CRIME PREVENTION USING DATA ANALYTICS AND MACHINE LEARNING

INTRODUCTION

Crime is one of the biggest and dominating problem in our society and its prevention is an important task. Daily there are huge numbers of crimes committed frequently. This require keeping track of all the crimes and maintaining a database for same which may be used for future reference. Crimes are of different type –robbery, murder, rape, assault, battery, false imprisonment, kidnapping, homicide. Since crimes are increasing there is a need to solve the cases in a much faster way. The crime activities have been increased at a faster rate and it is the responsibility of police department to control and reduce the crime activities. Crime prediction and criminal identification are the major problems to the police department as there are tremendous amount of crime data that exist. There is a need of technology through which the case solving could be faster.

Chicago has been in the news a lot in the last few years especially due to the recent up surge of murders resulting from gun violence. The objective of this project is to analyze dataset which consist of numerous crimes and predicting the type of crime which may happen in future depending upon various conditions. In this project, we will be using the technique of machine learning and data science for crime prediction of Chicago crime data set. The crime data is extracted from the official portal of Chicago police. It consists of crime information like location description, type of crime, date, time, latitude, longitude. Before training of the model data we shall be performing Data Analysis using Pandas and Visualizations libraries to understand tends and patterns of the crimes which are been committed. We shall then use Foursquare APIs to come up with more insights about the locations in which various crimes are getting committed.

This project will be useful for two types of Stakeholders:

1. Chicago Police Departments – Chicago PD will have a more clear picture about the trend of the crimes that are been perpetrated in the city. They might also get some more insights on the type of venues close to districts in which mostly crimes take place. Like Narcotics related crimes can tend to happen in a district which has more open spaces like parks.
2. Real Estate Buyers – For anyone who is planning to purchase a real estate, they would surely like to view the insights of the solution. This would give them a deep insight on the districts and the locations which are more prone to crimes

been committed. Surely a person planning to open a Jewelry store would not like to open one in a neighborhood where lot of theft occurs. Also they can check out the common venues in a location, which will help them in understanding the competition of the market which they are entering into.

Dataset

In this project, we will be using the technique of machine learning and data science for crime prediction of Chicago crime data set. The crime data is extracted from the official portal of Chicago police. The data is publicly available data from the City of Chicago's data portal linked below to explore crime in Chicago from January 2001 to February 2018

Data Sources:

- https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2
- https://data.cityofchicago.org/Public-Safety/Boundaries-Police-Beats-current-/aerh-rz74

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.

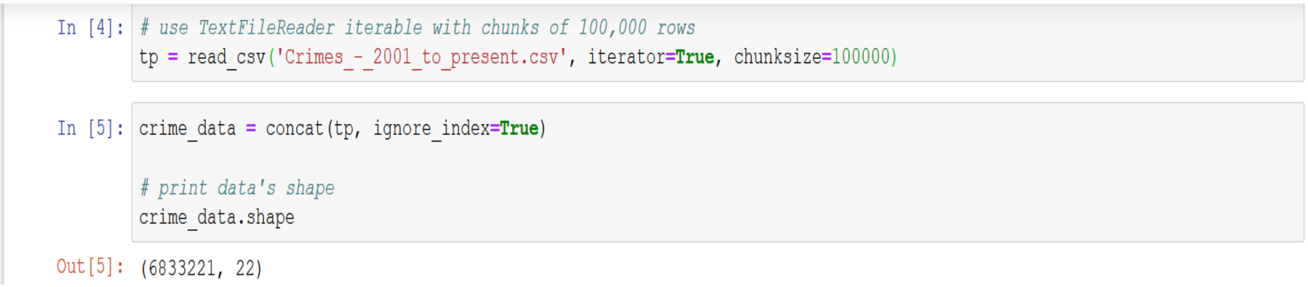
Columns in this Dataset:

Column Name	Description	Type
ID	Unique identifier for the record.	Number
Case Number	The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.	Plain Text
Date	Date when the incident occurred. this is sometimes a best estimate.	Date & Time
Block	The partially redacted address where the incident occurred, placing it on the same block as the actual address.	Plain Text
IUCR	The Illinois Unifrom Crime Reporting code. This is directly linked to the Primary Type and Description. See the list of IUCR codes at https://data.cityofchicago.org/d/c7ck-438e.	Plain Text

Column Name	Description	Type
Primary Type	The primary description of the IUCR code.	Plain Text
Description	The secondary description of the IUCR code, a subcategory of the primary description.	Plain Text
Location Description	Description of the location where the incident occurred.	Plain Text
Arrest	Indicates whether an arrest was made.	Checkbox
Domestic	Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.	Checkbox
Beat	Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts. See the beats at https://data.cityofchicago.org/d/aerh-rz74.	Plain Text
District	Indicates the police district where the incident occurred. See the districts at https://data.cityofchicago.org/d/fthy-xz3r.	Plain Text
Ward	The ward (City Council district) where the incident occurred. See the wards at https://data.cityofchicago.org/d/sp34-6z76.	Number
Community Area	Indicates the community area where the incident occurred. Chicago has 77 community areas. See the community areas at https://data.cityofchicago.org/d/cauq-8yn6.	Plain Text
FBI Code	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS). See the Chicago Police Department listing of these classifications at http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html.	Plain Text

