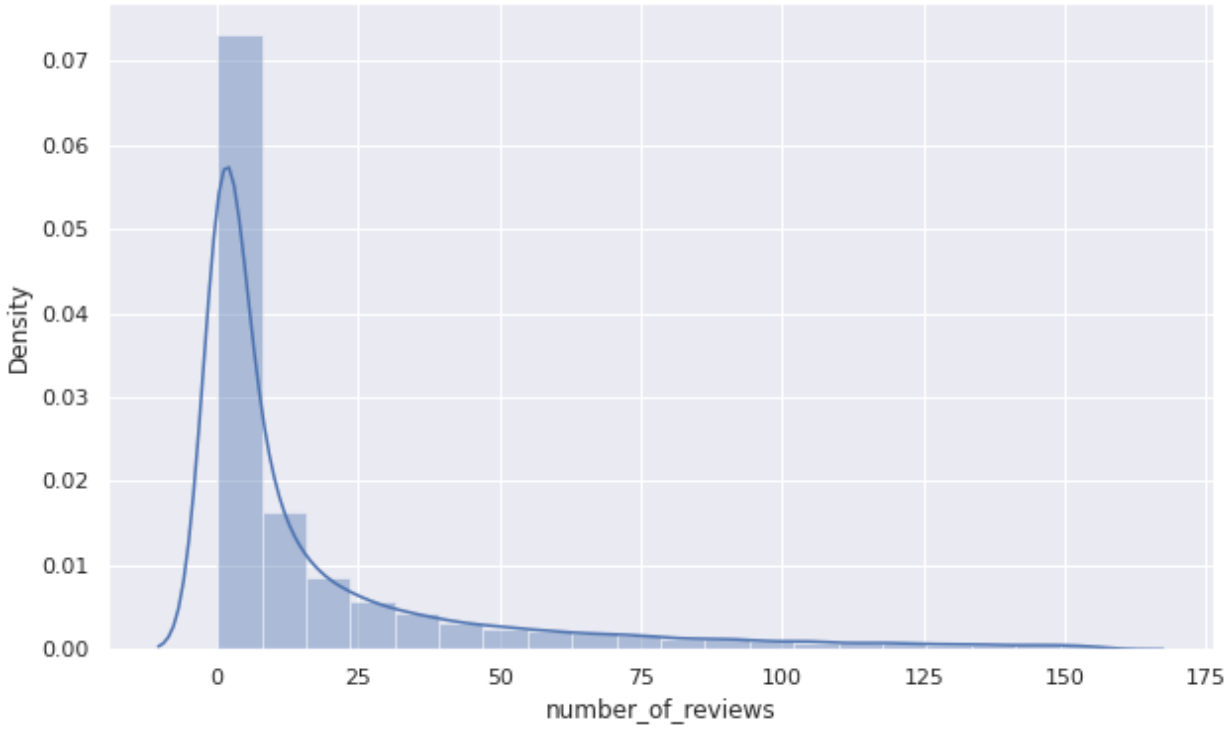
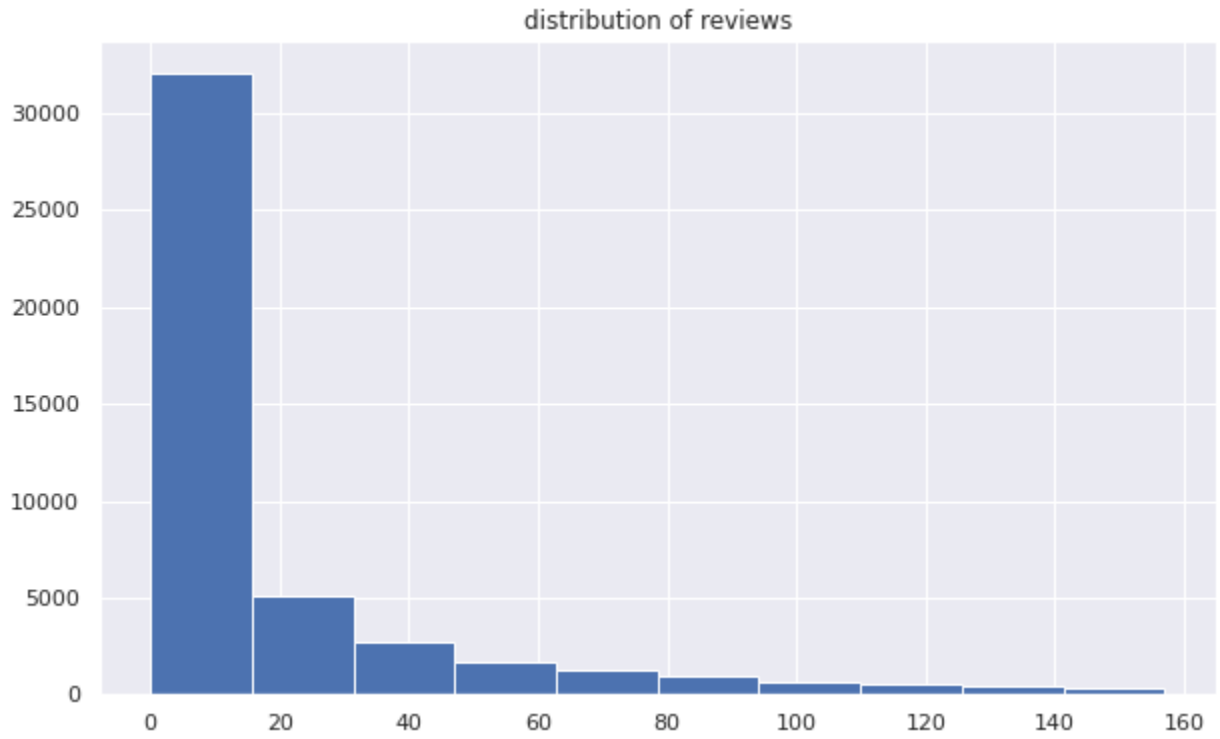
گزارش پروژه 1 علوم داده(بخش 1)

محمدرضا صیدگر-97222055

بخش اول پروژه راجع به داده هایی است از یک سامانه مسافرتی در نیویورک که اطلاعات مربوط به خانه هایی که به مسافران اجاره داده میشوند را در خود دارد.
اول از همه یه اطلاعات کلی از داده ها بدست آوردم که کل خانه ها 48895 است،اطلاعات مربوط به آخرین بازدید و بازدید خیلی از خانه ها در دسترس نیست و در آخر جنس هر متغیر که عددی است یا کتگوریکال.
در مرحله اول داده کاوی اطلاعات بسیار کلی ای از داده های عددی موجود بدست آوردم مثلا میانگین قیمت همه خانه ها حدودا 152 دلار است، حداقل شب برای اجاره میانگین 7 شب است،هر خانه ای به طور میانگین 23 تا بازدید داشته است و همینطور 112 روز از سال در دسترس و قابل book کردن هستن.
در مرحله بعدی داده های outlier را حذف کردم چون ممکن است این داده ها روی آماره ها و آنالیز کردن داده ها تاثیر منفی بگذارند اما این کار را روی عرض و طول جغرافیایی انجام ندادم چون ممکن است خانه های اطراف شهر نیویورک حذف شوند و فقط خانه های مرکز شهر باقی بمانند که به ما دید خوبی نمی دهد.درنهایت 45743 داده باقی ماندند.
از طرفی چون داده های ناموجود در دیتاست داشتیم ،داده های عددی را با میانگین آن ستون جایگزین کردم و داده های غیر عددی را با پرتکرار ترین عضو آن ستون.درنهایت دیگر هیچ داده ی null نداشتیم.
در ادامه آمدم ستون number of review را مورد بررسی قرار دادم چونکه این ستون میتوان دید خوبی به ما بدهد از بازخورد مسافران. در این ستون که دیگر داده پرت هم ندارد میانگین 18 است یعنی هرخانه میانگین 18 تا بازدید

دارد و خانه ای وجود دارد که بازدیدی نداشته و همینطور خانه ای که 157 بازدید داشته به عنوان کمترین و بیشترین بازدید و انحراف معیار هم 29 است.

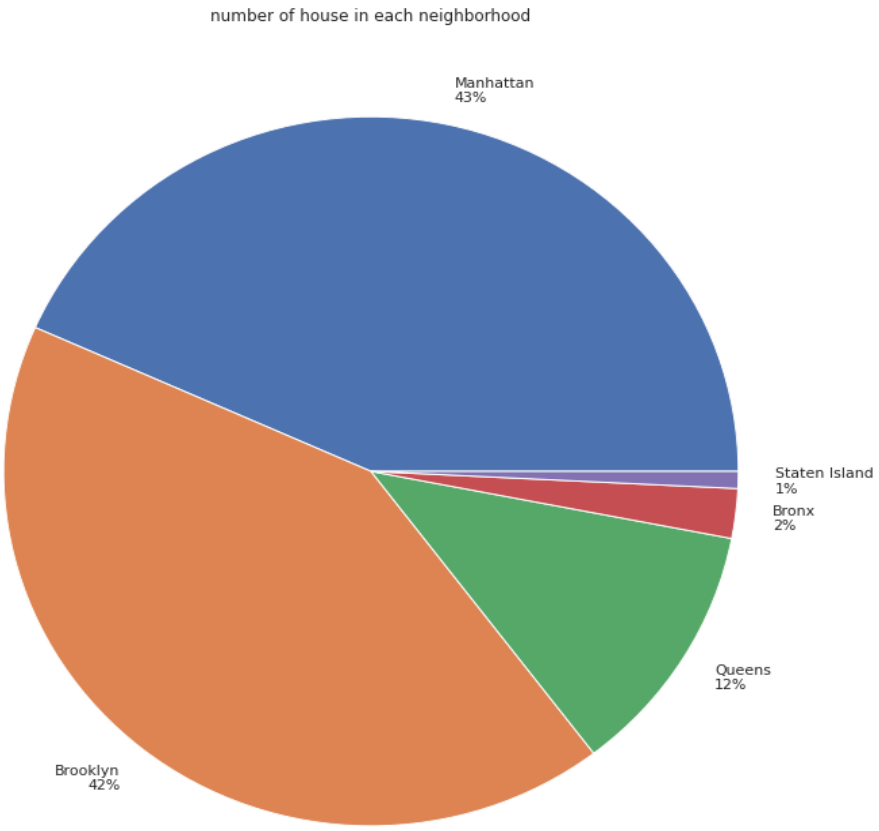




دو تصویر بالا توزیع داده های ستون بازدید ها را نشان میدهد و به شکل واضح مشخص است که این توزیع نرمال نیست.همین طور میانگین ،میانه و مد رو بررسی کردم که کلا اعداد متفاوتی بدست آمد.

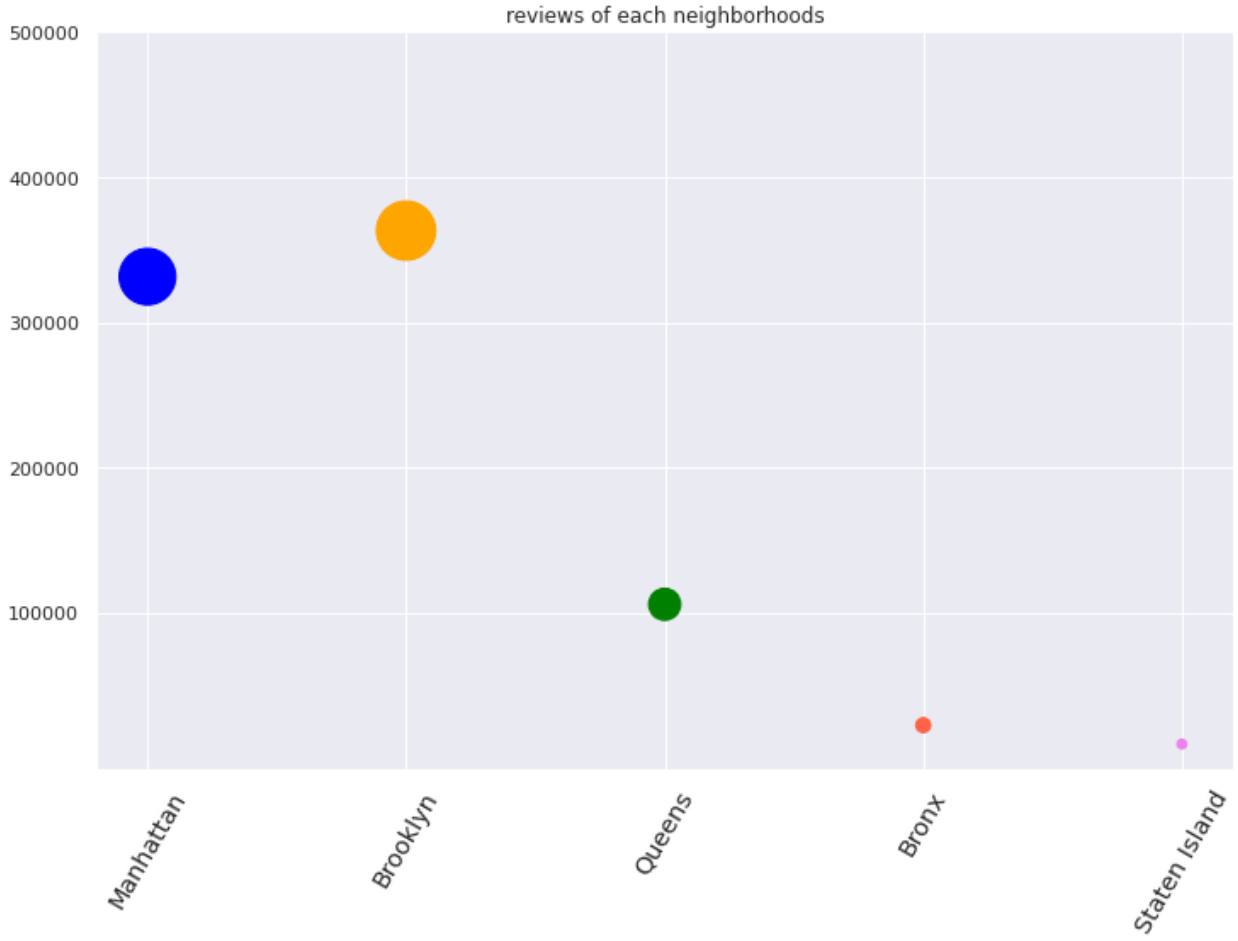
(because mean is 18.2 , median is 5 and mode is 0

بخش بعدی تعداد خانه ها را در محله های بزرگ بررسی کردم که 43% خانه های موجود از منهتن بود،42% از بروکلین،12% از کوییز،2% برونکس و 1% هم جزیره استاتن.



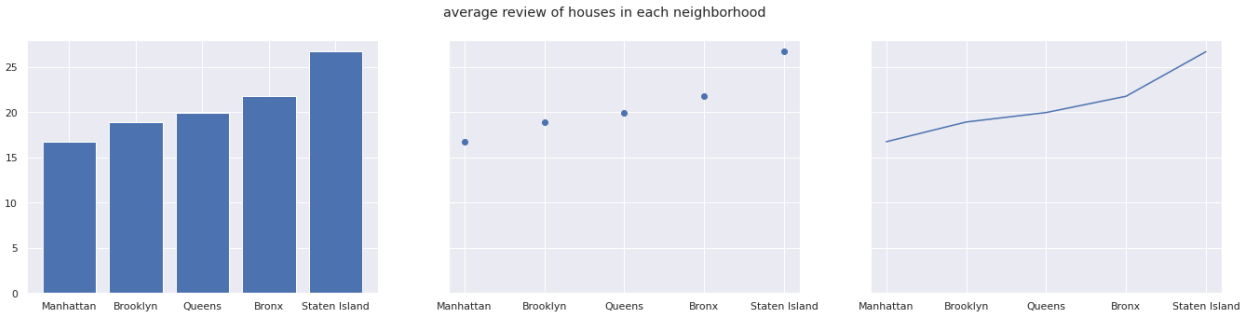
شکل بالا درک بهتری را به ما از میزان خانه های هر محله میدهد.
حال تعداد کل بازدید ها از خانه های هر محله را بررسی کردیم که طبق انتظار بازدید ها کلا از محله های منهتن و بروکلین بیشتر است و این ناشی از تعداد

خانه های آنجا است (البته اینطور میتوان برداشت کرد که چون مناطق منهتن و بروکلین در نیویورک، مناطق محبوب تری هستند اکثر خانه هم طبق انتظار مسافران آنجا ارائه میشود).

