A Data-Driven Analysis of Unidentified Flying Object (UFO) Sightings.

Name: Saurabh Sunil Kamble
School and Department: Rutgers - The State University of New Jersey,
School of Graduate Studies - CS Dept

Course Name: 17:610:560:90 Foundations of Data Science
Assignment Number:
Module 15 Project: Final Submission - Data Analysis Report

Submission Date: 12.20.2023

1. Study Title:

A Data-Driven Analysis of Unidentified Flying Object (UFO) Sightings.

2. Abstract:

Unidentified Flying Objects (UFOs) have fascinated humanity for generations. This study employs a data-driven approach to examine the patterns of UFO sightings, focusing on their shapes, durations, and geographical distributions. Utilizing a detailed dataset (Wright, 2023) spanning over a century, the analysis focuses on the distinctive characteristics of UFOs, such as shape and encounter duration and their geographical locations. By applying statistical methods and predictive modeling, the study aims to uncover potential correlations and predict UFO shape category, offering new insights into the global phenomena of UFO sightings.

3. Introduction:

The fascination with UFOs spans decades, with reports describing unusual light patterns and movements in the sky (Arranz A. , 2017). This research seeks to demystify aspects of these sightings through a methodical examination of a rich dataset comprising various sighting characteristics. By analyzing factors like sighting shapes, durations, and geographical data, the study attempts to decode potential patterns and regional trends in UFO sightings. Utilizing advanced statistical techniques and predictive modeling, it explores the likelihood of sighting shape categories in different geographical regions, potentially shedding light on environmental or demographic influences on UFO sighting occurrences. By combining statistical techniques with machine learning algorithms, the project aims to provide a nuanced understanding of UFO sightings.

4. Research Question/Hypothesis:

Research Question:

Can we predict UFO shape categories based on regional environmental and temporal data using machine learning models?

Hypothesis:

There is a statistically significant association between the shapes and durations of UFO sightings and their geographical locations. Furthermore, these characteristics can be used to predict the category of UFO shape with a high degree of accuracy using a machine learning model that takes into account regional environmental factors and the timing of sightings. Environmental factors, population density, or other regional attributes may influence this pattern. A predictive model is developed to predict the shape categories of sightings.

5. Previous Research:

The subject of Unidentified Flying Objects has long fascinated humanity, combining cultural intrigue with scientific inquiry. Arranz's 2017 article in the South China Morning Post captures this fascination, focusing on the existential question of extraterrestrial existence and humanity's search for life beyond Earth. Complementing this cultural perspective, the National UFO Reporting Center offers a structured approach to UFO phenomena, cataloging sightings by characteristics like shape in their 2023 index. This dual approach highlights UFOs as both a source of human curiosity and a subject for empirical, data-driven analysis, reflecting the complex nature of UFO sightings.

6. Data:

About the data:

The dataset utilized in this study provides an extensive collection of data on UFO sightings, compiled from Kaggle's UFO Sightings. It encompasses detailed records from 1906 to 2014, featuring around 80,000 documented incidents that facilitate a detailed analysis of sighting patterns across time and locations, supported by eyewitness descriptions. Originally, I planned to analyze data that included environmental and demographic factors alongside sighting specifics to explore deeper correlations. However, the dataset primarily focused on the details of the sightings without these additional contextual factors. Despite this, it still offered valuable insights into the patterns and trends of UFO sightings over time and by various characteristics.

The UFO Sightings dataset contains the following columns:

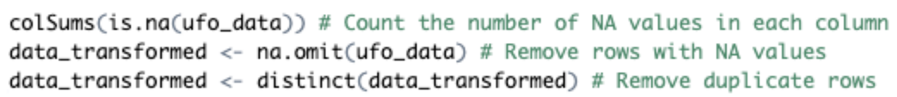
1. Date_time: The date and time of the UFO sighting.
2. date_documented: The date when the UFO sighting was documented.
3. Year: The year in which the UFO sighting occurred.
4. Month: The month in which the UFO sighting occurred.
5. Hour: The hour of the day at which the UFO sighting occurred.
6. Season: The season (e.g., spring, summer, fall, winter) of UFO sightings.
7. Country_Code: A country code where the UFO sighting is reported.
8. Country: The name of the country where the UFO sighting took place.
9. Region: The specific region where the UFO sighting occurred.
10. Locale: Locality information.
11. Latitude: The geographical latitude coordinate of the UFO sighting.
12. Longitude: The geographical longitude coordinate of the UFO sighting.
13. UFO_shape: The reported shape of the UFO.
14. Length_of_encounter_seconds: The duration of UFO sighting inseconds.
15. Encounter_Duration: Timing of the UFO sighting.
16. Description: A description of the UFO sighting.

Each entry within the dataset offers precise temporal data, including the date and time of the sighting, as well as the date of documentation. Geographic specificity is achieved through recorded latitude and longitude coordinates, country, and locale information, allowing for examining regional trends. The dataset also categorizes sightings by year, month, hour, and season, presenting a robust framework to examine temporal dynamics in UFO activity. With 16 variables, the dataset enables statistical methods to uncover correlations and insights into UFO sightings' global nature and distribution.

7. Methodology:

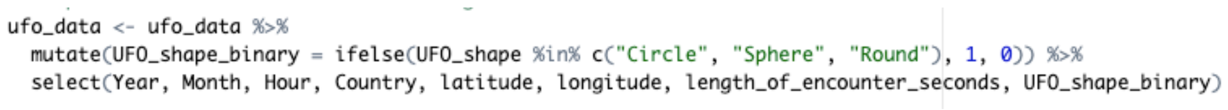
Data Collection and Pre-Processing:

The dataset includes a range of characteristics related to UFO sightings and has been sourced from Kaggle. In the data preprocessing stage, missing values were checked to be removed, to ensure data integrity. Any duplicate records identified were eliminated to guarantee the uniqueness of each sighting event.

