نام و نام خانوادگی: درسا الهوئی

شماره دانشجویی: ۹۷۲۲۲۰۰۵

شماره تمرین: ۱ **تاریخ تحویل: ۲ آبان ۱۴۰۰**

عنوان تمرین: تحلیل آماری داده های Airbnb در بخش اول و گزارش جرایم در بخش دوم

---

۱. مقدمه

هدف از این گزارش بررسی آماری داده ها همراه با استفاده از تست های آماری آموخته شده است.

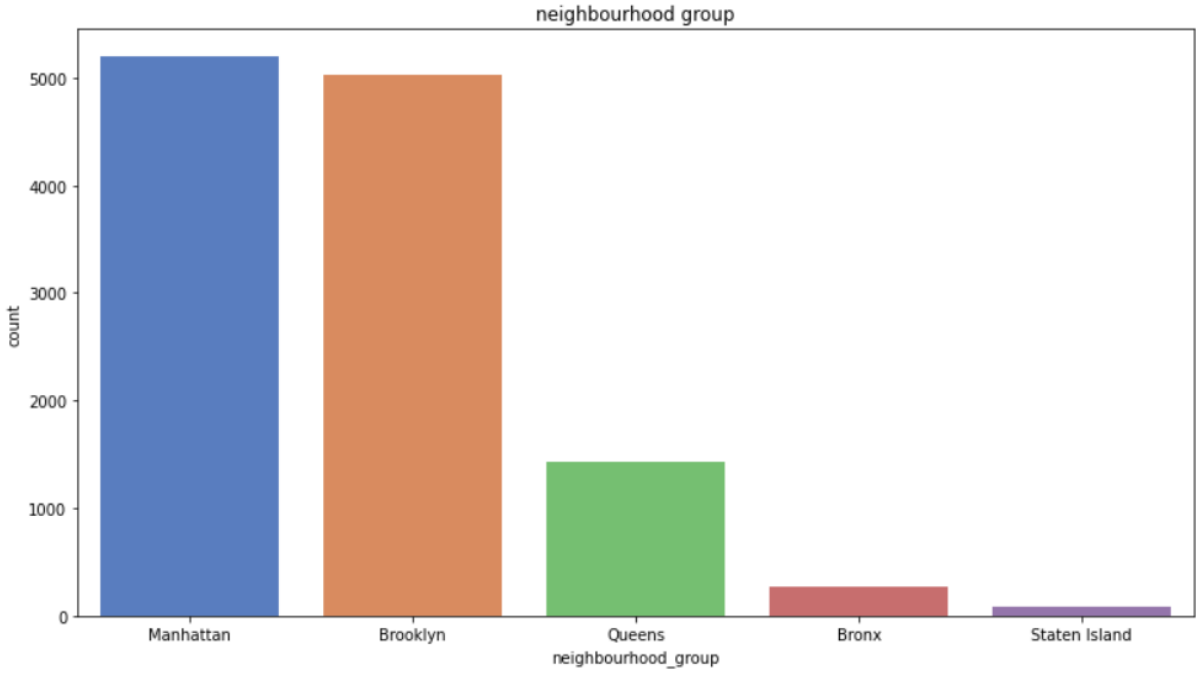
* لازم به ذکر است که تمامی تست های به کار گرفته شده، با فرض نرمال بودن داده ها است. (منبع این فرض نیز در قسمت منابع ذکر شده)

۲. تحلیل آماری داده های Airbnb

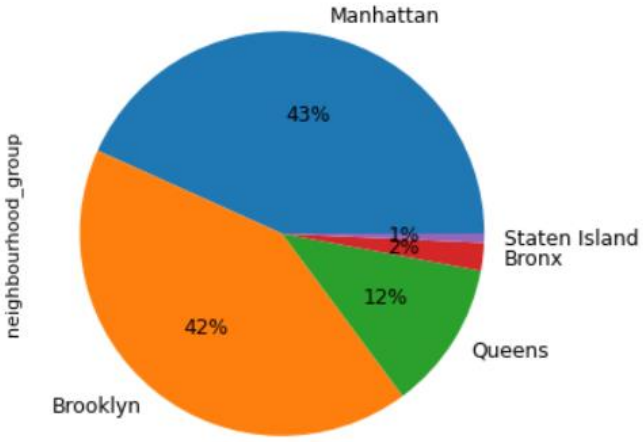
به منظور تحلیل آماری داده ها از آنها نمونه برداری می کنیم. قبل از این کار پردازش داده را انجام دادیم. برای مثال حذف ستون هایی مثل "name" و "host_name " که عملا کاربردی ندارند. همچنین حذف سطر مربوط ستون هایی که دارای تعداد زیادی داده null هستند. بعد از پایان پردازش داده ها، از آنها نمونه برداری می کنیم. در اینجا نمونه انتخابی ۱۲۰۰۰ عضو دارد.

در گام بعدی یک ماتریس پراکندگی (scatter) که در واقع مجموعه ای از نمودار های پراکندگی است رسم کردیم.

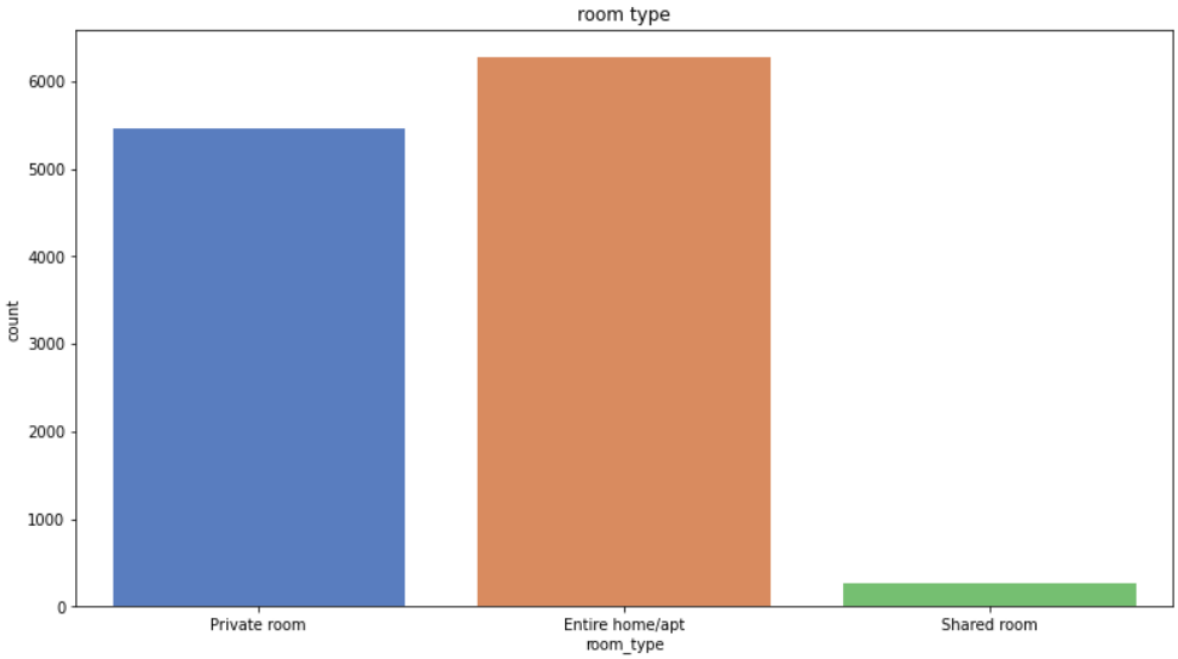
در مرحله بعد مناطقی که خانه ها در آنها پراکنده شدند را پیدا کردیم و نمودار تعداد خانه ها در هر محله را رسم کردیم. طبق این نمودار بیشترین تعداد خانه ها در منطقه منهتن و سپس با اختلاف کمی در منطقه بروکلین عرضه می شوند.



