Assignment 3: Data Exploration

Samantha Pace

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.

```
getwd() # This code chuck is checking the working directory, which is the EDE_Fall2023 folder
```

```
## [1] "/home/guest/R/EDE_Fall2023"
```

```
# install.packages("tidyverse") # This code chunk is installing tidyverse,
# but now is in a comment so that it will knit.
# install.packages("lubridate") # This code chunk installed lubridate.

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate) # These code chunks are loading the packages.
```

```
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE)
# This R code chunk is creating a new data frame called Neonics.
Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv", stringsAsFactors = TRUE)
# This R code chunk is creating a new data frame from the CSV file called Litter.
```

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 may be interested in the ecotoxicology of neonicotinoids on insects in order to understand if the neonicotinoids are functioning successfully as incecticides. We may want to know how well they are successfully killing insects and when they are unsuccessful or less successful, what is happening then. Furthermore, we may want to know what is the context around the successes and failures of neonicotinoids as insecticides.

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 may be interested in studying litter and woody debris that fall to the ground in forests because it affects the ecosystem, specifically how nutrients and energy moves through the ecosystem. Woody debris and litter affects and provides habitats for aquatic and terrestrial species, can indicate important trends in the forest health, and may show human interaction with the land.

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 sampled through raised traps, and woody debris through ground traps. The data reported is for a single trap at a given time. 2. Ground trap sample collection for woody debris occurs annually. Litter sampling is done once every 1-2 weeks in deciduous forests and once every 1-2 months in evergreen forests. The traps may be inaccessible during the winter so there may be months with out frequent collections. 3. Sampling occurs at NEON sites that are home to woody plants that are at least 2 meters tall in designated plots. These plots are 400 square meters large. There is at least one litter trap and one raised trap for each plot.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) # This r code chunk provides the number of columns and rows.
```

```
## [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?

```
summary(Neonics$Effect)
```

```
##     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
```

```
# This code provides a summary of the Effect column in the Neonics dataset.
```

Answer: The most common effects that are studied are Population (with 1803 observations), Mortality (with 1493 observations), Behavior (with 360 observations), and Feeding behavior (with 255 observations). These effects may be of interest because they are the most common effects and determine if neonicotinoids are effective as insecticides and exactly what the effect is.

