Analyzing trends in crime - 1975 to 2014

Angelina, Asha, Nikolai

Initial question:

How have trends in crime changed over the past 40 years in major U.S. cities?

Question for modeling:

How have clusters considering homicide per 100k and cities changed across the three years- 1975, 1995, and 2014 (looking at the initial, middle, and end year)?

Data:

Crime data of Major Cities in the United States over a span 40 year span from 1975 to 2014

Variables used for project:

- department name (string)
- Homicides per 100k (integer)

Source:

https://github.com/themarshallproject/city-crime

Tools used:

1) Google Colab was used to:

  2) Clean data

  3) Explore data (EDA)

  4) Model Data (Knn model)

2) Tableau

  1) Visualizations (Bubble Maps and Geo-Bubble Maps)

3) Jupyter Notebook and Pycharm

  1) Visualizations (Geo-Heatmap)

EDA

Process:

For the initial EDA, we created bivariate line graphs of each variable throughout the years. We did this because we are curious as to how crime changed throughout time. By looking at this data, it can help us gain a clearer picture of if crime was actually decreasing, the periods in which it did decrease and possible allow us to understand why this might have happened (political acts that may have affected crime, etc.) However, we also understand that the year itself was most likely not why crime changed, but nonetheless think the information this comparison will provide will be beneficial.

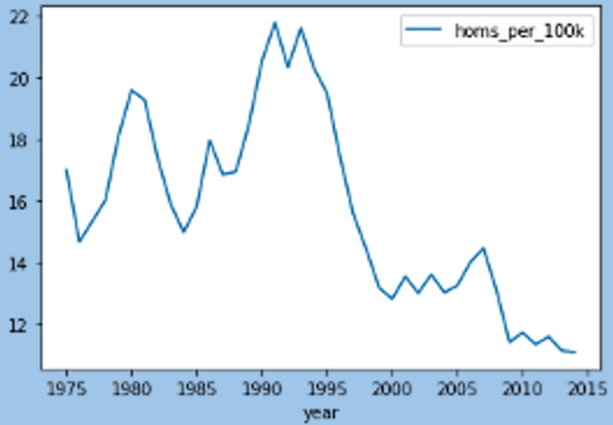
Our primary focus was on homes per 100k and Total homes. However, we analyzed all variables to gain a greater understanding of our variables and data and to ensure others could understand it as well.

Data Used:

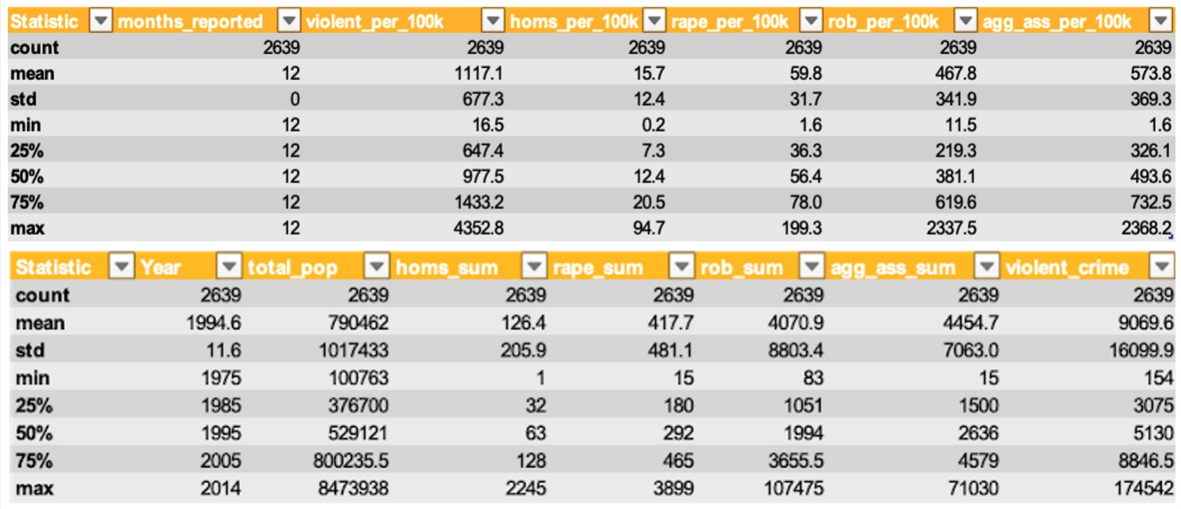
Total homicide: The range from the 25th to the 75th percentile is 32 to 128. However, there are many outliers far above this range (going past 2000).

