Assignment 3: Data Exploration

Rachael Stephan

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 in packages
library(tidyverse); library(lubridate); library(here)

#check working directory
getwd()
```

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

```r
#check current location
here()
```

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

```r
#upload data sets with strings as factors
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: Neonictinoids are a common pesticide that have recently been suspect of having deleterious effects on insect populations, including acting as neurotoxins. Neonictinoids have a high toxicity for invertebrates, and their broad spectrum application mean they will impact many taxa. This can become an issue when the pesticides begin to have impacts on nontarget species that provide important ecosystem services, such as pollinators. Understanding how common agriculture pesticides impact invertebrates can help us preserve healthy populations of plants and insects.

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: litter and woody debris may be important components of the carbon cycle and the ecosystem. The litter and woody debris will release carbon and organic matter into the soil as they decompose to facilitate new growth and uptake. Certain animals may also use them as a food source (i.e., decomposers) or as a habitat. Additionally, wer may be interested in these for safety reasons because litter and woody debris may increase fire risk as they can act as tinder.

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. Litterfall and woody debris are quantified by mass captured in aerial and ground traps /n 2. Masses < 0.01 g are reported for presence of functional groups but not quantified/n 3. finest temporal resolution is day of trapping; finest spatial resolution is one trap/n 4. latitude, longitude, and elevation of the plot center are recorded

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#get dimensions of Neonics
dim(Neonics)
```

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

Answer: 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 summary statistics of the effects of neonictinoids and sort them from high to low
sort(summary(Neonics$Effect),decreasing = TRUE)
```

```
##       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 top 5 most common effects from most common to less common are population, mortality, behaviour, feeding behaviour, and reproduction. These effects may be of specific interest because it is important to understand the mechanisms through which pesticides impact insects. Looking at commanalities between these effects can indicate possible target insect systems (e.g., nervous system) and the relative toxicity insecticides pose to these systems (i.e., more common effect meaning that insecticide poses a higher toxicity to the related organs/systems). The effects listed, such as mortality, are also severe effects that indicate this may be a significant issue to resolve.

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...]

```
#get summary statistics of the studied species
#note: sort is not needed with maxsum because maxsum automatically takes the most frequent levels.
#     n+1 levels are needed to be specified because one level is "other".
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 top six species are all bees or wasps. These are all important pollinators. Pollinators perform important ecosystem services of pollinating plants to ensure their successful reproduction. Without these species, or with impaired individuals, plant reproduction would falter and the parts of the ecosystem they support would also suffer.

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...]

```
#determine the class of the `Conc.1..Author.` column
class(Neonics$Conc.1..Author.)
```