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...]

```
summary(Neonics$Species.Common.Name)
```

```
##                          Honey Bee                      Parasitic Wasp
##                                667                                285
##              Buff Tailed Bumblebee                Carniolan Honey Bee
##                                183                                152
##                         Bumble Bee                   Italian Honeybee
##                                140                                113
##                    Japanese Beetle                  Asian Lady Beetle
##                                 94                                 76
##                     Euonymus Scale                           Wireworm
##                                 75                                 69
##                  European Dark Bee                  Minute Pirate Bug
##                                 66                                 62
##                Asian Citrus Psyllid                     Parastic Wasp
##                                 60                                 58
##             Colorado Potato Beetle                   Parasitoid Wasp
##                                 57                                 51
##                Erythrina Gall Wasp                      Beetle Order
##                                 49                                 47
##        Snout Beetle Family, Weevil          Sevenspotted Lady Beetle
##                                 47                                 46
##                     True Bug Order              Buff-tailed Bumblebee
##                                 45                                 39
##                       Aphid Family                    Cabbage Looper
##                                 38                                 38
##                Sweetpotato Whitefly                     Braconid Wasp
##                                 37                                 33
##                       Cotton Aphid                    Predatory Mite
```

```
##                                33                                33
##              Ladybird Beetle Family                        Parasitoid
##                                30                                30
##                       Scarab Beetle                    Spring Tiphia
##                                29                                29
##                        Thrip Order               Ground Beetle Family
##                                29                                27
##                  Rove Beetle Family                    Tobacco Aphid
##                                27                                27
##                       Chalcid Wasp            Convergent Lady Beetle
##                                25                                25
##                     Stingless Bee                 Spider/Mite Class
##                                25                                24
##                 Tobacco Flea Beetle                 Citrus Leafminer
##                                24                                23
##                    Ladybird Beetle                         Mason Bee
##                                23                                22
##                            Mosquito                    Argentine Ant
##                                22                                21
##                              Beetle          Flatheaded Appletree Borer
##                                21                                20
##                 Horned Oak Gall Wasp                 Leaf Beetle Family
##                                20                                20
##                   Potato Leafhopper          Tooth-necked Fungus Beetle
##                                20                                20
##                       Codling Moth           Black-spotted Lady Beetle
##                                19                                18
##                       Calico Scale                 Fairyfly Parasitoid
##                                18                                18
##                        Lady Beetle              Minute Parasitic Wasps
##                                18                                18
##                          Mirid Bug                    Mulberry Pyralid
##                                18                                18
##                           Silkworm                     Vedalia Beetle
##                                18                                18
##                Araneoid Spider Order                         Bee Order
##                                17                                17
##                      Egg Parasitoid                     Insect Class
##                                17                                17
##             Moth And Butterfly Order        Oystershell Scale Parasitoid
##                                17                                17
## Hemlock Woolly Adelgid Lady Beetle               Hemlock Wooly Adelgid
##                                16                                16
##                              Mite                      Onion Thrip
##                                16                                16
##                Western Flower Thrips                     Corn Earworm
##                                15                                14
##                   Green Peach Aphid                        House Fly
##                                14                                14
##                           Ox Beetle                 Red Scale Parasite
##                                14                                14
##                  Spined Soldier Bug               Armoured Scale Family
##                                14                                13
##                    Diamondback Moth                     Eulophid Wasp
```

```
##                                13                                13
##                 Monarch Butterfly                   Predatory Bug
##                                13                                13
##             Yellow Fever Mosquito               Braconid Parasitoid
##                                13                                12
##                       Common Thrip      Eastern Subterranean Termite
##                                12                                12
##                             Jassid                        Mite Order
##                                12                                12
##                          Pea Aphid                  Pond Wolf Spider
##                                12                                12
##           Spotless Ladybird Beetle            Glasshouse Potato Wasp
##                                11                                10
##                           Lacewing           Southern House Mosquito
##                                10                                10
##            Two Spotted Lady Beetle                       Ant Family
##                                10                                 9
##                       Apple Maggot                         (Other)
##                                 9                               670
```

```
# This code provides a summary of the column called Species.Common.Name in the Neonics dataset.
```

Answer: The six most commonly studied species in this dataset are the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carnolan Honey Bee, Bumble Bee, and Italian Honeybee. All of these species for the most part are important and effective pollinators (potentially less so for the Parasitic Wasp). These insects might be of interest over the other insects because the Neonics is used as an insecticide on plants and crops and if the insecticide harms important pollinators, that could be detrimental to both the plants/crops and the bee populations. Generally, the goal of insecticides is to reduce insects that are pests and not ones that play important roles to the growth of the plants and ecosystem. Understanding the effects of insecticides on pollinator species rather than non-pollinator species would be of interest.

