محمد زیاری - 97222047

بخش دوم - دیتاست Airbnb

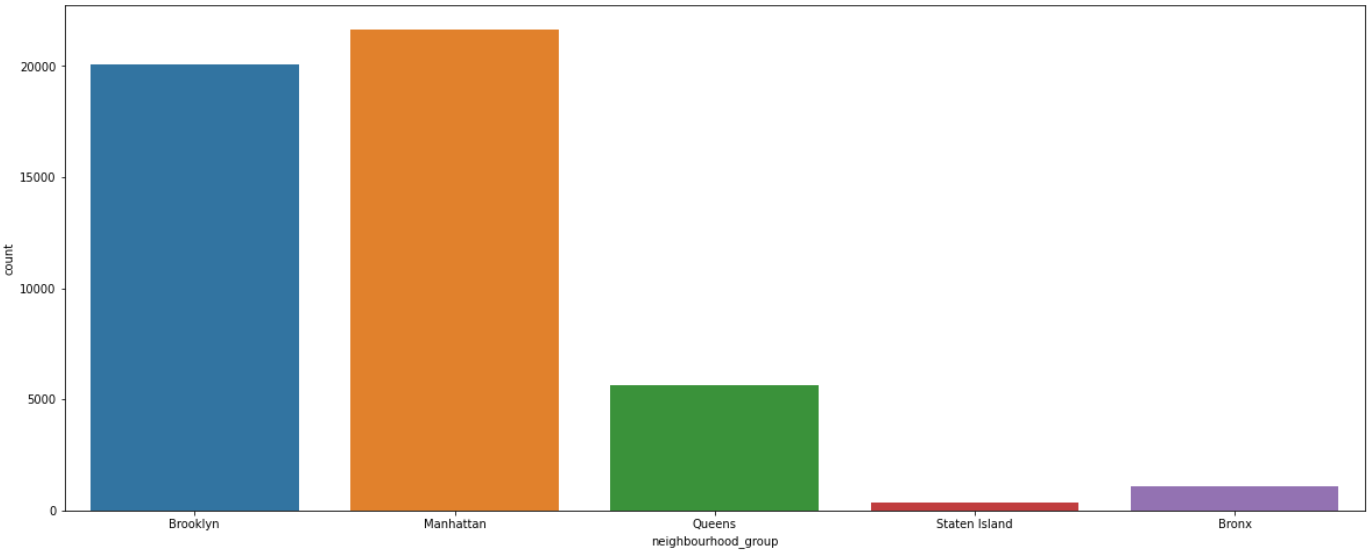
داده های ما در این سوال خانه ها و اقامتگاه هایی است که با بررسی داده ها با تست های آماری و هم چنین جدول ها و نمودارها ارتباط ستونهای مختلف را بررسی میکنیم.

قبل از شروع کار اصلی ، ابتدا داده ها را از kaggle میخوانیم و در بخش پاکسازی داده های با price=0 را از دیتاست پاک کردم زیرا معنی نداشت که خانه ای ، قیمتی برابر یا کوچکتر از صفر داشته باشد. البته انجام این کار در کارهای آتی هم به درد میخورد که به آن اشاره خواهم کرد.

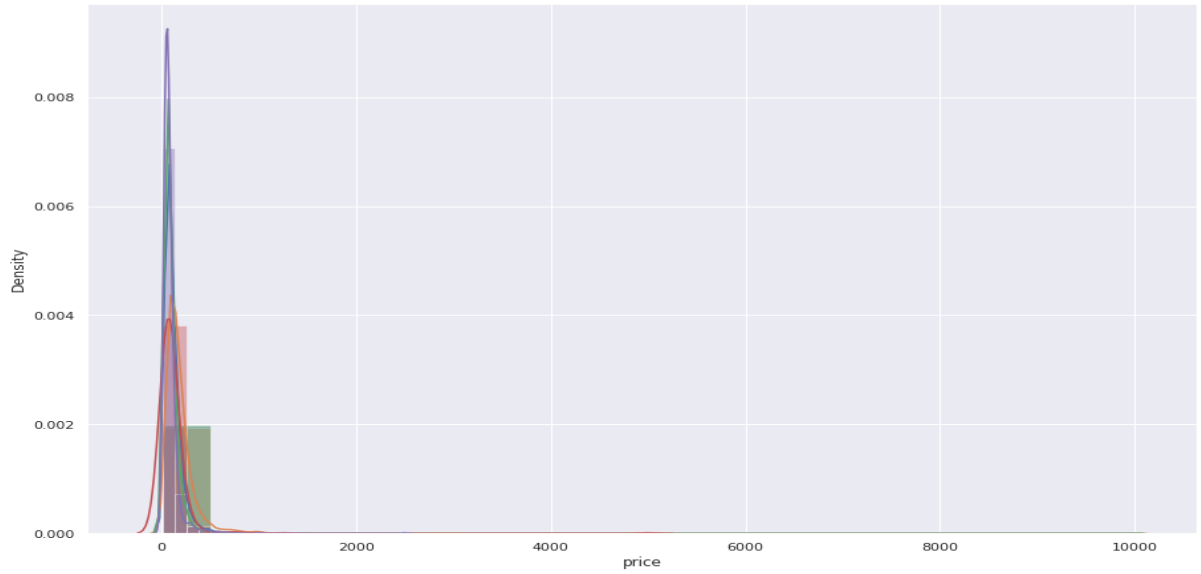
سپس داده های پرت که 3 برابر بیشتر از واریانس از میانگین فاصله داشتند (چه 3 برابر بیشتر و چه 3 برابر کمتر از میانگین) را به عنوان داده ی پرت از دیتاست حذف کردیم.

پس از حذف ستون های name و host-name به دلیل اینکه در این پروژه از nlp استفاده نمی کنیم که پردازش متن داشته باشیم و همچنین id که قاعدتا تاثیری در روند یادگیری ندارد تنها ستونهایی که مقادیر null دارند last-review و review-per-month بودند که انها را با استفاده از میانگین پر کردیم.

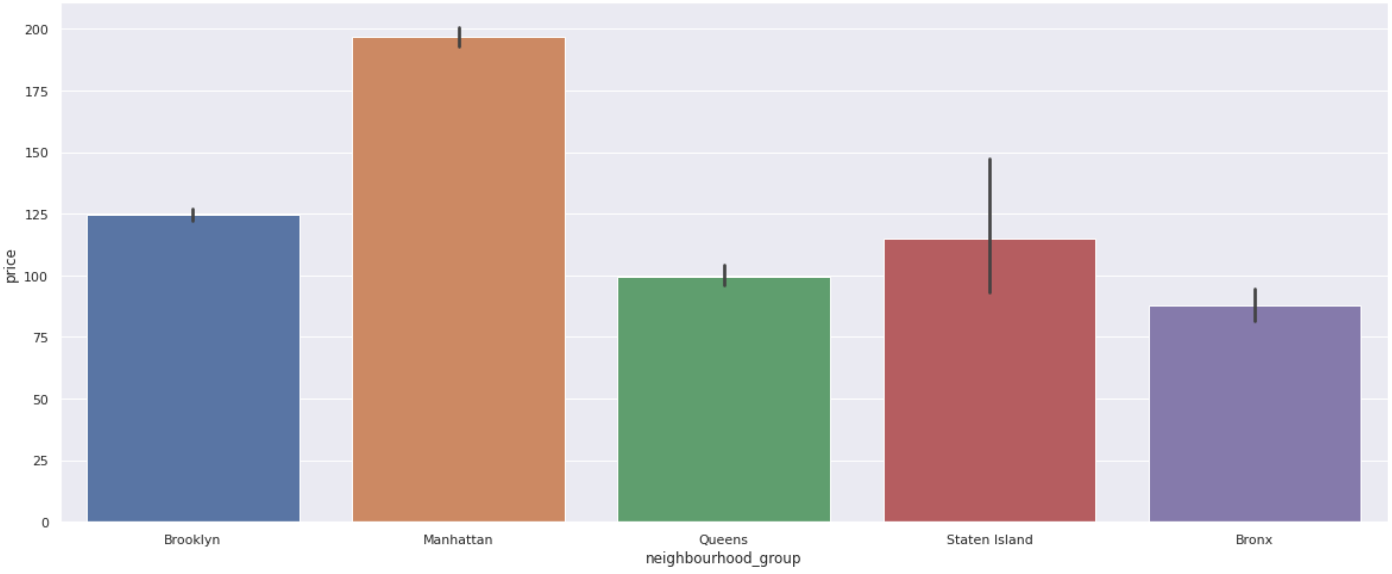
در بلاک بعدی ارتباط محله یا neighbourhood group های مختلف را با ستون یا feature های دیگر بررسی میکنیم. نمودار پایین ، تعداد خانه های هر منطقه را نشان می دهد.



مشخصا تعداد خانه های manhattan بیش از همه است. حالا وارد بررسی قیمت خانه ها می شویم تا ببینیم ، محله بر قیمت تاثیر دارد یا خیر. توزیع قیمت خانه های هر منطقه را در نمودار زیر می بینیم:



این رابطه از توزیع نرمال پیروی می کند ، اما ما تست Anova را بر روی neighbourhood Group های مختلف بررسی میکنیم و تست بر این اساس است که آیا قیمت خانه های مختلف در یک بازه است یا خیر که جوابی که بدست آمد p-value بسیار پایینی در حدود 10 به توان منفی 300 دارد که مشخصا فرض صفر برقرار است و قیمت خانه ها در نواحی مختلف متفاوتند. یکی از دلایلی که این عدد بسیار بزرگ را رقم زده میتواند تعداد بالای نمونه ها باشد اما زمانی که روی داده ها از describe استفاده کردیم که جدول در کد موجود است ، میبینیم میانگین قیمت ها به طور قابل توجهی با هم اختلاف دارند. نمودار قیمت خانه ها در زیر مشاهده می کنیم.