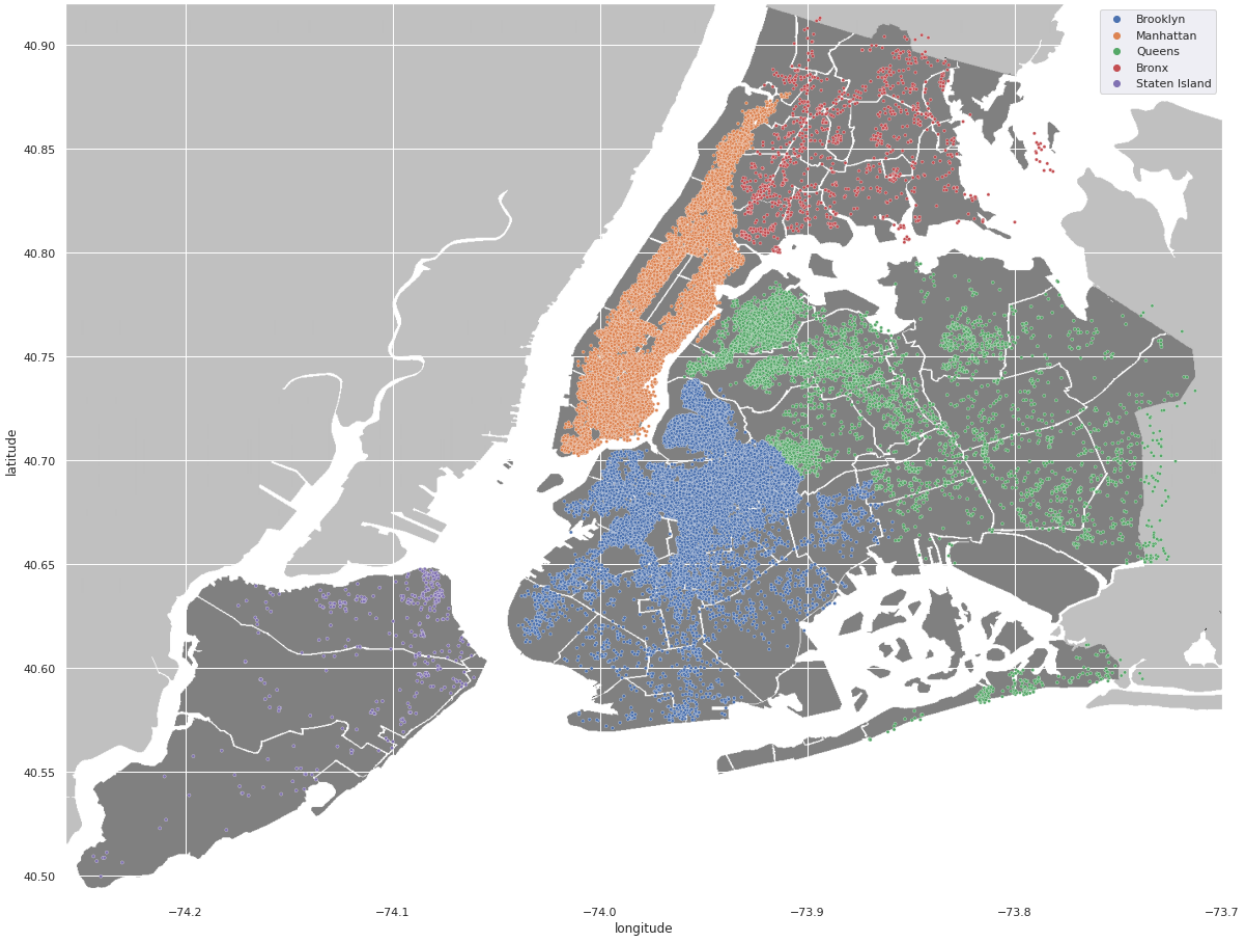


میتوان از شکل دقیق تر متوجه شویم که بروکلین و خانه های آن منطقه محبوب تر است نسبت به منهتن چونکه با توجه به اینکه تعداد خانه هایش کمتر است از آن ولی تعداد بازدید هایش بیشتر است.

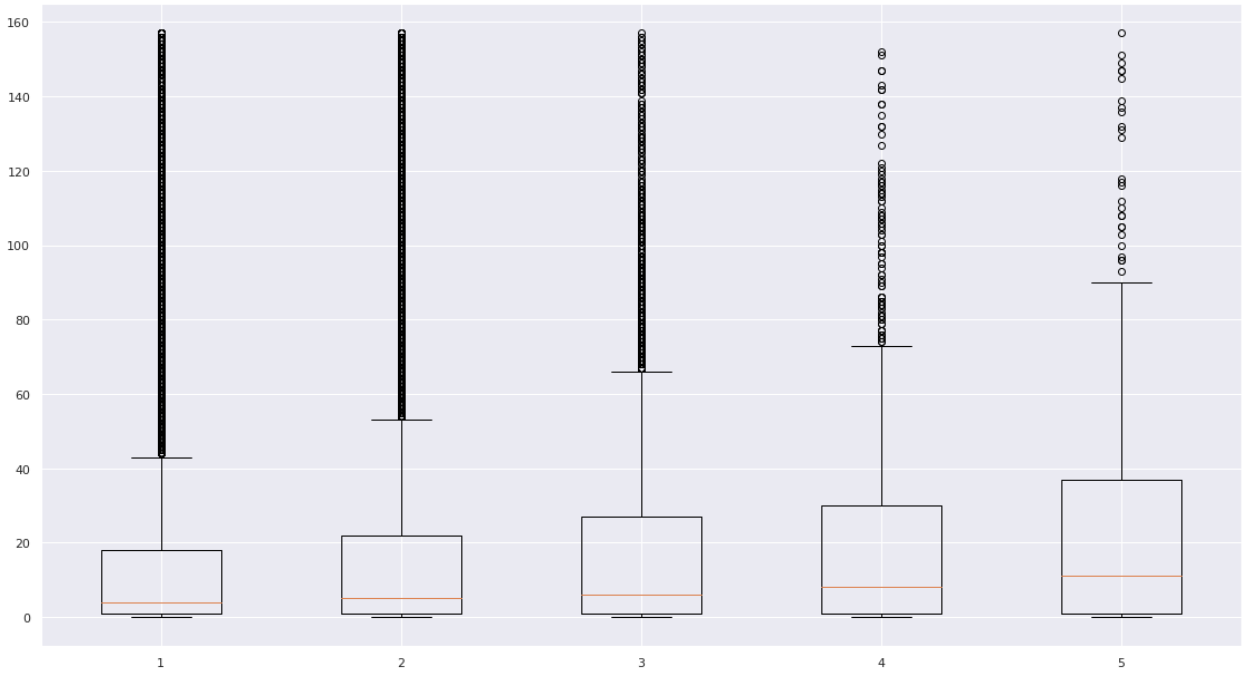
در قسمت بعدی اطلاعات بالا را تقسیم بر تعداد خانه ها کردیم که یک دید جدیدی به ما بدهد از میانگین بازدید هر خانه در منطقه های مختلف به این دلیل که کلا 2 محله منهتن و بروکلین با تعداد خانه بالا در همه آمار ها بالا هستند و نمی توان همه چیز را متوجه شد.



این نمودار ها به ما بازدید متوسط هر خانه در منطقه های مختلف را میدهد که نشان میدهد مناطقی که تعداد خانه ها کمتر است برعکس بازدید از هر خانه می تواند بیشتر باشد از طرفی بازدید از هر خانه در منهتن و بروکلین آمار کمتری نسبت به بقیه مناطق دارد و دلیلش این است که انقدر تعداد خانه ها زیاد است احتمال کمتری وجود دارد مسافری که به آن منطقه بیاید حتما از یه خانه مشخص بازدید کند.



این شکل به ما منطقه بندی شهر نیویورک با رنگ های متمایز و خانه های موجود را نشان میدهد.

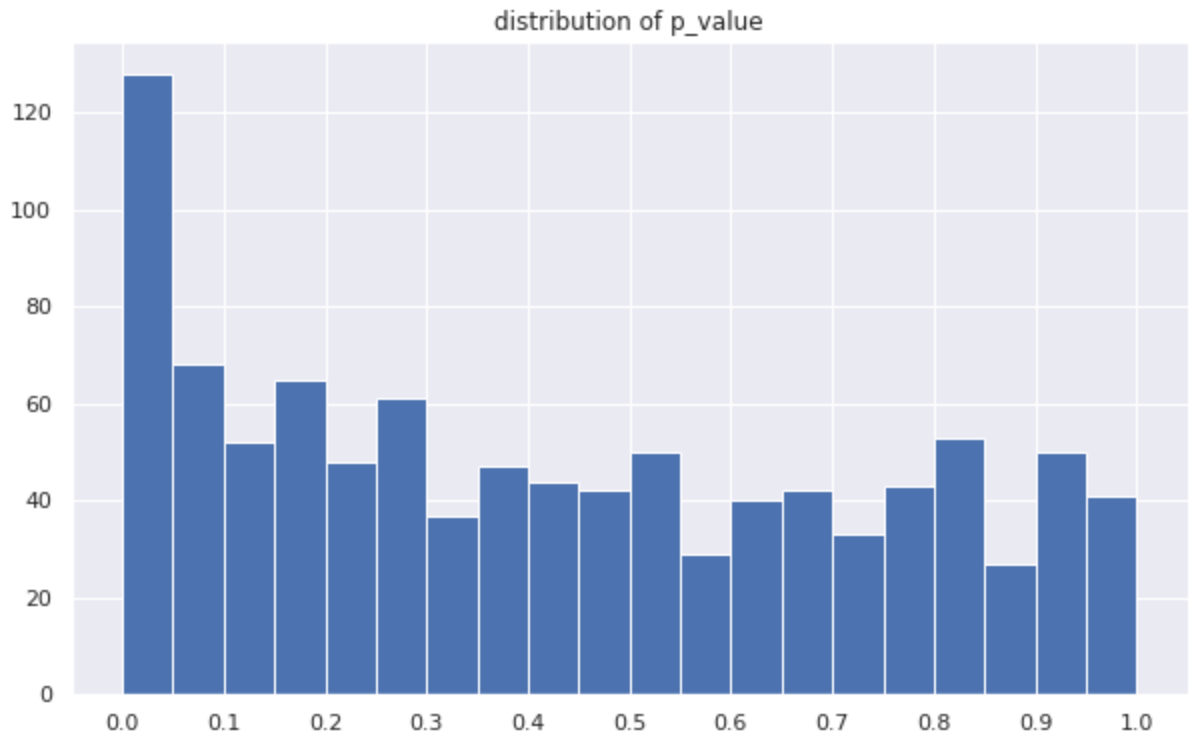


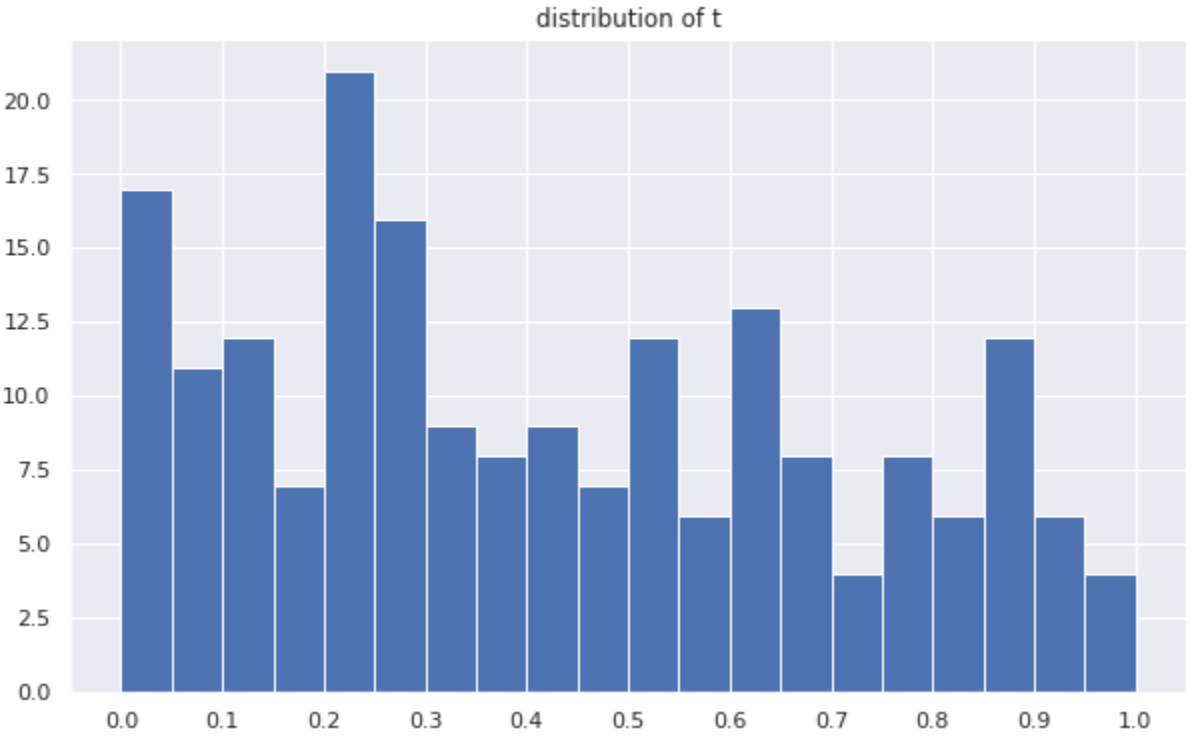
neighbourhood_group	count	mean	std	min	25%	50%	75%	max
Bronx	1034.0	21.738878	30.774726	0.0	1.0	8.0	30.0	152.0
Brooklyn	19221.0	18.897144	30.359137	0.0	1.0	5.0	22.0	157.0
Manhattan	19830.0	16.718406	28.732725	0.0	1.0	4.0	18.0	157.0
Queens	5304.0	19.941365	29.654287	0.0	1.0	6.0	27.0	157.0
Staten Island	354.0	26.661017	35.243118	0.0	1.0	11.0	37.0	157.0

و در نهایت این شکل هم توزیع بازدید های هر منطقه به ترتیب عدد(1-منهتن,2-بروکلین,3-کوییز,4-برونکس,5-جزیره استاتن)را نشان میدهد که باز مشخص است بازدید از خانه های منهتن فشرده شده در بازه 0 تا 20 است. جدول زیر نمودار هم توصیفات عددی از نمودار را نشان می دهد.

در ادامه از آماره توصیفی anova استفاده کردم برای اینکه ببینیم از نظر آماری هم آیا بازدید های این مناطق باهم متفاوت است یا خیر که در این تست p value بسیار کمتر از 0.05 بدست آمد نزدیک (10^{-25}) و آماره t=30.5 که در نتیجه یعنی این 5 منطقه از نظر تعداد بازدید باهم متفاوت هستند.

حال بررسی رو ریزتر کردیم در حد مقایسه منهتن و بروکلین که در قسمت بالا مشاهده کردیم که در کل مجموع تعداد بازدید های منهتن بیشتر بود حالا میخوایم ببینیم آیا از نظر آماری هم میتوان این نتیجه را گرفت یا خیر.
1000 بار سمپل های 200 تایی رو از دو منطقه مذکور برداشتم و ttest را روی این دو اجرا کردم که این 1000 تا اجرا نتایج متفاوتی داشت که در نمودار های زیر توزیع p_value و t را مشاهده می کنید.





