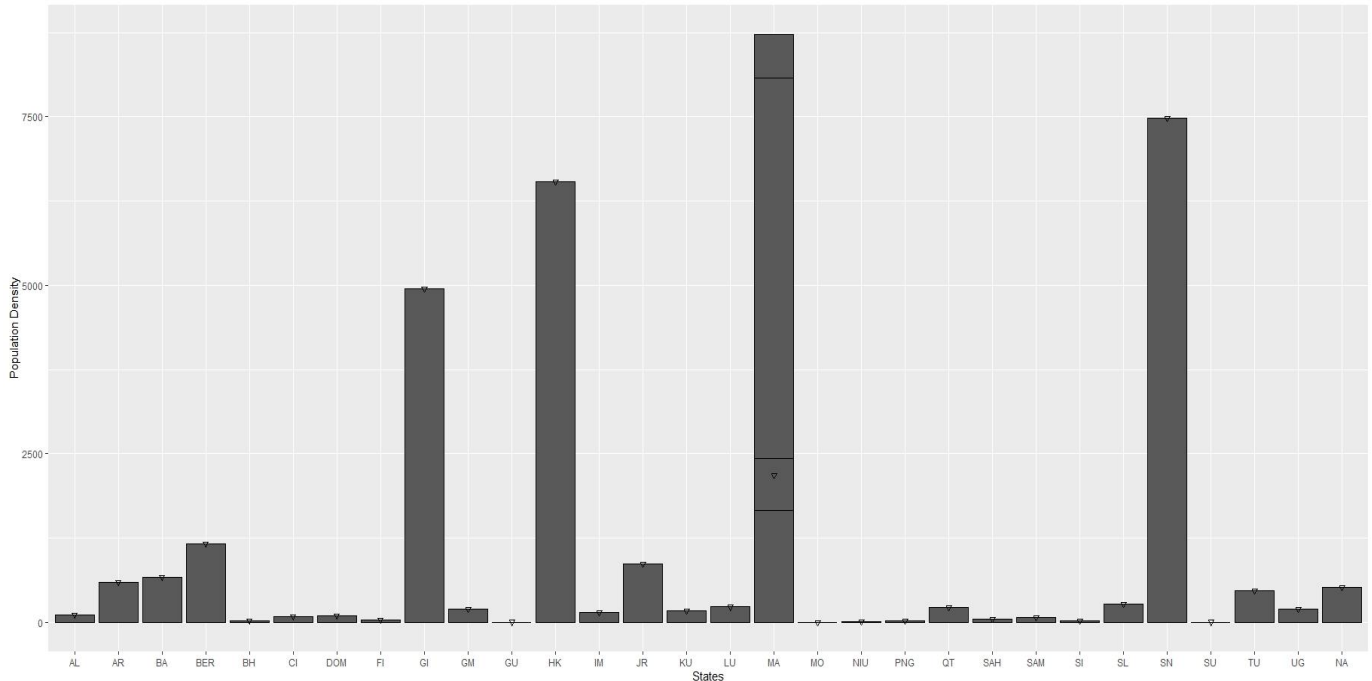


Figure 20: Bar chart of Population density by States

The bar plot shows that the mean pop density of Singapore is the largest among all the other states, while Guyana and Suriname have the least mean population density. I state the null hypothesis that mean population density is not the same for all the continents and the alternative hypothesis is the mean population density is the same for all the continents. From the p values in figure 21, it is also evident that the null hypothesis is not rejected as the p-value(0.73) is greater than 0.05. The mean of population density is represented by the small triangle inside the bar plot.

3.4.1 Is the mean distribution of the population density of all the states the same?

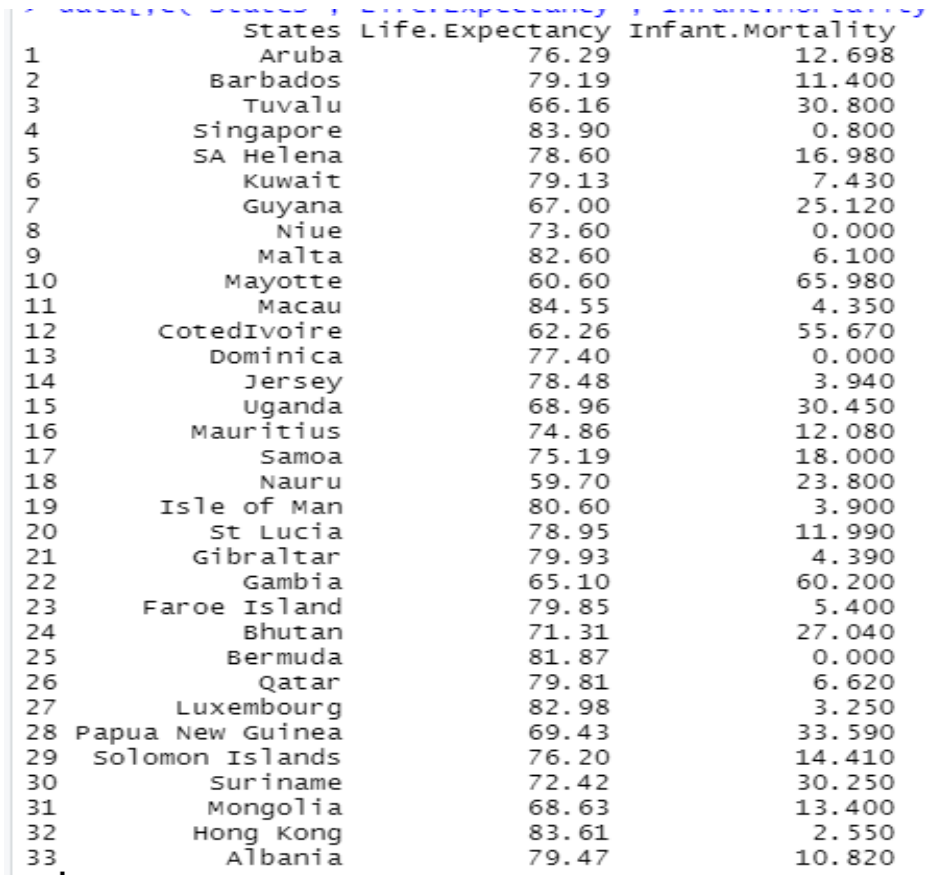
```
> model1 <- aov(data$pop.density ~ data$State.Abbreviations, data)
> summary(model1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data\$State.Abbreviations	28	110803028	3957251	0.711	0.738
Residuals	3	16692171	5564057		

Figure 21: ANOVA test for Population density by States

3.5 How are infant mortality and life expectancy related?

Now let us visualize the life expectancy and infant mortality for each state. Below is a plot of the same.



	States	Life.Expectancy	Infant.Mortality
1	Aruba	76.29	12.698
2	Barbados	79.19	11.400
3	Tuvalu	66.16	30.800
4	Singapore	83.90	0.800
5	SA Helena	78.60	16.980
6	Kuwait	79.13	7.430
7	Guyana	67.00	25.120
8	Niue	73.60	0.000
9	Malta	82.60	6.100
10	Mayotte	60.60	65.980
11	Macau	84.55	4.350
12	CotedIvoire	62.26	55.670
13	Dominica	77.40	0.000
14	Jersey	78.48	3.940
15	Uganda	68.96	30.450
16	Mauritius	74.86	12.080
17	Samoa	75.19	18.000
18	Nauru	59.70	23.800
19	Isle of Man	80.60	3.900
20	St Lucia	78.95	11.990
21	Gibraltar	79.93	4.390
22	Gambia	65.10	60.200
23	Faroe Island	79.85	5.400
24	Bhutan	71.31	27.040
25	Bermuda	81.87	0.000
26	Qatar	79.81	6.620
27	Luxembourg	82.98	3.250
28	Papua New Guinea	69.43	33.590
29	Solomon Islands	76.20	14.410
30	suriname	72.42	30.250
31	Mongolia	68.63	13.400
32	Hong Kong	83.61	2.550
33	Albania	79.47	10.820

Figure 22: Infant Mortality Vs Life Expectancy per each state

From the above plot, we can see that as the infant mortality rate increases the life expectancy decreases. For instance, states such as Mayotte have an infant mortality rate of above 60, however, the life expectancy of these states is less than 60. On the other hand, Hong Kong, Gibraltar, Niue, and many more have an infant mortality rate below 5, and their life expectancy is 84, 80, and 74 respectively, which constitute a lot of difference between Infant mortality and Life expectancy. From figure 18, we know that they are negatively correlated, hence the above visualization makes even more sense. There are a few questions that arise, as from the above plot we can infer that states such as Mayotte, and Cote d Ivoire have a lower life expectancy, does that mean on average the life expectancy of that particular continent will also be low?

Also, because of the high negative correlation between life expectancy and infant mortality rate, we will take a look at the distribution of life expectancy per continent and what we can infer from that particular plot.

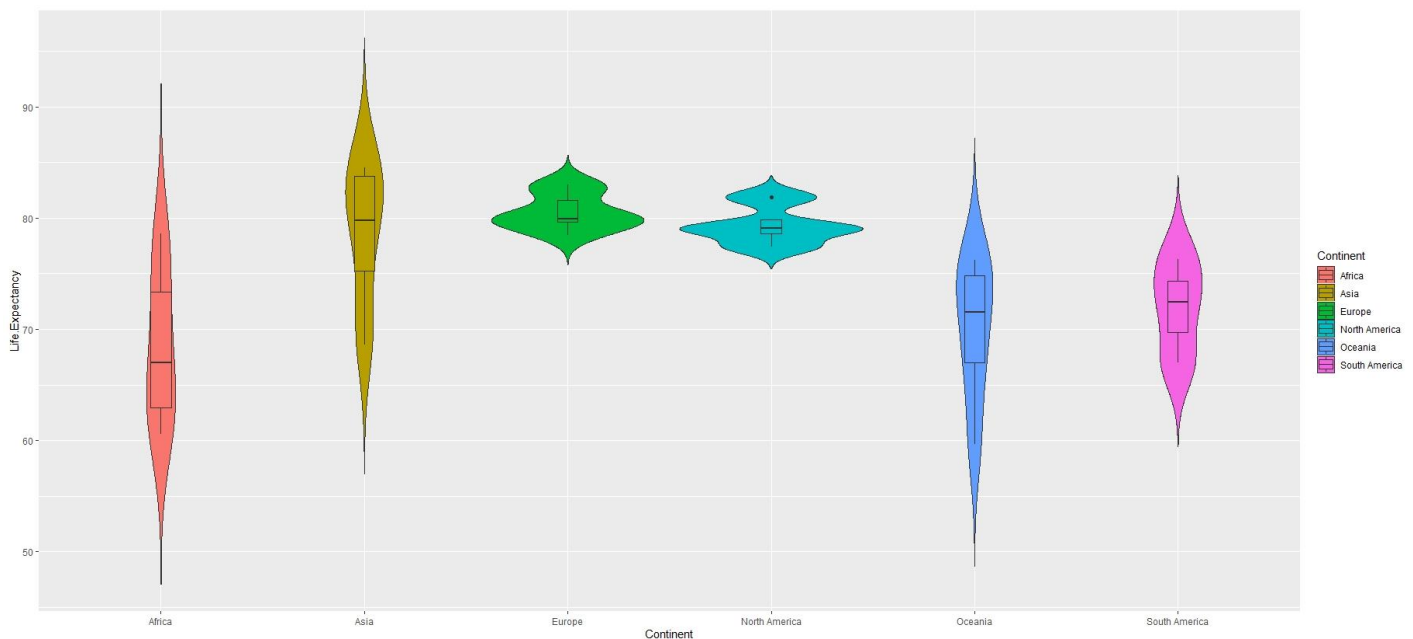


Figure 21: Life Expectancy per Continent

On average, Africa has the lowest life expectancy of about 68 which is lower than the other continents. Asia and Europe have a high life expectancy of about 80. However, we can see two long tails on each end for Africa, Asia, and Oceania which means that some states in these continents have a long life expectancy, while some have short life expectancy though they are on the same continent. Also, since Europe and North America have a wide shape in the middle of the violin plot indicates that the life expectancy of the states in these continents is highly concentrated around the median. From the above plot, we may ask questions like what are those states in these three continents that have a high and low life expectancy? We will try and answer this question from the next plot.

3.5.1 Is the mean distribution of life expectancy of all the continents the same?

```
> model2 <- aov(data$Life.Expectancy ~ data$Continent, data)
> summary(model2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data\$Continent	5	825.4	165.08	5.803	0.000919 ***
Residuals	27	768.1	28.45		

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

Figure 22: AOV test between Life Expectancy and Continent

Figure 22 gives us the AOV test between life expectancy and continent, which also gives us evidence to reject the null hypothesis as the p-value is 0.000919 which is less than 0.05. This means that at least one of the continent has a different mean life expectancy and not all the continents have a similar mean life expectancy.

