Milestone Report

**Problem:**

Since 1961, Yuri Gagarin started human’s first journey to the universe, humankind has never stopped exploring the universe. One of the important topics people always discuss is whether we are alone or there are other civilizations in the universe. There are hundreds of reports of sighting UFO every year. However, we still do not have conclusive evidences to prove Aliens exist. Therefore, me as a fan of science fiction, I want to do a prediction on the duration, of UFO sightings so that in the future, base on the prediction we can prepare for the UFO sighting that last longer than average, and hopefully we are able to record this moment and help scientists to do further research on it.

**Client:**

NASA, astronomer, physicist and biologist would love to capture clear videos or pictures of UFO to do more research and study on it. If we can narrow down the location, time and duration of sighting UFO, we will have higher chances of capturing high quality UFO videos or pictures.

**Dataset:**

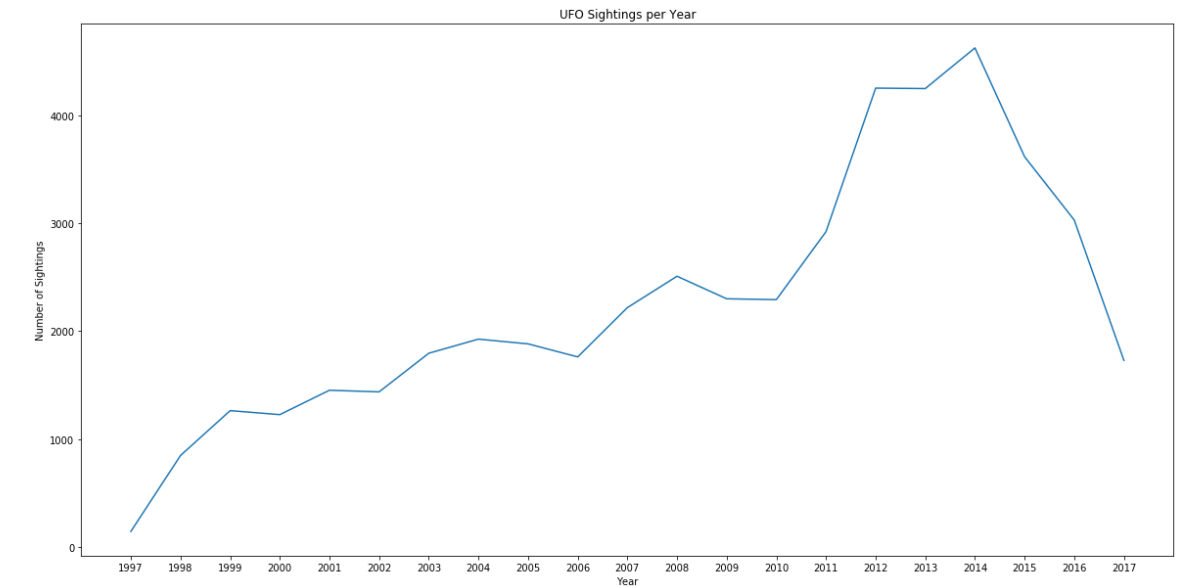
I used the “Consolidated UFO and Weather Data” dataset from <https://www.kaggle.com/emorelli/consolidated-ufo-weather-data>. This dataset contains over 51,871 reports of UFO sightings from 1997 to 2017 in 8845 cities, and it also contains historical weather data for every day from 1997 to 2017. Furthermore, I also obtained the population and population density data of 50 States from “Census Bureau” by using API.

