Analyzing UFO Sightings

Shealagh Brown & Sam Zimpfer

2025-05-01

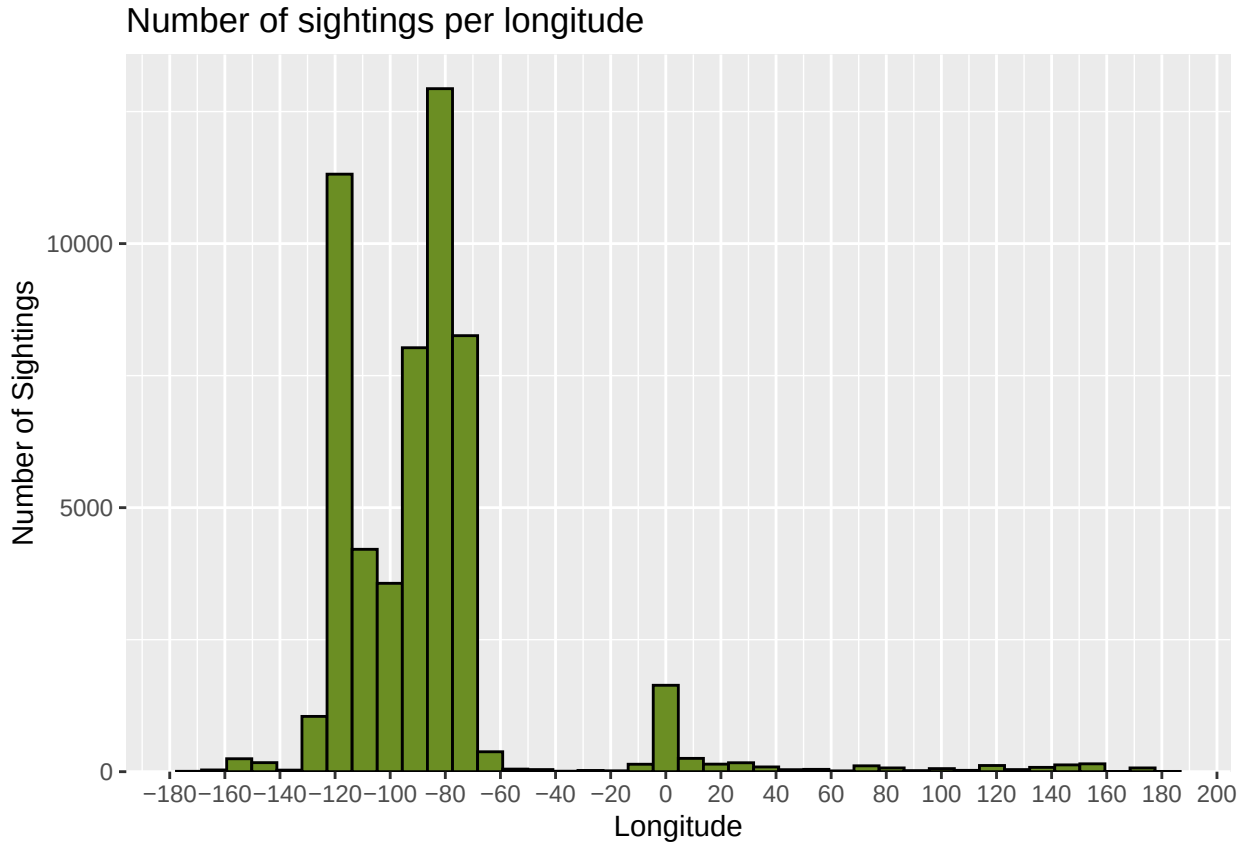
Introduction

The data we were working with came from a data set called "UFO sightings scrubbed" that was found on Kaggle.com from a user named Akhil Goyal. The data was last updated three months ago, making it quite recent data. It contains information on all UFO sightings dating back to 1906. The data could have some bias if ufo sightings from certain regions of the world were not recorded or included in this data set, additionally it is observational data collected by different people around the globe which can create large amounts of variation in the data.
This data is of interest because UFOs have been a topic of public debate for years. With increasing amounts of interest in space travel and extraterrestrials in more recent years, the fascinations with UFOs has only grown stronger. For this project we want to explore what influences sightings as this can be valuable knowledge for those trying to investigate UFOs.
In order to work with our data we had to clean it. This included converting the datetime column into year, month, day, seconds, minutes, hours format. Then we created a new data set where we added columns for years, seconds, and months and kept the updated datetime, city, state, country, longitude, and latitude columns. We then had to convert both longitude and latitude into numeric values in order to work with them. The code for this data cleaning follows:

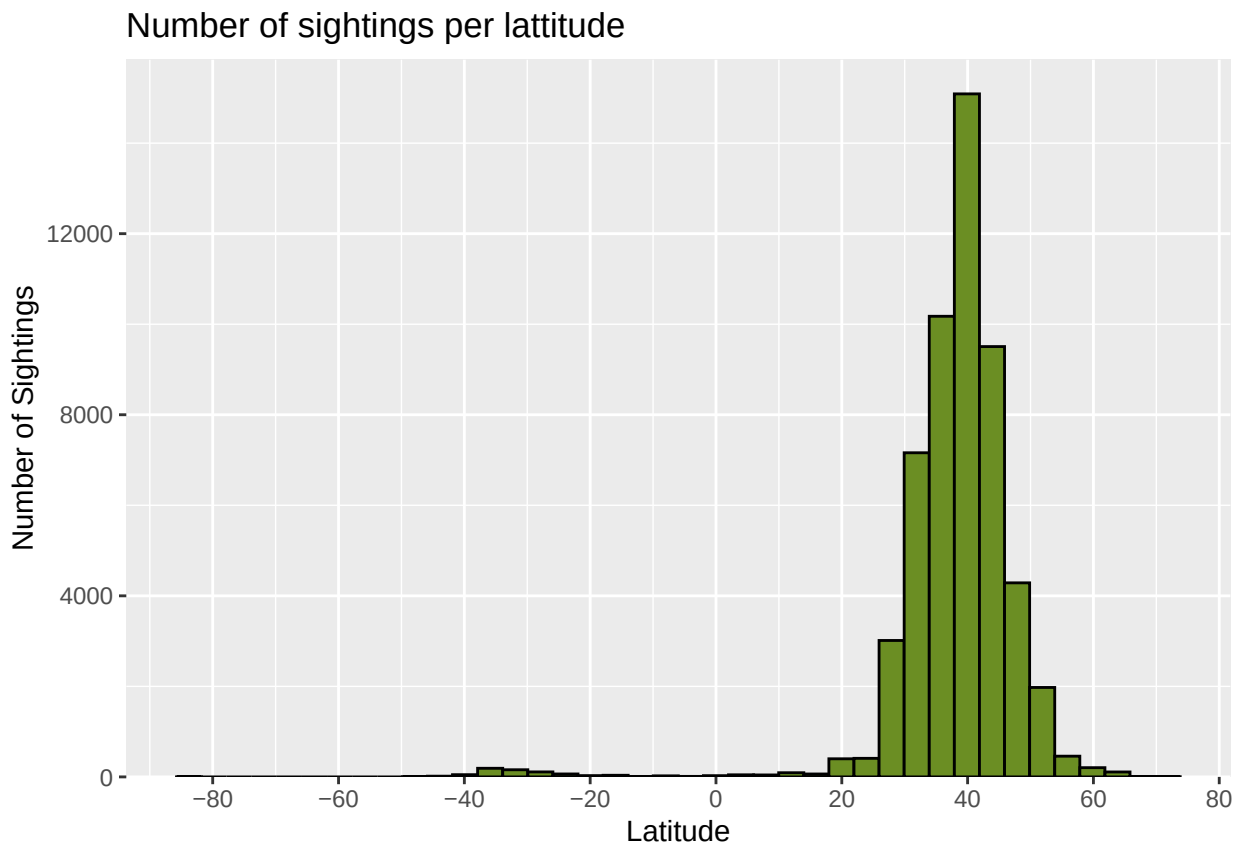
```
ufo_raw <- read.csv("ufo_sightings_scrubbed.csv")

#convert date time to ymd_hms format
ufo_raw$datetime <- ymd_hms(ufo_raw$datetime)

#cleaning data
ufo_raw|>
   mutate(seconds = duration..seconds.,
          year = year(datetime), #create year column
          month = month.name[month(datetime)])|> #create month column and convert to name
  select(datetime, city, state, country, seconds,latitude,longitude, year, month)|>
    filter(seconds <= 40000) |>
    # filter out badly formatted entries that could cause NA's during the following conversion
    filter(grepl("^[-]?[0-9.]+$", latitude),
           grepl("^[-]?[0-9.]+$", seconds)) -> ufo


ufo$latitude <- as.numeric(ufo$latitude) #changing lat to numeric
ufo$seconds <- as.numeric(ufo$seconds) #changing seconds to numeric

ufo <- ufo |> drop_na(longitude, latitude, seconds)
```

Data Analysis

The first question we had about UFO sightings is how the location around the world influenced the number of UFO sightings that occurred. We began by creating histograms for longitude and latitude. From there we created a scatter plot to compare latitude and longitude to see if there was a relationship between the two of them.



