Assignment 3: Data Exploration

Sam Campbell

Spring 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.

```
#check working directory
getwd()
```

```
## [1] "/home/guest/R/EDA-Spring2023"
```

```
#load tidyverse and lubridate packages
library(tidyverse)
library(lubridate)
#Rename datasets
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 the NRDC, Neonicotinoids permanently bind to insects' nerve cells, leading to overstimulation, paralysis, and death. If not targeted specifically to pests, I would guess these chemicals could harm ecologically important insects like pollinators.

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 provides important habitat structure, energy and nutrients for wildlife in 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 and woody debris are collected from elevated traps and ground traps and weighed to an accuracy of .01 grams. 2. Ground traps are sampled once per year, and elevated traps may be sampled once every 2 weeks in deciduous forest sites, or once every 1-2 months in an evergreen forest site. 3. Sampling occurs in tower plots, using 20 40m x 40m plots in forested sites and 4 40m x 40m plots plus 26 20m x 20m plots in sites with low-statured vegetation.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#Find the dimensions of Neonics dataset
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?

```
#Summary of Neonics dataset, Effect column
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
```

Answer: The most common effects being studied are Population, followed by Mortality and Behavior. These effects could tell us how the litter and woody debris are affecting other organisms in the area through changing those organisms' behavior, contributing to their death, etc.

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 of Neonics dataset, Species Common Name column
sort(summary(Neonics$Species.Common.Name))
```

```
##                         Ant Family                         Apple Maggot
##                                  9                                    9
##              Glasshouse Potato Wasp                            Lacewing
##                                 10                                   10
##             Southern House Mosquito              Two Spotted Lady Beetle
##                                 10                                   10
##           Spotless Ladybird Beetle                  Braconid Parasitoid
##                                 11                                   12
##                       Common Thrip         Eastern Subterranean Termite
##                                 12                                   12
##                             Jassid                           Mite Order
##                                 12                                   12
##                          Pea Aphid                    Pond Wolf Spider
##                                 12                                   12
##              Armoured Scale Family                    Diamondback Moth
##                                 13                                   13
##                      Eulophid Wasp                    Monarch Butterfly
##                                 13                                   13
##                      Predatory Bug                Yellow Fever Mosquito
##                                 13                                   13
##                       Corn Earworm                   Green Peach Aphid
##                                 14                                   14
##                          House Fly                            Ox Beetle
##                                 14                                   14
##                 Red Scale Parasite                  Spined Soldier Bug
##                                 14                                   14
##              Western Flower Thrips Hemlock Woolly Adelgid Lady Beetle
##                                 15                                   16
##              Hemlock Wooly Adelgid                                Mite
##                                 16                                   16
##                         Onion Thrip              Araneoid Spider Order
##                                 16                                   17
##                           Bee Order                     Egg Parasitoid
##                                 17                                   17
```

```
##                     Insect Class          Moth And Butterfly Order
##                               17                                17
##      Oystershell Scale Parasitoid         Black-spotted Lady Beetle
##                               17                                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
##                   Codling Moth         Flatheaded Appletree Borer
##                               19                                20
##              Horned Oak Gall Wasp                Leaf Beetle Family
##                               20                                20
##                Potato Leafhopper         Tooth-necked Fungus Beetle
##                               20                                20
##                   Argentine Ant                            Beetle
##                               21                                21
##                      Mason Bee                           Mosquito
##                               22                                22
##                Citrus Leafminer                  Ladybird Beetle
##                               23                                23
##                Spider/Mite Class              Tobacco Flea Beetle
##                               24                                24
##                    Chalcid Wasp            Convergent Lady Beetle
##                               25                                25
##                    Stingless Bee              Ground Beetle Family
##                               25                                27
##               Rove Beetle Family                     Tobacco Aphid
##                               27                                27
##                    Scarab Beetle                     Spring Tiphia
##                               29                                29
##                     Thrip Order            Ladybird Beetle Family
##                               29                                30
##                      Parasitoid                     Braconid Wasp
##                               30                                33
##                    Cotton Aphid                   Predatory Mite
##                               33                                33
##              Sweetpotato Whitefly                      Aphid Family
##                               37                                38
##                   Cabbage Looper             Buff-tailed Bumblebee
##                               38                                39
##                   True Bug Order           Sevenspotted Lady Beetle
##                               45                                46
##                     Beetle Order       Snout Beetle Family, Weevil
##                               47                                47
##               Erythrina Gall Wasp                Parasitoid Wasp
##                               49                                51
##             Colorado Potato Beetle                   Parastic Wasp
##                               57                                58
##               Asian Citrus Psyllid                Minute Pirate Bug
##                               60                                62
```

```
##                European Dark Bee                         Wireworm
##                               66                               69
##                   Euonymus Scale                Asian Lady Beetle
##                               75                               76
##                  Japanese Beetle                  Italian Honeybee
##                               94                              113
##                      Bumble Bee                Carniolan Honey Bee
##                              140                              152
##            Buff Tailed Bumblebee                     Parasitic Wasp
##                              183                              285
##                        Honey Bee                         (Other)
##                              667                              670
```

Answer: The six most commonly studied species in the dataset are Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. These species are all pollinators, so may be of interest because of their unique role in supporting plants' reproductive processes by carrying pollen between plants.

