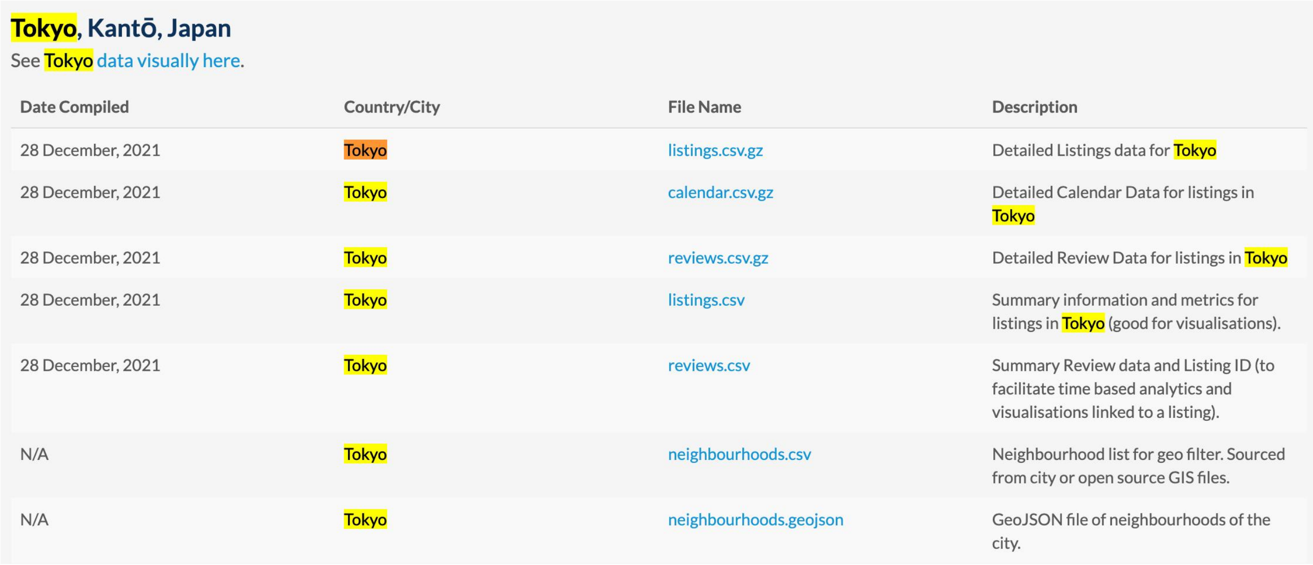
project 01: Write a Data Science Blog Post

01 choose a dataset of myself

i choosed a Tokyo AirBNB data from the following link.

http://insideairbnb.com/get-the-data.html

the screen capture is also attached here. including some listings, calendar and reviews data.



Tokyo, Kantō, Japan

See Tokyo data visually here.

Date Compiled	Country/City	File Name	Description
28 December, 2021	Tokyo	listings.csv.gz	Detailed Listings data for Tokyo
28 December, 2021	Tokyo	calendar.csv.gz	Detailed Calendar Data for listings in Tokyo
28 December, 2021	Tokyo	reviews.csv.gz	Detailed Review Data for listings in Tokyo
28 December, 2021	Tokyo	listings.csv	Summary information and metrics for listings in Tokyo (good for visualisations).
28 December, 2021	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.

02 Key Steps for Project

02.01 data and github url

downloaded the csv files for listings, calendar and reviews from above dataset. and included in the following github project:

https://github.com/allhanz/data_scientist_nanodegree.git

subproject name: 01_project_01

02.02 several issues for data analysis

1) how many host are avaliable in the tokyo airBNB market japan?
2) is there any relationship between price and the datetime?
3) which area have the most listing count and may be the most popular for customers?
4) what kinds of host are listed in the AirBNB tokyo market and how many lists per host?
5) how the price for different room type?
6) which area have the most review count number?

