Assignment 3: Data Exploration

Emma Wellbaum

OVERVIEW

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

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Salk_A03_DataExploration.Rmd") prior to submission.

The completed exercise is due on <>.

Set up your R session

1. Check your working directory, load necessary packages (tidyverse), 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.

```
# Check working directory
getwd()
```

```
## [1] "C:/Users/emmaw/Documents/ENV872/Environmental_Data_Analytics_2021/Assignments"
```

```
# Load necessary packages
library(tidyverse)
```

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

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 ecotoxicologoy of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: The ecotoxicology of neonicotinoids on insects is relevant to evaluating the efficacy of neonicontinoids in agricultural target species as well as non-target species. Neonicotinoids are the most widely used class of insecticide in the world, so we may interested in seeing how well it works. We may also be interested in seeing whether neonicotinoids are safe. In particular, studies have found that neonicotinoids can kill bees and other pollinators (i.e., non-target species).

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: Forest litter and woody debris are an indication of forest decomposition. We may be interested in studying litter and forest debris in Niwot Ridge in order to study the rate of forest decomposition, predict carbon cycling, or evaluate the available nutrients 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: * Litter and woody debris sampling occurred at NEON sites with wooded vegetation >2m tall. * NEON employed two methods of collecting litter and woody debris: elevated traps and ground traps. NEON deployed one elevated trap and one ground trap per 400 m^2 plot area. Depending on the vegetation type, the placement of traps may be random or targeted. * NEON conducted annual sampling of ground traps and targetted sampling of elevated traps (every 2 weeks - 2 months depending on the vegetation type).

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

```
## [1] 4623   30
```

6. Using the summary function on the “Effects” column, determine the most common effects that are studied. Why might these effects specifically be of interest?

```
Effects <- summary(Neonics$Effect)
head(sort(Effects, decreasing = TRUE))
```

```
##       Population        Mortality         Behavior Feeding behavior 
##             1803             1493              360              255 
##     Reproduction      Development 
##              197              136
```

   Answer: Population and mortality are by far the most common effects, which makes sense because both are important metrics of species health and survival. Population effects look at changes affecting the total population and mortality looks at death specifically. Mortality is also relevant to evaluating insecticide efficiacy, as an insecticide is designed to kill. The question is whether the insecticide is killing the target or non-target species.

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.

```
head(summary(Neonics$Species.Common.Name))
```

```
##            Honey Bee        Parasitic Wasp Buff Tailed Bumblebee 
##                  667                   285                   183 
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee 
##                  152                   140                   113
```

Answer: The six most commonly studied species in the dataset are all major pollinators. These species make sense to study over other insects not only because the important ecosytem service they provide through pollination, but also because pollination itself makes these species particularly vulnerable to neonicotinoids. Studies have found neonicotinoids in both pollen and nectar.

8. Concentrations are always a numeric value. What is the class of Conc.1..Author. in the dataset, and why is it not numeric?

```
class(Neonics$Conc.1..Author.)
```

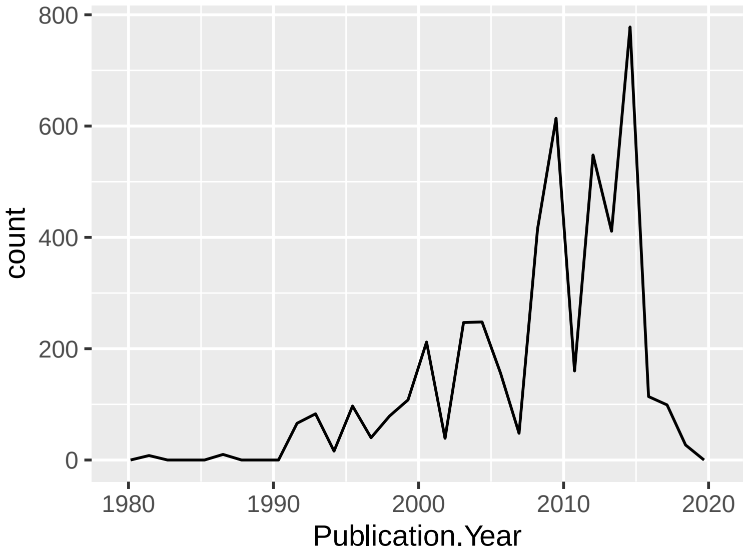
```
## [1] "factor"
```

Answer: Conc.1..Author represents the initial concentration of the chemical being studied, as observed by the author of the study. While concentrations are always numeric, R imported this data as factor or character data because the author concentrations contain symbols (e.g., ~, /, <, >) in addition to numbers. According to the metadata, these symbols allow the author to indicate an approximate concentration or signal a range in concentrations.

