Analysis of UFO Sighting Reports within the Last Century

Noelle Cafone

December 2, 2017

EM 624 – Informatics for Engineering Management

Instructor: Dr. Carlo Lipizzi

Project Goals and Conditions

The mystery of unidentified flying objects has existed in culture for the past century. Sightings of apparent objects in the sky that are not identifiable as planes or stars or any presently known aircraft are typically believed to be extraterrestrial space crafts. People all over the world report UFO sightings due to this belief. During the 1940s and 1950s, UFOs were referred to as "flying saucers", and the United States Air Force coined the term UFO in 1953. During the Cold War, UFOs became a popular interest due to increased concern for national security.

Since UFO sightings have been reported worldwide for several decades, the goal of this project is to analyze where UFO sightings are the most common and what types of objects are sighted. This analysis can help to gain further understanding of what the unidentified flying objects might be based on the location and the shape of the object. This project is limited by the data file containing the reports of UFO sightings since it contains some empty cells and might not contain the most updated information. This project is also has a time constraint for the due date of the project which is the final week of the course.

Business Understanding

The business requirements are to utilize accurate, reliable data and to obtain visual representations of the analysis. The business objective of this analysis are to determine which country has the most reports of UFO sightings, which can then be narrowed down to determine the areas of the country that are most likely to have UFO sightings. The United States government could use this information to see if there is any validity to the popular beliefs of extraterrestrial space crafts, as well as to determine if the objects are a threat to national security. Another objective is to determine if there is any correlation between the most common shapes reported overall and the most common shapes reported for each of the top cities. The government could use this information to further their investigation of the unidentified flying objects. In order to achieve these objectives, the data will be mined to create a list of the top countries in the UFO sighting reports, a list of the top cities in the UFO sighting reports, and a list of the most common shapes observed. This information can be graphed for visual representation that can help discover correlations in the data. The data can also be narrowed down to just the top cities so that the most common shapes in those cities that were sighted can be analyzed to see if they are in harmony with the most common shapes reported overall. This will allow for the government to determine if specific objects are observed over specific locations.

Data Understanding

The dataset "scrubbed.csv" found on kaggle.com titled "UFO Sightings" is a csv file containing reports of unidentified flying objects within the last century. The file was renamed as "ufo.csv" and contains columns of the date and time, the city, the state, the country, the shape of the object, the duration, the comments, the date the report was posted, the latitude, and the longitude of the UFO sighting. This dataset contains 80,333 lines. It was posted on Kaggle by the National UFO Reporting Center, but it was last updated a year ago so it is not as up to date as possible. However, it was scraped, geolocated and time standardized from NUFROC data. Although the scrubbed file supposedly does not contain any incomplete reports, some blank cells are found. However, the data seems to be reliable and a sufficient quality overall.

Data Preparation

First, the file will be opened as a pandas data frame in order to structure the data into an easy to analyze table, and therefore "pandas" was imported as "pd". The data frame is titled "ufo_file". Then, any rows containing empty cells will be removed. It was observed that there are many blank cells in the 'country' column, so the incomplete rows are removed from that column. This is done by importing "numpy" as "np" and using a replace feature, "ufo_file['country'].replace('',np.nan, inplace=True)" and dropping the rows containing the blank entries, "ufo_file.dropna(subset=['country'],inplace=True". This completes the data transformation that is required to begin the analysis of the data. The data frame will now contain complete entries for the countries and cities that will be analyzed to determine the most common locations.

Data Representation

In order to meet the business need of analyzing the areas of the country that are most likely to have UFO sightings, a list of the countries is created from the 'country' column in the file. Then, the Counter function is used to count the frequency of each country in the column, and the five most common countries are calculated using "Counter(CountriesList).most_common(5)". Next, the top five countries are structured into a pandas data frame and the columns are labeled to display the country and the frequency. The data frame is printed and the output can be seen below:

The top 5 countries with reports of UFO sightings are:

	Country	Frequency
0	us	65114
1	ca	3000
2	gb	1905
3	au	538
4	de	105

To obtain a visual representation of this information, a bar plot is created for the top five countries, which is listed under Figure 1.

