



## داده های مربوط به AirBnB

### پیش پردازش داده ها

با توجه به داده های موجود مشاهده میکنیم که ۴ ستون وجود دارد که دارای داده های null هستند. دو ستون تحت عنوان Name و host\_name هستند که اطلاعات زیادی به ما نمیدهند و نیازی به آنها نداریم، به همین دلیل این دو ستون را حذف میکنیم. دو ستون دیگر را نیز با استفاده از بیشترین value که در کل دیتاست دارند پر کردیم. بدین ترتیب داده های پوچ ما صفر میشوند.

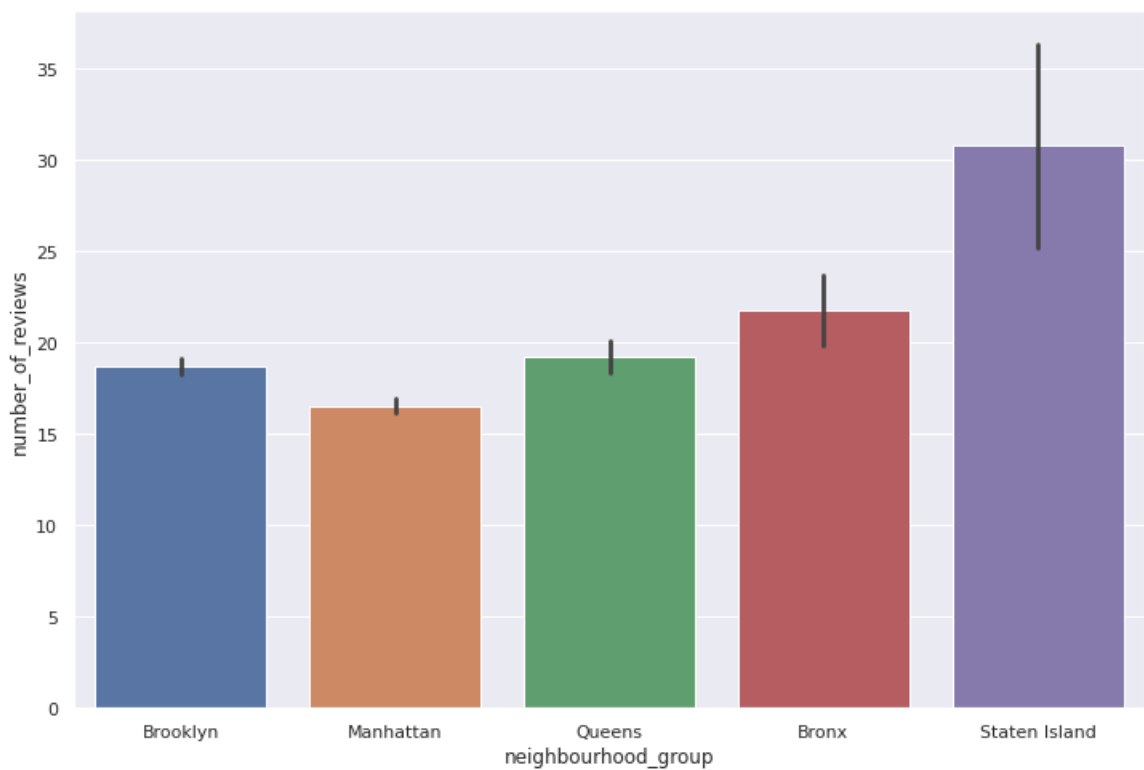
در مرحله بعد Outlier ها را نیز حذف میکنیم. آن داده هایی که در خارج از بازه  $mean() \pm 3 * std()$  هستند را از دیتاست حذف کردیم. حال میتوانیم به بررسی دقیق تر داده ها بپردازیم.

### بررسی تعداد بازدید ها

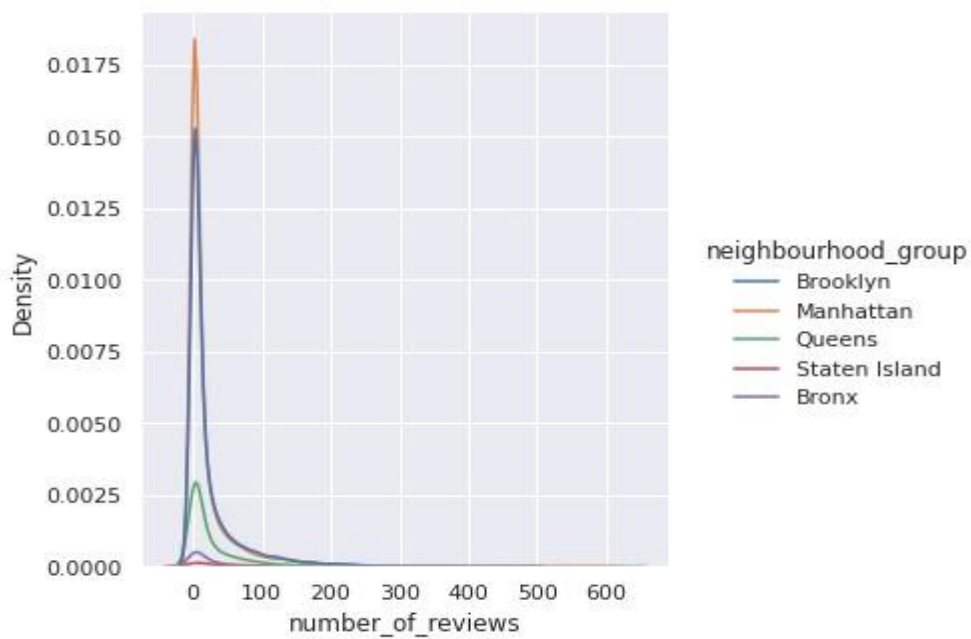
برای بررسی مجموعه داده های این دیتاست، ویژگی های ستون های مختلف و ارتباط آنها با یکدیگر را بررسی میکنیم. در ابتدا برای آنکه اطلاعاتی در مورد مناطق مختلف کسب کنیم، ستون neighbourhood\_group را که شامل ۵ منطقه منحصر بفرد است بر اساس تعداد بازدید ها و همچنین قیمت آنها بررسی میکنیم. اگر این مناطق را بر اساس review ها بررسی کنیم، مشاهده میکنیم که Staten Island به طور میانگین با اختلاف زیادی از بقیه مناطق تعداد بازدید های بیشتری داشته است. جدول زیر بیانگر همین مسئله میباشد:

