

# 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

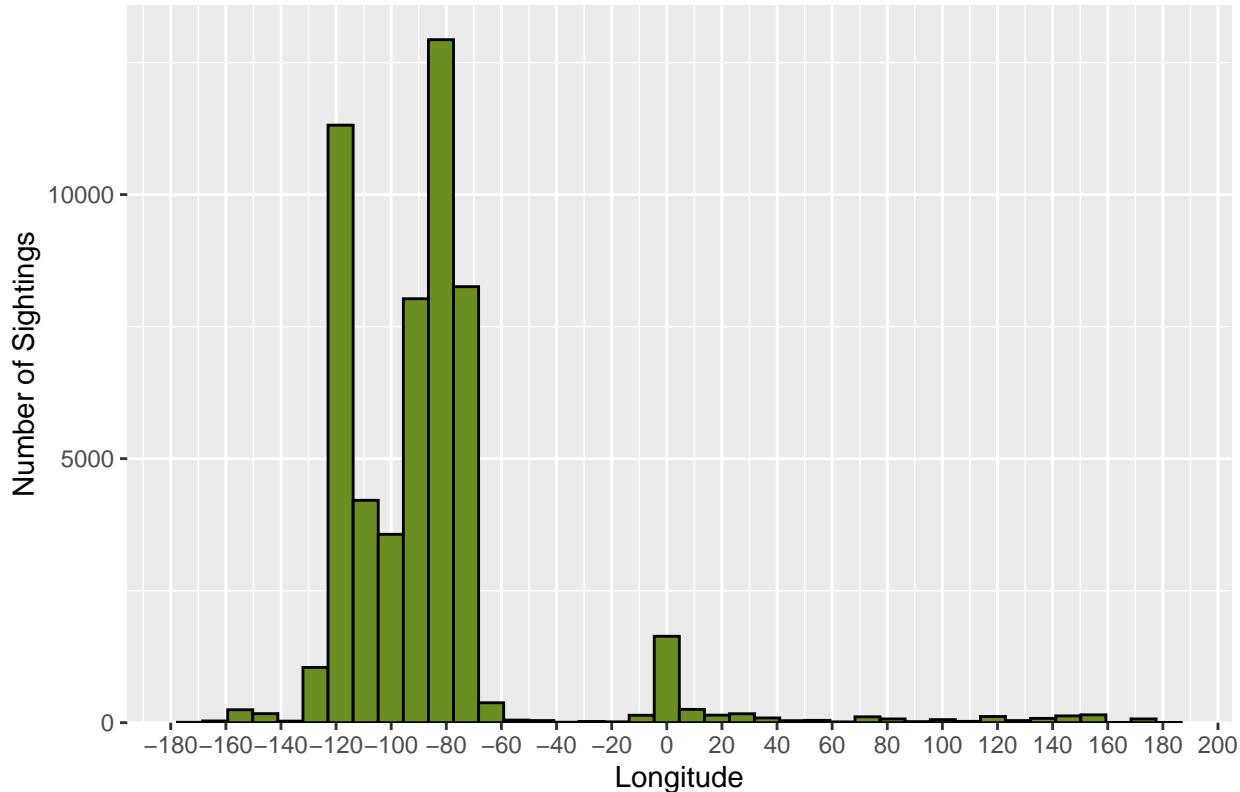
ufo$latitude <- as.numeric(ufo$latitude) #changing lat to numeric
ufo$seconds <- as.numeric(ufo$seconds) #changing seconds to numeric