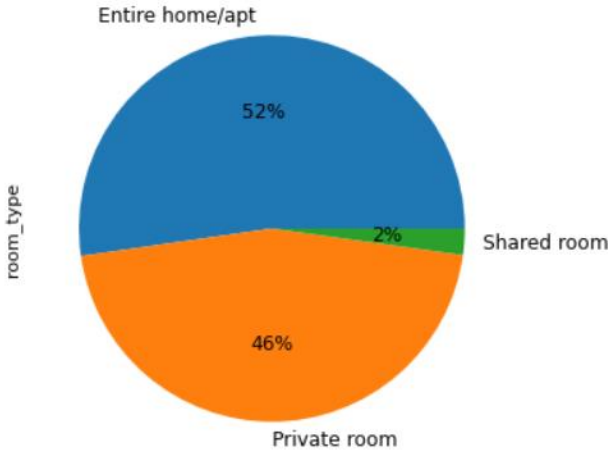
همچنین نمودار دایره ای (pie) آن را رسم کردیم. این نمودار درصد پراکندگی خانه ها را به ما نشان می دهد.



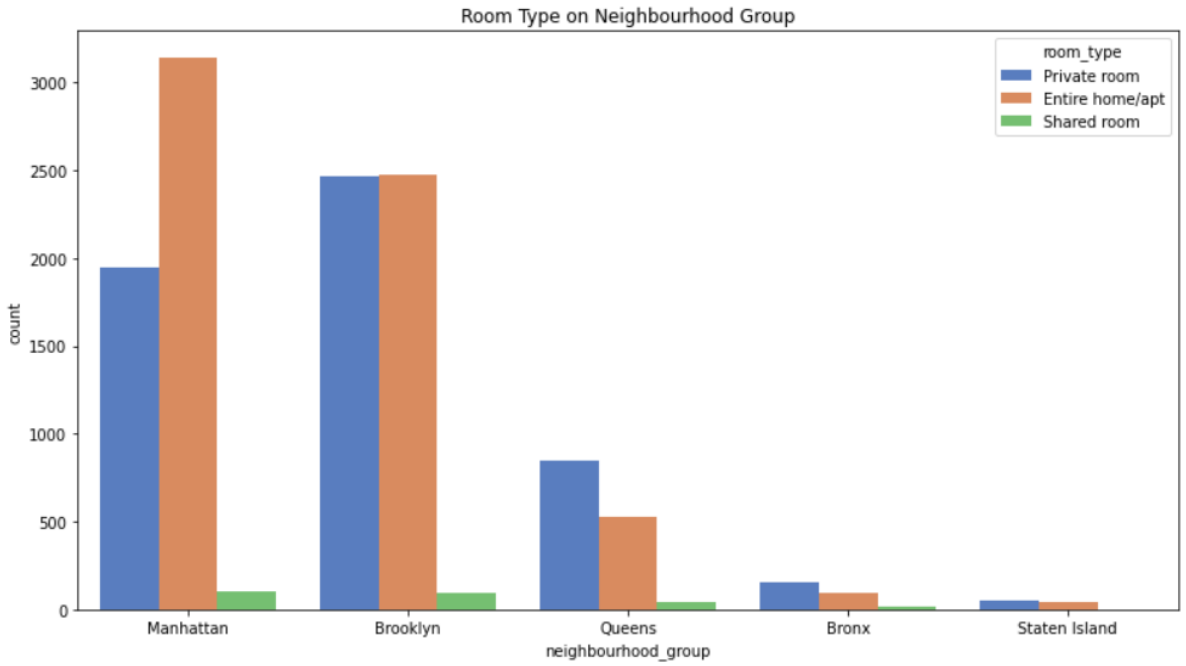
نمودار بعدی که رسم کردیم مربوط به پراکندگی خانه ها در سه نوع خانه موجود است که بیشترین آنها در دو نوع entire home/apt و private room قرار می گیرند.



نمودار دایره ای مربوط به این داده ها نیز این موضوع را تایید می کند.



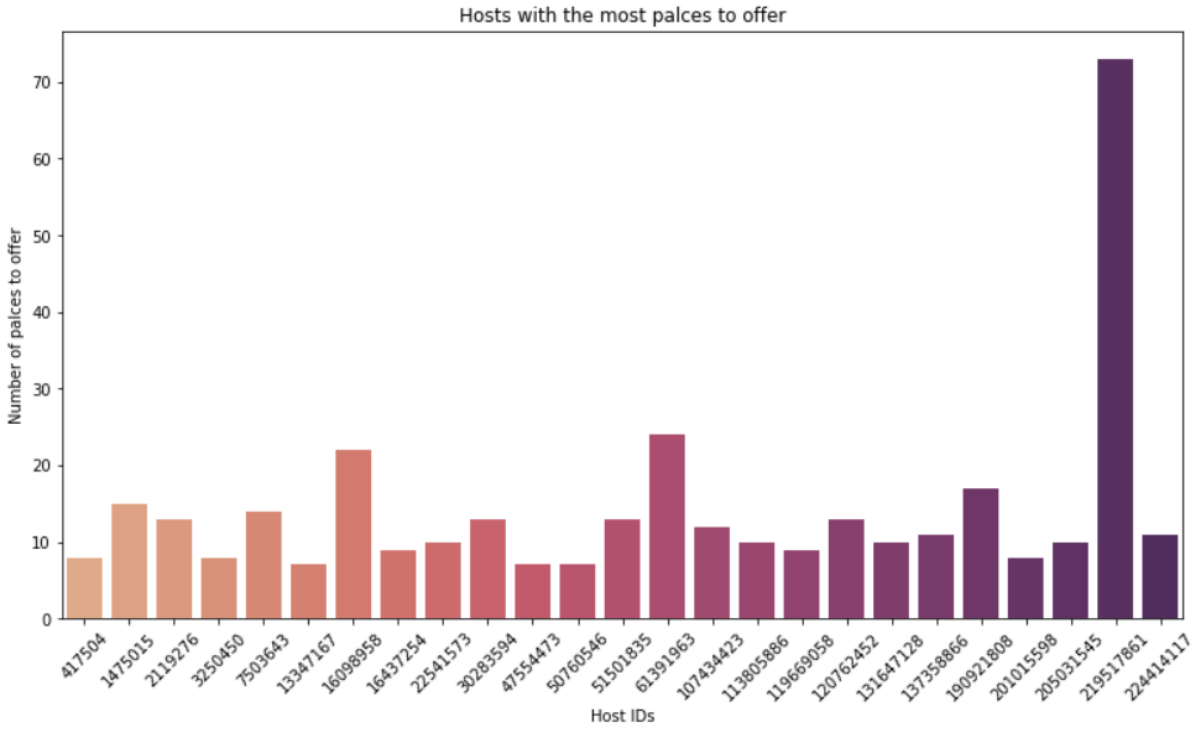
در گام بعدی ترکیبی از دو نمودار بالا را رسم کردیم. یعنی پراکندگی خانه های هر یک از سه نوع در هر یک از پنج منطقه.



