Assignment 3: Data Exploration

Student Name

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.

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

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

```
## [1] "/Users/chrissiepantoja/Library/CloudStorage/OneDrive-DukeUniversity/PHD DUKE/1 COURSES/3 FALL SE
```

```
# Upload the datasets
Neonics <- read_csv(here("Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"), col_types = cols(.default =
Litter <- read_csv(here("Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv"), col_types = cols(.defaul

# View the first few rows of each dataset
head(Neonics)
```

```
## # A tibble: 6 x 30
##   `CAS Number` `Chemical Name`           `Chemical Grade` Chemical Analysis Me~1
##   <fct>        <fct>                     <fct>            <fct>
## 1 58842209     Tetrahydro-2-(nitromethy~ Technical grade~ Unmeasured
## 2 58842209     Tetrahydro-2-(nitromethy~ Technical grade~ Unmeasured
## 3 58842209     Tetrahydro-2-(nitromethy~ Technical grade~ Unmeasured
## 4 58842209     Tetrahydro-2-(nitromethy~ Technical grade~ Unmeasured
## 5 58842209     Tetrahydro-2-(nitromethy~ Technical grade~ Unmeasured
## 6 58842209     Tetrahydro-2-(nitromethy~ Technical grade~ Unmeasured
## # i abbreviated name: 1: `Chemical Analysis Method`
## # i 26 more variables: `Chemical Purity` <fct>,
## #   `Species Scientific Name` <fct>, `Species Common Name` <fct>,
## #   `Species Group` <fct>, `Organism Lifestage` <fct>, `Organism Age` <fct>,
## #   `Organism Age Units` <fct>, `Exposure Type` <fct>, `Media Type` <fct>,
## #   `Test Location` <fct>, `Number of Doses` <fct>,
## #   `Conc 1 Type (Author)` <fct>, `Conc 1 (Author)` <fct>, ...
```

```
head(Litter)
```

```
## # A tibble: 6 x 19
##   uid              namedLocation domainID siteID plotID trapID weighDate setDate
##   <fct>            <fct>         <fct>    <fct>  <fct>  <fct>  <fct>     <fct>
## 1 7f065fec-bcb2-4~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-~ 2018-0~
## 2 88df210b-1445-4~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-~ 2018-0~
## 3 7f3c549c-1dfa-4~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-~ 2018-0~
## 4 97806ab5-42d2-4~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-~ 2018-0~
## 5 9d7c89f5-85f8-4~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-~ 2018-0~
## 6 6ca7a3e8-4d9e-4~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-~ 2018-0~
## # i 11 more variables: collectDate <fct>, ovenStartDate <fct>,
## #   ovenEndDate <fct>, fieldSampleID <fct>, massSampleID <fct>,
## #   samplingProtocolVersion <fct>, functionalGroup <fct>, dryMass <fct>,
## #   qaDryMass <fct>, remarks <fct>, measuredBy <fct>
```

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: The ecotoxicology of neonicotinoids on insects is crucial because these insecticides, while effective against pests, have been shown to harm non-target species like bees, which are essential pollinators. Their persistence in the environment can lead to prolonged exposure, affecting insect biodiversity and ecosystem health. Understanding these impacts helps balance agricultural benefits with the need to protect vital insect populations and maintain ecological balance.

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: Studying litter and woody debris in forests, such as at the Niwot Ridge LTER station, is crucial because these materials play significant roles in forest ecosystems. They contribute to nutrient cycling, providing essential nutrients to the soil as they decompose. Woody debris also serves as habitat for various organisms, enhancing biodiversity. Additionally, it acts as a carbon sink, helping to mitigate climate change by storing carbon. Understanding these processes helps in managing forests sustainably and preserving their ecological functions.

