Segmenting and Clustering Crimes Against Females in New York City

Amr Elrasad

1 Introduction

This project is utilizing the power of data science and machine learning to infer a safety maps of the New York City for women.

The project uses the data of incidents reported to New York police department to analyse the relation between police precincts locations and the sex of the victim. In this project, I focus on female sex to be able to come up with safety map showing how danger or safe each precinct is in New York City.

2 Business value

The project idea is interesting for many stakeholders such as:

1. Mobile apps development industry: this looks like a good app on mobile phone for each woman to avoid danger spots in the city. I can be also helpful for tourists.
2. Police department: the information inferred by this study will help the police department to better re-distribute and organize the patrol service
3. Employers: this would help employers better understands the risks of their business locations in relation to female employees.
4. Business Owners: they would utilize this information to direct their advertisement towards women if their business location is in the nearest areas.
5. Housing for females: the information might help planning for female students/employee housing developments to choose best spots of such projects.

3 Data

The data used in this project is available on (https://catalog.data.gov/dataset/nypd-complaint-data-historic). The data is provided by Data.gov which is managed and hosted by the U.S. General Services Administration.

The data provide several keys which are:

CRM_ATPT_CPTD_CD	Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely
HADEVELOPT	Name of NYCHA housing development of occurrence, if applicable
HOUSING_PSA	Development Level Code
JURISDICTION_CODE	Jurisdiction responsible for incident. Either internal, like Police(0), Transit(1), and Housing(2); or external(3), like Correction, Port

	Authority, etc.
JURIS_DESC	Description of the jurisdiction code
KY_CD	Three digit offense classification code
LAW_CAT_CD	Level of offense: felony, misdemeanor, violation
LOC_OF_OCCUR_DESC	Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of
OFNS_DESC	Description of offense corresponding with key code
PARKS_NM	Name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included)
PATROL_BORO	The name of the patrol borough in which the incident occurred
PD_CD	Three digit internal classification code (more granular than Key Code)
PD_DESC	Description of internal classification corresponding with PD code (more granular than Offense Description)
PREM_TYP_DESC	Specific description of premises; grocery store, residence, street, etc.
RPT_DT	Date event was reported to police
STATION_NAME	Transit station name
SUSP_AGE_GROUP	Suspect's Age Group
SUSP_RACE	Suspect's Race Description
SUSP_SEX	Suspect's Sex Description
TRANSIT_DISTRICT	Transit district in which the offense occurred.
VIC_AGE_GROUP	Victim's Age Group
VIC_RACE	Victim's Race Description
VIC_SEX	Victim's Sex Description
X_COORD_CD	X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Y_COORD_CD	Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Latitude	Midblock Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude	Midblock Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

4 Methodology

The phases of this project are as follows

4.1 Data Analysis

In this phase, the data is read and cleaned. Also, the relevant information for study is chosen.

The steps in details are:

1. Read data frame

	CMPLNT_NUM	CMPLNT_FR_DT	CMPLNT_FR_TM	CMPLNT_TO_DT	CMPLNT_TO_TM	ADDR_PCT_CD	RPT_DT	KY_CD
0	619128592	05/25/2017	11:00:00	05/25/2017	11:15:00	48	05/25/2017	351
1	699494668	05/25/2017	11:00:00	05/25/2017	12:00:00	71	05/25/2017	106
2	103321764	05/25/2017	11:00:00	05/25/2017	11:10:00	104	05/25/2017	352
3	137516053	05/25/2017	11:00:00	05/25/2017	11:05:00	113	05/25/2017	341
4	451320561	05/25/2017	11:00:00	NaN	NaN	5	05/25/2017	351

2. Choose relevant information

   I choose the columns to work on. Which are: Borough, crime category, precincts code, incident

