بسمه تعالی

دانشگاه شهید بهشتی
دانشکده علوم ریاضی
گروه علوم کامپیوتر

تمرین سری ۱ واحد درسی داده کاوی

سارا چرمچی
۴۰۰۴۲۲۰۶۶

استاد راهنما : جناب آقای دکتر فراهانی

دستیار آموزشی : آقای علی شریفی

فروردین ۱۴۰۱

مقدمه

کلیه تحلیل های این گزارش با استفاده از زبان برنامه نویسی پایتون اجرا شده است و کدها به پیوست موجود است.

تمرین اول

در این تمرین از دیتاست شرکت آمریکایی مربوط به اجاره منازل در شهر نیویورک استفاده می کنیم . این داده ها شامل نام و شماره خانه ها و میزبانان، منطقه و محله های شهری، اطلاعات جغرافیایی و قیمت و داده های ارزیابی آن ها توسط مهمانان و سایر اطلاعات است.

ابتدا کتابخانه های مورد استفاده را وارد می کنیم. سپس داده ها را به هر طریقی مناسب است می خوانیم (در اینجا داده های کگل را از طریق کولب فراخوانی می کنیم)

برای داشتن نمای کلی از دیتاست ابتدا داده ها را نمایش داده و اطلاعات کلی از قبیل نوع داده ها و تعداد و ... را بررسی می کنیم و اطلاعات آماری داده های عددی را نگاهی اجمالی می کنیم. هم چنین نمای کلی از تعداد سطر و ستون داده ها را می بینیم (۴۸۸۹۵, ۱۶).

```
id                                int64
name                             object
host_id                           int64
host_name                        object
neighbourhood_group              object
neighbourhood                    object
latitude                        float64
longitude                       float64
room_type                        object
price                             int64
minimum_nights                    int64
number_of_reviews                 int64
last_review                      object
reviews_per_month               float64
calculated_host_listings_count    int64
availability_365                  int64
dtype: object
```

۱.پاک سازی داده ها :

پاک سازی داده ها شامل حذف ستون های هجو، حذف داده های تکراری ، بررسی داده های ناموجود و رسیدگی به داده های پرت.

مرحله اول پاک سازی داده ها : حذف ستون های بی ارزش : در این دیتاست برخی از اطلاعات هجو اند مانند نام میزبان و تاریخ آخرین نظر. با توجه به آنکه تعداد نظرها و میانگین نظرها در ماه را داریم پس در این تحلیل نیازی به تاریخ آخرین نظر نیست

مرحله دوم پاک سازی داده ها : بررسی داده های ناموجود : null

ابتدا نمایی از تعداد کل داده های ناموجود در هر ستون را گرفته سپس تصمیم می گیریم. داده های null شامل

['name', 'reviews_per_month '] . با دستور fillna جای داده های ناموجود name را با NOname پر کرده و با توجه به آنکه اگر اطلاعاتی از نظرها در ماه موجود نیست پس نظری نیست و می نوان جای داده خالی را با همان روش و با صفر پر کرد

دوباره تعداد داده های ناموجود را بررسی کرده و میبینیم داده ی ناموجودی وجود ندارد

مرحله سوم پاک سازی داده ها : بررسی داده های تکراری ، در این دیتاست داده ی تکراری نداشتیم.

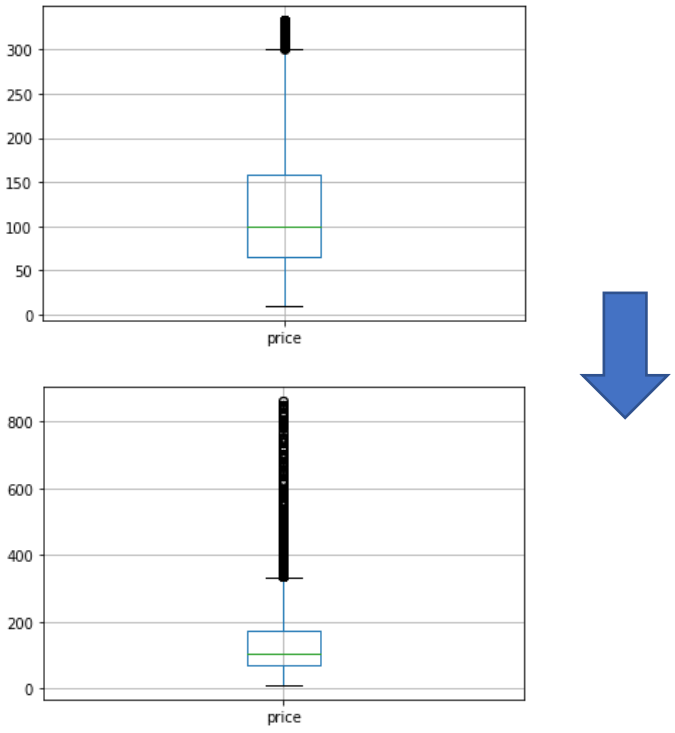
مرحله چهارم پاک سازی داده ها : رسیدگی به داده های پرت ، یکی از داده های عددی مهم این دیتاست price است. ابتدا اطلاعات اماری را با دستور describe() دریافت می کنیم و به کمینه صفر بر می خوریم. چون قیمت صفر به عنوان اجاره بها بی معنی است به عنوان داده اشتباه در نظر میگیریم و خذف می کنیم.

سپس داده های قیمت را با روش سنتی نرمالایز کردن داده بررسی می کنیم در این مورد، میانگین به اضافه یا منهای ۳ برابر انحراف معیار است که دادههای بزرگتر از میانگین به اضافه ۳ برابر انحراف معیار و کوچکتر از میانگین منهای ۳ برابر انحراف معیار، پرت محسوب می شوند. چون ایـن روش در سایر پارامترها نیز تحت تأثیر دادههای پرت است دوباره بررسی می کنیم.

می بینیم که همچنان داده های پرت زیادی وجود دارد پس از روش دیگری استفاده می کنیم:

روش نمودار جعبهای : این روش نموداری برای نشان دادن موقعیت، پراکندگی و چولگی دادهها به کار میرود و از فراوانی برای تشخیص دادههای پرت استفاده میکند. این نمودار با استفاده از یک مستطیل (Box)و دو خط یا میله در دو طرف مستطیل و بهوسیله میانه، چارکهای اول و سوم و کمترین و بیشترین مقادیر رسم میشود. طول مستطیل برابر با فاصله چارکی IQR یعنی تفاوت بین چارک سوم و چارک اول است. یا در یک نوع نمودار جعبهای که از آن برای تشخیص دادههای پرت استفاده میشود، دادههایی که کوچکتر $Q_1-(1.5)IQR$ یا بزرگتر از $Q_3+1.5(IQR)$ باشند جزء دادههای پرت خفیف و دادههایی که کوچکتر از $Q_1-(3)IQR$ یا بزرگتر از $Q_3+(3)IQR$ باشند جزء دادههای پرت قوی محسوب میشوند. شکل زیر نمودار جعبهای برای نمونههای مختلف را نشان میدهد؛

با تعریف تابعی با فرمول بالا و خدف داده های پرت خفیف به نموداری مطابق زیر رسیدیم :

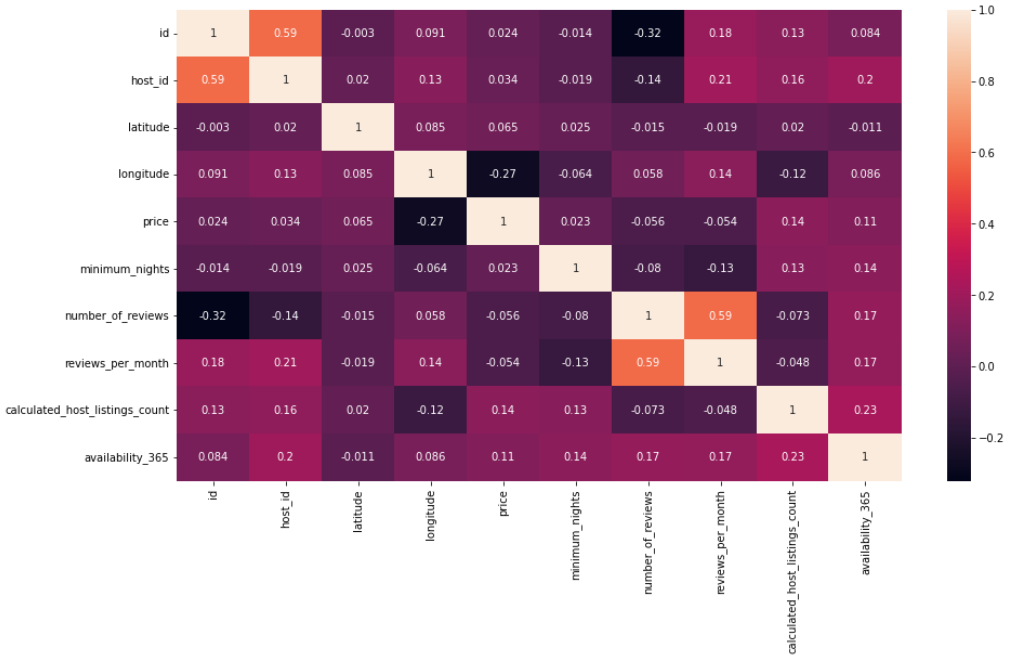


به وضوح داده های پرت تاحد زیادی حدف شده است

۲. مصور سازی داده ها

ارایه اطلاعات کلی در حالت تجمیعی در خصوص آگهی ها.

