Zane Alderfer

Dr. John Stinnent

IST 782

May 16, 2024

MS of Data Science Portfolio Draft

Data Encounters of the Third Kind: Unraveling the UFO Phenomenon

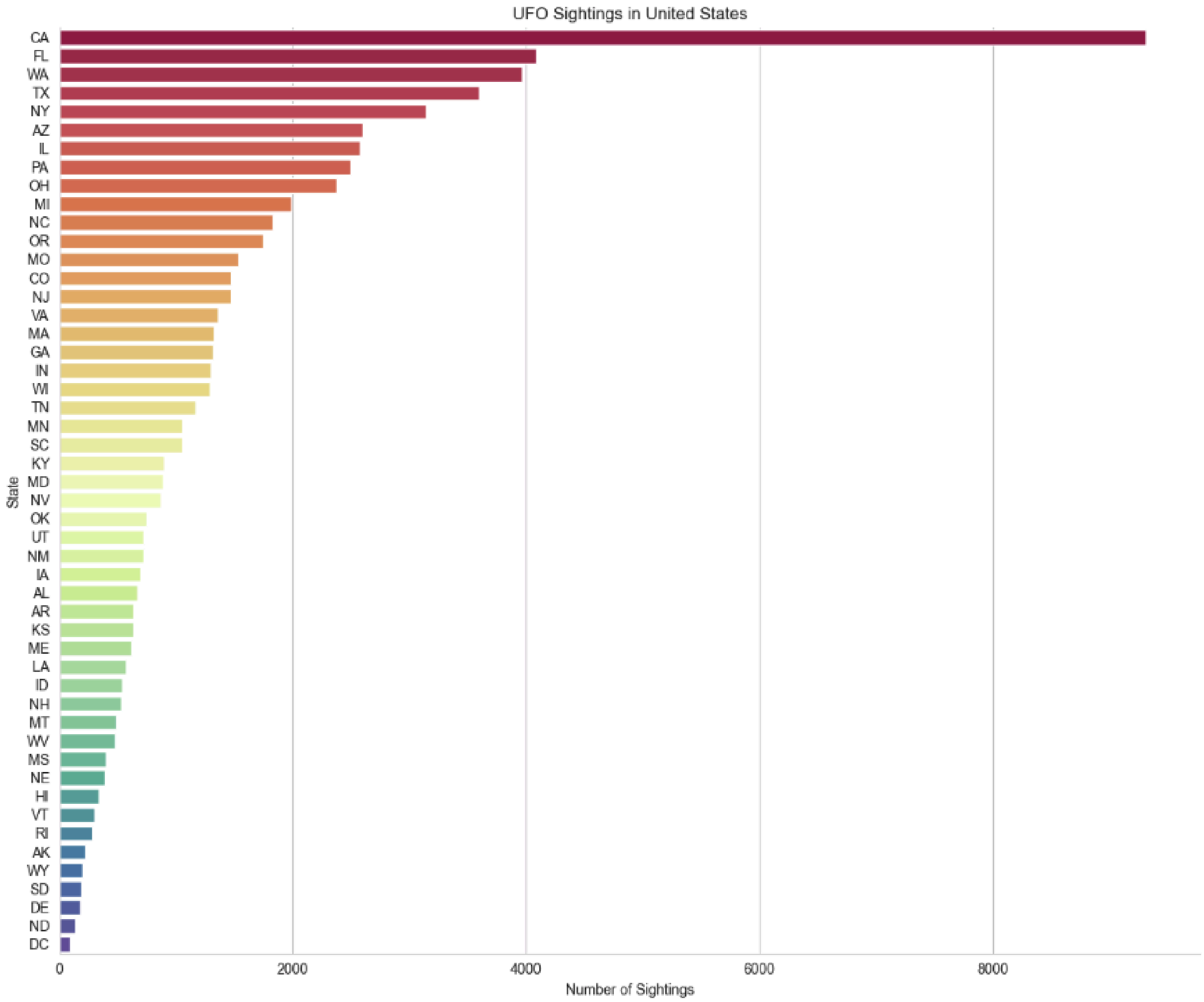
In the report "Data Encounters of the Third Kind: Unraveling the UFO Phenomenon," Zane Alderfer, Ben Heindl, and Victoria Haley present a thorough analysis of UFO sighting data. The primary aim of their study was to investigate patterns and trends that might provide insights into potential extraterrestrial activities or atmospheric phenomena. By employing various analytical techniques, the researchers sought to achieve a deeper understanding of UFO sightings and their possible correlations with different factors.

Data Preparation and Integrity

The initial phase of the project involved meticulous data cleaning and preparation to ensure the integrity and reliability of the dataset. This process included converting data types, handling missing values, and focusing the analysis on UFO sightings within the United States. These steps were essential to facilitate accurate analysis and interpretation of the data.

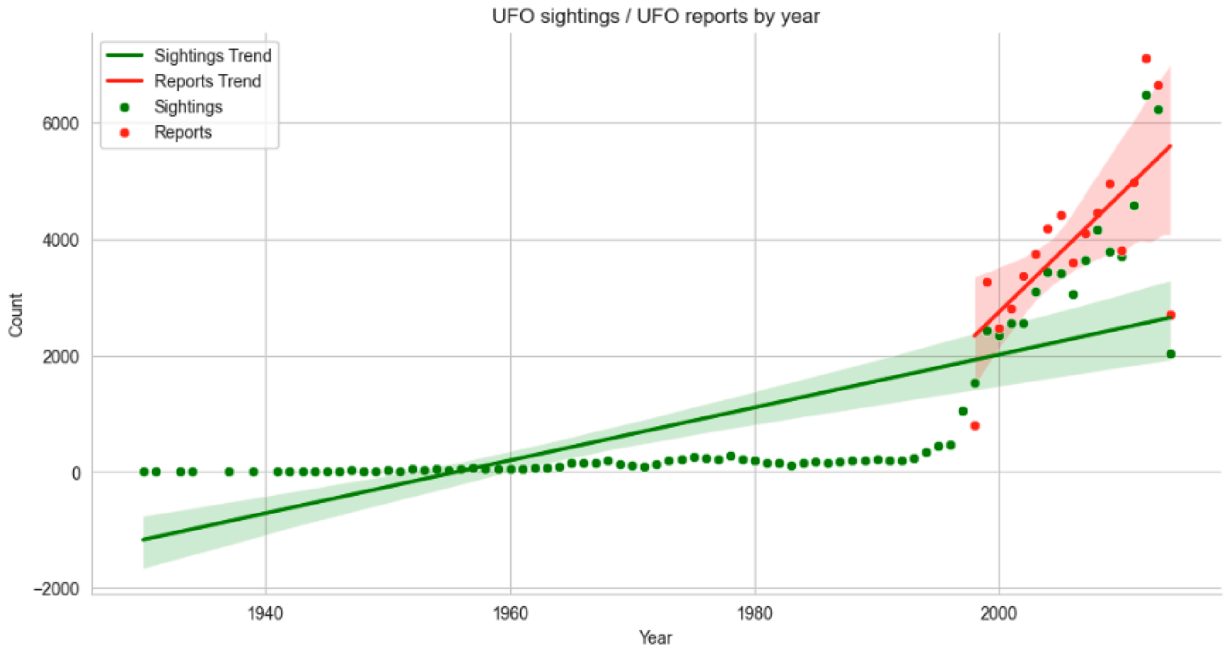
Spatial Analysis

The spatial analysis aimed to identify geographical patterns and hotspots of UFO activity across the United States. The analysis revealed that states such as California, Florida, Texas, and Washington reported the highest number of sightings. This suggests these states might experience higher UFO activity or have more active reporting systems compared to others. Furthermore, the study found that UFO sightings were more common near major airports, indicating that some sightings might be misidentified aircraft or influenced by airport activities. This spatial proximity emphasizes the need to consider environmental and contextual factors when analyzing UFO sightings. One of the key graphs in the study is a bar plot showing the total number of UFO sightings reported in each state.