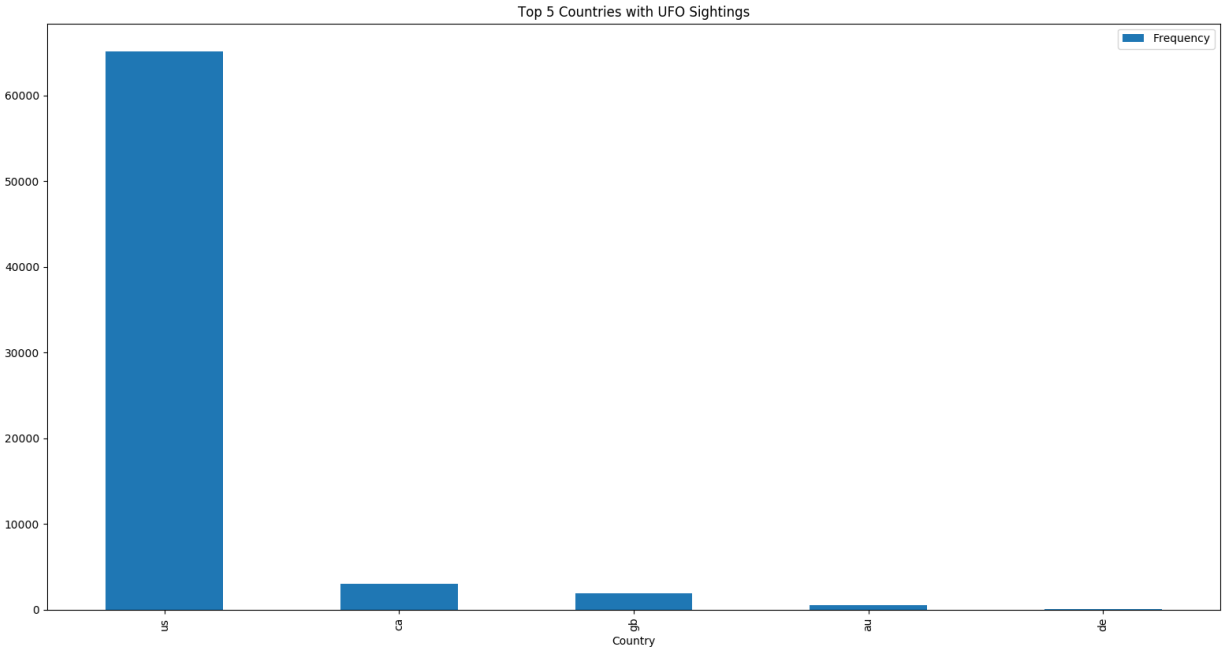

Figure 1: Top 5 Countries Bar Plot

A pie chart is also created to further emphasize that the United States has a significantly higher amount of UFO sighting reports, seen in Figure 2.

