Homework #2

CS 5890, Fall 2015

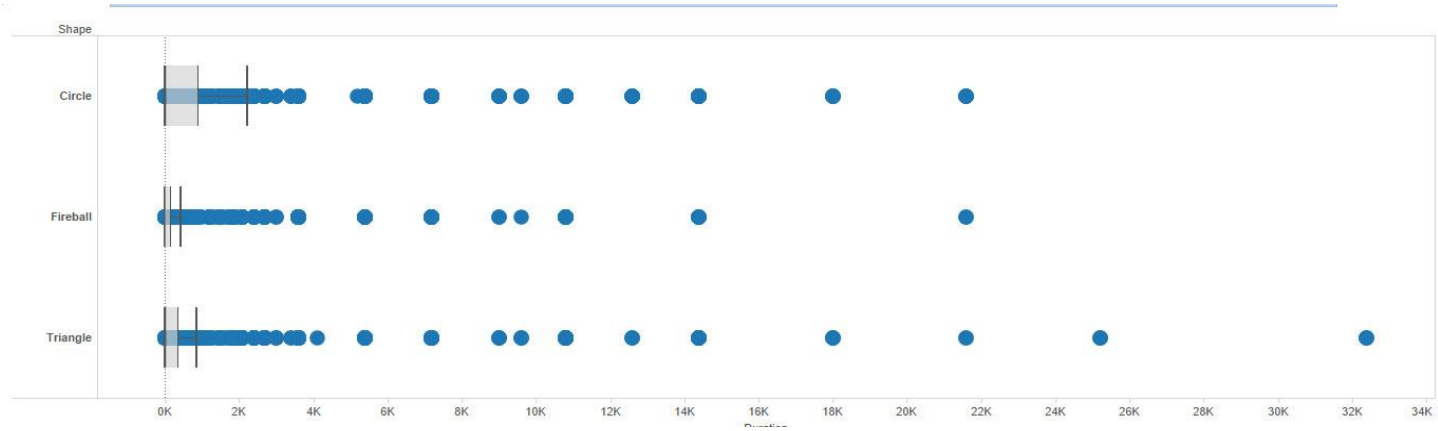
Task 1: UFO Data Collection, Cleaning, and Exploratory Analysis

Task: 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 n cleaning.

Answer)