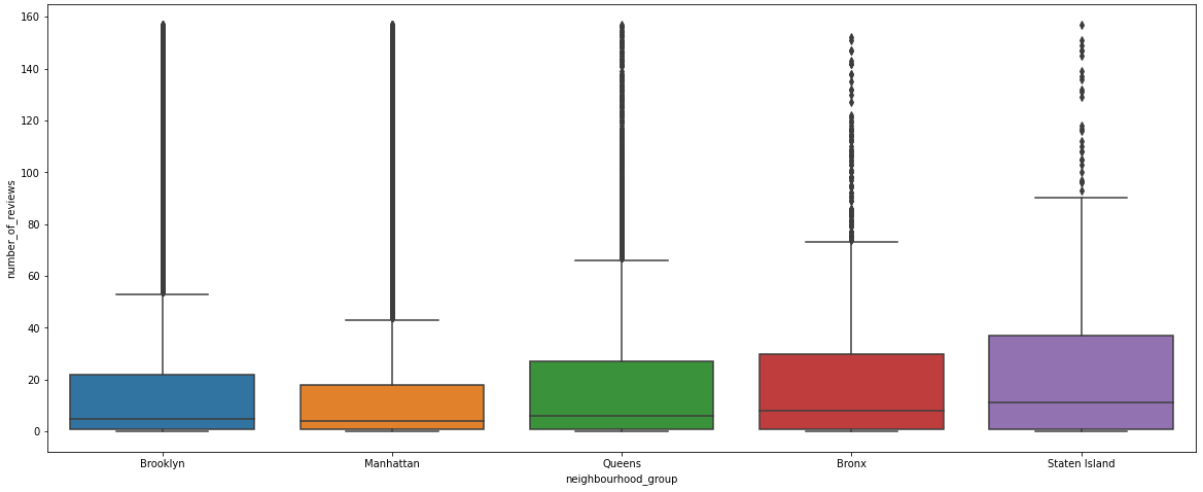


با توجه به میانگین قیمت ها ، فرض می گیریم که تفاوت بسزایی در میزان قیمت خانه ها در Staten Island و Brooklyn وجود دارد و T-test را روی قیمت این دو محله انجام می دهیم. نتیجه نمونه های متفاوتی که از این دو گرفته شد . مثلا آخرین نمونه p-value=0.4 داشت و کمترین مقدار p-value نیز 0.04 بود ، که به

نسبت اینکه هر نمونه تعداد 373 موجودیت دارد ، مقدار خیلی کمی به نظر نمیاد و فرض 0 ما رد میشود و قیمت این دو محله تفاوت بسزایی ندارند. بررسی های دیگر بر قیمت را پس از نرمال کردن تابع آن انجام می دهیم.

به سراغ تعداد بررسی یا Number of reviews می رویم. توزیع آن در کد موجود است که برای شلوغ نشدن زیاد گزارش آن را نمایش نمی دهم. وقتی روی این ستون نسبت به محله های مختلف anova میزنیم تا در واقع میزان تفاوت آن ها را بیش از حد ، نشان دهیم عدد 10 به توان منفی 28 را نشان میدهد که در نگاه اول بسیار بالاست اما وقتی به میانگین ها و همچنین نسبت به p-value قیمت ها مقایسه انجام میدهیم ، شاید آنقدر هم بالا نباشد. برای همین sample یا نمونه ای از neighbourhood group های مختلف میگیریم تا ببینیم باز هم همین نتیجه را می دهد یا خیر.

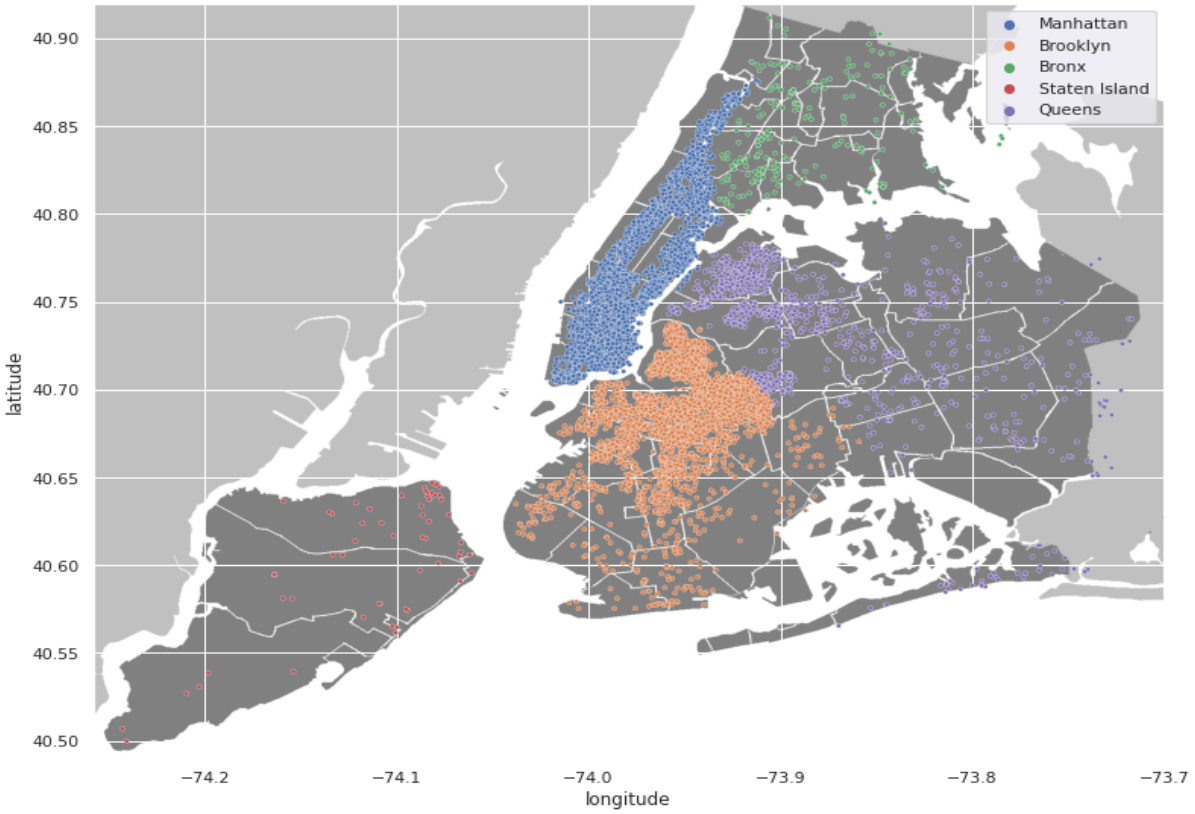
پس از گرفتن 300 نمونه از هر محله و بررسی چندین و چند باره ، اعداد مختلفی در بازه 0.001 تا 0.5 بدست آمد که خیلی واضح نمیتوان گفت فرض قبول است یا نه.تا حدودی می توان گفت میزان review ها در محله های مختلف تفاوت چشمگیری ندارد اما به طور خاص staten Island را بررسی میکنیم که چه چیزی باعث شده میانگین review بالاتری داشته باشد. به نظرم خانه های کم staten Island باعث شده ، خانه با review ی کم در آنجا کمتر باشد که این مورد ، روی میانگین این شهرها تاثیر بگذارد. حال باهم میانگین شهرهای مختلف را می بینیم.



همانطور که می بینیم ، با اینکه تعداد review بالا در شهر های دیگر بسیار بیشتر است اما تعداد زیاد خانه ها ، باعث شده خانه هایی با review کمتر در این شهرها بیشتر شود که این مورد نشان دهنده این است که میانگین تعداد review ها با تعداد خانه های یک شهر رابطه عکس دارد.

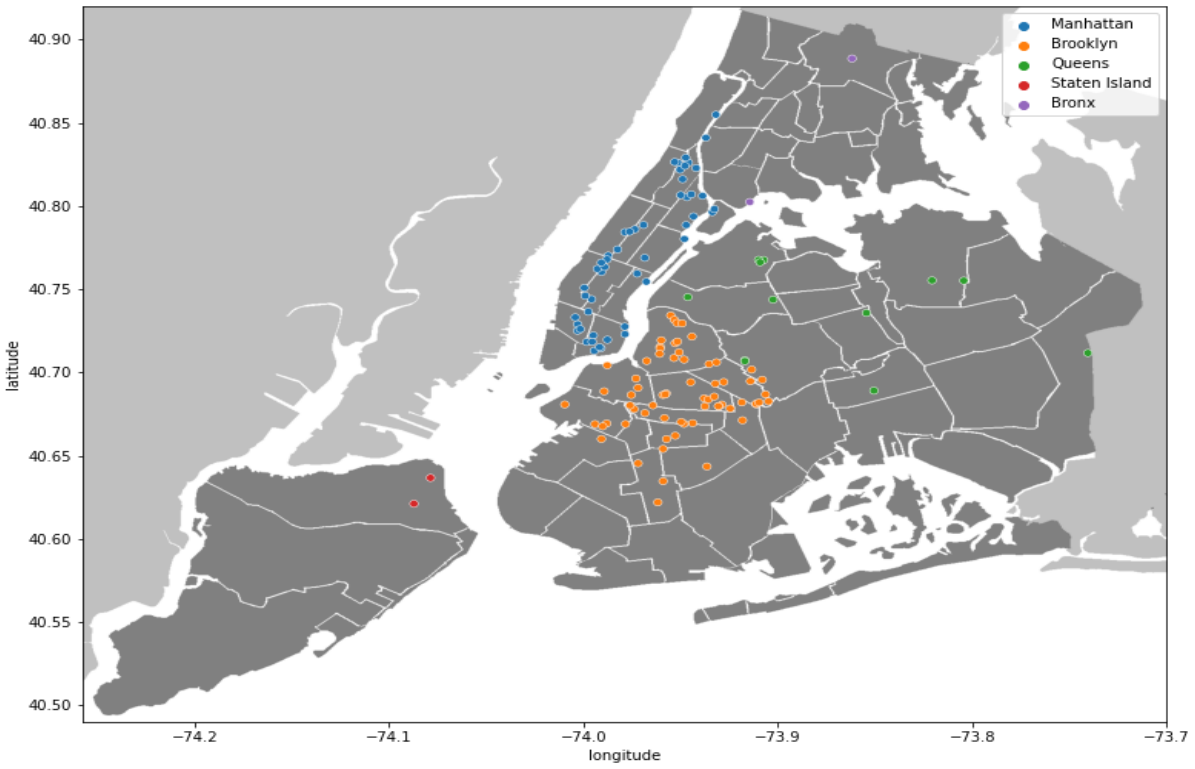
اگر با فرض بیشتر بودن review ها در staten island یک t-test جداگانه روی خانه های staten island و داده اصلی بزنیم مشاهده میکنیم که اکثرا اعدادی کوچک برای p-value به دست می آید که نشانگر آن است که review در staten island بیشتر است.

