**EDA On UFO Sightings**

**UFO**

An **unidentified flying object** (**UFO**), or **unidentified anomalous phenomenon** (**UAP**), is any perceived aerial phenomenon that cannot be immediately identified or explained. Upon investigation, most UFOs are [identified](https://en.wikipedia.org/wiki/Identification_studies_of_UFOs) as known objects or atmospheric phenomena, while a small number remain unexplained.

While unusual sightings have been reported in the sky throughout history, UFOs became culturally prominent after [World War II](https://en.wikipedia.org/wiki/World_War_II), escalating during the [Space Age](https://en.wikipedia.org/wiki/Space_Age). Studies and investigations into UFO reports conducted by governments (such as [Project Blue Book](https://en.wikipedia.org/wiki/Project_Blue_Book) in the United States and [Project Condign](https://en.wikipedia.org/wiki/Project_Condign) in the United Kingdom), as well as by organisations and individuals have occurred over the years without confirmation of the fantastical claims of believers. The [U.S. government](https://en.wikipedia.org/wiki/Federal_government_of_the_United_States) currently has two entities dedicated to UAP (or UFO) data collection and analysis: [NASA's UAP independent study team](https://en.wikipedia.org/wiki/NASA%27s_UAP_independent_study_team) and the [All-domain Anomaly Resolution Office](https://en.wikipedia.org/wiki/All-domain_Anomaly_Resolution_Office).

Scientists and sceptical organisations such as the [Committee for Sceptical Inquiry](https://en.wikipedia.org/wiki/Committee_for_Skeptical_Inquiry) have provided prosaic explanations for UFOs, namely that they are caused by natural phenomena, human technology, delusions, and hoaxes. Small but vocal groups of [ufologists](https://en.wikipedia.org/wiki/Ufologists) favour unconventional or [pseudoscientific](https://en.wikipedia.org/wiki/Pseudoscience) hypotheses, often claiming that UFOs are evidence of [extraterrestrial intelligence](https://en.wikipedia.org/wiki/Extraterrestrial_hypothesis), [technologically advanced cryptids](https://en.wikipedia.org/wiki/Cryptoterrestrial_hypothesis), [interdimensional contact](https://en.wikipedia.org/wiki/Interdimensional_UFO_hypothesis) or [future time travellers](https://en.wikipedia.org/wiki/Time-traveller_UFO_hypothesis), but even after decades of promotion of such ideas by believers and in popular media, [the kind of evidence required to solidly support such claims](https://en.wikipedia.org/wiki/Sagan_standard) has not been forthcoming. Beliefs surrounding UFOs have inspired parts of [new religions](https://en.wikipedia.org/wiki/UFO_religion) even as social scientists have identified the ongoing interest and storytelling surrounding UFOs as a modern example of [folklore](https://en.wikipedia.org/wiki/Folklore) and [mythology](https://en.wikipedia.org/wiki/Mythology) understandable with [psychosocial explanations](https://en.wikipedia.org/wiki/Psychosocial_UFO_hypothesis).

**EDA**

## 

Exploratory Data Analysis (EDA) refers to the method of studying and exploring record sets to apprehend their predominant traits, discover patterns, locate outliers, and identify relationships between variables. EDA is normally carried out as a preliminary step before undertaking extra formal statistical analyses or modelling.

### The Foremost Goals of EDA

**1. Data Cleaning**: EDA involves examining the information for errors, lacking values, and inconsistencies. It includes techniques including records imputation, managing missing statistics, and figuring out and getting rid of outliers.

**2. Descriptive Statistics**: EDA utilises precise records to recognize the important tendency, variability, and distribution of variables. Measures like suggest, median, mode, preferred deviation, range, and percentiles are usually used.

**3. Data Visualization**: EDA employs visual techniques to represent the statistics graphically. Visualisations consisting of histograms, box plots, scatter plots, line plots, heatmaps, and bar charts assist in identifying styles, trends, and relationships within the facts.

