Assignment 3: Data Exploration

Emma Brentjens

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. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.
6. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai.

The completed exercise is due on Sept 30th.

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. Be sure to include the subcommand to read strings in as factors.

```
# 1.  checking working directory
getwd()
```

```
## [1] "/home/guest/R/EDA-Fall2022/Data/Raw"
```

```
## loading packages
library(tidyverse)
library(ggplot2)

## uploading and naming datasets
Neonics <- read.csv("ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = T)
Litter <- read.csv("NEON_NIWO_Litter_massdata_2018-08_raw.csv", stringsAsFactors = T)
```

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: This research can provide information on the efficacy of neonicotinoids to protect crops from insect pests. It is important to understand which doses are effective as using too little of the

insecticide would not produce the desired effect but using too much may be harmful to nontarget species, like bees (Texas A&M AgriLife Extension, n.d.).
Citation: Texas A&M AgriLife Extension. (n.d.). What is a neonicotinoid? *Insects in the City*. Retrieved September 26, 2022, from https://citybugs.tamu.edu/factsheets/ipm/what-is-a-neonicotinoid/

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 amount of leaf litter and debris on the ground can influence soil moisture and nutrient cycling. As leaf litter decomposes, it is broken down into compounds that plants can absorb as they grow (Giweta, 2020). Leaf litter types can also provide information about plant composition in ecosystems (Giweta, 2020), and serve as habitat for insects.
   Citation: Giweta, M. (2020). Role of litter production and its decomposition, and factors affecting the processes in a tropical forest ecosystem: A review. *Journal of Ecology and Environment*, 44(1), 11. https://doi.org/10.1186/s41610-020-0151-2

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. Researchers place one elevated and one ground litter trap for each 400m^2 section of plot.
   2. Traps are placed randomly in plots with over 50% aerial vegetation cover while in sites with less than 50% aerial vegetation cover, traps are placed in specific locations near the vegetation of interest.
   3. Researchers sample ground traps once each year, while the elevated trap sampling varies by vegetation type (more frequent in deciduous forests when trees are shedding their leaves than in evergreen forests).

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# 5.  finding number of rows and columns
nrow(Neonics)
```

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

```
ncol(Neonics)
```

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

```
## The 'Neonics' dataset has 30 columns and 4623 rows/observations
```

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?

```
# 6.  obtaining summary of effects studied
summary(Neonics$Effect)
```

```
##     Accumulation        Avoidance         Behavior     Biochemistry
##               12              102              360               11
##          Cell(s)      Development        Enzyme(s) Feeding behavior
##                9              136               62              255
##         Genetics           Growth        Histology       Hormone(s)
```

```
##              82              38               5               1
##  Immunological    Intoxication     Morphology       Mortality
##              16              12              22            1493
##      Physiology      Population  Reproduction
##               7            1803             197
```

Answer:
Most common effects studied:
1. Population (1803)
2. Mortality (1493)
3. Behavior (360)

The researchers are interested in these effects because they are important measures of how the insecticides affect insects. Population abundance and mortality demonstrate how insecticides impact insect viability, which is relevant to understanding the efficacy of the insecticides in reducing insect populations and unintended effects on nontarget species. Insect behavior could include important life history activities like feeding and mating, which can have a more gradual effect on 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.

```
# 7.  obtaining summary statistics of species studied
summary(Neonics$Species.Common.Name)
```

```
##                      Honey Bee                  Parasitic Wasp
##                            667                             285
##             Buff Tailed Bumblebee             Carniolan Honey Bee
##                            183                             152
##                     Bumble Bee                Italian Honeybee
##                            140                             113
##                Japanese Beetle              Asian Lady Beetle
##                             94                              76
##                 Euonymus Scale                       Wireworm
##                             75                              69
##               European Dark Bee              Minute Pirate Bug
##                             66                              62
##            Asian Citrus Psyllid                  Parastic Wasp
##                             60                              58
##          Colorado Potato Beetle                Parasitoid Wasp
##                             57                              51
##            Erythrina Gall Wasp                  Beetle Order
##                             49                              47
##     Snout Beetle Family, Weevil        Sevenspotted Lady Beetle
##                             47                              46
##                  True Bug Order           Buff-tailed Bumblebee
##                             45                              39
##                   Aphid Family                Cabbage Looper
##                             38                              38
##            Sweetpotato Whitefly                 Braconid Wasp
##                             37                              33
##                   Cotton Aphid                Predatory Mite
##                             33                              33
##          Ladybird Beetle Family                   Parasitoid
##                             30                              30
##                   Scarab Beetle                 Spring Tiphia
```