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

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

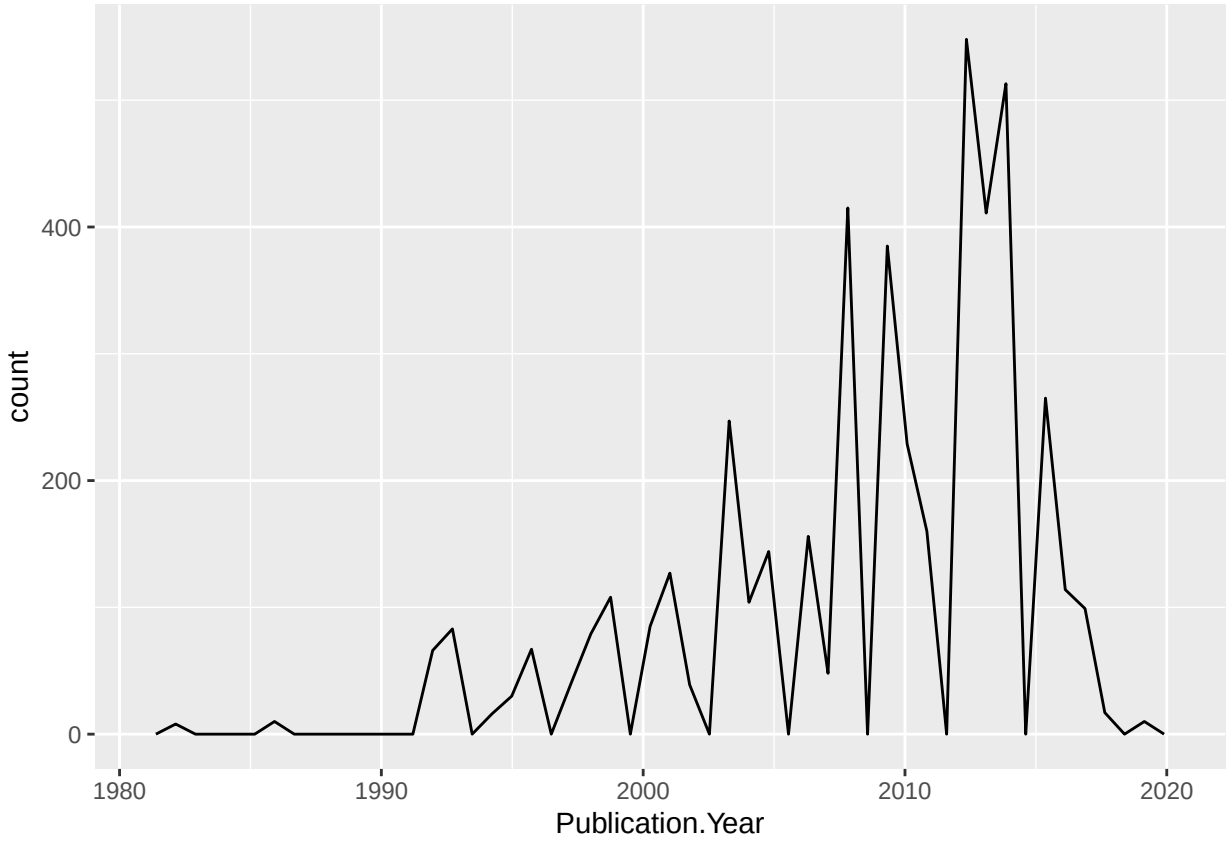
```
# This code determines the class of column 'Conc.1..Author'.
```

Answer: The class of Conc.1..Author is a factor. This is because when we read and imported the original raw data, the code included the phrase 'stringasfactors = true' so the columns are strings and not numeric.

Explore your data graphically (Neonics)