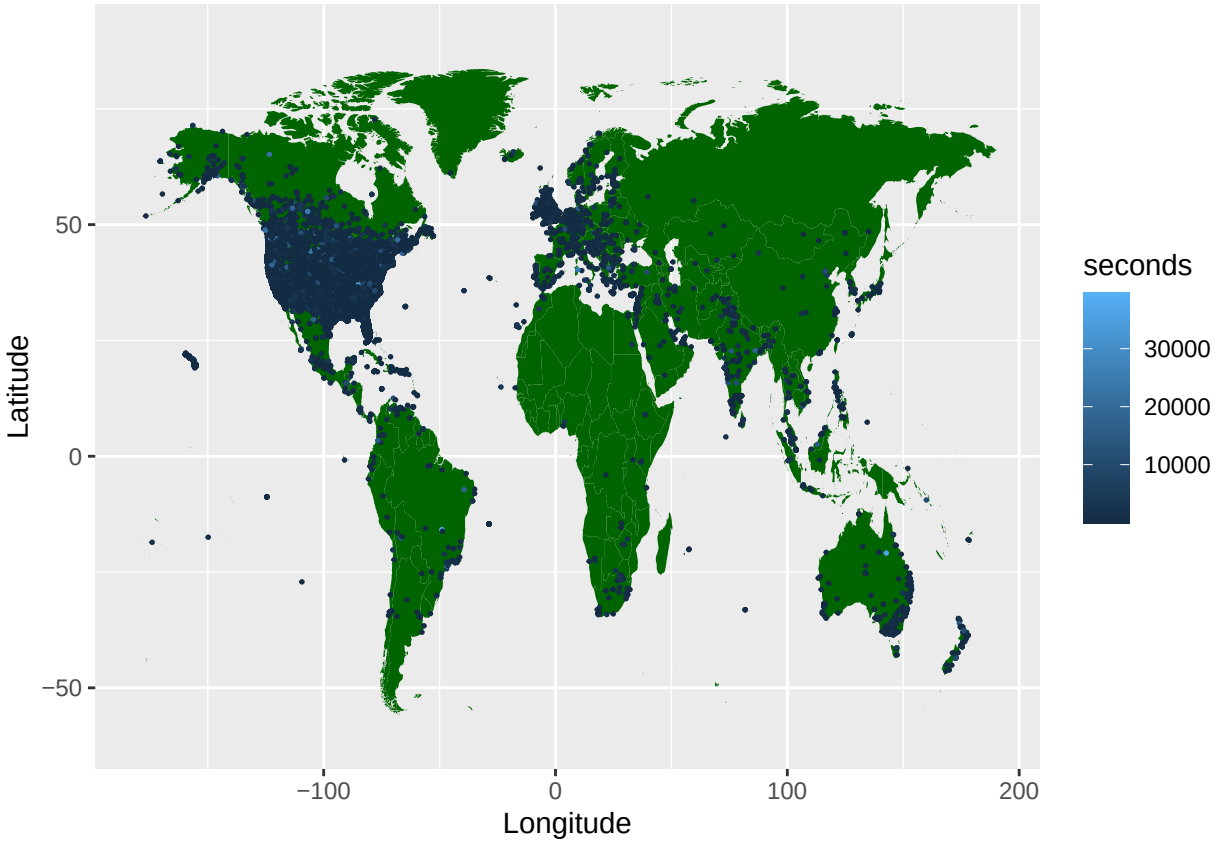
When looking at the histogram of longitude, we notice there are two main peaks within the spread. One peak is around -120 and the seconds is around -80. This make sense because these are the longitudes that correspond with highly populated areas of the US. The relative height of these peaks implies that there are considerably more sightings in the US than anywhere else in the world. This observation can be interpreted in multiple ways, either that UFO's are more commonly reported in the US than anywhere else (either accurately or inaccurately), or that there really are more UFO visits in the US than anywhere else. We can't make a solid determination between these two based on the data, but we can clearly see that there have been more reports around the US.



In the latitude histogram, we see than sightings are most prevalent around 40. Again, this corresponds to the coordinates of the US and also Europe, which is the second most frequently reported area of UFO sightings. This data supports the same conclusions we see from the longitude histogram.

