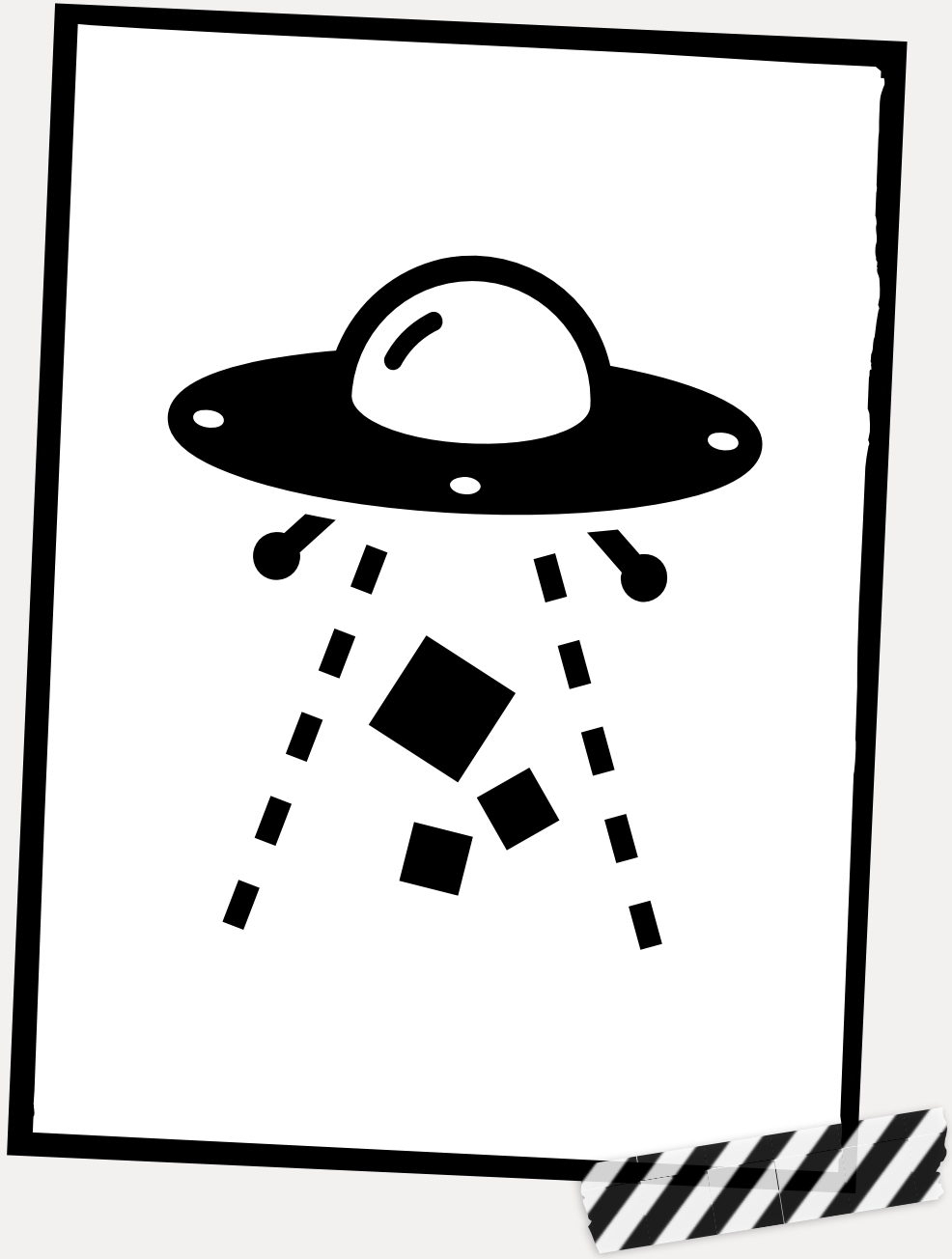
Data Encounters of the Third Kind: Unraveling the UFO Phenomenon

IST 718 Final Project

Zane Alderfer, Ben Heindl, Victoria Haley

Project Overview

- Objective:
  - Investigate patterns and trends in UFO sighting data.
  - Predict the likelihood and location of future UFO sightings.
  - Analyze correlations between UFO sightings and airports.
- Data Sources:
  - UFO Sighting Data from Kaggle.
  - Census Data for demographic information.
  - US Boundary Files for geographical boundaries.
  - Latitude/Longitude to FIPS Code conversion for location identification.
- Importance of the Project:
  - Understanding UFO sightings can provide insights into potential extraterrestrial activities or atmospheric phenomena.
  - Analyzing spatial and temporal patterns in UFO sightings can aid in public safety and awareness.
  - Examining the impact of proximity to major airports on reported UFO sightings to understand potential misidentifications and perceptual biases.

