CAPSTONE REPORT

Destination Birding, informing Knoxville Birders

DESTINATION BIRDING

Marilyn Long

06.23.2023

BrainStation Data Science Bootcamp

INTRODUCTION

Birding is estimated to be a $40B industry in the United States annually, and in 2011 created 660,000 jobs (USFWS). Birders, for the most part, are inherently adventurous and willing to travel to see new or beautiful birds to add to their life lists (which is a list of species they've seen – mine is nearly 600). Personally, I recently spent a month in Colombia. One of the main drivers for my trip was to visit the country with the most bird species in the world – about 20%!

My app, Destination Birding, is meant to inform birders on when and where to bird. It is currently limited to Knoxville, TN, but cities across the world could replicate this to promote bird tourism in their area, boosting their local economy by both travel-generated revenue (ex. Hotel taxes) and job creation.

MY DATA

I used data from eBird, an online database of bird observations created by the Cornell Lab of Ornithology. eBird provides researchers and amateur naturalists with real time data about bird distribution and abundance. I requested access to seven years of eBird data from Knox County, TN, which has its own R package, auk, to extract data from their txt file. The original data frame has almost 50 columns. I reduced it down to 7:

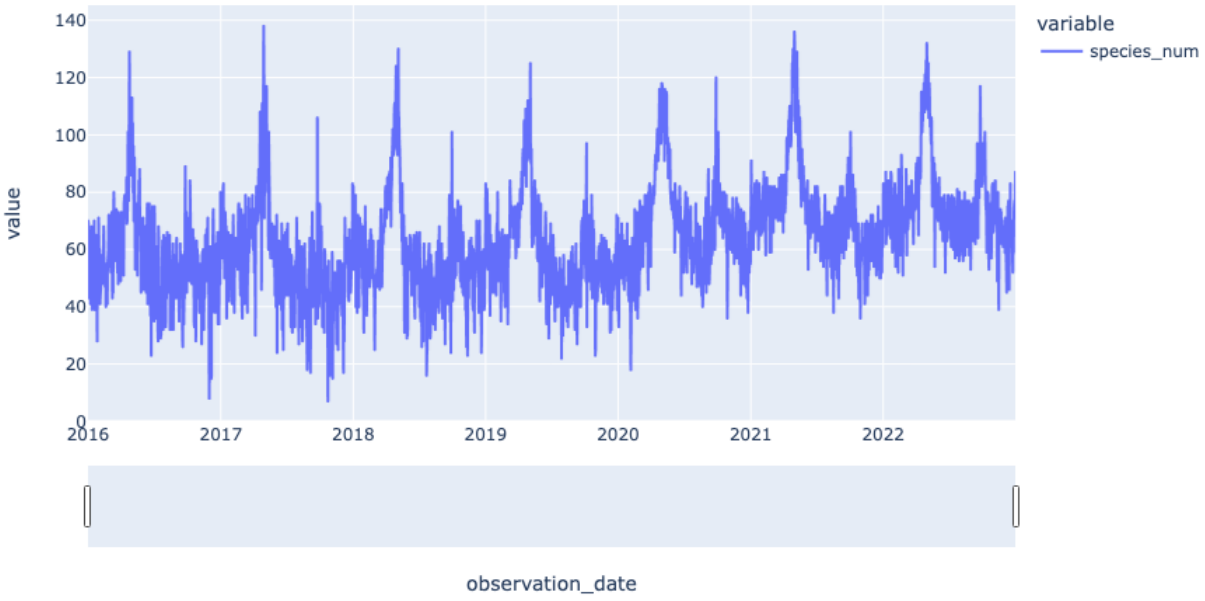
- Checklist ID
- Common Name
- Observation Count
- Locality
- Latitude
- Longitude
- Observation Date