	number_of_reviews							
	count	mean	std	min	25%	50%	75%	max
neighbourhood_group								
Bronx	963.0	21.743510	30.746189	0.0	1.0	8.0	30.0	152.0
Brooklyn	19165.0	18.720689	30.046985	0.0	1.0	5.0	22.0	155.0
Manhattan	19762.0	16.538913	28.432848	0.0	1.0	4.0	18.0	155.0
Queens	4311.0	19.193923	29.464950	0.0	1.0	5.0	25.0	155.0
Staten Island	170.0	30.782353	37.501317	0.0	2.0	17.5	45.0	151.0

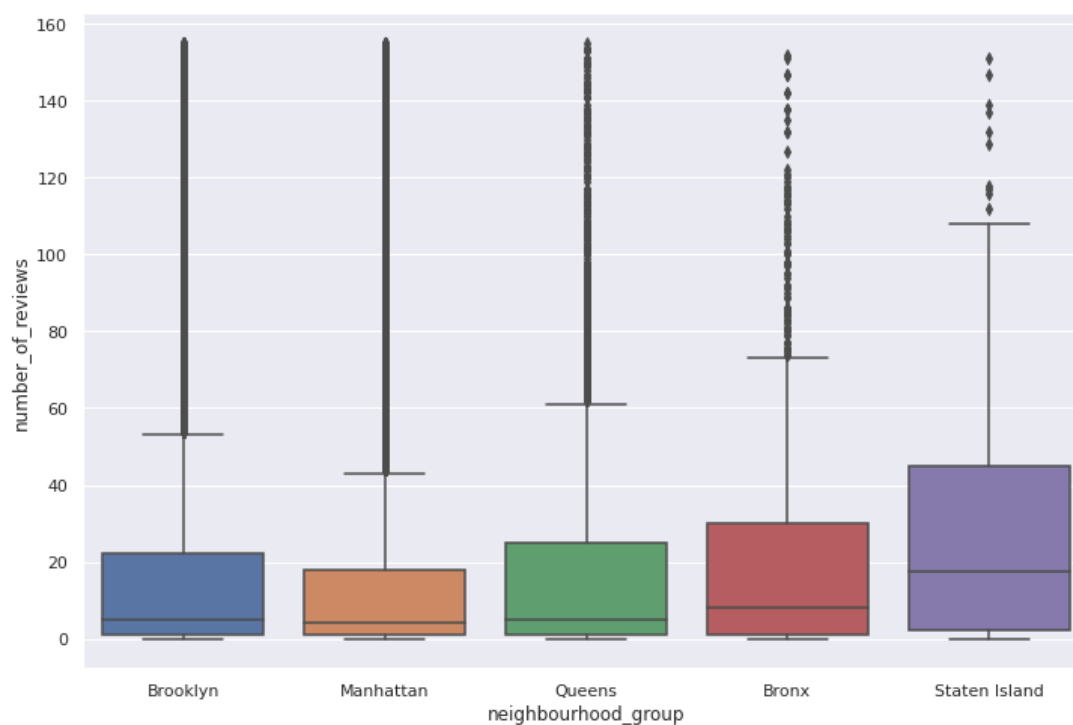
همچنین در نمودار زیر نیز این مسئله کاملاً مشخص است:



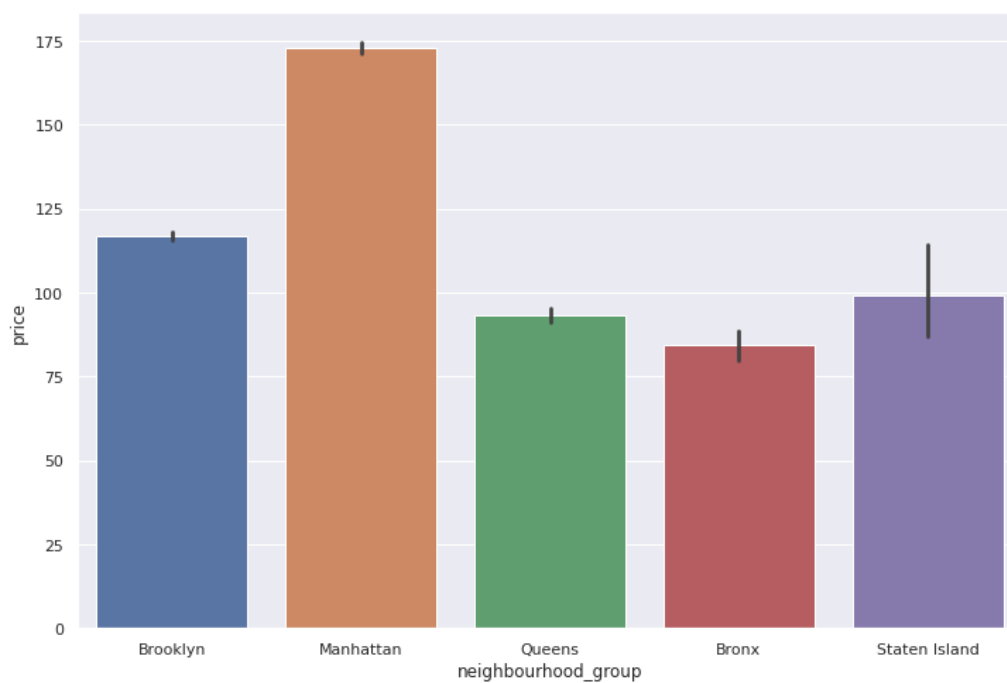
نمودار زیر نیز نشان دهنده توزیع تعداد بازدید های براساس neighbourhood های مختلف است، همانطور که میبینیم توزیع مورد نظر نرمال نیست، در ادامه تغییراتی در داده ها می‌دهیم تا توزیع را نرمال کنیم .



