Assignment 3: Data Exploration

Chunyi Xu

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.

```
#Noting down codes to install packages
#install.packages("tidyverse");
#install.packages("lubridate");
#install.packages("here")

#Loading necessary packages (tidyverse, lubridate, here)
library (tidyverse)
library (lubridate)
library (here)
```

```
## Warning: package 'here' was built under R version 4.3.3
```

```
#Checking working directory is the project folder
here ()
```

[1] "/Users/xuchunyi/Desktop/EDA_Spring2025/EDA_Spring2025"

```
#Uploading the ECOTOX neonicotinoid dataset
Neonics <- read.csv(
  file = here("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),
  stringsAsFactors = TRUE)

#Uploading the Niwot Ridge NEON dataset for litter and woody debris
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:The ecotoxicology of neonicotinoids on insects helps us better understand their ecological impacts. Although neonicotinoid insecticides were introduced because their relatively lower toxicity than other older insecticides, recent research has indicated that their environmental impact may also be devastating. Neonicotinoid can have harmful consequences for benefical insects, pollinator, and aquatic invertebrates. Source: https://www.xerces.org/pesticides/understanding-neonicotinoids#:~:text=While%20they%20were%20initially%20introduced,beneficial%20insects%2C%20and%20aquati

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 are essential components of forest and stream ecosystems. They are involved in the nutrient cycling and carbon budgets process, serving as a source of energy for aquatic ecosystems and providing habitat for aquatic and terrestrial organisms. They also influence the ecosystem's structure and roughness and could potentially change sediment transport and water flows. Source: https://research.fs.usda.gov/treesearch/20001

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 is sampled at terrestrial NEON sites that contain woody vegetation >2m tall. 2.The sampling plot's centers should be 50m from buildings and other non-NEON infrastructure. 3. The placement of litter traps within the plots may be either targeted or randomized, depending on the vegetation. For example, in sites with more than 50% aerial cover of woody vegetation >2m in height, the trap placement is randomized. In sites with less than 50% cover of woody vegetation and distributed vegetation, the placement is targeted.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#Checking the dimensions of the dataset Neonics. It has 4623 rows and 30 columns.
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? [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```
# Creating a new dataframe "Effect_Summary" for the summary of the "Effect" column
Effect_Summary <- summary (Neonics$Effect)
# Viewing the new dataframe "Effect_Summary"
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
```

```
# Sorting the new dataframe "Effect_Summary" in order of magnitude
sort (Effect_Summary)
```

```
##      Hormone(s)        Histology       Physiology          Cell(s)
##               1                5                7                9
##    Biochemistry     Accumulation     Intoxication    Immunological
##              11               12               12               16
##      Morphology           Growth        Enzyme(s)         Genetics
##              22               38               62               82
##       Avoidance      Development     Reproduction Feeding behavior
##             102              136              197              255
##        Behavior        Mortality       Population
##             360             1493             1803
```

Answer:The top six effects that are studied are Population, Mortality, Behavior, Feeding behavior, Reproduction, and Development. These effects are studied because they have a more significant impact on the biodiversity of an ecosystem than other impacts. For example, population is essential to maintaining habitat diversity, sustaining the balance of the food chain, and controlling the carbon cycle. However, studies have found that the use of neonicotinoids leads to increased population extinction rates. Therefore, mortality is worth investigating, too, as it is closely linked to the population of a species. Animal behavior, including feeding behavior, can potentially change ecosystem dynamics, so this topic becomes

specifically of interest, too. Animal behavior is critical to delivering ecosystem services, which determine the material benefits people receive, such as food and fuel, and the regulating benefits, such as climate regulation and flood control. Source: https://www.noble.org/regenerative-agriculture/wildlife/top-5-considerations-for-increasing-wildlife-diversity/ https://environment.yale.edu/news/article/protecting-wildlife-populations-can-enhance-natural-capture-capture https://www.sciencedirect.com/science/article/abs/pii/S0169534722002762 https://aithor.com/essay-examples/an-analysis-of-the-impact-of-feeding-habits-on-animal-behavior-and-ecosystems#3-the-ecological-consequences-of-feeding-habits

