**Tornadoes and Bird Detection**

Halyna Tymofeyev

Pierina Logroño

IST652 6/11/2023

Tornadoes are violently spinning, funnel-shaped columns of air that stretch from the dark thunderclouds that form in all the way to the ground. The wind from a tornado can top up to 250 miles an hour and can rip apart buildings, destroy bridges, flip trains and send cars flying. Tornadoes can occur anywhere in the world, but the United States has had the strongest and most significant number of tornadoes, with about a thousand twisters a year, in all shapes, sizes, and speeds. In the United States, most Tornadoes form in what is known as Tornado Alley, which is located in the central region of the U.S., extending from Texas through Oklahoma, Kansas, Nebraska, North and South Dakota, Iowa, Minnesota, Wisconsin, Illinois, Indiana, Missouri, Arkansas, Montana, and Ohio and winds can even reach the Canadian Prairies. Tornadoes typically develop with Thunderstorms, when cold, dry air is pushed over warm, humid air. The warm air then rises through the colder air, causing an updraft or change in wind direction. If winds from the thunderstorm vary significantly in speed or direction, the updraft will begin to rotate. As the rotating updraft draws in more warm air from a moving thunderstorm, its rotation speed increases, and a funnel cloud begins to form. As the twister gains strength, the funnel becomes longer. It is most dangerous when it touches the ground.1 To track these tornadoes, meteorologists use weather satellites, weather balloons, and buoys. These tools gather real-time data, like wind speed and temperature, to later analyze with supercomputers. Scientists have been researching efficient ways of tracking tornadoes to inform people within the areas to take cover quickly.

As humans use advanced technology to track tornadoes and major storms, the mystery of whether birds can predict these disasters have also created an outburst in research. In 2013, an article, "How do Tornadoes Affect Birds?" was published in Audubon.org, discussing how birds leave their nesting areas when they detect unusual weather patterns and return. The author also discusses how another author believes birds are the worst predictors, and they happen to move fast. Over the years, however, more research has been done, and there may be more proof that birds can accurately predict extreme weather. One article on Audubon.org, "Are These Birds Better Than Computers at Predicting Hurricane Seasons?" discusses how an associate professor at Delaware State University observed and tracked nesting and migration patterns on Veeries. He believes that these birds are better predictors of major hurricanes than meteorologists!

**Data description and sources:**

For this analysis, two datasets are to will be used from kaggle.com. The primary dataset is [U.S. Tornado Dataset 1950-2021](https://www.kaggle.com/datasets/danbraswell/us-tornado-dataset-1950-2021), and it comes in the form of a CSV file with 67,558 rows and 14 columns. NOAA's Storm Prediction Center produced the original data. However, some changes have been made, such as column deletions, changes in data types, and sorting by date.

**Column Definitions:**

* **yr**- 4-digit year
* **mn**- Month (1-12)
* **dy**- Day of the month
* **date**- Datetime object (e.g., 1950-01-01)
* **st**- State where tornado originated; 2-digit abbreviation
* **mag**- F rating thru Jan 2007; E.F. rating after Jan 2007 (-9 if unknown rating)
* **inj**- Number of injuries
* **fat**- Number of fatalities
* **slat**- Starting latitude in decimal degrees
* **slon**- Starting longitude in decimal degrees
* **elat**- Ending latitude in decimal degrees (value of 0 if missing)
* **elon**- Ending longitude in decimal degrees (value of 0 if missing)
* **len**- Length of track in miles
* **wid**- Width in yards

It is important to note that since February 2007, tornadoes have been rated on the enhanced Fujita Scale:

## The Enhanced Fujita Scale

| **Rating** | **Wind Speed** | **Damage** |
| --- | --- | --- |
| EF0 | 65–85 mph | Light damage |
| EF1 | 86–110 mph | Moderate damage |
| EF2 | 111–135 mph | Considerable damage |
| EF3 | 136–165 mph | Severe damage |
| EF4 | 166–200 mph | Devastating damage |
| EF5 | >200 mph | Incredible damage |
|  |  |  |

The secondary data is [Bird Songs Recordings from the United States](https://www.kaggle.com/datasets/gpreda/bird-songs-recordings-from-united-states), also found on Kaggle in the form of a CSV file. This dataset contains 26 columns and about 53,300 rows. The date range is 2008 to 2020. This dataset is derived from <https://xeno-canto.org/contributors> which is open to the public to contribute recordings of wildlife from anywhere in the world.

