Final Report – Navdeep Raj

UFO Sightings Analysis

INTRODUCTION & PROBLEM STATEMENT

Unidentified flying object (UFO), also called flying saucer, any aerial object or optical phenomenon not readily identifiable to the observer. UFOs became a major subject of interest following the development of rocketry after World War II and were thought by some researchers to be intelligent extraterrestrial life visiting Earth.

UFO reports have varied widely in reliability, as judged by the number of witnesses, whether the witnesses were independent of each other, the observing conditions (e.g., fog, haze, type of illumination), and the direction of sighting. Typically, witnesses who take the trouble to report a sighting consider the object to be of extraterrestrial origin or possibly a military craft but certainly under intelligent control. This inference is usually based on what is perceived as formation flying by sets of objects, unnatural—often sudden—motions, the lack of sound, changes in brightness or colour, and strange shapes.

In our case, our primary objective is:

1. to perform an in-depth EDA
2. Classify the UFO sightings as per the duration interval i.e Short (< 5 Minutes), Medium (5 – 60 Minutes) and Long (> 60 Minutes)
3. Find interesting trends and predict some behavior in terms of their shape, size, color, etc.

DATA

Data has been collected from NUFORC(National UFO Reporting Center) Website which has catalogued almost 90,000 reported UFO sightings over its history, most of which were in the United States.

Initial dataset has over 88000 UFO recordings and 11 attributes namely:

- Datetime – Date & time of occurrence
- City
- State
- Country
- Shape
- Duration (Seconds)
- Duration (Hours)
- Comments/Description
- Date Posted
- Latitude
- Longitude

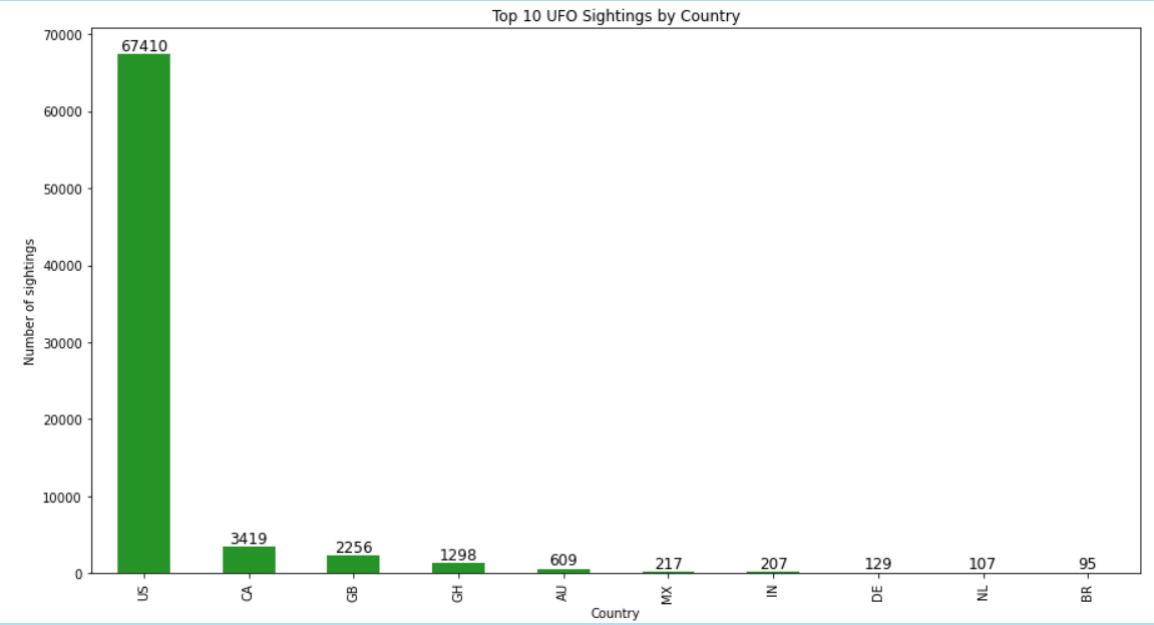
DATA WRANGLING

➔ Date Columns were converted to Datetime format
➔ Country, State and City Columns were dropped as they had lots of null values and instead, I used Reverse Geocoder (RG package) to get Country, State and City from Latitude and longitude tuple
➔ Using regex package, both duration columns were converted to single duration column in minutes after eliminating the null rows.
➔ For filling the missing values in shape, function was created to find if the unique shape keywords were found in the description. And rest of the missing rows were omitted
➔ Datetime column was used to generate month and year
➔ After final formatting and dealing with missing values, dataset had around 77500 records and 11 attributes namely:
    - Date_time,
    - Duration_minutes
    - Description
    - Date_posted
    - lat_long
    - Country
    - State
    - City
    - Shape_final
    - Year

