Homework #2
CS 5665, Fall 2016

1. Cleaning and Extracting data:

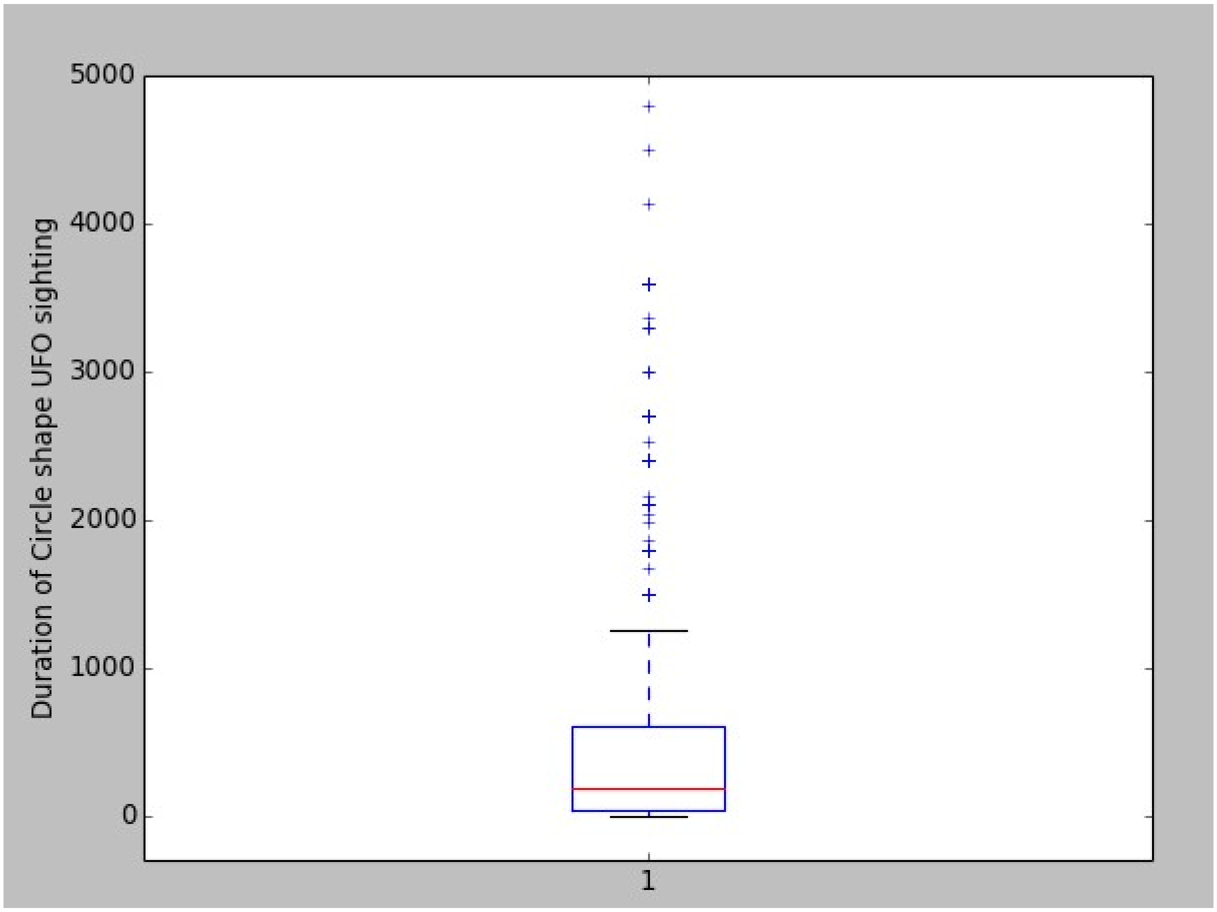
+ Cleaning : Below is the replacement I performed for cleaning duration data.

```
value.replace("<","")
value.replace(">", "")
value.replace("~", "")
value.replace("-", " ")
value.replace(".", "")
```

