Homework #2

CS 5665, Fall 2016

Task 1: UFO Data Collection, Cleaning, and Exploratory Analysis

- Our first task is to collect, clean, and explore data from the UFO sightings database. In particular, we are especially interested in UFOs corresponding to one of **three shapes**:
    - Circle
    - Triangle
    - Fireball

  For each UFO shape, we should collect all sightings from the list of UFO sightings made **between January 1, 2005 and September 22, 2016**. In other words, ignore any sighting made after September 22, 2016 or before January 1, 2005.
  You should represent each sighting by these eight features:
    - Date of Sighting
    - Time of Sighting
    - City
    - State
    - Shape
    - Duration
    - Summary
    - Posted Date (when the sighting was posted to the website)

  Answer:

  The extraction part was fairly easy. So, I used Microsoft Excel's "Get External Data from the Web" feature to extract the UFO data of each shape into individual excel spreadsheets. After this removed data not present in the range of the duration. Then I merged all the three sheets into one new sheet. Finally, I split the date column into date and time. Now the data cleaning process will start.

- As part of your data collection and cleaning, you should do your best to convert all Durations to seconds, whenever possible. Keep in mind a few guidelines:
    - If a duration has a "<" sign, you should simply ignore the "<" sign. For example, if the duration is specified as "< 1 minute", consider the duration to be "1 minute". You should subsequently convert "1 minute" to "60 seconds".
    - If a duration has a range, use the upper limit as its value. For example, if the duration is listed as "5-8 minutes", you should consider the duration as "8 minutes". (Again, you will need to eventually convert minutes into seconds).

- You may encounter some other oddities in the data. Do your best to extract maximum value from the messy data; be sure to explain to us the decisions you have made in terms of data extraction and cleaning.

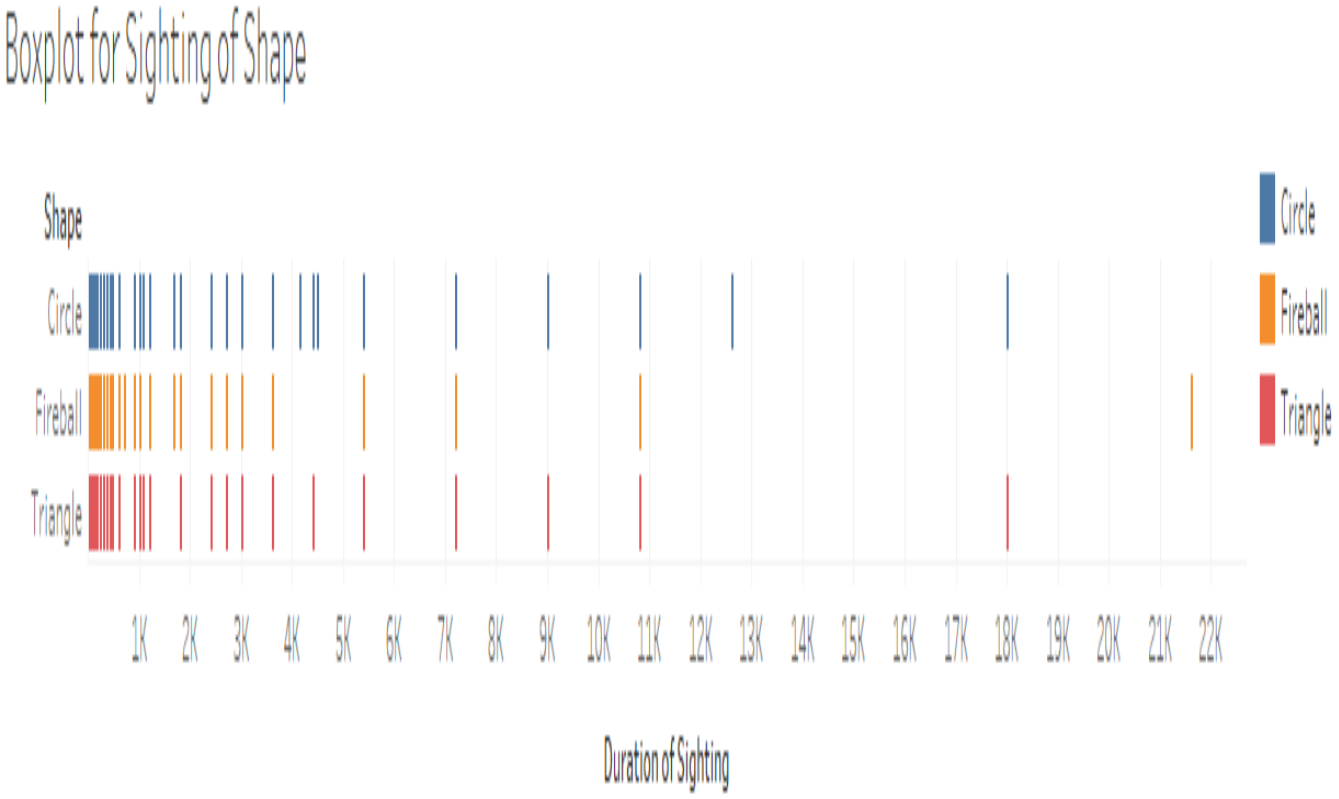
Answer:

