Chapter 1 : Datasets

Dataset	Link	Mime Type	Features Extracted	Explanation	Characteristics	Scripts
UFO Dataset with Latitude and Longitude	ufo_lat_long.json	application/json text/csv	Latitude Longitude	Used Nominatim and Geopy to populate Lat/Long for all locations in given ufo dataset	Sets the basis for location based matching and dataset merging	*ufo_add_lat_long.py* -Uses Nominatim and Geopy to find location for different sightings -Results of this are run with airport cities and string matching to allot lat-longs to those locations which were not processed by geopy
Airport Cities	Airport cities	text/csv	Airport name Airport distance	Name of the airport closest to the location of each ufo sighting instance and the distance between the location and the airport in miles	Finds the closest airports to the location to help infer if what people sighted as ufos where in fact airplanes	*airport_pollution_merger.py* -Calculates the distance between current location and all airports to find min
Air Pollution	Air Pollution	text/csv	SO2 mean CO mean O3 mean	Mean of parts per million per location within a given day	These gases cause both air pollution and result in global warming (as a result of greenhouse effect) so it helps make inferences on the effect of ufo sightings on air quality and global warming	*helper.py* -adds lat long to air pollution dataset *airport_pollution_merger.py* -merges ufo_lat_long with the air pollution dataset based on location (closest)
Drug Poisoning	Drug Poisoning Mortality Rate	application/json	Population County Death rate	Population of the county for each location of ufo sighting along with the death rate from drug poisoning (rates are deaths	This dataset suggests the mortality rate due to drug poisoning along for ufo sighting locations, so we may know if ufo presence	*integrate_drug.py* -Merges ufo_lat_long.py with drug dataset based on location and time (1996 - 2013)

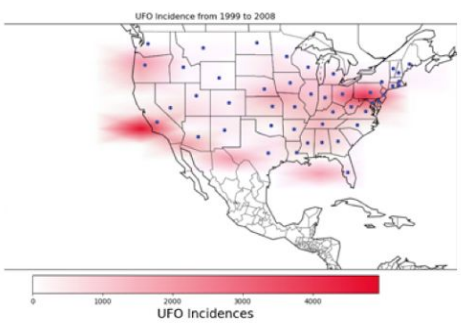
				per 100,000)	caused any kind of poisoning	
Cancer instances	Cancer	Proxy image (the owner has derived data by processing images) text/text	Cancer incidence counts in -allraces -white -hispanic	This dataset is matched based on time and state of the location for all ufo sightings that gives the number of incidences of cancer in different races	This dataset helps infer how ufo sightings affect the development of cancer in people. The motivation for choosing this dataset was mainly based on a thought that maybe, ufo's (if they exist) release some radiations (like cosmic rays) which prove to be fatal and cancerous to humans	*integrate_cancer.py* -Merges ufo_lat_long.py with cancer dataset based on location and time (1999 - 2014)

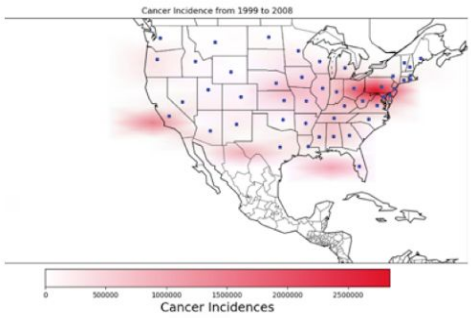
Chapter 2: Inferences

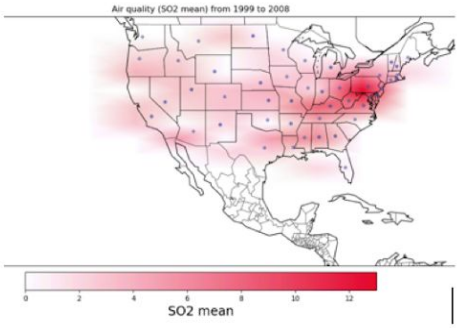
What you noticed about the dataset as you completed the tasks?
The ufo dataset came with 6 features.
We observed that of the 60095, 44892 reportings happened a day or a few days after the ufo was sighted. All reports have a description and only 2491 reportings are missing shape details about the sighted ufo.

