Chicago Crime Data Analysis

Introduction and background of the Data:

In this world, crimes are an inseparable part of our lives. Every day we hear about them and some of us are even involved in at least one of them during our life. Being cautious and improve safety is not a simple instruction anymore. We need to use modern technology and data science techniques to more wisely act against this problem. There are so many records and documentation in the police department that have been gathered during the years, which can be used as a valuable source of data for the data analytics tasks. Applying analytical task to these data bring us valuable information that can be used to increase the safety of our society and lower the crime rate.

Objectives of Study (Aim):

The main idea behind this project is to create a user story of the crimes Dataset, which involved geographical analysis, Crime Data analysis and the use of machine learning models on the Chicago crime dataset. Analyzing and examining of crimes gives an understanding of crime regions and can used to take the precautions to reduce the crime rates.By identifying the patterns will allow us to tackle problems. My approach involves the prediction crimes and visualization of patterns. Using of past data could help us to correlate factors which might help to understand the possibility of happening of the particular crime.

About Data:

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system.

- Dataset consists of 22 columns
- Consists of 6.3 million of rows
- Size of the data set 1.4 GB

Data Wrangling:

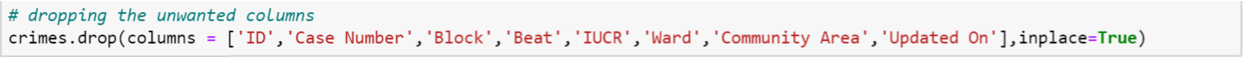
In this project, I will be dealing with the data set containing of all the crimes that are reported by the police in their directory from 2001-present.

This dataset is available in the Chicago city data repository. This dataset consisting of 7 million row of data by the columns of ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordinate, Year, Updated On, Latitude, Longitude, Location

On this data first objective was to remove the unwanted data, columns containing the unwanted data or un relevant information like ID, Case Number, Block, IUCR, Beat, Updated On and location were removed. The second objective was to convert some of the object type columns to integers. The questions we are associated to different types of crimes and how they are located.

After preprocessing, the data frame consists of following information with our dropping null values

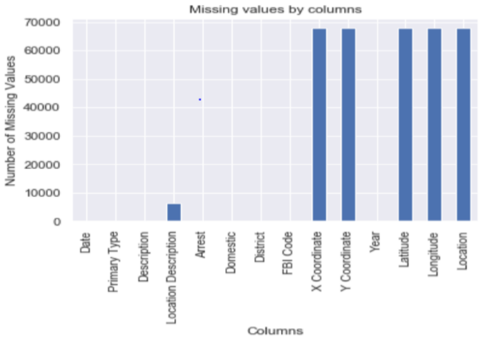
Dropped the following columns from the main Data Frame called crimes



```
# dropping the unwanted columns
crimes.drop(columns = ['ID','Case Number','Block','Beat','IUCR','Ward','Community Area','Updated On'],inplace=True)
```

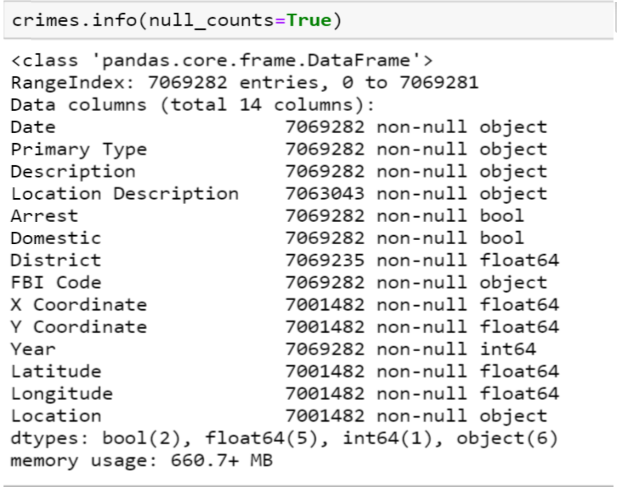
Difference between the columns containing null values and without null values (which is after dropping all null values)

Inspecting the features, we see that all the features that have a large count of missing values are features that relate to the geographical location of the crime scene. This is No Surprise as the Chicago Crime Dataset is based on firsthand accounts of people involved in or around the crime. It is not necessary that such firsthand reports need to contain the specific locations of the crime. We have 3,45,286 missing values in the whole dataset that are present in Location Description, Community, X Co-ordinate, Y Co-ordinate, Latitude, Longitude and Location.

