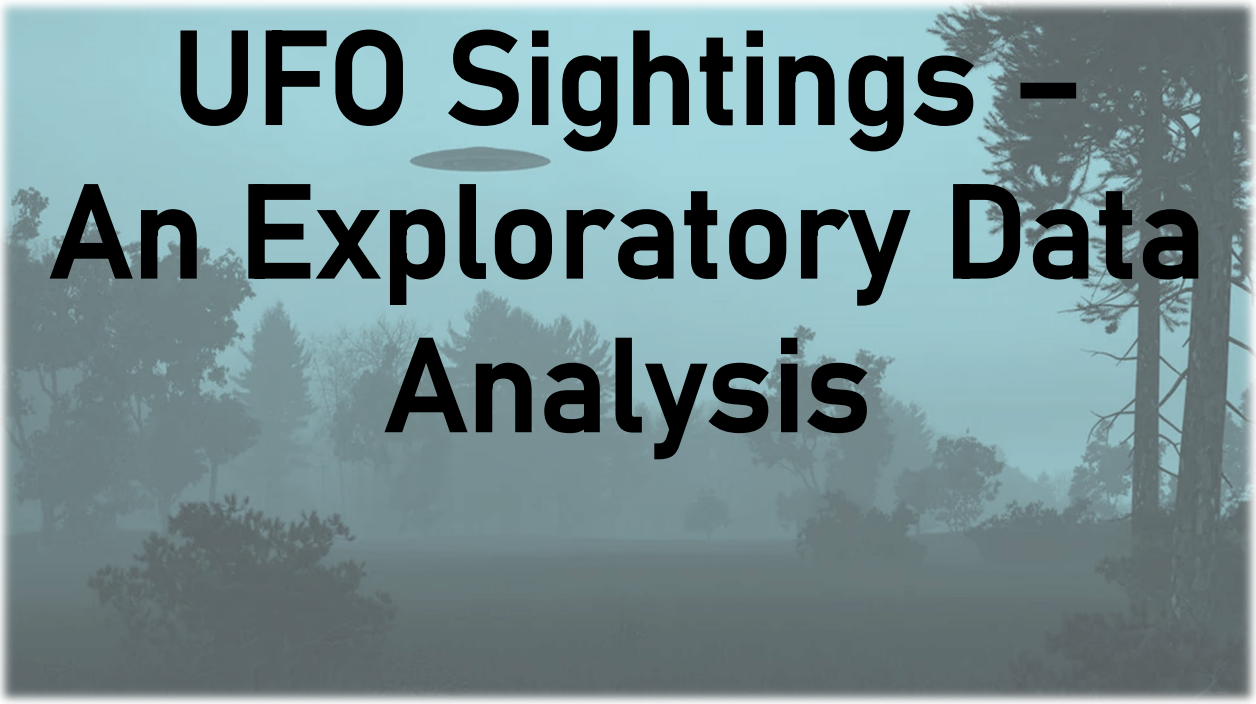


UFO Sightings – An Exploratory Data Analysis

Tenacious Data is:
Abi Chambers
Alejandro Perez
Carlos De La Rosa
Ann Ly

Project 1 - Group 2

Introduction

"The universe is under no obligation to make sense to you." -Neil deGrasse Tyson

But that doesn't keep us from trying; from looking up into the vastness of the dark, night sky, and asking questions about what we see.

The first UFO sightings were reported as early as the 1940s[1]. Since then, with the help of technology, UFO sightings reported from all over the world have been growing exponentially, although this data analysis will focus only on those sightings reported within the United States.

We chose to do our analysis on UFO sighting data out of genuine interest, as well as the accessibility of the data. Coming from Kaggle, the dataset we used was rated well amongst other users, and had been used before in several other EDAs[2], some of which also looked at correspondence with meteorite landings and the release of various UFO-related popular movies. In general, most previous EDAs resemble this one in their analysis of elements such as sighting locations, trends across the years, and the shapes of UFOs sighted.

