Capstone project

Introduction

Have you ever wondered what a life with no crimes will feel like?

The Crime rate have been growing rapidly in the last years, making the work of the police officers much harder. Imagine the time spent to surveillance areas, investigate crimes and understand the criminal behavior. We live in a world where technology solves many of our problems.

In this project I will be helping the Denver police department in analyzing previous crimes with the venues that surrounded the areas where they occurred. This analysis will assist them in predicting new crimes according to the time and location of previous crimes, and how will the type of venues surrounding crime areas affect their number.

The crime analysis will not only help the police officers in understanding the criminal behavior, it will also help in targeting the areas where crimes most probably will occur, saving the time and effort of the police officers.

Data

Two datasets will be used in the project:

- A dataset from Kaggle that contains historical data of Denver's city crimes (offense type, offense category, first occurrence, incident address, longitude, latitude, Neighborhood)
- Dataset extracted using the Foursquare API. Using the neighborhood from the crime dataset, a set of venues will be extracted (Arts and Entertainment, Fitness Center, Food, Medical Center, Nightlife Spot, and Shop and Service)

Methodology

In this part will discus the methods used to prepare, clean and extract all the needed features for the data analysis process.

Crime Data Cleaning and preparation

The Denver crime data consist of 512656 records, that have been recorded in the last 5 years (2015-2020). The dataset has 19 columns (incident ID, offense ID, offense code, offense code extension, offense type, offense category ID, first occurrence date, last occurrence date, reported date, incident address, geo x, geo y, geo lon, geo lat, district ID, precinct ID, neighborhood ID, is crime, is traffic)

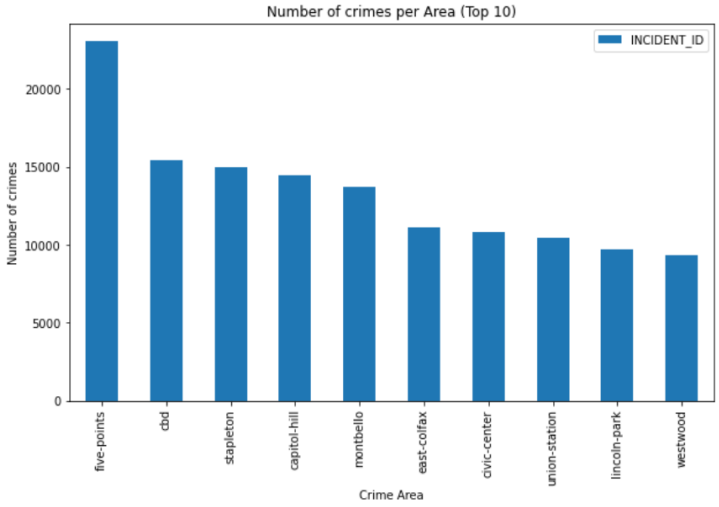
The following steps have been taking to clean and prepare the data for analysis:

- Since the main concern is offense that involves crimes, all rows that contains a value of is_crime=0 and is_traffic=1 will be dropped.

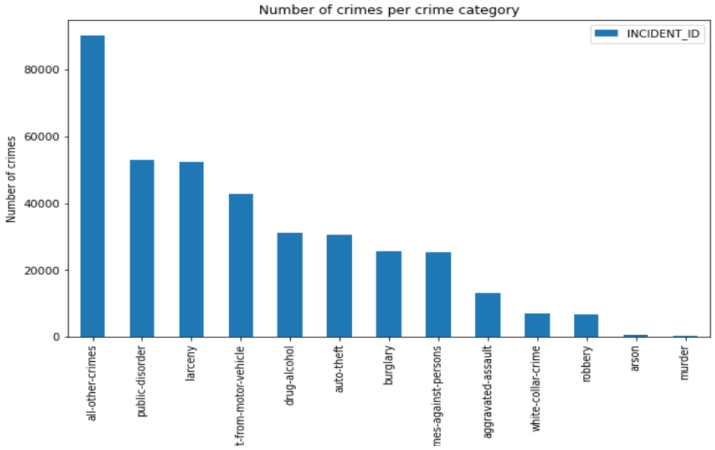
- Both the is_crime and is_traffic columns will be dropped since they only contain one value in all rows and will not be needed for further analysis.
- Dropping all the columns that will not be needed for analysis.
- Extracting the time components from the reported date column and adding all the components as columns to the data frame as shown below

