2021

DATA VISUALIZATION PROJECT

USING THE WORLDPOP DATASET

JAYASMITA CHAKRABORTY

Table of Contents

PART 1: VISUALISATION

INTRODUCTION

The data contains information on population, area of a country, GDP, the incoming international travellers, the number of flights operating from the country, and revenue generated from tourism.
With the data, let's try to answer two questions:

1. What is the relation between population, GDP and tourism revenue?
2. Which is the cheapest, but tourist-friendly, destination to visit for your next vacation?

FEASIBILITY OF THE DATA

The data contains a lot of missing values. For some nations there are many attributes that are entirely missing.
For example, the attribute **`hotels`** is missing from countries 3,6,7,11,14,15,17-27.
Similarly, the **`Flights-WB`** is also missing from countries 4,6,7,9-11,15,18-21,23,24. Even with pre-processing, it will be difficult to use these attributes.

The above-mentioned attributes are required to answer question 2. So, data imputation and pre-processing is required.

DATA PRE-PROCESSING

1. Predictive Modelling

Imputing the data requires predictive modelling. It's not accurate to use methods such as mean, median or mode to generate values for the missing data. This is because the missing attributes of one country are not dependent on the missing attributes of another country.

A better method would be to apply classification or regression trees to correlate the attributes of one., single nation and predict the missing attribute values for that nation.

Package **`mice`** is appropriate for achieving such data imputation.

2. Split, Apply, Combine

Using packages **`plyr`** and **`tidyr`** to summarise the mean value of the key attributes such as **`pop, gdpnom, ovnarriv, receipt`**, etc. will help to compare between the nations and find out patterns.

3. Normalisation

For answering question 2, I have used data normalising to scale the values of attributes like **`receipt, Flights-WB, ovnarriv,`** etc. to yield comparable values that can be plotted on the same graph without the data looking skewed. I have used unit length normalising

Question 1: What Is The Relation Between Population, GDP & Tourism Revenue?

Step 1: Analyse the Distribution of Population for the countries in the dataset

One of the first reasons to analyse the three parameters (Population, GDP and Tourism Revenue) before anything else is because for most countries, these 3 attributes have the least missing values in the dataset. This means that the predictive data imputation used will give more accurate results and less data imputation is required anyway for these attribute values.

Figure 1: Population per country

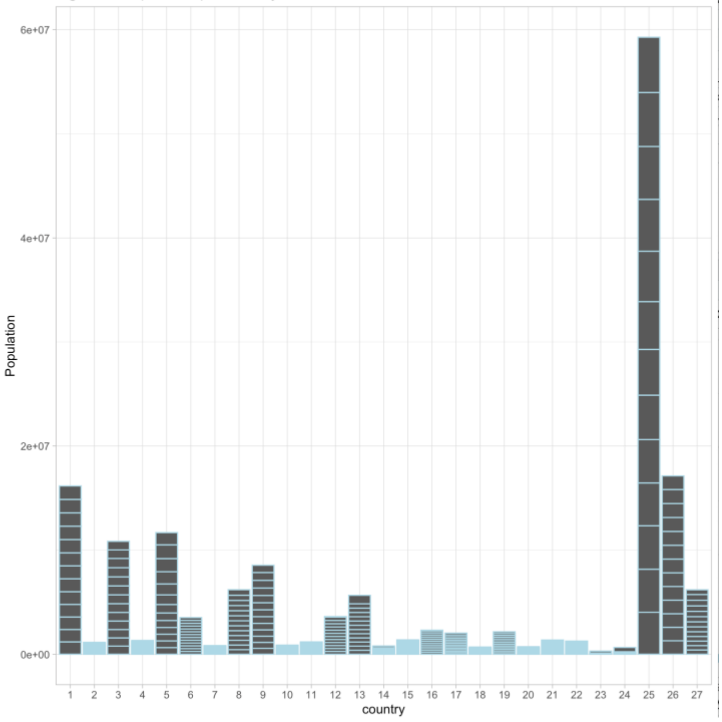
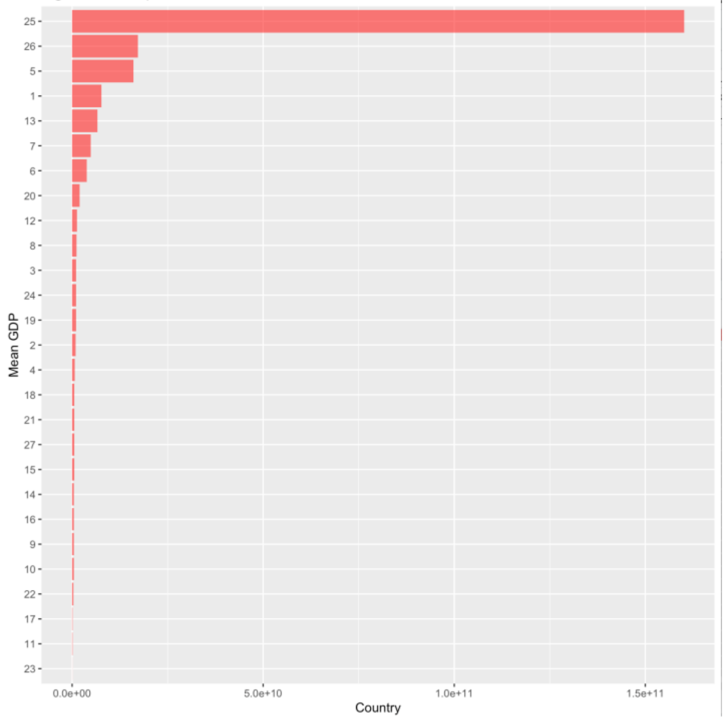


