Tokyo Airbnb: What do we have here among millions of observations?

Roxanne Chui

Short Bio

- Undergraduate in forensic anthropology
  - Bayesian Statistics
  - Predictive Shotgun Spread Pattern Analysis with RCMP
- Pharmacy Assistant for Shoppers Drug Mart
- Master’s candidate in Human-centred Data Science
  - Experimental Design
- Github: https://github.com/roax888/Tokyo_AirBnB_Analysis

Tokyo, Japan

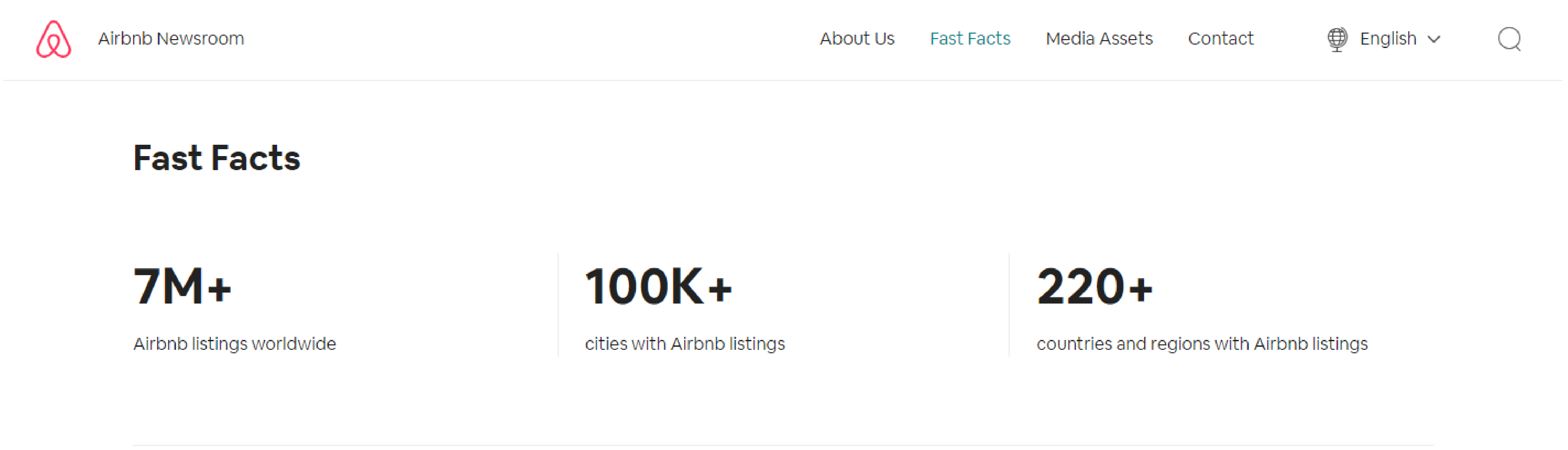
Background - Tokyo

- Peak travel season in Tokyo:
  - March & April – Sakura season
  - October & November – autumn foliage season.
- Wide range of accommodation
  - Airbnb, rustic guesthouses, luxury hotels, capsule hotels and more.
- International Olympic Committee partnered up with Airbnb in providing accommodation and experiences through to 2028 [2].
- Olympic Games Tokyo 2020 (Olympics and Paralympic) was postponed to 2021
  - July 24 - August 9, 2020 to 23 July - 8 August 2021
  - 25 August - 6 September 2020 to 24 August - 5 September 2021
- <u>How would Olympic games affect Airbnb listing prices during the Olympic year?</u>

[2] IOC, "IOC and Airbnb announce major global Olympic partnership," Olympic News, 19 April 2020. [Online]. Available: https://www.olympic.org/news/ioc-and-airbnb-announce-major-global-olympic-partnership. [Accessed 8 July 2020].

