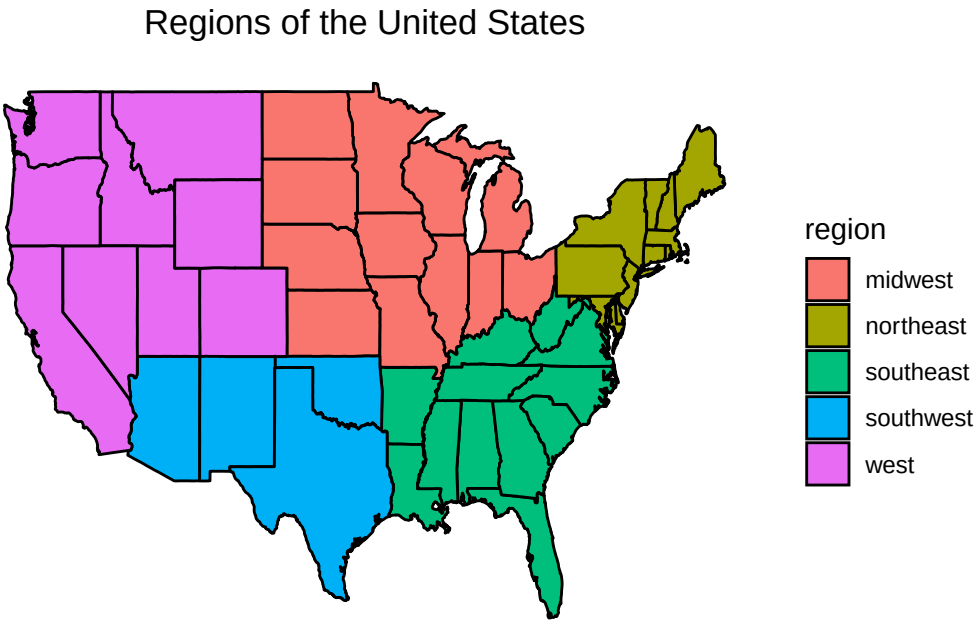
Q3 - Quantity of Birds Struck

Question

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	

Coefficient	Estimate	p-value	Significance Level
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	
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.

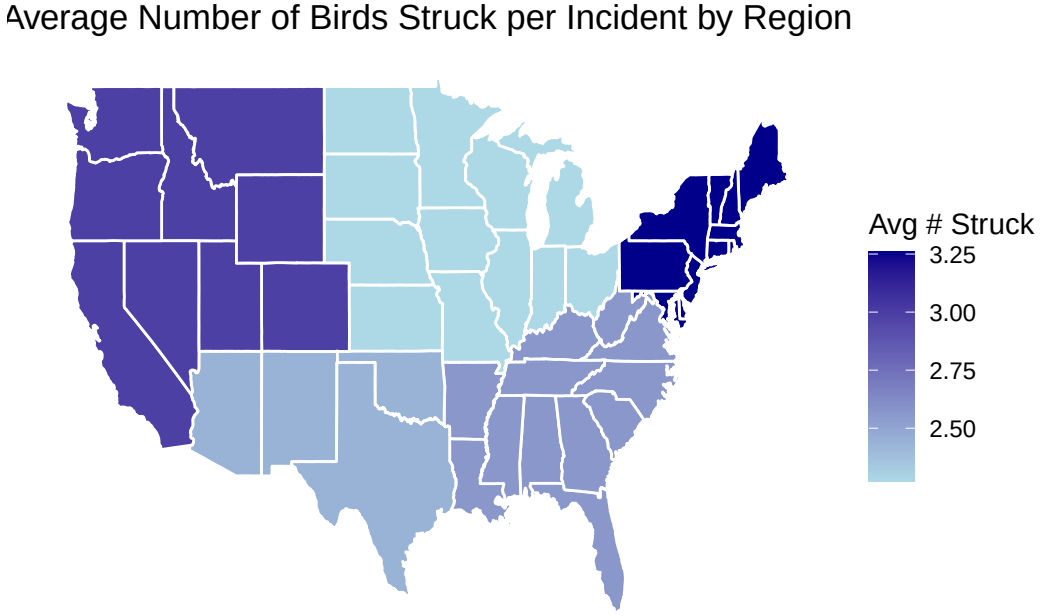
- Northeast & Midwest
  - Difference: 0.99 Birds

  - 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.