An RDD Approach to AirBnB Listing Price - Before, During, and After the 2020 Tokyo Olympics

Roxanne Chui

03/04/2020

Abstract

Diff-in-Diff and regression discontinuity design were used to analyze and assess the casualty of the Summer Olympics in Tokyo on the average AirBnB listing price in 2020. The results of Diff-in-Diff performed on this paper was identical to a similar paper published in March 2020. On RDD, three methods of dissecting and analyzing data produced different estimates but had similar features. Overall, the linear regression models performed individually to produce a better and accurate representation of the trend of 2020 listing prices than performing a multivariant linear regression model. RDD on this dataset has a couple of weaknesses due to multiple cut-off periods and a sharp threshold.

Introduction

On March 30, 2020, the Olympic Games Tokyo 2020 was announced and postponed from July 24 - August 9, 2020, to 23 July - 8 August 2021. According to WashingtonPost (2020), as Japan has a wide range of accommodation options for Olympics travellers, from Airbnbs to rustic guesthouses, luxury hotels, capsule hotels and more. Each accommodation type will have its policies regarding cancelling or rescheduling lodging. Although there is no new release of new Tokyo Inside AirBnB data as of April 3, 2020, the author would use regression discontinuity on of Tokyo AirBnB calendar and listings in 2020, especially before, during, and after the (presumptive) Olympic Games Tokyo 2020 for 2021.

Background

Diff-in-Diff of Tokyo 2019 & 2020

A previous paper from March used diff-in-diff to examine the change in the average AirBnB listing price during the summer Olympic season in Tokyo 2020. That paper observed and discussed that the average price difference between July

to August in 2020 (Treatment), when Olympic Games would be happening, and July to August in 2019 (Control) is YEN 29126.01, which is roughly CAD 380.41.