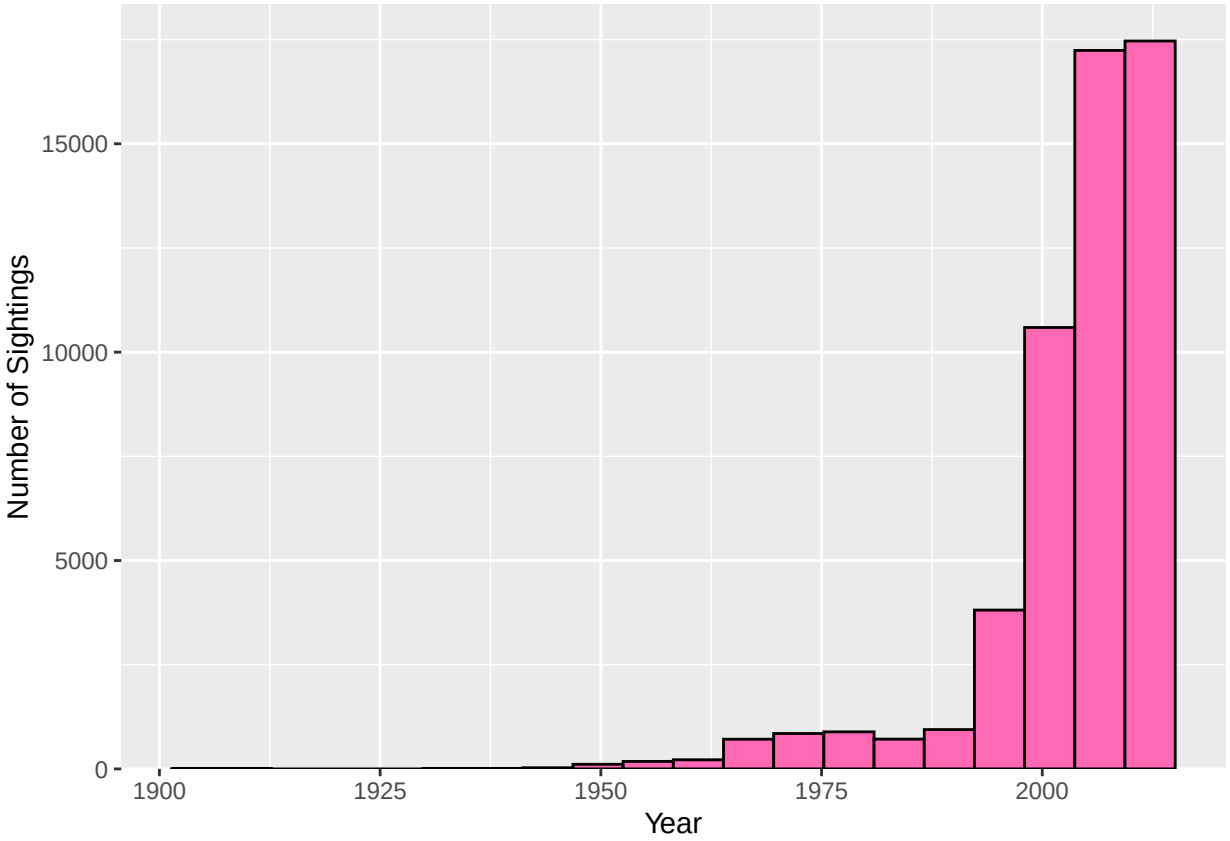
```
## Scale for y is already present.
## Adding another scale for y, which will replace the existing scale.
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



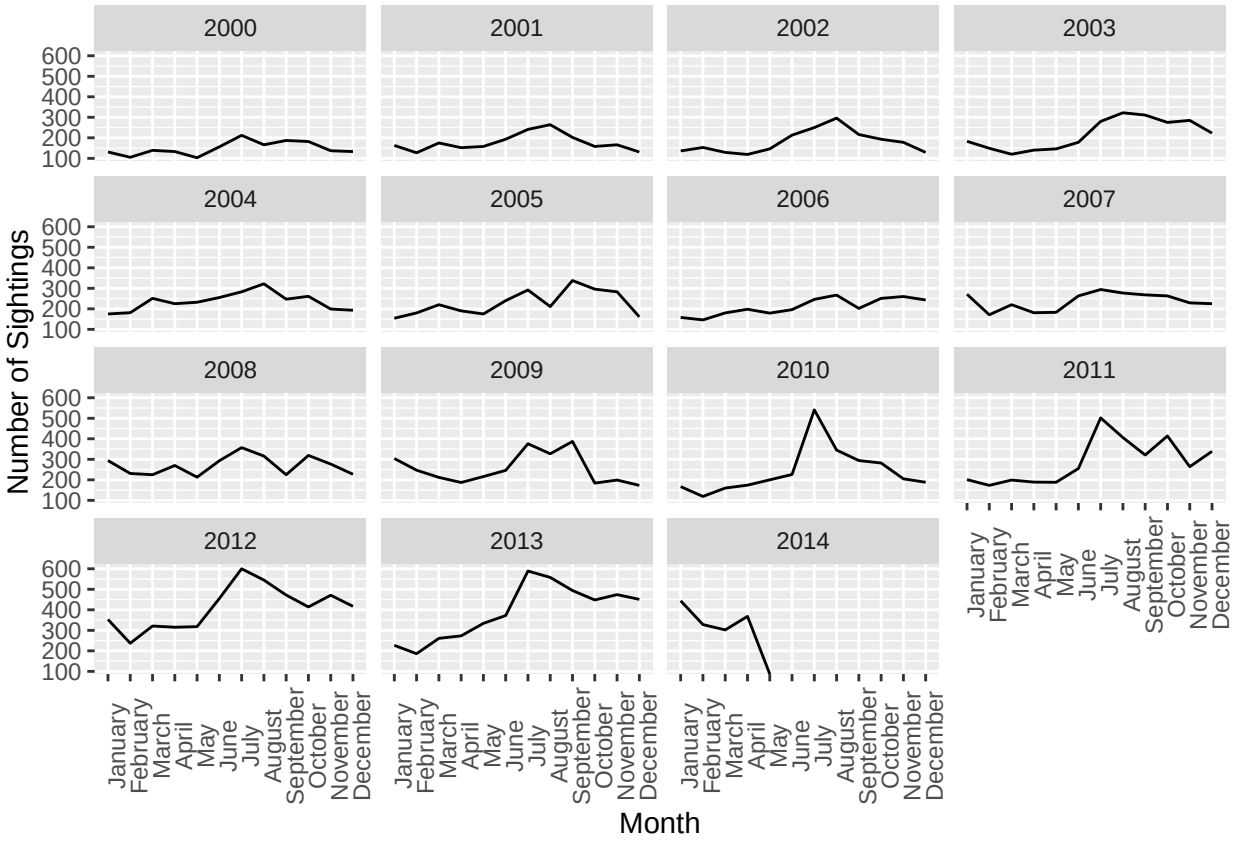
Here, we can see in greater detail the geographical dispersion of the sightings. The graph supports the same claims as the previous two graphs. We can also see from the color coding based on duration that most of the sightings are pretty short.

The next question we had was how the number of sightings had increase or decreased over the full epoch of time that the data set spans. We created a histogram of sightings per group of 5 years to answer this question.



The histogram shows that UFO sightings were relatively infrequent in the early 1900's. Sightings gradually started to increase during the 1950's and then there was a sharp increase at the end of the 1990's and into the 2000's. The number of sightings seems to start to plateau past 2005 also, so it would interesting to see whether the data plateaus in the long term after this point or if it would continue to increase beyond a certain point. This pattern might indicate developments in human history that could be of interest to extraterrestrial beings or have sparked human interest in the existance of extraterrestrials.

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```



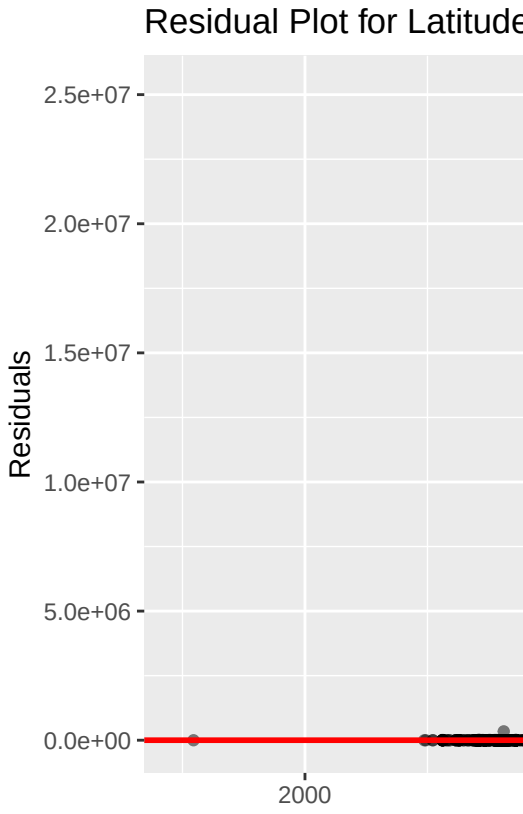
Here we have a small multiples graph which shows the seasonal trends in UFO data since the year 2000. Like the previous graph, these graphs also show that total UFO sightings have increased since 2000. Before 2009 sightings are generally consistent over the years. In 2009 we can see the first real peak during the summer. This trend seems to be reflected in the following years since then. It's possible that this is simply because more people are outdoors at this time of year, so more UFO visits are reported as sightings. Because of this highly plausible explanation, we can't necessarily determine from the data whether there was an actual increase in UFO visits in the summers.

Machine learning

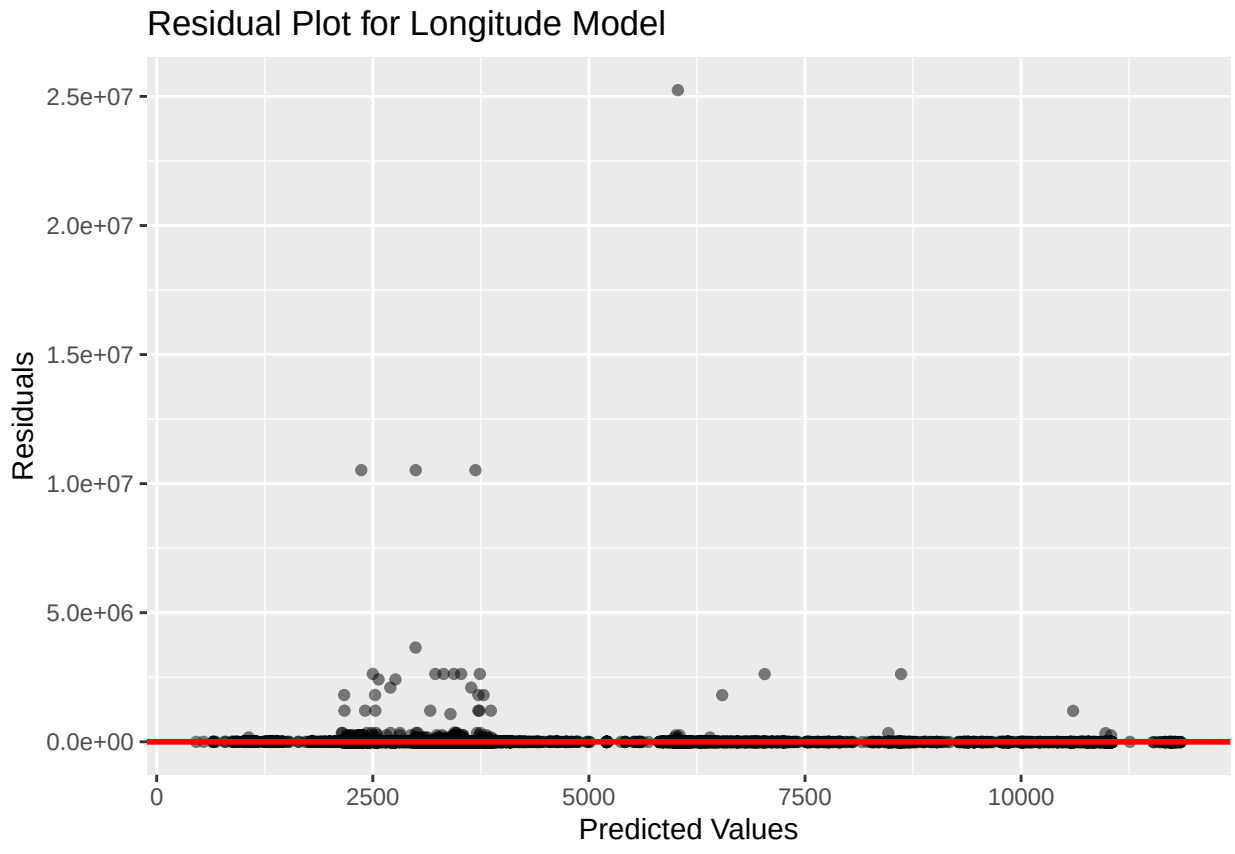
###Linear Regression

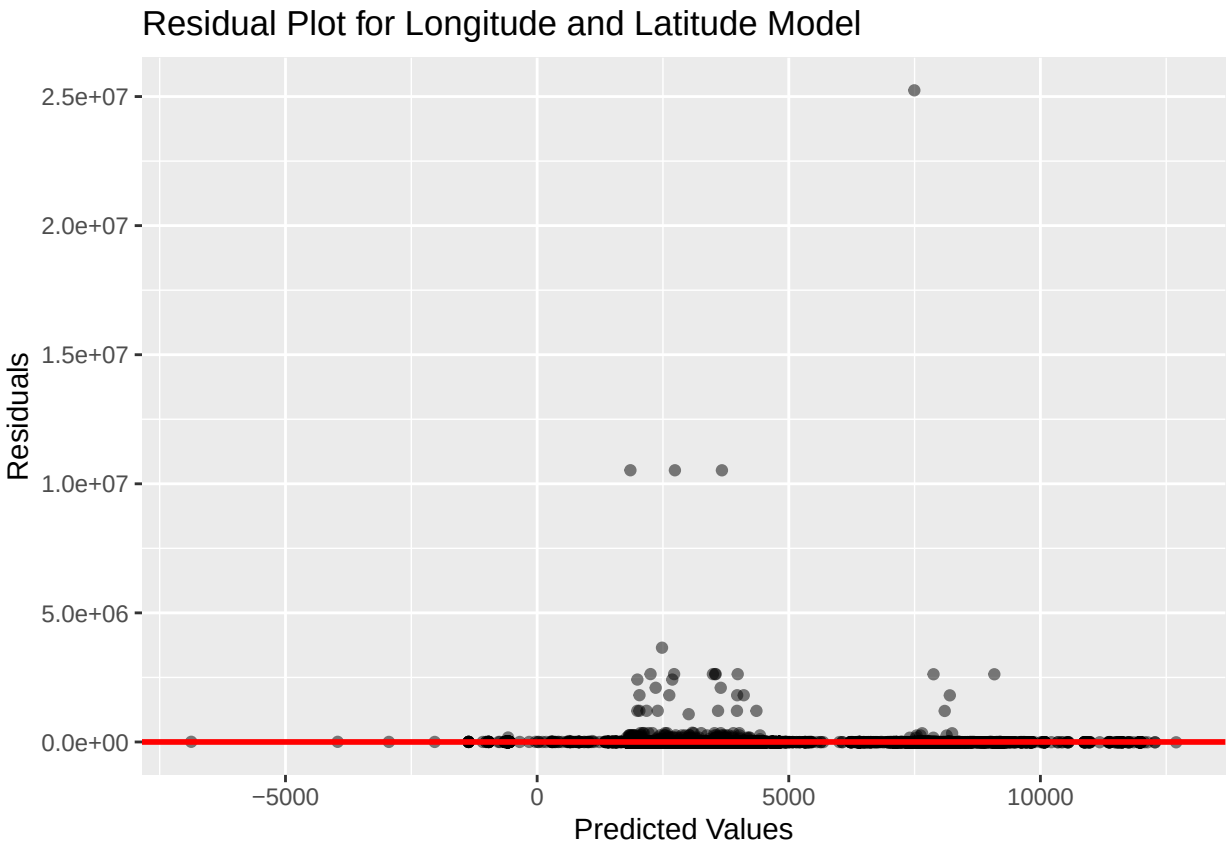
We wanted to use machine learning techniques to predict the duration of sightings and the longitude/latitude they took place at. We first chose to try to create a linear regression model. We thought this would be appropriate as we wanted to explore how our explanatory variables could explain the duration of sightings which is a numerical response variable. We also thought this would be a good technique to use as it is an eager learner and we wanted to create a model.