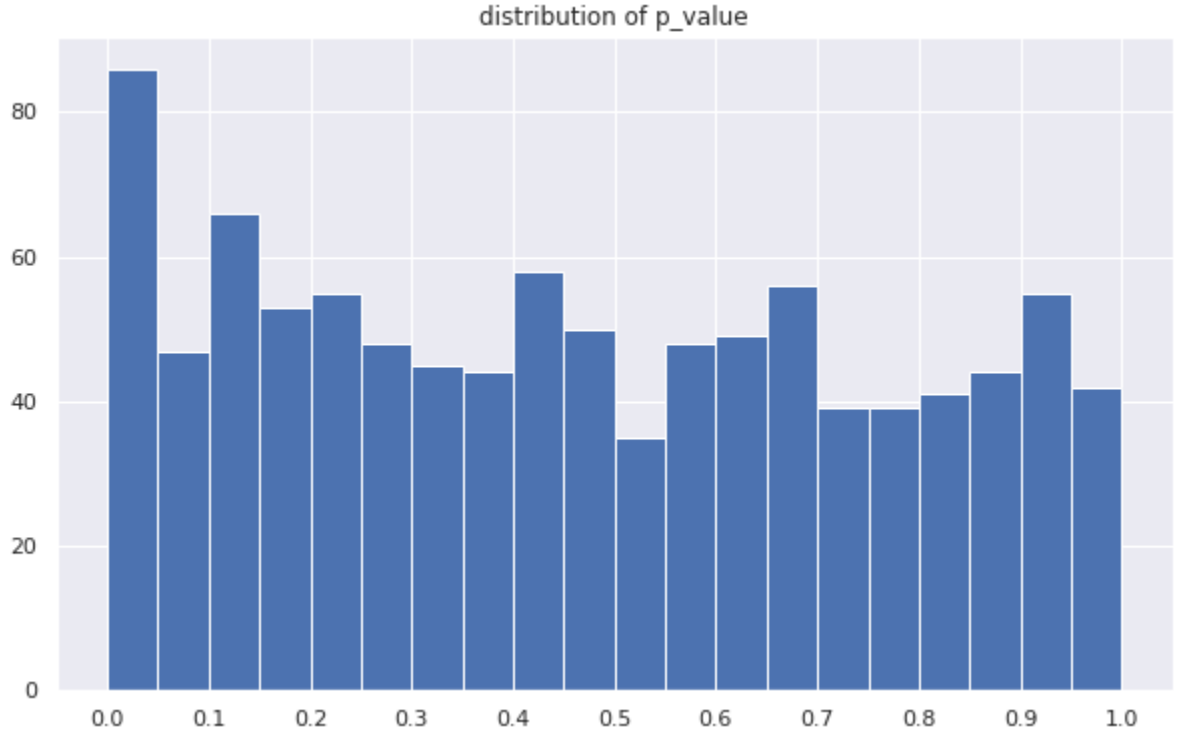
این نشان میدهد که این دو منطقه ممکن با هم تفاوتی نداشته باشند.

بار دیگر روی کل داده های این دو محله ttest را اجرا کردیم:

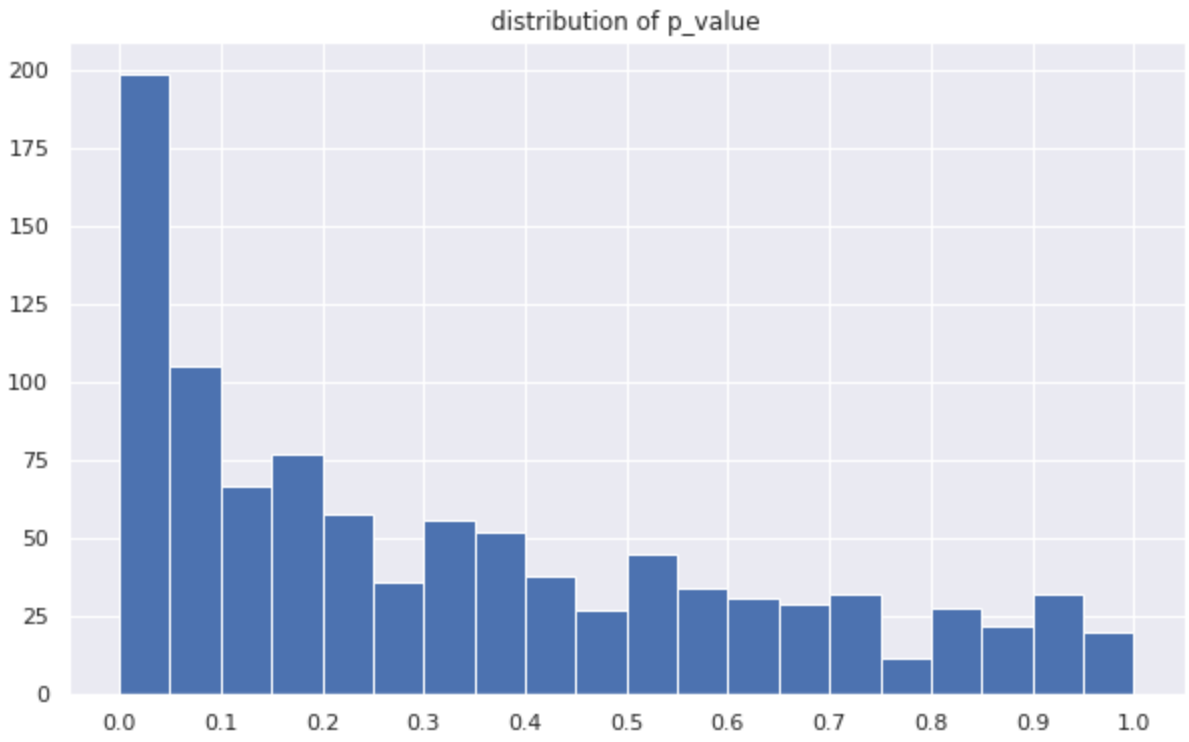
$P\_value = 10^{-13}$ و $t = -7.28$

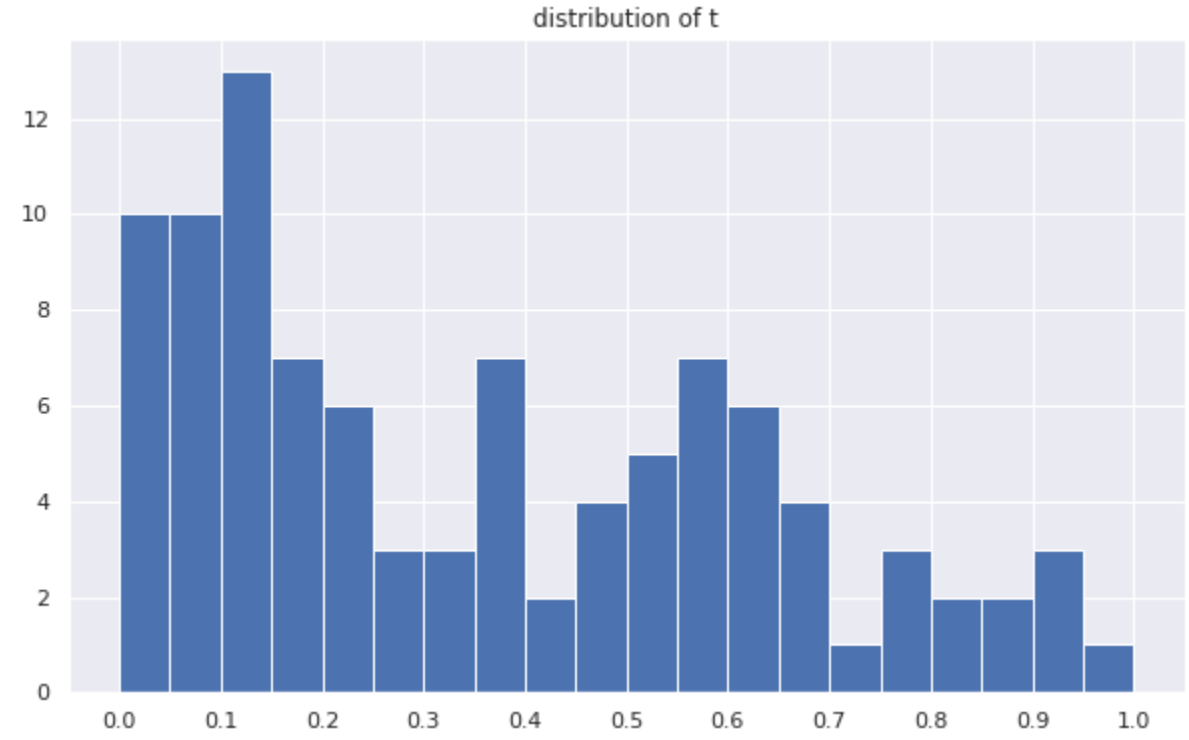
که این نشان می دهد فرض 0 رد میشود یعنی برابر نیستند اما آماره t نشان میدهد که این تفاوت خیلی هم زیاد نیست.

در قسمت بعد منطقه بخشی از منهتن به شکل random رو با بخشی از همه مناطق بررسی کردم که این دفعه بیشتر نتیجه تصادفی ای داشتیم شکل زیر:



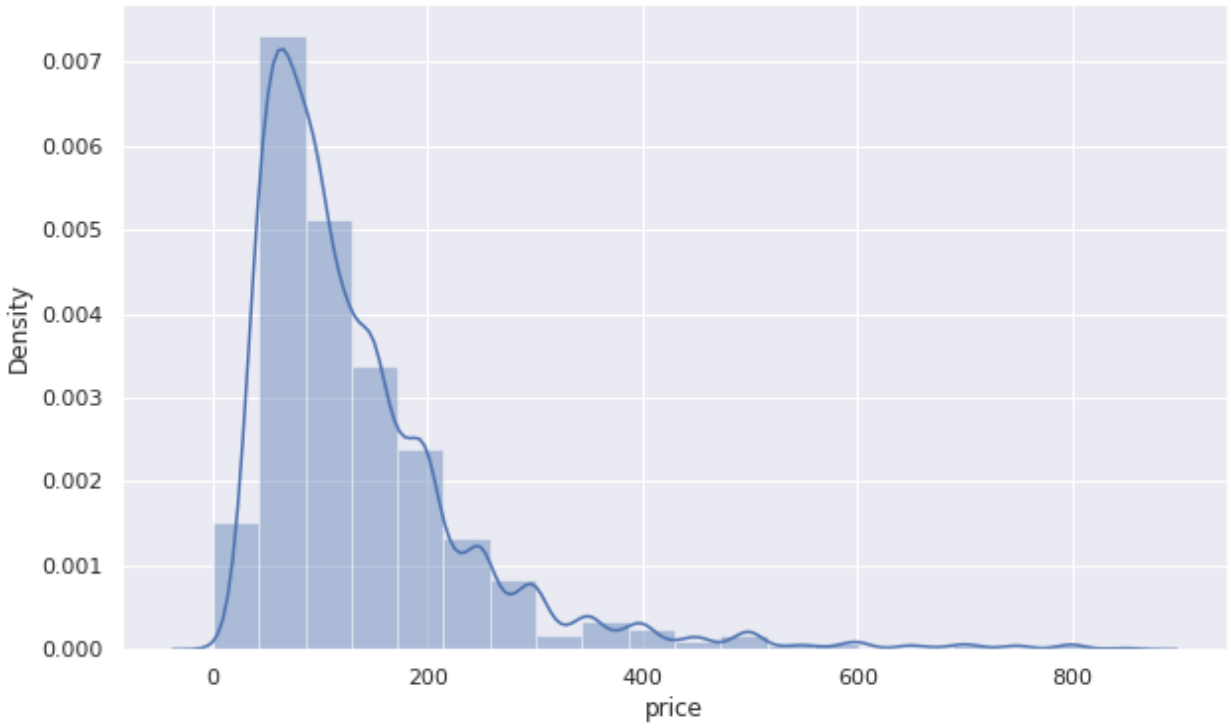
در ادامه دو محله منهتن و کوییز رو مورد بررسی قرار دادم که 1000 بار تصادفی سمپل 200 تایی گرفتم:





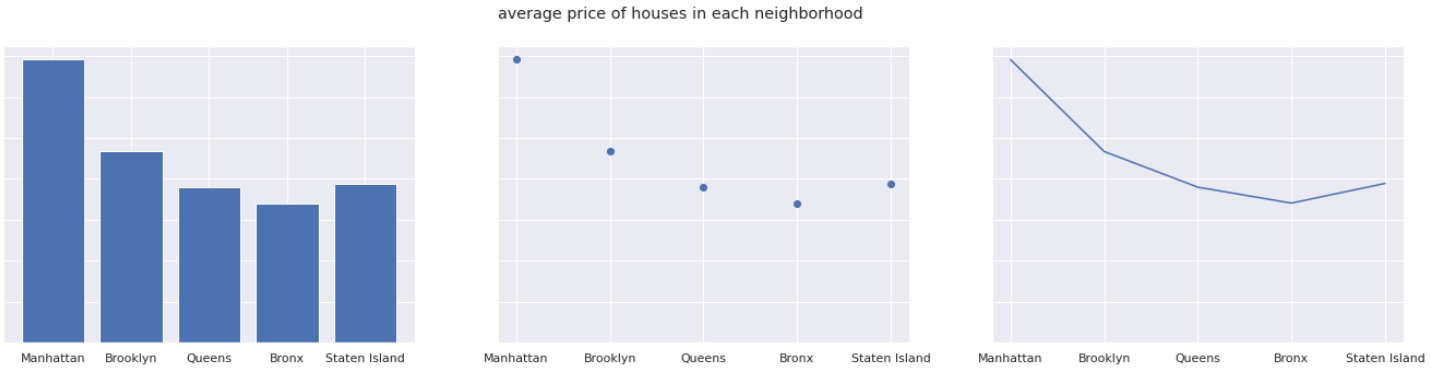
این دو نشان میدهد که احتمال تفاوت بین منهتن و کوییز بیشتر است.

حال بررسی را روی ستون قیمت یا price انجام دادم. اول از همه اطلاعات کلی بدست آوردم مثل قیمت هر خونه به طور میانگین 137 دلار است و خونه های رایگان وجود دارند و همینطور خونه ای با قیمت 860 دلار وجود دارد ، میانه قیمت هم 103 دلار است.
دوباره آمدم نمودار توزیع داده های قیمت را رسم کردم:

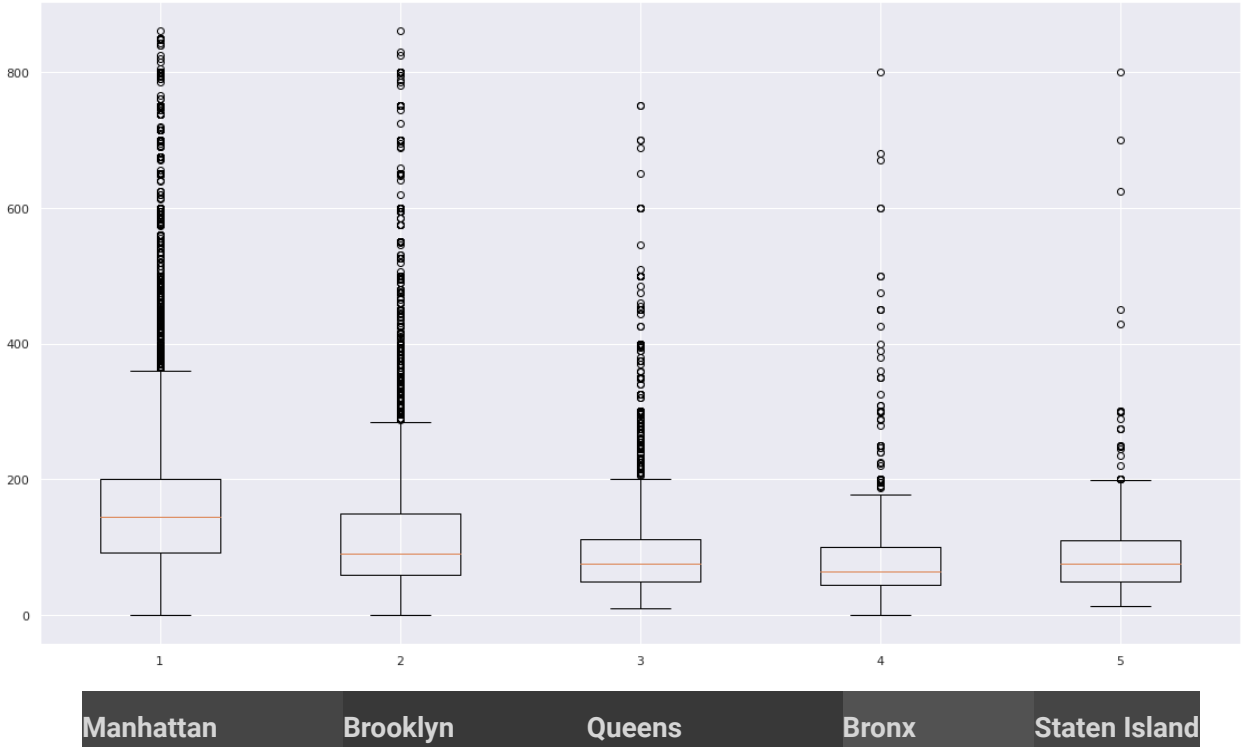


که باز هم مشخص است این داده ها توزیع نرمال ندارند.