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

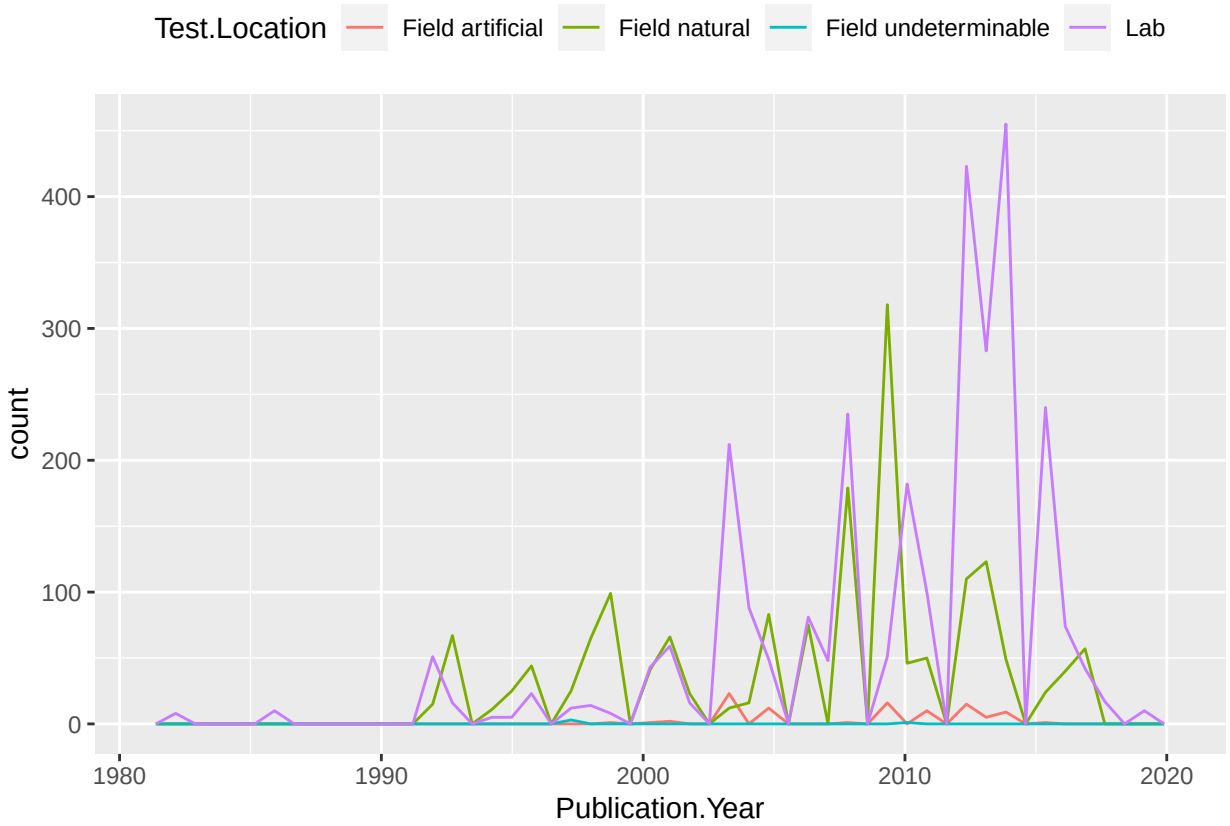
```
ggplot(Neonics) +
  geom_freqpoly(aes( x = Publication.Year), bins = 50)
```




```
# This code is creating a line graph of the number of studies conducted by
# publication year with 50 bins.
```

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

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 50) +
  theme(legend.position = "top")
```




