CR_SDAAN_9: MSc Data Science and Analytics

COMP8060_27330 Scientific Prog in Python

Python Group Assignment – Blue Sqaudron.

Thomas Flynn: R00242398

Edward Duffy: R00257226

Prem Vikas: R00241672

Akash: R00241122

Provia Kadusabe: R00241719

Chinmay: R00241525

Introduction.

The primary aim of this report is to analyse and interpret data related to Unidentified Flying Objects(UFOs) sightings which are often a hot topic people are curious about. The dataset contains over 80,000 recorded UFO sightings dating back as far as 1949. There are 16 variables used to describe these UFO sightings. By thoroughly analysing this large dataset of reported sightings, we aim to identify patterns, trends, and any discrepancies within the data that could offer new insights about UFO sightings.

Exploratory Data Analysis (EDA):

In this segment, our focus will be on delving into the dataset to unveil its statistical characteristics and conduct data pre-processing to eliminate any inconsistencies that could hinder our analysis. Subsequently, we aim to engage in hypothesis testing through thorough data exploration. Our objective is to reveal connections between variables and derive valuable insights from the data.

Data Analysis.

The dataset on UFO sightings is full of valuable information. It consists of 16 different variables that help describe each sighting. This variety gives us many different options to analyse the dataset. The 16 variables are:

- Date_time: Standardised date and time of sighting.
- date_documented: Date when the sighting was reported.
- Year: Year of sighting.
- Month: Month of sighting.
- Hour: Hour of sighting.
- Season: Season of the sighting.
- Country_Code: - Country code for the country of the sighting.
- Country: Country name.
- Region: Broader location area, like state or province.
- Locale: Specific location, such as city or town.
- latitude: Latitude.
- longitude: Longitude.
- UFO_shape: A one-word description of the spacecraft.
- length_of_encounter_seconds: Standardised to seconds, length of the observation of the UFO.
- Encounter_Duration: Raw description of the length of the encounter (shows uncertainty to previous column).
- description: Text description of the UFO encounter.

Data Description.

To Analyse the given dataset, we need to explore the data by using .info function to get the data types of the variables, the count of non-null values in each column and the column name. The below are the details of the given dataset.

Variables	Data Type	Data classification
Date_time	Numerical (Continous)	Object
date_documented	Numerical (Continuous)	Object
Year	Numerical (Discrete)	Float
Month	Numerical (Discrete)	Float

Hour	Numerical (Discrete)	Float
Season	Categorical (Nominal)	Object
Country_Code	Categorical (Nominal)	Object
Country	Categorical (Nominal)	Object
Region	Categorical (Nominal)	Object
Locale	Categorical (Nominal)	Object
latitude	Numerical (Continuous)	Float
longitude	Numerical (Continuous)	Float
UFO_shape	Categorical (Nominal)	Object
length_of_encounter_seconds	Numerical (Continuous)	Float
Encounter_Duration	Categorical (Nominal)	Object
description	Textual	Object

Table 1. Details of the given Dataset

Data Cleaning.

To carry out our analysis, we first start by making sure the data is cleaned and accurate. This means we will check for any errors or missing information and fix them. This is essential for a reliable analysis. This step is crucial in setting a strong foundation for our further analysis. Next, we can examine the basic patterns found in the data. This initial exploration helps us gain a better foundational understanding of the dataset and guides our further complex analysis.
By using the isnull function we check if there are any null values in the dataset.

```
Unnamed: 0                      0
Date_time                     796
date_documented               798
Year                          794
Month                         799
Hour                          799
Season                        793
Country_Code                 1053
Country                      1048
Region                       1356
Locale                       1249
latitude                      798
longitude                     793
UFO_shape                    2709
length_of_encounter_seconds   796
Encounter_Duration            798
Description                   814
dtype: int64
```

Figure 1. Output of the isnull function to check for null values count.

From the above figure we can observe that there are more than 15,000 null values in the dataset and it needs to treated.
To tackle the above we replace the mean of the column for null values for numeric variables and do the same for the categorical variables.
As the result of replacing all null values with mean and mode the total null values count is reduced to zero.

Data Summary.

The dataset summary can be generated using the describe function applied to the numeric columns of entire dataset. This function aids in computing essential statistical summaries that facilitate a better understanding of the data.

	Month	Hour	latitude	longitude	\
count	80328.000000	80328.000000	80328.000000	80328.000000	
mean	6.834902	15.554365	38.145767	-86.789185	
min	1.000000	0.000000	-82.862752	-176.658056	
25%	4.000000	10.000000	34.159167	-112.073333	
50%	7.000000	19.000000	39.414167	-87.903333	
75%	9.000000	21.000000	42.737043	-78.878005	
max	12.000000	23.000000	72.700000	178.441900	
std	3.218062	7.725925	10.395275	39.475666	

	length_of_encounter_seconds
count	8.032800e+04
mean	8.931977e+03
min	1.000000e-03
25%	3.000000e+01
50%	1.800000e+02
75%	6.000000e+02
max	9.783600e+07
std	6.198571e+05

Figure 2. Output of the describe function.

From the figure we can observe the different stats analysis:

- The minimum length of the encounter is 0.001 seconds and 97836000 is longest time length of encounter.
- The first year of encounter is 1906 and the latest year of encounter is 2014.