**Data Cleaning and Wrangling:**

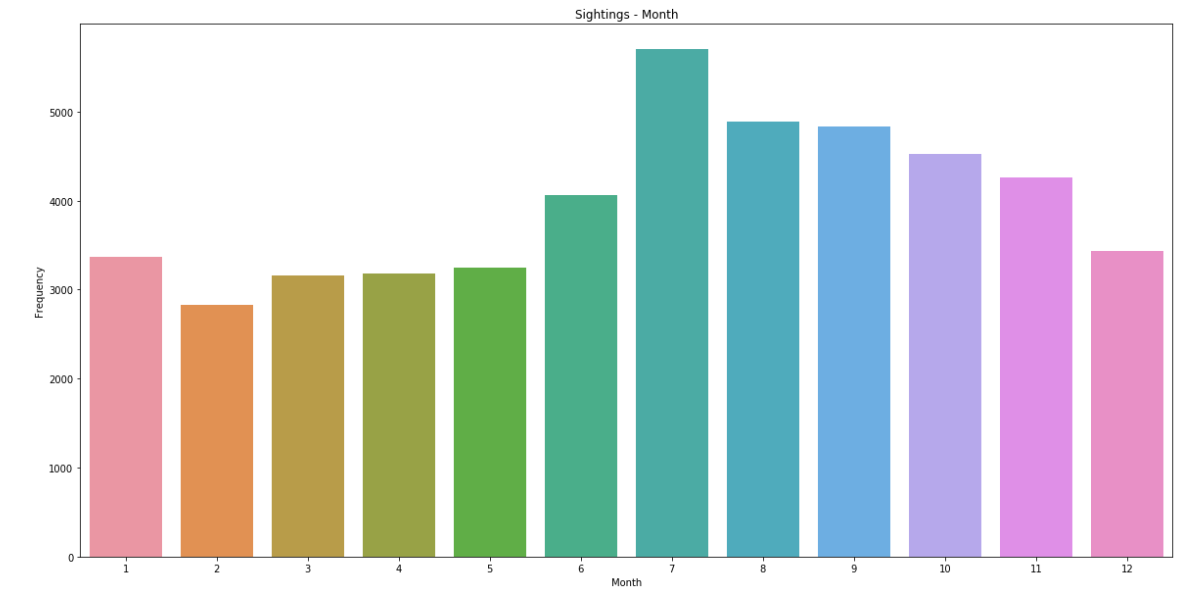
For data cleaning and wrangling, I mainly used “pandas” module in Python. First I dropped the duplicate columns of the dataset. Next I checked how many null values of every column in the dataset, and I dropped the columns that contain 80% or above null values. Also, in the “duration” column, the zero value rows are also null value so I dropped these rows also. Finally, I also merge the population and population density data into the original dataset.

**Data Story and Inferential Statistics:**

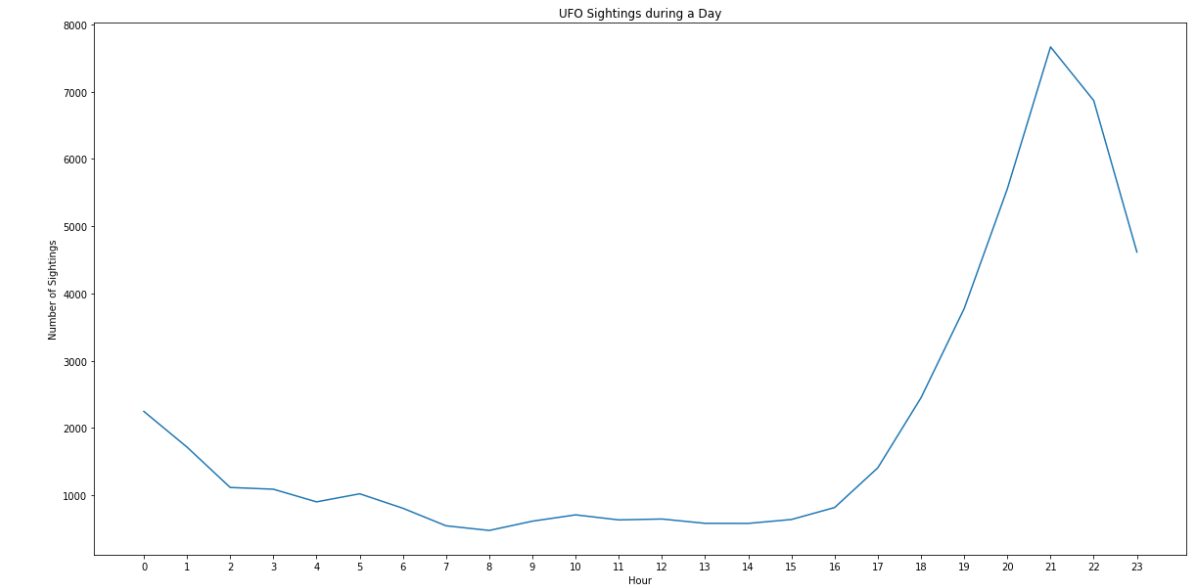
For data story telling, I used “numpy”, “matplotlib”, and “seaborn” modules in Python. First I plotted the trend of UFO sighting every years. From the plot we can see the number of UFO sighting has an upward trend from 1997 till 2014. 2014 was the year recorded sighting the most UFOs. The number of UFO sighting started decreasing after 2014.