	ADDR_PCT_CD	BORO_NM	Latitude	Longitude	LAW_CAT_CD	VIC_SEX
0	48	BRONX	40.846484	-73.893850	MISDEMEANOR	D
1	71	BROOKLYN	40.667757	-73.929829	FELONY	F
2	104	QUEENS	40.729690	-73.874856	MISDEMEANOR	F
3	113	QUEENS	40.678989	-73.791585	MISDEMEANOR	D
4	5	MANHATTAN	40.713989	-73.998714	MISDEMEANOR	E

   location, and victim sex

3. Choose victim sex

   I choose the rows where the female is the victim sex

	ADDR_PCT_CD	BORO_NM	Latitude	Longitude	LAW_CAT_CD
0	71	BROOKLYN	40.667757	-73.929829	FELONY
1	104	QUEENS	40.729690	-73.874856	MISDEMEANOR
2	48	BRONX	40.855168	-73.887904	MISDEMEANOR
3	45	BRONX	40.842202	-73.849757	MISDEMEANOR
4	24	MANHATTAN	40.794953	-73.971437	FELONY

4. One hot encoding
   One hot encoding to change categorial column of crime category as in figure below:

	ADDR_PCT_CD	BORO_NM	Latitude	Longitude	FELONY	MISDEMEANOR	VIOLATION
0	1	MANHATTAN	40.710229	-74.007746	0	0	1
1	1	MANHATTAN	40.720464	-74.006852	0	1	0
2	1	MANHATTAN	40.715529	-74.009240	0	1	0
3	1	MANHATTAN	40.722855	-74.003375	1	0	0
4	1	MANHATTAN	40.705024	-74.012978	0	1	0

5. Scoring
   A score column is added to give a single value metric to each index based on crime category.

```
1 df_grouped
```

	ADDR_PCT_CD	BORO_NM	Latitude	Longitude	FELONY	MISDEMEANOR	VIOLATION	Score
0	1	MANHATTAN	40.714436	-74.008120	193.0	284.0	113.0	2043.0
1	5	MANHATTAN	40.718126	-73.995498	160.0	254.0	87.0	1736.0
2	6	MANHATTAN	40.733790	-74.001023	233.0	205.0	78.0	1936.0
3	7	MANHATTAN	40.716364	-73.984518	163.0	294.0	132.0	1961.0
4	9	MANHATTAN	40.726165	-73.983405	273.0	369.0	102.0	2676.0
5	9	QUEENS	40.726554	-73.987828	1.0	0.0	0.0	5.0
6	10	MANHATTAN	40.747869	-74.000401	177.0	224.0	82.0	1721.0
7	13	MANHATTAN	40.738675	-73.985680	281.0	335.0	123.0	2656.0
8	14	MANHATTAN	40.752727	-73.988398	405.0	381.0	126.0	3420.0
9	17	MANHATTAN	40.752853	-73.971203	178.0	156.0	117.0	1592.0

6. Adding empty clusters
   Using a new copy of the data frame with new column for cluster labels is added. It is initialized with NaN.

	Cluster Labels	ADDR_PCT_CD	BORO_NM	Latitude	Longitude	FELONY	MISDEMEANOR	VIOLATION	Score
0	NaN	1	MANHATTAN	40.714436	-74.008120	193.0	284.0	113.0	2043.0
1	NaN	5	MANHATTAN	40.718126	-73.995498	160.0	254.0	87.0	1736.0
2	NaN	6	MANHATTAN	40.733790	-74.001023	233.0	205.0	78.0	1936.0
3	NaN	7	MANHATTAN	40.716364	-73.984518	163.0	294.0	132.0	1961.0
4	NaN	9	MANHATTAN	40.726165	-73.983405	273.0	369.0	102.0	2676.0

7. Cleaning repetitions
   A cleaning is needed to clean repeated rows with precincts in multiple boroughs.

8. Grouping:

Grouping is done based on precincts codes and the location coordinates are averaged.

```
1 df_grouped
```

