MIS 587: US Birdstrike Analysis Report
by Kirsten Fure

Tableau Dashboard Link:
https://public.tableau.com/profile/kirsten.fure#!/vizhome/BirdStrikesAnalysis/Dashboard1

Overview:

The goal of this report is to analyze bird strikes on US flights during the 12-year period of 2000 to 2011. The first thing I did was inspect the data in order to gather high-level insights into what types of data points we were given and the quality of the each attribute.

Data Quality Analysis:

Opening the file in excel, I notice this is a rather large dataset of 99,405 records and 37 columns, so I decide to use DQ Analyzer to inspect the data. Since there are 37 columns, I need to exclude some attributes as candidates for dimensions and assemble a smaller list of possible attributes.

Some of the attributes have a large of amount of NULL values, making the accuracy of those data points too low, so I decide to exclude them as possible dimensions. I also exclude attributes that have a large amount of distinct values or when values do not have any significant differences because it will be difficult to find statistical patterns from these.

Attributes I plan to exclude from dimension possibilities (but keep for detail drill-down):

- Location Freeform en route is 90% null.
- Reported date is 72% NULL.
- Aircraft Flight Number is 66% NULL with 6722 distinct values.
- Pilots warned of birds 59% NULL and Y/N too similar to be significant.
- Speed is 61% NULL and the other values have low percentages.
- When Time HHMM has 49% NULL values and all the other values have low percentages.
- Remarks have over 74k distinct values so I would not find any statistical patterns.
- Other Effects has 575 Distinct Values.
- Feet above ground is 44% NULL values, but 23% are 0 feet. There are many distinct values (other than 0 feet with) low percentages. The "When Phase of Flight" is a more useful categorized attribute for my analysis.
- Conditions Sky is similar to Conditions Precipitation with 50.5% NULL Values, but of remaining records, 24% are No Cloud(s) and 16% Some Cloud(s), so it seems less common in overcast weather. However the Conditions Precipitation attribute seems more useful with higher percentages so I exclude Conditions Sky.

- I decide to also exclude Aircraft Airline Operator. In order to know if birdstrikes are more common with specific carriers, I would need frequency data on all their flights to determine how often birdstrikes occur compared to how many do not occur. With only the dataset in this file, the airlines with the most birdstrikes may be the airlines that have the most overall US flights or have a hub in an area with more birdstrikes. I feel this could be misleading to include at this time.
- Effect Impact to Flight has 5 values: NULL(43%), None(50%), Precautionary Landing (3.7%), Other (1.4%), Aborted Take-off (1.3%), Engine Shut Down (.23%). I am going to assume NULL is equivalent to None in this case and assume they would track any unusual activity. Because of the large number of Null and None values, this field is more useful for drill-down detail than statistical analysis.
- Miles from airport has 48% NULL but 48% are 0 miles, so most bird strikes occur while plane is still at airport. Only 4% of the time are they miles away, but it can happen at any point in the flight. Due to the large number of NULL records, I would rather use When Phase of Flight for statistical analysis which has less null values, however still 38% null.
- Aircraft Make Model – most common was only 5.6% of records.

Location Nearby if en route has 98.7% NULL values, however I notice that the data does not provide one single data point with bird strike location. We have the origin state and the location if the airplane was on route (which is a state). I decided to combine these two in excel, creating a new column "Birdstrike Location State".

Number of human fatalities is 99% NULL values. But, due to the FAA's interest in loss of human life, I spend more time considering these. I decide to assume these NULL values are equivalent to 0 values (no deaths). I feel confident that if there were deaths, they would be tracked in the data. Number of people injured is similar wth 99.9% NULL values. I am going to make the same assumptions with this attribute and I am keeping them both as facts.

All of the Cost attributes have a high amount of NULL or None values (> 93%), but the FAA is interested in cost analysis of birdstrikes. I am assuming the NULL or None values are equivalent to 0 cost (that the birdstrike did not cost any significant damage). However, further inspection of the data shows that there are records with damage that report 0 cost, so the reporting of cost needs to be addressed. It is possible that the numbers in this report could be on the low side.

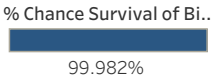
I considered doing calculations with Aircraft Time Out of Service (in hours) in order to get a dollar value. But after some research, I realize this dollar cost relies on many factors, such as plane type, number of seats, etc. So, I decide to keep it simple and leave it in hours.

Data Analysis in Tableau:

Because loss of human life seems the most concerning, I begin my analysis there. I learn that when a plane is en route or climbing and a bird strike occurs, it is the most dangerous!

When: Phas..	
Null	0
Approach	0
Climb	5
Descent	
En Route	12
Landing	
Landing Roll	1
Parked	0
Take-off run	0
Taxi	0

Overall Survival Probability in a Birdstrike

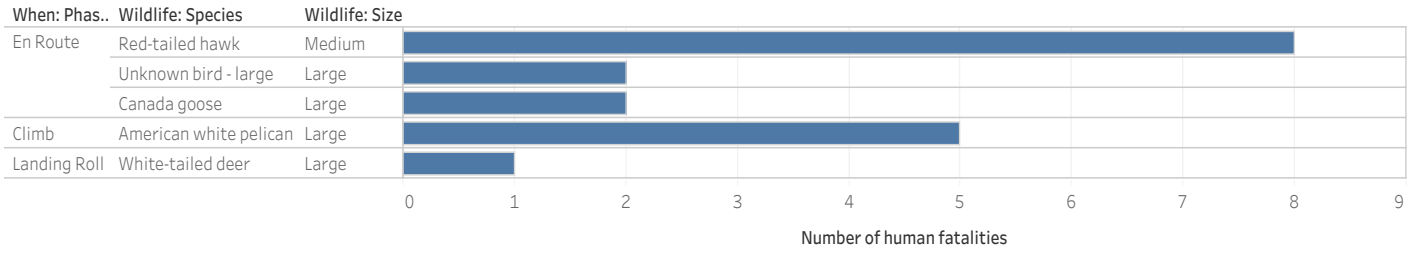


% Chance Survival of Birdstrike. Color shows % Chance Survival of Birdstrike. The marks are labeled by % Chance Survival of Birdstrike.

