EDA on Cherry Blossom Blooming Days

Ashlyn Jew

1 Initial Hypotheses or Questions

1.1 Motivation

Cherry blossoms, or sakura, are an icon of Japan's culture and natural scenery. There are millions of tourists that travel to Japan every year just to view the cherry blossom flowers. In Japan, cherry blossoms are associated with the arrival of spring, and are extremely popular even among Japan's residents. Across the Pacific Ocean in the Bay Area, cherry blossom celebrations occur in April, despite seeing cherry blossoms blooming earlier in the year. This leads to the question of when and where most cherry blossoms bloom. With global warming, are cherry blossoms blooming earlier in the year? If so, when and where is the best time to travel to Japan to see first blooms or full blooms? The following exploratory data analysis aims to answer these questions specifically relating to the dates of the first and full blooms, as well as the locations of these cherry blossoms.

1.2 Hypotheses

1. Cherry blossoms in Japan are having their first bloom and full bloom dates earlier in the year over time.
2. The time between first bloom and full bloom has shortened over time.
3. There is a regional pattern in blooming dates (i.e. cherry blossoms in cities that are nearby tend to have similar blooming dates).
    3.1. Cherry blossoms in the south of Japan bloom first before the Cherry blossoms further north.

1.3 Analysis Plan

My analysis plan is as follows:

1. Plot the proportion of bloom dates that occurred in each month for each year to determine if the proportion of early-month bloom dates has increased.
2. Calculate the number of days between first bloom and full bloom for all observations.
3. Plot the cities on a choropleth to see where each city lies. Determine if there may be some regional trend.

2 Data Source(s)

2.1 Description

The dataset I selected has information on cherry blossoms' (also known as Sakura) first and full bloom dates across different cities and areas in Japan. Data is available for the years 1953 to 2021. The final cleaned and transformed dataset I used for exploratory data analysis contains 7083 rows and 7 columns.

Table 1: Variables in Final Dataset

Variable Name	**Description**
Site Name	*[String]* Name of the city or area where the observation took place.
Currently Being Observed	*[Boolean]* Is this site still being observed?
Latitude	*[Number (Decimal)]* Latitude of Site Name city/area
Longitude	*[Number (Decimal)]* Longitude of Site Name city/area
Year	*[Date]* Year of Observation (from 1953 to 2021)
First Bloom Date*	*[Date]* Date of first bloom at Site Name
Full Bloom Date*	*[Date]* Date of full bloom at Site Name

*According to https://sakura.weathermap.jp/en.php,
"First bloom" means observation trees have 5 to 6 flowers that have bloomed.
"Full bloom" means 80% of the observation tree flowers have bloomed.

2.2 Source(s)

The data was downloaded from Kaggle at this page: https://www.kaggle.com/ryanglasnapp/japanese-cherry-blossom-data. The compiler of the data is Kaggle user Ryan Glasnapp. As stated in the dataset description, the dataset was scraped from the Japan Meteorological Agency's website. Most of the translation of the data from Japanese to English was done through Google Translate using a Python package, while certain translations were done manually.

2.3 Format

There were two .csv files available for download: one detailing the sakura first bloom dates, and the other detailing the sakura full bloom dates. Each dataset contained a column for the names

of the cities or areas that the observation was recorded, a column indicating whether the observation site was still being observed, and columns for each year from 1953 to 2020. Each .csv file had 102 different observation sites or rows and 72 columns.

2.4 Transformations

The original data on Kaggle covered the years 1953 to 2020, so I updated the data with 2021 dates manually by navigating to the Japan Meteorological Agency website. The Kaggle dataset also included a Python notebook that had instructions for adding longitude and latitude data, so I followed those instructions to add those two columns to the data. I also transformed the data to long format, so the final dataset I used contained 7038 rows and 7 columns (shown in Table 1).

3 Exploration

3.1 Exploring Hypothesis 1

I wanted to see if there were more first blossom dates earlier in the year with time on the y-axis. I determined that one way to see this is to count the number of dates that fall into each month for each year. Because the number of observations for each year is different, I got the percentage for each month instead of the raw counts.