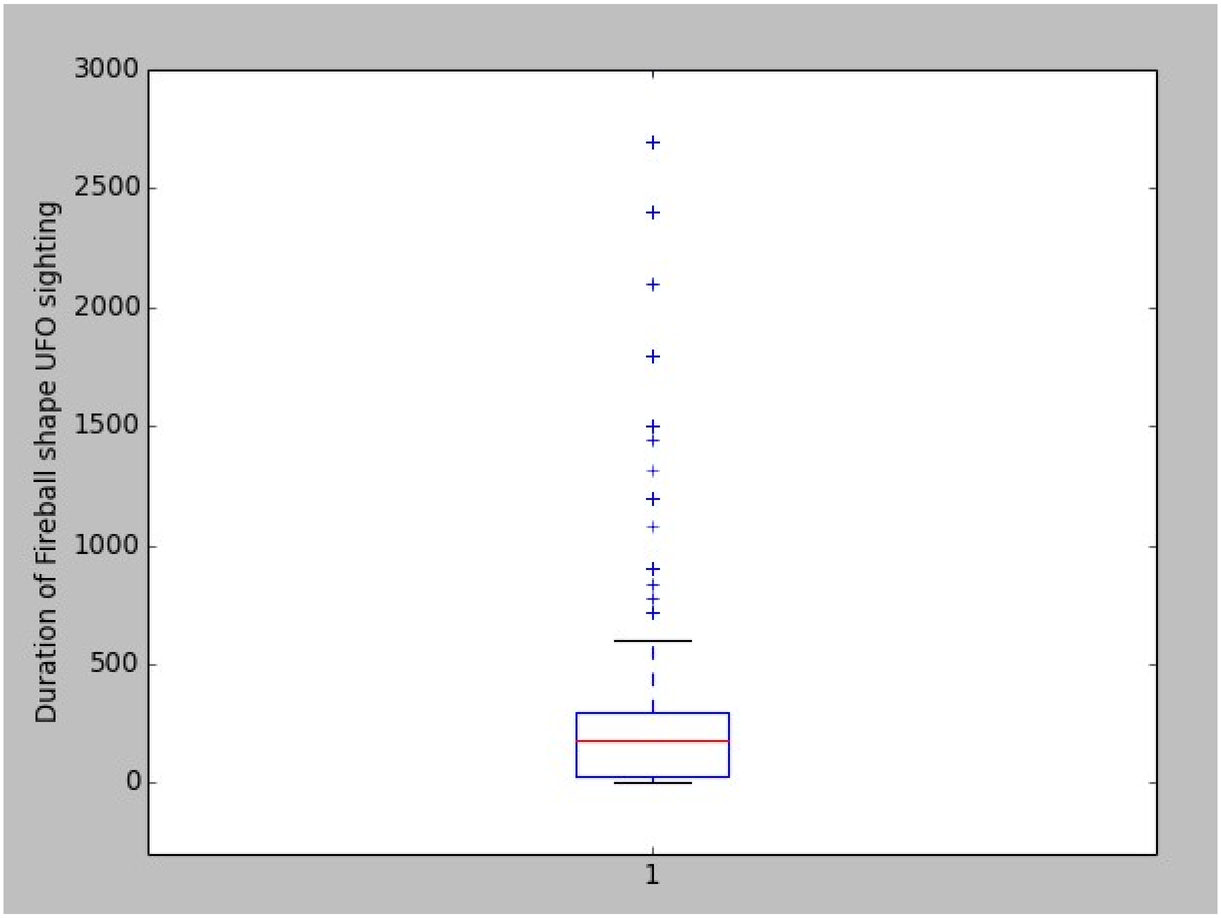
+ Consideration of several time span notation and converting everything to seconds

```
seconds = ["sec","seconds","secs","second"]
minutes = ["min","minutes", "minute","mins"]
hours = ["hours", "hours","hr"]
```

