

# 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 but only contains data from 1906 till 2014. 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.

This data is of interest because UFOs have been a topic of public interest 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. Additionally we had to remove some columns with improper formatting and we dropped any columns with NA's. The variables we used and code for this data cleaning follows:

**datetime**: the date and time of sighting in year, month, day, hours, minutes, seconds format.

**city**: city where sighting occurred

**state**: state where sighting occurred

**country**: country where sighting occurred

**seconds**: duration of sightings in seconds

**latitude**: latitude of sighting

**longitude**: longitude of sighting

**year**: year when sighting occurred

**month**: month when sighting occurred

```
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(grep1("^-?[0-9.]+$", latitude),
      grep1("^-?[0-9.]+$", seconds)) -> ufo

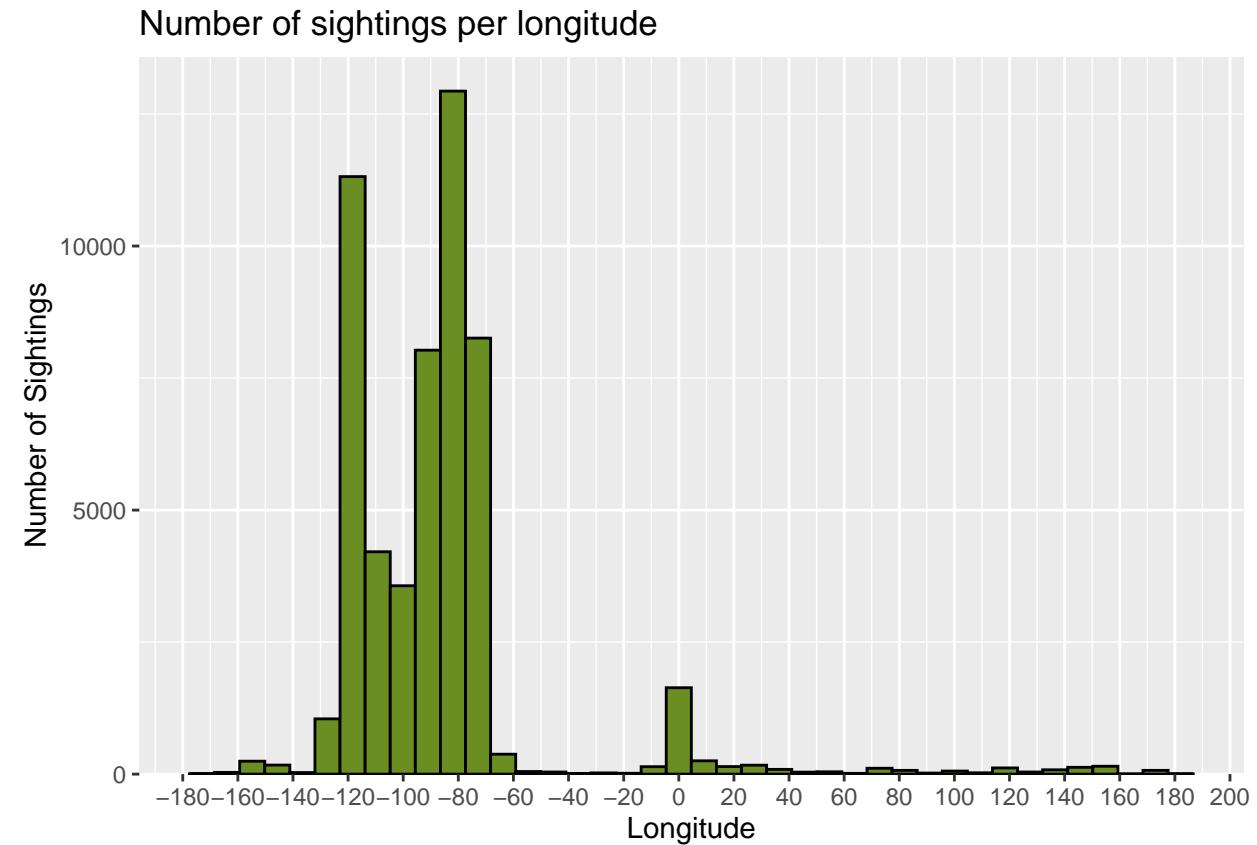
ufo$latitude <- as.numeric(ufo$latitude) #changing lat to numeric
ufo$seconds <- as.numeric(ufo$seconds) #changing seconds to numeric

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

