Crime prediction in Chicago

Final course of IBM Data Science Professional Certificate by Coursera

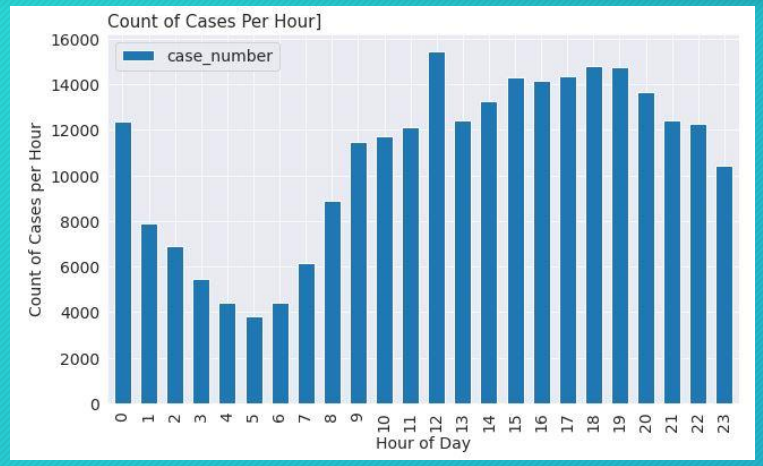
Real world case

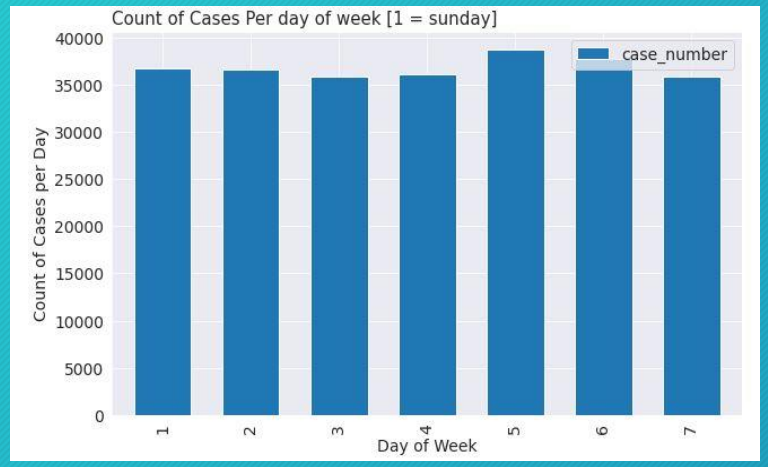
- In general, every traveler, whether for business or pleasure, intends to enjoy a little of the place and have fun. Inspired by a scenario in which some co-workers will go to Chicago on business, they will have a few more days to explore the city. Knowing them, I know they like to go out at night after work and drink some drikes and listen to good music. This project's main idea is to predict the potential for a crime to happen close to a nightclub search on foursquare. Let's go!