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?

```
#Searching for class of 'Conc.1..Author.' column in Neonics dataset
class(Neonics$Conc.1..Author.)
```

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

Answer: The class of the 'Conc.1..Author.' column is factor, because not all values in this column are numeric. Some values are characters, such as NR.

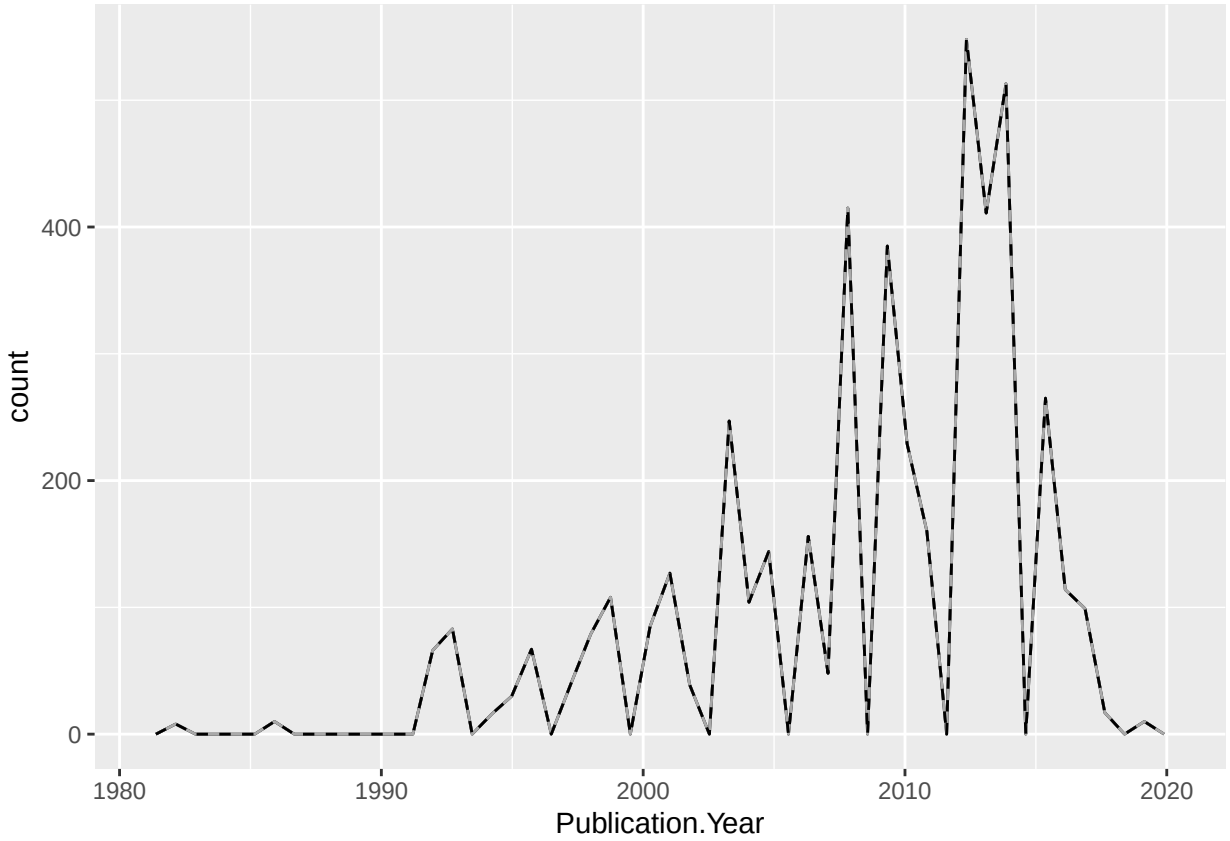
Explore your data graphically (Neonics)

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

```
#Initial try, creates error
geom_freqpoly(aes(Neonics$Publication.Year))
```