```

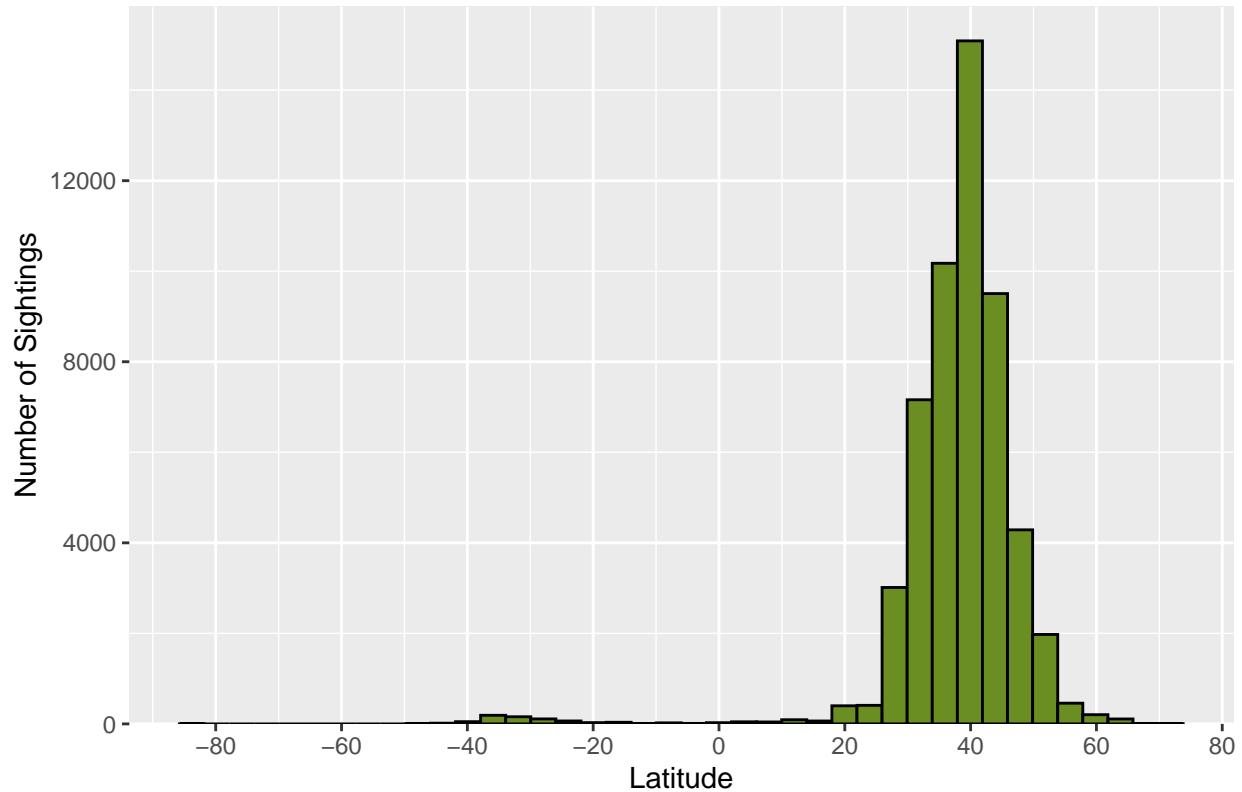
## Data Analysis

The first question we had about UFO sightings is how the location around the world influenced the number 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.

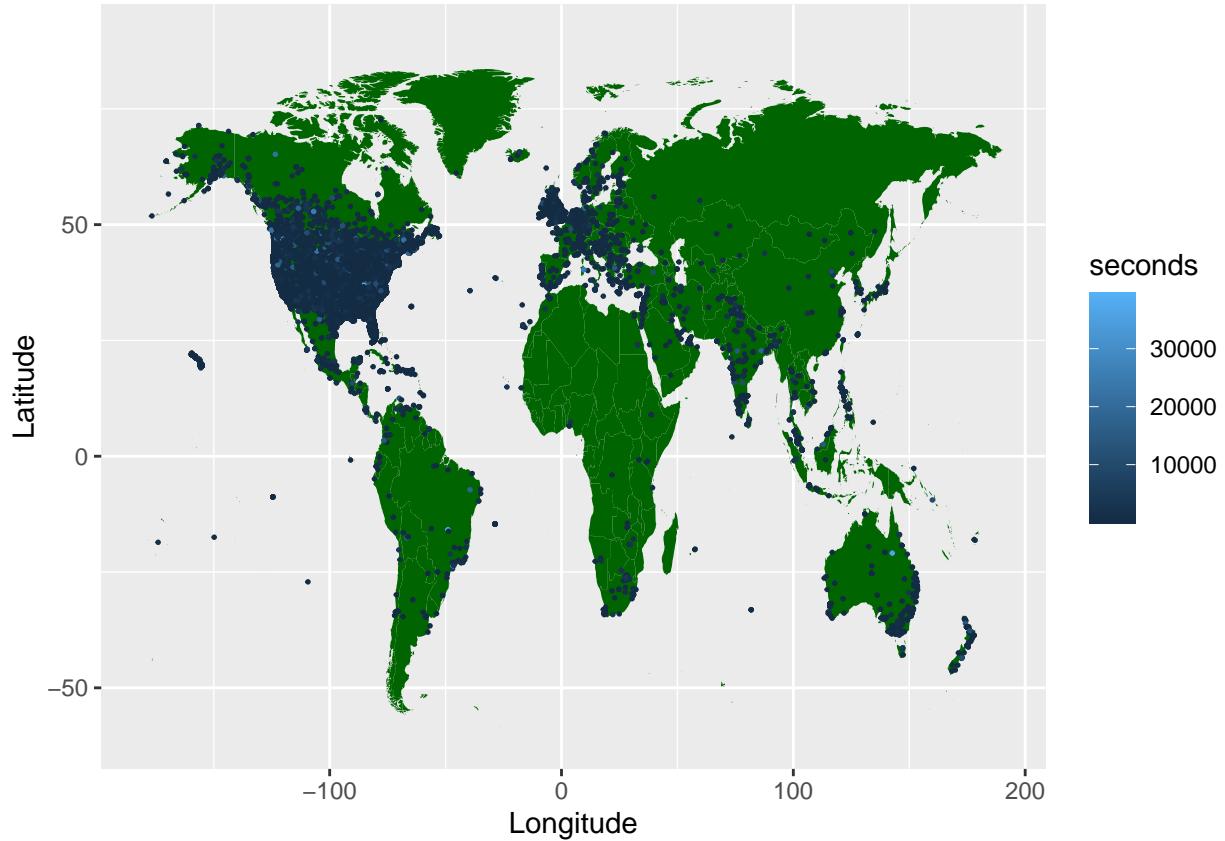


When looking at the histogram of longitude, we notice there are two main peaks within the spread. One peak is around -120 and the second is around -80. This makes sense because these are the longitudes that correspond with highly populated areas of the US. The relative height of these peaks implies that 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.