```
##                              29                              29
##                    Thrip Order            Ground Beetle Family
##                              29                              27
##               Rove Beetle Family                  Tobacco Aphid
##                              27                              27
##                    Chalcid Wasp          Convergent Lady Beetle
##                              25                              25
##                   Stingless Bee               Spider/Mite Class
##                              25                              24
##              Tobacco Flea Beetle               Citrus Leafminer
##                              24                              23
##                  Ladybird Beetle                     Mason Bee
##                              23                              22
##                         Mosquito                 Argentine Ant
##                              22                              21
##                          Beetle       Flatheaded Appletree Borer
##                              21                              20
##              Horned Oak Gall Wasp              Leaf Beetle Family
##                              20                              20
##                Potato Leafhopper       Tooth-necked Fungus Beetle
##                              20                              20
##                   Codling Moth        Black-spotted Lady Beetle
##                              19                              18
##                   Calico Scale             Fairyfly Parasitoid
##                              18                              18
##                    Lady Beetle          Minute Parasitic Wasps
##                              18                              18
##                      Mirid Bug                Mulberry Pyralid
##                              18                              18
##                       Silkworm                 Vedalia Beetle
##                              18                              18
##             Araneoid Spider Order                       Bee Order
##                              17                              17
##                  Egg Parasitoid                    Insect Class
##                              17                              17
##          Moth And Butterfly Order     Oystershell Scale Parasitoid
##                              17                              17
## Hemlock Woolly Adelgid Lady Beetle           Hemlock Wooly Adelgid
##                              16                              16
##                            Mite                     Onion Thrip
##                              16                              16
##             Western Flower Thrips                    Corn Earworm
##                              15                              14
##                Green Peach Aphid                       House Fly
##                              14                              14
##                        Ox Beetle                Red Scale Parasite
##                              14                              14
##                Spined Soldier Bug             Armoured Scale Family
##                              14                              13
##                 Diamondback Moth                   Eulophid Wasp
##                              13                              13
##                Monarch Butterfly                   Predatory Bug
##                              13                              13
##             Yellow Fever Mosquito              Braconid Parasitoid
```

```
##                             13                                 12
##                  Common Thrip       Eastern Subterranean Termite
##                             12                                 12
##                         Jassid                         Mite Order
##                             12                                 12
##                      Pea Aphid                  Pond Wolf Spider
##                             12                                 12
##       Spotless Ladybird Beetle            Glasshouse Potato Wasp
##                             11                                 10
##                      Lacewing           Southern House Mosquito
##                             10                                 10
##        Two Spotted Lady Beetle                        Ant Family
##                             10                                  9
##                  Apple Maggot                           (Other)
##                              9                                670
```

Answer:
Most common species studied:
1. Honey Bee (667)
2. Parasitic Wasp (285)
3. Buff Tailed Bumblebee (183)
4. Carniolan Honey Bee (152)
5. Bumble Bee (140)
6. Italian Honeybee (113)

All of these species are pollinators, which is likely why researchers are most interested in them. Pollinators are critical to maintaining plant populations, so negative effects of insecticides on these species are highly concerning.

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

