Assignment 3: Data Exploration

Courtney Horn

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.

```
## [1] "/Users/courtneyannehorn/Desktop/EDA/EDAfin/Assignments"
```

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: We might want to learn more about the ecotoxicology of neonicotinoids so that we can gain a greater understanding of possible damage the neonicotinoids could cause to ecosystems and human health. Similarly, I researched the ecotoxicology of anticoagulant rodenticides to help a government agency gain a greater understanding of the threat they pose to Southern California predators.

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: The litter and woody debris that fall on the ground in the forest may affect the nutrient compositions of the soil. This in turn may influence the plant composition of the forest by determining which species can grow there. The litter and woody debris that fall to the ground in the forest also may influence the community of invertebrates in the soil.

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: * The litter and woody debris are sampled at terrestrial NEON sites that contain woody vegetation >2m tall. * Sampling occurs only in tower plots, which are selected randomly. * Ground traps are sampled once per year.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

   Answer: There are 4623 rows and 30 columns

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?

   Answer: Population and mortality are the effects that are studied most commonly. These might be of interest, because it is important to study both population and mortality to determine how well a population is doing. Here, we are likely trying to figure out how the insecticides are affecting insect populations

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.

   Answer: The most commonly studied species are Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. These species seem like they may all be pollinators. Pollinators may be of greater interest to the group of researchers than other insects are, because pollinators are important for human food sources.

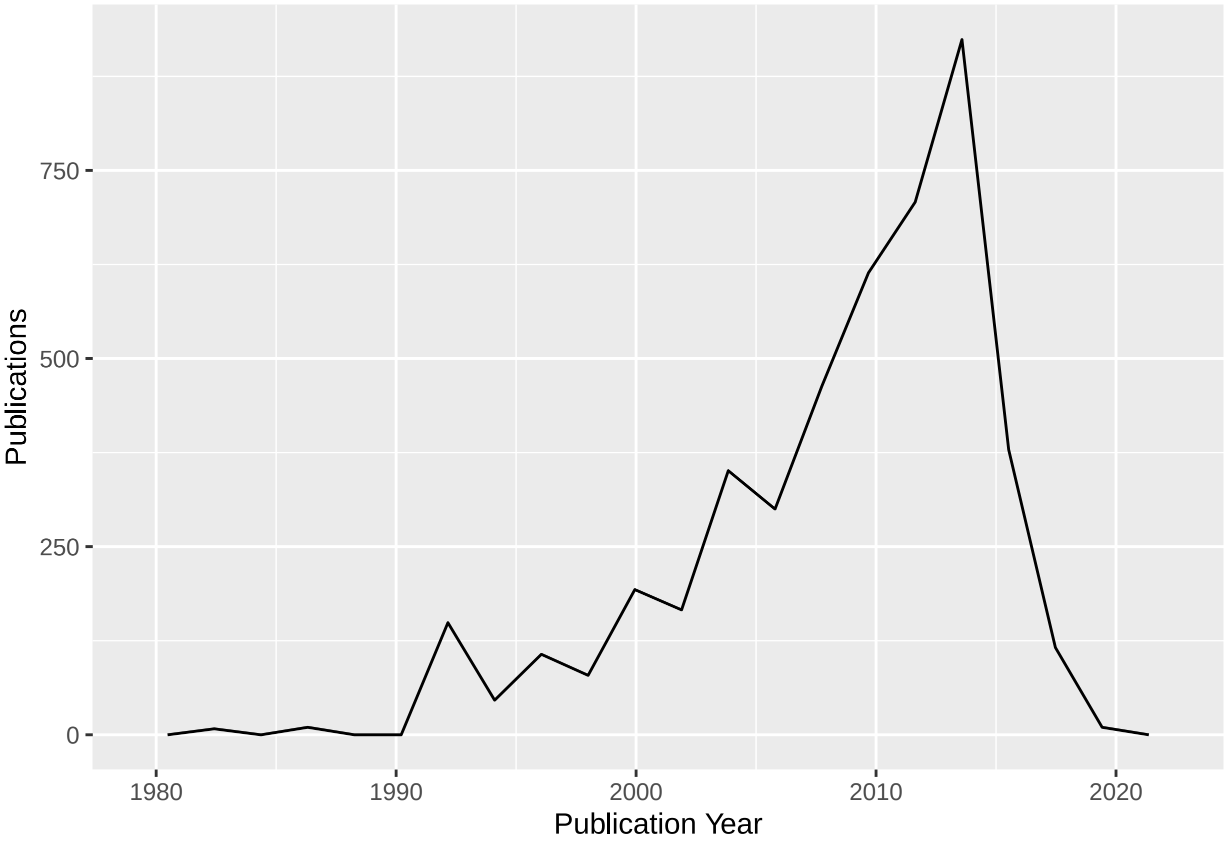
8. Concentrations are always a numeric value. What is the class of Conc.1..Author. in the dataset, and why is it not numeric?