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...]

```
# Creating a new dataframe "Effect_Species.Common.Name" for the summary of the "Effect" column
Effect_Species.Common.Name <- summary (Neonics$Species.Common.Name)
# Viewing the new dataframe "Effect_Species.Common.Name"
print (Effect_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
```

```r
# Sorting the new dataframe "Effect_Species.Common.Name" in order of occurrence
sort (Effect_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
```

```
##               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 bees. Bees are the primary pollinators in the ecosystem. They play important roles in maintaining and improving biodiversity by transporting flower pollen and fertilizing plants and have an impact on other insects, animals, and plants. Therefore, they are vital in sustaining the global food supply and a healthy ecosystem. However, a significant amount of studies indicate that neonics usage is a leading cause of massive bee die-offs around the world. Source: https://greenly.earth/en-us/blog/ecology-news/why-are-bees-so-important-for-people-and-the-environment https://www.nrdc.org/stories/neonicotinoids-101-effects-humans-and-bees

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. . .]

```r
#Figuring out the class of "Conc.1..Author", and the class is "factor"
class (Neonics$Conc.1..Author)
```

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

```r
#Viewing the dataframe
#View (Neonics)
#View(Neonics$Conc.1..Author)
```

Answer: The class of 'Conc.1..Author' is factor. By viewing the entire dataframe Neonics, the entries in the column "Conc.1..Author" includes both values and symbols "/," becoming string variables. When we uploaded the dataset, we read strings variables in as factors. Therefore, the column "Conc.1..Author" is classified as factor. Meanwhile, by viewing the column "Conc.1..Author" separately, the result also shows that the value for "Conc.1..Author" are non-numeric but are factors with 1006 levels <0.0004 <0.025 <0.088 <0.5 <1.5 <10/ <2.5/ <4.00 <5.00 <8.79 . . . NR/.

Explore your data graphically (Neonics)

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