Data Visualization

The graphical representation of data and information involves creating visual elements such as charts, graphs, other visual aids to present data patterns, trends, and insights. Effective data visualization can help in understanding complex datasets, identifying relationships between variables, and communicating findings more intuitively and efficiently.

Histogram of UFO sightings over the years.

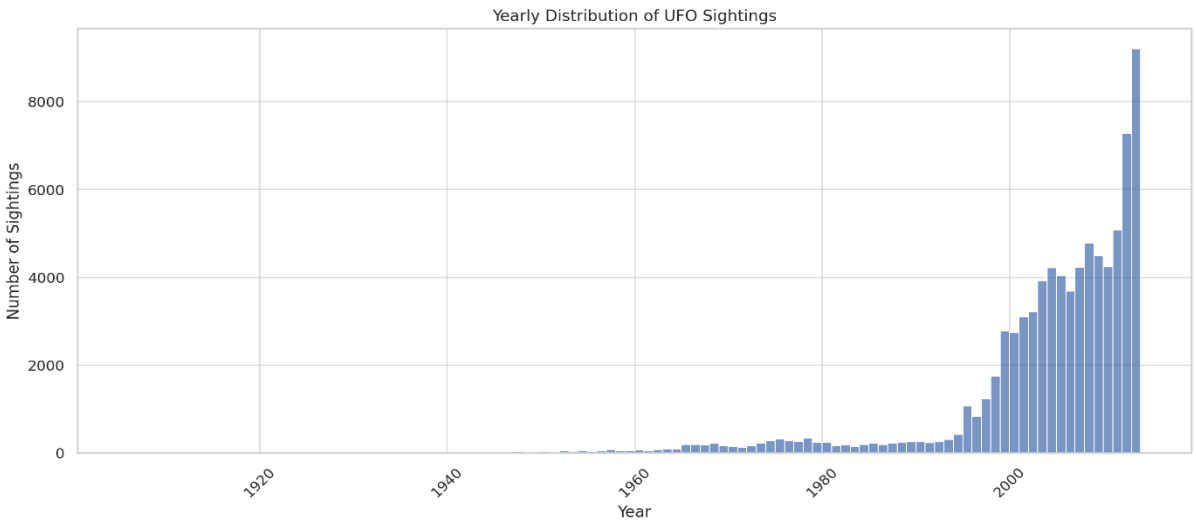


Figure 3. Histogram of UFO sightings over the years

The histogram above shows the distribution of UFO sightings across different years. It appears that the number of reported sightings has increased significantly in more recent years. This trend might be due to a variety of factors like more widespread use of technology to record and report sightings, or other sociocultural factors.

Barchat of UFO shape and Sightings count.

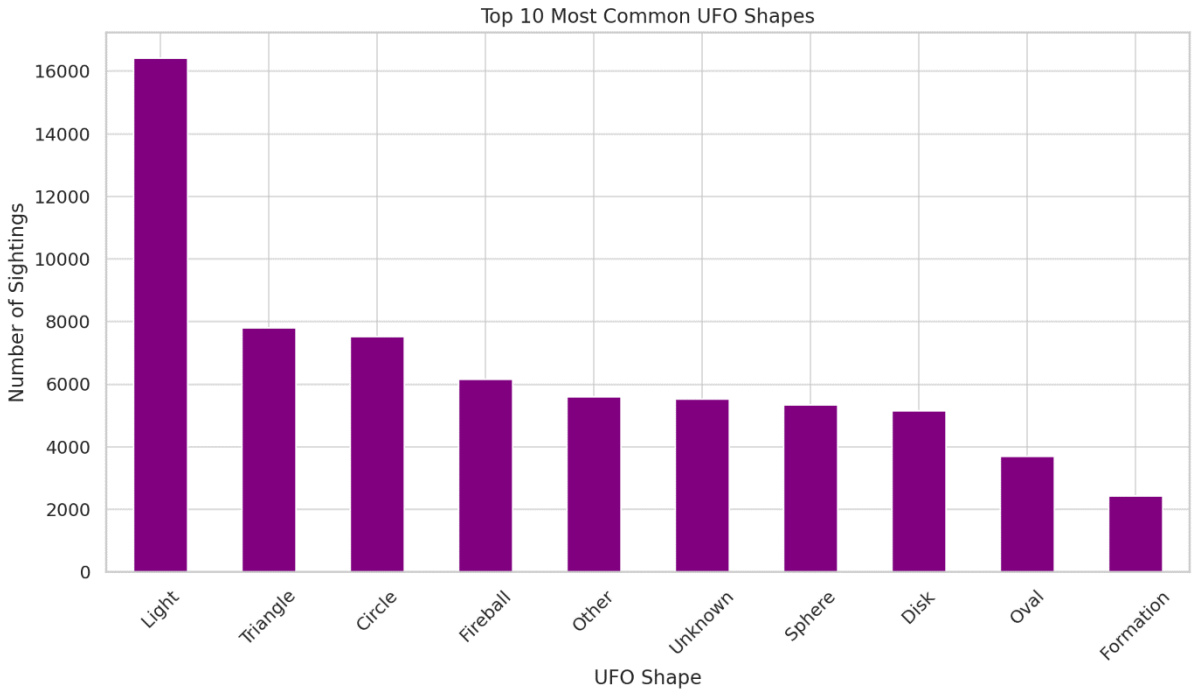


Figure 4. Barchart of UFO sightings and UFO shapes.