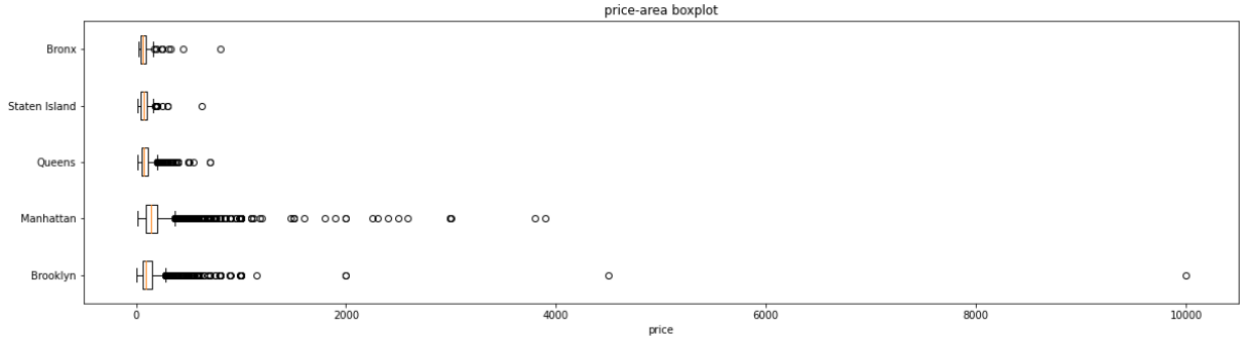
جدولی که در این گام کشیده شده است، بیانگر شناسه (ID) میزبانی می باشد که بیشترین تعداد خانه را عرضه می کند.

	hostid	count
0	219517861	73
1	61391963	24
2	16098958	22
3	190921808	17
4	1475015	15
5	7503643	14
6	51501835	13
7	120762452	13
8	2119276	13
9	30283594	13
10	107434423	12
11	224414117	11
12	137358866	11
13	205031545	10
14	113805886	10
15	131647128	10
16	22541573	10
17	119669058	9

در نمودار زیر نیز تعداد خانه های ۲۵ عرضه کننده برتر را رسم کردیم.



نمودار زیر نیز، نمودار جعبه ای مربوط به پراکندگی قیمت خانه ها در هر منطقه می باشد.



در این مرحله تست ANOVA را روی قیمت در هر منطقه انجام دادیم. این تست به این صورت است که میانگین قیمت خانه ها را در هر منطقه محاسبه می کند و اگر تنها یکی از میانگین ها با چهار میانگین دیگر برابر نباشد، فرض صفر رد می شود.

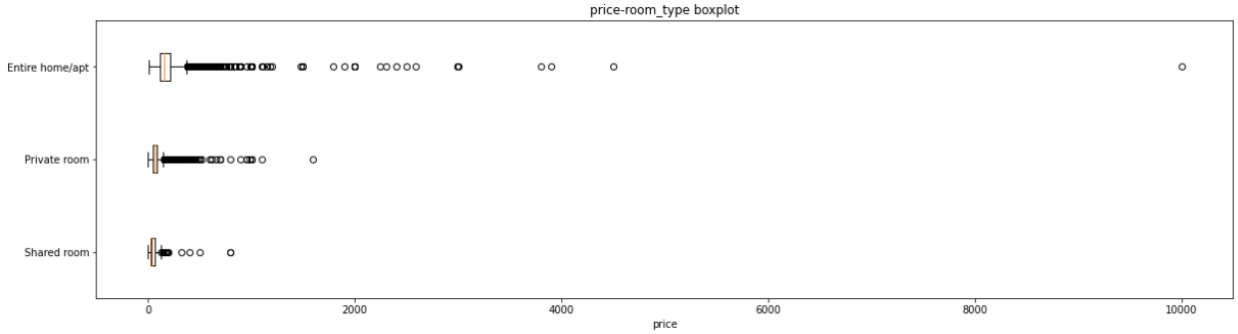
```
F-Statistic=119.572, p=0.0000
```

در اینجا p-value مقداری برابر صفر دارد (نه الزاما صفر مطلق. کوچکتر بودن آن از ۰,۰۵ کفایت می کند) که به این معناست که فرض صفر که فرض برابر بودن میانگین هر پنج دسته بود رد می شود.

مشاهده می کنید که با چاپ میانگین هر دسته این موضوع به وضوح مشهود است.

```
Average price in Brooklyn:        121.19109166832372
Average price in Manhattan:       179.3312788906009
Average price in Queens:          91.66199158485273
Average price in Staten Island:   94.17241379310344
Average price in Bronx:           78.57142857142857
```

نمودار جعبه ای زیر نیز پراکندگی قیمت بسته به هر سه نوع از خانه ها را نمایش می دهد.



طبق نتیجه تست ANOVA فرض صفر ما رد می شود. یعنی هر سه میانگین قیمت برابر نیستند.

```
F-Statistic=736.462, p=0.0000
```

مشاهده می شود که میانگین قیمت ها نیز با یکدیگر برابر نیست.

```
Average price of Shared room:       62.68821292775665
Average price of Private room:      81.63488286969253
Average price of Entire home/apt:   197.3252032520325
```

در دو نوع shared room و private room میانگین قیمت ها نزدیک به هم است. به منظور بررسی بیشتر یک two sample t test روی این دو اعمال کردیم. خروجی به صورت زیر است:

```
statistic=-4.476789404467649, pvalue=7.7244045054561e-06
```

فرض صفر، که فرض برابری میانگین دو دسته است رد میشود چراه p-value مقدار کوچکتری از ۰,۰۵ دارد.

در مرحله بعدی نمودار جعبه ای حداقل تعداد شب رزرو خانه به قیمت رسم شده است. این نمودار در کد قابل مشاهده می باشد.

