Final Project

Yuriko Schumacher

Introduction and purpose of the project

This spring, cherry blossoms saw the record earliest first-blooming in various locations in Japan. There are many news reports, including one by The Washington Post and BBC. These articles tell concerns about the trend of first-blooming getting earlier and earlier, which they say is a sign of climate change.

The purpose of this project is to understand the statistical trend of first-blooming dates of cherry blossoms in Japan, as well as to explore some theories regarding first-blooming dates.

Exploratory data analysis questions

The following are three exploratory data analysis questions I'm going to explore in this project:

1. **Are cherry blossoms' first-blooming date really getting earlier over time?** Cherry blossoms may have bloomed the earliest ever, but it could have happened by random chance. To determine what is really happening, I'm going to investigate the general trend regarding first-blooming date over time.

For this question, I'm first going to compare the average first blooming dates over the decades, followed by geographical analysis to show in which locations cherry blossoms were blooming on April 1st over time. I'm then comparing each year's actual first-blooming dates to 30-year-average in various locations in Japan.

2. **What is the geographical distribution about whether 400- and 600-degree theories hold true?** There are several theories to determine cherry blossom's first-blooming date. One is the *400-degree theory*, (the first-blooming happens when the cumulative daily **average** temperature since February 1st reaches 400 degrees). Another is the *600-degree theory* (first-blooming occurs when the cumulative daily **high** temperature since February 1st reaches 600 degrees).

In the midterm project, I examined these theories in six major locations in Japan and concluded these theories would hold true in some locations but not others. In this section, I'm going to conduct the same analysis for all observation locations and see if there is a geographical pattern of where these theories would be true.

Moreover, for places where these theories are found not to expect the actual first-blooming date, I'm going to calculate how many days before or after the expected date (based on the theories) first-blooming is likely to happen.

3. **Is there a significant correlation between difference in days and latitude?** Based on the analysis on Question 2, I'd like to determine if there is a significant correlation between latitude and difference in days (between actual and expected first-blooming date based on 400- and 600-theories).

Data

The main data set I'm using is about cherry blossom's first-bloom dates from 1953 to 2020 in more than 100 locations in Japan, published by Japan Meteorological Agency(JMA). I downloaded a version posted on Kaggle.com.

This data set includes cherry blossom's first-blooming dates in more than 100 locations in Japan where the first-blooming were ever observed by JMA. For this project, I'm going to focus on 58 locations where observations are currently conducted.

I'm also using geographical data about JMA's observation locations, available on the agency's list of observatories.

Finally, to analyze the second and third questions, I'm using data about daily high and average temperature in more than 50 locations in Japan. Those data sets were directly downloaded from JMA, by using the agency's downloading system.

For all tests in this project, I'm going to apply the alpha value of **0.05**.

Load data and conduct an initial analysis

First, I'm going to load the data set I'm going to mainly analyze on this project.

Load first-blooming date data

```
# load data set about cherry blossom's first-blooming date
sakura_raw <- read_csv("data/sakura_first_bloom_dates.csv")

# tidy data
sakura_gathered <- sakura_raw %>%
  select(-`30 Year Average 1981-2010`, -Notes) %>%
  gather(key = "year", value = "date", "1953":"2020") %>%
  rename(location = `Site Name`,
         now_observed = `Currently Being Observed`)

# filter out locations currently not observed, calculate date of the year
sakura <- sakura_gathered %>%
  filter(now_observed == TRUE) %>%
  mutate(
    date_x = ymd(date),
    # Not using lubridate::floor_date() function because some dates should be calculated as negative va
    floor_date = as.Date(paste(year, "-01-01", sep = ""), format = "%Y-%m-%d"),
    doy_x = interval(floor_date, date_x) %>% time_length(unit = "day")) %>%
  select(-floor_date, -now_observed, -date)

head(sakura)
```

```
## # A tibble: 6 x 4
##   location  year  date_x     doy_x
##   <chr>     <chr> <date>     <dbl>
## 1 Wakkanai  1953  1953-05-21   140
## 2 Asahikawa 1953  1953-05-11   130
## 3 Abashiri  1953  1953-05-24   143
## 4 Sapporo   1953  1953-05-07   126
## 5 Obihiro   1953  1953-05-15   134
## 6 Kushiro   1953  NA             NA
```

The cleaned data set has 3944 rows and 4 columns. Each row shows an observation, and columns are as follows:

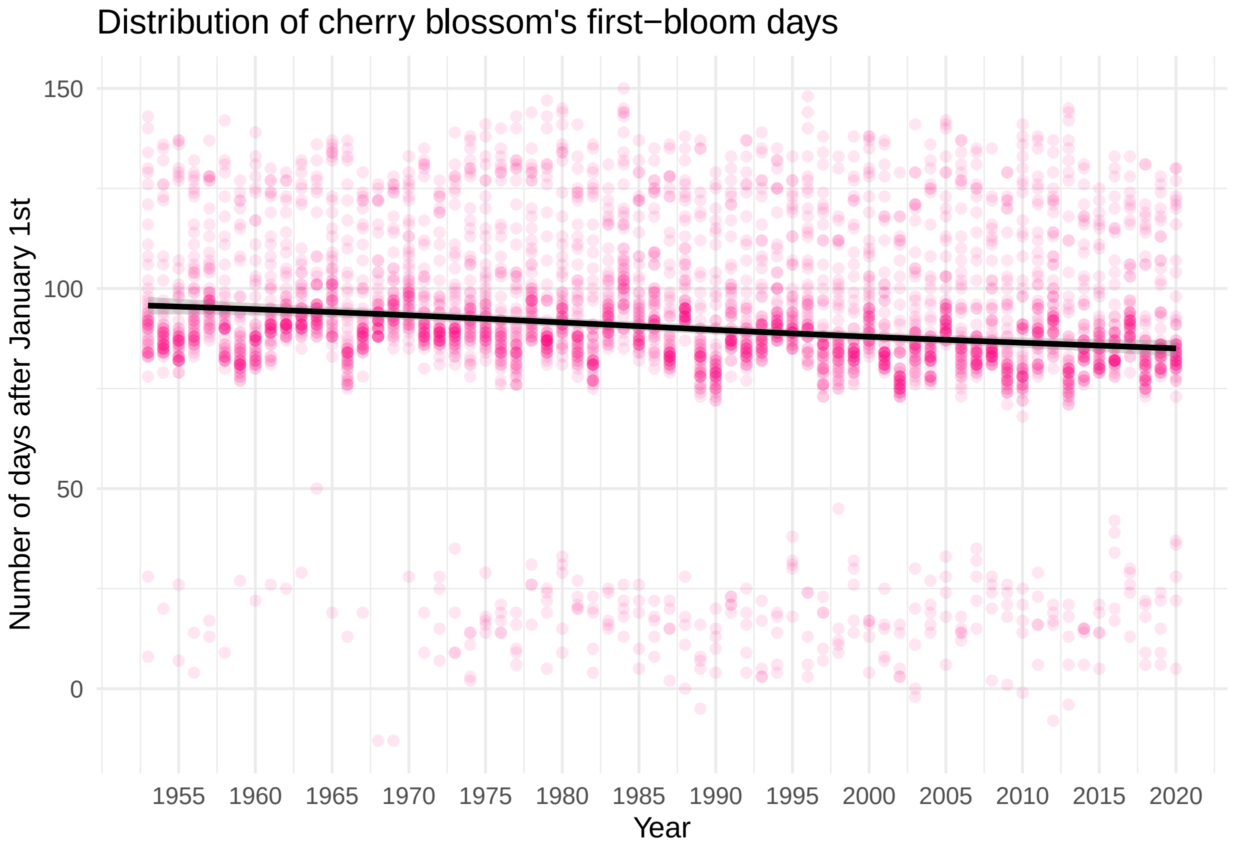
- **location**: the location the cherry blossom's first-blooming was observed
- **year**: the year of observation
- **date_x**: date of first-blooming (Year-Month-Day)
- **doy_x**: date of first-blooming, converted as the number of days after January 1st

I'm computing the dates as number of days after January 1st. Negative numbers mean the first-blooming was before January 1st (late December). I count those cases as first-blooming in the coming spring (thus negative number of days after January 1st).

Load observation location data and conduct an initial analysis

First, I'm going to plot the data to see overall distribution.

```r
# First exploratory plot
sakura %>%
  group_by(year) %>%
  ggplot(aes(x = as.numeric(year), y = doy_x)) +
  geom_point(color = "#ff0080", alpha = 0.1) +
  geom_smooth(color = "black") +
  scale_x_continuous(breaks = c(1955, 1960, 1965, 1970, 1975, 1980, 1985, 1990,
                                1995, 2000, 2005, 2010, 2015, 2020)) +
  labs(title = "Distribution of cherry blossom's first-bloom days",
       x = "Year", y = "Number of days after January 1st") +
  theme_minimal()
```



As the visualization above shows, there are a group of observations that are plotted far from the majority of observations. In those locations, first-blooming dates of the year are below 50 days after January 1st. Where are those locations?

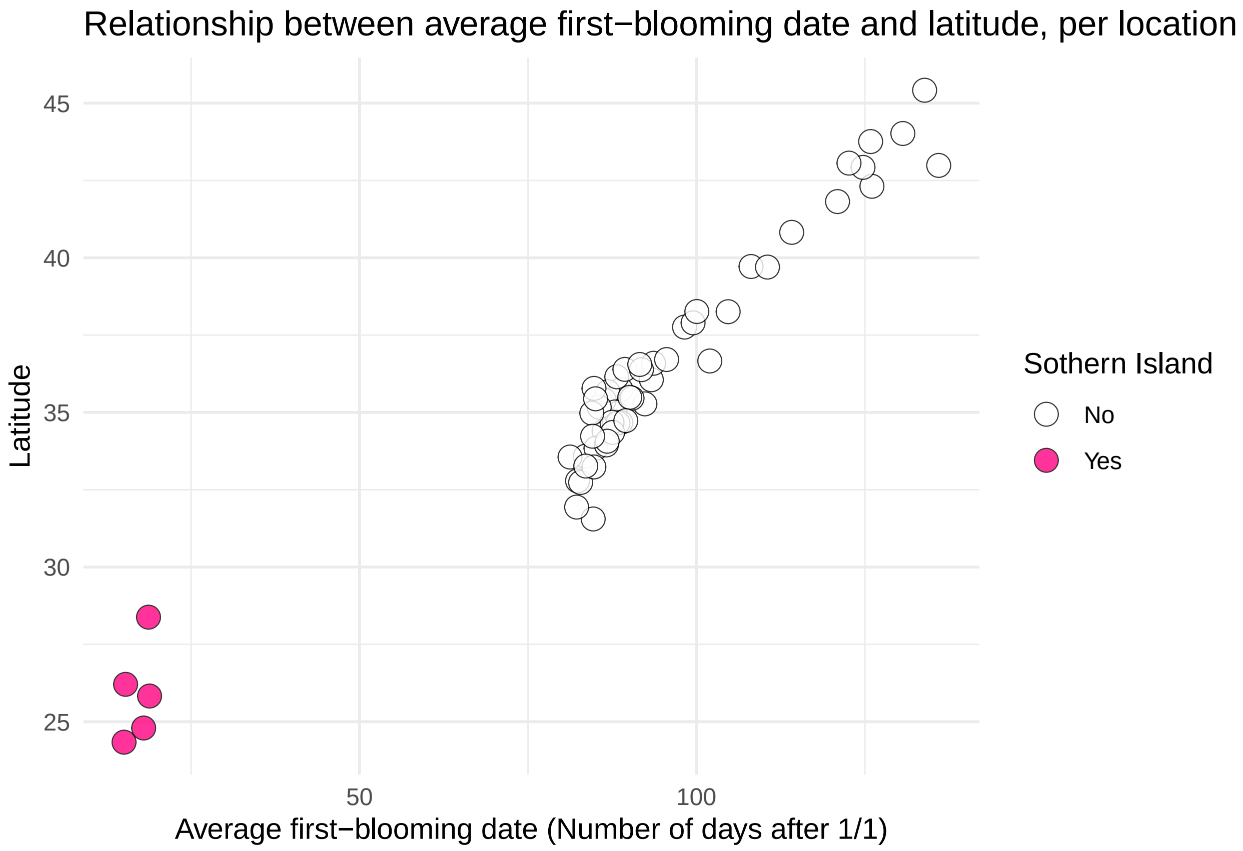
```
# Look for locations with extremely early first-blooming dates
sakura %>%
  arrange(doy_x)
```

```
## # A tibble: 3,944 x 4
##    location        year  date_x     doy_x
##    <chr>           <chr> <date>     <dbl>
##  1 Ishigaki Island 1968  1967-12-19   -13
##  2 Ishigaki Island 1969  1968-12-19   -13
##  3 Ishigaki Island 2012  2011-12-24    -8
##  4 Ishigaki Island 1989  1988-12-27    -5
##  5 Naha            2013  2012-12-28    -4
##  6 Miyakojima      2003  2002-12-30    -2
##  7 Naha            2010  2009-12-31    -1
##  8 Ishigaki Island 1988  1988-01-01     0
##  9 Naze            2003  2003-01-01     0
## 10 Naha            2009  2009-01-02     1
## # ... with 3,934 more rows
```

When we look at the data set, arranged by *doy_x*, we can see those extremely early first-blooming were observed in five sothern locations in Japan – Naze, Ishigaki Island, Miyakojima, Naha, and Minamidaitojima.