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. Spatial Sampling Design: Sampling is conducted at terrestrial NEON sites with woody vegetation taller than 2 meters. The sampling occurs in tower plots, which are randomly selected within the 90% flux footprint of the primary and secondary airsheds. In forested areas, sampling is done in 20 plots of 40m x 40m, while in areas with low-stature vegetation, it is done in 4 plots of 40m x 40m and 26 plots of 20m x 20m. Each plot contains 1-4 pairs of litter traps (one elevated and one ground trap) per 400 m^2. 2. Trap Placement: The placement of litter traps within plots can be either targeted or randomized, depending on the vegetation cover. In areas with more than 50% aerial cover of woody vegetation taller than 2 meters, trap placement is randomized. In areas with less than 50% cover or patchy vegetation, trap placement is targeted to areas beneath qualifying vegetation. 3. Temporal Sampling Design: Ground traps are sampled once per year, while elevated traps are sampled more frequently depending on the vegetation type. In deciduous forest sites, elevated traps are sampled every two weeks during senescence, and in evergreen sites, they are sampled every 1-2 months year-round. Sampling may be discontinued for up to six months during the dormant season in sites with deciduous vegetation or limited winter access.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# Get the dimensions
dimensions <- dim(Neonics)
cat("The dataset has", dimensions[1], "rows and", dimensions[2], "columns.\n")
```

```
## The dataset has 4623 rows and 30 columns.
```

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...]

```
# Get a summary of the "Effect" column
effect_summary <- summary(Neonics$Effect)

# Sort the summary in descending order
```

```
sorted_effects <- sort(effect_summary, decreasing = TRUE)

# Print the sorted effects
print(sorted_effects)
```

```
