COLORADO SCHOOL OF MINES

UFO Sightings and Wine Reviews

Jessicca Wolf, Joshua Hoskinson, Dani Barna, Sam Pauling, Amanda Giles

CSCI403 Final Project

December 10, 2017

1 Introduction

UFO sightings are mysterious things. Are they evidence of other intelligent life in the universe? Or are they simply drunken misinterpretations of rather mundane light sources? (i.e. streetlights, Christmas lights, the moon). Since the first question is rather difficult to answer, we tackled the second. Our analysis investigates whether drunken people see UFOs more.

There isn't data available for how intoxicated each UFO sighter was at the time of the sighting, so we instead used winery data by state, our rationale being that if there's a very highly rated winery in the same location as the UFO sighting then the person seeing the UFO probably consumes good wine, given the opportunity. We pulled both wine review and UFO sighting data off of Kaggle and created the ERD shown in Figure 2.3.

The Wine Reviews dataset contained the following information: winery name, description, variety, and price of wine as well as country and designation. The UFO sightings dataset contained latitude, longitude, and city of sighting, as well as the shape of the UFO. Our goal was to see if there was a correlation between UFO sightings and the number of wineries within a state.

The datasets used in our project do not have all the attributes of the original Kaggle datasets; we realized some of the information given in both the Wine Reviews and UFO Sightings data were not essential to our final data and therefore we did not include them. We also chose to only analyze data within the United States. This made it relatively simple to link the wine review and UFO data: each has a 1:N relationship with the state entity which makes it easy to track which state the UFO sighting and the wine review took place in.

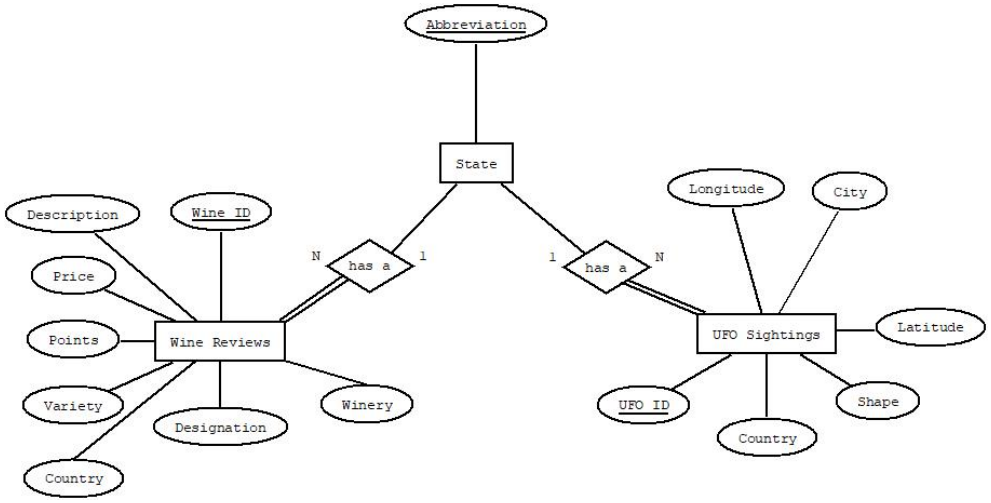

Figure 1.1: ERD Diagram Relating UFO Sightings and Wine Reviews

After making the ERD we were able to load the data into our database and use our final dataset to construct our visualization. Due to the mass quantity of UFO sightings we narrowed down the data we visualized to only sightings that were reported to last over an hour. Similarly, due to the mass quantity of data entries on U.S. wineries, we also narrowed down our count to highly rated wineries (90+ rating). After doing so we realized that some of

the data had special character or was incorrectly encoded.Since bad tuples made up a relatively small portion of the overall data we chose to simply delete the tuples from our data set. The query used to gather this subset of data and export it as a csv is as follows:

```
\COPY (SELECT ufo.latitude, ufo.longitude, ufo.city, ufo.duration, states.abbreviation
  FROM ufo, states WHERE duration > 5000 AND fk_ufo_state = abbreviation
  AND fk_ufo_state IN (SELECT abbreviation FROM states WHERE abbreviation
  IN (SELECT fk_wine_state FROM wine WHERE points > 90 AND price > 50)))
TO 'Z:/CSCI403/FINAL/output.csv' WITH CSV
```

