

Assignment 3: Data Exploration

Israel Golden, Section #2

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Exploration.

Directions

1. Change “Student Name, Section #” on line 3 (above) with your name and section number.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., “FirstLast_A03_DataExploration.Rmd”) prior to submission.

The completed exercise is due on <>.

Set up your R session

1. Check your working directory, load necessary packages (tidyverse), 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 add the stringsAsFactors = TRUE parameter to the function when reading in the CSV files.**

```
library(tidyverse)

# load data
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: Neonicotinoids are a type of pesticide that was developed and released in the 1990s to serve as an effective alternative to contemporary pesticides. The benefit of neonicotinoids over competing pesticides is that they are safe for human consumption. While effective in killing the targeted insect agricultural pests, neonicotinoids spread far beyond the farm into neighboring ecosystems. Neonicotinoid dust from agricultural operations easily spreads by the wind into other locales. In addition, its spread was (and is) aided by its solubility in water which allows for contamination of waterways and subterranean insect larvae. The cumulative effect of the spread of this pesticide is largescale devastation for insect populations anywhere near - or connected by wind currents or waterways - neonicotinoids are spread. To make matters worse, neonicotinoids have long lifetimes (between several months to a couple of years). This is undoubtedly bad news for ecosystem function in affected areas as insects play invaluable roles as a food source and for their role in pollination.

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: A better question is why WOOLN'T we be interested in litter and woody debris! Woody debris is important for understanding productivity, nutrient fluxes and carbon storage in forest ecosystems. Woody debris and litter also serves as some of the prime nutrient inputs for river and stream ecosystems. If we are to connect this dataset thematically to the neonicotinoids dataset, woody debris and litter are habitat for the insects that we are concerned about being affected by neonicotinoids!

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: * Litter trap pairs consist of one elevated trap and one ground trap. They are spaced such that one pair exists for every 400m² plot area which totals 1-4 traps per plot. * Depending on the vegetation of a given site, litter trap placement can either be targeted or randomized. * Ground traps are sampled once per year while elevated traps are checked more frequently based on the type of forest and differences in litter-fall that arise from phenophases (e.g., senescence in temperate forests during Fall).

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

```
# 4623 rows, 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
```

Answer: The four most common studied effects of neonicotinoids on insects are population (1803), Mortality (1493), Behavior (360), and Feeding behavior (255). These effects are specifically of interest to ecotoxicologists because they want to understand how this agricultural pesticide - released into neighboring ecosystems - affects population dynamics of insects (population, mortality). With respect to behavior and feeding behavior, I suspect they would want to know how feeding behavior - and behavior in general - exposes them to neonicotinoids.

- 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.

```
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 Woolly 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

Answer: The six most commonly studied insects in this dataset are the Honey Bee (667), the Parasitic Wasp (285), the Buff Tailed Honey Bee (183), the Carniolan Honey Bee (152), the Bumble Bee (140), and the Italian Honeybee (113). All of these species are in the hymenoptera order of insects and are prolific - and charismatic - pollinators. As for why they might be of more interest than other species of insects, I believe their prowess in pollination and overall charisma make them the worthy subjects of this focus and research.

8. Concentrations are always a numeric value. What is the class of Conc.1..Author. in the dataset, and why is it not numeric?