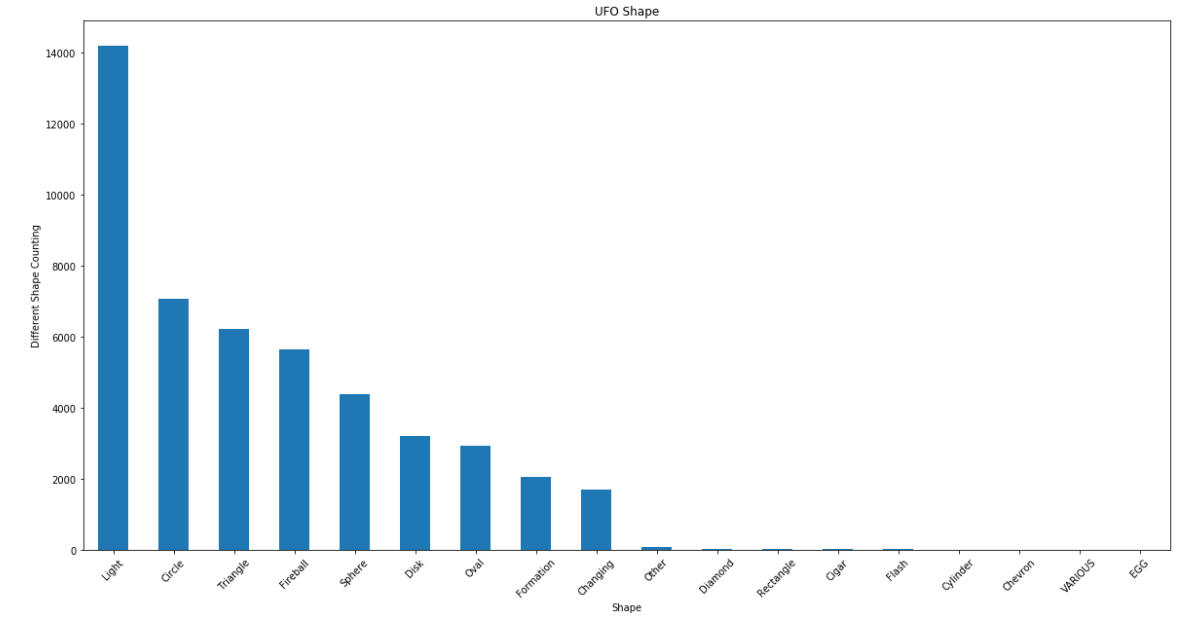
During a whole year period, summer and fall seasons have the higher frequency of UFO sighting.