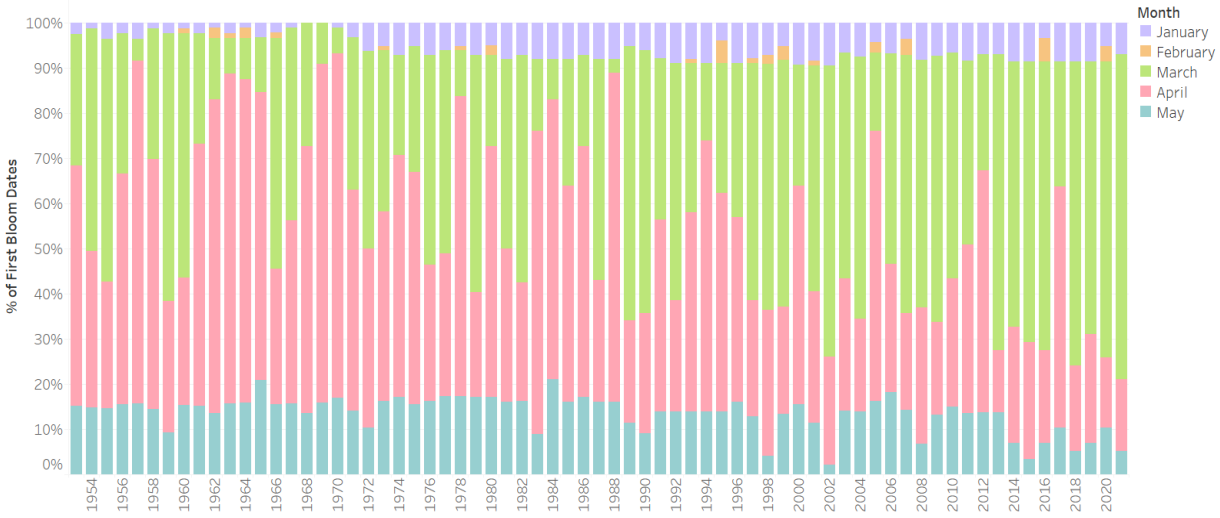

Figure 1: % of First Bloom Dates for Each Month (Stacked Bar Chart)

With this stacked bar chart, it allowed me to see that the months of March, April, and May consisted of the most first bloom dates. You could see a general decreasing trend for the month of April, but I chose to use a line chart to see the trend more clearly. I also only focused on those three months.