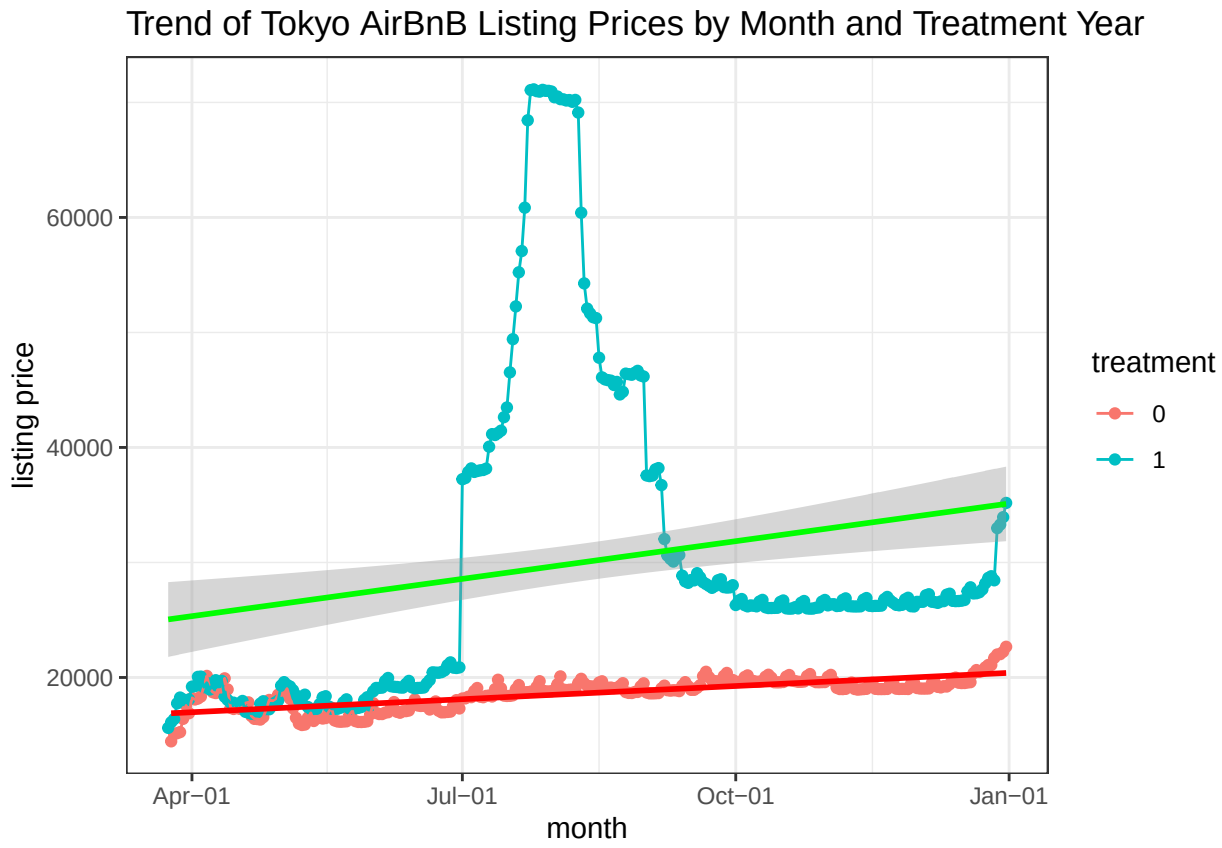

Figure 1. Difference between Listing Price between 2019 (Control) and 2020 (Treatment)

Table 3. 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

Regression Discontinuity of the Year 2020

One variable in that failed to factor in the March paper was the 2020 Summer Paralympics, which the games were originally planned to held between August 25 and September 6, 2020.

Table 4: Days of Olympics and Paralympics

Periods	Date (2020)	Day of the Year
Olympics	July 24 to August 9	183 to 221
Paralympics	August 25 to September 6	238 to 250