## Number of sightings per latitude

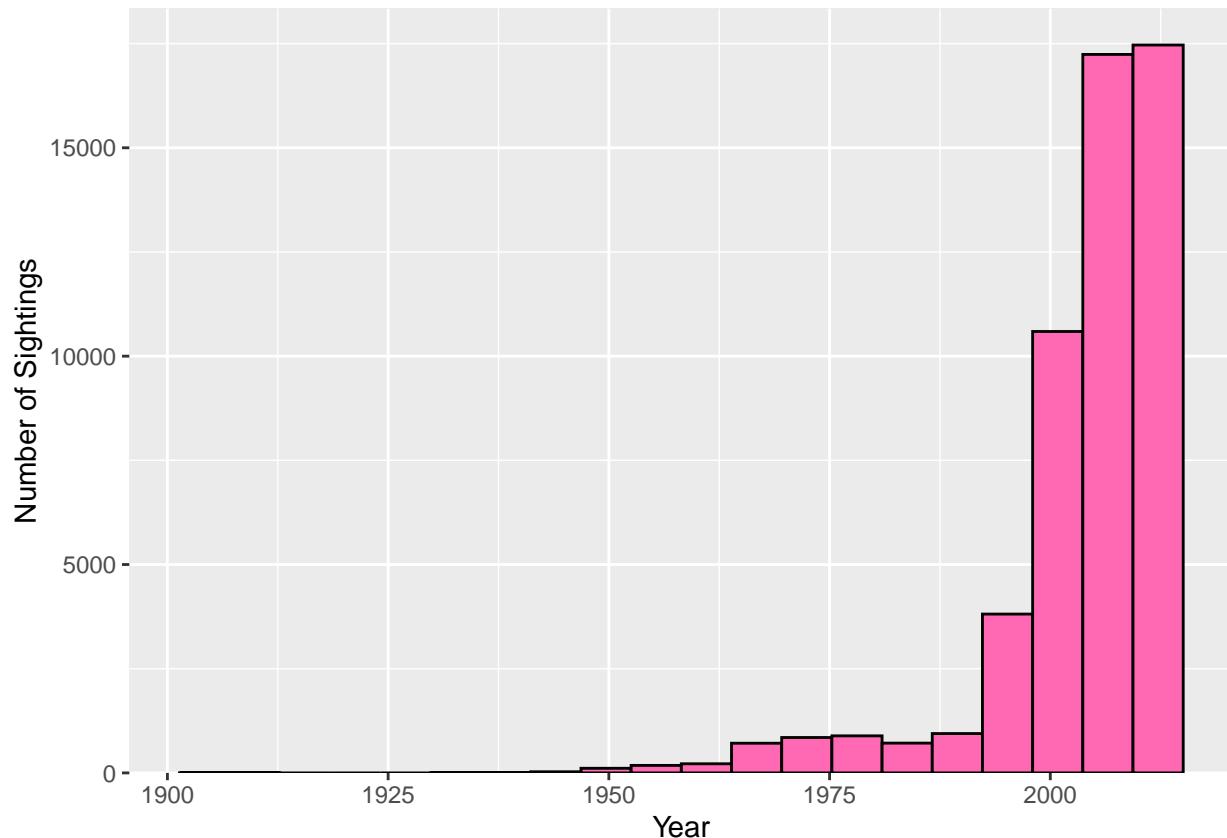


In the latitude histogram, we see than sightings are most prevalent around 40. Again, this corresponds to the coordinates of the US and also Europe, which is the second most frequently reported area of UFO sightings. This data supports the same conclusions we see from the longitude histogram.

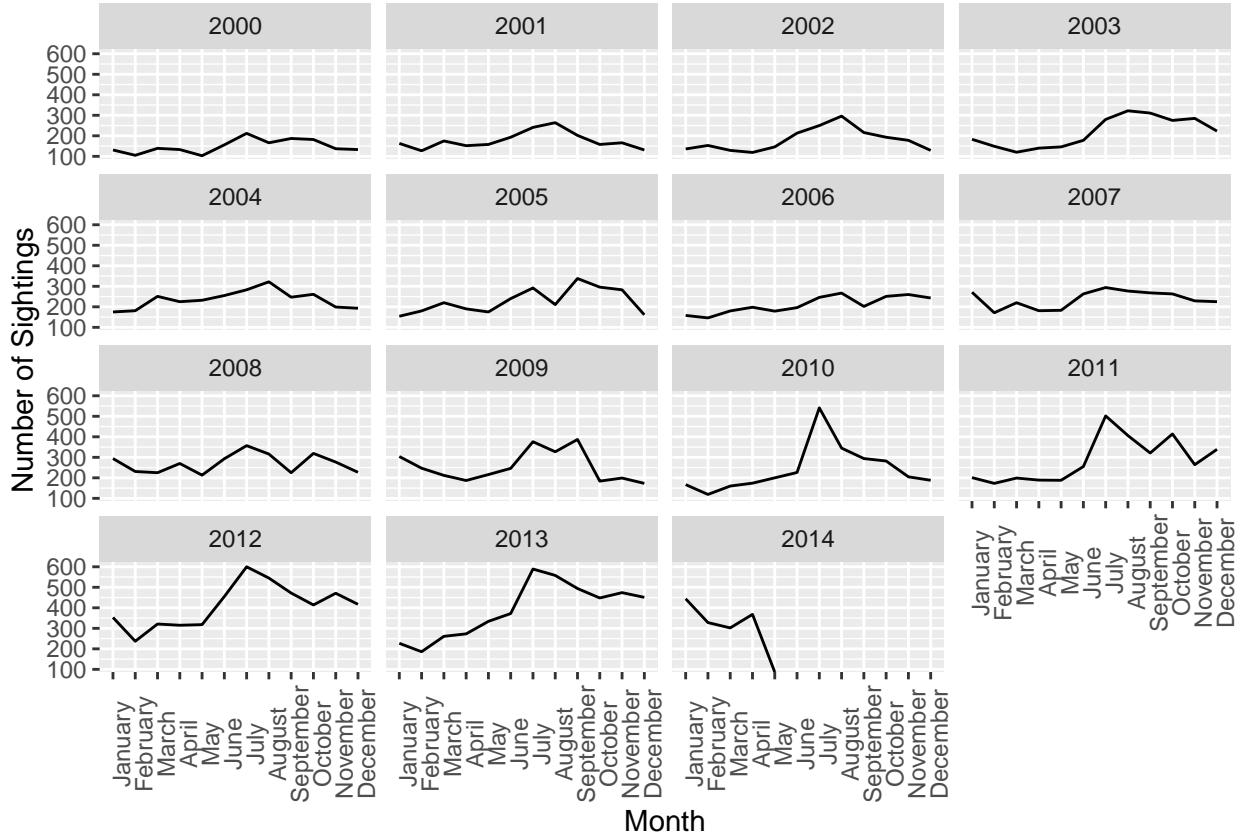


Here, we can see in greater detail the geographical dispersion of the sightings. The graph supports the same claims as the previous two graphs that there is a high volume of sightings in the United States. This graph also indicates other potential hot spots of sightings, specifically Europe and the eastern coast of Australia. The color coding makes it evident that most of the sightings are relatively short.