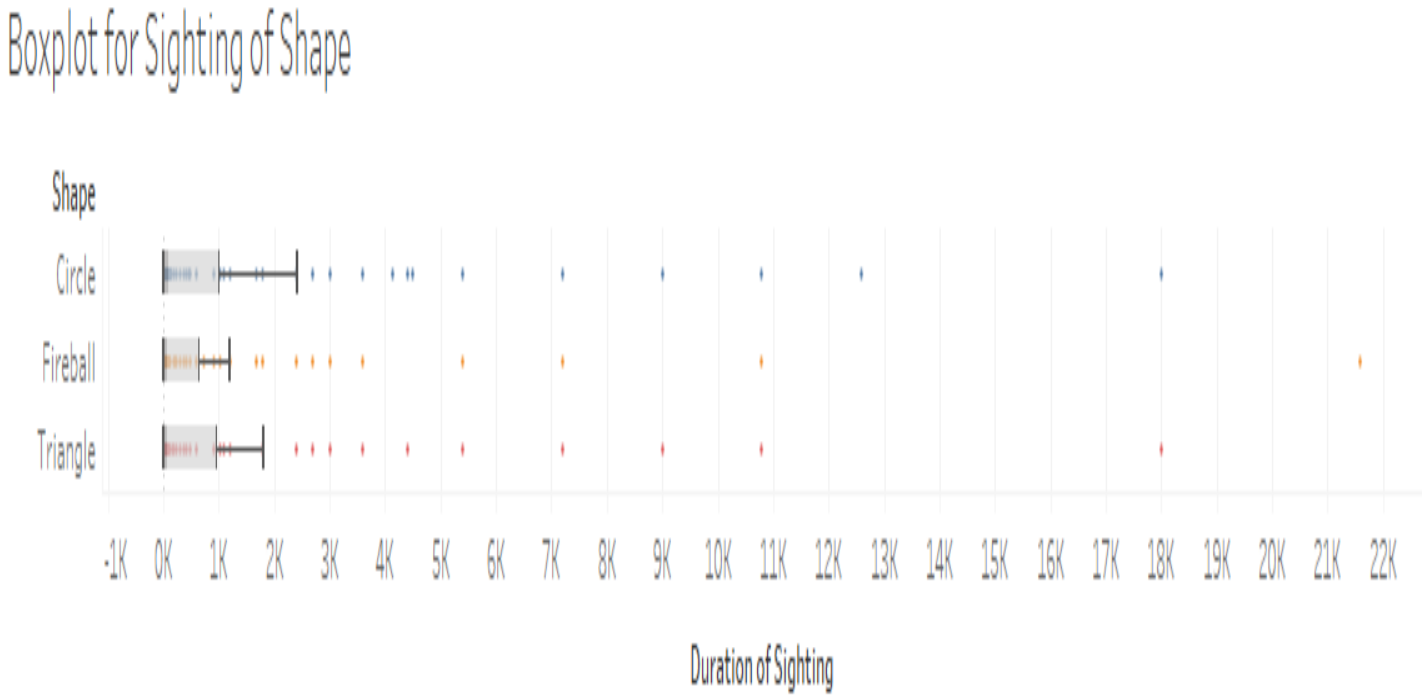
There were a lot of inconsistencies in the "Duration" column, some of the values were alphanumeric, contained special symbols, etc. To make the data consistent I used Microsoft Excel and OpenRefine.
There were lots of empty attributes too, which I deleted.

