Aircraft Wildlife Collisions

Alexander Flick & Vy Nguyen

21/07/2021

Contents

Introduction

In this paper we want to investigate a dataset of aircraft-wildlife collisions in the U.S. from the 90s. We start by introducing the data and some steps we took regarding data preparation. For our analysis we then first take a look at what kind of aircrafts are frequently involved in bird/wildlife strikes. After that we examine what kind of circumstances lead to such collisions and if there are any dependencies between certain variables. In the following chapter we then also take location into account to analyze if this has any influence on the kind of wildlife involved in the collisions. For the last research topic we focus on the time variable and put that in relation to the previous results.

We then finish with a summary of our findings and further research questions that could possibly be analyzed in the future.

1. The Data

1.1 The Birds Dataset

To begin, we first have a look at the dataset we're working with. This is the *birds* dataset from the *openintro* package (OpenIntro 2012): It is a collection of all **collisions between aircraft and wildlife** that were reported to the US Federal Aviation Administration **between 1990 and 1997**, with details on the circumstances of the collision. It consists of **19302 observations and 17 variables** which are given in the table below.

Table 1: Variables of the birds dataset

Variable	Description	Type
opid	Three letter identification code for the operator (carrier) of the aircraft.	Factor w/ 285 levels
operator	Name of the aircraft operator.	Factor w/ 285 levels
atype	Make and model of aircraft.	Factor w/ 284 levels
remarks	Verbal remarks regarding the collision.	Categorical
phase_of_flt	Phase of the flight during which the collision occurred.	Factor w/ 8 levels
ac_mass	Mass of the aircraft in kg classified.	Discrete (1-5)
num_engs	Number of engines on the aircraft.	Discrete (1-4)
date	Date of the collision. (MM/DD/YYYY 0:00:00)	Categorical
time_of_day	Light conditions.	Factor w/ 4 levels
state	Two letter abbreviation of the US state in which the collision occurred.	Factor w/ 58 levels
height	Feet above ground level.	Continuous (0-32500)
speed	Knots (indicated air speed).	Continuous (0-400)
effect	Effect on flight.	Factor w/ 5 levels
sky	Type of cloud cover, if any.	Factor w/ 3 levels
species	Common name for bird or other wildlife.	Factor w/ 241 levels
birds_seen	Number of birds/wildlife seen by pilot.	Factor w/ 3 levels
birds_struck	Number of birds/wildlife struck.	Factor w/ 5 levels

We have 11 factor variables, two categorical ones (with one of them being date information), two discrete numeric variables and two continuous variables. There are also some missing values which we will look at later.

1.2 Preparation

Data Preparation: Cleaning *birds_seen* variable: The 9412th observation had one wrong entry in the *birds_seen* variable which has been replaced with a *NA* value, because we did not want to introduce a new level for one entry.

Factor Levels: The factor levels of the following variables have been (re-)ordered for visualization purposes.

- Order from low to high:
  - birds_struck
  - birds_seen
  - sky (by cloudiness)
  - ac_mass
  - num_engs
- Individual ordering:
  - effect: Reordered to put "None" effect first and separate it from the remaining "actual effect," to achieve a separation between "None" effect and effects.
  - phase_of_flt: Ordered according from parking, over start of flight to end of flight

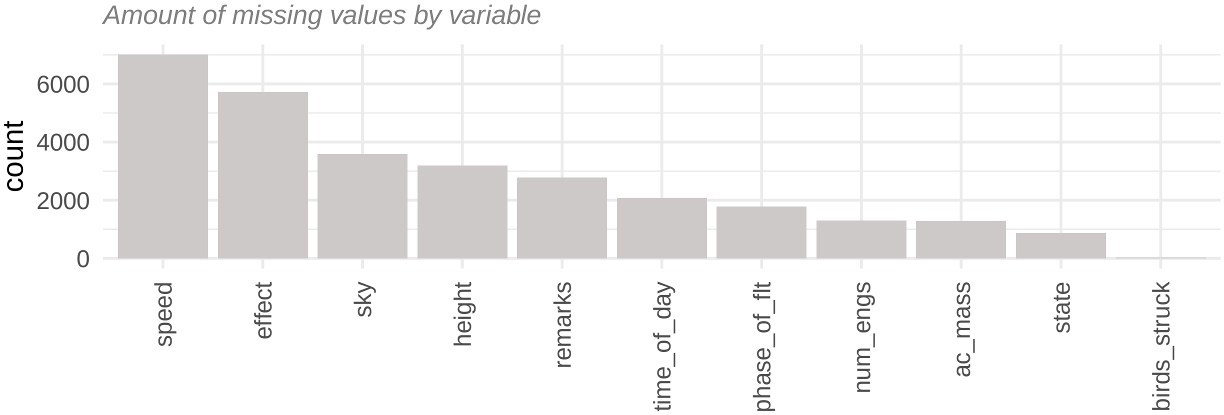
Preparation for Visualization: Since we wanted to visualize our findings and results in a simple and understandable manner, we were very considerate with the colors and type of plots we used. We looked at different color palettes (Data Novia 2018; Melanie Frazier 2020; aximaps 2021) and made sure that colors were distinguishable when comparing different classes, or sequential when in-/decrease was important. We also tried to pick colors that matched with what we wanted to display (e.g. yellow for daylight/no clouds etc.). For the states we found that using a map for visualization worked very well, for timelines a lineplot with points was the easiest to interpret, we used scatter, bar-, density-, spine- and mosaicplots for comparing amounts and proportions/ratios and boxplots to compare different levels of a factor variable with each other. And for certain questions wordclouds (E. Le Pennec 2019) worked the best. We then made sure that axis labels, plot titles and legends were concise and clear.

1.3 Handling missing values

There were no missing values for the following variables: atype, birds_seen, date, operator, opid and species.

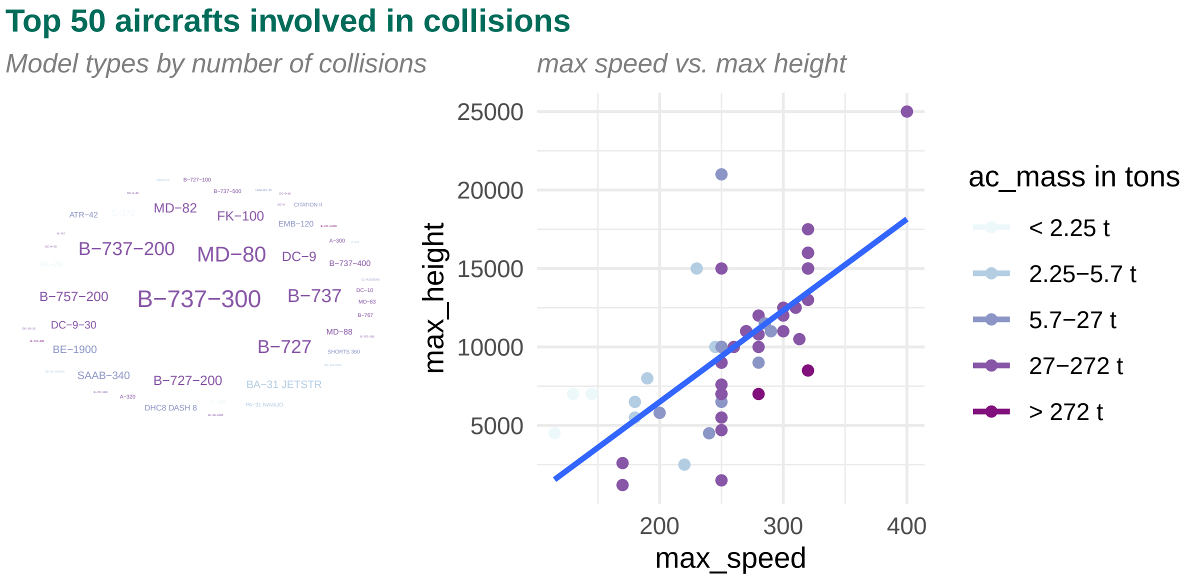
birds_struck: We are using the *remarks* variable to check for spare parts/hints in text from which we can determine that no wild life has been struck. By doing this we were able to decrease the missing values by 19 (from 39 to 20).
birds_seen: We assume that missing values in the *birds_seen* variable indicate that the pilot did not see wildlife before the strike occurred. Thus, we set all *NA* values in the *birds_seen* variable as *None*. This leads to 14539 values which are set as *None*.

The following barplot shows the amount of missing values for each of the variables that had missing values.



2. Aircraft models

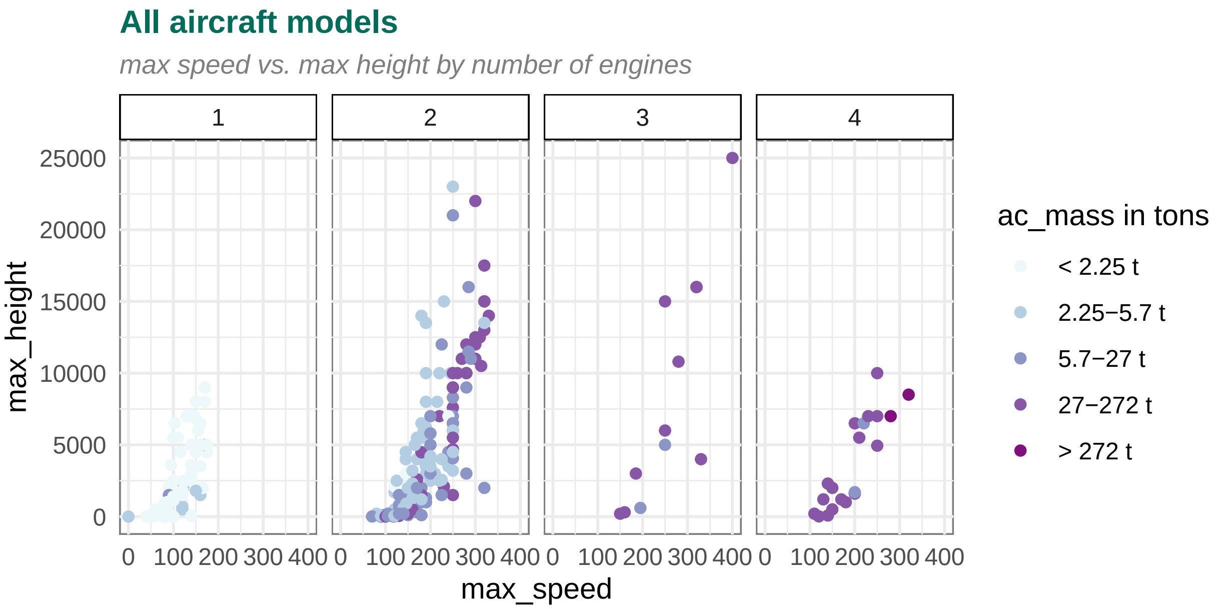
The dataset contains 284 different aircraft models. We extracted the overall maximum speed (*max_speed*) and height (*max_height*) for each aircraft to get an overview of the aircraft models and the relationship between speed and height. The following left plot shows the top 50 aircraft models that are causing the most strikes and on the right you can see the speed height relationship of the top 50 aircrafts colored by their *ac_mass*.