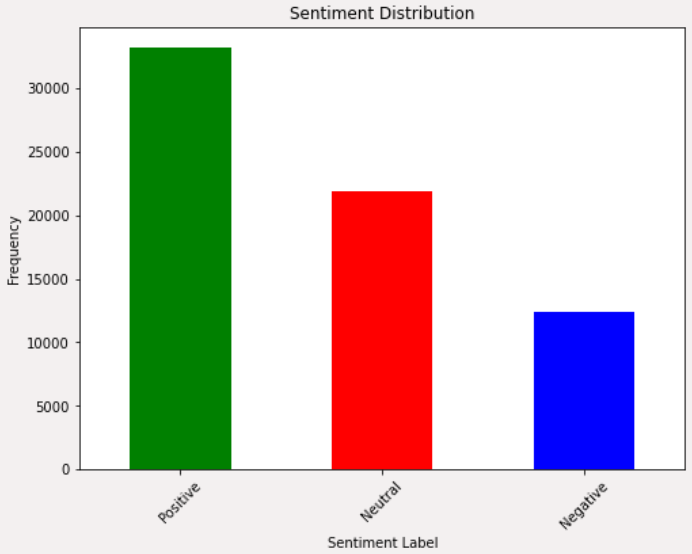


Data Description

- 67,401 rows x 18 columns
- Total sightings: 67,401
- Time range: June 1930 to May 2014
- Geographic coverage: 50 states, 14,875 cities in the US
- Most common shape: Light (14,320 sightings)
- Median household income: $77,790
- Poverty percentage: 12.41%
- Estimated population in 2022: ~835,205
- Births and deaths in 2022: 8,904 births, 7,351 deaths

Data Preparation

- Loaded the UFO sighting data from the CSV file.
- Cleaned the data by:
  - Converted 'latitude' and 'longitude' to numeric format, handling any errors with 'coerce'.
  - Converted 'country' to categorical data type.
  - Converted 'datetime' and 'date posted' columns to datetime format, handling any errors with 'coerce'.
  - Converted 'duration (seconds)' to numeric format, handling any errors with 'coerce'.
- Filtered the data for UFO sightings in the United States.
- Transformed additional columns:
  - Converted 'datetime' and 'date posted' columns to datetime format for further analysis.
  - Converted categorical columns ('State', 'country', 'shape', 'city', 'FIPS') to categorical data type.
  - Converted 'duration (seconds)' to integer data type for numerical analysis.

