به نام خدا

مهرانه مقتدائی فر ۹۷۲۲۲۰۸۶

تاریخ: ۲ / ۸ / ۱۴۰۰

گزارش تمرین سری اول

درس: مبانی علوم داده

استاد درس: دکتر خردپیشه

---

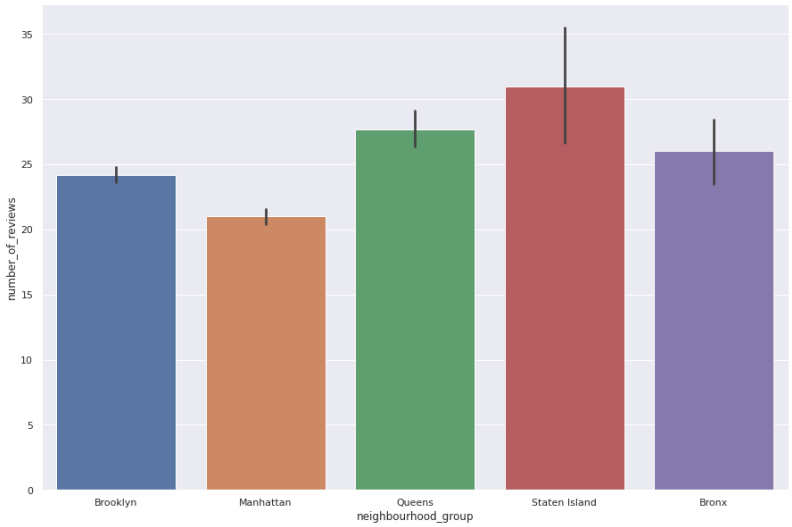
۱. مجموعه داده های Airbnb

بررسی تعداد بازدید ها

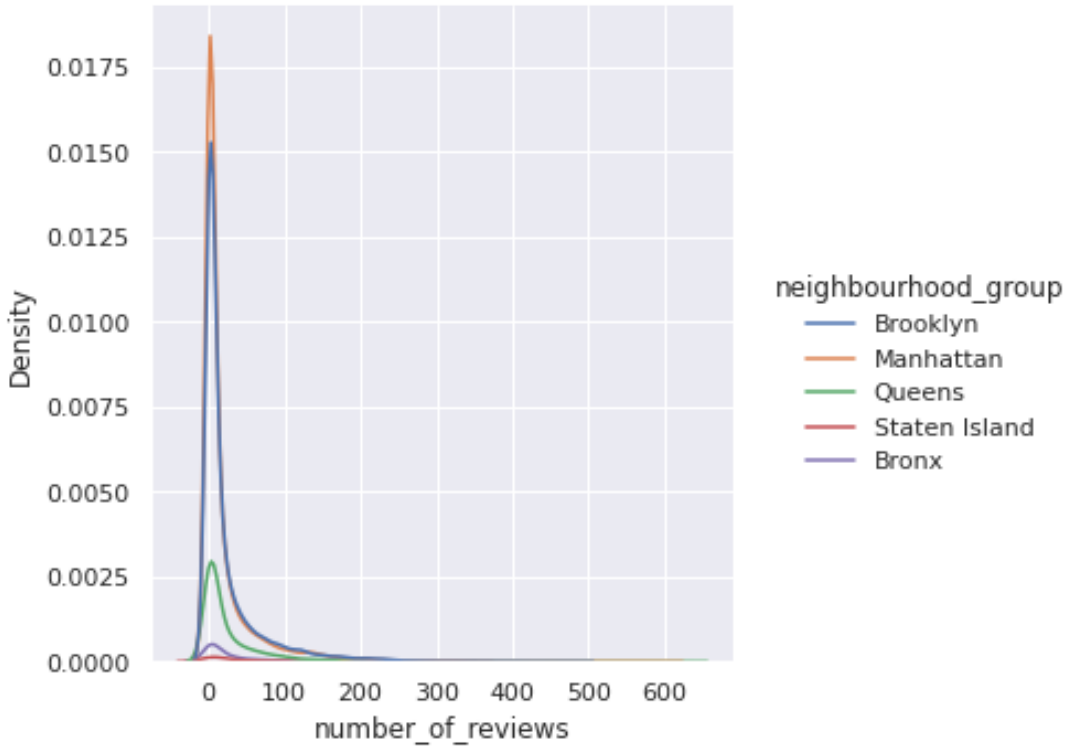
برای بررسی مجموعه داده های این دیتاست، ویژگی های ستون های مختلف و ارتباط آنها با یکدیگر را بررسی میکنیم. در ابتدا برای آنکه اطلاعاتی در مورد مناطق مختلف کسب کنیم، ستون neighbourhood_group را که شامل ۵ منطقه منحصر بفرد است بر اساس تعداد بازدید ها و همچنین قیمت آنها بررسی میکنیم. اگر این مناطق را بر اساس 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 بیان گر همین مسئله می‌باشد:

	number_of_reviews							
	count	mean	std	min	25%	50%	75%	max
neighbourhood_group								
Bronx	1091.0	26.004583	42.214774	0.0	1.0	9.0	32.0	321.0
Brooklyn	20104.0	24.202845	44.344868	0.0	1.0	6.0	25.0	488.0
Manhattan	21661.0	20.985596	42.572277	0.0	1.0	4.0	19.0	607.0
Queens	5666.0	27.700318	51.955853	0.0	1.0	7.0	32.0	629.0
Staten Island	373.0	30.941019	44.830766	0.0	1.0	12.0	42.0	333.0

همچنین در نمودار زیر نیز این مسئله کاملا مشخص است:



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



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 H0 ما این بود که این نواحی بایکدیگر از نظر تعداد بازدید تقریبا یکسان هستند، پس از بررسی های ابتدایی فرض یک H1 این بود که خیر، بازدید های ناحیه State Island از بقیه بیشتر است و در کل تعداد بازدید ها برابر نیست(مخالف H0) 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 H0 رد میشود، این همان نتیجه ای است که ما از اول هم انتظار داشتیم و با استفاده از تست آماری نیس 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	17.41	27.600357	0.0	1.00	9.5	20.5	192.0
Brooklyn	100.0	20.52	32.879480	0.0	1.00	4.0	24.0	151.0
Manhattan	100.0	24.33	47.173451	0.0	1.00	6.0	23.5	323.0
Queens	100.0	22.71	35.436184	0.0	1.00	8.5	31.0	183.0
Staten Island	100.0	39.68	60.503757	0.0	1.75	12.0	52.0	333.0

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 H0 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.

استفاده از کتابخانه researchpy نیز اطلاعات آماری جالبی را به ما میدهد.

```
(           Variable      N    Mean         SD        SE  95% Conf.   Interval
 0  number_of_reviews  100.0  38.530  50.268663  5.026866  28.555607  48.504393
 1  number_of_reviews  100.0  23.100  40.574784  4.057478  15.049082  31.150918
 2           combined  200.0  30.815  46.216454  3.267997  24.370652  37.259348,
                                 Independent t-test   results
 0  Difference (number_of_reviews - number_of_revi...   15.4300
 1                              Degrees of freedom =   198.0000
 2                                               t =     2.3885
 3                           Two side test p value =     0.0179
 4                          Difference < 0 p value =     0.9911
 5                          Difference > 0 p value =     0.0089
 6                                      Cohen's d =     0.3378
 7                                      Hedge's g =     0.3365
 8                                  Glass's delta =     0.3070
 9                                    Pearson's r =     0.1674)
```

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 برروی قیمت تاثیر دارد یا نه؟

سه نوع room type مختلف مشاهده میکنیم، که تعداد هرکدام و میانگین قیمت آنها مشخص است:

	price							
	count	mean	std	min	25%	50%	75%	max
room_type								
Entire home/apt	25409.0	211.794246	284.041611	0.0	120.0	160.0	229.0	10000.0
Private room	22326.0	89.780973	160.205262	0.0	50.0	70.0	95.0	10000.0
Shared room	1160.0	70.127586	101.725252	0.0	33.0	45.0	75.0	1800.0

مانند مرحله قبل، برروی این سه نوع نیز یک بار تست ANOVA و یک بار نیز بین دو نوع از خانه ها t-test انجام دادیم، با 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.