```
# 8.  determining class of 'Conc.1..Author' column
class(Neonics$Conc.1..Author.)
```

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

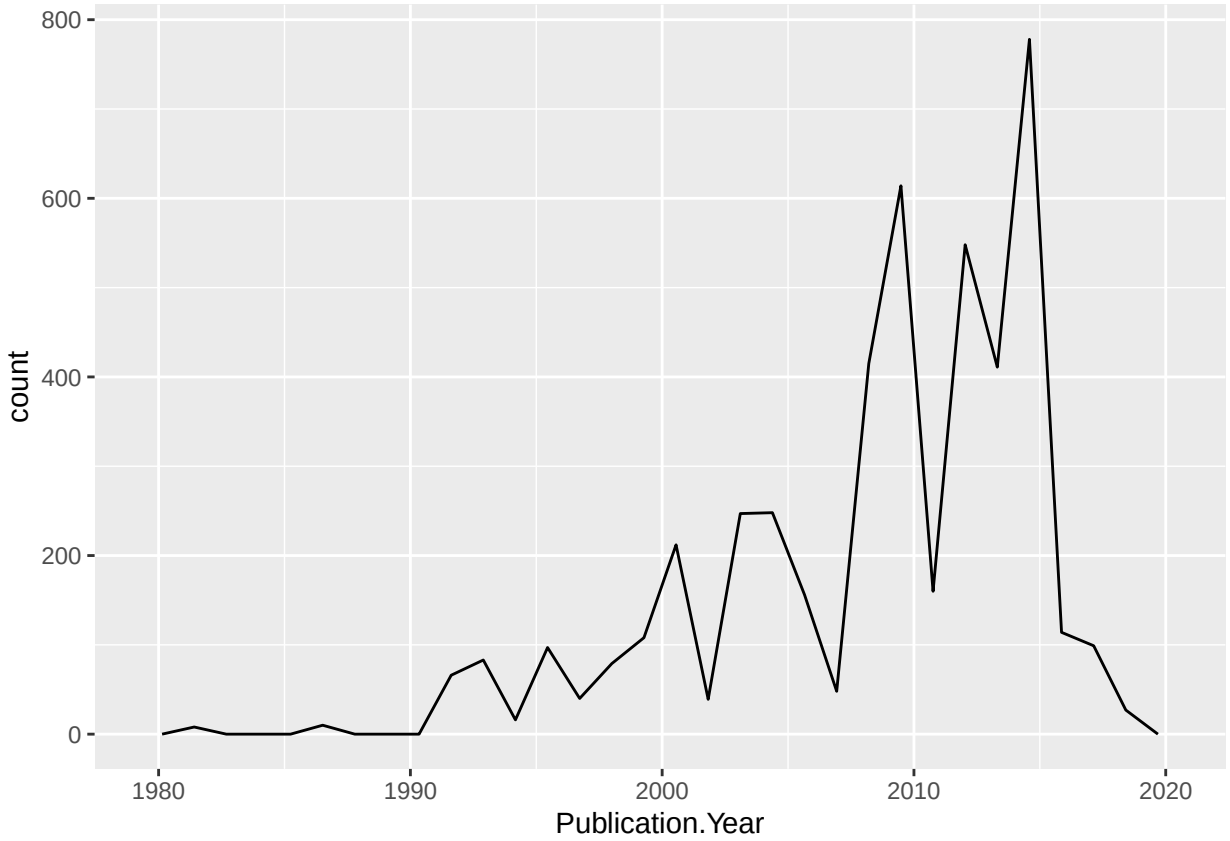
Answer: The class of "Conc.1..Author" is "factor." The column does not have all numbers, some are NR which means not recorded. Since there are characters in the column, R recorded this as a string. When loading the dataset, we told R to read strings as factors.

Explore your data graphically (Neonics)

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

```
# 9.  creating plot of studies by publication year
studies_by_year <- ggplot(data = Neonics, aes(x = Publication.Year)) + geom_freqpoly()

