DSCI550 HW1: Analysis of the Bigfoot Research Organization (BFRO) Sasquatch Reporting Database

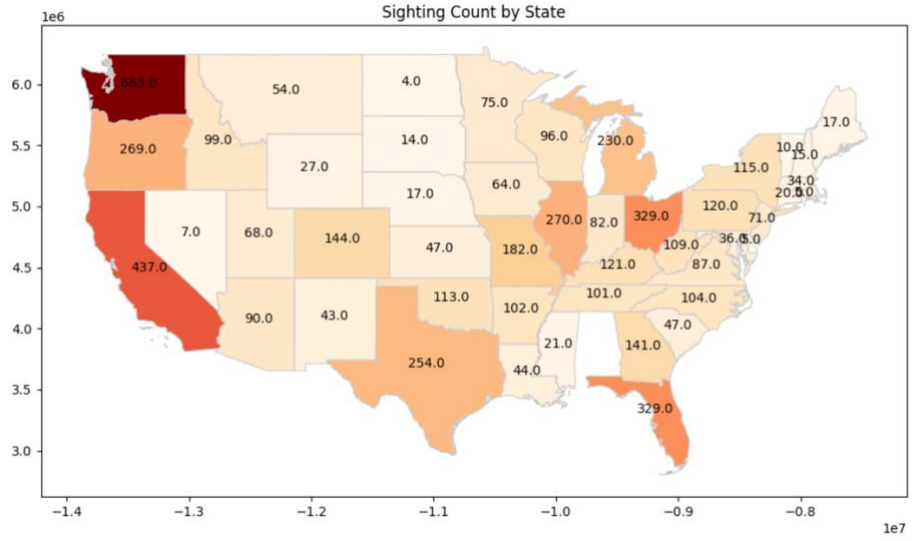
Siqi Xiao 2711223438, Nora Alwadaah 9284989783

1. Introduction

This report encapsulates a comprehensive analysis that merges the BFRO Sasquatch Reporting Database with three distinct datasets: SRTM Digital Elevation Data, Animal Distribution Data, and UFO Sighting Dataset. Our aim is to discern patterns and correlations that might offer insights into the reported sightings of the elusive creature known as Bigfoot.

2. Analysis and Findings

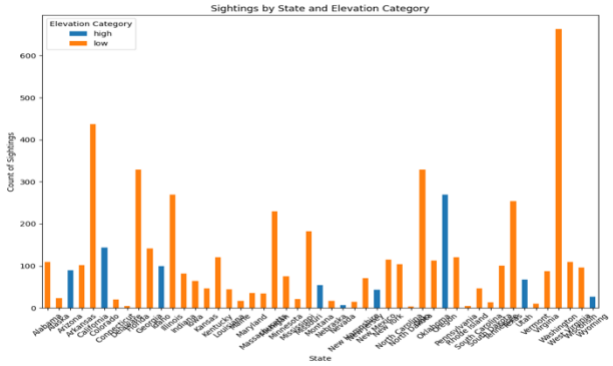
The map figure depicting state-wise counts of Bigfoot sightings provides an essential visual tool for understanding the geographic distribution of reported encounters throughout the United States. Each state is marked with the number of sightings, revealing patterns that suggest higher frequencies in certain areas, potentially correlated with extensive forested landscapes, demographic concentrations, and rich socio-cultural folklore surrounding the Bigfoot legend. Recognizing the significance of environmental, demographic, and socio-cultural factors in influencing the likelihood of reporting Bigfoot sightings, we meticulously selected our three datasets—land elevation, animal distribution, and UFO sightings—to comprehensively address these aspects in our subsequent analyses. This map not only frames our initial understanding of the distribution of sightings but also strategically guides the direction of our study, underpinning the analysis sections that follow in the report.

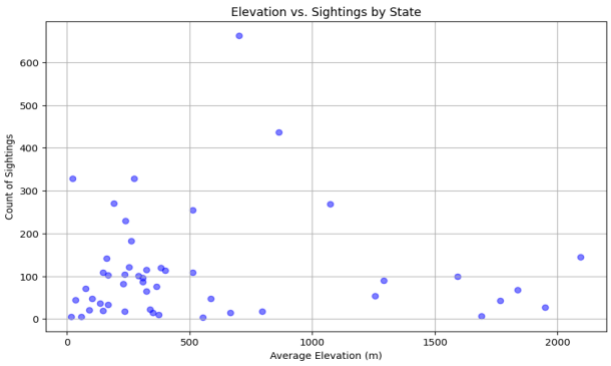


A. Correlation Between Bigfoot Sightings and Land Elevation in the United States

The first dataset we joined is the Shuttle Radar Topography Mission (SRTM) Digital Elevation Data. This dataset is offered by NASA, and is considered as of Image MIME type. It was originally produced to provide consistent, high-quality elevation data at near global scope. By calling the provided functions, users can get elevation information around U.S.

