Chris Huber, Tayyaba Arshad, Shiva Lakshmanan

INFO 3300

March 3, 2018

Project 1 Write Up

Description of Data

The UFO sightings data came from Kaggle. According to the author, The UFO dataset was scraped from the National UFO Reporting Center's website. The fields of the dataset included a timestamp, city, state, shape of the "UFO", duration of the sighting (MM:SS), a textual description, and a latitude/longitude coordinate pair locating the sighting. The dataset contained over 80,000 data points, which is too many to visualize on a map without extreme loading times. So, for our maps we took a random subset of the data to obtain around 12,000 data points to plot. To perform this filtering, we used Microsoft Excel. Firstly, we used the RAND function to randomly generate a number between 0 and 1 for every row of original UFO dataset. Then, we used the function =IF([RAND() > 0.15, 0, 1) which returns the number 0 if the randomly generated number was above 0.15 and return the number 1 if it was below 0.15. We then filtered the data and organized it so that every point under the IF function that returned a 1 would be separated. We were left with 11,880 data points in our filtered data set. It is important to note that the filtered data was only used for our map visualizations, due to the visual constraints. However, data points that were in between the years 1970 and 2013 were used for our line graph visualization, which was a little less than the full 80,000.

When choosing which fields to use in our project, we selected only the numeric fields to support our argument, trimming out the rest of the fields . We didn't want to confuse the user by using every piece of information. For our histogram, we displayed the number of UFO sightings that occurred every year. Therefore, we only looked at the "datetime" column, extracted only the

year, and aggregated all of the data from each year to plot in the bins of our histogram. The rest of the information in this dataset for this visualization was unnecessary to our argument. For our second visualization we wanted to plot on a world map where each sighting took place. In order to do this we used the latitude and longitude information for each UFO sighting from the dataset. In conclusion, for the entirety of the project we only used the year, latitude, and longitude of every UFO sighting. These pieces of information proved to be the most useful in conveying our story.

The second dataset that we used for this project was found on data.worldbank.org and provided information about airline flights. The fields in this dataset included the country name and code, indicator name and code, and the number of flights that departed from every country in every year starting in 1970. The main information that we needed was for our histogram visualization, where we needed to plot the total number of flight departures every year. Therefore, we summed up all of flight data over all the countries and came up with a total number for every year. By the end of the aggregation, we had the total flight departures every year from 1970 to 2016 (and saved the results in flights_by_year.csv).

The third dataset that we used was found on Kaggle and consisted of airport codes (JFK, ORD…) and corresponding latitude/longitude pairs.. We kept this dataset as is to use in our US map visualization.

A final dataset was a custom one. We looked up the locations of the busiest airports in the United States, and indexed them by name and airport code. The data for this “important_airports.csv” was found on Wikipedia.

An additional data file we used was a JSON geographic data file called “world-50m.json”, which allowed us to generate paths for both our US map and world map visualizations.

Description of Mapping

Our project contained three visualizations. The first was a map of the entire world which was used to visualize the UFO sightings dataset. We used the d3.geoEquirectangular() projection to display the world map in rectangular form. We then added the UFO sightings based on their latitude and longitude coordinates, using the projection's pathGenerator to convert latitude and longitude coordinates into pixel locations in the SVG. We represented each UFO sighting with a small light blue circle, with a low opacity. We used low opacity so that if sightings occurred in geographically similar areas, the overlap would produce increased brightness. Therefore, users would be able to visually see if an area had a greater number of sightings based on both the brightness of the area as well as the amount of area that the blue circles covered.

The second visualization is a line graph which illustrates the number of UFO sightings per year since 1970 with the number of plane departures per year since 1970. Since there were two separate datasets being used, we generated two different line graphs that were laid on top of each other so that they could be easily compared. The scale for our x-axis remained consistent, as both datasets were formatted to 1970 to 2013 with the filter function. However, the scaling for the y-axis differed. For our flight departure dataset, we used a linear scale that ranged from 48,238,800 to 287,885,038. Both the axis labeling and the line itself were represented with the color orange. For our UFO sighting dataset, we used a linear scale that ranged from 0 to 7,537. The axis labeling and line were represented here with the color light blue. We also maintained an opacity of 1 for all aspects of the graph so that all pieces of information would be equally considered. We picked blue and orange as colors because they are complementary and are thus easy to differentiate.

The third visualization is a map of the United States which displays both the geographical distribution of UFO sightings as well as airports. Similarly to the world map

mentioned above, both sightings and airports were added to the map based on their coordinates. Every UFO sighting was displayed with a small blue dot, while every airport was displayed with a small orange dot. On the map, we can see that the distribution of airports is fairly even throughout the country. UFO sightings are distributed throughout the country as well, but they are much more concentrated and densely packed in certain regions of the United States. We then decided to some bigger green circles to the map which highlighted some of the big clusters on the map. Seven of the areas that we highlighted with these circles were also near some of the country's largest and busiest airports. We used a larger radius of circle to indicate a feasible radius in which one might see a plane in the sky and misreport it as a UFO. One green circle was also added in the infamous region of the United States known as Area 51.

The Story

In the first visualization we can see the distribution of UFO sightings around the world, each of which is indicated by a blue dot. It was surprising to see that the highest concentration of UFO sightings was located in the United States and the United Kingdom. Besides these two areas, sightings are fairly rare in the rest of the world.

In the second visualization we can see the number of UFO sightings per year increasing over time. We can also see that the number of flight departures per year are increasing over time as well. There is a correlation between UFO sightings and flight departures, which can lead us to believe that although there is not sufficient evidence to prove that the increase in flight departures is directly causing UFO sightings, that there is at least a positive correlation. It is feasible to draw the conclusion that some UFO sightings were in actuality airplanes that were mistaken as UFOs, but it is impossible to get an estimate of how many or what percentage. It is also important to note that for a period of time from about 1980 to 1990 the rate of flight

departures was increasing steadily while the number of UFO sightings remained fairly consistent.

In the third visualization we zoom into the United States to see the distribution of UFO sightings across the country labeled as blue dots. It is visible now that although there are sightings across the country, that there are clusters of sightings in specific areas of the country as well. Additionally, the locations of airports across the country have been mapped as well with orange dots. The green circles on this map indicate some of the largest airports across the United States where a high frequency of flight departures would be concentrated. Within the radius of these circles we can also see that there is much higher concentration of UFO sightings in the area. This map makes it easier to visualize the geographical correlation between flight departures and UFO sightings. Again, since there is no proof of causation, another possible explanation of the increase in UFO sightings could be that major airports are generally located in areas with high population density. The clusters of UFO sightings could be due to the flight departures, but it could also be due to the high concentration of people and greater likelihood of a person “seeing” a UFO in that area.