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

## Data Analysis

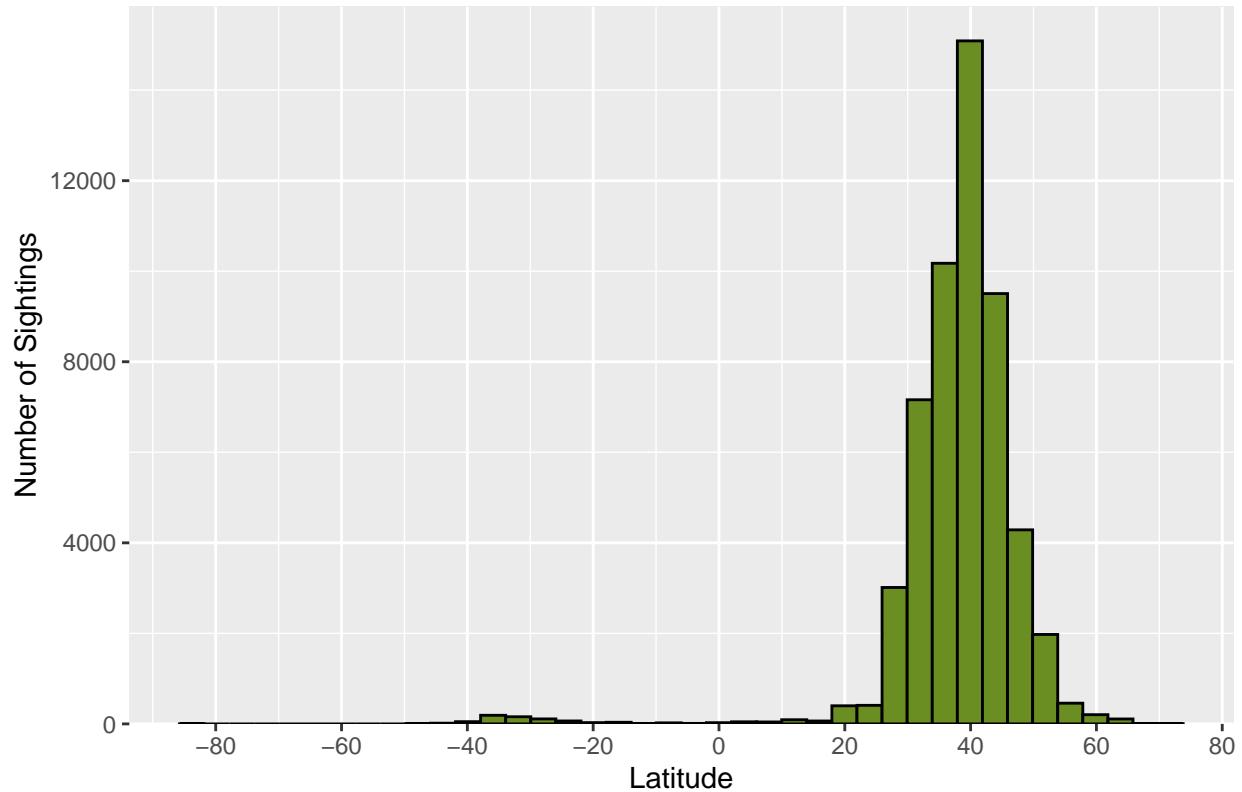
The first question we had about UFO sightings is how the location around the world influenced the number of UFO sightings that occurred. We began by creating histograms for longitude and latitude. From there we created a scatter plot to compare latitude and longitude to see if there was a relationship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

