UFO

Shealagh Brown & Sam Zimpfer

2025-03-20

Introduction

Cleaning the data

The code chunk below reads in our data about UFO sightings. We had to convert the datetime variable to year, month, date format. Then we created a new data frame containing only necessary variables for our purposes. Latitude had to then be convered into a numeric variable.

The variables used are: **datetime**: the date and time of sighting in year, month, day, hours, minutes, seconds format.
city: city where sighitng occurred
state: state where sighting occurred
country: country where sighting occurred
seconds: duration of sightings in seconds
latitude: latitude of sighting
longitude: longitude of sighing
year: year when sighing occurred
month: month when sighting occured

```
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) -> ufo

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

```
## Warning: NAs introduced by coercion
```

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

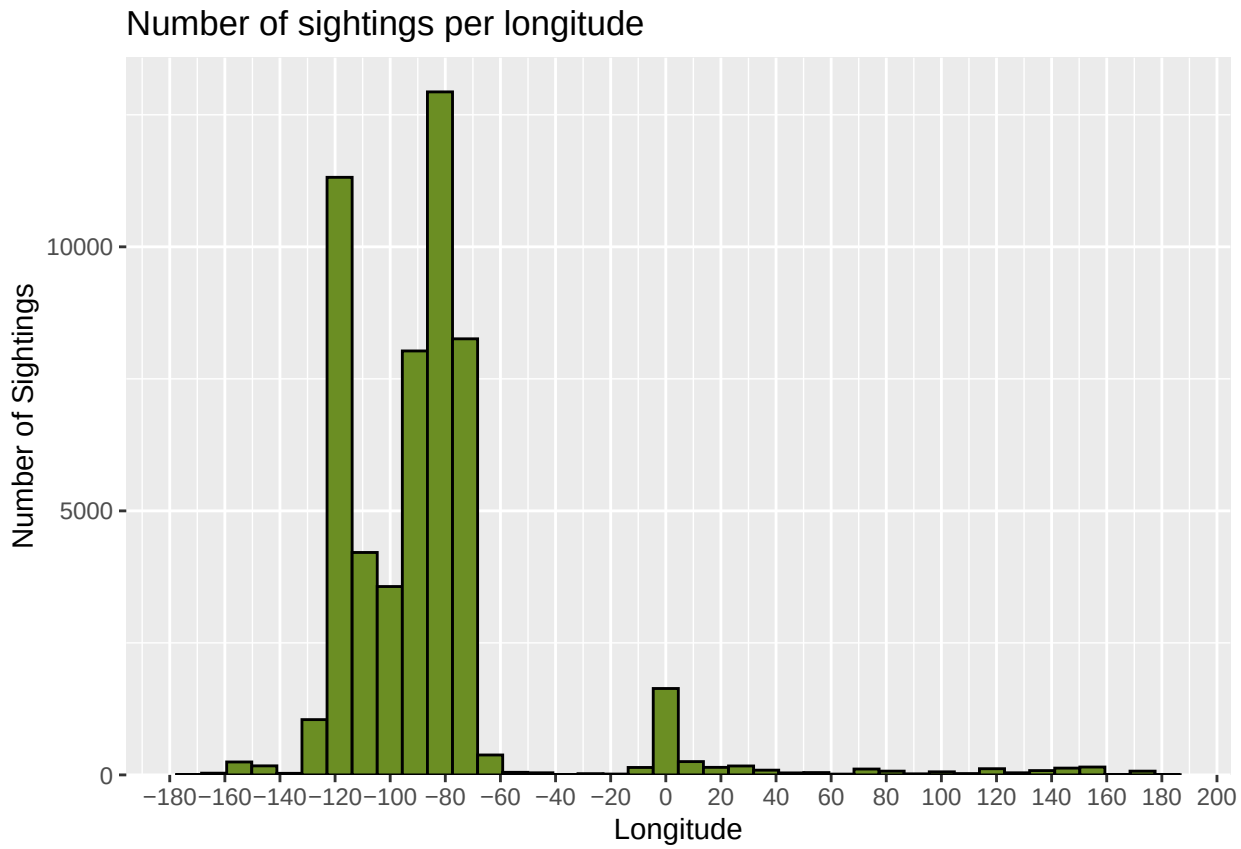
```
## Warning: NAs introduced by coercion
```

Data Summary

The first question we wanted to answer is where do most UFO sightings occur.

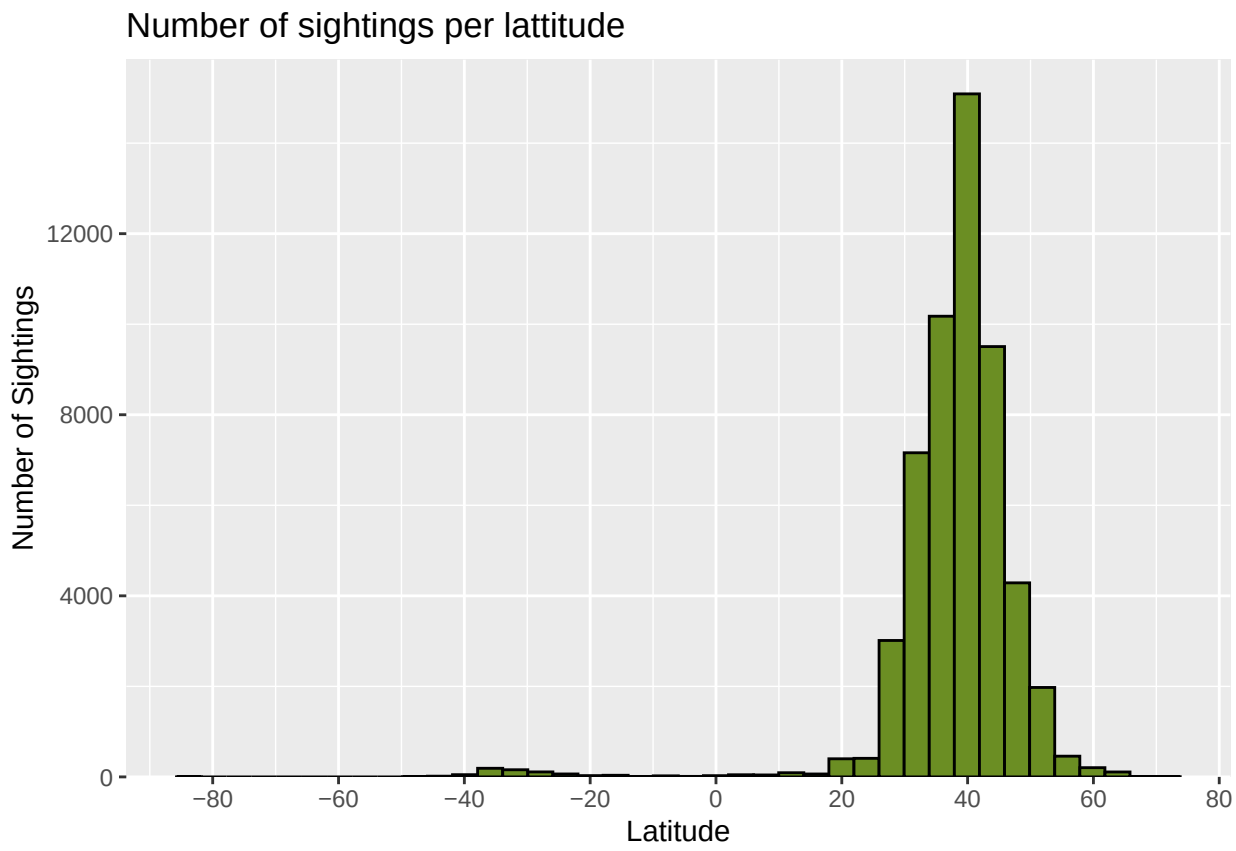
We created graphs of sighingts based on longitude and latitudue

