Procedure Followed For Cleaning:-

I have taken the given data into excel and cleaned it by applying a filter to duration column. I have converted all the units of duration into either minutes or seconds or hours. On the other hand, I replaced all the messy data with a figure as per my wish for each of the three shapes.

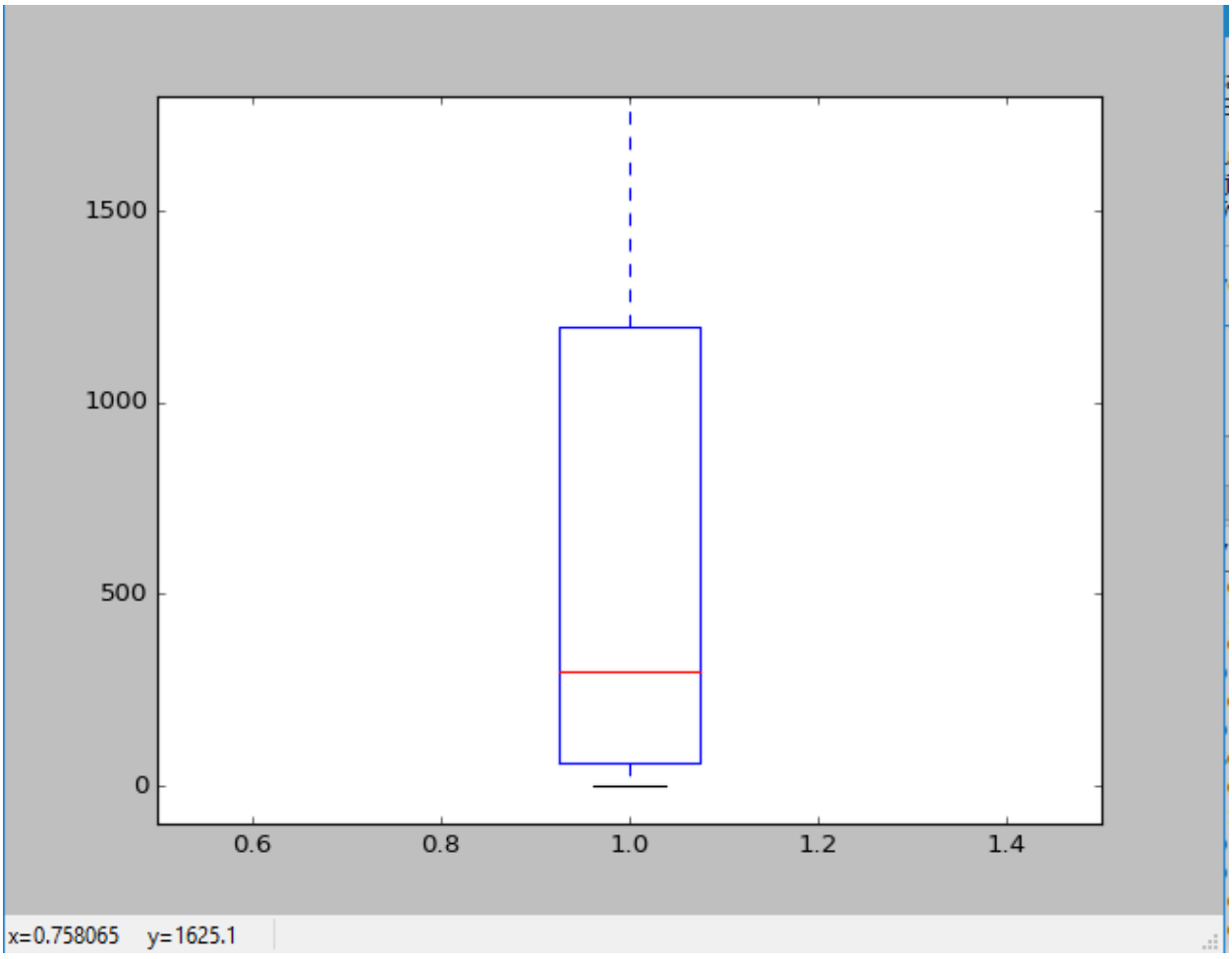
For CIRCLE:- 1 hours

For TRIANGLE:- I deleted all the messy rows