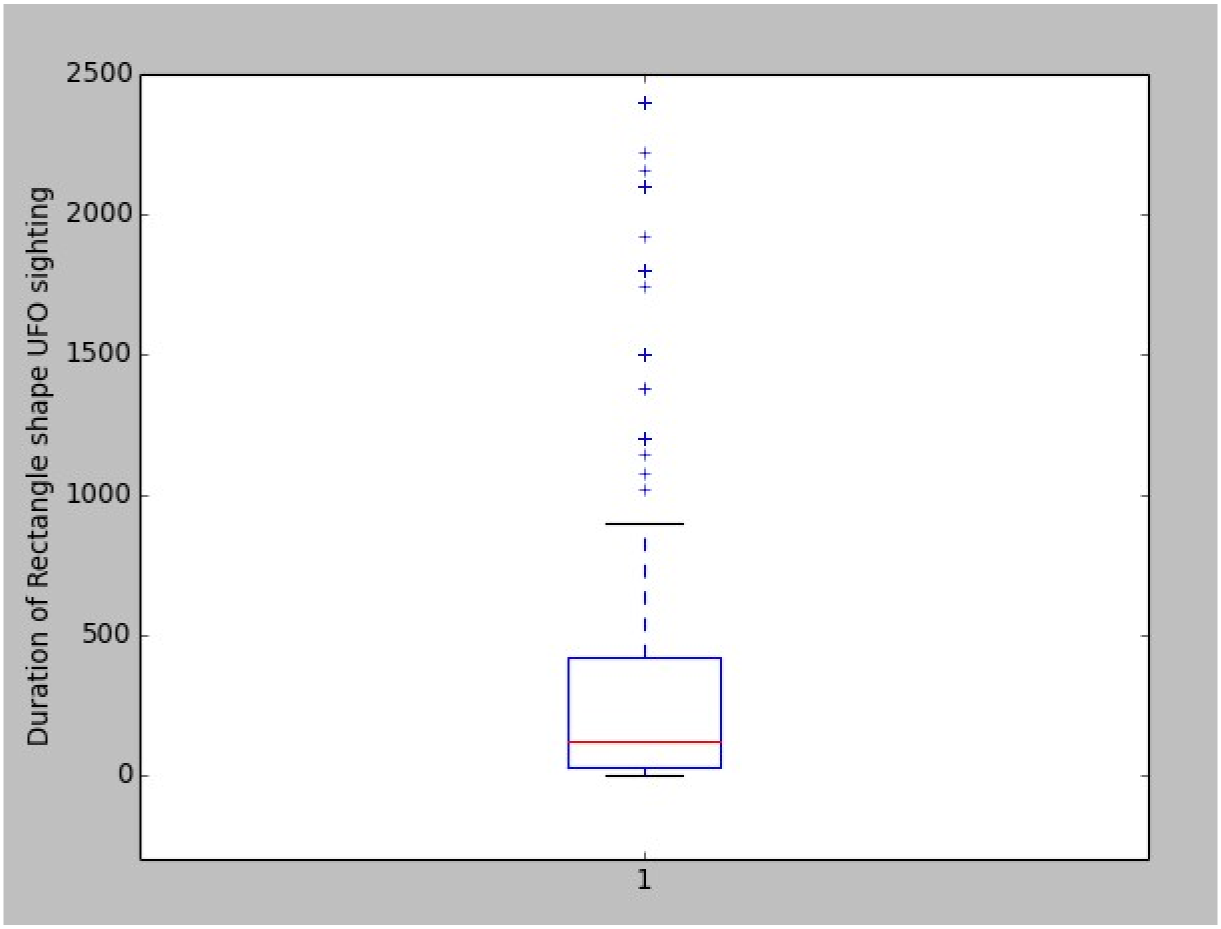
+ Given two numbers in a duration 5-8, considering 8.