نمودار زیر نیز بیانگر تعداد بازدید های هر منطقه است که با استفاده از **boxplot** نمایش داده شده، در این نمودار نیز به طور واضح بازدید های Staten Island از باقی مناطق بیشتر است.

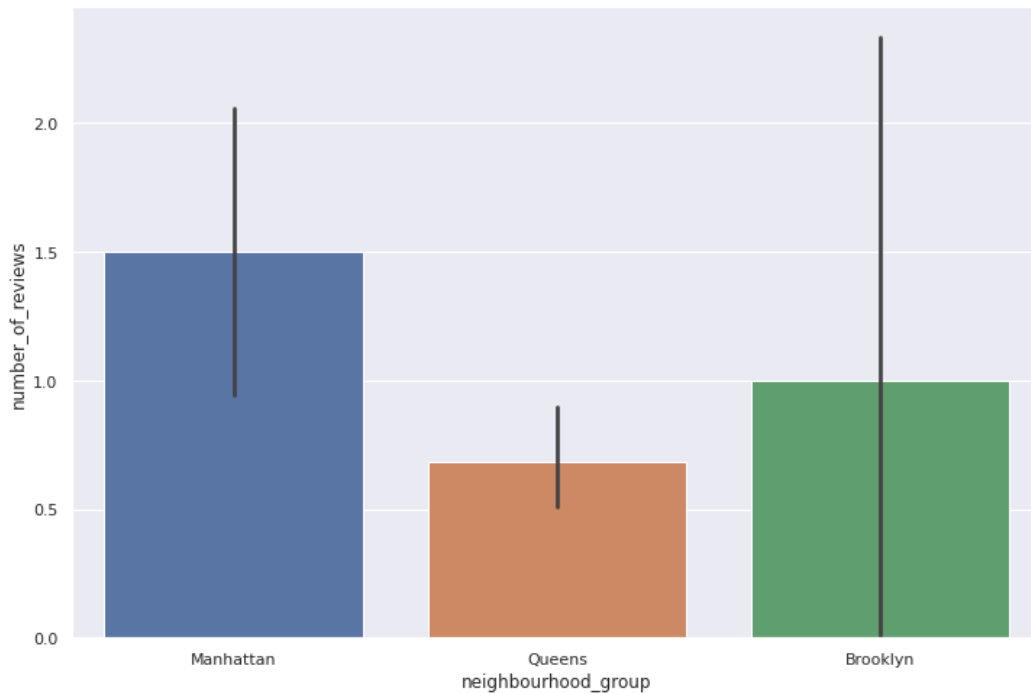


نمودار دیگری برای نمایش قیمت هر منطقه نیز داریم که در آن Manhattan از باقی بیشتر است.

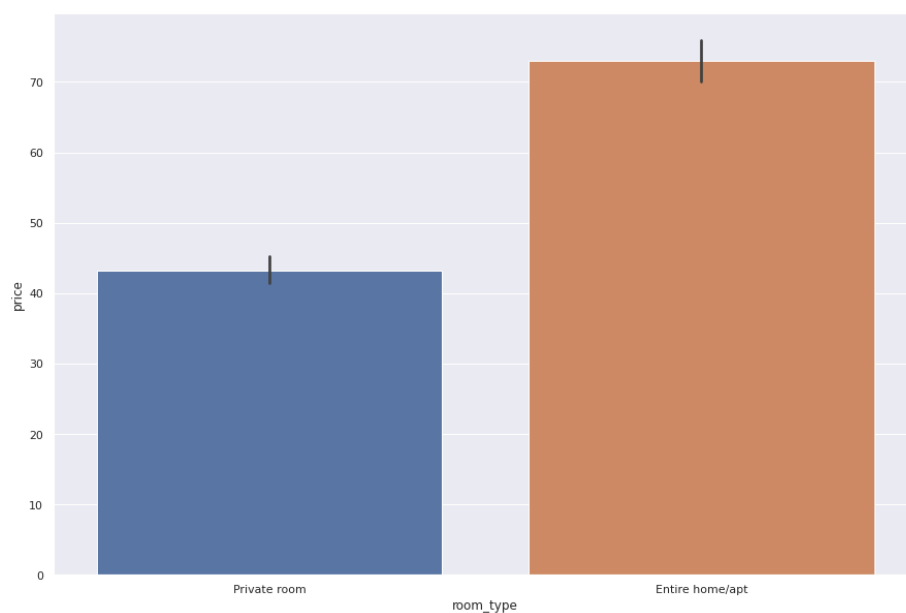


## بررسی صاحبان آگهی (Hosts)

در این بخش با استفاده از `host_id` ابتدا آن صاحبی که بیشترین تعداد آگهی را داشته است را پیدا کردیم. بیشترین تعداد ۱۰۳ مربوط به `host_id = ۱۳۷۳۵۸۸۶۶` است که شامل ۳ ناحیه `Manhattan`، `Brooklyn` و `Queens` میباشد. بیشترین تعداد بازدید مربوط به منهن بوده است.



در این بخش برخلاف کل آگهی ها، اتاق `shared_room` نداریم، همچنین قیمت خانه ها `Entire Home` بیشتر است.



پس از این نمودار ها، تست های مختلفی بر روی تعداد بازدید ها انجام دادیم:

اولین تست انجام شده را بر روی تعداد بازدید های هر منطقه زدیم. در اینجا چون 5 sample داریم از تست ANOVA استفاده کردیم. ابتدا بر روی تمامی داده ها انجام دادیم اما چون تعداد داده ها بسیار زیاد است، نتایج ما خیلی خوب نیست به همین علت پس از آن به تعداد تصادفی از هر منطقه ، ۱۰۰ داده را انتخاب کردیم .