Data Information and Cleaning

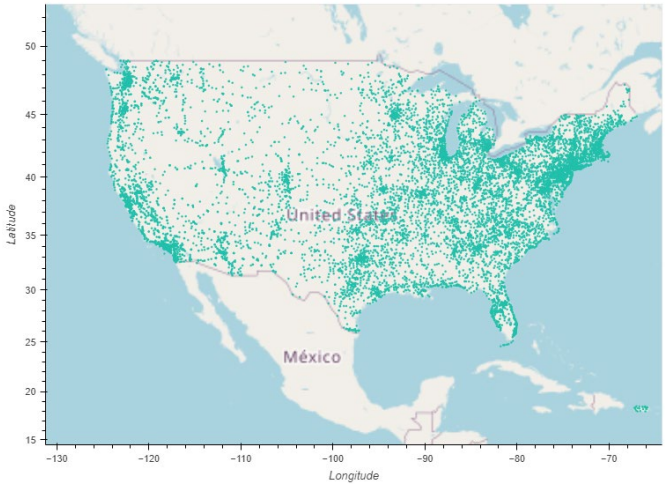
This analysis was based on a single CSV from the Kaggle website's publicly available datasets, titled 'UFO Sightings around the world'[3]. Using Jupyter Notebook, in conjunction with Pandas, we read in our CSV, containing eleven columns and over 80,000 rows of data and began the exploration and cleaning process. As this dataset was primarily self-reported data, there was much converting and tidying to be done. We eliminated a couple of columns as redundant, and therefore unnecessary, as well as a 'description' column that was strictly a text retelling of individual encounters. The raw data in the remaining columns was almost entirely in object data types, so reformatting columns containing information such as dates and times was high priority. Also, our data spanned around eight decades of time, so we separated our single column of date and time data out into a few additional new columns, and did some grouping and binning to better categorize our sightings over the years. We dropped rows with null entries in the location information, only reducing our data by less than 20%. It was determined that further narrowing of the data down to just those sightings reported in the United States was our best course of action, in terms of efficient and effective analysis, particularly given that U.S. sightings represented over 90% of sightings worldwide.

Research Questions & Analysis

To explore hypothesis and possible relationships present in our data, we again used the Jupyter Notebook platform, with various libraries imported, such as pandas, NumPy, SciPy, and Matplotlib, amongst others. We then posed, and attempted to answer through analysis and visualization, the following four research questions:

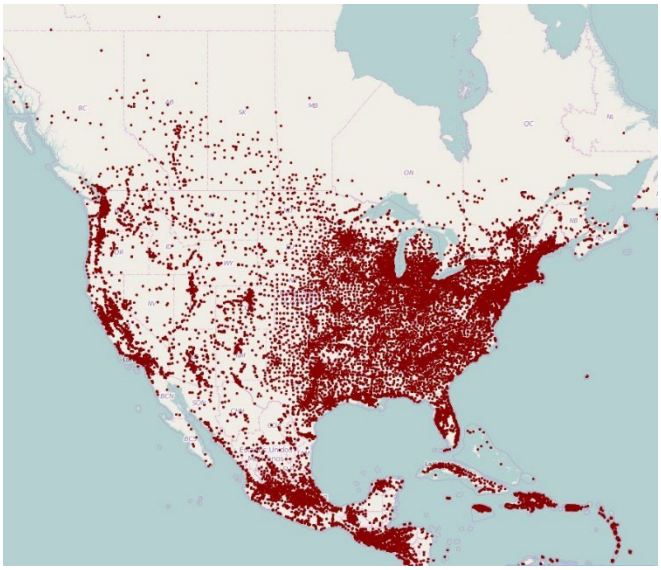
1) Are there trends in the locations of UFO sightings in the United States?

Plotting the latitude and longitude coordinates onto a map, we saw a typical trend with U.S. maps. The sighting locations were denser in cities/regions with high populations. It is easy to identify large cities (i.e., L.A., San Diego, Dallas, Miami) and high population areas such as the Pacific Northwest coast and the Northeast coast. We pulled a similar map from the Visual Capitalist website[4] that displayed a population density map of the U.S. There is either a high correlation of sighting locations and population density, or it could just be that sightings reflect a randomized sample of the total population. Hence, the plots look almost identical.



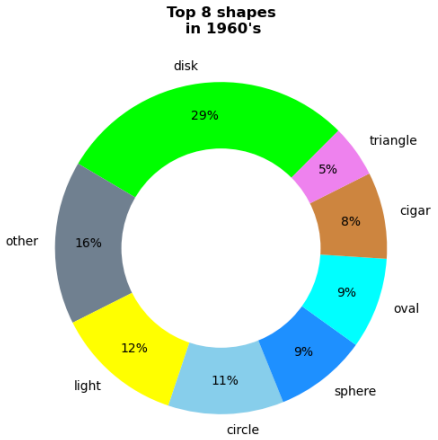
Tenacious Data UFO sightings map plot

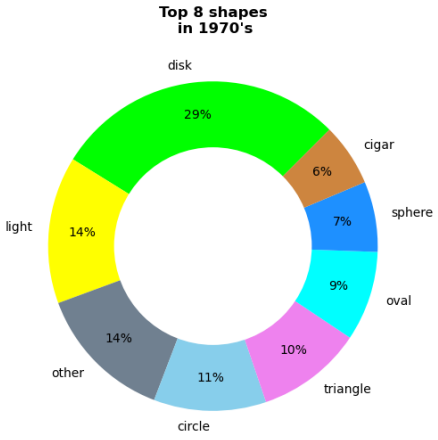
Visual Capitalist population density map