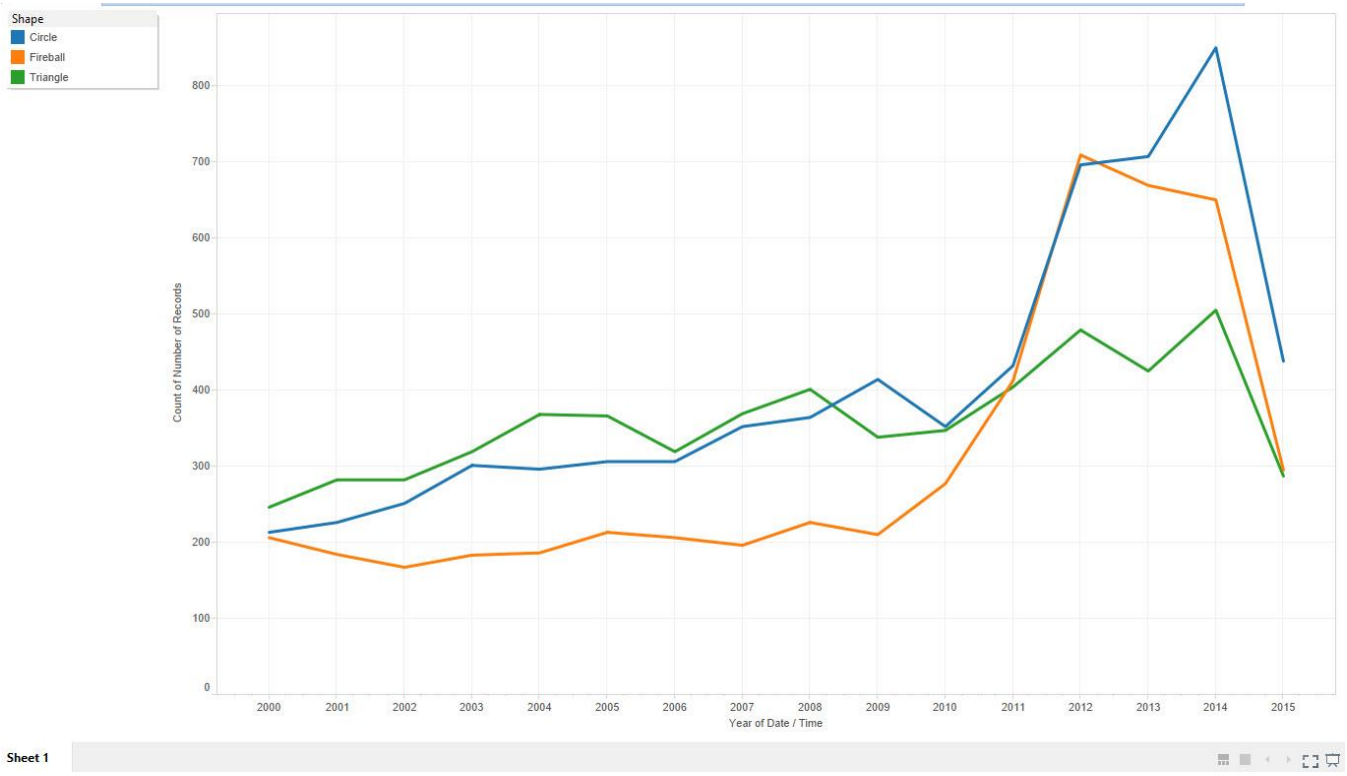
- The data extraction part was fairly easy, I used Excel's "Extract Data from Web" feature to extract the UFO data of each shapes into individual Excel spreadsheets. After that I merged all the three sheets into one file. Now the data cleaning step began.
- There was a lot of inconsistencies in the "length of duration" column, some of the time was in alphanumeric characters, and to make the data consistent I used Google Refine.
- There were lots of empty attributes too, I simply delete those rows.

Task: A boxplot of the duration of UFO sightings of each shape (one boxplot per shape).