   Answer: Conc.1..Author is a factor. These numbers refer to something (I beleive the numbers refer to who took the data). However, the numbers don't have any numerical meaning. They are just labels. Conc.1..Author is not a numeric variable because the values are just labels and don't have numerical meaning.

Explore your data graphically (Neonics)

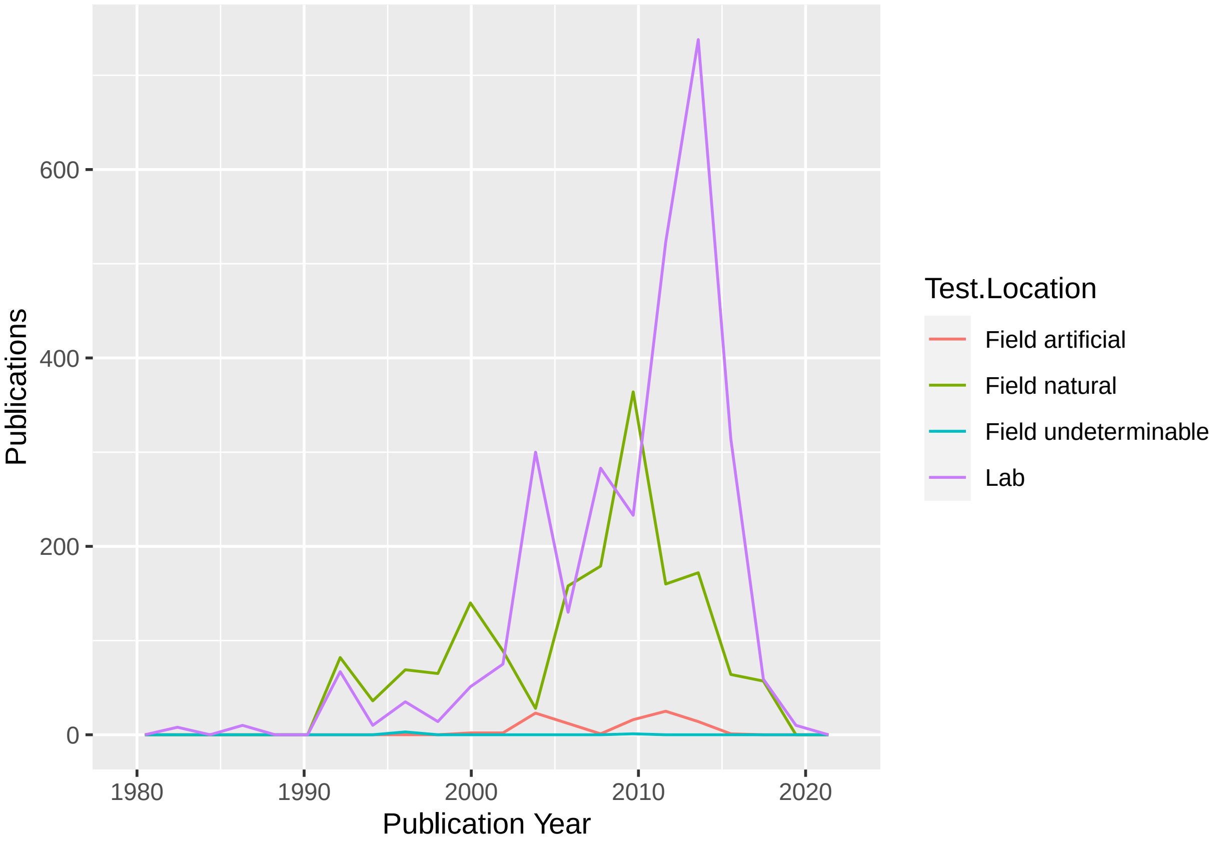
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
## Warning: Use of `Neonics$Publication.Year` is discouraged. Use
## `Publication.Year` instead.
```



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