- Month

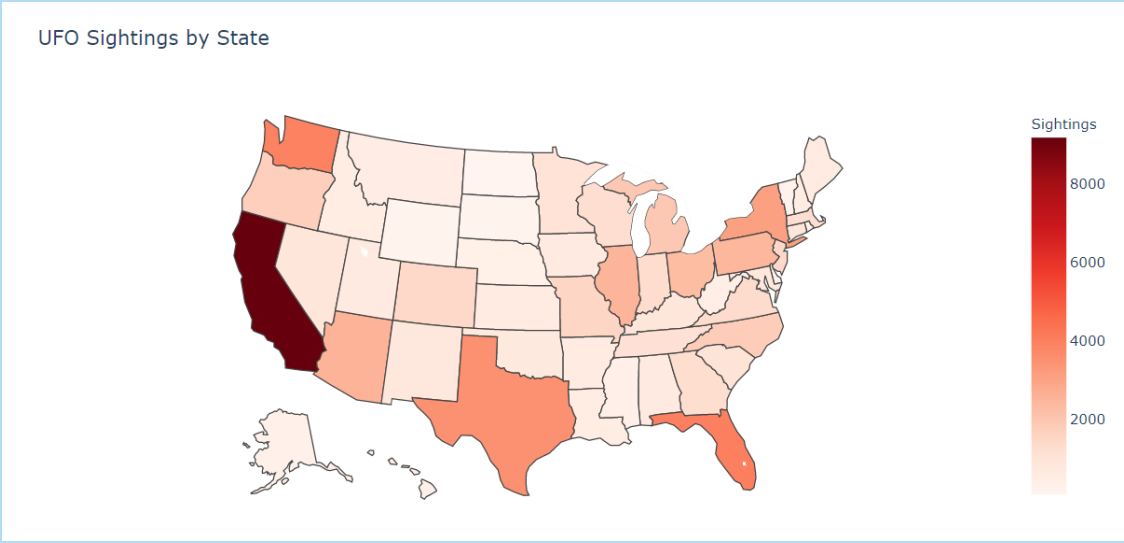
EXPLORATORY DATA ANALYSIS

- **UFO Sightings by Country**



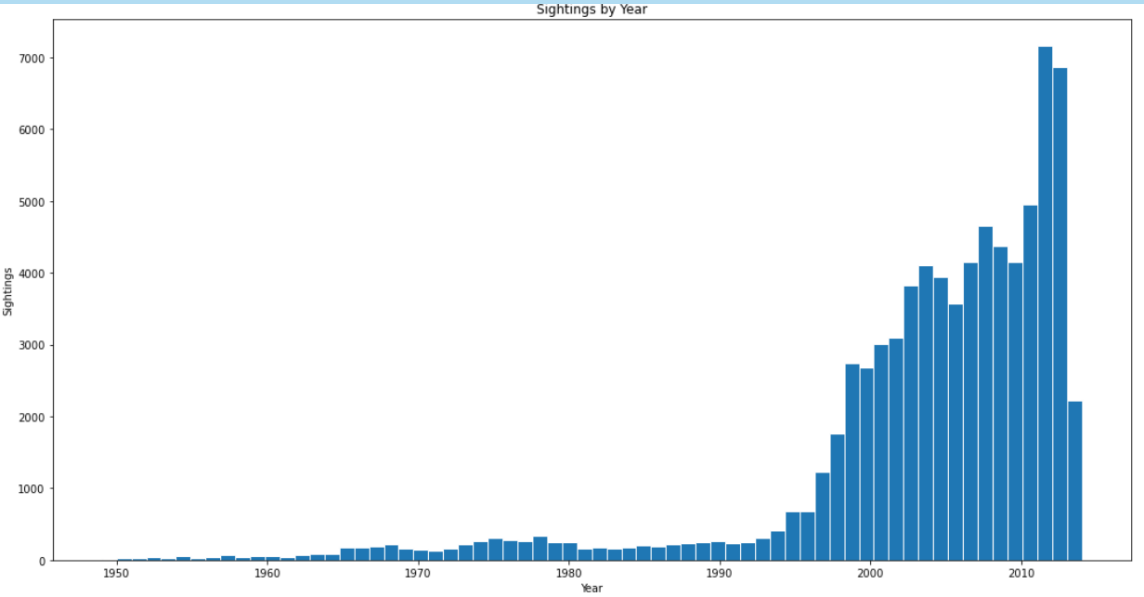
As seen, most of the sightings were in US followed by Canada, Great Britain, Ghana and Australia

- **UFO Sightings in US by State**



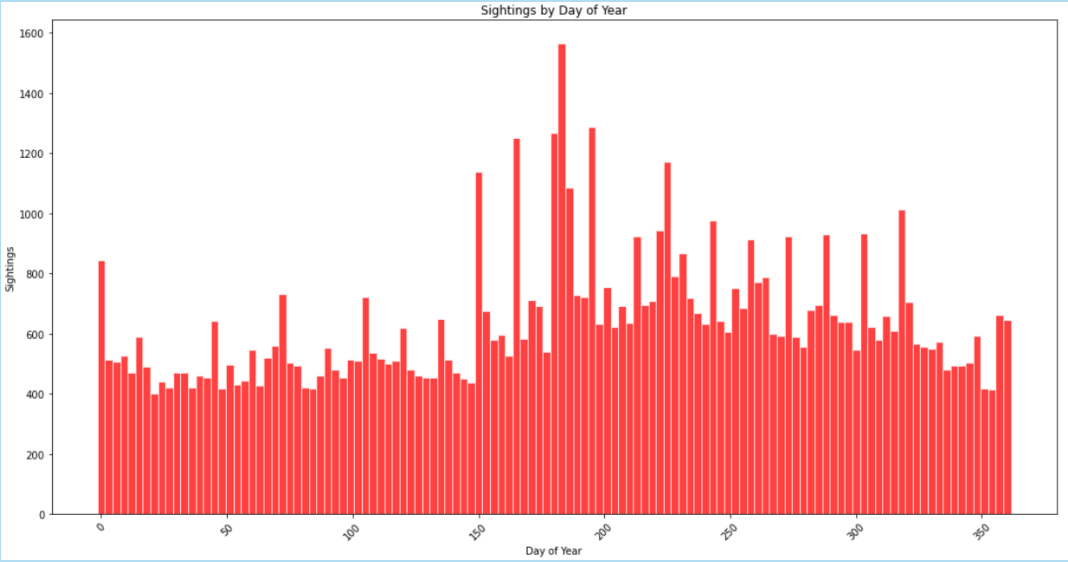
Most of the observations were in California (more than 9000) followed by Florida (around 4000) and Washington

