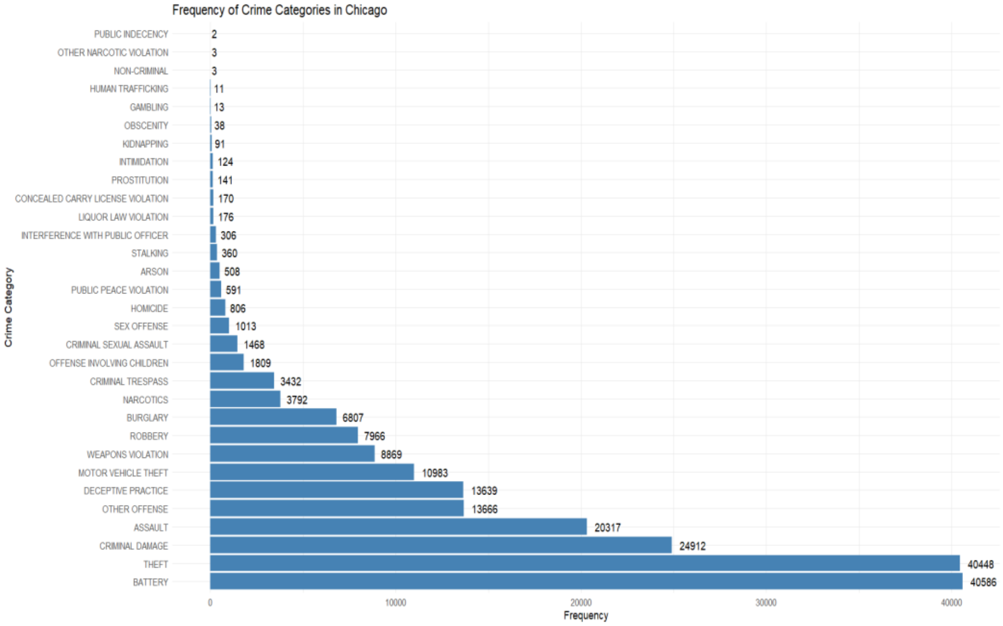
Project 5. Spatial Clustering and Structure of Crimes in the City of Chicago

Abstract

This project focused on data exploration of the crimes in Chicago, carious crime categories were implemented to find out the frequency and crimes and the temporal frequencies showed that summer tends to have more crimes than winter. In addition, there were more crimes occurred during midnight compared to morning. After that, he spatial clustering study of Chicago crimes (three key categories: Criminal Sexual Assault, Motor Vehicle Theft, and Robbery) showed the sensitivity of clustering outcomes to variations in DBSCAN parameters, with different algorithms such as the use of OPTICS and HDBSCAN. At last, this project analyzed the distance between each clusters and found that some crimes actually occurred in the same neighborhood.

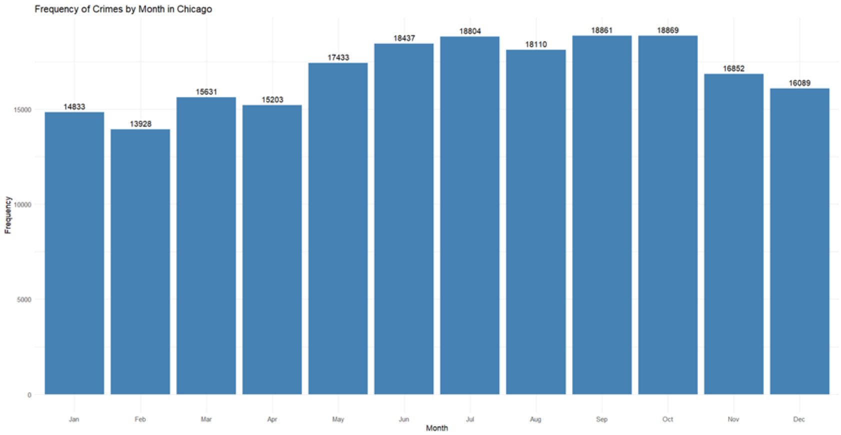
1. Task 01 Data Exploration

After implementing the data from the csv file provided by the professor, the bar chart shows the crimes by their frequency. While making the chart, the coordinates has been flipped as the x axis represents the frequency and the y axis represents crime category.



The bar chart indicates that Battery is the most frequent crime, with over 40,000 occurrences. Theft is a close second, also nearing 40,000 incidents. Additionally, two other crime categories—assault and criminal damage—exceed 20,000 occurrences. Conversely, several crimes have notably low frequencies, under 150. These include Prostitution, Intimidation, Kidnapping, Obscenity, Gambling, Human Trafficking, Non-Criminal, Other Narcotic Violation, and Public Indecency.

The bar chart below also shows the frequency of crimes in Chicago by Month.



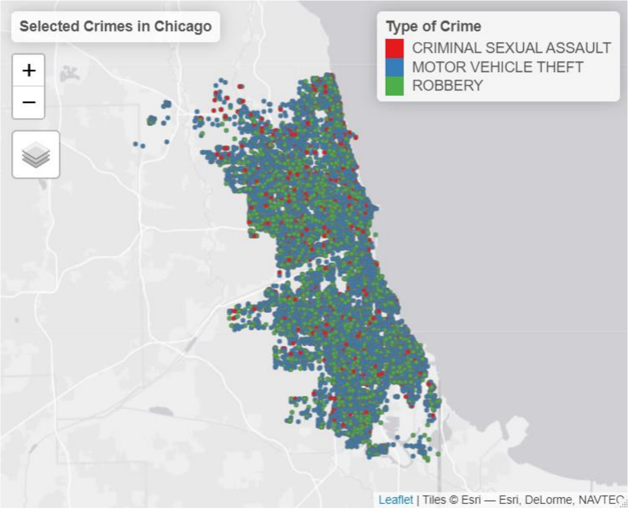
The city experiences a consistent level of crime throughout the year with monthly crime figures hovering between approximately 14,000 to 19,000 incidents. The lowest crime frequency is observed in February with 13,928 reported incidents. This could be attributed to the colder weather, which might deter some types of outdoor criminal activities. July witnesses the highest frequency of crimes with 18,847 incidents. Warmer months, especially summer, often see an uptick in various activities, including crime, as people are more likely to be outdoors.

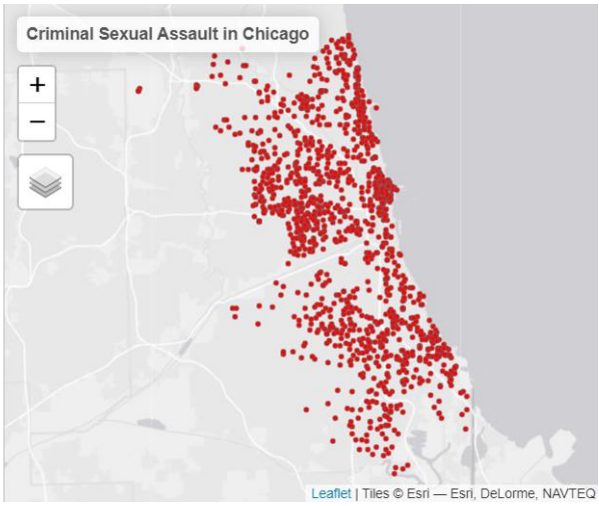
The bar chart of frequency of crimes by day of week in Chicago (Appendix) shows that the crimes are fairly evenly distributed across the days, with no single day showing a drastic difference in crime rate. This indicates that crime in Chicago is fairly consistent throughout the week.