**Column Definitions:**

* **id –** Bird identification number
* **gen –** Genus of the bird
* **sp -** Species
* **ssp -** Subspecies
* **en –** Bird type encoding/conservation status
* **rec –** Recorder name
* **cnt -** Country
* **loc -** Location
* **lat -** Latitude
* **lng -** Longitude
* **alt –** Altitude of the location
* **type –** Type of bird sound file
* **url -** The URL where the recording can be accessed.
* **file -** The name of the file.
* **file-name -** The full name of the file.
* **sono -** The unique identification number of the recording.
* **Lic -** The license under which the recording is made available.
* **q –** Score of the quality of the recording
* **length –** Length of the recording
* **time –** Time the recording was made
* **date -** Date the recording was made
* **uploaded -** date when the recording was uploaded
* **also -** Additional observations or recordings were made at the exact location and time.
* **rmk -** Remarks or notes about the observation or recording.
* **bird-seen –** Whether the bird was seen or only heard. (yes/no/unknown)
* **playback-used -** Whether playback was used to elicit the bird's response during the observation or recording.

**Method of Analysis:**

* Examine the data, prepare the data, and examine the patterns in the data.
* Identify trends or seasonal patterns in tornado activity.
* Calculate a statistical summary such as mean, median, and range of tornado characteristics by year or month.
* Create a pattern based on the geographic location to see the severity of tornadoes in a particular region based on the U.S. map.
* By regression analysis, we can identify if the specific type of tornado had the most fatalities by looking at the slat, slon, len, and wid to see if they affect fatalities or injuries.
* Merge the bird calls data with tornado data and provide geospatial analysis to visualize if there are patterns in the bird calls and tornado frequencies as well as the bird species types and the tornado frequencies.
* Correlation analysis will be used to determine if there is a correlation or no correlation between tornadoes and bird calls.

We expect most tornado occurrences to be in the central states of the USA and with most fatalities and injuries. We expect to see fewer bird callings before tornado occurrences and more bird callings after or when there are no tornado occurrences in the specific region.

**Data Questions:**

* Has the magnitude of tornadoes worsened over the years?
* Which tornado had the most fatalities?
* Which tornado had the most injuries?
* In which locations are tornadoes with the highest magnitude?
* In which states tornadoes are the widest?
* In which states Tornadoes are most frequent?
* Are bird calls detected before, during, or after tornados hit specific regions?
* Is there a specific bird species that can predict tornadoes? (birds heard=least tornadoes happened)

**Program Description:**

For the data preparation and data analysis following packages were used:

* Pandas - for data manipulation and analysis.
* CSV - to read and write CSV data in a tabular format
* matplotlib.pyplot - for creating plots and visualizations
* folium - for creating interactive maps and visualizations
* folium.plugins - subpackage of the folium library for creating advanced map visualizations using folium.
* Branca.colormap - module within the branch library used for creating custom color maps and color scales.
* sklearn.linear\_model - for implementing linear regression models.
* stats models. API - or performing various statistical analyses
* seaborn - for creating attractive and informative statistical graphics.
* re - for work with patterns and perform pattern matching
* us - for use functionalities related to United States geography and data.
* plotly.graph\_objects - for interactive and highly customizable visualizations.

Also, for both datasets, data cleaning was used by removing the missing values, converting the date column to datetime format, filtering, sorting, and mapping was performed. In the bird's dataset, the state column was abbreviated to be merged with the tornado dataset, and statistical analysis was done.

For the Birds dataset, additional cleaning was performed, such removed missing values, changing the date column to DateTime format, # Splitting the "loc" column by the last comma into "loc" and "state" in order to derive a state column separately, then the state abbreviation was performed to match the state column of tornado datasetfor merging.Filtering data by the United States was performed to focus only on the U.S. territory.

**Analyzed Results/Outputs:**

In this program, we started by converting the tornadoes dataset into a pandas data frame after importing packages.

**A screenshot of a computer

Description automatically generated**

We then check for null values within the dataset and convert the date column to DateTime format.