	ADDR_PCT_CD	BORO_NM	Latitude	Longitude	FELONY	MISDEMEANOR	VIOLATION	Score
0	1	MANHATTAN	40.720384	-74.006939	193.0	284.0	113.0	2043.0
1	5	MANHATTAN	40.716097	-73.997252	160.0	254.0	87.0	1736.0
2	6	MANHATTAN	40.733985	-74.005457	233.0	205.0	78.0	1936.0
3	7	MANHATTAN	40.716392	-73.983726	163.0	294.0	132.0	1961.0
4	9	MANHATTAN	40.726359	-73.988002	273.0	369.0	102.0	2676.0
5	10	MANHATTAN	40.742876	-73.998551	177.0	224.0	82.0	1721.0
6	13	MANHATTAN	40.736980	-73.982771	281.0	335.0	123.0	2656.0
7	14	MANHATTAN	40.753830	-73.995050	405.0	381.0	126.0	3420.0
8	17	MANHATTAN	40.756657	-73.970653	178.0	156.0	117.0	1592.0
9	18	MANHATTAN	40.765130	-73.985013	302.0	357.0	184.0	2949.0

4.2 Foursquare API

I used Foursquare API to get the location coordinates of each precinct. I created a list of all of them on my Foursquare account. I get list id query to get list details. A loop is used to parse list and nested dictionaries to get each list item location and updates the corresponding data frames.

4.3 Clustering

I used two clustering approaches:

1. K-Means: This algorithm helps clustering data into desired number of clusters and helps infer the relation between data samples.
2. DBSCAN: This algorithm can give insights about outliers and can help decide appropriate number of clusters.

More details are in the results section

5 Results & Discussion

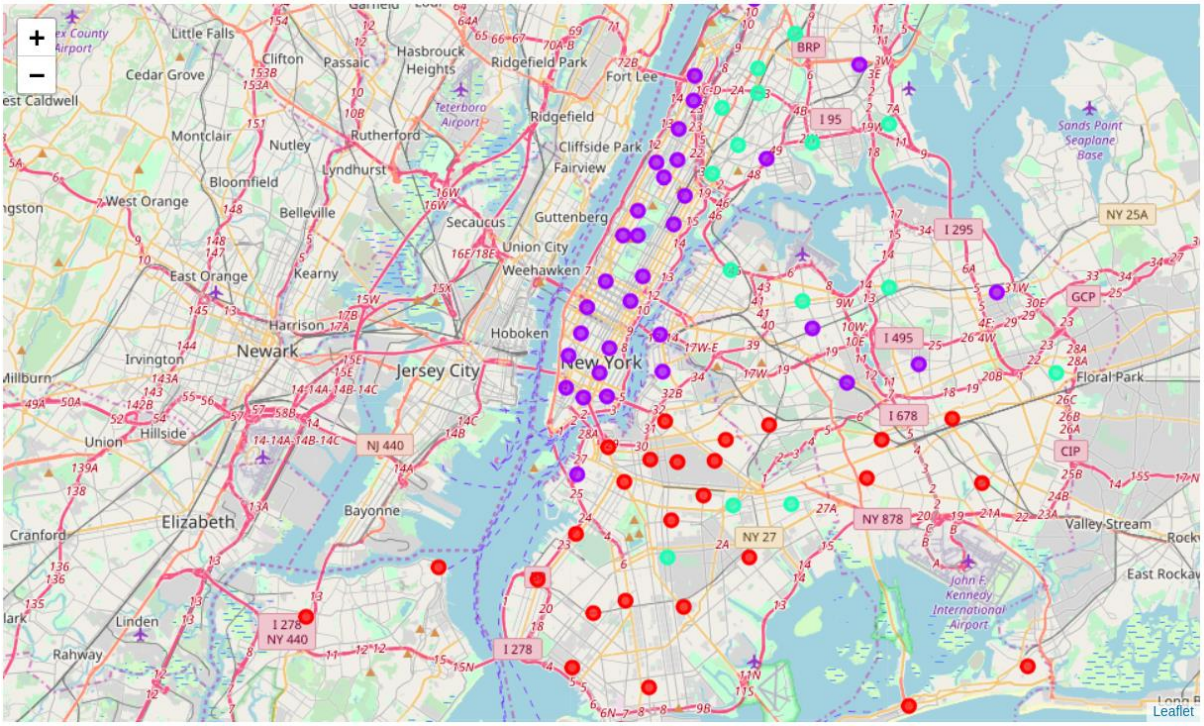
In this section, an overview of the results is presented.

