Project Milestone 2

Kimberly Adams

08/7/2022

A quick note before you start grading this document, I ran into significant issues trying to import/process my key dataset due to its size (even a slimmed down version limited to just one county is over 770 million rows of data) and thus much of the progress I wanted to present is absent. I apologize profusely and hope to have everything in order by the final milestone or at least something better to present than this document.

How to import and clean my data

• Data importing and cleaning steps are explained in the text and follow a logical process. Outline your data preparation and cleansing steps.

Import eBird Data

The data I downloaded was already filtered down to only Florida, United States observations, but it is still too large to view. I am going to have to filter down the observations to complete checklists (all species observed recorded in checklist) using stationary or traveling protocols I am also going to narrow my search to Orange County, Florida as this is a county around Orlando which is in the center of the state and contains a very active bird watching hotspot (Lake Apopka Wildlife Drive).

```
## Note: Auk is a package spcifically designed for working with eBird data.
## The following import steps attempt to follow published instructions found here:
## https://cornelllabofornithology.github.io/ebird-best-practices/ebird.html#ebird-extract
## Code show for example, will not run in markup due to time intensity
library(auk)
library(lubridate)
library(gridExtra)
library(tidyverse)

# resolve namespace conflicts
select <- dplyr::select

# setup data directory
dir.create("eBirdData", showWarnings = FALSE)

ebd <- auk_ebd("/Users/kimberlyadams/Documents/GitHub/DSC520-Statistics-and-R/data/Project/eBird Orange

ebd$col_idx$name

## Remove un-need Columns by saying which columns to keep
cols <- c("global unique identifier", "taxonomic order", "common name", "observation count", "county",
```

```
f_select <- "data/ebd_smaller.txt"
selected <- auk_ebd(f_ebd) %>%
  auk_select(select = cols, file = f_select) %>%
  read_ebd()
glimpse(selected)
# file size difference
file.size(f_ebd) / file.size(f_select)

## Creating a subset of the dataset
ebd_filters <- ebd %>%
    # restrict to the standard traveling and stationary count protocols
    auk_protocol(protocol = c("Stationary", "Traveling")) %>%
    # restrict to only observations from Orange County, Florida
    auk_complete()
ebd_filters

# output files
data_dir <- "eBirdData"
if (!dir.exists(data_dir)) {
    dir.create(data_dir) }

f_ebd <- file.path(data_dir, "ebd_US-FL-095_200201_202112_relJun-2022.txt")

# only run if the files don't already exist
if (!file.exists(f_ebd)) {
    f_sampling <- file.path(data_dir, "ebd_sampling_relJun-2022.txt")
    auk_filter(ebd_filters, file = f_ebd, file_sampling = f_sampling)
}
View(f_ebd)
```

I need to learn how to better use the AUK package to filter the eBird dataset to get the two types of values I need from it:

- number of checklists submitted each year
- total number of species seen each year

As of right now, I have failed miserably to be able to filter the dataset to a point were I can work with it. The original file for all of Florida is a 15 GB txt file that I cannot open with the TextEditor and Excel on my computer. Even a different file with just Orange County, Florida has so many rows in it that neither R nor Excel will open it. I will have to continue working with this to get it to filter. If I cannot do this, then I may have to change project topics.

Import the Precipitation, Temperature, and Hurricanes Datasets

Precipitation

I will need to extract the Annual values (from the Annual column) for the target years (Year column) for which I have eBird data. eBird only started in 2002, so I am going to remove entries from prior to that year.

Temperature

I will need to extract the Annual values (from the Annual column) for the target years (Year column) for which I have eBird data. eBird only started in 2002, so I am going to remove entries from prior to that year.

Hurricanes

This dataset will definitely need clean up for my analysis both in filtering for only hurricanes that hit Florida and then cleaning up the year the hurricane occurred so that I can then total how many hurricanes hit Florida each year.

Filtering for just hurricanes that hit Florida and cleaning up the Date column to extract Year:

eBird only started in 2002, so I am going to remove entries from prior to that year. Even though eBird might contain data from before, I am not sure how valid it is.