3.5.2 Which are the states that have both low and high life expectancy and which continent do they belong to?

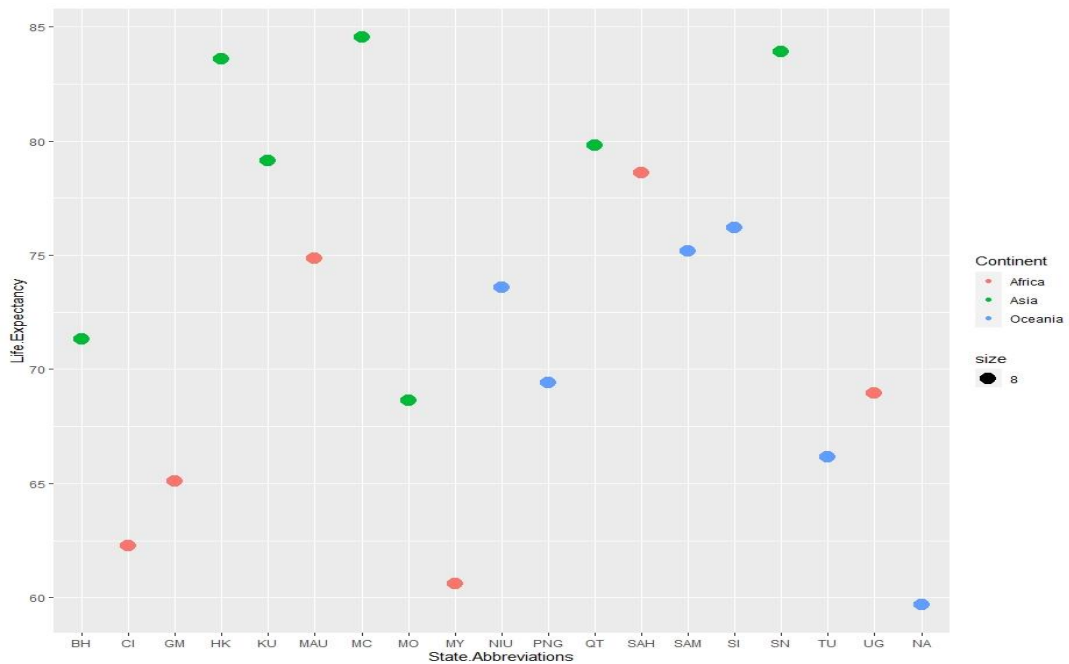


Figure 23: Scatterplot of Life Expectancy per state in each Continent

We discussed above that Africa has a lower life expectancy than the other three continents and it turns out to be true from the above plot. The states represented in orange are from Africa and states such as Cote d Ivoire(CI), Mayotte(MY), and Gambia(GM) all have a life expectancy between 60-70. However, from figure 21 we also saw that Africa has two long tails on each end to show that some states have low life expectancy whereas some have a high life expectancy and it is true which can be seen in figure 23. The states such as MY, CI, and GM have low expectancy whereas states such as Mauritius(MAU), and South Americaint Helena(SAH) have a higher life expectancy and all these states belong to the same continent Africa. Similar situations can be seen in Oceania as well where states such as Nauru(NA) have low life expectancy and other states have higher life expectancy. Finally, in Asia most of the states have high life expectancy except Mongolia which has low life expectancy as compared to other states in Asia.

3.5.3 Is the mean distribution of the life expectancy in the 3 continents the same?

```
> model3 <- aov(sub_data$Life.Expectancy ~ sub_data$Continent, sub_data)
> summary(model3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sub_data\$Continent	2	405.8	202.91	4.653	0.0255 *
Residuals	16	697.8	43.61		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>

Figure 24: ANOVA test for Life Expectancy vs Continent for the subset data

When we performed the AOV test for the subset that I created considering only 3 continents(Africa, Asia, Oceania), I got the statistic as shown in figure 24. As explained above, the null hypothesis is rejected here as well($0.02 < 0.05$) which means that at least one continent has different mean distribution.

Now that we are clear about the life expectancy of different states. Let us try and understand infant mortality rate. Previously, we discussed as infant mortality rate increases the life expectancy decreases. Now that we know life expectancy is high for which states and on an average for which continent, this should mean that infant mortality rate have to be lower for those states and on an average for the same continent as well. We will now try and find out does it really mean that way.

3.5.4 What is the distribution of the infant mortality rate per continent?

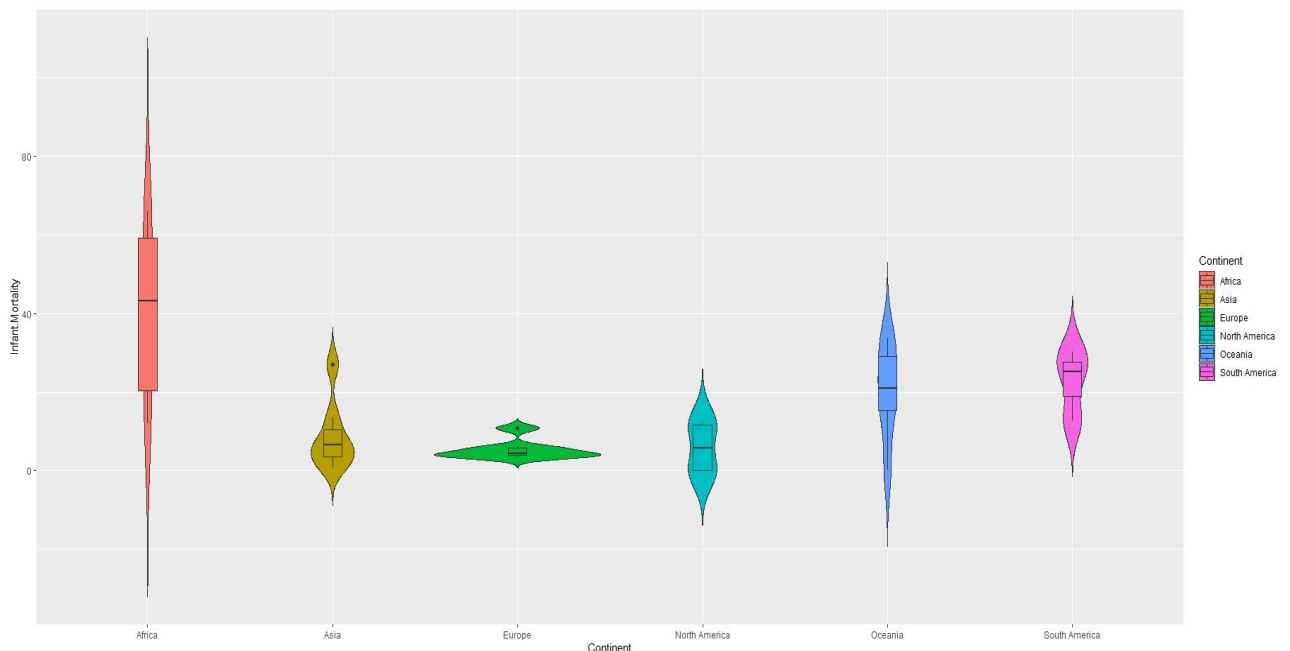


Figure 25: Infant Mortality Rate per Continent

On average, Africa has the highest infant mortality rate and it turns out exactly what we expected it to be. On the other hand, Europe and Asia have lower infant mortality rates as compared to North America, South America, and Oceania. However, Africa has long tails on either end which means that some states have high infant mortality and some have lower. It is similar to Oceania, which also has long tails on either side. Europe has a wide shape in the middle of the plot and as already explained, it indicates that infant mortality is highly concentrated around the median.

Now let us try and find out which states have high and low infant mortality in the continents of Africa, Oceania, and Asia, just like how we studied life expectancy in detail in figure 23.

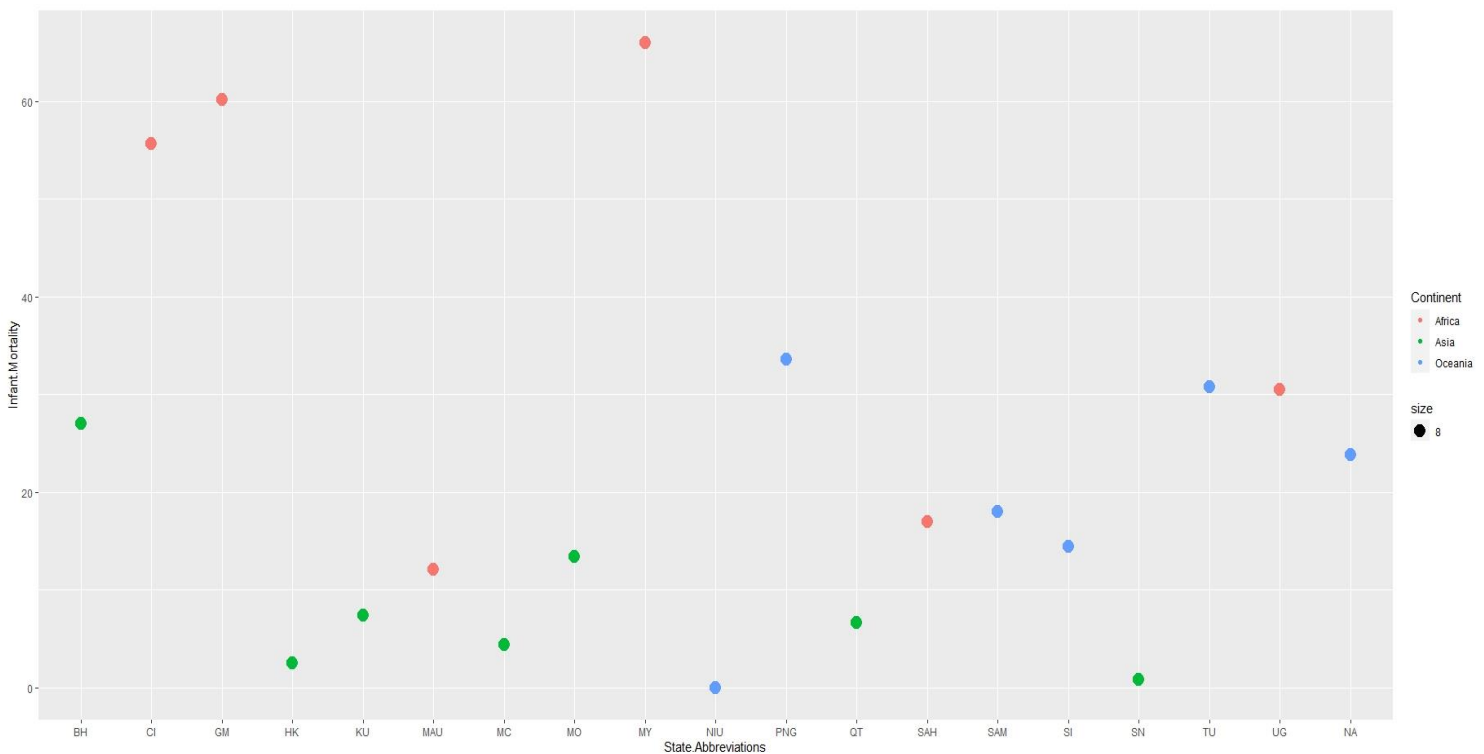


Figure 26: Infant Mortality for each state in 3 continents

From the plot, it is clear that Africa has 3 states that have high infant mortality rates and 3 states that have low infant mortality states, hence the long tail on both sides in plot 25. Oceania also has a similar trend that we expected. We can also infer that Asia has states that have the lowest infant mortality rate.

3.5.5 Is the mean distribution of the infant mortality rate same for all the continents?

```
> model4 <- aov(data$Infant.Mortality ~ data$Continent, data)
> summary(model4)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
data\$Continent	5	5290	1058.1	6.639	0.000377	***
Residuals	27	4303	159.4			

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 27: ANOVA test for Infant Mortality and Continent

We have used the AOV test as explained by (Ross & Willson, 2017) in the above figure, there is clear evidence that the null hypothesis is rejected as the p-value(0.000377) is less than 0.05. This means that there is at least one continent that has a different mean infant mortality rate.

Now we have understood and visualized the distribution of Population, landmass, life expectancy, and infant mortality in each state and each continent. From the correlation plot in figure 18, there are other variables like Educational Attainment, Economy, and Aircraft Movements. Further in this report we will try and understand these variables.

From figure 18 it is clearly visible that life expectancy and educational attainment are positively correlated. There are also studies such as articles or blogs which suggest that on average college graduates live longer than those without a degree. But how true is it in my case? Let's have look at the life expectancy vs educational attainment for each state and continent.

