# TEAM\_02\_CSCI599\_HW\_BIGDATA\_Report

## Input File: ufo\_awesome\_FINAL\_OUTPUT\_v2.tsv

## Output file: A folder “10visualizations” with all the individual d3 visualizations

## Data Preparation and Cleaning

We started with manually cleaning the date field in the v2 TSV file in order to make the British UFO sightings dates correspond to the same standard format as that of the UFO\_awesome sightings date format.

Once we were done with that, we wrote a script which converts the entire V2 TSV into json and also indexes the same into ElasticSearch. We then wrote 10 scripts to aggregate and extract the data we required for the 10 D3 Visualizations in TSV format. Once we had the 10 TSV’s, we wrote programs to put them into JSON and ingest it to their corresponding D3 visualizations.

### Our D3 Visualizations:

**The count of the number of UFO sightings based on year**

This is the first aggregation that comes to mind when deciding to collect useful information whether the number of sightings increased or decreased every year.

From the visualization it is evident that after remaining constant till almost 1995, the number of sightings reported grew almost exponentially by the year.

**The different shapes of the UFO sightings and its count**

With the data consisting of the different types of shapes of UFO sighted, we thought it would be a good thing if we could analyze of which shape did most people encounter their UFO sightings.  
  
From the visualization we can see that majority of the people found the shape of the UFO to be …. followed by ….

**The number of sightings per year within a 25-mile radius vs outside the 25-mile radius**

**The total duration of the sightings in each year**

One good result to view would be the total combined duration for all the sightings in that year which would give us a hint about how long were the UFO’s visible in the sky’s to people  
  
From the visualization we can see that ….

**The count of UFO sightings in densely populated areas vs sparsely populated areas by year**

As analyzed in Assignment 1, it seems to be a relevant metric to evaluate whether majority of the sightings took place in densely populated areas or whether they took place in sparsely populated areas.

As visible in the visualization we can say that highly sighted UFO locations are sparsely populated.

**The count of UFO sightings in a rural area vs an urban area by year**

In Assignment 1, we checked for whether the sightings took place in rural area or did they take place in urban areas. This gives us a good understanding of the mindset of such people and also gives us a brief idea about the literacy of the people involved in such sightings.

As is evident from the visualization, for a given Census year range, most of the sightings happened in rural areas. For the year range 1991 – 2010, rural sightings percentage is at least 74.9% ~75% (ranging up to 91.9%)

**How many Sci-Fi movies vs UFO sightings took place each year?**

Based on the year of release of sci-fi movies and the year of the UFO sightings, we can predict if whether the UFO sighting was a delusion or not.

As we can see in the visualization, in majority of the years, the ratio of the sightings to the number of Sci-Fi movies released was less than 2.  21.95% of the years where a UFO was reported could be a possibility of delusion where the person confused an aircraft for a UFO after watching a sci-fi movie released in the same year.

**The count of meteorites in rural vs urban areas by year**

We can analyze the count of meteorites landings that took place in rural area vs urban areas. Combining this with the results obtained from the count of UFO sightings in a rural area vs an urban area would give us a good understanding of whether people confused a meteorite landing to be a UFO or not.

As is evident from the visualization, …

**A Bubble graph based on the number of sightings in the states of USA**

Based on the city and state aggregation, we thought it would be a very nice if we could visualize the states which had the maximum visualizations. With the help of a bubble graph, the sizes of the bubbles of different states would indicate whether the number of sightings in that state was low or high.

As is evident from the visualization, …

**A graph representing who the sighting was reported by**

Another feature we thought that would be very useful is to recognize the gender and child vs adult nature of the person who reported the sighting. We do this using the description provided and checking for gender and age based keywords such as “he”, “her”, “his”, “adult” etc.

As is evident from the visualization, …