Cody Whitt
pkz325
CPSC 4530
Spring 2023
Assignment 3

Github for Project: https://github.com/codydev901/4530_assignment_3

# DataSet 1: Chattanooga DUI Arrests (Geospatial Point)

## Source and Inspiration

In assignment 1 I used Chattanooga crime data to explore criminal trends in relation to day of week and time of day. I remembered that the dataset also contained geospatial information (latitude and longitude), so I decided to use that part of the data for this assignment. I was interested primarily in visualizing the locations of DUI arrests within the Chattanooga area, and as an additional part of the analysis I also added a temporal component (season) to see if any additional relationships relating to that could be seen.

Link:
https://www.chattadata.org/Public-Safety/Police-Incident-Data/jvkg-79ss

Raw Data Filename: Police_Incident_Data.csv

## Dataset Type

This dataset is in the form of a flat table. It exists as a single .csv file which was downloaded from the above link and used locally from then on. In terms of this project it is being used statically, but again has potential to be used dynamically.

## Data Type

Data exists in this dataset in the form of items and attributes. Items exist as individual rows in the table and the attributes are defined as the column labels (the first row of the .csv). The data contains 516416 items and 19 attributes in its raw form, where each item could best be described as an entity holding information representing an individual occurrence of a crime. Temporal and geospatial information is provided for each crime through attributes holding a date, longitude, and latitude.

## Attribute Types and Semantics

As mentioned above, 19 attributes exist in the dataset, however only 4 were used for this analysis. The two geospatial attributes used were *Latitude* and *Longitude* which held floats and tied the item to a location on the Earth's surface/provided a location. I also used *Date_Incident* which represented a temporal attribute and contained a date string in the format %Y-%m-%d %H-%M-%S %p and *Incident_Description* which held and string and contained a simple description/classification of the type of crime that the item represents.

## Pre-Processing and Plotting

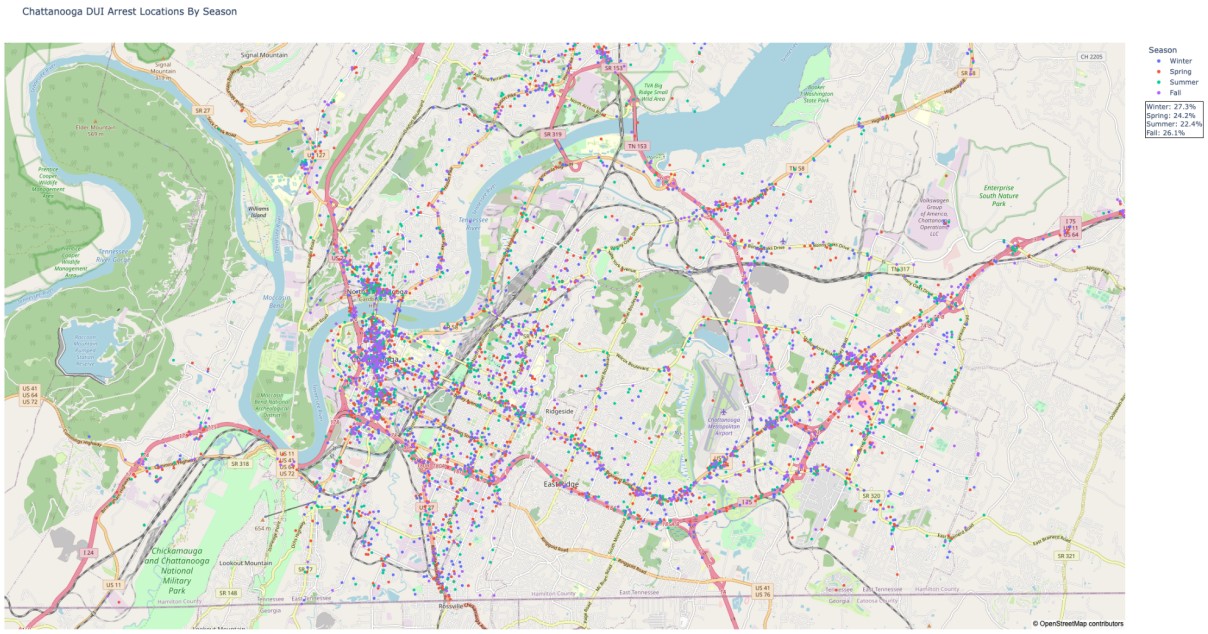
I used Python to process and plot the data, in particular Pandas and Plotly for this data set, with use of Plotly's integration with OpenStreetMap being the core part of the visualization. To process the data, the raw .csv file was read into a DataFrame and an initial check performed by called .head() and .info(), this then allowed me to determine which attributes would be needed. I then filtered the DataFrame to only include crimes that I was interested in (Driving Under The Influence) and dropped null values. This left 4493 items remaining of the original ~516k. Since I was also interested in grouping the items by season (Winter, Spring, Summer, Fall), I then wrote a quick function to create a new attribute *Season* from the information that existed in each item's *Date_Incident*. The resulting data from the above steps was then written to a separate file to be used in the plotting step.

To plot I used Plotly Express's scatter_mapbox() with an argument to utilize a geographic map provided by OpenStreetMap. The map itself is a conformal, cylinder projection. The items (instances of DUI Arrest at a particular location/season) are mapped to discrete points on the map by their associated latitude and longitude, with the categorical season attribute being encoded by color. The resulting figure could best be described as a dot map. Also to mention, I ended up adding an annotation to the figure underneath the legend that shows the seasonal distribution (which ended up being roughly uniform). There isn't a clear trend regarding that aspect, so I wanted to provide the viewer with some additional information there.