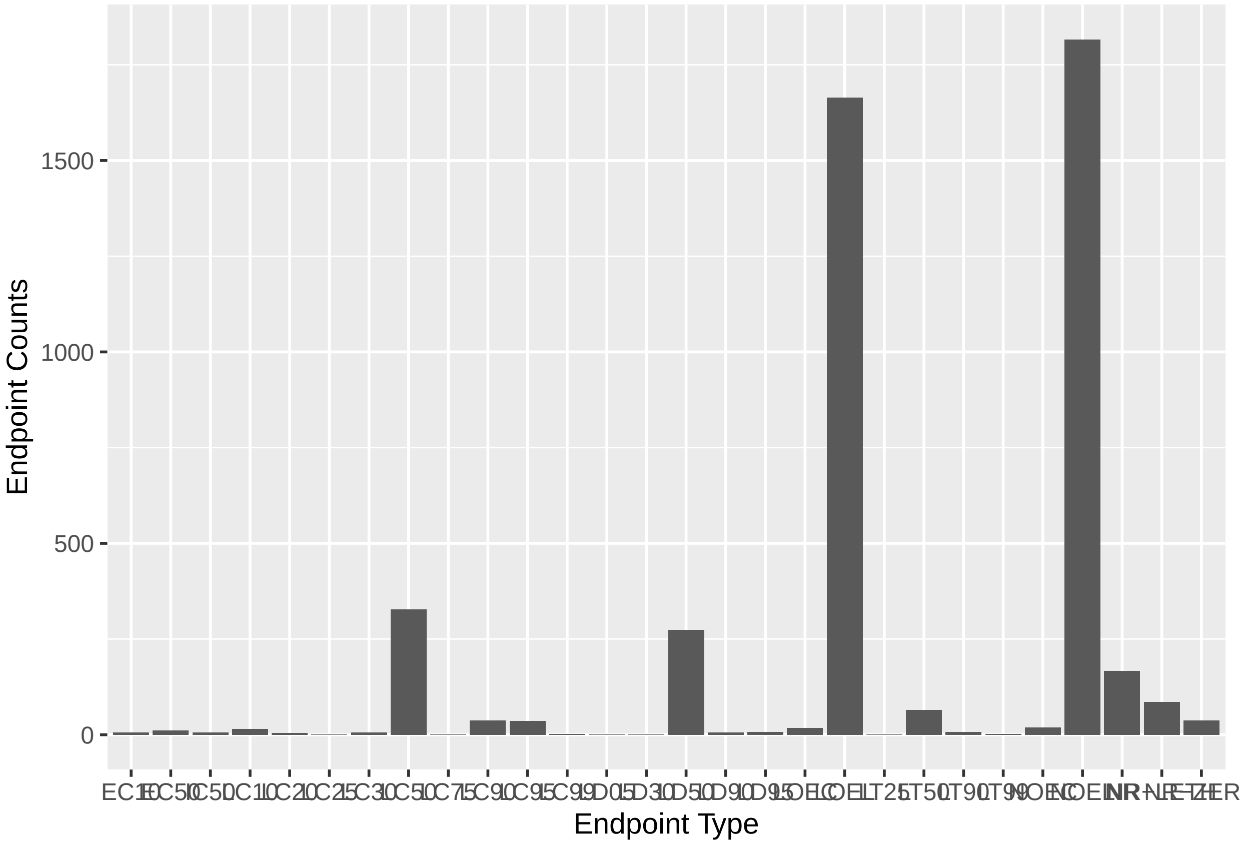
```
## Warning: Use of `Neonics$Publication.Year` is discouraged. Use
## `Publication.Year` instead.
```



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

> Answer: The lab and field sites for natural field studies are the most common test locations. The amount of Labs usage increases dramatically from about 2009 - 2015, and becomes much more common than use of field natural experiments during that time. Lab usage then dramatically decreased between 2015 and 2020.

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.



