Northeastern University
CS 7295 – Information Visualization
March 2018, Abby, Emily, Lydia and Peter

Milestone 1 - UFOs

1 Data acquisition and clean-up

How dirty was the data and what kind of clean-up was required? How easy was it to download / use API? What is the format of the data? What are the items and attributes in your dataset? What types of data are you working with (categorical, ordinal, quantitative)?

We acquired the data easily from the Reporting Database of the National UFO Reporting Center's website. The reports are organized in tables which can be accessed by the following indices: Event Date, State, Shape, Date Posted. When organized by dates, the website provides one table for each month. We downloaded the tables for each month of 2017 into one excel sheet. Since the report form of the website mostly provides the user with a specific set of options for each question, the data set was relatively clean. The attributes of each item, i.e. report, are:

- **Date** (date and time of sighting) - quantitative
- **City** (city and/or area of the sighting) - categorical
- **State** - categorical
- **Shape** (one of the 21* choices for craft shapes) - categorical
- **Duration** (statement of estimated duration of the sighting in text) - quantitative
- **Summary** (written description and comments on the sighting)
- **Date Posted** (date the report was posted) - quantitative

We note that there are missing values in the attributes Shape and Duration. We only kept the reports regarding sightings in the US since there were only a few reports outside the US and we can assume that these areas would not be well represented. After this, we are left with 4593 reports. We fixed a few obvious mistakes on the states probably caused by miss-clicking the state close to the right one on the drop-down menu. We changed all the craft shapes with value "Unknown" to missing values, assuming that a missing value and an "Unknown" value essentially mean the same thing. Finally, since we intend to use Duration as a quantitative attribute, we used the information of the text to derive the duration of the sighting in minutes. For this last part, we used the mean of a range to give a single number of minutes and we had to make some assumptions as to what people perceive as "a few minutes" or "several hours" for example. We believe that these assumptions do not alter the data significantly and would not affect our visualizations.

2 Exploring the data

Any interesting trends/observations? Any evidence of missing or dirty data? Any features or observations in the data that are confusing or unexpected?

We explored our data in Tableau and Google Explore. We noticed that there are particular dates with a higher number of UFO sightings, such as Dec 22, Jul 4, Dec 9, and Jan 1. Some of these reports could have been caused by celebratory events associated with certain holidays. For the vast majority of the reports, the sighting's duration was at most one hour but in general the range of the durations is very large (one second to 20 years). In addition, most sightings refer to objects with "Light" or "Circle" shape and all 20 shapes (except for "Unknown", which we cleaned) are present. This will probably make it difficult for us to represent this categorical variable. Finally,

although the time of a sighting is almost a uniformly distributed variable, there were a lot of sightings occurring at exactly 3:00 am.

*Changing, Chevron, Cigar, Circle, Cone, Cross, Cylinder, Diamond, Disk, Egg, Fireball, Flask, Formation, Light, Other, Oval, Rectangle, Sphere, Teardrop, Triangle, Unknown

3 Interview

Write a paragraph: how did the interview go? What did you learn? What were you surprised by during the interview? Has the interview changed your motivating questions?

After some initial trouble communicating with the director of the National UFO database, and trouble with sourcing a line of communication, we had a solid half hour interview with Peter Davenpor [1], Director of the National UFO Reporting Center. Davenport discussed with us many things regarding his database, most notable of which was his role in establishing the data collection method that the Center uses now. Davenport modernized the collection and organization of the data, both retroactively and for use in the future. I was surprised by how much the tone and attitude of the interview subject changed over the course of the scheduling process and interview; initially, Davenport was skeptical of our use of his data and interview. During the interview, it came to light that this was due to him being burned before by major publications and national news outlets. Our interview was incredibly informative but I do not believe that it changed our line of questioning or reasoning in this project.

4 Task analysis

Create a full list of domain tasks (i.e., what are the tasks a user wants to accomplish with the data using your visualization) and translate these into high/mid/low level tasks.

4.1 Domain Tasks

High level:
1. Analyze - Consume - Discover, Present, Enjoy

Medium level:
1. Search

Low level:
1. cluster - is there a cluster of UFO sightings?
2. filter - what years have UFO sightings?
3. find extremum - what state has the most UFO sightings?
4. correlate - is there a trend of increasing or decreasing UFO sightings over the years?

4.2 Sketches

[1] Interview Notes: `https://docs.google.com/document/d/1ejWb30b9zhYVn6H0adOAkQ_ZRSgHtJb8noliHprtfIw/edit`