**4. Feature Engineering**: EDA allows for the exploration of various variables and their adjustments to create new functions or derive meaningful insights. Feature engineering can contain scaling, normalisation, binning, encoding express variables, and creating interplay or derived variables.

**5. Correlation and Relationships**: EDA allows discovering relationships and dependencies between variables. Techniques such as correlation analysis, scatter plots, and pass-tabulations offer insights into the power and direction of relationships between variables.

**6. Data Segmentation**: EDA can contain dividing the information into significant segments based totally on sure standards or traits. This segmentation allows advantage insights into unique subgroups inside the information and might cause extra focused analysis.

**7. Hypothesis Generation**: EDA aids in generating hypotheses or studies questions based totally on the preliminary exploration of the data. It facilitates form the inspiration for in addition evaluation and model building.

**8. Data Quality Assessment**: EDA permits for assessing the niceness and reliability of the information. It involves checking for records integrity, consistency, and accuracy to make certain the information is suitable for analysis.

### **Types of EDA**

Depending on the number of columns we are analysing we can divide EDA into two types.

EDA, or Exploratory Data Analysis, refers back to the method of analysing and analysing information units to uncover styles, pick out relationships, and gain insights. There are various sorts of EDA strategies that can be hired relying on the nature of the records and the desires of the evaluation. Here are some not unusual kinds of EDA:

**1. Univariate Analysis**: This sort of evaluation makes a speciality of analysing character variables inside the records set. It involves summarising and visualising an unmarried variable at a time to understand its distribution, relevant tendency, unfold, and different applicable records. Techniques like histograms, field plots, bar charts, and precise information are generally used in univariate analysis.

**2. Bivariate Analysis**: Bivariate evaluation involves exploring the connection between variables. It enables finding associations, correlations, and dependencies between pairs of variables. Scatter plots, line plots, correlation matrices, and move-tabulation are generally used strategies in bivariate analysis.

**3. Multivariate Analysis**: Multivariate analysis extends bivariate evaluation to encompass greater than variables. It aims to apprehend the complex interactions and dependencies among more than one variable in a records set. Techniques inclusive of heatmaps, parallel coordinates, aspect analysis, and primary component analysis (PCA) are used for multivariate analysis.

**4. Time Series Analysis**: This type of analysis is mainly applied to statistics sets that have a temporal component. Time collection evaluation entails inspecting and modelling styles, traits, and seasonality inside the statistics through the years. Techniques like line plots, autocorrelation analysis, transferring averages, and ARIMA (AutoRegressive Integrated Moving Average) fashions are generally utilised in time series analysis.

**5. Missing Data Analysis**: Missing information is a not unusual issue in datasets, and it may impact the reliability and validity of the evaluation. Missing statistics analysis includes figuring out missing values, know-how the patterns of missingness, and using suitable techniques to deal with missing data. Techniques along with lacking facts styles, imputation strategies, and sensitivity evaluation are employed in lacking facts evaluation.

**6. Outlier Analysis**: Outliers are statistics factors that drastically deviate from the general sample of the facts. Outlier analysis includes identifying and knowing the presence of outliers, their capability reasons, and their impact at the analysis. Techniques along with box plots, scatter plots, z-rankings, and clustering algorithms are used for outlier evaluation.

**7. Data Visualization**: Data visualisation is a critical factor of EDA that entails creating visible representations of the statistics to facilitate understanding and exploration. Various visualisation techniques, inclusive of bar charts, histograms, scatter plots, line plots, heatmaps, and interactive dashboards, are used to represent exclusive kinds of statistics**.**

**Kaggle**

Kaggle is a [data science competition platform](https://en.wikipedia.org/wiki/Data_science_competition_platform) and online community of [data scientists](https://en.wikipedia.org/wiki/Data_science) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning) practitioners under [Google LLC](https://en.wikipedia.org/wiki/Google_LLC). Kaggle enables users to find and publish datasets, explore and build models in a web-based data science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges

UFO dataset is obtained from the kaggle website

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

**UFO dataset is analysed to find where and when the sightings are occurred**

**Tech Stack**

Different libraries of python are used:

**Numpy**

NumPy is a [library](https://en.wikipedia.org/wiki/Library_(computing)) for the [Python programming language](https://en.wikipedia.org/wiki/Python_(programming_language)), adding support for large, multi-dimensional [arrays](https://en.wikipedia.org/wiki/Array_data_structure) and [matrices](https://en.wikipedia.org/wiki/Matrix_(mathematics)), along with a large collection of [high-level](https://en.wikipedia.org/wiki/High-level_programming_language) [mathematical](https://en.wikipedia.org/wiki/Mathematics) [functions](https://en.wikipedia.org/wiki/Function_(mathematics)) to operate on these arrays. The predecessor of NumPy, Numeric, was originally created by [Jim Hugunin](https://en.wikipedia.org/wiki/Jim_Hugunin) with contributions from several other developers. In 2005, [Travis Oliphant](https://en.wikipedia.org/wiki/Travis_Oliphant) created NumPy by incorporating features of the competing Numarray into Numeric, with extensive modifications. NumPy is [open-source software](https://en.wikipedia.org/wiki/Open-source_software) and has many contributors. NumPy is a NumFOCUS fiscally sponsored project.

**Pandas**

Pandas (stylized as pandas) is a [software library](https://en.wikipedia.org/wiki/Software_library) written for the [Python programming language](https://en.wikipedia.org/wiki/Python_(programming_language)) for data manipulation and [analysis](https://en.wikipedia.org/wiki/Data_analysis). In particular, it offers [data structures](https://en.wikipedia.org/wiki/Data_structure) and operations for manipulating numerical tables and [time series](https://en.wikipedia.org/wiki/Time_series). It is [free software](https://en.wikipedia.org/wiki/Free_software) released under the [three-clause BSD licence](https://en.wikipedia.org/wiki/3-clause_BSD_license). The name is derived from the term "[panel data](https://en.wikipedia.org/wiki/Panel_data)", an [econometrics](https://en.wikipedia.org/wiki/Econometrics) term for [data sets](https://en.wikipedia.org/wiki/Data_set) that include observations over multiple time periods for the same individuals, as well as a play on the phrase "Python data analysis". [Wes McKinney](https://en.wikipedia.org/wiki/Wes_McKinney) started building what would become Pandas at [AQR Capital](https://en.wikipedia.org/wiki/AQR_Capital) while he was a researcher there from 2007 to 2010

