Homework: Analysis of the Haunted Places Dataset
Due: Friday, March 14, 2025 12pm PT

---

1. Overview

Figure 1: The Haunted Places dataset: mysterious reports strange sightings in and around the United States consisting of 21,983 rows x 10 columns. The full dataset can be found at https://www.kaggle.com/datasets/sujaykapadnis/haunted-places.

In this assignment we will explore several of the topics discussed in the early portion of class – Big Data – MIME types and their taxonomy – Data Similarity – and so forth. To do this, we will leverage the dataset highlighted in Figure 1 – a set of 21,983 reported mysterious haunted places. The posts contain several features which are highlighted below:

- **City** – the city in which the haunted place resides in.
- **Country -** The country where the place is located (always "United States").
- **Description -** A text description of the place. *The amount of detail in these descriptions is highly variable*.
- **Location** - A title for the haunted place.
- **State** - The US state where the place is located.
- **State_abbrev** - The two-letter abbreviation for the state.
- **Longitude** - Longitude of the place.
- **Latitude** - Latitude of the place.
- **City_longitude** - Longitude of the city center.
- **City_latitude** - Latitude of the city center.

The Haunted Places dataset is a rich dataset with high variation in its features and properties. For example, as can be seen from Fig 2., there are 4,386 unique cities, and 9,904 unique

locations in the dataset. These locations are always somewhere in the United States mainland, including Alaska and Hawaii.

