INFO 3300
Project 1
Albina Chowdhury, Cesar Cisneros, Yunjie Liu
October 6th, 2022

Project 1 Final Report

Final Screenshots of Visualizations

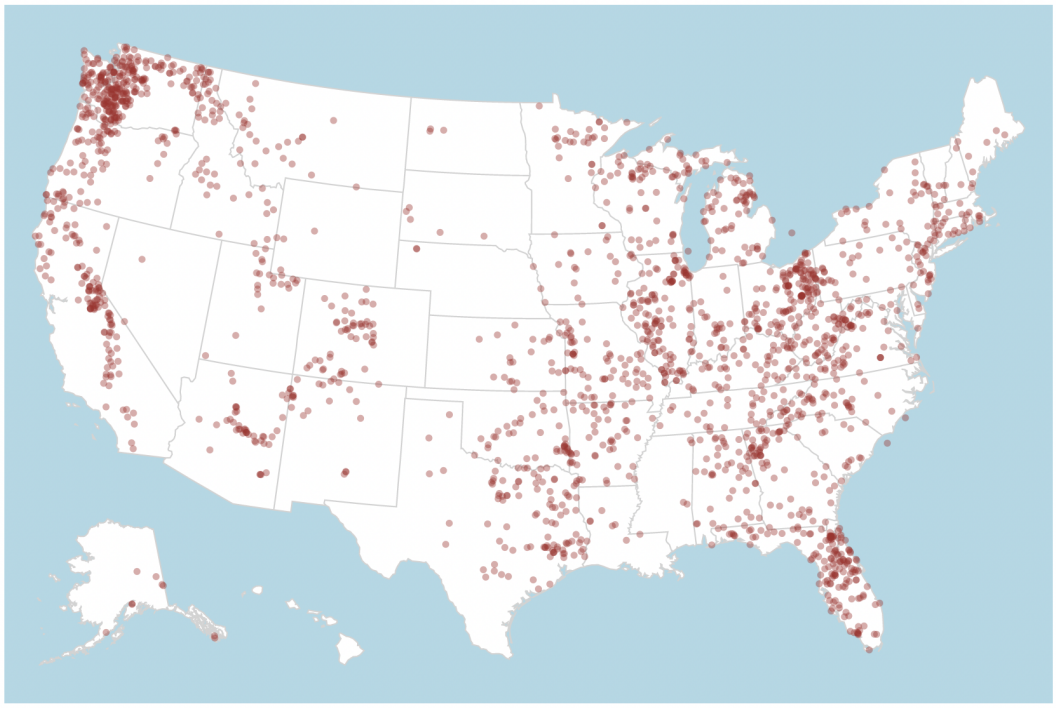

Figure 1: A U.S. Map With Circles Representing Bigfoot Sightings

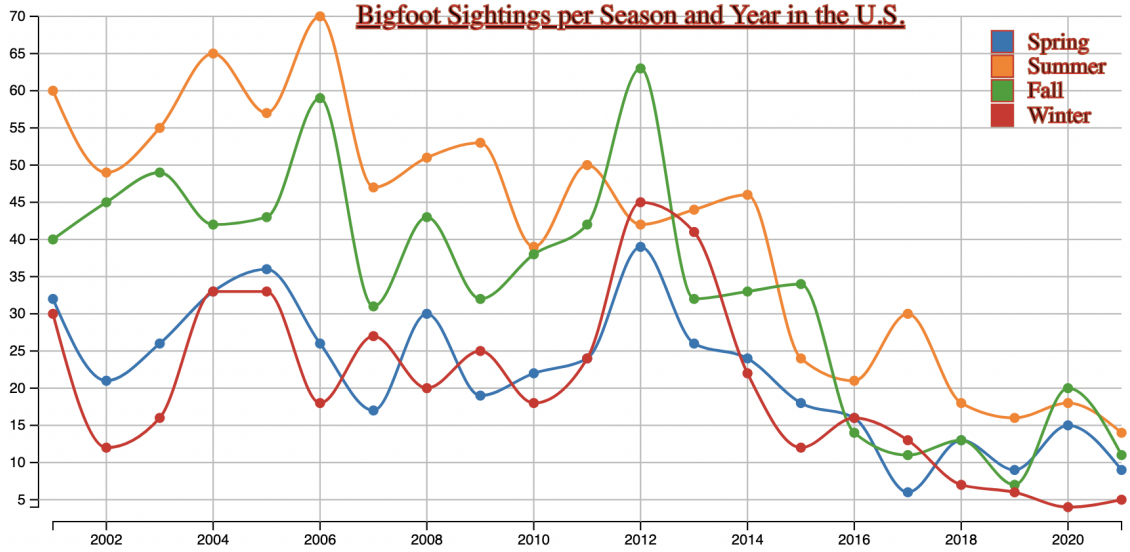

Figure 2: A Line Chart Visualizing the Relationship Between Year, Season, and Frequencies of Bigfoot Sightings