To confirm this, I'm going to visualize the relationship between latitude and average first-blooming dates per locations.

```r
# Load observation location data
observation.raw <- read_csv('data/observatory-locations.csv')

# Save location names for sothern islands
sothern_islands <- c("Naze", "Ishigaki Island", "Miyakojima", "Naha", "Minamidaitojima")

# Combine observation location data and first-blooming dates, plot the relationship
sakura %>%
  group_by(location) %>%
  summarize(avg_doy_x = mean(doy_x, na.rm = TRUE)) %>%
  cbind(observation.raw) %>%
  select(-3) %>%
  mutate(is_sothern = ifelse(location %in% sothern_islands, "Yes", "No")) %>%
  ggplot(aes(x = avg_doy_x, y = lat)) +
  geom_point(aes(fill = is_sothern),
             pch = 21, color = "black", size = 4, stroke = 0.3, alpha = 0.8) +
  scale_fill_manual(values = c("white", "#ff0080")) +
  labs(title = "Relationship between average first-blooming date and latitude, per location",
       x = "Average first-blooming date (Number of days after 1/1)", y = "Latitude",
       fill = "Sothern Island") +
  theme_minimal()
```



As shown in the visualization, as I suspected, those five locations are far from majority of data, thus can be considered as outliers. Therefore, for this project, I'm going to filter out those five locations from all analyses. So for this project, I'm going to conduct analyses over 53 locations in total.

```
# filter out those five locations from the data set
sakura <- sakura %>%
    filter(location != "Naze",
           location != "Naha",
           location != "Minamidaitojima",
           location != "Miyakojima",
           location != "Ishigaki Island")

observation <- observation.raw %>%
    filter(location != "Naze",
           location != "Naha",
           location != "Minamidaitojima",
           location != "Miyakojima",
           location != "Ishigaki Island")
```

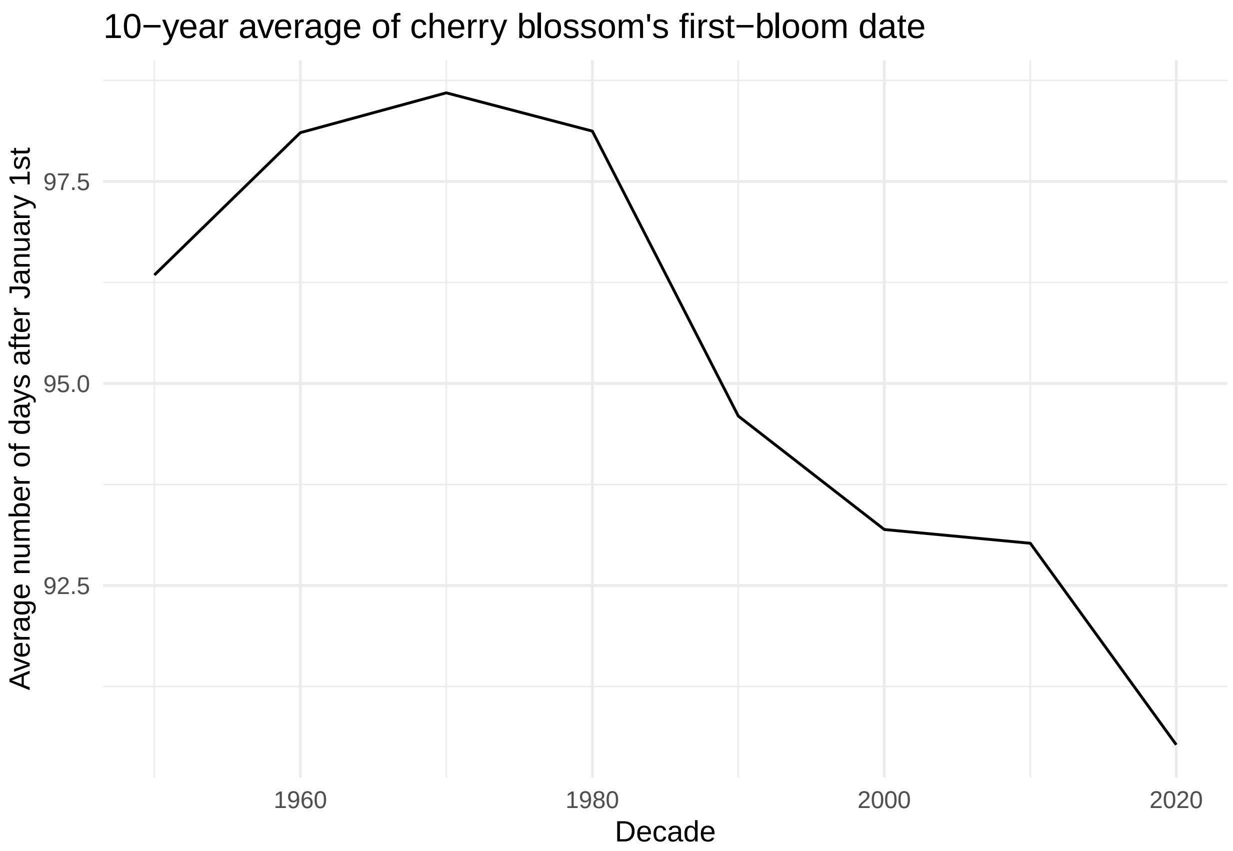
Question 1: Are cherry blossoms' first-blooming date really getting earlier over time?

In this section, I'm going to examine whether first-blooming is really getting earlier over time. First, I'm going to repeat my analysis as I did in the midterm project as an introduction.

Decade average

Here, I'm going to calculate the average first-blooming date per decades and draw a line chart.

```
# Create a new data set that contain the average per decades
sakura_decade <- sakura %>%
  mutate(decade = floor_date(date_x, unit = "10 years") %>% year()) %>%
  group_by(location, decade) %>%
  summarize(avg_doy_x = mean(doy_x, na.rm = TRUE))

# Draw a line chart
sakura_decade %>%
  group_by(decade) %>%
  summarize(avg_doy_x = mean(avg_doy_x)) %>%
  ggplot(aes(x = decade, y = avg_doy_x)) +
  geom_line() +
    labs(title = "10-year average of cherry blossom's first-bloom date",
       x = "Decade", y = "Average number of days after January 1st") +
  theme_minimal()
```



The result is different from what I showed in the midterm project, since I excluded five southern locations from this. From this visualization, we can see the average first-blooming dates actually became later during 1950s and 1960s, before they started getting earlier in the 1970s. During 1980s and 2010s, the average dates changed drastically and became much earlier.

Paired-sample t-test

Next, I'm going to conduct a paired-sample t-test to determine if there is a significant difference in average first-blooming dates over decades. I'm going to compare data about 1950s to 1960s, 1960s to 1970s, . . . , and 2010s to 2020s – repeat the test seven times. In this case, the observations average first-blooming dates in two different decades are paired by observation location.

The hypotheses are as follows:

- **Null hypothesis**: there is no significant difference in average first-blooming dates between decades
- **Alternative hypothesis**: there is a significant difference in average first-blooming dates between decades

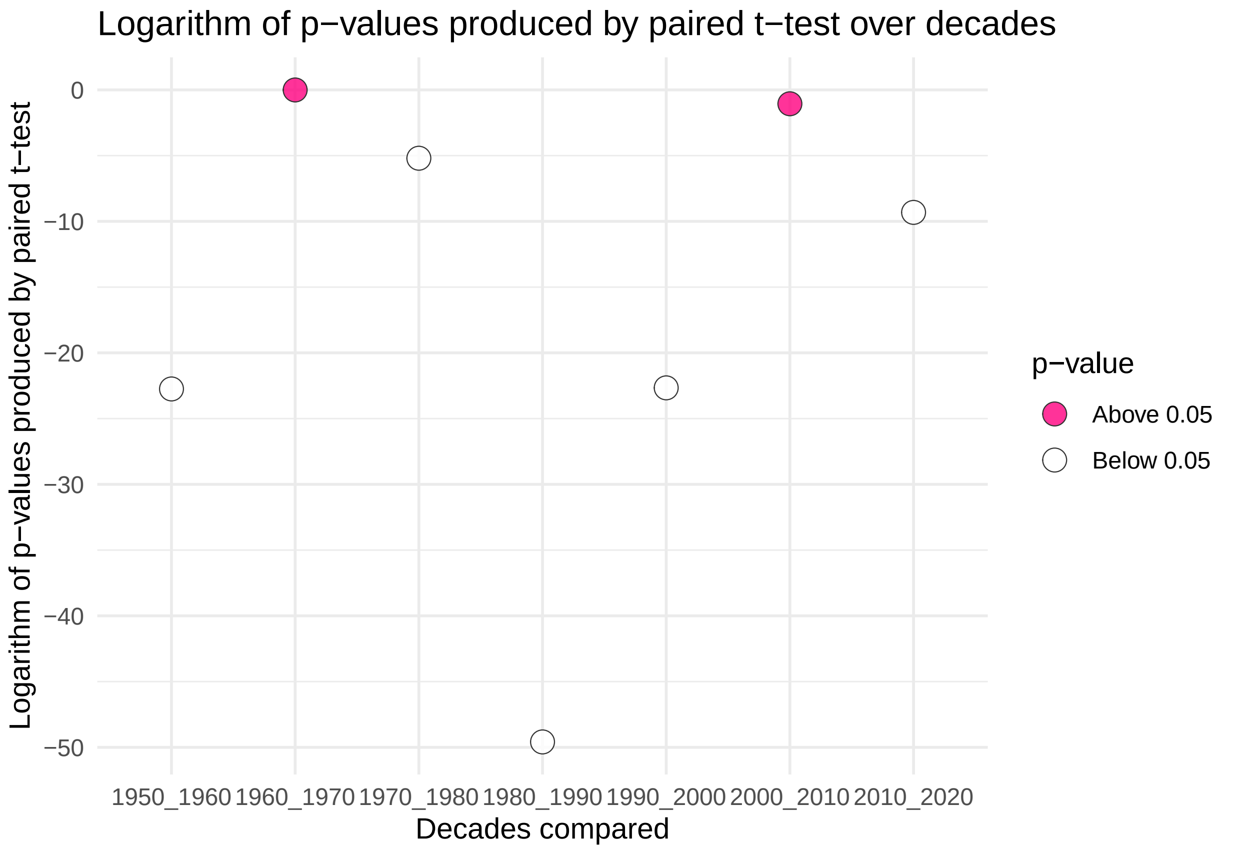
Note There are no data available for the locations in Kushiro and Miyazaki in the 1950s and 1960s (In Kushiro, available data are after 1972, while in Miyazaki, available data are after 1971). So when comparing data from 1950s to 1960s, and from 1960s to 1970s, I'm using data about 51 locations, excluding Kushiro and Miyazaki.

```
# Conduct paired t-test to compare data in the 1950s to 1960s
paired1950_1960 <- sakura_decade %>%
```

```r
  filter(decade %in% c(1950, 1960))
p_50_60 <- t.test(avg_doy_x ~ decade, data = paired1950_1960, paired = TRUE)$p.value # p-value = 1.323e

# Conduct paired t-test to compare data in the 1950s to 1960s
# Exclude data about Kushiro and Miyazaki in the 1970s
paired1960_1970 <- sakura_decade %>%
  filter(decade %in% c(1960, 1970)) %>%
  filter(location != "Kushiro") %>%
  filter(location != "Miyazaki")
p_60_70 <- t.test(avg_doy_x ~ decade, data = paired1960_1970, paired = TRUE)$p.value # p-value = 0.9933

# Conduct paired t-test to compare data in the 1970s to 1980s
paired1970_1980 <- sakura_decade %>%
  filter(decade %in% c(1970, 1980))
p_70_80 <- t.test(avg_doy_x ~ decade, data = paired1970_1980, paired = TRUE)$p.value # p-value = 0.0055

# Conduct paired t-test to compare data in the 1980s to 1990s
paired1980_1990 <- sakura_decade %>%
  filter(decade %in% c(1980, 1990))
p_80_90 <- t.test(avg_doy_x ~ decade, data = paired1980_1990, paired = TRUE)$p.value # p-value < 2.2e-1

# Conduct paired t-test to compare data in the 1990s to 2000s
paired1990_2000 <- sakura_decade %>%
  filter(decade %in% c(1990, 2000))
p_90_00 <- t.test(avg_doy_x ~ decade, data = paired1990_2000, paired = TRUE)$p.value # p-value = 1.44e-

# Conduct paired t-test to compare data in the 2000s to 2010s
paired2000_2010 <- sakura_decade %>%
  filter(decade %in% c(2000, 2010))
p_00_10 <- t.test(avg_doy_x ~ decade, data = paired2000_2010, paired = TRUE)$p.value # p-value = 0.3469

# Conduct paired t-test to compare data in the 2010s to 2020s
paired2010_2020 <- sakura_decade %>%
  filter(decade %in% c(2010, 2020))
p_10_20 <- t.test(avg_doy_x ~ decade, data = paired2010_2020, paired = TRUE)$p.value # p-value = 8.977e
```

Based on the p-values generated by t-test, I'm going to visualize which decades saw significant change in the average first-blooming dates.

```r
# Create data frame from p-values
decades <- c("1950_1960", "1960_1970", "1970_1980", "1980_1990", "1990_2000", "2000_2010", "2010_2020")
paired_pVal <- c(p_50_60, p_60_70, p_70_80, p_80_90, p_90_00, p_00_10, p_10_20)

# Plot p-values and whether they are above/below alpha level
data.frame(decades, paired_pVal) %>%
  mutate(is_above_alpha = ifelse(paired_pVal > 0.05, "Above 0.05", "Below 0.05")) %>%
  ggplot(aes(x = decades, y = log(paired_pVal))) +
  geom_point(aes(fill = is_above_alpha),
             pch = 21, color = "black", size = 4, stroke = 0.3, alpha = 0.8) +
  scale_fill_manual(values = c("#ff0080", "white")) +
  labs(title = "Logarithm of p-values produced by paired t-test over decades",
       x = "Decades compared", y = "Logarithm of p-values produced by paired t-test",
       fill = "p-value") +
  theme_minimal()
```



Based on the result, comparison between 1960s and 1970s, as well as between 2000s and 2010s generated p-values above alpha level of 0.05. Thus, for those comparisons, I accept the null hypothesis and conclude there is no significant difference between those decades. On the other hand, as for other five comparisons, p-values are below 0.05, so I reject the null hypothesis and accept the alternative hypothesis – there is a significant difference between those decades in the average first-blooming dates.

