Assignment 3: Data Exploration

Sujay Dhanagare

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 subcommand to read strings in as factors.

```
library(lubridate)
library(here)
library(tidyverse)

here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```r
#Importing datasets
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:We are interested in the ecotoxicology of neonicotinoids on insects because they can harm non-target species like bees, which are important for pollination. These insecticides can also affect ecosystems and reduce biodiversity, with long-lasting effects due to their persistence in the environment. Studying these impacts helps us balance agricultural use with the need to protect the environment.

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:We study litter and woody debris in forests because they allow us to track and identify changing patterns in forest ecosystems, especially in response to climate change. This material determines the soil type, and shifts in its accumulation or decomposition can signal broader ecological changes. Additionally, excessive debris can increase the risk of wildfires, making it important to understand its impact in fire-prone regions as climate change alters fire dynamics.

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.Trap Types and Placement: Litter and fine woody debris are collected using elevated PVC traps for litter and ground traps for fine woody debris. The elevated traps are 0.5m^2 in size and set about 80 cm above the ground, while the ground traps cover a rectangular area of 3 m x 0.5 m to capture larger debris that can't be collected by the elevated traps. 2.Sampling Frequency:The frequency of sampling varies depending on the site and vegetation. In deciduous forest sites, elevated traps are sampled frequently (every two weeks) during leaf senescence, while in evergreen forest sites, sampling is less frequent (every one to two months). Ground traps, on the other hand, are sampled only once per year. 3. Plot Design: The sampling is done within 20 m x 20 m or 40 m x 40 m plots, depending on the site. Trap pairs (one elevated and one ground trap) are deployed for every 400 m^2 of plot area, resulting in 1-4 trap pairs per plot. Trap placement is randomized in sites with >50% cover of woody vegetation, but targeted in areas with patchy vegetation.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
##Basic overview of Neonics dataset

#class(Neonics)
#colnames(Neonics)
#str(Neonics)
#length(Neonics)

print(dim(Neonics))
```

```
## [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? [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```
## Summary of the dataset

most_effeccts <- sort(summary(Neonics$Effect), decreasing = TRUE) #This gives the frequency of each eff

print(most_effeccts)
```

```
##       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:These effects are key indicators of the ecological risks associated with neonicotinoid use, particularly in agriculture, where pollinators and other beneficial insects are essential for crop production and biodiversity maintenance. For instance, population decline (1803 instances) is a key area of study as it reflects the broader ecological impact. Mortality (1493 instances) is another significant focus, as it measures the direct lethality of neonicotinoids on both target and non-target species. Behavioral changes (360 instances), including feeding behavior (255 instances), are of interest because they can disrupt essential activities such as foraging and reproduction, which can further impact population stability. Additionally, reproductive impairment (197 instances) is crucial as it affects long-term population dynamics, even when short-term mortality is low. These results provide a comprehensive understanding of both immediate and long-term consequences of neonicotinoid exposure, informing ecological risk assessments and regulatory decisions.

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...]