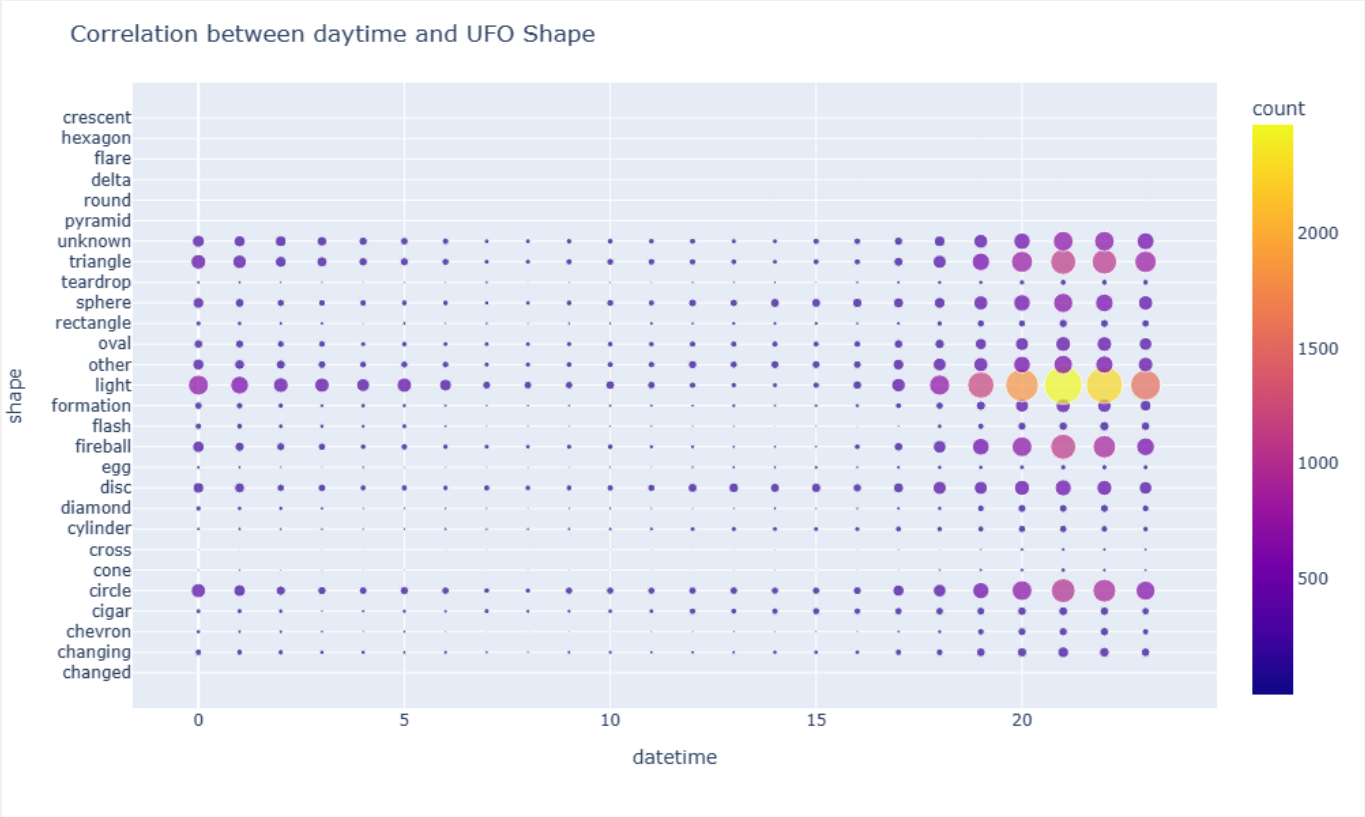


NLP Insights into UFO Sightings

- Conducted word cloud analysis on both UFO sighting comments and reported shapes, unveiling frequently occurring terms and shapes to explore prevalent themes and patterns in public perception and responses to UFO encounters.
- Sentiment analysis provided insights into the distribution of sentiments (positive, negative, neutral) within the comments, shedding light on public reactions.
- The analysis supplemented comprehension of the UFO phenomenon beyond numerical data alone, uncovering valuable insights from textual content.

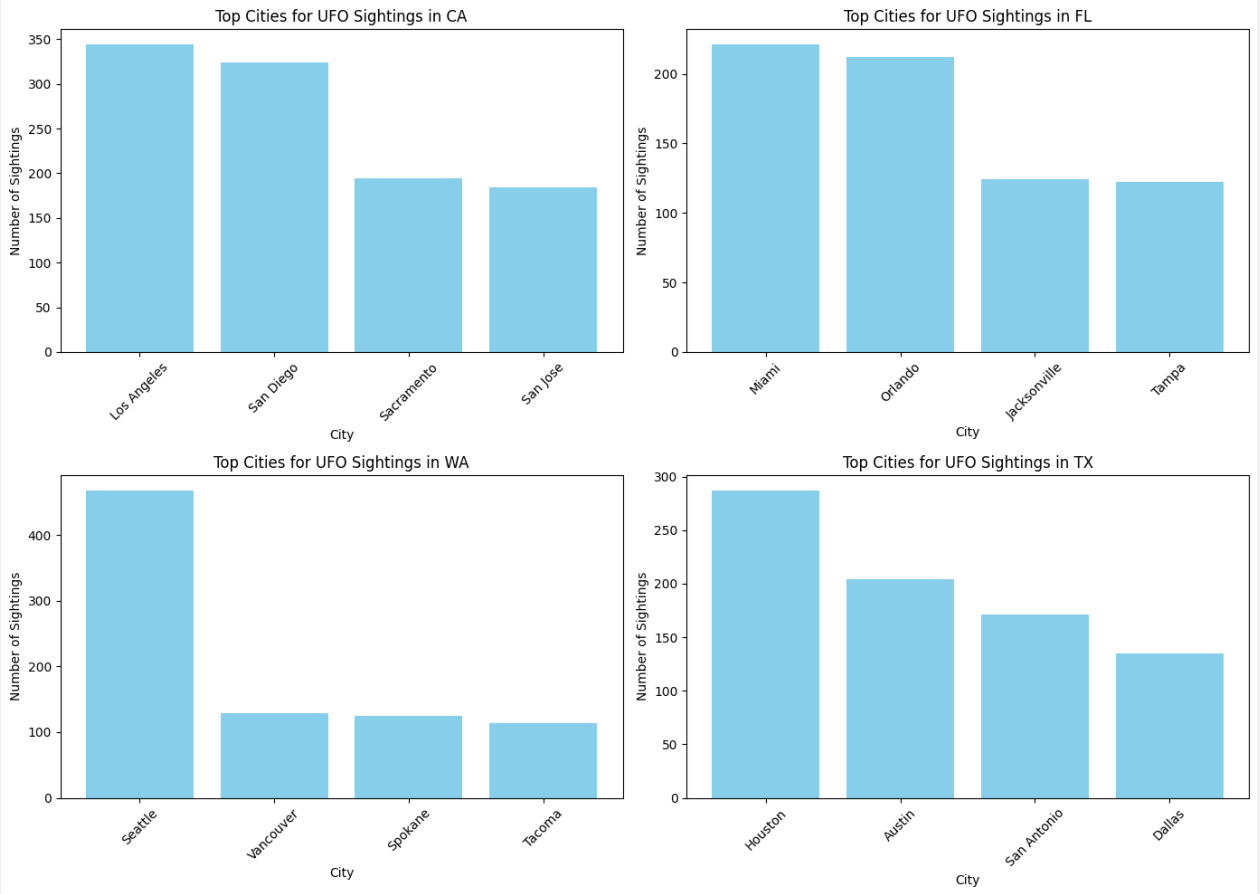
EDA: Correlation Analysis: Time of Day & UFO Shape

- Noticeable concentration of sightings for certain shapes like "light", "circle", and "triangle" across various times.
- "Light" seems to be the most frequently reported UFO shape, with the highest number of sightings occurring around the 20th hour of the day.
- The least reported shapes include "cone", "cross", and "changed", with very few sightings, mostly independent of the time of day.
- The graph suggests a potential pattern in UFO shape reports in relation to the time of day, particularly for the "light" shape.



