Bird-Aircraft Collision Analysis Report

Natalie Huante, Katherine Hansen, Sarah Fieck, Kelsey Hawkins

2024-12-13

Introduction

As air traffic becomes increasingly popular, it is important to observe and discuss the dangers that planes might encounter in the skies. An understanding of such risks can be beneficial to preparing and warning pilots, preventing and predicting high risk flights, etc. Of the many dangers that exist, one that we will focus on is bird strikes. A bird strike, also known as a bird-aircraft collision, simply refers to a collision between a bird of some type and a plane. They can often lead to damage to the aircraft and, in some cases, may even endanger the lives of passengers and crew on-board. The danger of bird strikes is often overlooked and generally not thought of to be a significant risk when flying, but it is in fact a plausible happening. This paper aims to discuss and analyze bird strikes to build a deeper understanding of how often it occurs, what factors might be influential, etc. and therefore be useful for the warning, prevention, and prediction of such events.

Any analysis mentioned in this paper is computed from the same data set containing logged information about when bird strikes occurred. The data was collected by aviation authorities around the world, although the majority of this particular data comes from flights originating in the United States. There are a total of 25,429 rows (or reported incidents) and 26 columns in this data set. The columns provide information about each entry (or bird strike incident recorded) and include but are not limited to some of the following: location, time, type of bird, extent of damage, weather conditions.

We focus on four questions throughout this paper and they are as follows:

1. Is there evidence that warnings of birds in the flight path have an effect on the number of birds hit? Does it even matter if pilots are warned?
2. Does the month of the year affect the count of bird strike incidents and, if so, which months are riskier? Are bird strikes increasing over the year?
3. What factors impact the number of birds struck per incident the most, and does the number of birds struck per incident change based on the region of the airport?
4. FIXME FIXME FIXME

By answering these questions, we can gain insights into the frequency, causes, and severity of bird strike incidents, providing us information to develop effective prevention measures in the future.

Notes: The data set is linked along with the GitHub repository in the useful links section for readers who would like to reference either the original data or the code for models used throughout this paper's analysis.

Q1 - Pilot Warnings & Strike Severity

Is there evidence that warnings of birds in the flight path have an effect on the number of birds hit? Does it even matter if pilots are warned?

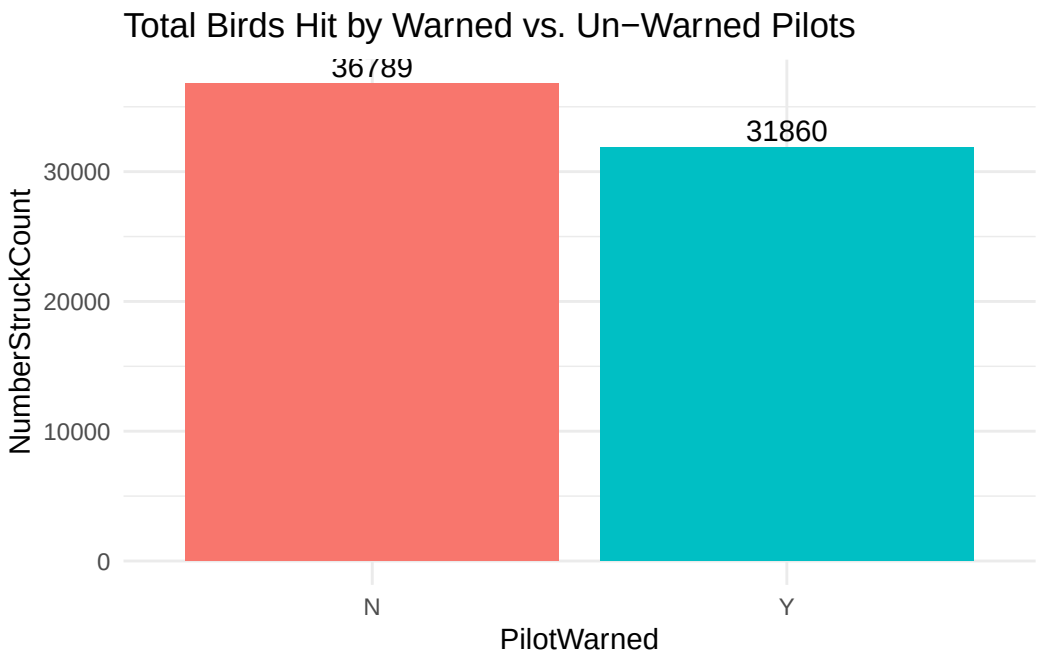
Exploratory Data Analysis

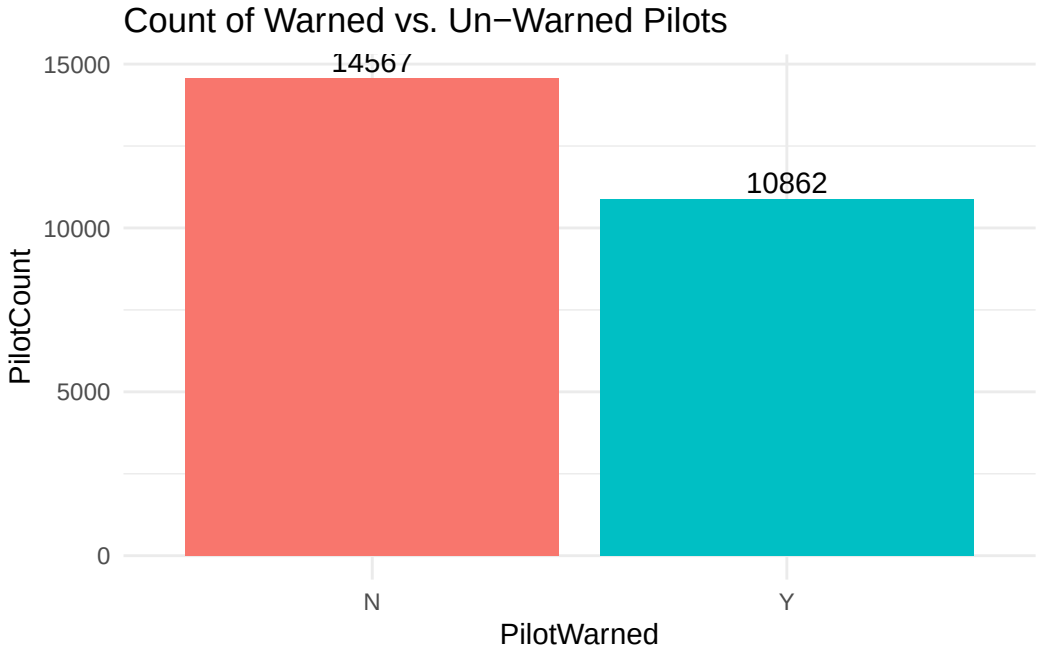
In order to understand the relationship between the two columns, we need to explore the shape of the data and prepare it for modeling. When looking at the summary, we can see that on average, 2.7 birds were hit. Based on the summary statistics and visuals below, collisions typically do not hit an extreme number of birds, primarily keeping the count under 100. Outliers may skew our data a little bit, as the maximum amount of birds hit was 942. When looking at the dataset, there were 14,567 collisions where pilots were not warned, and 10,862 collisions when pilots were warned. In total, 36,789 birds were hit by pilots who were not warned, and warned pilots hit 31,860 birds. There is not a clear trend of collision counts changing throughout specific years.

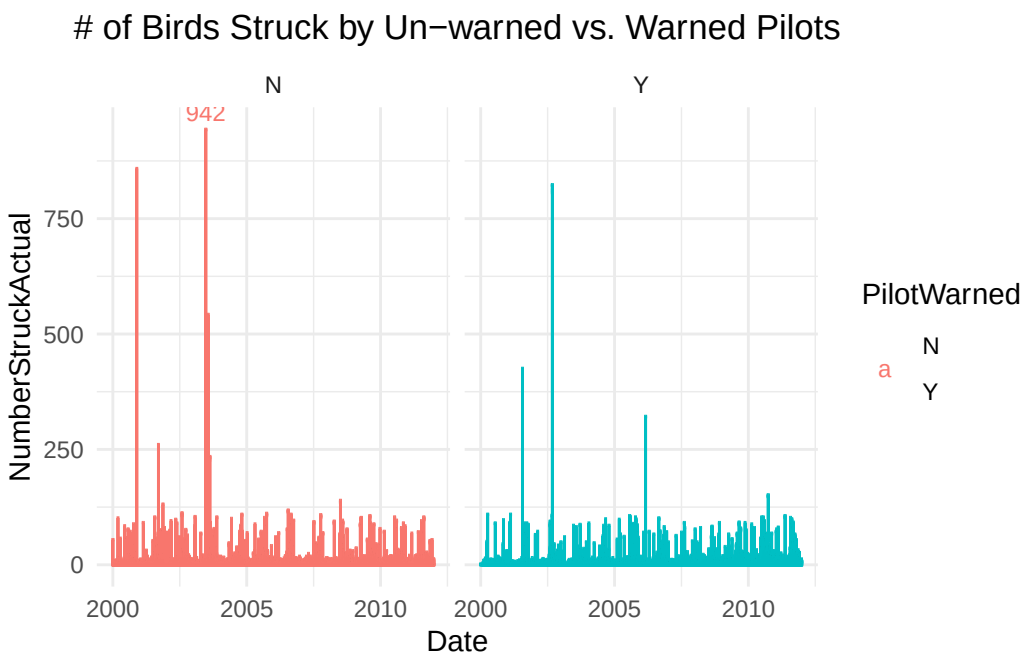
```
# Formatting the date column as a date instead of string. For EDA purposes
data$Date <- as.Date(data$FlightDate, format = "%m/%d/%y %H:%M")

# Mean, Median, Quartiles, Maximum, Minimum
summary(data$NumberStruckActual)

# Count of Instances - how many incidences were the pilots warned vs not
 ↪  warned about the birds in the sky?
table(data$PilotWarned)

data$WasWarned <- 0
data$WasWarned[data$PilotWarned == "Y"] <- 1
```

