GE 2324: The Art & Science of Data

Analysis of Crime Rates and Statistics in the USA

Project Report

Group Members

- Avi Malhotra (55773896)
- Chiu Pit Lai Billy (55226317)
- Lai Hok Him (55713248)
- Pratul Rajagopalan (55858290)
- Lav Asrani (56444472)
- Avinesha Kumar (56327741)
- Sukrit Kumar (56557780)

**Note: All group members contributed equally.*

Table of Contents

**Note: All figures referred in the report are attached in the Appendix. The main report is from pages 2-8.*

1. Introduction

For the group project of this course, we decided to conduct an inquiry into the crime rates in the demographic of the United States of America. Our purpose for choosing this topic stemmed from our intellectual curiosity given the abnormally high crime rates in the USA and their unique Second Amendment rights (The right to bear arms). Naturally, this demographic was interesting to examine to study and analyze any prevalent patterns using the conceptual knowledge gained throughout this course. The abundance of raw data available on a range of websites was beneficial as it helped us inquire into factors such as cities, crime types, and duration as variables from a variety of perspectives.

2. Raw Data

The data that our group has selected pertains to the state of Connecticut (a state in the North-Eastern United States of America) and other states that are in proximity to it - New York, New Jersey, Pennsylvania, Rhode Island and Massachusetts. The data is divided into groups based on the type of crime committed: Larceny-theft, burglary, motor vehicular crime, Murder, Violent crime and much more.

The data provided compares the crime rates in Connecticut with that of the neighbouring north-eastern states and the United States as a whole. This dataset contains data based on the years 1960-2007. Some of the more basic patterns that emerged were that Connecticut usually had a lower crime rate for each type of crime than the United States as a whole, and that the crime rate usually peaked in the period between 1975 - 1990.

3. Data Analysis

3.1 General trends in major cities

We sorted the data and have come up with this interesting line chart. The line chart in *Figure 1* shows the crime rate of 9 major cities of America over a period of 47 years.

In the year 1960, New York has the highest crime rate while New Hampshire has the lowest. Maine, Connecticut, New Jersey, and Massachusetts follow the same trend over the period. There is a significant decline in the crime rates of all the states from the year 1990 to 2007. Pennsylvania is the most stable when compared to all the other states. The crime rate of New York declined the most at the end (year 2007) and has become the second-lowest, whereas Maine has the most crime rates.

3.2 Impact of the 2008 stock market crash

To study the impact of the 2008 stock market crash, we have come up with two heat maps *(Figures 2 and 3):* for the year 2007 (a year before the crash) and for the year 2009 (a year after the crash). We gathered the data which had all the types of crime happening in different states in that particular year. In order to calculate the crime rates (per 100,000 people), we added all the crimes, divided them by population, and multiplied it by 100,000.

We discovered an interesting correlation when comparing crime rates across years. While crime rates in large cities declined until 2007, they abruptly rose after that (in 2009, for instance). This can be ascribed to the 2008 sock market crisis, which could be one of the factors contributing to the rapid rise in crime.

3.3 Investigating different crime types across cities

We partitioned the data into 4 clusters on Tanagra and we got the above clusters, as seen in *Figures 4 to 7*. Cluster 1 had only 3 states – South and North Dakota and New Hampshire (crime rate of 1821.5 - 2032.0). The next cluster contained states from New York to Wisconsin and included Connecticut. These states had crime rates ranging from 2392.7 - 3128.6. The next cluster included states from Minnesota to Kansas with crime rates from 3325.2 - 4131.3. The fourth cluster had states from Missouri to South Carolina where the crime rates were between the values of 4243.3 - 5060.0.

From the dendrogram, the cities were split into 3 major clusters, with cluster 1 having 19 cities, cluster 2 with 15 cities and cluster 3 with 16 cities.

Similarly, for the property crime statistics depicted in *Figures 8 to 10*, we divided the states into four clusters. Cluster 1 contained states from South Dakota to West Virgina which had

property crime rates (per 100,000 inhabitants) between 1652.3 - 2525.0. Cluster 2 included the states from Iowa to Mississippi with rates between 2615.6 - 3200.8. Cluster 3 had the states of Delaware to Nevada with crime rates between 3370.1 - 3777.8. The final cluster had states from Georgia to Arizona with crime rates between 3901.0 - 4414.0.

The dendrogram had 3 clusters with 16, 12 and 22 items in the clusters originally.

With the all the data, we tried to understand what the factors were in the similarity of crime rates among certain states. The states like North Dakota, Vermont and New Hampshire consistently had the lowest crime rates for any type of crime and almost always ranked as states in the top 5 for the lowest crime rate, while States like Arizona and South Carolina have very high crime rates and consistently rank towards the top in terms of crime rate in any category.

Experts state that the crime rate depends on population density, economic conditions and how urban or rural an area is. This could be why the northeastern states have such low crime rates, because these states have a relatively high average income with low levels of unemployment – leading to better socioeconomic conditions for the occupants of these regions. These states are also sparsely populated with only 17.2% of the US population. This low crime rate can be verified by looking at the cluster map of the United States (fig 7) - most states in the northeast are either in cluster 3 or 4 - the lower crime rate states.

Most of the states in Cluster 1 and 2 (the highest crime rate clusters) are in the South (fig 7) which contain a very large percentage of the United States population at 38%. These states have exceptionally high rates of people in low socioeconomic conditions, with a low income and a lower literacy rate than the national average.

3.4 Association Between Different Crimes in US

Finding association involves application of Apriori algorithm.

