CIS 530 - Advanced Data Mining

CRIME HOTSPOT DETECTION USING CLUSETING (CHICAGO)

**Data:**

[https://data.cityofchicago.org/](https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about_data)

**Code:**

# ---------------------------

# Step 1: Install & Load Packages

# ---------------------------

# Install required packages if not already installed

packages <- c("tidyverse", "lubridate", "ggplot2", "cluster",

"factoextra", "dendextend", "leaflet", "leaflet.extras")

install.packages(setdiff(packages, rownames(installed.packages())))

# Load libraries

library(tidyverse)

library(lubridate)

library(ggplot2)

library(cluster)

library(factoextra)

library(dendextend)

library(leaflet)

library(leaflet.extras)

library(readr)

# ---------------------------

# Step 2: Load & Explore Dataset

# ---------------------------

# Load dataset

crime\_data <- read\_csv("Crimes\_2025.csv")

# View structure

head(crime\_data)

glimpse(crime\_data)

# ---------------------------

# Step 3: Preprocess Data

# ---------------------------

# Select key columns and remove missing data

crime\_clean <- crime\_data %>%

select(Date, `Primary Type`, Latitude, Longitude, `Community Area`) %>%

drop\_na()

# Convert Date and create derived time variables

# Convert 'Date' column and filter last 5 years

crime\_clean <- crime\_data %>%

select(Date, `Primary Type`, Latitude, Longitude, `Community Area`) %>%

drop\_na() %>%

mutate(

Date = mdy\_hms(Date),

Hour = hour(Date),

Month = month(Date),

Year = year(Date), # Extract year

Crime\_Type\_Code = as.numeric(as.factor(`Primary Type`))

) %>%

filter(Date >= as.Date(Sys.Date()) - years(5)) # Only include last 5 years

# ---------------------------

# Step 4: K-Means Clustering

# ---------------------------

# Prepare clustering data

kmeans\_data <- crime\_clean %>%

select(Latitude, Longitude, Crime\_Type\_Code)

# Elbow method on a 5,000-point sample to find optimal k

set.seed(42)

sample\_data <- kmeans\_data %>% sample\_n(5000)

fviz\_nbclust(sample\_data, kmeans, method = "wss") +

geom\_vline(xintercept = 6, linetype = 2) +

labs(title = "Elbow Method - Optimal k for K-Means (Sample)")

# Run K-Means on full dataset with k = 6

set.seed(123)

kmeans\_result <- kmeans(kmeans\_data, centers = 6, nstart = 25)

crime\_clean$KMeans\_Cluster <- as.factor(kmeans\_result$cluster)

# ---------------------------

# Step 5: K-Means Visualization (Leaflet)

# ---------------------------

leaflet(data = crime\_clean) %>%

addTiles() %>%

addCircleMarkers(

lng = ~Longitude,

lat = ~Latitude,

color = ~KMeans\_Cluster,

radius = 2,

stroke = FALSE,

fillOpacity = 0.5,

popup = ~paste("Crime Type:", `Primary Type`)

)

# ---------------------------

# Step 6: Hierarchical Clustering

# ---------------------------

# Sample 500 records for efficiency

set.seed(42)

hier\_data <- kmeans\_data %>% sample\_n(500)

# Compute distance matrix and hierarchical clustering

dist\_matrix <- dist(hier\_data)

hc <- hclust(dist\_matrix, method = "ward.D2")

# Plot dendrogram

plot(hc, labels = FALSE, hang = -1, main = "Dendrogram - Hierarchical Clustering")

abline(h = 150, col = "red", lty = 2)

# Cut into 6 clusters and add to data

hc\_clusters <- cutree(hc, k = 6)

hier\_data$HCluster <- as.factor(hc\_clusters)

# Visualize clusters

ggplot(hier\_data, aes(x = Longitude, y = Latitude, color = HCluster)) +

geom\_point(alpha = 0.7) +

theme\_minimal() +

labs(title = "Hierarchical Clustering (500 Sample Points)",

x = "Longitude", y = "Latitude")

# Cophenetic correlation evaluation

coph <- cophenetic(hc)

cor\_val <- cor(dist\_matrix, coph)

print(paste("Cophenetic correlation coefficient:", round(cor\_val, 3)))