RDD Method 1: 3 Cut-offs and 4 Periods

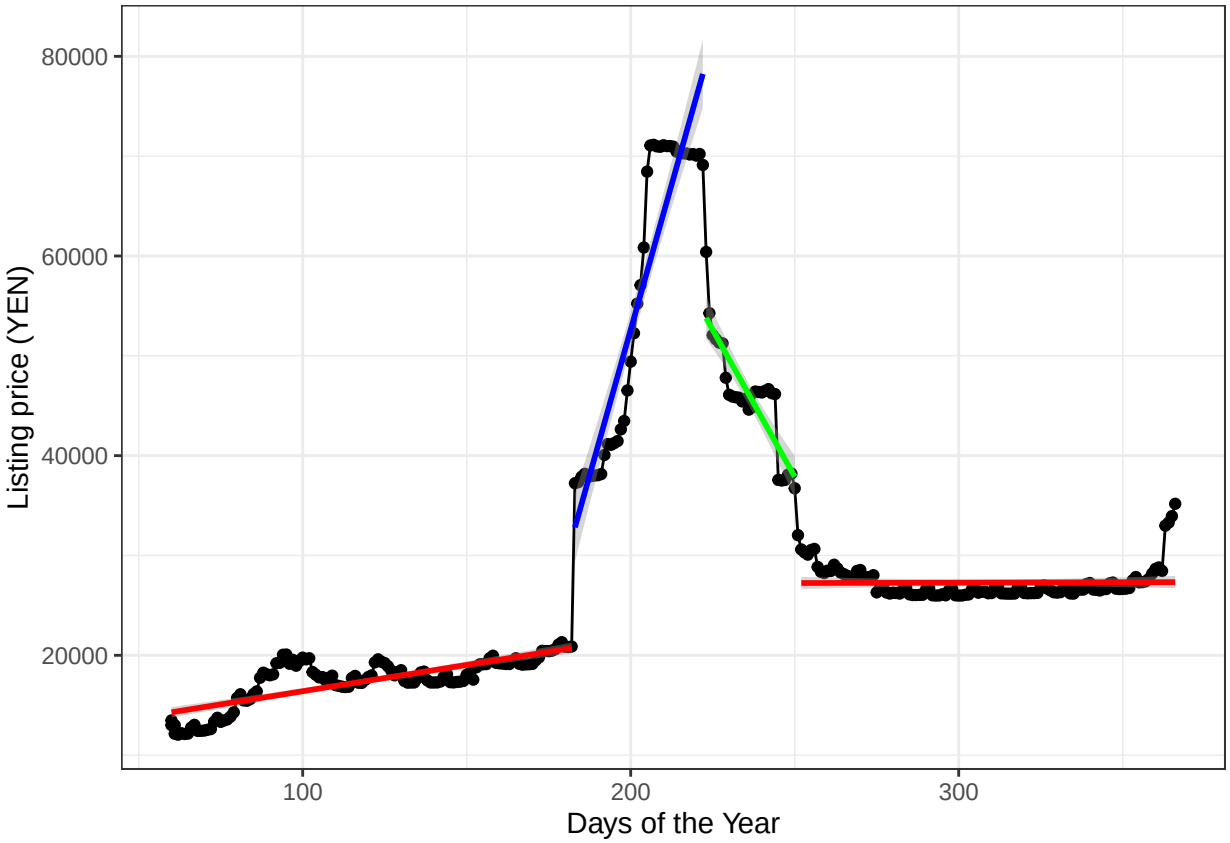

Figure 2. Effect of of Olympics and Paralympics Games on 2020

The start to end date of the data used begins from 2020-02-29 to 2020-12-31, the day that the 2020 data was scrapped and the end of the year, for better representation for future analysis during the analysis. Based on Figure 1, Figure 2 observe 2020 into three periods based on the AirBnB listing prices: before the month of the Olympic Games, during the Olympic Games, and after the Olympic Games. The middle period (during the Olympic Games) can further sub-categorized into Olympic events and Paralympic events. From Figure 2, the slopes of the four periods were calculated and resulted in Table 5 to Table 8.

Table 5: Regression Line 1: Before 2020-07-01 (Day 60 to 182)

term	estimate	std.error	statistic	p.value
(Intercept)	11128.	474.	23.5	2.86e-47
day_of_year	52.7	3.78	13.9	4.38e-27

Table 6: Regression Line 2: From 2020-07-01 to 2020-08-09 (Day 183 to 222)

term	estimate	std.error	statistic	p.value
(Intercept)	-180363.	15075.	-12.0	1.87e-14
day_of_year	1165.	74.3	15.7	3.56e-18

Table 7: Regression Line 3: From 2020-08-10 to 2019-09-06 (Day 222 to 250)

term	estimate	std.error	statistic	p.value
(Intercept)	185060.	15301.	12.1	3.52e-12
day_of_year	-589.	64.7	-9.10	1.44e-09

Table 8: Regression Line 4: After 2020-09-06 (Day 251 onwards)

term	estimate	std.error	statistic	p.value
(Intercept)	27034.	1460.	18.5	5.08e-36
day_of_year	0.81	4.70	0.17	0.86

Table 9: Regression Discontinuity on Period 1, 2, 3, and 4

term	estimate	std.error	statistic	p.value
(Intercept)	12275.	1361.	9.02	2.21e-17
day_of_year	43.2	10.5	4.11	5.07e-05
period2	34183.	1361.	25.1	3.70e-76
period3	24175.	1699.	14.2	1.42e-35
period4	1732.	2113.	0.820	0.413

RDD Method 2: 3 Cut-offs and 3 Periods

The second method of RDD is similar to the first method, having the four cut-offs but group by three periods: Olympics, Paralympics, and no Olympics.

Table 10: Regression Discontinuity on Olympics, Paralympics, and no Olympics

term	estimate	std.error	statistic	p.value
(Intercept)	11411.	856.	13.3	3.16e-32
day_of_year	51.3	3.65	14.0	7.25e-35
period Olympics	33731.	965.	35.0	2.50e-108
period Paralympics	22295.	1131.	19.7	3.16e-56

RDD Method 3: 1 cut-off and 2 Periods

The last method of RDD towards the data is by removing the Olympic periods and only focus on two periods: before 2020-07-01 as period 0 and after 2020-09-13 as period 1. Period 0 and 1 would represent the listing prices before and after the Olympic games affecting the 2020 listing price trend.