CLEANING & EDA SUMMARY

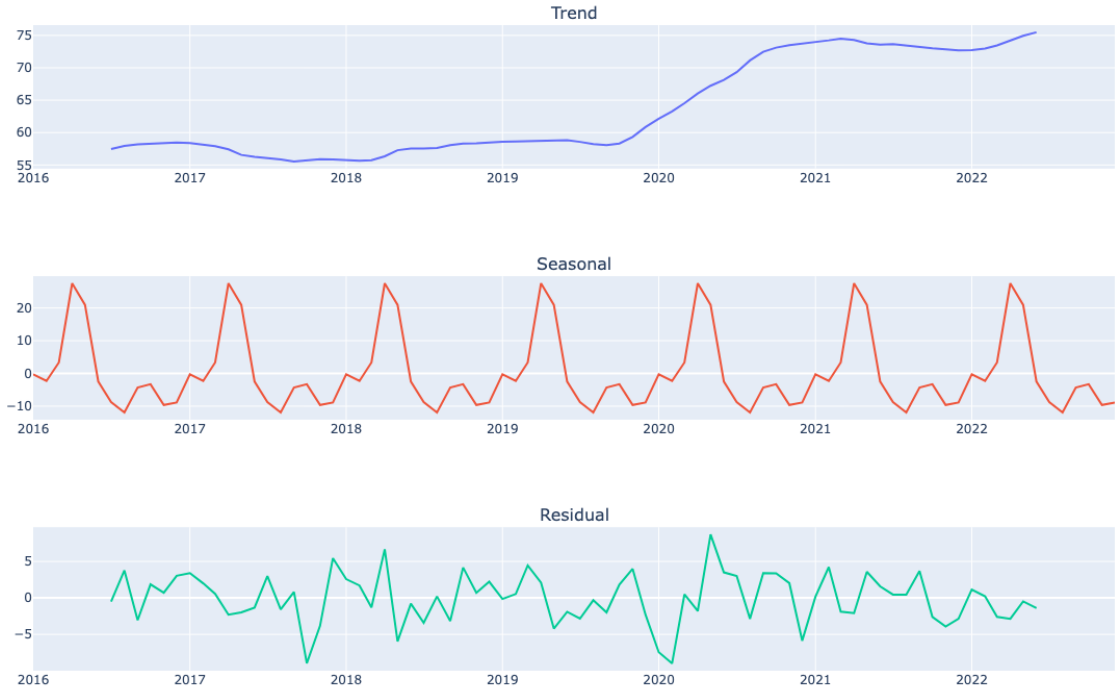
The dataser was pretty clean when acquired, but I did remove non-naturally occurring species (ex. peacocks). The main issue I had to deal with was deciding how to treat NaNs in the observation count column. If a user hears a certain bird species, but doesn't see it, they may mark "heard not seen," which results in an X for the observation count. There

were around 18,000 cases of this, representing about 2% of the data. I decided to fill these with ones, thinking that if they heard the bird, there must at least be one of them present. Upon further improvements of the project, I'd like to experiment with imputing the nulls using Funk Single Value Decomposition to predict the most likely observation count.

One of the first things I decided to explore was species diversity over time. The dataset had no gaps in time, which I found quite impressive and exciting as a fellow birder. That meant that for every single day from 2016-2022, at least one person submitted a checklist to eBird in Knox County. Next I grouped by observation date and summed the number of unique species observed daily.



It looks like from 2016-2020 there's an annual average of about 60 unique species. However from 2020-2022, the average appears to be a bit higher, around 70. Decomposing, or breaking down, the time series into an overall trend, seasonal variation, and residuals helps make this more interpretable.



Now it is very clear that there was an increase in species diversity starting at the end of 2019. As we all know, COVID-19 changed people's behaviors in many ways. People were spending a lot more time at home, and a lot of people became birders during this time period (NYT). My intuition told me that species diversity hasn't actually increased in the past few years; it rose as a factor of more people birding. I confirmed this by going through the same process but instead of counting unique bird species, I counted unique checklists IDs. The graphs look almost identical.