homs_per_100k decreased over the years with a few upward spikes in a few years (based on graph and correlation coefficient = -0.1615689232252936).

EDA - Examples



Bivariate with year and homicides per 100k

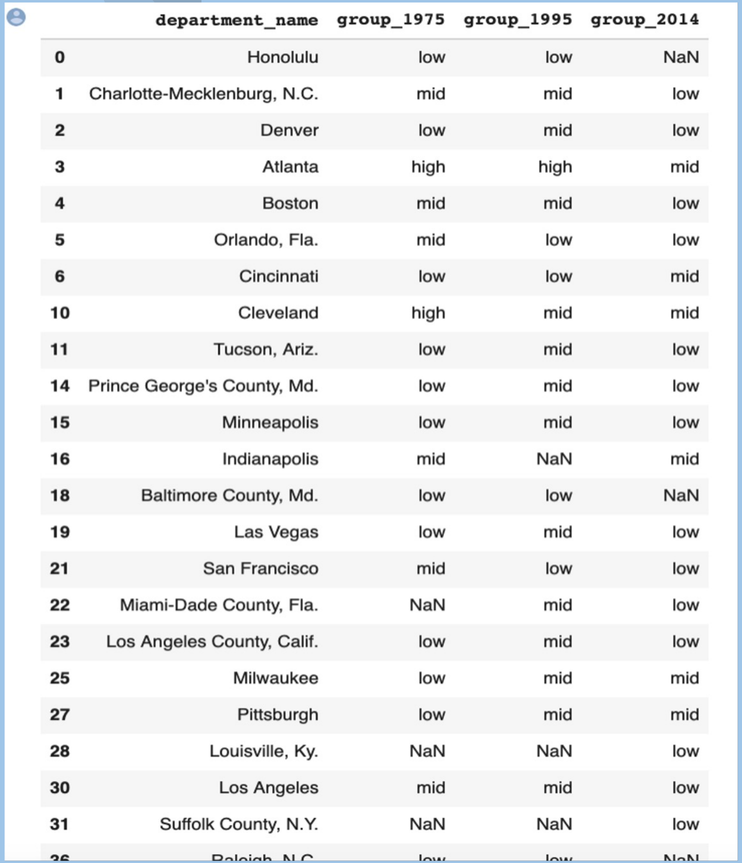


Statistic	months_reported	violent_per_100k	homs_per_100k	rape_per_100k	rob_per_100k	agg_ass_per_100k
count	2639	2639	2639	2639	2639	2639
mean	12	1117.1	15.7	59.8	467.8	573.8
std	0	677.3	12.4	31.7	341.9	369.3
min	12	16.5	0.2	1.6	11.5	1.6
25%	12	647.4	7.3	36.3	219.3	326.1
50%	12	977.5	12.4	56.4	381.1	493.6
75%	12	1433.2	20.5	78.0	619.6	732.5
max	12	4352.8	94.7	199.3	2337.5	2368.2

Statistic	Year	total_pop	homs_sum	rape_sum	rob_sum	agg_ass_sum	violent_crime
count	2639	2639	2639	2639	2639	2639	2639
mean	1994.6	790462	126.4	417.7	4070.9	4454.7	9069.6
std	11.6	1017433	205.9	481.1	8803.4	7063.0	16099.9
min	1975	100763	1	15	83	15	154
25%	1985	376700	32	180	1051	1500	3075
50%	1995	529121	63	292	1994	2636	5130
75%	2005	800235.5	128	465	3655.5	4579	8846.5
max	2014	8473938	2245	3899	107475	71030	174542

