John Senft

Course Project

DS 350: Methods for Data Analysis

March 14, 2017

Introduction

For this project, I started with a Kaggle file from National UFO Reporting Center (NUFORC) covering reported UFO sightings from 1949 through December 2016. Then I collected Twitter posts containing the string “ufo” from mid-February through March 13, 2017. Generally, I grabbed 1000 tweets every couple of days, amounting to about 4000 a week for the 5 weeks. The goal is to see if the NUFORC sighting reports are a predictor of Twitter UFO postings or not.

Method / Results

I built a data frame from the NUFORC sighting data and then aggregated the information by number of sightings per state over the entire time frame. From this, California had the most reported sightings at 9655, followed by Washington with 4268 (see Table 1, Figure 1 & 2). I also thought it would be insightful to normalize the data by taking a look at sightings as a percentage of state populations. I pulled in census data from 2010 and filtered NUFORC data for sightings from 2010 on. Vermont, Montana, Alaska, Washington, and Maine were the top 5 states with reported sightings as a percentage of state population (see Figure 3).

Starting in mid-February, I used the Twitter API to pull in tweets using the search term: “ufo”. After filtering out re-tweets and removing duplicates, I ended up with 17927 tweets for this project last collected on March 13, 2017. I originally considered using geolocations contained in the tweets to find where sightings came from but it turns out that very few tweets have this information. To mitigate this issue, I searched the text of these tweets for state names or abbreviations. After viewing many of these tweets, it was clear that if the text described a sighting, it also had some indication of the location of the sighting. Using this, I was able to tally up the number of times a state name appeared in these tweets by pulling the text into a word corpus. Each tally was considered a Twitter sighting for this project. Of the 17927 tweets, 2007 state sightings were found using this method. Some tweets claimed sightings for non-US locations. Those were not used here.

Once state names were substituted into the tweet texts, I created a corpus from it to get all of the word counts. After preprocessing and conversion to a document term matrix, I was able to sum the word columns and remove states and counts against a vector containing state names which yielded a total count for all 50 states (see Table 2 & Figure 4). Indiana had the highest count of sightings, which stands out very clearly in the graphic depicting tweet sighting positive counts (see Figure 5).

Since this project comprises two discrete data sets of counted occurrences, I chose using a Chi-squared test to compare them for “goodness of fit.” Using the NUFORC data from 2010 on, I calculated percentages for each state of the total number of sightings (23898). These were used as the expected probabilities for the Twitter counts of sightings. Running the Chi-squared test on Twitter sighting counts produced a minuscule p-value of 2.2e-16. Since the p-value is very small, we reject the null hypothesis that the NUFORC data is a predictor of UFO sighting tweets.

Since the degrees of freedom for the Chi-squared test is 49, it was very likely that the test would fail to support the null hypothesis. An improvement to this analysis might be to group the states into regions and re-run the analysis with far fewer degrees of freedom. Additionally, some discussion of error should be made. During processing of the tweet text, substitutions were made for state abbreviations. I searched the text for “OR”, “OK”, “ME”, and “IN” (surrounded by spaces) and swapped with full state names to avoid having the words “in”, “me”, “ok”, and “or” falsely tagged as a sighting. Obviously, this is imperfect, and the fact that Indiana had the most sighting positives by a wide margin, does raise suspicion, and should be examined further. Also, user typos in the state spelling would not have been counted as a sighting positive, and some tweets may have more than one state name contained in the text string which would have resulted in incorrect counts.

Data Sources:

\* https://www.census.gov/library/publications/2011/compendia/statab/131ed/population.html

\* https://www.kaggle.com/NUFORC/ufo-sightings

\* http://www.fonz.net/blog/archives/2008/04/06/csv-of-states-and-state-abbreviations/

\* Twitter search via twitter api

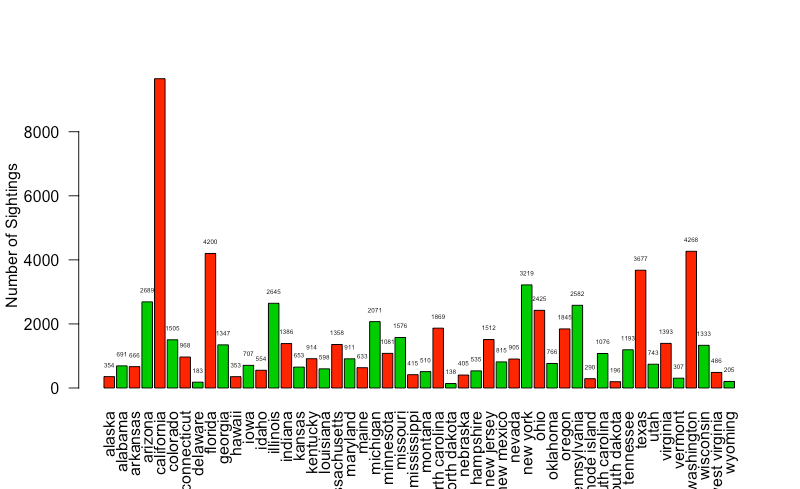
Figure 1: UFO Sightings per state (1949 -2016) according to NUFORC data

