# Difference-In-Difference on AirBnB Listing Price - 2020 Tokyo Olympics Spike

Roxanne Chui

02/03/2020

**Abstract**

The paper looked at the change in the average AirBnB listing price during the Summer Olympics in Tokyo. There is an overall significant difference in the average listing price between July to August in 2019 against 2020 which the average price spiked up by YEN 32096 during the Olympic season. 2020 listing price in Tokyo was also measured against 2020 Paris, where the 2024 Olympics would be held in the City of Paris. Although the paper is for practicing statistical analysis on Difference-inDifference, the analysis is an insight for summer travellers and Olympics fans in terms of selecting accommodation and adjusting travel budget during the season.

## Introduction - 2020 Tokyo Olympics

2020 Tokyo Olympics starts on July 22, 2020, with the opening ceremony on July 24, 2020, and ends on August 9, 2020, and Japan is hoping to welcome millions of visitors and guests attending the event live. Typical Tokyo accommodations for tourists and guests are hotels, ryokan (traditional Japanese inn), guesthouses, capsule hotels, as well as AirBnB. At the same time, summer is often a travelling time for students and teachers, as the school term just ended and both parties have an extended holiday. A budget traveller would preferably choose AirBnB as a cheaper accommodation when travelling to Japan during the summer season, even though the author is not an Olympic fan. Which lead to the purpose of this paper, to consider the effect of the 2020 Tokyo Olympics on Tokyo's AirBnB Pricing.

Inside AirBnB reports and scrapes AirBnB listings data, including the information on price adjustments on each calender days made by the host. AirBnb calender data not only displays a listing's regular price but the adjusted price determined by the host on future calender days, such that the price listed on AirBnB today may be different in the future. Adjusted price is a variable that would be explored for this paper.

## Methodology and Results

We would first measure the average listing price by day in 2020, where 2020 Olympics would be happening in Tokyo, and compare it to the average listing price in 2019, where Olympics

did not happen in Tokyo. We then measure Tokyo's 2020 AirBnB listing price again and measure Paris's 2020 AirBnB listing price. That difference in the differences would then be our estimate of the effect of Olympics to AirBnB prices at the hosting country. The treatment is the presence of Olympics.

### Measure of Central Tendency for Tokyo AirBnB Listings

*Table 1. Central Tendency for AirBnB Lisitngs in 2019 and 2020*

| Year | # Unique Listings | Mean | SD | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|---|---|
| 2019 | 9768 | 18814 | 59713 | 350 | 7000 | 12000 | 19000 | 1199999 |
| 2020 | 15551 | 29123 | 94102 | 146 | 7400 | 12000 | 22000 | 3500000 |

According to Inside AirBnB as of Feb 2020, Tokyo has more than 15,000 AirBnB active listings in contrast to 9700 listings in 2019. There was a 59% increase in the number of listings and the average listing price between 2019 and 2020 also increase by 55%.