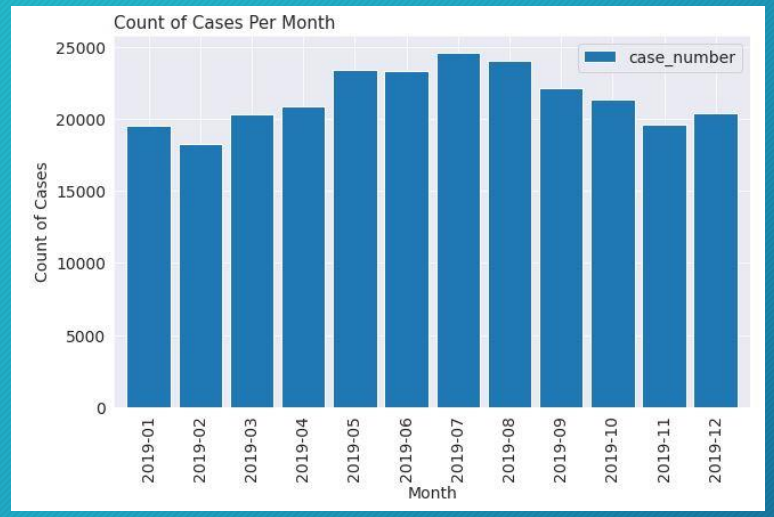
Data acquisition and cleaning

- The data were accessed at: https://data.cityofchicago.org/Public-Safety/Crimes-2019/w98m-zvie , where the API was generated to download in https://data.cityofchicago.org/api/views/w98mzvie/rows.csv?accessType=DOWNLOAD&api_foundry=true
- After some procedures for cleaning the data and creating new columns (which can be seen in detail in the attached notebook, remember, we are in the "Readme"), a new dataframe was created for the crimes that occurred in 2019.

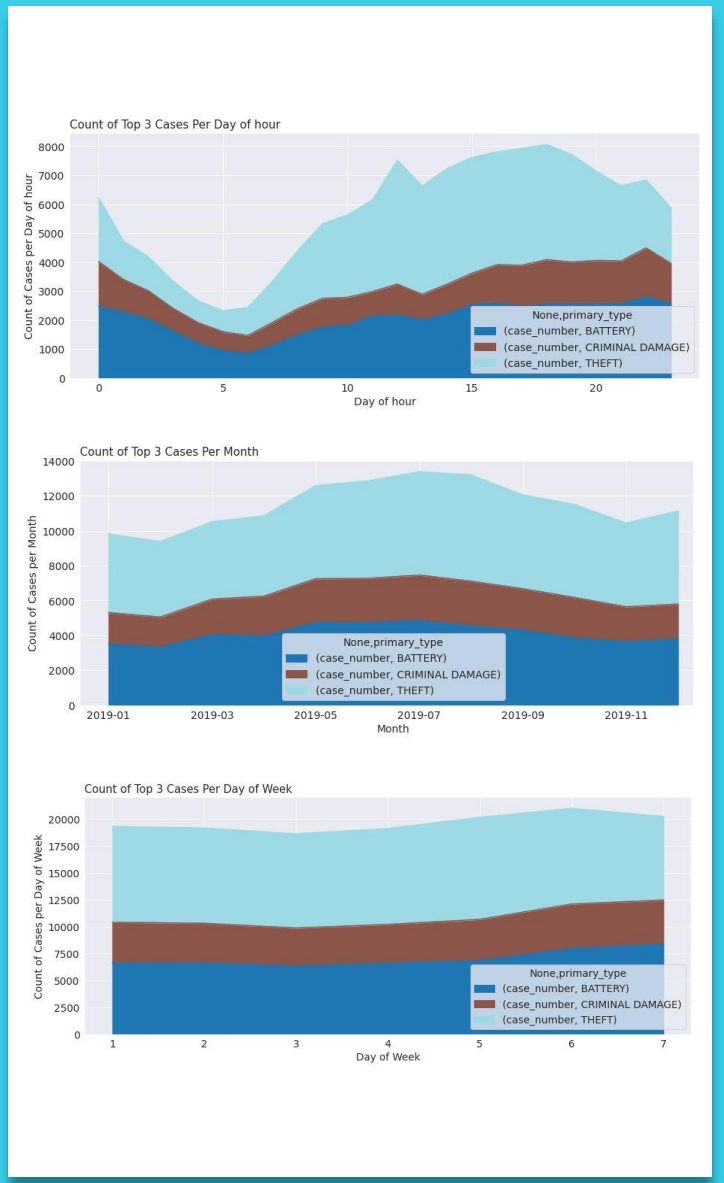
Descriptive statistics of Chicago crimes (Graphs)