Airbnb

From https://news.airbnb.com/fast-facts/



Fast Facts

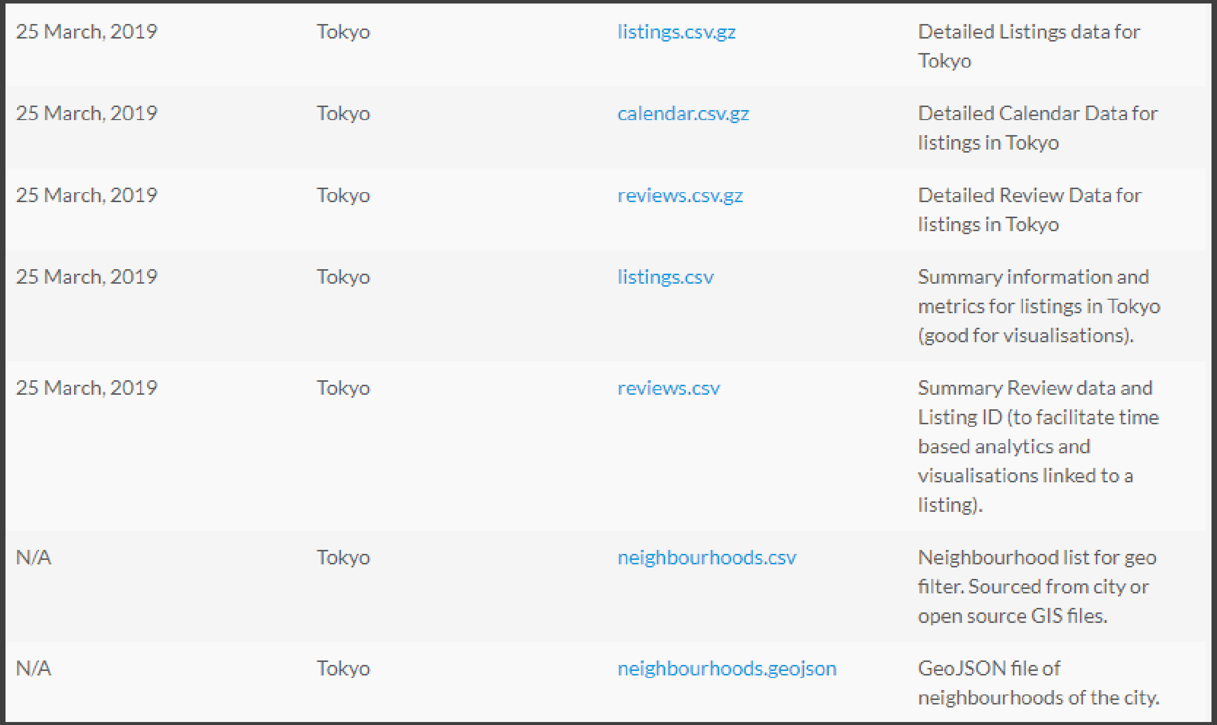
7M+
Airbnb listings worldwide

100K+
cities with Airbnb listings

220+
countries and regions with Airbnb listings

About Airbnb data

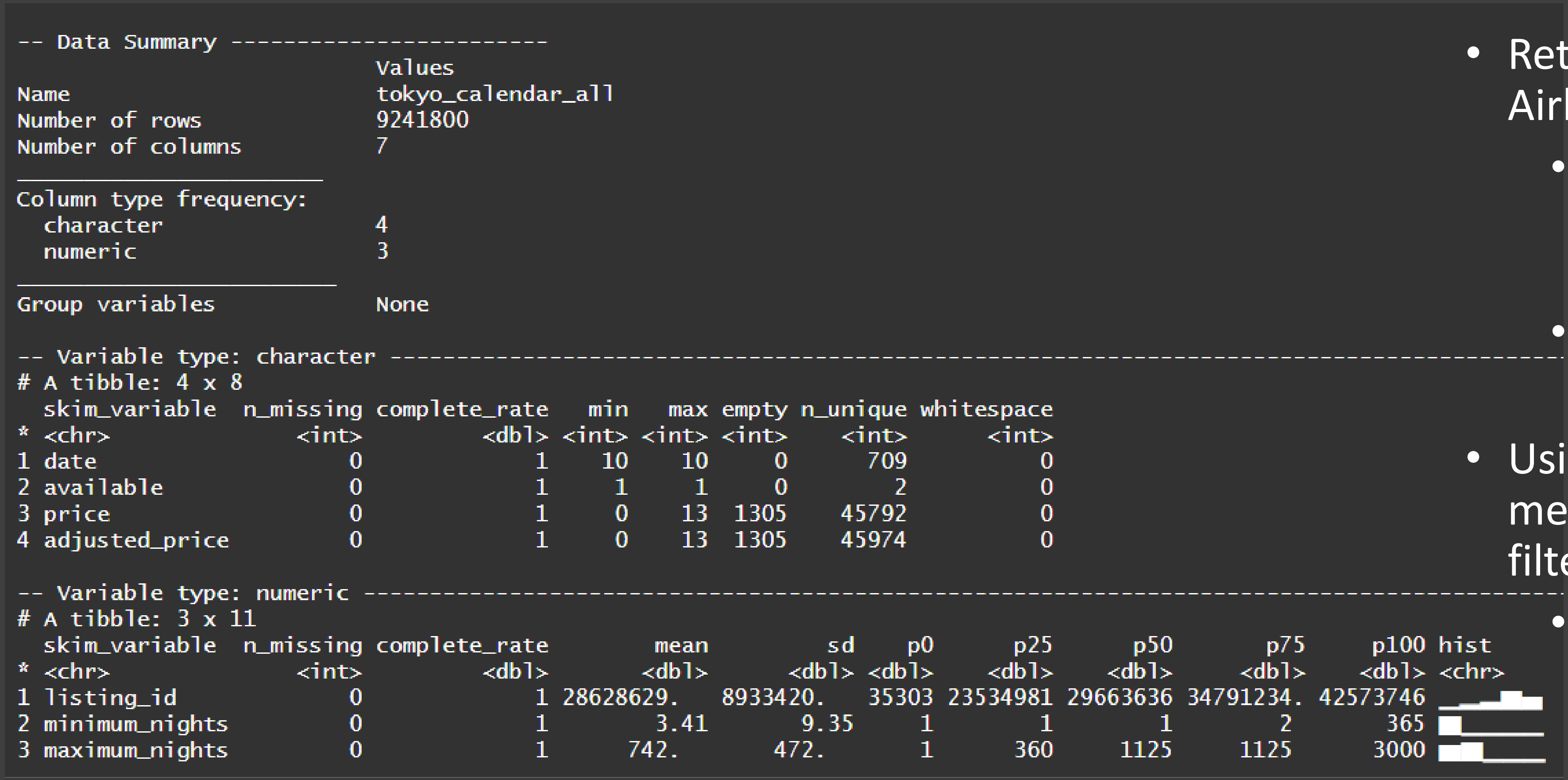
- Inside Airbnb[1] by Murray Cox
- 83 major cities across North America, Latin America, Europe, Africa, Asia/Pacific
- Scraped and publish every month
- Earliest data scraped varies between cities
- Data types:
  - Calendar (detailed)
  - Listing csv (detailed & summary)
  - Review (detailed & summary)
  - Neighbourhood (csv & geojson)



25 March, 2019	Tokyo	listings.csv.gz	Detailed Listings data for Tokyo
25 March, 2019	Tokyo	calendar.csv.gz	Detailed Calendar Data for listings in Tokyo
25 March, 2019	Tokyo	reviews.csv.gz	Detailed Review Data for listings in Tokyo
25 March, 2019	Tokyo	listings.csv	Summary information and metrics for listings in Tokyo (good for visualisations).
25 March, 2019	Tokyo	reviews.csv	Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing).
N/A	Tokyo	neighbourhoods.csv	Neighbourhood list for geo filter. Sourced from city or open source GIS files.
N/A	Tokyo	neighbourhoods.geojson	GeoJSON file of neighbourhoods of the city.

[1] M. Cox, "About Inside Airbnb," Inside Airbnb, [Online]. Available: http://insideairbnb.com/.

Dataset – calendar.csv



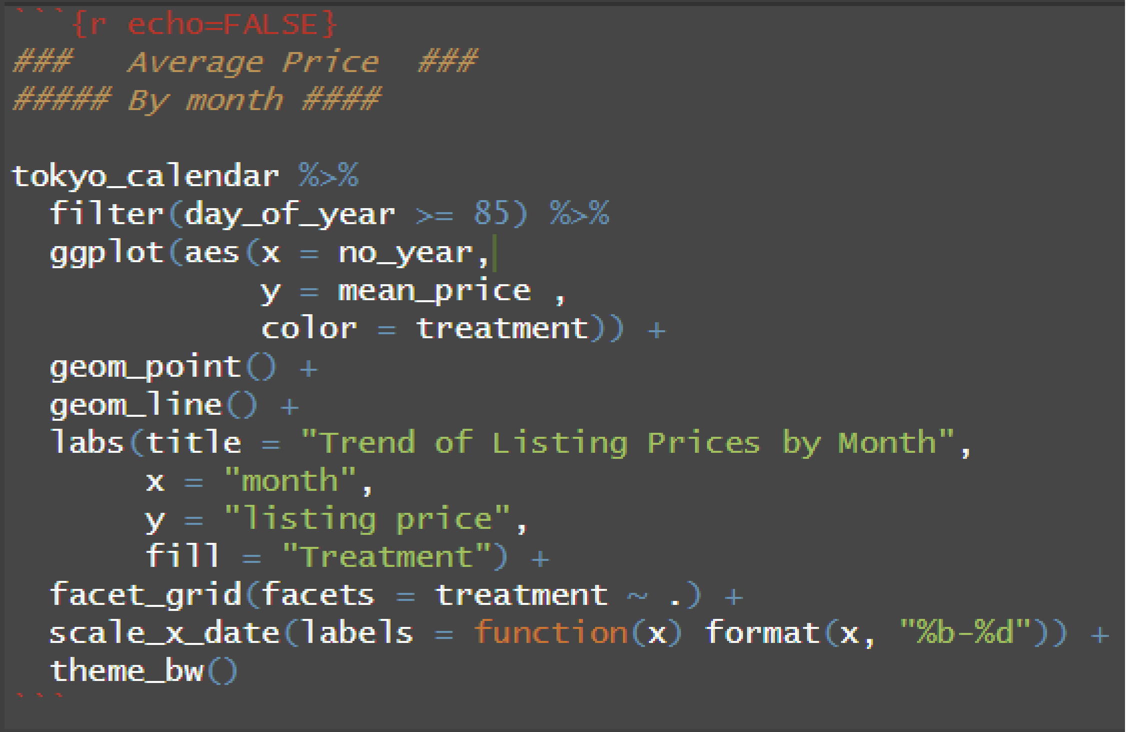
```
-- Data Summary ------------------------
                           Values
Name                       tokyo_calendar_all
Number of rows             9241800
Number of columns          7
_______________________
Column type frequency:
  character                4
  numeric                  3
________________________
Group variables            None

-- Variable type: character ------------------------------------------------------------------------------------
# A tibble: 4 x 8
  skim_variable  n_missing complete_rate   min   max empty n_unique whitespace
* <chr>              <int>         <dbl> <int> <int> <int>    <int>      <int>
1 date                   0             1    10    10     0      709          0
2 available              0             1     1     1     0        2          0
3 price                  0             1     0    13  1305    45792          0
4 adjusted_price         0             1     0    13  1305    45974          0

-- Variable type: numeric --------------------------------------------------------------------------------------
# A tibble: 3 x 11
  skim_variable  n_missing complete_rate       mean         sd    p0      p25      p50       p75     p100 hist
* <chr>              <int>         <dbl>      <dbl>      <dbl> <dbl>    <dbl>    <dbl>     <dbl>    <dbl> <chr>
1 listing_id             0             1 28628629.   8933420.  35303 23534981 29663636 34791234. 42573746 ▁▂▃▇▅
2 minimum_nights         0             1       3.41       9.35     1        1        1         2      365 ▇▁▁▁▁
3 maximum_nights         0             1     742.       472.       1      360     1125      1125     3000 ▆▇▁▁▁
```

- Retrieved 2 datasets from Inside Airbnb: Tokyo
  - Calendar data from March 2019 and February 2020 (~ 1 year)
  - Overlapping period March 25 to December 31
- Using skimr::skim() on the merged calendar data before filtering
  - A total of 9,241,800 observations (rows) and 7 variables