## output for studies by year plot
studies_by_year
```

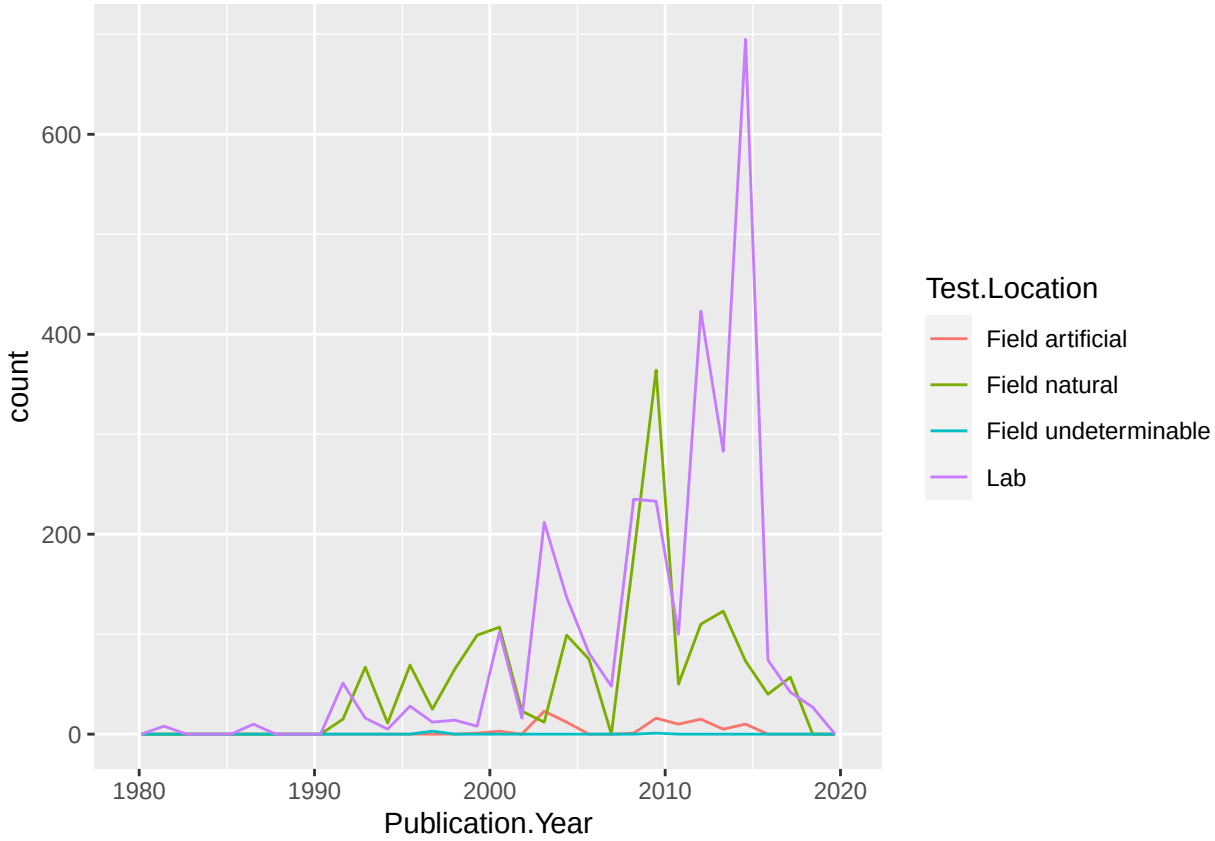
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
# 10.  creating plot of studies by publication year and test location
studies_by_year_and_location <- ggplot(data = Neonics, aes(x = Publication.Year,
    color = Test.Location)) + geom_freqpoly()

## output for studies by year and test location plot
studies_by_year_and_location