- Based on your cleaned data, you should perform a basic exploratory data analysis to better understand what you've got. Specifically, we expect to see the following:
  - A boxplot of the duration of UFO sightings of each shape (one boxplot per shape).

Answer:

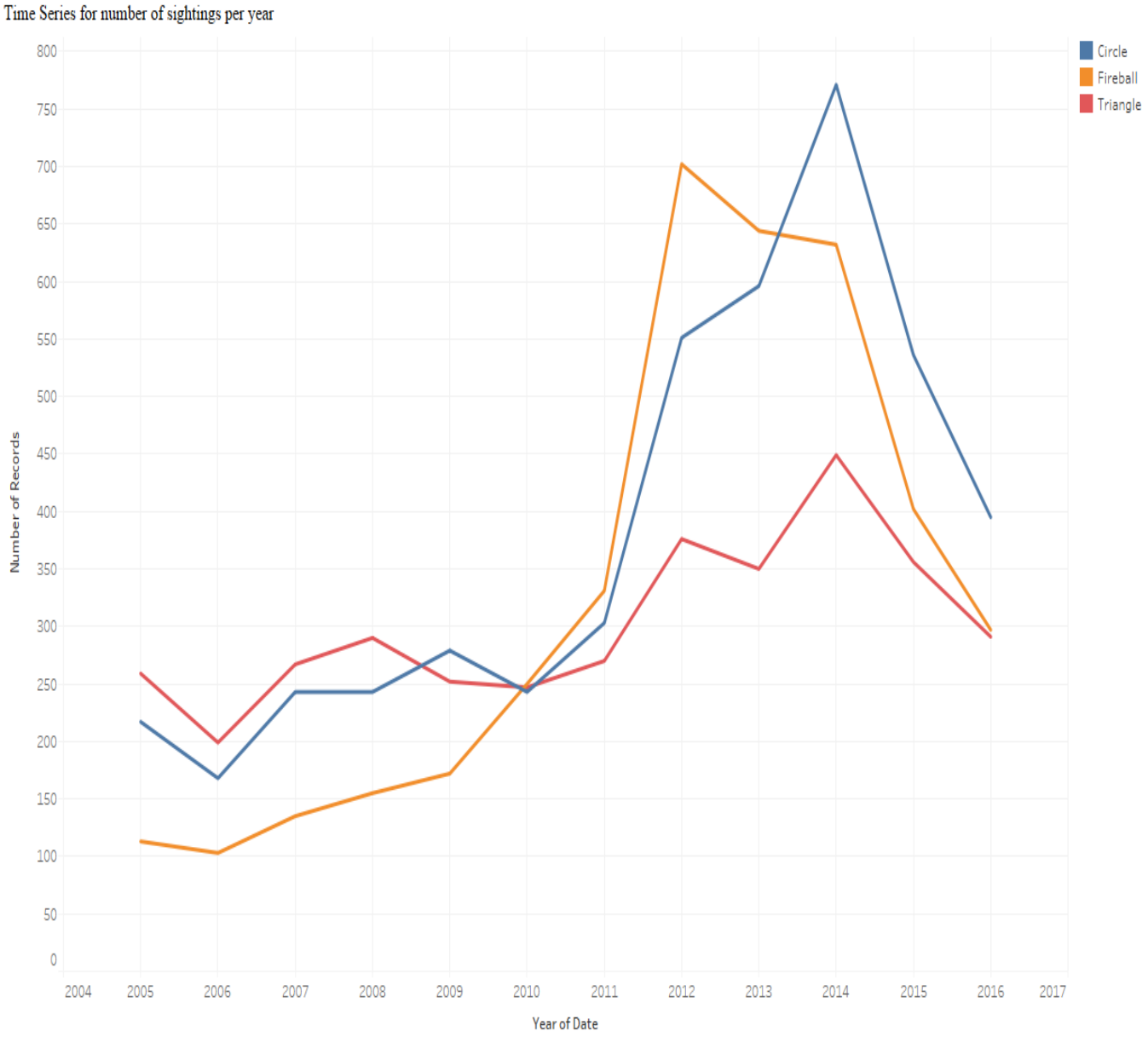
For performing basic exploratory I have used Tableau which reads data from the excel created.





