Assignment 3: Data Exploration

Livia Hoxha

Fall 2023

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.

```
#First step is to check my working directory
getwd()
```

```
## [1] "C:/Users/Lenovo/Desktop/EDE_Fall2023"
```

```
#Here I import the datasets,
#Then I name them according to the directions of the assignment
#I also included the subcommand to read strings in as factors

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)
```

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: According to online sources, Neonicotinoids are one of the most widely used classes of insecticides in agriculture today. Understanding their toxicity and environmental impact is crucial.There is concern that neonicotinoids may be linked to declining insect populations, particularly bees and other pollinators. Studying their ecotoxicology can shed light on any harmful effects.Insects play vital roles in ecosystems, such as pollination, decomposition, and as food sources for other species. If neonicotinoids are significantly impacting insect populations, this could have far-reaching ecological consequences. Neonicotinoids act on the nervous systems of insects. Researching their mode of action and metabolism in different insect species can help predict toxicity and exposure risks. There are knowledge gaps around the chronic and sublethal effects of neonicotinoids. Ecotoxicology studies can provide important data to regulatory agencies tasked with weighing the costs and benefits of these pesticides. Understanding how neonicotinoids move through the environment (water, soil, plants etc) and impact non-target insects can lead to insights on mitigating any negative effects.

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: Some reasons why the study of forest litter and woody debris is important are: - Fallen leaves, twigs, and larger woody material are key components of forest nutrient cycling. As this debris decomposes, nutrients are returned to the soil to support new plant growth. Understanding litter dynamics informs models of forest ecosystem function. - The amount and composition of litter and woody debris influences forest regeneration and growth. Decomposition releases nutrients but also can form physical barriers to seedling establishment. - Litter depth and woody debris influence forest habitat structure, impacting many animal species. For example, fallen logs provide habitat for insects, fungi, amphibians, and small mammals.- Forest litter and woody debris affect wildfire risk and behavior. Greater biomass accumulation provides more fuel for potential intense fires. Studying debris decomposition rates and moisture content can help predict fire danger. - Changes in litter and woody debris over time, due to factors like climate change or invasive pests, can indicate broader forest health issues. These datasets provide baselines for monitoring ecosystem change.- Regional variations in litter and woody debris related to tree species, climate, topography etc. can improve understanding of broader ecological patterns and processes. So, analyzing the accumulation and decomposition of forest litter and woody debris provides insight into fundamental ecosystem functions, habitats, and natural disturbance regimes in forests. The Niwot Ridge dataset will help characterize these dynamics in Colorado subalpine forests.

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 is collected from elevated traps that are 0.5 mˆ2 PVC baskets elevated 80 cm above the ground. Fine woody debris is collected from 3 m x 0.5 m rectangular ground traps.

2.In forested sites, trap placement is randomized from a grid of possible locations. In non-forested sites, trap placement is targeted beneath areas of woody vegetation. 3. Traps are sampled at varying frequencies depending on site - from monthly to annually. Deciduous forests are sampled more frequently during leaf fall. Each trap collects litter sorted into functional groups like leaves, twigs, seeds, etc.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# Here we see the dimensions of the dataset

dim(Neonics)
```

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

```
dim
```

```
## function (x)  .Primitive("dim")
```

```
# 4623 are the number of rows in the dataset,whereas 30 is the number of 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?

```
#Here I use the "summary" function on the Effect column.
#We print the summary to see the most common effects

effect_summary <- summary(Neonics$Effect)
print(effect_summary)
```

```
##     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: Mortality (1493 occurrences): Mortality is a paramount concern because it represents the direct and immediate impact of neonicotinoid pesticides on insect populations. High mortality rates can lead to population declines, potentially affecting crop pollination and ecosystem balance.Also, Population (1803 occurrences): Studying the population effects of neonicotinoids helps assess how these pesticides influence the abundance and dynamics of insect populations. Large-scale population impacts can disrupt ecosystems and agricultural systems.Additionally, Behavior (360 occurrences): Behavioral changes, such as alterations in foraging and navigation patterns, are essential to study because they can affect an insect's ability to survive and reproduce. Behavioral disruptions can lead to reduced efficiency in resource acquisition and survival challenges.

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...]