Initially, we had a dataset which contained different crime rates of United States of different years. Unfortunately, Apriori algorithm was not applicable to it as it is only applicable to the table of true and false. In order to achieve it, we took out the average of every crime rate. It gave us the average rate of that particular crime in a year as shown in *Figure 11*.

With the help of this average, we came out with a table which told us the years in which the rate of that crime was greater than the average. This table is depicted in *Figure 12* of the appendix.

Now, with the help of the table, we found association rules using Weka.

- Aggravated assault rate=TRUE 29 ==> Forcible rape rate=TRUE 29 <conf:(1)>
- Property crime rate=TRUE 25 ==> Larceny-theft rate=TRUE 25 <conf:(1)>
- Burglary rate=TRUE 24 ==> Murder and nonnegligent manslaughter rate=TRUE 24 <conf:(1)>
- Robbery rate=TRUE Larceny-theft rate=TRUE 24 ==> Property crime rate=TRUE 24 <conf:(1)>
- Robbery rate=TRUE Property crime rate=TRUE 24 ==> Larceny-theft rate=TRUE 24 <conf:(1)>
- Murder and nonnegligent manslaughter rate=TRUE 26 ==> Robbery rate=TRUE 25 <conf:(0.96)>
- Property crime rate=TRUE 25 ==> Robbery rate=TRUE 24 <conf:(0.96)>
- Property crime rate=TRUE Larceny-theft rate=TRUE 25 ==> Robbery rate=TRUE 24 <conf:(0.96)>
- Property crime rate=TRUE 25 ==> Robbery rate=TRUE Larceny-theft rate=TRUE 24 <conf:(0.96)>
- Forcible rape rate=TRUE 31 ==> Aggravated assault rate=TRUE 29 conf:(0.94)

From these Rules, we can get to know many associations. One of them can be, if in any particular year, the Aggravated Assault rate was high than in that year Forcible Rape Rate was also high.

Moreover, one of the interesting associations (interest = |confidence - expectations|) is that, if in any year, the Burglary rate was high than Murder and nonnegligent manslaughter rate was also high.

3.5 Network analysis in New York Police Precincts

The network graph in *Figures 13 and 14* represents the map for the New York City. There are 77 nodes and 174 edges in the network. The nodes represent the police precincts allocated in New York city and the edges represent the adjacent connection to the precinct. Then we can observe the betweenness and closeness of the graph.

In the graph illustrated in *Figure 15*, it shows the betweenness of each precinct by the size of node. With larger size of node, those precincts will have more betweenness. The result tells us that the precinct 104 has a large betweenness and precinct 71 has a small betweenness. We assume that larger betweenness will have more crime.

Figure 16 shows the closeness of each precinct. Larger size of node will have larger closeness. Not surprisingly, most of the centered nodes have large closeness and the border precincts have smaller closeness.

Based on *Figure 17* and after observing the betweenness and the closeness of the precincts of New York city, and comparing to the crime report, it seems that betweenness and the closeness are not significantly affecting the crime rate of New York City. Therefore, betweenness and closeness are not the factor of high crime rate, but the network graphs is still useful to arrange the human resource in each police precinct for efficiency.

In addition, to allocate the few centers of the precincts, we can also use the k-mean clustering method, as shown in *Figures 18 and 19*. By setting k=3, we can obtain the result of 3 centroid which would be the center of those precincts to handle the crime cases of the city efficiently.

3.6 Investigating patterns in New York and New Hampshire

As the crime rate dataset we are working with has a collection of ratio variables, we start by calculating the Pearson correlation value between the two variables, population (X) and the total number of reported crimes (Y). The value of R is -0.8397, as calculated in *Figures 20 and 21*.

This is a strong negative correlation, which means that high X (population) variable scores go with low Y (crimes reported) variable scores (and vice versa), as shown in *Figure 22*.

The strongly negative correlation value implies that with the increase of population from 18 million in 1965 to 19.2 million in 2007, i.e., over 42 years, the state of New York has seen a drastic decrease in the number of reported crimes, which sets it aside from the usual linear trend.

However, it would be unfair to say that the correlation value is a perfect indicator of an increase or decrease in the number of reported crimes in a particular state or nation. In the dataset, we can observe irregularities that contradict the calculated correlation value. For example, New York saw a massive increase in crime during the mid-70s to early-80s and then late-80s to mid-90s, with more than a hundred thousand crimes being reported every year during these periods. However, these irregularities can be backed up with factual data. The homicide rate in New York City more than doubled during the 70s, with one of the main reasons being a reduction in the city's police force due to a financial crisis and an increase in the number of street gangs and mafia organizations. Similarly, the crack epidemic surge or the increase in exposure to cocaine in New York is believed to be the primary reason for the crime surge in the 80s.

Nevertheless, things have started to turn better in the state of New York. The attacks of 9/11 triggered a fight against crime and terrorism, with a massive increase in the annual security budget and a more significant number of cops deployed on the streets. New York City has one of the lowest crime rates of major cities in the United States (2019).

Moving onto *Figures 23 to 25*, we start by calculating the Pearson correlation value of the two ratio variables, population (X) and the total number of reported crimes (Y). The value of R is 0.676.

This is a moderate positive correlation, which means there is a tendency for high X (population) variable scores go with high Y (reported crimes) variable scores (and vice versa).

The moderately positive correlation value implies that with the increase of population from 600 thousand in 1960 to 1.3 million in 2007, i.e., over 47 years, the state of New Jersey has seen a high increase in crime rates.