The left plot shows that especially the **B-737 class** (including B-737-200 and B-737-300) causes the most strikes together with MD-80 and B-727. "B" stands here for Boeing and MD-80 is an aircraft from McDonnell Douglas. The model causing the most strikes is the B-737-300 which caused in total 1328 strikes. The B-737 is a narrow-body airliner produced by Boeing at its Renton Factory in Washington and was developed for short and thin routes. Unfortunately we do not have the amount of flights and duration of the flight in the data, so we can not do further analysis on this. But we assume, that the aircraft models with higher amount of strikes had also more flights and or longer flights.

From the right plot we can see, that aircraft **models with higher *max_speed* are also able to fly higher**. This can be also seen from the linear model, that has been estimated by geom_smooth. The correlation of *max_speed* and *max_height* for the top 50 models is: 0.688, which is positive and quite high. In addition we can see via the coloring by *ac_mass* that **lighter aircrafts tend to have less max_speed and max_height**. We can also see that the lighter aircrafts under 5700 kg or even 27000 kg are less prominent in the top 50. The group with an **aircraft mass between 27001-272000 kg is the most prominent group** under the top 50 aircrafts. Two aircraft models (B-747-1/200 and B-747-400) with *ac_mass* of more than 272000 kg are also Boeings and present in the top 50 which are the only aircrafts of that weight class overall.

The next plot shows again the *max_speed* by *max_height* splitted by number of engines and colored by *ac_mass* but for all 260 aircraft models. Since some models had very few strikes it can happen that they have been assigned an *max_height* of about 0.



From the plot we can see, that again higher *max_speed* causes higher *max_height* and most of the aircraft models have two engines. Interesting is that for aircrafts with one engine, there is no model with a speed greater 200 knots and height greater 10000 foot. So we could conclude from that, that this is the maximum height and speed possible for aircrafts with one engine. We can also see that lighter machines tend to have less engines:

- first group includes nearly only machines with aircraft mass less than 2250 kg,
- second group is mixed with aircraft mass between 2251-272000 kg,
- the last two groups include mostly aircrafts with a mass of 27001-272000 kg.

Looking at the contingency table also confirms that:

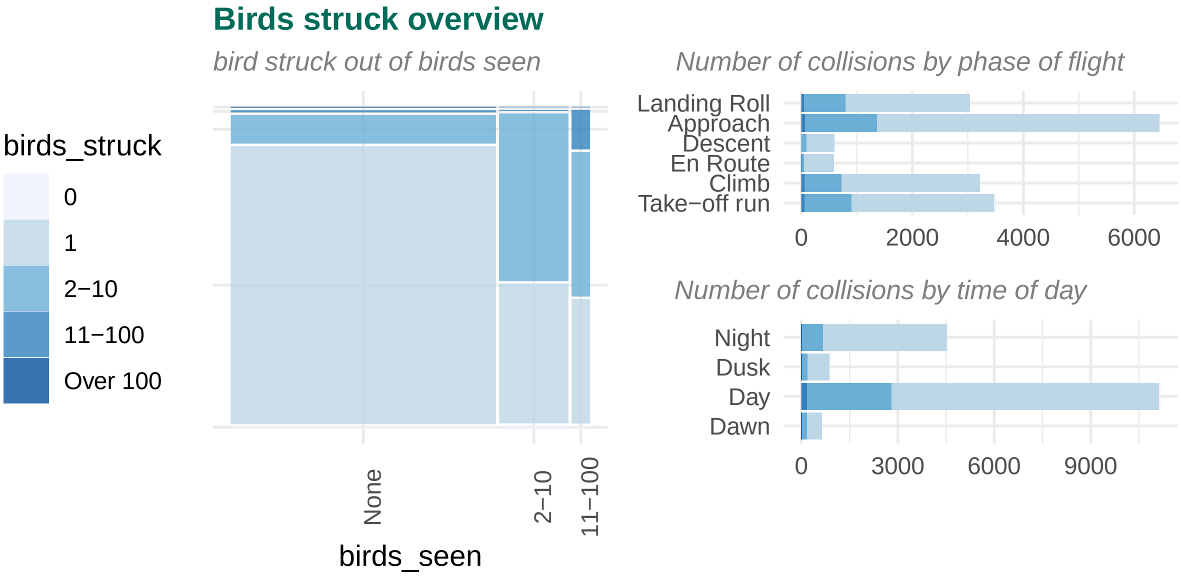
```
##          ac_mass
## num_engs  1  2  3  4 5
##        1 57  9  1  0 0
##        2  8 56 48 38 0
##        3  0  0  2 10 0
##        4  0  0  2 16 2
```

3. Analysis of influences for birds_struck, birds_seen and effect

In the following analysis we will go more into detail about the different conditions regarding the amount of birds that have been struck (variable: *birds_struck*), the amount of birds that have been seen by the pilot before a strike happened (variable: *birds_seen*), and the different effects a strike had on the aircraft (variable: *effect*). Most of the time we will use barplots with position fill, mosaicplot and spineplots as we want to compare proportions and check for dependencies. We will first start with the analysis of strikes.

3.1 Analysis of strikes (birds_struck)

The following plot on the left shows the amount of birds struck out of birds seen. On the top right we visualized the number of collisions by each phase of flight and on the bottom right the plot shows the amount of collisions by time of day. For coloring we use the *birds_struck* variable.

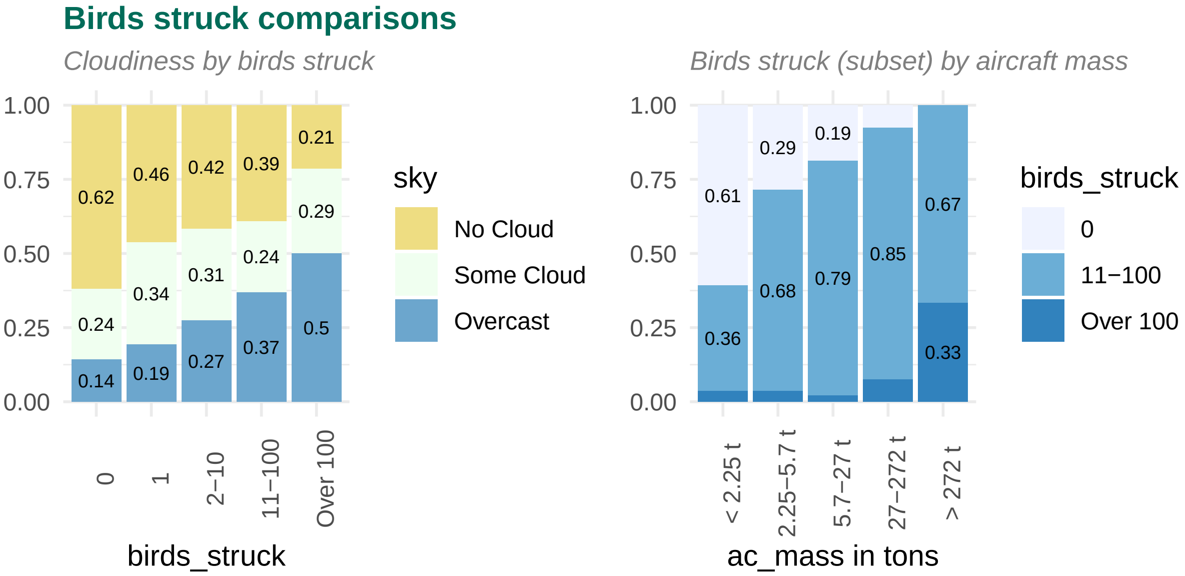


The mosaicplot shows that the largest group of birds seen is "None" followed by "2-10" and "11-100." So **most often we do not see birds before a strike** happens and most often only 1 bird is struck. But if the pilot sees something, then according to the data he only sees birds in groups, where larger groups of birds are less likely than smaller ones. If the pilot sees birds, **quite often less birds were hit than seen**. For example if the pilot sees 2-10 birds, he hits about 40% of the time less birds and for the group *birds_seen* between 11-100, about 85% of the time the pilot hits less birds than seen. The event of hitting less birds than seen will later be analyzed more in detail in chapter 3.3.

The top right plot shows that collisions mostly occur at the beginning and end of a flight in phases Take-Off, Climb, Approach and Landing Roll. The number of collisions in the Approach phase is the highest and about 2 times higher than in the other start and end phases. During the middle of the flight in phases "En Route" and "Descent" less collisions occur.

From the bottom right plot we can see that strikes happen most often during the day with over 10500 collisions followed by night with about 4500. Interesting to see is that strikes with larger groups of birds (11-100) happen mostly during the day (small values are not visible in the plot).

We will now have a look on the influence of the sky variable for the different amounts of *birds_struck* and analyze the influence of the aircraft mass.

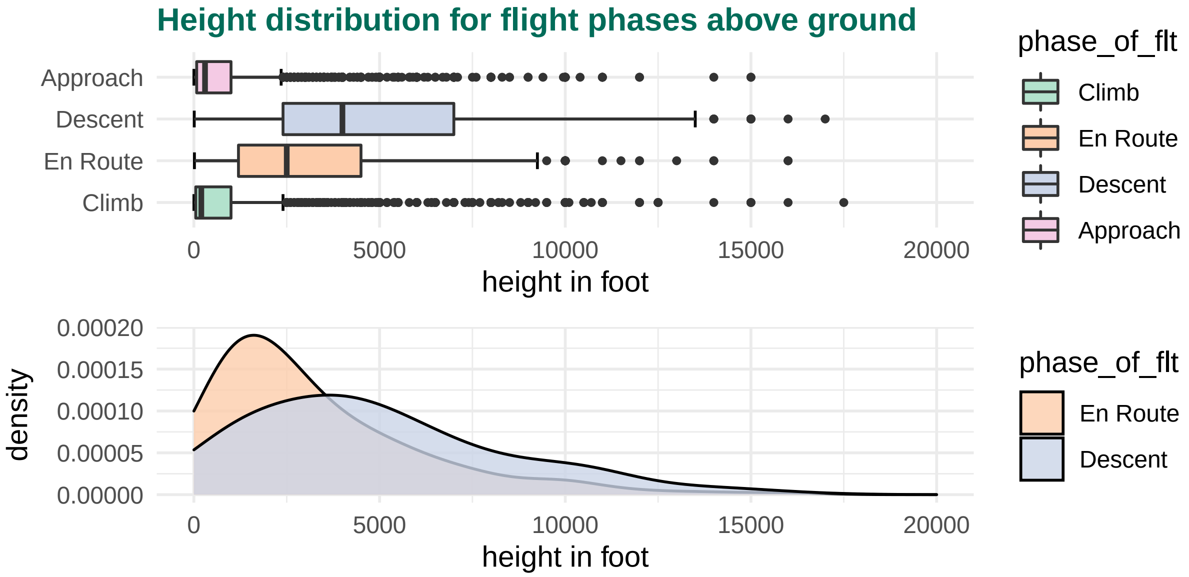


From the left plot we can see that with **increasing number of *birds_struck* the proportion of overcast increases** and the proportion of no clouds decreases. The proportion of some clouds is more or less the same across all groups. From that we can conclude that the higher the amount of birds struck the more clouds and thus we would assume that birds are flying in bigger groups more often when it is cloudy or the pilot as well as the birds are not able to see each other because of the clouds and so nobody can try to prevent the collision by dodging the other one.