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

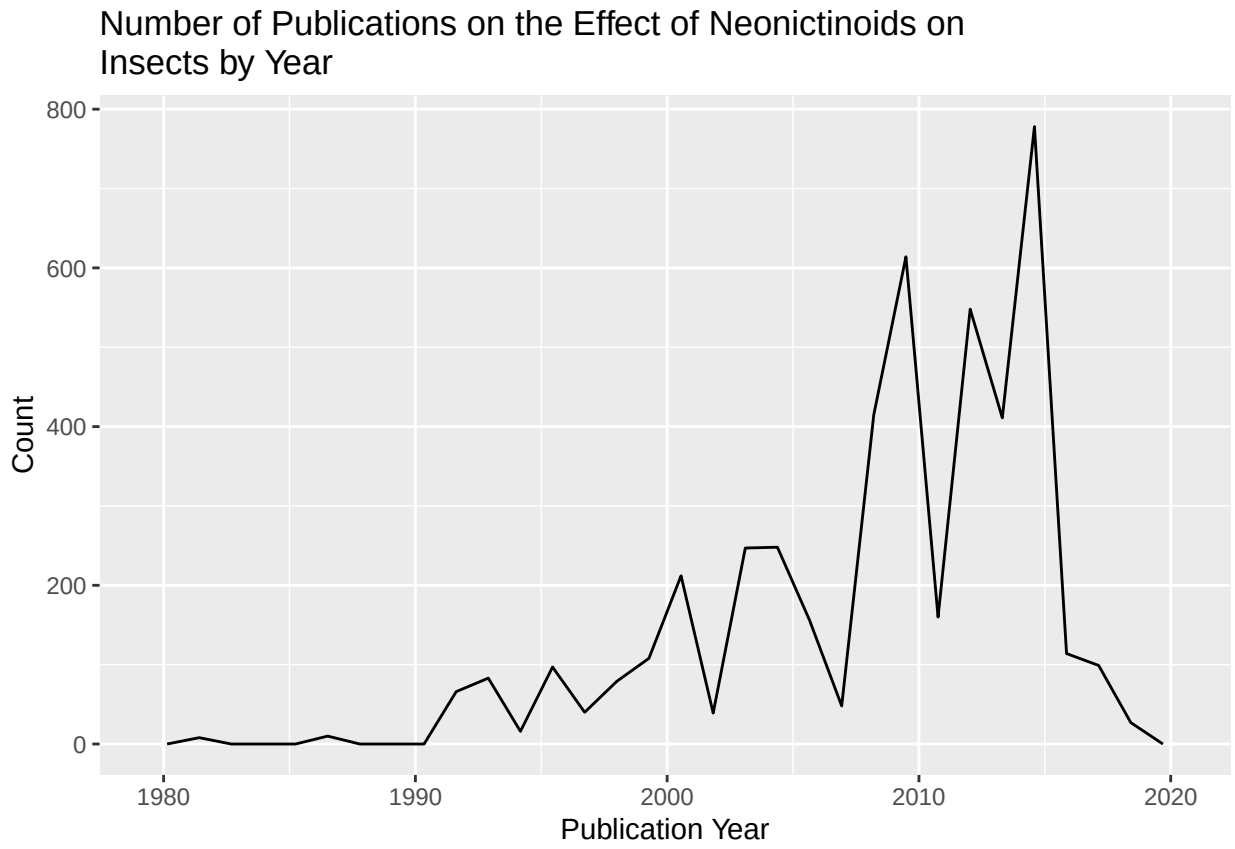
Answer: The column has a class of factor. The column represents the concentration of the chemicals, but the scales of these concentrations are not all the same. Some values are numeric (although it would not be helpful to have all of these values on the same numerical axis because their units are different and the relative relationships would not be accurate). However, others are not numerical and include characters (e.g., 144/, NR, >=13.66, and ~41). Since all strings were specified to be factors, the class of this column is factor.

Explore your data graphically (Neonics)

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

```
#create a frequency polygon graph to show how many studies were conducted per year
ggplot(Neonics)+
  geom_freqpoly(aes(x = Publication.Year))+
  labs(
    title = "Number of Publications on the Effect of Neonictinoids on\nInsects by Year",
    x = "Publication Year",
    y = "Count")
```