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

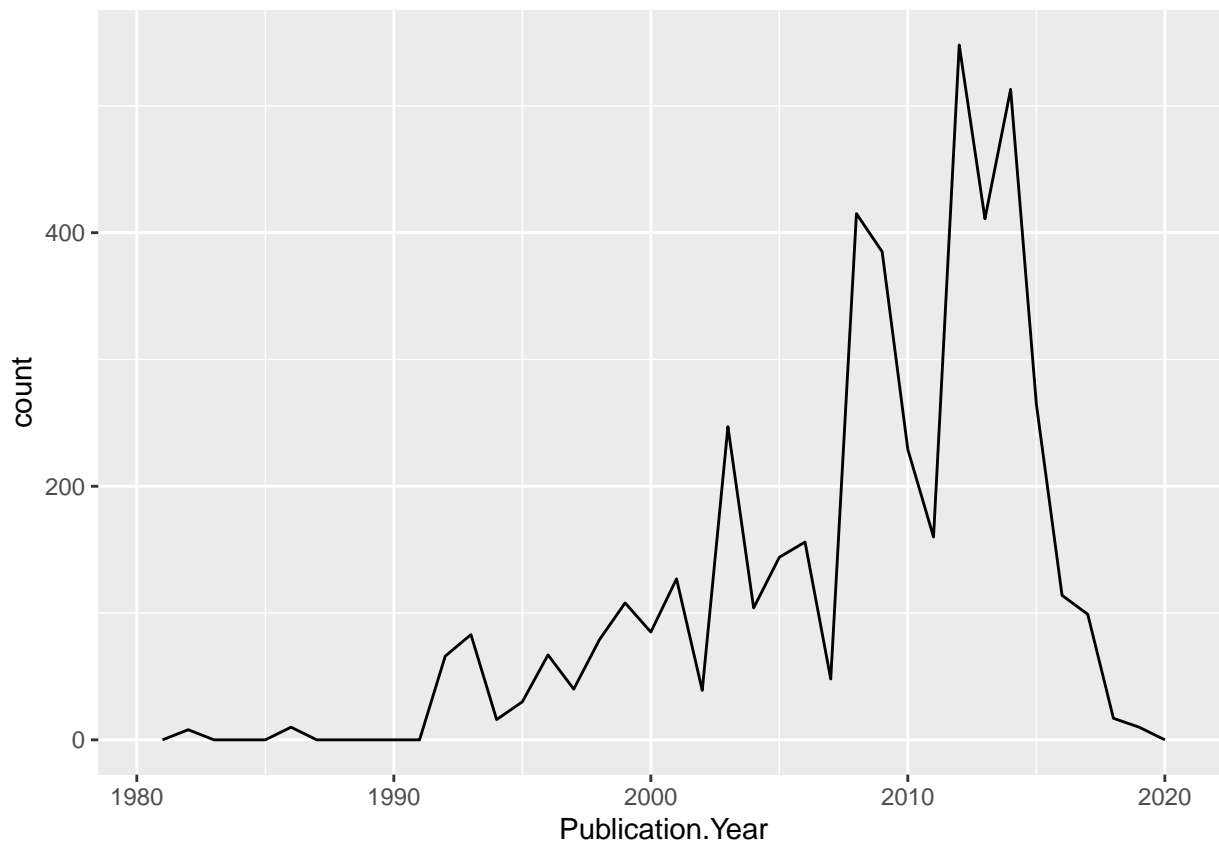
```
## [1] "factor"
```

Answer: The Conc.1..Author column holds factor data. In most cases, this column would be numeric except that it holds more than numbers. Throughout the column are ‘/’ to indicate approximate doses as well as ‘NR’ values which likely stand for not recorded. Since there is a mixture of numeric, grammatical, and alphabetical characters, R interprets this column as a string. Finally, because we included the command `stringsAsFactors = TRUE` when we imported the .csv file, it reads as a factor.

Explore your data graphically (Neonics)