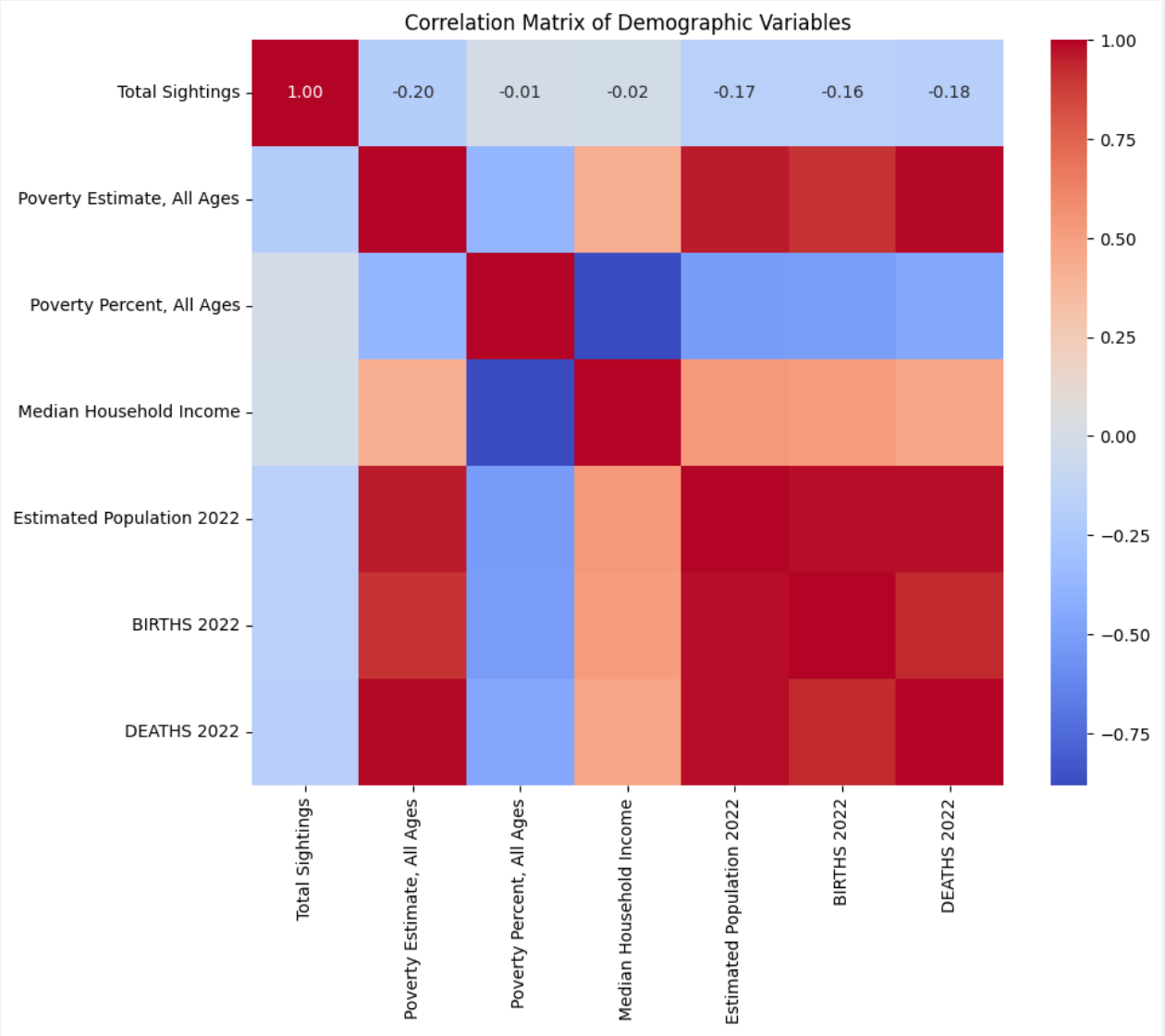
EDA: UFO Sightings Analysis Highlights

- Year with the highest number of sightings: 2012
  - 6466 sightings
- State with the highest number of sightings: California
- Most common City: Seattle, WA
  - Number of Sightings: 469



