Assignment 3: Data Exploration

Lily Zhang

Spring 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 Sakai.

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

```
#get current working directory with the absolute filepath and set the working directory
getwd()
```

```
## [1] "/Users/lilyzhang/Desktop/EDA_Spring2024/Assignments"
```

```
setwd("/Users/lilyzhang/Desktop/EDA_Spring2024/Data/Raw")

#check and load tidyverse and lubridate packages
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.1
## Warning: package 'readr' was built under R version 4.3.1
## Warning: package 'dplyr' was built under R version 4.3.1
## Warning: package 'lubridate' was built under R version 4.3.1
```

```r
library(lubridate)

#upload two datasets and assign names as "Neonics" and "Litter" respectively
Neonics <- read.csv("ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE)
Litter <- read.csv("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: First, neonicotinoids are a class of synthetic, neurotoxic insecticides that function as agonists at the insect nicotine acetylcholine receptor (nAChRs) in neurons. The nAChRs are a common Na+/K+ pathway in insects neuron, which suggests that neonicotinoid may have indiscriminate effect on exposed targets beyond pests, including butterflies and bees. Second, neonicotinoid have the potential for bioaccumulation, posing a potential threat to both agriculture and the surrounding ecosystem dynamics. Understanding the ecotoxicology of neonicotinoids on insects can help assess their impact on insect populations, environmental fate, and overall ecosystem health. This knowledge contributes to improvement of regulations as well as conservation efforts. Reference: Steve M. Ensley, Chapter 40 - Neonicotinoids, Editor: Ramesh C. Gupta. Veterinary Toxicology (Third Edition), Academic Press, 2018,Pages 521-524.https://doi.org/10.1016/B978-0-12-811410-0.00040-4.

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, primarily composed of foliage, along with woody debris, constitutes a vital component in both aquatic and terrestrial ecosystems due to its rich organic content. The accumulation and decomposition of these organic materials integrate food webs and play a pivotal role in nutrient cycling within forest ecosystems. In addition, litter and woody debris serve as valuable resources for diverse biotas, including fungi, invertebrates, and vertebrates in some cases. Information on the presence, characteristics, and spatial-temporal dynamics of litter and woody debris inform us the resilience and biodiversity of these environments, thereby supporting ecological management and conservation efforts.

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. Litter and fine woody debris sampling at terrestrial NEON sites with woody vegetation exceeding 2 meters takes place in randomly selected tower plots within the 90% flux footprint of airsheds. 2. The number of plots is limited by spacing requirements, and trap placement may be targeted or randomized based on vegetation cover. 3. Ground traps are sampled annually and elevated trap frequency varies by vegetation: 1x every 2 weeks during deciduous forest senescence and 1x every 1-2 months in evergreen sites.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset? > Answer: Neonics has 4623 rows and 30 columns

```r
#Retrieve and print the dimensions of the Neonics dataset
Neonics_dimensions <- dim(Neonics)
print(Neonics_dimensions)
```

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

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?

```r
#summarize the "Effect" column in Neonics dataset,
#arrange it in a decreasing order, and print the result
effect_summary <- summary(Neonics$Effect)
decreasing_effect_summary <- sort(effect_summary, decreasing = TRUE)
print(decreasing_effect_summary)
```

```
##       Population        Mortality         Behavior Feeding behavior
##             1803             1493              360              255
##     Reproduction      Development        Avoidance         Genetics
##              197              136              102               82
##        Enzyme(s)           Growth       Morphology    Immunological
##               62               38               22               16
##     Accumulation     Intoxication     Biochemistry          Cell(s)
##               12               12               11                9
##       Physiology        Histology       Hormone(s)
##                7                5                1
```

Answer: The most prevalent effect studied is population effects, which, In contrast to individual-level impacts, offer a more representative and comprehensive understanding of real-world scenarios, particularly in agricultural settings. This focus is vital for assessing potential disruptions to ecosystem functioning, and further contributing to long-term ecological risk assessments. In addition, mortality is also a common area of study due to the direct interactions of neonicotinoid. These studies not only reveal potential pathways of absorption, distribution, metabolism, and excretion but also provide meaningful data, including LD50 values, essential for the production, application and regulation of such insecticides.

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: The `sort()` command can sort the output of the summary command...]