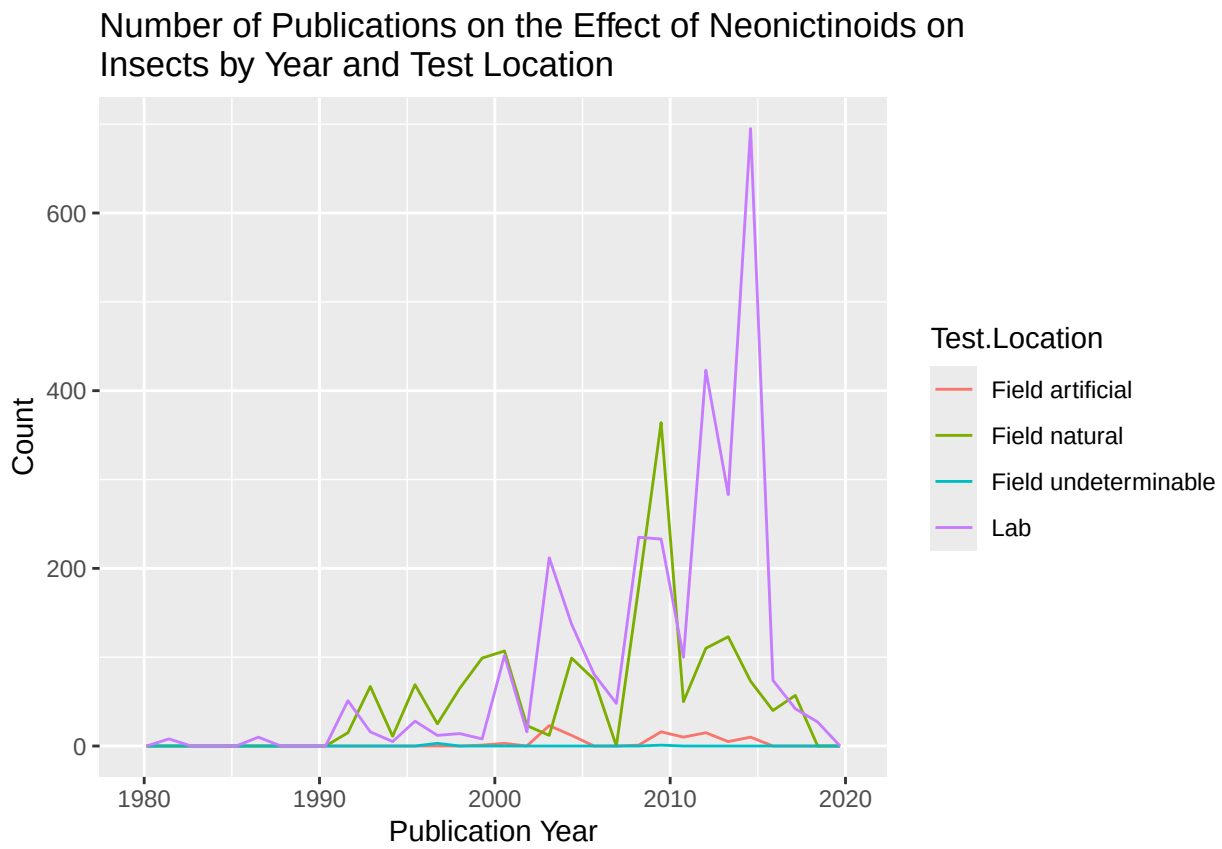
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
#create a frequency polygon graph as above but specify the colour is based on location
ggplot(Neonics)+
  geom_freqpoly(aes(x = Publication.Year, colour = Test.Location))+
  labs(
    title = "Number of Publications on the Effect of Neonictinoids on\nInsects by Year and Test Locatio
    x = "Publication Year",
    y = "Count")
```

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



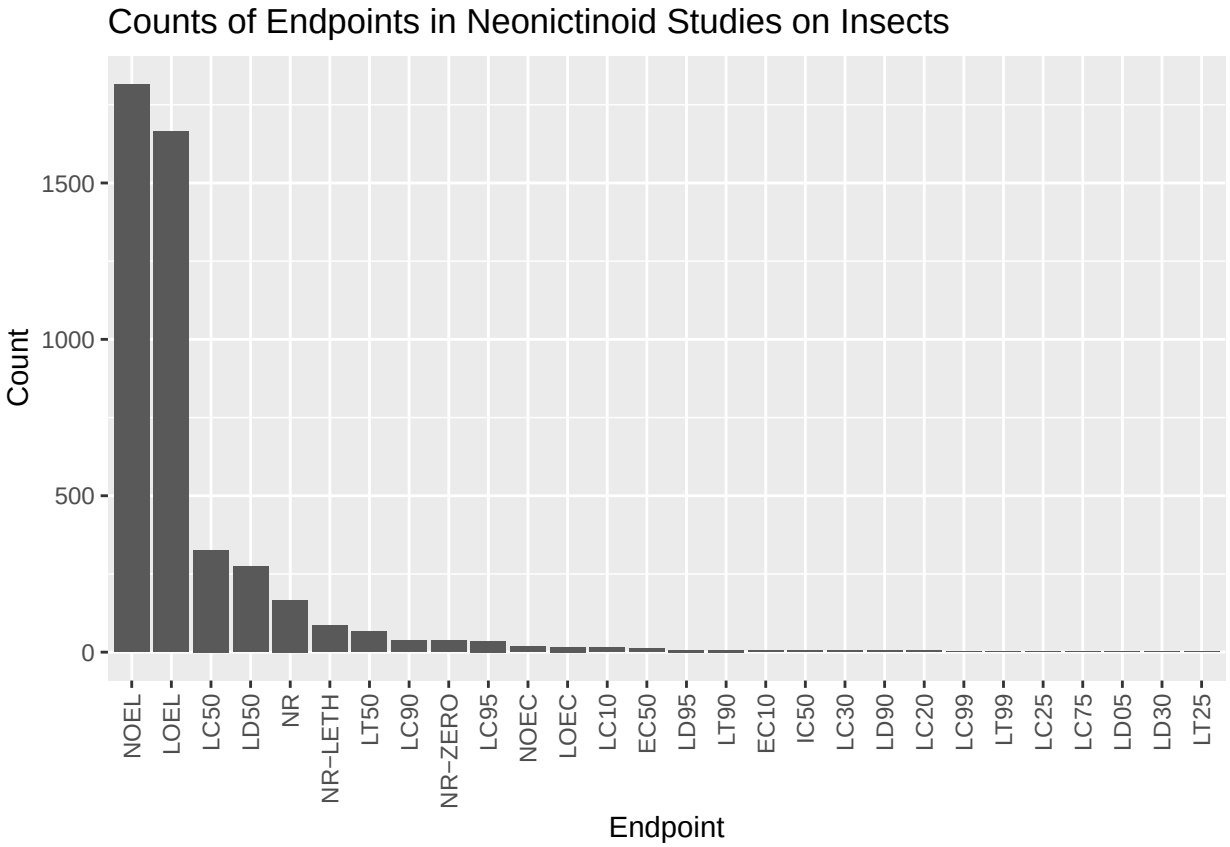
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test location is the laboratory. Although laboratory studies have always been popular, there were some periods of time (i.e., ~1993-2001 and 2009) where natural field studies were more popular (or more published). Artifical field and undeterminable field studies have not been very popular throughout this data set.

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 graph that shows the common endpoints with the most common at the origin
ggplot(Neonics)+
  geom_bar(aes(x = fct_infreq(Endpoint)))+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  labs(
    title = "Counts of Endpoints in Neonictinoid Studies on Insects",
    x = "Endpoint",
    y = "Count")
```