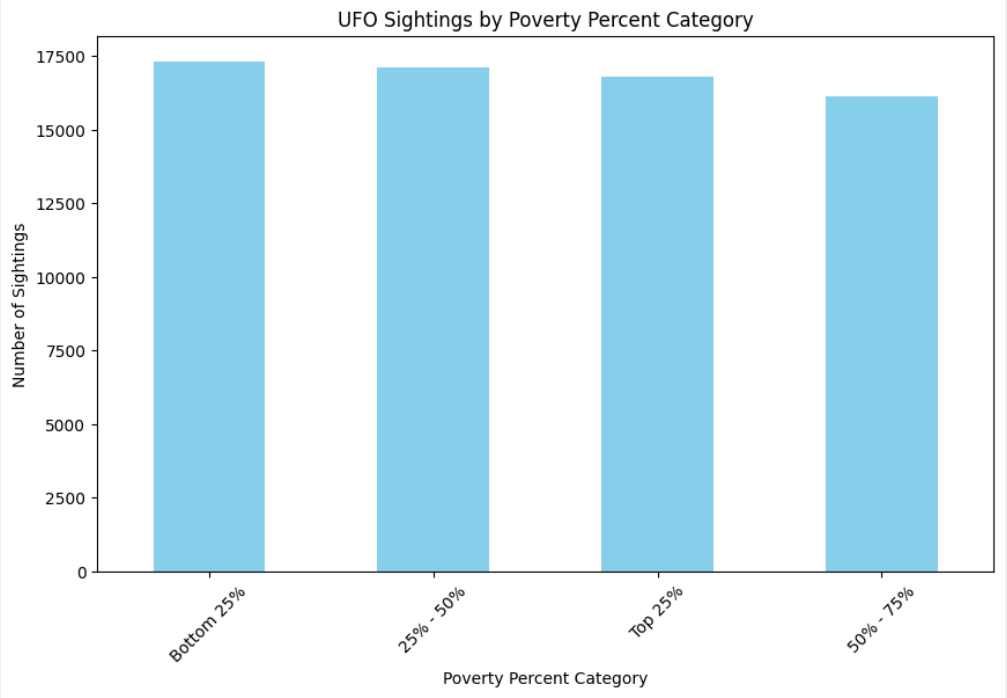
EDA: Demographic Analysis

- Poverty Estimate, All Ages:
  + P-Value: 0.170
- Poverty Percent, All Ages:
  + P-Value: 0.930
- Median Household Income:
  + P-Value: 0.901
- Estimated Population 2022:
  + P-Value: 0.226
- BIRTHS 2022:
  + P-Value: 0.253
- DEATHS 2022:
  + P-Value: 0.216



EDA: Statistical Analysis of UFO Sightings by Poverty Percentage

- **Kruskal-Wallis Test Results:**
  - H Statistic: 47.79
  - p-value: 2.36e-10
- **Insights:**
  - Statistically significant differences in occurrences of UFO sightings across poverty percent categories.
- **Observations:**
  - Highest occurrence in bottom 25% poverty category.
  - Followed by 25-50% category, then top 25%, and 50-75% category.
  - However, only slight decrease in sightings from bottom 25% to 50-75% category.
- **Conclusion:**
  - Suggests potential correlation between poverty percentage categories and UFO sighting frequency.
- **Implications:**
  - Contributes to understanding socioeconomic influences on reported phenomena.
  - Highlights need for further investigation into underlying factors.



EDA: Comparison of UFO Sightings in States with Highest and Lowest Numbers

STATE WITH THE HIGHEST NUMBER OF SIGHTINGS: CALIFORNIA

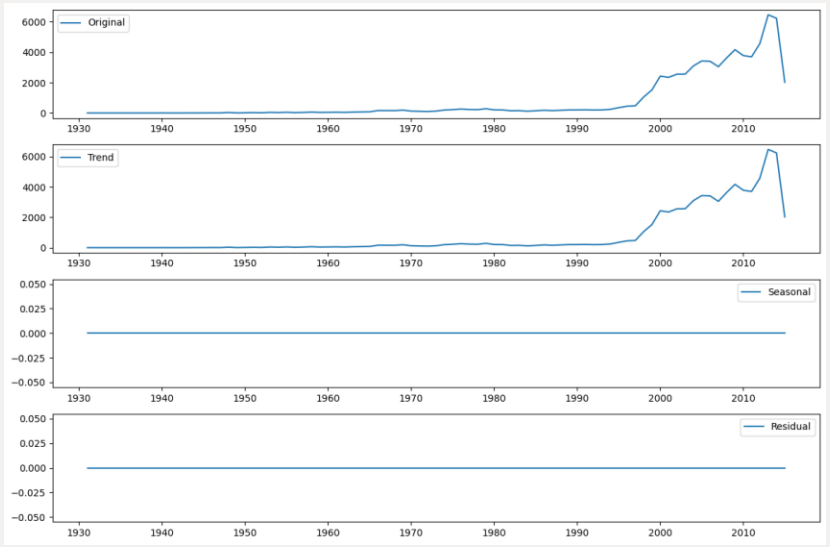
- Number of sightings: 9310
- Average Time Between Sightings (Days): 2.65
- Mean census values:
  + Poverty Percent: 12.21%
  + Median Household Income: $92,595.89
  + Estimated Population 2022: 3,203,719
  + BIRTHS 2022: 33,509
  + DEATHS 2022: 25,897