In comparison between other decades, the average first-blooming dates are significantly different. Given the line chart generated in the previous chunk, we can say, from 1970s to 1980s, from 1980s and 1990s, from 1990s and 2000s, and from 2010s and 2020s, the average first blooming dates became significantly earlier.

Mapping what April 1st looked like in the past

Next, assuming the first-blooming dates are getting earlier over decades, I'm going to conduct a geographical analysis about whether cherry blossoms were blooming on the April first over the decades.

I'm using the library "maps" to draw a Japanese map. Since I'm excluding data about southern islands from this project I'm only mapping a map above 31 degrees north.

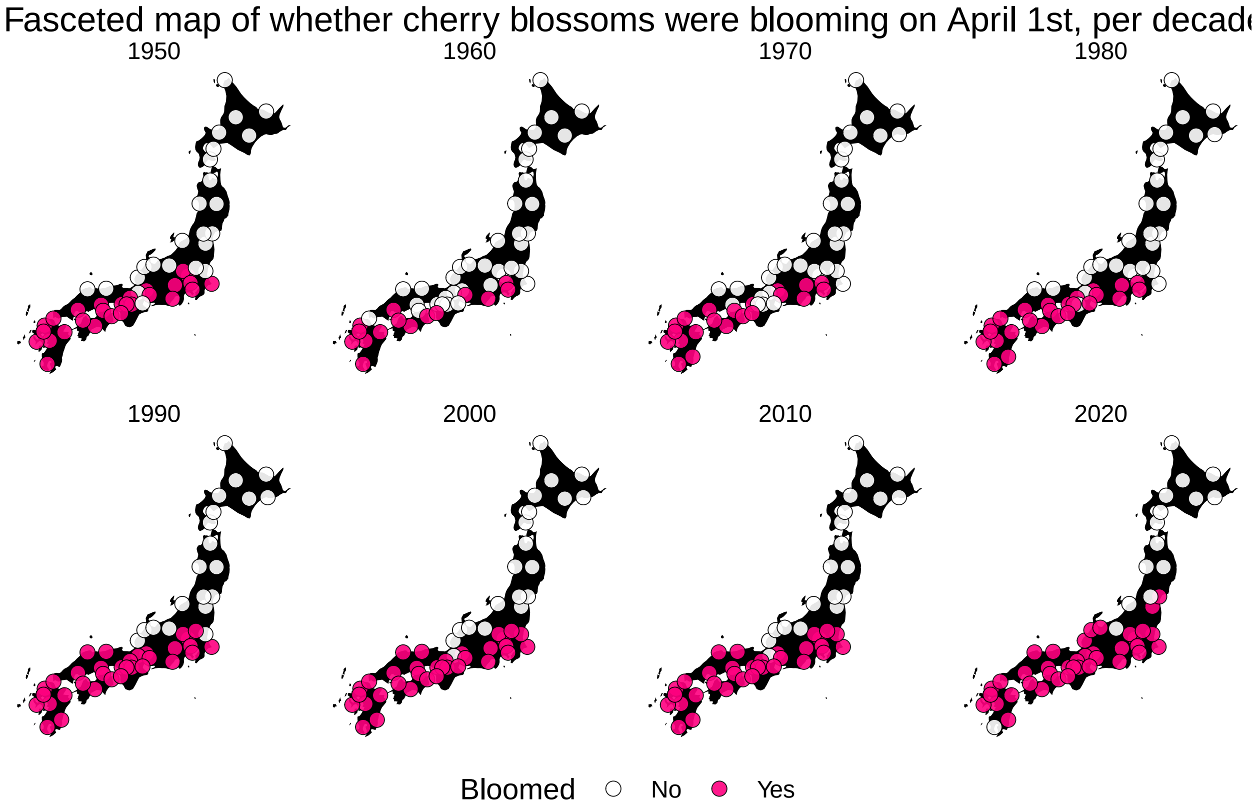
```
# Store Japanese map data into an object, using maps::map_data() function
japan <- map_data("world") %>%
  filter(region == "Japan")

# Filter out data from southern islands
japan <- japan %>%
  filter(lat > 31)
```

Next, I'm going to merge the location data to average first-blooming date data. I'm binding location data for each decade, before merging all the data to plot points on the map.

```r
# Bind location data to each decade data
sakura1950 <- sakura_decade %>%
  ungroup() %>%
  filter(decade == 1950) %>%
  add_row(location = "Kushiro", decade = 1950, avg_doy_x = NA) %>%
  add_row(location = "Miyazaki", decade = 1950, avg_doy_x = NA) %>%
  arrange(location) %>%
  cbind(observation) %>%
  select(-4)

sakura1960 <- sakura_decade %>%
  ungroup() %>%
  filter(decade == 1960) %>%
  add_row(location = "Kushiro", decade = 1960, avg_doy_x = NA) %>%
  add_row(location = "Miyazaki", decade = 1960, avg_doy_x = NA) %>%
  arrange(location) %>%
  cbind(observation) %>%
  select(-4)

sakura1970 <- sakura_decade %>%
  ungroup() %>%
  filter(decade == 1970) %>%
  arrange(location) %>%
  cbind(observation) %>%
  select(-4)

sakura1980 <- sakura_decade %>%
  ungroup() %>%
  filter(decade == 1980) %>%
  arrange(location) %>%
  cbind(observation) %>%
  select(-4)

sakura1990 <- sakura_decade %>%
  ungroup() %>%
  filter(decade == 1990) %>%
  arrange(location) %>%
  cbind(observation) %>%
  select(-4)

sakura2000 <- sakura_decade %>%
  ungroup() %>%
  filter(decade == 2000) %>%
  arrange(location) %>%
  cbind(observation) %>%
  select(-4)

sakura2010 <- sakura_decade %>%
  ungroup() %>%
  filter(decade == 2010) %>%
  arrange(location) %>%
  cbind(observation) %>%
  select(-4)

sakura2020 <- sakura_decade %>%
  ungroup() %>%
  filter(decade == 2020) %>%
  arrange(location) %>%
  cbind(observation) %>%
  select(-4)

# Merge all the data in order to plot points on the map
sakura_decade_merged <- rbind(
  sakura1950, sakura1960, sakura1970, sakura1980,
  sakura1990, sakura2000, sakura2010, sakura2020
) %>%
  mutate(Bloomed = ifelse(avg_doy_x > 90, "No", "Yes"))

# Exclude NA values
sakura_decade_merged <- sakura_decade_merged[complete.cases(sakura_decade_merged), ]
```

Finally, I'm going to plot the map.

```
# Plot the fasceted map
ggplot() +
  geom_polygon(data = japan %>% filter(lat > 31),
               aes(x = long, y = lat, group = group),
               fill = "#000000") +
  geom_point(data = sakura_decade_merged,
             aes(x = long, y = lat, fill = Bloomed),
             pch = 21, color = "black", size = 2.5, stroke = 0.25, alpha = 0.9) +
  scale_fill_manual(values = c("#ffffff", "#ff0080", "#AAAAAA")) +
  facet_wrap(vars(decade), ncol = 4) +
  theme_void() +
  theme(legend.position = "bottom") +
  labs(title = "Fasceted map of whether cherry blossoms were blooming on April 1st, per decade") +
  coord_map()
```


Fasceted map of whether cherry blossoms were blooming on April 1st, per decad
1950
1960
1970
1980
1990
2000
2010
2020
Bloomed
○ No
● Yes

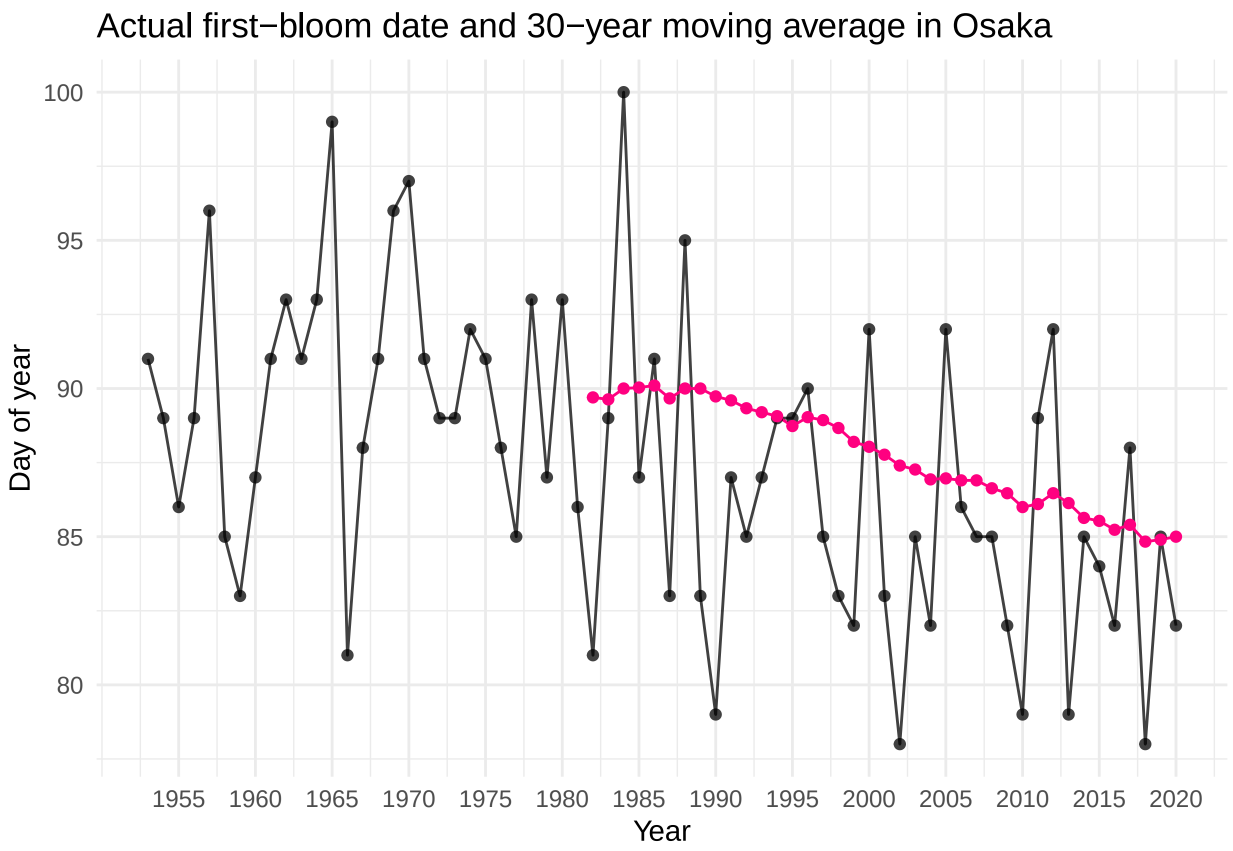
The visualization shows whether cherry blossoms, on average, were blooming on April 1st per decade, over time. From the visualization, we can see in the 1960s and 1970s, many locations in Chugoku, Kinki, and Kanto area are colored as white, meaning cherry blossoms were yet to have first-blooming. However, in recent decades, they are mostly past-blooming. Chubu-Hokuriku area are mostly pre-blooming even in recent years, but in 2020 (and also 2021), they were already blooming on April 1st.

Compare actual first-blooming dates and 30-year moving average

Next, I'm going to look at the change in first-blooming dates from another point of view. I'm going to calculate the 30-year-average date per location, set the date as an expected first-blooming date for the next year, and conduct a one-sample t-test to see if those expectations would hold true.

