Question2: Seasonality & Trends

Q2 - Seasonality & Trends:

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?

Loading the Dataset

Loading in the dataset and printing out the first few rows to verify that it was loaded in correctly.

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
```

```r
strikes_by_month <- data %>%
  group_by(Year, Month) %>%
  summarise(Count = n())
```

```r
# 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. We will set the count of bird strikes for each month-year as the outcome and the predictors as 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)

# 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:

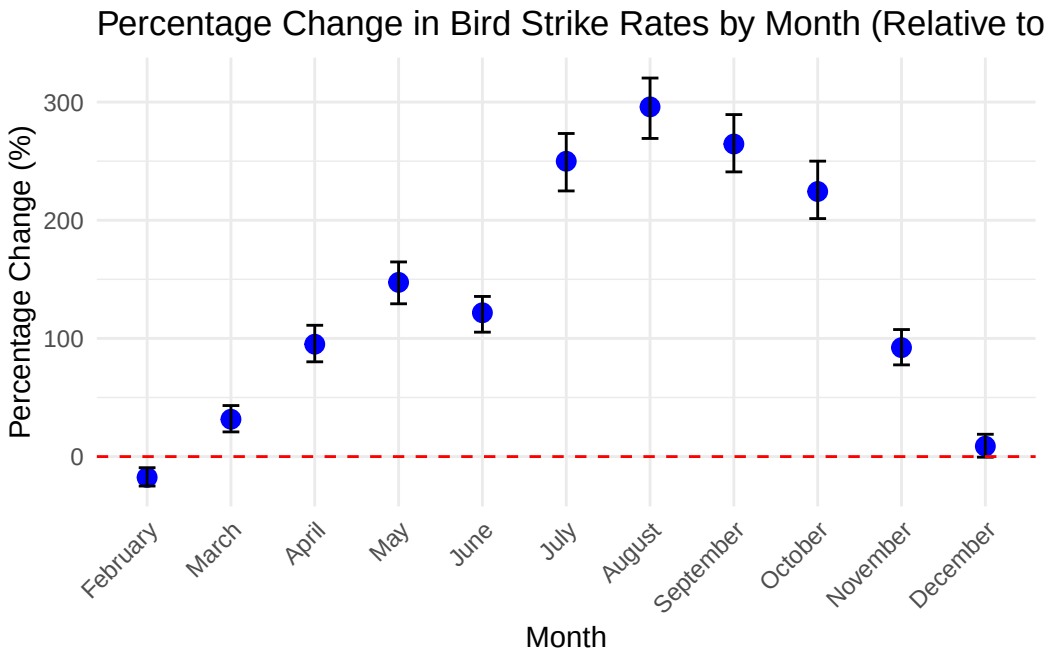
Month	Percentage Relative Rate	Percentage Difference
01 - January (Intercept/Baseline)	1 (Baseline)	0% (Baseline)
02 - February	$(e^{-0.1937}) * 100 = 82.4\%$	-17.6%
03 - March	$(e^{0.27452}) * 100 = 131.6\%$	+31.6%
04 - April	$(e^{0.6683}) * 100 = 195.1\%$	+95.1%
05 - May	$(e^{0.9058}) * 100 = 247.4\%$	+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.

```r