We calculated the average elevation of every state in U.S. to explore whether there is a relationship or a pattern between average elevation and number of bigfoot sightings, from a topography perspective.Two figures below illustrate the found relationship between elevation and number of sighting reported. The left figure below reflect how many sighting was reported through the years within each state in U.S. The right figure uses (average elevation, count sighting of bigfoot) as the coordinate, positioning every state in the coordinate. To reflect the pattern better, bars are colored in orange if the state is of low elevation, and in blue if is of high elevation (using 1000m as the threshold) in the left figure. From the figure, an observation can be that almost all of the states with high reporting number are of low elevation, indicating a feature of sighting -- more common in low elevation place.

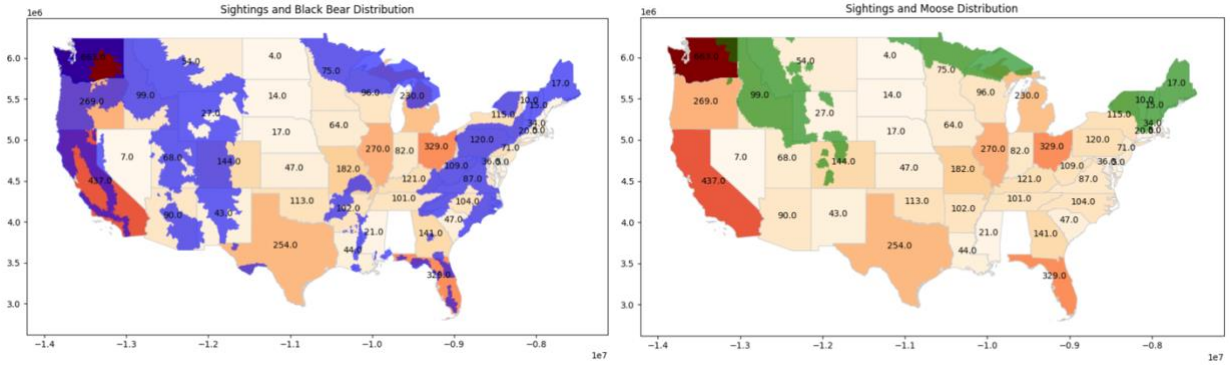




B. Correlation Between Bigfoot Sightings and Possible Animal Misidentification in the United States

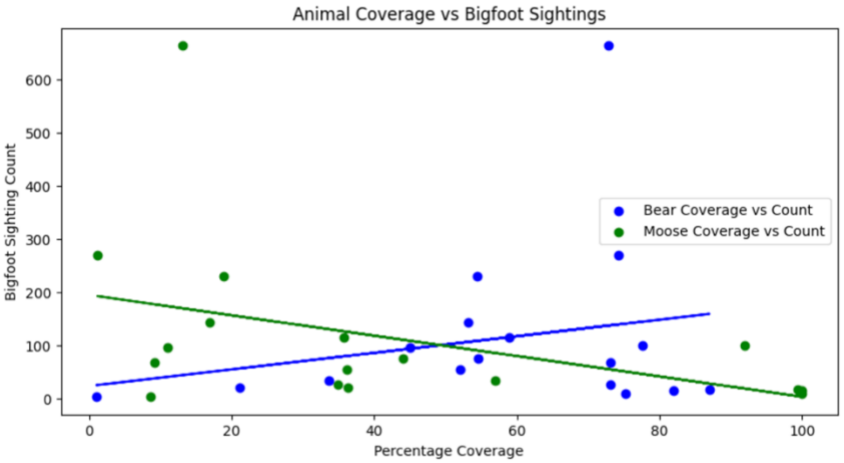
As part of our analysis, we explored the potential for misidentification of Bigfoot based on the presence of black bears and moose in various U.S. states. By cross-referencing the number of Bigfoot sightings with animal distribution data, we aimed to uncover any significant correlations that might suggest sightings of Bigfoot could, in some cases, be encounters with these animals instead. Our visual analysis consisted of two parts:

1. A **heatmap** which depicted the percentage of habitat coverage by black bears and moose across the U.S. The colors ranged from light to dark, signifying an increase in the likelihood of these animals' presence in each state. The categorization of confusion levels was as follows: states shaded lighter indicated a lower probability ('Low' <20%), medium shades indicated a 'Medium' probability (<40%), darker shades indicated a 'High' probability (<60%), and the darkest shades indicated a 'Very High' probability (>60%) of habitat coverage.