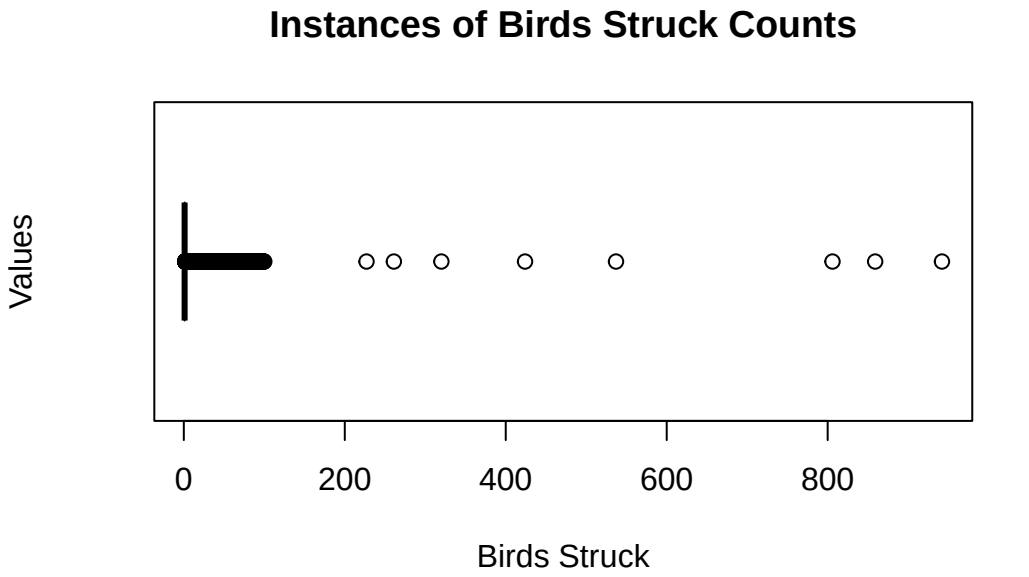
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0     1.0     1.0     2.7     1.0   942.0
```

```
    N     Y
14567 10862
```









Regression Modeling

Now that we have a deeper understanding of the data, we can model the relationship to understand if there are any effects of warnings on the number of birds struck. We used both Bayesian and Frequentist modeling methods to obtain thorough and consistent results. While they both serve the same purposes, in their foundations they approach hypothesis testing with different expectations. Bayesian models incorporate prior information to give a probabilistic

framework that updates beliefs about parameters as we gain new information. Frequentist models rely on observed data via random sampling. Combining these two approaches gives us a more cohesive idea of the relationship between the two variables.

Bayesian Process

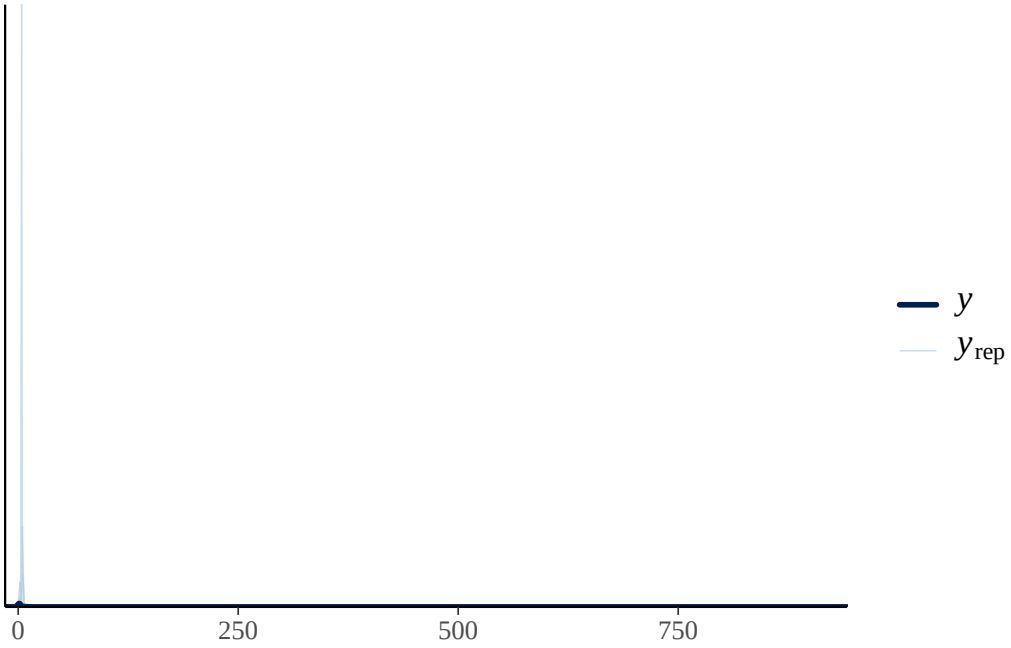
When we begin Bayesian Hypothesis Testing, we must start by using prior distributions that make sense. Bayesian models can experience two kinds of sources of predictive distribution variability. One of which is distributional variability. This kind of variability questions the certainty of our parameters. In order to test our parameters, we included predictive checks of the prior distribution to ensure our samples from the prior match the reality. We wanted to use a somewhat informed prior, containing a normal distribution for the coefficient and the log-normal for the intercept, to prevent negative predictions, as it is impossible to hit a negative number of birds. Based on the prior predictive check, the prior was not as informative as hoped. There are a few samples that fall into the negative range. Mostly, the samples feature values consistent with the prior data.

The posterior predictive check gives us more hope for the model, as the values sampled from the posterior distribution seem much more in line with the true values of the posterior. The same prior is used. No negative values are sampled, which makes sense in our scenario. The trace plots below the predictive checks also indicate convergence across sampling chains, a desirable trait for modeling reliability.

```
# Lower bound idea:
 ↪  https://discourse.mc-stan.org/t/checking-understanding-of-bounds-on-priors/8940
 ↪  since we at least need a collision value of 1. Stop predicting negatives!
# Log normal idea to introduce non-negativity, took out variance

# Prior distributions
priors <- c(prior("normal(0, 2)", class = "b", lb = 0),
            prior("lognormal(log(2.7), 0.5)", class = "Intercept"))

mod_1 <- brm(NumberStruckActual ~ WasWarned, data = data, prior = priors,
 ↪  sample_prior = "only")
```

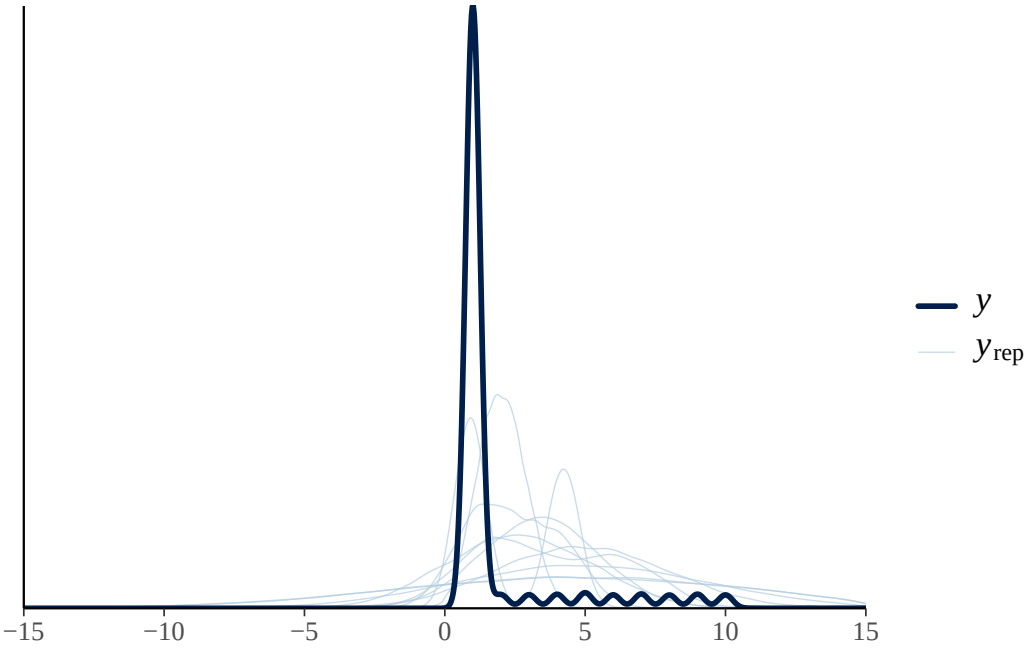
```
get_prior(mod_1)