### Tokyo 2019 & 2020 Pricing Data Visualization

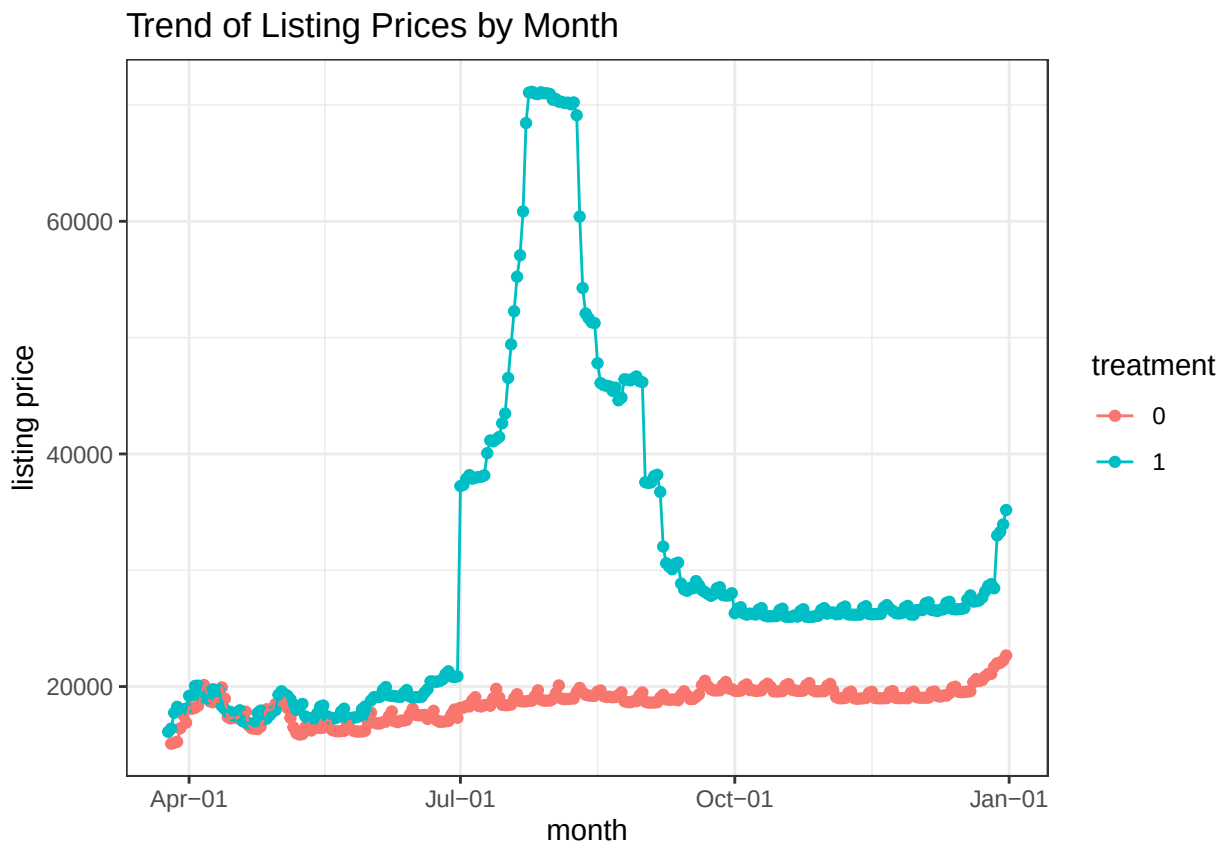


*Figure 1. Difference between Listing Price between 2019 (Control) and 2020 (Treatment)*

The dataset reported price adjustment by host starting March 25, 2019, hence the trends would start on day 85. As seen from Figure 1 and Table 1, AirBnB listing price has a large spike from July to August and dropping down in September. When focused on the price differences of June 30, 2020, and July 01, 2020, the difference in median price is YEN 4000 (15000 - 11000) and the difference in average price is YEN 16361.31 (37238.19 - 20876.88). The average listing price started to drop on August 16, 2020, and remained slightly fluctuating around YEN 30000 by mid-September 2020.

*Table 2. Logistics Regression on Price by Day of the Year in 2019*

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 18196. | 125. | 146. | 0 |
| day_of_year | 3.38 | 0.592 | 5.70 | 2.43e-08 |

*Table 3. Logistics Regression on Price by Day of the Year in 2020*

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 24338. | 1344. | 18.1 | 8.13e-53 |
| day_of_year | 25.9 | 0.592 | 4.06 | 5.90e- 5 |

*Table 4. Estimate for Listing Price by Treatment (Olmypic Year) and Day of the Year*

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 18196. | 963. | 18.9 | 4.32e-65 |
| day_of_year | 3.38 | 4.56 | 0.740 | 4.60e-01 |
| treatmentTreatment | 6142. | 1357. | 4.53 | 7.01e-06 |
| day_of_year:treatmentTreatment | 22.5 | 6.43 | 3.50 | 4.97e-04 |

## Data Analysis - Difference-In-Difference

### Difference between July 01 and August 31 of 2019 and 2020

The estimate of the logistic regression from the previous tables only reflects the general differences in the year with and without the Olympic Games. Based on Figure 1, we further focus on the difference-in-difference of the two months when the summer Olympics happened around, July 01 and August 31, of a different year.