**Matplotlib**

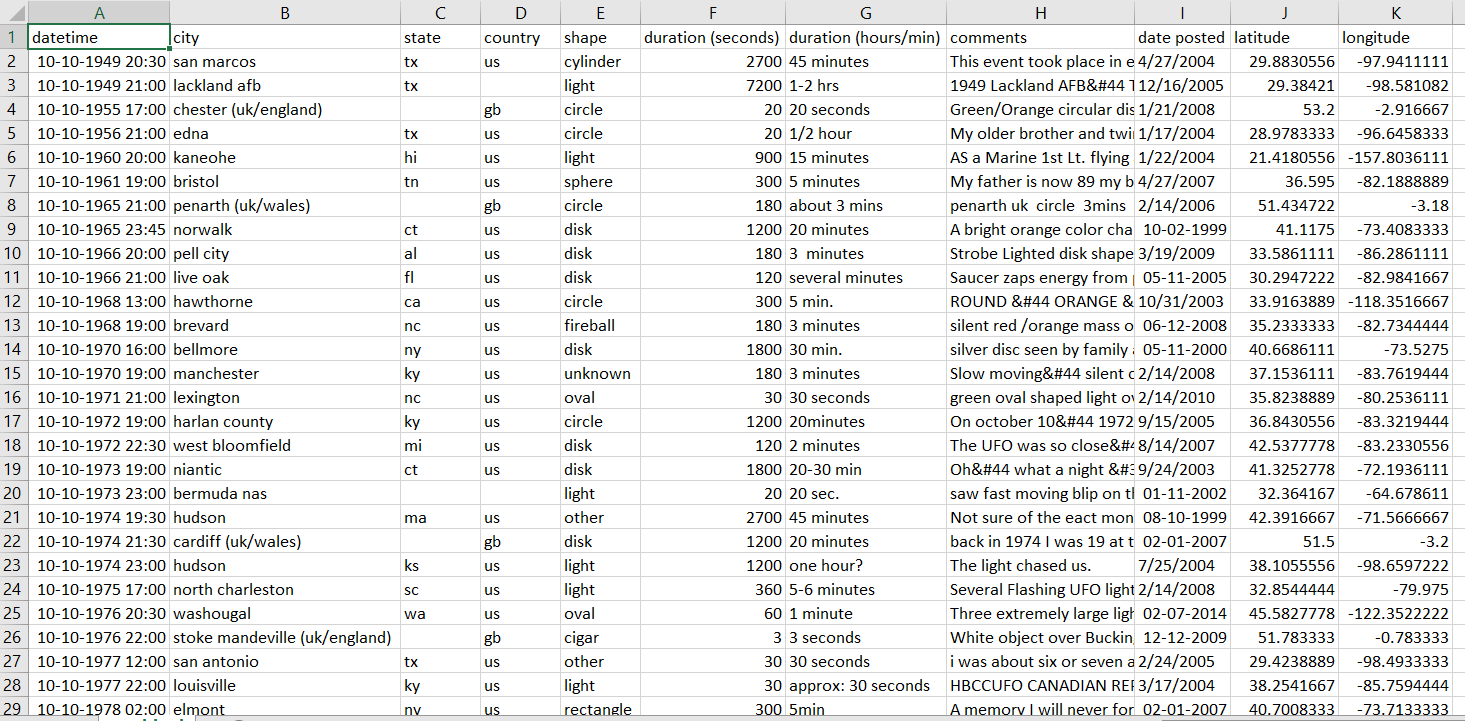
Matplotlib is a [plotting](https://en.wikipedia.org/wiki/Plotter) [library](https://en.wikipedia.org/wiki/Library_(computer_science)) for the [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) [programming language](https://en.wikipedia.org/wiki/Programming_language) and its [numerical mathematics](https://en.wikipedia.org/wiki/Numerical_analysis) extension [NumPy](https://en.wikipedia.org/wiki/NumPy). It provides an [object-oriented](https://en.wikipedia.org/wiki/Object-oriented_programming) [API](https://en.wikipedia.org/wiki/API) for embedding plots into applications using general-purpose [GUI toolkits](https://en.wikipedia.org/wiki/GUI_toolkit) like [Tkinter](https://en.wikipedia.org/wiki/Tkinter), [wxPython](https://en.wikipedia.org/wiki/WxPython), [Qt](https://en.wikipedia.org/wiki/Qt_(software)), or [GTK](https://en.wikipedia.org/wiki/GTK). There is also a [procedural](https://en.wikipedia.org/wiki/Procedural_programming) "pylab" interface based on a [state machine](https://en.wikipedia.org/wiki/State_machine) (like [OpenGL](https://en.wikipedia.org/wiki/OpenGL)), designed to closely resemble that of [MATLAB](https://en.wikipedia.org/wiki/MATLAB), though its use is discouraged[]](https://en.wikipedia.org/wiki/Matplotlib#cite_note-3) [SciPy](https://en.wikipedia.org/wiki/SciPy) makes use of Matplotlib.

**Seaborn**

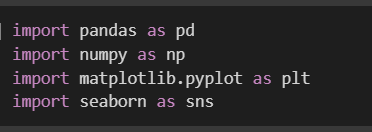
Seaborn is a Python data visualisation library based on [matplotlib](https://matplotlib.org/). It provides a high-level interface for drawing attractive and informative statistical graphics.

