Assignment 3: Data Exploration

Gaby Czarniak

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() # checking my working directory
```

```
## [1] "/home/guest/gaby-cz_EDE_Fall2023"
```

```
print(getwd()) # printing (getwd()) so that TAs can see output
```

```
## [1] "/home/guest/gaby-cz_EDE_Fall2023"
```

```
# tinyverse and lubridate were already installed
# commenting out the install commands given that I need to knit and
# including library command to load the packages

# install.packages("tidyverse")
# install.packages("lubridate")

# library(tidyverse)
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 error
```

```
# library(lubridate)
library(lubridate)
```

```
# Uploading dataset on ecotoxicology
# Assigning name
# Reading strings in as factors
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",stringsAsFactors = TRUE)
# Neonics

# Uploading and renaming dataset on litter and woody debris
# Assigning name
# Reading strings in as factors
Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",stringsAsFactors = TRUE)
# 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: I did a brief internet search to learn more about neonicotinoids, and information at https://www.pnas.org/doi/10.1073/pnas.2017221117 described neonicotinoids as an exceptionally toxic class of insecticides, containing compounds that target specific receptors is insects' bodies. While intended to protect crops, these toxic substances are harming non-target species (including beneficial pollinators and arthropods that contribute to healthy soils) and traveling through food webs. They are threatening food webs and biodiversity.

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 can affect carbon and nutrient cycling in forests. It can also affect soil moisture content, impacting the water available to forest plants, and create/maintain habitats for other organisms.

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, "defined as material that is dropped from the forest canopy and has a butt end diameter <2cm and a length <50 cm," is "collected in elevated 0.5m2 PVC traps. 2. Woody debris, which they define as"material that is dropped from the forest canopy and has a butt end diameter <2cm and a length >50 cm," is collected in ground traps since it's longer and can't be collected in elevated traps. 3. Spatial sampling design indexes on the size of plots, whereas temporal sampling design indexes on ground traps that are sampled once per year.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

```
print(dim(Neonics))
```

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

```
# The Neonics dataset has 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?

```
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
```

```
print(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: Population is the most common of the effects studied (from "Effect" column), followed by mortality, behavior, feeding behavior, reproduction, and development–in that order. In the associated metadata file "ECOTOX_CodeAppendix.pdf", "Population" is defined as "measurements and endpoints relating to a group of organisms or plants of the same species occupying the same area at a given time." This effect is measured as abundance, so it makes sense that it would be commonly studied, as someone looking for patterns would be looking for data at a group level for a given species in a given time and place, and to draw conclusions at a population level, one would need many data entries. The next most common effect is mortality, which is defined as "measurements and endpoints where the cause of death is by direct action of the chemical." It makes sense to me that this would be commonly studied, because on the pyramid of severity of effect, with least severe at the bottom and increasing severity toward the top, mortality tops the pyramid. I can imagine someone interested in limiting severe effects of neonicotinoids would be looking to determine protection requirements near the most severe cases, assuming that those protections would be conservative in contributing to the protection of less severe cases, too. "Behavior" is defined as "overt activity of an organism represented by three effect groups - avoidance, general behavior, and feeding behavior. All measurements related to reproductive behavior are listed under the major effect group REP." Because behavior is an aggregate of three effects, it makes sense that it would appear commonly studied, especially that feeding behavior, itself, comes in as next-most-common. Reproduction is key to species survival, and development of an organism (defined in the metadata file as covering "toxicant effects on tissue organization in growing progeny") could affect reproductive capacity.