```
#Now I use the summary function on the column "Common Name".
#Sort the summary results in descending order, because by doing that we focus on the values/categories
#Now I select the top six most commonly studied species.
#Lastly, I print the top species

species_counts <- summary(Neonics$Species.Common.Name)

sorted_species_counts <- sort(species_counts, decreasing = TRUE)

top_species <- head(sorted_species_counts, n = 6)

print(top_species)
```

```
##               (Other)             Honey Bee        Parasitic Wasp
##                   670                   667                   285
## Buff Tailed Bumblebee   Carniolan Honey Bee            Bumble Bee
##                   183                   152                   140
```

Answer:The species mentioned have several things in common, such as: being crucial pollinators, supporting plant reproduction and food production; found in diverse ecosystems, making them suitable for ecological research; and their presence in agricultural and urban settings makes them relevant for studies on human impacts on ecosystems. They might be of interest over other insects, because these species exhibit diverse behaviors and ecological roles, offering insights into ecosystem dynamics. Furthermore, they are extensively studied, facilitating research. Also, Honey Bees, in particular, play a vital role in crop pollination, contributing significantly to the economy.Lastly, declines in pollinator populations, including Bumble Bees, highlight the need for conservation efforts.

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?

```
class("Conc.1.Author")
```

```
## [1] "character"
```

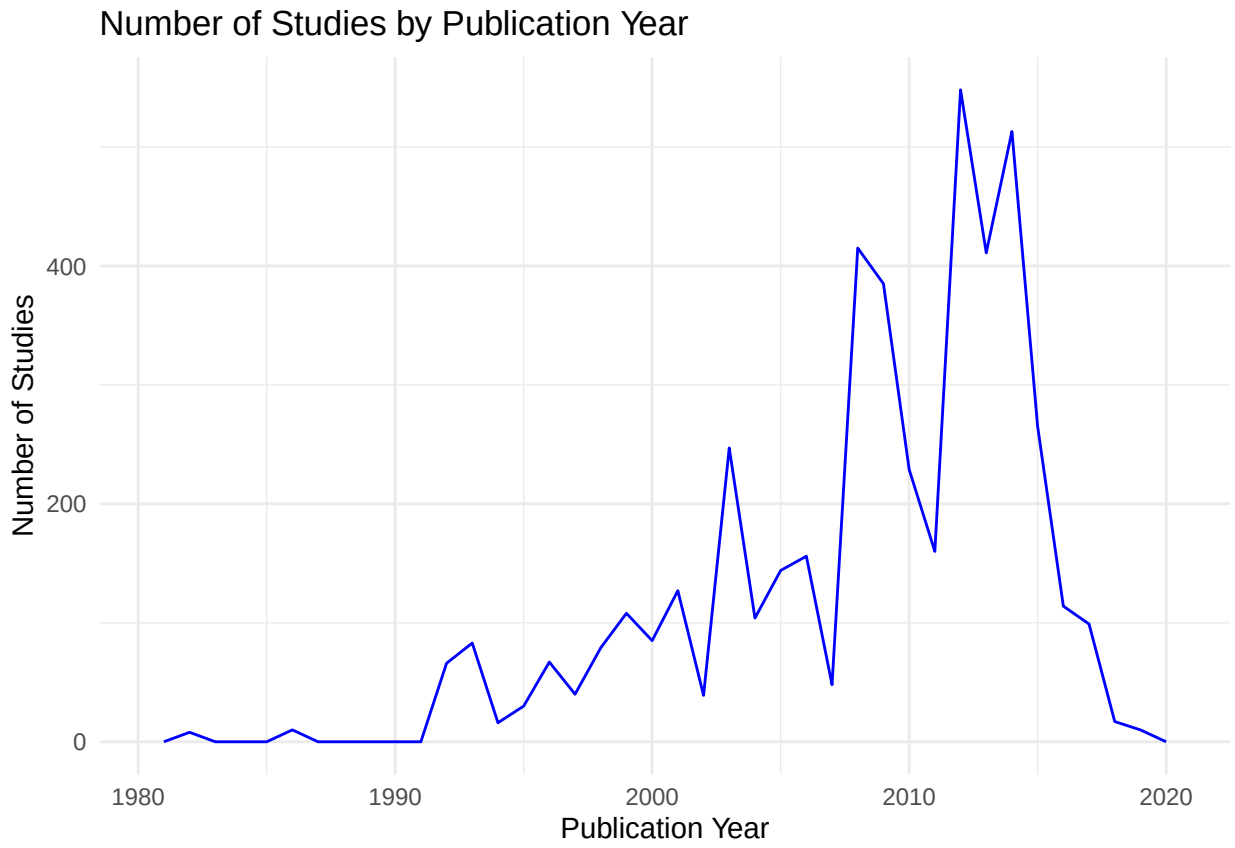
Answer: The class of "Conc.1.Author" is character. The reason why "Conc.1.Author" is not numeric, is because in our dataset, this column contains also non-numeric characters or symbols.

Explore your data graphically (Neonics)

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

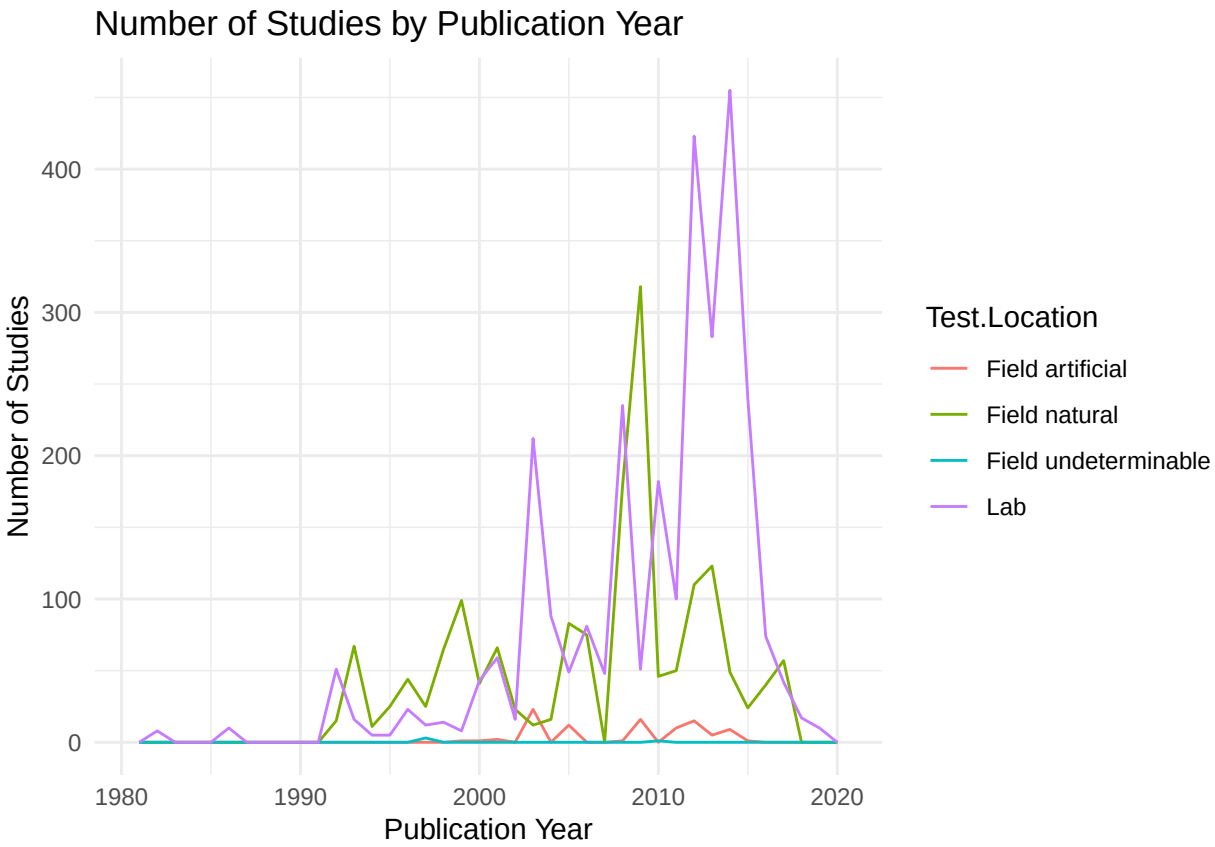
```
#First I load the package
library(ggplot2)