3.6 How is Life Expectancy related to Educational Attainment?

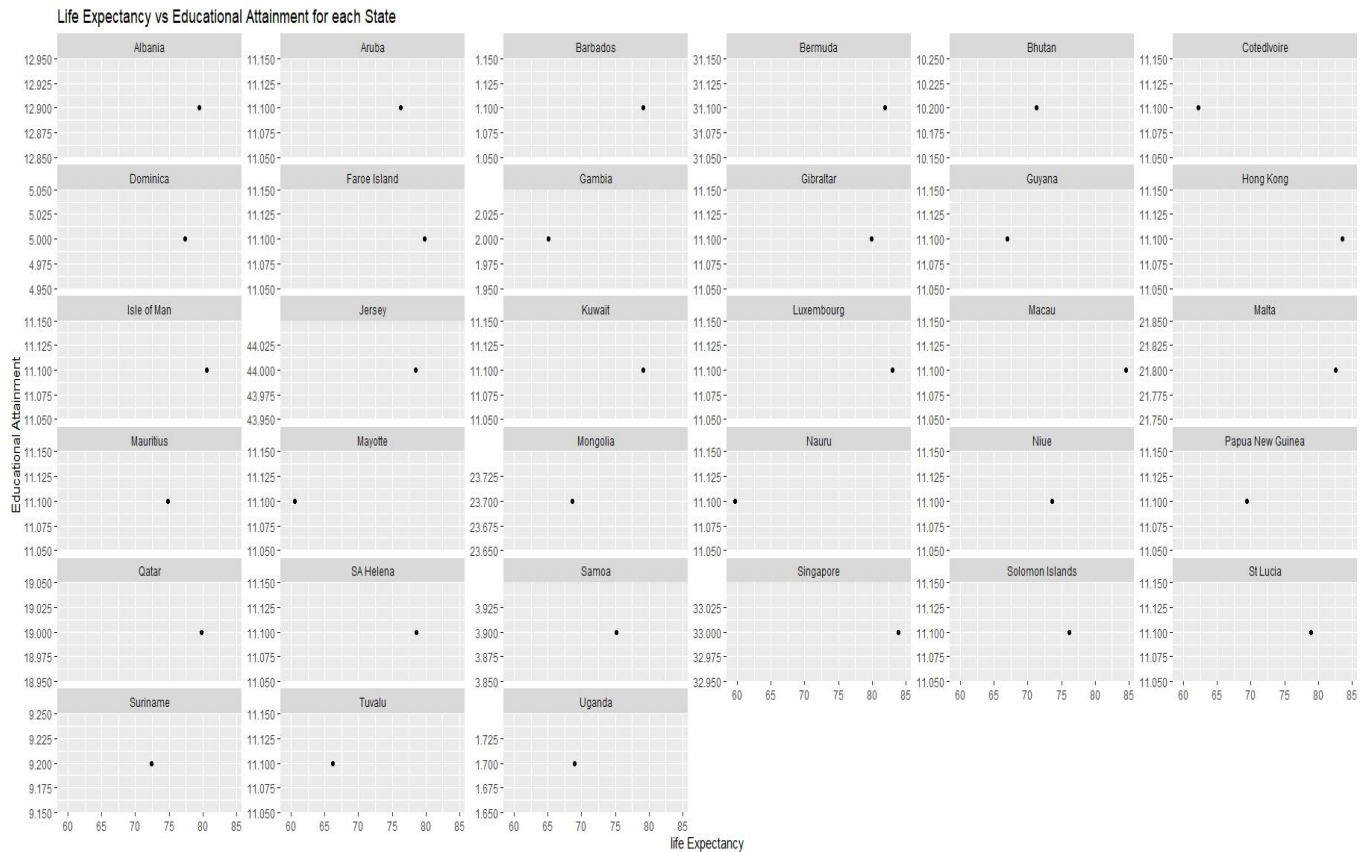


Figure 28: Life Expectancy Vs Educational Attainment per each state

The plot above is pretty weird, we expected that those states that have high life expectancy will have high educational attainment. There are states like Singapore, Bermuda, and New Jersey that have a high life expectancy and high educational attainment. However, there are a lot of states like Luxembourg, Macau, St Lucia, and many more that have a high life expectancy of about 70-80 but their educational attainment is about 11%. There are also states that have a low life expectancy and low educational attainment of about 1% which is what we expected but there are also states that have a low life expectancy of 55-65 and high educational attainment of about 11%. This is what is weird and a bit confusing. While there are a lot of states that have a life expectancy between 65-70 and educational attainment of about 11%, there is one state that has the same life expectancy between 65-70 but educational attainment of about 1% which is the least among all the states.