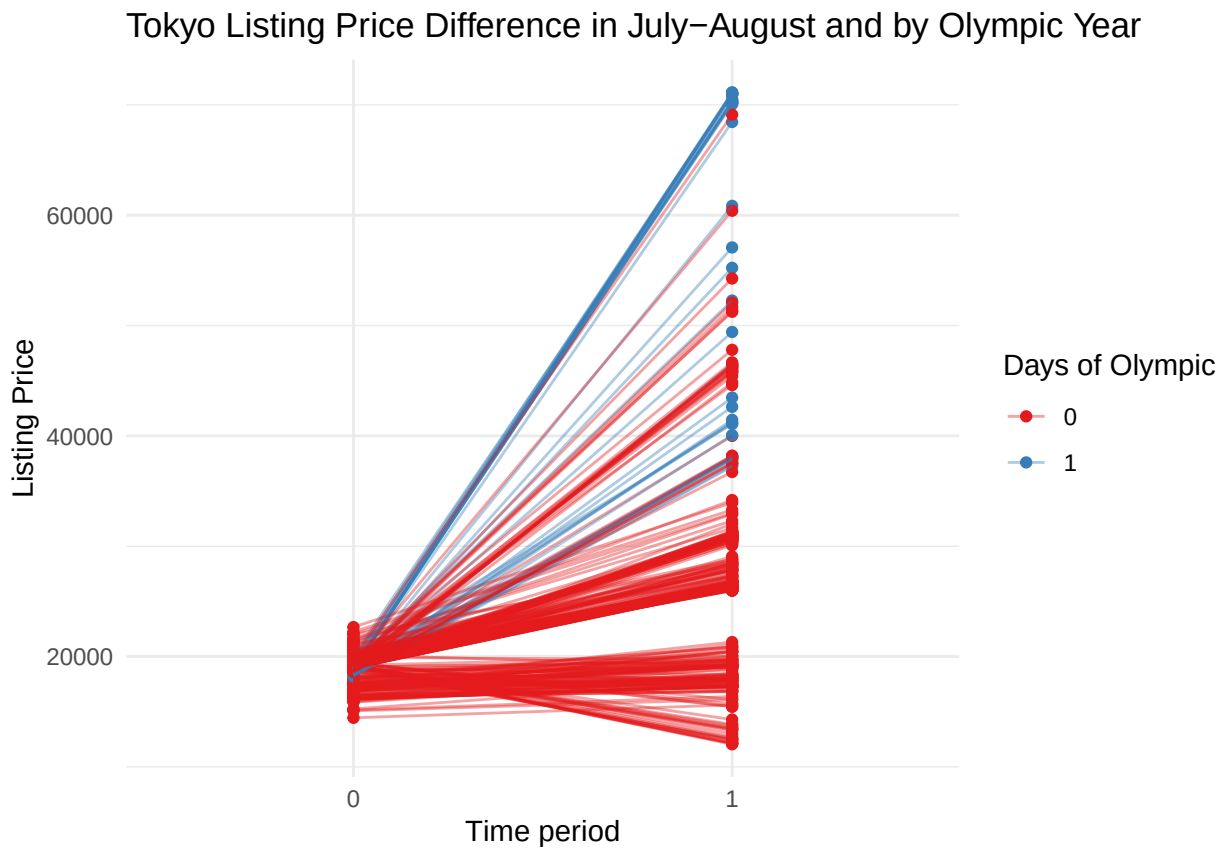


*Figure 2. Tokyo AirBnB Listing Price between May-June and July-August by Olympic Year*

*Table 5. Estimate Price by Treatment (Olmypic Year) and Period (Olmypic Season)*

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 18810. | 380. | 49.6 | 8.13e-57 |
| treatment_group1 | 32.3 | 1161. | 0.03 | 1.09e-01 |
| period1 | 7211. | 537. | 1.72 | 8.75e-02 |
| treatment_group1:period1 | 29126. | 1642. | 20.0 | 3.35e-53 |

By focusing on the two periods, the average price difference between July to August in 2020 (Treatment) and July to August in 2019 (Control) is YEN 29126.01, which is roughly CAD 380.41. The presence of the Olympic Games during the two months has a greater significant impact to the average AirBnB listing prices in Tokyo than during the year.