```
long_hist <- ggplot(
  data = ufo,
  mapping = aes( x= longitude
  )
)+
  geom_histogram(fill = "olivedrab",
                 color = "black",
                 bins = 40)+
  labs(
    title = "Number of sightings per longitude",
    x = "Longitude",
    y = "Number of Sightings"
  )+
  scale_x_continuous(breaks = seq(-200, 200, 20))+
  scale_y_continuous(expand = c(0, 0, 0.05, 0))

lat_hist <- ggplot(
  data = ufo,
  mapping = aes( x= latitude
  )
)+
  geom_histogram(fill = "olivedrab",
                 color = "black",
                 bins = 40)+
  labs(
    title = "Number of sightings per lattitude",
    x = "Latitude",
    y = "Number of Sightings"
  )+
  scale_x_continuous(breaks = seq(-100, 100, 20))+
  scale_y_continuous(expand = c(0, 0, 0.05, 0))

long_hist
```




```
lat_hist
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_bin()‘).
```



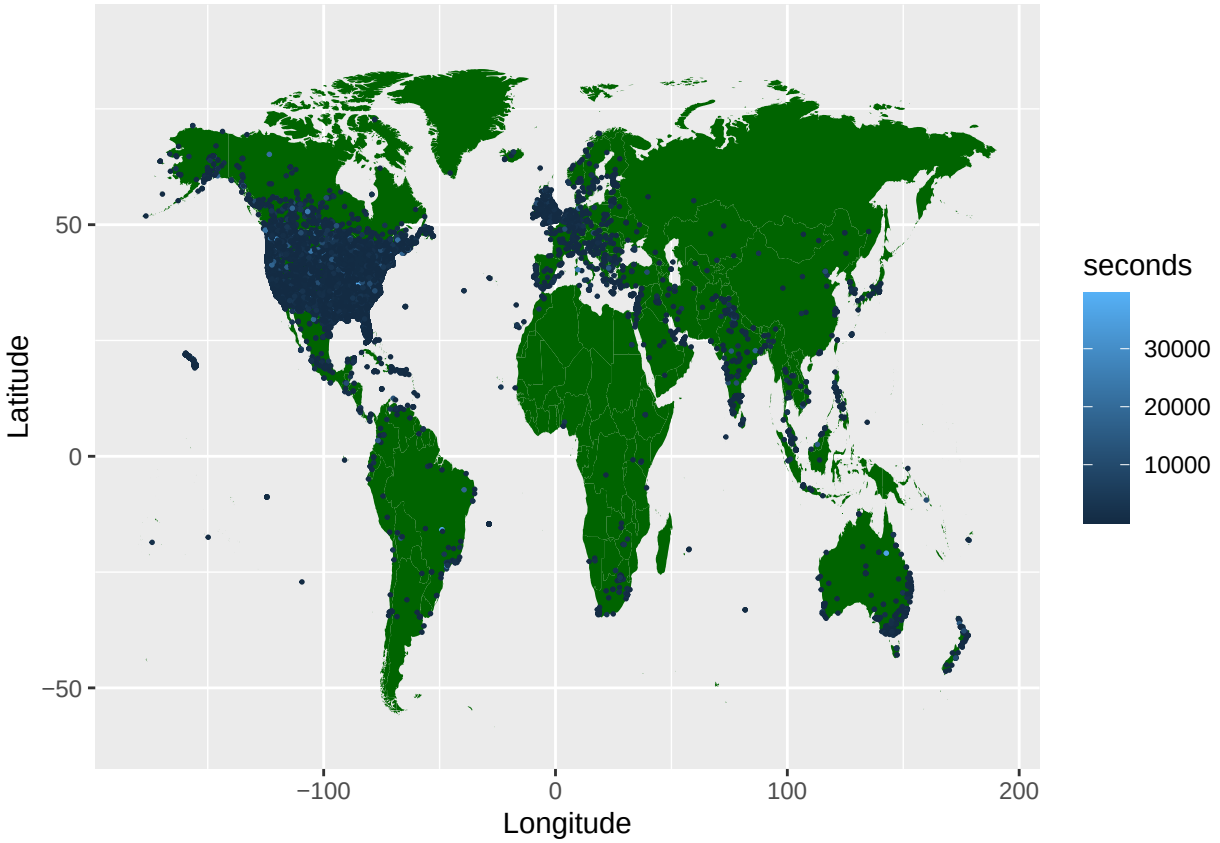
Scatter plot of sighting locations

```
map <- map_data("world")

map_plot <- ggplot()+
  geom_polygon(data = map,
               mapping = aes(x= long,
                             y = lat,
                             group = group),
               fill = "darkgreen")+
  geom_point(data = filter(ufo, seconds < 40000),
             mapping = aes(x = longitude,
                           y = latitude,
                           color = seconds),
             size = 0.3
           )+
  labs(
    x = "Longitude",
    y = "Latitude"
  )+
scale_y_continuous(expand = c(0, 0, 0.05, 0))+
  ylim(-60,90)
```

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