ابتدا رابطه کلی داده های عددی را نمایش می دهیم :

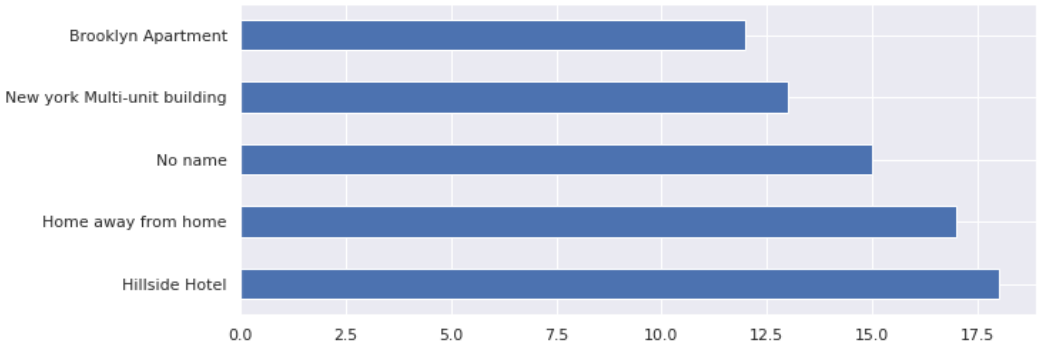


بدیهی است که رابطه ی هر مولفه با خودش حداکثری است (۱) و تا تیره ترین حالت که به صفر میل کرده و عدم ارتباط دو مولفه را نشان می دهد مانند رابطه تعداد نظرات و شناسه.

سپس به بررسی و مصور سازی هر ویژگی می پردازیم:

تعداد آگهی ها : تعداد آگهی های هر خانه را با دستور df.name.value_counts و ۱۰ تای اول را چاپ می کنیم.

سپس هیستوگرامی از ۵ اقامتگاه اول از حیث تعداد آگهی رسم می کنیم . اقامتگاه Hillside Hotel ۱۸ بار در آگهی ها آمده است.



میبینیم که ۱۵ اقامتگاه فاقد نام منحصربفردی در دیتاست بوده است.

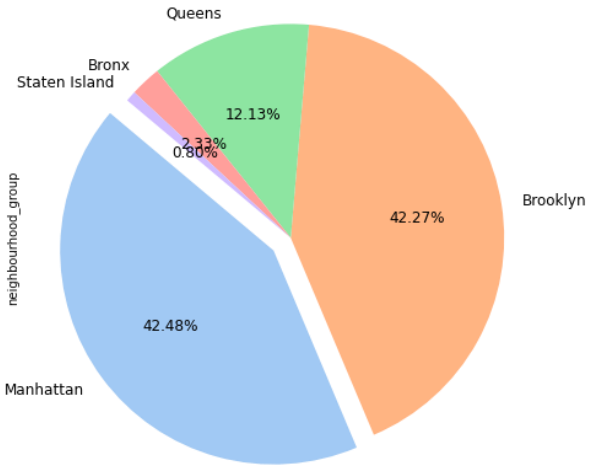
تعداد آگهی ها در هر منطقه جغرافیایی:

با توجه به انکه مناطق جغرافیایی جزو داده های categorical است می توان با دستور unique() آماری از کلیه داده ها بدست آورد

و نموداری از تجمع آگهی ها در مناطق ۵ گانه رسم می نماییم:

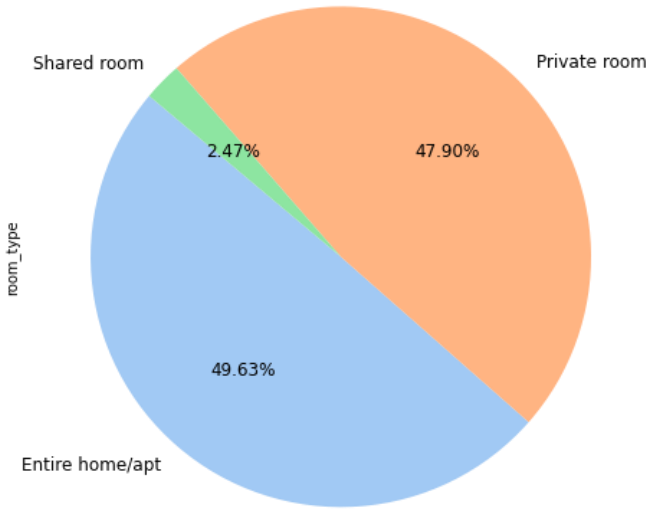
Manhattan	19500
Brooklyn	19406
Queens	5567
Bronx	1069
Staten Island	365

Name: neighbourhood_group, dtype: int64



محله Manhattan با بیش از ۴۲٪ فراوانی پر تقاضاترین محله نیویورک و Staten Island با کم تر از ۱٪ کم تقاضاترین محله است.

تقاضای نوع اتاق ها چگونه است ؟ نموداری از درصد اگهی ها برحسب نوع اتاق ها رسم می کنیم، تقاضای خانه دربست اول است :

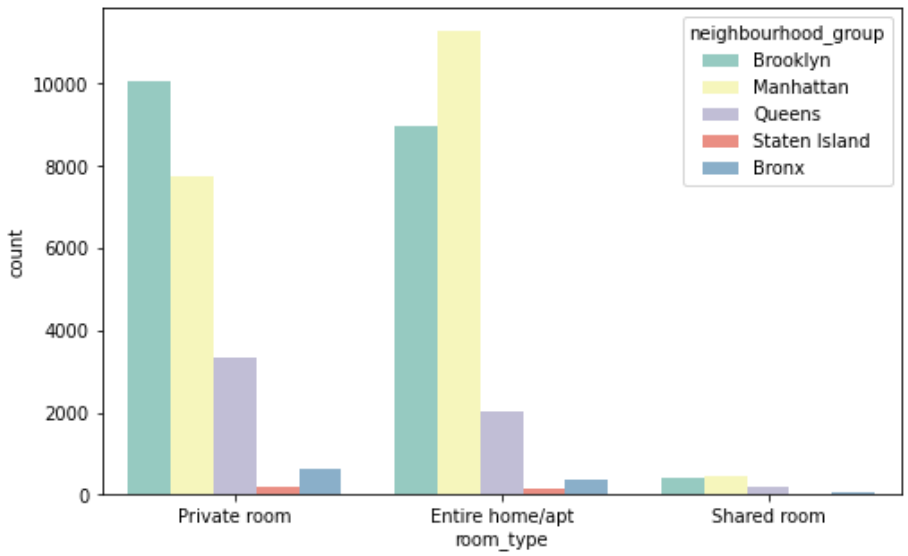


اما سوال : چه رابطه ای بین محله های ۵ گانه نیویورک و نوع اتاق ها در آگهی ها وجود دارد؟

با استفاده از تکنیک seaborn.countplot رابطه معناداری بین دو مولفه ی محله های ۵ گانه نیویورک و نوع اتاق ها پیدا می کنیم. count plot نوعی هیستوگرام است که به جای داده های کمی و متغیرها با داده های categorical کار می کند.

از plot می توان دریافت :

- محله Manhattan بیشترین آگهی خانه دربست را بین سایرین دارد، پس از آن محله Brooklyn در رتبه دوم قرار دارد.
- در مقابل خانه، اتاق های شخصی رتبه Brooklyn بیشتر از Manhattan است. یعنی در مورد اتاق های شخصی محله Brooklyn پرتقاضا تر است یا Manhattan به اندازه محله Brooklyn اتاق شخصی اگهی نشده است.
- همانطور که از piechart پیشین دیدم اتاق های مشترک کمترین آگهی ها را شامل می شدند که همچنان منطقه Manhattan پرتقاضاترین است.
- محله های Brooklyn, Queens, Bronx بیشترین اتاق های شخصی و محله Manhattan که در کل پرتقاضاترین است رتبه اول در تعداد اگهی های خانه دربست را دارد.