### Difference between Tokyo 2019, Tokyo 2020, and Paris 2020

Since 2024 Olympics will be held in Paris, we would integrate Paris 2020 price listings and compare the trend with the trend of 2019 Tokyo and 2020 Tokyo. As Paris is not holding the 2020 Olympics, the trend of Paris 2020 AirBnB price should be similar to the the trend of Tokyo 2019 AirBnB price.

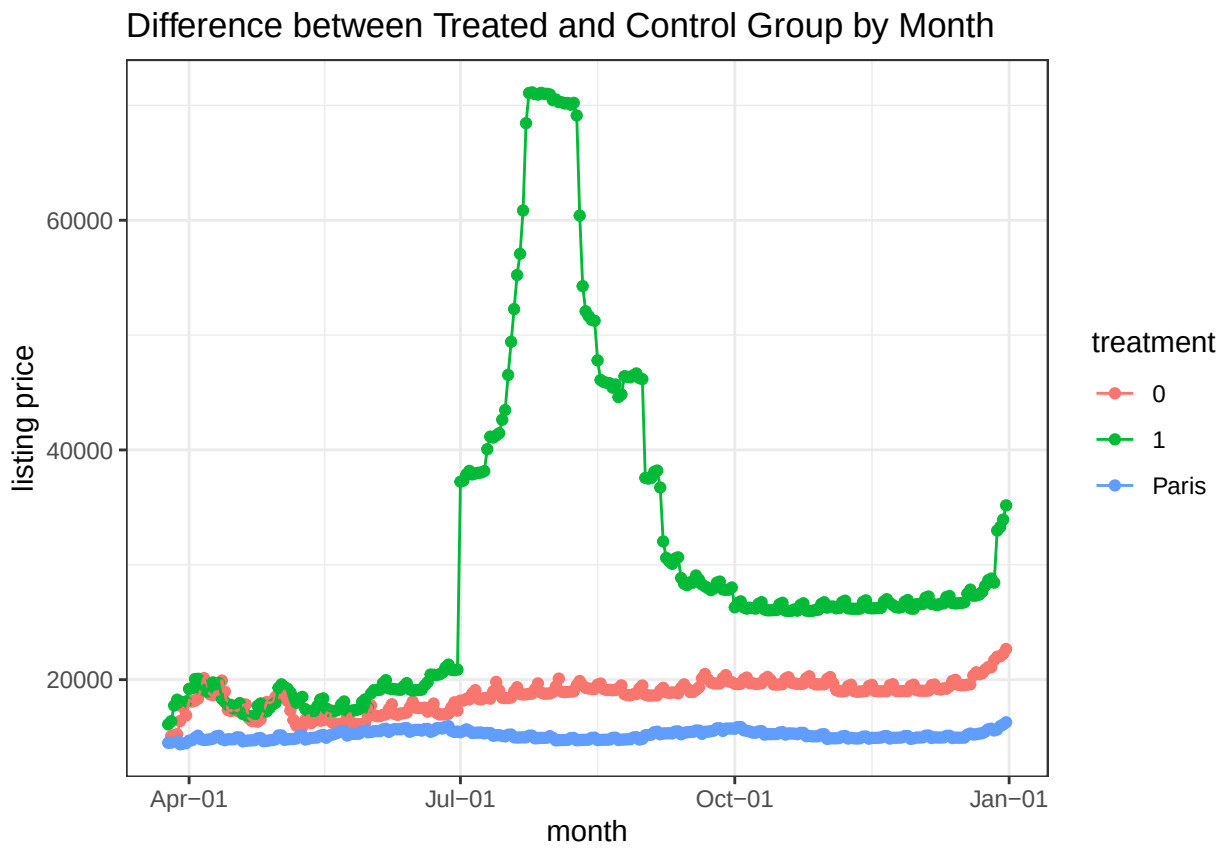


*Figure 3. Differences in Price between Tokyo 2019, Tokyo 2020, and Paris 2020*

*Table 6. Logistics Regression on Price by Day of the Year in 2020 Paris*

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 14784. | 112. | 132. | 2.015e-315 |
| day_of_year | 1.40 | 0.537 | 2.61 | 9.32e-03 |

*Table 7. Difference in Estimate for Listing Price by Treatment between Tokyo 2019, Tokyo 2020, and Paris 2020*

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 18196. | 786. | 23.2 | 5.91e-97 |
| day_of_year | 3.38 | 3.72 | 0.907 | 3.65e-01 |
| treatment1 | 6142. | 1107. | 5.55 | 3.64e-08 |
| treatmentParis | -3411. | 1093. | -3.12 | 1.85e-03 |
| day_of_year:treatment1 | 22.5 | 5.24 | 4.29 | 1.97e-05 |
| day_of_year:treatmentParis | -1.97 | 5.20 | -0.379 | 7.05e-01 |

*Table 8. Difference in Estimate for Listing Price by City, Treatment, and Day of the Year*

| term | Paris 2020 estimate | Tokyo 2019 estimate | Tokyo 2020 estimate |
|---|---|---|---|
| (Intercept) | 14784. | 18196. | 24338. |
| day_of_year | 1.40 | 3.38 | 25.9 |

As seen in Table 8, although the intercept and the slope for Paris 2020 are lower than Tokyo 2019, the trend of listing price in Paris for 2020 (1.40) is more similar to 2019 Tokyo (3.38) than 2020 Tokyo (25.9).

## Discussion

The paper had used diff-in-diff to examine the change in the average AirBnB listing price during the summer Olympic season in Tokyo 2020. From the statistical analysis, the treatment, whether the city is hosting an Olympic event, observed to be a significant factor with an increase of AirBnB listings and an increase of the average listing price.