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. . .]

```
# Calling summary function to get first six most common species
# by common name
summary(Neonics$Species.Common.Name, 6)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee               (Other)
##                   152                   140                  3196
```

```
print(summary(Neonics$Species.Common.Name, 6))
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee               (Other)
##                   152                   140                  3196
```

```
# Because the sixth-most-common was "Other",
# I also wanted to call the summary function using an argument
# of 7, so that I could get a full six most common species
# rather than five, plus "Other"
summary(Neonics$Species.Common.Name, 7)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee
##                   152                   140                   113
##               (Other)
##                  3083
```

```
print(summary(Neonics$Species.Common.Name, 7))
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee
##                   152                   140                   113
##               (Other)
##                  3083
```

```
# sorting across all Species.Common.Name
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
```

```
# sorting across just the first seven Species.Common.Name
sort(summary(Neonics$Species.Common.Name, 7))
```

```
##      Italian Honeybee            Bumble Bee   Carniolan Honey Bee
##                   113                   140                   152
## Buff Tailed Bumblebee        Parasitic Wasp             Honey Bee
##                   183                   285                   667
##               (Other)
##                  3083
```

```
print(sort(summary(Neonics$Species.Common.Name, 7)))
```

```
##      Italian Honeybee            Bumble Bee   Carniolan Honey Bee
##                   113                   140                   152
## Buff Tailed Bumblebee        Parasitic Wasp             Honey Bee
##                   183                   285                   667
##               (Other)
##                  3083
```

Answer: The six most commonly studied species in the dataset, by common name, are Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. Although "Parasitic Wasp" was more than 58 entries less than "Honey Bee", I did notice that there seemed to be a typo in the Species.Common.Name entry of "Parastic Wasp", which could mean even more entries should be listed under "Parasitic Wasp" if it was, indeed, a typo. It seems like most of these species are important pollinators, which are key to our food chain. Parastic wasps also seem to have agricultural importance, due to their use as biological control agents against pests, according to https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4516919/.

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?

```
# Asking r to tell me the class of the column's data
class(Neonics$Conc.1..Author.)
```

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

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

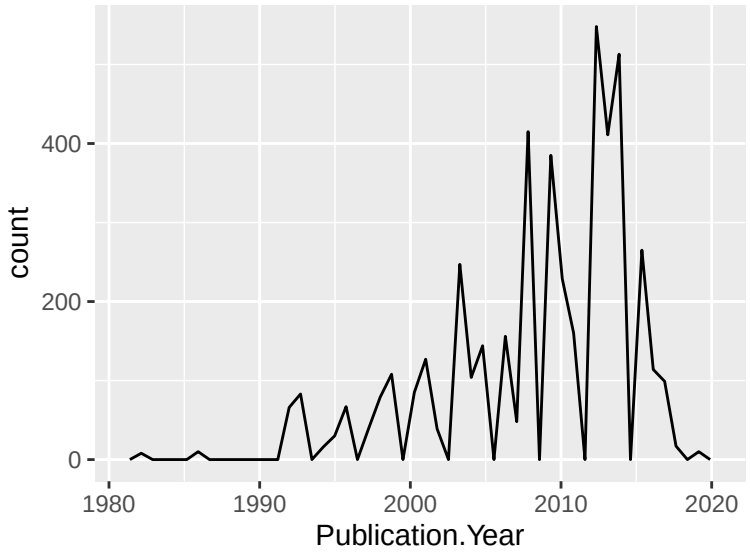
```
## [1] "factor"
```

Answer: The Conc.1..Author. column in the dataset is a factor. It's not numeric because I asked r to read the string in as factors. Storing the data as factors rather than numeric variables allows us to do more and different things with the data.

Explore your data graphically (Neonics)

