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

## Data Preparation and Cleaning

We started with manually cleaning the date field in the TSV\_v2 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 TSV\_v2 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 TSVs, we wrote programs to output the data into JSON, CSV or TSV files and ingest them to their corresponding D3 visualizations.

### 10 D3 Visualizations:

1. **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 every year.

1. The number of different shapes of the UFO sightings

With the data consisting of the different types of shapes of the UFO sightings, we thought it would be a good thing if we could analyze how many types of shapes (1D, 2D, 3D or Unknown) did people observed the UFO sightings.  
From the visualization we can see that majority of the people found the shape of the UFO to be are 1 dimensional which include flash, light, line, diagonal, zigzag line etc.

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

We aggregated the count of sightings per year where the sighted location is within 25-mile radius of an airport or not, mostly indicating the UFO Sighting as False positive.

1. **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 to people.

From the visualization we can see that, maximum number of reported sightings were in the duration of minutes followed by seconds.

1. **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.

In the visualization, we decided to compare individual donut charts of densely populated areas and sparsely populated areas. As visible in the visualizations, the percentage of sightings per year for highly sighted UFO locations are sparsely populated.

1. **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 areas or in urban areas. This gives us a good understanding of the mindset of such people and also gives us a brief idea about the education level 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 ~75% (ranging up to ~92%)

1. **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.  ~22% of the sightings 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.

1. **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, meteoroids occurrence in rural areas is more compared to urban areas.

1. **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 which states had the maximum number of visualizations. With the help of a bubble graph, the size of the bubbles of different states indicates whether the number of sightings in that state was low or high.

From the visualization, the size of bubble per each state in US represents the count of UFO Sightings happened from the given dataset. It is evident that maximum happened in California state.

1. **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 “male”, “he”, “her”, “his”, “adult”, “female”, “kids”, “boy” etc.

As is evident from the visualization, majority of the sightings are reported by Male. There are some sightings reported by kids which can also be ignored.

### ElasticSearch Visualizations:

We started with feeding our TSV\_v2 dataset of ~62,000 records to elasticsearch and we stored the data under the index = “bigdata”. We also indexed the coordinates of all the locations present in the records, which we later use to map the locations on world map D3 visualization.

1. We create the elasticsearch client and establish the connection with host <http://localhost:9200/>
2. An HTML dropdown has been provided where the user can dynamically select the reported year.
3. Once the user selects the Year, the elasticsearch queries the indexed data and provides the filtered output to the D3 visualizations.
4. We then extract the location and its corresponding coordinates from the filtered output.
5. The world-map D3 visualization then displays the resulting UFO sightings coordinates. The tooltip points to the corresponding location.