TITLE: EMBARKATION & DISEMBARKATION TRENDS AT SINGAPORE'S CHANGI AIRPORT
STUDENT: Mudit Sharma (30340071)
TUTOR: Fahimeh Sadat Saleh (Section 19)

1 - INTRODUCTION

According to Wikipedia, Singapore attracted approximately 19.1 million visitors in 2019 with receipts at S$27.1 billion, according to preliminary figures by the Singapore Tourism Board. This makes it clear that tourism is a prominent driver for the Singaporean economy. Having established this, I intend to analyse trends in embarkation and disembarkation at Changi Airport.

Specifically, the questions I hope to answer are as follows:

1. *Over the years, which regions and countries do outgoing passengers depart for?*

    1.1. *Are there any significant trends that may be observed? For example, are there certain regions or countries that initially created a heavy outflow, but have now tapered off?*

2. *Over the years, which regions are countries do incoming passengers arrive from?*

    2.1. *Are there significant trends that may be observed? For example, are there certain regions or countries that initially created a heavy influx, but have now tapered off?*

3. *By combining the datasets on the common fields of region and country, for which regions or countries is the number of incoming passengers consistently or notably larger than the number of outgoing passengers? Are any of these trends seasonal?*

2 - DATA WRANGLING

1. Air Passenger Departures - Total by Region and Selected Country of Disembarkation (Monthly from January 1961 to February 2020) as reported by the Civil Aviation Authority of Singapore (CAAS), a statutory board under the Ministry of Transport of the Government of Singapore.

    1.1. **DESCRIPTION**: Tabular data with 5 columns and 7811 columns. 1 column for year and month, 1 column for region of destination (South East Asia, Europe, North East Asia etc.), 1 column for country of destination (Japan, United Kingdom, Hong Kong), and 1 column for number of passengers. 1 column lists simple text describing the data represented by each row.

2. Air Passenger Arrivals - Total by Region and Selected Country of Embarkation (Monthly from January 1961 to August 2019) as reported by the Civil Aviation Authority of Singapore (CAAS), a statutory board under the Ministry of Transport of the Government of Singapore.

    2.1. **DESCRIPTION**: Tabular data with 5 columns and 7745 columns. 1 column for year and month, 1 column for region of origin (South East Asia, Europe, North East Asia etc.), 1 column for country of origin (Japan, United Kingdom, Hong Kong), and 1 column for number of passengers. 1 column lists simple text describing the data represented by each row.

Question 1 above concerns itself with the countries passengers embark for, and any significant trends that can be seen in the number of outgoing passengers over time. To answer this, we need only consider ourselves with a single dataset: Air Passenger Departures - Total by Region and Selected Country of Disembarkation

Similarly, Question 2 concerns itself with similar data for disembarking passengers, which only requires us to use [Air Passenger Arrivals - Total by Region and Selected Country of Embarkation]

However, it is more practical to combine the datasets early on because such a dataset can be easily leveraged to answer all 3 questions. Upon importing these datasets, we first make sure that our column names and types are easily understood. For some reason, the datasets in question use *month* to refer to month and year. The other columns are named *level_1, level_2*, and *value*. We can specify data types, and even rename columns within Tableau so as to be able to easily understand the data.

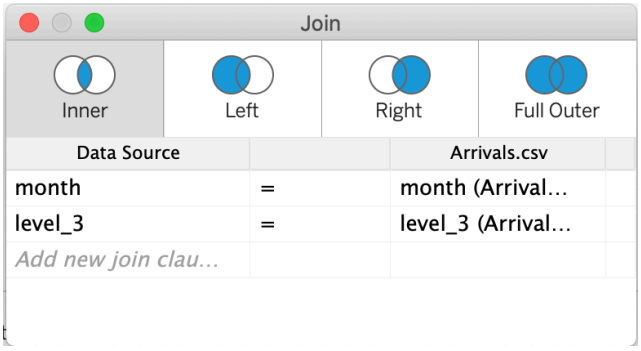


Figure 1

Before we do this, however, we must JOIN the two tables along *month*, which represents the month and year when the data was accumulated over. We also join along *level_3*, which represents the country of embarkation or disembarkation. Here, we would opt for an Inner Join, because the datasets don't exactly cover the same dates. CAAS has recorded Arrivals up-to August 2019, but Departures are recorded up-to February 2020. In order to avoid any unnecessary discrepancies, an Inner Join would not include any date after August 2019.

In addition, we use Tableau to Filter the data to include dates up-to December 2018 only. This ensures that we only have data that describes *full* years, as an Inner Join would include only 8 months of 2019.

Now, we can go ahead the rename columns and change them into the appropriate data types. We can also hide certain columns such as *level_2*, which was describing geographical *region* (South East Asia, Europe etc.) rather than country, which is what we are actually interested in. We can also hide one of the *Date* columns, and one of the *Country* columns. The dataset now looks like this:

Arrivals.csv Date	Arrivals.csv Country	Arrivals.csv Arrivals	Departures.csv Departures
1/1/1961	Malaysia	*null*	*null*
1/1/1961	Indonesia	1687	1878
1/1/1961	Thailand	1172	919
1/1/1961	Philippines	139	122
1/1/1961	Vietnam	*null*	*null*

Figure 2

3 - DATA CHECKING

We can now see that there are some NULL values in the data. We can rectify this using the following:

```
CASE [Arrivals]
WHEN NULL THEN 0
ELSE [Arrivals]
END
```

We perform the same operation on *Departures*. We rename our new calculated fields, hide the columns with 'dirty' data, and our dataset now looks like this:

Arrivals.csv Date	Arrivals.csv Country	Calculation Arrivals	Calculation Departures
1/1/1961	Malaysia	0	0
1/1/1961	Indonesia	1,687	1,878
1/1/1961	Thailand	1,172	919
1/1/1961	Philippines	139	122
1/1/1961	Vietnam	0	0

Figure 3

Since this is a dataset that is provided and maintained by the Government of Singapore, it is unlikely (but not impossible) that is contains any *errors* apart from the NULL values which we have just imputed. However, in the interest of transparency, we can take some steps to ensure that this is the case.