02.03 data analysis details by creating jupyter notebook

02.03.01 requirements lib for data analysis

pandas
numpy
plotly
datatable
matlibplot

02.03.02 some insight for data analysis

imaging i am a host who will decide to put some houses into the tokyo AirBNB market. i'd better to understand how the market situation. for example, which location, which area, what kind of rom etc will be the popular to the local customers and for Foreign customers. with these kinds of questions, to check is there any answer infotmation included in this tokyo AirBNB dataset via data analysis.

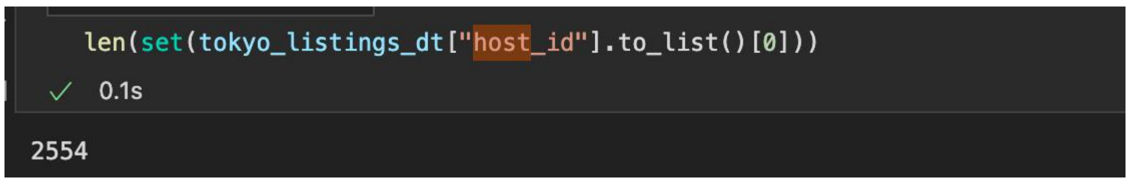
02.03.03 data analysis details

1) to check how many host are avaliable in tokyo AirBNB market and what kinds of host are listed in the AirBNB tokyo market and how many lists per host?

to count the number of data items included in the "host_id" column in the listings dataset and also calculate the percentage for each host abou how many list are avaiabler.

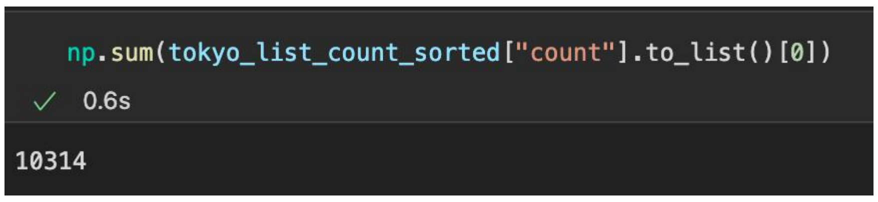
01. the host count numer for tokyo AirBNB market

2554 hosts have totally 10314 lists(houses) in tokyo AirBNB market.



```
len(set(tokyo_listings_dt["host_id"].to_list()[0]))
```