#Then I create the plot
```

```r
ggplot(data = Neonics, aes(x = Publication.Year)) +
  geom_freqpoly(binwidth = 1, color = "blue") +
  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(data = 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") +
  theme_minimal()
```



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

> Answer: The most common test locations are Lab, and as we can see, the number of studies conducted in Lab increase significantly as years pass. Then, the second most common test location is Field Natural, which reaches its peak just before the year 2010, then the rate declines.

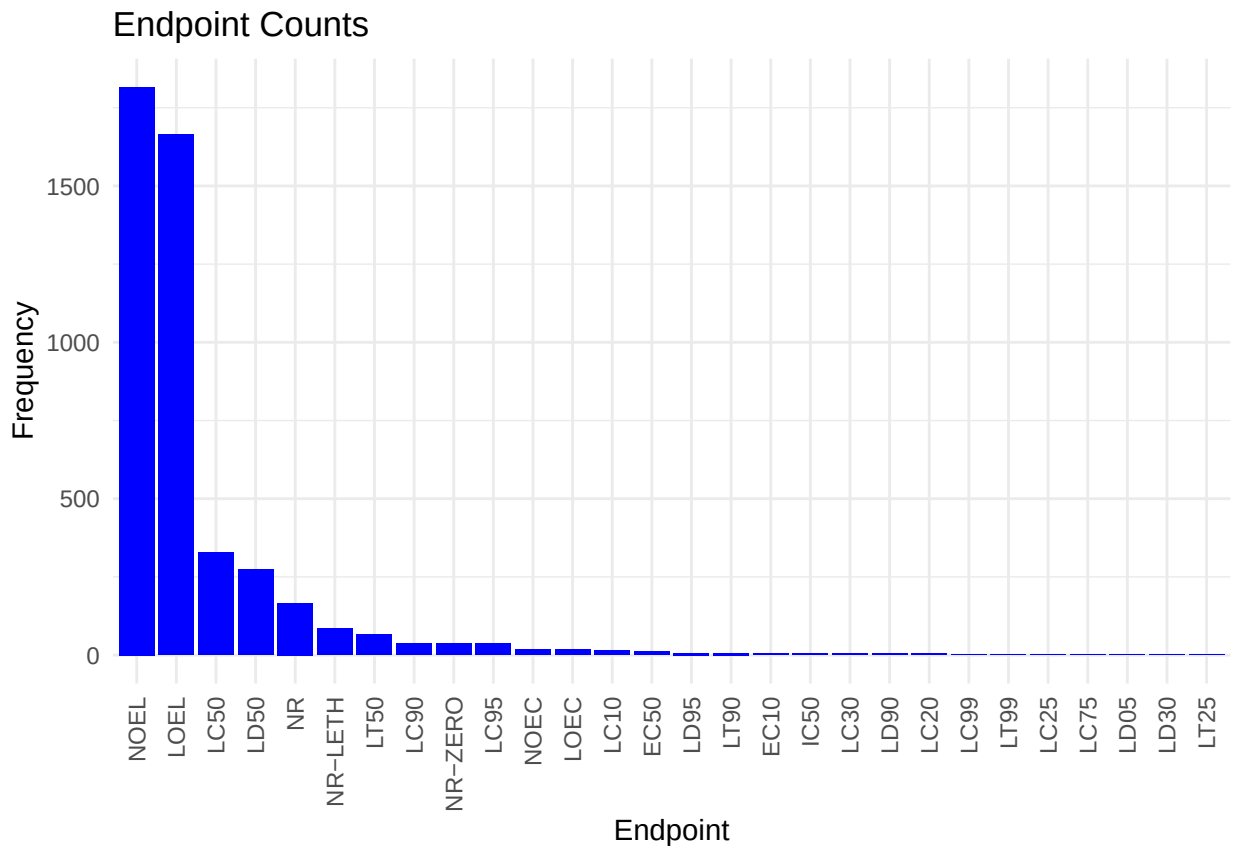
11. Create a bar graph of Endpoint counts. What are the two most common end points, and how are they defined? Consult the ECOTOX_CodeAppendix for more information.

[**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...]

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Here I calculate the counts of each unique endpoint
endpoint_counts <- table(Neonics$Endpoint)

# Now we convert the counts to a data frame for plotting
endpoint_counts_df <- data.frame(
  Endpoint = names(endpoint_counts),
  Frequency = as.numeric(endpoint_counts))

# Here we sort the data frame by frequency in descending order
endpoint_counts_df <- endpoint_counts_df %>%
  arrange(desc(Frequency))

# Now I create the bar graph
ggplot(data = endpoint_counts_df, aes(x = reorder(Endpoint, -Frequency), y = Frequency)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Endpoint Counts",
       x = "Endpoint",
       y = "Frequency") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



Answer:The two most common endpoints are NOEL and LOEL. NOEL (No Observed Effect Level):is the highest dose or concentration of a substance at which no observable adverse effects or significant changes in the studied organisms are detected during a specific exposure period. NOEL serves as a safety threshold, indicating when exposure to a substance does not result in discernible

harm or significant changes in the organisms being studied. LOEL (Lowest Observed Effect Level):is the lowest dose or concentration of a substance at which observable adverse effects or significant changes in the studied organisms are detected during a specific exposure period.LOEL represents the point at which adverse effects become evident, signaling that exposure to the substance is causing harm or significant alterations in the organisms.

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.