در اینجا با توجه به حداقل شب و قیمت (داده کتگوریکال و عددی) از تست pearson استفاده کردیم. نتایج به صورت زیر است:

```
correlation=0.024, pvalue=0.00923
```

با توجه به مقدار p-value فرض صفر (فرض بر قرار نبودن رابطه خطی بین پارامترهای مورد بررسی) رد می شود. با توجه به مقدار correlation، رابطه مستقیم ضعیفی بین آنها برقرار است.

در ادامه نیز نمودار پراکندگی قیمت با توجه به حداقل شب رزرو رسم شده است. این نمودار در کد قابل مشاهده است.

با توجه به نمودار جعبه ای قیمت، نمودار دارای چولگی می باشد. خط نمایانگر میانگین (خط آبی) بر خط نمایانگر میانه (خط نارنجی) منطبق نمی باشد.

*نکته: عدم همخوانی تصاویر نمودارها با کد به علت اجرای مجدد کد است. در هر مرحله نمونه انتخابی متفاوت می باشد.

۳. تحلیل آماری داده های مربوط به جرایم

در این بخش دیتای مربوط به پنج سال گذشته را از سایت دانلود کردیم. این دیتا شامل فایل هایی از سال ۲۰۱۵ تا ۲۰۱۹ بود که در فایل ۲۰۱۹ علاوه بر دیتا مربوط به آن سال، دیتا مربوط به سال ۲۰۲۰ و ۲۰۲۱ نیز قرار داشت. در گام اول، سطرهایی که دارای ستونی با مقدار null بودند را حذف کردیم. در مرحله بعد sample ای با تعداد عضو ۱۰۰۰۰۰ از کل داده ها برداشتیم.

در گام بعدی، هدف رسم تعداد رویداد هر جرم در سال های مختلف بود اما به نکته جالبی بر خوردیم آن هم متفاوت بودن فرمت داده های مربوط به تاریخ (ستون های date_occured و date_reported) بود. پس برای آنکه بتوانی از آن دیتا استفاده کنیم، تمامی فرمت های ممکن را لیست کردیم تا با استفاده از try-catch بتوانیم فرمت تاریخ را به فرمت درست تبدیل کنیم. سپس برای هر یک از دیتایی که در ستون های تاریخ رویداد و تاری ثبت گزارش بودند، سه ستون ایجاد کردیم.

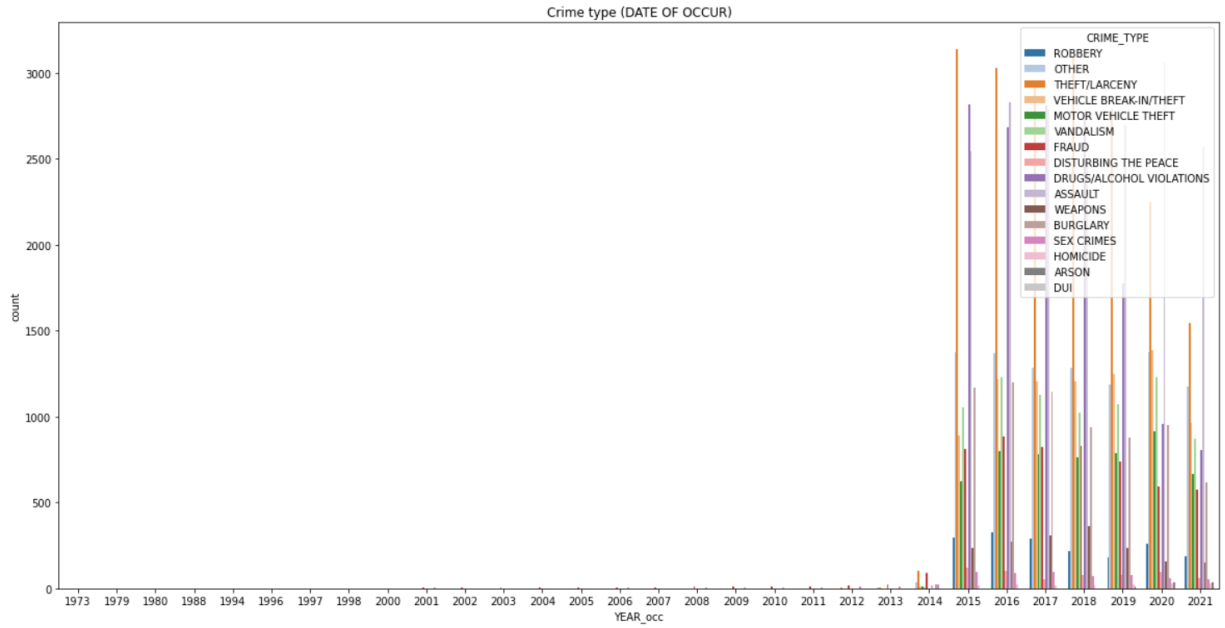
YEAR_occ, MONTH_occ, TIME_occ مربوط به زمان، ماه و سالی هستند که آن جرم در آن زمان اتفاق افتاده است و YEAR_rep, MONTH_rep, TIME_rep مربوط به زمان، ماه و سالی هستند که آن جرم در آن زمان گزارش داده شده است.

بعد از چاپ کردن ستون data.YEAR_occ.value_counts() (بیانگر تعداد جرم های رخ داده در هر سال است) به نتیجه جالبی رسیدیم!

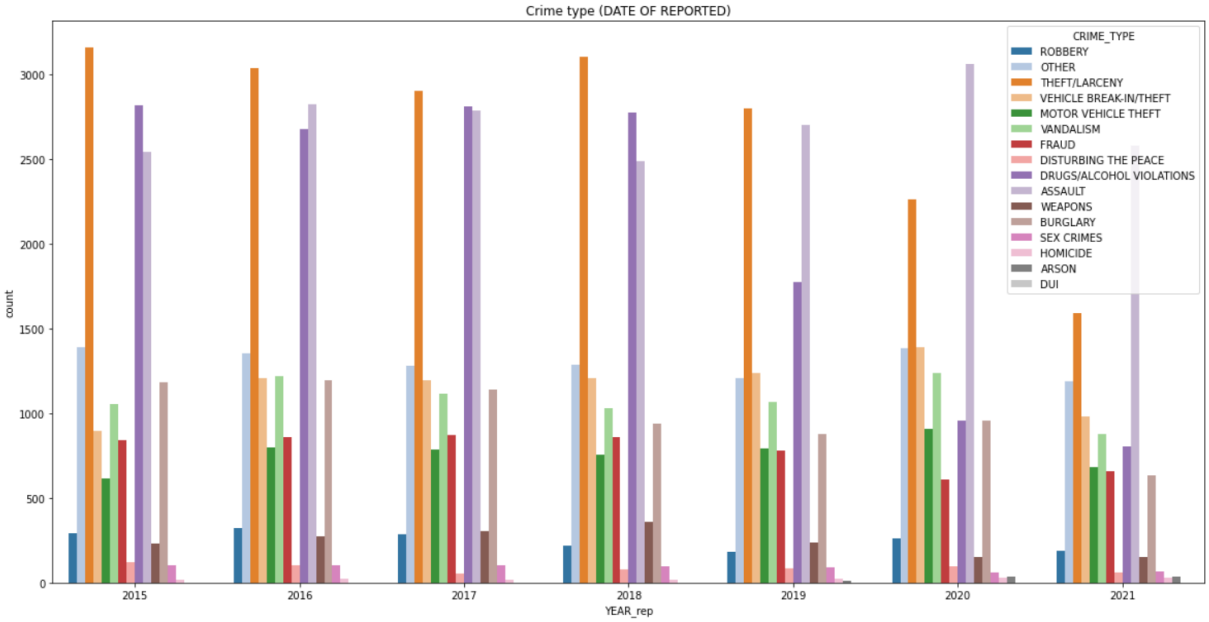
2016	16059
2017	15644
2015	15196
2018	15143
2019	13776
2020	13342
2021	10295
2014	324
2013	53
2012	34
2010	21
2008	19
2011	18
2009	17
2006	11
2007	10
2001	6
2005	5
2004	5
2003	5
2000	4
2002	3
1996	2
1994	2
1997	1
1980	1
1998	1
1979	1
1973	1
1988	1