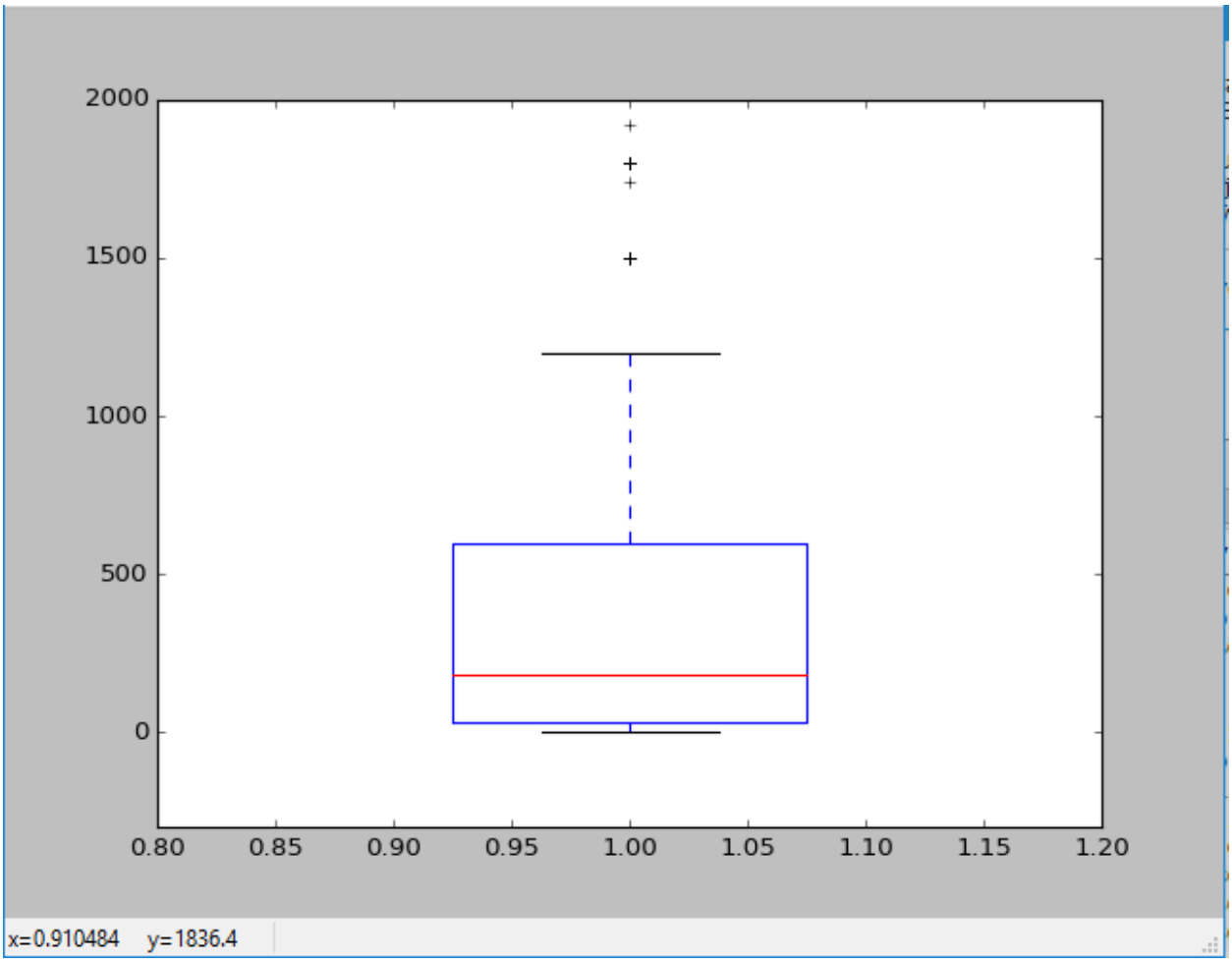
For FIREBALL:- 50 seconds

1.1 : BOX PLOTS

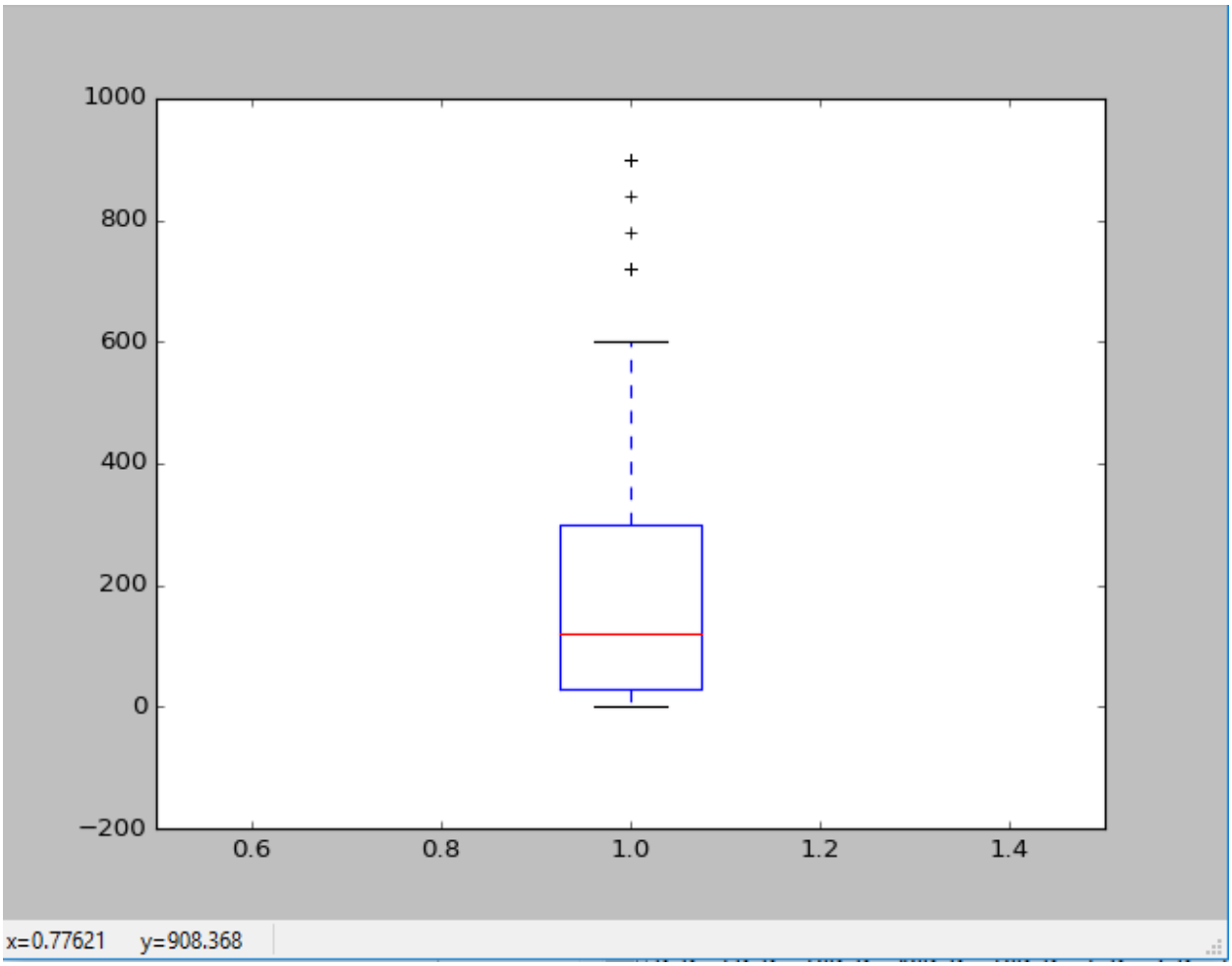
A) **BOX PLOT** of duration of UFO sightings of **CIRCLE**