The next question we had was how the number of sightings had increased 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 existence of extraterrestrials.



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 June and July. This trend seems to be reflected in the following years since then. It's possible that this is simply because most sightings occur in the northern hemisphere and during these months it is summer. This means 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

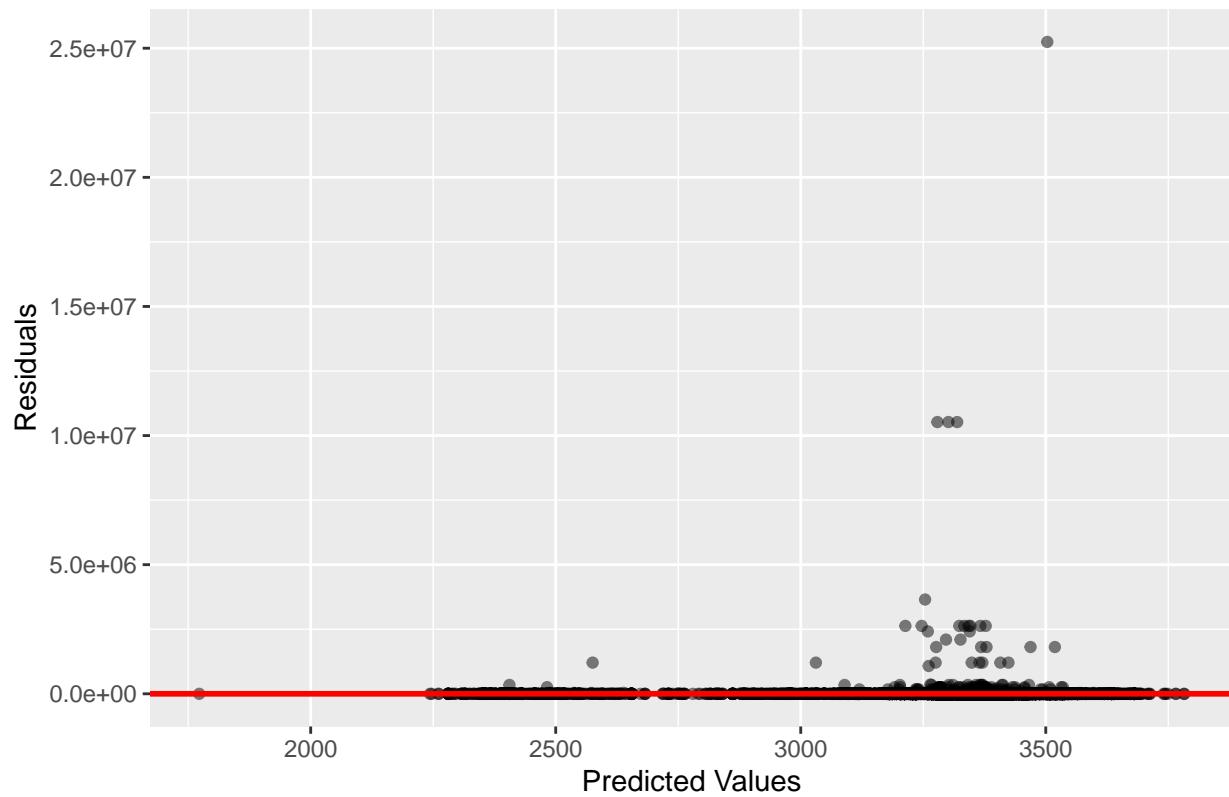
### Linear Regression

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.

