EDA Writeup : Cherry Blossom Viewing Data from TripAdvisor

Johannes Yu

1 Initial Hypotheses or Questions

1.1 Motivation

I would like to know how region will affect the individual's rating of cherry blossom experience. My assumption is that urban areas, such as Tokyo metropolitan (東京都), with higher population density, will naturally have more locations for cherry blossom viewing, and higher counts of reviews posted, which give us an impression that we can get the best cherry blossom viewing experience in urban areas. However, I believe that the excessively crowded environment will be a negative factor for the rating for the experience of cherry blossom viewing, and therefore I aim to find out the relationship between the locations and the ratings. On top of that, I plan to gain some insight about whether group type, and time point in the year can affect tourists' behavior in rating.

1.2 Hypotheses

Hypothesis 1

Highly populated areas come with more user reviews.

Hypothesis 2

Cherry blossom viewing is a seasonal activity, which takes place in spring.

Hypothesis 3

Different group types should have different preferences regarding cherry blossom viewing.

Hypothesis 4

Tourists visiting different regions should have different levels of rating over different times of a year.

1.3 Analysis Plan

- Collect user reviews from TripAdvisor, including:
    - Location (address, coordinates)
    - Group type
    - Region
    - Date of review

    - Individual rating
- Thorough comparison between any 2 variables.

2 Data Source(s)

2.1 Description

Dataset 1 : Location Overview

- 1000+ rows of data regarding cherry blossom viewing locations in Japan.
- Metadata: location_id (str), name(str), rate(int), review_count(int), region(str), city(str), address(str), latitude(float), longitude(float), href(str)
- region: Kanto(関東地方), Kinki(近畿地方), Kyushu-Okinawa(九州・沖縄地方), Hokkaido(北海道地方), Chubu(中部地方), Shikoku(四国地方), Tohoku (東北地方), Chugoku(中国地方)

Dataset 2 : Location Reviews

- 2500+ rows collected from the first 50 reviews of the first 50 locations from the search results.
- Metadata: location_id(str), review_id(str), review_text(str), date(date), group(str), rate(int)
- group: one-person(一人), family(ファミリー), friends(友達), couple-spouse(カップル・夫婦), business(ビジネス)

2.2 Source(s)

- TripAdvisor - By using “花見” (Cherry blossom viewing in Japanese) as a keyword, I conducted a thorough web scraping against all the results (about 1000 results) to collect the addresses, overall ratings, regions and users’ individual reviews including their group types and dates when they paid their visits.
- Google Map API - The geographical coordinates can be determined by a location’s address using Google Map API

2.3 Format

.html

2.4 Transformations

Python html parsing and web scraping using BeautifulSoup and Selenium (python libraries)

3 Exploration

- Map each datapoint to its coordinates, and compare the distribution with known urban areas.

- Explore single variables and their distribution.
    - Datapoint distribution by region
    - Datapoint distribution by group type
    - Datapoint distribution by date (year and month)
    - Datapoint distribution by rating
- Compare 2 variables at a time and find their correlation.
    - Rating
        - Helps to gain insight of tourists' preference/behavior
    - Non-rating
        - Helps to understand context
- Compare multiple variables at a time and find their correlation.

