Alysa Xu, Catherine Lu, Justin Huh, Rohan Rane, Vishal Menon
DSCI 550: Data Science at Scale
13th March 2025

DSCI 550 Group 8 Homework 1 Report

I. Observations and Insights from the Dataset

While completing the tasks, we noticed the volume of the data (21,984 rows and 10 columns) and the variety of column names and the data types. Throughout the course of the project, we also noticed how several columns have multiple null values, specifically city, country, description, and location. The description column is quite unstructured, which makes text extraction difficult. For the time of day we were able to extract, we noticed descriptions often mention nighttime activity, which may point to the belief that hauntings are more commonly perceived in low-light conditions. We also noticed that many locations share the same city but have different descriptions, suggesting possible duplicate/related reports.

II. Answering New Questions with Joined Datasets

By integrating the Crime, Census, and Moon Phase datasets, we were able to introduce new dimensions to the analysis that were previously unanswerable using the Haunted Places dataset alone. Some of the key insights now possible include:

1.) Crime Dataset

a.) Are hauntings more frequently reported in areas with unsolved crimes?
b.) Do ghost sightings correlate more with male or female victims of past crimes?
c.) Are haunted locations linked to violent crimes involving weapons, such as firearms or knives?

2.) Census Dataset

a.) Are hauntings more common in highly populated areas or more isolated places?
b.) Do cities with older populations report more hauntings (perhaps due to cultural beliefs or historical significance)?
c.) Is there a correlation between abandoned buildings, vacant housing units, and paranormal activity?

3.) Moon Dataset

a.) Does the likelihood of a ghost sighting increase during a full moon?
b.) Are hauntings reported more frequently when the moon is closer to Earth?
c.) Are ghost sightings correlated with supermoons, when the moon appears largest in the sky?

III. Clusters Revealed in Data

The clustering results revealed distinct patterns across different similarity metrics. Jaccard similarity formed two dominant clusters, suggesting that haunted locations shared highly overlapping categorical attributes, such as crime type, location, or haunting descriptions. Edit distance, on the other hand, produced highly fragmented clusters, indicating that small textual differences in descriptions significantly impacted grouping, often creating too many small clusters that may not accurately reflect meaningful similarities. Cosine similarity generated one large cluster, which implies that numerical attributes (e.g., crime rates, population size, moon phase data) were highly correlated across haunted locations, causing most locations to be grouped together based on numerical patterns rather than categorical differences.

IV. Evaluating Similarity Metrics for Grouping

The most accurate similarity metric depends on the type of features being analyzed. Jaccard similarity was effective in clustering locations with shared categorical features, such as the type of crime or haunting, making it useful for analyzing recurring patterns in hauntings. Cosine similarity was best for grouping locations with similar numerical trends, such as cities with comparable crime rates, moon phase values, and population densities, making it ideal for identifying statistical patterns in hauntings. Edit distance, however, was the least effective as it over-fragmented the data, treating even minor variations in text descriptions as distinct clusters, leading to unnecessary separation of similar locations. Therefore, for categorical data, Jaccard is best, while for numerical patterns, Cosine similarity is more appropriate.

V. Unintended Consequences Suggested by Additional Datasets

The integration of crime, census, and lunar data suggests unintended correlations and potential biases in haunted location reports:

1.) Crime Dataset

- a.) Locations with unsolved violent crimes may be perceived as haunted due to local folklore, trauma, or psychological imprinting.
- b.) The fact that female victims are linked to certain crimes may shape how hauntings are reported and whether ghosts are perceived as malevolent or benign.
- c.) Some locations may market haunted reputations to tourists while ignoring the history of violence.

2.) Census Dataset

- a.) Cities with older populations and historical buildings may report more hauntings due to nostalgia, cultural beliefs, and preserved architecture.
- b.) Areas with high vacancy rates and economic decline might have more abandoned buildings, which are frequently labeled as "haunted."

3.) Moon Dataset

- a.) If people believe a full moon affects supernatural activity, they may be more likely to report hauntings on such nights.
- b.) The brightness and size of the moon may influence how people interpret strange noises, shadows, and light reflections, leading to increased paranormal reports.