```r
#summarize the "Speices.Common.Name" column in Neonics dataset,
#arrange it in a decreasing order, and print the result
commonname_summay <- summary(Neonics$Species.Common.Name)
decreasing_commonname_summary <- sort(commonname_summay, decreasing = TRUE)
print(decreasing_commonname_summary)
```

```
##                   (Other)                Honey Bee
##                       670                      667
##             Parasitic Wasp      Buff Tailed Bumblebee
##                       285                      183
##          Carniolan Honey Bee               Bumble Bee
##                       152                      140
##             Italian Honeybee           Japanese Beetle
##                       113                       94
##           Asian Lady Beetle            Euonymus Scale
##                        76                       75
##                  Wireworm         European Dark Bee
```

```
##                          69                          66
##          Minute Pirate Bug          Asian Citrus Psyllid
##                          62                          60
##              Parastic Wasp       Colorado Potato Beetle
##                          58                          57
##           Parasitoid Wasp          Erythrina Gall Wasp
##                          51                          49
##              Beetle Order  Snout Beetle Family, Weevil
##                          47                          47
##     Sevenspotted Lady Beetle              True Bug Order
##                          46                          45
##        Buff-tailed Bumblebee               Aphid Family
##                          39                          38
##              Cabbage Looper         Sweetpotato Whitefly
##                          38                          37
##              Braconid Wasp               Cotton Aphid
##                          33                          33
##              Predatory Mite       Ladybird Beetle Family
##                          33                          30
##                 Parasitoid               Scarab Beetle
##                          30                          29
##              Spring Tiphia                Thrip Order
##                          29                          29
##        Ground Beetle Family          Rove Beetle Family
##                          27                          27
##              Tobacco Aphid               Chalcid Wasp
##                          27                          25
##      Convergent Lady Beetle             Stingless Bee
##                          25                          25
##           Spider/Mite Class          Tobacco Flea Beetle
##                          24                          24
##            Citrus Leafminer           Ladybird Beetle
##                          23                          23
##                  Mason Bee                    Mosquito
##                          22                          22
##              Argentine Ant                      Beetle
##                          21                          21
##   Flatheaded Appletree Borer        Horned Oak Gall Wasp
##                          20                          20
##          Leaf Beetle Family          Potato Leafhopper
##                          20                          20
##   Tooth-necked Fungus Beetle              Codling Moth
##                          20                          19
##    Black-spotted Lady Beetle              Calico Scale
##                          18                          18
##         Fairyfly Parasitoid               Lady Beetle
##                          18                          18
##       Minute Parasitic Wasps                 Mirid Bug
##                          18                          18
##            Mulberry Pyralid                  Silkworm
##                          18                          18
##              Vedalia Beetle       Araneoid Spider Order
##                          18                          17
##                  Bee Order              Egg Parasitoid
```

```
##                              17                                  17
##                    Insect Class           Moth And Butterfly Order
##                              17                                  17
##      Oystershell Scale Parasitoid Hemlock Woolly Adelgid Lady Beetle
##                              17                                  16
##             Hemlock Wooly Adelgid                               Mite
##                              16                                  16
##                     Onion Thrip             Western Flower Thrips
##                              16                                  15
##                     Corn Earworm                 Green Peach Aphid
##                              14                                  14
##                        House Fly                         Ox Beetle
##                              14                                  14
##                Red Scale Parasite                Spined Soldier Bug
##                              14                                  14
##             Armoured Scale Family                  Diamondback Moth
##                              13                                  13
##                     Eulophid Wasp                 Monarch Butterfly
##                              13                                  13
##                     Predatory Bug             Yellow Fever Mosquito
##                              13                                  13
##                Braconid Parasitoid                       Common Thrip
##                              12                                  12
##      Eastern Subterranean Termite                             Jassid
##                              12                                  12
##                        Mite Order                         Pea Aphid
##                              12                                  12
##                   Pond Wolf Spider           Spotless Ladybird Beetle
##                              12                                  11
##             Glasshouse Potato Wasp                          Lacewing
##                              10                                  10
##            Southern House Mosquito            Two Spotted Lady Beetle
##                              10                                  10
##                         Ant Family                       Apple Maggot
##                               9                                   9
```

Answer: The most commonly studied species, including honey bee (667), parasitic wasp (285), buff tailed bumblebee (183), carniolan honey bee (152), bumble bee (140), Italian honeybee (113), all belong to the order Hymenoptera. Most of these invertebrates are critical polinators and contributors to complex ecological structures, providing essential pollination services for crops. However, the extensive use of neonicotinoids in agriculture raises concerns as it poses potential threats to these bees, consequently impacting crop yields. Beyond their ecological significance, these species are chosen for study possibly due to their accessibility and availability in both ecologcial and agricultural contexts.

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?