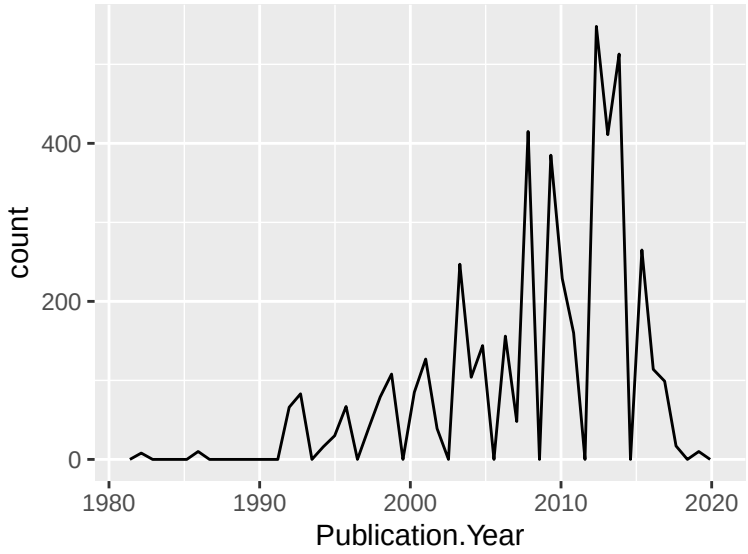
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
# Plotting a frequency line graph of the number of studies (y axis)
# by Publication Year (x axis)
#
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50)
```

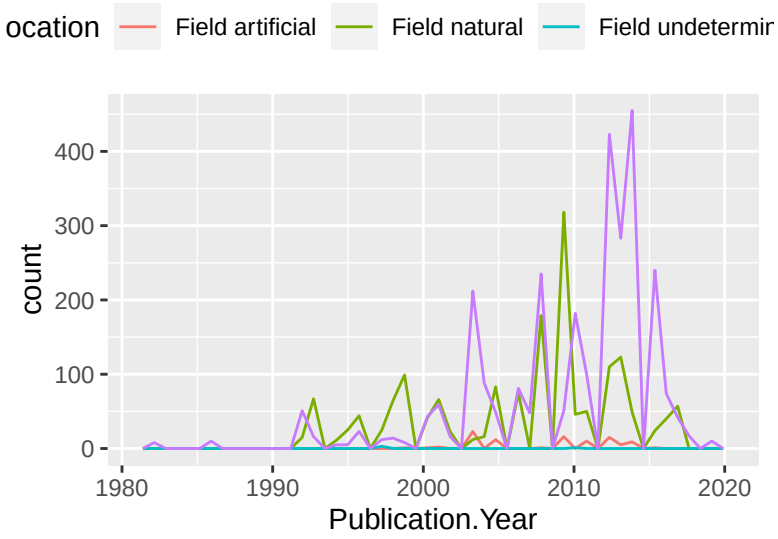


10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
# Plotting a frequency line graph of the number of studies (y axis)
# by Publication Year (x axis)
#
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50)
```




```
# Color-coding the count of number of studies by their Test Location
#
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 50) +
  theme(legend.position = "top")
```




```
# Confirming most common test locations
sort(summary(Neonics$Test.Location))
```

```
## Field undeterminable     Field artificial        Field natural
##                    4                   96                 1663
##                  Lab
##                 2860
```

```
print(sort(summary(Neonics$Test.Location)))
```

```
## Field undeterminable     Field artificial        Field natural
##                    4                   96                 1663
##                  Lab
##                 2860
```

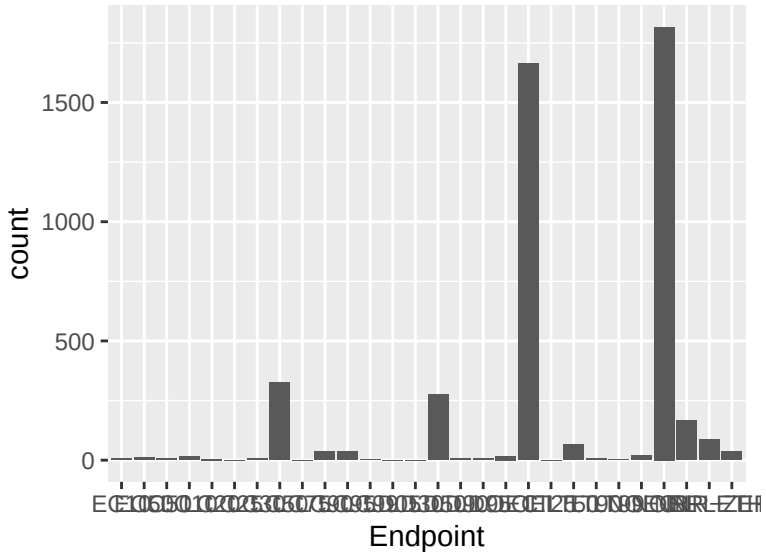
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: Overall, the most common test locations are lab locations, followed by field natural, then field artifical, then field undeterminable. They do differ over time–lab locations have definitely been more common in the 2010s, but there have been moments in the mid-to-late 1990s where field natural were the most common test locations, as well as in 2009, when they were more common than lab locations.