```
## # A tibble: 3 x 4
##   explanatories          r.squared adj.r.squared   sigma
##   <chr>                      <dbl>         <dbl>   <dbl>
## 1 latitude             0.000000881    -0.0000177 142621.
## 2 longitude            0.0000784       0.0000598 142615.
## 3 latitude + longitude 0.0000999       0.0000627 142615.
```



Residual Plots for Duration of Sightings vs Longitude and Lattitude



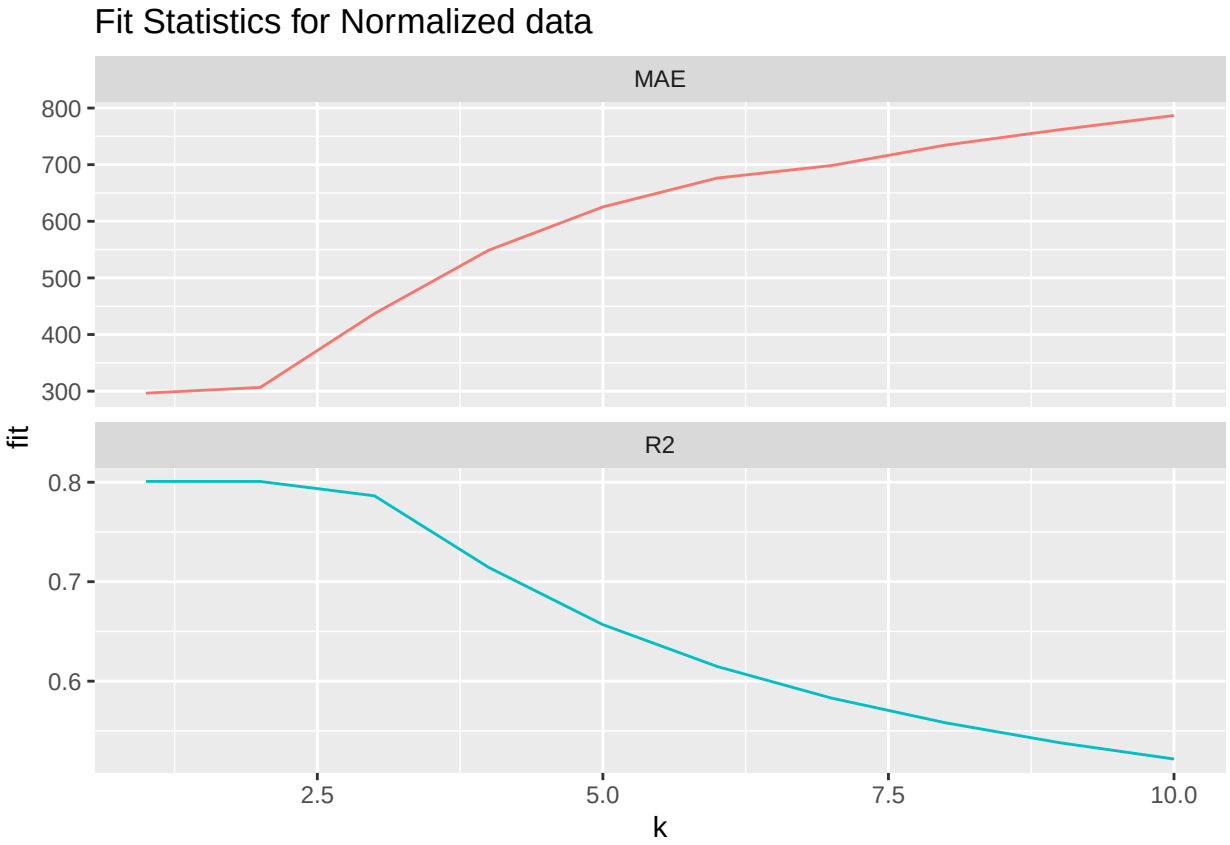


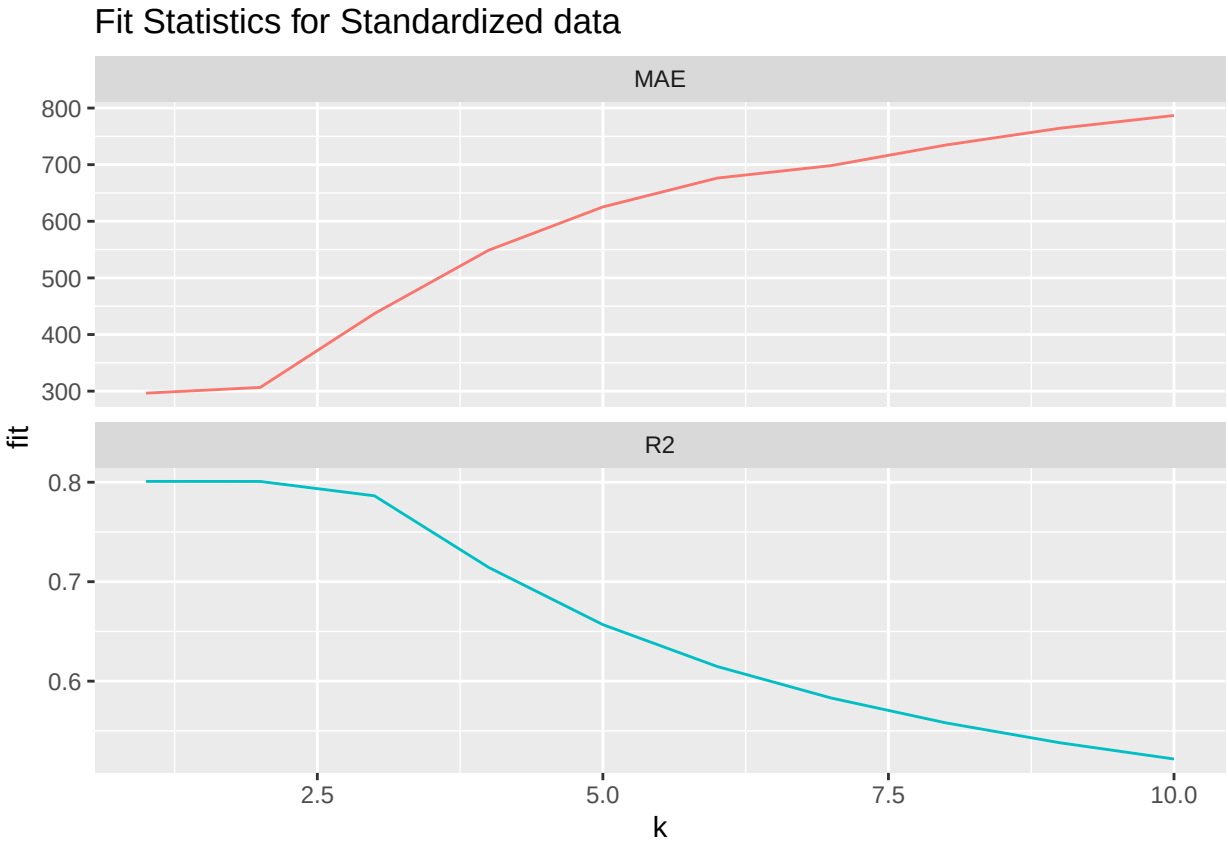
We created three different linear regression models, one with longitude as the explanatory variable, one with latitude as the explanatory variable, adn the third with both. We wanted to compare which model would be the most accurate. From the table above we can see that all three models had R-squared values that were very close to zero. This indicates that our models were not a good predictor of duration of UFO sightings.

We then made residual graphs to see what else we could notice about our models and the data. These plots also indicate that the linear model is not a good predictor of duration based on longitude or latitude. In each plot there are clusters of outliers that are significantly higher than the majority of the data.

kNN regression

Since our linear regression proved to not be effective in predicting duration of sightings based on longitude or latitude we decided to try a k Nearest Neighbors regression. We noticed that a large portion of the data was acting differently than the rest, and thought kNN regression could account for this.


Fit Statistics for Normalized data
MAE
800
700
600
500
400
300
R2
0.8
0.7
0.6
fit
2.5
5.0
7.5
10.0
k