```
#Determine the class of "Conc.1..Author." in the Neonics dataset
class(Neonics$Conc.1..Author.)
```

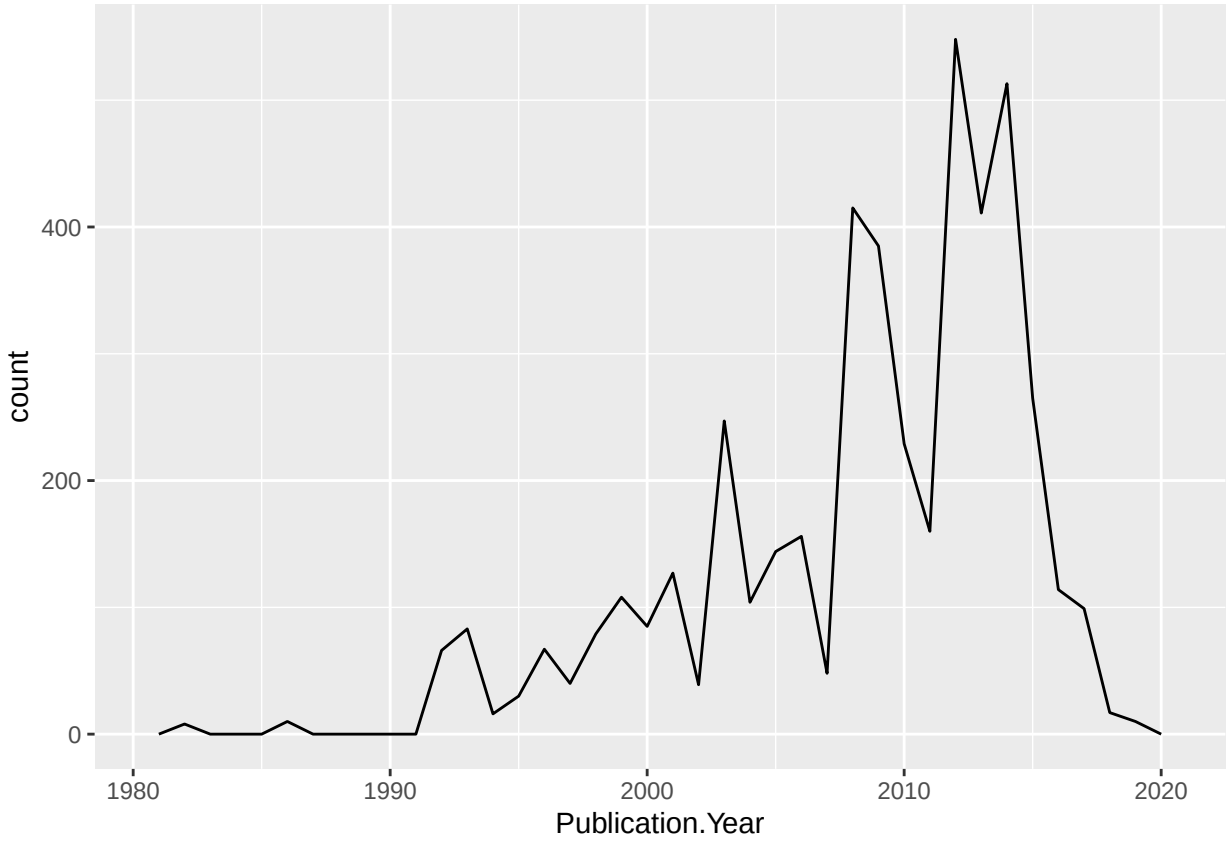
```
## [1] "factor"
```

Answer: It is a factor data. There is "/" in the column, which is considered as a character data. Numeric data can only contain integers and real numbers.

Explore your data graphically (Neonics)