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

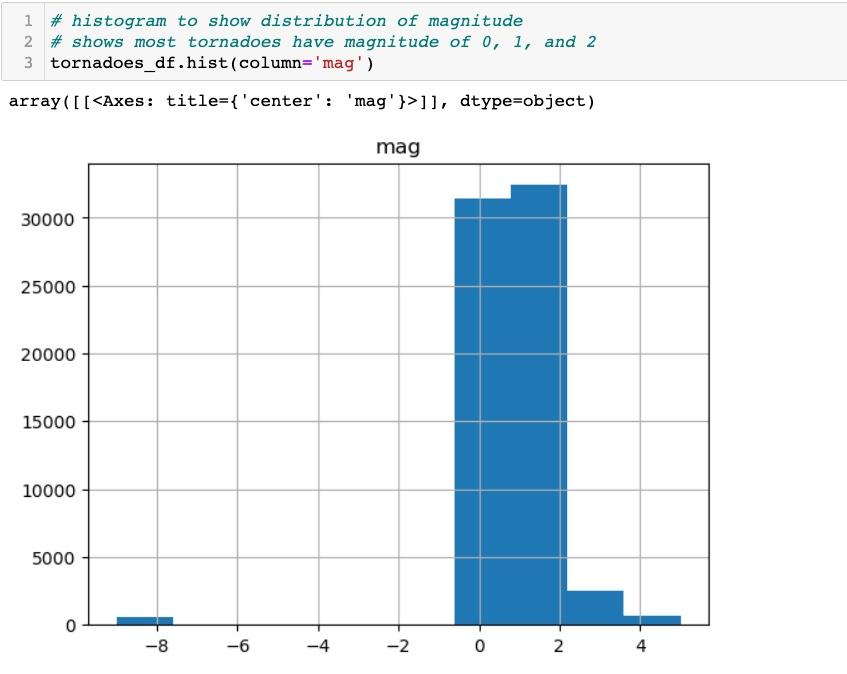
To obtain descriptive statistics of the tornado dataset, we use the describe() function. This shows us how this dataset is skewed, considering most tornadoes were of low magnitudes, with no injuries or fatalities. Also, the length in miles for most tornadoes is less than 50, and the width for most is about 500 yards or less. To better understand the distribution for each variable, we also constructed histograms for each.

