Yingyi Zhong, Mallory Xu, Yafei Wang, Ruohan Xu, Sean Iredell
Dr. Chris Mattmann
DSCI 550: Data Science at Scale
March 14, 2025

Haunted Places Dataset Analysis Report

Overview of The Haunted Places, Alcohol Abuse and Daylight Duration Datasets:

The original haunted places dataset shows diversity in paranormal evidence, with some locations having audio or visual proof while others lack documentation, possibly due to limited investigations or credibility concerns. Witness reports also vary significantly, with some locations having multiple witnesses and others none, indicating differences in documentation or personal experiences. Apparition types range from common ones like “Ghost” and “Male” spirits to rarer sightings such as “Orb” and “UFO,” while some reports remain vague or unclassified. Locations with more witnesses may have a higher likelihood of recorded evidence, and certain apparition types might be more frequently associated with sites where stronger paranormal evidence is documented. For the newly introduced feature Haunted Places Date, some of the event descriptions do not mention time or date-related information, and some data are Unknown. The feature Time of Day extracts the time when the event occurred from the event description. In general, most supernatural phenomena occur at night or in the early morning when the light is weak. In the feature Event type, remove the Unknown part, and Supernatural accounts for almost half. The original data set structure also provides direction for future data analysis. Use geographic data to understand the "haunted hotspot map" in the United States, or further analyze the correlation between the number of witnesses and audio or impact evidence.

In the joined haunted places dataset, total death and death percentage under 21 due to alcohol abuse for each state was joined with no missing values. The values are copied from the table on the website into an excel spreadsheet. In the excel spreadsheet, there is one column indicating “State” names, formatted in the same way as the “haunted_places.csv”’s “state”, and other two columns “Alcohol Abuse Total Deaths” and “Alcohol Abuse Death% Under 21”. The same excel spreadsheet was used to document sunrise time, sunset time and daylight duration for each state for convenience. These features require manual entry or effort to collect information. To keep the data consistent and the process efficient, each state capital’s sunrise time, sunset time and daylight duration is collected instead of referring to all the cities included in the original haunted_places.csv. All sunrise and sunset time are collected on February 26th 2025 while the annual average daylight duration of 2025 is collected by calculating the average of everyday of 2025’s daylight hours in the query result.

Three Additional Datasets:

1) **Excel Dataset: The Federal Real Property Profile**

The Federal Real Property Profile (FY23 Federal Real Property Profile, FRPP) is an annual public data set released by the General Services Administration (GSA) of the United States, recording detailed information on real estate (such as buildings, land, facilities, etc.) owned or managed by the federal government in fiscal year 2023.

By merging this dataset with the original paranormal dataset, we can analyze the spatial association between federal real estate and "haunted places", the relationship between historical buildings and "haunted places". We chose the three core features of Age of Property, Utilization, and Replacement Value and added them to the original dataset to better understand price discounts, market liquidity, legal disclosure rules, and cultural perceptions.

The analysis of Age of Property is to explore the association between the year of construction and paranormal reports. Overall, the age of buildings in different states in the United States varies greatly.

The buildings in some states are relatively new, while the buildings in some states have a long history. The buildings in eastern or early industrialized states such as Massachusetts and New York are older, probably because these areas developed cities earlier and have more historical buildings. Some southern and western states, such as Arizona, Nevada, and Florida, have relatively new buildings, probably due to late urbanization or frequent demolition and reconstruction. According to the different ages of the buildings, older buildings have a greater probability of reporting paranormal events. The analysis of Utilization is to explore the association between property utilization and paranormal events. If the housing utilization rate in a region is low, it means that the houses are vacant or have not been taken care of for many years, which may affect the visual perception of passers-by at night and lead to reports of paranormal events.The purpose of analyzing Replacement Value is to explore the relationship between real estate replacement value and paranormal events. The higher the real estate replacement value in a region, the higher the quality and value of the house itself, which represents a comprehensive evaluation of the quality of the house. Houses with higher comprehensive quality have a lower probability of reporting paranormal events.

2) HTTP Dataset with Features: Happiness score, Mental Illness Rate (%), Suicide Rate (%

To enrich the Haunted Places dataset, additional three datasets from World Population Review (2024) were incorporated, including Happiness Score, Mental Health Statistics, and Suicide Rates by State. These datasets provide key indicators related to emotional well-being, work environment, community conditions, mental illness prevalence, and suicide rates per 100,000 people. By merging these datasets, it became possible to analyze potential relationships between hauntings and psychological well-being on a state-by-state level.