The bar chart and the data show the top 10 most commonly reported UFO shapes. The shape reported most frequently is "Light," followed by "Triangle," "Circle," "Fireball," and other shapes such as "Sphere," "Disk," and "Oval."

These findings suggest a diverse range of descriptions for UFOs, with "Light" being particularly common. This could be due to the fact that lights in the sky are often unidentifiable and can be perceived differently depending on the observer's perspective and environmental conditions.

Barchart of Country and UFO sightings.

The bar chart and the accompanying data show the top 10 countries with the most UFO sightings. It's evident that the United States has a significantly higher number of reported sightings compared to other countries, with over 70,000 instances. This is followed by Canada, the United Kingdom, Australia, and other countries, but with a much lower count.

This disparity could be due to several factors:

Cultural Factors: There might be more interest and a greater tendency to report UFO sightings in the United States.

Reporting Mechanisms: The ease of reporting, technology to receive any unusual signals and the prevalence of reporting activities in the United States might contribute to the higher numbers.

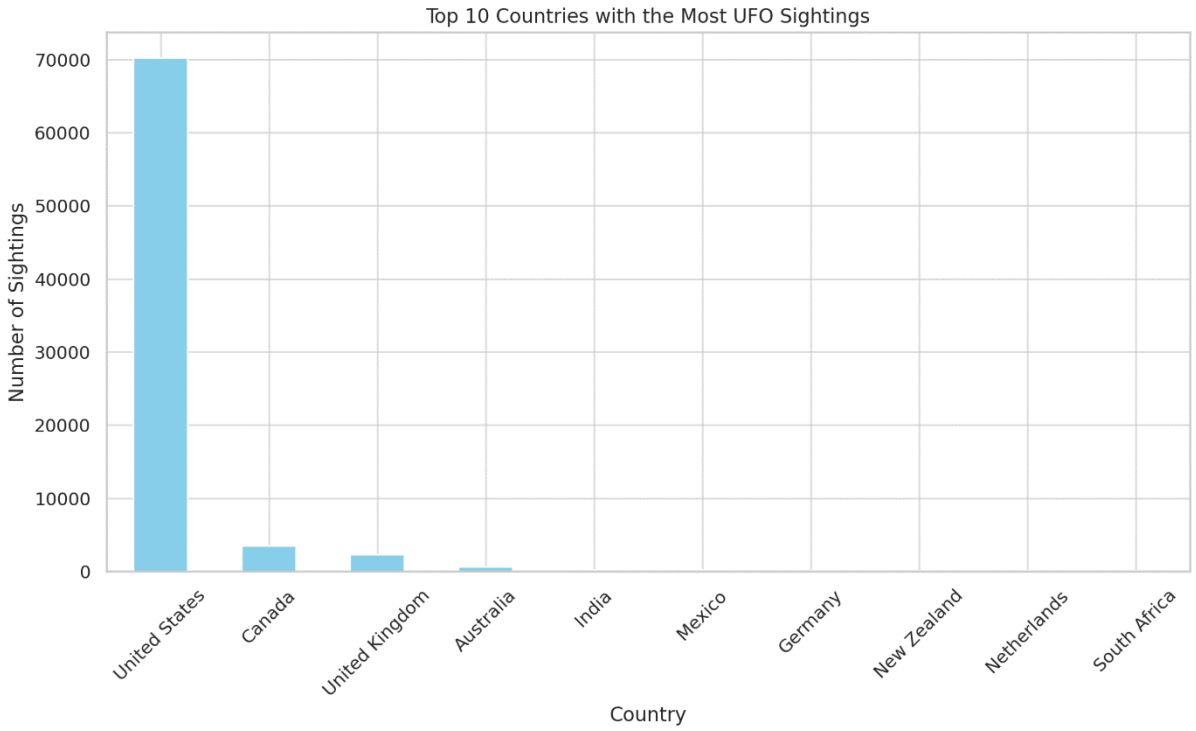


Figure 5. Barchart of UFO sightings and Country

Methods

In this section we try to analyse the analytic questions that need to be examined to get an overview of the UFO sightings and its factors.

For further exploration and analysis of the given dataset in a comprehensive manner, we used the following methods of python which belong to different packages.

1. For reading and finding information about dataset
    - read_csv() : This method from pandas library used to read a CSV file (e.g., ufo-sightings-1.csv) and create a dataframe named 'ufo_dataset' in our case.
    - head() : This method is used to display the first few rows of the dataframe, providing a quick overview of the data structure and content.
    - describe() : This method generates descriptive statistics of the numeric columns in the DataFrame, such as count, mean, standard deviation, minimum, and maximum. This gives a statistical summary of the dataset.
    - info() : This method provides a concise summary of the DataFrame, including the number of non-null values and data types of each column. It's useful for understanding the completeness of the dataset and the data types.

