Wei Deng
Info247
Exploratory Data Analysis

1 Initial Hypotheses/Questions

1.1 Motivation

As part of our group's final project, we are looking at the annual phenomenon of cherry blossoms in Japan. During the data gathering phase of our project, we compiled tourism data from Japan to see if cherry blossoms in Japan had any effect on the country's annual tourism, and specifically, from which foreign countries. This EDA serves to explore the annual Japanese tourism data, which give us the number of annual visitors of each country by month/year to Japan.

1.2 Hypotheses

- Japan attracts more foreign visitors in the months of March - May than other months.
- Japan attracts more visitors from East Asian countries compared to other countries.
- The difference of foreign visitors in Japan in the months of March - May is higher for non-East Asian countries than East Asian countries.
  - In other words, more visitors outside of East Asia go to Japan specifically during the cherry blossom season.

1.3 Analysis Plan

1. Group data into ”blooming season” (March - May) and ”non-blooming season” (other months).
2. Analyze whether Japan attracts foreign visitors at a higher rate during blooming season compared to non-blooming season by creating some exploratory plots.
3. Compare the difference in visitors coming from East Asian countries and those that are not.
4. Calculate the percent of tourism activity from each country in the blooming season and compare East Asian countries and non-East Asian countries. Create plots that illustrate this difference.

2 Data Source

2.1 Description

The data has columns: Country/Area, Month, Year, Visitor Arrivals (int), Growth Rate, and Season (blooming/non-blooming). The dataset's range of Years is 1990 - 2022. Each of the columns of the data are important for our analysis.

2.2 Source

The Japanese tourism data was collected from the Japan National Tourism Organization (JNTO), which is operated by the Japanese government.

2.3 Format

The dataset is given as a CSV and downloadable directly as such from the source above.