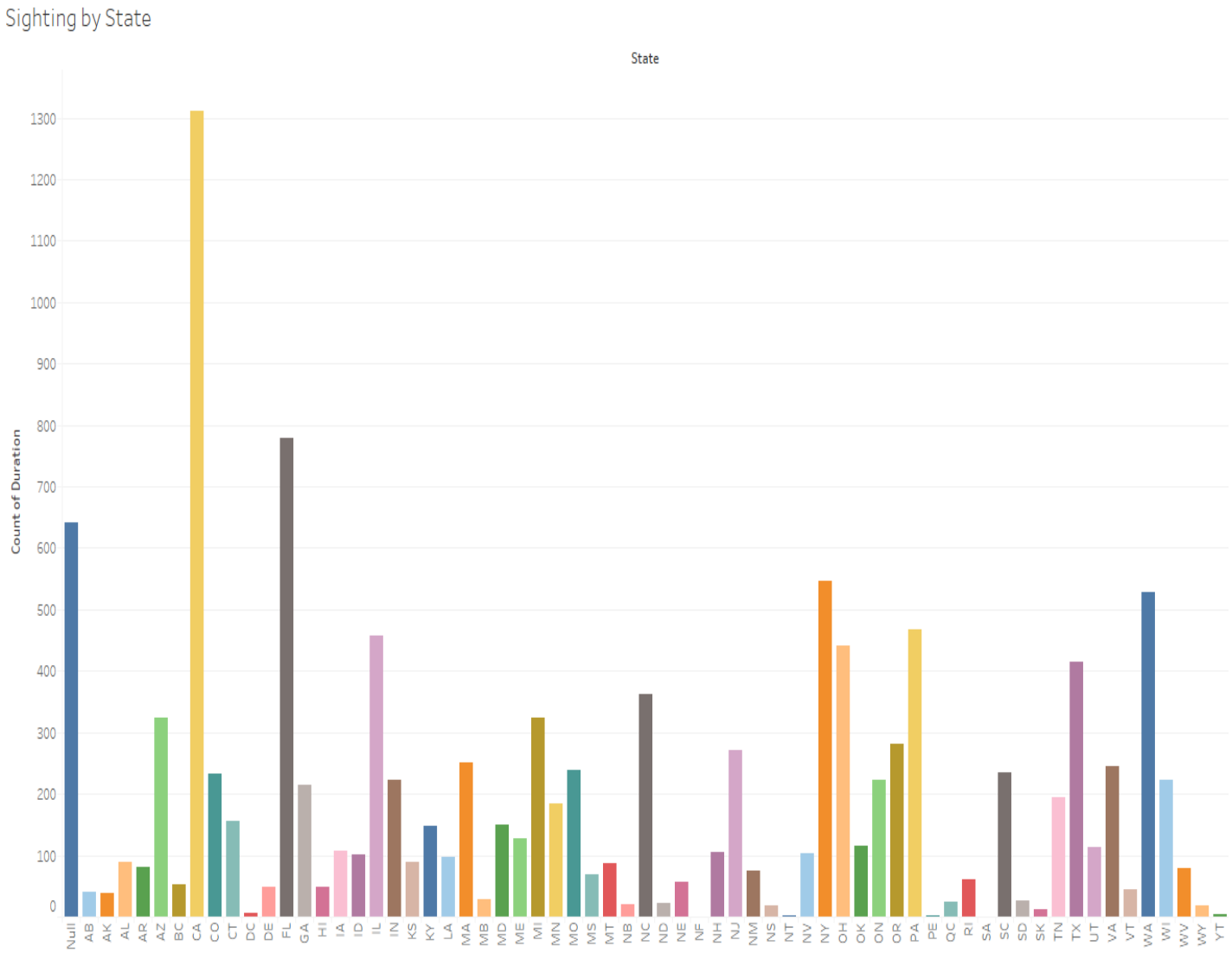
I created two boxplots

- A time series figure with the number of sightings per year (one line per shape).