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

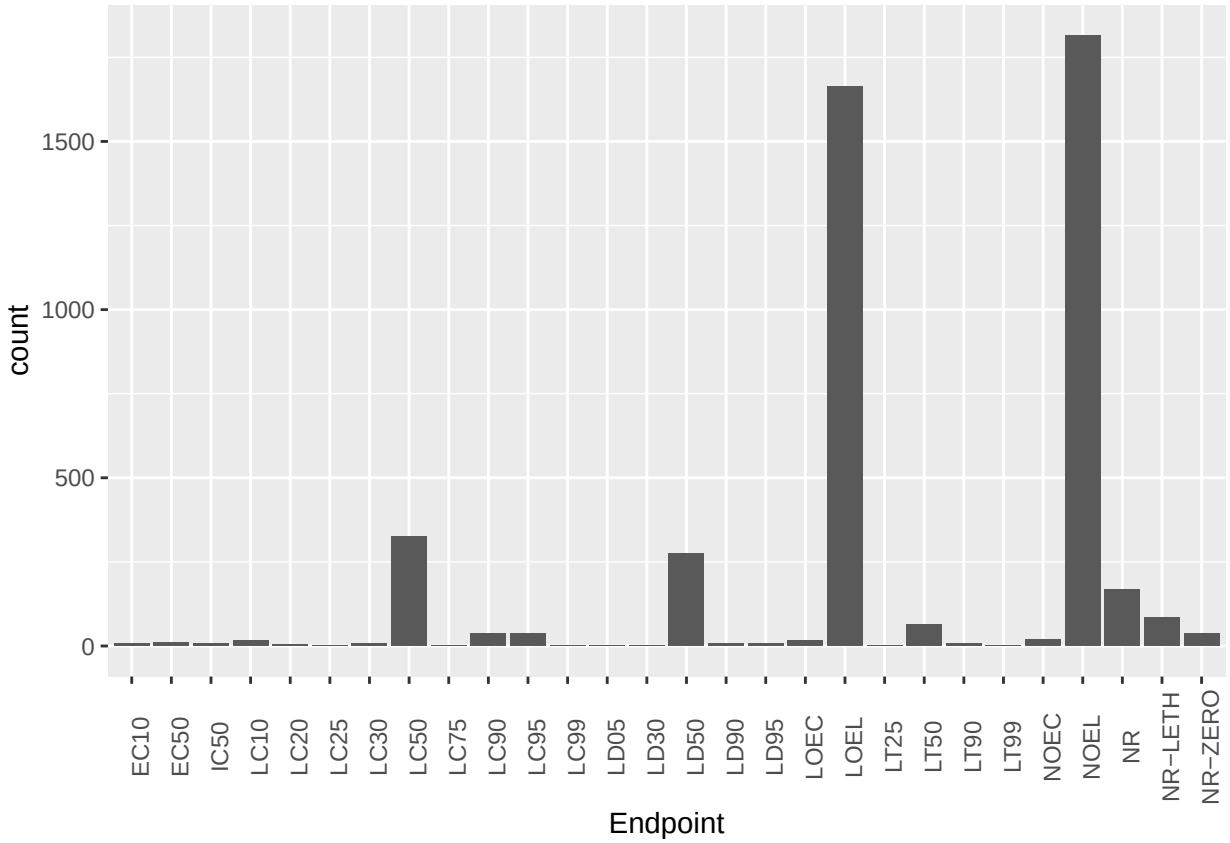


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are labs and natural fields. The number of lab studies generally seems to have increased over time (before 2020) while the number of natural field studies peaked at around 2010 and has declined since then (perhaps due to increasing popularity of lab tests).

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.

