



Shahid Beheshti
University

دانشگاه شهید بهشتی
دانشکده علوم ریاضی
گروه علوم کامپیوتر

گزارش تمرین های سری اول
تمرین اول (دیتاست خانه های اجاره ای)
درس داده کاوی

اساتید محترم:

جناب آقای دکتر هادی فراهانی و جناب آقای دکتر سعید رضا خردپیشه

آموزشیار محترم: جناب آقای علی شریفی

جواد تدین ۹۹۴۲۲۰۴۱


بهار ۱۴۰۰

به نام خدا

در این تمرین هدف تقویت توانایی در استفاده از تحلیل ها است و استنتاج های اماری.

این تمرین در مورد داده های جمع اوری شده از خانه های اجاره ای برای اقامت کوتاه مدت در شهر نیویورک امریکا است و در آن اطلاعاتی در مورد میزبان ها مهمان ها مکان اقامت و .. گفته شده است.

حال با توجه به دیتا ست موجود با خواندن دیتا ست توسط یک فایل CSV به بررسی سولات خواسته شده می پردازیم.

Jupyter Airbnb-Tadayon-99422041 Last Checkpoint: Last Thursday at 10:01 AM (autosaved)  Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

In [25]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sea
from scipy import stats
import urllib
```

In [26]:

```
data_set = pd.read_csv('AB_NYC_2019.csv')
data_set.head(20)
```

Out[26]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	
2	3647	THE VILLAGE OF HARLEM... NEW YORK!	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	
		Entire Apt.										

در ابتدا داده ها و فیچرها را مورد بررسی قرار می دهیم.

مجموعه داده ای که در اینجا در نظر گرفته ایم یک لیست از خانه های اجاره ای موجود است به علاوه مقداری از ویژگی هارا هم بررسی کرده ایم که این اطلاعات در جدول قابل مشاهده است.

jupyter Airbnb-Tadayon-99422041 Last Checkpoint: an hour ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

```
In [25]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sea
from scipy import stats
import urllib
```

```
In [26]: data_set = pd.read_csv('AB_NYC_2019.csv')
data_set.head(20)
```

Out[26]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	
2	3647	THE VILLAGE OF HARLEM....NEW YORK!	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	
5	5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hill	40.74767	-73.97500	Entire home/apt	200	3	
6	5121	BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesant	40.68688	-73.95596	Private room	60	45	
7	5178	Large Furnished Room Near B'way	8967	Shunichi	Manhattan	Hell's Kitchen	40.76489	-73.98493	Private room	79	2	
8	5203	Cozy Clean Guest Room - Family Apt	7490	MaryEllen	Manhattan	Upper West Side	40.80178	-73.96723	Private room	79	2	
9	5238	Cute & Cozy Lower East Side 1 bdrm	7549	Ben	Manhattan	Chinatown	40.71344	-73.99037	Entire home/apt	150	1	
10	5295	Beautiful 1br on Upper West Side	7702	Lena	Manhattan	Upper West Side	40.80316	-73.96545	Entire home/apt	135	5	
11	5441	Central Manhattan/near	7989	Kate	Manhattan	Hell's Kitchen	40.76076	-73.98867	Private room	85	2	

که سطر اول تا چهارم را مشاهده می کنیم و دید کلی نسبت به دیتا ست پیدا می کنیم.

تعدادی از ویژگی های مجموعه داده مثل id اسم خانه " منطقه ای که خانه در آن واقع شده است و... موجود است.

اول از همه ستون هایی که داده های عددی دارند را از داده های کتگوریکال جدا می کنیم. و شروع میکنیم پارامترهای اماری را مثل میانگین واریانس مینیمم و ماکزیمم و انحراف معیار را بدست می آوریم .

jupyter Airbnb-Tadayon-99422041 Last Checkpoint: Last Thursday at 10:01 AM (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

23	8110	MAISON DES SIRENES1.bohemian apartment	22486	Lisel	Brooklyn	Park Slope	40.68001	-73.97865	Private room	110	2
24	8490	SIRENES1.bohemian apartment	25183	Nathalie	Brooklyn	Bedford-Stuyvesant	40.68371	-73.94028	Entire home/apt	120	2