Explore your data graphically (Neonics)

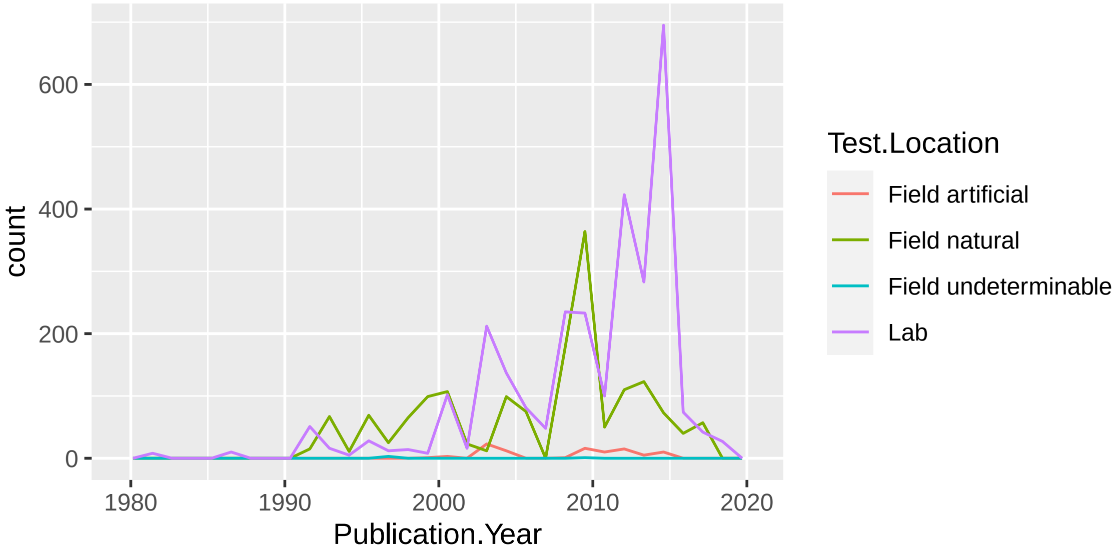
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 30)
```



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

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 30)
```