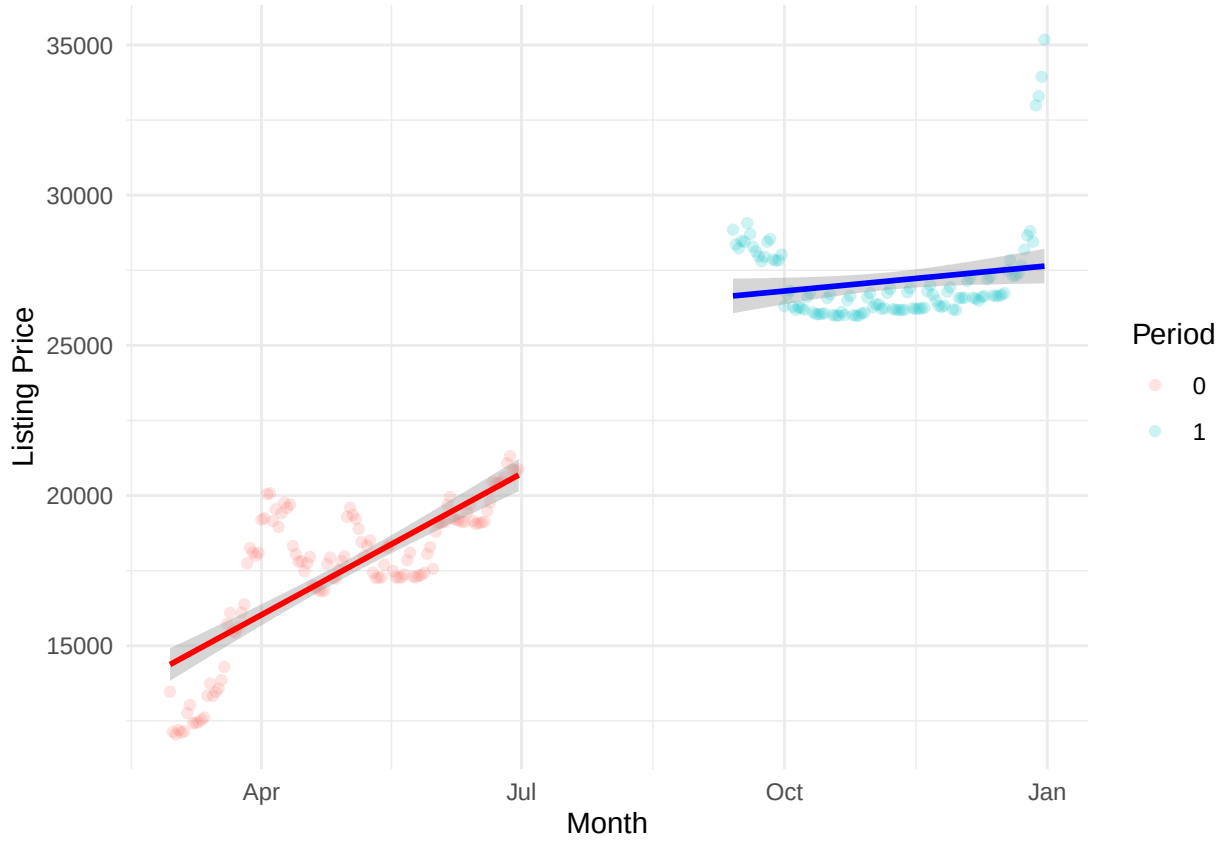

Figure 3. RDD Effect of without Olympics

Table 11: Regression Discontinuity before and after Olympics

term	estimate	std.error	statistic	p.value
(Intercept)	13423.	424.	31.7	7.73e-86
day_of_year	33.9	3.27	10.4	5.92e-21
period	3147.	661.	4.76	3.40e-06

Discussion

The paper used difference-in-difference and regression discontinuity design to analyze the effect of the Olympic Games on the AirBnB Listings in Tokyo 2019 and 2020.