```r
#Generating a plot for the number of studies conducted by publication year by using the dataset Neonics
ggplot(Neonics) +
geom_freqpoly(aes(x = Publication.Year), bins = 50)
```
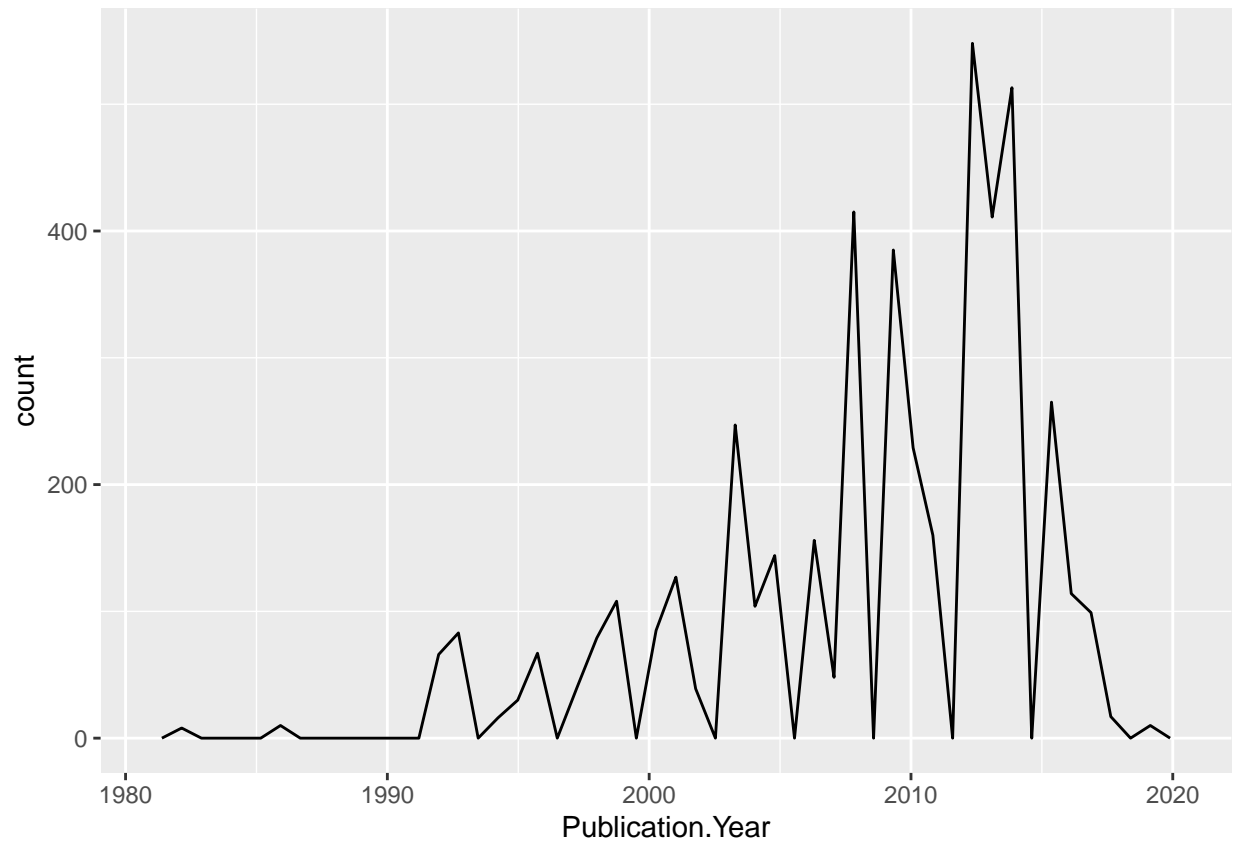
10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
# Reproducing the same graph but adding a color aesthetic according to the variable Test.Location
ggplot(Neonics) +
geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 50) +
theme(legend.position = "left")
```
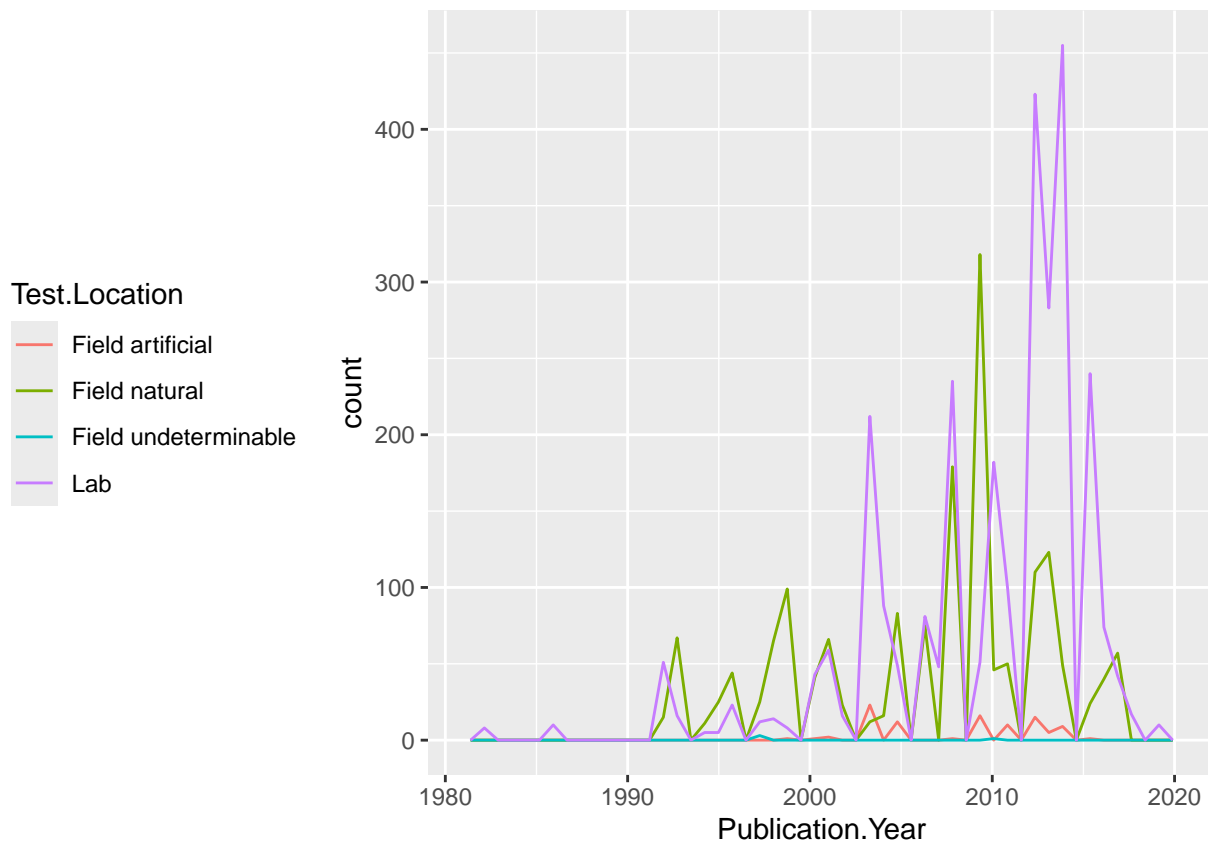
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer:The most common test location is the Lab. The second common test location is Field
> Natural. They differ over time. For example, from around 1993 to 2000, and from around 2007
> to 2010, Field Nature was a more common test location than Lab. However, from around 2003
> to 2005, and from around 2012 to 2020, Lab has become a more common test location than Field
> Natural. From 2000 to around 2003, both locations were equally common.