2 Visualization

To visualize our data we used python within *jupyter notebooks*. Along with that, the libraries we used were *plotly* and *pandas*. We plotted the UFO sightings on an image of a United States based off of its latitude and longitude. The size of the circle to mark the location is sized based off of the duration of the UFO sighting. The color of the markers for each sighting depends on the number of wineries in the state: the markers are darker in states that have a larger amount of highly rated wineries and lighter in states with few to zero highly rated wineries. Since the interface we used to make the graphic (jupyter notebooks) could not handle the volume of data in our database, we only plotted data for five states.

In jupyter notebooks, our visualization allows for zooming in and hovering over a marker for more information. The information displayed for each marker includes the city, state, and duration of the UFO sighting.

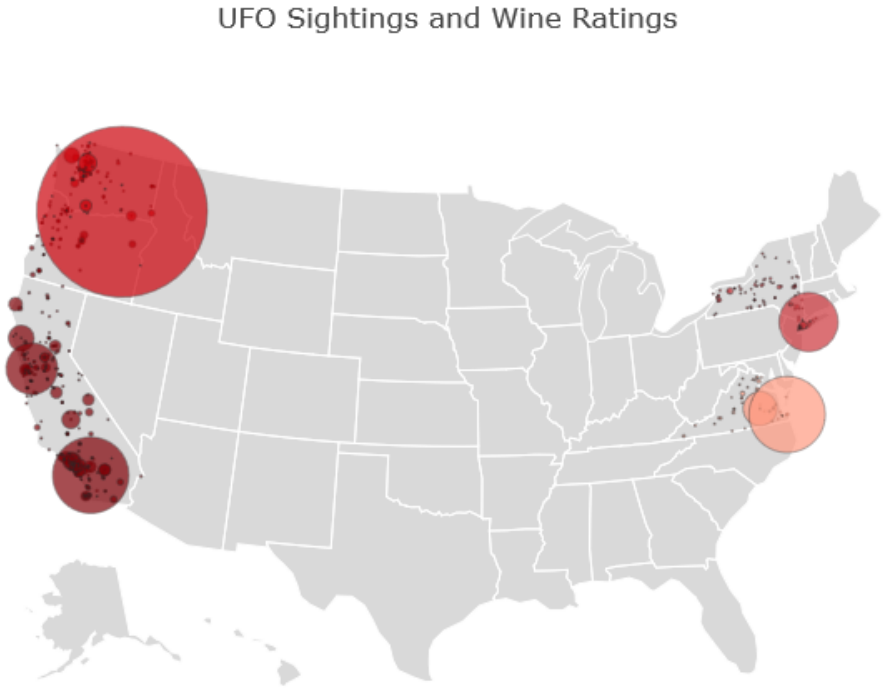

Figure 2.1: UFO Sightings in the U.S.

Figure 2.2: Color based on number of highly rated wineries in state

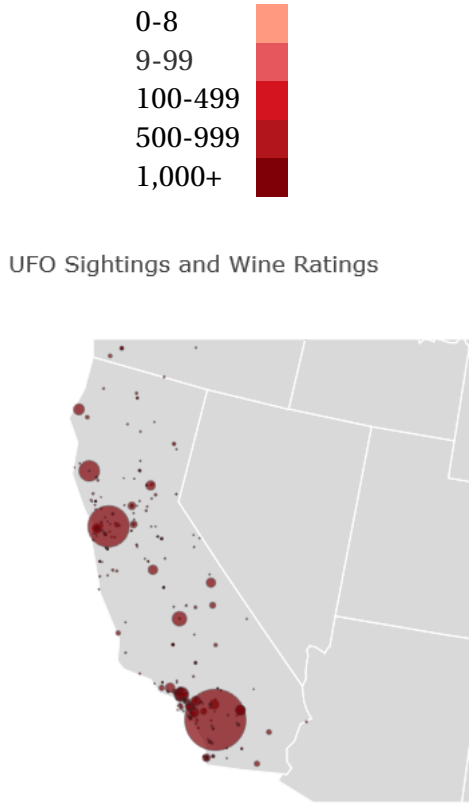

Figure 2.3: UFO Sightings in California

3 Challenges

Our biggest challenge while forming our database was deciding how to merge the two datasets. The raw datasets we found contained a lot of data that proved to be insignificant so it took a while to sort through and find the data we wanted to use. We also struggled with how the datasets related. We ended up narrowing down the datasets so the data was only from wineries and sightings within the United States. This made it easier to related the datasets and join them by state.