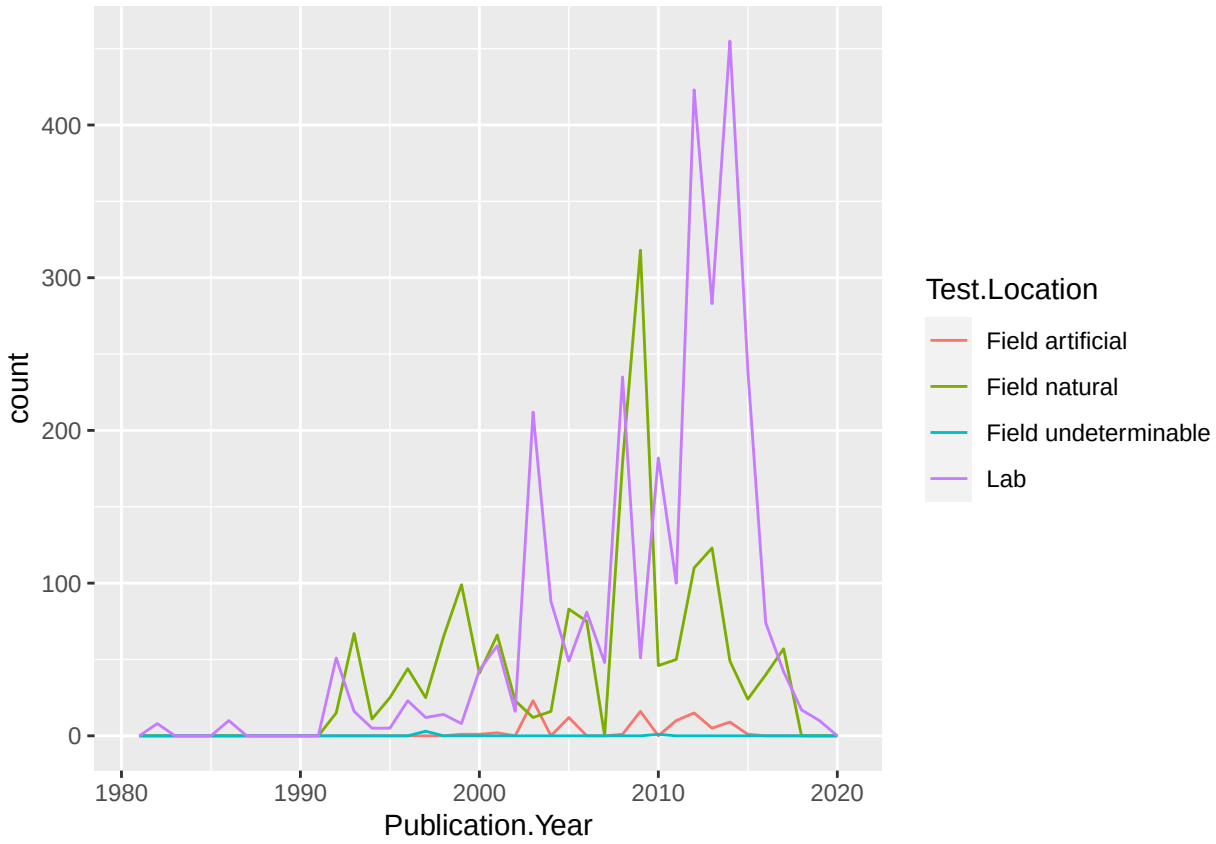
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
#check and load ggplot2 packages
library(ggplot2)
#Use ggplot to initializes the plot and specifies the x-axis as the publication year.
#Use binwidth = 1 to represent that each year should be a seperate bin.
ggplot(Neonics, aes(x = Publication.Year)) +
  geom_freqpoly(binwidth = 1, show.legend = FALSE)
```