Mean = 1014.89924623
Median = 180.0
Mode = 300.0

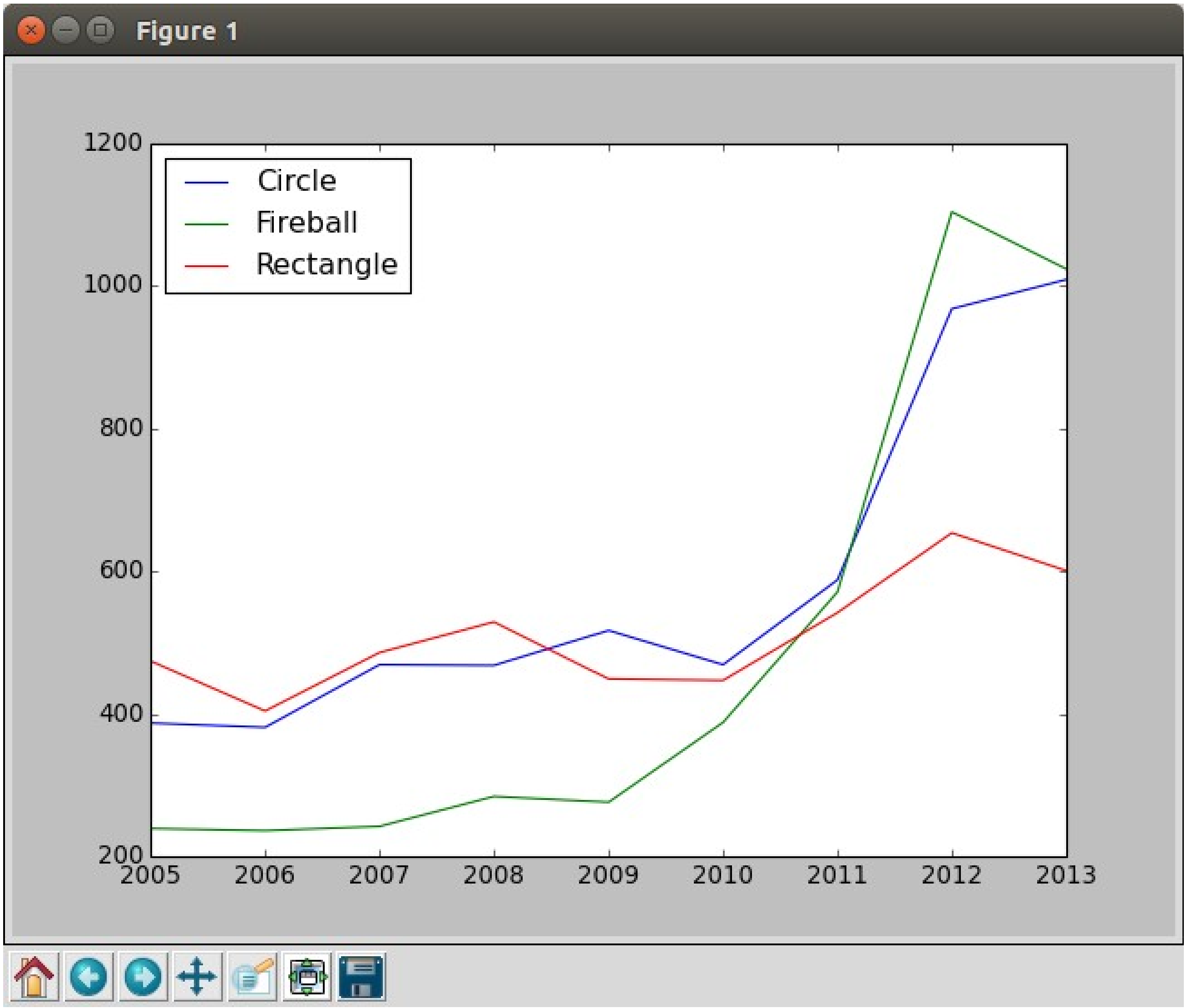


Mean = 473.277380297
Median = 180.0
Mode = 300.0

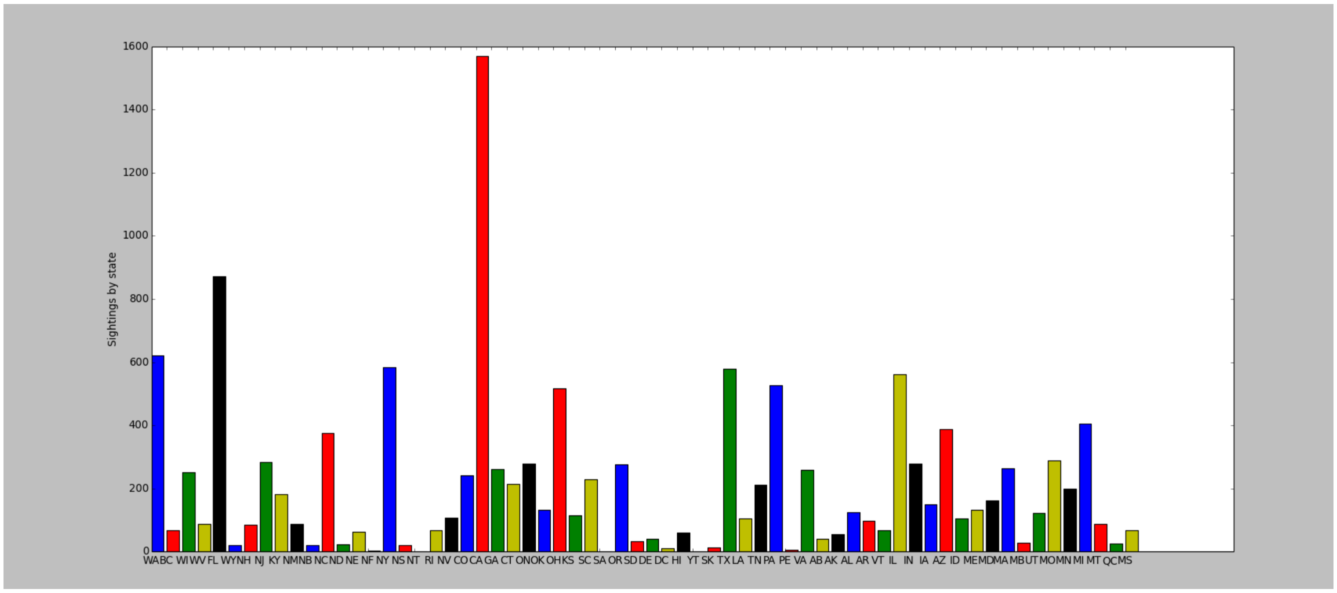


Mean = 920.24880248
Median = 120.0
Mode = 300.0

+ A time series figure with the number of sightings per year (one line per shape).



+A bar chart for sightings by state.