```
# This code chuck is plotting the number of studies conducted by
# publication year based on test location. It includes a legend
```

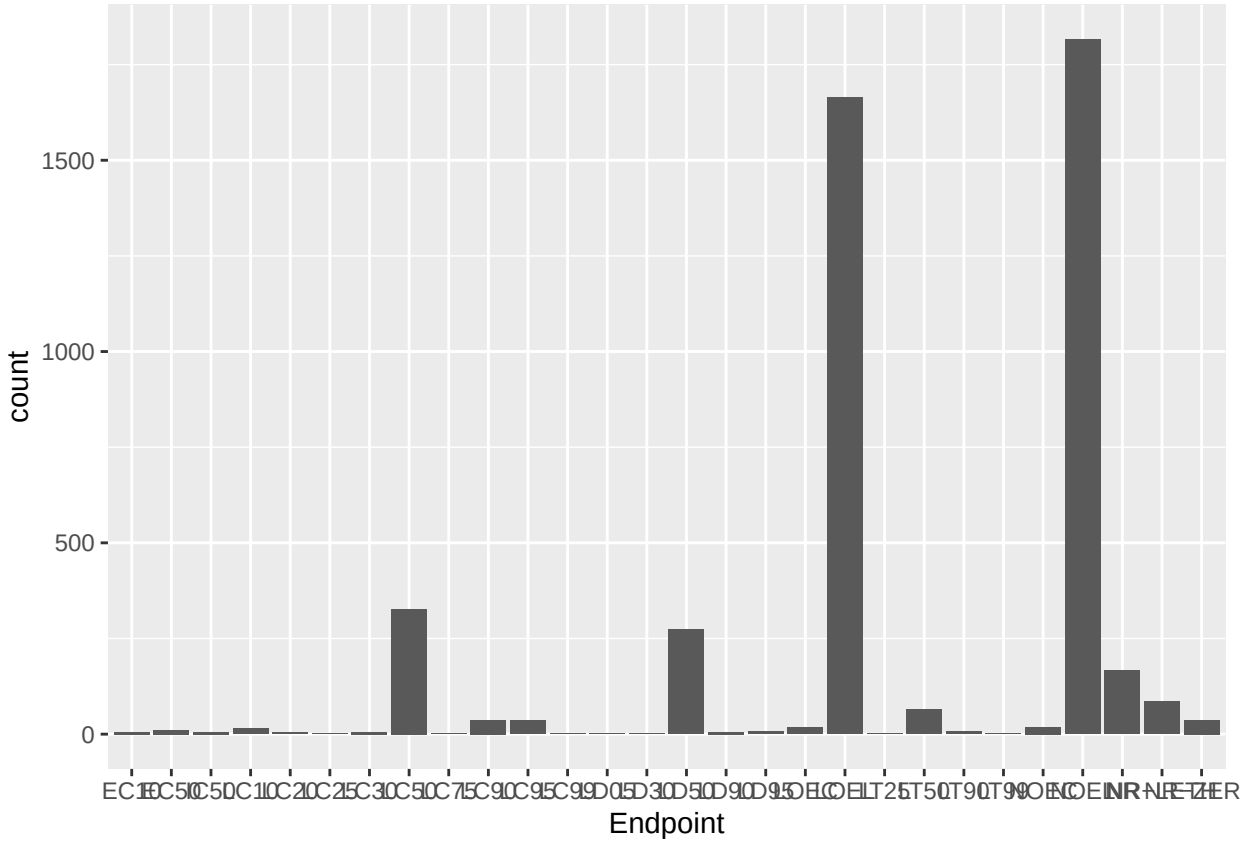
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test location are the Lab and Field natural. They do differ over time. Before about 2002, there were more natural field test locations and the count was about 100 or less studies. There were lab tests at this time but less. Between 2000-2010, there was an increase in the number of lab studies, but an even larger increase in the number of natural field studies. From 2010 to 2020, the number of lab studies continued to increase and the number of field studies declined per year.

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...]

```
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar()
```




```
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

```
## List of 1
##  $ axis.text.x:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 1
##   ..$ vjust        : num 0.5
##   ..$ angle        : num 90
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

```
  # This chunk is creating a bar chart of the Endpoint categories by count with ggplot.
```

Answer: The two most common endpoints are LOEL and NOEL. LOEL is 'Lowest-observable-effect-level'. This is the lowest concentrations that had effects noticeably distinguishable from the control groups. NOEL is 'No-observable-effect-level' which is the highest concentration that created relatively similar results to the control groups.

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.

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

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

```
# This code chunk determines the class of the column collectDate.
# It is a factor, not a date.
```

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

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

```
unique(Litter$collectDate)
```

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

```
# This code chuck is first converting the collect.Date column to be a date
# rather than a factor. Then it is checking that it is a date now. The last code
# is finding the unique collection dates in August 2018.
```

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(Litter$plotID)
```