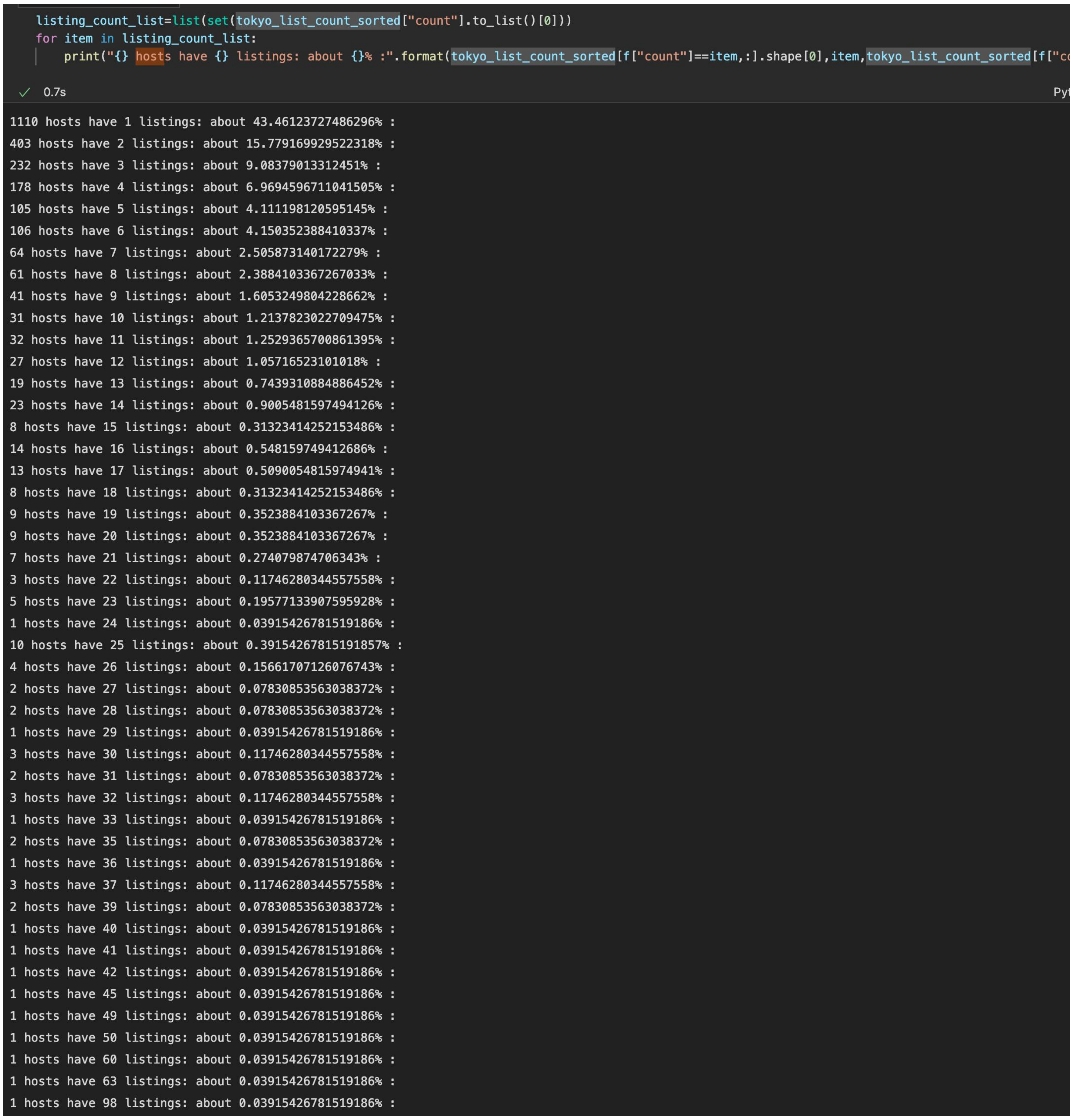
```
2554
```



```
np.sum(tokyo_list_count_sorted["count"].to_list()[0])
```

```
10314
```

