Data Encounters of the Third Kind: Unraveling the UFO Phenomenon

Zane Alderfer, Ben Heindl, Victoria Haley

Spring 2024, IST 718 Final Project

Summary of findings:

This report presents a comprehensive analysis of UFO sighting data aimed at investigating patterns and trends to gain insights into potential extraterrestrial activities or atmospheric phenomena. The analysis utilized various analytical techniques, including spatial, temporal, shape-related, textual, and demographic analyses, to achieve the project objectives. Data cleaning and preparation ensured the integrity and reliability of the dataset, followed by spatial analysis revealing geographical patterns and hotspots of UFO activity. Temporal analysis unveiled fluctuations in sightings over time, while shape analysis identified common UFO shapes reported. Text analysis of comments associated with sightings revealed positive sentiments. Demographic analysis explored correlations between reported UFO sightings and socioeconomic variables, yielding weak correlations and highlighting data quality concerns. Spatial proximity analysis revealed a significant association between UFO sightings and proximity to major airports. The findings underscore the multifaceted nature of UFO phenomena, influenced by temporal trends, demographic correlations, and spatial proximity to major airports.

Recommendations:

Continued research and analysis, improved data quality, public awareness and education efforts, integration of multidisciplinary approaches, enhanced reporting and documentation practices, and mitigation of misidentifications to advance understanding and promote informed public discourse on UFO phenomena.

Specification

Problem Statement:

The project aims to investigate patterns and trends in UFO sighting data to gain insights into potential extraterrestrial activities or atmospheric phenomena. Additionally, it seeks to predict the likelihood and location of future UFO sightings. Understanding these patterns can contribute to public safety and awareness by identifying areas with higher UFO activity. Moreover, the analysis aims to explore potential correlations between reported UFO sightings and demographic variables obtained from census data, such as median household income, poverty percentage, and estimated population. Furthermore, the project examines the impact of proximity to major US airports on reported UFO sightings to understand potential misidentifications and perceptual biases.

Hypotheses:

The project will test several hypotheses based on the available data. Firstly, it hypothesizes that certain geographic regions may exhibit higher frequencies of UFO sightings compared to others, potentially indicating underlying patterns or phenomena specific to those areas. Secondly, it predicts that temporal trends in UFO sightings will reveal fluctuations or clusters over time, suggesting periodic patterns or events influencing sighting occurrences. Moreover, the project anticipates developing predictive models to forecast the likelihood and location of future UFO sightings based on historical data and identified patterns, contributing to proactive measures for public safety and awareness. Additionally, the project aims to investigate whether demographic variables such as median household income, poverty percentage, and estimated population correlate with the frequency or distribution of reported UFO sightings. It aims to explore whether there is a relationship between the number of reported UFO sightings and various demographic factors, without presuming any specific demographic variables to be more likely to report a UFO sighting.

Data Sources and Description:

The analysis utilizes multiple datasets, including UFO sighting data sourced from Kaggle, census data providing demographic information, US boundary files for geographical boundaries, and latitude/longitude to FIPS code conversion for precise location identification. The UFO sighting dataset comprises 67,401 rows and 18 columns, covering sightings from June 1930 to May 2014 across 50 states and 14,875 cities in the US. The most reported shape of UFOs is 'Light,' with 14,320 sightings. The average duration of sightings is 5,444 seconds (about 1 and a half hours). Additionally, the dataset includes demographic indicators such as median household income ($77,790), poverty percentage (12.41%), estimated population in 2022 (~835,205), and births and deaths in 2022 (8,904 births, 7,351 deaths). The examination of census data will provide additional demographic variables for analysis, contributing to understanding potential correlations with reported UFO sightings. Furthermore, the project will incorporate data on major US airports to explore correlations between proximity to airports and reported UFO sightings.

Observation

The analysis of UFO sighting data has revealed several intriguing patterns and trends, shedding light on the phenomenon and its potential connections. Our exploration encompassed various aspects, including temporal trends, demographic correlations, sentiment analysis, and the influence of proximity to major airports on reported sightings.

Firstly, the temporal analysis uncovered notable fluctuations in UFO sightings over the decades, coinciding with the historical timeline outlined in the article 'A Brief Cultural History of UFOs: From Secret Soviet Weapons to Alien Visitors' on PBS.org. The gradual increase in sightings until 1980, followed by sharper rises around 1994 and 2009-2010, aligns with the periods of heightened public interest and media attention to UFO phenomena documented in the article. The peak year for sightings

in our data, 2012, also corresponds with significant cultural and scientific discussions surrounding UFOs during that time. This synchronization between our findings and the historical narrative provided by PBS.org underscores the cultural and societal influences shaping UFO reporting behavior over time.

Secondly, our examination of demographic correlations aimed to elucidate any relationships between socioeconomic variables and UFO sightings. Surprisingly, weak correlations were observed between poverty percentage, median household income, estimated population, births, deaths, and the number of reported sightings. This suggests that factors beyond traditional demographic indicators may contribute to the frequency of UFO sightings, emphasizing the complexity of the phenomenon. Additionally, our analysis revealed statistically significant differences in the number of sightings across poverty percentage categories, as indicated by the results of the Kruskal-Wallis test. However, it is important to note a significant limitation in our analysis due to discrepancies found in the census data. While we meticulously prepared the data by addressing nulls and duplicates, some states had identical census/demographic information that appeared to be unlikely. Unfortunately, due to time constraints, we were unable to thoroughly investigate this issue, which raises concerns about the reliability of our demographic analyses. As a result, the findings regarding demographic correlations should be interpreted with caution, and further investigation into the accuracy of the census data is warranted should the project be continued. Despite this limitation, our analysis provides valuable insights into potential relationships between socioeconomic factors and UFO sightings, albeit with the caveat of data quality concerns

Additionally, sentiment analysis of comments accompanying UFO sightings revealed mostly positive sentiments, followed by some neutral and negative sentiments. This indicates a positive or neutral perception of UFO sightings among individuals reporting them, further highlighting the diverse nature of public attitudes towards this phenomenon.