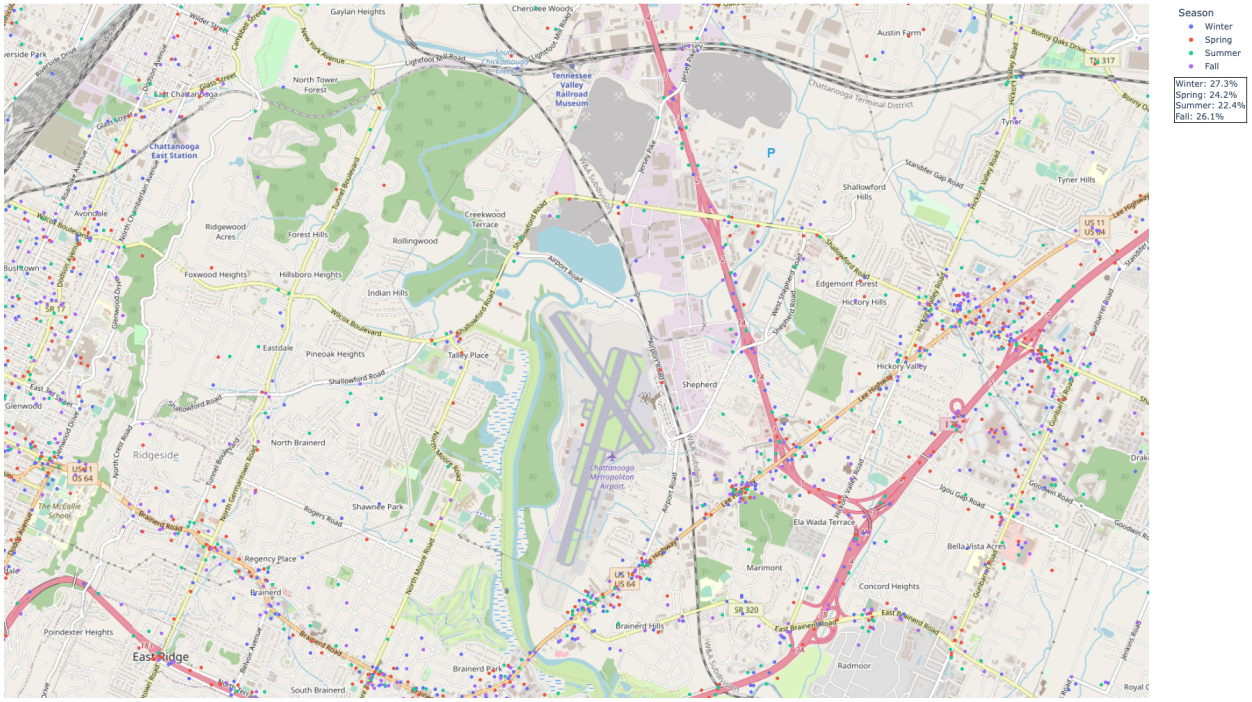
Parsed Filename: chattanooga_dui_season_parsed.csv

The parse and plot steps for this DataSet exist as separate functions and are contained in the parse_plot_dui.py file.

## Figure




Full Figure / Default Zoom




Zoom to the region mentioned in the analysis (Wilcox Boulevard/Shallowford Road in middle, Brainerd Road/Lee Highway at bottom, left to right)

## Analysis

So the primary objective of the visualization is to display the locations where DUI arrests have occurred within the Chattanooga area. The density of points within certain sub-areas compared to others shows which parts of the city have more (or fewer) DUI arrests. As perhaps expected, the DUI arrests seem to be mostly concentrated downtown and are present along major roadways (which I suppose is good in terms of “trusting” the underlying data/pre-processing, it would be interesting, but potentially suspect, if the points were not distributed this way). What I was hoping to find, and do potentially see a case of, is a difference in DUI arrests occurring along paths between two high-traffic areas. For example, when moving between Downtown and Hamilton Place (roughly left to right), one has a number of options. Two of these include Wilcox Boulevard/Shallowford Road and Brainerd Road/Lee Highway, where the first route appears to have much fewer DUI arrests than the second route (Shown in zoomed sub-section of Figures above). This could suggest that if one wanted to avoid drunk drivers, it would be better to take the first route (or in a more “evil” use, that if one was a drunk driver, it would be better to take the first route to minimize chance of arrest). The topic could benefit from further exploration in terms of relating to overall traffic on the roads in order to potentially normalize the arrest statistics.

In terms of the secondary objective (seasonal analysis), the arrests ended up being uniformly distributed, so the visualization didn’t provide too much value there in terms of seeing a trend relating to area (though I guess no-trend is still a trend). I added the seasonal distribution annotation to provide additional context for that part. It does show that DUI’s are slightly more common in Fall/Winter than in Spring/Summer, which perhaps is due to the winter holidays.

# DataSet 2: WW2 Allied Bombing in Europe (Geospatial Area)

## Source and Inspiration

Similar to DataSet 1, I ended up using different attributes of an earlier dataset that I had found interesting/wanted to continue to work with. There is an Allied aerial bombing dataset from WW2 on Kaggle and among many things, it contains geospatial information relating a magnitude of bombs dropped on a particular geographic point at a given time. I decided to use this in a visualization showing the progression of Allied bombing activity in Europe between the years 1940 - 1945.

Link:
https://www.kaggle.com/datasets/usaf/world-war-ii?select=operations.csv

Raw Data Filename: operations.csv

## Dataset Type

This dataset is in the form of a flat table. It exists as a single .csv file which was downloaded from the above link and used locally from then on. This one is completely static, since WW2 is over. I suppose it does have the potential to change if errors are found or new records uncovered.

## Data Type

Data exists in this dataset in the form of items and attributes. Items exist as individual rows in the table and the attributes are defined as the column labels (the first row of the .csv). The data contains 178281 items and 46 attributes in its raw form, where each item could best be described as an entity holding information representing the occurrence of a mission flown by a particular type of aircraft on some date, somewhere, dropping some amount of bombs in WW2. The “somewhere” mentioned in the previous sentence provided the geospatial information needed for this analysis in the form of latitude and longitude.