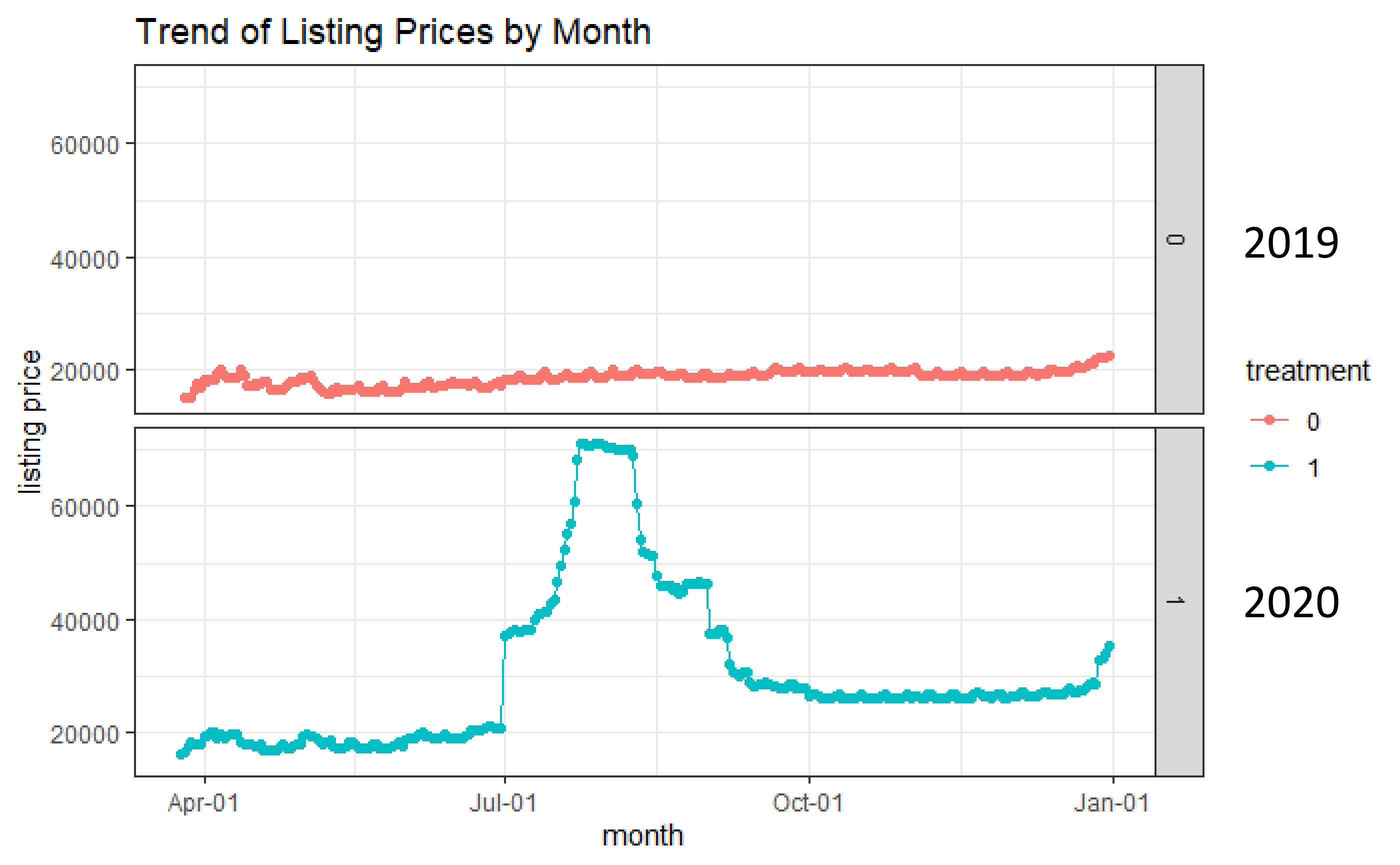
Dataset – Cleaning

- Data cleaning:
  - Add treatment group as 0 and 1
  - lubridate()
  - price format from str to int (and removing all the “,”)
    - One listing was ¥1,000,000.00, equivalent to $12609.09.
  - drop_na (1305 with missing prices)
  - group_by(listing_id, treatment, as.numeric(format(date, "%j“)))
  - summarize (mean_price = mean(adjusted_price)
- Trick: Run data analysis with a super computer or occupy yourself during downtime

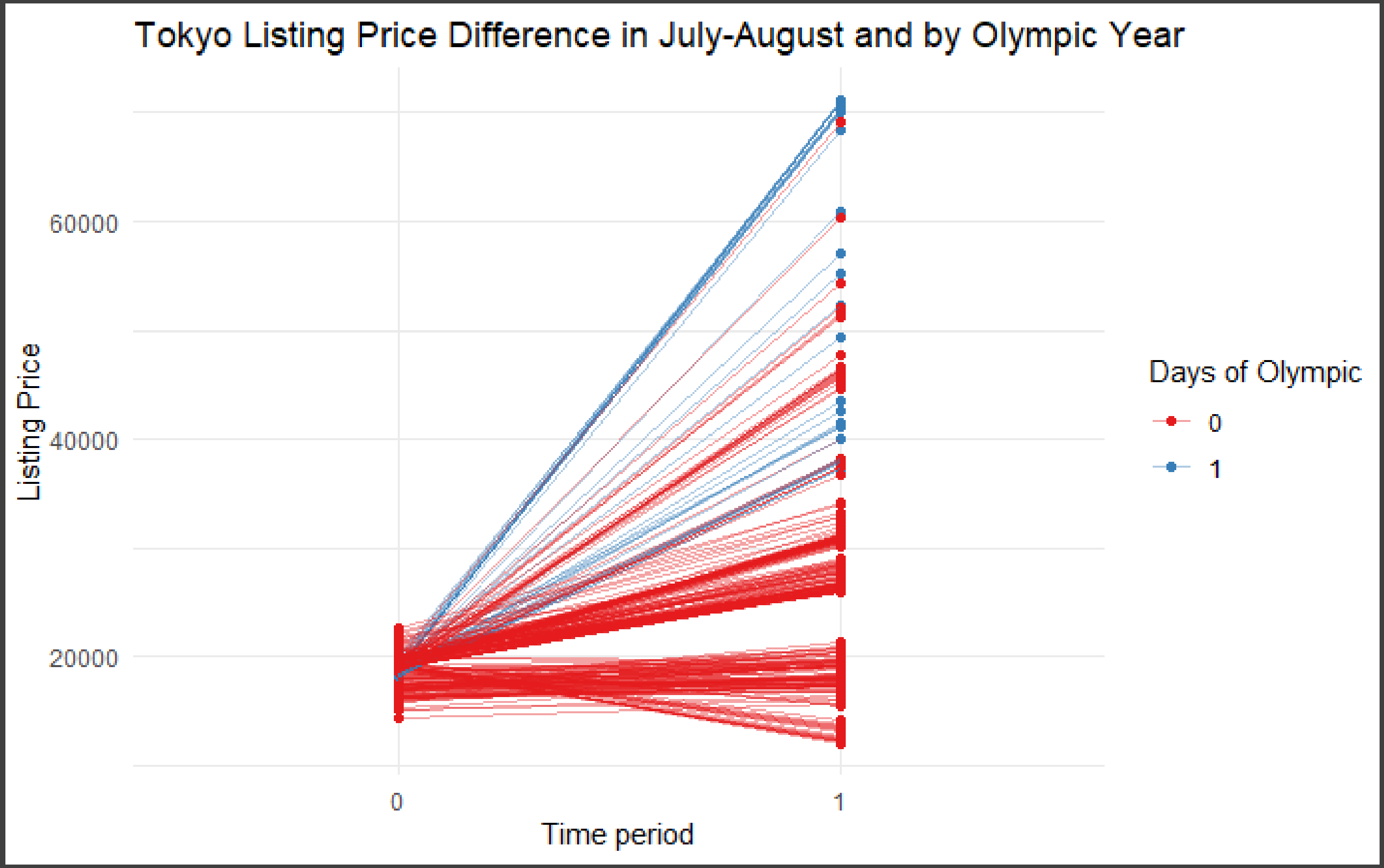
Dataset – ggplot()



~~~
```{r echo=FALSE}
###   Average Price   ###
##### By month ####

tokyo_calendar %>%
  filter(day_of_year >= 85) %>%
  ggplot(aes(x = no_year,
             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()
```
~~~