Answer: The top endpoint is NOEL. NOEL stands for 'no observable effect level' and is used in the database for terrestrial usage. NOEL is defined as the highest concentration at which there is no statistical difference from the control. The second top endpoint is LOEL. LOEL stands for 'lowest observable effect level' and is used in the database for terrestrial usage. LOEL is defined as the lowest concentration (dose) at which there is a statistical difference 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.

```
#determine the class of collectdate using the class function
class(Litter$collectDate)
```

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

```
#check the format of the date
head(Litter$collectDate)
```

```
## [1] 2018-08-02 2018-08-02 2018-08-02 2018-08-02 2018-08-02 2018-08-02
## Levels: 2018-08-02 2018-08-30
```

```
#The class is changed to date using lubridate
Litter$collectDate <- ymd(Litter$collectDate)

#check class
class(Litter$collectDate)
```

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

```
#check correct year month day is saved correctly
head(format(Litter$collectDate, format = "%B %d, %Y"), 3)
```

```
## [1] "August 02, 2018" "August 02, 2018" "August 02, 2018"
```

```
#check which dates litter was sampled
format(unique(Litter$collectDate), format = "%B %d, %Y")
```

```
## [1] "August 02, 2018" "August 30, 2018"
```

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`?

```
#check the unique plots at Niwot ridge
unique(Litter$plotID)
```

```
##  [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_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

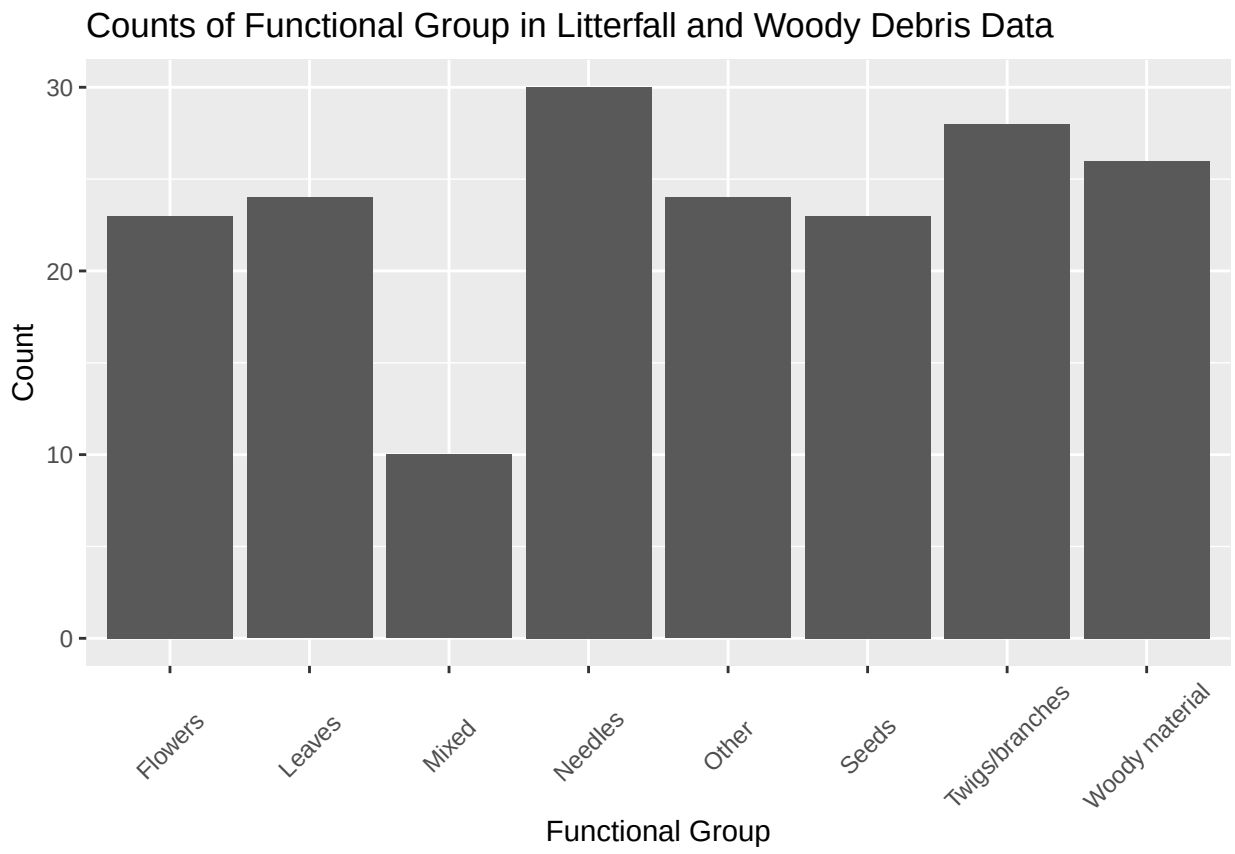
```
#check the summary of the plots at Niwot ridge
summary(Litter$plotID)
```

```
## 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 different plots sampled at Niwot Ridge. The unique function provides the levels of plots at Niwot Ridge. The summary function provides the count for each level. The summary function therefore provides all of the information the unique function provides and more (i.e., count).