Firstly, we check for errors in the *Date* column:

- Using Pandas, we can see that there are 696 unique dates in this dataset, which makes sense because the dataset spans 58 years (1961 - 2008), and because the number of arrivals and departures is reported once a month, on the 1st of each month. $58 \times 12 = 696$, meaning that each month in these 58 years is accounted for.
- We can also see that there are 7,656 rows that include a value in the Date column. This also makes sense because on the 1st of each month, CAAS reports arrivals and departures for 11 countries, and $696 \times 11 = 7{,}656$
- Finally, using Regular Expressions, we can ascertain that each of these dates is correctly formatted, and does not contain outlandish dates like 45/67/2198. Therefore, we can conclude that the Date column is free of errors.

```
import pandas as pd
import re

df = pd.read_csv('df.csv')
```

```
print ("Unique Dates:",len(set(df.Date)))
print ("Unique Countries:",len(set(df.Country)))
print ("Rows:",len(df.Date))
```

```
Unique Dates: 696
Unique Countries: 11
Rows: 7656
```

```
regex='[1]{1}/[1-9]{1}[0-2]{0,1}/1{0,1}9{1}[6-9]{1}[0-9]{1}|[1]{1}/[1-9]{1}[0-2]{0,1}/2{1}0{1}0{0,1}1{0,1}[0-9]{1}'
None in [re.search(regex,x) for x in df.Date]
```

```
False
```

Figure 4

Next, we verify that the *Country* column doesn't contain any errors such as mistyped country names, non-existent countries, or missing values. We can see that each country is mentioned 696 times, meaning each of these countries has an arrivals/departure value for each of the 696 months covered in the dataset. Some of these values were NULL values, which have since been imputed with 0.

Having verified that the *Date* column and the *Country* column do not contain errors, we can now start visualising our data to answer our questions.

5 - DATA EXPLORATION

5.1 - QUESTION 1

1. *Over the years, which regions and countries do outgoing passengers depart for?*

    1.1. *Are there any significant trends that may be observed? For example, are there certain regions or countries that initially created a heavy outflow, but have now tapered off?*

We have already the domain of countries using a Python Set as shown in Figure 5. There are 11 countries: Malaysia, Indonesia, Thailand, Philippines, Vietnam, China, Hong Kong, Japan, United Kingdom, France, and Germany.

```
df.Country.value_counts()
```

```
Philippines       696
Hong Kong         696
China             696
Indonesia         696
France            696
United Kingdom    696
Vietnam           696
Germany           696
Malaysia          696
Japan             696
Thailand          696
Name: Country, dtype: int64
```