- **UFO Sightings trend by Year**



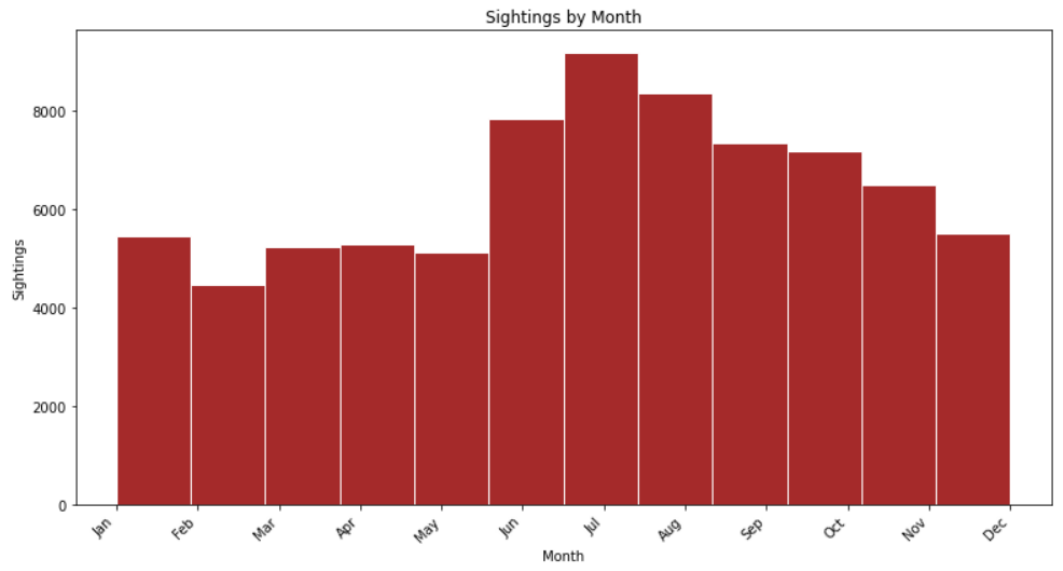
For the UFO sighting trend over the last 70 years, it is observed that since 2000 onwards there has been a huge increase in sightings starting from close to 3000 and going up to almost 7000 in 2012 and 2013

- **UFO Sightings Trend by Day of the Year**



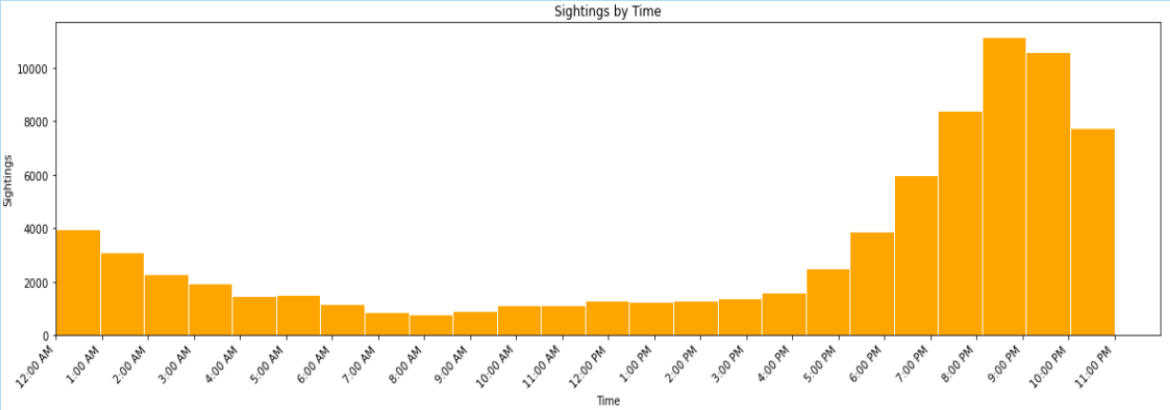
If we see distribution by day if the year it doesn't really give much insight but more sightings are recorded during the middle of the year

- UFO Sightings Distribution by Month



For UFO distribution by months, it can be seen that sightings begin to rise after half of the year is done and then again decrease by the end of the year, It is at peak during July and August with almost more than 8000 sightings