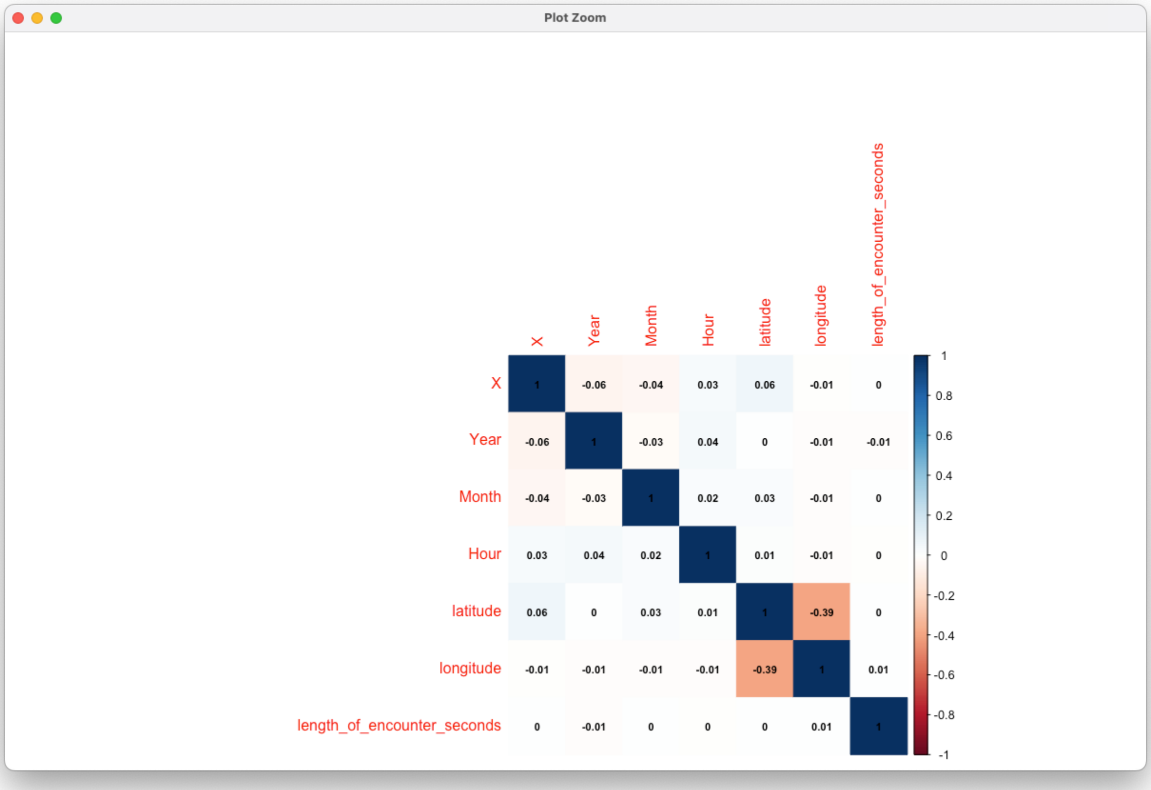


```
colSums(is.na(ufo_data)) # Count the number of NA values in each column
data_transformed <- na.omit(ufo_data) # Remove rows with NA values
data_transformed <- distinct(data_transformed) # Remove duplicate rows
```

The 'UFO_shape' variable was converted into a binary format, outlining UFO shapes as either 'circular' or 'non-circular', to align with the requirements of logistic regression modeling. Additionally, standardization was applied to date-time and geographic coordinates, ensuring consistency across these critical variables for subsequent analytical processes. This rigorous data preparation laid the groundwork for the reliable application of machine learning techniques in the study.



```
ufo_data <- ufo_data %>%
  mutate(UFO_shape_binary = ifelse(UFO_shape %in% c("Circle", "Sphere", "Round"), 1, 0)) %>%
  select(Year, Month, Hour, Country, latitude, longitude, length_of_encounter_seconds, UFO_shape_binary)
```



	X	Year	Month	Hour	latitude	longitude	length_of_encounter_seconds
X	1	-0.06	-0.04	0.03	0.06	-0.01	0
Year	-0.06	1	-0.03	0.04	0	-0.01	-0.01
Month	-0.04	-0.03	1	0.02	0.03	-0.01	0
Hour	0.03	0.04	0.02	1	0.01	-0.01	0
latitude	0.06	0	0.03	0.01	1	-0.39	0
longitude	-0.01	-0.01	-0.01	-0.01	-0.39	1	0.01
length_of_encounter_seconds	0	-0.01	0	0	0	0.01	1

Figure 1: Correlation Matrix of UFO Sightings Data

Correlation Matrix: The correlation matrix plot provides a initial look at the relationships between various features in the dataset.

Exploratory Data Analysis (EDA):

In the EDA phase, the study deployed a suite of visual tools to interrogate and interpret the UFO sightings dataset. Stacked bar plots were generated to illustrate the frequency of UFO shapes over time, offering insights into trends across different periods. A density plot was constructed to visualize the distribution of sightings throughout the years, while word clouds provided a qualitative assessment of the most commonly reported UFO shapes. These visualizations, grounded in thorough statistical summaries, enabled the identification of outliers, patterns, and key relationships that informed the subsequent modeling phase.