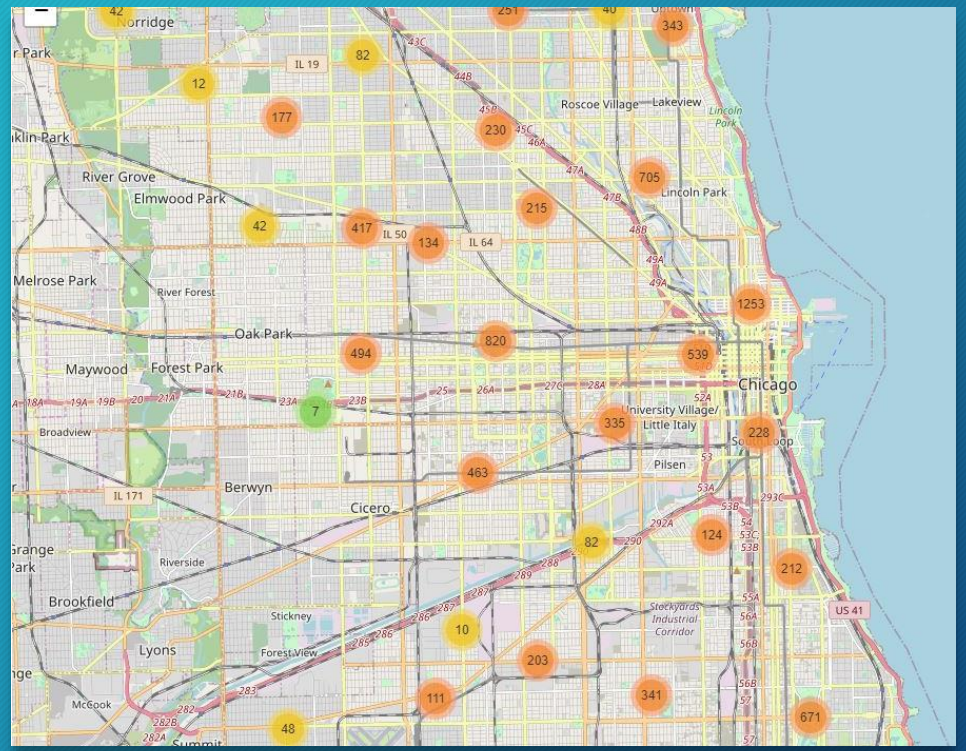


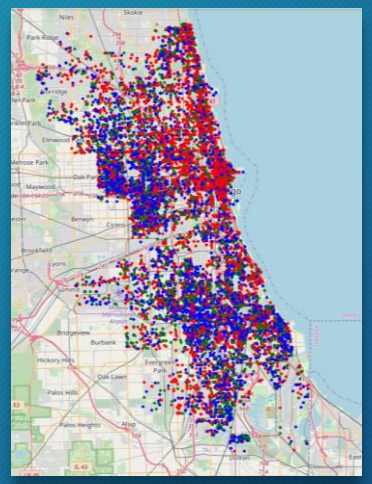
3 most commonly occurring ones

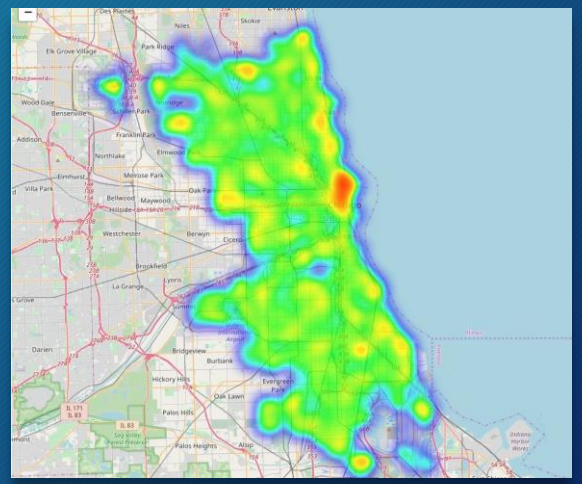


Chicago crimes (Maps)