شکل زیر خانه هایی را نشان می دهد که تعداد Review ها در آن صفر است. فرض ما برای کم بودن review صفر در staten island تا حدودی پذیرفته است.

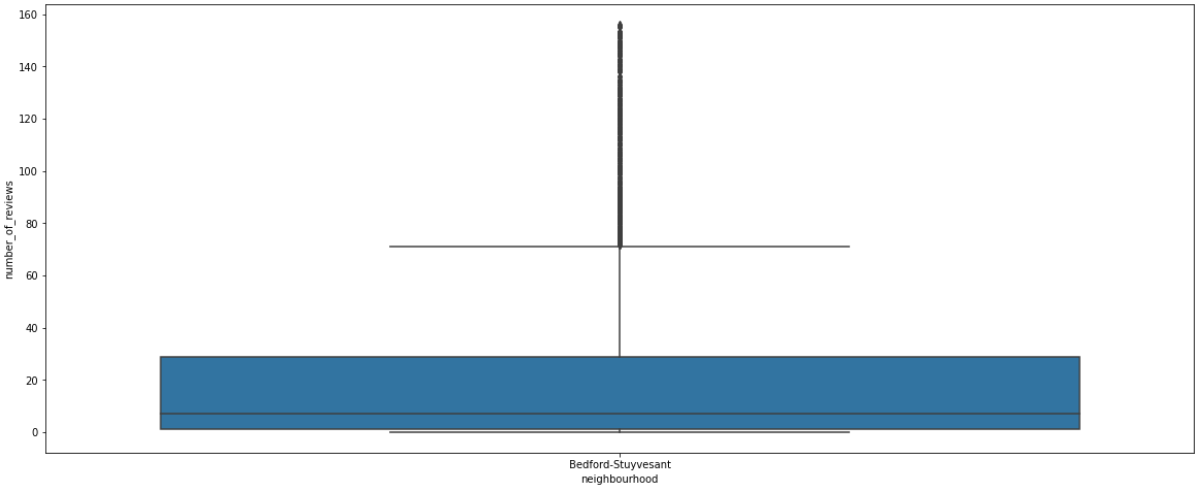


اما در ارتباط با این سوال که بیشترین مشتری ها را کدام میزبانها دارند ، می توانیم از دو ستون کمک بگیریم.

ابتدا می توانیم تعبیر کنیم مکان هایی که بیشترین review را داشته باشند ، پرمشتری ترینند. برای اینکار خانه هایی که بالای 150 review دارند را جدا میکنیم و روی نقشه نمایش می دهیم.



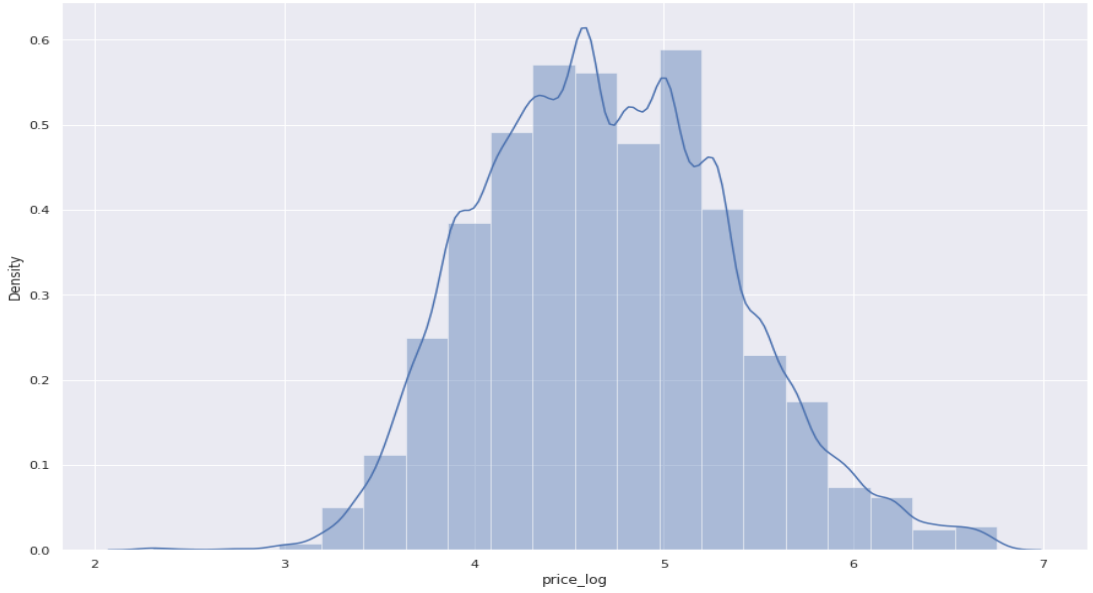
حالا با چند neighbourhood مواجه میشویم که خانه های زیادی از آن در این نقشه هستند. برای مثال ما `Bedford-Stuyvesant` را جدا کردیم و نمودار جعبه ای تعداد review های آن به شکل زیر است.



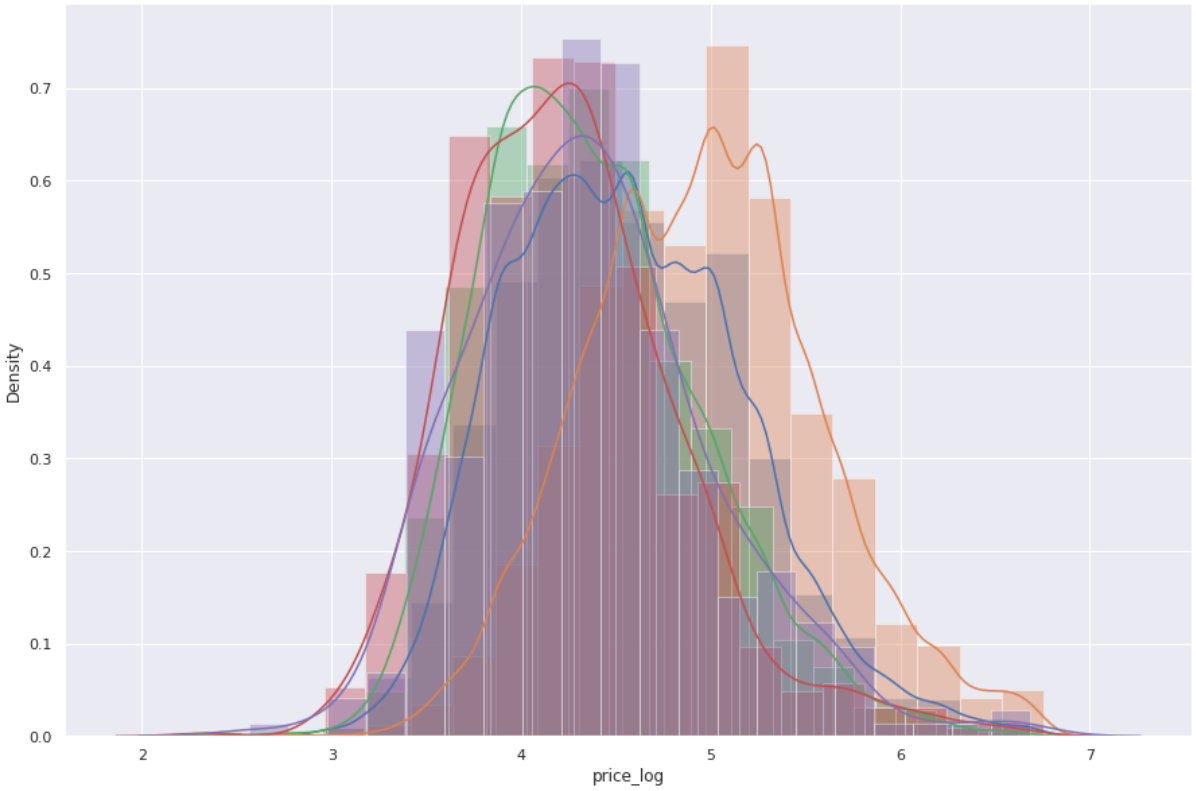
که اگر با بقیه مقایسه شود مقدار بالایی دارد. از این جهت می توانیم با این فرض روی آنها T-test اجرا کنیم و فرض آن است که میزان review های این neighbourhood تفاوت زیادی با داده اصلی دارد. وقتی این کار را روی چند sample انجام دادم ، مشخص است که این فرض درست بوده و میزان p-value بسیار پایین در حد 10 به توان -10 خواهد بود.

