Final Project Data Sheet

Motivation

UFO Sightings 2015

- Why was this dataset created?
  - To keep track of UFO sightings in the United States
- Was there a specific task in mind?
  - For people in these areas to self-report UFO sightings so there can be a record maintained of unusual, possibly UFO-related events. The primary reporting mechanism is a telephone hot-line. There is also guaranteed anonymity to callers.
- Who created this dataset and on behalf of which entity?
  - Creator: National UFO Reporting Center
  - Publisher: National UFO Reporting Center
  - Publisher of Record: Wolfram Research
- Who funded the creation of the dataset?
  - The National UFO Reporting Center is a non-profit Washington State corporation and it is funded through subscription revenues from its monthly replica watches uk newsletter, sales of its video tapes, sales of general information packets, and honoraria for public presentations.

US Military Bases

- Why was this dataset created?
  - To have more data used to protect the US and have information of our bases.
- Was there a specific task in mind?
  - To have a record of information on US military bases.
- Who created this dataset and on behalf of which entity?
  - The Military Bases dataset is part of the U.S. Department of Transportation (USDOT)/Bureau of Transportation Statistics's (BTS's) National Transportation Atlas Database (NTAD).
- Who funded the creation of the dataset?
  - The US Government.

US Airports

- Why was this dataset created?
  - Crowdfunded by people for airport location
- Was there a specific task in mind?
  - To have a place to store information about airports so people could have access for studies, info, etc.
- Who created this dataset and on behalf of which entity?
  - Crowdfunded community data

- Who funded the creation of the dataset?
    - It was crowdfunded

US Population Density

- Why was this dataset created?
    - To contribute to the US Census
- Was there a specific task in mind?
    - See above
- Who created this dataset and on behalf of which entity?
    - United States Census
- Who funded the creation of the dataset?
    - United States Census Bureau

Composition

UFO Sightings 2015

- What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?
    - UFO Sightings and types of instances/interactions between them-
        - Date (month, day, hour, minute, second)
        - City
        - State
        - Position (Latitude and Longitude)
        - Shape (of the UFO)
        - Duration (Seconds, Minutes, Hours, Days, Weeks, Months)
        - Summary (description of the UFO sighting)
- How many instances are there in total (of each type, if appropriate)?
    - 4,163 instances in total
- Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
    - Data set contains all possible instances
- What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features?
    - Date (month, day, hour, minute, second)
        - In a JSON-type date object
    - City
        - In a JSON-type entity object
    - State
        - In a JSON-type entity object
    - Position (Latitude and Longitude)
        - In a JSON-type GeoPosition object

  - Shape (of the UFO)
    - In a JSON-type Shape object
  - Duration (Seconds, Minutes, Hours, Days, Weeks, Months)
    - In a JSON-type duration object
  - Summary (description of the UFO sighting)
    - Unprocessed text
  -
- Is there a label or target associated with each instance?
  - Yes, see above.
- Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable).
  - There are some instances in the shape and duration columns that have missing input because it was unavailable.
- Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?
  - Yes, there are associated instances for each UFO sighting
- Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
  - No there are not.
- Are there any errors, sources of noise, or redundancies in the dataset?
  - The state is actually repeated in the city column, which is a redundancy.
- Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
  - Self-contained
- Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?
  - No, there are no instances of confidential date of those who reported these sightings anywhere in the dataset
- Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
  - No
- Does the dataset relate to people?
  - No

US Military Bases

- What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?
  - Military bases, locations, and sizes
    - Base Name
    - Base State
    - Base Country

        - Base Size
- Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
    - “This list does not necessarily represent a comprehensive collection of all Department of Defense facilities, and only those in the fifty United States and US Territories were considered for inclusion”
    - Not necessarily a complete set - likely for safety reasons but full sets do not exist for public use
- What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.
    - All text-
        - OBJECTID - Unique ID (int)
        - COMPONENT- Type of base (string)
        - SITE_NAME - Name of the military base (string)
        - JOINT_BASE - Name of joint base, otherwise ‘N/A’ (string)
        - STATE_TERR - States plus Puerto rico (string)
        - COUNTRY - US plus Puerto rico (string)
        - OPER_STAT - If the base is active (string)
        - SharpeSTArea - Area of base (long)
        - ShapeSTLength - Length of base (long)
- Are there any errors, sources of noise, or redundancies in the dataset? Missing values?
    - We used Object ID, and State_terr primarily, and there were no missing, erroneous, or inconsistent values in these fields. The only issues were the formatting of the state names including spaces which was not the same for other data sets.
- Confidentiality, ethical concerns, relate to people?
    - Does not relate to people and is an open database by the US government for public use

US Airports

• What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

- US airports and data about each airport
    - ID number
    - Type of airport
    - Name of airport
    - Latitude and Longitude
    - Elevation
    - Municipality

- GPS Code
- Local Code
- Keywords

• How many instances are there in total (of each type, if appropriate)?

- 22,768

• Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

- The dataset contains all possible instances of US airports, it is not a sample.

• What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

- ID number
- Type of airport -- large, medium, small, helicopter, seaplane
- Name of airport -- Full name of airport
- Latitude and Longitude -- coordinates
- Elevation -- the elevation of the airport in feet
- Municipality -- City where it is located
- GPS Code -- Airport code (KLAX for LA airport)
- Local Code -- Airport code that is more common (LAX for LA)
- Keywords -- keywords that are associated with airport

• Is there a label or target associated with each instance? If so, please provide a description.

