Purpose

Through the first three scaffolded activities you processed your data, explored your data, likely processed it again, and analyzed your data.  In the final step of the process, you will “make meaning” of your data through storytelling, both visually and in writing.

Process

Although this is the 4th scaffold, please submit your project in its entirety.  If you had separate colab notebooks, please combine them into one.

You may choose to create data visualization of your MODEL RESULTS (note this is not necessarily the same as your EDA from scaffold 2, though you may find some of your EDA may work here as well) using either or combining Python and Tableau (use whichever you feel most comfortable with).

You will also communicate the results of your project through writing.  You are free to adopt any citation/formatting style of your choosing.  You may write in first person or for a formal audience.  Regardless you are expected to be communicating at a college level.  This means that your written work will demonstrate editing to ensure there are no substantive grammatical, spelling or formatting errors.

Prompts

1. Repeat the context for your project, why did your group choose this topic?

We chose this topic because we thought it would be interesting and possibly informative to analyze crime changes in major US cities. While we initially considered a project that was more focused on incarceration, we decided to look at trends in crime as there was more data available that would allow us to create a meaningful and clear project.

1. Repeat the initial research question

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

1. Include a written discussion of your model results.
2. What (if any) modifications to your data were done?

NaN values were dropped - despite this we still had a significant number of observations to draw conclusions from the visualizations only show cities/department names whose cluster category did not remain constant for all 3 sample years. This creates a boundary on our conclusions that we must be mindful of.

1. Interpret in clear everyday language that most people can understand, what the findings of the analysis

Out of the major U.S. cities that did change 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.

1. Produce visual illustrations of your model results. These can be produced in Tableau or Python

**Interactive Visualizations**

* [https://public.tableau.com/profile/asha7569#!/vizhome/data0200-final-AngelinaAshaNikolai/data0200](https://public.tableau.com/profile/asha7569" \l "!/vizhome/data0200-final-AngelinaAshaNikolai/data0200)
* <https://public.tableau.com/profile/nikolai1458#!/vizhome/Book1_16068939698790/GeoHeatmaps-Data200?publish=yes>

Map

Description automatically generated

Geo Heat Map Comparing Homicides per 100k in Major Cities Across the United States in 1975

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

Map

Description automatically generated

Geo Heat Map Comparing Homicides per 100k in Major Cities Across the United States in 1995

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.

Map

Description automatically generated

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

1. In this section, you can very carefully begin to draw potential implications of your findings and propose possible explanations for your findings.

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. Provided this, we were limited in the implications drawn.

While year was our explanatory variable, we do not believe that the year can describe why the changes in homicide per 100k occurred. Additionally, we do not believe that the trends observed necessarily imply that homicide rates will continue to decrease based on year alone.

This leads us to question what variables (that we did not test for) might be responsible for the shifting of many cities into lower crime clusters over the years. This falls into domain knowledge.

1. What remaining questions arise from your analysis?
   1. Given what you found, what questions are still unanswered?  What might be the “next phase” of the project?

* What factors may be associated with the decline of homicides in the following years 1975, 1995, and 2014?
* 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?
* The next phase may be to look at more recent years which would be more relevant and maybe do a multiple linear regression to analyze if there are any variables associated with the changes in homicides. Some variables may be policing (overstaffed, understaffed, staffed, not staffed) and demographic details about the cities.
  1. What limitations does your data/model have that would have been helpful in better understanding your research question?

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.

Limitations in our data exist from the many of the choices we’ve made along the process. For example, the many cities we’ve had to drop due to inconsistent reported data - this limits the cities we’ve been able to analyze and impacts the generalization of our results. Additionally, thinking to the K-means model, one limitation stems from the fact that our clusters are not fixed. Our clusters are relative - for each year the homicide rates that qualify a city for the ‘high, med, and low’ clusters differ. So, when we say that cities dropped from the high to med cluster - that doesn't necessarily mean that the number of homicides dropped - rather in *comparison* to other cities, it had a med rate of homicides.

1. Share you google co-lab code that is appropriately commented and/or your Tableau story

**Code**

* <https://colab.research.google.com/drive/1R2UD0Jszwqn_vXB-sDRo_IPmttEPEpJ2#scrollTo=i5SIXeS52uV1>

**Visualizations (Interactive)**

* <https://public.tableau.com/profile/asha7569#!/vizhome/data0200-final-AngelinaAshaNikolai/data0200>
* <https://public.tableau.com/profile/nikolai1458#!/vizhome/Book1_16068939698790/GeoHeatmaps-Data200?publish=yes>

**Geo Heatmaps (Tableau)**

\*Code was created in Pycharm and uploaded into Jupyter Notebook as I was unable to get the following extensions to work properly in Google Colab

#Code Below Requires following extensions/plugins installed:

Gmaps

Jupyter Notebook

numpy

pandas

opencage

voila

#Imports cvs and splits department\_name into two columns

import pandas as pd

crime = pd.read\_csv('/Users/nikolaistambler/Library/Application Support/JetBrains/PyCharmEdu2020.1/scratches/crime.csv')

# Splits department\_name value into two new columns (City and State).

crime[['City','State']] =  crime.department\_name.str.split(",", expand=True)