Number of sightings per longitude



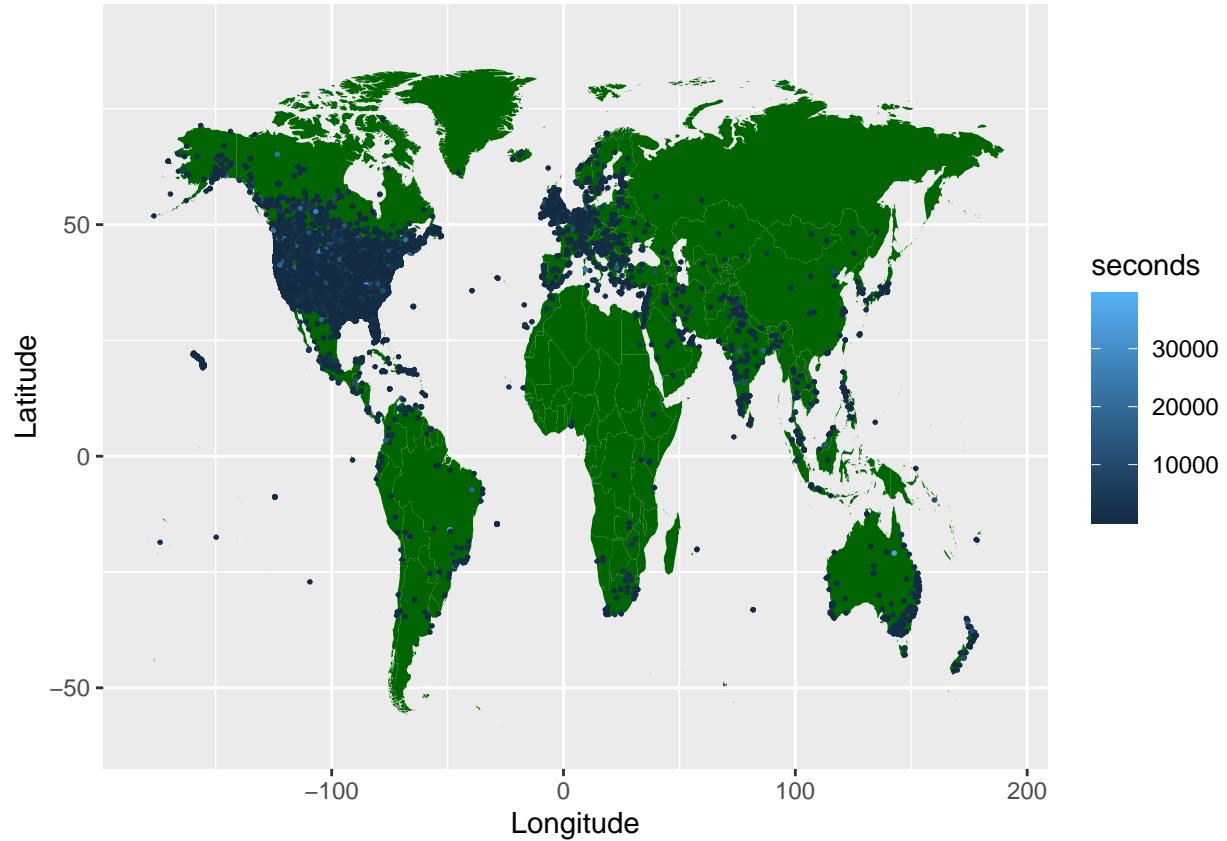
When looking at the histogram of longitude, we notice there are two main peaks within the spread. One peak is around -120 and the second is around -80. This makes sense because these are the longitudes that correspond with highly populated areas of the US. The relative height of these peaks implies that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



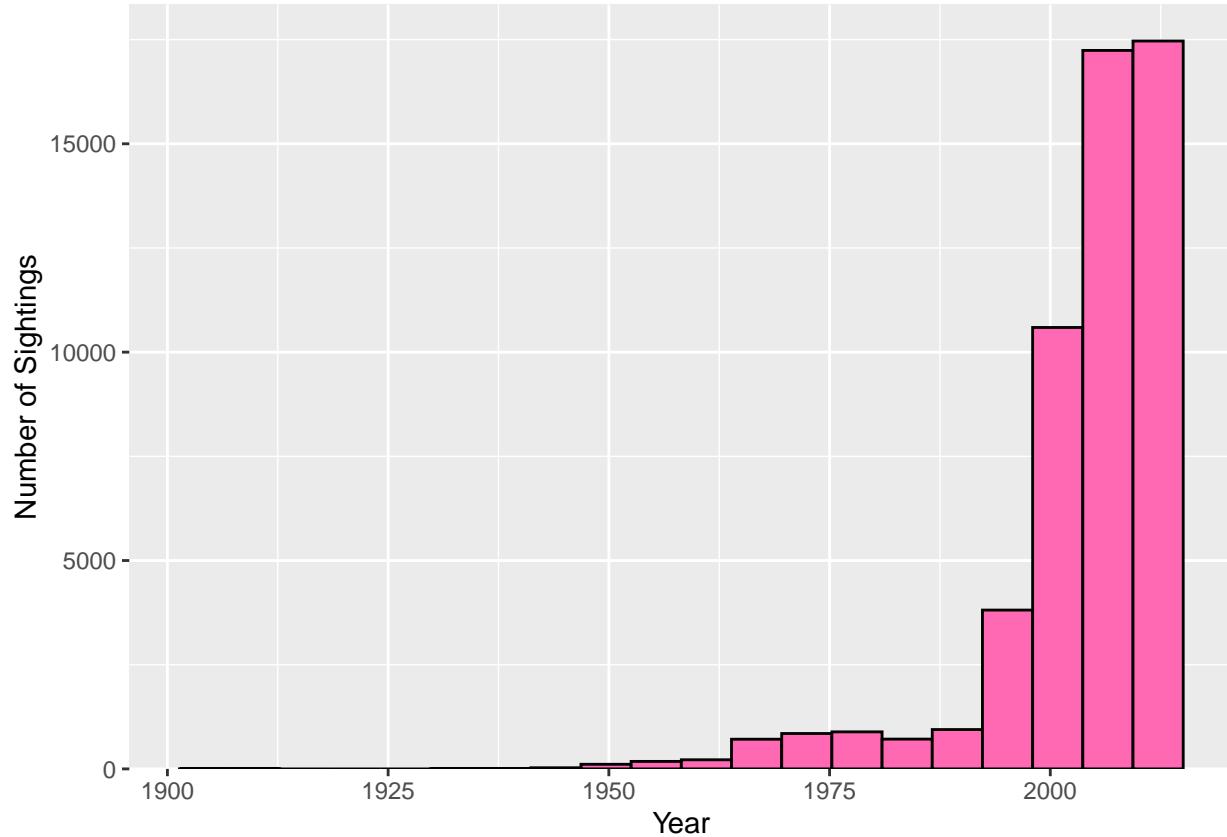
In the latitude histogram, we see than sightings are most prevalent around 40. Again, this corresponds to the coordinates of the US and also Europe, which is the second most frequently reported area of UFO sightings. This data supports the same conclusions we see from the longitude histogram.

```
## 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()').
```