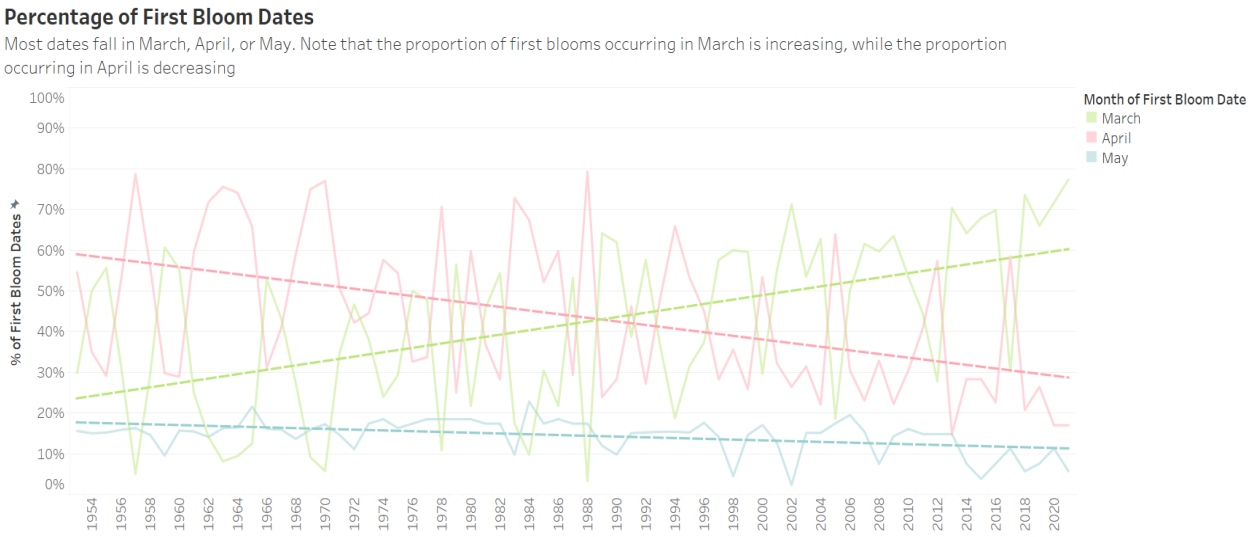

Figure 2: % of First Bloom Dates Falling in March, April, and May (Trend Lines)

After I plotted the line chart over time for those three months, you could see the trends for each month. However, I wanted to make those trends more clear because the data does fluctuate a lot, so I added trend lines to highlight the trends.

Since the trend line for March shows an increase over time while the trend lines for April and May show a decrease over time, this indicates that more cherry blossoms are first blooming in earlier months. This supports the idea that cherry blossoms in Japan have been blooming earlier in the year since 1953. This is evidence that supports our first hypothesis.

3.2 Exploring Hypothesis 2

Making use of the full bloom date column as well, I wanted to see how the time between first bloom and full bloom changed over time. I created a calculated field that subtracted the days between the full bloom date and the first bloom date for each observation. Then, I plotted the average of the time to full bloom over time (Figure 2).

With the line chart I was able to see the trend over time clearly. The graph shows a fluctuating trend over the years where most of the average days to full bloom stayed within the 6 to 8 days range. This indicates that the time between first bloom and full bloom has stayed consistent over time. Thus, this graph does not support our hypothesis that the time to full bloom has shortened over time.

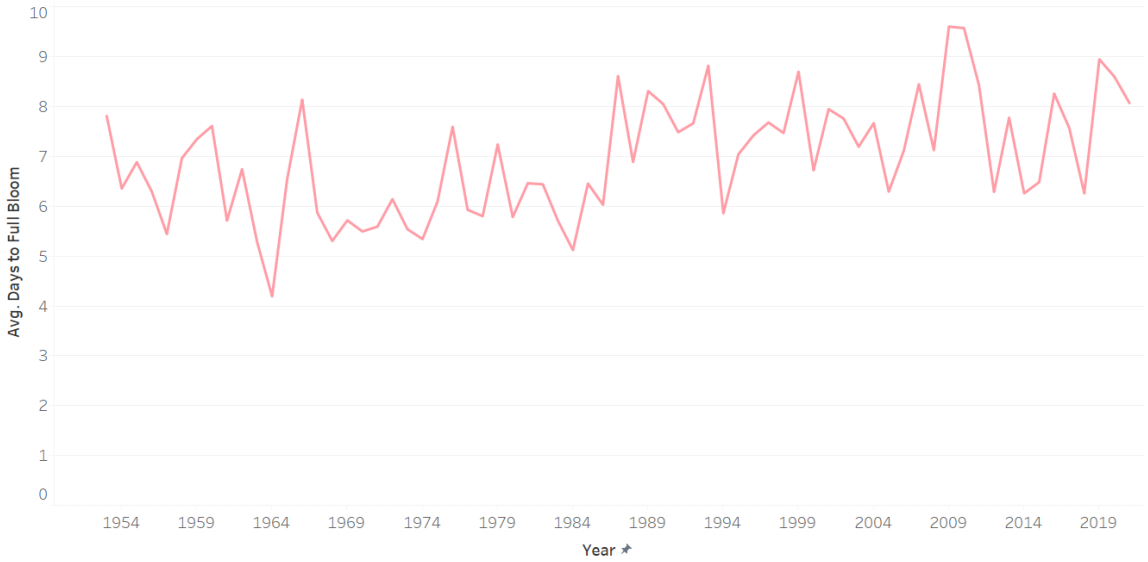

Figure 2: Average Days From First Bloom to Full Bloom (Trend Line)

3.3 Exploring Hypothesis 3

The data also included information about where the observation was recorded, and the sites were also spread across Japan. Hence, I wanted to explore any geographical trends over the years.