On the right side we can see a subset of the *birds_struck* groups, which focuses on the less frequent categories in our dataset. From that we can see that the aircraft mass has an influence on the amount of birds struck. For the lightest aircraft class up to 2250 kg the proportion of 0 birds struck is very high, which means that the pilot has seen a bird and was likely to hit one, but he did not. For this category we would expect that lighter machines are better able to initiate unplanned evasive maneuvers as they are more agile. The relationship of striking less birds than seen will be analyzed more in detail in chapter 3.3. It also stands out, that the proportion of striking over 100 birds is significantly higher for the heaviest aircrafts above 272000 kg. This will most likely be due to the fact that heavier machines are also larger in size and thus they will hit more birds when flying through a flock of birds compared to lighter and thus smaller machines.

3.1.1 Height of collisions above ground

We will now have a look on the height variable together with phase of flight. The top plot shows a boxplot of height for flight phases, which are above ground. We defined Climb, En Route, Descent and Approach as phases which are above ground level. On the bottom plot we will then have a closer look on the height for the flight phases “En Route” and “Descent” by visualizing the density.



The height distribution for “Climb” and “Approach” are very similar and their medians are very close to the ground. The other two boxplots for “En Route” and “Descent” are more far away from the ground and show more variance in the data. Thus, we will focus on these two. Interesting here is that the 25% quantile of the “Descent” boxplot is above the one from “En Route,” same for the median and 75% quantile. We will investigate more the differences of the two distributions by using a density plot.

Both flight phases have the highest density for lower heights up to about 5000 foot. The curve for “En Route” is much steeper, but a little more narrow compared to “Descent.” Both density cross at about 3750 foot, which is the highest point of “Descent” phase. From that height upwards the probability to strike a birds is higher when being in “Descent” phase. This will be most likely because of the angle of the aircraft at which it is flying. In phase “En Route” it will be most likely parallel to the ground, so the height of the airplane does not change much during that phase, but in “Descent” phase the aircraft has an negative angle and is loosing height to prepare, for example, for “Approach” phase.

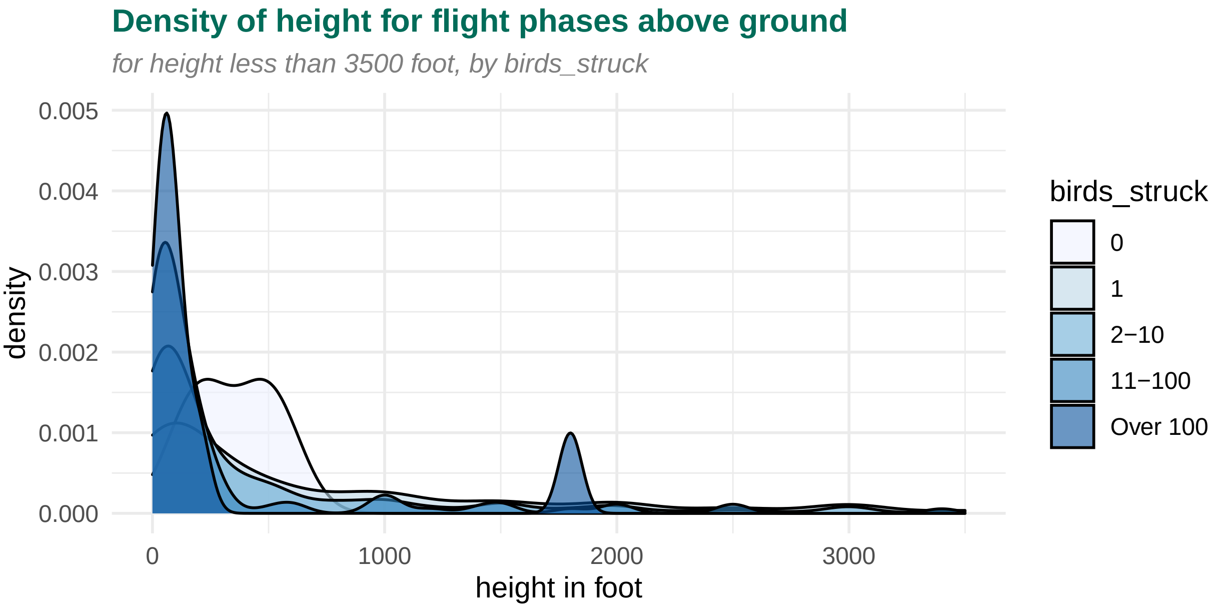
We are performing a t-test to compare the two group heights (“Descent” and “En Route”), so that we can confirm that both distributions are not equal.

```
##
##  Welch Two Sample t-test
##
## data:  height by phase_of_flt
## t = -5.4321, df = 764.3, p-value = 7.491e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1871.7879  -878.0419
## sample estimates:
## mean in group En Route  mean in group Descent
##               3609.116               4984.031
```

The t-test also shows that the **group means are different** from each other, because the p-value is less than 5% or even 1% level. The **mean in group “Descent” is higher** than mean of “En Route.” One would actually expect that the height is smaller than in “En Route” because the Descent phase happens after the “En Route” phase, where the aircraft is loosing height to prepare for approach or during flight in general. So since all aircrafts mostly have to be in the phase “En Route” before going over to “Descent” we can say that it is more likely to hit a bird in greater height (greater 3750 foot) when being in phase “Descent” rather than in “En Route.” The main difference between these flight phases is that the aircraft is not flying parallel to the ground when being in phase “Descent.” So the **angle of the aircraft during the flight is most likely having an influence on hitting birds**. And especially a negative angle increases the probability of hitting birds, based on the findings here and from chapter 3.1.1 where we found out that the “Approach” phase has the highest number of strikes.

3.1.2 Density of height by birds_struck for flight phases above ground

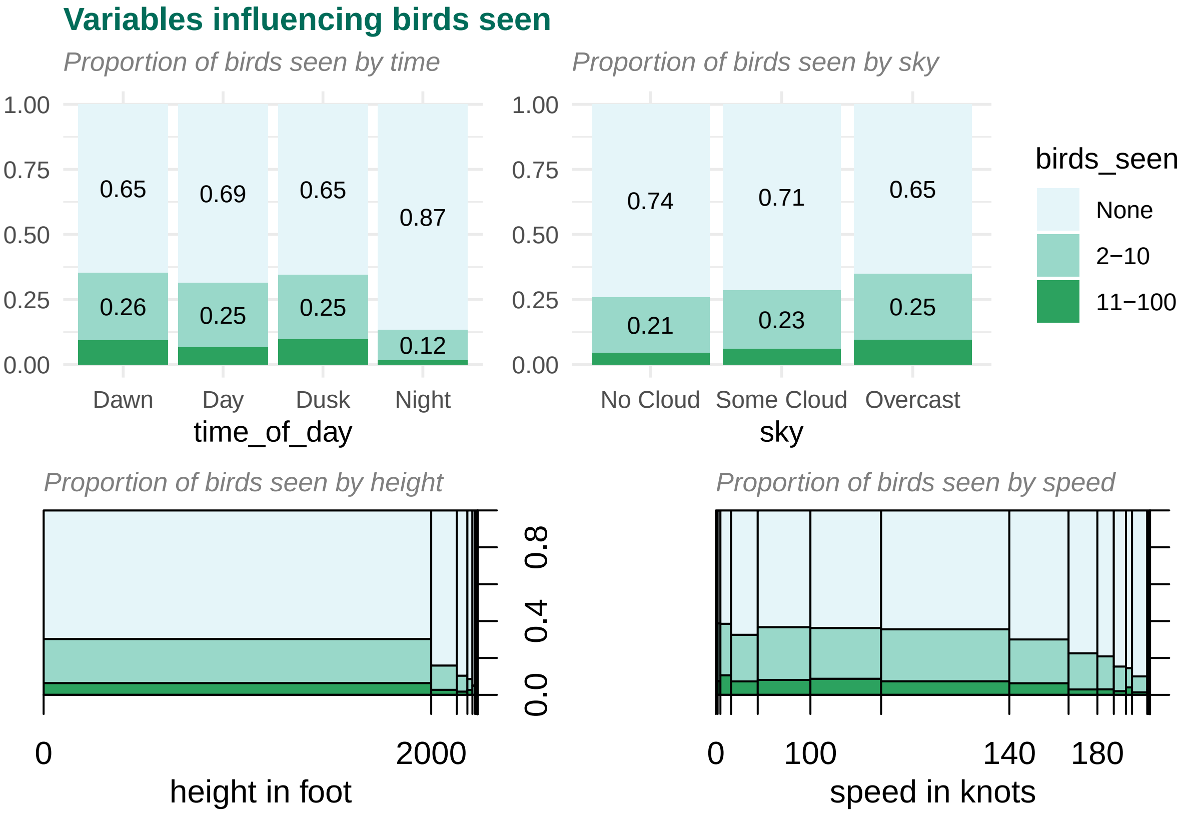
We will now have another look on the density of *height*, but this time split by the different numbers of birds involved in the collision. The following plot shows exactly this.



The density plot shows that the probability for striking a bird is highest near ground level. The density starts with a **very steep and narrow peek at the beginning at about 100 foot** for all categories of birds_struck greater 0. So most of the collisions happen near ground level, which we have already seen at the beginning of chapter 3.1.1 that the flight phases near ground level (Take-Off, Climb, Approach and Landing Roll) have the highest number of strikes (see 3.1 Number of collisions by flight phase). For *birds_struck* equal to 0 the distribution is a little bit different from the others with a density of about 0.0016 from 200 to 500 foot. Remarkable is also the **peek for “Over 100” birds struck at a *height* of about 1800 foot**. For the “Over 100” category we have the lowest amount of observations in the dataset but nevertheless 4 out of 8 strikes happened here. So for this category we have the same probability of hitting birds near ground level and at about 1800 foot. One of the involved species here was a **flock of about 500 gulls**. Unfortunately the other observations had no information about the *species* that were involved in the collision.

3.2 Analysis of birds seen

There originally were only two levels in the data available for birds seen and we added one level for the missing values. The two plots at the top show the proportion of birds seen by the *time_of_day* (left plot) and the cloudiness of the *sky* (right plot). At the bottom you can see two spineplots that show the proportion of *birds_seen* along the two continuous variables *height* and *speed*.



Time of Day: Based on the assumption that missing values in the data for *birds_seen* stands for the fact that the pilot did not see wildlife/birds before a strike happened, the average proportion (unweighted) of birds seen before a strike happened is 1-(0.65+0.69+0.65+0.87)/4=0.285. So most often pilots do not see birds before a collision. The proportion of overall **birds seen at night is much lower** with just 0.13 compared to the other times (Dawn, Day and Dusk).