```
map_plot
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

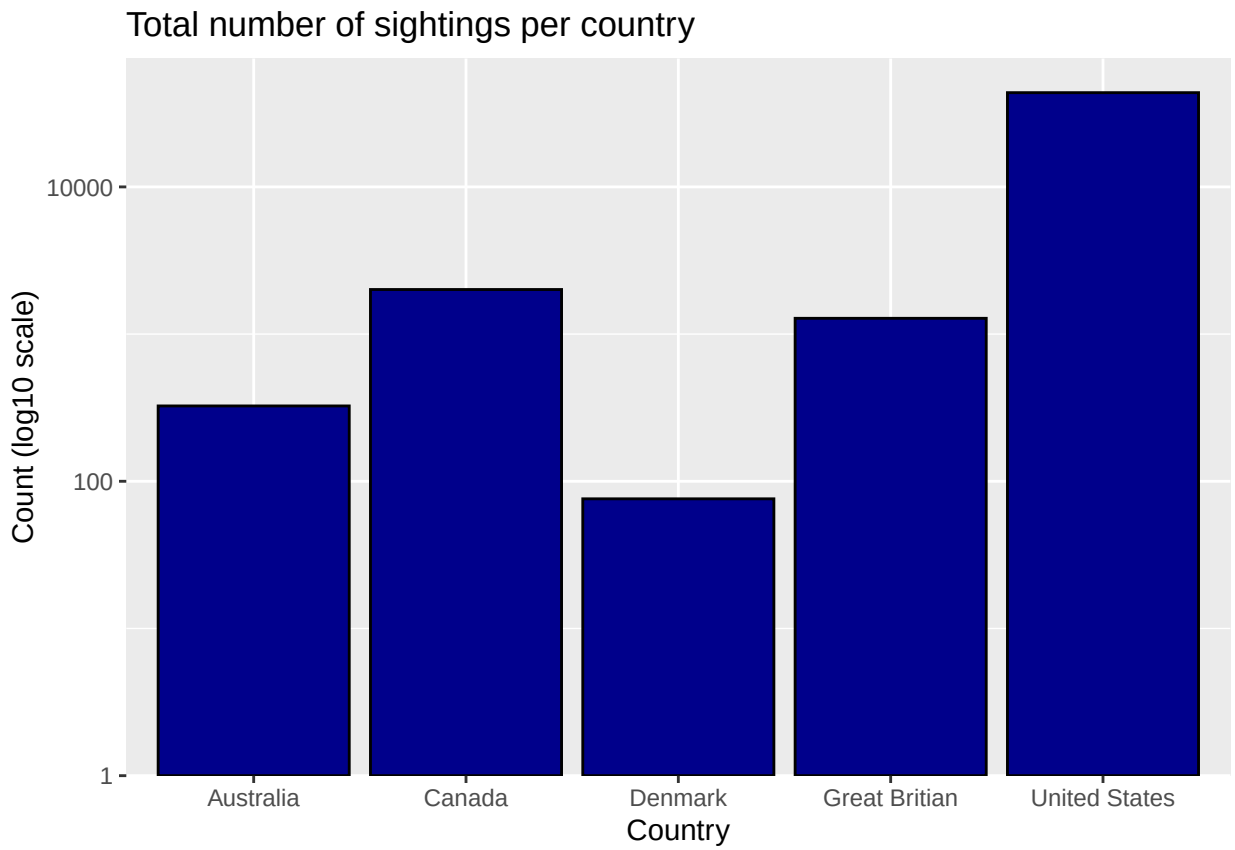


Observations

From the histograms, we noticed that most ufo sightings are concentrated around -120 through -70 longitude and 30 through 50 latitudes. When longitude and latitude are graphed against each other and over layed with a map of the word we can see that these values correspond with 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.

Bar graph of sightings per country

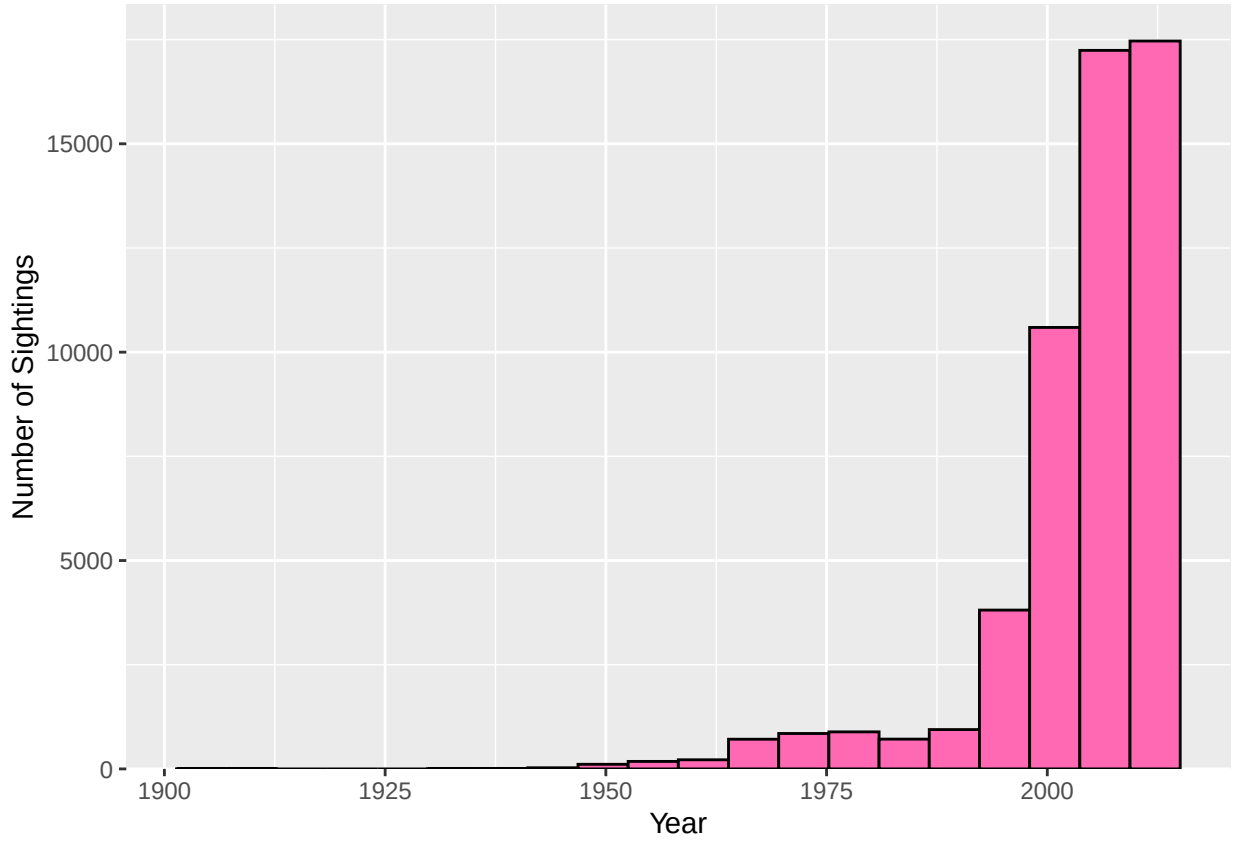
```r
modified_data = ufo |> filter(country != "")

ggplot(data = modified_data,
       mapping = aes(x = country))+
  geom_bar(fill = "blue4",
           color = "black",
           na.rm = T)+
scale_y_log10(expand = c(0, 0, 0.05, 0))+
  labs(title = "Total number of sightings per country",
       x = "Country",
       y = "Count (log10 scale)")+
scale_x_discrete(labels = c("Australia", "Canada", "Denmark", "Great Britian", "United States"))
```



From this graph we can determine that the United States is the country with the most sightings. There is a signficantly higher amount of sightings in the U.S. as opposed to the other countries as it is the only country with greater than 10,000 sightings. Of the five highest countries with sighitngs, Denmark has the lowest amount of sightings. They have just around 100 sighitngs reported.

