DSCI 550: Big Data Assignment - Haunted Places Analysis Report

Team Members: Zili Yang, Chen Yi Weng, Aadarsh Sudhir Ghiya, Niromikha Jayakumar, Yung Yee Chia, Colin Leahey

1. Introduction

1.1 Objective

This project explores the Haunted Places dataset (21,983 entries) to uncover hidden patterns in reported supernatural phenomena across the U.S. Key goals include:

- Extracting structured features (e.g., witness counts, apparition types) from unstructured text.
- Joining external datasets (e.g., alcohol abuse, daylight duration) to identify correlations.
- Clustering haunted places using Tika-Similarity to analyze relationships between features like time of day, witness reliability, and geographic proximity.

1.2 Datasets Overview

The Haunted Places dataset includes 21,983 entries with attributes such as City, State, Description, and Location. Descriptions vary widely in detail, ranging from explicit events (e.g., "78 miners died in 1890") to vague anecdotes (e.g., "strange lights seen at night"). This variability makes feature extraction and similarity analysis critical for structured insights.

2. Data Collection and Preprocessing

2.1 Haunted Places Dataset

- Source: Downloaded via a Dropbox link provided.
- Preprocessing :
    - File Format Conversion: Converted the dataset from CSV to TSV using pandas for compatibility with Tika-Similarity.
    - Handling Missing Values:
        - Replaced empty strings and whitespace with NaN for proper handling.
        - Removed fully empty rows.
        - Filled missing latitude/longitude values using city-level coordinates where available.
        - Dropped rows still missing location data after imputation.
    - Text Standardization:
        - Lowercase all text fields.
        - Removed punctuation to standardize descriptions.
    - Feature Engineering:
        - Audio and Visual Evidence: Identified mentions of sounds, voices, and visual phenomena using regex-based keyword searches.
        - Date Extraction: Used the datefinder library to extract dates from descriptions. Defaulted to "2025-01-01" when no clear date was found.
        - Witness Count: Used number-parser to extract numerical values from written descriptions, replacing vague phrases (e.g., "many" → 10, "a few" → 3).
        - Time of Day: Extracted keywords like "morning," "evening," and "dusk" to categorize the time of reported hauntings.
        - Apparition Type & Event Type: Identified keywords related to ghostly appearances (e.g., "orb," "ghost," "UFO") and events (e.g., "murder," "supernatural").

This preprocessing ensure data consistency, improved feature extraction, and made the dataset more suitable for similarity analysis and clustering.

2.2 Additional Datasets and Data Integration

- **Dataset 1:** Alcohol abuse by state
    - Why Chosen: To explore correlations between alcohol abuse and unreliable witness reports.

    - Features: Extracted state-level alcohol abuse statistics and merged on State_abbrev.
- **Dataset 2:** Amount of daylight by state
    - The daylight dataset was obtained through web scraping from "Time and Date" using Selenium and BeautifulSoup, extracting tables of sunrise and sunset times for U.S. states. It was chosen to analyze how daylight duration influences witness reliability, paranormal reports, and alcohol consumption patterns. The extracted HTML tables were processed into a TSV file (daylight_hours_full.tsv), with sunrise and sunset times converted from AM/PM format to 24-hour numeric values for accurate calculations.
- **Dataset 3:** Religious Adherents Census Data by county
    - This dataset is combined with two datasets. This first one is the GeoJSON file representing the shapes of all counties in the United States. The second dataset is census data that includes the population(2020), the number of religious adherents (who have a religious faith), and the percentage of that proportion for each county in .xlsx format.
    - Why it was chosen? Understanding the relationship between religious adherence and reported haunted house locations may provide insights into potential sociocultural or psychological influences on paranormal activity reports. Higher or lower religious adherence levels might correlate with the prevalence of haunted house claims, influencing beliefs in the supernatural and ghostly encounters.
- **Dataset 4:** Kaggle - US crime rates by county
    - The crime dataset was chosen to analyze the relationship between crime rates, haunted locations, and alcohol consumption patterns at the county level. The dataset, originally sourced from ICPSR Crime Data (2016) and U.S. Census Population Data (2013), was formatted as TSV and merged using state abbreviations (state_abbrev) and county names (county). A left join ensured that all haunted place records were retained while crime data was added where available, allowing for deeper insights into potential crime-haunting correlations.
- **Dataset 5:** Historic Places
    - The dataset was chosen to see if a certain place with historical significance (i.e., a battlefield or old church) would be associated with reports of haunted places. The dataset is from the National Park Service. Extracted were about 80,000 names of historic places, their location, and a date if applicable. The data was briefly converted to a geojson file in order to see the places' proximity to haunted reports. A radius of 5 miles was used to determine proximity.

3. Feature Engineering

3.1 Extracted Features from Haunted Places Dataset

Feature Name	Description
Audio Evidence	Identifies whether the description mentions noises or sounds
Image/Video/Visual Evidence	Detects mentions of cameras, photos, or written messages
Haunted Places Date	Extracted or defaulted date
Haunted Places Witness Count	Extracted number of witnesses from descriptions
Time of Day	Categorized as morning, afternoon, evening, dusk, or unknown
Apparition Type	Categorized as ghost, orb, UFO, UAP, or unknown
Event Type	Identifies if the event involved a murder, death, or supernatural occurrence

3.2 Features from Required Datasets

- **Dataset 1: Alcohol Abuse by State**
    - Binge_Drinking_Rate (%): The percentage of the adult population in each state that engages in binge drinking. This helps measure alcohol consumption intensity.