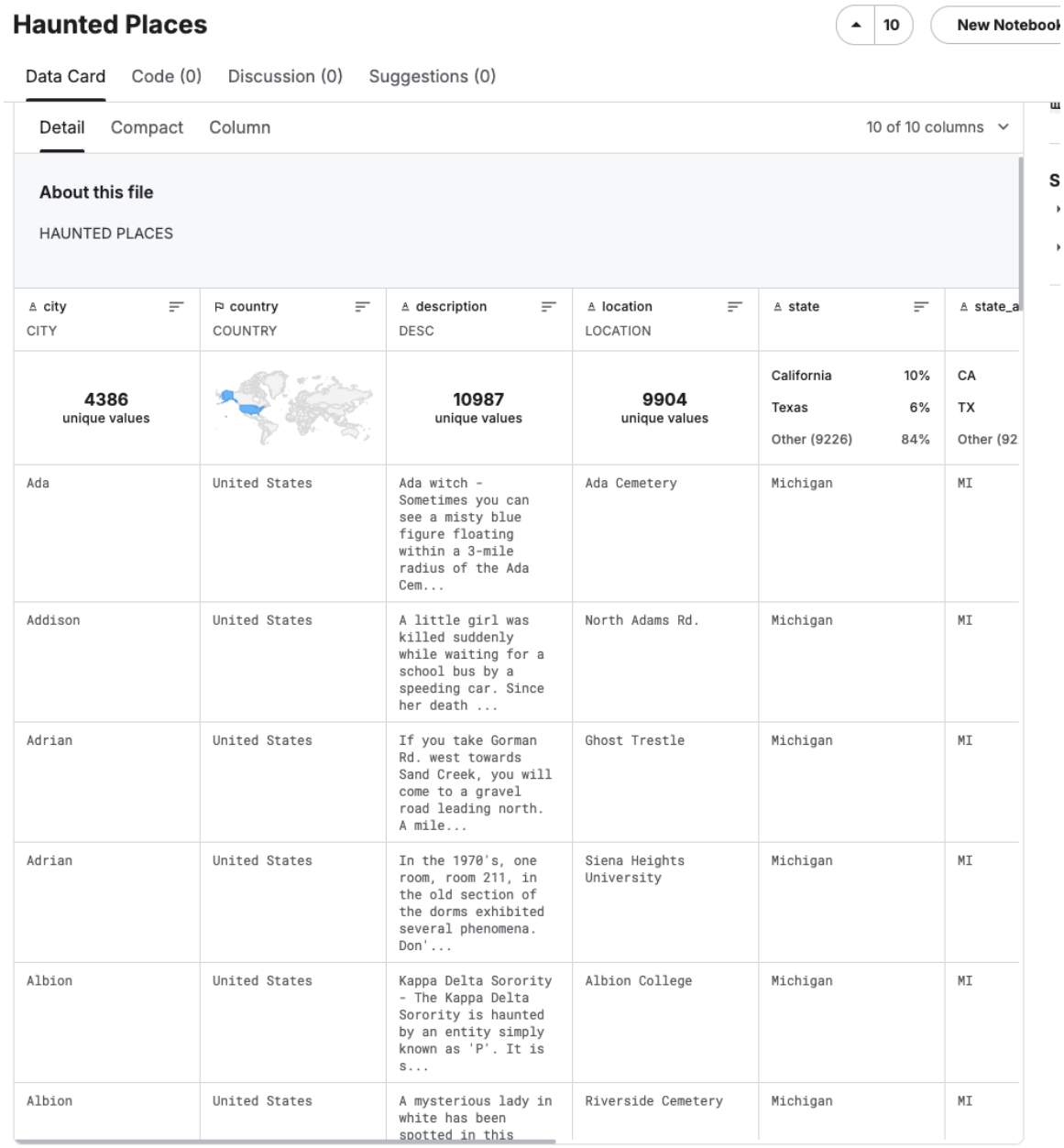

Figure 2. Distribution of features in the Haunted Places dataset.

One of the other important elements of the Haunted Places dataset is the description of the sighting. There are roads, streets, homes, commercial buildings, and other potentially haunted areas, with high variation. The descriptions of the haunted places gives the reader some inclination as to the modality of why the place is haunted, such as “If you take a camera…[sic] and take pictures, you’ll see orbs everywhere”, or “during the constructions of the train trestle, two men were found inexplicably hanging by suspension cords used to lift steel beams. Passersby have noted the sound of snapping necks…”, or

“This street had so much activity that the neighbors were comparing stories that at night they experienced that beings were laying down on top of them”, and so on.

These descriptions have rich modalities in the way that they were observed (sound, sight, smell, etc.) Additionally, some describe events that occurred (murder?) whereas others describe supernatural objects inhabiting dark areas (“orbs” near trees). Others describe eye witness accounts of what happened, and others even more so temporal characteristics that influence the evidence, such as happening during the day; or at night, or early evening. Some are repeating events, while others occurred once only, maybe twice. And some of the events have multiple witnesses whereas some are described as a story from a singular witness or an “ancient tale”.

Geographic areas, proximity to city center are also useful features. Did this sighting occur far out in the woods away from a city center where people are? Does that affect the witness reports? Or was this an event that happened right in metropolitan area. How should that effect how you weigh the evidence?

Take the cardinality of witnesses for example. Are single witness sightings more likely to have a follow up performed? You could use the number-parser Python tool to extract out the numerical value of number of witnesses: https://github.com/scrapinghub/number-parser as an example.

Additional insight can be gleaned from exploring the temporal properties of sighting, which are semi-structured and only present in the description. However, using a tool such as https://github.com/akoumjian/datefinder you can easily discern if there is information about when these sightings occurred provided the information is extractable. Are there particular days of the week with sightings? What happens if you slice down sightings related to geographic area *and* date?

2. Objective

Exploring the Haunted Places dataset may lead you to ask several questions to build up a profile of evidence related to the event. You can group these questions into the following characteristics:

Haunted Place Modality

- Evidence
  - audio evidence? (As discerned by “noises”, “sound of snapping neck”, “nursery rhymes”)
  - image/video or visual evidence? (As discerned by keywords in the description like “cameras” and “take pictures”, “names of children written on walls”, etc.)
- Is it haunted due to some event? (“When the railroad was built, <something> happened”)

Temporal characteristics
- Can time of day be discerned? Evening, afternoon, morning?
- Can date be discerned? (Year, month, day?)
- If the date/time is not directly present, can you find it by searching through description or location name on Google?
    - If not, perhaps set year to 2025, day to 01 and month to 01

Apparition type
- Is it a ghost?
    - Is there a ghost description (male/female/child?)
- Is it an orb? Unidentified Flying Object (UFO)? Unidentified Aerial Phenomena (UAP)?

Event type
- Did someone die here? Murder? Natural?
- How many people?
- How many witnesses?

Figure 3: Evidentiary questions you could ask about the Haunted Places dataset, deriving features that aren't initially present.