B) **BOX PLOT** of duration of UFO sightings of **TRIANGLE**

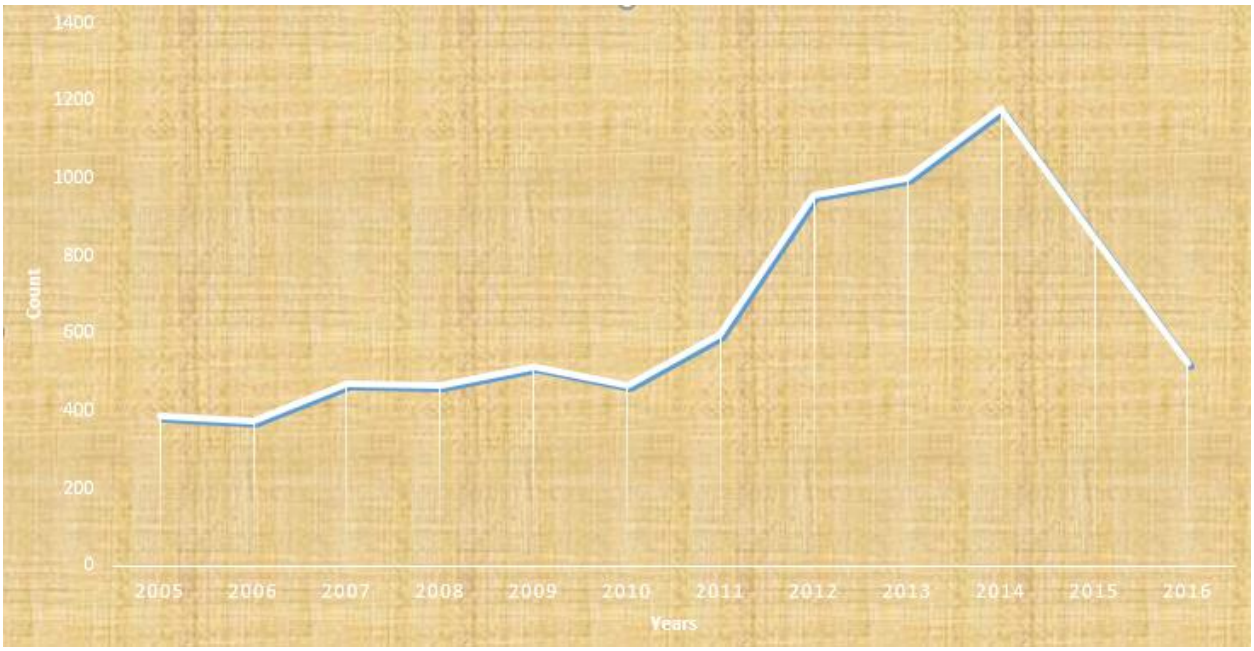


C) **BOX PLOT** of duration of UFO sightings of **FIREBALL**