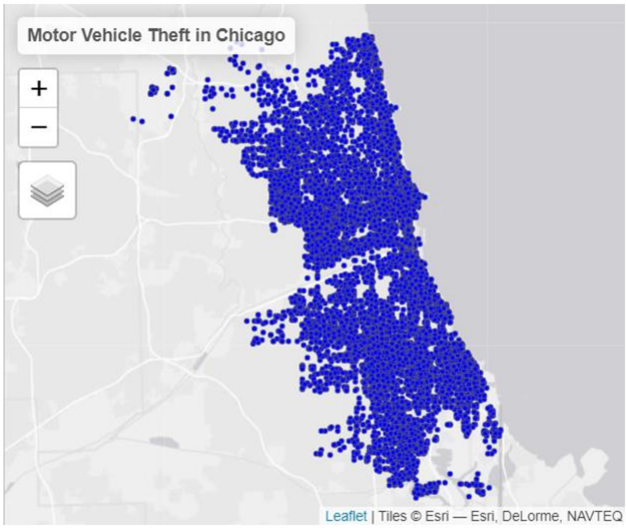
The bar chart of frequency of crimes by season in Chicago (Appendix) shows that summer has the highest frequency of crimes with 55,351 occurrences. This could be attributed to the warmer weather and increased outdoor activities, which might result in higher opportunities for crimes to take place. Autumn follows closely behind with 54,582 reported crimes. The close proximity of these numbers to the summer season might indicate sustained activity levels as the weather begins to cool. Winter has the lowest frequency of crimes with 44,850 incidents. The decrease in crime during this season might be due to colder temperatures, which could reduce the number of people outdoors, and thus potentially reduce the opportunities for certain crimes to occur.

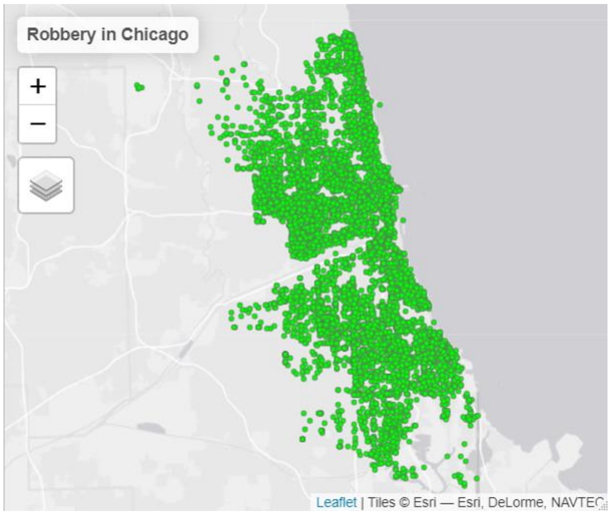
The bar chart of crime frequency in Chicago by hour (Appendix) shows that the hour with the highest frequency of crime is at midnight (0 hour) with a total of 13,637 incidents. This could be linked to various factors such as late-night activities, reduced visibility, or fewer witnesses. Starting from 10 AM, the frequency of crime begins to climb and remains relatively consistent throughout the afternoon and evening hours, fluctuating between approximately 8,500 to 11,000 incidents. This consistency could be attributed to regular daytime and evening activities when the city is most active.

Based on the exploration, three categories of crime were selected for the future analysis. The following maps shows the distribution of three selected crimes: Robbery, Motor Vehicle Theft, and Criminal Sexual Assault. The following maps shows the distribution.


Selected Crimes in Chicago
Type of Crime
🟥 CRIMINAL SEXUAL ASSAULT
🟦 MOTOR VEHICLE THEFT
🟩 ROBBERY
+
−
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ


Criminal Sexual Assault in Chicago
+
−
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

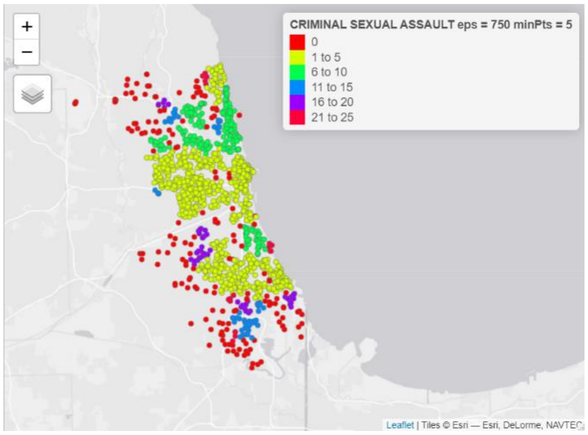

Motor Vehicle Theft in Chicago
+
−
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

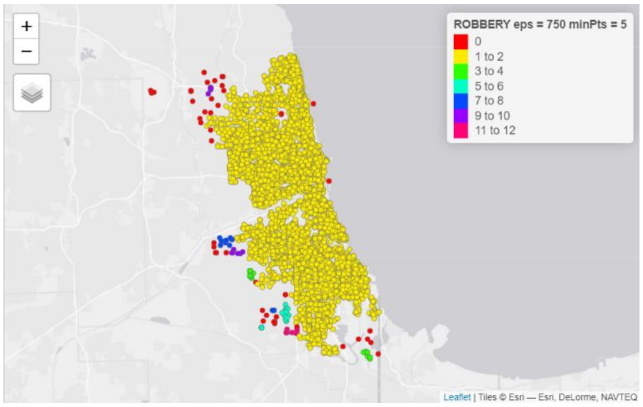


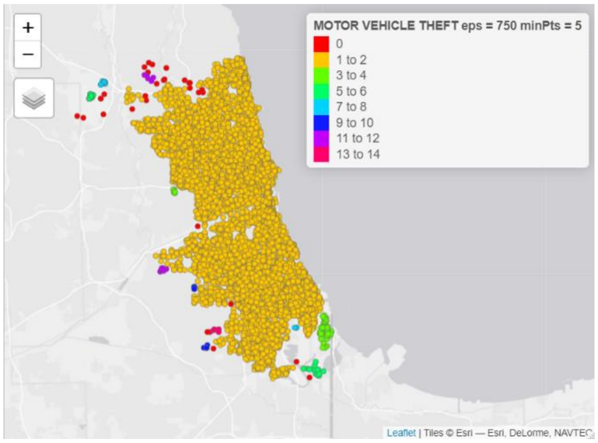
Robbery in Chicago appears to be fairly distributed throughout the city; however, there are pronounced hotspots where the density of reports is higher. The western region and the northeastern corner seem to have a higher concentration of cases compared to other areas. As for motor vehicle theft, both the southern and northern parts of Chicago display a significant number of incidents, somewhat mirroring the distribution pattern of robberies. In contrast, when comparing the south to the north side of Chicago, the southern area has fewer instances of criminal sexual assault and robbery. Furthermore, it's evident that the northwestern region of the city tends to have the fewest reported cases of these crimes.

2. Task 02 Spatial Clustering

The DBSCAN was initially used for the spatial clustering analysis for three types of crime. During this process, the "eps" also stands for epsilon means the distance between two points for one to be considered as in the neighborhood of other. The minPts represents the minimum number of points required to form a dense region. For a point to be considered a core point (a central point in a cluster), it should have at least "minPts" number of points within the "eps" distance. Different parameters setting will have different visualizations. For example, the following figures shows the DBSCAN clustering of parameters setting epsilon equals 750 meters (converted from degrees). And minimum point of 5 (rest of the clustering map showing in Appendix).