- **Sightings by Time of Day**

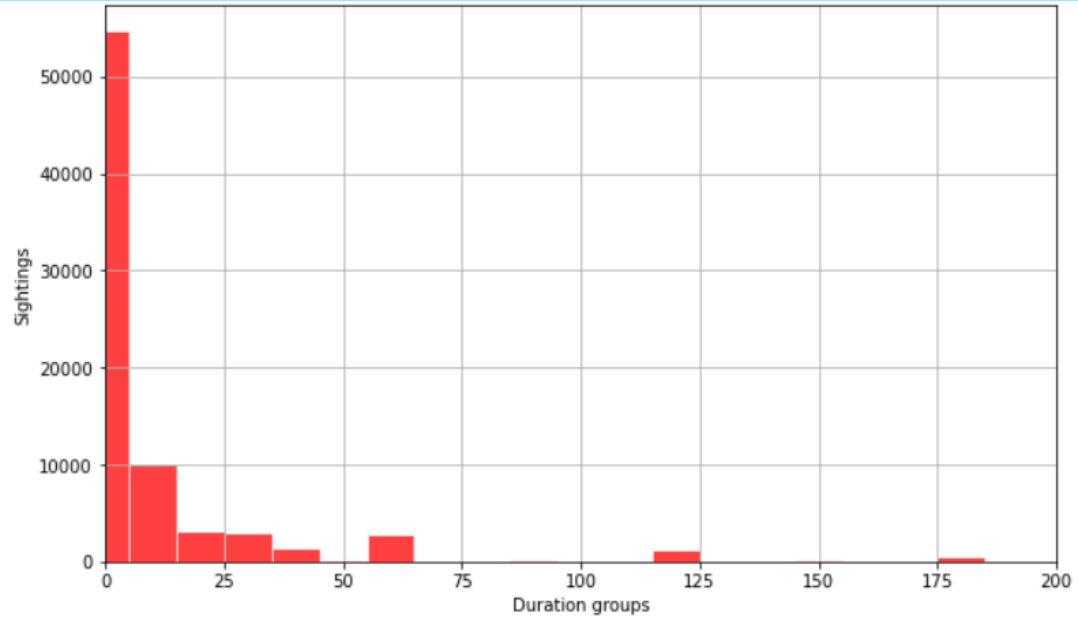


To map sightings by time of the day, this is more practical and common as observed that most of the sightings are between 7:00 PM and 11:00 PM, during mornings and afternoon, it drops to less than 2000 whereas evening onwards it rises up to 10000

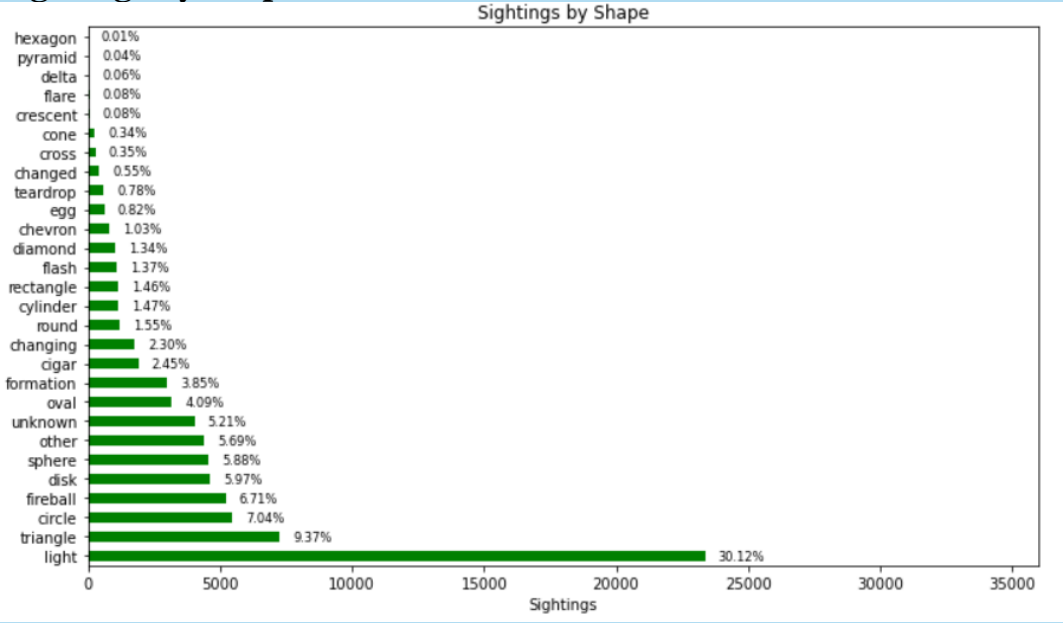
- **Duration distribution**

For Duration, I observed a massive outlier with 137800 minutes, which I though didn't makes sense so after deleting it,

a) We observe we still have max value of 5760 minutes which also is suspicious but I didn't delete it and viewing the boxplot and summary, it can be seen that avg duration is around 15 minutes with std of 72
b) If we look at the IQR its between 0.5 minutes and 10 minutes which is clearly very less as compared to max value
c) I created another boxplot with data with 3 standard deviations from mean, it can be observed that with Max values of 201 minutes, our mean drops down to 11.2 and std of 24 minutes with IQR between 0 and 10 minutes
d) If we look at the duration histogram most of the sightings lie with first 5 bins that is less than 10 minutes



- **Sightings by Shape**



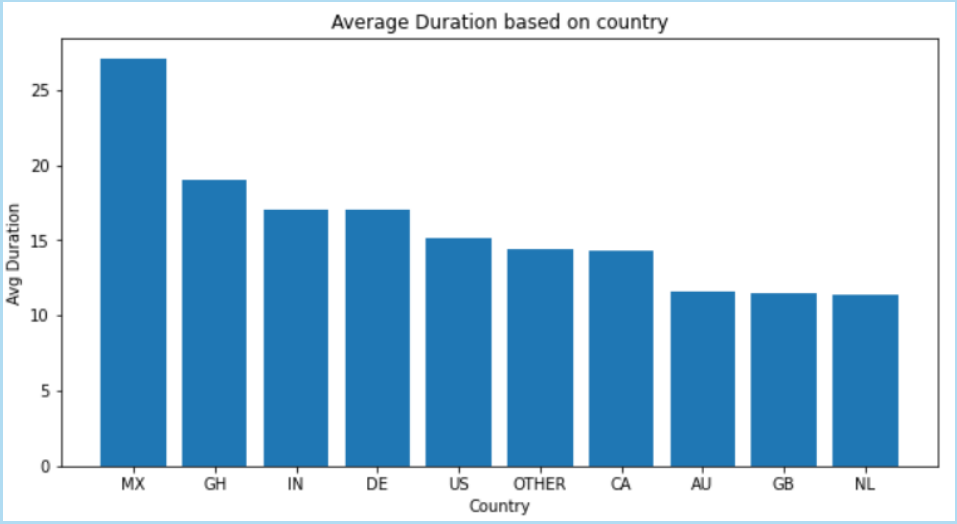
After getting shapes form the data, it can be seen that 30% of UFO sightings have light shape followed by triangle and circle at 9% and 7% respectively