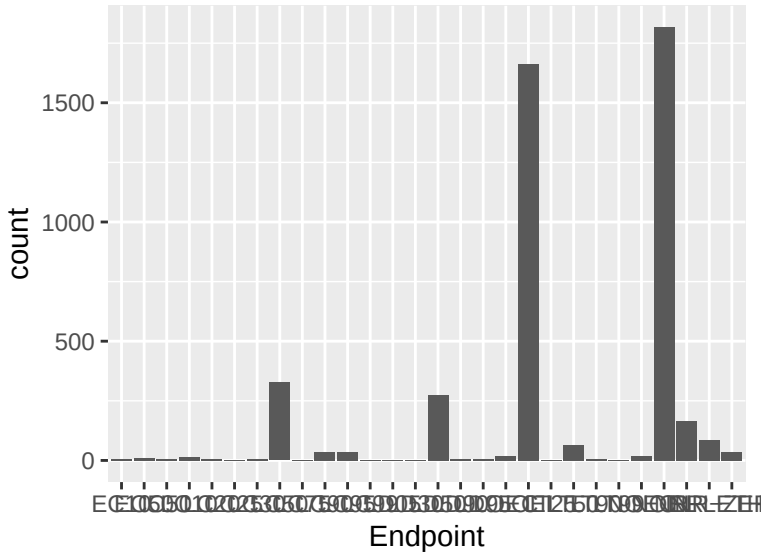
11. Create a bar graph of Endpoint counts. What are the two most common end points, and how are they defined? Consult the ECOTOX_CodeAppendix for more information.

[**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...]

```
# Plotting Endpoint counts as bar graph
ggplot(Neonics, aes(x = Endpoint)) +
  #theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
  #may be an error with theme because when I run it, none of my data is visible
  geom_bar()
```



```
print(ggplot(Neonics, aes(x = Endpoint)) +
  #theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
  #may be an error with theme because when I run it, none of my data is visible
  geom_bar())
```



Answer: NOEL and LOEL are the two most common end points. They are defined in ECOTOX_CodeAppendix as: for LOEL, "Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC)"), and for NOEL, "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)" – both for Terrestrial database usage.

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.

```
# Checking class of collectDate in Litter
# class tells me that it is a factor
class(Litter$collectDate)
```

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

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

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

```
# I double check that it's not a date
is.Date(Litter$collectDate)
```

```
## [1] FALSE
```

```r
print(is.Date(Litter$collectDate))
```

```
## [1] FALSE
```

```r
#date_obj_collectDate <- (Litter$collectDate)
#format(date_obj_collectDate, format = "%y - %m - %d")
#date_obj_collectDate <- ymd(Litter$collectDate)
date_obj_collectDate <- as.Date(Litter$collectDate, format = "%y-%m-%d")
date_obj_collectDate
```

```
##   [1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [26] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [51] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
##  [76] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## [101] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## [126] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## [151] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## [176] NA NA NA NA NA NA NA NA NA NA NA NA NA
```

```r
is.Date(Litter$collectDate)
```

```
## [1] FALSE
```

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`?

```r
# Removing duplicates from plotID
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
```

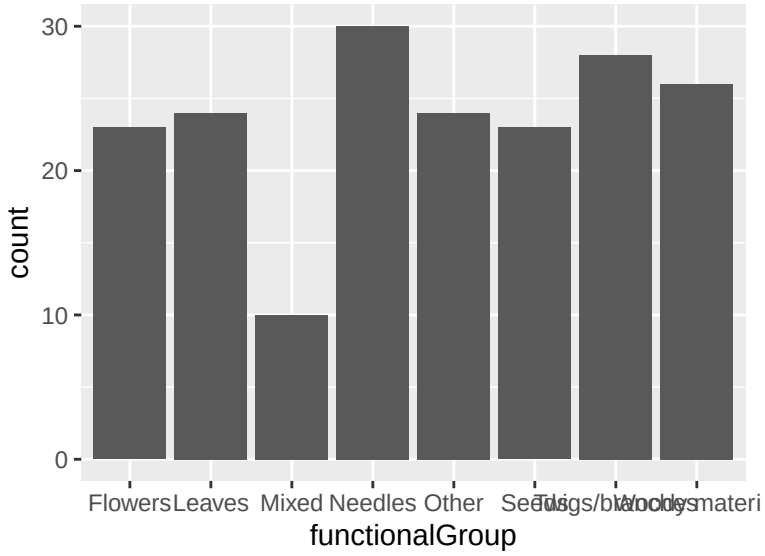
```r
print(unique(Litter$plotID))
```

```
##  [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
```