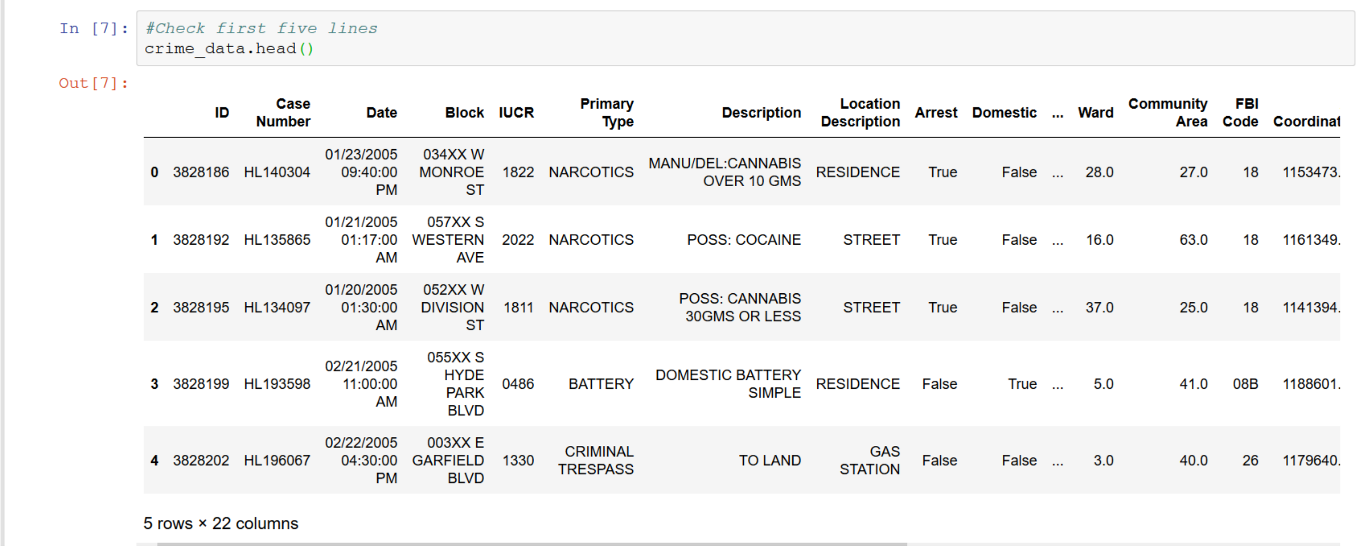
Column Name	Description	Type
X Coordinate	The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.	Number
Y Coordinate	The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.	Number
Year	Year the incident occurred.	Number
Updated On	Date and time the record was last updated.	Date & Time
Latitude	The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.	Number
Longitude	The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.	Number
Location	The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.	Location

The original dataset from the city data portal was in .CSV format but was too large (at 1.5 GB) for my laptop’s resources. I could not fit the whole file in memory so I used pandas' *TextFileReader* function which allowed me to load the large file in chunks of 100,000 rows and then concatenate the chunks back together into a new data frame.



```
In [4]: # use TextFileReader iterable with chunks of 100,000 rows
        tp = read_csv('Crimes_-_2001_to_present.csv', iterator=True, chunksize=100000)

In [5]: crime_data = concat(tp, ignore_index=True)

        # print data's shape
        crime_data.shape

Out[5]: (6833221, 22)
```

Here is a snippet of the first few rows of the dataset:



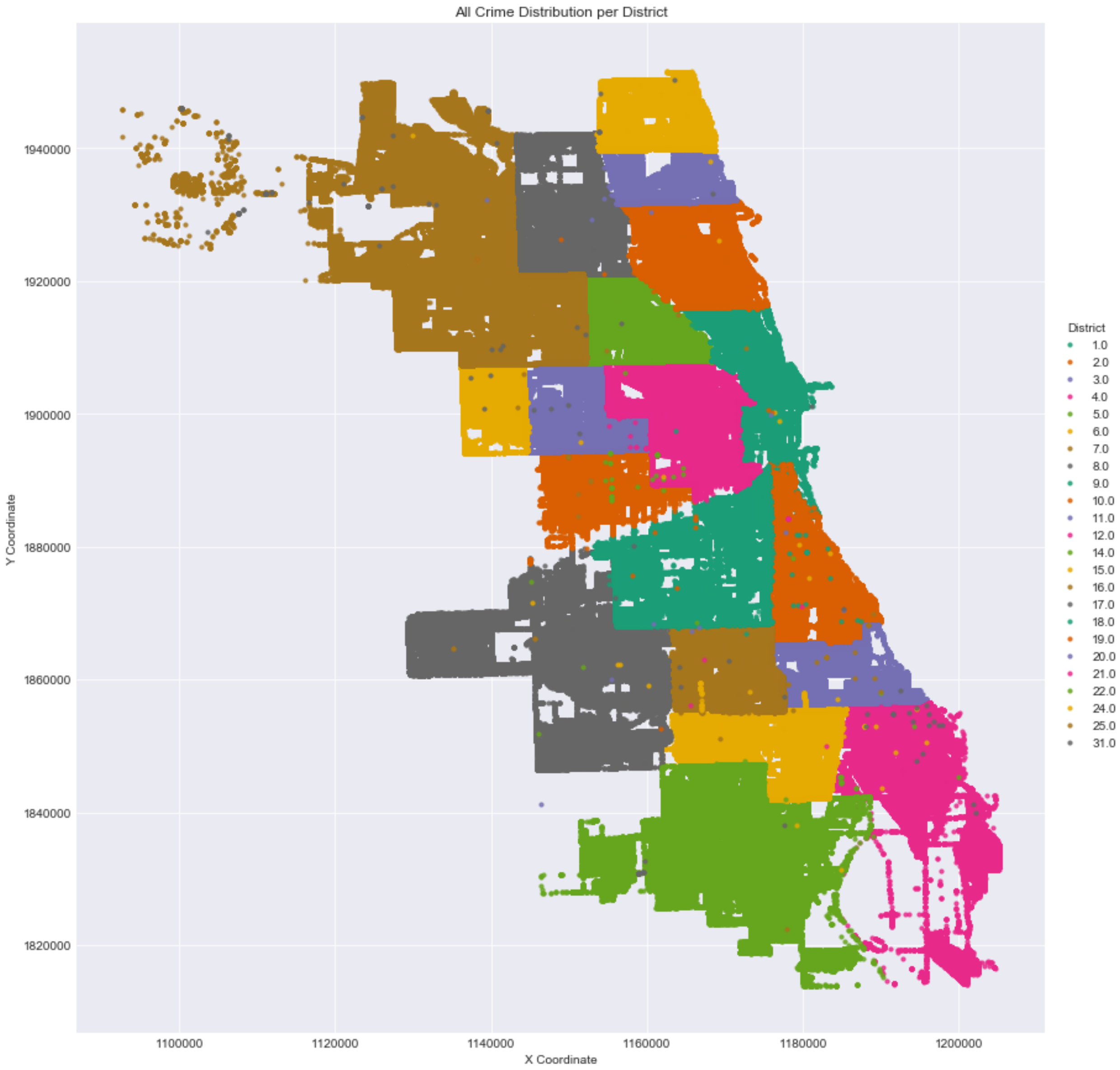
```
In [7]: #Check first five lines
        crime_data.head()

Out[7]:
```

	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	...	Ward	Community Area	FBI Code	Coordinat
0	3828186	HL140304	01/23/2005 09:40:00 PM	034XX W MONROE ST	1822	NARCOTICS	MANU/DEL:CANNABIS OVER 10 GMS	RESIDENCE	True	False	...	28.0	27.0	18	1153473.
1	3828192	HL135865	01/21/2005 01:17:00 AM	057XX S WESTERN AVE	2022	NARCOTICS	POSS: COCAINE	STREET	True	False	...	16.0	63.0	18	1161349.
2	3828195	HL134097	01/20/2005 01:30:00 AM	052XX W DIVISION ST	1811	NARCOTICS	POSS: CANNABIS 30GMS OR LESS	STREET	True	False	...	37.0	25.0	18	1141394.
3	3828199	HL193598	02/21/2005 11:00:00 AM	055XX S HYDE PARK BLVD	0486	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE	False	True	...	5.0	41.0	08B	1188601.
4	3828202	HL196067	02/22/2005 04:30:00 PM	003XX E GARFIELD BLVD	1330	CRIMINAL TRESPASS	TO LAND	GAS STATION	False	False	...	3.0	40.0	26	1179640.