# Prior Predictive Check
mod_1 |> pp_check()
```

```
Using 10 posterior draws for ppc type 'dens_overlay' by default.
```




```
# PP Check zoomed in
mod_1 |> pp_check() + xlim(-15, 15)
```

```
Using 10 posterior draws for ppc type 'dens_overlay' by default.
```

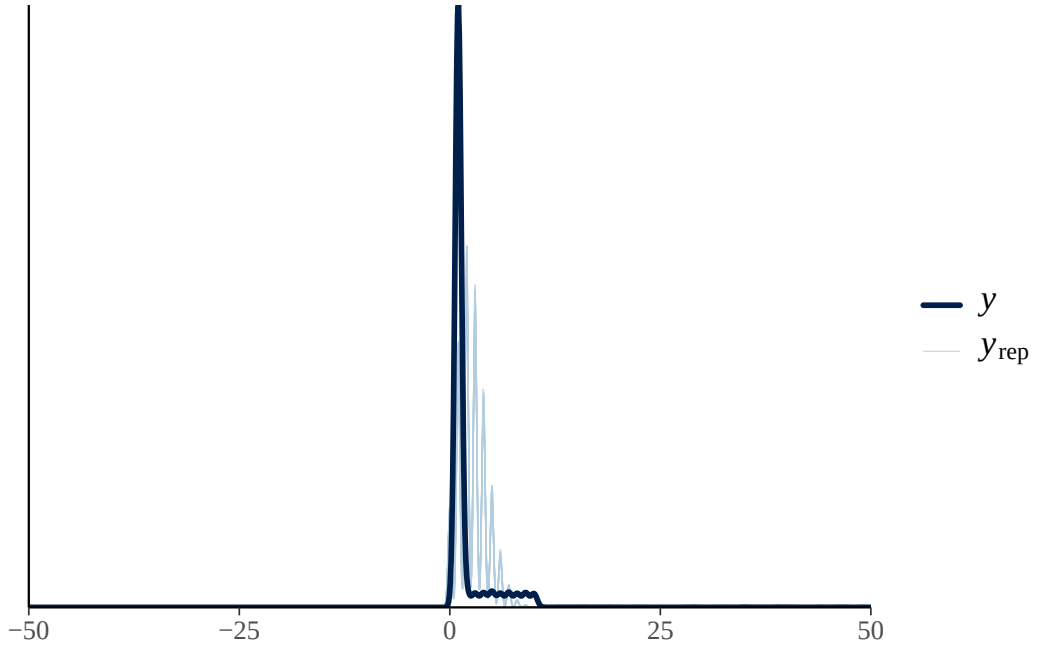


```
                  prior     class      coef group resp dpar nlpar lb ub
            normal(0, 2)         b                                   0
            normal(0, 2)         b WasWarned                         0
 lognormal(log(2.7), 0.5) Intercept
     student_t(3, 0, 2.5)     sigma                                   0
       source
         user
 (vectorized)
         user
      default
```

```
# Posterior Distribution Check
# poisson likelihood function for non-negativity count (count of birds hit)
mod_1 <- brm(NumberStruckActual ~ WasWarned, data = data, prior = priors,
 ↪  family = poisson())
```

```
mod_1 |> pp_check() + xlim(-50, 50)
```

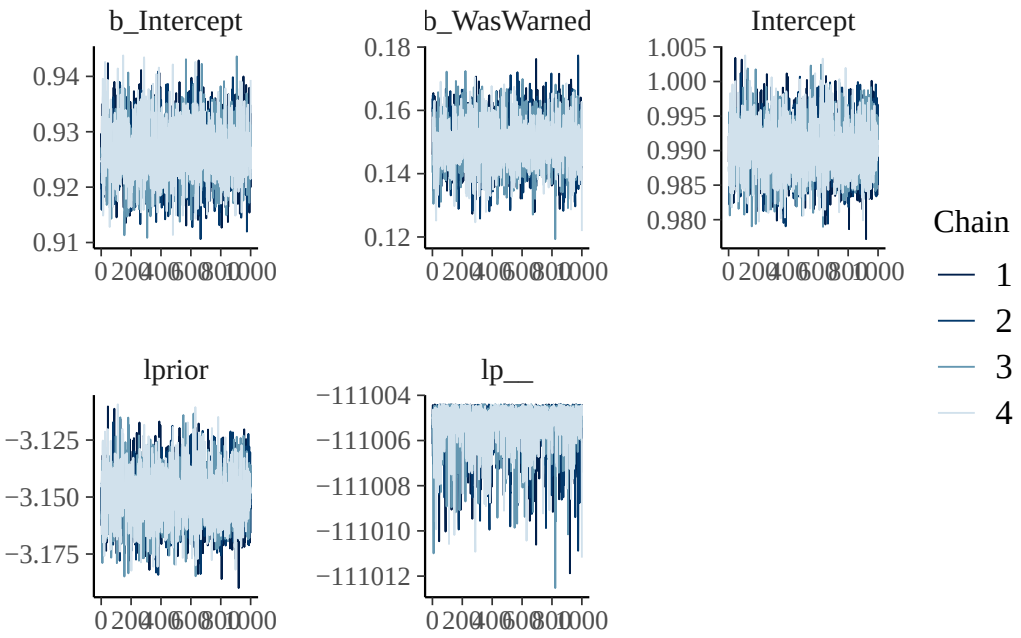
```
Using 10 posterior draws for ppc type 'dens_overlay' by default.
```



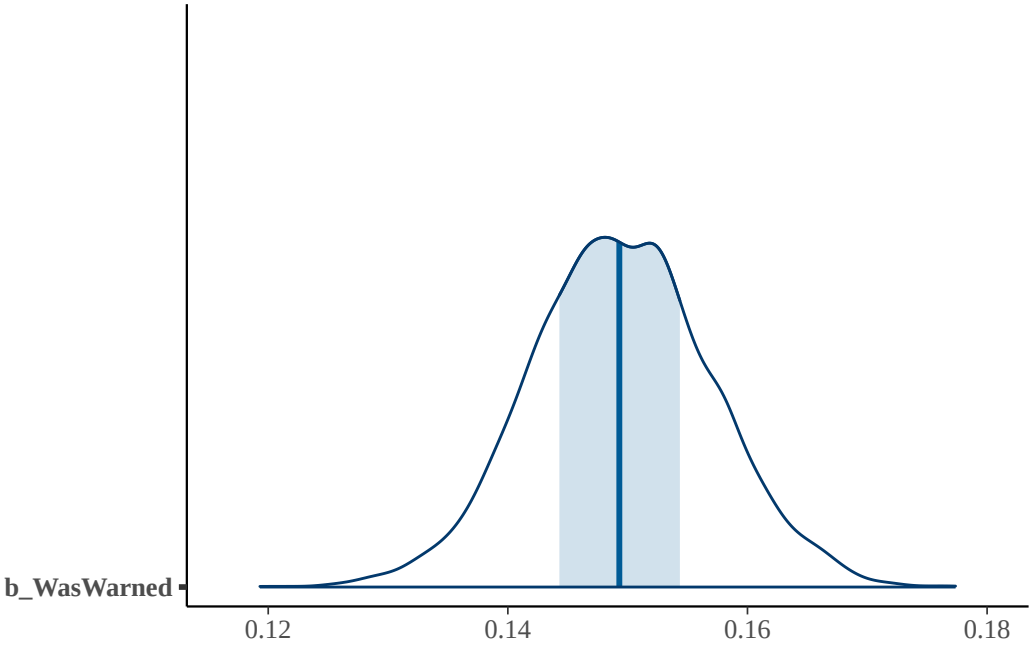
```
# Summary of Values
summary(mod_1)
```

```
 Family: poisson
  Links: mu = log
Formula: NumberStruckActual ~ WasWarned
   Data: data (Number of observations: 25429)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Regression Coefficients:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     0.93      0.01     0.92     0.94 1.00     2798     2577
WasWarned     0.15      0.01     0.13     0.16 1.00     2393     1937

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
# Measure convergence & metrics
mcmc_trace(mod_1)
```