BIVARIATE ANALYSIS

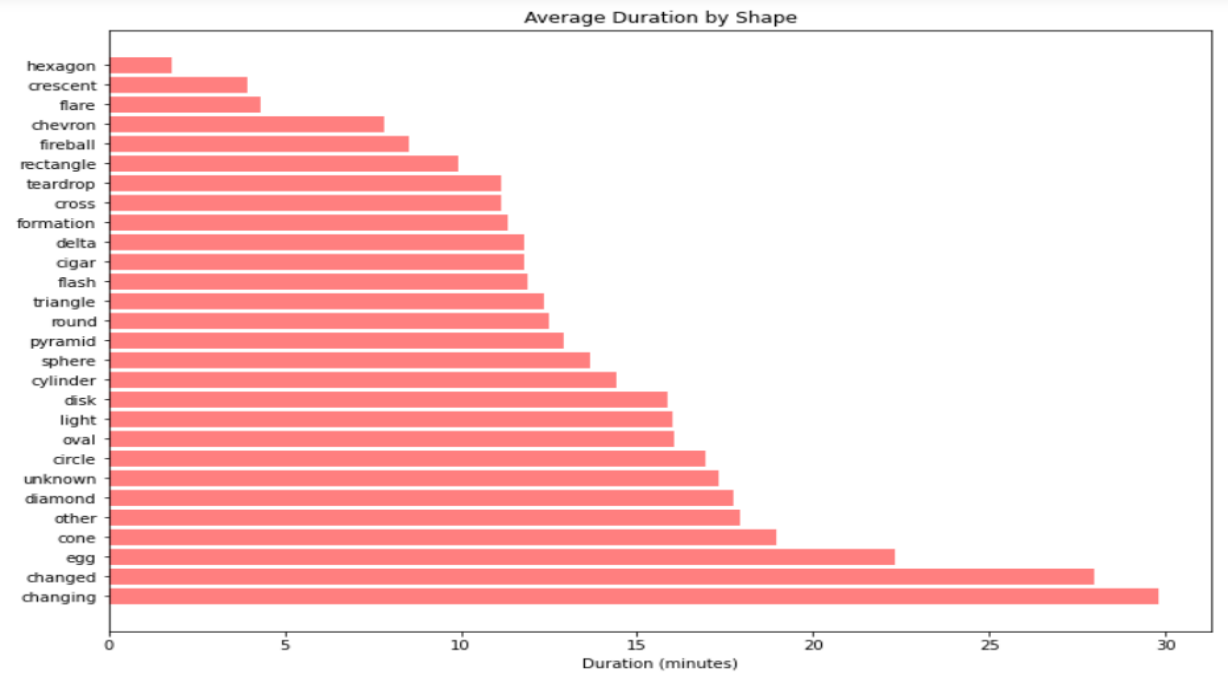
- **Avg Duration by Country**

It shows that for most of the countries avg duration is lee than 20 minutes, for some popular countries: -

Country	Max_Duration	Avg_Duration
MX	1080.0	27.089585
GH	1440.0	19.036849
IN	300.0	17.041643
DE	840.0	17.024341
US	5760.0	15.151452