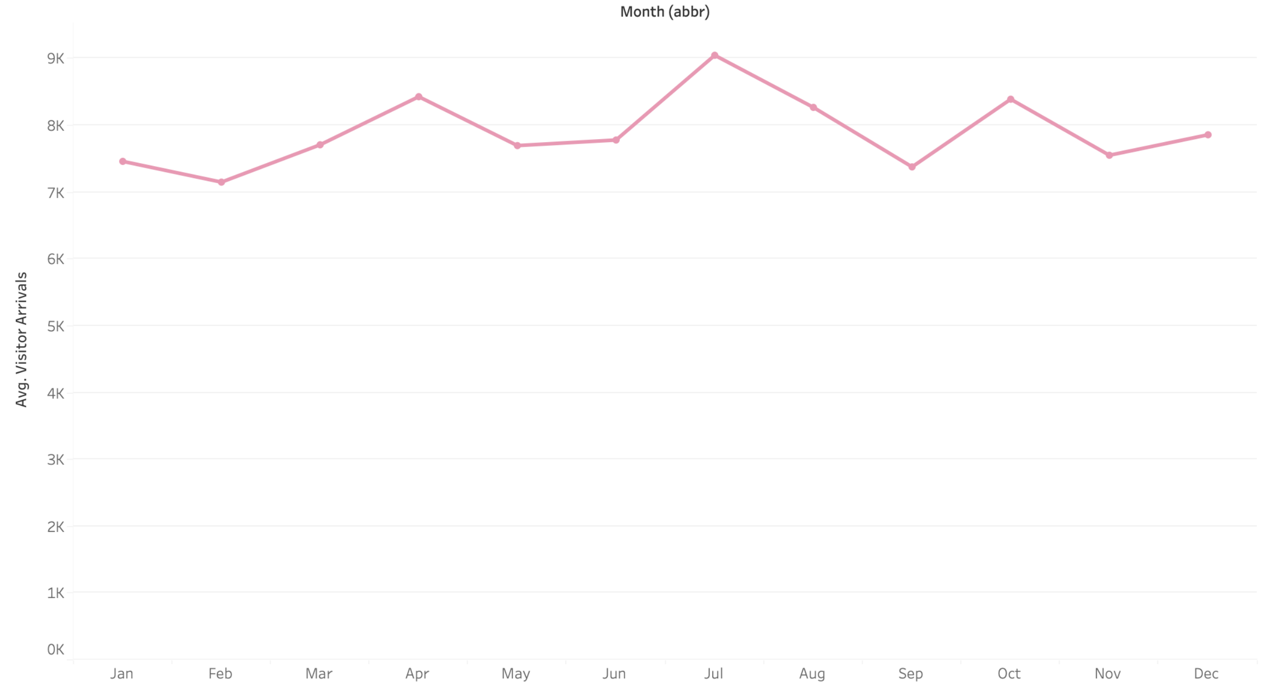
2.4 Transformations

I created a new column (Season) as a classifier of whether the month was in what I defined to be as the cherry blossom blooming season (March - May). I also added an ”East Asian” column to indicate whether the country is in East Asia, where I classified East Asia as China, South Korea, North Korea, Mongolia, and Taiwan.

3 Exploration

For each of the following visualizations, I subsetted the data from years 2000 to 2019. This is due to the amount of missing data for many countries for many years before 2000 and due to the effects of the pandemic after 2020.

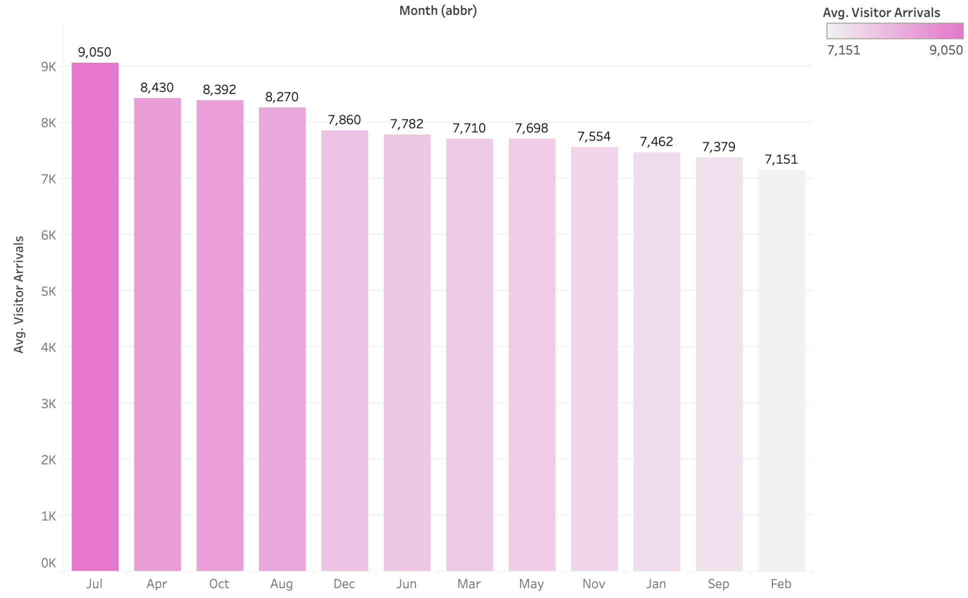
Average Visitor Arrivals by Month (2000-2019)



The trend of average of Visitor Arrivals for Month (abbr) Month. The data is filtered on Year, which ranges from 1/1/2000 to 1/1/2019.

