به نام خدا

محمد زیاری – (97222047)

تاریخ: 1400/8/2

گزارش تمرین سری اول

درس: مبانی علوم داده

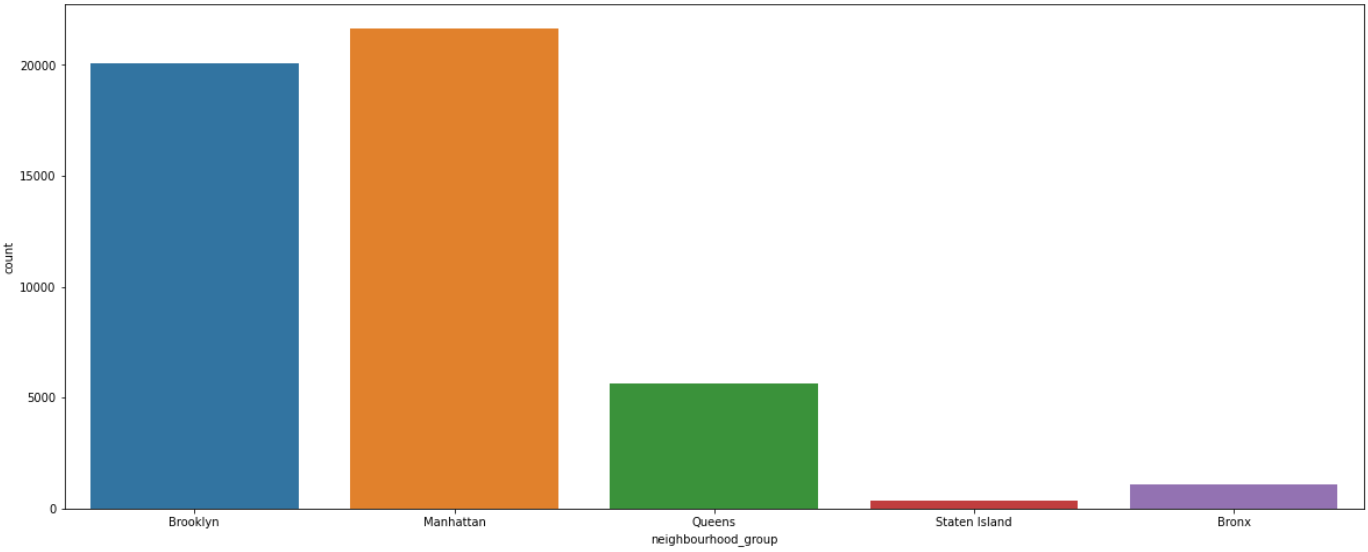
استاد درس: دکتر خردپیشه

1. مجموعه داده های Airbnb

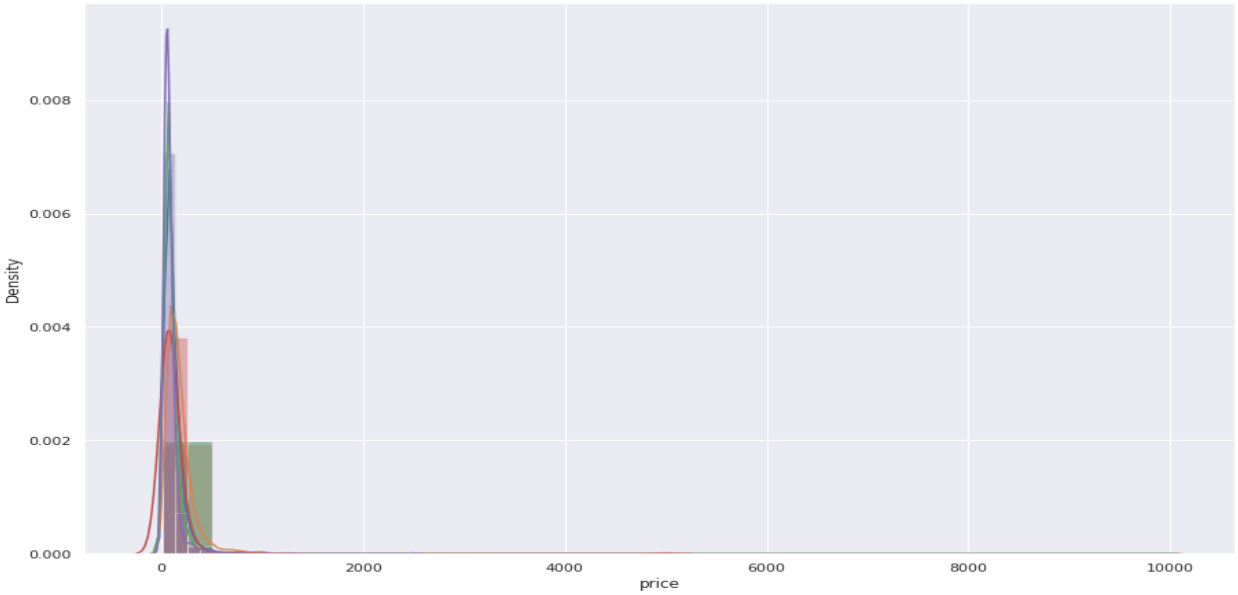
داده های ما در این سوال خانه ها و اقامتگاه هایی است که با بررسی داده ها با تست های آماری و هم چنین جدول ها و نمودارها ارتباط ستونهای مختلف را بررسی میکنیم.

قبل از شروع کار اصلی ، ابتدا داده ها را از kaggle میخوانیم و در بخش پاکسازی داده ها تنها کاری که انجام دادم مربوط به پاک کردن داده های با price=0 بود که 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.

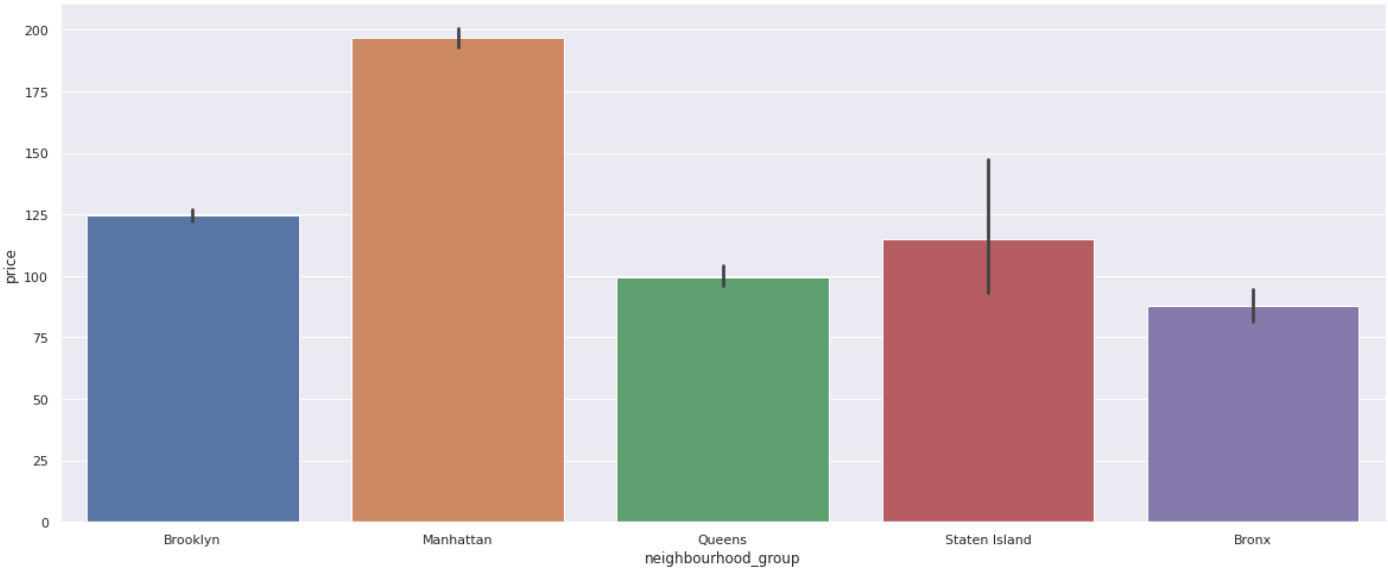
در بلاک بعدی ارتباط محله یا neighbourhood group های مختلف را با ستون یا feature های دیگر بررسی میکنیم. نمودار پایین ، تعداد خانه های هر منطقه را نشان می دهد.



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 نمی 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.

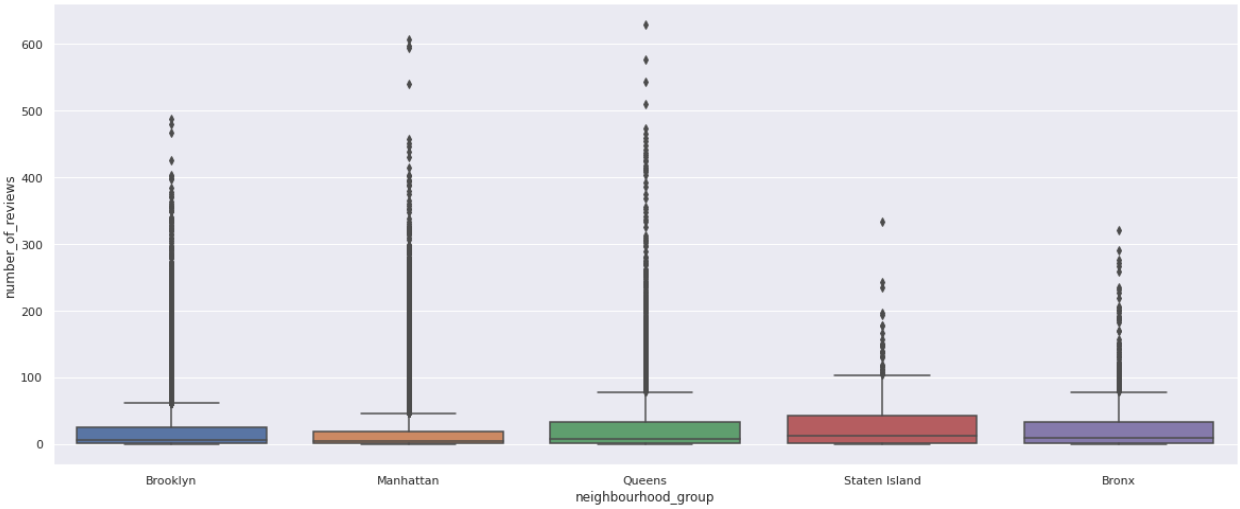


گرفته شد . مثلا آخرین نمونه p-value=0.4 داشت و کمترین مقدار p-value نیز 0.04 بود ، که به نسبت اینکه

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.