Figure 1 and Table 1 to 3 illustrate 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. Given that the alpha was 0.05, the null hypothesis was rejected with a p-value close to zero. Therefore, the presence of the Olympic Games during July and August has a significant impact on the average Airbnb listing prices in 2020 than 2019 for the City of Tokyo before the postponement.

There are three methods to dissect 2020 into different periods. Method one cuts the year at 3 days of the year: Day 183, Day 223, and Day 251. Before Day 183 was the period before the spike of the average listing price, where the average listing price has 52.7 YEN increase by the end of June 30, 2020. Between Day 183 to 222, which the Olympics Games would be happening, has an 1165 YEN increase. By Day 223 to 250, which the Paralympics Games would be happening, average listing prices dropped by 589 YEN. On Day 251 onwards, after the end of all Olympic events, the listings prices would insignificantly increase the prices until the end of the year. In Table 12, the four separate linear regression reflects the directionality of the trend of 2020, however, when performing linear regression on the four periods together as seen in Table 13, the estimates are compared against the linear regression without linear discontinuity design.

Table 12. Difference in Estimate for 4 Individual Periods

term/estimate	Pre-Day 183	Day 183 to 222	Day 223 to 250	Post-Day 250
(Intercept)	11128.	-180363.	185060.	27034.
day_of_year	52.7	1165.	-589.	0.81

Table 13: Regression Discontinuity on 4 Individual Periods Combined

term	estimate	std.error	statistic	p.value
(Intercept)	12275.	1361.	9.02	2.21e-17
day_of_year	43.2	10.5	4.11	5.07e-05
period2	34183.	1361.	25.1	3.70e-76
period3	24175.	1699.	14.2	1.42e-35
period4	1732.	2113.	0.820	0.413

Method 2 also cuts the year into three days: Day 183, Day 223, and Day 251. Instead of having four individual periods, the Day 183 to 222 is the "Olympics" period, Day 223 to 250 is the "Paralympics" period, while the rest of the year is the "No Olympics" period. Table 10 reflect that during the non-Olympic season the average listing price in Tokyo would have a 52 YEN increase throughout the year. During the "Olympics" and "Paralympics" periods the average listing price would have a further 33731 YEN and 22295 YEN increase. The three periods used as cut-off had a significant impact on the difference in listing price throughout 2020.

Method 3, unlike Method 1 and 2, forgo the peak periods and focused on the significance of

discontinuity between pre- and post- Olympics. Although removing the Olympic period left a gap in the data, the difference in average listing price between pre- and post-Olympics is 3147 YEN. Compared to 2019, given the same dating conditions, the difference in average listing price between pre- and post-Olympics for 2019 is 2226 YEN. Therefore, the difference-in-difference between periods before July 1 and periods after September 13 is 921 YEN.

Table 14: Regression Discontinuity before and after "Olympics" for 2019

term	estimate	std.error	statistic	p.value
(Intercept)	17025.	281.	60.7	6.49e-133
day_of_year	1.47	2.01	0.73	5.92e-21
period	2226.	378.	5.89	3.40e-06

Disclaimer

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.

Weaknesses

Throughout the statistical analysis and discussion, there is a significant weakness of using RDD on this dataset. First, there are multiple slopes across the periods, mainly the two points of RDD between July 1 and August 8. Although the days of the "discontinuous spike and drops" were identifiable, another analyst may consider the different thresholds for the data. Second, as seen as the three methods of approach to RDD, different methods of "dissecting" the period produce different estimates. Although the author attempted to answer the casualty of Olympic games to that listing prices in Tokyo 2020, experimenting with different methods of dissecting the periods had led to various results that cannot confidently conclude the implications of the results.

Ethics

The following ethics issues were addressed: 1. The personal privacy of individuals is limited to what is published on AirBnB website. As the analysis is overall trend analysis, business and guest name and id were omitted from the result analysis. The analysis does not target business, company, or guest personnel but an overview insight. 2. Beneficial wise, the problem set is an academic exercise to LDD statistical analysis and AirBnB guests for casual reading. 3. All codes and links to datasets are referenced and attached as appendices.