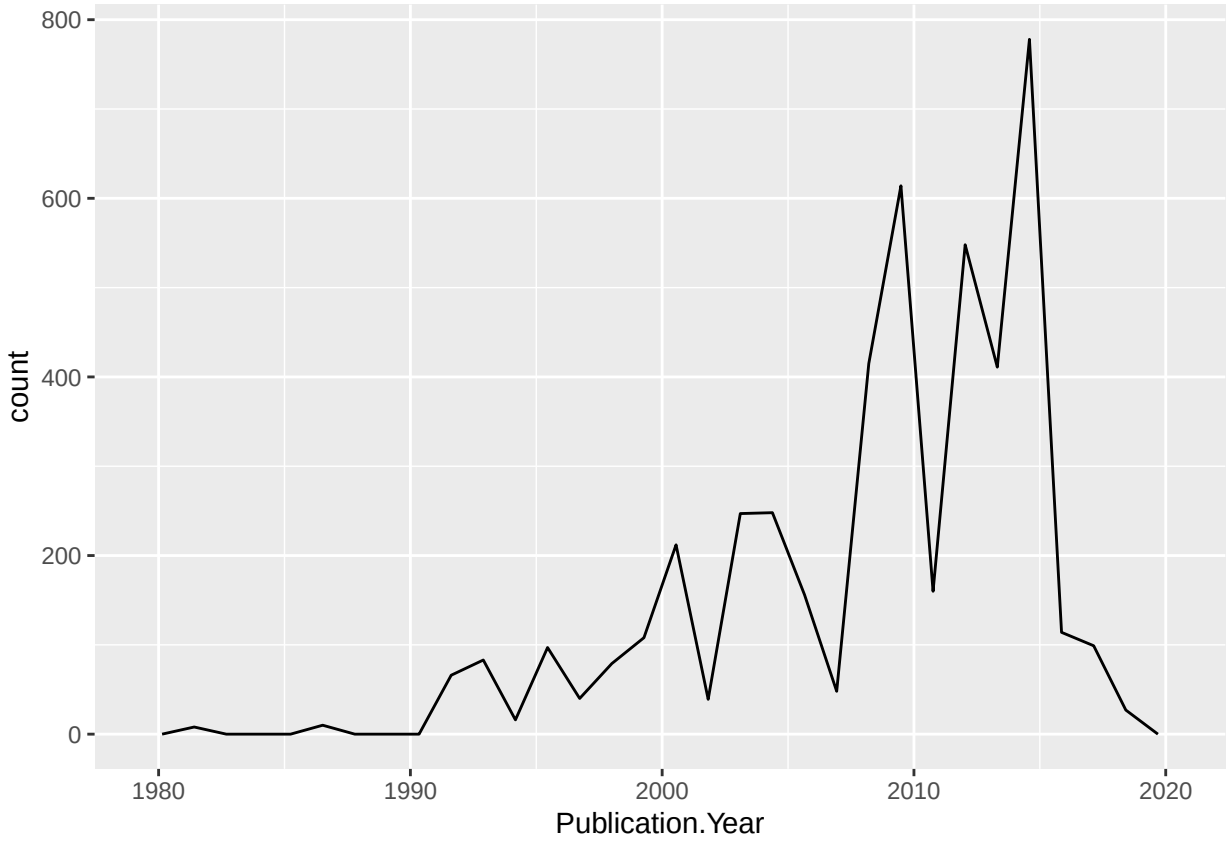
```
## mapping: x = ~Neonics$Publication.Year
## geom_path: na.rm = FALSE
## stat_bin: na.rm = FALSE, pad = TRUE
## position_identity
```

```
#Second try, using code from the lesson
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50, color = "darkgray") +
  geom_freqpoly(aes(x = Publication.Year), bins = 50,  lty = 2)
```



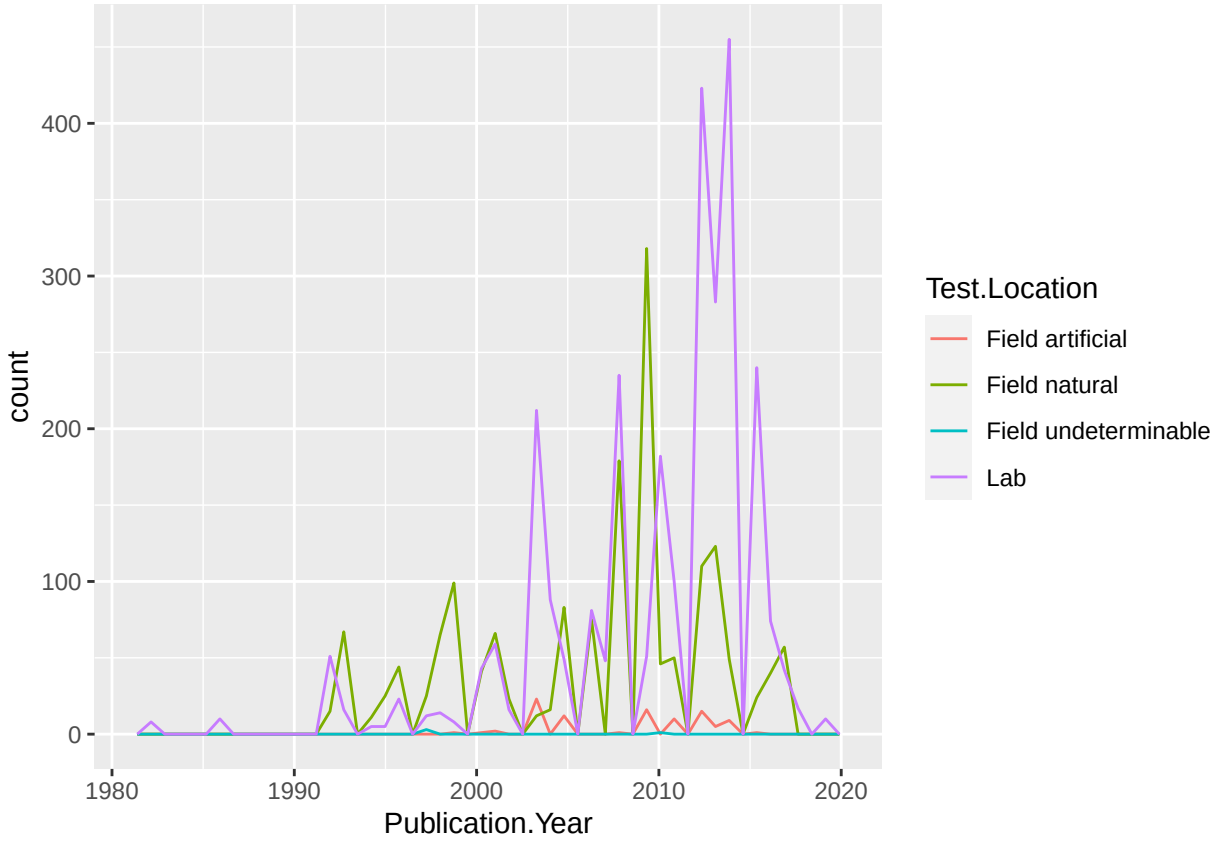

```
#3rd try, simplified version of code from lesson
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year))
```

```
## `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.