**Dataset**

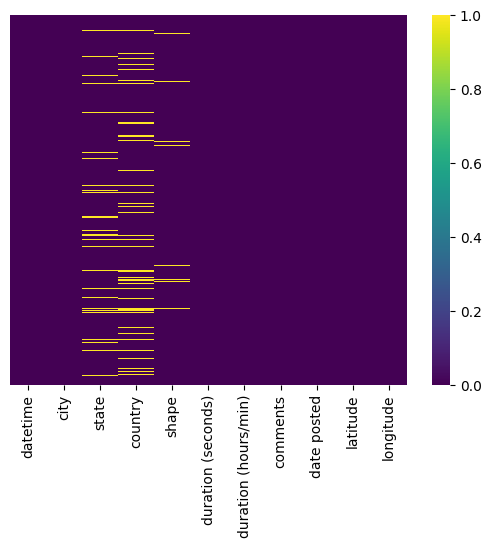
Sample of dataset is given below.Details about date on which UFO sighted, City ,State,Country where it is founded is included. ALso duration, latitude,longitude etc also given in the CSV file



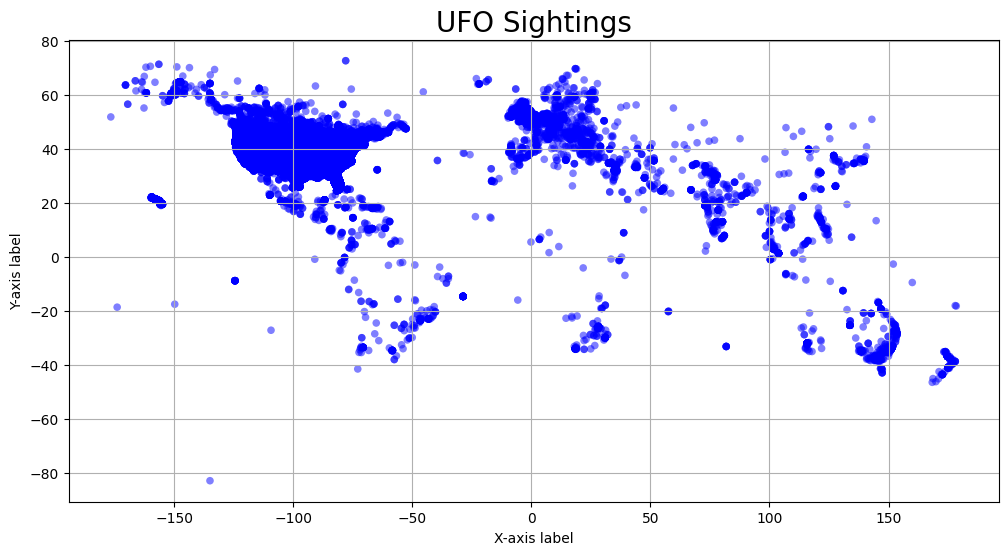
1.First required libraries are imported



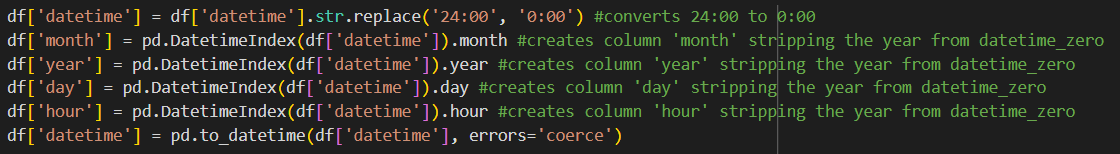
2 Missing values are visualised using heatmap. A few values are missing in state ,country and shape