MODELING & RESULTS

The app consists of three main parts:

1. A species diversity map
    a. Use case: It's June. Where should I go to see the most species?
2. Species-specific time series analyses
    a. Use case: when is the American Coot most likely to be in Knoxville?
3. A species-specific bar chart and species and time-specific map
    a. Use case: when should I look for Bald Eagles? Where are they found?

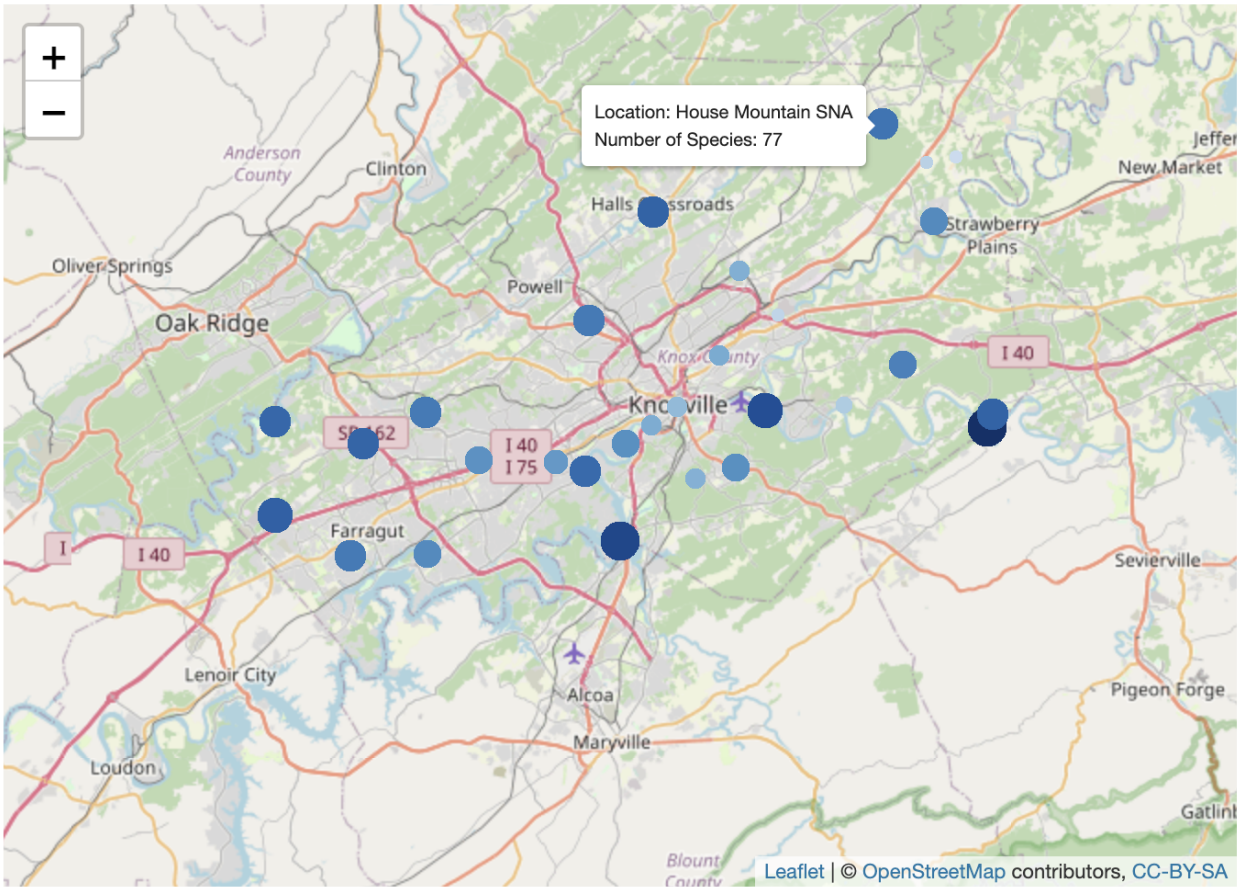
I'll dig into each of these in more detail.