##       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
##     Intoxication     Accumulation     Biochemistry          Cell(s)
##               12               12               11                9
##       Physiology        Histology       Hormone(s)
##                7                5                1
```

Answer: The most common effects are population, mortality, behavior, feeding behavior and reproduction because they provide comprehensive insights into the health and sustainability of ecosystems. By studying these aspects, researchers can identify early signs of environmental stress, understand the mechanisms behind ecological changes, and develop strategies for conservations and management.

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...]

```
# Six most commonly studied species
summary(Neonics$`Species Common Name`, maxsum = 7)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee           Bumble Bee      Italian Honeybee
##                   152                   140                   113
##               (Other)
##                  3083
```

Answer: The six most commonly studied species are —Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee—primarily belong to the order Hymenoptera, which includes bees, wasps, and ants. These insects are notable for their roles as pollinators, with honey bees and bumblebees being particularly crucial for the pollination of many crops and wild plants. Additionally, many of these species exhibit complex social behaviors, living in colonies with a division of labor among workers, queens, and drones. This social structure allows for efficient functioning and survival of the colony.

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...]

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

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

Answer: This column class is a factor, it means that R has interpreted this column as categorical data or non-numeric characters that have a fixed number of unique values, known as levels.

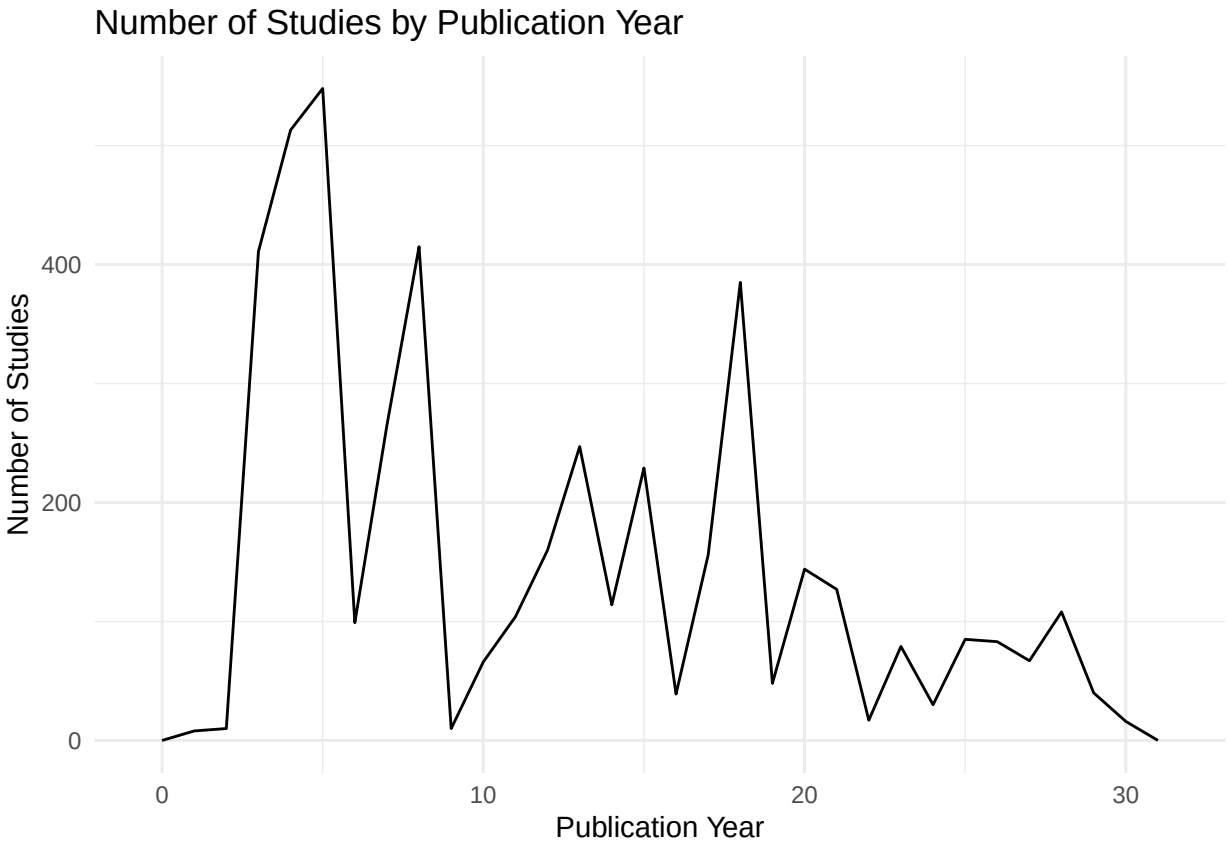
Explore your data graphically (Neonics)

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