Bar graph of sightings per year

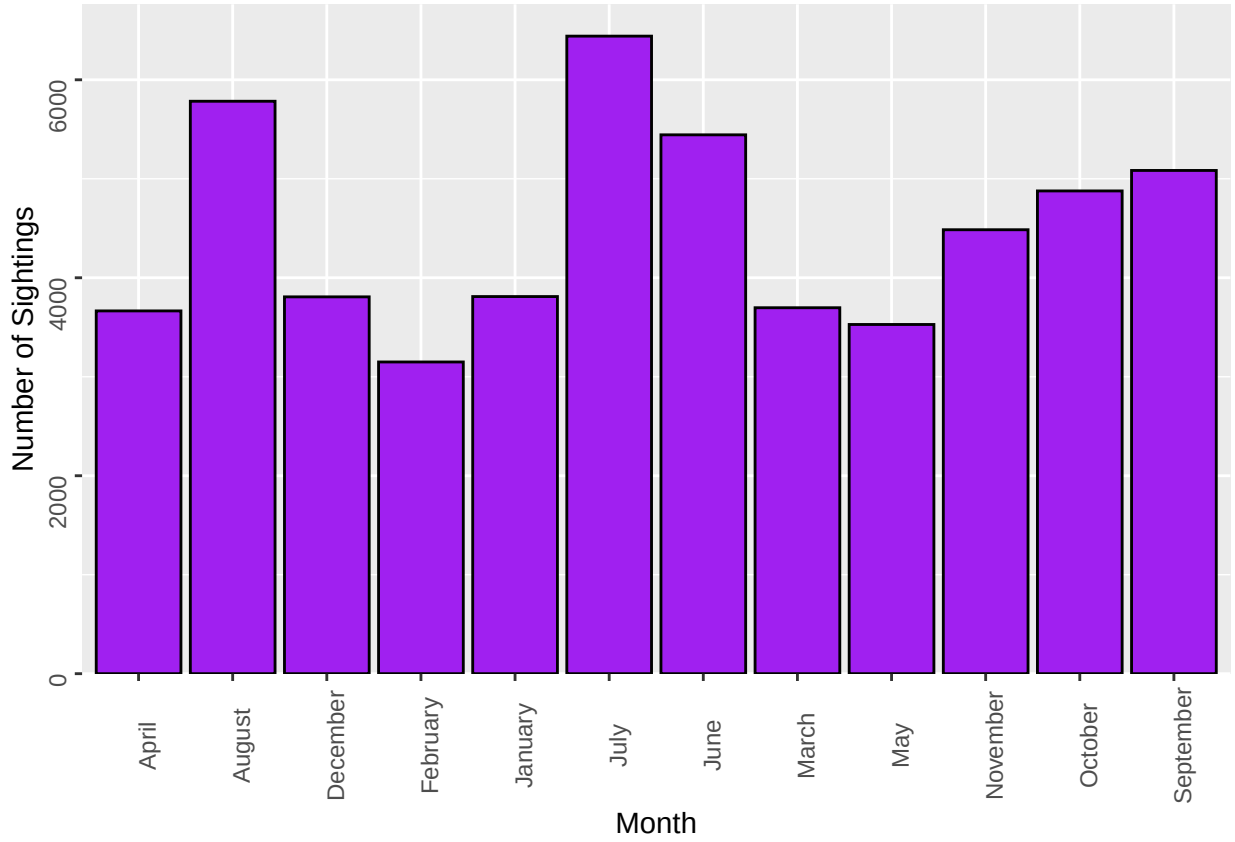
```
ggplot( data = ufo,
        mapping = aes(x = year))+
  geom_histogram(fill = "hotpink",
                 color = "black",
                 bins = 20)+
  labs(
    x = "Year",
    y = "Number of Sightings"
  )+
scale_y_continuous(expand = c(0, 0, 0.05, 0))
```



Observations 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.

```
by_month <- ggplot( data = ufo,
        mapping = aes(x = month))+
  geom_bar(fill = "purple",
                 color = "black")+
  labs(
    x = "Month",
    y = "Number of Sightings"
  )+
scale_y_continuous(expand = c(0, 0, 0.05, 0))+
theme(
  axis.text = element_text( angle = 90
```

```
    )
)
by_month
```



Observations Over the years, UFO sightings have been most common during June, July, and August. The fewest occurances happend during February.

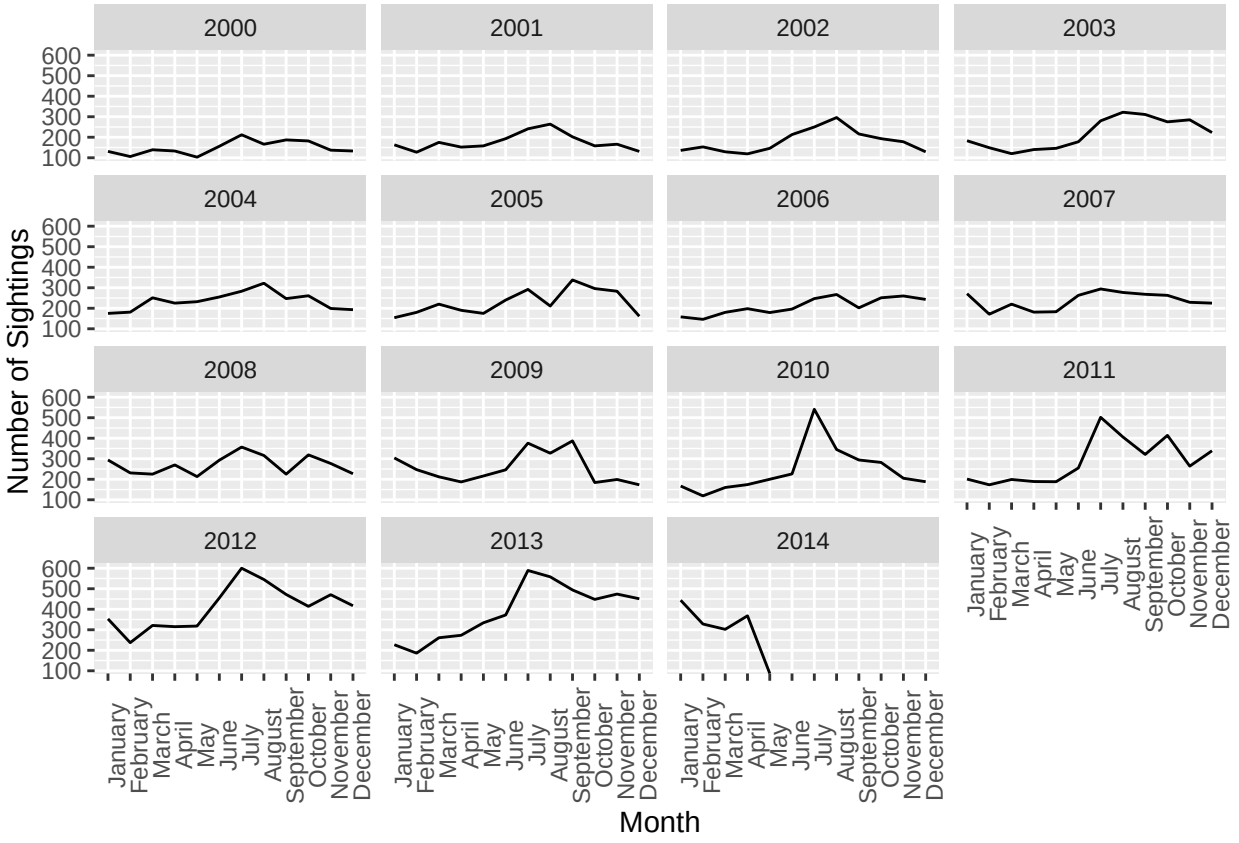
Small Multiples Graph of Number of Sightings Per Month Since Year 2000

```
# create table of year, month, number of sightings
summary <- ufo |> filter(year >= 2000) |> group_by(year, month) |> summarise(count = n())
```