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

```
#Add color aesthetics and a frequency polygon
#visualize the distribution of studies over publication years
ggplot(Neonics, aes(x = Publication.Year, color = Test.Location)) +
  geom_freqpoly(binwidth = 1, show.legend = TRUE)
```



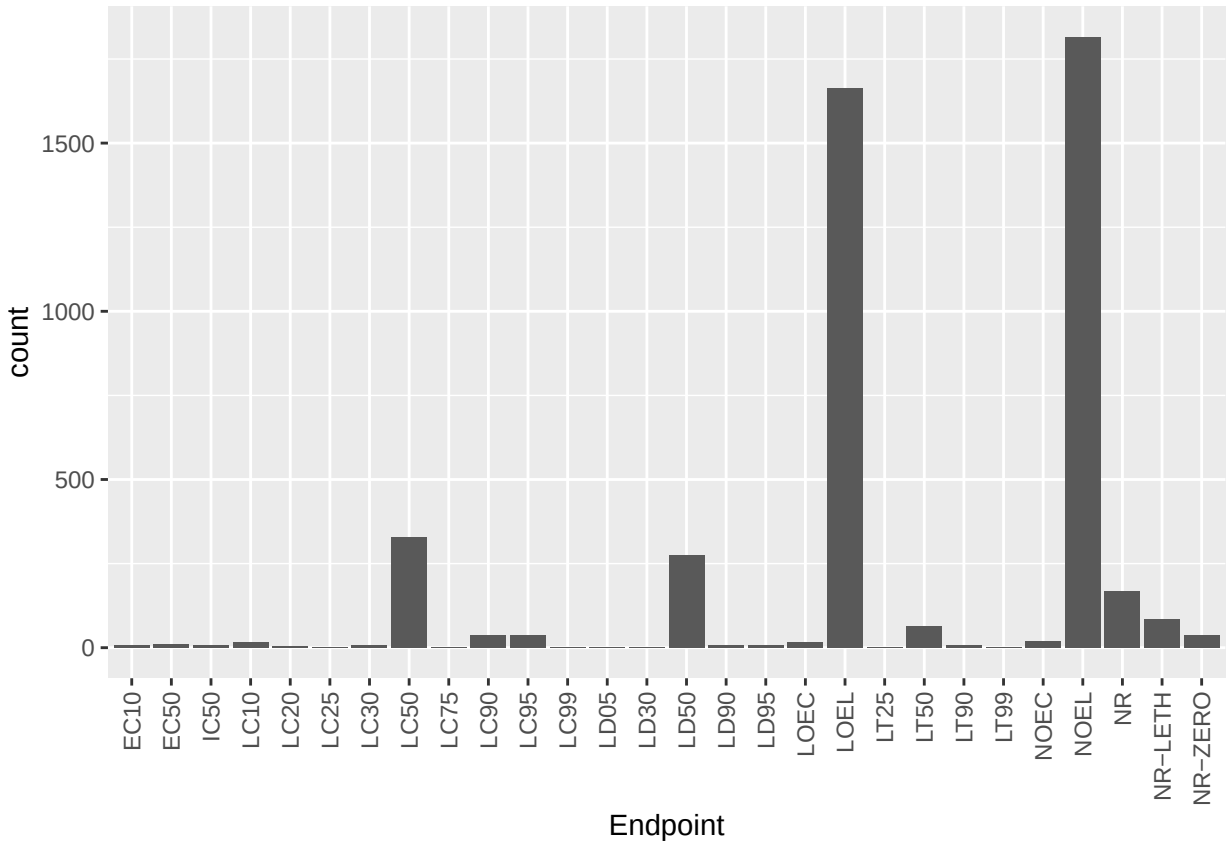
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are in laboratory but around 1995 and 2009, there was an increase in studies conducted in natural field locations.

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...]

```
#create a bar chart of endpoint counts in Neonics dataset using ggplot,
#endpoint is the column for x-axis, apply theme adjustments
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: The two most common endpoint is LOEL (lowest observable effect level) and NOEL (no observable effect level). LOEL represents the lowest dose causing effects significantly different from controls, while NOEL represents the highest dose where effects are not significantly different from controls.

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.