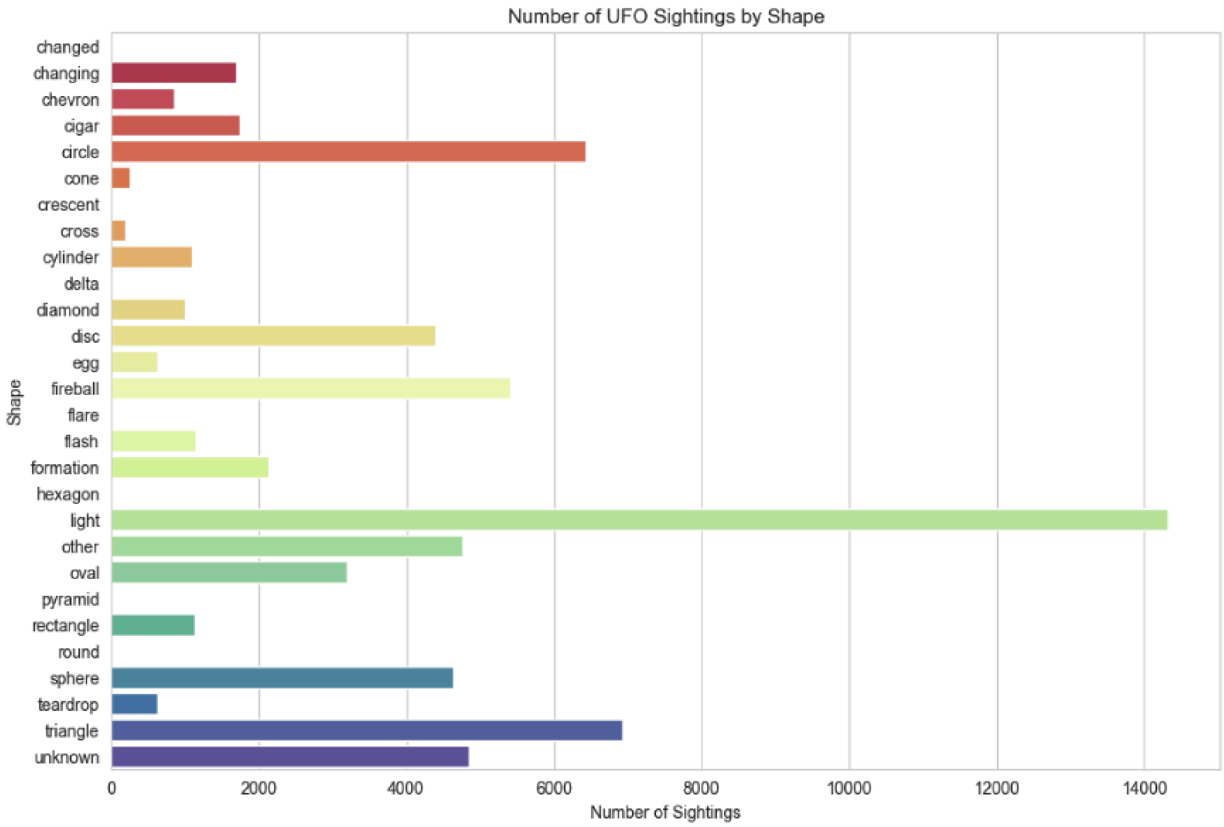
Temporal Analysis

Temporal analysis uncovered significant fluctuations in UFO sightings over time. The data showed notable peaks around the years 1980, 1994, and 2009-2010, with the highest number of sightings occurring in 2012. These peaks likely correlate with periods of heightened media attention and public interest in UFO phenomena. Understanding these temporal trends is crucial for recognizing periods of increased UFO activity and correlating them with historical and cultural contexts. The temporal distribution of sightings is illustrated through a line graph that depicts the number of sightings over time.



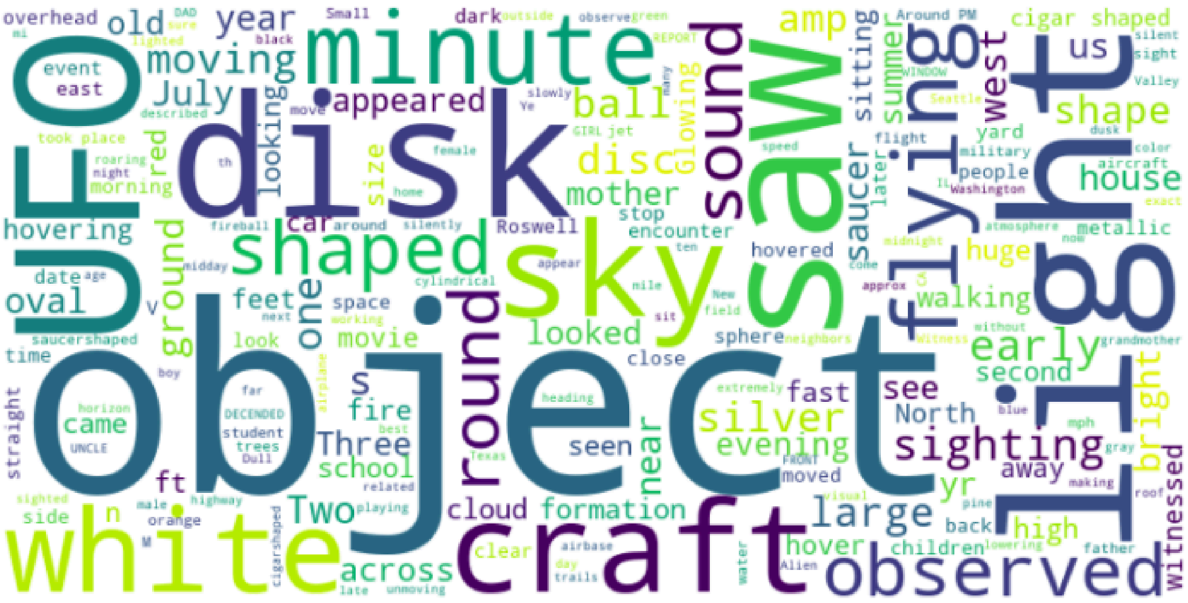
Shape Analysis

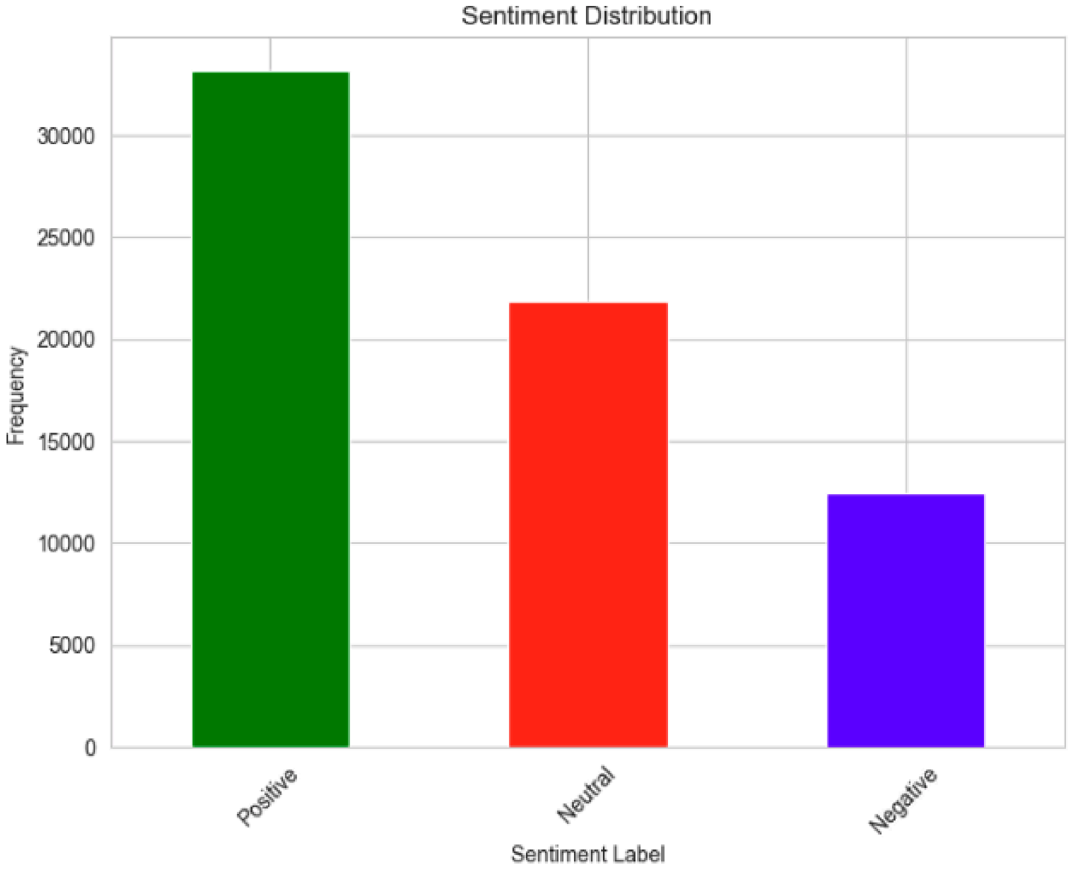
An analysis of the reported shapes of UFOs provided additional insights. The most frequently observed shape was 'Light,' followed by 'Circle' and 'Triangle.' Identifying common shapes is instrumental in distinguishing between potential misidentifications and genuine sightings. This shape analysis contributes to a more nuanced understanding of the characteristics of UFO sightings as reported by witnesses. Another insightful graph categorizes UFO sightings based on the shape reported by witnesses.