```
# Here I check the class of the collectDate variable
class(Litter$collectDate)
```

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

```
# Since it was a factor, now I change it to a date
Litter$collectDate <- as.Date(Litter$collectDate)
#Check the class of the variable after the change and it is now a date
class(Litter$collectDate)
```

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

```
#Now I use the unique function to find dates when litter was sampled in August 2018
unique_dates <- unique(Litter$collectDate)
august_2018_dates <- unique_dates[format(unique_dates, "%Y-%m") == "2018-08"]
august_2018_dates
```

```
## [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`?

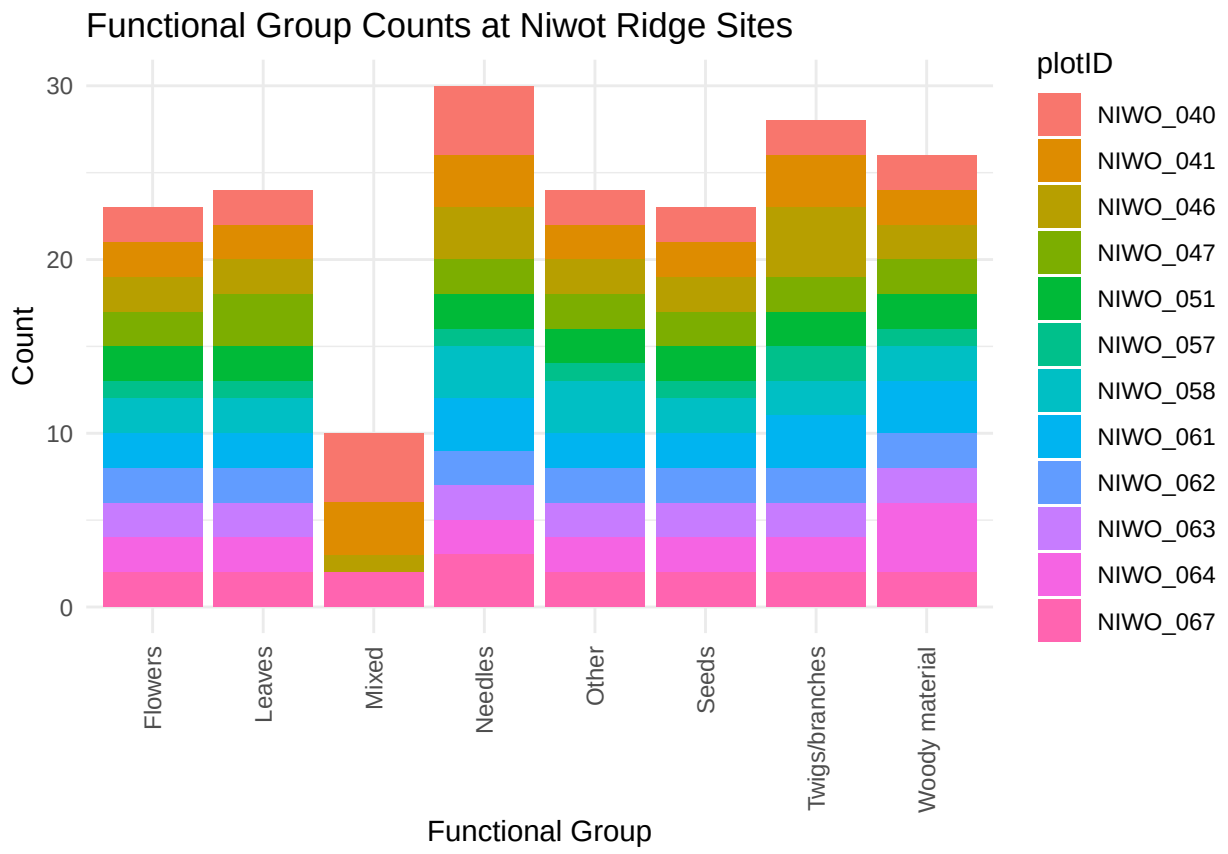
```
unique_plots <- unique(Litter$PlotID)
num_plots <- length(unique_plots)
num_plots
```

```
## [1] 0
```

Answer: Unique is used to find and list unique values, while summary provides a broader range of statistics and information about a variable.