```
##   rescale k       R2     MAE
## 1    norm 1 0.800855 296.486
```

When looking at the fit statistic plots for both the normalized and standardized we can see the R-squared values are at their max and the MAE is minimized at values close to one. We then searched the data to find the best k value, which ended up being a k value of one when using the normalized data. This k value is unusual and leads up to believe that this model is likely overfit to the data that we used to train it.

Both the linear regression and kNN regression failed to be good techniques for the effect of longitude and latitude on duration. This could be due to a lack of relationship between these variables. There is a large portion of the data where the sightings have a very short duration. There is also a second portion of the data that has much longer duration. Additionally, the many sightings are concentrated in the United States and Europe. These nuances of the data could have influenced their behaviors with regression.

kNN Classifications

We were also interested in using machine learning to see if we could use latitude and longitude to predict which month a sighting happened in. We started by using kNN classification because our response variable is categorical and we thought it would be an effective method of predicting month based on longitude and latitude.

```
## Confusion Matrix and Statistics
##
##           predicted
```

```
## actual      April August December February January July June March  May
##   April       522    398      186       91     191  681  386   221  138
##   August      212   1651      265      131     220 1101  540   250  192
##   December    175    397      658      105     206  642  381   202  162
##   February    128    329      193      372     215  558  345   189  116
##   January     162    399      225      120     683  621  423   193  138
##   July        194    694      253      133     251 2604  610   254  187
##   June        201    581      263      111     256 1005 1434   281  207
##   March       167    394      199      120     208  624  381   643  144
##   May         160    418      191       84     160  636  383   205  480
##   November    185    474      236      126     219  764  451   210  178
##   October     184    521      250      131     231  871  494   244  168
##   September   176    542      238      116     249  976  535   257  174
##           predicted
## actual      November October September
##   April          255     300       296
##   August         360     414       447
##   December       287     288       304
##   February       213     259       231
##   January        260     303       283
##   July           360     457       443
##   June           333     384       388
##   March          239     285       293
##   May            258     301       251
##   November       924     378       340
##   October        303    1093       387
##   September      335     381      1105
##
## Overall Statistics
##
##                Accuracy : 0.2263
##                  95% CI : (0.2228, 0.2299)
##     No Information Rate : 0.2061
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.1468
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: April Class: August Class: December Class: February
## Sensitivity              0.211679       0.24287         0.20843        0.226829
## Specificity              0.938734       0.91203         0.93778        0.946745
## Pos Pred Value           0.142428       0.28549         0.17284        0.118170
## Neg Pred Value           0.961199       0.89274         0.94998        0.974950
## Prevalence               0.045865       0.12643         0.05872        0.030502
## Detection Rate           0.009709       0.03071         0.01224        0.006919
## Detection Prevalence     0.068164       0.10756         0.07081        0.058549
## Balanced Accuracy        0.575206       0.57745         0.57310        0.586787
##                      Class: January Class: July Class: June Class: March
## Sensitivity                 0.22111     0.23495     0.22537      0.20419
## Specificity                 0.93830     0.91013     0.91541      0.93967
## Pos Pred Value              0.17927     0.40435     0.26341      0.17392
```

```
## Neg Pred Value              0.95184     0.82084     0.89800      0.94995
## Prevalence                  0.05745     0.20613     0.11834      0.05857
## Detection Rate              0.01270     0.04843     0.02667      0.01196
## Detection Prevalence        0.07086     0.11978     0.10125      0.06876
## Balanced Accuracy           0.57970     0.57254     0.57039      0.57193
##                       Class: May Class: November Class: October Class: September
## Sensitivity             0.210158         0.22389        0.22569           0.23175
## Specificity             0.940815         0.92826        0.92266           0.91879
## Pos Pred Value          0.136093         0.20602        0.22411           0.21735
## Neg Pred Value          0.964092         0.93501        0.92330           0.92476
## Prevalence              0.042480         0.07676        0.09007           0.08868
## Detection Rate          0.008927         0.01719        0.02033           0.02055
## Detection Prevalence    0.065598         0.08342        0.09071           0.09456
## Balanced Accuracy       0.575487         0.57608        0.57417           0.57527
```

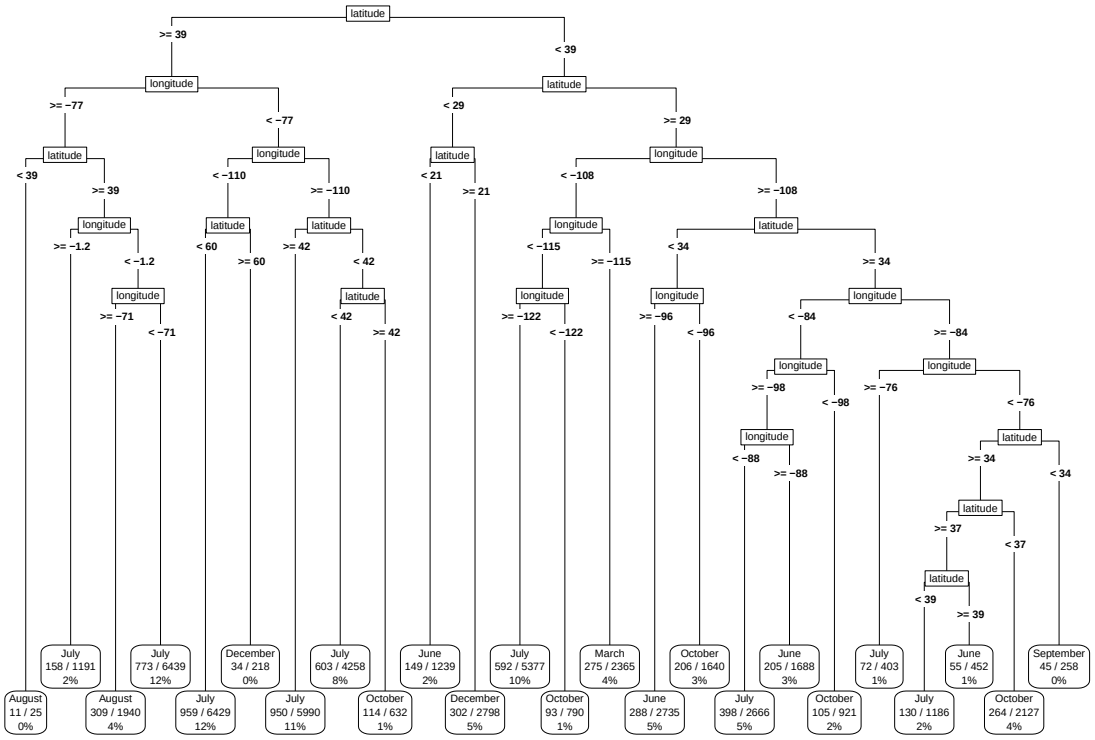
From the confusion matrix displayed above we see that the accuracy of this model is 22.5 %. While this is a relatively low accuracy it is still a 2.1% higher accuracy rate than if predictions were made based on no information. The matrix also shows that this difference is statisitcally signficant, with a p value of 2.2e-16. So while the model isn't super accurate its still better than if there was no information to base preditions on.

Classification Tree

The accuracy of our kNN classification was relatively low so we decided to make a classification tree to make a model of longitude, latitude, and the month of a sighting. We thought this may work better as it is a very large data set and classification trees are able to run quicker. Additonally it would provide a visualization and steps to take when determining the month of a sighting.

