

Assignment 3: Data Exploration

Yao Yao

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] "/Users/yaoyao/Desktop/ENV872/EDA-Spring2023"
```

```
#Load the two packages  
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr 0.3.4
## v tibble 3.1.8       v dplyr 1.0.10
## v tidyr 1.2.1        v stringr 1.4.1
## v readr 2.1.2        v forcats 0.5.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
library(lubridate)
```

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

```
#Upload two datasets and name the 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: Neonicotinoid is a kind of pesticides that permanently binds to the nerve cells of insects, overstimulating and destroying the insects. It is commonly used to help agriculture and cultivate crops. When the insects are degraded in the soil or eaten by other organisms, it will pose toxic effects on other organisms and contaminate the environment close by. Neonicotinoid will also be toxic to bees (they can help the pollination process) and other insects that have important economic and ecological values. Thus, it is important to research on this chemical and its effects on insects population, like bees.

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: Forest litter and woody debris are important components of healthy soil. The decomposition of them releases essential nutrients into the soil and also helps the soil to keep moist. They are also essential materials for nesting, hiding and protected spots for animals. It is important to research on them to learn about the soil and environmental quality in general.

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. Temporal Sampling Design: The sampling frequency varies by the kind of vegetation that is on different sites. For deciduous forest during senescence, the frequency is 1x every 2 weeks. For evergreen sites, the frequency is 1x every 1-2 months. 2. Spatial Sampling Design: The litter sampling took place in 20 40m x 40m plots for sites that have forested tower airsheds. On the other hand, for sites that have low-statured vegetation over the tower airsheds, the sampling took place in 4 40m x 40m tower plots plus 26 20m x 20m plots. 3. Spatial Sampling Design: Locations of tower plots are selected randomly within the 90% flux footprint of the primary and secondary airsheds and the litter and woody debris are sampled at terrestrial NEON sites that have woody vegetation >2m tall.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#Dimensions of the 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?

```
#Effect columns
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 that are studied include mortality, population, behavior and feeding behavior. Mortality is directly related to the research of toxicity of neonicotinoid to insects. By seeing the fluctuation of population when neonicotinoid is given, the effect of neonicotinoid on insects can also be discovered. Behavior and feeding behavior may be important factors that closely related to how neonicotinoid is taken by the insects and how much are taken in, affecting insects' survival.

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...]

```
#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 Woolly 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: Except from the “Other” category, honey bee, parasitic wasp, buff tailed bumblebee, carniolan honey bee, bumble bee and italian honeybee are the top six most commonly studied species in the dataset.They are all very important to agriculture and crops and plants’ growth.

Different kinds of bees are heavily researched upon, since they create important ecological and economic value in the ecosystem as crop pollinators and the use of pesticides will heavily affect their population. Parasitic wasphelp farmers and gardeners in naturally controlling crops by killing those insects that are harmful to the crops, so if neonicotinoid can threat their survivalship, they couldn't control the harmful insects anymore.

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

```
#Class of "Conc.1..Author"
class(Neonics$Conc.1..Author.)
```