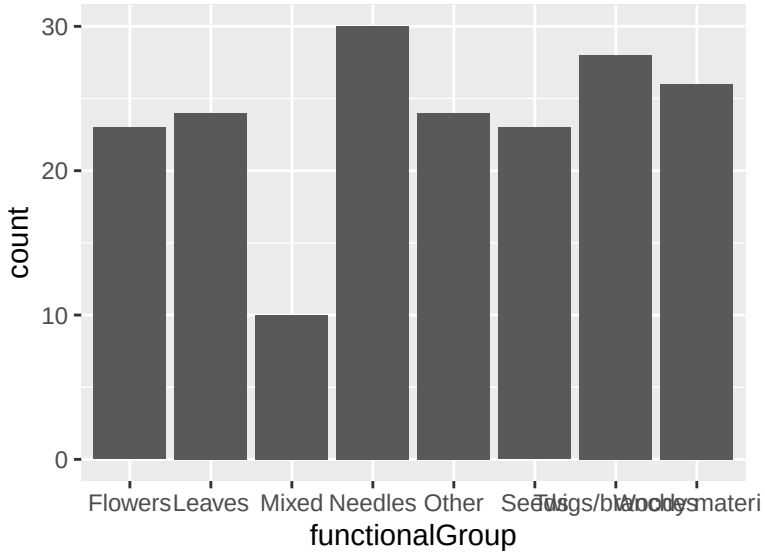
Answer: Twelve plots were sampled at Niwot Ridge. The unique function removes duplicates from plotID.

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.

```r
# Plotting functionalGroup counts as bar graph
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar()
```

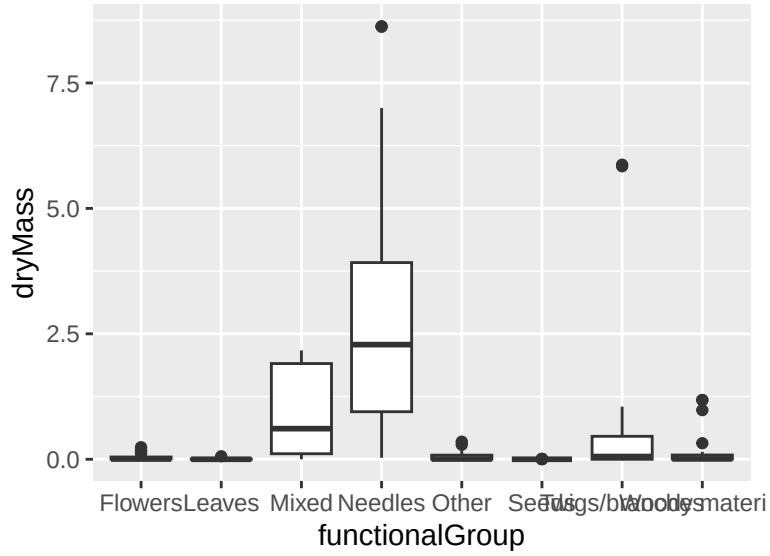



```
print(ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar())
```