```
class(Neonics$`Publication Year`)
```

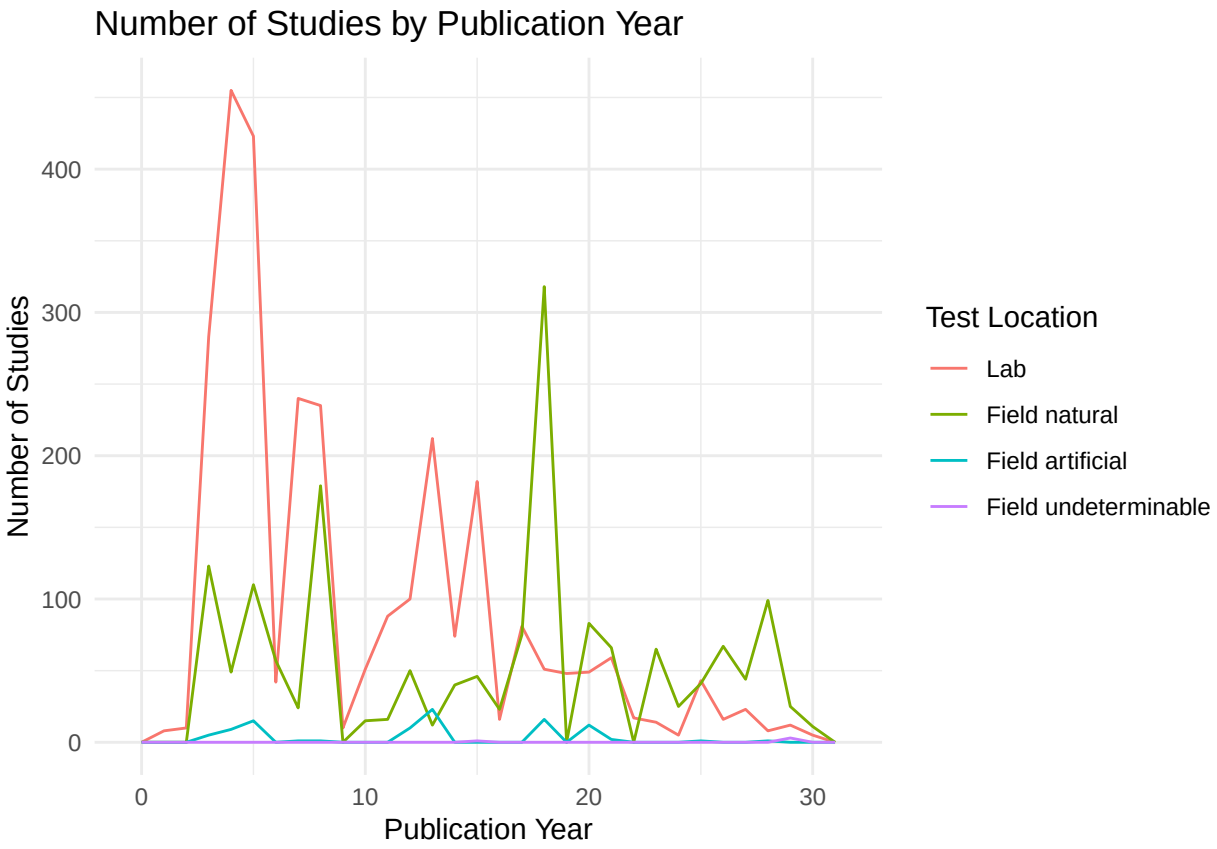
```
## [1] "factor"
```