```
##  xerror cutoff cp prune value
##   0.9951362587   0.0003380734
```

```
## Warning: All boxes will be white (the box.palette argument will be ignored) because
## the number of classes in the response 12 is greater than length(box.palette) 6.
## To silence this warning use box.palette=0 or trace=-1.
```




```
##      month  Apr Aug Dec Feb Jan Jul Jun Mar May Nov Oct Sep
##     August [.06 .16 .07 .05 .07 .11 .09 .07 .06 .07 .08 .11] when latitude >=       39 & longitude is
##     August [.00 .44 .12 .00 .08 .08 .00 .08 .12 .08 .00 .00] when latitude is 39 to 39 & longitude >=
##   December [.08 .08 .11 .07 .10 .08 .09 .09 .06 .09 .09 .08] when latitude is 21 to 29
##   December [.06 .05 .16 .08 .08 .05 .04 .10 .03 .12 .11 .12] when latitude >=       60 & longitude <
##       July [.06 .09 .10 .06 .07 .11 .09 .08 .08 .09 .07 .08] when latitude is 37 to 39 & longitude is
##       July [.07 .10 .08 .06 .09 .11 .10 .07 .07 .09 .08 .09] when latitude is 29 to 39 & longitude is
##       July [.06 .12 .07 .06 .07 .12 .10 .07 .06 .09 .09 .10] when latitude >=       39 & longitude is
##       July [.07 .11 .06 .07 .09 .13 .11 .06 .08 .06 .07 .08] when latitude >=       39 & longitude >=
##       July [.07 .11 .06 .04 .05 .14 .10 .07 .07 .08 .10 .10] when latitude is 39 to 42 & longitude is
##       July [.07 .13 .06 .06 .06 .15 .10 .07 .06 .06 .08 .10] when latitude is 39 to 60 & longitude <
##       July [.06 .09 .08 .04 .07 .15 .10 .06 .08 .09 .10 .08] when latitude is 34 to 39 & longitude is
##       July [.06 .14 .05 .05 .05 .16 .11 .05 .06 .08 .08 .10] when latitude >=       42 & longitude is
##       July [.07 .12 .07 .04 .08 .18 .12 .05 .05 .08 .06 .08] when latitude is 34 to 39 & longitude >=
##       June [.08 .09 .09 .06 .09 .09 .11 .07 .06 .10 .09 .07] when latitude is 29 to 34 & longitude >=
##       June [.09 .07 .10 .08 .10 .07 .12 .10 .09 .08 .06 .06] when latitude <  21
##       June [.07 .09 .07 .06 .06 .11 .12 .06 .07 .11 .10 .09] when latitude is 34 to 39 & longitude is
##       June [.07 .07 .08 .08 .08 .07 .12 .06 .04 .09 .12 .12] when latitude is 39 to 39 & longitude is
##      March [.07 .08 .07 .08 .08 .08 .09 .12 .08 .07 .09 .09] when latitude is 29 to 39 & longitude is
##    October [.08 .09 .06 .05 .06 .08 .11 .08 .06 .09 .11 .11] when latitude is 34 to 39 & longitude is
##    October [.05 .07 .10 .05 .09 .07 .10 .08 .07 .09 .12 .10] when latitude is 29 to 39 & longitude <
##    October [.07 .07 .07 .06 .08 .10 .09 .08 .07 .09 .12 .10] when latitude is 34 to 37 & longitude is
##    October [.08 .09 .05 .06 .07 .10 .09 .06 .07 .10 .13 .10] when latitude is 29 to 34 & longitude is
##    October [.07 .14 .04 .03 .04 .11 .11 .06 .06 .06 .18 .10] when latitude is 42 to 42 & longitude is
##  September [.05 .10 .08 .03 .05 .16 .12 .03 .03 .06 .11 .17] when latitude is 34 to 34 & longitude is
```

```
##            Overall
## latitude  866.2078
## longitude 582.6532


## Call:
## rpart(formula = month ~ longitude + latitude, data = ufo, method = "class",
##     parms = list(split = "information"), minsplit = 2, minbucket = 1,
##     cp = -1)
##   n= 53767
##
##          CP nsplit rel error xerror    xstd
## 1  0.001134      0     1.000  1.000 0.00159
## 2  0.000993      3     0.997  0.998 0.00160
## 3  0.000740      6     0.994  0.998 0.00160
## 4  0.000634      7     0.993  0.997 0.00161
## 5  0.000507      8     0.992  0.996 0.00161
## 6  0.000430      9     0.992  0.996 0.00161
## 7  0.000380     14     0.990  0.996 0.00161
## 8  0.000373     16     0.989  0.996 0.00161
## 9  0.000359     19     0.988  0.995 0.00161
## 10 0.000338     23     0.986  0.995 0.00162
##
## Variable importance
##  latitude longitude
##        57        43
##
## Node number 1: 53767 observations,    complexity param=0.00113
##   predicted class=July       expected loss=0.88  P(node) =1
##     class counts:  3665  5783  3807  3148  3810  6440  5444  3697  3527  4485  4877  5084
##    probabilities: 0.068 0.108 0.071 0.059 0.071 0.120 0.101 0.069 0.066 0.083 0.091 0.095
##   left son=2 (27122 obs) right son=3 (26645 obs)
##   Primary splits:
##       latitude  < 39.3  to the right, improve=309.0, (0 missing)
##       longitude < -110  to the left,  improve= 39.1, (0 missing)
##   Surrogate splits:
##       longitude < -76.6 to the right, agree=0.633, adj=0.26, (0 split)
##
## Node number 2: 27122 observations,    complexity param=0.00113
##   predicted class=July       expected loss=0.862  P(node) =0.504
##     class counts:  1697  3444  1654  1497  1674  3736  2774  1707  1716  2079  2376  2768
##    probabilities: 0.063 0.127 0.061 0.055 0.062 0.138 0.102 0.063 0.063 0.077 0.088 0.102
##   left son=4 (9595 obs) right son=5 (17527 obs)
##   Primary splits:
##       longitude < -77.2 to the right, improve=52.5, (0 missing)
##       latitude  < 56    to the left,  improve=46.2, (0 missing)
##   Surrogate splits:
##       latitude < 50.5  to the right, agree=0.691, adj=0.126, (0 split)
##
## Node number 3: 26645 observations,    complexity param=0.00113
##   predicted class=July       expected loss=0.899  P(node) =0.496
##     class counts:  1968  2339  2153  1651  2136  2704  2670  1990  1811  2406  2501  2316
##    probabilities: 0.074 0.088 0.081 0.062 0.080 0.101 0.100 0.075 0.068 0.090 0.094 0.087
##   left son=6 (4037 obs) right son=7 (22608 obs)
##   Primary splits:
```

```
##       latitude  < 29.3  to the left,  improve=64.9, (0 missing)
##       longitude < -111  to the left,  improve=28.3, (0 missing)
##   Surrogate splits:
##       longitude < -74.5 to the right, agree=0.879, adj=0.199, (0 split)
##
## Node number 4: 9595 observations,    complexity param=0.000373
##   predicted class=August     expected loss=0.875  P(node) =0.178
##     class counts:   580  1198   634   571   717  1144   962   640   626   773   780   970
##    probabilities: 0.060 0.125 0.066 0.060 0.075 0.119 0.100 0.067 0.065 0.081 0.081 0.101
##   left son=8 (25 obs) right son=9 (9570 obs)
##   Primary splits:
##       latitude  < 39.4  to the left,  improve=17.2, (0 missing)
##       longitude < -1.23 to the right, improve=16.8, (0 missing)
##
## Node number 5: 17527 observations,    complexity param=0.000359
##   predicted class=July       expected loss=0.852  P(node) =0.326
##     class counts:  1117  2246  1020   926   957  2592  1812  1067  1090  1306  1596  1798
##    probabilities: 0.064 0.128 0.058 0.053 0.055 0.148 0.103 0.061 0.062 0.075 0.091 0.103
##   left son=10 (6647 obs) right son=11 (10880 obs)
##   Primary splits:
##       longitude < -110  to the left,  improve=51.1, (0 missing)
##       latitude  < 45.4  to the right, improve=47.6, (0 missing)
##   Surrogate splits:
##       latitude < 45.2  to the right, agree=0.838, adj=0.574, (0 split)
##
## Node number 6: 4037 observations,    complexity param=0.000507
##   predicted class=December   expected loss=0.894  P(node) =0.0751
##     class counts:   339   310   427   280   384   320   388   360   278   344   324   283
##    probabilities: 0.084 0.077 0.106 0.069 0.095 0.079 0.096 0.089 0.069 0.085 0.080 0.070
##   left son=12 (1239 obs) right son=13 (2798 obs)
##   Primary splits:
##       latitude  < 21.2  to the left,  improve=19.9, (0 missing)
##       longitude < -156  to the left,  improve=14.3, (0 missing)
##   Surrogate splits:
##       longitude < -80   to the right, agree=0.909, adj=0.703, (0 split)
##
## Node number 7: 22608 observations,    complexity param=0.000993
##   predicted class=July       expected loss=0.895  P(node) =0.42
##     class counts:  1629  2029  1726  1371  1752  2384  2282  1630  1533  2062  2177  2033
##    probabilities: 0.072 0.090 0.076 0.061 0.077 0.105 0.101 0.072 0.068 0.091 0.096 0.090
##   left son=14 (8532 obs) right son=15 (14076 obs)
##   Primary splits:
##       longitude < -108  to the left,  improve=42.7, (0 missing)
##       latitude  < 36.2  to the left,  improve=28.2, (0 missing)
##
## Node number 8: 25 observations
##   predicted class=August     expected loss=0.56  P(node) =0.000465
##     class counts:     0    11     3     0     2     2     0     2     3     2     0     0
##    probabilities: 0.000 0.440 0.120 0.000 0.080 0.080 0.000 0.080 0.120 0.080 0.000 0.000
##
## Node number 9: 9570 observations,    complexity param=0.000373
##   predicted class=August     expected loss=0.876  P(node) =0.178
##     class counts:   580  1187   631   571   715  1142   962   638   623   771   780   970
##    probabilities: 0.061 0.124 0.066 0.060 0.075 0.119 0.101 0.067 0.065 0.081 0.082 0.101
```