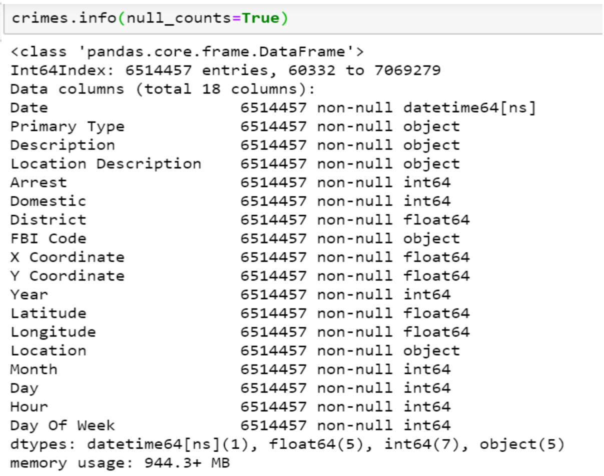


Since, these features are not direct numeric values, we can't use summary statistical functions to fill in the missing values. Hence I thought to remove all values containing null using dropna().92.5 percent of data is retained after dropping null values.

Below figure shows the comparison of no-null values counts before and after dropping null values

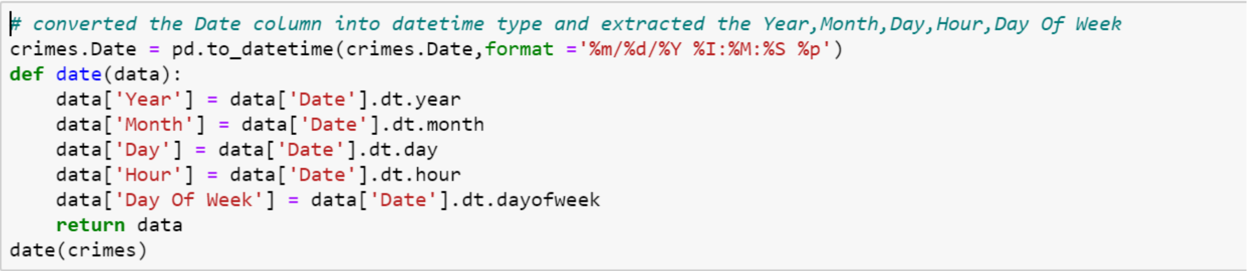


```
crimes.info(null_counts=True)

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7069282 entries, 0 to 7069281
Data columns (total 14 columns):
Date                    7069282 non-null object
Primary Type            7069282 non-null object
Description             7069282 non-null object
Location Description    7063043 non-null object
Arrest                  7069282 non-null bool
Domestic                7069282 non-null bool
District                7069235 non-null float64
FBI Code                7069282 non-null object
X Coordinate            7001482 non-null float64
Y Coordinate            7001482 non-null float64
Year                    7069282 non-null int64
Latitude                7001482 non-null float64
Longitude               7001482 non-null float64
Location                7001482 non-null object
dtypes: bool(2), float64(5), int64(1), object(6)
memory usage: 660.7+ MB
```