I have done four clustering experiments as follows:

5.1 K-means I

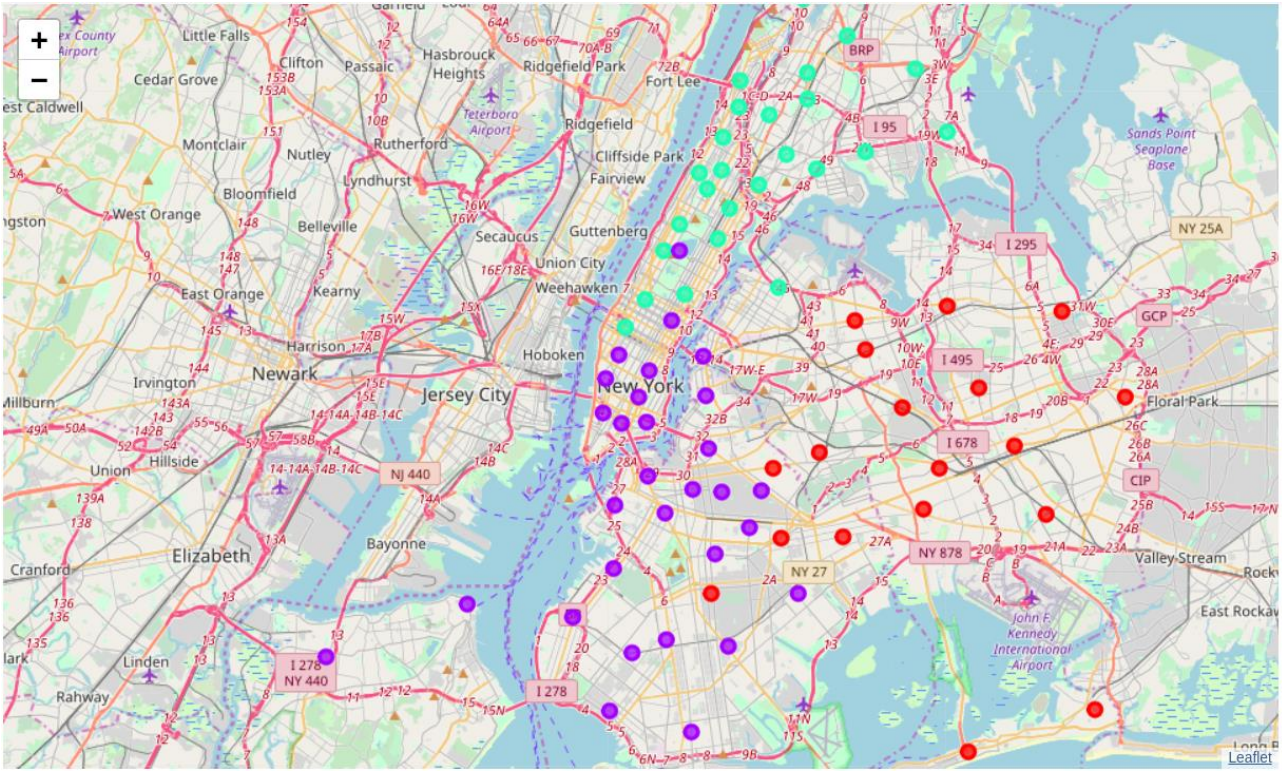
In this experiment, I used the crime category count for each type and the location coordinates to cluster these data points. The results from clustering is as below

There is some geographical over lapping between clusters. This is due to the fact that coordinates are among the data passed to clustering algorithm.