Moreover, our analysis uncovered an interesting relationship among UFO sightings, major airports, and missing persons reports. UFO sightings were approximately 23 times more likely within 10 miles of a major airport compared to random locations in the US. This finding emphasizes the significance of considering environmental and contextual factors in understanding UFO sightings. Additionally, our examination observes the presence of missing persons locations in similar areas, coinciding with high population density. However, it's important to note that correlation doesn't imply causation, and the overlap between UFO sightings and missing persons incidents may not indicate a direct relationship. This further underscores the importance of holistic analysis when investigating phenomena like UFO sightings.

Summarizing our comprehensive analysis of UFO sighting data, we have gained valuable insights into this phenomenon. The temporal trends reveal a fluctuating pattern of sightings over time, while the weak demographic correlations and sentiment analysis findings underscore the multifaceted nature of factors influencing UFO encounters. Furthermore, the strong association between sightings and major airports highlights the significance of contextual considerations in interpreting UFO sighting data. However, it is crucial to acknowledge the limitations posed by discrepancies in census data, which may impact the reliability of certain analyses. Thus, further research and analysis are warranted to delve deeper into the intricate dynamics of UFO phenomena and their underlying causes, while also addressing data quality concerns.

Analysis

Data Cleaning and Preparation:

The initial phase involved meticulous data cleaning and preparation to ensure the integrity and reliability of the dataset. The UFO sighting data was loaded from a CSV file into a pandas DataFrame, and columns pertinent to location, time, shape, and duration of sightings were selected. Data cleaning

steps included converting data types, handling missing values, and filtering the data to focus on UFO sightings in the United States. These steps were essential to facilitate accurate analysis and interpretation of the data.

Spatial Analysis:

The spatial analysis undertaken in this project aimed to explore geographical patterns and hotspots of UFO activity. Initially, the dataset was aggregated by state to calculate the number of UFO sightings per state. This allowed for a comprehensive overview of UFO activity across different regions of the United States. The resulting bar plot visually depicted the distribution of sightings, highlighting states with the highest frequencies of UFO reports. Moreover, to delve deeper into the spatial distribution of sightings within top states, further analysis was conducted at the city level. For each of the top 10 states with the highest number of sightings, the analysis identified the top cities with the most reported UFO sightings. By visualizing this data through individual bar plots for each state, we gained insights into localized UFO activity within regions known for high overall sighting counts. Additionally, plotting sightings on a map of the United States using longitude and latitude coordinates facilitated the identification of spatial clusters and trends. This spatial representation enhanced our understanding of regional variations in UFO activity and provided valuable insights into the geographic distribution of sightings.

Temporal Analysis:

The temporal analysis conducted in this project aimed to reveal patterns and fluctuations in UFO sightings over time. Initially, the dataset was examined to discover correlations between the hour of the day and the frequency of sightings, shedding light on diurnal variations in UFO activity. Moreover, an exploration of correlations between the hour of the day and reported UFO shapes provided insights into potential temporal trends in sightings characteristics. Furthermore, trend analysis over time was

performed to uncover long-term changes and fluctuations in UFO activity. By identifying the year with the highest number of sightings and examining associated details such as the state with the highest and lowest sightings, the most common shape observed, and the longest duration of a sighting, we gained valuable insights into the temporal dynamics of UFO encounters. These analyses not only deepen our understanding of temporal patterns in UFO sightings but also inform future research directions in this field.

Predictive Analysis and Implications:

The analysis aimed to predict the date and location of future UFO sightings by leveraging various techniques on historical sighting data. Firstly, the data was grouped by datetime to quantify the number of sightings for each period. Time series decomposition and k-means clustering were then employed to identify temporal patterns and segment the data into clusters representing distinct periods. The clusters revealed a pattern of slow activity from 1930 to 1994, followed by a spike from 1995 to 2008, and a significant increase from 2009 to 2014. From these clusters, the one with the most recent sightings was selected, and the average time between sightings within this cluster was calculated. Utilizing this average time, the next likely sighting date was predicted. The predicted date assumes that sightings occur at regular intervals within the chosen cluster. Finally, to provide an up-to-date prediction, the analysis can be rerun from the current date, using the same methodology to project the next sighting date from that point onward. As of the date this report was finalized, the next likely UFO sighting will occur on April 28, 2024.

Furthermore, we predicted the latitude and longitude coordinates for the next UFO sighting by training linear regression models separately for both latitude and longitude. The linear regression model outperformed alternative models such as random forest regression and support vector regression, as evidenced by lower Mean Squared Error (MSE) scores.

Predicted Location:

- Latitude: 46.9966667
- Longitude: -120.5466667
- City: Ellensburg, WA

Model Evaluation (Linear Regression):

- Mean Squared Error (Latitude): 7.741127231675609e-31
- Mean Squared Error (Longitude): 9.85739125556265e-29

Alternative Models:

- Random Forest Regression MSE: Latitude: 1.3190029768825625e-05, Longitude: 0.00016442051192016774
- Support Vector Regression MSE: Latitude: 0.03419574237205408, Longitude: 0.06464941401996009

These predictions provide valuable insights into the temporal trends and the spatial distribution of UFO sightings. By combining both temporal and spatial predictive analyses, we can enhance our understanding of UFO phenomena and contribute to proactive measures for public safety and awareness.