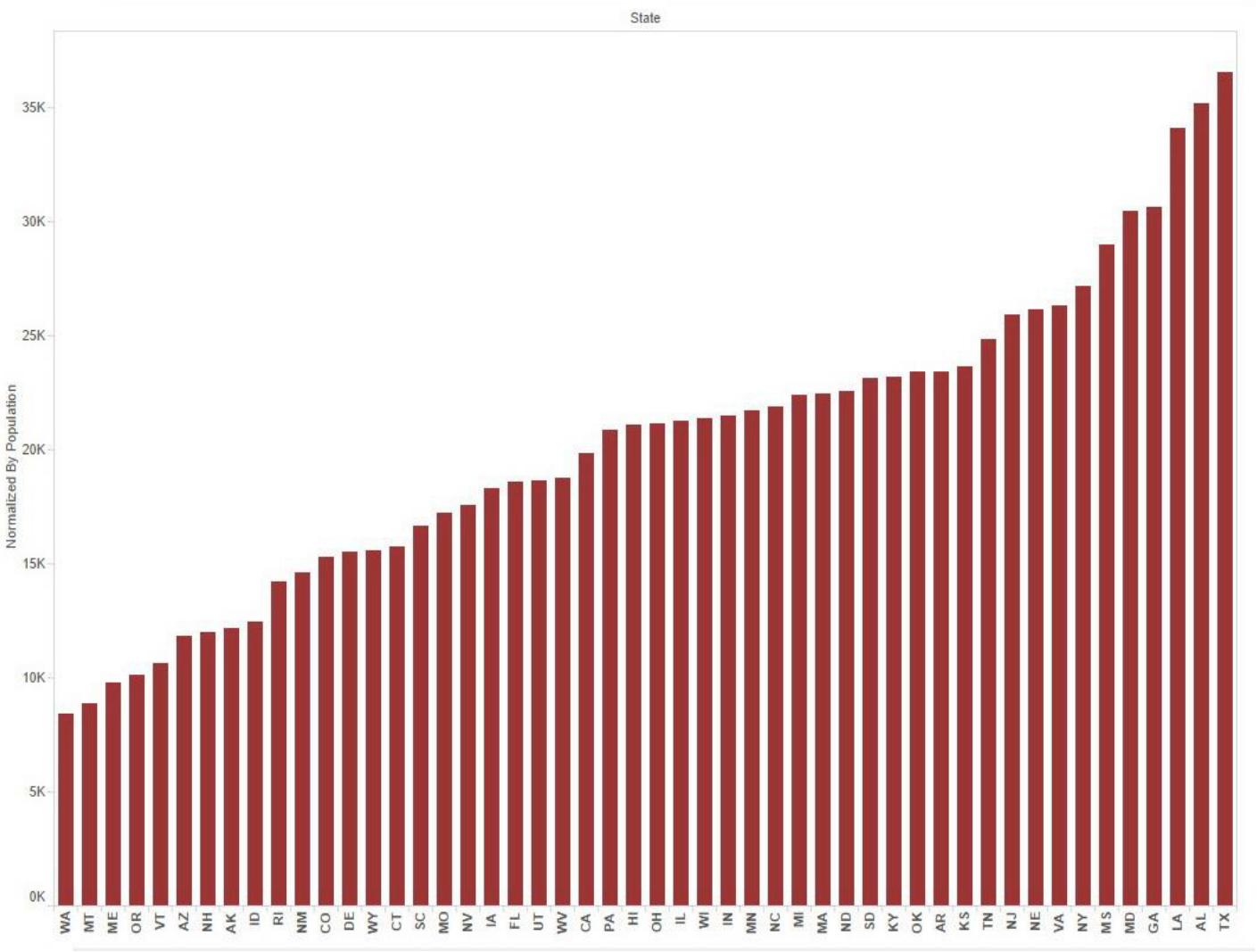
- From the line chart above we can infer that the number of sightings in 2014 for Circle is maximum, while its minimum for Triangle
- In 2012 Fireball shape UFO were seen the most.
- From the chart above we can also infer that Triangle shaped UFO were sighted the least from between the duration of 2005 to 2016

- A bar chart for sightings by state.