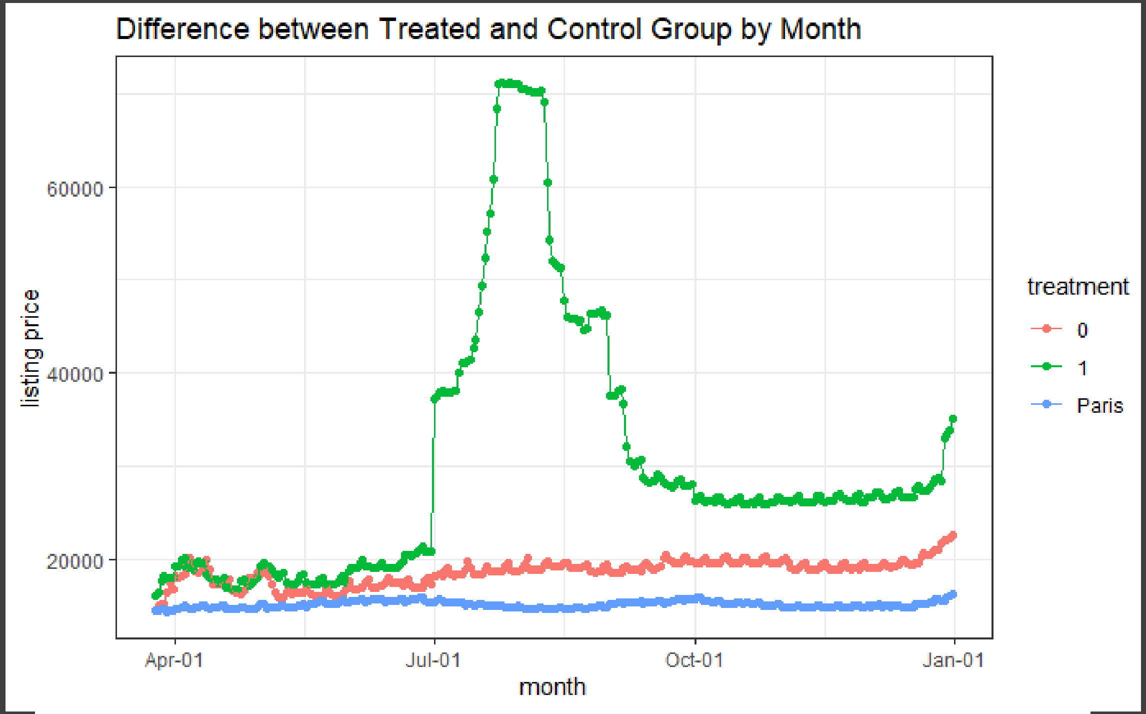


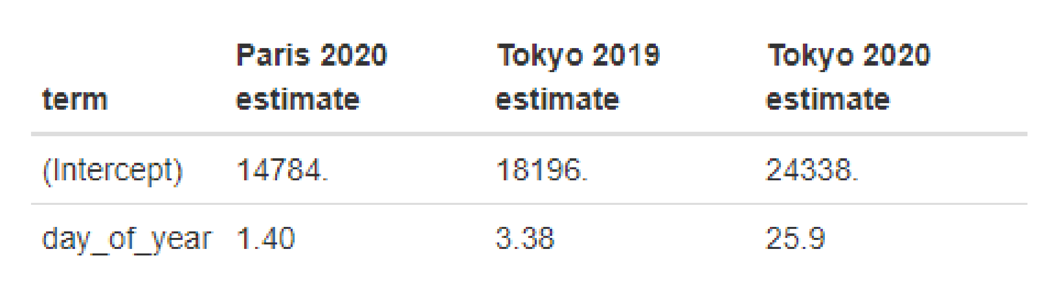
Analysis 1 – Difference-in-Difference[3]



- Compare differences of average listing prices of 2019 and 2020 during Olympic days
  - estimate for the average price difference between July to August in 2020 (blue) and July to August in 2019 (red)
  - ¥ 29,126.01 ($366.78).

[3] "Difference in differences," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Difference_in_differences. [Accessed 8 July 2020].





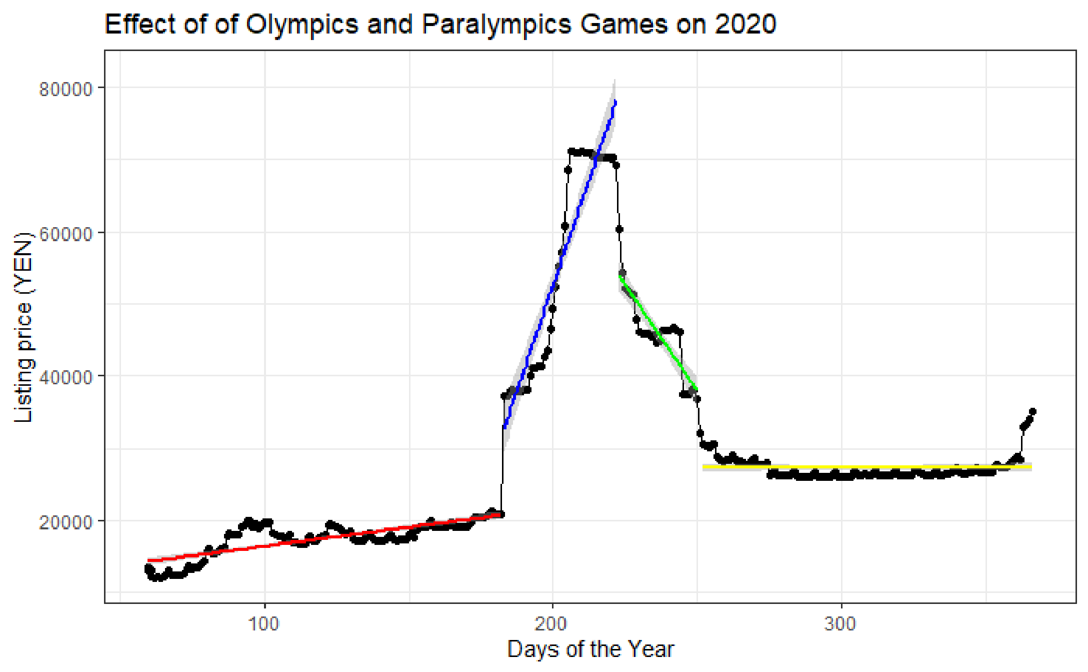
term	Paris 2020 estimate	Tokyo 2019 estimate	Tokyo 2020 estimate
(Intercept)	14784.	18196.	24338.
day_of_year	1.40	3.38	25.9

Analysis 1 – Difference-in-Difference

- 2024 Paris Olympics
- Further compare with average listing prices with Paris 2020 (next Olympic games host)
- Intercept and slope for Paris 2020 are lower than Tokyo 2019 and much lower than Tokyo 2020
- Trend of listing price in Paris for 2020 (blue) is more similar to 2019 Tokyo (red) than 2020 Tokyo (green)

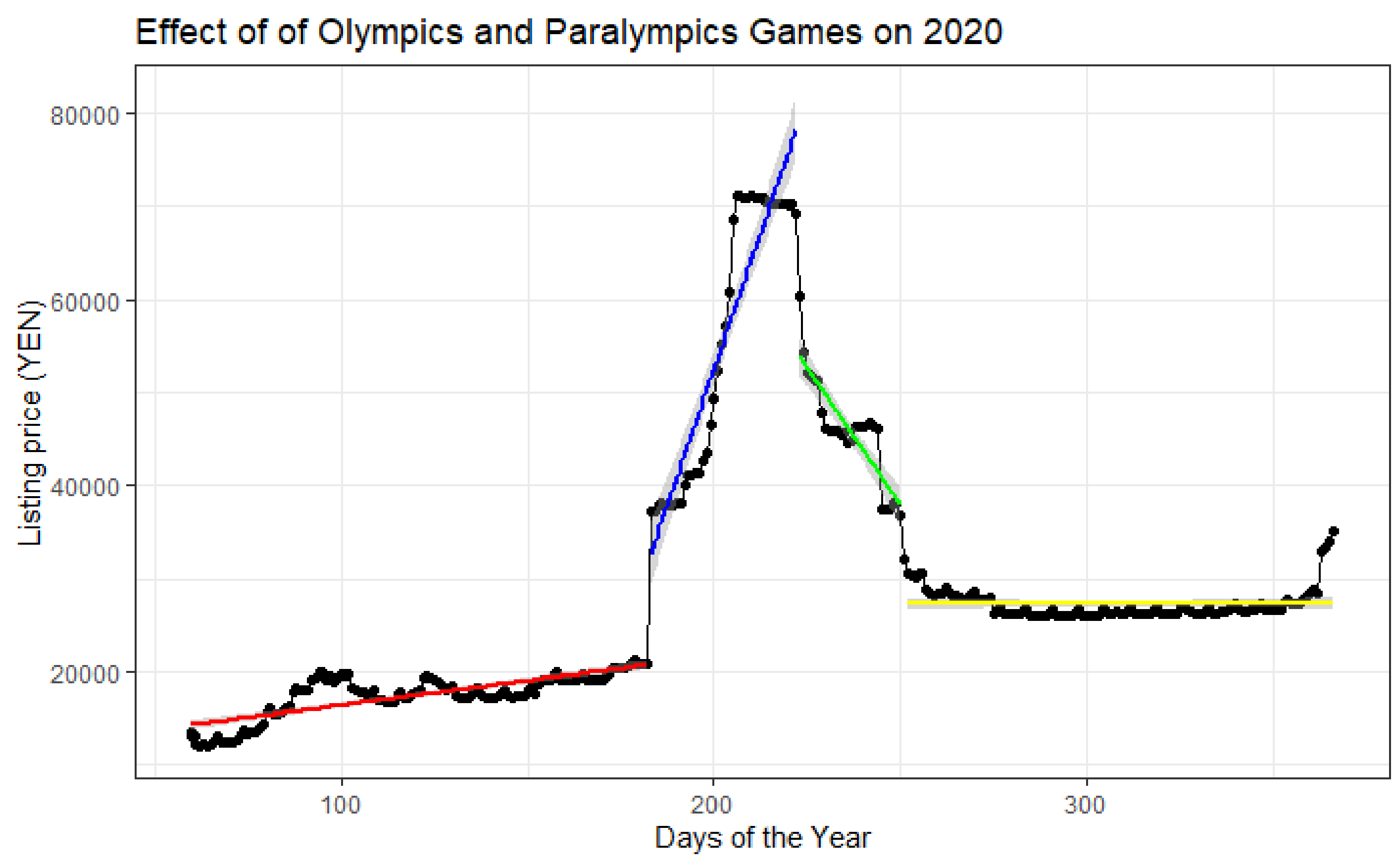
Analysis 2 – Regression Discontinuity[4]