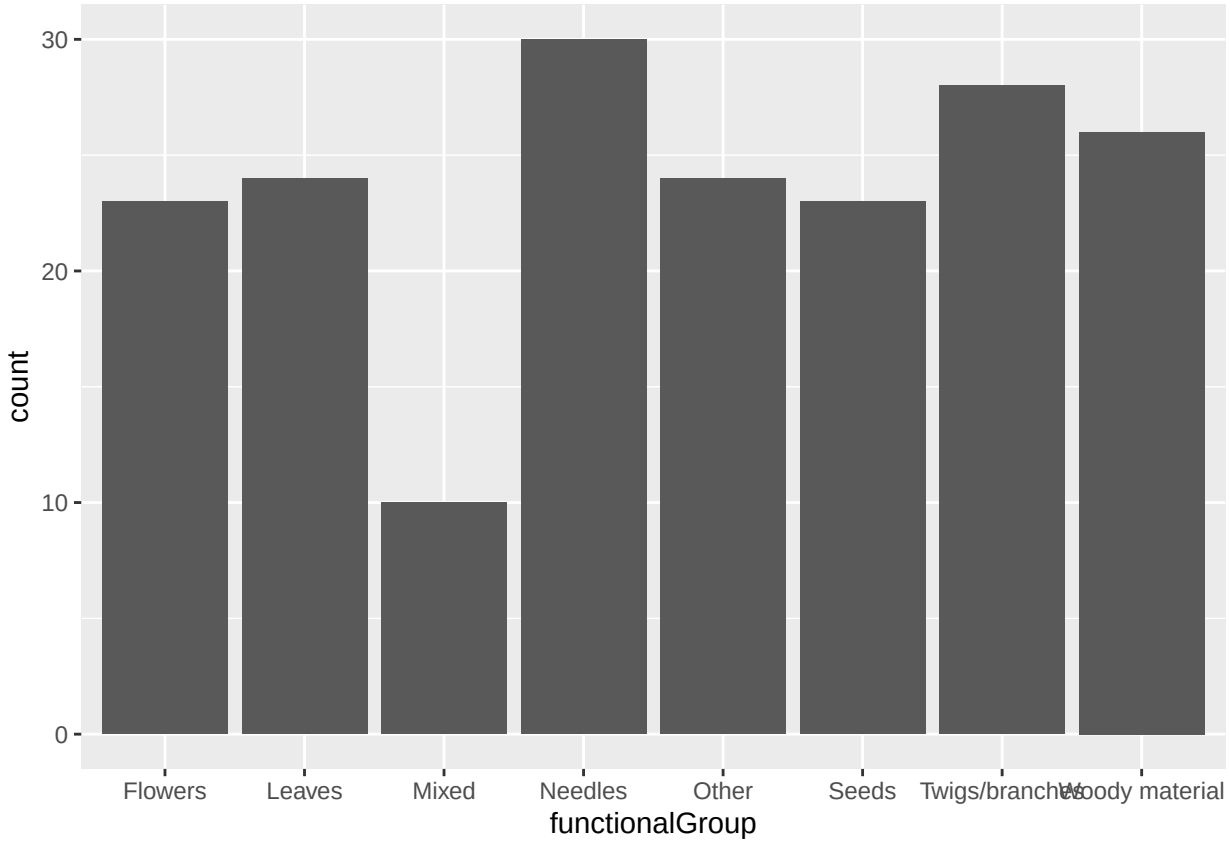
```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