Answer: NOEL and LOEL are the two most common end points. NOEL stands for 'No-observable-effect-level'. This is the highest dose for which responses are not significantly different than the responses resulting from the control according to statistical analysis. NOEL is a statistically defined end point. LOEL stands for 'Lowest-Observable-effect-level'. This means the lowest concentration producing effects that were significantly different from the control.

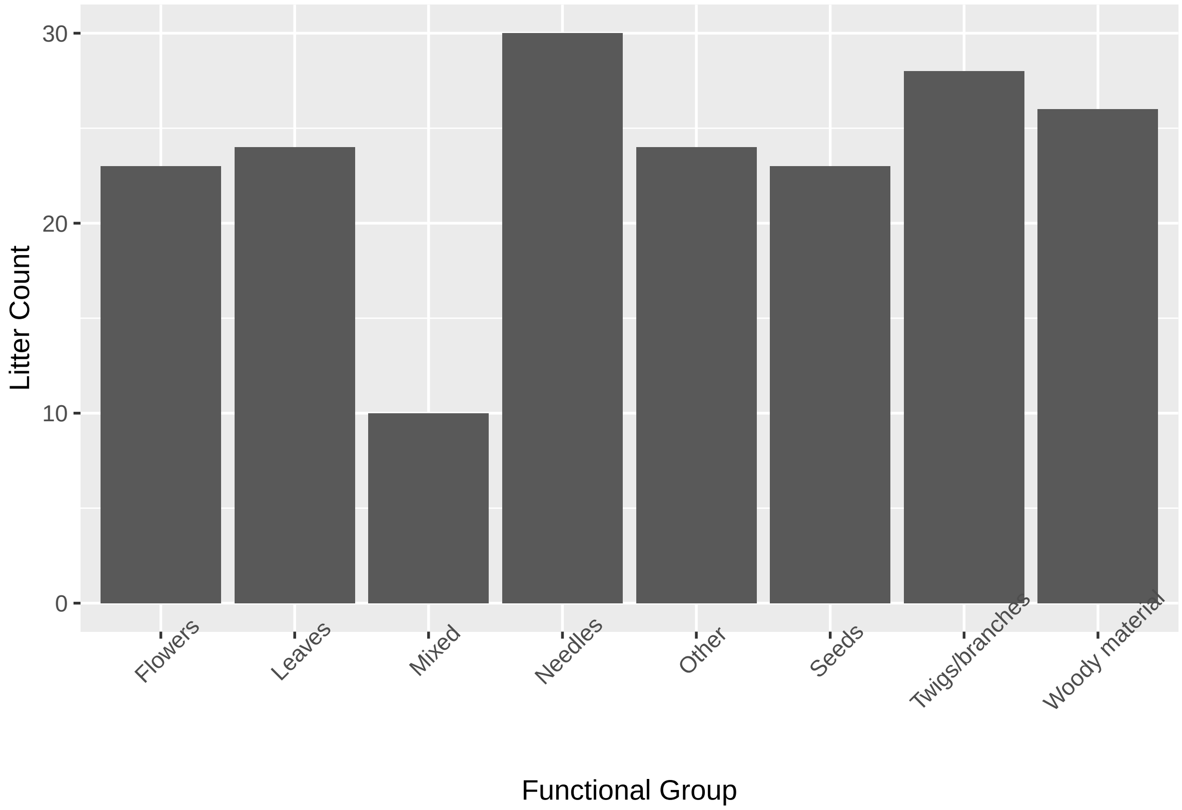
Explore your data (Litter)

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.

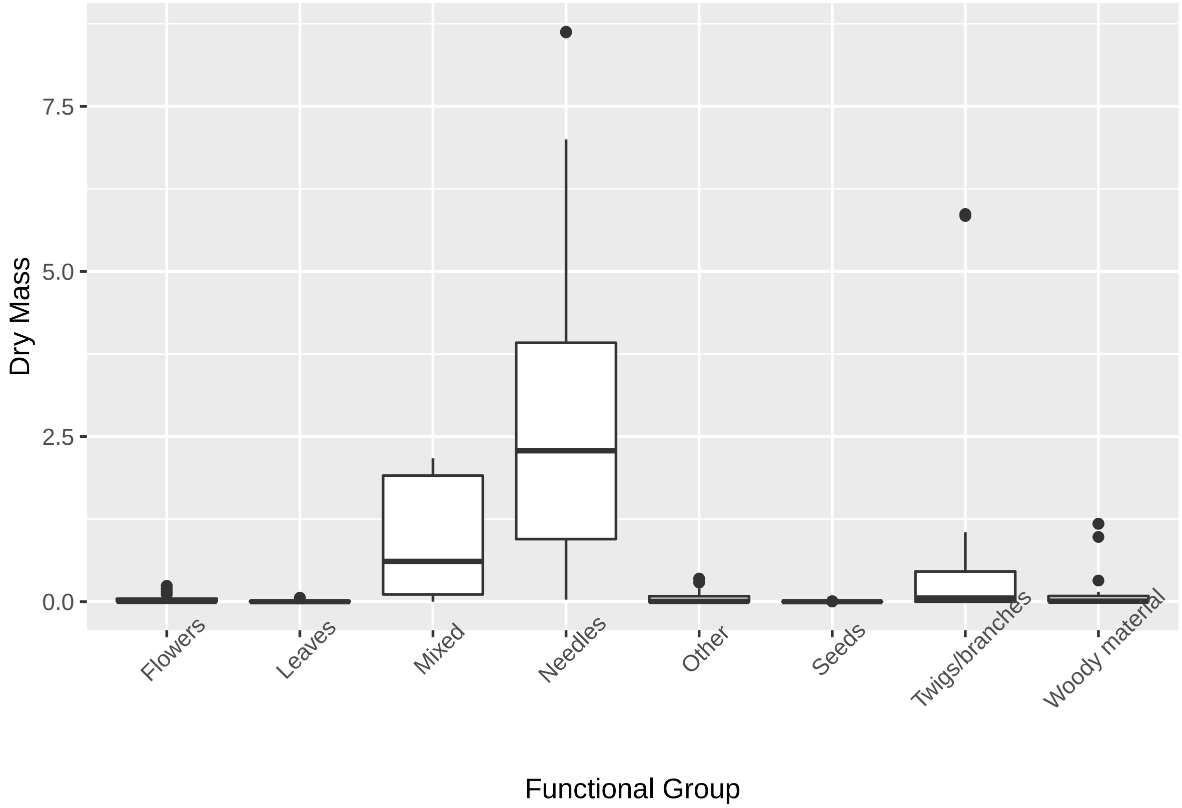