Shape Analysis:

An analysis of the frequency of various UFO shapes reported was conducted to identify common appearances of UFOs. By quantifying the number of sightings associated with each shape, we aimed to uncover the most frequently reported UFO shapes, providing insights into public perceptions and descriptions of UFO appearances. This analysis is instrumental in discerning prevalent shapes reported by witnesses, which can potentially aid in distinguishing between misidentifications and genuine

sightings. Understanding the prevailing shapes observed contributes to a more nuanced interpretation of UFO encounters and enhances our ability to discern patterns and trends within the data.

Text Analysis:

Text analysis was conducted on comments accompanying UFO sightings, employing various techniques including word cloud generation and sentiment analysis This analytical approach was selected to uncover prevalent themes and patterns in public perception and responses to UFO encounters, thereby supplementing our comprehension of the phenomenon beyond numerical data alone. By delving into the textual content associated with sightings, we aimed to extract valuable insights into the collective interpretation and reactions to UFO sightings, contributing to a more comprehensive understanding of this mysterious phenomenon.

Demographic Analysis:

The investigation explored correlations between reported UFO sightings and demographic variables to uncover potential influences on sighting patterns. This analysis aimed to grasp the societal context and factors shaping UFO activity variations across demographic groups. Initially, a comparative analysis inspected various demographic variables alongside UFO sightings across US states. Utilizing visualizations like bar charts and histograms aided in assessing potential associations. Statistical tests, including the Kruskal-Wallis H Test, underscored significant disparities in sighting durations among different poverty percent groups, hinting at socioeconomic influences. Pearson correlation analysis revealed weak negative correlations between UFO sightings and demographic variables, yet none reached statistical significance. Furthermore, a multiple linear regression model failed to identify any significant relationship between demographic variables and total UFO sightings. Lastly, a pair plot visually explored potential correlations, offering valuable insights into the intricate interplay between

UFO sightings and demographics. Nevertheless, it is crucial to interpret these findings cautiously due to earlier concerns about data reliability.

Spatial Proximity Analysis:

The analysis examined the spatial dynamics between reported UFO sightings and major US airports, aiming to reveal any noticeable trends or correlations. By quantifying the distances between UFO sighting locations and major airport hubs, interesting insights were obtained into how the presence of aviation infrastructure may influence the perception and reporting of UFOs. This methodological approach was pivotal in addressing potential biases coming from misidentifications or heightened awareness due to airport activities, thus enhancing the comprehension of UFO sighting phenomena within the US context. Additionally, the examination revealed that UFO sightings were approximately 23 times more likely to occur within 10 miles of a major airport compared to random locations, signifying a 2226% higher likelihood of sightings in these areas relative to the overall region. Such findings emphasize the significance of considering spatial proximity to airports when analyzing UFO sighting data, shedding light on previously overlooked aspects of the phenomenon.

As such, the comprehensive analysis of UFO sighting data, encompassing spatial, temporal, shape-related, textual, and demographic dimensions, has provided valuable insights into potential extraterrestrial activities or atmospheric phenomena. Understanding these patterns and correlations contributes to public safety and awareness while advancing our knowledge of the UFO phenomenon. Further research and analysis in this field hold the promise of uncovering deeper insights into the nature of UFO sightings and their societal implications.

Recommendation

Based on the findings of our analysis of UFO sighting data, several recommendations can be made to further enhance understanding, research, and public awareness of this phenomenon.

Continued Research and Analysis:

Our analysis has provided valuable insights into the patterns and trends of UFO sightings. We recommend further research and analysis in this field to dive deeper into understanding the underlying causes and factors contributing to UFO encounters. This includes exploring additional datasets, conducting more advanced statistical analyses, and collaborating with experts in relevant fields such as astronomy, psychology, and sociology.

Improved Data Quality:

The discrepancies found in census data highlight the importance of ensuring data quality and reliability for future analyses. We recommend conducting thorough data validation and verification processes to address inconsistencies and inaccuracies in demographic datasets. We also suggest exploring alternative sources or methodologies to supplement census data, which may provide a more comprehensive understanding of demographic trends and reduce reliance solely on governmental datasets.

Public Awareness and Education:

Our analysis has revealed the multifaceted nature of UFO sightings, influenced by factors such as temporal trends, demographic correlations, and spatial proximity to major airports. We recommend increasing public awareness and education about UFO phenomena to promote informed and rational discourse. This could involve educational campaigns, public forums, and media outreach efforts to provide accurate information and debunk myths surrounding UFO sightings.

Integration of Multidisciplinary Approaches:

Given the complexity of UFO phenomena, we recommend integrating multidisciplinary approaches in future research and analysis. This includes collaboration between scientists, researchers, policymakers, and the public to explore various perspectives, theories, and methodologies. Incorporating insights from fields such as astronomy, psychology, sociology, and data science could provide a more holistic understanding of UFO phenomena.

Enhanced Reporting and Documentation:

Our analysis has highlighted the importance of accurate reporting and documentation of UFO sightings, including detailed information about location, time, shape, and duration. We recommend enhancing reporting mechanisms and standardizing data collection practices to improve the quality and reliability of UFO sighting databases. This could involve developing standardized reporting forms, leveraging emerging technologies for data collection, and promoting transparency and openness in reporting processes.

Mitigation of Misidentifications:

The significant association between UFO sightings and proximity to major airports underscores the need to mitigate misidentifications and perceptual biases in reporting. We recommend implementing measures to educate the public about common misidentifications of aircraft and celestial objects, as well as promoting critical thinking and skepticism in evaluating UFO reports. This could help reduce the number of false positives and enhance the credibility of UFO sighting data.

By implementing these recommendations, we can further advance our understanding of UFO phenomena, promote informed public discourse, and contribute to scientific research and exploration in this intriguing field.