Nightclubs and crimes

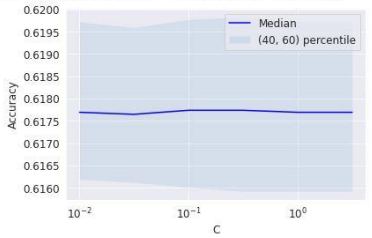






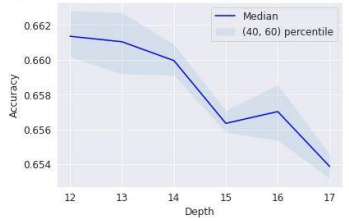
Models

```
C:  0.01   2020-04-14 11:24:01.366071
C:  0.03162277660168379   2020-04-14 11:24:10.169375
C:  0.1   2020-04-14 11:24:18.775032
C:  0.31622776601683794   2020-04-14 11:24:27.451008
C:  1.0   2020-04-14 11:24:36.063854
C:  3.1622776601683795   2020-04-14 11:24:44.727775
```



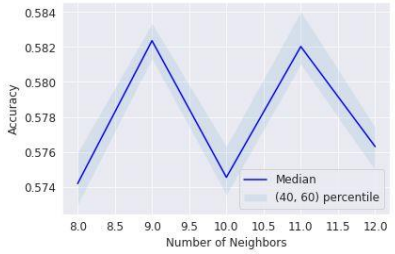

```
LR_model = LogisticRegression(C = 0.1, solver = 'liblinear').fit(X, y)
```

```
Depth:  12   2020-04-14 11:22:37.171828
Depth:  13   2020-04-14 11:22:49.587633
Depth:  14   2020-04-14 11:23:01.754843
Depth:  15   2020-04-14 11:23:13.878943
Depth:  16   2020-04-14 11:23:26.430570
Depth:  17   2020-04-14 11:23:39.539926
```



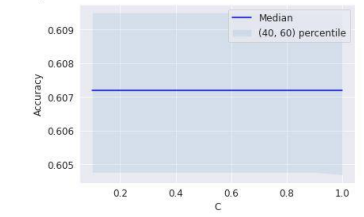

```
[ ]  Tree_model = DecisionTreeClassifier(criterion = "entropy", max_depth = 12).fit(X, y)
```

```
Heighbours:  8   2020-04-14 11:39:33.296261
Heighbours:  9   2020-04-14 11:58:57.334631
Heighbours:  10   2020-04-14 12:18:30.476200
Heighbours:  11   2020-04-14 12:37:24.502584
Heighbours:  12   2020-04-14 12:56:40.718368
```




```
[ ]  KNN_model = KNeighborsClassifier(n_neighbors = 9).fit(X, y)
```

```
Alpha:  0.1   2020-04-14 11:25:58.705666
Alpha:  0.2   2020-04-14 11:26:03.876223
Alpha:  0.30000000000000004   2020-04-14 11:26:09.050006
Alpha:  0.4   2020-04-14 11:26:14.183219
Alpha:  0.5   2020-04-14 11:26:19.371613
Alpha:  0.6   2020-04-14 11:26:24.533896
Alpha:  0.7000000000000001   2020-04-14 11:26:29.713495
Alpha:  0.8   2020-04-14 11:26:34.871969
Alpha:  0.9   2020-04-14 11:26:40.041951
Alpha:  1.0   2020-04-14 11:26:45.203379
```

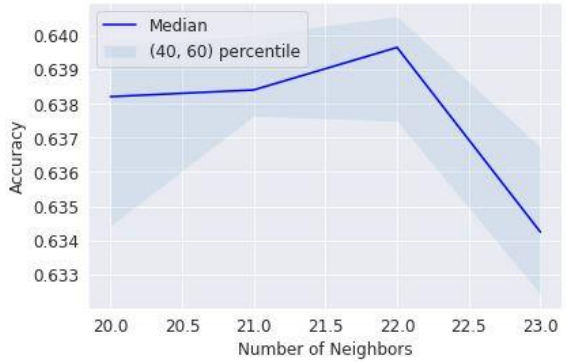



```
[ ]  NB_model = BernoulliNB(alpha=a).fit(X, y)
```

