VG101 23SU Project1 - Chicago Crime Analysis

1. General rules

This project is designed to hone your programming skills in data visualization. It is strongly recommended that you do not leave the entire project to the last minute. It is an individual project. The project will be submitted electronically as **a typed report** in the format of pdf. Handwritten submission will not be accepted! You should use a professional type-setting program (Markdown or LATEX) for your report. **You should submit your code for each task to JOJ.** The deadline is **11:59 pm 7/17 2023**. No late submission will be accepted.

1.1 Task List

Tasks	Requirements	Points
Task 1: Dataset Split	data cleaning	5
Task 2: Error Detection	data cleaning	10
Task 3: Accumulated Crime cases	line plot	6
Task 4: Yearly Crime cases	bar plot	6
Task 5: Primary Type	pie chart	10
Task 6: Secondary Description	word cloud	10
Task 7: Crimes distribution in Police District	geobubble plot	10
Task 8: Boundaries of Community Areas	geoplot	10
Task 9: Density Plot with Community Area Boundaries	geodensity plot	5
Task 10: Do it yourself!	-	8

1.2 Grading Policy

- Correctness (80%): 6 figures and 3 tasks. In particular, common mistakes are missing labels, inconsistency in bounding box and grid, missing title, missing figure name, etc. (2 points each).
- Style (10%): The code follows a good style (naming, commenting, indentation).
- Description (10%): You need to describe what each graph means and what's your observations in your report.

1.3 Guidelines for plotting

Please follow these guidelines when you plot:

- Plot the right figure required by each task.
- Figure title should concisely describe the purpose of this figure. Axis labels are clearly labeled.
- Figure legends are explainable.
- The line should be clearly visible. Usually, the line width has to be adjusted rather than using the default size.
- All font sizes should be appropriate and readable. Usually, it has to be adjusted rather than using the default size.
- Tic marks may also need to be adjusted since the default tic marks may be too detailed.
- Colors should be distinguishable enough. Following a certain color pallete(scheme) might be a good idea.
- Your figures should be also readable and distinguishable between different cases in the gray-scale mode. It is also suggested being user-friendly to those color-blinded readers.

2. Introduction

2.1 Why we are doing this

Data analysis is widely used in all types of disciplines. Analysing data and visualizing it are important skills that allow us to understand the measured system more.

As one of the most common and powerful tools in today's world, data analysis is widely used in city planning for understanding the data and casting new light on city policies. Chicago city government has been using data to analyze crimes, which is also made public for research. We want to use this opportunity to let you have a closer look at it.

You would encounter many interesting tasks and would plot various types of figures, like bar, word cloud, density plot on actual maps etc.

2.2 Dataset Introduction

We will use two datasets for this project. They are both taken from Chicago Data Portal. The main dataset **Crime_2001_to_Present.csv** is taken from Crimes - 2001 to present and the crime data it contains will be useful for **all** of the tasks in this project. The dataset **CommAreas.csv** is taken from Boundaries - Community Areas (current) and irrelevant columns of data are deleted. It contains information about boundaries of the Chicago community areas and would be used for three tasks.

Because the dataset of recorded crime cases is very large, it takes minutes to run programs operating on the dataset even after we split it later. You are recommended to first test it fully on smaller dataset like Crimes_2022.csv or even a small sample of your own choosing. When the program is tested successful, try the larger dataset and patiently wait for a few minutes and output the final figure.

2.2.1 Crime_2001_to_Present.csv

This dataset provides reported incidents of crime that occurred in the City of Chicago from 2001 to 2023.5.5 with the exception of murders. Each case contains the date, type, description and location of the incidence.

The data have been anonymized in several ways to protect the privacy of crime victims: name of the victims are not shown, addresses are shown at the block level only and specific locations are not identified. However, even with these transformations of the data, researchers will be able to do data characterizations and simulation.

The data are structured as blank-separated columns. Each row reports on the information of a single reported incident of crime.