```
Neonics$`Publication Year` <- as.numeric(Neonics$`Publication Year`)

library(ggplot2)

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_minimal()
```



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

```r
ggplot(Neonics, aes(x = `Publication Year`, color = `Test Location`)) +
  geom_freqpoly(binwidth = 1) +
  labs(title = "Number of Studies by Publication Year",
       x = "Publication Year",
       y = "Number of Studies",
       color = "Test Location") +
  theme_minimal()
```



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

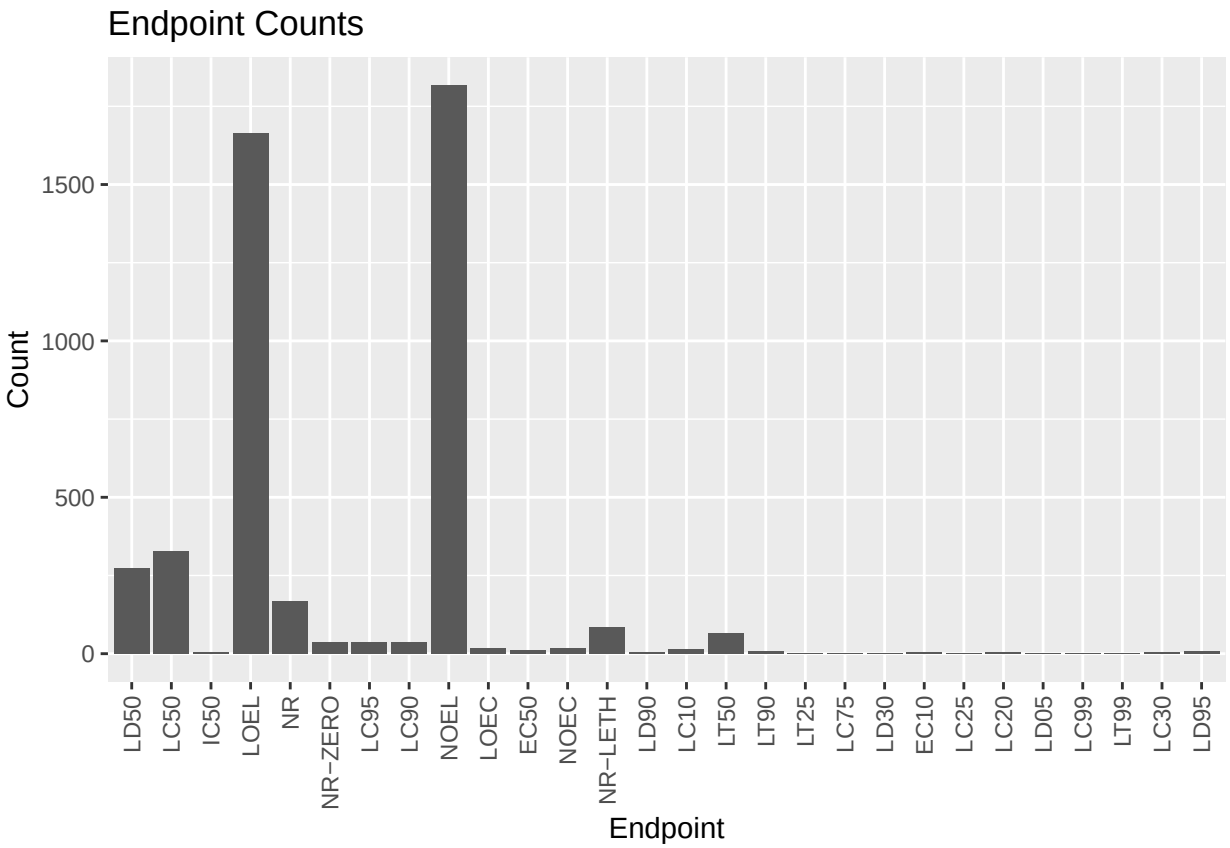
> Answer: Common test locations are laboratories, natural fields, artificial fields, and undeterminable fields. These vary over time due to technological advancements, research focus, and environmental changes.

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
library(dplyr)
library(ggplot2)

# Count the occurrences of each endpoint and select the top 2
endpoint_counts <- Neonics %>%
```

```
  count(Endpoint) %>%
  arrange(desc(n))

# Plot the bar graph
ggplot(endpoint_counts, aes(x = as.factor(Endpoint), y = n)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  labs(title = "Endpoint Counts", x = "Endpoint", y = "Count")
```