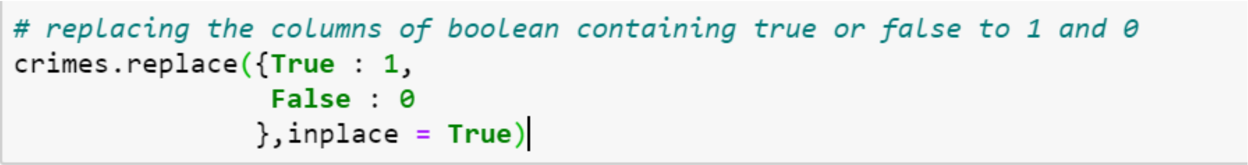
```
crimes.info(null_counts=True)

<class 'pandas.core.frame.DataFrame'>
Int64Index: 6514457 entries, 60332 to 7069279
Data columns (total 18 columns):
Date                    6514457 non-null datetime64[ns]
Primary Type            6514457 non-null object
Description             6514457 non-null object
Location Description    6514457 non-null object
Arrest                  6514457 non-null int64
Domestic                6514457 non-null int64
District                6514457 non-null float64
FBI Code                6514457 non-null object
X Coordinate            6514457 non-null float64
Y Coordinate            6514457 non-null float64
Year                    6514457 non-null int64
Latitude                6514457 non-null float64
Longitude               6514457 non-null float64
Location                6514457 non-null object
Month                   6514457 non-null int64
Day                     6514457 non-null int64
Hour                    6514457 non-null int64
Day Of Week             6514457 non-null int64
dtypes: datetime64[ns](1), float64(5), int64(7), object(5)
memory usage: 944.3+ MB
```

I have created 3 additional columns Day, Hour, Day Of Week by converting the Date column to the datetime type and then slicing the required column attributed for the Date. Below figure gives the code that I used for slicing. This will be helpful for me to identify the crime pattern in different sections of day



```
# converted the Date column into datetime type and extracted the Year,Month,Day,Hour,Day Of Week
crimes.Date = pd.to_datetime(crimes.Date,format ='%m/%d/%Y %I:%M:%S %p')
def date(data):
    data['Year'] = data['Date'].dt.year
    data['Month'] = data['Date'].dt.month
    data['Day'] = data['Date'].dt.day
    data['Hour'] = data['Date'].dt.hour
    data['Day Of Week'] = data['Date'].dt.dayofweek
    return data
date(crimes)
```

The columns [‘Arrest’, ‘Domestic’] contains the values in terms Boolean values, I have converted to their respective 1’s and 0’s in the column places using the following code



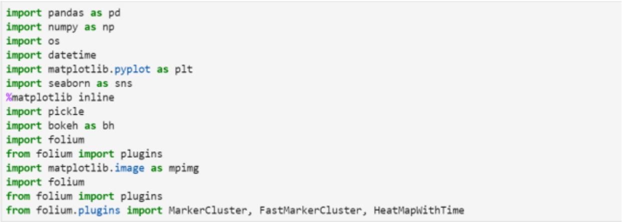
```
# replacing the columns of boolean containing true or false to 1 and 0
crimes.replace({True : 1,
                False : 0
               },inplace = True)
```

All the changes done to the Data Frame is saved in the form of CSV file and PICKLE files.

Chicago crime data analysis EDA

Imported Packages:

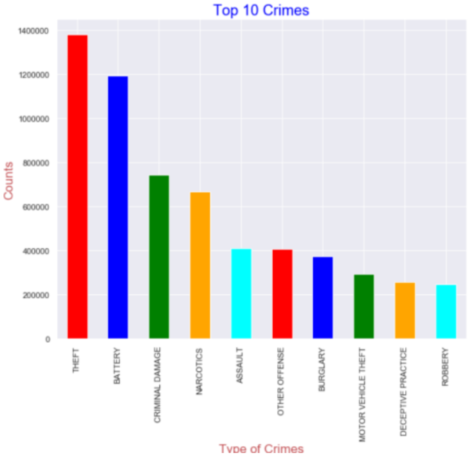
Here is the list of packages that are Imported to do my Exploratory Data Analysis. I have used Matplotlib, Seaborn for my visualization of data and Folium for the interactive maps