# ---------------------------

# Step 7: Additional Visualizations

# ---------------------------

# Top 10 Crime Types

crime\_clean %>%

count(`Primary Type`) %>%

arrange(desc(n)) %>%

top\_n(10) %>%

ggplot(aes(x = reorder(`Primary Type`, n), y = n)) +

geom\_col(fill = "steelblue") +

coord\_flip() +

labs(title = "Top 10 Crime Types in 2020", x = "Crime Type", y = "Count") +

theme\_minimal()

# Crime Trend by Hour

crime\_clean %>%

count(Hour) %>%

ggplot(aes(x = Hour, y = n)) +

geom\_line(color = "tomato", size = 1.2) +

geom\_point(color = "black") +

labs(title = "Crime Trend by Hour of the Day", x = "Hour (0–23)", y = "Number of Crimes") +

theme\_minimal()

# Interactive Crime Heatmap

leaflet(data = crime\_clean) %>%

addTiles() %>%

addHeatmap(

lng = ~Longitude,

lat = ~Latitude,

blur = 20,

max = 0.05,

radius = 10

) %>%

addLegend(

position = "bottomright",

colors = "red",

labels = "Higher Density",

title = "Crime Heatmap (2020)"

)

**Output:**

> # Install required packages if not already installed

> packages <- c("tidyverse", "lubridate", "ggplot2", "cluster",

+ "factoextra", "dendextend", "leaflet", "leaflet.extras")

> install.packages(setdiff(packages, rownames(installed.packages())))

> # Load libraries

> library(tidyverse)

── Attaching core tidyverse packages ──────────────────────────────────────────────── tidyverse 2.0.0 ──

✔ dplyr 1.1.4 ✔ readr 2.1.5

✔ forcats 1.0.0 ✔ stringr 1.5.1

✔ ggplot2 3.5.2 ✔ tibble 3.2.1

✔ lubridate 1.9.4 ✔ tidyr 1.3.1

✔ purrr 1.0.4

── Conflicts ────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──

✖ dplyr::filter() masks stats::filter()

✖ dplyr::lag() masks stats::lag()

ℹ Use the conflicted package to force all conflicts to become errors

> library(lubridate)

> library(ggplot2)

> library(cluster)

> library(factoextra)

Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

> library(dendextend)

---------------------

Welcome to dendextend version 1.19.0

Type citation('dendextend') for how to cite the package.

Type browseVignettes(package = 'dendextend') for the package vignette.

The github page is: https://github.com/talgalili/dendextend/

Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendextend/issues

You may ask questions at stackoverflow, use the r and dendextend tags:

https://stackoverflow.com/questions/tagged/dendextend

To suppress this message use: suppressPackageStartupMessages(library(dendextend))

---------------------

Attaching package: ‘dendextend’

The following object is masked from ‘package:stats’:

cutree

> library(leaflet)

> library(leaflet.extras)

> library(readr)

> # Load dataset

> # https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about\_data

> crime\_data <- read\_csv("Crimes\_2025.csv")

Rows: 60835 Columns: 22

── Column specification ────────────────────────────────────────────────────────────────────────────────

Delimiter: ","

chr (12): Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Beat, Dis...

dbl (8): ID, Ward, Community Area, X Coordinate, Y Coordinate, Year, Latitude, Longitude

lgl (2): Arrest, Domestic

ℹ Use `spec()` to retrieve the full column specification for this data.

ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

> # View structure

> head(crime\_data)

# A tibble: 6 × 22

ID `Case Number` Date Block IUCR `Primary Type` Description `Location Description` Arrest

<dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <lgl>

1 13807860 JJ220558 04/13/202… 005X… 0710 THEFT THEFT FROM… STREET FALSE

2 13805171 JJ216644 04/13/202… 008X… 0281 CRIMINAL SEXU… NON-AGGRAV… APARTMENT FALSE

3 13807001 JJ218639 04/13/202… 016X… 1320 CRIMINAL DAMA… TO VEHICLE RESIDENCE - GARAGE FALSE

4 13804755 JJ216959 04/13/202… 064X… 0910 MOTOR VEHICLE… AUTOMOBILE STREET FALSE

5 13804041 JJ216041 04/13/202… 131X… 4387 OTHER OFFENSE VIOLATE OR… STREET TRUE

6 13805942 JJ217659 04/13/202… 003X… 0810 THEFT OVER $500 OTHER (SPECIFY) FALSE

# ℹ 13 more variables: Domestic <lgl>, Beat <chr>, District <chr>, Ward <dbl>, `Community Area` <dbl>,

# `FBI Code` <chr>, `X Coordinate` <dbl>, `Y Coordinate` <dbl>, Year <dbl>, `Updated On` <chr>,

# Latitude <dbl>, Longitude <dbl>, Location <chr>

> glimpse(crime\_data)

Rows: 60,835

Columns: 22

$ ID <dbl> 13807860, 13805171, 13807001, 13804755, 13804041, 13805942, 13804208, 1…

$ `Case Number` <chr> "JJ220558", "JJ216644", "JJ218639", "JJ216959", "JJ216041", "JJ217659",…

$ Date <chr> "04/13/2025 12:00:00 AM", "04/13/2025 12:00:00 AM", "04/13/2025 12:00:0…

$ Block <chr> "005XX E 38TH PL", "008XX W WELLINGTON AVE", "016XX W BRYN MAWR AVE", "…

$ IUCR <chr> "0710", "0281", "1320", "0910", "4387", "0810", "1320", "0760", "0630",…

$ `Primary Type` <chr> "THEFT", "CRIMINAL SEXUAL ASSAULT", "CRIMINAL DAMAGE", "MOTOR VEHICLE T…

$ Description <chr> "THEFT FROM MOTOR VEHICLE", "NON-AGGRAVATED", "TO VEHICLE", "AUTOMOBILE…

$ `Location Description` <chr> "STREET", "APARTMENT", "RESIDENCE - GARAGE", "STREET", "STREET", "OTHER…

$ Arrest <lgl> FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, TR…

$ Domestic <lgl> FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FA…

$ Beat <chr> "0212", "1933", "2012", "0724", "0433", "1831", "0222", "1935", "1421",…

$ District <chr> "002", "019", "020", "007", "004", "018", "002", "019", "014", "019", "…

$ Ward <dbl> 4, 44, 40, 16, 10, 42, 4, 43, 26, 44, 37, 19, 4, 48, 15, 22, 7, 8, 18, …

$ `Community Area` <dbl> 35, 6, 77, 68, 55, 8, 39, 7, 22, 6, 25, 74, 36, 77, 61, 30, 43, 43, 66,…

$ `FBI Code` <chr> "06", "02", "14", "07", "26", "06", "14", "06", "05", "06", "08B", "14"…

$ `X Coordinate` <dbl> 1180392, 1169981, 1164294, 1170751, 1201735, 1173634, 1186381, 1173188,…

$ `Y Coordinate` <dbl> 1879578, 1920130, 1937257, 1862151, 1818522, 1904417, 1873151, 1918772,…

$ Year <dbl> 2025, 2025, 2025, 2025, 2025, 2025, 2025, 2025, 2025, 2025, 2025, 2025,…

$ `Updated On` <chr> "04/20/2025 03:42:02 PM", "04/20/2025 03:42:02 PM", "04/20/2025 03:42:0…

$ Latitude <dbl> 41.82483, 41.93634, 41.98346, 41.77722, 41.65677, 41.89314, 41.80705, 4…

$ Longitude <dbl> -87.61370, -87.65071, -87.67112, -87.64958, -87.53747, -87.63775, -87.5…

$ Location <chr> "POINT (-87.613698729 41.824825134)", "POINT (-87.650710119 41.93633563…

> # Select key columns and remove missing data

> crime\_clean <- crime\_data %>%

+ select(Date, `Primary Type`, Latitude, Longitude, `Community Area`) %>%

+ drop\_na()

> # Convert Date and create derived time variables

> # Convert 'Date' column and filter last 5 years

> crime\_clean <- crime\_data %>%

+ select(Date, `Primary Type`, Latitude, Longitude, `Community Area`) %>%

+ drop\_na() %>%

+ mutate(

+ Date = mdy\_hms(Date),

+ Hour = hour(Date),

+ Month = month(Date),

+ Year = year(Date), # Extract year

+ Crime\_Type\_Code = as.numeric(as.factor(`Primary Type`))