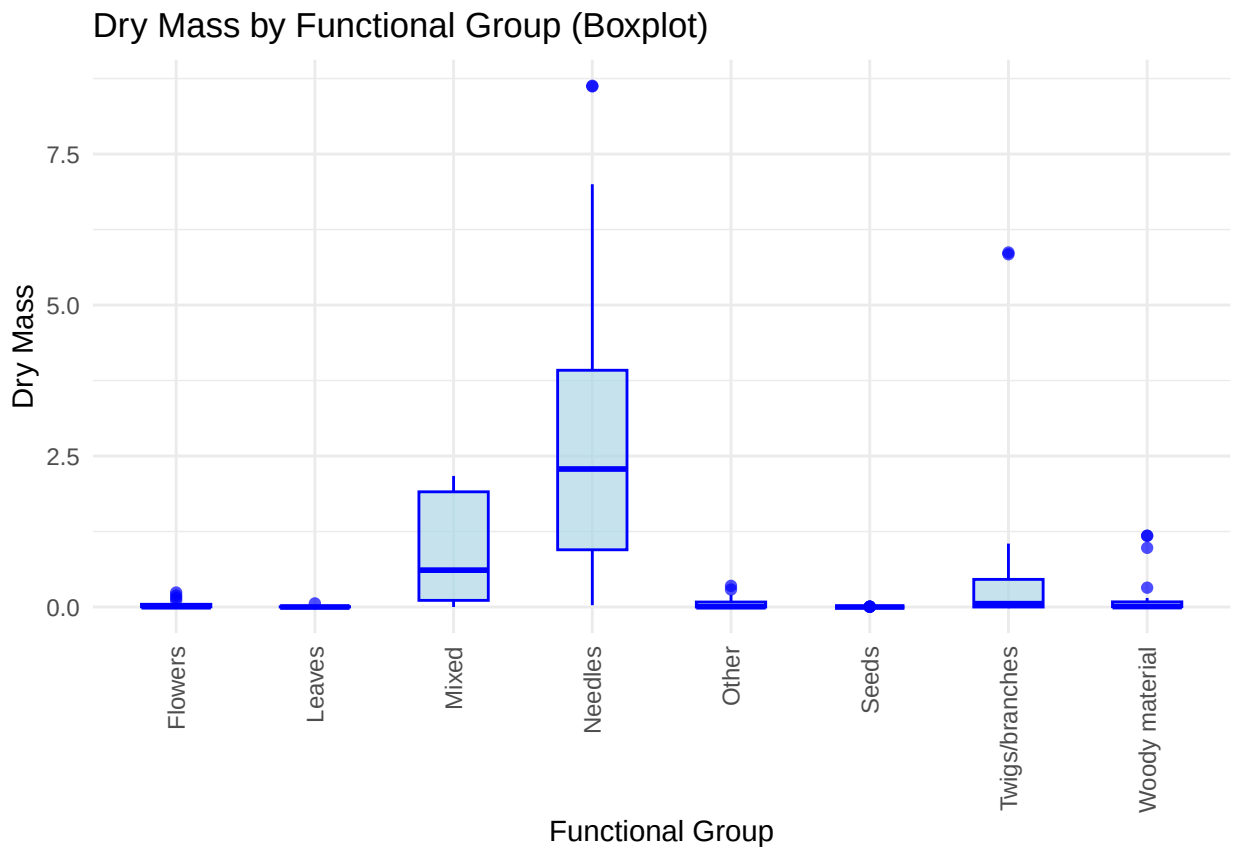
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x = functionalGroup, fill = plotID)) +
  geom_bar() +
  labs(title = "Functional Group Counts at Niwot Ridge Sites",
       x = "Functional Group",
       y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```