To begin, I'm going to focus on one location: Osaka.

```
# Filter data about Osaka and compute 30-year-average
osaka <- sakura %>%
  filter(location == "Osaka") %>%
  mutate(doy_30yr = rollmean(doy_x, k = 30, fill = NA, align = "right"))

# Draw a line chart
osaka %>%
  ggplot(aes(x = as.numeric(year))) +
  geom_line(aes(y = doy_x, group = 1), alpha = 0.75) +
  geom_line(aes(y = doy_30yr, group = 1), color = "#ff0080") +
  geom_point(aes(y = doy_x), alpha = 0.75) +
  geom_point(aes(y = doy_30yr), color = "#ff0080") +
    scale_x_continuous(breaks = c(1955, 1960, 1965, 1970, 1975, 1980, 1985, 1990,
                                   1995, 2000, 2005, 2010, 2015, 2020)) +
  labs(title = "Actual first-bloom date and 30-year moving average in Osaka",
       x = "Year", y = "Day of year") +
  theme_minimal()
```



The pink dots and line show 30-year-average, while the black dots and line show actual observed first-blooming dates. From this visualization, it looks like in recent years, 30-year-average first-blooming dates are getting earlier.

Next, I'm going to set the 30-year-average as an expected first-blooming date, and compare the actual blooming date to determine if there is a significant difference between 30-year-average and actual first-blooming date. In particular, since I'm assuming the actual first-blooming dates are earlier than the expected date, I'm conducting a one-tailed one-sample t-test on this.

The hypothesis are: * **Null hypothesis**: the difference between the actual first-bloom date and 30-year-average of first-bloom date is 0 days * **Alternative hypothesis**: the difference between the actual first-bloom date and 30-year-average of first-bloom date is greater than 0 days

```
# Conduct one-sample t-test
osaka_diff <- osaka %>%
  mutate(diff = doy_30yr - doy_x)

t.test(x = osaka_diff$diff, mu = 0, alternative = "greater") # p-value = 0.002531
```

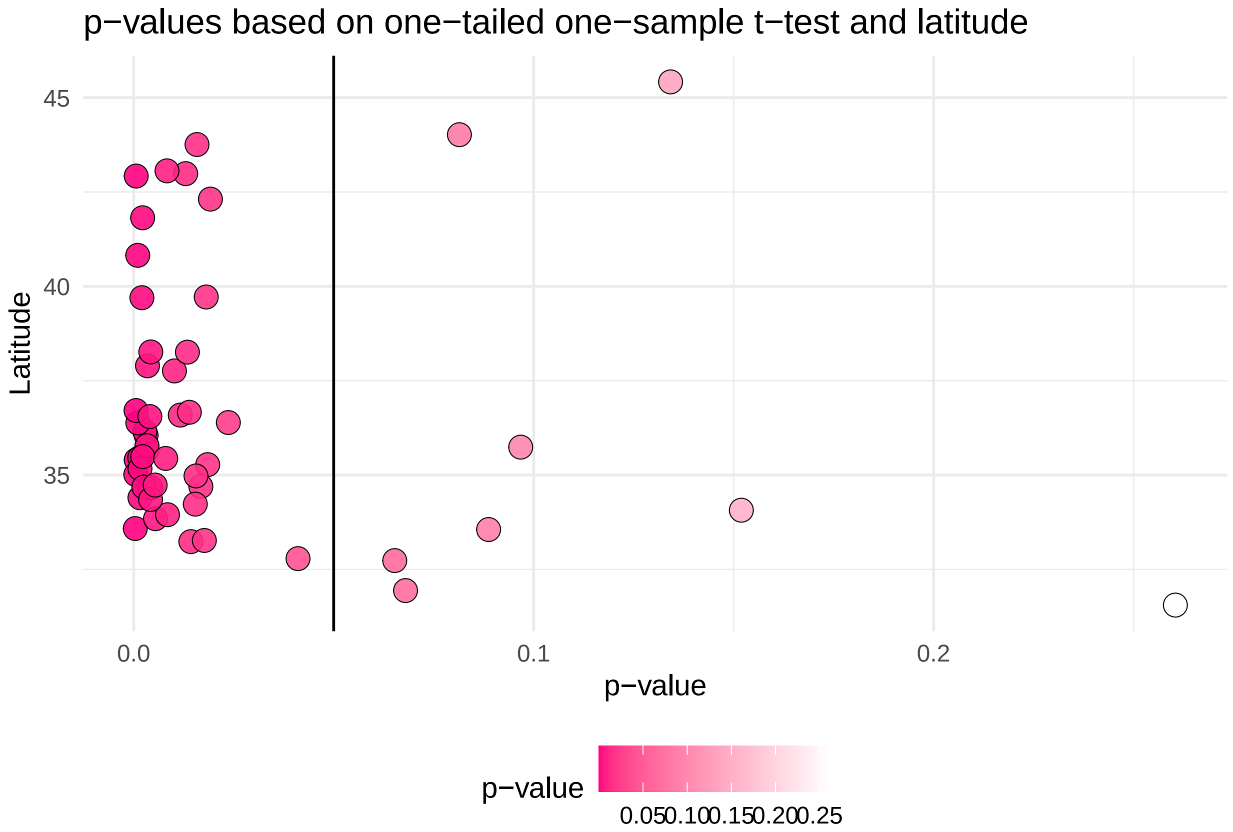
```
##
```

```
##  One Sample t-test
##
## data:  osaka_diff$diff
## t = 2.9757, df = 38, p-value = 0.002531
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  0.9349992       Inf
## sample estimates:
## mean of x
##  2.157265
```

Here, the p-value is below alpha level, so I reject the null hypothesis and accept the alternative hypothesis. For the 39 years of data available, the first-blooming dates are significantly earlier than the previous 30-year-average in Osaka.

Now, I'm going to expand this to all 53 locations in Japan.

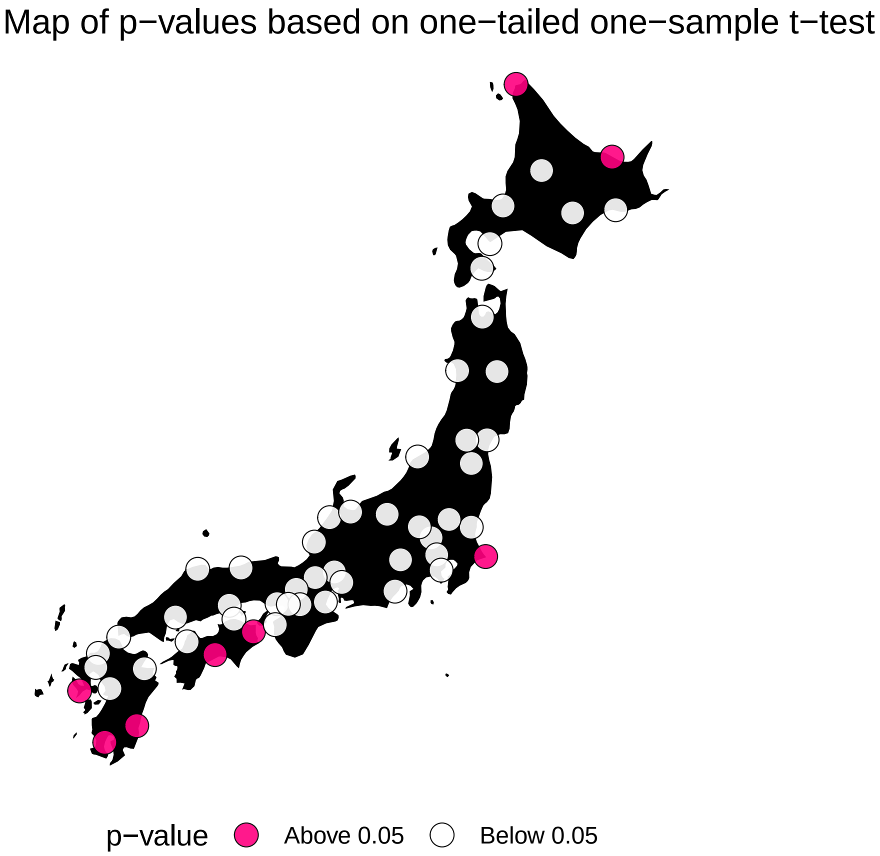
```r
sakura_30yr <- NULL
pVal <- NULL

for(i in 1:53) {
  sakura_30yr[[i]] <- sakura %>%
    filter(location == observation$location[i]) %>%
    mutate(doy_30yr = rollmean(doy_x, k = 30, fill = NA, align = "right")) %>%
    mutate(diff = doy_30yr - doy_x)
  pVal[i] <- t.test(x = sakura_30yr[[i]]$diff, mu = 0, alternative = "greater")$p.value
}

sakura_30yr_pVal <- observation %>%
  cbind(pVal) %>%
  mutate(is_above_alpha = ifelse(pVal > 0.05, "Above 0.05", "Below 0.05"))

sakura_30yr_pVal %>%
  ggplot() +
  geom_vline(xintercept = 0.05) +
  geom_point(aes(x = pVal, y = lat, fill = pVal),
             pch = 21, color = "black", size = 4, stroke = 0.3, alpha = 0.9) +
  scale_fill_gradient(low = "#ff0080", high = "white") +
  labs(title = "p-values based on one-tailed one-sample t-test and latitude",
       fill = "p-value", x = "p-value", y = "Latitude") +
  theme_minimal() +
  theme(legend.position = "bottom")
```



Based on the calculation and visualization, as a result of one-sample t-test, p-values are above alpha level in eight locations in Japan – in those locations, based on the years of data available, there is no significant difference between the actual first-blooming dates and the moving 30-year-average in recent 30 years. On the other hand, for all other locations, I reject the null hypothesis and accept the alternative hypothesis, concluding the actual first-blooming dates are earlier than the average date of 30 recent years for the location.

I'm now going to map the p-values to see geographical distribution about it.

```
ggplot() +
  geom_polygon(data = japan %>% filter(lat > 31),
               aes(x = long, y = lat, group = group),
               fill = "black") +
  geom_point(data = sakura_30yr_pVal,
             aes(x = long, y = lat, fill = is_above_alpha),
             pch = 21, color = "black", size = 4, stroke = 0.3, alpha = 0.9) +
  scale_fill_manual(values = c("#ff0080", "white")) +
  labs(title = "Map of p-values based on one-tailed one-sample t-test",
       fill = "p-value") +
  theme_void() +
  theme(legend.position = "bottom") +
  coord_map()
```



Based on the map, the locations with p-values above the alpha level of 0.05 are, in general, in either northern or southern end of Japan, except for an eastern city in Kanto area, Choshi, Chiba. It is reasonable to say that in most places, the first-blooming date is earlier than its recent 30-year average. There might be factors that are causing earlier first-blooming in recent decades, but those northern or southern locations might not get influenced as much as other locations.

Question 2: What is a geographical distribution of whether 400- and 600- theories hold true?

In the midterm project, I concluded 400- and 600- theories hold true in some locations but not others. Now, I'm going to conduct the same one-sample t-test for all 53 locations to analyze the geographical distribution about whether those theories are true.

- **400-degree theory**: The first-blooming happens when the cumulative daily **average** temperature since February 1st reaches 400 degrees.
- **600-degree theory**: Cherry blossom's first-blooming occurs when the cumulative daily **high** temperature since February 1st reaches 600 degrees.