+ ) %>%

+ filter(Date >= as.Date(Sys.Date()) - years(5)) # Only include last 5 years

> # Prepare clustering data

> kmeans\_data <- crime\_clean %>%

+ select(Latitude, Longitude, Crime\_Type\_Code)

> # Elbow method on a 5,000-point sample to find optimal k

> set.seed(42)

> sample\_data <- kmeans\_data %>% sample\_n(5000)

> fviz\_nbclust(sample\_data, kmeans, method = "wss") +

+ geom\_vline(xintercept = 6, linetype = 2) +

+ labs(title = "Elbow Method - Optimal k for K-Means (Sample)")

> # Run K-Means on full dataset with k = 6

> set.seed(123)

> kmeans\_result <- kmeans(kmeans\_data, centers = 6, nstart = 25)

> crime\_clean$KMeans\_Cluster <- as.factor(kmeans\_result$cluster)

> leaflet(data = crime\_clean) %>%

+ addTiles() %>%

+ addCircleMarkers(

+ lng = ~Longitude,

+ lat = ~Latitude,

+ color = ~KMeans\_Cluster,

+ radius = 2,

+ stroke = FALSE,

+ fillOpacity = 0.5,

+ popup = ~paste("Crime Type:", `Primary Type`)

+ )

> # Sample 500 records for efficiency

> set.seed(42)

> hier\_data <- kmeans\_data %>% sample\_n(500)

> # Compute distance matrix and hierarchical clustering

> dist\_matrix <- dist(hier\_data)

> hc <- hclust(dist\_matrix, method = "ward.D2")

> # Plot dendrogram

> plot(hc, labels = FALSE, hang = -1, main = "Dendrogram - Hierarchical Clustering")

> abline(h = 150, col = "red", lty = 2)

> # Cut into 6 clusters and add to data

> hc\_clusters <- cutree(hc, k = 6)

> hier\_data$HCluster <- as.factor(hc\_clusters)

> # Visualize clusters

> ggplot(hier\_data, aes(x = Longitude, y = Latitude, color = HCluster)) +

+ geom\_point(alpha = 0.7) +

+ theme\_minimal() +

+ labs(title = "Hierarchical Clustering (500 Sample Points)",

+ x = "Longitude", y = "Latitude")

> # Cophenetic correlation evaluation

> coph <- cophenetic(hc)

> cor\_val <- cor(dist\_matrix, coph)

> print(paste("Cophenetic correlation coefficient:", round(cor\_val, 3)))

[1] "Cophenetic correlation coefficient: 0.878"

> # Top 10 Crime Types

> crime\_clean %>%

+ count(`Primary Type`) %>%

+ arrange(desc(n)) %>%

+ top\_n(10) %>%

+ ggplot(aes(x = reorder(`Primary Type`, n), y = n)) +

+ geom\_col(fill = "steelblue") +

+ coord\_flip() +

+ labs(title = "Top 10 Crime Types in 2020", x = "Crime Type", y = "Count") +

+ theme\_minimal()

Selecting by n

> # Crime Trend by Hour

> crime\_clean %>%

+ count(Hour) %>%

+ ggplot(aes(x = Hour, y = n)) +

+ geom\_line(color = "tomato", size = 1.2) +

+ geom\_point(color = "black") +

+ labs(title = "Crime Trend by Hour of the Day", x = "Hour (0–23)", y = "Number of Crimes") +

+ theme\_minimal()

Warning message:

Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.

ℹ Please use `linewidth` instead.

This warning is displayed once every 8 hours.

Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was generated.

> # Interactive Crime Heatmap

> leaflet(data = crime\_clean) %>%

+ addTiles() %>%

+ addHeatmap(

+ lng = ~Longitude,

+ lat = ~Latitude,

+ blur = 20,

+ max = 0.05,

+ radius = 10

+ ) %>%

+ addLegend(

+ position = "bottomright",

+ colors = "red",

+ labels = "Higher Density",

+ title = "Crime Heatmap (2020)"

+ )

>