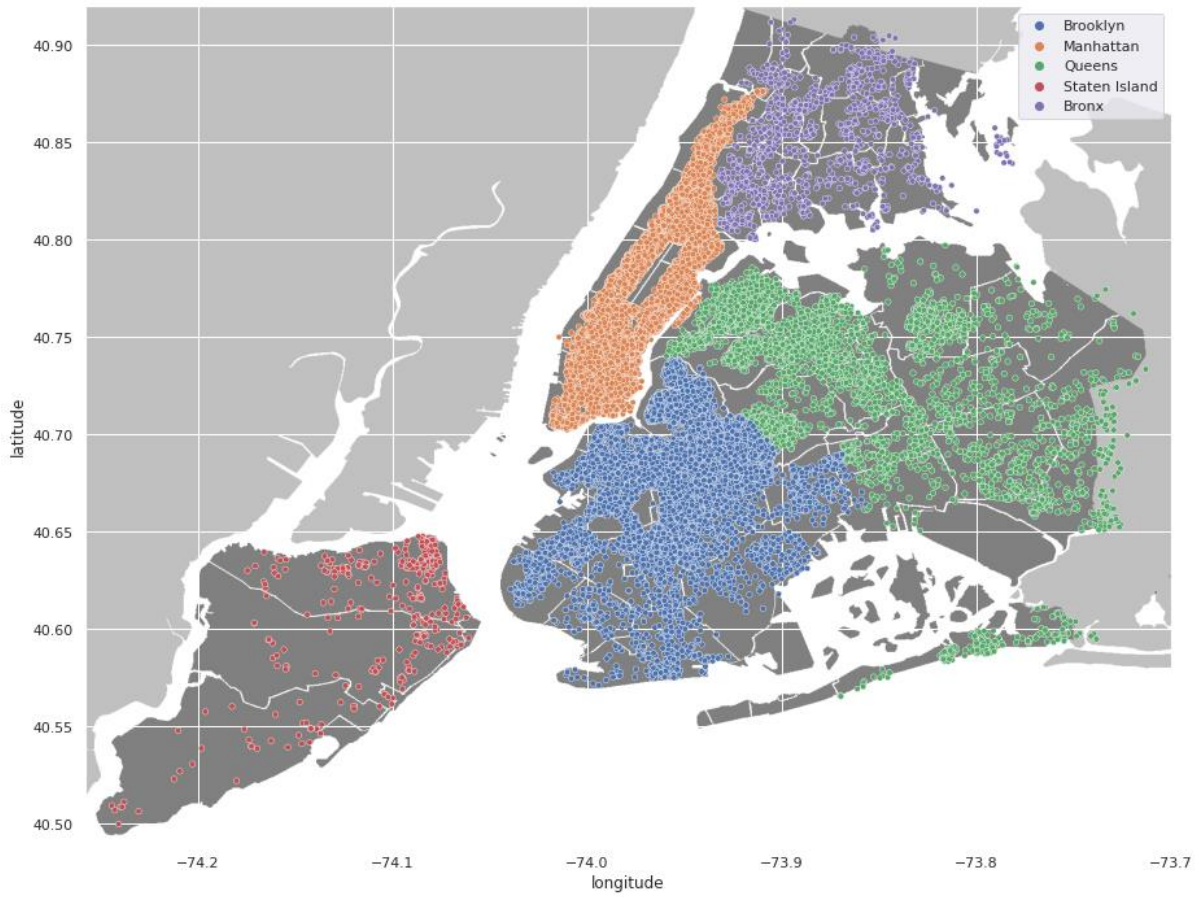
حداقل تعداد شب ماندن و قیمت

سپس به بررسی دوتا داده عددی پرداختیم. آیا تعداد حداقل شب ها بر روی قیمت خانه ها تاثیر گذار است یا خیر؟

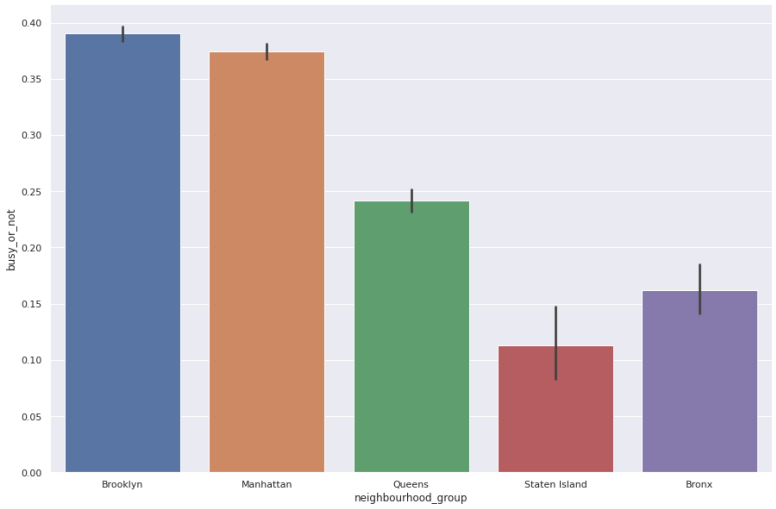
فرض H0 این است که این دو مقدار به هم مربوط نیستند و قیمت برای تمامی یکسان است. این دو متغیر هردو عددی هستند و برخلاق تست های قبلی که معمولا یکی عددی و دیگری categorical بود نیست. به همین علت تست های آماری مناسب برای این داده ها استفاده از correlation ها هست. هر دو تست pearson و spearman را تست کردیم 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 H0 ما رد میشود و مورد قبول نیست و در نتیجه میتوانیم بگوییم که تعداد minimum night ها برروی 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:

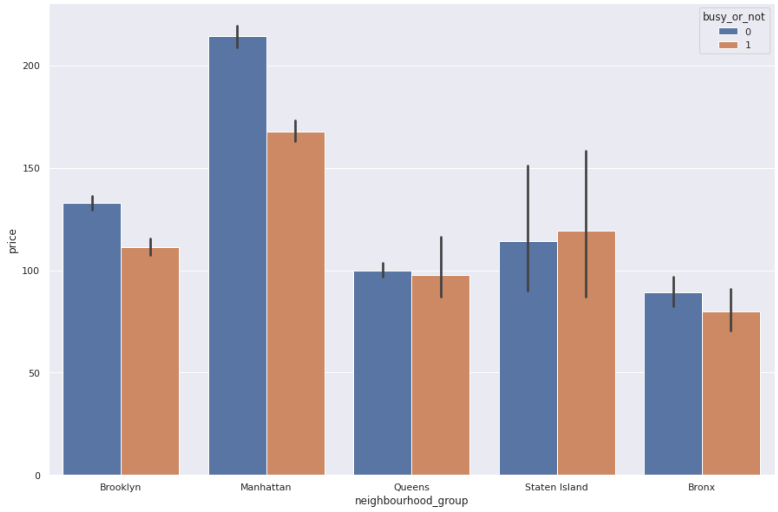


همانطور که در نقشه نیز مشخص است، تراکم برای نقاط نارنجی و آبی بیشتر از بقیه هست یعنی دو شهر Manhattan و Brooklyn بیشترین خانه ها را دارند، همچنین این وضعیت را نیز میتوان از روی availibility نیز مشخص نمود. برای آنکه availibility را بهتر بتوانیم مشاهده کنیم، اینگونه بررسی کردیم که هرچی از ۳۶۵ روز سال تعداد کمتری availibility داشته باشیم به این معناست که قطعا آنجا شلوغ تر بوده و بیشتر رزرو شده است. پس میتوانیم آن host هایی که تعداد availibility 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.