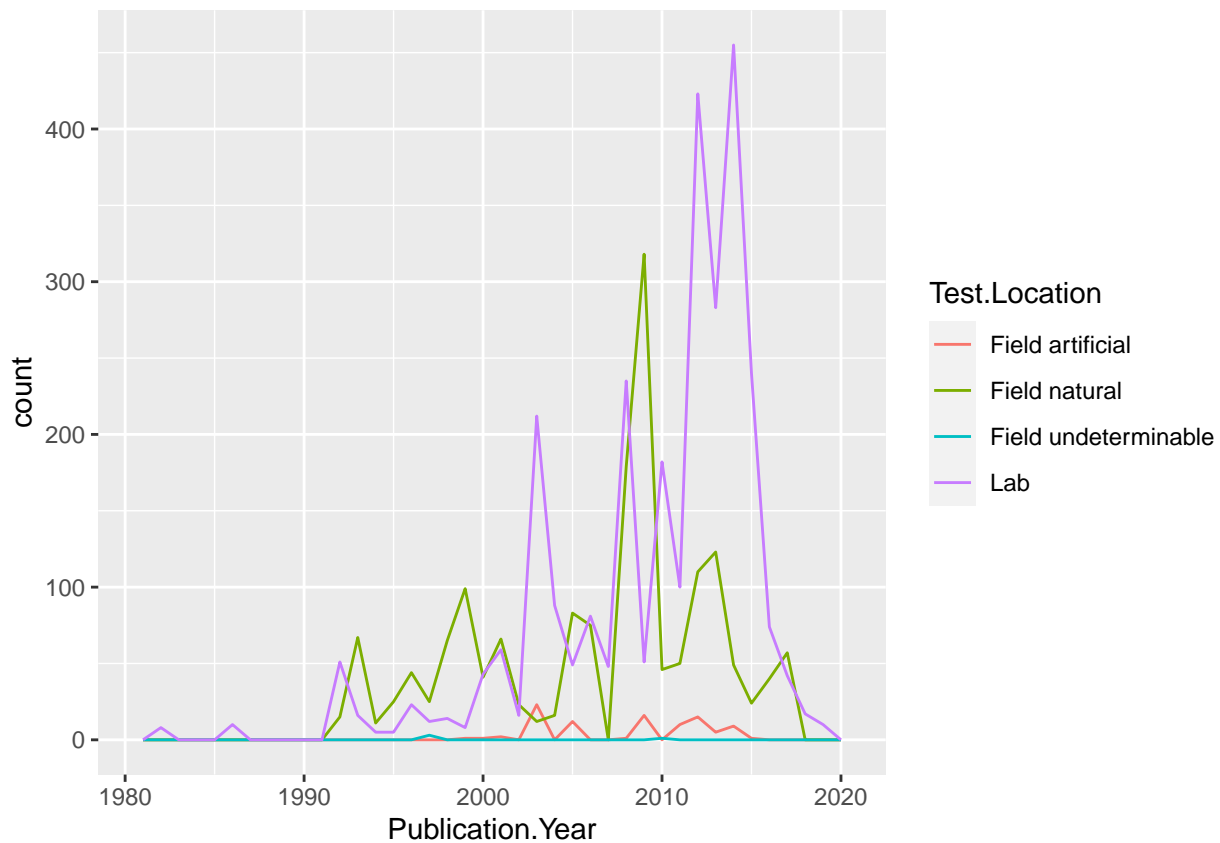
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 38)
```



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 = 38)
```

