# General Sir John Kotelawala Defence University

## Faculty of Management, Social Sciences and Humanities

## Department of Languages

## BSc in Applied Data Science Communication


D L S NADAVI - D/ADC/23/0028

Y M S S B SENAVIRATHNA - D/ADC/23/0030

A G A U S GUNASEKARA - D/ADC/23/0034

M L B T S PERERA - D/ADC/23/0047


Fundamentals of Data Mining / LB 2114

Year 2: Semester 1

# Implementing clustering in data mining for Crimes in the city of Los Angeles

## 1) Introduction

Los Angeles is an American metropolis that serves as Southern California's cultural, financial, and economic core. With a population of approximately 3.98 million, it is the second most populous city in the United States and the most populous in the state of California. Los Angeles, noted for its diverse population and historic sites, suffers a number of criminal issues. Property offenses such as burglary and vehicle theft are common, as are violent crimes like robbery and assault. The city is dealing with gang-related issues, which have contributed to several criminal acts. Law enforcement activities seek to address these concerns and improve public safety in Los Angeles' large metropolitan terrain.

This report illustrate key metrics and insights into the crime statistics in Los Angeles. Also this report includes how to do how to do cluster analysis on a dataset. It also contains information about k-means clustering, elbow method and visualization tools available in R for clustering. This report clearly and orderly describes all the steps in doing k-means clustering to do the cluster analysis for the dataset using R. The aim of creating this report is to identify the crimes in Los Angeles done by teenagers.

## 2) Data Set

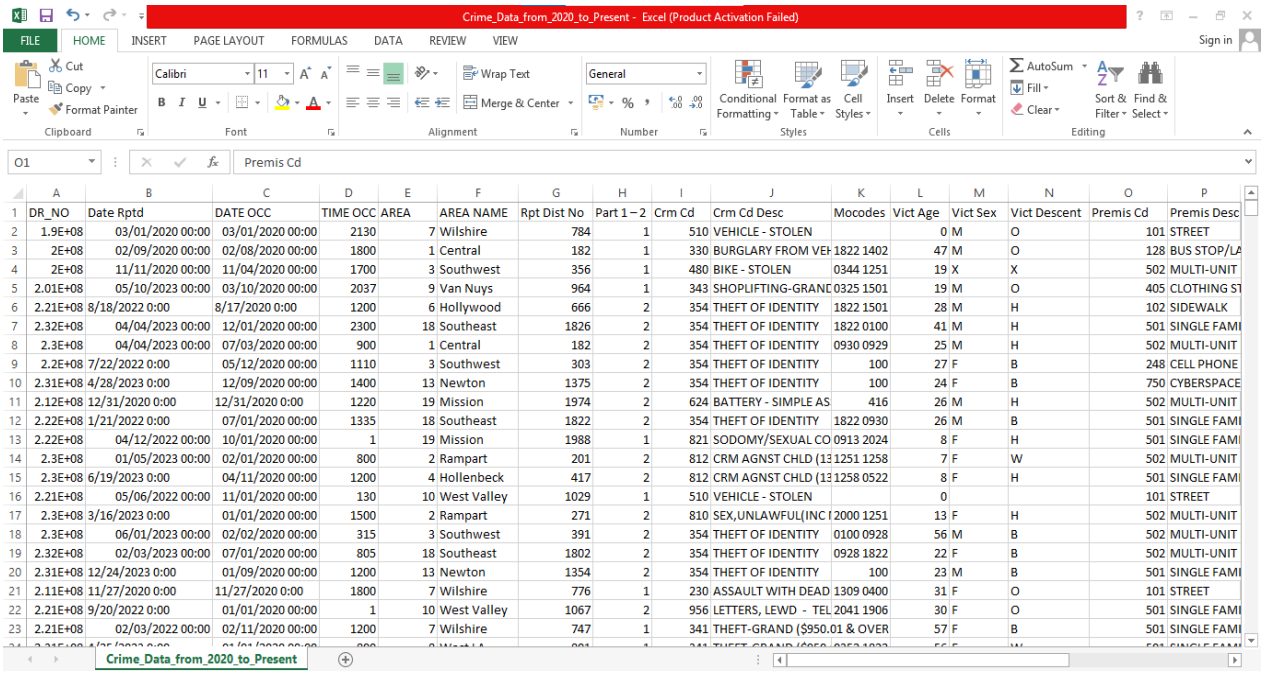
The data set was taken from: https://catalog.data.gov/dataset/crime-data-from-2020-to-present/resource/5eb6507e-fa82-4595-a604-023f8a326099

This dataset contains information about various crimes in the city of Los Angeles dating back to 2020. There are 28 columns and 892935 rows in the data set.

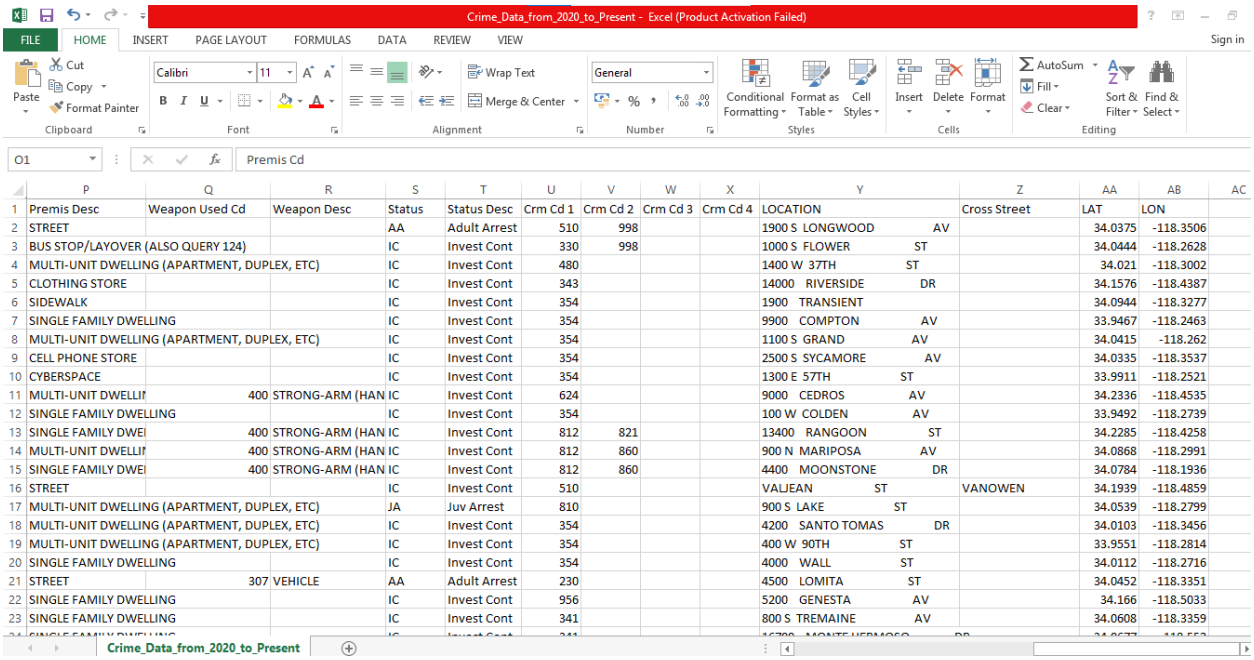
Attributes of the data set are,

1. DR_NO
2. Date Rptd - Date of the crime reported
3. DATE OCC - Date of the crime occurred
4. TIME OCC - Time of the crime occurred
5. AREA – Area code
6. AREA NAME – Name of the area
7. Rpt Dist No – Reported District No
8. Part 1 – 2
9. Crm Cd – Crime code
10. Crm Cd Desc – Crime code description

11. Mocodes
12. Vict Age – Age of the victim
13. Vict Sex – Sex of the victim
14. Vict Descent
15. Premis Cd – Premises code
16. Premis Desc – Premises description
17. Weapon Used Cd – Weapon code used for the crime
18. Weapon Desc – Weapon description
19. Status – AA, IC
20. Status Desc – Adult Arrest (AA), Invest Cont (IC)
21. Crm Cd 1 – Crime code 1
22. Crm Cd 2 - Crime code 2
23. Crm Cd 3 - Crime code 3
24. Crm Cd 4 - Crime code 4
25. LOCATION
26. Cross Street
27. LAT -Latitude
28. LON – Longitude




| DR_NO | Date Rptd | DATE OCC | TIME OCC | AREA | AREA NAME | Rpt Dist No | Part 1-2 | Crm Cd | Crm Cd Desc | Mocodes | Vict Age | Vict Sex | Vict Descent | Premis Cd | Premis Desc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.9E+08 | 03/01/2020 00:00 | 03/01/2020 00:00 | 2130 | 7 | Wilshire | 784 | 1 | 510 | VEHICLE - STOLEN | | 0 | M | O | 101 | STREET |
| 2E+08 | 02/09/2020 00:00 | 02/08/2020 00:00 | 1800 | 1 | Central | 182 | 1 | 330 | BURGLARY FROM VEH | 1822 1402 | 47 | M | O | 128 | BUS STOP/LA |
| 2E+08 | 11/11/2020 00:00 | 11/04/2020 00:00 | 1700 | 3 | Southwest | 356 | 1 | 480 | BIKE - STOLEN | 0344 1251 | 19 | X | X | 502 | MULTI-UNIT |
| 2.01E+08 | 05/10/2023 00:00 | 03/10/2020 00:00 | 2037 | 9 | Van Nuys | 964 | 1 | 343 | SHOPLIFTING-GRAND | 0325 1501 | 19 | M | O | 405 | CLOTHING ST |
| 2.21E+08 | 8/18/2022 0:00 | 8/17/2020 0:00 | 1200 | 6 | Hollywood | 666 | 2 | 354 | THEFT OF IDENTITY | 1822 1501 | 28 | M | H | 102 | SIDEWALK |
| 2.32E+08 | 04/04/2023 00:00 | 12/01/2020 00:00 | 2300 | 18 | Southeast | 1826 | 2 | 354 | THEFT OF IDENTITY | 1822 0100 | 41 | M | H | 501 | SINGLE FAMI |
| 2.3E+08 | 04/04/2023 00:00 | 07/03/2020 00:00 | 900 | 1 | Central | 182 | 2 | 354 | THEFT OF IDENTITY | 0930 0929 | 25 | M | H | 502 | MULTI-UNIT |
| 2.2E+08 | 7/22/2022 0:00 | 05/12/2020 00:00 | 1110 | 3 | Southwest | 303 | 2 | 354 | THEFT OF IDENTITY | 100 | 27 | F | B | 248 | CELL PHONE |
| 2.31E+08 | 4/28/2023 0:00 | 12/09/2020 00:00 | 1400 | 13 | Newton | 1375 | 2 | 354 | THEFT OF IDENTITY | 100 | 24 | F | B | 750 | CYBERSPACE |
| 2.12E+08 | 12/31/2020 0:00 | 12/31/2020 0:00 | 1220 | 19 | Mission | 1974 | 2 | 624 | BATTERY - SIMPLE AS | 416 | 26 | M | H | 502 | MULTI-UNIT |
| 2.22E+08 | 1/21/2022 0:00 | 07/01/2020 00:00 | 1335 | 18 | Southeast | 1822 | 2 | 354 | THEFT OF IDENTITY | 1822 0930 | 26 | M | B | 501 | SINGLE FAMI |
| 2.22E+08 | 04/12/2022 00:00 | 10/01/2020 00:00 | 1 | 19 | Mission | 1988 | 1 | 821 | SODOMY/SEXUAL CO | 0913 2024 | 8 | F | H | 501 | SINGLE FAMI |
| 2.3E+08 | 01/05/2023 00:00 | 02/01/2020 00:00 | 800 | 2 | Rampart | 201 | 2 | 812 | CRM AGNST CHLD (13 | 1251 1258 | 7 | F | W | 502 | MULTI-UNIT |
| 2.3E+08 | 6/19/2023 0:00 | 04/11/2020 00:00 | 1200 | 4 | Hollenbeck | 417 | 2 | 812 | CRM AGNST CHLD (13 | 1258 0522 | 8 | F | H | 501 | SINGLE FAMI |
| 2.21E+08 | 05/06/2022 00:00 | 11/01/2020 00:00 | 130 | 10 | West Valley | 1029 | 1 | 510 | VEHICLE - STOLEN | | 0 | | | 101 | STREET |
| 2.3E+08 | 3/16/2023 0:00 | 01/01/2020 00:00 | 1500 | 2 | Rampart | 271 | 2 | 810 | SEX,UNLAWFUL(INC I | 2000 1251 | 13 | F | H | 502 | MULTI-UNIT |
| 2.3E+08 | 06/01/2023 00:00 | 02/02/2020 00:00 | 315 | 3 | Southwest | 391 | 2 | 354 | THEFT OF IDENTITY | 0100 0928 | 56 | M | B | 502 | MULTI-UNIT |
| 2.32E+08 | 02/03/2023 00:00 | 07/01/2020 00:00 | 805 | 18 | Southeast | 1802 | 2 | 354 | THEFT OF IDENTITY | 0928 1822 | 22 | F | B | 502 | MULTI-UNIT |
| 2.31E+08 | 12/24/2023 0:00 | 01/09/2020 00:00 | 1200 | 13 | Newton | 1354 | 2 | 354 | THEFT OF IDENTITY | 100 | 23 | M | B | 501 | SINGLE FAMI |
| 2.11E+08 | 11/27/2020 0:00 | 11/27/2020 0:00 | 1800 | 7 | Wilshire | 776 | 1 | 230 | ASSAULT WITH DEAD | 1309 0400 | 31 | F | O | 101 | STREET |
| 2.21E+08 | 9/20/2022 0:00 | 01/01/2020 00:00 | 1 | 10 | West Valley | 1067 | 2 | 956 | LETTERS, LEWD - TEL | 2041 1906 | 30 | F | O | 501 | SINGLE FAMI |
| 2.21E+08 | 02/03/2022 00:00 | 02/11/2020 00:00 | 1200 | 7 | Wilshire | 747 | 1 | 341 | THEFT-GRAND ($950.01 & OVER | | 57 | F | B | 501 | SINGLE FAMI |




| | P | Q | R | S | T | U | V | W | X | Y | Z | AA | AB | AC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Premis Desc | Weapon Used Cd | Weapon Desc | Status | Status Desc | Crm Cd 1 | Crm Cd 2 | Crm Cd 3 | Crm Cd 4 | LOCATION | Cross Street | LAT | LON | |
| 2 | STREET | | | AA | Adult Arrest | 510 | 998 | | | 1900 S LONGWOOD AV | | 34.0375 | -118.3506 | |
| 3 | BUS STOP/LAYOVER (ALSO QUERY 124) | | | IC | Invest Cont | 330 | 998 | | | 1000 S FLOWER ST | | 34.0444 | -118.2628 | |
| 4 | MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC) | | | IC | Invest Cont | 480 | | | | 1400 W 37TH ST | | 34.021 | -118.3002 | |
| 5 | CLOTHING STORE | | | IC | Invest Cont | 343 | | | | 14000 RIVERSIDE DR | | 34.1576 | -118.4387 | |
| 6 | SIDEWALK | | | IC | Invest Cont | 354 | | | | 1900 TRANSIENT | | 34.0944 | -118.3277 | |
| 7 | SINGLE FAMILY DWELLING | | | IC | Invest Cont | 354 | | | | 9900 COMPTON AV | | 33.9467 | -118.2463 | |
| 8 | MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC) | | | IC | Invest Cont | 354 | | | | 1100 S GRAND AV | | 34.0415 | -118.262 | |
| 9 | CELL PHONE STORE | | | IC | Invest Cont | 354 | | | | 2500 S SYCAMORE AV | | 34.0335 | -118.3537 | |
| 10 | CYBERSPACE | | | IC | Invest Cont | 354 | | | | 1300 E 57TH ST | | 33.9911 | -118.2521 | |
| 11 | MULTI-UNIT DWELLI | 400 | STRONG-ARM (HAN | IC | Invest Cont | 624 | | | | 9000 CEDROS AV | | 34.2336 | -118.4535 | |
| 12 | SINGLE FAMILY DWELLING | | | IC | Invest Cont | 354 | | | | 100 W COLDEN AV | | 33.9492 | -118.2739 | |
| 13 | SINGLE FAMILY DWEI | 400 | STRONG-ARM (HAN | IC | Invest Cont | 812 | 821 | | | 13400 RANGOON ST | | 34.2285 | -118.4258 | |
| 14 | MULTI-UNIT DWELLI | 400 | STRONG-ARM (HAN | IC | Invest Cont | 812 | 860 | | | 900 N MARIPOSA AV | | 34.0868 | -118.2991 | |
| 15 | SINGLE FAMILY DWEI | 400 | STRONG-ARM (HAN | IC | Invest Cont | 812 | 860 | | | 4400 MOONSTONE DR | | 34.0784 | -118.1936 | |
| 16 | STREET | | | IC | Invest Cont | 510 | | | | VALJEAN ST | VANOWEN | 34.1939 | -118.4859 | |
| 17 | MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC) | | | JA | Juv Arrest | 810 | | | | 900 S LAKE ST | | 34.0539 | -118.2799 | |
| 18 | MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC) | | | IC | Invest Cont | 354 | | | | 4200 SANTO TOMAS DR | | 34.0103 | -118.3456 | |
| 19 | MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC) | | | IC | Invest Cont | 354 | | | | 400 W 90TH ST | | 33.9551 | -118.2814 | |
| 20 | SINGLE FAMILY DWELLING | | | IC | Invest Cont | 354 | | | | 4000 WALL ST | | 34.0112 | -118.2716 | |
| 21 | STREET | 307 | VEHICLE | AA | Adult Arrest | 230 | | | | 4500 LOMITA ST | | 34.0452 | -118.3351 | |
| 22 | SINGLE FAMILY DWELLING | | | IC | Invest Cont | 956 | | | | 5200 GENESTA AV | | 34.166 | -118.5033 | |
| 23 | SINGLE FAMILY DWELLING | | | IC | Invest Cont | 341 | | | | 800 S TREMAINE AV | | 34.0608 | -118.3359 | |

## 3) Explanation and Preparation of Data Set

### a. Data Preprocessing

Crime data set has been used for the clustering task. To get a proper data set for clustering, we have filtered the NULL values from the dataset. We selected only 13 columns from the dataset to do the cluster analysis as dealing with 28 columns seems to be difficult.

### b. Data Explanation

| *Variable* | *Explanation* |
|---|---|
| 1) DATE OCC | Date of the crime occurred |
| 2) TIME OCC | Time of the crime occurred |
| 3) AREA | Area code |
| 4) AREA NAME | Name of the area where the crime occurred |
| 5) Part 1 – 2 | Part 1 or 2 |
| 6) Crm Cd | Crime code |
| 7) Crm Cd Desc | Crime code description |
| 8) Vict Age | Age of the victim |
| 9) Vict Sex | Sex of the victim |
| 10) Premis Cd | Premises code |
| 11) Premis Desc | Premises description |
| 12) LAT | Latitude |
| 13) LON | Longitude |

## 4) Data Mining

***Clustering***

Clustering is a data science and machine learning technique that groups comparable rows in a data set. The clustering method aims to separate groups with similar qualities and assign them to clusters. After executing a clustering algorithm, a new column appears in the data set to show which group each row of data belongs to the most.

***K means Clustering***

K-Means clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into distinct, non-overlapping clusters. Each cluster is represented by its center. The algorithm begins by randomly initializing cluster centroids and iteratively assigns data points to the nearest centroid based on a chosen distance metric, commonly the Euclidean distance.

This is how the k-means clustering works.

- **Step 1:** Choose k randomly selected locations to serve as cluster centers.
- **Step 2:** Find the Euclidean distance between each data point and each cluster center.
- **Step 3:** Assign the data point to the cluster that has the least Euclidean distance.
- **Step 4:** Recalculate the cluster centroids using the data points provided in step 3.
- **Step 5:** Repeat steps 2-4 until the centroids stop updating or the maximum number of iterations is reached.

***Elbow Method***

The elbow method is a graphical approach to determining the ideal K value in a k-means clustering algorithm. The elbow graph displays the within-cluster-sum-of-squares (WCSS) values on the y-axis, which correlate to different K values (on the x-axis). The ideal K value is found at the point when the graph forms an elbow. The elbow method uses the k-means clustering algorithm over a range of k values.

1. A score is calculated for each k value.
2. By default, the distortion score is calculated as the sum of the squared distances between each data point and its assigned center.

3. When we plot the value of k vs this score, we get an elbow-shaped plot.
4. The elbow point is where the distortion drops the most, and this is where the ideal value of k can be obtained.