Answer: The two most common endpoints ae LOEL and NOEL. LOEL is the Lowest-Observable-Effect-Level: lowest dose (concentration) producing by effects that were significantly (as reported by authors) from responses of controls. NOEL is the No-Observable-Effect-Level: highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test.

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.

```
# Check the class of collectDate
class(Litter$collectDate)
```

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

```r
# collectDate is a factor, convert it to Date
if (!inherits(Litter$collectDate, "Date")) {
  Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
}

# Confirm the new class of collectDate
class(Litter$collectDate)
```

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

```r
# Determine unique dates in August 2018
august_2018_dates <- unique(Litter$collectDate[format(Litter$collectDate, "%Y-%m") == "2018-08"])

print(august_2018_dates)
```

```
## [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`?

```r
# Determine the unique plots sampled at Niwot Ridge
unique_plots <- unique(Litter$plotID[Litter$siteID == "NIWO"])
print(unique_plots)
```

```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 ... NIWO_057
```

```r
# Count the number of unique plots
num_unique_plots <- length(unique_plots)
print(num_unique_plots)
```

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

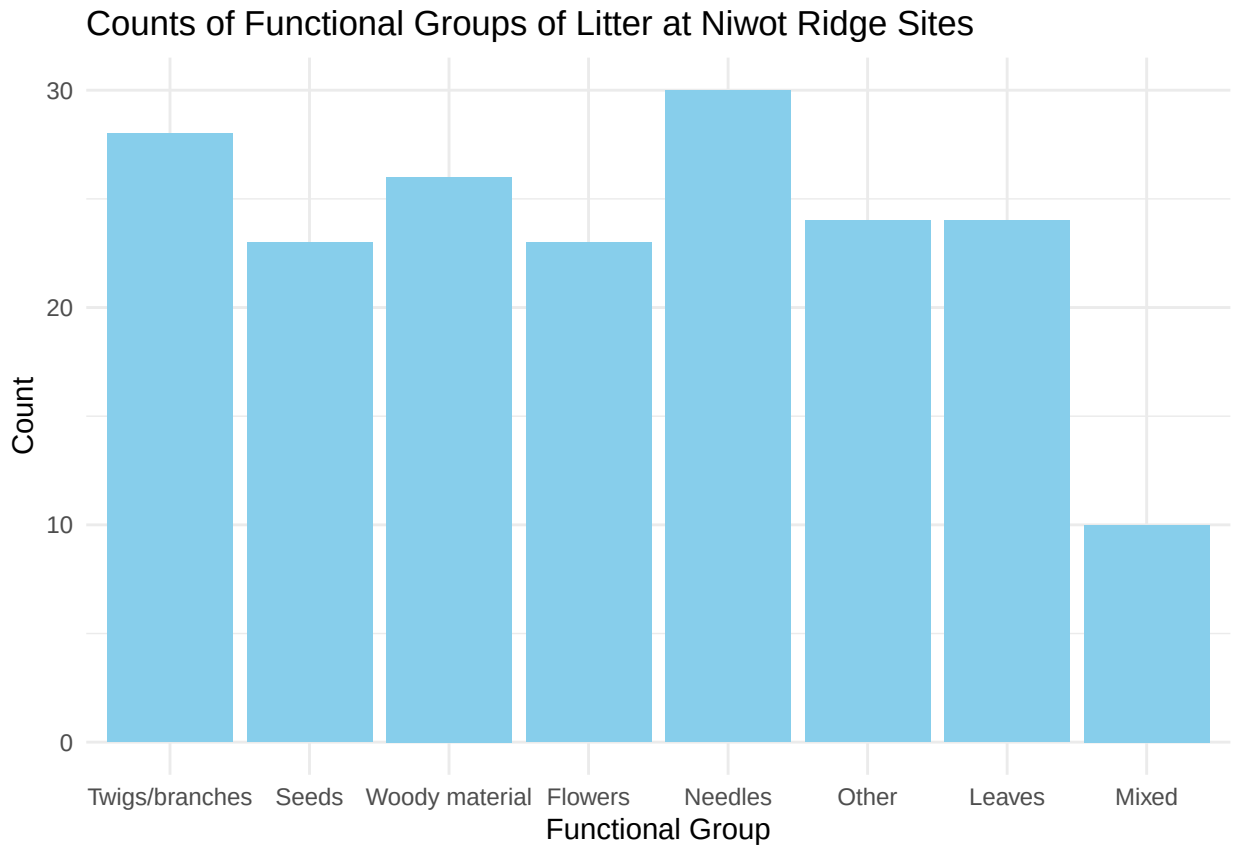
```r
# Using summary to get a statistical summary of PlotID
summary_plots <- summary(Litter$plotID[Litter$siteID == "NIWO"])
print(summary_plots)
```

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

Answer: The dataset includes 12 distinct plots. The summary gives a count of plots by name, while the unique function simply lists the plot names.

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.

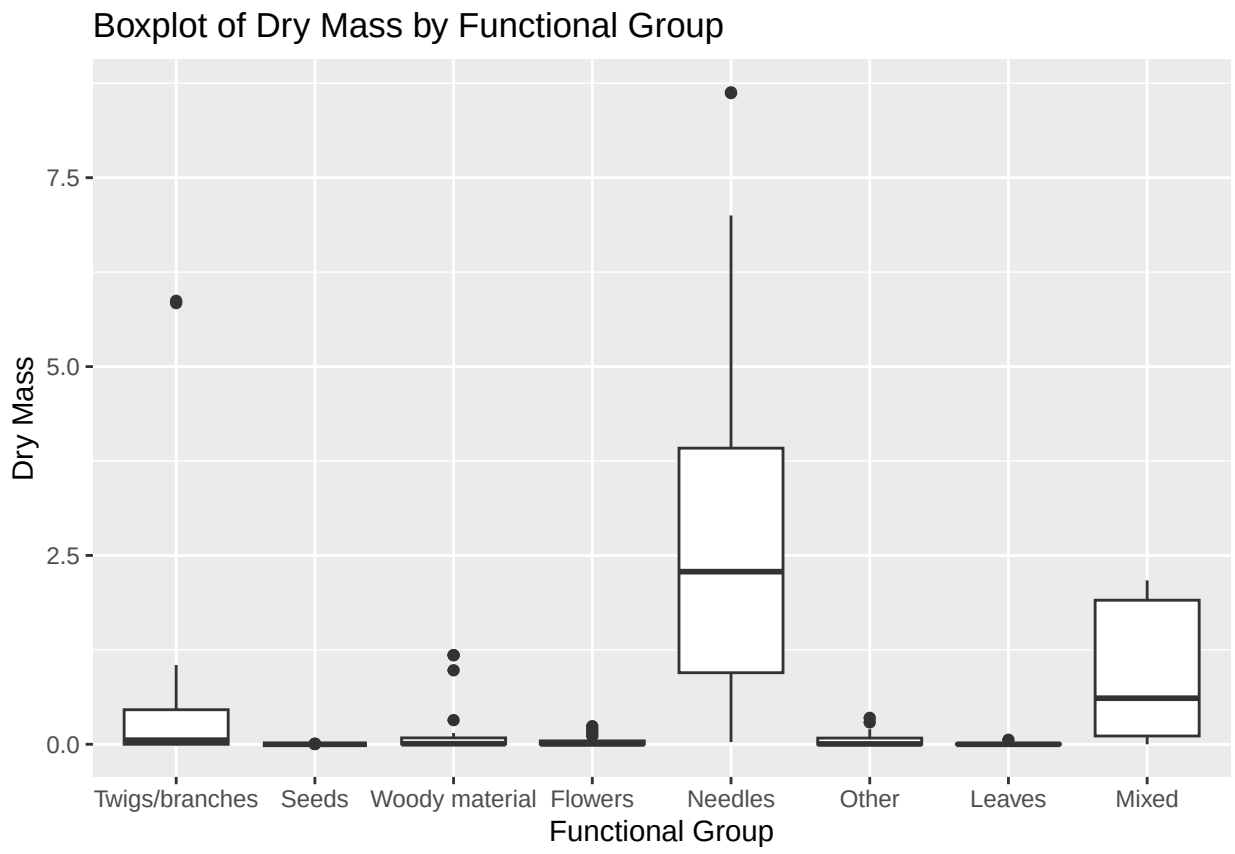
```r
# Summarize the counts of each functional group
functional_group_counts <- Litter %>%
  count(functionalGroup)