2. For filter the dataset to deal with missing values and imputing those data

- isna() : This method returns a DataFrame of the same shape as ufo_dataset with boolean values, where True indicates missing values (NaN) and False indicates non-missing values.
- sum() : This method then sums up the boolean values along each column, resulting in a Series containing the count of missing values for each column.
- copy() : Creates a copy of the original dataset (ufo_dataset) named imputed_ufo_dataset. This is done to avoid modifying the original dataset during the imputation process.
- fillna() : For each numeric column, it fills missing values (NaN) with the median of that column using this method. For each categorical column, it fills missing values (NaN) with the mode (most frequently occurring value) of that column using this method.

3. For plot encounters in a world map by using the latitude and longitude
    - scatter_geo: This function creates a scatter plot on a geographical map. It takes the DataFrame (imputed_ufo_dataset) as the data source.
    - lat='latitude', lon='longitude': These parameters specify which columns in the DataFrame should be used as latitude and longitude coordinates for plotting points on the map.
    - hover_name='Locale': This parameter specifies that the 'Locale' column should be used as the hover information when you hover over each data point on the map.
    - symbol='UFO_shape': This parameter assigns different symbols to different UFO shapes, making it easier to distinguish between them on the map.
    - update_geos: This method is used to update the map projection type. In this case, it sets the projection type to "natural earth".
    - update_layout: This method is used to update the layout of the entire plot. It sets the title of the plot to 'UFO sightings worldwide'.
    - show(): This function displays the interactive plot. When we run this code, an interactive map will be generated, allowing us to zoom in, hover over data points for details, and interact with the plot.

4. For finding longest encounters per country and longest encounters overall
    - to_numeric(): This function is used to convert the 'length_of_encounter_seconds' column to numeric values. The errors='coerce' parameter is used to replace any non-numeric values with NaN (Not a Number).
    - groupby('Country'): This method groups the DataFrame by the 'Country' column.
    - idxmax(): This finds the index of the row with the maximum within each group (country). In this case the row is 'length_of_encounter_seconds'.
    - loc[]: This is used to select and filter rows based on the indices obtained from idxmax().

5. For finding the trend between UFO shape and year
    - Grouping Data:
        - ufo_shape_trends: We are grouping the original dataset (imputed_ufo_dataset) by 'Year' and 'UFO_shape' using the groupby method. The size() function is then used to count the occurrences within each group, and reset_index is used to convert the result into a DataFrame with the columns 'Year', 'UFO_shape', and 'Counts'.
    - Creating an Interactive Line Plot:

- px.line(): This function creates an interactive line plot. We specify the DataFrame (ufo_shape_trends) and the variables for the x-axis ('Year'), y-axis ('Counts'), and color ('UFO_shape'). Labels and titles are also provided for better visualisation.

- Updating Traces (Lines and Markers):
  - update_traces(): This method is used to update the properties of the traces in the plot. Here, we set the mode to 'lines+markers' to display both lines and markers in the plot.
  - hovertemplate: This attribute sets the content of the hover box that appears when we hover over data points. Here, it displays the count (%{y}) and the year (%{x}) in bold.
- Customising Layout:
  - update_layout(): This method is used to customise the layout of the plot. We set the titles for the x-axis, y-axis, and legend.
- Showing the Plot:
  - show(): Finally, this function is used to display the interactive plot.

6. To find the best season to spot UFO and differentiate the result between geographically

- pycountry Library:
  - import pycountry: Imports the pycountry library, which provides information about countries, including ISO country codes.
- pycountry_convert Library:
  - import pycountry_convert as pc: Imports the pycountry_convert library, an extension of pycountry that includes additional functionality, such as converting between different country-related codes.
- get_continent_from_country Function:
  - Uses the pycountry library to get the country details (including the alpha-2 country code) from the country name.
  - Uses the pycountry_convert library to convert the alpha-2 country code to the continent code.
  - Returns the continent code.
- classify_region Function:
  - Takes the continent code as input and classifies it into 'Americas', 'Europe', or 'Other' based on certain conditions.
- Applying Functions to Create New Columns:
  - Uses the apply method to apply the get_continent_from_country and classify_region functions to the 'Country' column and create new columns 'Continent' and 'Region'.
- get_season Function:
  - This function takes a month as input and returns the corresponding season based on the Northern Hemisphere. It divides the months into Spring (March to May), Summer (June to August), Autumn (September to November), and Winter (December to February).
- Applying the Function to Create a Season Column:
  - The apply method is used to apply the get_season function to the 'Month' column of the dataset, creating a new column named 'Season' that represents the season for each sighting.
- Grouping and Counting Sightings:

- The groupby method is used to group the dataset by both 'Region' and 'Season'.
- The size method is applied to get the count of sightings in each group.
- The reset_index method is used to convert the result into a DataFrame and rename the count column to 'Count'.
- The resulting DataFrame, seasonal_sightings, contains the count of sightings for each combination of region and season.

- Creating a Bar Plot:
  - The seaborn library is used to create a bar plot (sns.barplot) with the 'Season' on the x-axis, 'Count' on the y-axis, and different colors for each 'Region'.
  - The hue parameter is set to 'Region' to differentiate the bars based on regions.
- Adding Titles and Labels:
  - Titles and labels are added to the plot using plt.title, plt.xlabel, plt.ylabel, and plt.legend to provide context and clarity to the viewer.
- Displaying the Plot:
  - The plt.show() method is used to display the finalized bar plot.