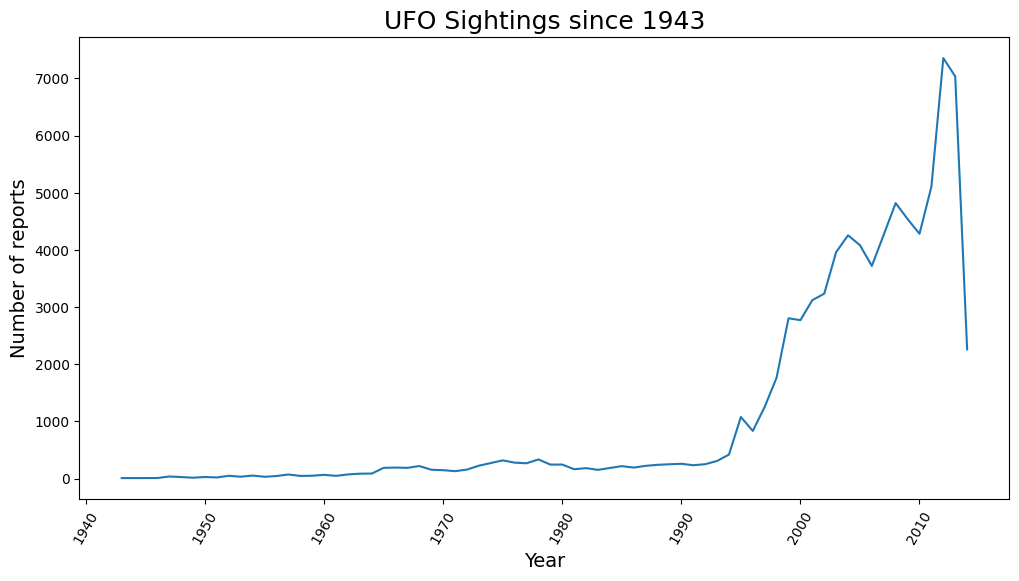
3. Scatter plot is made to show the sightings in latitude and longitude



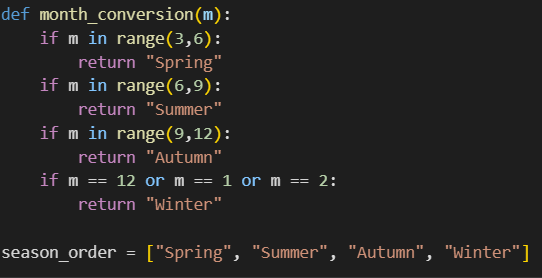
4.From the single column “datetime” different columns are made for date,time,year etc



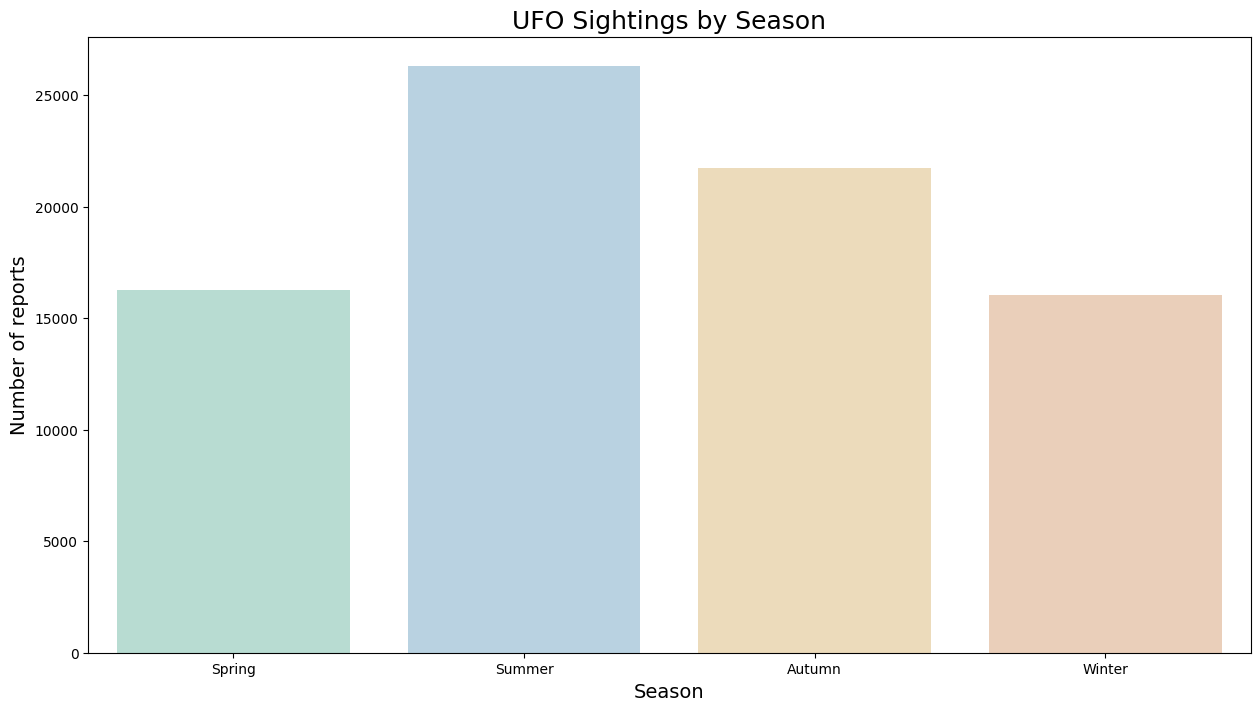
5. Line Plot is made between years and Number of sightings: A remarkable increase in sightings after 1990 is observed.