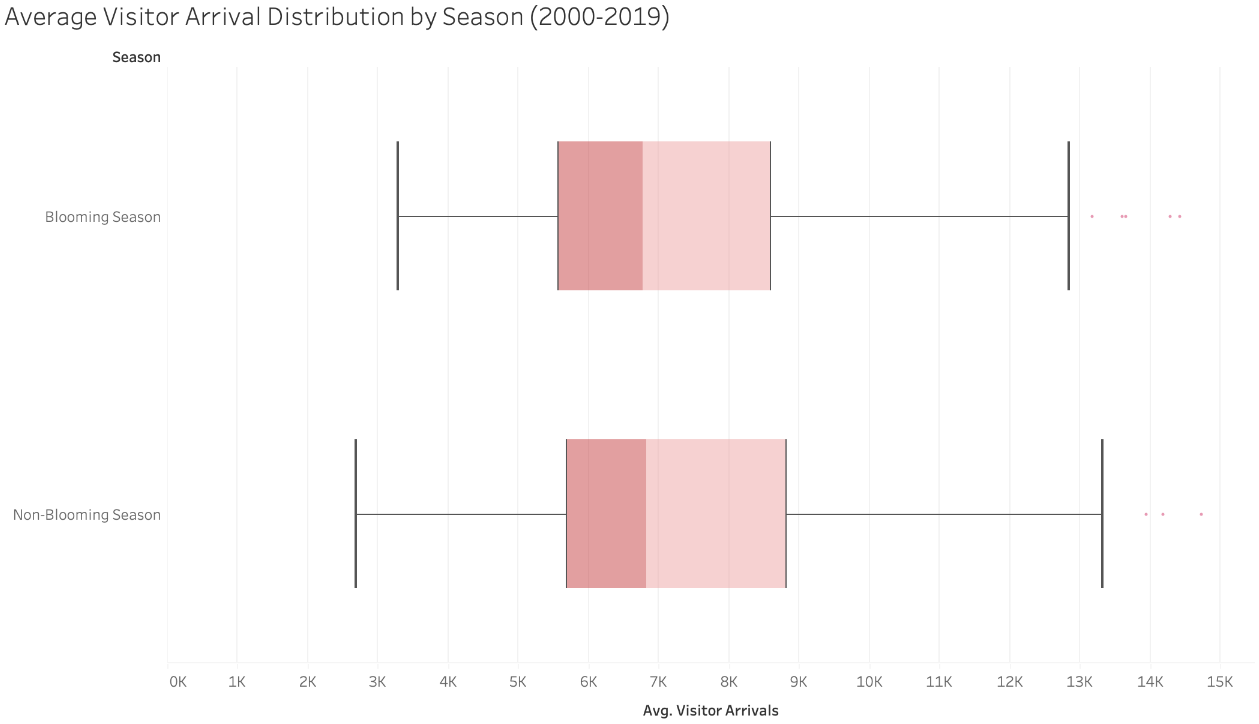
In exploring visitor arrivals, I first wanted to look at the average visitor arrivals for each country by month. This is to confirm our suspicion that there are more visitors to Japan during the blooming season (months March to May). We see that although the number of visitors started increasing after February, March-May do not constitute as the months with the highest average visitors. Next, I wanted to see by order, which months experienced the highest average number of visitors.

Average Visitor Arrivals by Month (2000-2019)



Average of Visitor Arrivals for each Month (abbr) Month. Color shows average of Visitor Arrivals. The data is filtered on Year, which ranges from 1/1/2000 to 1/1/2019.

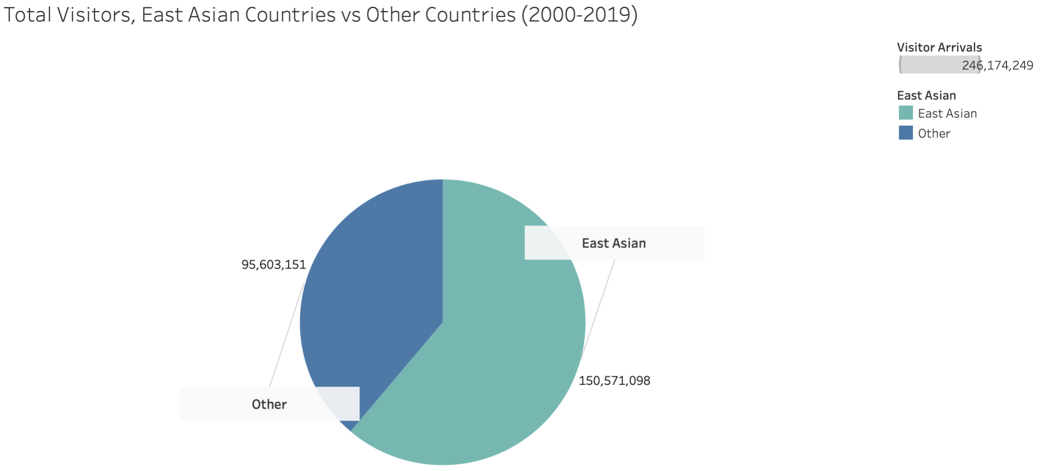
The above plot shows us the highest number of average visitors sorted by month. This shows that July is the month experiencing the highest number of visitors on average, with April following and March and May toward the middle of the distribution. Lastly, I wanted to check the distribution of average visitors by month and year for all countries by creating a barplot.