دومین روشی که میتوان شهرهای پرترافیک را پیدا کرد استفاده از availability است اما قبل از آن دو کار انجام میدهیم.

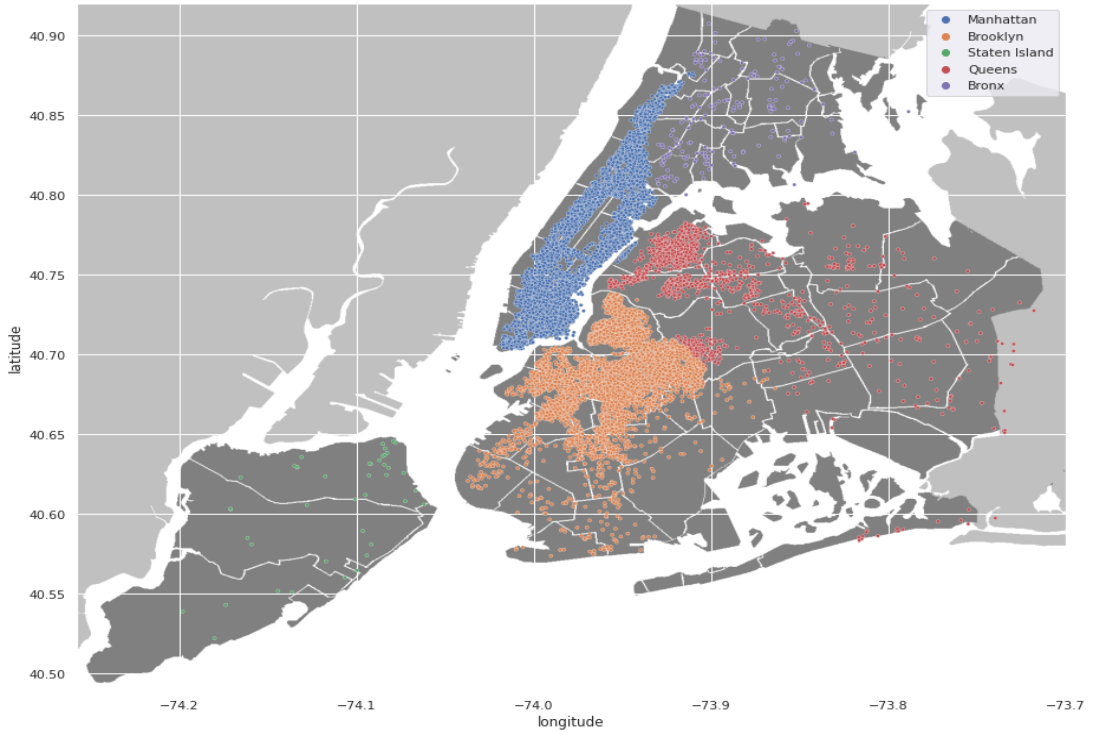
همانطور که گفتیم ستون price از توزیع نرمال پیروی نمی کرد پس آن را با استفاده از log transform نرمال میکنیم تا از آن در تست های آتی استفاده کنیم. نمودار توزیع آن پس از نرمال شدن به شکل زیر است.



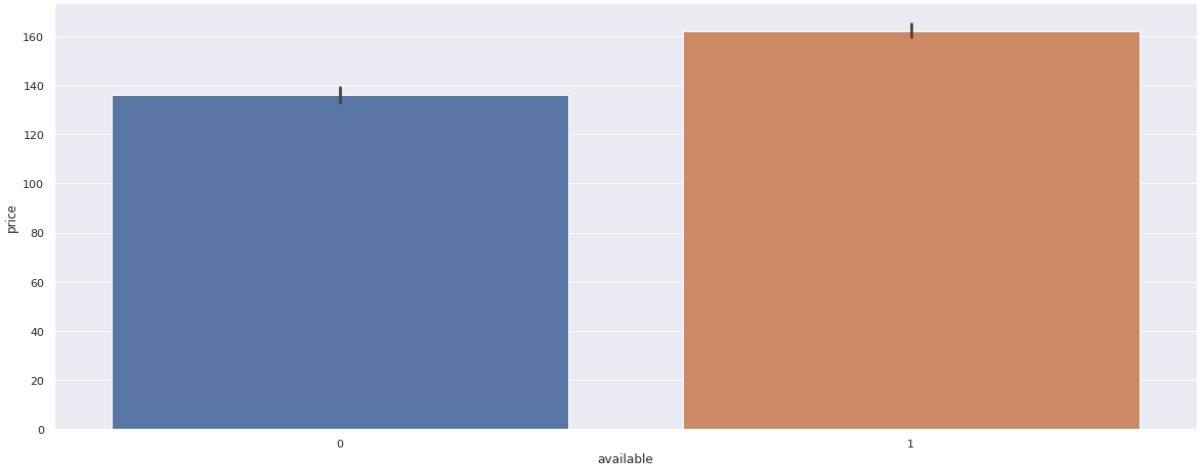
همچنین توزیع قیمت در مناطق مختلف نیز به صورت زیر جداسازی شده است.



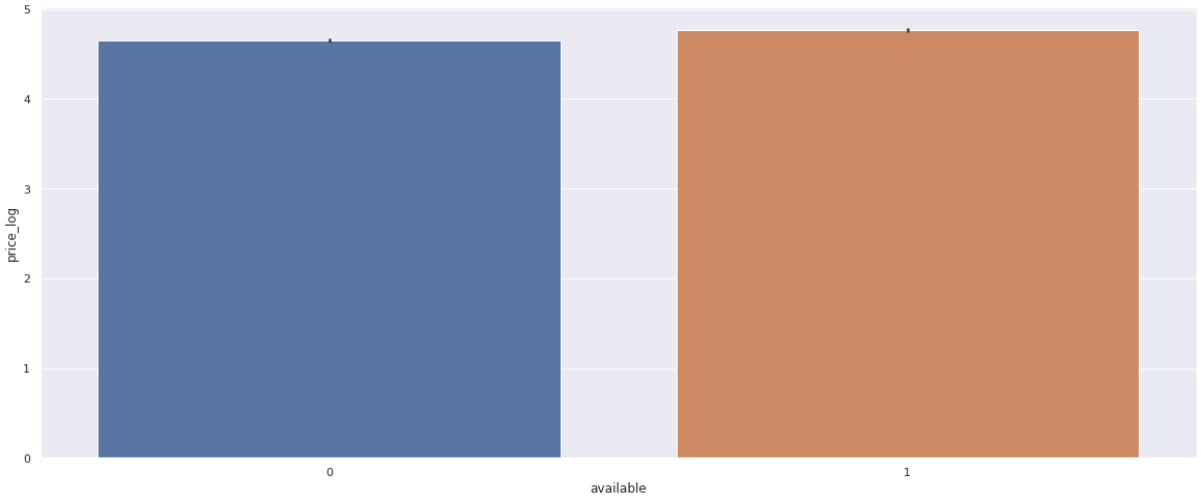
کار دیگری که خواهیم کرد آن است که busy ترین میزبان ها را پیدا کنیم . آنهایی که Availability=0 دارند در واقع هیچ زمان خالی ندارند و از همه شلوغ تر هستند. پس آن ها را جدا کرده و در نقشه نشان می دهیم که به شکل زیر است:



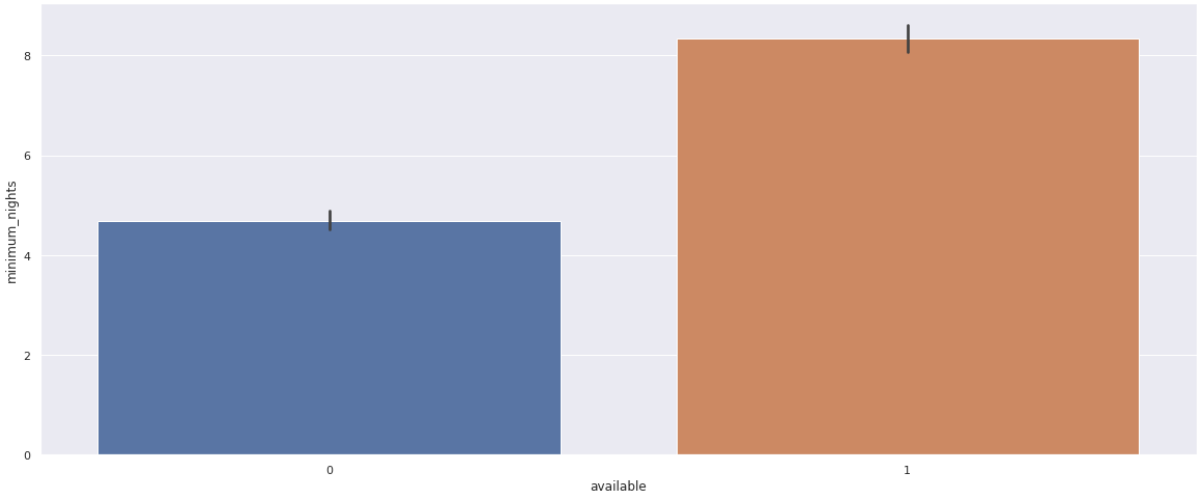
حال شروع به کشیدن نمودار ها برای بررسی تفاوت شلوغ ترین ها نسبت به بقیه میکنیم. نمودار زیر قیمت را نشان میدهد که تفاوت بسیار قابل توجهی بین این دو وجود ندارد ولی به طور کل خانه های اشغال شده ، قیمت کمتری دارند. وقتی از t-test هم با فرض تفاوت فاحش قیمت ها استفاده کردیم نیز p-value معمولا در sample ها مقدار بالایی داشت.