Figure 5

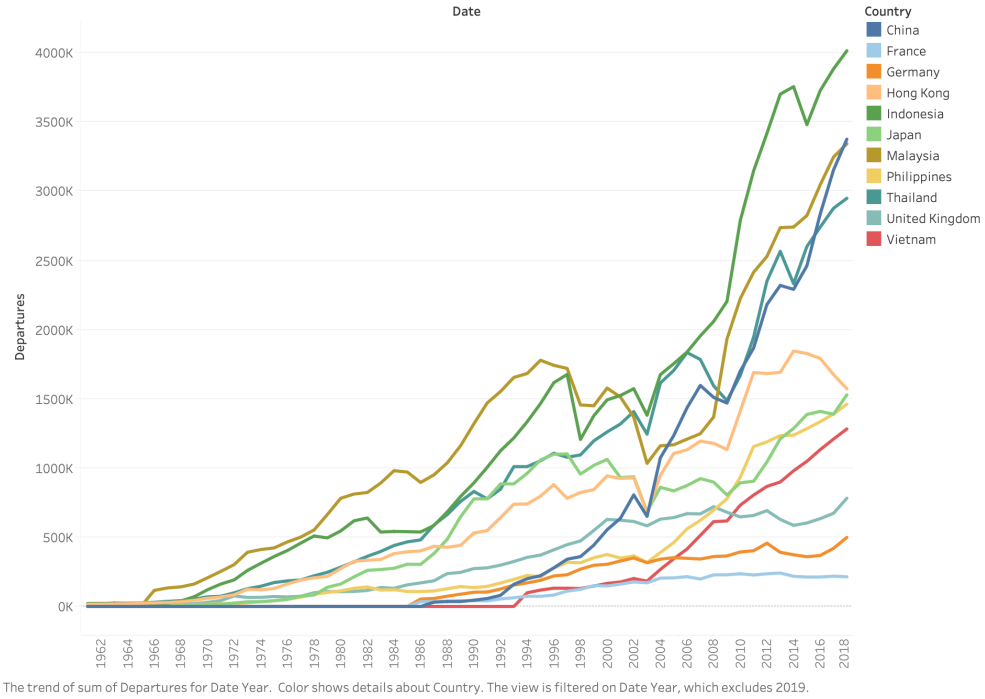


Figure 6

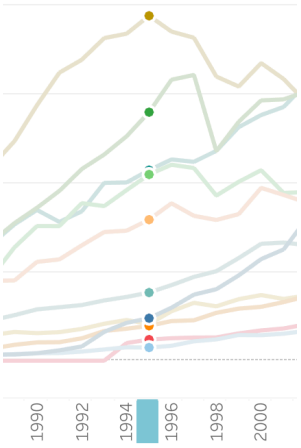
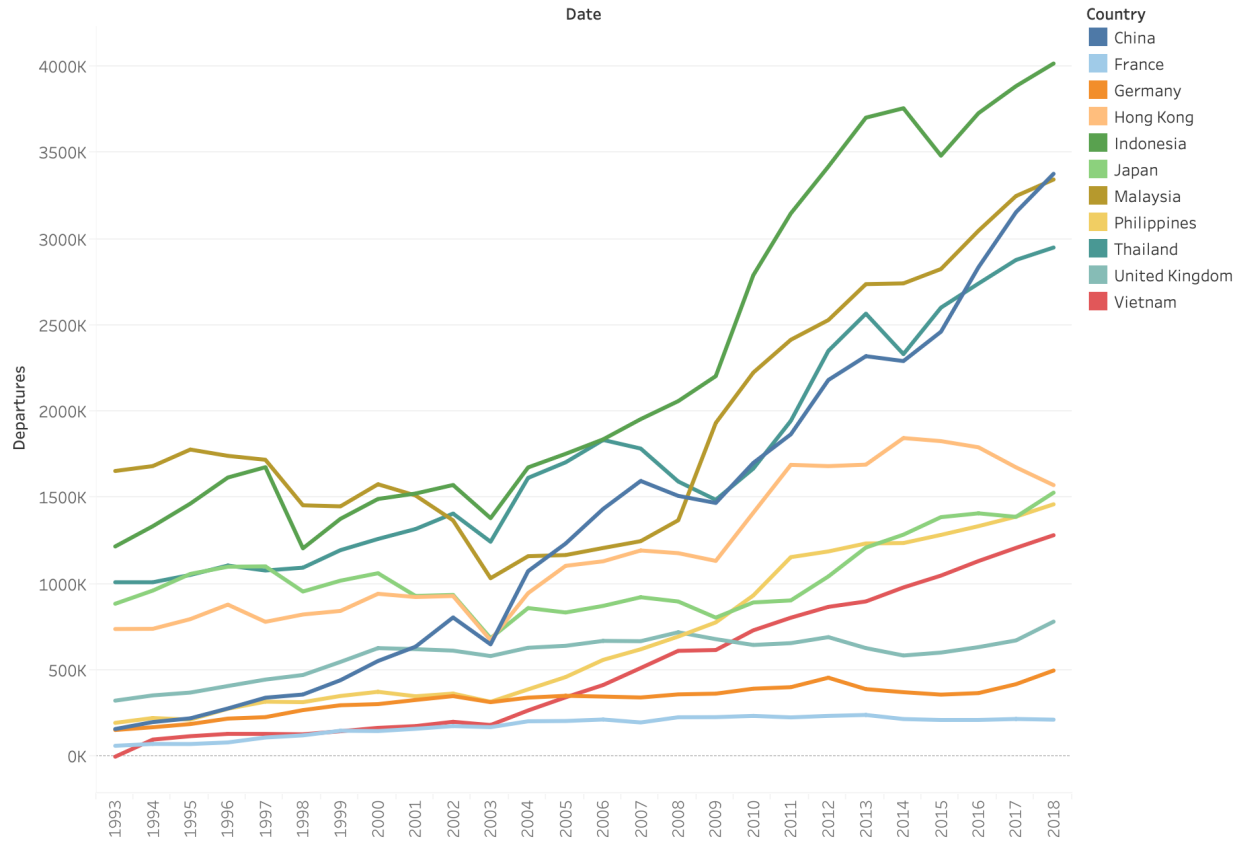


Figure 7

A basic line graph in Tableau allows us to deduce that Departures sees a local maxima no sooner than 1995 for all countries. While Figure 6 shows the overall trend in Departures, Figure 7 confirms that 1995 is the earliest known occurrence of a local maxima. Having determined, this, and keeping in mind the fact that
Question 1.1 specifically asks for countries to which a heavy outflow has now tapered off, it would be simpler to analyse data post-1993 only (Departures to Vietnam are not measured until 1994).



The trend of sum of Departures for Date Year. Color shows details about Country. The view is filtered on Date Year, which keeps 26 of 59 members.

Figure 8

This visualisation puts us in a much better position to answer Question 1.1, enabling us to note (with greater granularity) that following a local maxima at or after 1995, the next significant 'dip' was in 2003 where there is evidently a confounding variable at play. This was in fact the case, as the SARS coronavirus (SARS-CoV) outbreak was first reported to the World Heath Organisation by the Chinese regime in February 2003. Singapore, as a popular tourist destination and transit hub, was also affected by the outbreak, and the Government took measures such as extended school closures, and mandatory and enforced quarantines. Due to this, it is likely that Singaporeans averse to travelling abroad in crowded airplanes to other popular tourist destinations like Thailand, Indonesia, and China. In March 2003, the United States Centre for Disease Control (CDC) even issued a travel advisory recommending against travel to affected countries, including Singapore and Hong Kong, meaning that Departures to Hong Kong from Singapore would also be affected. The 2008-2009 period also sees a slight dip in most countries, which can likely be attributed to the global recession at the time.