***Visualization tools available in R for clustering***

R allows us to generate visually appealing data visualizations by adding a few lines of code. To accomplish this, we employ many of R's features. Data visualization is an effective way to examine and interpret data.

1. **Cluster plot -** Clusters are small groups of data points that appear in scatter plots. K-means clustering is the most frequent unsupervised machine learning strategy for partitioning a given data set into k groups, where k is the number of established categories by the analyst. Cluster analysis and fact analysis are ways for categorizing items into smaller components in Clusters.
2. **Scatterplot matrix -** Shows the relationships between multiple variables. The matrix can indicate links between variables after charting two-way combinations of the variables to highlight which associations are likely to be important. Outliers in numerous scatter plots can also be identified using the matrix.
3. **Distance Matrix** - This graphic shows the distances between rows in a matrix or data set. This technique makes use of the 'factoextra' R package, which makes it easier to extract and visualize the results of exploratory multivariable data analysis.
4. **Dot plot** - A dot plot or dot chart is similar to a scatter plot. However, the primary distinction is that in R, the dot plot displays the index (each category) on the vertical axis, allowing you to evaluate the value of each observation.

## 5) Implementation R

***Packages Used***

1) **dplyr :** This package is used for data manipulation. It provides a set of functions that allow you to filter rows, select columns, group data, and perform various operations on data frames.
2) **ggplot2:** This is a powerful and flexible plotting system. It allows you to create complex and customized visualizations using a layered grammar of graphics. It's particularly popular for creating publication-quality graphics.
3) **stats:** This package is part of the R base distribution and contains fundamental statistical functions and distributions. It includes functions for basic statistical calculations, probability distributions, and hypothesis testing, linear models.
4) **ggfortify:** This package extends ggplot2 to create visualizations for various statistical models in a consistent and user-friendly manner. It provides functions to create ggplot2 plots for objects created by statistical modeling functions.
5) **cluster** : This package includes functions for plotting cluster-related visualizations. It is often used in conjunction with hierarchical clustering.
6) **factoextra** : This package is designed to extract and visualize information from the output of multivariate analysis, including clustering results. It provides functions to create informative plots such as scatter plots.

***Explanation of the experimental procedure and Visualization of the results***

**Step 01**

Install and activate packages.

```
#install packages
install.packages("dplyr")
install.packages("ggplot2")
install.packages("stats")
install.packages("ggfortify")
install.packages("cluster")
install.packages("factoextra")

#load required libraries
library(dplyr)
library(ggplot2)
library(stats)
library(ggfortify)
library(cluster)
library(factoextra)
```

**Step 02**

Install 'readxl' package for importing the data set which saved as Excel file.

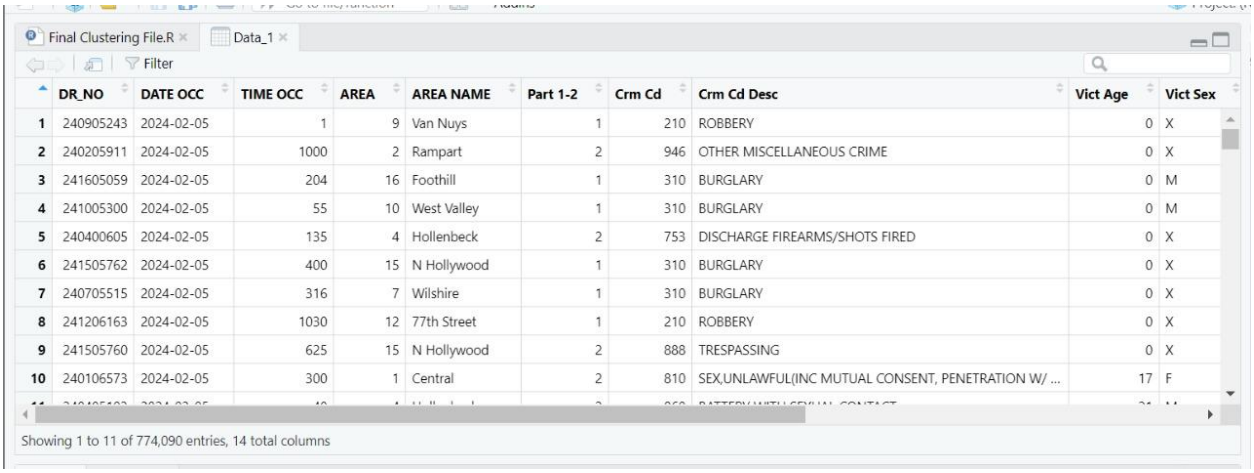
Then, import the dataset.

```
install.packages("readxl")
library(readxl)

Data=read_excel("C:\\Users\\ASUS\\Documents\\2nd Yr 1st Sem\\Data Mining\\Assignment 1\\Crime_Data_from_2020_to_Present.xlsx")
View(Data)
```

**Step 03**

Delete NA rows from the data set.

```
#delete NA raws
Data_1 <- na.omit(Data)
View(Data_1)
```



Final Clustering File.R | Data_1

Filter

| | DR_NO | DATE OCC | TIME OCC | AREA | AREA NAME | Part 1-2 | Crm Cd | Crm Cd Desc | Vict Age | Vict Sex |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 240905243 | 2024-02-05 | 1 | 9 | Van Nuys | 1 | 210 | ROBBERY | 0 | X |
| 2 | 240205911 | 2024-02-05 | 1000 | 2 | Rampart | 2 | 946 | OTHER MISCELLANEOUS CRIME | 0 | X |
| 3 | 241605059 | 2024-02-05 | 204 | 16 | Foothill | 1 | 310 | BURGLARY | 0 | M |
| 4 | 241005300 | 2024-02-05 | 55 | 10 | West Valley | 1 | 310 | BURGLARY | 0 | M |
| 5 | 240400605 | 2024-02-05 | 135 | 4 | Hollenbeck | 2 | 753 | DISCHARGE FIREARMS/SHOTS FIRED | 0 | X |
| 6 | 241505762 | 2024-02-05 | 400 | 15 | N Hollywood | 1 | 310 | BURGLARY | 0 | X |
| 7 | 240705515 | 2024-02-05 | 316 | 7 | Wilshire | 1 | 310 | BURGLARY | 0 | X |
| 8 | 241206163 | 2024-02-05 | 1030 | 12 | 77th Street | 1 | 210 | ROBBERY | 0 | X |
| 9 | 241505760 | 2024-02-05 | 625 | 15 | N Hollywood | 2 | 888 | TRESPASSING | 0 | X |
| 10 | 240106573 | 2024-02-05 | 300 | 1 | Central | 2 | 810 | SEX,UNLAWFUL(INC MUTUAL CONSENT, PENETRATION W/ ... | 17 | F |

Showing 1 to 11 of 774,090 entries, 14 total columns

**Step 04**

Create a new vector to format ‘TIME OCC’ column values correctly.

```
# Create a vector of numeric time values for make time correctly
new_times <- c(Data_1$`TIME OCC`)

# Convert to character format with a colon separator
formatted_times <- sprintf("%02d:%02d", new_times %/% 100, new_times %% 100)
```

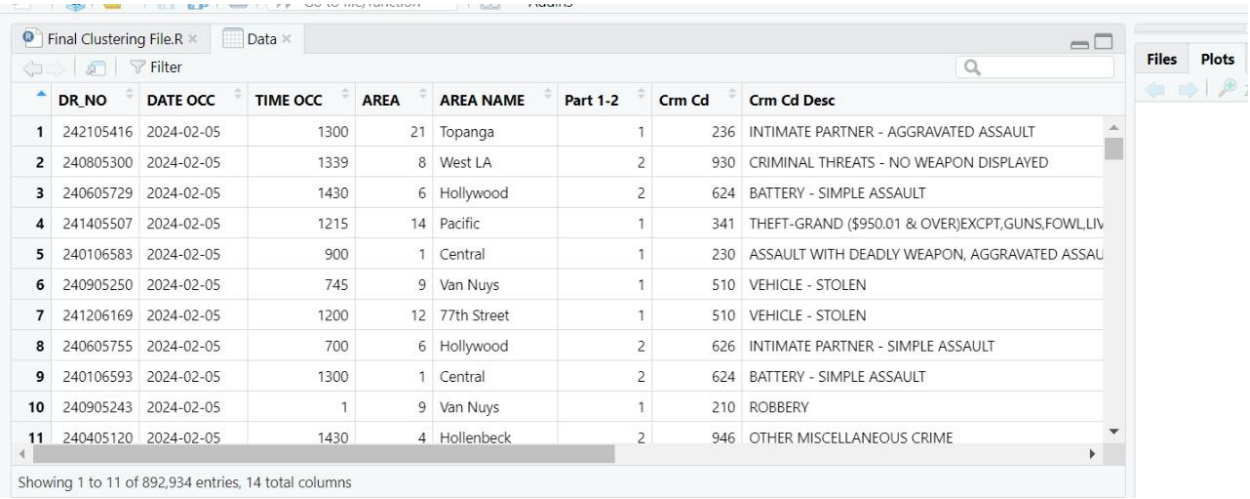
Create a data frame with the formatted times.

```
Time_OCC<- data.frame(`TIME OCC`= formatted_times)
View(Time_OCC)
```

| | TIME.OCC |
|---|---|
| 1 | 00:01 |
| 2 | 10:00 |
| 3 | 02:04 |
| 4 | 00:55 |
| 5 | 01:35 |
| 6 | 04:00 |
| 7 | 03:16 |
| 8 | 10:30 |
| 9 | 06:25 |
| 10 | 03:00 |
| 11 | 00:40 |

Remove the previous ‘TIME OCC’ column from the dataset and add new column ‘TIME_OCC’ to the dataset. Then define the new order of columns.

```
Data_1[ ,-3]

Data_2=cbind(c(Data_1[ , -3]),Time_OCC)
View(Data_2)

# Define the new order of columns
Data_3=Data_2[ ,c(1,2,14,3:13)]
View(Data_3)
```




| | DR_NO | DATE OCC | TIME OCC | AREA | AREA NAME | Part 1-2 | Crm Cd | Crm Cd Desc |
|---|---|---|---|---|---|---|---|---|
| 1 | 242105416 | 2024-02-05 | 1300 | 21 | Topanga | 1 | 236 | INTIMATE PARTNER - AGGRAVATED ASSAULT |
| 2 | 240805300 | 2024-02-05 | 1339 | 8 | West LA | 2 | 930 | CRIMINAL THREATS - NO WEAPON DISPLAYED |
| 3 | 240605729 | 2024-02-05 | 1430 | 6 | Hollywood | 2 | 624 | BATTERY - SIMPLE ASSAULT |
| 4 | 241405507 | 2024-02-05 | 1215 | 14 | Pacific | 1 | 341 | THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIV |
| 5 | 240106583 | 2024-02-05 | 900 | 1 | Central | 1 | 230 | ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAU |
| 6 | 240905250 | 2024-02-05 | 745 | 9 | Van Nuys | 1 | 510 | VEHICLE - STOLEN |
| 7 | 241206169 | 2024-02-05 | 1200 | 12 | 77th Street | 1 | 510 | VEHICLE - STOLEN |
| 8 | 240605755 | 2024-02-05 | 700 | 6 | Hollywood | 2 | 626 | INTIMATE PARTNER - SIMPLE ASSAULT |
| 9 | 240106593 | 2024-02-05 | 1300 | 1 | Central | 2 | 624 | BATTERY - SIMPLE ASSAULT |
| 10 | 240905243 | 2024-02-05 | 1 | 9 | Van Nuys | 1 | 210 | ROBBERY |
| 11 | 240405120 | 2024-02-05 | 1430 | 4 | Hollenbeck | 2 | 946 | OTHER MISCELLANEOUS CRIME |

Showing 1 to 11 of 892,934 entries, 14 total columns

Filter the rows of the dataset as follows to get the crimes done by teenagers between the dates from 2023-12-01 to 2023-12-31:

```
# Get rows according to sorted excel file (only 2023)
Data_3$`DATE OCC` <- as.Date(Data_3$`DATE OCC`, format = "%Y-%m-%d")

# Filter rows between specific dates
Data_4<- subset(Data_3, `DATE OCC` >= "2023-12-01" & `DATE OCC` <= "2023-12-31")
View(Data_4)
```

```
#Get only teenagers
Crime_Data=Data_4[which(Data_4$`Vict Age` >=13 & Data_4$`Vict Age`<=19), ]
View(Crime_Data)
```

**Step 05**

Explore the dataset as follows:

Use the ‘name ()’ function to get the column names of the dataset.

Use ‘head ()’ and ‘tail ()’ functions to get first and last 6 rows in the dataset.

Use the ‘summary ()’ function to get the summary of the dataset.

Use the ‘str ()’ function to get the structure of the dataset.

```
names(Crime_Data)
head(Crime_Data)
tail(Crime_Data)
summary(Crime_Data)
str(Crime_Data)

nrow(Crime_Data)
ncol(Crime_Data)
dim(Crime_Data)
```

```
> head(Crime_Data)
          DR_NO   DATE OCC TIME.OCC AREA   AREA NAME Part 1-2 Crm Cd                                Crm Cd Desc Vict Age Vict Sex
16183 240504116 2023-12-31    16:50    5      Harbor        2    627        CHILD ABUSE (PHYSICAL) - SIMPLE ASSAULT       13        F
16184 241304515 2023-12-31    17:00   13      Newton        2    956        LETTERS, LEWD  -  TELEPHONE CALLS, LEWD       14        F
16185 232118269 2023-12-31    12:50   21     Topanga        1    210                                      ROBBERY       16        M
16186 241204006 2023-12-31    13:34   12 77th Street        2    626             INTIMATE PARTNER - SIMPLE ASSAULT       16        M
16187 241204008 2023-12-31    13:45   12 77th Street        2    626             INTIMATE PARTNER - SIMPLE ASSAULT       17        F
16188 231821477 2023-12-31    13:30   18   Southeast        1    230 ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT       18        F
      Premis Cd                          Premis Desc     LAT       LON
16183       203                       SINGLE FAMILY DWELLING 34.0808 -118.2655
16184       501                            OTHER PREMISE 34.1232 -118.2003
16185       502 MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC) 34.1486 -118.4063
16186       402                                    STREET 34.0726 -118.3463
16187       502                                    STREET 34.1685 -118.4019
16188       501                                  SIDEWALK 34.2314 -118.3973
> tail(Crime_Data)
          DR_NO   DATE OCC TIME.OCC AREA  AREA NAME Part 1-2 Crm Cd                          Crm Cd Desc Vict Age Vict Sex Premis Cd
31736 240505203 2023-12-01    14:00    5     Harbor        2    930 CRIMINAL THREATS - NO WEAPON DISPLAYED       17        F       102
31737 231820193 2023-12-01    20:50   18  Southeast        1    210                                ROBBERY       17        M       101
31738 241704304 2023-12-01    13:00   17 Devonshire        1    440     THEFT PLAIN - PETTY ($950 & UNDER)       18        M       101
31739 231917325 2023-12-01    12:55   19    Mission        1    440     THEFT PLAIN - PETTY ($950 & UNDER)       18        M       210
31740 231614601 2023-12-01    21:08   16   Foothill        2    903                CONTEMPT OF COURT       19        F       101
31741 230516471 2023-12-01    21:30    5     Harbor        1    330            BURGLARY FROM VEHICLE       19        M       101
                           Premis Desc     LAT       LON
31736             SINGLE FAMILY DWELLING 34.0377 -118.2607
31737                          STREET 34.2647 -118.4517
31738                   OTHER PREMISE 34.0982 -118.2787
31739                        DRIVEWAY 34.0377 -118.4408
31740                           ALLEY 34.1549 -118.4585
31741 ABANDONED BUILDING ABANDONED HOUSE 34.1299 -118.3770
>
```

```
> summary(Crime_Data)
     DR_NO               DATE OCC            TIME.OCC              AREA        AREA NAME            Part 1-2         Crm Cd
 Min.   :230125718   Min.   :2023-12-01   Length:503         Min.   : 1.00   Length:503         Min.   :1.000   Min.   :110.0
 1st Qu.:230517220   1st Qu.:2023-12-07   Class :character   1st Qu.: 5.00   Class :character   1st Qu.:1.000   1st Qu.:310.0
 Median :231224924   Median :2023-12-14   Mode  :character   Median :12.00   Mode  :character   Median :2.000   Median :624.0
 Mean   :231997811   Mean   :2023-12-14                      Mean   :10.85                      Mean   :1.551   Mean   :532.2
 3rd Qu.:231821019   3rd Qu.:2023-12-23                      3rd Qu.:17.00                      3rd Qu.:2.000   3rd Qu.:627.0
 Max.   :242104028   Max.   :2023-12-31                      Max.   :21.00                      Max.   :2.000   Max.   :956.0
 Crm Cd Desc          Vict Age       Vict Sex           Premis Cd     Premis Desc             LAT             LON
 Length:503         Min.   :13.00   Length:503         Min.   :101.0   Length:503         Min.   : 0.00   Min.   :-118.6
 Class :character   1st Qu.:15.00   Class :character   1st Qu.:101.0   Class :character   1st Qu.:34.01   1st Qu.:-118.4
 Mode  :character   Median :17.00   Mode  :character   Median :203.0   Mode  :character   Median :34.06   Median :-118.3
                    Mean   :16.85                      Mean   :297.9                      Mean   :34.01   Mean   :-118.1
                    3rd Qu.:19.00                      3rd Qu.:501.0                      3rd Qu.:34.17   3rd Qu.:-118.3
                    Max.   :19.00                      Max.   :946.0                      Max.   :34.32   Max.   :   0.0
> str(Crime_Data)
'data.frame':   503 obs. of  14 variables:
 $ DR_NO      : num  2.41e+08 2.41e+08 2.32e+08 2.41e+08 2.41e+08 ...
 $ DATE OCC   : Date, format: "2023-12-31" "2023-12-31" "2023-12-31" "2023-12-31" ...
 $ TIME.OCC   : chr  "16:50" "17:00" "12:50" "13:34" ...
 $ AREA       : num  5 13 21 12 12 18 19 18 6 5 ...
 $ AREA NAME  : chr  "Harbor" "Newton" "Topanga" "77th Street" ...
 $ Part 1-2   : num  2 2 1 2 2 1 1 1 1 2 ...
 $ Crm Cd     : num  627 956 210 626 626 230 330 230 310 745 ...
 $ Crm Cd Desc: chr  "CHILD ABUSE (PHYSICAL) - SIMPLE ASSAULT" "LETTERS, LEWD  -  TELEPHONE CALLS, LEWD" "ROBBERY" "INTIMATE PARTNER - SIMPLE ASS
AULT" ...
 $ Vict Age   : num  13 14 16 16 17 18 18 18 18 18 ...
 $ Vict Sex   : chr  "F" "F" "M" "M" ...
 $ Premis Cd  : num  203 501 502 402 502 501 101 501 501 108 ...
 $ Premis Desc: chr  "SINGLE FAMILY DWELLING" "OTHER PREMISE" "MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)" "STREET" ...
 $ LAT        : num  34.1 34.1 34.1 34.1 34.2 ...
 $ LON        : num  -118 -118 -118 -118 -118 ...
```

## Step 06

Use the ‘dim ()’ function to get the dimension of the data set which includes the number of rows and columns in the data set.

```
> nrow(Crime_Data)
[1] 503
> ncol(Crime_Data)
[1] 14
> dim(Crime_Data)
[1] 503  14
>
```

## Step 07

Get the structure of the dataset using ‘str ()’ function to check the variable type of column names.

```
> str(Crime_Data)
'data.frame':   503 obs. of  14 variables:
 $ DR_NO      : num  2.41e+08 2.41e+08 2.32e+08 2.41e+08 2.41e+08 ...
 $ DATE OCC   : Date, format: "2023-12-31" "2023-12-31" "2023-12-31" "2023-12-31" ...
 $ TIME.OCC   : chr  "16:50" "17:00" "12:50" "13:34" ...
 $ AREA       : num  5 13 21 12 12 18 19 18 6 5 ...
 $ AREA NAME  : chr  "Harbor" "Newton" "Topanga" "77th Street" ...
 $ Part 1-2   : num  2 2 1 2 2 1 1 1 1 2 ...
 $ Crm Cd     : num  627 956 210 626 626 230 330 230 310 745 ...
 $ Crm Cd Desc: chr  "CHILD ABUSE (PHYSICAL) - SIMPLE ASSAULT" "LETTERS, LEWD  -  TELEPHONE CALLS, LEWD" "ROBBERY" "INTIMATE PARTNER - SIMPLE ASS
AULT" ...
 $ Vict Age   : num  13 14 16 16 17 18 18 18 18 18 ...
 $ Vict Sex   : chr  "F" "F" "M" "M" ...
 $ Premis Cd  : num  203 501 502 402 502 501 101 501 501 108 ...
 $ Premis Desc: chr  "SINGLE FAMILY DWELLING" "OTHER PREMISE" "MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)" "STREET" ...
 $ LAT        : num  34.1 34.1 34.1 34.1 34.2 ...
 $ LON        : num  -118 -118 -118 -118 -118 ...
>
```