Summary Statistics

K-means clustering - cities in each cluster by year



	department_name	group_1975	group_1995	group_2014
0	Honolulu	low	low	NaN
1	Charlotte-Mecklenburg, N.C.	mid	mid	low
2	Denver	low	mid	low
3	Atlanta	high	high	mid
4	Boston	mid	mid	low
5	Orlando, Fla.	mid	low	low
6	Cincinnati	low	low	mid
10	Cleveland	high	mid	mid
11	Tucson, Ariz.	low	mid	low
14	Prince George's County, Md.	low	mid	low
15	Minneapolis	low	mid	low
16	Indianapolis	mid	NaN	mid
18	Baltimore County, Md.	low	low	NaN
19	Las Vegas	low	mid	low
21	San Francisco	mid	low	low
22	Miami-Dade County, Fla.	NaN	mid	low
23	Los Angeles County, Calif.	low	mid	low
25	Milwaukee	low	mid	mid
27	Pittsburgh	low	mid	mid
28	Louisville, Ky.	NaN	NaN	low
30	Los Angeles	mid	mid	low
31	Suffolk County, N.Y.	NaN	NaN	low
36	Raleigh, N.C.	low	low	NaN

We chose to use K-means clustering because we thought it would do a good job of categorizing and describing our data. We didn’t necessarily want a model that would predict outcomes as we felt that our explanatory variable (year) alone was not sufficient to explain the trends in crime.

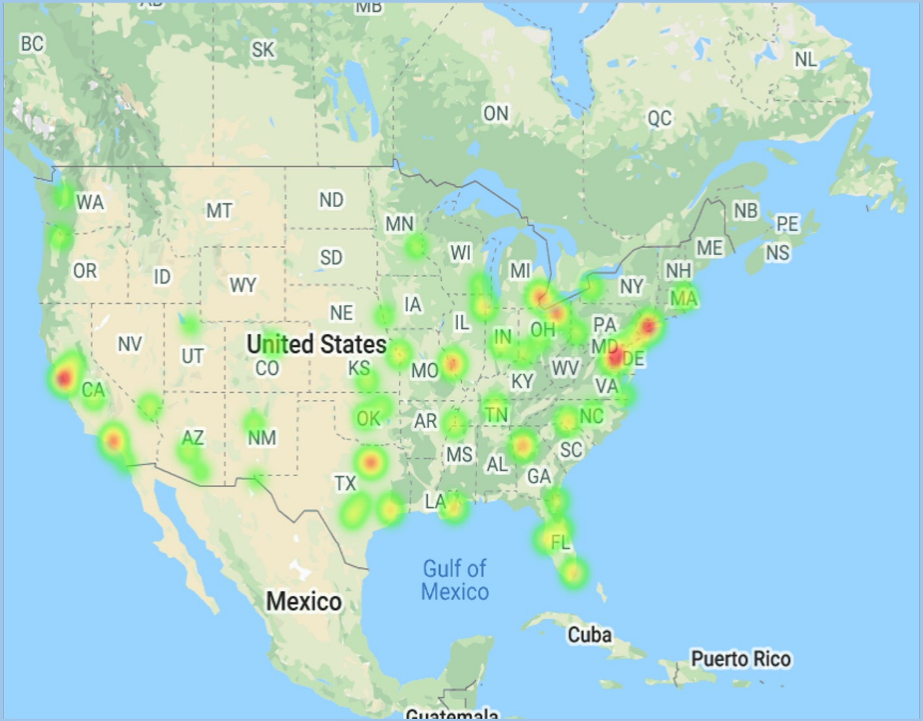
Our model created 3 cluster (0, 1, 2). We then assigned descriptive values to these clusters so that our data would be interpretable. In the image, you can see some of the cities and their cluster for the three sample years we choose.