I am curious how dangerous bird strikes are, so I create a calculation field and find out that the chance of survival in the event of a birdstrike is pretty high (99.982%). That's great news! I notice that flights with fatalities do not seem to have a higher cost, so I am assuming that this dataset is lacking the $ cost of fatalities. I would think there would be legal costs or insurance ramifications.

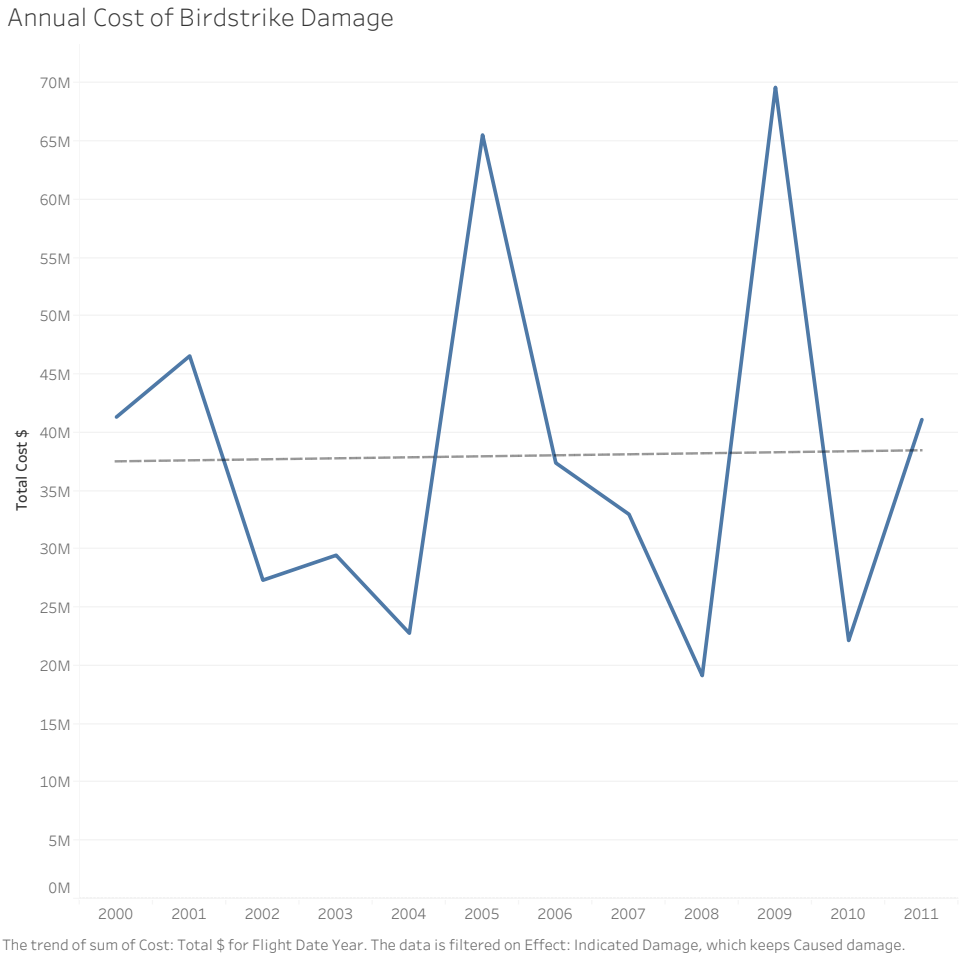
I check to see if there is a certain bird associated with these fatalities, but there isn't a trend. There were 5 flights with fatalities and they all had a different species: Canada goose, Large Unknown bird, Red-tailed hawk, American white pelican and a White-tailed deer! The only thing in common was that they were either a large or medium sized species.