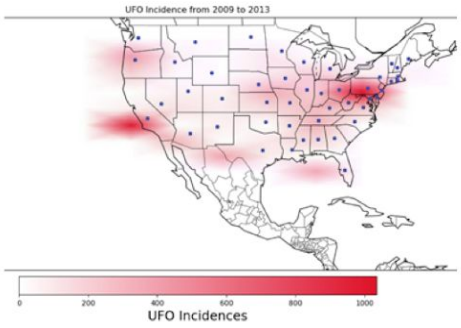
What questions did your new joined datasets allow you to answer about the UFO sightings previously unanswered?
By joining the Cancer incidences dataset with the UFO dataset we hoped to understand how (if yes) UFO sightings are related to the cancer incidences. UFO occurrences may have influenced the environment which may cause cancers.
We explored the relationship between UFO and cancer incidence from 1999 to 2013 for each state in the US. We plot the UFO and cancer incidence in the map (using matplotlib) to see the relationship. The figures on the left (below) are maps for UFO sighting and the figures on the middle (below) are the corresponding maps for cancer incidence. We can see that the densities of UFO sighting and cancer incidence have positive correlation. UFO sighting may cause high radiation which causes higher cancer incidences. We still need more data and analysis to validate this relationship.
By joining with the air pollution dataset with the ufo dataset, we hoped to find a relation between ufo sightings and air quality levels. We wanted to see if places with many sightings had poor air quality or otherwise.
There are mean value of SO2, O3, and CO to indicate the quality of air in the air quality dataset. We use SO2 mean as an indicator for the air quality and plot the average of SO2 mean (1999-2008, 2009-2013

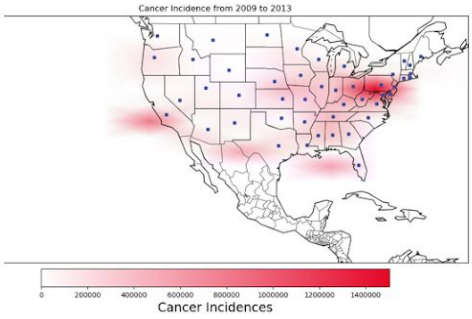
respectively) in maps. The figures in the right column below show the air quality in US. From figures, we can see there is no causal relationship between UFO incidences and the air quality.

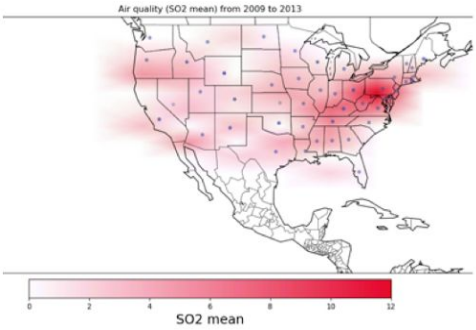












What clusters were revealed?

1. Where drug poisoning death rates are high, more UFO sightings are observed in that cluster
2. Where ozone(mean ppm) is high, UFO sightings were less
3. UFO Sightings are less in rural areas (defined by population size)
4. Number of cancer incidences were high where drug poisoning death rate was low
5. CO was high where drug poisoning death rate was low
6. Drug poisoning death rate was high where UFO sighting lasted a longer time (more than 3 minutes)

What similarity metrics produced more (in your opinion) accurate measurements? Why?

Jaccard produced more accurate measurements as it was easy to cluster based on features. Whereas edit distance/ cosine similarity measures didn't support a clear clustering measure along features.

What did the additional datasets suggest about "unintended consequences" related to UFO data?

The cancer incidence dataset indicates that the UFO sightings and high cancer incidence have positive correlation.

The air quality (SO2, O3, CO) and ufo sightings show weakly positive correlation. We can see that California has less air pollution as compared to the North East whereas both places have high density of ufo sightings.

UFO sightings are higher in big cities as opposed to rural areas or small cities. This inference is made from the population field extracted from the drug data set.

Do UFO sightings only occur in rural areas?

We define rural area based on the population. The total amount of population in US is about 46.2 million. There are 3007 counties in US. Based on this fact, if the population in a county is lower than 15,000, we call the county as rural area.

