Bird Species

The `OrdwayBirds` data table is a historical record of birds captured and released at the Katharine Ordway Natural History Study Area, a 278-acre preserve in Inver Grove Heights, Minnesota, owned and managed by Macalester College. Originally written by hand in a field notebook, the entries have been transcribed into electronic format under the supervision of Jerald Dosch, Dept. of Biology, Macalester College.

Due to mistakes in data entry, the `SpeciesName` variable needs some fixing. `SpeciesName` is intended to identify the species of each of the birds, but the spelling often varies among birds of the same biological species. This leads to mis-classification of birds. There are also problems with the `Month` and `Day` variables; they are supposed to be numerical, but mistakes prevent them from being correctly identified as such.

Fortunately, all these errors are easy to correct. The data table `OrdwaySpeciesNames` collects together all the variant spellings. Entry by entry, each mis-spelling was translated (by a human) into a standardized spelling. Thus, `join()` can be used to correct the mis-spellings in the `OrdwayBirds` table.

You are going to look at the month-to-month presence of different species. Think of your assignment as creating a manual for birders to guide them to the correct time of year to visit Ordway to see a particular species.

Some simple data cleaning

There are many variables that you won't need for this activity, and you still have to fix the `Month` and `Day` variables. To keep things simple, cut and paste this command into a chunk at the start of the document.

Step o *Get the data table*

```
OrdwayBirds <-
  OrdwayBirds %>%
```

```
select( SpeciesName, Month, Day ) %>%
mutate( Month=as.numeric(as.character(Month)),
        Day=as.numeric(as.character(Day)))
```

The `mutate()` step is part of the data cleaning process, arranging `Month` and `Day` as numerical variables as originally intended by the folks entering the data.

Task 1. Including mis-spellings, how many different species are there in the `OrdwayBirds` data?

Make a data table that gives the number of distinct species in the `SpeciesNameCleaned` variable in `OrdwaySpeciesNames`.
You will find it helpful to use `n_distinct()` a reduction function, which counts the number of unique values in a variable.

Task 2 Use the `OrdwaySpeciesNames` table to create a new data table that corrects the mis-spellings in `SpeciesNames`. This can be done easily using the `inner_join()` data verb.

```
Corrected <-
  OrdwayBirds %>%
  inner_join( OrdwaySpeciesNames ) %>%
  select( Species=SpeciesNameCleaned, Month, Day ) %>%
  na.omit() ## cleaned up the missing ones
```

Look at the names of the variables in `OrdwaySpeciesNames` and `OrdwayBirds`.

- Which variable(s) was used for matching cases.
- What were the variable(s) that will be added .

Task 3 Count how many bird captures there are of each of the (corrected) species? You can call the variable that contains the count `count`. Arrange this into descending order from the species with the most birds, and look through the list.

Define for yourself a "major species" as a species with more than a particular threshold count. Set your threshold so that there are 5 or 6 species designated a major.

Filter to produce a data table with only the birds that belong to a major species. Save the output in a table called `Majors`.

Task 4 When you have correctly produced `Majors`, write a command that produces the month-by-month count of each of the major species. Call this table `ByMonth`.

Hint: Remember `n()`. Also, one of the arguments to one of the data verbs will be `desc(count)` to arrange the cases into descending order. Display the top 10 species in terms of the number of bird captures.

Hint: Remember that summary functions can be used case-by-case when filtering or mutating a data table that has been grouped.

Display this month-by-month count with a bar chart arranged in a way that you think tells the story of what time of year the various species appear. You can use `barGraphHelper()` to explore different possibilities. Use the "Show Expression" button in `mBar()` to create an expression that you can cut and paste into a chunk in your Rmd document, so that the graph gets created when you compile it.

Warning: `barGraphHelper()` should never be a statement in your Rmd file, it needs to be used interactively from the console.

Once you have the graph, use it to answer these questions:

1. Which species are present year-round?
2. Which species are migratory, that is, primarily present in one or two seasons?
3. What is the peak month for each major species?
4. Which major species that are seen in good numbers for at least 6 months of the year? (Hint: `n_distinct()` and `>= 6`.)