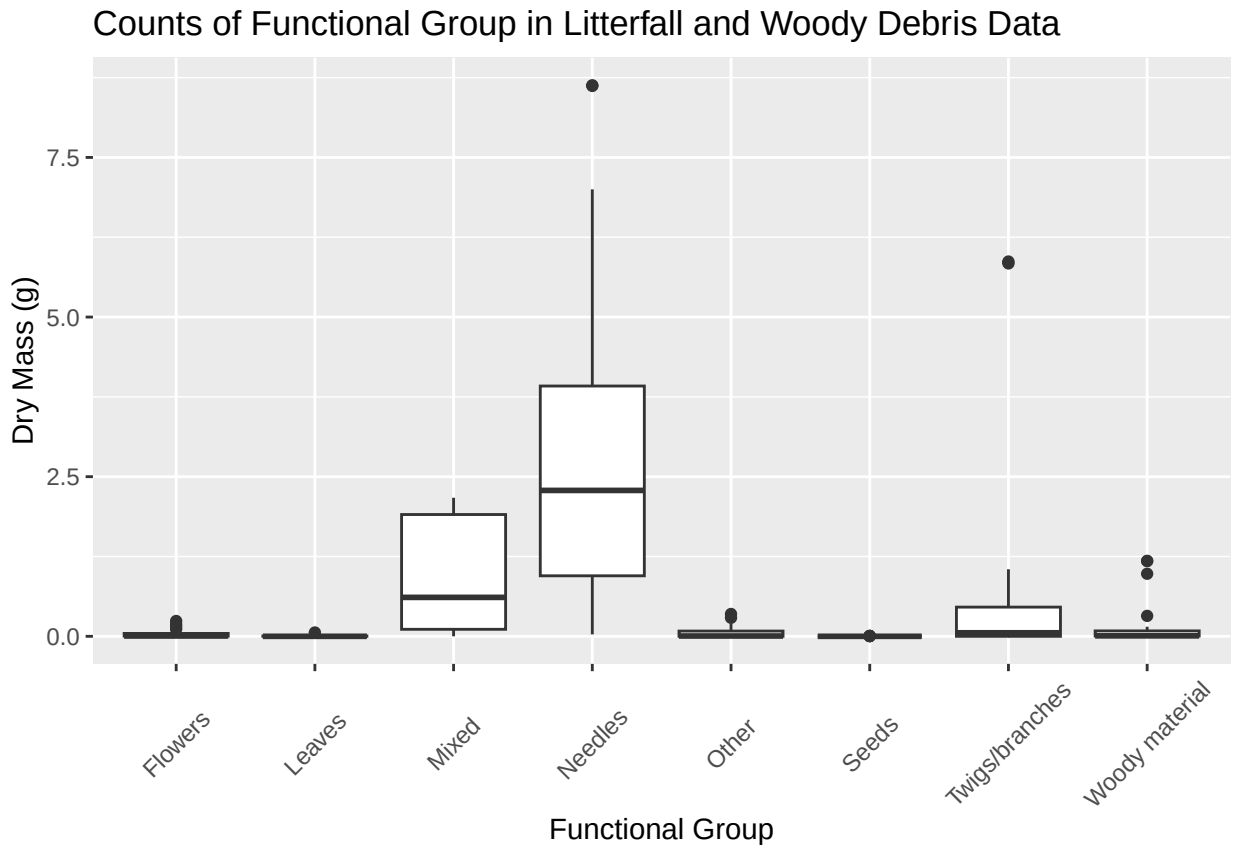
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#create bar graph specifying axes and titles
ggplot(data = Litter, aes(x = functionalGroup))+
  geom_bar() +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5))+
  labs(
    title = "Counts of Functional Group in Litterfall and Woody Debris Data",
    x = "Functional Group",
    y = "Count")
```