7. To perform Hypothesis Testing
   - Datetime Conversion:
     - The 'Date_time' column is converted to datetime format using pd.to_datetime. The errors='coerce' parameter is used to handle any errors by converting them to NaT (Not a Time).
   - Hour Extraction:
     - The 'Hour' column is created by extracting the hour from the 'Date_time' using ufo_dataset['Date_time'].dt.hour.
   - Defining Day and Night:
     - A new column 'Time_of_day' is created and initialized with 'Day'. For entries where the hour is between 20 and 6 (inclusive), it is set to 'Night'.
   - Counting UFO Sightings:
     - The value_counts method is used to count the occurrences of each unique value in the 'Time_of_day' column, providing the number of UFO sightings during the day and night.
   - Chi-Squared Test:
     - The chi2_contingency function from the scipy.stats module is used to perform a chi-squared test of independence on the observed counts of UFO sightings during the day and night against the expected counts.
   - Interpreting the p-value:
     - The obtained p-value is compared to a significance level (commonly 0.05).
     - If the p-value is less than the significance level, the null hypothesis is rejected, indicating a significant difference in UFO sightings between day and night.
     - If the p-value is greater than or equal to the significance level, the null hypothesis is not rejected, suggesting no significant difference.

Results and Analysis.

For the above-mentioned methods to solve the analytic questions, this section provides the results, output and visualizations to help understand the dataset.
The Key findings for questions are:

3) Plot the encounters in a world map using Plotly

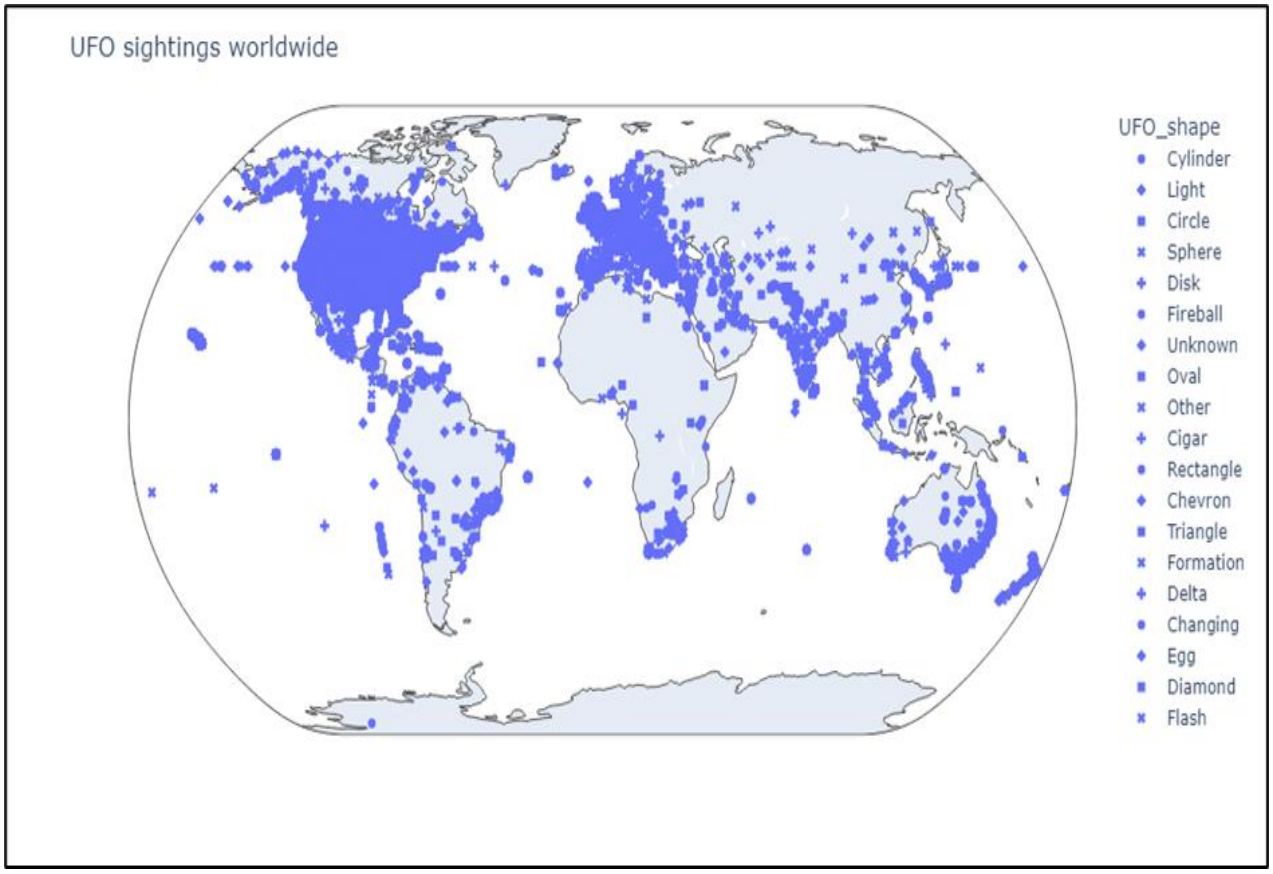


Figure 6: Global Distribution of UFO Sightings with Shape Categorisation.