Still, even after narrowing our dataset down, we had to do some serious cleaning before our data could be a functional database. For example, the winery data spelled out each state's name, while the UFO data used abbreviations. This meant we had to either write a query so SQL knew that "Arizona" and "AZ" were the same thing, or do a find and replace in Excel. We also had this issue with country name. Beyond that, there were nonsensical values ("America" is not a state) and nulls in a lot of tuples. Cleaning and normalizing the data was a significant challenge.

There were several challenges that we faced while visualizing the data as well. First, finding the best library to use was tricky because many data visualization packages exist and have different benefits. We ended up using *matplotlib* and went modeled our visualization off of

an example online at the following website: https://plot.ly/python/bubble-maps/. Next, jupyter notebooks would not accept extremely large amounts of data. Therefore, we had to limit the data we displayed in the ways mentioned above. Finally, deciding effective ways to demonstrate a possible correlation was complicated because the information we were hoping to get across was the location of the UFO sightings, the duration of the sighting, and how some quantity relating to wine consumption. Based off of our data we think that the number of highly rated wineries is an indicator of wine consumption in that state. Overall, we were happy with our visualization and think it leads the viewer to draw some interesting conclusions.

4 Conclusion

Although we cannot assume causation, there have been many more UFO sightings in states with more highly rated wineries. Unsurprisingly, the locations of UFO sightings have occurred more often and for longer durations near major cities. This makes sense because there are more artificial lights near major cities that could be mistaken for UFOs, there are more people so more chances for sightings, and as evidenced by our visualization, we like to believe that the exquisite wine from the highly rated wineries improve one's ability to detect UFOs and gives them the courage to report it.

If we had more time and resources, we would collect data that was more in line with the actual amount of alcohol consumption, as well as the consumption of other substances like hallucinogenics. Along with that, we would plot more data points in an effective way that still highlights the more significant sightings. Finally, we would like to expand the project to also compare the amount of UFO sightings per country while comparing a country's alcohol consumption.

5 SQL CODE FOR DATABASE CREATION

```
DROP TABLE IF EXISTS wine CASCADE;
DROP TABLE IF EXISTS states CASCADE;
DROP TABLE IF EXISTS location CASCADE;
DROP TABLE IF EXISTS ufo CASCADE;
DROP TABLE IF EXISTS xref_wine CASCADE;
DROP TABLE IF EXISTS xref_ufo CASCADE;
CREATE TABLE states (
abbreviation TEXT PRIMARY KEY
);
CREATE TABLE wine (
country TEXT,
designation TEXT,
points INTEGER,
price INTEGER,
fk_wine_state TEXT,
region_1 TEXT,
region_2 TEXT,
variety TEXT,
winery TEXT,
FOREIGN KEY (fk_wine_state) REFERENCES states (abbreviation)
);
CREATE TABLE ufo (
city TEXT,
fk_ufo_state TEXT,
country TEXT,
shape TEXT,
duration NUMERIC,
latitude NUMERIC,
longitude NUMERIC,
FOREIGN KEY (fk_ufo_state) REFERENCES states (abbreviation)
);
\COPY states FROM 'C:/Users/Josh Hoskinson/Documents/CSCI403/states.csv' WITH (FORMAT csv);
\COPY wine FROM 'C:/Users/Josh Hoskinson/Documents/CSCI403/wine_data.csv' WITH (FORMAT csv);
\COPY ufo FROM 'C:/Users/Josh Hoskinson/Documents/CSCI403/ufo_data.csv' WITH (FORMAT csv);
ALTER TABLE wine ADD COLUMN wine_id SERIAL PRIMARY KEY;
ALTER TABLE ufo ADD COLUMN ufo_id SERIAL PRIMARY KEY;
```