References

Airport Data

Source: Esri

URL: https://hub.arcgis.com/datasets/esri::1000000-or-more/explore?location=31.537347%2C-96.116266%2C4.58

Citation: Esri (n.d.). Airport Data. Retrieved from https://hub.arcgis.com/datasets/esri::1000000-or-more/explore?location=31.537347%2C-96.116266%2C4.58

Eghigian, Greg. "A Brief Cultural History of Ufos, from Secret Soviet Weapons to Alien Visitors." PBS, July 8, 2021. https://www.pbs.org/newshour/nation/a-brief-cultural-history-of-ufos-from-secret-soviet-weapons-to-alien-visitors.

Lat/Long to FIPS Code API

Source: Federal Communications Commission (FCC)

URL: https://geo.fcc.gov/api/census/

Citation: Federal Communications Commission (FCC) (n.d.). Lat/Long to FIPS Code API. Retrieved from https://geo.fcc.gov/api/census/

Missing Persons Report

Source: Bugmaster (Data World)

URL: https://data.world/bugmaster/missing-persons-john-doe-jane-doe-known-females

Citation: Bugmaster (n.d.). Missing Persons Report. Retrieved from https://data.world/bugmaster/missing-persons-john-doe-jane-doe-known-females

US Boundary Files - Shapefile

Source: Natural Earth

URL: https://www.naturalearthdata.com/downloads/

Citation: Natural Earth (n.d.). US Boundary Files - Shapefile. Retrieved from https://www.naturalearthdata.com/downloads/

US Estimated Population, Birthrate, Deathrate

Source: U.S. Census Bureau

URL: https://www2.census.gov/programs-surveys/popest/datasets/2020-2022/counties/totals/

Citation: U.S. Census Bureau (n.d.). US Estimated Population, Birthrate, Deathrate. Retrieved from https://www2.census.gov/programs-surveys/popest/datasets/2020-2022/counties/totals/

US Poverty & Median Household Income

Source: U.S. Census Bureau

URL: https://www.census.gov/data/datasets/2022/demo/saipe/2022-state-and-county.html

Citation: U.S. Census Bureau (n.d.). US Poverty & Median Household Income. Retrieved from https://www.census.gov/data/datasets/2022/demo/saipe/2022-state-and-county.html

UFO Sighting Data

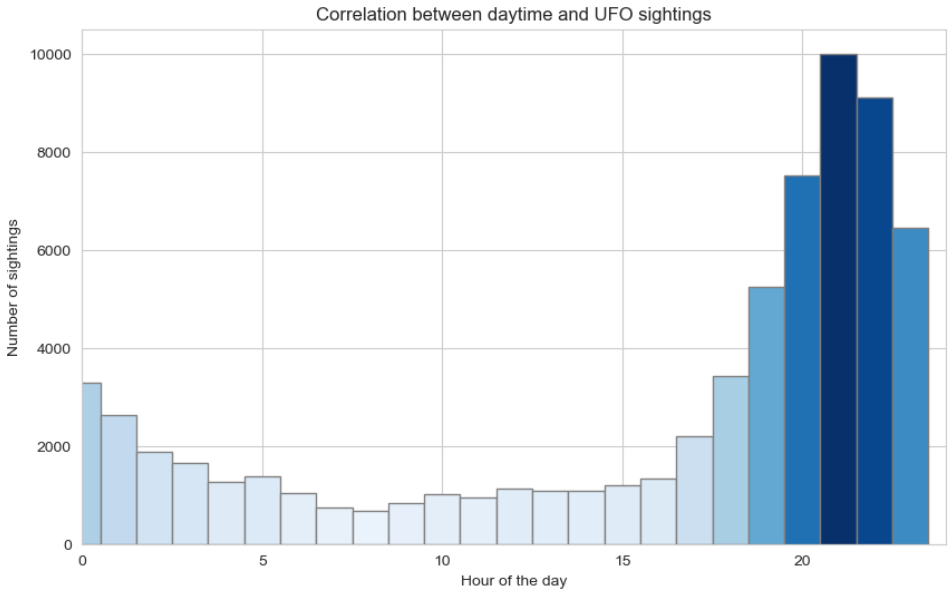
Source: National UFO Reporting Center (NUFORC)

URL: https://www.kaggle.com/datasets/NUFORC/ufo-sightings

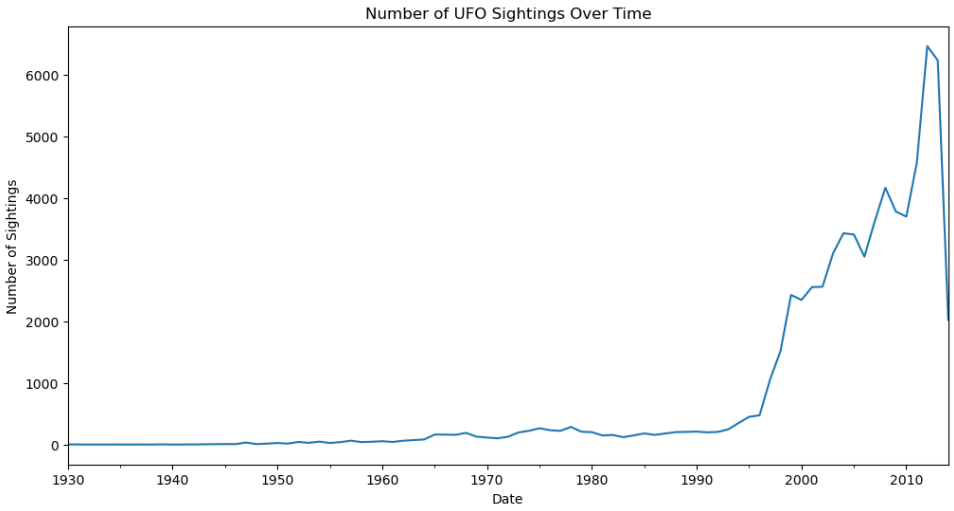
Citation: National UFO Reporting Center (NUFORC) (n.d.). UFO Sighting Data. Retrieved from https://www.kaggle.com/datasets/NUFORC/ufo-sightings