Species Diversity Map

The species diversity map reacts to the user input, the month. It filters the data for that month, for example June. Because there were almost 7,000 unique latitude and longitude points on the map, I had to cluster, or group, the data together for it to be useful. I used k-means to group nearby points into 30 distinct clusters around Knoxville.

The next challenge was labeling the points. On eBird, the user can name their location or use an already established one. This can make naming the locality tricky. Ultimately, the point on the map should reflect the name of the place at which people are going to bird. For example, Seven Islands State Birding Park is a popular spot to look for birds. Some people use the full name, some call it "7 islands", "park", etc. My approach was to go with the masses, meaning I'd look for the most frequent locality name in the cluster and use that.

Here's an example output for the month of June.



Species-specific Time Series Analyses

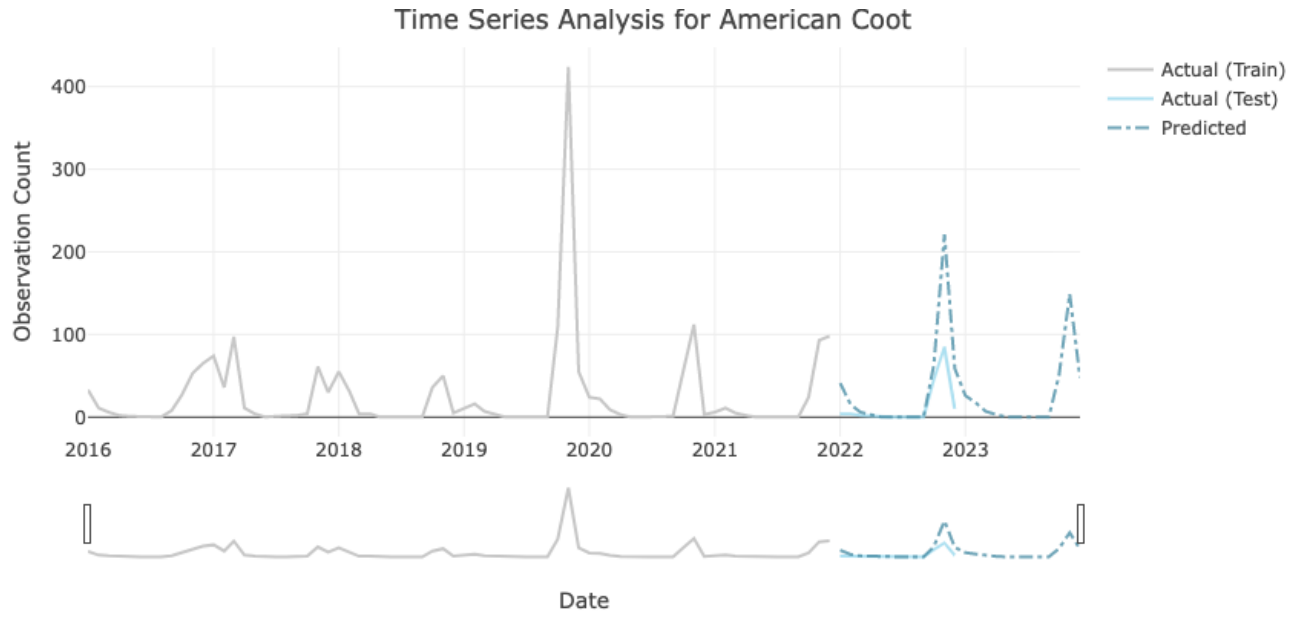
The Time Series Analysis tab in Destination Birding can be used to predict future observations of species of interest. The tab is interactive and will show the analysis for the species selected in the dropdown. The data is first filtered for the selected species.

Next, the series is completed, meaning any dates that are missing from the start and end date are added and the observation counts are filled with zeroes. My thought here was that if the bird wasn't observed, then it was not present. (This is another instance of where Funk SVD could be used to impute the missing values, however the SARIMA still yielded good results for most species).