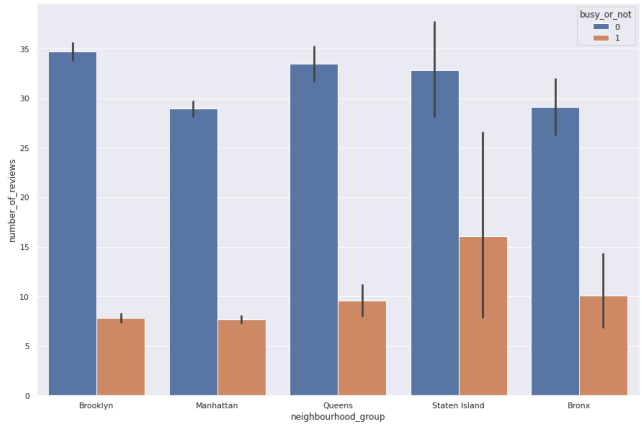
رابطه شلوغ بودن با قیمت و تعداد بازدیدها

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

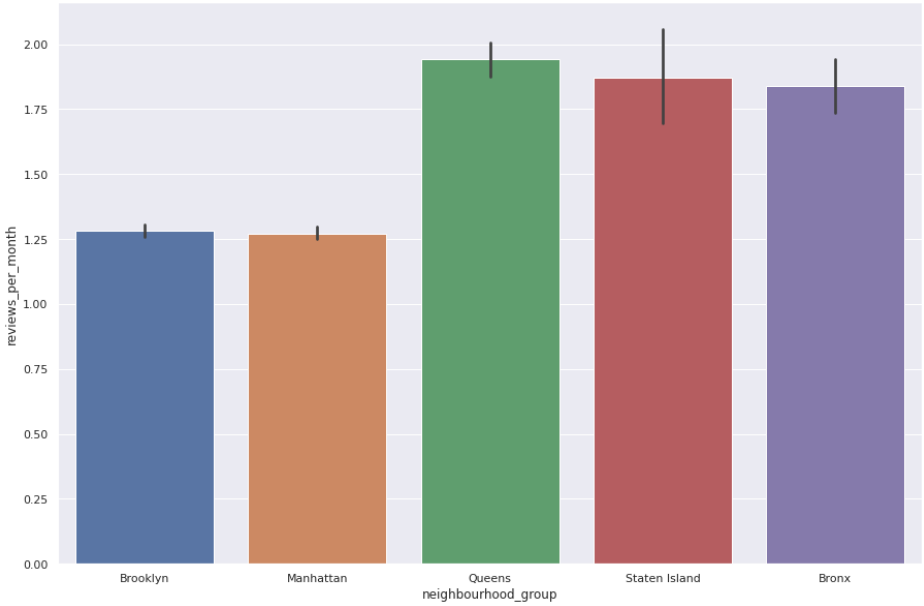
با توجه به t test انجام شده بین محله هایی که availibility آنها برابر با صفر است و قیمت خانه ها، مشاهده میکنیم که p value مقداری کمتر از 0.05 دارد و این به این معنیست که فرض H1 ما مورد قبول است و یعنی محله‌ای که ترافیک بالاتری دارد، قیمت بیشتری نیز دارد. این مسئله در نمودار بالا نیز قابل مشاهده بود.

سپس با تست بعدی نیز بررسی کردیم که تعداد بازدید ها چطور؟ 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 بازدید های 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 queens 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

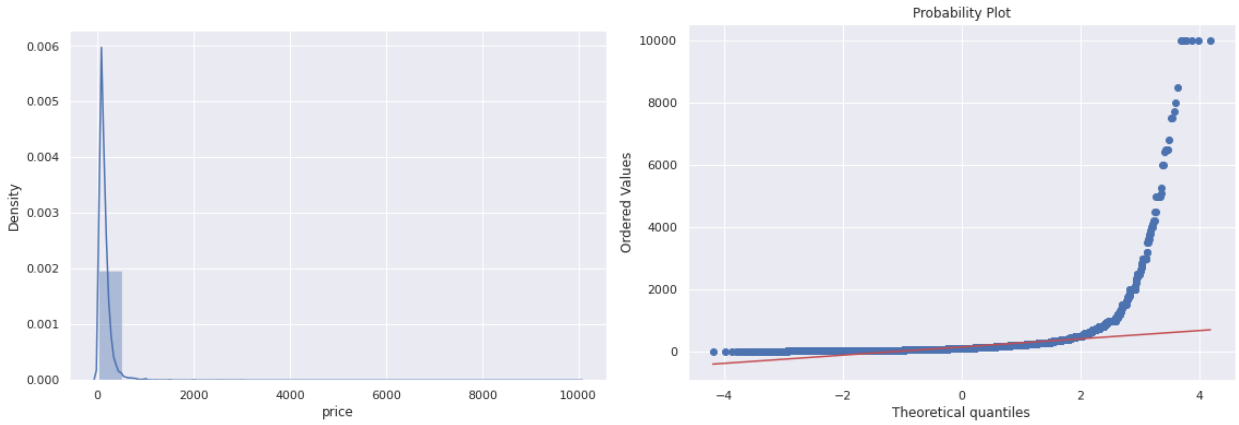
Bronx → Melrose

نرمال کردن توزیع price

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

منبع : https://www.analyticsvidhya.com/blog/2021/05/how-to-transform-features-into-normal-gaussian-distribution/

در ابتدا خانه هایی راکه price = 0 داشتند ، از dataset حذف کردیم. سپس دو نمودار زیر راکه 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.