هدف از انجام این تست بررسی فرضی بود که در ابتدا در مورد تعداد بازدید ها بررسی کرده بودیم. فرض صفر  $H_0$  ما این بود که این نواحی بایکدیگر از نظر تعداد بازدید تقریباً یکسان هستند، پس از بررسی های ابتدایی فرض یک  $H_1$  این بود که خیر، بازدید های ناحیه State Island از بقیه بیشتر است و در کل تعداد بازدید ها برابر نیست (مخالف  $H_0$ ) حال با استفاده از این تست و به دست آوردن  $p$ -value ، چون این مقدار در هر دو صورت (با در نظر گرفتن تمامی داده ها و انتخاب برخی از آنها) کمتر از 5٪ گزارش شد، پس میتوان نتیجه گرفت که فرض  $H_0$  رد میشود، این همان نتیجه ای است که ما از اول هم انتظار داشتیم و با استفاده از تست آماری نیز آن را متوجه شدیم.

همانطور که در جدول زیر هم قابل مشاهده است، میبینیم که میانگین ها خیلی از هم فاصله دارند و همین نشان دهنده نرمال نبودن توزیع است.

	number_of_reviews							
	count	mean	std	min	25%	50%	75%	max
neighbourhood_group								
Bronx	100.0	23.78	33.790974	0.0	2.0	9.0	28.50	147.0
Brooklyn	100.0	21.43	32.092749	0.0	1.0	4.0	28.75	149.0
Manhattan	100.0	15.92	26.046842	0.0	1.0	4.0	15.50	129.0
Queens	100.0	15.77	25.527507	0.0	1.0	5.0	20.25	141.0
Staten Island	100.0	31.67	38.511721	0.0	2.0	19.0	46.75	151.0

سپس برای منطقه Staten Island و مقایسه آن با همه نواحی از  $t$ -test و همچنین تست Wilcoxon استفاده کردیم. برای اینکار نیز مانند قبل، از ۱۰۰ نمونه از این داده های استفاده کردیم و با استفاده از این دو تست نیز با توجه به کم بودن  $p$ -value متوجه میشویم که فرض  $H_0$  ما که به معنی یکسان بودن بازدید ها بود رد میشود و مشخص خواهد شد که بازدید staten island از تمامی مناطق دیگر بیشتر است.

سپس دو منطقه Brooklyn و Staten Island را نیز مقایسه کردیم و دیدیم که باز هم نتایج نشان دهنده آن است که این دو از نظر تعداد بازدید باهم متفاوت هستند.

## بررسی نوع اتاق ها و قیمت

پس از آنکه در مورد تعداد بازدید ها اطلاعات مناسبی را بدست آوردیم؛ به بررسی نوع اتاق ها پرداختیم. قطعا نوع اتاق انتخابی، قیمت های متفاوتی را دارد . بررسی کردیم آیا نوع اتاق بر روی قیمت تاثیر دارد یا نه ؟ سه نوع room type مختلف مشاهده میکنیم، که تعداد هر کدام و میانگین قیمت آنها مشخص است:

room_type	price							
	count	mean	std	min	25%	50%	75%	max
Entire home/apt	22853.0	190.844047	115.430517	0.0	120.0	160.0	225.0	860.0
Private room	20430.0	84.121439	61.364480	0.0	50.0	70.0	96.0	848.0
Shared room	1088.0	66.965074	72.995032	0.0	33.0	45.0	75.0	800.0

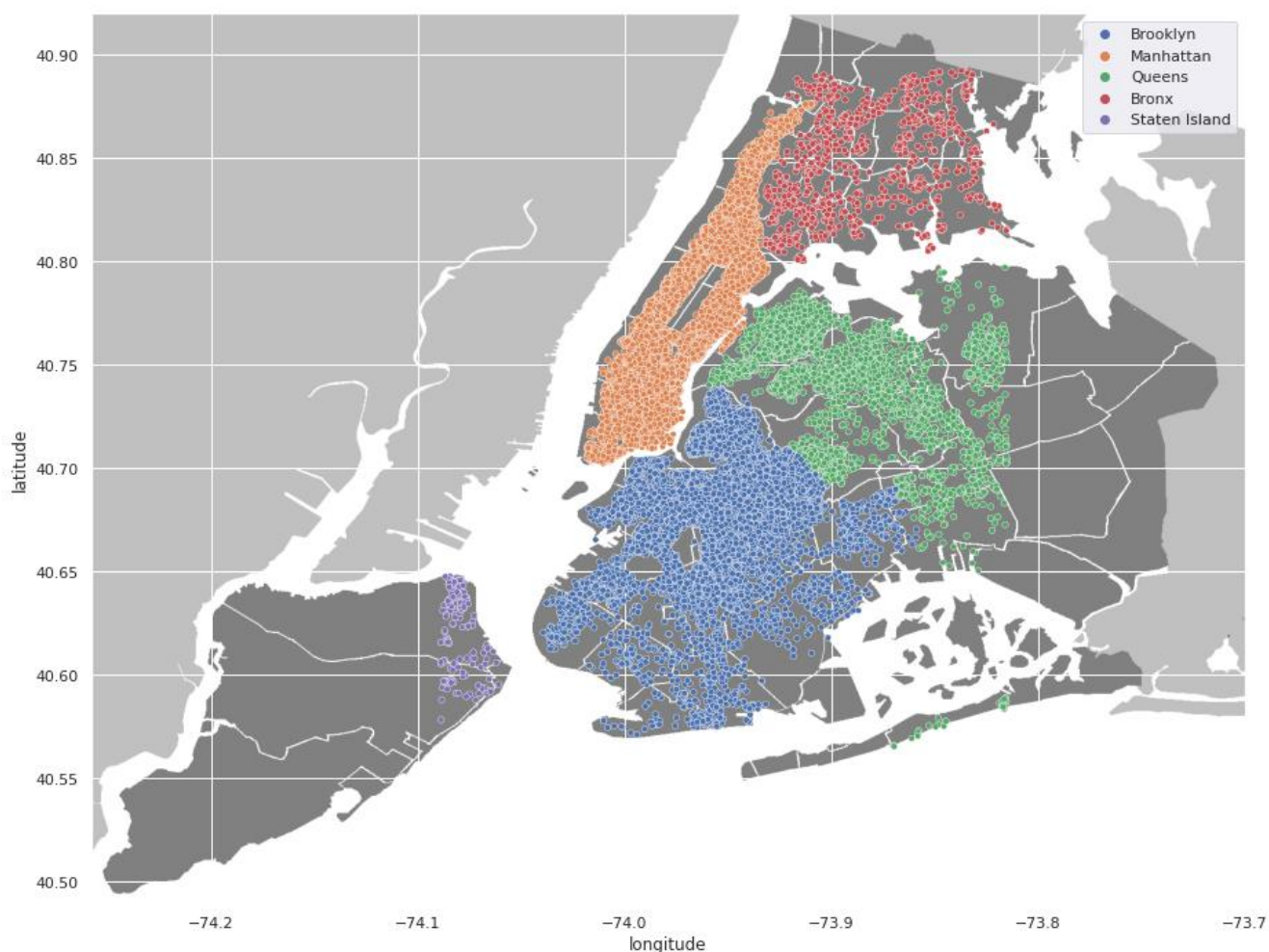
توجه به اختلاف میانگین بسیار بالا نتایج کاملا واضح است و خانه هایی که به صورت entire home گرفته شوند قطعا از قیمت بالاتری برخوردار هستند.