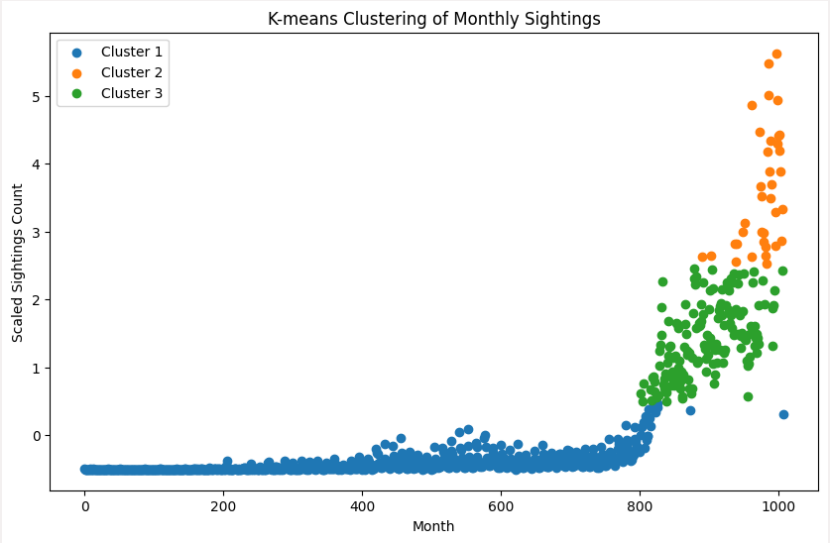
STATE WITH THE LOWEST NUMBER OF SIGHTINGS: NORTH DAKOTA

- Number of sightings: 131
- Average Time Between Sightings (Days): 222.5
- Mean census values:
  + Poverty Percent: 11.05%
  + Median Household Income: $73,723.53
  + Estimated Population 2022: 56,762
  + BIRTHS 2022: 719
  + DEATHS 2022: 521

When is the next UFO sighting likely to occur?

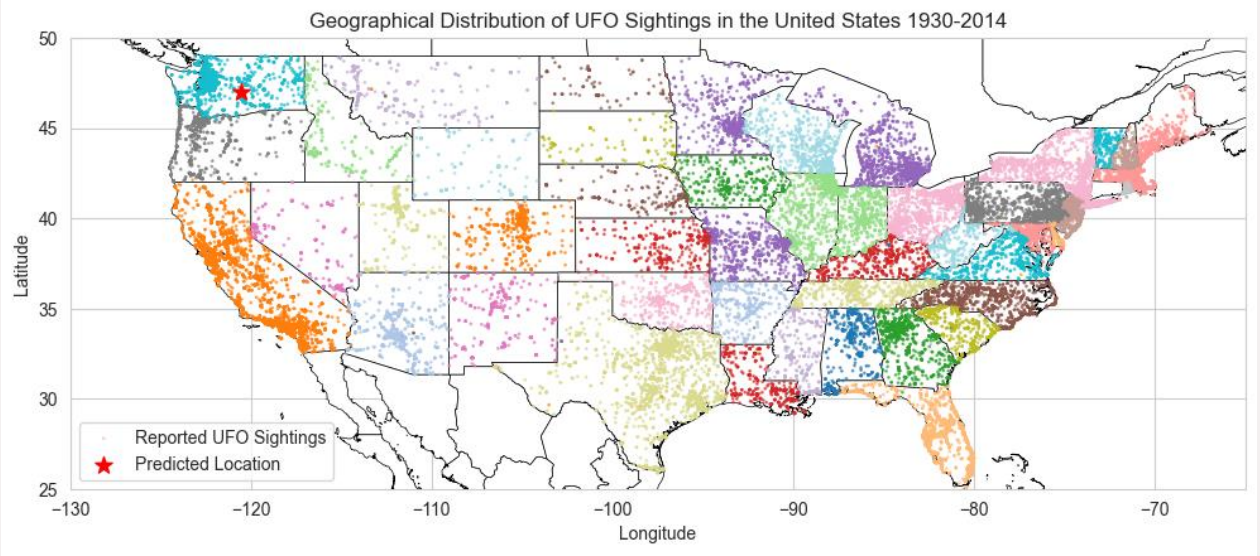
- Data Analysis:
  - Analyzed historical UFO sighting data.
  - Highest number of sightings in a month: 997 in July 2013
  - Average amount of sightings in a month: 67
- Model Development:
  - Time series decomposition to understand trends, seasonality, and residuals.
  - K-means clustering to identify patterns in monthly sightings.
- Predictive Analysis:
  - Identified the cluster with the highest likelihood of the next sighting.
  - Calculated the average time between sightings in the selected cluster.
- Prediction Results:
  - Predicted next sighting date: April 28, 2024.
- Implications:
  - Provides insight into potential patterns of UFO sightings.
  - Offers a basis for preparedness and monitoring.