```
In [29]: data_set['price'].describe()
```

```
Out[29]: count    48884.000000
mean       152.755053
std        240.170260
min         10.000000
25%         69.000000
50%        106.000000
75%        175.000000
max        10000.000000
Name: price, dtype: float64
```

تا به اینجای کار مرحله اول ما تمام می شود و وارد مرحله دوم می شویم .

در مرحله دوم دادهای از دست رفته را مورد بررسی قرار می دهیم طبق مواردی که پیشتر آموختیم اگر دیتا ست ما داده ی (نال) وجود داشت از چند طریق میتوانیم مقادیر دیگری از جمله میانگین را جایگزین کنیم و دیگر در دیتا ست دیتای نال نداشته باشیم.

در ستون قیمت آپارتمان مشاهده می کنیم که قیمت صفر هم در دیتا ست موجود است و این به هیچ عنوان امکان پذیر نیست.

یعنی امکان ندارد که شخصی خانه خود را رایگان اجاره دهد

پس ستون هایی که در آن قیمت آپارتمان صفر گزارش شده است را حذف میکنیم با انجام این کار نوعی پاکسازی در دیتا ست انجام می دهیم

jupyter Airbnb-Tadayon-99422041 Last Checkpoint: Last Thursday at 10:01 AM (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

19	7750	Huge 2 BR Upper East Cental Park	17985	Sing	Manhattan	East Harlem	40.79685	-73.94872	Entire home/apt	190	7
----	------	----------------------------------	-------	------	-----------	-------------	----------	-----------	-----------------	-----	---