```
#Adding colors for the column Test.Location. Bins defaulted to 30, had to adjust to 50 to make visible
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 50)
```



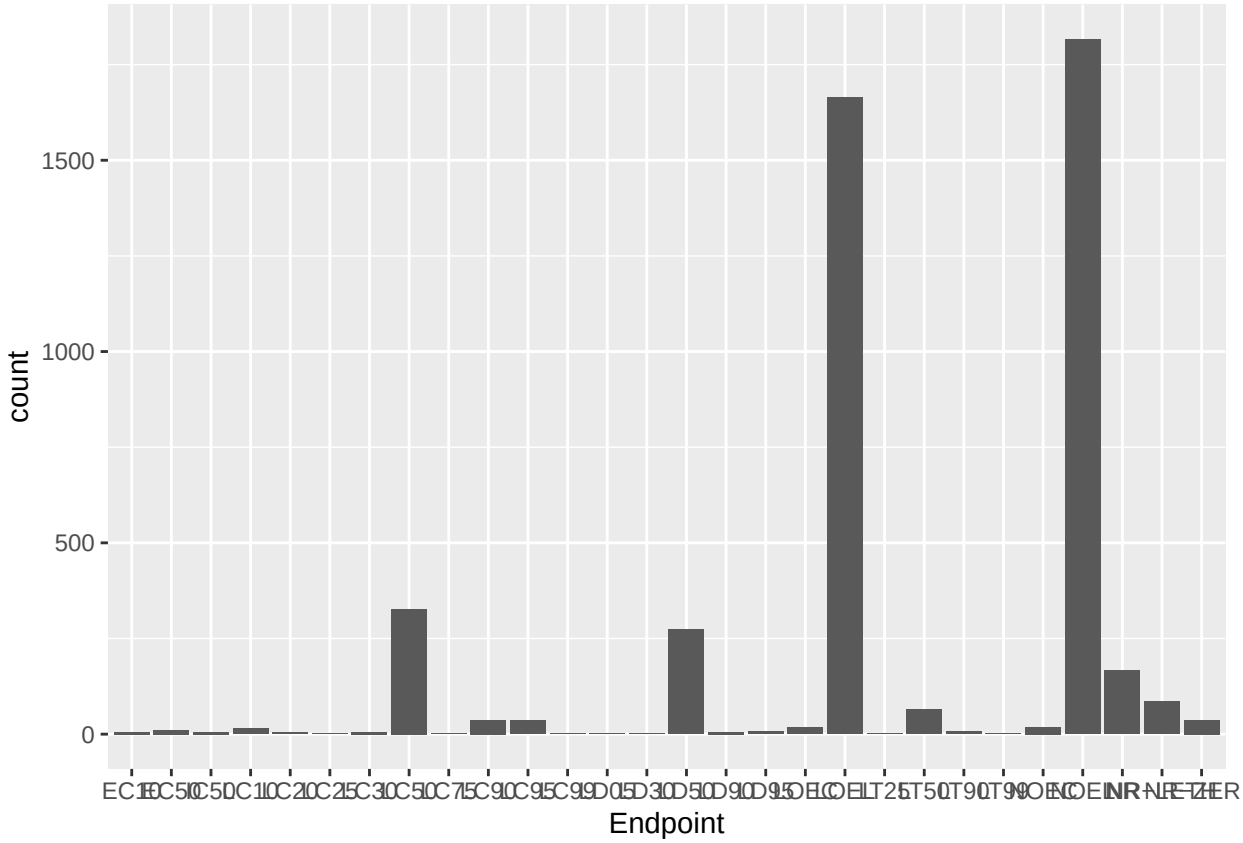
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are Lab and Field natural, with one of these exceeding the other depending on the year.

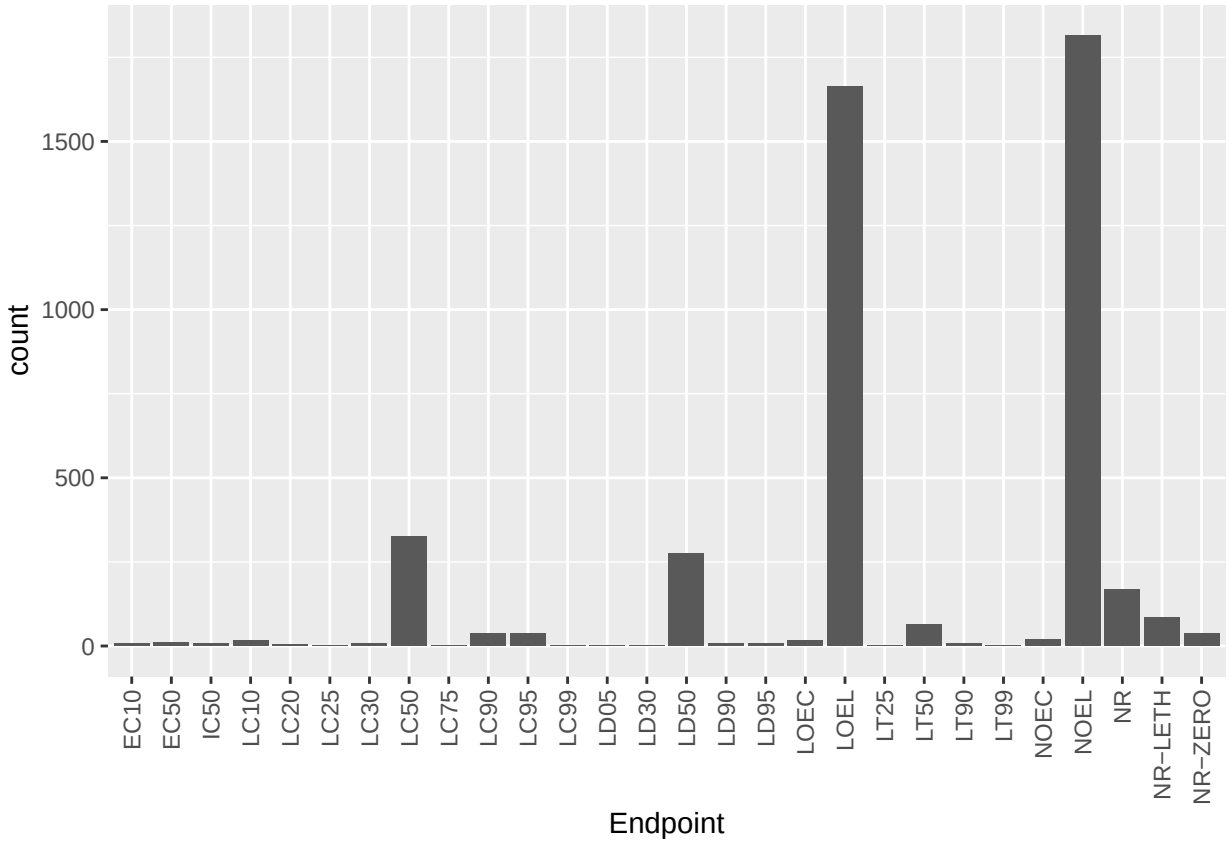
11. Create a bar graph of Endpoint counts. What are the two most common end points, and how are they defined? Consult the ECOTOX_CodeAppendix for more information.

[**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...]