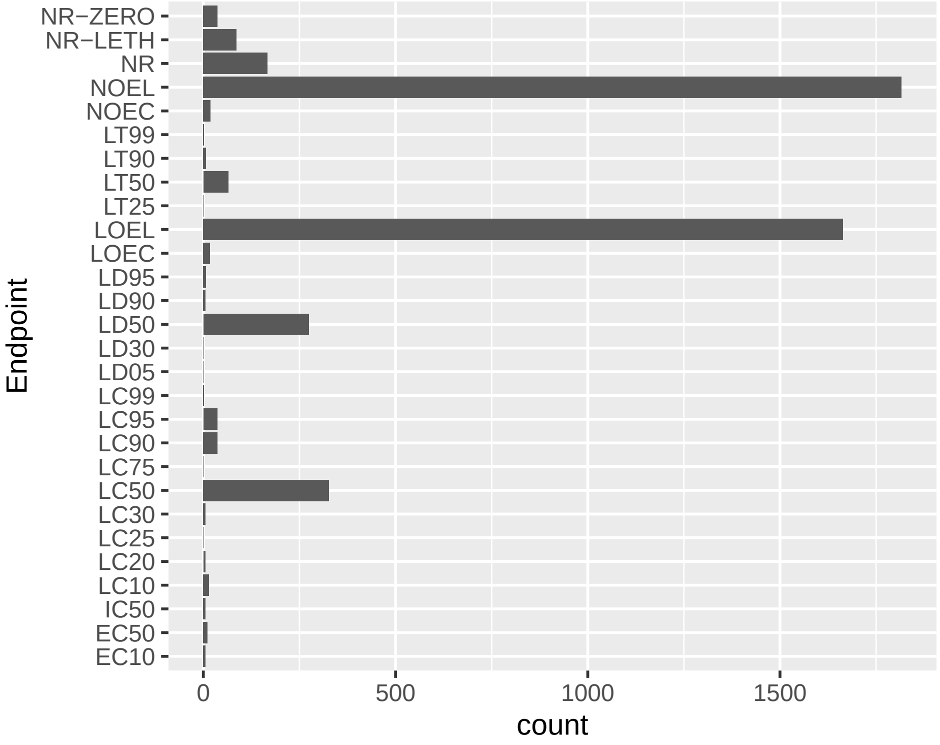


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test location is "Lab," followed "Field Natural." The frequency of lab studies generally increased since the early 1990s (shortly after neonicontinoids were developed). While there is similar trend in the frequency of natural field studies, there has fewer natural field studies in the past decade (perhaps due to changes in neonicontinoid use/regulation).

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.

```
ggplot(Neonics, aes(y = Endpoint)) +
  geom_bar()
```



Answer: The two most common end points are NOEL (no-observable-effect-level) and LOEL (lowest-observable-effect-level). NOEL is the highest dose concentration producing effects not significantly different from control responses and LOEL is the lowest dose concentration productimg significantly different effects from control. Both endpoints are determined by the author according to the author's statistical analysis.

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)
```

```
## [1] "character"
```

```
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
```

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

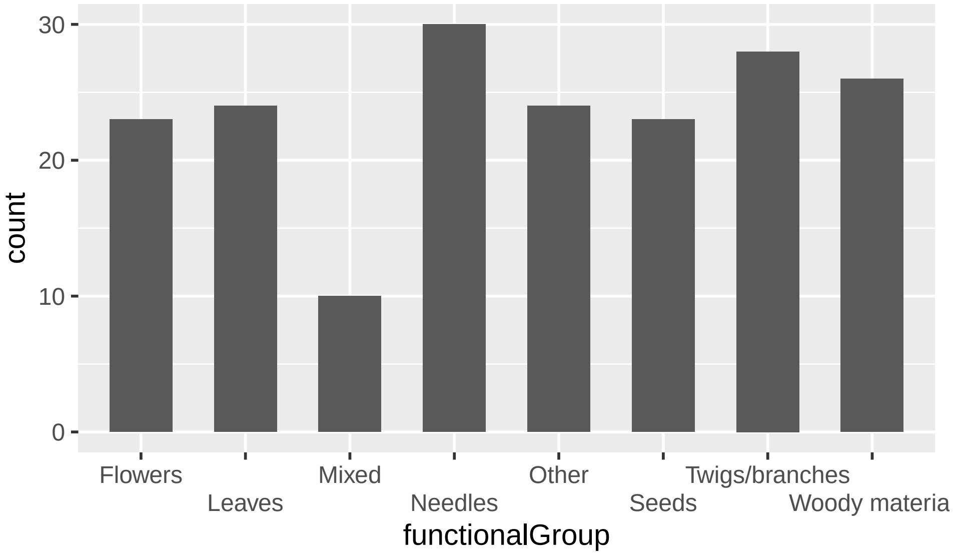
```
length(unique(Litter$plotID))
```