The general crime trend and the correlation value of New Jersey differ significantly from New York. Unlike New York, the state of New Jersey has seen an increase in reported crimes with the increase in the state population, which is quite reasonable if we consider the linear correlation. However, the crime statistics of New Jersey is proportioned differently than that of New York. New Jersey experiences a higher number of reported thefts and burglaries and surprisingly lower numbers of aggressive assaults and murders than New York.

4. Conclusion

From our investigation, we were able to:

- examine general trends in major cities more carefully,
- compare the varying crime rates across cities,
- consider the impact of nation-wide events on crime rates,
- inquire into intra police-precinct relations,
- explain the association and correlations between crime types.

Nonetheless, this project is by no means exhaustive. While our research is thorough and well supported by our evidence and analysis, there are still certain questions that require further inquiry, and we list them as possible areas of exploration in the future.

These questions are:

1. Are some inter-city precincts more important than others?
2. Are some age groups/genders targeted more often for crimes?
3. What is the transportation monetary expenditure incurred by precincts?

Finally, the group project of this course has truly been an intellectually stimulating experience. It gave us all an opportunity to explore the applications of the knowledge gained throughout this course in previously unforeseen circumstances.

5. References

*All raw data has been collected from the The US government's open data website.

*All analysis has been done using the concepts and software taught in class by the course leader Dr. Helena Wong and tutorial instructor Dr. Kenneth Lee.

6. Appendix

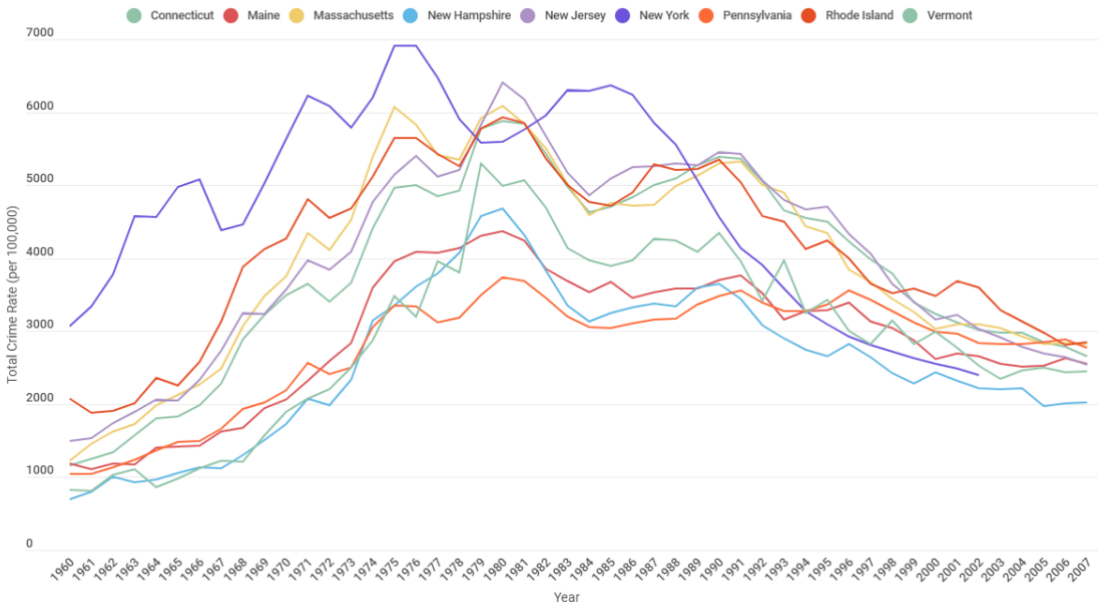
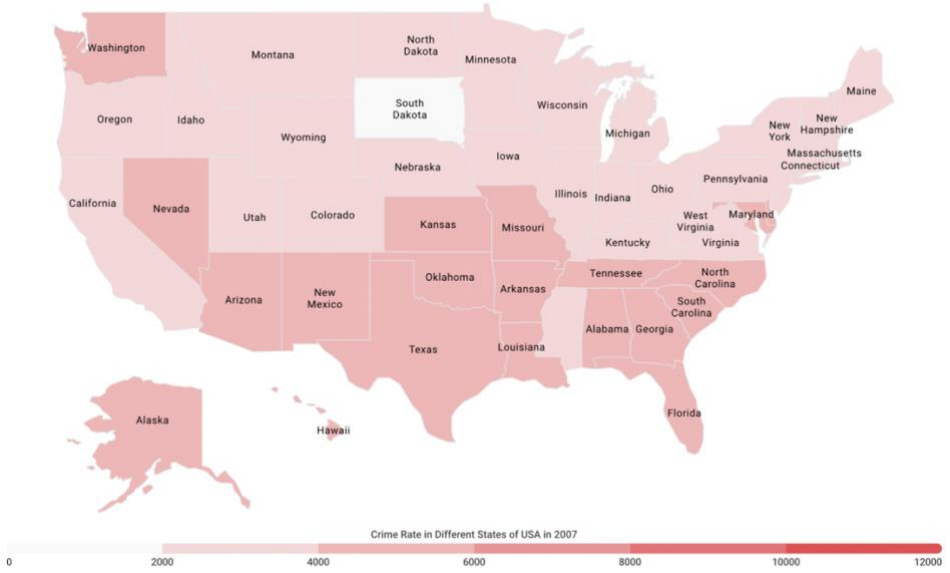
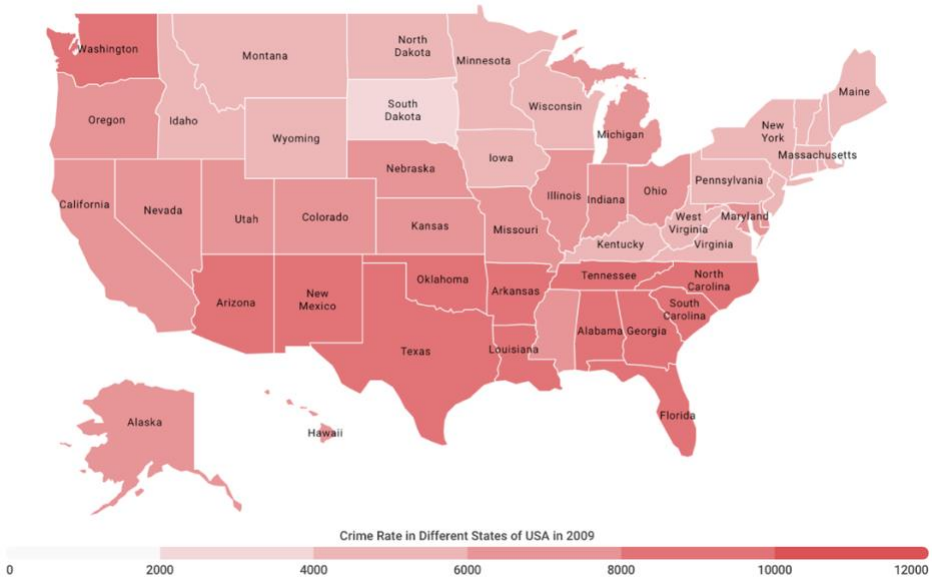
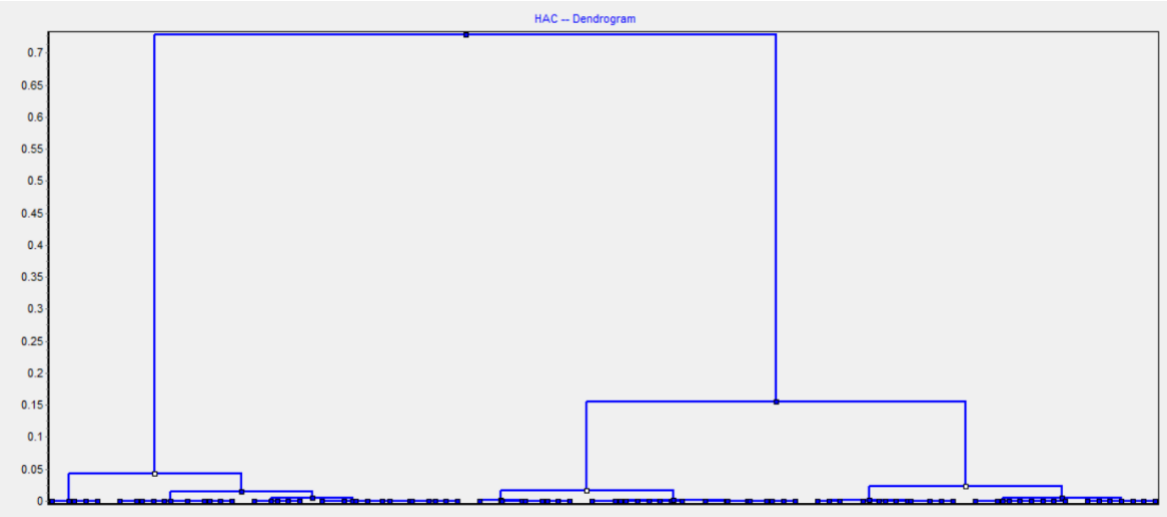