- Focus on the Feb 2020 calendar data
- Multiple cuts of discontinuity in 2020 price trend
- 3 Methods of Regression Discontinuity in the Year 2020
    - 3 Cut-offs and 4 Periods
    - 3 Cut-offs and 3 Periods
    - 1 cut-off and 2 Periods



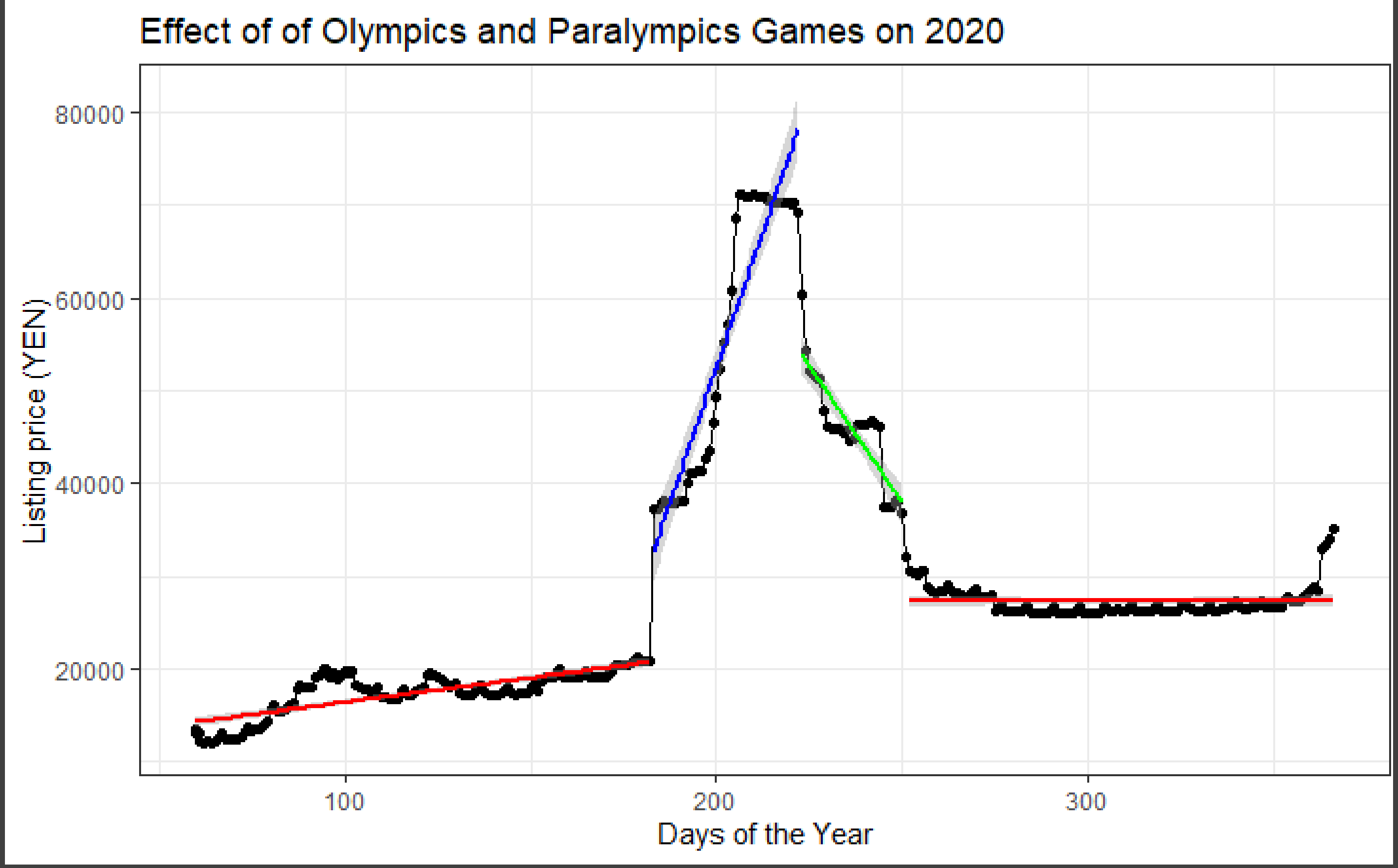
[4] "Regression discontinuity design," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Regression_discontinuity_design. [Accessed 8 July 2020].

Regression Discontinuity Method 1



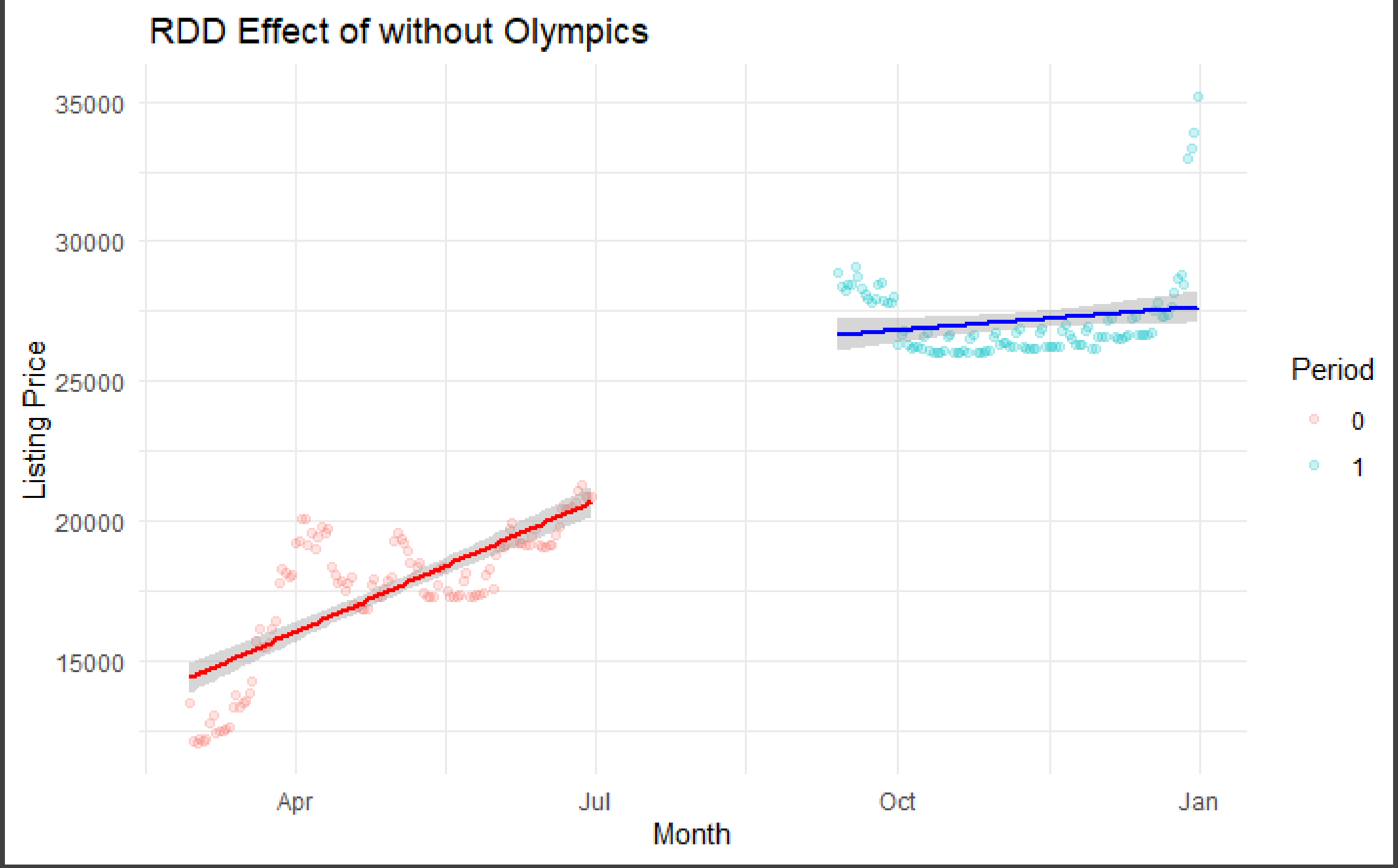
- 3 Cut-offs and 4 Periods
  - Before Day 183 – before the month of the Olympic Games
    - ↑ ¥ 52.7 ($0.66)
  - Day 183 to 222 – during the Olympic events
    - ↑ ¥1,165 ($14.66)
  - Day 223 to 250 – during the Paralympic events
    - ↓ ¥589 ($7.41)
  - Day 251 onwards – after the Olympic Games
    - ↑ ¥16 ($0.20)