- This graph is to compare the sightings occurred in various states.
- From the graph we can infer that the most number of sightings occurred in California.

- Normalize the sightings by state population. What do you observe? Anything interesting?



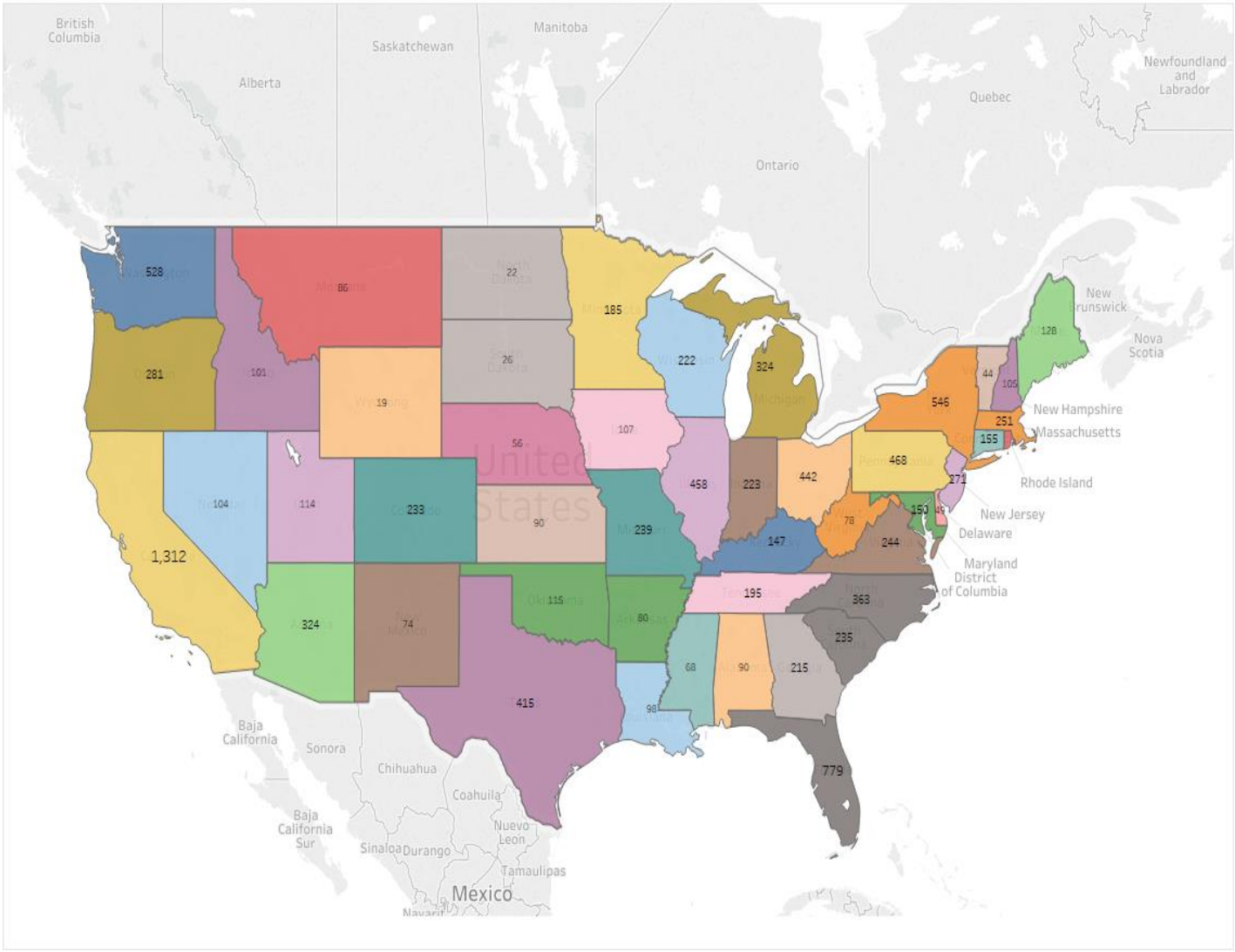
- For normalizing the sightings by state population, I have applied the formula:

Normalized Sighting of each state = (Population of each state / Sightings of that state)

- The above graph shows the normalized sighting for each state, i.e. the number of people per sighting.
- Washington has least number of population per sightings, which means that more number of people have seen the UFOs in Washington according to the population. Similarly, Texas has most number of people per sightings, which means that lesser number of people have seen UFOs in Texas according to the population.
- This observation is quite revealing because in the previous bar graph, California has most number of sightings, but when we normalize by population, Washington tops and California goes in the middle.