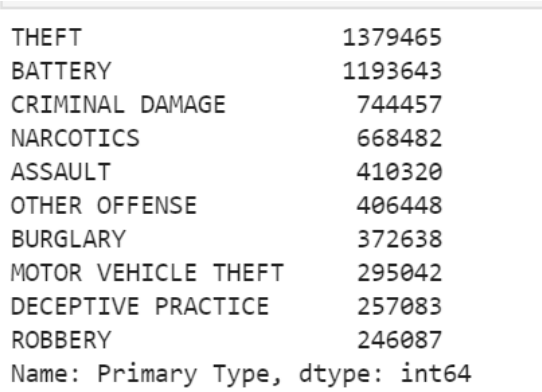


```
import pandas as pd
import numpy as np
import os
import datetime
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import pickle
import bokeh as bh
import folium
from folium import plugins
import matplotlib.image as mpimg
import folium
from folium import plugins
from folium.plugins import MarkerCluster, FastMarkerCluster, HeatMapWithTime
```

1.All About Crimes in Chicago:

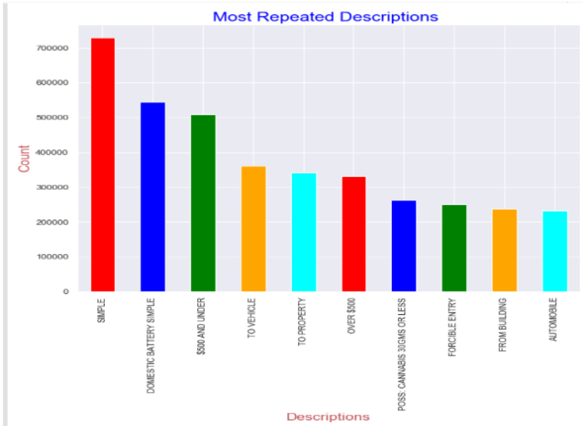
By seeing the data, I was very curious to find the topmost crimes in the Chicago. Below figure represents the top 10 crimes in the Chicago. Out of the top 10 crimes, Theft was the most occurring crimes with an count of 1379465.Higher counts of Battery and criminal Damage indicates the presence of physically violent community.

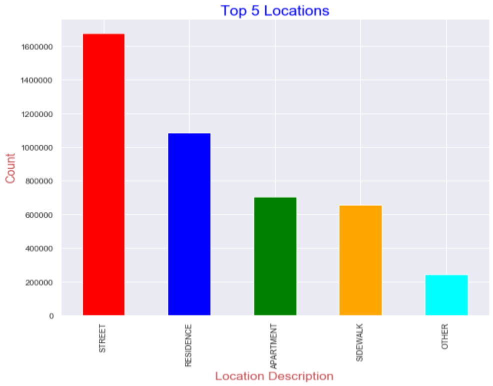




```
THEFT                  1379465
BATTERY                1193643
CRIMINAL DAMAGE         744457
NARCOTICS               668482
ASSAULT                 410320
OTHER OFFENSE           406448
BURGLARY                372638
MOTOR VEHICLE THEFT     295042
DECEPTIVE PRACTICE      257083
ROBBERY                 246087
Name: Primary Type, dtype: int64
```

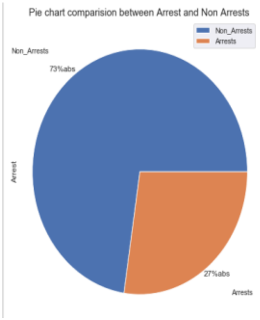
The figure represents the Descriptions that are reported along with the crime. Majority of the crimes are reported as Simple and the battery crimes in the domestic reported as simple, some were under 500$ which could be Theft or Robbery or Burglary. Most of the crimes are at street level and at house level

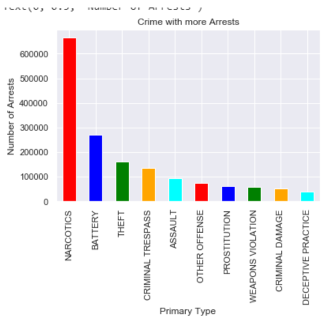


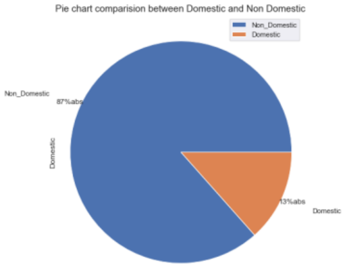


2.Arrests in city of Chicago:

73% of the crimes see No Arrests, means only 23 percent of the crimes has been solved from 2001 to present. As we see in the bar chart which represents the crimes with a greater number of arrests, Arrests related to Narcotics are more. 99%of the Narcotics cases we arrested (total narcotics cases were 67000(approximately). As we overall, they are 1379465 theft cases but only 18000 cases are solved, May be police were not taking serious to cases which are involved with the theft and Battery. 87 percent of the crimes are related to Non-Domestic areas.

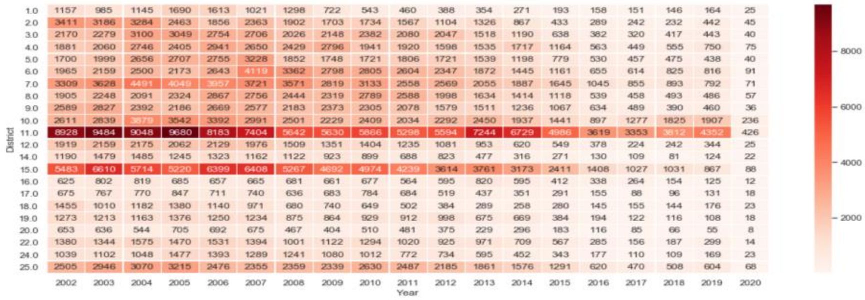






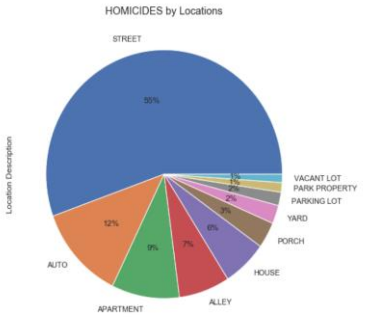
Let's how the Narcotics arrests distributed across the districts.District 11 is more involved with the crimes.There is down trend of narcotics arrests over the years except some hikes in the year 2013-2014

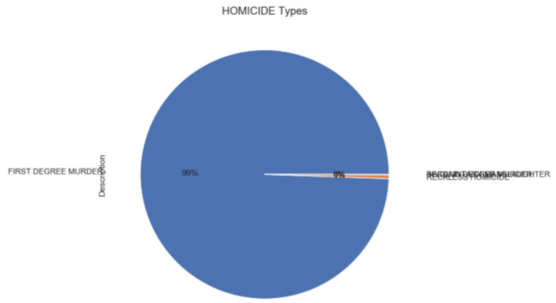
```
<matplotlib.axes._subplots.AxesSubplot at 0x231e226bec8>
```