```r
mod_1$fit

# Posterior Examination - parameter estimation for the coefficient
# plot a parameter for coefficient to see densitiy of posteriors for each
 ↪  values

mcmc_areas(mod_1, pars = c("b_WasWarned"))
```




```
Inference for Stan model: anon_model.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

                 mean se_mean   sd       2.5%        25%        50%        75%
b_Intercept      0.93    0.00 0.01       0.92       0.92       0.93       0.93
b_WasWarned      0.15    0.00 0.01       0.13       0.14       0.15       0.15
Intercept        0.99    0.00 0.00       0.98       0.99       0.99       0.99
lprior          -3.15    0.00 0.01      -3.17      -3.16      -3.15      -3.14
lp__        -111005.38    0.02 0.99 -111008.06 -111005.75 -111005.08 -111004.68
                97.5% n_eff Rhat
b_Intercept      0.94  2784    1
b_WasWarned      0.16  2375    1
Intercept        1.00  3971    1
lprior          -3.13  3935    1
```

```
lp__        -111004.42  1982    1

Samples were drawn using NUTS(diag_e) at Fri Dec 13 12:50:39 2024.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Frequentist Process

We can also use Frequentist Hypothesis testing methods to possibly confirm our findings on the relationship between warnings and birds hit. A generalized linear model was used, with a Poisson distribution to account for the fact that we are working with count data (# of birds hit). We also wanted to see if this effect is equivalent to the null, meaning that while there is an effect, it is basically equivalent to zero. Lastly, we wanted to confirm our findings using a similar hypothesis test method, chi-square, to measure independence between birds hit and warnings to pilots.

```r
# poisson for count
mod_2 <- glm(NumberStruckActual ~ WasWarned, data = data, family = poisson())

summary(mod_2)
confint(mod_2)
```

```
Waiting for profiling to be done...
```

```
Call:
glm(formula = NumberStruckActual ~ WasWarned, family = poisson(),
    data = data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.926440   0.005214  177.70   <2e-16 ***
WasWarned   0.149641   0.007653   19.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 163605  on 25428  degrees of freedom
Residual deviance: 163224  on 25427  degrees of freedom
```

```
AIC: 222003

Number of Fisher Scoring iterations: 6

                2.5 %    97.5 %
(Intercept) 0.9162042 0.9366413
WasWarned   0.1346382 0.1646378
```

```r
TOST_result <- tsum_TOST(
  m1 = mean_diff,
  mu = 0,
  sd1 = std_error,
  n1 = nrow(data),
  low_eqbound = -0.1*std_error,
  high_eqbound = 0.1*std_error
)

TOST_result$decision
```

```
$TOST
[1] "The equivalence test was non-significant, t(25428) = 8478.352, p = 1e+00"

$ttest
[1] "The null hypothesis test was significant, t(25428) = 8494.299, p = 0e+00"

$combined
[1] "NHST: reject null significance hypothesis that the effect is equal to zero \nTOST: don't
```

```r
# Clean data to where theres a dummy var for if the pilot is warned, if the
 ↪  pilot is not warned
# Measure the independence of warnings and birds being hit
hit <- data %>%
  group_by(WasWarned) %>%
  summarise(BirdsHit = sum(NumberStruckActual))

chisq_test <- chisq.test(hit$BirdsHit)
print(chisq_test)
```

```
    Chi-squared test for given probabilities
```

```
data:  hit$BirdsHit
X-squared = 353.9, df = 1, p-value < 2.2e-16
```

Results

In order to determine if warning has an effect or not, we look at the estimate calculated and presented in the Bayesian model summary. Strangely enough, we see that a warning to a pilot increases the amount of birds hit by ~15%. When looking at the 95% credible interval to account for uncertainty, this increase can range between 14-16%. This uncertainty interval does not include zero or negative numbers, meaning that this model provides compelling evidence that there is a positive effect on birds hit when a pilot is warned of birds in the flight path. The posterior examination visual shows the range of effect, spanning across the credible interval.

Similarly, the Frequentist model shows us that warnings have a positive effect on birds hit, with about a 15% increase. When looking at the confidence interval, we see that this value ranges between about 13% and 16%. The very low p-value indicates significance in this relationship, and gives more evidence against the null hypothesis. This result is very much in line with what we got from the Bayesian model. While this is a positive effect, it is not the largest. A two-sided T-test to measure equivalence to the null hypothesis was used. We set the lower and upper bound using the standard deviations. Based on the results, we found that we have evidence to reject the null hypothesis that there is no effect on birds hit and warnings. However, this effect does not entirely mean it's the most impactful relationship. Based on the test's results, the effect that warning has is not the strongest, and could be equivalent to no effect.

Lastly, we wanted to use a separate Frequentist hypothesis testing method to see how far away our sample statistics are from the null hypothesis' statistics. A Chi-squared test was used to specifically measure the independence between birds hit by pilots who were warned versus those who were not. This test further assesses whether there is a statistically significant relationship between warnings and bird count. After running it, we received a high chi-squared score of 353.9, indicating the observed data is very different from if there was no effect between the number of birds hit and pilots warned. The chi-squared result also shares the same p-value from the model results, further conveying significance. Due to the chi-squared scores difference, we have further evidence that there is an effect of warnings on the number of birds hit.

Discussion

Based on the findings from the Bayesian and Frequentist Hypothesis Testing methods, we have found that there is a positive effect on the number of birds hit if the pilot is warned about them. Based on the TOST results, there is evidence that while there is an effect greater than zero, it might be equivalent to zero, meaning that warnings may not increase the number that

much. This seems counter-intuitive, as the point of a warning is to prevent more birds being hit, keeping passengers safe, and preserving the engine of the aircraft. It could be that when pilots know about birds in the path, they overcompensate with safety maneuvers and end up still hitting birds. Or maybe the warnings occur when the birds are too close to the aircraft and it is far too late to do anything. However, It could be worthwhile to explore covariates and introduce other factors from the data that may play a bigger role in effecting the number of birds hit. Understanding this relationship can help pilots prepare for birds and other wildlife in their path.

Q2 - Seasonality & Trends Over Time

Does the month of the year affect the count of bird strikes and, if so, which months are riskier? Are bird strikes increasing over the years?

Aggregate Data

The first step is to summarize our bird strike data by month and year. Since we are concerned with the number of bird strikes that has been recorded within each month, we should calculate the total strike counts by month and year.

The original data set has a variable, "FlightDate", that contains the date in which the flight took place. We will take this value and extract the month and year. We will then aggregate the data on those values and store the monthly counts in a new table.

```
# library import
library(dplyr)

# let's make sure the variable is in date format
data$FlightDate <- as.Date(data$FlightDate, format = "%m/%d/%y %H:%M")

# extract the month and the date from this variable and store in new columns
data$Year <- format(data$FlightDate, "%y")
data$Month <- format(data$FlightDate, "%m")

# tables of aggregated counts
strikes_by_month <- data %>%
  group_by(Year, Month) %>%
  summarise(Count = n())
```

```
# verify the new table was created correctly
print(strikes_by_month)
```

```
# A tibble: 144 x 3
# Groups:   Year [12]
   Year  Month Count
   <chr> <chr> <int>
 1 00    01       45
 2 00    02       57
 3 00    03       80
 4 00    04      103
 5 00    05      118
 6 00    06      117
 7 00    07      158
 8 00    08      189
 9 00    09      185
10 00    10      176
# i 134 more rows
```

Regression Model

Now that we have aggregated our data, we want to fit our regression model. The outcome will be the count of bird strikes by month and year and the predictors will be the month and year in which the flight took place. This will allow us to calculate the effects of the month and the year on the total count of strikes that month.

We will use a poisson regression model because we are working with count data, so a poisson distribution is most appropriate. For the link function, we include a log function.

As a quick note, the count of bird strikes is treated as a continuous variable, while the month and year is a categorical variable. So, the coefficients of our model will indicate the following:

- The coefficients for *Month* will represent the relative risk of bird strikes compared to the baseline, which will be the month of January, or Month01. This will help us look at seasonality.
- The coefficients for *Year* will represent the expected increase or decrease in bird strikes for each year compared to the baseline, which will be the year 2000. This happens to be the earliest year in our dataset. This will help us look at trends over time and to identify if birds strikes are currently on the decline or the incline.