CRIMINAL SEXUAL ASSAULT eps = 750 minPts = 5
0
1 to 5
6 to 10
11 to 15
16 to 20
21 to 25
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

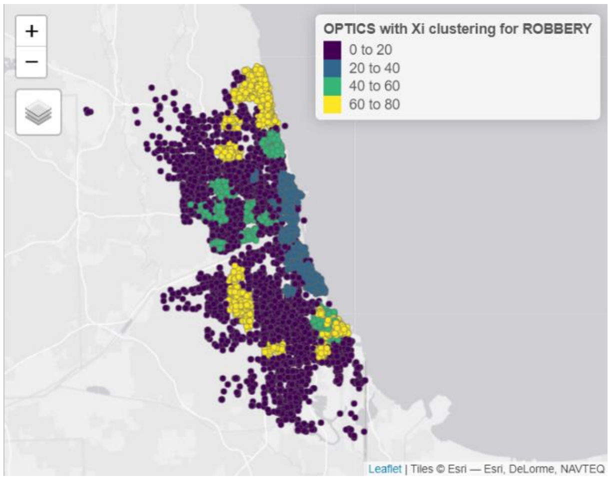

ROBBERY eps = 750 minPts = 5
0
1 to 2
3 to 4
5 to 6
7 to 8
9 to 10
11 to 12
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

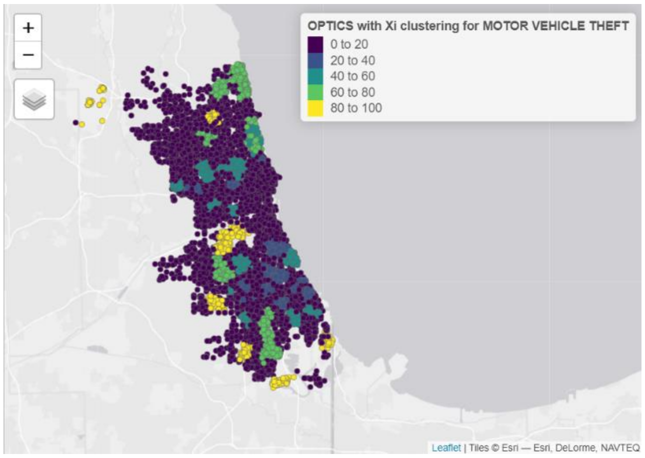

MOTOR VEHICLE THEFT eps = 750 minPts = 5
0
1 to 2
3 to 4
5 to 6
7 to 8
9 to 10
11 to 12
13 to 14
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

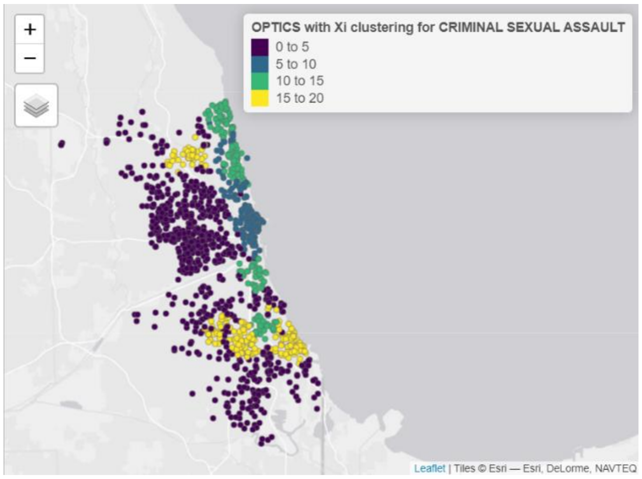
As observed from the maps, each dot represents a crime incident, and their color indicates the density or the number of incidents nearby. It's evident that both the Motor Vehicle Theft and Robbery maps have a large area dominated by yellow (gold) dots. This suggests that, under these parameters, criminal cases aren't in close proximity to other cases, preventing the formation of a cluster. Conversely, the Criminal Sexual Assault map, with the same parameter settings, displays a more varied color distribution, indicating a wider range of crime densities. These settings provide a clearer depiction for people and city authorities to identify and address crime hotspots. Maps with different parameter settings are provided in the Appendix.

OPTICS

In addition to DBSCAN, OPTICS was employed for further analysis. During this analysis, the minimum point was set to 30. The subsequent figures display the spatial clustering of the three mentioned crimes using OPTICS.


OPTICS with Xi clustering for ROBBERY
0 to 20
20 to 40
40 to 60
60 to 80
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ


OPTICS with Xi clustering for MOTOR VEHICLE THEFT
0 to 20
20 to 40
40 to 60
60 to 80
80 to 100
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

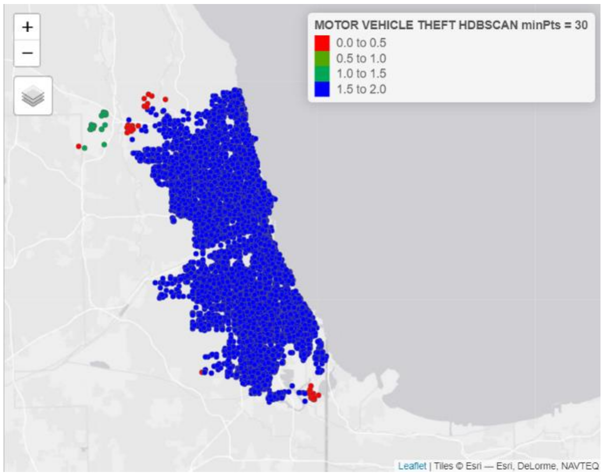

OPTICS with Xi clustering for CRIMINAL SEXUAL ASSAULT
0 to 5
5 to 10
10 to 15
15 to 20
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

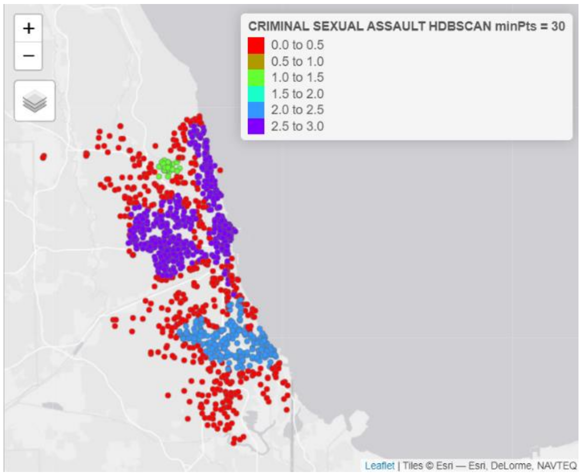
For the Motor Vehicle Theft, most of the Chicago region shows most areas (purple dots) of motor vehicle thefts are in 0 to 20 occurrences. Yellow dots (with 80 to 100 occurrences) are mostly located at the edge of southern Chicago. In contrast, we can see yellow dots (and gree) located in the northern (also southern) and central Chicago areas. It shows that robbery occurred more in the northern area as compared to the crime of motor vehicle theft. For Criminal Sexual Assault, it is interesting to see that the eastern edge and southern Chicago has more cases compared to the western areas. Most people would believe that north and east Chicago would be "safer" compared to the southern (where has a lot of yellows) and western side. This can probably relate to that the eastern side of Chicago has higher population density, another reason can be that residents in east and north Chicago would have more reporting rate once such crimes occurred.