A screenshot of a computer

Description automatically generated

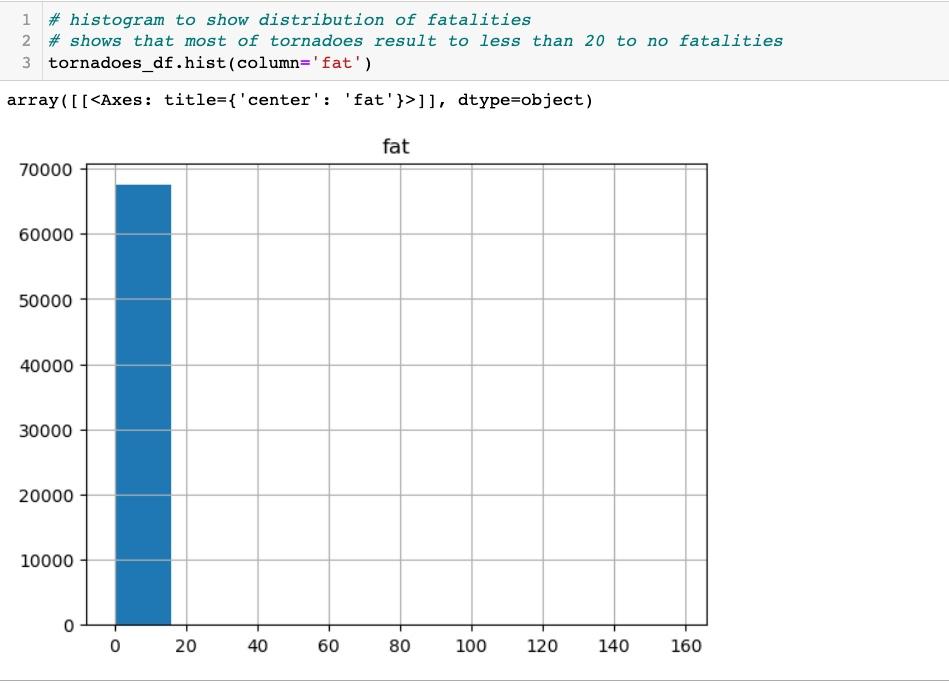
A table of numbers with numbers

Description automatically generated



**A picture containing text, screenshot, line, number

Description automatically generated**

****

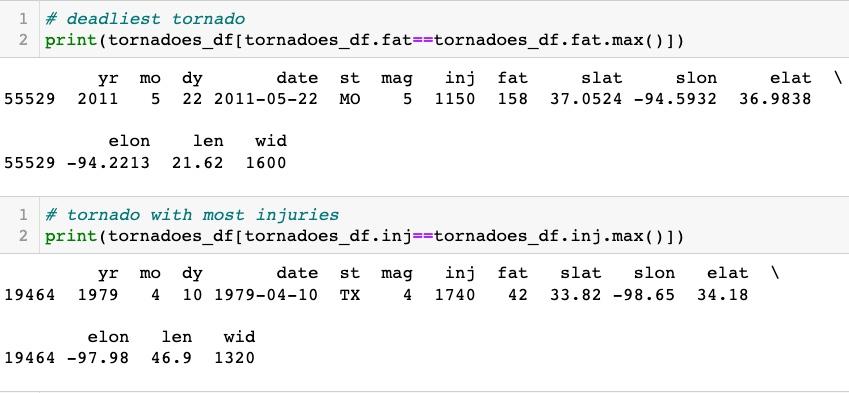
**A picture containing text, screenshot, line, plot

Description automatically generated**

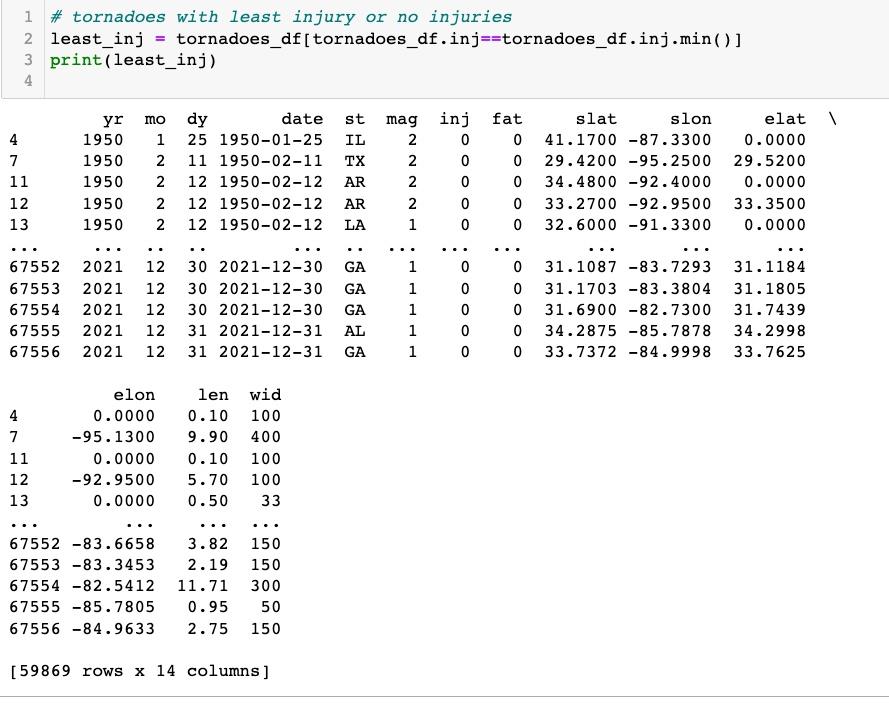
**A picture containing text, screenshot, line, plot

Description automatically generated**

To find the deadliest tornado and the tornado with the most injuries, we use the max() function and apply it to those columns.

****

To see tornadoes with the least number of injuries- in this case, 0-we apply the min() function to the injuries column. This shows us that there are 59,869 tornadoes with no injuries in the dataset.



To see what years these tornadoes took place and their magnitudes, we constructed another histogram in which the hue of the bars is based on their magnitudes.

A picture containing text, screenshot, plot

Description automatically generated

Next, we look at tornadoes with the highest magnitude of 5. These were the tornadoes in the dataset that had the most injuries. There is a total of 59 tornadoes that had magnitude 5.

\*\*\*This shows just the beginning and end of the output\*\*\*A screenshot of a computer

Description automatically generated

**A screenshot of a computer

Description automatically generated**

To visualize the highest magnitude tornadoes (in this case, is tornadoes of magnitude 5) we first aggregate the data to group by the year and state to get the counts for each year, and state. Then, we create pivots so we can plot the aggregated magnitude 5 tornadoes in a stacked bar graph by year in which the different colors represent the states. Here we can see that the most magnitude 5 tornadoes occurred in 1974—7 total occurrences. 3 of them took place in Alabama, 2 in Indiana and 1 in Kentucky and Ohio.

