Final Project Writeups

Walker Babich, Henry Grob, Julia Makela & Bradley Woodruff
IS590DV, Data Visualization, Fall 2017
December 8, 2017

Contact Information

Walker Babich	walkerbabich@gmail.com
Henry Grob	grob2@illinois.edu
Julia Makela	jpmakela@illinois.edu
Bradley Woodruff	bsw@illinois.edu

Component 1: Transportable Array Interactive

Reason for the Approach

Slider Component
The sliders individually index time and station, giving the user multiple dimensions in which to move around the data.

Map
Using a map to plot the locations of each station is logical as each station is accompanied by coordinates. The map provides a visualization of the transportable array in a manner that is understable, geographically.

Oscillogram
The decision to incorporate an oscillogram is based upon the first component's requirements, but also it is an easy to understand visualization.

Spectrogram
We chose a hexbin plot to represent the spectrogram, plotting the absolute value of the seismograph readings to give an impression of the total intensity, rather than show a hexbin

version of what is already shown in the oscillogram above it. We borrowed this idea from a ResearchGate[1] post on Spectrograms and Oscillograms.

Audio

Although the audio component works, we were unable to implement it into the figure that contains our map, line plot, and spectrogram. The audio component is included in the bottom of component 1 to add another dimension to the data, an audio format of the station's intensity reading.

Strengths of the Approach

This approach shows a multidimensional view of earthquake data: location, intensity, and movement. It takes a complex set of data and breaks it into digestible components.

Weaknesses of the Approach

Several weaknesses arose during the course of compiling component 1. One major weakness is the map component. Although we wanted to originally plot every station and update the color based on each respective station's intensity readings over time, we were unable and resorted to selecting a specific station, which would geographically appear on the map and the time slider would change the station (marker's) intensity. Another weakness is arranging and coordinating the multiple plots into one figure. The solution we found was to use matplotlib gridspec.

Additional Wishes

We thought it would have been useful to turn the time slider into a playable widget, where the user could set the slider where they like and click 'Play' to start the visualization. Additionally, the station slider could have more utility. As it is, you drag the slider and hope to land on the station you want. An option that lets the user pick a specific station would be more useful.

Contributions from the Group

Contribution	**Group Member(s)**
Primary coding	Walker
Additional contributions to code	Julia
Strategizing, review and feedback	Bradley, Henry, Julia, Walker

[1] Retrieved from https://www.researchgate.net/figure/267827408_fig2_Figure-2-Spectrograms-and-Oscillograms-This-is-an-oscillogram-and-spectrogram-of-the

Commenting and documentation	Walker

Component 2: Transportable Array Movie

Reason for the Approach

Our approach was to use Jupyter to create individual .png image files of seismic activity measured at each station at each point in time. The image files were then zipped and exported, so that we could use MEncoder to stitch them together into a movie, along with a simple cover slide that we made using Microsoft Paint to create an .png image file that was the same size and shape as our figures.

To provide context, the image files were set on a basemap that centers the focus on the continental United States. Stations were plotted by longitude and latitude. Color was used to indicate seismic activity at a particular point in time, with a diverging colormap selected to indicate both positive and negative values. The min and max on the color scale were set to include all altitude values in the dataset that are within two standard deviations of the mean. (This choice of outlier removal was made after observing a version of the video that had set the min and max color at the min and max of the altitude values. We found that outlier altitude values were expanding our color bars so far that it was difficult to see variations in color for most of the event. Changing to remove outliers made for a much more impactful video.) The neutral, central color, which shows up as grey in our colorbar, was set at "0" (or no change in altitude).

In our first attempt at creating the video, we ran our code for all 14,400 individual image files. Using an “industry standard” (personal communication, J. Makela, November 26, 2017) of 24 frames per second, this would lead to a long video -- 10 minutes. And, the processing time was also very lengthy (e.g., over 6 hours for only the first step of creating the images). After reviewing the initial images, we decided to focus on the single, most active hour in the dataset, which fell between 20 minutes and 80 minutes.