5 rows × 22 columns

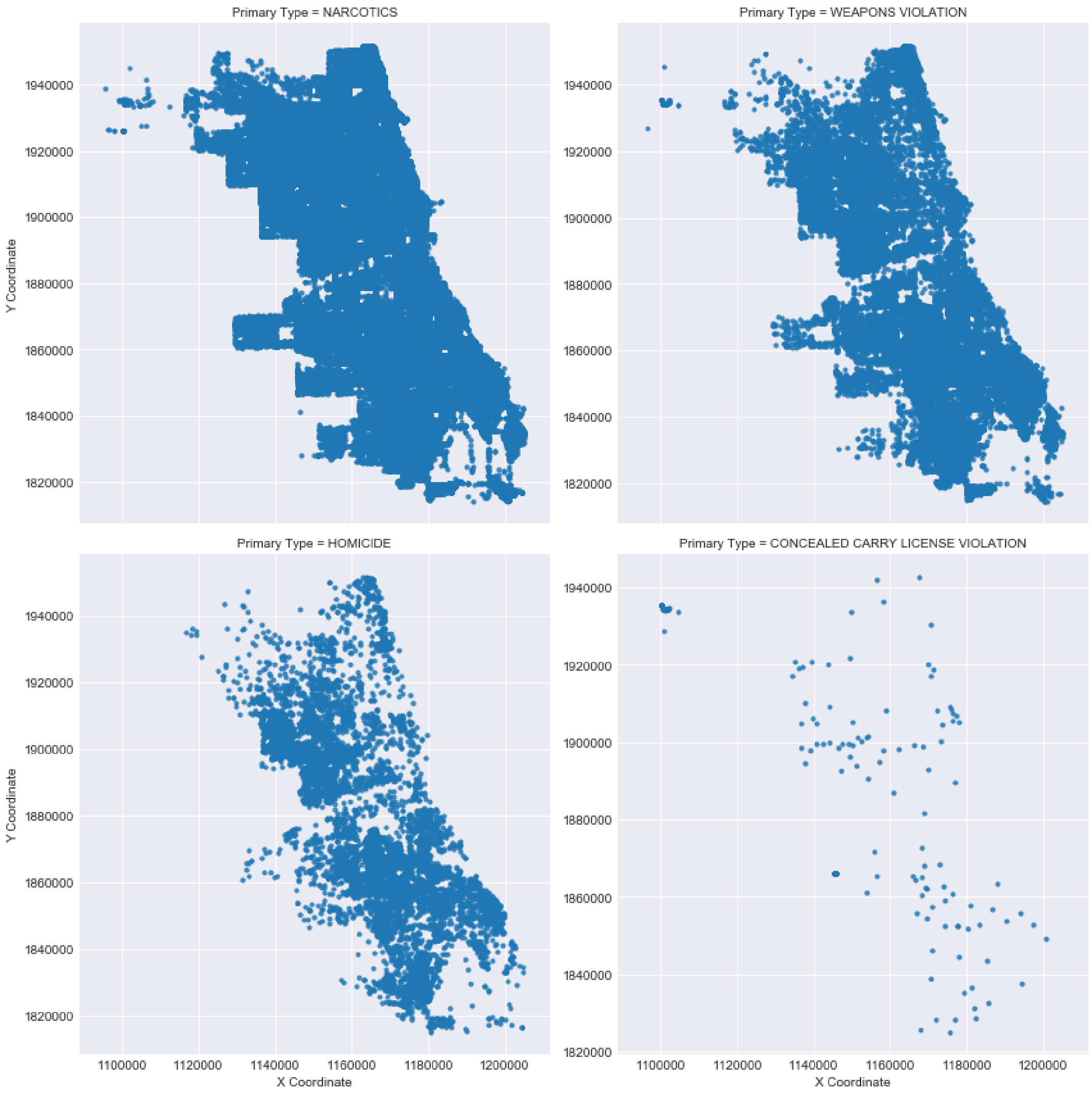
METHODOLOGY

Let us find how all crime in general is distributed across the whole city. For this a scatter plot is created mapping all crime geo locations (X and Y coordinates) in the dataset and plotted it on the city's geographic map.



There are four major crimes which usually happen in Chicago. These four crimes are homicide, concealed weapon violations, narcotics and weapons violation and their plots are shown below.

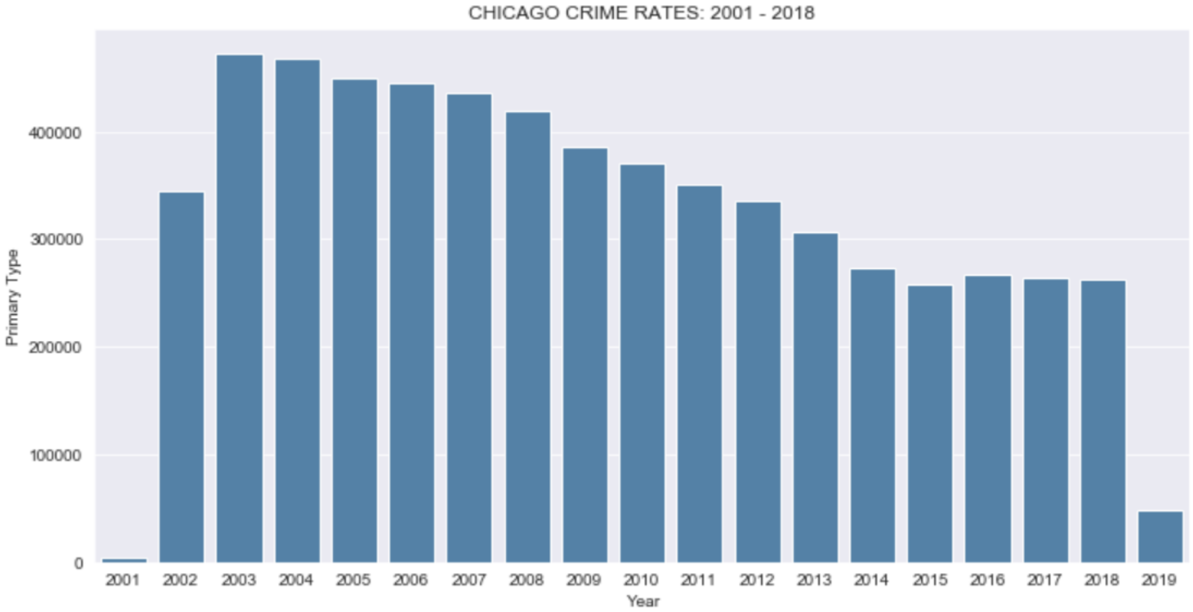
Next I visualized each geographical distribution scatter plot for each of the crimes in the group to understand how the 4 crimes are distributed across the city.