# 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)
# 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 = "Percentage 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.

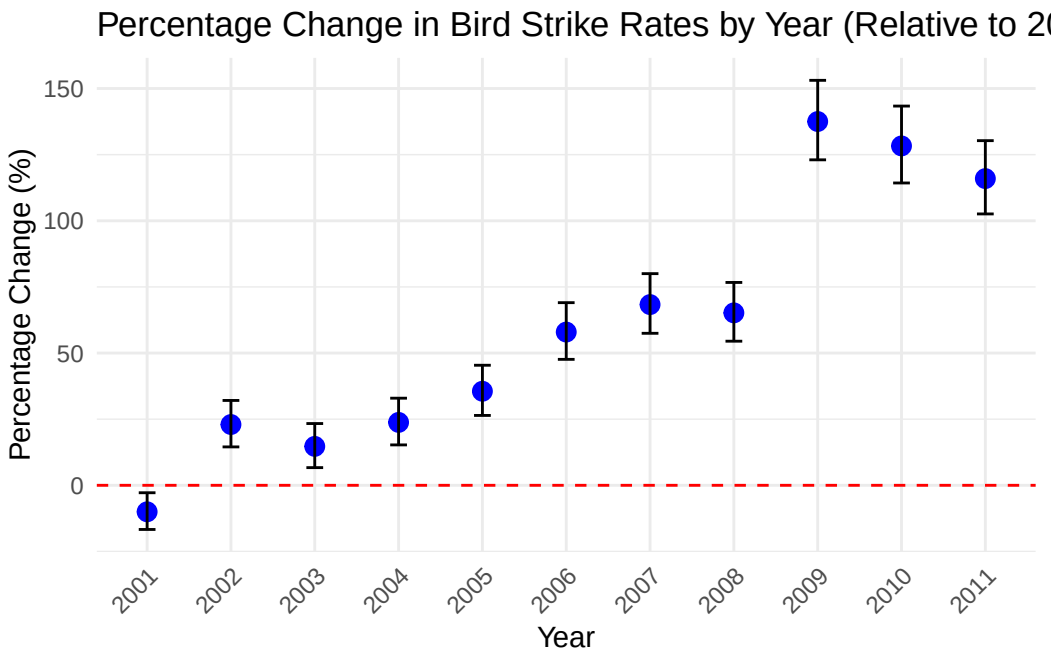
Year	Percentage Relative Rate	Percentage Difference
2000 - Intercept/Baseline	1 (Baseline)	0% (Baseline)
2001	$(e^{-0.1056}) * 100 = 90\%$	-10.0%
2002	$(e^{0.2068}) * 100 = 123\%$	+23.0%
2003	$(e^{0.1372}) * 100 = 114.7\%$	+14.7%
2004	$(e^{0.2133}) * 100 = 123.8\%$	+23.8%
2005	$(e^{0.3042}) * 100 = 135.5\%$	+35.5%
2006	$(e^{0.4570}) * 100 = 158.0\%$	+58.0%

Year	Percentage Relative Rate	Percentage Difference
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)
```

```r
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 = "Percentage 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%.

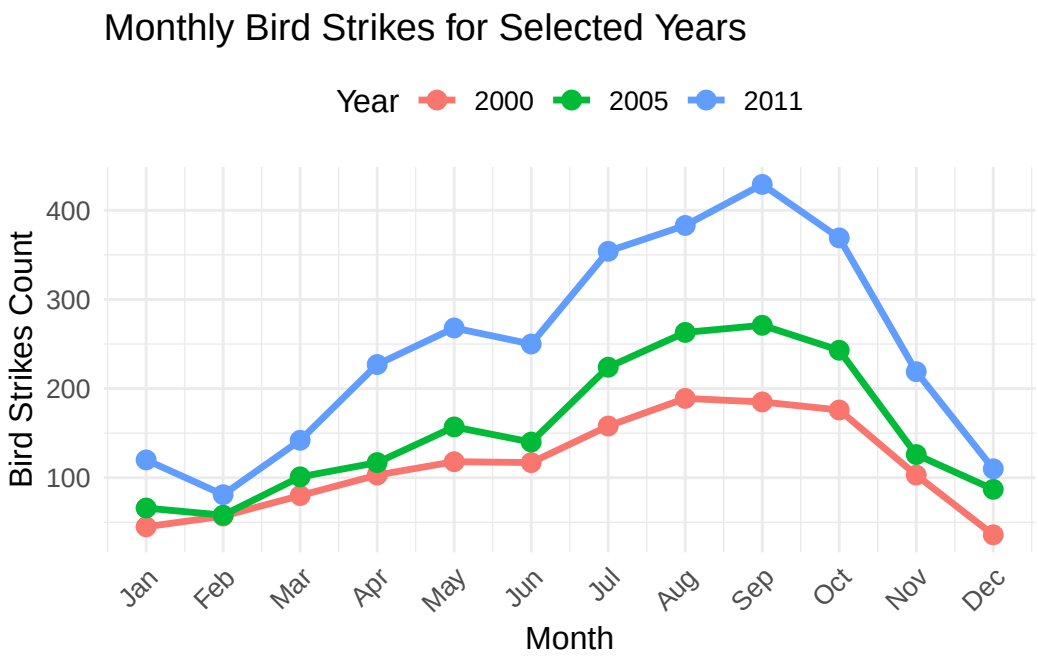
Question 2: Are bird strikes increasing over the years?

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)
```

```r
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 probablility of a bird strike occurring?