



5 APRIL 2021

FELONIES INVOLVING A DANGEROUS WEAPON NEAR NEW YORK SCHOOLS IN 2019

ANTHONY ALBERTYN

Anthony Albertyn BSc(Hons) MSc

Data Scientist

1. INTRODUCTION

BACKGROUND

A quick online search will reveal that crime in New York state has been falling since the 1990's but in 2019 there were still a significant number of felony-level crimes committed with a dangerous weapon. Teenagers can be on both ends, as perpetrators as well as victims, and can potentially suffer trauma and injury if they are about when these crimes are committed. The NYPD does publish crime statistics and various organisations create different kinds of crime maps to visualise the data. I could find various media reports on schools with highest crime rates, but these usually focused on a specific area, like Brooklyn, or are out of date, or did not break down the crimes down sufficiently or did not plot the schools on a map. The idea for this research was born when I realised that there was a gap in this information, and because of my perception of the need to do more to help teenagers in New York state avoid violence.

AIM

The aim of the research is to analyse crime data in New York city and relate that to nearby high schools to find the most at-risk schools where children could get caught up in serious violent crime.

OBJECTIVES

1. Cluster high schools in New York state by number of felonies committed in 2019, within a 500m radius of the school between the hours of 7am to 9am and 2pm to 6pm
2. Find the top-twenty high schools in New York state with the highest amount of crimes determined in objective one

RESEARCH QUESTION

Which high schools in New York state had the highest number of felonies, involving a dangerous weapon, in 2019?

AUDIENCE

The research may be of interest to law enforcement, education authorities and social support organisations in New York state who have an interest in protecting teenagers from serious crime committed near their schools.

VALUE

The research will be important to this audience because it can help identify areas and high schools where school children are at risk of being harmed by serious crime, and allow the authorities to identify where best to focus limited resources to tackle the problem.

2. DATA

DATA SOURCES

The data sources used for this project are:

Publisher: NYC Open Data

Data set: NYPD Complaint Data Historic

Resource link: <https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i>

Columns of data: 35

Rows of data: 6.5 million

Format: CSV

Published: Foursquare

Data: From developer API

Resource link: <https://foursquare.com/developers>

Format: JSON

DATA LIMITATION

The Foursquare data appears to be of a lower quality than the NYPD data because the Foursquare data is submitted by Foursquare users, and due to the volume of the data received, Foursquare probably cannot verify all the data. Some problems encountered with Foursquare data was duplication of venues, region in city name, city name left blank, city name in state field. The Foursquare data might not include all high schools in New York state either. This capstone project specifically requires the use of Foursquare location data, and if this was a commercial project, the author would use an alternative more comprehensive data set for New York high schools.

EXAMPLE NYPD COMPLAINT DATA

| DATE | TIME | OFFENCE | CATEGORY | LATITUDE | LONGITUDE |
|------------|----------|---------------------|----------|-----------|------------|
| 01/01/2019 | 16:25:00 | Dangerous Weapon | Felony | 40.701528 | -73.943227 |
| 01/02/2019 | 17:40:00 | Dangerous Weapon | Felony | 40.70376 | -73.79915 |

EXAMPLE FOURSQUARE DATA

```
JSON = """"venue": {
  "id": "44e9fddaf964a520bb371fe3",
  "name": "Stuyvesant High School",
  "contact": {
    "phone": "2123124800",
    "formattedPhone": "(212) 312-4800",
    "twitter": "stuyvesanthigh",
    "facebook": "112007138816093",
    "facebookName": "Stuyvesant High School"
  },
  "location": {
    "address": "345 Chambers St",
    "crossStreet": "at West St.",
    "lat": 40.717608879529976,
    "lng": -74.01320281198944,
    "labeledLatLngs": [
      {
        "label": "display",
        "lat": 40.717608879529976,
        "lng": -74.01320281198944
      }
    ]
  },
  "distance": 6406,
  "postalCode": "10282",
  "cc": "US",
  "neighborhood": "Battery Park City",
  "city": "New York",
  "state": "NY",
  "country": "United States",
  "contextLine": "Battery Park City",
  "contextGeoId": 7620,
  "formattedAddress": [
    "<span itemprop=\\"streetAddress\\">345 Chambers St</span> (at West St.)",
    "<span itemprop=\\"addressLocality\\">New York</span>, <span itemprop=\\"addressRegion\\">NY</span> <span itemprop=\\"addressCountry\\">United States"
  ]
}, """
```

Figure 1. JSON data from Foursquare

3. METHODOLOGY

DATA EXTRACTION

The data was be extracted as follows:

1. Filtered the NYPD Complaint data by year (2019), level of crime (felony), description of crime (dangerous weapon), time of crime (between 7am to 9pm and between 2pm to 6pm) and crime completed (true).
2. The latitude and longitude data for each of these felonies were used to do an API call to Foursquare to find venues with category 'high school' within a radius of 500 m from the crime.
3. A list of these high schools, including their venue id, name, city, latitude and longitude data was compiled. This list contained no high schools that did not have any felonies committed near them because they were found by their distance from a felony.
4. The list of schools was cleaned and grouped by venue id, and a count column was added - so it is known how many crimes linked to a specific this school as a venue by proximity to it.