date	day	day_	month	month_	year	year_	hour	min	time_occurred
27	4	Thursday	12	December	2018	2018-12	16	51	16:51:00
12	1	Monday	6	June	2017	2017-06	8	44	08:44:00
09	1	Monday	12	December	2019	2019-12	13	35	13:35:00
22	6	Saturday	12	December	2018	2018-12	22	0	22:00:00
20	5	Friday	4	April	2018	2018-04	13	33	13:33:00

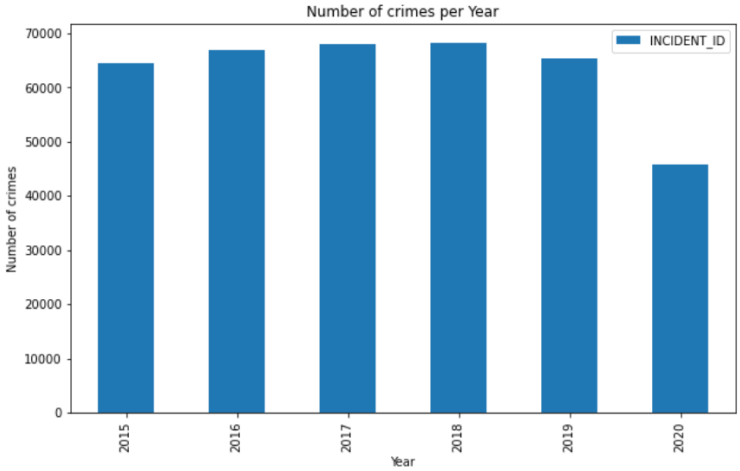
- Drop all the rows that contains NULL values in the GEO_LON and GEO_LAT columns
- Investigating the effect of crime area in crime rates, to decide on which features to include in further analysis

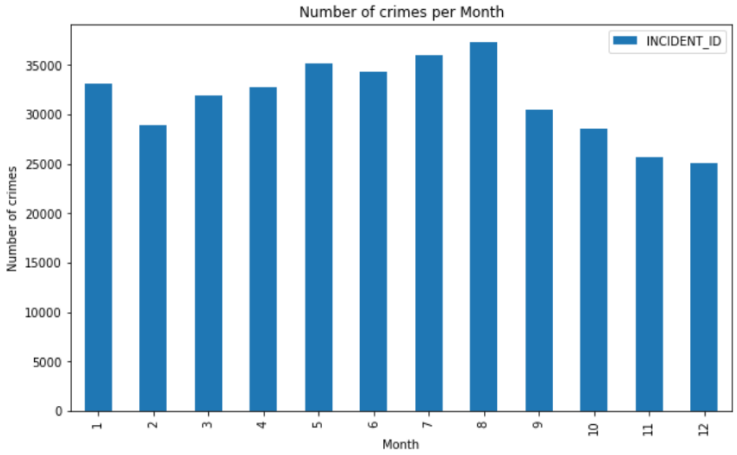


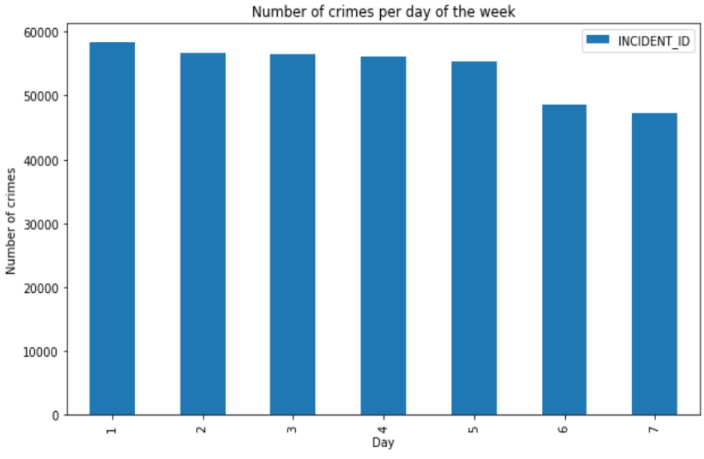
- Investigating the effect of crime category in crime rates, to decide on which features to include in further analysis