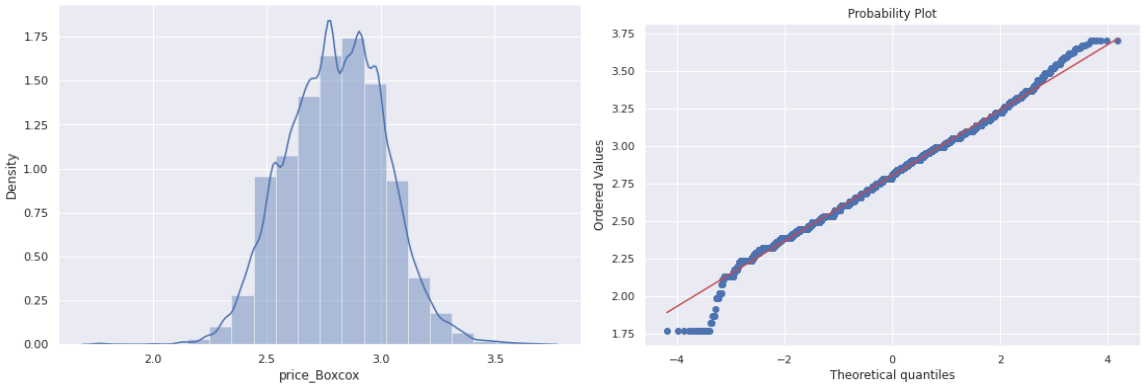
نمودار سمت راست نمودار Q-Q است ، که محور x شامل مقادیر quantile و محور y شامل مقادیر price است. اگر مقادیر داده price ما نزدیک به خط x=y باشد در آنصورت توزیع ما نرمال است، اما در اینجا همانطور که مشاهده میکنید توزیع نرمال نیست چون مقادیر ما از خط مورد نظر فاصله دارند.

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

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

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

هرگاه که $\lambda = 0$ باشد، نیز مقدار لگاریتم طبیعیه 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 تقریبا توزیع ما کاملا نرمال شده، حتی اگر mode،median و mean را نیز محاسبه کنیم، مشاهده میکنیم که کاملا نزدیک به یکدیگر هستند.

Mean → 2.802

Median → 2.800

Mode → 2.781

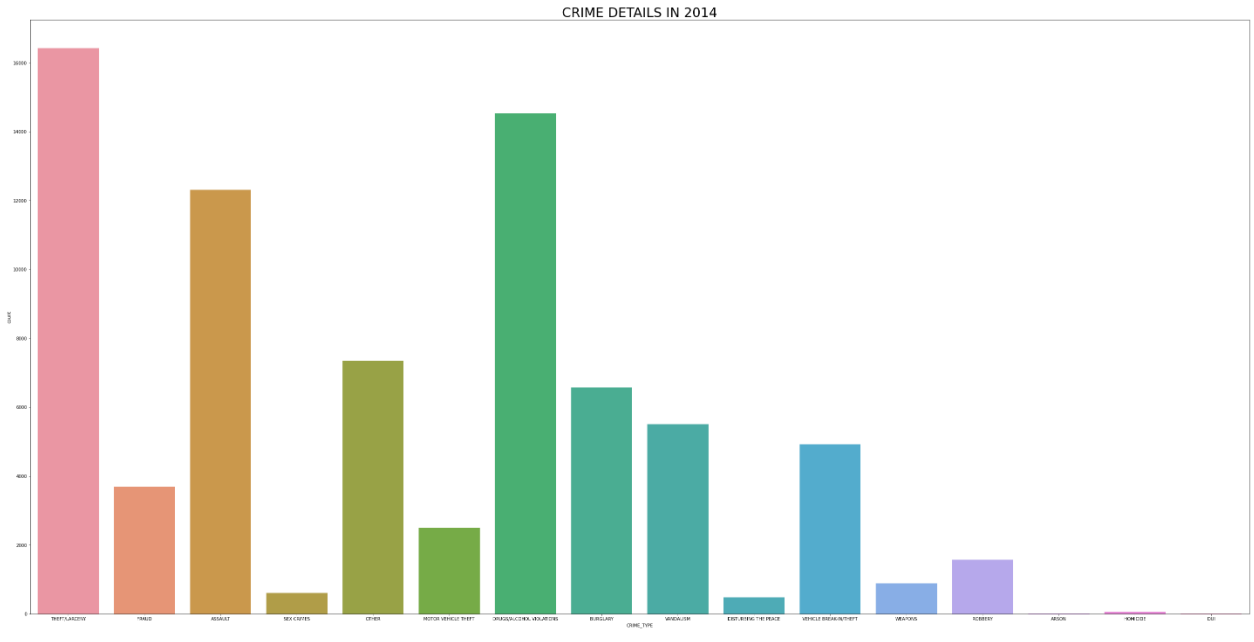
پس از آنکه توزیع ما بسیار به توزیع نرمال نزدیک شد، نتیجه تست هایی که در ادامه داریم بسیار دقیق تر خواهد بود.

۲. مجموعه داده های گزارش های جنائی

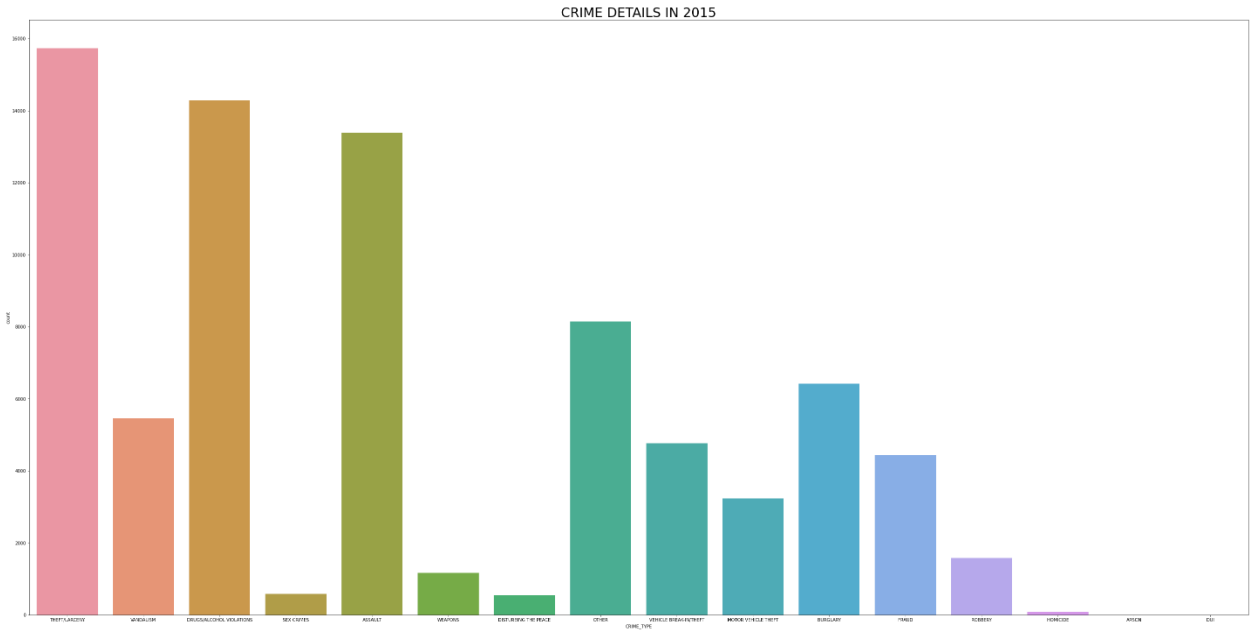
برای این مجموعه داده، داده های ۵ سال متوالی ۲۰۱۴ تا ۲۰۱۸ را درنظر گرفتم

ابتدا برای آنکه نشان دهیم چه جرمی بیشتر از همه انجام شده، در این ۵ سال هر کدام را به صورت جداگانه محاسبه و در نهایت نیز با concat کردن هر ۵ دیتا ست، با استفاده از نمودار ها و description اطلاعات لازم را بدست آوردیم.