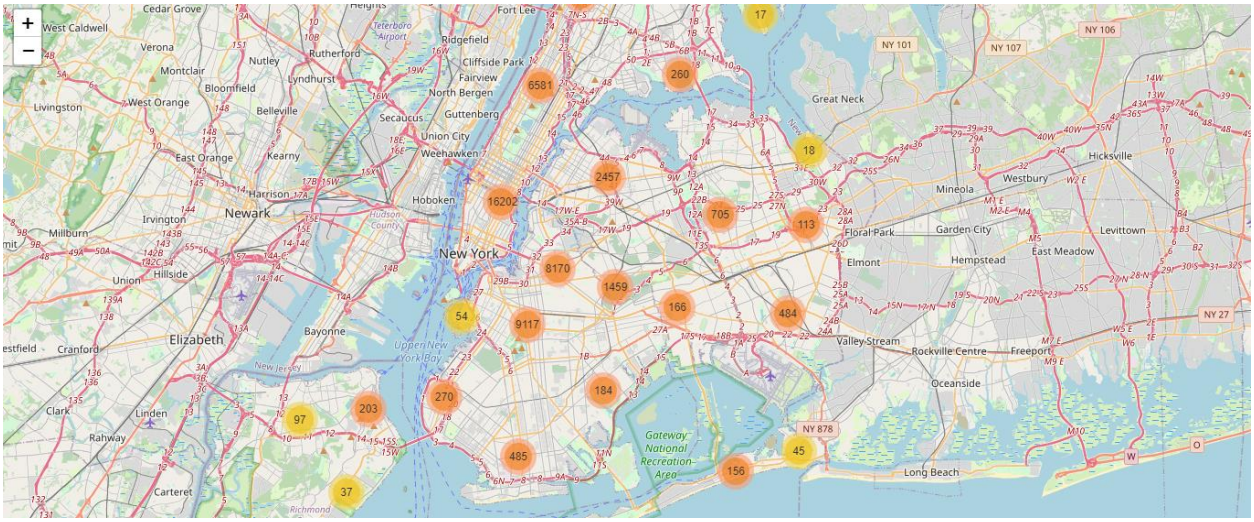


اطلاعات آماری روی نقشه

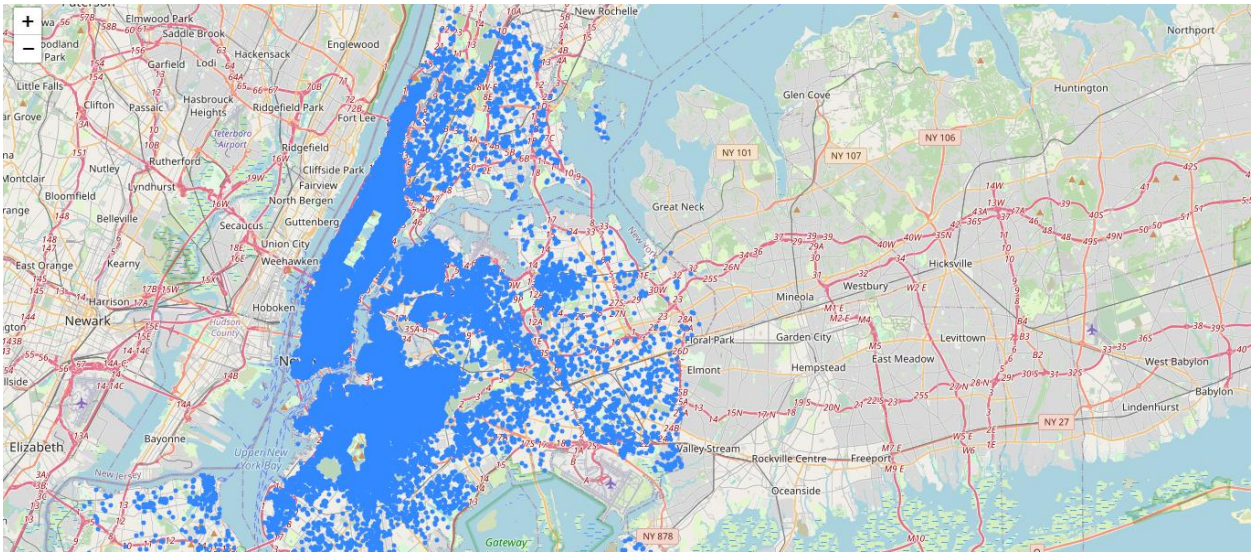
در مورد محله های جغرافیایی بحث کردیم اما بهتر است برای دید بهتر این اطلاعات آماری را روی نقشه نیویورک پیاده سازی کنیم:

با استفاده از کتابخانه Folium که کتابخانه ای پایتونی است که در پس زمینه از ماژول های جاوا استفاده می کند و در خروجی نقشه interactive می دهد.

به عنوان نقطه مرکزی نمایش مختصات شهر نیویورک را می دهیم و لیستی از مختصات های دیتاست و نمایش نقاط محله های نیویورک را به تابع MarkerCluster() می دهیم به این ترتیب خروجی یک فایل html است که روی گوگل کروم قابل فرخوانی است و نقشه interactive از نیویورک می دهد که روی تمام مناطق اگهی دار علامت گذاری شده است و تعداد اگهی هر منطقه روی نقشه دیده می شود.
نمونه از خروجی بدست امده (در قالب screenshot) :

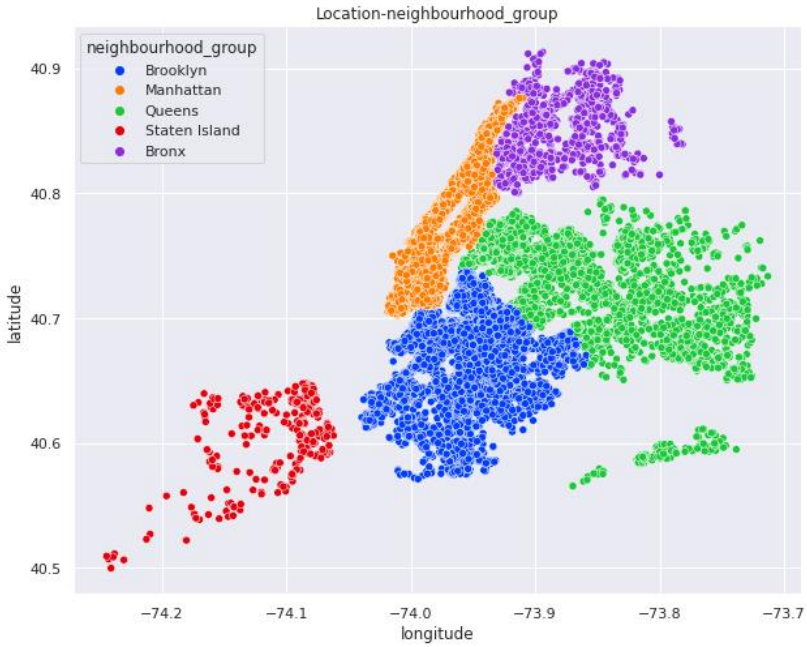


می توان با استفاده از تابع CircleMaker() پراکندگی نقاط را با دایره روی نقشه نمایش داد. نمونه از خروجی بدست امده (در قالب screenshot) :



نقشه تراکم زیاد اقامتگاه ها در مناطق گران قیمتی مثل Manhattan را به خوبی نشان می دهد.

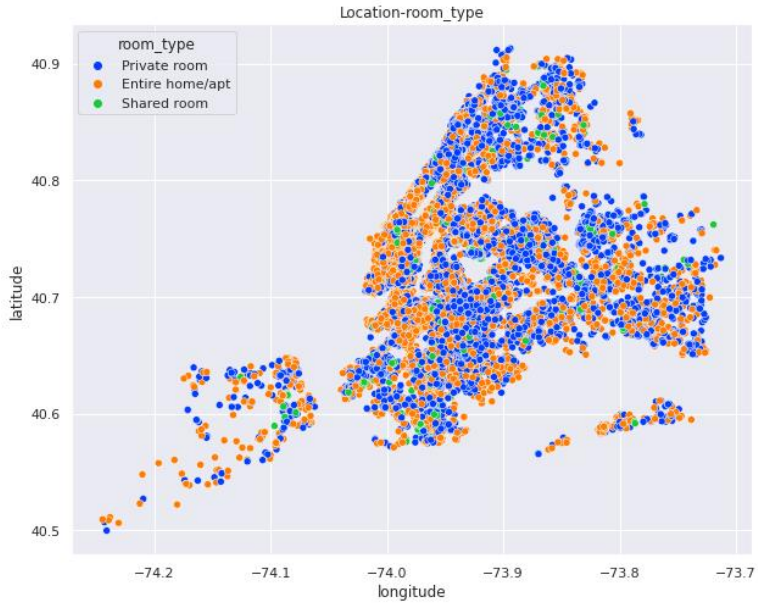
حال با استفاده از کنابخانه seaborn نقشه ای از پراکندگی آگهی ها روی مناطق ۵ گانه نیویورک بدست می آوریم به این ترتیب که داده کمی ورودی داده های مختصات جغرافیایی است و داده categorical مناطق ۵ گانه نیویورک است.