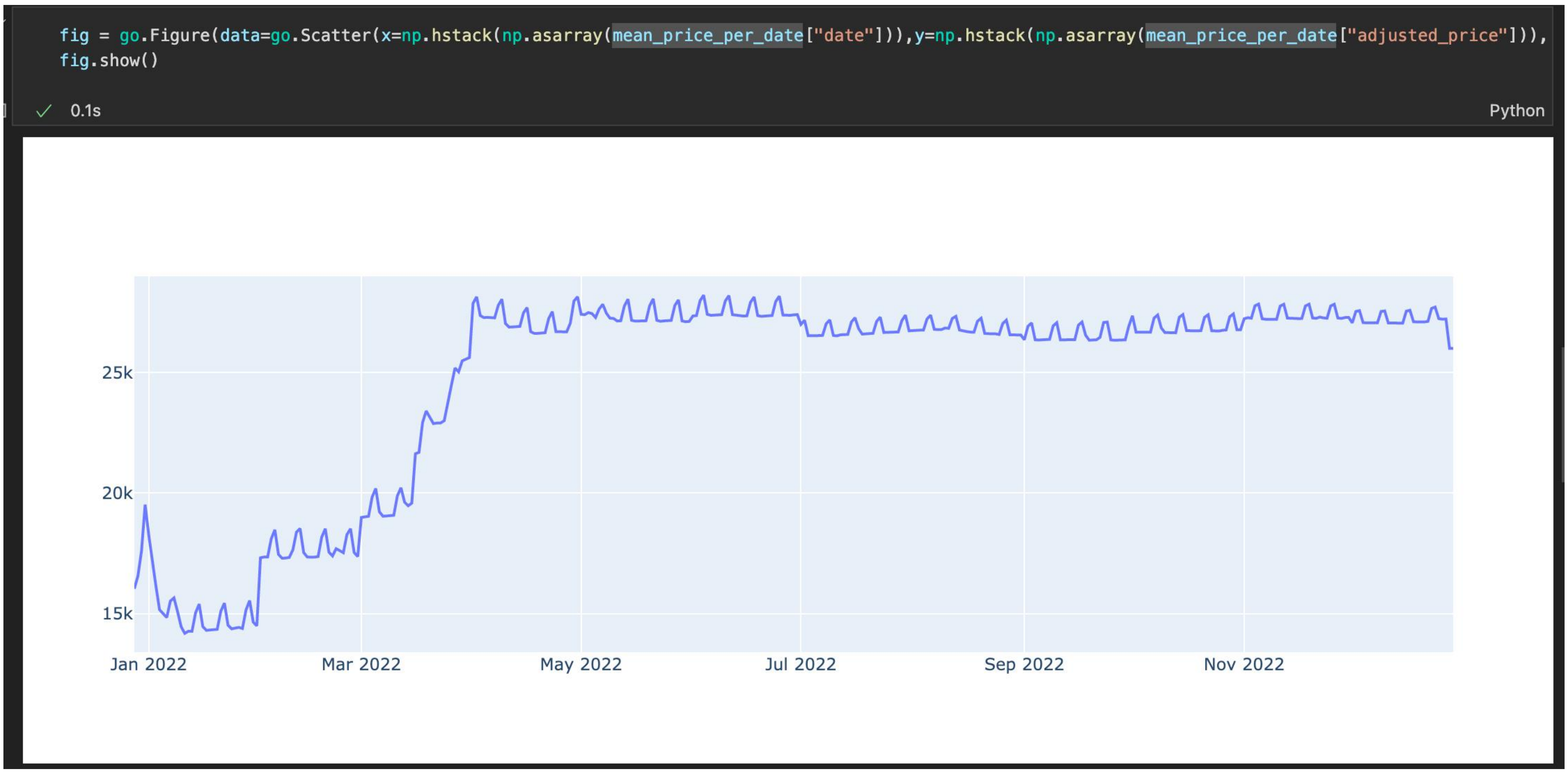
02. the percentage information for how many lists are haved by each kind of host
we can see that around 70% hosts have less than 3 lists(houses) in the tokyo AirBNB market.



```
listing_count_list=list(set(tokyo_list_count_sorted["count"].to_list()[0]))
for item in listing_count_list:
    print("{} hosts have {} listings: about {}% :".format(tokyo_list_count_sorted[f["count"]==item,:].shape[0],item,tokyo_list_count_sorted[f["c
```