+ Observation was California has highest number of sighting
+ According to the data there are 64 states in USA.
+ There were some give location name, which doesn't not exists ex. NF, SA, YT, PE in USA states list.

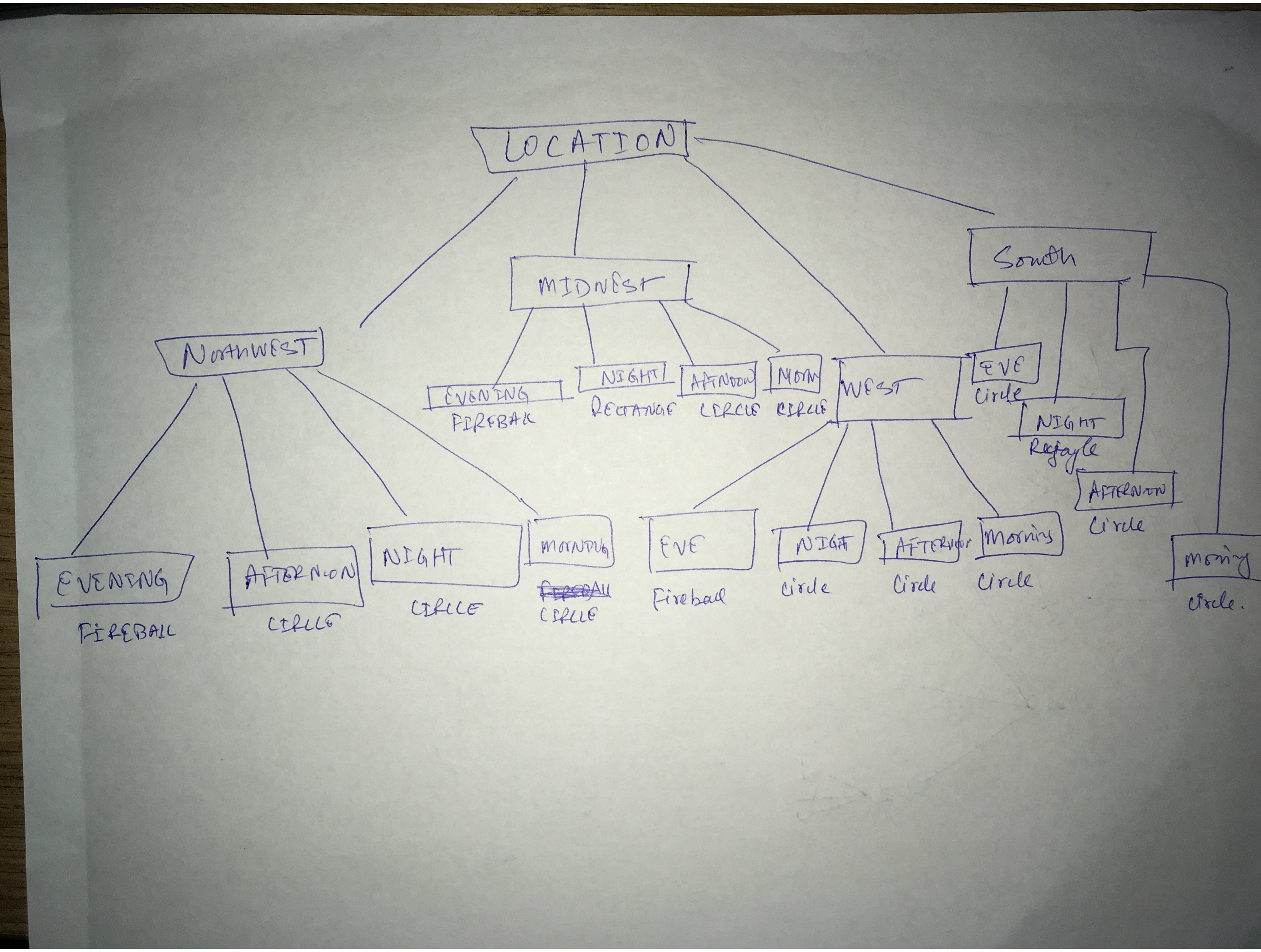
2. Report accuracy of the decision tree classifier using Gini Impurity