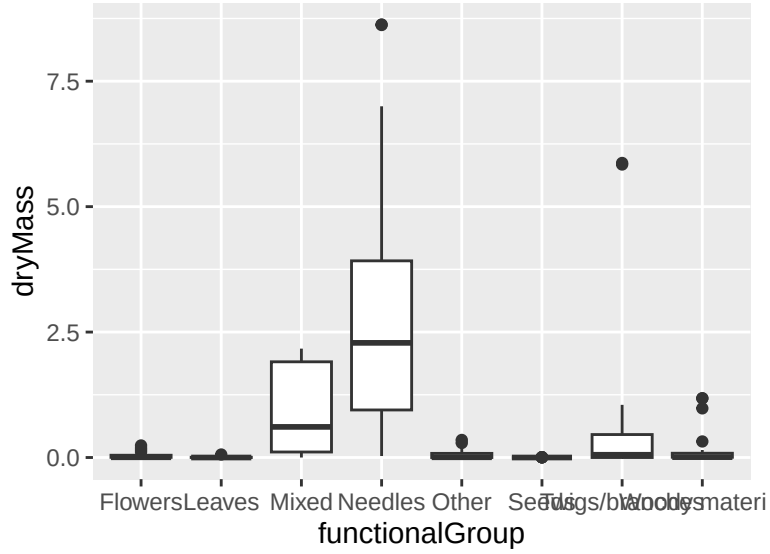


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

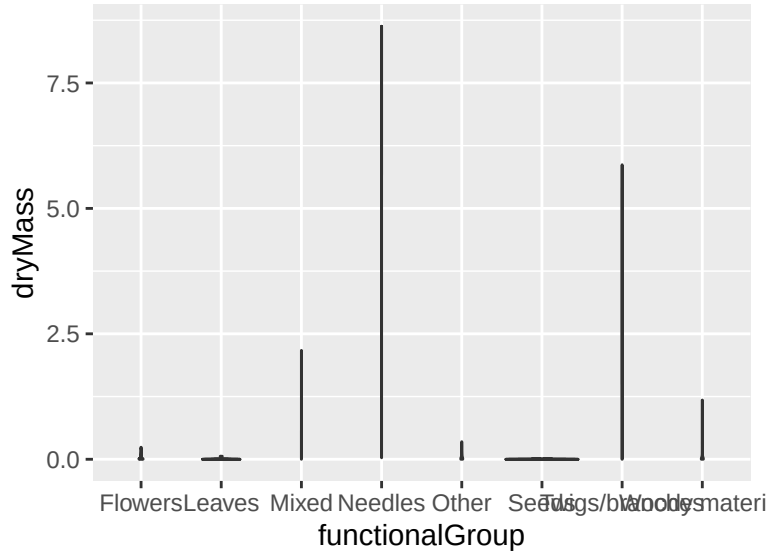
```
#
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass))
```



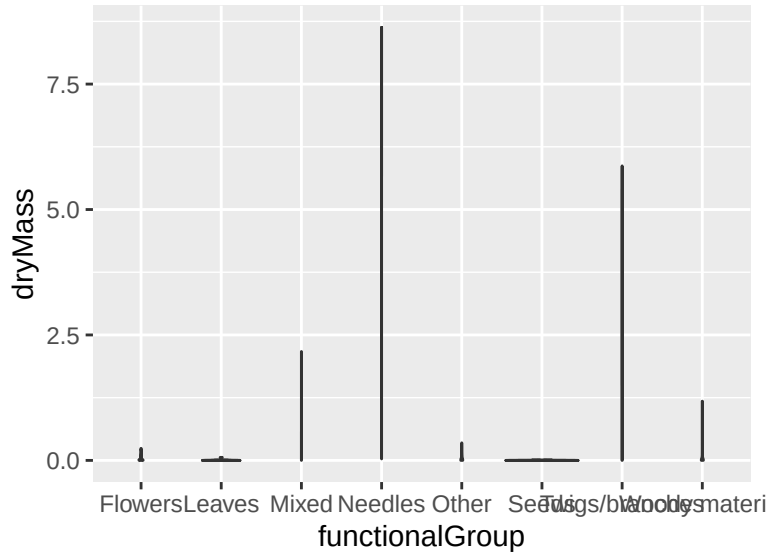

```
print(ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass)))
```




```
#
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75))
```




```
print(ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              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: The dryMass is being shown by functionalGroup and functionalGroup is made up of elements like flowers and leaves, whose masses don't have associated probability densities.

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

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