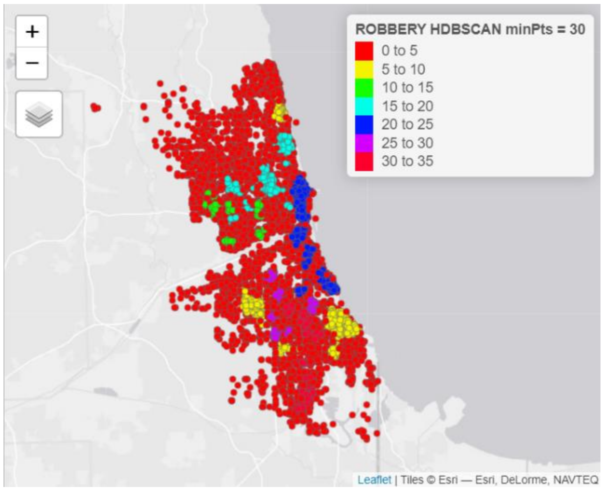
HDBSCAN

The next step is to conduct HDBSCAN for the spatial clustering. HDBSCAN builds upon the concept of DBSCAN but introduces a hierarchical approach. It examines the density levels at various scales, allowing clusters of varying densities to be identified. The following map shows the spatial clustering of HDBSCAN.

.


MOTOR VEHICLE THEFT HDBSCAN minPts = 30
0.0 to 0.5
0.5 to 1.0
1.0 to 1.5
1.5 to 2.0
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ


CRIMINAL SEXUAL ASSAULT HDBSCAN minPts = 30
0.0 to 0.5
0.5 to 1.0
1.0 to 1.5
1.5 to 2.0
2.0 to 2.5
2.5 to 3.0
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ


ROBBERY HDBSCAN minPts = 30
0 to 5
5 to 10
10 to 15
15 to 20
20 to 25
25 to 30
30 to 35
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

For motor vehicle thefts, it's evident that incidents occur across almost all regions of Chicago. However, the airport area appears to have fewer cases compared to other urban areas in the city. In the case of criminal sexual assault, the clustering suggests a clear demarcation, almost like a horizontal line, splitting Chicago into northern and southern regions. Notably, the northern part of Chicago seems to have a higher number of such cases than the southern areas. The robbery map is the most diverse of the three, displaying the highest number of clusters scattered across various regions of the city. Despite this diversity, it's still discernible that the eastern edge of Chicago has a higher concentration of cases compared to the western edge.

While performing DBSCAN, the algorithm does not explicitly determine the number of clusters. The outcome depends on the data and the choice of parameters. Thus, during the DBSCAN process, multiple epsilon values were tested to determine the best parameters. In contrast, HDBSCAN dynamically determines the number of clusters based on the hierarchical structure of the data. Comparing the results, it's evident that the Motor Vehicle Theft maps are almost identical for both algorithms. Similarly, for Criminal Sexual Assault, the distribution of clusters is largely consistent, with only minor differences in clustering. However, for the Robbery map, clear distinctions are noticeable in the southern Chicago area between the DBSCAN and HDBSCAN results. Another observation is that OPTICS tends to identify more clusters for Motor Vehicle Theft compared to both HDBSCAN and DBSCAN, offering a more detailed visualization of the cases (with HDBSCAN being more detailed than DBSCAN). Comparing these three methods, we can conclude that

HDBSCAN excels in certain areas, such as detecting clusters of varying densities due to its hierarchical approach. OPTICS also effectively identifies clusters of different densities, leveraging its reachability plot, which can be segmented at various levels to discern clusters of varying densities.

3. Spatial Structure

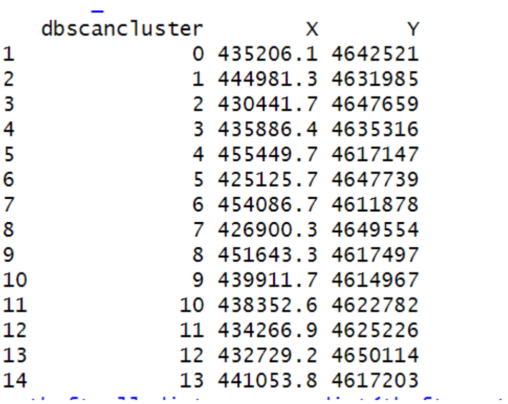
Find out Centroids

The initial step is to identify the centroid of each cluster. For this purpose, I utilized the DBSCAN clustering method. By examining the spatial coordinates of the crime data, the DBSCAN algorithm effectively groups them into clusters. Once these clusters are determined, the centroid (the geometric center) of each cluster is computed by averaging the X and Y coordinates of all points within that cluster.

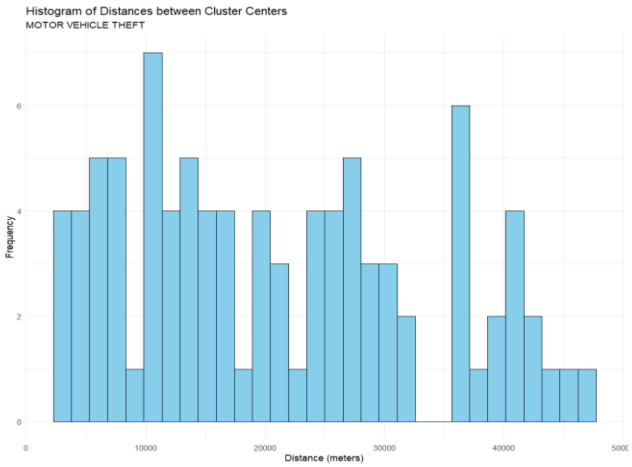
Identifying the Closest Centroids

Having determined the centroids, the next step is to identify the pairs that are closest to each other. To achieve this, I implemented a loop that systematically iterates through each possible pair of centroids. For each pair, the Euclidean distance—a standard measure of spatial distance—is computed. By comparing these distances, I was able to identify and highlight the pair of centroids that were the closest.

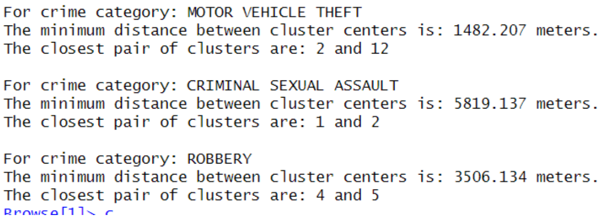
The following screenshot provides an illustrative example of the X and Y coordinates derived from the DBSCAN clusters. After obtaining these coordinates, the distances between the closest centroids are computed. Furthermore, a histogram is presented to show the frequency distribution of distances between these cluster centers. Comprehensive details of the DBSCAN cluster coordinates for other categories, as well as additional distance frequency histograms, are listed in the Appendix.



```
   dbscancluster        X       Y
1              0 435206.1 4642521
2              1 444981.3 4631985
3              2 430441.7 4647659
4              3 435886.4 4635316
5              4 455449.7 4617147
6              5 425125.7 4647739
7              6 454086.7 4611878
8              7 426900.3 4649554
9              8 451643.3 4617497
10             9 439911.7 4614967
11            10 438352.6 4622782
12            11 434266.9 4625226
13            12 432729.2 4650114
14            13 441053.8 4617203
```



The screenshot below shows the minimum distance between cluster centers. For motor vehicle theft, the minimum distance is 1,482 meters, for criminal sexual assault, the minimum distance is 5,819 meters, and for robbery, the closest distance is 3,506 meters. A key notice there is that the parameter setting can significantly affect the result of closest distance. Thus, the result can be affected by the setting of epsilon and minimum points.