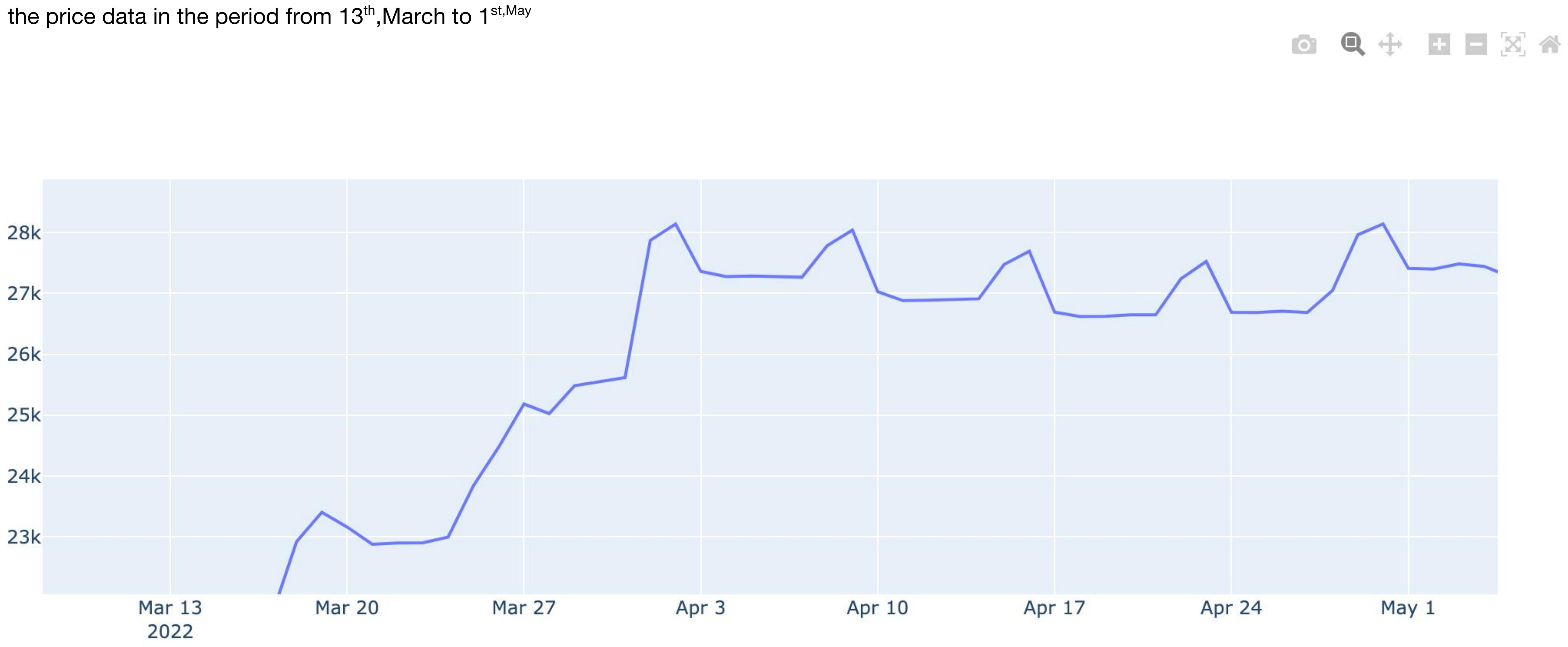
```
1110 hosts have 1 listings: about 43.46123727486296% :
403 hosts have 2 listings: about 15.779169929522318% :
232 hosts have 3 listings: about 9.08379013312451% :
178 hosts have 4 listings: about 6.9694596711041505% :
105 hosts have 5 listings: about 4.111198120595145% :
106 hosts have 6 listings: about 4.150352388410337% :
64 hosts have 7 listings: about 2.505873140172279% :
61 hosts have 8 listings: about 2.3884103367267033% :
41 hosts have 9 listings: about 1.6053249804228662% :
31 hosts have 10 listings: about 1.2137823022709475% :
32 hosts have 11 listings: about 1.2529365700861395% :
27 hosts have 12 listings: about 1.05716523101018% :
19 hosts have 13 listings: about 0.7439310884886452% :
23 hosts have 14 listings: about 0.9005481597494126% :
8 hosts have 15 listings: about 0.31323414252153486% :
14 hosts have 16 listings: about 0.548159749412686% :
13 hosts have 17 listings: about 0.5090054815974941% :
8 hosts have 18 listings: about 0.31323414252153486% :
9 hosts have 19 listings: about 0.3523884103367267% :
9 hosts have 20 listings: about 0.3523884103367267% :
7 hosts have 21 listings: about 0.274079874706343% :
3 hosts have 22 listings: about 0.11746280344557558% :
5 hosts have 23 listings: about 0.19577133907595928% :
1 hosts have 24 listings: about 0.03915426781519186% :
10 hosts have 25 listings: about 0.39154267815191857% :
4 hosts have 26 listings: about 0.15661707126076743% :
2 hosts have 27 listings: about 0.07830853563038372% :
2 hosts have 28 listings: about 0.07830853563038372% :
1 hosts have 29 listings: about 0.03915426781519186% :
3 hosts have 30 listings: about 0.11746280344557558% :
2 hosts have 31 listings: about 0.07830853563038372% :
3 hosts have 32 listings: about 0.11746280344557558% :
1 hosts have 33 listings: about 0.03915426781519186% :
2 hosts have 35 listings: about 0.07830853563038372% :
1 hosts have 36 listings: about 0.03915426781519186% :
3 hosts have 37 listings: about 0.11746280344557558% :
2 hosts have 39 listings: about 0.07830853563038372% :
1 hosts have 40 listings: about 0.03915426781519186% :
1 hosts have 41 listings: about 0.03915426781519186% :
1 hosts have 42 listings: about 0.03915426781519186% :
1 hosts have 45 listings: about 0.03915426781519186% :
1 hosts have 49 listings: about 0.03915426781519186% :
1 hosts have 50 listings: about 0.03915426781519186% :
1 hosts have 60 listings: about 0.03915426781519186% :
1 hosts have 63 listings: about 0.03915426781519186% :
1 hosts have 98 listings: about 0.03915426781519186% :
```