- Investigating the affect of time in crime rates, to decide on which features to include in further analysis



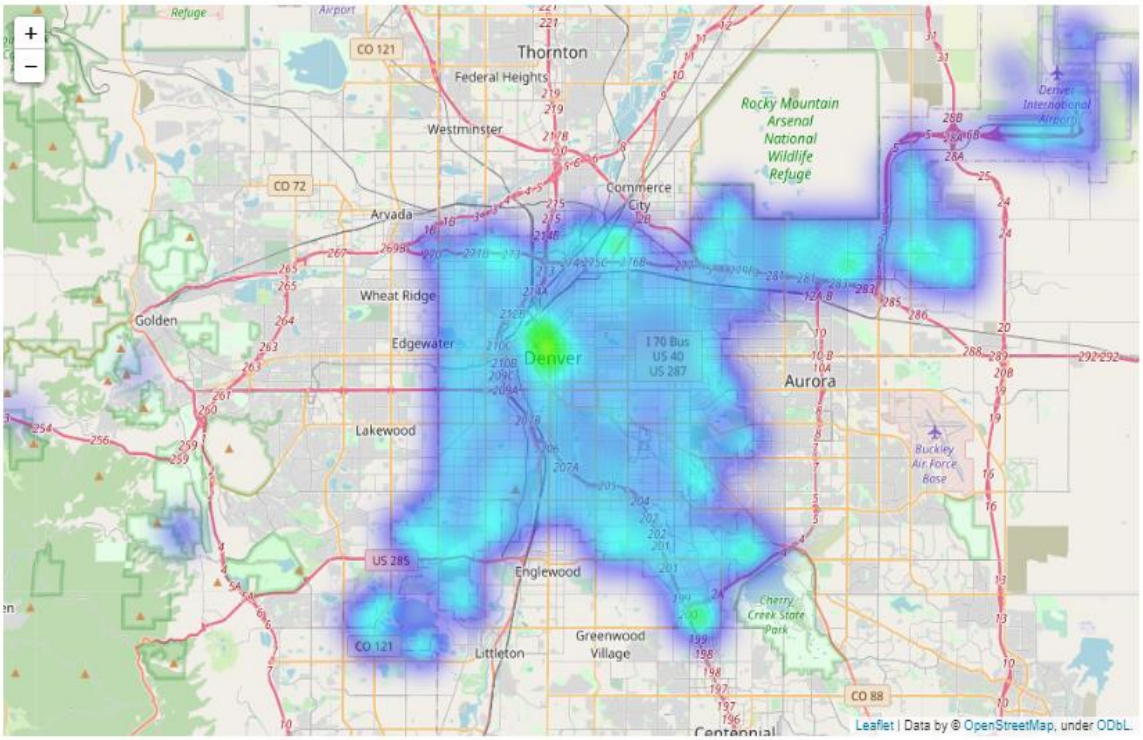




Interesting trends appeared in the analysis of Crime rates comparing to time. However, as our focus is the comparison of crime rates with area, a further investigation will not be done.

- Plotting all crime locations using folium maps and cluster markers.