Figure 2: GDP per Nation



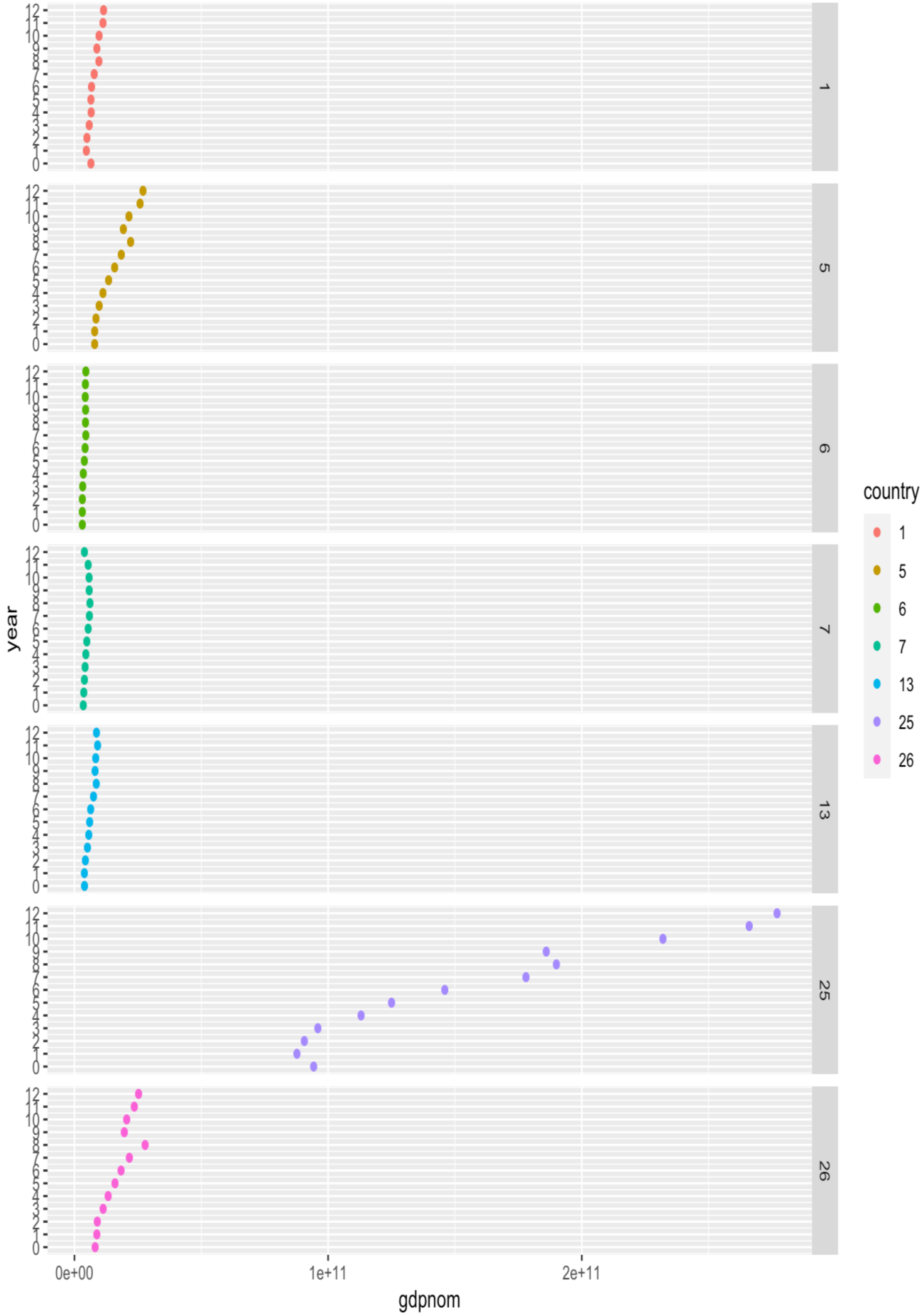
From the above figures, Country 25 (Singapore) stands out and shows a huge year-on-year increase in population.

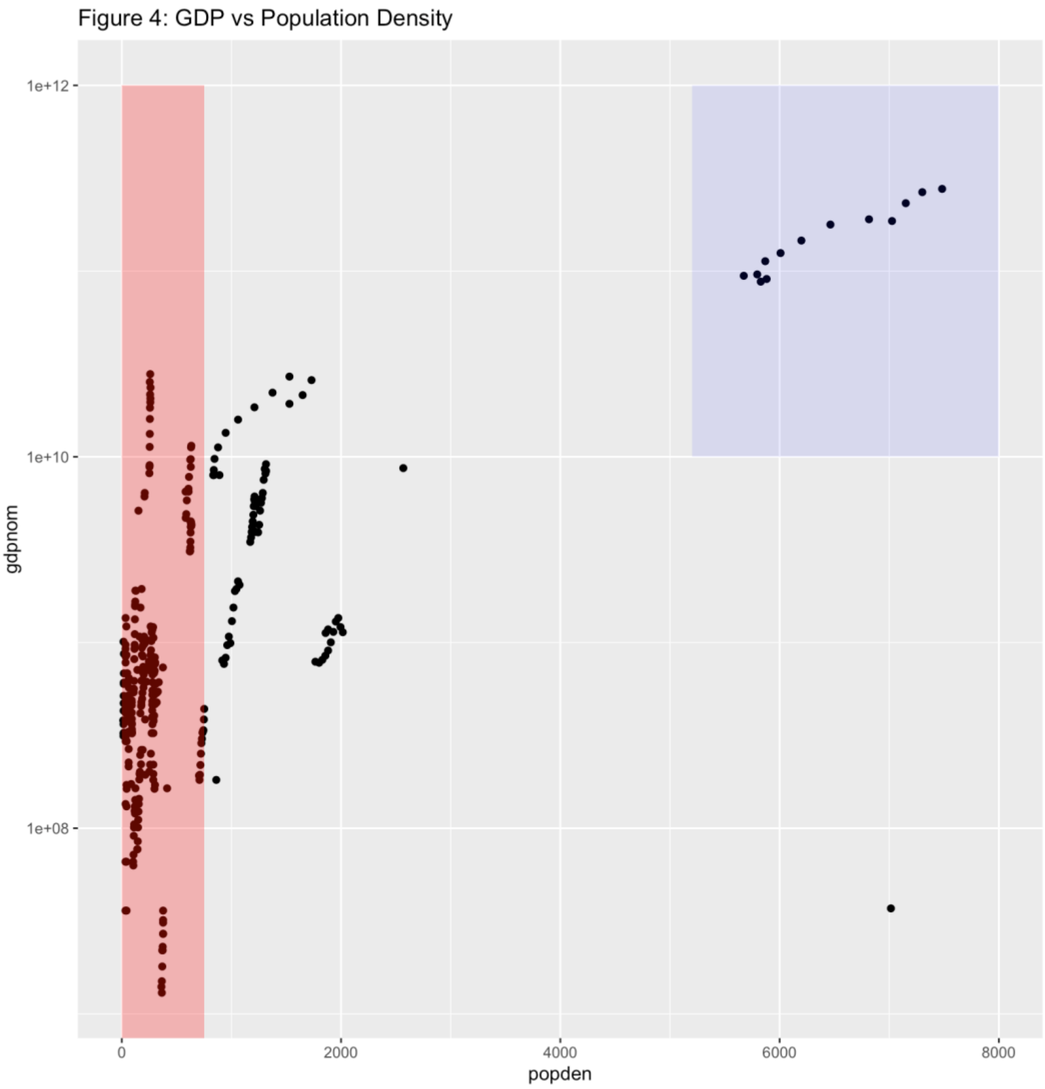
Also, shown by the graph below, Singapore is having the maximum GDP.
Figure 3 shows that the countries with increasing GDP also correspond with the countries with higher population in shown in Figure 1.
Does this imply that higher population leads to higher GDP?

Step 2: Introducing a new metric population density = population/area in sq. km.

Using the above metric to observe patterns between population and GDP – illustrated in Figure 4.

Figure 3: 13 Year GDP Trend





There are two parts to this plot which are significant, which are annotated in blue and red.

1. <u>Blue</u>: Increase in population density increases the GDP rapidly
   This trend is only applicable to country 25 – Singapore
2. <u>Red</u>: Increase in GDP for a certain **constant** population density

Step 3: Analysis for GDP & Tourism Revenue for Country 25 where population density and GDP Increase Proportionally

The scatter plot (Figure 5) shows that GDP is continuously increasing for country 25 - Singapore as population density increases.
How much has tourism increased over the 13 year time frame? Is it dependent on GDP growth?

Figure 6 shows that as GDP increased, the tourism revenue also increased. Maybe it contributed to the increase in GDP. Tourism revenue flattens for year 12. What happened in year 12?

The most important factors in the dataset that have a direct impact on tourism revenue are – **`ovnarriv, dayvisit, arram , arreur, arraus`**. Figure 7 shows how these factors have been changing over the thirteen year time frame.