So, to review, our baseline will be January 2000 and each coefficient will represent a relative increase or decrease to this month.

```
# make sure that month is categorical and not continuous
strikes_by_month$Month <- as.factor(strikes_by_month$Month)
```

```r
# regression model
poisson_model <- glm(Count ~ Month + Year,
                     data = strikes_by_month,
                     family = poisson(link = "log"))

# output the coefficients
summary(poisson_model)
```

```
Call:
glm(formula = Count ~ Month + Year, family = poisson(link = "log"),
    data = strikes_by_month)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.91941    0.04195  93.440  < 2e-16 ***
Month02     -0.19370    0.04861  -3.985 6.75e-05 ***
Month03      0.27452    0.04334   6.334 2.38e-10 ***
Month04      0.66829    0.04018  16.633  < 2e-16 ***
Month05      0.90578    0.03871  23.398  < 2e-16 ***
Month06      0.79648    0.03935  20.241  < 2e-16 ***
Month07      1.25231    0.03704  33.805  < 2e-16 ***
Month08      1.37610    0.03656  37.638  < 2e-16 ***
Month09      1.29383    0.03688  35.085  < 2e-16 ***
Month10      1.17627    0.03737  31.478  < 2e-16 ***
Month11      0.65341    0.04028  16.222  < 2e-16 ***
Month12      0.08389    0.04526   1.854  0.06380 .
Year01      -0.10560    0.03930  -2.687  0.00721 **
Year02       0.20677    0.03642   5.677 1.37e-08 ***
Year03       0.13718    0.03700   3.707  0.00021 ***
Year04       0.21329    0.03637   5.865 4.49e-09 ***
Year05       0.30419    0.03565   8.532  < 2e-16 ***
Year06       0.45703    0.03456  13.222  < 2e-16 ***
Year07       0.52073    0.03415  15.249  < 2e-16 ***
Year08       0.50186    0.03427  14.645  < 2e-16 ***
Year09       0.86511    0.03224  26.832  < 2e-16 ***
Year10       0.82553    0.03243  25.453  < 2e-16 ***
Year11       0.76986    0.03272  23.532  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 8364.97  on 143  degrees of freedom
Residual deviance:  230.09  on 121  degrees of freedom
AIC: 1260.5

Number of Fisher Scoring iterations: 4
```

```r
# look at the confidence intervals for uncertainty
confint(poisson_model)
```

```
Waiting for profiling to be done...

                   2.5 %      97.5 %
(Intercept)  3.836585281  4.00101859
Month02     -0.289129788 -0.09856028
Month03      0.189730827  0.35964226
Month04      0.589869854  0.74738634
Month05      0.830300142  0.98206893
Month06      0.719717968  0.87398949
Month07      1.180179464  1.32540913
Month08      1.304946734  1.44828421
Month09      1.222046027  1.36661747
Month10      1.103495011  1.24999452
Month11      0.574784341  0.73270052
Month12     -0.004776411  0.17267408
Year01      -0.182693975 -0.02861931
Year02       0.135468498  0.27824736
Year03       0.064708318  0.20977669
Year04       0.142097558  0.28466808
Year05       0.234423055  0.37419782
Year06       0.389446860  0.52495053
Year07       0.453979378  0.58785199
Year08       0.434872580  0.56921949
Year09       0.802187095  0.92858205
Year10       0.762224218  0.88937292
Year11       0.705989935  0.83424256
```

Given the above output, we can interpret not only the coefficients, but also the confidence intervals of our results in order to assess the certainty of those results.

Month Coefficients: As an example, let's first take Month02's (February's) coefficient of -0.1937, which represents the log-relative rate of how many bird strikes are expected in February compared to January. If we convert this log-relative rate to a percentage relative rate, we can simply raise e to the power of the coefficient and multiply by 100. By doing this calculation, it may be easier to interpret that the the relative rate of bird strikes expected in February is $e^{-0.1937}$ or 0.824 compared to January (our baseline). In other words, February's expected rate is 82.4% the rate in January. Applying this calculation to each month and it's coefficient, we calculate the following:

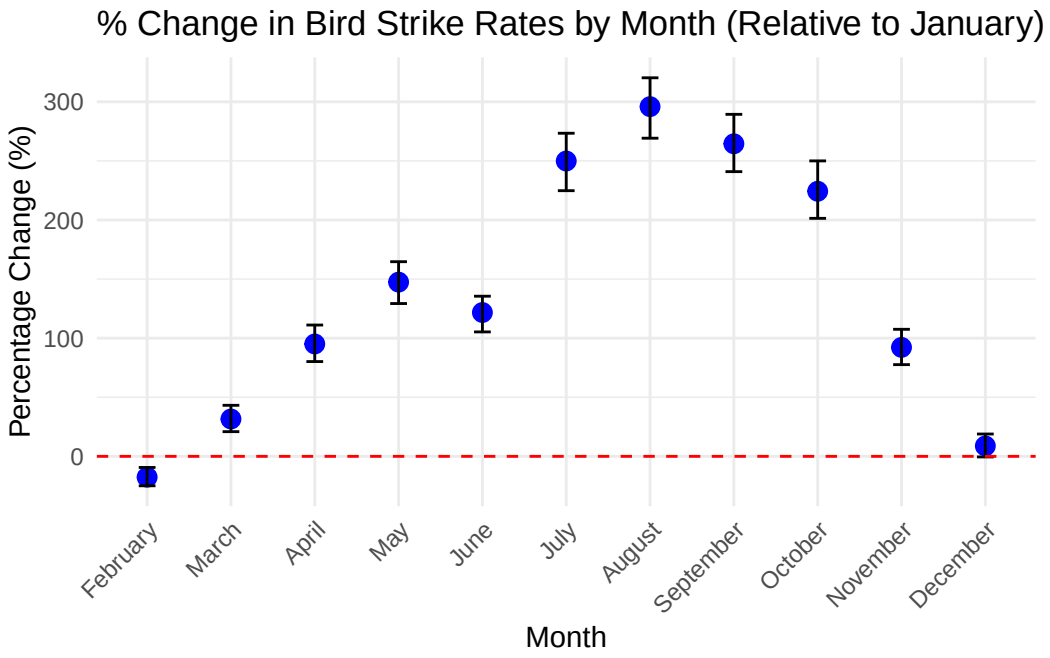
Month	Percentage Relative Rate	Percentage Difference
01 - January (Intercept/Baseline)	1 (Baseline)	0% (Baseline)
02 - February	$(e^{-0.1937}) * 100 = 82.4\%$	-17.6%
03 - March	$(e^{0.27452}) * 100 = 131.6\%$	+31.6%
04 - April	$(e^{0.6683}) * 100 = 195.1\%$	+95.1%
05 - May	$(e^{0.9058}) * 100 = 247.4\%$	+147.4%
06 - June	$(e^{0.7965}) * 100 = 221.7\%$	+121.7%
07 - July	$(e^{1.2523}) * 100 = 349.9\%$	+249.9%
08 - August	$(e^{1.3761}) * 100 = 395.9\%$	+295.9%
09 - September	$(e^{1.2938}) * 100 = 364.5\%$	+264.5%
10 - October	$(e^{1.1763}) * 100 = 324.3\%$	+224.3%
11 - November	$(e^{0.6534}) * 100 = 192.2\%$	+92.2%
12 - December	$(e^{0.0839}) * 100 = 108.8\%$	+8.8%

Now, while the coefficients suggest one conclusion, we also calculate the 95% confidence intervals for each month's effect. This is important to consider because if any month's confidence interval contains the value of 0, then that means that the possibility of the month having no effect in comparison to our baseline is within our range of likely effect values. Therefore, we would be more uncertain about that particular month's observed effects being significant and enough to conclude there is an effect.