```
In [27]: data_set.shape
```

```
Out[27]: (48895, 18)
```

```
In [28]: data_set = data_set[data_set['price']>0]
data_set.head(25)
```

```
Out[28]:
```


	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_o
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149		1
1	2595	Skyliit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225		1
2	3647	THE VILLAGE OF HARLEM...NEW YORK!	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150		3
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89		1
4	5022	Entire Apt. Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80		10

با توجه به درس هایی که در داده کاوی آموخته ایم دیتای موجود در دنیای واقعی اصطلاحاً دیتای کیفی است و باید قبل از شروع کار روی دیتا پاکسازی انجام دهیم

کارهایی که به عنوان پاکسازی دیتا شناخته می شود مانند حذف نویز از دیتا است یا شامل حذف دیتای ناسازگار

پس ما در این مرحله دیتا را پاکسازی کردیم و قیمت های صفر را حذف کردیم.

بعد از این کار باید تعداد خانه های موجود در هر منطقه را بدست آوریم.

jupyter Airbnb-Tadayon-99422041 Last Checkpoint: an hour ago (autosaved)  Logout

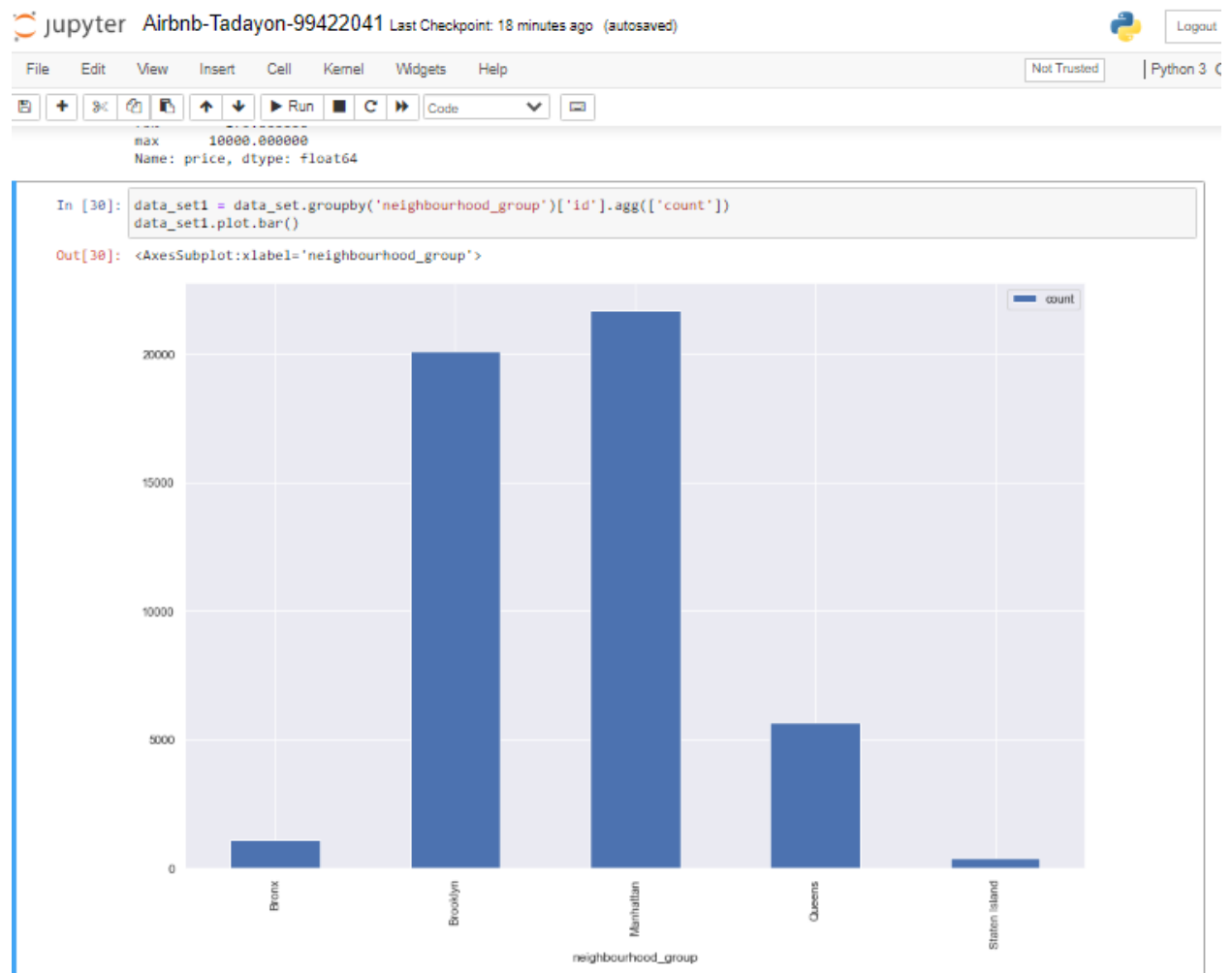
File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

In [28]: `data_set = data_set[data_set['price']>0]
data_set.head(25)`

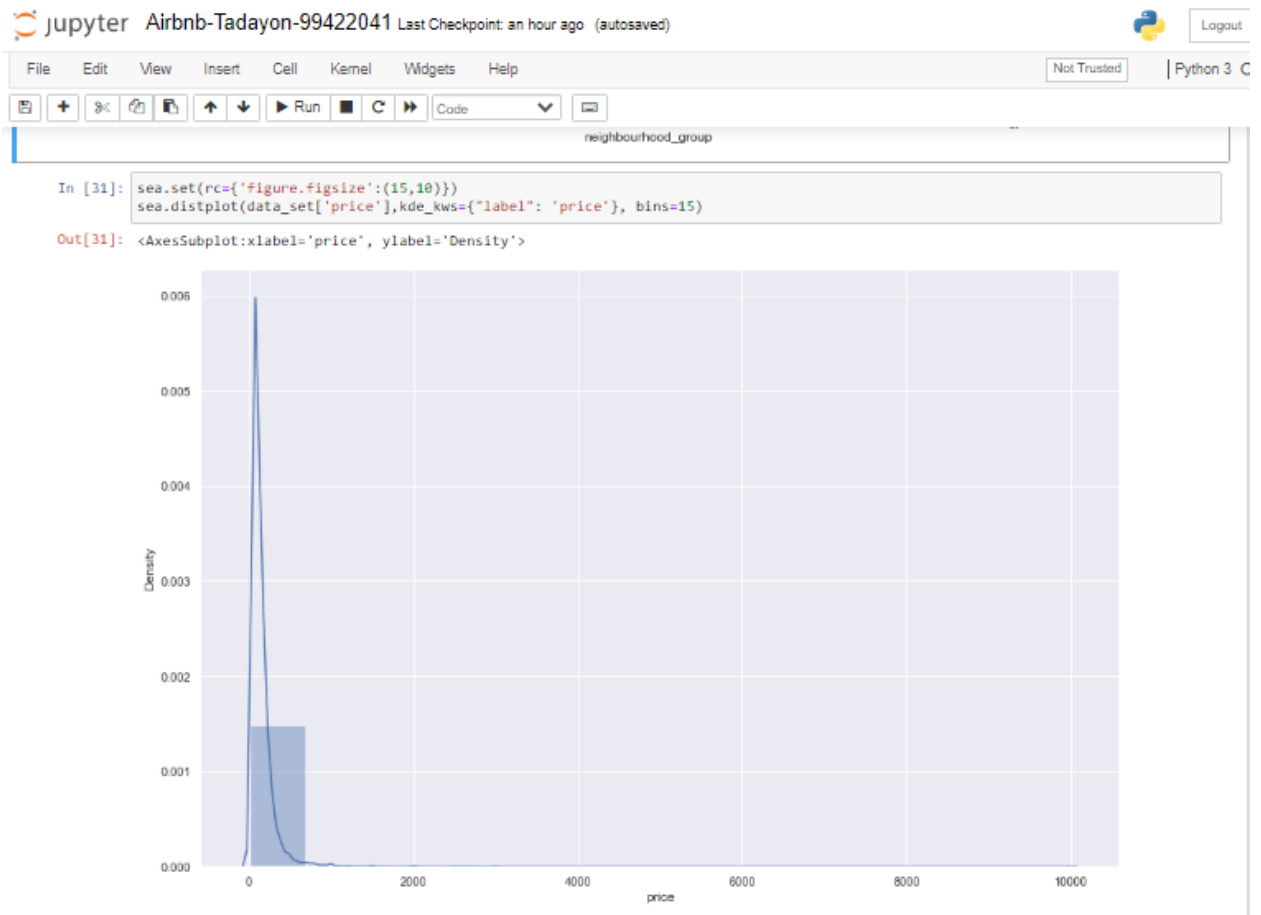
Out[28]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149		1
1	2595	Skyliit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225		1
2	3647	THE VILLAGE OF HARLEM...NEW YORK!	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150		3
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89		1
4	5022	Entire Apt. Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80		10
5	5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hill	40.74767	-73.97500	Entire home/apt	200		3
6	5121	BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesant	40.68688	-73.95596	Private room	60		45
7	5178	Large Furnished Room Near B'way	8967	Shunichi	Manhattan	Hell's Kitchen	40.76489	-73.98493	Private room	79		2