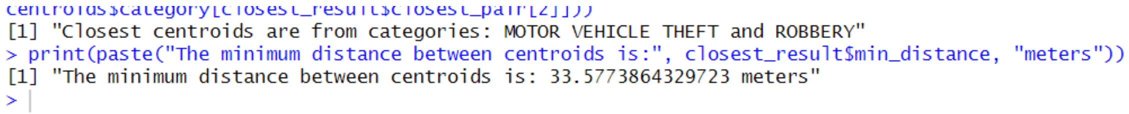


```
For crime category: MOTOR VEHICLE THEFT
The minimum distance between cluster centers is: 1482.207 meters.
The closest pair of clusters are: 2 and 12

For crime category: CRIMINAL SEXUAL ASSAULT
The minimum distance between cluster centers is: 5819.137 meters.
The closest pair of clusters are: 1 and 2

For crime category: ROBBERY
The minimum distance between cluster centers is: 3506.134 meters.
The closest pair of clusters are: 4 and 5
Browse[1]> c
```

For finding the distance between closet cluster center of all three crime categories. The first step is to find the distance between clusters. The coordinates were converted into matrix, during this process, the zero distance will be removed. After that the paired distance will be calculated and then find out the minimum distance between two categories. The screenshot below shows that the closest distance of two clusters are between a motor vehicle theft and robbery. The minimum distance is 33.58 meters.



```
centroids$category[closest_result$closest_pair[2]]))
[1] "Closest centroids are from categories: MOTOR VEHICLE THEFT and ROBBERY"
> print(paste("The minimum distance between centroids is:", closest_result$min_distance, "meters"))
[1] "The minimum distance between centroids is: 33.5773864329723 meters"
>
```

To test out the hypothesis: “different types of crimes in Chicago are actually also clustered. or mixed, i.e., they actually happen in the same neighborhood.” It is needed to understand the minimum distance we found from this task. In the first part of the minimum distance is between the same category. Meaning that the minimum distance between vehicle theft cluster 1 and vehicle cluster 2. The second part of minimum distance we found is the closest distance of any two clusters from these three clusters we found. The fact that robbery and vehicle theft clusters can be as close as 33 meters suggests a strong spatial overlap between where these crimes occur. This distance is significantly smaller than the minimum distance between clusters within the "Robbery" or “Motor Vehicle Theft” category itself (1,400, 5,800 meters). This finding supports the hypothesis that different types of crimes (criminal sexual assault, robbery, and vehicle theft) can happen in the same or closely neighboring locations.

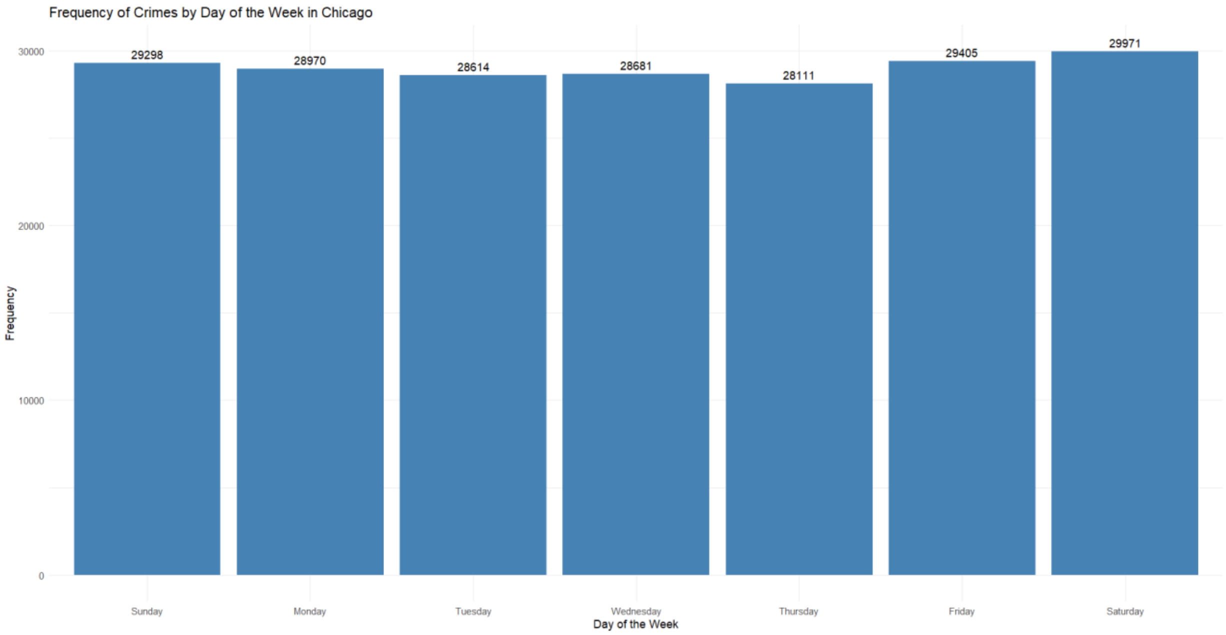
4. Summary/Conclusion/Concluding Remarks

In this project, we conducted a spatial clustering analysis of crimes in Chicago, focusing on three selected categories: Criminal Sexual Assault, Motor Vehicle Theft, and Robbery. The results revealed that varying the parameters of DBSCAN can lead to significantly different clusters. Additionally, the use of OPTICS and HDBSCAN also yielded distinct spatial clustering outcomes. For the third task, it was observed that the minimum distance between two clusters from different crime categories is notably smaller than the minimum distance between two clusters of the same crime category. This observation supports the hypothesis that these crimes often occur within the same neighborhoods.