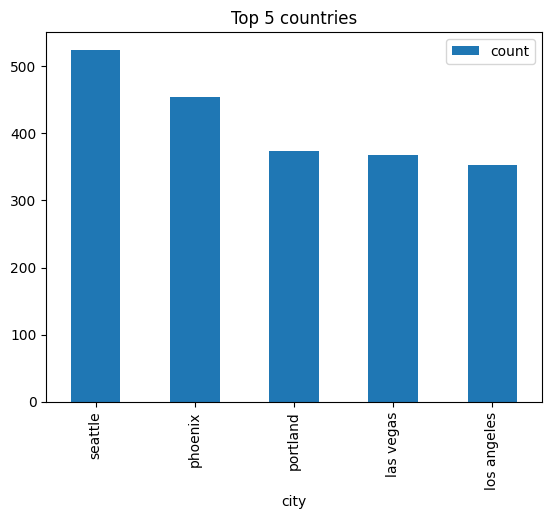
6. Months are divided to seasons



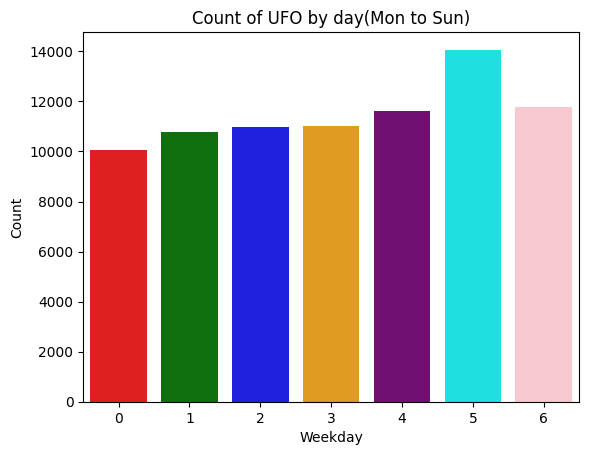
7. Bar Plot is plotted between number of sightings and season. It is observed that summer season have more number of sightings



8. Top 5 cities where UFO sightings observed plotted.Seattle had the most



9. Bar Plot of count of sightings and day in a week plotted. It is observed that Saturday has the most number of sightings



**Project Link**

https://colab.research.google.com/drive/1LQTWD7P-rv37pBAEieHFF5rAt1TIzRtA?usp=sharing

<https://github.com/jisnaAnfil/scifor_jisna/blob/mini_project/ufoproject.py>

https://github.com/jisnaAnfil/scifor\_jisna/blob/mini\_project/EDA%20On%20UFO%20Sightings.pptx

**Inference:**

From the data it is observed that Seattle has the most number of UFO sightings . Moreover, UFO sightings are more often in the Summer season and on Saturdays. After 1990 there is a remarkable increase in the UFO sightings