Text and Sentiment Analysis

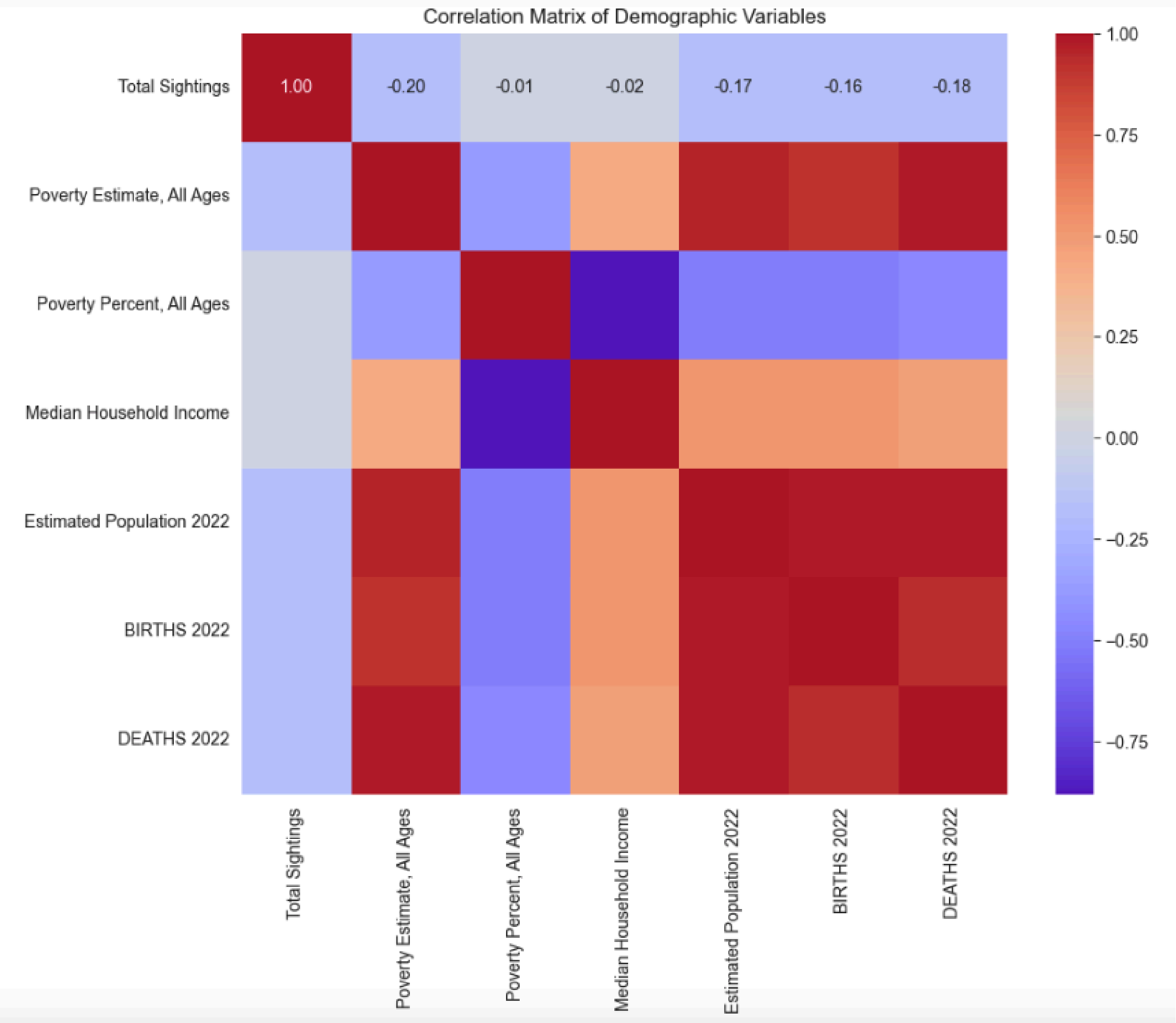
Text analysis of comments associated with UFO sightings revealed mostly positive sentiments. Common terms used in the comments included 'light,' 'object,' and 'saw.' This sentiment analysis indicates that individuals reporting UFO sightings generally have a positive or neutral perception of their experiences. Understanding the language and sentiments expressed in these reports provides valuable context for interpreting public attitudes toward UFO phenomena. The sentiment analysis of comments accompanying UFO sightings is depicted through a word cloud and a sentiment distribution chart.





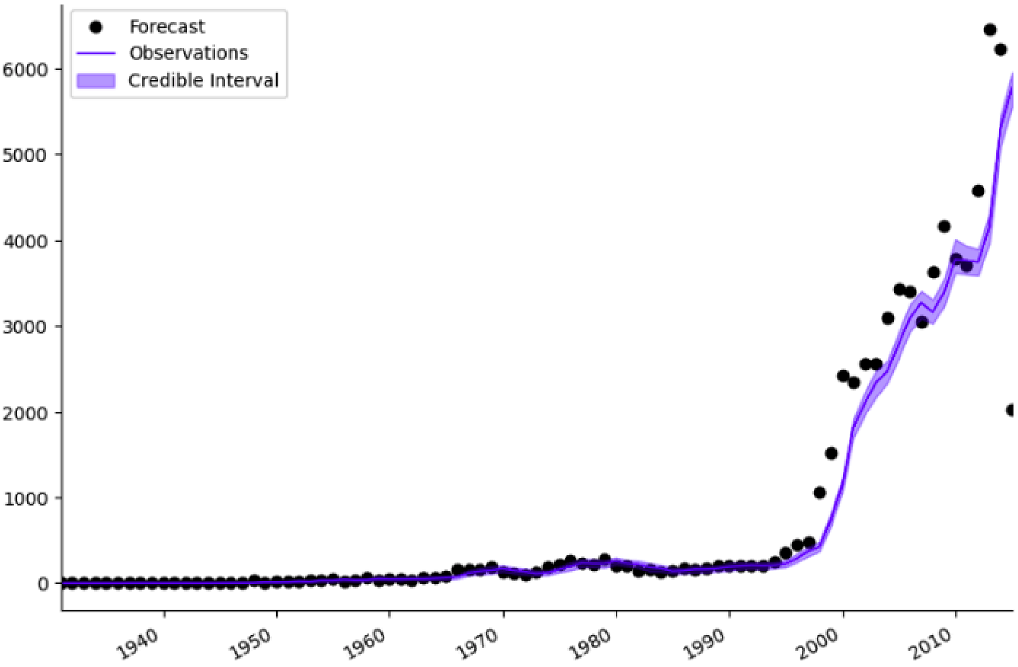
Demographic Analysis

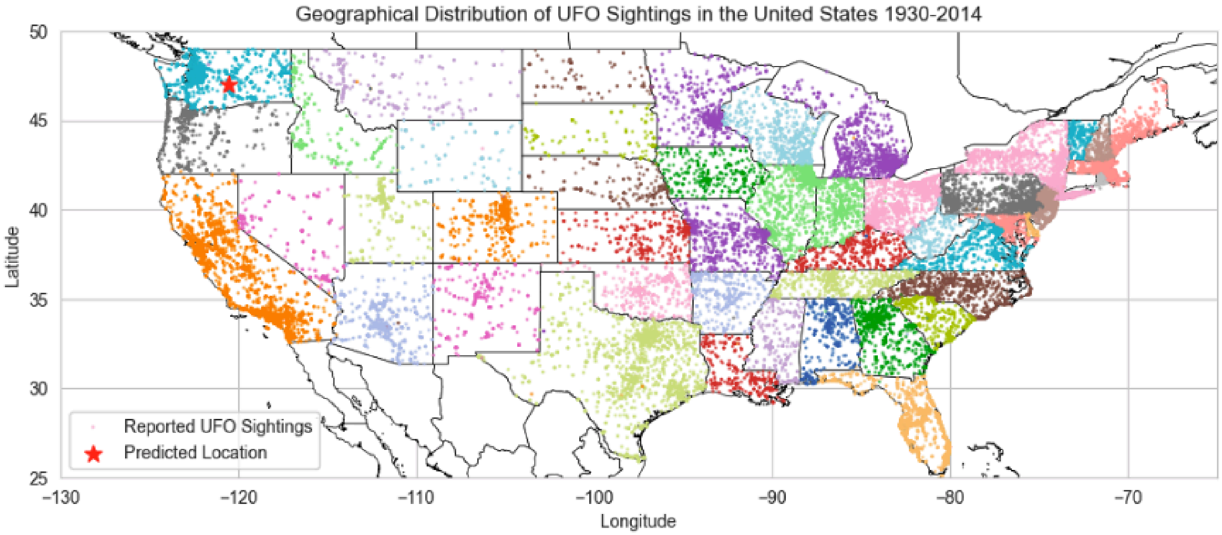
The demographic analysis explored correlations between UFO sightings and various socioeconomic variables. Surprisingly, the analysis found weak correlations between the number of sightings and factors such as median household income and poverty percentage. This suggests that traditional demographic indicators may not significantly influence UFO sighting frequency. The findings highlight the complexity of the phenomenon and the need to consider a broader range of factors. The heatmap of demographic variables and sightings explores the correlation between various socioeconomic factors and the number of UFO sightings. A scatter plot visualizing the relationship between the number of sightings and various demographic variables underscores the weak correlations identified in the heatmap.