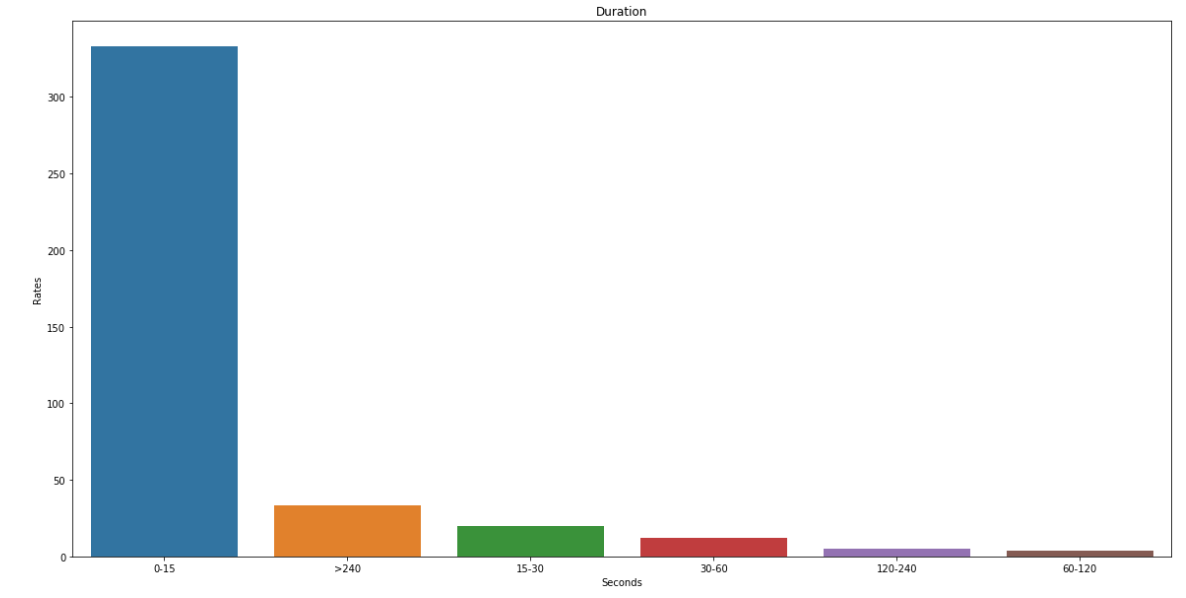
During a day period, UFO sightings mostly happened during night time from 6 p.m. to 12 a.m. The trend starts to drop down after 12 since people go to sleep.



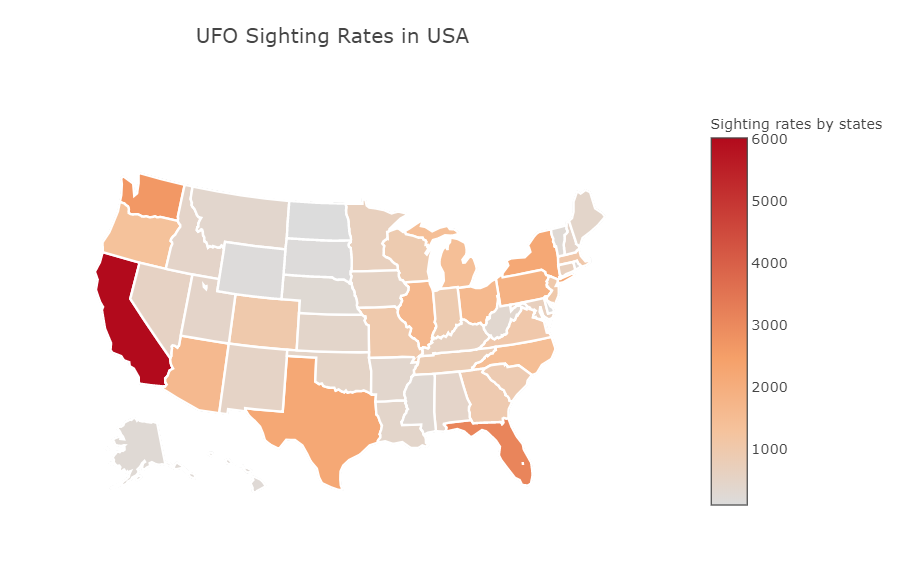
About 14000 of 51,871 reports, people said the UFO shape they saw was light. The second most common shape is circle, there are about 7000 reports said they saw circle shape UFO.