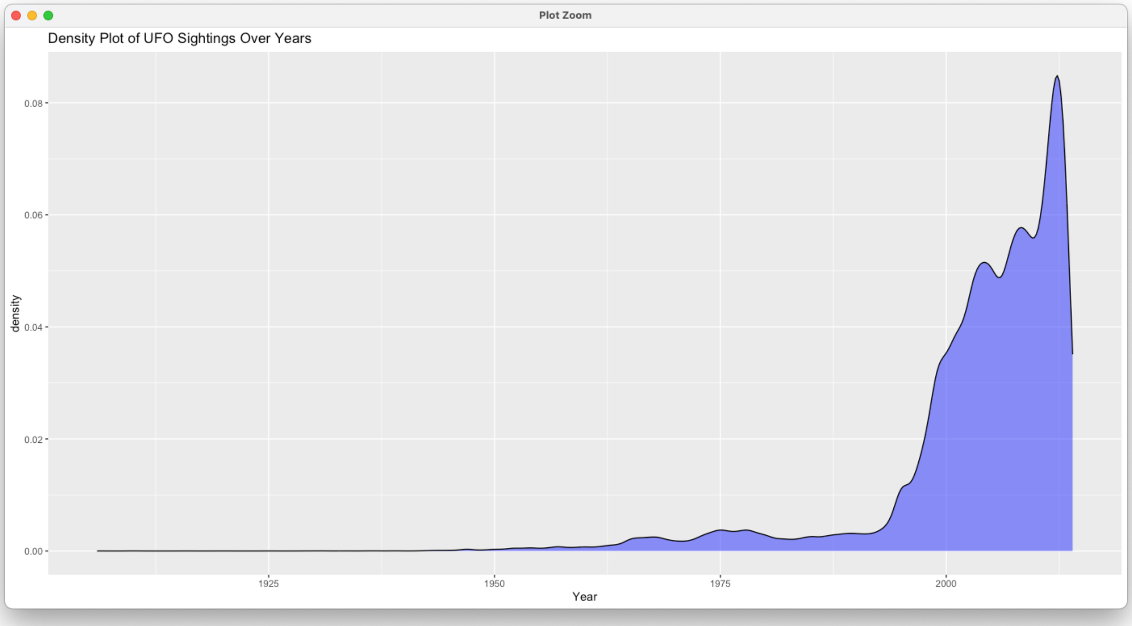


Figure 2: Trend of UFO Sightings Frequency by Year

Density Plot of UFO Sightings Over Years: The density plot serves as a tool to analyze the distribution of sightings over time, highlighting periods of increased reports. This is pivotal in understanding whether there are specific time frames that show a surge in sightings, possibly correlating with external events.

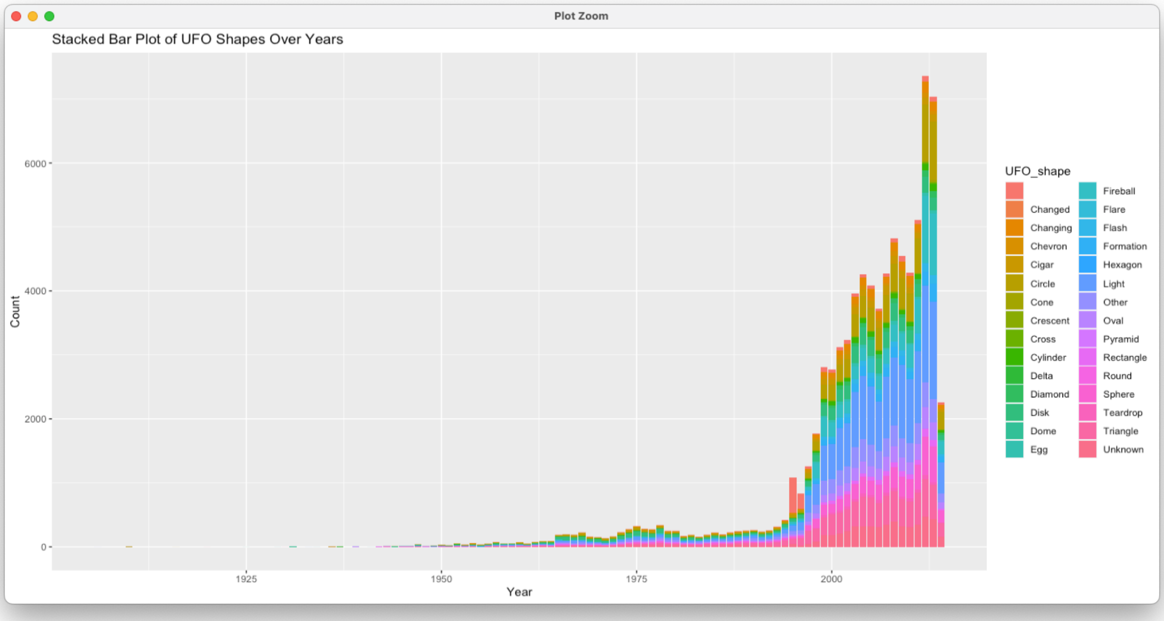


Figure 3: Distribution of Reported UFO Sightings each Year

Stacked Bar Plot of UFO Shapes Over Years: This plot provides a visual representation of the frequency and variety of UFO shapes reported over the years. It offers a clear view of how certain shapes have appeared more frequently over time.