Sheet 18

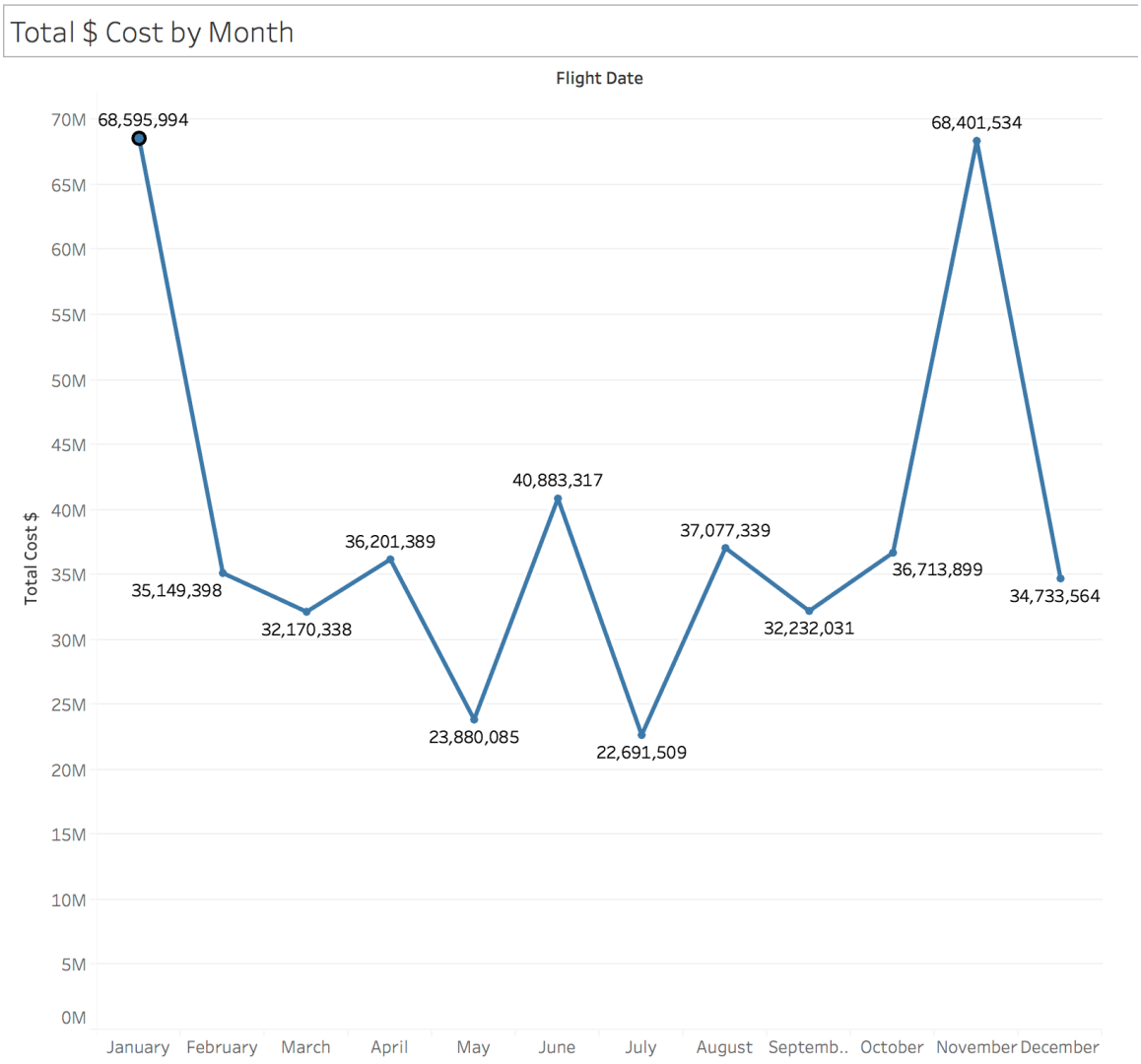


Sum of Number of human fatalities for each Wildlife: Size broken down by When: Phase of flight and Wildlife: Species. Details are shown for Wildlife: Number struck. The data is filtered on Number of human fatalities, which includes values greater than or equal to 1.

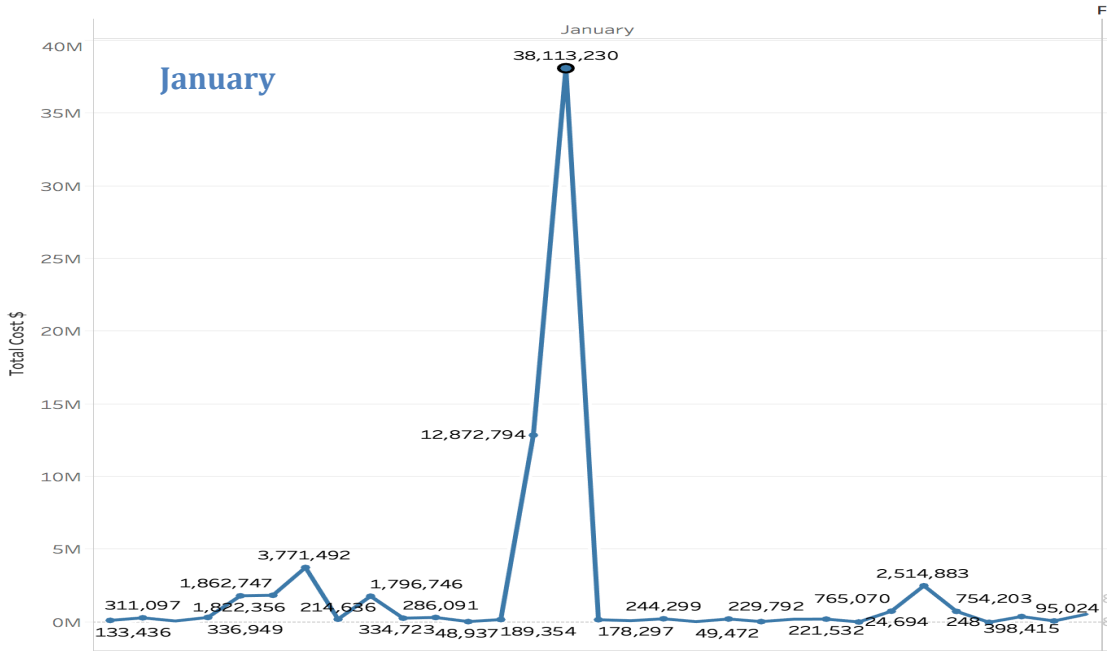
Next, I look at annual cost due to birdstrikes. They average about $38M per year, but can be as low as $19M and as high as $70M in a single year.



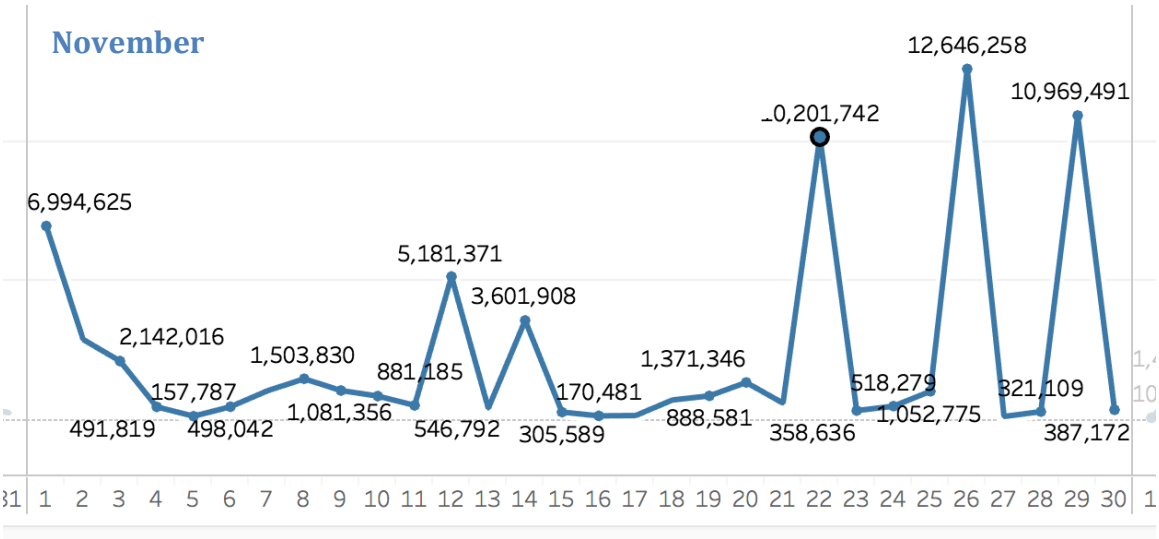
I look at cost over time to see if there are trends. I compared by year and by quarter per year. I found that overall, January and November have been the most expensive months.