با اینکه دیتا ما شامل داده هایی از سال ۲۰۱۵ تا ۲۰۲۱ است اما تاریخ رخداد تعدادی از جرم ها به سال ها قبل باز می گردد! یعنی فرد شاکی، پس از گذشت چندین سال اقدام به ثبت جرم کرده است.

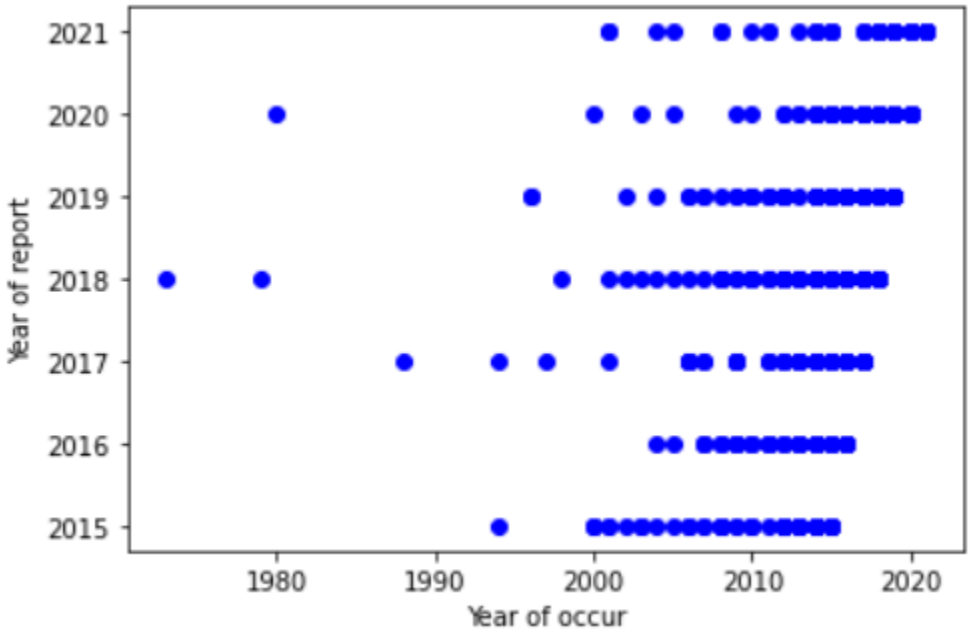
نمودار زیر مربوط به تعداد رخداد هر جرم در هر سال می باشد.



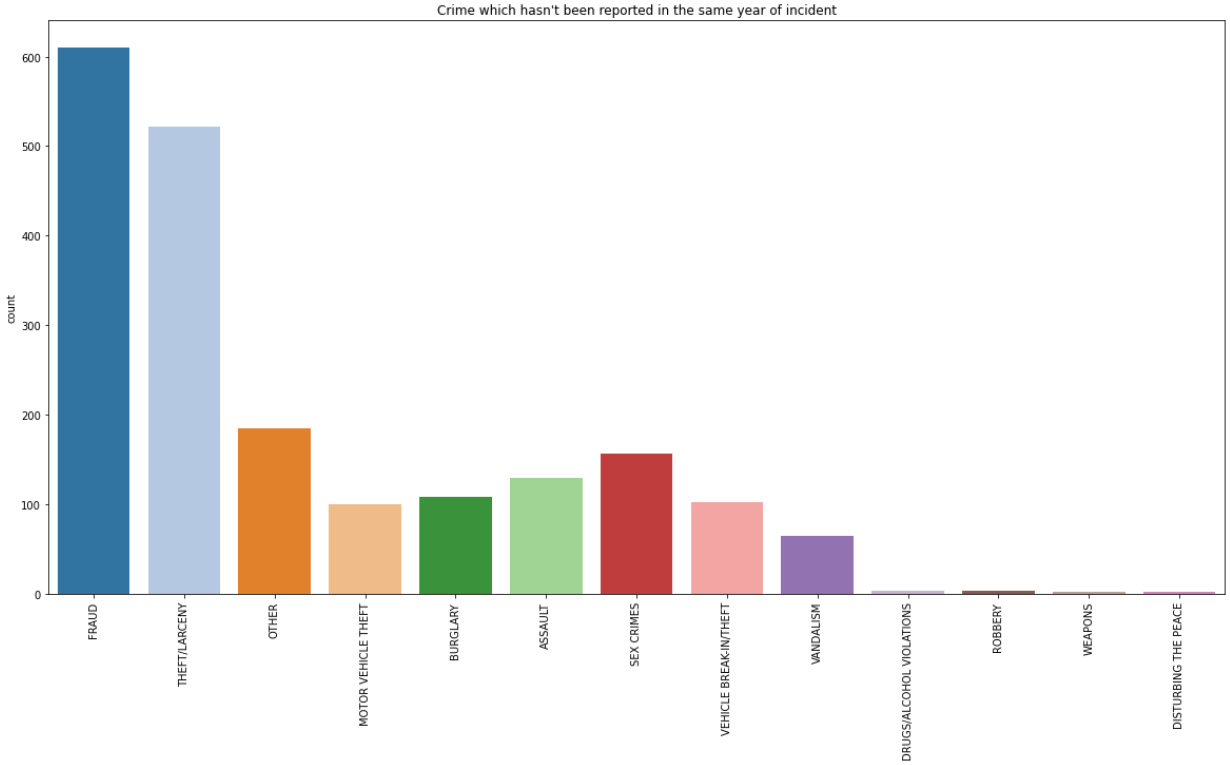
نمودار زیر نیز مربوط به تعداد گزارش هر جرم در هر سال است.



در گام بعدی، نمودار مربوط به زمان رخداد هر جرم نسبت به زمان گزارش آن را رسم کردیم.
ستون افقی زمان رخداد و ستون عمودی زمان گزارش هر جرم را نشان می دهد.