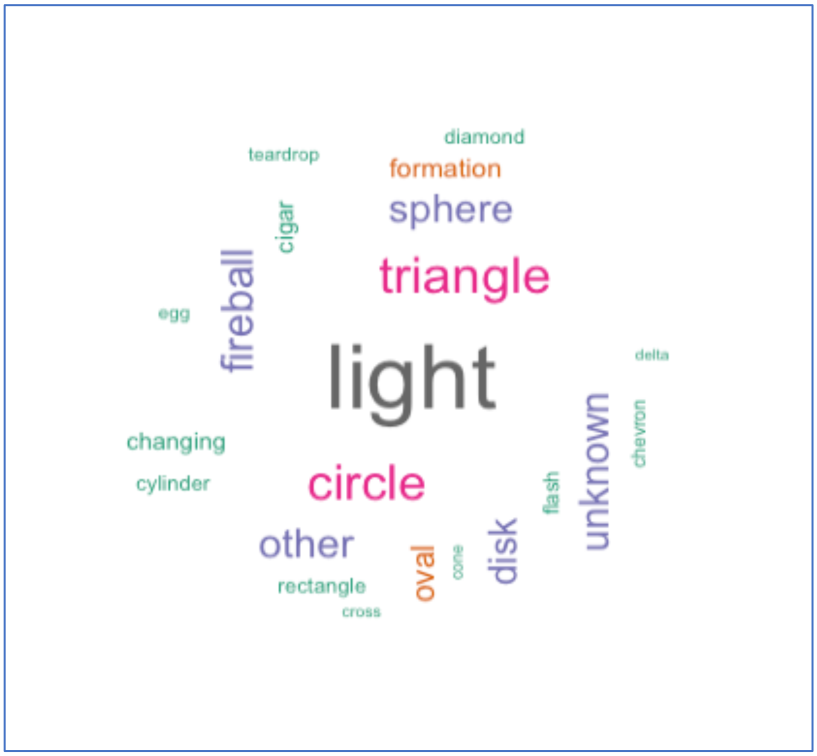


Figure 4: Word Cloud of Reported UFO Shapes

Word Cloud of Reported UFO Shapes: The word cloud is an intuitive way to visually highlight the most frequently reported UFO shapes. It complements quantitative analysis with a qualitative touch, giving a quick sense of which shapes dominate the eyewitness reports.

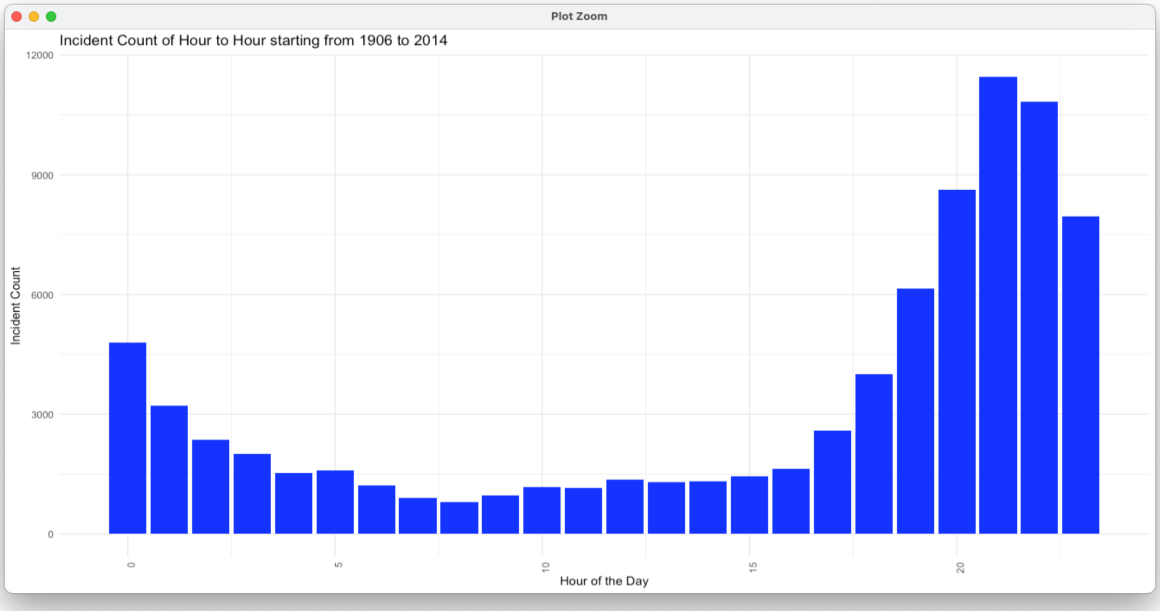


Figure 5: UFO Sightings Incident Count of Hour to Hour

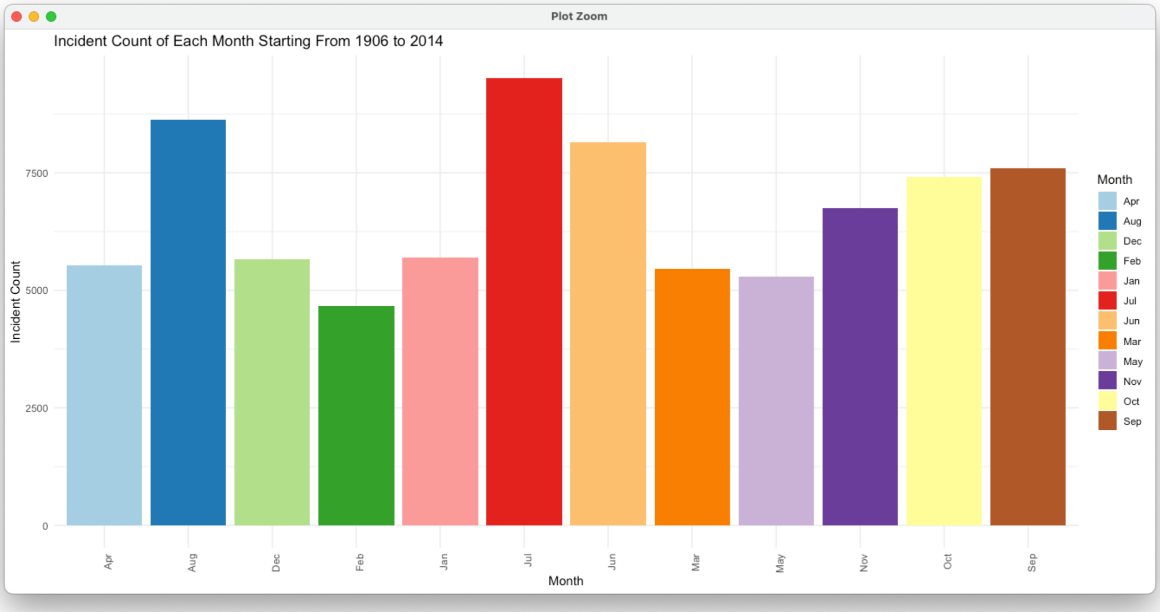


Figure 6: UFO Sightings Incident Count of Each Month

Incident Count of Hour to Hour and Month to Month: These plots help in cracking any temporal patterns, such as specific times of the day or months of the year when sightings are more prevalent. Such patterns could suggest environmental factors or human activities influencing the likelihood of a sighting being reported.