Then the data was split into train and test sets, pre and post 2022. The predicted values start in 2021 and extend to 2023. The purpose of this is to be able to compare the model's predicted values against real observations in the test set, and still see future predictions.

Finally, the train data was run through R's forecast::auto.arima() which uses min and max parameters values to automatically select the best parameters based on AIC, a statistical evaluation metric. In more general terms, the model is optimized for each unique species without having to manually change the parameters.

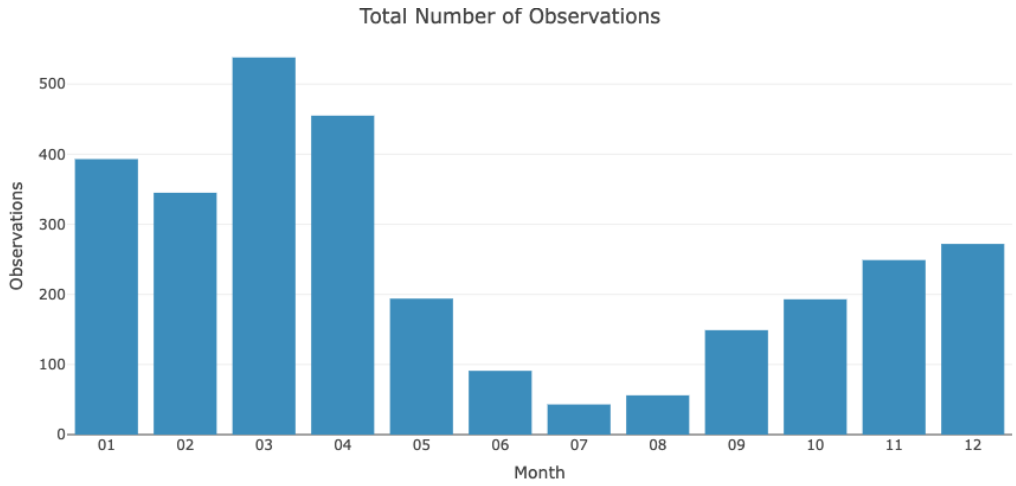
The image below shows the output for the American Coot.



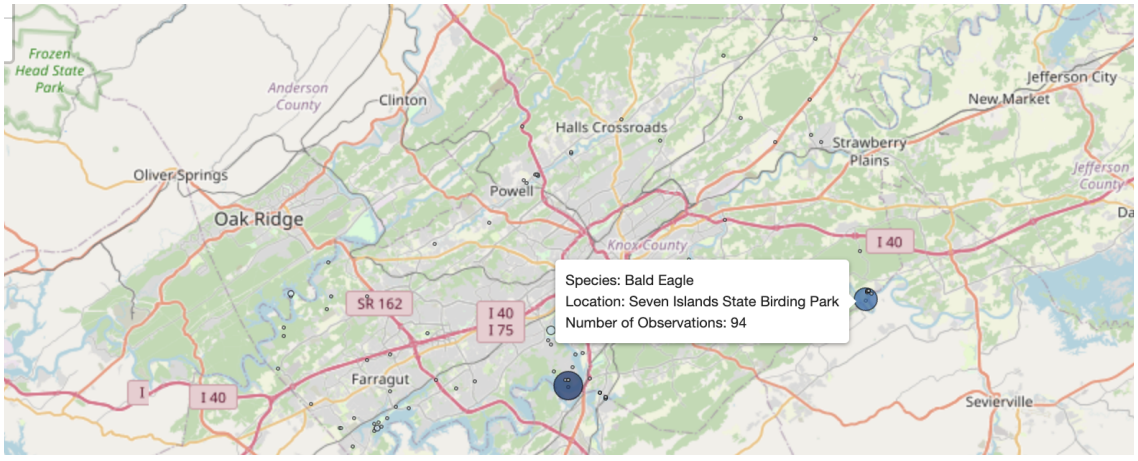
When and Where to bird: Map & Bar Chart

The final part of the web app shows both a map and bar chart. The bar chart shows the sum of observations for the species of interest and the map shows where those observations occurred. The bar chart is reactive only to the species input, while the map needs both the species and month(s) to be selected. There is also a user-input option to cluster the data or not, which uses dbscan. The bar chart is meant to be used as a reference. With a quick glance, it's apparent when and where a certain species is most frequently observed in Knox County.