First, I plotted the data on the map of Japan and colored the points by the month of their full bloom date. I included the years as pages so that we could see how the colored regions changed over time.

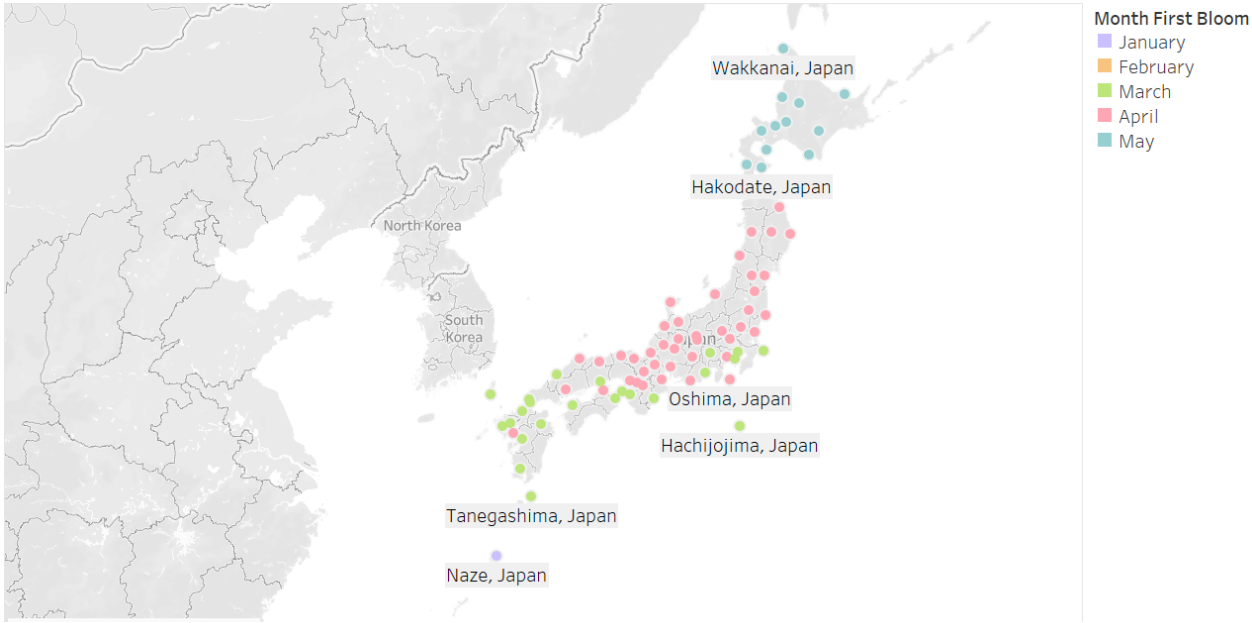

Figure 3: First Bloom Month Choropleth (for the year 1953)

You can clearly see that the observations in the southern region of Japan are colored for the earlier months of the year: January, February, and March. Whereas the middle region of Japan is colored for April and the northern region of Japan is colored for May. These patterns are similar over each year. This suggests that cherry blossoms bloom first in the southern regions of Japan and then in the middle and northern regions. This supports my sub-hypothesis that cherry blossoms in Japan bloom in the south first.

Next, I plotted the data on the map of Japan and colored the points by the average time it takes from first bloom to full bloom.