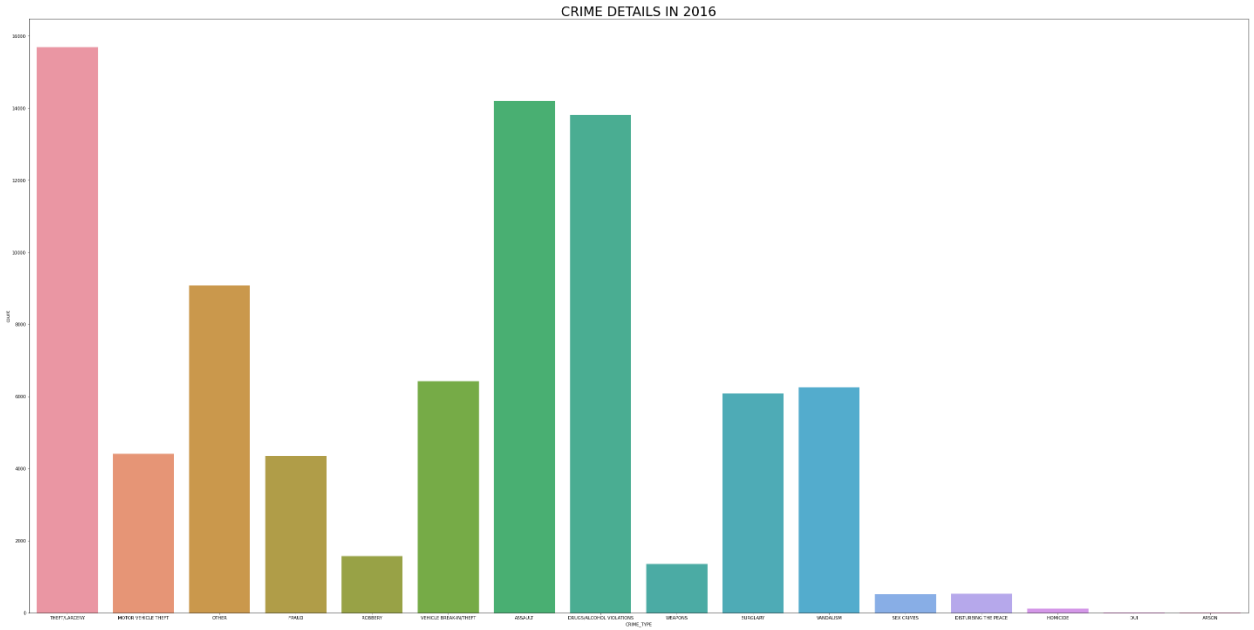
نمودار های زیر برای هر سال نشان میدهد که چه جرمی بیشتر در آن سال انجام شده:



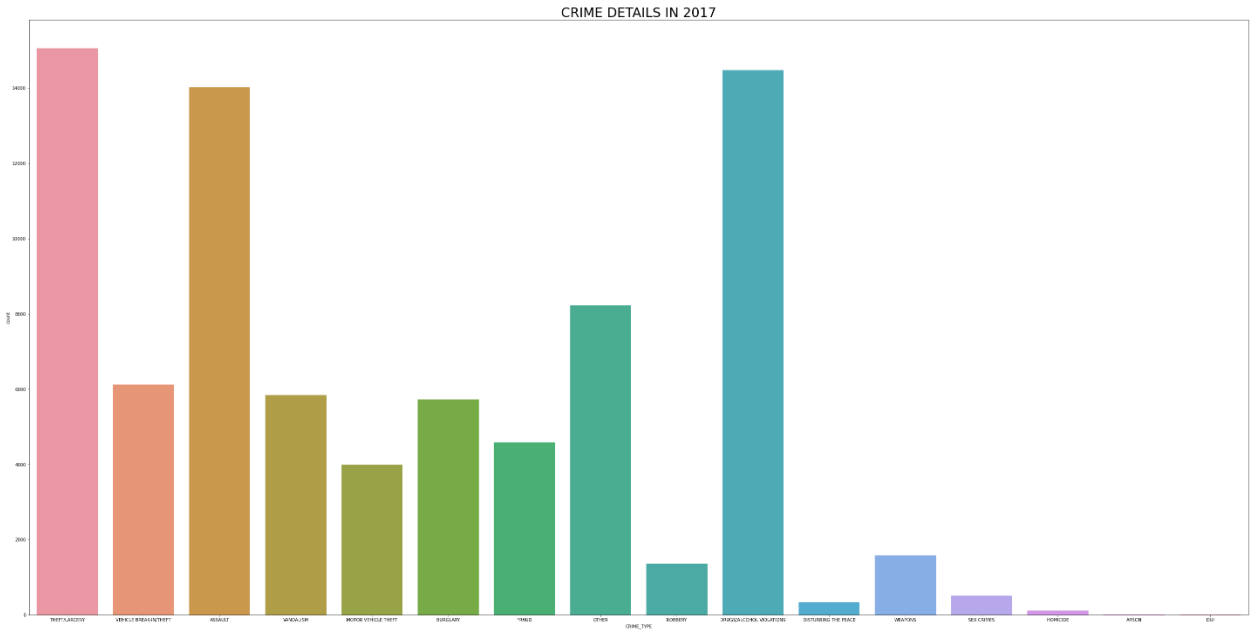
در سال ۲۰۱۴ مشاهده میشود که جرم THEFT/LARCENY بیشترین وقوع را داشته و پس از آن DRUGS/ALCOHOL VIOLATION بوده است.



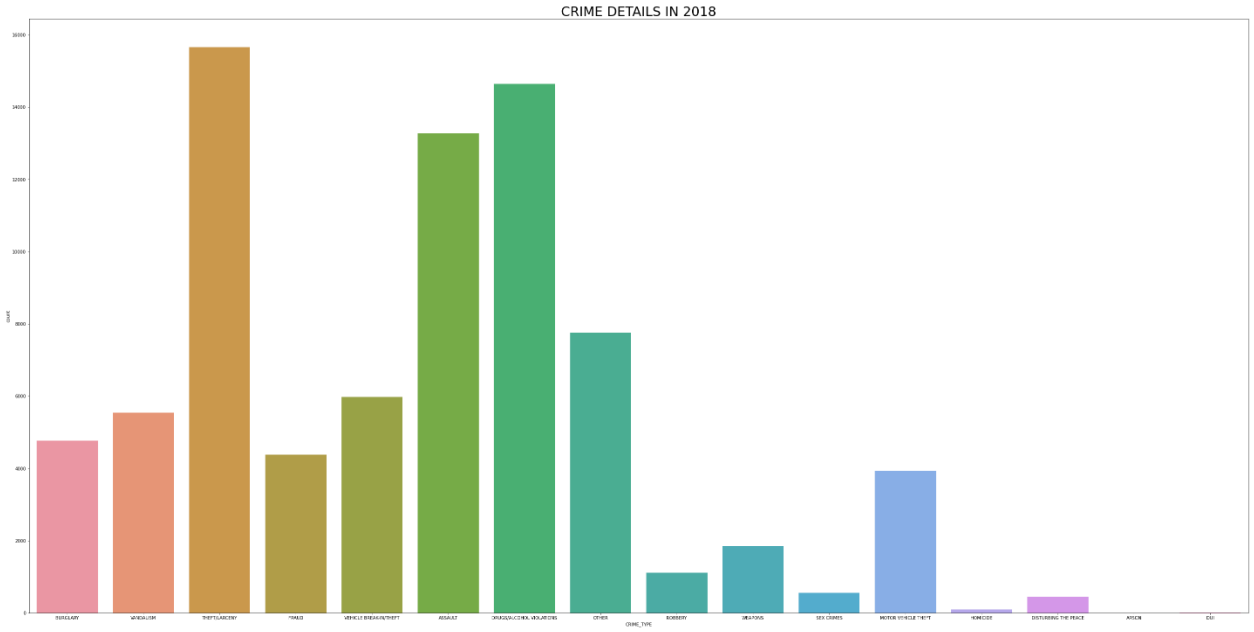
نتایج سال ۲۰۱۵ نیز مانند سال ۲۰۱۴ بوده است.