2. **Scatter plots** with regression lines provided a more precise look at the relationship between animal coverage and Bigfoot sightings:

- **Bear Coverage vs Sightings**: Blue points represents each state, plotting the bear habitat coverage against reported sightings of Bigfoot. The accompanying blue regression line shows moderate positive correlation
- **Moose Coverage vs Sightings**: In contrast, the green points plotted moose habitat coverage against Bigfoot sightings, with the green regression line indicating a mild negative correlation.



The visualizations show a trend where increased bear habitat correlates with more Bigfoot sightings, hinting at possible misidentification, but the spread of the data points also suggests other contributing factors of sighting reports. Conversely, regions with more moose habitat see fewer sightings, potentially due to better animal recognition. This suggests a moderate relationship, meriting further study into other influencing factors.

C. Correlation Between Bigfoot Sightings and UFO Sightings in the United States

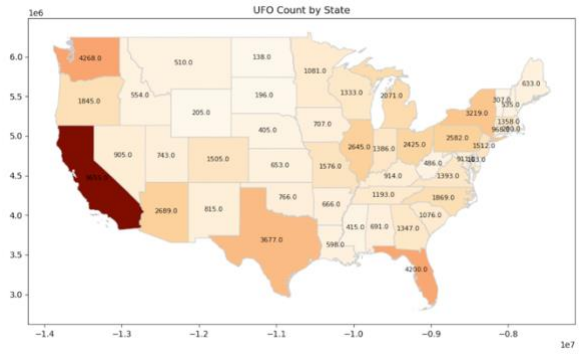
The third dataset we chose to merge with is the UFO Sighting Dataset, by mapping the number of report within each state in U.S. to the map, a stunning similarity between the map for Bigfoot sighting and the map for UFO sighting appeared. California, Washington, and Florida lead the rank in both situations, and the distributions of reports in both maps seems to be highly correlated. Several factors may contribute to this trend:

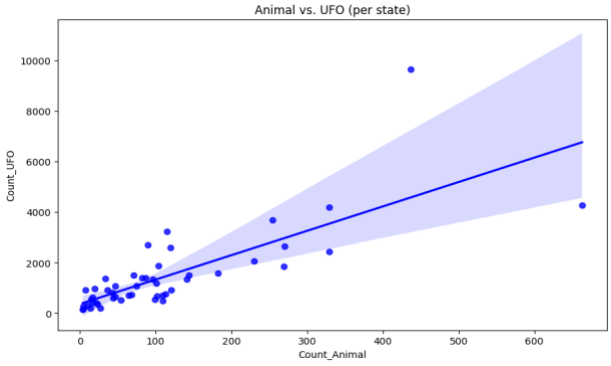
Environmental Influences: The diverse landscapes in these states, ranging from urban areas to remote wilderness, create conducive conditions for sightings. For example California, owns a vast and varied landscape, including urban areas, deserts, and mountainous regions, may contribute to a higher number of reported UFO sightings and bigfot sightings. Similarly, the dense forests and wilderness areas of Washington, while Florida's swamps and forests offer a similarly secluded environment for such reports.

Population and Tourism: The large populations and high tourist activity increase the likelihood of people observing and reporting unusual occurrences. California's and Florida's large population and status as a tourist destination may result in more people being outdoors and potentially observing unusual phenomena.

Cultural Impact: A rich cultural attraction with the paranormal, bolstered by local myths and media, may encourage reporting behaviors. For instance, the Pacific Northwest, including Washington, has a rich tradition of Bigfoot lore, which may influence the number of reports. Also, contributed by Confirmation Bias, meaning people in areas with a history of sightings may be more inclined to interpret ambiguous observations as aligned with local lore (UFOs in California, Bigfoot in Washington and Florida), leading to a self-reinforcing cycle of reports.

The relationship between UFO and Bigfoot sightings is underscored by a linear regression analysis, revealing that geographical and cultural dynamics significantly influence the frequency of reports in certain regions. This suggests a possible link between the tendency to report unexplained events and regional characteristics.





E. Tika-Similarity Analysis Overview

The Tika-similarity analysis across four progressive runs utilized Jaccard, Cosine, and Edit Distance measures to analyze the textual data from the Bigfoot sighting records. Below is the summary of runs conducted:

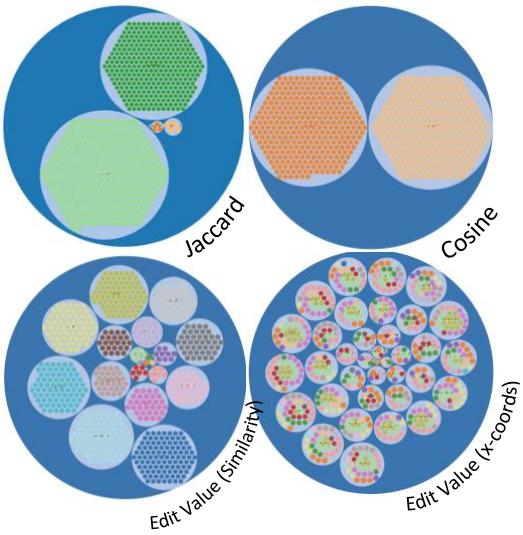
1. **Initial Run (500 Records)**: Analyze small sample within single directory - textual consistency baseline.
2. **Medium Run (2,000 Records)**: Examined a larger set across multiple directories, confirming diversity without significant repetition. Organized into five directories, each containing up to ten JSON files with 40 records each.
3. **Full Dataset Analysis (First Method - 5,000+ Records):** A thorough scan of the entire dataset in larger chunks, detecting common themes and possible frequent sighting hotspots.
4. **Full Dataset Analysis (Second Method - 5,000+ Records):** Dissected the dataset into smaller segments, revealing a broader variance in similarity scores, indicative of unique sighting details.

In the first run with 500 records, the Tika-similarity metrics likely revealed a high level of homogeneity with tight clustering, which could be expected in a smaller, more controlled sample. This initial dataset likely contained less variation, possibly due to a narrow range of sources or a limited geographical spread in sightings.

However, with the second run of 2000 records, the similarity analysis suggested an increase in variability. As the dataset expanded, the clusters became more dispersed across the similarity measures. This can be interpreted as a reflection of the broader diversity of Bigfoot sighting reports when scaling up the sample size. More distinct narratives and descriptors likely began to emerge, illustrating the richer complexity of a larger dataset.

In the detailed runs involving over 5000 records, two approaches were applied:

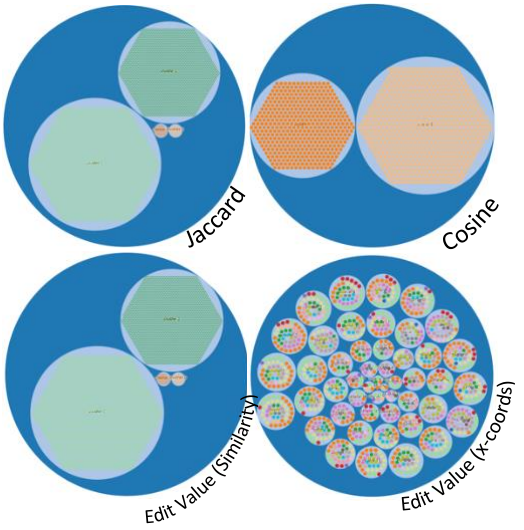
Third run, divided into three directories, with each holding up to 15 JSON files of 150 records each showing distinct patterns in the data:



- **Jaccard Similarity**: Reveals two clear clusters, indicating groups of reports with similar descriptions, suggesting either shared sighting elements or common reporting language.
- **Cosine Similarity**: Displays well-defined clusters, pointing to recurrent phrasing and keywords within the sighting reports, hinting at a common descriptive language/vocabulary.
- **Edit Distance Using Similarity**: Exhibits a diverse range of similarity, from tight clusters indicating shared narratives to isolated points that denote unique reports.
- **Edit Distance With X-Coordinates**: Offers a spatial view that suggests geographical influences on the similarity of reports, with wider clustering indicating regional reporting variations.

This run highlighted the dataset's multifaceted nature, with a mix of shared descriptions and unique reports, affirming the richness of the sighting narratives.

Fourth run, structured two directories, each containing up to 100 JSON files with smaller chunks of 50 records, we observed:



- **Jaccard Similarity**: Spread of clusters reflect diversity of text, indicating unique and varied sighting reports.
- **Cosine Similarity**: Strong clustering patterns, a high degree of narrative similarity within smaller data segments.
- **Edit Distance Using Similarity**: A range of clustering, with tight groups signifying narrative similarities and isolated points hinting at unique reports.
- **Edit Distance With X-Coordinates**: Cluster placements suggesting regional influences on sighting reports, with the potential impact of local characteristics on descriptions.

Tight clusters showed common descriptors, and dispersed groupings reflected the unique nature of individual sightings. Considering both similarity and spatial factors, further suggest a potential influence of geographical contexts on the sighting reports.

Removing nearly identical or repetitive textual columns could potentially reveal more interesting analysis; however, due to time constraints, this exploration remains for future work.

Thoughts on Tika:

Using Tika for text extraction was straightforward, but applying Tika similarity for analysis presented challenges. Although it's powerful for comparing document similarities, the resulting graphs were not user-friendly and hard to interpret, requiring extensive research to make sense of the data.