Figure 5:Population Density vs. GDP (Year-on-Year)

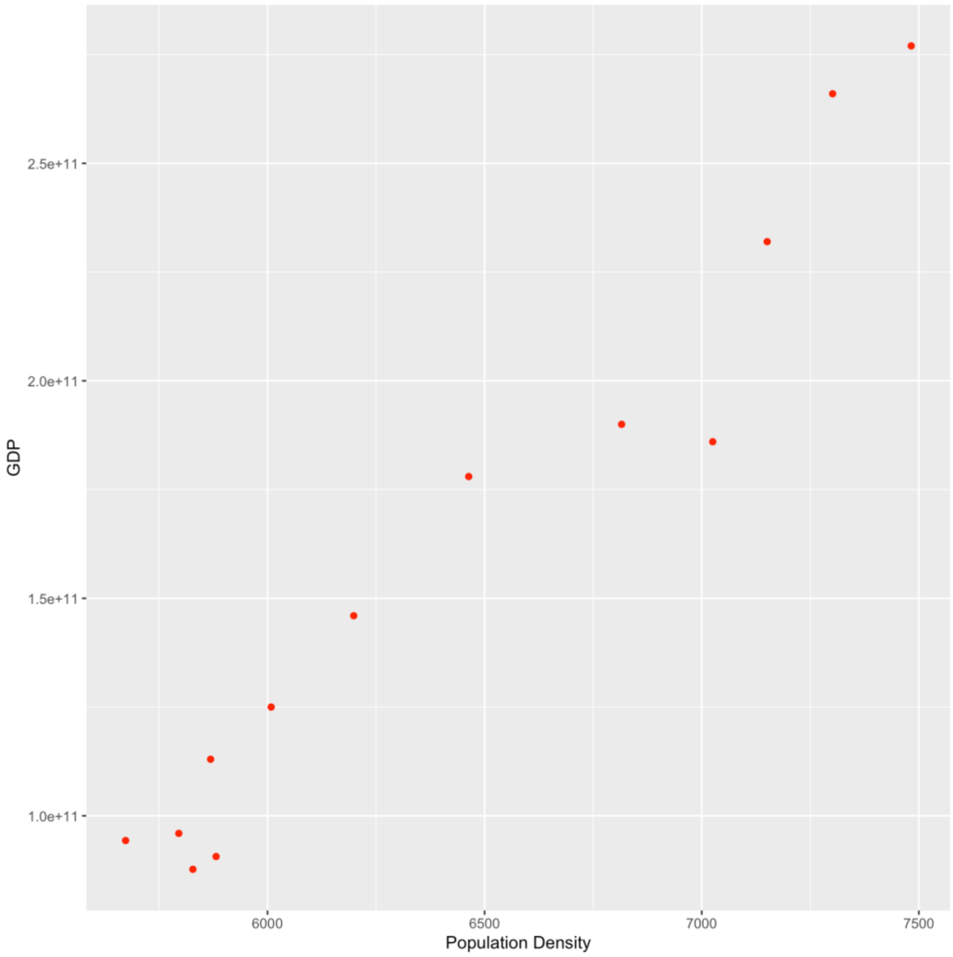


Figure 6: Tourism Revenue vs. GDP

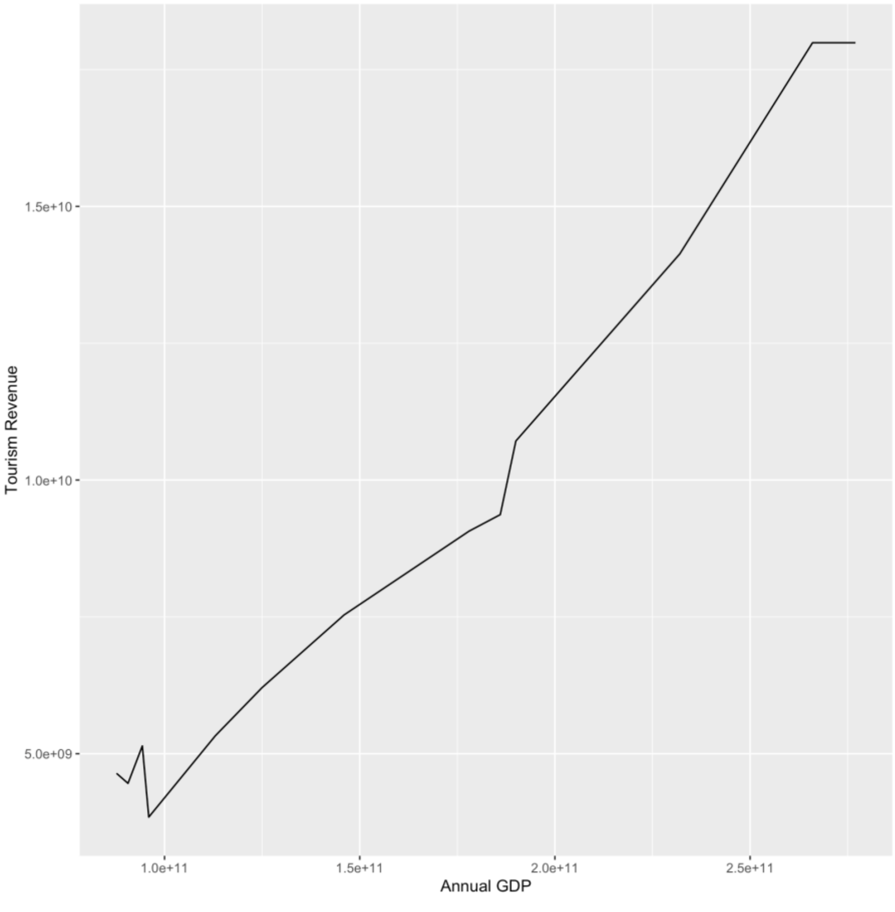
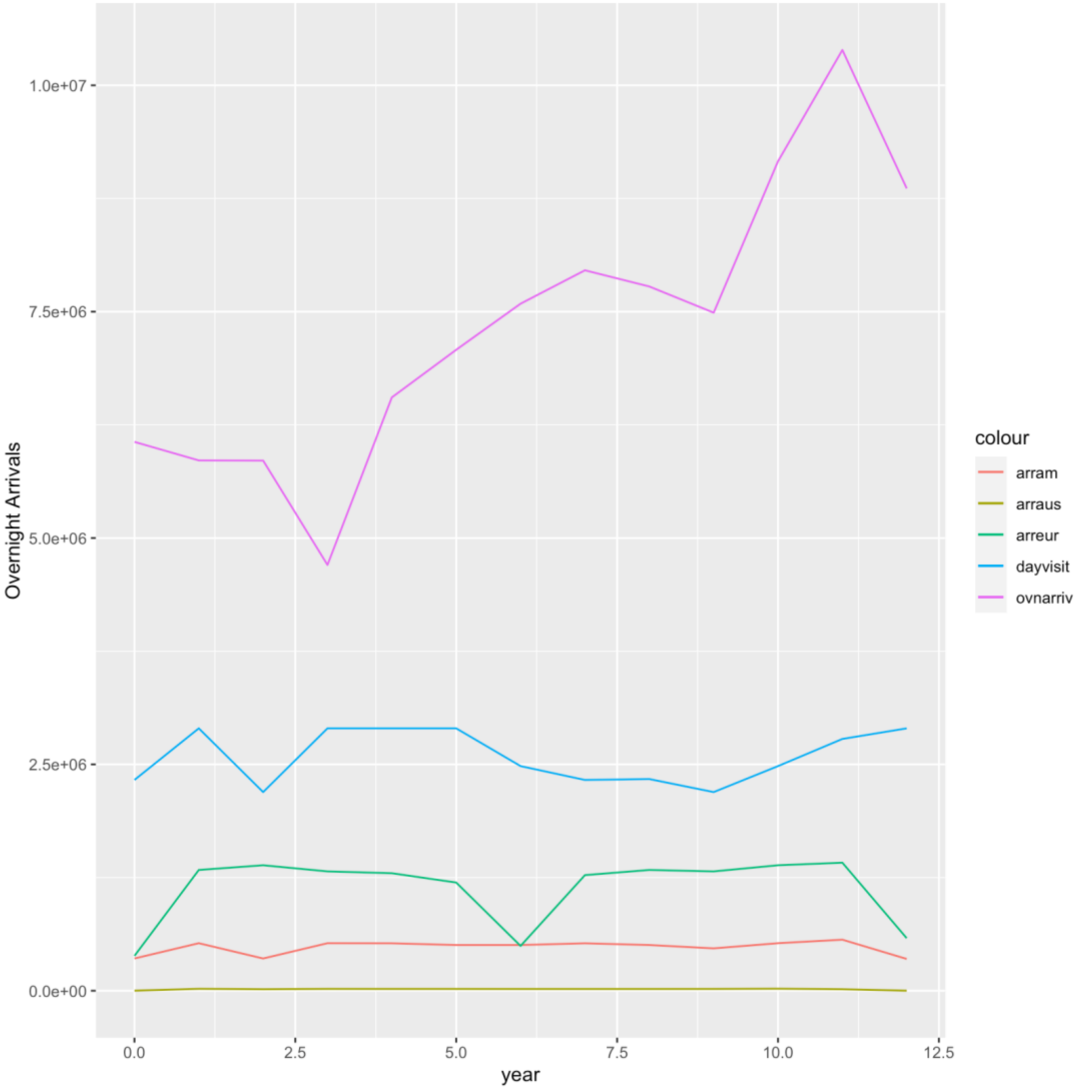