Average of Visitor Arrivals for each Season. Details are shown for Month (abbr) Month and Year Year. The data is filtered on Year, which ranges from 1/1/2000 to 1/1/2019.

From the boxplot above, we see that the distributions for the different groups of months are slightly different but overall pretty similar. These three plots adds some doubt to the idea that there are more visitors to Japan during the blooming season.

Next, we will take a look at the data from the angle of exploring our other major hypotheses (is the distribution of East Asian visitors different from those that aren't?).

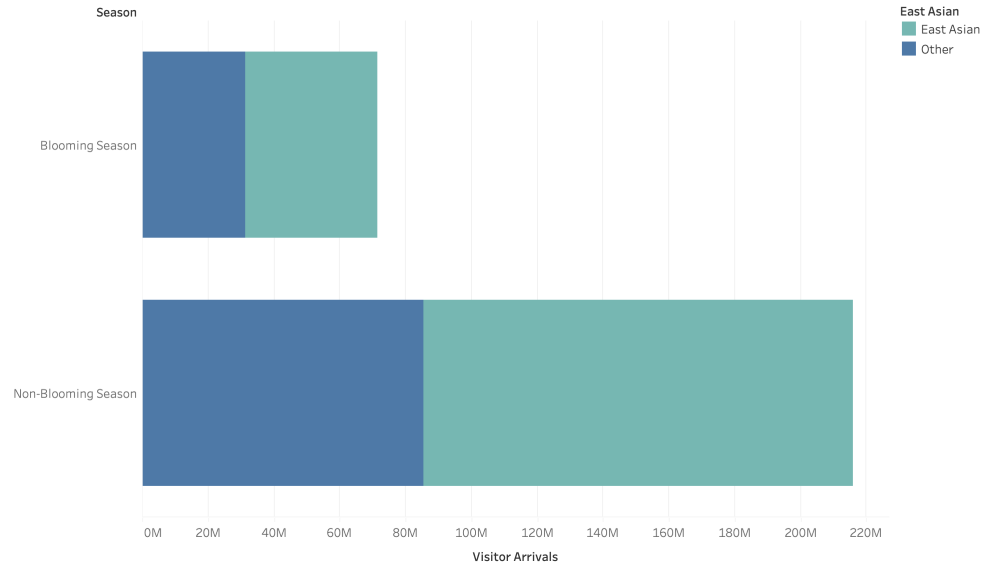


East Asian (color) and sum of Visitor Arrivals (size). The data is filtered on Year, which ranges from 1/1/2000 to 1/1/2019.

From this pie graph, we see that East Asian countries account for most of the tourism in Japan, even compared to the other countries combined. This confirms our suspicion that East Asian countries account for most of the tourism in Japan in general.

Next, we explore whether there is a difference in the visitor arrivals for Japan during the blooming season for East Asian countries vs non-East Asian countries.