District	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
1.0	1157	985	1145	1690	1613	1021	1298	722	543	460	388	354	271	193	158	151	146	164	25
2.0	3411	3186	3284	2463	1856	2363	1902	1703	1734	1567	1104	1326	867	433	289	242	232	442	45
3.0	2170	2279	3100	3049	2754	2706	2026	2148	2382	2080	2047	1518	1190	638	382	320	417	443	40
4.0	1881	2060	2746	2405	2941	2650	2429	2796	1941	1920	1598	1535	1717	1164	563	449	555	750	75
5.0	1700	1999	2656	2707	2755	3228	1852	1748	1721	1806	1721	1539	1198	779	530	457	475	438	40
6.0	1965	2159	2500	2173	2643	4119	3362	2798	2805	2604	2347	1872	1445	1161	655	614	825	816	91
7.0	3309	3628	4491	4049	3957	3721	3571	2819	3133	2558	2569	2055	1887	1645	1045	855	893	792	71
8.0	1905	2248	2091	2324	2867	2756	2444	2319	2789	2588	1998	1634	1414	1118	539	458	493	486	57
9.0	2589	2827	2392	2186	2669	2577	2183	2373	2305	2078	1579	1511	1236	1067	634	489	390	460	36
10.0	2611	2839	3879	3542	3392	2991	2501	2229	2409	2034	2292	2450	1937	1441	897	1277	1825	1907	236
11.0	8928	9484	9048	9680	8183	7404	5642	5630	5866	5298	5594	7244	6729	4986	3619	3353	3812	4352	426
12.0	1919	2159	2175	2062	2129	1976	1509	1351	1404	1235	1081	953	620	549	378	224	242	344	25
14.0	1190	1479	1485	1245	1323	1162	1122	923	899	688	823	477	316	271	130	109	81	124	22
15.0	5483	6610	5714	5220	6399	6408	5267	4692	4974	4239	3614	3761	3173	2411	1408	1027	1031	867	88
16.0	625	802	819	685	657	665	681	661	677	564	595	820	595	412	338	264	154	125	12
17.0	675	767	770	847	711	740	636	683	784	684	519	437	351	291	155	88	96	131	18
18.0	1455	1010	1182	1380	1140	971	680	740	649	502	384	289	258	280	145	155	144	176	23
19.0	1273	1213	1163	1376	1250	1234	875	864	929	912	998	675	669	384	194	122	116	108	18
20.0	653	636	544	705	692	675	467	404	510	481	375	229	296	183	116	85	66	55	8
22.0	1380	1344	1575	1470	1531	1394	1001	1122	1294	1020	925	971	709	567	285	156	187	299	14
24.0	1039	1102	1048	1477	1393	1289	1241	1080	1012	772	734	595	452	343	177	110	109	169	23
25.0	2505	2946	3070	3215	2476	2355	2359	2339	2630	2487	2185	1861	1576	1291	620	470	508	604	68

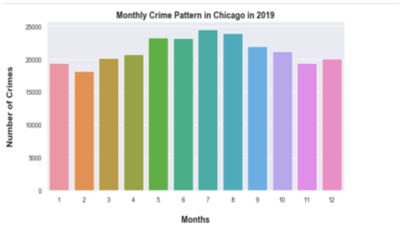
Out of 9421 cases of Homicides,99 percent of the cases are involved with First Degree Murder.56% of homicides are occurring in the streets and followed by Auto, Apartment which contributes 10 percent cases each.

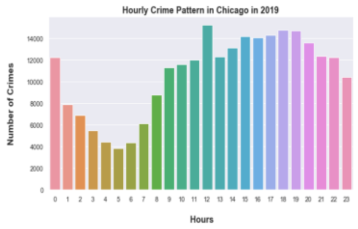


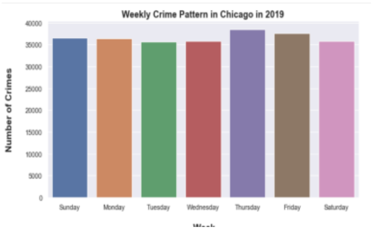


3. Crime vs Time:

Here is the Monthly, Hourly, Weekly crime patterns of the crimes in 2019.We see more crime in the summer months, may more people coming out for vacations that might lead to increase in the crimes count. Crime rate increases with the day, needs some rest to crime, less crimes rates between 2 am to 7 pm. coming to weekly part all the days reports the same number of crimes but Thursday and Friday reports slightly greater than remaining days