نمودار available نسبت به price_log (همان توزیع نرمال price) هم به صورت زیر است که باز هم تفاوت آنچنانی ندارد.



اما مواردی که در شلوغ بودن تاثیر داشتند. اول از همه minimum nights بود که با فرض اینکه هر چقدر minimum nights کمتر باشد ، خانه ها شلوغ تر هستند پیش رفتیم و t-test را روی 500 sample اجرا کردیم که p-value معمولا عددی کوچکتر از 0.05 بود که نشان دهنده آن است که هرچقدر شب های کمتری برای مینیمم اجاره کردن در نظر بگیریم ، اقامتگاه شلوغ تر خواهد بود. نمودار هم آن را نشان می دهد.

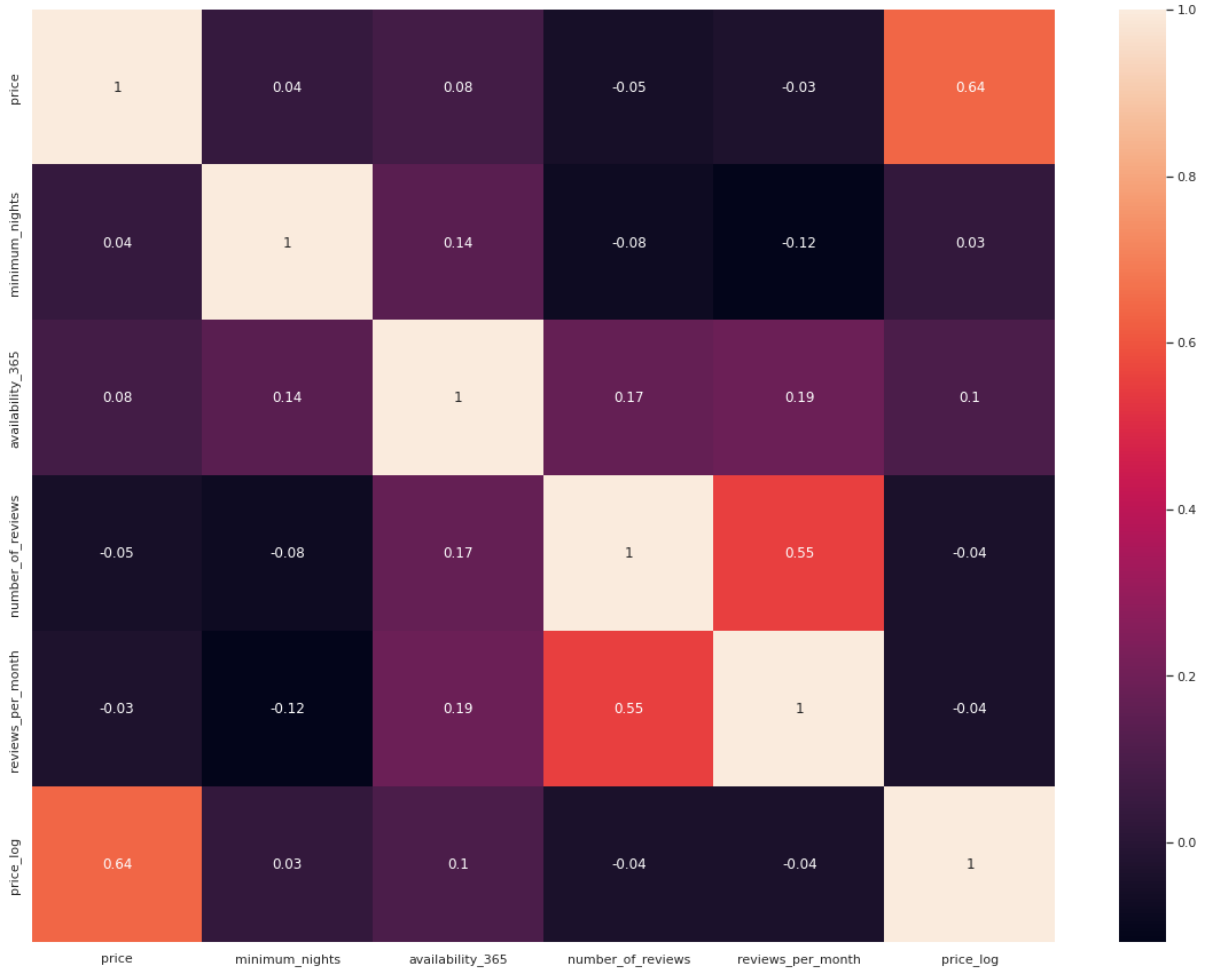


پس از آن calculated_host_listings_count را در نظر میگیریم و مانند ستون قبلی روی آن T-test و wilcoxon میزنیم که باز هم میزان بدست آمده بسیار کم است (در حدود 0.0004) و فرض مورد قبول است و calculated_host_listings_count در حالتی که همه روزها پر هستند بسیار کمتر از حالت عادی است. یعنی میزبان هایی که خانه های کمتری دارند معمولا خانه هایشان پر است. نمودار آن نیز در کد موجود است.

البته بیشتر تست های بالا با تغییر بر روی دیتا فریم انجام شدند و می توان با اندکی تغییر خیلی از آنها را با correlation نیز بررسی کرد که در پایین میبینیم.

تاثیر نوع اتاق بر availability هم با متد describe بررسی شد که به نظر تفاوت چندانی نکرده است و فرضیه ای روی آن نداشتم که تست کنم.

نمودار Correlation ستون ها نیز به صورت زیر است:



اگر بخواهیم توصیفی کوتاه در ارتباط با آن داشته باشیم ، واضحترین ارتباطات بین price_log و price همچنین بین review_per_month و number_of_reviews است که دلیل آن بسیار واضح است.

اما روابط دیگری که قابل بررسی است، تاثیر عکس review_per_month و minimum_night است که هر چه شب های کمتری قابل دسترسی باشد ، review های بیشتری انجام خواهد شد.

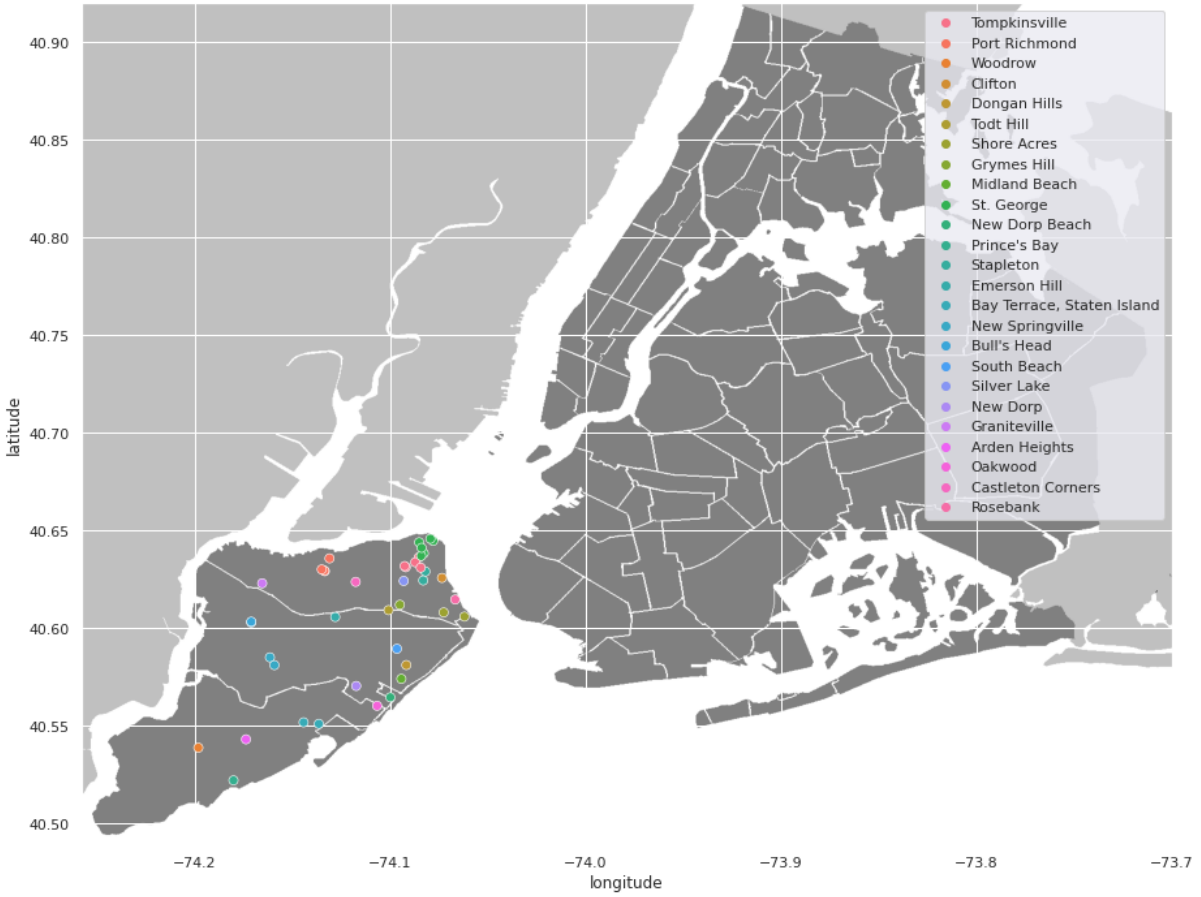
نکته جالب و عجیب آن است که هرچه review ها بیشتر باشد ، availability نیز بیشتر بوده است که این کمی سخت قابل درک است. حدسی که من دارم این است که شاید خانه هایی که همیشه پر هستند توسط چند مشتری خاص پر میشوند که باعث میشود تعداد نظرات افزایش نیابد.