Comparing the data between 2019 and 2020, there is an overall 59% increase in the number of Tokyo listings and a 55% increase in the average listings price per night. From Figure 1, the average listing price increased by 78% ((37238.19 - 20876.88)/20876.88) on from June 30, 2020, to July 01, 2020, and started to drop after on August 16, 2020, to YEN 30000 by mid-September 2020. During the peak season, the start of July and the end of August is also the 2020 Tokyo Summer Olympics, which takes place from July 22 to August 9, 2020. By focusing the diff-in-diff price on the two periods between July and August, the estimate for the average price difference between July to August in 2020 (Treatment) and July to August in 2019 (Control) is YEN 29126.01, which is roughly CAD 380.41. As the p-value for the diff-in-diff logistic regression model is close to 0 and the alpha value is 0.05, we rejected our null hypothesis, and establish that the presence of Olympic Games during the July and August has a significant impact to the average AirBnB listing prices in 2020 than 2019 for City of Tokyo.

We also measured Tokyo's 2019 and 2020 AirBnB listing price against Paris's 2020 AirBnB listing price. The logistic regressional model reflects that the trend of the average listing price in Paris for 2020 is more similar to 2019 Tokyo than 2020 Tokyo. The model for Paris and the presence of the 2020 Tokyo Olympics resulted in a p-value of 0.7, hence we failed to reject that the 2020 Tokyo Olympics does not affect the listing prices in the City of Paris.

### Ethics

The datasets from Inside AirBnB utilizes public information compiled from the Airbnb website including the availability calendar for 365 days in the future, and the reviews for each listing. Data is verified, cleansed, analyzed and aggregated. Inside AirBnB and the author of this paper is not endorsed by AirBnB or AirBnB's competition, and that the paper is only for practicing statistical analysis and discussion. The dataset that was used does not contain or publish the host or guest information due to privacy concerns.