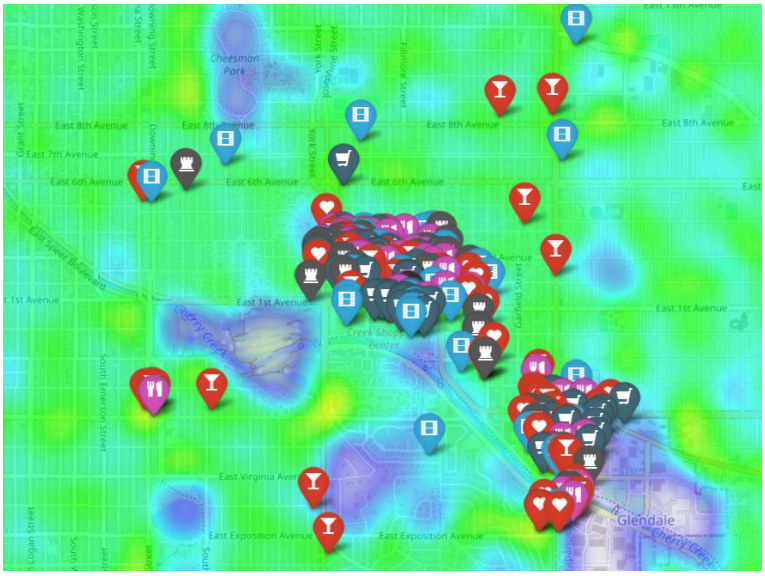
- Plotting all crime locations using folium maps and heat map plugin.



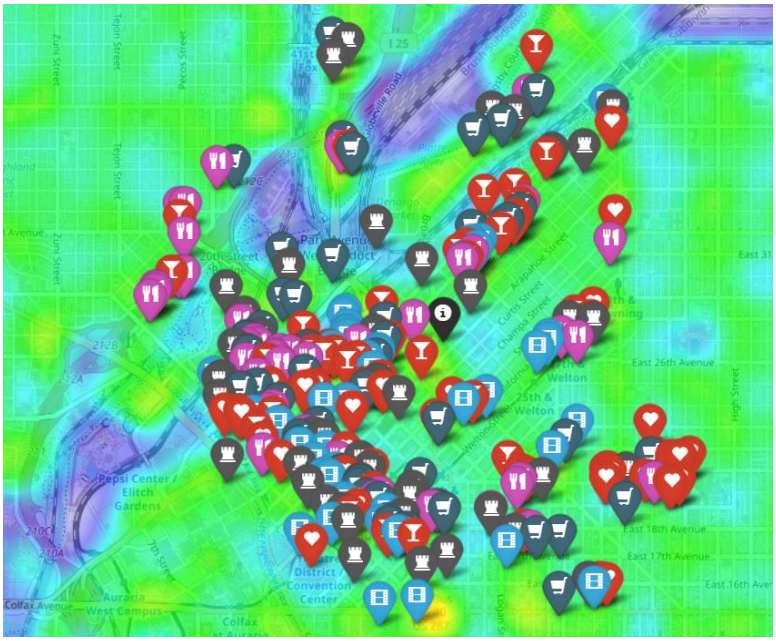
Visualize crime data with venues data in every neighborhood in Denver city

Since there are about 79 neighborhoods in Denver will be only showcasing the ones with high crime rate and venue count. The crime rate will be mapped using a heat map and the venues will be shown as folium map markers.

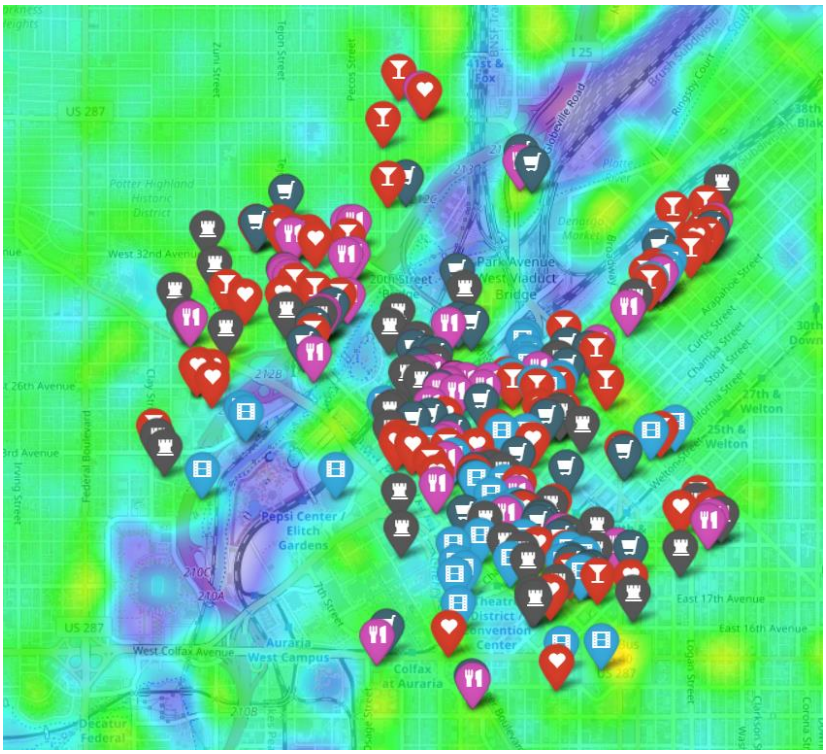
Cheery creek neighborhood:



Five points neighborhood:



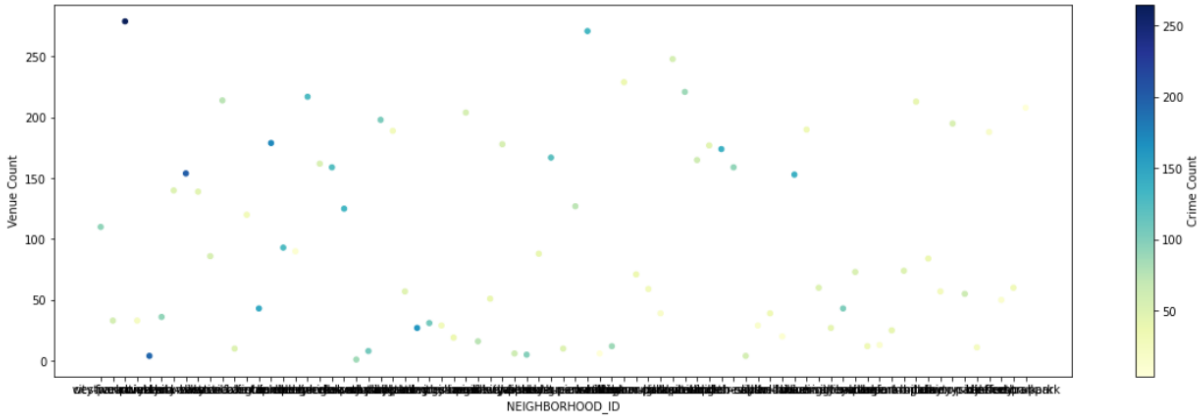
Union station neighborhood:



Visualize relationship between the number of venues and crimes by neighborhood

The following steps have been taking to visualize the relationship:

- Calculate the total business burglary crime for every neighborhood.
- Calculate the total venues for every neighborhood
- Define a data frame that contains the above values.
- Visualize the relationship



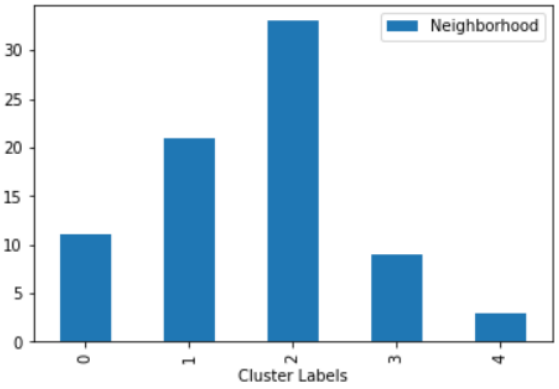
Clustering the data

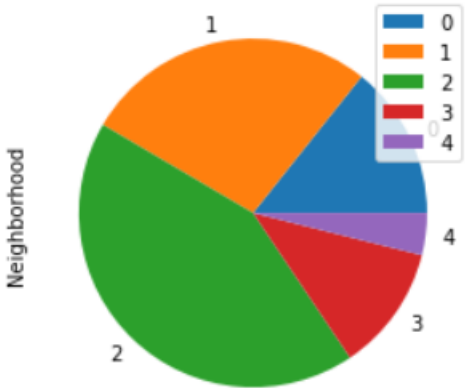
The following steps have been taking to prepare and cluster the data :

- One hot encodes all venues types for each neighborhood
- Get the mean of each encoded value by neighborhood
- Get the most common venues in each neighborhood
- Adding the crime count, venue count for each neighborhood to the data frame used for modeling
- Using K-means clustering with K=5 to cluster the data

Results

The clustering algorithm resulted in five different neighborhood groups. The following visualization will show the distribution of the neighborhoods within the clusters.