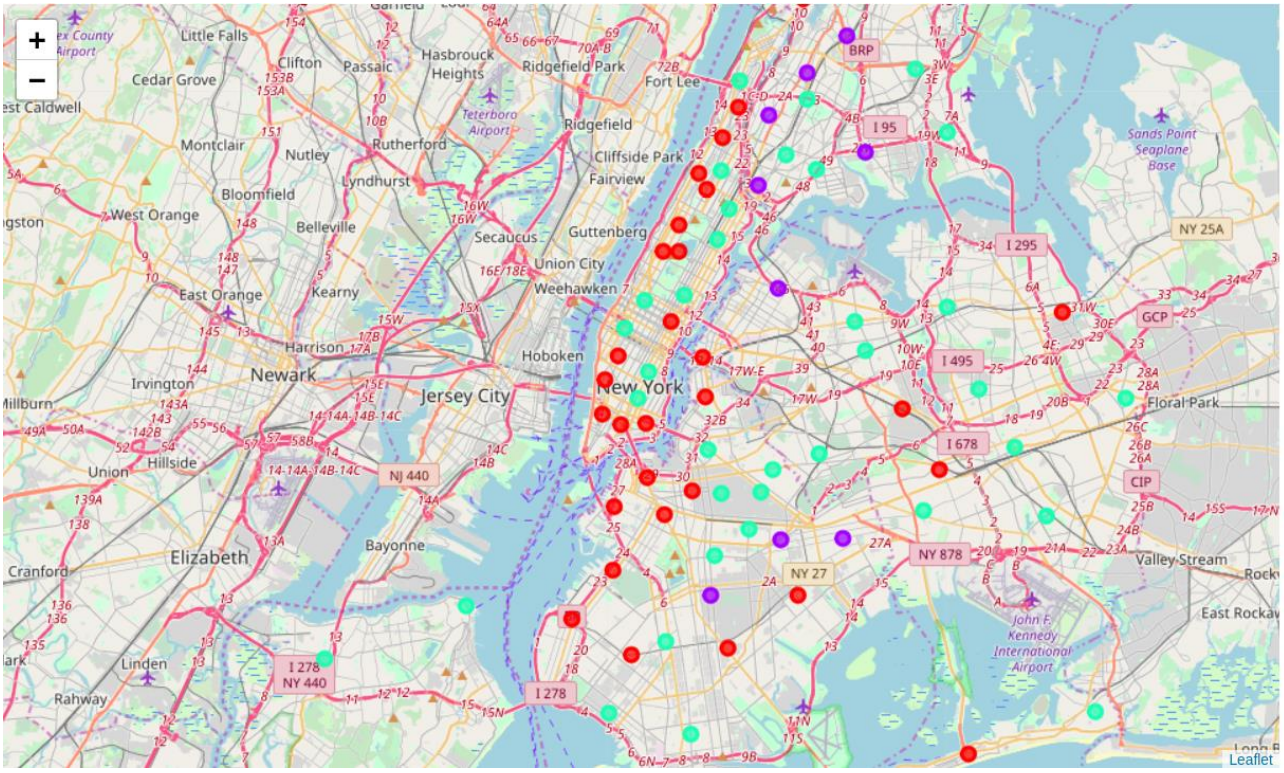
5.2 K-Means II

In this experiment, I used the score and the location coordinates to cluster these data points. The results from clustering is as below



In this case, less overlapping among clusters is noticed. This is due to the fact that only one metric in this experiment (score) is passed to the algorithm unlike the previous experiment

5.3 K-Means III



In this experiment, I used score only for clustering.