Predictive Analysis

The study also included predictive modeling to forecast the likelihood and location of future UFO sightings. Based on historical data and identified patterns, the researchers predicted the next likely sighting to occur on April 28, 2024, in Ellensburg, Washington. This prediction was derived from time series decomposition and clustering techniques, providing valuable insights into potential future trends and helping to inform proactive public safety measures. The time series decomposition of sightings provides a detailed view of the trend, seasonality, and residuals in UFO sighting data over time.





Spatial Proximity Analysis

The spatial proximity analysis revealed that UFO sightings were approximately 23 times more likely to occur within 10 miles of a major airport. This finding underscores the importance of considering the influence of aviation infrastructure on the perception and reporting of UFOs. By addressing potential biases from misidentifications or heightened awareness due to airport activities, the analysis enhances our understanding of UFO sighting phenomena. A particularly significant finding is illustrated through a map plotting UFO sightings in relation to major US airports.

Recommendations

Based on their findings, the researchers made several recommendations to further enhance understanding, research, and public awareness of UFO phenomena. They emphasized the need for continued research and analysis, suggesting the exploration of additional datasets and advanced statistical methods. They also highlighted the importance of improved data quality, advocating for thorough validation and verification processes to address inconsistencies in demographic data.

The researchers recommended increasing public awareness and education to promote informed and rational discourse about UFO sightings. They called for the integration of multidisciplinary approaches, encouraging collaboration among scientists, researchers, policymakers, and the public to explore various perspectives and methodologies. Additionally, they suggested enhancing reporting and documentation practices to ensure accurate and reliable UFO sighting databases.

To mitigate misidentifications, the researchers advised implementing measures to educate the public about common misidentifications of aircraft and celestial objects. Promoting critical thinking and skepticism in evaluating UFO reports could help reduce the number of false positives and enhance the credibility of UFO sighting data.

Conclusion

The analysis conducted in this study provides a comprehensive overview of UFO sighting data, highlighting critical spatial and temporal trends, common shapes, sentiment analysis, and demographic correlations. These graphs collectively enhance our understanding of the multifaceted nature of UFO phenomena and guide future research and public awareness efforts. By identifying patterns and trends, this study contributes to the broader discourse on UFO sightings and lays the groundwork for further exploration and investigation into this intriguing field.

Machine Learning in Energy Data Analysis: A Comprehensive Study

In their IST 707 Final Project, Zane Alderfer and Chris Snyder explore the application of machine learning algorithms to analyze and predict patterns in energy consumption and generation. The study utilizes a dataset from Kaggle titled "Hourly Energy Demand Generation and Weather," containing 29 variables and over 35,000 rows. The primary goal is to demonstrate how machine learning models can be applied to energy data to enhance understanding and efficiency in energy management.

Introduction

Machine learning algorithms offer powerful tools for analyzing relationships within datasets and making predictions. These algorithms fall into two main categories: supervised learning models, such as association-rule mining and clustering techniques, and unsupervised models, including decision trees, random forests, Naive Bayes, and k-nearest neighbors (kNN). The ability of these algorithms to process extensive datasets rapidly and accurately has significant implications for various industries, including the energy sector.

In the context of energy data, machine learning algorithms can organize and predict patterns in energy consumption and output, which is crucial for understanding current usage and making future projections. For instance, support vector machines (SVMs) can forecast energy demand, aiding utility companies in efficiently managing power generation and distribution. These predictions are particularly valuable for optimizing the use of renewable energy sources, ensuring cleaner and more cost-effective energy solutions.

Data and Preprocessing

The dataset used in this study spans from 2015 to 2018 and includes metrics for various energy sources, weather forecasts, and energy outputs recorded in megawatts (MW). Initial preprocessing steps involved downloading the data, handling null values, and creating subsets for analysis. The data was cleaned using the na.omit() function to remove rows with null values. Subsets were created to focus on specific aspects, such as combining weather data with energy outputs from Valencia, and aggregating energy sources based on emission levels.

Analysis and Models

The research process involved several steps and the application of various machine learning models:

1. Exploratory Data Analysis (EDA):
    a. Plots were created to visualize energy output forecasts versus actual outputs, showing a clear linear relationship with high accuracy in Spain's predictive models.
    b. Figures illustrating energy outputs from high-emission and low-emission sources indicated cyclical patterns with fluctuations in demand throughout the year.
2. Decision Trees:

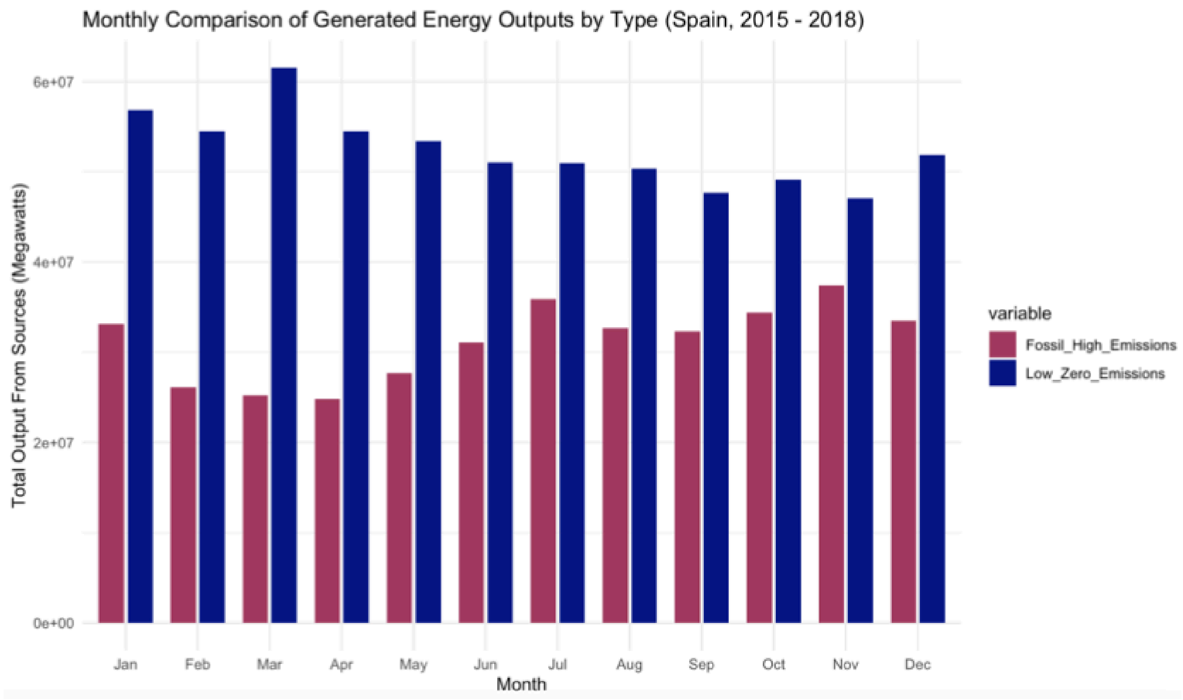
   a. Decision trees were constructed to analyze the factors influencing high and low emission energy outputs.
   b. The key splitting variables for high emissions were identified as hard coal and fossil gas, while wind and hydropower were significant for low emissions.

3. Association Rules:
   a. Association rules were generated to explore the impact of weather on energy output.
   b. Rules indicated that high temperatures and solar generation were correlated, and cloudy days were associated with higher overall energy output.
4. Naive Bayes Prediction:
   a. The Naive Bayes model was used to predict the month based on energy output and weather indicators.
   b. January and July were the most frequently predicted months, possibly due to similarities in weather patterns with adjacent months.

Results and Interpretations

The study produced several key findings:

1. Energy Output Predictions:
   a. There was a high degree of accuracy in forecasting energy outputs, with minor discrepancies due to overestimation.
   b. Both high-emission and low-emission energy sources exhibited cyclical patterns in output, influenced by seasonal variations.