2) is there any relationship between price and the datetime?
by calculating the average price per day, we get the following figure data. we can see from around April the average price is becoming higher and higher, and between the April and december, the average of price in Saturday is the most expensive in one week and the price in Friday is the second. and the average price in January is the cheapest during one year, Feburay is the second, march is the third, and from Aprial the average price is returned to the normal level.



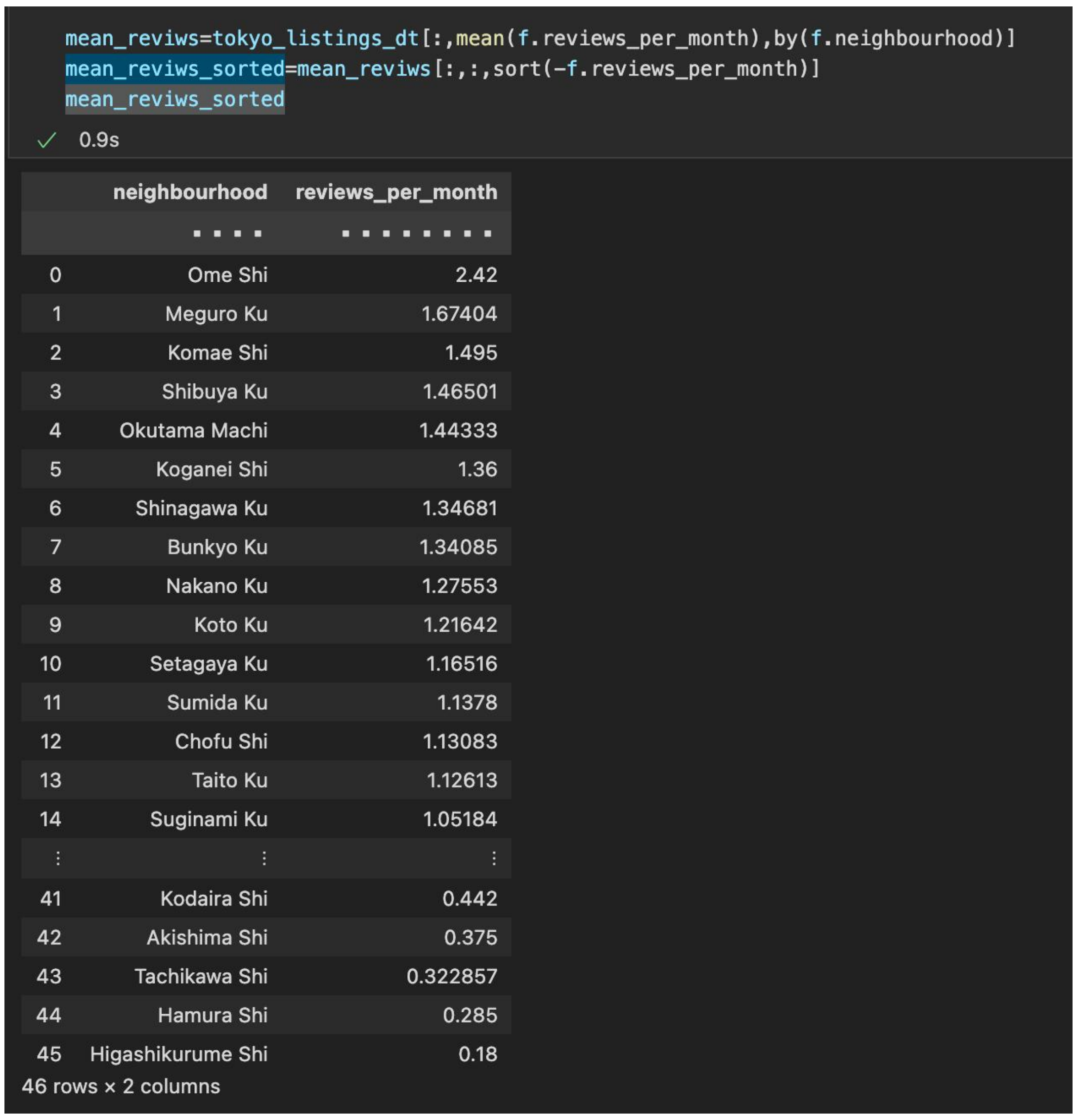
```
fig = go.Figure(data=go.Scatter(x=np.hstack(np.asarray(mean_price_per_date["date"])),y=np.hstack(np.asarray(mean_price_per_date["adjusted_price"])),
fig.show()
```