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

```
# order data by natural month ordering
summary$month <- factor(summary$month, levels = month.name)
summary <- summary |> arrange(year, month)

ggplot(data = summary,
       mapping = aes(x = month,
```

```
                    y = count,
                    group = 1)) +
  geom_line() +
  labs(x = "Month",
       y = "Number of Sightings") +
  scale_y_continuous(expand = c(0, 0, 0.05, 0)) +
  theme(axis.text.x = element_text(angle = 90)) +
  facet_wrap(~ year)
```



Observations

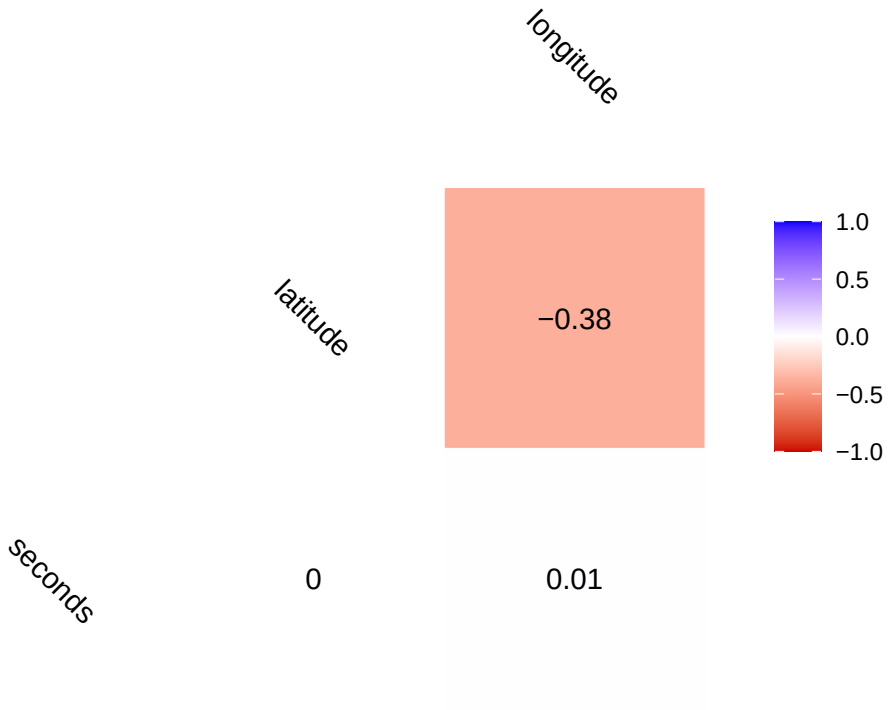
When we compare how sightings per month has changed over time we see that generally the amount of sighitngs has increased. Before 2009 sightinga are generally conisistent over the year. In 2009 we can see the first real peak during the summer. This trend seems to be reflected in the following years since then.

#Machine Learning

We want to conduct linear regression to determine the effect that longitude and latitude have on the durration of sighitngs. Before we create a model we will check for

Correlation Plot for lattitude, longitude, and seconds

```
ufo|>
  select(seconds:longitude)|>
ggcorr(
  low = "red3",
  mid = "white",
  high ="blue",
  label = T,
  label_round = 2,
  angle = -45
)
```




```
lat_lm<- lm(
  formula = seconds ~ latitude,
  data = ufo
)

long_lm<- lm(
  formula = seconds ~ longitude,
  data = ufo
)

lat_long_lm<- lm(
```

```
  formula = seconds ~ latitude + longitude,
  data = ufo
)

bind_rows(
  glance(lat_lm),
  glance(long_lm),
  glance(lat_long_lm)
  ) |>

  mutate(
    explanatories = c(as.character(formula(lat_lm))[3],
                      as.character(formula(long_lm))[3],
                      as.character(formula(lat_long_lm))[3])
  ) |>
select(explanatories, r.squared, adj.r.squared, sigma)
```

```
## # 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 142614.
## 3 latitude + longitude 0.0000999       0.0000627 142615.
```

```
lat_long_lm |>
  tidy()
```

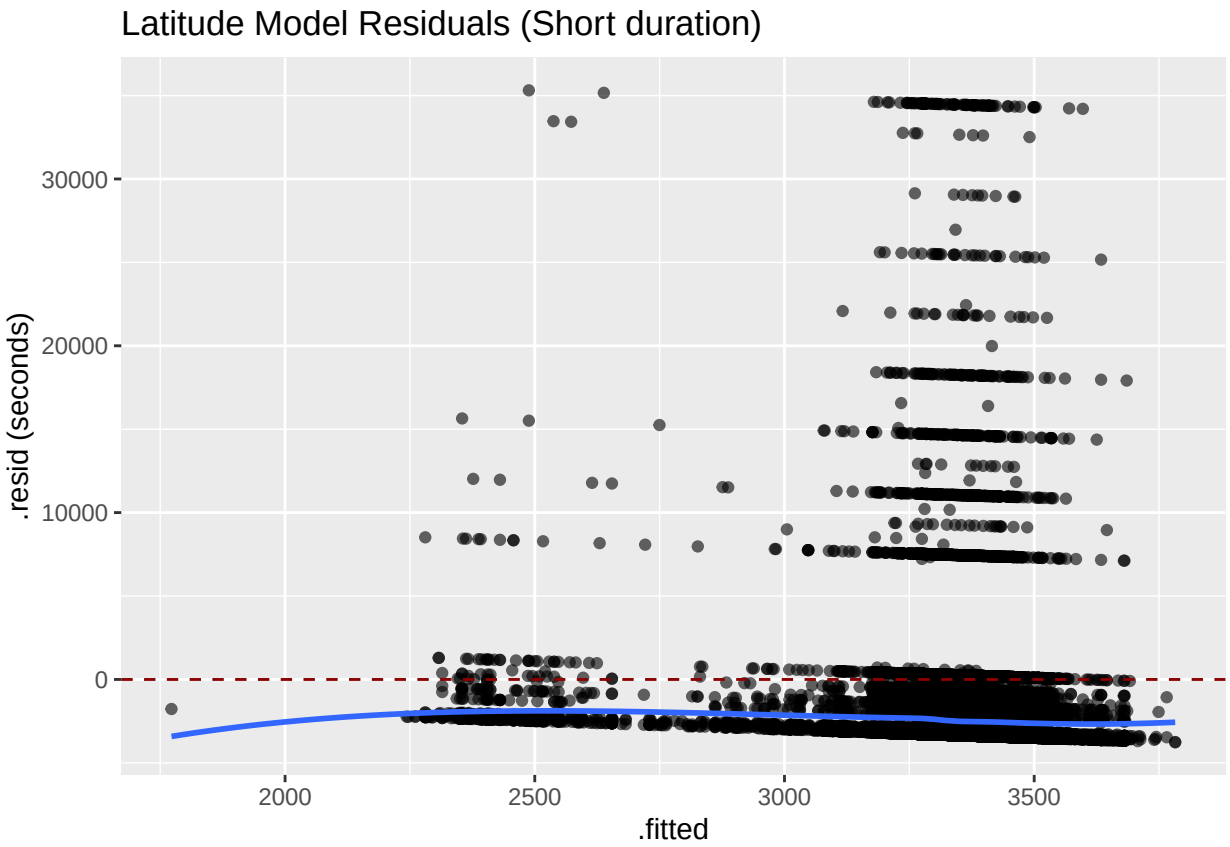
```
## # A tibble: 3 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   4093.     2409.       1.70  0.0893
## 2 latitude        68.9      64.1      1.07  0.283
## 3 longitude       39.0      16.9      2.31  0.0210
```