```
# 11.  creating bar graph for endpoint counts
endpoint_counts_graph <- ggplot(data = Neonics, aes(x = Endpoint)) + geom_bar() +
    theme(axis.text.x = element_text(angle = 90))

## output for endpoint counts graph
endpoint_counts_graph
```



Answer: The two most common endpoints are NOEL (no-observable-effect-level) and LOEL (lowest-observable-effect-level). The NOEL is defined as the greatest concentration of chemical that does not cause an effect significantly different than the control. The LOEL refers to the lowest chemical concentration that causes an effect that varies significantly 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.

```
# 12.  determining the class of the litter collect date
class(Litter$collectDate)  ##class = 'factor'
```

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

```
## Converting collectDate from factor to date
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")

## determining new class of collectDate
class(Litter$collectDate)
```

```
## [1] "Date"
```

```
## Determine August 2018 sampling days
unique(Litter$collectDate)
```

```
## [1] "2018-08-02" "2018-08-30"
```

```
## The days sampled in August 2018 were August 2nd and August 30th
```

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`?

```
# 13.  determining number of plots sampled at Niwot Ridge
length(unique(Litter$plotID))
```

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

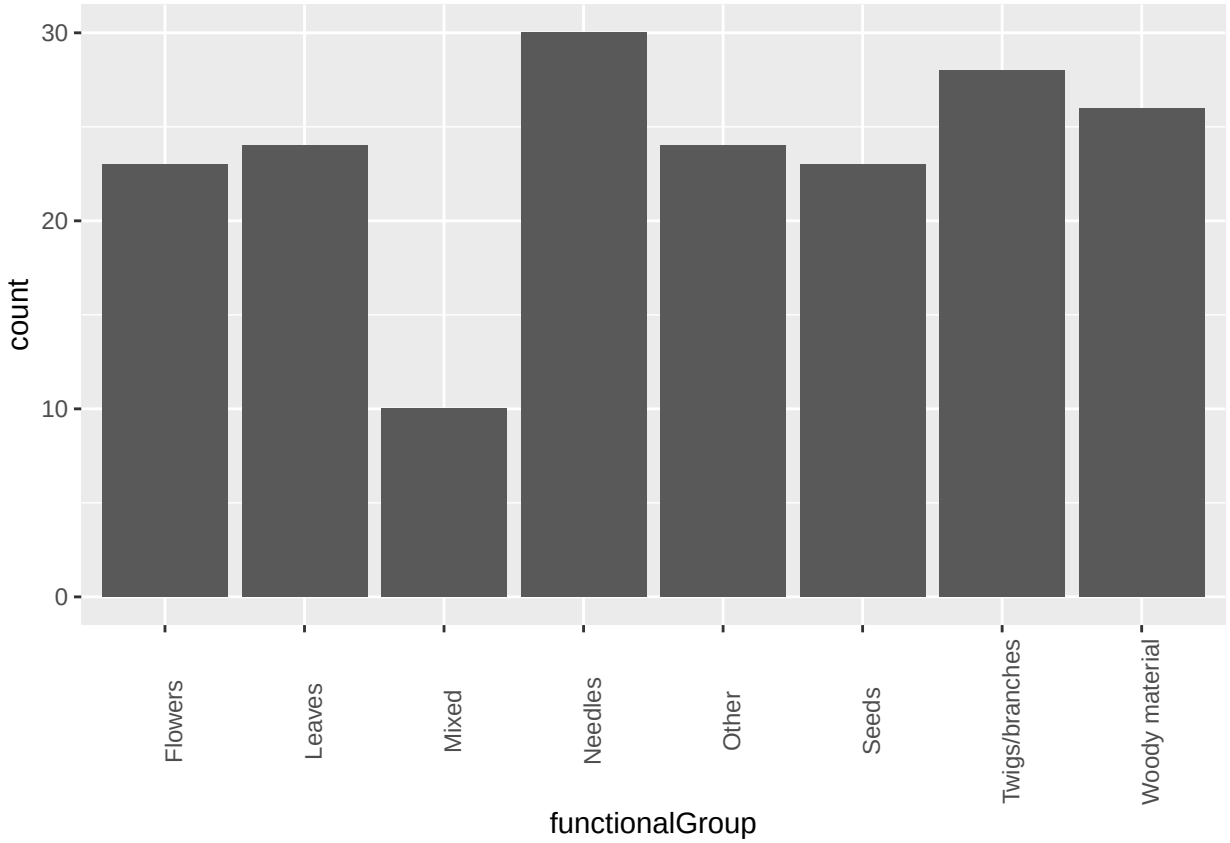
```
## running summary command on plots
summary(Litter$plotID)
```

```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

Answer: There were 12 plots sampled at Niwot Ridge. A summary of Litter$plotID gives you the number of observations at each plot while the unique function provides the number of plots studied.