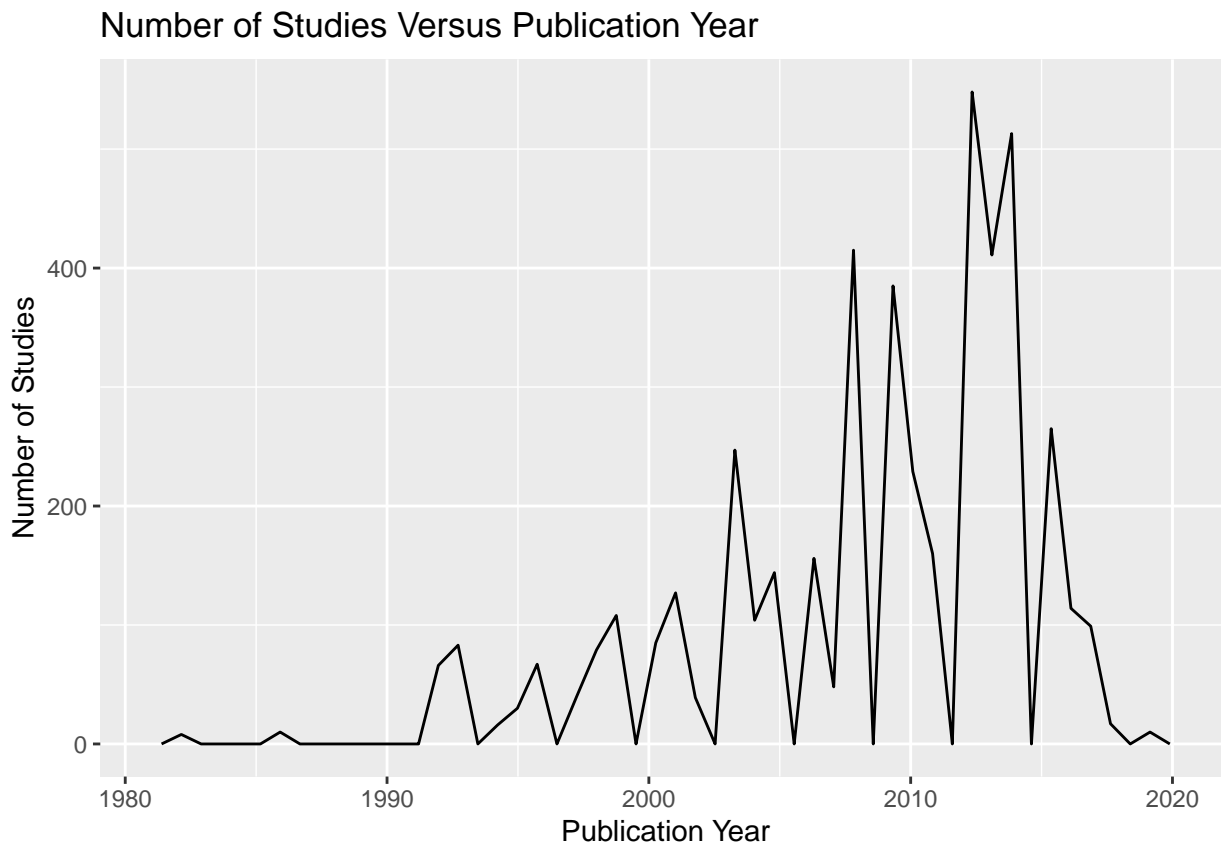
```
## [1] "factor"
```

Answer:It is a factor. By looking into the dataset, we can see that many entries are like “<0.025” which is a range rather than a specific number. In addition, there are also “NR” - not recorded data. As a resultmm the column can't be numeric.

Explore your data graphically (Neonics)

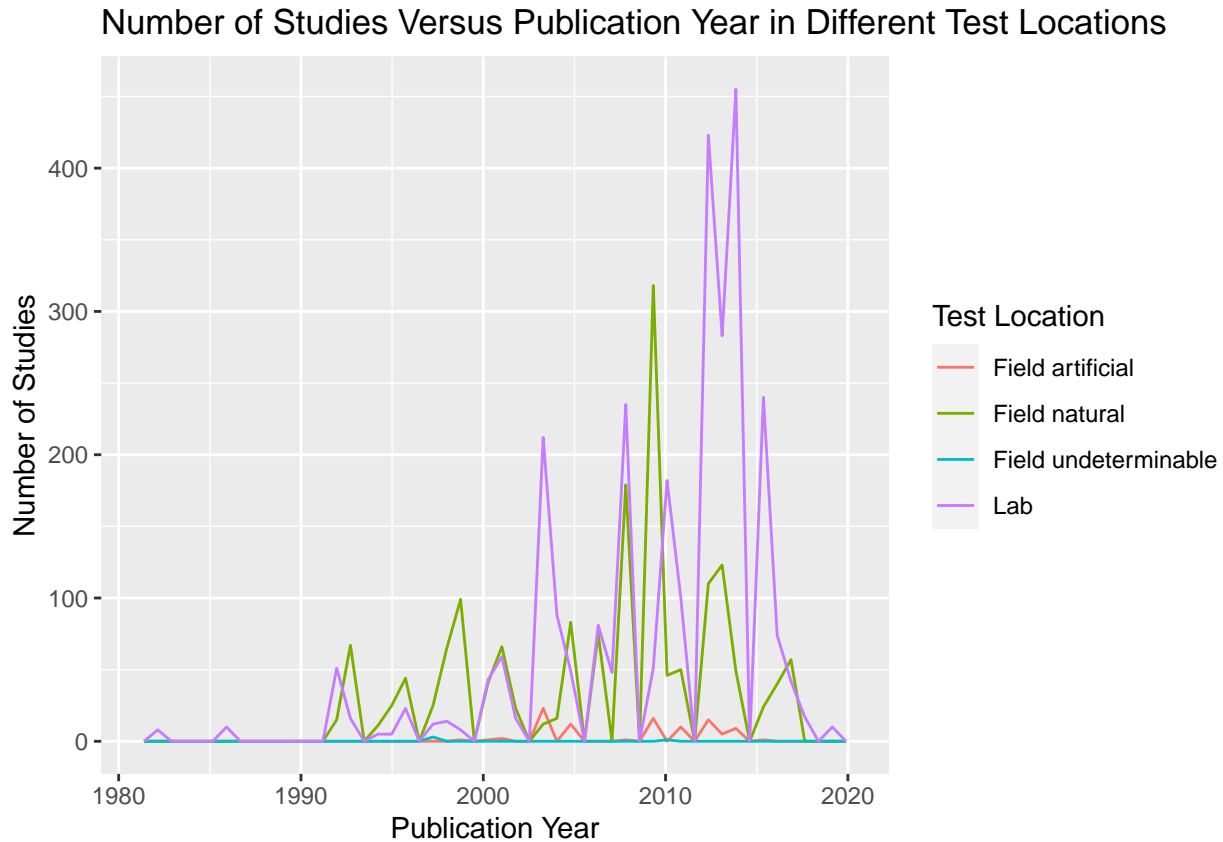
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
#Plot of the number of studies conducted by publication year
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50) + labs(y= "Number of Studies", x = "Publication Y
```



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