```
# This code chunk is finding the unique plotID results.
```

Answer: There were a total of 12 plots sampled at Niwot Ridge. The information obtained from the 'unique' function is counting the total number of unique observations are found within the column. There are twelve and the results list the names of the plot IDs. The summary function outputs the number per unique plotID to summarize not only how many unique plotIDs there are, but how many observations were at each.

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.

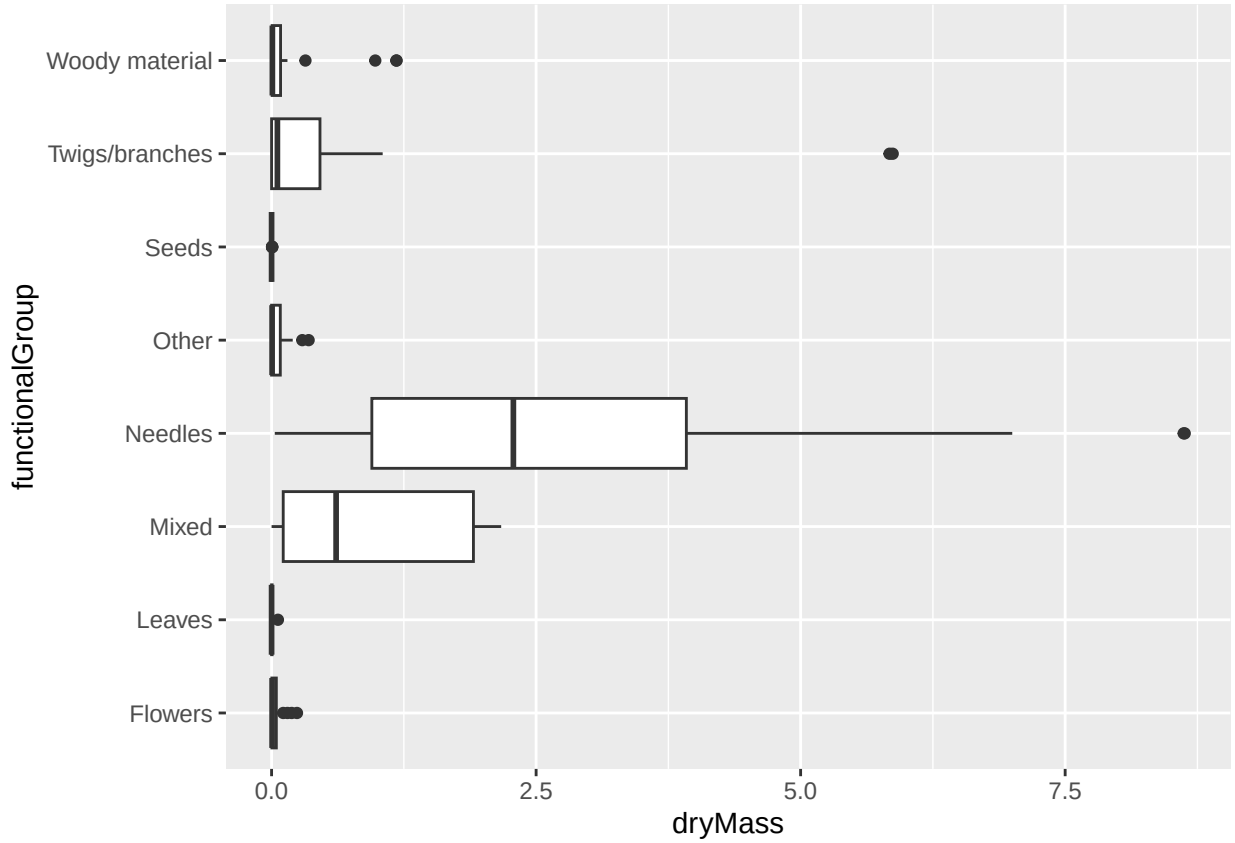
```
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar()
```



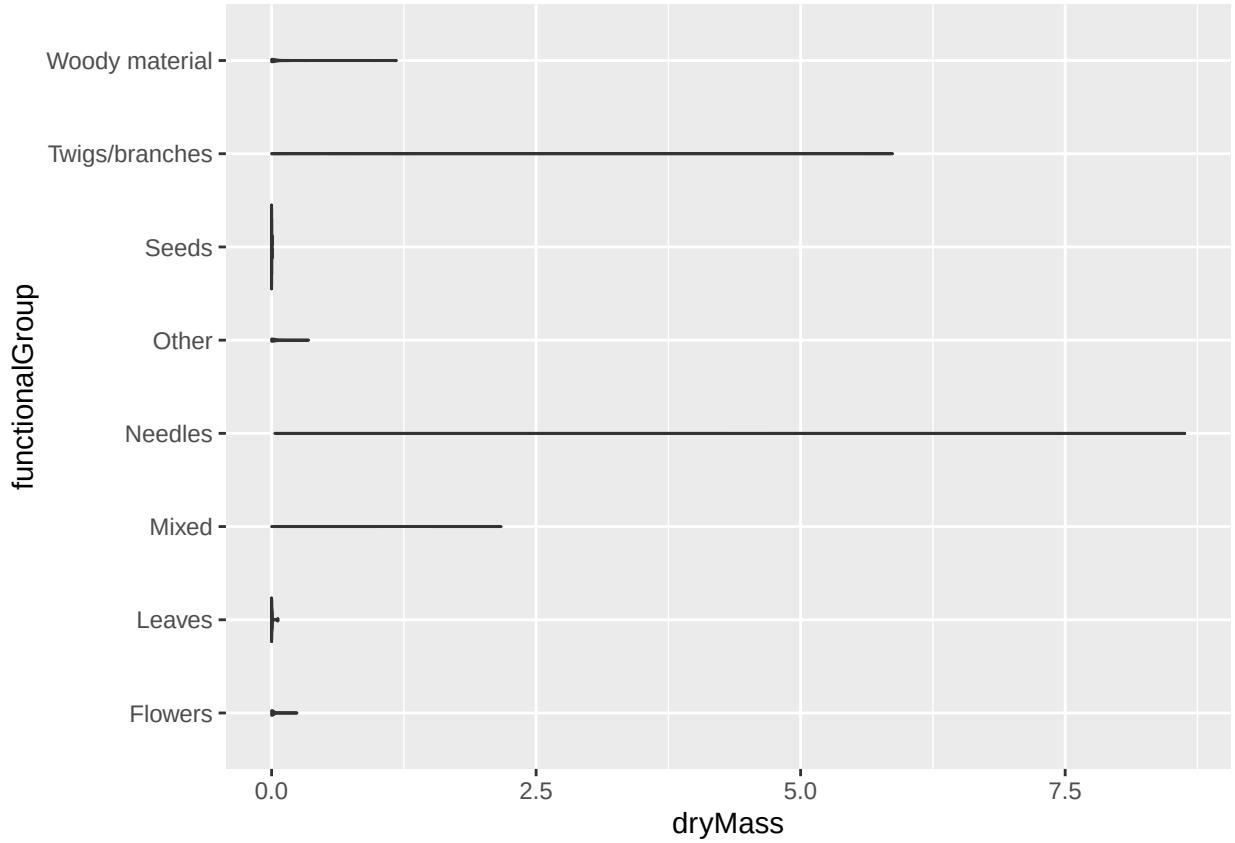

```
# This code chunk creates a bar chart of the functional group counts with ggplot.
```

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

```
# This code chunk is creating a boxplot of dryMass and functionalGroup
ggplot(Litter) +
  geom_boxplot(aes(x = dryMass, y = functionalGroup))
```




```
# This code chunk is creating a violin plot of dryMass and functionalGroup.
ggplot(Litter) +
  geom_violin(aes(x = dryMass, y = functionalGroup), draw_quantiles = c(0.25, 0.5, 0.75))
```



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

> Answer: A lot of the information in the violin plot is lost in the compression of the lines, and more information is communicated and comparable in the boxplot. The width of the violin plot is determined by frequency of the observation, and there were limited frequencies so the lines were primarily thin and lack much communication besides showing the comparable ranges of the functional groups. The boxplot communicates the median, IQR, range, and outliers for each group well.

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 the Mixed functional group.