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer screen

Description automatically generated**

Next, we look at which state had the most tornadoes, and which had the least. Using the group by() and count() function on the “st" column gives us the counts of each state. Since we assigned the result to "state," we applied the max() function to obtain the state with the most tornadoes. We then apply the min() function to obtain the state with the least. In the dataset, Texas had the most tornadoes, 9,149 in total, and "VI," representing the Virgin Islands, had the least number of tornadoes, with one within this dataset.

A screenshot of a computer

Description automatically generated with medium confidence

We also graphed the Count of Tornadoes by State and Magnitude, where we have states on the X-axis and count of tornadoes on the Y-axis, and colors are used to differentiate amongst magnitudes. We can see most tornadoes took place in Texas and nearly 5,000 of them were of a magnitude of 0. We followed the same process as the high magnitude tornadoes bar graph above with aggregating the counts by state and magnitude, creating a pivot and then a stacked bar graph.

**A screen shot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

\*Complete result is not shown in screenshot above\*

A screenshot of a computer screen

Description automatically generated

To answer the question of whether the magnitude of tornadoes has worsened over the years, we grouped the data by year and magnitude to get counts for each. We plotted the results in a line plot and got the results below.

A graph showing the weather forecast

Description automatically generated with medium confidence

It’s a little difficult to see changes in the line graph with the except of the last few years, in which the magnitude is unknown. To get a clearer visual, we took the approach of creating a pivot to get the counts of tornadoes by year and magnitude, and then created a line graph plotting the counts by year in which the colors represent the magnitudes (see visual below). In this visual, the trend lines look steady for tornadoes with magnitudes of 3, 4 and 5 through the span of 1950-2020. In fact, the visual shows how rare it is for tornadoes of magnitudes 4 and 5 to occur. However, for tornadoes of magnitude 2, from around 1955 to 1975, it was more common for about 200 or more to take place. After 1975, less than 200 took place each year.

Tornadoes with magnitudes of 1 show an increase at around 1975, a steady up and down trend between 400 and 200 counts—no less than 200, and then a spike of more than 600 at around 2010 and before 2020. The zero magnitude tornadoes had a huge uptick in 1990 and remained steadily above 550 counts per year after then, with then another uptick of about 1,200 around 2005 and another uptick of about 1,000 before 2010. This is a clear sign of the advancement of technology over the year after the invention of super computers in 1990’s. There’s also a sudden uptick of tornadoes with unknown magnitudes around 2019-2020, which was when the Covid-19 world pandemic took place.

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

Next, we wanted to see what year most tornadoes occurred. For this, we used the mode() function. We also use code to print some of those tornadoes and how many. In 2004, there were a total of 1,817 tornadoes in this dataset.

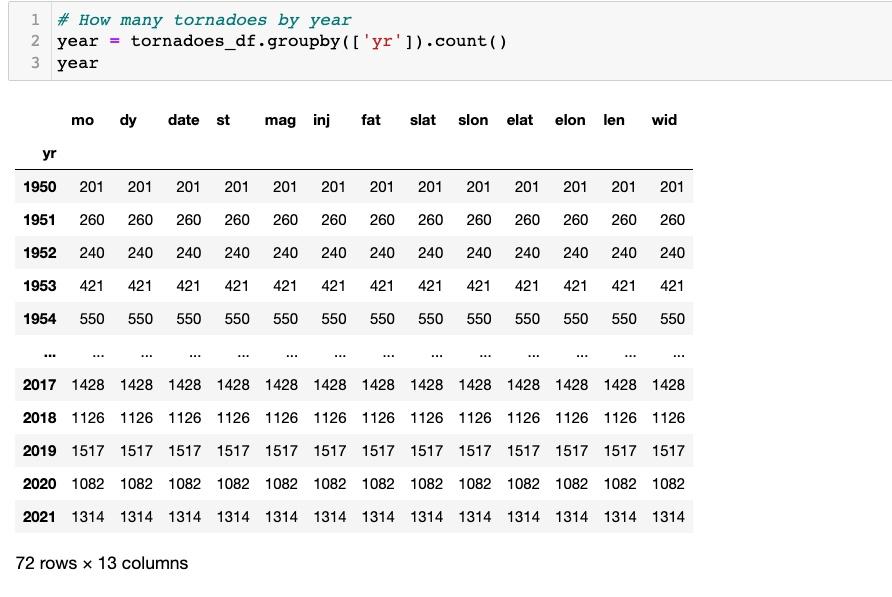
A picture containing text, font, screenshot

Description automatically generated

**A screenshot of a computer

Description automatically generated with medium confidence**

We also used group by () to extract counts of tornadoes by year and construct a line plot to show the trend of total tornadoes for each year.