Residual Plots for Duration of Sightings vs Longitude and Lattitude

```
show_resid_plot <- function(model, name) {
  augment_columns(x = model,
                  data = ufo_trimmed) |>
    ggplot(mapping = aes(x = .fitted,
                         y = .resid)) +
    geom_point() +
    geom_hline(yintercept = 0,
               color = "darkred",
               linetype = "dashed",
               linewidth = 1) +
    geom_smooth(method = "loess",
                se = F) +
```

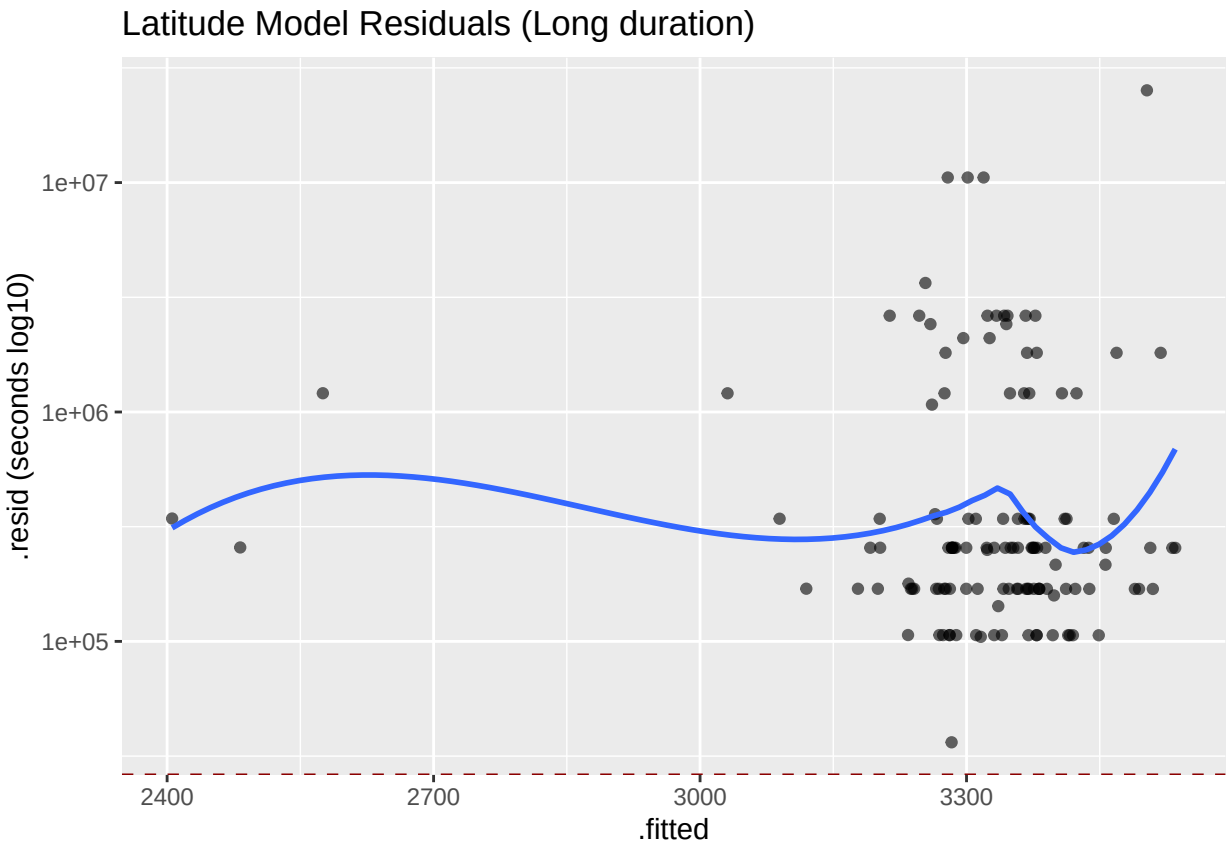
```
    labs(title = paste(name, "Residuals"),
         y = ".resid (seconds)")
}

show_resid_plots_split <- function(model, name) {
  # classify data as short duration or long duration
  augmented <- augment_columns(x = model,
                               data = ufo) |>
    mutate(long = abs(.resid) > 0.25 * sd(.resid, na.rm = TRUE))

  graph_short <- augmented |>
    filter(!long) |>
    ggplot(aes(x = .fitted,
               y = .resid)) +
    geom_point(alpha = 0.6) +
    geom_hline(yintercept = 0,
               linetype = "dashed",
               color = "darkred") +
    geom_smooth(method = "loess",
                se = FALSE) +
    labs(title = paste(name, "Residuals (Short duration)"),
         y = ".resid (seconds)")

  print(graph_short)


  graph_long <- augmented |>
    filter(long) |>
    ggplot(aes(x = .fitted,
               y = .resid)) +
    geom_point(alpha = 0.6) +
    geom_hline(yintercept = 0,
               linetype = "dashed",
               color = "darkred") +
    geom_smooth(method = "loess",
                se = FALSE) +
    labs(title = paste(name, "Residuals (Long duration)"),
         y = ".resid (seconds log10)") +
    scale_y_log10()

  print(graph_long)
}

show_resid_plots_split(lat_lm, "Latitude Model")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



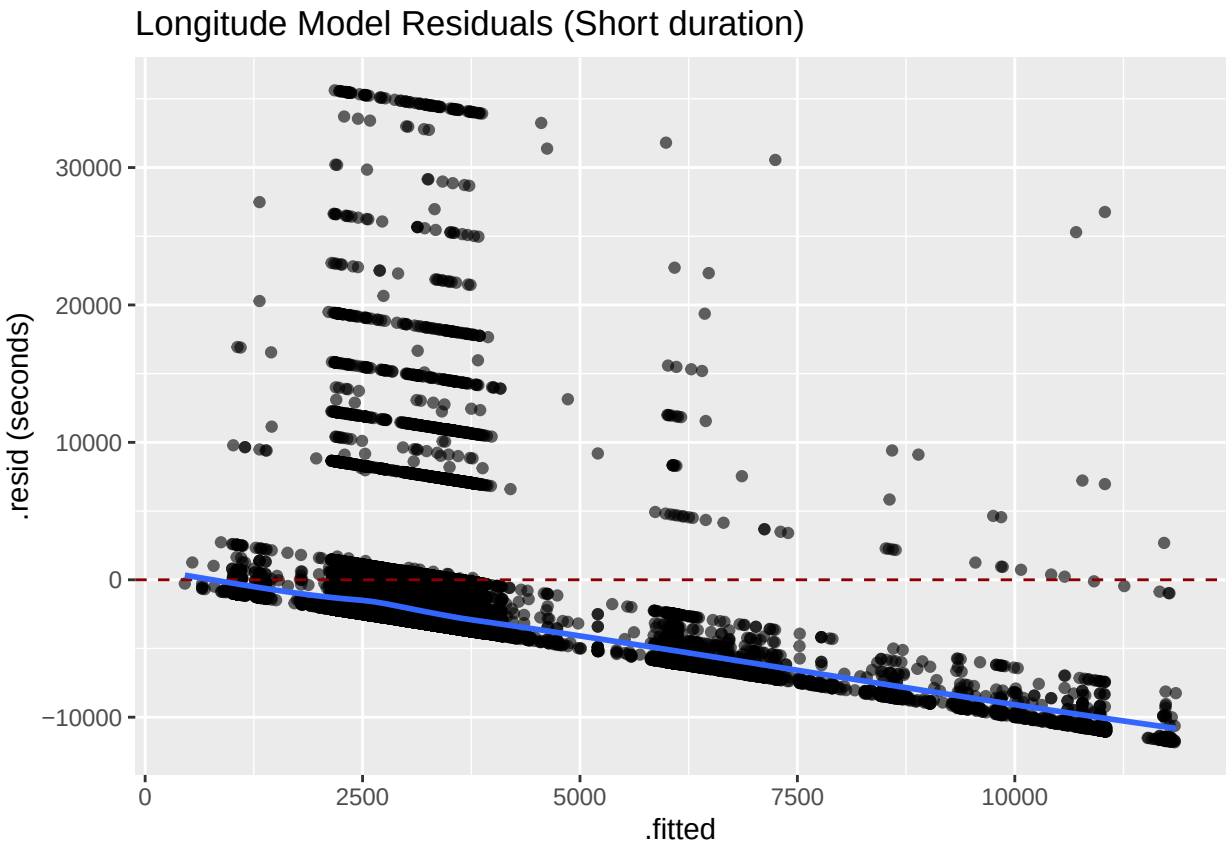