Data extraction was performed using Python, with requests sending HTTP requests to retrieve data, BeautifulSoup parsing HTML content to extract relevant table information, and pandas structuring the dataset for further analysis. The Happiness Score dataset included columns such as “state”, “total happiness score”, “emotional & physical well-being rank”, “community & environment rank”, and “work environment rank”. The Mental Health Statistics dataset contained “state”, “rates of mental illness”, “adults with anxiety or depression”, “adults with severe mental illness”, and “overall mental health standing (youth & adults)”. The Suicide Rates dataset included “state”, “suicide rate (per 100k)”, and “suicides”. After cleaning the data, the three datasets were merged using state names as the common key. From each dataset, the most representative column was selected and renamed to create a new DataFrame, resulting in a final dataset containing four key columns: “state”, “happiness score”, “mental illness rate (%)”, and “suicide rate (%)”. To ensure consistency, “happiness score” was converted to a numerical float for quantitative comparisons, “mental illness rate (%)” had percentage symbols removed before being transformed into numerical format, and “suicide rate (%)” was standardized by converting per 100k population rates into percentages. The final cleaned and structured dataset was then saved as a CSV file for further analysis.

3) State Capital Climate – CSV/Text File

The same data collection approach which focused on gathering information for each state capital was used for the additional dataset, a csv text file that belongs to one of the eight primary MIME types. The dataset includes climate information for each capital city for each month. The csv file is downloaded from Local Climatological Data (LCD) and Summary of Monthly Normals(NMLY), provided by National Centers for Environmental Information (NCEI) and National Oceanic and Atmospheric Administration (NOAA). LCD from 2024 is used for this project, which provides information about Monthly Total Liquid Precipitation, Monthly Mean Temperature and Daily Average Wind Speed. Monthly Average Wind Speed is calculated from the average of the sum of daily average wind speed divided by the total number of the days for each month. Missing values on Monthly Total Liquid Precipitation and Monthly Mean Temperature are manually filled by using information provided by NMLY as a second reference. Missing values from Monthly Average Wind

Speed are filled by using information provided by Weather Spark. After the averages of wind speed, precipitation and temperature for each state capital is collected or calculated, they are featurized based on the below standard:

- Monthly Mean Temperature is categorized into three levels based on its value. If the temperature is below 50°F, it is considered cold. When the temperature falls between 50°F and 79°F, it is classified as mild. If the temperature reaches 80°F or higher, it is regarded as hot.
- Monthly Total Liquid Precipitation is also classified into three levels. When the recorded precipitation is less than 2 inches, the condition is labeled as wet. If the precipitation falls between 2 and 5 inches, it is considered normal. Any amount exceeding 5 inches is categorized as dry.
- Monthly Average Wind speed grouped into 3 categories as well. A wind speed of less than 4 mph is described as gentle. If the wind speed ranges between 4 and 8 mph, it is considered windy. When the wind speed exceeds 8 mph, it is classified as strong.

Weather conditions can greatly influence perceptions of haunted places. Factors like temperature, precipitation, and wind can distort senses and heighten fear. Cold enhances eerie sensations, making chills feel supernatural, while heat can cause fatigue and hallucinations. Rain and fog obscure vision, leading to misinterpretations of vague shapes, while wind moves objects unpredictably, creating the illusion of ghostly activity. Sounds like dripping water or rustling leaves can further amplify eerie experiences.

Analysis Approach Summary:

In the Jaccard Similarity two distinct clusters appeared instantly, one cluster that had Audio evidence and one cluster that did not. This single distinct feature proved to be decisive as all similarities calculated with a .63 similarity (the highest) contained audio evidence. In edit distance there were several very specific groups dealing with some of the added features including one specific group (cluster 4) that was very climate focused with all of the events being murders in cold weather while it was raining. Another edit distance cluster (cluster 1) dealt with very specific geographical similarity with all of the events taking place in desert areas of California. This cluster made me question some of the accuracy of the similarity because of repeated collinear features. For example, with this California cluster all of the weather effects were the same, with the same drinking, and daylight allotments. This suggests that possibility to me that this cluster was too heavily influenced by locational factors and not enough by some of the more important features including apparition type, date, or event type.