## Attribute Types and Semantics

I ended up using 4 of the native attributes for this analysis. For geospatial information, *Target Latitude* and *Target Longitude* held floats and tied the item to a location on the Earth’s surface/provided a location. *Mission Date* held a string which provided temporal information that I would later pull the *year* from. And finally *High Explosive Weight (Tons)* which held a float and would end up being used to provide a weighting magnitude encoded in the visualization. This allowed for an attempt at visualizing the change in both bombing locations and magnitudes over time.

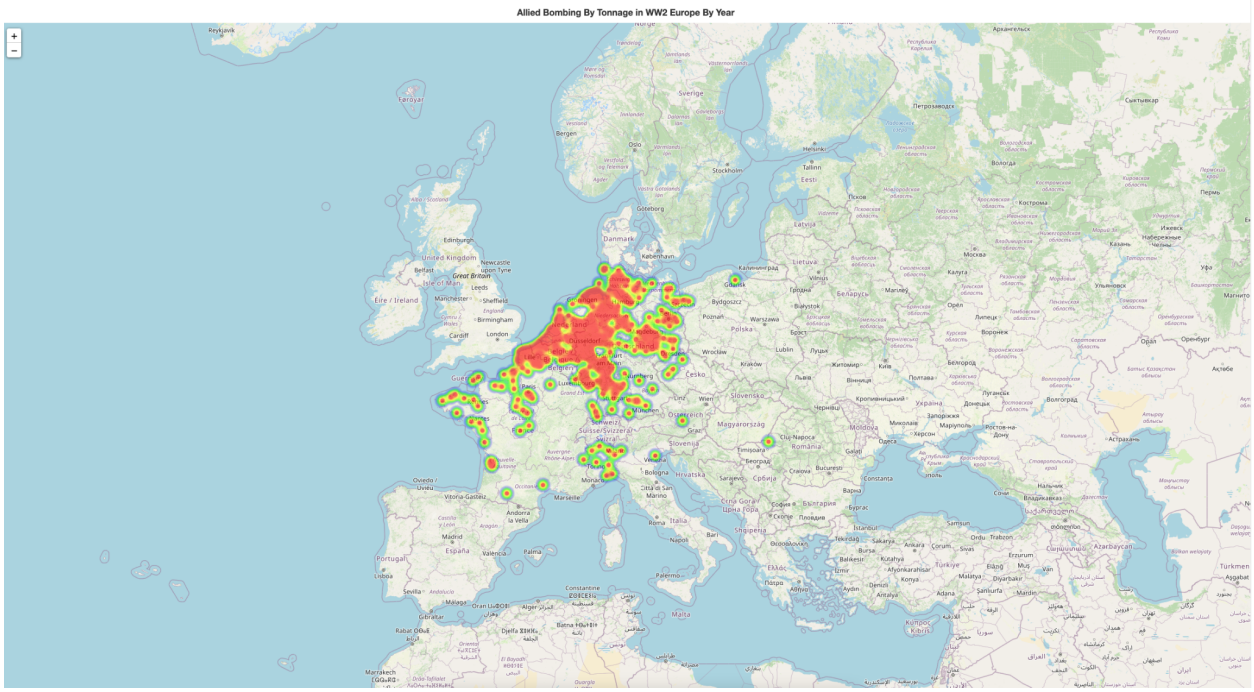
## Pre-Processing and Plotting

I used Python to process and plot the data, in particular Pandas and Folium. After seeing Folium suggested in the assignment requirements, it looked like its HeatMapWithTime() would lend itself well to this dataset in terms of providing a continuous area-based spatial coverage along with weighting provided by the bomb tonnage. It was also a good opportunity to try something other than Plotly which I have used most often so far. The data was first loaded into a DataFrame and filtered down to the attributes mentioned above and null values were then dropped. This left ~139k of ~178k items. Since I was only interested in Europe for this, I then filtered again on Latitude and Longitude to remove items not falling within a certain boundary (Europe). Google suggested using (34, 72) -> (-25, 45), so I used that (and it seemed to work fine when looking at the visualization later on). This removed a further ~38k items that represented missions occurring outside of the geospatial area I was interested in. Finally I wanted to quickly check on the quality of the *High Explosives Weight (Tons)* field, I looked at the min and max values present and saw a min of 0.0 and max of 999.0. The max seemed somewhat reasonable, but I didn't want to use 0.0 values, so I then dropped those. There only ended up being ~300 or so 0.0 values there. So finally I was left with ~101k items, where I then wrote a quick transformation function to extract a *Mission Year* from the *Mission Date*. This resulting DataFrame was then written to .csv to be used in the plotting step.