- Visualize the distributions on a map. Do you notice anything peculiar?

Sheet 5



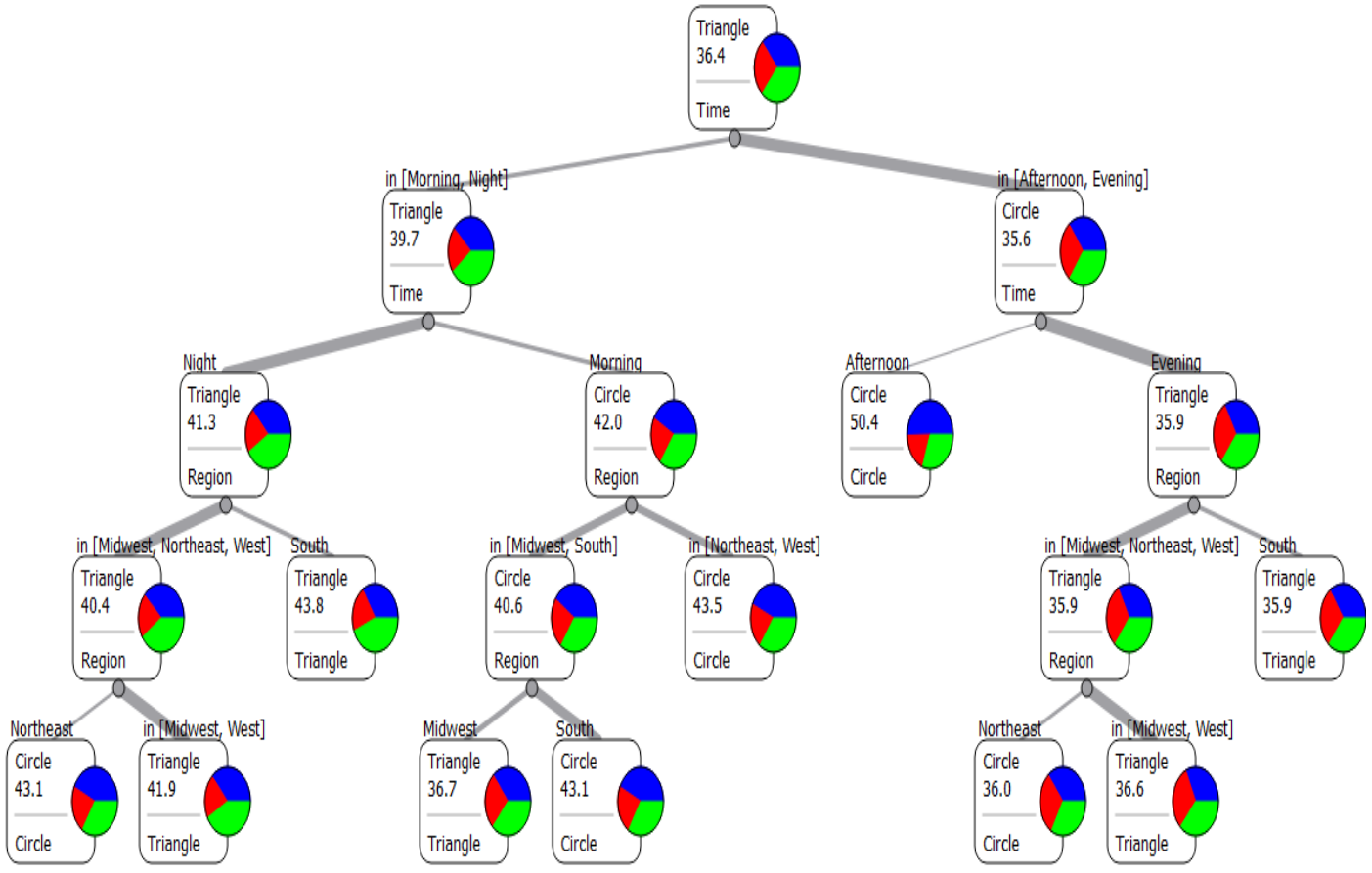
- The above graph shows, non-normalized sightings on the map.
- California has most number of UFO sightings.
- Midwest region has least number of UFO sightings BY COUNT