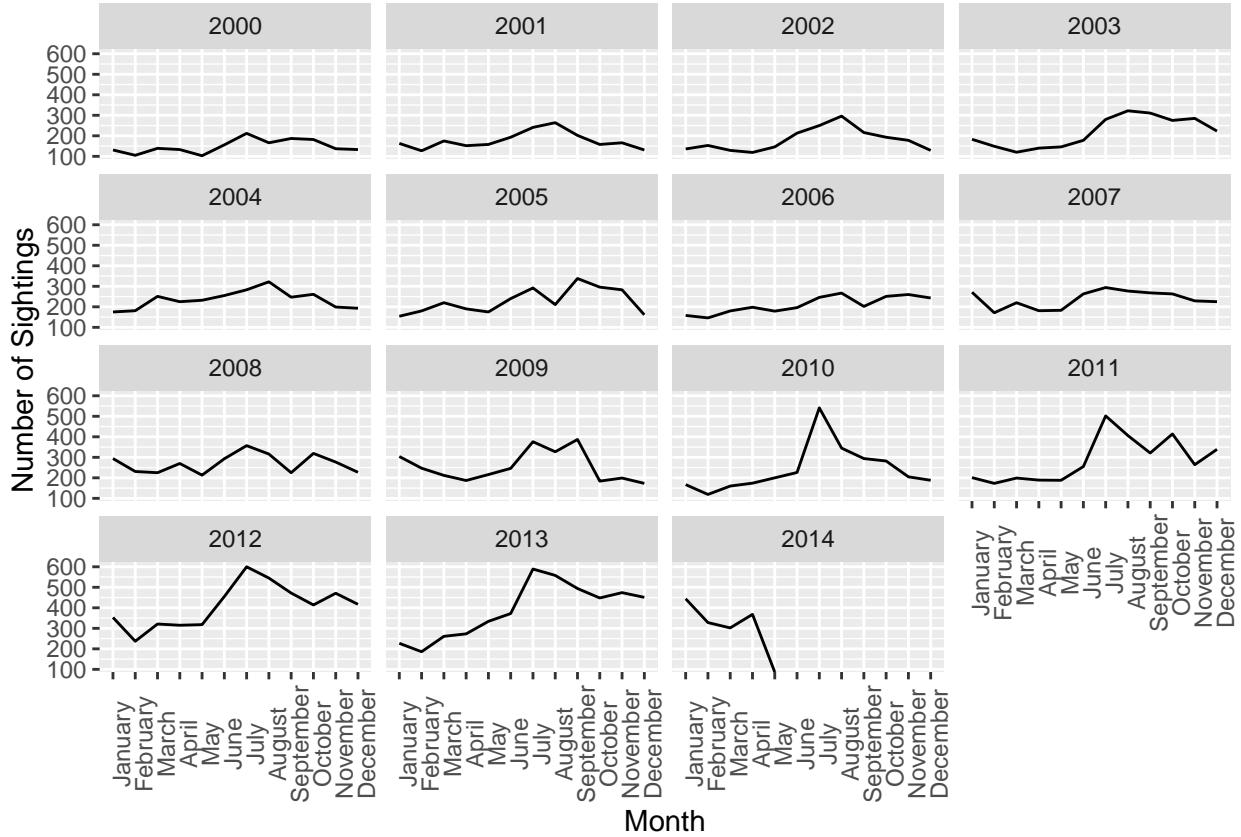
Here, we can see in greater detail the geographical dispersion of the sightings. The graph supports the same claims as the previous two graphs. We can also see from the color coding based on duration that most of the sightings are pretty short.