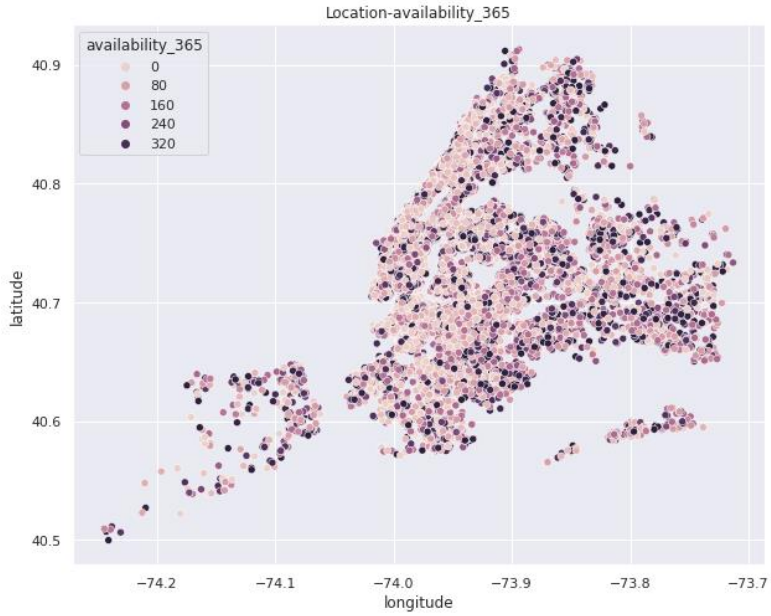
جالب است که منطقه پر تقاضای Manhattan که پیشتر بررسی کردیم منطقه نسبتا کوچکی در شهر است

حال پراکندگی نوع اتاق ها را با همین روش روی نقشه پیاده سازی می کنیم.

همانطور که پیشتر دیدم اتاق های مشترک که کمتر ۳٪ اقامتگاه ها را شامل می شود با پراکندگی زیاد و تعداد اندک روی نقشه می بینیم.



سپس موجود بودن اتاق ها در طی سال را با مولفه availability_365 روی نقشه پیاده سازی می کنیم.



شاخص های کلی قیمت

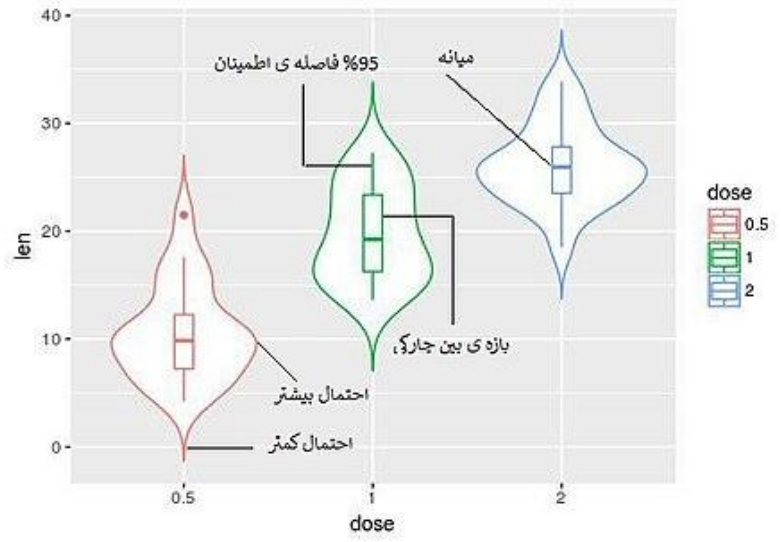
ابتدا اطلاعات آماری ستون price را دریافت می کنیم.

میانگین هزینه اقامت در نیویورک 119.975755 است و حداقل هزینه لازم 10.000000 است. گران ترین اقامتگاه در نیویورک 333.000000 هزینه دارد مانند اقامتگاهی در منهتن به نام Midtown Sleep 6 Central Convenience

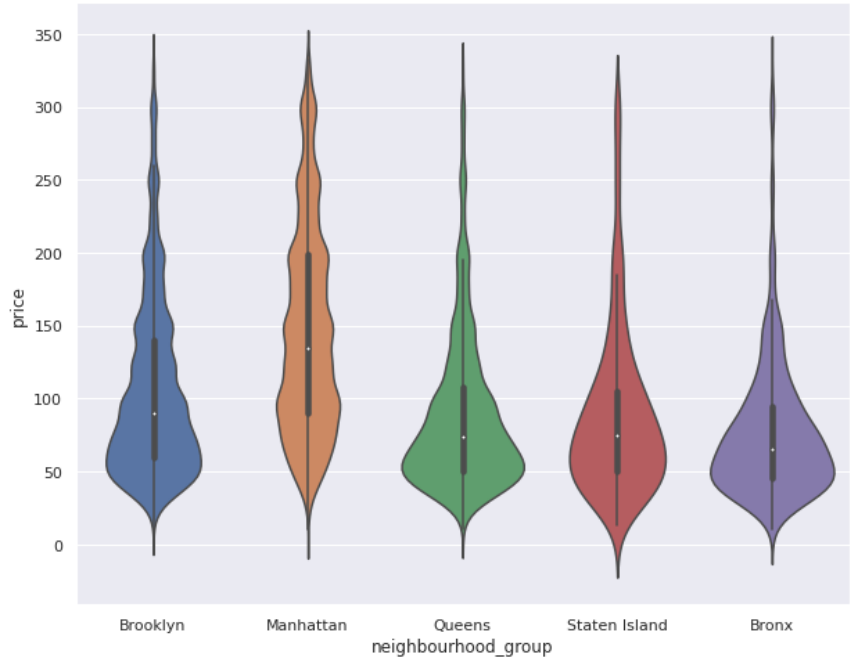
نگاهی به پرهزینه ترین اقامتگاه ها در نیویورک می اندازیم : (بخشی از اطلاعات گران ترین اقامتگاه ها)

	id	name	host_id	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	reviews_p
13798	10403199	Mid-Century Museum Sleepover	29065752	Brooklyn	Williamsburg	40.71863	-73.94528	Entire home/apt	333	7	0	
15150	12093494	Midtown Sleep 6 Central Convenience	62031986	Manhattan	Midtown	40.75237	-73.98769	Entire home/apt	333	1	246	
19641	15728780	SOMMwhere in NYC/ a unique, conscious artists ...	148108	Manhattan	Lower East Side	40.72297	-73.98946	Private room	333	1	40	
25161	20154569	Flatiron Loft 3BR/1.5 Bath Best Location\n30 days	35635299	Manhattan	Midtown	40.74353	-73.98364	Entire home/apt	333	30	66	
28062	21919913	27 FLR VIEWS!LINCOLN SQR-LUXURY 2BR MIDTOWN W	76104209	Manhattan	Upper West Side	40.77055	-73.98615	Entire home/apt	333	30	0	

کدام مناطق گران تر هستند؟ برای پاسخ به این سوال باید توزیع قیمت در هر منطقه را بررسی کرد. برای این منظور می توان از نمودار ویالونی استفاده کرد. نمودار ویولنی روشی برای رسم کردن داده‌های عددی است، گاهی میانگین و میانه برای درک و شناخت یک مجموعه داده به تنهایی کافی نیستند



با کتابخانه seaborn می توان نمودار ویالونی را رسم کرد :

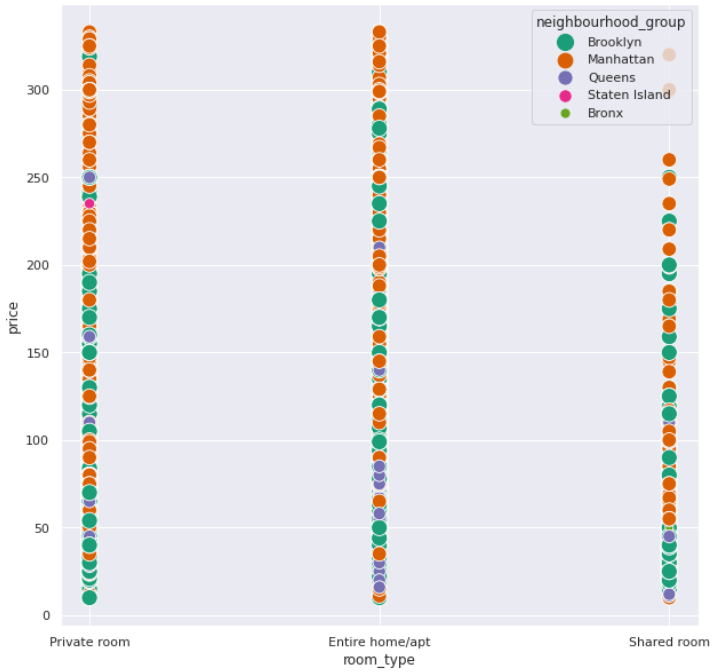


هر کدام از دو طرف خط رسم شده عمودی وسط نمودار، یک برآورد چگالی برای نمایش توزیع شکل داده‌هاست. بخش‌های عریض‌تر نمودار نشان‌دهندهاین است که نمونه‌ها در داده مورد نظر با احتمال بیشتری این مقدار را می‌توانند بگیرند و هر چه برای یک مقدار این عرض

کوچکتر باشد احتمال آن کمتر است. همانطور که انتظار داشتیم Manhattan گران ترین منطقه نیویورک است. و در مقابل منطقه Bronx را داریم که در قیمت های پایین نمودار عریض‌تر شده و با احتمال زیاد ارزان ترین منطقه برای اقامت است.