Appendices

Appendix A – Correlation between Time of Day and UFO Sightings



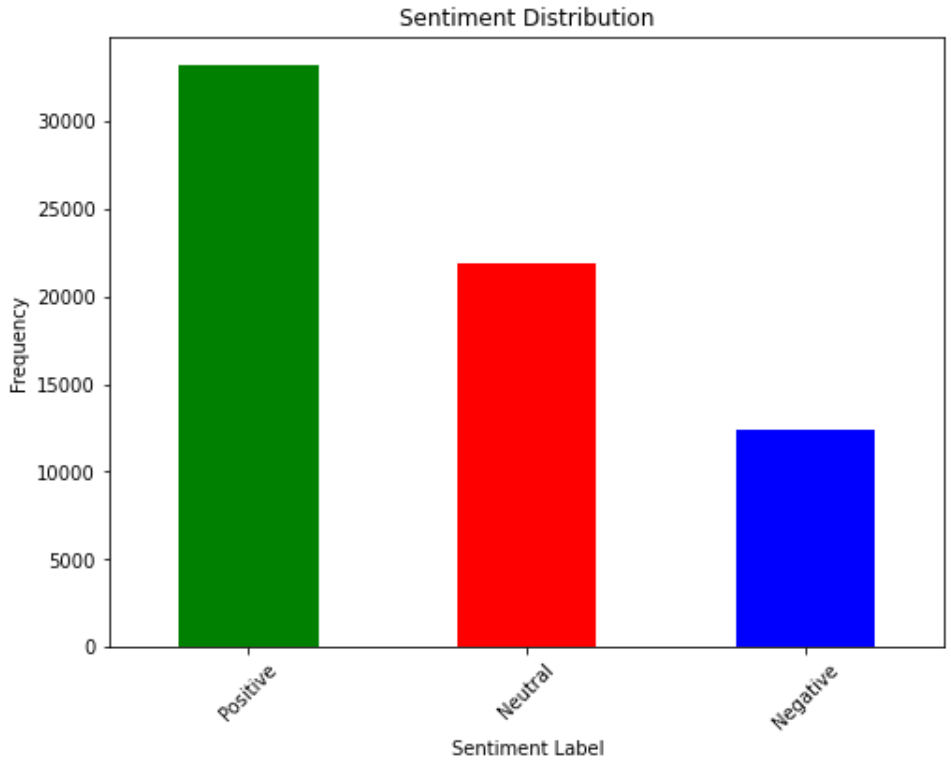
Appendix B – Time Series Plot of UFO Sightings