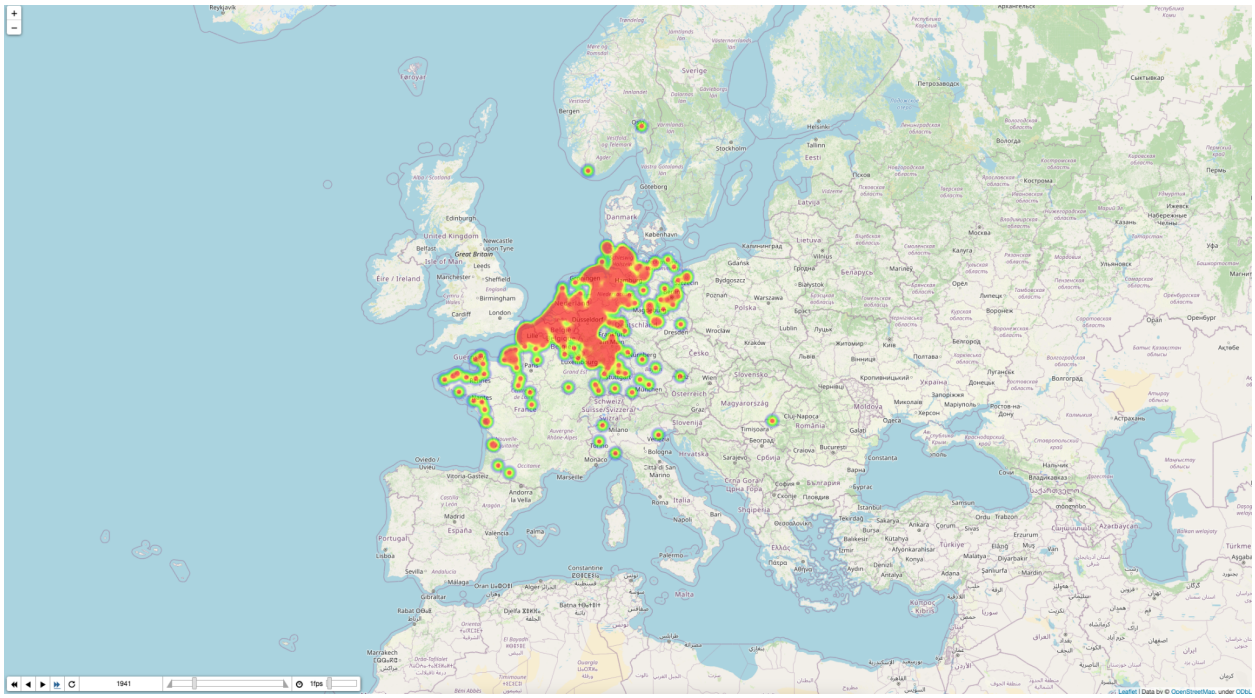
To plot, I used Folium's HeatMapWithTime() functionality where latitude and longitude provided the location for each item, explosive tonnage provided an associated weight, and the time interval was the year the item/mission occurred. The figure was set to center on Europe, and then written to .html, where I then viewed in the browser to interact with. The visualization itself could be described as an isarithmic map, where bombing activity in a geospatial area is represented by the contours of the heatmap overlaid upon it. To capture the progression of the time component, I included figures for each of six years below.