## حداقل تعداد شب ماندن و قیمت

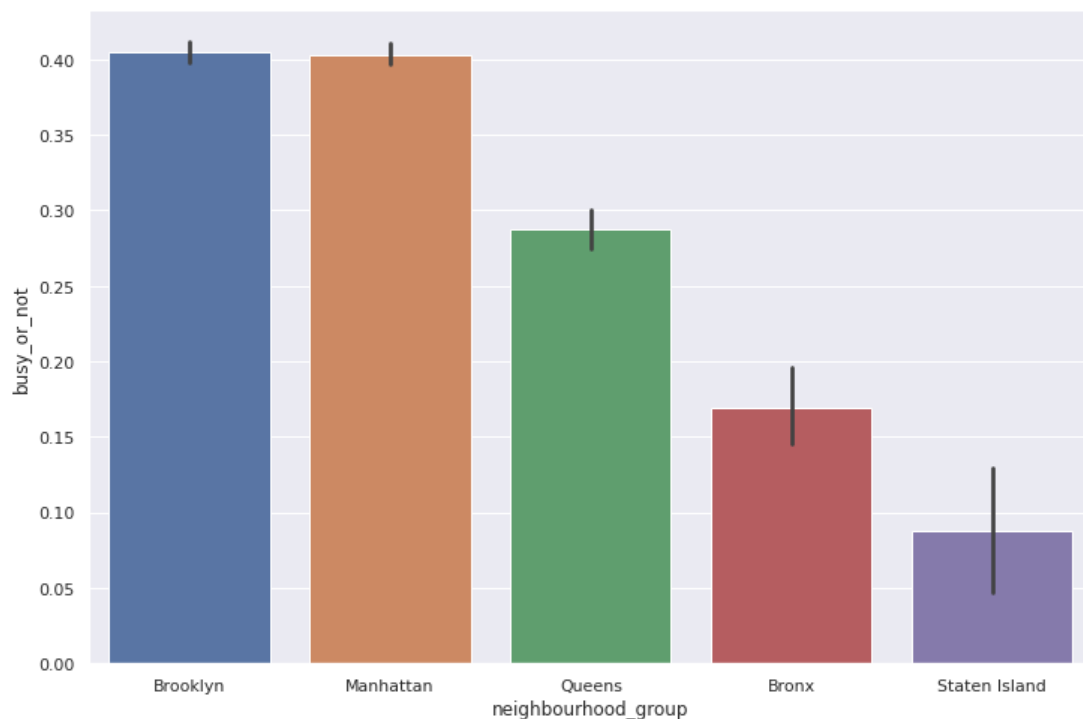
سپس به بررسی دوتا داده عددی پرداختیم. آیا تعداد حداقل شب ها بر روی قیمت خانه ها تاثیر گذار است یا خیر؟ فرض  $H_0$  این است که این دو مقدار به هم مربوط نیستند و قیمت برای تمامی یکسان است. این دو متغیر هر دو عددی هستند و برخلاف تست های قبلی که معمولا یکی عددی و دیگری categorical بود نیست. به همین علت تست های آماری مناسب برای این داده ها استفاده از correlation ها هست. هر دو تست pearson و spearman را تست کردیم و باز هم در ابتدا بر روی کل داده ها این کار را انجام دادیم و مشاهده میکنیم که نتایج قابل توجه است و  $p$ -value مقدار بسیار کمی دارد در نتیجه میتوان گفت که فرض  $H_0$  ما رد میشود و مورد قبول نیست و در نتیجه میتوانیم بگوییم که تعداد minimum night ها بر روی قیمت خانه تاثیر داشته است.

## بررسی ترافیک شهر ها

برای بررسی آنکه ترافیک کدام شهر ها بیشتر است از دو روش استفاده کردیم. اول نشان دادن بر روی نقشه نیویورک بود که در شکل زیر هم مشاهده میکنید:

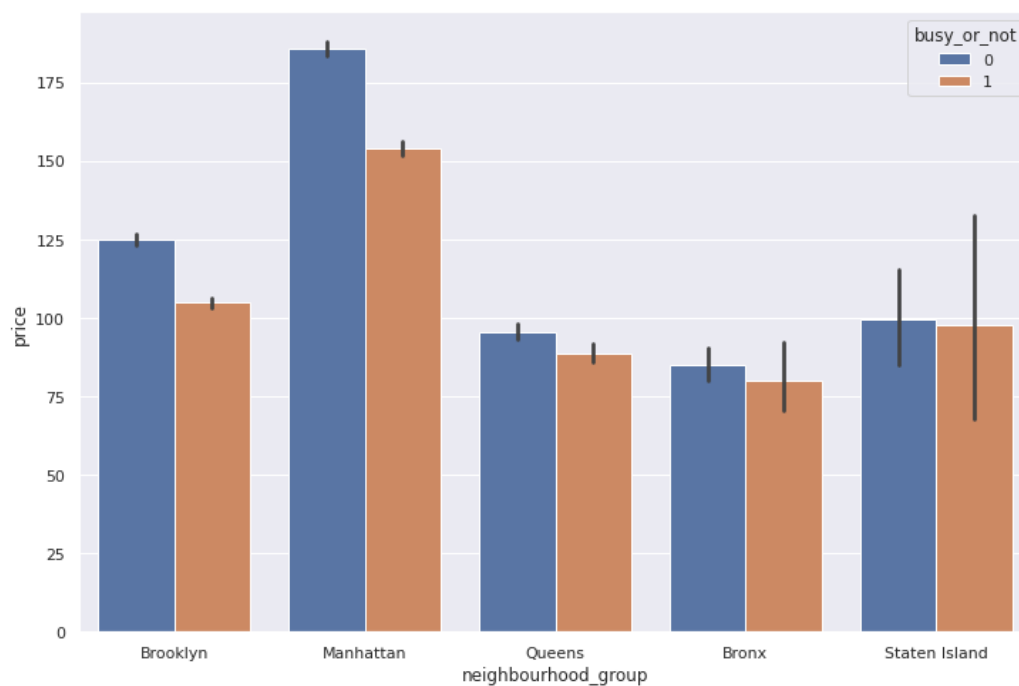


همانطور که در نقشه نیز مشخص است، تراکم برای نقاط نارنجی و آبی بیشتر از بقیه هست یعنی دو شهر Brooklyn و Manhattan بیشترین خانه ها را دارند، همچنین این وضعیت را نیز میتوان از روی **availability** نیز مشخص نمود. برای آنکه **availability** را بهتر بتوانیم مشاهده کنیم، اینگونه بررسی کردیم که هرچی از ۳۶۵ روز سال تعداد کمتری **availability** داشته باشیم به این معناست که قطعا آنجا شلوغ تر بوده و بیشتر رزرو شده است. پس میتوانیم آن **host** هایی که تعداد **availability** آنها برابر با صفر است را به عنوان **host** های **busy** در نظر بگیریم و بقیه را خیر. بدین ترتیب به نمودار زیر رسیدیم:



همانطور که در نمودار هم مشخص است، Brooklyn و Manhattan جز شلوغ ترین شهر ها محسوب میشوند بدین معنی که مردم بیشتر در این شهر ها رفت و آمد دارند و بیشتر خانه رزرو کردهاند.

نمودار زیر نشان میدهد، با اینکه قیمت خانه ها در Manhattan بیشتر از بقیه محله ها هست، اما تعداد مراجعات به این شهر بیشتر است.





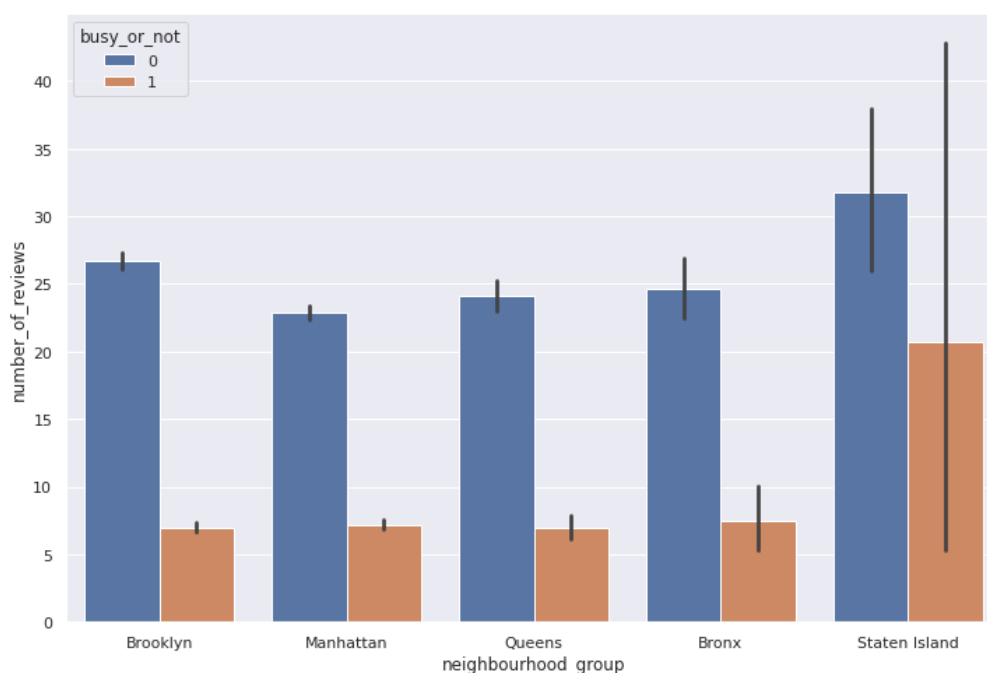
## رابطه شلوغ بودن با قیمت و تعداد بازدیدها

سپس میخواهیم بررسی کنیم که آیا محله هایی که busy هستند، قیمت بالاتری نیز دارند یا خیر؟