In the test, I'm comparing the expected first-blooming date based on the theories and actual blooming date. The hypotheses are:

- **Null hypothesis**: the difference between the actual first-blooming date and the expected first-blooming date based on 400-degree/600-degree theory is 0 days

- **Alternative hypothesis**: the difference between the actual first-blooming date and the expected first-blooming date based on 400-degree/600-degree theory is not 0 days

```
file_paths <- dir_ls("data/daily-temp")
file_contents <- NULL
file_names <- NULL

# load daily temperature data about 53 locations
for(i in 1:53) {
  file_names[[i]] <- str_sub(file_paths[[i]], start = 17, end = -5)
  file_contents[[i]] <- read_csv(file_paths[[i]])
}
file_contents <- set_names(file_contents, file_names)

theory400_date <- NULL
theory600_date <- NULL

# calculate expected date
for(i in 1:53) {
  file_contents[[i]] <- file_contents[[i]] %>%
    mutate(date = ymd(date),
           avg_temp = as.numeric(avg_temp),
           high_temp = as.numeric(high_temp),
           year = year(date)) %>%
  group_by(year) %>%
  mutate(cum_high_temp = cumsum(high_temp),
         cum_avg_temp = cumsum(avg_temp))

  theory400_date[[i]] <- file_contents[[i]]
  theory600_date[[i]] <- file_contents[[i]]

  theory400_date[[i]] <- theory400_date[[i]] %>%
  filter(cum_avg_temp > 400) %>%
  filter(cum_avg_temp == min(cum_avg_temp)) %>%
  mutate(
    date_mu = date,
    floor_date = floor_date(date, unit = "year"),
    doy_mu = interval(floor_date, date_mu) %>% time_length(unit = "day")
  ) %>%
  select(date_mu, doy_mu, year)

  theory600_date[[i]] <- theory600_date[[i]] %>%
  filter(cum_high_temp > 600) %>%
  filter(cum_high_temp == min(cum_high_temp)) %>%
  mutate(
    date_mu = date,
    floor_date = floor_date(date, unit = "year"),
    doy_mu = interval(floor_date, date_mu) %>% time_length(unit = "day")
  ) %>%
  select(date_mu, doy_mu, year)
}

# calculate difference between the actual date and expected date
for(i in 1:53) {
  theory400_date[[i]] <- sakura %>%
    filter(location == file_names[[i]]) %>%
    merge(theory400_date[[i]]) %>%
    mutate(diff = doy_x - doy_mu)

  theory600_date[[i]] <- sakura %>%
    filter(location == file_names[[i]]) %>%
    merge(theory600_date[[i]]) %>%
    mutate(diff = doy_x - doy_mu)
}

# conduct t-test
theory400_pVal <- NULL
theory400 <- NULL
theory600_pVal <- NULL
theory600 <- NULL

for (i in 1:53) {
  theory400_pVal <- rbind(theory400_pVal,
                          t.test(x = theory400_date[[i]]$diff, mu = 0)$p.value)
  theory400[[i]] <- cbind(location = file_names[[i]], pVal = theory400_pVal[[i]])
  theory600_pVal <- rbind(theory600_pVal,
                          t.test(x = theory600_date[[i]]$diff, mu = 0)$p.value)
  theory600[[i]] <- cbind(location = file_names[[i]], pVal = theory600_pVal[[i]])
}

head(theory400)
```

```
## [[1]]
##      location   pVal
## [1,] "Abashiri" "3.69192190994362e-45"
##
## [[2]]
##      location pVal
## [1,] "Akita"  "6.7183234573113e-43"
##
## [[3]]
##      location pVal
## [1,] "Aomori" "8.2111476877462e-46"
##
## [[4]]
##      location    pVal
## [1,] "Asahikawa" "3.88168360998572e-55"
##
## [[5]]
##      location pVal
## [1,] "Choshi" "1.87562188352897e-16"
##
## [[6]]
##      location pVal
## [1,] "Fukui"  "6.35736213780031e-32"
```

```
head(theory600)
```

```
## [[1]]
##      location   pVal
## [1,] "Abashiri" "2.18740972963854e-30"
##
## [[2]]
##      location pVal
## [1,] "Akita"  "2.06272723561901e-17"
##
## [[3]]
##      location pVal
## [1,] "Aomori" "2.09134341570129e-13"
##
## [[4]]
##      location    pVal
## [1,] "Asahikawa" "1.35793500841999e-30"
##
## [[5]]
##      location pVal
## [1,] "Choshi" "2.67244573698321e-12"
##
## [[6]]
##      location pVal
## [1,] "Fukui"  "0.00478752257743315"
```

Based on the calculation, for the analysis about 400-degree theory, p-values are above alpha level in only four locations (Takamatsu, Tokyo, Tsu, Yokohama), while p-values associated with the t-test about 600-degree theory are above alpha in only three locations (Kanazawa, Sendai, Toyama).

For tests that produced these higher p-values, I accept the null hypothesis and conclude that there is no significant difference between expected first-blooming date based on 400- or 600-degree theory, meaning the theory holds true in those locations. On the other hand, for all other locations, I reject the null hypothesis and accept the alternative hypothesis: there is a significant difference between expected date and actual date, thus theories won't be true there.

Next, I'm going to merge these p-values with location data.

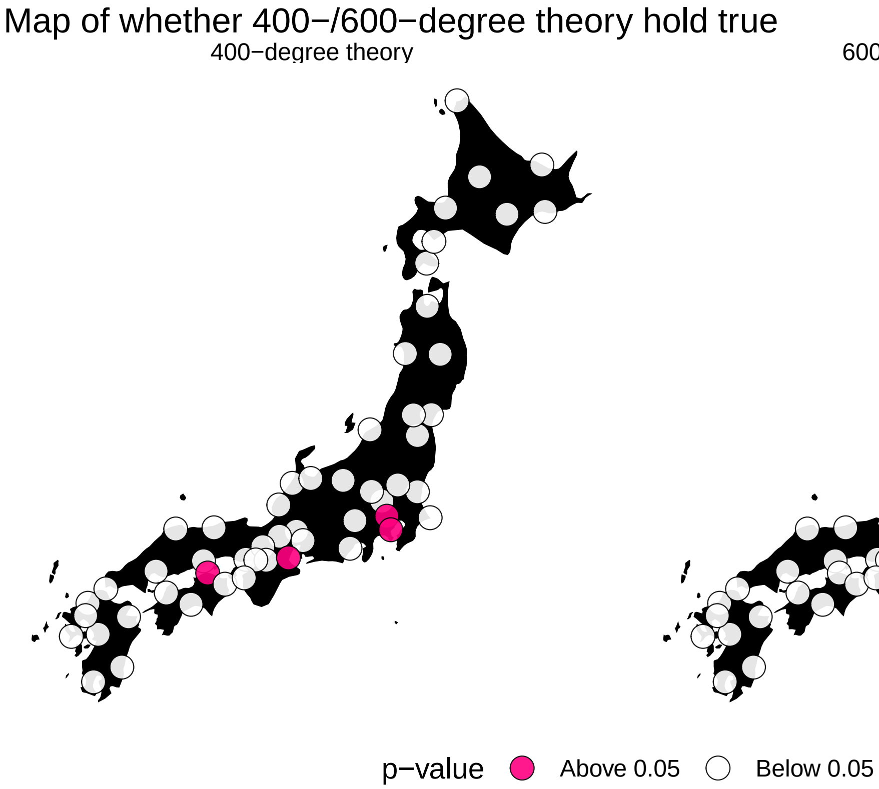
```
# merge with location data
theory400_location <- as.data.frame(do.call(rbind, theory400)) %>%
  mutate(pVal = as.numeric(format(pVal, scientific = FALSE))) %>%
  mutate(is_above_alpha = ifelse(pVal >= 0.05, "Above 0.05", "Below 0.05"))
theory400_merged <- cbind(theory400_location, observation) %>%
  select(-4) %>%
  mutate(theory = "400-degree theory")

theory600_location <- as.data.frame(do.call(rbind, theory600)) %>%
  mutate(pVal = as.numeric(format(pVal, scientific = FALSE))) %>%
  mutate(is_above_alpha = ifelse(pVal >= 0.05, "Above 0.05", "Below 0.05"))
theory600_merged <- cbind(theory600_location, observation) %>%
  select(-4) %>%
  mutate(theory = "600-degree theory")
```

```
theory400_600_merged <- rbind(theory400_merged, theory600_merged)
head(theory400_600_merged)
```

```
##    location         pVal is_above_alpha      lat     long            theory
## 1  Abashiri 3.691922e-45     Below 0.05 44.01885 144.2809 400-degree theory
## 2     Akita 6.718323e-43     Below 0.05 39.71794 140.1000 400-degree theory
## 3    Aomori 8.211148e-46     Below 0.05 40.82214 140.7685 400-degree theory
## 4 Asahikawa 3.881684e-55     Below 0.05 43.75783 142.3731 400-degree theory
## 5    Choshi 1.875622e-16     Below 0.05 35.73978 140.8583 400-degree theory
## 6     Fukui 6.357362e-32     Below 0.05 36.05586 136.2231 400-degree theory
```

Now, I’m going to geographically map the values.

```
ggplot() +
  geom_polygon(data = japan %>% filter(lat > 31),
               aes(x = long, y = lat, group = group),
               fill = "black") +
  geom_point(data = theory400_600_merged,
             aes(x = long, y = lat, fill = as.character(is_above_alpha)),
             pch = 21, color = "black", size = 4, stroke = 0.3, alpha = 0.9) +
  scale_fill_manual(values = c("#ff0080", "white")) +
  theme_void() +
  theme(legend.position = "bottom") +
  labs(title = "Map of whether 400-/600-degree theory hold true",
       fill = "p-value") +
  facet_wrap(vars(theory)) +
  coord_map()
```

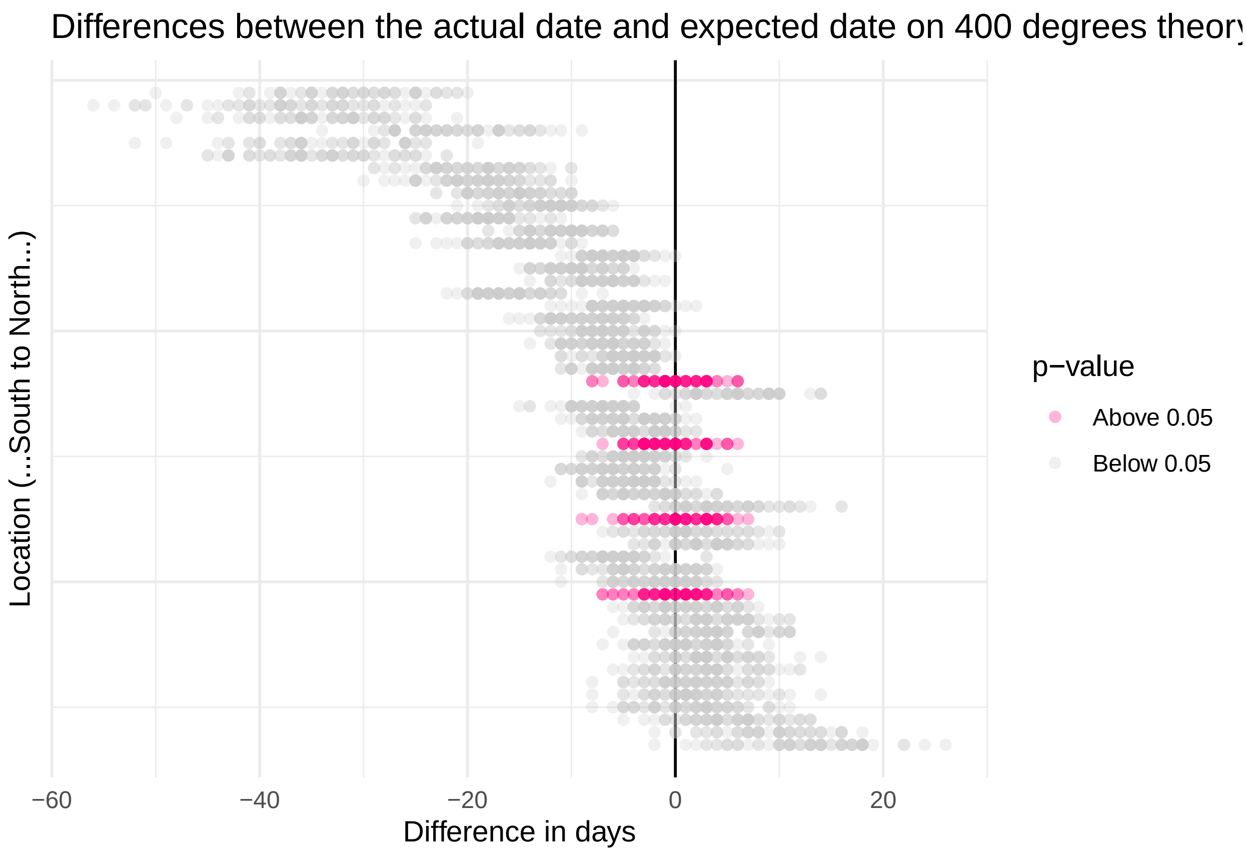


Based on the analysis, there are quite limited number of locations where those theories actually hold true. Additionally, except for the fact that places with higher p-values are close to each other in terms of latitude, there are no other specific patterns suggested from this visualization about what geographical characteristics are related to whether those theories are true.