```
#Use class to determine the class of collectDate in the data set Litter.
#It is a factor data, not a date.
class(Litter$collectDate)
```

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

```
#Convert collectDate to year-month-day format and confirm the new class as date
Litter$collectDate <- as.Date(Litter$collectDate)
class(Litter$collectDate)
```

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

```
#Determine which date litter was sampled in August 2018
unique(Litter$collectDate[format(Litter$collectDate, "%Y-%m") == "2018-08"])
```

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

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`?

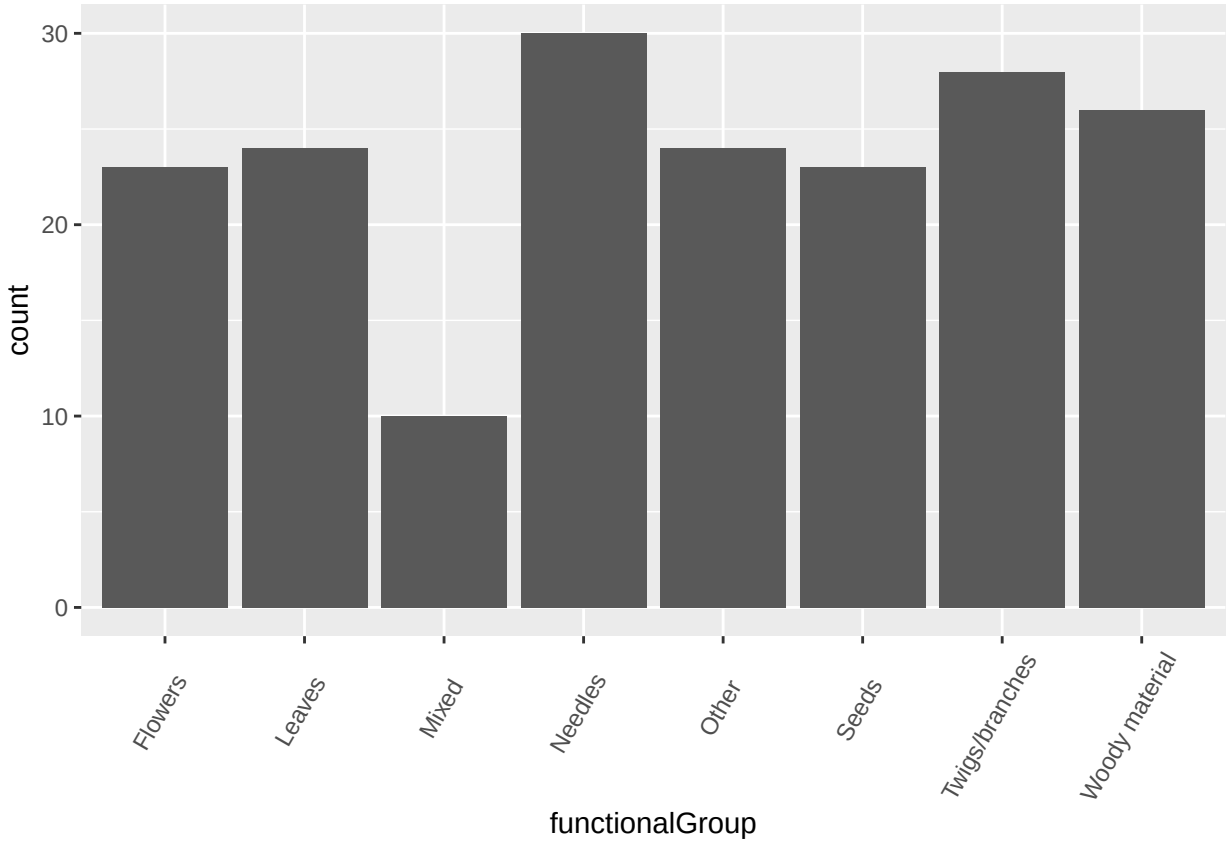
```r
#Subset unique plotIDs for NIWO (Niwot Ridge) sites using unique function,
#print the unique plotIDs sampled at Niwot Ridge
NIWO_plots <- unique (Litter$plotID [Litter$SiteID == "NIWO"])
print(NIWO_plots)
```

```
## factor()
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

Answer: 12 levels were sampled at Niwot Ridge. In this case, using 'unique' helps identify the specific plot values sampled at Niwot Ridge, while 'summary' provide an overall count of each unique plot across all sites.

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.