Below is a graph that will visualize the relative percentages listed compared to the baseline of January, along with the confidence intervals for each month.

```
# library imports
library(ggplot2)

# month names
month_names <- c("February", "March", "April", "May", "June", "July",
 ↪  "August", "September", "October", "November", "December")
# month coefficients converted to relative percentage difference
month_relative_percentage_diff <- c(-17.6, 31.6, 95.1, 147.4, 121.7, 249.9,
 ↪  295.9, 264.5, 224.3, 92.2, 8.8)
```

```r
# month lower confidence interval converted to relative percentage difference
lower_ci_percentage_diff <- c(-24.9, 20.9, 80.2, 129.3, 105.3, 224.8, 269.2,
 ↪  240.9, 201.4, 77.6, -0.5)
# month upper confidence interval converted to relative percentage difference
upper_ci_percentage_diff <- c(-9.4, 43.2, 111.1, 164.7, 135.5, 273.4, 320.3,
 ↪  289.4, 250.0, 107.5, 18.9)

# create data frame for the plot
percentage_diff_data <- data.frame(
                          Month = month_names,
                          PercentDiff = month_relative_percentage_diff,
                          LowerCI = lower_ci_percentage_diff,
                          UpperCI = upper_ci_percentage_diff)

# make sure the order of months remains intact (feb --> dec)
percentage_diff_data$Month <- factor(percentage_diff_data$Month,
                                     levels = c("February", "March", "April",
                                      ↪  "May", "June", "July", "August",
                                      ↪  "September", "October", "November",
                                      ↪  "December"))

# plot
ggplot(percentage_diff_data, aes(x = Month, y = PercentDiff)) +
    geom_point(size = 3, color = "blue") +
    geom_errorbar(aes(ymin = LowerCI, ymax = UpperCI), width = 0.2, color =
     ↪  "black") +
    geom_hline(yintercept = 0, linetype = "dashed", color = "red") + # no
     ↪  relative difference
    labs(
      title = "% Change in Bird Strike Rates by Month (Relative to January)",
      x = "Month",
      y = "Percentage Change (%)"
    ) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



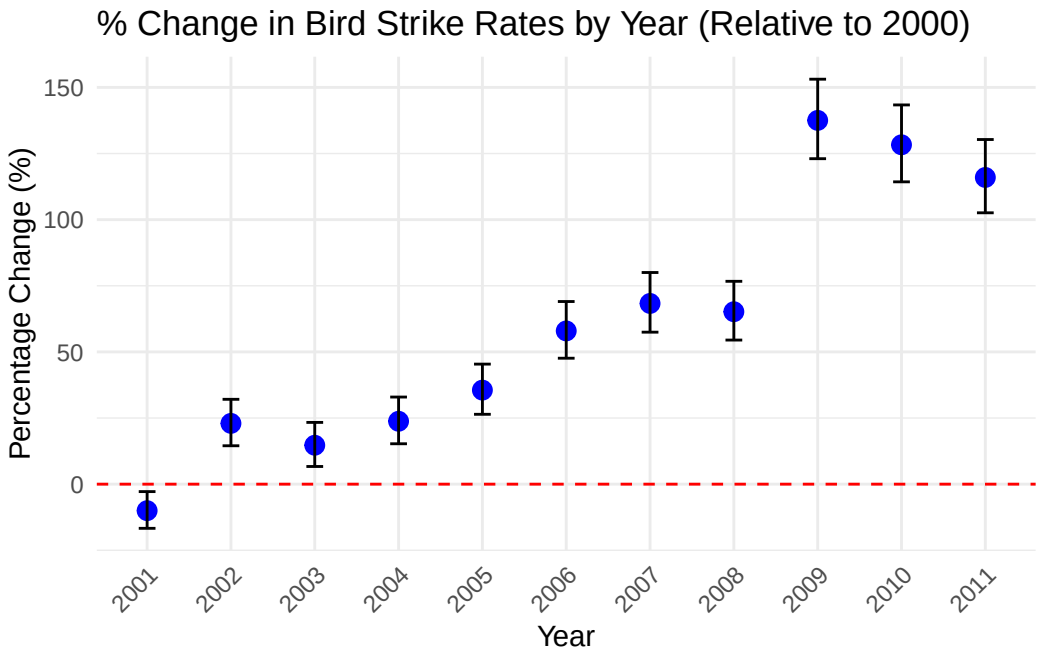
Year Coefficients: We can take a similar approach with the coefficients for each year in the model's output. Like the month coefficients, the year coefficients also represent the log-relative rate of bird strikes for each year compared to the baseline year, which is the year 2000. So, each coefficient represents the relative increase or decrease in expected bird strikes for that year. We can look at the group of coefficients chronologically to then make conclusions and/or observations about whether there has been a trend of increasing/decreasing rates over time or if there are any years that showed unexpected rates. To make these coefficients easier to interpret, we will apply the same conversion from log-relative rates to percentage relative rates, which will give us the difference in percentages, compared to 2000. Below, we include these calculations.

Year	Percentage Relative Rate	Percentage Difference
2000 - Intercept/Baseline	1 (Baseline)	0% (Baseline)
2001	$(e^{-0.1056}) * 100 = 90\%$	-10.0%
2002	$(e^{0.2068}) * 100 = 123\%$	+23.0%
2003	$(e^{0.1372}) * 100 = 114.7\%$	+14.7%
2004	$(e^{0.2133}) * 100 = 123.8\%$	+23.8%
2005	$(e^{0.3042}) * 100 = 135.5\%$	+35.5%
2006	$(e^{0.4570}) * 100 = 158.0\%$	+58.0%
2007	$(e^{0.5207}) * 100 = 168.3\%$	+68.3%
2008	$(e^{0.5019}) * 100 = 165.2\%$	+65.2%
2009	$(e^{0.8651}) * 100 = 237.5\%$	+137.5%
2010	$(e^{0.8255}) * 100 = 228.3\%$	+128.3%
2011	$(e^{0.7699}) * 100 = 215.9\%$	+115.9%

Below is a graph that will visualize the relative percentages listed compared to the baseline of 2000, along with the confidence intervals for each year.

```
# library imports
library(ggplot2)

# year names
year_names <- c("2001", "2002", "2003", "2004", "2005", "2006", "2007",
 ↪  "2008", "2009", "2010", "2011")

# month coefficients converted to relative percentage difference
year_coeffs = c(-0.1056, 0.2068, 0.1372, 0.2133, 0.3042, 0.4570, 0.5207,
 ↪  0.5019, 0.8651, 0.8255, 0.7699)
# month lower confidence interval converted to relative percentage difference
lower_ci_coeffs <- c(-0.1827, 0.1355, 0.0647, 0.1421, 0.2344, 0.3894, 0.4540,
 ↪  0.4349, 0.8022, 0.7622, 0.7060)
# month upper confidence interval converted to relative percentage difference
upper_ci_coeffs <- c(-0.0286, 0.2782, 0.2098, 0.2847, 0.3742, 0.5250, 0.5879,
 ↪  0.5692, 0.9286, 0.8894, 0.8342)

# convert the coefficients
year_rate_ratios <- exp(year_coeffs)
percentage_changes <- (year_rate_ratios - 1) * 100
lower_percentage <- (exp(lower_ci_coeffs) - 1) * 100
upper_percentage <- (exp(upper_ci_coeffs) - 1) * 100

# create data frame for the plot
year_percentage_diff_data <- data.frame(
                          Year = year_names,
                          PercentDiff = percentage_changes,
                          LowerCI = lower_percentage,
                          UpperCI = upper_percentage)

# make sure the order of months remains intact (feb --> dec)
year_percentage_diff_data$Year <- factor(year_percentage_diff_data$Year,
                                    levels = c("2001", "2002", "2003",
                                     ↪  "2004", "2005", "2006", "2007",
                                     ↪  "2008", "2009", "2010", "2011"))

# plot
ggplot(year_percentage_diff_data, aes(x = Year, y = PercentDiff)) +
    geom_point(size = 3, color = "blue") +
```

```
  geom_errorbar(aes(ymin = LowerCI, ymax = UpperCI), width = 0.2, color =
   ↪  "black") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") + # no
   ↪  relative difference
  labs(
    title = "% Change in Bird Strike Rates by Year (Relative to 2000)",
    x = "Year",
    y = "Percentage Change (%)"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Results

Question 1: Does the month of the year affect the count of bird strikes and, if so, which months are riskier?

For each month (with the exclusion of December), we can see that their confidence interval does not contain the percentage difference of 0%, which would have indicated that no difference between that month and the baseline (January) would be plausible. However, since this is not the case, we can conclude that each month does have an effect on what the expected rate is. Each month (with the exclusion of February) has a varying level of increase in the relative rate, indicating there is some seasonality.

So, we can conclude that, from the Poisson regression results, that the month of the year does affect the relative rate of bird strikes we expect to see. Summer and early fall months (July, August, September, October) have the highest relative increases compared to January at +249.9%, +295.9%, +264.5% and +224.3% respectively, so they could be considered the "riskiest" months for bird strikes. Meanwhile, winter and early spring months (December, February, March) had the lowest relative rates at +8.8%, -17.6%, and +31.6%.

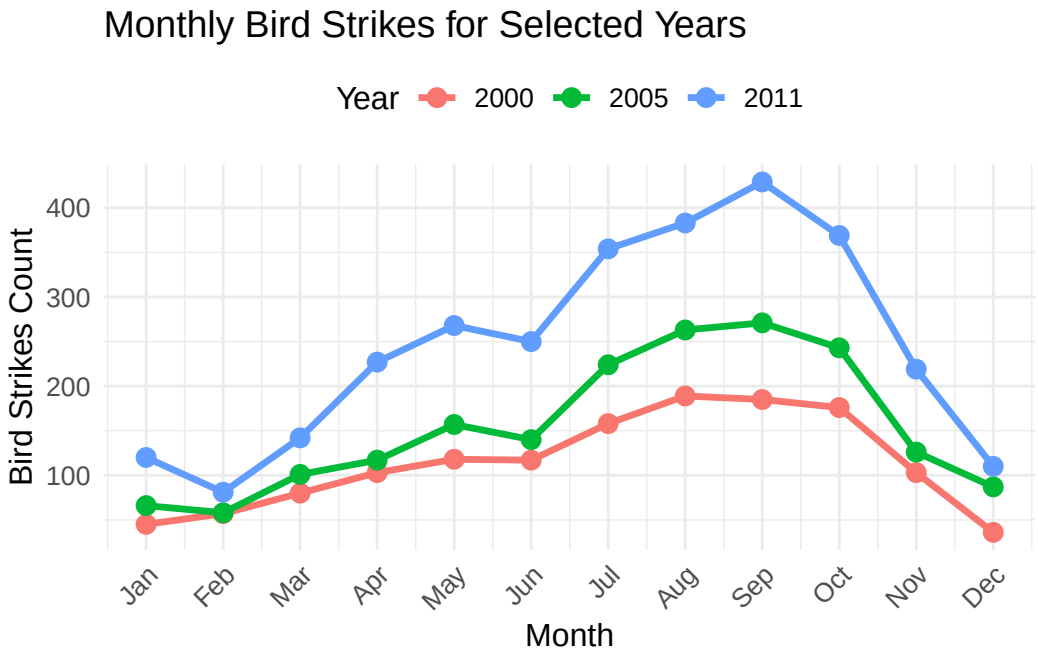
Question 2: Are bird strikes increasing over the years?

