HW-2

Sample Solution

Solution-1.

Task 1: UFO Data Collection, Cleaning, and Exploratory Analysis

Q: Do your best to extract maximum value from the messy data; be sure to explain to us the decisions you have made in terms of data extraction and cleaning

Data Cleaning: Data cleaning is done as one of following ways:

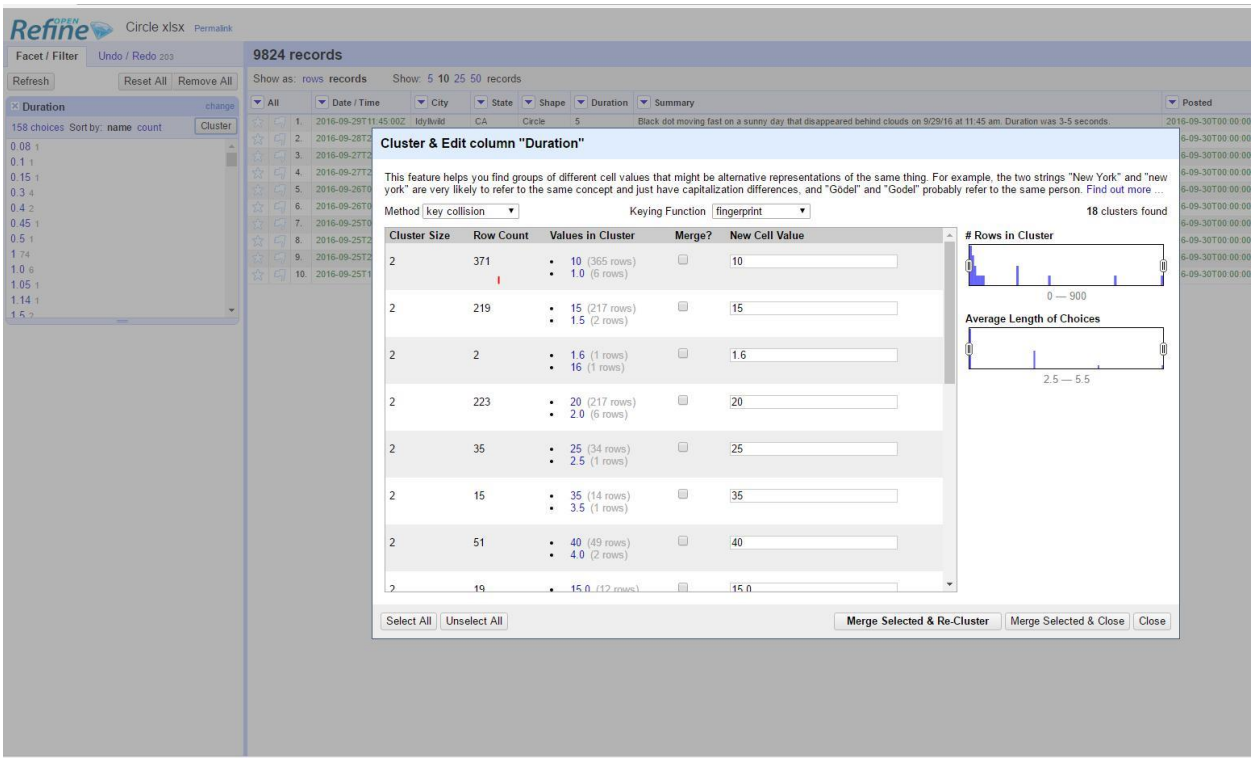
The data is segregated on the basis of UFO shapes as follows: Circle, Triangle, Fireball into an Excel format.

The first column containing sighting date is sorted and cleaned between the dates 1st Jan 2005 and Sep 2016.

This data is loaded in open refine and cleaned using the functionalities of clustering. (Fig 1)

Blank rows and messy data is cleaned. All the values are converted to seconds.