DATA ANALYSIS

The data was analysed as follows:

1. K-means ML analysis was applied to the high school data to cluster the schools by count of nearby felony crimes and a cluster label was be added to the school data in a new column.
2. The high school data was then sorted in descending order by number of felonies and the top twenty records were extracted.

DATA VISUALISATION

The data was visualised as follows

1. Plot felony crimes on its own map using Folium.
2. Plot high schools clusters on a map sing Folium.
3. Print top twenty high schools including the total number of crimes for each school.
4. Plot the top twenty high schools on a map using Folium.

4. ANALYSIS

The NYPD data set returned 6.4 million crimes, and after filtering them by year, time, type of crime and level of crime, these results were reduced to 1009 felonies. These felonies were visualised on a map to confirm the geographic location.

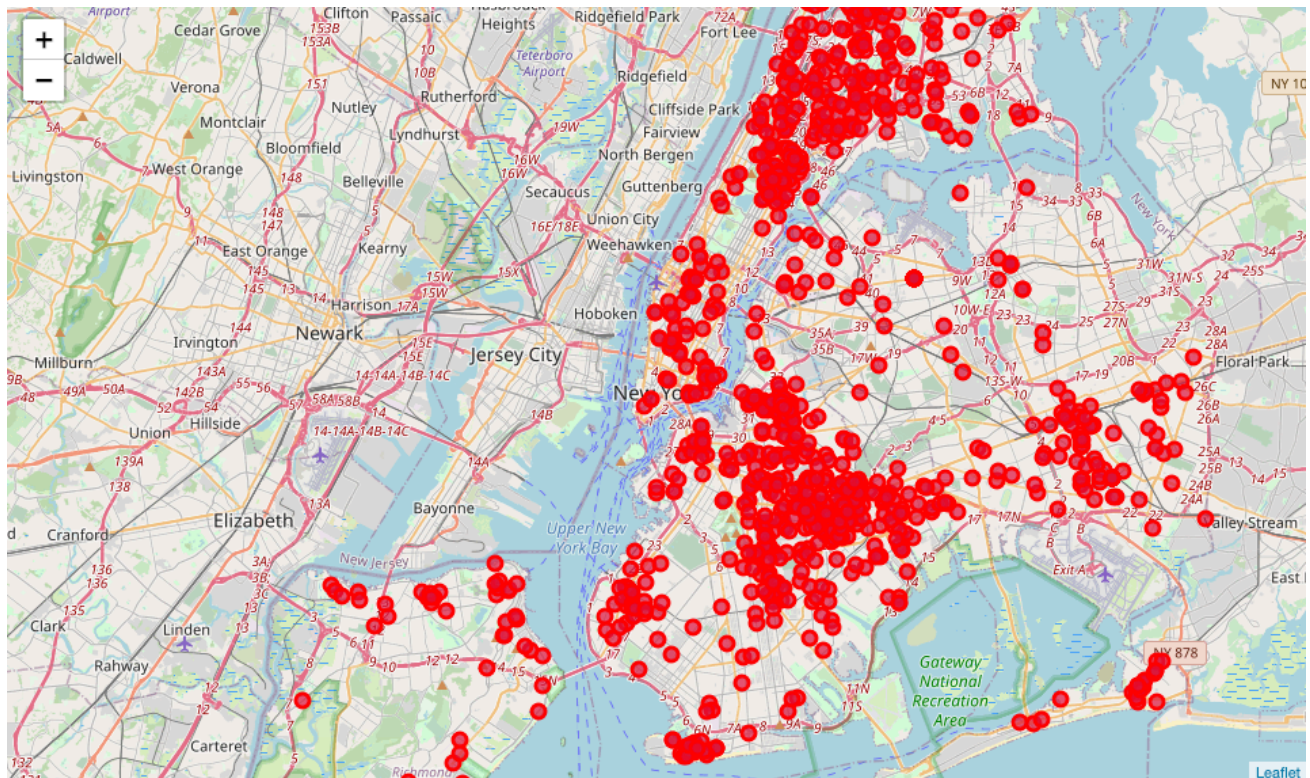


Figure 2. Map of felonies with a dangerous weapon - 2019

From the map it can be seen that many of the crimes were in New York city and Brooklyn.

The NYPD crime data did not include data on nearby schools, so the author used Foursquare API to find nearby high schools and count the number of crimes near each high school.

After reducing the number of crimes to those that took place near high schools, the total number of crimes at different time of the day varied as follows:

| TIME | FELONY COUNT |
|------------|--------------|
| 7am to 8am | 77 |
| 8am to 9am | 62 |
| 2pm to 3pm | 166 |
| 3pm to 4pm | 175 |
| 4pm to 5pm | 248 |
| 5pm to 6pm | 281 |

The high schools were sorted by highest count of crime and the top twenty schools with the highest number of nearby felonies were mapped to obtain a visualisation of their locations.