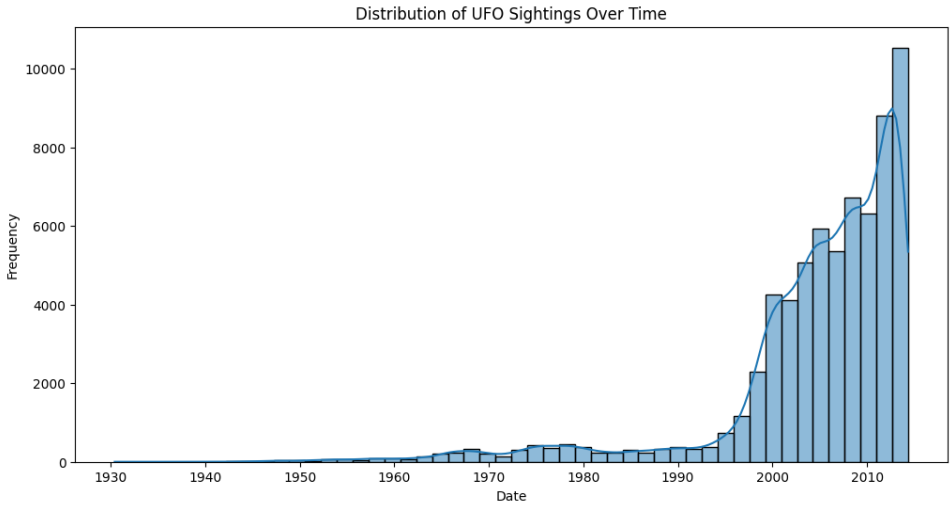
Appendix C – Comments Word Cloud

Appendix D – Shape Word Cloud

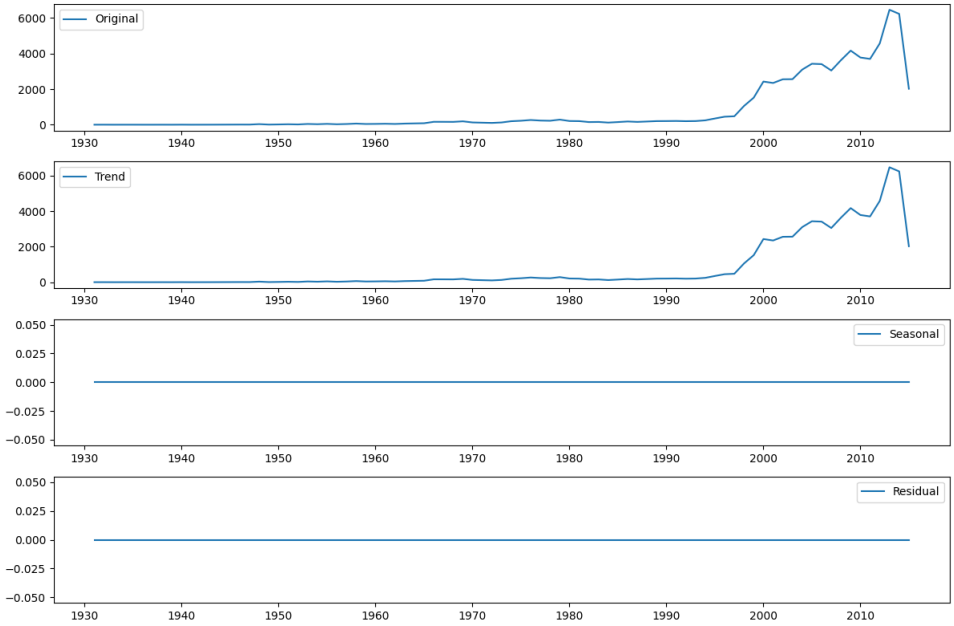
Appendix E – Sentiment Distribution



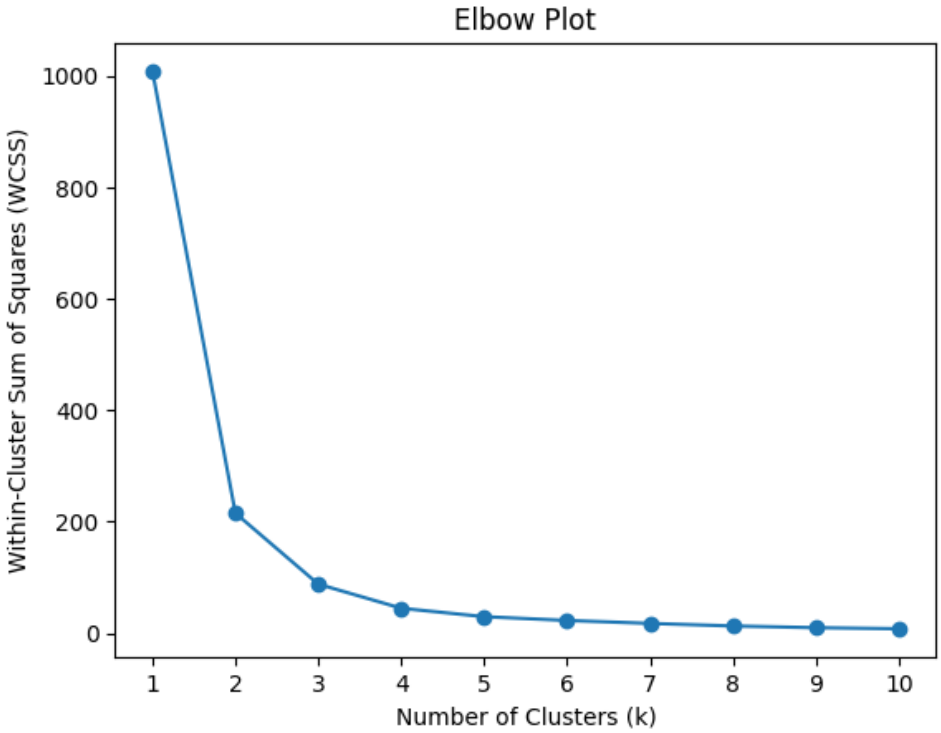
Appendix F – Distribution of UFO Sightings Over Time



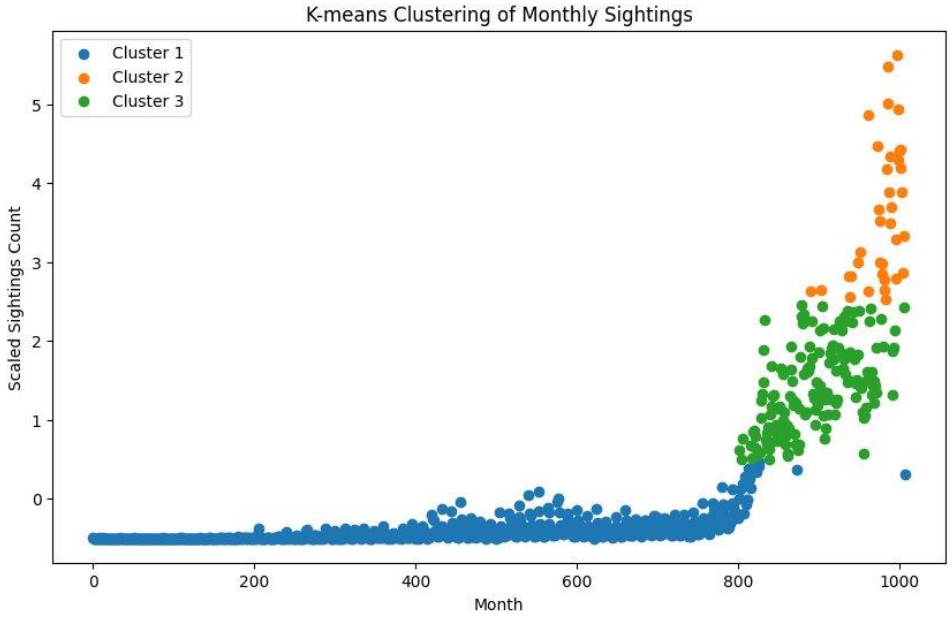
Appendix G – Time Series Decomposition of UFO Sightings