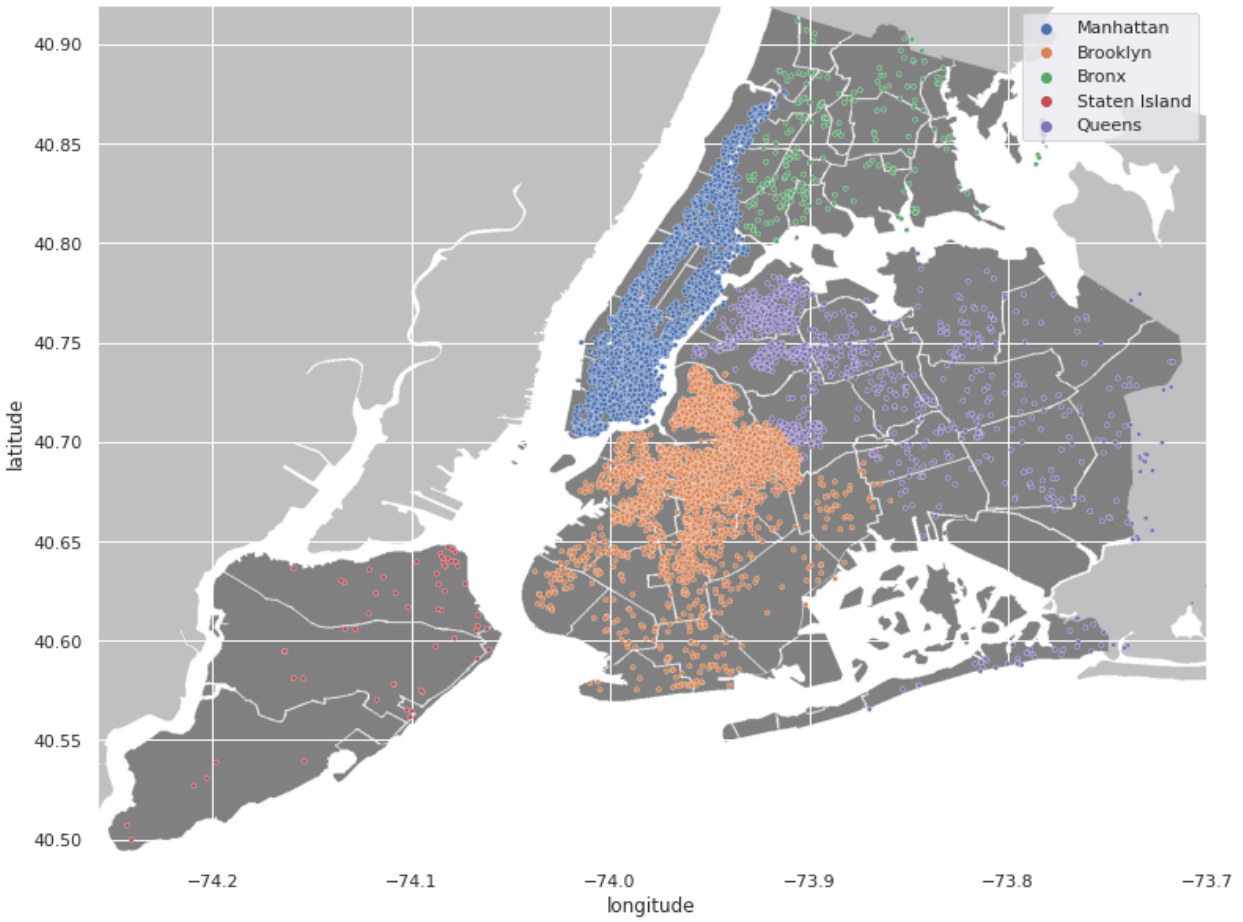
پس از گرفتن 300 نمونه از هر محله و بررسی چندین و چند باره ، اعداد مختلفی در بازه 0.001 تا 0.5 بدست آمد که خیلی واضح نمیتوان گفت فرض قبول است یا نه.تا حدودی می توان گفت میزان review ها در محله های مختلف تفاوت چشمگیری ندارد اما به طور خاص staten Island را بررسی میکنیم که چه چیزی باعث شده میانگین review 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.

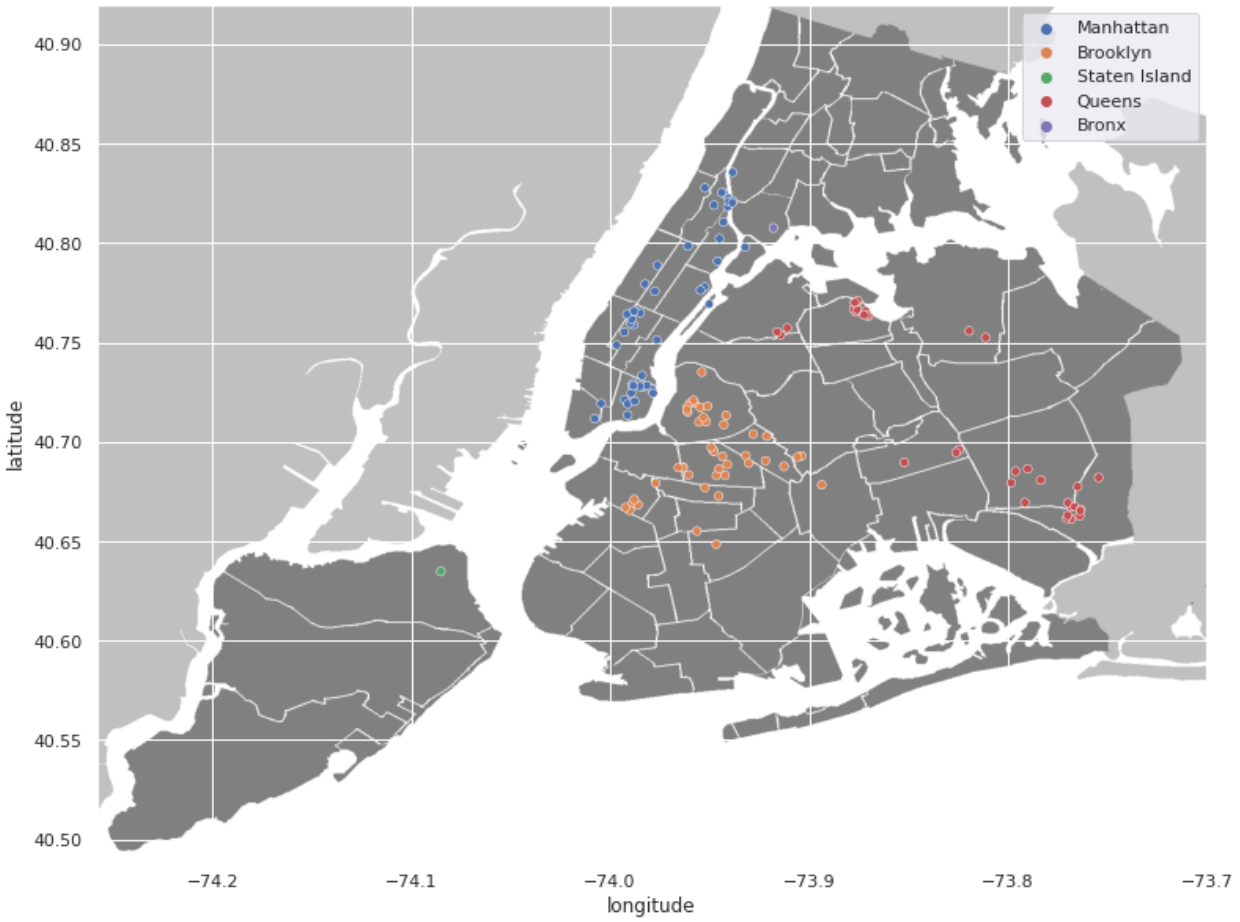
اگر با فرض بیشتر بودن review ها در staten island یک t-test جداگانه روی خانه های staten island و داده اصلی بزنیم مشاهده میکنیم که اکثرا اعدادی کوچک برای p-value به دست می آید که نشانگر آن است که review در staten island بیشتر است.

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.

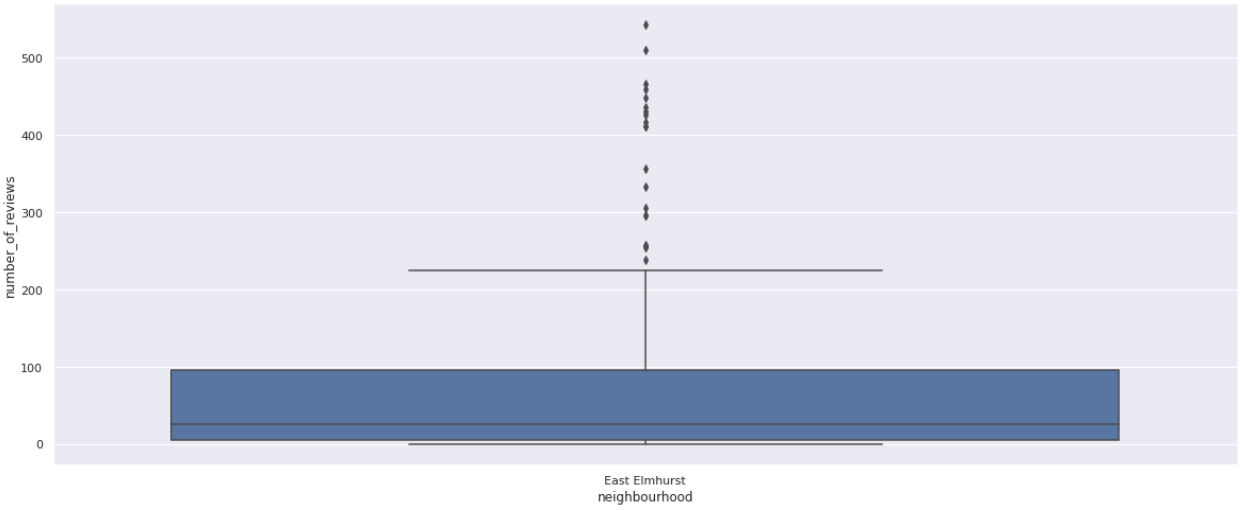


اما در ارتباط با این سوال که پرترافیک ترین قسمت ها کجاست ، می توانیم از دو ستون کمک بگیریم.

ابتدا می توانیم تعبیر کنیم مکان هایی که بیشترین review را داشته باشند ، پرترافیک ترینند. برای اینکار خانه هایی که بالای 300 review دارند را جدا میکنیم و روی نقشه نمایش می دهیم.



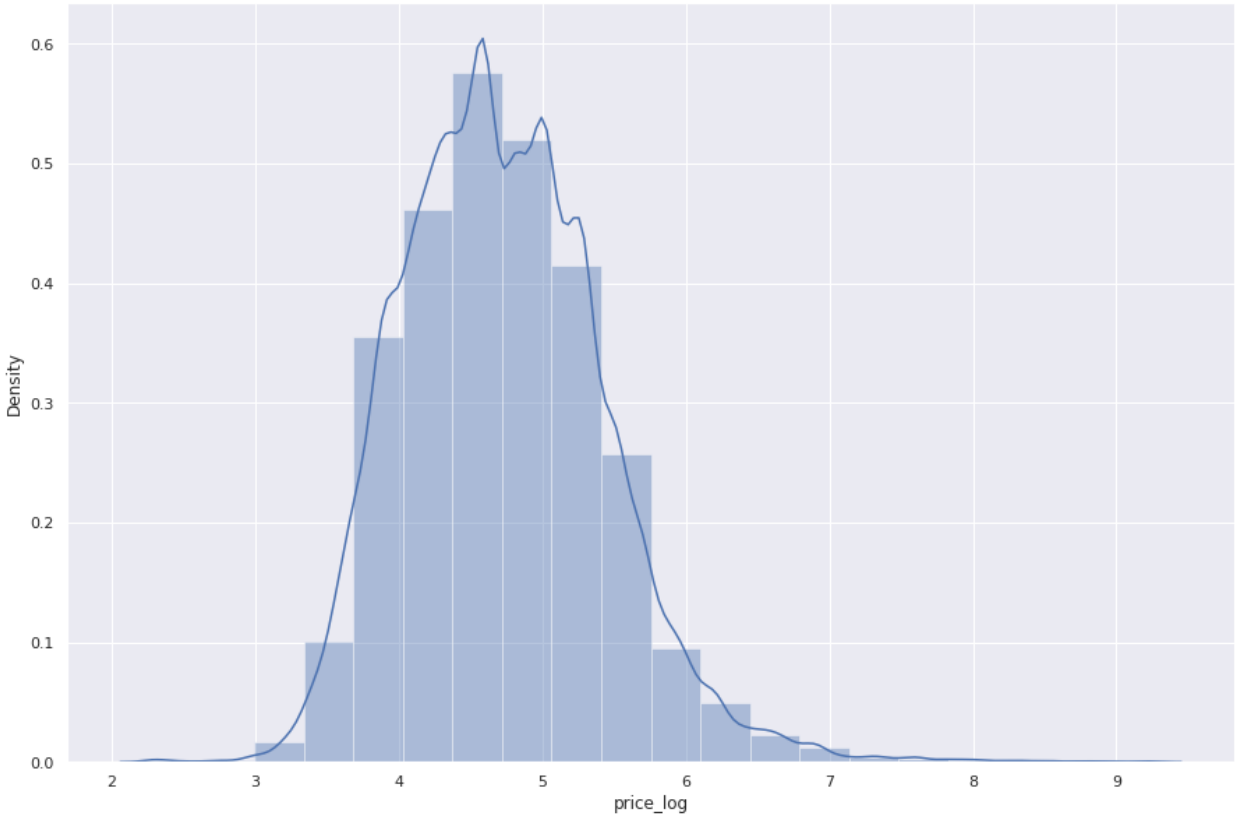
حالا با چند neighbourhood مواجه میشویم که خانه های زیادی از آن در این نقشه هستند. برای مثال ما East Elmhurst را جدا کردیم و نمودار جعبه ای تعداد review های آن به شکل زیر است.