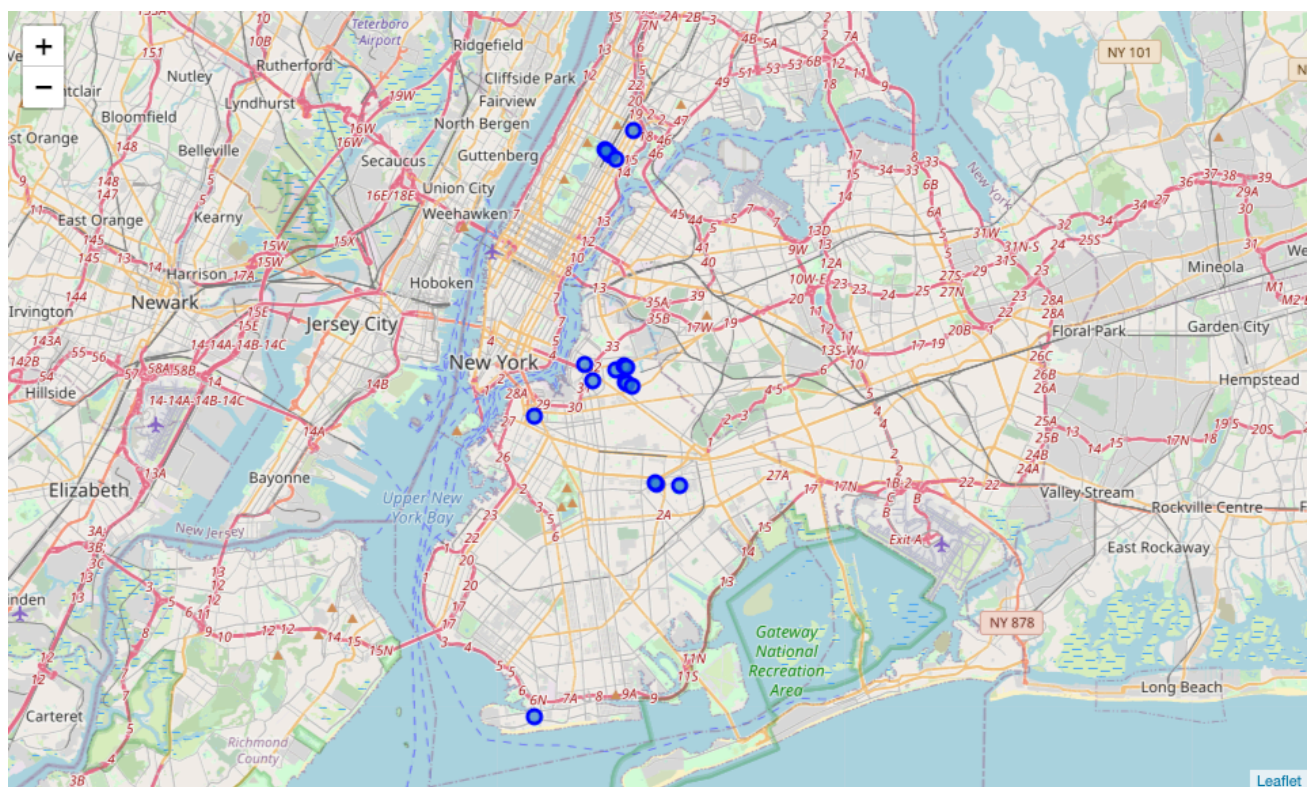


Figure 3. Map of twenty high schools with highest number of nearby felonies in 2019

The high school data, that included the total felonies near each school was then segmented into six clusters using the un-supervised K-means clustering machine learning algorithm.

To determine the optimal number of clusters to segment the data by, the author used The Elbow Method to determine the distortion error for different number of clusters. Six clusters appeared to be a sensible choice.