نمودار زیر نیز نوع و تعداد جرایمی را که در سال مشابه رخداد آن گزارش نشده اند را نمایش می دهد.



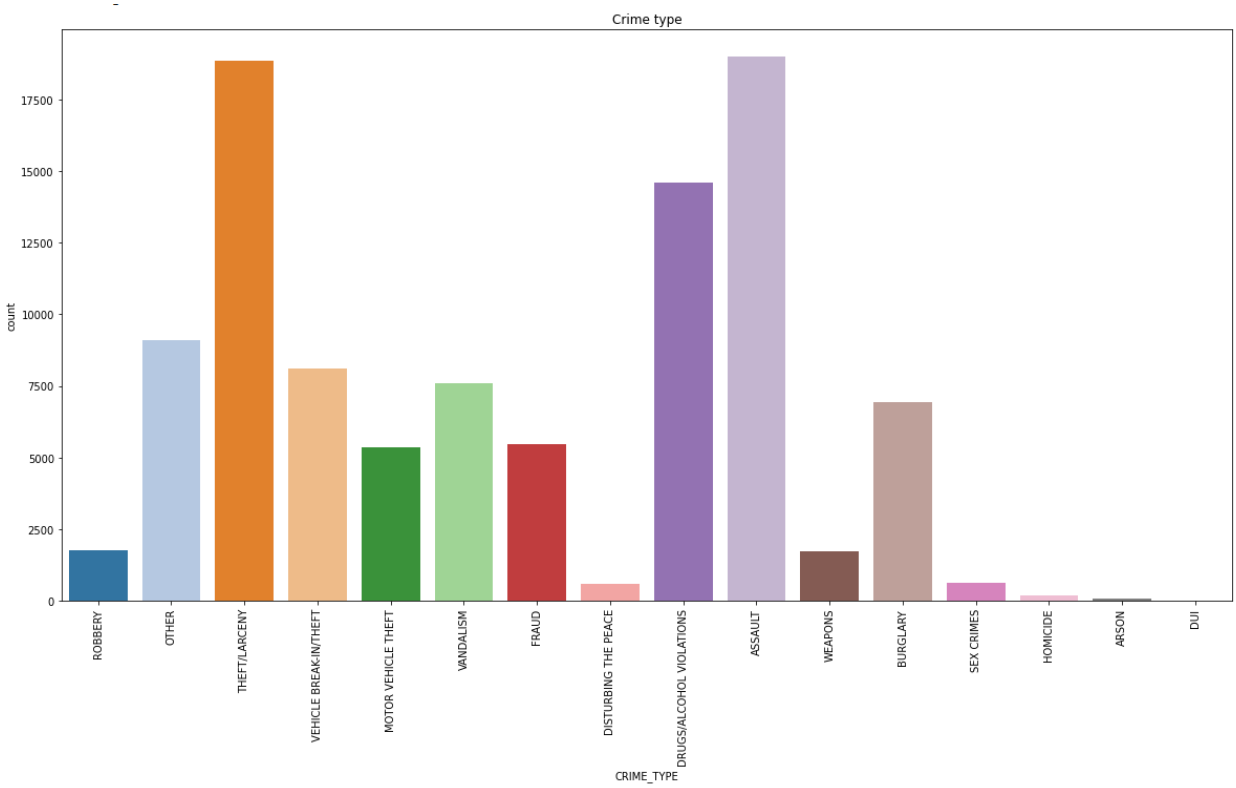
طبق این نمودار دو مورد از بیشترین جرایمی که در سال مشابه گزارش نشده اند، FRAUD و THEFT / LARCENY می باشند.

در دسته جرایم، سه مورد از آنها همگی در سال مشابه رخداد، گزارش داده شده اند. این سه دسته در زیر قابل مشاهده می باشند:

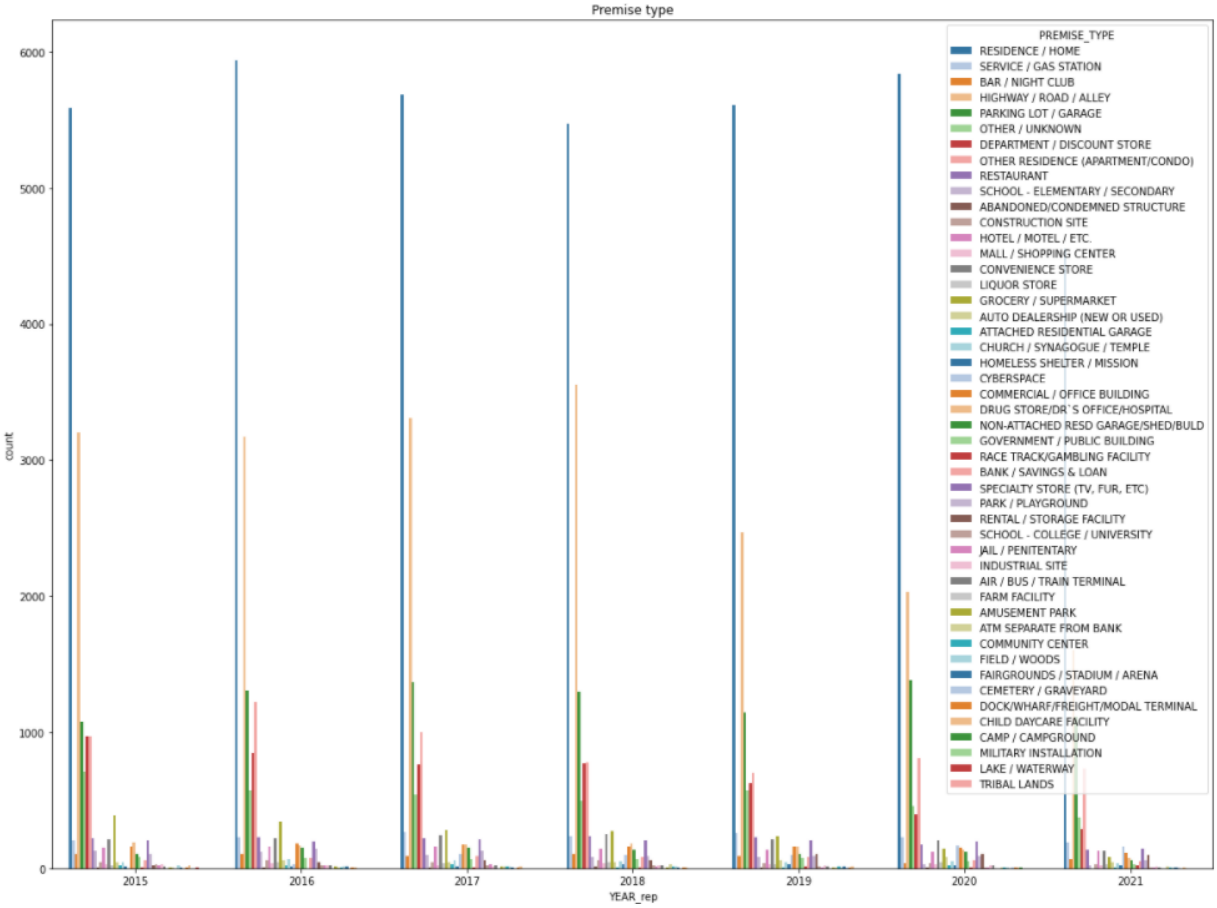
```
#Crimes which were reported the same year of the incident.
t=[]
for i in set(data.CRIME_TYPE):
  if i not in set(c):
    t.append(i)
print(t)
```

```
['ARSON', 'DUI', 'HOMICIDE']
```