Figure 1 highlights the locations of UFO sightings categorised by the shape of the UFO's. Each shape has a corresponding icon to represent it as indicated by the legend on the right-hand side of the map. Figure 1 indicates that sightings are most frequent in North America and Europe compared to the other continents. This aligns with the earlier findings in EDA as the United states, Canada, and United Kingdom are the countries with the largest number of sightings.

4) Investigation on which country had the longest encounter and the shape of that UFO Summary showing longest encounter in a country with the identified shape of it.

```
Longest Encounters per Country:
           Country  length_of_encounter_seconds UFO_shape
13690  Afghanistan                       1200.0     Light
38241      Albania                        120.0     Light
6185       Algeria                         60.0     Light
60728    Argentina                      86400.0      Oval
34882      Armenia                        600.0    Circle
...            ...                          ...       ...
45627   Uzbekistan                       6900.0    Sphere
78059    Venezuela                       3600.0     Light
63215      Vietnam                      10800.0    Sphere
45784       Zambia                        240.0  Fireball
68365     Zimbabwe                       2700.0     Light

[151 rows x 3 columns]

Country with the Longest Encounter:
Country                          United Kingdom
length_of_encounter_seconds        97836000.0
UFO_shape                              Sphere
Name: 559, dtype: object
```

Figure 7: Summary showing longest encounter in a country with the identified shape.

From Figure 2, we can say that the longest encounter was seen in United Kingdom which lasted for around 978,260,000 seconds where the shape of the UFO was spherical. Canada followed with the encounter that lasted for about 82,800,000 seconds and the UFO shape was described as Other. Then followed United states with an encounter that lasted for about 66,276,000seconds with a Sphere UFO shape.

5) Investigation of the trend between UFO shapes and year

Investigation of the trend between UFO shapes and year reveals: firstly, we notice a sharp increase in the number of sightings from around 1990s to 2010s generally. The UFOs shape Light has the highest number of sightings from the 1990s to 2010s. The most sightings in this time period are noticed in categories of Light, Circle, Flash and Fireball. Before the 1990s, the most sighted shapes were Disk and Light and then we also see emergence of new shapes from the 1960s with the peak being in 2000s.

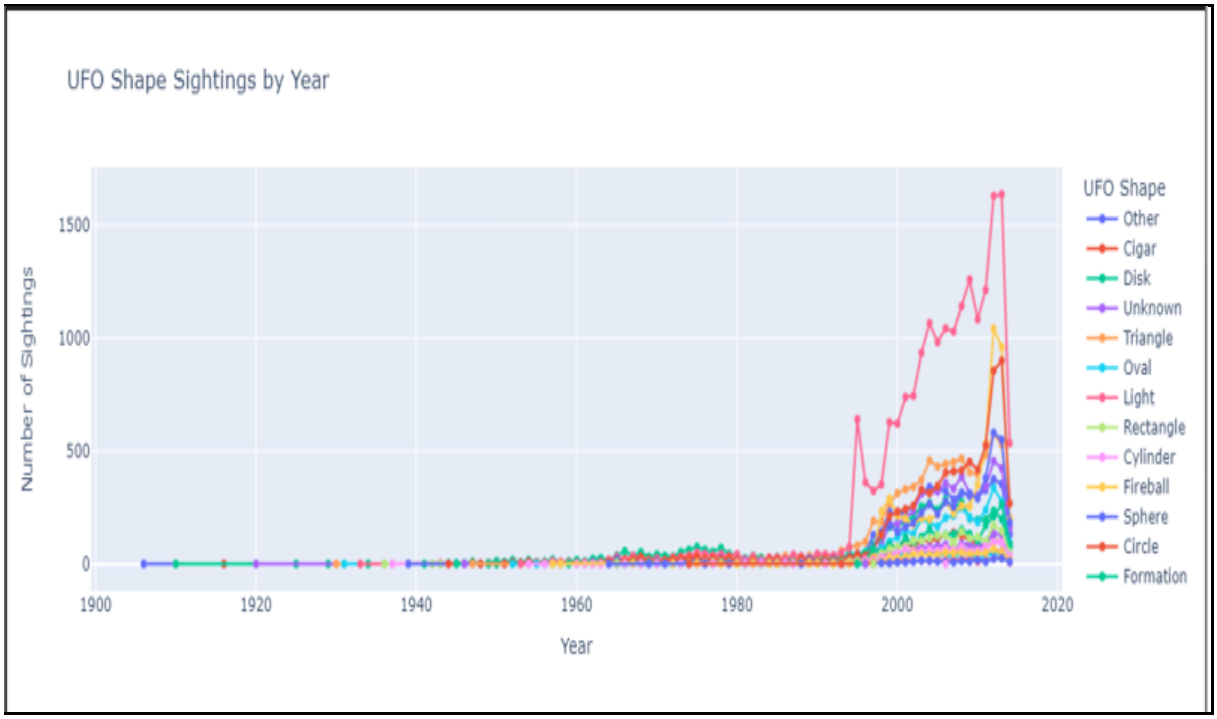
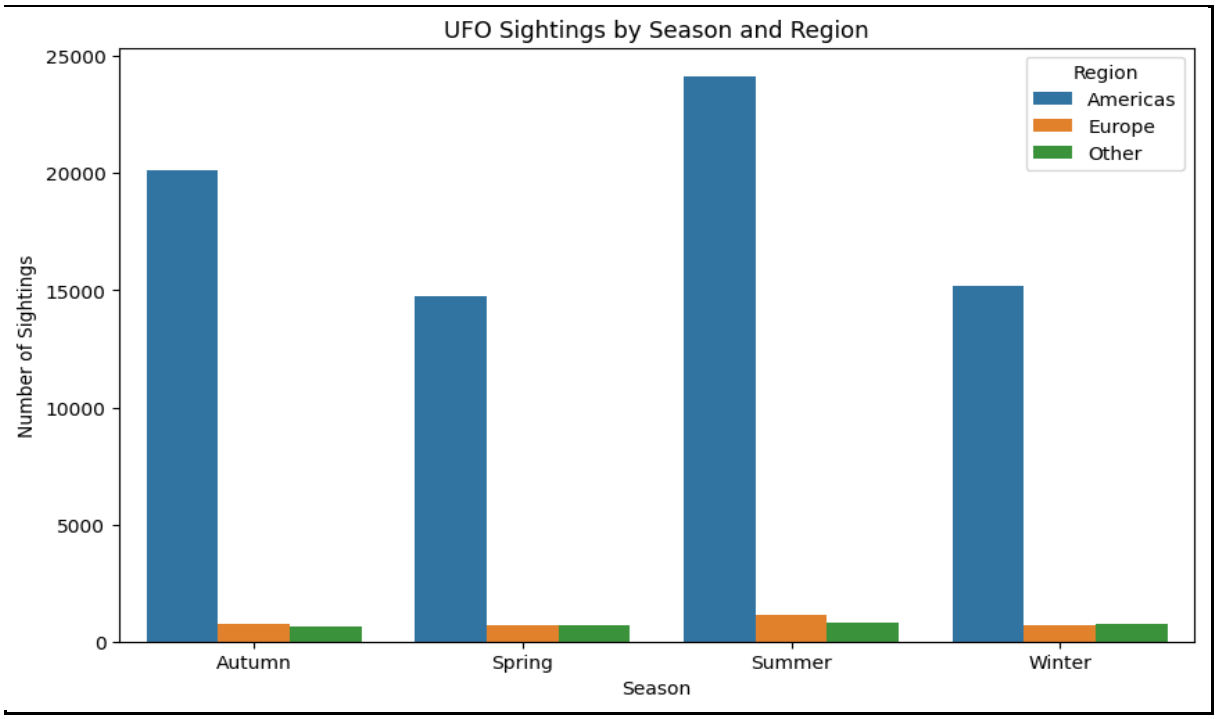