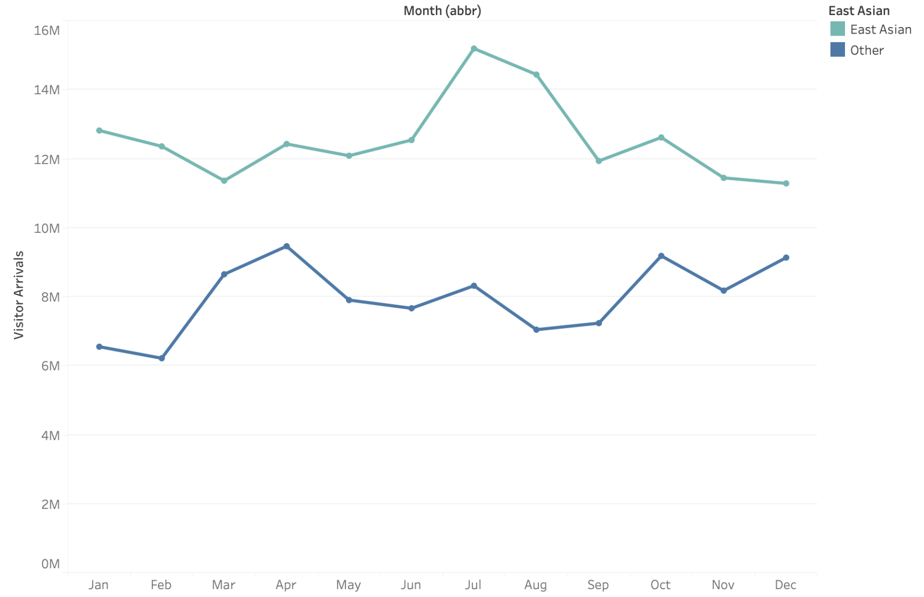
Visitor Arrivals by Blooming Season and East Asian (vs Non-East Asian)



Sum of Visitor Arrivals for each Season. Color shows details about East Asian.

From this plot, we see that while most of the tourism to Japan for East Asian countries occurs during the non-blooming season (about two-thirds), the proportion is different for non-East Asian countries (closer to one-half. This confirms our suspicion that there is a difference between when visitors go to Japan depending on whether they're from an East Asian country or not.

Sum of Visitor Arrivals by Month between East Asian Countries and Others (2000-2019)