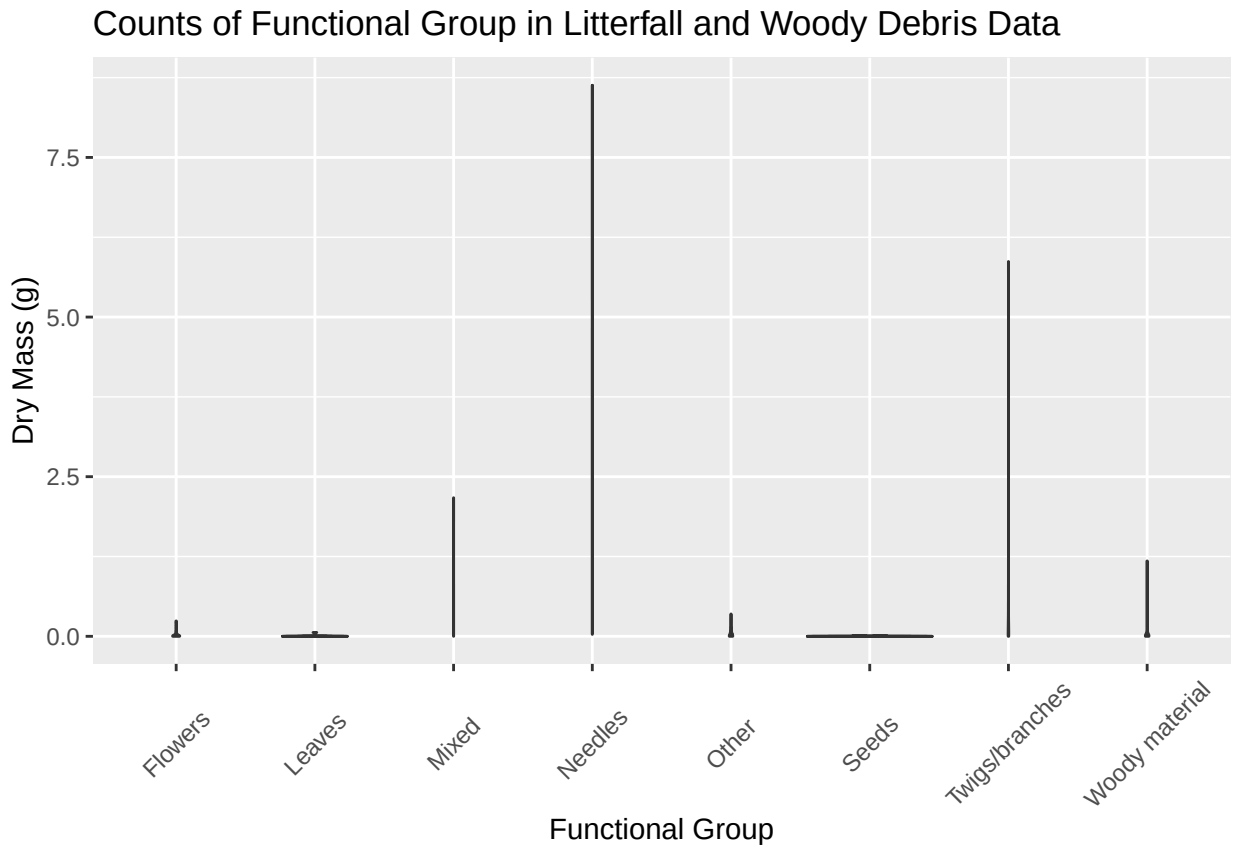


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# create a boxplot
ggplot(data = Litter)+
  geom_boxplot(aes(x = functionalGroup, y = dryMass))+
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5))+
  labs(
    title = "Counts of Functional Group in Litterfall and Woody Debris Data",
    x = "Functional Group",
    y = "Dry Mass (g)")
```