4 Dashboards

4.1 Hypothesis 1: Highly populated areas come with more user reviews.

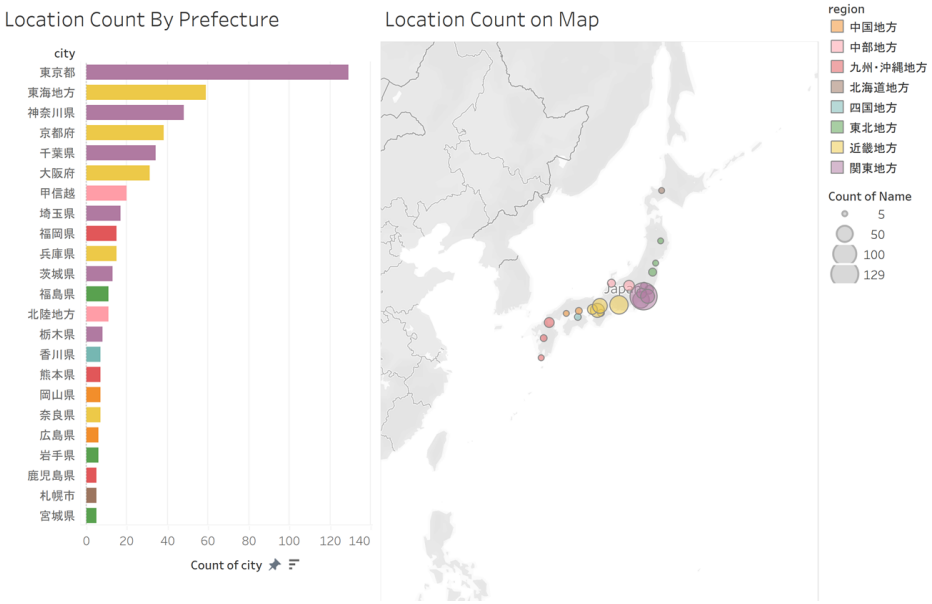

Figure 1: Geographical distribution of user review

In this bar chart on the left, we can see that prefectures from 2 of the major metropolitan areas of Japan, Kanto and Kinki, are taking the top 6 places that have the most user reviews recorded, which is assisted by the map on the right to show more accurate locations. Therefore, this hypothesis is supported by the data.

4.2 Hypothesis 2: Cherry blossom viewing is a seasonal activity, which takes place in spring.

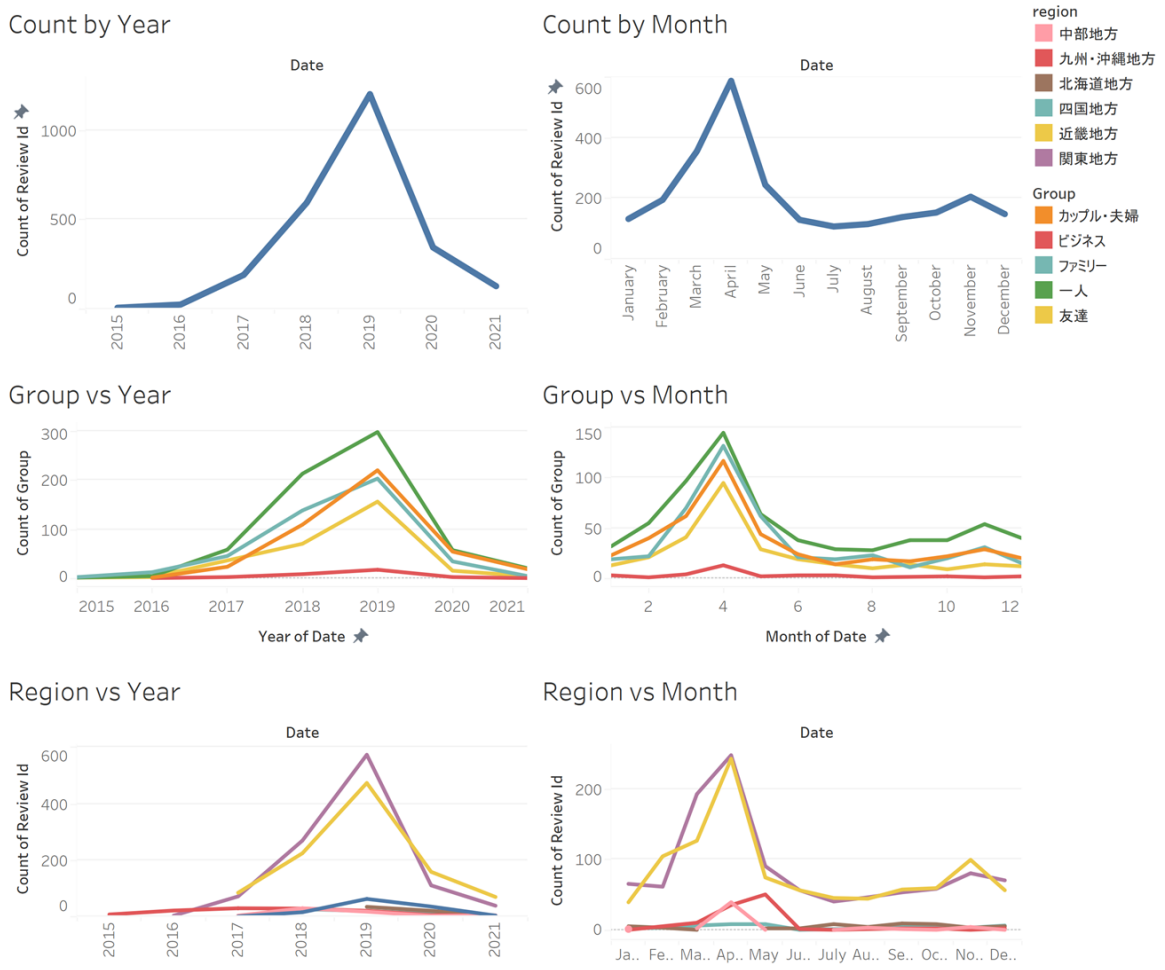

Figure 2: Review counts distribution by date

From this dashboard, we can see 2 peaks.