The next significant 'dip' was in the 2014-2015 period, but this seems to have affected only Thailand and Indonesia. It is still worth mentioning here because they are among the highest exporters of passengers. In addition to the Indian Ocean Tsunami of 2004, Thailand underwent a turbulent coup d'état in 2014, the effects of which lingered until early March 2019. These may be the cause of this dip, and we can see that the rate of increase of passenger departures has not yet recovered to pre-coup levels, and in fact appears to be approaching a plateau as of the end of 2018.

Across the ocean in Indonesia, the 2014-2015 period was also quite turbulent with many natural disasters such as volcanic eruptions and monsoon flooding which displaced more than 200,000 people. 2015 in particular saw many aviation disasters, with not only passenger casualties but also Indonesian Air Force casualties. In addition, a sharp uptick in terrorist incidents was also a likely contributor to the decline in departures for Indonesia.

5.2 - QUESTION 2

2. *Over the years, which regions are countries do incoming passengers arrive from?*

    2.1. *Are there significant trends that may be observed? For example, are there certain regions or countries that initially created a heavy influx, but have now tapered off?*

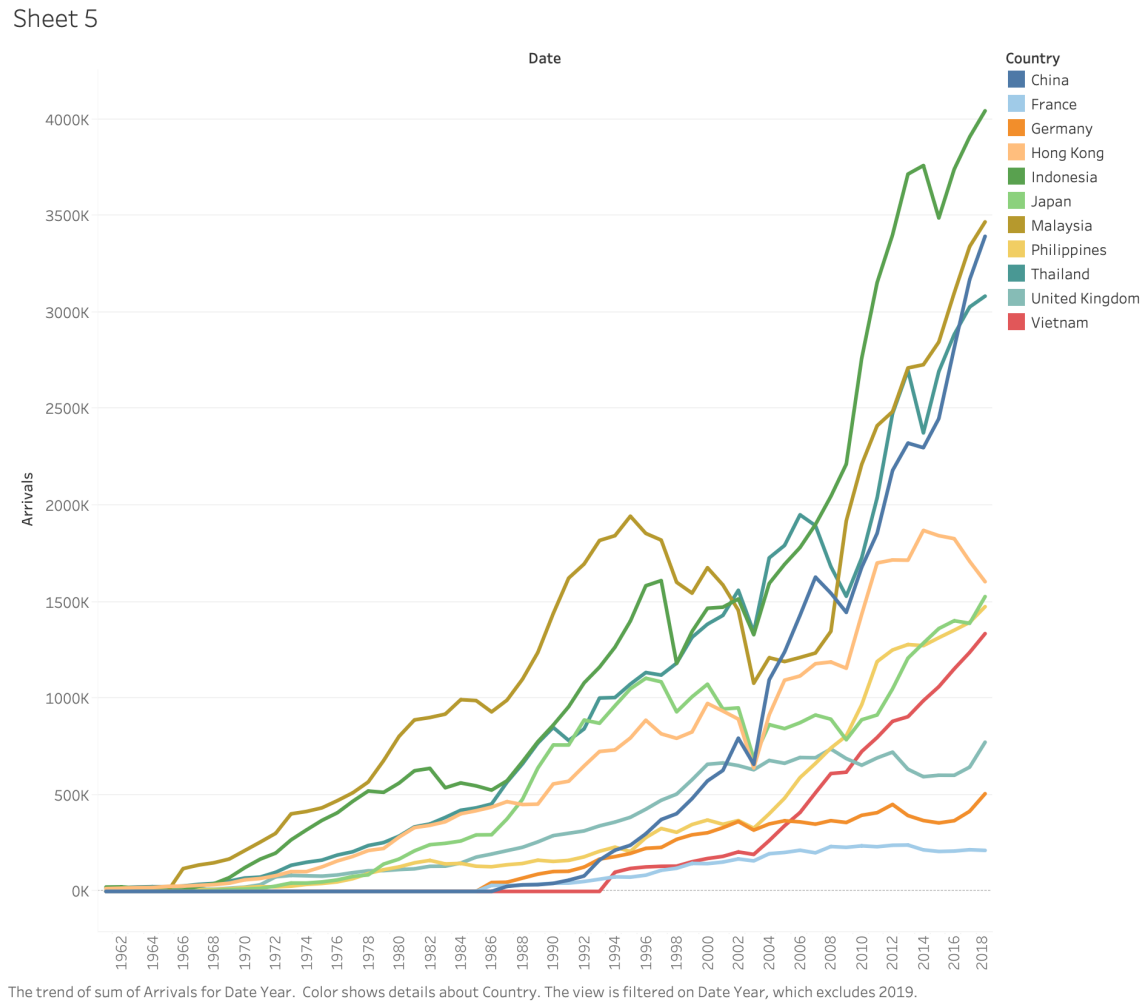


Figure 9

Similar to Departures, Arrivals sees a local maxima no sooner than 1995 for all countries. Having determined, this, and keeping in mind the fact that Question 2.1 specifically asks for countries from which a heavy influx has now tapered off, it would be simpler to analyse data post-1993 only (Arrivals from Vietnam are not measured until 1994).