Parsed Filename: ww2_bombing_parsed.csv

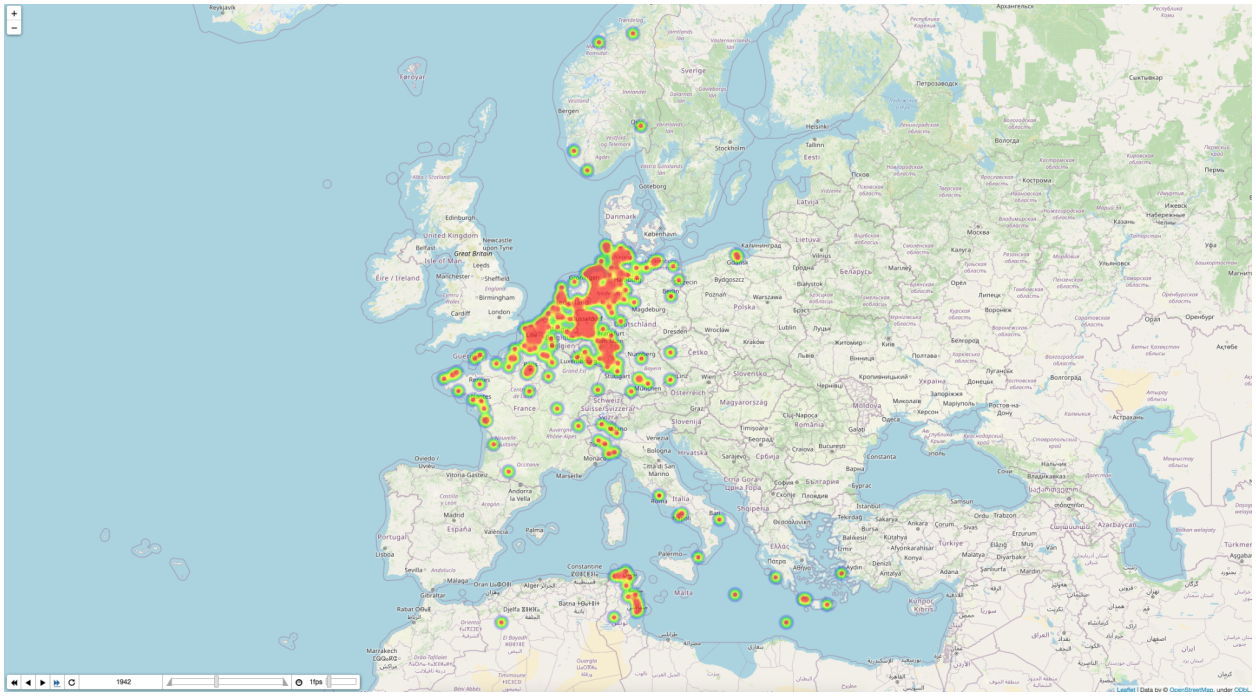
Figure



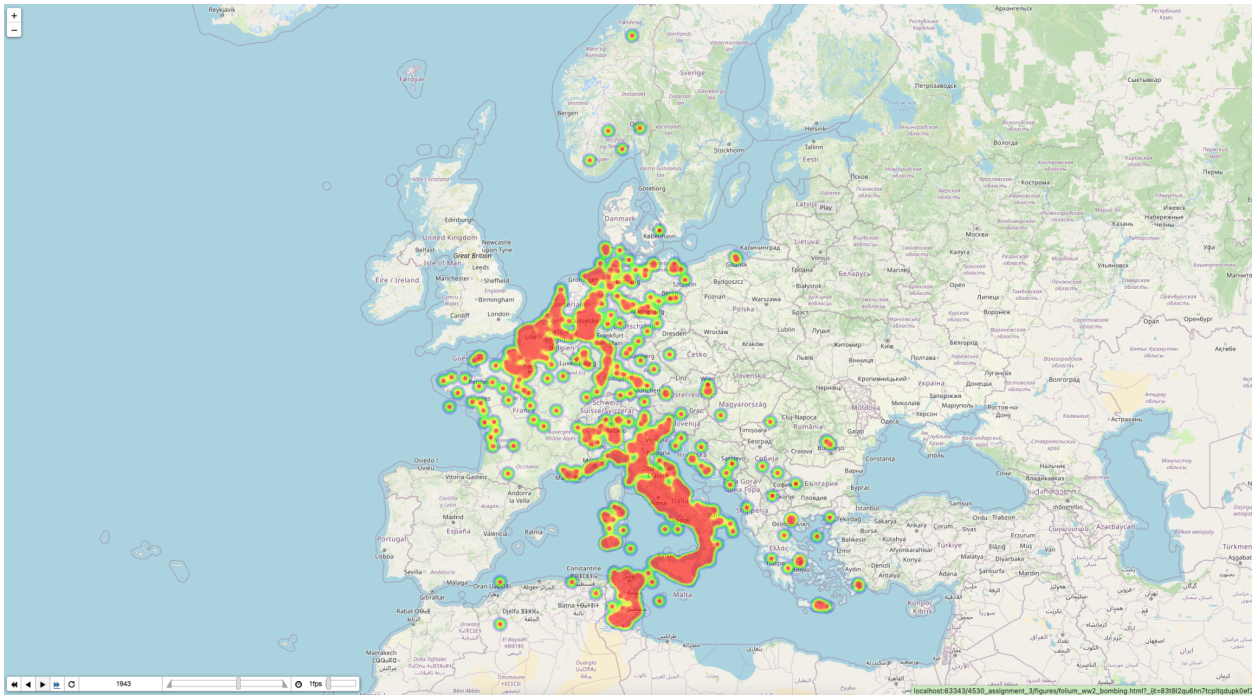

1940



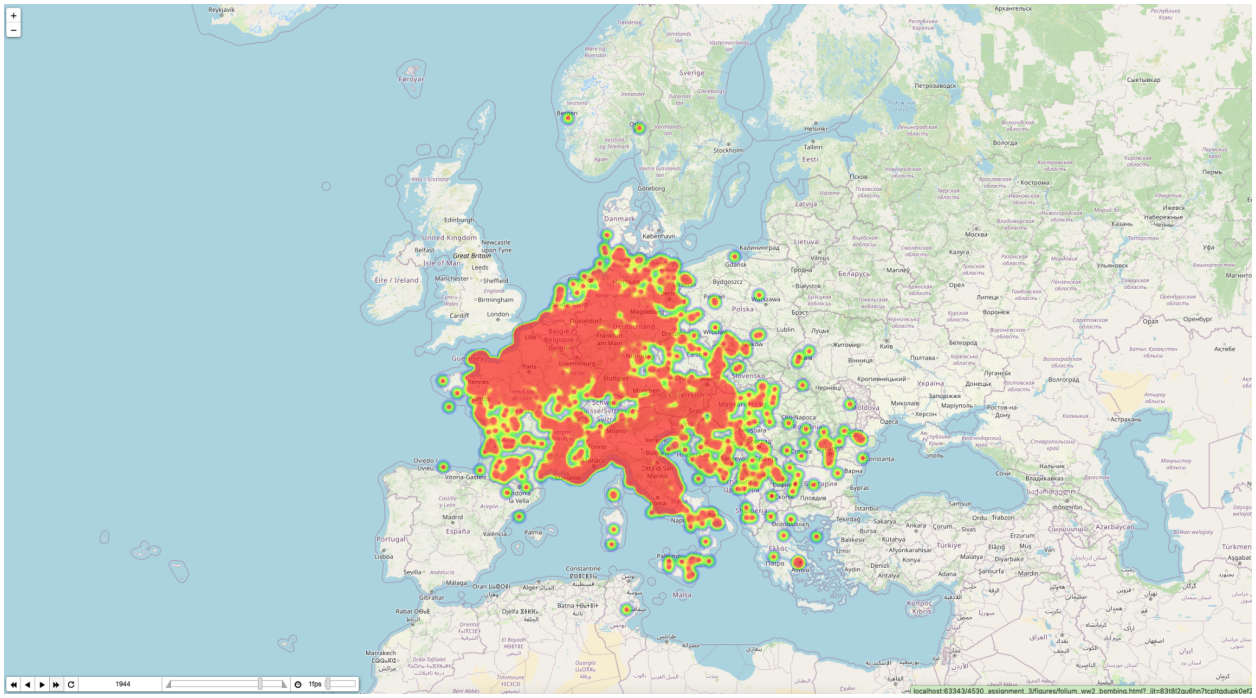

1941



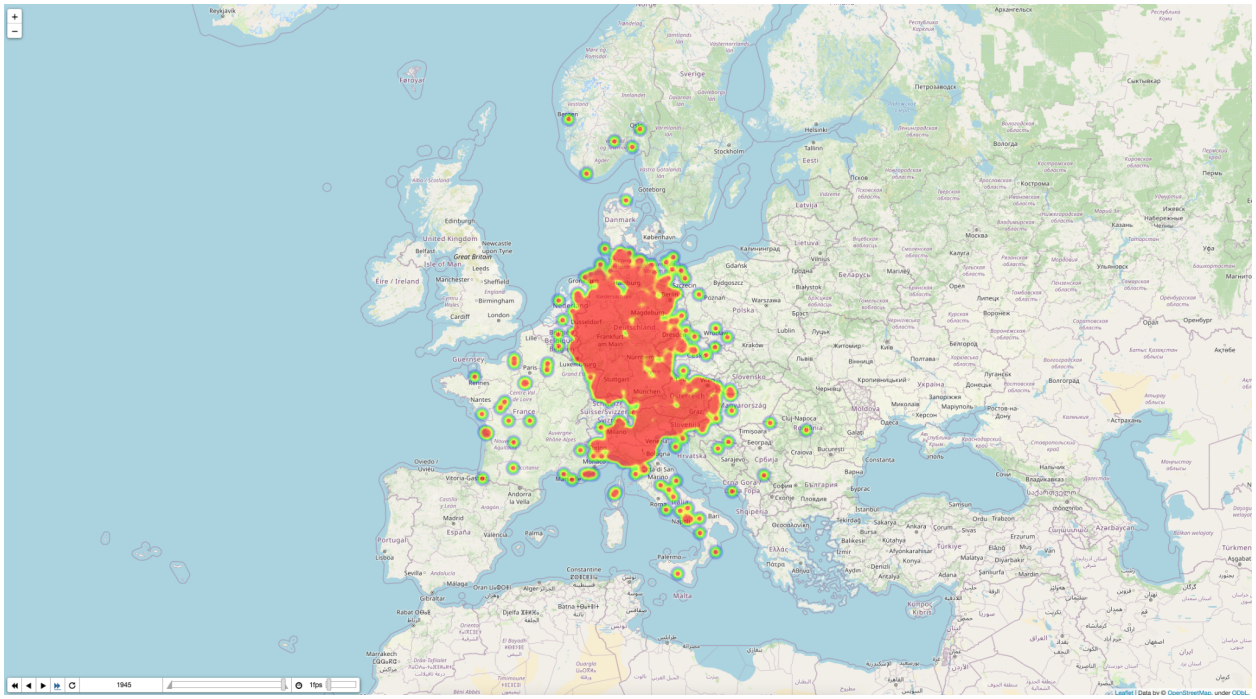

1942




1943




1944




1945

## Analysis

In terms of purpose/audience, this visualization would perhaps best be appreciated most by someone with existing background knowledge of WW2 in Europe and serve to reinforce that existing knowledge. For example, seeing bombing activity concentrated in Northern Germany/Belgium/Netherlands in 1940-1941 makes sense since that was the active area of the war during those years. Then in 1942-1943, you can see activity start to pick up in North Africa/Mediterranean/Italy which is where the allies first focused after the United States entered the war. It does look like the areas of Northern Germany/Belgium/Netherlands saw reduced pressure in 1942 compared to 1940-1941, which could potentially be explained due to resources now being used further south.

Then in 1944 you can see major activity throughout Northern France/Germany that likely coincided with the invasion of Normandy, as well as activity in Italy now being concentrated in the northern portion (as the Italian campaign progressed). Finally in 1945 we see bombing activity concentrated within Germany/Northern Italy/Austria which lines up with the Axis being pushed back to those borders by then.

Not that important, but interesting to look at, we can see neutral Switzerland in the middle of the sea of red in 1944, so that lines up with what one would expect there.