#Drops all unnecessary columns

crime = crime.drop(columns = ["Unnamed: 0", "department\_name", "total\_pop", "homs\_sum", "rape\_sum", "rob\_sum", "agg\_ass\_sum", "violent\_crime", "months\_reported","violent\_per\_100k","rape\_per\_100k", "rob\_per\_100k", "agg\_ass\_per\_100k"])

#Creates three different datasets (based on year)

crime\_1975 = crime[(crime.year == 1975)]

crime\_1995 = crime[(crime.year == 1995)]

crime\_2014 = crime[(crime.year == 2014)]

# Imports opencage and configures for use (finding lon/lat values given a city name and/or state)

from opencage.geocoder import OpenCageGeocode

key = 'b59b13e931c54e949b2155ad995008ad'

geocoder = OpenCageGeocode(key)

# Create empty lists

list\_lat\_75 = []

list\_long\_75 = []

# Iterates through crime\_1975 dataset and populates list\_lat\_75 and list\_long\_75 with lat/long values of cities within the crime\_1975 dataset

for index, row in crime\_1975.iterrows():

    City = row['City']

    results = geocoder.geocode(City)

    lat = results[0]['geometry']['lat']

    long = results[0]['geometry']['lng']

    list\_lat\_75.append(lat)

    list\_long\_75.append(long)

# Create new columns from lists

crime\_1975['lat'] = list\_lat\_75

crime\_1975['long'] = list\_long\_75

# Create empty lists

list\_lat\_95 = []

list\_long\_95 = []

# Iterates through crime\_1995 dataset and populates list\_lat\_95 and list\_long\_95 with lat/long values of cities within the crime\_1995 dataset

for index, row in crime\_1995.iterrows():

    City = row['City']

    results = geocoder.geocode(City)

    lat = results[0]['geometry']['lat']

    long = results[0]['geometry']['lng']

    list\_lat\_95.append(lat)

    list\_long\_95.append(long)

# Create new columns from lists

crime\_1995['lat'] = list\_lat\_95

crime\_1995['long'] = list\_long\_95

# Create empty lists

list\_lat\_14 = []

list\_long\_14 = []

# Iterates through crime\_2014 dataset and populates list\_lat\_14 and list\_long\_14 with lat/long values of cities within the crime\_2014 dataset

for index, row in crime\_2014.iterrows():

    City = row['City']

    results = geocoder.geocode(City)

    lat = results[0]['geometry']['lat']

    long = results[0]['geometry']['lng']

    list\_lat\_14.append(lat)

    list\_long\_14.append(long)

# Create new columns from lists

crime\_2014['lat'] = list\_lat\_14

crime\_2014['long'] = list\_long\_14

# Drops year column from crime datasets

crime\_1975 = crime\_1975.drop(columns = ["year"])

crime\_1995 = crime\_1995.drop(columns = ["year"])

crime\_2014 = crime\_2014.drop(columns = ["year"])

# Creates new dataset containing only lat and long values for each of the three years (1975, 1995, 2014)

loc\_1975 = crime\_1975[['lat','long']].copy()

loc\_1995 = crime\_1995[['lat','long']].copy()

loc\_2014 = crime\_2014[['lat', 'long']].copy()

# Converts dataset values into an array

import numpy as np

locations\_75 = np.array(loc\_1975)

locations\_95 = np.array(loc\_1995)

locations\_14 = np.array(loc\_2014)

# Imports and configures gmaps

import gmaps

import gmaps.datasets

gmaps.configure(api\_key="AIzaSyBblVh51qwkxg6mhZcbWmjCd8WA8-qW3X4")

# Creates a geo heatmap of homicides per 100k in 1975. Uses homs\_per\_100k values to calculate density of map.

weights = crime\_1975["homs\_per\_100k"]

fig = gmaps.figure()

heatmap\_layer = gmaps.heatmap\_layer(locations\_75, weights= weights)

fig.add\_layer(heatmap\_layer)

fig

# Creates a geo heatmap of homicides per 100k in 1995. Uses homs\_per\_100k values to calculate density of map.

weights = crime\_1995["homs\_per\_100k"]

fig = gmaps.figure()

heatmap\_layer = gmaps.heatmap\_layer(locations\_95, weights= weights)

fig.add\_layer(heatmap\_layer)

fig

# Creates a geo heatmap of homicides per 100k in 2014. Uses homs\_per\_100k values to calculate density of map.

weights = crime\_2014["homs\_per\_100k"]

fig = gmaps.figure()

heatmap\_layer = gmaps.heatmap\_layer(locations\_14, weights= weights)

fig.add\_layer(heatmap\_layer)

fig

# Imports base64 and HTML

import base64

from IPython.display import HTML

# Function that exports csv file of modified crime data for use in Tableau

def create\_download\_link(df, title = "Download CSV file", filename = "df-heatmap.csv"):

    csv = df.to\_csv()

    b64 = base64.b64encode(csv.encode())

    payload = b64.decode()

    html = '<a download="{filename}" href="data:text/csv;base64,{payload}" target="\_blank">{title}</a>'

    html = html.format(payload=payload,title=title,filename=filename)

    return HTML(html)

# Creates the download link to get new CSV file

create\_download\_link(crime\_1975)

create\_download\_link(crime\_1995)

create\_download\_link(crime\_2014)