In regards to the content of the analysis, the latest calendar datasets were last scraped by Insider AirBnB on February 28, 2020. The calendar data was only able to capture listings information that hosts updated before the data scraped such that that accuracy of the dataset may not completely reflect the information on AirBnB website. Besides, the data has two prices: price and adjusted price. This paper looked at the latter over the former as the price is only the baseline price for a listing, while the adjusted price displayed more variation throughout the year. The underlying reason for hosts to adjust the listing prices are, however, indeterminable. Unless we perform text analysis on the listings' title and description or survey the hosts, we cannot assume that Summer Olympics is the sole factor for the spurge of accommodation availability and not from other biases. Furthermore, about 0.2% of the data was also dropped from analysis due to missing price information for the diff-in-diff analysis and timeline comparison (observations that happened on February 29,2020, were dropped), but posed little impact to the overall analysis.

The model used for the analysis is logistic regression which measured the probability of Olympic games pose an effect on AirBnB listing price. The binary response is suited for the analysis as the author uses the diff-in-diff to determined to the Olympic games did or did not affect listing prices. Other statistical model was not used for this paper, due to beyond the scope of the practice, but is open to exploring.

## Reference


- Airbnb by the Numbers: Usage, Demographics, and Revenue Growth. (2020, February 18). Retrieved from https://muchneeded.com/airbnb-statistics/
- David Robinson and Alex Hayes (2019). broom: Convert Statistical Analysis Objects into Tidy Tibbles. R package version 0.5.3. https://CRAN.R-project.org/package=broom
- Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. http://www.jstatsoft.org/v40/i03/.
- H. Wickham., (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org
- Sam Firke (2019). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 1.2.0. https://CRAN.R-project.org/package=janitor
- Paris. Adding data to the debate. (2020, March 15). Retrieved from http://insideairbnb.com/paris/
- Schaal, D., O'Neill, S., & Skift. (2017, January 4). Airbnb Is Becoming an Even Bigger Threat to Hotels Says a New Report. Retrieved from https://skift.com/2017/01/04/airbnb-is-becoming-an-even-bigger-threat-to-hotels-says-a-new-report/
- Schaal, D., & Exhibition Bureau. (2018, November 14). Airbnb's Growth Is Slowing Amid Increasing Competition From Booking and Expedia: Report. Retrieved from https://skift.com/2018/11/14/airbnbs-growth-is-slowing-amid-increasing-competition-from-booking-and-expedia/
- Tokyo. Adding data to the debate. (2020, February 29). Retrieved from http://insideairbnb.com/tokyo/

- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Vacation Rentals, Homes, Experiences & Places. (n.d.). Retrieved from https://www.airbnb.ca/


# Appendix

## Code

### Tokyo AirBnB datasets import and manipulation

```
library(tidyverse)
library(ggplot2)
library(janitor)
library(broom)
library(lubridate)

calendar2020 <- read.csv("datasets/tokyo_calendar.csv",
                         stringsAsFactors = FALSE)
calendar2019 <- read.csv("datasets/tokyo_calendar2019.csv",
                         stringsAsFactors = FALSE)

calendar2019$treatment <- "0"
calendar2020$treatment <- "1"

tokyo_calendar_all <- rbind(calendar2019,calendar2020)
#skimr::skim(tokyo_calendar_all)

tokyo_calendar_all <- tokyo_calendar_all %>%
  mutate(price = str_remove(price, "\\$"),
         price = str_remove(price, ","),
         price = str_remove(price, ","), #removing the second "," of $1,000,000
         price = as.integer(price),
         adjusted_price = str_remove(adjusted_price, "\\$"),
         adjusted_price = str_remove(adjusted_price, ","),
         adjusted_price = str_remove(adjusted_price, ","),
         adjusted_price = as.integer(adjusted_price)) %>%
  drop_na() #9241800 - 1305 = 9240495

# table(is.na(tokyo_calendar_all$price))
# there are 1305 obs where price is absent for a specific listing_id (29647305)

# table(is.na(tokyo_calendar_all$adjusted_price))
# there are 1305 obs where price is absent for a specific listing_id (29647305)
```

```
tokyo_calendar_all <- tokyo_calendar_all %>%
  mutate(date = ymd(date),
         year = year(date),
         month = month(date),
         day = day(date),
         day_of_year = as.numeric(format(date, "%j")))

tokyo_calendar <- tokyo_calendar_all %>%
  select(listing_id,
         date,
         year,
         month,
         day_of_year,
         price,
         adjusted_price,
         treatment) %>%
  group_by(date, month, year, treatment, day_of_year) %>%
  summarise(median_price = median(adjusted_price),
            mean_price = mean(adjusted_price),
            counts = n())

#replace all years to 0 to group by months and day
tokyo_calendar$no_year <- tokyo_calendar$date
year(tokyo_calendar$no_year) <- 0
```