# DataSet 3: US and Syrian Civil War Wound Locations (Scientific)

## Source and Inspiration

So after browsing through some of the examples/information provided in context of the Scientific part of the assignment, I decided I wanted to try to do something with a PointCloud/Open3D. The human skeleton 3D model on the Artec 3D website seemed to be a good starting point, so I then thought about how to use that. I came up with the idea of trying to map location of wounds suffered in war to the skeleton model and was eventually able to find a couple tables holding information on that (but the quality/how I interpreted the data is questionable, which I'll discuss more) ~ that information seems hard to find. In either case, it did provide an exercise in the process. I ended up using the visualization to express where "surviving wounded" were wounded in the United States Civil War and Syrian Civil War.

Syrian Civil War Info: https://militaryhealth.bmj.com/content/jramc/166/4/261.full.pdf (Table 3)
US Civil War Info: https://achh.army.mil/history/book-wwii-woundblstcs-chapter5 (Table 63)
Human Skeleton Model: https://www.artec3d.com/3d-models/human-skeleton-hd

## Dataset Type

There were two Datasets used in this one. The data holding information on the war/wound information exists as a flat table, .csv file. I ended up making this dataset myself using information from the two references above.

The human skeleton 3D model was used in .PLY format. This format stores three-dimensional data (x, y, z) coordinates which is then used to construct a collection of polygons modeling the overall object. This would fall under the Geometry dataset type.

## Data Type

For the war/wound dataset, data exists in the form of items. Each row holds information representing the combination of war, location on the human body a wound occurred, and the relative survival percent associated with that war/location.

The human skeleton 3D model exists in the form of positions, where information exists to reference a position in 3D space.