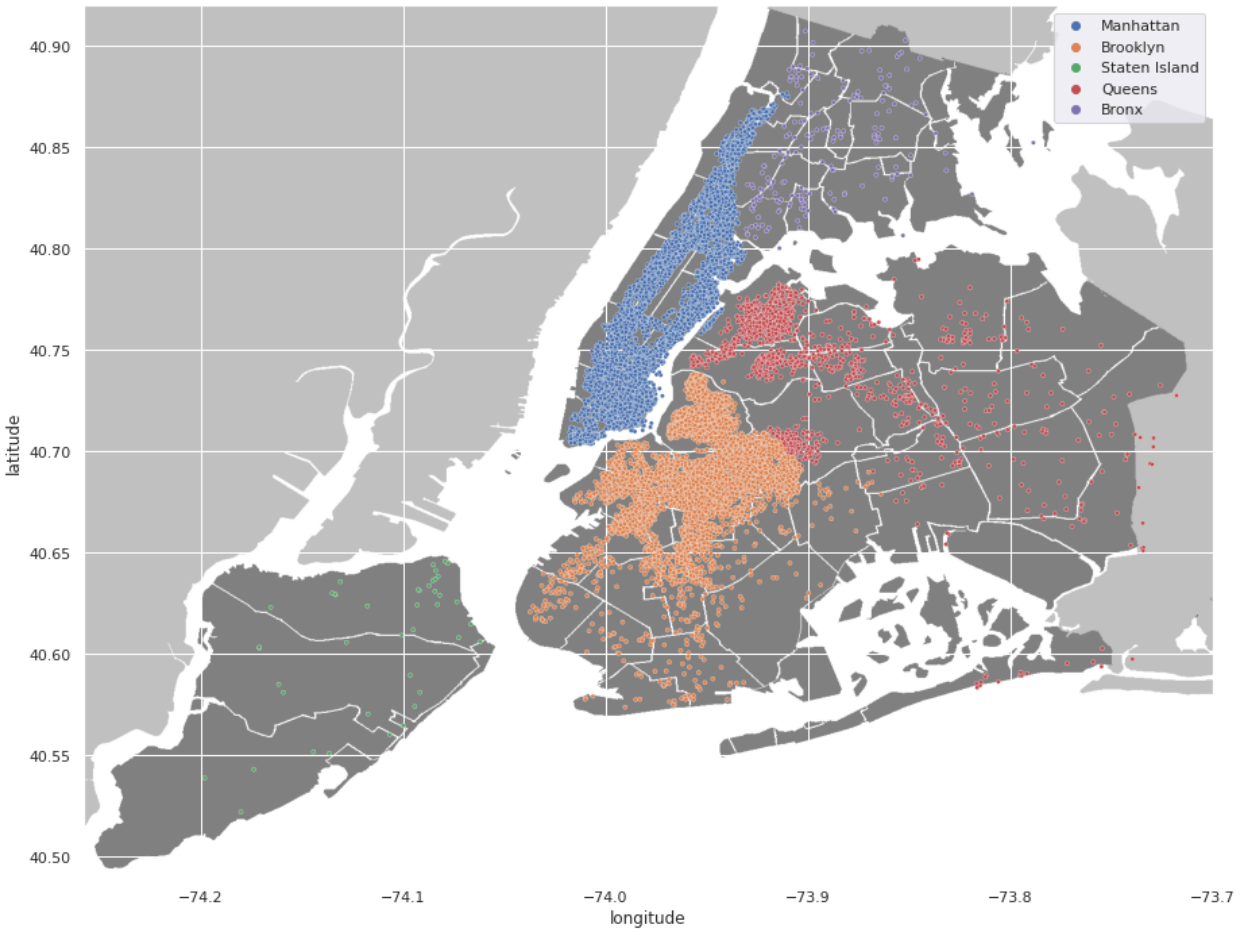
که اگر با بقیه مقایسه شود مقدار بالایی دارد. از این جهت می توانیم با این فرض روی آنها T-test اجرا کنیم و فرض آن است که میزان review های این neighbourhood تفاوت زیادی با داده اصلی دارد. وقتی این کار را روی چند sample انجام دادم ، مشخص است که این فرض درست بوده و میزان p-value بسیار پایین در حد 10 به توان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

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

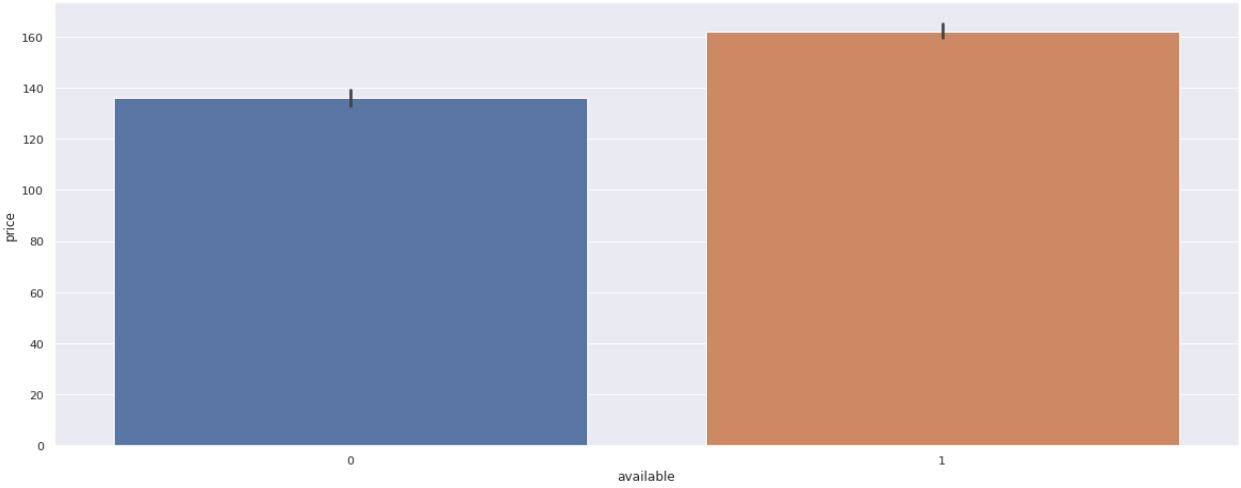
همانطور که گفتیم ستون price از توزیع نرمال پیروی نمی کرد پس آن را با استفاده از log transform نرمال میکنیم تا از آن در تست های آتی استفاده کنیم. نمودار توزیع آن پس از نرمال شدن به شکل زیر است.



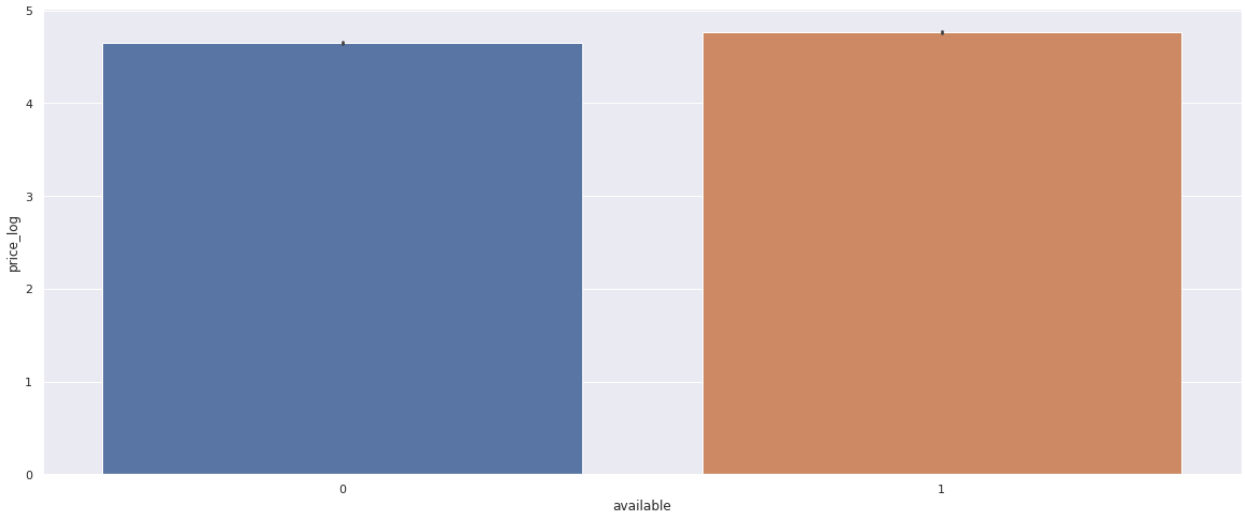
کار دیگری که باید بکنیم آن است که busy ترین میزبان ها را پیدا کنیم . آنهایی که Availability=0 دارند در واقع هیچ زمان خالی ندارند و از همه شلوغ تر هستند. پس آن ها را جدا کرده و در نقشه نشان می دهیم که به شکل زیر است:



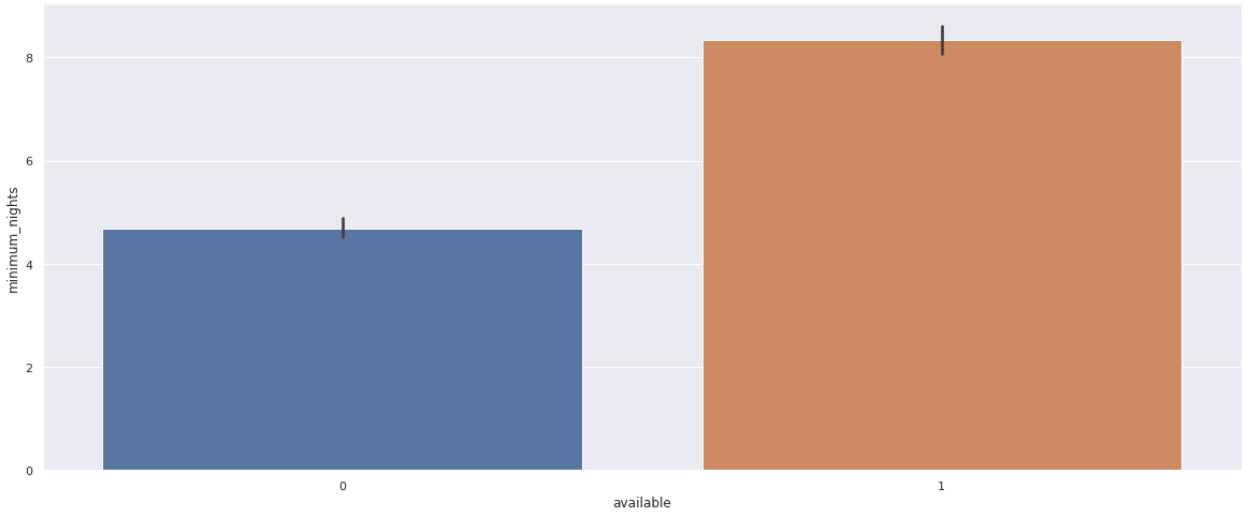
حال شروع به کشیدن نمودار ها برای بررسی تفاوت شلوغ ترین ها نسبت به بقیه میکنیم. نمودار زیر قیمت را نشان میدهد که تفاوت بسیار قابل توجهی بین این دو وجود ندارد ولی به طور کل خانه های اشغال شده ، قیمت کمتری دارند. وقتی از t-test هم با فرض تفاوت فاحش قیمت ها استفاده کردیم نیز p-value معمولا در sample ها مقدار 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 بود 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 sample 500 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.