```
#Fourth try, successful
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar()
```




```
#Fifth try, adding in TIP code with success. Use + then put TIP code on the next line after geom_bar()
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: The two most common endpoints are NOEL and LOEL. NOEL is defined as No-observable-effect-level, highest dose producing effects not significantly different from responses of controls. LOEL is defined as Lowest-observable-effect-level, lowest dose producing effects that were significantly different from responses of controls.

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"
```

```
#'collectDate' is not a date class, it is a factor
#Inputting code to change collectDate class to date
Litter$collectDate <- as.Date(Litter$collectDate)
#Confirming new class
class(Litter$collectDate)
```

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

```
#Determining which dates litter was sampled in August 2018
unique(Litter$collectDate)
```

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

```
#Litter sampled on dates 2 and 30 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`?

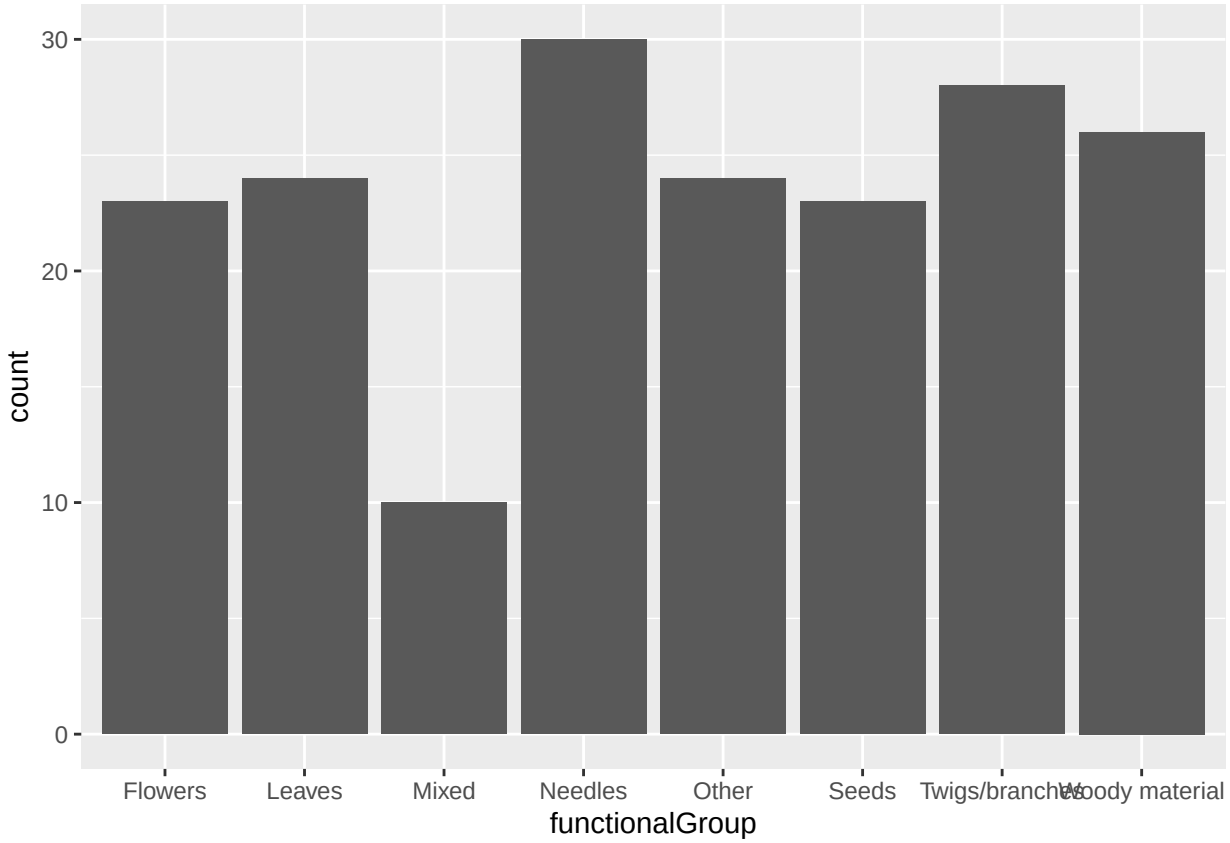
```
#Using unique function to check which plots were sampled at Niwot Ridge
unique(Litter$plotID)
```

```
##  [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
```