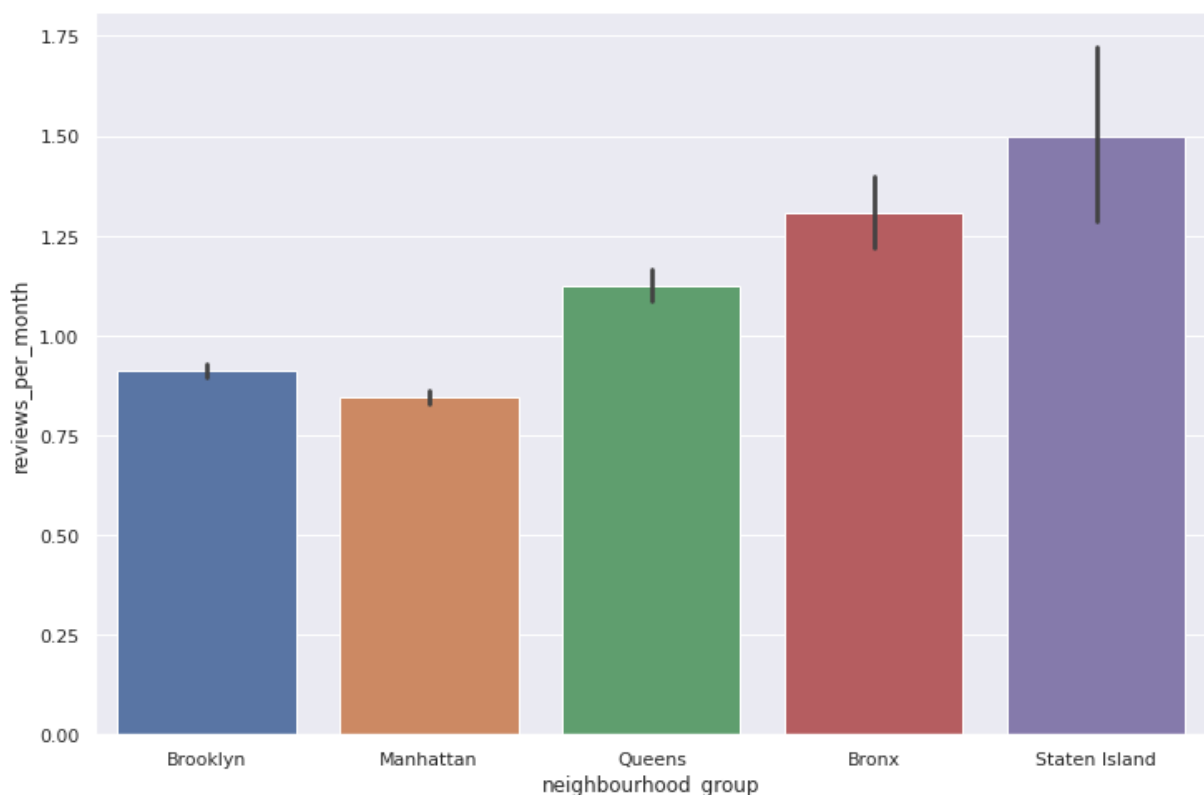
با توجه به t test انجام شده بین محله هایی که availability آنها برابر با صفر است و قیمت خانه ها، مشاهده میکنیم که p value مقداری کمتر از 05.0 دارد و این به این معنیست که فرض H1 ما مورد قبول است و یعنی محلهای که ترافیک بالاتری دارد، قیمت بیشتری نیز دارد. این مسئله در نمودار بالا نیز قابل مشاهده بود.

سپس با تست بعدی نیز بررسی کردیم که تعداد بازدیدها چگونه است؟ آیا آن هم بیشتر است؟

در این تست نیز به همین نتیجه میرسیم. یعنی فرض H0 ما رد میشود. از روی مقدار P-VALUE در sample های مختلف این مسئله نمایان است. همچنین در نموداری که در زیر داریم نیز این مسئله مشخص میشود. اگر دقت کنیم Staten Island تعداد بازدیدهای آن زمانی که  $busy = 1$  است بیشتر از بقیه هست. پس این فرض نیز برقرار است.



در نمودار زیر نیز تعداد بازدید های هر ماه را برای مناطق مختلف مشاهده میکنیم. همانطور که میبینیم، میزان بازدید ها در هر ماه برای منطقه Staten Island از بقیه بیشتر بوده است.



### بررسی neighbourhood های خاص

سپس برای هر neighbourhood خاص نیز میزان available بودن را بررسی کردیم. تعداد neighbourhood ها برای هر منطقه خاص زیاد است اما با توجه به تفسیری که قبلا داشتیم، هر چه تعداد روز های available کمتر باید به این معنیست که رزرو بیشتری انجام شده ، پس برای هر neighbourhood میانگین کمتر را در نظر میگیریم و بدین ترتیب در هر منطقه میتوانیم به نتایج زیر برسیم:

Manhattan -> Morningside Heights

Brooklyn -> Downtown Brooklyn

Staten Island -> Rossville

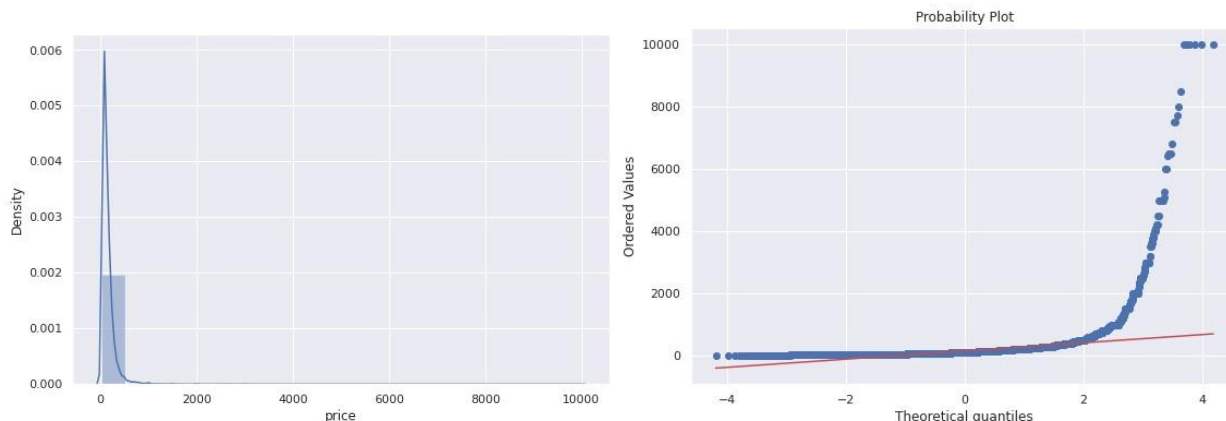
Queens -> Little Neck

Bronx -> Melrose

## نرمال کردن توزیع price

در مرحله بعدی، بررسی هایی بر روی قیمت انجام دادیم اما ایندفعه توزیع price رو به توزیع نرمال تبدیل کردیم:

در ابتدا خانه هایی را که price = 0 داشتند ، از dataset حذف کردیم. سپس دو نمودار زیر را که نشان دهنده توزیع price پیش از نرمال شدن است را داریم:



نمودار سمت چپ توسط KDE plot رسم شده و به صورت curve توزیع داده را نشان میدهد.