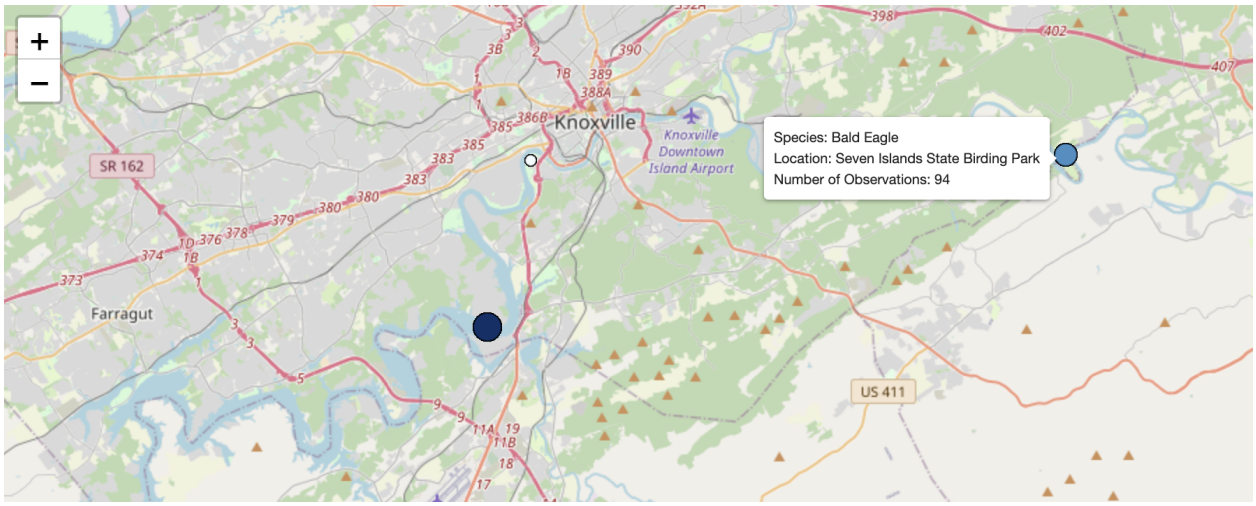
The bar chart filters for the selected species, groups observations by month and adds up the total observation counts. Here's an example of what that looks like for the Bald Eagle.



The map filters by month(s) and species. Optionally, it can show clustered/grouped data or not. The marker size and color intensity indicate higher observation counts. Here's an example for the Bald Eagle in March **not** clustered.



And here's an example of what the map looks like for the same bird and time, but **with** clustered data:



The eye is easily drawn to the best places to look for this bird without all the noise.

CONCLUSION

My app let's users intuitively and quickly visualize eBird data in the Knoxville area. Cities could use this process to explore bird data in their area and then use that information to entice birders to visit their city, thereby boosting their local economy by collecting revenue from hotel taxes and creating jobs.

For birders, the species diversity map lets them see the most diverse places to go birding given the time of year. The time series analysis lets them see when a species is most likely to be seen in Knoxville. The When & Where bar chart and map provides species-specific mission information. For example, say someone really wants to see a Tennessee Warbler this year. Now they can easily see when and where to go looking.

For future improvements of this project, I'd like to experiment with using Funk SVD to impute NaN observation counts. I'd also like to try different clustering techniques. Instead of clustering using k-means with 30 centers, I could try to set birding destination centers (ex: parks and natural areas in Knoxville) and group observations within a certain radius. I'd also like to use a silhouette score to evaluate dbscan to choose the best parameters.