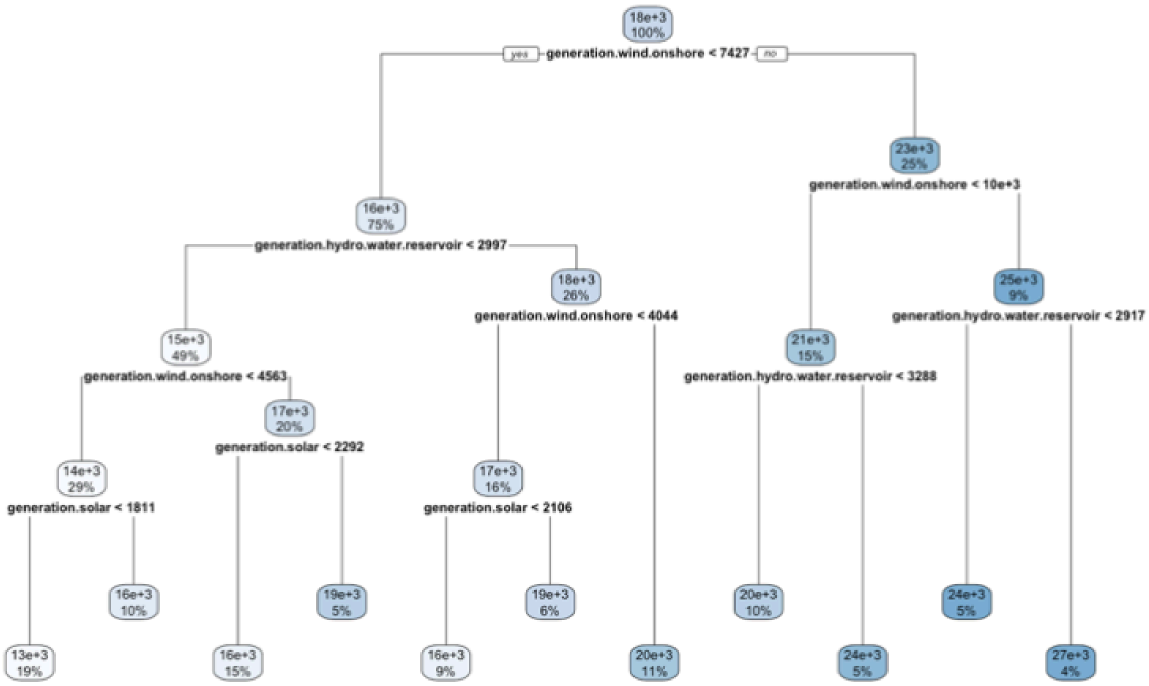


2. Decision Tree Insights:

    a. The primary contributors to high-emission outputs were hard coal and fossil gas, while wind and hydropower were crucial for low-emission outputs.

    b. The models highlighted the seasonal dependencies of different energy sources, with renewable sources peaking during wetter months.

Decision Tree 2: Predicting Outputs from Low/Zero Emission Sources

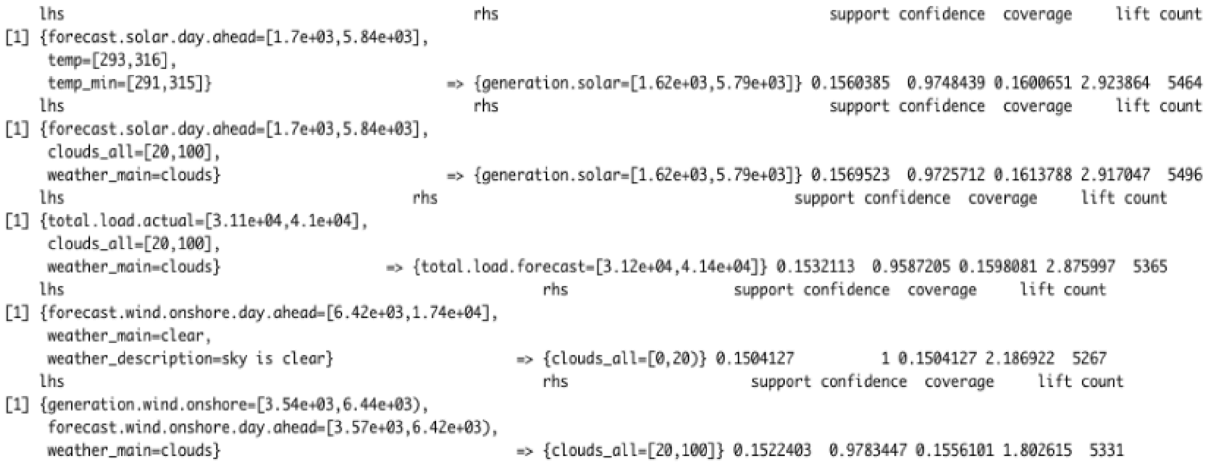

18e+3
100%
yes
generation.wind.onshore < 7427
no
16e+3
75%
generation.hydro.water.reservoir < 2997
23e+3
25%
generation.wind.onshore < 10e+3
15e+3
49%
generation.wind.onshore < 4563
18e+3
26%
generation.wind.onshore < 4044
21e+3
15%
generation.hydro.water.reservoir < 3288
25e+3
9%
generation.hydro.water.reservoir < 2917
14e+3
29%
generation.solar < 1811
17e+3
20%
generation.solar < 2292
17e+3
16%
generation.solar < 2106
13e+3
19%
16e+3
10%
16e+3
15%
19e+3
5%
16e+3
9%
19e+3
6%
20e+3
11%
20e+3
10%
24e+3
5%
24e+3
5%
27e+3
4%

3. Association Rules:

    a. Weather significantly impacted energy output, with clear correlations between temperature, solar generation, and wind energy output.

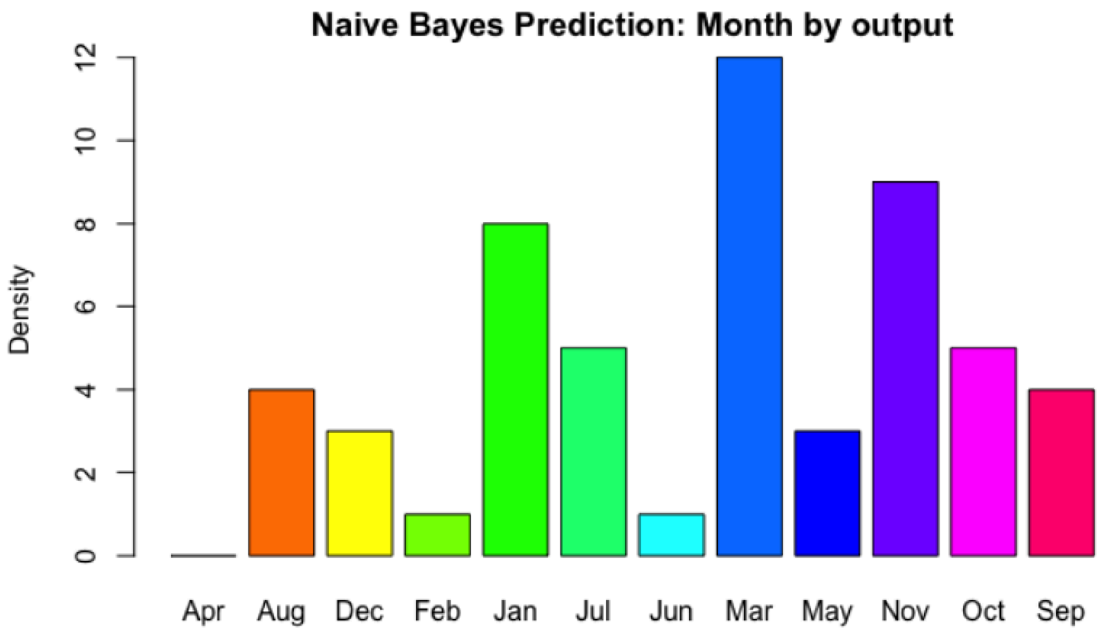
    b. Cloudy days tended to drive higher energy output overall.