Each row will contain the following information:

1. ID(int): Unique identifier for the record
2. CaseNumber (string) - unique Chicago Police Department RD(record division) number
3. Date (datetime) - Date when the incident occurred, sometimes the best estimate
4. Block(string) - the address where the incident occurred, not specified for privacy
5. IUCR(int) - Illinois uniform crime reporting code, which directly links to PrimaryType and Description.
6. PrimaryType(string): primary description of the IUCR code 2
7. Description (string) - secondary description of the IUCR code
8. LocationDescription(string) - location where the incident happened
9. Arrest(string) - indicates whether an arrest is made
10. Domestic(string) - whether an incident is domestic related
11. Beat(int) - Indicates the beat where the incident occurred. A beat is the smallest police geographic area
12. District(int)-Indicates the police district where the incident occurred.
13. Ward(int) - The ward (City Council district) where the incident occurred. See the wards at
14. Community Area(int)-Indicates the community area where the incident oc- curred. Chicago has 77 community areas.
15. FBI Code(int)-Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).
16. X Coordinate(int)-The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
17. y Coordinate(int)-The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
18. Year(int)-Year the incident occurred
19. Updated(datetime)- On Date and time the record was last updated.
20. Latitude(int)-The latitude of the location where the incident occurred.
21. Longitude(int)-The longitude of the location where the incident occurred.
22. Location(string)-The latitude and longitude of the location, allow for map

2.2.2 CommAreas.csv

Chicago is divided into 77 community areas and this dataset provides current community area boundaries in Chicago.

Each row will contain the following information:

1. the_geom(string): the boundary points of the community area
2. AREA_NUMBE: the area number of the community area
3. COMMUNITY: name of the community area

In particular, the field the_geom takes the form of

```
'MULTIPOLYGON (((-87.60914087617894 41.84469250265398, -87.61616875933031
41.84514162682185, -87.61613940345198 41.8451280515347, ......, -87.60937658092277
41.84515338263664, -87.60914087617894 41.84469250265398)))'
```

with each pair of two doubles as latitude and longitude of a point on the borderline of a community area, e.g.(*-87.60914087617894 41.84469250265398*).

In other words, if the_geom of a community area A is 'MULTIPOLYGON (((0 0, 1 0, 1 1, 0 1, 0 0)))', the boundary points of A is (0,0) (1,0) (1,1) (0,1). You would be asked to connect these boundary points in later tasks.

3. Tasks

3.1 Data cleaning

Data cleaning is often an important early step in the data analysis process. It helps improve data quality, reduce errors and noise, ensuring accurate and reliable results.

By cleaning the data, we can remove duplicates, handle missing data, address outliers and erroneous entries, standardize data formats, etc., providing a solid foundation for subsequent data analysis and modeling.

3.1.1 Dataset Split

The dataset Crime_2001_to_Present.csv is very large and the dated records inside would greatly slow down our program.

To operate on the dataset more effectively, you can split the dataset to get crime records for more recent years. For tasks in 3.2, we will use crime records from 2015 to 2022 while for tasks in 3.3, we will use crime records in 2022 for efficiency.

Write a function called `splitData` to write the crime records from year **2015 to 2022** into a file named *Crime_2015_to_2022.csv* and write the crime records of the year **2022** into a file named *Crime_2022.csv*.

The function should have 1 input (string): the input file name

Sample Input:

```
splitData("Crime_2001_to_Present.csv")
```

It may require 5 to 10 minutes to finish running. So you see why we need to split the dataset.

3.1.2 Error Detection

The dataset CommAreas.csv contains some errors in the field **the_geom**. In a few community areas, the_geom is not in the consistent format as described in 2.2.2. Because the_geom of a community area often contains more than thousands of boundary points, it's difficult to identify the errors with one's eyes *(I bet you can't)*. These errors could be wrongly formatted numbers like --83, or unwanted characters in the middle like !,#,(, etc.

Luckily in our case, if we can find the first position of the errors in the_geom of a community error, we can take the boundary point data before the error and it would still provide enough information for us to draw the community area boundaries.

Write a function called `correctBoundary` and it would

1. identify which the_geom contain error among the community areas
2. find the first occurrence of error in the_geom
3. only use the boundary points before the first error and rewrite the_geom accordingly
4. Output the correct format data to **CommArea_fixed.csv**.

The function would have 1 input (string): the input filename

Sample Input:

```
splitData("CommAreas_Sample.csv")
```

CommAreas_Sample.csv

the_geom	AREA_NUMBE	COMMUNITY
'MULTIPOLYGON (((0 0, 1 0, 1 1, 0 1, 0 0, 1 0, ! 3 0, 1 1)))'	10	'A'

Sample output:

CommArea_fixed.csv

the_geom	AREA_NUMBE	COMMUNITY
'MULTIPOLYGON (((0 0, 1 0, 1 1, 0 1, 0 0, 1 0,)))'	10	'A'

- Hint1: you may find matlab function: `readtable` and `writetable` useful.

- Hint2: In Matlab, deleting multiple lines at the same time is much faster than deleting one line at a time.
- Hint3: in finding out where the community area data is wrong, bisection method and `str2num` would be useful.

3.2 Analysis of crime numbers by year

How many crimes are committed each year and in total? Is Chicago getting safer by year?