```r
#Gives the most commonly studied species

species_mostcommon <- summary(Neonics$Species.Common.Name, maxsum = 7) #maxsum = 7 as the last number i
```

Answer:The six most commonly studied species—Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee—are of interest because they are key pollinators or natural pest controllers. Pollinators like honey bees and bumblebees are crucial for crop production and ecosystem health, while parasitic wasps play an important role in biological pest control. These species are sensitive to neonicotinoids, making them central to understanding the broader ecological risks and ensuring sustainable agricultural practices.

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...]

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

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

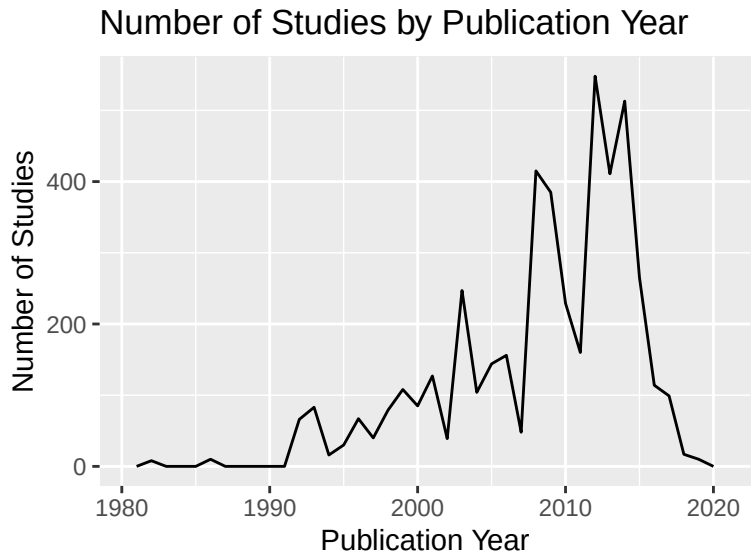
Answer: The class of `Conc.1..Author.`is "factor" because it contains non-numeric values, such as "NR" (not reported) and entries with characters like "/", "<", and ">". These non-numeric symbols prevent the column from being classified as numeric.

Explore your data graphically (Neonics)

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

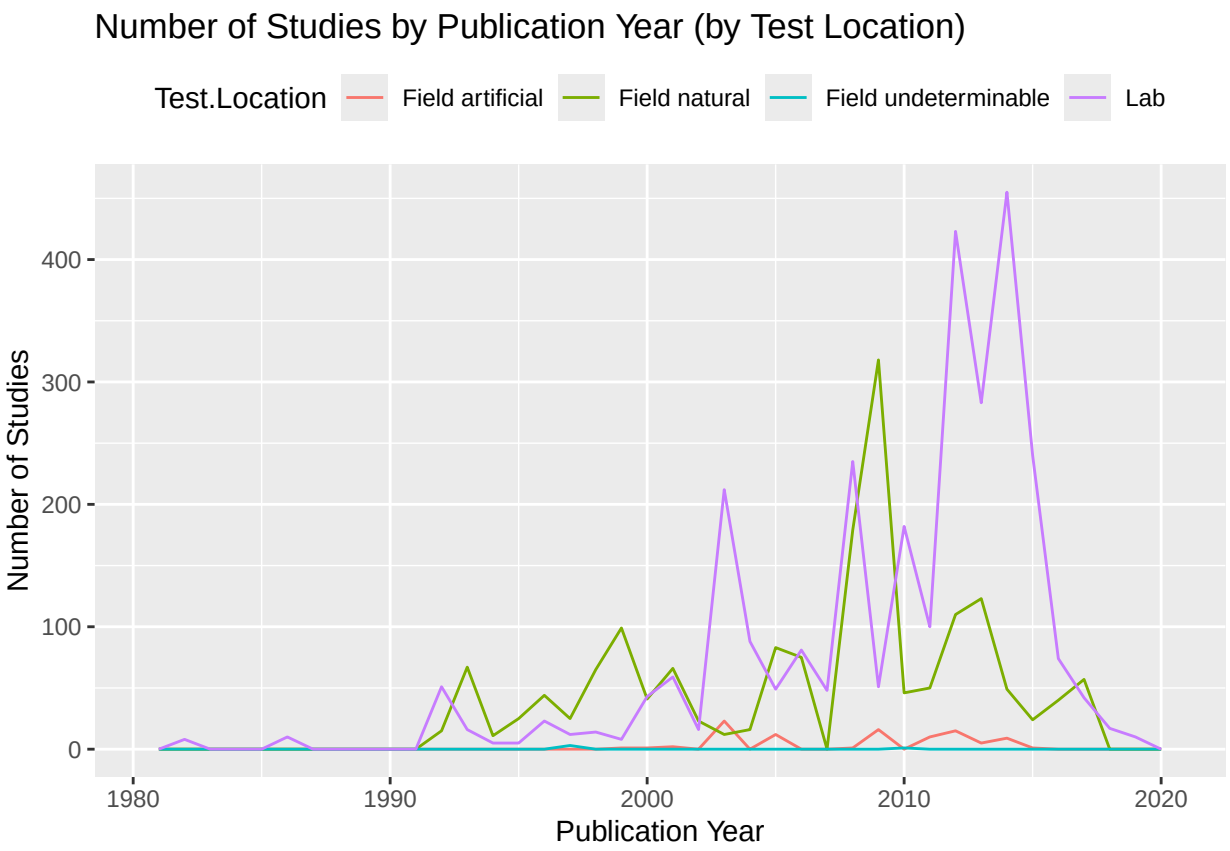
```r
#