What ways could you explore this, by looking beyond this rich dataset and by applying lessons learned from class thus far where we have been studying the 5 V's, MIME types of associated datasets, and large datasets and their characteristics? Could you join for example, the Alcohol abuse by state data from https://drugabusestatistics.org/alcohol-abuse-statistics/ and then determine if there is correlation between population level abuse of alcohol and the presence of these Haunted places? What about a dataset that gave you information about visibility at that city/state/location during all times of the year to determine the quality of the visual evidence and observations taken by the reporters of these Haunted places? There are multiple datasets providing evidence of amount of daylight per state, such as https://www.timeanddate.com/astronomy/usa and https://aa.usno.navy.mil/data/Dur_OneYear.

What about police reports? Does the amount of police reports in a community have any affect as to the connection between the community and murders and other violent crime that may suggest a correlation between these supernatural events? For example you could look at violent crime and other police reports by state with this dataset, https://projects.csgjusticecenter.org/tools-for-states-to-address-crime/50-state-crime-data/.

And finally, recently, a website TarotCards.io, recently studied and revealed the most Haunted states by Google search volume and published the following statistics:

Key Findings:

Most Haunted State (Total Search Volume): California takes the top spot with 2,860 average monthly searches related to hauntings, followed by Texas (2,680) and New York (2,150).
Least Haunted State (Total Search Volume): Alaska ranks last, with just 570 searches per month.
Most Haunted State (Per Capita): Wyoming leads the way, with 98.89 searches per 100,000 residents, followed closely by Vermont (92.62) and North Dakota (81.12).
Least Haunted State (Per Capita): California is the least haunted by search interest when adjusted for population size, with only 7.25 searches per 100,000 people.

The Most Haunted States by Total Search Volume

California – 2,860 searches
Texas – 2,680 searches
New York – 2,150 searches
Florida – 2,080 searches
Ohio – 2,070 searches

The Least Haunted States by Total Search Volume

46. Hawaii – 650 searches
47. North Dakota – 640 searches
48. Vermont – 600 searches
49. Wyoming – 580 searches
50. Alaska – 570 searches

The Most Haunted States Per Capita (Per 100,000 Inhabitants)

Wyoming – 98.89 searches
Vermont – 92.62 searches
North Dakota – 81.12 searches
Alaska – 77.71 searches
South Dakota – 77.52 searches

The Least Haunted States Per Capita (Per 100,000 Inhabitants)

46. Pennsylvania – 13.59 searches
47. New York – 10.86 searches
48. Florida – 9.05 searches
49. Texas – 8.56 searches
50. California – 7.25 searches

The full dataset can be found here: https://docs.google.com/spreadsheets/d/1-ok5MWfRfGpO2nJkL3zcyDaw-PEEFKGOgiynJ3YiXS0/edit?gid=1333584719#gid=1333584719.

What other features can you think of that would be useful to join to the Haunted places dataset?

You will choose at least three additional publicly accessible datasets along these lines to join the Haunted Places dataset to, and you must add at least three new features per dataset that you join. The datasets you select may not all belong to the same MIME top level type – that is – you must pick a different MIME top level type for each of the three datasets you are joining to this Haunted Places dataset.

Once the data is joined properly, you will explore the combined dataset using Apache Tika and an associated Python library called Tika-Similarity. Using Tika Similarity, you can evaluate data *similarity* (as discussed during the Deduplication lecture in class; and also during data forensics discussions). Tika similarity will allow you to explore and test different distance metrics (Edit-Distance; Jaccard similarity; Cosine similarity, etc.). And it will give you an idea of how to cluster data, and finally it will let you visualize the differences between different clusters in your new combined dataset. So, you can figure out how similar Haunted Places are, given their locations, descriptions, witnesses, apparition types, geographic proximity to certain towns, and other features, and explore your new augmented dataset. For example, you may ask, how many Haunted places were actually ghost sightings that occurred with low light visibility in the Fall season and if the sightings were nearby any airports and included reports from at least two witnesses?

The assignment specific tasks will be specified in the following section.

3. Tasks

1. Download and install Apache Tika
    a. Chapter 2 in your book covers some of the basics of building the code, and additionally, see https://tika.apache.org/2.9.1/index.html
    b. Install Tika-Python, you can pip install tika to get started.
        i. Read up on Tika Python here: http://github.com/chrismattmann/tika-python
2. Download the Haunted Places dataset
    a. We will provide you a Dropbox link in Slack for each validated team
    b. Make a copy of the original dataset (because you are going to modify/add to it in this assignment)
3. Create a combined TSV file for your Haunted Places dataset
    a. Convert the CSV to TSV (here's a simple example of how to do this with Python https://unix.stackexchange.com/questions/359832/converting-csv-to-tsv)
4. Add and expand the dataset with the following features
    a. Add a new feature called "Audio Evidence". Set it to True if you match text like "noises", "sound of snapping neck", "nursery rhymes", False otherwise.
    b. Add a new feature called "Image/Video/Visual Evidence". Set it to True if you match text like "cameras" and "take pictures", "names of children written on walls", etc.