در ارتباط با همین موضوع `pearsonr` و `spearmanr` را برای 100 نمونه انجام میدهیم که معمولا p-value کوچکی میداد و نمی توان فرض را به طور کامل پذیرفت.

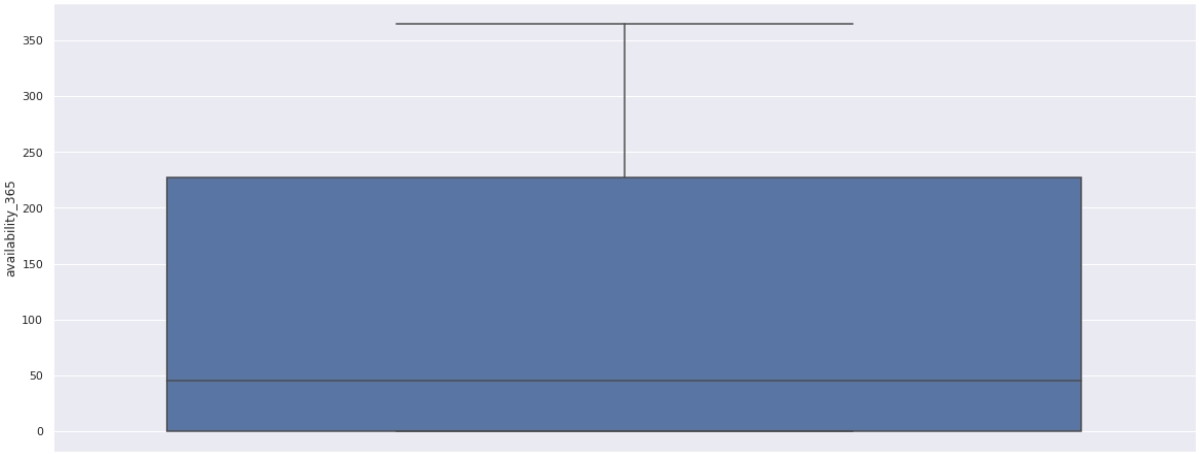
همانطور که گفتیم ابتدا review ها را بررسی کردیم و چند شهر که ترافیک بالایی بر اساس review دارند را در کد مشخص کردیم و روی یکی از آنها تستی هم زدیم.

اما برحسب availability ، من همه محله هایی که خانه ای با availability=0 داشتند را در نقشه نمایش دادم و سپس با تعداد ک خانه در آن محله یا neighbourhood بررسی کردم که هرکدام تغییر قابل توجهی داشتند ، بررسی آماری و

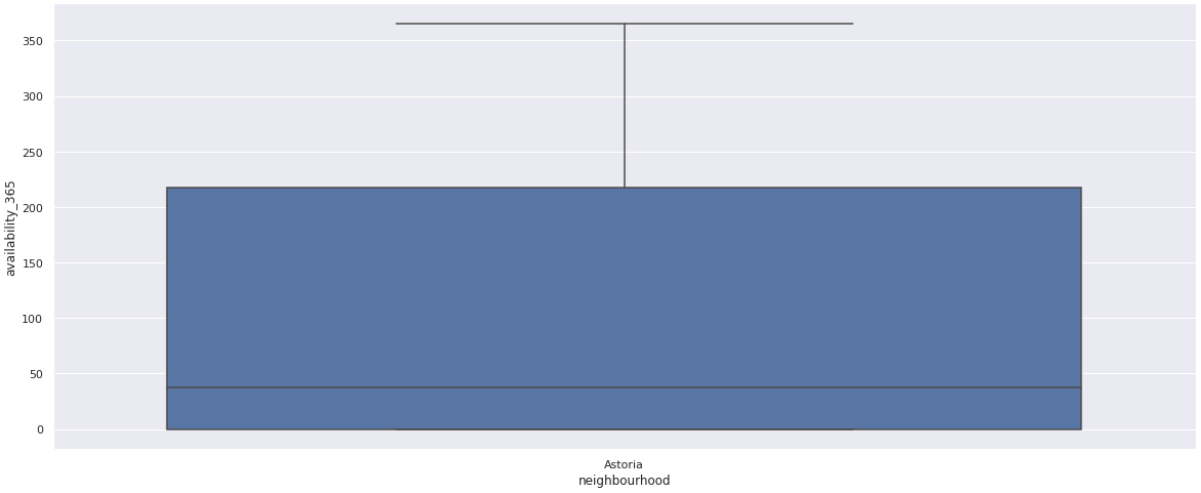
تست روی آنها انجام دهم. همه نقشه ها و تعداد خانه ها در کد موجود است ولی برای نمونه نقشه خانه های پر Staten_island به شکل زیر است :



اول از همه در زیر ، boxplot ستون avalibilty_365 را با هم میبینیم تا با شهر های دیگر مقایسه کنیم :

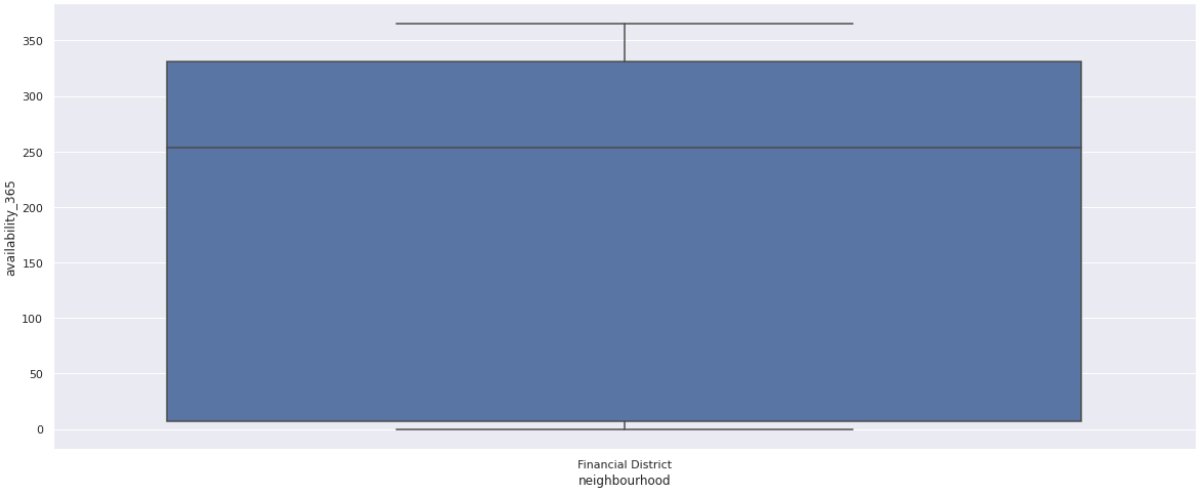


اولین شهری که بررسی می کنیم Astoria ست که به خاطر بررسی های روی جدول به نظر باید availability پایینی داشته باشد اما boxplot آن به شکل زیر است:



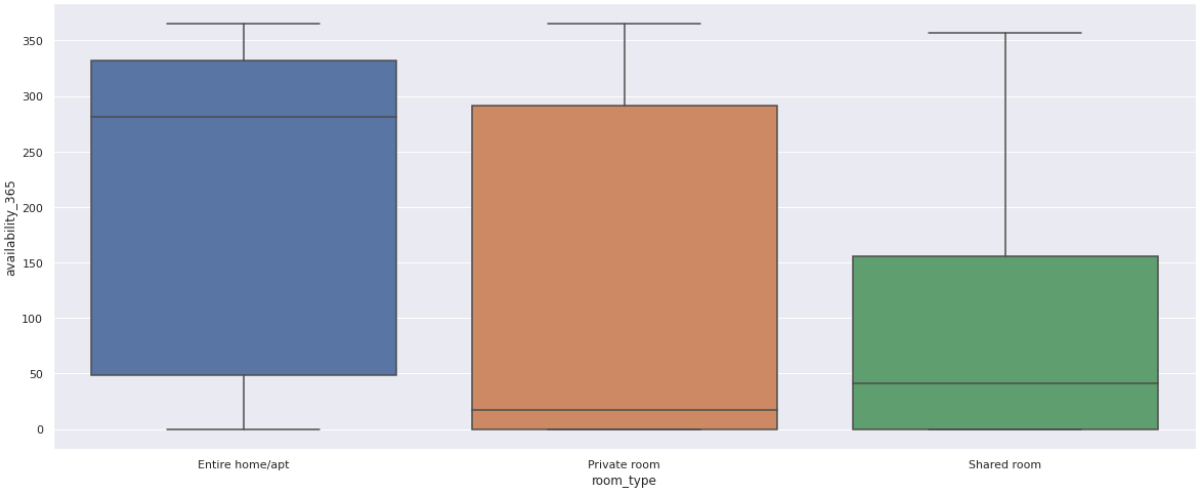
که به نظر خیلی تفاوت فاحشی با داده اصلی ندارد و و وقتی با داده اصلی با استفاده از t-test مقایسه میشود، به علت p-value بالا فرض اولیه ما رد و astoria شهر با ترافیک بالا محسوب نمیشود.

اما شهر Financial District در منهتن بررسی بعدی ماست که انتظار ما با توجه به مشاهدات آن است که availability بالایی داشته باشد. boxplot آن به شکل زیر است:



این بار احتمالا باید T-test ما pvalue پایینی داشته باشد که همینطور هم هست و اولین رقم اعشار آن در حدود ده یا بیست قابل دیدن است پس این فرض قبول است و این شهر از ترافیک زیادی برخوردار نیست و اکثر مواقع available می باشد.