ggplot(Neonics, aes(x = Publication.Year)) +
  geom_freqpoly(binwidth = 1) +
  labs(title = "Number of Studies by Publication Year",
       x = "Publication Year",
       y = "Number of Studies") +
  theme(legend.position = "top")
```



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

```
ggplot(Neonics, aes(x = Publication.Year, color = Test.Location)) +
  geom_freqpoly(binwidth = 1) +
  labs(title = "Number of Studies by Publication Year (by Test Location)",
       x = "Publication Year",
       y = "Number of Studies") +
  theme(legend.position = "top")
```



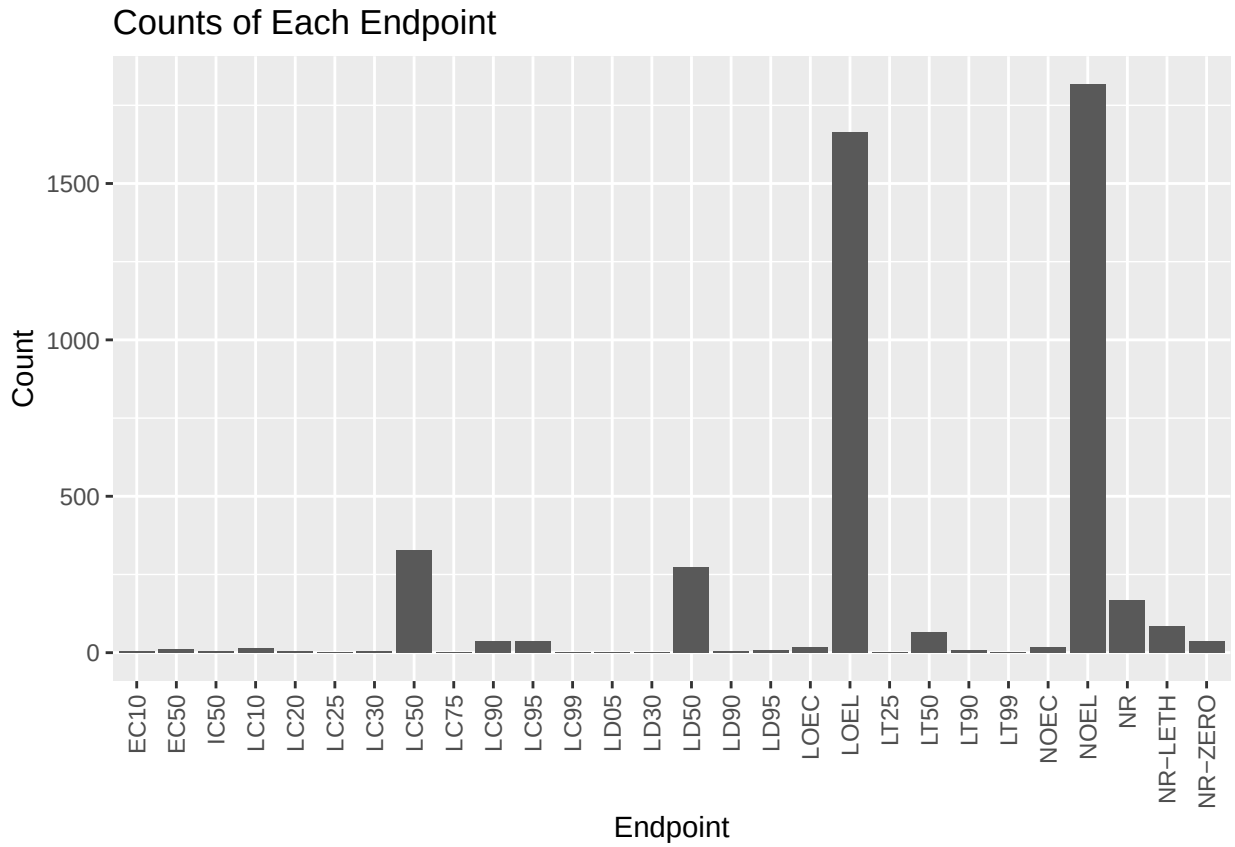
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer:The most common test location over time is Lab, particularly from 2000 to 2015, indicating a heavy focus on controlled, laboratory-based studies during this period. Field natural is the second most common location, showing several peaks of interest, especially around 2010. Field undetermined also shows some research activity but less so compared to Lab and Field natural. Field artificial remains the least utilized location for studies. Over time, research has shifted between these environments, with a significant decline in studies after 2015.

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...]