For tasks in 3.2, use the dataset Crime_2015_to_2022.csv.

3.2.1 Accumulated Crime cases

To characterize the growth of crime, we want to investigate the trend in accumulated crime numbers by year from 2015-2022. Plot the accumulated number of recorded crimes by year using a line plot. X axis is year, and y axis is accumulated number of crimes.

Hint: you may find function `sortrows` useful

Write a script called 'Analysis_of_crimeNumber_accumulated.m' and plot the figure as 'Accumulated_crimeNumber_line.jpg'

3.2.2 Yearly Crime cases

The number and change of crimes reported each year is an indication of the city public safety and influenced by many other factors. For example, we can expect that due to the covid19 pandemic lockdowns, the crime cases reported in 2020 and 2021 might be fewer. Plot the number of reported crimes each year from 2015-2022 using a bar plot. X axis: year, and Y axis: yearly number of crimes.

Write a script called 'Analysis_of_crimeNumber_yearly.m' and plot the figure as 'Yearly_crimeNumber_bar.jpg'.

3.3 Analysis of crime type

Which type of crime happens the most in Chicago? What are the hot keywords Chicago police use in describing crimes?

For task 3.3.1, use the dataset Crime_2015_to_2022.csv and for task 3.3.2, use the dataset Crime_2022.csv.

3.3.1 Primary Type

Plot the distribution of crime primary type using a pie chart. Plot the top 10 crime primarytype from 2015-2022 and the others as another category called "others" on the pie chart. Remember to label each category.

Write a script called 'analysis_of_crime_type.m' and plot the figure as 'crime_type_pie.jpg'.

3.3.2 Secondary Description

Plot the crime description of 2022 using a wordcloud. Crime description offers more details than primarytype, and we want to use a wordcloud to see the specific characteristics of crime in chicago. Some of the description contains primary type, like 'aggrevated battery' contains 'battery'. Exclude these primary type words from description when drawing a wordcloud.

Write a script called 'analysis_of_crime_description.m' and plot the figure as 'crime_description_wordcloud.jpg'.

Hint: for word cloud, use function `wordcloud`. The function `extractWords` defined in `wordcloud` help document could also be useful.

3.4 Analysis of crime location

Where is the most dangerous place in Chicago? Where is the safest? We want to analyze the location of the crimes by regions in Chicago and visualize it on a map.

For geographic plots, use geobasemap topographic. For task 3.4.1 and 3.4.2.2, use dataset Crime_2022.csv and for task 3.4.2.1, use dataset CommArea_fixed.csv.

3.4.1 Crimes distribution in Police District

Plot the distribution of crimes of 2022 in Chicago **districts** using geobuble plot. The bubble of a district could be place anywhere as long as it is within the region of the district. You should label the cases number and districts number.

Write a script called 'analysis_of_crime_location_bubble.m' and plot the figure as 'crime_distribution_bubble.jpg'

- Hint1: you may find function `geobubble` useful.
- Hint2: when choosing the location of the bubble of a district, you can put the bubble in the location of the first appearance of recorded case of the district if the position is valid.

3.4.2 Crimes distribution in Community Area

Plot the distribution of crimes of 2022 in different **community areas** using a geodensity plot and plot the community area boundaries on the map as well using `geoplot`.

3.4.2.1 Boundaries of Community Areas

Plot the boundaries of the community areas on the chicago map and label each community number (AREA_NUMB) in the middle of a community area.

Write a function called `drawBoundary` and you should call this function in 3.4.2.2 to add community areas boundaries to the map. The plotted line should be in the same color.

Function input(string): input file name

Sample Input:

```
drawBoundary("CommArea_fixed.csv");
```

- Hint1: `geoplot` would be useful for drawing boundaries.
- Hint2: draw one community area at a time, and hold the previous drawings
- Hint3: label each community area use `testm` (need to install `mapping toolbox`)

3.4.2.2 Density Plot with Community Area Boundaries

Plot the density of crime using `geodensity`. Call `drawBoundary` written in 3.4.2.1 and hold the drawings.

Write a script called "analysis_of_crime_location_density.m" and plot the figure as "crime_location_density.jpg".

-Hint: use `geolimit(manual)` to fix the frame of map after you have drawn the boundaries of community areas and want to plot density.

3.5 Do it Yourself!

Define another problem that interests you. Give the relating plot.

You may look at unexamined relationships in this dataset or checkout Chicago Data Portal and find some other data and problems relating to crimes. What could influence crimes in Chicago?

4. Reference

"VG101 22SU Project1-Google Datacenter Workload Characterization", VG101 22SU Teaching Group