Regression Discontinuity Method 2



- 3 Cut-offs and 3 Periods
  - no Olympics (rest of the year)
    - ↑ ¥52 ($0.66)
  - Olympics to no Olympics
    - ↑ ¥ 33,731 ($424.59)
  - Paralympics to no Olympics
    - ↑ ¥22,295 ($280.64)

Regression Discontinuity Method 3



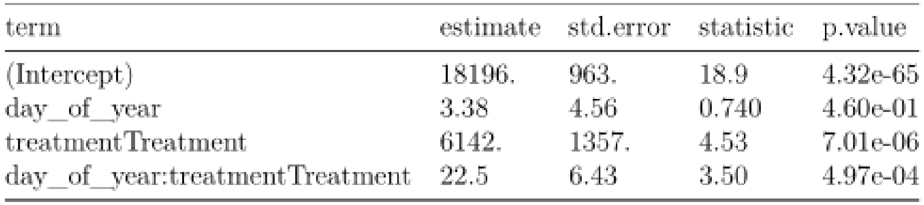
- 1 cut-off and 2 Periods
  - Period 0 – before July 1
  - Period 1 – after Sept 3
- 2020: ↑¥3,147 ($39.61)
- 2019: ↑¥2,226 ($28.02)
- Difference ¥921 ($11.59)

Linear Regression Models & Weakness

- 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.
- Improvement on RDD weakness
  - multiple slopes across the periods
  - different thresholds
  - different methods of "dissecting" the period

Bivariant LRM:

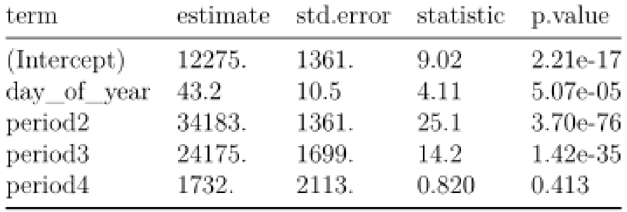
2019 and 2020



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

RDD Method 1:

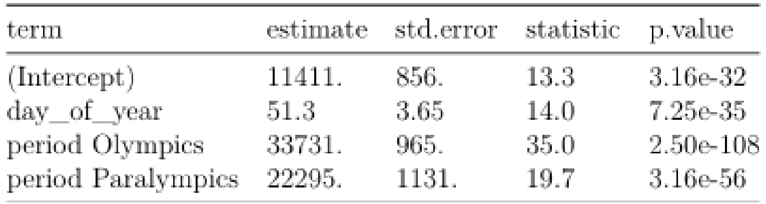
3 Cut-offs and 4 Periods



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:

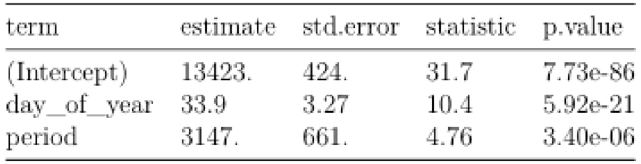
3 Cut-offs and 3 Periods



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

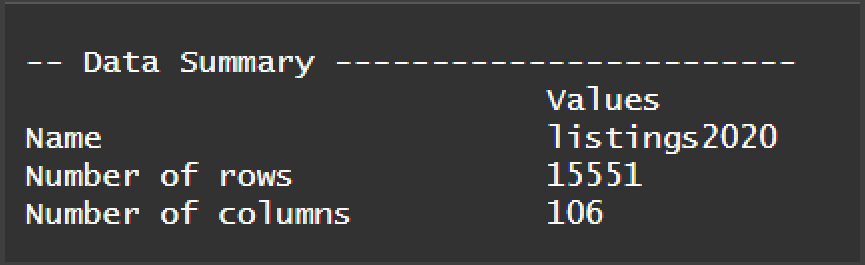
RDD Method 3:

1 Cut-offs and 2 Periods



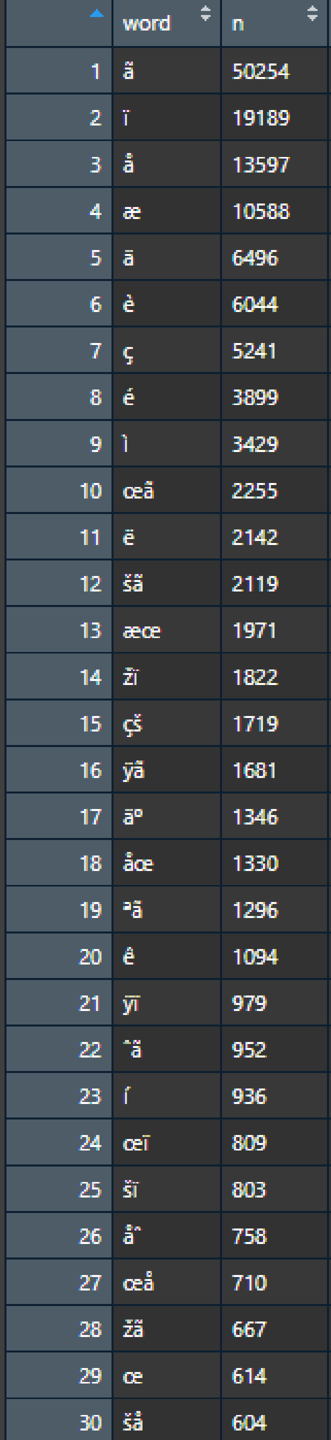
term	estimate	std.error	statistic	p.value
(Intercept)	13423.	424.	31.7	7.73e-86
day_of_year	33.9	3.27	10.4	5.92e-21
period	3147.	661.	4.76	3.40e-06