```r
# Create the bar graph of Endpoint counts
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar() +
  labs(title = "Counts of Each Endpoint",
       x = "Endpoint",
       y = "Count") +
 theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))  # Rotating x-axis labels for bett
```




```r
# Counts of each Endpoint in descending order
endpoint_counts <- sort(table(Neonics$Endpoint), decreasing = TRUE)

# The two most common Endpoints
print(head(endpoint_counts, 2))
```

```
##
## NOEL LOEL
## 1816 1664
```

Answer:The two most common end points are NOEL and LOEL. Terrestrial LOEL (Lowest-observable-effect-level): Definition: The lowest dose (or concentration) that produces effects significantly different from the control group, according to the authors' report. It is a critical endpoint in ecotoxicological studies because it marks the threshold where observable effects begin to appear in the tested organisms (LOEAL/LOEC).

Terrestrial NOEL (No-observable-effect-level): Definition: The highest dose (or concentration) that does not produce effects significantly different from the control group, according to the authors' statistical tests. It represents the threshold below which no observable effects occur (NOEAL/NOEC).

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.

```
print(class(Litter$collectDate))
```

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

```
#The class of collectDate is factor. It needs to be converted to Date

Litter$collectDate <- as.Date(Litter$collectDate)
print(class(Litter$collectDate))
```

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

```
# Use the unique function to find dates in August 2018
unique_dates_august_2018 <- unique(Litter$collectDate[format(Litter$collectDate, "%Y-%m") == "2018-08"])

print(unique_dates_august_2018)
```

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

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`?

```
#Determine unique plots at Niwot Ridge by using plot ID

unique_plots_Niwot <- unique(Litter$plotID)

print(length(unique_plots_Niwot)) # Number of plots sampled at Niwot Ridge
```

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