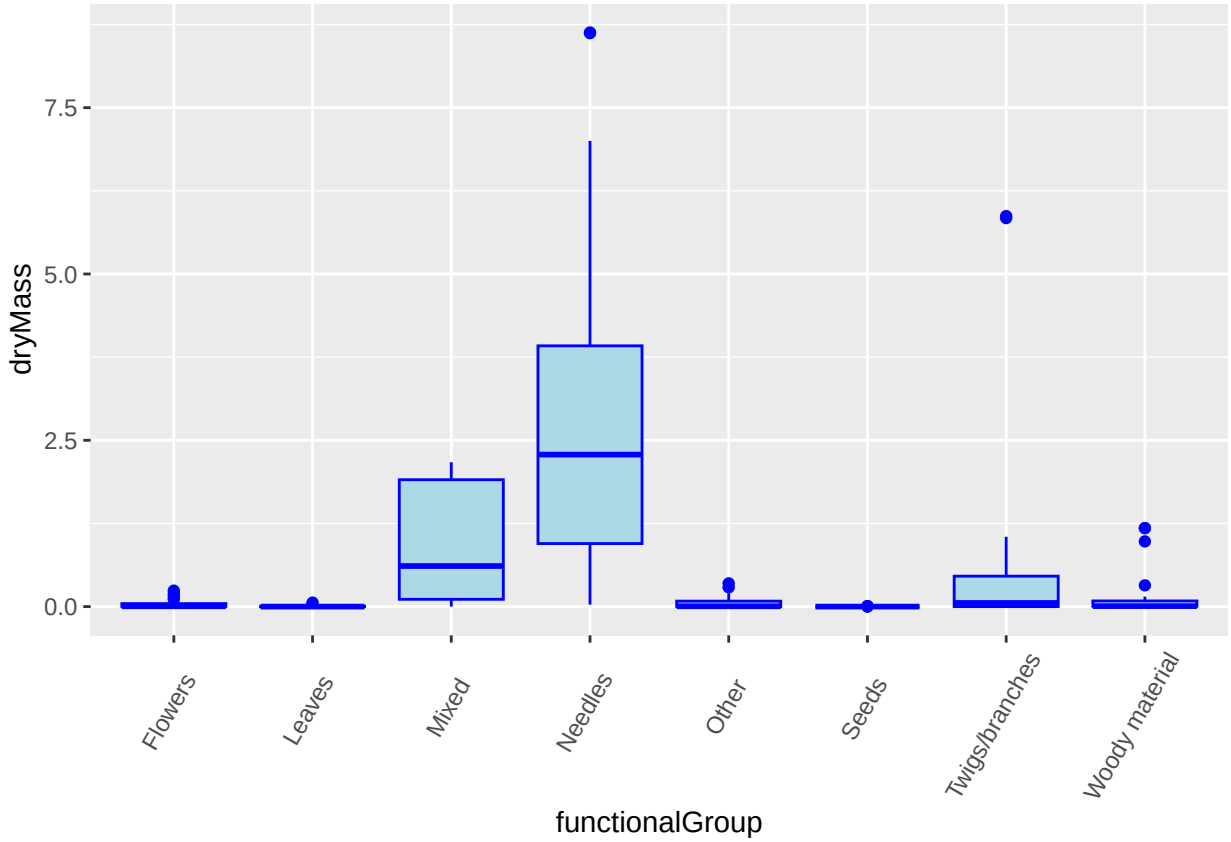
```r
#Use ggplot to create a bar graph of functionalGroup counts
ggplot(Litter, aes(x = functionalGroup))+
  geom_bar() +
#add theme adjustments for better readability
  theme(axis.text.x = element_text(angle = 60, vjust = 0.7, hjust=0.7))
```