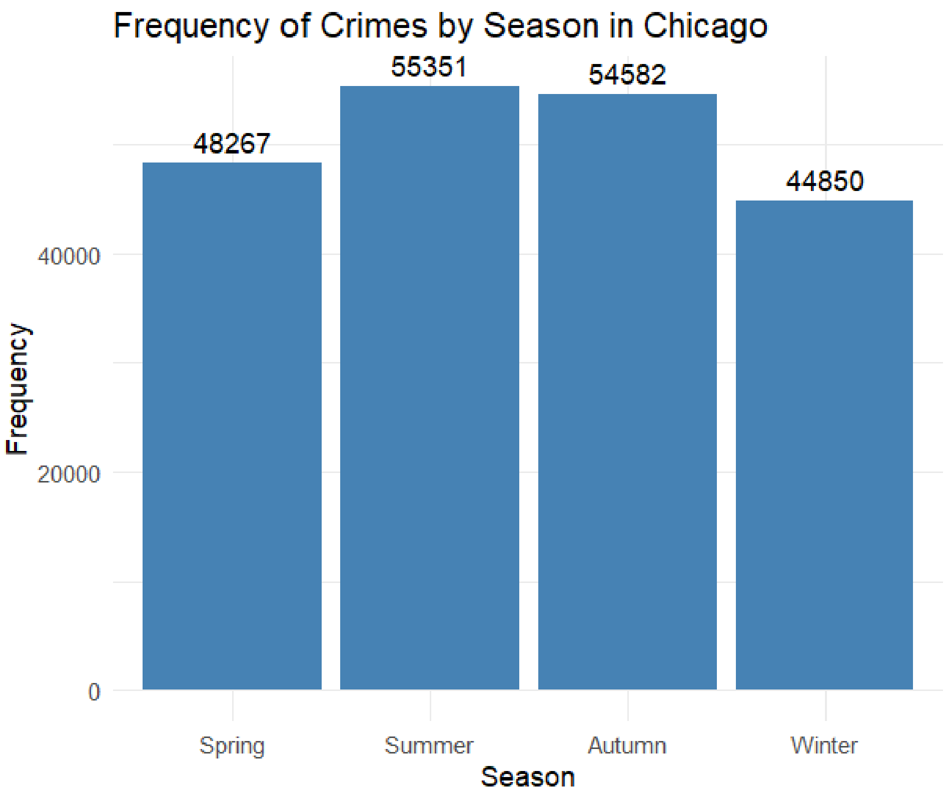
Acknowledgement

References

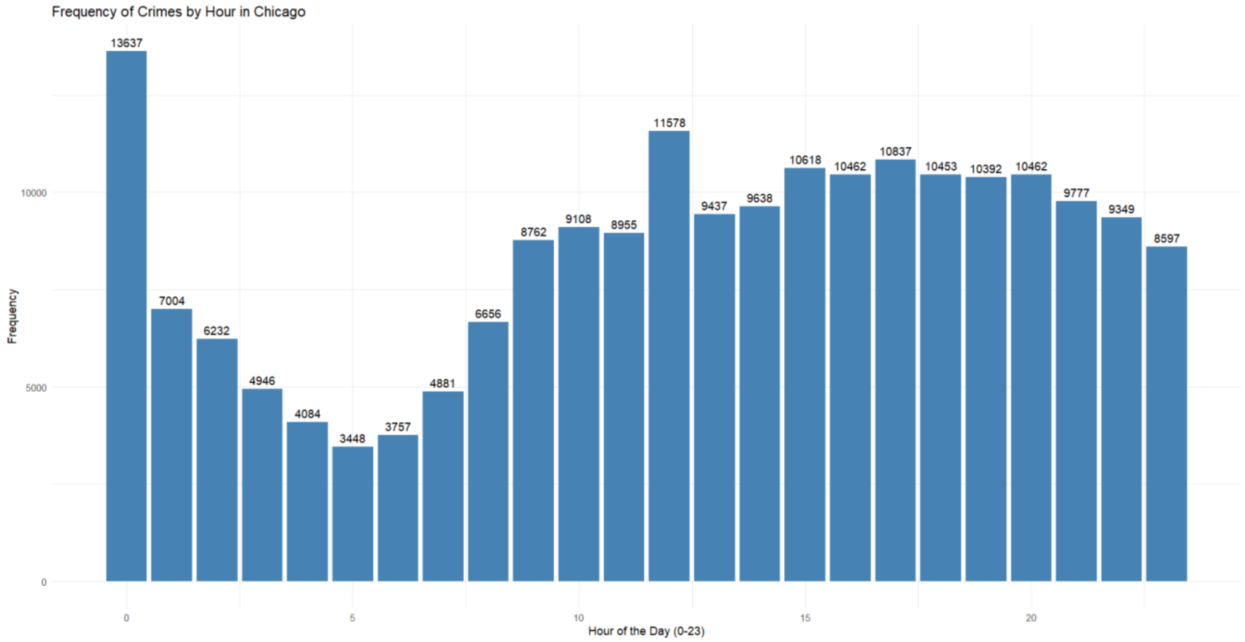
Day of Week Crime Frequency Chicago



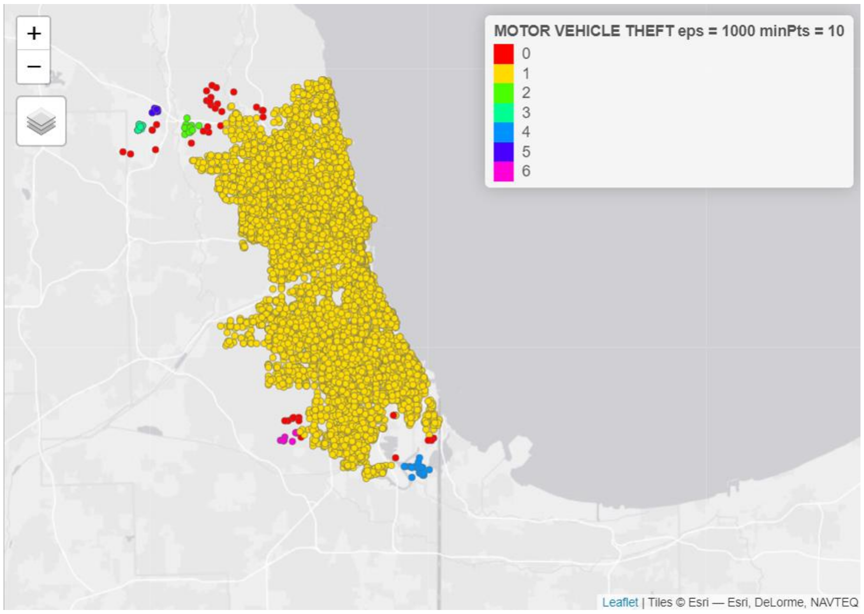
Season Crime Frequency in Chicago

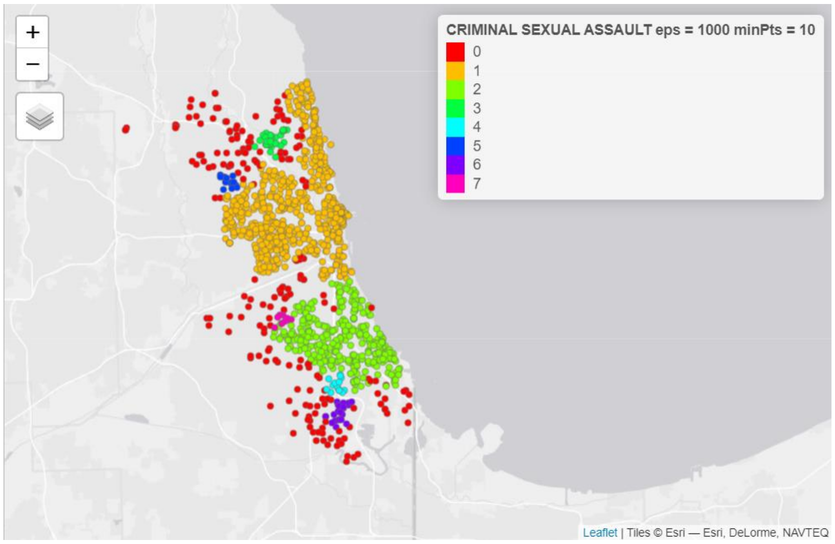


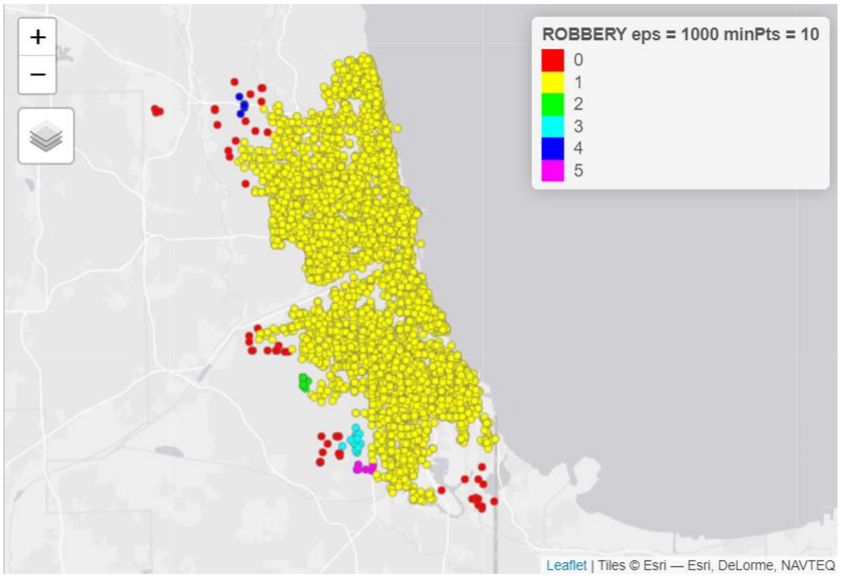
Hour Crime Chicago

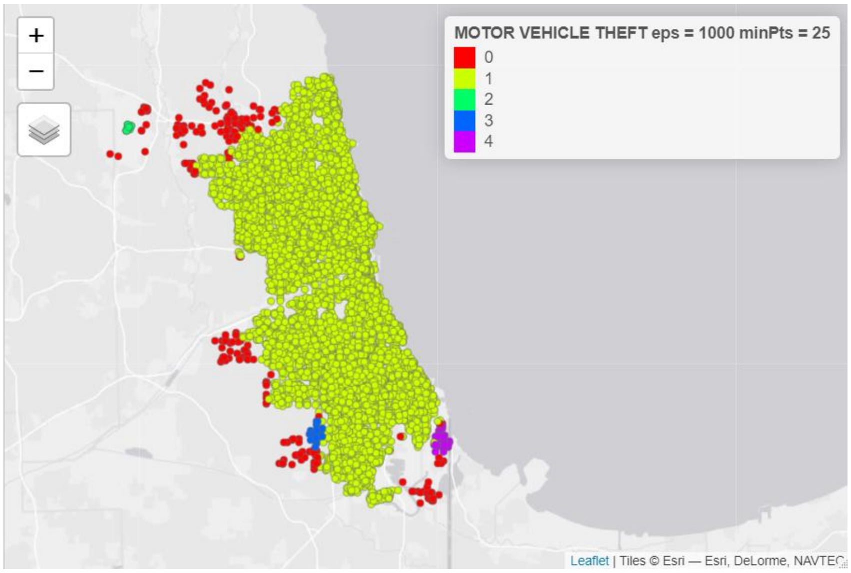


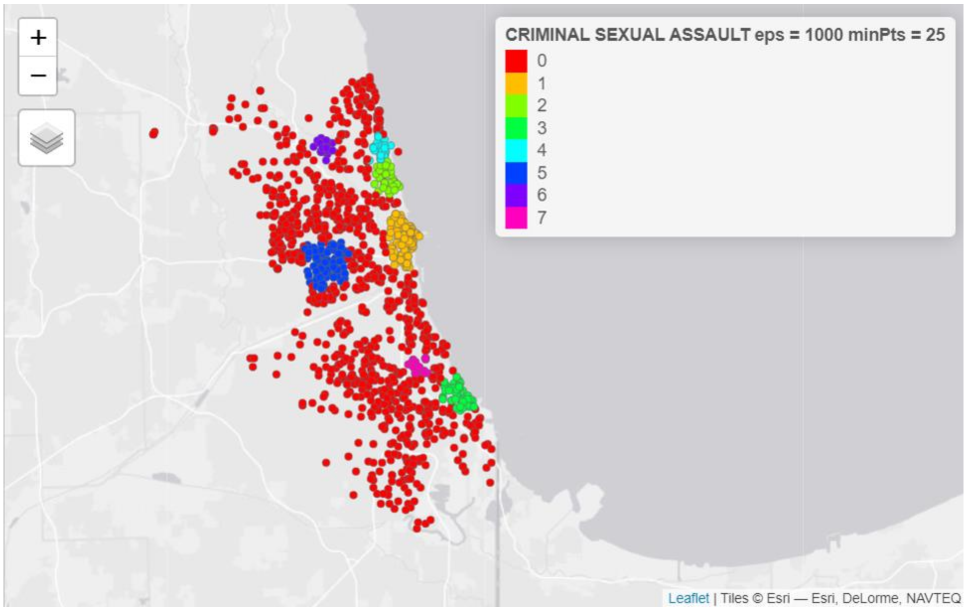
DBSCAN

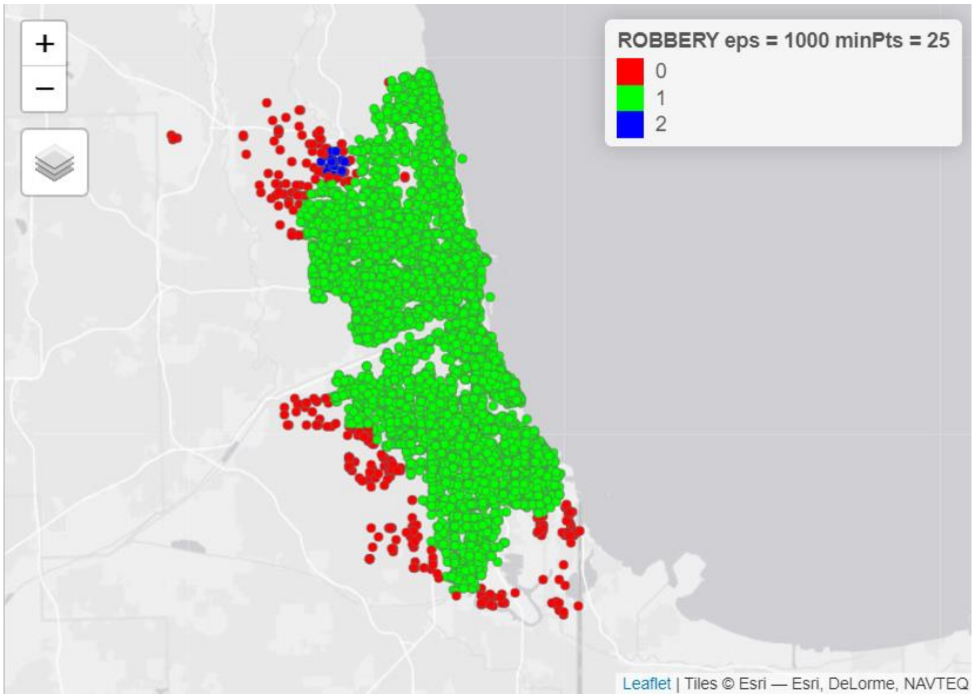