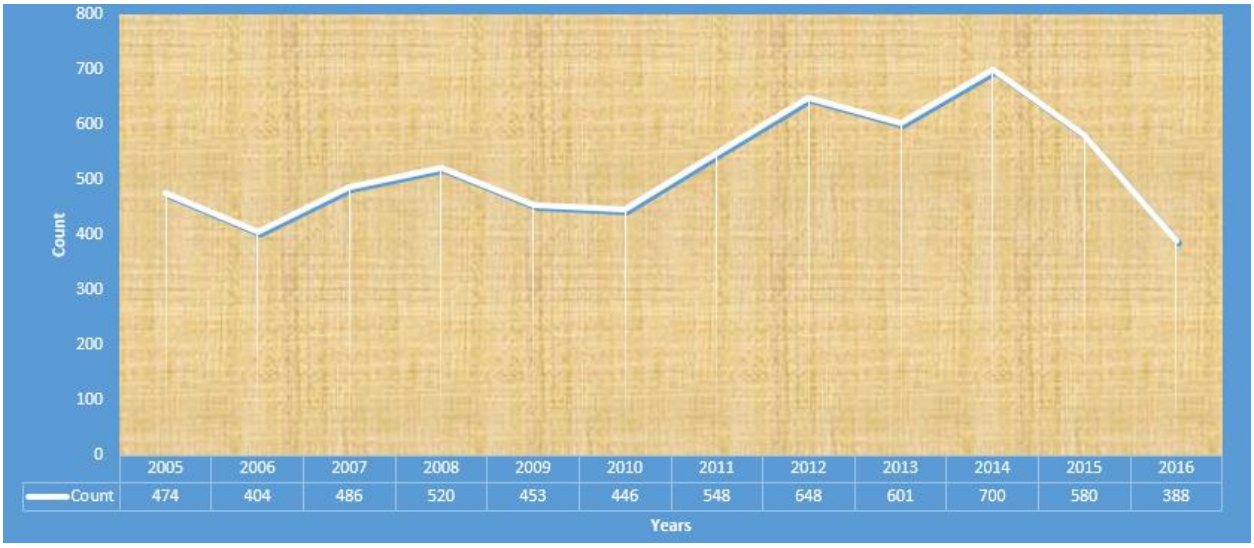


1.2: TIME SERIES

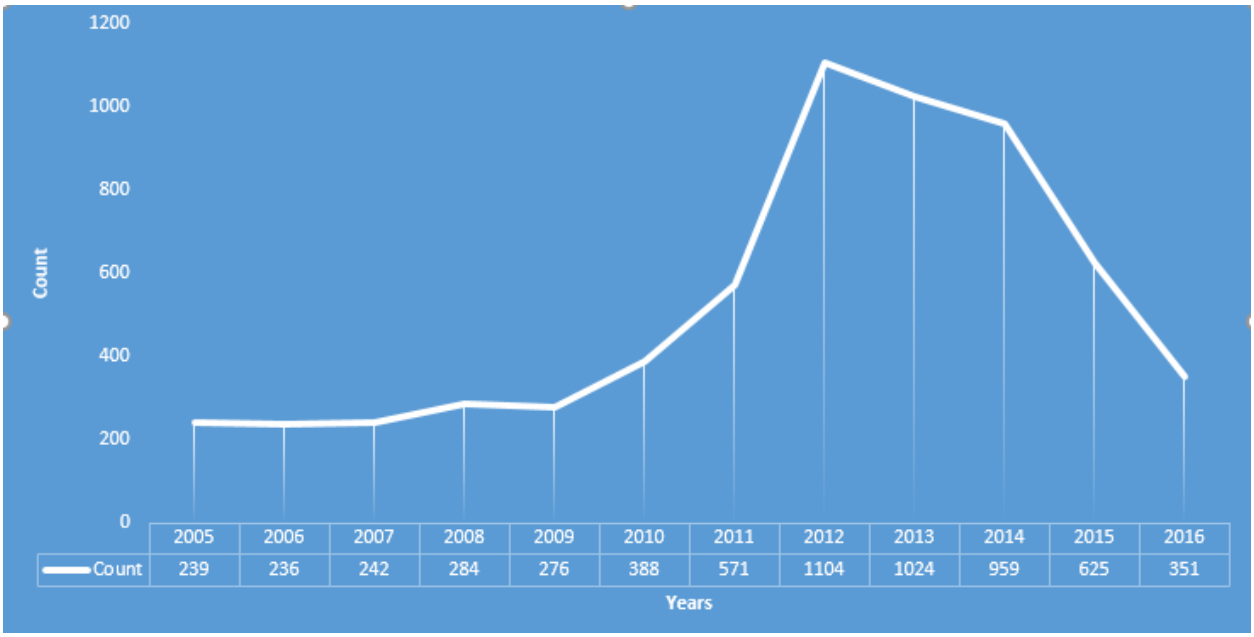
A) **TIME SERIES** with sightings of **CIRCLE**



