**Analyzing UFO Sightings Dataset Using Python**

**Dataset** – NUFORC UFO Sightings (<https://www.kaggle.com/datasets/NUFORC/ufo-sightings>)

**Tools** – Python, PyCharm

**Author** – Isuri Perera Balasooriya

**Date** – 13th March 2025

**City** – Colombo, Sri Lanka

Contents

[Table of Figures 2](#_Toc192783295)

[1 Introduction 2](#_Toc192783296)

[1.1 Overview 2](#_Toc192783297)

[1.2 The Dataset 3](#_Toc192783298)

[1.3 Tools 3](#_Toc192783299)

[2 Exploratory Data Analysis of UFO Sightings Data 4](#_Toc192783300)

[2.1 Data Cleaning 4](#_Toc192783301)

[2.2 Analytical Queries and Data Visualization 8](#_Toc192783302)

[3 Key Findings 11](#_Toc192783303)

[3.1 Limitations of the Data 12](#_Toc192783304)

[3.2 Future Studies 12](#_Toc192783305)

[4 Conclusion 12](#_Toc192783306)

# Introduction

## Overview

Unidentified Flying Object (UFO) sightings have been reported across the world for decades, sparking curiosity and speculation. While many reports remain anecdotal, analyzing structured data can provide valuable insights into patterns and trends in these sightings.

This project aims to conduct an exploratory data analysis (EDA) of a UFO sightings dataset using Python and Pandas to examine the distribution, frequency, and characteristics of reported sightings. Key objectives include identifying temporal and geographical trends, analyzing the most commonly reported UFO shapes, and visualizing potential patterns. By leveraging data visualization techniques, this study seeks to provide a structured analysis of UFO sighting records, offering insights into potential correlations and anomalies within the dataset.

## The Dataset

The dataset used for this project is from Kaggle which can be found at <https://www.kaggle.com/datasets/NUFORC/ufo-sightings>. The dataset consists of 80,332 entries and includes the following fields,

* Datetime
* City
* State
* Country
* Shape
* duration (seconds)
* duration (hours/min)
* comments
* date posted
* latitude
* longitude

## Tools

The following tools were used to analyze the data,

* Python – for data preparation, analysis and visualization
* PyCharm

# Exploratory Data Analysis of UFO Sightings Data

The aim of this study is to use Python to analyze a large dataset and derive key insights and patterns in the data. Firstly, the environment for data analysis was setup with the following assumptions,

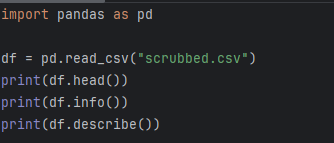
* Python is already installed in the system
* PyCharm is already installed in the system
* A GitHub repository is already created to store the code
* Git is already installed in the system
* The dataset is downloaded in csv format

Author has created a folder in GitHub which can be referred to get any files related to this project.

Link - <https://github.com/isuri-balasooriya2/TheMathLab/tree/main/UFO_Sightings>

## Data Cleaning

The dataset, which was downloaded in csv format, was loaded using pandas and using several pandas functions, the dataset was explored to identify which type of data is been used.

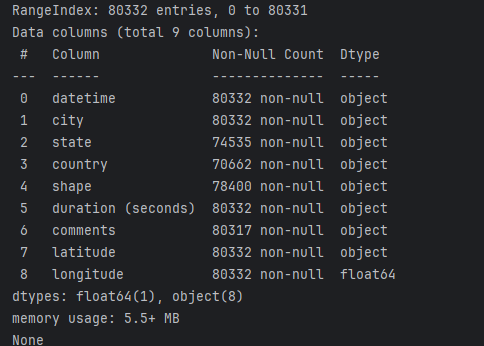


The first step in data cleaning is to identify if the data we are working with is valid and useful for the analysis task. By exploring the different columns in the dataset, it was identified that the columns “duration (hours/min)” and “date posted” are not required for the planned analysis task. Therefore, the above mentioned columns were dropped using the below code.

A computer screen with text

Description automatically generated

After deleting unwanted columns, the dataset looks as follows,



The next step in data cleaning is to identify missing values. Before handling missing data, its crucial to identify which columns contain missing values and how many.