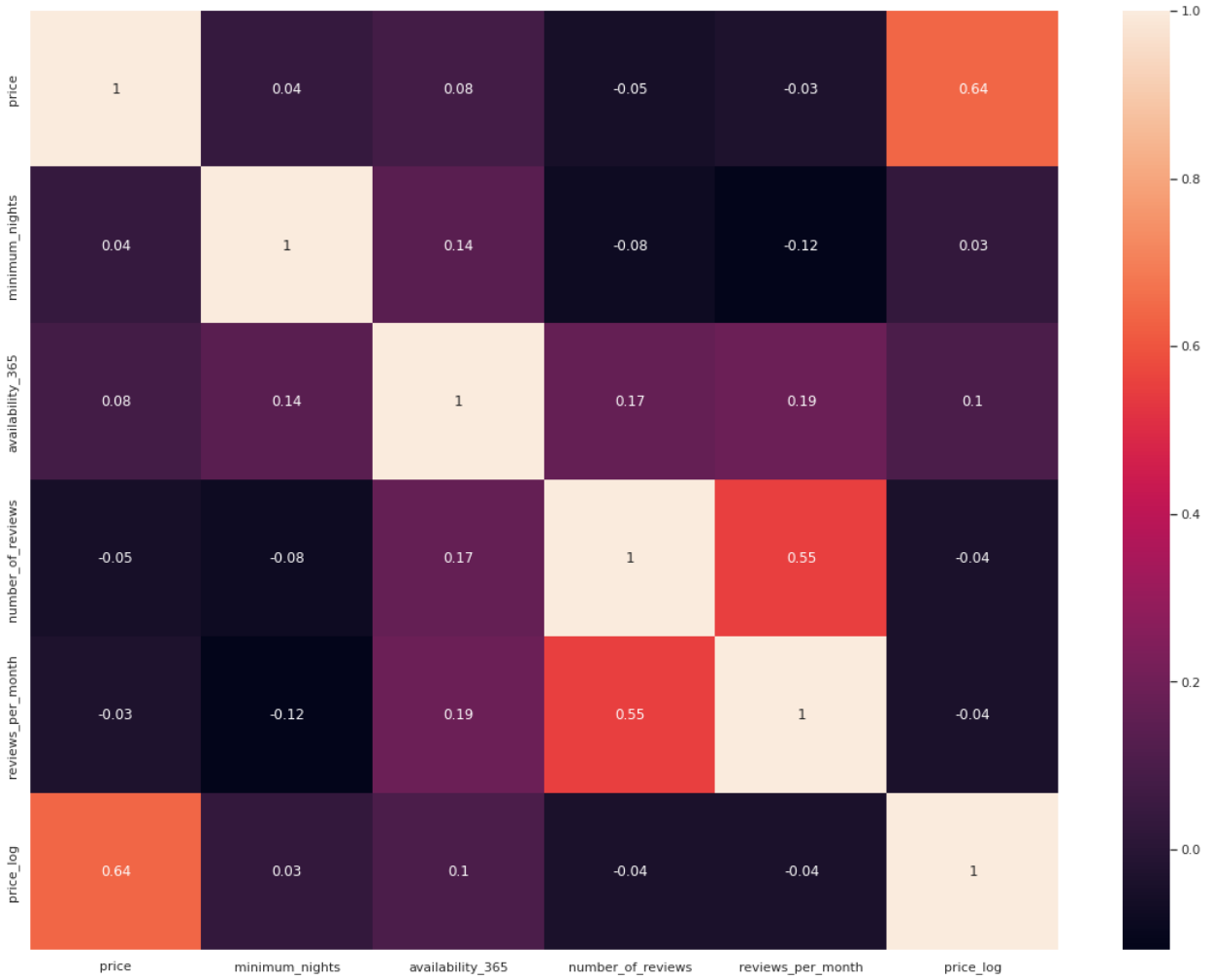
پس از آن calculated_host_listings_count را در نظر میگیریم و مانند ستون قبلی روی آن T-test و wilcoxon میزنیم که باز هم میزان بدست آمده بسیار کم است (در حدود 0.0004) و فرض مورد قبول است و

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.

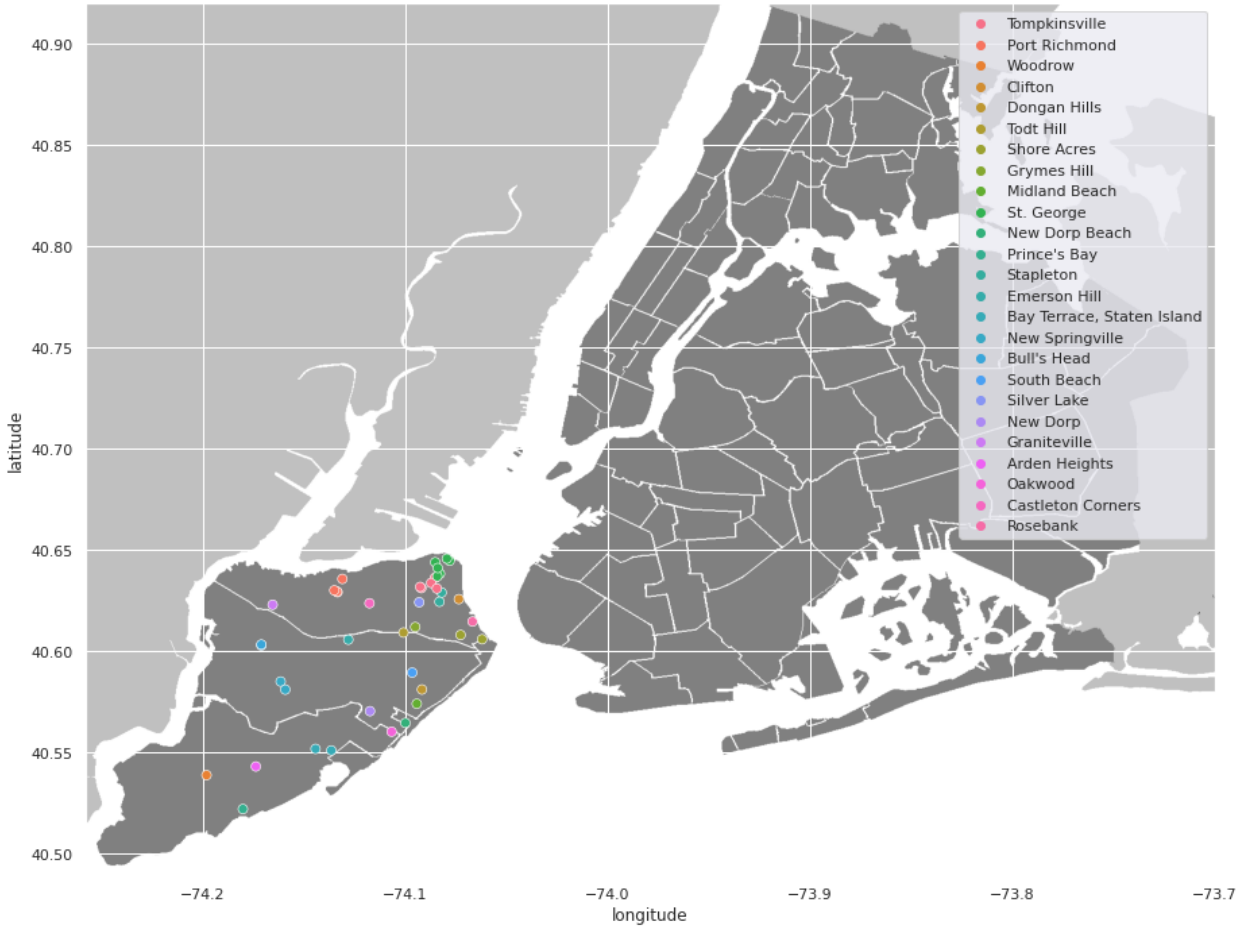
اما روابط دیگری که قابل بررسی است، تاثیر عکس review_per_month و minimum_night است که هر چه شب های کمتری قابل دسترسی باشد ، review های بیشتری انجام خواهد شد.

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.

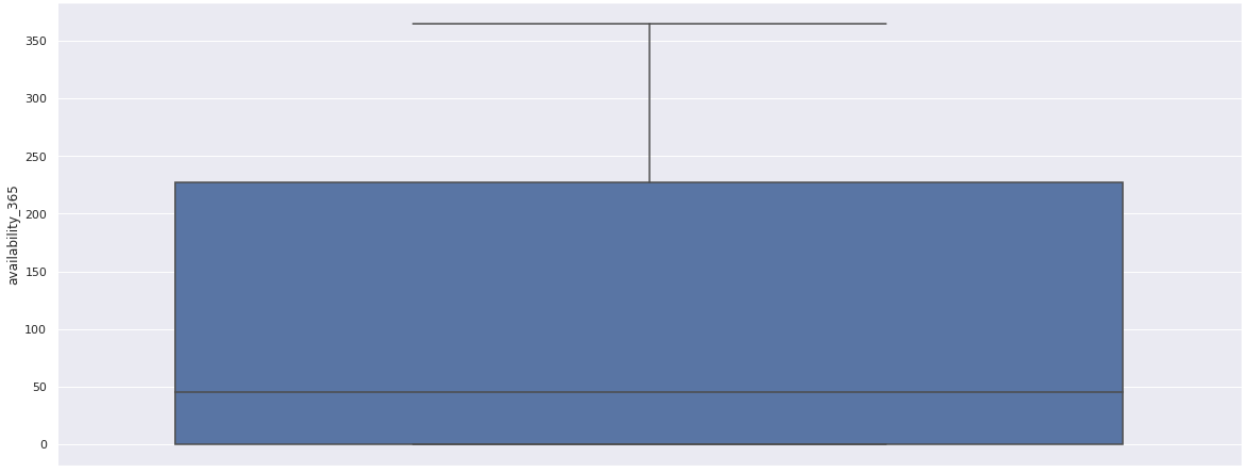
در ارتباط با همین موضوع `pearsonr` و `spearmanr` را برای 100 نمونه انجام میدهیم که معمولا p-value کوچکی میداد و نمی توان فرض را به طور کامل پذیرفت.

در بخش آخر میخواهیم بیشترین ترافیک را پیدا کنیم. 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.

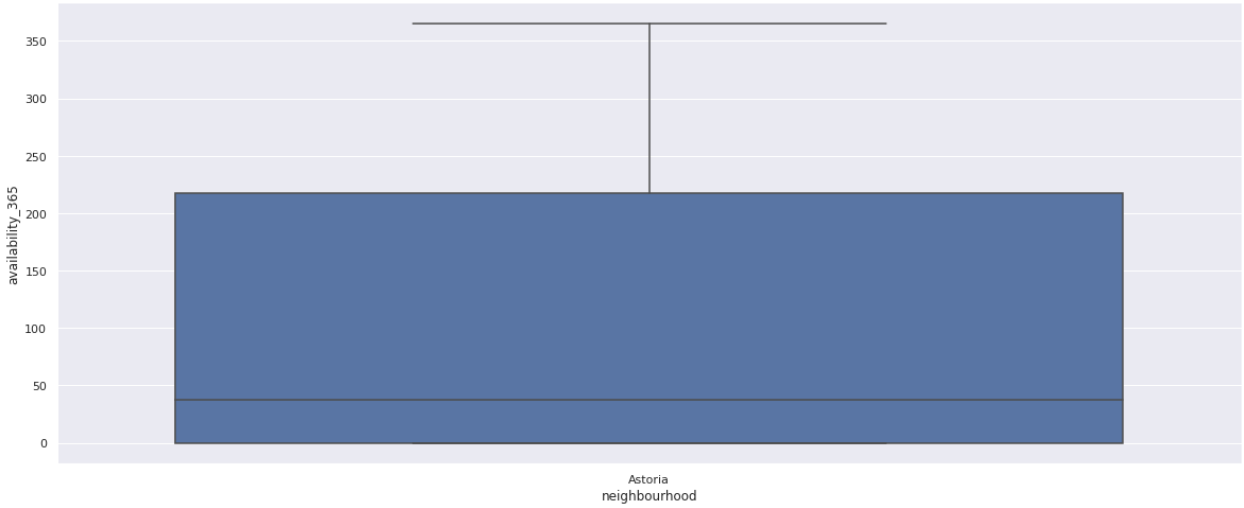
اما برحسب availability ، من همه محله هایی که خانه ای با availability=0 داشتند را در نقشه نمایش دادم و سپس با تعداد ک خانه در آن محله یا neighbourhood بررسی کردم که هرکدام تغییر قابل توجهی داشتند ، بررسی آماری و تست روی آنها انجام دهم. همه نقشه ها و تعداد خانه ها در کد موجود است ولی برای نمونه نقشه خانه 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 :