**Step 08**

Convert character columns into numerical columns as follows:

```
#Convert character columns into numerical

Crime_Data$`DATE OCC`= as.factor(Crime_Data$`DATE OCC`)
Crime_Data$`DATE OCC`= as.numeric(Crime_Data$`DATE OCC`)
table(Crime_Data$`DATE OCC`)

Crime_Data$TIME.OCC=as.factor(Crime_Data$TIME.OCC)
Crime_Data$TIME.OCC=as.numeric(Crime_Data$TIME.OCC)
table(Crime_Data$TIME.OCC)

Crime_Data$`AREA NAME`= as.factor(Crime_Data$`AREA NAME`)
Crime_Data$`AREA NAME`= as.numeric(Crime_Data$`AREA NAME`)
table(Crime_Data$`AREA NAME`)


Crime_Data$`Crm Cd Desc`= as.factor(Crime_Data$`Crm Cd Desc`)
Crime_Data$`Crm Cd Desc`= as.numeric(Crime_Data$`Crm Cd Desc`)
table(Crime_Data$`Crm Cd Desc`)

Crime_Data$`Vict Sex`= as.factor(Crime_Data$`Vict Sex`)
Crime_Data$`Vict Sex`= as.numeric(Crime_Data$`Vict Sex`)
table(Crime_Data$`Vict Sex`)

Crime_Data$`Premis Desc`= as.factor(Crime_Data$`Premis Desc`)
Crime_Data$`Premis Desc`= as.numeric(Crime_Data$`Premis Desc`)
table(Crime_Data$`Premis Desc`)


str(Crime_Data)
```

**Step 09**

Get the structure of the dataset again using 'str ()' function to check whether the character columns of the dataset are correctly converted into numeric columns for clustering.

```
> str(Crime_Data)
'data.frame':   503 obs. of  14 variables:
 $ DR_NO       : num  2.41e+08 2.41e+08 2.32e+08 2.41e+08 2.41e+08 ...
 $ DATE OCC    : num  31 31 31 31 31 31 31 31 31 31 ...
 $ TIME.OCC    : num  123 125 77 85 87 84 125 61 112 23 ...
 $ AREA        : num  5 13 21 12 12 18 19 18 6 5 ...
 $ AREA NAME   : num  5 10 17 1 1 15 8 15 7 5 ...
 $ Part 1-2    : num  2 2 1 2 2 1 1 1 1 2 ...
 $ Crm Cd      : num  627 956 210 626 626 230 330 230 310 745 ...
 $ Crm Cd Desc: num  11 27 34 25 25 1 10 1 9 48 ...
 $ Vict Age    : num  13 14 16 16 17 18 18 18 18 18 ...
 $ Vict Sex    : num  1 1 2 2 1 1 2 2 1 2 ...
 $ Premis Cd   : num  203 501 502 402 502 501 101 501 501 108 ...
 $ Premis Desc: num  48 38 34 50 50 47 43 50 48 53 ...
 $ LAT         : num  34.1 34.1 34.1 34.1 34.2 ...
 $ LON         : num  -118 -118 -118 -118 -118 ...
>
```

**Step 10**

Check if there are any missing values to input.

```
#checking for any not available in the dataset

anyNA(Crime_Data)
```

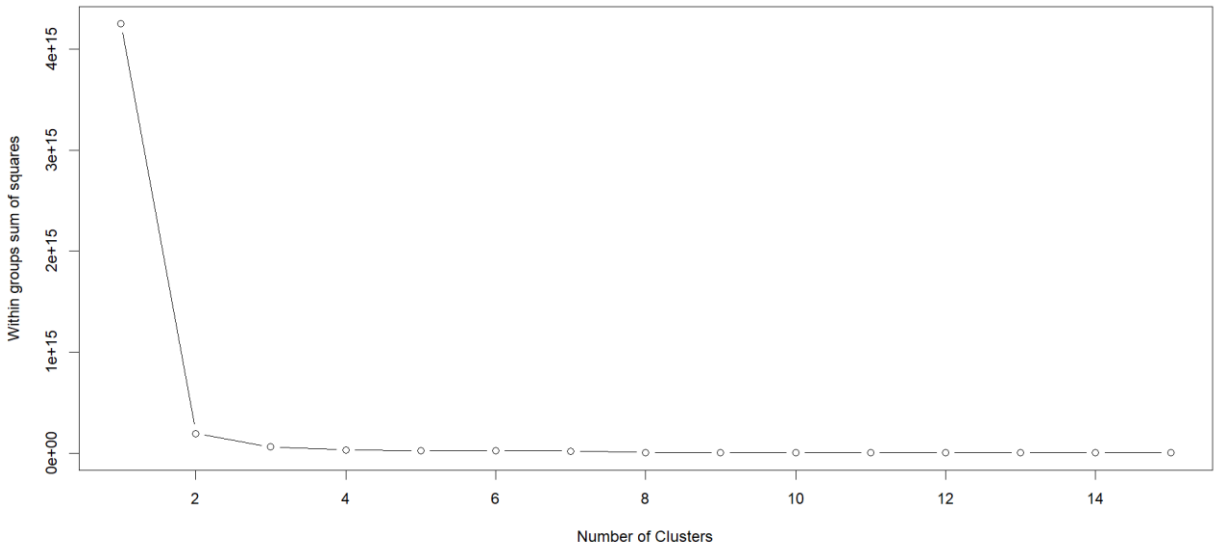
```
> anyNA(Crime_Data)
[1] FALSE
>
```

**Step 11**

Elbow method

Use wss plot to identify the maximum number of clusters can be created from the dataset.

```
#wss plot to choose maximum number of clusters

wssplot <- function(data, nc=15, seed=123){
  wss <- (nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc){
    set.seed(seed)
    wss[i] <- sum(kmeans(data, centers=i)$withinss)}
  plot(1:nc, wss, type="b", xlab="Number of Clusters",
       ylab="Within groups sum of squares")
  wss
}

wssplot(Crime_Data)
```




**Step 12**

K-Means clustering

Use the following line of code to create k means clustering.

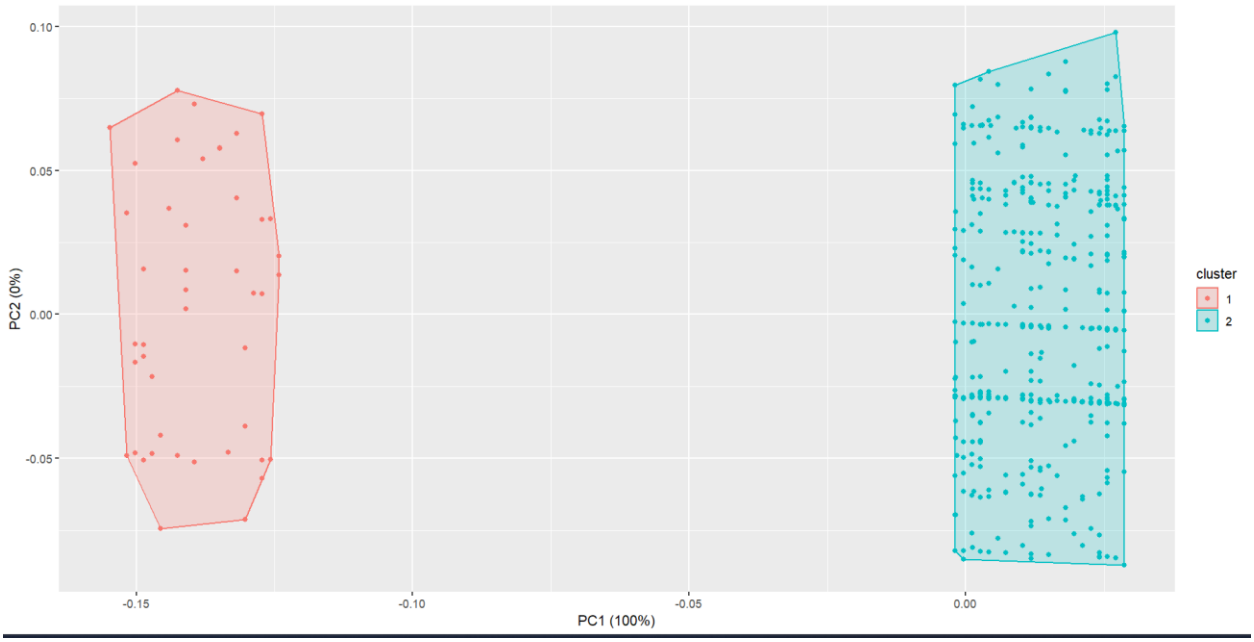
```
#Spotting the k means in the curve in order to choose the optimum number of cluster=2

KM=kmeans(Crime_Data,2)
KM
```

```
> KM
K-means clustering with 2 clusters of sizes 45, 458

Cluster means:
      DR_NO DATE OCC TIME.OCC     AREA AREA NAME Part 1-2   Crm Cd
1 241060003 20.13333 104.7778 10.55556  9.622222 1.466667 581.4222
2 231107421 14.49127 102.0218 10.88428 10.096070 1.558952 527.3996
  Crm Cd Desc Vict Age Vict Sex Premis Cd Premis Desc      LAT
1    26.11111 16.84444 1.422222  297.1111    41.11111 34.08797
2    16.17249 16.85153 1.469432  297.9934    39.75983 34.00365
        LON
1 -118.3410
2 -118.0983

Clustering vector:
16183 16184 16185 16186 16187 16188 16189 16190 16191 16192 16193 16194
    1     1     2     1     1     2     1     2     1     1     1     1
16195 16196 16680 16681 16682 16683 16684 16685 16686 16687 16688 16689
    1     1     2     2     1     2     2     2     1     2     2     1
16690 16691 16692 16693 17178 17179 17180 17181 17182 17183 17184 17185
    2     1     2     2     2     2     2     2     2     2     2     1
17186 17187 17649 17650 17651 17652 17653 17654 17655 17656 17657 17658
    2     1     2     2     2     2     1     2     2     2     2     2
17659 17660 18108 18109 18110 18111 18112 18113 18114 18115 18116 18117
    2     2     2     2     1     2     2     2     2     2     2     1
18118 18586 18587 18588 18589 18590 18591 18592 18593 18594 18595 18996
    2     2     2     2     2     2     1     2     2     2     2     2
18997 18998 18999 19000 19001 19002 19003 19004 19005 19006 19007 19008
    2     2     2     2     2     2     2     2     2     2     2     2
19009 19010 19011 19012 19013 19014 19475 19476 19477 19478 19479 19480
    2     2     2     2     2     2     2     2     2     2     2     2

    2     2     2     2     2     1     1     2     2     2     2     2
28888 28889 28890 28891 28892 28893 28894 28895 28896 28897 28898 28899
    2     2     2     2     2     2     2     2     2     2     2     2
28900 28901 28902 28903 28904 28905 28906 28907 29418 29419 29420 29421
    2     2     2     2     2     2     2     2     2     1     2     2
29422 29423 29424 29425 29426 29427 29428 29429 29430 29431 29432 29433
    2     2     2     1     2     2     2     2     2     2     2     2
29434 29435 29436 29437 29438 29439 29440 29441 29442 29968 29969 29970
    2     2     2     1     2     2     2     2     2     2     2     1
29971 29972 29973 29974 29975 29976 29977 29978 29979 29980 29981 29982
    2     2     2     2     2     2     2     2     2     2     2     2
30487 30488 30489 30490 30491 30492 30493 30494 30495 30496 30497 30498
    2     2     2     2     2     2     2     2     2     2     2     2
31051 31052 31053 31054 31055 31056 31057 31058 31059 31060 31061 31062
    2     2     2     2     2     2     2     2     2     2     2     2
31063 31064 31065 31066 31723 31724 31725 31726 31727 31728 31729 31730
    2     2     2     2     2     2     2     2     1     2     2     2
31731 31732 31733 31734 31735 31736 31737 31738 31739 31740 31741
    2     2     2     2     2     1     2     1     2     2     2

Within cluster sum of squares by cluster:
[1] 1.614255e+13 1.783839e+14
 (between_SS / total_SS =  95.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
>
```

## Step 13

Plot the result.

```
#Evaluating cluster Analysis

library(ggfortify)
autoplot(KM,Crime_Data,frame=TRUE)  #cluster plot
```



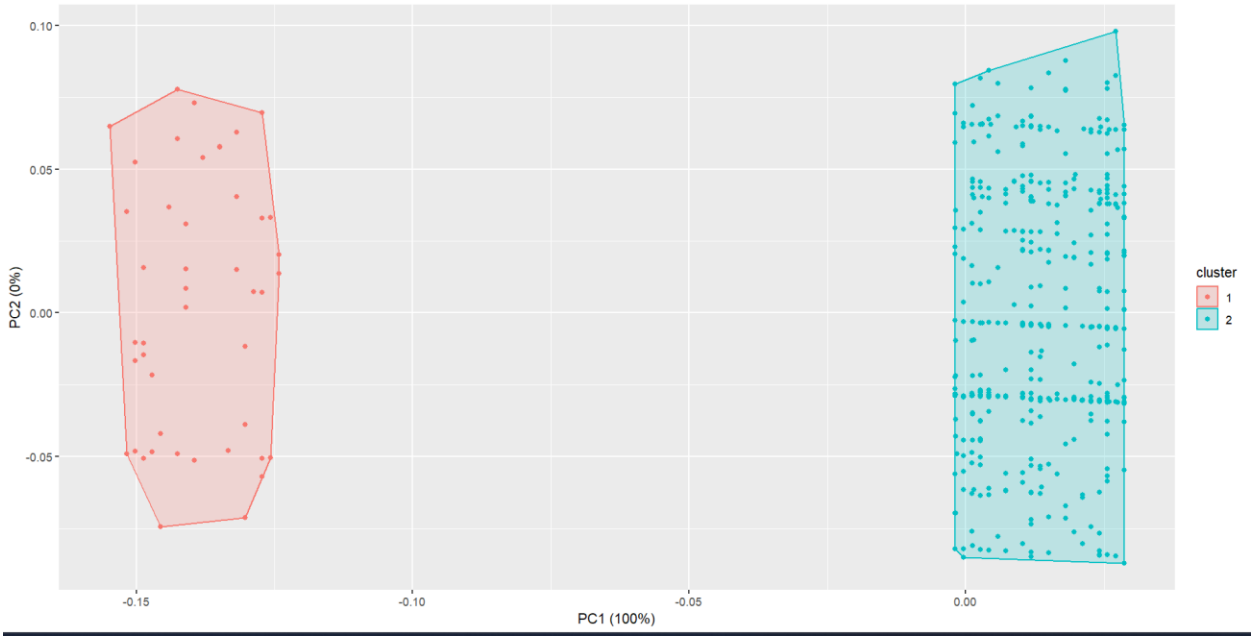

## Step 14

Get the average positions of the clusters.

```
> KM$centers
      DR_NO DATE OCC TIME.OCC     AREA AREA NAME Part 1-2   Crm Cd Crm Cd Desc Vict Age Vict Sex
1 241060003 20.13333 104.7778 10.55556  9.622222 1.466667 581.4222    26.11111 16.84444 1.422222
2 231107421 14.49127 102.0218 10.88428 10.096070 1.558952 527.3996    16.17249 16.85153 1.469432
  Premis Cd Premis Desc      LAT       LON
1  297.1111    41.11111 34.08797 -118.3410
2  297.9934    39.75983 34.00365 -118.0983
>
```

## 6) Result analysis and Discussion

In this section, we will describe the result of the clustering plot which we have obtained using the k – means clustering and the elbow method.




According to the plot, there are two clusters which represents the crimes done by teenagers in the city of Los Angeles from 2023-12-01 to 2023-12-31. Cluster 1 represents less no of points compared to the cluster 2. With the identification of two distinct clusters, we can get a significant understanding of the diverse nature of criminal incidents. These clusters may signify different crime profiles, suggesting the need for targeted intervention strategies.

## 7) Conclusion

Crime in Los Angeles is a complex issue influenced by various factors. While overall crime rates have fluctuated, efforts in community policing and social programs have shown some positive impact. The cluster analysis of the crime dataset in Los Angeles has disclosed valuable insights into the underlying patterns within the city's criminal activities. This cluster analysis provides actionable insights, enabling law enforcement to deploy resources strategically and address the specific challenges associated with different crime clusters in Los Angeles.

## 8) References

https://catalog.data.gov/dataset/crime-data-from-2020-to-present/resource/5eb6507e-fa82-4595-a604-023f8a326099

# Power BI Dashboard for Crimes in the city of Los Angeles

## 1) Introduction

Los Angeles is an American metropolitan city, known for its diverse population and historic buildings, has a lot of criminal challenges. Burglary and vehicle theft, as well as serious crimes including robbery and assault, are prevalent. This dashboard is based on the crime incidents done by teenagers in the city of Los Angeles from 2023-12-01 to 203-12-31. The importance of this dashboard is that some visualization in this are based on the R codes that we have used in the previous clustering task (Task 02). Through this dashboard, viewers can get a clear understanding about crimes occurred in Los Angeles.

## 2) Exploration of the dataset

To create the dashboard we used the same dataset which we have used for the clustering analysis in task 02. The data set had 28 columns and 892935 rows including NULL values. To get a proper data set, we have filtered the NULL values from the dataset.

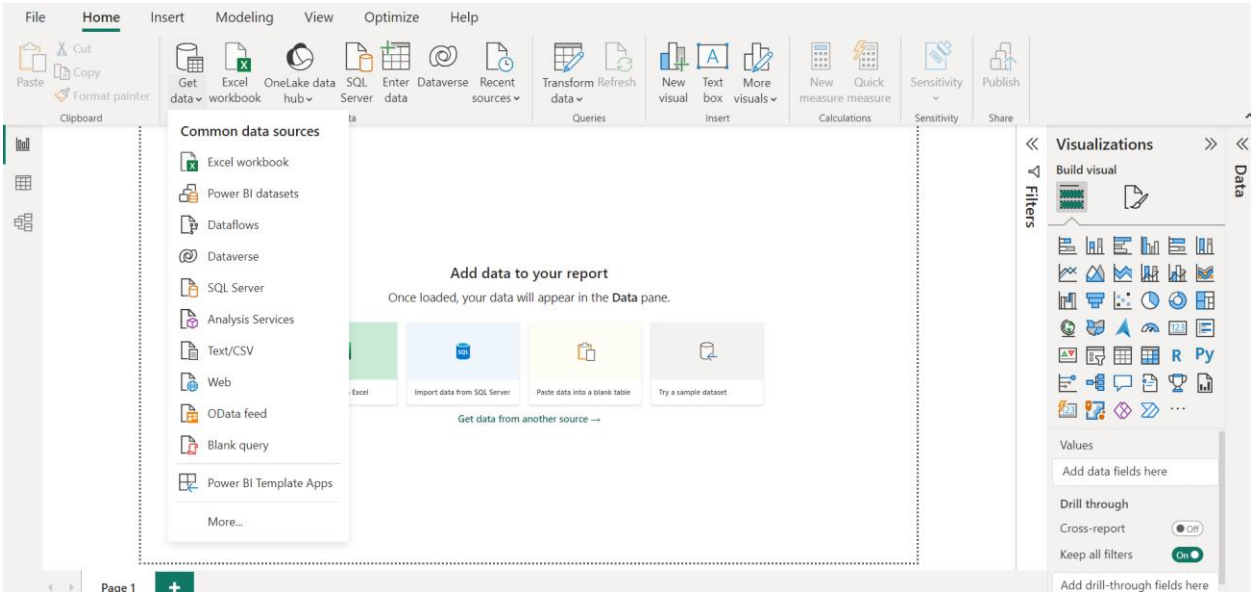
## 3) Dashboard Design & Implementation

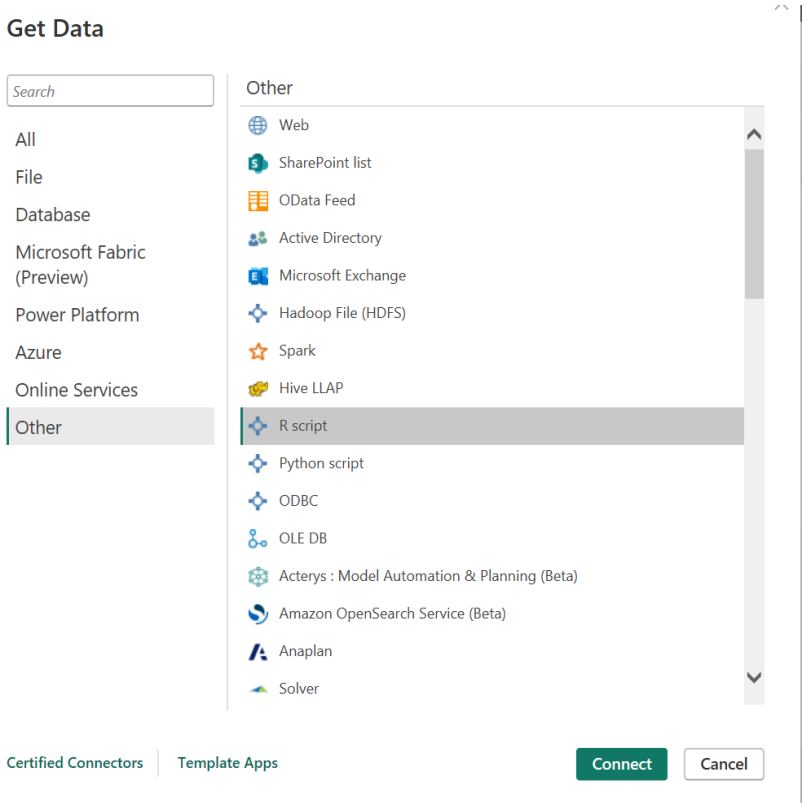
The following are the main steps that we followed to create the dashboard.

