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

## Data Preparation

We started with analyzing the UFO sightings data. The most important column identified was the location as it should be used to get the latitude and longitude using geocode API. We observed following nature of the location field in the data which were later exploited to decode the ISO region for the location.

1. Total number of sightings are 61067. Number of non-US sightings based on a cursory analysis is 9019. Out of these there are around 21k unique locations.
2. Most of the US sightings followed the notation of (county\_name, state\_code) For example: - Santa Cruz, CA.
3. Some US sightings are mostly random and has Freeway/Expressway names
4. There is no similar notation used in non-US sightings except for the countries Canada, Australia and some others. These countries have the country name in the location string.

### Finding Nearest Airport

We used the following airport dataset**: -** <http://ourairports.com/data/airports.csv>

This file contains around 53k airports with the longitude, latitudes, iso\_country and iso\_region. First thing we had to do was get the latitudes and longitudes for the UFO sightings dataset and then compare it to the airports dataset to get the closest airport. The first instinct was to compare all the locations with all the airports. But soon we could see that it will be too many comparisons. Then we found that as most of the sightings are in US and there are around 22k airports corresponding to US region, we need to have a better strategy for reducing number of comparisons. So, we came up with the following strategy.

1. For all the US sightings, we will assume that it could be related to an airport in that state or the neighboring state. For this we created a key-value pair data structure which had every state’s neighboring state. Using the state codes of the neighboring states we formed ISO region codes such as US-CA (for USA California). We used these codes to narrow down on the list of airports to be compared to the UFO sighting location.
2. For non-US sightings we assume that the sightings will be related to the airports in that country alone. While getting the latitude and longitudes from geocode API, we also fire the reverse query to capture the ISO country code of that location. This helps reducing the number of comparisons in non-US sightings.

### Issues faced while getting longitude and latitudes from geocoding API

1. The OpenMaps geocode API would throw Too many requests error if overwhelmed with too many geocode requests. For this we had to store all the unique locations in pyMongo DB and then call the geocode API with a sleep of 1 second.
2. Some cities in US share names with cities in Canada and Mexico for such cases to get best results we had to include the name of country in the location query string.
3. Same city names across different states in USA. For these we had to include the state name in the location query string.

## Data Sets for additional 9 features:

### Dataset 1

Mime Type - JSON

Meteorite Landing - based on year and location we can get to know if people confused a meteorite for a ufo

F1=name of closest meteorite

F2=distance of closest meteorite to each city for that particular year the UFO was sighted

F3=Is it possible that they mistook it (ie: sighting happened at < 50 miles)

### Dataset 2 – Census Data

**MIME Type – text/CSV**

census - based on year and state, county we can get to know population density and housing density

F1=housing density

F2 =population density

F3=county

Source for Census data: <https://factfinder.census.gov/>

Additional data source: CSV of all counties and respective cities under each county across all the states of US. Open Source at <https://github.com/grammakov/USA-cities-and-states?files=1>. We have added below entries to this data file.

|  |  |  |
| --- | --- | --- |
| **City** | **County** | **State** |
| Iowa | Johnson | Iowa |
| Kansas | Wyandotte | Kansas |
| Oregon | Clackamas | Oregon |
| Plant | Hillsborough | Florida |
| Carson | Carson | Nevada |
| Watauga | Tarrant | Texas |
| Murphy | Collin | Texas |
| Bloomington | Hennepin | Hennepin |

#### Issues

* We could find the census data county wise for 2000 and 2010 years. For the years before 2000, the census data format is different and the demographic information is classified under different ethnic races.
* Problem with alaska state, cities not in the counties data

#### Handling issues - input data sources / code:

* Cleaned the data -- changed the column name
* changed PR col format -- added united states suffix
* Cleaned up data in county cities list :
* Parcel Return Service|DC|Washington D.C.|DISTRICT OF COLUMBIA|Parcel Return Svc
* PRS|DC|Washington D.C.|DISTRICT OF COLUMBIA|Prs
* Parcel Return Service|DC|Washington D.C.|DISTRICT OF COLUMBIA|Parcel Return Service
* Parcel Return Service|DC|Washington D.C.|DISTRICT OF COLUMBIA|Fedex
* Removed data related to DISTRICT OF COLUMBIA as no city names mentioned.
* Junk row N/A removed
* added missing data IOWA city of Johnson county of IOWA state.
* sighted at , reported at issues
* ---Incorrect locations of UFO sightings :
* "Laporte, WA"
* Ambiguous locations:
* Silver Beach, NY
* Spelling mistakes:
* [' Seatle', ' WA'] // Seattle
* [' Lewiston', ' MT'] //Lewistown
* [' Hollywood', ' CA'] // it is supposed to be WestHollywood
* [' Bluff City', ' AR'] //it is Pine Bluff
* [' Tawas', ' MI'] // East Tawas
* [' Tajunga', ' CA']
* [' Capistrano Beach/Dana Point', ' CA']
* [' St. Louis (Sappington/Crestwood)', ' MO']
* [' Ft. Worth', ' TX']
* [' Polk City', ' IA']
* [' Salt Lake City', ' UT']
* [' Terre haute outskirts', ' IN']
* [' Miami Township', ' OH']
* [' Salt Lake City', ' UT']
* [' Salt Lake City', ' UT']
* [' Troy to Montgomery', ' AL']
* 51547 sightings for valid 50 US states
* Incorrect sighting locationsn which we could not map to a county : 5883 out of 51547
* Issues with suburbs , cities &towns within suburbs

#### What features we extracted and joined with UFO data

#### What we inferred:

### Dataset 3

MIME Type - HTML

Sci-Fi Movies/shows (ie: star wars, star trek) released

F1 = Year in which the movies were released and in the same year the number of sightings took place

F2 = ratio of number of movies released to number of sightings that took place in that year

F3 = based on the ratio, if it possible that due to a high influence of sci-fi movies, people imagined aircrafts to be UFO