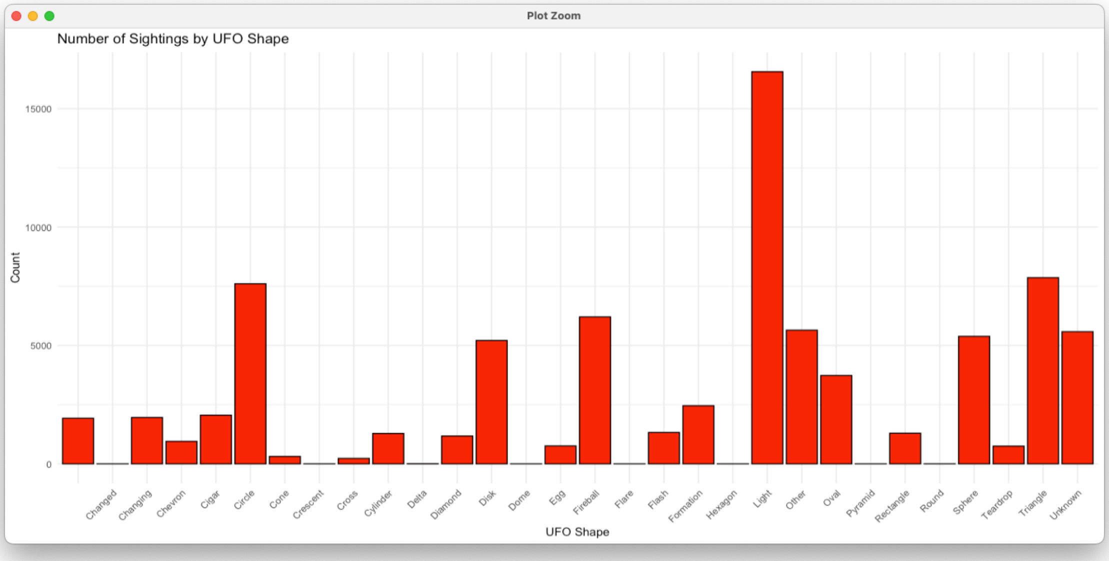


Figure 7: Number of Sightings by UFO Shape

Number of Sightings by UFO Shape: This detailed bar plot allows for a more granular look at the frequency of each specific UFO shape reported, which is crucial for understanding the diversity of sightings.

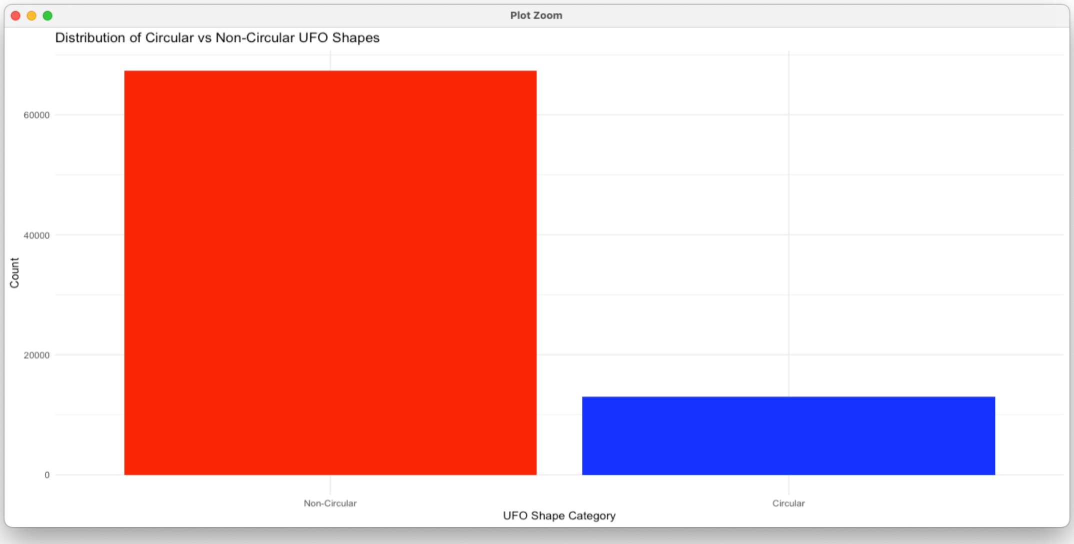
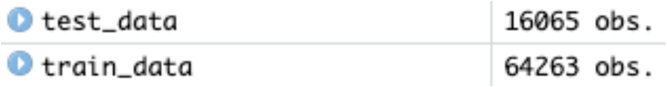


Figure 8: Distribution of Circular vs Non-Circular UFO Shapes

Distribution of Circular vs Non-Circular UFO Shapes: By categorizing UFO shapes into binary groups, this plot was instrumental in preparing the data for logistic regression modeling. It also helps to quickly identify the most commonly reported shapes, which is essential for the predictive analysis.

Model Implementation:

The study employed logistic regression via the glmnet package for its predictive analysis. The process involved transforming the 'UFO_shape' variable into a binary format, crucial for the binary logistic regression approach. The dataset was split into an 80-20 ratio for training and testing, ensuring a robust validation process. This setup was crucial for implementing the glmnet package, which facilitated the logistic regression analysis. The focus was on tuning the model parameters, particularly regularization parameters, through cross-validation techniques, thereby optimizing the model's predictive capabilities and ensuring its generalizability to new data.



test_data	16065 obs.
train_data	64263 obs.

Model Evaluation:

The model's performance was evaluated using a combination of statistical metrics and visualization techniques. The precision and F1 scores were calculated. The ROC and Precision-Recall curves were plotted to visualize the model's classification capability, and a calibration curve was employed to examine the relationship between predicted probabilities and outcomes. These comprehensive evaluation methods confirmed the model's predictive power and reliability in the context of UFO shape category classification.