Interactive Visualizations

Our bubble map most clearly visualizes our results from our Kmeans model. The Geo-Bubble maps allow us to visualize our data in the real world context.

1) **Bubble Maps**
   https://public.tableau.com/profile/asha7569#!/vizhome/data0200-final-AngelinaAshaNikolai/data0200
2) **Geo-Bubble Map**
   https://public.tableau.com/profile/nikolai1458#!/vizhome/Book1_16068939698790/GeoHeatmaps-Data200?publish=yes

Geo Heatmap - 1975

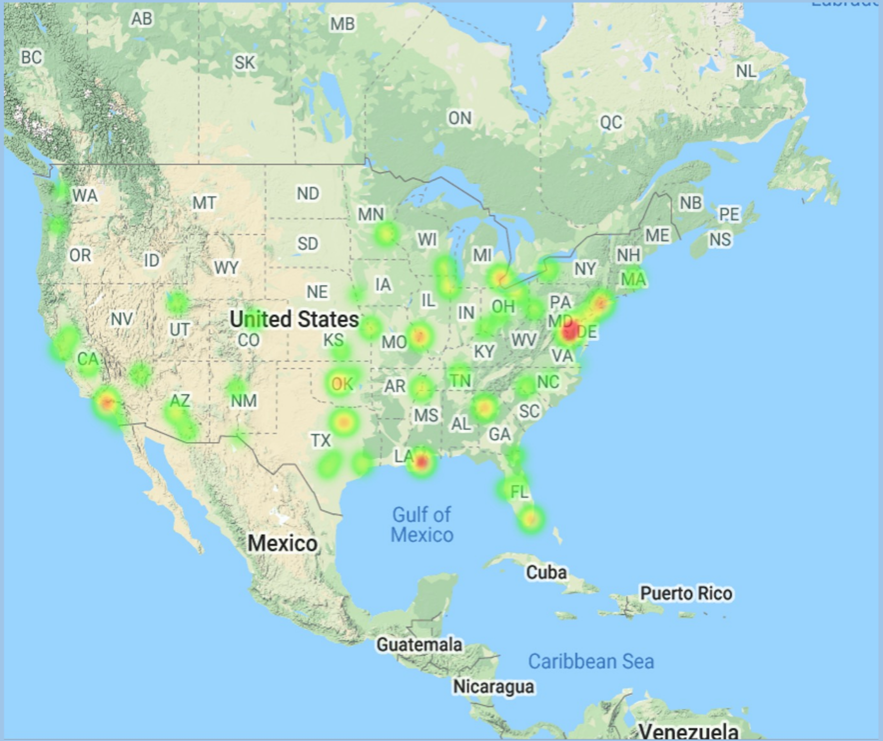


Geo Heat Map Comparing Homicides per 100k in Major Cities Across the United States.

In 1975 homicide rate varied between 1.64 and 44.19 per 100k.

The redder the color of the city, the more homicides occurred. As a result, this model shows which cities belong to which k mean cluster (low, mid, high). While there are many reddish bubbles, these are on the lighter side indicating relatively low rates nationwide.