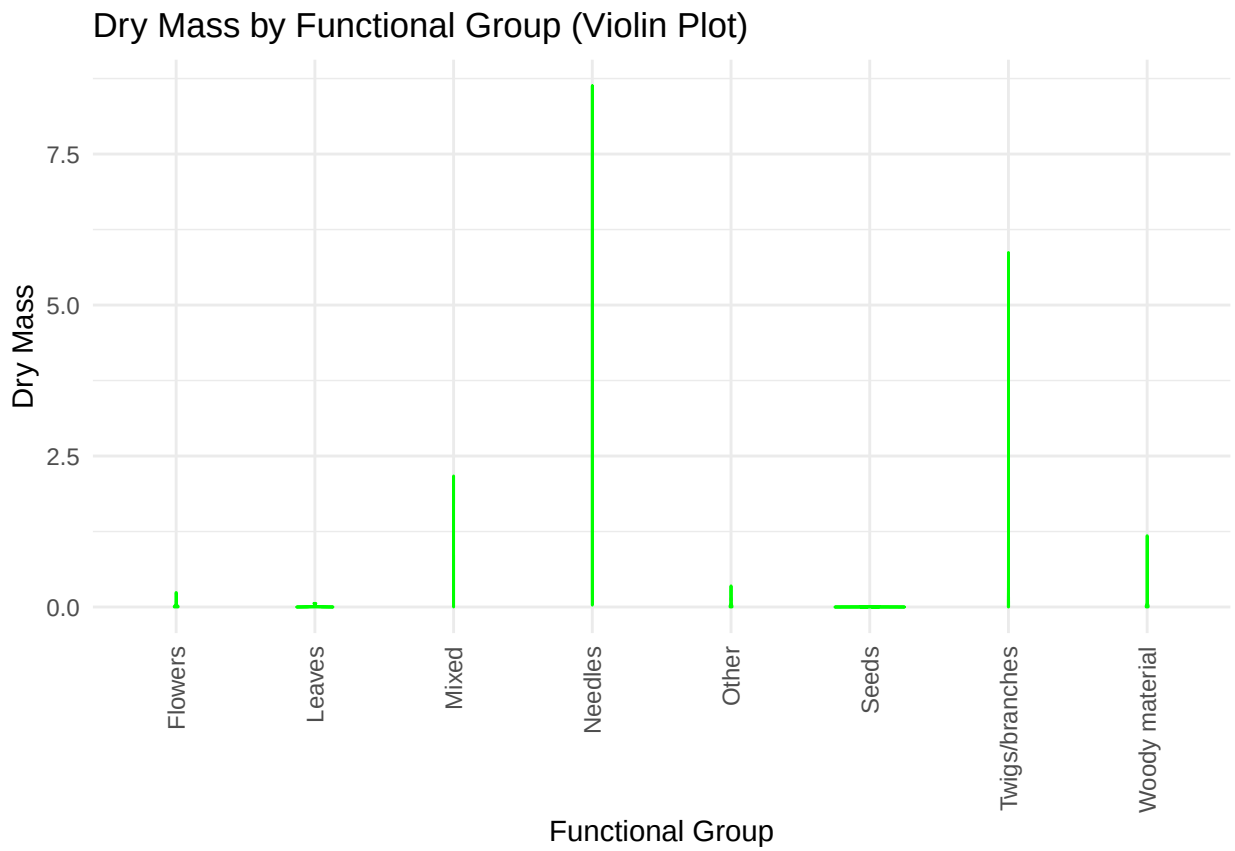


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# Boxplot for dryMass by functionalGroup
boxplot_plot <- ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_boxplot(fill = "lightblue", color = "blue", alpha = 0.7, width = 0.5) +
  labs(title = "Dry Mass by Functional Group (Boxplot)",
       x = "Functional Group",
       y = "Dry Mass") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
print(boxplot_plot)
```




```r
# Violin plot for dryMass by functionalGroup
violin_plot <- ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_violin(fill = "lightgreen", color = "green", alpha = 0.7, width = 0.5) +
  labs(title = "Dry Mass by Functional Group (Violin Plot)",
       x = "Functional Group",
       y = "Dry Mass") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
print(violin_plot)
```



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

> Answer: In our case, the distribution of dry mass values within each functional group.The boxplot reveals central tendency (the horizontal line represents the median for a specific functional group), spread (the height of each box represents the interquantile range), and also outliers (as we can see from our plot, we can notice individual data points that fall outside the range). Whereas, the violin plot shows only the median, which we already saw in the boxplot.

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

> Answer: The groups with the highest median dry mass(horizontal line) have on average the highest biomass. In our case, it's the Needles and Twigs/branches.