The column on the left shows that the number of user reviews peaked in 2019. The assumption is that the growth before 2019 came from TripAdvisor's expansion of user base, and the decline afterwards indicates that Covid-19 pandemic has lowered the tourists' interest to travel and upload reviews.

On the other hand, the accumulated review counts distribution by months shows that cherry blossom viewers are the most active from march to may, which supports my hypothesis that spring is the season for cherry blossom viewing.

4.3 Hypothesis 3: Different group types should have different preferences regarding cherry blossom viewing.

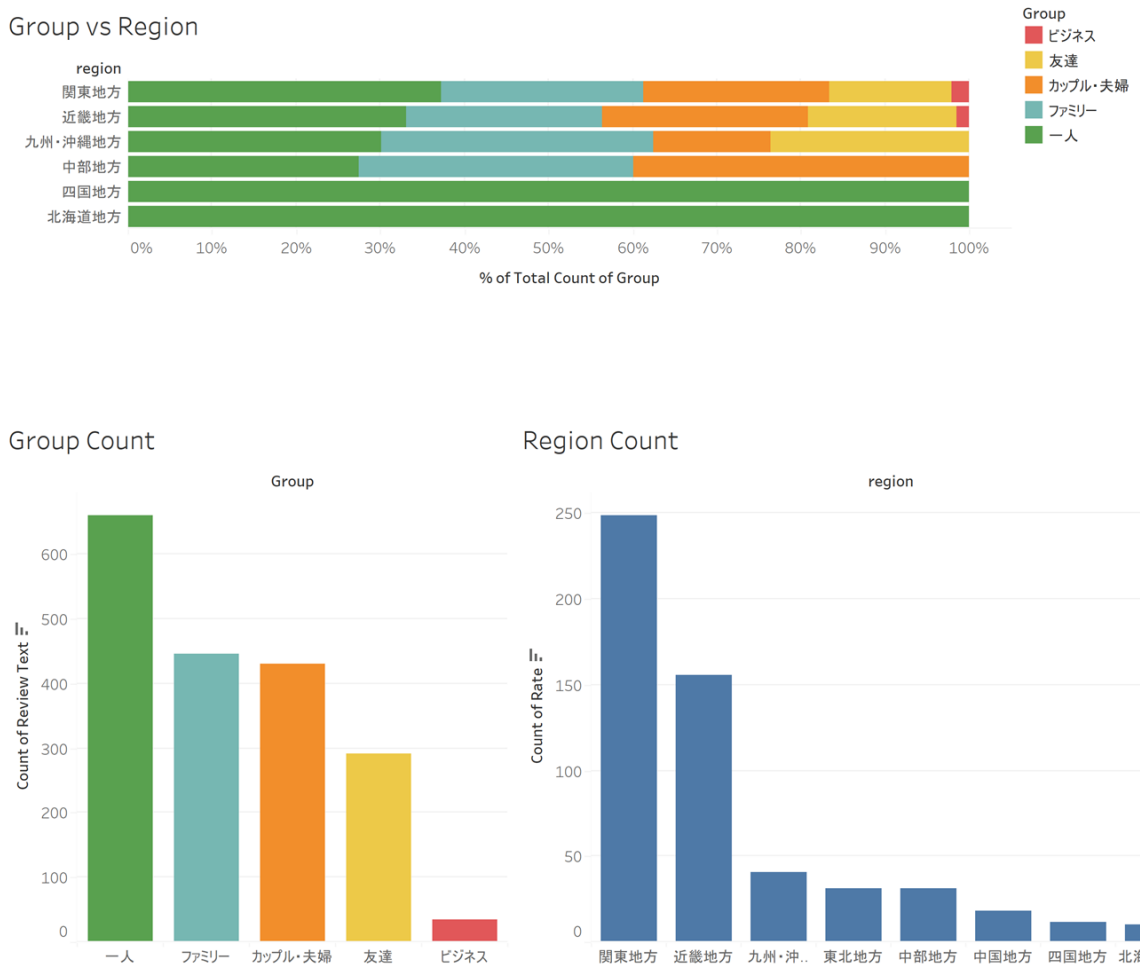

Figure 3: Group type composition by regions

From the lower left chart, we can see that the number of one-person travelers have taken the first place among all the group types, followed by the family, couple-spouse, friends, and business traveler. Also, from the bar chart on the top, we can see that there are some slight differences between the group type compositions among different regions, where in Hokkaido and Shikoku tourists are predominantly one-person tourist, and Kanto, and Kinki have a small percentage of business travelers. However, given that regions except Kanto and Kinki are having rather small sample sizes, this dataset might not be representative.