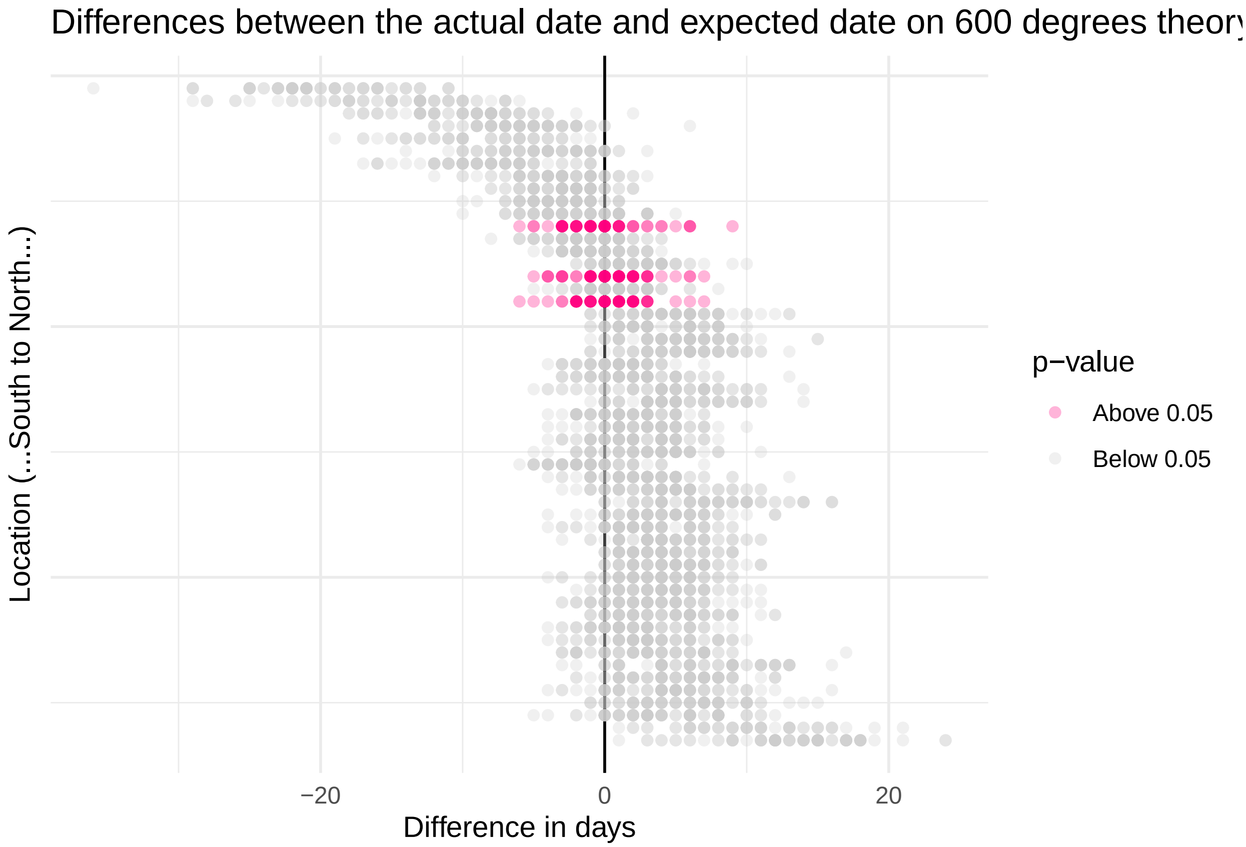
However, since the 400-degree theory holds true in the capital city of Tokyo and second biggest city of Yokohama, it is understandable why people value this theory.

Next, I'm going to visualize the distribution of all difference in days in all locations.

```
location_order <- c("Wakkanai", "Abashiri", "Asahikawa", "Sapporo", "Kushiro", "Obihiro", "Muroran", "Ha

theory400_to_plot <- do.call(rbind.data.frame, theory400_date) %>%
  mutate(is_above_alpha = ifelse(location %in% c("Takamatsu", "Tokyo", "Tsu", "Yokohama"), "Above 0.05"
  mutate(location = fct_relevel(location, location_order))

theory400_to_plot %>%
  ggplot(aes(x = diff, y = desc(location)), size = 0.1) +
  geom_vline(xintercept = 0) +
  geom_point(aes(color = as.character(is_above_alpha)), alpha = 0.3) +
  scale_color_manual(values = c("#FF0080", "#cccccc"))  +
    labs(title = "Differences between the actual date and expected date on 400 degrees theory",
      x = "Difference in days", y = "Location (←South to North→)",
      color = "p-value") +
  theme_minimal() +
  theme(axis.text.y = element_blank())
```



```
theory600_to_plot <- do.call(rbind.data.frame, theory600_date) %>%
  mutate(is_above_alpha = ifelse(location %in% c("Kanazawa", "Sendai", "Toyama"), "Above 0.05", "Below
  mutate(location = fct_relevel(location, location_order))

theory600_to_plot %>%
  ggplot(aes(x = diff, y = desc(location)), size = 0.1) +
  geom_vline(xintercept = 0) +
  geom_point(aes(color = as.character(is_above_alpha)), alpha = 0.3) +
  scale_color_manual(values = c("#FF0080", "#cccccc")) +
    labs(title = "Differences between the actual date and expected date on 600 degrees theory",
       x = "Difference in days", y = "Location (←South to North→)",
       color = "p-value") +
  theme_minimal() +
  theme(axis.text.y = element_blank())
```



Each horizontal line represents distribution of difference between expected date (backed by theories) and actual first-blooming dates for one location. Vertically, locations are arranged by latitude, from north (upper side on the chart) to south (lower side on the chart). The pink-highlighted lines are the distributions of difference in days where p-values are above alpha level (where those theories are true).

By comparing these two visualizations, the distribution about 400-degree theory shows a wider range. Especially in northern locations, the difference is generally about -30 to -40 days – the cumulative daily temperature reaches more than a month after cherry blossoms bloom for the first time each year.

Additionally, especially from the visualization regarding 600-degree theory, there is a curved pattern in the relationship between latitude and difference in days. In many locations below (south of) three pinked lines in this visualization, the distributions are similar across locations – generally, the difference in days seems to be about a few days after the expected date.

Inversely calculate "real" first-blooming date based on theories

For the locations where those theories won't hold true, these visualizations suggest different "real" expected date, based on those theories. Next, I'm going to inversely calculate the expected first-blooming dates, based on those theories.

I'm using qt() function to calculate what t-score would result in a probability of at least half the alpha level (0.025). Based on the range of t-score, I calculate the range of "mu" values which would generate p=value of at least 0.05 in the t-test.

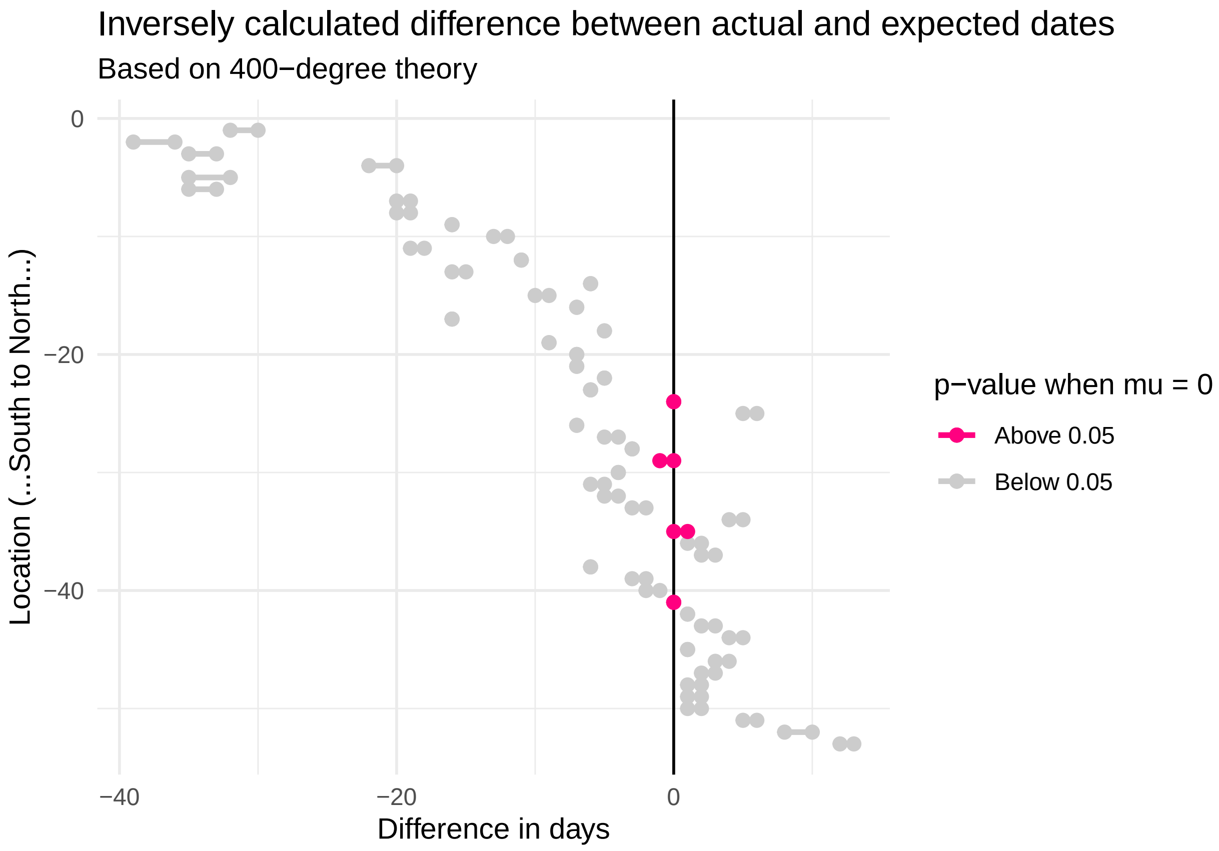
```
t400 <- NULL
t600 <- NULL
inversed_mu400 <- NULL
```

```
inversed_mu600 <- NULL

# calculate range of t-score and convert them to integer range in difference in days
# 400-degree theory
for (i in 1:53) {
  t400[[i]] <- qt(p = 0.025, df = nrow(theory400_date[[i]]) - 1)
  mean <- mean(theory400_date[[i]]$diff, na.rm = TRUE)
  sd <- sd(theory400_date[[i]]$diff, na.rm = TRUE)
  size <- nrow(theory400_date[[i]])
  se <- sd / sqrt(size)
  mu_max <- -t400[[i]] * se + mean
  mu_min <- t400[[i]] * se + mean
  mu_max_floor <- floor(mu_max)
  mu_min_ceiling <- ceiling(mu_min)
  inversed_mu400[[i]] <- list(min = mu_min_ceiling, max = mu_max_floor)
}

# 600-degree theory
for (i in 1:53) {
  t600[[i]] <- qt(p = 0.025, df = nrow(theory600_date[[i]]) - 1)
  mean <- mean(theory600_date[[i]]$diff, na.rm = TRUE)
  sd <- sd(theory600_date[[i]]$diff, na.rm = TRUE)
  size <- nrow(theory600_date[[i]])
  se <- sd / sqrt(size)
  mu_max <- -t600[[i]] * se + mean
  mu_min <- t600[[i]] * se + mean
  mu_max_floor <- floor(mu_max)
  mu_min_ceiling <- ceiling(mu_min)
  inversed_mu600[[i]] <- list(min = mu_min_ceiling, max = mu_max_floor)
}

inversed_mu400 <- as.data.frame(do.call(rbind, inversed_mu400))
inversed_mu600 <- as.data.frame(do.call(rbind, inversed_mu600))

inversed_mu400_merged <- cbind(observation, inversed_mu400) %>%
  mutate(max = as.numeric(max),
         min = as.numeric(min))
inversed_mu600_merged <- cbind(observation, inversed_mu600) %>%
  mutate(max = as.numeric(max),
         min = as.numeric(min))
```

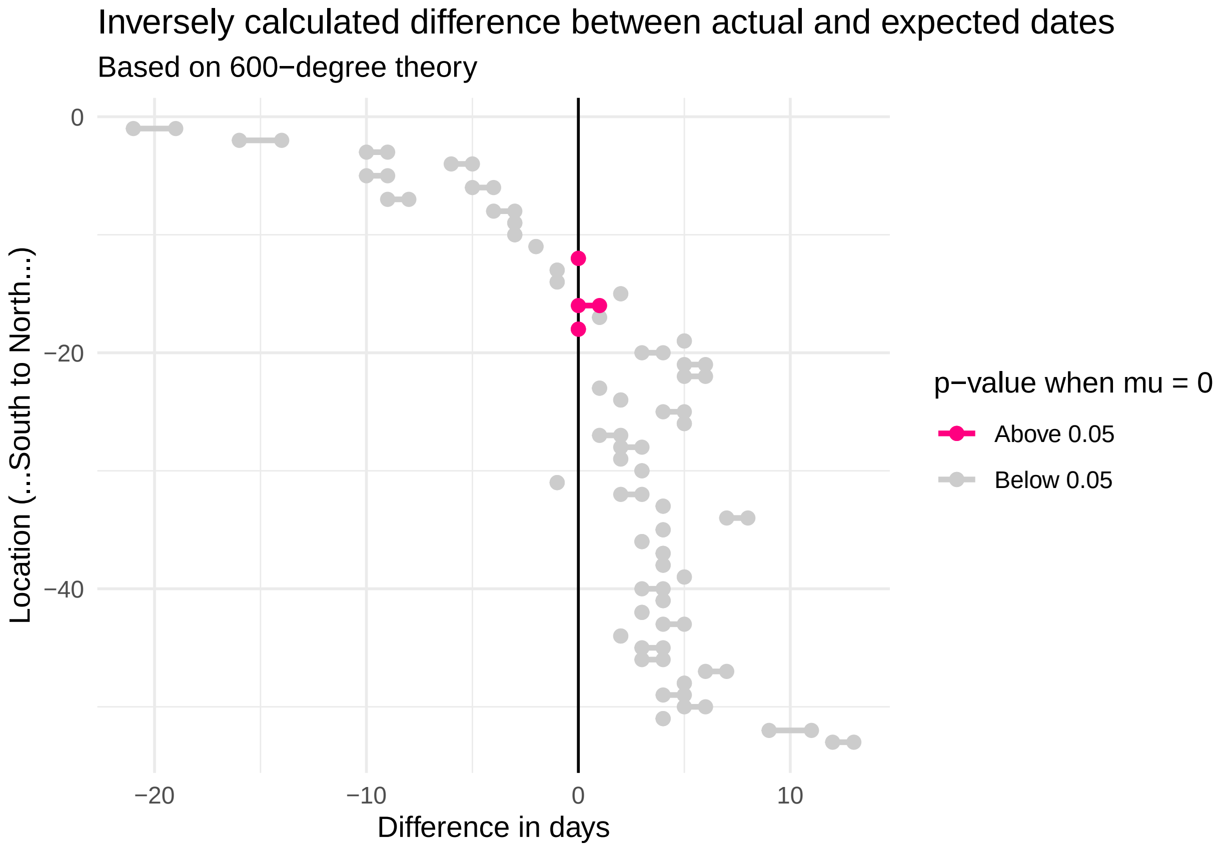
In all locations, there is a range in difference in days where people can expect cherry blossoms' first-blooming before or after the expected based on cumulative daily temperature.