It looks like *Narcotics* and *Weapons Violation* crimes are common all over the city but *Homicide* and *Concealed Carry Violation* crimes have a specific geographic pattern starting to emerge.

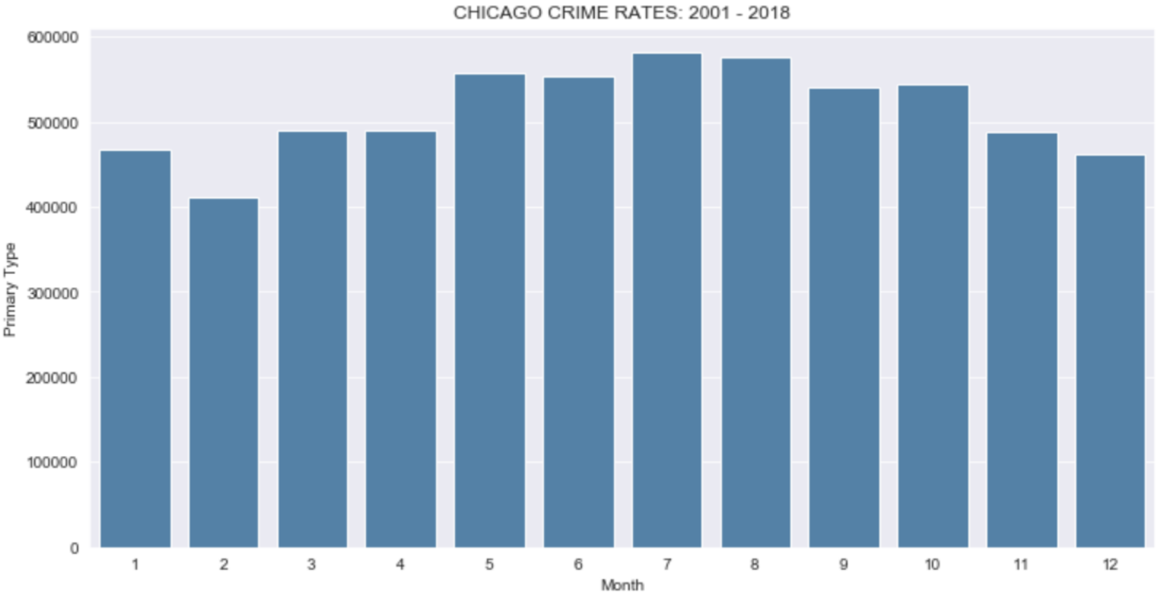
Then we proceed with the data cleaning and the data extractions so that we get a leaner pandas dataframe with just the relevant data from the original crime dataframe

Next I created some crime vs time visualization and made notes on the observations thereafter.



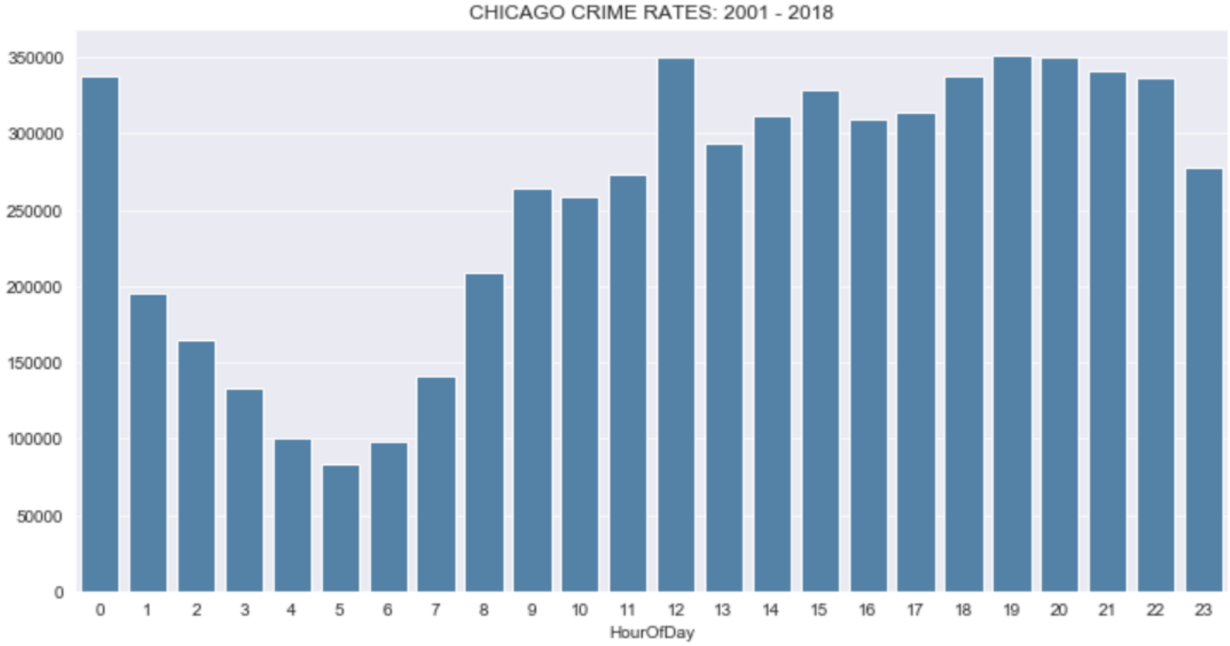
We see that the trend on the decrease in the crime count continued till 2015 post that again there has been a slight spike in the crime counts and the crime rate has not been decreasing further.