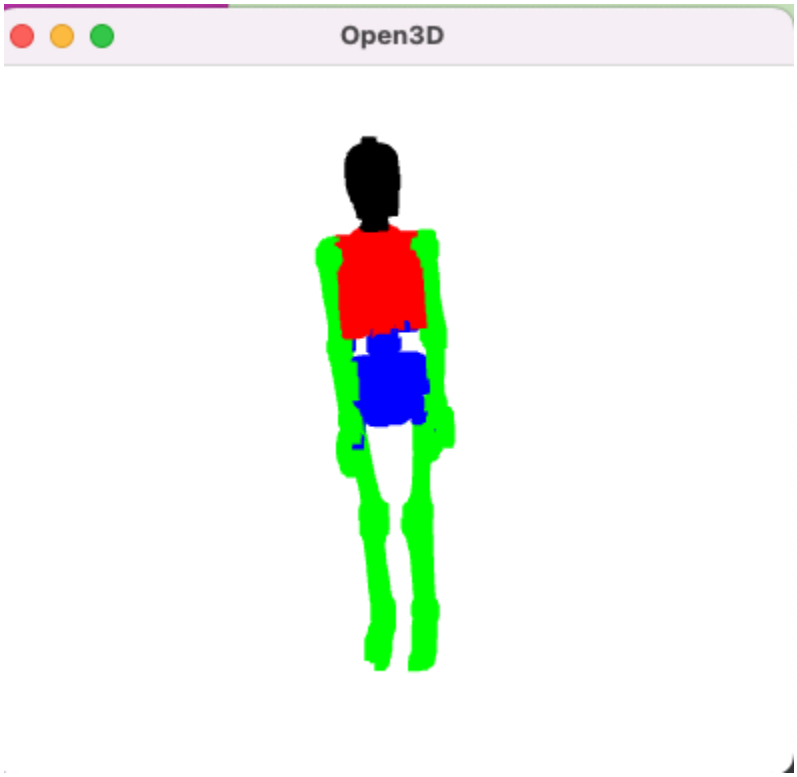
## Attribute Types and Semantics

For the war/wound dataset, I will discuss in the next section how I constructed it, but it contains the following attributes: *civil_war, wound_location*, *relative_survival_rate*, *relative_survival_rate_adjusted*. The *civil_war* attribute holds a string, which in this case is a categorical classification (United States or Syrian). The next attribute, *wound_location* also holds a categorical classification (head, chest, abdomen, or extremity). Finally the two *survival_rate* attributes hold floats that represent cumulative percentages relative to the group of items in each *civil_war*. For example, for a single item: (United States, extremity, 0.732) states that "Of surviving wounded in the United States Civil War, 73.2% suffered an injury to their arms or legs".

For the human skeleton 3D model, the data provides information to construct a visual 3D representation of a skeleton. While not stored in that .ply file in this context, I ended up associating the 4 *wound_location* categories mentioned above to their respective x, y, z points during runtime, which allowed me to color certain parts of the skeleton model in that context.

## Pre-Processing and Plotting

So for this part of the assignment I used Python, with Pandas, Numpy, and Open3D in particular. Unlike earlier DataSets, I started this one off with the exploration in the visualization step to make sure I would end up being able to do what I wanted to do. Using Open3D, I loaded the skeleton model and iterated through the points, using trial and error to eventually map the four regions of the human body to independent colors using the x, y, z coordinates, for example:




Is how I mapped head/chest/abdomen/extremities to locations within the 3D object. I left this code present in the relevant submission .py file for reference.

I mentioned above that I created the war/wound dataset. This was done by copying information from the following two tables referenced above into a single table.

TABLE 63.-*Comparison of wounds in living wounded of two past wars and World War II with casualties of Bougainville campaign, 15 February to 21 April 1944, inclusive, by anatomic location*

| Anatomic location | Living wounded | | | | | Bougainville campaign | Dead[1] |
|---|---|---|---|---|---|---|---|
| | Civil war | World War I | | World War II | | Living wounded[1] | |
| | | U.S. Army | British Army | U.S. Army | Russian Army | | |
| Head, face, neck | 9.1 | 11.4 | 16.8 | 16.1 | 9.1 | 20.7 | 49.0 |
| Chest | 11.7 | 3.6 | 7.8 | 9.8 | 11.4 | 12.4 | 29.6 |
| Abdomen | 6.0 | 3.4 | 4.7 | 5.6 | 6.2 | 5.7 | 16.3 |
| Upper extremities | 36.6 | 36.2 | 30.4 | 28.2 | 28.0 | 27.4 | .3 |
| Lower extremities | 36.6 | 45.4 | 40.3 | 40.3 | 45.3 | 33.8 | 4.8 |
| Total | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

**Table 3** Anatomical site of injury of patients traumatically injured in the Syrian Civil War according to Abbreviated Injury Scale[62]

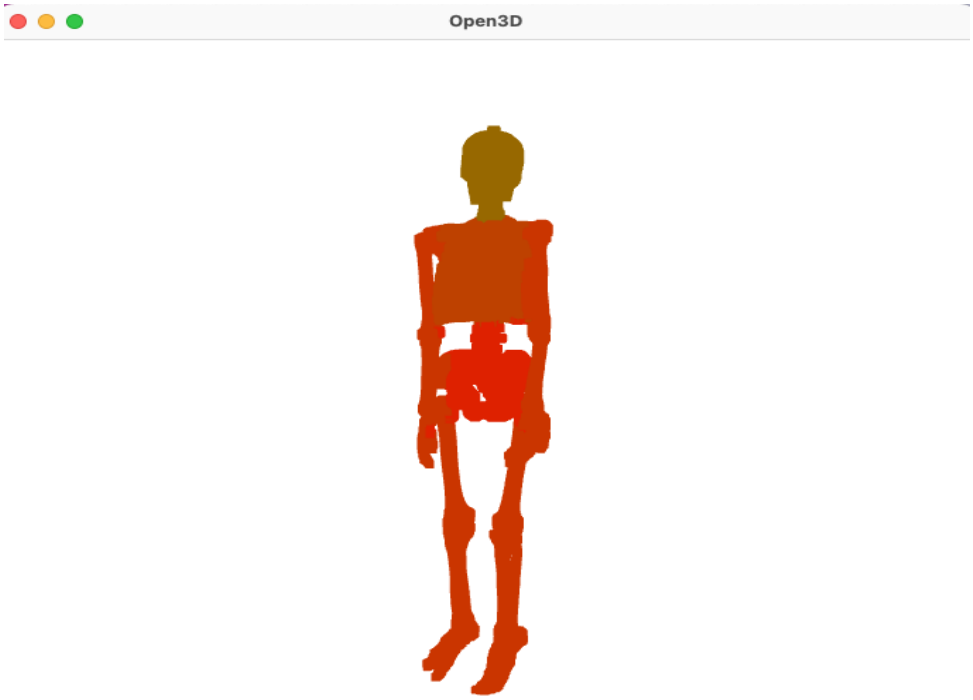
| Anatomical site of injury | | |
|---|---|---|
| Head | n | 719 |
| | % | 26.6 |
| Face | n | 240 |
| | % | 8.9 |
| Neck | n | 37 |
| | % | 1.4 |
| Chest | n | 613 |
| | % | 22.7 |
| Abdomen/pelvis | n | 316 |
| | % | 11.7 |
| Extremity | n | 508 |
| | % | 18.8 |
| Spine | n | 89 |
| | % | 3.3 |
| Other | n | 179 |
| | % | 6.6 |