```
##   left son=18 (1191 obs) right son=19 (8379 obs)
##   Primary splits:
##       longitude < -1.23 to the right, improve=16.8, (0 missing)
##       latitude  < 45.4  to the right, improve=14.9, (0 missing)
##   Surrogate splits:
##       latitude < 57.7  to the right, agree=0.882, adj=0.054, (0 split)
##
## Node number 10: 6647 observations,    complexity param=0.000359
##   predicted class=July       expected loss=0.854  P(node) =0.124
##     class counts:   442   844   441   435   424   969   625   439   378   430   553   667
##    probabilities: 0.066 0.127 0.066 0.065 0.064 0.146 0.094 0.066 0.057 0.065 0.083 0.100
##   left son=20 (6429 obs) right son=21 (218 obs)
##   Primary splits:
##       latitude  < 59.8  to the left,  improve=41.0, (0 missing)
##       longitude < -134  to the right, improve=38.3, (0 missing)
##   Surrogate splits:
##       longitude < -135  to the right, agree=0.995, adj=0.862, (0 split)
##
## Node number 11: 10880 observations,    complexity param=0.000359
##   predicted class=July       expected loss=0.851  P(node) =0.202
##     class counts:   675  1402   579   491   533  1623  1187   628   712   876  1043  1131
##    probabilities: 0.062 0.129 0.053 0.045 0.049 0.149 0.109 0.058 0.065 0.081 0.096 0.104
##   left son=22 (5990 obs) right son=23 (4890 obs)
##   Primary splits:
##       latitude  < 41.7  to the right, improve=34.6, (0 missing)
##       longitude < -87.7 to the right, improve=21.6, (0 missing)
##   Surrogate splits:
##       longitude < -105  to the right, agree=0.596, adj=0.101, (0 split)
##
## Node number 12: 1239 observations
##   predicted class=June       expected loss=0.88  P(node) =0.023
##     class counts:   112    82   125    93   118    87   149   122   106    98    74    73
##    probabilities: 0.090 0.066 0.101 0.075 0.095 0.070 0.120 0.098 0.086 0.079 0.060 0.059
##
## Node number 13: 2798 observations
##   predicted class=December   expected loss=0.892  P(node) =0.052
##     class counts:   227   228   302   187   266   233   239   238   172   246   250   210
##    probabilities: 0.081 0.081 0.108 0.067 0.095 0.083 0.085 0.085 0.061 0.088 0.089 0.075
##
## Node number 14: 8532 observations,    complexity param=0.000993
##   predicted class=July       expected loss=0.901  P(node) =0.159
##     class counts:   619   815   642   565   720   847   821   715   595   733   708   752
##    probabilities: 0.073 0.096 0.075 0.066 0.084 0.099 0.096 0.084 0.070 0.086 0.083 0.088
##   left son=28 (6167 obs) right son=29 (2365 obs)
##   Primary splits:
##       longitude < -115  to the left,  improve=36.5, (0 missing)
##       latitude  < 37.4  to the left,  improve=21.0, (0 missing)
##   Surrogate splits:
##       latitude < 33.5  to the right, agree=0.791, adj=0.245, (0 split)
##
## Node number 15: 14076 observations,    complexity param=0.000993
##   predicted class=July       expected loss=0.891  P(node) =0.262
##     class counts:  1010  1214  1084   806  1032  1537  1461   915   938  1329  1469  1281
##    probabilities: 0.072 0.086 0.077 0.057 0.073 0.109 0.104 0.065 0.067 0.094 0.104 0.091
```

```
##   left son=30 (4375 obs) right son=31 (9701 obs)
##   Primary splits:
##       latitude  < 33.6  to the left,  improve=26.7, (0 missing)
##       longitude < -76   to the right, improve=23.4, (0 missing)
##   Surrogate splits:
##       longitude < -94.9 to the left,  agree=0.731, adj=0.133, (0 split)
##
## Node number 18: 1191 observations
##   predicted class=July       expected loss=0.867  P(node) =0.0222
##     class counts:    83   136    77    78   107   158   132    72    98    73    80    97
##    probabilities: 0.070 0.114 0.065 0.065 0.090 0.133 0.111 0.060 0.082 0.061 0.067 0.081
##
## Node number 19: 8379 observations,    complexity param=0.000373
##   predicted class=August     expected loss=0.875  P(node) =0.156
##     class counts:   497  1051   554   493   608   984   830   566   525   698   700   873
##    probabilities: 0.059 0.125 0.066 0.059 0.073 0.117 0.099 0.068 0.063 0.083 0.084 0.104
##   left son=38 (1940 obs) right son=39 (6439 obs)
##   Primary splits:
##       longitude < -70.9 to the right, improve=17.5, (0 missing)
##       latitude  < 43    to the right, improve=15.7, (0 missing)
##   Surrogate splits:
##       latitude < 45.7  to the right, agree=0.913, adj=0.624, (0 split)
##
## Node number 20: 6429 observations
##   predicted class=July       expected loss=0.851  P(node) =0.12
##     class counts:   429   833   407   417   406   959   616   418   371   404   528   641
##    probabilities: 0.067 0.130 0.063 0.065 0.063 0.149 0.096 0.065 0.058 0.063 0.082 0.100
##
## Node number 21: 218 observations
##   predicted class=December   expected loss=0.844  P(node) =0.00405
##     class counts:    13    11    34    18    18    10     9    21     7    26    25    26
##    probabilities: 0.060 0.050 0.156 0.083 0.083 0.046 0.041 0.096 0.032 0.119 0.115 0.119
##
## Node number 22: 5990 observations
##   predicted class=July       expected loss=0.841  P(node) =0.111
##     class counts:   330   850   302   286   294   950   672   315   387   484   499   621
##    probabilities: 0.055 0.142 0.050 0.048 0.049 0.159 0.112 0.053 0.065 0.081 0.083 0.104
##
## Node number 23: 4890 observations,    complexity param=0.000359
##   predicted class=July       expected loss=0.862  P(node) =0.0909
##     class counts:   345   552   277   205   239   673   515   313   325   392   544   510
##    probabilities: 0.071 0.113 0.057 0.042 0.049 0.138 0.105 0.064 0.066 0.080 0.111 0.104
##   left son=46 (4258 obs) right son=47 (632 obs)
##   Primary splits:
##       latitude  < 41.5  to the left,  improve=26.6, (0 missing)
##       longitude < -87.7 to the right, improve=23.0, (0 missing)
##
## Node number 28: 6167 observations,    complexity param=0.00074
##   predicted class=July       expected loss=0.895  P(node) =0.115
##     class counts:   442   616   485   387   536   650   603   440   412   561   497   538
##    probabilities: 0.072 0.100 0.079 0.063 0.087 0.105 0.098 0.071 0.067 0.091 0.081 0.087
##   left son=56 (5377 obs) right son=57 (790 obs)
##   Primary splits:
##       longitude < -122  to the right, improve=23.2, (0 missing)
```

```
##       latitude  < 37.4  to the left,  improve=19.9, (0 missing)
##
## Node number 29: 2365 observations
##   predicted class=March      expected loss=0.884  P(node) =0.044
##     class counts:   177   199   157   178   184   197   218   275   183   172   211   214
##    probabilities: 0.075 0.084 0.066 0.075 0.078 0.083 0.092 0.116 0.077 0.073 0.089 0.090
##
## Node number 30: 4375 observations,    complexity param=0.000634
##   predicted class=October    expected loss=0.894  P(node) =0.0814
##     class counts:   359   377   328   279   371   401   442   272   281   440   464   361
##    probabilities: 0.082 0.086 0.075 0.064 0.085 0.092 0.101 0.062 0.064 0.101 0.106 0.083
##   left son=60 (2735 obs) right son=61 (1640 obs)
##   Primary splits:
##       longitude < -96.1 to the right, improve=22.5, (0 missing)
##       latitude  < 31.1  to the left,  improve=12.9, (0 missing)
##   Surrogate splits:
##       latitude < 29.4  to the right, agree=0.64, adj=0.04, (0 split)
##
## Node number 31: 9701 observations,    complexity param=0.00043
##   predicted class=July       expected loss=0.883  P(node) =0.18
##     class counts:   651   837   756   527   661  1136  1019   643   657   889  1005   920
##    probabilities: 0.067 0.086 0.078 0.054 0.068 0.117 0.105 0.066 0.068 0.092 0.104 0.095
##   left son=62 (5275 obs) right son=63 (4426 obs)
##   Primary splits:
##       longitude < -84.2 to the left,  improve=18.5, (0 missing)
##       latitude  < 36.6  to the right, improve=14.5, (0 missing)
##   Surrogate splits:
##       latitude < 34.2  to the right, agree=0.563, adj=0.042, (0 split)
##
## Node number 38: 1940 observations
##   predicted class=August     expected loss=0.841  P(node) =0.0361
##     class counts:   114   309   130   101   145   211   178   131   112   143   146   220
##    probabilities: 0.059 0.159 0.067 0.052 0.075 0.109 0.092 0.068 0.058 0.074 0.075 0.113
##
## Node number 39: 6439 observations
##   predicted class=July       expected loss=0.88  P(node) =0.12
##     class counts:   383   742   424   392   463   773   652   435   413   555   554   653
##    probabilities: 0.059 0.115 0.066 0.061 0.072 0.120 0.101 0.068 0.064 0.086 0.086 0.101
##
## Node number 46: 4258 observations
##   predicted class=July       expected loss=0.858  P(node) =0.0792
##     class counts:   298   465   254   189   214   603   444   277   285   355   430   444
##    probabilities: 0.070 0.109 0.060 0.044 0.050 0.142 0.104 0.065 0.067 0.083 0.101 0.104
##
## Node number 47: 632 observations
##   predicted class=October    expected loss=0.82  P(node) =0.0118
##     class counts:    47    87    23    16    25    70    71    36    40    37   114    66
##    probabilities: 0.074 0.138 0.036 0.025 0.040 0.111 0.112 0.057 0.063 0.059 0.180 0.104
##
## Node number 56: 5377 observations
##   predicted class=July       expected loss=0.89  P(node) =0.1
##     class counts:   402   560   409   345   466   592   522   374   355   487   404   461
##    probabilities: 0.075 0.104 0.076 0.064 0.087 0.110 0.097 0.070 0.066 0.091 0.075 0.086
##
```

