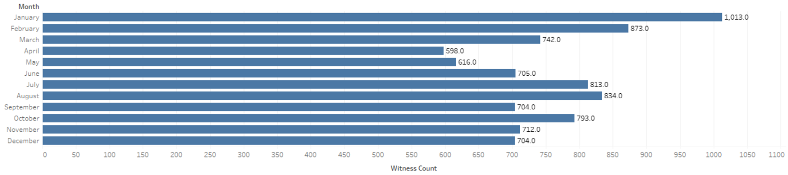
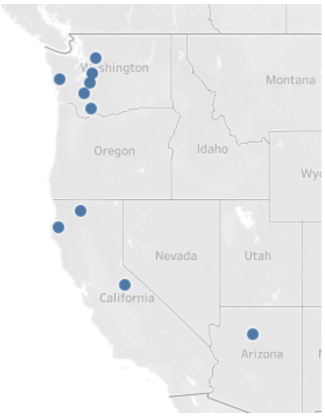
Report of observations and analysis about Bigfoot

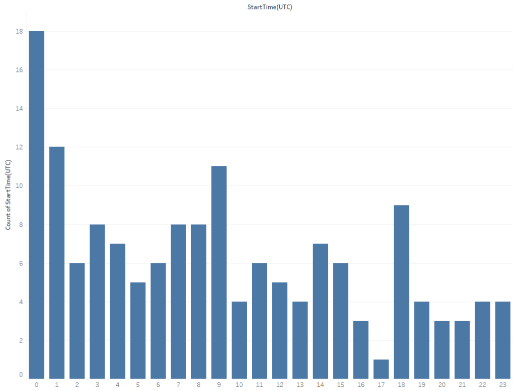
In analyzing the data pertaining to Bigfoot sightings (as represented in the task 4 dataset), we have observed several notable patterns and correlations that can help us better understand the conditions under which these sightings are most likely to occur. The frequency of Bigfoot sightings varies significantly with the time of year. The data indicates a marked increase in sightings during the winter months, with January and February showing the highest numbers, peaking at 1,013 and 873 witness counts respectively. This seasonal trend may suggest that Bigfoot is more active or visible during specific times of the year, or alternatively, that conditions such as bare trees in winter may make it easier to spot such creatures.



There is a fascinating correlation between sightings with multiple witnesses and specific weather conditions. Particularly, sightings where rain or fog was present seem to be more common. This connection could imply that Bigfoot is more likely to be seen during poor weather conditions, or that these conditions increase the chance of misidentification of other animals or phenomena as Bigfoot. When pinpointing the geographic distribution of sightings, Washington counties emerge as hotspots. This could potentially be attributed to the region's higher precipitation rates, aligning with the observed trend that sightings are more common in wet weather. Washington's diverse ecosystems, including dense forests and mountainous terrain, may also provide a suitable habitat that aligns with the commonly believed environment for Bigfoot. Furthermore, sightings are overwhelmingly more frequent in the earlier hours, with a significant drop as the day progresses. The highest number of sightings is reported at the very start of the day, potentially due to the nocturnal or crepuscular nature of Bigfoot, assuming the reports are accurate.



The additional datasets suggest that the high number of Bigfoot sightings during the rainy months of January and February may have unintended consequences. Specifically, the increase in sightings during adverse weather conditions like rain or fog could be linked to lower visibility and the psychological effects of such environments on people, potentially leading to misidentifications or heightened imaginative perceptions. Moreover, the concentration of sightings in Washington's counties, which have higher precipitation levels, hints at a correlation between weather patterns and sighting frequencies. These patterns could imply that environmental factors significantly influence the likelihood of reported Bigfoot encounters, which may not necessarily indicate the presence of such a creature, but rather the conditions under which people believe they have witnessed it.



To enrich the ecological background, the data on state-by-state environmental feature images were manually entered into the first dataset. These static features provide context for understanding the general ecological landscape of the sightings. 'Total Disasters' are the sum of all-natural disasters in a state from 1994 to 2021. (Figure 4, NOAA) The data was pulled from the Nation Centers for Environmental Information's annual state-by-state disaster pictures. This feature allows exploration of potential increases in Bigfoot sightings in years with higher numbers of natural disasters. Natural disasters occur that result in the possible relocation of wildlife, and such events may result in more frequent Bigfoot sightings. 'Forest Cover Rate' is the percentage of land covered by forest within each state, extracted from the image. This feature could support the analysis of Bigfoot's habitat preferences. Densely forested areas may report more sightings, which may indicate that Bigfoot relies on forested environments for hiding or sustenance. ‘Number of Lakes’ is to extract the number of lakes in each state from the picture and classify the number. This feature provides a way to examine the relationship between water resources and Bigfoot sightings, assuming that more lakes may correspond to more sightings due to Bigfoot's need for water.