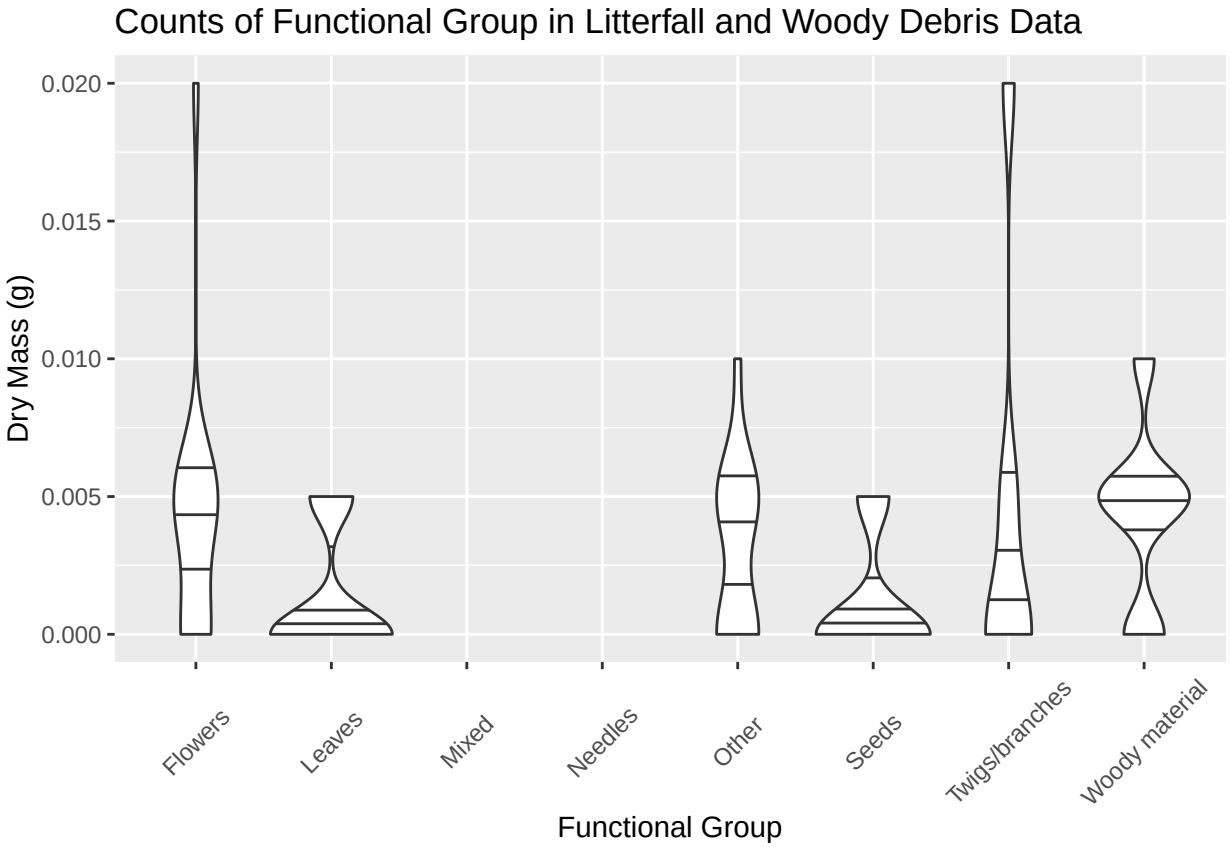

```
# create a violin plot
ggplot(data = Litter)+
  geom_violin(aes(x = functionalGroup, y = dryMass),
              drop = FALSE,
              draw_quantiles = c(0.25, 0.5, 0.75))+
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5))+
  labs(title = "Counts of Functional Group in Litterfall and Woody Debris Data",
    x = "Functional Group",
    y = "Dry Mass (g)")
```




```
# The violin plot distribution for many functional groups seems to be concentrated
# at low mass levels.Using ylim, look at the lowest masses to see if this is the
# case for these distributions
ggplot(data = Litter)+
  geom_violin(aes(x = functionalGroup, y = dryMass),
              drop = FALSE,
              draw_quantiles = c(0.25, 0.5, 0.75))+
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5))+
  labs(title = "Counts of Functional Group in Litterfall and Woody Debris Data",
    x = "Functional Group",
    y = "Dry Mass (g)")+
  ylim(0, 0.02)
```

```
## Warning: Removed 82 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
```

```
## Warning: Cannot compute density for groups with fewer than two datapoints.
```



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

> Answer: The violin plots does not well illustrate the distributions. The values are mostly all concentrated in between masses of 0g and 0.02g except for mixed, needles, and twig/branches functional groups, which have larger distributions. These smaller distributions are not well illustrated due to the orders of magnitude of difference between the distributions. There are also a couple groups with designated 'outliers'. This is shown on the violin plot as a straight line connecting these outliers to the majority of the data. It is not clear how many outliers lie between the largest mass outlier and most of the values. The box plot more clearly represents the distribution of the functional groups at larger and smaller scales as well as better illustrates the outliers.

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

> Answer: Needles tend to have the highest biomass at these sites followed by mixed and then twig/branches.