Next we see the crime trend for each month



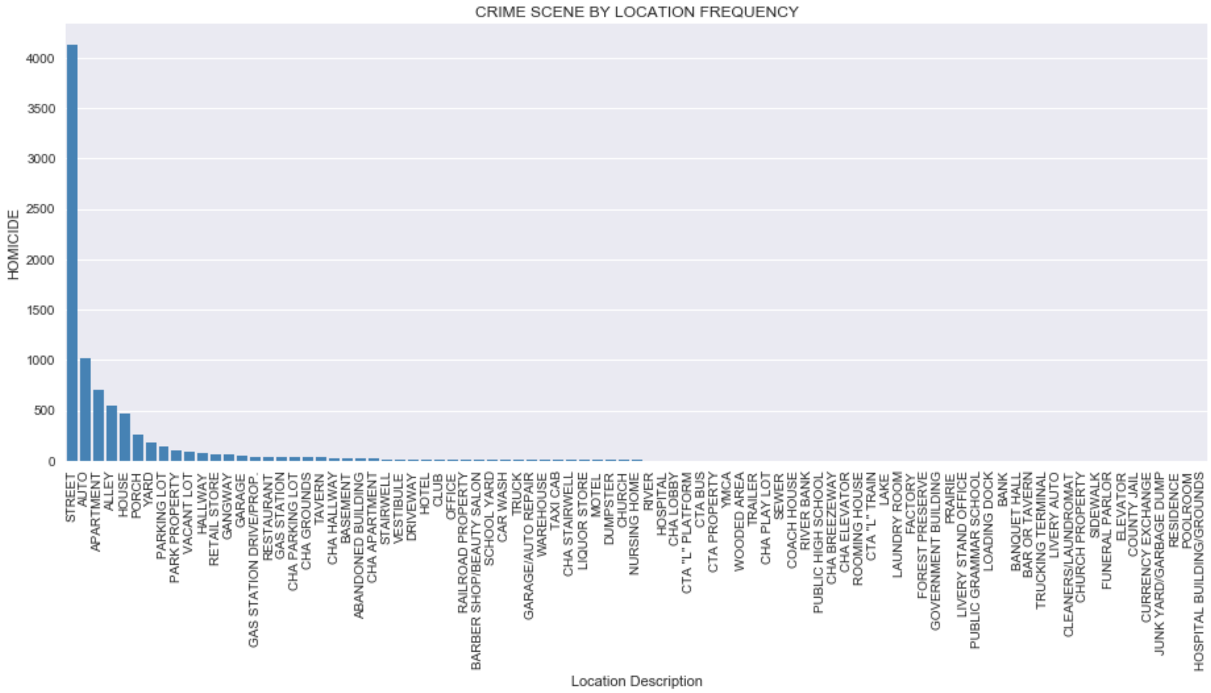
We see that crimes tend to spike during summer months from May-August, and we see the least crime during the holidays of Christmas-New Year

Going further granular we check the crime trend on the hour of the day



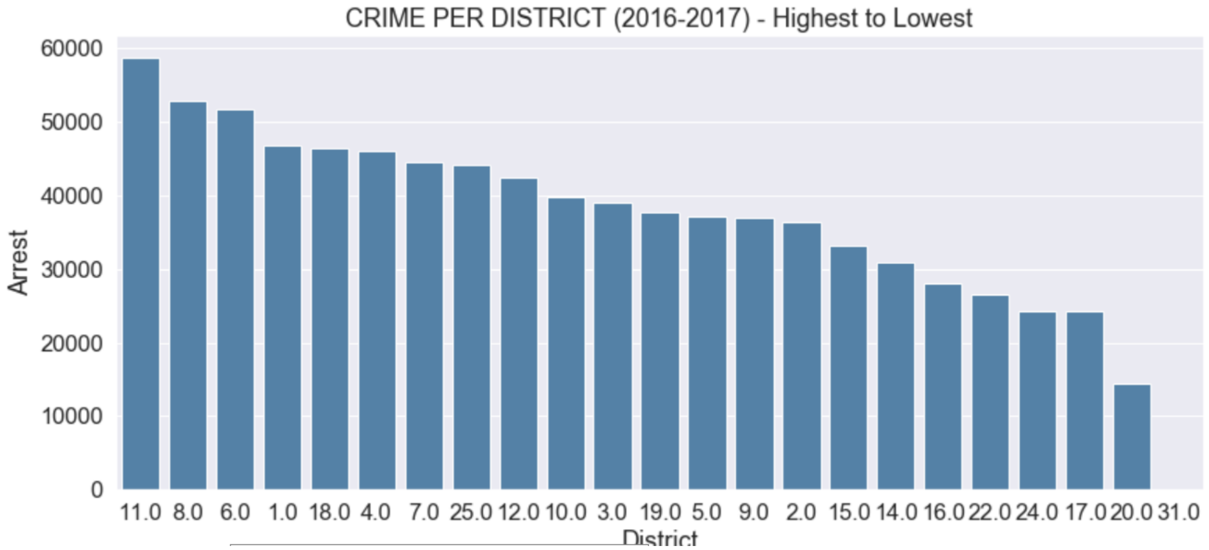
Continuing with the time frame analysis, it makes sense to look at crime rates in Chicago from a time of the day perspective. It becomes evident that crime occur more at night than during the day