    - Annual_Alcohol_Related_Deaths: The number of deaths per year attributed to alcohol abuse. This highlights the impact of excessive drinking on public health.
    - Cost_Per_Drink (USD): The average cost per alcoholic drink in each state. This can be used to analyze economic factors related to alcohol consumption.
- **Dataset 2: Daylight Hours**
    - Avg_Daylight_Hours: The average number of daylight hours per state
    - Sunrise_Variability: The standard deviation of sunrise times within each state
    - Daylight_Hours_Range: The difference between the maximum and minimum daylight hours

3.3 Features from External Three Datasets

- **Dataset 3: Religious Adherents Census Data (by County)**
    - Adherents: Number of religious adherents for each county
    - Adherents as % of Population: Percentage of religious adherents population for each county
    - Haunted Houses Count per County: The count of haunted houses for each county
- **Dataset 4: Kaggle - US crime rates by county**
    - crime_rate_per_100000: Crime rate per 100,000 people in the county.
    - MURDER: Number of reported murder cases.
    - ROBBERY: Number of reported robbery cases.
    - BURGLARY: Number of reported burglary cases
- **Dataset 5: Historic Places**
    - Name of historic place
    - Year of nearest historic place
    - Number of historic places within 5 miles

4. Similarity Analysis Using Tika-Similarity

4.1 Installation and Setup

- **Steps taken to install Apache Tika, Tika-Python, and ETLLib**:
    - Installation of Apache Tika was straightforward using the command pip install tika to enable text extraction and similarity analysis.
    - Installed ETLLib for data transformation tasks.
    - Verified installations by running sample text extraction and similarity checks.
- **Conversion of TSV to JSON using ETLLib**:
    - Transformed the TSV data into JSON format for processing and analysis using tsvtojson.

4.2 Distance Metrics and Clustering

- **Jaccard Similarity**: Clusters formed using Jaccard Similarity grouped places with similar thematic elements, such as haunted hotels or historical sites.
- **Cosine Similarity**: Clusters formed using Cosine Similarity grouped places with similar term frequencies, such as those emphasizing "ghost sightings" or "paranormal activity."
- **Edit Distance**: Edit Distance highlighted textual similarities in descriptions, capturing minor variations in phrasing. Edit Distance grouped places with nearly identical descriptions, such as those sharing common legends or historical accounts.

5. Results and Discussion

5.1 Key Findings

- The analysis revealed that external factors like cultural context, environmental conditions, and historical events play a significant role in the prevalence and nature of haunted place reports.
- Alcohol abuse and low-light conditions emerged as potential amplifiers of paranormal experiences, either through psychological effects or cultural storytelling.

5.2 Interpretation of Similarity Clusters

- Jaccard Similarity focused on word overlap, forming clusters based on shared vocabulary.
- Edit Distance created clusters where minor text variations appeared frequently.

- Cosine Similarity identified semantically related locations, leading to better clustering in terms of narrative similarity.

6. Conclusion and Future Work

6.1 Summary

This study analyzed the Haunted Places dataset, integrating external data sources to examine potential correlations between hauntings and external factors such as alcohol abuse, daylight duration, crime rates, and historical significance. By structuring unstructured text and applying similarity-based clustering, the study provided a more data-driven approach to understanding supernatural reports. The findings suggest:

- States with higher binge drinking rates reported more hauntings, suggesting a possible link between substance use and perception of supernatural experiences
- States with fewer daylight hours correlated indicates a higher reports of haunting, suggesting the idea that visibility conditions of the environment might influence perceptions
- High crime rate areas particularly near historically violent sites, abandoned buildings and urban legends shows a higher concentration of hauntings
- Using Tika-Similarity, distinct clustering patterns emerged, revealing groupings of haunted locations based on apparition type, geographic proximity, and external characteristics.

6.2 Limitations

Merging external datasets required extensive preprocessing due to inconsistent location formats. Missing or ambiguous geographic identifiers limited data integration accuracy. Aggregated state-level data restricted analysis at finer spatial levels. Text similarity methods, while effective for grouping haunted places, struggled with long descriptions and required optimization for large datasets. The study relied primarily on text-based clustering, which may overlook additional spatial or environmental factors influencing hauntings. Despite these limitations, integrating structured data provided a more systematic approach to analyzing paranormal reports.

6.3 Future Enhancements

Here are some potential improvements that will help

- Expanded datasets: Integrating latitude/longitude data for precise location mapping, weather conditions to assess environmental effects, and mental health statistics to explore psychological influences.
- Advanced data processing: Utilizing BERT/LDA for NLP to improve text feature extraction, refining geospatial clustering for better spatial analysis, and applying time-series analysis if timestamped data becomes available.
- Enhanced visualization: Developing interactive heatmaps to illustrate geographic trends, network graphs to map relationships between haunted locations, and spatial overlays to compare hauntings with external factors like crime rates and historical sites.

7. References

Kaggle. “United States Crime Rates by County,” December 28, 2016. https://www.kaggle.com/datasets/mikejohnsonjr/united-states-crime-rates-by-county.

“Maps and Data Files for 2020 | U.S. Religion Census | Religious Statistics & Demographics.” n.d. https://www.usreligioncensus.org/node/1639.

“National Register of Historic Places.” 2014. NPS Data Store. May 2014. https://irma.nps.gov/DataStore/Reference/Profile/2210280.

U.S. Department of Commerce, U.S. Census Bureau, Geography Division, Spatial Data Collection and Products Branch. 2024. “Counties - United States of America.” June 6, 2024. https://public.opendatasoft.com/explore/dataset/georef-united-states-of-america-county/export/?flg=en-us&disjunctive.ste_code&disjunctive.ste_name&disjunctive.coty_code&disjunctive.coty_name.