سال ۲۰۱۶ نیز مانند ۲۰۱۵ بوده است و همچنان TEFTH از بقیه جرم ها بیشتر اتفاق افتاده اما جرم ASAULT از DRUG VIOLATION پیشی گرفته و بیشتر شده است.



سال ۲۰۱۷ نیز تقریبا مانند ۲۰۱۶ بوده است اما دوباره DRUG VIOLATION زیاد شده و از ASSAULT بیشتر شده.



سال ۲۰۱۸ نیز مانند سال قبلی بوده است.

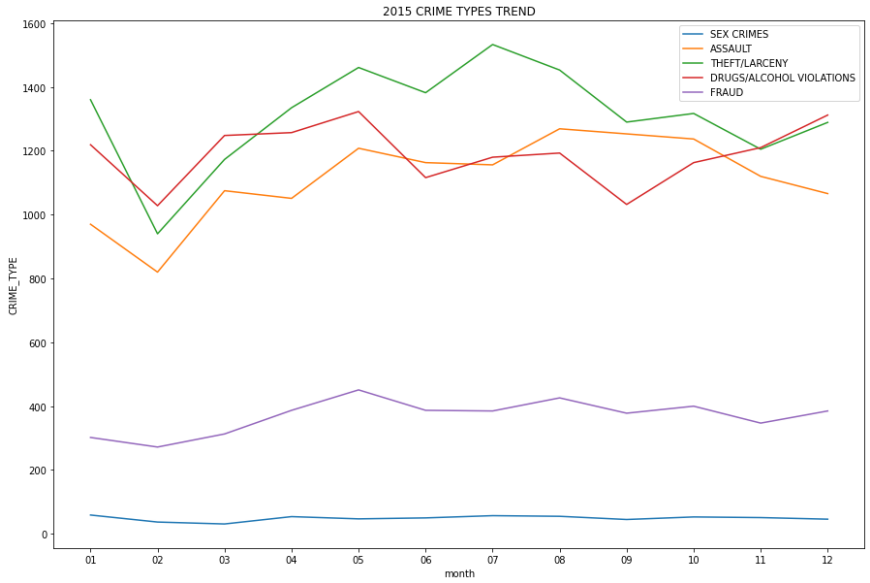
وقتی تمامی این ۵ سال را در کنار هم قرار دهیم میبینیم که THEFT/LARCENY با تعداد ۴۰۳۷۰۶ تا از بقیه جرم ها بیشتر اتفاق افتاده است.

در مورد اینکه در کدام ZIP CODE بیشتر جرم اتفاق افتاده است، همانطور که مشخص است، در ZIP CODE = 40203.0 بیشتر تعداد جرم و جنایت انجام شده است همچنین اطلاعاتی که بدست آوردیم، مشاهده میکنیم که بیشترین جنایت که THEFT/LARCENY بود در ZIP CODE = 40219.0 اتفاق رخ داده است.

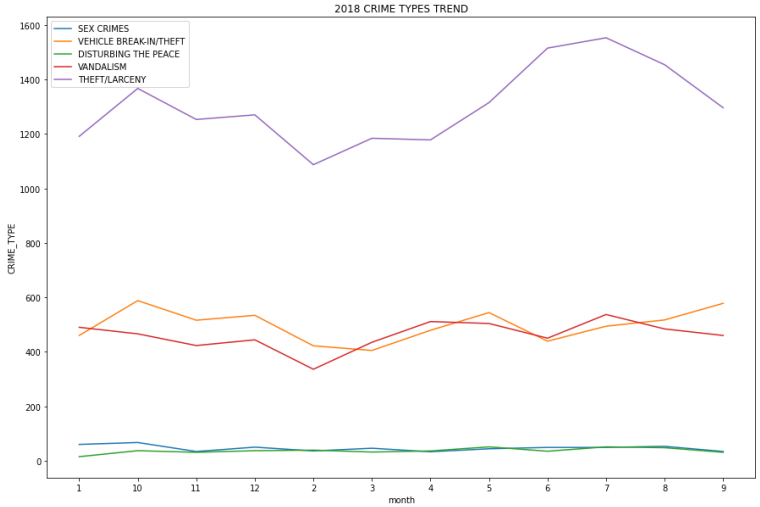
اطلاعات جالب آن است که بیشترین جنایتی که در ZIP CODE = 40203.0 رخ داده است مربوط به DRUGS/ALCOHOL VIOLATIONS بوده است.

برای بررسی میزان افزایش یا کاهش جرایم، بررسی ها را ابتدا بر اساس ماه بر روی دو دیتاست مربوط به سال های ۲۰۱۵ و ۲۰۱۸ انجام داده‌ایم بدین صورت که از زمانی که جرم مورد نظر گزارش شده، تاریخ آن و سپس ماه آن را جدا کردیم و بر اساس ماه ها و تعداد وقوع جرایم مختلف نمودار های متفاوتی را رسم نمودیم.

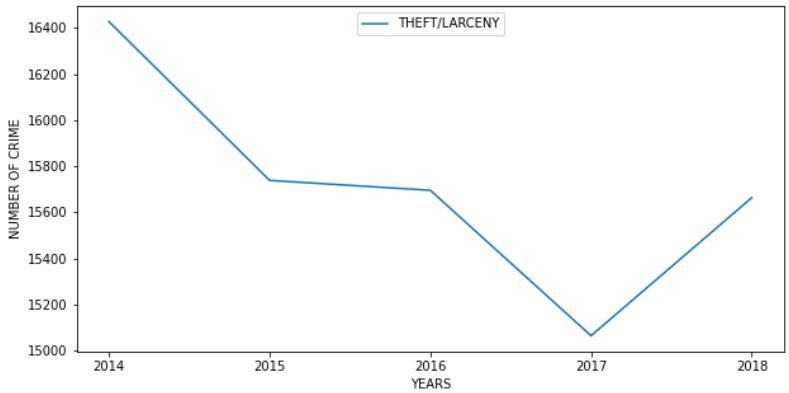
در سال ۲۰۱۵ تعداد وقوع جرایم SEX CRIMES, ASSAULT, 'THEFT/LARCENY', DRUGS/ALCOHOL VIOLATIONS, FRAUD را در ۱۲ ماه بررسی کردیم. نمودار زیر trend این جرایم را نشان میدهد:

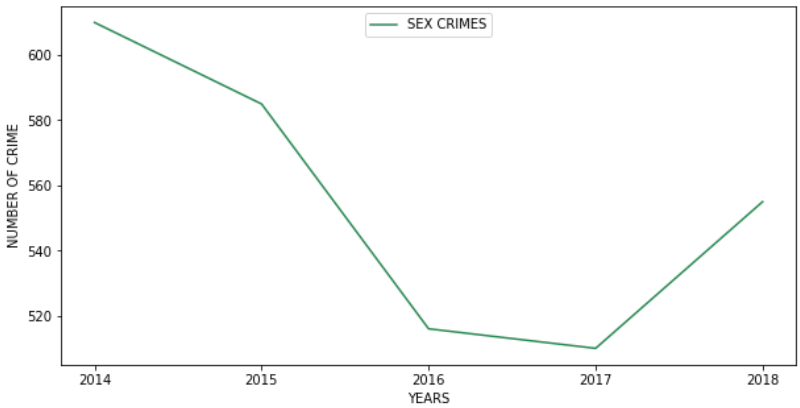


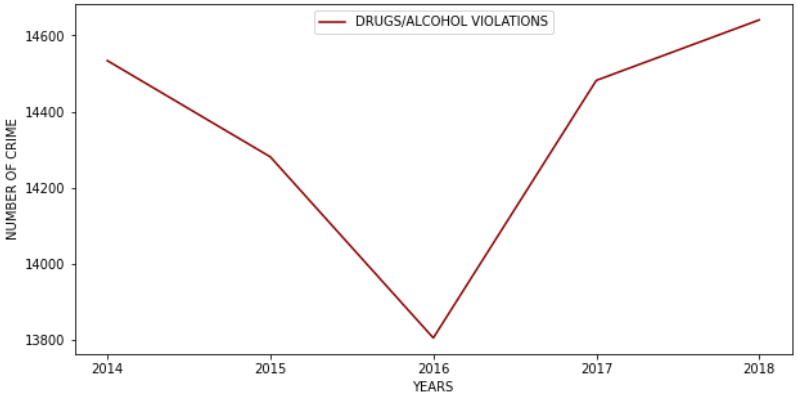
در سال ۲۰۱۸ نیز تعداد وقوع جرایم SEX CRIMES, VEHICLE BREAK-IN/THEFT, DISTURBING THE PEACE, VANDALISM, THEFT/LARCENY را در ۱۲ ماه بررسی کردیم:

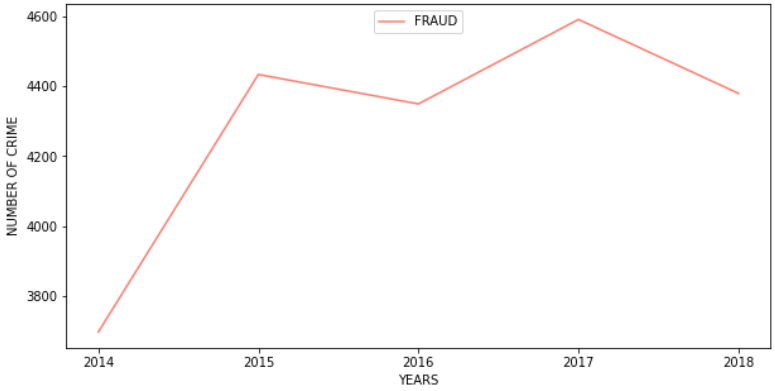


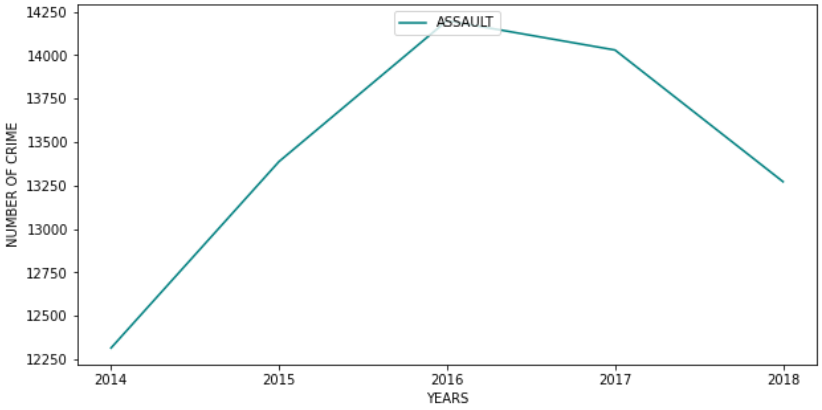
سپس بررسی را سالیانه انجام دادیم و برای سال های متوالی ۲۰۱۴ الی ۲۰۱۸ به بررسی ترند جرایم THEFT/LARCENY, SEX CRIMES, DRUGS/ALCOHOL VIOLATIONS, FRAUD, ASSAULT, ROBBERY پرداختیم که نمودار هر یک از آنها را در زیر باهم مشاهده میکنیم:

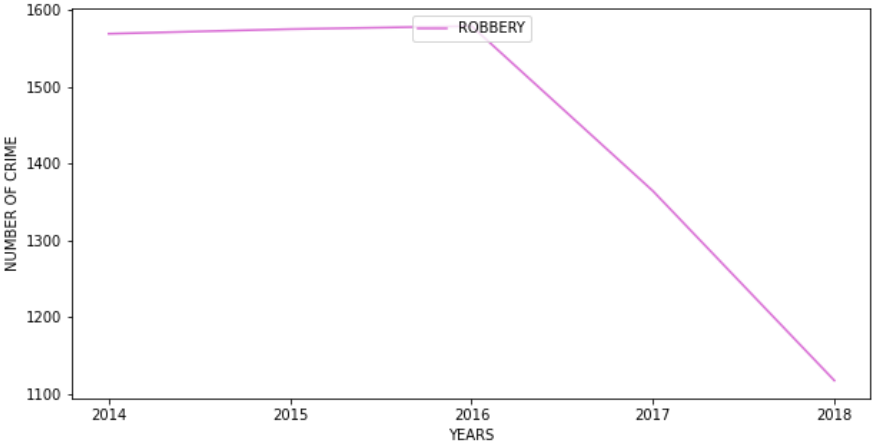





DRUGS/ALCOHOL VIOLATIONS
NUMBER OF CRIME
14600
14400
14200
14000
13800
2014
2015
2016
2017
2018
YEARS


FRAUD
NUMBER OF CRIME
4600
4400
4200
4000
3800
2014
2015
2016
2017
2018
YEARS


ASSAULT
NUMBER OF CRIME
14250
14000
13750
13500
13250
13000
12750
12500
12250
2014
2015
2016
2017
2018
YEARS

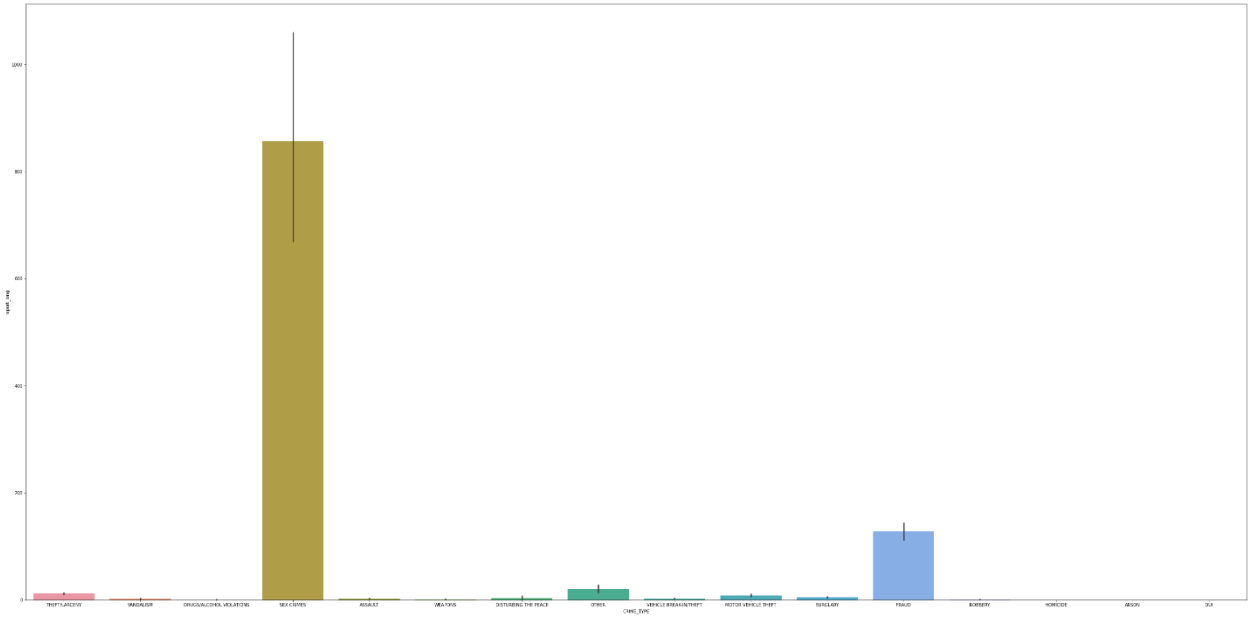

ROBBERY
NUMBER OF CRIME
1600
1500
1400
1300
1200
1100
2014
2015
2016
2017
2018
YEARS

همانطور که در نمودار ها نیز مشخص است میتوان گفت که در این ۵ سال، جرم ROBBERY روند نزولی و جرائم FRAUD و ASSAULT روند صعودی داشته‌اند.

برای بررسی آنکه ببینیم کدام جرم بیشترین زمان را تا reportکردن داشته است از دو داده DATE_REPORTED و DATE_OCCOURD استفاده کردیم. در واقع اختلاف زمانی که جرم اتفاق افتاده تا زمانی که گزارش شده را محاسبه کردیم. این دو داده نوع آنها string است و نمیتوانستیم اختلاف آنها را محاسبه کنیم به همین علت مجبور بودیم این داده را به نوع dateکه برای خود pandas است تبدیل کنیم و پس از محاسبه اختلاف آنها دوباره آن را به str تبدیل کردیم تا بتوانیم بخش اول آن که فقط تعداد روز طول کشیده را نشان میدهد بیابیم.

در دیتاست اول چون 3 داده پوچ موجود بود اول آنها را حذف کردیم و سپس عملیات را انجام داده‌ایم.

همانطور که مشاهده میکنیم با مقایسه روز هایی که این اتفاقات طول کشیده تا گزارش شوند، میبینیم که جرم SEX CRIME با اختلاف بیشترین زمان را برده تا گزارش شود و این قابل حدس نیز بود.

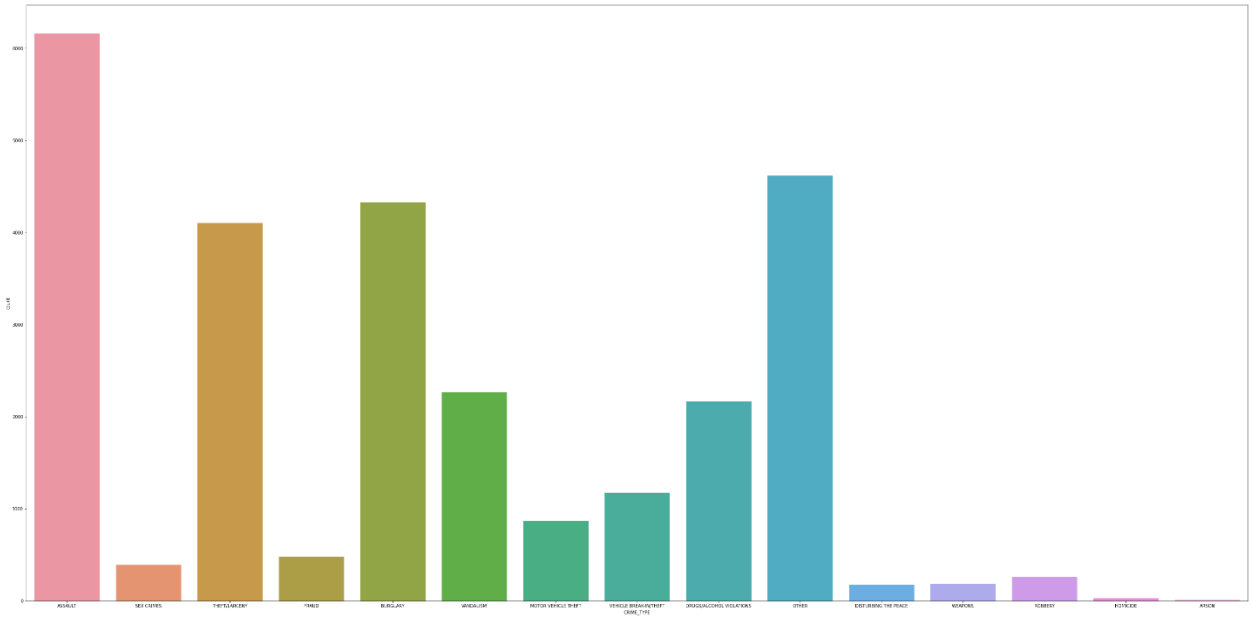


سپس به بررسی اطلاعات دیگر پرداختیم. مانند آنکه بیشترین جرم در کدام شهر ها اتفاق افتاده و چه جرمی بوده است؟

تعداد شهر هایی که در آنها جرم ها اتفاق افتاده حدود ۱۱۰ تا بود، در این شهر ها همانطور که میبینیم، ۷۳۷۰۴ تا مربوط به شهر LOUISVILLE بوده و پس از آن نیز با اختلاف شهر LVIL است. در این دو شهر با توجه به نمودار ها میبینیم که در شهر LOUISVILLE جرم THEFT/LARCENY بوده و در شهر LVIL بیشترین جرم مربوط به DRUGS/ALCOHOL VIOLATIONS بوده است.

در ادامه نیز به بررسی PREMISE_TYPE پرداختیم. این ستون به معنی جاییست که جرم در آن اتفاق افتاده چه نوع مکانی بوده است. مانند خانه، خیابان، بزرگراه، رستوران و ...

با توجه به description اعمال شده میتوانیم متوجه شویم که بیشتر جرایم در RESIDENCE / HOME اتفاق افتاده است. حال در این مکان چه جرمی بیشتر رخ داده؟ همانطور که در نمودار زیر نیز مشاهده میکنیم، ASSAULT بیشتر اتفاق افتاده است.



همچنین دومین مکانی که بیشتر جرائم در آنجا اتفاق افتاده مربوط به خیابان و بزرگراه ها یا درواقع HIGHWAY / ROAD / ALLEY بوده است. که با توجه به نمودار زیر میتوانیم متوجه شویم که در این مکان ها بیشتر جرائم مربوط به تصادفات و خشونت ها در اثر مصرف الکل و مواد یا همان DRUGS/ALCOHOL VIOLATIONS بوده است.