### Measure of Central Tendency for Tokyo AirBnB Listings

```
tokyo_calendar_all %>%
#  filter(treatment == "0") %>%   #calendar 2019
  filter(treatment == "1") %>%  #calender 2020
  mutate(listing_id = as.factor(listing_id)) %>%
  skimr::skim()
```

### Tokyo Data Visualization

```
###   Median Price  ###
##### By day of the year ####

tokyo_calendar %>%
  ggplot(aes(x = day_of_year, y = median_price , color = treatment)) +
  geom_point() +
  geom_line() +
```

```
  labs(x = "Day of Year",
       y = "listing price",
       fill = "Treatment") +
  theme_bw()

###   Median ###

tokyo_calendar %>%
  ggplot(aes(x = no_year,
             y = median_price ,
             color = treatment)) +
  geom_point() +
  geom_line() +
  labs(x = "month",
       y = "listing price",
       fill = "Treatment") +
  facet_grid(facets = treatment ~ .) +
  scale_x_date(labels = function(x) format(x, "%b-%d"),#) +
               limits = as.Date(c("0000-01-01","0000-12-31"))) +
  theme_bw()

###   Average Price  ###
##### By day of the year ####

tokyo_calendar %>%
  ggplot(aes(x = day_of_year, y = mean_price , color = treatment)) +
  geom_point() +
  geom_line() +
  labs(x = "Day of Year",
       y = "listing price",
       fill = "Treatment") +
  theme_bw()

###   Average Price  ###
##### By month ####

tokyo_calendar %>%
#  filter(no_year >= as.Date("0000-03-25")) %>%
  filter(day_of_year >= 85) %>%
  ggplot(aes(x = no_year, #replace all years to 0 to group by months and day
             y = mean_price ,
             color = treatment)) +
  geom_point() +
  geom_line() +
```

```
  labs(title = "Trend of Listing Prices by Month",
       x = "month",
       y = "listing price",
       fill = "Treatment") +
#  facet_grid(facets = treatment ~ .) +
  scale_x_date(labels = function(x) format(x, "%b-%d")) +
  theme_bw()
```

### Logistic Regression Chunk

```
###  Logistics Regression on Price by Day of the Year in 2019*
tokyo_lm1 <- tokyo_calendar%>%
  filter(treatment == "0")

lm1 <- lm(mean_price ~ day_of_year,
          tokyo_lm1)
tidy(lm1)

### Logistics Regression on Price by Day of the Year in 2020*

tokyo_lm2 <- tokyo_calendar%>%
  filter(treatment == "1")

lm2 <- lm(mean_price ~ day_of_year,
          tokyo_lm2)
tidy(lm2)

### Estimate for Listing Price
#### by Treatment (Olmypic Year or Not) and Day of the Year

lm3 <- lm(mean_price ~ day_of_year*treatment,
          tokyo_calendar)
tidy(lm3)
```