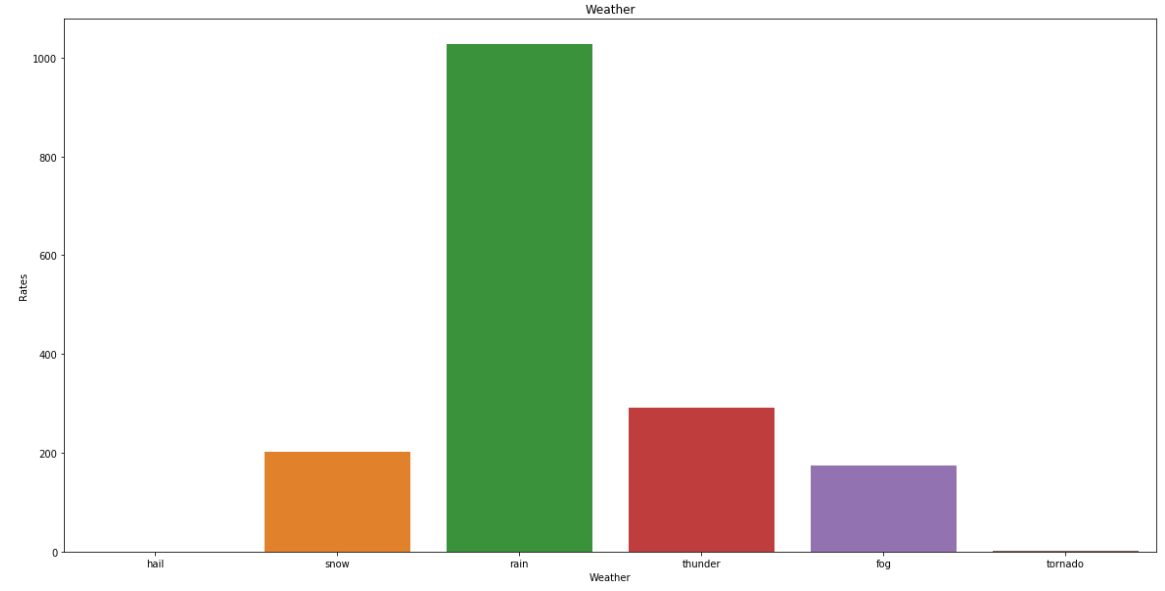
Next, I want to see how long did the UFO sighting usually last. Over 300 reports of UFO sighting only last zero to fifteen seconds. The second highest rates of UFO sighting last for over 240 seconds.



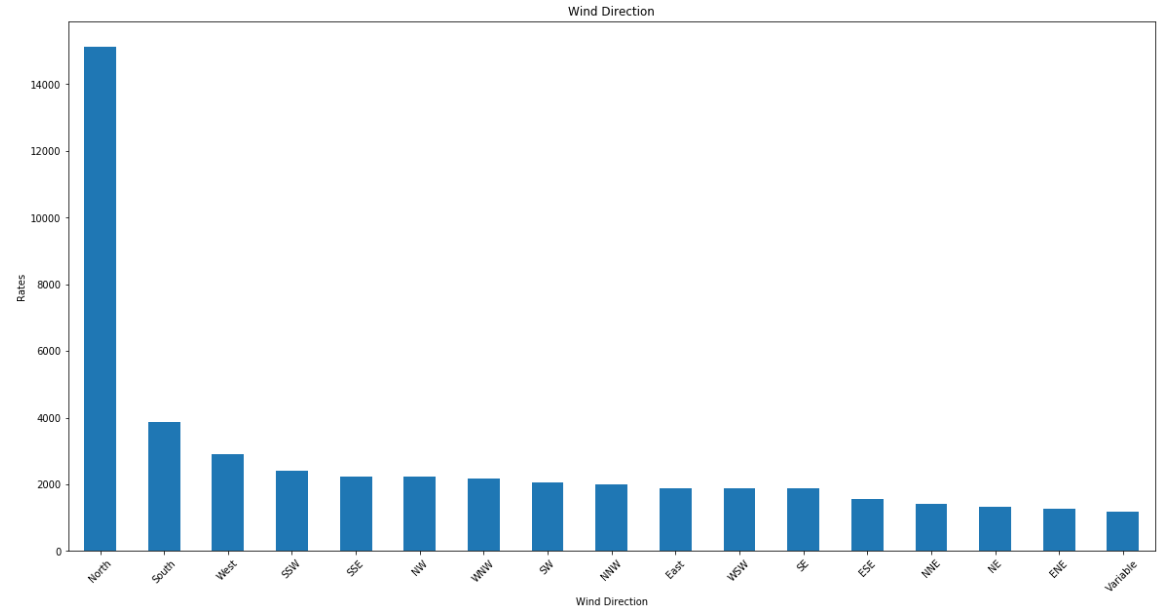
Next, I want to know where the UFO sighting occurred the most. Over 6000 of the UFO sighting happened in California, the next highest is in Florida about 4000 reports.