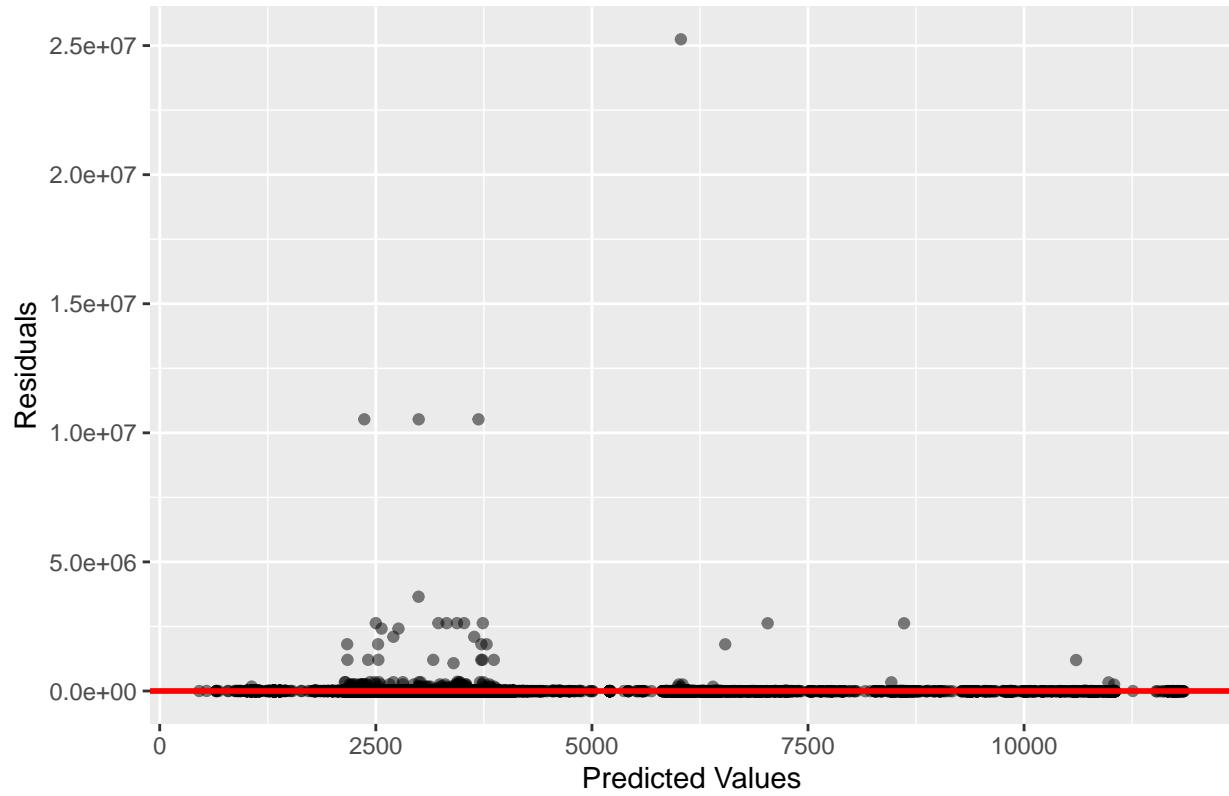
```
## # A tibble: 3 x 4
##   explanatories      r.squared adj.r.squared    sigma
##   <chr>                <dbl>            <dbl>    <dbl>
## 1 latitude        0.000000881 -0.0000177 142621.
## 2 longitude       0.0000784   0.0000598 142615.
## 3 latitude + longitude 0.0000999  0.0000627 142615.
```