بعد از این در این قسمت امدم قیمت متوسط هر خانه در هر منطقه را بدست آوردم:



مشخص است که به طور میانگین هر خانه در منهتن گران تر است از بروکلین و بروکلین گران تر است از جزیره استاتن و او هم گران تر از کوییز و در آخر هم برونکس که برای مسافران خانه های با قیمت کمتر را ارائه می دهد.

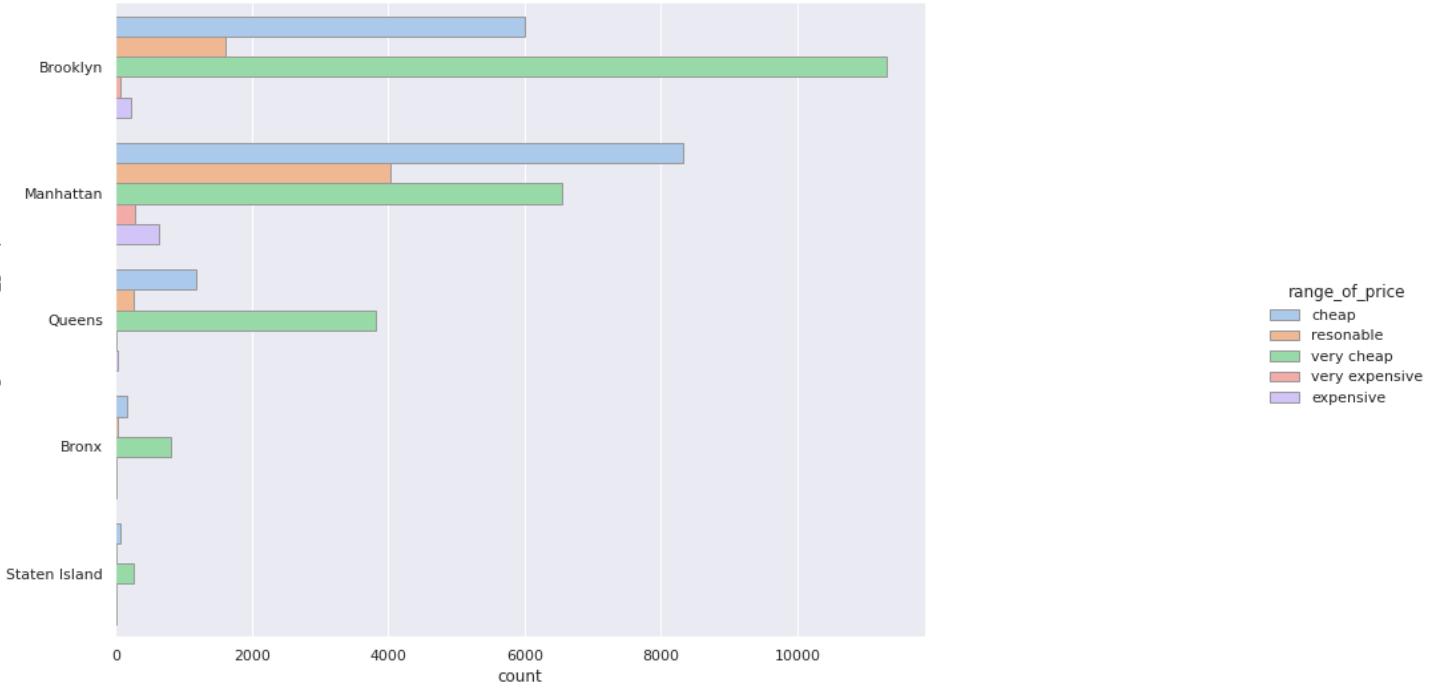


neighbourhood_group	count	mean	std	min	25%	50%	75%	max
Bronx	1034.0	85.219536	74.140347	0.0	45.0	65.0	99.75	800.0
Brooklyn	19221.0	116.709745	89.185967	0.0	60.0	90.0	150.00	860.0
Manhattan	19830.0	172.746193	121.769897	0.0	92.0	145.0	200.00	860.0
Queens	5304.0	94.949284	70.214812	10.0	50.0	75.0	111.25	750.0
Staten Island	354.0	97.225989	85.473153	13.0	50.0	75.0	109.75	800.0

شکل بالا نمودار های جعبه ای هر منطقه را بر روی قیمت ارائه می دهد و جدول هم توصیفات عددی نمودار است.

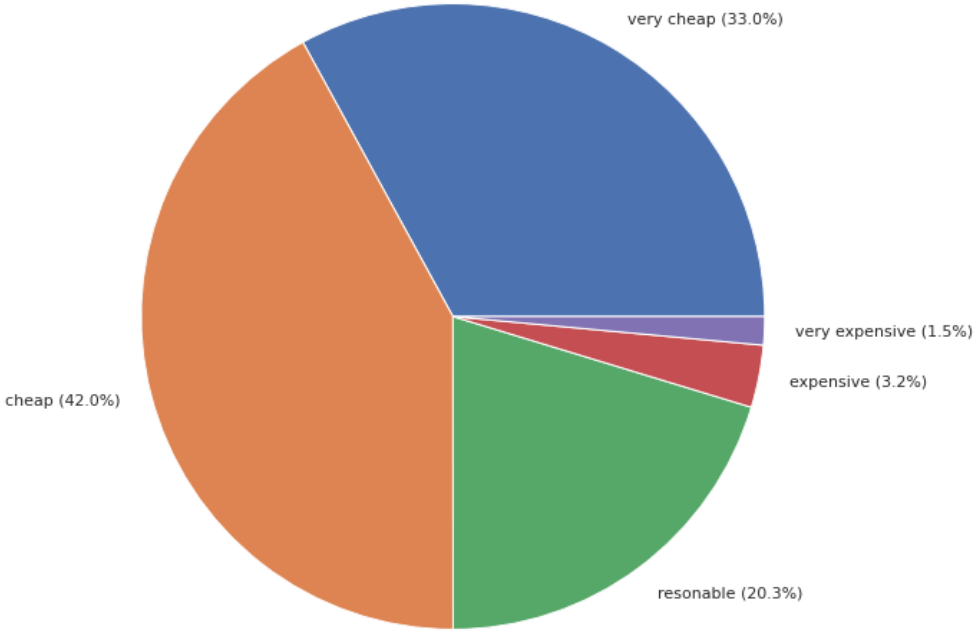
تو قسمت بعد یک ستون جدید به اسم range of price برای خانه ها در نظر گرفتم که همان قیمت است فقط کتگوریکال است و به range آن قیمت که ارزان یا گران است اشاره می کند.

نمودار زیر دید کلی نسبت به نوع قیمتی خانه های در مناطق مختلف را به ما می دهد.

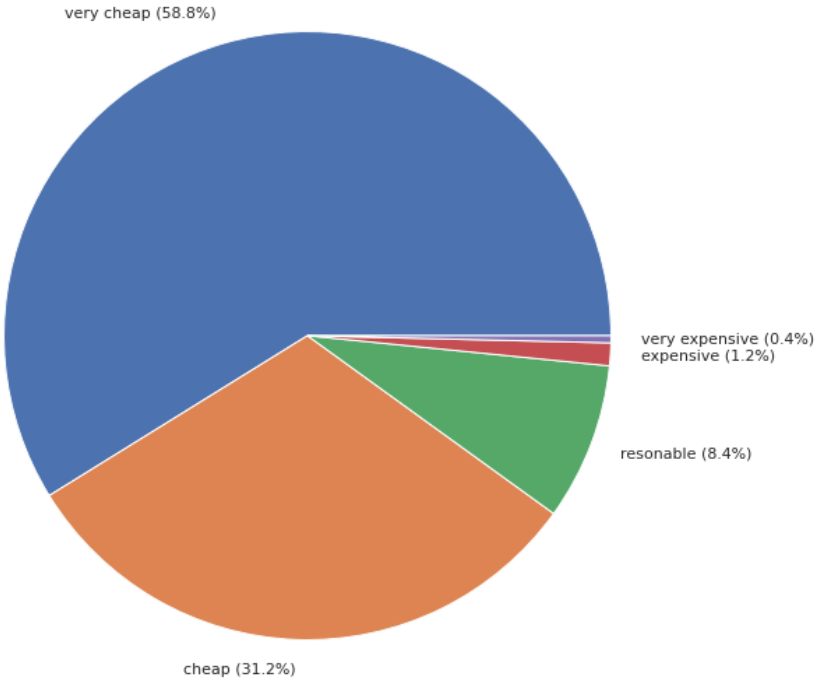


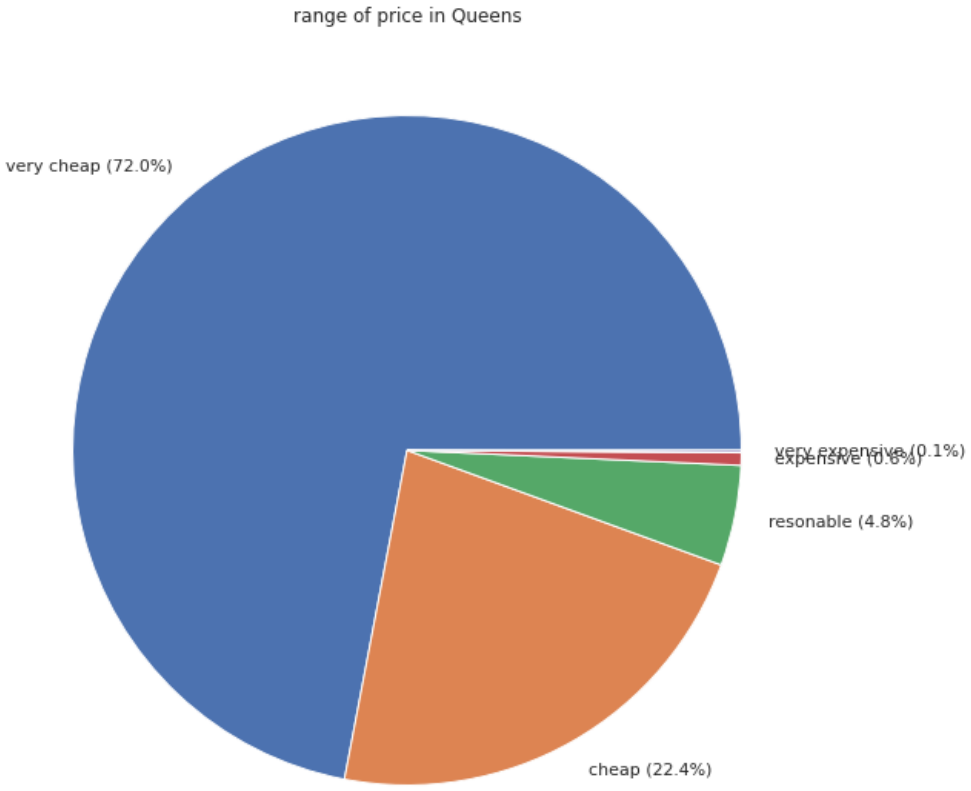
اما نمودار های زیر هر منطقه را به شکل جدا بررسی می کند:

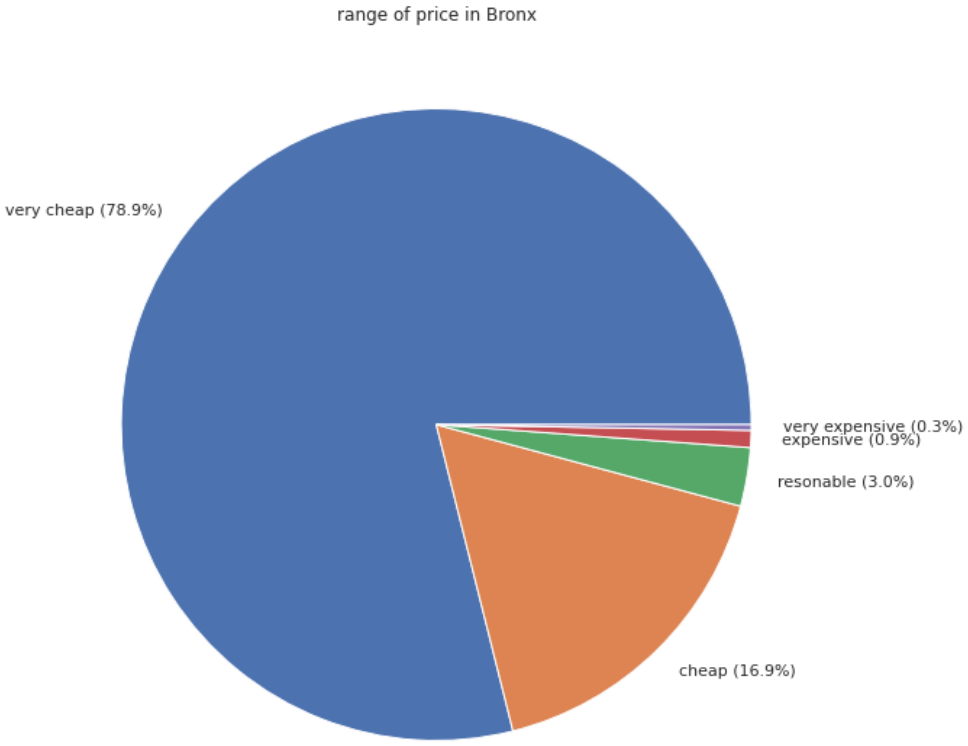
range of price in Manhattan

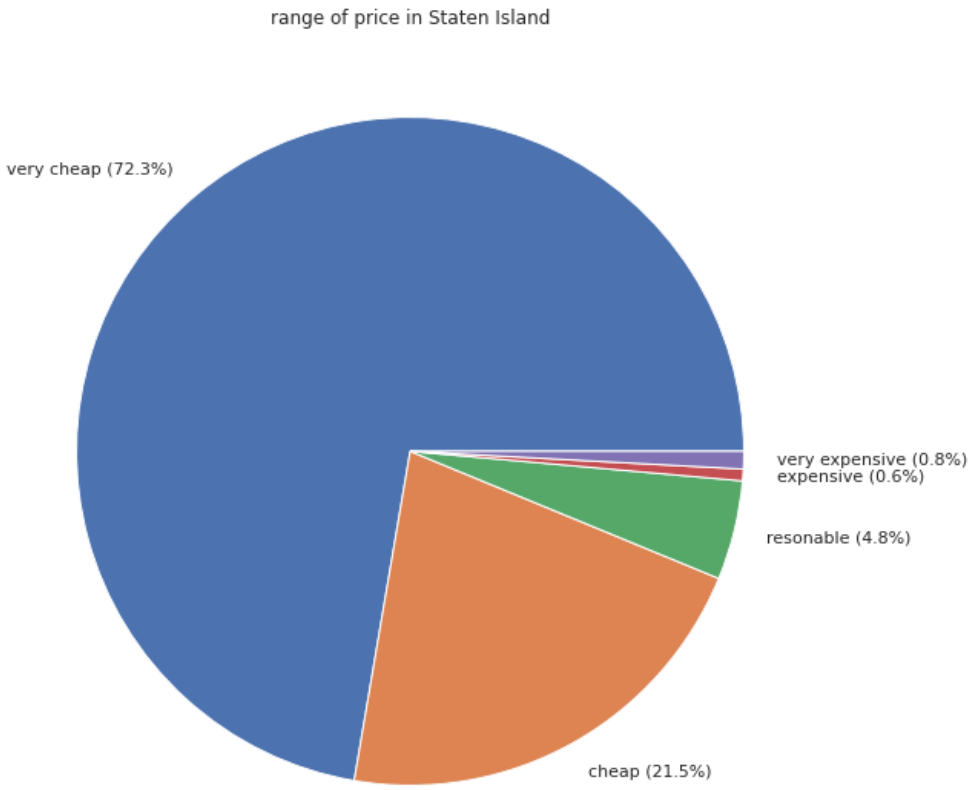


range of price in Brooklyn

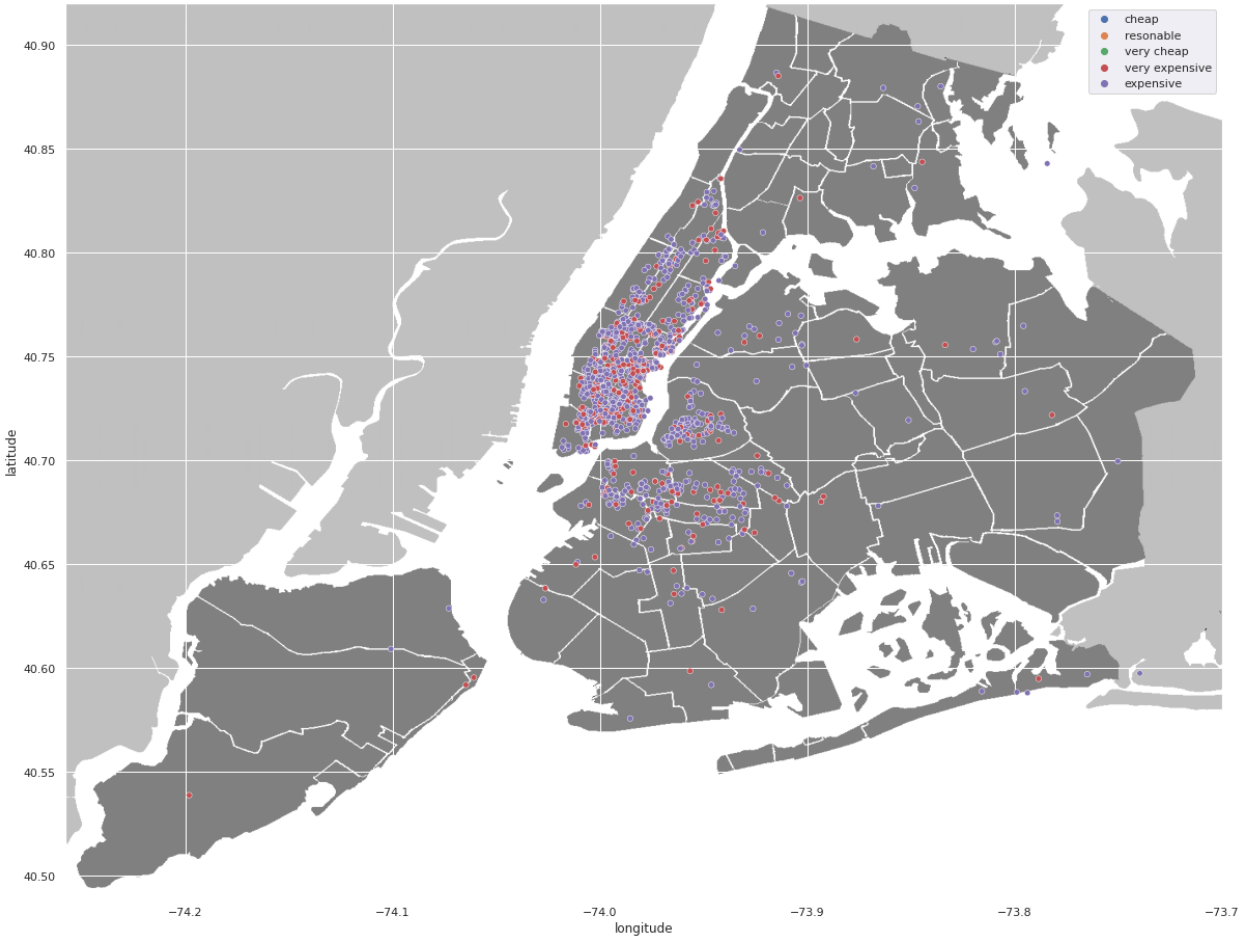



range of price in Queens
very cheap (72.0%)
very expensive (0.1%)
expensive (0.8%)
resonable (4.8%)
cheap (22.4%)

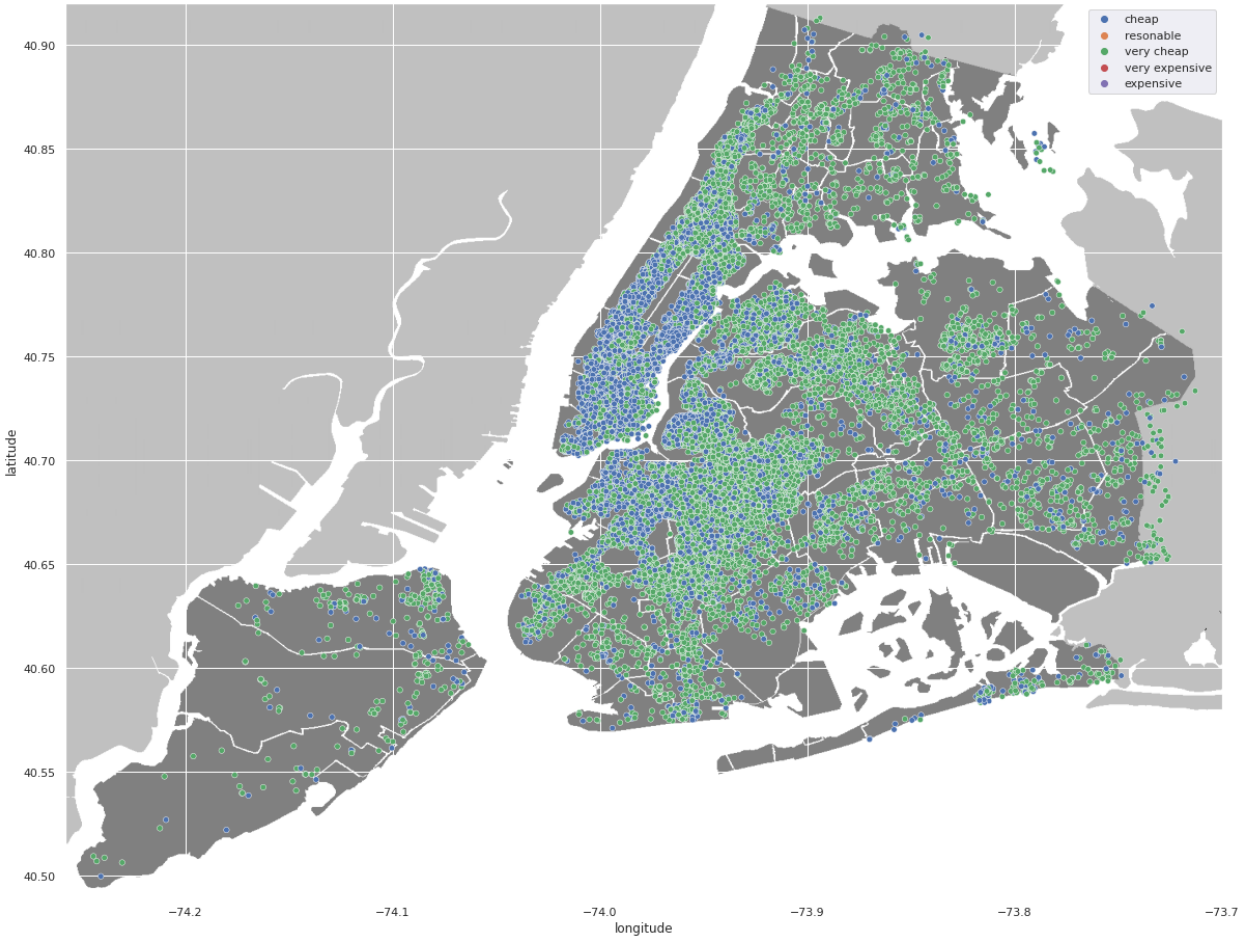

range of price in Bronx
very cheap (78.9%)
very expensive (0.3%)
expensive (0.9%)
resonable (3.0%)
cheap (16.9%)



از تفاوت های محسوس و قابل توجه میتوان گفت که در بروکلین تعداد خانه های خیلی ارزان بیشتر است از منهتن و از طرفی تعداد خانه های گران و خیلی گران منهتن از همه جا ها بیشتر است.



این نقشه هم میتواند دیدی را نسبت به خانه های گران و بسیار گران به ما بدهد که مشخصا اکثرا در منهتن و بروکلین هستند.



این نقشه هم خانه های ارزان و بسیار ارزان را برای ما مشخص کرده است که نتایج قبلی که گرفتیم اینجا هم کاملا قابل رویت است.

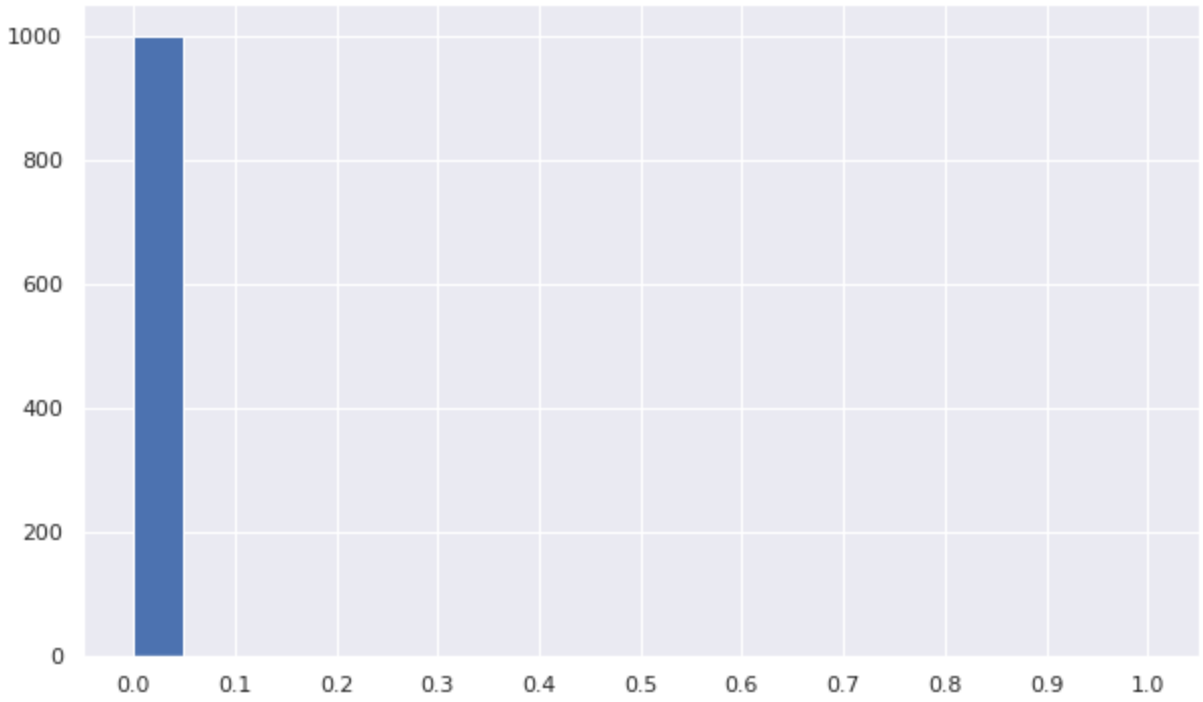
در ادامه تست های آماری انجام دادیم:

تست anova رو هر 5 منطقه :

(F_onewayResult(statistic=1091.5493275441468, p_value=0.0

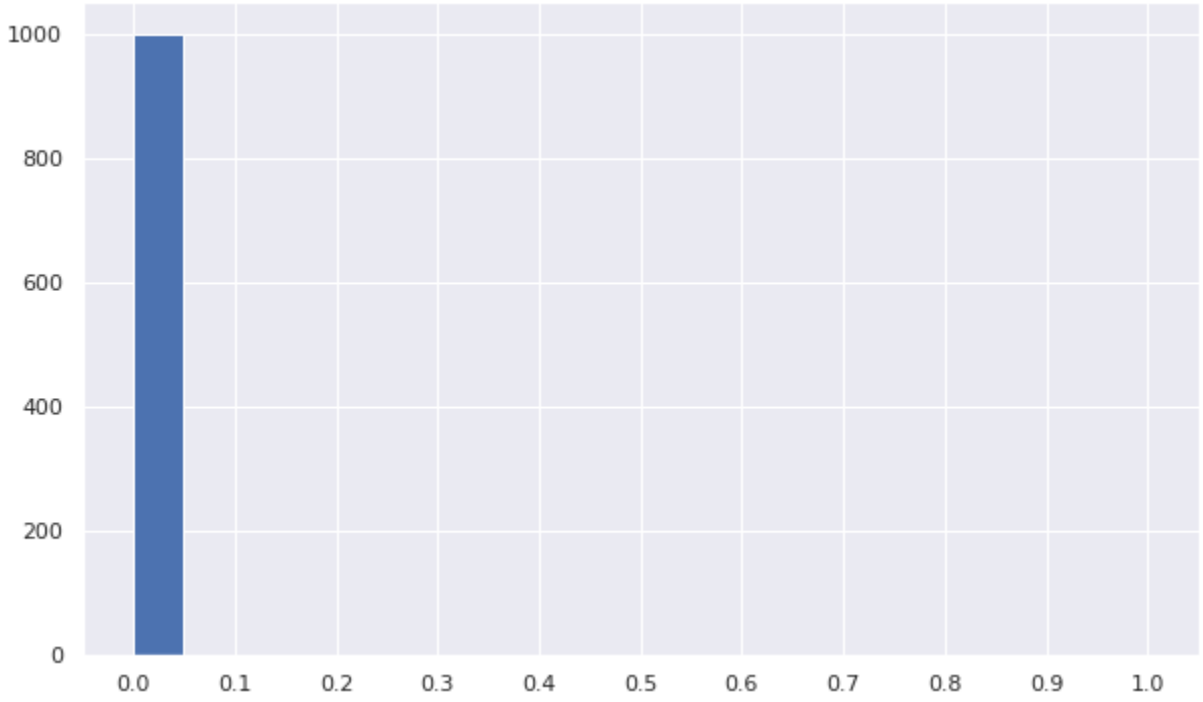
که یعنی باهم بسیار متفاوت هستند

تست ttest روی منهتن و بروکلین:



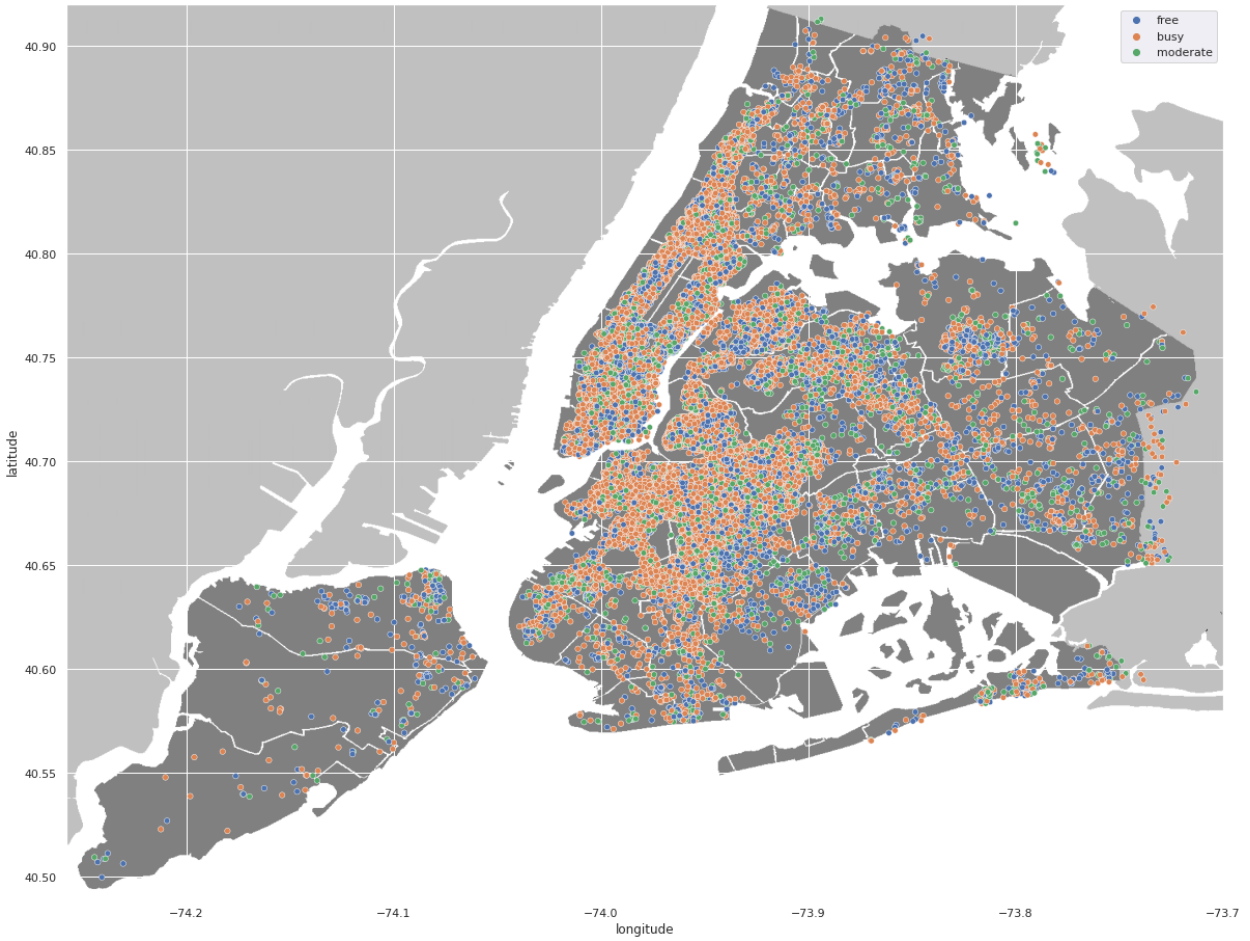
که همه 1000 بار p_value کمتر از 0.05 بوده.

تست ttest روی منهتن و کوییز:

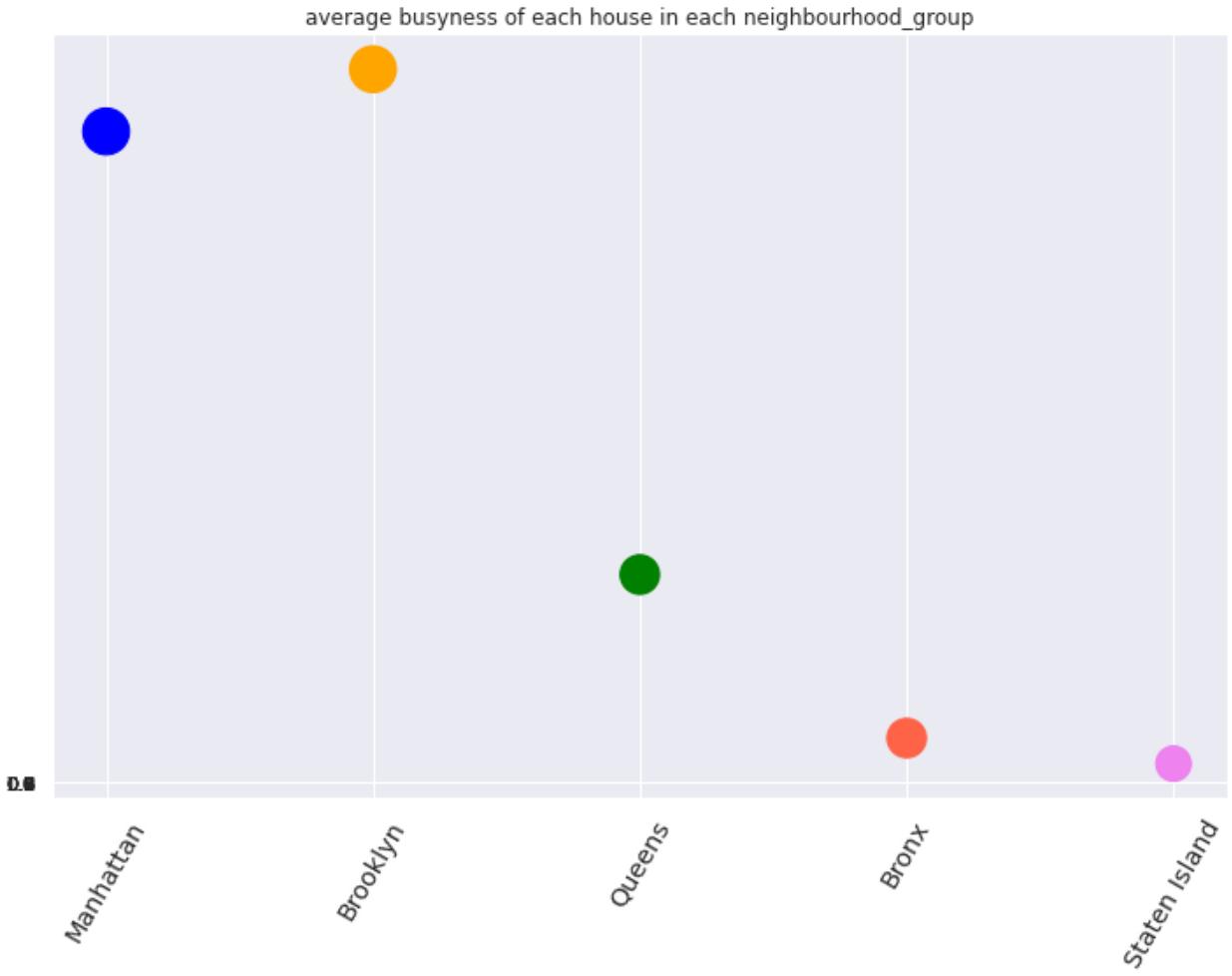


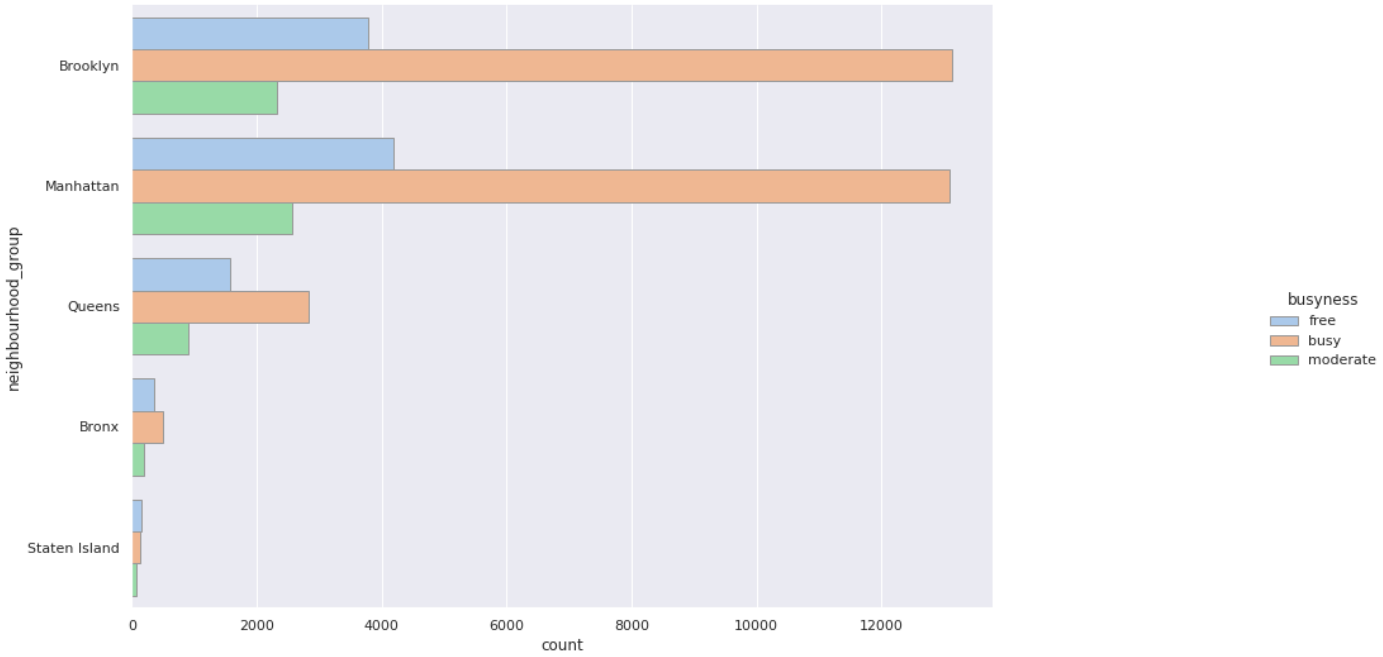
که همه 1000 بار p_value کمتر از 0.05 بوده.

در بخش بعدی شلوغ بودن خانه ها را بررسی کردم که میتوان از ستون availability_365 استفاده کرد چون آن خانه ای که خلوت است تعداد روز های بیشتری در دسترس است و خانه هایی که شلوغ اند تعداد روز کمتر در دسترس هستند.
باز هم مثل بالا در دسترس بودن را به 3 بخش free و moderate و busy تقسیم کردیم و این ستون رو که شمال 3 دسته است به داده ها اضافه کردیم تا بر اساس این تصمیم گیری کنیم.



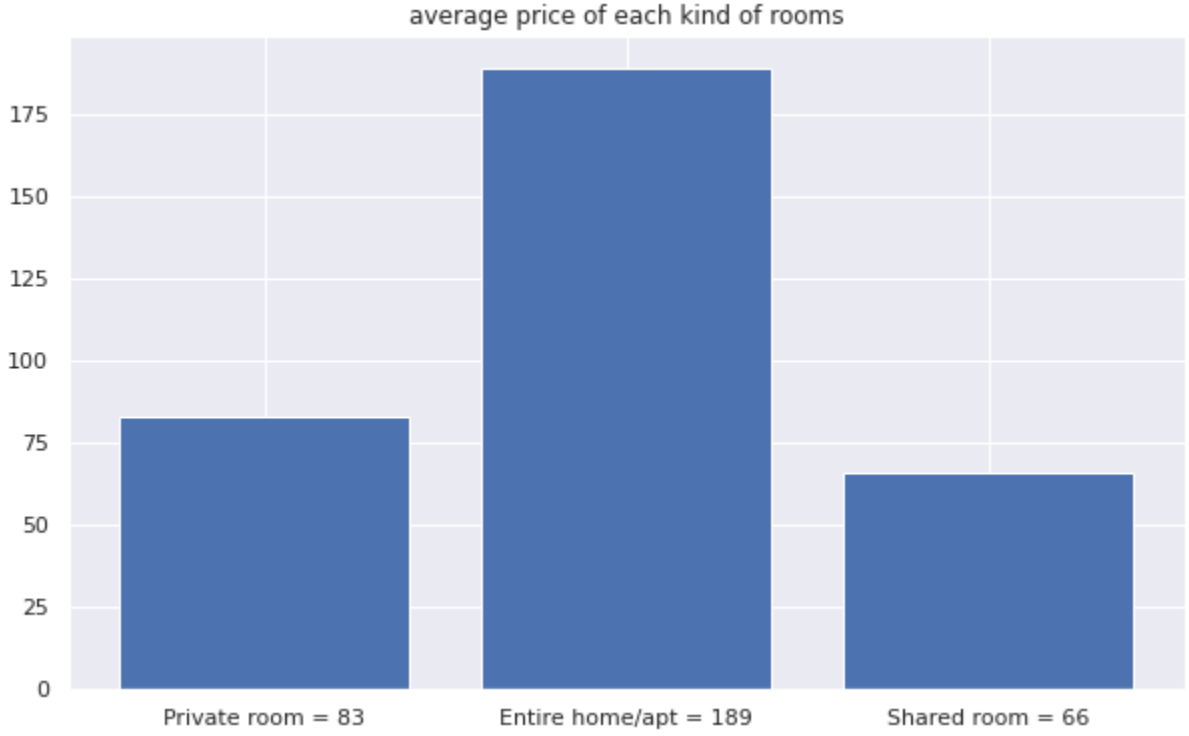
این طور که از این شکل می بینیم باز هم بیشتر خانه های مهتن و بروکلین کمتر در دسترس هستند و سرشان شلوغ تر است.


average busyness of each house in each neighbourhood_group
Manhattan
Brooklyn
Queens
Bronx
Staten Island


Brooklyn
Manhattan
Queens
Bronx
Staten Island
neighbourhood_group
0
2000
4000
6000
8000
10000
12000
count
busyness
free
busy
moderate

این نمودارها هم به ما دقیق نشان میدهد که هر خانه در هر منطقه چقدر احتمال دارد که busy باشد و سرش شلوغ باشد.

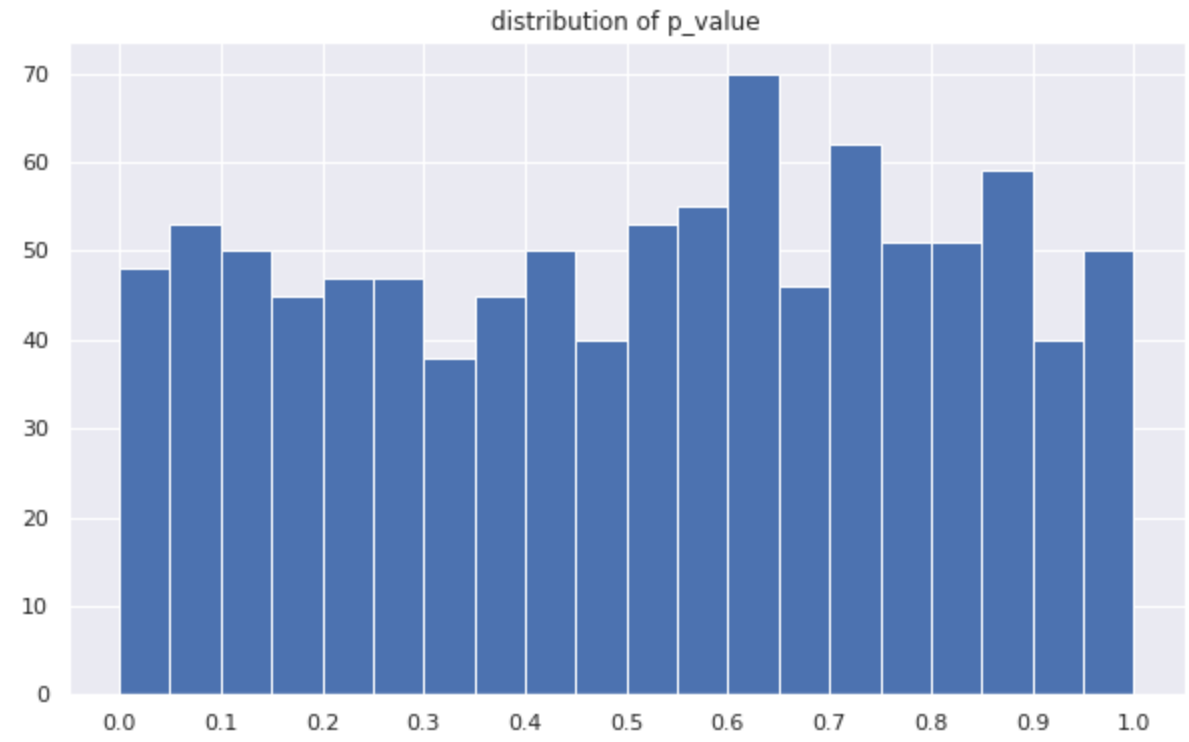
در بحث بعدی نوع هر خانه را با قیمت آن بررسی کردیم