A black background with white text

Description automatically generated

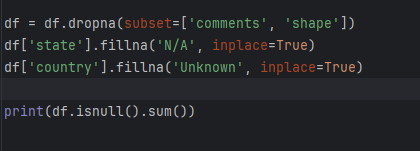
The above code gives the following output for the dataset that is been used.

A computer screen shot of a black screen

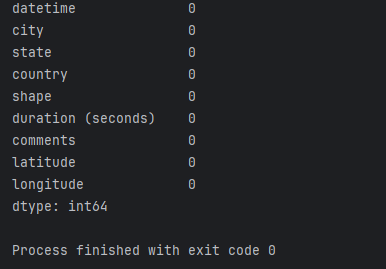
Description automatically generated

The specific method to handle missing values depends on the dataset and the purpose of the analysis. In this dataset which has 80,332 entries, removing the 9670 entries with missing country makes a bigger impact compared to removing 15 entries with missing comments value. In this case, the following was done to handle missing data,

* Remove the entries with missing comments
* Remove the entries with missing shape
* Replace missing state with N/A
* Replace missing country with Unknown



Once the above is done, the dataset was checked again for missing values and returned the following output.



From print(df.info()) it was identified that latitude and longitude are of different data types. To avoid confusions later, both columns were set to numeric.



A screenshot of a computer

Description automatically generated

The dataset has been cleaned and prepared as required for the exploratory data analysis.

## Analytical Queries and Data Visualization

* Trend of UFO sightings over the years

A computer screen with text on it

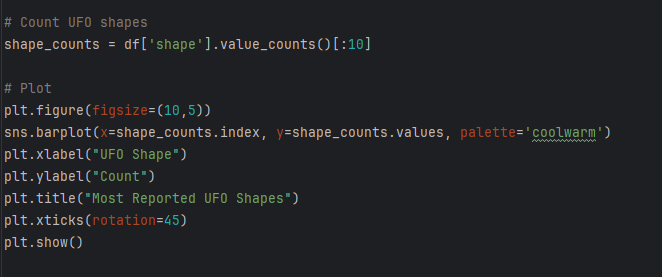
Description automatically generated

The datetime value was converted to proper datetime format and then the year was extracted and counted per year. A line plot was used to display the trend over the years.

A graph with a line

Description automatically generated

* The most common shapes of UFOs



the number of distinct UFO shapes were counted first and a bar chart was used to display the findings.

A graph of different colored bars

Description automatically generated

* Countries with most UFO sightings



A bar chart was used to display the top countries with most UFO sightings.

A graph of a number of people

Description automatically generated

* Most common time for UFO sightings

A computer screen shot of a program

Description automatically generated

A graph of a number of people

Description automatically generated with medium confidence

# Key Findings

Based on this exploratory data analysis, the following key insights were identified,

* The number of UFO sightings have increased since 1990, reaching highest number of sightings in 2013
* Compared to the recorded sightings until 1980, 1993-2014 shows a steady growth in the number of sightings
* The most reported UFO shape is light, triangle and circle, not disk shape as popularized in recent years
* USA is the country with most UFO sightings by a considerable margin. Canada comes seconds followed by Great Britain, Australia and Denmark
* The most popular time for UFO sightings is between 8PM-9PM
* Most UFO sightings seem to occur during night time

## Limitations of the Data

This analysis is based on publicly available UFO sighting data, which may contain inconsistencies, biases, and missing information. The dataset relies on self-reported sightings, making it susceptible to subjective interpretations, misidentifications, and reporting bias. Additionally, variations in data collection methods over time and across different regions may impact the accuracy and completeness of the findings. External factors, such as media influence or cultural perceptions, could also affect the frequency of reported sightings, making it difficult to establish objective patterns.

## Future Studies

Future research could incorporate additional datasets to enhance the analysis, such as meteorological data, air traffic records, and astronomical observations to rule out natural explanations for sightings. Machine learning techniques could be applied to classify and predict sighting patterns based on geographical and temporal factors. Expanding the dataset by integrating global reports and conducting sentiment analysis on public reactions to sightings could provide deeper insights into the cultural and psychological aspects of UFO phenomena.

# Conclusion

This study provided an exploratory analysis of UFO sightings, highlighting key trends in time, location, and reported characteristics. While intriguing patterns emerged, the limitations of self-reported data emphasize the need for further investigation using more comprehensive and validated sources. By combining data science techniques with interdisciplinary research, future studies may offer more robust conclusions about the nature of these sightings and their potential explanations.