Next we look at the crime rate by the Location



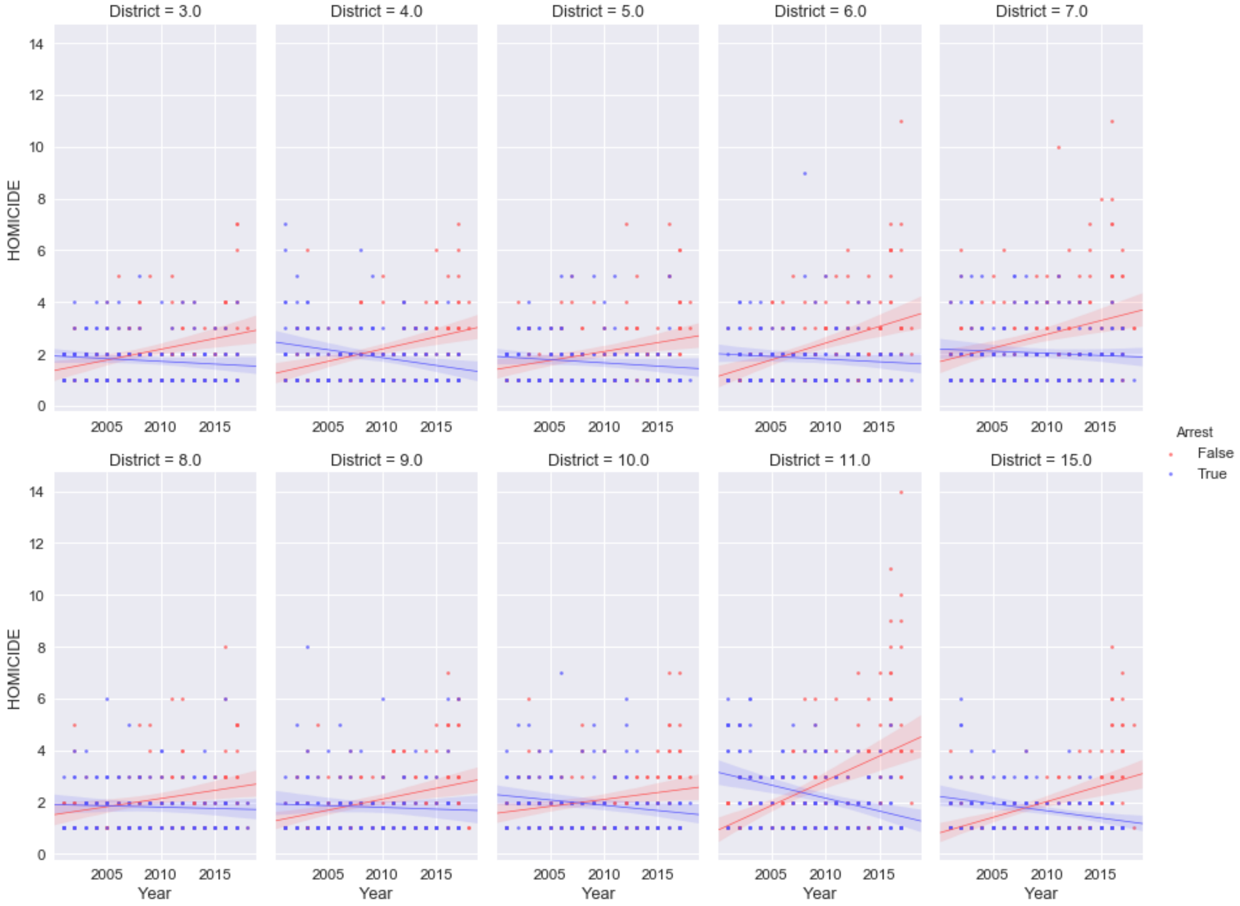
From our Seaborn FactorPlot above, we can observe that majority of the murders (more than 60%) happen on the street. So it seems that street are no safe in Chicago

Then we look into the crime rates based on the Police Districts



District 11, 8 and 6 are the ones most prone to crime and the districts 17,20 and 31 being the ones with the least crime rate.

Then we plot to show regression lines of both arrests and non-arrests on the same axis for the top 10 most dangerous districts



A further break down of the arrest vs. non-arrest on a district level, of just the high crime districts, shows a very interesting observation of decreasing number of arrests over the years in the dataset.

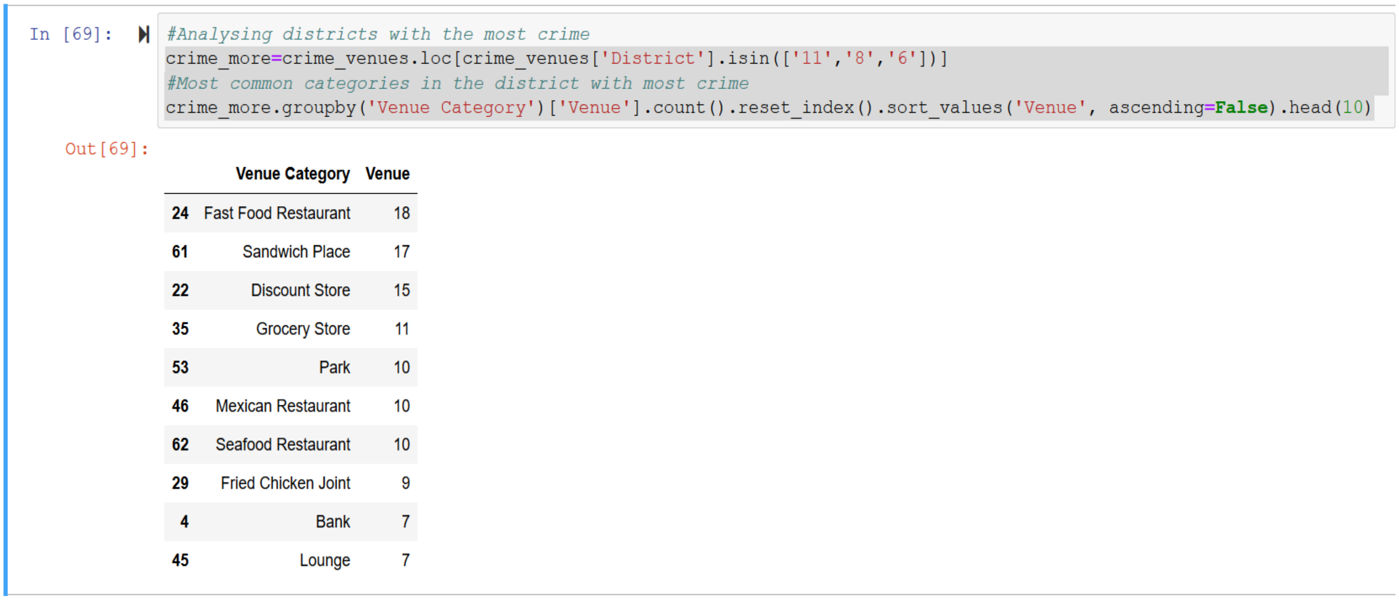
The trend of police making less arrests than the previous year had already began before 2006 and by 2010 the tables flipped in all districts to making arrests on less than half of all crimes.

It may well be that the odds of getting arrested for a homicide are less than half. The odds get much better if you are a criminal living in a high crime district.

ANALYSIS OF THE LOCATIONS OF THE DISTRICT USING FOURSQUARE API

Now you have an insight on the trends of the crime and the Districts in which most and the least crimes take. With those insights how a potential investor can approach his plan of setting up business is what we shall look in this section. If an investor decides to set up his business in a district with high crime count it is imperative for him to give a higher importance to security also it would be useful for him to know what kinds of business are most thriving in those districts. Similar plans can also be made up for a district with a low crime count.

During our exploratory analysis we found that district 11, 8 and 6 were the ones with the most crime rates. Let us see the most common venues in these districts.