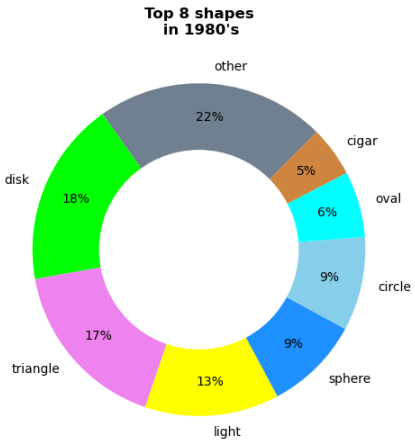


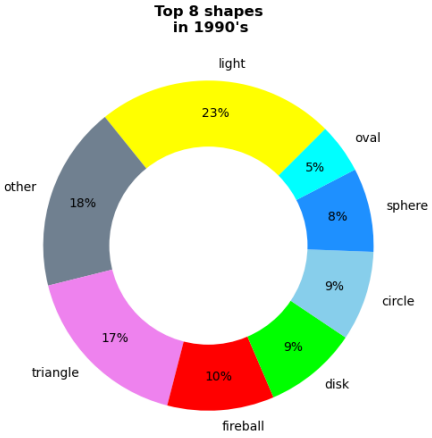
2) If so, do those trends extend to UFO shapes sighted, or the lengths of the encounters?

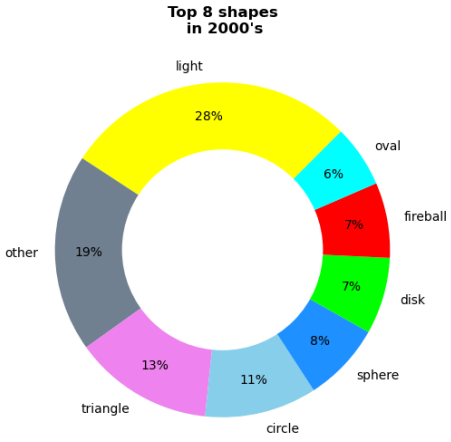
We proceeded to filter our data so that we could look at the popularity of UFO shapes. Given that we had 19 shapes and over 100 years of data, we opted to look at the top eight shapes across six decades. The decades prior to the 1960s were a small portion of our population, and we opted to not include them in our plots. We were able to see a couple of trends. The number of disk sightings decreased in percentage of sightings with respect to the top 8 shapes. This may be due to early radio programs, movies, or media displaying a disk shape.