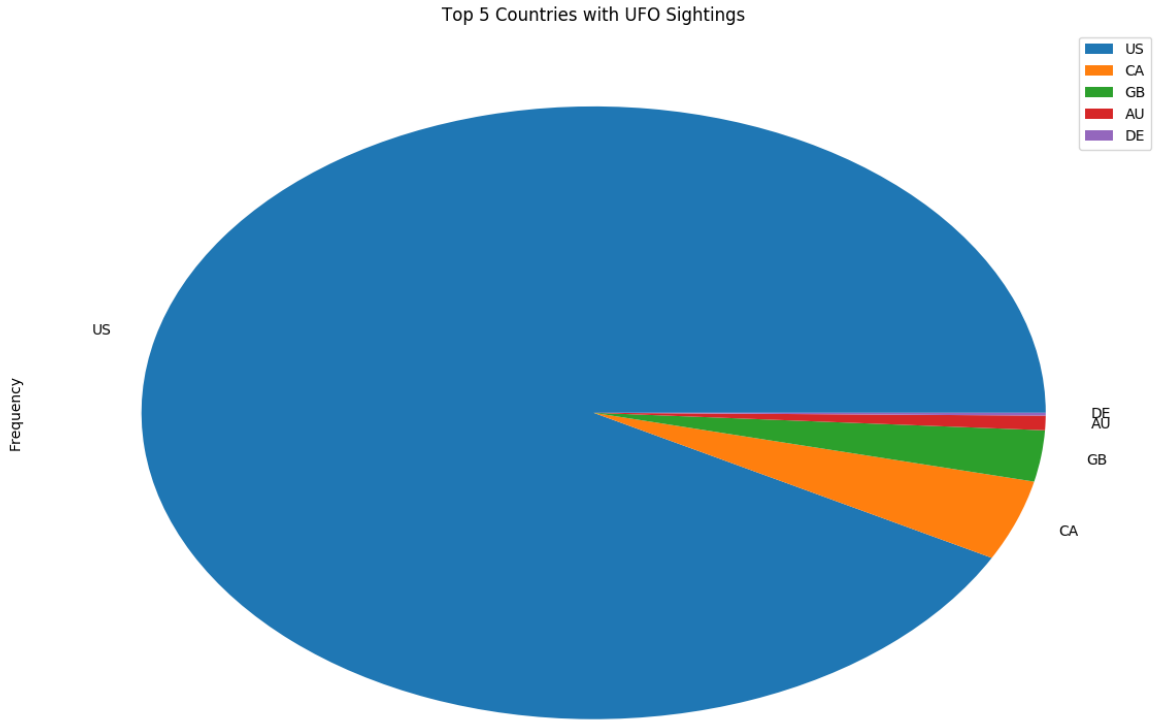

Figure 2: Top 5 Countries Pie Chart

Next, a similar analysis is done on the top cities with reports of UFO sightings. The top five cities in the dataset are all in the United States, so the analysis can focus on the country with the most UFO reports and narrow down to the top cities in order to determine a common location. A list is created of the entries in the 'city' column in the file. The top cities are then calculated in the same way as the top countries, by counting the frequency of each city in the list and then taking the five most common cities. This data is then structured as a pandas data frame, with columns labeled 'City' and 'Frequency'. The script gives the output as follows:

The top 5 cities with reports of UFO sightings are:

	City	Frequency
0	seattle	524
1	phoenix	454
2	portland	373
3	las vegas	367
4	los angeles	352

Visual representation of this data is also graphed to aid the analysis of the top cities with reports of UFO sightings, as seen in Figures 3 and 4.