Figure-1



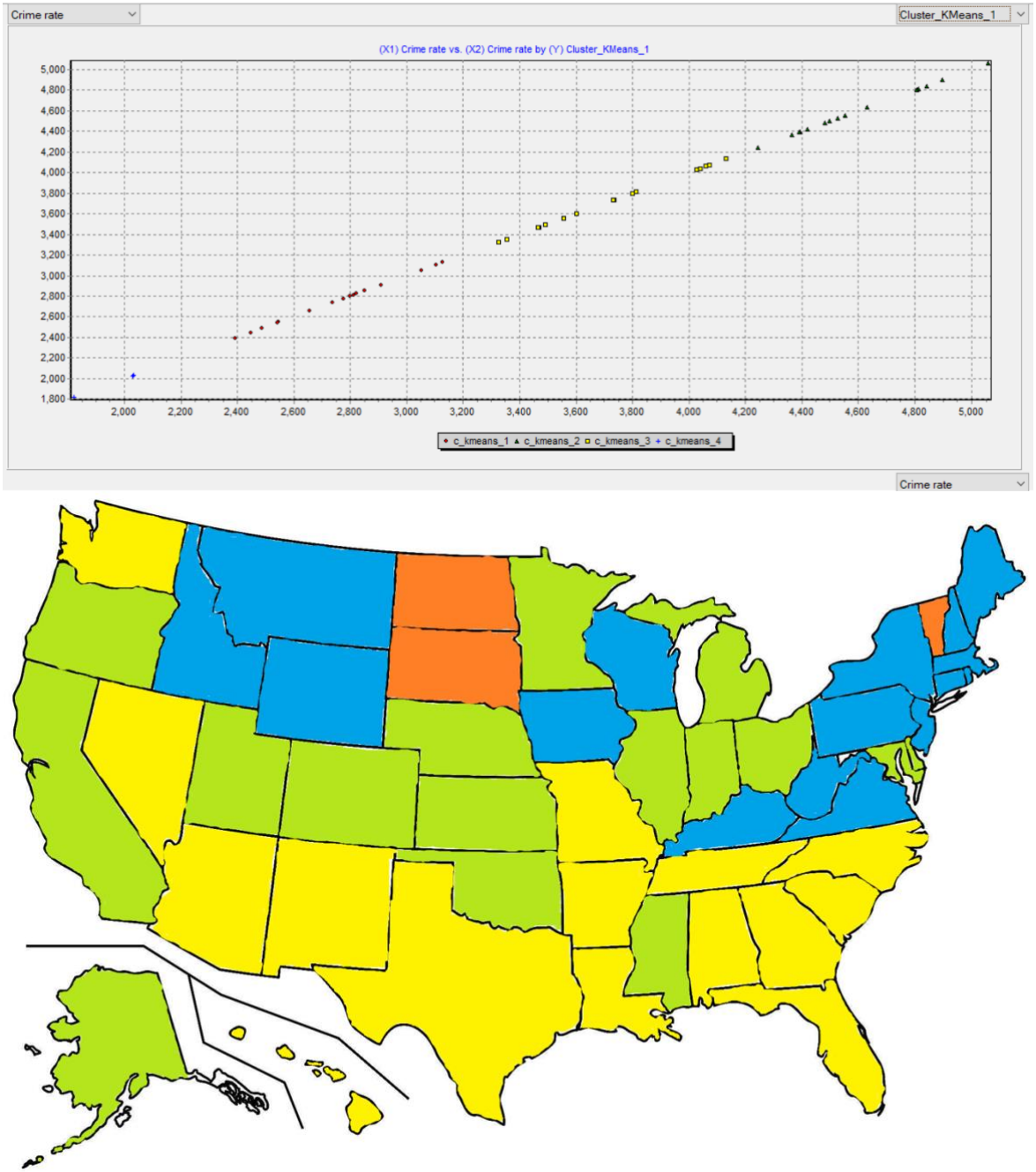


Figures 2 and 3

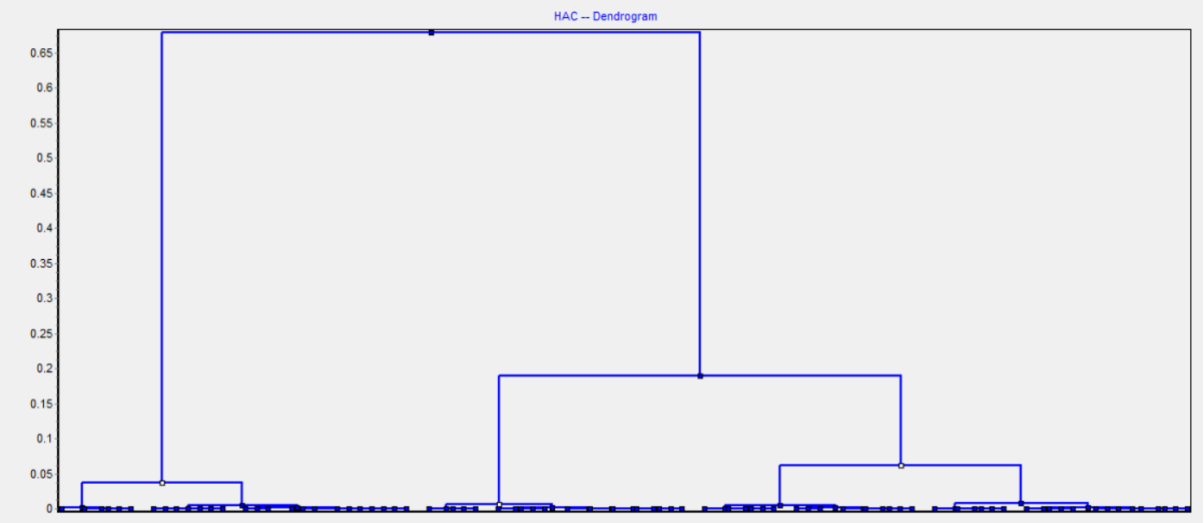


Clustering results

Clusters	From the dendrogram	After one-pass relocation
cluster n°1	19	19
cluster n°2	15	15
cluster n°3	16	16