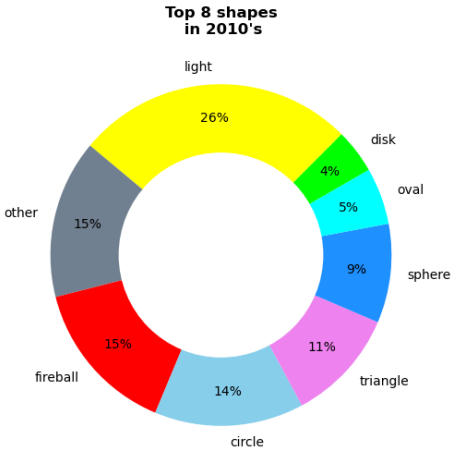






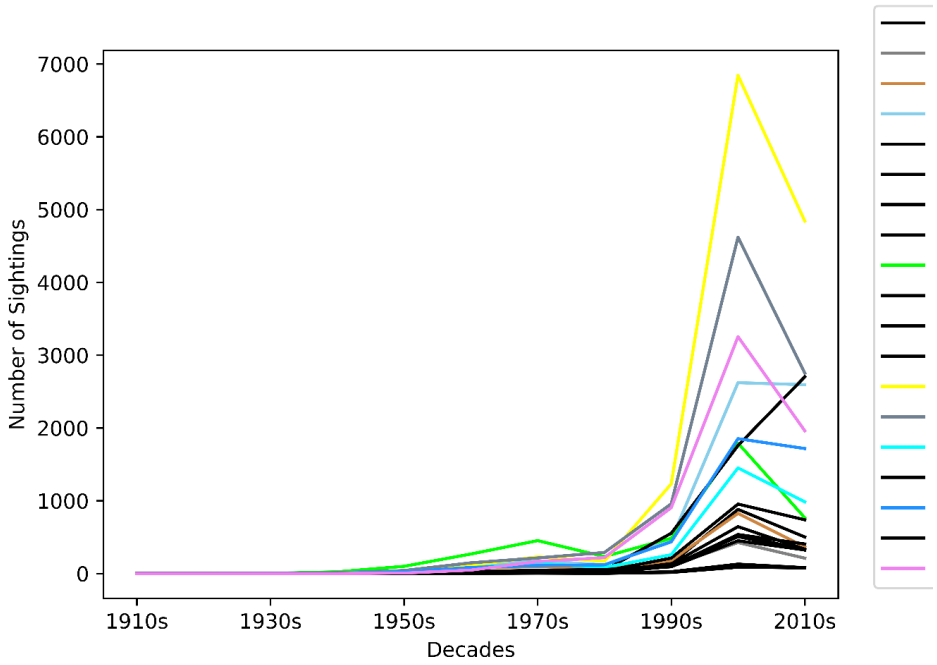






Another trend is how the “light” shape grows throughout these decades and becomes the most frequently sighted shape from the 1990s onward. The sightings of triangle shaped UFO’s is around the time when advancements in aeronautics saw an increase in triangle shaped aircraft such as the Lockheed’s F-117 and Northrop Grumman’s B-2. Aircraft like these benefit from their triangle shape because of stealth technology. Finally, an interesting thing we see is the cigar shaped sightings in the 60s to 70s, and then they taper off. This is likely UFO sightings being reported after seeing zeppelins, which flew during this era.

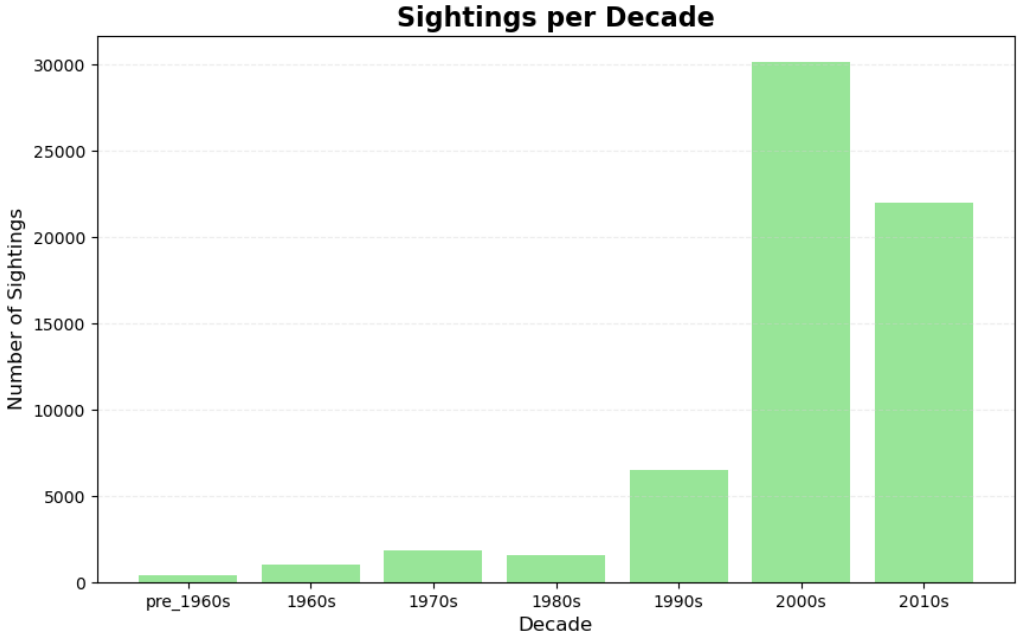
	1960s	1970s	1980s	1990s	2000s	2010s
0	Disk	Disk	Other	Light	Light	Light
1	Other	Light	Disk	Other	Other	Other
2	Light	Other	Triangle	Triangle	Triangle	Fireball
3	Circle	Circle	Light	Fireball	Circle	Circle
4	Sphere	Triangle	Sphere	Disk	Sphere	Triangle
5	Oval	Oval	Circle	Circle	Disk	Sphere
6	Cigar	Sphere	Oval	Sphere	Fireball	Oval
7	Triangle	Cigar	Cigar	Oval	Oval	Disk