Figure 7: Factors Affecting Tourism Revenue



From the above graph, overnight arrivals, arrivals from USA and Europe reduced in year 12. But there was an increase in the number of visit days. Perhaps this balanced out the decrease in the incoming passengers and flattened the tourism revenue curve for year 12.

Step 4: Analysis for GDP & Tourism Revenue for Countries with Constant Population Density

Figure 8 shows a magnified view of the part annotated in blue in Figure 4.
Most of these countries show a very rapid increase in GDP for a constant population density. The most noticeable increases are for countries 23(Tuvalu), 24(Palau) & 26(Trinidad & Tobago), as shown in Figure 8.1. It seems that country 24(Palau) has shown the largest increase in GDP in the 13 year time frame.

NOTE:
Country 24 had no "gdpnoom", "dayvisit" and "arrus" information in the original dataset.
Most of the visualizations below is from imputed data, hence may be inconclusive.

Figure 8: GDP Increase for Near Constant Population Density

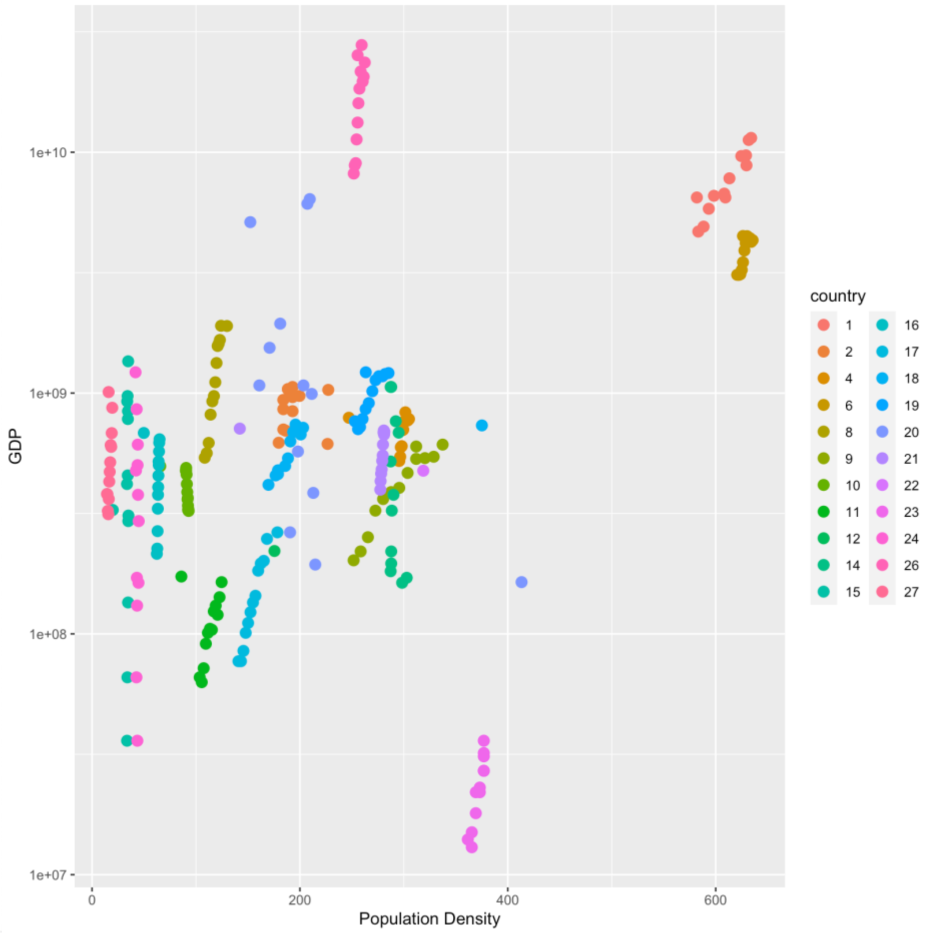
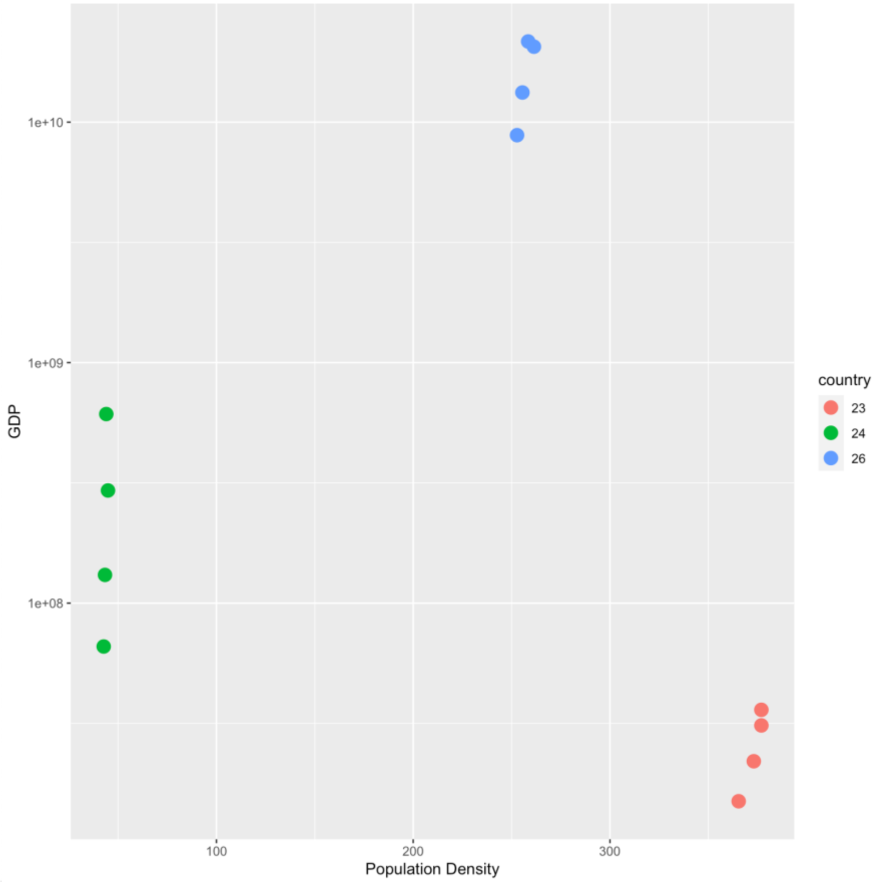