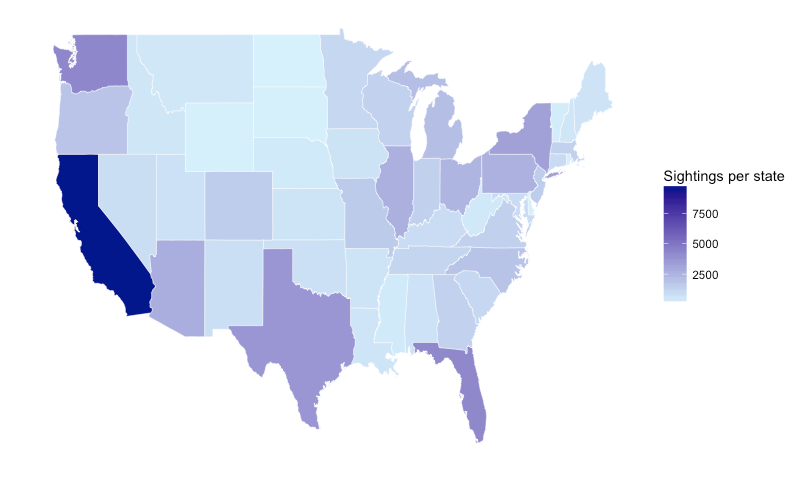
Figure 2: Heatmap of UFO Sightings Reported to NUFORC (1949 - 2016)

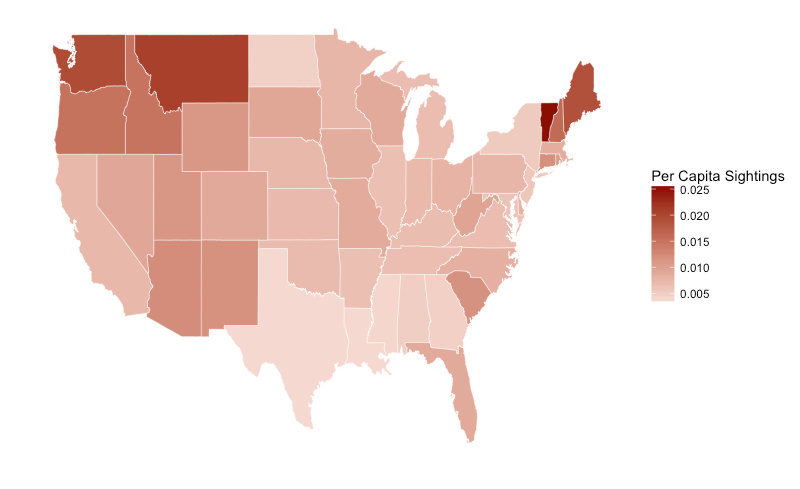
Figure 3: Heatmap of UFO Sightings as Percent of State Population since 2010

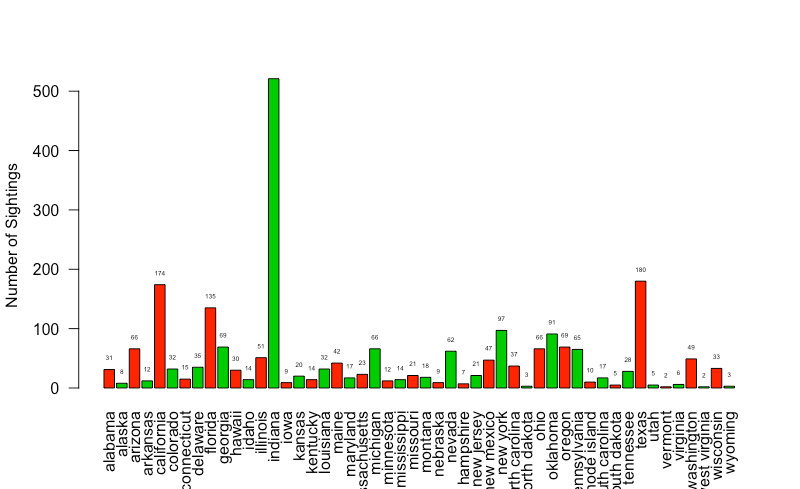
Figure 4: Number of state matches found in tweet text searched by “ufo” (Feb/March 2017)

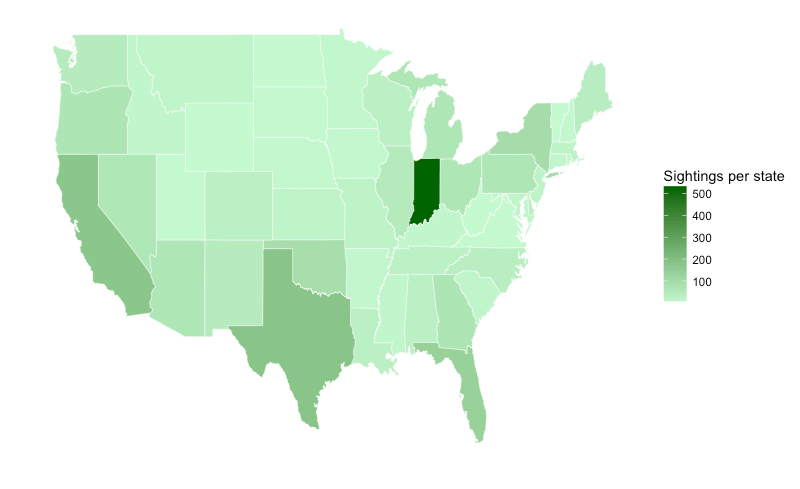
Figure 5: Heatmap of “ufp” tweet counts per state (Feb/March 2017)

Table 1: Top 5 UFO Sightings per state (1949 -2016) according to NUFORC data

|  |  |
| --- | --- |
| California | 9655 |
| Washington | 4268 |
| Florida | 4200 |
| Texas | 3677 |
| New York | 3219 |

Table 2: Top 5 UFO Sightings per state (Feb/March 2017) from Twitter search

|  |  |
| --- | --- |
| Indiana | 521 |
| Texas | 180 |
| California | 174 |
| Florida | 135 |
| New York | 97 |