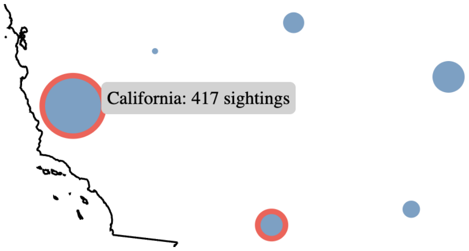
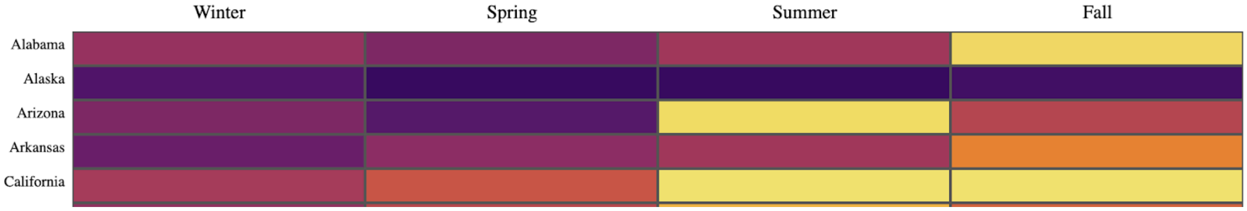
Team 11 - DSCI 550 Assignment #3 Report

1. How are your 5 D3 visualizations answering and showing off your features from assignments 1 and 2 and the work you did?

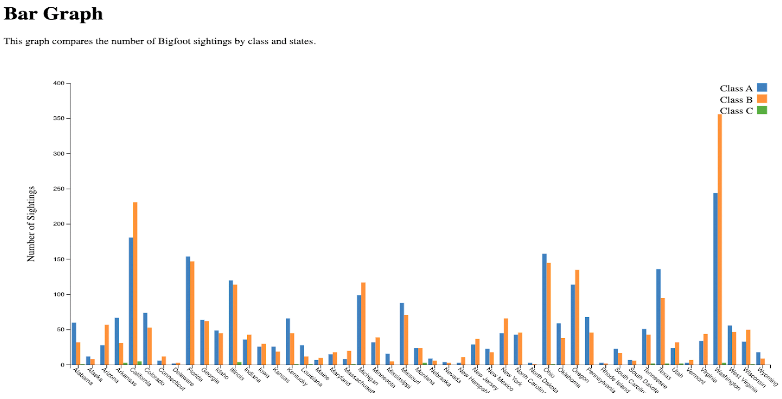
The first D3 visualization we chose was a Bubble Map of the United States designed to showcase the frequency of sightings per state, with a red border around the Bigfoot Hotspot states (as seen below). We thought it was a good idea to showcase an overall visual of where the sightings occurred, with the states with many more sightings having larger bubbles. We were able to leverage our Bigfoot Hotspot feature from Assignment 1, along with the lat and long coordinates, to provide an informative visual of our data.



We also designed an interactive heatmap of each state's total number of witnesses by season. This is looking at the reports with multiple witnesses, utilizing the (hard-earned) Witness Count feature that we added back in Assignment 1. This visualization was to see similar features with multiple witness reports. In Assignment 1, we noticed summer and fall as the most popular seasons with reports of multiple witnesses, so this helps us visually see that by the colors, the lighter colors have higher witness counts. Below, we have a small snippet of the visualization, where one can then see just that:

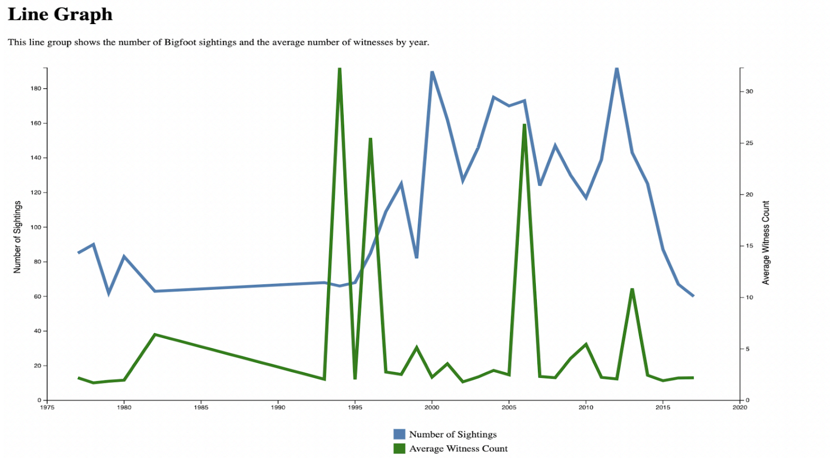


The third visualization we chose was a grouped bar chart to show the distribution of different classes of sightings between different states. With this graph, we wanted to see which regions of the US may have more or less 'trustworthy' reports, leveraging the location data we obtained from the previous assignments.


Bar Graph

This graph compares the number of Bigfoot sightings by class and states.