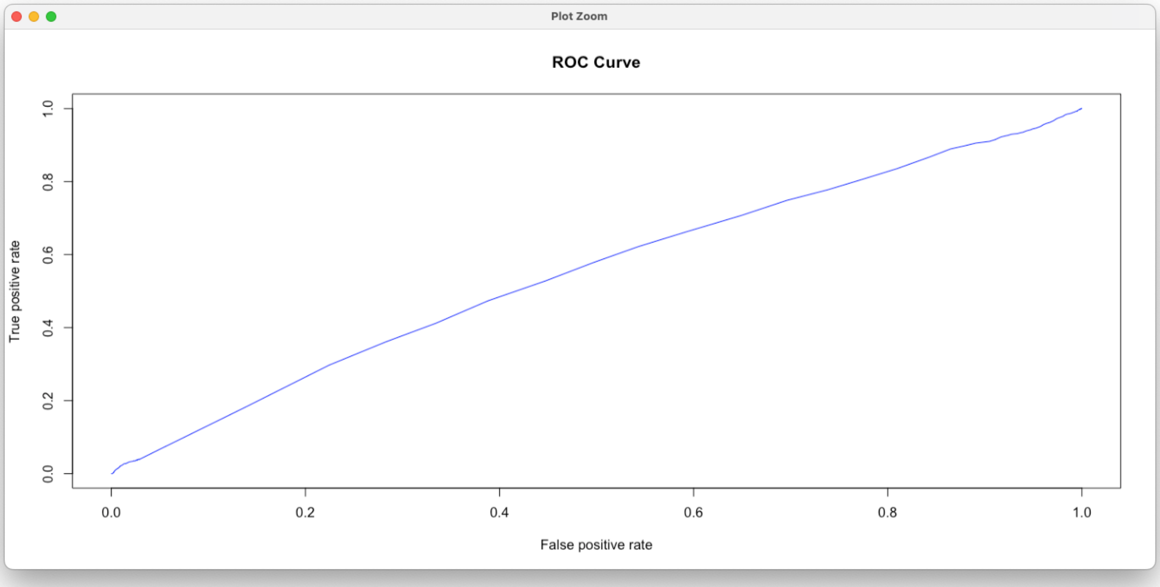


Figure 9: ROC Curve

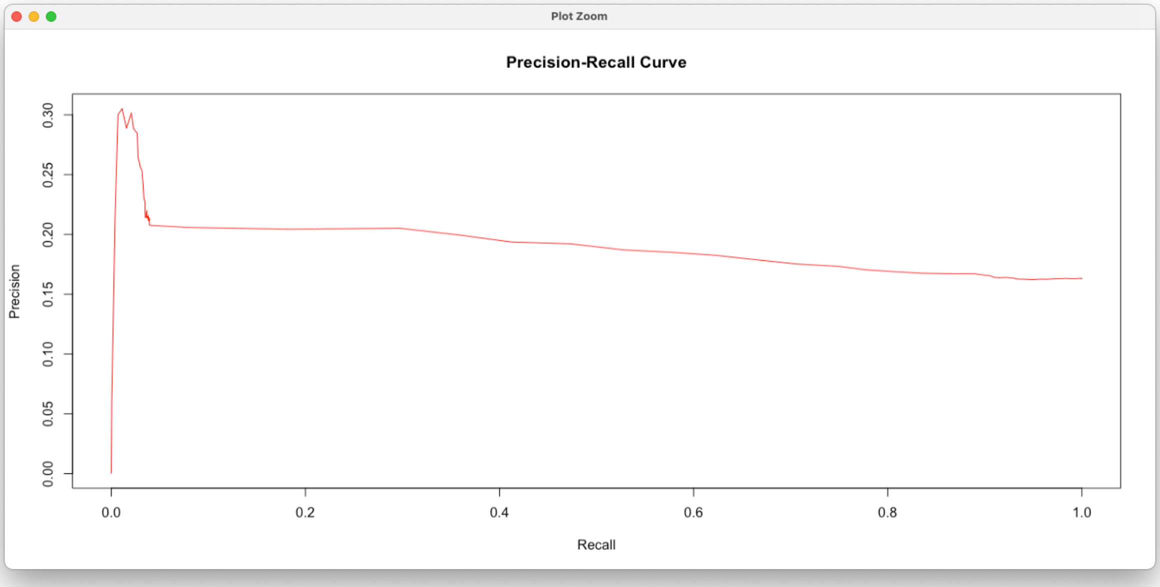


Figure 10: Precision-Recall Curve

Precision-Recall and ROC Curves: These plots are essential for evaluating the performance of the logistic regression model. They provide insights into the trade-offs between precision and recall and the overall ability of the model to distinguish between the binary classes of UFO shapes.

8. Results:

The model evaluation utilized a logistic regression model with regularization, trained to distinguish between circular and non-circular UFO shape categories. The Receiver Operating Characteristic (ROC) curve showed a moderate true positive rate over the false-positive rate, indicating that the model can reasonably differentiate between the two classes. However, the Precision-Recall Curve indicated that the model has room for improvement, as the precision declined with increasing recall. The model demonstrated a high accuracy of 83.68%. The F1 Score, a harmonic mean of precision and recall, was found to be approximately 0.911, which would usually suggest a good balance between precision and recall.