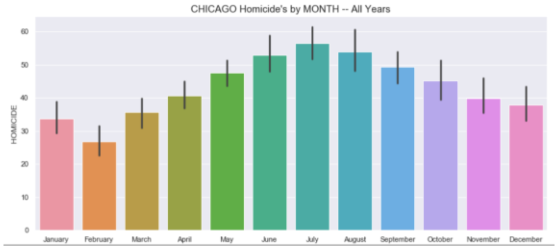


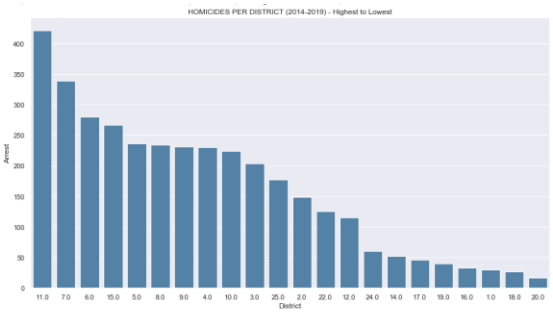




Homicides crime rates vs years

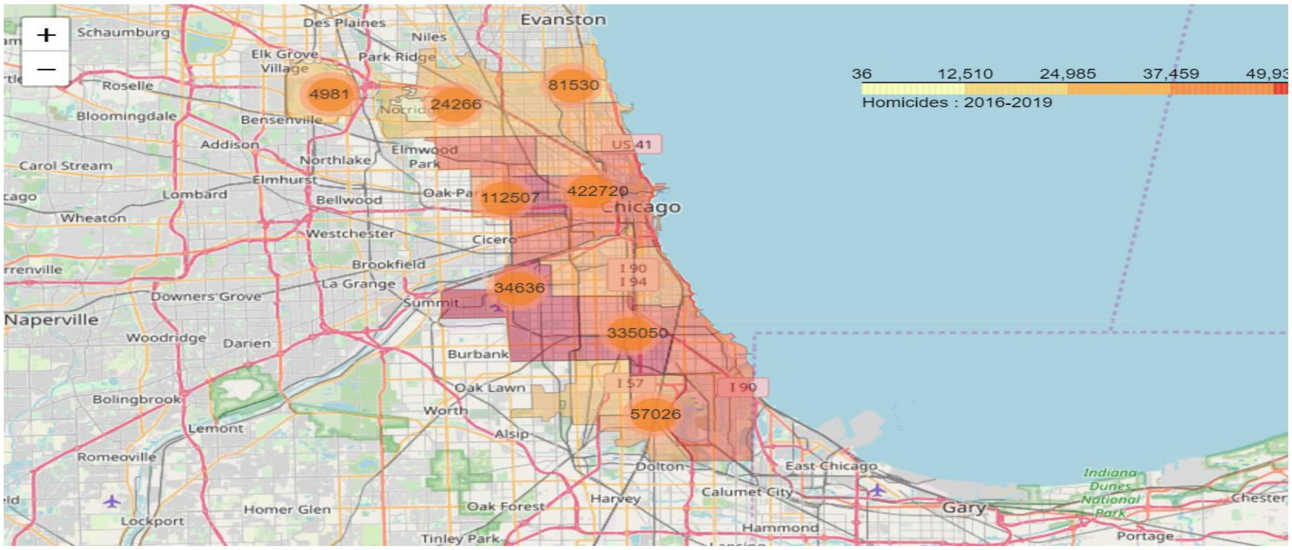
Most of the homicides were happening in the night between 5pm to morning 5am, records very low homicide rate between 7 to 11 am. Districts 11,7,6,15 has more arrests towards homicides compared to remaining districts.