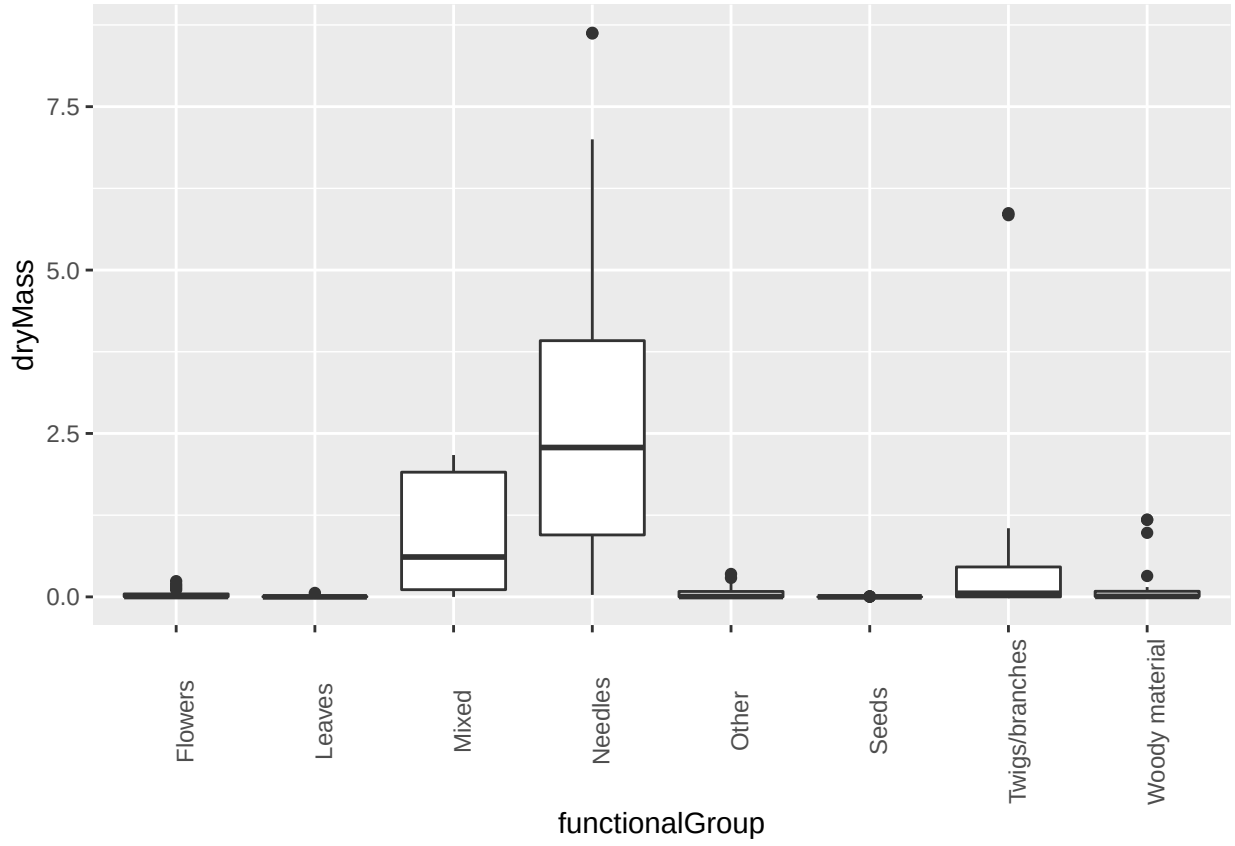
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
# 14.  creating functionalGroup counts bar graph
functionalGroup_bar_graph <- ggplot(data = Litter, aes(x = functionalGroup)) + geom_bar() +
    theme(axis.text.x = element_text(angle = 90))