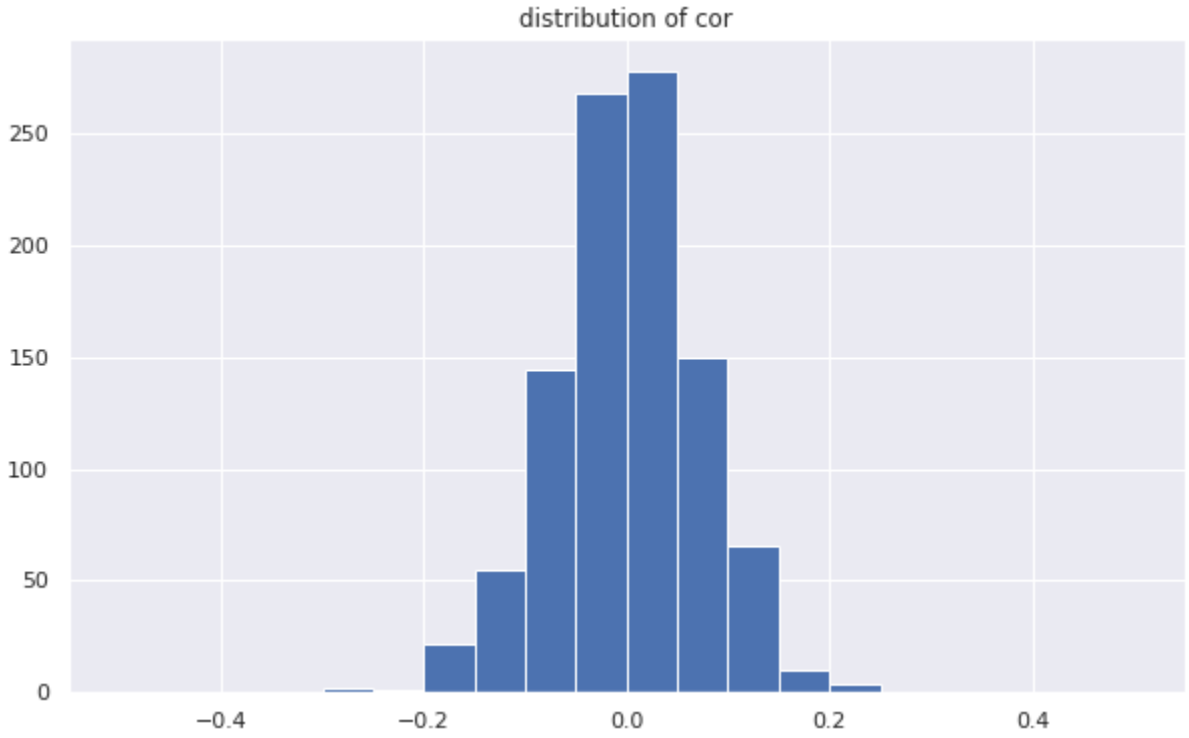


نمودار بالا نشان می دهد که نوع اتاق به طور متوسط چه قیمتی می تواند باشد. که یعنی آنهایی که به شکل خانه خصوصی و آپارتمانی هستند قیمت بالاتر و بعد بعد اتاق ها خصوصی و در آخر هم اتاق ها اشتراکی هم کمترین قیمت ها را دارند و برای مسافرانی که قیمت کمتر برایشان مهم است این گزینه آخر مورد مناسب تری است.

این هم نمودار تعداد بازدید از هر نوع اتاق به طور متوسط.

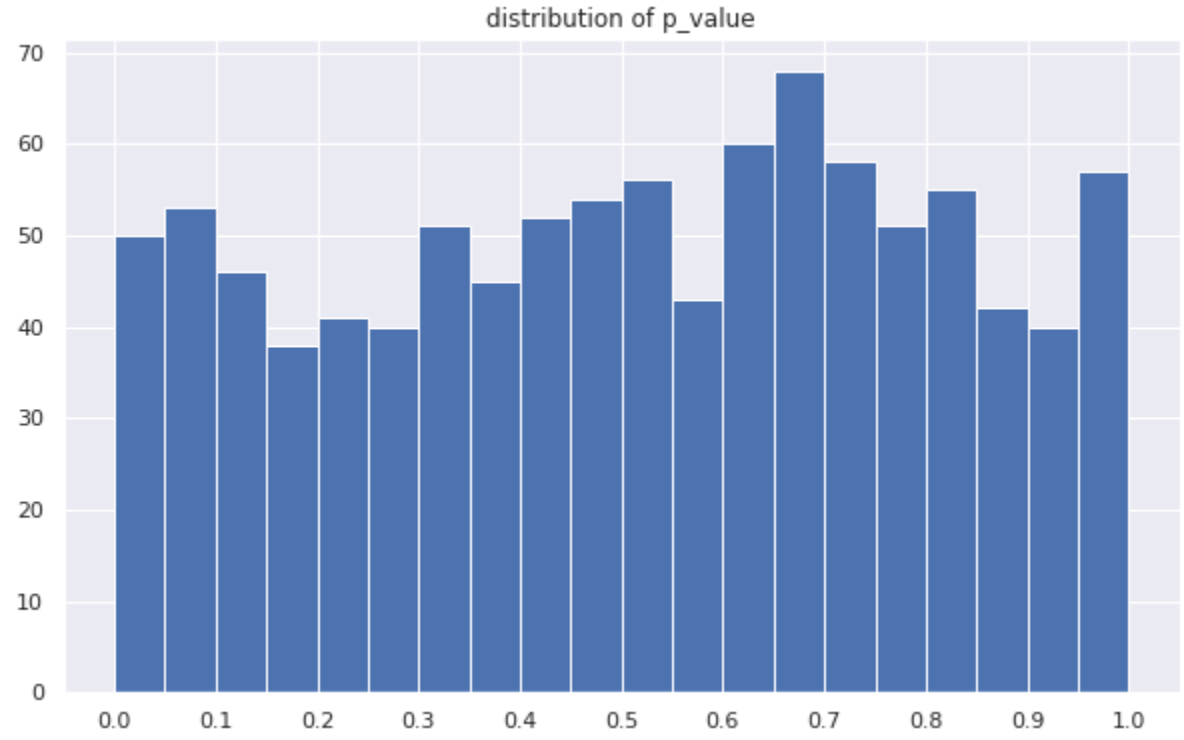
در ادامه تست های کورولیشن بین متغیر های عددی زدم:
بین قیمت و تعداد بازدید:

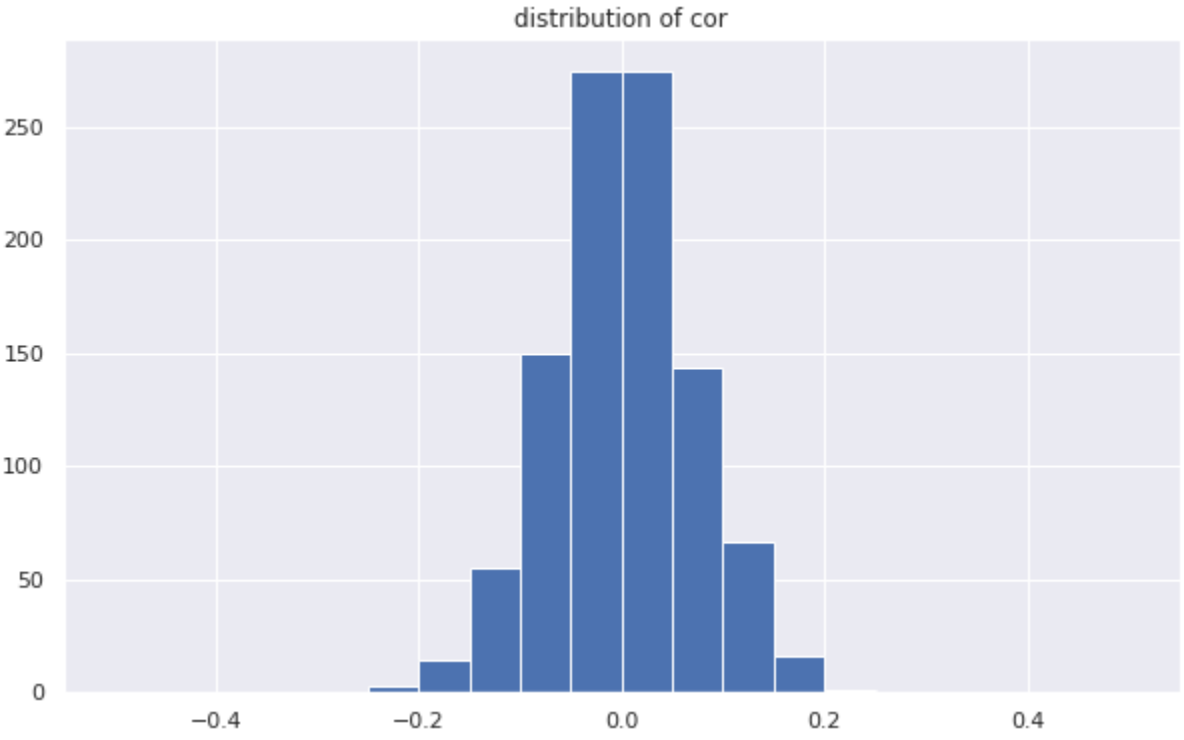




این 2 توزیع نشان می دهد که همبستگی خاصی بین این 2 متغیر وجود ندارد.

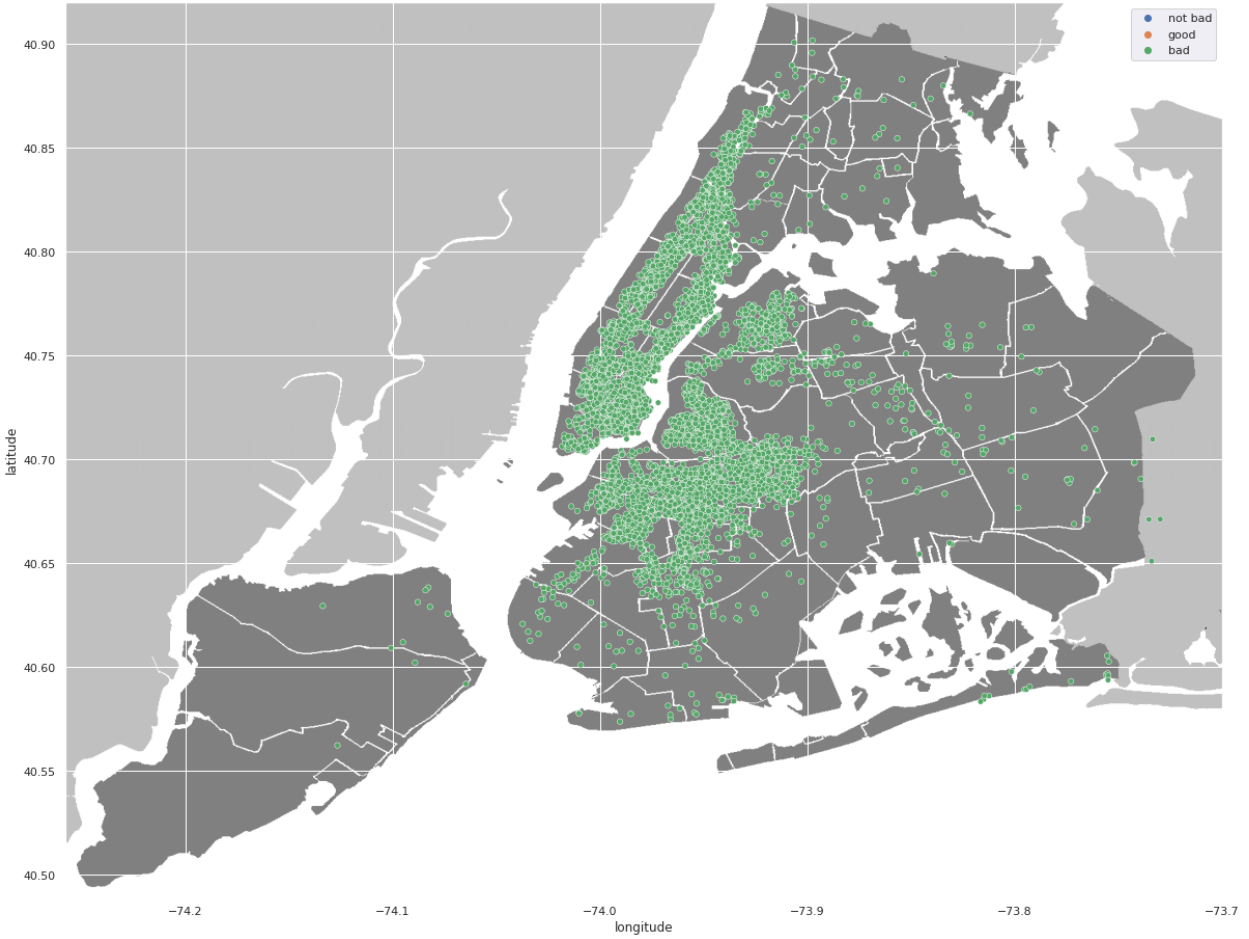
بین در دسترس بودن و تعداد بازدید:





این 2 توزیع نشان می دهد که همبستگی خاصی بین این 2 متغیر وجود ندارد.

در بخش بعدی آخرین بازدیدی که از هر خانه شد را بررسی کردیم و براساس سال آخرین بازدید آنهایی که 2019 بازدید داشتند برچسب good گرفتند ،2018 برچسب not bad ، گرفتند و 2017 برچسب bad گرفتند با این منطق که خانه ای که آخرین بازدید از آن مربوط به 2 سال قبل است احتمالا خانه خوبی نخواهد بود.



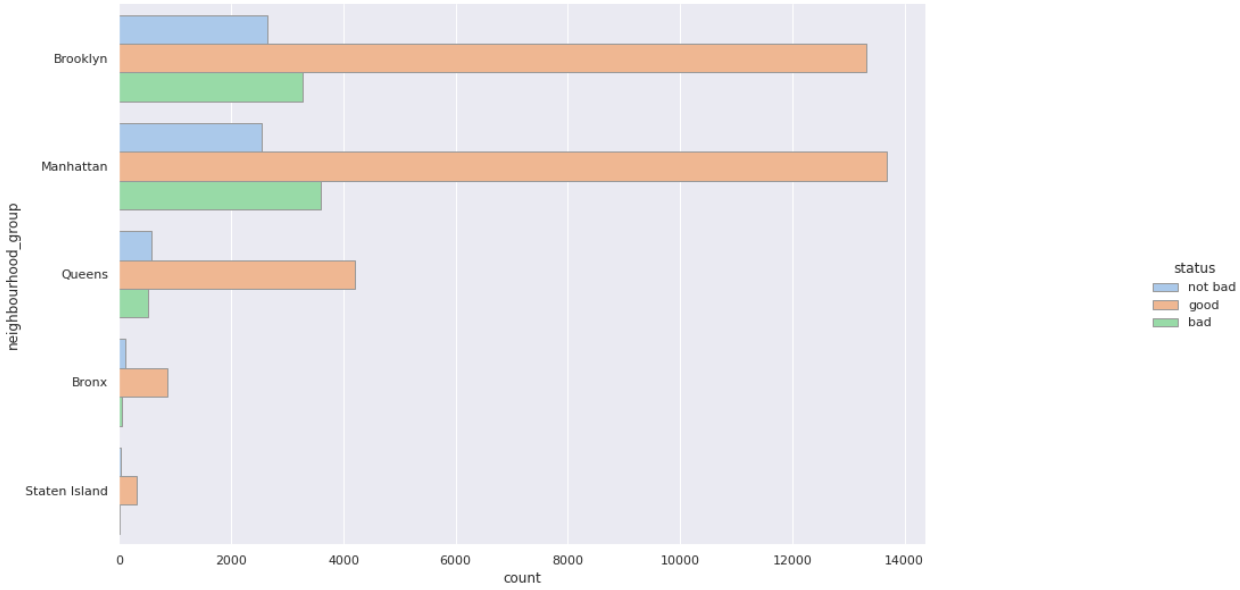
این هم خانه هایی که برچسب بد را گرفته اند.