c. Add a new feature called "Haunted Places Date" and use https://github.com/akoumjian/datefinder on the Description text to pull out dates. If you can't find them, set the current date to 2025/01/01.
d. Add a new feature called "Haunted Places Witness Count" and use Number Parser : https://github.com/scrapinghub/number-parser to obtain the number of witnesses (if possible to identify in the description). If unable to identify, set count to 0.
e. Add a new feature called "Time of Day", and try to discern "Evening", "Morning", or "Dusk" from the text. If not discernable, set to "Unknown".
f. Add a new feature called "Apparition Type" and discern from the description if it is a "Ghost", "Orb", "UFO", UAP", or what type of apparition the Haunted place is inhabited by. Perhaps it is a "Male", "Female", "Child", or "Several Ghosts". Parse this from the description and set to "Unknown" if not discernable.
g. Add a new feature called "Event type". Was it a murder? Did someone die here? Was it a supernatural phenomenon? Discern this by parsing the "Description" text and searching for keywords.
h. Join the Alcohol Abuse by State dataset, here https://drugabusestatistics.org/alcohol-abuse-statistics/.
i. Join the amount of daylight by state dataset, here:
    i. https://www.timeanddate.com/astronomy/usa
    ii. https://aa.usno.navy.mil/data/Dur_OneYear

5. Identify at least three other datasets, each of different top level MIME type (can't all be e.g., text/*)
    a. Check out places including: https://catalog.data.gov/dataset (Data.gov)
    b. For each dataset, develop a Python program to join the data to your new Haunted Places dataset
    c. For each non text/* dataset, be prepared to describe how you featurized the dataset
    d. Each dataset that you join must contribute at least three features (in addition to the features you are adding described in part 5)
    e. For each feature you add, be prepared to discuss what types of queries it will allow you to answer and also how you computed the feature
6. Download and install Tika-Similarity
    a. Read the documentation
    b. You can find Tika Similarity here (http://github.com/chrismattmann/tika-similarity)
    c. You will also need to install ETLLib, here (http://github.com/chrismattmann/etllib)
    d. Convert the TSV dataset into JSON using ETLLib's tsv2json tool
    e. Compare Jaccard similarity, edit-distance, and cosine similarity using Tika Similarity
    f. Compare and contrast clusters from Jaccard, Cosine Distance, and Edit Similarity – do you see any differences? Why?
    g. How do the resultant clusters generated highlight the features you extracted? Be prepared to identify this in your report.

7. Package your data up by combining all of your new JSONs with additional features into a single TSV (tab separated values) file where the columns represent the features and the rows are the instances of your sightings.
8. (**EXTRA CREDIT**) Add some new D3.js visualizations to Tika Similarity
    a. Currently Tika Similarity only supports Dendrogram, Circle Packing, and combinations of those to view clusters, and relative similarities between datasets
    b. Download and install D3.js
        i. Visit http://d3js.org/
        ii. Review Mike Bostock's Visual Gallery Wiki
        iii. https://github.com/mbostock/d3/wiki/Tutorials
        iv. Consider adding
            1. Feature related visualizations, e.g., time series, bar charts, plots
            2. Add functionality in a generic way that is not specific to your dataset
            3. See gallery here: https://github.com/d3/d3/wiki/Gallery
            4. Contributions will be reviewed as Pull Requests in a first come, first serve basis (check existing PRs and make sure you aren't duplicating what some other group has done)

4. Assignment Setup

4.1 Group Formation

You can work on this assignment in groups sized at **minimum 2, and maximum 6**. You may reuse your existing groups from discussion in class. Please fill out the group details in the form provided after class. Only one form submission per team. If you have any questions, contact your TA/Course Producer via their email address with the subject: DSCI 550: Team Details.

4.2 Haunted Places dataset

Access to the data is provided by a Dropbox link. The dataset itself is approximately 5.4Mb unzipped. You may want to distribute the data between your team-mates since the data is fairly small (for now).