B)**TIME SERIES** with sightings of **TRIANGLE**

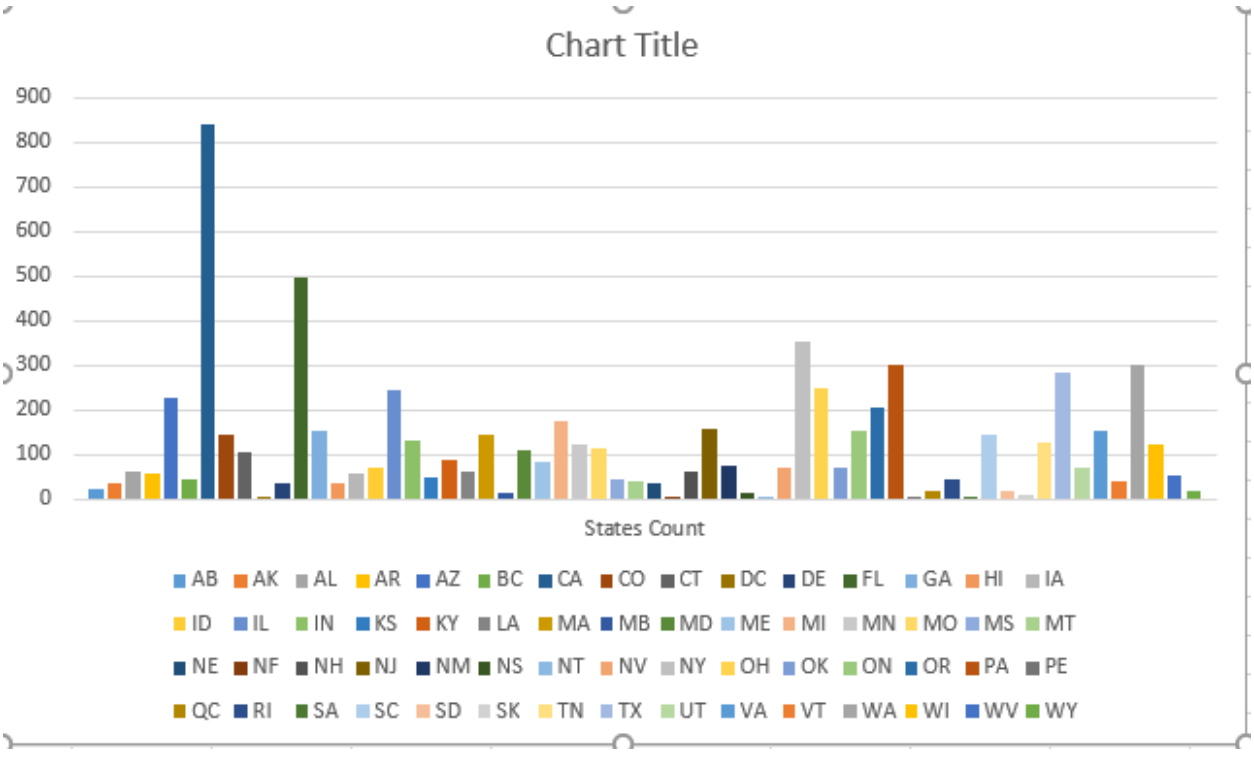


C) **TIME SERIES** with sightings of **FIREBALL**

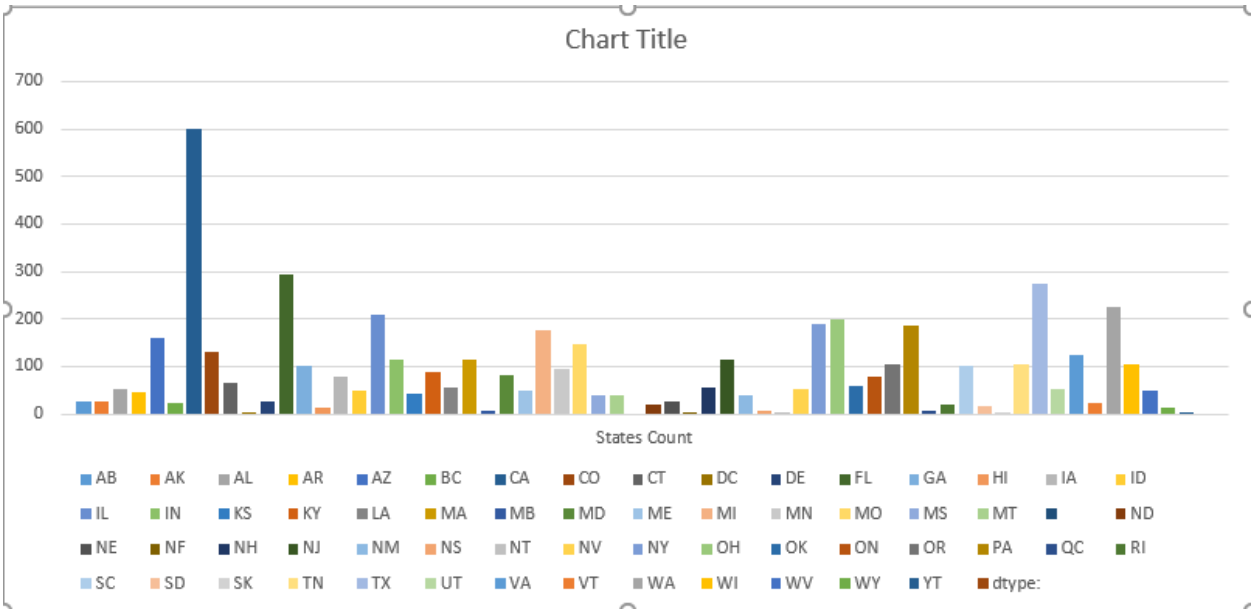


1.3: **BAR CHARTS**

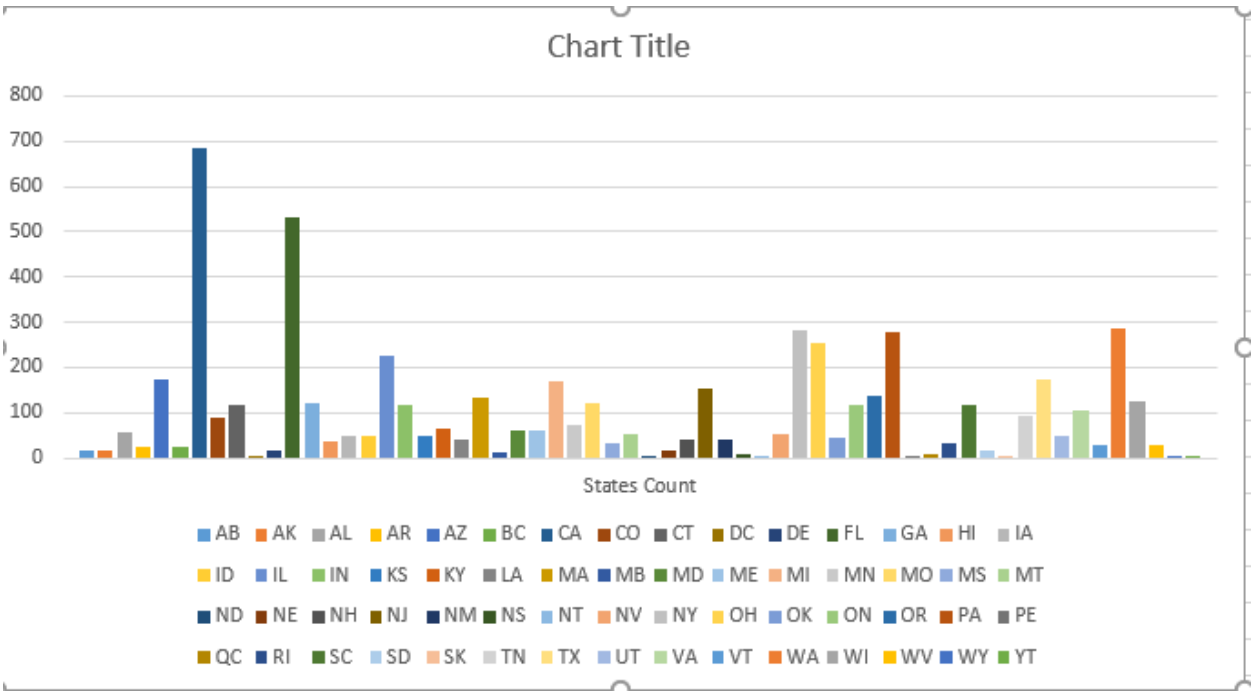
A) **BAR CHART** for sightings of **CIRCLE**



B) **BAR CHART** for sightings of **TRIANGLE**



C) **BAR CHART** for sightings of **FIREBALL**



1.4: Normalize by state population

I have calculated the count to population ratio of states for each shape(Circle, Triangle and Fireball). Later, I have arranged the list in ascending order and took the least and largest ratio states for each shape.