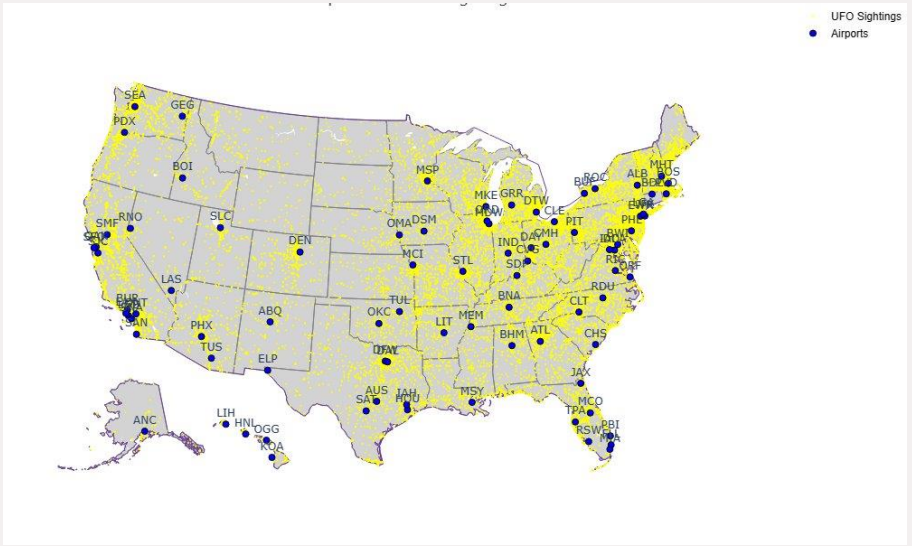
Where is the next UFO sighting likely to occur?

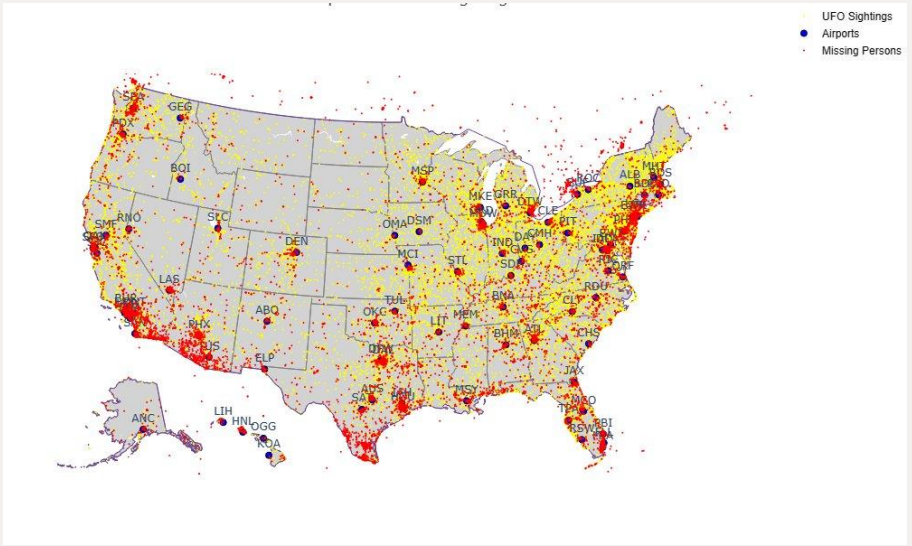
- **Data Analysis:**
  - Grouped data by datetime, latitude, and longitude to train location prediction models.
- **Location Prediction Model:**
  - Trained linear regression models for latitude and longitude separately.
  - Predicted latitude and longitude for the next UFO sighting.
- **Predicted Location:**
  - **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: 0.0001644205119201677 4
  - Support Vector Regression MSE: Latitude: 0.03419574237205408, Longitude: 0.06464941401996009



Spatial Analysis: UFO Sightings, Major US Airports, and Missing Persons Locations

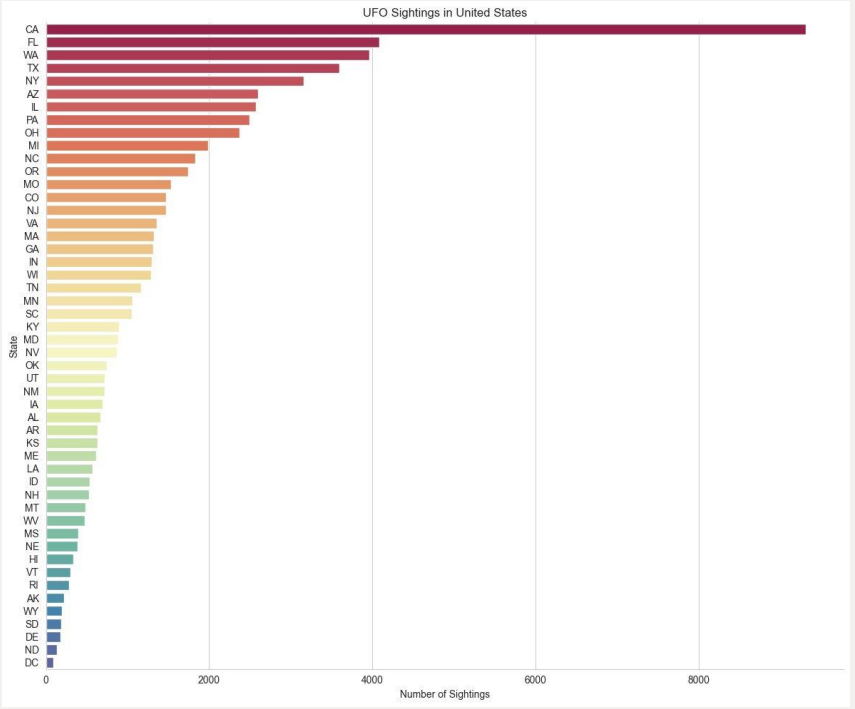
- Analysis of spatial relationship between UFO sightings, major US airports, and missing persons locations
- UFO sightings 23 times more likely within 10 miles of major airports
- Signifies a 2226% higher likelihood of sightings near airports compared to overall regions
- Methodological approach mitigated biases from misidentifications or heightened awareness near airports
- Emphasizes importance of considering airport proximity in UFO sighting analysis, while acknowledging the presence of missing persons locations in similar areas, which may also be influenced by high population density