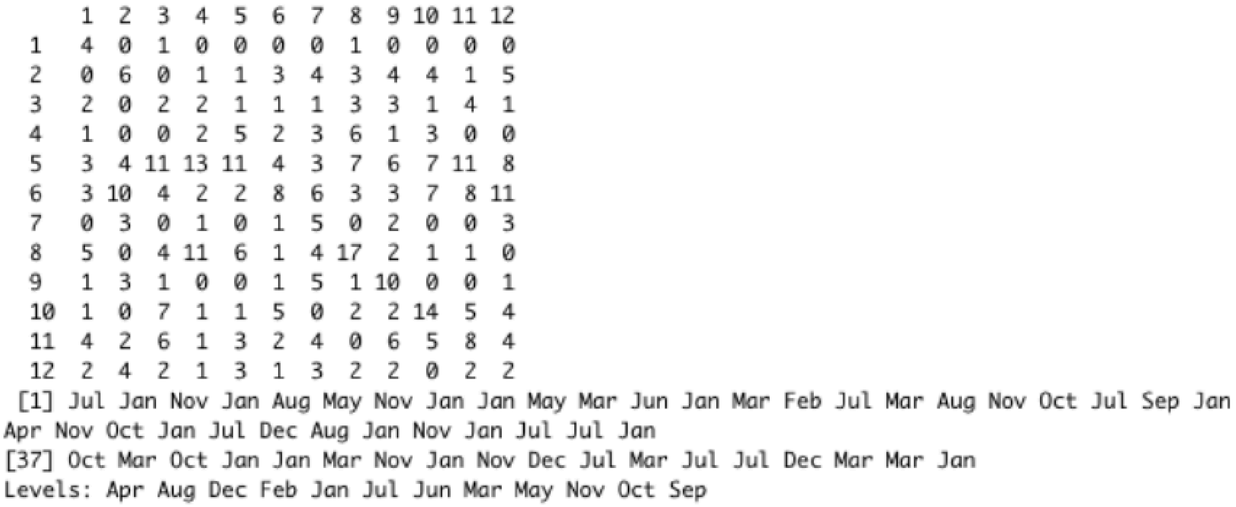
5 Association Rules: Interesting Implications



```
    lhs                                                rhs                                       support confidence  coverage     lift count
[1] {forecast.solar.day.ahead=[1.7e+03,5.84e+03],
     temp=[293,316],
     temp_min=[291,315]}                            => {generation.solar=[1.62e+03,5.79e+03]} 0.1560385  0.9748439 0.1600651 2.923864  5464
    lhs                                                rhs                                       support confidence  coverage     lift count
[1] {forecast.solar.day.ahead=[1.7e+03,5.84e+03],
     clouds_all=[20,100],
     weather_main=clouds}                           => {generation.solar=[1.62e+03,5.79e+03]} 0.1569523  0.9725712 0.1613788 2.917047  5496
    lhs                                         rhs                                         support confidence  coverage     lift count
[1] {total.load.actual=[3.11e+04,4.1e+04],
     clouds_all=[20,100],
     weather_main=clouds}                    => {total.load.forecast=[3.12e+04,4.14e+04]} 0.1532113  0.9587205 0.1598081 2.875997  5365
    lhs                                                        rhs                   support confidence  coverage     lift count
[1] {forecast.wind.onshore.day.ahead=[6.42e+03,1.74e+04],
     weather_main=clear,
     weather_description=sky is clear}                      => {clouds_all=[0,20)} 0.1504127          1 0.1504127 2.186922  5267
    lhs                                                        rhs                     support confidence  coverage     lift count
[1] {generation.wind.onshore=[3.54e+03,6.44e+03),
     forecast.wind.onshore.day.ahead=[3.57e+03,6.42e+03),
     weather_main=clouds}                                   => {clouds_all=[20,100]} 0.1522403  0.9783447 0.1556101 1.802615  5331
```

4. Naive Bayes Predictions:
    a. The model struggled to accurately predict specific months, highlighting the complexity of correlating energy output and weather with calendar months.
    b. January and July were common predictions, suggesting these months had distinctive energy and weather patterns.





```
     1  2  3  4  5  6  7  8  9 10 11 12
  1  4  0  1  0  0  0  0  1  0  0  0  0
  2  0  6  0  1  1  3  4  3  4  4  1  5
  3  2  0  2  2  1  1  1  3  3  1  4  1
  4  1  0  0  2  5  2  3  6  1  3  0  0
  5  3  4 11 13 11  4  3  7  6  7 11  8
  6  3 10  4  2  2  8  6  3  3  7  8 11
  7  0  3  0  1  0  1  5  0  2  0  0  3
  8  5  0  4 11  6  1  4 17  2  1  1  0
  9  1  3  1  0  0  1  5  1 10  0  0  1
  10 1  0  7  1  1  5  0  2  2 14  5  4
  11 4  2  6  1  3  2  4  0  6  5  8  4
  12 2  4  2  1  3  1  3  2  2  0  2  2
 [1] Jul Jan Nov Jan Aug May Nov Jan Jan May Mar Jun Jan Mar Feb Jul Mar Aug Nov Oct Jul Sep Jan
Apr Nov Oct Jan Jul Dec Aug Jan Nov Jan Jul Jul Jan
[37] Oct Mar Oct Jan Jan Mar Nov Jan Nov Dec Jul Mar Jul Jul Dec Mar Mar Jan
Levels: Apr Aug Dec Feb Jan Jul Jun Mar May Nov Oct Sep
```

Conclusion

The study demonstrates the effectiveness of machine learning algorithms in analyzing and predicting energy data. The clear correlation between weather and energy output underscores the potential for these tools to enhance energy management and efficiency. However, the lack of strong correlation between specific months and energy output indicates the need for more refined models and data from regions with more variable weather patterns.

Future research could benefit from extending the analysis to countries with more erratic weather and considering the evolving energy landscape with increased reliance on renewable sources. The insights gained from such studies will be crucial as the world continues to transition to cleaner energy solutions.

By leveraging machine learning, the energy sector can optimize resource utilization, improve efficiency, and contribute to sustainable energy practices.

Walmart Sales Analysis: Insights from Team A's Final Project

In their IST 687 final project, Team A, consisting of Megan, Nate, and Zane, conducted a detailed analysis of Walmart sales data to understand how various factors influence sales across different stores and departments. The project utilized various data science techniques, including data cleaning, exploratory data analysis, and the application of machine learning models. This comprehensive study aimed to provide actionable insights for optimizing sales strategies at Walmart.

Introduction

The primary goal of this project was to analyze Walmart's sales data, incorporating external factors such as weather, economic indicators, and holiday periods. Machine learning algorithms were employed to uncover patterns and relationships within the data, ultimately leading to better-informed business decisions. The team leveraged a dataset from Kaggle, which included variables such as store and department IDs, weekly sales figures, dates, and various economic indicators.

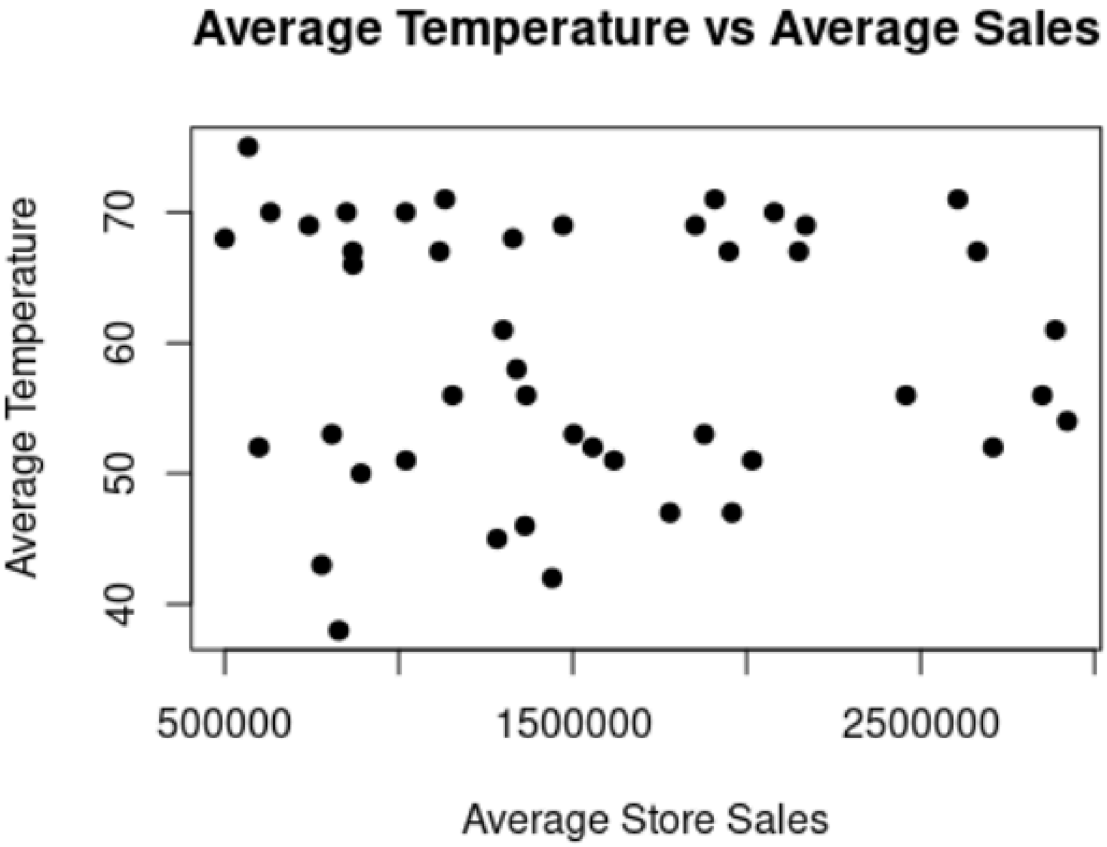
Data Preparation and Cleaning

The initial step involved reading the sales data and additional features data into RStudio, followed by extensive data cleaning and preprocessing. This included handling null values, standardizing date formats, and merging datasets to create a comprehensive dataset for analysis. The team created several subsets of the data to focus on different aspects of sales performance.