****

**A graph with blue lines

Description automatically generated**

Next, we do more coding to find the tornadoes with unknown magnitude, totaling 605.

A screenshot of a computer

Description automatically generated

We find the longest tornado in miles and the widest tornado in yards using max(), giving us the same result. This tornado was a magnitude 2 in 1953 in Louisiana and was not the deadliest or with the most injuries.

A screenshot of a computer code

Description automatically generated with low confidence

The tornado with the most fatalities was one that occurred in Missouri in May 22nd, 2011, with 158 fatalities.

A screenshot of a computer

Description automatically generated

The tornado that caused the most injuries took place in Texas on April 10th, 1979, causing 1,740 injuries.

A screenshot of a computer

Description automatically generated

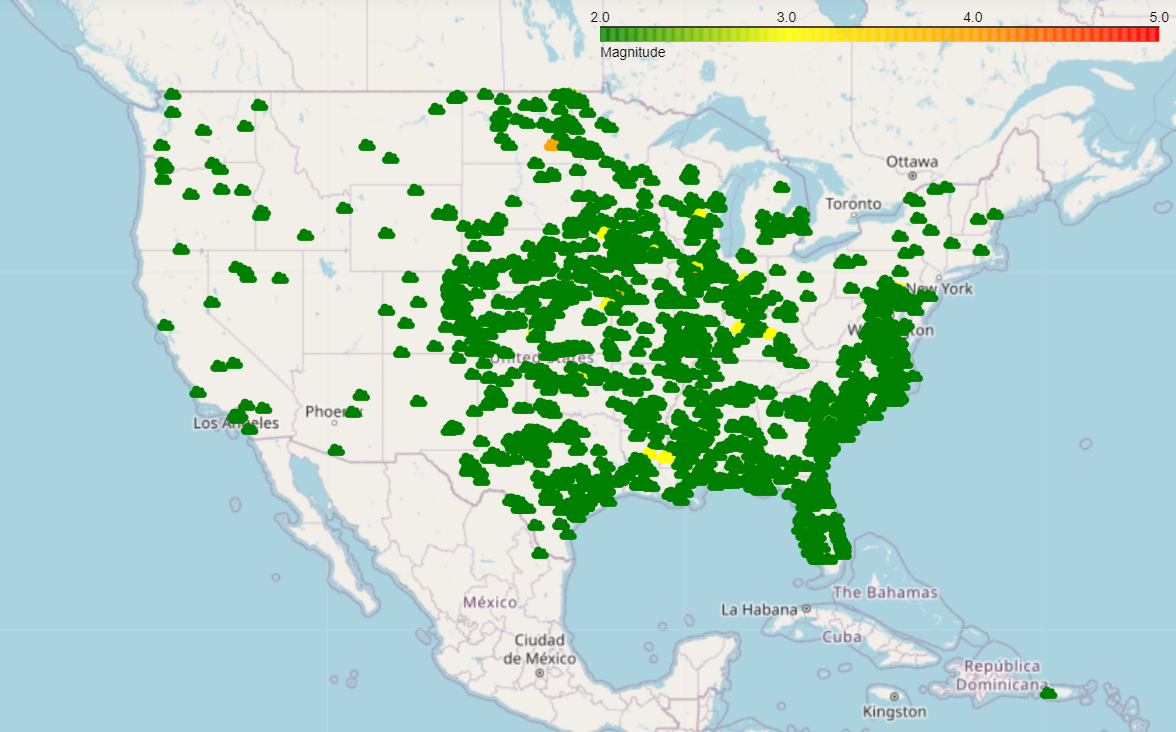
As mentioned before, most tornadoes took place in the year 2004. The code with the map below shows the tornadoes:

A screenshot of a computer code

Description automatically generated

A map of the world

Description automatically generated



A map of the united states

Description automatically generated

We can see many tornadoes in 2004, but only a few with high magnitude; a magnitude of 4 was recorded in North Dakota and Missouri. Also, there were a few tornadoes in Louisiana, Oklahoma, Kansas, Kentucky, Indiana, and Illinois.