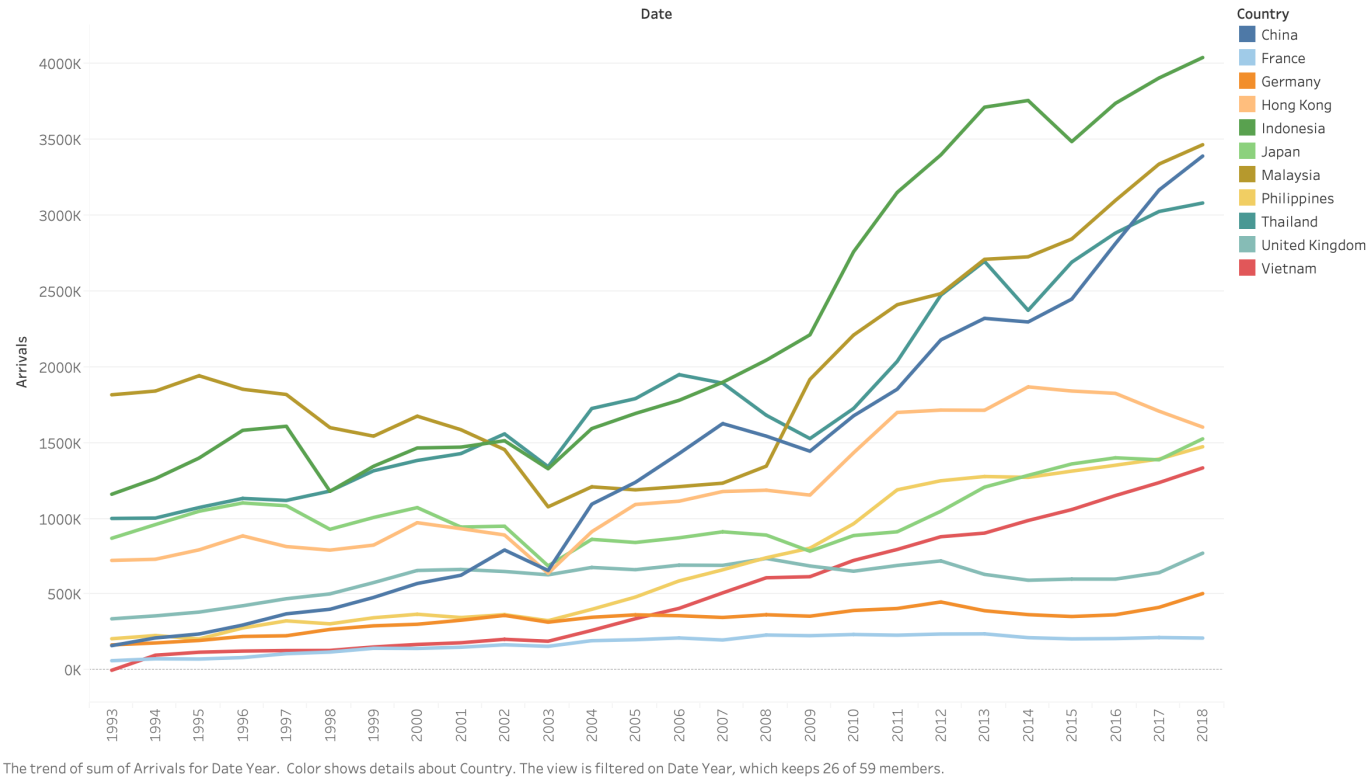


Figure 10

It is important to note that from 1993-2018, the Arrivals and Departures graphs look almost identical, with dips and rises occurring at the same times, and almost at the same rate.

This visualisation puts us in a much better position to answer Question 2.1, enabling us to note (with greater granularity) that following a local maxima at or after 1995, the next significant 'dip' was in 2003 where there is evidently a confounding variable at play. This was in fact the case, as the SARS coronavirus (SARS-CoV) outbreak was first reported to the World Heath Organisation by the Chinese regime in February 2003. Singapore, as a popular tourist destination and transit hub, was also affected by the outbreak, and the Government took measures such as extended school closures, and mandatory and enforced quarantines. In March 2003, the United States Centre for Disease Control (CDC) even issued a travel advisory recommending against travel to affected countries, including Singapore.

Again, the 2008-2009 period also sees a slight dip in most countries, which can likely be attributed to the global recession at the time.

As with Arrivals, Departures once again sees a dip from 2014-2015 mainly affecting Thailand and Indonesia. When discussing Arrivals, I had mentioned the 2004 tsunami, and political turbulence in Thailand during that period, and in Indonesian, further natural disasters, terrorist incidents, aviation disasters, and a devaluation of the Indonesian Rupiah. It is likely that the causes for the simultaneous drop in Arrivals *and* Departures are the same, even if my explanation does not cover *all* of the actual confounding variables responsible.

5.3 - QUESTION 3

3. *By combining the datasets on the common fields of region and country, for which regions or countries is the number of incoming passengers consistently or notably larger than the number of outgoing passengers? Are any of these trends seasonal?*

Using Tableau, we first create a new Calculated Field *Net Influx = Arrivals – Departures*. Figure 12 enables us to answer the question of which countries seem to have notably high Net Influxes, which is the first half of Question 3. Here, we can note that almost every single year sees a positive Net Influx of passengers, with some exceptions like Indonesia which sees more than 20 years of a negative Net Influx. Japan is in the same boat, almost always seeing a negative Net Influx. Hong Kong for the most part sees a positive Net Influx, but this was not always the case, notably in the early 2000's. Other countries like China and France sporadically see a negative Net Influx, but usually not over extended periods.