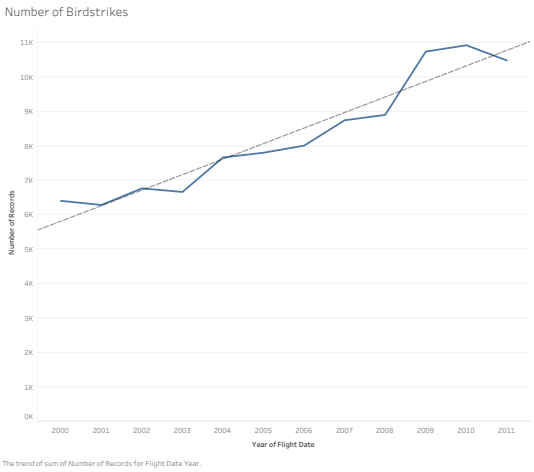


This is not a trend for January year after year though. Digging into the records, I found that there was one very costly $38M birdstrike incident on January 15, 2009 on the Hudson River in New York.

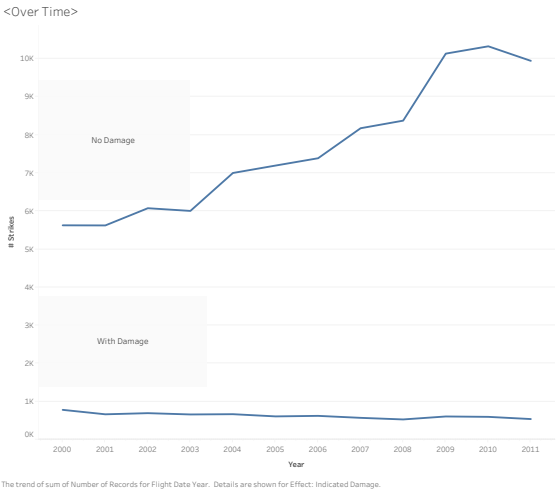


But, November had several costly incidents. So it seems that November is more costly over time.



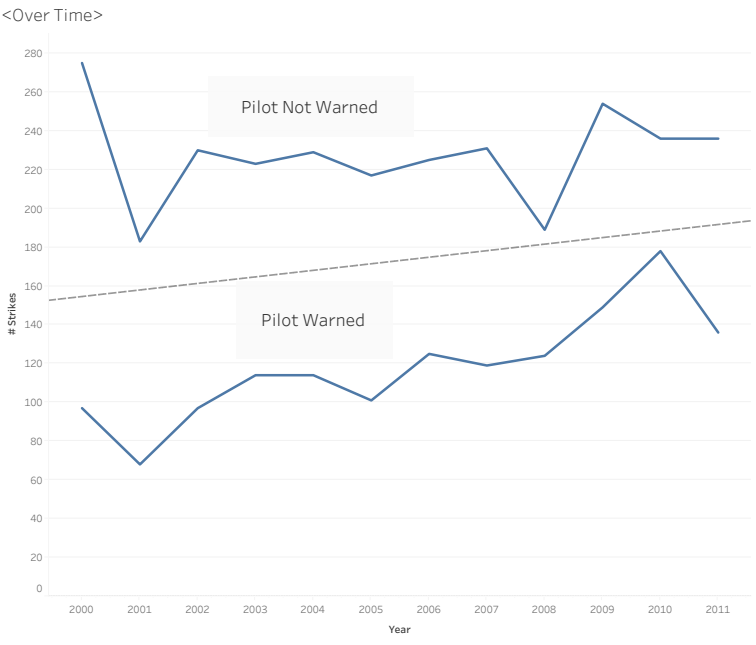


After looking at cost over time, I am curious if the frequency of birdstrikes has a pattern over time. It does increase over time! Why would that be? So I decide to separate out birdstrikes that had damage vs. those that did not.