# Create the bar graph
ggplot(functional_group_counts, aes(x = functionalGroup, y = n)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Counts of Functional Groups of Litter at Niwot Ridge Sites",
       x = "Functional Group",
       y = "Count") +
  theme_minimal()
```



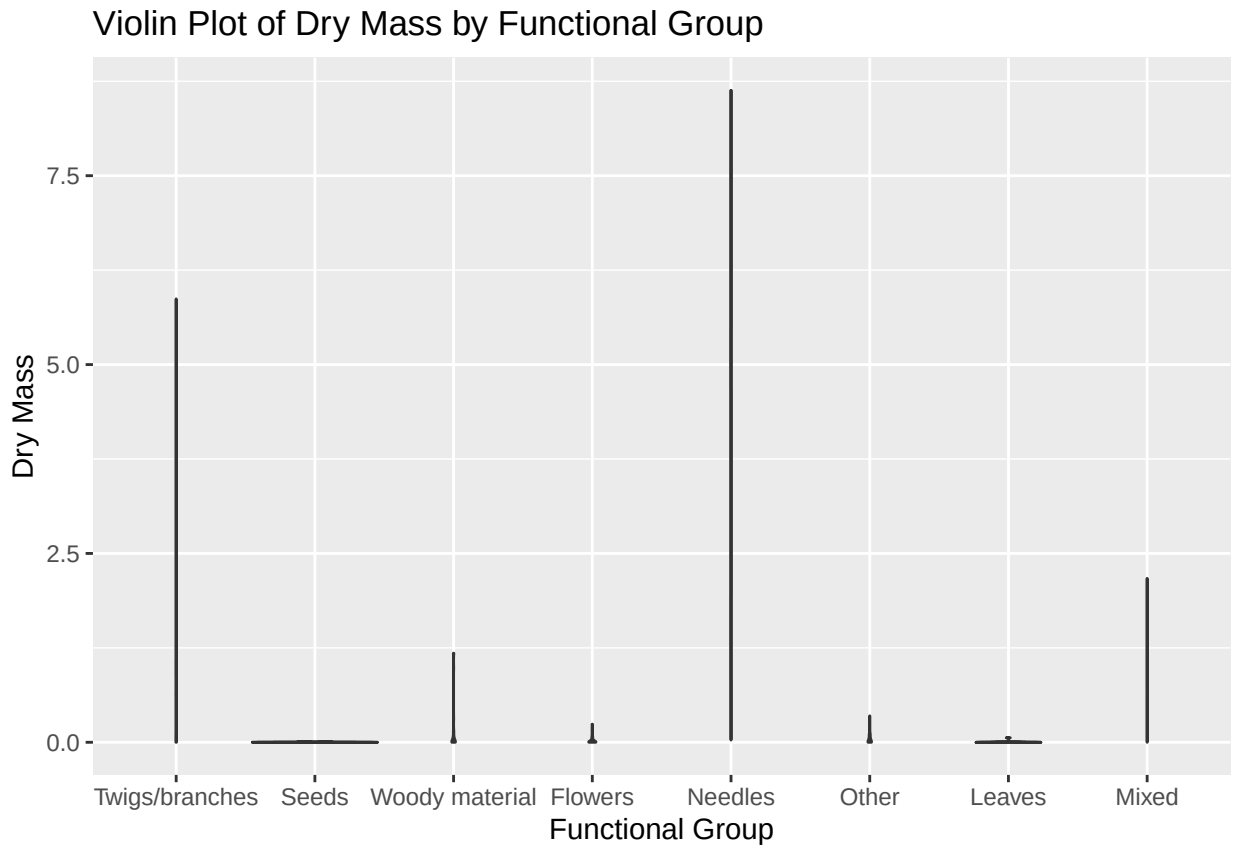
15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```r
# Ensure dryMass is numeric
Litter$dryMass <- as.numeric(as.character(Litter$dryMass))

# Check for NA values
if (any(is.na(Litter$dryMass))) {
  warning("There are NA values in dryMass after conversion.")
}

# Create boxplot with titles
boxplot <- ggplot(Litter) +
```

```
  geom_boxplot(
    aes(
      x = functionalGroup,
      y = dryMass,
      group = cut_width(functionalGroup, 1)
    )
  ) +
  ggtitle("Boxplot of Dry Mass by Functional Group") +
  xlab("Functional Group") +
  ylab("Dry Mass")

# Create violin plot with titles
violin_plot <- ggplot(Litter) +
  geom_violin(
    aes(
      x = functionalGroup,
      y = dryMass
    ),
    draw_quantiles = c(0.25, 0.5, 0.75)
  ) +
  ggtitle("Violin Plot of Dry Mass by Functional Group") +
  xlab("Functional Group") +
  ylab("Dry Mass")

# Print the plots
print(boxplot)
```




```
print(violin_plot)
```



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

> Answer: Boxplot is more effective visualization option than Violin plot. Boxplots are particularly effective for comparing distributions across different groups and work well with smaller datasets and can effectively highlight outliers and the spread of the data. While violin plot provides more detailed information and include the density distribution, which can be useful but might not be necessary for all analyses.

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

> Answer: Twigs/branches and needles generally have the highest biomass at these sites. This conclusion is based on the outliers observed, with needles having approximately 8.3 units of biomass and twigs/branches having around 5.8 units of biomass.