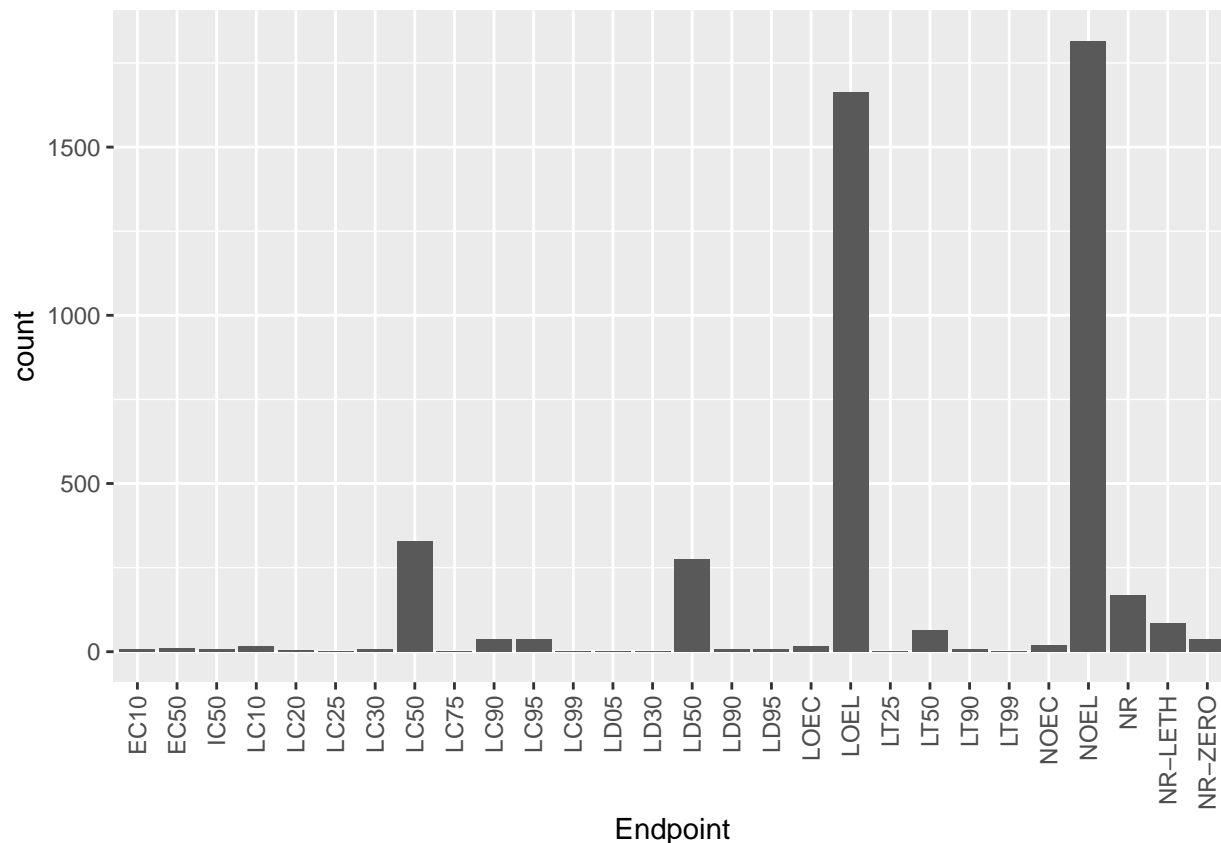


Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: The most common test locations are the lab and the natural field. Field studies seemed to have peaked around 2009 meanwhile lab studies peaked (higher than field studies) around 2014.

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.

```
endpoints <- ggplot(data = Neonics, aes (x = Endpoint))
endpoints + geom_bar(stat = "count") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: The two most common endpoints are LOEL and NOEL. LOEL is defined in the ECOTOX_CodeAppendix as “Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC)”. Similarly, ECOTOX_CodeAppendix defines NOEL as “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)

- 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"
```

```
# Not a date!
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##   date, intersect, setdiff, union
```



```
Litter$collectDate <- ymd(Litter$collectDate)
class(Litter$collectDate)
```

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

```
# Hurray, it's a date!
unique_dates <- unique(Litter$collectDate)
unique_dates
```

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

```
# Litter was sampled on 2018-08-02 and 2018-08-30
```

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_plots <- unique(Litter$plotID)
unique_plots
```

```
## [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
```

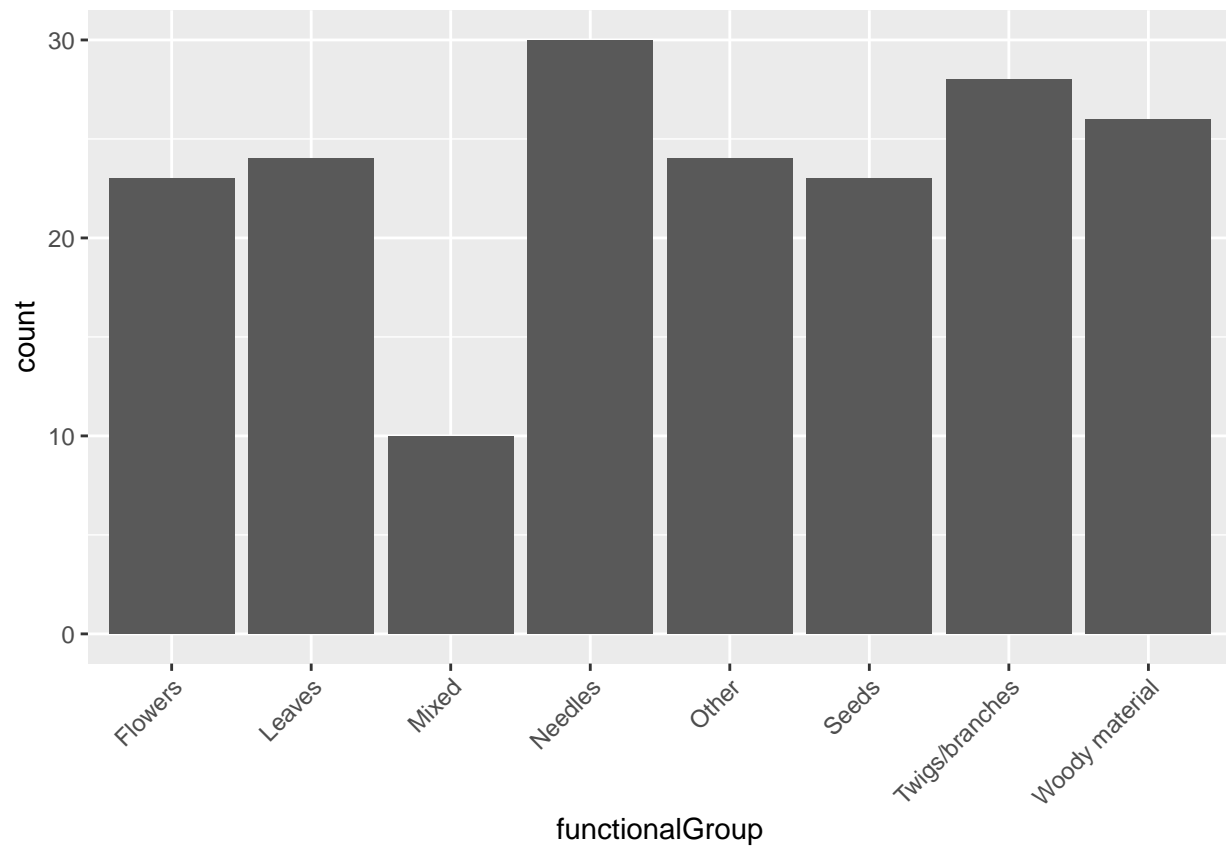
```
summary(Litter$plotID)
```