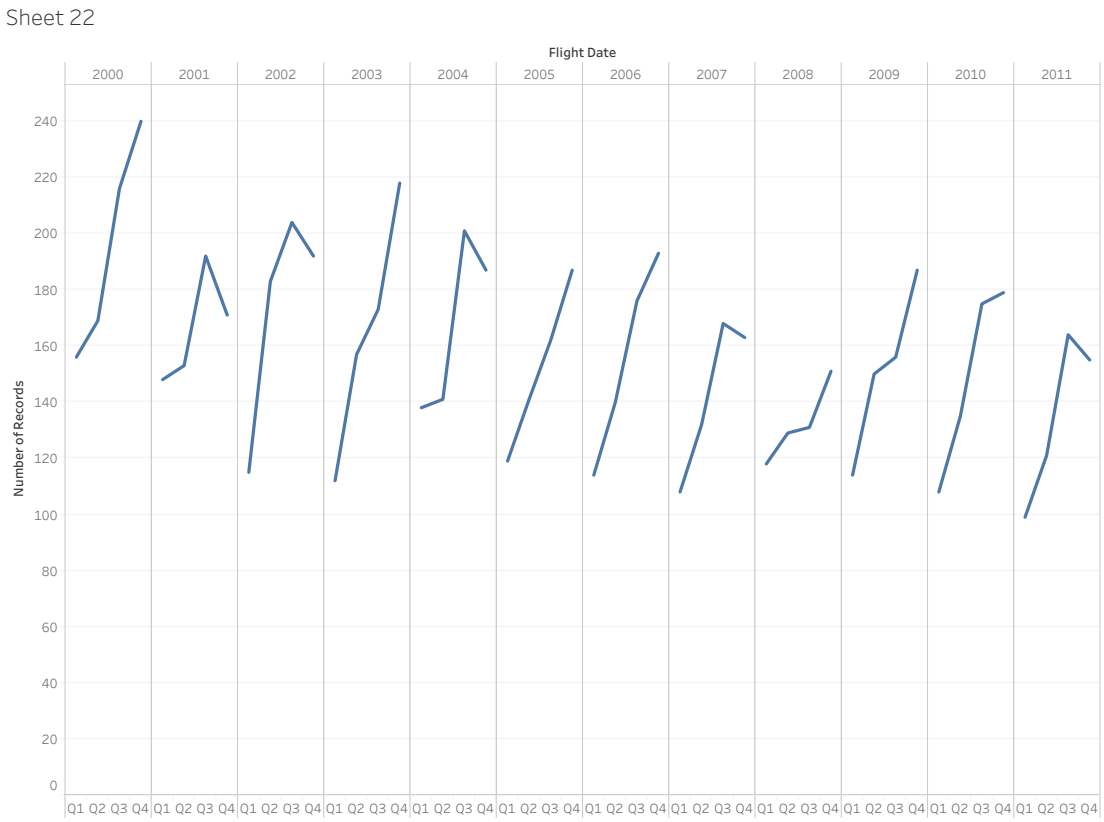


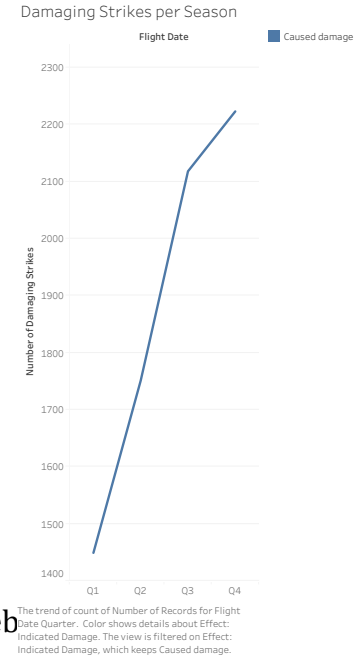
And the data shows that the reporting of birdstrikes did go up, which is a good thing. With more reporting, we can do more analysis to help understand the phenomenon. But, the number of damaging strikes has actually gone down over time by almost 30%. I am wondering if warning pilots has contributed to the drop in damage counts.

Surprisingly, the data shows the the number of damaging incidents when pilots were warned has increased over time.



I am curious if the frequency of birdstrikes correlates with the season, like how we noticed that November has a higher cost. Because we know that the reporting of birdstrikes has increased over time, I limit this analysis only to the strikes that caused damage. And I map them by Quarter, year after year. I definitely notice a trend. The frequency of birdstrikes tends to increase from Q1 to Q2 to Q3 to Q4.

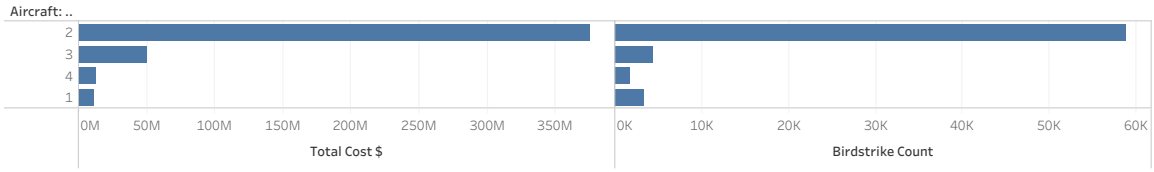




I remove the year and map the sum of birdstrikes per quarter overall (for all years) and the pattern definitely trends up from Q1 to Q4. Interesting! So, the end of the year has the most frequent damaging birdstrikes. This would make sense with bird migration in the winter.

I looked at number of engines and noticed that airplanes with 2 engines have a much higher cost due to birdstrikes, but I also noticed that they correlate with a much larger number of birdstrikes overall. So, dual engine planes are probably the most common and being flown the most, so these numbers are not statistically significant.