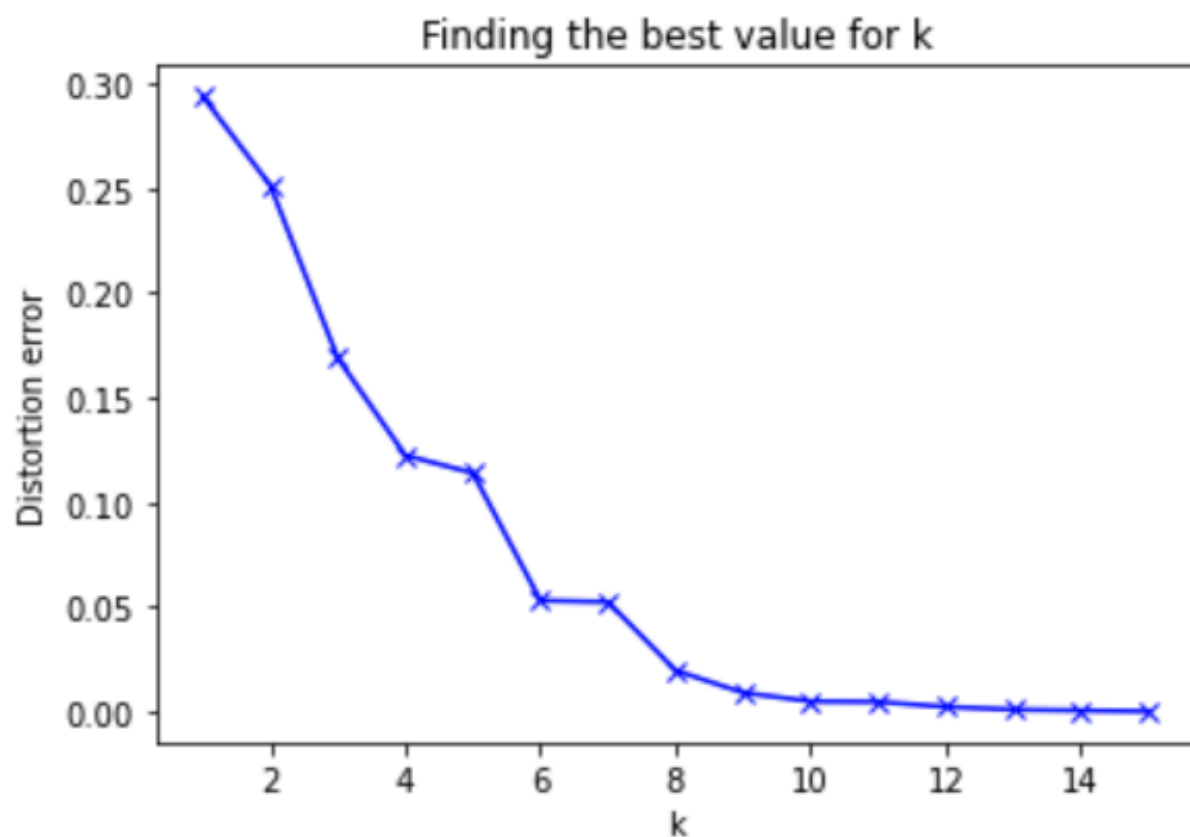


Figure 4. The Elbow Method to determine k value of K-means algorithm

After clustering the high-schools in clusters numbered 0 to 5, the distribution of the number of high schools in each cluster were are follows:

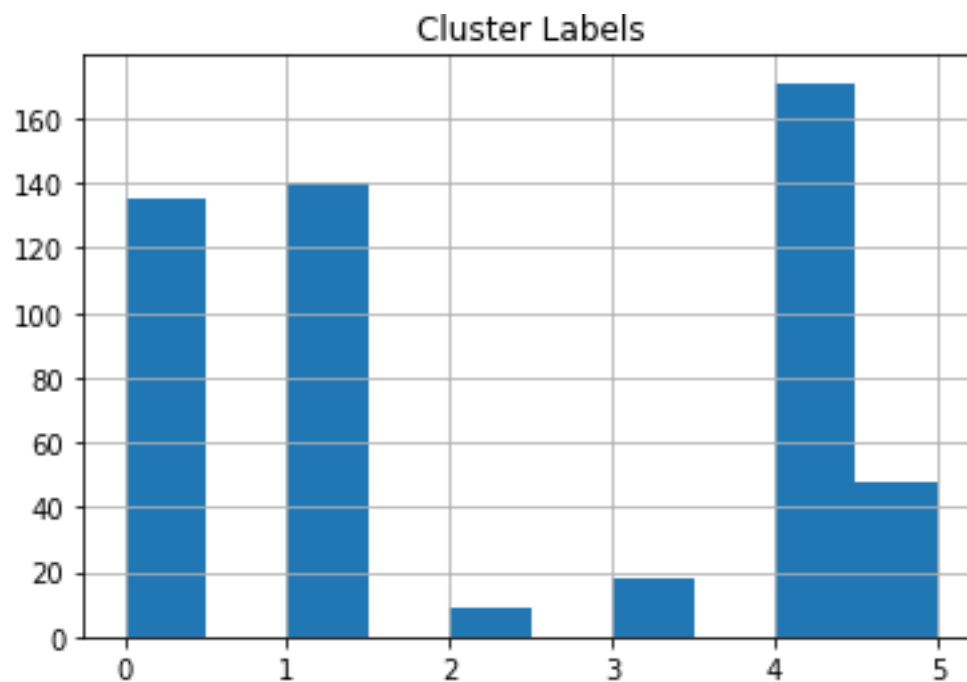


Figure 5. Distribution of high schools by cluster number

The distribution of the number of crimes per school for each cluster can be visualised in box plates as follows:

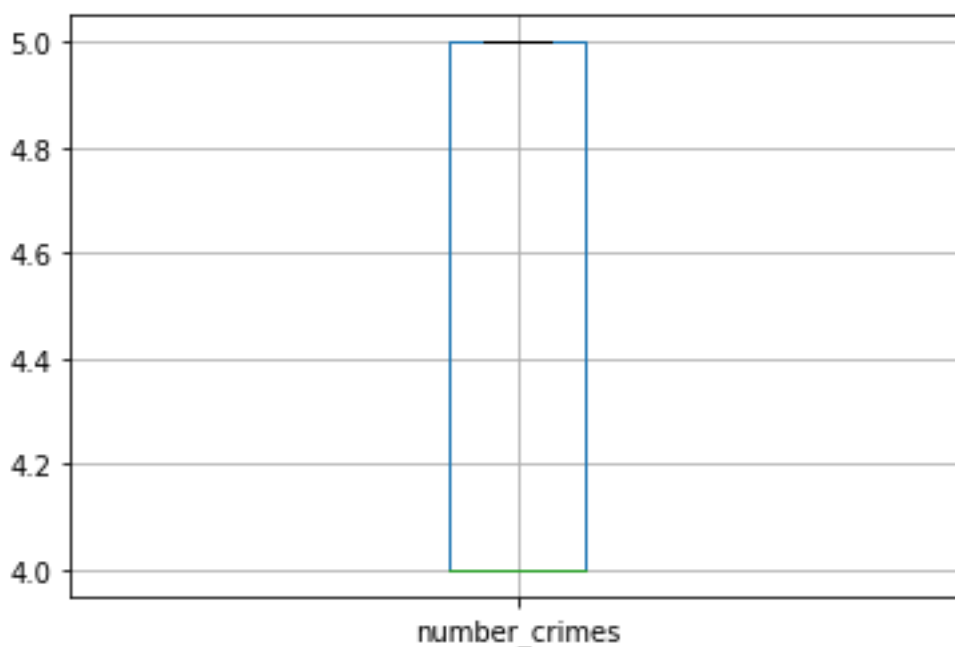


Figure 6. Cluster-0 box plot

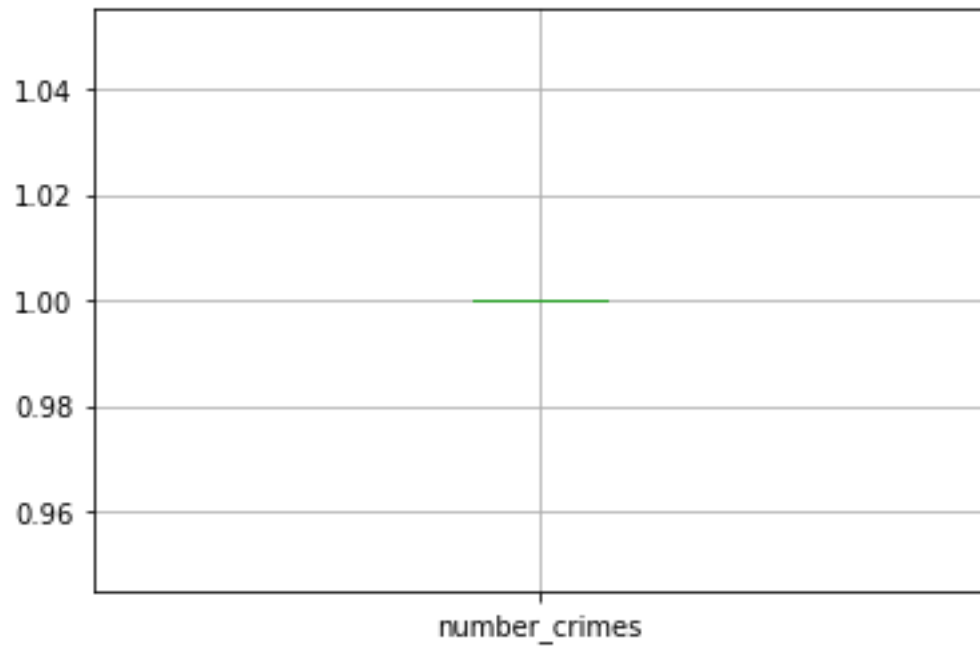


Figure 7. Cluster-1 box plot

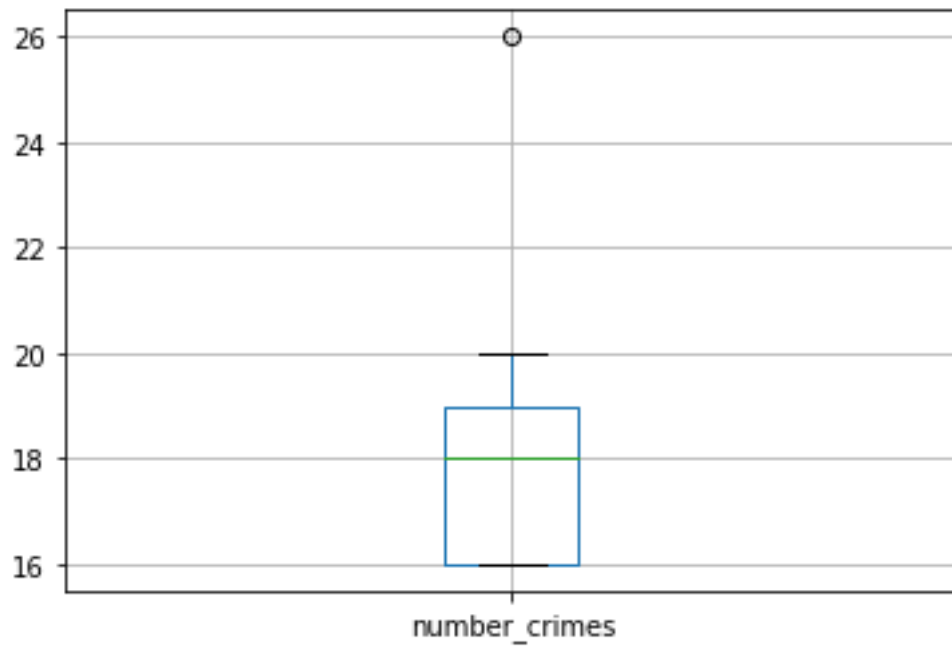


Figure 8. Cluster-2 box plot

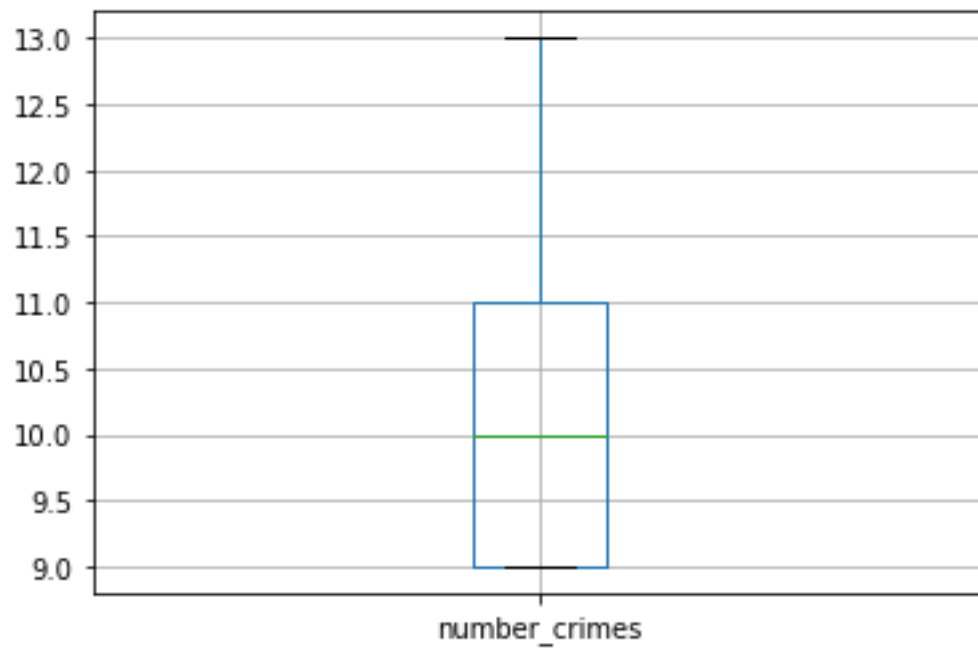