Best model

Decision Forest using a Random Forest

```
Estimator:  20   2020-04-14 11:26:58.111537
Estimator:  21   2020-04-14 11:27:36.607867
Estimator:  22   2020-04-14 11:28:17.126355
Estimator:  23   2020-04-14 11:28:59.199545
```

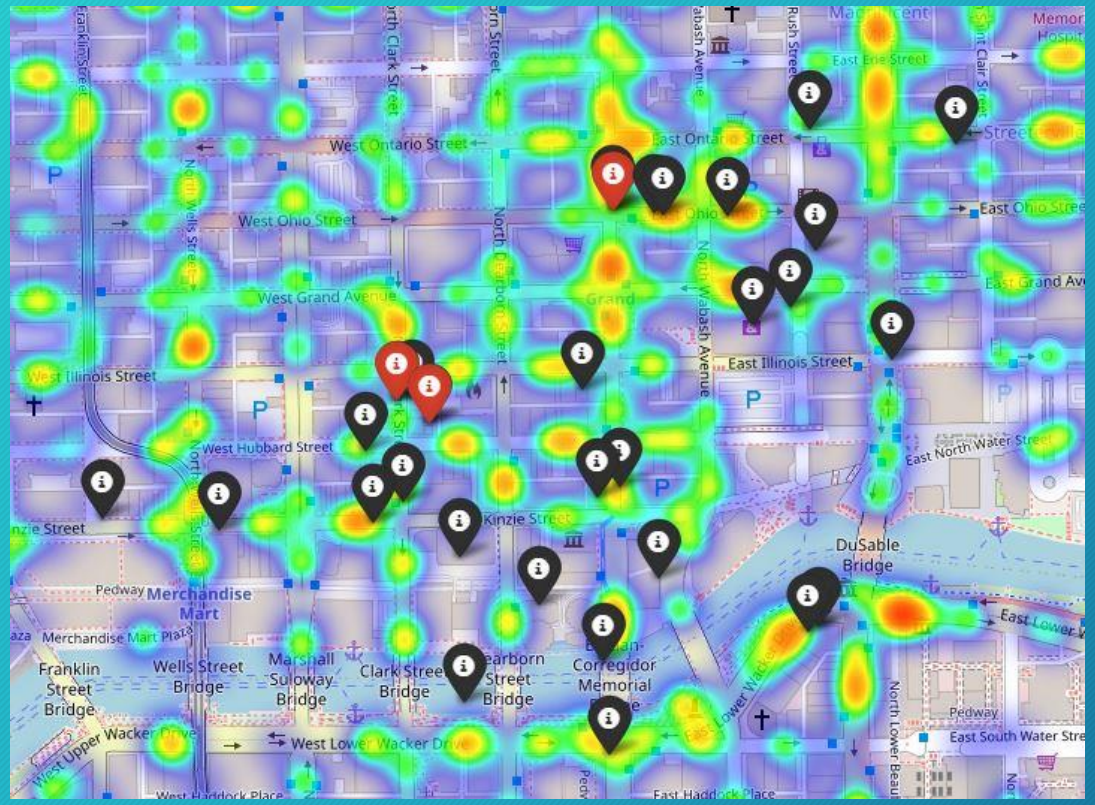



```
[ ]  Forest_model = RandomForestClassifier(n_estimators = 22, max_features = 'sqrt').fit(X, y)
```

Algorithm	Jaccard	F1-Score	LogLoss
KNN	0.699465	0.729000	10.380224
Decision Tree	0.695396	0.710063	10.520755
Bernoulli Naive Bayes	0.612360	0.664197	13.388800
Logistic Regression	0.621705	0.685819	13.066041
Random Forest	0.995606	0.996037	0.151749