Furthermore, other than sunny weather, UFO sighting mostly occurred during raining weather

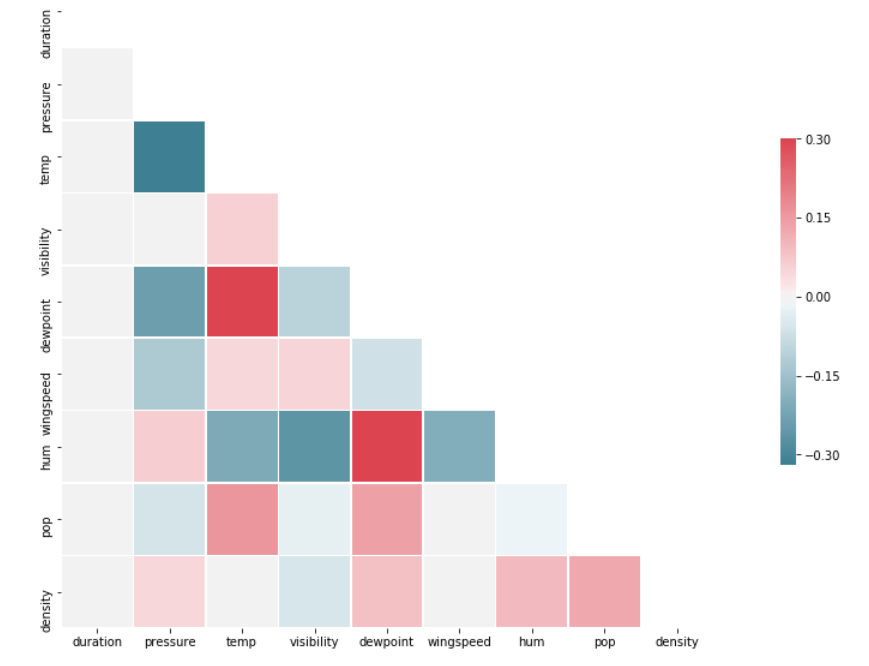


Surprisingly, UFO sighting happened heavily during north wind.



For inferential statistics part, first, I test if the duration of UFO sighting is normal distributed, since the p-value is equal to 0 which we can reject the null hypothesis and conclude that the distribution of duration is not normal.

Next, I want to find are there strong correlations between pairs of independent variables or between an independent and a dependent variable. I have done a correlation heat map before to see if duration is correlated with weather factors, population, and population density.



From the map we can saw that there are no correlation relationship between duration and weather factors/population. I test one of the correlation relationship between duration and population using bootstrap method, and since p-value is we cannot reject the null hypothesis and conclude that there is no correlation between populations and the sighting duration. Therefore, the test tells the same result as the map.

Next, I also want to test if there is correlation relationship between sighting rates and population/population density. I also use bootstrap method on both test. For the correlation relationship between sighting rates and population test, since the p value is less than 0.05, we can reject the null hypothesis and conclude that there is correlation between populations and the sighting rates. For the correlation relationship between sighting rates and population density test, Since the p value is greater than 0.05, we cannot reject the null hypothesis and conclude that there is no correlation between populations density and the sighting rates.