Relative to the year 2000, we can see that bird strikes have overall increased over time, with some noise and fluctuations in between. For example, from the year 2008 to 2009 we see a rather large jump from a percentage difference of +65.2% to +137.5%, but outside of this noise, we can observe a steady increase since 2000, with the rates more than doubling from the baseline (2000) to 2011.

Again, we also include the 95% confidence intervals for these results as well and can observe that, for each year, neither CI includes the percentage difference of 0%. This means that for each year, the difference in the percentage relative rates does not include the possibility of that difference to be 0, or there being no difference. So, we can more confidently conclude that there is a difference in those years, and combined with the coefficients, that that difference is steadily increasing over time.

```
library(ggplot2)

year_comparison_data <- data.frame(
      Year = rep(c(2000, 2005, 2011), each = 12),
      Month = rep(1:12, 3),
      Count = c(45, 57, 80, 103, 118, 117, 158, 189, 185, 176, 103, 36, #
       ↪  2000
                66, 58, 101, 117, 157, 140, 224, 263, 271, 243, 126, 87, #
                 ↪  2005
                120, 81, 142, 227, 268, 250, 354, 383, 429, 369, 219, 110)) #
                 ↪  2011)

ggplot(year_comparison_data,
      aes(x = Month, y = Count, group = Year, color = as.factor(Year))) +
      geom_line(size = 1.2) +
      geom_point(size = 3) +
      scale_x_continuous(breaks = 1:12, labels = c("Jan", "Feb", "Mar",
       ↪  "Apr",
                                                   "May", "Jun", "Jul", "Aug",
                                                   "Sep", "Oct", "Nov", "Dec"))
                                                    ↪  +
      labs(
```

```
    title = "Monthly Bird Strikes for Selected Years",
    x = "Month",
    y = "Bird Strikes Count",
    color = "Year") +
  theme_minimal() +
  theme(
    text = element_text(size = 12),
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "top")
```



To further demonstrate our conclusions, we have included a graph above that provides a visual representation of the count data from the original dataset, but only includes the year 2000, 2005, and 2011. This graph aligns with the results of our model and showcases both the seasonality and overall trend of bird strikes through the years. We can see that each of the three years have a similar seasonal pattern where summer and early fall months see an increase in total bird strikes and we see this in the graph where all three lines rise from months June to about October. Likewise, we can see a fall for all three lines in the winter and early spring months, representing a decrease in bird strikes, which also aligns with the conclusions drawn from the model.

Discussion

Having seen the results and having drawn the conclusions from them, there are some factors worth discussing in terms of why the seasonality and upwards trend are present and what could be causing them.

For the monthly effects and the observed seasonality of increased strikes during summer/fall months and decreased strikes in the winter/spring, this could be due to seasonal migration patterns where birds tend to be more active in the skies during their typical migration season. This also could be due to the number of flights occurring in each season. Perhaps more people travel during the summer, which could lead to more flights and more bird strikes.

For the yearly effects and the observed upwards trend of bird strikes per year, this could be to several environmental and cultural shifts. Increased air traffic throughout the years from 2000 to 2011 has surely increased, so there would be many more flights in more recent years compared to a decade prior. Therefore, with more flights, there is more of an opportunity for a bird strike to occur. Another factor could be the method and frequency in which bird strikes are reported accurately. So, there could be more accurate or accessible ways to track bird strikes with new policies or technology, leading to recent years appearing as if they have "more" bird strikes compared to previous years.

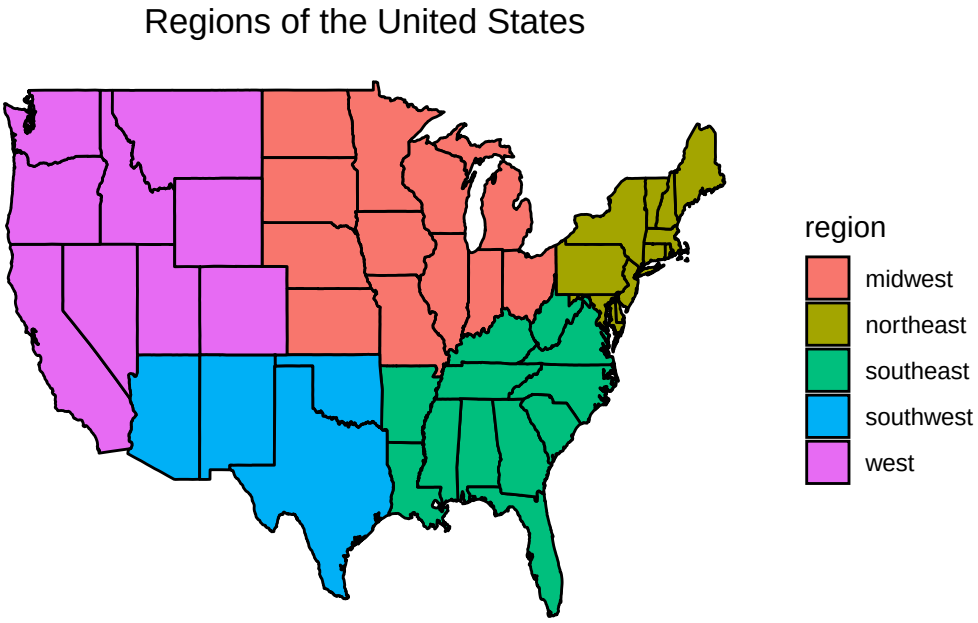
If we were to elaborate on and continue exploring the seasonality and trends of this data, some considerations and questions we would observe are stated below: - Whether it would make sense to control for weather, # of flights, etc. as aggregated co-variates in some way or if that is implied by the seasonality of months. In one way, this would complicate the model given that we would have to aggregate the co-variate data to be able to be applied to the aggregated counts. For some variables, this could result in losing the variability and therefore losing their meaningfulness at the monthly-level. For those that don't, it is worth exploring whether it really matters to include them or not. For example, does knowing the month of the year in which a flight takes place not come with certain assumptions of what type of weather is expected, if rain is plausible, etc. regardless of what region you are in. - Has the relative rates of months (the seasonality) changed over the years? - Do some months see more larger bird strikes than others? The original dataset has a variable that represents the number of birds involved in the bird strike recorded. So, one bird strike might represent 100 birds colliding with a plane where another might represent only 1 bird colliding with a plane. In our results, we treated all bird strikes equally regardless of how many birds were involved in the incident, so it would be interesting to see whether certain types of bird strikes are more probable in some months versus others. - The data that we use only contains information about flights where bird strikes did occur and were reported, but nothing about flights where bird strikes did not occur or they weren't reported. So, if we happened to have that data available, we would ask the following: What percentage of strikes are actually being reported and has that percentage changed throughout the year? What is the direct effect of the month and year in which a flight takes place on the probability of a bird strike occurring?

Q3 - Influences on Strike Severity & Regional Variations

What factors impact the number of birds struck per incident the most, and does the number of birds struck per incident change based on the region of the airport?

Analysis

The column State was feature engineered to represent different regions of the US. This was done using domain knowledge of the geographical and cultural regions of the US, as well as consulting sources on the topic. Five regions of the US were defined: Northeast, Midwest, Southeast, Southwest, and West. The regions were defined as follows:



To answer the first part of the question, a linear regression model was run with the features AircraftType, Altitude, Engines, FlightPhase, ConditionsPrecipitation, ConditionsSky, PilotWarned, and the newly feature engineered Region to predict NumberStruckActual. Then, the magnitude and sign of the coefficients was studied, along with the significance level, to determine which predictors have the largest impact.

Then, to answer the second part of the question, a one-way ANOVA was run to see if the mean of birds hit per strike, NumberStruckActual, was the same across all regions. The null hypothesis hypothesized that the mean of the number of birds stuck per incident is the same between all regions. The alternative hypothesis was that the mean of the number of birds struck per incident is not the same between all regions. The significance level, alpha, will be set to 0.05. The p-values returned from the ANOVA model were studied, and if they were less

than 0.05, the null hypothesis was rejected and evidence would've been found that the means between the regions are not all the same. If differences were found, Tukey's HSD test was used to determine which regions have differences.

This will not be a causal estimate, as there are many other things not available in the dataset that may make the number of birds struck different including time of day, and the number of birds present and available to strike. A map visualization was used to show the average number of bird strikes per airport via a map ggplot.

Results

Part A - What factors impact the number of birds struck per incident the most?