## Residual Plots for Duration of Sightings vs Longitude and Latitude

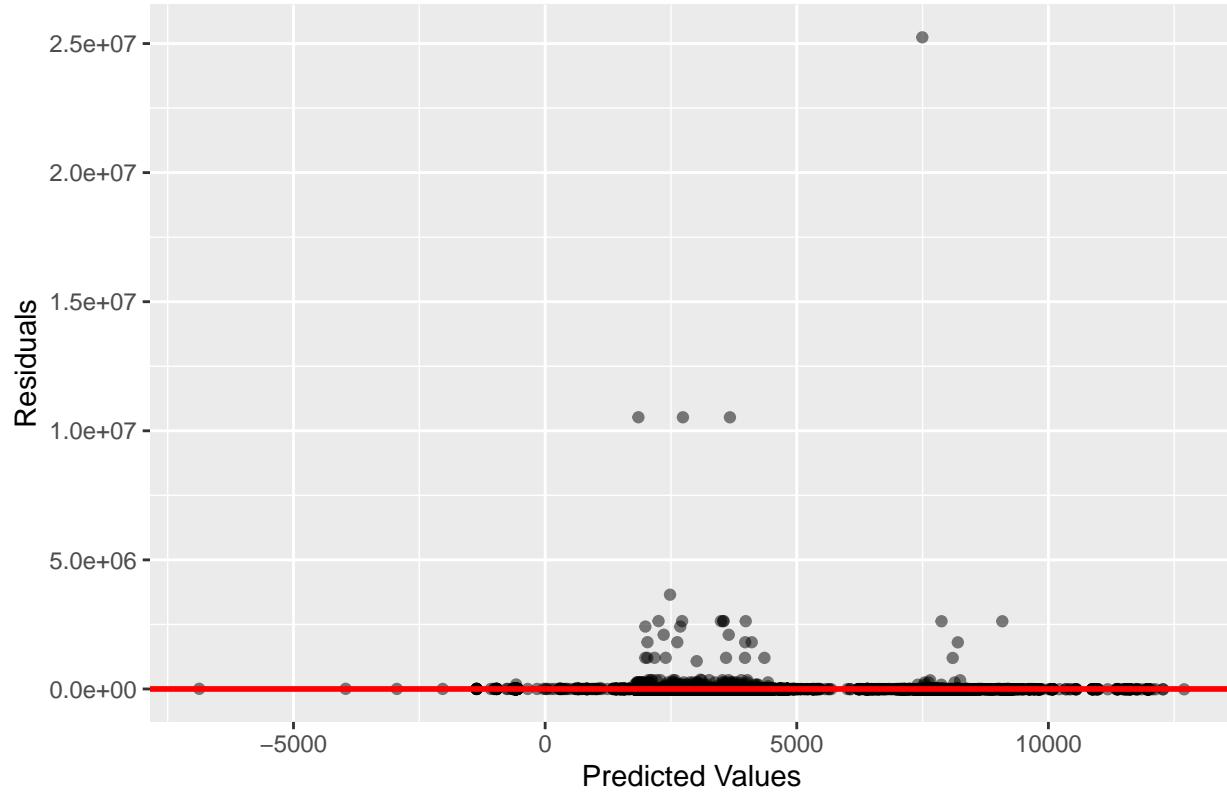
Residual Plot for Latitude Model



Residual Plot for Longitude Model



## Residual Plot for Longitude and Latitude Model



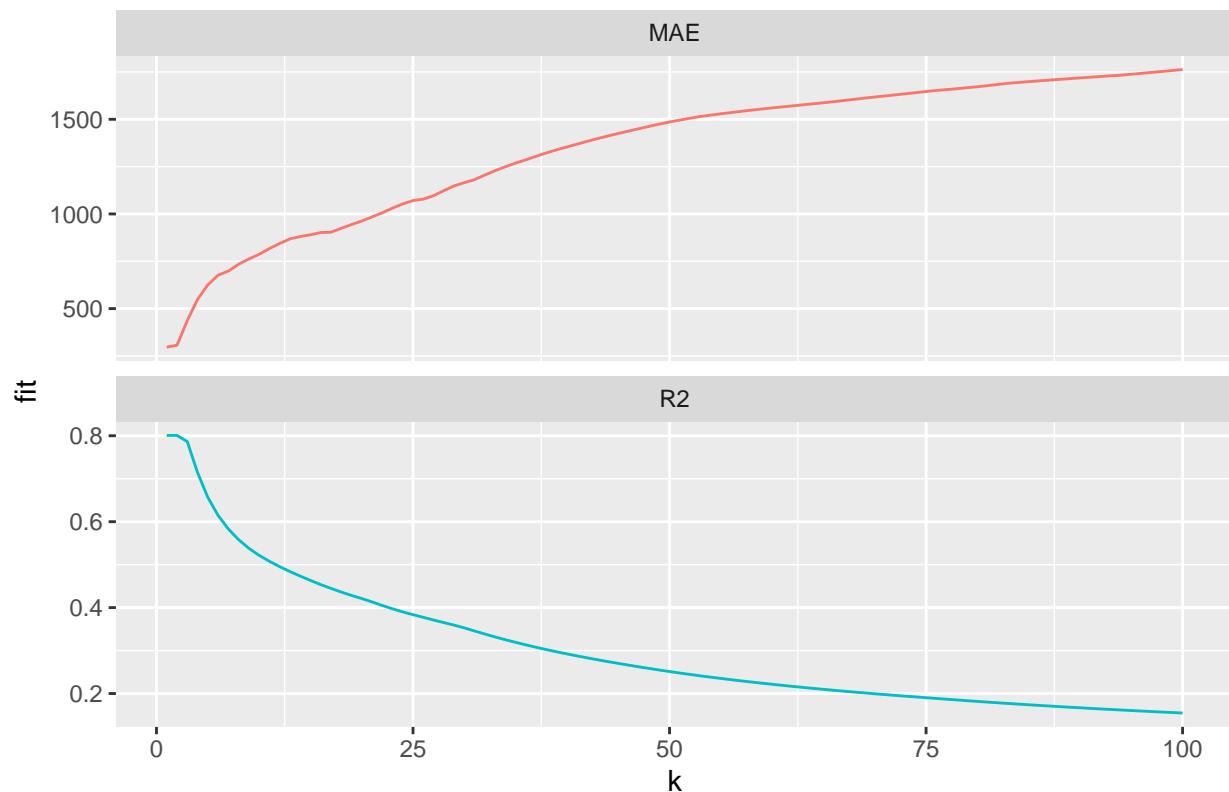
We created three different linear regression models, one with longitude as the explanatory variable, one with latitude as the explanatory variable, and the third with both. We wanted to compare which model would be the most accurate. From the table above we can see that all three models had R-squared 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 so the residuals dont have an even spread.