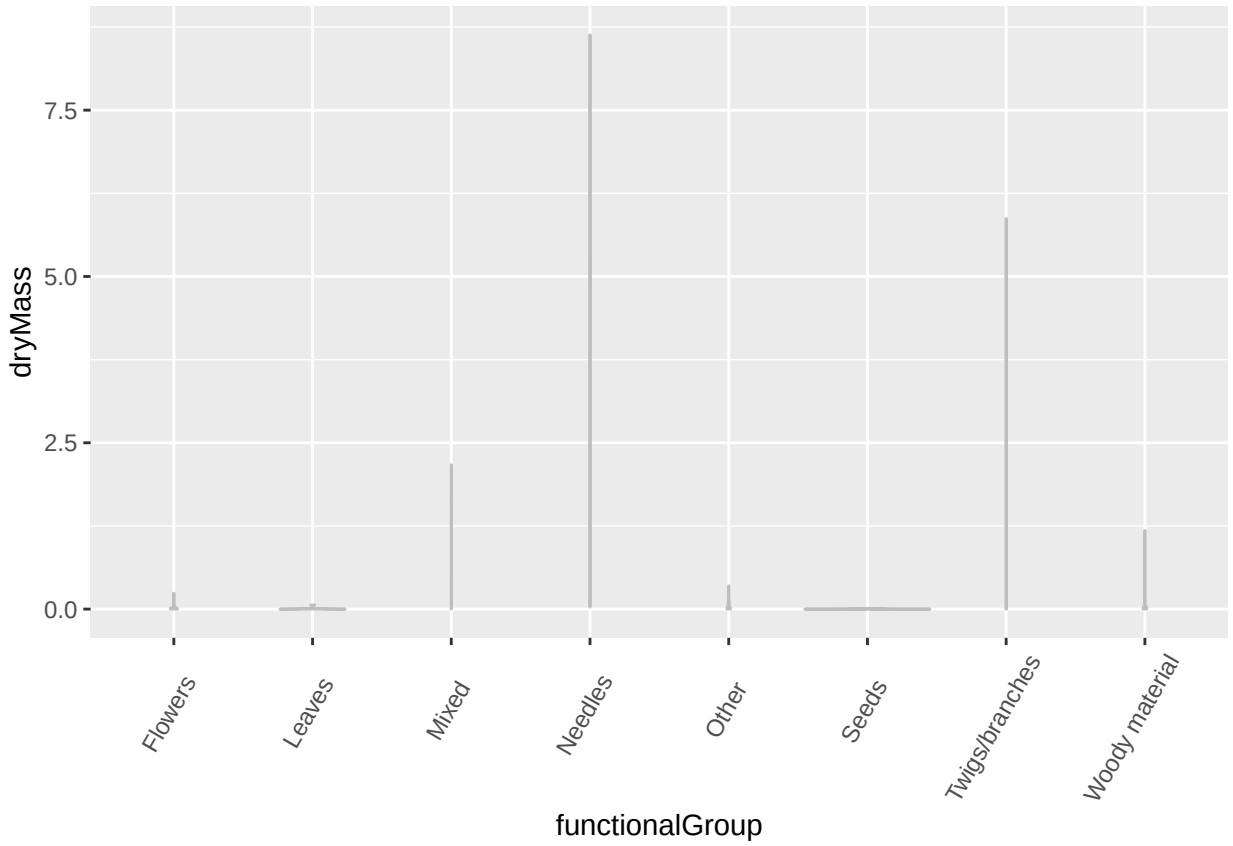
15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```r
#Use ggplot to create a boxplot, x-axis as functionalGroup, y-axis as dryMass
#add color for better visualization and theme adjustment for better readability
#set alpha as 1 (fully opaque) to better show outlier
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_boxplot(fill = "lightblue", color = "blue", alpha = 1) +
```

```
  theme(axis.text.x = element_text(angle = 60, vjust = 0.7, hjust=0.7))
```




```
#Use ggplot to create a violin plot, x-axis as functionalGroup, y-axis as dryMass
#add color for better visualization and theme adjustment for better readability
ggplot(Litter, aes(x = functionalGroup, y = dryMass))+
  geom_violin(fill = "lightgrey", color = "grey", alpha = 1) +
  theme(axis.text.x = element_text(angle = 60, vjust = 0.7, hjust=0.7))
```



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

> Answer: The boxplot is a more effective choice for visualizing this data because it provides a clear representation of outliers and extreme values, and it highlights the interquartile range. On the other hand, the data does not have an apparent shape, making the violin plot less effective in contrast to the box plot.

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

> Answer: Needles have the highest biomass. Mixed littler also tend to have higher biomass.