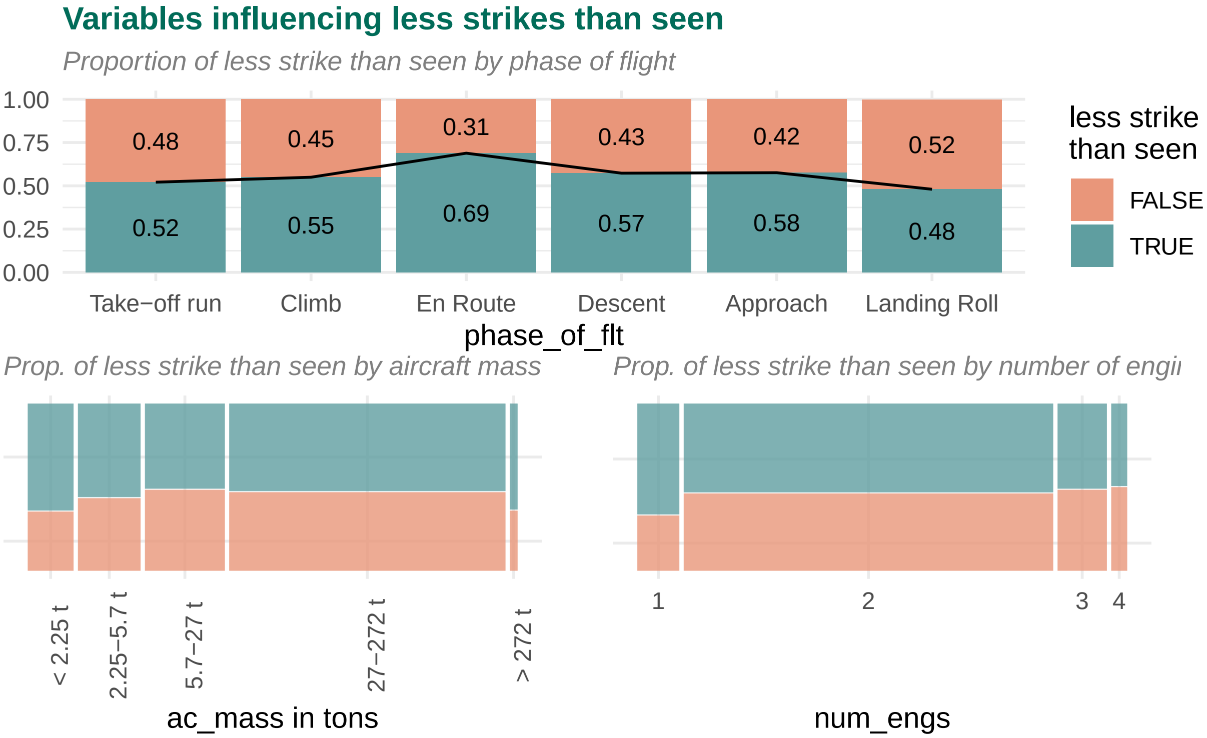
Sky: The proportion of **wildlife seen increases with cloudiness**. This is a **contradiction** to what we would have expected. We would have thought that with increasing cloudiness the amount of wildlife/birds seen decreases, due to the fact that the pilot can not see as good as when there are less or no clouds. So one reason for this could be that most of the pilots aren't within the area of clouds when the collision happens, but have recorded the cloudiness of the sky. Another reason could be that birds fly more often when its cloudy.

From the spineplots we can see again that it is less likely to see birds before a collision, but the higher the *speed* & *height*, its becoming even less likely to see birds before a collision. For heights greater 2000 foot and for speed greater 140 knots, the probability to see birds before a collision decreases.

3.3 Influences when number of birds struck is less than birds seen

In the following we will have a closer look on a self-generated variable, which compares the amount of *birds_seen* with the amount of *birds_struck*. It will be true if the number of birds struck is smaller than birds seen and false if not. In the beginning (chapter 3.1.1) we have already seen, that the number of *birds_seen* does not always infer that the same amount of birds were struck. By using this variable, we want to have a look at what influences the proportion of striking less birds than seen.

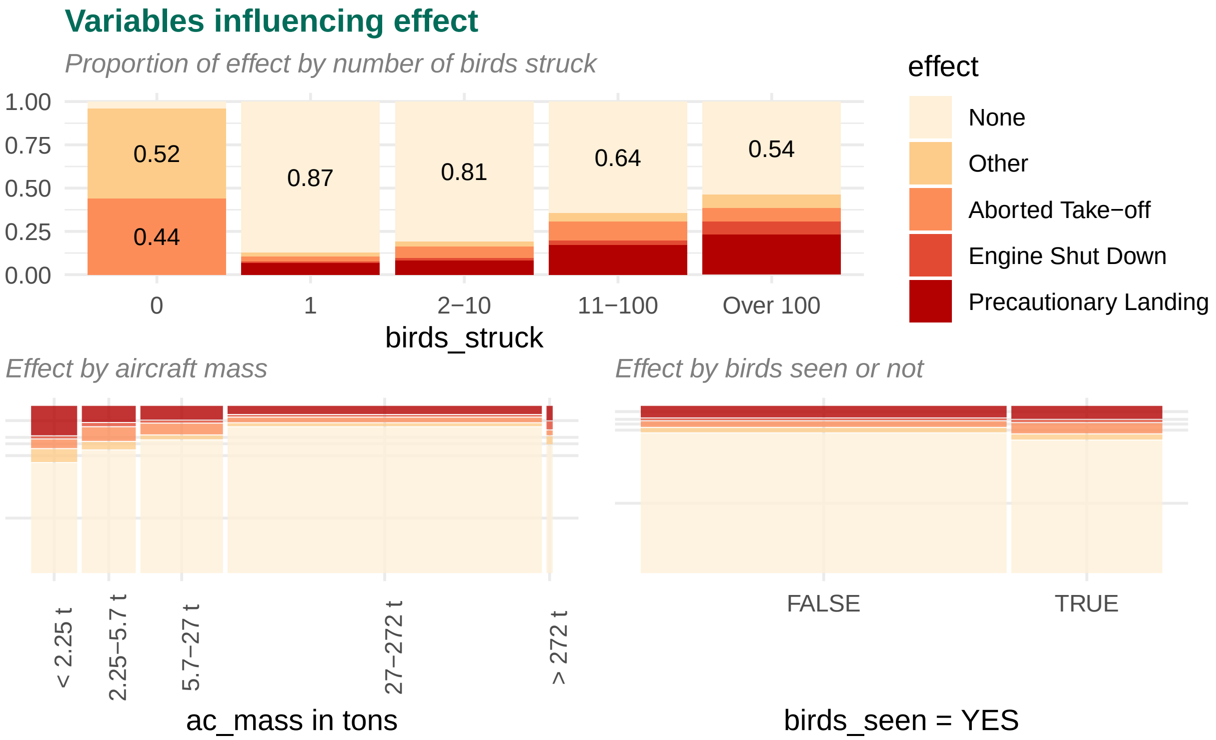
The top plot shows the proportion of less strikes than seen by phase of flight. On the bottom left we can see a comparison of the proportion of less strikes by the aircraft mass and on the bottom right we compare it by different number of engines.



For the top barplot we observe a relationship based on the phase of flight, like we did earlier in chapter 3.1.1 when comparing the number of collisions by flight phases over the start, middle and landing of the flight. The proportion of less strikes is increasing from the start of the flight (Take-Off Run) and reaches the highest proportion of 0.69 in the phase "En Route" and then decreases again, where it reaches a minimum of 0.48 in the last phase "Landing Roll." So we see again, that the angle of the flight must have an influence on striking birds, this time especially on striking less birds than seen. The number of engines and aircraft mass also have an influence of how much birds are hit out of the birds seen. Lighter machines and machines with less engines tend to hit less wildlife out of the seen wildlife. This might be again due to better agility of the lighter and thus smaller aircrafts. Most prominent are aircrafts with ac_mass = 4 (27001-272000 kg) and two engines which matches also with the proportions of unique aircraft models from chapter 2.

3.4 Analysis of effect variable

The description of the *effect* variable from the dataset is insufficient. We can not really find out what is meant by the *effect* variable. Does for example “Precautionary landing” and “Aborted take-off” mean that the pilot did this to prevent striking birds, or did he had to do this because of the collision. We interpreted this variable as it really is the effect on the aircraft because a collision occurred. The top plot shows the different proportions of *effect* for each category of *birds_struck*. The Bottom left plot shows the proportions of each *effect* by the aircraft mass and on the bottom right we compare effect with *birds_seen* or not seen.



The greater the number of *birds_struck*, the more likely that the collision has an *effect* on the aircraft. Especially the proportion of “Precautionary Landing” increases. For *birds_struck* equals 0 we have very small proportion of “None” effect and high proportion of “Other” and “Aborted Take-off,” which would actually speak for the *effect* variable representing an action that was done in order to prevent the collision. So the interpretation of the *effect* variable needs further research or the publisher of the dataset should be contacted asked for a more precise description. For the aircraft mass, we can see that the larger the aircraft mass the less likely the collision has an effect on the aircraft. This might be due to the fact, that heavier machines are more robust. But for aircrafts with a mass greater 272000 kg we can see a an increase in the proportion of *effect*, which might be due to the fact that these machines have a higher probability to strike over 100 birds, which will then also have an effect on the heavier aircrafts.