Analysis and Visualization

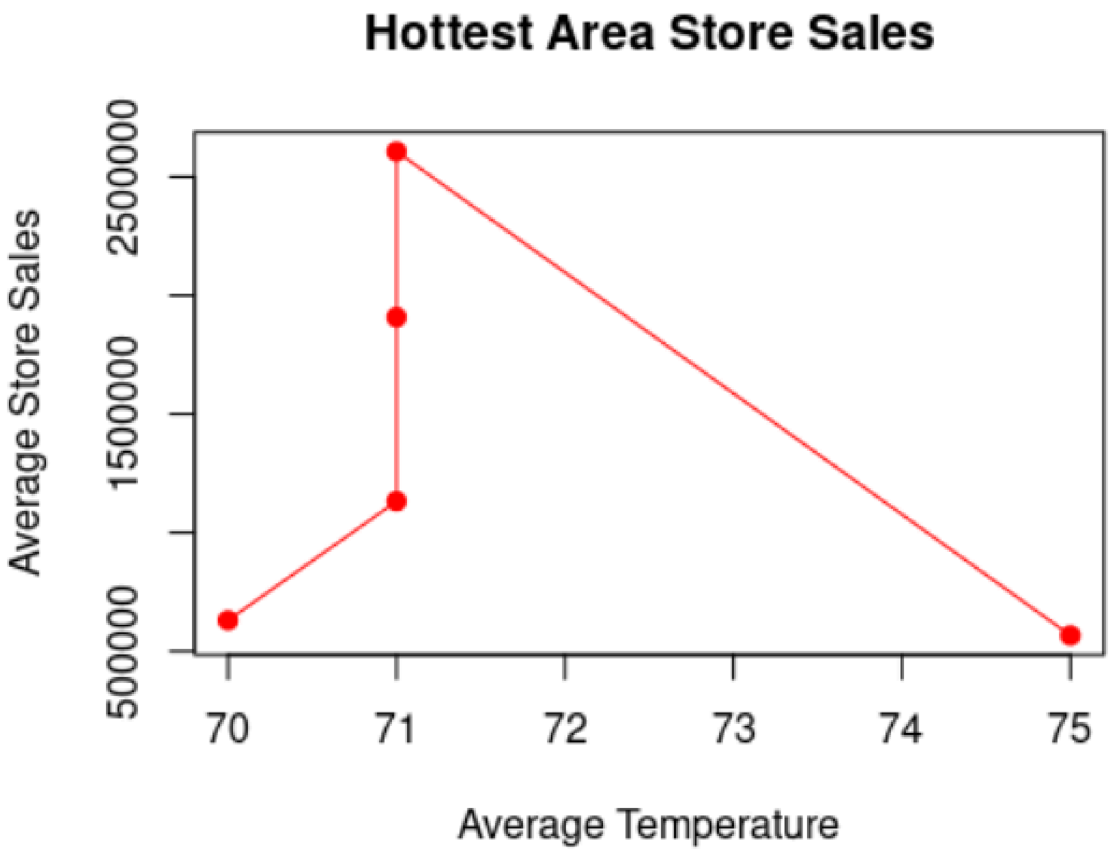
The analysis was divided into three main sections, each addressing a specific business question:

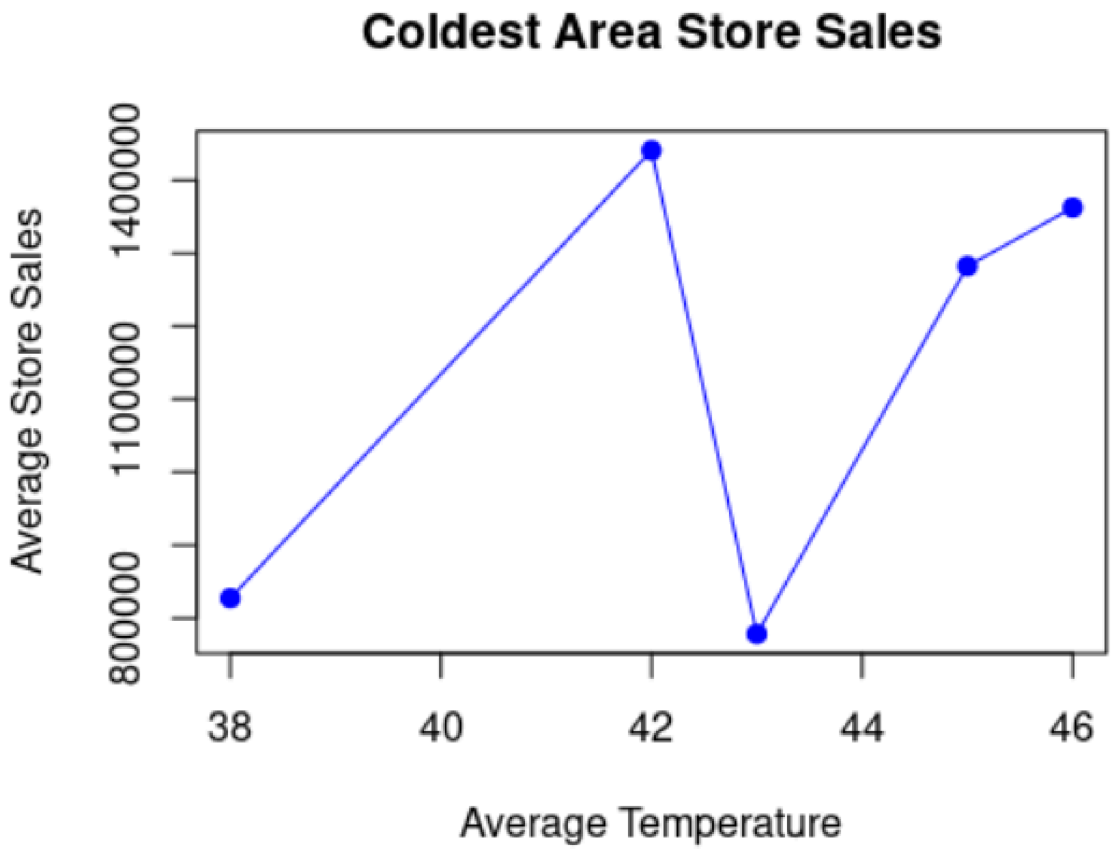
1. Impact of Weather on Store Sales:
    a. The team examined how weather conditions affect sales across 45 different stores. They calculated average weekly sales for each store and created scatter plots to visualize the relationship between average temperature and sales. The analysis revealed that stores in hotter areas tend to have higher average sales, with notable variations across different stores.