zoom in the price data in the period from 13th,March to 1st,May



3) which area have the most listing count and may be the most popular for customers

ranked the revies_per_mount count number for each neighbourhood and each area as following figure. and it was supriced me, the ume shi seems the most popular area for the customers because this area so far away from the shijyuku and shubuya ku. maybe the average price for each list is cheaper or some other reason? still unknow, need to research more in further.



```
mean_reviws=tokyo_listings_dt[:,mean(f.reviews_per_month),by(f.neighbourhood)]
mean_reviws_sorted=mean_reviws[:,:,sort(-f.reviews_per_month)]
mean_reviws_sorted
```

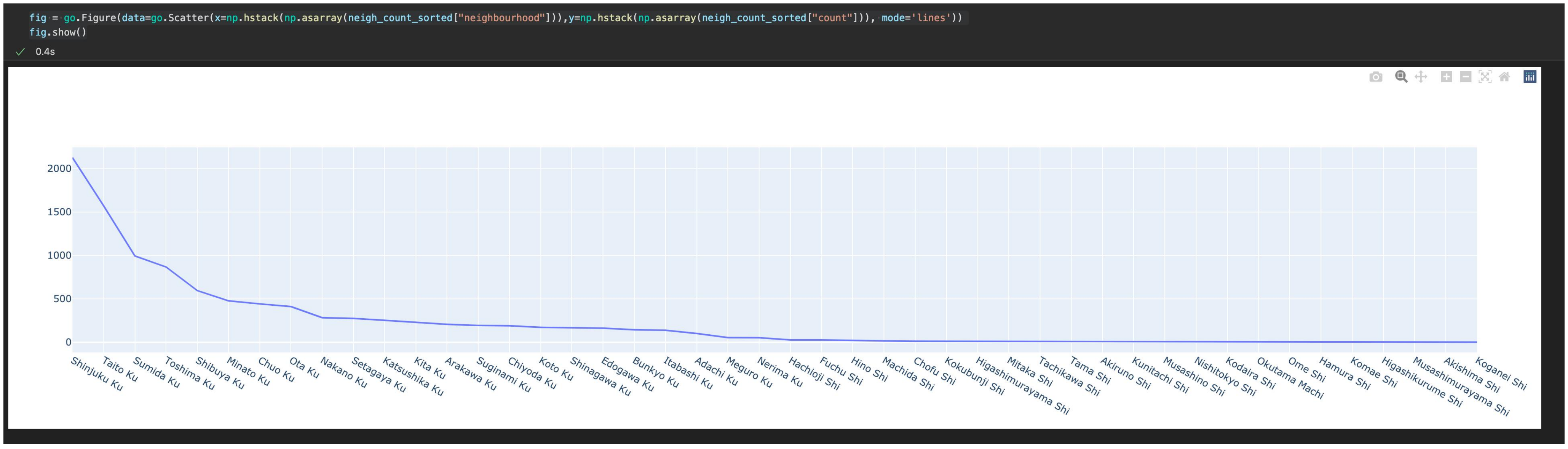
✓ 0.9s

	neighbourhood	reviews_per_month
	▪▪▪▪	▪▪▪▪▪▪▪▪
0	Ome Shi	2.42
1	Meguro Ku	1.67404
2	Komae Shi	1.495
3	Shibuya Ku	1.46501
4	Okutama Machi	1.44333
5	Koganei Shi	1.36
6	Shinagawa Ku	1.34681
7	Bunkyo Ku	1.34085
8	Nakano Ku	1.27553
9	Koto Ku	1.21642
10	Setagaya Ku	1.16516
11	Sumida Ku	1.1378
12	Chofu Shi	1.13083
13	Taito Ku	1.12613
14	Suginami Ku	1.05184
⋮	⋮	⋮
41	Kodaira Shi	0.442
42	Akishima Shi	0.375
43	Tachikawa Shi	0.322857
44	Hamura Shi	0.285