Outliers are removed for creating box plots. (Values above Q3+1.5*I QR are removed)



Total rows after cleaning – 18897

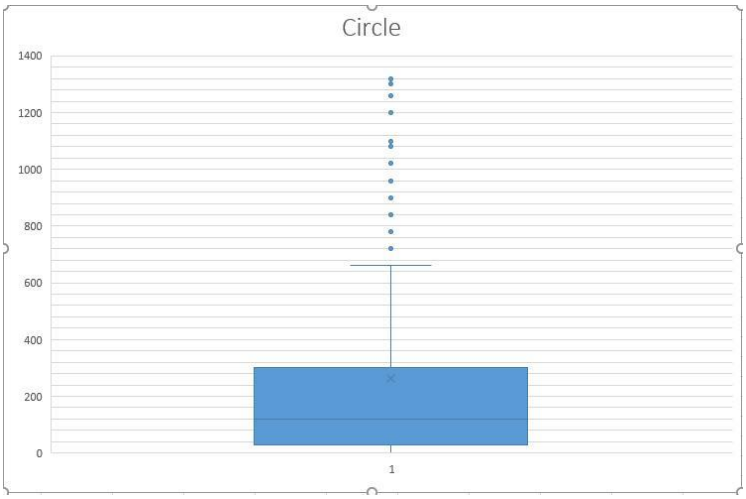
CI RCLE sightings = 7147

Fireball Sightings = 5968

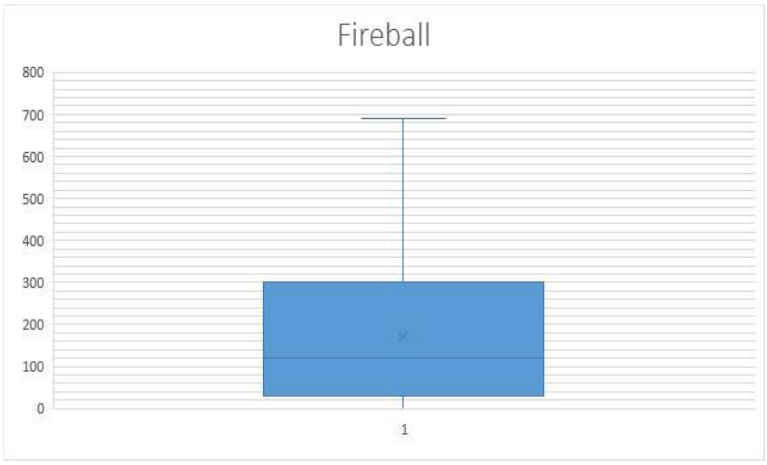
Triangle Sightings = 5784

Q: A boxplot of the duration of UFO sightings of each shape (one boxplot per shape).

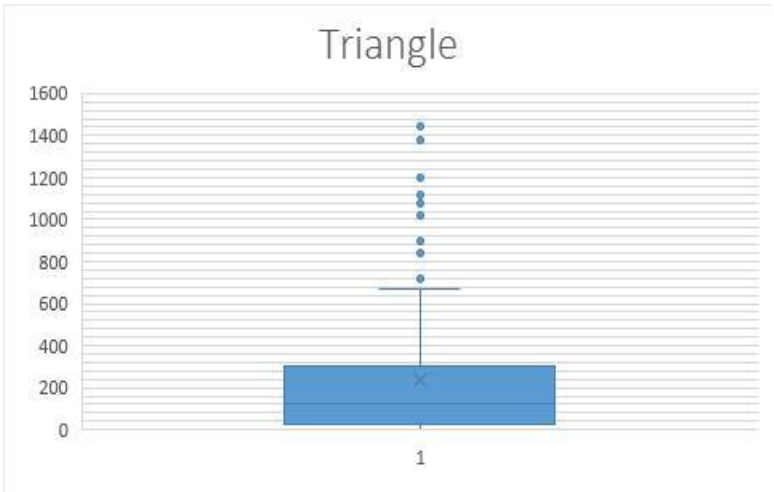
BOXPLOT for **CIRCLE** shaped UFO sightings:



BOXPLOT for **FIREBALL** shaped UFO sightings:



BOXPLOT for **TRIANGLE** shaped UFO sightings:

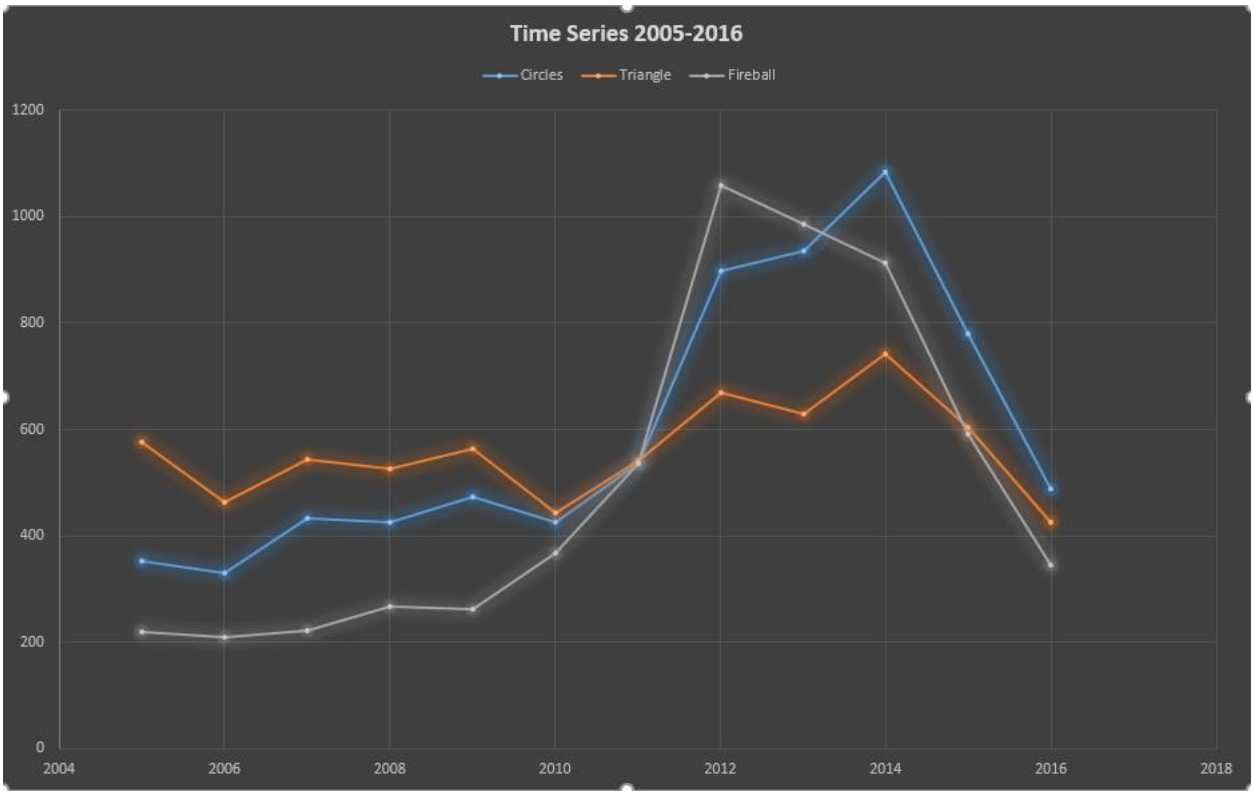


Q-2: A time series figure with the number of sightings per year (one line per shape).

Number of sightings per year:

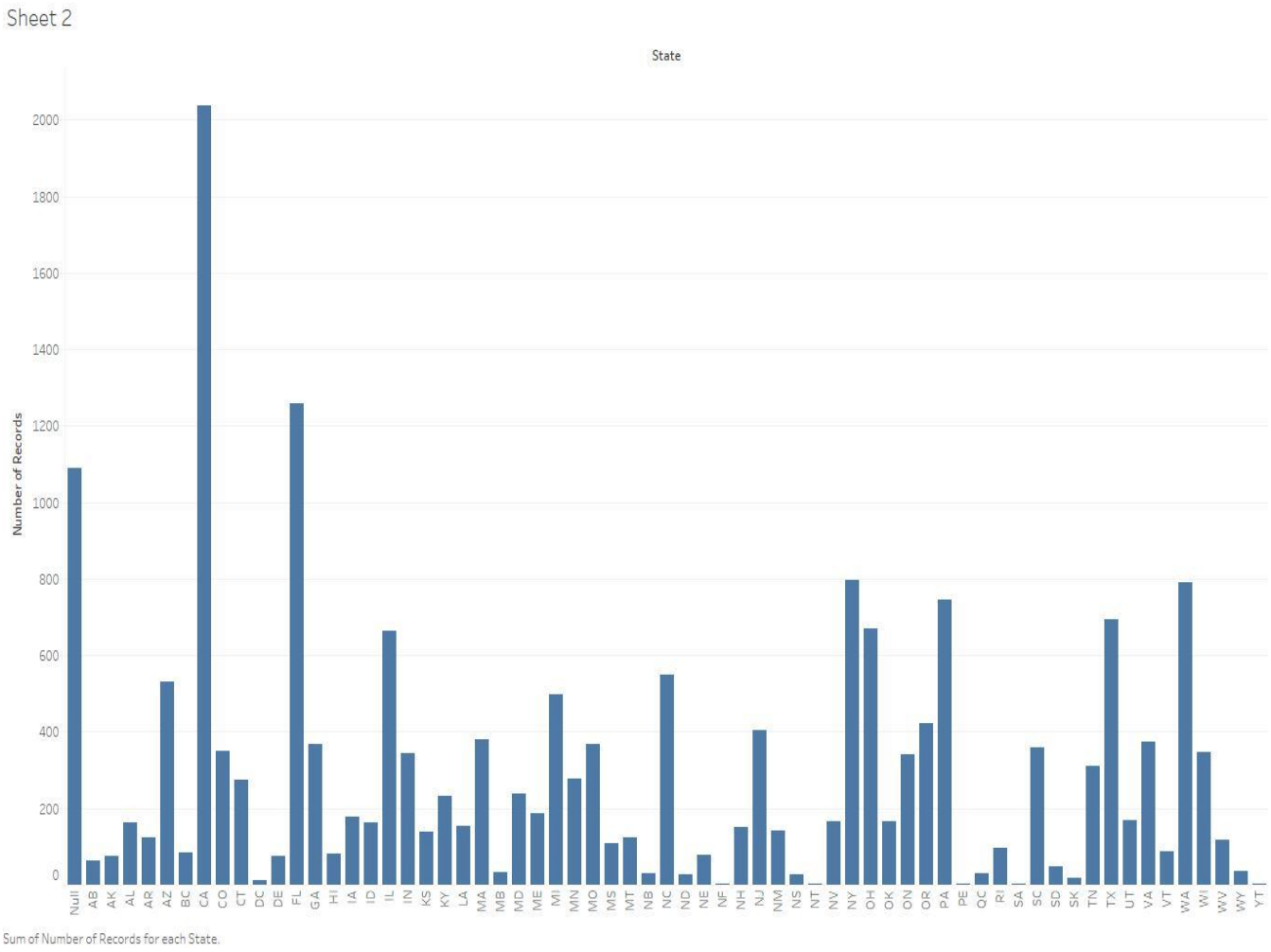
Year	Circles	Triangle	Fireball
2016	487	424	345
2015	779	605	590
2014	1085	743	912
2013	935	628	986
2012	898	670	1059
2011	536	542	537
2010	424	443	368
2009	472	563	262
2008	425	526	268
2007	432	543	221
2006	330	463	209
2005	352	576	220

TIME SERIES



Q: A bar chart for sightings by state:

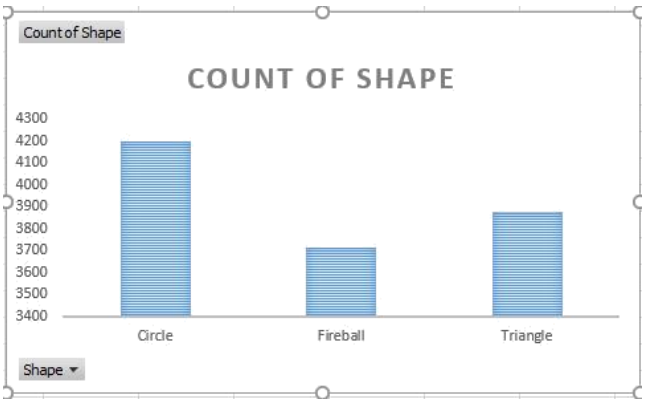
BAR CHART:



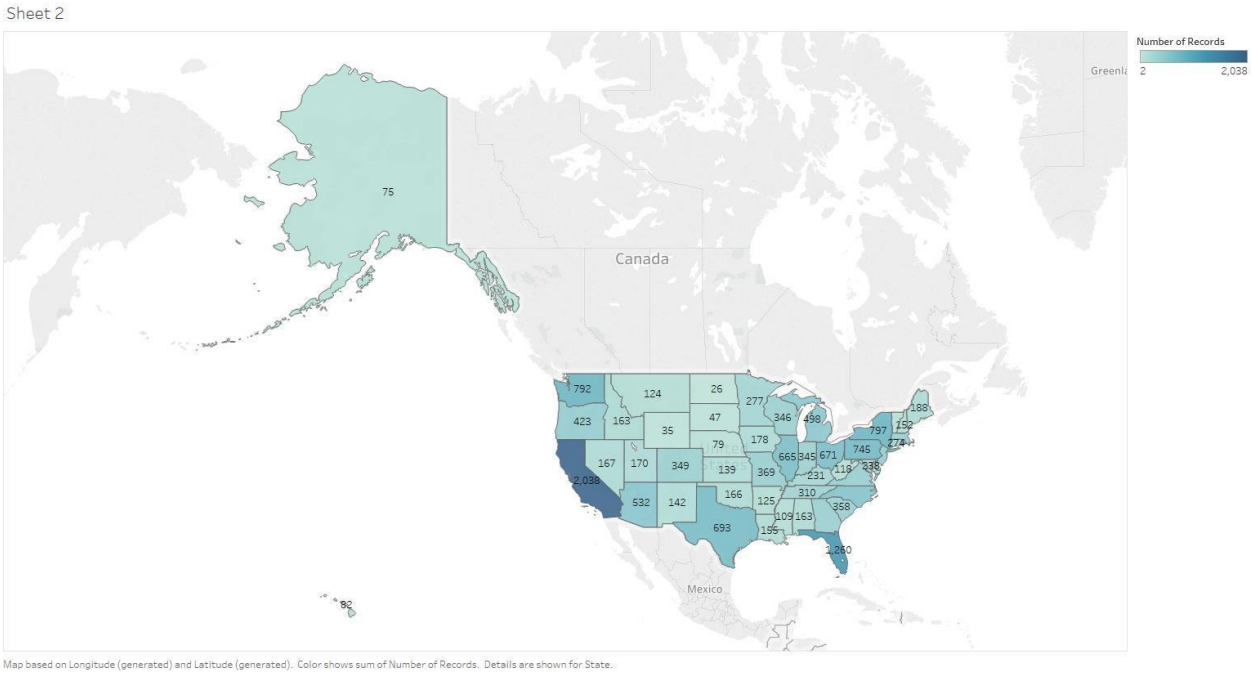
Q. In addition, you should also identify some other interesting insights from the data. You are encouraged to explore the data based on your own intuitions, but you are required to ask and answer at least one additional question beyond the basic data analysis we require above.

a) we can analyse with the above data that the number of sightings are maximum in California (2038) and Florida (1260).

b) we can also say that number of sightings for CI RCLE shape has always been maximum:

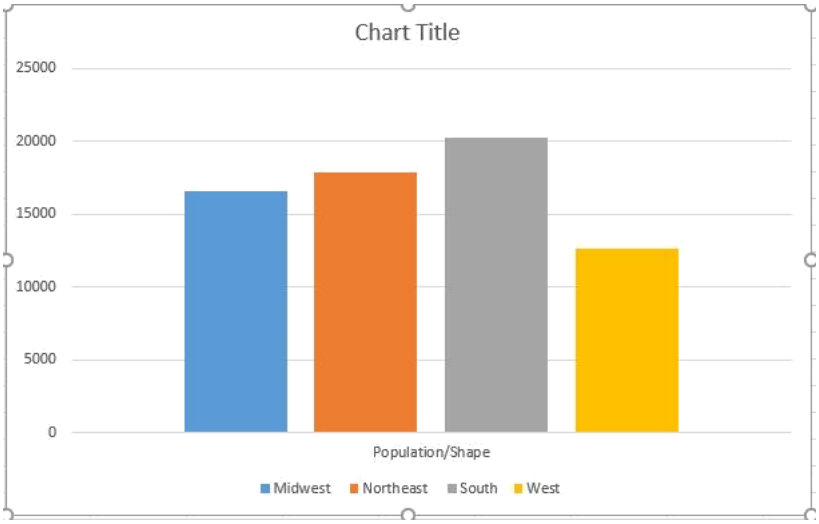


c) We can also visualize this entire data on the physical map as below:



d) we can also divide our data in 4 regions of USA and analyze the same based upon the population of each region:

Region	Shapes	Population	Population/Shape
Midwest	3948	65.37m	16557
Northeast	3127	55.94m	17889
South	5661	114.6m	20243
West	4504	56.9m	12633



- Here we can see that sightings, when normalized with the population of regions, gives us a better picture.
- Sightings per Person in Northeast are more than WEST. However, by the count sightings are more in the West than Northeast (in above table).