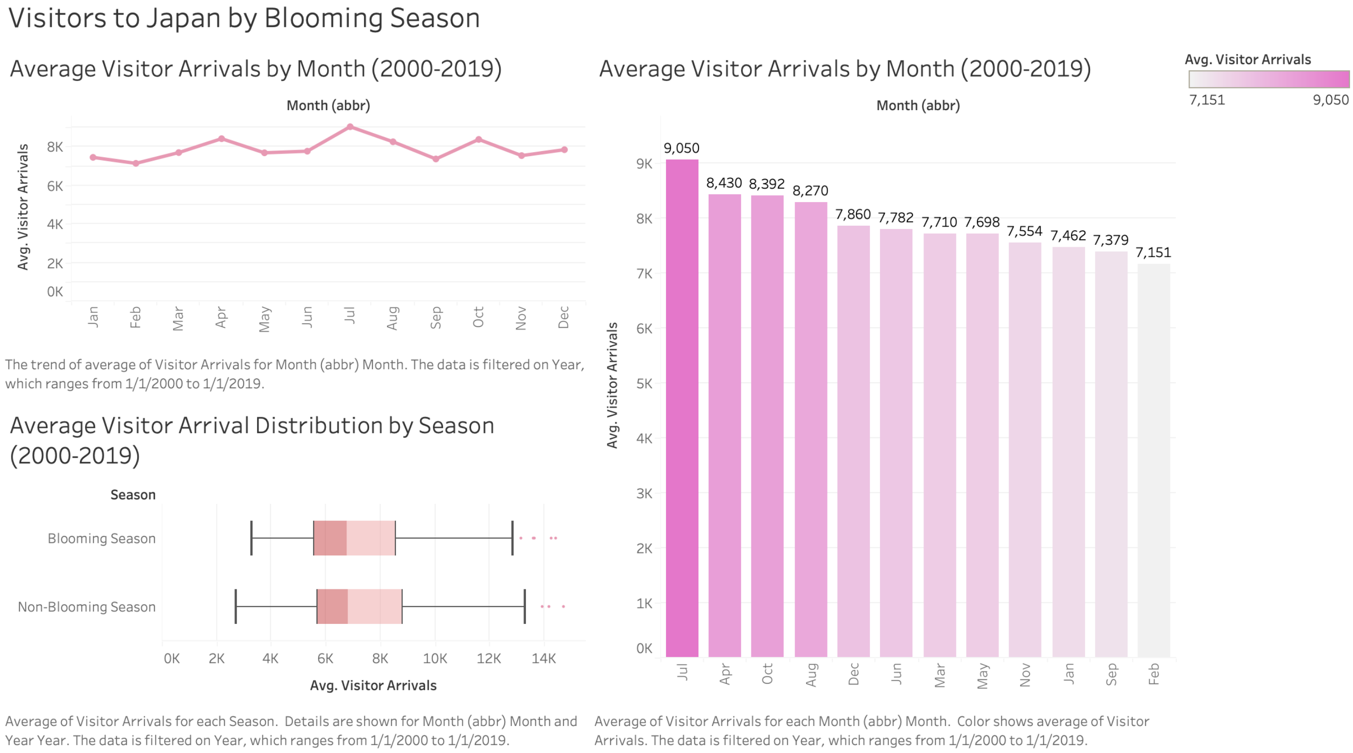


The trend of sum of Visitor Arrivals for Month (abbr) Month. Color shows details about East Asian. The data is filtered on Year, which ranges from 1/1/2000 to 1/1/2019.

From this plot, we see that for the start of the blooming season (March), there is an increase in tourism from non-East Asian countries compared to the decrease in visitors from East Asian countries.

4 Dashboards

4.1 Hypothesis 1



Visitors to Japan by Blooming Season

Average Visitor Arrivals by Month (2000-2019)

The trend of average of Visitor Arrivals for Month (abbr) Month. The data is filtered on Year, which ranges from 1/1/2000 to 1/1/2019.

Average Visitor Arrival Distribution by Season (2000-2019)

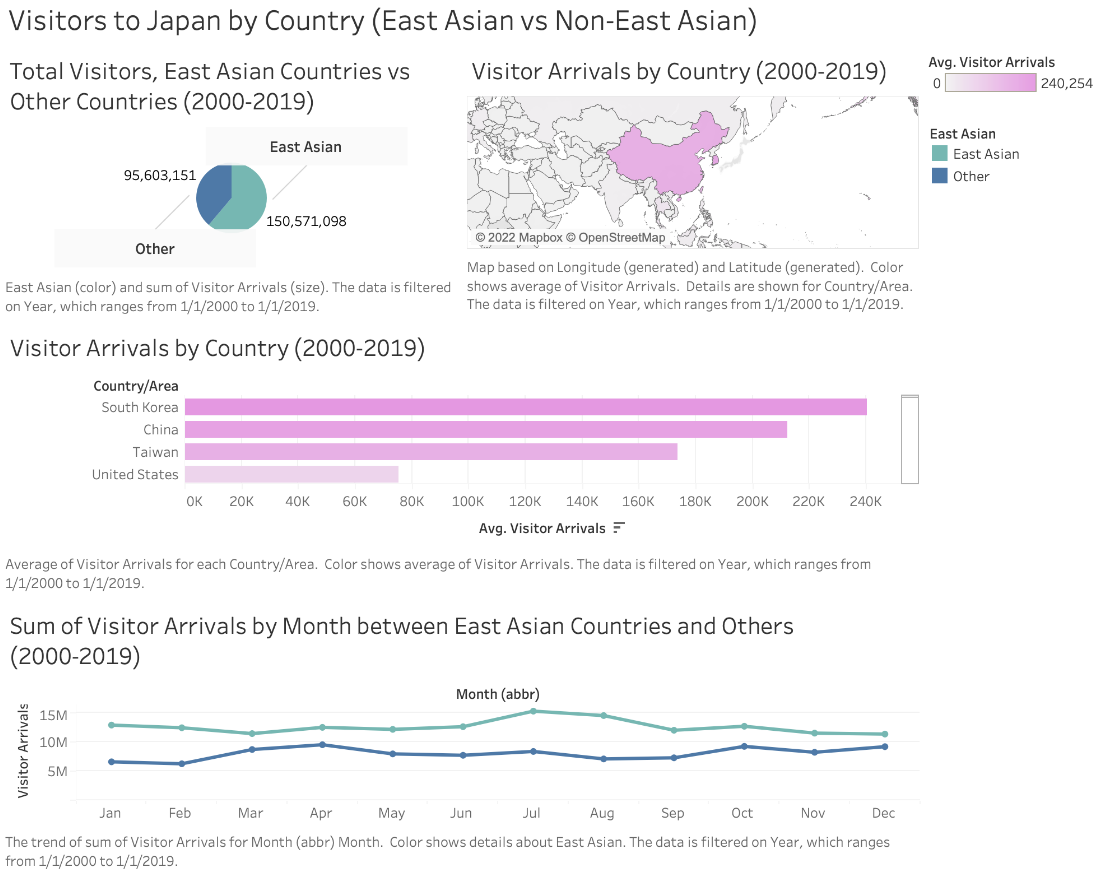
Average of Visitor Arrivals for each Season. Details are shown for Month (abbr) Month and Year Year. The data is filtered on Year, which ranges from 1/1/2000 to 1/1/2019.