Geo Heatmap - 1995

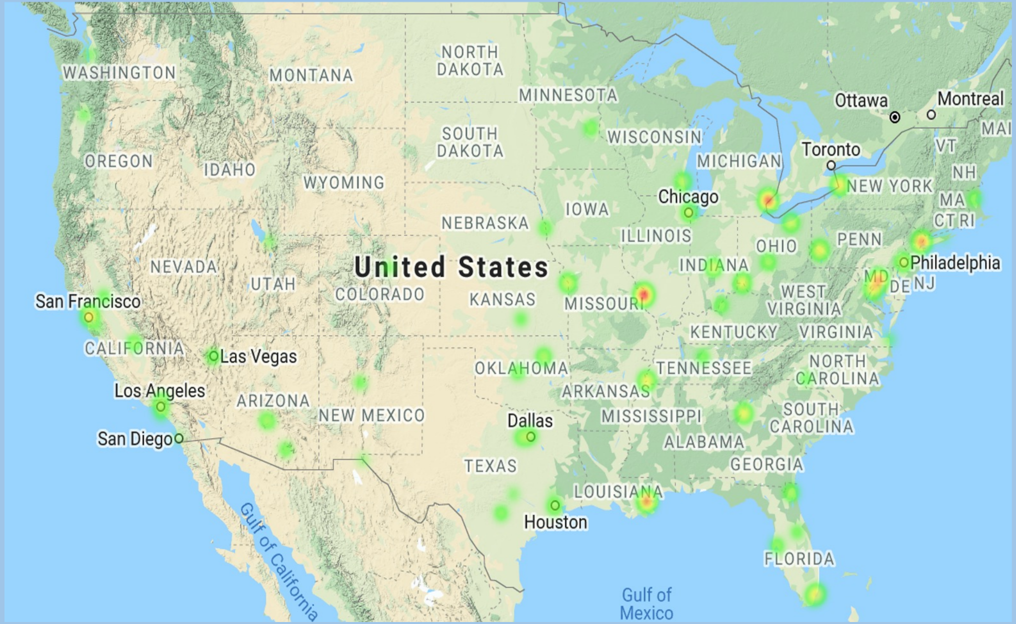


Geo Heat Map Comparing Homicides per 100k in Major Cities Across the United States.

In 1975 homicide rate varied between 1.66 and 74.51 per 100k. This was the highest rates of all three years and indicates an uptick in crime and homicide from 1975.

The redder the color of the city, the more homicides occurred. As a result, this model shows which cities belong to which k mean cluster (low, mid, high). In this case, while fewer red bubbles than 1975, these are of greater intensity thus indicating higher homicide rates.

Geo Heatmap - 2014



Geo Heat Map Comparing Homicides per 100k in Major Cities Across the United States.

In 2014 homicide rate varied between 0.57 and 49.91 per 100k. On average the homicide rates were the lowest of the three years indicating a decrease in crime and homicide since 1995.

The final map allows us to visualize our conclusion that many cities dropped to a lower crime cluster. This is evidenced in the decrease of redder bubbles on the map

Results

Out of the major U.S. cities changed clusters, the majority of them moved to a lower crime cluster (considering the years 1975, 1995, and 2014).

This means that the majority of cities dropped in relative homicide rates. It was observed that from 1975 to 1995 either cities remained constant (low to low or mid to mid like for example: Los Angeles) or they decreased in homicides per 100k (Cleveland for example). Then from 1995 to 2014 it was observed that most cities, even those that saw an uptake in homicide rates from 1975 to 1995, shifted and decreased (Salt Lake City for example). Many of the cities have changed from a higher crime group to a lower crime group over the three periods, 1975, 1995, and 2014. While there are certainly outliers that saw an increase, New Orleans for example, the majority did not. Many of the cities have changed from a higher crime group to a lower crime group over the three periods, 1975, 1995, and 2014.

Real World Applications and Implications

This is where domain knowledge plays a huge role!

We acknowledge that even though year was our explanatory variable, it DOES NOT explain why homicide rates have changed. Also, we do not assume that the trends observed necessarily imply that homicide rates will continue to decrease based on year alone.

The dataset was limited to crime rates and numbers ,however, we did not have data on policing, demographic information and other potential variables that may have been associated with crime which would have allowed us to make interference about are data and interferes about potential real world implications.

Next Steps

- Look into more recent years which would be more relevant
- Look into variables that may be associated with the changes/trends in homicides for cities For example:policing, demographics,etc.
- Potential Future Questions: What are the trends/changes in homicides in more recent years? What factors may be associated with the trends/changes in homicides in more recent years?