45	Higashikurume Shi	0.18

46 rows × 2 columns

4) which area have the most review count number?

here is the list(house) ranking for each area. we can see shinjyuku ku has the most lists(house). shibuya ku is only in 4th place. but taito ku is in the second place, maybe it is more closer than other area to the narita international Airpot.

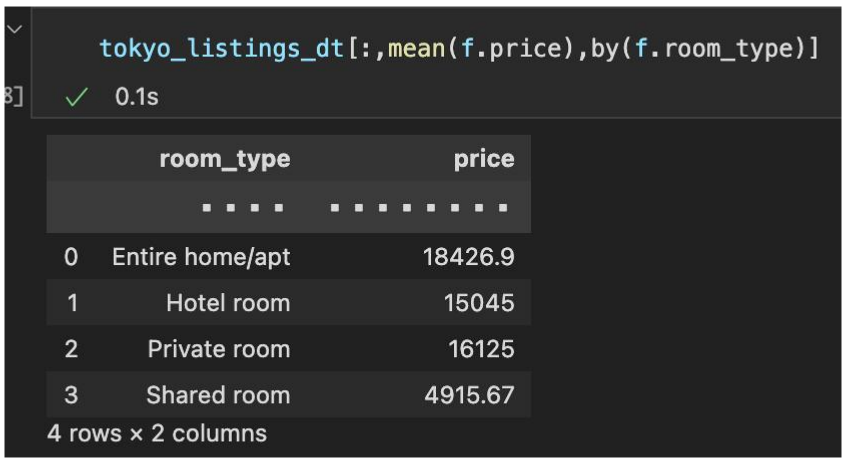


```
fig = go.Figure(data=go.Scatter(x=np.hstack(np.asarray(neigh_count_sorted["neighbourhood"])),y=np.hstack(np.asarray(neigh_count_sorted["count"])), mode='lines'))
fig.show()
```

✓ 0.4s

5) how the price for different room type?

the following are the average price for each type room. we can see that the entire room is the most expensive, private room is in the second place, hotel room is the 3rd. and of course the shared room is the cheapest. this is the same as my imagination.



```
tokyo_listings_dt[:,mean(f.price),by(f.room_type)]
```

8] ✓ 0.1s

	room_type	price
	▪▪▪▪	▪▪▪▪▪▪▪▪
0	Entire home/apt	18426.9
1	Hotel room	15045
2	Private room	16125
3	Shared room	4915.67

4 rows × 2 columns

for more details, you can check in the notebook file which is “data_analysis.ipynb”