```
## Warning in scale_y_log10(): log-10 transformation introduced infinite values.
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



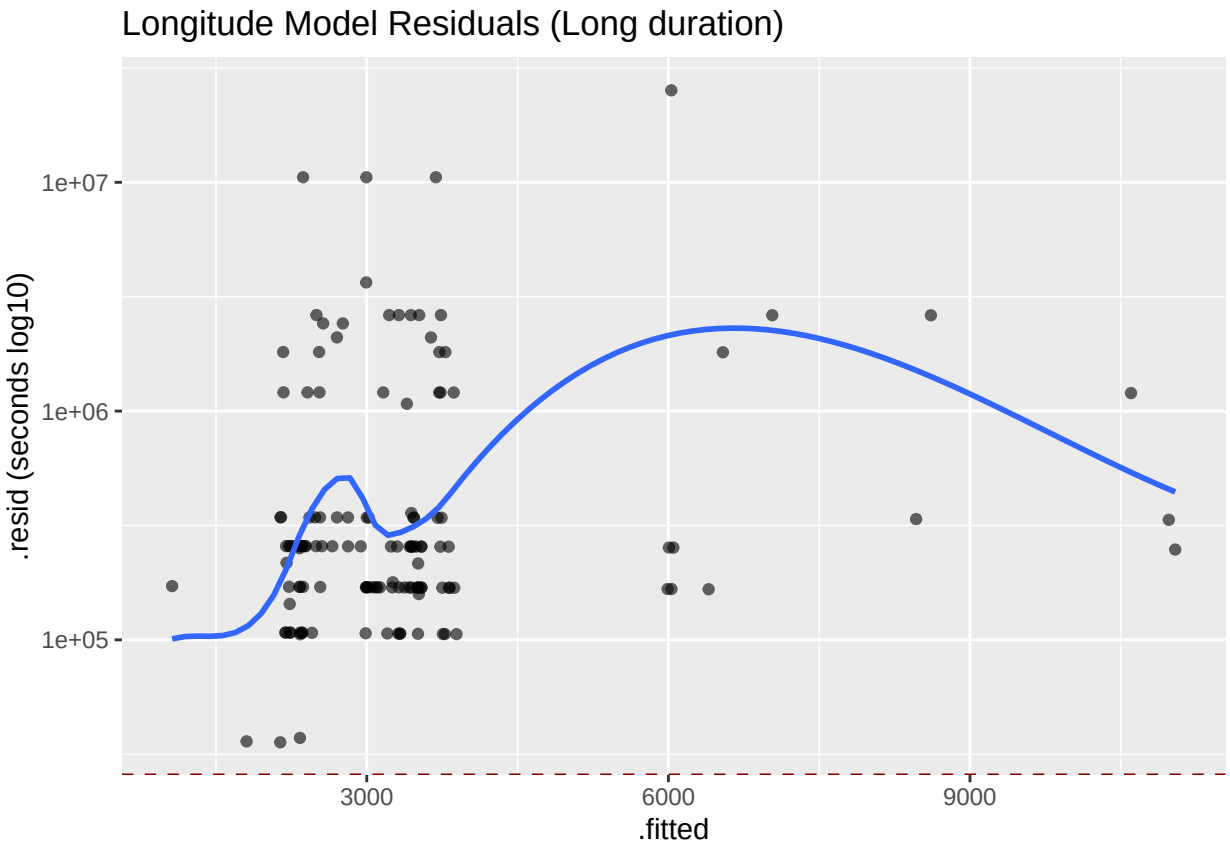

```
show_resid_plots_split(long_lm, "Longitude Model")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



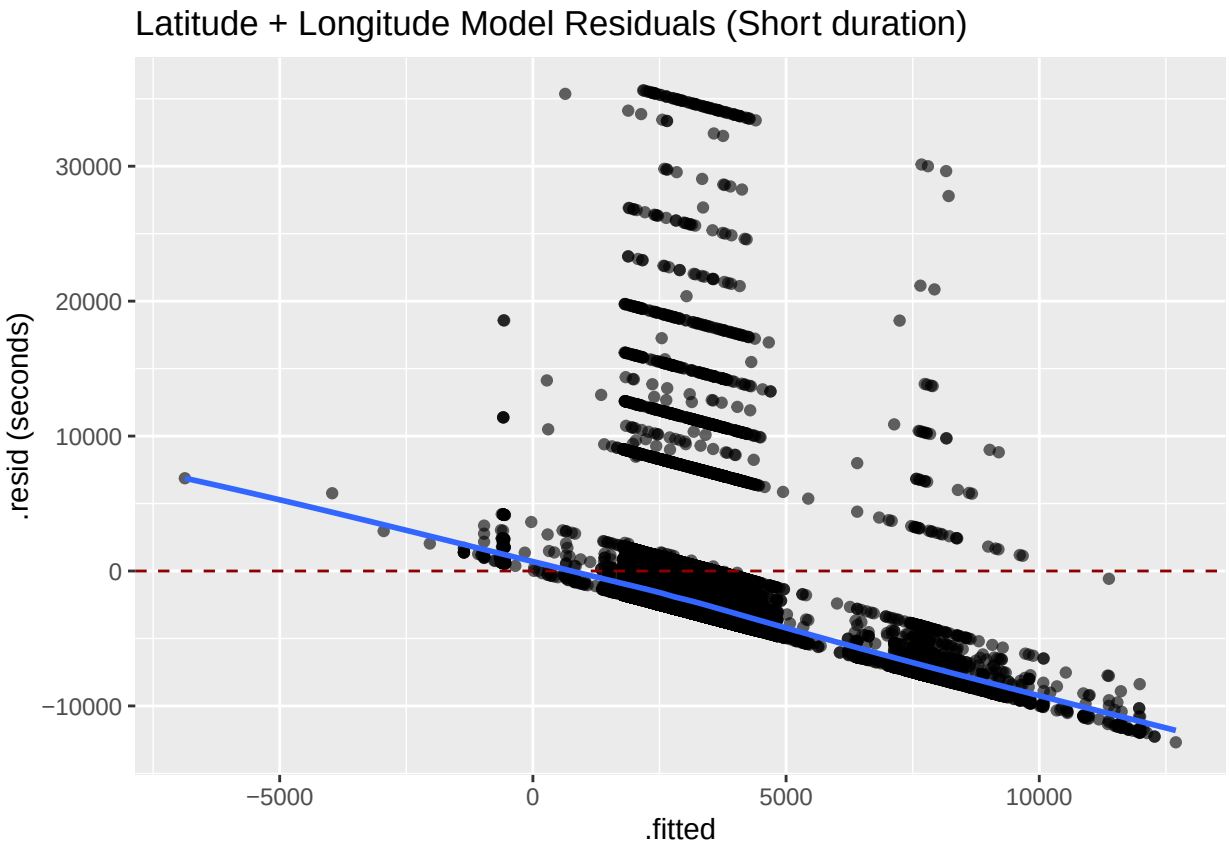

