INFO 3300 - Data Driven Web Applications
Mary Kolbas (mck86), Tammy Zhang (tz332), Daisy Tseng (dt369), Luke Ellis (lke8)
15 November 2022

Project 2 Final Report

Students across Cornell have recently seen an increase in "Crime Alert" emails that outline crimes relating to or taking place on Cornell's campus. While there is certainly a lot to be said about some of the recent crimes that have occurred (especially the more well-known ones like the Ganędagǫ arsons and the stolen letters of the Cornell sign), it is best to discuss this issue with data to back up these claims. We decided to take a look at what we could do with Cornell's crime data to help students better understand safety on our campus, and to encourage them to maintain awareness. Cornell is not without criminal threats, so a sense of caution is still important.

Data Description

Our idea was clear from the start: we would generate a map with points representing the location of different crimes. Different colors would represent different types of crimes, and the map would potentially help us to discover "hotspots" for criminal activity. The idea seemed straightforward, but the hurdles for making this dataset quickly became clear. For our first dataset, the Cornell University Police Department only publishes the last 60 days of criminal reports at once, so it would be hard to create a large dataset. We solved this by referring to the Wayback Machine from Archive.org. This website takes "screenshots" of websites and lets the user view the changes. We were able to find archives of the CUPD Daily Crime Log website that were mostly 60 days or fewer apart, which we used to compile our final crime dataset.

The next challenge was figuring out a way to convert the CUPD's locational statement (e.g. Ganedago Hall, Physical Sciences Building, etc.) into geographical coordinates. We found a free Geocoding API that allowed us to convert addresses or building names into coordinates and returned the data as JSON. So our web scraping process was complete: we would scrape the Wayback Machine for historical crime data, then we would run the locations through the API to get our coordinates. While this worked for most of the data points, there were about 300 points without coordinates. We considered these points important for our story, so we decided to manually input their coordinates from Google Maps and Cornell's Location Directory search.

Next, for our second and third datasets, we needed to gather map data to make a TopoJSON map of Ithaca. We found a GeoJSON data file for Tompkins County municipality boundaries and a GeoJSON of most of the buildings on Cornell's campus and in Tompkins County. The county boundaries were cropped down to Ithaca city and Ithaca town limits using ArcGIS Pro, as we noticed most of the data points were concentrated in this area. These GeoJSON files were converted from a NAD1983 projection to WGS84 projection with GIS, then converted to TopoJSON using mapshaper.org and simplified to greatly reduce the number of paths in the file.

Dataset Source Links

1) CUPD web scraped dataset:
    - https://dailycrimelog.cupolice.cornell.edu/
    - https://web.archive.org/web/20220000000000*/https://dailycrimelog.cupolice.cornell.edu/
2) Tompkins county GeoJSON
    - https://cugir.library.cornell.edu/catalog/cugir-008030
3) Ithaca buildings GeoJSON
    - https://cugir.library.cornell.edu/catalog/cugir-008163

Final combined dataset: crime-data.csv

Field name	**Type**	**Description**
IncidentType	Categorical String	CUPD's categorization of criminal offense
ReportNumber	String	CUPD's unique identifier for reports; used to remove duplicates during web scraping, first two numbers are year of incident
Reported	String	Time of the report
Reported_date	Datetime	Converted "Reported" string into datetime
Occurred	String	Alleged estimated time of the crime
Location	String	An address or building name where the crime allegedly took place
Narrative	String	CUPD's description of the events of the report
Disposition	Categorical string	The police response/status of the case at the time of the web scraping. Includes: "arrest," "closed," "unfounded," and a few more categories.
coords	String	A latitude and longitude in degrees in WGS84 projection as determined by the Geocode API or manually, in the format ([Latitude], [Longitude])
incident_bin	Categorical string	The categories used to reduce the amount of categories of IncidentType. Rather than having 40+ categories, we reduced it to 7 categories using crime

		categories from Justia.com and *Social Problems: Continuity and Change* from UMN's library.

Design Rationale

We chose to modify our Tompkins County and Ithaca Buildings TopoJSONs to reduce their geographic range and fidelity using ArcGIS Pro and mapshaper.org. This was an important part of usability because it significantly reduced the lag when rendering the page and interacting with pan/zoom, although there still is a delay. This means we had to remove data points outside of this geographic range, which was a decent tradeoff because we wanted our visualization to focus on crimes near Cornell's campus as opposed to all of CUPD's external agency assistance off campus.

Our marks are circles or dots. We chose to display our crime data points as dots on a map as opposed to using a contour plot or other aggregate display methods. We decided that it was important for the user to be able to select specific data points and see additional information, which would not be plausible if all data was aggregated already. By using a slightly translucent opacity, data points can be layered and give the user a sense of density. We also implemented a small jitter function so that data points at the exact same location could be distinguished and clicked, while keeping all data points on their respective building. We straddled the line between making the data points clickable without misrepresenting the data point's true location with too much randomization. Thus, we priortized having our final map show all data points on their correct building.

For our channels, we have vertical and horizontal aligned positioning due to our three data layers using the same map projection (WGS84). In addition, we also utilized hue as another channel. We chose distinct muted ordinal colors for our incident type "bins"/filter categories given our serious topic of CUPD and Cornell campus crime, avoiding any colors that could rely on stereotypes or have misleading connotations. The filter buttons provide discoverability and act as a legend, matching with the corresponding colors of the data points.