```
summary(Litter$plotID)
```

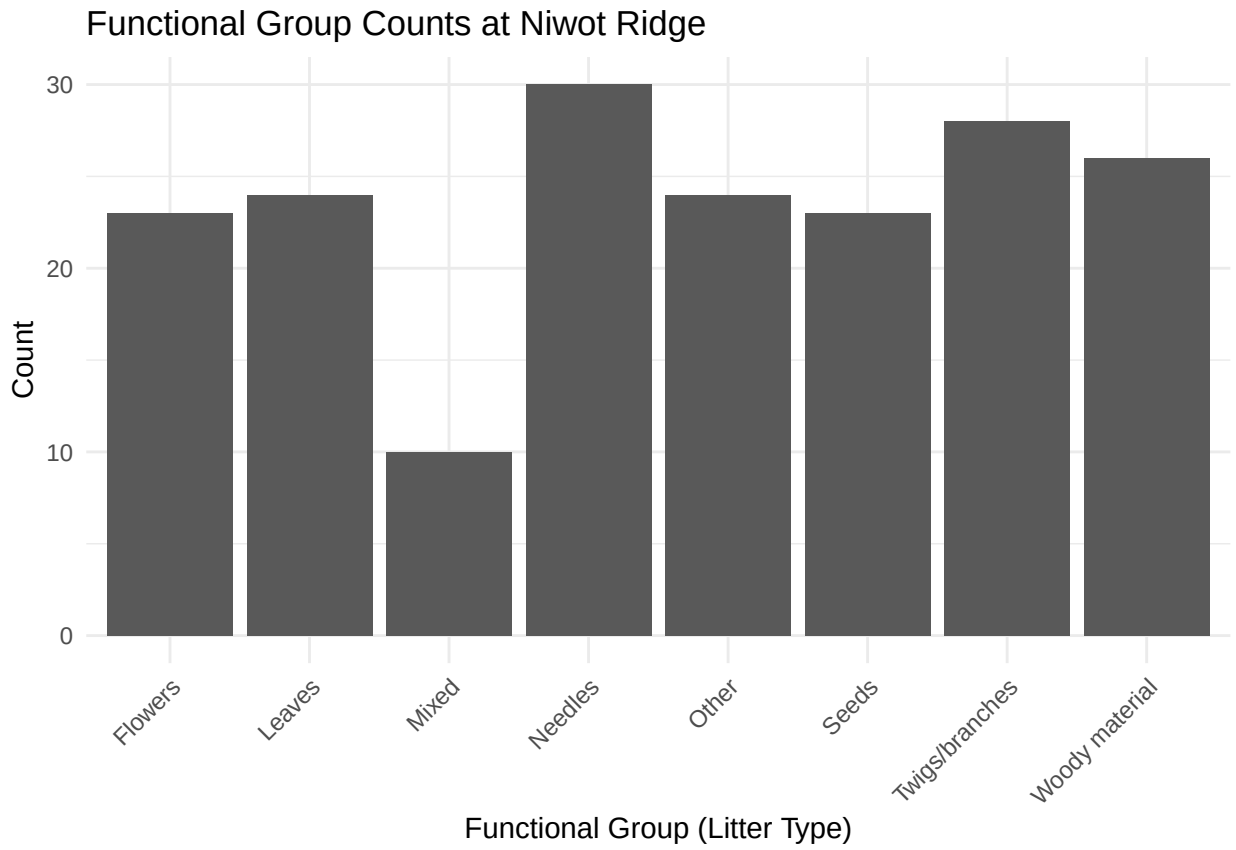
```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

Answer:Unique here tells us of how many plots were sampled, whereas summary tells us how many 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.