After performing the aov test with a p-value of 0.1 we have clear evidence that the null hypothesis is rejected which means life expectancy and educational attainment are significantly correlated. Below is an image of the above test that has been performed.

```
> model31 <- aov(data$Life.Expectancy ~ data$Educational.Attainment, data)
> summary(model31)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data\$Educational.Attainment	1	134.1	134.10	2.848	0.102
Residuals	31	1459.4	47.08		

Figure 29: ANOVA test between Life expectancy and educational attainment

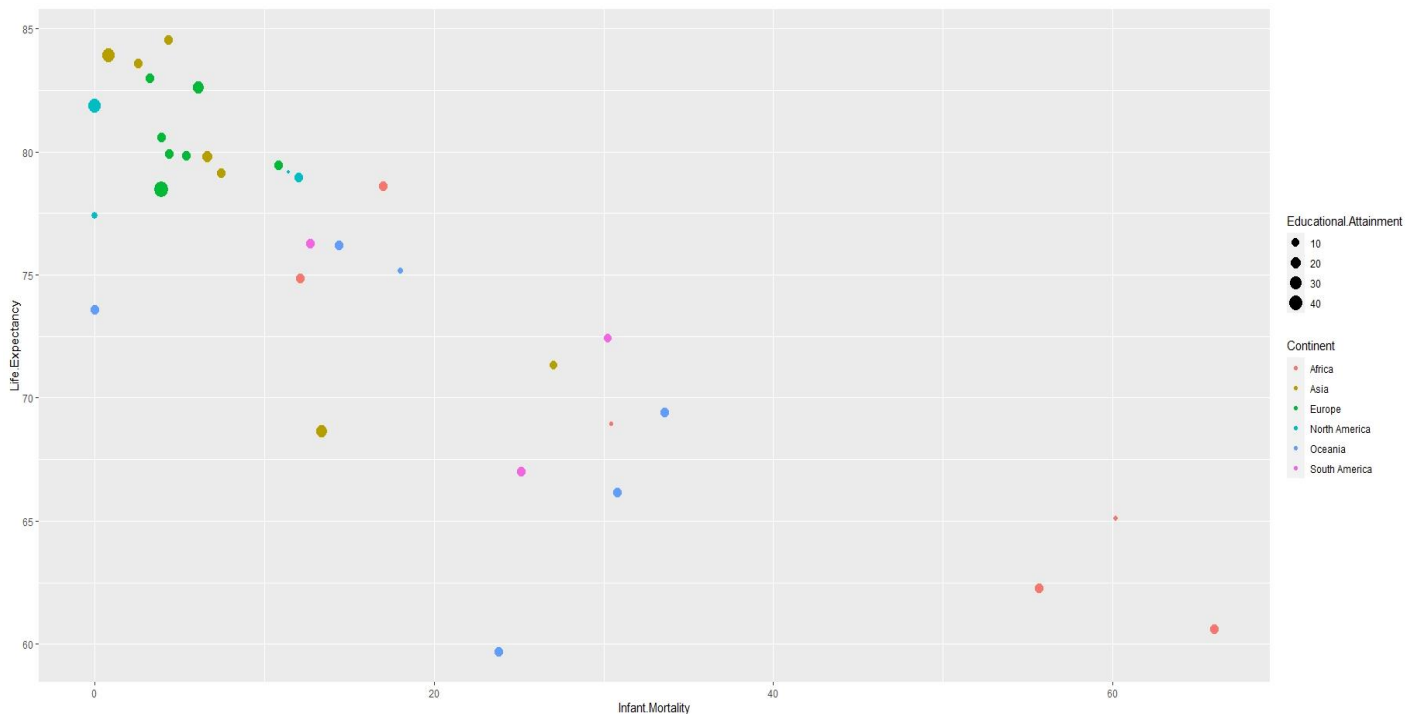


Figure 30: Infant Mortality Vs Life Expectancy, educational attainment

From the above plot, it is visible that life expectancy and infant mortality are negatively correlated. Some states that have infant mortality rates over 50 have small symbols, which means that their educational attainment is low about 10-20%, and these states are colored orange which means they all belong to the same continent which is Africa.

Most of the continents have lower educational attainment, and lower life expectancy but higher infant mortality rate, while states in the other continents have relatively higher educational attainment, lower infant mortality rate, and high life expectancy.

We are clear with the relationship between life expectancy, and infant mortality rate with all the other variables. Let us now try to understand the relationship between educational attainment per population area per state in each continent.

3.7 How are educational attainment related to other variables?

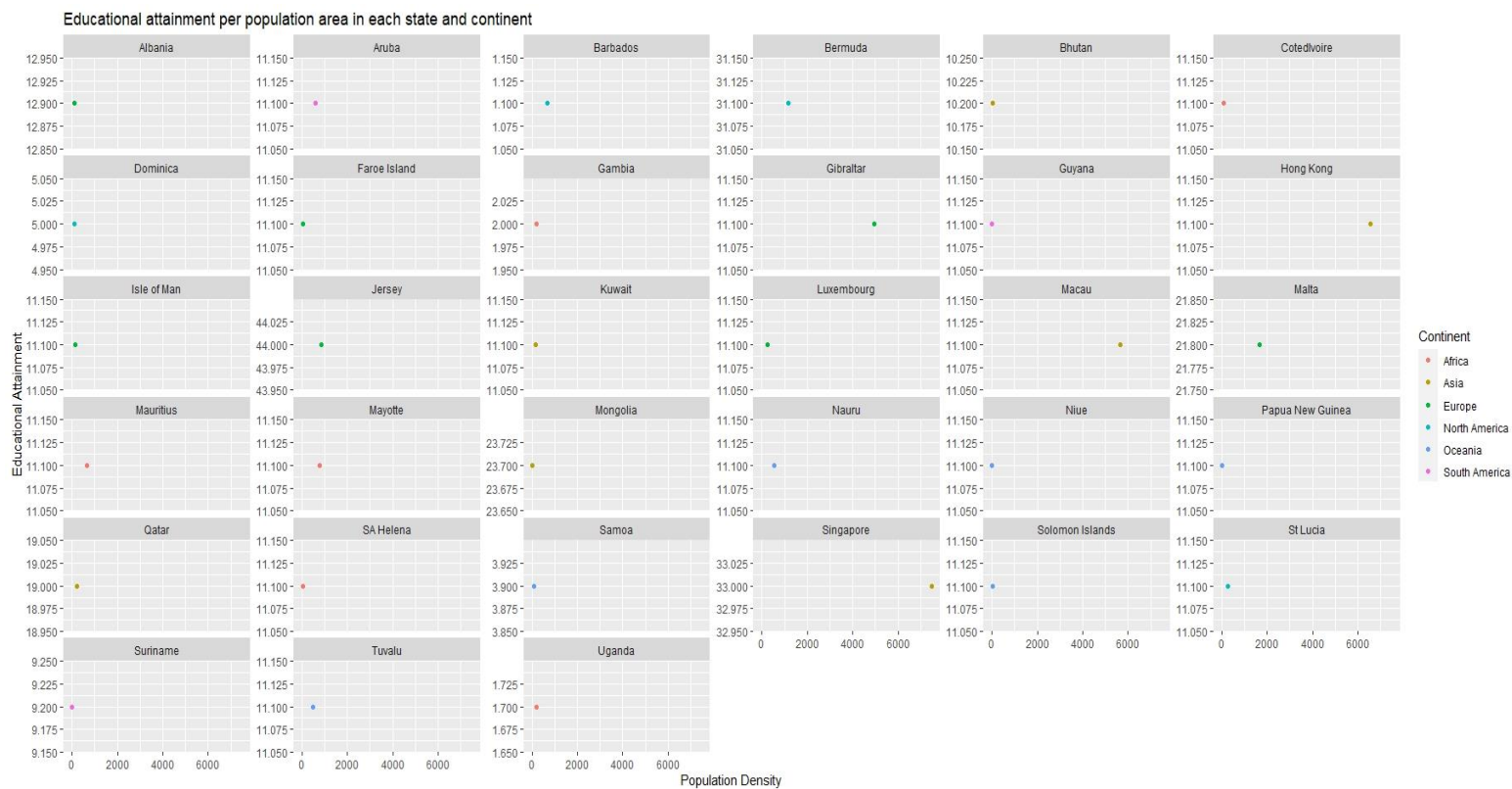


Figure 31: Educational attainment per population area

From the above plot, we can infer that almost all the states in all the continents have a low population per area and low educational attainment as well. However, there is one state that is jersey which is in the North American Continent that has a low population per area but it has higher educational attainment. Among all the states, jersey is the only state that has the highest educational attainment followed by Singapore in the Asian Continent. We can also see that there are a lot of states that have a high and low population per area but their education attainment still remains around 11%. While Jersey has the highest educational attainment, Uganda has the least educational attainment which is about 1% and the population per area is also pretty low, probably somewhere in the 100s. It can also be seen that a lot of states have the same educational

attainment, this is because of the method we used for imputing missing values where we replaced the missing values with the median.

In the correlation plot from figure 18, we can see that educational attainment and the economy of the country are positively correlated. We will next try to find how are they positively correlated. How much effect does education attainment help on the country's economy? Do the states that have low educational attainment have a low economy? Is the economy majorly dependent on the education of the country or are there other factors as well that could improve a state's economy? Why do workers with college degrees earn so much more than those without degrees?



Figure 32: Educational Attainment Vs Economy per each state

The educational attainment data has been obtained from (The World Bank Data Group, 2021). Figure 32 suggests that all the states that have educational attainment of about 10% or above have a high economy. However, there are also states that have 10% educational attainment but their economy is low as compared to other states that have approximately the same educational attainment. Does this mean that education is not that important for economic growth for those countries and are there other factors that might be helping in the economic growth of the country?

One reason might be that the states that have educational attainment of 10% and a high economy might be having an excess supply of workers that do not require any sort of degree or any specialized training as well. There might also be

case where a single state could specialize only in one particular industry. However, a developed economy will generally include various industries. From figure 32, states such as Singapore and Hong Kong have educational attainment of about 33% and 11% respectively and their economy is high as well. In some cases, highly skilled workers that have qualified above the degree level might be concentrated in a specific geographic region. Also, education benefits not just the individual but benefits society as a whole. Graduates are more environmentally conscious, they have healthier habits, and have a higher level of civic participation. Also, increased tax revenues from higher earnings, healthier children, and reduced family size together are involved in building a stronger nation and economy. In short, education prepares each individual not only by providing them with job skills but also teaches them to be active members of their communities and societies. Hence educational attainment could be a significant factor in the economy of a country.

Next, we will look at the relationship between educational attainment and aircraft movements. Do the states that have high educational attainment mean they have high aircraft movements data? Does that mean more number of people are arriving and departing to and from the state for education purposes? Does it mean that more people are moving to different states for pursuing education? Let us visualize the relationship between these two features.