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...]

```
# Creating a bar graph of Endpoint counts and adjusting the X-axis labels
ggplot(Neonics)+
geom_bar(aes(x = Endpoint))+
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```
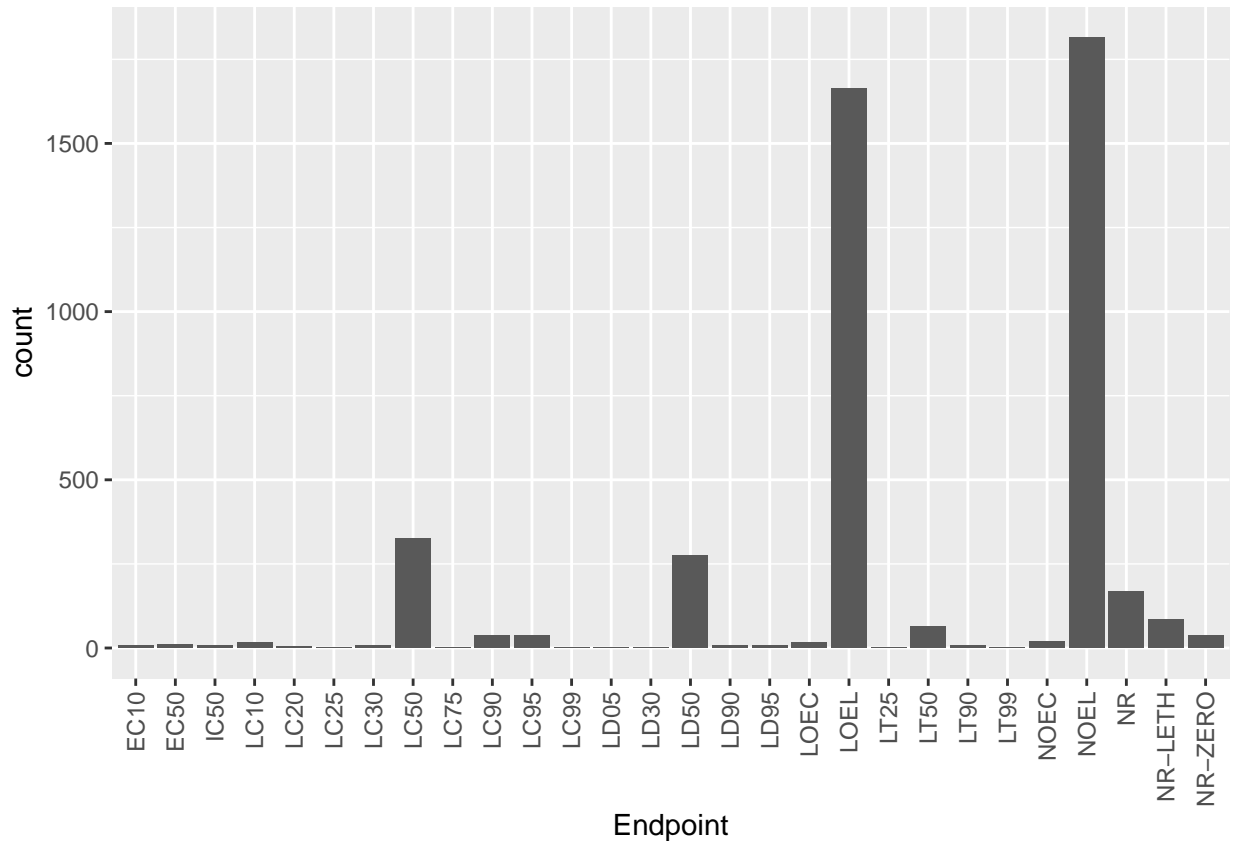
Answer: The two most common end points are NOEL and LOEL. LOEL is for the terrestrial database usage. It means lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC). NOEL is for terrestrial database usage too. It means no-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test (NOEAL/NOEC).

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.