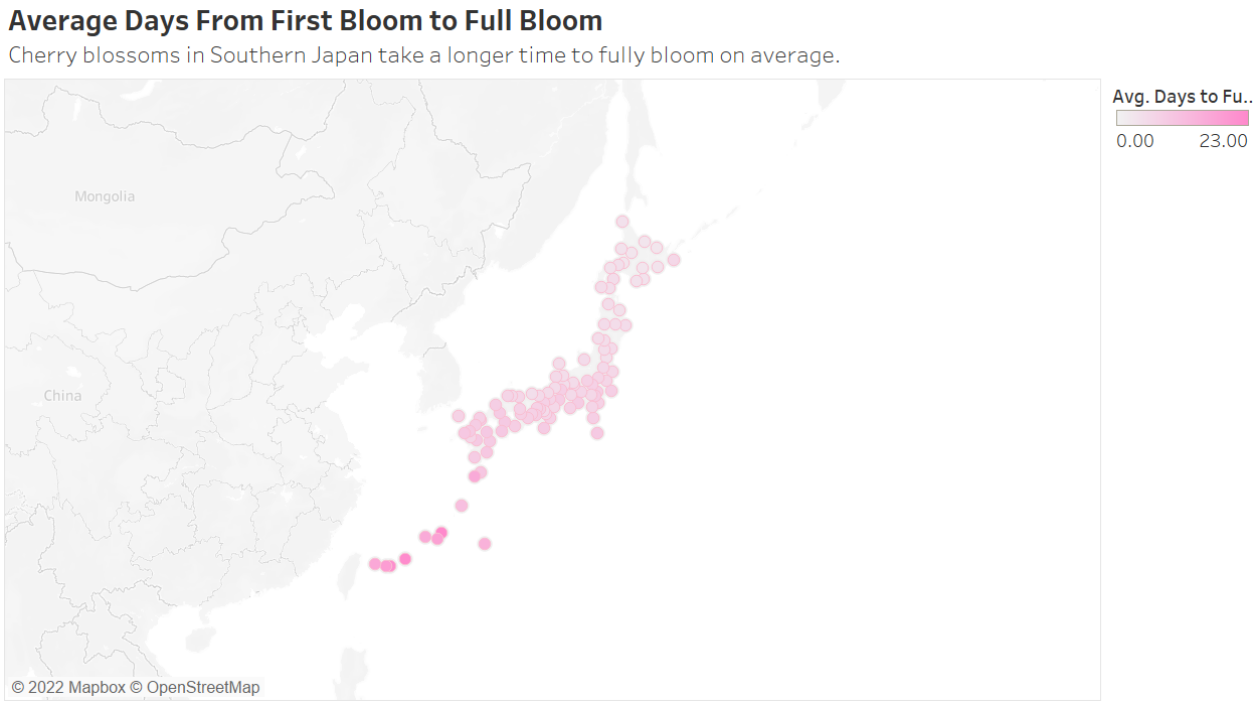

Figure 4: Average Days From First Bloom to Full Bloom (Map)

The choropleth above shows that the southern islands of Japan have slightly darker pink colors, while the rest of Japan is slightly lighter pink. This indicates that cherry blossoms in those southern islands tend to take a longer time to fully bloom on average. On the other hand, the cherry blossoms on the main island take about a week to fully bloom on average.