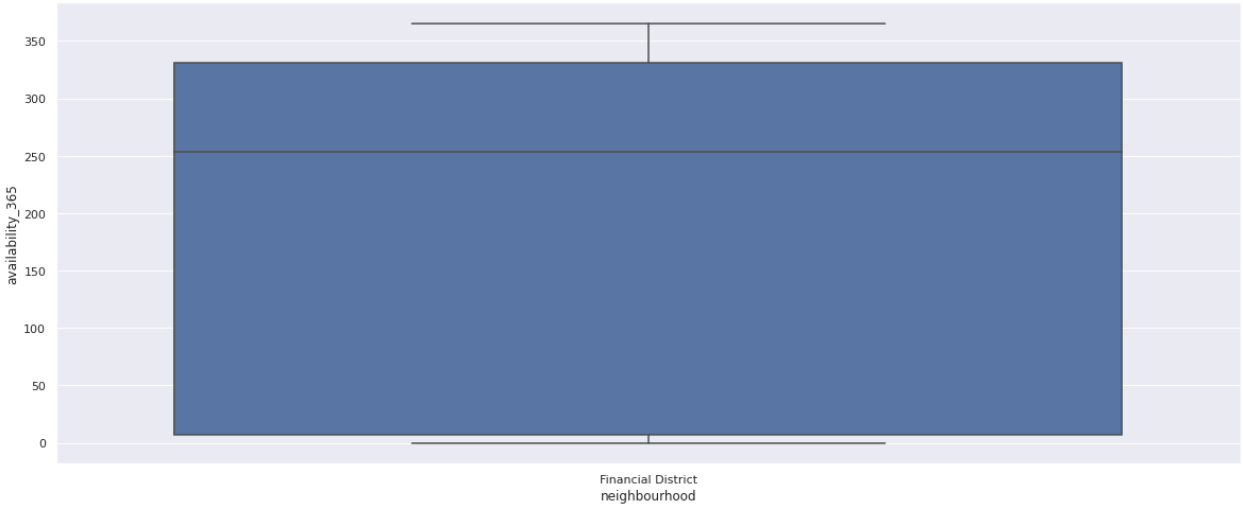


اولین شهری که بررسی می کنیم Astoria ست که به خاطر بررسی های روی جدول به نظر باید avalibility پایینی داشته باشد اما boxplot آن به شکل زیر است:

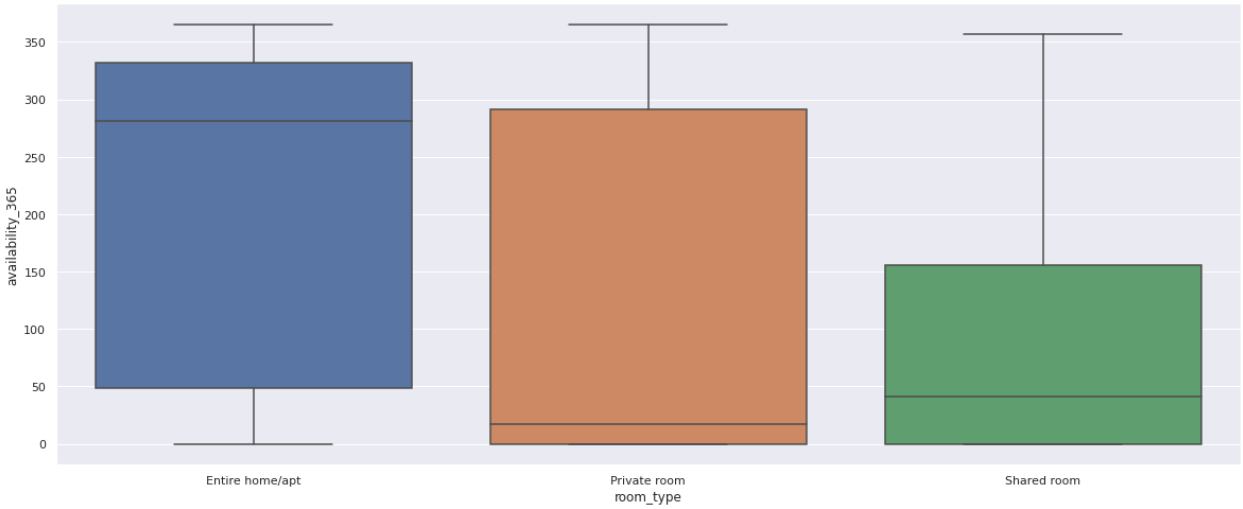


که به نظر خیلی تفاوت فاحشی با داده اصلی ندارد و و وقتی با داده اصلی با استفاده از t-test مقایسه میشود، به علت p-value بالا فرض اولیه ما رد و astoria شهر با ترافیک بالا محسوب نمیشود.

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:



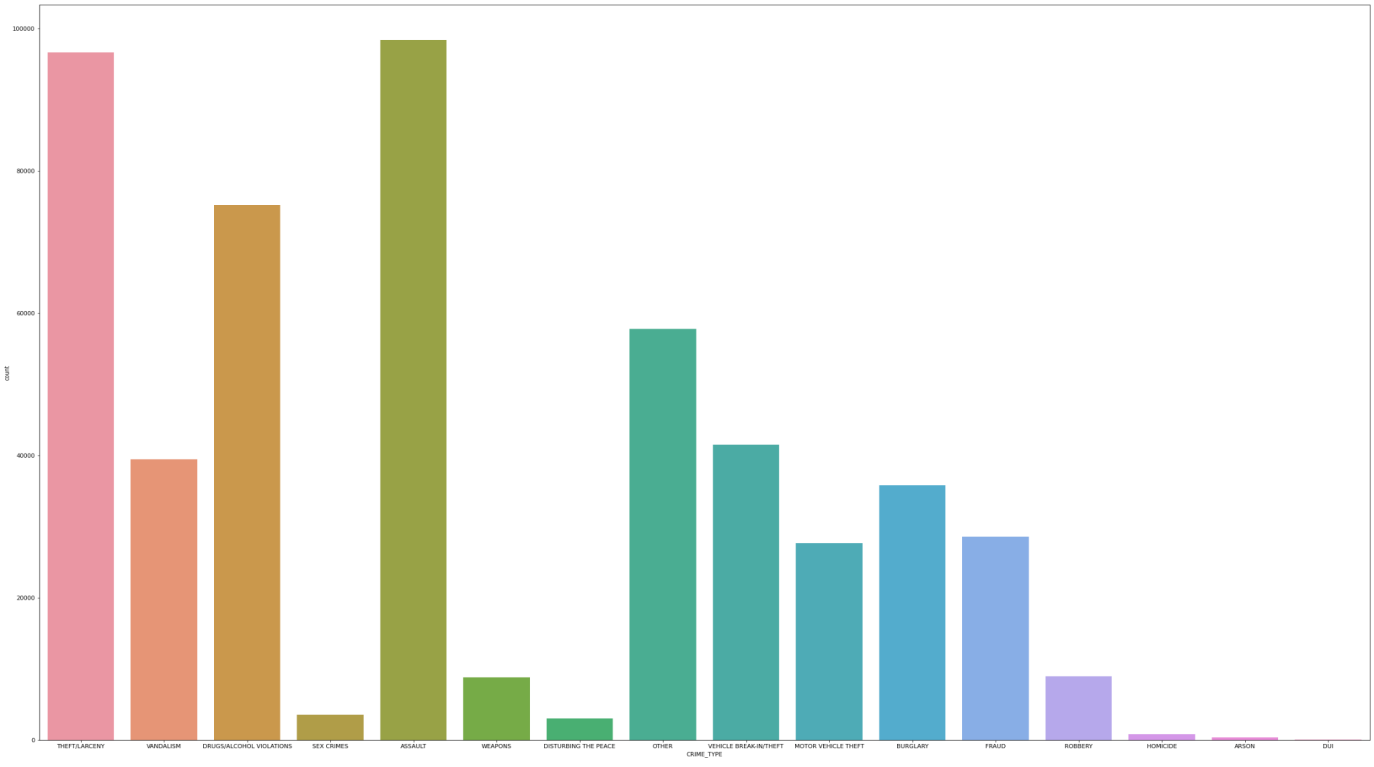
این بار احتمالا باید T-test ما pvalue پایینی داشته باشد که همینطور هم هست و اولین رقم اعشار آن در حدود ده یا بیست قابل دیدن است پس این فرض قبول است و این شهر از ترافیک زیادی برخوردار نیست و اکثر مواقع available می باشد.
حدسی که میتوانم برای کم بودن ترافیک در آن بزنم آن است که private room ها در آن خیلی 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 مین چند نمونه neighbourhood بسنده می کنم.

2. مجموعه داده های گزارش های جنائی

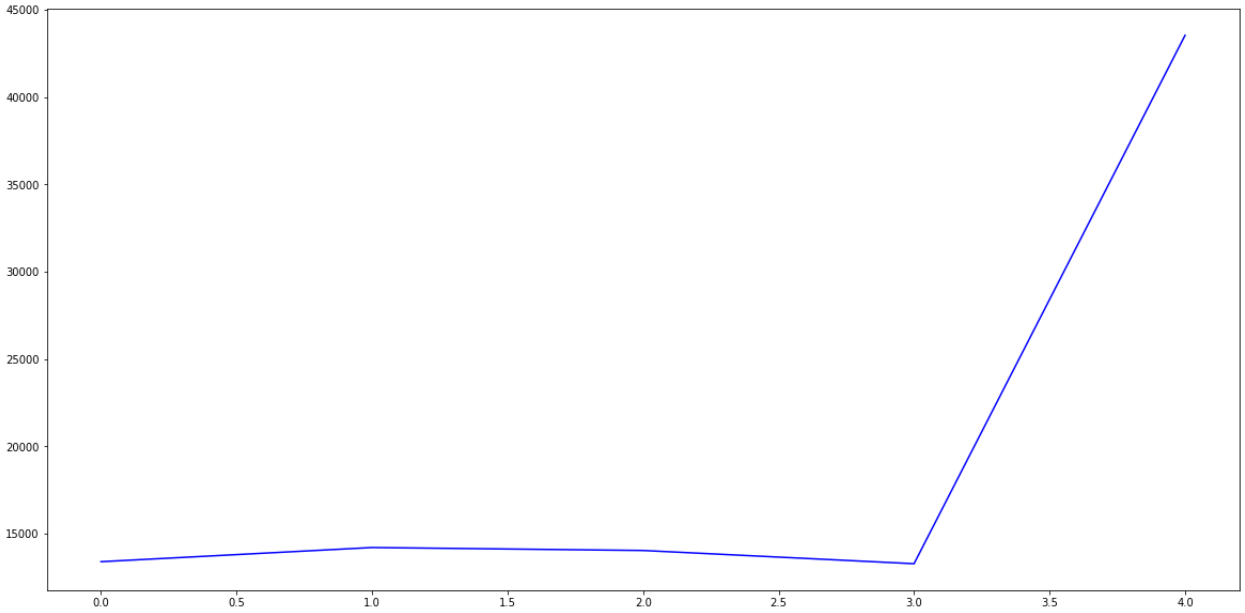
در این دیتاست به دنبال بررسی گزارش های مطرح شده از یک زندان در شهر Louisville در ایالت کنتاکی هستیم که پس از خواندن داده های 5سال اخیر که از سال 2015 تا 2019 است و concat کردن آنها به یکدیگر دیتا فریمی ساختیم که بیشترین جرمی که انجام شده است را مشاهده کنیم که Assault بوده و نمودار فراوانی حمله ها به شکل زیر است. (اگر وضوح تصویر کم است، در کد نیز موجود است.)



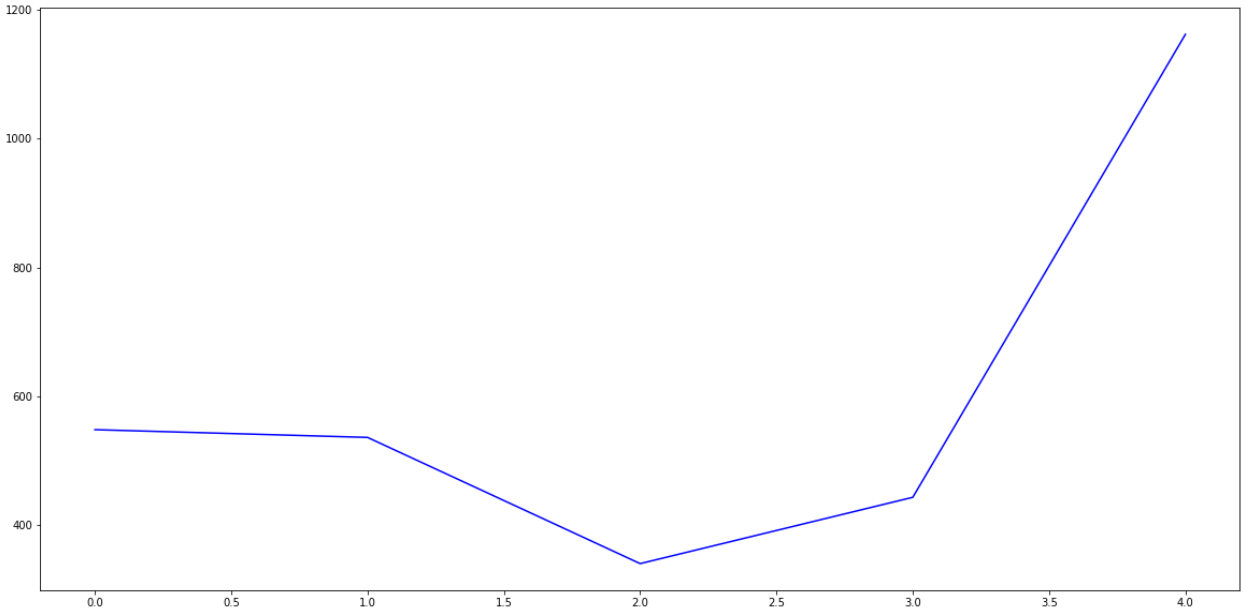
برای به دست آوردن zip-code ای که بیشتر حوادث در آنجا رخ می دهد نیز ، کافی است describe بزنیم که top را به ما میدهد که zip_code = 40214 است که 23656 جرم در آن رخ داده است.