```r
# Determining the class of collectDate, and the result shows that it is not a date. It is "factor" inst
class (Litter$collectDate)
```

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

```r
#Changing collectDate to a date
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")

#Confirming the new class of collectDate is date
class (Litter$collectDate)
```

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

```
# Determining which dates litter was sampled in August 2018, the result shows that 2018-08-02 and 2018-
unique (Litter$collectDate)
```

```
## [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`?

```
#Using the unique function to determine how many different plots were sampled at Niwot Ridge (NIWO), an
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
```

```
# Comparing the result given by the unique function and the summary function
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 are 12 different plots were sampled at the Niwot Ridge, since there 12 levels in the plotID column. While the information obtained from 'unique' only shows the unique values that one factor can take, the information obtained from 'summary' gives the detailed counts of each unique value.

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 counts and adjusting the X-axis labels
ggplot(Litter)+
geom_bar(aes(x = functionalGroup))+
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```
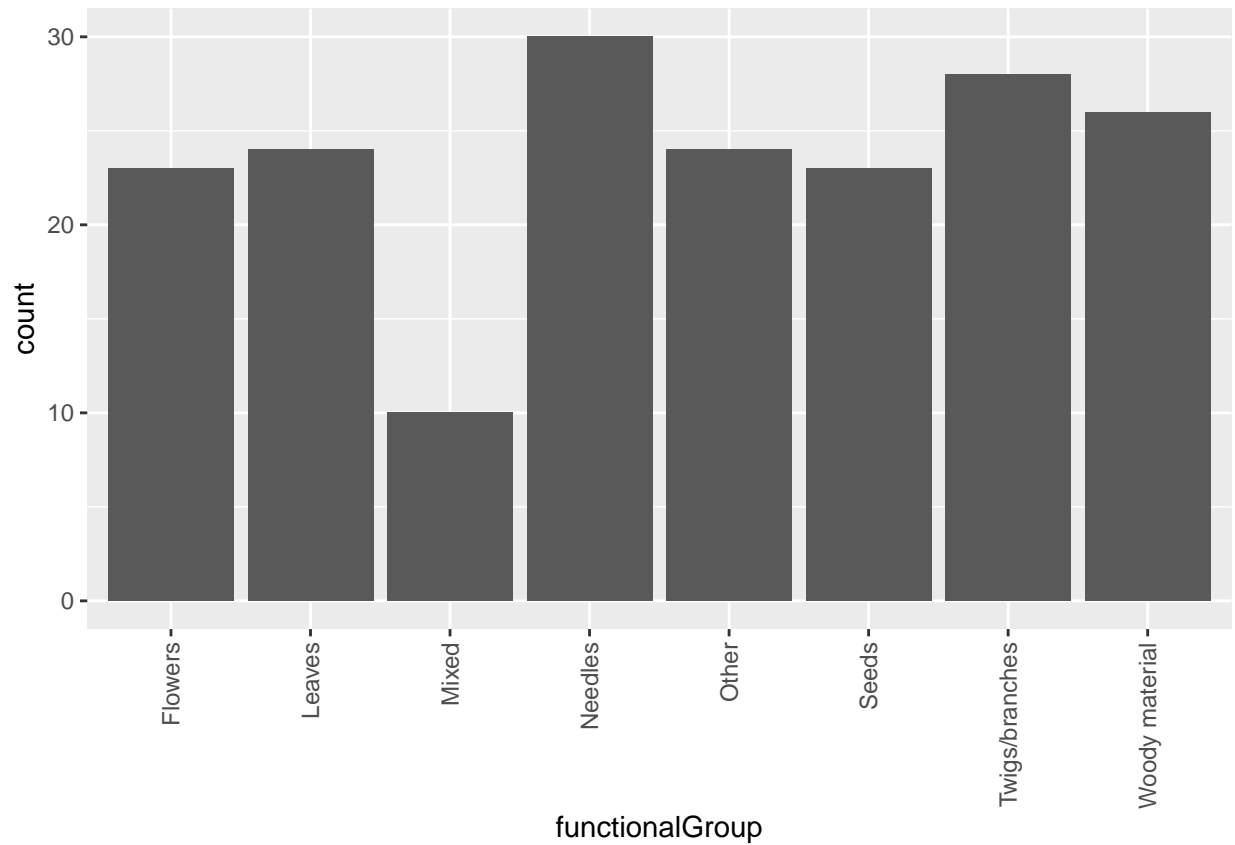
15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functional-Group.