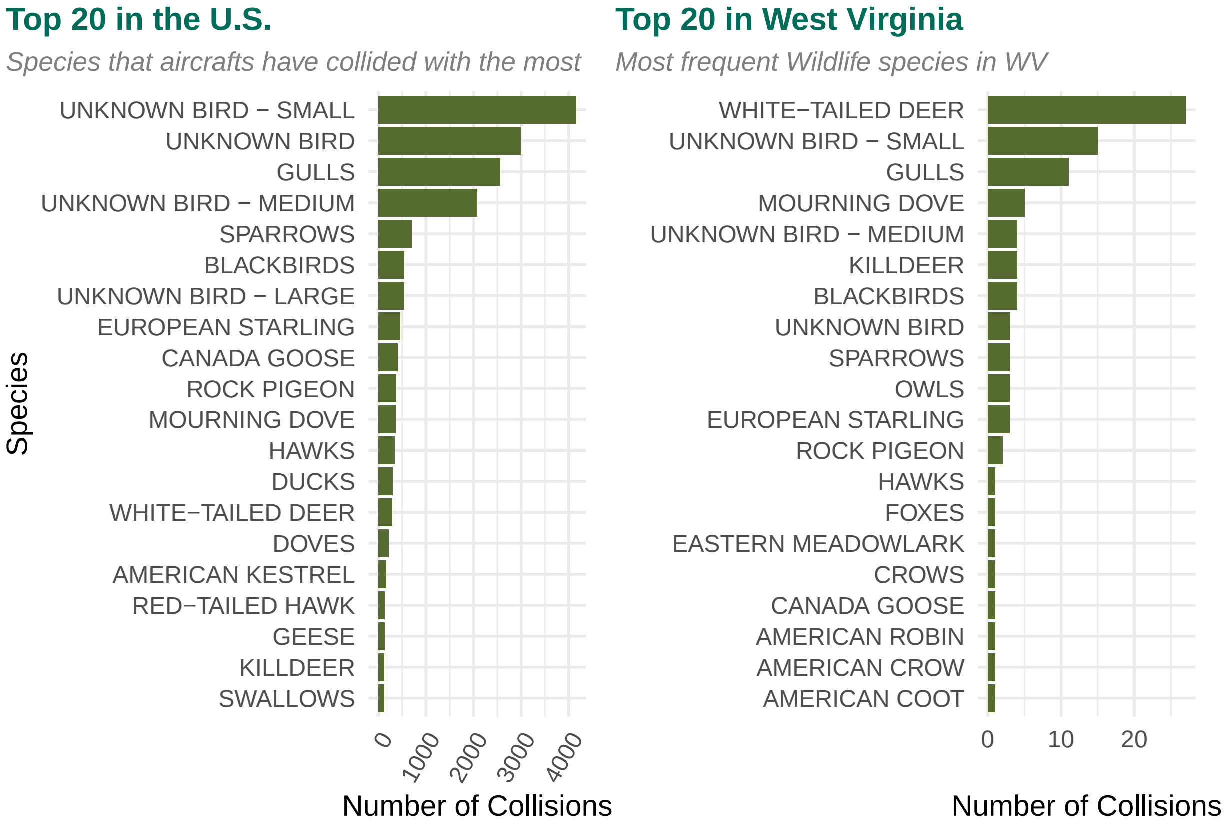
Having a look on the last plot, we can see that the “None” proportion of *effect* is a little smaller when the pilot has seen birds before the strike and “Engine Shutdown” and “Other” are a little higher. We would have actually expected to have less effects on the aircraft when the pilot has seen birds before the strike, because the pilot could then try to prevent the collision, but this is not the case here. So this again would speak for a different interpretation of the *effect* variable.

4. Is there a Relationship between State and Wildlife?

For the next research question we wanted to explore if there are different kinds of wildlife that aircrafts collide with depending on the state they're in: So essentially the relationship between the State and the Species.

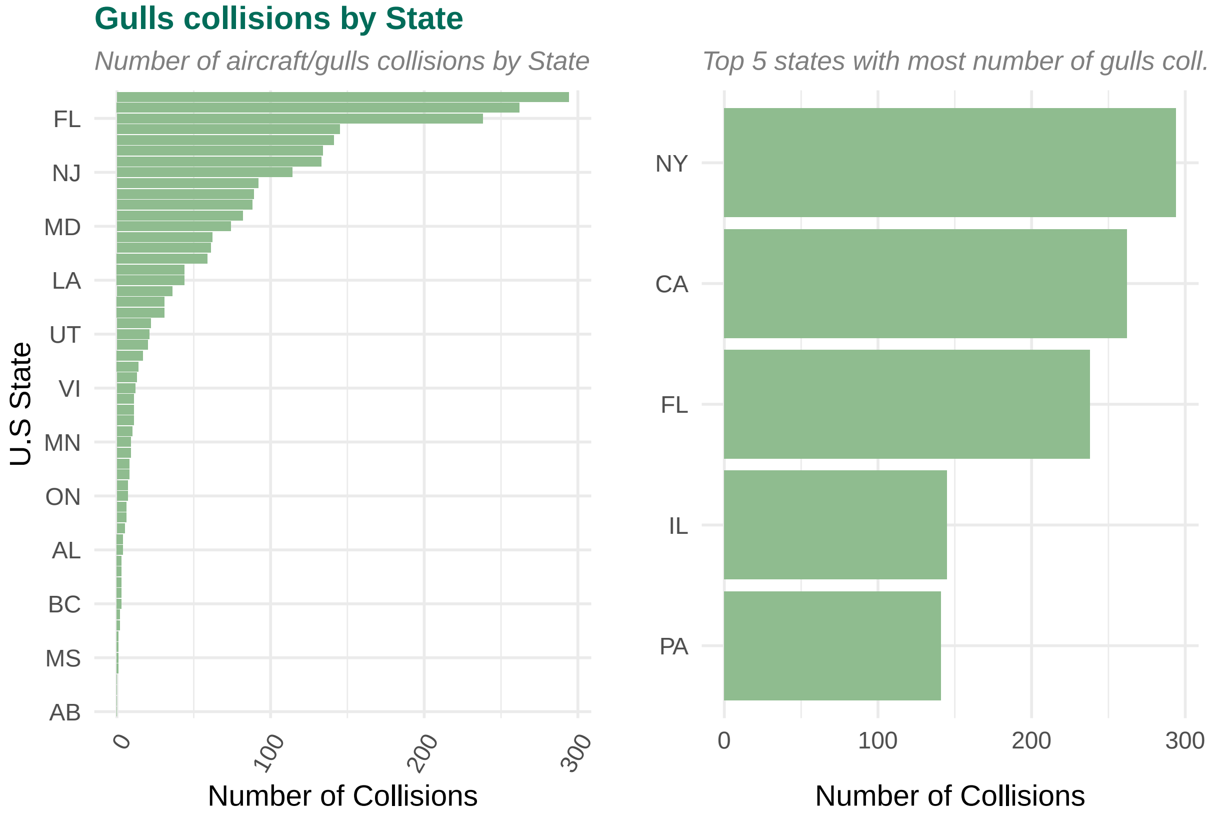
4.1 Species by State

To begin we looked at the **20 most frequent species** in the whole US. The figure shows the number of animals of a specific type which has been displayed in a sorted barplot. We can see that the most frequent species that appear in the data are birds which could not be identified further (e.g. over 4000 collisions for small unknown birds). Apart from those we can verify the fact that most bird strikes involve birds with big populations, particularly geese and gulls in the United States (Wikipedia 2021a). The third most frequent animal, for example, are gulls with over 2500 collisions.



When we compare that with the most frequent species in West Virginia (WV), the most frequent species that aircrafts have collided there with is surprisingly not a bird but the white-tailed deer (Over 30 collisions). And foxes are also among the list. For the rest there are minor differences, for example European Starlings, Rock Pigeons, Sparrows etc. are still one of the most frequent wildlife animals. However, one can see that for example ducks are not that frequent on in WV compared to the nationwide average. Owls also appear more there than in the US in general.

To further examine the state and species variables we could, for example look at all the collisions where gulls where involved and check what states those happened in the most. This is again displayed in barplots:



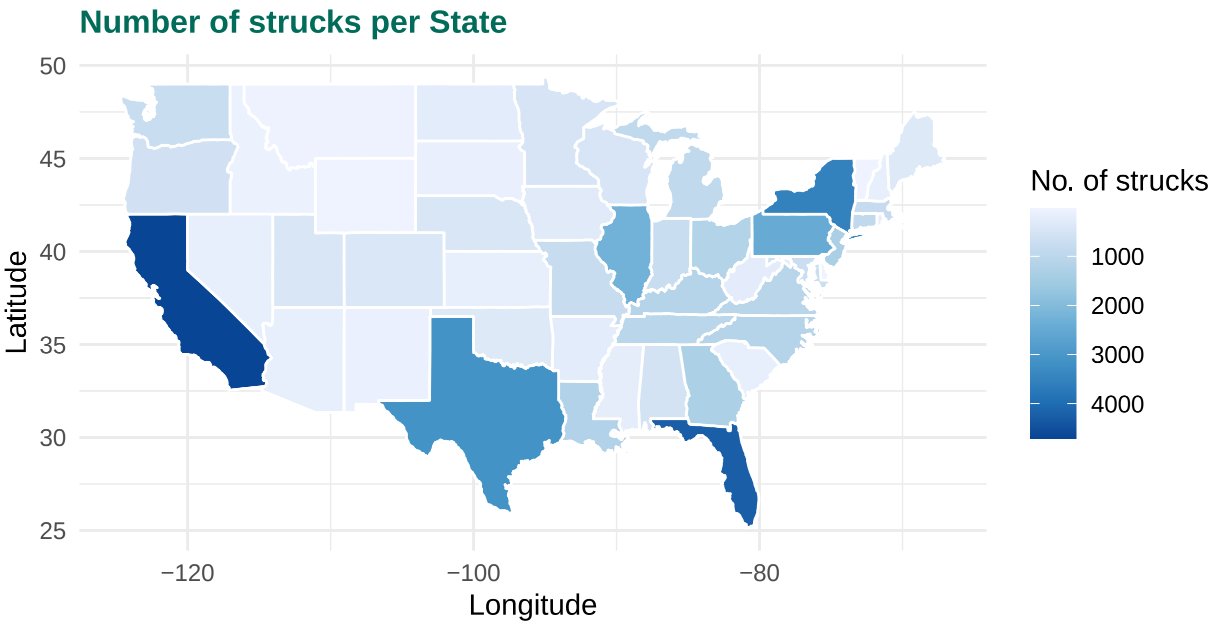
When we look at the top states we see that those are all mainly US states that lie on the coastline, for example California, New York and Florida with between 200 to 300 gulls collisions each. This therefore makes sense that collisions with gulls happen there a lot since those birds are often found near where the sea is (Wikipedia 2021b).

4.2 Collisions per State

We also further examined the **number of collisions per state** which is displayed below.
For this, we wanted to display our findings in a map for better visualization (Zhiyi Guo and Fan Wu 2019; ggplot2 tidyverse 2021). In the following we have displayed the number of collisions, so the variable *birds_struck*, depending on the state. Since this is a factor variable we have unclassed it and mapped the levels to the numeric values below. We then grouped the data by state and summed those up to display in the U.S. map.