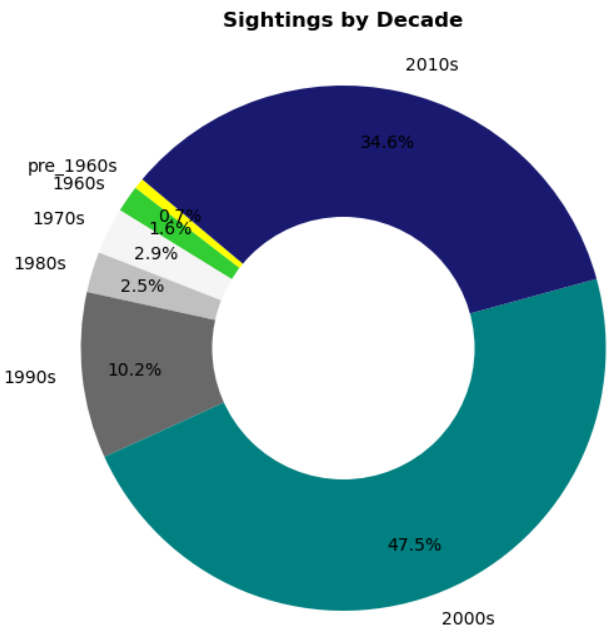
3) Are there any changes in sighting frequency or locations over time?

We found that our data spanned a great deal of time, over many, many individual dates, so we decided early on to bin according to decades, and even then, the earliest sightings in our data were few and far between, but stretched across several decades. That is how the decision was made to group all data prior to 1960 in a pre_1960s category.

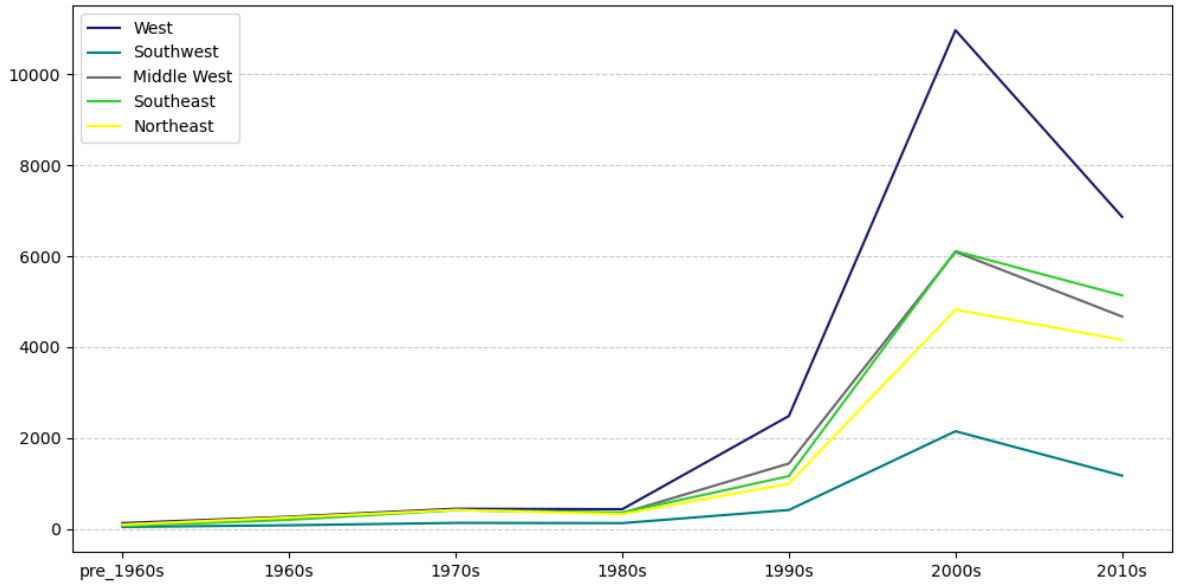
It seemed obvious to theorize that sightings would steadily increase over the years, and to an extent, we found that to be true, but not as literally as we initially suspected. While the number of sightings does see a consistent incline up until the 1970s, the data reflects a bit of a decline in the 1980s. As we move into the 1990s, we see a jump from about 1,600 sightings over the 80s, to 6,500 sightings; a growth that makes sense, given the rise of the internet in American households. The most glaring and significant escalation, as shown in the below bar chart, happens as we move from the 1990s to the 2000s, when the number of sightings rockets to over 30,000 across the decade. At this point, the internet has become much more common and accessible to the average person[5], which provides us with a partial insight. But there was another interesting influence our team discovered in its research. There is a relatively well-known, and certainly well-documented, phenomenon society sees as the result of popular culture's influence, known as the Scully Effect, that was the result of television series, The X-Files, which aired from 1993-2002[6] (with various limited returns, and movies through 2018, though we clearly see a bit of a decrease in sightings as we get into the 2010 decade).



We get a further-clarified idea of the proportion of sightings over the decades in the Sightings by Decade donut chart, where it is visually clear that the 2000s were the most active years for UFO sightings in the United States.