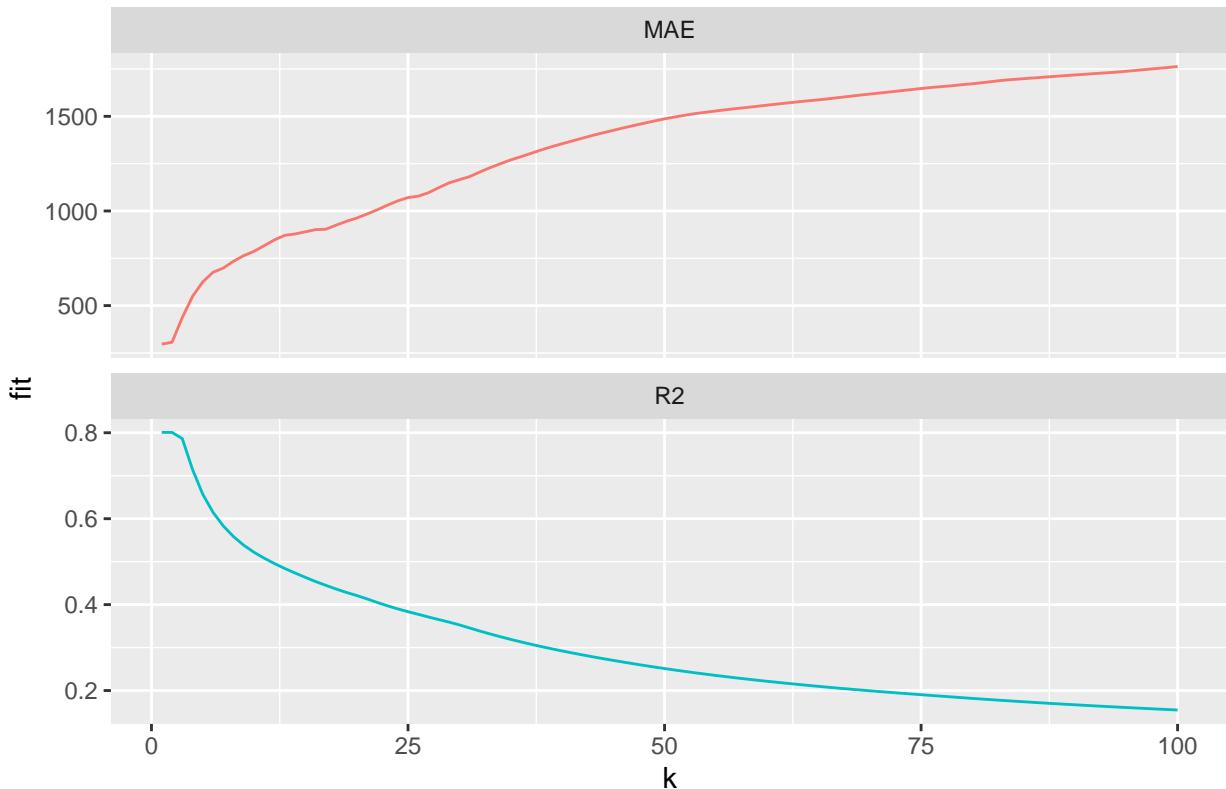
### kNN regression

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



## Fit Statistics for Standardized data



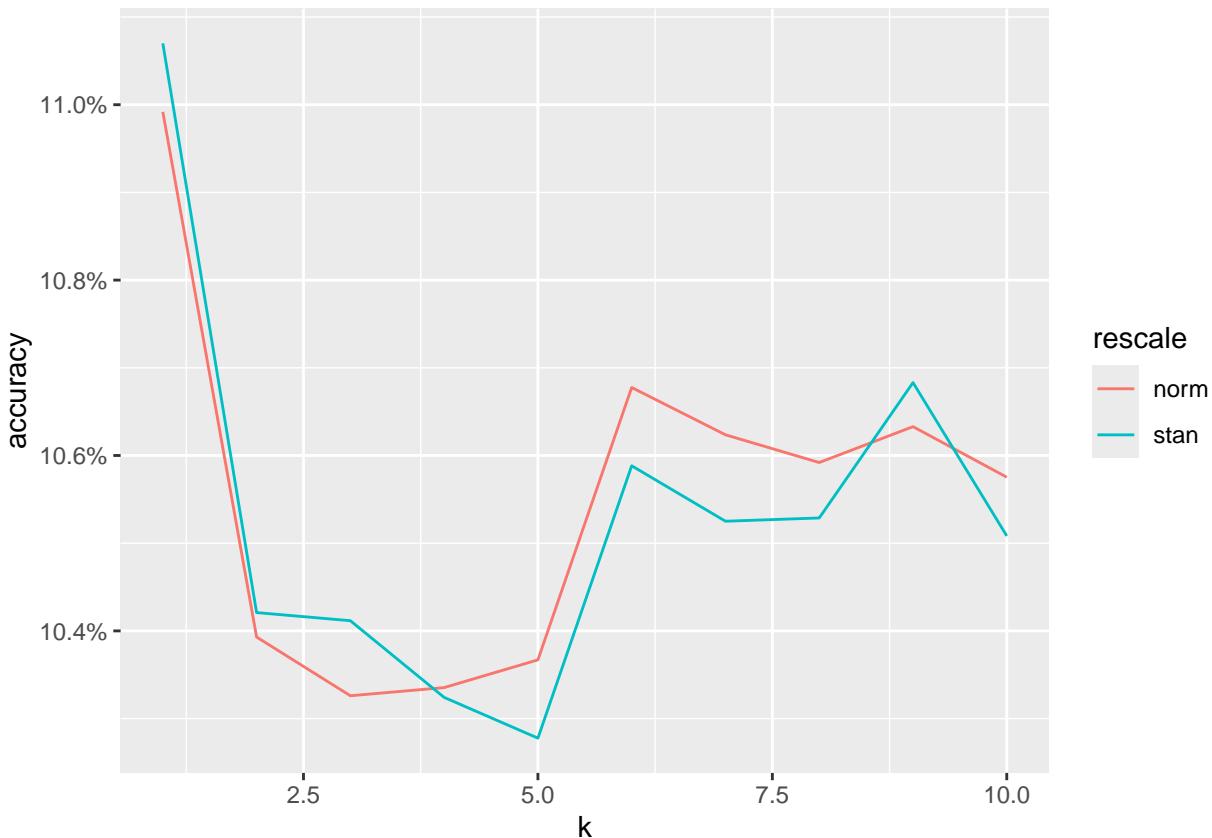
```
##   rescale k      R2      MAE
## 1    norm 1 0.800855 296.486
```

When looking at the fit statistic plots for both the normalized and standardized data 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, 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.



```

## # A tibble: 1 x 3
##       k rescale_method accuracy
##   <int> <chr>          <dbl>
## 1     1 stan_acc      0.111

## Confusion Matrix and Statistics
##
##           predicted
## actual      April August December February January July June March May
##   April      221    446     213    161    232   581   408   209   189
##   August     291    794     277    230    294  1060   636   287   276
##   December    193    435     290    151    248   578   437   240   194
##   February    185    366     224    115    202   505   349   187   138
##   January     223    461     263    171    276   581   399   242   208
##   July        326    885     359    227    367  1290   667   328   318
##   June        291    712     302    211    294   902   707   323   271
##   March       204    442     250    168    245   545   424   241   175
##   May         214    462     219    104    227   627   360   193   193
##   November    246    520     289    162    286   673   477   291   212
##   October     260    587     286    189    268   800   524   253   259
##   September   281    687     282    189    327   848   577   269   273
##
##           predicted
## actual      November October September
##   April        311     340     354
##   August       460     570     608