VI. Additional Datasets Descriptions & Feature Extraction Method:

1.) Crime Datset (text/csv)

- a.) The crime dataset obtained from Kaggle is a record of crimes in the US from 1980 to the present. It includes 24 different columns that describe the crime including details of the perpetrator, victim, when the crime occurred, and type of crime. We added three features of crime solved (yes or no), victim sex, and weapon
- b.) **For the crime solved (yes/no)**, we obtained the data from the US Crime kaggle dataset. Some queries that can be answered based on this feature are what percentages of crimes remain unsolved over time, are certain types of crimes more likely to remain unsolved, and does the solving of the crime vary by location, year, or victim/perpetrator demographics? **For the victim sex feature**, some queries that can be answered are if certain crimes are more likely to have male vs. female victims, are female victims more likely to be associated with domestic crimes vs. public violence, and does the crime-solving rate differ based on the victim's gender? **For the weapon feature,** some queries that can be answered are what the most commonly used weapon in crimes over time is, are gun-related crimes more likely to be solved than non-gun crimes, and are certain weapons more likely to be used against male vs. female victims?

2.) Census Dataset (application/xml)

- a.) This dataset contains city-level census data from the American Community Survey (ACS). It includes demographic and housing-related information for different cities in the United States. The dataset was retrieved from an API xml file and includes the following key attributes: The name of the city, the median age residents in the city, the population or total number of residents, and the total number of housing units available in the city. This dataset was cleaned and formatted before merging it with the haunted places dataset to provide additional demographic context.
- b.) Extracting the features of the dataset involved the following steps: First the city name was cleaned; specifically was extracted from the "NAME" column, removing state names and extra identifiers like "CDP", "village", "town", or "borough". Then, the text was normalized to lowercase, capitalized, and duplicates were removed to ensure each city appears only once. Next, unnecessary columns were dropped because the dataset originally contained "state" and "place" codes, which were removed because they weren't needed for merging. Lastly, they were merged with the haunted places dataset. The "city" column was used as the key for joining this dataset with the haunted places dataset.
- c.) **For the median age**, we obtained this data by extracting directly from the ACS API dataset as DP05_0018E. The data represents the median age of all residents in a given city. Some queries

that can be answered with this feature include the following: Do cities with higher median ages have different crime rates, are hauntings reported more frequently in older or younger groups, how does the median age of a city correlate with weapon-related crimes? **For the population feature,** the data was extracted from the ACS API as DP05_001E, representing the total number of people in each city. Some queries that can be answered are if more crimes are reported in highly populated cities, do larger cities have higher crime-solving rates, and lastly do hauntings occur more frequently in densely populated areas? **For the housing units,** the data was extracted from the ACS API as DP04_0002E, representing the total number of housing units in each city. Some queries this feature can answer are as follows: do cities with a large number of housing units experience higher or lower crime rates, are hauntings more common in areas with older housing units, and lastly does the number of free housing units correlate with high/low economic activity in haunted locations?

3.) Moon Dataset (image/tiff)

a.) According to Vedic Astrology, the moon is extremely important in determining the well-being of humanity and what occurs that night. Auspicious ties to the phases of the moon are extremely common. For example, some religions believe the full moon brings good luck to a household, while others will not look at a full moon as they believe this is inauspicious. We wanted to understand whether there was a correlation between the phases of the moon, distance from the Earth, and how much light is given by the moon through the visible diameter with the prevalence of haunted instances.

b.) We looked into the metadata of the NASA phases of the moon data set which allowed us to query any date and time to produce an image of the moon and its phase. The metadata includes phase, distance from Earth, and diameter of visibility which correlates based on the image shown from that day. For the Moon Phase Image dataset, for each haunting event date, a request was sent to NASA's Dial-a-Moon API. We retrieved information such as the moon phase, moon diameter, and moon distance. **For the moon phase**, we obtained the values from the NASA Dial-a-Moon API and used the data to fetch the moon phase corresponding to the Haunted Places Date. The value represents the phase of the moon at that time in degrees where 0° is a New Moon, and 180° is a Full Moon. Some queries that can be answered for the moon phase feature data are if ghost sightings increase during a full moon, are specific types of hauntings more frequent during certain moon phases, and does the visibility of the moon (New vs. Full Moon) affect how people perceive paranormal events? **For the moon diameter**, we obtained the values from the NASA API, specifically the apparent diameter of the moon (in arcseconds) from Earth on the given date. Some queries that can be answered with the moon diameter data feature are as follows: do people report more hauntings when the moon appears larger in the sky, is there a correlation between supermoons and paranormal activity and are certain ghost sightings linked to times when the moon is visually smaller? **For the moon distance,** we obtained the values from the NASA API, specifically the distance from the Earth to the Moon in kilometers on the given haunting date. Some queries that can be answered with this moon distance data feature are as follows: Are hauntings more common when the moon is closer to Earth, Does the gravitational effect of the moon play a role in supernatural beliefs and Is there any connection between extreme lunar distances and increased ghost sightings?