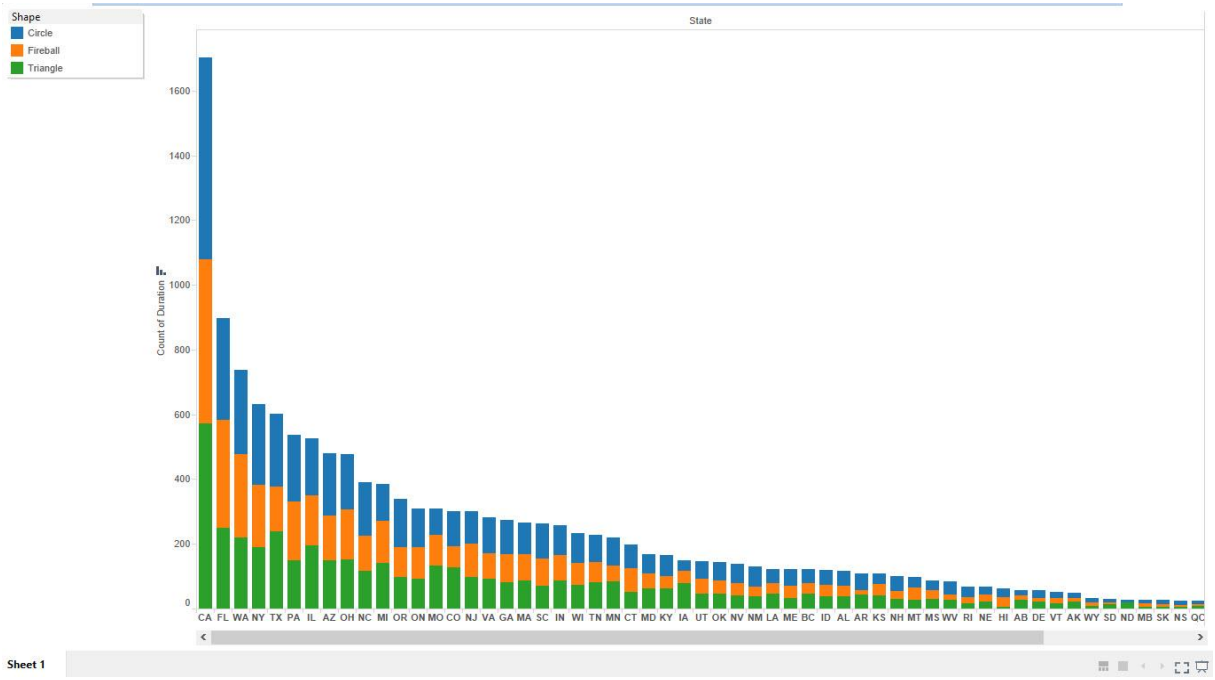


(Go to next page)

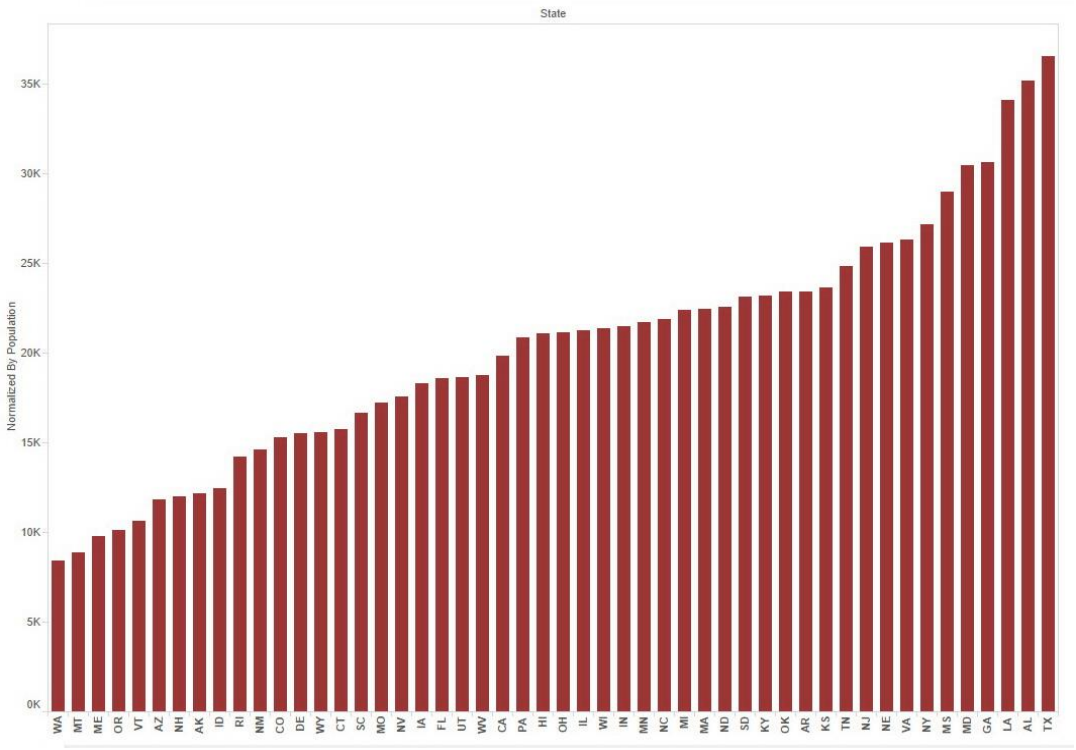
Task: **A time series figure with the number of sightings per year (one line per shape).**



Task: **A bar chart for sightings by state.**



Task: 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.