Also, Hurricane Irma is listed twice because it changed intensity categories so I will keep the higher category one.

- What do you not know how to do right now that you need to learn to import and cleanup your dataset?

I need to figure out how to merge data sets properly where 0's are added for years with no hurricanes and also my summarizing needs work as I am ending up with duplicated years when I try to merge all the datasets together.

What does the final data set look like?

```
# Rename columns as necessary
names(precipAnnual.df)[names(precipAnnual.df) == 'Annual'] <- 'AnnualPrecip'
names(tempAnnual.df)[names(tempAnnual.df) == 'Annual'] <- 'AnnualTemp'

weatherData.df <- merge(precipAnnual.df,tempAnnual.df, by = "Year")
FinalData.df <- merge(weatherData.df,HurricanesAnnual.df, by = "Year")
```

My goal for my final data set is to combine information from the different datasets and include the only the following columns:

- Year = probably ranging from 2002 (when eBird started) to 2021 (2022 excluded due to incomplete future data). The year range is dependent on how far back the eBird data goes.
- AnnualPrecip = total precipitation for the year
- AnnualTempe = Average temperature for the year
- HurriNum = number of hurricanes that hit Florida during the year
- HurriInt = average hurricane Category during that year
- HurriMaxWind = max hurricane wind speed during the year
- eBirdChks = number of checklists submitted to eBird during the year
- eBirdSpsPerChk = average number of different species seen within indivudal checklists during the year

What information is not self-evident?

I will need to calculate the values for the eBird data in regards to the total number of checklists and species observed each year. I will also need to calculate the hurricane data for number that hit Florida each year and the average Category and max wind speed for the year.

What are different ways you could look at this data?

I could pick a single species of bird to look at weather impacts. This is a can of worms though because I feel I would have to pick a year-round resident species, a migratory species, and a vagrant species (rarely seen in Florida) to really get a good idea on the impact of weather on individual bird populations.

A value I would love to be able to extract is the number of unique species seen within each year. This would require summarizing the eBird data by year and by species and then counting the listed species. Considering how difficult it is for me to manipulate the eBird data at the moment, this will have to wait.

I am sure there are other variables I could create when looking deeper into this dataset, but I tried to not let this project sprawl any more than it already has.

- Could summary statistics at different categorical levels tell you more?

This could definitely be the case if I kept all the columns in the eBird dataset as there is a lot to explore in it. Again the difficulty I have working with that massive dataset makes me less inclined to explore further than needed.

The weather data (Precipitation, Temperature, and Hurricane) data could be looked at on a monthly level along with the eBird data tie in. The reason I have decided on a yearly interval is to help smooth out seasonal variability, but comparing exact months to each other could be possible. With this level of study, you could also look at Spring and Fall bird migration changes.

How do you plan to slice and dice the data?

- Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.

I plan to extract useful variables from their respective data frames and then combine them into a new data frame where each year has its respective variable values. I have explained my goal in the previous section describing what the final data set would look like.

- How can you incorporate the pipe (%>%) operator to make your code more efficient?

Pipes are best used when trying to filter the eBird data with multiple criteria, but could be used more with some of the weather data too. They are useful in just about any multi-step process. Personally, I find it easier to read the code without them, but that is just because I am so new. I can definitely see how they would save code space.

What types of plots and tables will help you to illustrate the findings to your questions?