Further, we performed a geospatial analysis of the tornadoes with magnitudes 2,3,4, and 5 that happened between 2008-2020 and during the months of April, May, June, and July when tornadoes were most common. We generated a heat map below. We can see those tornadoes most commonly occurred in the Eastern states, most located in the South- East. Some are happening in Florida and California.

A map of the united states

Description automatically generated with low confidence

A regression analysis was made on the tornadoes dataset to see if we can identify any relationships and measure magnitude of changes. First, we ran a regression using fatalities as the dependent variable, and latitude, longitude, length and width as the independent variables.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

Looking at the regression analysis results for fatalities, we can see that the p-value of the F-statistic (Prob F-statistic) is less than 0.05, which tells us the model is significant, and we have confidence to interpret. However, the R-square is admittedly low, 0.065, which tells us 6.5% of the change in fatalities is explained by the changes in the coefficients that are statistically significant. In this case, only the length and width have p-values lower than 0.05 and are statistically significant. Both coefficients are positive meaning as say the length of the tornado increases, fatalities an increase by 0.0358 and that is the same with width but by 0.0007, which is a much smaller change.

**A picture containing text, screenshot, number, font

Description automatically generated**

For injuries, we found that the p-value of the F-statistic is also less that 0.05, which also gives us confidence in this model. R-squared is a bit higher than this time, 0.076, which tells us 7.6% of the change in injuries is explained by the changes in the coefficient. In this regression analysis, we have 3 coefficients that have a p-value less than 0.05 and are therefore statistically significant. The longitude, length, and width. They also have positive coefficients letting us know that as each coefficient increases, the number of injuries can increase as well. The length seems to have the strongest magnitude of change indicating that as the length increases, the number of injuries can increase by 0.4785. Of course, we do have to keep in mind that the percentage of change, 7.6%, is still low.

Birds Dataset added

We also added the Birds dataset to the tornadoes dataset to compare if birds would be good predictors of tornadoes.

A screenshot of a computer

Description automatically generated

Data cleaning was performed, such removed missing values, date column to datetime format, # Splitting the "loc" column by the last comma into "loc" and "state" to derive a state column separately, then the state abbreviation was performed to have an abbreviated form of the state column to merge with tornadoes dataset**.** Filtering data by the United States was completed to focus only on the US territory.

**A screenshot of a computer program

Description automatically generated with medium confidence**

****

**A screenshot of a computer code

Description automatically generated with medium confidence**

The statistical description of the bird songs dataset was done, and we can see that there are few integer values in this dataset, and we can analyze on the frequency of the variables, states, and locations.

****

**A picture containing text, screenshot, font, number

Description automatically generated**

We generated separate histograms for the 'lat' and 'lng' columns of the birds\_df DataFrame, showing the distribution of values in each column.

Latitude:

A picture containing diagram, screenshot, plot, line

Description automatically generated

Longitude:

A picture containing screenshot, diagram, text, plot

Description automatically generated

Most bird recordings were found in 2015, which we used to analyze data further.

A picture containing text, font, screenshot, line

Description automatically generated

A screenshot of a computer program

Description automatically generated with medium confidence

In 2008, the most tornadoes happened, and in 2015, we have the most bird recordings. Also, recent years 2019 and 2020, were used for data merging and comparison analysis.

A picture containing text, font, screenshot, line

Description automatically generated

A picture containing text, font, screenshot, line

Description automatically generated

A close-up of numbers

Description automatically generated with low confidence

A screenshot of a computer code

Description automatically generated with low confidence

A screen shot of a computer code

Description automatically generated with low confidence

In the comparative analysis of the Tornado fatalities vs. Bird Callings in 2008, we can see the most bird callings with the least tornado fatalities.

A map of the united states

Description automatically generated

A map of the united states

Description automatically generated

The same Python code but with different years was used for 2015, 2019, and 2020. We also can see that there are fewer tornado fatalities in the states where birds were located, and where there are tornado fatalities, the birds are not recorded greatly. This contributes to the hypothesis that birds can detect tornadoes ahead of time and may not fly to the affected area at that time or may fly to different locations.