MOTOR VEHICLE THEFT eps = 1000 minPts = 10
0
1
2
3
4
5
6
+
−
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ


+
−
CRIMINAL SEXUAL ASSAULT eps = 1000 minPts = 10
0
1
2
3
4
5
6
7
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

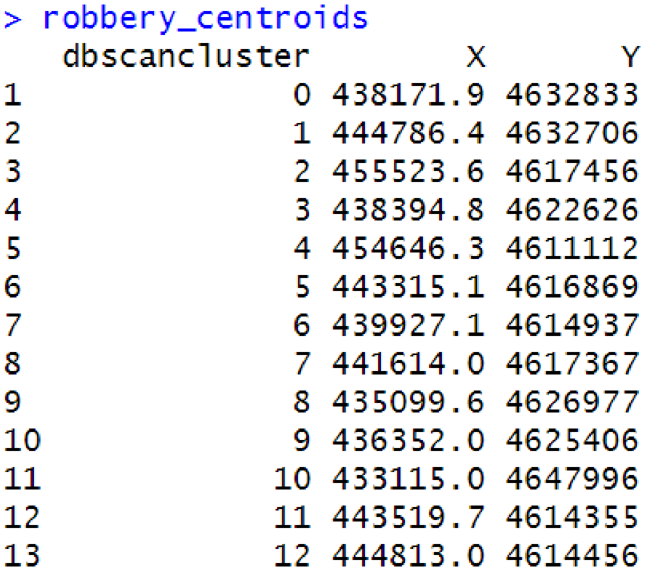

+
−
ROBBERY eps = 1000 minPts = 10
0
1
2
3
4
5
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ


+
−
MOTOR VEHICLE THEFT eps = 1000 minPts = 25
0
1
2
3
4
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