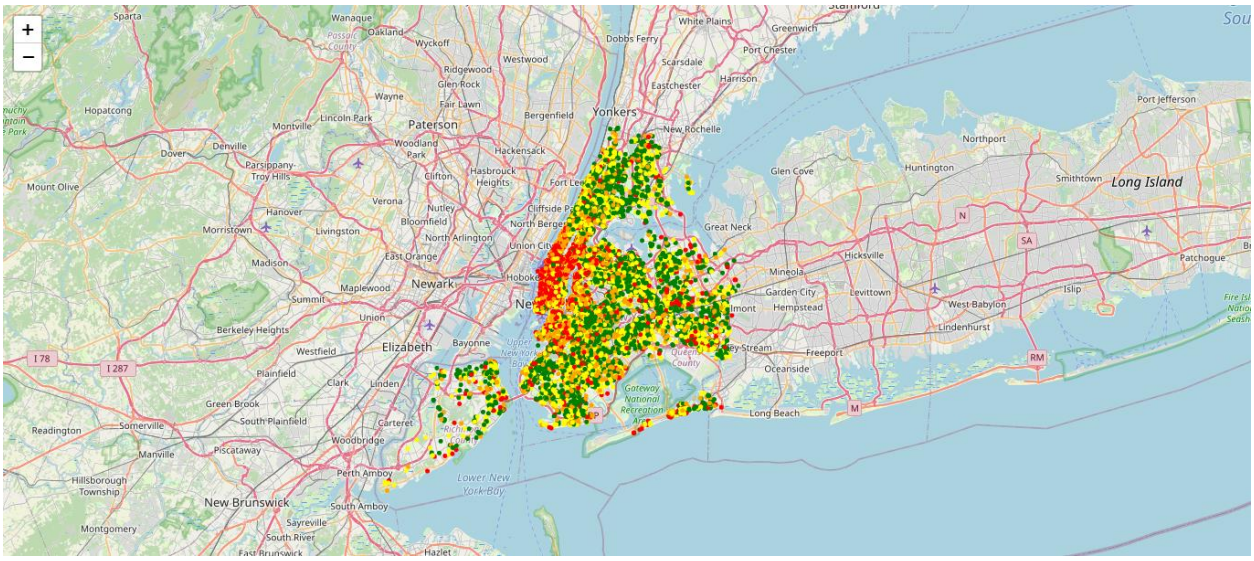
در ادامه علاوه بر شاخص منطقه جغرافیایی، نوع اتاق را هم مد نظر قرار داده و توزیع قیمت را بررسی می کنیم:

دو داده categorical داریم و داده قیمت به عنوان داده کمی وارد می شود : توزیع قیمت بر اساس نوع اتاق در مناطق ۵ گانه با استفاده از seaborn.scatterplot :



همانطور که میبینیم توزیع قیمتی در منطقه Manhattan برای یک خانه کامل بسیار متنوع است. در مقابل اتاق های اشتراکی با بازه قیمتی کمتر و قیمت های بیشتر در حتی در اتاق های اشتراکی مناطق Manhattan و Brooklyn دیده می شود. و همانطور که انتظار داریم Manhattan جزو گران ترین مناطق اقامتی شهر است.

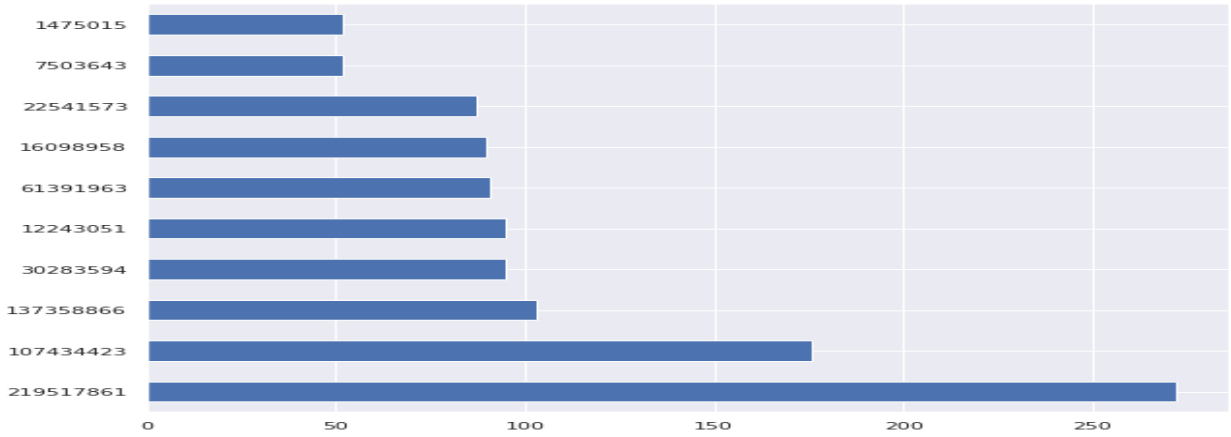
در ادامه توزیع قیمتی در مناطق جغرافیایی را روی نقشه با روش پیشتر توضیح داده شده folium بررسی می کنیم: رنگ های مختلفی را برای محدوده های قیمتی تعیین می کنیم. به این منظور چارک های قیمتی را در نظر گرفته و چهار رنگ را برای بازه های : {بیشینه قیمتی، ۷۵ ٪، ۵۰٪ ، ۲۵٪ ، کمینه قیمتی} دو به دو در نظر می گیریم و نقشه حاصل (در قالب screenshot) :



می بینیم که تراکم رنگ قرمز در منطقه Manhattan بیشتر به چشم می خورد که قرمز نشاندهنده بازه قیمتی حداکثری است.

۳. بررسی صاحبان آگهی و گزارشی از تعداد خانه های مرتبط با هر صاحب آگهی

تعداد اگهی های مربوط به هر میزبان را با دستور value_counts() بدست می آوریم. نمودار زیر نمایشی از ۱۰ میزبان پر آگهی است :



از روی نمودار مشخص است که میزبان به شناسه ۲۱۹۵۱۷۸۶۱ با تعداد ۲۷۲ آگهی پر آگهی ترین میزبان در این دیتاست است.

حال به بررسی اقامتگاه ها و صاحبین هر اقامتگاه می پردازیم و با گروهبندی خانه های مرتبط با هر صاحب آگهی را می یابیم

و میبینیم که میزبان با شناسه ۲۱۹۵۱۷۸۶۱ که حدود ۲۷۲ آگهی دارد ۲۶۰ اقامتگاه دارد.

حال تعداد اقامتگاه های هر میزبان را می یابیم : (۱۰ میزبان با بیشترین تعداد خانه)

```
top10_host.head()
```

host_id		id	name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	reviews_per_month	cal
219517861	**nunique**	272	260	1	7	222	232	2	93	2	19	125	
107434423	**nunique**	176	175	2	19	172	168	1	72	4	3	16	
137358866	**nunique**	103	103	3	10	101	100	2	37	1	5	35	
12243051	**nunique**	95	95	1	5	90	91	1	51	1	5	22	
30283594	**nunique**	95	95	1	8	93	88	1	23	2	5	12	