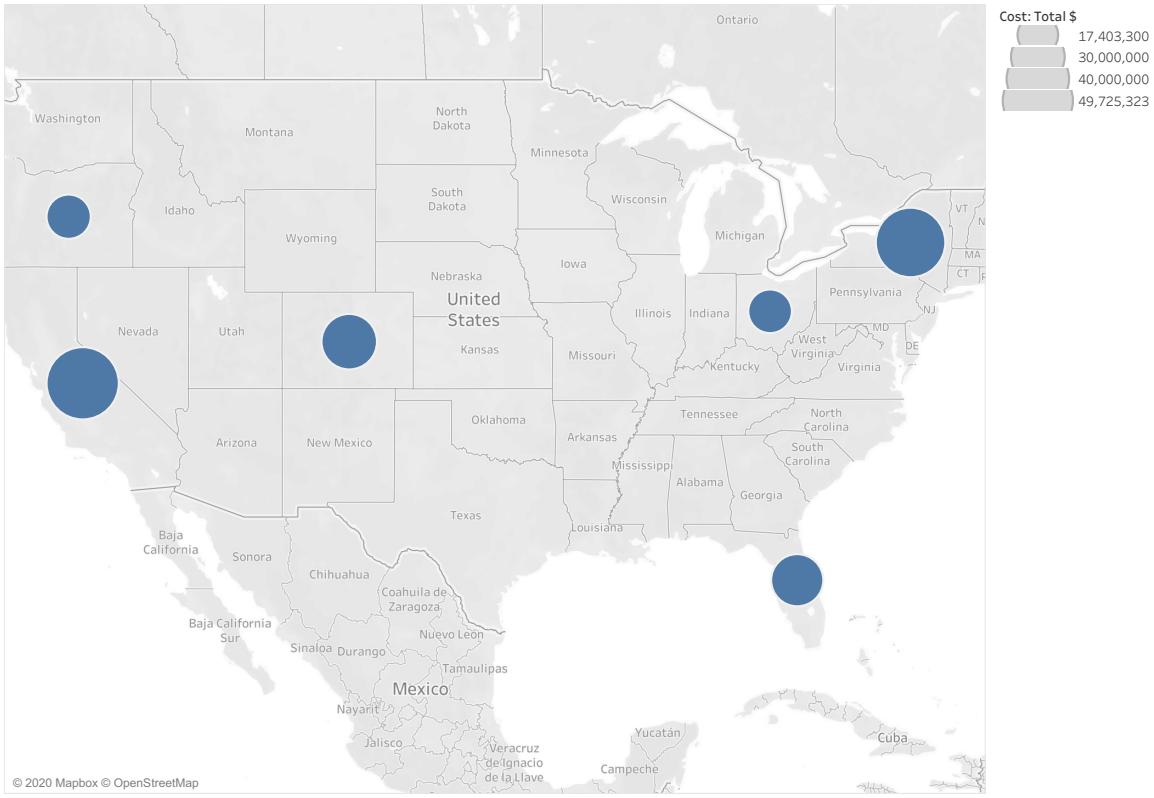
Number of Engines



Sum of Cost: Total $ and sum of Number of Records for each Aircraft: Number of engines?. The view is filtered on Aircraft: Number of engines?, which keeps 1, 2, 3 and 4.

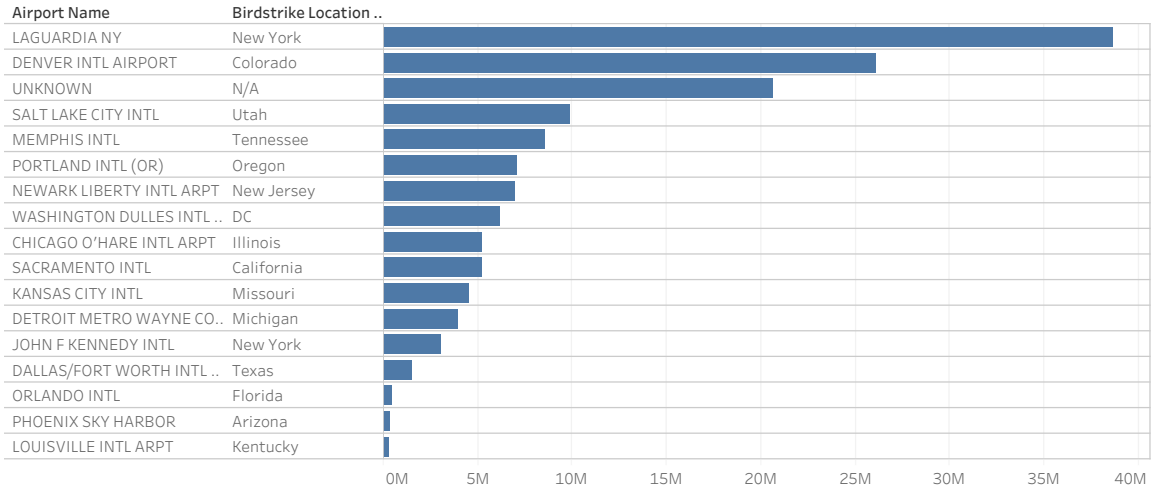
I am curious if certain states have a higher cost due to birdstrikes, and California has the highest cost. However, there are 3 large airports in California.

Most Costly $ Birdstrike States



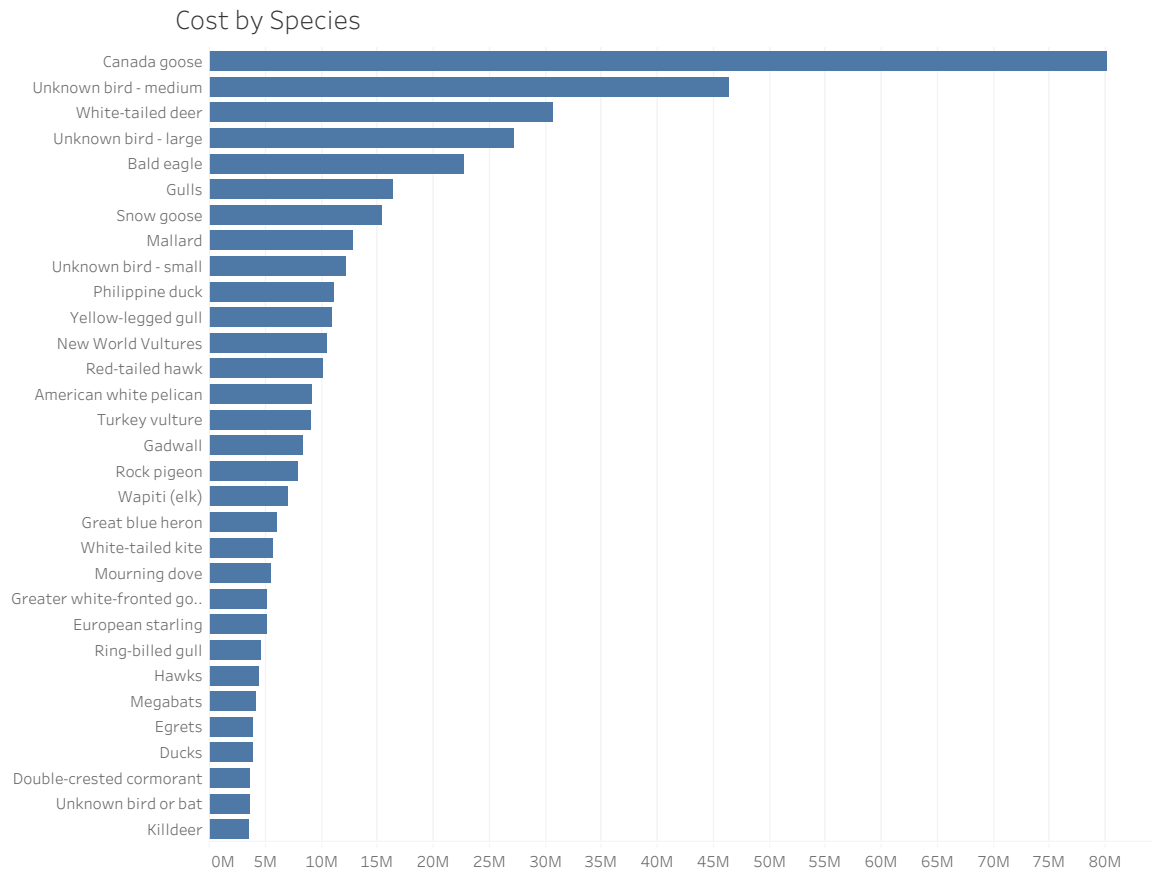
Map based on Longitude (generated) and Latitude (generated). Size shows sum of Cost: Total $. Details are shown for Birdstrike Location State. The view is filtered on sum of Cost: Total $ and Birdstrike Location State. The sum of Cost: Total $ filter includes values greater than or equal to 14,755,797. The Birdstrike Location State filter keeps 65 of 87 members.