Regarding 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 prices. 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/
- Compton, N. B. (2020, March 31). How to cancel, change or make new travel plans for the Olympics in 2021. Retrieved from https://www.washingtonpost.com/travel/tips/how-cancel-change-or-make-new-travel-plans-olympics/
- 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
- 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-increasingcompetition-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

```
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)

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

tokyo_calendar <- tokyo_calendar_all %>%
  select(listing_id,
         date,
         day_of_year,
         price,
         adjusted_price,
         treatment) %>%
  group_by(date,treatment, day_of_year) %>%
  summarise(mean_price = mean(adjusted_price))
```

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

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

tokyo_calendar %>%
  filter(day_of_year >= 84) %>%
  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() +
  geom_smooth(data = tokyo_calendar %>%
                filter(treatment == "0") %>%
                filter(day_of_year >= 84),
              method='lm',
              color = "red") +
    geom_smooth(data = tokyo_calendar %>%
                filter(treatment == "1") %>%
                filter(day_of_year >= 84),
              method='lm',
              color = "green") +
  labs(title = "Trend of Tokyo AirBnB Listing Prices by Month and Treatment Year",
       x = "month",
       y = "listing price",
       fill = "Treatment") +
  scale_x_date(labels = function(x) format(x, "%b-%d")) +
  theme_bw()
```

Table 1. Logistics Regression on Price by Day of the Year in 2019*

```r
##### Regression Line 1: Year 2019
lm1red <- tokyo_calendar %>%
  filter(treatment == "0") %>%
  filter(day_of_year > 84)

lm(mean_price ~ day_of_year, data = lm1red) %>%
  tidy()
```

```
## # A tibble: 2 x 5
```

```
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  15896.    150.        106.  4.20e-227
## 2 day_of_year     12.3     0.629      19.5 4.56e- 54
```

Table 2. Logistics Regression on Price by Day of the Year in 2020*

```
##### Regression Line 2: Year 2020
lm2green <- tokyo_calendar %>%
  filter(treatment == "1") %>%
  filter(day_of_year > 84)

lm(mean_price ~ day_of_year, data = lm2green) %>%
  tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  22251.     2438.       9.13 1.43e-17
## 2 day_of_year     34.9      10.2      3.43 6.89e- 4
```

```
### Logistics Regression on Price by Day of the Year in 2020
lm2green <- tokyo_calendar %>%
  filter(treatment == "1") %>%
  filter(day_of_year > 84)

lm(mean_price ~ day_of_year, data = lm2green) %>%
  tidy()

## Estimate for Listing Price by Treatment (Olmypic Year) and Day of the Year*

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

tidy(lm3)
```

RDD Method 1: 3 Cut-offs and 4 Periods

```
## Effect of of Olympics and Paralympics Games on 2020
tokyo_calendar %>%
```

```
  filter(treatment =='1') %>%
  filter(day_of_year >= 60)%>%
  ggplot(aes(x = day_of_year,
             y = mean_price)) +
  geom_point() +
  geom_line() +
  geom_smooth(data = tokyo_calendar %>% #before 2020-07-01
                filter(treatment == "1") %>%
                filter(day_of_year >= 60 & day_of_year < 183),
              method='lm',
              color = "red") +
  geom_smooth(data = tokyo_calendar %>% # 2020-07-01 to 2020-08-09
                filter(treatment == "1") %>%
                filter(day_of_year >= 183 & day_of_year <= 222),
              method='lm',
              color = "blue") +
  geom_smooth(data = tokyo_calendar %>% # 2020-08-10 to 2020-09-06
                filter(treatment == "1") %>%
                filter(day_of_year > 222 & day_of_year <= 250),
              method='lm',
              color = "green") +
  geom_smooth(data = tokyo_calendar %>% #after 2020-09-06
                filter(treatment == "1") %>%
                filter(day_of_year > 251),
              method='lm',
              color = "red") +
  theme_bw() +
  labs(x = "Days of the Year",
       y = "Listing price (YEN)")