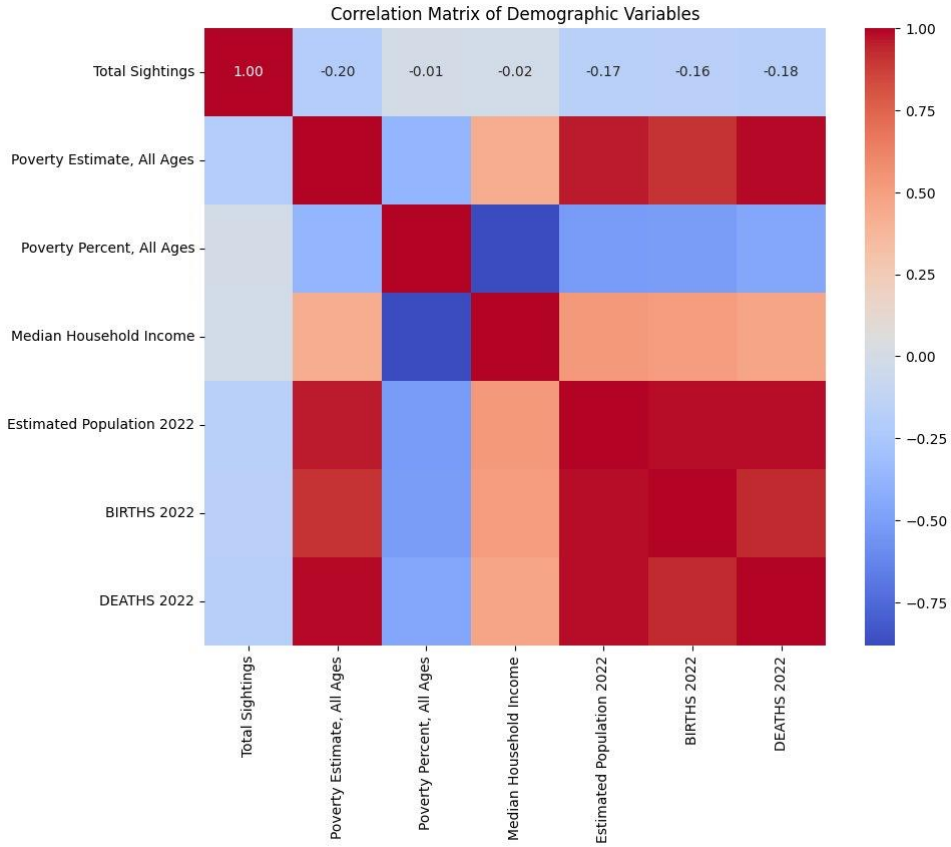
Appendix H – Elbow Plot to determine number of clusters



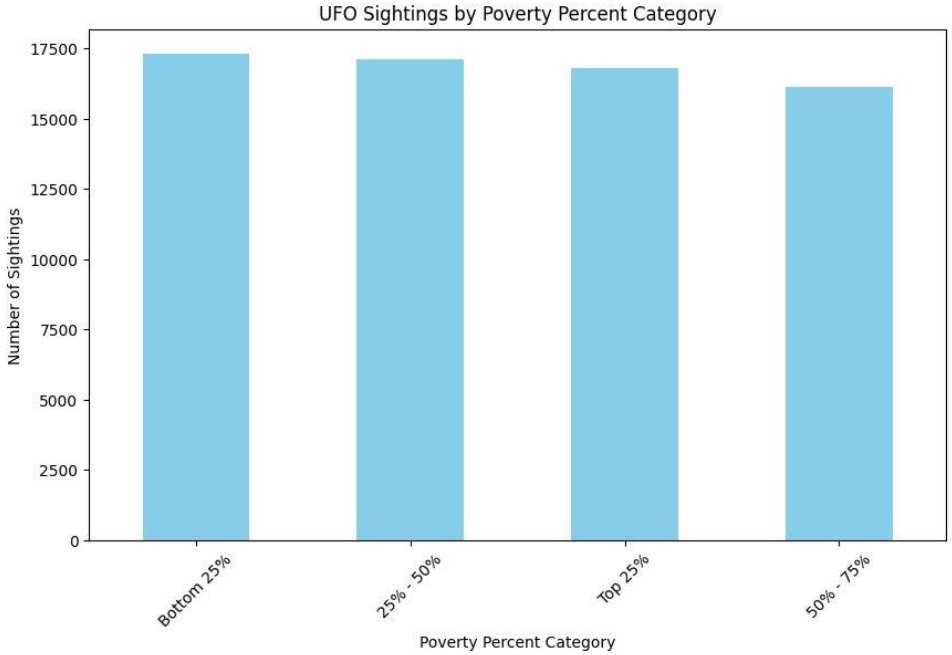
Appendix I – Plot of K-Means Clusters of Monthly Sightings



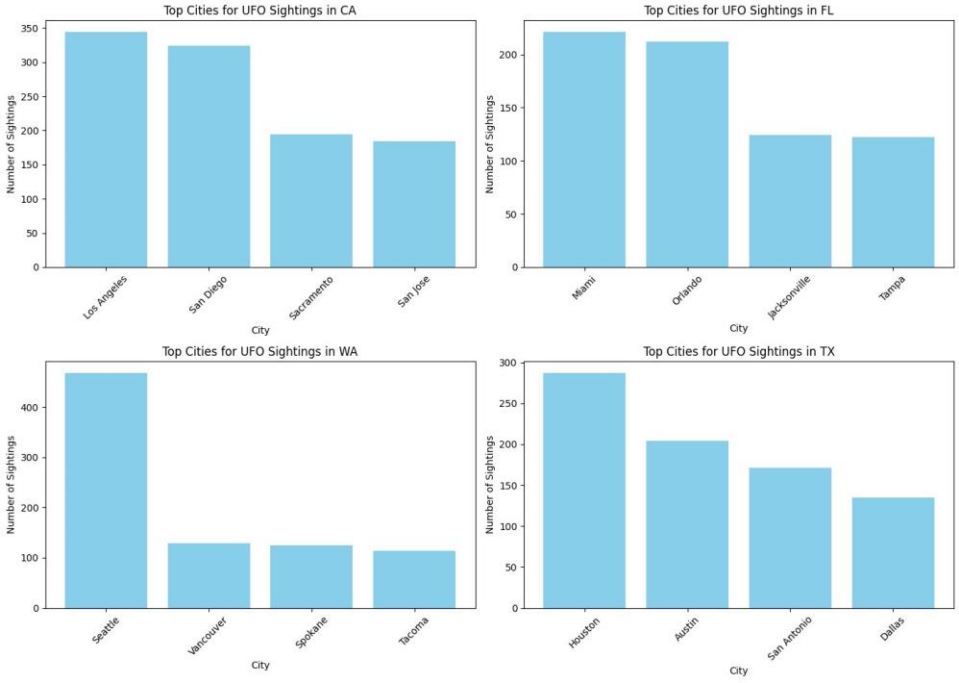
Appendix J – Correlation Heatmap of Demographic Variables and Total Sightings