The following table will show the disturbuation of the niubourhood along side their total burgulary crime rate and venue count:

Cluster 0

	Neighborhood	Crime Count	Venue Count
0	cbd	88	12
1	harvey-park	87	1
2	highland	146	43
3	mar-lee	75	16
4	montbello	107	31
5	overland	194	4
6	ruby-hill	107	8
7	university-hills	93	36
8	valverde	99	5
9	virginia-village	102	43
10	washington-virginia-vale	162	27

Cluster 1

	Neighborhood	Crime Count	Venue Count
0	athmar-park	104	198
1	barnum	30	120
2	bear-valley	50	140
3	capitol-hill	88	221
4	central-park	6	208
5	city-park	28	189
6	gateway-green-valley-ranch	74	127
7	globeville	93	159
8	goldsmith	58	248
9	kennedy	17	188
10	lowry-field	55	178
11	marston	46	139
12	montclair	41	213
13	regis	56	204
14	rosedale	63	165
15	southmoor-park	37	229
16	sunnyside	54	195
17	university	33	190
18	west-colfax	75	214
19	west-highland	54	162
20	windsor	46	177

Cluster 2

	Neighborhood	Crime Count	Venue Count
0	auraria	33	12
1	barnum-west	28	59
2	belcaro	54	73
3	berkeley	48	57
4	chaffee-park	14	50
5	cheesman-park	42	25
6	city-park-west	59	33
7	civic-center	50	10
8	clayton	27	57
9	cole	40	88
10	congress-park	29	29
11	cory-merrill	37	19
12	country-club	23	33
13	dia	36	84
14	east-colfax	55	86
15	fort-logan	50	74
16	hale	25	11
17	harvey-park-south	65	55
18	hilltop	18	39
19	indian-creek	11	90
20	jefferson-park	31	60
21	north-capitol-hill	59	4
22	north-park-hill	13	13
23	skyland	15	29
24	sloan-lake	34	39
25	south-park-hill	50	10
26	sun-valley	50	60
27	university-park	39	51
28	villa-park	64	6
29	washington-park	43	27
30	washington-park-west	37	71
31	wellshire	11	20
32	whittier	4	6

Cluster 3

	Neighborhood	Crime Count	Venue Count
0	baker	140	153
1	college-view-south-platte	133	125
2	elyria-swansea	199	154
3	hampden	128	93
4	hampden-south	139	174
5	lincoln-park	124	159
6	northeast-park-hill	120	167
7	stapleton	173	179
8	westwood	94	110

Cluster 4

	Neighborhood	Crime Count	Venue Count
0	cherry-creek	128	217
1	five-points	265	279
2	union-station	132	271

Disscussion

Each cluster have a special characteristic that requires the attention of police officers.

- Cluster 0: Contains a very high burglary crime count even though the number if venues is small. Thus, the neighborhoods in this cluster requires a more extensive surveillance.
- Cluster 1: have a very low burglary crime rate even though it's populated with venues. Police officers may try to understand why the crime rate is lower in this neighborhoods comparing to others
- Cluster 2 and 3: the number of crimes and venues is relatively proportional.
- Cluster 4: have a very high burglary crime rate and a high venues count. And a further investigation of the area is needed to understand the causes.

Conclusion

Finally, there's an endless approches and wasys to analyse these dataset and a further more questions to be answered. The joy of data science is never getting out of ideas and creative ways to analyze your data. As a futuer plan for this project, I believe a further analysis about the relationship between the crime rate and time is very promising.