Figure 8.1



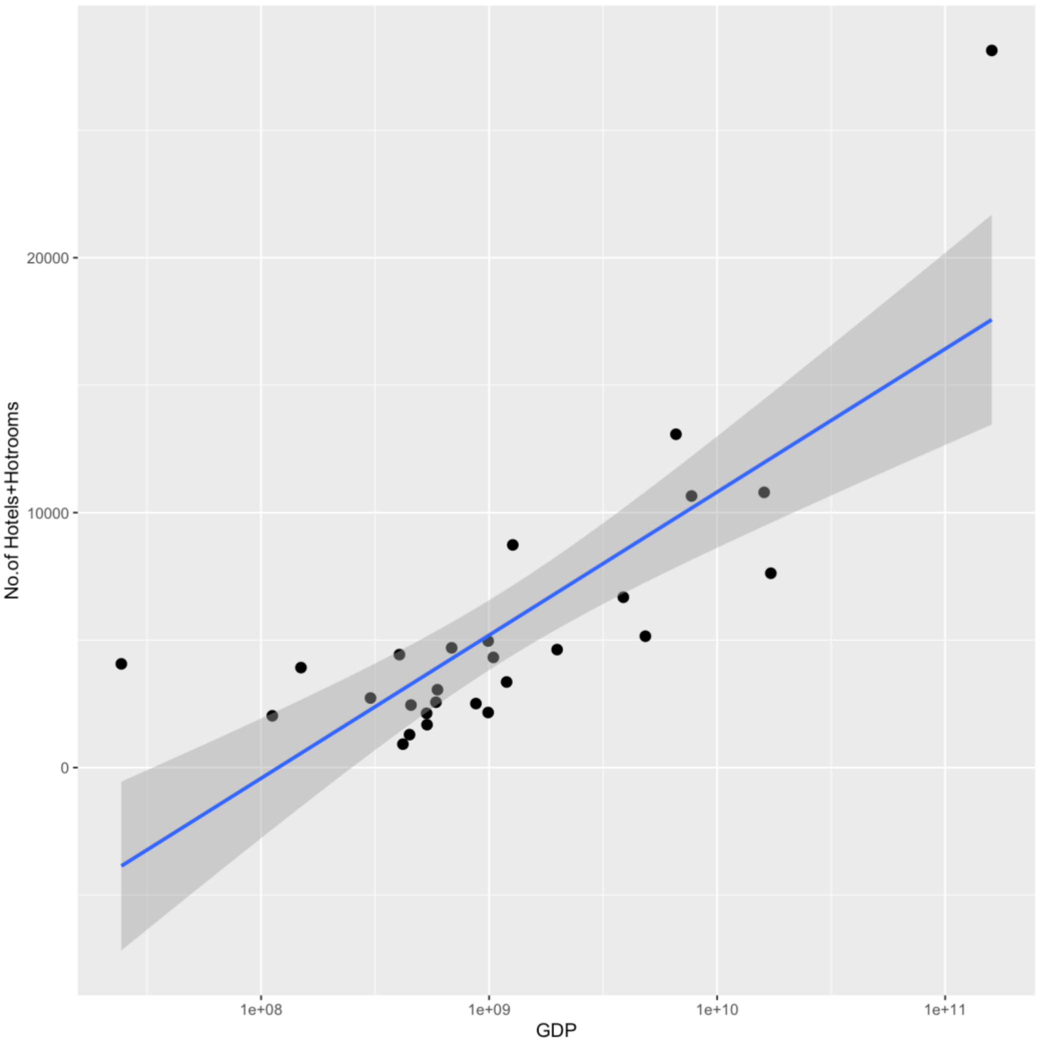
Question 2: Which is the Cheapest, but Tourist-Friendly, Destination for Your Next Vacation?

For answering this question the selected country should be inexpensive, but also has good infrastructure for tourists, such as regular flights, hotels, etc. GDP can be considered as a measure to check if the country has the capacity to spend on tourism infrastructure.

ASSUMPTION: Countries with high GDP will tend to spend more on tourism infrastructure
Can we verify this assumption graphically?

After aggregating the data from the whole dataset, we get Figure 9 below which shows that a high national average GDP(for 13 years) does imply greater number of hotels and hot-rooms in that country.

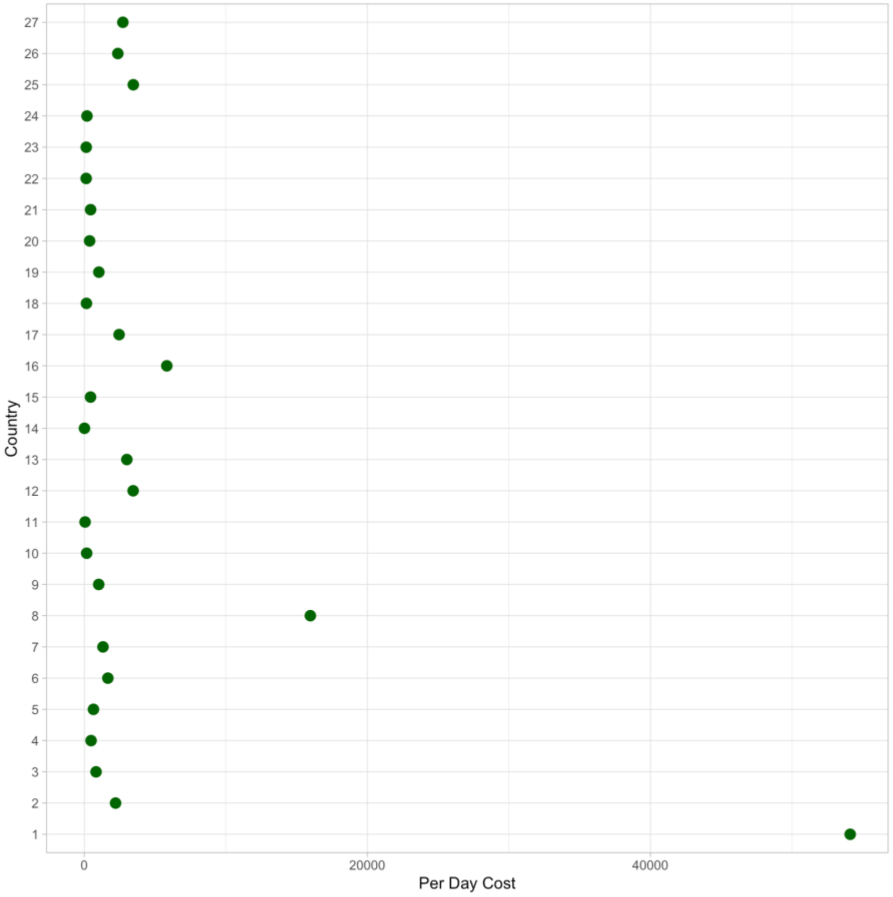
Figure 9: No. of Hotels + Hotrooms vs. National GDP



Step 1: Finding the per day cost

This can be achieved by introducing a new metric total <u>tourism revenue/ number of visit days i.e. receipt/dayvisit.</u>
To reduce the effect of data imputation and missing values, using average of the 13 years' timeframe is better. Figure 10 below shows the per day cost of each country in the dataset.

