Assignment 3: Data Exploration

Mason Ibrahim

Fall 2024

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Exploration.

Directions

1. Rename this file `<FirstLast>_A03_DataExploration.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Assign a useful **name to each code chunk** and include ample **comments** with your code.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.
7. After Knitting, submit the completed exercise (PDF file) to the dropbox in Canvas.

TIP: If your code extends past the page when knit, tidy your code by manually inserting line breaks.

TIP: If your code fails to knit, check that no `install.packages()` or `View()` commands exist in your code.

Set up your R session

1. Load necessary packages (tidyverse, lubridate, here), check your current working directory and upload two datasets: the ECOTOX neonicotinoid dataset (ECOTOX_Neonicotinoids_Insects_raw.csv) and the Niwot Ridge NEON dataset for litter and woody debris (NEON_NIWO_Litter_massdata_2018-08_raw.csv). Name these datasets "Neonics" and "Litter", respectively. Be sure to include the sub-command to read strings in as factors.

```
#load packages
library(tidyverse)
library(lubridate)
library(here)

#read data
Neonics <- read.csv(
  file = here("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),
  stringsAsFactors = TRUE)
Litter <- read.csv(
  file = here("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv"),
  stringsAsFactors = TRUE)
```

Learn about your system

2. The neonicotinoid dataset was collected from the Environmental Protection Agency's ECOTOX Knowledgebase, a database for ecotoxicology research. Neonicotinoids are a class of insecticides used widely in agriculture. The dataset that has been pulled includes all studies published on insects. Why might we be interested in the ecotoxicology of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: In addition to the insects targeted for with pesticides, neonicotinoids could also have effects on beneficial insects such as pollinators. It could also help to see which insects it is most effective at killing.

3. The Niwot Ridge litter and woody debris dataset was collected from the National Ecological Observatory Network, which collectively includes 81 aquatic and terrestrial sites across 20 ecoclimatic domains. 32 of these sites sample forest litter and woody debris, and we will focus on the Niwot Ridge long-term ecological research (LTER) station in Colorado. Why might we be interested in studying litter and woody debris that falls to the ground in forests? Feel free to do a brief internet search if you feel you need more background information.

   Answer: Litter and woody debris can serve as an important part of the nutrient cycle in the forest. By adding nutrients back into the ground while decomposing, keeping the soil moist, and providing shelter for insects and other small creatures on the forest floor, litter and woody debris serve many purposes. This information can help calulate productivity in the forest.

4. How is litter and woody debris sampled as part of the NEON network? Read the NEON_Litterfall_UserGuide.pdf document to learn more. List three pieces of salient information about the sampling methods here:

   Answer:

   1. Ground traps were sampled once per year.

   2. Target sampling frequency for elevated traps varied by vegetation present at the site, with frequent sampling (1x every 2weeks) in deciduous forest sites during senescence, and infrequent, year-round sampling (1x every 1-2 months) at evergreen sites.

   3. Litter and fine woody debris sampling was executed at terrestrial NEON sites that contained woody vegetation >2m tall.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset? 4623 rows and 30 columns

```
dim(Neonics) #use dim() for dimensions
```

6. Using the `summary` function on the "Effect" column, determine the most common effects that are studied. Why might these effects specifically be of interest? [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```
neo_effect <- summary(Neonics$Effect) #create list of values in specifically in the effect
column
sort(neo_effect, decreasing = TRUE) #sort by decreasing value to make it simple to find
the most common effect
```

Answer:Population with a count of 1803 and mortality with a count of 1493 were the most common effects studied. These effects would specifically be of intest because it can illustrate the (either intentional or non-intentional) deaths the exposure to neonicotinoids cause.

7. Using the `summary` function, determine the six most commonly studied species in the dataset (common name). What do these species have in common, and why might they be of interest over other insects? Feel free to do a brief internet search for more information if needed.[TIP: Explore the help on the `summary()` function, in particular the `maxsum` argument…]

```
summary(Neonics$Species.Common.Name, maxsum = 6) #generating a summary with the top six
most commonly studied species with maxsum()
```

Answer: Other, Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee. Besides the 'Other" category, the rest of the insects are in the order of insects "Hymenoptera", which are known pollinators, making them important to the ecosystem function. Unintentional killing of these species could result in lower pollination rates.

8. Concentrations are always a numeric value. What is the class of `Conc.1..Author.` column in the dataset, and why is it not numeric? [Tip: Viewing the dataframe may be helpful…]

```
class(Neonics$Conc.1..Author.) #use class() to find class
```

Answer: Factor, it is not numeric because the values entered include other symbols such as </> and /.

Explore your data graphically (Neonics)

9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
#some studies repeat, just with different species, have to sort out the repeating
information

unique(Neonics$Title) #checking how many unique papers there are to make sure plot is
correct
neo_unique <- distinct(Neonics, Title, Publication.Year, Test.Location) #sort out
repeating papers

#plot the unique information

ggplot(neo_unique, aes(x = Publication.Year)) +
  geom_freqpoly(binwidth = 1) #bin = 1 to make sure we look at one bin per year
```

10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
ggplot(neo_unique, aes(x = Publication.Year, color = Test.Location)) + #adding color to
aes and setting it to Test.Location
         geom_freqpoly(binwidth = 1)
```

Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer:Test location was most common in the lab. Lab locations became more popular after the year 2000. Natural field sites were the most common before the year 2000, but decreased in popularity over the past couple decades.

11. Create a bar graph of Endpoint counts. What are the two most common end points, and how are they defined? Consult the ECOTOX_CodeAppendix for more information.

[**TIP**: Add `theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))` to the end of your plot command to rotate and align the X-axis labels…]

```
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar()+ #make it a bar graph with geom_bar()
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) #to rotate and align
the X-axis labels for ease of visulization
```

> Answer:NOEL and LOEL NOEL: No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test (NOEAL/NOEC) LOEL: Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC)

Explore your data (Litter)

12. Determine the class of collectDate. Is it a date? If not, change to a date and confirm the new class of the variable. Using the `unique` function, determine which dates litter was sampled in August 2018.

```
class(Litter$collectDate) #check class
Litter$collectDate <- ymd(Litter$collectDate) #change to date with lubridate
class(Litter$collectDate) #check class again
unique(Litter$collectDate) #unique values in collectDate column
```

13. Using the `unique` function, determine how many different plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

```
unique(Litter$plotID) #12 levels

summary(Litter$plotID)#compare to summary
```

> Answer: 12 different plots, unique doesn't count how many values are in each plotID, but it does tell you how many levels (or unique values) are in that column.

14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x = functionalGroup))+
  geom_bar()
```

15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter, aes(x = functionalGroup, y = dryMass))+ #use y= to change y axis
  geom_boxplot()

ggplot(Litter, aes(x = functionalGroup, y = dryMass))+ #same as above, but with violin
  geom_violin()
```

Why is the boxplot a more effective visualization option than the violin plot in this case?

> Answer: The dataset contains alot of zeros, with a few other higher values, causing the violin plot does not display a meaningful shape. It appears like a straight line since the bandwith isn't appropriate.

What type(s) of litter tend to have the highest biomass at these sites?

> Answer: Needles