Task 2: Predicting UFO Shape

The given data is modified on the basis of below points:

- Training data for creating the decision tree has been selected from Jan 2005 till Dec 2013
- Test data to calculate the accuracy of the system has been selected from Jan 2014 till Sep 2016
- Time of sightings is divided into following 4 categories – Morning, Evening, Afternoon, Night
- Region of sightings is divided into 4 categories: Midwest, Northeast, South, West
- Below table gives us the entire picture of Training data and its GI NI I ndex

Time	Region	CI RCLE	FI REBALL	TRI ANGLE
Night	Midwest	146	111	193
Night	West	223	136	213
Night	South	216	145	281
Night	Northeast	130	69	101
Morning	Midwest	64	52	72
Morning	West	109	42	74
Morning	South	133	46	97
Morning	Northeast	57	38	42
Afternoon	Midwest	73	34	60
Afternoon	West	162	53	83
Afternoon	South	149	73	84
Afternoon	Northeast	70	31	53
Evening	Midwest	630	733	676
Evening	West	660	646	628
Evening	South	870	876	820
Evening	Northeast	509	587	399

GINI INDEX of entire training data: [1-(PC) ^2-(PT) ^2-(PF) ^2)] = ***0.665635***

PC - Probability of Circle

PF - Probability of Fireball

PT – Probability of Triangle

Now, we need to decide which should be the root node in our decision tree based on the GI NI I ndex:-

Taking 'TIME of Sighting' attribute as the root node:

TI ME as the root	Circle	Fireball	Triangle	total	PC	PF	PT
Night	715	461	788	1964	0.364052953	0.401222	0.2347251
Morning	363	178	285	826	0.439467312	0.21549637	0.3450363
Afternoon	454	191	280	925	0.490810811	0.20648649	0.3027027
Evening	2669	2842	2523	8034	0.332213094	0.35374658	0.3140403
Total	4201	3672	3876	11749			

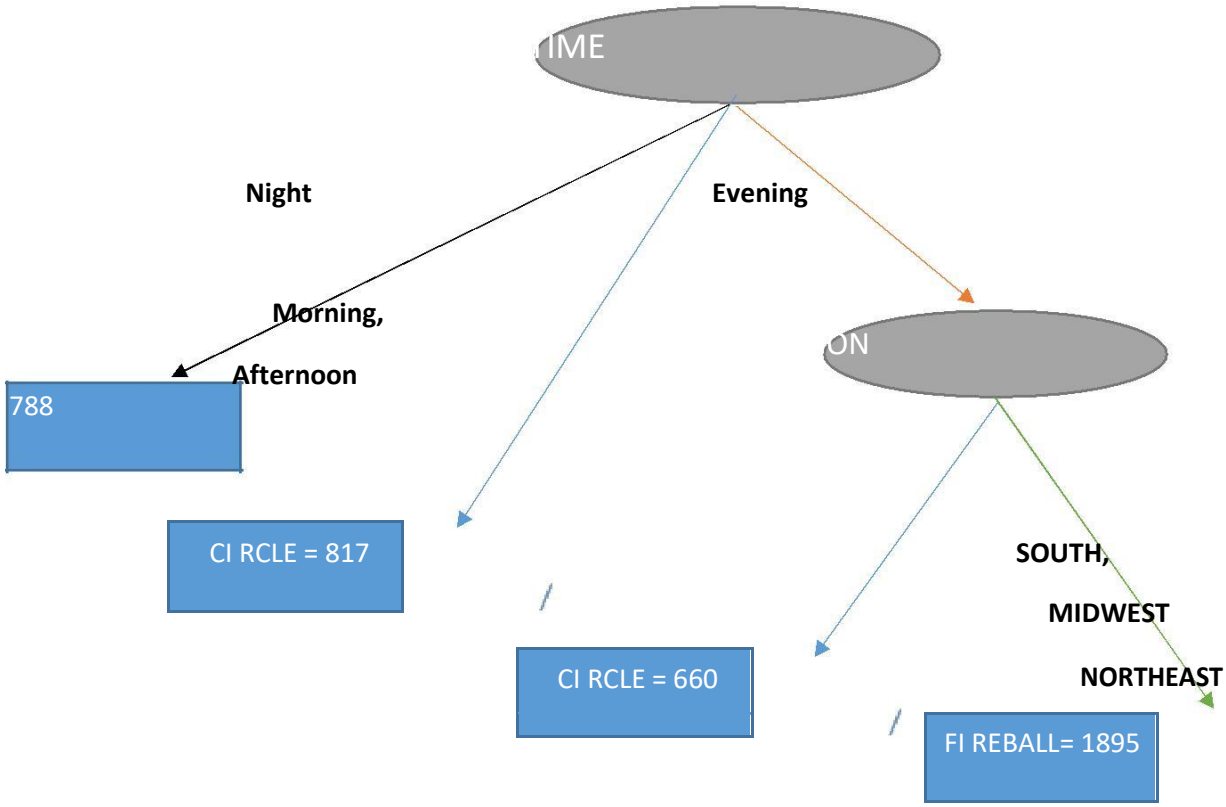
GINI INDEX – 0.3363

Taking 'REGION attribute as the root node:

REGI ON as root	Circle	Fireball	Triangle	total	PC	PF	PT
Midwest	913	930	1001	2844	0.321026723	0.32700422	0.3519691
West	1154	877	998	3029	0.380983823	0.2895345	0.3294817
South	1368	1140	1282	3790	0.360949868	0.30079156	0.3382586
Northeast	766	725	595	2086	0.367209971	0.34755513	0.2852349
Total	4201	3672	3876	11749			

GINI INDEX – 0.366

- Based on the above GI NI values, we can infer that TIME as the root node has a better accuracy (low GI NI value) than REGION as the root node.
- The best possiblescenario on the basis of GI NI impurity values, gives us the below decision tree:


IME
Night
Evening
Morning,
Afternoon
788
CI RCLE = 817
ON
SOUTH,
MIDWEST
NORTHEAST
CI RCLE = 660
FI REBALL= 1895

Classification accuracy of the DECISION TREE using TEST Data:

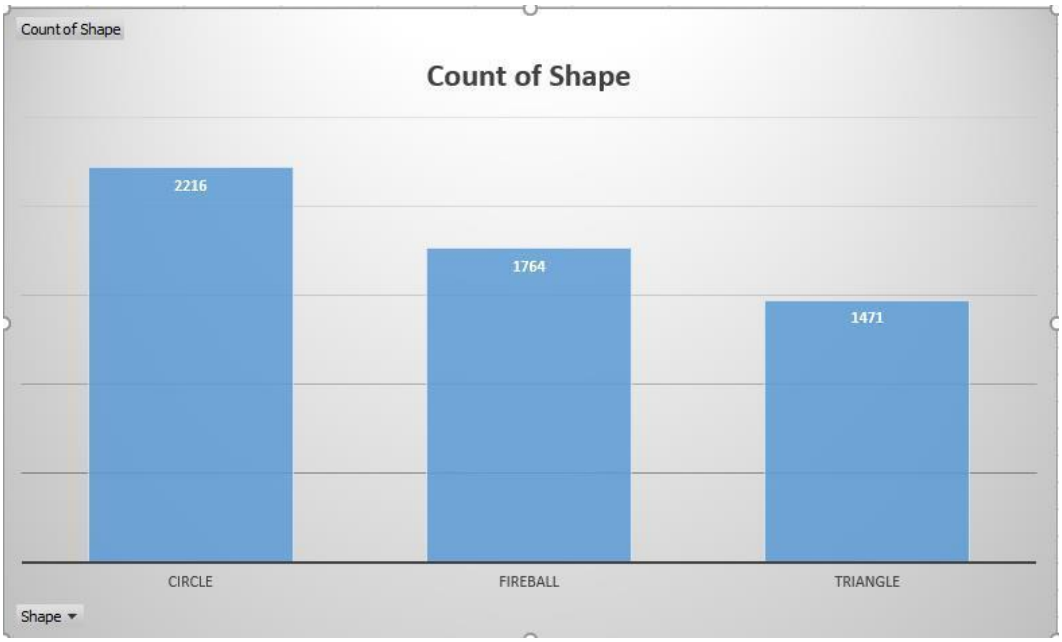
With the help of above Decision Tree, we can predict the shape of UFO by providing the test data (only 2 attributes Region and Time) as below:

1. **Example 1** (Evening, Midwest) → FIREBALL (using above tree) -- correct
2. **Example 1** (Evening, WEST) → CIRCLE (using above tree) -- incorrect
3. **Example 1** (Morning, Midwest) → CIRCLE (using above tree) -- correct
4. **Example 1** (Night, Northeast) → TRIANGLE (using above tree) -- correct
5. **Example 1** (Afternoon, South) → CIRCLE (using above tree) -- incorrect

- Above observation shows that the decision tree *does not always give* the correct values when we test it with our test data

Summary of all instances of test data:

Total instances in test data – 5452



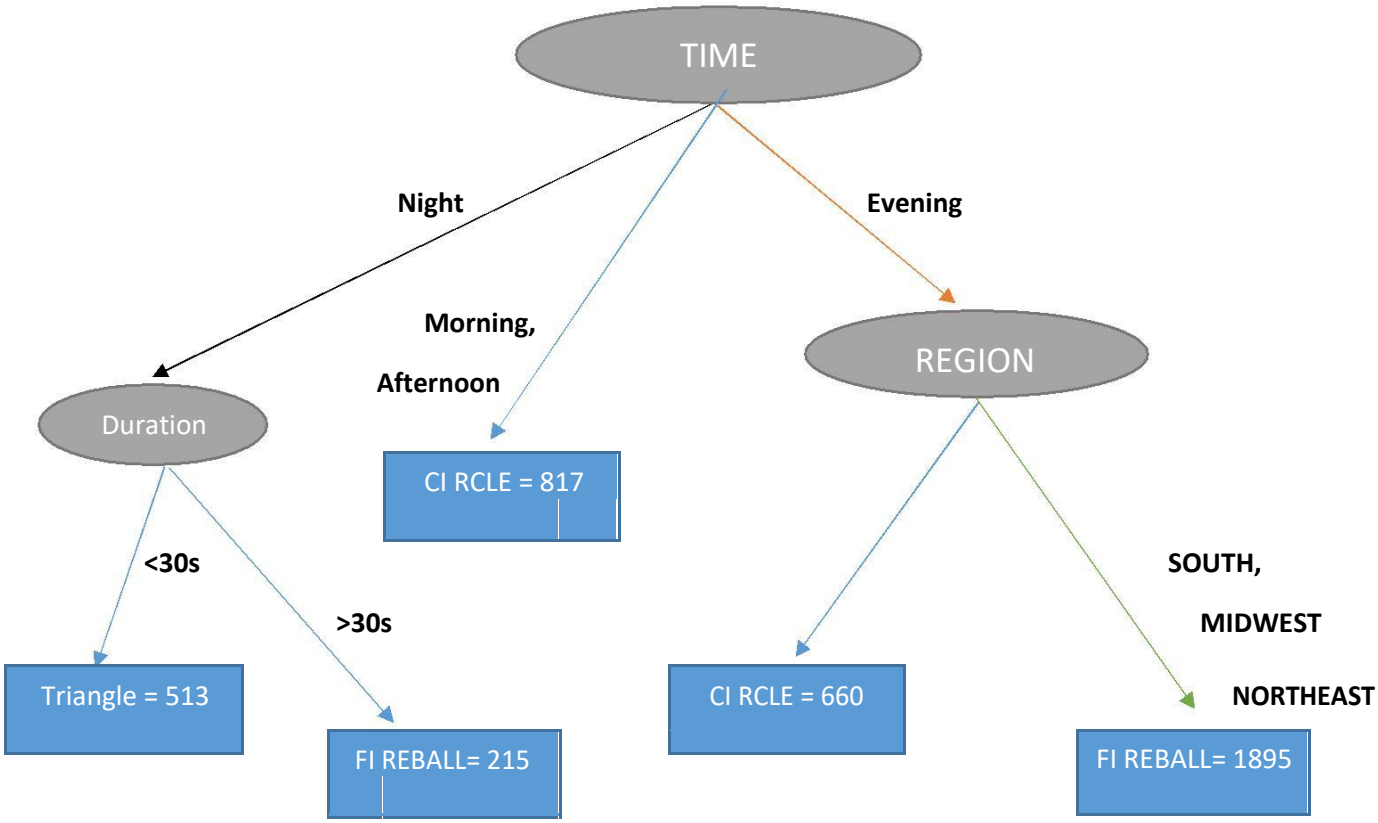
Correctly Classified Instances	2310	**42.3734** %
Incorrectly Classified Instances	3142	**57.6266** %

ACCURACY ~ 43%

Task 3: Can you improve your prediction rate (accuracy) over what you got from Task 2? You may use raw features instead of the two features or even combining all features. Or something else?

- The accuracy can be increased by adding more attributes to our decision tree. We can see that whenever the TI ME is *'Night',* our decision tree gives the prediction as **TRIANGLE.** However, if us we add some more attribute to our training data, ***duration of sighting,*** it will improve the like – prediction more.
- Example: Sightings during the NI GHT with duration more than 30seconds are Triangles and Sightings with duration < 30s are Fireballs. This will add one more branch to our decision tree but the prediction percentage will improve.

• Describe how you can achieve better result compared with Task 2's result. • Report what result you got.



- **Total instances in test data – 5452**
  - Correctly Classified Instances 2597 **47.64** %
  - Incorrectly Classified Instances 2855 **52.36** %

- **ACCURACY ~ 47.6%**

- **What feature is the most important feature to distinguish shapes of UFOs?**

    - TIME of sighting is the most important attribute to distinguish the shape of UFO. It has the lesser GINI impurity value as compared to the REGION attribute.
    - GINI (time) = 0.3363