Next, I'm going to visualize these ranges in the same format as the whole distribution visualization.

```
inversed_mu400_merged %>%
    mutate(is_above_alpha = ifelse(location %in% c("Takamatsu", "Tokyo", "Tsu", "Yokohama"), "Above 0.05
  mutate(location = fct_relevel(location, location_order)) %>%
  ggplot() +
  geom_vline(xintercept = 0) +
  geom_segment(aes(x = min, xend = max, y = desc(location), yend = desc(location), color = is_above_alpl
  geom_point(aes(x = min, y = desc(location), color = is_above_alpha), size = 2) +
  geom_point(aes(x = max, y = desc(location), color = is_above_alpha), size = 2) +
  scale_color_manual(values = c("#FF0080", "#cccccc")) +
  labs(title = "Inversely calculated difference between actual and expected dates",
       subtitle = "Based on 400-degree theory",
       x = "Difference in days", y = "Location (←South to North→)",
       color = "p-value when mu = 0") +
  theme_minimal()
```



```
inversed_mu600_merged %>%
    mutate(is_above_alpha = ifelse(location %in% c("Kanazawa", "Sendai", "Toyama"), "Above 0.05", "Belo
  mutate(location = fct_relevel(location, location_order)) %>%
  ggplot() +
  geom_vline(xintercept = 0) +
  geom_segment(aes(x = min, xend = max, y = desc(location), yend = desc(location), color = is_above_alp
  geom_point(aes(x = min, y = desc(location), color = is_above_alpha), size = 2) +
  geom_point(aes(x = max, y = desc(location), color = is_above_alpha), size = 2) +
  scale_color_manual(values = c("#FF0080", "#cccccc")) +
  labs(title = "Inversely calculated difference between actual and expected dates",
       subtitle = "Based on 600-degree theory",
       x = "Difference in days", y = "Location (←South to North→)",
       color = "p-value when mu = 0") +
  theme_minimal()
```



Same with the previous visualizations, vertically, locations are arranged by latitude, from north (upper side on the chart) to south (lower side on the chart). Horizontal axis represents difference between the expected and actual first-blooming dates. The pink-highlighted points (and lines) represent locations where these theories hold true – p-values are above alpha level when we assume the difference between expected and actual dates is 0 days.

By comparing the visualizations, as we expected above, there is smaller variation in difference in days based on 600-degree theory, than that based on 400-degree theory.

But it is exciting to know that people can expect first-blooming based on cumulative daily high/average temperatures. For example, here is the calculation about Osaka, where neither theories hold true.

```
# show data about Osaka
inversed_mu400_merged %>%
  filter(location == "Osaka")
```

```
##   location      lat     long min max
## 1    Osaka 34.68196 135.5187   2   3
```

```
inversed_mu600_merged %>%
  filter(location == "Osaka")
```

```
##   location      lat     long min max
## 1    Osaka 34.68196 135.5187   4   4
```

In Osaka, people can expect first-blooming 2 to 3 days after cumulative daily high temperature reaches 400 degrees, or 4 days after cumulative daily average temperature reaches 600 degrees.

Question 3: Correlation between latitude and difference in days

Here, I'm going to test if there is a correlation between latitude and average difference in days based on the theories. For this analysis, I'm going to focus on distribution regarding 600-degree theories.

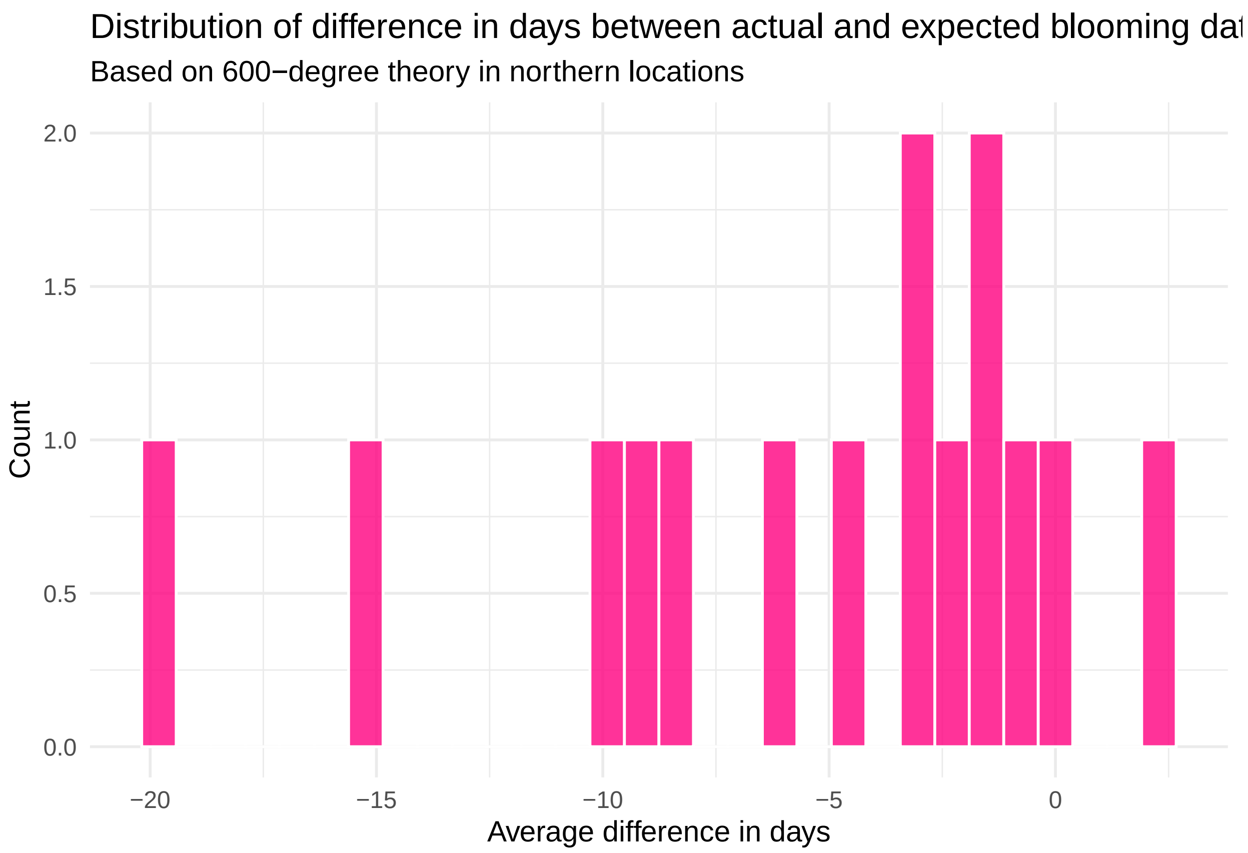
The whole distribution about differences between the actual date and expected date suggests curbed line. This makes sense because in northern two thirds of locations, latitude changes drastically, whereas in other locations, latitudes stay similar, while longitudes change instead.

Thus, I'm going to subgroup the distributions based on latitude for this analysis. I decided to group locations above "Fukushima" as northern locations, and others as western locations.

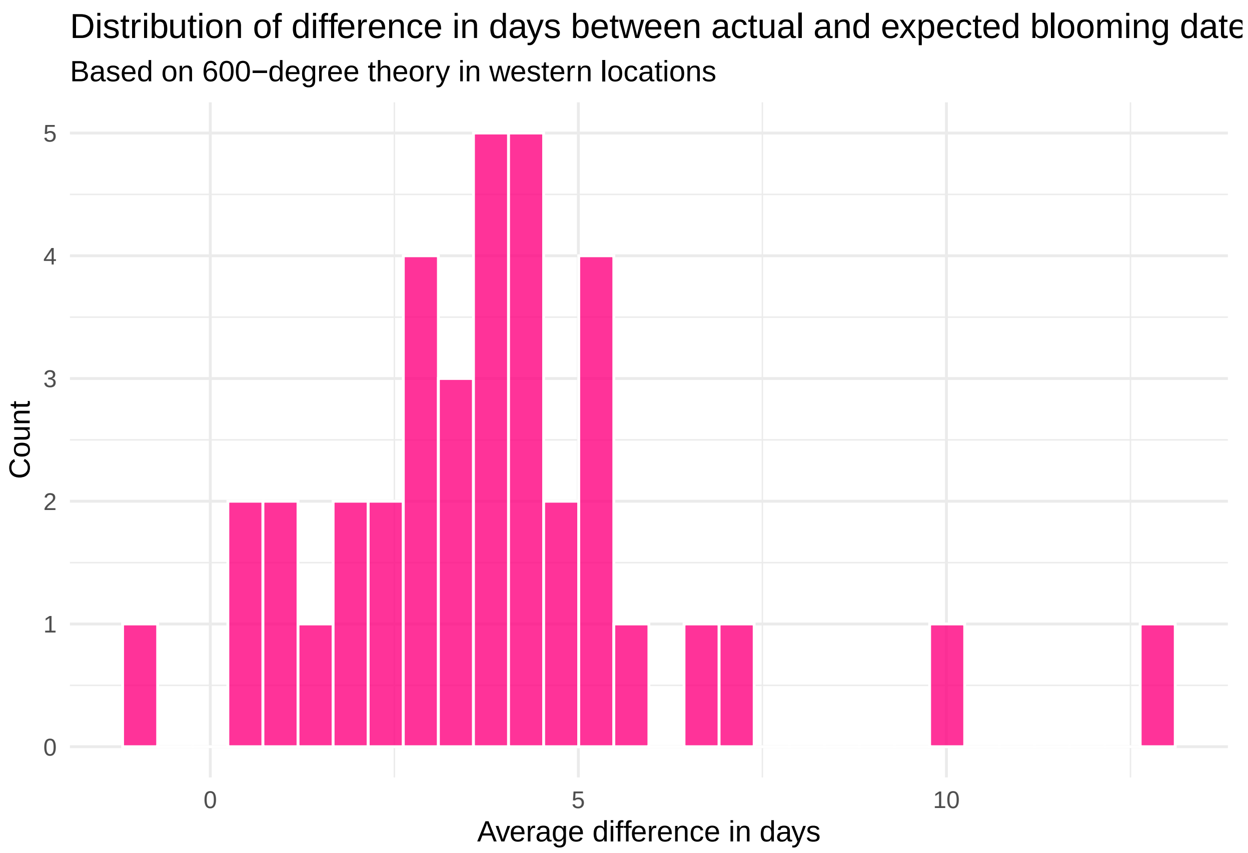
```r
# subgroup by latitude (location)
north <- c("Wakkanai", "Abashiri", "Asahikawa", "Sapporo", "Kushiro", "Obihiro", "Muroran", "Hakodate",

west <- c("Toyama", "Nagano", "Kanazawa", "Utsunomiya", "Maebashi", "Mito", "Kumagaya", "Fukui", "Tokyo

theory600_north <- do.call(rbind.data.frame, theory600_date) %>%
  filter(location %in% north)

theory600_west <- do.call(rbind.data.frame, theory600_date) %>%
  filter(location %in% west)
```

Next, I'm going to calculate the average difference in days between actual and expected dates for each group, and visualize the distribution. I'll then test for normality.

```r
# calculate average difference in days
theory600_north_ct <- theory600_north %>%
  group_by(location) %>%
  summarize(mean = mean(diff, na.rm = TRUE))

theory600_west_ct <- theory600_west %>%
  group_by(location) %>%
  summarize(mean = mean(diff, na.rm = TRUE))

# draw histogram
theory600_north_ct %>%
  ggplot(aes(x = mean)) +
  geom_histogram(color = "white", fill = "#ff0080", alpha = 0.8) +
  labs(title = "Distribution of difference in days between actual and expected blooming date",
       subtitle = "Based on 600-degree theory in northern locations",
       x = "Average difference in days", y = "Count") +
  theme_minimal()
```



```
theory600_west_ct %>%
  ggplot(aes(x = mean)) +
  geom_histogram(color = "white", fill = "#ff0080", alpha = 0.8) +
    labs(title = "Distribution of difference in days between actual and expected blooming date",
       subtitle = "Based on 600-degree theory in western locations",
       x = "Average difference in days", y = "Count") +
  theme_minimal()
```



Based on the visualizations, the distribution in northern locations doesn't have much significant shape. There is a slight hump slightly below 0 days of difference, but it seems that there are not enough observations available to judge if the distribution is normal or not, yet.