برای آنکه محاسبه کنیم ، چه جرم هایی کاهش یا افزایش داشته است ، ذکر این نکته ضروری است که با توجه به اینکه داده ی سال آخر بسیار کامل تر و بیشتر از سال اول است ، همه جرم ها افزایش داشته است. زیرا داده های سال آخر از 2019 شروع میشود تا سال 2021. اما میتوان فهمید بعضی از جرم ها افزایش چشمگیری داشته و بعضی از جرم ها افزایش بسیار کمتر.

برای مثال این گراف رخداد assault در 5سال اخیر است که در 4سال ابتدایی ، تعداد بسیار نزدیکی دارد اما در سال آخر به وضوح ، افزایش چشمگیری داشته است.

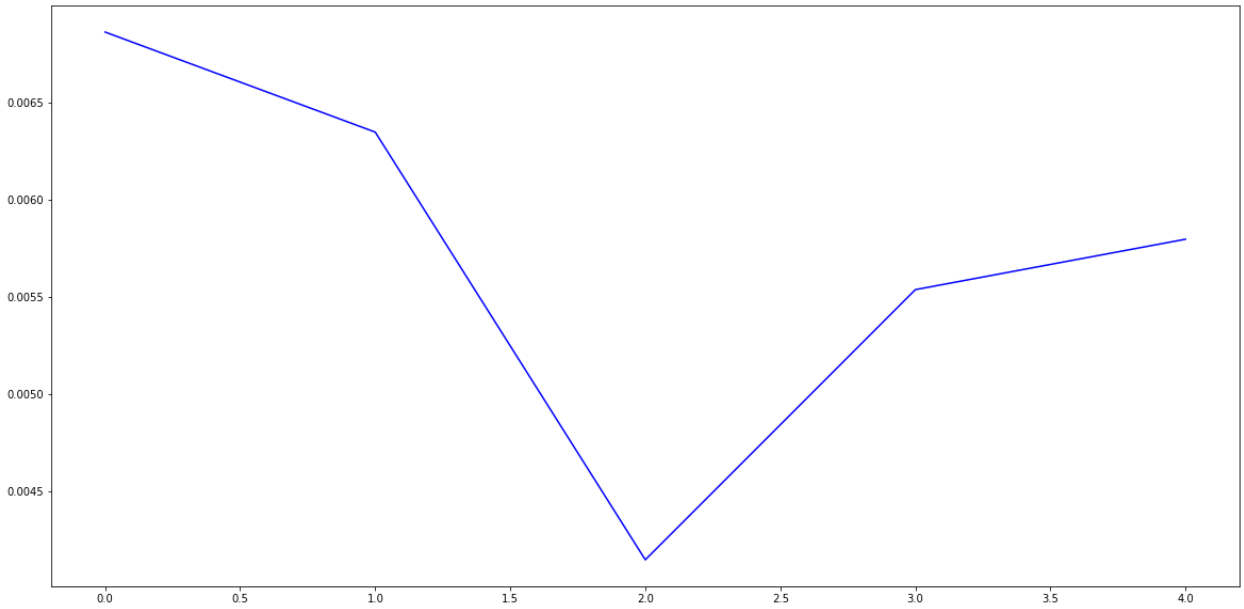


نمودار زیر نیز ، نمودار DISTURBING THE PEACE است که در دو سال کاهش داشت اما پس از آن شکل صعودی به خود گرفته است.



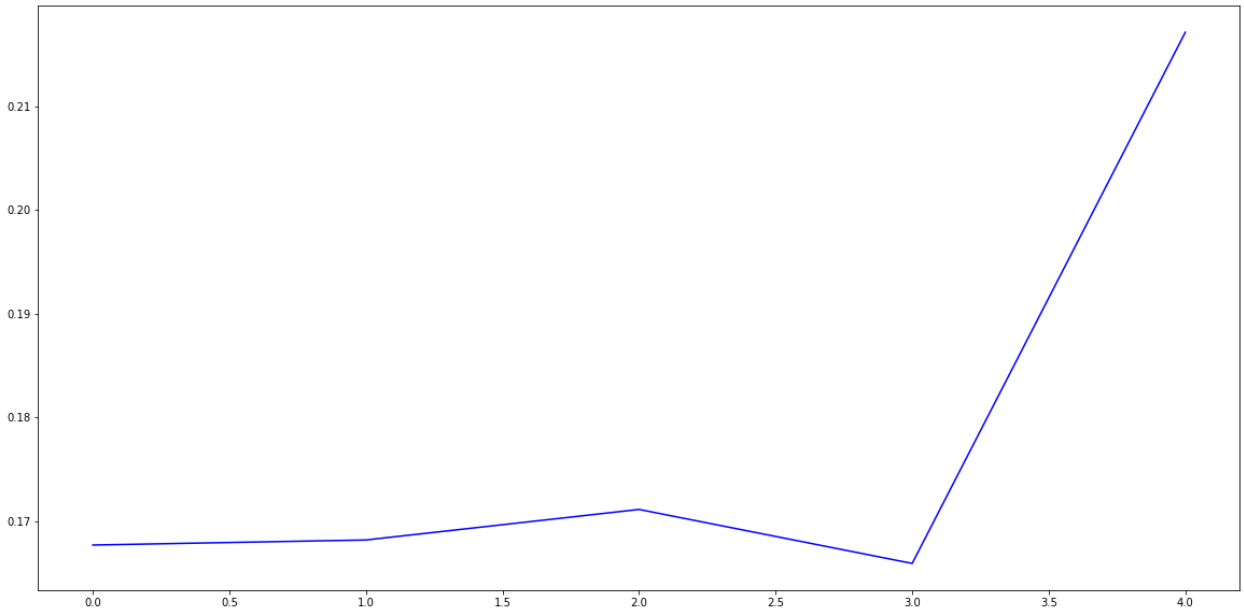
اما همانطور که گفتم ، همه ی این جرم ها در آخر صعودی خواهند شد. برای اینکه نسبت افزایش جرم را در طی سال ها بفهمیم، بررسی میکنیم به نسبت کل جرم های انجام شده این جرم چند بار تکرار شده است.

ابتدا از همان DISTURBING THE PEACE شروع میکنیم که همانطور که در نمودار پایین ملاحظه میکنید ، با اینکه پس از سال سوم دوباره، حالت صعودی می گیرد اما باز هم به سال اول نمی رسد.

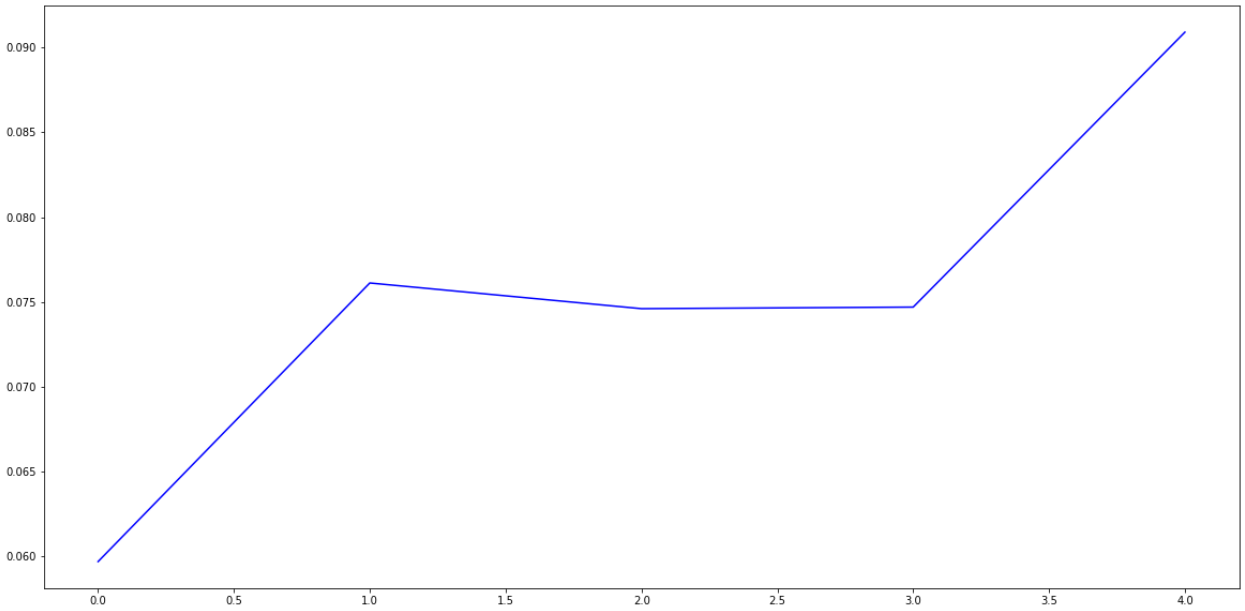


حال با این روش ، روی همه جرم ها بررسی می کنیم کدام یک در طی این سالیان همواره صعودی بود و کدام یک همواره نزولی.

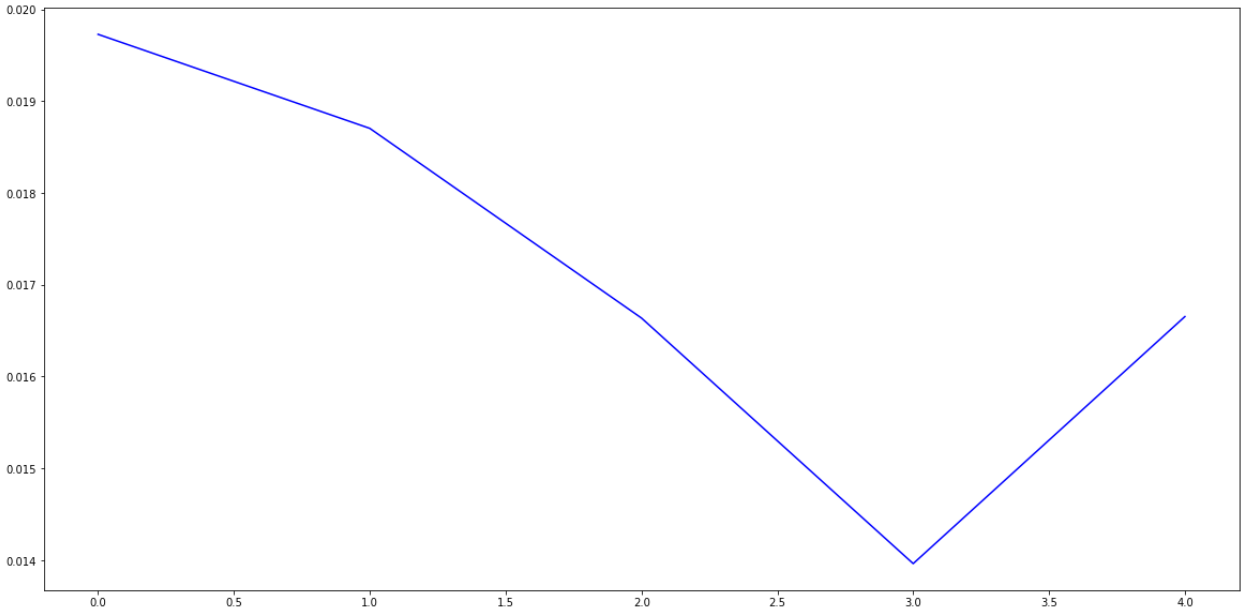
در زیر نمودار assault را مشاهده میکنیم که به جز سال 2018 در بقیه موارد روند صعودی را طی می کند و میتوان گفت این جرم در حال افزایش سالیانه است.