Figure 4: Rating by group types and regions

In this dashboard we can see that the rating behaviors are rather distinct among different region and group types. For example, family, one-person and friend groups rate higher in Kyushu-Okinawa than Shikoku, while business travelers rate Kanto highest, which does not contradict the hypothesis.

4.4 Hypothesis 4: Tourists visiting different regions / of different group types should have different levels of rating over different times of a year.

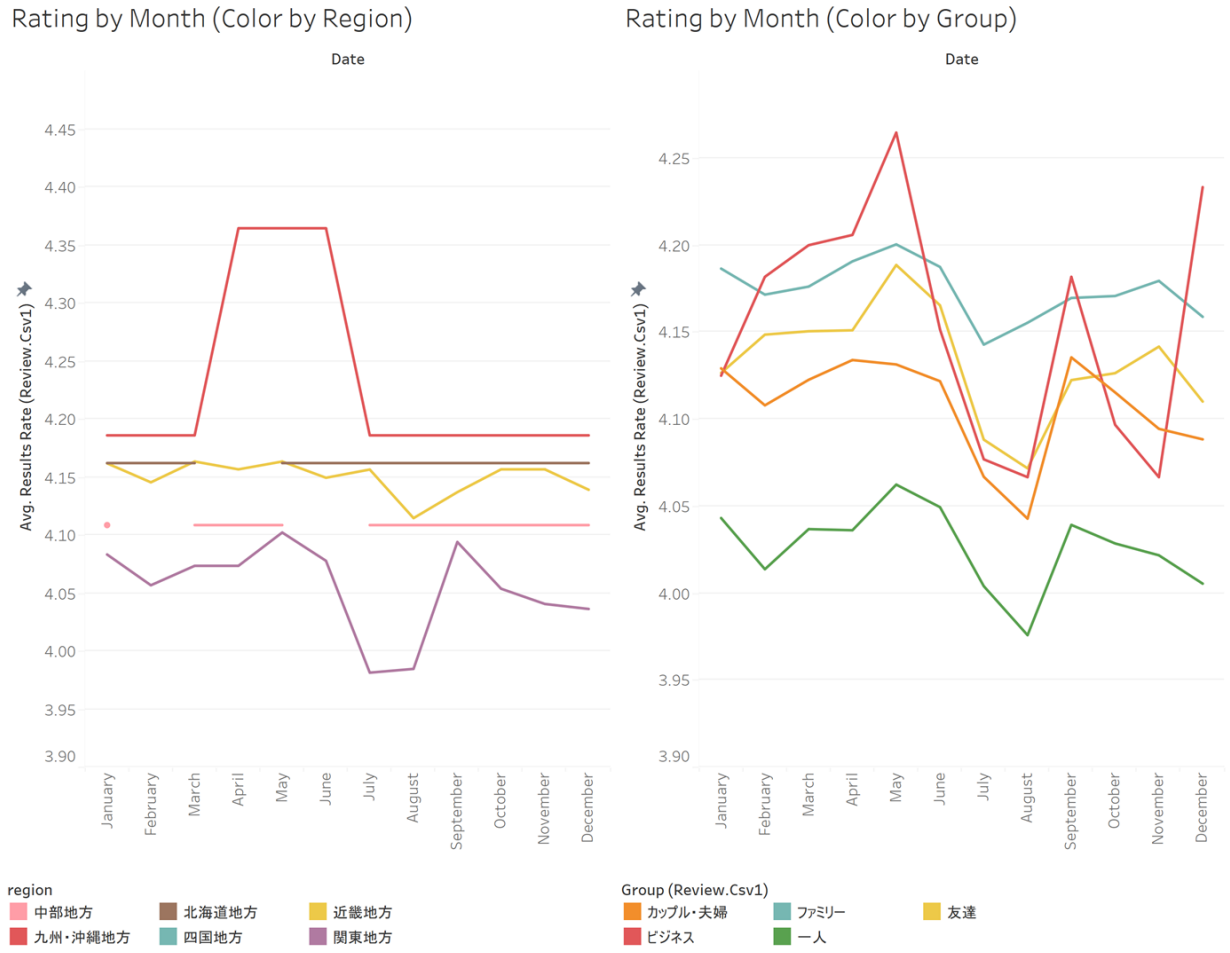

Figure 5: Rating of group types and regions by months

In the charts above, we can see that the time in a year is also a critical factor that affects the tourist rating. For example, Kyushu-Okinawa has a much higher rating during spring, whereas Kanto has a dip in rating during summer.