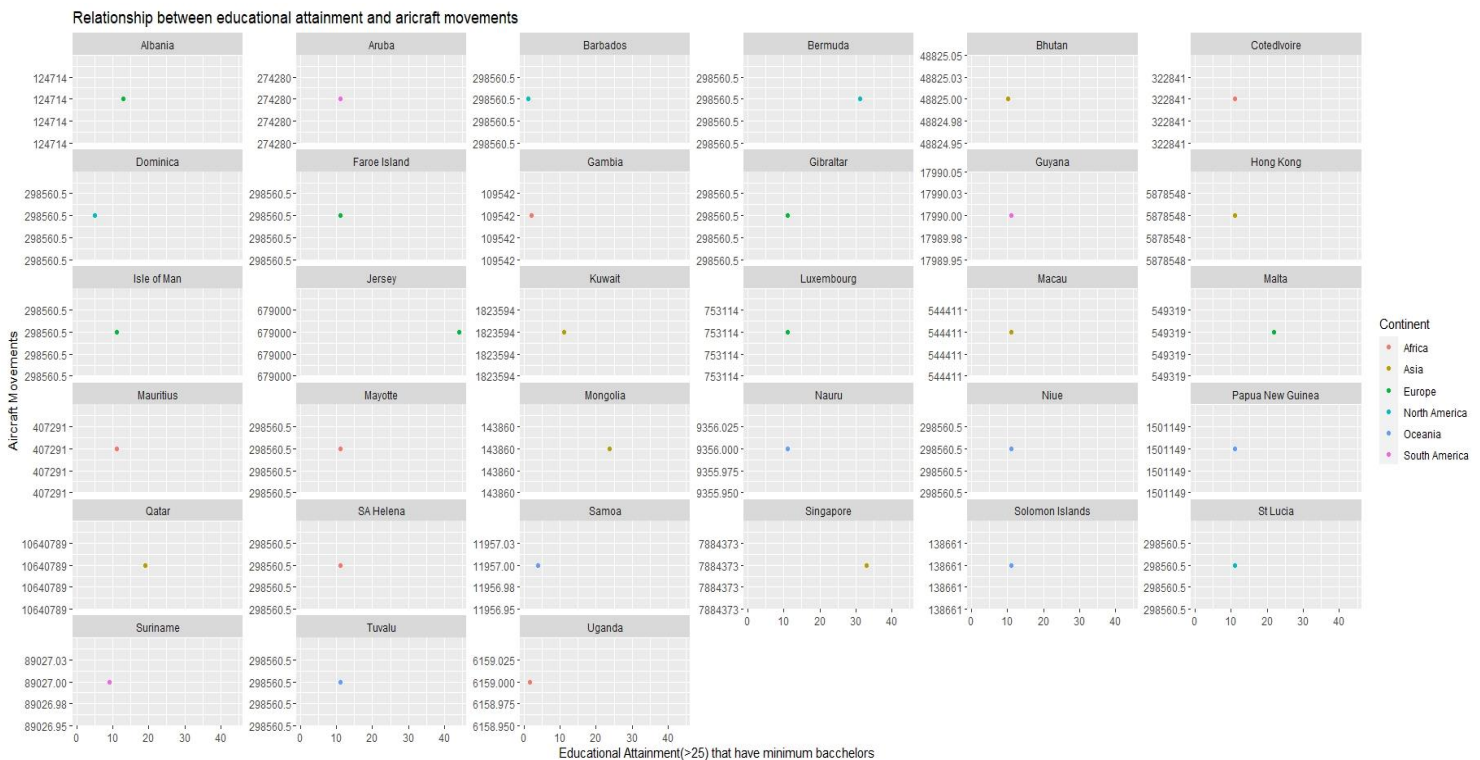


Figure 33: Relation between educational attainment and aircraft movements

The aircraft movement data has been collected from (The World Bank Data Group, 2021). We can infer from the above plot that states that have educational

attainment of 10% or below have aircraft movements in the 1000s. However, there are also states such as Dominica that has educational attainment of about 5% but their aircraft movements of about 0.3 million. This means that the passengers arriving or departing to and from the state are not only for educational purposes. This also does not mean that passengers moving to other states are just for educational purposes. Education could also be one of the reasons for aircraft movements to increase. There are also studies that talk about the importance of distinguishing between education acquired in the home country before migrating and education acquired in the host country. Just like educational attainment, a lot of states have a similar number of aircraft movements and that is also because of the method we used for imputing values.

Before we move further, let us look at the educational attainment per population area in each state.

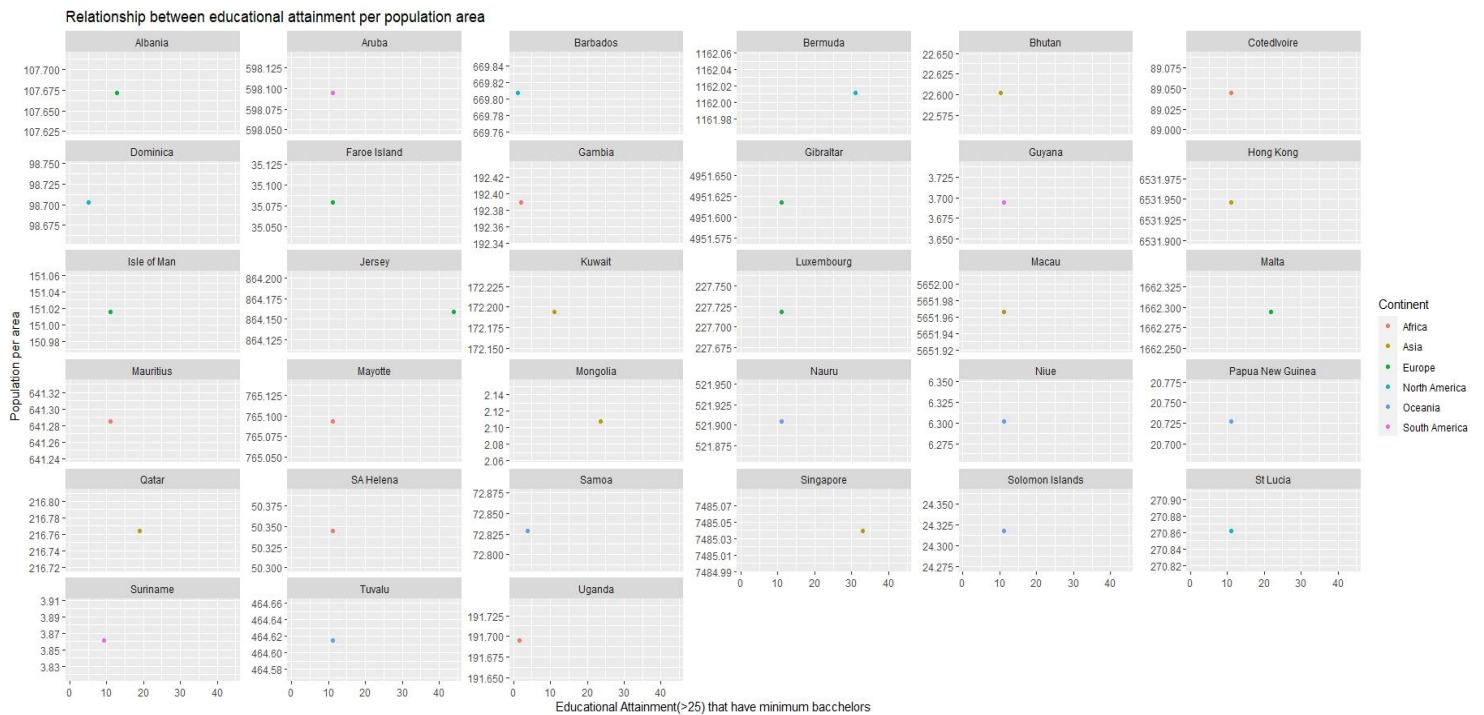


Figure 34: Educational Attainment per population area in each state

In the previous plot, we saw the educational attainment vs aircraft movements for the entire state as a whole. In the above plot, we have visualized the education per population area, the education per population area is pretty bad for all the states. However, states such as Mongolia have educational attainment of about 30% for an area of about 2km which I think is excellent whereas Singapore has an education of about 35% for an area of 7845km. On the other hand, there are also states such as Bermuda which has the same educational attainment of 35% but for a smaller area of 1200km which is the best among all.

Now we will visualize the relationship between aircraft movements and other variables. Firstly, we will see the distribution of the aircraft movements per state.

3.8 How are the Aircraft Movements related to other variables?

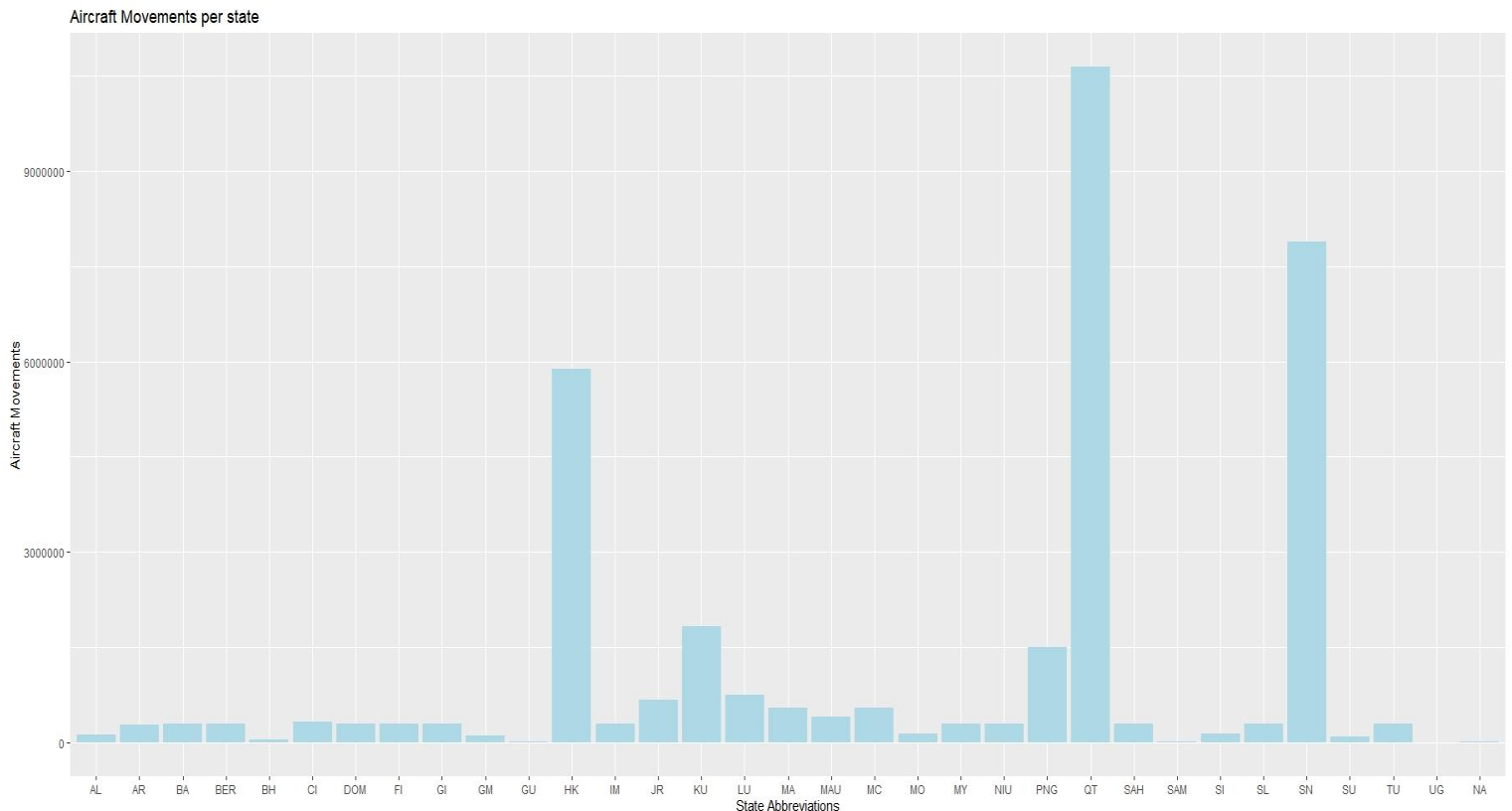


Figure 35: Aircraft Movements per State

The plot shows that Qatar(QT) has the highest number of aircraft movements at about 1 million followed by Singapore(SN) and Hong Kong(HK). States such as Barbados(BA), Bermuda(BER), Cote d Ivoire(CI), Dominica(Dom), Faroe Islands(FI), Gibraltar(GI), and Aruba(AR) have similar aircraft movements of below 0.05 million. Similarly, Bhutan(BH), Albania(AL), Suriname(SU), and Nauru(NA) have very low aircraft movements. What could be the reason for such low aircraft movements from these states? It could mean that these are not states that the tourists are generally interested in and there are not many places for tourists to visit as well. One more reason could be that the facilities in these states are not very great and hence a lot of people are departing from these states to other countries. While these states have at least some aircraft movements but are low,

there are states such as Guyana(GU), Samoa Islands(SAM), and Uganda which have very low aircraft movements such that they are not even visible in the above plot. We saw states that have high aircraft movements and extremely low aircraft movements but we do not which continents these states belong to. Let us have a look at that first and then we will check if all states have similar mean distributions of the aircraft movements per continent.

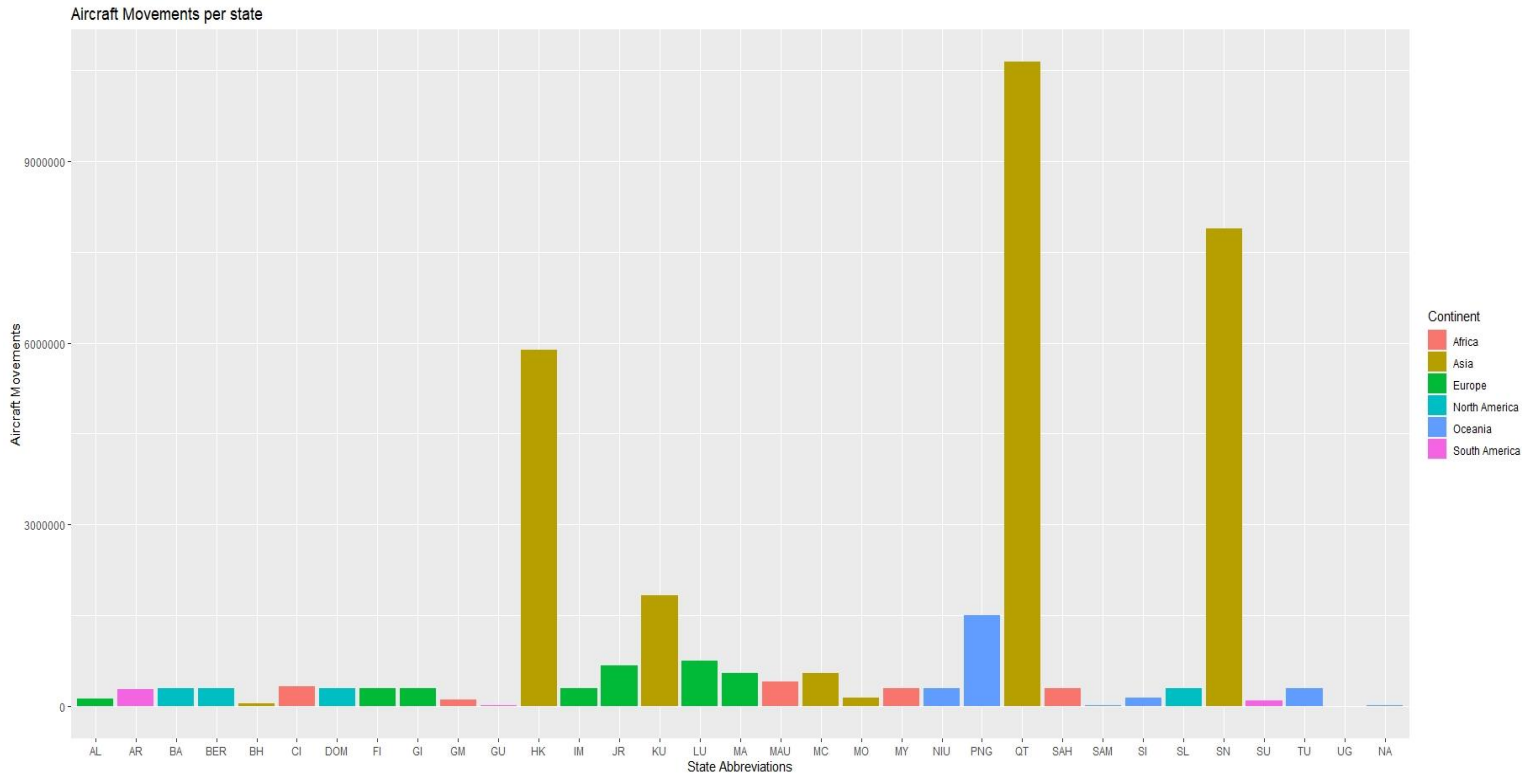


Figure 36: Aircraft Movements per State in each Continent

From the above plot, we can see that the highest number of aircraft movements are from Asia whereas North America which is represented by light blue has similar aircraft movements. After Asia, Europe and North America have a good amount of aircraft movements but Europe has a slightly higher number of aircraft movements as more states have higher aircraft movements than North America. Africa has a lower number of aircraft movements than Europe and North America which is represented by pink in the above plot. The continent that Uganda belongs to is not visible in the above plot but from figure 15 we know that it belongs to Africa as well. Oceania has a few aircraft movements where Papua New Guinea has the highest number of aircraft movements in Oceania. The above plot could also mean that the countries in Asia are more suitable for tourists and there are more tourism places in that particular continent.

Next, we will look at the mean distribution of the aircraft movements data per continent and see if they are similar.

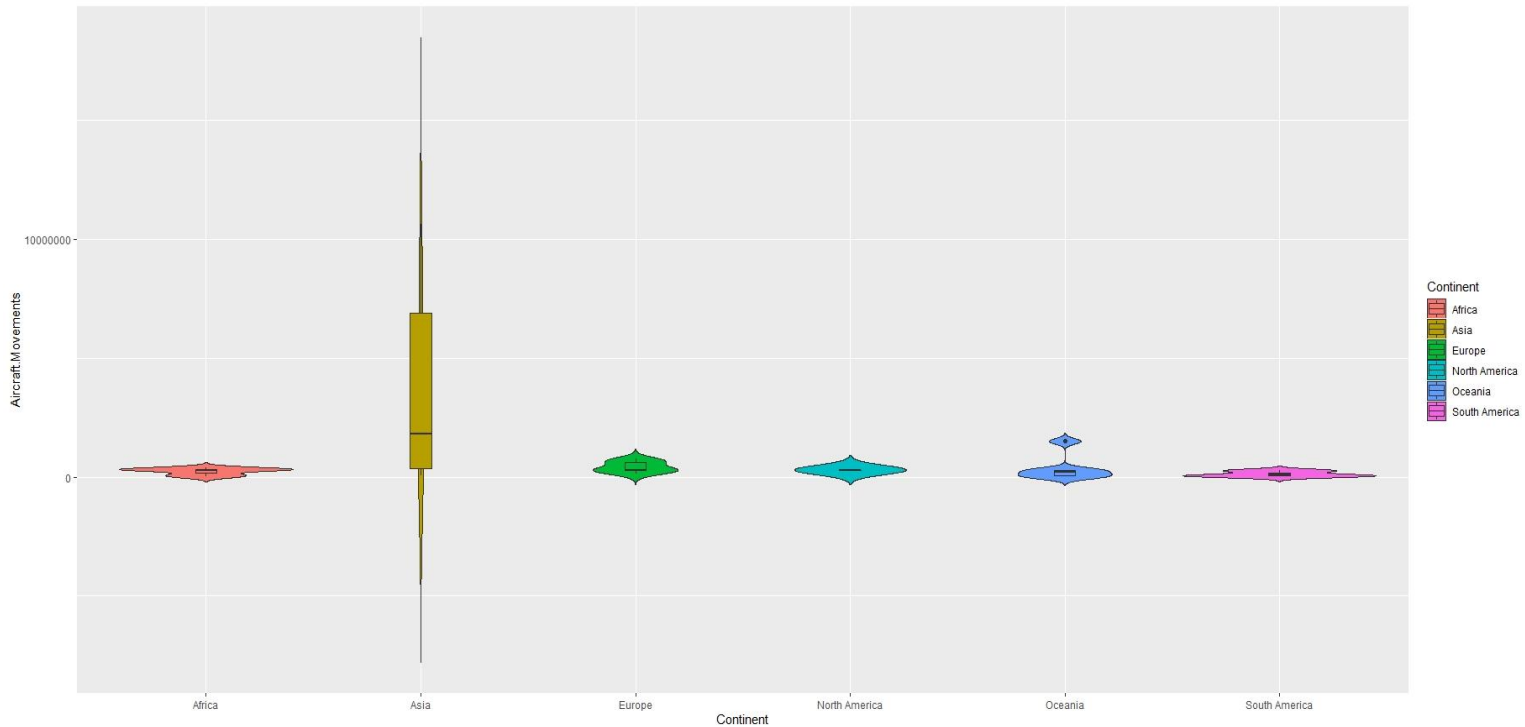


Figure 37: Mean Distribution of aircraft movements per continent

From the violin plot above we can infer that almost all the continents except for Asia has a similar mean distribution of aircraft movements. However, all the continents have a wide shape in the middle which means that the distribution of aircraft movements is concentrated around the median whereas Asia has long tails on either end of the plot which is represented by light yellow and the long tails as already explained suggest that there are states in Asia that have high and low aircraft movements. Let us check with an AOV test regarding our hypothesis and It will be more clear with the help of a p-value.