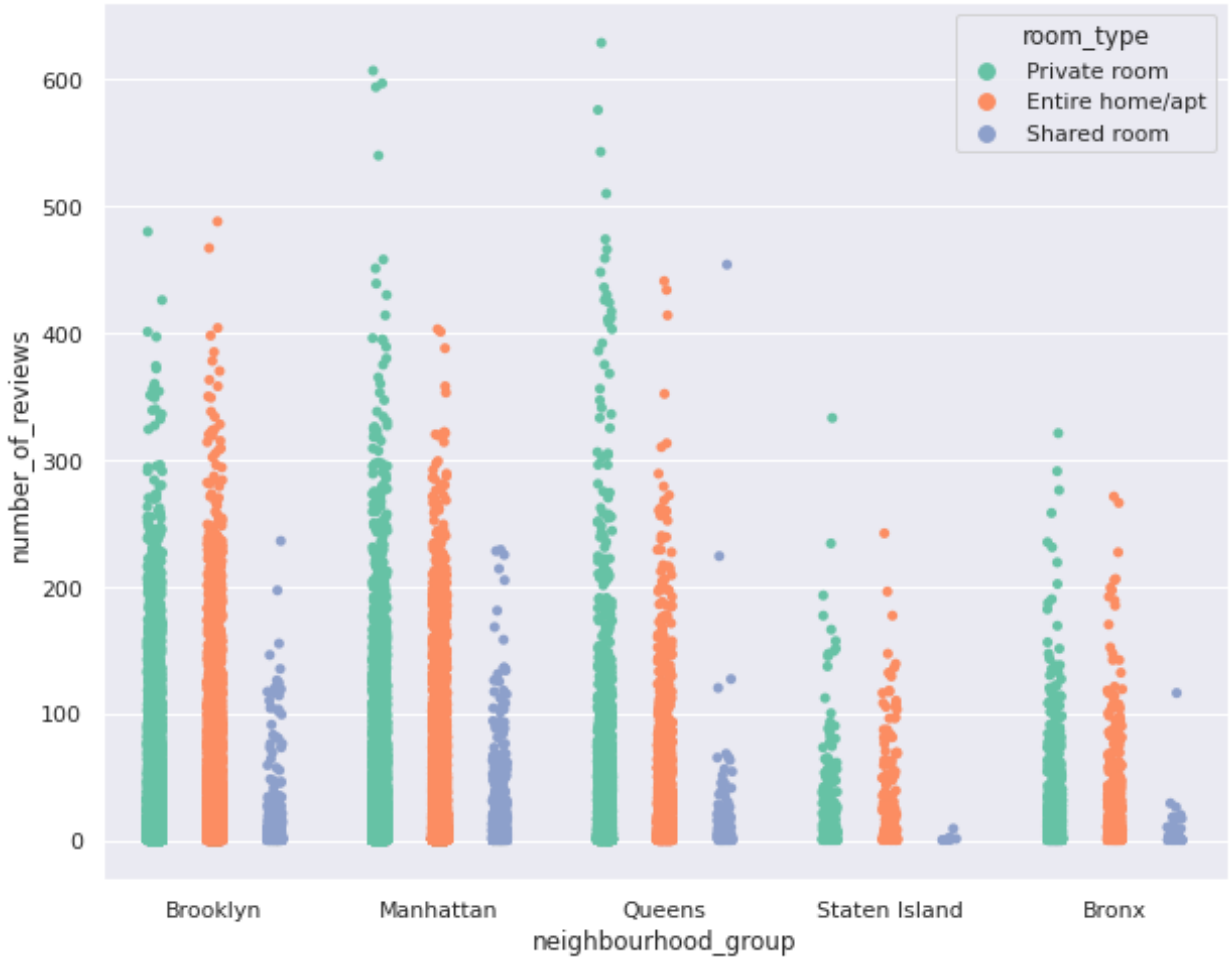
۴. صاحبان اگهی با بیشترین مشتری

ابتدا اطلاعات کلی آماری از قبیل تعداد کل نظرات و میانگین و حداقل تعداد نظرات را بدست می اوریم و میبینیم میزبانی با اقامتگاه Room near JFK Queen Bed بیشترین تعداد نظرات یعنی ۶۲۹ نظر یا به نوعی بیشترین مشتری را داشته است.

و ۵ میزبان با بیشترین تعداد نظرات را نمایش می دهیم :

	id	name	host_id	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	reviews_per_mor
11759	9145202	Room near JFK Queen Bed	47621202	Queens	Jamaica	40.66730	-73.76831	Private room	47	1	629	14
2031	903972	Great Bedroom in Manhattan	4734398	Manhattan	Harlem	40.82085	-73.94025	Private room	49	1	607	7
2030	903947	Beautiful Bedroom in Manhattan	4734398	Manhattan	Harlem	40.82124	-73.93838	Private room	49	1	597	7
2015	891117	Private Bedroom in Manhattan	4734398	Manhattan	Harlem	40.82264	-73.94041	Private room	49	1	594	7
13495	10101135	Room Near JFK Twin Beds	47621202	Queens	Jamaica	40.66939	-73.76975	Private room	47	1	576	13

در ادامه رابطه تعداد مشتری های اقاتگاه ها را با محله های ۵ گانه نیویورک و نوع اتاق آن ها نمایش می دهیم (با استفاده از کتابخانه seaborn) :



می بینیم که کمترین نظرات را منطقه Staten Island و برای اتاق های اشتراکی داشته.

بررسی می کنیم دلیل پر مشتری بودن ۱۰ میزبان اول در چیست؟

ابتدا نگاهی به مناطق جغرافیایی و نوع اتاق های اقامتگاه های برتر می اندازیم :

neighbourhood_group	
Manhattan	4
Queens	4
Brooklyn	2
dtype: int64	

room_type	
Private room	9
Entire home/apt	1
dtype: int64	

بیشترین مربوط به منطقه پرتراکم Manhattan و اتاق های شخصی است.

آزمونی را مطرح می کنیم که ببینیم ایا رابطه ای بین نوع محله و تعداد مشتری ها وجود دارد؟ هم چنین ایا رابطه ای بین نوع اتاق و تعداد مشتری ها وجود دارد؟

فرض آزمون اول :

H0 (null hypothesis): variance(private_room) = variance(shared_room) = variance(entire_home)

H1 (alternate hypothesis): variance(private_room) != variance(shared_room) != variance(entire_home)

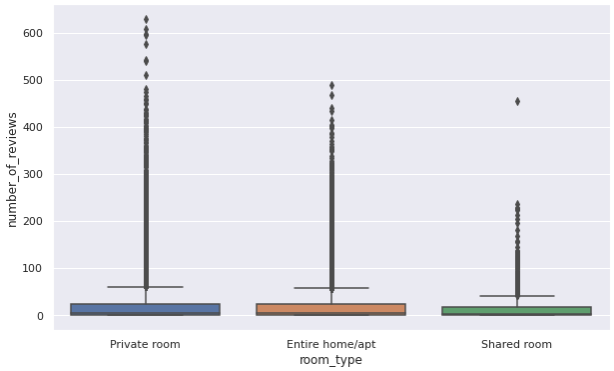
و

میبینیم نتیجه :

LeveneResult(statistic=14.596990768180596, pvalue=4.59856641400037e-07)

جواب p-value کمتر از الفا (۰.۰۵) و نزدیک به صفر است و فرض صفر ما رد می شود و می توان گفت واریانس تعداد مشتری ها برای انواع متفاوت اتاق ها برابر نیست

و این مسیله را با نمودار جعبه ای زیر می توان دریافت :



فرض آزمون دوم :

H0 (null hypothesis): mean_reviews(Brooklyn) = mean_ reviews (Manhattan) = ... = mean_ reviews (Bronx) (see categories above)

H1 (null hypothesis): mean_ reviews (Brooklyn) != mean_ reviews (Manhattan) != ..... != mean_ reviews (Bronx)

جواب :

KruskalResult(statistic=221.1279817666322, pvalue=1.0719953553621434e-46)

جواب p-value کمتر از الفا (۰.۰۵) و نزدیک به صفر است و فرض صفر ما رد می شود و می توان گفت تعداد مشتری ها برای محلات مختلف برابر نیست

۵. آزمون فرض دلخواه

آزمون اول : آیا میانگین قیمت در محله های مختلف نیویورک یکسان است ؟

از روش ANOVA استفاده می کنیم، اساس آزمون تحلیل واریانس، تجزیه واریانس به دو بخش واریانس یا «پراکندگی بین گروهی (Between Group Variability)و واریانس یا «پراکندگی درون گروهی (Within Group Variability) «است.

فرض :

H0 (null hypothesis): mean_price(Brooklyn) = mean_price(Manhattan) = ... = mean_price(Bronx)

H1 (null hypothesis): mean_price(Brooklyn) != mean_price(Manhattan) != ..... != mean_price(Bronx)

جواب :

F_onewayResult(statistic=24.524367116502283, pvalue=2.627963864762465e-20)

وقتی p-value نزدیک به صفر است و کمتر از آلفا است فرض رد می شود و همانطور که انتظار داشتیم بین میانگین قیمت و محله ارتباط وجود دارد و برخی محلات حتی در میانگین نیز گران قیمت ترند.