2. Departmental Sales Trends:

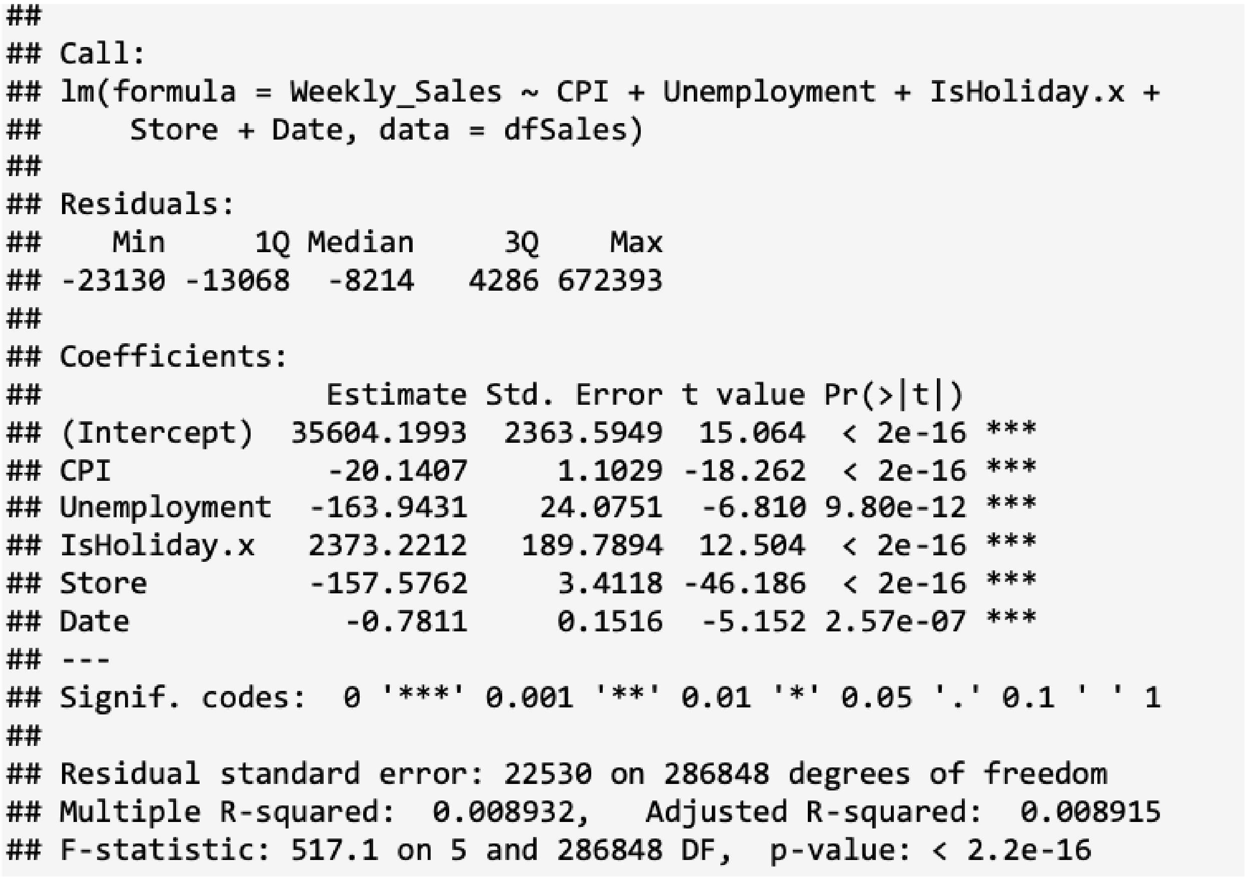
    a. This section focused on how weather impacts the profitability of 99 departments within Walmart stores. The team calculated total sales for each department based on temperature and created line graphs to show sales trends. They identified the top and worst-performing departments in both hot and cold weather, providing insights into which departments might benefit from targeted promotions or relocations.


Hottest Area Store Sales
Average Store Sales
500000
1500000
2500000
70
71
72
73
74
75
Average Temperature



3. Influence of External Factors on Weekly Sales:

    a. The final section explored how various inputs, such as the Consumer Price Index (CPI), unemployment rate, fuel prices, and holiday flags, affect weekly sales. Bivariate scatter plots and multivariate regression models were used to analyze these relationships. The results indicated that CPI, unemployment, and holiday flags significantly impact weekly sales, while temperature and fuel prices were less influential.



```
##
## Call:
## lm(formula = Weekly_Sales ~ CPI + Unemployment + IsHoliday.x +
##     Store + Date, data = dfSales)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -23130 -13068  -8214   4286 672393
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35604.1993  2363.5949  15.064  < 2e-16 ***
## CPI            -20.1407     1.1029 -18.262  < 2e-16 ***
## Unemployment  -163.9431    24.0751  -6.810 9.80e-12 ***
## IsHoliday.x   2373.2212   189.7894  12.504  < 2e-16 ***
## Store         -157.5762     3.4118 -46.186  < 2e-16 ***
## Date            -0.7811     0.1516  -5.152 2.57e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22530 on 286848 degrees of freedom
## Multiple R-squared:  0.008932,   Adjusted R-squared:  0.008915
## F-statistic: 517.1 on 5 and 286848 DF,  p-value: < 2.2e-16
```

Results and Interpretations

The findings from the analysis provided several key insights:

1. Weather and Sales:
    a. Stores in warmer climates generally experienced higher sales, suggesting that these stores might benefit from specific sales strategies tailored to their local weather conditions.
2. Department Performance:
    a. Certain departments, such as groceries and electronics, performed consistently well across different weather conditions, while others showed significant variability. This information can help Walmart optimize product placement and promotional efforts to maximize sales.

3. Economic Indicators:
    a. The multivariate regression analysis revealed that economic indicators like CPI and unemployment significantly impact sales. This suggests that Walmart's sales strategies should consider broader economic conditions to remain effective.

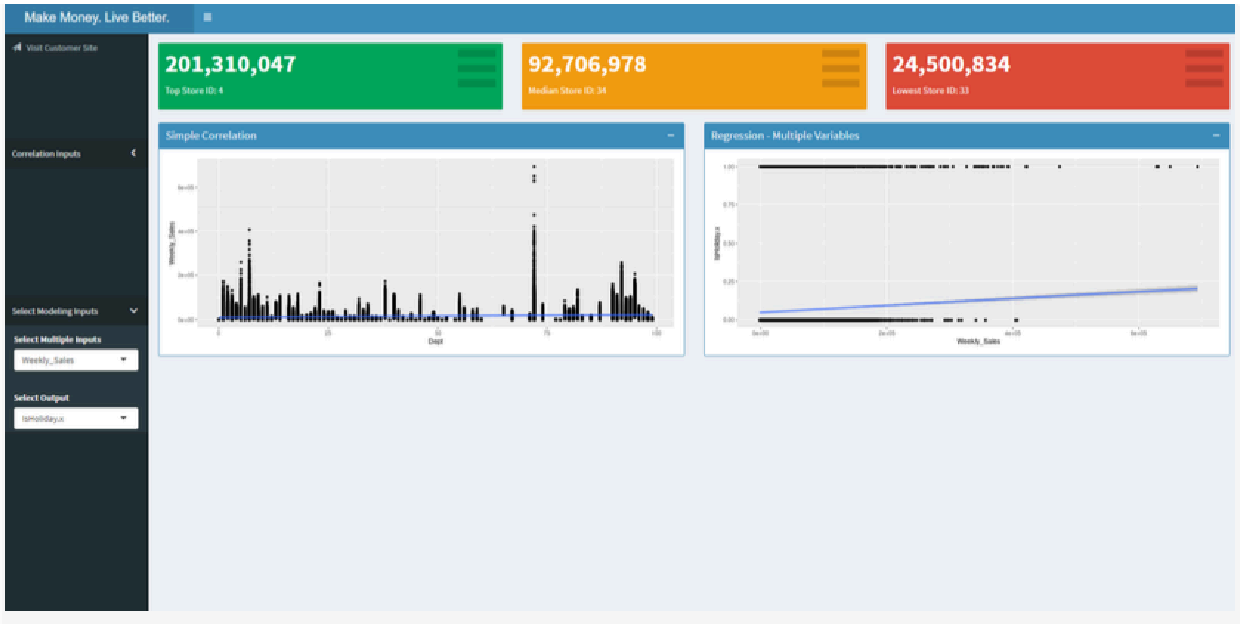
Recommendations

Based on the analysis, the team made several recommendations:

1. Tailored Sales Strategies:
    a. Develop sales plans that consider local weather patterns, focusing on stores in warmer areas to capitalize on higher sales potential.
2. Departmental Adjustments:
    a. Adjust product placement and promotional efforts based on departmental performance in different weather conditions. Departments that perform well in specific conditions should be highlighted during those times.
3. Economic Considerations:
    a. Monitor economic indicators closely and adjust sales strategies accordingly. Understanding the impact of CPI and unemployment on sales can help Walmart make more informed decisions.
4. Enhanced Data Analysis:
    a. Incorporate additional data sources, such as geographic coordinates and more detailed sales data, to refine the analysis further. This could involve developing a dynamic dashboard to track real-time sales trends and economic conditions.

Shiny App Dashboard

As part of the project's future phases, the team proposed developing a Shiny app dashboard. This interactive tool would allow Walmart to dynamically select inputs and monitor sales trends over time using real-time data. The prototype dashboard includes functionalities for visualizing simple correlations and multiple regression models, providing a user-friendly interface for ongoing sales analysis.



Make Money. Live Better.

Visit Customer Site

Correlation Inputs

Select Modeling Inputs

Select Multiple Inputs: Weekly_Sales

Select Output: IsHoliday.x

201,310,047
Top Store ID: 4

92,706,978
Median Store ID: 34

24,500,834
Lowest Store ID: 33

Simple Correlation

Regression - Multiple Variables

Conclusion

Team A's final project on Walmart sales analysis provides valuable insights into how various factors, including weather, economic indicators, and holiday periods, influence sales across different stores and departments. The use of machine learning models and comprehensive data analysis offers actionable recommendations for optimizing sales strategies. By implementing these insights, Walmart can enhance its sales performance and better respond to changing market conditions.