A map of the united states

Description automatically generated

A map of the united states

Description automatically generated

A map of the united states

Description automatically generated

A map of the united states

Description automatically generated

**A map of the united states

Description automatically generated**

**A map of the united states

Description automatically generated**

Lastly, correlation analysis was performed to determine if there is a particular bird genus that can determine if the tornado is happening.

**A screenshot of a computer

Description automatically generated with low confidence**

By looking at the correlation analysis, we can see that there is no positive correlation and a slight negative correlation - birds are not heard when tornadoes are happening.

Bird\_Euphagus has -0.183367, and bird\_Tympanuchus has -0.377432 correlation with tornado mag. This is the strongest negative correlation that may mean these birds are most likely not to be heard when tornadoes happen.

This shows that this type of bird genus could be a possible predictor for tornadoes. However, the correlation is relatively small, only -37%, so we cannot be certain, and further studies on these species need to be performed.

**A picture containing text, screenshot, font, line

Description automatically generated**

Using the following codes:

A picture containing text, font, screenshot

Description automatically generated

A picture containing text, font, screenshot

Description automatically generated

Above, We have an example of pattern detection done in both datasets. We pinpointed a Tornado on April 9th, 2015, in Texas in coordinates 32 Latitude and -95.1992 Longitude. In the Bird Song dataset, two bird calls were recorded within less than 200 miles of the tornado's location. One bird recording was detected on April 8th, 2015, in coordinates 31.2032 Latitude, -97.6124, and the other was detected after the tornado on April 10th, 2015, in coordinates 32.0320 Latitude and -97.1212 Longitude.

**Conclusion:**

By statistical analysis, we determined seasonal patterns of tornadoes; they mostly happen in March, April, May, June, July, and August. Every year we see an increase in tornadoes; however, they are not in very high magnitudes. We believe this is due to technological advancements that make tracking more tornadoes and their magnitudes possible. Most tornadoes happen in states such as Texas, Oklahoma, Kansas, Nebraska, North and South Dakota, Iowa, Minnesota, Wisconsin, Illinois, Indiana, Missouri, Arkansas, Montana, and Ohio. We found that the most tornadoes took place in Texas, and the two other states affected are Kansas and Colorado. And the greatest magnitude of tornadoes is happening in Texas, Kansas, Oklahoma, and Alabama. By the regression analysis, we have determined that the higher the starting latitude of the tornadoes, the fewer fatalities and fewer injuries the tornado brings. The greater the tornado's length or width, the more injuries it can get.

We obtained a geospatial analysis by merging the bird data with the tornado data. We determined that the birds do not happen to be heard in the areas where the tornadoes are happening the most, and birds migrate from where tornadoes are about to happen. By pattern detection, we found that birds were heard a couple of months before the tornado in the particular area and a couple of months after. This supports our idea that birds could be predictors of tornadoes or severe weather. The correlation analysis determined that the bird genus bird\_Euphagus has -0.183367, which is the strongest negative correlation between bird callings and tornado magnitudes. This tells us that bird callings are not likely to happen when tornadoes happen, which supports our theory.

|  |  |
| --- | --- |
| **Name** | **Tasks performed** |
| Halyna Tymofeyev | Finding Birds dataset, cleaning birds data set, contributing to questions, coming up with analyses, statistical analysis on birds and tornado datasets, merging data sets, geospatial analysis, regression and correlation analyses.  Writing the second part of the final report, method of analyses, program description, and conclusion.  PowerPoint- Second half and conclusion. |
| Pierina Logroño | Set up Google collaboration documents, finding tornado dataset, cleaning tornadoes dataset, coming up with questions, researching articles, exploratory analysis, determining trends, statistical analyses on birds and tornado datasets. Writing introduction and first half of the final report, data description and sources, checking for grammatical errors.  PowerPoint- Second introduction and first half. |

References:

1. Goertzel, Laura. “Tornadoes.” *National Geographic Kids.* Date unknown. <https://kids.nationalgeographic.com/science/article/tornado>

1. Bryce, Emma. “How do Tornadoes Affect Birds?” *Audubon.org*. 24 May 2013. <https://www.audubon.org/news/how-do-tornadoes-affect-birds#:~:text=Trees%20where%20birds%20nest%20are,wildlife%20deaths%20after%20a%20storm>.

1. McGlashen, Andy. “Are These Birds Better Than Computers at Predicting Hurricane Seasons?” *Audubon.org.* 13 August 2019. <https://www.audubon.org/news/are-these-birds-better-computers-predicting-hurricane-seasons>