**Step 01**

Importing the dataset to Power BI by using R script.

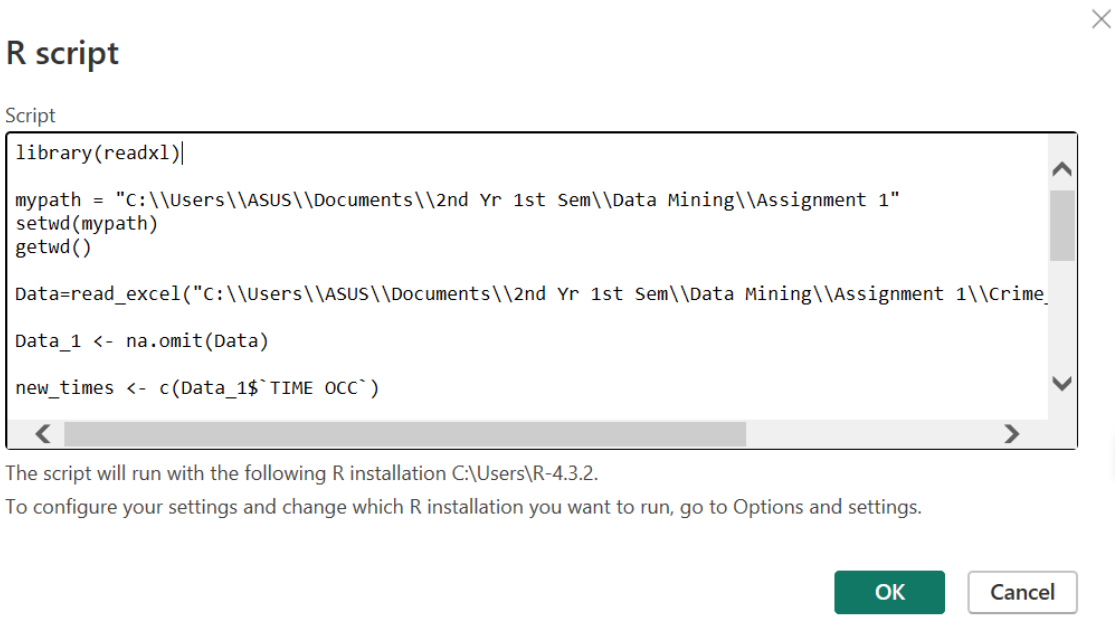
Start POWER BI → GET DATA → MORE → OTHER → R SCRIPT → CONNECT






**Step 02**

Enter the following R code to import the dataset and click OK.

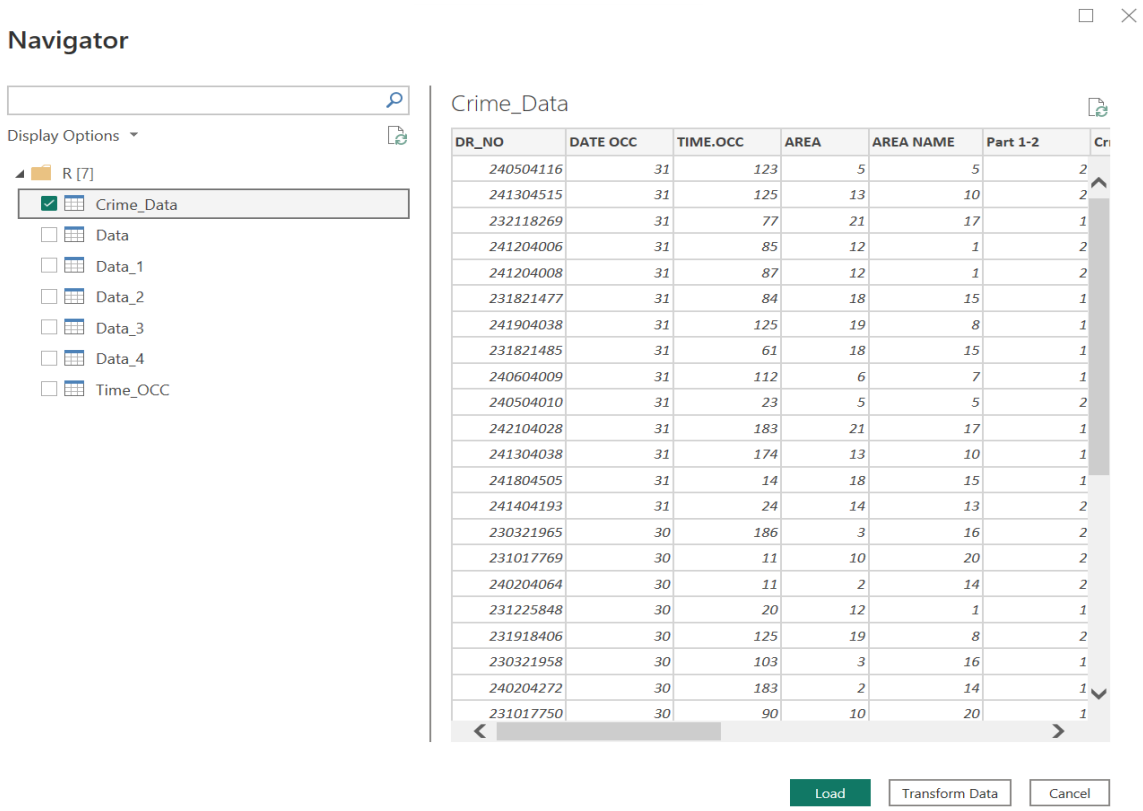


```r
library(dplyr)
library(ggplot2)
library(stats)
library(ggfortify)
library(cluster)
library(factoextra)

library(readxl)

mypath = "C:\\Users\\ASUS\\Documents\\2nd Yr 1st Sem\\Data Mining\\Assignment 1"
setwd(mypath)
getwd()

Data=read_excel("C:\\Users\\ASUS\\Documents\\2nd Yr 1st Sem\\Data Mining\\Assignment 1\\Crime_Data_from_2020_to_Present.xlsx")

Data_1 <- na.omit(Data)

new_times <- c(Data_1$`TIME OCC`)

formatted_times <- sprintf("%02d:%02d", new_times %/% 100, new_times %% 100)

Time_OCC<- data.frame(`TIME OCC`= formatted_times)

Data_1[ ,-3]

Data_2=cbind(c(Data_1[ , -3]),Time_OCC)

Data_3=Data_2[ ,c(1,2,14,3:13)]

Data_3$`DATE OCC` <- as.Date(Data_3$`DATE OCC`, format = "%Y-%m-%d")
```

```r
Data_4<- subset(Data_3, `DATE OCC` >= "2023-12-01" & `DATE OCC` <= "2023-12-31")

Crime_Data=Data_4[which(Data_4$`Vict Age` >=13 & Data_4$`Vict Age`<=19), ]

Crime_Data$`DATE OCC`= as.factor(Crime_Data$`DATE OCC`)
Crime_Data$`DATE OCC`= as.numeric(Crime_Data$`DATE OCC`)
table(Crime_Data$`DATE OCC`)

Crime_Data$TIME.OCC=as.factor(Crime_Data$TIME.OCC)
Crime_Data$TIME.OCC=as.numeric(Crime_Data$TIME.OCC)
table(Crime_Data$TIME.OCC)

Crime_Data$`AREA NAME`= as.factor(Crime_Data$`AREA NAME`)
Crime_Data$`AREA NAME`= as.numeric(Crime_Data$`AREA NAME`)
table(Crime_Data$`AREA NAME`)


Crime_Data$`Crm Cd Desc`= as.factor(Crime_Data$`Crm Cd Desc`)
Crime_Data$`Crm Cd Desc`= as.numeric(Crime_Data$`Crm Cd Desc`)
table(Crime_Data$`Crm Cd Desc`)

Crime_Data$`Vict Sex`= as.factor(Crime_Data$`Vict Sex`)
Crime_Data$`Vict Sex`= as.numeric(Crime_Data$`Vict Sex`)
table(Crime_Data$`Vict Sex`)

Crime_Data$`Premis Desc`= as.factor(Crime_Data$`Premis Desc`)
Crime_Data$`Premis Desc`= as.numeric(Crime_Data$`Premis Desc`)
table(Crime_Data$`Premis Desc`)

anyNA(Crime_Data)
```

**Step 03**

Select our final dataset (after editing from R) and load the dataset from the Navigator.

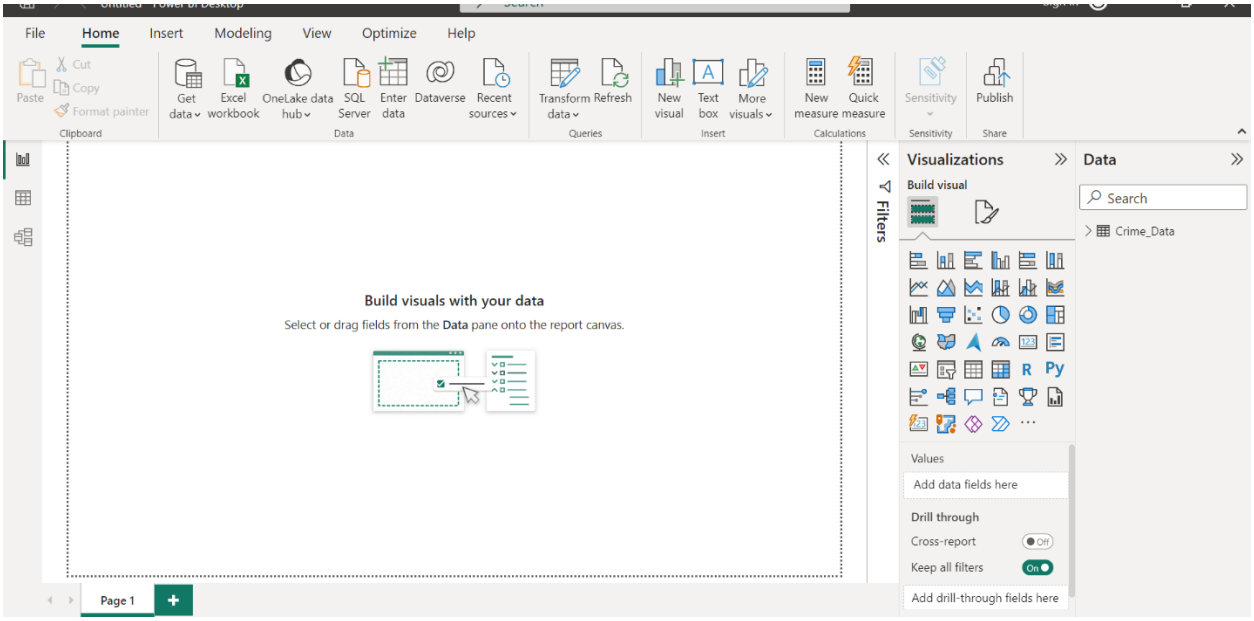


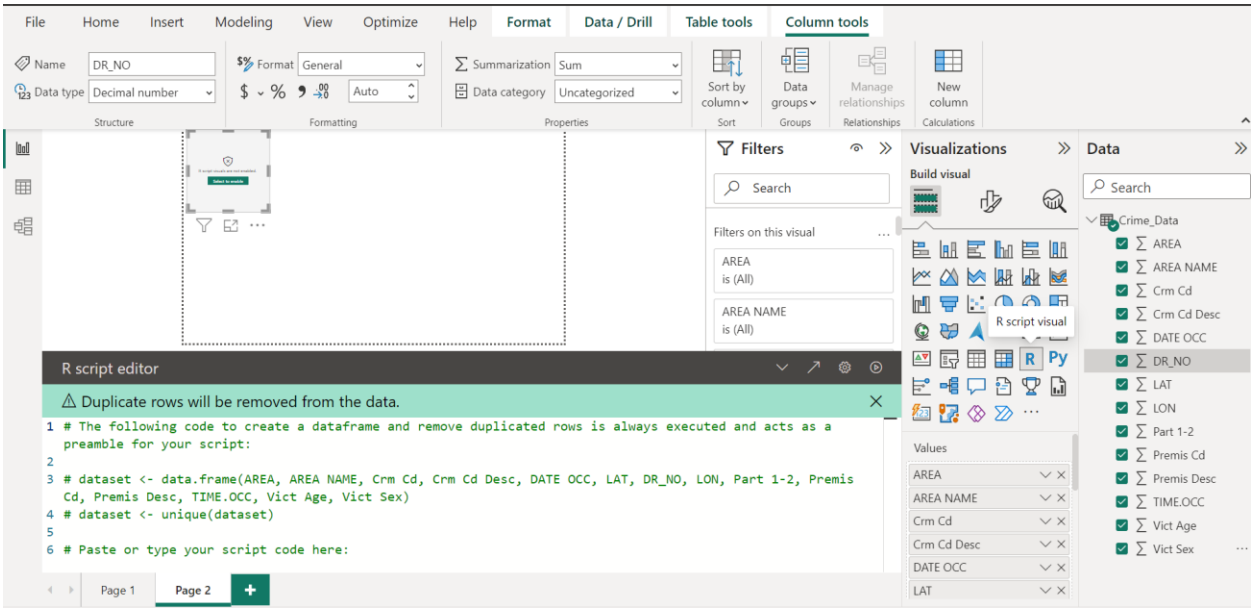

**Step 04**

**Get the WSS plot in clustering**

Open R Script Visual in the visualizations side and select all Numeric Columns in the Data Set.

Then enter the full R code in the R script editor box. (Don't remove any sentence when the R script editor opens.)



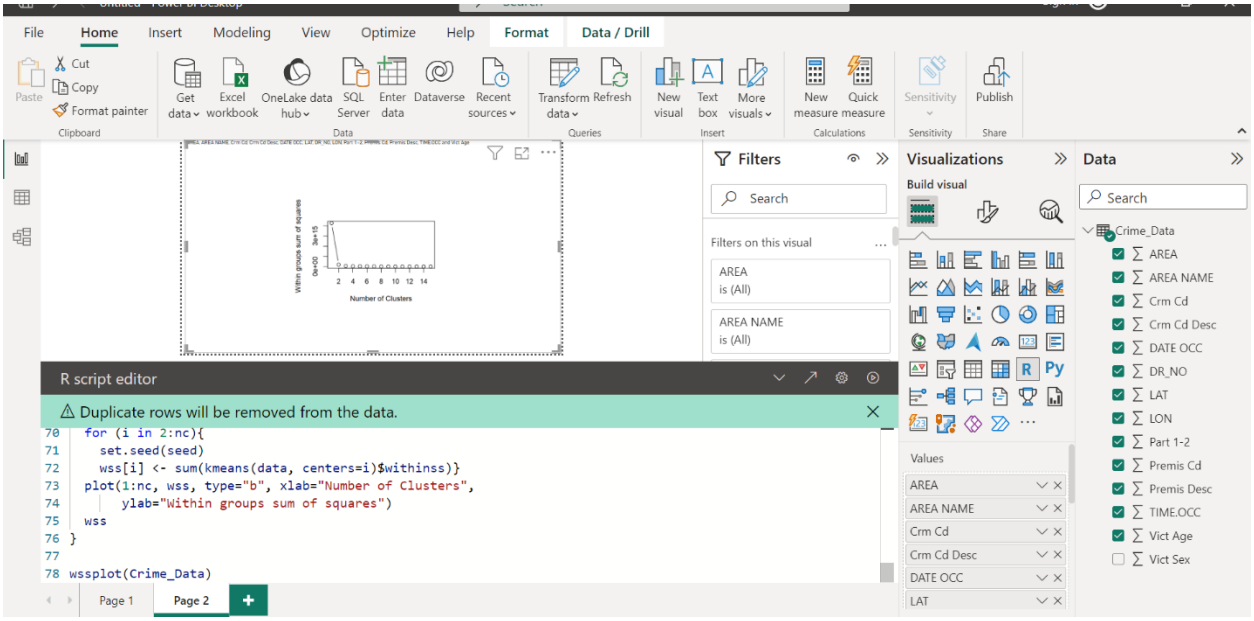




Enter the following code in the R script editor to get the WSS plot in clustering.

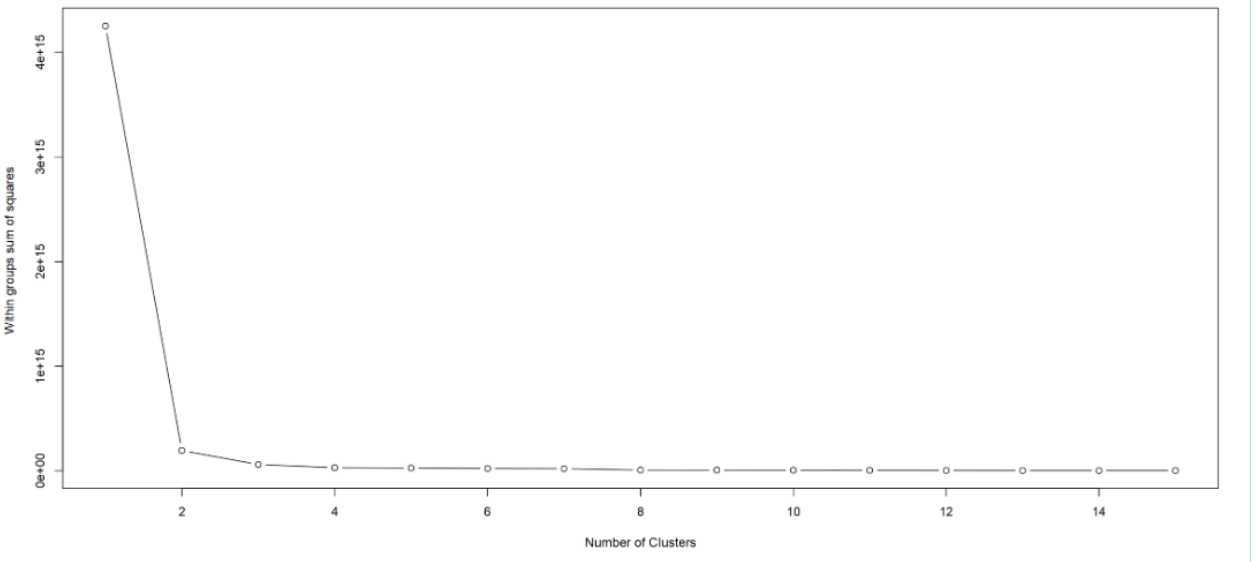
```
library(dplyr)
library(ggplot2)
library(stats)
library(ggfortify)
library(cluster)
library(factoextra)

library(readxl)

mypath = "C:\\Users\\ASUS\\Documents\\2nd Yr 1st Sem\\Data Mining\\Assignment 1"
setwd(mypath)
getwd()

Data=read_excel("C:\\Users\\ASUS\\Documents\\2nd Yr 1st Sem\\Data Mining\\Assignment 1\\Crime_Data_from_2020_to_Present.xlsx")

Data_1 <- na.omit(Data)

new_times <- c(Data_1$`TIME OCC`)

formatted_times <- sprintf("%02d:%02d", new_times %/% 100, new_times %% 100)

Time_OCC<- data.frame(`TIME OCC`= formatted_times)

Data_1[ ,-3]

Data_2=cbind(c(Data_1[ , -3]),Time_OCC)

Data_3=Data_2[ ,c(1,2,14,3:13)]

Data_3$`DATE OCC` <- as.Date(Data_3$`DATE OCC`, format = "%Y-%m-%d")
```

```
Data_4<- subset(Data_3, `DATE OCC` >= "2023-12-01" & `DATE OCC` <= "2023-12-31")

Crime_Data=Data_4[which(Data_4$`Vict Age` >=13 & Data_4$`Vict Age`<=19), ]

Crime_Data$`DATE OCC`= as.factor(Crime_Data$`DATE OCC`)
Crime_Data$`DATE OCC`= as.numeric(Crime_Data$`DATE OCC`)
table(Crime_Data$`DATE OCC`)

Crime_Data$TIME.OCC=as.factor(Crime_Data$TIME.OCC)
Crime_Data$TIME.OCC=as.numeric(Crime_Data$TIME.OCC)
table(Crime_Data$TIME.OCC)

Crime_Data$`AREA NAME`= as.factor(Crime_Data$`AREA NAME`)
Crime_Data$`AREA NAME`= as.numeric(Crime_Data$`AREA NAME`)
table(Crime_Data$`AREA NAME`)


Crime_Data$`Crm Cd Desc`= as.factor(Crime_Data$`Crm Cd Desc`)
Crime_Data$`Crm Cd Desc`= as.numeric(Crime_Data$`Crm Cd Desc`)
table(Crime_Data$`Crm Cd Desc`)

Crime_Data$`Vict Sex`= as.factor(Crime_Data$`Vict Sex`)
Crime_Data$`Vict Sex`= as.numeric(Crime_Data$`Vict Sex`)
table(Crime_Data$`Vict Sex`)

Crime_Data$`Premis Desc`= as.factor(Crime_Data$`Premis Desc`)
Crime_Data$`Premis Desc`= as.numeric(Crime_Data$`Premis Desc`)
table(Crime_Data$`Premis Desc`)

anyNA(Crime_Data)
```

```
wssplot <- function(data, nc=15, seed=123){
  wss <- (nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc){
    set.seed(seed)
    wss[i] <- sum(kmeans(data, centers=i)$withinss)}
  plot(1:nc, wss, type="b", xlab="Number of Clusters",
       ylab="Within groups sum of squares")
  wss
}

wssplot(Crime_Data)
```




The WSS plot can be seen as follows:




A WSS plot is a way to visualize the within-cluster sum of squares for different numbers of clusters in k-means clustering. It can help you choose the optimal number of clusters based on the elbow method. One package that can help you make a WSS plot in R is the "factoextra package". Another package that can make a WSS plot in R is the "cluster package". You can see that the elbow point is at k = 2, which means that 2 clusters are the optimal number for the Crime data set we used for clustering.

