* Brief description of the data set and a summary of its attributes
* Initial plan for data exploration
* Actions taken for data cleaning and feature engineering
* Key Findings and Insights, which synthesizes the results of Exploratory Data Analysis in an insightful and actionable manner
* Formulating at least 3 hypothesis about this data
* Conducting a formal significance test for one of the hypotheses and discuss the results
* Suggestions for next steps in analyzing this data
* A paragraph that summarizes the quality of this data set and a request for additional data if needed

Datasets used:

* UFO sightings -> reports of unidentified flying object reports in the last century.
* Question should be in what country are most likely to have UFO sightings?
* The UFo sightings datasets consists of 2 parts: completed and scrubbed.
* Both contains datetime, city, state, country, shapre, duration in seconds, duration in hours or mins, comments, date posts, latitude, longitude.

data are: city, state, time, description and duration of each sighting.

We are predicting: what country

Imported datasets.

|  | **datetime** | **city** | **state** | **country** | **shape** | **duration (seconds)** | **duration (hours/min)** | **comments** | **date posted** | **latitude** | **longitude** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 10/10/1949 20:30 | san marcos | tx | us | cylinder | 2700 | 45 minutes | This event took place in early fall around 194... | 4/27/2004 | 29.8830556 | -97.941111 |
| **1** | 10/10/1949 21:00 | lackland afb | tx | NaN | light | 7200 | 1-2 hrs | 1949 Lackland AFB&#44 TX. Lights racing acros... | 12/16/2005 | 29.38421 | -98.581082 |
| **2** | 10/10/1955 17:00 | chester (uk/england) | NaN | gb | circle | 20 | 20 seconds | Green/Orange circular disc over Chester&#44 En... | 1/21/2008 | 53.2 | -2.916667 |
| **3** | 10/10/1956 21:00 | edna | tx | us | circle | 20 | 1/2 hour | My older brother and twin sister were leaving ... | 1/17/2004 | 28.9783333 | -96.645833 |
| **4** | 10/10/1960 20:00 | kaneohe | hi | us | light | 900 | 15 minutes | AS a Marine 1st Lt. flying an FJ4B fighter/att... | 1/22/2004 | 21.4180556 | -157.803611 |

Data.columns:

Index(['datetime', 'city', 'state', 'country', 'shape', 'duration (seconds)',

'duration (hours/min)', 'comments', 'date posted', 'latitude',

'longitude'],

dtype='object')

data.dtypes:

datetime object

city object

state object

country object

shape object

duration (seconds) object

duration (hours/min) object

comments object

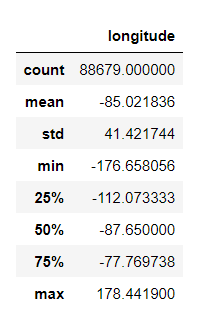
date posted object

latitude object

longitude float64

dtype: object

data.describe()



Data.isnull().sum()

datetime 0

city 0

state 7409

country 12365

shape 2922

duration (seconds) 2

duration (hours/min) 3019

comments 35

date posted 0

latitude 0

longitude 0

dtype: int64

# we have missing values, therefore we need to clean it up:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 80332 entries, 0 to 80331

Data columns (total 11 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Date\_time 80332 non-null object

1 city 80332 non-null object

2 state/province 74535 non-null object

3 country 70662 non-null object

4 UFO\_shape 78400 non-null object

5 length\_of\_encounter\_seconds 80332 non-null object

6 described\_duration\_of\_encounter 80332 non-null object

7 description 80317 non-null object

8 date\_documented 80332 non-null object

9 latitude 80332 non-null object

10 longitude 80332 non-null float64

dtypes: float64(1), object(10)

In order to convert data types:

Do 3 basic operations:

* 1. Use atype() to force an appropriate dtype
  2. Create a custom function to convert the data
  3. Use pandas functions such as to\_numeric() or to\_datetime()