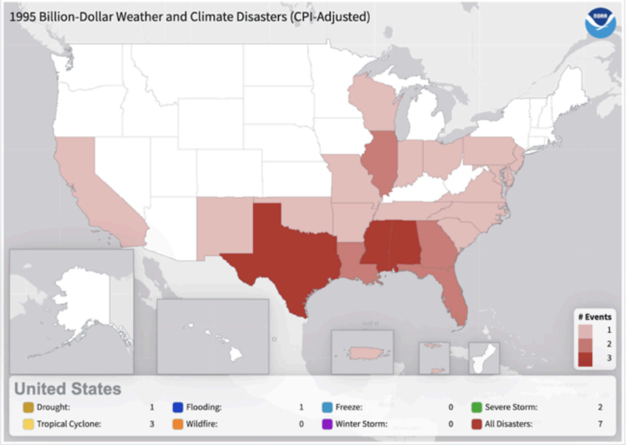


Figure 4. Total Disasters in 1994

At the same time, the second dataset about species distribution also provides dynamic, county-specific ecological information. This dataset is sourced from the New York State's public API and extracted after processing. ‘Wildlife Diversity’ quantifies the number of unique species recorded within each county as a direct measure of biodiversity. This characteristic reflects the ecological richness and habitat diversity of an area. This feature suggests that sightings of BigFoot are more common in areas with higher biodiversity. Bigfoot prefers or needs complex ecosystems with abundant resources and hiding places. The diverse species may also indicate less human disturbance, providing a safer haven for elusive creatures. ‘Conservation concerns’ indicate areas where a species is determined as endangered or threatened based on state or federal conservation status. This binary indicator highlights counties with severe ecological vulnerability and the need for conservation measures. Areas marked as ‘Conservation Concern’ may have a higher frequency of Bigfoot sightings. ‘Last Documented Presence’ records the most recent year for any species observed in the county, providing a temporal dimension to biodiversity data. This feature can provide insight into recent ecological changes, conservation efforts, or species migrations. Analyzing Bigfoot sightings under this feature can reveal patterns related to ecological or environmental disturbances. An increase in recent species recordings may indicate habitat loss for wildlife, affecting Bigfoot visibility or movement patterns.

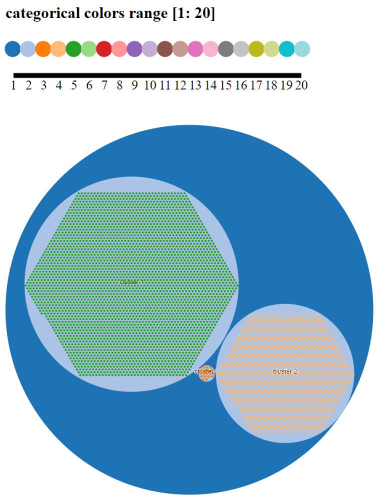
In areas where sightings have been reported, especially those with sensitive ecosystems, increased human activity may inadvertently threaten local wildlife and destroy habitat. This is particularly concerning in counties labeled ‘Conservation Concern’. Also, in states with dense forest cover, the likelihood of mistaking other animals or phenomena for Bigfoot may be higher, resulting in a possible increase in reported sightings. Instead, public interest in Bigfoot sightings offers an unconventional way to promote environmental protection. Sightings in areas of ecological value can draw attention to the need to protect these landscapes, potentially inspiring support for conservation initiatives.

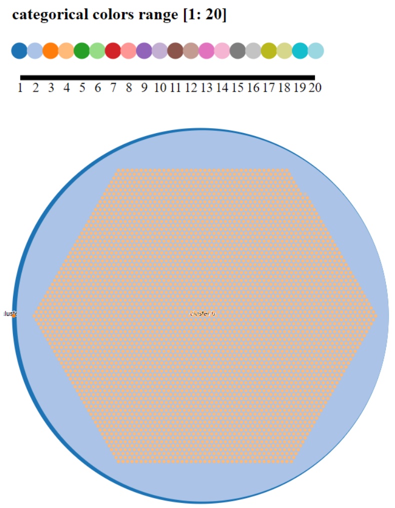
Meanwhile, the features from the unemployment dataset provide insights into the general size and socioeconomic status of areas with reported Bigfoot sightings. The features were extracted by merging the BFRO dataset with county-level unemployment data obtained from the US Department of Agriculture. Initially, relevant columns were filtered and reshaped to retain information such as area names, states, and unemployment rates for various years. Data cleaning steps were performed, and then a left merge was done on 'County,' 'Year,' and 'State' columns to combine the datasets, resulting in the extraction of the desired features for analysis. The column labeled ‘City/Suburb/Town/Rural’ assigned each area with one of those locale categories. While analyzing locale proportions, it was found that almost half were labeled as ‘city’ while nearly 17% were designated as ‘rural’. This finding challenges the traditional notion that Bigfoot sightings are primarily confined to remote regions. However, this phenomenon may be attributed to the higher population density of urban areas, resulting in more potential witnesses. Furthermore, forests and other natural habitats adjacent to cities may serve as locations for Bigfoot sightings. Such sightings occurring near urban areas may be included in the tally for the city, contributing to a higher number of reported sightings. The ‘Unemployment rate’ feature lists the unemployment rate for every county. This provides insights into how the socioeconomic status of an area can be related to Bigfoot sightings. An analysis of the dataset revealed a median unemployment rate of 6.3% across all counties. As of January 2024, the national unemployment rate is 3.7%, which is noticeably lower than 6.3%, suggesting that there is a correlation between areas with higher unemployment rates and increased sightings. This could be attributed to unemployed individuals having more leisure time for outdoor activities, potentially leading to more sightings in natural environments near cities. Meanwhile, the ‘Civilian labor force’ feature provides the number of people age 16 and older who are categorized as either employed or unemployed. While the civilian labor force does not provide the total population size of an area, it indirectly indicates population density, with areas containing larger civilian labor forces likely having greater overall populations. As mentioned before, areas with larger populations may have more reported Bigfoot sightings due to the greater number of potential witnesses.