4 Dashboards

4.1 Dashboard for Hypothesis 1

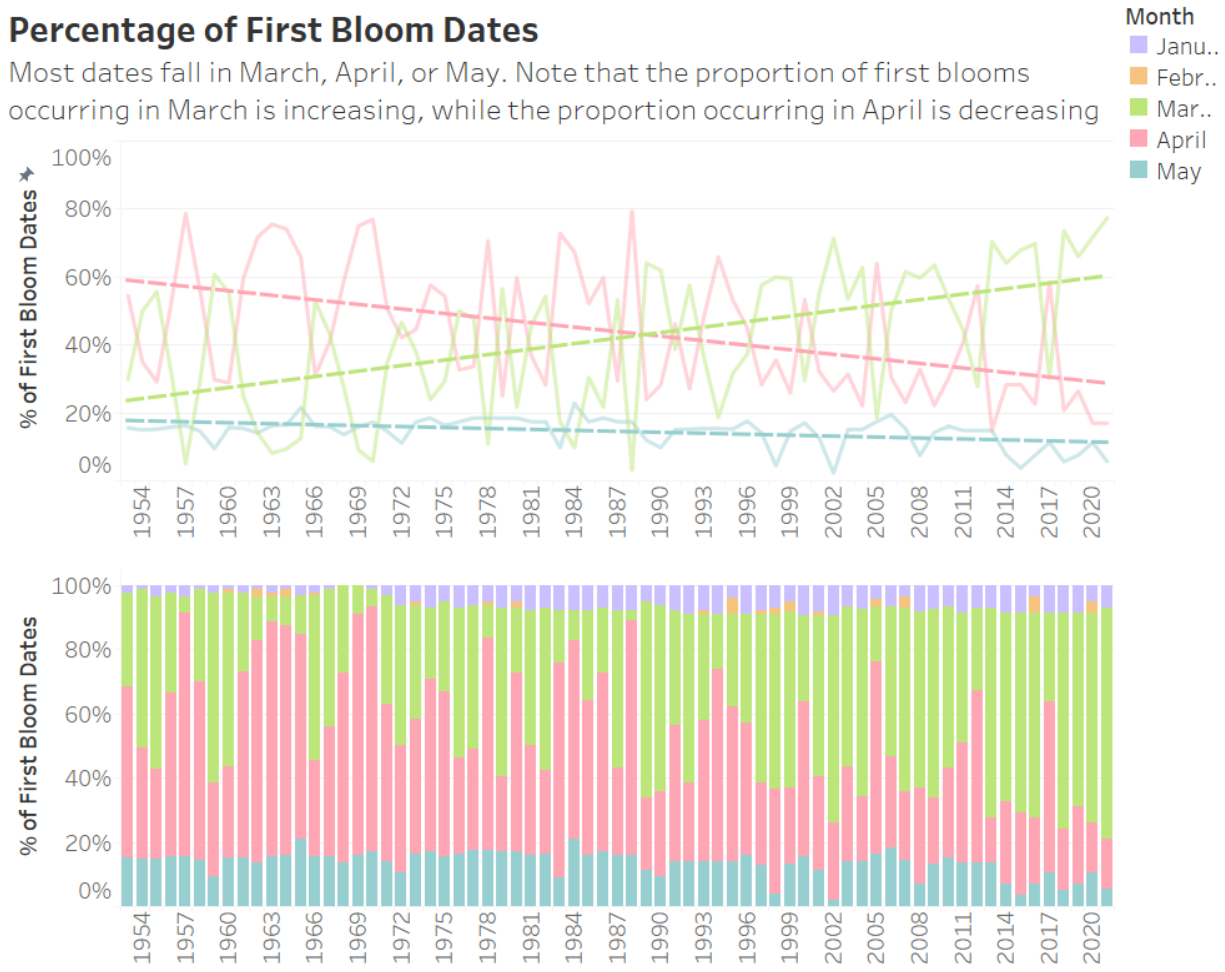

Figure 5: Dashboard 1 - Percentage of First Bloom Dates per Month

The chart in the top half shows the specific trends for March, April, and May while the chart in the bottom half shows the trends for the months January to May. The dashboard supports our hypothesis that cherry blossoms are blooming earlier in the year over time.