```
#Plot of the number of studies conducted by publication year at different test locations
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 50) + labs(y= "Number of Studies")
```



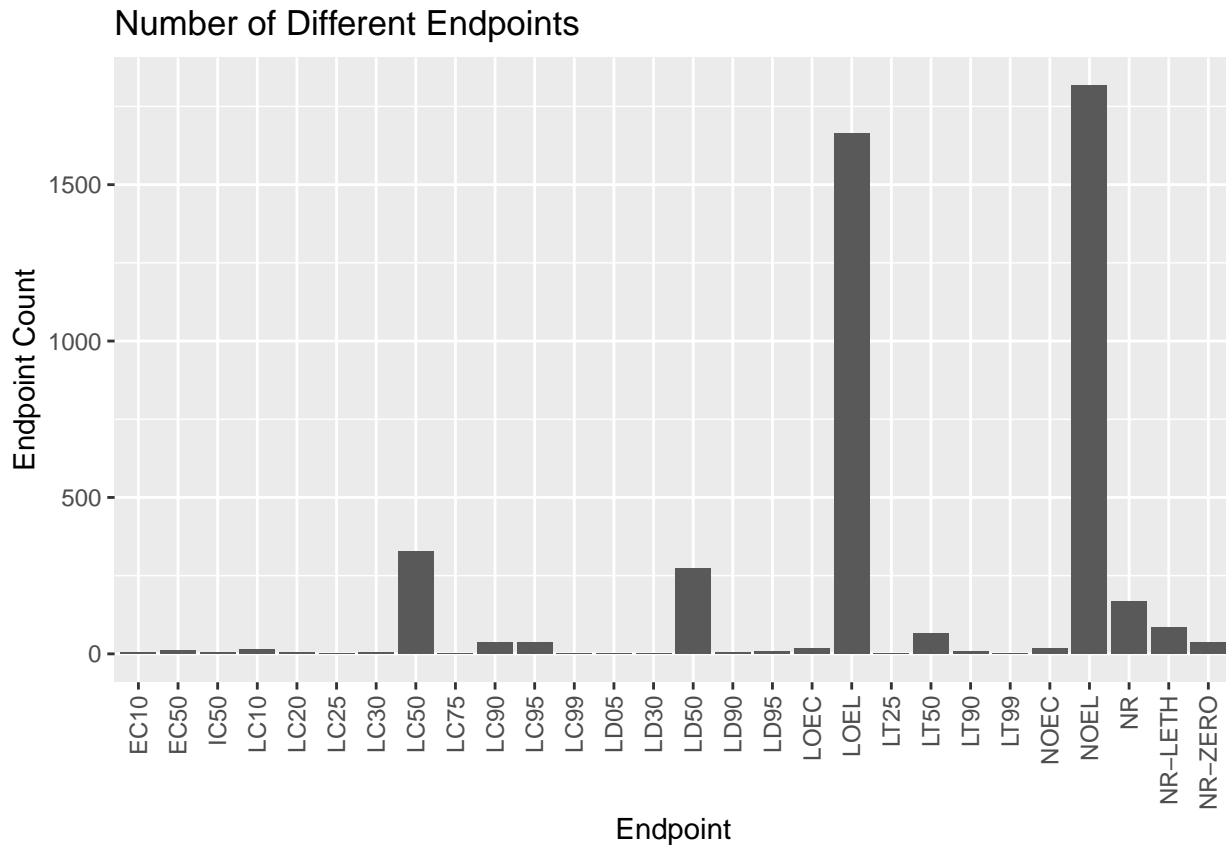
Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: The most common test locations are lab and natural fields, since they have the most counts of publications throughout the year. There are fluctuations on the number of publications throughout the year. Test in labs become very prevalent through 2010-2015 and went down from 2015-2020. Number of research on both locations increase gradually from 1980-2010 with a peak of number of publications conducted in natural field at around 2010. The number of publications in natural field then gradually went down after 2010.

11. Create a bar graph of Endpoint counts. What are the two most common endpoints, 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...]