The next question we had was how the number of sightings had 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 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

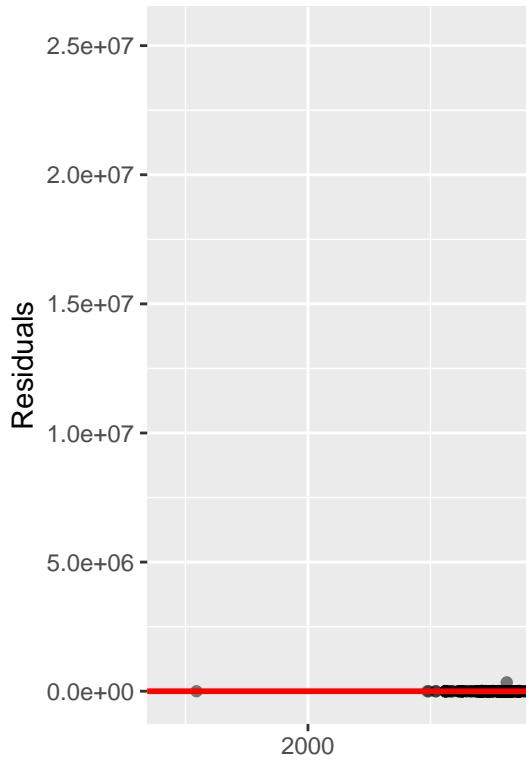
### ###Linear Regression

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
##   explanories      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.
```

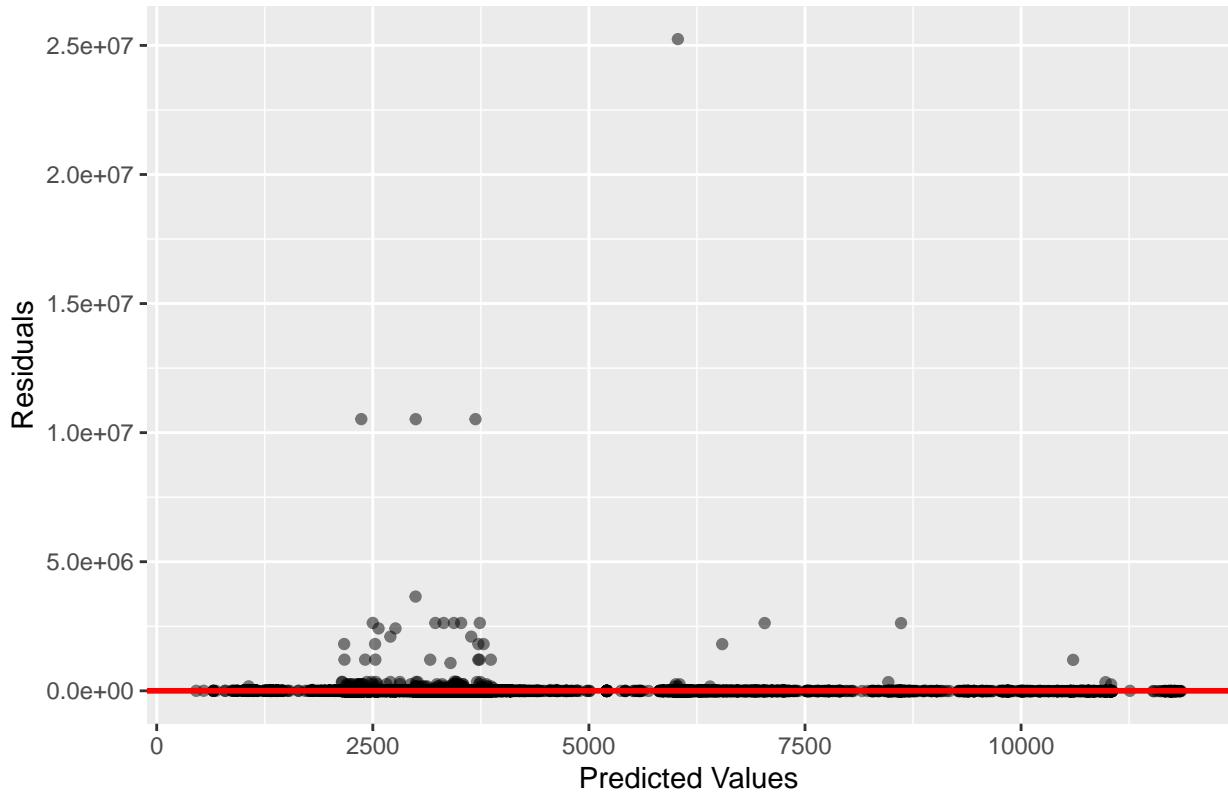


Residual Plot for Latitude

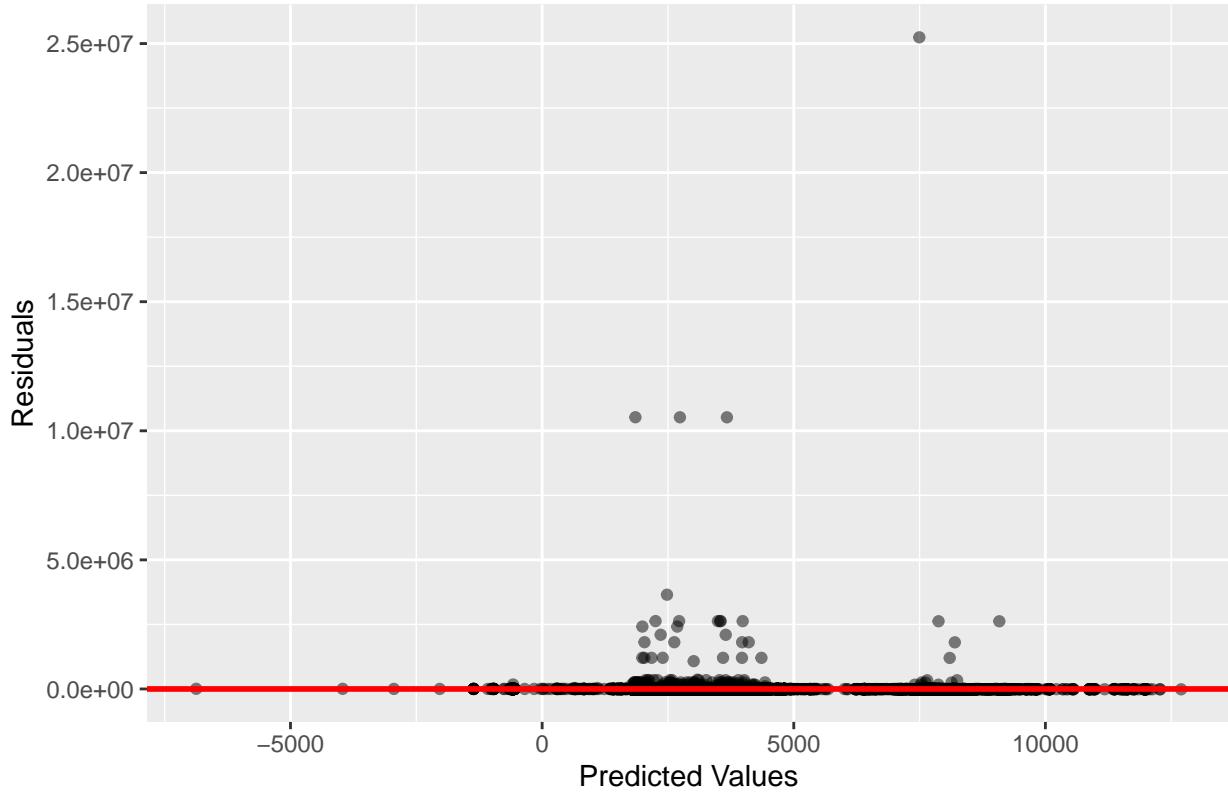


Residual Plots for Duration of Sightings vs Longitude and Lattitude