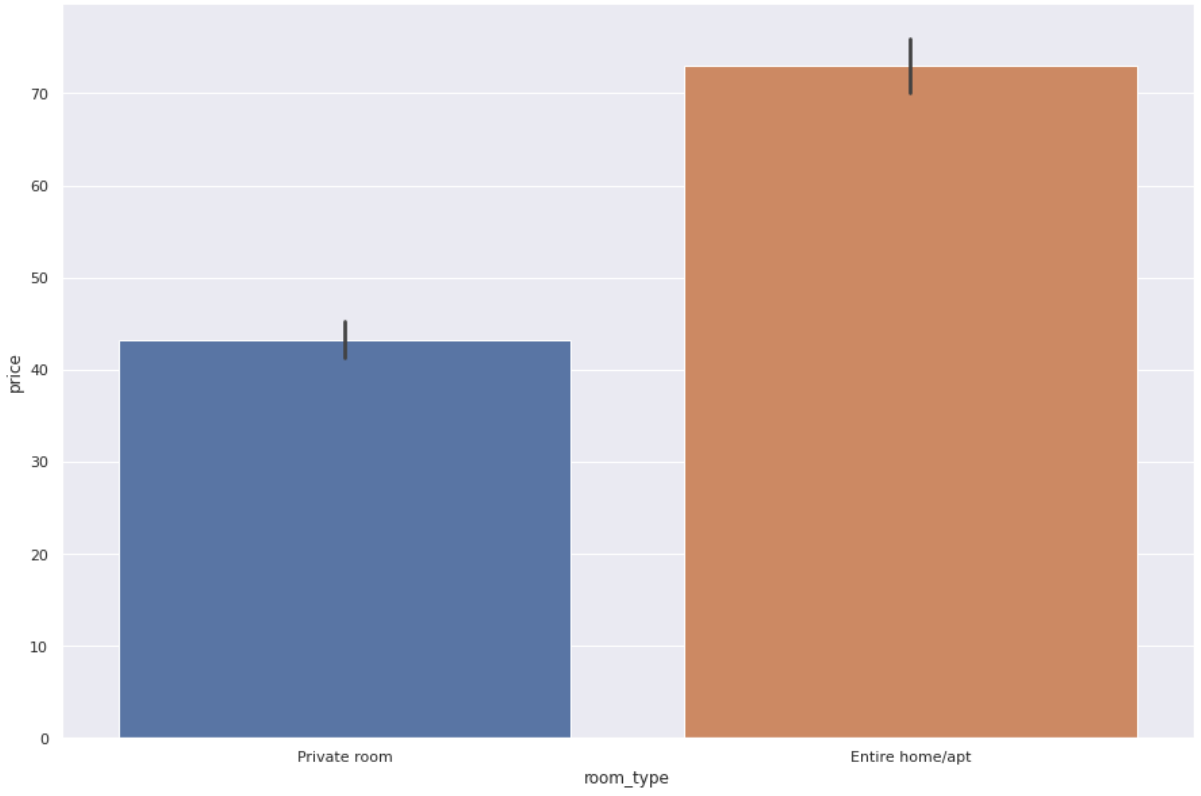
حدسی که میتوانم برای کم بودن ترافیک در آن بزنم آن است که private room ها در آن خیلی کم است. به طور کلی تعداد خانه ی کامل یا آپارتمان در کل داده نزدیک به تعداد اتاق خصوصی است ولی در این شهر بسیار بیشتر است و همانطور که در نمودار زیر میبینیم ، همین آپارتمانها باعث availability بالاست.



به طور کل با استفاده از جدول و میانگین ها می توان فرض گرفت و با بررسی های آماری آن را اثبات کرد که من به این چند نمونه neighbourhood بسنده میکنم.

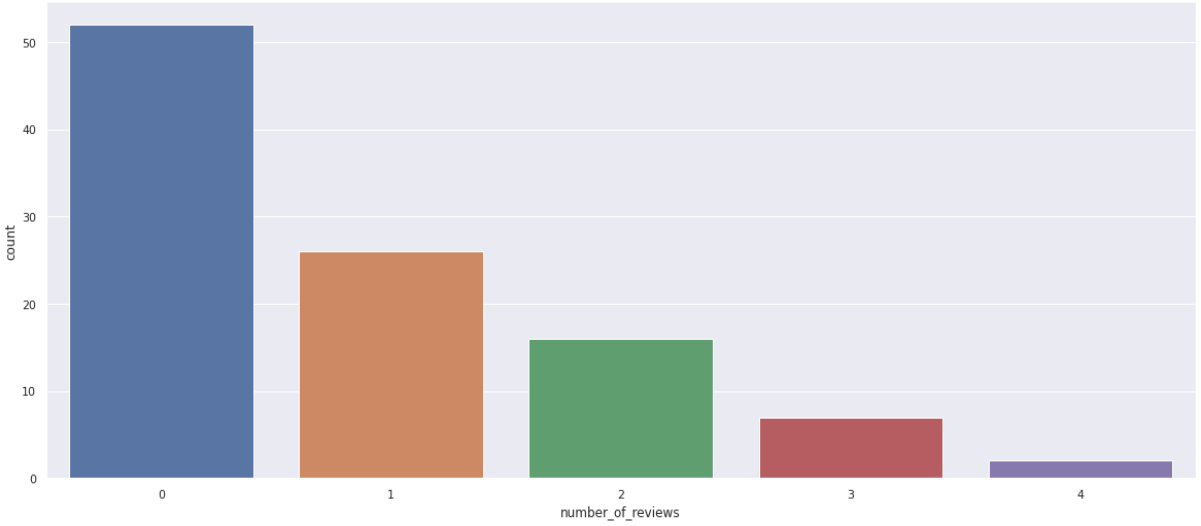
برای بررسی میزبان های متفاوت تعداد واحدی که هر میزبان در اختیار دیگران قرار می دهد را بررسی می کنیم. بیشترین واحد مربوط به میزبان با ایدی 137358866 می باشد که 103 خانه برای اجاره دارد. من در این بخش برای همین میزبان مقادیر مختلف ستونها و در واقع مشخصات خانه هایش را بررسی کردم.

این میزبان 101 اتاق خصوصی و تنها 2 خانه کامل یا اپارتمان دارد و تفاوت میزان قیمت این 2 مدل خانه اش در نمودار زیر مشخص شده است.

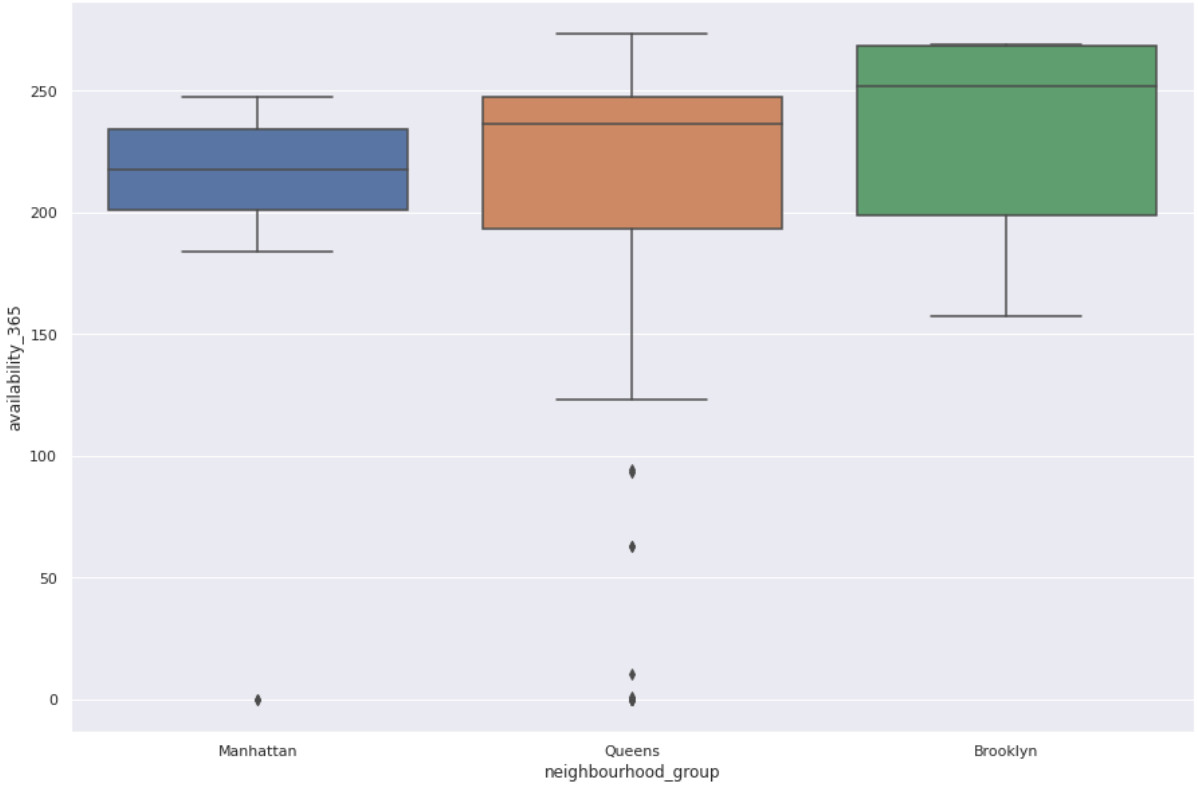


با توجه به میزان قیمت در کل داده ها که نمودارش در کد اصلی قابل مشاهده است می توان اینگونه نتیجه گرفت که خانه های این میزبان از قیمت کمتری برخوردار است.

تعداد بررسی های خانه های این میزبان تیز در نمودار زیر آمده است که بیشترین 4 تا بوده است و تعدادش نیز حالت نزولی دارد.