Figure 10: Per Day Cost in Each Country



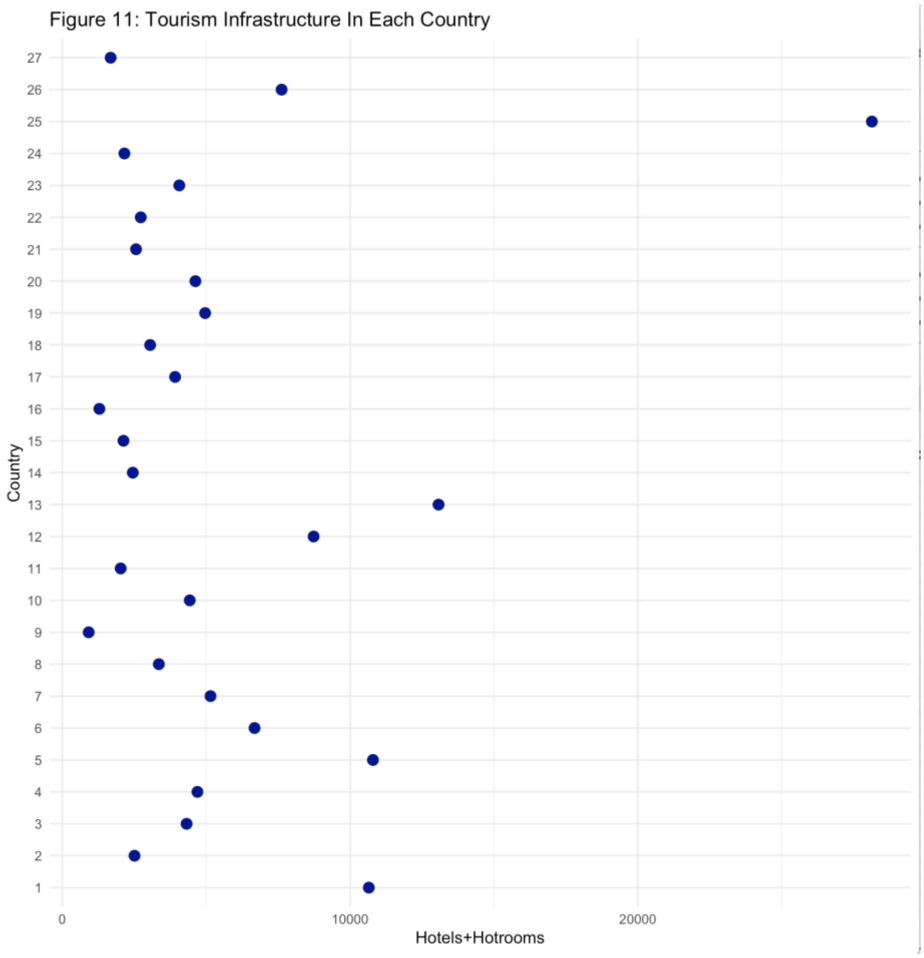
The most expensive countries to visit are:
1. Country 1 – Mauritius
2. Country 8 – Cape Verde
3. Country 16 – Samoa

The other countries are more or less equivalent in their per day expenses, and a lot lower.

Step 2: Analysing tourism infrastructure (in terms of hotels & hot rooms)

From Figure 11 below, we can see that country 25 (Singapore) has the largest number of hotels. Followed by country 13 (Malta), country 5(Bahrain) , country 1(Mauritius), country 12 (Maldives) and country 26(Trinidad & Tobago). The remaining countries don't seem to have a lot of options for hotels and hot rooms.

It seems that tourism infrastructure is most developed in country 25 -Singapore.



Step 3: Accessibility (in terms of flight options)

Countries with more number of flights are more easily accessible and guarantees more flight options and regular flights.

Figure 12 below shows again, that country 25(Singapore) has the most flights. Followed by countries 5(Bahrain), but still a lot less than country 25. The other countries with large number of flights are countries 26(Trinidad & Tobago), 3(Antigua & Barbuda) & 9(Comoros).

Step 4: Check the number of international arrivals to assess popularity of the country

From figure 13 in the next page, it seems that countries 25(Singapore) & 5(Bahrain) have the most international passengers from flights, followed by country 26(Trinidad & Tobago) and 1(Mauritius).

Can we come to direct conclusion that these countries are most popular for tourists?
No, because not all flight passengers come for tourism, some may come for work.
It's better way would be to compare the tourism revenue (attribute: "receipt") of both these nations to check if indeed these countries are popular tourist destinations.

The analysis for Singapore (country 25) is complete and the relevant graph is Figure 6.

Figure 12: Flights in the Countries

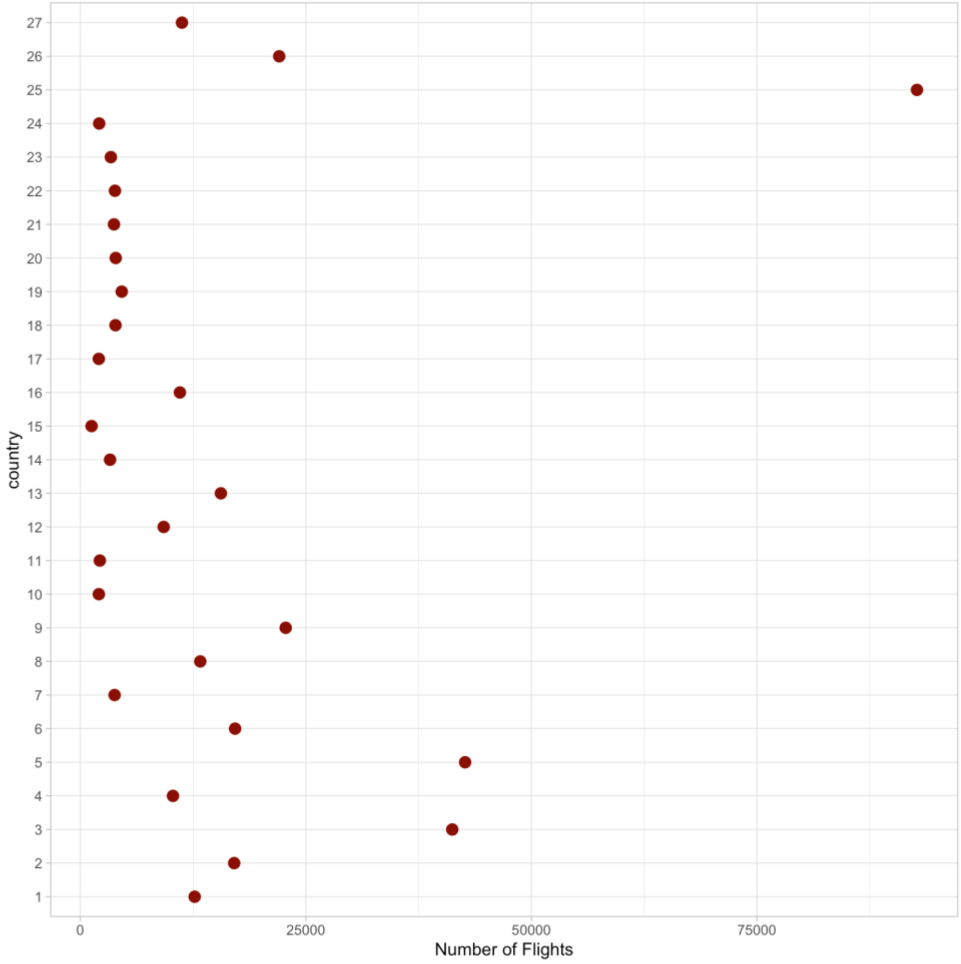
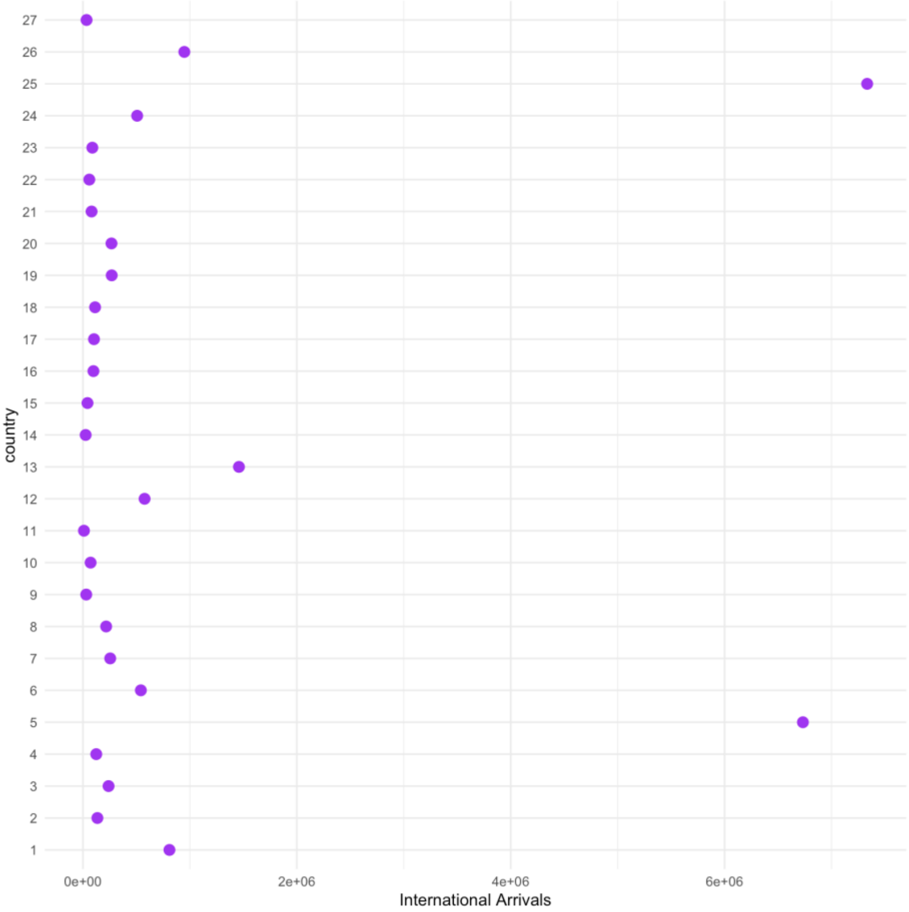