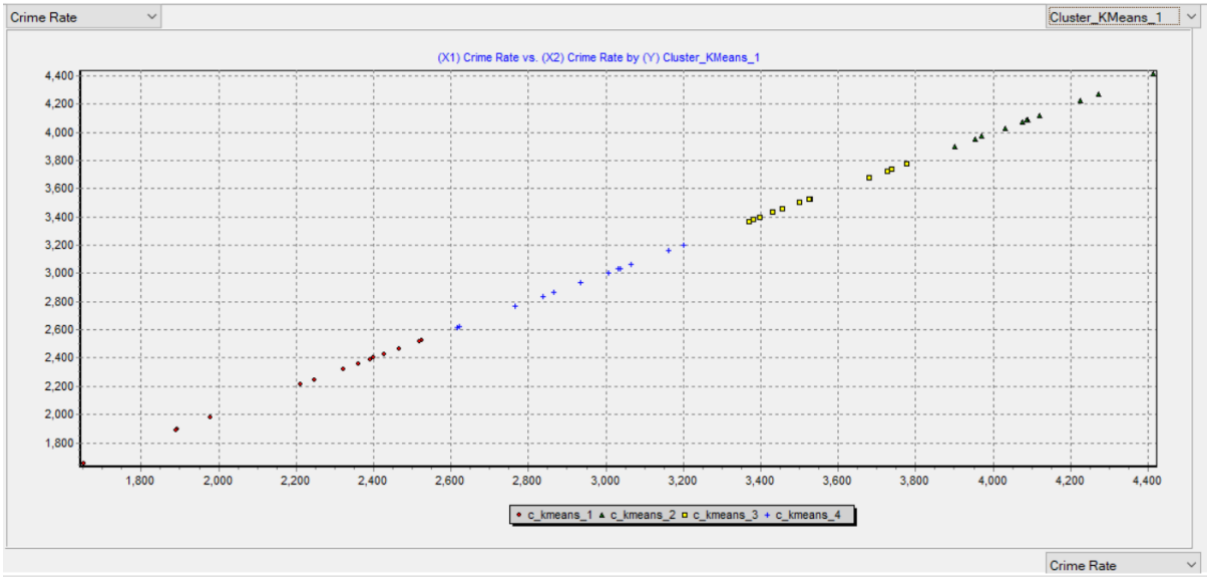


Figures 4 to 7



Clustering results

Clusters	From the dendrogram	After one-pass relocation
cluster n°1	16	17
cluster n°2	12	15
cluster n°3	22	18

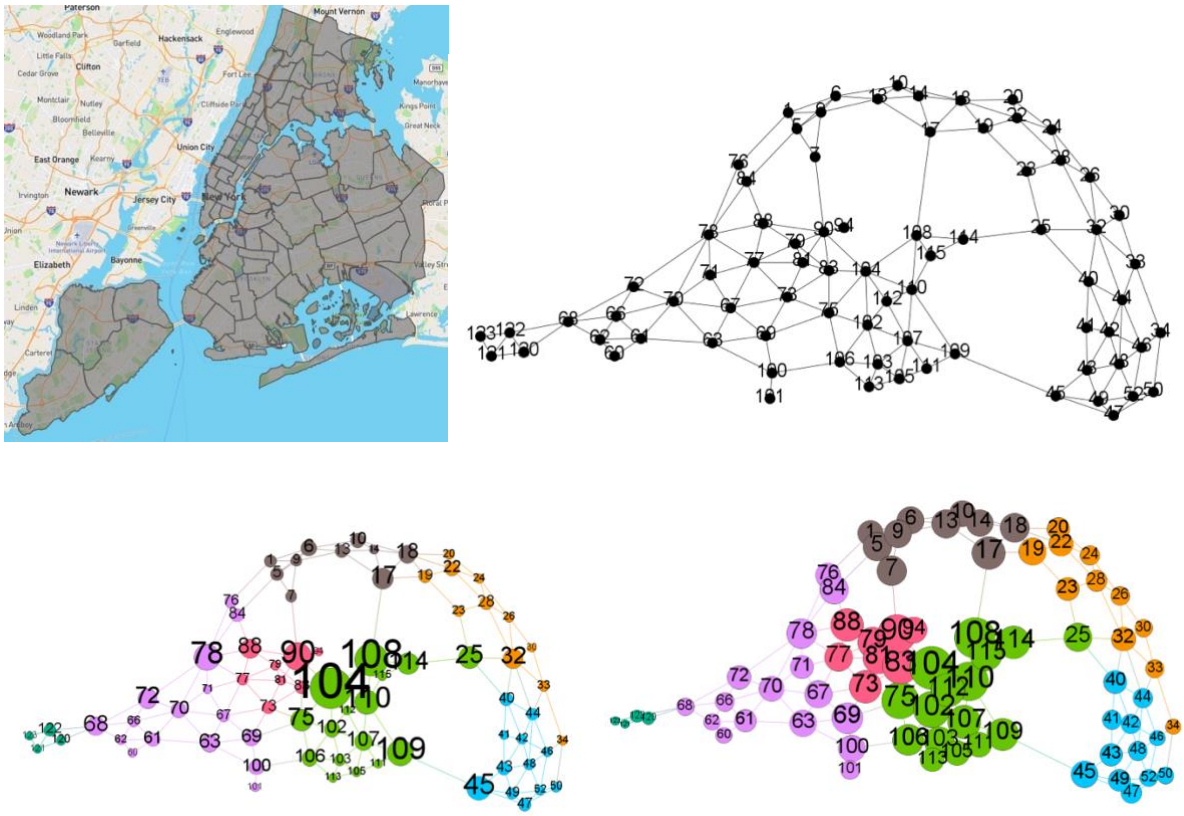


Figures 8 to 10

Year	State	Murder and nonnegligent manslaughter rate	Forcible rape rate	Robbery rate	Aggravated assault rate	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle thef rate
1999	United States	5.7	32.8	150.1	334.3	3743.6	770.4	2550.7	422.5
2000	United States	5.5	32.0	145.0	324.0	3618.3	728.8	2477.3	412.2
2001	United States	5.6	31.8	148.5	318.5	3656.1	740.8	2484.6	430.6
2002	United States	5.6	33.1	146.1	309.5	3630.6	747.0	2450.7	432.9
2003	United States	5.7	32.3	142.5	295.4	3591.2	741.0	2416.5	433.7
2004	United States	5.5	32.4	136.7	288.6	3514.1	730.3	2362.3	421.5
2005	United States	5.6	31.8	140.8	290.8	3431.5	726.9	2287.8	416.8
2006	United States	5.7	31.0	149.4	287.5	3334.5	729.4	2206.8	398.4
2007	United States	5.6	30.0	147.6	283.8	3263.5	722.5	2177.8	363.3
	AVERAGE	7.5	28.9	175.8	270.4	3946.3	1050.3	2452.9	443.1