13. Using the `unique` function, determine how many plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

    Answer: 12 plots were sampled at Niwot Ridge. The unique function returns a vector, data frame, or array with duplicate elements or rows removed. Here, unique reported all of the levels of the plotID column. However, it didn't report how many times each plot was sampled. The summary function reported all of the values of the plotID column, and the number of times each plot was sampled.

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.



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

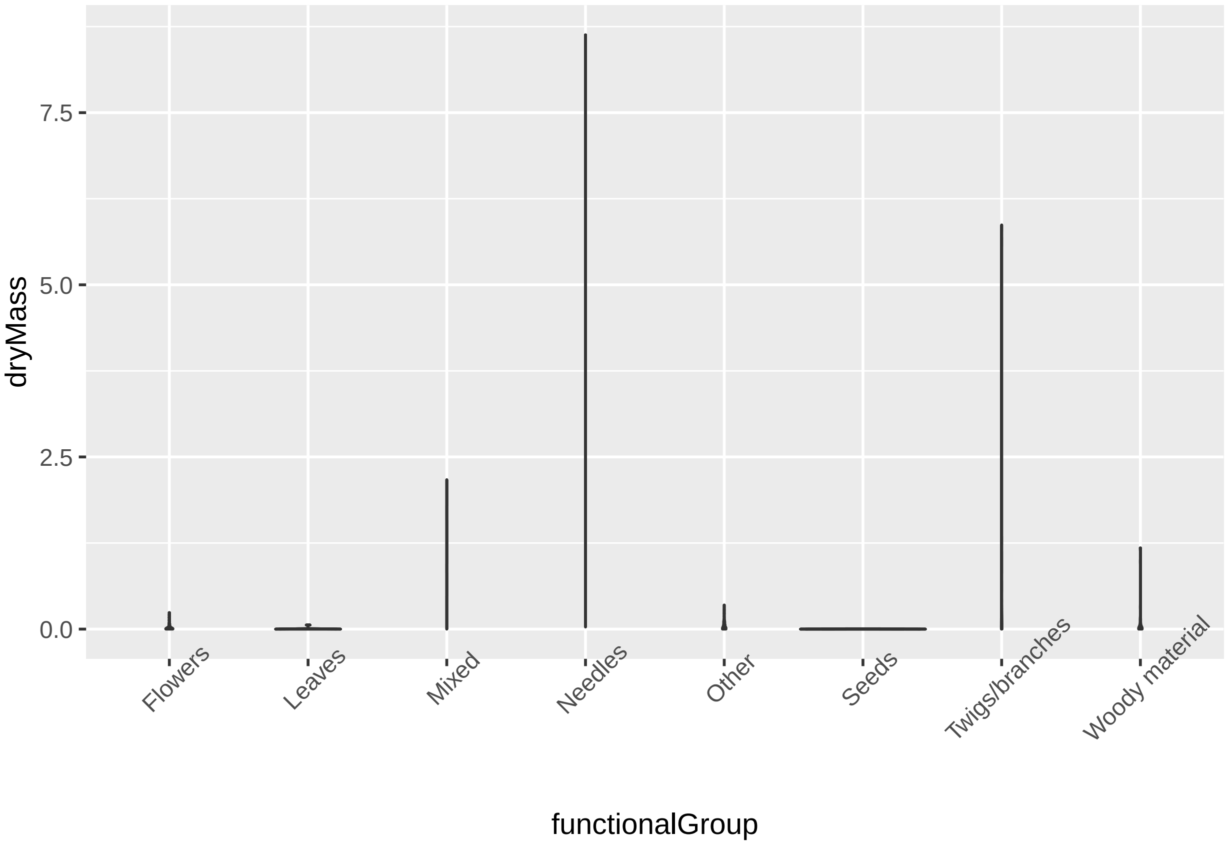