The coefficients from the linear regression model were as follows:

Coefficient	Estimate	p-value	Significance Level
(Intercept)	1.583e+00	0.049869	*
Altitude	-2.511e-04	2.29e-05	***
Engines	5.780e-01	0.013584	*
FlightPhaseClimb	9.277e-01	9.34e-05	***
FlightPhaseDescent	5.833e-01	0.297063	
FlightPhaseLanding Roll	1.531e-01	0.519334	
FlightPhaseParked	-1.163e+00	0.794215	
FlightPhaseTake-off run	3.823e-01	0.112498	
FlightPhaseTaxi	-1.171e+00	0.484559	
ConditionsPrecipitationFog, Rain	-6.343e-01	0.686320	
ConditionsPrecipitationFog, Rain, Snow	-3.028e+00	0.734469	
ConditionsPrecipitationFog, Snow	-2.662e+00	0.673824	
ConditionsPrecipitationNone	6.750e-01	0.269846	
ConditionsPrecipitationRain	-4.093e-01	0.553174	
ConditionsPrecipitationRain, Snow	-5.178e-01	0.920311	
ConditionsPrecipitationSnow	9.471e-01	0.541484	
ConditionsSkyOvercast	3.788e-01	0.162548	
ConditionsSkySome Cloud	-1.437e-01	0.446219	
PilotWarnedY	2.973e-01	0.085486	.
regionnortheast	9.747e-01	0.000192	***
regionsoutheast	3.589e-01	0.134046	

Coefficient	Estimate	p-value	Significance Level
regionsouthwest	1.284e-01	0.656828	
regionwest	4.024e-01	0.115306	

Legend:

*** : value is significant at the $p < 0.001$ level

** : value is significant at the $p < 0.01$ level

* : value is significant at the $p < 0.05$ level

. : value is significant at the $p < 0.1$ level

Because the p-value threshold was set to be $p = 0.05$ for this problem, only those coefficients with 1 or more *s are considered significant. The significant features are as follows:

- For every increase of 1 foot in altitude, the number of birds predicted to be struck decreases by 2.511e-04.
- For every increase of 1 engine that a plane has, the number of birds predicted to be struck increases by 5.780e-01.
- When a flight is in the "Climb" phase, the number of birds predicted to be struck increases by 9.277e-01, compared to a flight that is in the "Approach" phase.
- When a flight originates in the northeast, the number of birds predicted to be struck increases by 9.747e-01, compared to a flight that originates in the mideast.

No other variables have a significant impact on the predicted number of birds hit by a strike.

Part B - Does the number of birds struck per incident change based on the region of the airport?

The results of the ANOVA model are as follows:

```
               Df  Sum Sq Mean Sq F value  Pr(>F)
region          4    2884   720.9   4.562 0.00111 **
Residuals   23418 3700218   158.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These results show that there *is* evidence that the difference between the means of birds struck across regions is not zero, because the p-value is less than 0.05. The null hypothesis was rejected in favor of the alternative hypothesis, that there *is* a difference between the mean number of birds struck per incident by region. Because of this result, Tukey's HSD was run to determine which regions had evidence of a difference in means.

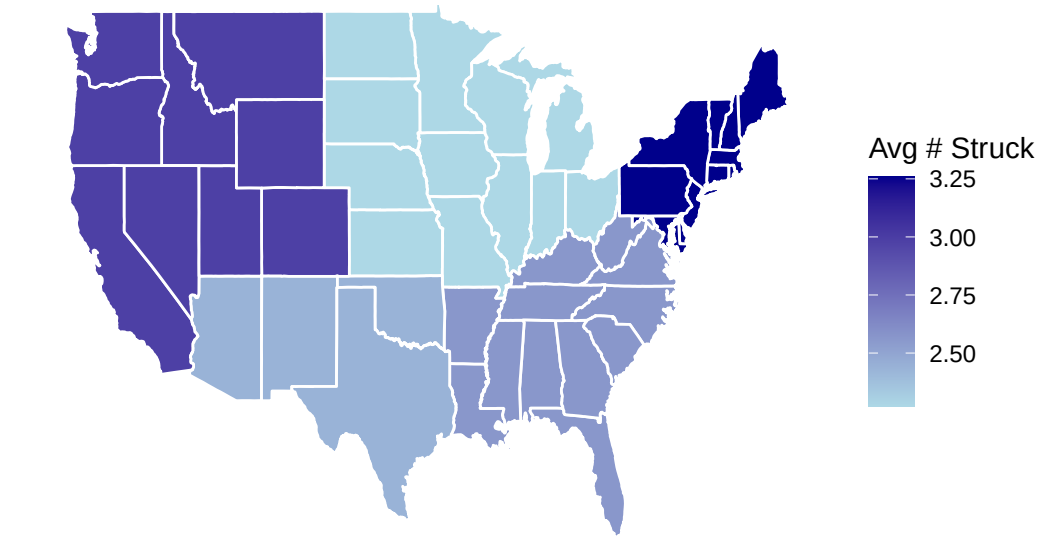
```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = NumberStruckActual ~ region, data = data_us_only)

$region
                           diff        lwr        upr     p adj
northeast-midwest     0.98636706  0.2794889  1.6932452 0.0013307
southeast-midwest     0.32135701 -0.3242327  0.9669467 0.6547171
southwest-midwest    -0.05456861 -0.8313067  0.7221695 0.9996995
west-midwest          0.38121727 -0.3082690  1.0707035 0.5570262
southeast-northeast -0.66501005 -1.3485498  0.0185297 0.0610559
southwest-northeast -1.04093567 -1.8494915 -0.2323798 0.0040717
west-northeast      -0.60514979 -1.3302924  0.1199929 0.1524848
southwest-southeast -0.37592562 -1.1314863  0.3796350 0.6551133
west-southeast       0.05986026 -0.6056780  0.7253985 0.9992015
west-southwest       0.43578588 -0.3576101  1.2291818 0.5635131
```

When a 95% CI contains 0, it is possible that the mean difference is 0, so the result is not significant. Only two sets of regions do not contain 0 in their 95% CI, so only two sets of regions have significant differences in means.

- Northeast & Midwest
  - Difference: 0.99 Birds
  - 95% CI: [0.28, 1.69]
- Southwest & Northeast
  - Difference: -1.04 Birds
  - 95% CI: [-1.84, -0.23]

This means that the difference in the average number of birds struck in the Northeast is 0.99 birds more than the average in the Midwest. Similarly, the difference in the average number of birds struck per strike in the southwest is 1.04 birds fewer than in the Northeast. A map outlining the differences by region is shown below.

```
ggplot(merged_data, aes(x=long, y=lat, fill = region_avg, group=group)) +
  geom_polygon(color = "white") +
  ggtitle('Average Number of Birds Struck per Incident by Region') +
  scale_fill_continuous(low = "lightblue", high = "darkblue", name = "Avg #
   ↪  Struck") +
  #  scale_fill_viridis(name = "Avg # Struck", limits = c(1, 13)) +
  theme_minimal() +
  theme(axis.title.x = element_blank(),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.title.y = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank(),
        panel.grid = element_blank(),
        plot.title = element_text(hjust = 0.5))
```



Discussion

There are many factors that impact how many birds are struck per incident. The result that no weather conditions had a significant impact on the number of birds struck per incident was surprising, as intuitively, I'd have expected the weather conditions to impact the presence of certain quantities of birds.

Although the result that altitude has a negative effect on number of birds strike may seem counter-intuitive for lower altitudes, because planes reach altitudes of approximately 30,000 feet and birds very rarely have the capability of flying that high, bird strikes at very high altitudes are next to impossible. This result confirms that more bird strikes occur at lower altitudes, which aligns with the reasonable range for birds to be flying in.

When planes have more engines, they are also expected to hit more birds per strike, which aligns with the idea that planes with more engines tend to be bigger. When a plane has more surface area available, it is possible for it to hit more birds in one strike.

When a plane is in the "Climb" phase, there is a statistically significant impact on number of birds struck compared to a plane in the "Approach" phase. Although these flight phases have some overlap in Altitude, the Climb phase is much longer than the Approach phase, which may explain why more birds are struck per incident in this phase - they simply have more time to be struck.

A plane taking off from the Northeast compared to the Midwest has a significant impact on the number of birds struck as well. This is related to part B and is discussed in more detail below.

In the future, it'd be interesting to study how the time of year has an impact on the number of birds struck per incident. In certain months, birds tend to travel in larger packs due to migration patterns, so it is possible that bird strikes with larger numbers of birds are more common then. It'd be interesting to combine some of the results from questions 2 and 3 and include the month as a predictor in the regression to see if that has an impact.

For part B of this question, only two sets of regions were found to have significant differences. Consultation with experts with more understanding of bird population distributions might be necessary for a full understanding of these difference, but an initial investigation into the bird population patterns showed that since 1970, the Midwest has seen a much higher bird population decline than the Northeast. This could be a reason that the number of birds hit per strike is higher in the Northeast, because there may be more birds present. The Southwest region in this dataset is made up of grasslands, arid lands, and western forests. A majority of the land is made up of grasslands and western forests, which have seen higher declines in bird population than have been seen in the mostly 'eastern forest' land that makes up the Northeast. So a similar conclusion can be drawn that perhaps there are not as many birds readily available in the Southwest compared to the Northeast.

In the future, it'd be helpful to gain a deeper understanding of the populations of birds in each region of the US. It'd be important to understand how different types of birds behave and whether or not they travel in flocks. If a region is made up of birds that mostly travel in flocks, it'd be possible that that region may see higher quantities of birds struck per incident.

Q4 - KELSEY

Useful Links

- Bird Strikes Data Set
- Project GitHub