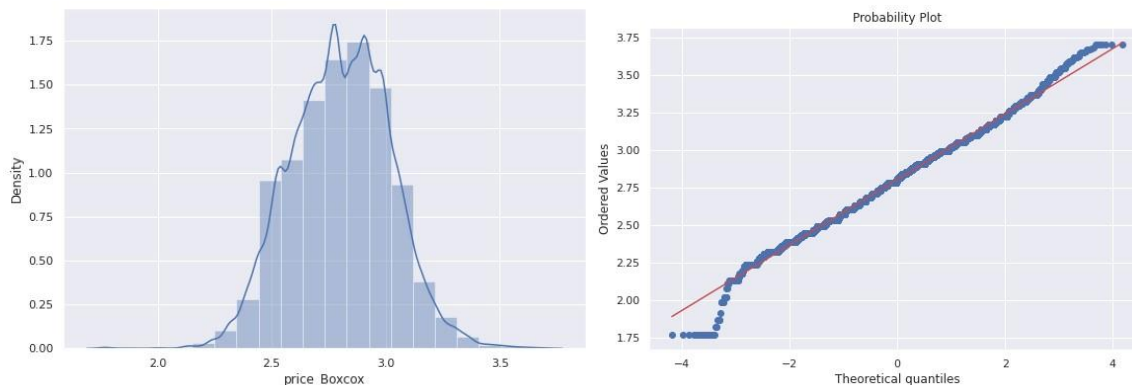
نمودار سمت راست نمودار Q-Q است ، که محور x شامل مقادیر quantile و محور y شامل مقادیر price است. اگر مقادیر داده price ما نزدیک به خط  $x=y$  باشد در آنصورت توزیع ما نرمال است، اما در اینجا همانطور که مشاهده میکنید توزیع نرمال نیست چون مقادیر ما از خط مورد نظر فاصله دارند .

حال برای تبدیل کردن این توزیع به توزیع نرمال از تبدیل Boxcox استفاده کردیم. این تبدیل به صورت زیر عمل میکند:

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y_i) & \text{if } \lambda = 0, \end{cases}$$

که در این تبدیل،  $y$  متغیر پاسخ و  $\lambda$  پارامتر تبدیل است.  $\lambda$  میتواند از -5 تا 5 متغیر باشد. در طول تبدیل ، همه مقادیر  $\lambda$  در نظر گرفته می شود و مقدار بهینه/بهترین برای متغیر انتخاب می شود. بهینه ترین مقدار آن مقداری است که نتیجه آن بهترین تقریب از منحنی توزیع نرمال را به ما بدهد.

هرگاه که  $\lambda = 0$  باشد، نیز مقدار لگاریتم طبیعی  $y$  محاسبه میگردد. این توزیع نسبت به بقیه توزیع ها بهتر برای price عمل کرد. در شکل زیر نمودار های بالا را پس از تبدیل میبینیم:



همانطور که میبینید این تبدیل بسیار خوب عمل کرده و تقریباً توزیع ما کاملاً نرمال شده، حتی اگر **mode**، **mean** و **median** را نیز محاسبه کنیم، مشاهده میکنیم که کاملاً نزدیک به یکدیگر هستند.

Mean = 2.802

Median = 2.800

Mode = 2.781

پس از آنکه توزیع ما بسیار به توزیع نرمال نزدیک شد، نتیجه تست هایی که در ادامه داریم بسیار دقیق تر خواهد بود.


### مدلسازی برای پیشبینی قیمت:

برای پیشبینی قیمت از مدل **regression** استفاده کردیم. با توجه به آنکه تعداد فیچر های ما زیاد است پس مدل ما غیرخطی خواهد بود. برای پیشبینی قیمت نیاز است تا این ویژگی را به عنوان **target** و بقیه ویژگی ها را که بر اساس آنها میتوان پیشبینی خوبی داشت و مرتبط به قیمت هستند را به عنوان **features** در نظر میگیریم.

این مدل سعی میکند با در اختیار داشتن آن فیچر ها، نزدیک ترین منحنی را به هدف اصلی پیدا کند تا بتواند بهترین پیشبینی را داشته باشد.

در اینجا پیش از آنکه اطلاعات را به مدل بدهیم نیاز بود تا دو ویژگی **categorical** را تبدیل به داده **numeric** کنیم تا بتوانیم به مدل بدهیم. این عمل با استفاده از دستور **get\_dummies** که کار **One hot encoding** را انجام میدهد، انجام دادیم. همچنین بخشی از داده در اختیار را به عنوان داده تست کنار گذاشتیم تا بتوانیم با توجه به آن دقت مدل را بسنجیم.

مدل **regression** را با استفاده از پکیج **sklearn** ساختیم و سپس برروی فیچر هایی که در اختیار داریم ران کردیم. نتایج **predict** شده در زیر آمده است:



	test	pred
0	300	181.764622
1	255	235.586644
2	179	175.780003
3	70	50.327367
4	115	130.203099

همانطور که مشخص است، مدل خیلی دقیق نیست و خطی به نسبت زیادی دارد. در برخی موارد به خوبی پیشبینی شده اما در برخی دیگر فاصله زیادی با داده تست دارد. همچنین مقدار  $MSE$  برابر با ۷۹۲۲ است که به نسبت عدد بزرگی است اما باز هم خوب عمل کرده است.