In order to make the *wound_location* categories line up I ended doing a few transformations. Upper/Lower extremities US Civil War data was combined to a single item, and Head/Face/Neck combined for the Syrian Civil War. Since the Spine/Other categories of the Syrian dataset were not reflected in the US one and represented only ~10% of total, I used the existing weighting of defined categories to distribute that remaining 10% between them in order to remove those two categories while preserving the relationship between the ones I kept. This was all done on the fly in Google Sheets, with the resulting data saved to the project folder and represented what I term the “parse” step in the other datasets. The resulting table is as follows:

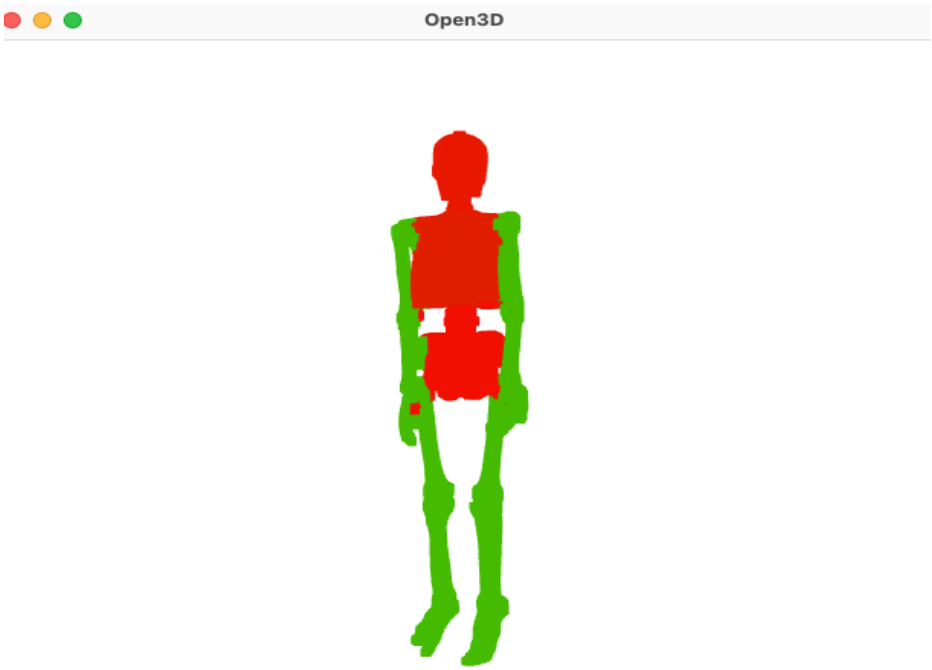
| | A | B | C | D |
|---|---|---|---|---|
| | civil_war | wound_location | relative_survival_rate_raw | relative_survival_rate_adjusted |
| | United States | head | 0.091 | 0.091 |
| | United States | chest | 0.117 | 0.117 |
| | United States | abdomen | 0.06 | 0.06 |
| | United States | extremity | 0.732 | 0.732 |
| | Syria | head | 0.369 | 0.4076 |
| | Syria | chest | 0.227 | 0.2567 |
| | Syria | abdomen | 0.117 | 0.1287 |
| | Syria | extremity | 0.188 | 0.207 |

To plot, both the war/wound table and the 3D skeleton model were loaded in, I ended up making the plot function run off the appropriate argument passed to determine which of the two civil wars to generate a visualization from. A quick helper function was written to convert *relative_survival_rate_adjusted* to a Red/Green scale RGB, then the 4 categories were assigned their respective RGB values. Red was assigned by 1.0 - wound_percent (so lower relative survival rate == more red), Green was wound_percent, and B as 0.0. The 3D skeleton was then colored and displayed.

Figure




Syrian Civil War




US Civil War

## Analysis

The figures are rather rough and don't really convey information that couldn't be seen in the associated table, but they do provide an example of the exercise of mapping spatial attributes in this way and do offer a few points of comparison that can be discussed. The viewer can see that of surviving wounded in the US Civil War, the majority had wounds to their extremities, while in the Syrian Civil War it appears that head wounds are potentially both more common and/or more survivable. In both wars there were low amounts of surviving wounded with abdomen wounds, suggesting that is a dangerous area to be wounded in. I made the assumption that all the wounded in the Syrian Civil War data were surviving wounded, which is important to mention, since they were tabulated in that way that they had at least survived to be offered medical treatment/data recorded in that way.

Another observation is that the more "uniform coloring" of the Syrian Civil War figure when compared to the US Civil War figure suggests that you really only had a decent chance of surviving a wound in the United States Civil War if it was to your extremities, vs. in the Syrian Civil War, the distribution is more balanced indicating that wounds to other parts of the body are more survivable now than back then. This potentially makes sense in the context of ~150 years of medical advances.

I considered changing how the color scheme was applied (making it ranked/discrete rather than continuous), but decided to leave it how it is since it does reflect the underlying data. But perhaps that could have been handled differently, since it is somewhat hard to distinguish some of the reds without close inspection.