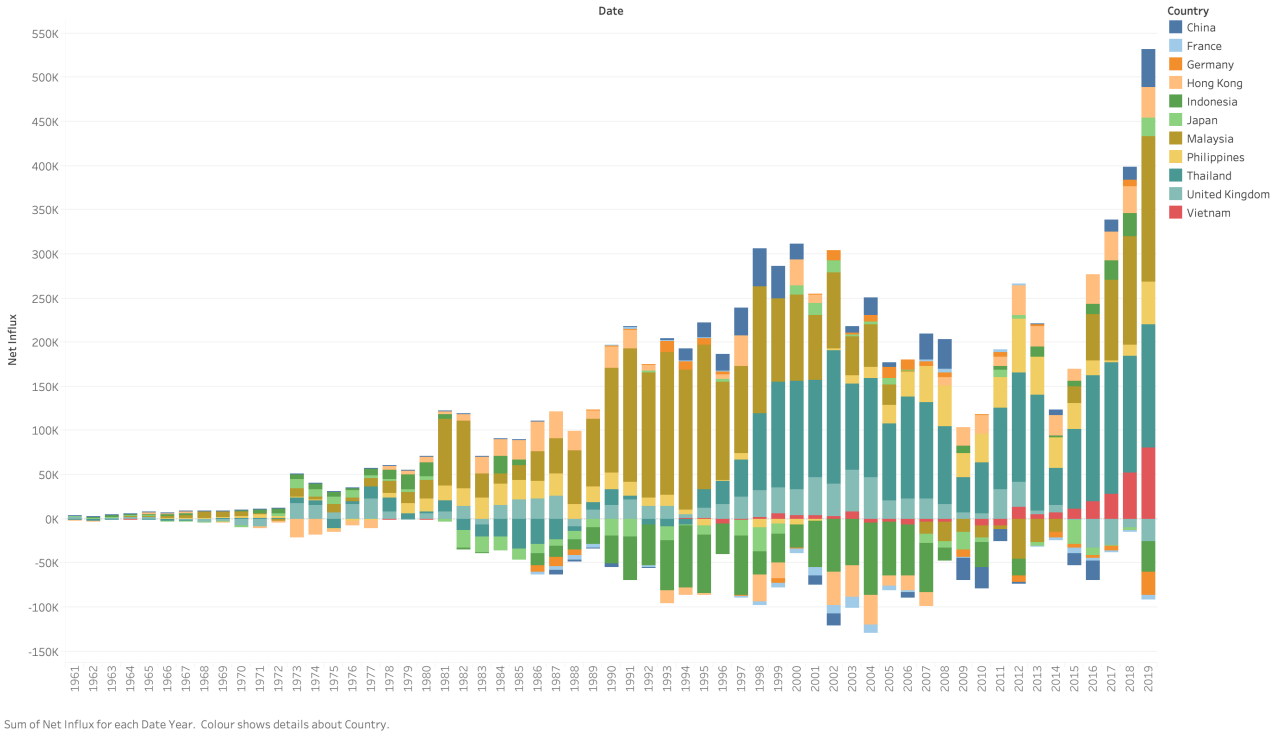


Figure 12

It is important to remember that a low Net Influx does not necessarily mean low Arrivals or low Departures in that period but instead shows the *difference* between the two measurements in the given time period. It is entirely possible that a low Net Influx in a given time period saw record-setting Arrivals *and* Departures. During the course of answering Questions 1 and 2, we can see that 2014 saw very high numbers for both Arrivals and Departures across all 11 countries. However, Figure 12 shows a minuscule Net Influx for this same year, illustrating that Net Influx only measures Arrivals and Departures *in relation* to each other.

The second half of Question 3 asks about the seasonality of Net Influx, which we can analyse using Figure 13:

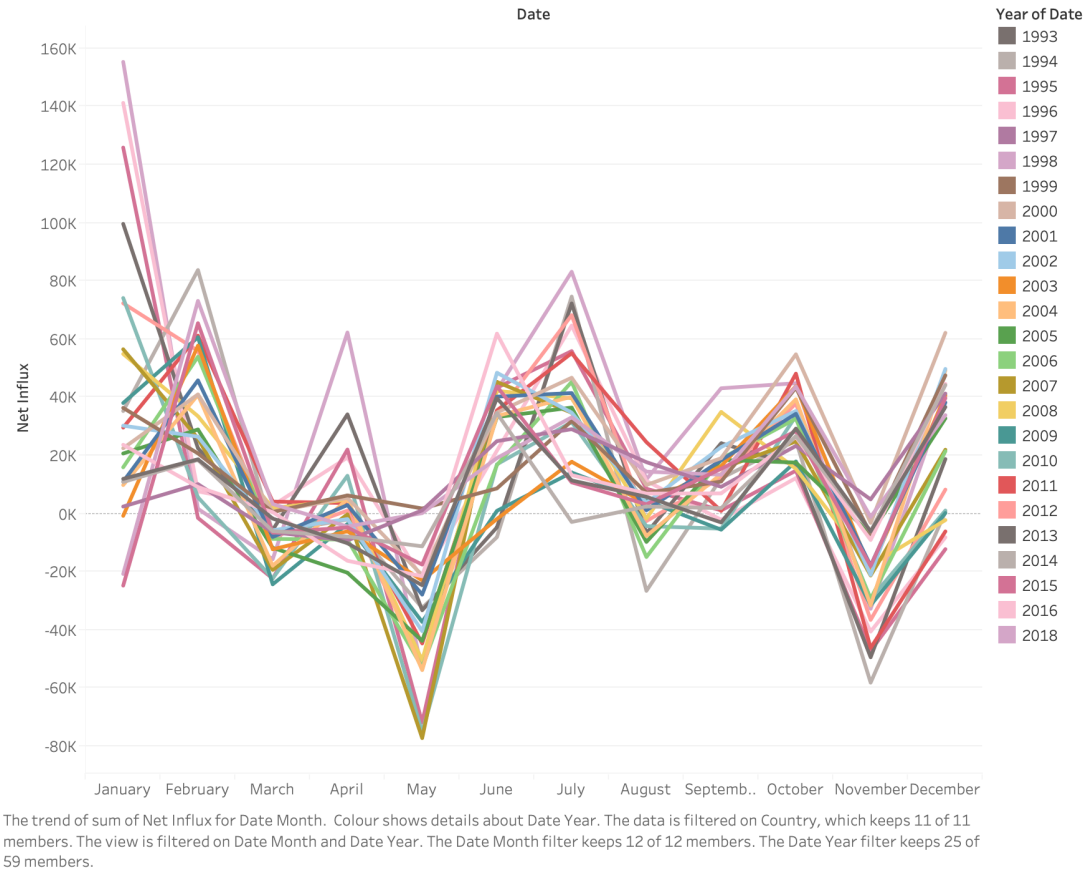


Figure 13

This graph plots the monthly trend in Net Influx for every year from 1993-2018. This range was introduced so that we are better able to see the rises and falls of a mature tourism industry, rather than observing trends in the 1960's which would see double or triple digit values of Net Influx, leaving us unable to form meaningful opinions. In this filtered range, we observe in the tens of thousands.

Figure 13 allows us to easily see that there are in fact trends in Net Influx, despite the graphs for Arrivals and Departures looking virtually identical at first glance. There are clear peaks in February, June, July, October, and December. Conversely, there are clear dips in March, May, August, and November. Using these insights, we can say that trends in Net Influx are in fact seasonal over the years, and that 25 years of data continues to support this seasonality.

However, we cannot yet reliably say that Net Influx from each individual country also follows this trend. It is possible that these trends derive from a select few countries with extremely high values in terms of Arrivals and Departures throughout the years. To verify this without overwhelming ourselves with data, we can plot the total Arrivals and average Departures on a monthly basis, while also introducing a filter that *only considers positive values on Net Influx* on a per-country, and per-month basis:

Using this graph which I had edited for clarity, we can see that hardly any country is responsible for a consistently positive Net Influx, except for Malaysia and Thailand (except for November and December). The United Kingdom and the Philippines see a positive Net Influx for approximately 6 months out of a year, but this is not as consistent as the trend seen with Malaysia and Thailand. In our analysis of only Arrivals and Departures data, we noted that China and Indonesia had notably high levels of Arrivals and Departures, but Figure 14 clarifies that in fact, the two have almost always seen higher Departures than Arrivals.