Figure 8: Plot showing the relationship between UFO shapes and evolution of the years

6. Relationship between UFO sightings and seasons in Americas and Europe



Graph 9: A bar chart of UFO sightings by season

Analysis between UFO sightings and seasons reveals that generally the sightings are higher in Spring and Autumn and are least in spring . Both Americas and Europe show the number of sightings being highest in Summer, followed by Autumn and Winter and lastly Spring. Americas has the highest count

of sightings in each season compared to Europe. The analysis suggests that if one is looking to see an UFO, summer might be the right season and also its mostly likely in the Americas.

Summarising from section 7: we statistically prove that there was a significant difference in the frequency of the UFO sightings between the day and night time. After performing chi-square test we stated that UFO sightings were seen mostly during the Night time. We could use this information to see if we can observe that, is there any common shapes during night or in some different regions of the world in both (day/ night) time frequencies.

Hypothesis Testing

Hypothesis is a tentative explanation or predictions of a proposed statement which can be tested either visualising the given dataset or by formulating a declarative statement and examining the statement.

To formulate a hypothesis, data exploration is the key. Understanding the relations between the given attributes and analysing patterns if any present leads us to form a declarative statement which then can be tested for its integrity.

Exploring the given UFO dataset, many questions like the duration of the UFO, which country has the majority of UFO sightings, in what seasons the UFO are familiar to be seen, etc and all these are already covered by above scenarios and explained with visualization.

But there is one investigation that concerns us and which is not covered in the above sections are the UFO sightings appearing during what time of the given day.

To explore more on this, let's first create a visualization and see if there are any patterns that help us understand the above statement.

Hypothesis:

There are significantly more UFO sightings during the night (defined as 8 PM to 6 AM) than during the day (6 AM to 8 PM).

The next step is to test the above hypothesis using a statistics method and verify our statement.

The below will be our statements to test our hypothesis using statistics:

Null Hypothesis H_0: There is no significant difference in UFO sightings between day and night.

Alternate Hypothesis H_1: There is a significant difference in UFO sightings between day and night.

The statistical approach **Chi-Squared Test for Independence** test is used to validate the above statements.

Since we have categorised the 'Day' and 'Night' duration from hours and want to verify the observed frequency against the expected frequency of the hypothesis statements, **Chi-Squared Test for**

Independence is the best fit model to use since the association or dependency of two frequencies needs to be verified.

To perform the Chi-squared test we need to import chi2_contingency function from the scipy.stats module to compare the two frequencies.

Since we have the observed value of 55036 and 25292 sightings during 'Nights' and 'Day' respectively. The summation of both above values are the total observed value.

The expected value for each case is exactly half of the total value i.e. expected_night = total_sightings / 2 and expected_day = total_sightings / 2.

Now, we can perform Chi-squared test on observed and expected count to get the p-value.