آزمون دوم : آیا برای هر اقامتگاه محله آن و میزبان آن تعیین کننده قیمت است ؟

از روش Two Way ANOVA استفاده می کنیم. برای آزمون برابری میانگین بین چند جامعه که توسط بیش از دو متغیر طبقه‌ای معرفی شوند، باید از تحلیل واریانس دو طرفه استفاده کرد.

	sum_sq	df	F	PR(>F)
neighbourhood_group	2.447747e+07	4.0	1492.811395	0.000000
host_id	3.701216e+04	1.0	9.029069	0.002659
Residual	1.881584e+08	45901.0	NaN	NaN

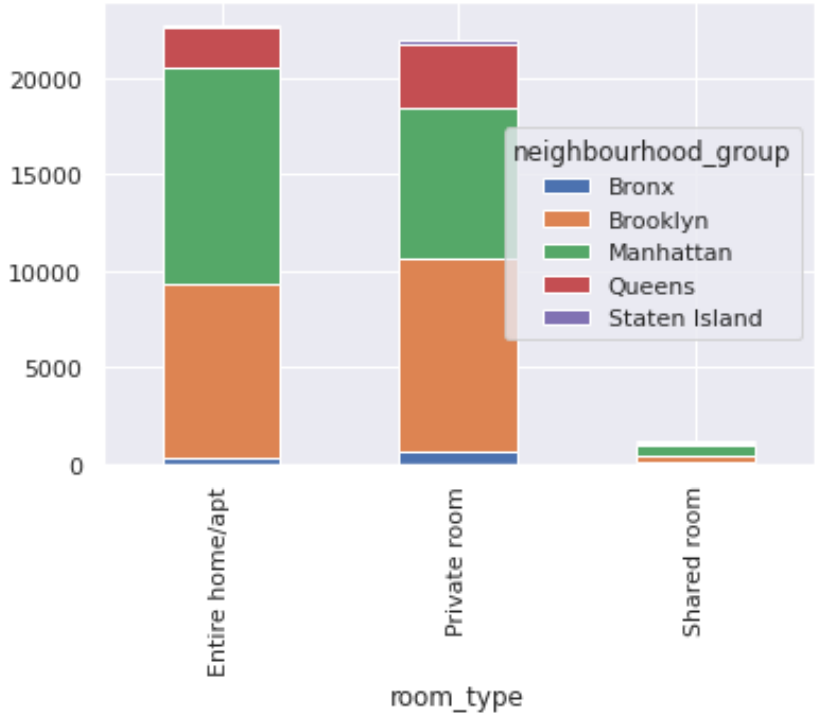
با توجه به مقدار p-value می توان گفت بین قیمت یک اقامتگاه و صاحب آن و محله ان ارتباطی وجود دارد یعنی برای یک اقامتگاه در دو منطقه متفاوت و صاحب متفاوت قیمت متفاوتی بدست می اید.

آزمون سوم :

از آزمون Chi Square Test استفاده می کنیم ، برعکس آزمون‌های t که پارامتری است و با توجه به توزیع داده‌ها صورت می‌گیرد، «ازمون کای ۲» (Chi Sqaure Test) از نوع آزمون‌های ناپارامتری است. به این معنی که برای داده‌ها یا جامعه آماری هیچ توزیعی احتمالی در نظر گرفته نمی‌شود. از این آزمون برای تعیین ارتباط بین دو متغیر طبقه‌ای استفاده می‌شود.

فرض:

جواب : p-value= 1.1127147476863028e-251 که باز هم از افا کوچک تر است و یعنی بین نوع اتاق و محله ارتباط وجود دارد یعنی نسبت وجود یک اتاق از نوعی خاص بستگی به محله ای دارد که به دنبال اتاق در ان هستیم. همانطور که از نمودار پیداست می توان گفت اگر به دنبال خانه دربست باشیم در محله Manhattan گزینه های بیشتری داریم نسبت به محله Bronx.



آزمون چهارم: آیا واریانس قیمت در انواع اتاق ها برابر است؟

یکی از آماره‌های استنباطی برای سنجش برابری واریانس در چند جامعه مستقل، استفاده از آماره لون و اجرای آزمونی به نام «آزمون لون » (Levene’s Test)است. فرض صفر در آزمون لون، یکسان بودن واریانس‌ها است.

فرض:

H0 (null hypothesis): variance(private_room) = variance(shared_room) = variance(entire_home)

H1 (alternate hypothesis): variance(private_room) != variance(shared_room) != variance(entire_home)

جواب :

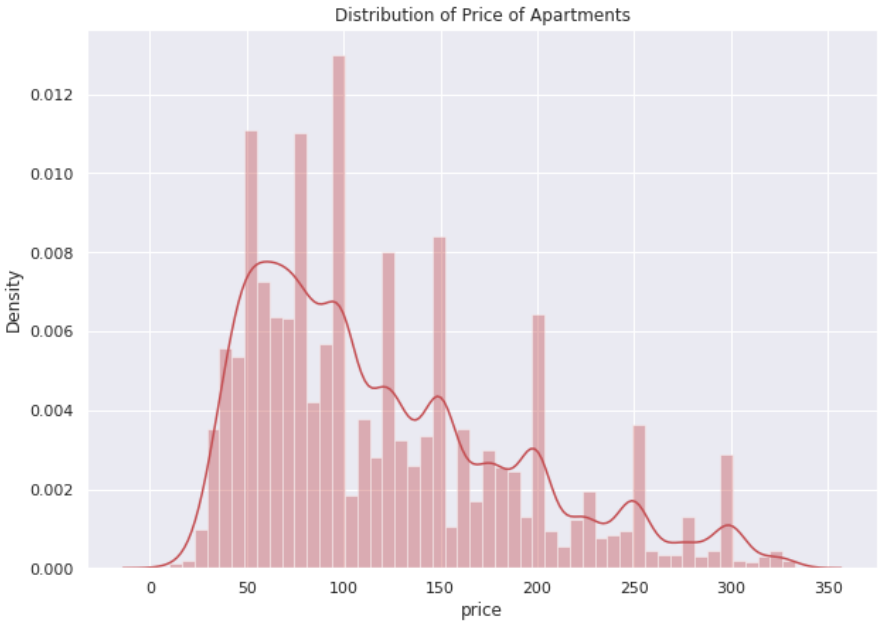
LeveneResult(statistic=2373.430140290372, pvalue=0.0)

با این مقدار فرض صفر رد شده است و یعنی واریانس قیمتی برای انواع اتاق ها برابر نیست و به عنوان مثال واریانس قیمتی در اتاق های مشترک بیشتر از خانه های دربست است.

آزمون پنجم: آیا توزیع قیمت نرمال است ؟

«آزمون شاپیرو ویلک» (Shapiro-Wilk Test) از آزمون‌های برازش توزیع نرمال محسوب می‌شود. به کمک این آزمون و آماره آن می‌توانید مشخص کنید که آیا داده‌ها از توزیع نرمال پیروی می‌کنند یا خیر.

ابتدا توزیع قیمت را رسم می کنیم :



به وضوح نرمال نیست و چولگی دارد، اما مورد آزمون Shapiro قرار می دهیم

جواب : (0.9190240502357483, 0.0)

آزمون نیز نرمال نبودن توزیع قیمت را تایید کرد.

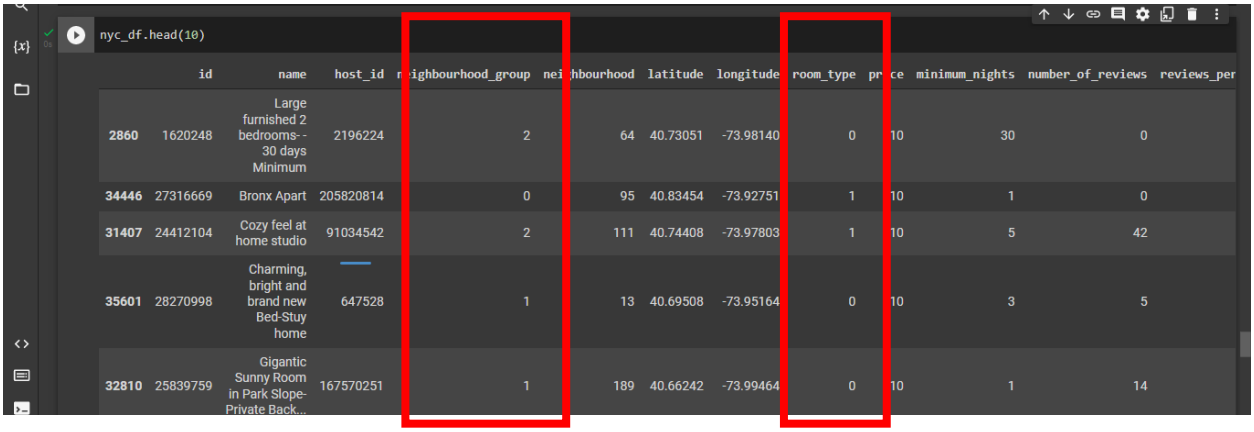
۶. تلاش در ساخت مدل جهت پیش بینی پارامترهایی مانند قیمت

ایا می توان مدلی ارایه کرد که براساس شاخصه های دیتاست، قیمت اقامتگاه ها را پیش بینی کند؟

از مدل رگرسیون خطی استفاده می کنیم. اما پیش از مدل سازی، بخش پیش پردازش داده ها را تکمبل می کنیم تا مدل پاسخ بهتری دهد. منظور از پیش پردازش داده ها encoding است. مدل های یادگیری ماشین نیاز دارند متغیرهای عددی به عنوان ورودی دریافت کنند بنابراین اگر در دیتاست متغیرهای categorical داشته باشیم برای فیت کردن و ارزیابی مدل پیش از مدل سازی داده ها را encode می کنیم.

ابتدا با روش sklearn.preprocessing.LabelEncoder برچسب را بین 0 and n_classes-1 کد گذاری می کنیم. و فیت می کنیم.