8	5203	Cozy Clean Guest Room - Family Apt	7490	MaryEllen	Manhattan	Upper West Side	40.80178	-73.96723	Private room	79		2
9	5238	Cute & Cozy Lower East Side 1 bdrm	7549	Ben	Manhattan	Chinatown	40.71344	-73.99037	Entire home/apt	150		1
10	5295	Beautiful 1br on Upper West Side	7702	Lena	Manhattan	Upper West Side	40.80316	-73.96545	Entire home/apt	135		5
11	5441	Central Manhattan/near Broadway	7989	Kate	Manhattan	Hell's Kitchen	40.76076	-73.98867	Private room	85		2
12	5803	Lovely Room 1, Garden, Best Area, Legal rental	9744	Laurie	Brooklyn	South Slope	40.66829	-73.98779	Private room	89		4
13	6021	Wonderful Guest Bedroom in Manhattan for SINGLES	11528	Claudio	Manhattan	Upper West Side	40.79826	-73.98113	Private room	85		2
14	6090	West Village Nest - Superhost	11975	Alina	Manhattan	West Village	40.73530	-74.00525	Entire home/apt	120		90
15	6848	Only 2 stops to Manhattan studio	15991	Allen & Irina	Brooklyn	Williamsburg	40.70837	-73.95352	Entire home/apt	140		2

نمودار زیر نشان می‌دهد که دو منطقه ی منهتن و بروکلین دارای بیشترین تعداد خانه است