از الگوی نمودار زیر که مربوط به VEHICLE BREAK-IN/THEFT است نیز برداشت میشود که ، سال 2016 نسبت به سال قبلی افزایش این جرم را داشتیم اما پس از آن 2سال تقریبا به هم اندازه این جرم تکرار شد تا اینکه در سال 2019 باز هم افزایش چشمگیری در انجام این جرم به وجود آمده است.

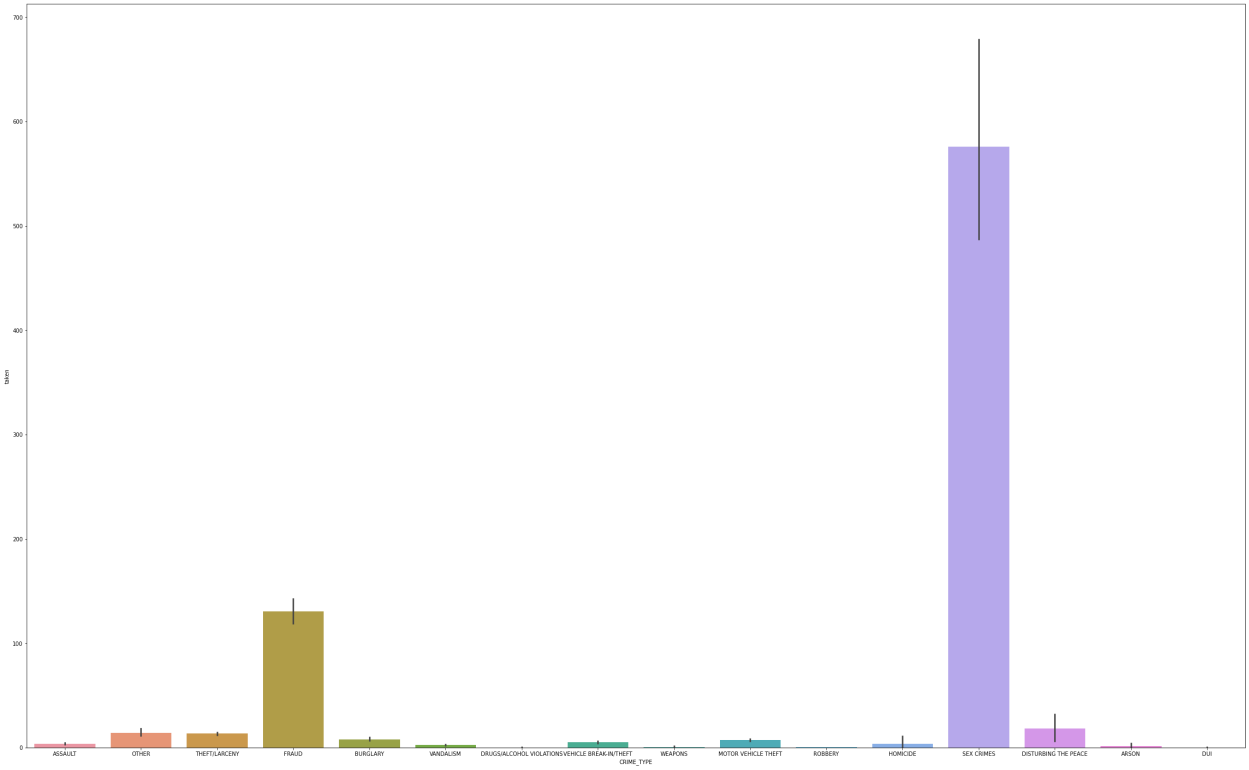


در آخر هم نمودار Robbery در طی این 5 سال را مشاهده میکنیم که طی 4 سال ابتدایی تنها کاهش داشته اما سال آخر به میزان اندکی بیشتر شده است که قابل توجه است.



در بخش بعدی بررسی می کنیم ، چه جرمی دیرتر از همه گزارش شده است. برای این کار باید مدت زمان فاصله بین عمل انجام شده و زمان گزارش شده را بررسی کنیم.

در این بخش من روی دیتاست سال 2019 کار کردم تا جواب را بررسی کنم. پس از ساختن ستونی جدید و قرار دادن تفاضل دو ستون دیگر روی آن ، نمودار زمان گزارش کردن را مشاهده میکنیم که به صورت زیر است:



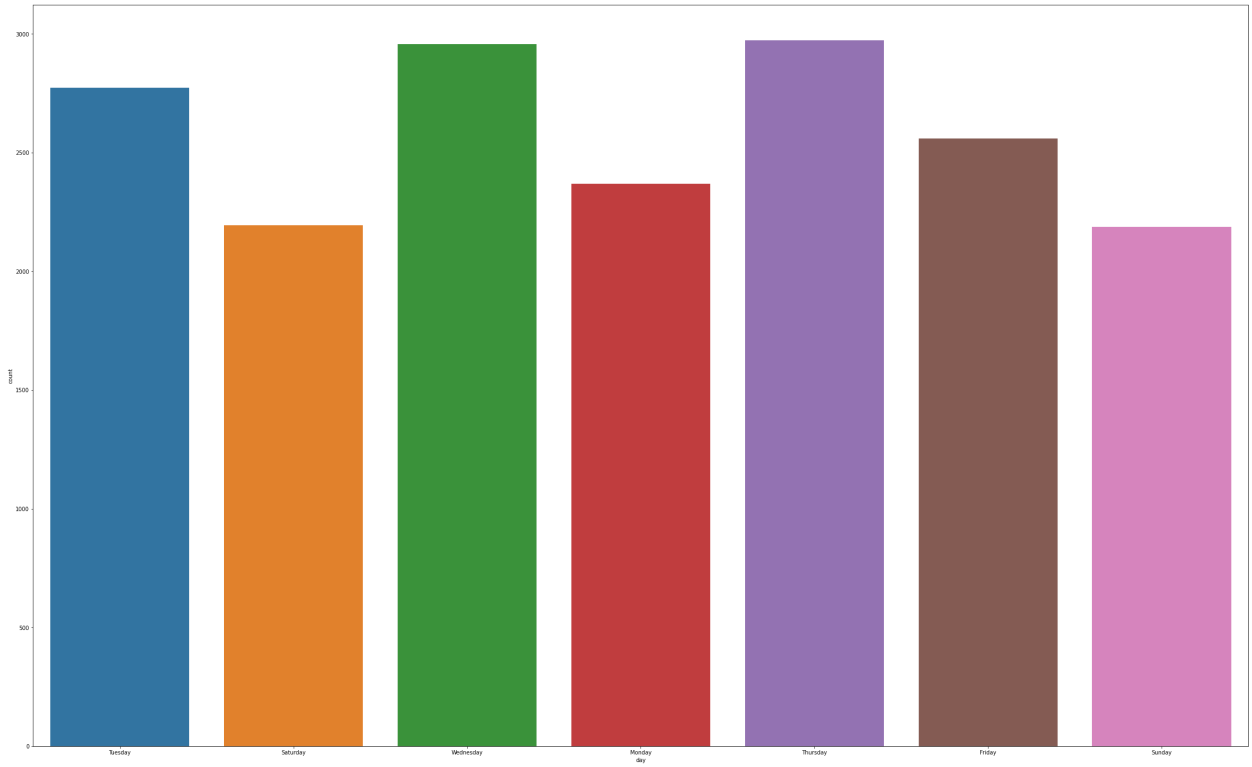
مشخصا بیش از همه Sex Crimes زمان برده است که دلیل آن می تواند خجالت یا ترس و اضطراب از قضاوت جامعه باشد. پس از آن نیز Fraud است که شاید فردی که مورد ضرر واقع شده ، دیرتر متوجه این اقدام شده است.

مواردی نیز که به مدت کمتری گزارش را ثبت کرده اند مربوط به دزدی یا خشونت ناشی از الکل و مواد مخدر بوده است.

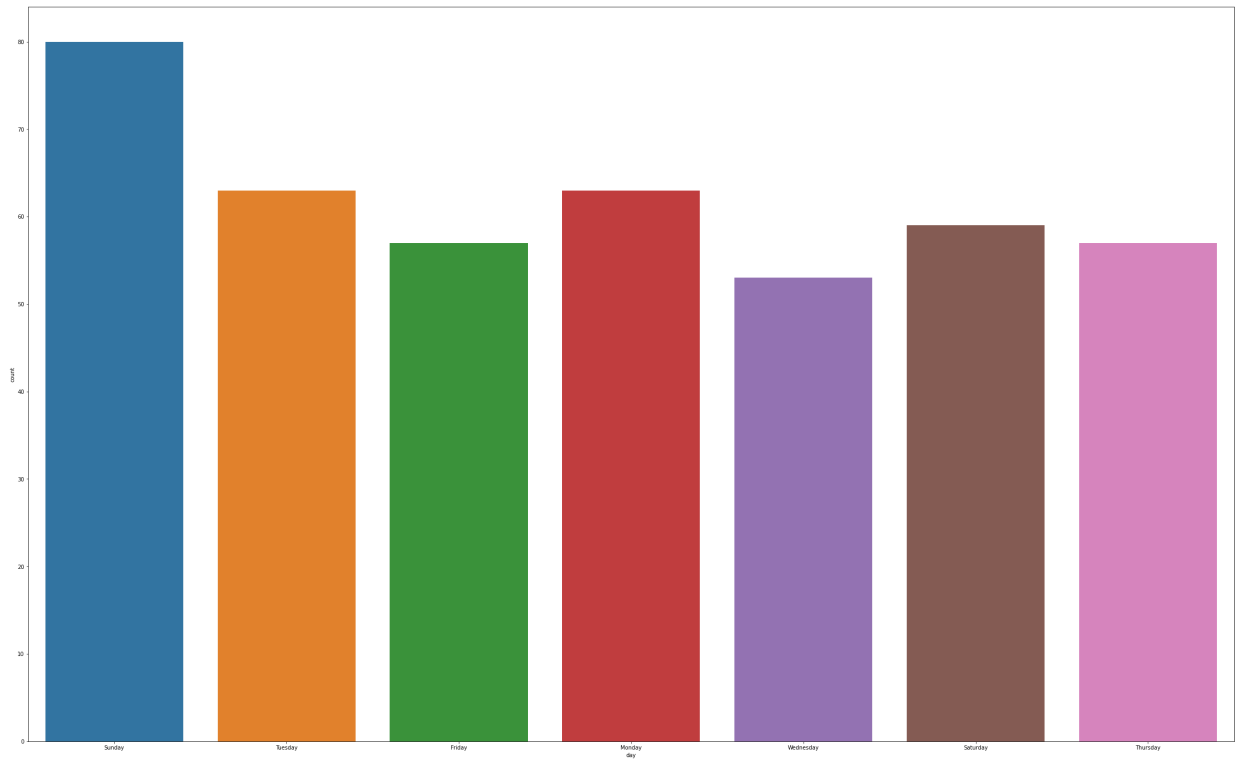
پس از آن ستون جدیدی تحت عنوان day میسازیم که روی آن روزی که این جرم اتفاق افتاده است را ، نشان می دهد که نمودار جرم های صورت گرفته بر اساس روز های هفته به صورت زیر است:

همانطور که مشخص است، تفاوت زیادی بین روزهای هفته ، وجود ندارد اما شاید در جرم های خاص، تفاوت بین روزهای هفته وجود داشته باشد که آن را هم بررسی می کنیم.