The utilization of the features from the third dataset also reveal unintended consequences related to Bigfoot sightings. The correlation between higher unemployment rates and increased sightings suggests a socioeconomic dimension to the phenomenon, bringing to light underlying social issues in areas with higher unemployment rates. Furthermore, the analysis uncovered that a greater proportion of sightings occur in urban areas, which can be tied into the observation that densely populated areas tend to report more sightings. This raises concerns about the credibility of such reports, as skepticism may arise since witnesses can mistake humans for other entities, particularly under unclear visual conditions. The insights gained from these features highlight the multifaceted relationships between socioeconomic factors, human activity, and environmental dynamics in understanding Bigfoot phenomena.

The cluster visualizations generated using Apache Tika's Tika-Similarity library provided insightful groupings of the BFRO sightings data. The features we choose to compute similarity are Report Type, Class, Year, Season, Month, State, County, Nearest Town, Time And Conditions, Date, National Park Visitation Count, and Witness Count. These features were selected based on their potential to provide a comprehensive overview of each report's context, including when and where sightings occurred, the environment during the event, and witness credibility indicators. Three different similarity metrics – Jaccard, Cosine Distance, and Edit Distance – are employed, resulting in distinct clustering patterns(Figure 1-3, 100 jsons per folder).

Jaccard similarity considers the intersection and union of feature sets, which works well for the categorical and Boolean features present in this dataset, such as report class, season, location details, and witness count. The clusters formed using Jaccard similarity tended to group reports from the same geographic regions like states or counties together. These location-based clusters make intuitive sense, as sightings from nearby areas are likely to share many common characteristics.





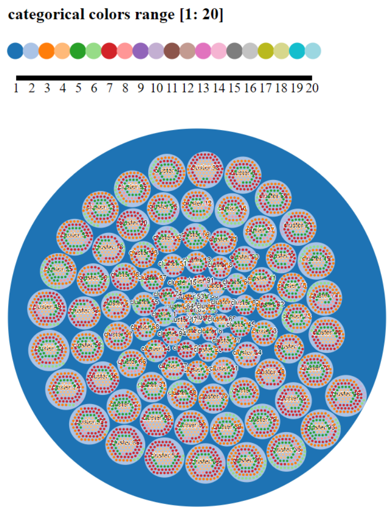


Figure 1. Jaccard Similarity

Figure 2. Cosine Distance

Figure 3. Edit Similarity

Given the diverse range of information available for each sighting, Jaccard's ability to holistically assess similarity led to highly meaningful clusters.

While Cosine Distance and Edit Distance also revealed some logical groupings, their respective focuses on specific fields or string comparisons resulted in less cohesive clusters compared to Jaccard. These metrics potentially separated events that were conceptually similar but differed in certain text attributes. Cosine Distance is more appropriate for datasets with continuous numeric features, while Edit Similarity is better suited for textual data. Since the BigFoot dataset primarily contains categorical variables, these metrics may not have captured the similarities as effectively as Jaccard.

Notably, there were distinct tight clusters of reports that had multiple witnesses listed. The presence of corroborating witnesses adds credibility to a sighting, so it is reasonable that these reports formed their own well-defined groups across all three similarity metrics. This distinct cluster warrants further investigation to understand if there are any common patterns or characteristics shared among the multi-witness events.

Overall, Apache Tika's clustering capabilities enable us to explore the information embedded within the BFRO dataset from multiple angles. By flexibly adjusting the similarity computation methods in tandem with the data's actual characteristics, we can achieve optimal grouping effects and further extract valuable insights.

However, the use of Tika-Similarity for clustering required significant data pre-processing efforts, including converting the data to JSON format using the ETLLib library. During the process of utilizing the ETLLib library, we encountered compatibility issues on the Windows operating system. This incompatibility necessitates the download of a virtual machine and also has the limitation of using only Python 2. These hurdles significantly drained our resources. Additionally, while attempting to run Tika Similarity on the entire dataset, we aimed to divide the data into folders containing 500 jsons each. Nonetheless, we consistently faced issues where the Tika server would crash, forcing us to reduce the number of jsons per folder below 300, the picture of which can be seen in the tika_result_images folder. This limitation posed a significant challenge to our analysis process.

Overall, Tika is a powerful tool for handling diverse data formats and computing similarities, but it may have a slight learning curve for unfamiliar users.