Edit distance provided the best groupings because it provided both the most clusters in terms of numbers, as well as, providing the most robust clusters. This might because both cosine and jaccard were not able to parse beyond some of the limitations of the preprocessing and developed large clusters around information that was not relevant (ie. dates of “2025-01-01” and event type “Unknown”). Edit distance was able to circumvent these limitations because it was more equally weighted when evaluating features because it did not include unions or products of different features.

The process of using Apache Tika was smooth and seamless, providing instant results. A particularly enjoyable aspect was testing the installed server, where files could be easily dragged and dropped to receive immediate metadata. In contrast, ETLlib presented more challenges due to its outdated Python dependencies and unclear interactions with Apache, particularly in how Tika utilized it for generating visuals. Methods with a more straightforward logic—such as defining an input directory, an output directory, and executing the transformation locally—are preferable.

Findings:

Based on our preprocessing, there is no significance associated with the time of day. However, there was also not an optimal ratio of time of days that could be extracted to time of days that could not be

extracted. Overall, it did not seem like a very relevant feature as, for the most part, the events happened at night, or at least towards the end of the day when light was limited.

Additional datasets reveal unintended consequences related to haunted places. The three most important are alcohol abuse, climate, and property age. Two key factors drive more reports: alcohol impairing logical reasoning and suboptimal sensory conditions. The alcohol abuse dataset is straightforward, while climate and property age offer deeper insights. Weather conditions like wind and rain can lead to misinterpretations, such as mistaking a slamming door for paranormal activity. Poor visibility due to rain likely contributed to many false reports. The most intriguing factor is property age—older buildings often have dim lighting and structural issues that create unsettling noises, increasing the likelihood of hauntings being reported.

Specific locations are more likely to be influenced by alcohol abuse, just by general geographic statistics, but that would be impossible to tell in this case because the lowest granularity that we have access to is state level data. Alcohol abuse has much more regional and local influences, but it is not effective to estimate its state level prevalence over such wide and regionally diverse areas. For example, alcoholism is probably a much larger contributor to haunted sittings in a rural town in Texas than it is in a major city like Dallas or Austin. Using state level data effectively erases this correlation by having larger cities normalize smaller more regional effects. When collecting data in the future it may prove more beneficial to pick one level of analysis and list effects as such. For example, having regional effects of alcohol, only using regional level reports of haunted places and throwing away more local information to be able to better analyze the features and response variable and their relation.

The keywords that were the most common and became the biggest indicators of apparition type was an obvious one like “ghost”, but also many apparition types were attributed human characteristics and called “boy”, “girl”, “child”, “man” and “woman”. For the purposes of our research we broke these into three different categories with ghosts representing more paranormal unexplainable phenomena and children / male / female all being used as separate categories. One interesting study could be to combine all of these keywords into one category because ultimately it breaks down to a semantic difference in what the definition of a ghost actually is. It might be interesting to see if there is a sharp difference between events that featured apparitions with human characteristics and those that did not. At least implicitly to me it make sense for humans to recognize human characteristics in their surroundings even if they are not there because of the way our biology and senses were designed, so it almost gives more credibility to the apparitions that did not have human elements present because they were so glaringly different that they could not be attributed to the more common “human ghost.”

Based on the dataset I would say the feature with the highest likelihood of defining a haunted place is the apparition type. This is because during preprocessing it showed the highest success rate of identification with over 3/4 of the instances having a non “unknown” value. Additionally, this was the feature that I noticed separated legitimate hauntings from what I would consider illegitimate hauntings. Illegitimate hauntings are usually reports where the reporting party hears sounds that they can not identify or has feelings that they are not alone, however these instances are significantly weaker in evidence than specific cases where someone describes seeing a ghost, orb, or UFO. This is not to fully discount audio evidence and put a premium on visual evidence, I am just suggesting that for many of the apparition types for legitimate cases there are instances that are cited with specific recurring images. For many of the cases with sound alone the descriptions are vague with reports of doors shutting, voices talking, or footsteps. These auditory clues are much more attributable to building decay, wind, or animals. It is much harder to fool someone’s eyes than their ears. For this reason I think that the easiest way to define a haunted place from a non-haunted place is by apparition type and description, with the more specific the visual description the more believable.