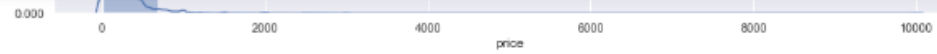


حال با توجه به تعداد خانه ها نمودار توزیع قیمت را رسم می کنیم.

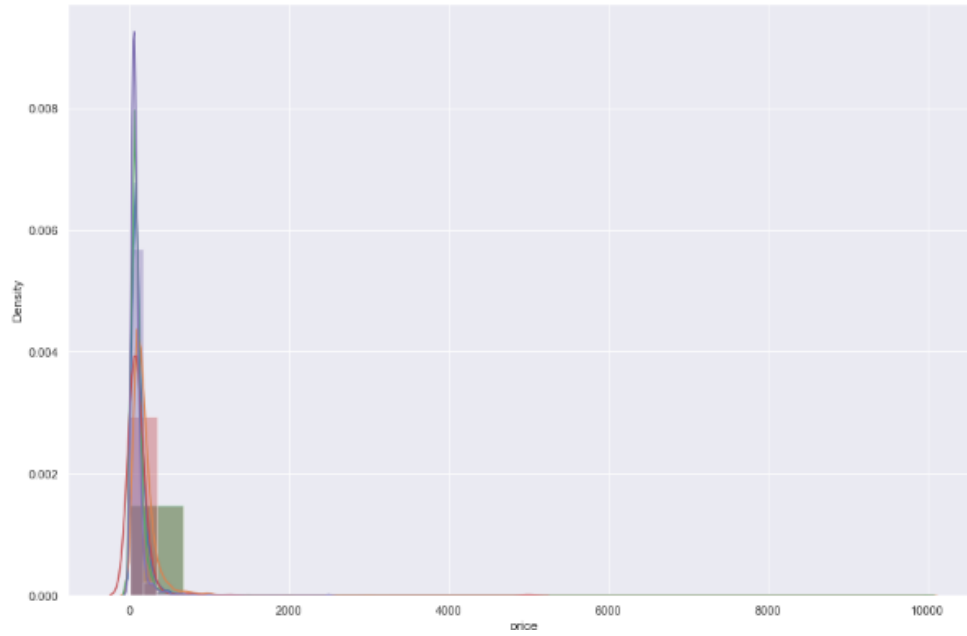


با رسم این نمودار آن چه مشاهده می شود این است که توزیع قیمت آپارتمان ها نرمال نیست .

پس باید توزیع را نرمال سازی کرده و مجدد نمودار را رسم کنیم .



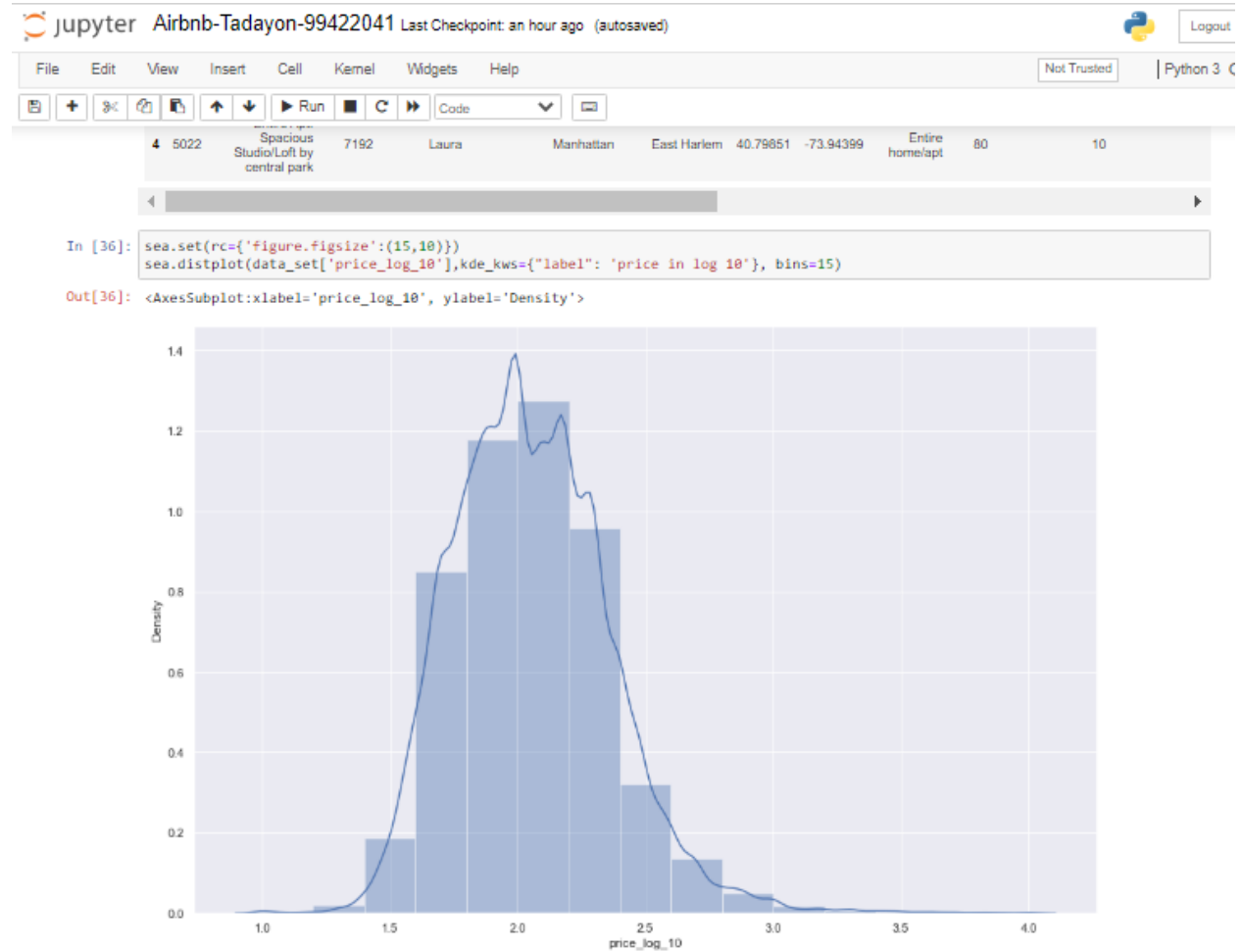
```
In [32]: sea.set(rc={'figure.figsize':(15,10)})
for groups in data_set.neighbourhood_group.unique():
    sea.distplot(data_set.price[data_set['neighbourhood_group']==groups],kde_kws={"label": groups},bins=15)
```



حال باید بررسی کنیم که آیا رابطه ای بین مناطق مختلف و قیمت خانه به چه صورت خواهد بود.



با استفاده از نمودار ویالون پلات که مشابه نمودار باکس پلات است بالا تریت قیمت خانه را بررسی می کنیم



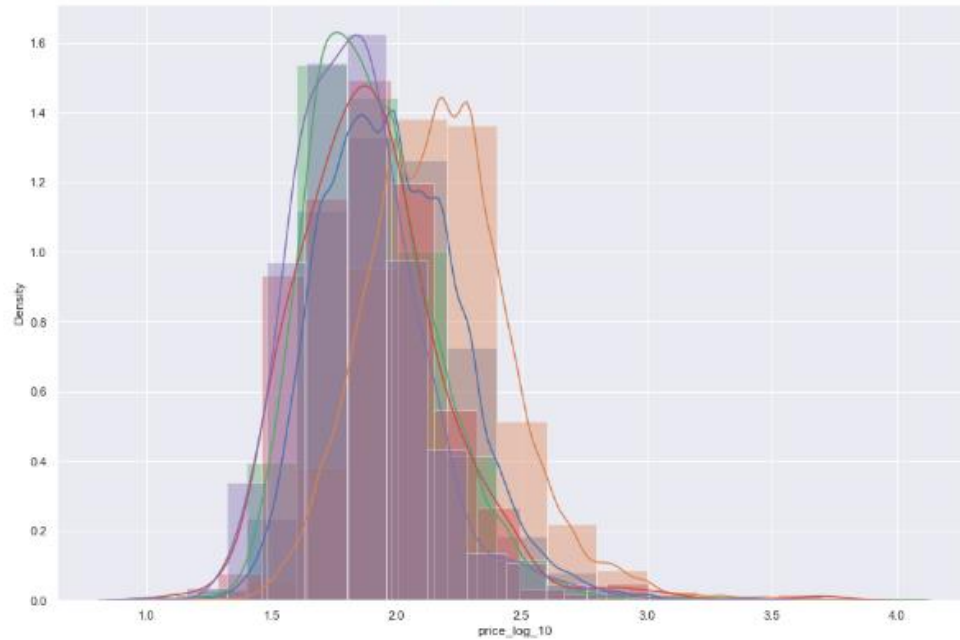
این نمودار ها و جدول ها به ما نشان می دهند که منطقه ی منهتن و بروکلین قیمت بالاتری نسبت به بقیه مناطق دارند .

حال در مرحله بعد ارتباط بین نوع اتاق ها و قیمت را بررسی میکنیم