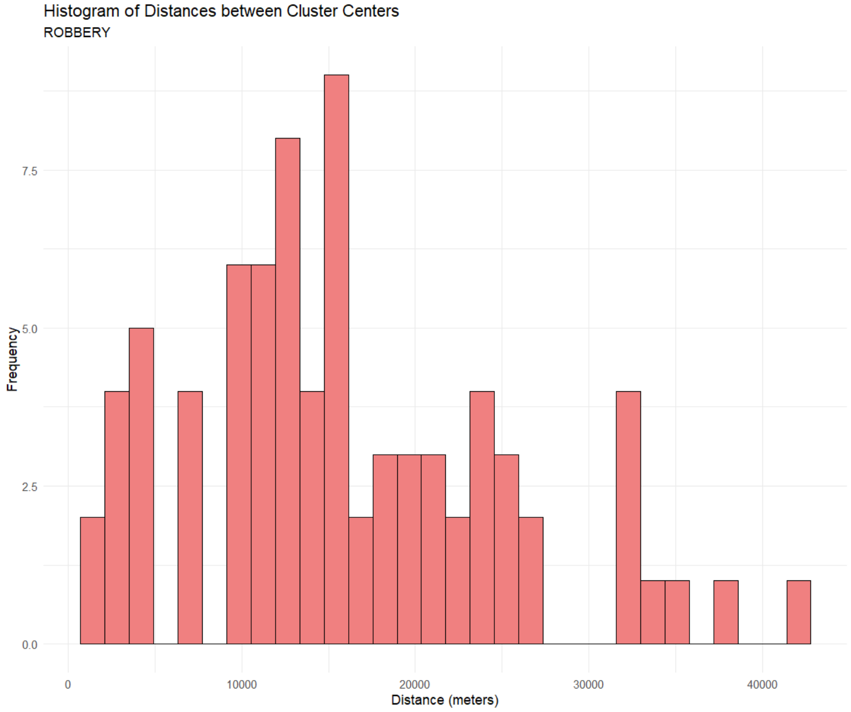

+
−
CRIMINAL SEXUAL ASSAULT eps = 1000 minPts = 25
🟥 0
🟧 1
🟩 2
🟩 3
🟦 4
🟦 5
🟪 6
🟪 7
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

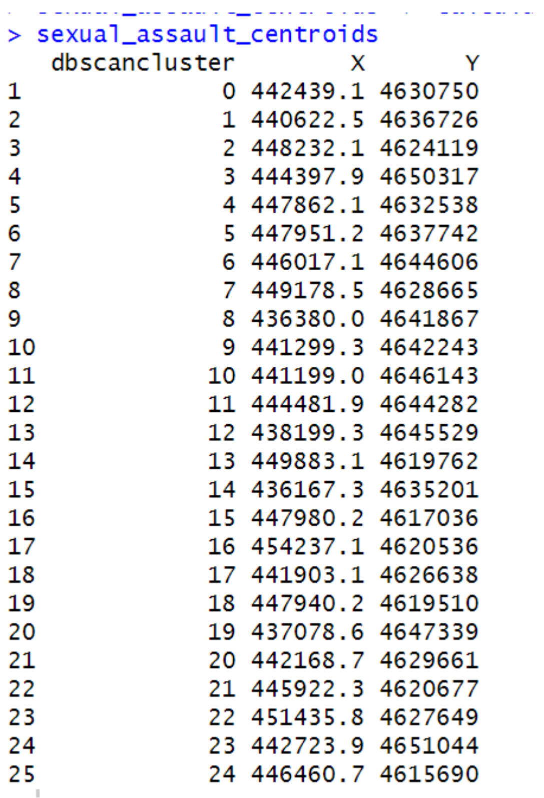

+
−
ROBBERY eps = 1000 minPts = 25
🟥 0
🟩 1
🟦 2
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

DBSCAN Cluster X/Y Coordinates

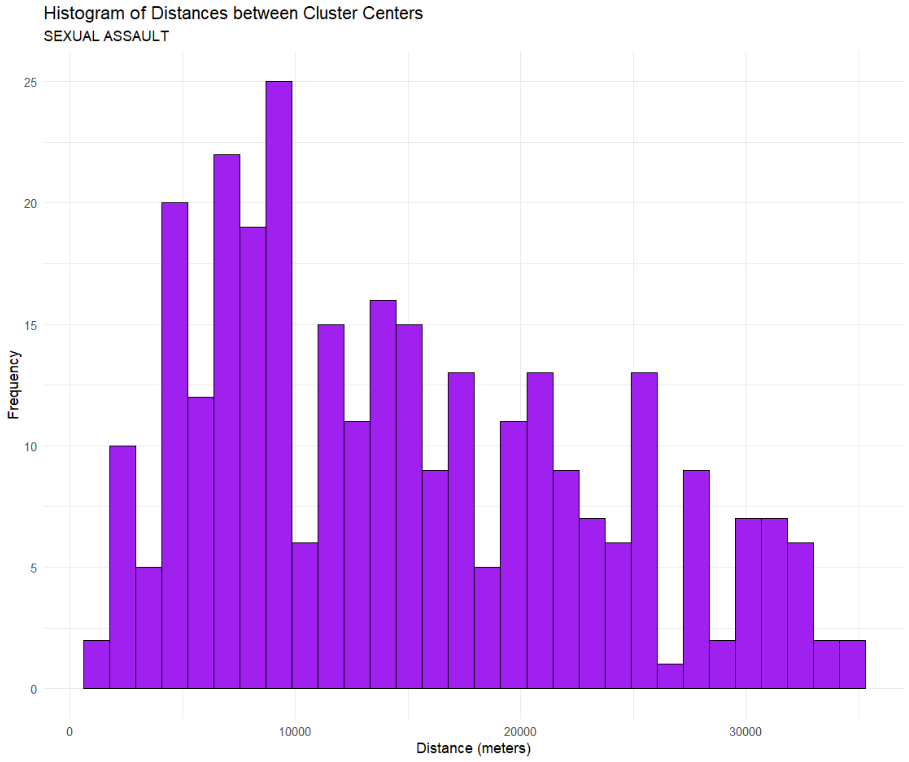


```
> robbery_centroids
   dbscancluster        X       Y
1              0 438171.9 4632833
2              1 444786.4 4632706
3              2 455523.6 4617456
4              3 438394.8 4622626
5              4 454646.3 4611112
6              5 443315.1 4616869
7              6 439927.1 4614937
8              7 441614.0 4617367
9              8 435099.6 4626977
10             9 436352.0 4625406
11            10 433115.0 4647996
12            11 443519.7 4614355
13            12 444813.0 4614456
```





```
> sexual_assault_centroids
   dbscancluster        X       Y
1              0 442439.1 4630750
2              1 440622.5 4636726
3              2 448232.1 4624119
4              3 444397.9 4650317
5              4 447862.1 4632538
6              5 447951.2 4637742
7              6 446017.1 4644606
8              7 449178.5 4628665
9              8 436380.0 4641867
10             9 441299.3 4642243
11            10 441199.0 4646143
12            11 444481.9 4644282
13            12 438199.3 4645529
14            13 449883.1 4619762
15            14 436167.3 4635201
16            15 447980.2 4617036
17            16 454237.1 4620536
18            17 441903.1 4626638
19            18 447940.2 4619510
20            19 437078.6 4647339
21            20 442168.7 4629661
22            21 445922.3 4620677
23            22 451435.8 4627649
24            23 442723.9 4651044
25            24 446460.7 4615690
```


Histogram of Distances between Cluster Centers
SEXUAL ASSAULT
Frequency
Distance (meters)