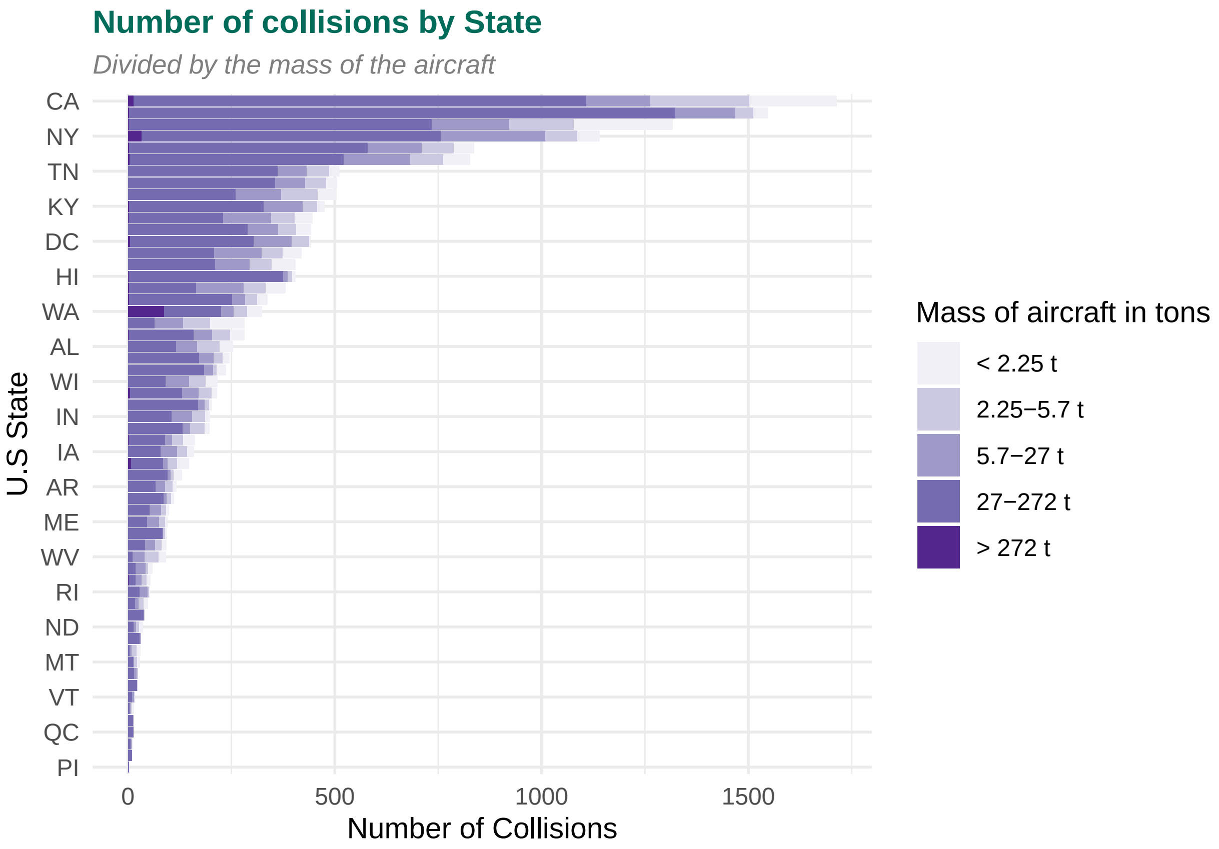
Table 2: Mapping of the factor variable 'birds_struck'

factor	level	mapped.value
1	0	0
2	1	1
3	2-10	5
4	11-100	50
5	Over 100	100



The states with the most collisions seem to be California, Florida, New York and Texas (each approx. around 4000 strucks). This is reasonable since those states are the most known ones and there is a lot of air traffic going on: People travel a lot to and from those states so it makes sense that there are the most collisions.

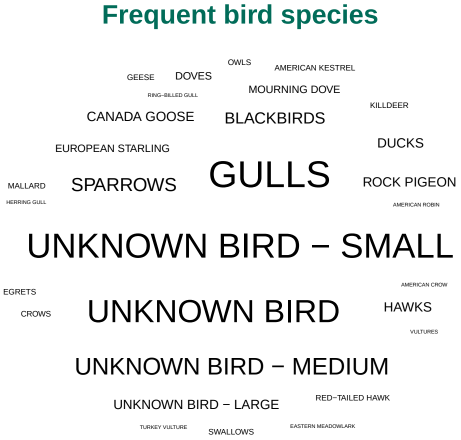
We then also included the variable *ac_mass* to look at:

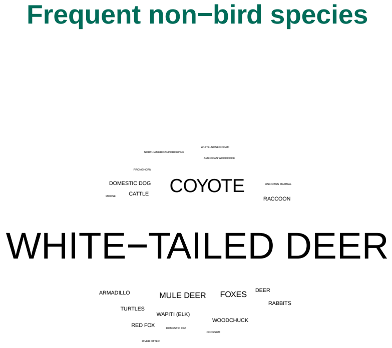


Here, what's interesting to observe is that collisions of aircrafts with a high mass (above 272000 kg) seem to mostly happen in the State of Washington, New York and California. This could be due to the fact that those states are frequently flown to and from, so bigger aircrafts like big passenger planes are used there more than in other states. Another reason could be that there's a big aerospace industry in Washington and the Boeing Everett factory (Aviation: Benefits Beyond Borders 2012). It's said to be the best state for building boeings (WA Governor's office 2018). This would also fit with our findings from earlier where the most frequent aircraft models in the heaviest *ac_mass* category were Boeings.

4.3 Bird collisions vs. non-bird collisions

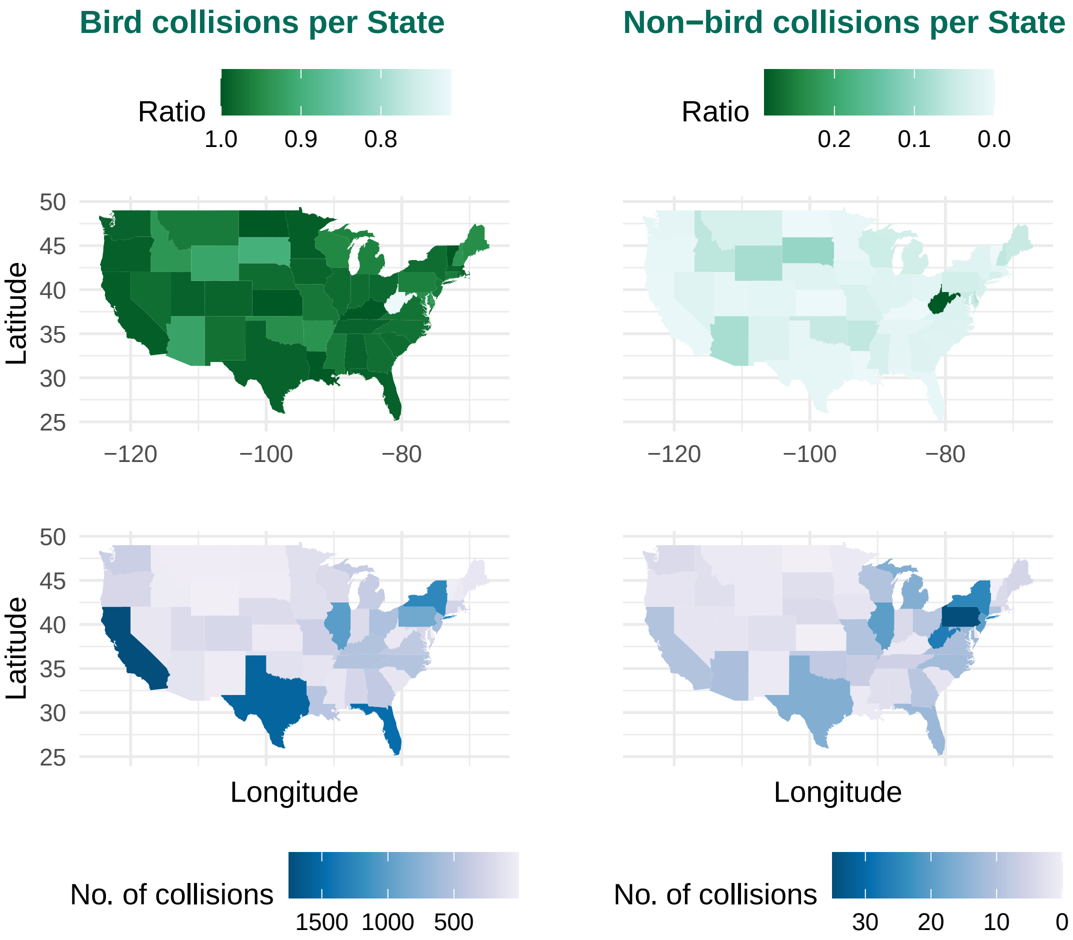
We further investigated by dividing the data into **strucks with birds and strucks with non-birds/other wildlife**. To do this we created another variable called *is.bird*: When "bird" appeared in the *remarks* or the *species* variable (along with other words), or when the height where the collision happened was more than 7 Feet above ground level, we categorized those collisions as bird collisions. The remaining ones were then classified as non-bird ones. In the following we see wordclouds of the bird and non-bird species, the size of the words in regard to their respective frequency in the data. (The number of the species/words have been limited due to size and readability, only the most frequent ones are displayed.)


Frequent bird species


Frequent non−bird species

We again can observe what we already saw earlier in the barplot for the most frequent species in general: The most frequent birds are unknown, but also gulls and a loth of other bird types can be found hee again. For the non-bird wildlife we see the white-tailed deer again which is, by far, the most frequent one in this category (visualized by the size of the word), followed by coyotes, mule deers and foxes.

Next, we want to find out if there's a difference in bird and non-bird collisions depending on what state the aircraft is in. To do this, we displayed the number of collisions and used the *is.bird* variable we created to divide the data. This is visualized in the maps below: The left column shows the bird strikes, the right one the non-bird ones. The upper maps show the ratio, so the proportion of bird/non-bird collisions in relation to total collisions by state. The two bottom maps show the absolute values for bird and non-bird encounters.

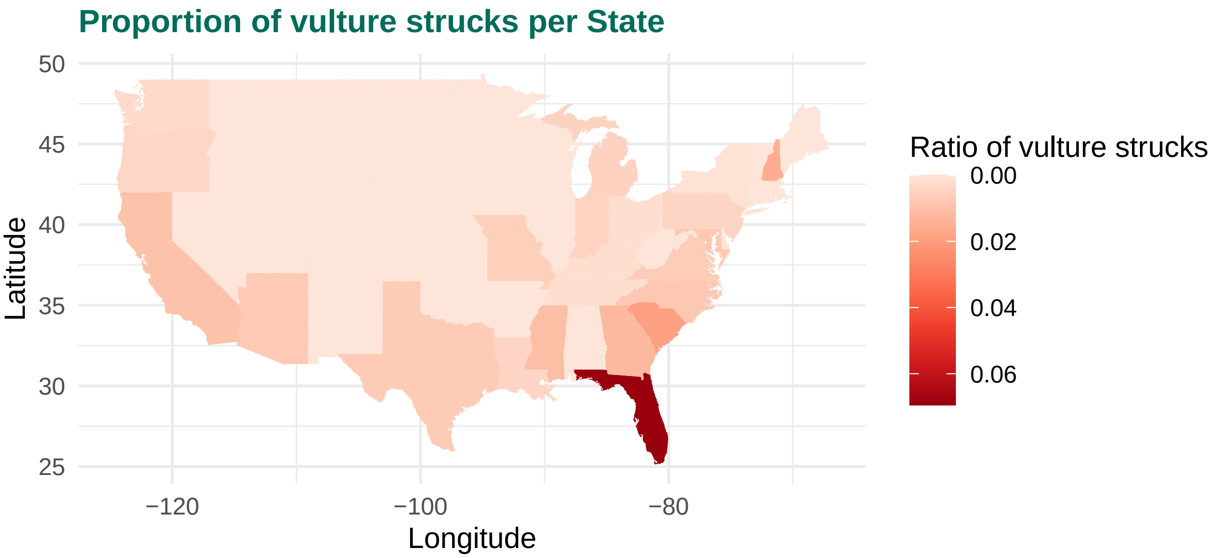


We see that, all in all, the states that have the most bird strikes also have the most non-bird strikes. Though states in the midwest and ones that do not lie on the coastline have a higher ratio of non-bird collisions than those that do.