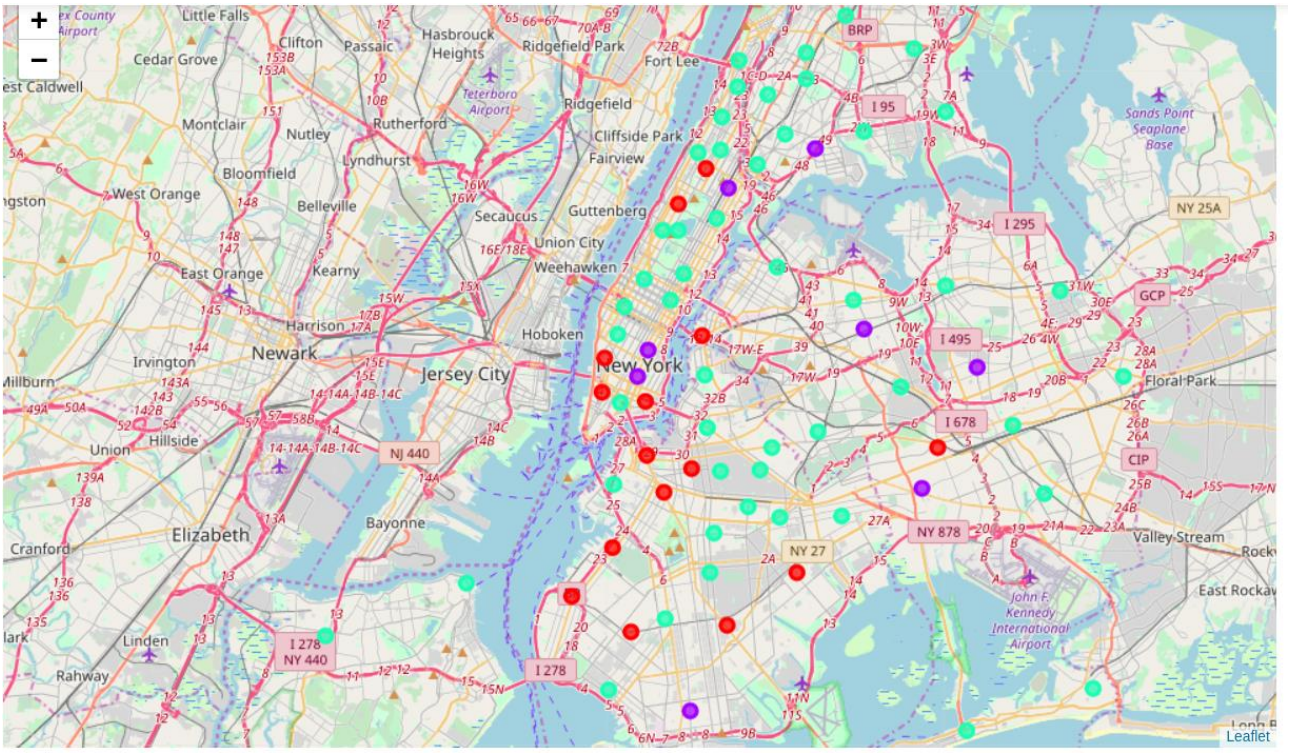
A large geographical scattering is noticed in this case. This is expected as the geographical coordinates are not used in clustering this time.

5.4 DBSCAN

In this experiment, I explored the power of DBSCAN algorithm to make use of:

1. Outlier detection
2. Ability to select number of clusters

The results are as below:



The algorithm classified data into 3 clusters based on the score. The geographical coordinates are not used in this experiment.

6 Conclusion

This study provided insights about crime against women in New York City. The study used the geographical locations, the crime category for clustering

Four different experiments were made. The four outcomes can be of potential interests for many stakeholders and governmental agencies.

I hope this study will provide safety of each woman in New York and would help increase crime avoidance and prevention rates.