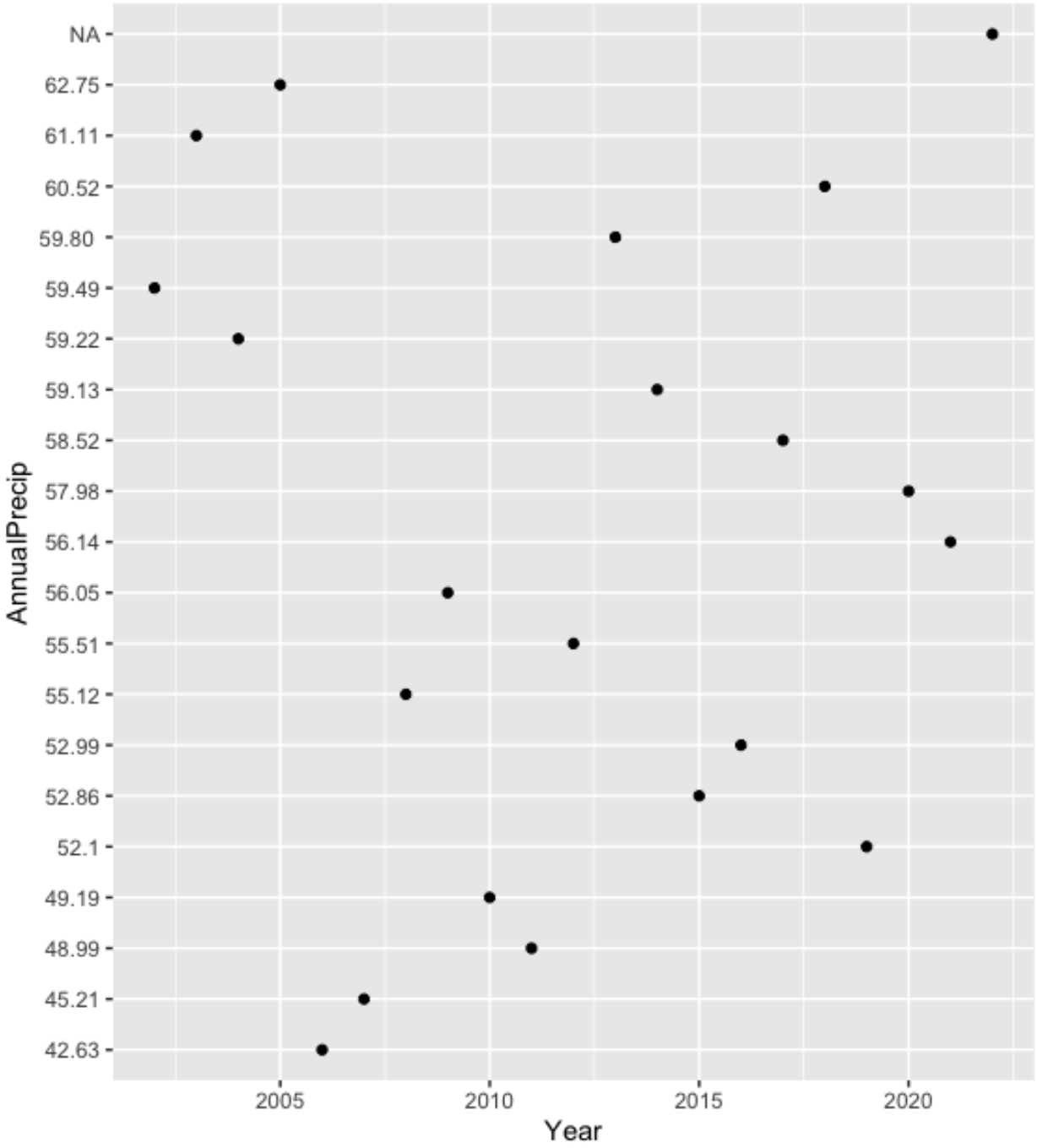
Scatterplots can help show the patterns of each variable. Note how there isn't really a pattern in the weather data:

```
library(ggplot2)
ggplot(precip.df, aes(x = Year, y = Annual)) + geom_point() + ggtitle("Annual Precipitation") + ylab =

# Code will not run in markdown due to error "could not find function "+\<-"
```

- What do you not know how to do right now that you need to learn to answer your questions?

The more I dive into the project, the more I am finding I can't get things to work. I need a better understanding overall and expecially need to be able to figure out why things that did work no longer do. For example, why did my precipitation graph work and give me the output on the next page and then suddenly start giving me this error: "could not find function"+<-" "?



Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

I haven't gotten that far honestly. My gut instinct is that I won't be able to show enough of a correlation to make it worth building any sort of model or to apply machine learning.