What we hope to for individuals to gain out of this visualization is an understanding of how the measurement of seismic activity moves across the distance of the United States over time. This is demonstrated by a “wave” of color that moves from west to east across the continent – moving from near the earthquake epicenter to further away.

Strengths of the Approach

The movie demonstrates movement over time in a structured space of the continental United States. The viewer is able to see both seismic activity and time variation, as required by the assignment.

Weaknesses of the Approach

Some information is lost in the complexity of this representation. For example, the station colors provide a sense of positive or negative seismic measurements, however, exact numbers are not available for individual stations in this presentation.

Additional Wishes

Looking at our final production, we are excited about what we were able to accomplish. However, one particular aspect that could use discussion and improvement is the way that the time is presented above the graph. Currently, it shows time elapsed in the following format: “0 days HH:MM:SS”. What is odd about our representation is that it uses the times directly from the file, and the video starts from 20 minutes, and ends at 80 minutes. That would not make much sense to a viewer outside of our IS590DV course. This was an easy shorthand for our programming. But, perhaps it would have made more sense to present the full date and time here? Or maybe to adjust the time elapsed to start at zero and run to 60 minutes, so that it would show an hour within the context of our visualization. So, these would be possible directions for improvement.

Contributions from the Group

Contribution	**Group Member(s)**
Primary coding	Julia
Additional contributions to code	Walker
Strategizing, review and feedback	Bradley, Henry, Julia, Walker
Commenting and documentation	Julia
Transition from Jupiter code into video	Julia

Component 3: UFO Database and Supplemental Data

Reason for the Approach

We wanted to visualize the number UFO sightings per state up against NIH data on gallons of alcohol consumed per year per capita since 1977. We decided to ask: what is the relationship between UFO sightings and alcohol consumption on a state-by-state basis?

To do this, we had to bring in a tertiary dataset. The alcohol data was already normalized by state population, so we need to “denormalized” the data so it would match the UFO data. Some things that we considered before we did this: census data is collected every 10 years, and the alcohol dataset was collected from 14-year-olds (14yo) and older. Due to time and data constraints, we found the census data for the year closest to the average, and we found the percentage of those 14 years old and older. Using that data, we found the percentage of 14yo’s and older for each state, and multiplied it by the gallons of alcohol consumed per capita for each state, resulting in the estimated total gallons of alcohol consumed, by state, for each year. Additional data cleaning was need to make sure that the year ranges for the UFO and alcohol data sets were the same.

Strengths of the Approach

This approach allowed us to visualize sightings for each state and a state selector which allows the user to choose a state and see graphs of total duration per year, and total sightings per year. Our additional alcohol data set allows the user to additionally visualize total gallons per capita per each state, and gallons per sighting per capita for each year for the selected state.

Weaknesses of the Approach

As mentioned above, we recognized a challenge with the additional alcohol data set that we chose - the data were already normalized as gallons of alcohol per capita. In order to normalize our UFO sightings by gallons of alcohol, we first needed to remove the per capita from the original data set. This meant finding a third dataset to integrate - U.S. census population statistics by state. In an ideal world, this would have been obtained for each year, and “denormalization” of the alcohol data would have been done by yearly population statistics. However, due to time limitations for this project, this was not possible. Instead, we used 1990 census data as an approximation in our data analyses. 1990 was selected because it was close to the mid-point of the dataset that we were analyzing. If we were able to include annual data, then this would allow us to provide a more accurate presentation of the total gallons of alcohol consumed. Note that for Component 4, when we present only US total numbers for a shortened period of time (1998 - 2014), we do complete analyses using annual population statistics, to address this weakness.

Additionally, the required time slider and total time vs total sightings slider is missing. We chose to omit these parts due to our completion of component 2.

Additional Wishes

If possible, we would include further interactive tooltips for the scatter plots. It would be interesting to be able to hover over a point and see to which year the point corresponded. With the additional population data, it would help create more accurate data and visualizations.