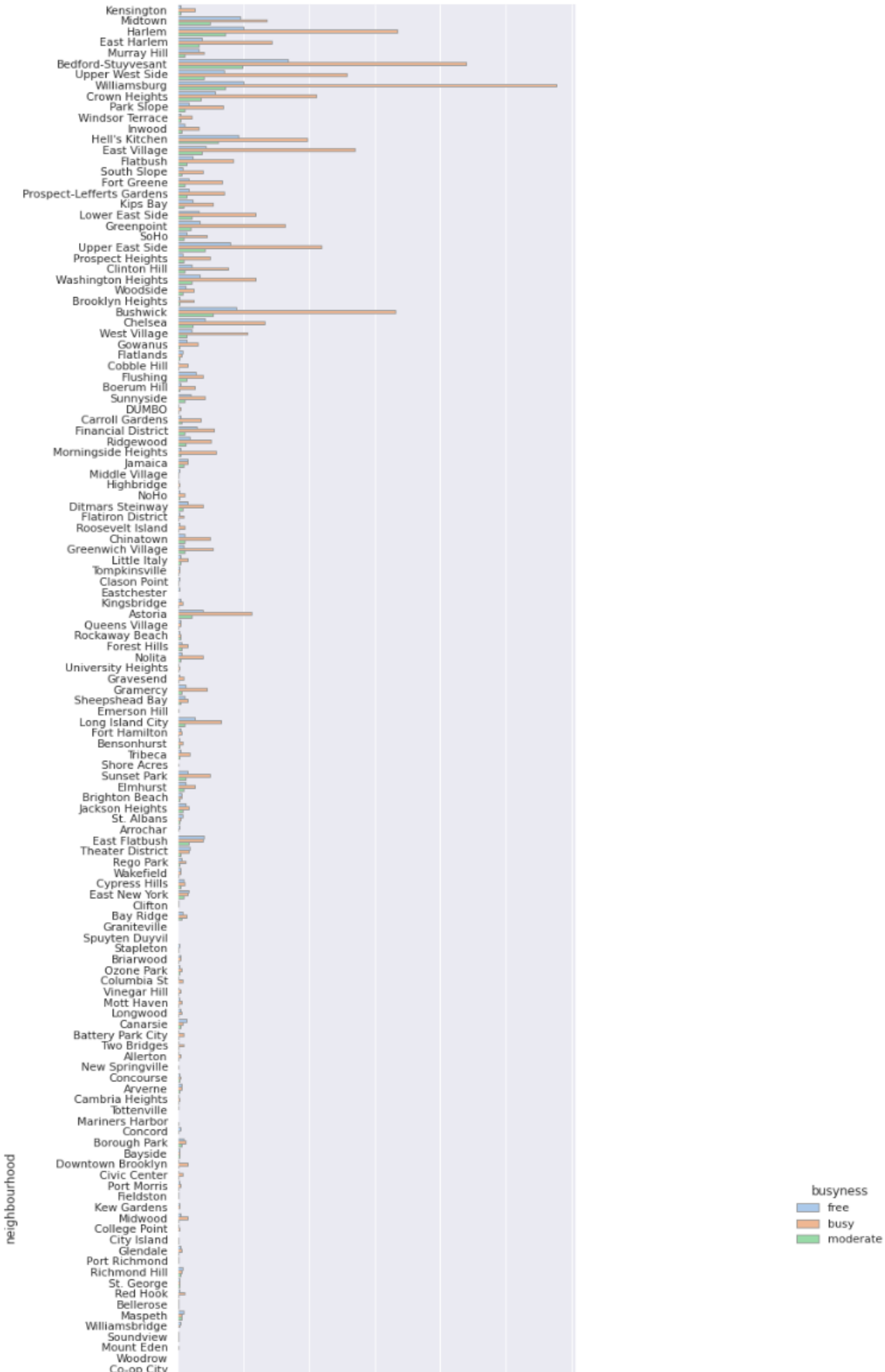
	status			
	count	unique	top	freq
neighbourhood_group				
Bronx	1034	3	good	862
Brooklyn	19221	3	good	13317
Manhattan	19830	3	good	13689
Queens	5304	3	good	4200
Staten Island	354	3	good	307

در هر منطقه به طور جدا خانه های خوب تعداد بیشتری دارند.

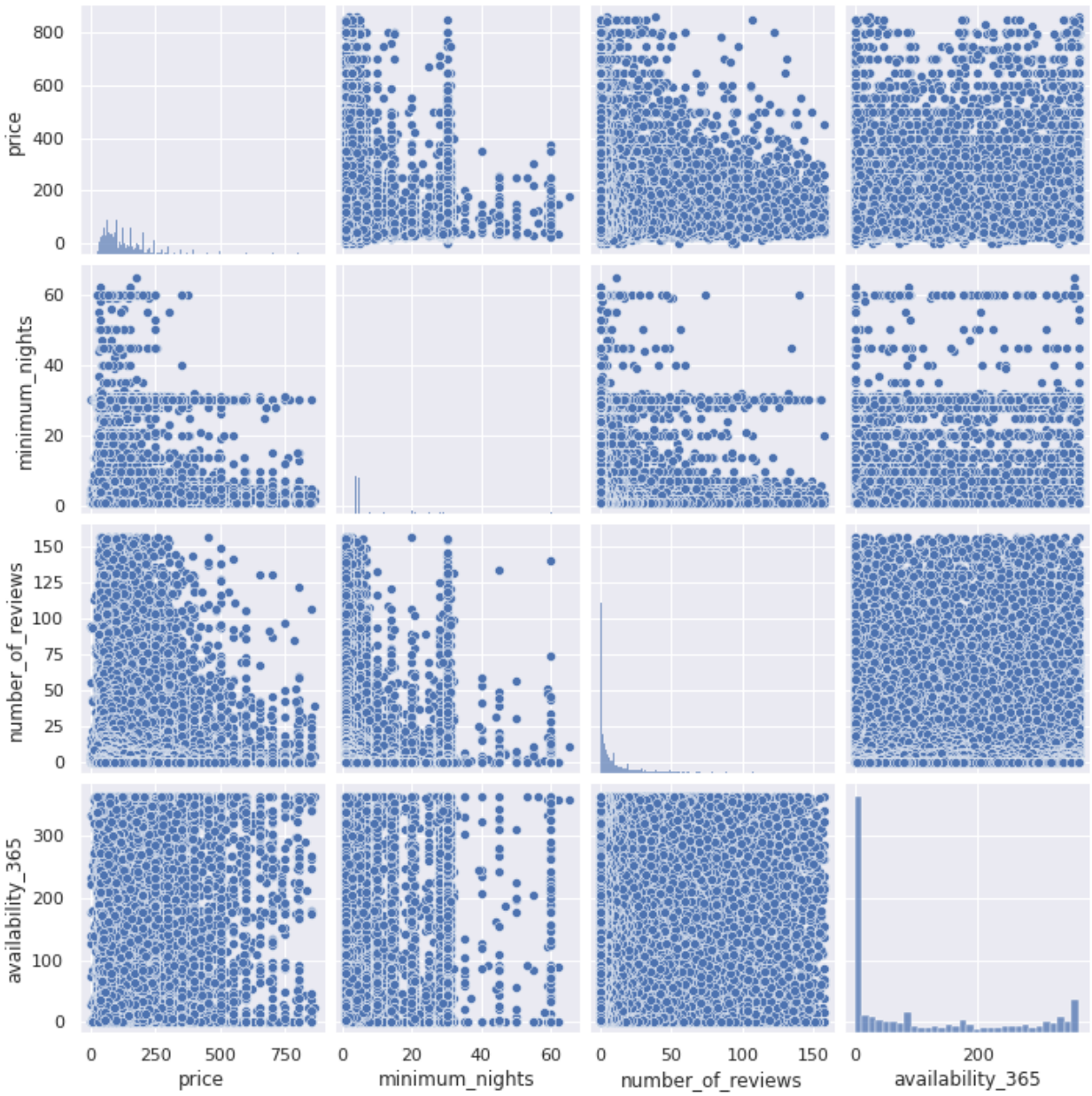
	neighbourhood_group			
	count	unique	top	freq
status				
bad	7448	5	Manhattan	3591
good	32375	5	Manhattan	13689
not bad	5920	5	Brooklyn	2638

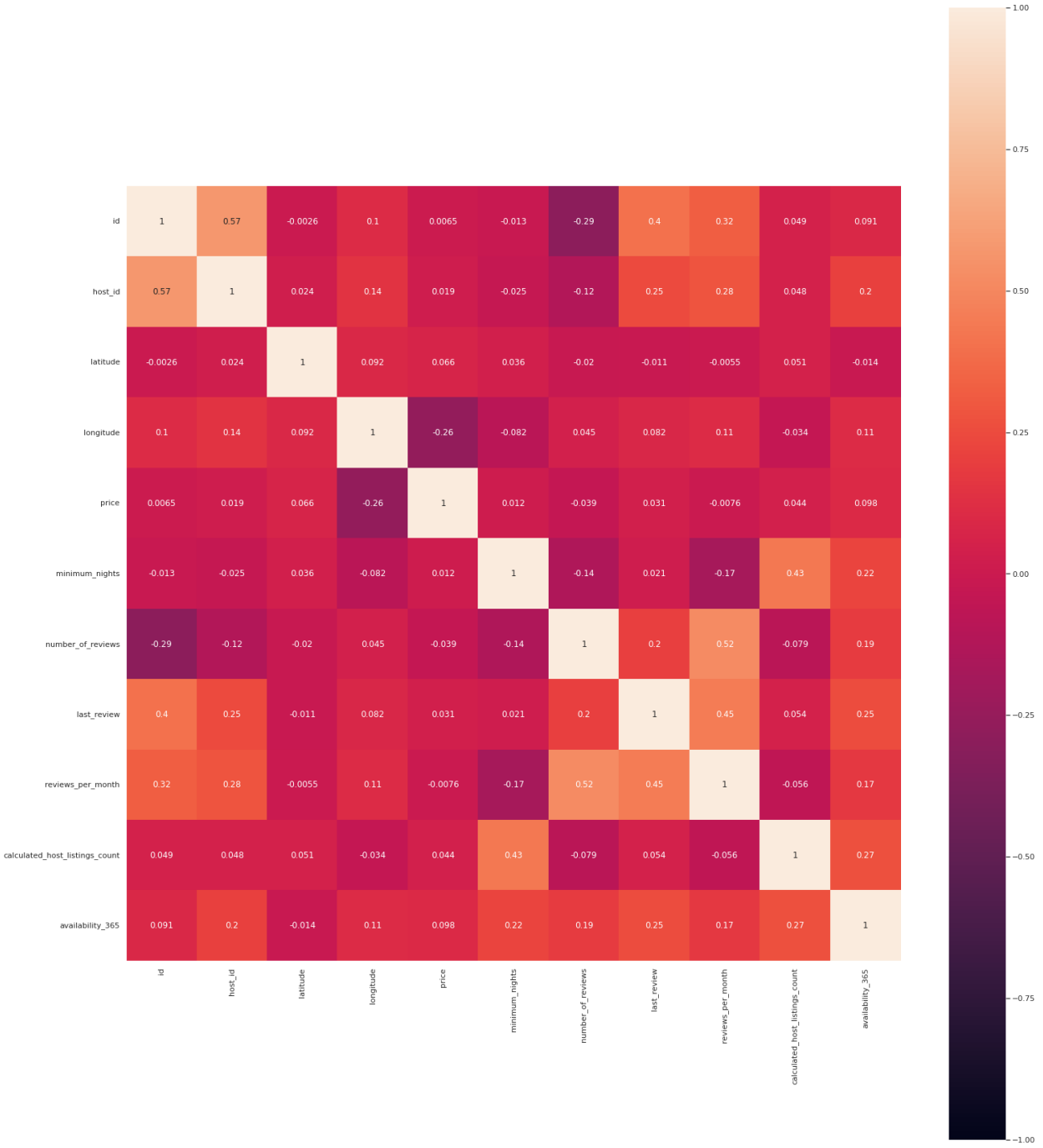
از 7448 خانه بد،3591 خانه در منهتن است بیشتر از هر منطقه دیگری.
از 32375 خانه خوب هم 13689 خانه در منهتن است بیشتر از هر منطقه دیگر.
البته ممکن است نتایج این بخش منطقی و درست نباشد چون که خیلی از داده ها در این ستون خالی بود.


Brooklyn
Manhattan
Queens
Bronx
Staten Island
neighbourhood_group
0
2000
4000
6000
8000
10000
12000
14000
count
status
not bad
good
bad


neighbourhood
Kensington
Midtown
Harlem
East Harlem
Murray Hill
Bedford-Stuyvesant
Upper West Side
Williamsburg
Crown Heights
Park Slope
Windsor Terrace
Inwood
Hell's Kitchen
East Village
Flatbush
South Slope
Fort Greene
Prospect-Lefferts Gardens
Kips Bay
Lower East Side
Greenpoint
SoHo
Upper East Side
Prospect Heights
Clinton Hill
Washington Heights
Woodside
Brooklyn Heights
Bushwick
Chelsea
West Village
Gowanus
Flatlands
Cobble Hill
Flushing
Boerum Hill
Sunnyside
DUMBO
Carroll Gardens
Financial District
Ridgewood
Morningside Heights
Jamaica
Middle Village
Highbridge
NoHo
Ditmars Steinway
Flatiron District
Roosevelt Island
Chinatown
Greenwich Village
Little Italy
Tompkinsville
Clason Point
Eastchester
Kingsbridge
Astoria
Queens Village
Rockaway Beach
Forest Hills
Nolita
University Heights
Gravesend
Gramercy
Sheepshead Bay
Emerson Hill
Long Island City
Fort Hamilton
Bensonhurst
Tribeca
Shore Acres
Sunset Park
Elmhurst
Brighton Beach
Jackson Heights
St. Albans
Arrochar
East Flatbush
Theater District
Rego Park
Wakefield
Cypress Hills
East New York
Clifton
Bay Ridge
Graniteville
Spuyten Duyvil
Stapleton
Briarwood
Ozone Park
Columbia St
Vinegar Hill
Mott Haven
Longwood
Canarsie
Battery Park City
Two Bridges
Allerton
New Springville
Concourse
Arverne
Cambria Heights
Tottenville
Mariners Harbor
Concord
Borough Park
Bayside
Downtown Brooklyn
Civic Center
Port Morris
Fieldston
Kew Gardens
Midwood
College Point
City Island
Glendale
Port Richmond
Richmond Hill
St. George
Red Hook
Bellerose
Maspeth
Williamsbridge
Soundview
Mount Eden
Woodrow
busyness
free
busy
moderate

در اینجا هم همانطور که در شکل بالا میبینیم خانه های هر محله کوچک از نظر شلوغ و خلوت بودن مورد بررسی قرار گرفت که میتوان فهمید شلوغ ترین محله williamsburg در بروکلین است یا Bedford-Stuyvesant که این هم در بروکلین است.
در آخر یک سری بررسی روی همبستگی انواع داده ها انجام داده شد:


price
minimum_nights
number_of_reviews
availability_365
0
200
400
600
800
0
20
40
60
0
25
50
75
100
125
150
0
100
200
300
0
250
500
750
0
20
40
60
0
50
100
150
0
200



مثلا می توان به این اشاره کرد که id با last review یک رابطه ای دارند که اگر id یک خانه بیشتر باشد(احتمالا این خانه اخیرا به این سامانه اضافه شده و جدید است) آخرین بازدید آن هم بیشتر است یعنی اخیرا مورد بازدید قرار گرفته است.