Interactive Elements

For our first interactive element, we chose to implement pan and zoom for the map due to the large amount of data and large geographic area we chose to present. This allows the user to see all points from afar, but also pan and zoom into specific areas and more carefully identify and click points of interest.

As users hover their mouse over the points, a small box appears with the location of the crime as noted by the CUPD. Since none of the TopoJSON we used included street names or building names on the map, the tooltip is essential for easy navigation. We found it important to implement this hover interaction displaying location because we realized it was difficult for the user to get a sense of where they are with just the building outlines. Therefore, with the tooltip, the user knows what building they are looking at. We made them discoverable by showing the location box below the circle over which the mouse hovers.

In addition, we also chose to allow the user to click on the data points, which reveals more information on the details of that report in the sidebar. By default, the sidebar provides instructions to the user to inform them that they are able to click to interact with the data points and what they should expect to appear. When a point is clicked, a black border outlines the data point to indicate which report is being viewed. This is a common user interaction that is often instinctive by users viewing online maps thanks to tools like Google Maps.

Furthermore, we added filters to our visualization that support data points that fit into multiple categories. Since there were originally 40+ crime types, we decided to create 7 bins of overarching categories of crime, enabling better usability. This was only a small tradeoff as the original incident type can still be seen in the details sidebar when a datapoint is clicked. However, for filtering, it would be more usable to have only a few categories that the user can filter by. We utilized online resources to determine the criteria for the 7 bins we wanted to focus on for the filters as detailed above. We also utilized affordances for the buttons as they have hover properties to indicate they are interactive, making them discoverable and usable. The buttons also have the same color as their respective data points and become bold/heavier and filled to indicate selection. The process of deciding this was that before, we didn't have the color correlation and found it difficult to match up the filtering to the points. Likewise, it wasn't clear that the buttons could be clicked on. Thus, we made the appropriate changes to create our final filtering interaction.

We went through several iterations for the time slider. We decided to visualize time like one does a timeline: a horizontal bar with the past traveling towards the present from left to right. First, we made a cumulative time slider that added data points as the user moved the slider. However, we thought it would be more informative if the user could see separate chunks of time, since the cumulative graph became cluttered and more difficult to read. Thus, we binned the dates into semesters rather than allow the user to specifically select a time period. This was a tradeoff, but ultimately our data visualization story wanted to capture campus crime and it was appropriate to divide time in a way parallel to campus activities. Using a smooth slider, if a user wanted to compare crime counts for each semester, they would have to manually find the end and start dates of each semester and enter them into the filter. This method allows them to jump between semesters and see a bigger picture more easily.

The Story

As expected, this map gives us several important insights into crime at Cornell. First, crime reports are more concentrated around Cornell's campus. While this is certainly due to the fact that this is CUPD data (and they normally respond to student calls), our visualization also provides additional information to explore the most prevalent types and locations of crimes. And if users were not already aware of these trends, perhaps they would think to be more cautious. Another important takeaway is that the "Crimes against Property" category of police reports is the most abundant form of crime. In other words, vandalism and property theft are among the

frontrunners of incidents at Cornell. This is a surprising insight and serves as a reminder to students to take good care of their valuable possessions.

This visualization also interestingly reflects the effect of the COVID-19 pandemic on student activity. Fall of 2020 was the first semester of on-campus instruction after the start of the pandemic, and student activity was extremely limited by the University. Parties did not occur frequently, classes were mostly on Zoom, large group gatherings were not permitted, students remained more isolated than before, and the in-person semester ended before Thanksgiving weekend. As a result, it makes sense that not nearly as many crimes were reported, since students were more scarce and less active. Compared to the Fall or Spring of 2022, the Fall of 2020 semester has noticeably fewer points on the map. In the Spring of 2021, the COVID-19 vaccines started being distributed to students, and we saw campus begin to open up more. Vaccinated students certainly felt safer to go out, pandemic fatigue set in, and the weather became warmer, which all invited more activity in general. And this semester also saw more crime than the previous "locked-down" semester, which was a fascinating and surprising trend conveyed through our visualization.

Team Contributions

- Mary Kolbas:
    - Data cleaning (manually adding/correcting coordinates) - 1.5 hours
    - Static map and minor interactions (basic pan/zoom, css for click/hover) - 2 hours
    - Binning data and implementing data colors (pre-processing) - 3 hours
    - Filtering by incident bin - 3 hours
        - Supporting data points that fit multiple bins - 3 hours
    - Time slider - 3 hours
    - Responsiveness for slider, button usability - 1 hour
    - Final report - 2.5 hours
- Luke Ellis:
    - Web scraping: scraping all the pages with crime data and finding their corresponding coordinates. - 3 hours
    - Data cleaning: manually editing the rows and columns to make sure data is accurate - 1 hour
    - Writing of final report - 2.5 hours
- Daisy Tseng:
    - Data cleaning (coordinates) - 2 hours
    - Mouseover hover tooltip interaction - 3 hours
    - Experimentation, debugging, colors - 1 hour
    - Final report - 2.5 hours
- Tammy Zhang:
    - Converting GeoJSONs to usable TopoJSONs and putting data into svg elements (cropping shapefiles and points in ArcGIS Pro, applying projections) - 6 hours

- Coding web page layout - 3 hours
- Implementing clickable points - 1 hour
- Debugging and styling - 3 hours