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.

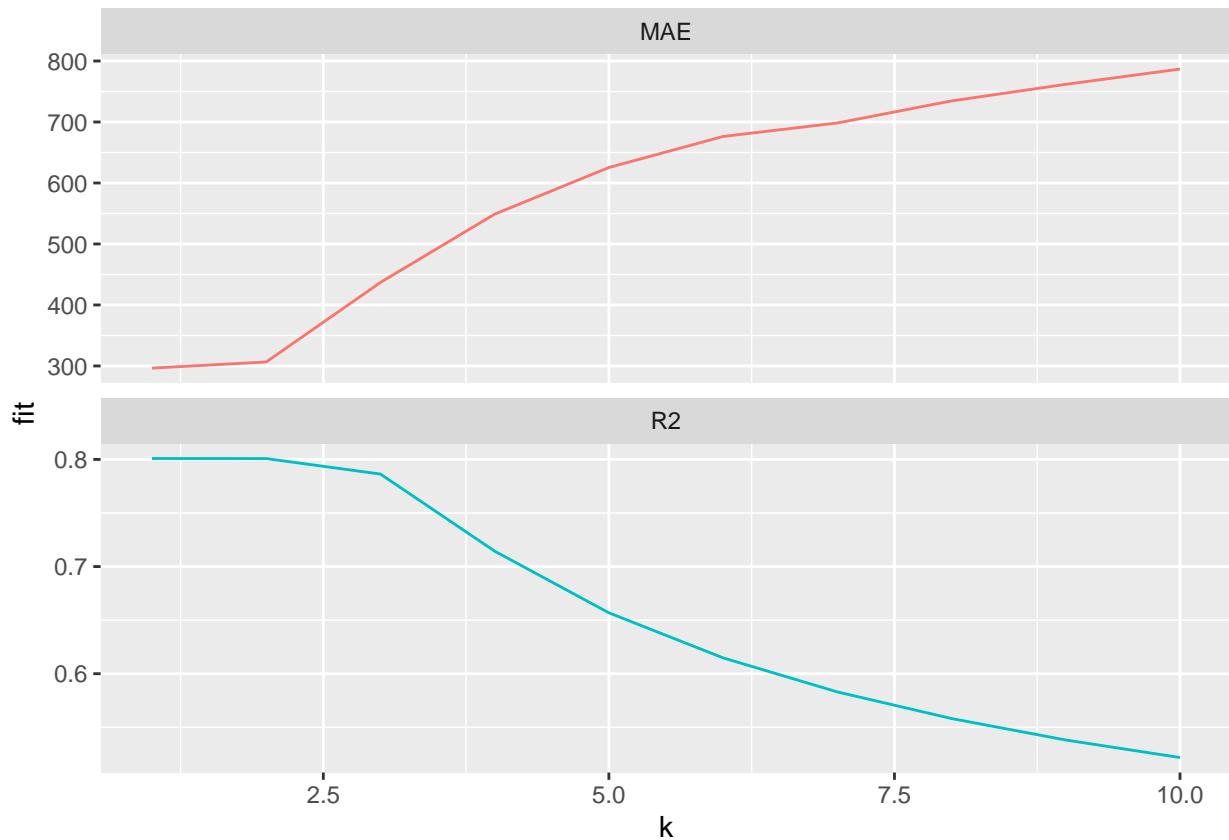
We then made residual graphs to see what else we could notice about our models and the data. These plots also indicate that the linear model is not a good predictor of duration based on longitude or latitude. In each plot there are clusters of outliers that are significantly higher than the majority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

### kNN regression

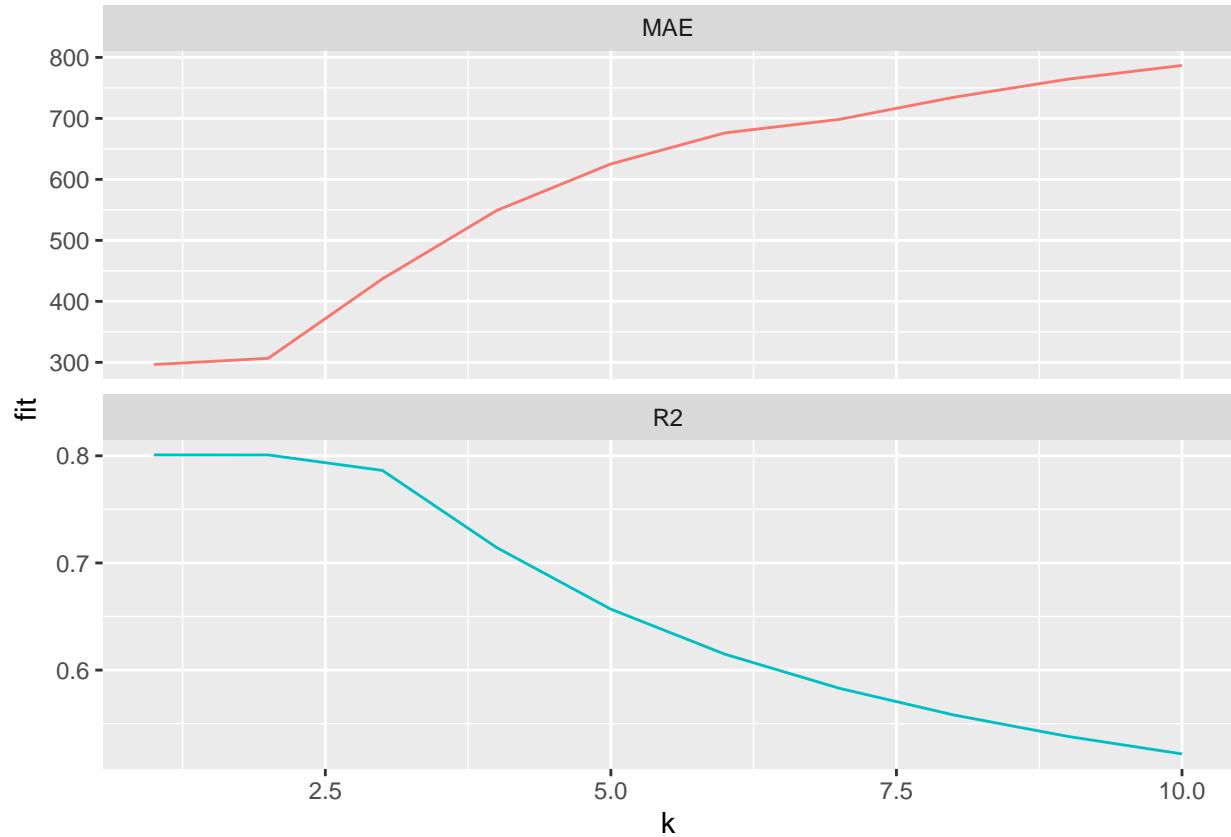
Since our linear regression proved to not be effective in predicting duration of sightings based on longitude or latitude we decided to try a k Nearest Neighbors regression. We noticed that a large portion of the data was acting differently than the rest, and thought kNN regression could account for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

```
##      k      R2      MAE
## 1    1 0.8008550 296.4860
## 2    2 0.8007957 306.5553
## 3    3 0.7863550 436.9400
## 4    4 0.7143796 548.9220
## 5    5 0.6568458 625.3485
## 6    6 0.6147682 676.2930
## 7    7 0.5830238 698.3884
## 8    8 0.5581418 734.5842
```

```
## 9 9 0.5381227 761.7924
## 10 10 0.5217339 786.8153
```



```
##      k      R2      MAE
## 1 1 0.8008550 296.4874
## 2 2 0.8007957 306.5569
## 3 3 0.7863550 436.9426
## 4 4 0.7143796 548.9410
## 5 5 0.6568458 625.3444
## 6 6 0.6147682 676.2891
## 7 7 0.5830238 698.3867
## 8 8 0.5581418 734.5817
## 9 9 0.5381054 764.3026
## 10 10 0.5217339 786.8307
```



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

The fifth table show that the best k value to use is 1, which is typically not a value that produces accurate regression models. Because of this, we see that kNN regression is also not a viable method of machine learning.