##### Regression Line 3: Day 60 to 182
lm1black <- tokyo_calendar %>%
  filter(treatment == "1") %>%
  filter(day_of_year >= 60 & day_of_year < 183)

lm(mean_price ~ day_of_year, data = lm1black) %>%
  tidy()

##### Regression Line 4: Day 183 to 221
lm2blue <- tokyo_calendar %>%
  filter(treatment == "1") %>%
  filter(day_of_year >= 183 & day_of_year <= 222)
lm(mean_price ~ day_of_year, data = lm2blue) %>%
  tidy()
```

```
##### Regression Line 5: Day 222 to 256 (2nd Week of September)
lm3green <- tokyo_calendar %>%
  filter(treatment == "1") %>%
  filter(day_of_year > 222 & day_of_year <= 250)
lm(mean_price ~ day_of_year, data = lm3green) %>%
  tidy()

##### Regression Line 6: Day 256 onwards
lm4black <- tokyo_calendar %>%
  filter(treatment == "1") %>%
  filter(day_of_year > 251)
lm(mean_price ~ day_of_year, data = lm4black) %>%
  tidy()

#  Regression By Periods
lm_period1 <- tokyo_calendar %>%
  filter(treatment == "1") %>%
  mutate("period" = case_when(
    day_of_year >= 60 & day_of_year < 183 ~ "1",
    day_of_year >= 183 & day_of_year <= 221 ~ "2",
    day_of_year >221 & day_of_year <= 250 ~ "3",
    day_of_year > 250 ~  "4" ))

lm(mean_price ~ day_of_year + period, data = lm_period1) %>%
  tidy()
```

RDD Method 2: 3 Cut-offs and 3 Periods

```
# By Event timeline
lm2020 <- tokyo_calendar %>%
  filter(treatment == "1") %>%
  filter(date >= as.Date("2020-02-29") &
           date <= as.Date("2020-12-31"))%>%
  mutate("period" = case_when(
    day_of_year >= 183 & day_of_year <= 222 ~ "Olympics",
    day_of_year >222 & day_of_year <= 250 ~ "Paralympics",
    TRUE ~ "No Olympics",)
    )

lm(mean_price ~ day_of_year + period, data = lm2020) %>%
  tidy()
```

RDD Method 3: 1 cut-off and 2 Periods

```
nolympics <- tokyo_calendar %>%
  filter(treatment == "1") %>%
  filter(date >= as.Date("2020-02-29") &
           date <= as.Date("2020-12-31"))%>%
  mutate("period" = case_when(
    date < as.Date("2020-07-01") ~ 0,
    date >= as.Date("2020-09-13") ~ 1,
    TRUE ~ 2)
    ) %>%
  filter(period != 2)

nolympics %>%
  filter(period != 2) %>%
  ggplot(aes(x = date,
             y = mean_price,
             colour = as.factor(period))) +
  geom_point(alpha = 0.2) +
  geom_smooth(data = nolympics %>% filter(period == 0),
              method='lm',
              color = "red") +
  geom_smooth(data = nolympics %>% filter(period > 0),
              method='lm',
              color = "blue") +
  theme_minimal() +
  labs(x = "Month",
       y = "Listing Price",
       colour = 'Period')

# Regression Discontinuity before and after "Olympics" for 2020
lm(mean_price ~ day_of_year + period, data = nolympics) %>%
  tidy()

## Regression Discontinuity before and after "Olympics" for 2019
lm2019 <- tokyo_calendar %>%
  filter(treatment == "0") %>%
  filter(date >= as.Date("2019-03-25") &
           date <= as.Date("2019-12-31"))%>%
  mutate("period" = case_when(
    date < as.Date("2019-07-01") ~ 0,
    date >= as.Date("2019-09-13") ~ 1,
    TRUE ~ 2)
    ) %>%
```

```
  filter(period != 2)

lm(mean_price ~ day_of_year + period, data = lm2019) %>%
  tidy()
```