Year	Country	Murder and nonnegligent manslaughter rate	Forcible rape rate	Robbery rate	Aggravated assault rate	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
1999	United States	false	true	false	true	false	false	true	false
2000	United States	false	true	false	true	false	false	true	false
2001	United States	false	true	false	true	false	false	true	false
2002	United States	false	true	false	true	false	false	false	false
2003	United States	false	true	false	true	false	false	false	false
2004	United States	false	true	false	true	false	false	false	false
2005	United States	false	true	false	true	false	false	false	false
2006	United States	false	true	false	true	false	false	false	false
2007	United States	false	true	false	true	false	false	false	false

Figures 11 and 12



Figures 13 to 16

Police Department
City of New York

Bill de Blasio
Mayor

Dermot Shea
Police Commissioner

Volume 28 Number 28 — *CompStat* — 104th Precinct

Report Covering the Week
7/12/2021 Through 7/18/2021

Crime Complaints

	Week to Date 2021	2020	% Chg	28 Day 2021	2020	% Chg	Year to Date* 2021	2020	% Chg	2 Year % Chg	11 Year % Chg	28 Year % Chg
Murder	0	0	***.*	0	0	***.*	1	0	***.*	-50.0	0.0	-90.0
Rape	1	1	0.0	1	2	-50.0	10	9	11.1	-33.3	11.1	-41.2
Robbery	5	6	-16.7	12	16	-25.0	63	99	-36.4	-29.2	-53.3	-87.8
Fel. Assault	5	8	-37.5	22	29	-24.1	130	167	-22.2	36.8	3.2	-19.8
Burglary	5	5	0.0	16	20	-20.0	127	164	-22.6	22.1	-49.0	-88.6
Gr. Larceny	12	8	50.0	38	34	11.8	252	309	-18.4	-26.3	12.0	-24.8
G.L.A.	2	6	-66.7	13	19	-31.6	111	73	52.1	30.6	-44.5	-94.3
TOTAL	*30*	*34*	*-11.76*	*102*	*120*	*-15.00*	*694*	*821*	*-15.47*	*-5.19*	*-26.56*	*-83.16*

Volume 28 Number 28 — *CompStat* — 77th Precinct

Report Covering the Week
7/12/2021 Through 7/18/2021

Crime Complaints

	Week to Date 2021	2020	% Chg	28 Day 2021	2020	% Chg	Year to Date* 2021	2020	% Chg	2 Year % Chg	11 Year % Chg	28 Year % Chg
Murder	0	2	-100.0	0	4	-100.0	4	10	-60.0	-33.3	-60.0	-90.2
Rape	1	0	***.*	2	0	***.*	13	11	18.2	30.0	30.0	-68.3
Robbery	5	3	66.7	14	13	7.7	82	89	-7.9	-19.6	-27.4	-93.0
Fel. Assault	7	11	-36.4	28	36	-22.2	160	136	17.6	20.3	-5.3	-76.4
Burglary	1	6	-83.3	10	13	-23.1	59	80	-26.2	-4.8	-58.5	-93.6
Gr. Larceny	7	6	16.7	17	31	-45.2	131	157	-16.6	-31.8	-5.8	-56.2
G.L.A.	2	3	-33.3	11	10	10.0	49	41	19.5	53.1	-34.7	-88.2
TOTAL	*23*	*31*	*-25.81*	*82*	*107*	*-23.36*	*498*	*524*	*-4.96*	*-7.26*	*-24.32*	*-86.02*