```
> model4 <- aov(data$Aircraft.Movements ~ data$Continent, data)
> summary(model4)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data\$Continent	5	69119288989139	13823857797828	3.339	0.0178 *
Residuals	27	111798924450626	4140700905579		

```
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Figure 38: AOV test between Aircraft Movements and Continent

With a p-value of 0.01 which is below 0.05, it is clearly evident that our null hypothesis is rejected, which means that at least one continent has a different mean distribution of aircraft movements dataset. Next, we will look at the aircraft movements per population area.

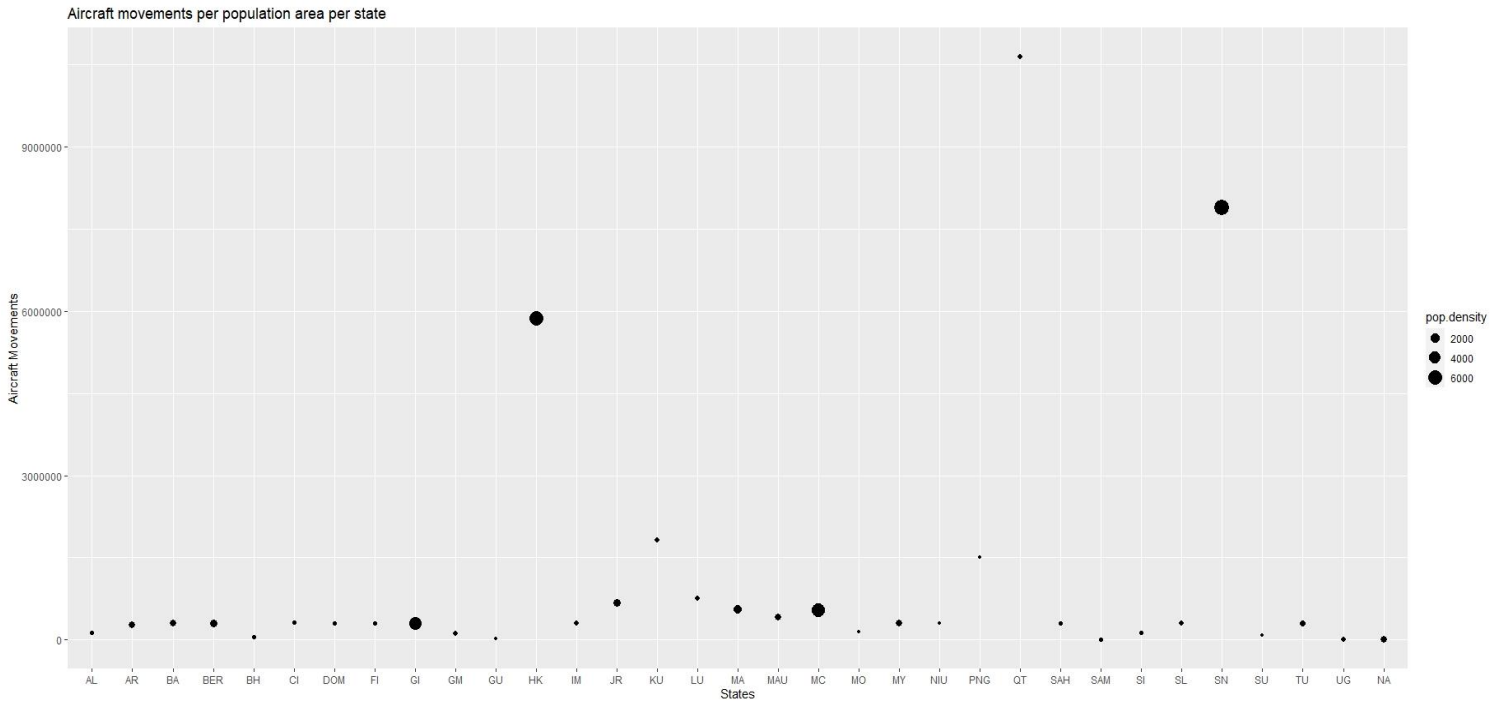


Figure 39: Aircraft Movements per population area in each Continent

From the figure, we can see that all the states have low aircraft movements and population density per area. States such as Singapore and Hong Kong have the highest number of aircraft movements and their population per area is also over 6000. There is one strange country that has a very low population per area but the number of aircraft movements is the highest among all the countries. This means that the passengers arriving in this country are mostly tourists because if they were not tourists and those people are residents of the country then the population per area of the country should be higher. We can also see that there are a lot of countries that have similar aircraft movements and their population per area is also below 2000. The number of states that has a population per area between 2000-4000 is also low. Do the states that have very low aircraft movements mean that tourism in that particular country is low? Does it also mean that the numbers shown for population per area in those states are only due to the residents in that state? The state that has the least aircraft movements and population per area are Guyana(GU), Uganda(UG), and Samoa Islands(SAM) and these 3 states also have similar aircraft movements and population per area.

The above plot clearly does not suggest that the states that have the highest population per area also have high aircraft movements. Next, we will look at how aircraft movements and economy are related which is a very important aspect of any country.

To conclude I would like to say that, on an average Asia has the highest number of aircraft movements and from figure 33 we saw that the aircraft movements are not just for educational attainment. Also, we saw that aircraft movements and population density are not highly correlated.

Next, we will try to answer if the states that have high aircraft movements also have a high economy and if aircraft movement is a significant factor in determining the economy of small states. Firstly, let us plot the visualization of the economy per state.

3.9 Which state has the highest economy and how is economy related to other variables?

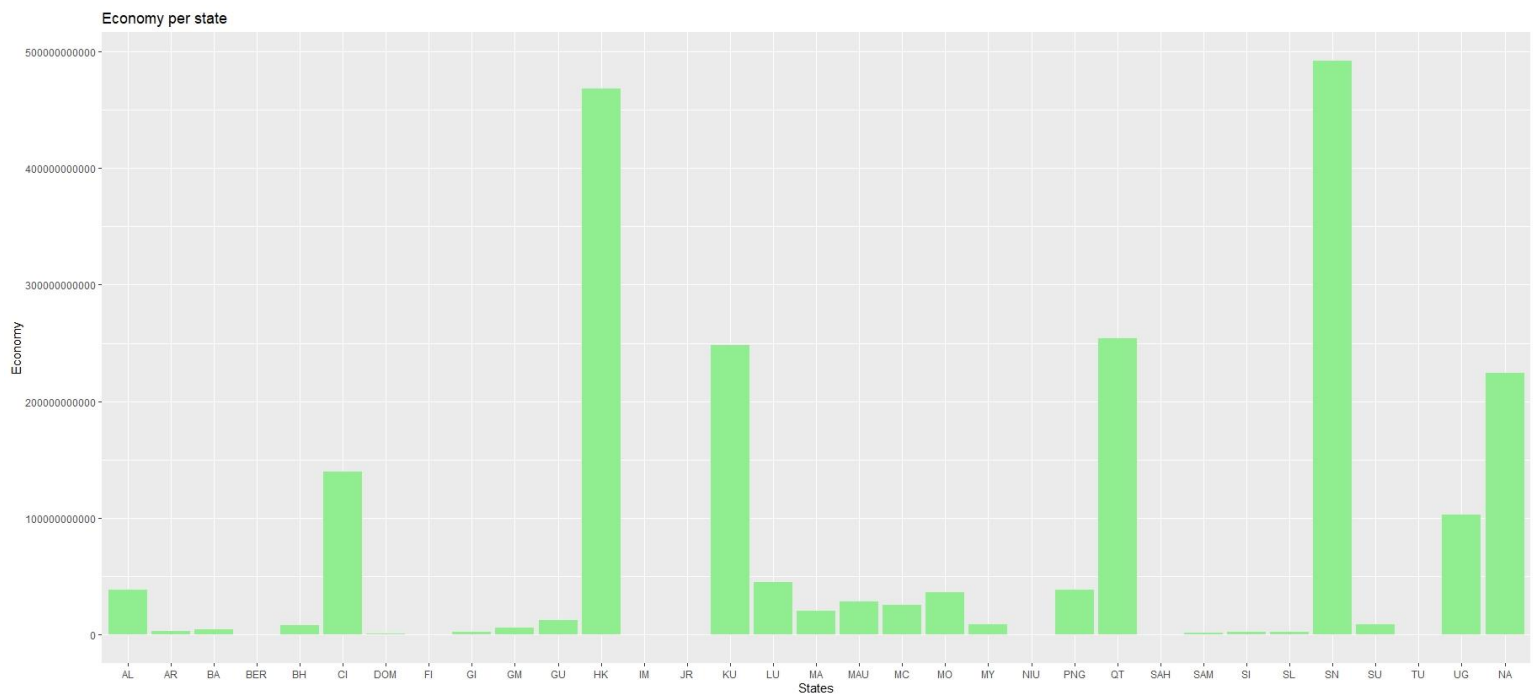


Figure 40: Economy per State

From the above plot, it is visible that Singapore(SN) has the highest economy followed by Hong Kong(HK). Kuwait(KU) and Qatar(QT) have almost the same economy. In terms of lower economies, states such as Luxembourg(LU), Mayotte(MY), Papua New Guinea(PNG), Samoa Islands(SAM), Suriname(SU), and there are many more have lower economies. The states Dominica(DOM), Bermuda(BER), Faroe Islands(FI), Jersey(JR), and Tuvalu(TU) have the least

economy which is not even visible in the above plot. They could be having an economy of probably thousands which is pretty bad, especially with Jersey and Luxembourg which have a decent population and a fair bit of aircraft movements as well. From the above scenario does it mean that states that have the highest economy have the highest aircraft movements and the ones that have low economy have lower aircraft movements? But before we visualize that, let us first try to visualize which continent these states that have lower and higher economies belong to.

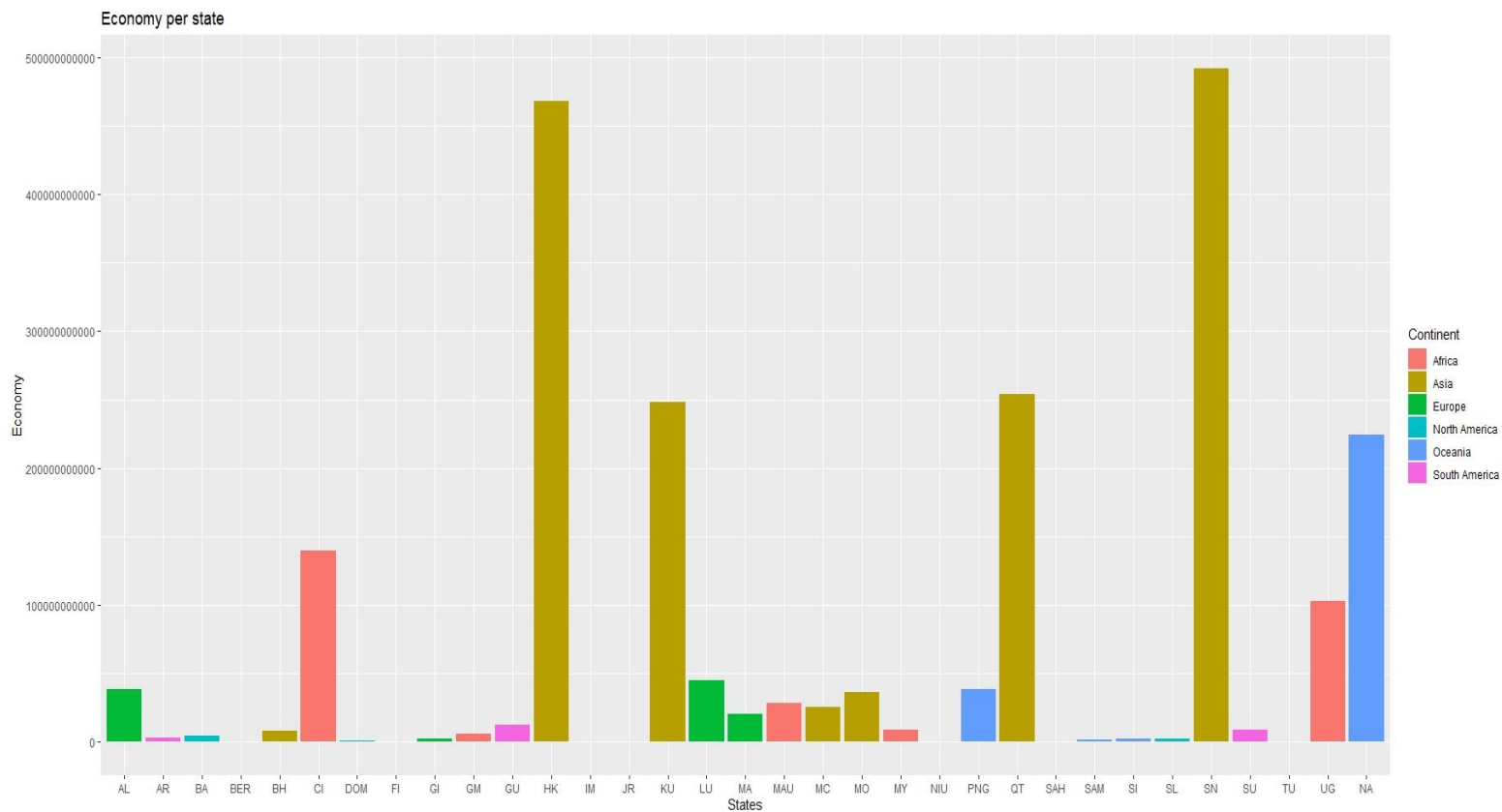


Figure 41: Economy Per state in each Continent

Again just like all the other variables that we have analyzed till now, the economy is also the highest in Asia and is above par as compared to all the other continents. North America and South America have the least economy in each state and because the difference between the largest and least economy is so big, the economy of some states is not even visible. We will try to find the economy of those states by sub-setting the data and plotting for that dataset. Oceania has only 1 state which is Nauru(NA) and which has the fifth highest economy after Singapore, Hong Kong, Qatar, and Kuwait. Africa has a fair bit of economy which is represented by green and it is at least visible which suggests that the economy of that continent is at least better as compared to North America or South America.

Next, we will look at those states that have very low economies and are not even visible in the above plot. I am guessing most of those states are in North America and South America but let us try by visualizing the same.

3.9.1 Which is the state that has the least economic and which continent does it belong to?

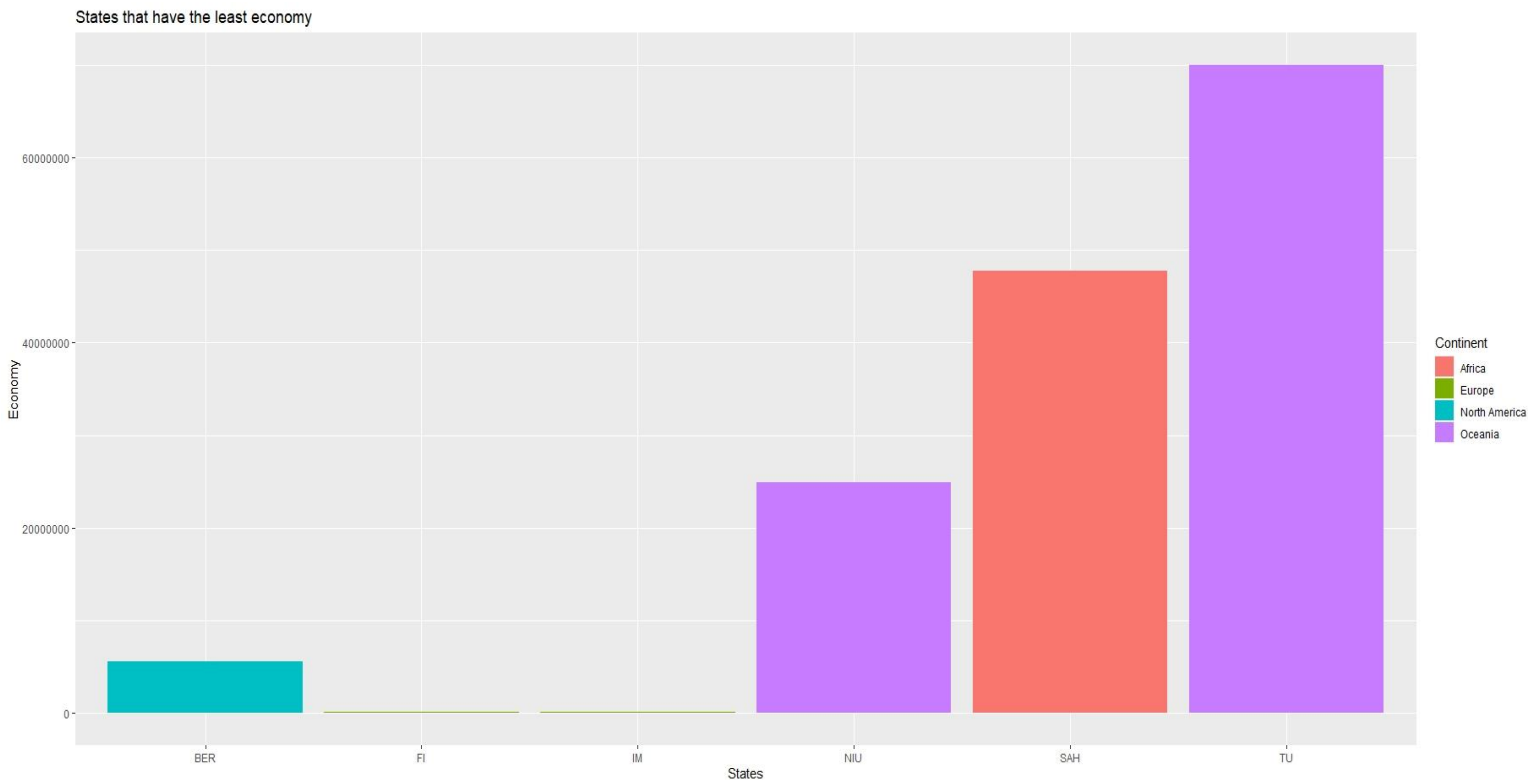


Figure 42: States that have the least economy among all the States

From the above plot it is clear that we were wrong with our assumption, we thought that the states that have the least economy were either from North America or South America, however, we see that there is no state that belongs to South America. There are 2 states in Oceania, 2 states in Europe, 1 state in Africa, and 1 state in North America. The maximum economy in the above plot is just above 60 million which is very less than the starting economy in figure 40 and that is the reason none of the states in figure 41 was visible in figure 42. Hence the above plot gives us a clear picture of what the economy of those states looks like.

Now that we are clear about the economy per state and each continent they belong to, we will look at the economy vs aircraft movements for each state and how they are related.

3.9.2 How important are the aircraft movements for these small islands?

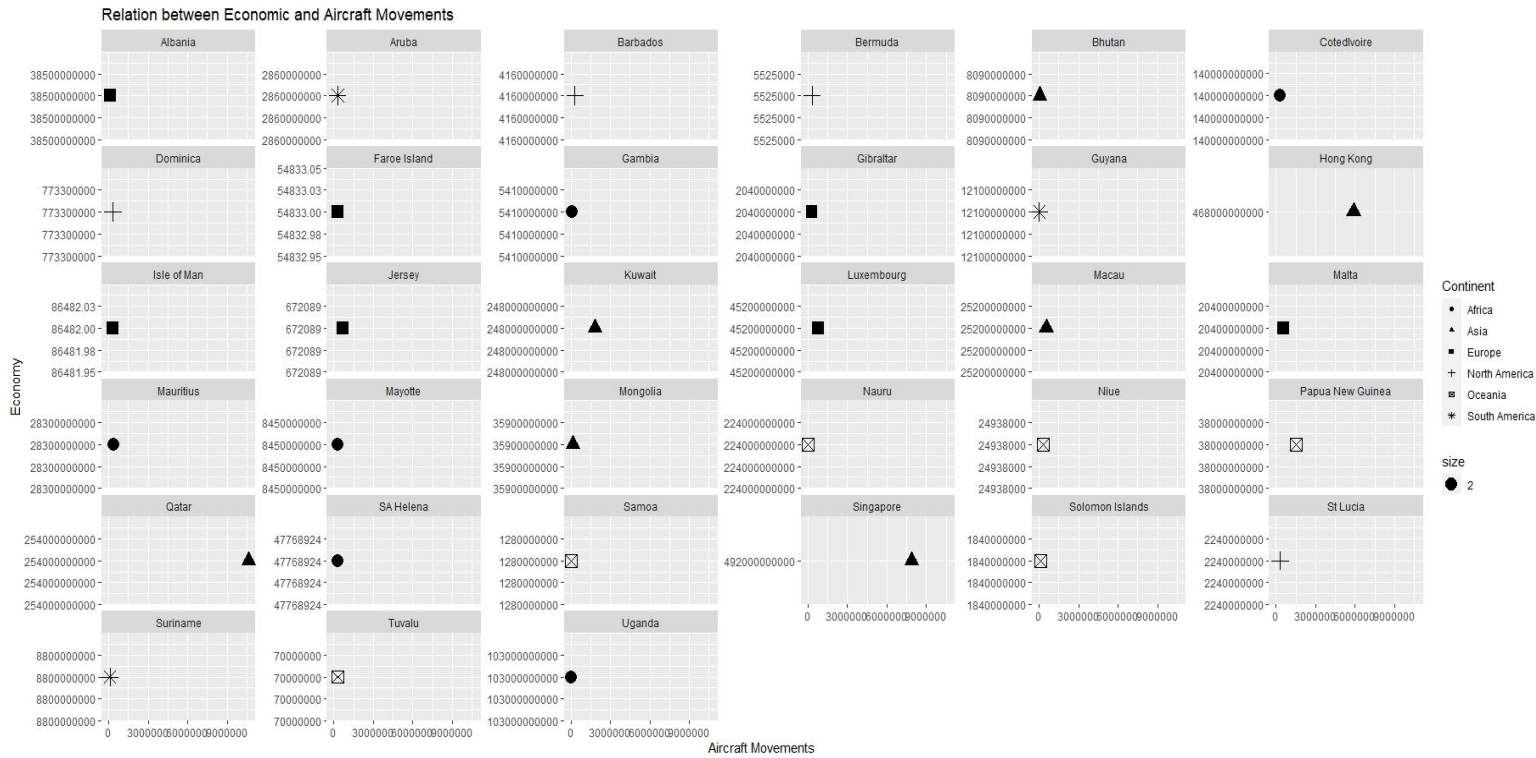


Figure 43: Relation between Economy and Aircraft Movements per state