We have population data from 1996 to 2013. We count the number of times UFO sightings in rural areas and the total number of times UFO sightings in US per year from 1996 to 2013. The ratio of UFO sightings in rural areas from 1996 to 2013 to the total sightings is 15%. Hence, most of UFO sightings do not occur in rural areas, which is contrary to our assumption.

Are UFO sightings mostly (greater than 75%) occurring in areas within 25 miles of an airport?

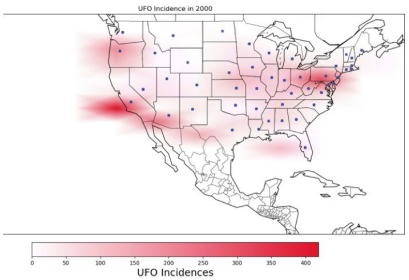
Yes. 99.54% UFO sightings happened in areas within 25 miles of an airport.

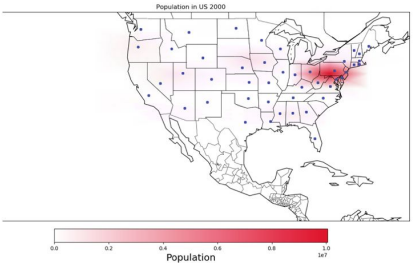
What do population demographics tell us about the areas in which UFOs occur?

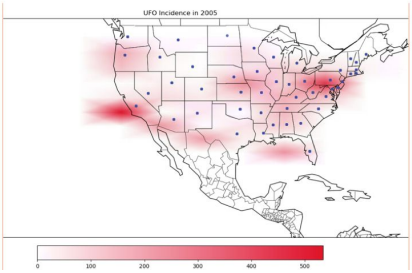
a. Densely populated?

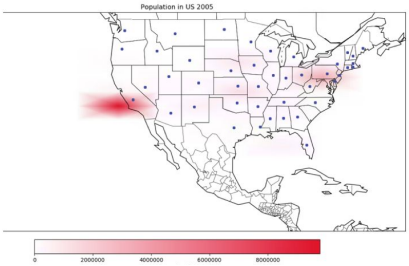
b. Sparsely populated?

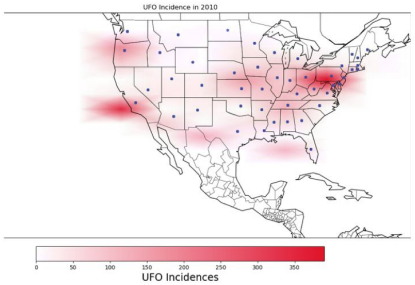
We have population data from 1996 to 2013. We have selected year 2000, 2005, 2010 to plot population density maps. We have also plotted maps for UFO sightings using the same years as the demographic density maps. The figures on the left (below) are maps for UFO sightings and the figures in the right column below are maps for population. In the figures, the population demographics tell us that the UFO sightings happen in both densely populated areas and sparsely populated areas.

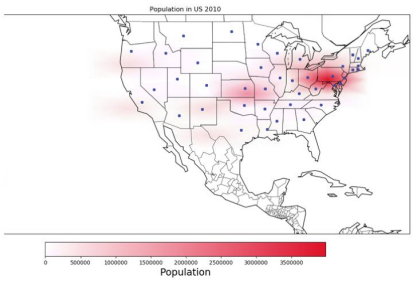












What insights do the "indirect" features you extracted tell us about the data?

First, extracted features from the cancer dataset tell us that the UFO sightings may have positive correlation with cancer incidences.

Second,extracted features from airport dataset tell us that most of the UFO sightings happen in areas which is very close to airports.

What clusters of sightings made the most sense? Why?

Where drug poisoning death rates are high, more UFO sightings are observed in that cluster. This made more sense because death rates were

high and air pollution measures also showed high pollution metrics, thus showing

Also include your thoughts about Apache Tika – what was easy about using it? What wasn't?
Apache Tika-Similarity was interesting to explore. We list the pros and cons (in our opinion) of the package here:-
Pros:

1. Readable code
2. Well organized
3. Easy to understand the flow

Cons:

1. Can add more parameters to make functions more flexible.
2. Doesn't run on python3 (need to change all functions like print to print() , etc)