Contributions from the Group

Contribution	**Group Member(s)**
Identification of supplemental dataset	Julia, Henry
Formatting of supplemental dataset	Henry
Primary coding	Henry
Additional contributions to code	Walker
Strategizing, review and feedback	Bradley, Henry, Julia, Walker
Commenting and documentation	Henry

Component 4: Infographic

Reason for the Approach

We decided the UFO data combined with alcohol consumption would lend itself well to the storytelling style of an infographic. We found a moderate positive correlation between alcohol consumed and UFO reportings (r = .68 for across the US, with moderate correlations existing across states with a variety of patterns of alcohol consumption (e.g., consistently low consumption in comparison to national averages - WV, r = .67, decreasing alcohol consumption over time - CA, r = .44; broad variation in alcohol consumption per capita over time - NY, r = .56). While this does not mean there is a causal link between reporting UFO sightings and alcohol consumption, the infographic simply acknowledges a relationship in a playful way. In designing this infographic, we were thinking about creating an eye-catching, and perhaps a bit off-the-wall, public service message -- one that would be memorable, perhaps in the way that the Shark Attack example shared in class stuck with us.

A particularly interesting aspect to the analysis that we conducted as we discovered the story for this infographic has do with with how we came to understand the content of the UFO dataset.

We noticed that, while UFO sightings are reported to have occurred as early as 1906, the actual year when these sightings were recorded in the UFO database begins in 1998. Starting in 1998, there is a large increase in the number of reported UFO sightings across the United States recorded in this dataset. The increase is so large that it seems to overwhelm analyses that attempt to cover both pre-1998 and post-1998 time periods. As we came to understand the dataset, we connected this to what we know of social science research -- people are much better at recalling and describing current and recent experiences than the are at recalling experiences in the distant past. Perhaps one contribution to the considerable change in number of reports at this time break is that individuals simply do not recall all instances of UFO sightings in the past, before the database was available to record their reports. For that reason, we limited our analysis to 1998 and forward in order to be more confident that we were not hindered by the "noise" of memory gaps in our data. When we do this, the relationship between UFO sightings and alcohol consumption (which we hypothesized may exist) emerged much more clearly from our dataset.

Strengths of the Approach

We have come up with a playful way to use two disparate datasets to present a single, memorable message in a unified and succinct way. With a few more iterations, additional time, and feedback from target audiences, perhaps a message like this could evolve into a tongue-in-cheek public service message that has a foundation in data and communicates an important idea, but does so in a way that is memorable and gets people talking. There is a large amount of data behind this infographic, and it took several iterations to come to this story. But, we believe we achieved the goals of a simple, clear message that is easy to interpret and will stay with our audience.

Weaknesses of the Approach

It was difficult to translate sightings per person into numbers that would make sense to an average person looking at our graph. What does 0.00001115 UFO sightings per capita even mean? So, we left those numbers off. It looks better, and the exact amounts wouldn't add much to the story, but we've lost something. If we were really interested in bringing this infographic to a broader audience, it would be important to test our ideas with some target members of that audience. We would need to find some ways to represent our complicated X and Y scale values in a way that would make sense to our audience.

Additional Wishes

We had the choice to use the graphics Canva had to offer (fast and easy), or create the images we had in mind and upload them (time consuming). In the end, we opted to use the design elements available in Canva even though they were different from our original concept. Given infinite time, we would have designed a more custom look.

We also found this correlation holds across states and it would be fun to present that data in another format, but that data is probably better suited for an interactive display.

Contributions from Group

Contribution	**Group Member(s)**
Brainstorming ideas	Julia, Bradley
Research	Julia, Henry
Data analysis	Henry, Julia
Storyboarding / Messaging	Bradley, Julia
Layout / Design	Bradley
Review and feedback	Bradley, Henry, Julia, Walker
Documentation	Bradley