From the above plot, we can see that states have high aircraft movements and also has a high economy and those states are Qatar, Hong Kong, and Singapore. Those states that have aircraft movements between 0 and 3 million have a lower economy. If we compare continents wise we see that most of the states that have both high economy and aircraft movements are from Asia. This means that most of the states that tourists like are in Asia. However, we can also see that states that have aircraft movements between 0 and 3 million have a very low economy of probably around thousands. This could also mean that these states have lower tourists and their economies are not largely dependent on tourists. Also, since Singapore, Qatar and Hong Kong have high aircraft movements and more and more people are visiting the country, it means that their economy could high because of that. In all the other continents, some states have higher and some states have lower aircraft movements, and economy. For an instance, if we consider Europe which is represented by a square has 3 states Jersey, Isle of Man, and Faroe island which have a low economy and aircraft movement whereas Malta, Luxembourg, and Gibraltar have a high economy and aircraft movement. Hence, we can say that there are some states that activities or places for tourists in each continent and for those states tourism is an important factor for their economy and

there are states where there are not many interesting activities or places for the tourists to visit. From figure 18, we see that aircraft movements and economy are highly correlated and it is proven from the above plot. Now that we have seen the relationship of aircraft movements with other variables before I conclude we will have a look at the official language that is spoken in most of the states and try to understand which is the most common language that is used in the majority of the states or continents.

3.10 What is the official language of each state?

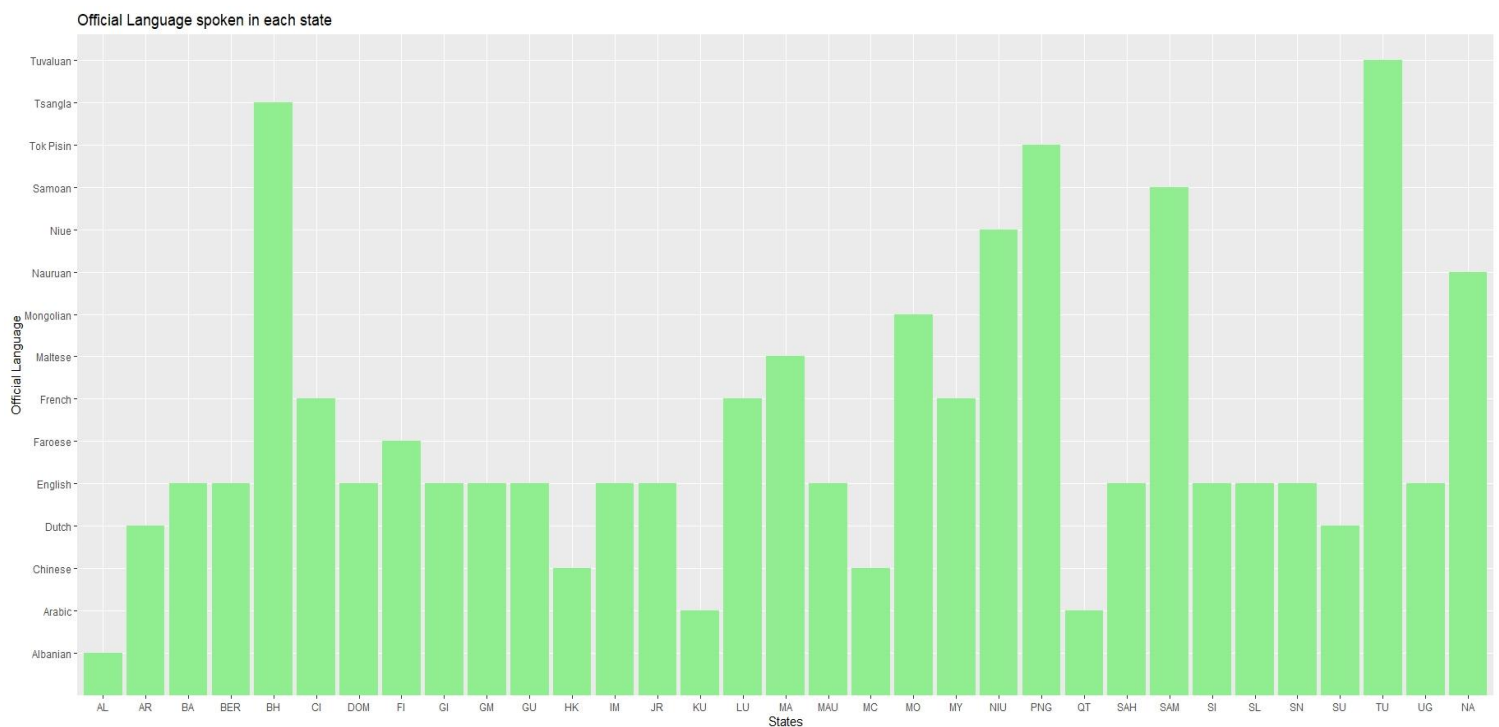


Figure 44: Official Language spoken in each State

From the above plot, we can see that there are about 15 languages that are spoken across different states and are represented along the y-axis. Also, Almost 50% of the states have English as the official language. This means that English has been spoken in almost 14 states by the majority of the population and not by every single person in that state. All the other states have their own languages where that language is not spoken by more than 2 states. After English, French is spoken in 3 states, and Chinese, Arabic, and Dutch are spoken in 2 states. The rest of the states have their own languages and it varies from state to state. Further, in this paper, the next chapter focuses on Conclusions from the analysis that we have conducted until now and the challenges that I have faced in accomplishing this project.

3.10.1 What is the language spoken in the state that has the highest population?

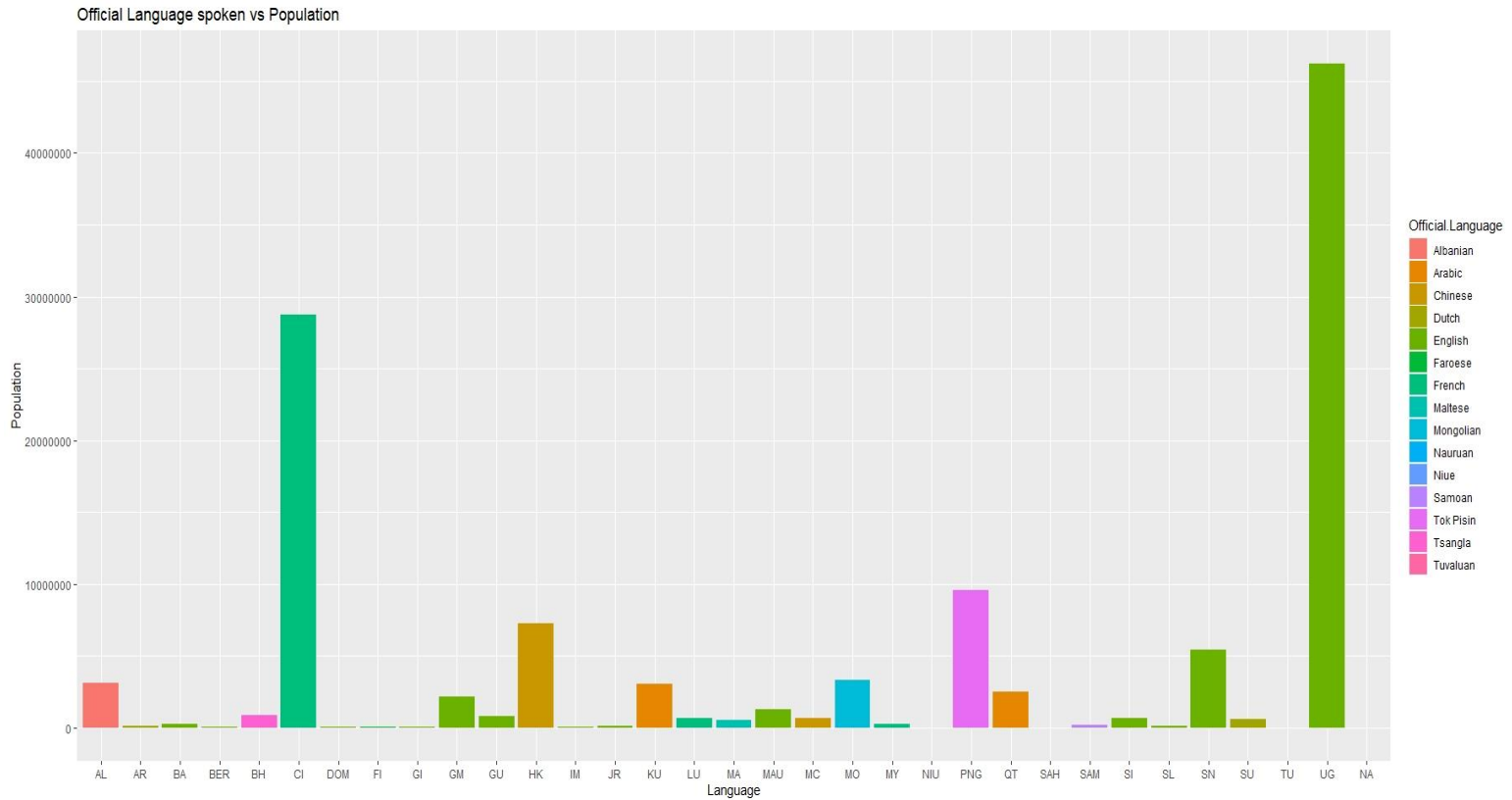


Figure 45: Official Language Vs Population

From the above figure we saw that English is the language that is mostly spoken by most of the states but which could be the state that has the highest population and that speak English. From the above figure we can see that Uganda is that state which has the highest population which means that majority of the population of Uganda speaks English. Those states that have very less population which are not visible in the above plot but we know the language that those states speak in figure 44. After Uganda, CI is the state that has the highest population and the majority of the population in CI speak French. Other than English, French, Dutch, most of the other states have their own languages.

4 Conclusion

We have explored many aspects of this work in the preceding chapters, beginning with what we meant by little islands. What gives them that name? The conversation then went on to the data that I had collected and whether there were any missing values or anomalies. What are the various approaches to dealing with missing numbers and outliers? I've also discussed the way I used to deal with missing values, as well as why I didn't use any method to deal with outliers. Following the data description, we proceeded to analyze our data by visualizing it. A brief summary is provided below.

We began by visualizing and attempting to comprehend the distribution of the numerical variables and discovered that almost all of them are skewed to the left or right. Then we moved on to identifying the number of states in each continent and the names of the states in each continent, where we discovered that Asia and Europe have the most states. Furthermore, we attempted to comprehend the population increase in each state in comparison to the population in 1972, and we discovered that Uganda has the largest population growth, while Niue has the lowest. Then we looked at how the numerical factors in our data correlated and discovered that the economy and aircraft movements are strongly correlated, whereas infant mortality and life expectancy are highly negatively correlated. We utilised the ANOVA test to evaluate our hypothesis, and I defined it as well. Then I added a new column to the dataset called population density, which is people per area, and tested the distribution of the states' mean population density. Then we moved on to the fascinating analysis of life expectancy and mortality rate, and we further examined the link of both life expectancy and mortality rate with all of the other variables. When we visualised life expectancy distributions, we saw that Asia, Africa, and Oceania had both high and low life expectancy distributions. We noticed the same thing with the infant mortality rate. Following that, we observed how educational attainment is related to different other qualities, and because we saw that educational achievement and economy are somewhat correlated, I posed the question as to why this association exists. We saw that education benefits the country as a whole in the cases of Singapore and Hong Kong.