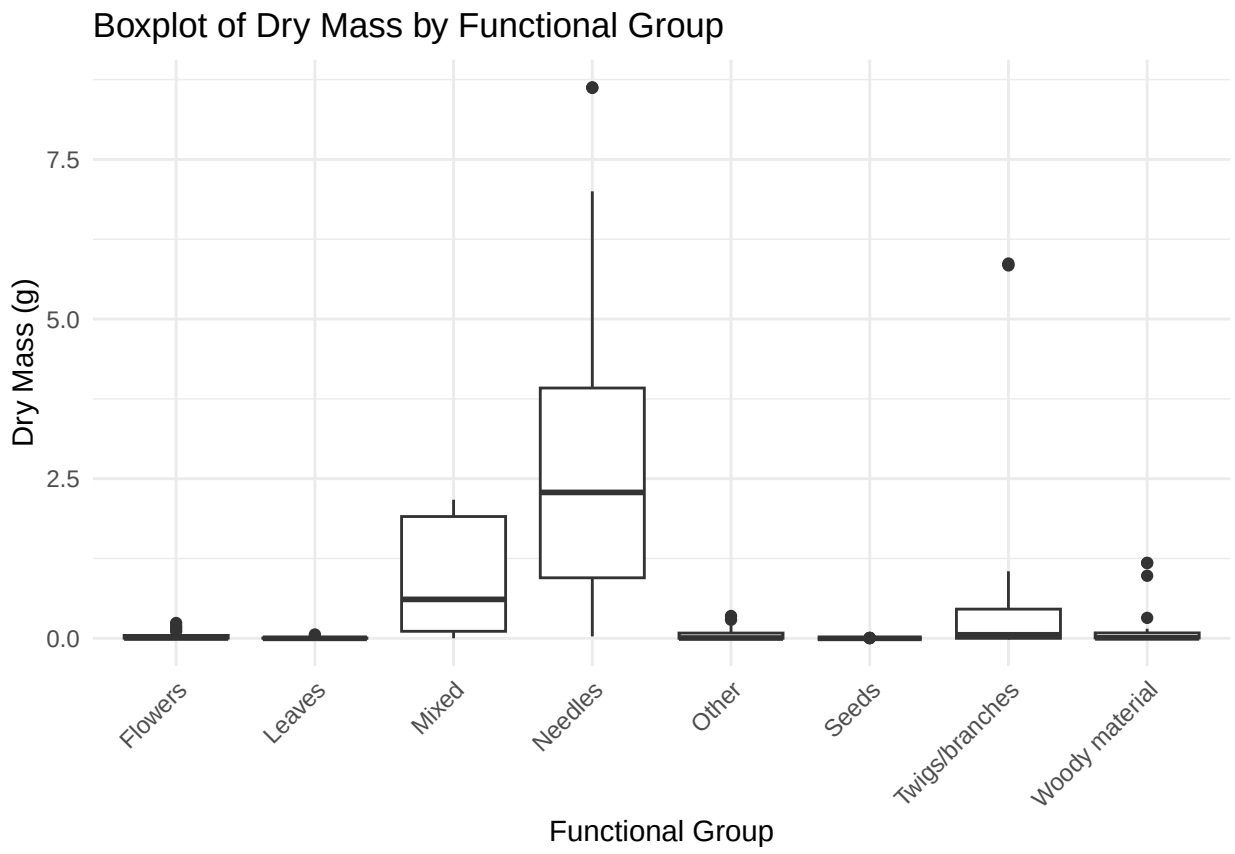
```r
# bar graph of functionalGroup counts

ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar() +
  labs(title = "Functional Group Counts at Niwot Ridge",
       x = "Functional Group (Litter Type)",
       y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for readability
```