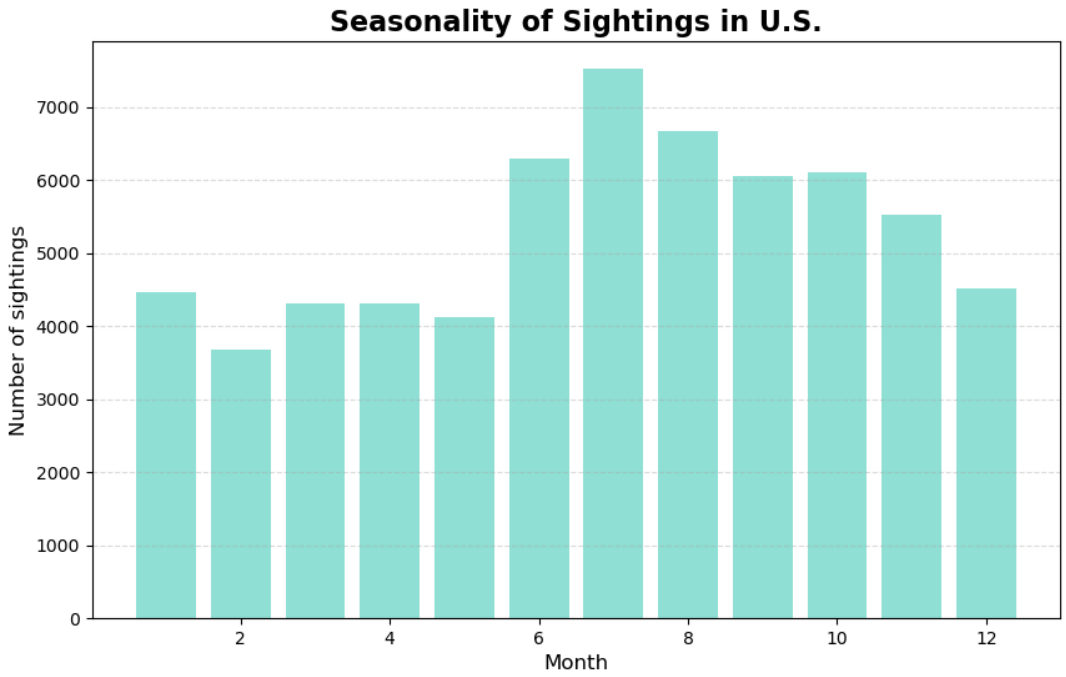
In terms of the number of sightings over the decades as they relate to location, for the purpose of clarity, we organized reported sightings into five major U.S. regions: West, Middle West, Southwest, Northeast, and Southeast. The data still shows us the most significant increase in number of UFO sightings from the 1990s to the 2000s overall, but in this line chart, it can be noted the West region of the U.S. reported the greatest increase, the Southwest saw the smallest increase, with the other three regions falling in the middle.



4) Are there any significant relationships between sightings and particular times of the year, or points in history?

Through an exploratory data analysis, we conducted further research to identify relationships between sightings and specific times or dates. To achieve this, we initially binned the data into U.S. geographical regions, which required additional exploration. One crucial step involved cleaning the column name 'state/province' to 'state_province' for consistency.

Once the data cleaning step was completed, we delved into the data by focusing on the month, region, and state_province. This allowed us to narrow down the information and extract the reported sightings per month for our regional dataset. By creating a new data frame, we could observe the frequency of values in these columns and further investigate the monthly data. We grouped the data based on month and region to facilitate the creation of a bar chart that visually displayed the significant months with the highest number of sightings.