Text Analysis



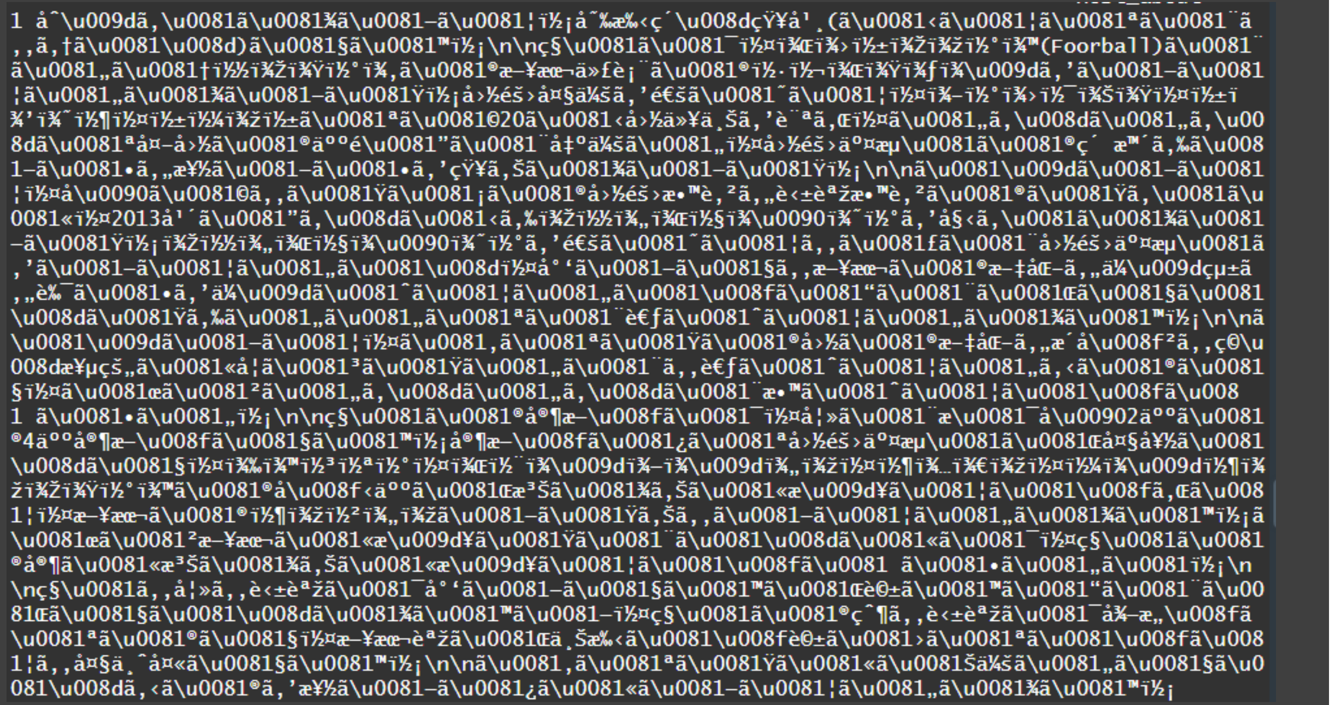
```
-- Data Summary ------------------------
                           Values
Name                       listings2020
Number of rows             15551
Number of columns          106
```

- Tokyo 2020 Listing data: 15,551 observations and 106 variables
- Merged listing data with my summarized calendar data by listing_id
  - Before filter: 11,352,230 observations and 114 variables
  - Host bio, host since, neighbourhood
  - Types of hosts adjust their price during Olympic period
  - tidytext::unnest_tokens()
- Japanese characters became unrecognized letters
  - Cannot just filter English letters and numbers



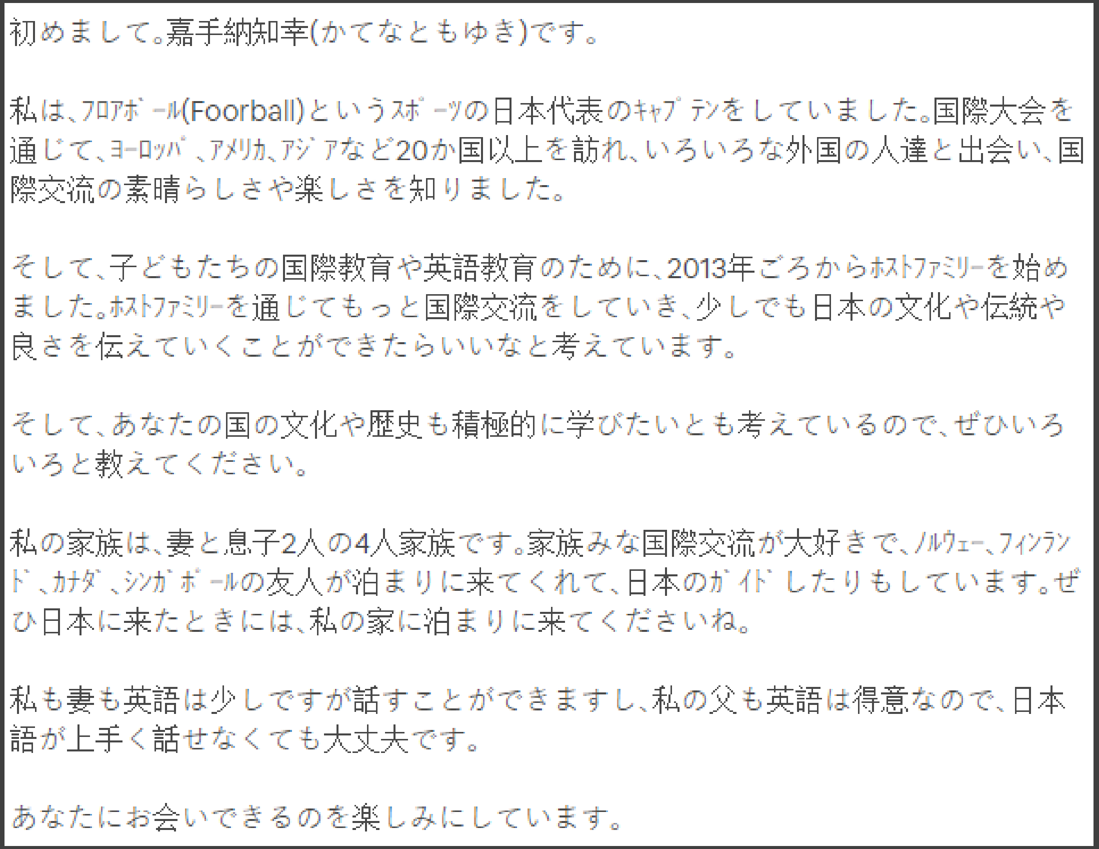
	word	n
1	ã	50254
2	ï	19189
3	å	13597
4	æ	10588
5	ä	6496
6	è	6044
7	ç	5241
8	é	3899
9	ì	3429
10	œã	2255
11	ë	2142
12	šã	2119
13	æœ	1971
14	žï	1822
15	çš	1719
16	ÿã	1681
17	ãº	1346
18	åœ	1330
19	ªã	1296
20	ê	1094
21	ÿï	979
22	ˆã	952
23	í	936
24	œï	809
25	šï	803
26	åˆ	758
27	œå	710
28	žã	667
29	œ	614
30	šå	604