If we look by airport, New York's LaGuardia Airport has the highest cost overall. I also limited this list by frequency of birdstrikes. For this list, I am curious about costly airports that continue to have frequent birdstrikes.



Sum of Cost: Total $ for each Birdstrike Location State broken down by Airport Name. The view is filtered on sum of Number of Records, which includes values greater than or equal to 947.

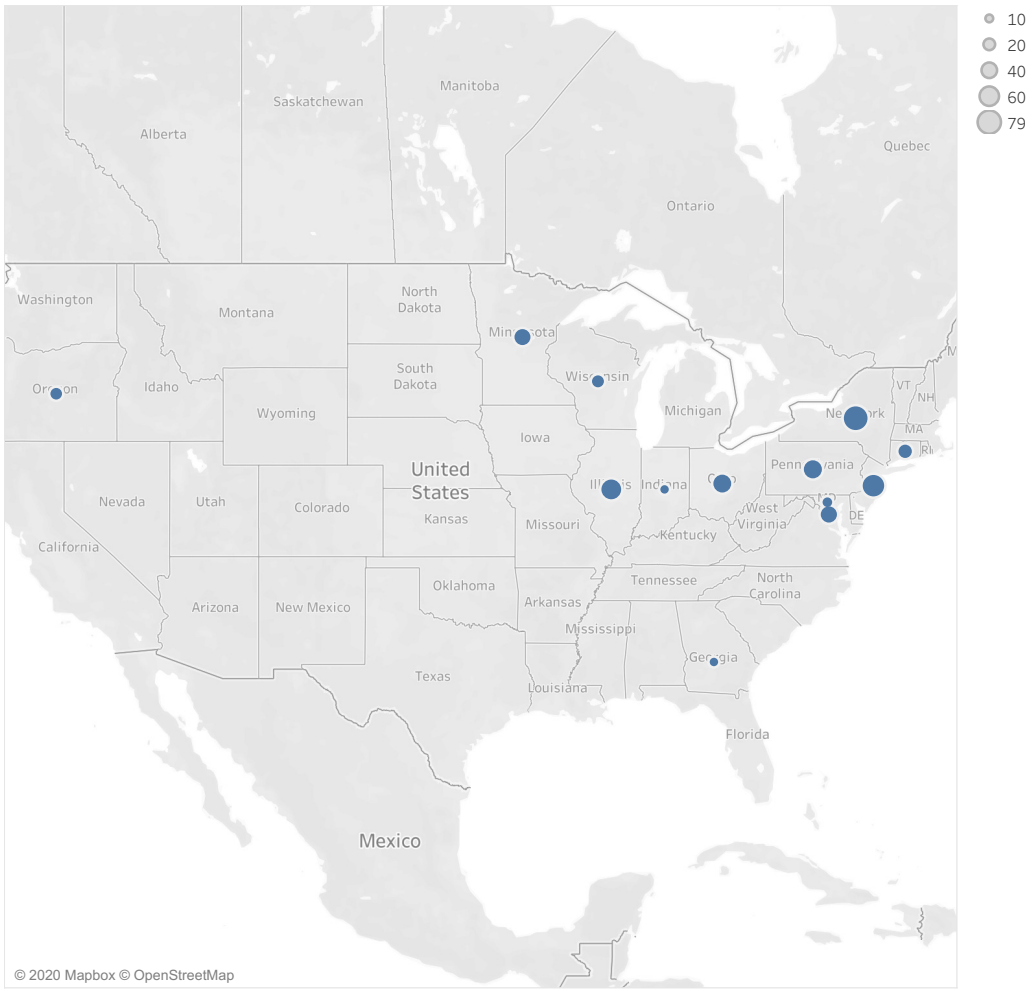
I see that New York, Colorado and California have costly birdstrikes. I am curious if there are certain birds causing these problems, so I look at the cost by wildlife species. The Canada goose is by far the most expensive problem! Also interesting is that the white-tailed deer is the third most costly. I would imagine that if a plane and a deer collide on the runway, there would be damage! Good thing deer don't fly!



Sum of Cost: Total $ for each Wildlife: Species. The view is filtered on sum of Cost: Total $, which includes values greater than or equal to 3,481,773.

I am curious which states have the biggest problems with the Canada Goose so I map by location and limit the data to only strikes related to the Canada Goose specifically. New York and New Jersey have the highest damage costs associated with this bird.