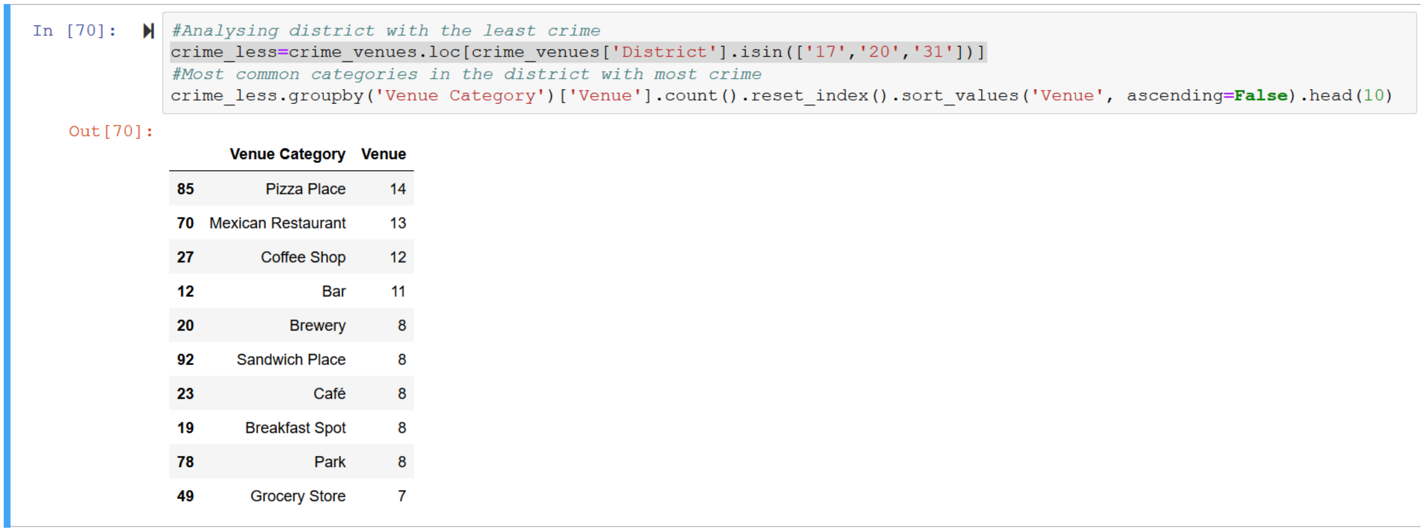
In [69]:

```
#Analysing districts with the most crime
crime_more=crime_venues.loc[crime_venues['District'].isin(['11','8','6'])]
#Most common categories in the district with most crime
crime_more.groupby('Venue Category')['Venue'].count().reset_index().sort_values('Venue', ascending=False).head(10)
```

Out[69]:

	Venue Category	Venue
24	Fast Food Restaurant	18
61	Sandwich Place	17
22	Discount Store	15
35	Grocery Store	11
53	Park	10
46	Mexican Restaurant	10
62	Seafood Restaurant	10
29	Fried Chicken Joint	9
4	Bank	7
45	Lounge	7

Similarly let’s take a look at the districts with the least crime counts.



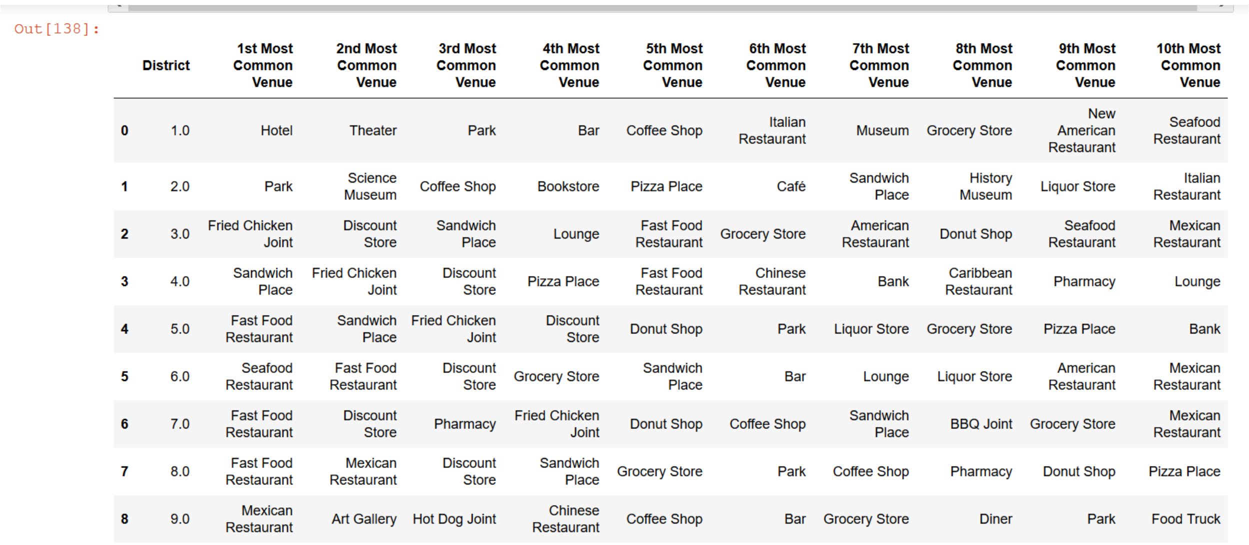
In [70]:

```
#Analysing district with the least crime
crime_less=crime_venues.loc[crime_venues['District'].isin(['17','20','31'])]
#Most common categories in the district with most crime
crime_less.groupby('Venue Category')['Venue'].count().reset_index().sort_values('Venue', ascending=False).head(10)
```

Out[70]:

	Venue Category	Venue
85	Pizza Place	14
70	Mexican Restaurant	13
27	Coffee Shop	12
12	Bar	11
20	Brewery	8
92	Sandwich Place	8
23	Café	8
19	Breakfast Spot	8
78	Park	8
49	Grocery Store	7

Then we perform a one hot encoding to look into the most common venues in all the districts.