4.2 Dashboard for Hypothesis 2 and 3

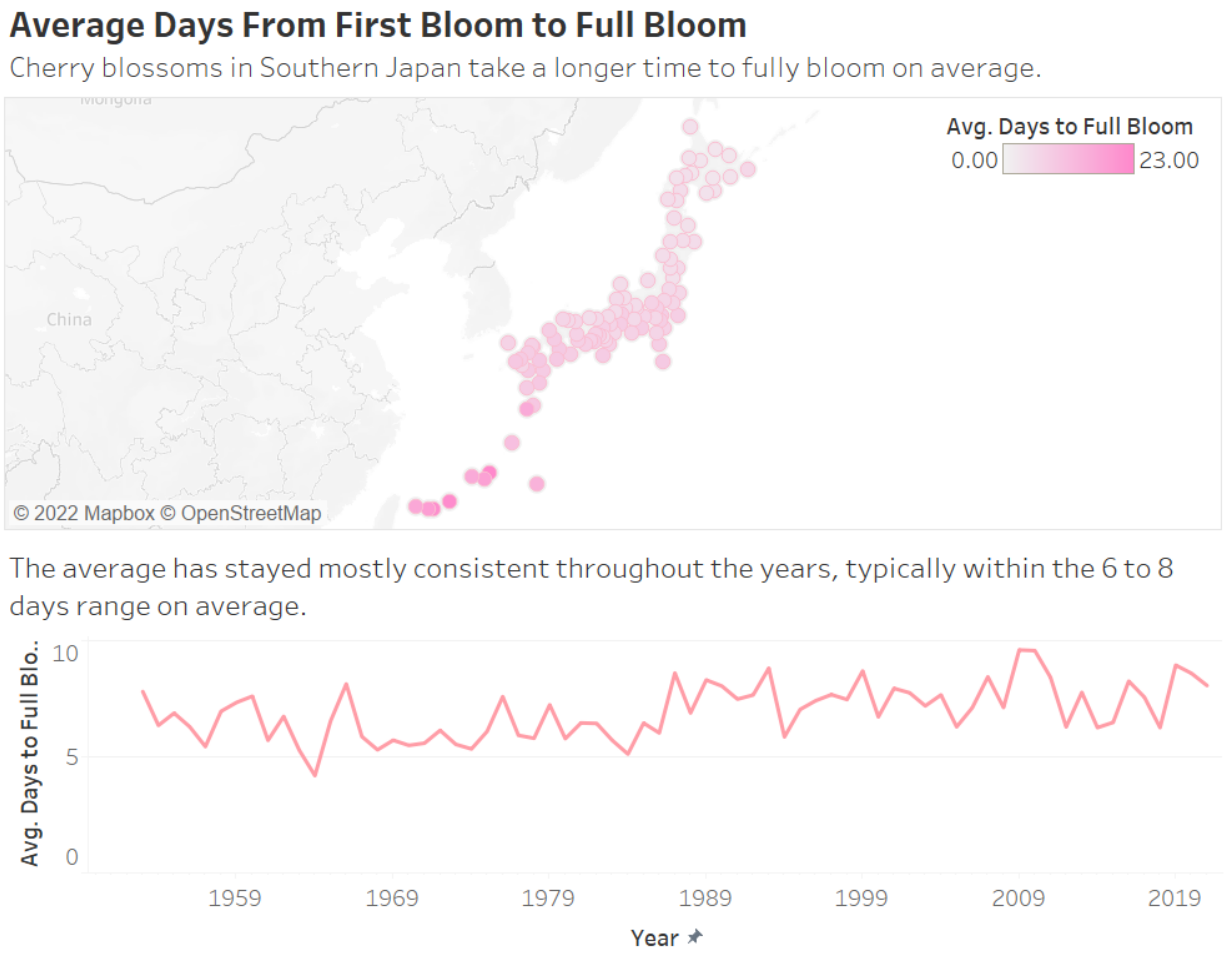

Figure 6: Dashboard 2 - Average Days from First Bloom to Full Bloom

The chart in the top half shows the average number of days to full bloom over the map of Japan, while the chart in the bottom half shows the trends over the years for the average days to full bloom. The dashboard supports the hypothesis that there are geographical trends for the blooming dates. However, it does not support the hypothesis that the cherry blossoms in the south of Japan bloom first before the cherry blossoms further north.

5 Conclusion

Although not all of my hypotheses were supported, I am not surprised at the results of the exploratory data analysis. I believe my visualizations showed clear evidence for each hypothesis because there were not many outliers or gray areas that I am concerned about. Although I cannot say what may be causing the earlier blooming of cherry blossoms, there can be other variables that I can add to this data to further my exploration. For example, I can add weather data such as temperature to determine if weather may have an association with the early blooms. Another idea for further analysis is that I can focus on particular cities or areas with high and low tourist volume. I can also compare the cherry blossom bloom dates to other flowers such as the plum blossom in Japan. There are also many connections I can make between the EDA of my other group members on tourism data.