```
## [1] 12
```

Answer: Twelve plots were sampled at Niwot Ridge. The information obtained from unique is different from the summary function in that the unique functions removes duplicate rows. In contrast, summary analyzes all rows of data.

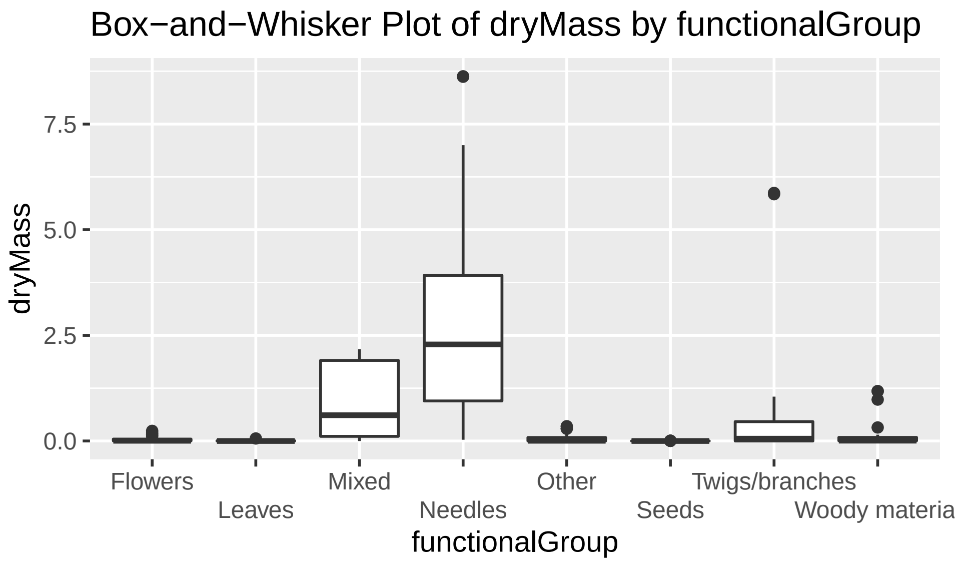
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar(width = .6) +
  scale_x_discrete(guide = guide_axis(n.dodge=2))
```



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

```
# Box-and-Whisker Plot of dryMass by Functional Group
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass)) +
  scale_x_discrete(guide = guide_axis(n.dodge=2)) +
  labs(title = "Box-and-Whisker Plot of dryMass by functionalGroup")
```

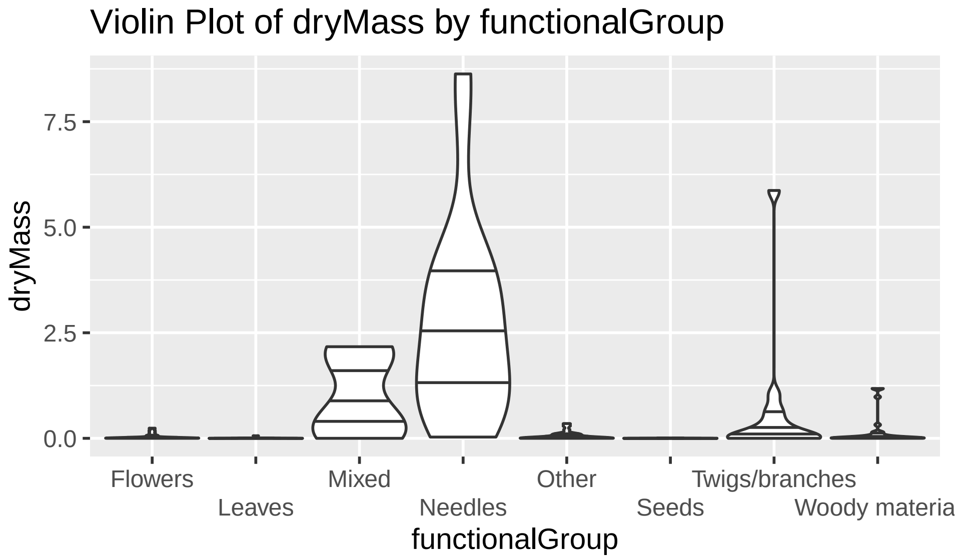


```
# Violin Plot of dryMass by functionalGroup
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass), scale = "width",
              draw_quantiles = c(0.25, 0.5, 0.75)) +
  scale_x_discrete(guide = guide_axis(n.dodge=2)) +
  labs(title = "Violin Plot of dryMass by functionalGroup")
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```



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

> Answer: A boxplot is more helpful here because it shows the summary statistics the numberical variable we are looking at – dry mass – more clearly. While the violin shapes contain density information, they weren't super helpful here unless scaled so that they were visible. Even then, it's very difficult to see what the mean dry mass of each functional group is on the violin plot. Additionally, the violin plot seems to be throwing an error about "collapsing non-unique x-values" and "missing ties" that most of the internet seems to think is a bug.

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

> Answer: Needles and mixed litter/debris have the highest biomass (potentially because mixed litter/debris contains a decent amount of needles). This makes sense since evergreen trees shed needles all-year round to some extent and needles generally take longer to decompose than decidulous matter.