Prediction

1 - Enolo Wine Cafe | 2 - Three Dots and Dash | 3 - Pops for Champagne



	name	date	latitude	longitude	prediction
0	Broken Shaker	2019-12-24 22:12:00	41.892488	-87.627336	0
1	London House Rooftop Bar	2019-05-28 22:17:00	41.888031	-87.625264	0
2	Gilt Bar	2019-02-07 20:15:00	41.889236	-87.635377	0
3	The Berkshire Room	2019-08-16 05:27:00	41.892495	-87.627476	0
4	Raised Rooftop at Renaissance Hotel	2019-01-06 19:43:00	41.886726	-87.628110	0
5	Bar Sótano	2019-12-07 04:00:00	41.890518	-87.630947	0
6	La Mez Agave Lounge	2019-04-17 14:28:00	41.889196	-87.631510	0
7	The Smith	2019-01-22 16:28:00	41.889417	-87.631070	0
8	Watershed	2019-02-13 20:08:00	41.892597	-87.628059	0
9	Pops for Champagne	2019-10-07 15:08:00	41.892530	-87.628038	1
10	Three Dots and a Dash	2019-07-05 10:29:00	41.890270	-87.630690	1
11	Lobby Bar	2019-11-28 11:40:00	41.892106	-87.625147	0
12	Flora Fauna	2019-01-25 13:26:00	41.890622	-87.628495	0
13	Side Door	2019-12-24 16:45:00	41.893398	-87.625235	0
14	Barrio	2019-02-14 01:50:00	41.888826	-87.630251	0
15	Rossi's Liquors	2019-08-26 17:16:00	41.889576	-87.627965	0
16	Celeste	2019-11-10 22:59:00	41.889958	-87.631607	0
17	Enolo Wine Cafe	2019-07-17 08:28:00	41.890501	-87.631149	1
18	Copper Fox Gastropub	2019-02-14 23:58:00	41.893238	-87.623130	0
19	Foundation Room	2019-10-04 14:19:00	41.888319	-87.629124	0
20	Highline Bar + Lounge	2019-09-13 18:18:00	41.889119	-87.633716	0
21	Travelle	2019-05-08 23:35:00	41.888598	-87.627413	0
22	Birreria	2019-05-12 19:46:00	41.892466	-87.626423	0
23	Tiny Tapp	2019-08-23 14:30:00	41.887271	-87.630189	0
24	ENO	2019-08-23 23:08:00	41.890937	-87.624066	0
25	On 21	2019-12-05 14:54:00	41.888088	-87.625118	0
26	Public House	2019-05-04 04:41:00	41.889474	-87.628274	0
27	Habitant - Nordstrom Michigan Avenue	2019-06-07 19:20:00	41.891308	-87.626047	0
28	Upstairs at The Gwen	2019-10-20 19:05:00	41.891489	-87.625514	0
29	Wollensky's Grill	2019-11-22 21:08:00	41.887703	-87.628191	0

Conclusion and Discussion

First of all, I have to mention the growth of my skills after all the courses of the IBM Data Science Professional Certificate, all the challenges proposed were of great value for the accomplishment of this project.

The objectives were achieved despite some problems midway through. The basic plan of the Foursquare API for Developers has a very short limit of consultations per day and until I get used to it, I exceeded that limit several times, making me have to wait for the next day and later having to pass the data to my personal drive and introducing it again.

Chicago is to be congratulated for providing open data with this quality and with an extremely interactive and simple platform, even having an area for developers.

Today (2020), we are experiencing the Coronavirus pandemic and I have observed many platforms developed for this cause, with humanitarian aid from data scientists to carry out projects on top of the data.

With the example of Chicago and the available data on Cornonavirus ... Shouldn't they look more at the importance of data and its availability?