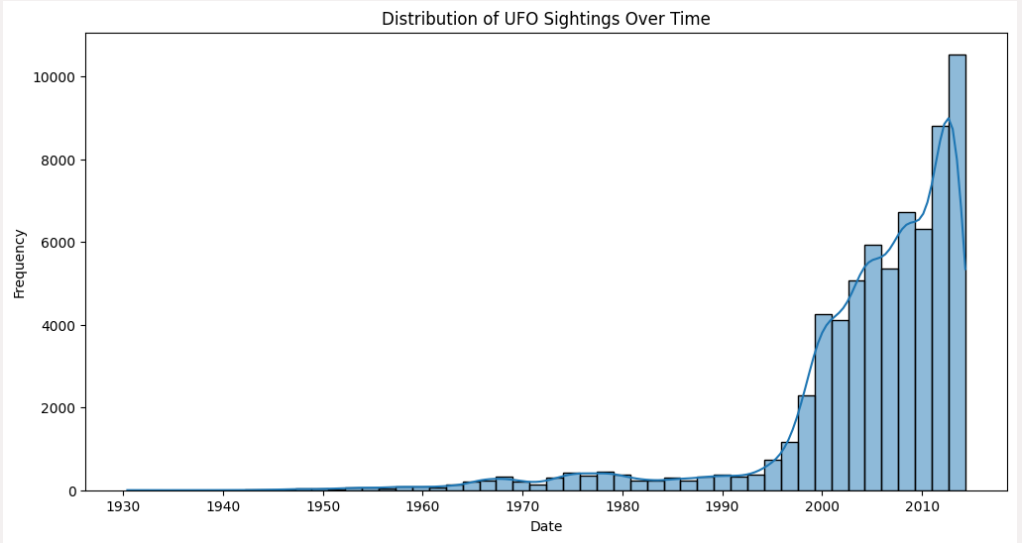




Key Findings and Insights

- Spatial Analysis:
  - + Identified geographical patterns and hotspots of UFO activity.
  - + Revealed significant association between UFO sightings and proximity to major airports.
- Temporal Analysis:
  - + Uncovered fluctuations in UFO sightings over time, aligning with historical trends.
  - + Identified temporal patterns in sightings characteristics, aiding in predictive modeling.
- Shape Analysis:
  - + Quantified frequency of various UFO shapes reported, highlighting common appearances.
- Text Analysis:
  - + Uncovered mostly positive sentiments in comments accompanying UFO sightings.
- Demographic Analysis:
  - + Observed weak correlations between socioeconomic variables and UFO sightings.
  - + Acknowledged limitations due to discrepancies in census data.





Implications and Challenges

- Implications and Significance:
  - Understanding of UFO phenomena influenced by various factors: temporal trends, demographic correlations, and spatial proximity to major airports.
  - Insights contribute to public safety, awareness, and informed discourse on UFO phenomena.
- Limitations and Challenges:
  - Data quality concerns due to discrepancies in census data.
  - Time constraints limited thorough investigation into data reliability.
  - Potential biases in reporting and misidentifications of sightings.

Recommendation for Future Research and Improvements

- Continued research to understand underlying causes and factors contributing to UFO encounters.
- Enhanced data quality through thorough validation and collaboration with relevant authorities.
- Public awareness campaigns to promote informed dialogue and reduce misidentifications.
- Integration of multidisciplinary approaches to explore diverse perspectives and methodologies.
- Standardization of reporting practices and documentation to improve reliability of UFO sighting databases.

Conclusion

- Recap of Main Points:
  - Comprehensive analysis of UFO sighting data.
  - Insights into temporal, spatial, and demographic trends.
  - Implications of proximity to major airports.
  - Challenges in data quality and reliability.
- Importance of the Project:
  - Contributes to understanding UFO phenomena.
  - Enhances public safety and awareness.
  - Advances knowledge in the field of ufology.

Thank you!

- 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
  - at/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