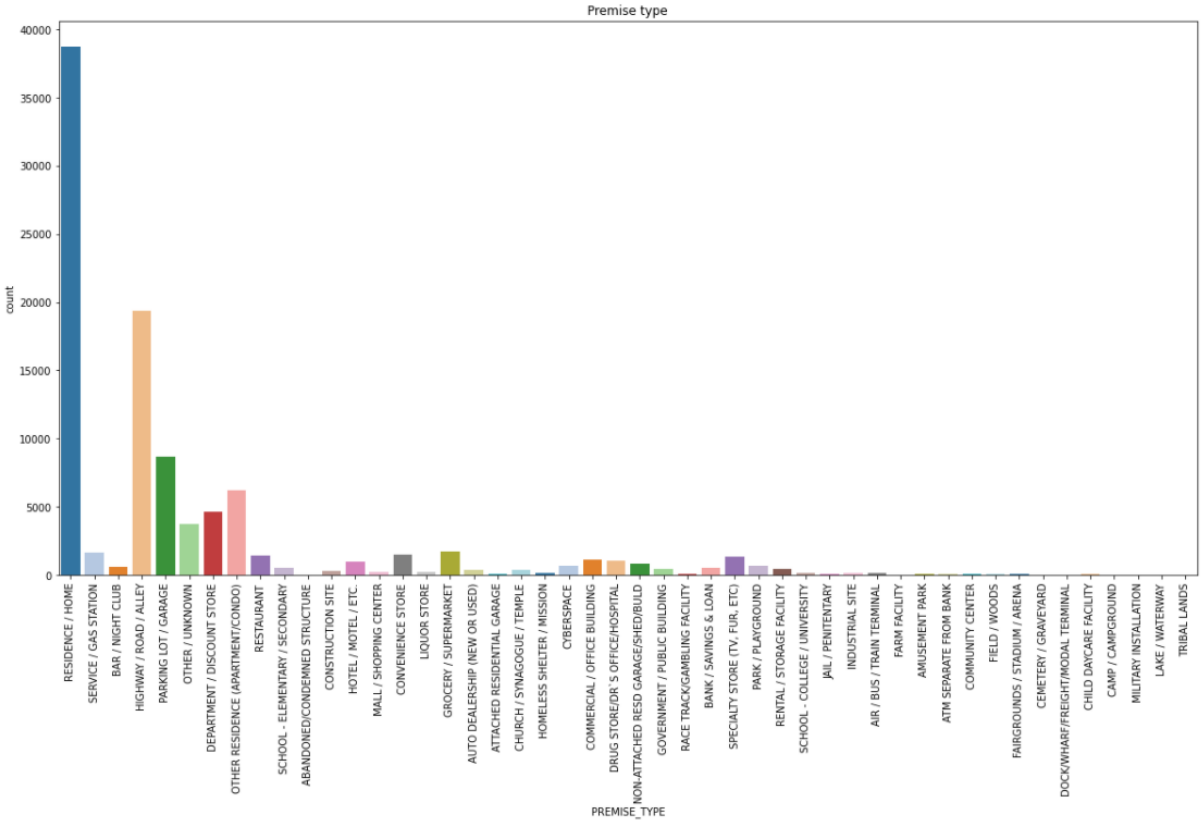
نمودار زیر تعداد رخداد هر یک از جرایم از سال ۲۰۱۵ تا ۲۰۲۱ را نشان می دهد. مشاهده می شود که در نمونه ما بیشتر جرایم از نوع THEFT / LARCENY و ASSAULT و DRUGS / ALCOHOL VIOLATIONS می باشند.



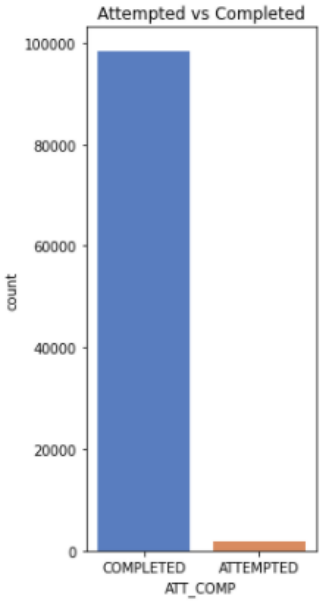
نمودار بعدی نیز تعداد گزارش هر یک از انواع جرایم در هر سال نسبت به محل رخداد آن جرم را نشان می دهد.



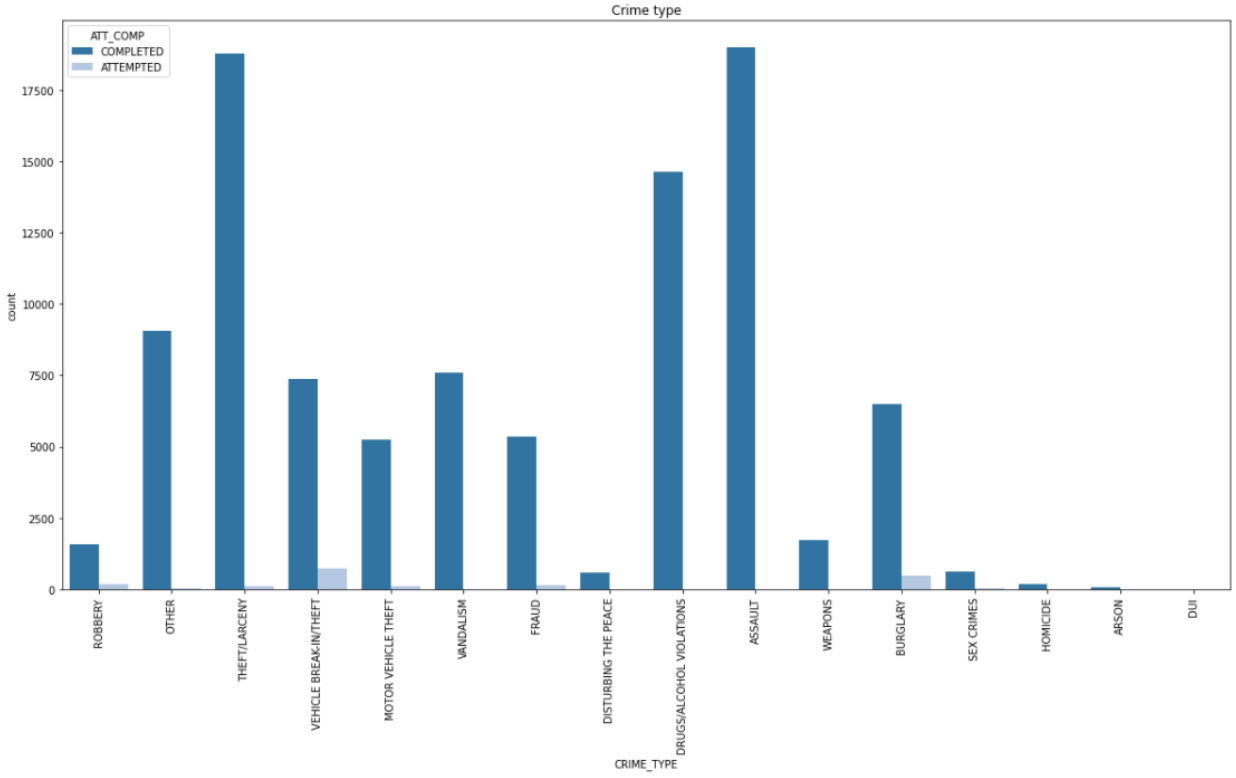
نمودار زیر نیز محل به نسبت تعداد جرایم را نشان می دهد. طبق نمودار بیشترین محلی که در آن جرم رخ داده است، RESIDENCE / HOME می باشد. HIGHWAY / ROAD / ALLEY نیز در جایگاه دوم قرار دارد. مشاهده می شود که میزان رخداد جرم در این دو محل با اختلاف زیادی از بقیه بیشتر است.