For CIRCLE:- **North Dakota**- Least ratio

Montana- Highest ratio

For TRIANGLE:- **North Dakota**- Least ratio

Montana- Highest ratio

For FIREBALL:- **North Dakota**- Least ratio

Montana- Highest ratio

Observation:- All the three figures have shown the same states as the least and highest ratio with least being ND(North Dakota) and MT(Montana).

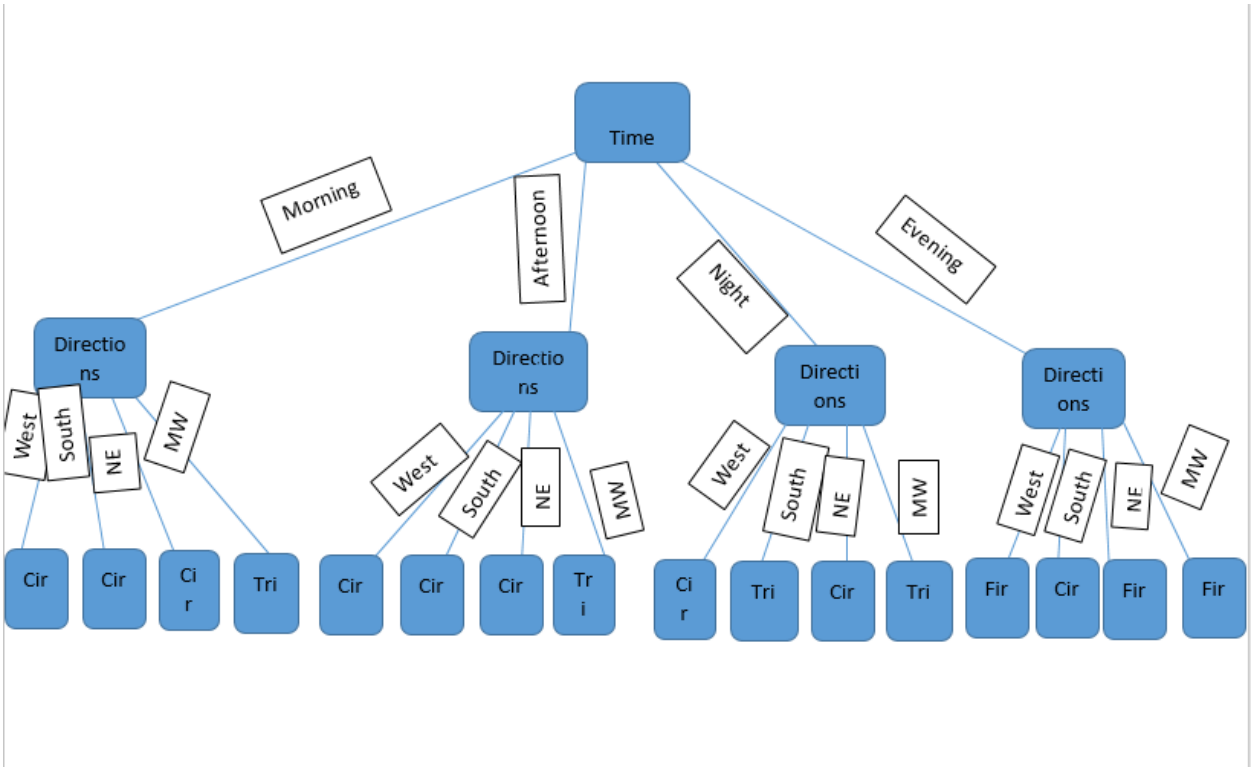
2.1 and 2.2:

I have divided the 50 states of US into four regions (West, South, North East and Middle East) and day into four parts(Morning, Afternoon, Night and Evening)..

Training Set: All the sightings between January 1st,2005 and December 31st,2013.

Test Set: All the sightings between January 1st,2014 and September 22nd,2016.

I have taken the training data according to the requirement mentioned in the question, made a document out of it for all the three shapes and calculated the gini index for each. Then I drew a figure which came out as follows:



Accuracy:-

When I compared the training set data with the test set data, the accuracy that I got was **36.71%**.