San Francisco Crime: Descriptive Statistics and Geographic Visualization

Legal Studies 123, UC Berkeley

February 11, 2022

1 Introduction

For this problem set, we will be working with the San Francisco Police Department's Incident Database. The dataset contains up-to-date information on incidents reported to the SFPD. Each observations is tagged with information about the incident's location, type of infraction, and date/time. In this problem set you will:

1. Explore the data with summary and descriptive statistics
2. Map the incidents

Make sure to start early and ask lots of questions! The due date is March 4.

2 Get the Data

Download the data from bcourses. You should have:

1. `Police_Department_Incident_Reports__2018_to_Present.csv`
2. `SF Find Neighborhoods.geojson`

These datasets, along with much more publicly available data, are also available at:
`https://data.sfgov.org/d/wg3w-h783/data`

3 Descriptive Statistics

3.1 Plot the number of incidents per year from 2018 to present (choose the appropriate type of plot). Have crime rates increased or decreased in general?

3.2 To get a more granular look, plot the number of incidents per month per year from 2018 to present. How does added granularity change your previous analysis of crime rate increase or decrease?

3.3 Check to see if these relationships change when looking at particular types of crime.

Plot and explain your findings.

3.4 Looking just at 2019, what proportion of the total does each type of crime constitute? Use at least one table and at least one plot to support your answer.

Use at least one table and at least one plot to support your answer.

3.5 Is there a relationship between day of week, time, and whether an incident occurs?

3.6 Is there a relationship between day/time and particular types of incidents? What about time of year?

3.7 What neighborhoods experience the most crime? Do different neighborhoods experience different types of crimes at different rates, or is the distribution of crime spatially consistent across neighborhoods?

Use the "Police District" column for neighborhood information. Consider using grouping techniques to illustrate your answer through plots.

3.8 Discuss two other interesting findings from your data.

4 Geographic Data

4.1 Plot individual incidents in 2019 as points on a map of San Francisco

You'll want to use folium (which you saw in lab) and `geopandas`, which extends DataFrames from pandas to GeoDataFrames, which include geographic information. Find the documentation on GeoDataFrames here.

1. Does crime seem randomly distributed in space, or do incidents tend to cluster close together?
2. Shade the points by type of crime and analyze whether certain neighborhoods experience certain types of crime more often.
3. Propose social scientific explanations for the patterns that you find

Hint: Use a random sample of points if all of the points are too numerous to plot.

4.2 Merge the incidents data with the GeoJSON file which contains the information on the boundaries of neighborhoods in San Francisco.

Hint: Merging the incidents data, which includes the location of each incident, with the SF neighborhoods data, which describes each SF neighborhood and the geographic region it occupies, can be thought of as a spatial *join* across the two tables. Look back at the `GeoDataFrame` documentation to perform this join. In the next problems, you will use this merged data to make visualizations of the frequency of various crimes by neighborhood.

Important Note: When running the spatial join, you may get weird exceptions from inside of geopandas. This can happen when necessary libraries are not installed. To be sure you have what you need, run the following:

```
# If you are on MacOS, make sure you have the XCode developer tools:
$ xcode-select --install
# If you are on MacOS, install homebrew (at https://brew.sh/) and run:
$ brew install spatialindex
# If you are on Ubuntu Linux, do the equivalent with:
$ sudo apt-get update && apt-get install -y libspatialindex-dev
# Everyone, be sure the python dependencies are installed:
$ pip install rtree
$ pip install geopandas
```

If you spend 30 minutes or more on errors coming from importing geopandas, or other internal errors in geopandas, ask in office hours, in lab, or on Piazza.

Hint: When making a GeoDataFrame, note that there is a `crs` attribute (the *coordinate reference system*) that you should take care to set to {`'init': 'epsg:4326'`}. You can do this by either assigning to it directly (like `mygeodataframe.crs = {'init': 'epsg:4326'}`) or by using the keyword argument to the constructor (`GeoDataFrame(..., crs={'init': 'epsg:4326'})`). This short geopandas guide explains what this means and what the CRS is for.

5 Mapping Incidents

5.1 Construct a choropleth map, coloring in each neighborhood by how many incidents it had in 2019. Then, construct several maps that explore differences by day of week, time of year, time of day etc.

5.2 Do you notice any patterns? Are there particular neighborhoods where crime concentrates more heavily?

5.3 Construct a heat map of crime.

How does the heat map compare to the choropleth map? Are neighborhoods a reasonably good proxy for the actual concentration of crime?

6 Discussion Questions

6.1 Based on the evidence from this lab assignment, why do you think "hotspots" policing became more popular in the last few decades? What are the pros and cons to this kind of approach?

6.2 Comment on what sorts of incidents get reported in this database

For instance, do you see a lot of reports about things like white collar crime? How do you think incident categories are selected? As data scientists, what kinds of ethical and legal concerns should we be aware of when we construct these sorts of datasets?

6.3 What other sorts of data would help improve your analysis?