به این ترتیب داده های categorical مورد نیاز به داده عددی تبدیل می شوند :



```
nyc_df.head(10)
```

	id	name	host_id	r ighbourhood_group	nei hbourhood	latitude	longitude	room_type	pr ce	minimum_nights	number_of_reviews	reviews_per
2860	1620248	Large furnished 2 bedrooms- - 30 days Minimum	2196224	2	64	40.73051	-73.98140	0	10	30	0	
34446	27316669	Bronx Apart	205820814	0	95	40.83454	-73.92751	1	10	1	0	
31407	24412104	Cozy feel at home studio	91034542	2	111	40.74408	-73.97803	1	10	5	42	
35601	28270998	Charming, bright and brand new Bed-Stuy home	647528	1	13	40.69508	-73.95164	0	10	3	5	
32810	25839759	Gigantic Sunny Room in Park Slope-Private Back...	167570251	1	189	40.66242	-73.99464	0	10	1	14	

بعد از آماده سازی داده ها، مدل سازی را شروع می کنیم. داده‌ها را به دو بخش آموزشی و آزمایشی تفکیک می کنیم. همانطور که مشخص است ٪۸۰ از داده‌ها برای بخش آموزش و ٪۲۰ برای آزمایش در نظر گرفته شده‌اند. به کمک کتابخانه sklearn و کلاس LinearRegression، محاسبات مربوط به مدل رگرسیونی در متغیرl_reg ذخیره می کنیم. و مدل را train می کنیم.

سپس کارایی مدل را ارزیابی می کنیم. سهمی که مدل رگرسیونی از تغییرات متغیر وابسته دارد را با مربع آر (R square) می‌شناسند. مقدار R^2 در بازه ۰ تا ۱ اندازه گیری می شود و هر چه به یک نزدیک تر باشیم کارایی مدل بهتر است. R^2 برابر صفر بدترین حالت و خطای مدل را نشان می دهد.

سایر معیارهای ارزیابی مانند : میانگین خطای مطلق (mean absolute error - MAE) که مجموع میانگین اختلاف مطلق بین مقادیر واقعی و پیش بینی است و خطای میانگین مربع (mean square error - MSE) که شبیه MAE است، با این تفاوت که مربع تفاوت مقادیر خروجی واقعی و پیش بینی شده را به جای استفاده از مقدار مطلق، قبل از محاسبه مجموع همه آنها محاسبه می کند.

با استفاده از کتابخانه sklearn بدون محاسبات اضافه می توان این مقدار را بدست آورد : sklearn.metrics.r2_score

Mean Squared Error: 50.36316945754358

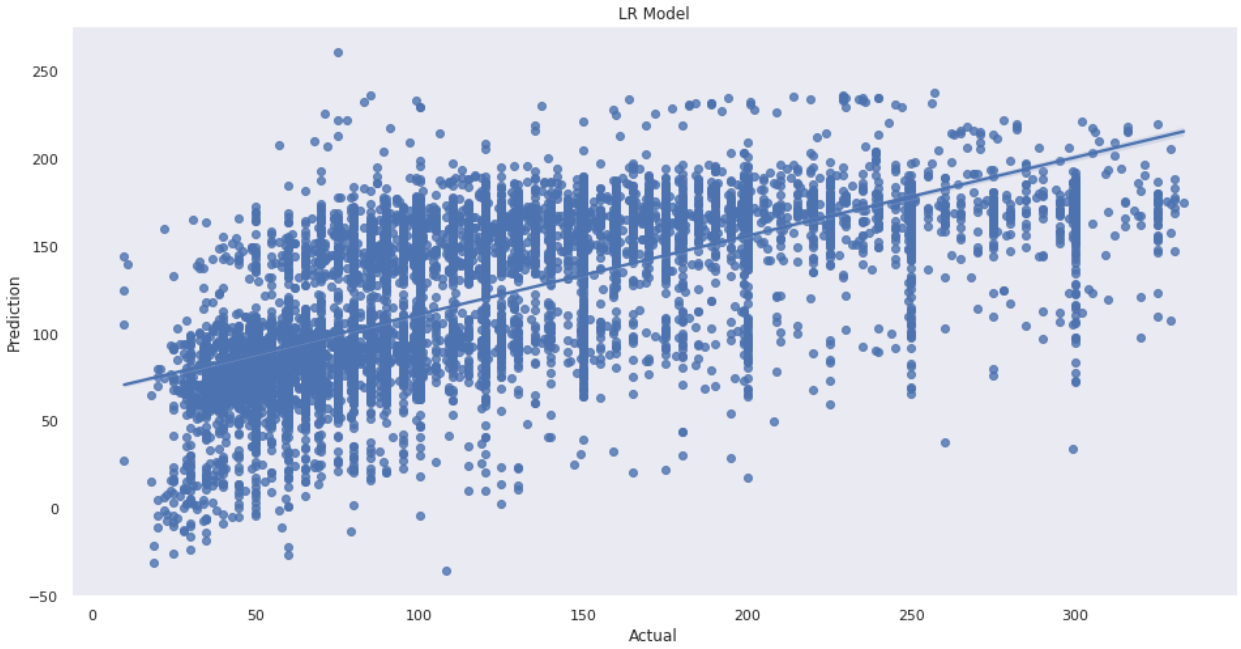
R2 Score: 44.76432513427423

Mean Absolute Error: 37.674965226176624

Mean Squareroot Error: 2536.44883780925

دقت مدل حدود ۴۴ ٪ درصد است.

در ادامه نمودار میله ای تفاوت مقادیر واقعی قیمت و مقادیر پیش بینی شده ی ۲۰ داده اول را نمایش می دهیم و با کتابخانه seaborn نمودار داده ها و مقدار فیت بودن مدل را مصور می کنیم :



چگونه می توان مدل را بهبود داد؟

همانطور که پیشتر دیدیم نمودار قیمت نرمال نبود و ممکن است روی پیش بینی موثر باشد. معمولا تبدیل لگاریتمی توزیع داده ها را نرمال یا نزدیک به نرمال می کند. پس از تبدیل کل مراحل را مجددا تکرار می کنیم و به جواب زیر می رسیم :

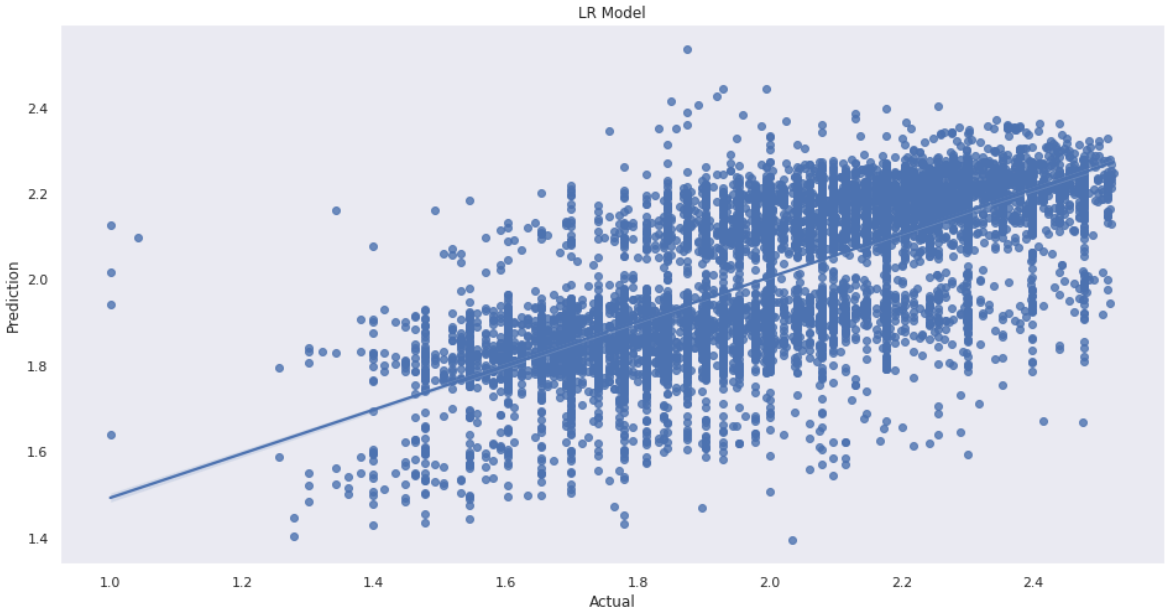
Mean Squared Error: 0.1753615772621385

R2 Score: 51.43839455556738

Mean Absolute Error: 0.13625872654730267

Mean Squareroot Error: 0.030751682779864967

دقت مدل به ۵۱ ٪ درصد رسید یعنی تنها با نرمال کردن توزیع قیمت مدل ۷ درصد بهبود یافت :


LR Model
Prediction
Actual
2.4
2.2
2.0
1.8
1.6
1.4
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4