```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
```

```
## 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: Violin plots display density distributions. This doesn't work well here, because the dry mass from each of the functional groups varied substantially per collections.

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

> Answer: Needles and twigs/branches.

Appendix

```
  ##1 Set up

#setwd("/Users/lmm89/OneDrive/Duke_University/7_Spring2021/ENV872_EDA/GitRepo_EDA_S2021/Environmental_D
#setwd("/Users/courtneyannehorn/Desktop/EDA/EDAfin")
#tinytex::reinstall_tinytex()
library(tinytex)
#setwd("/Users/courtneyannehorn/Desktop/EDA/EDAfin")
getwd()
#setwd("/Users/courtneyannehorn/Desktop/EDA/EDAfin")
library(tidyverse)
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", stringsAsFactors = TRUE)

  ##5. Obtain basic summaries of your data
dim(Neonics)
nrow(Neonics)
ncol(Neonics)
#There are 4623 rows and 30 columns

  ##6
summary(Neonics$Effect)

  ##7
colnames(Neonics)
summary(Neonics$Species.Common.Name)

  ##8
class(Neonics$Conc.1..Author.)

  ##9
View(Neonics)
colnames(Neonics)
class(Neonics$Publication.Year)

#ggplot(na.omit(Neonics), aes(x = Neonics$Publication.Year), bins = 20))

#ggplot(Neonics) + geom_freqpoly(aes(x = na.omit(Neonics$Publication.Year)), bins = 50)

#ggplot(Neonics) + geom_freqpoly(aes(x = Neonics$Publication.Year), bins = 50)

#ggplot(na.omit(Neonics)) + geom_freqpoly(aes(x = Neonics$Publication.Year), bins = 20)

  #plot of studies conducted by publication year
ggplot(na.omit(Neonics)) + geom_freqpoly(aes(x = Neonics$Publication.Year), bins = 20) + labs(x="Publica

  ##10
#Neonics$Test.Location
ggplot(na.omit(Neonics)) + geom_freqpoly(aes(x = Neonics$Publication.Year, color = Test.Location), bins


  ##11
Neonics$Endpoint
summary(Neonics$Endpoint)
endpointsum <- summary(Neonics$Endpoint)
endpointsum
sort(endpointsum)

  #bar graph code
ggplot(na.omit(Neonics), aes(x = Endpoint)) +
  geom_bar() + labs(x="Endpoint Type", y ="Endpoint Counts")


  ##12
View(Litter)
```

```r
colnames(Litter)
str(Litter)
dim(Litter)

  ##13
unique(Litter$plotID)
summary(Litter$plotID)

  ##14
  #bar graph of functional group counts
litter_type_bar <- ggplot(na.omit(Litter), aes(x = Litter$functionalGroup)) + geom_bar() + labs(x="Func

litter_type_bar + theme(axis.text.x = element_text(angle = 45))

  ##15
#boxplot of dry mass by functional group

funct_group_mass <- ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass)) + labs(x="Functional Group", y ="Dry Mass")

funct_group_mass + theme(axis.text.x = element_text(angle = 45))

  #Violin plot of dry mass by functional group

ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75))

litter_violin <- ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75))

litter_violin + theme(axis.text.x = element_text(angle = 45))
unique(Litter$functionalGroup)
twig <- ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75))
```