نمودار زیر تعداد جرایم نسبت به اینکه مجرم آن را به طور کامل انجام داده و یا قصد آن را داشته است نمایش می دهد. این نمودار به این معنی است که تعدادی از جرایم قبل آنکه به مرحله عمل برسند، تشخیص داده شده اند.



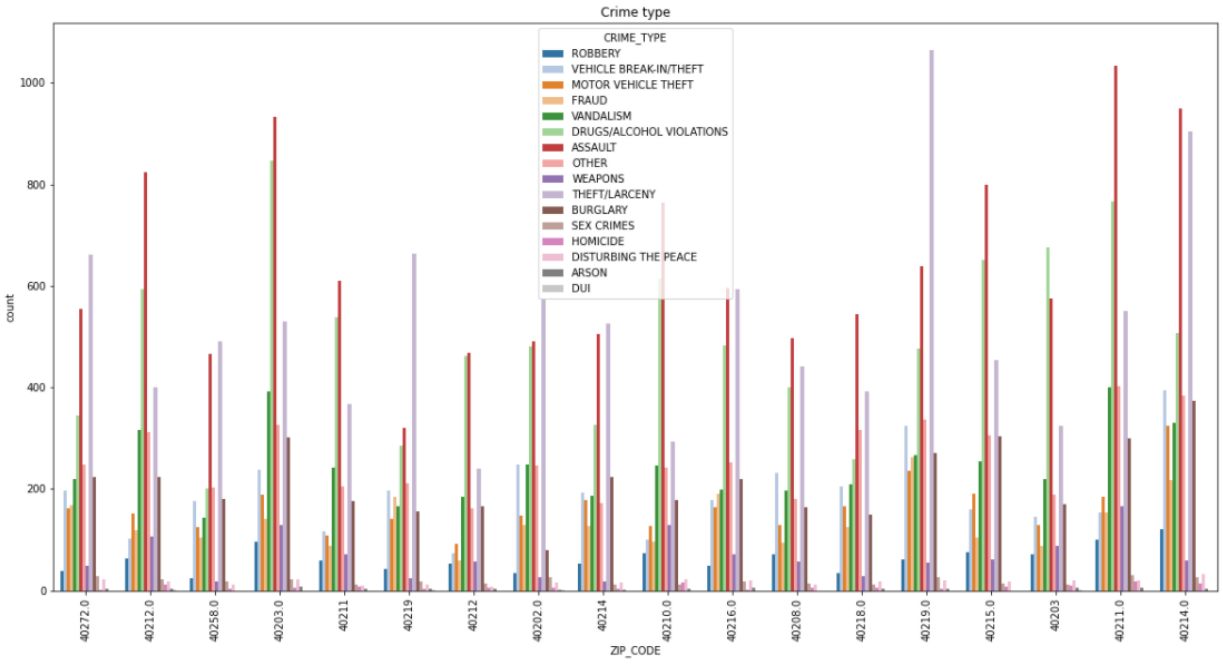
نمودار زیر نشان می دهد که چه مقدار از هر نوع جرم به طور کامل انجام شده و چه تعداد قرار بر انجام آن بوده است اما قبل از انجام جرم توسط مجرم شناسایی شده است.



در مرحله بعد جدولی برای ZIP_CODE هایی که بیشتریین تعداد جرم در آنها رخ میدهد رسم کردیم. برای این منظور از ZIP_CODE هایی که تعداد جرایم رخ داده در آنها بیش از ۲۰۰۰ است، استفاده کردیم.

	ZIP_CODE	count
0	40214	4638
1	40211	4284
2	40203	4181
3	40219	4045
4	40215	3397
5	40212	3267
6	40216	3040
7	40272	2921
8	40210	2918
9	40202	2755
10	40203	2721
11	40211	2612
12	40214	2539
13	40208	2494
14	40218	2468
15	40219	2427
16	40258	2163
17	40212	2048

نمودار زیر نیز پراکندگی هر نوع جرم در ZIP_CODE هایی که تعداد جرایم در آنها بیشتر بود را نمایش می دهد.



در آخرین مرحله نیز از تست chi square برای بررسی جرایم BURGLARY و ROBBERY نسبت به ماه رخداد آن جرم، استفاده کردیم. جدول تعداد جرم های صورت گرفته از هر یک دو نوع نسبت به ماه رخداد آن در زیر آمده است.

MONTH_occ	ROBBERY	BURGLARY
1	164	595
2	117	473
3	118	521
4	126	562
5	176	675
6	167	661
7	159	639
8	167	649
9	159	571
10	144	594
11	124	478
12	139	509

نتایج بدست آمده به صورت زیر است:

```
The statistical value: 5.892 & the p-value: 0.880 & the degree of freedom are: 11
```

با توجه به مقدار p-value فرض صفر ما تایید می شود که آن این است که رخ دادن این دو جرم به ماه های سال مرتبط نمی باشد.

درجه آزادی برابر ۱۱ است. مقدار آن از فرمول زیر بدست می آید:

(تعداد ستون - ۱)*(تعداد سطر - ۱)

۴. منابع

https://stats.stackexchange.com/questions/9573/t-test-for-non-normal-when-n50

https://www.pythonfordatascience.org