Description of the Data

The data that we are using consists of three files: bfro_report_locations.csv, bfro_reports.json, and bfro_reports_geocoded.csv from the source *data.world*. The columns we used from these files are the year bigfoot was sighted (column year), the latitude and longitude of where he was sighted (columns lat and long), and the season in which he was sighted (column season). Once we filtered out these specific columns, we were able to clean the data further within these columns in order for us to use them for our graphs. Using SQL, we were able to get rid of null or NaN values as well as filter out all the data prior to 2000 so that we could focus on more relevant data in the 21st century. We constructed a new column called "freq," which counted the frequency of bigfoot sightings for each season given a year. Additionally, we did an inner join with the reports CSV and the locations CSV in order to retrieve the longitude and latitude for each sighting on the specific report number (the primary key). Finally, we created two CSV files, one for each graph: one that consisted of the year, season, and freq columns and the other that consisted of the long and lat columns. For the line graph that we created, we had to convert the CSV file into a JSON file.

Design Rationale

After looking through the data, the first thing that came to our mind was making a map which has data as circles positioned at a certain Latitude and Longitude, using the columns lat and long provided. In the end product, we want to convey useful information such as where the most bigfoot sightings happen, which is the region with the most circles. Since our dataset primarily

focuses on bigfoot sightings in the U.S., we made sure to use a USA-specific projection so that the circles are better positioned on our map to achieve the results we want. We also made sure to use contrasting colors separating the outline of the map, the ocean, and the circles. Specifically, to emphasize the circles, we used a lighter gray for the outline and a darker brown for the circles. The mark we used is circles and the visual channels we used are positions on the map.

We also wanted to make a line graph displaying the frequency of bigfoot sightings in relation to year, where we will have four lines, each representing a season of a specific year. In this line graph, the x-axis will be the year, and the y-axis will be the frequency. Although bar charts will be more straightforward in comparing seasons in a particular year, we decided to use a line graph because we want to track changes and trends over periods of times for multiple groups, in this case seasons. Each line will have a distinct color so that the audience can easily separate them. The circles, representing each data point, will also have the same color as the line. The marks we used here are lines and circles while the visual channels are colors and vertical positions.

The Story

As previously mentioned, our goal with this project was twofold. First, we wanted to delve into the locations of these sightings [per frequency] as a way to gain a better understanding of the past history of this folklore legend. Secondly, we were interested in understanding the frequency of bigfoot sightings per season and year in the U.S.

Overall, our map demonstrates that the east coast has more scattered circles while the west coast has more concentrated areas of sightings. For example, there is an apparent cluster in the northwest of the U.S., particularly in Washington. This can be attributed to the numerous parks and forests in the state in comparison to other states. Also, the cluster in California appeared to be along the Sacramento Valley. By contrast, the circles on the east coast pretty much appear everywhere. Although there are some clusters in Ohio and Florida, there aren't as many circles as in the cluster in Washington. In addition, much of the midwest is void of any sightings, which isn't surprising considering its geographic features like flat plains. Finally, what's also interesting is that there are actually some circles in Alaska. We wonder how bigfoot could've survived in such harsh conditions.

Our season line plot highlights summer as the season with the most sightings, followed by fall. This is not surprising because we would expect bigfoot to be the most active in warmer seasons compared to cooler ones. However, we also noticed that the sighting frequencies in spring and winter are generally close, which contradicts our prior assumption that there might be more sightings in the warmer spring. Overall, there appears to be a decline in sightings over the years, albeit still maintaining the same trends observed per season.

Team Contributions

This group project has a few phases: idea brainstorming, data processing, and construction of data visualizations. All three members of the group contributed to the brainstorming of ideas as well as finding relevant datasets. We all came together and evaluated the datasets and thought about the potential visualizations we can do for each of them. Unanimously, we decided to use the bigfoot dataset because of the data provided by the CSV and JSON files and because it was the most interesting to us.

In particular, Cesar helped process the dataset used for the line graph while Albina and Yunjie processed the data for the map. Cesar and Albina constructed the line graph while Yunjie made the map with the help of Cesar and Albina. All three members helped each other with debugging and went to office hours when additional help was needed.

Overall, we spent around 2 hours brainstorming, 2 hours finalizing ideas, an hour cleaning data, and around 5 hours coding the data into visualizations. The coding took the most time because we had issues with the formatting of the CSV file – which we later converted to a JSON file – when trying to plot the line graph, and there were issues with data point scaling for the map. In total, to complete this project, we spent a total of around 10 hours.