### Difference between July 01 and August 31 of 2019 and 2020

```
olympics <- tokyo_calendar_all%>%
  select(day_of_year, adjusted_price, treatment) %>%
  filter(day_of_year != 366) %>%
  mutate(treatment_group = case_when(
    day_of_year >= 183 & day_of_year <= 221 ~ 1,
    TRUE ~ 0
```

```
  )) %>%
  mutate(period = treatment) %>%
  group_by(day_of_year, period, treatment_group) %>%
  summarise(average_price = mean(adjusted_price)) %>%
  select(day_of_year, period, treatment_group, average_price)

olympics$treatment_group <- as.factor(olympics$treatment_group)
olympics$period <- as.factor(olympics$period)

olympics %>%
  ggplot(aes(x = period,
             y = average_price,
             color = treatment_group)) +
  geom_point() +
  geom_line(aes(group = day_of_year), alpha = 0.4) +
  theme_minimal() +
  labs(title = "Tokyo Listing Price Difference in July-August and by Olympic Year",
       x = "Time period",
       y = "Listing Price",
       color = "Days of Olympic") +
  scale_color_brewer(palette = "Set1")

olympics_pivot <- olympics %>%
  pivot_wider(names_from = period,
              values_from = average_price,
              names_prefix = "price_") %>%
  mutate(difference = price_1 - price_0) %>%
  group_by(treatment_group) %>%
  summarise(average_difference = mean(difference))

olympics_pivot$average_difference[2] - olympics_pivot$average_difference[1]

lm4 <- lm(average_price ~ treatment_group*period,
                          data = olympics)
tidy(lm4)
```

### Paris Data

```
paris_calendar <- read.csv('datasets/paris_calendar.csv',
                           stringsAsFactors = FALSE)

paris_calendar$treatment <- "Paris"
```

```r
paris_calendar <- paris_calendar %>%
  mutate(price = str_remove(price, "\\$"),
         price = str_remove(price, ","),
         price = as.integer(price),
         price = price * 118.05, #exchange rate on March 21
         adjusted_price = str_remove(adjusted_price, "\\$"),
         adjusted_price = str_remove(adjusted_price, ","),
         adjusted_price = as.integer(adjusted_price),
         adjusted_price = adjusted_price * 118.05,
         )%>%
  drop_na()

paris_calendar <- paris_calendar %>%
  mutate(date = ymd(date),
         year = year(date),
         month = month(date),
         day = day(date),
         day_of_year = as.numeric(format(date, "%j")))

paris_calendar <- paris_calendar %>%
  select(listing_id,
         date,
         year,
         month,
         day_of_year,
         price,
         adjusted_price,
         treatment) %>%
  group_by(date, month, year, treatment, day_of_year) %>%
  summarise(median_price = median(adjusted_price),
            mean_price = mean(adjusted_price),
            counts = n())

paris_calendar$no_year <- paris_calendar$date
year(paris_calendar$no_year) <- 0

tpmerge <- rbind(tokyo_calendar,paris_calendar)

tpmerge %>%
  filter(day_of_year >= 85) %>%
  ggplot(aes(x = no_year,
             y = mean_price ,
             color = treatment)) +
  geom_point() +
```

```
  geom_line() +
  labs(title = "Difference between Treated and Control Group by Month",
       x = "month",
       y = "listing price",
       fill = "Treatment") +
#  facet_grid(facets = treatment ~ .) +
  scale_x_date(labels = function(x) format(x, "%b-%d")) +
  theme_bw()

### Logistics Regression on Price by Day of the Year in 2020 Paris
tp_lm <- tpmerge%>%
  filter(treatment == "Paris")

lm5 <- lm(mean_price ~ day_of_year,
          tp_lm)

tidy(lm5)
```

**Difference in Estimate for Listing Price by Treatment between Tokyo 2019, Tokyo 2020, and Paris 2020**

```
### Difference in Estimate for Listing Price
#### by Treatment between Tokyo 2019, Tokyo 2020, and Paris 2020
lm6 <- lm(mean_price ~ day_of_year*treatment,
          tpmerge)

tidy(lm6)
```