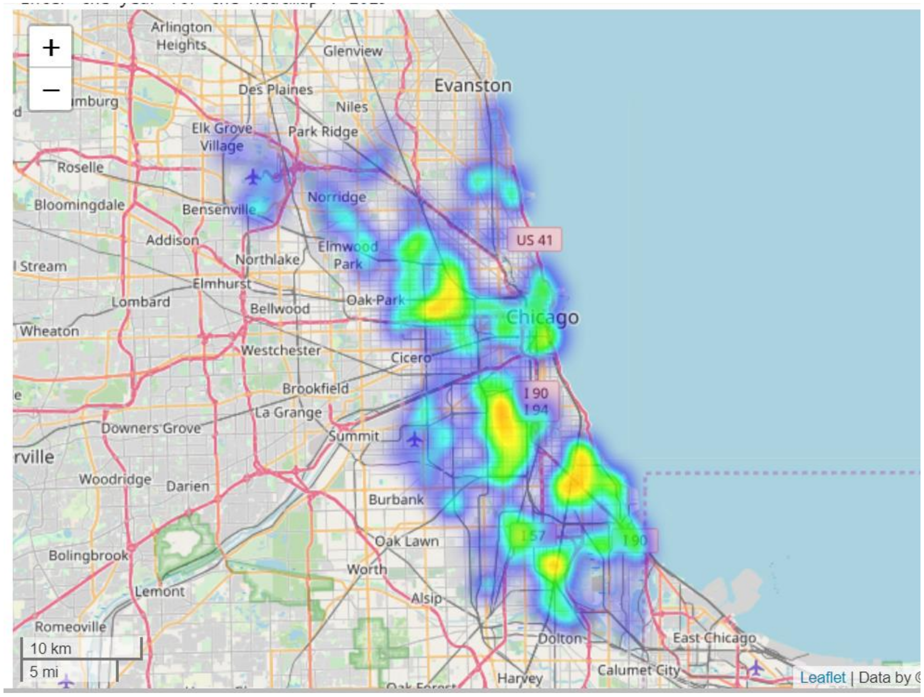


4.Crimes with Maps

Below chart shows the arrest count per police district



Heat map of all crimes for the year 2019



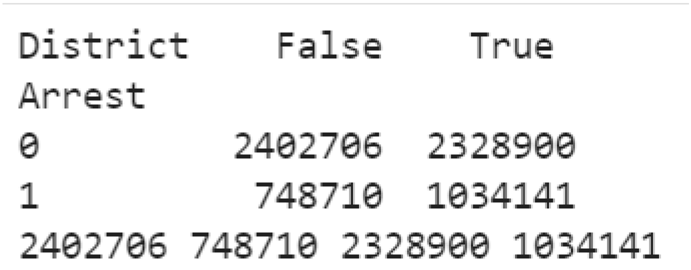
Inferential Statistics

Is there any association between the arrest and Districts : Inorder to find the relation between the districts, divide the districts which have more blacks population and the districts with other populations which includes white,asian and etc.The districts 15.0,11.0,10.0,21.0,2.0,7.0,9.0,3.0,6.0,4.0,5.0 majority blacks populated districts,where as rest of the districts are other populations.Applied hypothesis testing inorder find the whether there is an association between the Arrest in black majority populated districts and other.

Hypothesis Testing :

H0 : There is no Association with arrests between the Majority blacks districts and others.

H1 : There is an association with arrest between the districts.



```
District    False    True
Arrest
0         2402706  2328900
1          748710  1034141
2402706 748710 2328900 1034141
```

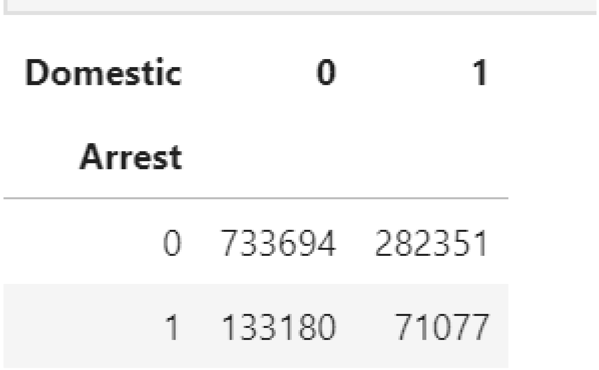
True represents the districts with majority black's districts

1 indicates the arrests Counts and 0 represents the Non arrests Counts

By using the chi-squared test, p value = 0.000. Since the P value < 0.05, hence we reject the null hypothesis. Arrest has an association between the districts with more black's and without.

- These districts 2,3,4,5,6,7,9,10,11,15,21 have more arrests compared to other districts.
- There might be racial disparities among the blacks than others.
- Police may be more interested in arresting the blacks.

Is there any association between the Arrest and Domestic majority populated Black's:



Domestic	0	1
Arrest		
0	733694	282351
1	133180	71077

Count of arrests are more in the more in the Non Domestic areas