```
#Creating a boxplot of dryMass by functionalGroup and adjusting the X-axis labels
ggplot(Litter) +
geom_boxplot(aes(x = functionalGroup, y = dryMass))+
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```
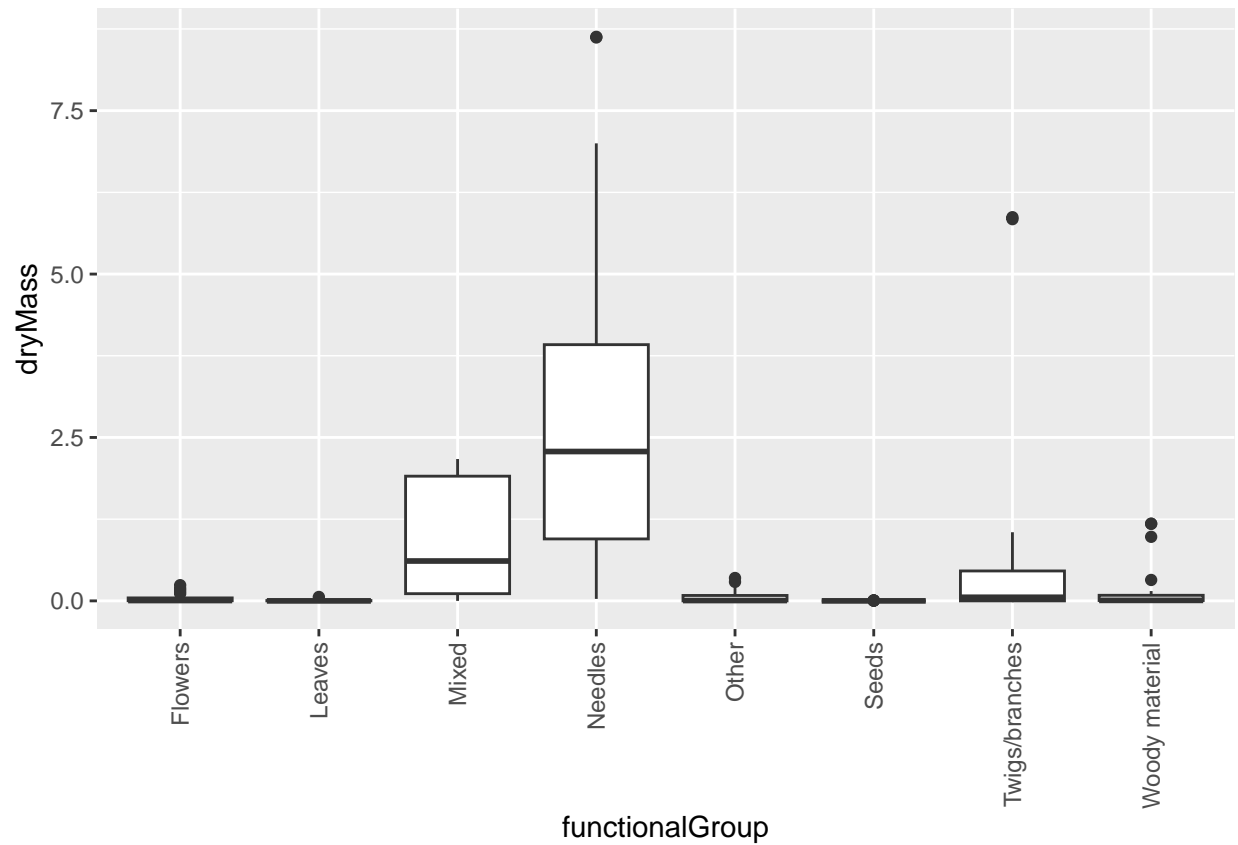
```
#Creating a violin plot of dryMass by functionalGroup and adjusting the X-axis labels
ggplot(Litter) +
geom_point(aes(x = functionalGroup, y = dryMass))+
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```
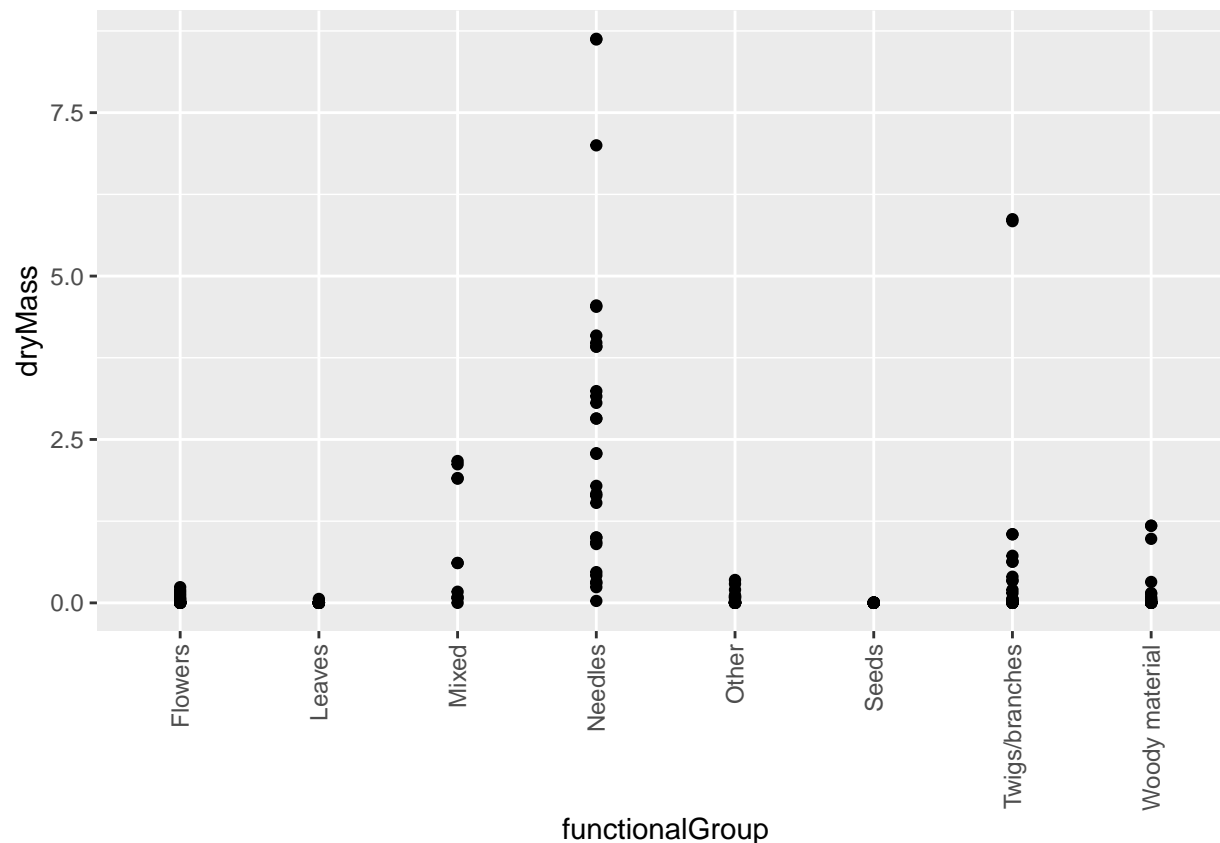
Why is the boxplot a more effective visualization option than the violin plot in this case?

Answer: The width of the violin plot is proportional to the number of entries in a column, making the plot look like a frequency poly at the same time. However, in this violin plot, the number of entries for the dryMass by functionalGroup is not enough to generate a violin plot with the evident distribution. Therefore, the violin plot is not informative and cannot showcase the median and the interquartile range. However, the boxplot can still showcase all helpful information, such as the median and the interquartile range, regardless of the distribution.

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

Answer: The Needles and Mixed types of litter tend to have the highest biomass at these sites. This is because they have higher median and higher interquartile range.