## output for functionalGroup counts bar graph
functionalGroup_bar_graph
```

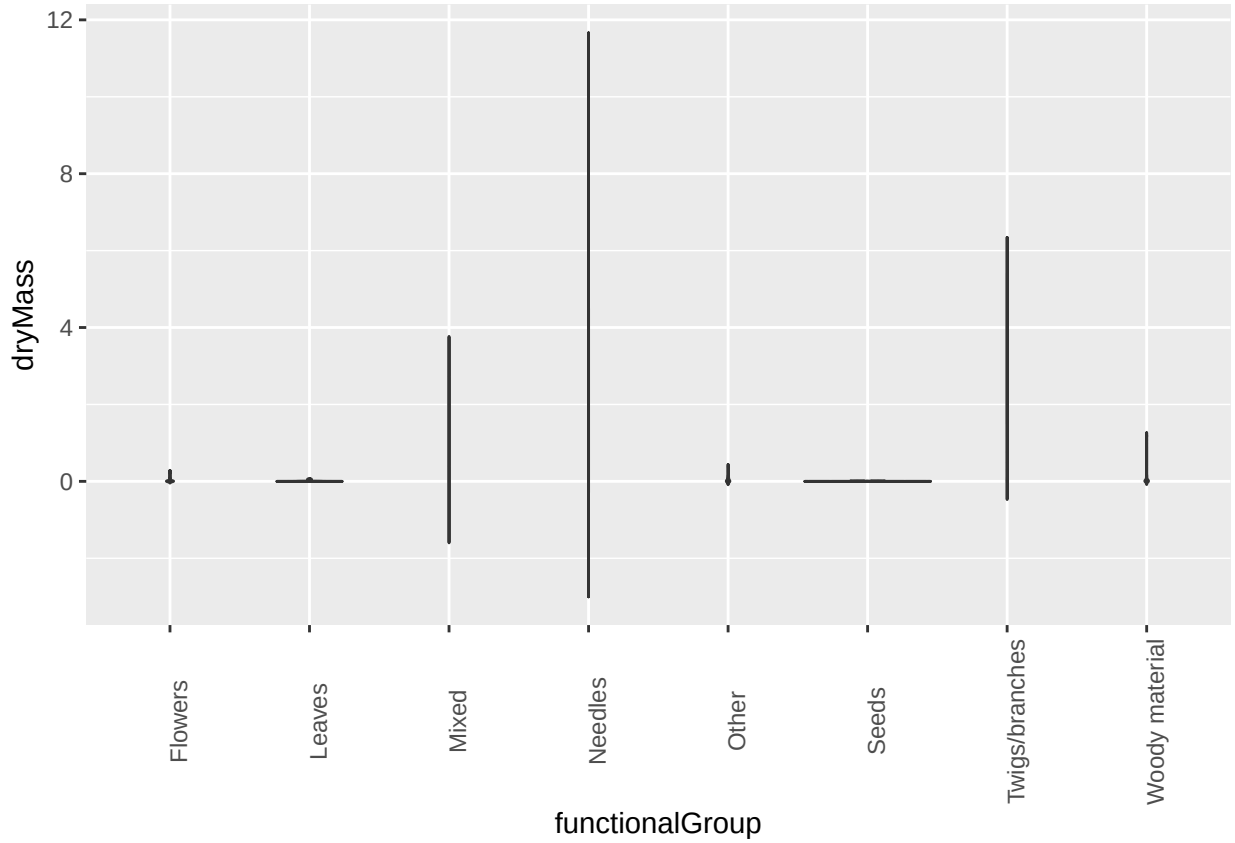


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# 15.  functionalGroup boxplot
functionalGroup_boxplot <- ggplot(data = Litter, aes(x = functionalGroup, y = dryMass)) +
    geom_boxplot() + theme(axis.text.x = element_text(angle = 90))

## boxplot output
functionalGroup_boxplot
```




```
## functionalGroup violin plot
functionalGroup_violin <- ggplot(data = Litter, aes(x = functionalGroup, y = dryMass)) +
    geom_violin(trim = F) + theme(axis.text.x = element_text(angle = 90))

## violin plot output
functionalGroup_violin
```



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

> Answer: Dry mass is a continuous variable, so the violin plot doesn't give us much information because there are not multiple observations with the exact same dry mass, except at zero. Thus, the boxplot is a better way of visualizing the data because it only shows the range of values, not the frequency of observations at different values.

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

> Answer: The type of litter with the highest biomass is needles, followed by mixed litter.