The Chi-Squared Test p-value: 0.0.

Getting a p-value 0 often indicates that the p-value is extremely small, typically approaching zero but not precisely zero. This extremely small p-value suggests strong evidence against the null hypothesis.

By above evidence we can reject null hypothesis and state that "there is significant difference in UFO sightings between day and night"

Hence, we conclude that the p-value we got would suggest that UFO sightings are more frequent at night compared to during the day form both visualization and Hypothesis testing

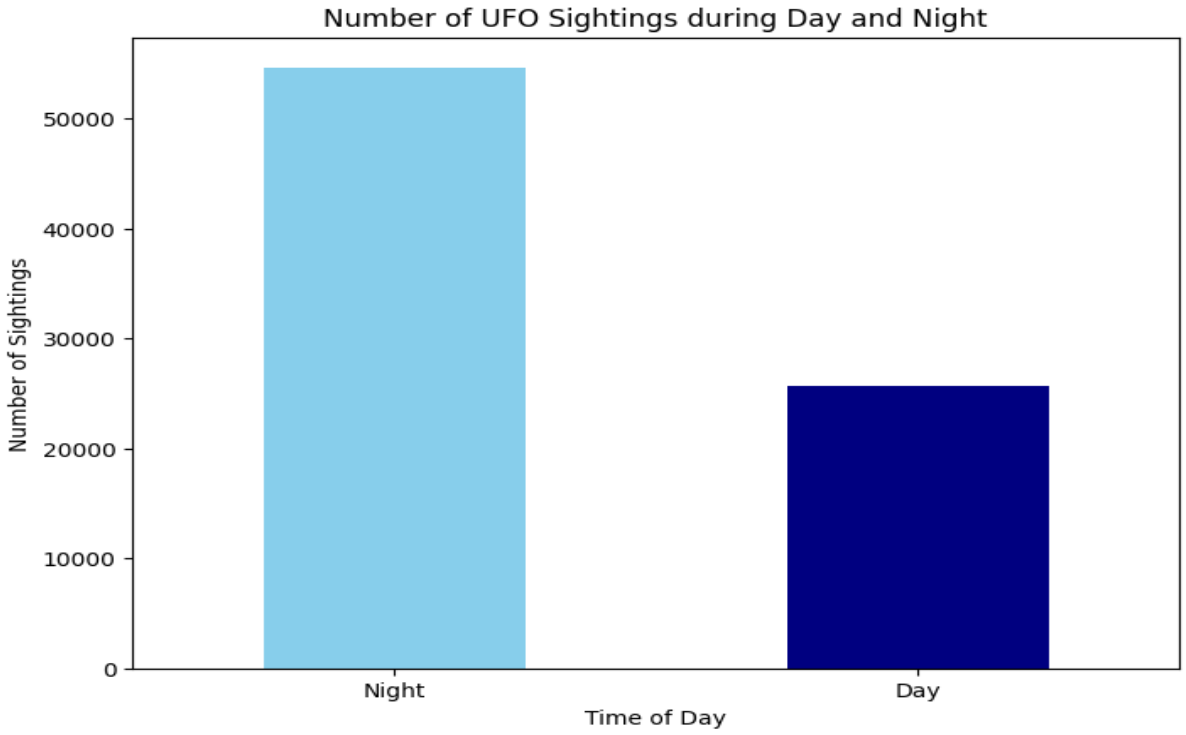


Figure 10. Barchat of Time of day and UFO sightings.

From the above barplot we observe that the UFO sightings appear is significantly more during the night time of the day then day time which verify our hypothesis.

Discussions

The analysis of the UFO sightings dataset has uncovered some interesting patterns. We examined where these sightings took place, how long they lasted, and also observed how the shapes of the UFO changed over time. The global map of sightings illustrated the locations where UFOs were spotted.

After analyzing the data, most UFO sightings happen in North America and Europe and we found that the United Kingdom had the longest encounter with a UFO, lasting around 978,360,000 seconds, with the UFO being spherical in shape. Speaking about UFO shapes, there were very few observed from the 1990s to the 1960s. However, in the mid-1960s, new shapes emerged with more sightings, reaching a peak in the year 2000. But the question here is, what factors could have contributed to these UFOs being seen more frequently?

We then looked into which seasons had the most UFO sightings and discovered that spring and autumn were the seasons with the greatest activity. Moreover, America outperforms Europe in all seasons. The Summer season is where we see a lot of potential in UFO visibility. This information can help us study different environmental factors that would have caused to have such UFO appearances.

The data on UFO sightings reveals that UFOs are more common at night compared to the day, adding a solid foundation to our common belief. The hypothesis test confirms that UFOs are more frequently seen during the night time from 8 pm to 6 am.

Conclusion

To conclude, most UFO sightings happen in North America and Europe, with the United Kingdom reporting the longest one. We noticed more types of UFOs being seen starting from the 1960s, and there were a lot more sightings from the 1990s. We also found that you're more likely to see a UFO in the summer, especially in America. Lastly, our study proved that UFOs are seen more at night. This comprehensive analysis of UFO sightings has shed light on various aspects, from time-based patterns to global preferences. These findings add details about UFO myths and conversations. By using figures and graphs to illustrate our insights, we are making a contribution towards understanding UFO sightings, sparking more curiosity for further exploration.