## Step 05

## Get the cluster plot

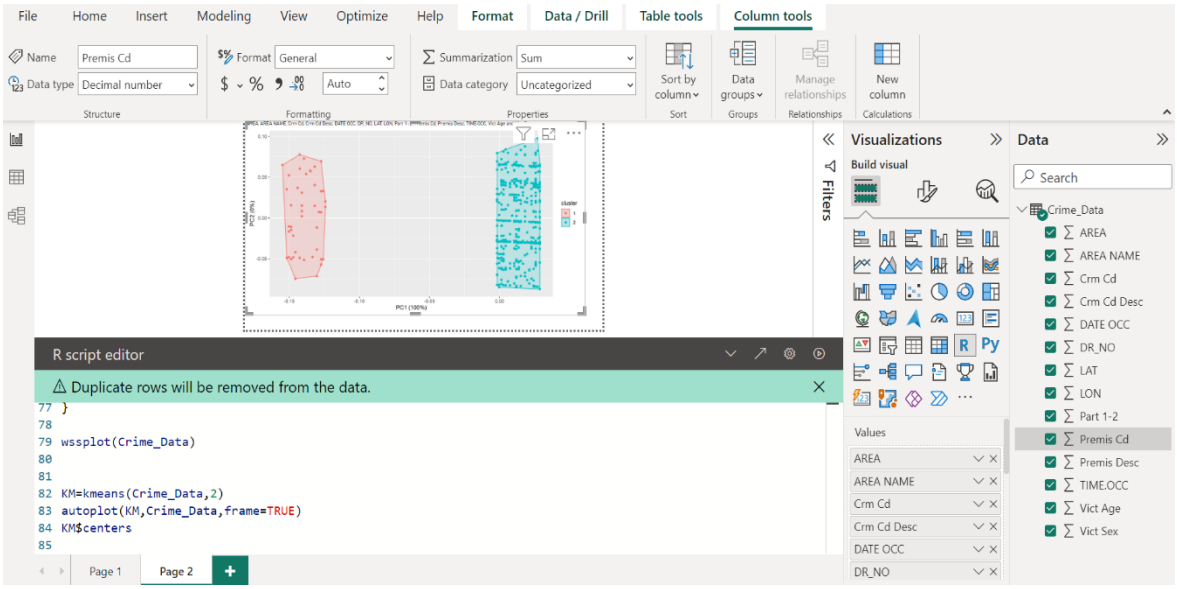
Enter the following code in the R script editor to get the cluster plot.

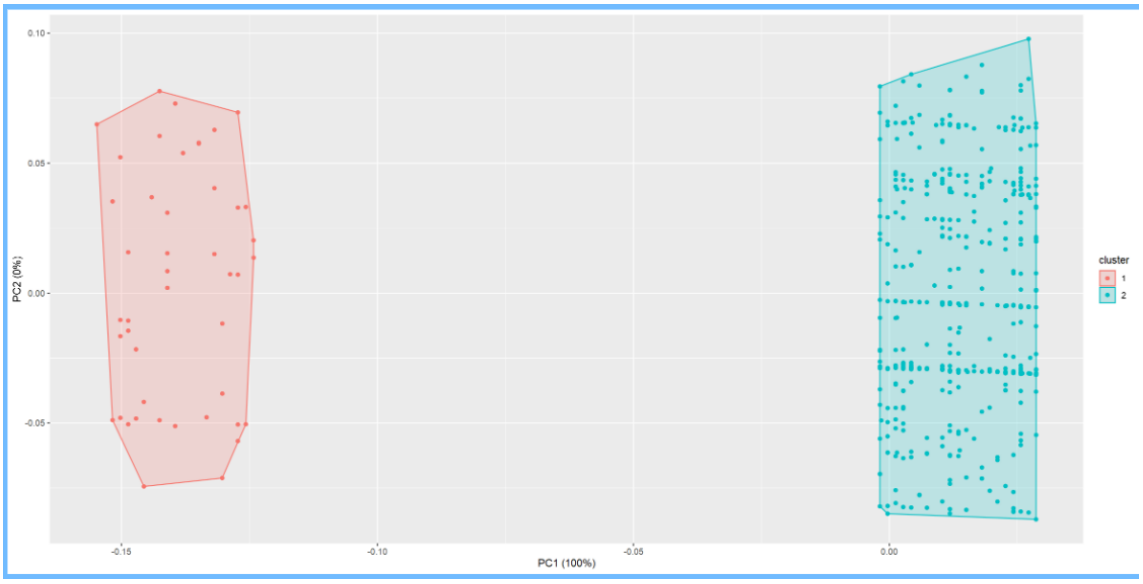
```
library(dplyr)
library(ggplot2)
library(stats)
library(ggfortify)
library(cluster)
library(factoextra)

library(readxl)

mypath = "C:\\Users\\ASUS\\Documents\\2nd Yr 1st Sem\\Data Mining\\Assignment 1"
setwd(mypath)
getwd()

Data=read_excel("C:\\Users\\ASUS\\Documents\\2nd Yr 1st Sem\\Data Mining\\Assignment 1\\Crime_Data_from_2020_to_Present.xlsx")

Data_1 <- na.omit(Data)

new_times <- c(Data_1$`TIME OCC`)

formatted_times <- sprintf("%02d:%02d", new_times %/% 100, new_times %% 100)

Time_OCC<- data.frame(`TIME OCC`= formatted_times)

Data_1[ ,-3]

Data_2=cbind(c(Data_1[ , -3]),Time_OCC)

Data_3=Data_2[ ,c(1,2,14,3:13)]

Data_3$`DATE OCC` <- as.Date(Data_3$`DATE OCC`, format = "%Y-%m-%d")
```

```
Data_4<- subset(Data_3, `DATE OCC` >= "2023-12-01" & `DATE OCC` <= "2023-12-31")

Crime_Data=Data_4[which(Data_4$`Vict Age` >=13 & Data_4$`Vict Age`<=19), ]

Crime_Data$`DATE OCC`= as.factor(Crime_Data$`DATE OCC`)
Crime_Data$`DATE OCC`= as.numeric(Crime_Data$`DATE OCC`)
table(Crime_Data$`DATE OCC`)

Crime_Data$TIME.OCC=as.factor(Crime_Data$TIME.OCC)
Crime_Data$TIME.OCC=as.numeric(Crime_Data$TIME.OCC)
table(Crime_Data$TIME.OCC)

Crime_Data$`AREA NAME`= as.factor(Crime_Data$`AREA NAME`)
Crime_Data$`AREA NAME`= as.numeric(Crime_Data$`AREA NAME`)
table(Crime_Data$`AREA NAME`)


Crime_Data$`Crm Cd Desc`= as.factor(Crime_Data$`Crm Cd Desc`)
Crime_Data$`Crm Cd Desc`= as.numeric(Crime_Data$`Crm Cd Desc`)
table(Crime_Data$`Crm Cd Desc`)

Crime_Data$`Vict Sex`= as.factor(Crime_Data$`Vict Sex`)
Crime_Data$`Vict Sex`= as.numeric(Crime_Data$`Vict Sex`)
table(Crime_Data$`Vict Sex`)

Crime_Data$`Premis Desc`= as.factor(Crime_Data$`Premis Desc`)
Crime_Data$`Premis Desc`= as.numeric(Crime_Data$`Premis Desc`)
table(Crime_Data$`Premis Desc`)

anyNA(Crime_Data)
```

```r
#wss plot to choose maximum number of clusters

wssplot <- function(data, nc=15, seed=123){
  wss <- (nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc){
    set.seed(seed)
    wss[i] <- sum(kmeans(data, centers=i)$withinss)}
  plot(1:nc, wss, type="b", xlab="Number of Clusters",
       ylab="Within groups sum of squares")
  wss
}

wssplot(Crime_Data)

#Spotting the k means in the curve in order to choose the optimum number of cluster=2

KM=kmeans(Crime_Data,2)


#Evaluating cluster Analysis
autoplot(KM,Crime_Data,frame=TRUE)  #cluster plot

#cluster centers
KM$centers
```

The cluster plot can be seen as follows:



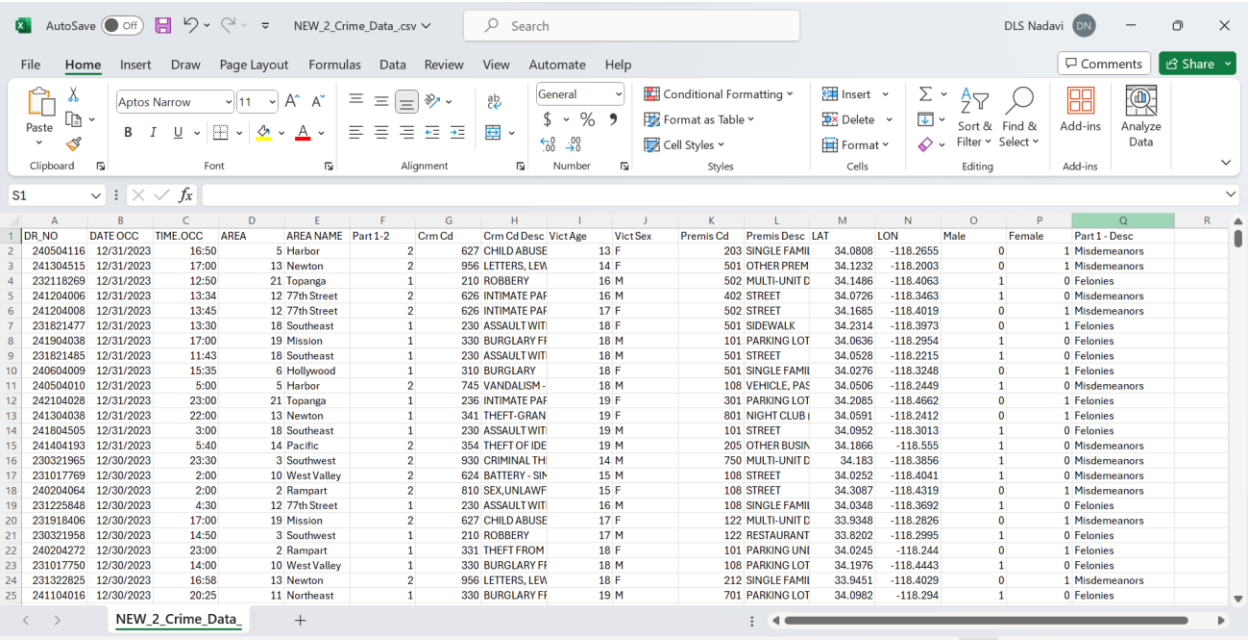



The Cluster plot is a type of graphical display that shows the relationship between data points in a set. A cluster plot can help identify groups or clusters of similar data points based on some criteria, such as distance, similarity, or density. A cluster plot can also show outliers or anomalies that do not belong to any cluster. The cluster plot also shows the variation and diversity within each cluster, such as the different styles of writing the same digit.

## Step 06

## Visualize the charts

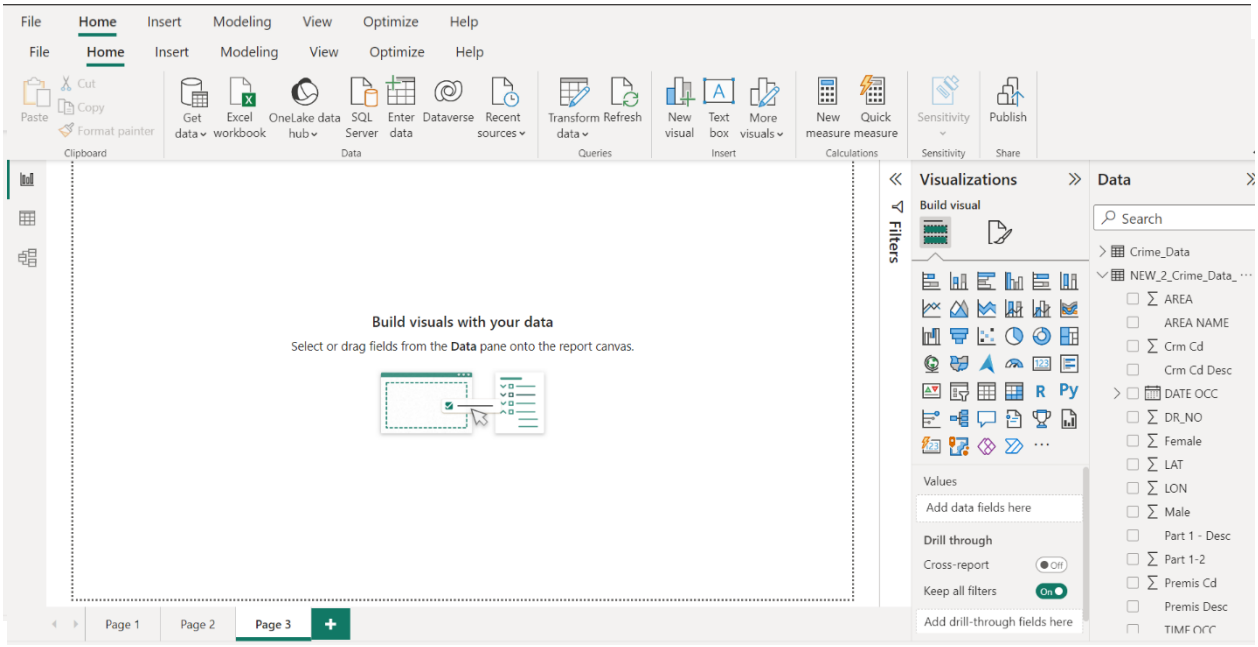
a) Add pages from the icon below the dashboard & add new three columns at last to the EXCEL sheet as follows.