4.3 Downloading and Installing Apache Tika

The quickest and best way to get Apache Tika up and running on your machine is to grab the `tika-app.jar` from: http://tika.apache.org/download.html. You should obtain a jar file called `tika-app-2.9.1.jar`. This jar contains all of the necessary dependencies to get up and running with Tika by calling it your Java program.

Documentation is available on the Apache Tika webpage at http://tika.apache.org/. API documentation can be found at http://tika.apache.org/2.9.1/api.

Since you will be using Tika Python, you will want to read up on the Tika REST API, here: https://cwiki.apache.org/confluence/display/TIKA/TikaServer. The Tika Python library is a robust REST client to the Java-side REST API.

You can also get more information about Tika by checking out the book written by Professor Mattmann called "Tika in Action", available from: http://manning.com/mattmann/.

5. Report

Write a short 4-page report describing your observations, i.e. what you noticed about the dataset as you completed the tasks. What questions did your new joined datasets allow you to answer about the Haunted Places data and its sightings and additional features previously unanswered? What clusters were revealed? What similarity metrics produced more (in your opinion) accurate groupings? Why? What did the additional datasets suggest about "unintended consequences" related to Haunted Places? You should also clearly explain which datasets you used to join the Haunted Places and how you extracted the new features from each dataset.

Thinking more broadly, do you have enough information to answer the following:

1. Are there clusters of Haunted Places with similar features, and all are murders occurring in the evening?
2. Does the time of day of the Haunted Place original sightings matter?
3. Are specific locations more likely to be influenced by alcohol abuse that cause more Haunted Places to be reported?
4. Are specific keywords bigger indicators of the apparition type related to a Haunted Place?
5. Is there a set of frequently co-occurring features that define a particular Haunted Place?
6. What insights do the "indirect" features you extracted tell us about the data?
7. What clusters of Haunted Places made the most sense? Why?

Also include your thoughts about Apache Tika – what was easy about using it? What wasn't?

Note: Report should be written using 11 pt Times New Roman font, single column with single spacing.

6. Submission Guidelines

This assignment is to be submitted ***electronically, by 12pm PT*** on the specified due date, via Gmail dsci550.2025a@gmail.com for the Thursday class, or dsci550.2025b@gmail.com for the Tuesday class. Use the subject line: DSCI 550: Mattmann: Spring 2025: BIGDATA Homework: Team XX. So, if your team was team 15, and you had the Thursday class, you would submit an email to dsci550.2025a@gmail.com with the subject "DSCI 550: Mattmann: Spring 2025: BIGDATA Homework: Team 15" (no quotes). **Please note only one submission per team**.

- All source code is expected to be commented, to compile, and to run. You should have at least a few Python scripts that you used to join three other datasets, and what you used to extract additional features.
- Use relative paths {not absolute paths} when loading your data files so that we can execute your script/notebook files without changing everything.

- If using a notebook environment, use markdown cells to indicate which tasks/questions you are solving.
- Include your updated dataset TSV. We will provide a Dropbox or Google Drive location for you to upload to {you don't need to attach it inside the zip file}.
- Also prepare a readme.txt containing any notes you'd like to submit.
- If you used external libraries other than Tika Python and Tika Similarity, you should include those jar files in your submission, and include in your readme.txt a detailed explanation of how to use these libraries when compiling and executing your program.
- Save your report as a PDF file (TEAM_XX_BIGDATA.pdf) and include it in your submission.
- Compress all of the above into a single zip archive and name it according to the following filename convention:

TEAM_XX_DSCI550_HW_BIGDATA.zip

Use only standard zip format. Do **not** use other formats such as zipx, rar, ace, etc.

- If your homework submission exceeds the Gmail's 25MB limit, upload the zip file to Google drive and share it with dsci550.2025a@gmail.com (Thursday class) or dsci550.2025b@gmail.com (Tuesday class).

When submitting, please organize your code and data file as the directory structure shown:

```
Data
      dataset1 {leave it empty for now}
Source Code
      script1
      notebook
Readme.txt
Requirements.txt
```

Important Note:

- Make sure that you have attached the file when submitting. Failure to do so will be treated as non-submission.
- Successful submission will be indicated in the assignment's submission history. We advise that you check to verify the timestamp, download and double check your zip file for good measure.
- Again, please note, only **one submission per team**. Designate someone to submit.

6.1 Late Assignment Policy

- -10% if submitted within the first 24 hours
- -15% for each additional 24 hours or part thereof