And, what was already quite obvious in the beginning: There are a lot more aircraft-wildlife collisions with birds than other animals (Over 90%). West Virginia has the lowest ratio of all states with approximately 71.28%. This fits with our findings from earlier where we looked at the most frequent species of that state: The white-tailed deer, which is obviously a non-bird animal, was the most common one during aircraft collisions. Though West Virginia doesn't have the most non-bird collisions in absolute values - this spot is taken by Pennsylvania - it has the most in relation to the total number of collisions of the given state.

4.4 Vulture collisions per State

Lastly, for the state/wildlife relationship, we will look at a specific animal, specifically vultures. We look for all **collisions with vultures** and see which states those happened in. We will look at the proportional values, so the percentage of collisions in a state where vultures were involved in:



Vultures are more frequent in the South and warm regions (Scott 2021b) which fits with our map above. Their habitat is not in the midwest and hence, there aren't any vulture collisions happening there.

So all in all, we can see that there is a relationship between wildlife and the state. Depending on where the aircraft collided with wildlife, the most frequent species differ from each other. We can also extract information about the habitat of certain wildlife with the given data (e.g. vulture example).

5. Timeline - Does time matter?

5.1 The date variable

For the last research question we involved the *date* variable of the dataset. We wanted to investigate the following: Have there been any **changes over time regarding the wildlife collisions** or are there **differences depending on the time of year**?
To do this we first had to perform some feature engineering on *date*: We removed the time part using Regex and formatted it into an actual date variable. We then extracted the month, the year and the weekday into separate variables.

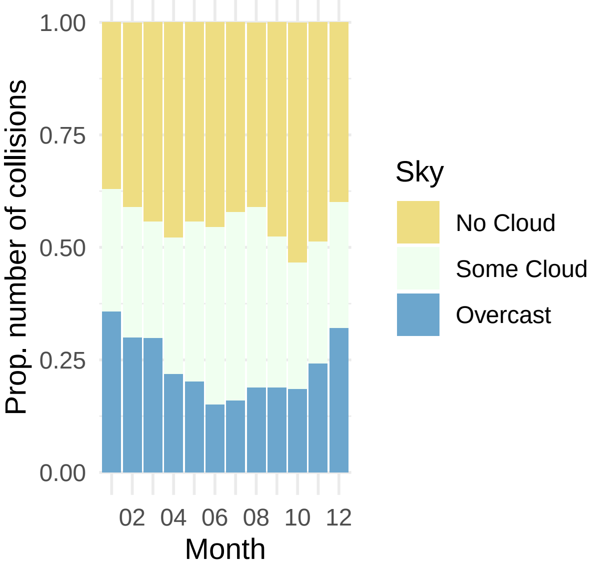
Table 3: Examples for Feature Engineering of the date variable

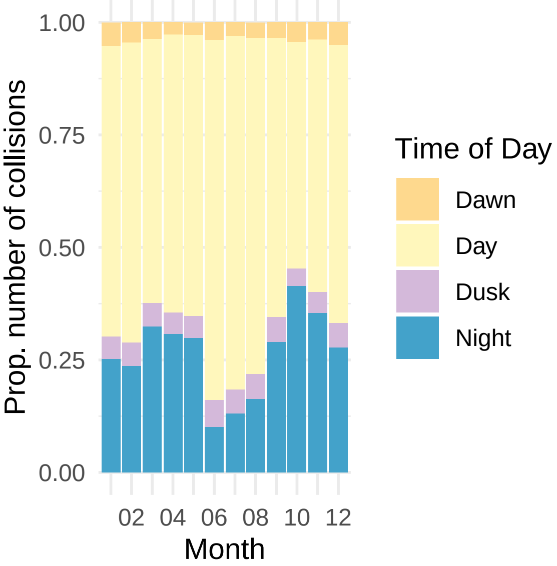
original.date	date	year	month	day	calendar_week
09/30/1990 0:00:00	09/30/1990	1990	09	So	39
11/29/1993 0:00:00	11/29/1993	1993	11	Mo	48

The above table shows the original *date* variable in the far left column, the other columns are the new variables that *date* has been mapped to. We now use those to investigate the data further.

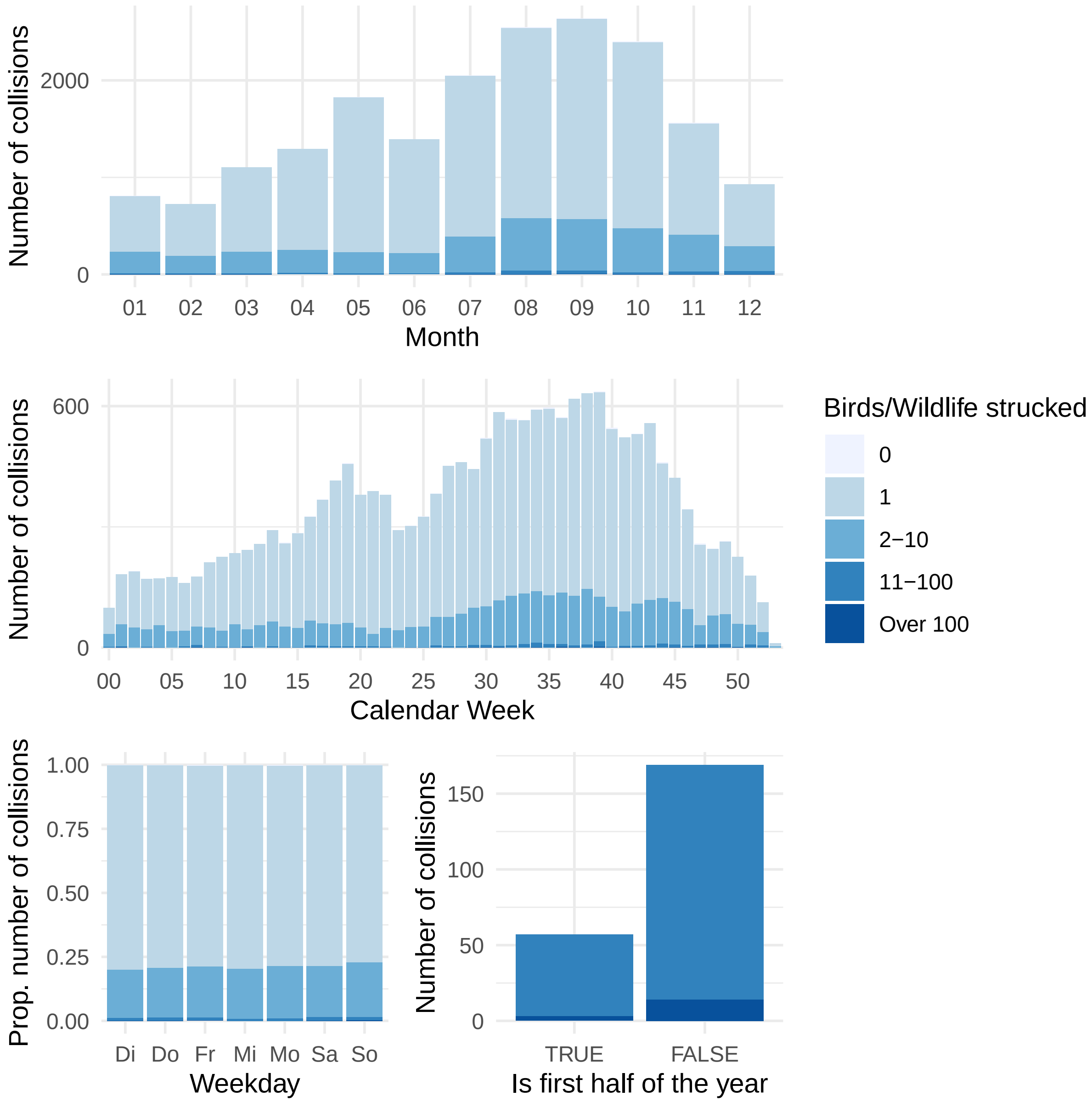
5.2 Monthly and weekly grouping

The number of collisions is used as reference against the different date/time variables and also in regards to the *sky*, the *time_of_day* and the *birds_struck* variable.





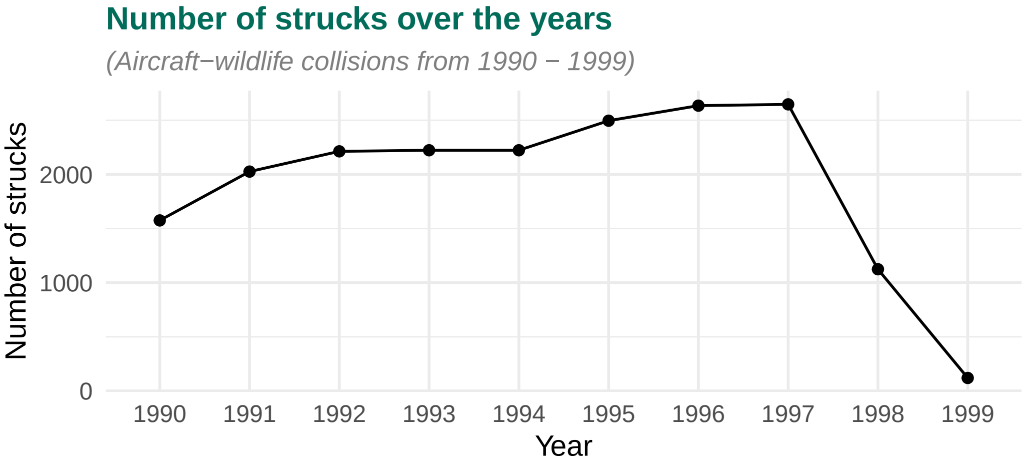
We can observe that the *sky* variable changes depending on the month: There is less overcast during summer months which is reasonable since generally during summer it rains less.
We also see that there are less collisions during the night in the summer. One possible explanation here could be that, at least in the U.S., during summer days are longer and it gets darker later. This would also match with the observation we made jut before with the *sky* variable. There are correspondingly also more collisions during the day in the summer which could also be because there are more flights happening compared to the winter time. Therefore the probability of more collisions is higher then, too.
So just based on this dataset alone we can also see changes in the weather depending on the season.