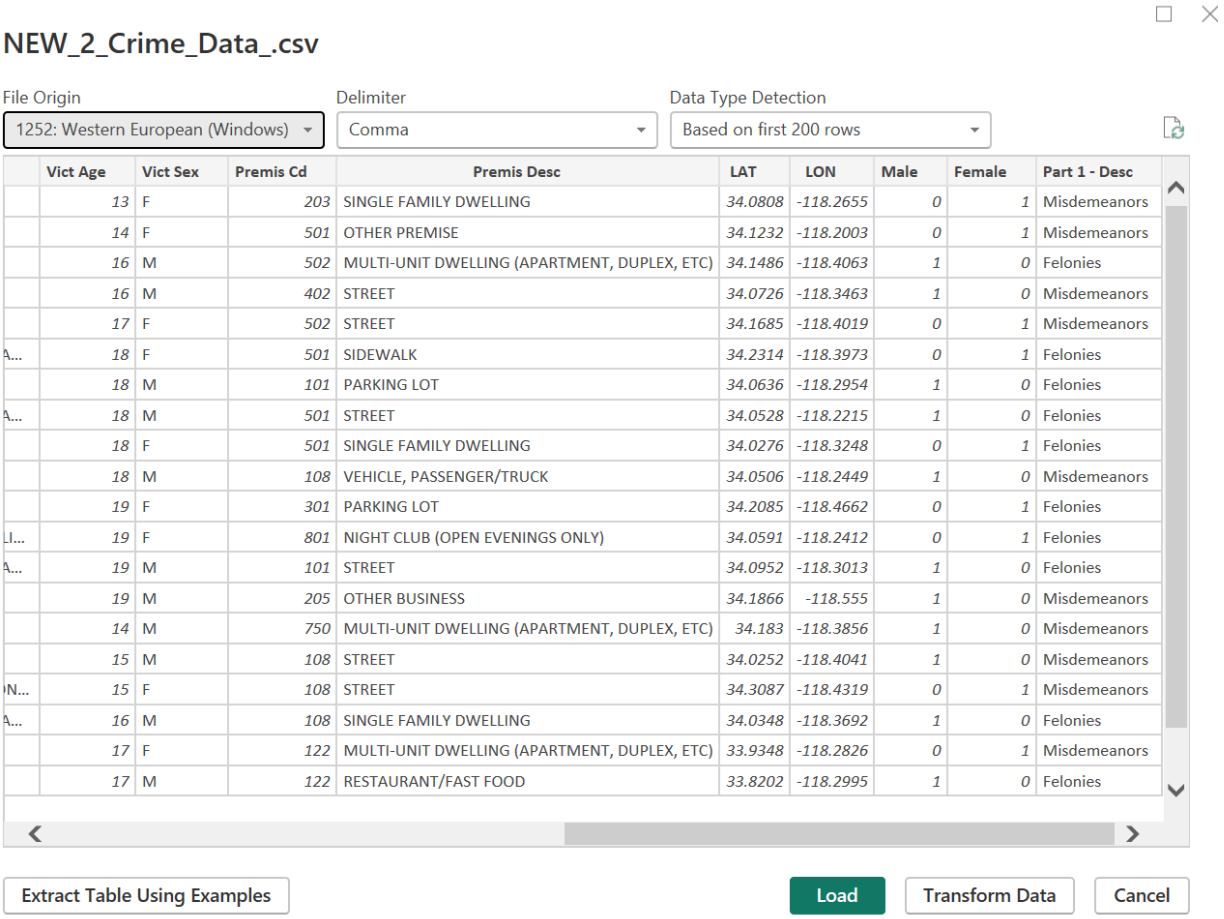



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | DR_NO | DATE OCC | TIME.OCC | AREA | AREA NAME | Part 1-2 | Crm Cd | Crm Cd Desc | Vict Age | Vict Sex | Premis Cd | Premis Desc | LAT | LON | Male | Female | Part 1 - Desc | |
| 2 | 240504116 | 12/31/2023 | 16:50 | 5 | Harbor | 2 | 627 | CHILD ABUSE | 13 | F | 203 | SINGLE FAMII | 34.0808 | -118.2655 | 0 | 1 | Misdemeanors | |
| 3 | 241304515 | 12/31/2023 | 17:00 | 13 | Newton | 2 | 956 | LETTERS, LEV | 14 | F | 501 | OTHER PREM | 34.1232 | -118.2003 | 0 | 1 | Misdemeanors | |
| 4 | 232118269 | 12/31/2023 | 12:50 | 21 | Topanga | 1 | 210 | ROBBERY | 16 | M | 502 | MULTI-UNIT D | 34.1486 | -118.4063 | 1 | 0 | Felonies | |
| 5 | 241204006 | 12/31/2023 | 13:34 | 12 | 77th Street | 2 | 626 | INTIMATE PAF | 16 | M | 402 | STREET | 34.0726 | -118.3463 | 1 | 0 | Misdemeanors | |
| 6 | 241204008 | 12/31/2023 | 13:45 | 12 | 77th Street | 2 | 626 | INTIMATE PAF | 17 | F | 502 | STREET | 34.1685 | -118.4019 | 0 | 1 | Misdemeanors | |
| 7 | 231821477 | 12/31/2023 | 13:30 | 18 | Southeast | 1 | 230 | ASSAULT WIT | 18 | F | 501 | SIDEWALK | 34.2314 | -118.3973 | 0 | 1 | Felonies | |
| 8 | 241904038 | 12/31/2023 | 17:00 | 19 | Mission | 1 | 330 | BURGLARY FF | 18 | M | 101 | PARKING LOT | 34.0636 | -118.2954 | 1 | 0 | Felonies | |
| 9 | 231821485 | 12/31/2023 | 11:43 | 18 | Southeast | 1 | 230 | ASSAULT WIT | 18 | M | 501 | STREET | 34.0528 | -118.2215 | 1 | 0 | Felonies | |
| 10 | 240604009 | 12/31/2023 | 15:35 | 6 | Hollywood | 1 | 310 | BURGLARY | 18 | F | 501 | SINGLE FAMII | 34.0276 | -118.3248 | 0 | 1 | Felonies | |
| 11 | 240504010 | 12/31/2023 | 5:00 | 5 | Harbor | 2 | 745 | VANDALISM - | 18 | M | 108 | VEHICLE, PAS | 34.0506 | -118.2449 | 1 | 0 | Misdemeanors | |
| 12 | 242104028 | 12/31/2023 | 23:00 | 21 | Topanga | 1 | 236 | INTIMATE PAF | 19 | F | 301 | PARKING LOT | 34.2085 | -118.4662 | 0 | 1 | Felonies | |
| 13 | 241304038 | 12/31/2023 | 22:00 | 13 | Newton | 1 | 341 | THEFT-GRAN | 19 | F | 801 | NIGHT CLUB ( | 34.0591 | -118.2412 | 0 | 1 | Felonies | |
| 14 | 241804505 | 12/31/2023 | 3:00 | 18 | Southeast | 1 | 230 | ASSAULT WIT | 19 | M | 101 | STREET | 34.0952 | -118.3013 | 1 | 0 | Felonies | |
| 15 | 241404193 | 12/31/2023 | 5:40 | 14 | Pacific | 2 | 354 | THEFT OF IDE | 19 | M | 205 | OTHER BUSIN | 34.1866 | -118.555 | 1 | 0 | Misdemeanors | |
| 16 | 230321965 | 12/30/2023 | 23:30 | 3 | Southwest | 2 | 930 | CRIMINAL TH | 14 | M | 750 | MULTI-UNIT D | 34.183 | -118.3856 | 1 | 0 | Misdemeanors | |
| 17 | 231017769 | 12/30/2023 | 2:00 | 10 | West Valley | 2 | 624 | BATTERY - SIM | 15 | M | 108 | STREET | 34.0252 | -118.4041 | 1 | 0 | Misdemeanors | |
| 18 | 240204064 | 12/30/2023 | 2:00 | 2 | Rampart | 2 | 810 | SEX,UNLAWF | 15 | F | 108 | STREET | 34.3087 | -118.4319 | 0 | 1 | Misdemeanors | |
| 19 | 231225848 | 12/30/2023 | 4:30 | 12 | 77th Street | 1 | 230 | ASSAULT WIT | 16 | M | 108 | SINGLE FAMII | 34.0348 | -118.3692 | 1 | 0 | Felonies | |
| 20 | 231918406 | 12/30/2023 | 17:00 | 19 | Mission | 2 | 627 | CHILD ABUSE | 17 | F | 122 | MULTI-UNIT D | 33.9348 | -118.2826 | 0 | 1 | Misdemeanors | |
| 21 | 230321958 | 12/30/2023 | 14:50 | 3 | Southwest | 1 | 210 | ROBBERY | 17 | M | 122 | RESTAURANT | 33.8202 | -118.2995 | 1 | 0 | Felonies | |
| 22 | 240204272 | 12/30/2023 | 23:00 | 2 | Rampart | 1 | 331 | THEFT FROM | 18 | F | 101 | PARKING UNI | 34.0245 | -118.244 | 0 | 1 | Felonies | |
| 23 | 231017750 | 12/30/2023 | 14:00 | 10 | West Valley | 1 | 330 | BURGLARY FF | 18 | M | 108 | PARKING LOT | 34.1976 | -118.4443 | 1 | 0 | Felonies | |
| 24 | 231322825 | 12/30/2023 | 16:58 | 13 | Newton | 2 | 956 | LETTERS, LEV | 18 | F | 212 | SINGLE FAMII | 33.9451 | -118.4029 | 0 | 1 | Misdemeanors | |
| 25 | 241104016 | 12/30/2023 | 20:25 | 11 | Northeast | 1 | 330 | BURGLARY FF | 19 | M | 701 | PARKING LOT | 34.0982 | -118.294 | 1 | 0 | Felonies | |

b) Import the Excel sheet (Get Data →Text/CSV →Load)

c) Imported data set in Data part.



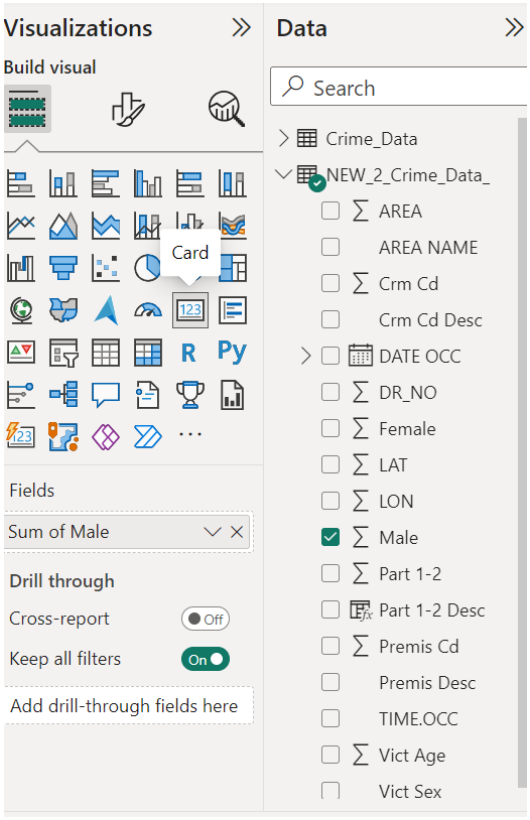




| Vict Age | Vict Sex | Premis Cd | Premis Desc | LAT | LON | Male | Female | Part 1 - Desc |
|---|---|---|---|---|---|---|---|---|
| 13 | F | 203 | SINGLE FAMILY DWELLING | 34.0808 | -118.2655 | 0 | 1 | Misdemeanors |
| 14 | F | 501 | OTHER PREMISE | 34.1232 | -118.2003 | 0 | 1 | Misdemeanors |
| 16 | M | 502 | MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC) | 34.1486 | -118.4063 | 1 | 0 | Felonies |
| 16 | M | 402 | STREET | 34.0726 | -118.3463 | 1 | 0 | Misdemeanors |
| 17 | F | 502 | STREET | 34.1685 | -118.4019 | 0 | 1 | Misdemeanors |
| 18 | F | 501 | SIDEWALK | 34.2314 | -118.3973 | 0 | 1 | Felonies |
| 18 | M | 101 | PARKING LOT | 34.0636 | -118.2954 | 1 | 0 | Felonies |
| 18 | M | 501 | STREET | 34.0528 | -118.2215 | 1 | 0 | Felonies |
| 18 | F | 501 | SINGLE FAMILY DWELLING | 34.0276 | -118.3248 | 0 | 1 | Felonies |
| 18 | M | 108 | VEHICLE, PASSENGER/TRUCK | 34.0506 | -118.2449 | 1 | 0 | Misdemeanors |
| 19 | F | 301 | PARKING LOT | 34.2085 | -118.4662 | 0 | 1 | Felonies |
| 19 | F | 801 | NIGHT CLUB (OPEN EVENINGS ONLY) | 34.0591 | -118.2412 | 0 | 1 | Felonies |
| 19 | M | 101 | STREET | 34.0952 | -118.3013 | 1 | 0 | Felonies |
| 19 | M | 205 | OTHER BUSINESS | 34.1866 | -118.555 | 1 | 0 | Misdemeanors |
| 14 | M | 750 | MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC) | 34.183 | -118.3856 | 1 | 0 | Misdemeanors |
| 15 | M | 108 | STREET | 34.0252 | -118.4041 | 1 | 0 | Misdemeanors |
| 15 | F | 108 | STREET | 34.3087 | -118.4319 | 0 | 1 | Misdemeanors |
| 16 | M | 108 | SINGLE FAMILY DWELLING | 34.0348 | -118.3692 | 1 | 0 | Felonies |
| 17 | F | 122 | MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC) | 33.9348 | -118.2826 | 0 | 1 | Misdemeanors |
| 17 | M | 122 | RESTAURANT/FAST FOOD | 33.8202 | -118.2995 | 1 | 0 | Felonies |

d) Display Charts

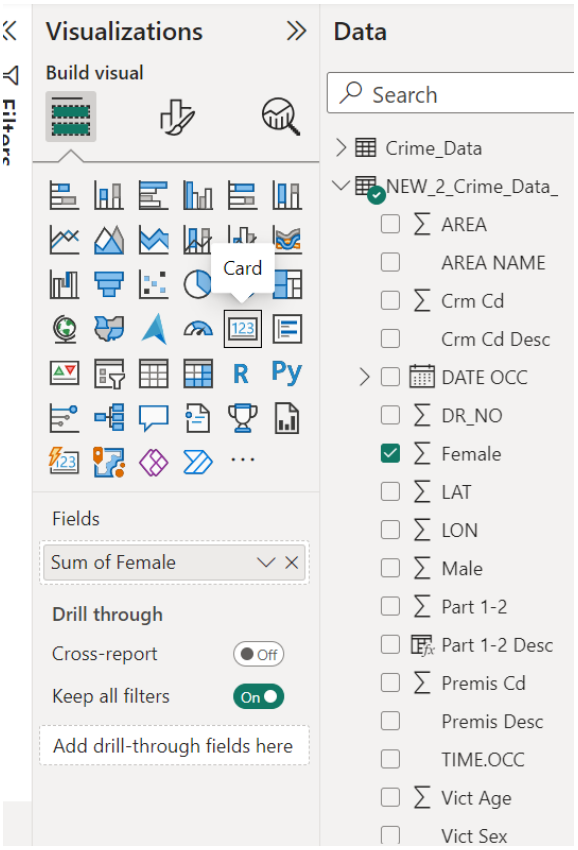
***1. Cards***

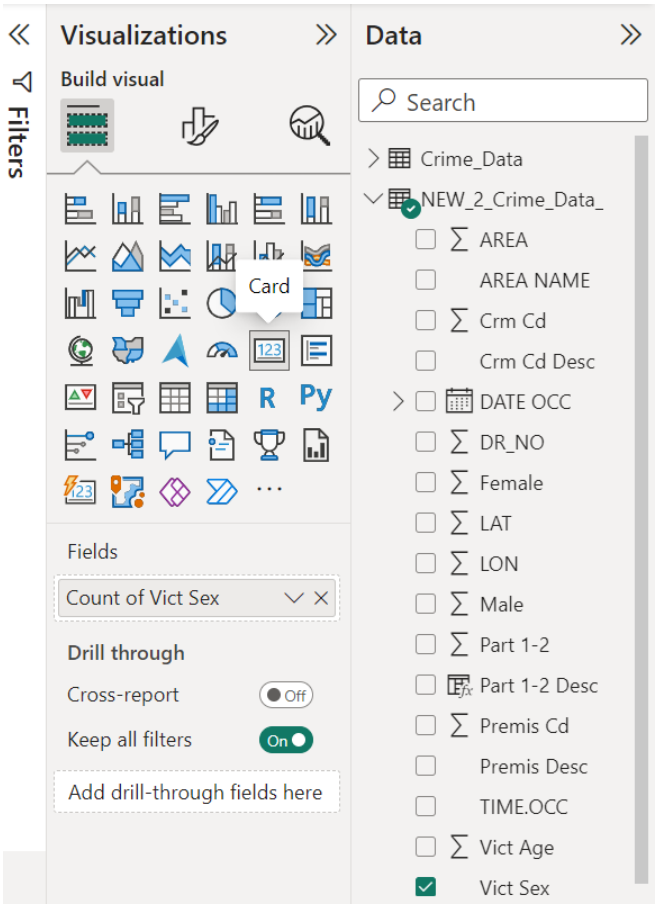
In here there are three cards. The first one represents the number of Male victims and second one shows the number of Female victims, and the third one presents the total number of victims. Overall, there are 502 people and among them there are 228 Males and 274 Females as victims.





Number of Female Offenders
274

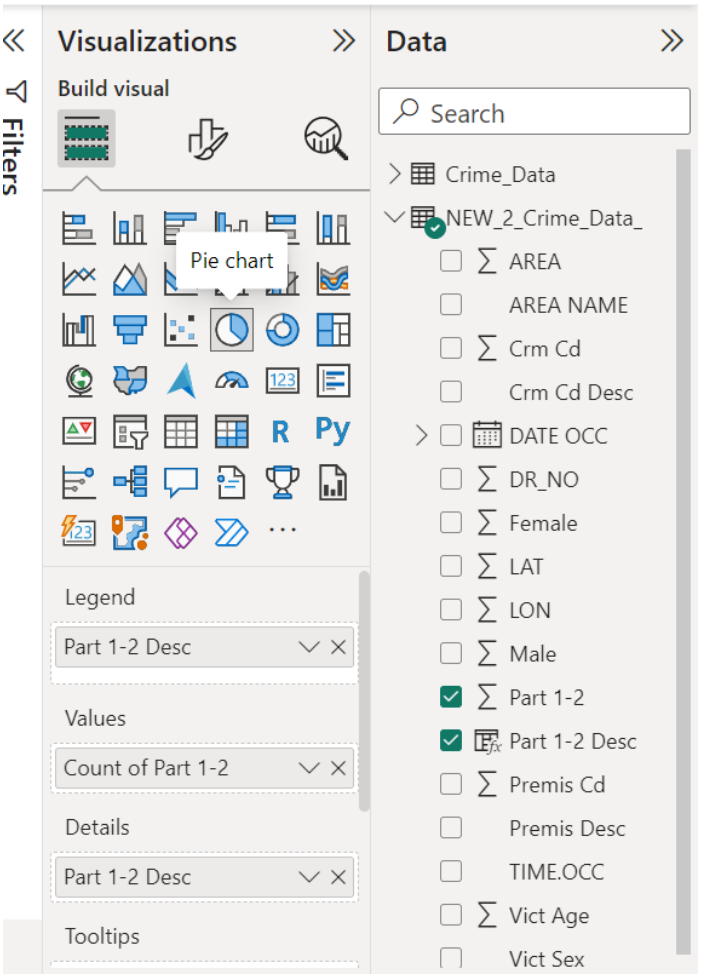

Visualizations
Build visual
Card
Fields
Sum of Female
Drill through
Cross-report
Off
Keep all filters
On
Add drill-through fields here
Data
Search
Crime_Data
NEW_2_Crime_Data_
AREA
AREA NAME
Crm Cd
Crm Cd Desc
DATE OCC
DR_NO
Female
LAT
LON
Male
Part 1-2
Part 1-2 Desc
Premis Cd
Premis Desc
TIME.OCC
Vict Age
Vict Sex

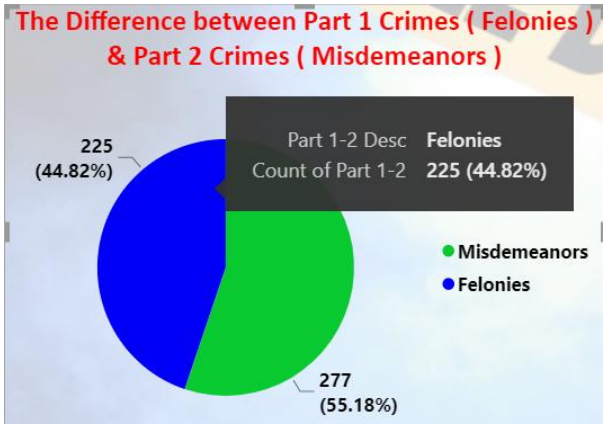

Filters
Visualizations
Build visual
Card
Fields
Count of Vict Sex
Drill through
Cross-report
Off
Keep all filters
On
Add drill-through fields here
Data
Search
Crime_Data
NEW_2_Crime_Data_
AREA
AREA NAME
Crm Cd
Crm Cd Desc
DATE OCC
DR_NO
Female
LAT
LON
Male
Part 1-2
Part 1-2 Desc
Premis Cd
Premis Desc
TIME.OCC
Vict Age
Vict Sex


Total Number Of Offenders
502

### *2. Pie Chart*

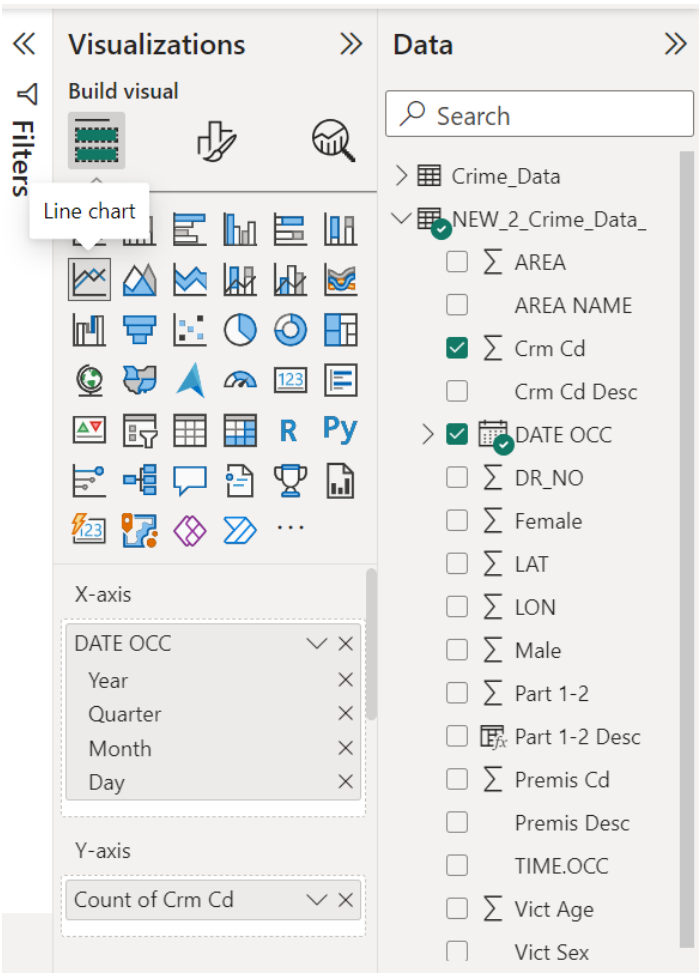
The pie chart with two sections labeled "Misdemeanors"(Part 2) and "Felonies"(Part 1) in our data set. The chart shows that the total percentage of crimes consists of 44.82% misdemeanors and 55.18% felonies. Felonies represent a smaller portion of the total crimes, making up less than half of all crimes. On the other hand, Misdemeanors are more prevalent, accounting for over half of all crimes. This information suggests that there is a lower number of serious crimes (felonies) being committed compared to less serious crimes (misdemeanors).

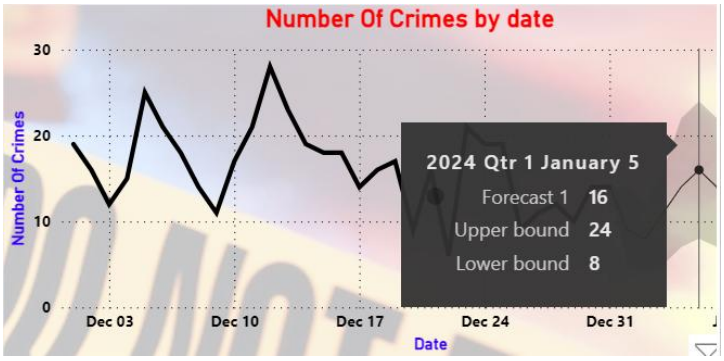


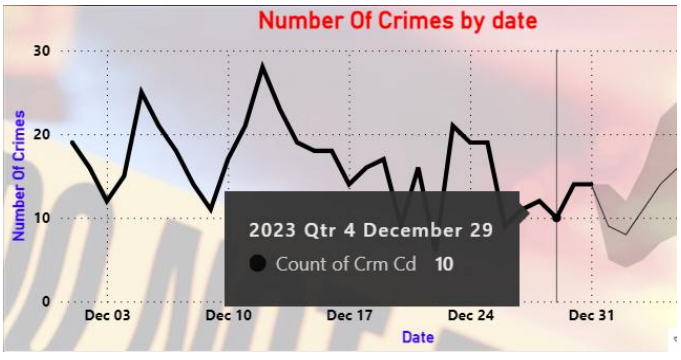



### *3. Line Chart*

The Line chart illustrates the number of crimes by date over a specific period, ranging from December 3rd to January 7th. The x-axis of this chart shows the date of the crime, and the y-axis shows the number of crimes. Notably, there was a peak in crime numbers on December 12th, followed by a significant decline. The chart gives a forecast from December 31 to January 10. The graph provides insights into crime patterns during this time frame.