- Yes, see above

• Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

- There are some instances in the keywords column where an airport does not have key words because there were no significant associated words with the airport.

• Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.

- Yes, there are relationships between all of the aspects because they are all part of one airport for each row.

• Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

- No

• Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

- The local and gps code are almost identical, and the name of the airport and municipality are very similar, which is a type of redundancy (although slightly different).

• Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

- Self contained

• Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

- No

• Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

- No

• Does the dataset relate to people? If not, you may skip the remaining questions in this section.

- No

US Population Density

- What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?
  - Per state: Population values, population density, housing density
- Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
  - Full set (census)
- What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
  - All text-
    - Numeric ID
    - State
    - Population
    - Housing Units
    - Area in sq miles
    - Population density

        - Housing Unit density
- Are there any errors, sources of noise, or redundancies in the dataset? Missing values?
    - No
- Confidentiality, ethical concerns, relate to people?
    - Does relate to people at a high level but no person is identifiable from the data. Presented on a state by state level.

Collection Process

UFO Sightings 2015

- How was the data associated with each instance acquired?
    - Through a telephone Hot-Line that The National UFO Reporting Center runs 24/7 and through an online UFO report form.
- Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)?
    - The data was reported by subjects over the phone and through a form online
- If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified?
    - According to their website, hoax and joke reports are ignored and not added to datasets.
- What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?
    - Human over the phone or a form online
- How were these mechanisms or procedures validated?
    - See above.
- Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
    - The hotline is staffed 24/7, however, it is unknown if they are compensated or not.
- Over what timeframe was the data collected?
    - The entirety of 2015
- Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?
    - Yes
- Were any ethical review processes conducted
    - Unknown

US Military Bases

- How was the data associated with each instance acquired?
    - Put together by the government - already has access to this information, and were not transparent about how it was compiled

US Population Data

- How was the data associated with each instance acquired?
    - Mandatory collection by the government every 10 years
- Over what timeframe was the data collected?
    - Collected in 2010 and projected for 2015
- What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?
    - Voluntary self-reporting and then manual labor / programs to fill gaps

US Airports

- How was the data associated with each instance acquired?
    - Through research of different airports and accumulation of data from each one by crowdsourcing
- Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)?
    - The data was reported by subjects through crowdsourcing from communities
- If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified?
    - According to their website, it is unverified community data
- What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?
    - People fill in the gaps of information voluntarily by creating and maintaining the data
- How were these mechanisms or procedures validated?
    - See above
- Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
    - Anyone with access to computer and information on airports, not compensated.
- Over what timeframe was the data collected?
    - From 2007 on.
- Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?
    - Yes
- Were any ethical review processes conducted
    - Unknown

Preprocessing/cleaning/labeling

UFO Sightings 2015

- Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?
  - The dataset information was put into JSON data objects in their respective columns as described above, but these were not cleaned and did not process missing values.

US Military Bases

- Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?
  - Organized into a CSV and mostly clean values, the main things that had to be changed later were to match across sets

US Airports

- Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?
  - Data from http://ourairports.com/countries/US/ was organized and formatted in csv into categories. The data however was not cleaned or processed for missing information.

US Population Density

- Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?
  - Yes, these values are predictions based on the 2010 census
  - Specific information about how this was done is no longer available on the website: see https://factfinder.census.gov/faces/affhelp/jsf/pages/metadata.xhtml?lang=en&type=table&id=table.en.PEP_2015_PEPANNRES#main_content

Uses

UFO Sightings 2015

- Has the dataset been used for any tasks already?
    - On Wolfram, there are examples of outputs like top 5 states and most common months, but other uses has been unknown. The National UFO Reporting Center publishes datasets every month and has these readily available to the public on their website.
- Is there a repository that links to any or all papers or systems that use the dataset?
    - No, there is not.
- What (other) tasks could the dataset be used for?
    - Information about conspiracies / people interested in sci-fy
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description.
    - No

US Military Bases

- Has the dataset been used for any tasks already?
    - Likely, for any analysis done on military bases in the US or military presence this would be a valuable and reliable database.
- Is there a repository that links to any or all papers or systems that use the dataset?
    - No, there is not.
- What (other) tasks could the dataset be used for?
    - Could be used to track presence of military in certain areas to see influence in those areas
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description.
    - Yes, that it is not necessarily a complete set
- Are there tasks for which the dataset should not be used?
    - Anything that could compromise safety

US Airports

- Has the dataset been used for any tasks already?
    - Possibly. The data from many of the airports has been viewed by over 100 people, but it is unclear whether they actually used it for any other tasks or not.
- Is there a repository that links to any or all papers or systems that use the dataset?
    - No, there is not.
- What (other) tasks could the dataset be used for?

    - The dataset could be used to plan trips, as there is location data and size of airport, which could be used to see what airports are close and approximate costs.
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description.
    - No
- Are there tasks for which the dataset should not be used?
    - Not particularly, maybe any task that has nothing to do with airports, travel, or populations. Alternatively, anything that doesn't have to do with things that could be guessed using this data, like population numbers (since larger airports=larger cities=more people)

US Population Density

- Has the dataset been used for any tasks already?
    - Yes definitely, any data analysis that needs population data by state since 2010 might use this set
- Is there a repository that links to any or all papers or systems that use the dataset?
    - No, there is not.
- What (other) tasks could the dataset be used for?
    - Any analysis done on a state level that needs to be adjusted based on population in each state
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description.
    - No