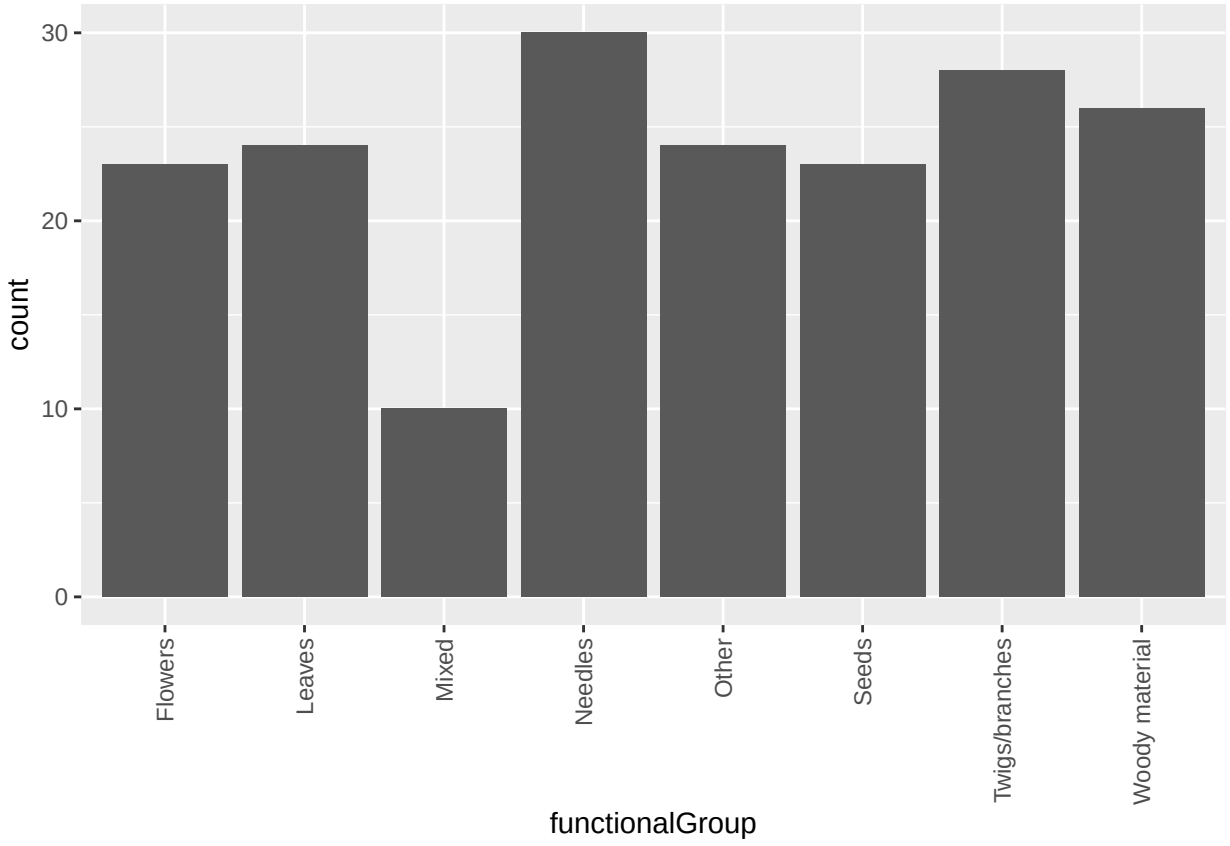
Answer: 12 plots were sampled at Niwot Ridge. Unique is different from Summary in that it doesn't check for statistics like mean and median, which are numerical values. Instead, it checks for which dates/characters/etc. show up throughout the data set.

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.

```
#Creating a bar graph of functionalGroup
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar()
```