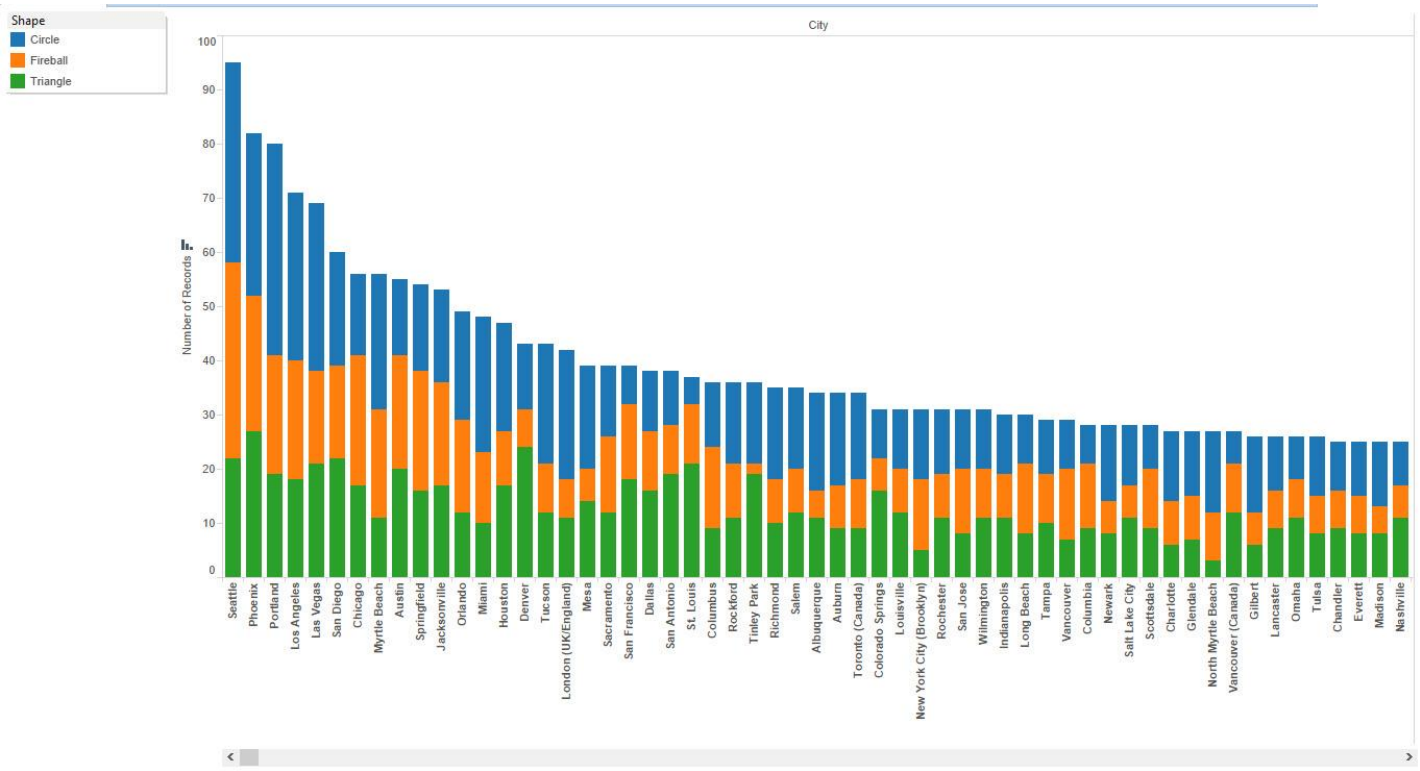
(Go to next page)

Task: Visualize the distributions on a map (e.g., using Tableau, Google Maps API, basemap, or D3). Do you notice anything peculiar?

- The above graph shows, non-normalized sightings on the map.
- The darker color of green show more number of sightings. California has darkest color, meaning that it has most number of UFO sightings.
- The pie charts show the distribution of shapes, we can see that high percentage of fireballs were observed in Florida. Likewise, more number of triangles were observed in Texas, and so on.
- Midwest region has least number of UFO sightings, as depicted by shallow green color and small size of pie charts.

(Go to next page)

Task: These are just two suggestions. You are encouraged to explore the data based on your own intuitions, but you are required to ask and answer at least one additional question beyond the basic data analysis we require above (boxplots, time series, bar chart). Remember, the people of Earth depend on you to draw interesting insights from the data!