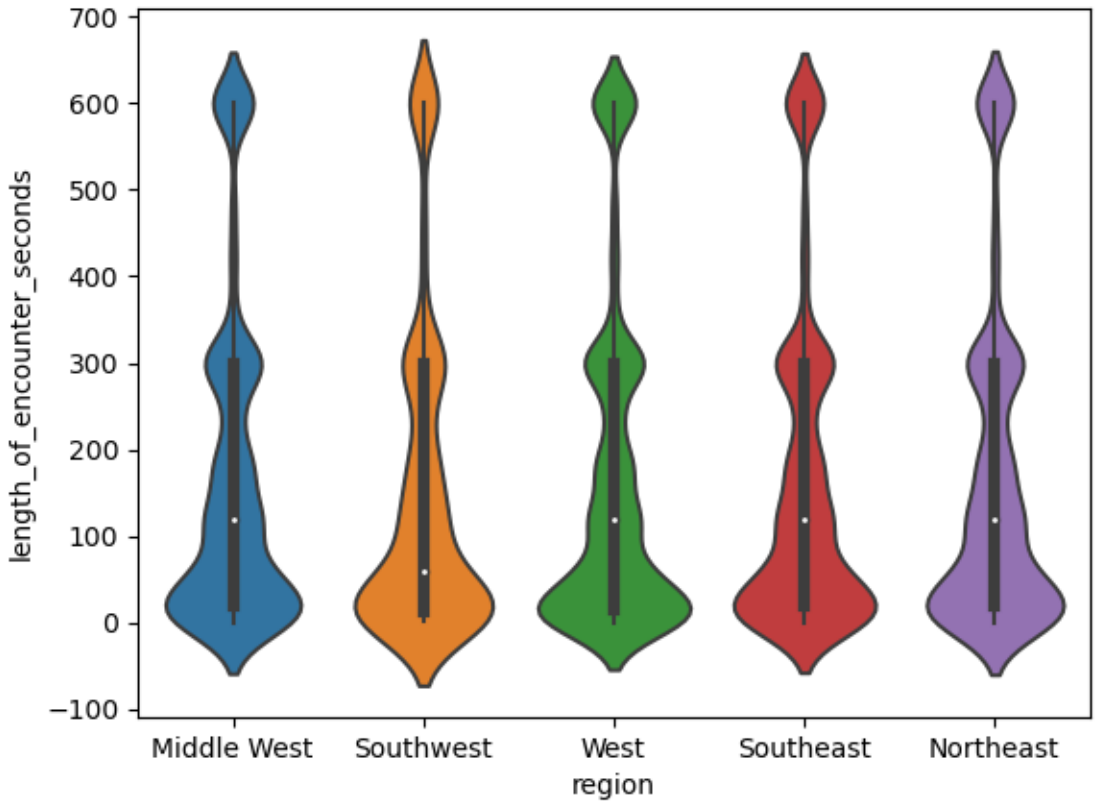


To enrich our analysis, we conducted additional internet research and discovered a correlation between UFO sightings and the summer months. This information was recorded by NUFORC (National UFO Reporting Center) and reported by R.D. Adams in an article featured on www.Beaumontenterprise.com titled "UFO Sightings"[7].

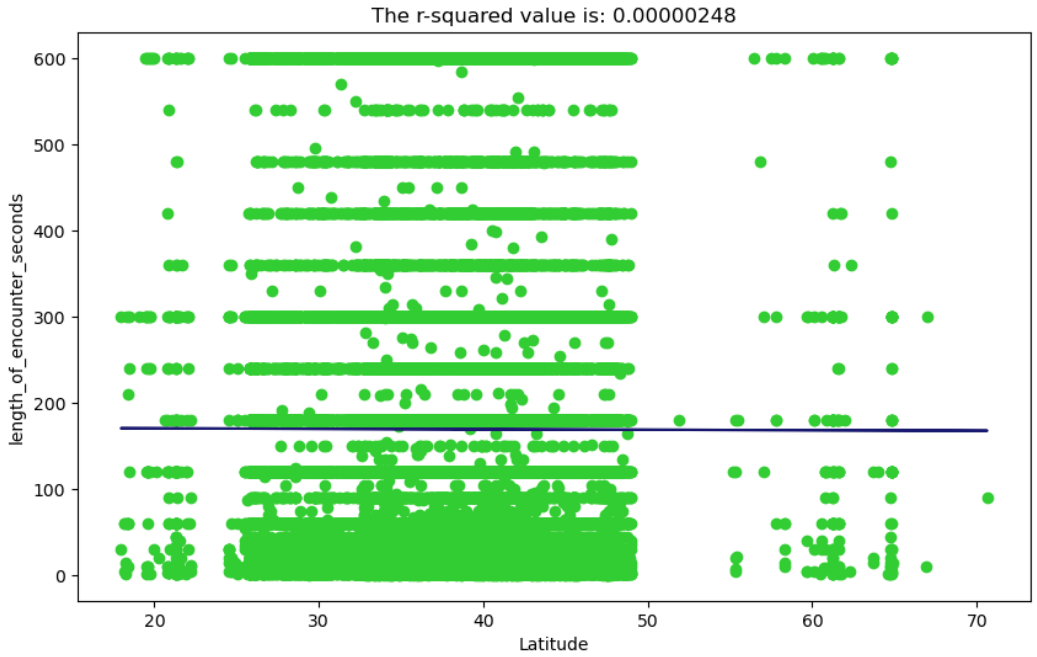
Regression

From our violin plots we can see that the average time duration of observations is relatively close across the 5 regions.