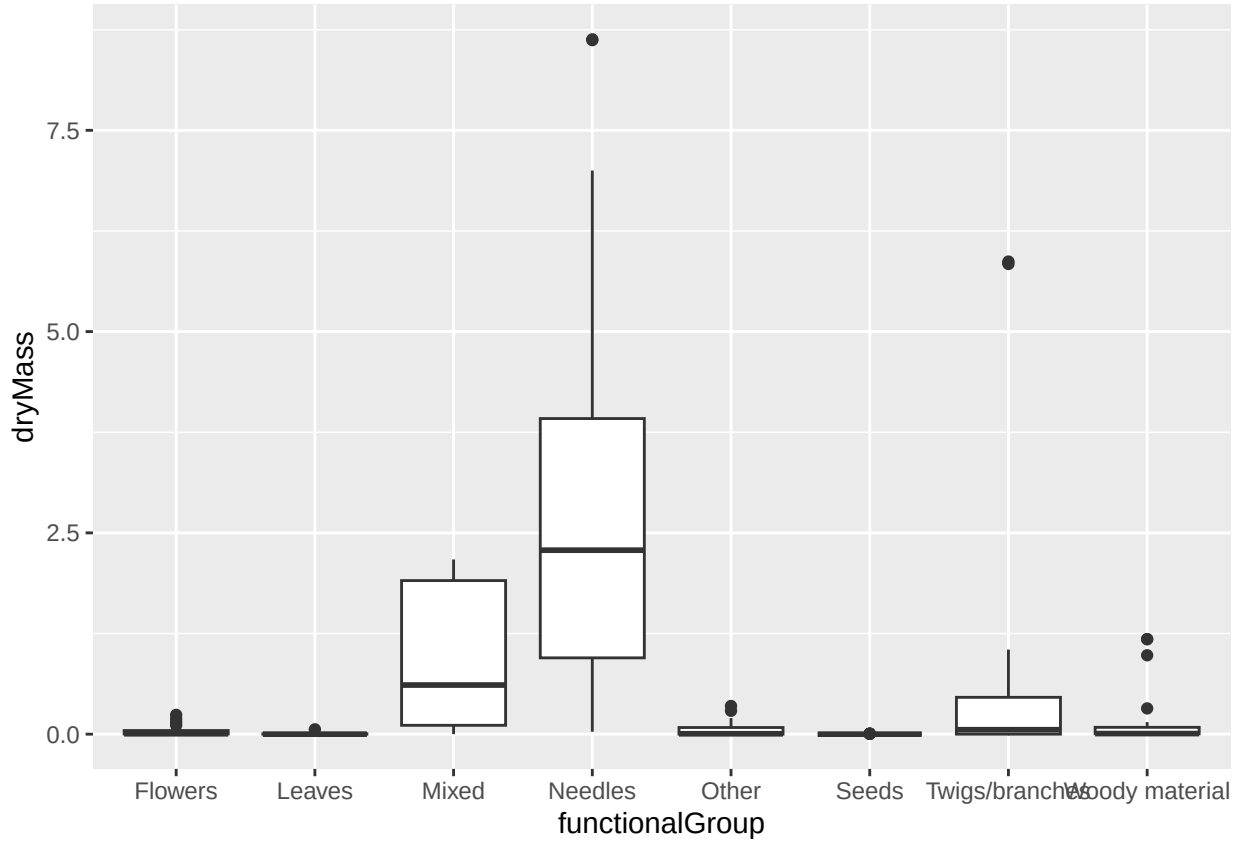



```
#Adding code so that names don't overlap in the bar graph x axis
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

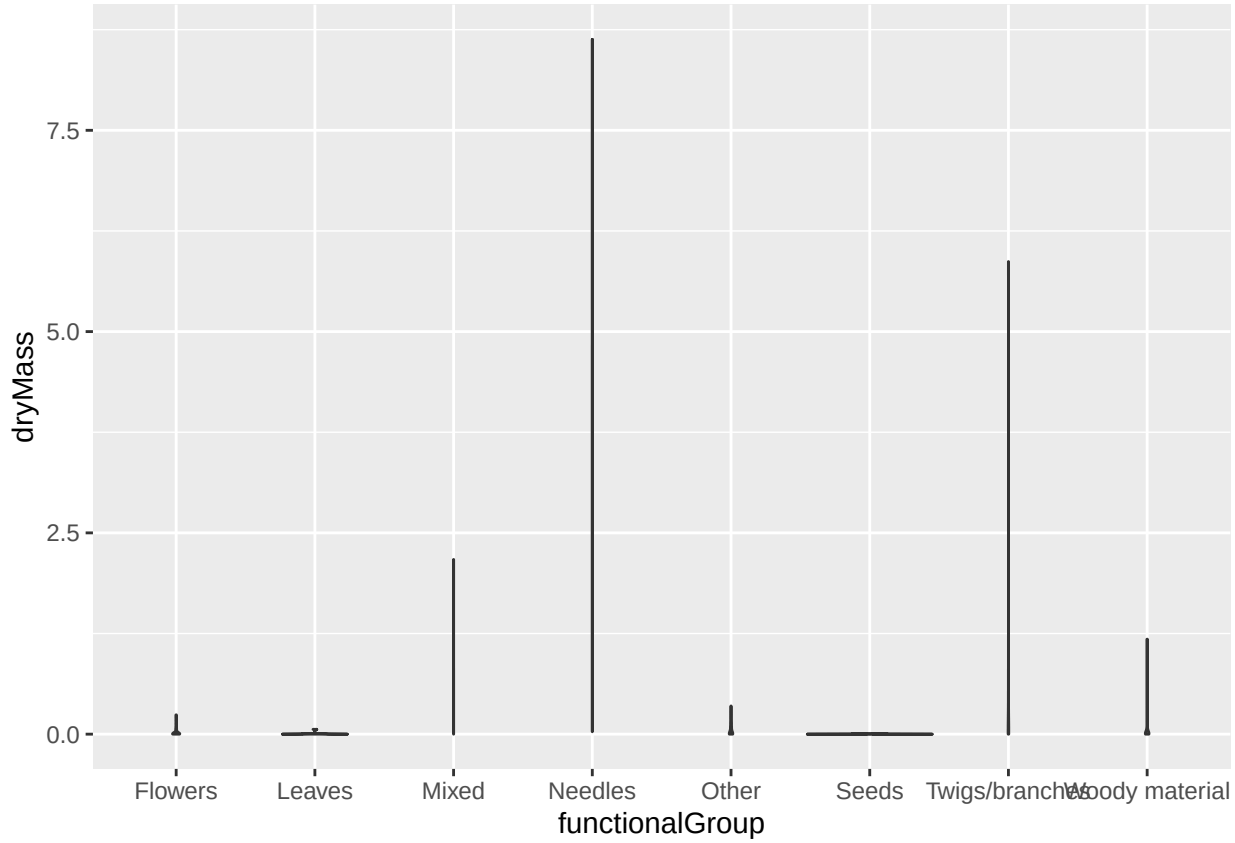


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#Using geom_boxplot with functionalGroup on x axis and dryMass on y axis
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass))
```




```
#Using geom_violin with functionalGroup on x axis and dryMass on y axis
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass))
```



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

> Answer: In this case the boxplot is a more effective visualization option than the violin plot because the width of each box better enables us to see which functionalGroups had greatest dry mass, whereas the violin plot for this data only shows lines with no width.

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

> Answer: Needles and Mixed litter tend to have the highest biomass at these sites.