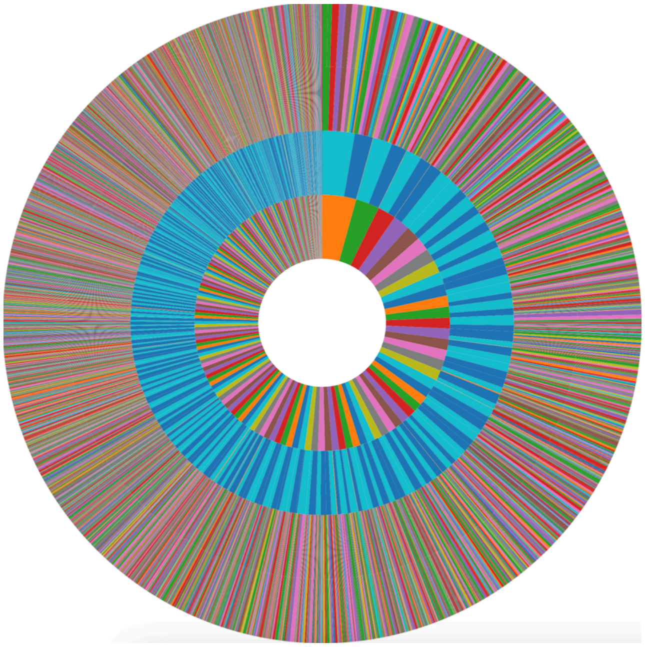
The fourth visualization we chose was a line graph showing trends in sighting and witness counts over time. Using witness counts found from assignment one and pure sighting numbers, we wanted to measure the distribution in these statistics over time to highlight any patterns or fluctuations within our data.



Line Graph

This line group shows the number of Bigfoot sightings and the average number of witnesses by year.

Lastly, we chose an interactive sunburst plot to explore a range of features related to Bigfoot sightings. These features include the closest parks, report classification, month, and severity of weather conditions. The sightings are ranked by frequency, from most to least frequent. By leveraging these features, we established connections to the nearest national parks for each sighting and integrated weather data from our assignment. This visualization allows us to delve into various feature levels, enabling a deeper analysis of our findings from previous assignments. A sunburst chart is a powerful visualization tool representing hierarchical datasets using nested rings. We thought it was an intuitive way to analyze relationships and proportions between categories and subcategories, making it valuable for data analysis.



2. Did Image Space allow you to find any similarity between the BFRO generated Bigfoot images that previously were not easily discernible?

ImageSpace does help in finding similarities between images. To process the similarity between sightings, we would have to analyze the text and features of each report. Using ImageSpace, one can discern how common certain features are using search functionality as well as SMQTK similarity. It needs to be clarified how much fidelity there is between the prompt text and the images. For example, how often does the scene describe snow and how often is that then depicted in the images needs to be clarified. Using the search functionality, searching for the word snow produces 323 images. Whereas the word, "snow" only appears in 225 observations. However, it is possible that words like "skiing," "snow-mobile," and "winter," all common words used to describe snow sightings, can influence the images produced; this is still useful as it gives you a sense of the sightings visually.

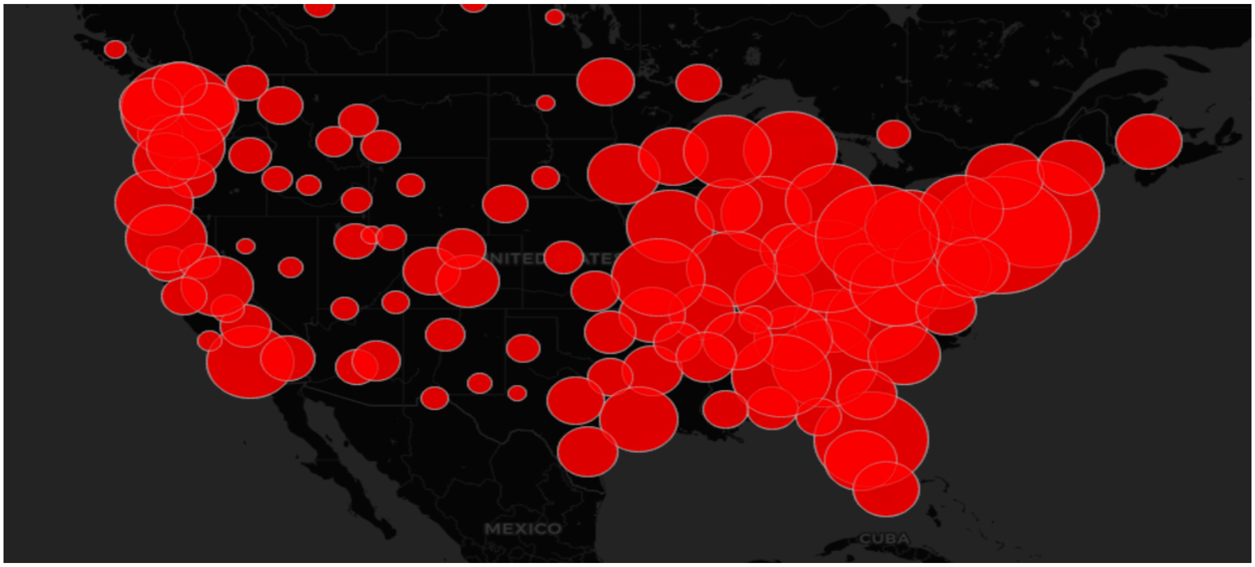
When testing out SMQTK similarity, for example, images with certain key features had highly consistent results. Images with bodies of water surrounded by trees or vegetation were very consistent with SMQTK similarity. The results were just as consistent when completing similarity searches on images with roads. Words like "road" or "highway" were frequently used to describe encounters. Therefore, many roadside scenes were created by the image generator. Our image collection has many depictions of an isolated trail, road, or path in the foreground or background. Occasionally, the search would confuse images that depict a road or a highway with one depicting a trail. The distinction being that many of these trails represent a grass or a dirt walking path. Likewise, for images with snow. The similarity search had results where 85-78% of those images had snow in the picture. However, many green nature scenes were also included.