Figure 9. Cluster-3 box plot

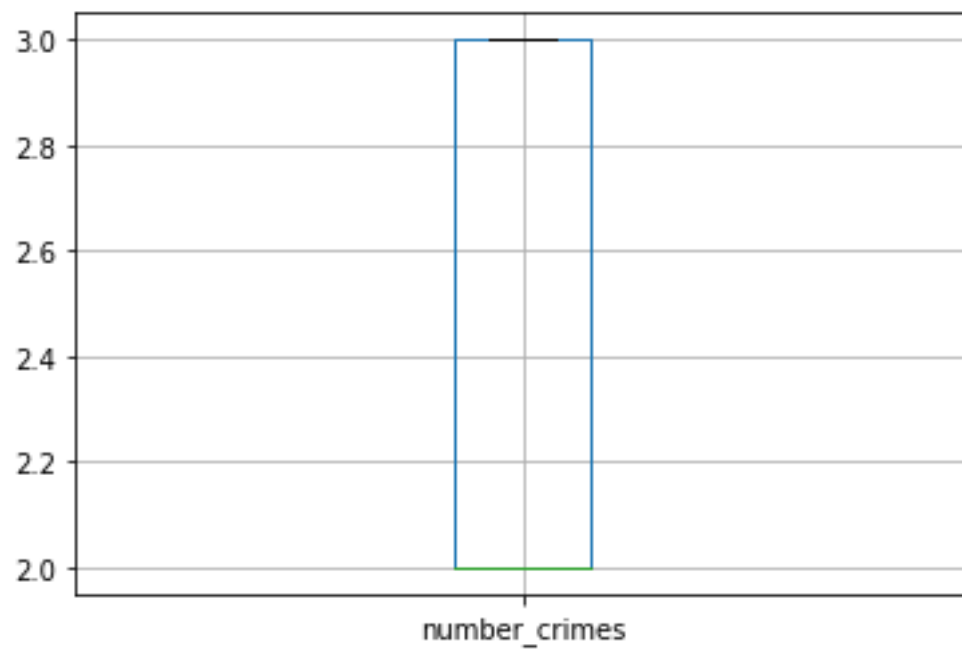


Figure 10. Cluster-4 box plot

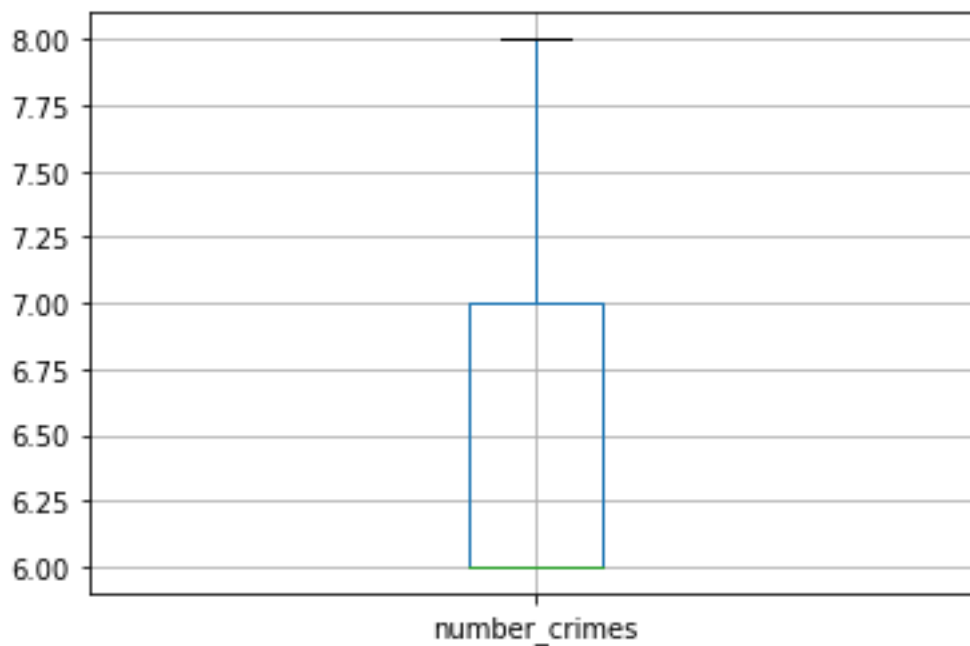


Figure 11. Cluster-5 box plot

The top-twenty high schools, sorted by descending number of crimes near each high school are as follows:

Top twenty most at-risk high schools in 2019

| | high_school_name | number_crimes | city |
|----|---|---------------|----------|
| 0 | The Brooklyn Latin School | 26 | Brooklyn |
| 1 | The Young Women's Leadership School of East Ha... | 20 | New York |
| 2 | The Heritage School | 19 | New York |
| 3 | Cristo Rey High School | 19 | New York |
| 4 | Magistri Lounge | 18 | New York |
| 5 | Central Park East 1 | 17 | New York |
| 6 | Williamsburg HS | 16 | Brooklyn |
| 7 | Williamsburg High School of Arts + Technology | 16 | Brooklyn |
| 8 | Central Park East High School | 16 | New York |
| 9 | Liberation High School | 13 | Brooklyn |
| 10 | Unity Preparatory Charter School - High School | 12 | Brooklyn |
| 11 | Grand Street Campus | 11 | Brooklyn |
| 12 | Academy For Young Writers | 11 | Brooklyn |
| 13 | Brownsville Academy | 11 | Brooklyn |
| 14 | High School For Enterprise, Business And Techn... | 11 | Brooklyn |
| 15 | El puente academy | 10 | Brooklyn |
| 16 | Park East High School | 10 | New York |
| 17 | Saint Joseph's High School | 10 | Brooklyn |
| 18 | Democracy Prep Endurance High | 10 | New York |
| 19 | Teachers Preparatory High School | 10 | Brooklyn |

Figure 12. Top-twenty high schools by level of crime nearby

5. RESULTS

The segmented high schools with crime nearby for 2019 can be visualised as follows:

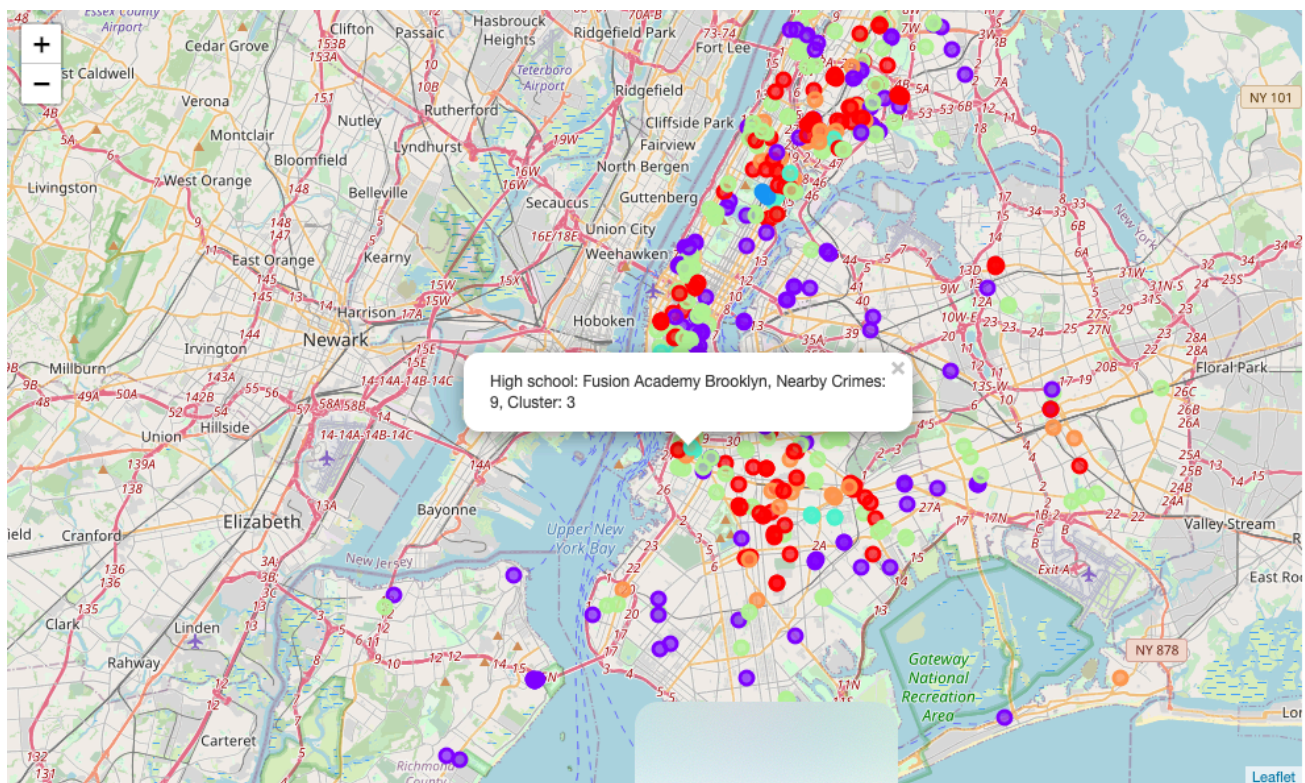


Figure 13. High schools segmented into 6 clusters

Map Keys

- Cluster 1 ——— 1 crime per school
- Cluster 4 ——— 2-3 crimes per school
- Cluster 0 ——— 4-5 crimes per school
- Cluster 5 ——— 6-8 crimes per school
- Cluster 3 ——— 9-13 crimes per school
- Cluster 2 ——— 16-26 crimes per school