```

```

## December      343      345      353
## February     284      277      316
## January      303      336      347
## July         469      550      654
## June          422      477      532
## March         343      320      340
## May           275      323      330
## November     421      493      415
## October       386      589      476
## September    375      463      513
##
## Overall Statistics
##
##               Accuracy : 0.1051
##             95% CI : (0.1025, 0.1077)
##   No Information Rate : 0.1672
## P-Value [Acc > NIR] : 1
##
##               Kappa : 0.0152
##
## McNemar's Test P-Value : <2e-16
##
## Statistics by Class:
##
##               Class: April Class: August Class: December Class: February
## Sensitivity      0.07530     0.11682     0.089121    0.055342
## Specificity      0.93225     0.89378     0.930374    0.941322
## Pos Pred Value   0.06030     0.13730     0.076175    0.036531
## Neg Pred Value   0.94583     0.87490     0.940673    0.961220
## Prevalence        0.05459     0.12642     0.060520    0.038648
## Detection Rate   0.00411     0.01477     0.005394    0.002139
## Detection Prevalence 0.06816     0.10756     0.070806    0.058549
## Balanced Accuracy 0.50377     0.50530     0.509748    0.498332
##
##               Class: January Class: July Class: June Class: March
## Sensitivity      0.084507    0.14349     0.11852     0.078681
## Specificity      0.930021    0.88499     0.90090     0.931840
## Pos Pred Value   0.072441    0.20031     0.12987     0.065188
## Neg Pred Value   0.940149    0.83730     0.89119     0.943639
## Prevalence        0.060744    0.16720     0.11094     0.056968
## Detection Rate   0.005133    0.02399     0.01315     0.004482
## Detection Prevalence 0.070861    0.11978     0.10125     0.068760
## Balanced Accuracy 0.507264    0.51424     0.50971     0.505260
##
##               Class: May Class: November Class: October Class: September
## Sensitivity      0.07132     0.09586     0.11588     0.097938
## Specificity      0.93471     0.91769     0.91192     0.905809
## Pos Pred Value   0.05472     0.09387     0.12077     0.100905
## Neg Pred Value   0.94998     0.91942     0.90808     0.902944
## Prevalence        0.05033     0.08169     0.09454     0.097420
## Detection Rate   0.00359     0.00783     0.01095     0.009541
## Detection Prevalence 0.06560     0.08342     0.09071     0.094556
## Balanced Accuracy 0.50301     0.50677     0.51390     0.501874

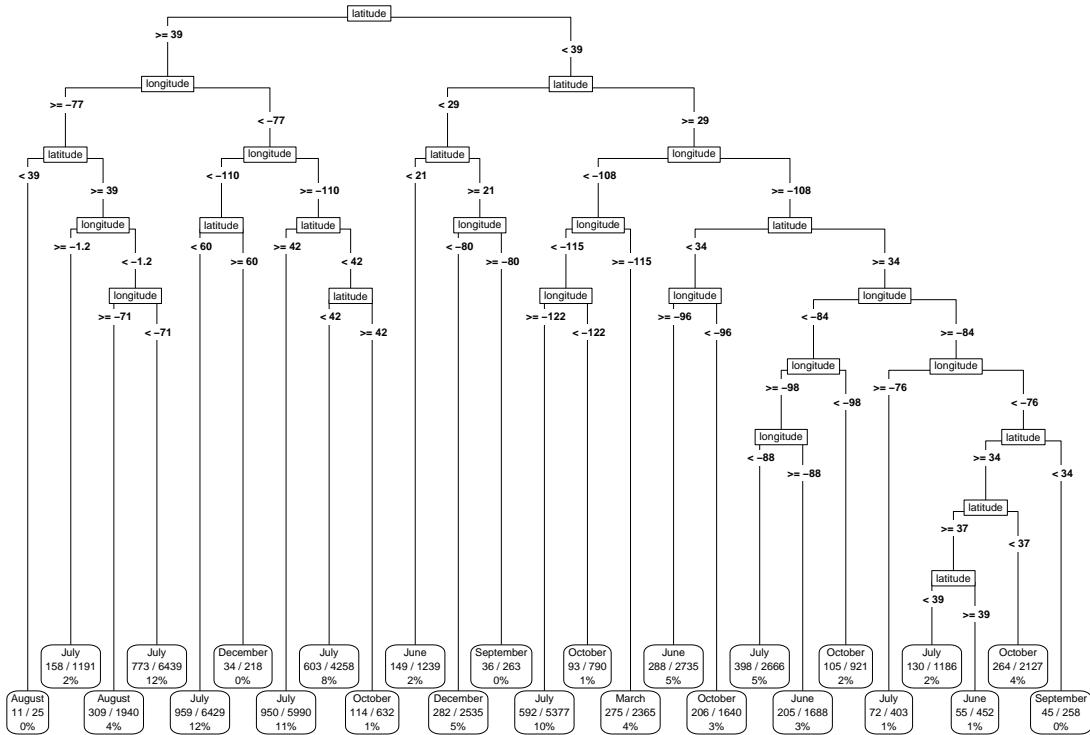
```

From the confusion matrix displayed above we see that the accuracy of this model is 42.78%. This is about 25% more accurate than predictions made based on no information. The matrix also shows that this

difference is statistically significant, with a p value of 2.2e-16. One odd thing about the model is that a k value of 1 was chosen to be the best. This might indicate that the model is overfit to the data because its only relying on the most close neighbor.

## 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. Additionally it would provide a visualization and steps to take when determining the month of a sighting.



```

##          Overall
## latitude  880.4461
## longitude 598.0308

```

The pruned classification tree is displayed above. This tree has a total of 24 leaves. The leaf that contains the most data is latitude greater than 39 and longitude less than -77. This node contains 33% of the data and classifies the month as July. The fractions on each leaf correspond to how many that were classified into that leaf were classified correctly. Most of them have low rates of correct classification which is likely due to there being a low accuracy of the model as a whole. From the results of the variable importance test we can determine that latitude is a more important predictor of month than longitude is.

## #Conclusion

Through our data analysis and machine learning we came to a few conclusions. The first is that the vast majority of UFO sightings occur in the United States with additional hot spots of sightings in places like

Europe and the eastern coast of Australia. We also found that the increase in UFO sightings began around the 1950s followed by a drastic increase in the early 2000s. Another conclusion we came to was that since 2000, a peak in sightings during the months of June and July has developed. From our regression machine learning techniques we conclude that longitude and latitude were not good predictors for sighting duration and could not produce a good regression model. Finally, we learned that classification techniques result in low accuracy for predicting the month of sightings based on longitude and latitude, but they are better than no information.

##Conclusion A major limitation of this project was the observational aspect of the data. This means our results are more varied and potentially less accurate than if there was a consistent method of recording sightings that could make this data more accurate. In the future other models could be made for the effect between longitude and latitude that had the ability to deal with the apparent gap between long and short duration sightings. Additionally, future research into the topic could collect more accurate data for how sightings change over the seasons or years by focusing in on a smaller range of the globe. Now that we have identified the US and parts of Europe as hot spots, UFO data collection could be concentrated to these locations for the highest probability of collecting a lot of data.