```
In [37]: stats.normaltest(data_set["price_log_10"])
```

```
Out[37]: NormaltestResult(statistic=3926.8040825694948, pvalue=0.0)
```

```
In [38]: sea.set(rc={'figure.figsize':(15,10)})
for groups in data_set.neighbourhood_group.unique():
    sea.distplot(data_set.price_log_10[data_set['neighbourhood_group']==groups],kde_kws={"label": groups}, bins=15)
```



حال تعداد خانه های هر نوع اتاق را بررسی میکنیم.



entire home

این نمودار نشان میدهد که خانه هایی که اتاق

دارند بیشتر هستند .

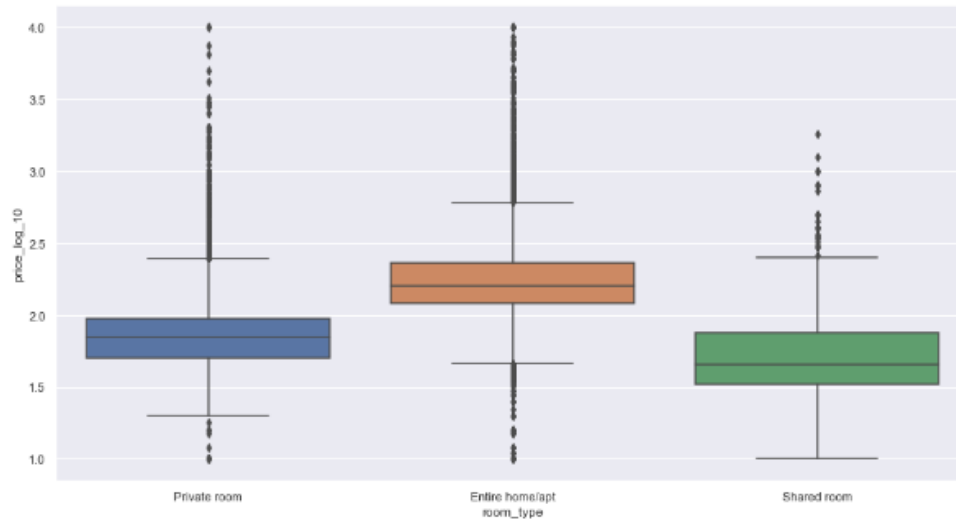
و بعد از آن اتاق خصوصی دارای بیشترین تعداد است.

Run Code



```
In [45]: plt.figure(figsize=(15,8))
seaborn.boxplot("room_type", "price_log_10", data=data_set)
```

```
Out[45]: <AxesSubplot: xlabel='room_type', ylabel='price_log_10'>
```



```
In [46]: fstat_nval = stats.f_oneway(*[data_set.price_log_10[data_set.room_type == c]
```