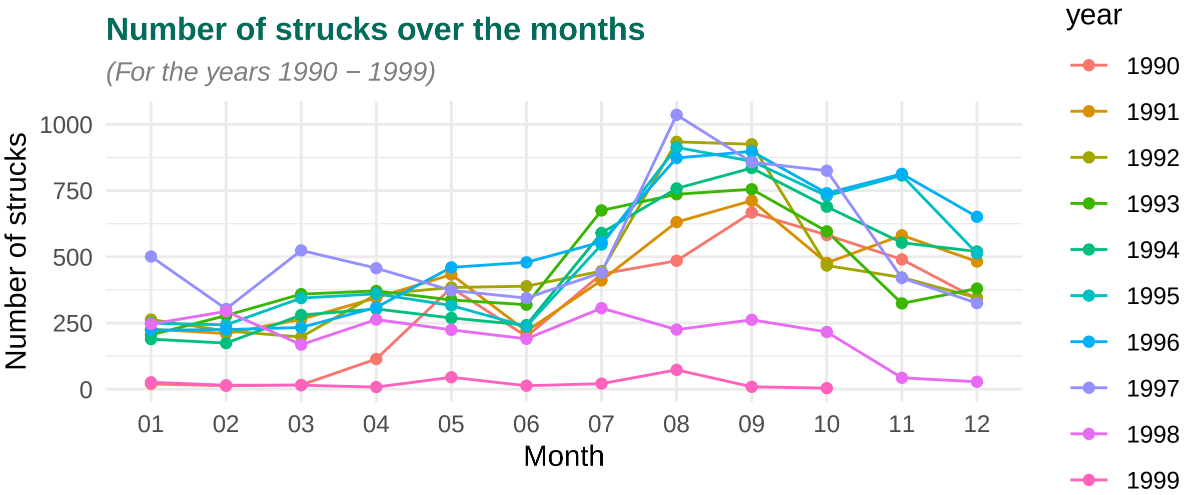


When looking at the number of collisions over the year we can see that there is an increase in collisions during autumn/fall which is especially evident when looking at the collisions where the amount of birds struck was high (where *birds_struck* is “11-100” or “Over 100”). This is displayed in the plot in the bottom right corner: There are way more collisions in the second half of the year than there are in the first half.
This could be connected to the fact that a lot of birds are migratory birds: So once it gets colder they migrate in big flocks which then leads to more collisions with a higher amount of birds during the fall months. However, there is also already an increase during the summer (July, August) and also a slight increase in May, which could again be because of spring migration. For the former, people tend to travel a lot more during summer so this could also explain the increase.
So with the *birds* dataset we are also able to visualize the migratory behavior of birds over the year.
We also examined the proportional number of collisions based on the different weekdays but these seem, as expected, to have no influence. So the number of bird/wildlife collisions is not dependent on the days of the week.

5.3 Grouping per Year

For the **grouping over the years** we will again use the *struck_num* variable from earlier which was based on the *birds_struck* column. We sum up the number of strucked birds/wildlife for each year and display this in a line/point plot. We also do the same again but also divided by each month. Both plots are displayed below.



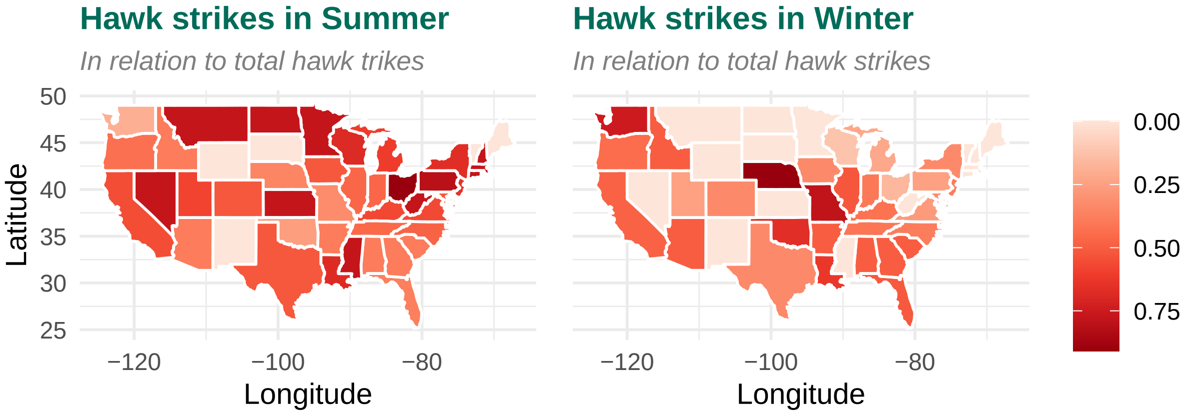


We again see that for almost all years there is an increase in strucked birds in the fall and less collisions in the first half of the year. The reasons for this are the same as above. With each year the number of collisions increases. However, the result for the years "1998" and "1999" are quite low compared to the previous years. After some research we found there was no valid reason for this observation as normally wildlife strikes tend to increase over time like as we have seen with the previous years. Therefore we assume that, at least in this dataset, for the years "1998" and "1999" there are data/reports missing. Especially since the number of collisions for 1999 in total is very close to 0 which is very unlikely to be real. When looking at the documentation of the Birds dataset (OpenIntro 2012) it also says data from 1999 to 1997. So this is another indicator supporting the claim.

Apart from this, our observation fit with general information that can be found with other sources (Wikipedia 2021a): Wildlife strikes increase with each year and within a year there are more collisions during the fall/bird migration months.

5.4 Hawk collisions over Time and by State

Finally, we pick a wildlife species as example: We want to investigate whether we can see changes for **hawk collisions depending on the season (summer/winter)** and also if there are any significant observations when we also look at the **location**.
To do this we look for all entries in the dataset where hawks were involved, so where "hawk" appears in *species*. We then divide the data into two dataframes: One where the month is >= 4 and < 10 (so the summer months from April to September) and the other one with the remaining entries. For each set we look at the *state* variable too see if the number of strikes changes depending on location. (We again use the *struck_num* variable from earlier.) The values shown are the proportional hawk strikes, so e.g. the hawk strikes in summer in relation to the total hawk strikes per state.



We observe that in the northern U.S. states the proportion of hawk strikes during the summer is the highest. Compared to the winter months the ratio of hawk strikes there is the lowest. And the ratio in the Southeast region increases slightly. And overall there seem to be less hawk strikes during the winter. This again can be explained with what was mentioned earlier: Hawks are migratory birds as well (Scott 2021a). During the winter months they travel to the southern regions, hence why there are almost no hawk strikes in the North during winter.

By visualizing our findings in a map and dividing by the date variable we were able to further show the migratory behavior of the birds in the U.S. and how this influences the number of collisions and strikes with aircrafts.

Summary

We can see that we can find out a lot about bird strikes and aircraft collisions based on this dataset.

B-737 class causes the most strikes. It includes aircrafts from Boeing which were developed for short and thin routes. Aircraft models with higher max speed are also able to fly higher. Lighter machines tend to have less engines. With one engine max. speed and max. height is limited to 200 knots and 10000 foot.

Most of the time we do not see wildlife/birds before a collision and if the pilot sees something, he most often hits less amount of birds/wildlife than seen. Collisions mostly occur at daytime and at the beginning and end of a flight in phases Take-Off, Climb, Approach and Landing Roll. The proportion of overcast increases with the number of birds struck, this might be because of birds flying in flocks more often when it is cloudy or due to the lack of vision which prevents evasive maneuvers. The mass of the aircraft has an influence on the amount of birds struck, lighter machines are most likely more agile and smaller in size, thus they can better initiate evasive maneuvers, especially compared to machines with a mass greater 275000 kg, which have an about three times higher proportion of striking over 100 birds. Bird strikes have a higher probability to occur near ground level and the probability of hitting birds is dependent on the phase of the flight. The probability to hit a bird is higher when the aircraft is having an negative angle and looses height.

Most often we do not see birds before a collision and at night its even less likely to see birds. Surprisingly the proportion of wildlife seen increases with cloudiness, this phenomenon needs to be further analyzed. Greater speed and height decreases the probability to see a bird.

If the aircraft is flying parallel to the ground the proportion of pilots striking less birds out of the seen is higher compared to all other flight phases. The greater the number of engines and aircraft mass the less likely less birds are hit out of the birds seen.

The greater the number of birds struck, the more likely that the collision has an effect on the aircraft. The larger the aircraft mass the less likely the collision has an effect on the aircraft. There is less proportion of effect when birds were seen, which is a contradiction to our interpretation of the effect variable.

We saw that states that people travel to and from the most also have the most collisions. And while birds make up most part of the encountered species, there is a small percentage of strikes that involve non-bird wildlife. Depending on the state the species differ.
We were also able to show the habitat of certain wildlife in the U.S. and how this influences the number of collisions and strikes with aircrafts by visualizing this with location data. By taking the time component into account we could also detect that wildlife strikes increase with each year and within a year there are more collisions during the fall/bird migration months. By combining this further with the location data we could also visualize the migratory behavior of the birds.

Outlook

For further data exploration one could consider the following points:

- Use another dataset with definite categorization for distinguishing birds and non-bird wildlife,
- Use a dataset with the number of flights of airlines (in the U.S.) in general to have proper values for the proportions/ratios of flights with and without collisions.
- Look for more information about the flights like start and destination for having direction, length of the flight etc. which could be included in further investigation,
- Compare these results with those of other countries and see if there are any significant differences or similarities or
- for more clean data: Have a closer look at the missing values and see if anything could be further replaced or imputed.
- The meaning of the "effect" variable is unclear, investigate more or contact the publisher of the dataset.

References

Aviation: Benefits Beyond Borders. 2012. "Washington State: The Ultimate Aerospace Cluster." Aviation: Benefits Beyond Borders. https://aviationbenefits.org/case-studies/washington-state-the-ultimate-aerospace-cluster/.

aximaps. 2021. "COLORBREWER 2.0 - Color Advice for Cartography." https://colorbrewer2.org/#.

Data Novia. 2018. "TOP r COLOR PALETTES TO KNOW FOR GREAT DATA VISUALIZATION." https://www.datanovia.com/en/blog/top-r-color-palettes-to-know-for-great-data-visualization/.

E. Le Pennec. 2019. "Ggwordcloud: A Word Cloud Geom for Ggplot2." https://cran.r-project.org/web/packages/ggwordcloud/vignettes/ggwordcloud.html.

ggplot2 tidyverse. 2021. "Polygons from a Reference Map." https://ggplot2.tidyverse.org/reference/geom_map.html.

Melanie Frazier. 2020. "R Color Cheatsheet." https://www.nceas.ucsb.edu/sites/default/files/2020-04/colorPaletteCheatsheet.pdf.

OpenIntro. 2012. "Aircraft-Wildlife Collisions - Documentation." https://www.openintro.org/data/index.php?data=birds.

Scott. 2021a. "16 Types of Hawks Found in the United States! (2021)." Bird Watching HQ. https://birdwatchinghq.com/hawks-in-the-united-states/.

———. 2021b. "The 3 Types of Vultures Found in the United States! (2021)." Bird Watching HQ. https://birdwatchinghq.com/vultures-in-the-united-states/.

WA Governor's office. 2018. "Report: Washington Is Best State for Building Boeing's New Mid-Market Aircraft." https://medium.com/wagovernor/report-washington-is-best-state-for-building-boeings-new-mid-market-aircraft-9e81d2d294c.

Wikipedia. 2021a. "Bird Strike." https://en.wikipedia.org/wiki/Bird_strike.

———. 2021b. "Gull." https://en.wikipedia.org/wiki/Gull#Distribution_and_habitat.

Zhiyi Guo and Fan Wu. 2019. "Different Ways of Plotting u.s. Map in r." https://jtr13.github.io/cc19/different-ways-of-plotting-u-s-map-in-r.html.