The similarity search performed worse with images with uncommon elements in our collection. For example, in one instance, the image generator created a landscape skyline view of a city at sunset. Most bigfoot sightings occurred in isolated and wooded areas. Therefore, city skylines are primarily absent from our collection. However, the skyline image was, because of the sunset, warmer and slightly ethereal in its lighting and color palette. Thus, the similarity search found other images that had a warmer color palette and more ethereal lighting, but were not cityscapes. The SMQTK similarity search functionality seems to be able to find some set of common conditions among most images, even if not the main object of the image.

Many of our sightings are from the southeast United States, particularly in parts of Texas, Alabama, and Florida, where the environment around many of these sightings is swamp-like. One image clearly depicts a swamp region with a slight morning fog, a boat, and trees rising above the swamp water. In this instance, the first two "similar" images are of a person seeing a figure through a car windshield and a person walking into a shack just beyond some trees. I would say that the most similar aspect of these three images is their color palate and not their content.

3. What type of location data showed up in your data? Any correlations not previously seen?

GeoParser was very easy to use using Docker pull. However, there were no ways to refine the results that the GeoParser was identifying in the BFRO dataset using this method. This led to a large number of results (close to 1300) outside of North America, where all of the reports are known to originate from. Of the results that did populate in North America, we noticed many reports originating from the eastern US, which corresponds with heavily forested areas, confirmed with the vegetation index calculated from Assignment 1. Then, there is a drop in reports to the west of the Mississippi River to the Rocky Mountains, which again makes sense, where there are more open plain areas. Finally, there were many reports along the Western US coast, with a concentration in the Pacific Northwest Region, where it is again very heavily wooded. The GeoParser tool allowed us to inspect the number of reports visually and where they were coming from, which helped put the entire dataset into perspective. This type of location data can be used to conclude many things that are not readily apparent when only given access to the printout of the location's latitude and longitude. It allows for much easier and faster analysis if one overlays the heat map that the GeoParser displays with satellite images of topography and vegetation. Recommendations to make GeoParser better would be the ability to import satellite images in the interface effortlessly and import maps showing major roads or towns instead of the default black landmasses.



Also include your thoughts about Image Space and ImageCat – what was easy about using them? What wasn't?

ImageSpace and ImageCat proved challenging to set up initially. After much troubleshooting on Slack, Augusto Partido helped us find our solution; the script "enable-imagespace.sh" was modified and updated in the local repository. We were able to update the "docker-compose.yml" on our own successfully. Running it on a Silicon-based Mac, it took a while to realize that the default platform in Docker had to be changed to Rosetta for access to "linux/amd64" both on the Docker Desktop app and through the command line. Users who were running Intel-based Macs did not run into this issue. The rest of the setup was simple to execute according to the quick start guide on the wiki. The process that took the longest was the second line of code in the wiki that deploys the ImageSpace script and gains access to all 5467 generated BigFoot images. This process took about 21 min on a Silicon-based M2 with 16GB (8GB allocated to Docker), while it took over an hour and a half on an Intel-based Mac with 16GB (10GB allocated to Docker).

Once ImageSpace with ImageCat was deployed, it was easy to query through our images. However, it was evident that our query results were only as accurate as the metadata from which our images were generated. For example, when querying for a "cat," many images did not contain a feline, but the metadata mentioned cats. This finding and similar findings suggest that a change may be necessary in our parameters of image generation or even further refinement of the prompt given to the image generator. SMQTK similarity was somewhat successful but only showed the first 20 results. We could not determine how to change this within the source code. Overall, ImageSpace and ImageCat were intuitive and straightforward after much troubleshooting on setup and deployment. However, using it uncovered and reconfirmed an underlying issue of our image generation - it is not as true to its given caption. Further refinement in this portion of the project would be further explored if time allowed.