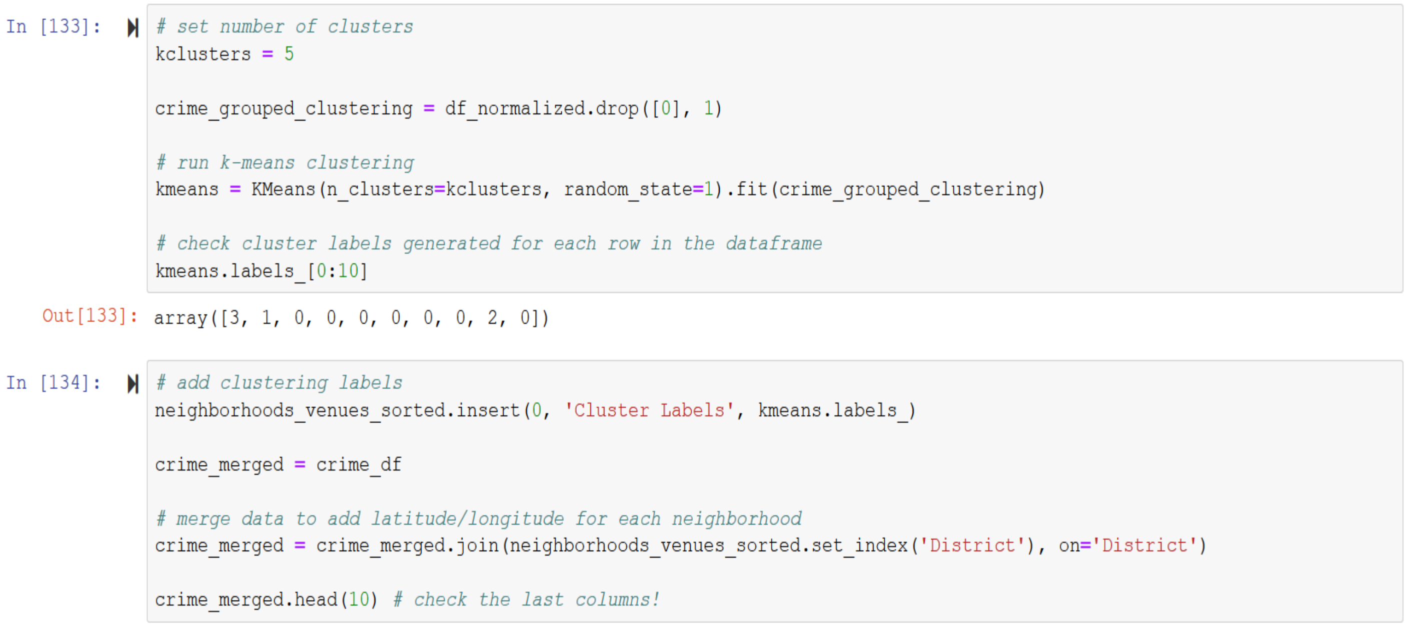


Out[138]:

	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	1.0	Hotel	Theater	Park	Bar	Coffee Shop	Italian Restaurant	Museum	Grocery Store	New American Restaurant	Seafood Restaurant
1	2.0	Park	Science Museum	Coffee Shop	Bookstore	Pizza Place	Café	Sandwich Place	History Museum	Liquor Store	Italian Restaurant
2	3.0	Fried Chicken Joint	Discount Store	Sandwich Place	Lounge	Fast Food Restaurant	Grocery Store	American Restaurant	Donut Shop	Seafood Restaurant	Mexican Restaurant
3	4.0	Sandwich Place	Fried Chicken Joint	Discount Store	Pizza Place	Fast Food Restaurant	Chinese Restaurant	Bank	Caribbean Restaurant	Pharmacy	Lounge
4	5.0	Fast Food Restaurant	Sandwich Place	Fried Chicken Joint	Discount Store	Donut Shop	Park	Liquor Store	Grocery Store	Pizza Place	Bank
5	6.0	Seafood Restaurant	Fast Food Restaurant	Discount Store	Grocery Store	Sandwich Place	Bar	Lounge	Liquor Store	American Restaurant	Mexican Restaurant
6	7.0	Fast Food Restaurant	Discount Store	Pharmacy	Fried Chicken Joint	Donut Shop	Coffee Shop	Sandwich Place	BBQ Joint	Grocery Store	Mexican Restaurant
7	8.0	Fast Food Restaurant	Mexican Restaurant	Discount Store	Sandwich Place	Grocery Store	Park	Coffee Shop	Pharmacy	Donut Shop	Pizza Place
8	9.0	Mexican Restaurant	Art Gallery	Hot Dog Joint	Chinese Restaurant	Coffee Shop	Bar	Grocery Store	Diner	Park	Food Truck

One logical fallacy always remains, suppose a district is having a lower crime ratio but that district is adjacent to a one which is having a higher crime ratio. Then can we go ahead and say that district won't be susceptible to higher crime in the future or shall we take precautionary measures to make sure that criminals from adjacent crime prone districts does not start perpetrating their crimes in that district.

So here we use K-Means clustering to cluster the districts based on their location as well their crime count. This I think will help us in truly fragmenting the locations based on their crimes.



```
In [133]:  # set number of clusters
           kclusters = 5

           crime_grouped_clustering = df_normalized.drop([0], 1)

           # run k-means clustering
           kmeans = KMeans(n_clusters=kclusters, random_state=1).fit(crime_grouped_clustering)

           # check cluster labels generated for each row in the dataframe
           kmeans.labels_[0:10]

Out[133]:  array([3, 1, 0, 0, 0, 0, 0, 0, 2, 0])

In [134]:  # add clustering labels
           neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

           crime_merged = crime_df

           # merge data to add latitude/longitude for each neighborhood
           crime_merged = crime_merged.join(neighborhoods_venues_sorted.set_index('District'), on='District')

           crime_merged.head(10) # check the last columns!
```

Once we examine the Clusters we get the insight that the Cluster 0 and 1 comprise of the districts with high crime count and they are close to each other, similarly the rest of the clusters have lesser crime counts.

We further examine the clusters and check the 10 most common venues in them.

CONCLUSIONS

Throughout this notebook, I used data compiled by the Chicago police department to extract some insights on homicides in Chicago. While the data analysis performed here is devoid of a national perspective my key findings can be distilled to a few key points below:

- The number of crimes vary greatly by police district. Districts with high numbers seem to always have high numbers year after year and vice versa.
- Crime numbers were largely unchanged from 2014 to 2018 which is a good thing and the Chicago PD must be congratulated for this.
- There is a correlation between weather temperatures and number of crimes per month. The warmer the month, the more crimes are committed.
- Day of the week also affects how many murders are committed with weekends seeing higher numbers.
- Time of the day also affects the amount of crimes committed with 9pm to 2am being the most dangerous and 7am to 1pm being the safest hours.
- District 11, 7 and 15 had the highest homicides in 2016–2017 period
- An overwhelming majority of crimes are committed on the street accounting for more than 60% of all crime scenes.
- We also had a picture as to what kind of venues are common in the districts having a high crime count and the districts having a low crime count.

While these observations are very illuminating, I should mention that they don’t paint the whole picture and comparisons should be made with national or other comparable data to give them more perspective. For example, while the number of crimes per year seem high, we don’t know if that is within the national average or too high. Such comparisons would help us build more compelling arguments in our data story telling by using them to support our local findings.

With that being said, our findings are still very relevant to local decision makers because they are very clear on the where and the when but not the why.