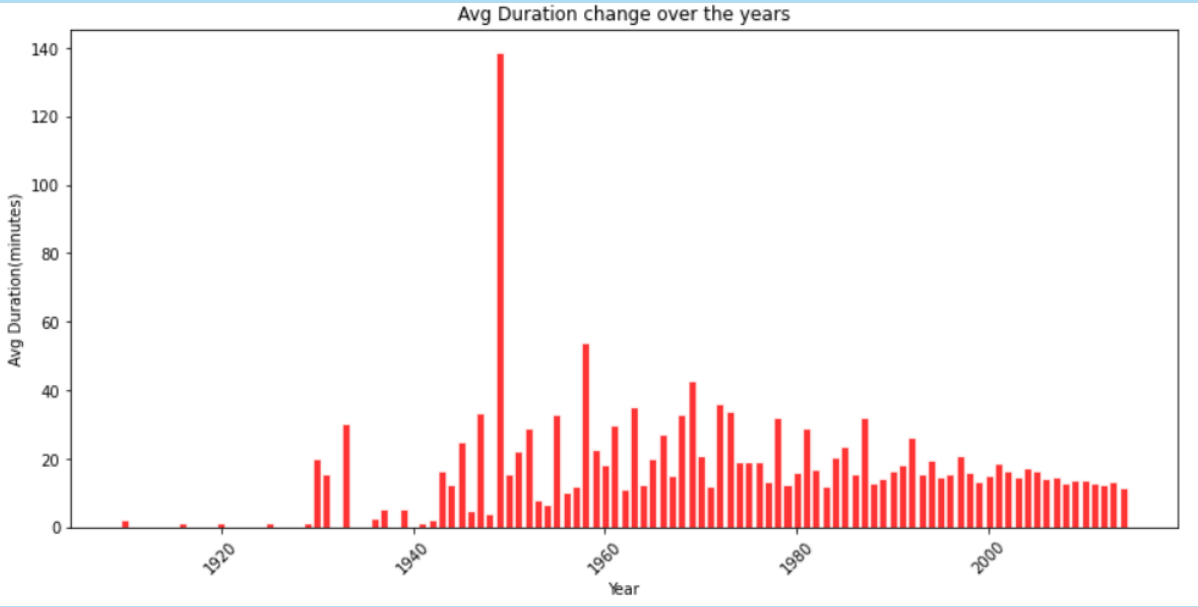
- **Average Duration by Shape**



When it comes to the average duration by shape, its observed that Changing shape has the longest average duration of around 30 minutes followed by Egg at 22.3 minutes and cone shape at 19 minutes

sh	Max_Duration	Avg_Duration
changing	2640.0	29.822803
changed	4320.0	27.965222
egg	4320.0	22.329778
cone	300.0	18.932337

- **Avg Duration over the Years**



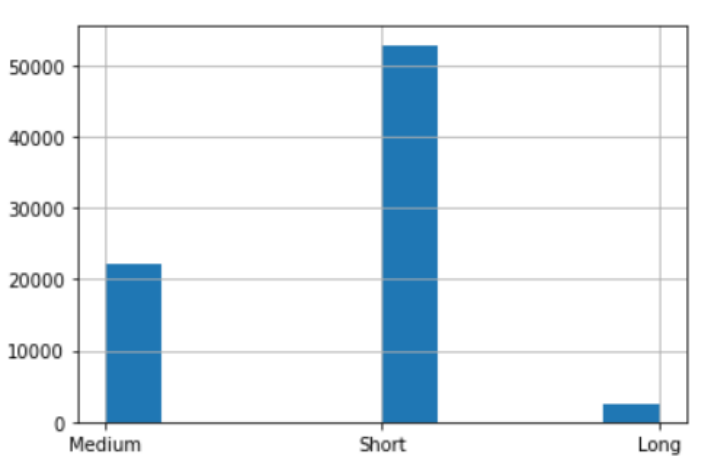
If we map average duration trend over the past years, we can visualize that for last 10-12 years average duration of UFO Sighting has been in range of 15 - 20 minutes but it also depends on the number of sightings in that year

MACHINE LEARNING MODEL SELECTION

For modelling purposes, Duration column was divided into 3 classes: -

1. Short – less than 5 minutes
2. Medium – 5 – 60 minutes
3. Long – greater than 60 minutes

Distribution of target variable in our data



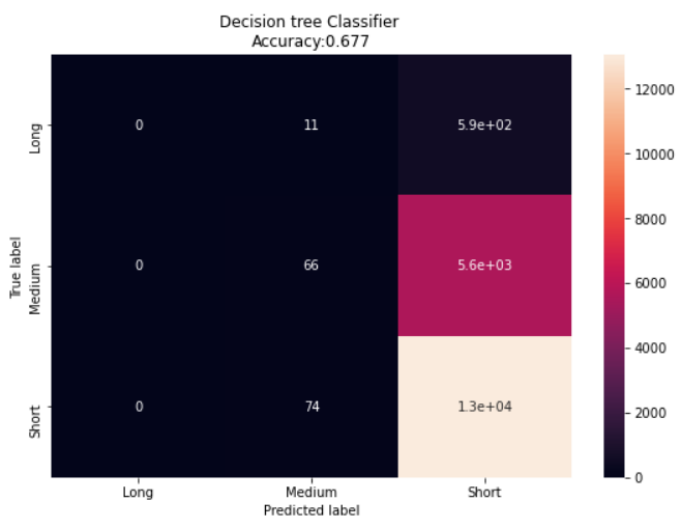

```
Short     52826
Medium    22119
Long       2542
Name: Duration, dtype: int64
```

It can be observed our data is not completely unbalance and major records fall into short duration category with more than 50000 sightings, followed by Medium at 22000 and Long at around 2500

Our dataset is almost categorical for all the features so after applying appropriate encoding technique i.e get_dummies() method, we finally chose 5 algorithms for initial training and prediction and the results are as follows:

Classification Algorithms & Metrics

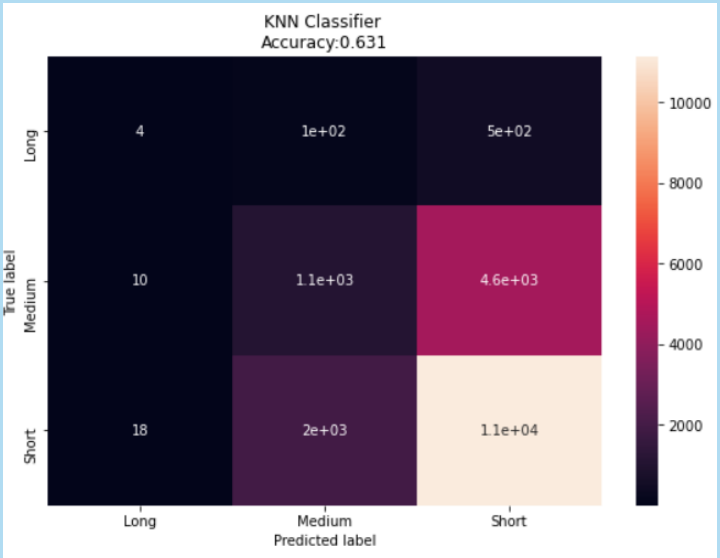
Decision Tree Classifier



	precision	recall	f1-score	support
Long	0.00	0.00	0.00	605
Medium	0.44	0.01	0.02	5635
Short	0.68	0.99	0.81	13132
accuracy			0.68	19372
macro avg	0.37	0.34	0.28	19372
weighted avg	0.59	0.68	0.55	19372