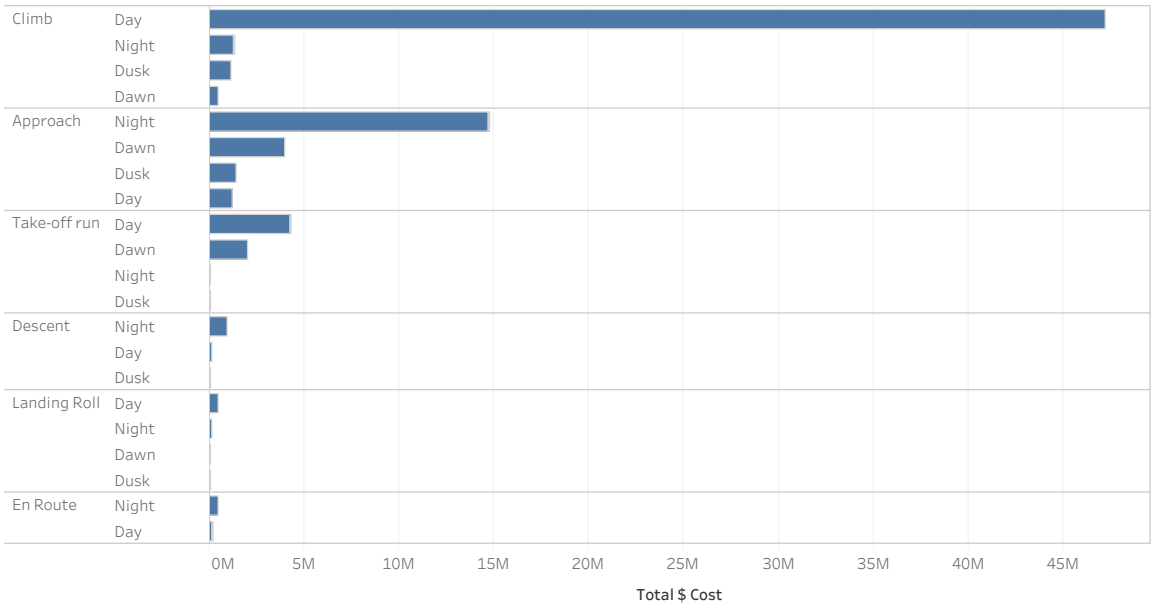
Problem States - Canada goose



Map based on Longitude (generated) and Latitude (generated). Size shows sum of Number of Records. Details are shown for Birdstrike Location State. The data is filtered on Wildlife: Species, which keeps Canada goose. The view is filtered on sum of Cost: Total $, which includes values greater than or equal to $569,643.

I am curious if there any trends about the Canada Goose that I can find. I map data specifically for the Canada Goose, flight phase and time of day. I find out that these geese are most problematic during a plane's climb during the day.

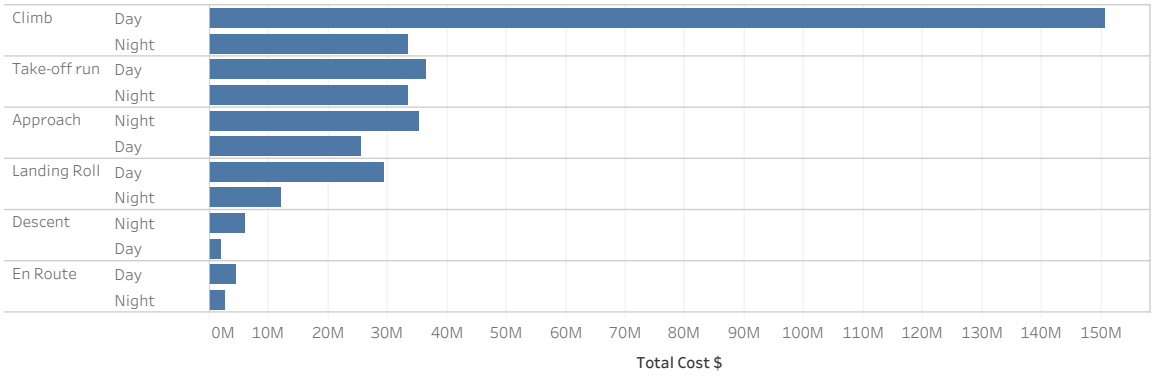
Canadian Geese during a Day Climb



Sum of Cost: Total $ for each When: Time of day broken down by When: Phase of flight. Details are shown for Wildlife: Species and Wildlife: Size. The view is filtered on Wildlife: Species, When: Phase of flight and When: Time of day. The Wildlife: Species filter keeps Canada goose. The When: Phase of flight filter keeps 8 of 8 members. The When: Time of day filter keeps Dawn, Day, Dusk, Night and UNKNOWN.

I run this analogy with all species and it turns out this is the most costly combo in general (climb during the day).

Most Costly Conditions



Sum of Cost: Total $ for each When: Time of day broken down by When: Phase of flight. The data is filtered on Wildlife: Species, which keeps 609 of 609 members. The view is filtered on When: Phase of flight and When: Time of day. The When: Phase of flight filter keeps 6 of 8 members. The When: Time of day filter keeps Day and Night.

I am curious if some airports are doing a better job of reporting birdstrikes. But I found that the number one airport reporting birdstrikes is UNKNOWN. Many other fields were null or none, so hopefully the FAA can work with aviation personnel to capture more accurate and detailed data to help improve costs and safety due to birdstrikes.

For my dashboard, I ultimately focused on providing visualizations that present data in 3 main areas: fatalities, cost and frequency of damaging birdstrikes. For fatalities, I decided to provide 2 tools, one that gives the number of deaths and associated flight phases and the other the overall survival rate in the event of a birdstrike. I feel these both give the most necessary complementary data in a simple way. For cost, I provided a graph of overall costs per year related to birdstrikes and show a trend line. This can alert personnel to any major trends up or down over time. To complement the annual cost data, I gave 2 tools to analyze cost in more detail, one by airport and the other by wildlife species. For frequency of damaging birdstrikes, I chose a graph showing damage strike counts and how it relates to the season. Lastly, I provide a map widget where the user can explore interactively by certain species, a combination of species or all species. They can limit the data to the top costly states and hover over them to get more detailed data. In all of the graphs in my dashboard, the user has the ability to drill-down into detailed records.