Task 2: Predicting UFO Shape

- Given your understanding of the data, your goal is now to build a decision tree classifier to predict the shape of a UFO. You have three target classes: circle, triangle, and fireball. For this task, you need only consider two simple features to represent each sighting:
    - **Region of the country:** We shall divide the 50 states into the four Census Bureau-designated areas: Northeast, Midwest, South, and West.
    - **Time of Day:** We shall consider only four parts of the day. Night (00:00-05:59), Morning (06:00-11:59), Afternoon (12:00-17:59), and Evening (18:00-23:59).

  That is, each sighting will simply be represented by its region and time of day. E.g., (South, Morning) or (Midwest, Evening).
  Next, split your dataset to training set and test set. Training set consists of all sightings made between **January 1, 2005 and December 31, 2013**. Test set consists of all sightings made between **January 1, 2014 and September 22, 2016**.
  Based on these two features with the training set, you should implement a decision tree classifier that uses Gini Impurity to determine the best feature at each step.

Answer:



- You should provide an illustration of the decision tree (built based on your training set). You may use a graphing toolkit (like networkx) or you may draw the tree manually.

Answer:

- The decision tree above is for the training set data between January 1, 2005 and December 31, 2013, and has been calculated using the genie Impurity method.
- Here green color is for triangle shaped UFO, blue is for circle and red for fireball.
- Initially, at top of the tree, the split is made based on the time, in sets of {Morning, Night} and {Afternoon, Evening}. On the left split, triangle is classified as 39.7%, circle about 38.4% and fireball 22%. Which is not the optimal split because the percentage of triangle and circle is very close, and it will not be clear that the UFO is a triangle or a circle. Same from all the other splits in the other nodes of the tree.
- We can observer from the graph that the decision tree predictor can somewhat predict circle and triangle, but it is not able to predict fireball.
- In conclusion we can say that, for this particular dataset, using only two attributes for prediction does not give a good classification tree, i.e., it is not able to predict shapes clearly using only two attributes. It may be wise to include more attributes to form the decision tree.

- You should report the classification accuracy for your decision tree using the test set.

Answer:

- After the giving the test data set to the model created by training dataset. There is a cumulative error of around 12-13%.
- The reason could be the increase in fireball sightings in the test data set or may be decrease in the number of triangles (as shown in the below graph). In short we can say that our model in somewhat accurate.
- From 2013 onwards in our training dataset, there could be increase in the fireball sightings at night because of the increase in the number flight at night, which somewhat look like fireballs?
- Or may be due to increased pollution, the sightings of triangles, which usually are seen in the morning have reduced? We don't know!

Task 3: Improving your Accuracy

- Can you improve your prediction rate (accuracy) over what you got from Task 2? You may use raw features instead of the two features or even combining all features. Or something else?
    - Describe how you can achieve better result compared with Task 2's result.
      Answer:

      We might increase the prediction accuracy by taking more attributes into consideration while developing the classification tree model, for example, we can include the duration of sightings, and group them into groups like <10 seconds, 11-60 seconds, 61-300 seconds, 300-750 seconds, 750-3600 seconds, 3601+ seconds, etc. This might make the decision tree more specific and help predict the shapes even better. We can also increase branching of the tree by categorizing data or converting attributes into categorical to reduce the height or depth of tree which will also make the tree traversal faster.

    - Report what result you got.
      Answer:

      Before including the "sighting duration" as an attribute in the decision tree, the shapes were getting classified in percentage between 34-40%, but after including "sighting duration", this percentage has slightly increased to around 37-44%. This is not a big difference, but our prediction has increased, which is a good thing.

    - What feature is the most important feature to distinguish shapes of UFOs?
      Answer:

      The most distinguishing shapes. It gave a good split in around the region 44-50%.