Figure 13: International Arrivals per Country

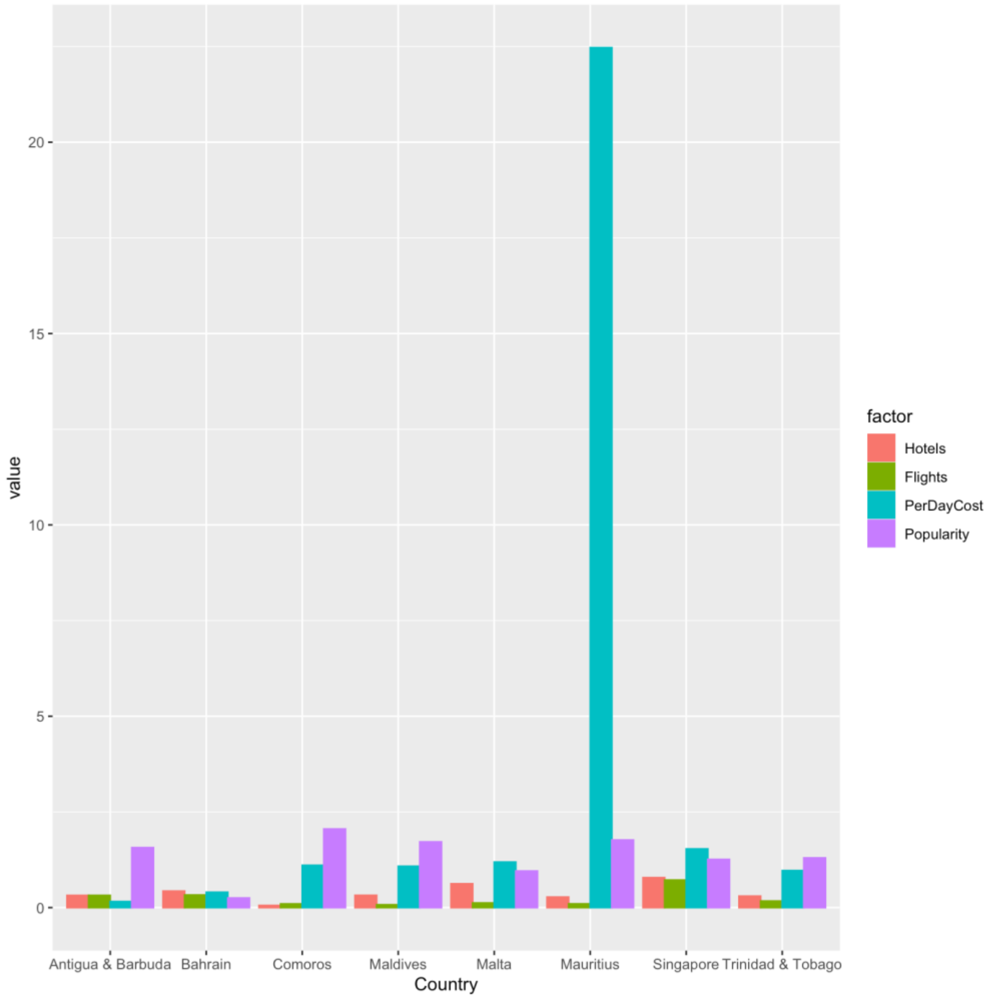


From the above graphs, Figures 10,11,12,13, tourist-friendly nations can be summarised as below:

- Mauritius
- Antigua & Barbuda
- Bahrain
- Maldives
- Malta
- Singapore
- Trinidad & Tobago

The Figure 14 summarizes what analyses have been done in this report for Question 2.