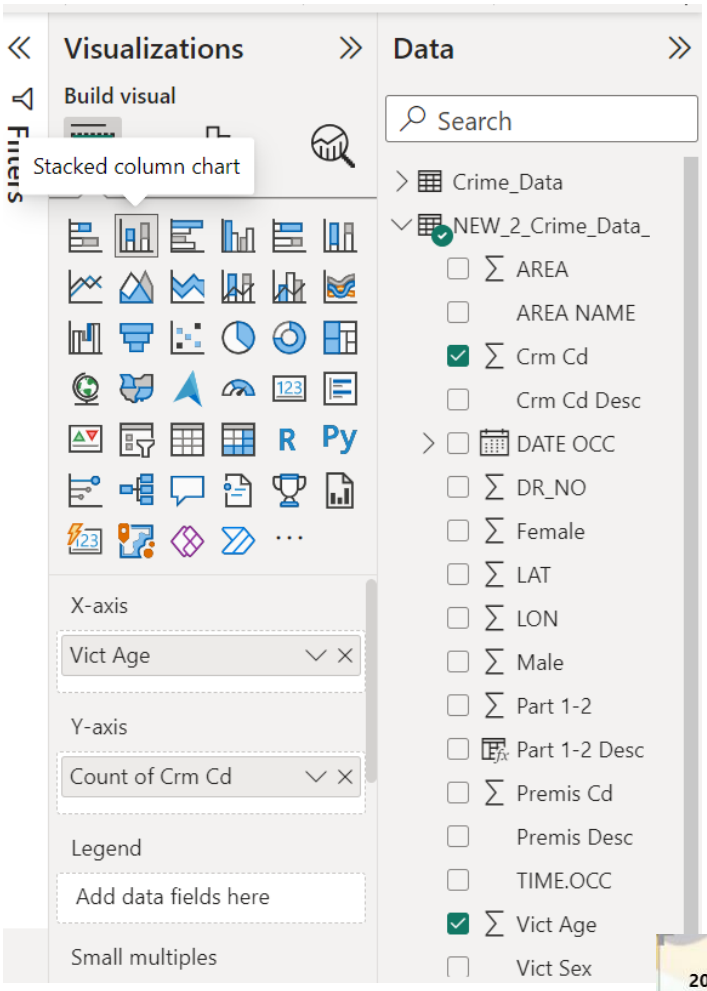



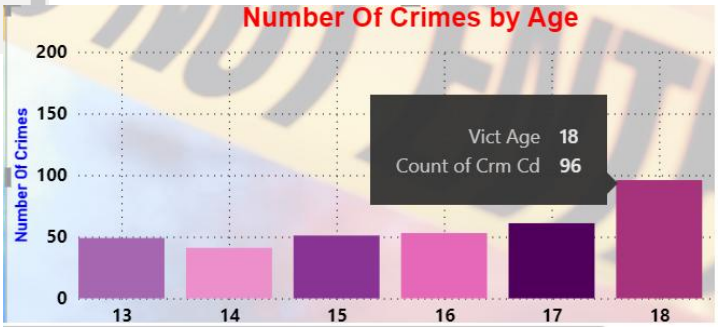



### *4. Stacked Column Chart*

The Column Bar chart represents the number of crimes by age for individuals aged 13 to 19. The chart is displayed as a bar graph with different colors representing each age group. The Y-axis represents the number of crimes, marked at intervals of fifty up to one hundred and fifty. The X-axis denotes the ages of victims (ranging from 13 to 19).

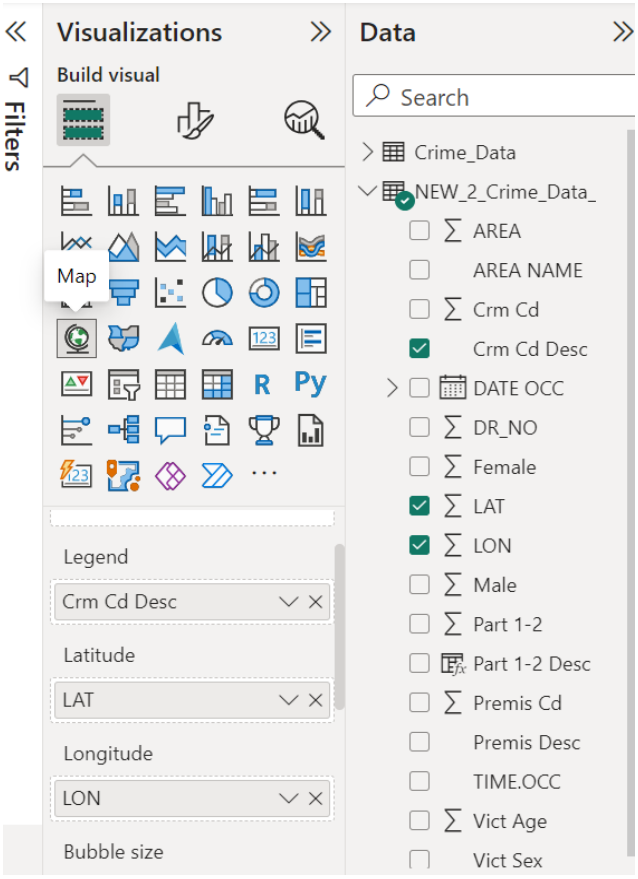
- Ages 13 to 16: These age groups have significantly fewer reported crimes.
- Age 17: There is a noticeable increase in crime numbers.
- Age 18: The highest crime rate is observed in this age group.
- Age 19: Although lower than age 18, the crime count remains significantly higher than ages 13 to 17.

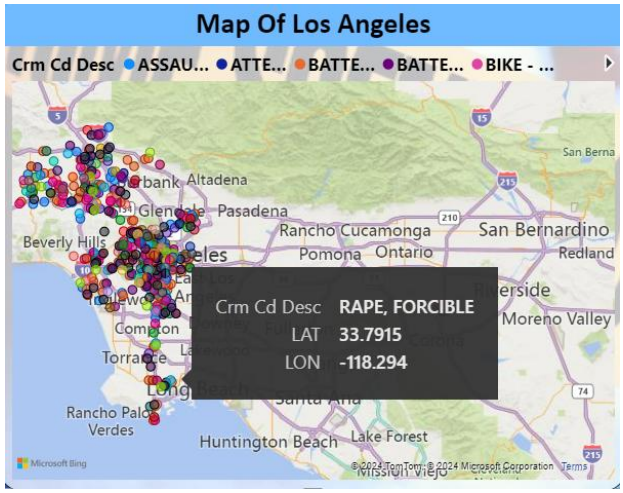


### *5. Map*

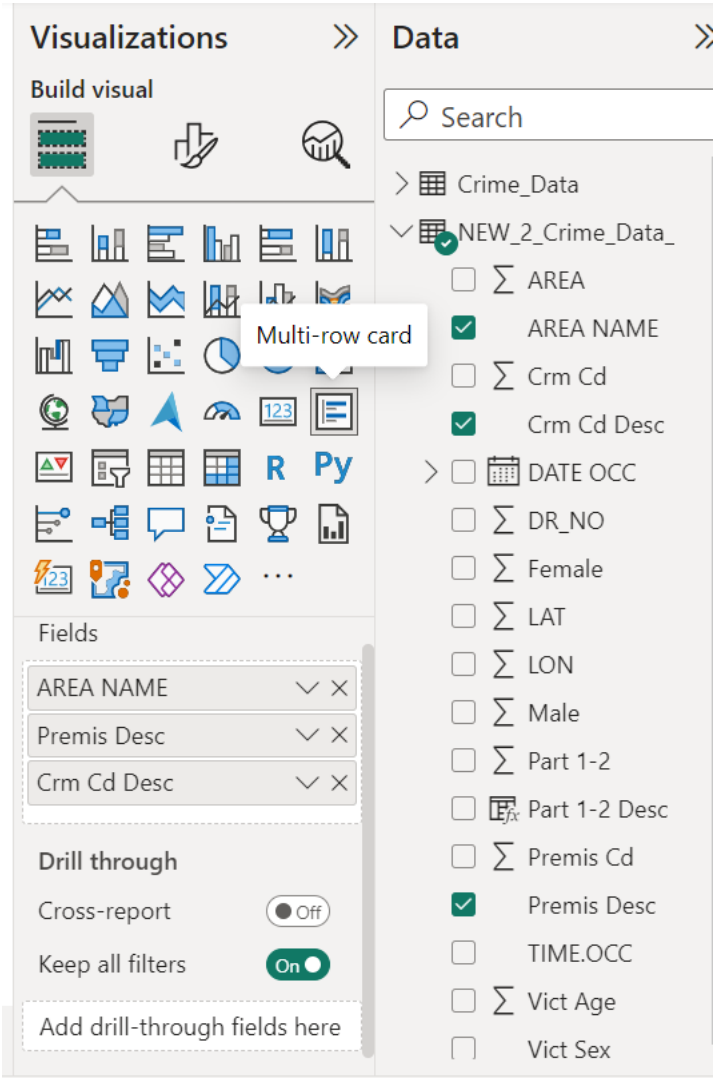
The map of Los Angeles shows the places where the crimes happened around the city of Los Angeles. Various colored circles are spread across different parts of the city representing each crime. Each color likely represents different types of crimes. From the map, we can observe that some areas have a higher concentration of certain crimes compared to others. Overall, the map provides a visual representation of crime concentration in different areas of Los Angeles, which can help law enforcement agencies and policymakers allocate resources and implement strategies to address crime issues in the region.



### *6. Multi – Row Card*

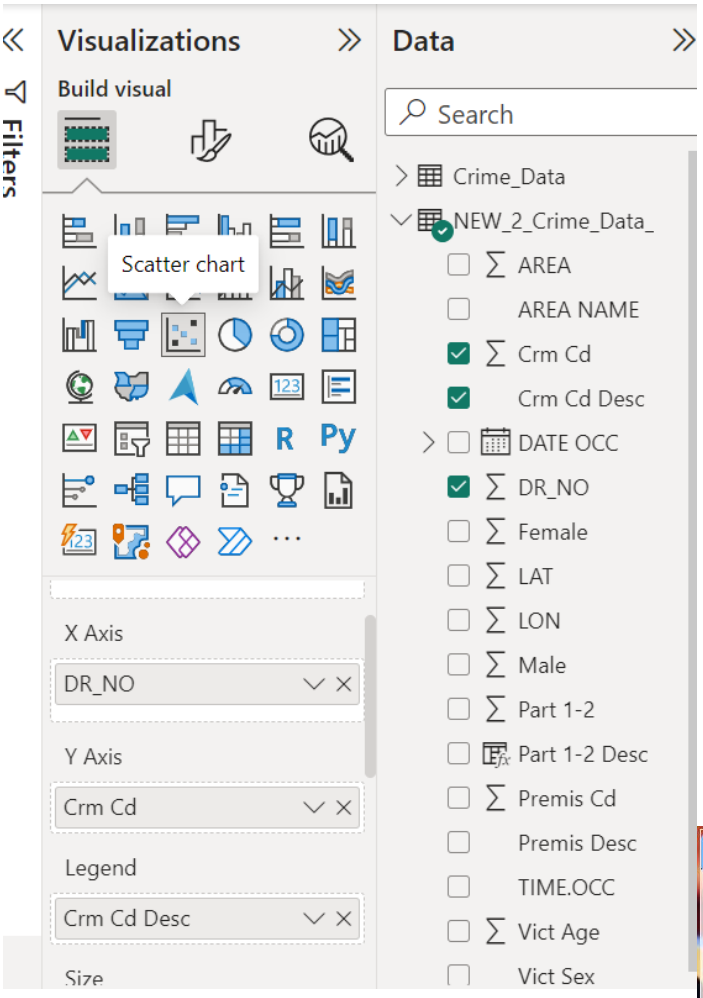
This represents the Area Name, Premise Description, and Crime code description.

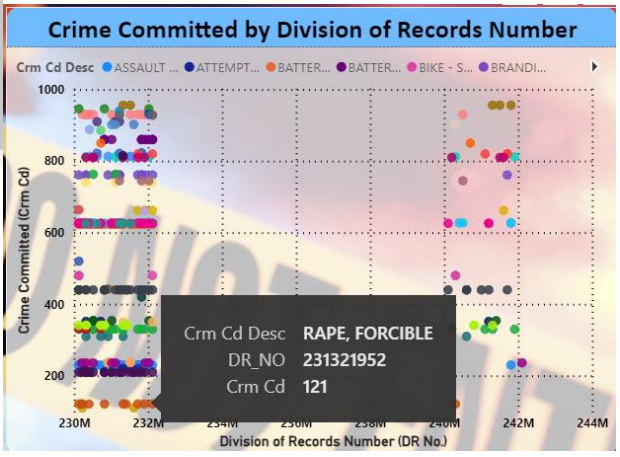



| Rampart | VEHICLE, PASSENGER/TRUCK | RAPE, FORCIBLE |
|---|---|---|
| AREA NAME | Premis Desc | Crm Cd Desc |

### *7. Scatter Chart*

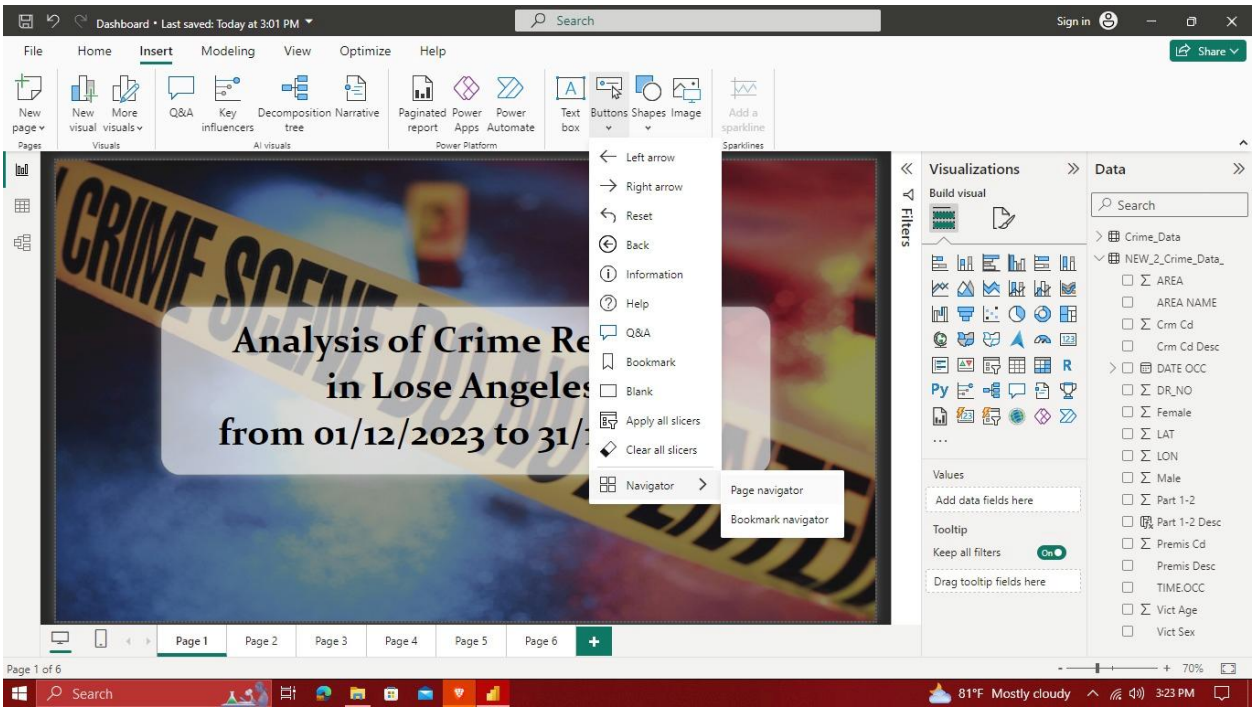
This chart shows the crime committed (Crm Cd) by division of records number (DR No.). There are various crime categories, such as assault, theft, burglary, and narcotics, along the Y-axis, and division of records numbers along the X-axis. Here, the DR No (division of records number) on the X-axis represents a unique identifier for each crime incident reported by the Los Angeles Police Department (LAPD). The first two digits of DR No. means the year in which the crime was reported. Although our dataset contains data about crimes that occurred in 2023, this graph shows that some of those crimes were reported in 2024. The chart provides a visual representation of crime data by division of records number, which can help law enforcement agencies and policymakers identify areas with high crime rates, allocate resources, and implement strategies accordingly.

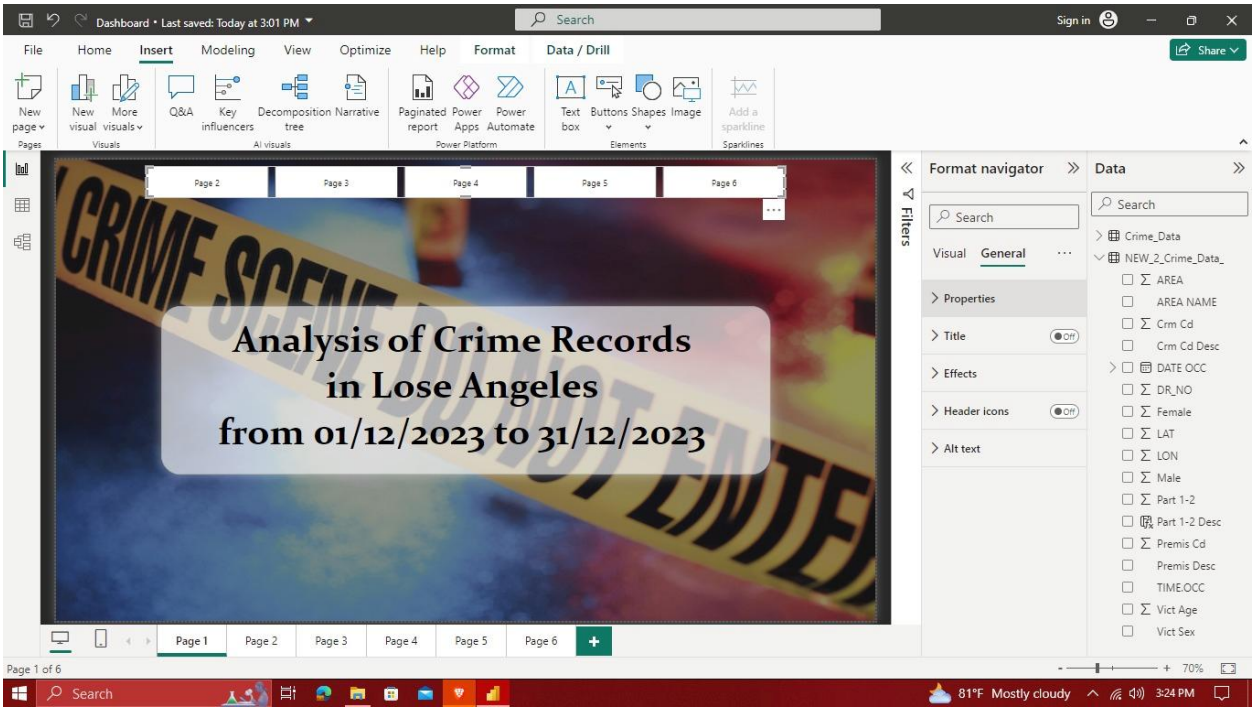


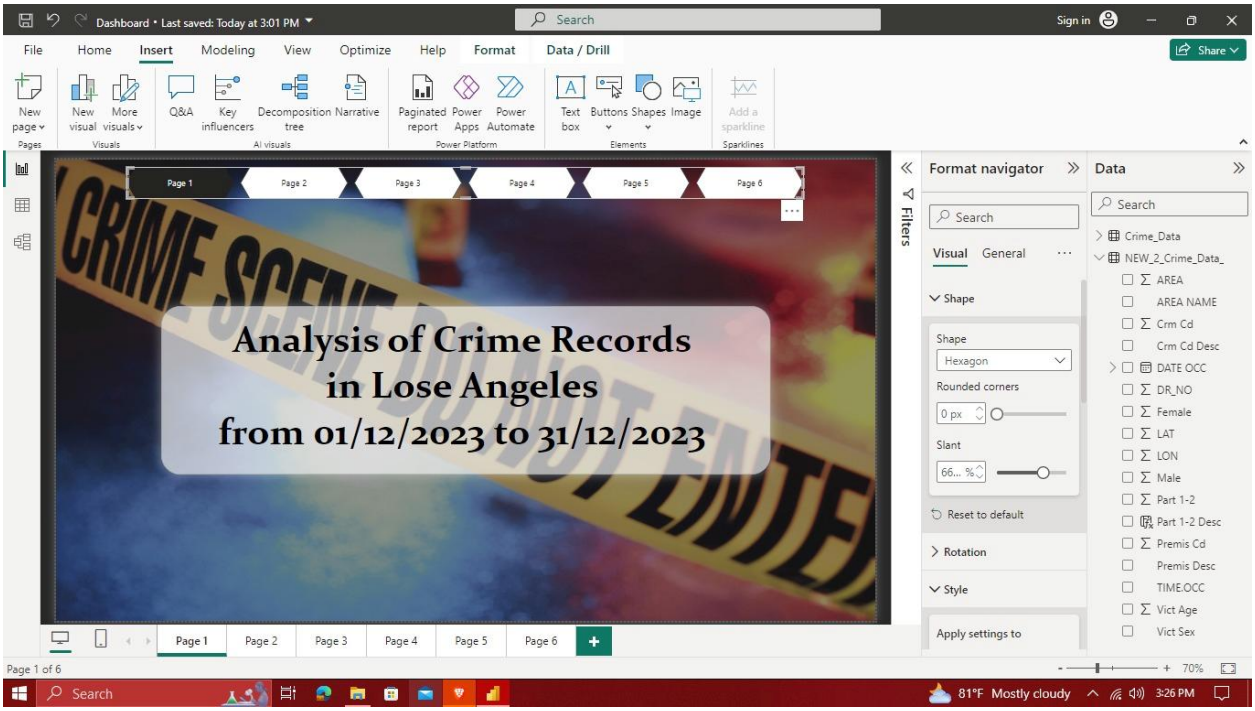



e) Edit the Dashboard.