Volume 28 Number 28 — *CompStat* — 34th Precinct

Report Covering the Week
7/12/2021 Through 7/18/2021

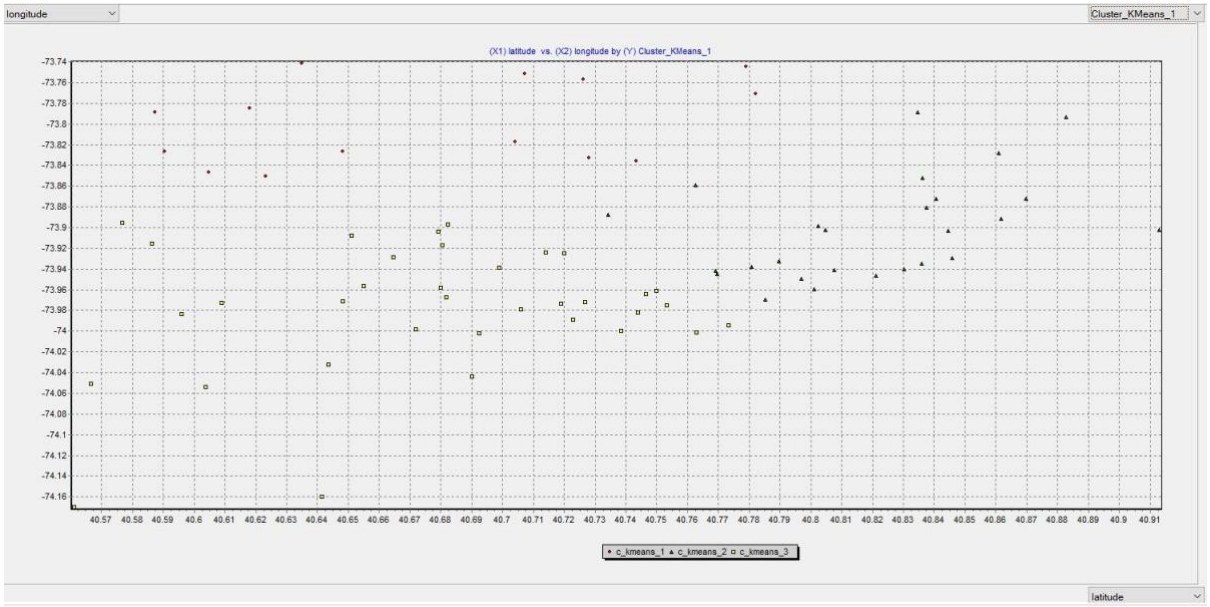
Crime Complaints

	Week to Date 2021	2020	% Chg	28 Day 2021	2020	% Chg	Year to Date* 2021	2020	% Chg	2 Year % Chg	11 Year % Chg	28 Year % Chg
Murder	0	0	***.*	1	0	***.*	7	2	250.0	133.3	250.0	-81.6
Rape	0	1	-100.0	0	3	-100.0	8	9	-11.1	-38.5	-27.3	-82.2
Robbery	5	6	-16.7	20	16	25.0	108	88	22.7	3.8	-26.5	-86.7
Fel. Assault	8	3	166.7	20	27	-25.9	148	137	8.0	-9.2	27.6	-72.0
Burglary	1	6	-83.3	7	12	-41.7	58	104	-44.2	-9.4	-38.3	-94.9
Gr. Larceny	6	7	-14.3	47	29	62.1	309	210	47.1	5.1	72.6	-53.5
G.L.A.	4	1	300.0	9	6	50.0	97	54	79.6	234.5	131.0	-91.4
TOTAL	*24*	*24*	*0.00*	*104*	*93*	*11.83*	*735*	*604*	*21.69*	*9.70*	*24.37*	*-83.17*

Figure 17

Cluster centroids

Attribute	Cluster n?	Cluster n?	Cluster n?
longitude	-73.798134	-73.903217	-73.978882
latitude	40.676956	40.820956	40.678958



Figures 18 and 19

Result Details & Calculation

X Values
$\sum$ = 784237798
Mean = 18238088.326
$\sum(X - M_x)^2 = SS_x$ = 11051445134771.4

Y Values
$\sum$ = 37138399
Mean = 863683.698
$\sum(Y - M_y)^2 = SS_y$ = 2358483160043.07

X and Y Combined
N = 43
$\sum(X - M_x)(Y - M_y)$ = -4286768203560.77

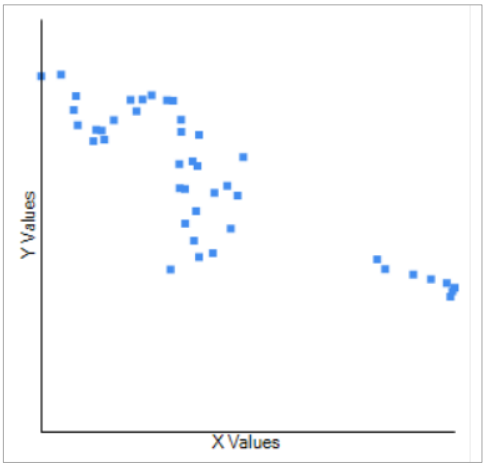
R Calculation
$r = \sum((X - M_y)(Y - M_x)) / \sqrt{((SS_x)(SS_y))}$

r = -4286768203560.77 /
√((11051445134771.4)
(2358483160043.07)) = -0.8397

Meta Numerics (cross-check)
r = -0.8397

Key

X: X Values
Y: Y Values
M_x: Mean of X Values
M_y: Mean of Y Values
$X - M_x$ & $Y - M_y$: Deviation scores
$(X - M_x)^2$ & $(Y - M_y)^2$: Deviation Squared
$(X - M_x)(Y - M_y)$: Product of Deviation Scores



Figures 20 to 22

Result Details & Calculation

X Values
$\sum$ = 46490756
Mean = 968557.417
$\sum(X - M_x)^2 = SS_x$ = 2425800057799.67

Y Values
$\sum$ = 1235163
Mean = 25732.562
$\sum(Y - M_y)^2 = SS_y$ = 6257252481.812

X and Y Combined
N = 48
$\sum(X - M_x)(Y - M_y)$ = 83287626919.75

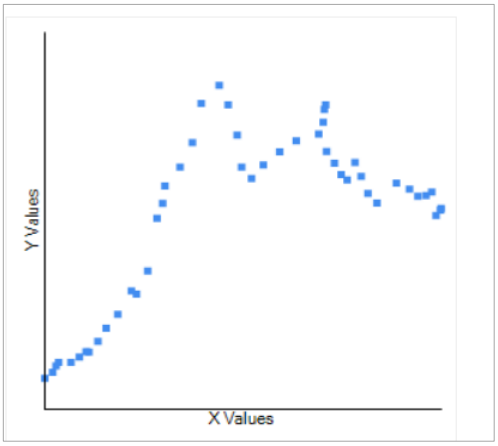
R Calculation
$r = \sum((X - M_y)(Y - M_x)) / \sqrt{((SS_x)(SS_y))}$

$r = 83287626919.75 / \sqrt{((2425800057799.67)(6257252481.812))} = 0.676$

Meta Numerics (cross-check)
r = 0.676

Key

X: X Values
Y: Y Values
M_x: Mean of X Values
M_y: Mean of Y Values
$X - M_x$ & $Y - M_y$: Deviation scores
$(X - M_x)^2$ & $(Y - M_y)^2$: Deviation Squared
$(X - M_x)(Y - M_y)$: Product of Deviation Scores



Figures 23 to 25