No of Circle data 2372
No of Rectangle data 1545
No of Fireball data 1833
Total number of reords 5750
True Positive 2251
Accuracy 0.39147826087 ~ 39%

Gini gain for
Location : 0.10857160268
Time : 0.09864254445

So came to conclusion to split on Location first and then Time. Below is the tree/rule

```
## RULE/Tree
# "NORTHWEST","EVENING","FIREBALL"
# "NORTHWEST","AFTERNOON","CIRCLE"
# "NORTHWEST","NIGHT","CIRCLE"
# "NORTHWEST","MORNING","CIRCLE"
# "MIDWEST","EVENING","FIREBALL"
# "MIDWEST","NIGHT","RECTANGLE"
# "MIDWEST","AFTERNOON","CIRCLE"
# "MIDWEST","MORNING","CIRCLE"
# "WEST","EVENING","FIREBALL"
# "WEST","NIGHT","CIRCLE"
# "WEST","AFTERNOON","CIRCLE"
# "WEST","MORNING","CIRCLE"
# "SOUTH","EVENING","CIRCLE"
# "SOUTH","NIGHT","RECTANGLE"
# "SOUTH","AFTERNOON","CIRCLE"
# "SOUTH","MORNING","CIRCLE"
```



Wanted to share wrote a generic function which can be used to calculate Gini impurity:

```
##  Find Occurence
def occurence(data, conditionLoc, conditionTime, conditionTruth):
    count = 0
    for row in data.iterrows():
        if conditionLoc != "" and conditionTime != "":
            if row[1][0] == conditionLoc and row[1][1] == conditionTime and row[1][2] ==
conditionTruth:
                count += 1
        elif conditionLoc != "":
            if row[1][0] == conditionLoc and row[1][2] == conditionTruth:
                count += 1
        elif conditionTime != "":
            if row[1][1] == conditionTime and row[1][2] == conditionTruth:
                count += 1
    return count
```

Using below value can be calculated, which is more like writing queries:

```
countOccurence(df,"NORTHWEST","EVENING","CIRCLE")
```

3) I didn't implement/code the improvement in the tree:

+ My idea was to include text data as another feature
+ Analyzing text which contains Circle or Rectangle or Fireball.
+ I am positive it will improve accuracy since it will improve the GINI Gain by giving homogeneity, but I don't have any concrete evidence to show it.