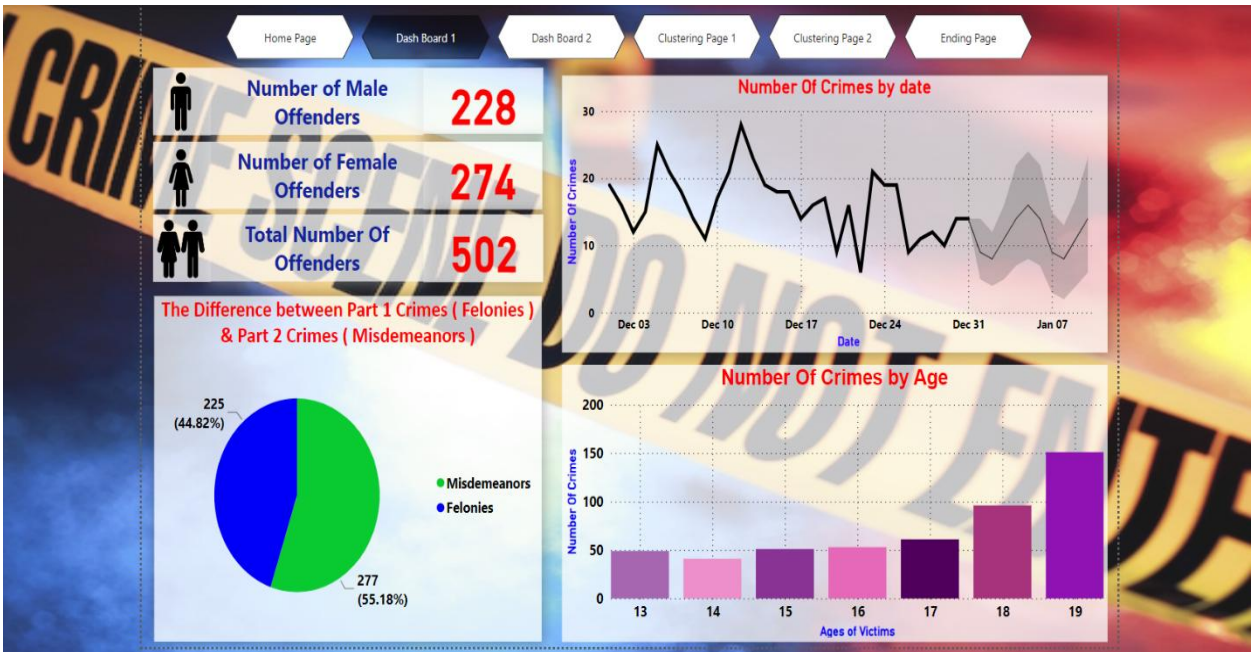
1. We edited the font sizes, font colors, backgrounds, images in the visualization parts.
2. As the next step, we arranged six pages in the whole Dashboard Document. The 1st page is for the cover page, 2nd & the 3rd pages are for a normal dashboard, 4th & 5th pages are for the clustering part and the last page is for the end.
3. Then, we created Buttons to move the pages.

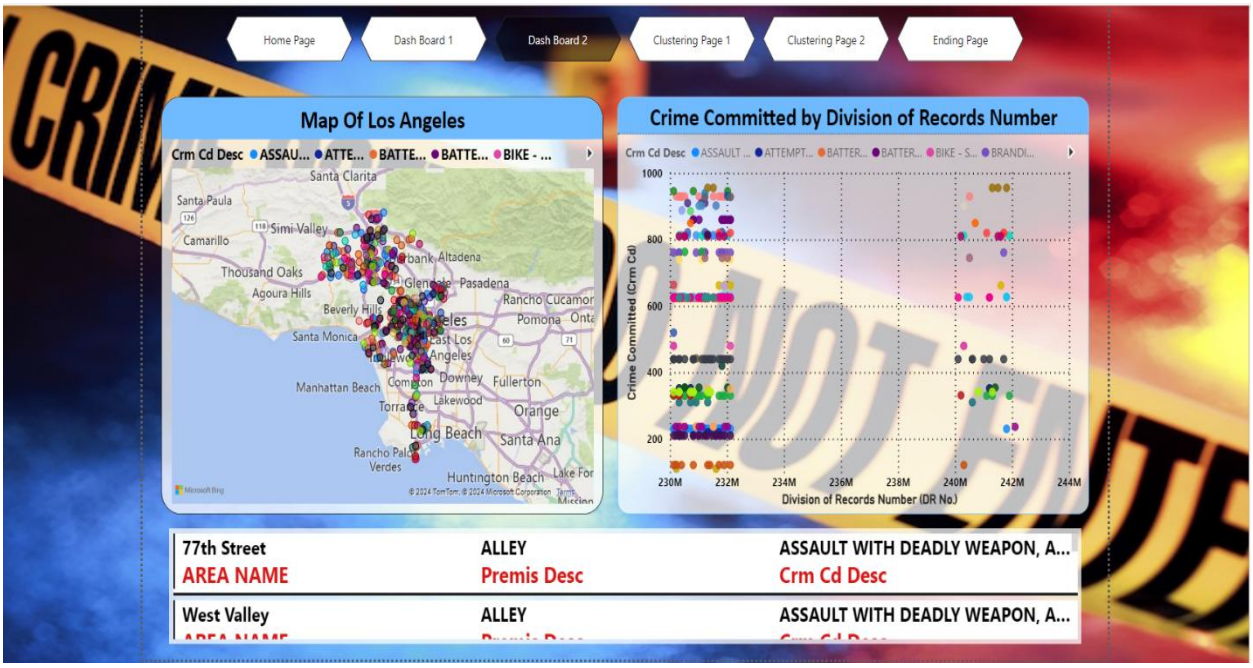


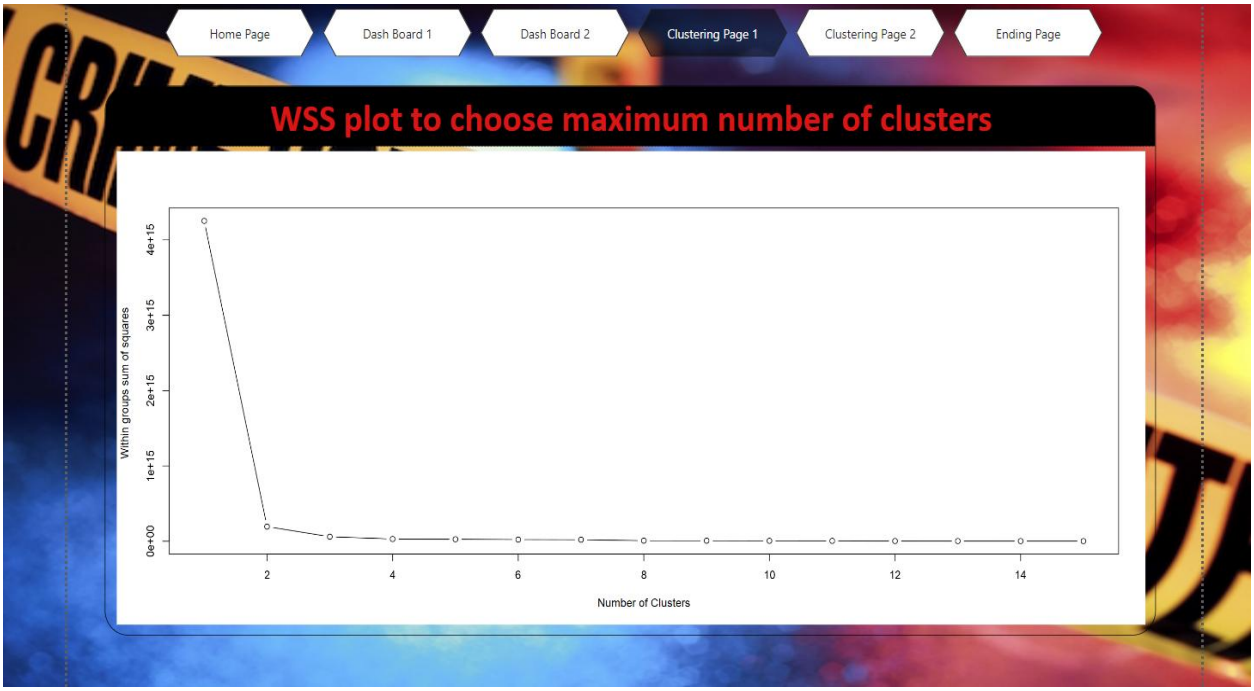

Dashboard • Last saved: Today at 3:01 PM
Search
Sign in
File
Home
Insert
Modeling
View
Optimize
Help
Format
Data / Drill
Share
Analysis of Crime Records
in Lose Angeles
from 01/12/2023 to 31/12/2023
Format navigator
Data
Visual
General
Properties
Title
Effects
Header icons
Alt text
Crime_Data
NEW_2_Crime_Data_
AREA
AREA NAME
Crm Cd
Crm Cd Desc
DATE OCC
DR_NO
Female
LAT
LON
Male
Part 1-2
Part 1-2 Desc
Premis Cd
Premis Desc
TIME OCC
Vict Age
Vict Sex
Page 1
Page 2
Page 3
Page 4
Page 5
Page 6
81°F Mostly cloudy
3:24 PM

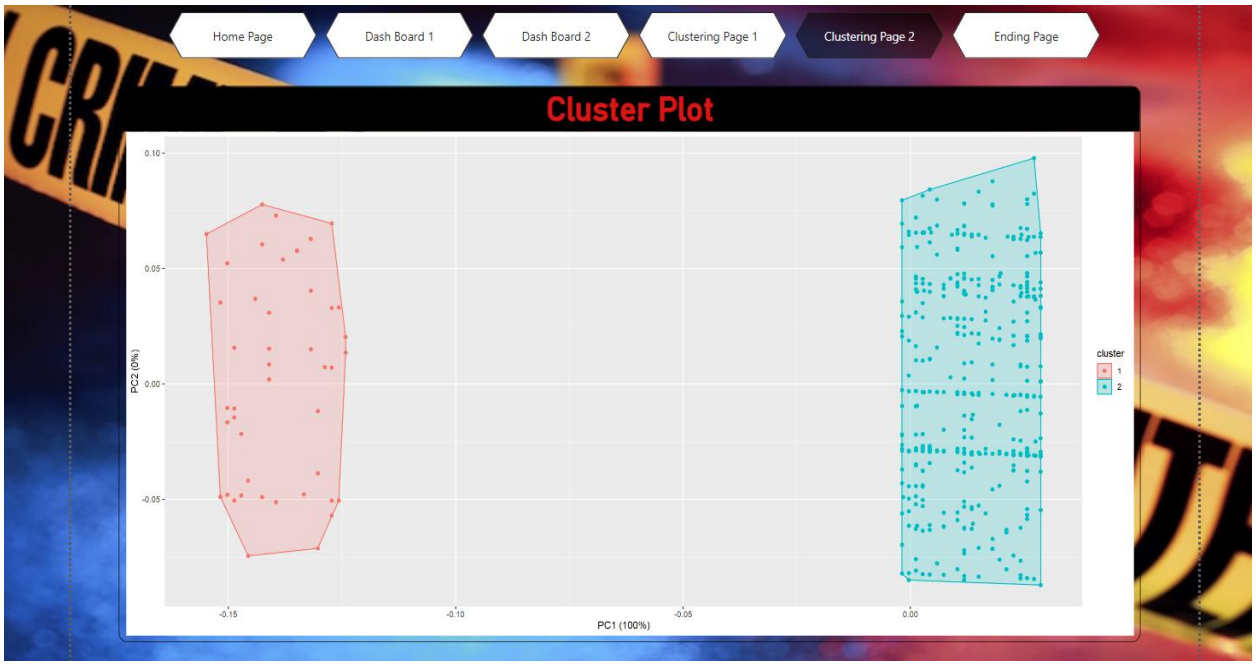

Dashboard • Last saved: Today at 3:01 PM
Search
Sign in
File
Home
Insert
Modeling
View
Optimize
Help
Format
Data / Drill
Share
Analysis of Crime Records
in Lose Angeles
from 01/12/2023 to 31/12/2023
Format navigator
Data
Visual
General
Shape
Shape
Hexagon
Rounded corners
0 px
Slant
66... %
Reset to default
Rotation
Style
Apply settings to
Crime_Data
NEW_2_Crime_Data_
AREA
AREA NAME
Crm Cd
Crm Cd Desc
DATE OCC
DR_NO
Female
LAT
LON
Male
Part 1-2
Part 1-2 Desc
Premis Cd
Premis Desc
TIME OCC
Vict Age
Vict Sex
Page 1
Page 2
Page 3
Page 4
Page 5
Page 6
81°F Mostly cloudy
3:26 PM

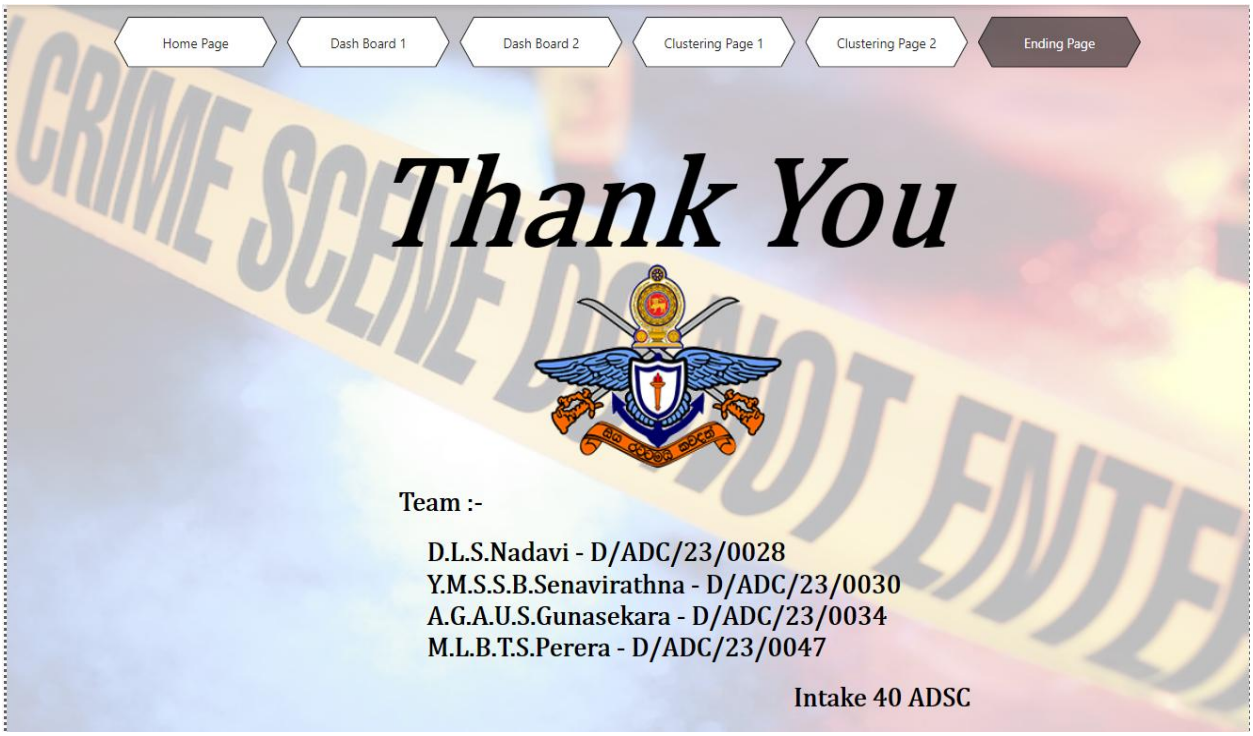
f) Finalized Dashboard.


Home Page
Dash Board 1
Dash Board 2
Clustering Page 1
Clustering Page 2
Ending Page
Map Of Los Angeles
Crm Cd Desc ASSAU... ATTE... BATTE... BATTE... BIKE - ...
Crime Committed by Division of Records Number
Crime Committed (Crm Cd)
Division of Records Number (DR No.)
77th Street
AREA NAME
ALLEY
Premis Desc
ASSAULT WITH DEADLY WEAPON, A...
Crm Cd Desc
West Valley
ALLEY
ASSAULT WITH DEADLY WEAPON, A...


Home Page
Dash Board 1
Dash Board 2
Clustering Page 1
Clustering Page 2
Ending Page
WSS plot to choose maximum number of clusters
Within groups sum of squares
Number of Clusters


Home Page
Dash Board 1
Dash Board 2
Clustering Page 1
Clustering Page 2
Ending Page
Cluster Plot
PC2 (0%)
PC1 (100%)
cluster
1
2


Home Page
Dash Board 1
Dash Board 2
Clustering Page 1
Clustering Page 2
Ending Page
Thank You
Team :-
D.L.S.Nadavi - D/ADC/23/0028
Y.M.S.S.B.Senavirathna - D/ADC/23/0030
A.G.A.U.S.Gunasekara - D/ADC/23/0034
M.L.B.T.S.Perera - D/ADC/23/0047
Intake 40 ADSC

## 4) Conclusion

A dashboard provide a comprehensive and easily digestible overview of data, fostering better decision-making, collaboration, and performance monitoring to the viewers. The above dashboard summarizes the data in the crime dataset in the form of graphs and plots. Therefore, by looking at the above dashboard viewers can get an idea about the crimes done by teenagers in the city of Los Angeles. According to the pie chart, information suggests that there is a lower number of serious crimes (felonies) being committed compared to less serious crimes (misdemeanors). The Line Chart gives a forecast from December 31, 2023, to January 10, 2024.

Overall, the analysis of the Column Bar Chart underscores the importance of age as a significant factor influencing criminal behavior among youth. The insights gleaned from this data could inform targeted interventions and policies aimed at reducing crime rates among adolescents and young adults. The Map provides a visual representation of crime concentration in different areas of Los Angeles, which can help law enforcement agencies and policymakers allocate resources and implement strategies to address crime issues in the region. The Scatter chart provides a visual representation of crime data by division of records number, which can help law enforcement agencies and policymakers identify areas with high crime rates, allocate resources, and implement strategies accordingly. According to this, we can see that the delay in reporting crimes can be a reason for criminals being on the loose and increasing crime incidents.

## 5) References

https://youtu.be/wVp_vq0cbIo?si=f2izda6AwfLpZKd-