As Seen from the Confusion Matrix , Around 600 values which are long were predicted as Short whereas 5500 Medium were predicted as short and 13000 were predicted correctly so around 68% were predicted correctly for short whereas only 66 were predicted correctly for Medium duration and 11 for long duration, overall Accuracy is 0.68 which is not bad but still predictability is not that good for medium and long

KNN Classifier



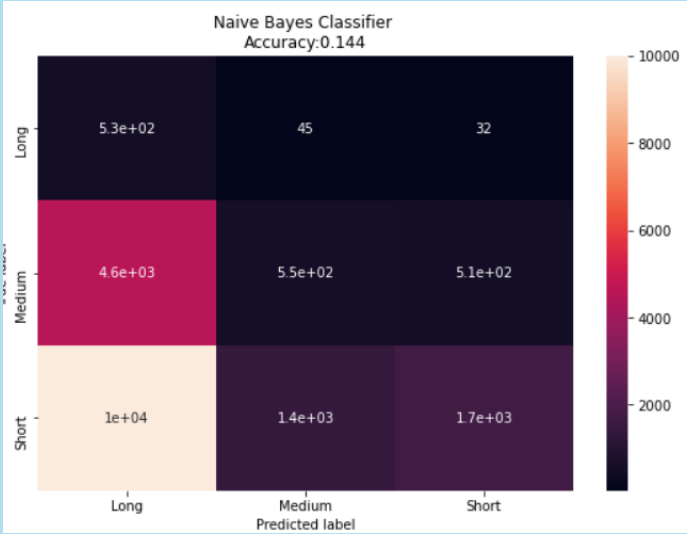
	precision	recall	f1-score	support
Long	0.12	0.01	0.01	605
Medium	0.34	0.19	0.24	5635
Short	0.69	0.85	0.76	13132
accuracy			0.63	19372
macro avg	0.38	0.35	0.34	19372
weighted avg	0.57	0.63	0.59	19372

As Seen from the Confusion Matrix, although accuracy is bit less at 63% but it shows improvement in predicting as compared to decision tree classifier. Around 497 values which are long were predicted as Short whereas 104 as Medium and 4 correctly as long.

In case of medium, its improved to 1000 values being predicted correctly and 4500 as short

In Short category, although the number is bit less, around 11100 values being predicted correctly, but it still is better with other categories

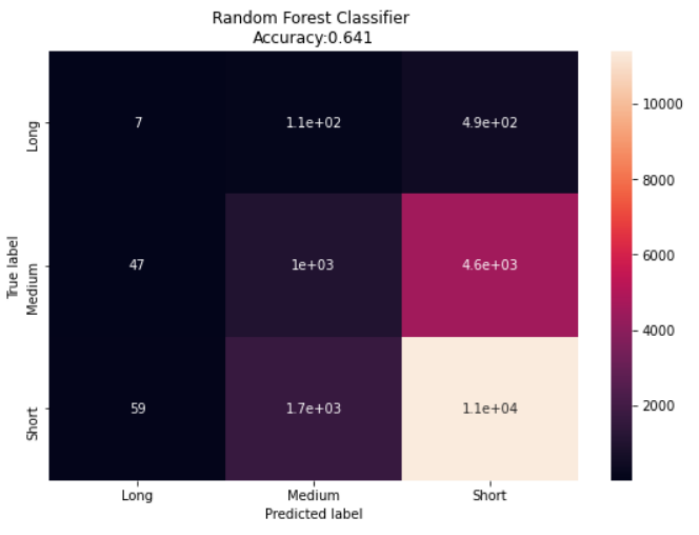
Naïve Bayes Classifier



	precision	recall	f1-score	support
Long	0.03	0.87	0.07	605
Medium	0.28	0.10	0.14	5635
Short	0.76	0.13	0.22	13132
accuracy			0.14	19372
macro avg	0.36	0.37	0.15	19372
weighted avg	0.60	0.14	0.20	19372

As Seen from the Confusion Matrix, this is pretty worse when it comes to accuracy just .14 and its reveres of previous models, For long this is predicting 528 correct vales out of 605 values but for other classes the prediction is pretty poor. 4500 medium incorrectly as long and 508 medium as short. 10000 short values incorrectly as long and 1400 short as medium while the correct predicting for short is very less compared to previous models i.e. only 1700 values

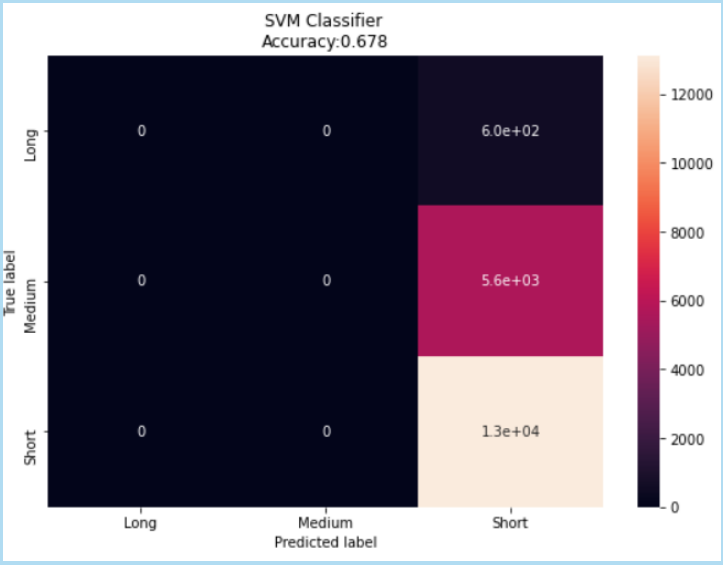
Random Forest Classifier



	precision	recall	f1-score	support
Long	0.06	0.01	0.02	605
Medium	0.36	0.18	0.24	5635
Short	0.69	0.87	0.77	13132
accuracy			0.64	19372
macro avg	0.37	0.35	0.34	19372
weighted avg	0.58	0.64	0.59	19372