As for the distribution in western locations, the visualization suggests unimodality and positive skew.

Next, I'm going to calculate skewness and kurtosis for each distribution.

```
# calculate skewness and kurtosis of average difference in days for northern locations
skewness(theory600_north_ct$mean) # -1.052358
```

```
## [1] -1.052358
```

```
kurtosis(theory600_north_ct$mean) # 3.381491
```

```
## [1] 3.381491
```

```
# calculate skewness and kurtosis of average difference in days for western locations
skewness(theory600_west_ct$mean) # 1.162122
```

```
## [1] 1.162122
```

```
kurtosis(theory600_west_ct$mean) # 5.866828
```

```
## [1] 5.866828
```

Based on these calculations, as for northern locations, the distribution is slightly negatively skewed but not too far from zero, and kurtosis is grater than 3 (leptokurtic), but not too far from three.

On the other hand, as for western locations, the distribution is slightly positively skewed as I assumed from the shape. The kurtosis is much higher than 3 and the distribution is heavily leptokurtic.

Finally, I'm going to conduct shapiro test to test for normality for both visualizations. The hypotheses are: * **Null hypothesis**: the distribution of number of dates is normally distributed * **Alternative hypothesis**: the distribution of number of dates is not normally distributed

```
shapiro.test(theory600_north_ct$mean) # p-value = 0.1042
```

```
##
##  Shapiro-Wilk normality test
##
## data:  theory600_north_ct$mean
## W = 0.90257, p-value = 0.1042
```

```
shapiro.test(theory600_west_ct$mean) # p-value = 0.006334
```

```
##
##  Shapiro-Wilk normality test
##
## data:  theory600_west_ct$mean
## W = 0.91365, p-value = 0.006334
```

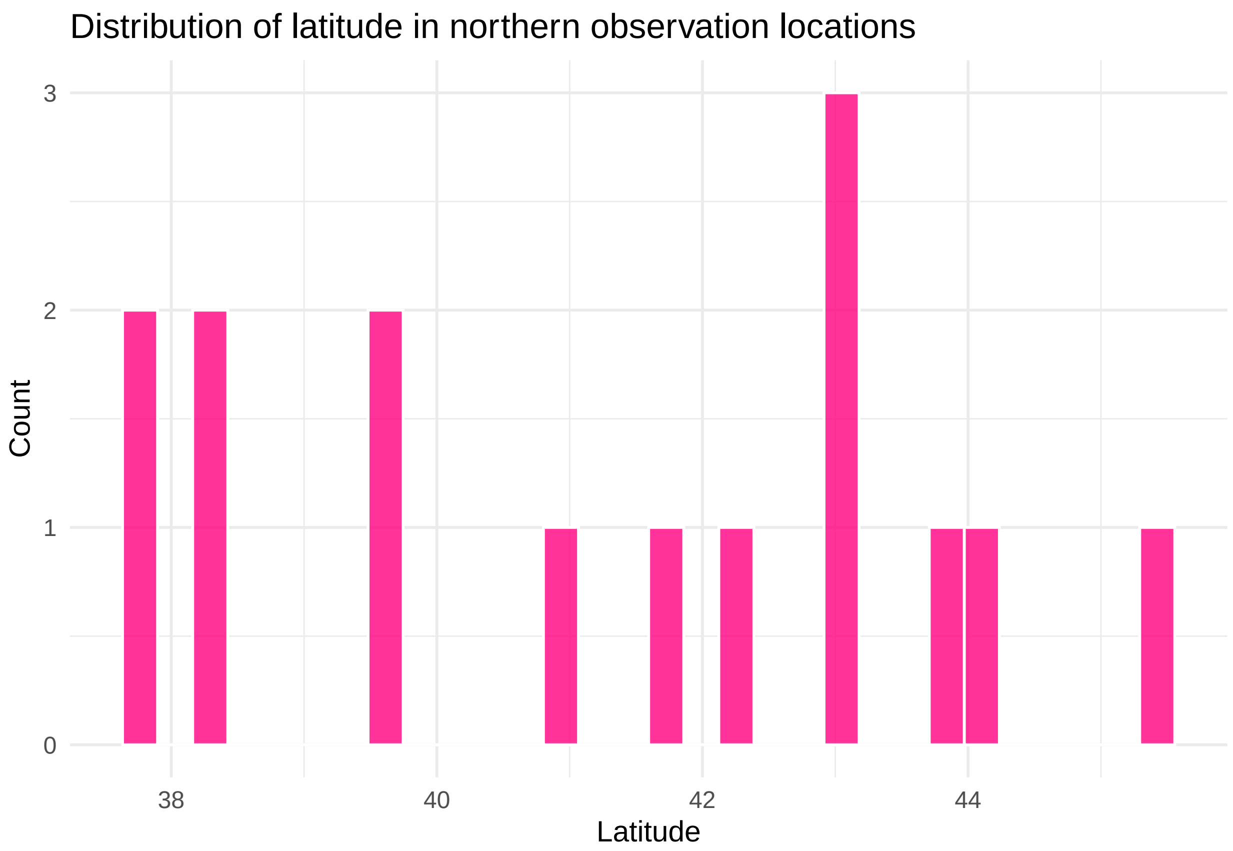
As a result of the test, for northern locations, p-value is above alpha level of 0.05. Thus I accept the null hypothesis and conclude that the distribution is normal.

As for western locations, p-value is below alpha level. I reject the null hypothesis and accept the alternative hypothesis – the distribution is not normal.

From now on, I'm only going to focus on northern locations.

Next, I'm going to test for normality of distribution of latitude in those northern locations.

```
observation_north <- observation %>%
  filter(location %in% c("Wakkanai", "Abashiri", "Asahikawa", "Sapporo", "Kushiro", "Obihiro", "Muroran

observation_north %>%
  ggplot(aes(x = lat)) +
  geom_histogram(color = "white", fill = "#ff0080", alpha = 0.8) +
  labs(title = "Distribution of latitude in northern observation locations",
       x = "Latitude", y = "Count") +
  theme_minimal()
```



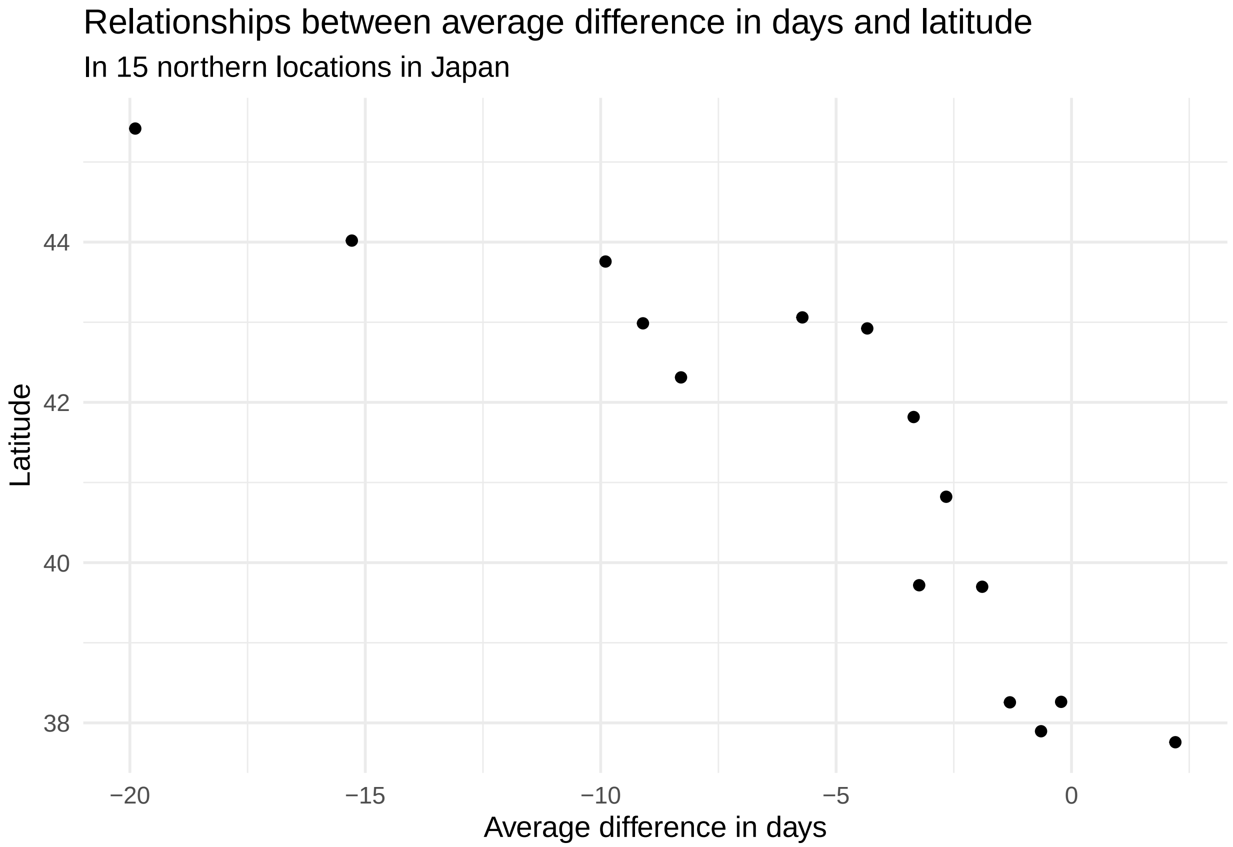
```
shapiro.test(observation_north$lat) # p-value = 0.2384
```

```
##
##  Shapiro-Wilk normality test
##
## data:  observation_north$lat
## W = 0.9261, p-value = 0.2384
```

Even though there are limited number of scores, the result of the test shows this is normally distributed.

Now, I'm going to draw a scatterplot to look for a linear relationship.

```
north600 <- observation_north %>%
  cbind(theory600_north_ct) %>%
  select(-4)

north600 %>%
  ggplot(aes(x = mean, y = lat)) +
  geom_point() +
  labs(title = "Relationships between average difference in days and latitude",
       subtitle = "In 15 northern locations in Japan",
       x = "Average difference in days", y = "Latitude") +
  theme_minimal()
```



This suggests slightly curved relationship, rather than linear. Even though this violates one of the assumptions behind correlation analysis, I'm going to conduct test to determine the average difference in days and latitude have significant correlation.

Since both variables are continuous variables, I'm using Pearson correlation coefficient. They are also both normally distributed, as well as they are paired.

The hypotheses are: * **Null hypothesis**: the correlation coefficient between the average difference in days and latitude is 0 – there is no significant correlation * **Alternative hypothesis**: the correlation coefficient between the average difference in days and latitude is not equal to 0 – there is a significant correlation

```
cor.test(x = north600$mean, y = north600$lat) # p-value = 1.188e-05, cor = -0.8846256
```

```
##
##  Pearson's product-moment correlation
##
## data:  north600$mean and north600$lat
## t = -6.8401, df = 13, p-value = 1.188e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9612759 -0.6809365
## sample estimates:
##        cor
## -0.8846256
```

Based on the result, the p-value is extremely small, thus I reject the null hypothesis and accept the alternative hypothesis. If we can assume there is a linear relationship between two variables, the test shows there is a significant correlation between these variables. Based on the scatterplot and the “cor” value, that correlation is negative. Moreover, based on the absolute number of the “cor” value, that negative correlation is very strong.

Summary

Findings from the analysis

- Based on temporal analysis, average first-blooming dates are getting earlier over decades, and the difference is statistically significant in several decades. Especially during 1990s and 2000s, the average first-blooming dates changed drastically and became much earlier.
- Whether the cherry blossoms are blooming on before April 1st looks pretty different now from decades ago.
- In many locations, first-blooming dates are significantly earlier than recent 30-year-moving average over decades.
- 400-degree theory only holds true in four locations, whereas 600-degree theory only holds true in three locations. This suggests that cherry blossom’s first blooming dates are affected not only by daily average/high temperatures, but other factors.
- Based on 400- and 600-degree theories, the expected first-blooming date can be calculated as how many days before or after the expected date that theories claim.
- As long as we can assume that there is a linear relationship between average difference in days based on 600-degree theories in 15 northern locations in Japan, there is a very strong negative correlation between these variables.

What’s next

- Since we now know the cherry blossoms are blooming earlier over decades on average, I’d like to explore if this trend is related to other factors, such as climate change as in temperature.
- As the test results about 400- and 600-degree theories suggest other geographical factors (than daily temperatures) to determine cherry blossoms’ first-blooming dates, I’d like to explore if possible other factors (such as wind speed and precipitation) influence the first-blooming dates.
- My correlation analysis is incomplete based on apparent lack of linear relationship between variables. It seems very complicated to expect cherry blossom’s fist-blooming dates. I’d like to look into models that are used by agencies like JMA, and study how the models are established and investigate if anything could be improved.

Finally, I’d like to give a special thanks to Steven Braun for all valuable feedbacks!