On the other hand, we can see that the rating difference in group types is less obvious. However, it's clear that one-person tourists give the lowest rating, the trend of business travelers' rating is much more fluctuating than the other types. Therefore, the hypothesis is supported.

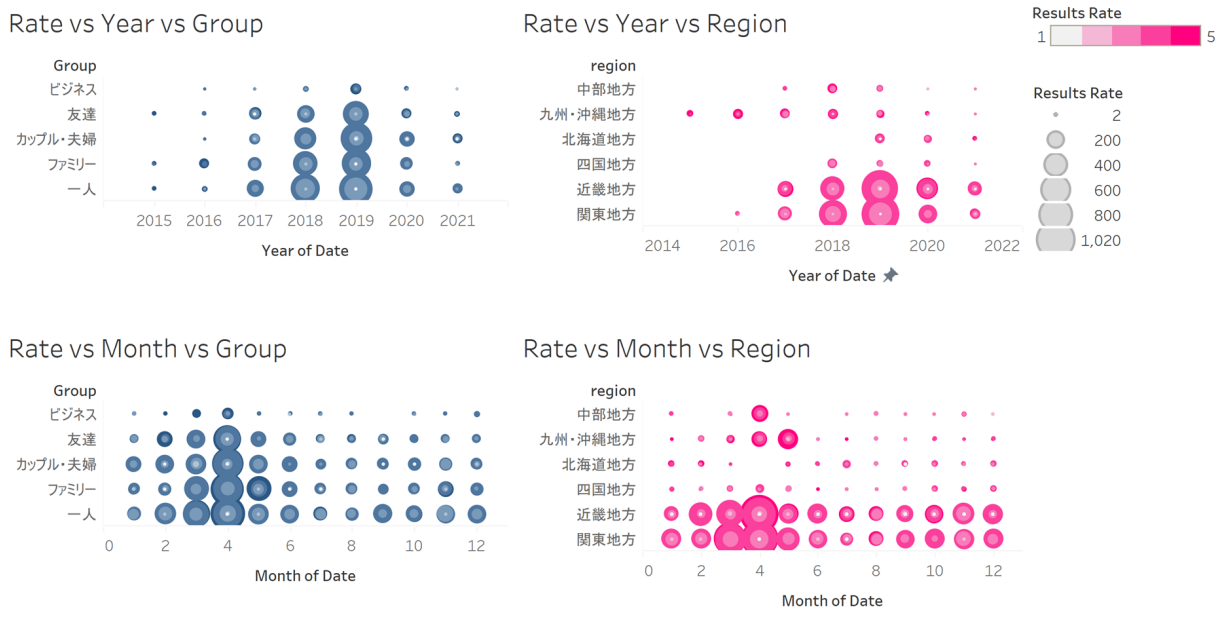

Figure 7: Rating counts of group types and regions by month and year

From these charts, it's hard to identify any insights more than what we've found from the previous visuals. However, it's still able to reiterate the overall distributions. For example, we can see that spring has the most reviews, so do Kanto and Kinki. Also, we can see that the users tend to give more ratings higher than 4, and hardly any lower than 3.

5 Conclusion

In conclusion, all of the hypotheses are not contradicted, and arguably supported by the dataset I collected. The hypotheses are as follows:

- Hypothesis 1: Highly populated areas come with more user reviews.
- Hypothesis 2: Cherry blossom viewing is a seasonal activity, which takes place in spring.
- Hypothesis 3: Different group types should have different preferences regarding cherry blossom viewing.
- Hypothesis 4: Tourists visiting different regions should have different levels of rating over different times of a year.

For the next steps, it's important to collect more data for regions outside Kanto and Kinki, so that the dataset can be more representative. Also, I'd like to conduct a comparative study the data collected with and without 花見 as the search keyword, in order to make sure the results and finding of this study are indeed relevant to cherry blossom viewing as a tourist activity.