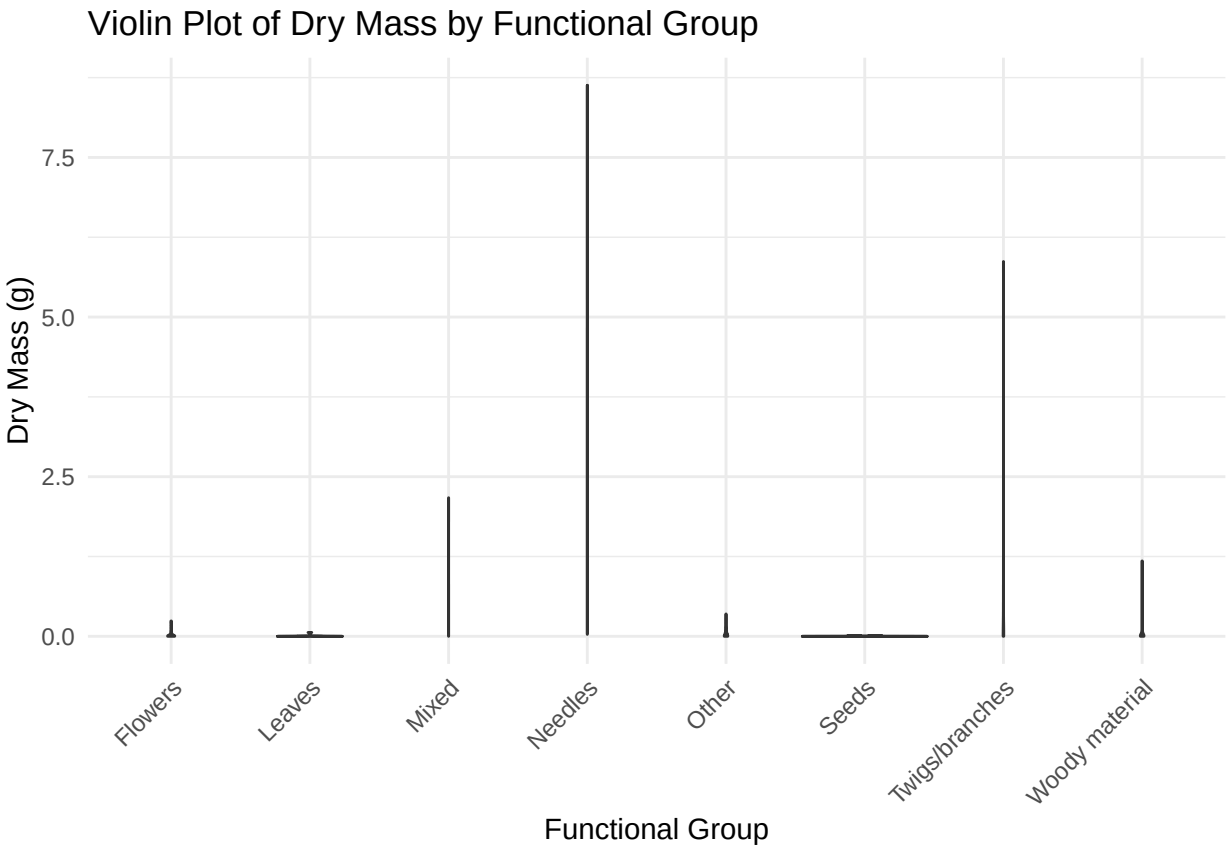


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```r
# Boxplot of dryMass by functionalGroup
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_boxplot() +
  labs(title = "Boxplot of Dry Mass by Functional Group",
       x = "Functional Group",
       y = "Dry Mass (g)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readabili
```




```
# Violin plot of dryMass by functionalGroup
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_violin() +
  labs(title = "Violin Plot of Dry Mass by Functional Group",
       x = "Functional Group",
       y = "Dry Mass (g)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readabili
```



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

Answer:The boxplot is more effective here because it provides a clear summary of the key statistics (median, quartiles, and outliers) that are more relevant to this dataset, where most of the functional groups have low variation in dryMass but with some significant outliers. The violin plot adds complexity without providing additional meaningful insights due to the compressed distributions.

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

Answer: Based on the boxplot and violin plot, the types of litter that tend to have the highest biomass at the Niwot Ridge sites are: 1. Needles: This functional group consistently shows the highest dry mass values, with a wide interquartile range and several significant outliers, indicating that needles tend to accumulate more biomass compared to other litter types. 2. Mixed: This group also has a relatively higher biomass, though not as much as needles. The range of dry mass values in this group is broader than most other litter types. 3. Woody material and Twig/branches: These groups show some higher dry mass values, though with fewer observations compared to needles. Woody material also includes some outliers, indicating that it occasionally accumulates significant biomass. Overall, Needles stand out as the litter type with the highest and most variable biomass, while Mixed, Woody material, and Twig/branches also contribute to relatively higher biomass at these sites.