```
## 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 unique plots being sampled at Niwot Ridge. Summary returns the names of all 12 plots as well as the number of times each plot has been sampled. Unique only tells you how many unique plots there are.

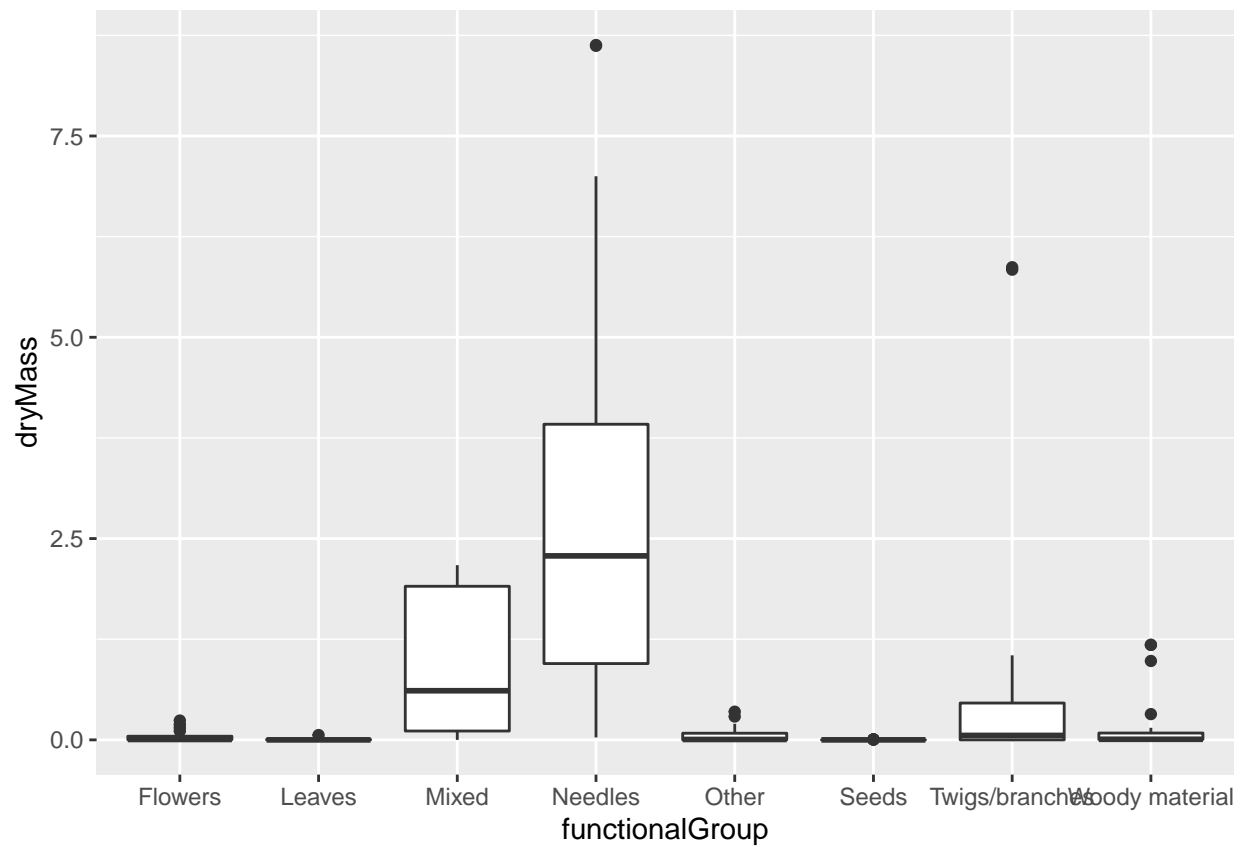
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
funcGroupCounts <- ggplot(data = Litter, aes (x = functionalGroup))
funcGroupCounts + geom_bar(stat = "count") +
  theme(axis.text.x = element_text(angle = 45, hjust=1))
```

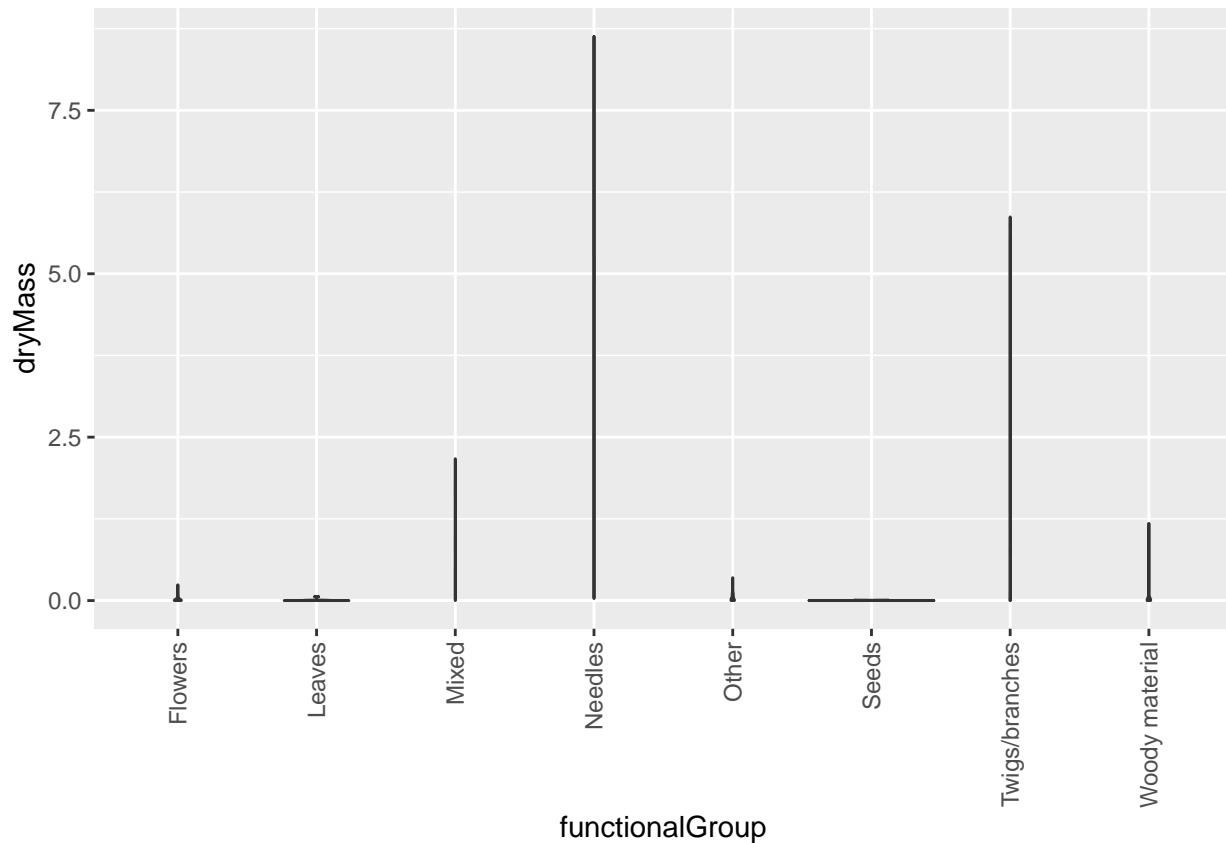


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by `functionalGroup`.

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



```
ggplot(Litter) +
  geom_violin(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: Boxplots show summary statistics like median, IQR, and outliers. Violin plots also show these but also show distribution of the data and are thus especially useful for multimodal datasets. However, when there are many unique values in a range, violin plots will have no width and appear as a straight line (as they do here). In this instance, the boxplots show the range far better and the mark of the median on the boxplot clearly shows the distribution of values.

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

Answer: Needles! This makes sense because there's often an abundance of pine needles on evergreen forest floors above any other kind of woody debris and our research site is in such a forest. The next highest biomass functional group are twigs and branches - while less abundant than needles, they are certainly heavier when they do occur.