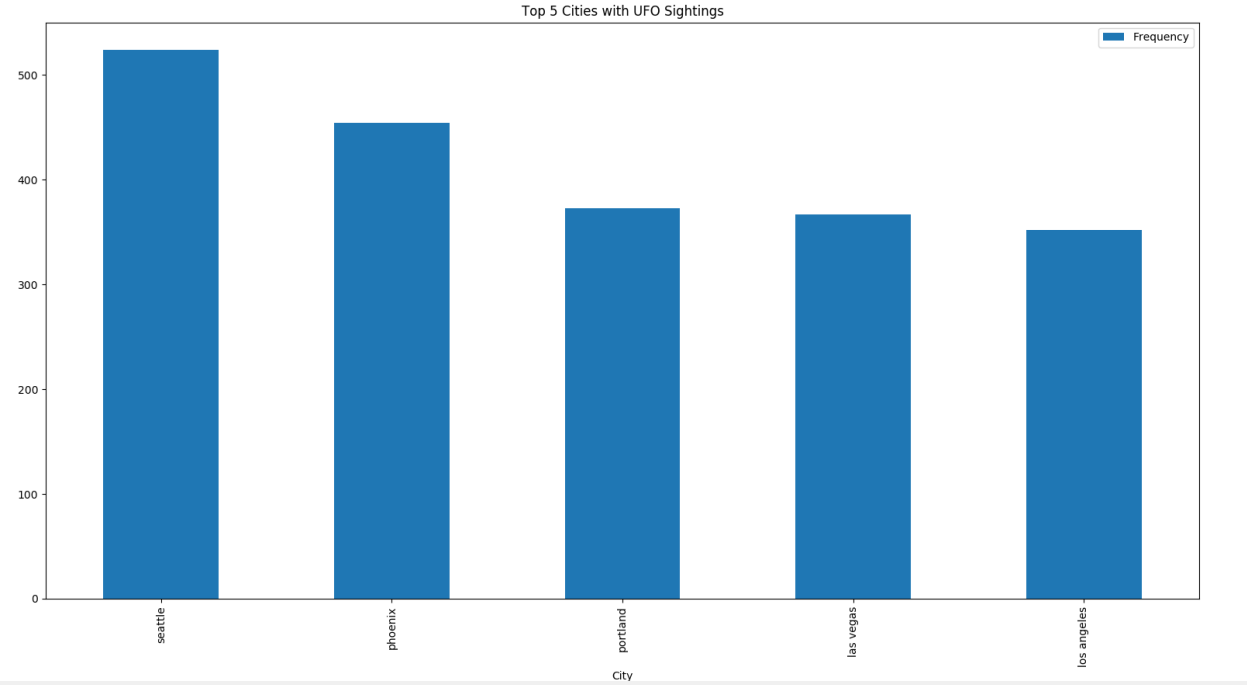

Figure 3: Top 5 Cities Bar Plot

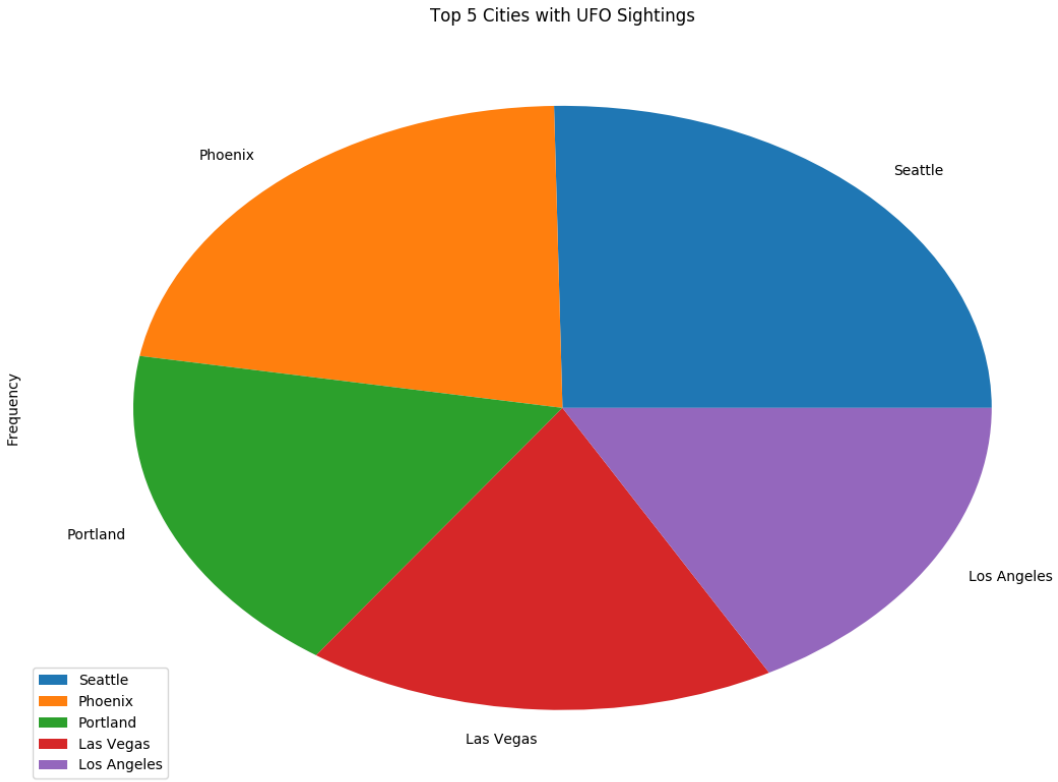

Figure 4: Top 5 Cities Pie Chart

Since the top five cities are all relatively similar in frequency, they will each be analyzed for the most common shapes observed at these locations. First, the top 15 shapes in the entire dataset are determined by counting the values of each shape, which lists the shapes in order of frequency, and then printing the first 15 elements. The script gives the output seen below:

The following is the count for the top 15 shapes reported:

light	14628
triangle	7038
circle	6717
fireball	5526
unknown	4899
other	4889
sphere	4732
disk	4477
oval	3285
formation	2171
cigar	1795
changing	1710
rectangle	1157

flash 1154
cylinder 1122

In order to determine the most common shapes reported for UFO sightings in Seattle, a new file is opened to write, and for loop is used to check each line in the original UFO file and add any row that contains ‘Seattle’ to the new file. The new file of Seattle data is then read into a pandas data frame and the columns titles are created as the same titles in the original UFO file. A list is then created for the ‘shape’ column in the Seattle data frame and the five most common shapes are calculated in the same way that the top five countries and cities were. The same process is done for the rest of the most common five cities; Phoenix, Portland, Las Vegas, and Los Angeles. The script gives the following output:

The following are the top shapes for UFO Sightings in Seattle:
[('light', 44), ('fireball', 22), ('other', 21), ('unknown', 20), ('circle', 17)]

The following are the top shapes for UFO Sightings in Phoenix:
[('light', 113), ('triangle', 51), ('other', 46), ('sphere', 41), ('circle', 36)]

The following are the top shapes for UFO Sightings in Portland:
[('light', 97), ('circle', 41), ('sphere', 38), ('unknown', 29), ('fireball', 29)]

The following are the top shapes for UFO Sightings in Las Vegas:
[('light', 85), ('circle', 42), ('sphere', 34), ('unknown', 34), ('other', 34)]

The following are the top shapes for UFO Sightings in Los Angeles:
[('light', 78), ('circle', 43), ('triangle', 42), ('disk', 36), ('fireball', 34)]