اول از همه خشونت ناشی از مواد مخدر و الکل را بررسی می کنیم که به صورت زیر است:

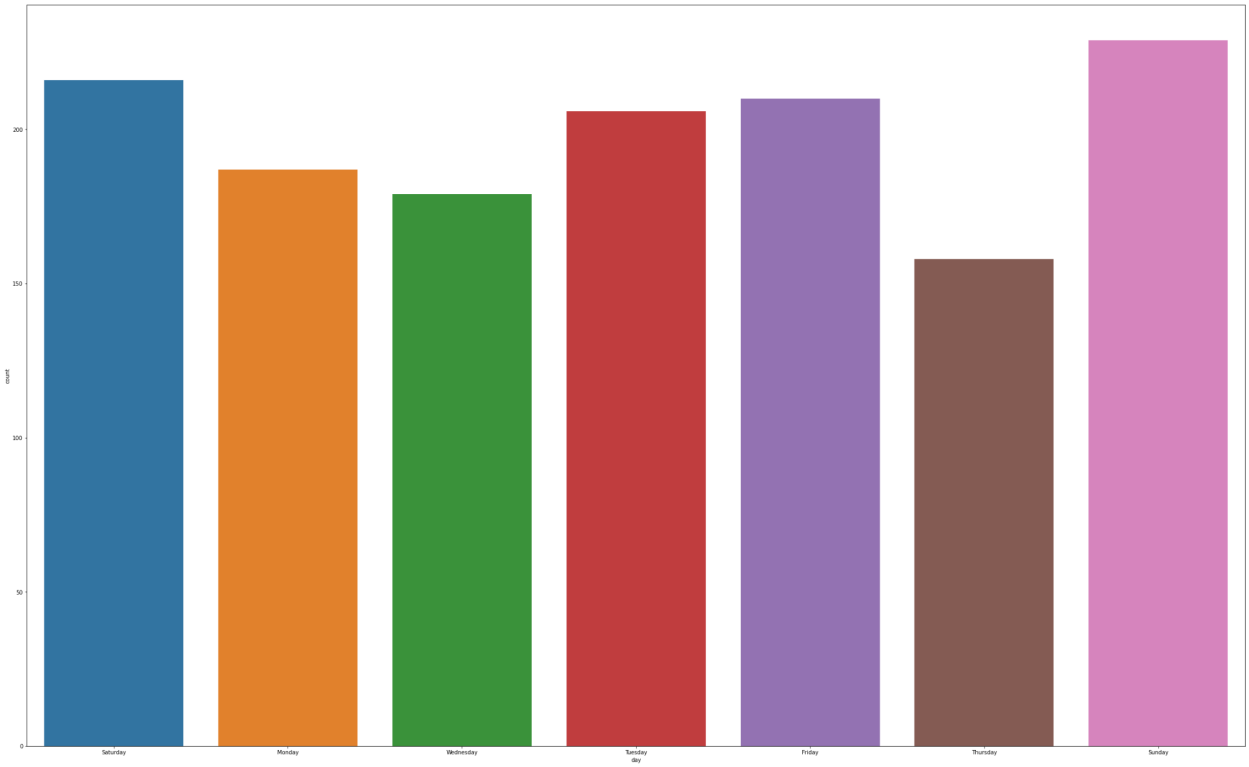


بیشتر این اتفاقات در روزهای وسط هفته یعنی دوشنبه و پنج شنبه رخ داده است و روز آخر هفته کمترین میزان را داراست.
حال برای Homicide نیز نمودار بر اساس روزهای هفته خواهیم داشت:

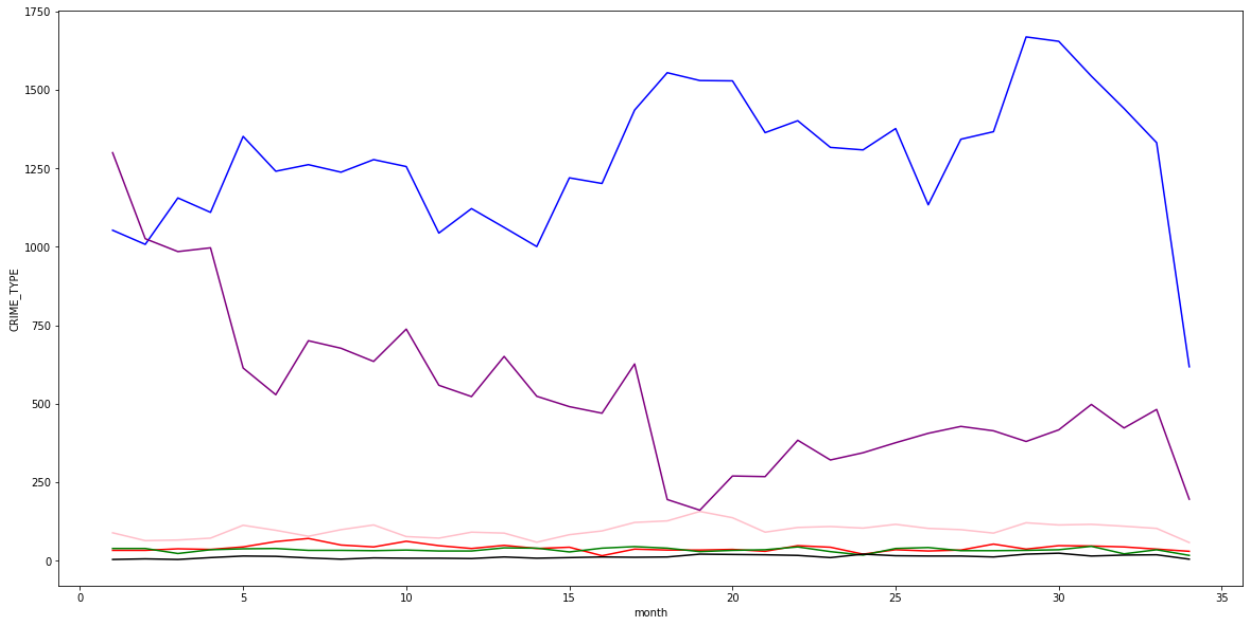


مشخص است در روز یکشنبه بیشترین میزان قتل را داشتیم که داده ای مثمر ثمر میباشد.

آخرین نموداری که بررسی می شود نیز مربوط به Sex crime است که بیشترین روزهایی که این اتفاق می افتد روزهای شنبه و یکشنبه بوده است.

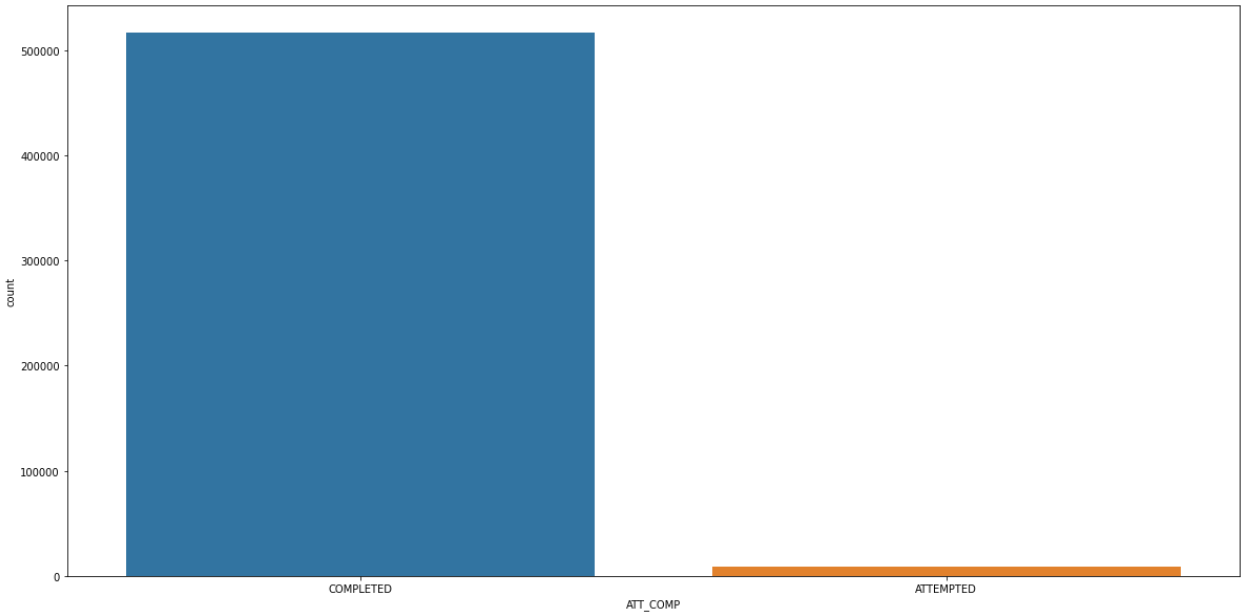


سپس ستونی تحت عنوان Month برای دیتا فریم 2019 به بعد در نظر گرفتیم و به ترتیب ماه ها ، روند نزولی یا صعودی بودن را برای 6 جرم مختلف نمایش دادیم که به شکل زیر است :

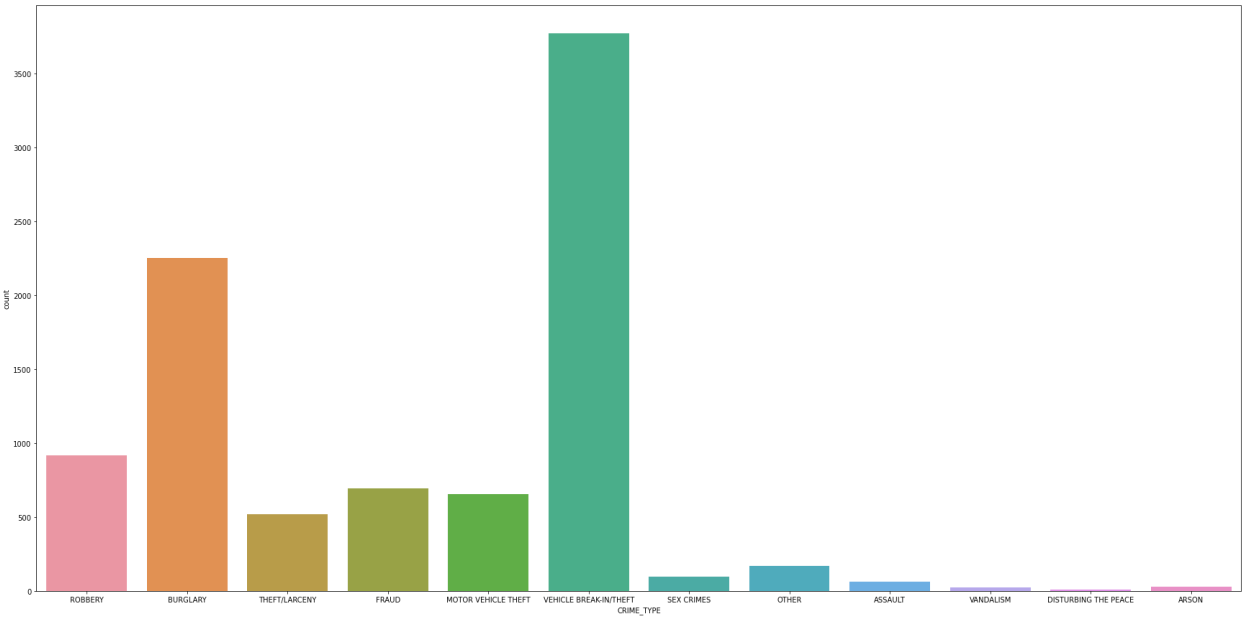


به طور کلی robbery , homicide , disturbing peace , Sex crimes روندی تقریبا مشابه داشتند و افزایش یا کاهش آنچنانی طی این 35 ماه نداشتند . رنگ آبی نیز که مربوط به Assault است خیلی روند مشخصی ندارد و ماه به ماه مقدار مختلفی دارد و نمی توان تعبیر خاصی از آن داشت ، ولی در آخر Drugs/alcohol violations به میزان قابل توجهی در ماه های آخر در حال کاهش است و نمودارش روند نزولی دارد.

مورد آخری که بررسی کردیم ستون ، Att-Comp است که مشخص میکند جرم به صورت کامل عملی شده یا تلاش برای انجام جرم انجام شده و به طور کامل انجام نشده است. نمودار زیر نشان دهنده تعداد جرم های کاملا انجام شده و نشده است.



مشخص است بیشتر جرم ها کاملا انجام شده ، اما جالب است که چه جرم هایی بیشتر ناقص یا در همان حد ATTEMPTED مانده اند. در نمودار زیر تعداد جرم های ATTEMPTED بر اساس جرم انجام شده آمده است :



بیشترین جرم هایی که ناقص مانده اند مربوط به VEHICLE BREAK-IN/THEFT , Burglary می باشد که با توجه به این که این جرم ها ، جرم هایی نیستند که خیلی زیاد انجام شده باشند و جزو جرم هایی با تراکم زیاد نبودند ، این مسئله می تواند اطلاعات خوبی باشد که ما از داده بیرون کشیده ایم.