Issue with Text Analysis on host_bio



```
1 å^\u009dã,\u0081ã\u0081¾ã\u0081—ã\u0081¦ï½¡å˜‰æ‰‹ç´\u008dçŸ¥å¹¸(ã\u0081‹ã\u0081¦ã\u0081ªã\u0081¨ã
,,ã,†ã\u0081\u008d)ã\u0081§ã\u0081™ï½¡\n\nç§\u0081ã\u0081¯ï½¤ï¾Œï¾›ï½±ï¾Žï¾žï½°ï¾™(Foorball)ã\u0081¨
ã\u0081„ã\u0081†ï½½ï¾Žï¾Ÿï½°ï¾‚ã\u0081®æ—¥æœ¬ä»£è¡¨ã\u0081®ï½·ï½¬ï¾Œï¾Ÿï¾ƒï¾\u009dã,’ã\u0081—ã\u0081
¦ã\u0081„ã\u0081¾ã\u0081—ã\u0081Ÿï½¡å›½éš›å¤§ä¼šã,’é€šã\u0081˜ã\u0081¦ï½¤ï¾–ï½°ï¾›ï½¯ï¾Šï¾Ÿï½¤ï½±ï
¾’ï¾˜ï½¶ï½¤ï½±ï½¼ï¾žï½±ã\u0081ªã\u0081©20ã\u0081‹å›½ä»¥ä¸Šã,’è¨ªã,Œï½¤ã\u0081„ã,\u008dã\u0081„ã,\u00
8dã\u0081ªå¤–å›½ã\u0081®äººé\u0081”ã\u0081¨å‡ºä¼šã\u0081„ï½¤å›½éš›äº¤æµ\u0081ã\u0081®ç´ æ™´ã,‰ã\u008
1—ã\u0081•ã,„æ¥½ã\u0081—ã\u0081•ã,’çŸ¥ã,Šã\u0081¾ã\u0081—ã\u0081Ÿï½¡\n\nã\u0081\u009dã\u0081—ã\u0081
¦ï½¤å\u0090ã\u0081©ã,,ã\u0081Ÿã\u0081¡ã\u0081®å›½éš›æ•™è‚²ã,„è‹±èªžæ•™è‚²ã\u0081®ã\u0081Ÿã,\u0081ã\u
0081«ï½¤2013å¹´ã\u0081”ã,\u008dã\u0081‹ã,‰ï¾Žï½½ï¾„ï¾Œï½§ï¾\u0090ï¾˜ï½°ã,’å§‹ã,\u0081ã\u0081¾ã\u0081
—ã\u0081Ÿï½¡ï¾Žï½½ï¾„ï¾Œï½§ï¾\u0090ï¾˜ï½°ã,’é€šã\u0081˜ã\u0081¦ã,,ã\u0081£ã\u0081¨å›½éš›äº¤æµ\u0081ã
,’ã\u0081—ã\u0081¦ã\u0081„ã\u0081\u008dï½¤å°‘ã\u0081—ã\u0081§ã,,æ—¥æœ¬ã\u0081®æ–‡åŒ–ã,„ä¼\u009dçµ±ã
,„è‰¯ã\u0081•ã,’ä¼\u009dã\u0081ˆã\u0081¦ã\u0081„ã\u0081\u008fã\u0081“ã\u0081¨ã\u0081Œã\u0081§ã\u0081
\u008dã\u0081Ÿã,‰ã\u0081„ã\u0081„ã\u0081ªã\u0081¨è€ƒã\u0081ˆã\u0081¦ã\u0081„ã\u0081¾ã\u0081™ï½¡\n\nã
\u0081\u009dã\u0081—ã\u0081¦ï½¤ã\u0081,ã\u0081ªã\u0081Ÿã\u0081®å›½ã\u0081®æ–‡åŒ–ã,„æ´å\u008f²ã,,ç©\u
008dæ¥µçš„ã\u0081«å¦ã\u0081³ã\u0081Ÿã\u0081„ã\u0081¨ã,,è€ƒã\u0081ˆã\u0081¦ã\u0081„ã,‹ã\u0081®ã\u0081
§ï½¤ã\u0081œã\u0081²ã\u0081„ã,\u008dã\u0081„ã,\u008dã\u0081¨æ•™ã\u0081ˆã\u0081¦ã\u0081\u008fã\u008
1 ã\u0081•ã\u0081„ï½¡\n\nç§\u0081ã\u0081®å®¶æ—\u008fã\u0081¯ï½¤å¦»ã\u0081¨æ\u0081¯å¦\u00902äººã\u0081
®4äººå®¶æ—\u008fã\u0081§ã\u0081™ï½¡å®¶æ—\u008fã\u0081¿ã\u0081ªå›½éš›äº¤æµ\u0081ã\u0081Œå¤§å¥½ã\u0081
\u008dã\u0081§ï½¤ï¾‰ï¾™ï½³ï½ªï½°ï½¤ï¾Œï½¨ï¾\u009dï¾—ï¾\u009dï¾„ï¾žï½¤ï½¶ï¾…ï¾€ï¾žï½¤ï½¼ï¾\u009dï½¶ï¾
žï¾Žï¾Ÿï½°ï¾™ã\u0081®å\u008f‹äººã\u0081Œæ³Šã\u0081¾ã,Šã\u0081«æ\u009d¥ã\u0081¦ã\u0081\u008fã,Œã\u008
1¦ï½¤æ—¥æœ¬ã\u0081®ï½¶ï¾žï½²ï¾„ï¾žã\u0081—ã\u0081Ÿã,Šã,,ã\u0081—ã\u0081¦ã\u0081„ã\u0081¾ã\u0081™ï½¡ã
\u0081œã\u0081²æ—¥æœ¬ã\u0081«æ\u009d¥ã\u0081Ÿã\u0081¨ã\u0081\u008dã\u0081«ã\u0081¯ï½¤ç§\u0081ã\u0081
®å®¶ã\u0081«æ³Šã\u0081¾ã,Šã\u0081«æ\u009d¥ã\u0081¦ã\u0081\u008fã\u0081 ã\u0081•ã\u0081„ã\u0081ï½¡\n
\nç§\u0081ã,,å¦»ã,,è‹±èªžã\u0081¯å°‘ã\u0081—ã\u0081§ã\u0081™ã\u0081Œè©±ã\u0081™ã\u0081“ã\u0081¨ã\u00
81Œã\u0081§ã\u0081\u008dã\u0081¾ã\u0081™ã\u0081—ï½¤ç§\u0081ã\u0081®ç^¶ã,,è‹±èªžã\u0081¯å¾—æ„\u008fã
\u0081ªã\u0081®ã\u0081§ï½¤æ—¥æœ¬èªžã\u0081Œä¸Šæ‰‹ã\u0081\u008fè©±ã\u0081›ã\u0081ªã\u0081\u008fã\u008
1¦ã,,å¤§ä¸^å¤«ã\u0081§ã\u0081™ï½¡\n\nã\u0081,ã\u0081ªã\u0081Ÿã\u0081«ã\u0081Šä¼šã\u0081„ã\u0081§ã\u0
081\u008dã,‹ã\u0081®ã,’æ¥½ã\u0081—ã\u0081¿ã\u0081«ã\u0081—ã\u0081¦ã\u0081„ã\u0081¾ã\u0081™ï½¡
```

Bio from Inside Airbnb file



初めまして。嘉手納知幸(かてなともゆき)です。

私は、ﾌﾛｱﾎﾞｰﾙ(Foorball)というｽﾎﾟｰﾂの日本代表のｷｬﾌﾟﾃﾝをしていました。国際大会を通じて、ﾖｰﾛｯﾊﾟ、ｱﾒﾘｶ、ｱｼﾞｱなど20か国以上を訪れ、いろいろな外国の人達と出会い、国際交流の素晴らしさや楽しさを知りました。

そして、子どもたちの国際教育や英語教育のために、2013年ごろからﾎｽﾄﾌｧﾐﾘｰを始めました。ﾎｽﾄﾌｧﾐﾘｰを通じてもっと国際交流をしていき、少しでも日本の文化や伝統や良さを伝えていくことができたらいいなと考えています。

そして、あなたの国の文化や歴史も積極的に学びたいとも考えているので、ぜひいろいろと教えてください。

私の家族は、妻と息子2人の4人家族です。家族みな国際交流が大好きで、ﾉﾙｳｪｰ、ﾌｨﾝﾗﾝﾄﾞ、ｶﾅﾀﾞ、ｼﾝｶﾞﾎﾟｰﾙの友人が泊まりに来てくれて、日本のｶﾞｲﾄﾞしたりもしています。ぜひ日本に来たときには、私の家に泊まりに来てくださいね。

私も妻も英語は少しですが話すことができますし、私の父も英語は得意なので、日本語が上手く話せなくても大丈夫です。

あなたにお会いできるのを楽しみにしています。

Bio from Airbnb

Data Science Opportunities

- Kaggle Competition
- Airbnb recruitment in 2016: Airbnb New User Bookings
  - Predict in which country a new user will make his or her first booking
- Topics [4]:
  - Occupancy Model
  - Demand and Price Analysis
  - Sentiment analysis on Reviews
  - Spatial Data Analysis
  - What makes a superhost?

[5] A. Peshin, S. Gupta and A. Ankita, "Exploratory Data Analysis and Visualization of Airbnb Dataset," Columbia University, 10 December 2018. [Online]. Available: http://www.columbia.edu/~sg3637/airbnb_final_analysis.html. [Accessed 8 July 2020].

Personal Lessons

- Airbnb dataset provides a fantastic source to better understand human pattern in a hosting landscape across the world.
  - Interaction between people and between information
  - Ethics considerations and limitation
- Find data projects that you find interesting
  - Inspired from experience, from current events, from media
  - You learn more about data during EDA
- Expect slower computer performance for large data when using R
  - More CPU & RAM, faster analytical speed
  - Filter and output to a small dataset

Reference

[1] M. Cox, "About Inside Airbnb," Inside Airbnb, [Online]. Available: http://insideairbnb.com/.

[2] IOC, "IOC and Airbnb announce major global Olympic partnership," Olympic News, 19 April 2020. [Online]. Available: https://www.olympic.org/news/ioc-and-airbnb-announce-major-global-olympic-partnership. [Accessed 8 July 2020].

[3] "Difference in differences," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Difference_in_differences. [Accessed 8 July 2020].

[4] "Regression discontinuity design," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Regression_discontinuity_design. [Accessed 8 July 2020].

[5] A. Peshin, S. Gupta and A. Ankita, "Exploratory Data Analysis and Visualization of Airbnb Dataset," Columbia University, 10 December 2018. [Online]. Available: http://www.columbia.edu/~sg3637/airbnb_final_analysis.html. [Accessed 8 July 2020].