The top five cities with reports of UFO sightings contain reports of ‘light’, ‘triangle’, ‘circle’, ‘fireball’, ‘unknown’, and ‘other’ for the shapes of the unidentified flying objects. These are also the six most common shapes observed overall.

Furthermore, all of these cities are on the west of the United States. One theory could be that UFOs are mostly seen in the western states because of the location of Area 51. This could be extraterrestrial space crafts or other countries spying on the United States. However, most likely is that the unidentified flying objects are meteors. This is supported by the shapes reported as lights or fireballs as well as unknown, if the meteors do not present a clear shape. The meteor showers are probably more easily visible on the west coast due to less city lights. Although the top cities are large cities on the west coast, they still have more visible open sky space than do the more congested cities on the east coast. This analysis successfully determined the most common locations for UFO sightings as well as the most commonly observed shapes and therefore determined that the sightings are most likely meteors.

Conclusions

The data set containing reports of UFO sightings over the past century was analyzed to determine which country has the most reports of UFOs, as well as to then determine what areas

in the top country are the most common for UFO sightings. It was also analyzed to determine a correlation between the shapes of the UFO sightings to help identify the unidentified flying objects. The data set obtained from the National UFO Reporting Center was of sufficient reliability and quality to complete this analysis. The data was transformed into pandas data frames in order to create lists and calculate the most common countries, cities, and shapes reported. The bar plots and pie charts of the most common cities help to determine the relative proportions of the most frequent locations for UFO sightings. This analysis allows the United States government to conclude that there is most likely no threat to national security by determining the locations and shapes of frequent UFO sightings and analyzing the results. Observations of “lights” and “fireballs” in the western states are most likely sightings of meteor showers.