As Seen from the Confusion Matrix, with accuracy of .641 another tree-based algorithm seems to be performing same as decision tree but with little improvement with correct values being predicted for Long, Medium and Short are 7, 1006 and 11400 respectively. which is acceptable.

SVC Classifier (OVO Mapping)



	precision	recall	f1-score	support
Long	0.00	0.00	0.00	605
Medium	0.00	0.00	0.00	5635
Short	0.68	1.00	0.81	13132
accuracy			0.68	19372
macro avg	0.23	0.33	0.27	19372
weighted avg	0.46	0.68	0.55	19372

Here i have used SVC for multiple classes using One over One classification method , as per confusion matrix, its only predicting values as sort there is no prediction for medium and large, So although accuracy is better bur again prediction is poor.

Metrics Comparison

Class	Support	F1 Score	Precision	Recall
SVC Classifier	Accuracy	0.68		
Short	13132	0.81	0.68	1.0
Medium	5635	0	0	0
Long	605	0	0	0
Decision Tree Classifier	Accuracy	0.68		
Short	13132	0.81	0.68	0.99
Medium	5635	0.023	0.44	0.012
Long	605	0	0	0
KNN Classifier	Accuracy	0.63		
Short	13132	0.76	0.69	0.85
Medium	5635	0.24	0.34	0.19
Long	605	0.01	0.12	0.01
Naives Bayes Classifier	Accuracy	0.14		
Short	13132	0.22	0.76	0.13
Medium	5635	0.14	0.28	0.10
Long	605	0.07	0.03	0.87
Random Forest Classifier	Accuracy	0.64		
Short	13132	0.77	0.69	0.87
Medium	5635	0.24	0.36	0.18
Long	605	0.02	0.06	0.01

As Seen from the Comparison Chart for Various Models

* Although SVC is higher in accuracy but, its prediction for medium and Long duration is 0, so it doesn't look suitable

* In my opinion, **KNN and Random Forest** seem to be much better models for our data, although accuracy is around 0.64 for both but their prediction is better than other models

Model Hyperparameter Tuning using GridSearchCV

I used GridsearchCV on four of our algorithms to see we can improve our previous models, Below is the summary of the tuning and best score along with best parameters

- **Model - Decision Tree**
  - Best Criterion - Entropy
  - Max Depth = 3
  - Score = 0.682
- **Model - KNN**
  - Best Neighbors - 20
  - Score = 0.656
- **Model - Random Forest**
  - Best Criterion - Gini
  - Max features = 3
  - Best n_Estimators = 500
  - Score = 0.616
- **Model - Multinomial NB**
  - Best Alpha = 0.1
  - Score = 0.681

In my opinion **Random Forest** seems to be the best model after tuning parameters, although the accuracy is less at 0.62 but its prediction is much better than other models when using the best parameters

* Although Multinomial Nb is the best in accuracy but its correct prediction for medium and Long is very less

WINNER – RANDOM FOREST CLASSIFIER

FINAL TAKEAWAYS

- Random Forest seems to be the best performing algorithm for our problem statement
- After carefully examining and exploring the data, this data is more suitable fir EDA and visual analysis rather than predictions
- A lot of recordings are unreliable as many records were unrealistic especially in case of duration going over to 2000 or 5000 minutes which doesn't seem correct
- A much better mining problem can be to analyze the comments/description and differentiate fake sightings from real ones in terms of credibility
- We need some more attributes to make our data more reliable if we want to predict their sightings or duration
- Some of the trends found were:

    - Most of the durations fall in short category that is less than 5 minutes
    - Most observed shape was just light which can be misleading at night time so records may be unreliable
    - There is no validation of the sightings as it's just what's seen by the observers who can be researchers or just common people
    - Data mainly caters to US as more than 50000 sightings are recorded there

Could such sightings be authentic? Of course, it's possible; many things are possible. The question is not what is possible but what is probable—what evidence and logic suggest. Before jumping to conclusions about ETs in spacecraft, we must look at the most likely explanations. Without some independent confirmation or other evidence, it's hard to know what observers might have seen. But is it more likely that they saw an optical illusion, or that a large, unknown object hovering or light flares, etc,

UFO reports have varied widely in reliability, as judged by the number of witnesses, whether the witnesses were independent of each other, the observing conditions (e.g., fog, haze, type of illumination), and the direction of sighting. Typically, witnesses who take the trouble to report a sighting consider the object to be of extraterrestrial origin or possibly a military craft but certainly under intelligent control. This inference is usually based on what is perceived as formation flying by sets of objects, unnatural—often sudden—motions, the lack of sound, changes in brightness or colour, and strange shapes.

UFOs are still a trick subject where nothing is concrete although their have been many conspiracy theories and government's secret projects but nothing scientific to prove its reality.