Average Visitor Arrivals by Month (2000-2019)

Average of Visitor Arrivals for each Month (abbr) Month. Color shows average of Visitor Arrivals. The data is filtered on Year, which ranges from 1/1/2000 to 1/1/2019.

From this dashboard, we see that each of the three plots refutes are hypothesis that there is more tourism in Japan during the blooming season of cherry blossoms. The plots refute this because although there is generally an upward trend in visitors during springtime, the blooming season months do not account for the highest number of average visitors. Additionally, the distributions of visitor arrivals are similar for both the blooming season and the non-blooming season.

4.2 Hypothesis 2



East Asian (color) and sum of Visitor Arrivals (size). The data is filtered on Year, which ranges from 1/1/2000 to 1/1/2019.

Map based on Longitude (generated) and Latitude (generated). Color shows average of Visitor Arrivals. Details are shown for Country/Area. The data is filtered on Year, which ranges from 1/1/2000 to 1/1/2019.

Average of Visitor Arrivals for each Country/Area. Color shows average of Visitor Arrivals. The data is filtered on Year, which ranges from 1/1/2000 to 1/1/2019.

The trend of sum of Visitor Arrivals for Month (abbr) Month. Color shows details about East Asian. The data is filtered on Year, which ranges from 1/1/2000 to 1/1/2019.

For the second hypothesis, in assessing whether the highest number of tourists comes from East Asia, the plots above confirm our suspicions. This can be seen on the choropleth map as well as the horizontal bar chart that the top countries for visitor arrivals to Japan are East Asian. For our other hypothesis, we also confirm that there is a difference in the average visitors for East Asian and non-East Asian countries during and not during the blooming season. This can be seen in the stacked horizontal barchart shown in the exploration section.

5 Conclusion

In our initial analysis of the tourism data in relation to the blooming season for cherry blossoms (which is the basis for our overall final project), we see that from the plots we created, the average number of visitors does not change depending on whether Japan is in the blooming season. This disputes are original hypothesis.

For our other hypotheses, we confirm that most of the tourism in Japan comes from other East Asian countries, which makes sense due to proximity. In relation to our first hypothesis, we do see that there is a slight difference in tourism during blooming season for non-East Asian countries, where there are slightly higher proportion of visitors going to Japan during the blooming season.

In assessing whether we should move forward with the tourism component of our final project, we may decide that confirming the second and third hypotheses are not compelling enough to make interesting visualizations. Thus, we will meet as a group and discuss together following this exploratory data analysis.