9. Limitations:

The present analysis study, while comprehensive in its approach to analyzing UFO sightings data, encounters several limitations that must be acknowledged:

- Data Quality and Reporting Bias: The dataset's integrity is only as reliable as the reporting mechanisms. Inconsistencies in reporting, potential misclassifications by observers, and varying levels of detail across reports can introduce noise and bias into the analysis.
- Temporal and Geographical Coverage: The data may not uniformly cover all temporal and geographical regions. Certain periods or locations might be overrepresented due to higher public interest, media coverage, or reporting practices, which can skew the analysis.
- Algorithmic Limitations: The choice of logistic regression with regularization may not be the most effective approach for the data at hand. More sophisticated models or ensemble methods might offer improved discrimination between classes and better handle imbalanced datasets.
- Generalizability: The findings and model performance are specific to the dataset used and may not generalize to other UFO sighting reports or datasets with different characteristics.
- Unaccounted Variables: There may be external factors or variables not included in the dataset that could significantly influence UFO sightings, such as weather conditions, celestial events, or changes in surveillance technology.

10. Conclusion:

The comprehensive analysis of UFO sightings data, over a century of UFO sighting reports has shed light on notable patterns within this global phenomenon. Data preprocessing and logistic regression modeling highlighted an increased frequency of sightings over time, with specific shapes like circles, lights, and triangles being reported more frequently. The logistic regression model distinguished between circular and non-circular shape categories with success, achieving an accuracy of 83.68%. However, the model's bias toward the majority class and potential limitations in data reporting and representativeness suggest that further refinement is needed. Despite these challenges, the study demonstrates the value of machine learning in analyzing complex phenomena. The insights gained pave the way for future studies, which could benefit from incorporating broader variables and leveraging more sophisticated algorithms to deepen our understanding of UFO sightings. In conclusion, this study provides a stepping stone for ongoing exploration in the study of UFO sightings, with data science proving to be a crucial tool in unraveling the mysteries of the skies.

11. Future Scope:

- Multinomial Classification: Expanding the binary classification of UFO shapes to a multinomial approach could offer more nuanced insights into the varied shapes of UFO sightings reported and potentially reveal more intricate patterns.
- Integration of Environmental and Demographic Data: Incorporating additional datasets that include environmental variables like weather patterns and demographic information such as population density could enhance the predictive modeling and provide a deeper understanding of the factors influencing sightings.
- Temporal Deep Dive: A focused analysis on temporal patterns, possibly using time-series analysis, to determine cyclical behaviors or correlations with known celestial events.
- Geospatial Analysis: Applying advanced geospatial analytics to understand regional hotspots and movements of UFO sightings over time, potentially linking these to geopolitical or natural phenomena.
- Technological Advancements in Detection: Exploring the impact of advancements in technology and surveillance on the frequency and accuracy of UFO reports.

12. References: APA Style

Arranz, A. (2017, December 24). *Are we alone?* Retrieved from South China Morning Post.: https://multimedia.scmp.com/culture/article/ufo/index.html

National UFO Reporting Center. (2023, July 22). *UFO Report Index by Shape of Craft*. Retrieved from NUFORC Reports by Shape: https://nuforc.org/ndx/?id=shape

Wright, J. &. (2023). *UFO Sightings* . Retrieved from Kaggle: https://doi.org/10.34740/KAGGLE/DSV/6913786