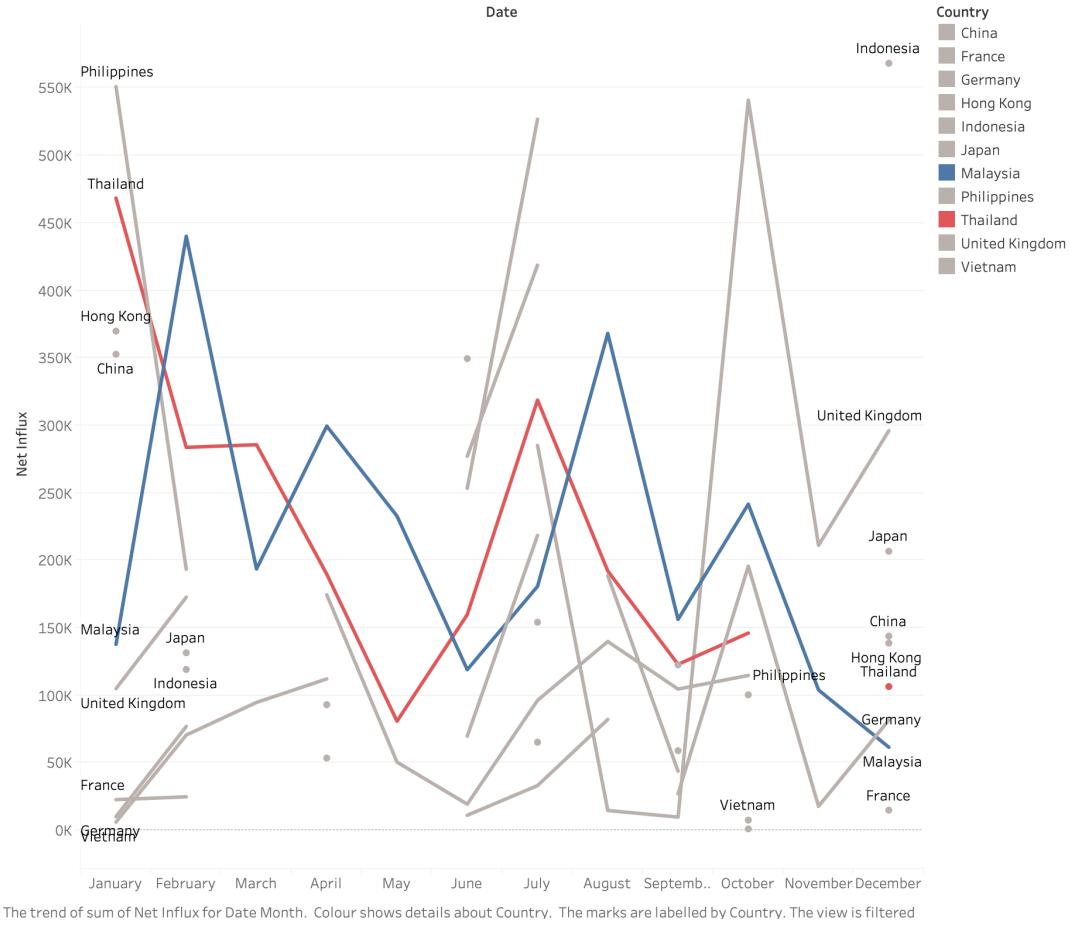


Figure 14

With the insights provided by Figure 14, we can also conclude that the clear-cut seasonality noted in Figure 13 was largely driven by only a handful of countries, namely Malaysia, Thailand, and also on the Philippines and the United Kingdom to an extent. We also see that in terms of the trends in Net Influx, Thailand seems to lag approximately 1 month behind Malaysia, at least from February to November. In conclusion, we can say that the Net Influx of passengers into Singapore is in fact seasonal, varying quite sharply from month to month, and is largely dependent on a select few countries.

6 - CONCLUSION

- **QUESTION 1**: Using visualisations, we were able to determine that though Departures was on a general upward trend, there were some notable dips that could be attributed to the 2003 SARS Virus outbreak, and the short-lived global recession from 2008-2009, both of which led to significant but short-lived tapering off for almost every country.

- **QUESTION 2**: Similarly, visualisations enabled us to observe similar (almost identical) trends in Arrivals, also reflecting dips attributable to the 2003 SARS Virus outbreak, 2008-2009 global recession, again leading to significant dips in almost every country.

- **QUESTION 3**: This question dealt with the Net Influx (Arrivals-Departures), and asked for which countries was this number notably or consistently high? I first approached this using a graph that aggregated on a yearly basis, enabling the identification of countries for which Net Influx varied notably, and also years during which high performers saw sudden drops. Following this, in order to determine the seasonality of trends in Net Influx, I opted for a monthly visualisation of Net Influx, where each year was represented by a different line on a line graph. This showed us the overall seasonality (by month) of Net Influx, allowing us to identify some very unmistakable highs and lows. Finally, I plotted a monthly visualisation while filtering for *positive* values of Net Influx, which showed that only a single country (Malaysia) was responsible for a year-long positive Net Influx, with Thailand coming in at a close second. I also analysed the seasonality of a positive Net Influx, noting that Thailand's trend line seemed to roughly trail Malaysia's trend line by 1 month.

7 - REFLECTION

During the course of this assignment, I spent a lot of time on settling for the best way to actually depict the data. I was unsure of which types of graphs to use, and also how best to approach Question 3 specifically. Questions 1 and 2 were not challenging to deal with, and I spent the bulk of time actually analysing the visualisation rather than generating it.

For Question 3, however, I struggled initially with my own broad wording, mainly because I used vague terminology like 'consistently', 'notably', and 'seasonal'. As we can see from Figures 12-14, there are no countries which consistently generated a positive Net Influx, which is why I was relegated to actually exploring 'notable' generators of positive Net Influx i.e the biggest players in the game, so to speak.

In addition, I regretted the broad question of which countries saw a seasonal trend in Net Influx, because seasonal might mean monthly, yearly, or a winter/summer/autumn/spring perspective. In the end, I went with a monthly overview to avoid any confusion and explanations that I would not be able to provide.

In the future, when faced with a similar task, I would definitely avoid vague terminology such as this, and try to write questions that could be definitely answered with a yes or a no, or create a more open-ended question asking for a simple visualisation and an overview of the trends observed. Also, I was fortunate enough to use a dataset that required very minimal wrangling and cleaning, which I was interested in learning more about especially since Tableau seems to have its own language which can be used on columns, and even in the Marks panel.

8 - BIBLIOGRAPHY

- Wikipedia contributors. (2020, April 29). 2002–2004 SARS outbreak. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:27, April 30, 2020, from https://en.wikipedia.org/w/index.php?title=2002%E2%80%932004_SARS_outbreak&oldid=953910192

- Wikipedia contributors. (2020, April 29). 2004 Indian Ocean earthquake and tsunami. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:27, April 30, 2020, from https://en.wikipedia.org/w/index.php?title=2004_Indian_Ocean_earthquake_and_tsunami&oldid=953873766

- Wikipedia contributors. (2019, September 30). 2004 in Thailand. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:27, April 30, 2020, from https://en.wikipedia.org/w/index.php?title=2004_in_Thailand&oldid=918845732

- Wikipedia contributors. (2018, December 18). 2014 in Indonesia. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:28, April 30, 2020, from https://en.wikipedia.org/w/index.php?title=2014_in_Indonesia&oldid=874327178

- Wikipedia contributors. (2019, October 1). 2015 in Indonesia. In *Wikipedia, The Free Encyclopedia*. Retrieved 16:28, April 30, 2020, from https://en.wikipedia.org/w/index.php?title=2015_in_Indonesia&oldid=919049681