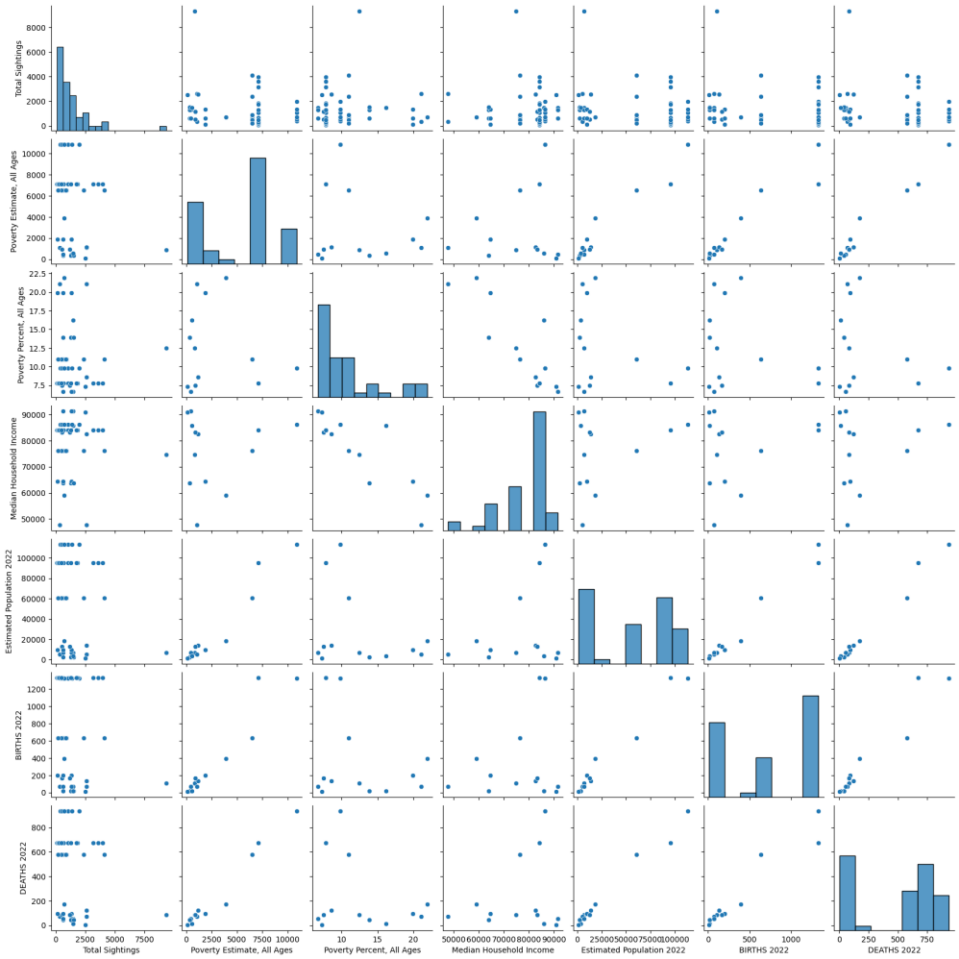
Appendix K – Bar plot of UFO Sightings by Poverty Percent Category



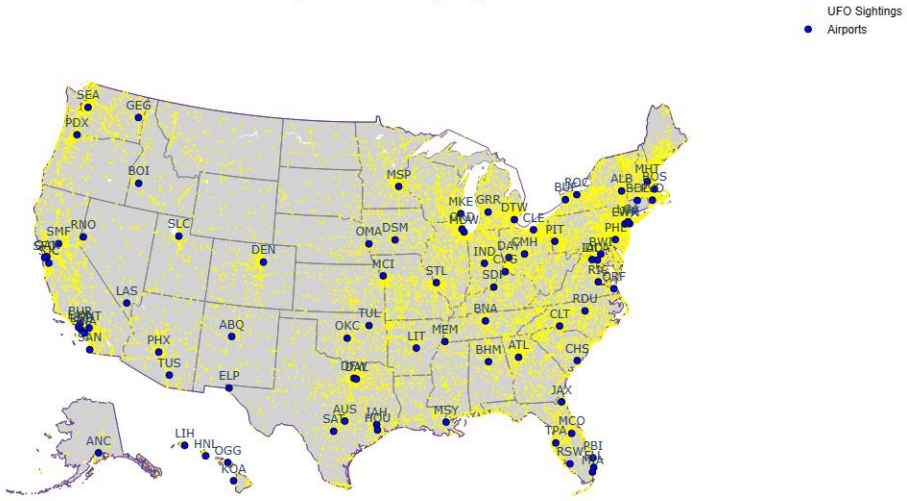
Appendix L – Top 4 Cities of Top 4 States by Number of Reported Sightings



Appendix M – Pair plot of Demographic Variables with Total Sightings



Appendix N - Map of UFO Sightings with Airports



Appendix O – Map of UFO Sightings with Airports and Reported Missing Persons