Middle West:	173.89 seconds
Southwest:	154.83 seconds
West:	170.38 seconds
Southeast:	168.05 seconds
Northeast:	169.19 seconds



Our null hypothesis is that the difference in means for each pair of regions is zero; in other words, they have the same average time of observation. We will reject the null hypothesis if the p-value is less than 5%. Therefore, we concluded that the means between the Southwest and the other four regions are NOT the same. We also concluded that the means of Middle West and Southeast are not the same.



Some Interesting History about UFO Sightings[8]

The first UFO sighting reported was assumed to be from the Puritans in March of 1639. They said that they saw "a great light in the night sky" while rowing a boat in the Muddy River. These events of UFO sightings continued through the years but did not get officially recorded and investigated by the government until the 1940s. In 1947, Kenneth Arnold reported seeing 9 objects moving at high speeds through the skies over Washington's Mount Rainier, which lead the government to form an organization called "Operation Sign." The purpose of this organization was to look into all of the UFO sightings reported in the United States. Operation Sign later changed its name to Project Blue Book in 1952. The organization was disbanded in 1969, making it the longest running government-based UFO investigation project of all time. During its time, the organization investigated over 12,000 cases; 90% of them were classified as "identified", and the remaining 10% were considered too vague or missing information.

Interestingly, one of the locations that has the most sightings reported in the United States is around the perimeter of Area 51. Area 51 is in the southern portion of Nevada, about 83 miles north-northwest of Las Vegas. Area 51 is officially used to construct and test aircraft for the military. The majority of the sightings around this area are debunked by experts, who believe that most sightings are simply aircraft test trials and maybe they were too "high tech" for normal people to grasp or comprehend. The fascination around these UFO sightings prompted television and movies such as The X-files, Independence Day, The Twilight Zone, Scooby-Doo and the Alien Invaders, and Paul.

Conclusions

Overall, it can be said that team Tenacious Data came to some expected conclusions, as well as one or two less expected. We fully anticipated that we would find trends in the locations of sightings but did not necessarily predict the bulk of those to be exclusive to the West region of the United States. We also presumed the data would show patterns in the number of sightings over time, though we were struck by the mild bi-modality in those numbers. Some of the trends our data shows are based on relatively obvious circumstances, such as population, ability to report a sighting, and mass media influence, but it was interesting to learn of trends that also exist among the shapes of UFO sighted over time. The clear summer seasonality of sightings was another observation we had not predicted in advance of analyzing the data, though it does make sense.

In terms of predictability based on this data analysis, we recommend a trip to the west coast some July if the hope is to catch a glimpse of something unidentified in the sky.

Limitations & Bias

Our data was not without some notable limitations. Perhaps most basically, it is presumed that the majority of our data was self-reported. This presents reliability concerns, due to the margin for human error, as well as the potential for individual biases to affect the information reported. Further, the dates of reported sightings stop in 2014. It would have given us a more accurate picture of trends and patterns if we had been able to include the most current data in our analysis. And finally, through a couple of different data mapping attempts, we learned that some of the location data is corrupt; that is, in the absence of assistance from geo API data, it would seem that some of the city data does not necessarily coordinate with the latitude and longitude data provided. That said, it is this team's opinion that corrupt data represents a small enough proportion of the overall data that the broad conclusions drawn from this analysis are still valid.

Future Work

As alluded to previously in this write-up, there was plenty of further research Tenacious Data would have been interested in stemming from this dataset. For instance, importing API census and geo data could allow for more precise location analysis, including map visualization. Also, in the vein of API or other additional datasets, looking at weather events, celestial happenings, and even star and/or planetary alignments in conjunction with the UFO sighting data could prove intriguing.

Works Cited

1. https://www.history.com/news/americas-first -ufo-sighting
2. https://www.kaggle.com/code/jonathanbouchet/e-t-phone-home-but-mostly-after-8-00pm
3. https://www.kaggle.com/datasets/camnugent/ufo-sightings-around-the-world
4. https://www.visualcapitalist.com/mapping-population-density-dot-town/
5. https://internetpkg.com/internet-history/#:~:text=The%20internet%20first%20became%20popular%20in%201991%2C%20when,by%20university%20scientists%20with%20limited%20access%20to%20it.
6. https://en.wikipedia.org/wiki/The_X-Files
7. Adams, R. D. UFO-sightings-by-state-best-places-to-see-UFO. (2022, July 27). https://www.beaumontenterprise.com/news/article/UFO-sightings-by-state-best-places-to-see-UFO-17330047.php (2022, July 27).
8. https://time.com/5627694/area-51-history
9. Aurora borealis – the Northern Lights. Chena Hot Springs, Alaska, 2013. LCDR Gary Barone, NOAA Corps (ret.), photographer. NOAA Photo Library.