We then moved on to the most essential component of our data, the aircraft movements. We discovered that Asia had the biggest number of aircraft movements, and we inferred that Asian states are the most popular with tourists. We also looked at the distribution of aircraft movements per population area in each continent, and Hong Kong and Singapore were among the cities with the most aircraft movements. Furthermore, we comprehended the economies of each state on each continent. Later, we observed the most intriguing study since they were substantially positively associated, and we could deduce that in this situation, states in Asia have greater values on average than states in other continents. Finally, before ending, we looked at the language spoken by the majority of people in each state.

4.1 Challenges Faced

Whenever someone is doing a project it is obvious that they will go through challenges in order to achieve the objective. The challenges will be different for different people, I would list some of the challenges that I faced.

- **Choice of Language:** With so many different languages available for data analysis such as Python, MATLAB, R, and many more, I found it difficult to choose between them. I looked at the various advantages and disadvantages like how easy is the language to understand, and whether a third person is able to understand what I am doing and I decided to go with R as my preferred choice of language.
- **Data Collection:** Since this particular project did not have a ready-made dataset, I had to surf the internet and collect the data manually as there were not many countries that had just 1 airport on today's date. This part of data collection was a tedious process.
- **Data Cleaning:** Since my data contained a lot of missing values that have already been discussed, I had to choose a method to deal with it and I had to also decide if I had to deal with the outliers in my case.

4.2 Steps to run the R file

- Install R studio and install the necessary libraries such as `install.packages("package name")`
- Once installed, store the dataset and the R file in the same folder
- Set your working directory where you stored the dataset and the R file using `setwd()`, in my case it is in MyProject hence I have given the path of MyProject in my `setwd()`.
- Then go on executing it line by line.
- You can also run 3-4 lines at once depending on if they are of the same plot. I have given clean gaps between each statement in the R file which indicates the statement belonging to each plot. For instance if there are 2 lines consecutively then they could be executed together and it is similar with 3 lines consecutively and so on, if there is only line of plot then they have to be executed individually.
- When there is a warning which says run it solving `dev.off()`, please go to the line where `dev.off()` is and run it from there onwards.
- All the files which are required are included in the directory MyProject.zip and the main R file is named as Dis.R in the directory MyProject.

- Also, Aviation_islands_original is the original CSV as depicted in figure 1 and clean_missing_aviation is the CSV which has been obtained after dealing with the missing values. I have used the clean_missing_aviation dataset for my analysis. If needed you can use the original dataset and impute the missing values, and carry forward the analysis.

5 References

- S, C., S, S. & M, R.-R., 2015. Multiple imputation: a mature approach to dealing with missing data. *Intensive care medicine*, 41(2), pp. 348-350.
- Aviation, U., 2018. *Uniting Aviation*. [Online] Available at: <https://unitingaviation.com/news/economic-development/aviation-is-at-the-heart-of-sustainable-development-for-small-island-countries/> [Accessed 3 August 2022].
- Briguglio, L., 1995. *Small island developing states and their economic vulnerabilities*, s.l.: World Development.
- Cheema, J., 2014. A review of missing data handling methods in education research. *Review of Educational Research*, 84(4), pp. 487-508.
- clark, W. r., 2018. *Small Island Developing States: Aviation Essential to Achieving UN Global Goals..* [Online] Available at: <https://www.linkedin.com/pulse/small-island-developing-states-aviation-essential-un-raillant-clark/> [Accessed 3 August 2022].
- Easterly, W. & Kraay, A., 2000. Small states, small problems? Income, growth, and volatility in small states. *World development*, pp. 2013-2027.
- Frost, J., 2021. *Guidelines for removing and handling outliers in data*. [Online] Available at: <https://statisticsbyjim.com/basics/remove-outliers/> [Accessed 10 08 2022].
- Gill, M., 2014. *Aviation's Important Role in Small Island State Economies*. [Online] Available at: <https://aviationbenefits.org/newswire/2014/09/aviation-s-important-role-in-small-island-state-economies/ll-island-states-building-a-sustainable-future-through-air-transport/> [Accessed 3 August 2022].
- Gill, M., 2014. *Small Island States: Building a Sustainable Future through Air Transport*. [Online] Available at: <https://aviationbenefits.org/blog/2014/09/small-island-states-building-a-sustainable-future-through-air-transport/> [Accessed 3 August 2022].
- Heijden, V. D., Stijnen, G. J. & K, . M. G., 2006. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10), pp. 1087-1091.
- Hindmarsh, J. H., 1996. How do we define small states and islands? A critical analysis of alternative conceptualizations.. p. 36.
- macola, I. g., 2020. *How Has Covid-19 Affected Islands That Depend on Aviation?*. [Online] Available at: <https://www.airport-technology.com/analysis/covid19-affected-islands-depend-aviation/> [Accessed 3 August 2022].

PhD, S. Y. & MD, G. B., 2016. Outliers. *The Southwest Respiratory and Critical Care Chronicles* , 4(13), pp. 52-56.

Ross, A. & Willson, V., 2017. *One-way anova. In Basic and advanced statistical tests* , Rotterdam: SensePublishers.

R, S. V., 2021. *End-to-End Introduction to handling missing values*.

[Online] Available at: <https://www.analyticsvidhya.com/blog/2021/10/end-to-end-introduction-to-handling-missing-values/> [Accessed 2022 08 10].

S., V. B., 2018. *Flexible imputation of missing data*. s.l.:CRC press.

The World Bank Data Group, 2021. *Air Transport, passengers carried*.

[Online] Available at: <https://data.worldbank.org/indicator/IS.AIR.PSGR> [Accessed 3 August 2022].

The World Bank Data Group, 2021. *Educational attainment, at least bachelor's or equivalent, population 25+, total (%)*. [Online] Available at:

<https://data.worldbank.org/indicator/SE.TER.CUAT.BA.ZS?end=2021&start=2000> [Accessed 3 August 2022].

Wikipedia, 2022. *Welcome to Wikipedia*. [Online] Available at:

https://en.wikipedia.org/wiki/Main_Page [Accessed 3 08 2022].