```
## Node number 57: 790 observations
##   predicted class=October    expected loss=0.882  P(node) =0.0147
##     class counts:    40    56    76    42    70    58    81    66    57    74    93    77
##    probabilities: 0.051 0.071 0.096 0.053 0.089 0.073 0.103 0.084 0.072 0.094 0.118 0.097
##
## Node number 60: 2735 observations
##   predicted class=June       expected loss=0.895  P(node) =0.0509
##     class counts:   229   237   240   174   253   241   288   178   166   273   258   198
##    probabilities: 0.084 0.087 0.088 0.064 0.093 0.088 0.105 0.065 0.061 0.100 0.094 0.072
##
## Node number 61: 1640 observations
##   predicted class=October    expected loss=0.874  P(node) =0.0305
##     class counts:   130   140    88   105   118   160   154    94   115   167   206   163
##    probabilities: 0.079 0.085 0.054 0.064 0.072 0.098 0.094 0.057 0.070 0.102 0.126 0.099
##
## Node number 62: 5275 observations,    complexity param=0.00043
##   predicted class=July       expected loss=0.875  P(node) =0.0981
##     class counts:   361   477   389   259   329   660   588   320   370   498   545   479
##    probabilities: 0.068 0.090 0.074 0.049 0.062 0.125 0.111 0.061 0.070 0.094 0.103 0.091
##   left son=124 (4354 obs) right son=125 (921 obs)
##   Primary splits:
##       longitude < -97.6 to the right, improve=18.6, (0 missing)
##       latitude  < 36.4  to the right, improve=13.7, (0 missing)
##   Surrogate splits:
##       latitude < 33.6  to the right, agree=0.828, adj=0.015, (0 split)
##
## Node number 63: 4426 observations,    complexity param=0.00043
##   predicted class=July       expected loss=0.892  P(node) =0.0823
##     class counts:   290   360   367   268   332   476   431   323   287   391   460   441
##    probabilities: 0.066 0.081 0.083 0.061 0.075 0.108 0.097 0.073 0.065 0.088 0.104 0.100
##   left son=126 (403 obs) right son=127 (4023 obs)
##   Primary splits:
##       longitude < -76   to the right, improve=24.4, (0 missing)
##       latitude  < 33.8  to the right, improve=21.7, (0 missing)
##
## Node number 124: 4354 observations,    complexity param=0.00038
##   predicted class=July       expected loss=0.866  P(node) =0.081
##     class counts:   287   392   333   209   271   585   484   250   315   413   440   375
##    probabilities: 0.066 0.090 0.076 0.048 0.062 0.134 0.111 0.057 0.072 0.095 0.101 0.086
##   left son=248 (2666 obs) right son=249 (1688 obs)
##   Primary splits:
##       longitude < -87.8 to the left,  improve=18.8, (0 missing)
##       latitude  < 36.3  to the right, improve=15.1, (0 missing)
##   Surrogate splits:
##       latitude < 34.7  to the right, agree=0.668, adj=0.145, (0 split)
##
## Node number 125: 921 observations
##   predicted class=October    expected loss=0.886  P(node) =0.0171
##     class counts:    74    85    56    50    58    75   104    70    55    85   105   104
##    probabilities: 0.080 0.092 0.061 0.054 0.063 0.081 0.113 0.076 0.060 0.092 0.114 0.113
##
## Node number 126: 403 observations
##   predicted class=July       expected loss=0.821  P(node) =0.0075
##     class counts:    28    49    27    16    34    72    47    20    19    32    25    34
```

```
##    probabilities: 0.069 0.122 0.067 0.040 0.084 0.179 0.117 0.050 0.047 0.079 0.062 0.084
##
## Node number 127: 4023 observations,    complexity param=0.00043
##   predicted class=October    expected loss=0.892  P(node) =0.0748
##     class counts:   262   311   340   252   298   404   384   303   268   359   435   407
##    probabilities: 0.065 0.077 0.085 0.063 0.074 0.100 0.095 0.075 0.067 0.089 0.108 0.101
##   left son=254 (3765 obs) right son=255 (258 obs)
##   Primary splits:
##       latitude  < 33.8  to the right, improve=25.5, (0 missing)
##       longitude < -83.3 to the right, improve=12.2, (0 missing)
##
## Node number 248: 2666 observations
##   predicted class=July       expected loss=0.851  P(node) =0.0496
##     class counts:   173   236   220   106   178   398   279   152   202   228   271   223
##    probabilities: 0.065 0.089 0.083 0.040 0.067 0.149 0.105 0.057 0.076 0.086 0.102 0.084
##
## Node number 249: 1688 observations
##   predicted class=June       expected loss=0.879  P(node) =0.0314
##     class counts:   114   156   113   103    93   187   205    98   113   185   169   152
##    probabilities: 0.068 0.092 0.067 0.061 0.055 0.111 0.121 0.058 0.067 0.110 0.100 0.090
##
## Node number 254: 3765 observations,    complexity param=0.00043
##   predicted class=October    expected loss=0.892  P(node) =0.07
##     class counts:   248   286   319   244   286   363   352   294   261   343   407   362
##    probabilities: 0.066 0.076 0.085 0.065 0.076 0.096 0.093 0.078 0.069 0.091 0.108 0.096
##   left son=508 (1638 obs) right son=509 (2127 obs)
##   Primary splits:
##       latitude  < 36.9  to the right, improve=12.5, (0 missing)
##       longitude < -83.3 to the right, improve=11.6, (0 missing)
##   Surrogate splits:
##       longitude < -77.7 to the right, agree=0.744, adj=0.411, (0 split)
##
## Node number 255: 258 observations
##   predicted class=September  expected loss=0.826  P(node) =0.0048
##     class counts:    14    25    21     8    12    41    32     9     7    16    28    45
##    probabilities: 0.054 0.097 0.081 0.031 0.047 0.159 0.124 0.035 0.027 0.062 0.109 0.174
##
## Node number 508: 1638 observations,    complexity param=0.00038
##   predicted class=June       expected loss=0.898  P(node) =0.0305
##     class counts:   101   137   161   111   121   160   167   126   109   152   143   150
##    probabilities: 0.062 0.084 0.098 0.068 0.074 0.098 0.102 0.077 0.067 0.093 0.087 0.092
##   left son=1016 (1186 obs) right son=1017 (452 obs)
##   Primary splits:
##       latitude  < 39    to the left,  improve=17.3, (0 missing)
##       longitude < -81.2 to the right, improve=10.9, (0 missing)
##   Surrogate splits:
##       longitude < -76.9 to the left,  agree=0.759, adj=0.128, (0 split)
##
## Node number 509: 2127 observations
##   predicted class=October    expected loss=0.876  P(node) =0.0396
##     class counts:   147   149   158   133   165   203   185   168   152   191   264   212
##    probabilities: 0.069 0.070 0.074 0.063 0.078 0.095 0.087 0.079 0.071 0.090 0.124 0.100
##
## Node number 1016: 1186 observations
```

```
##   predicted class=July       expected loss=0.89  P(node) =0.0221
##     class counts:    69   105   123    76    85   130   112    99    90   112    88    97
##    probabilities: 0.058 0.089 0.104 0.064 0.072 0.110 0.094 0.083 0.076 0.094 0.074 0.082
##
## Node number 1017: 452 observations
##   predicted class=June       expected loss=0.878  P(node) =0.00841
##     class counts:    32    32    38    35    36    30    55    27    19    40    55    53
##    probabilities: 0.071 0.071 0.084 0.077 0.080 0.066 0.122 0.060 0.042 0.088 0.122 0.117
```