```
## Warning in scale_y_log10(): log-10 transformation introduced infinite values.
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



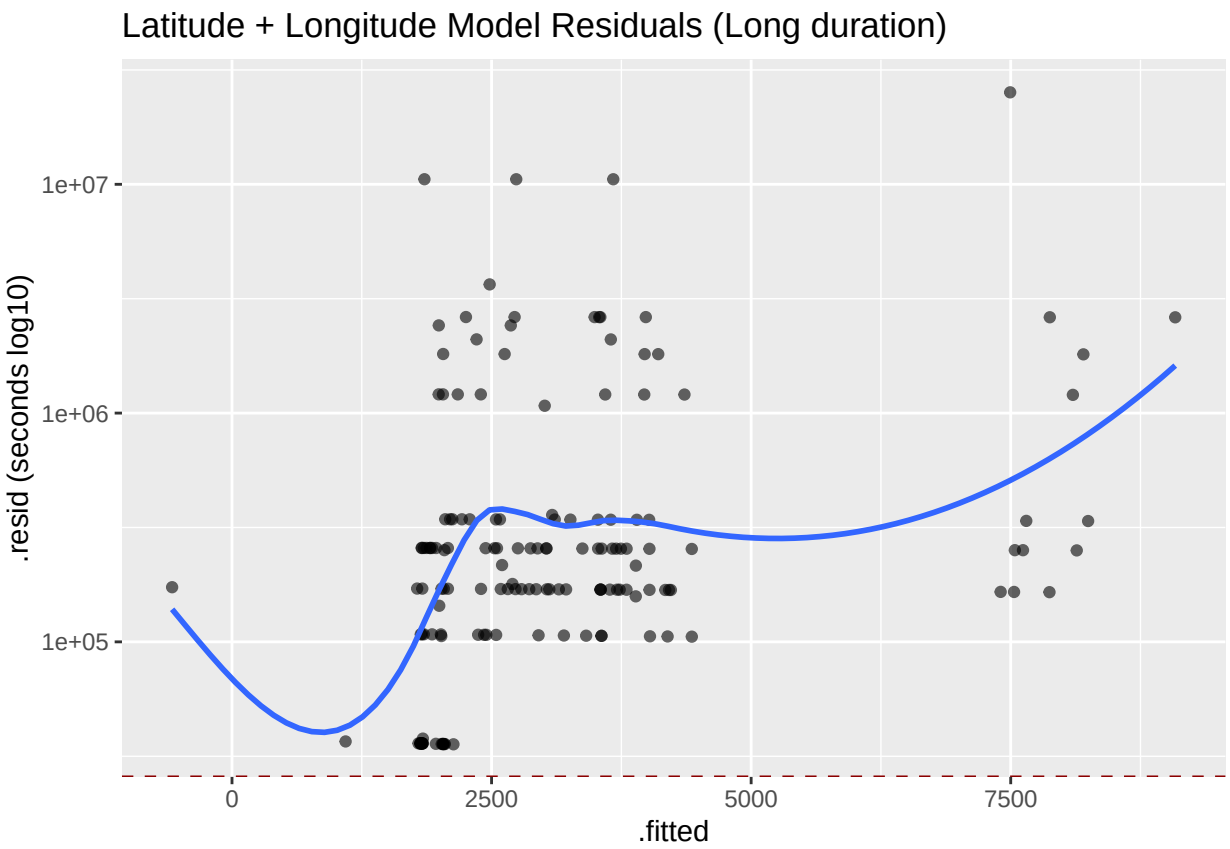

```
show_resid_plots_split(lat_long_lm, "Latitude + Longitude Model")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## Warning in scale_y_log10(): log-10 transformation introduced infinite values.
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Observations

- The model behaves slightly strangely, likely because sightings that are concentrated around certain longitudes and latitudes are being interpreted linearly
- There are much more positive residuals than negative residuals, however the line of best fit through the points is close to or less than 0. This indicates that most of the sightings were very short, and a much smaller portion was significantly longer

---

Using the Model