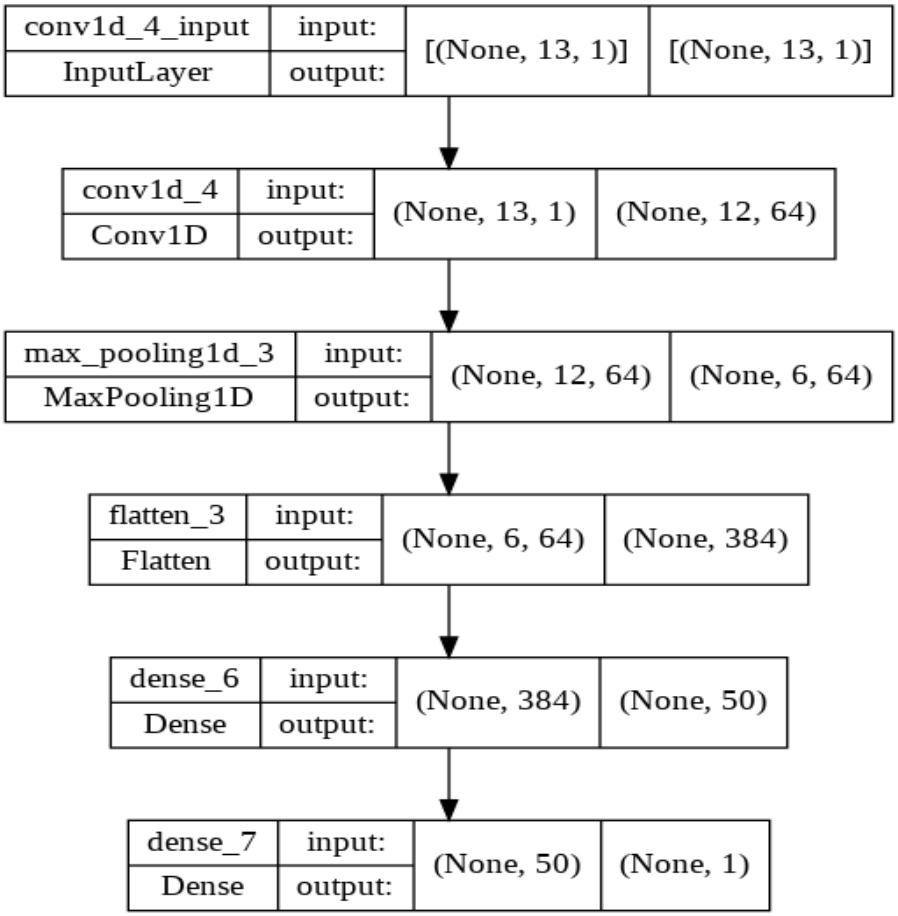


تعداد روزهایی که خانه ها در دسترس هستند را در ارتباط با مناطقی که این میزبان در انها خانه دارد در نمودار زیر منعکس شده است که روزهای در دسترس بالا هستند. با توجه به این مورد و نمودار بالا که نشان دهنده تعداد ریویوهای پایین خانه های این میزبان داشت می توان اینگونه نتیجه گرفت که این میزبان پرطرفدار نیست و خانه های شلوغی ندارد که یک از نشانه هایش پایینی قیمت است و شاید نشان می دهد خانه های این میزبان کیفیت کافی را ندارند.

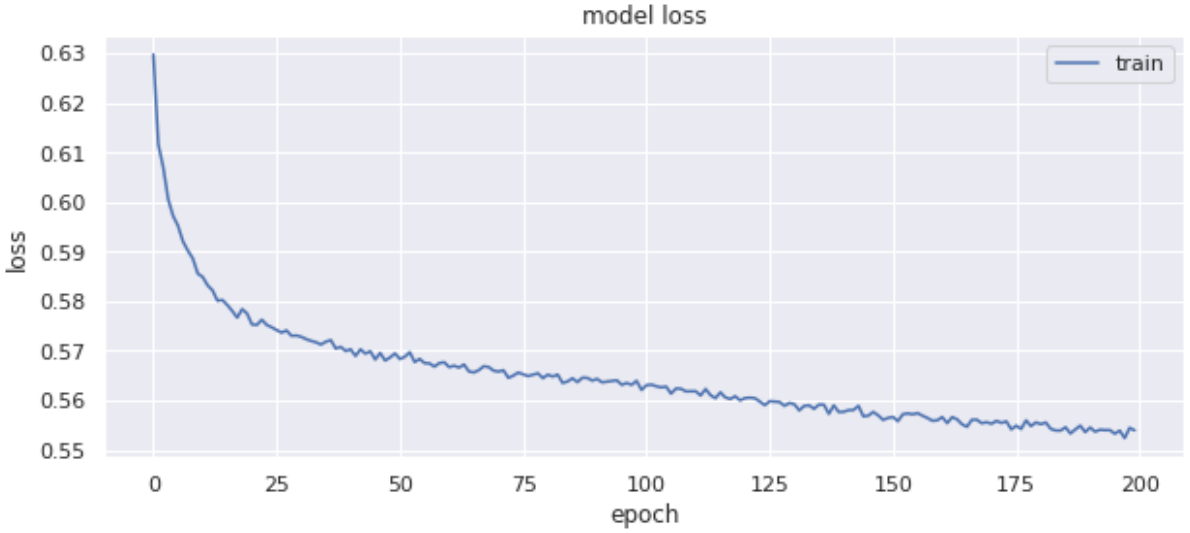


برای ساختن مدلی برای پیش بینی از 2 شبکه عصبی استفاده کردیم. اولین شبکه ای که از آن استفاده کردیم cnn بود. قبل از آن باید داده ها را برای دادن به مدل برای پیش بینی اماده می کردیم. ستونهای که پراکندگی بالایی داشت و یا غیرمرتبط بود از دیتاست حذف شد.

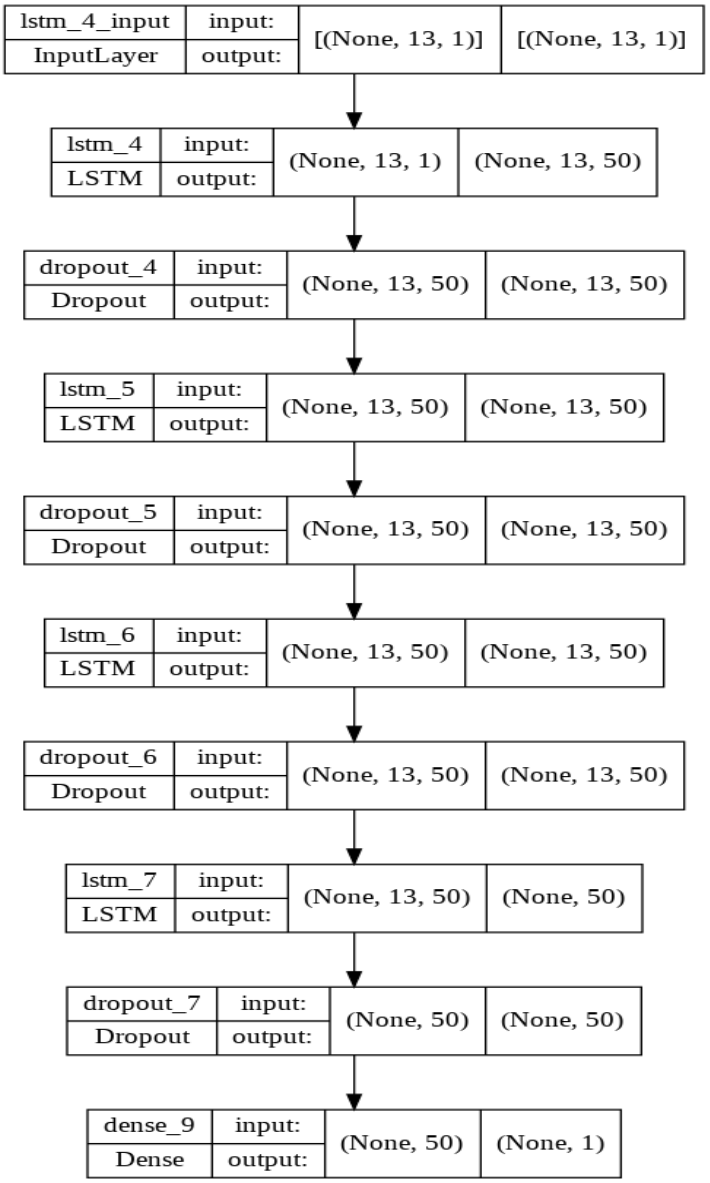
سپس neighbourhood group و room type را مقادیر کتگوریکال دارند را با one hot انکد می کنیم. پس از جداسازی داده ی هدف یا تارگت که price می باشد . همچنین اسکیل کردن داده ها با استفاده از استاندارد اسکیلر داده ها برای دادن به مدل آماده هستند. شمای کلی مدل cnn ما به شکل زیر است.



با استفاده از اپتیمایزر ادام مدلمان را فیت می کنیم و خطا را نیز mse در نظر میگیریم که نمودار خطا در ایپاک ها به شکل زیر است که در حال کاهش است و مدل در حال یادگیری می باشد.



مدل بعدیمان نوع خاصی از RNN ها با نام LSTM می باشد که شمای کلی آن به صورت زیر می باشد.



این مدل را نیز با اپتیمایزر ادام فیت می کنیم و خطا را برحسب MSE می سنجیم. در این مدل از متد EARLY stopping نیز استفاده می کنیم که پس از 12 امین ایپاک مدل متوقف می شود که نمودار خطاهای آن به شکل زیر است که در حال کاهش و یادگیری می باشد.