R Code:

```
# Library Installation ----------------------------------------------------------------------
----------------------------------------------
install.packages("gridExtra")
install.packages("CatEncoders")
install.packages("corrplot")
install.packages("ggplot2")
install.packages("tm")
install.packages("wordcloud")
install.packages('dplyr')
install.packages('corrplot')
install.packages("glmnet")
install.packages("rpart")
install.packages("caret",dependencies = TRUE)
install.packages("ROCR")
install.packages("Rcpp")
install.packages("e1071",dependencies = TRUE)
install.packages("gplots")

# Library Loading -------------------------------------------------------------------------
-------------------------------------------------
library(dplyr)
library(readr)
library(dplyr)
library(caret)
library(ggplot2)
library(glmnet)
library(corrplot)
library(ROCR)
library(wordcloud)
library(RColorBrewer)

# Data Loading into a vector------------------------------------------------------------
----------------------------------------------------
```

```
ufo_data <- read.csv("~/Desktop/Fall 23/FDS-
560/FDS_FINAL_PROJECT/ufo-sightings-transformed.csv")

# Some Initial Data Exploration --------------------------------------------------------------------
-----------------------------------------------------------------

head(ufo_data) # Display the first 5 rows of the data
print(dim(ufo_data)) # Print the dimensions (rows, columns) of the data
print(names(ufo_data)) # Print the column names of the data
print(str(ufo_data)) # Print the structure and data types of columns
print(summary(ufo_data)) # Print summary statistics for each columns

# Data Cleaning ----------------------------------------------------------------------------------------------
------------------------------------------------------------

colSums(is.na(ufo_data)) # Count the number of NA values in each
column
data_transformed <- na.omit(ufo_data) # Remove rows with NA values
data_transformed <- distinct(data_transformed) # Remove duplicate rows

# Data Visualization----------------------------------------------------------------------------------------
--------------------------------------

#Stacked Bar Plot for UFO Shapes Over Time
ggplot(ufo_data, aes(x = Year, fill = UFO_shape)) +
  geom_bar(position = "stack") +
  labs(title = "Stacked Bar Plot of UFO Shapes Over Years", x = "Year", y
= "Count")

#Density Plot for Year
ggplot(ufo_data, aes(x = Year)) +
  geom_density(fill = "blue", alpha = 0.5) +
  labs(title = "Density Plot of UFO Sightings Over Years", x = "Year")

# Define a color palette
color_palette <- brewer.pal(8, "Dark2")
```

```
# Create the word cloud
wordcloud(words = ufo_data$UFO_shape,
          max.words = 100,
          random.order = FALSE,
          rot.per = 0.35, # 35% of words are displayed at an angle
          scale = c(3, 0.5), # Scale between most and least frequent words
          colors = color_palette)

# Correlation matrix
numeric_var <- data_transformed %>% select_if(is.numeric)
cat_var <- data_transformed %>% select_if(is.factor)
corr_matrix<-cor(numeric_var) #correlation matrix for the numerical data
corrplot(corr_matrix, method = "color", addCoef.col = "black",
number.cex = 0.7)

# Create a summary count of incidents by hour
hourly_counts <- ufo_data %>%
  group_by(Hour) %>%
  summarise(Count = n()) %>%
  arrange(Hour)

# Create a summary count of incidents by month
monthly_counts <- ufo_data %>%
  group_by(Month) %>%
  summarise(Count = n()) %>%
  arrange(Month)

# Convert month number to month name for better readability
monthly_counts$Month <- month.abb[monthly_counts$Month]

# Plotting the hourly count bar plot
ggplot(hourly_counts, aes(x = Hour, y = Count)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Incident Count of Hour to Hour starting from 1906 to
2014",
      x = "Hour of the Day", y = "Incident Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Plotting the monthly count bar plot
ggplot(monthly_counts, aes(x = Month, y = Count, fill = Month)) +
  geom_bar(stat = "identity") +
  scale_fill_brewer(palette = "Paired") + # Using a color palette for
differentiation
  labs(title = "Incident Count of Each Month Starting From 1906 to 2014",
      x = "Month", y = "Incident Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Count shapes
shape_counts <- table(ufo_data$UFO_shape)

# Display the counts
print(shape_counts)

# Ensure 'UFO_shape' column is a factor for proper ordering in the bar
plot
ufo_data$UFO_shape <- as.factor(ufo_data$UFO_shape)

# Create a bar plot for UFO shapes
ggplot(ufo_data) +
  aes(x = UFO_shape) +
  geom_bar(fill = "red", color = "black") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Number of Sightings by UFO Shape", x = "UFO Shape", y
= "Count")


# Preprocess the Data for Model Building -------------------------------------
-----------------------------------------------
```

```
ufo_data <- ufo_data %>%
  mutate(UFO_shape_binary = ifelse(UFO_shape %in% c("Circle",
"Sphere", "Round"), 1, 0)) %>%
  select(Year, Month, Hour, Country, latitude, longitude,
length_of_encounter_seconds, UFO_shape_binary)

# Bar Plot to visualize the distribution of Circular vs Non-Circular shapes
ggplot(ufo_data, aes(x = as.factor(UFO_shape_binary), fill =
as.factor(UFO_shape_binary))) +
  geom_bar(show.legend = FALSE) + # Hide legend
  scale_x_discrete(labels = c("0" = "Non-Circular", "1" = "Circular")) +
  scale_fill_manual(values = c("0" = "red", "1" = "blue")) + # Assign
colors to each category
  labs(title = "Distribution of Circular vs Non-Circular UFO Shapes",
       x = "UFO Shape Category",
       y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0))

# Dummy Variable Creation and Standardization ------------------------------
-------------------------------------------------------
# Convert 'Country' to factor and create dummy variables
library(caret)
ufo_data$Country <- as.factor(ufo_data$Country)
dummy_model <- dummyVars(" ~ .", data = ufo_data)
ufo_data_transformed <- predict(dummy_model, ufo_data)

# Convert to data frame and ensure UFO_shape_binary is a column
ufo_data_transformed <- data.frame(ufo_data_transformed)
ufo_data_transformed$UFO_shape_binary <-
ufo_data$UFO_shape_binary

standardization <- function(x) {
  (x - mean(x)) / sd(x)
}
```

```
ufo_data_transformed <-
data.frame(lapply(ufo_data_transformed,standardization))
ufo_data_transformed$UFO_shape_binary <-
ufo_data$UFO_shape_binary



# Split the data into training and testing sets--------------------------------------------------
---------------------------------------------------------------
set.seed(123)
splitIndex <-
createDataPartition(ufo_data_transformed$UFO_shape_binary, p = 0.8,
list = FALSE)
train_data <- ufo_data_transformed[splitIndex,]
test_data <- ufo_data_transformed[-splitIndex,]


# Model Training -------------------------------------------------------------------------------------------
------------------------------------------------------------
train_data <- train_data[, colnames(test_data)]

# Train the logistic regression model with regularization
x <- model.matrix(UFO_shape_binary ~ . - 1, data = train_data) # -1 to
exclude the intercept
y <- train_data$UFO_shape_binary

# Fit the model using cross-validation
cv_model <- cv.glmnet(x, y, family = "binomial", alpha = 1) # alpha = 1
for Lasso
best_lambda <- cv_model$lambda.min

# Train the final model
final_model <- glmnet(x, y, family = "binomial", lambda = best_lambda,
alpha = 1)

# Make predictions on the test data
```

```
test_x <- model.matrix(UFO_shape_binary ~ . - 1, data = test_data)
predictions <- predict(final_model, newx = test_x, type = "response", s =
best_lambda)
predicted_class <- ifelse(predictions > 0.5, 1, 0)

# Model Evaluation and Visualization --------------------------------------------
-------------------------------------------------------------------


#ROC curve
roc_pred <- prediction(predictions, test_data$UFO_shape_binary)
roc_perf <- performance(roc_pred, "tpr", "fpr")
plot(roc_perf, col = "blue", main = "ROC Curve")

#Precision Recall Curve
pr_pred <- prediction(predictions, test_data$UFO_shape_binary)
pr_perf <- performance(pr_pred, "prec", "rec")
plot(pr_perf, col = "red", main = "Precision-Recall Curve")

# F1 Score, Precision, and Recall
conf_matrix <- confusionMatrix(predicted_class_factor,
actual_class_factor)
f1_score <- 2 * (conf_matrix$byClass["Precision"] *
conf_matrix$byClass["Recall"]) / (conf_matrix$byClass["Precision"] +
conf_matrix$byClass["Recall"])
precision <- conf_matrix$byClass["Precision"]

cat("F1 Score:", f1_score, "\n")
cat("Precision:", precision, "\n")

# Calculate and Print Model Accuracy
accuracy <- sum(predicted_class == test_data$UFO_shape_binary) /
length(test_data$UFO_shape_binary)
print(paste("Accuracy:", round(accuracy, 4)))
```