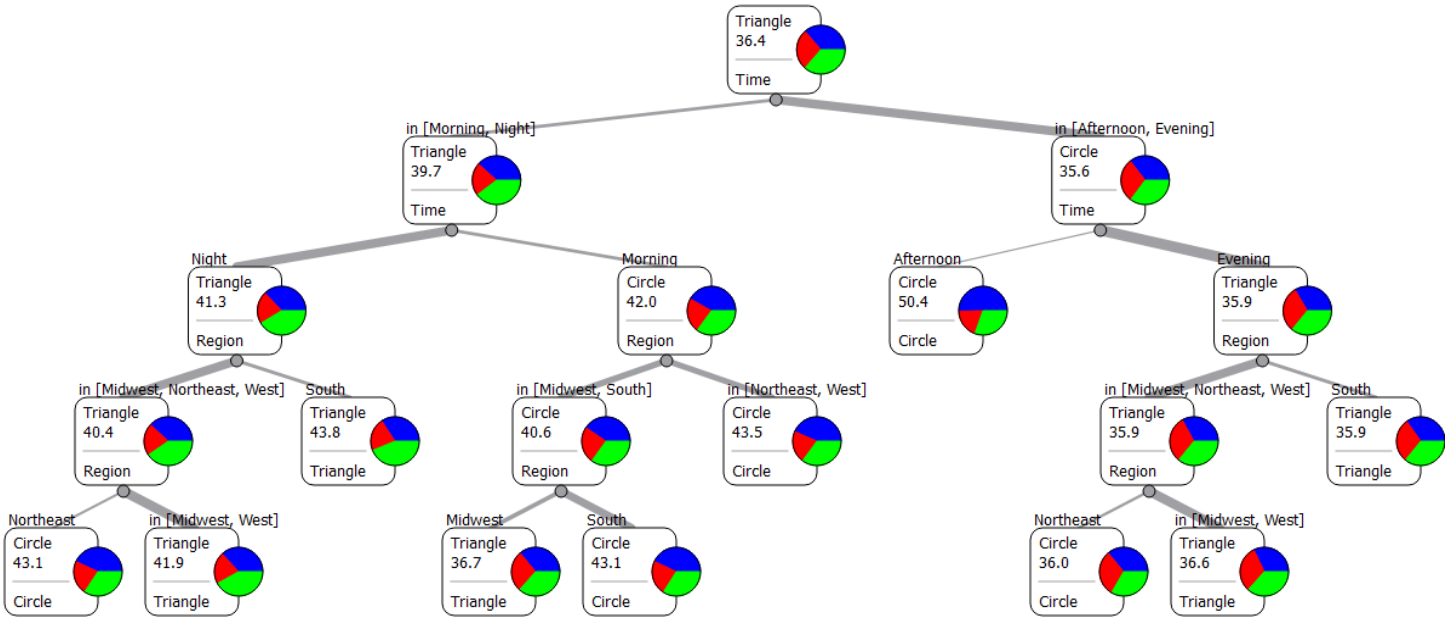
- The above bar graph answers the question: “Which cities observed most number of UFO sightings?”
- We can clearly see that Seattle observed most number of UFOs followed by Phoenix and Portland.
- It can be observed that Denver observed most percentage of Triangle shaped UFOs and very less number of Fireball UFOs.
- Columbus and Chicago observed most number of Fireball UFOs.
- Miami, Auburn and London observed most Circle UFOs.

---------- X ----------

(Go to next page)

Task 2: Predicting UFO Shape:

Task: 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.



- The decision tree above is for the training set data between January 1, 2000 and December 31, 2012, 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.

(Go to next page)

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

- After the feeing 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!

---------- X ----------

(Go to next page)

Task 3: Improving Your Accuracy (Additional 5 points. This task is optional)

- **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, like 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.

- **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 35-40%, not after including "sighting duration", this percentage has slightly increased to around 38-44%. This is not a big difference, but it our prediction has increased, which is a good thing never the less.

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

Answer) **"Time of Day"** feature gave the most distinguishing shapes. It gave a good split in around the region 43-50%.

---------- X ----------

(End)