<u>VII. Assessment of Sufficient Information Obtained to Answer Additional Questions</u>

1.) Are there clusters of Haunted Places with similar features, and all are murders occurring in the evening?

a.) Based on the cosine similarity clustering, we can see that many Haunted Places were placed in the same cluster, and likely have similar features. The Jaccard similarity clustering also produced two clusters, so it shows that there is some similarity between the Haunted Places’ features.

2.) Does the time of day of the Haunted Place original sightings matter?

   a.) We can examine Haunted Places as a DataFrame and see what proportion of rows occurred at a certain time of day. If more Haunted Places were reported at a certain time of day, we could infer that the time of day mattered.

3.) Are specific locations more likely to be influenced by alcohol abuse that cause more Haunted Places to be reported?

   a.) We can examine Haunted Places as a DataFrame and see if more Haunted Places were reported in states with higher rates of alcohol abuse. If certain locations of Haunted Places are reported with high rates of alcohol abuse, we could infer that those locations may be correlated with a higher concentration of reports of Haunted Places.

4.) Are specific keywords bigger indicators of the apparition type related to a HauntedPlace?

   a.) We can examine Haunted Places as a DataFrame and count how many rows have a specific keyword for an apparition type. For example, the keyword "spirit" could be a significant indicator for the Ghost apparition type.

5.) Is there a set of frequently co-occurring features that define a particular Haunted Place?

   a.) We can examine Haunted Places as a Dataframe and see if a particular Haunted Place was reported multiple times. We can compare the features of each report and see if any were frequently co-occurring.

6.) What insights do the "indirect" features you extracted tell us about the data?

   a.) The "indirect" features (such as audio, visual, date, witness count, time of day, apparition type, event type) we extracted help categorize the Haunted Places into certain types of reports. It added more detail to the origin of information of the data. When clustering the data, these features provide similarity between different Haunted Places. One potential insight is if external factors like proximity to cemeteries or crime rates correlate with hauntings, they may provide alternative explanations for paranormal reportings. A regression or correlation analysis between hauntings and indirect factors such as population density, historical landmarks, and crime data could be run to confirm results.

7.) What clusters of Haunted Places made the most sense? Why?

   a.) We don't have enough information about which Haunted Places were placed into the same cluster to determine which clusters made sense. From the clustering visualizations, we weren't able to see information about the data in each cluster. If we could figure out which Haunted Places were placed into each cluster, we could determine what features they have in common to see if the clusters made sense.

VIII. Reflections on Using Apache Tika

The tutorials on using Apache Tika were straightforward and helpful. However, because we were using Google Colab to run our code, we often had to modify certain commands to work with Colab. Since this project involved finding datasets and using Python to join them, we wanted to be able to work on the code collaboratively. However, this led to issues later on when we needed to have files on our local machine to run tika-similarity. Additionally, Python SimpleHTTPServer doesn't work on Colab, as the server cannot be accessed externally. We had to download the CSV, JSON, and HTML for jaccard_similarity, editdistance, and cosine_similarity onto our local machines in order to see the visualizations.

Interpreting the visualizations were also unclear: the Haunted Places TSV was converted into JSON files with random ids, and we could not tell which Haunted Places were placed into each cluster. It was also unclear whether Tika was clustering by metadata or by features.

The biggest challenge we ran into was scaling tika-similarity to work with the full Haunted Places dataset. We were able to set up the pipelines for running 100 JSON files, but trying to run the full dataset or even 500 JSON files caused the Tika server to stop responding. We attempted to change the jaccard_similarity.py file to call Tika's parser less, and were able to obtain the jaccard.csv file. However, this file was 8 GB, and we could not run the edit-cosine-circle-packing.py or edit-cosine-cluster.py files. They timed out on Google Colab due to memory limits, and we could not find a solution.