The top-twenty high schools with the highest nearby crime can be visualised as follows:

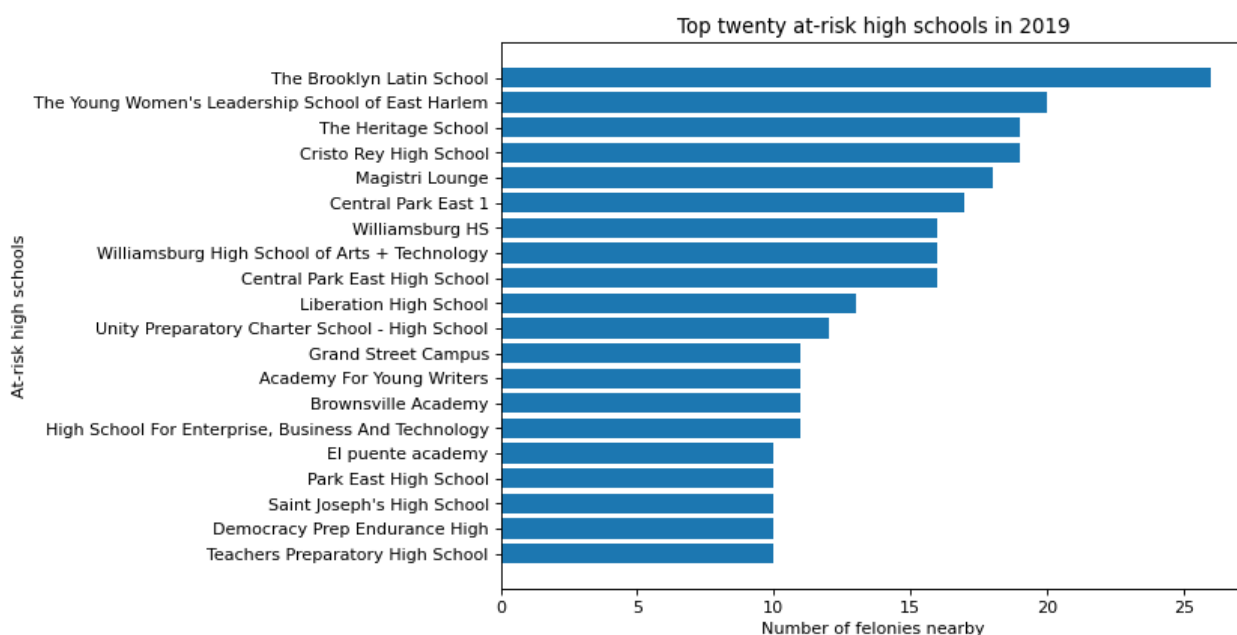


Figure 14. Top-twenty at-risk high schools in 2019

6. CONCLUSION

SUMMARY

The author set out to identify the top-twenty most at-risk high schools in New York state, using crime data from the NYPD and school location data from Foursquare and identify what areas of New York state are most affected. The audience that this research may be useful to is law enforcement, public bodies and social organisations that may be interested in using this research as input into decision-making relating to where to focus limited resources.

The author has identified 521 high schools in New York state that had at least one felony involving a dangerous weapon that was within a 500 m radius from the school in 2019. These schools were clustered using a K-means machine learning algorithm into six clusters and mapped these schools. The author then identified the top-twenty most at risk high schools based on number of felonies nearby. For the top-twenty schools, the number of felonies involving a dangerous weapon ranged from 10 to 26 felonies per school. Eight of these schools were in Brooklyn and 12 of these schools were in New York. The top 3 at risk high schools were: The Brooklyn Latin School, The Young Women's Leadership School of East Harlem and The Heritage School.

LIMITATIONS

The limitation of this research are:

1. Foursquare data may not have included all schools in New York state, so some at-risk schools may not have been identified.
2. The data relates to crimes in 2019, pre-Covid, and data for 2020 and 2021 is likely to have changed due to various forms of lock-down and restriction of movement.
3. The author attempted to capture crimes before and after school hours, but did not distinguish between school-days and non-school days, like weekends, public holidays and school holidays.
4. Due to time constraints, a thorough literature review was not carried out, nor was this research peer-reviewed.

FUTURE RESEARCH

A potential future direction for related research could be to consider only school-days and to compare this data over time to see how crime is changing over time. It might also be useful to include other crimes and then segment neighbourhoods by level of crime mix near schools.

7. APPENDIX

The complete analysis for this research can be found at:

https://github.com/anthonyalbertyn/Coursera_Capstone/blob/main/IBM-Data-Science-Capstone-Project-Anthony-Albertyn.ipynb

Note: GitHub will not display some of the maps, so to view the analysis in a user-friendly way, use this link instead:

https://nbviewer.jupyter.org/github/anthonyalbertyn/Coursera_Capstone/blob/main/IBM-Data-Science-Capstone-Project-Anthony-Albertyn.ipynb