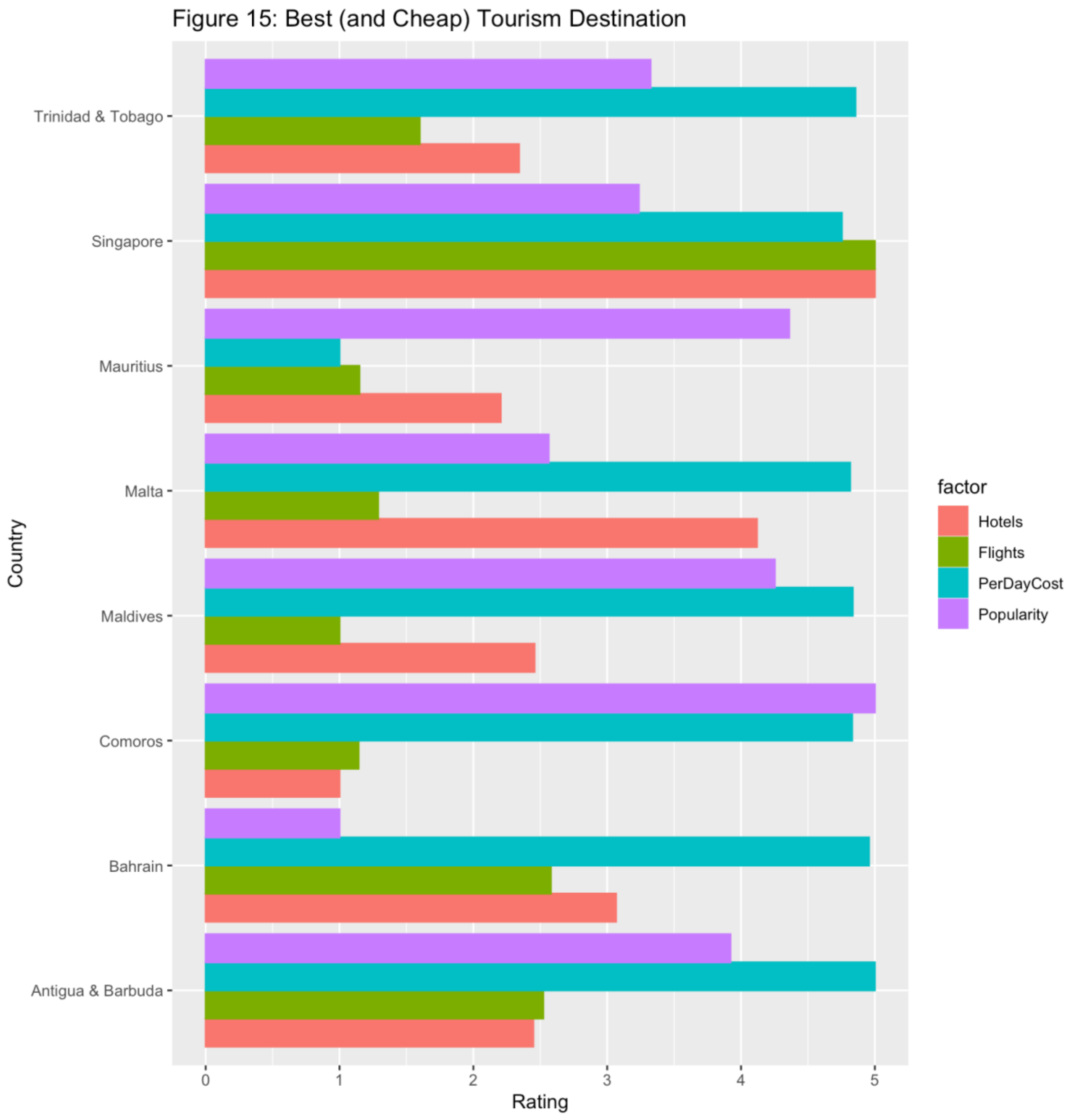
Figure 14: Best (and Cheap) Tourism Destination



However, this graph is not very visually appealing, and it doesn't give a clear picture to the readers as to the meaning of the values. A better representation would be to transform the values of the factors to a rating on a scale of 1 to 5.

Conclusion

After applying a rating system between 1 to 5 (1 for the least favourable and 5 for most favourable), the Figure 15 below gives a clear picture as to which countries have the most favourable attributes for a good and inexpensive vacation.

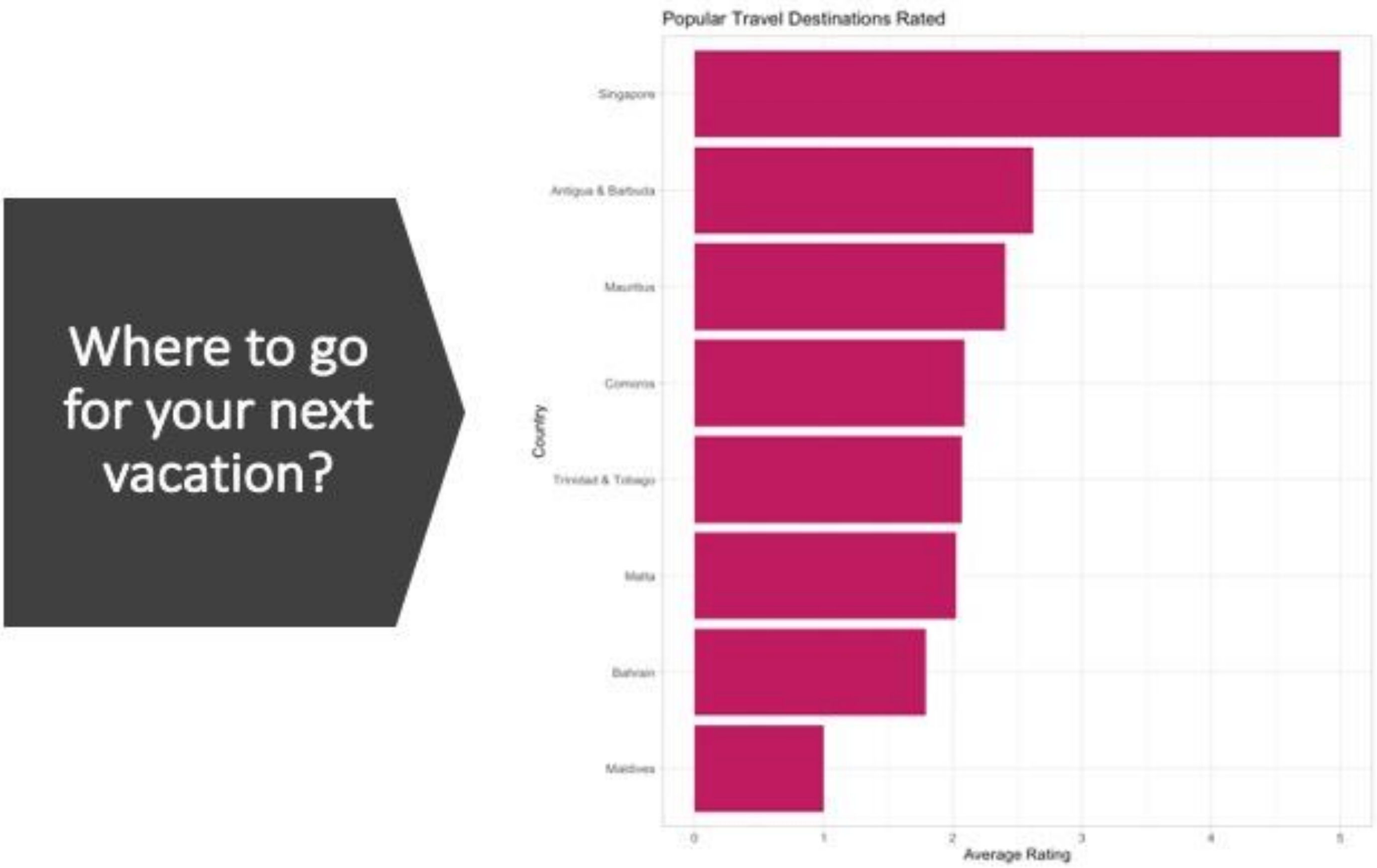


And quite clearly, Singapore has the highest rating for all factors that a tourist should look into.

NOTE:
For the factor "PerDayCost", a lower rating was given to countries that have a high per day cost. For all other factors, the rating is proportional to the value of the attributes seen in the visualisations before.

PART 2: PERSUASION

WHERE TO GO FOR YOUR NEXT VACATION?



Where to go for your next vacation?

The above slide rates the popular countries derived previously.
The factors for rating are :

1. Tourism infrastructure
   Given by averaging the `hotels` and `hotrooms` of each country form the imputed dataset
2. Accessibility: How fast can you get there?
   Given by the attribute `Flights-WB` from the dataset
3. Popularity
   Given by the number of international passengers `ovnarriv` and also by the attribute `receipt` (which shows how much money people spend on tourism)
4. Per Day Cost
   Cheaper is better – so if average daily spending is low, the country is better for travel

Based on the derivations, Singapore has the most balanced offerings and is having a rating of 5/5. Maldives has the lowest rating because it is lacking in most of four factors described above.