```
#Plot of the number of different Endpoints
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar() + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + labs(y= "Endpoint Count")
```



Answer: NOEL and LOEL. NOEL is no-observable-effect-level. It means that dose with highest concentration produces effects that is not significantly different from responses of controls according to author's reported statistical test. LOEL is lowest-observable-effect-level. It means that dose with lowest concentration produces effects that is significantly different from responses of controls according to author's reported statistical test.

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.

```
#Determine the class of collectDate
class(Litter$collectDate)
```

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

```
#It is a factor
```

```
#Change to date
```

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

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



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

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

```
#It is 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`?

```
#Determine how many 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
```

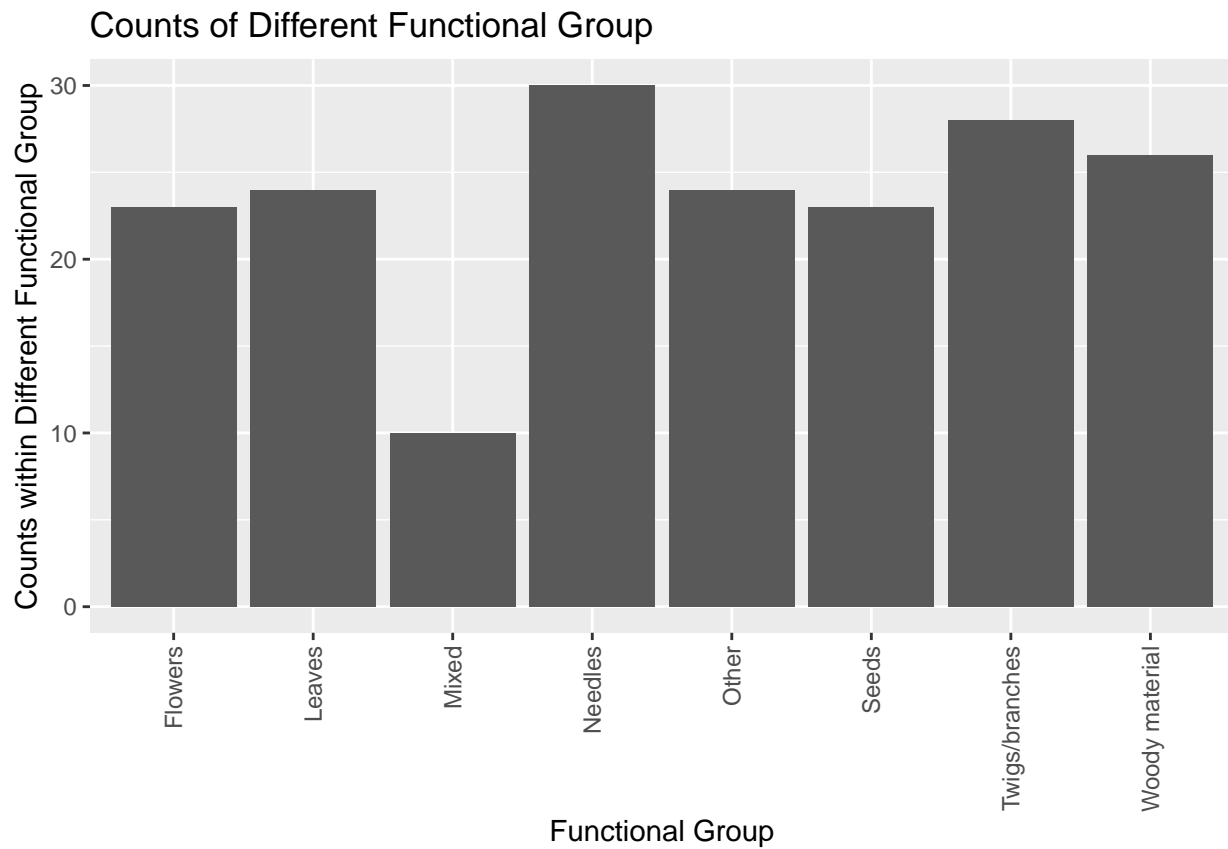
```
summary(Litter$plotID)
```

```
## 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: 12 plots were sampled. Unique function gives you the number of unique values and what they are. Summary function gives use the unique vlaues with the number count of each value in the dataset.

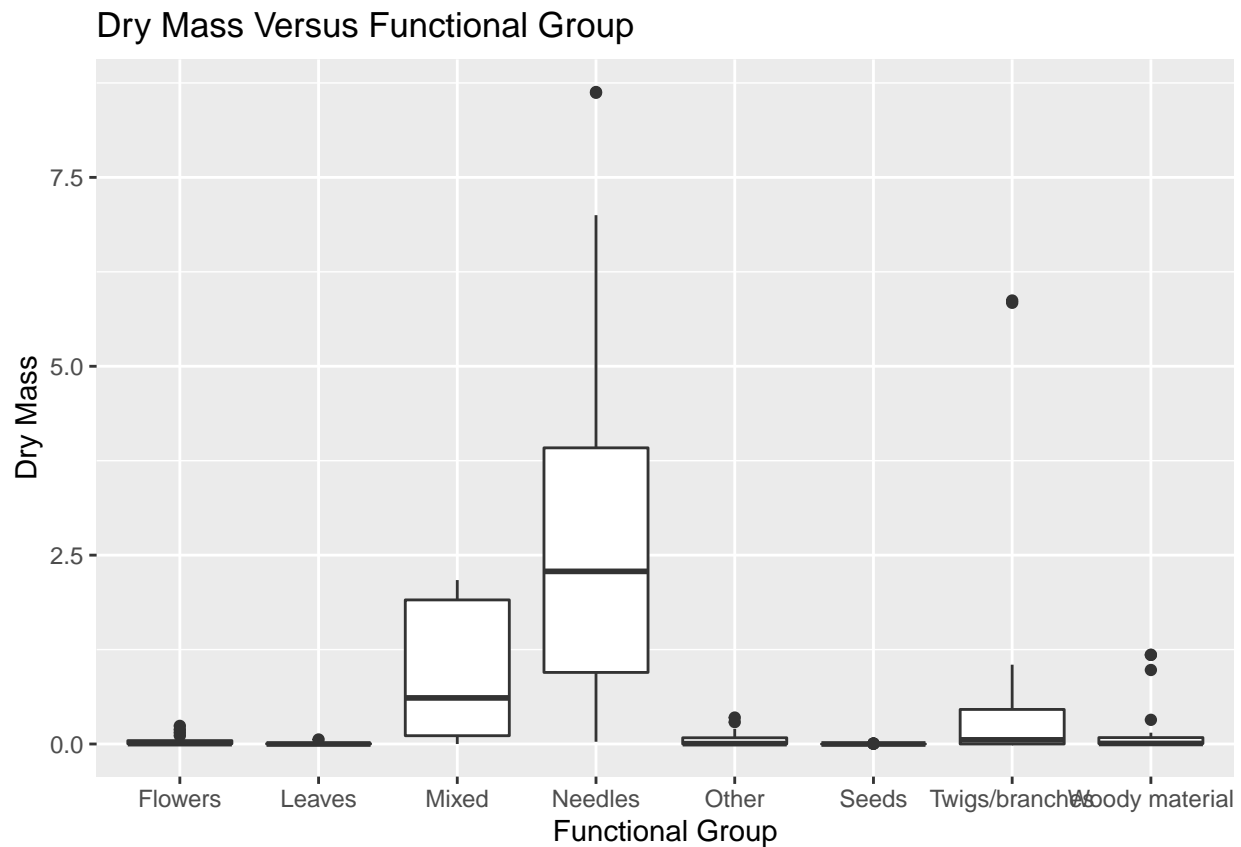
14. Create a bar graph of `functionalGroup` counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#Bar graph of functionalGroup
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar() + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) + labs(y= "Counts wi
```



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

```
#Boxplot of dryMass by functionalGroup  
ggplot(Litter) +  
  geom_boxplot(aes(x = functionalGroup, y = dryMass)) + labs(y= "Dry Mass", x = "Functional Group") + 1
```

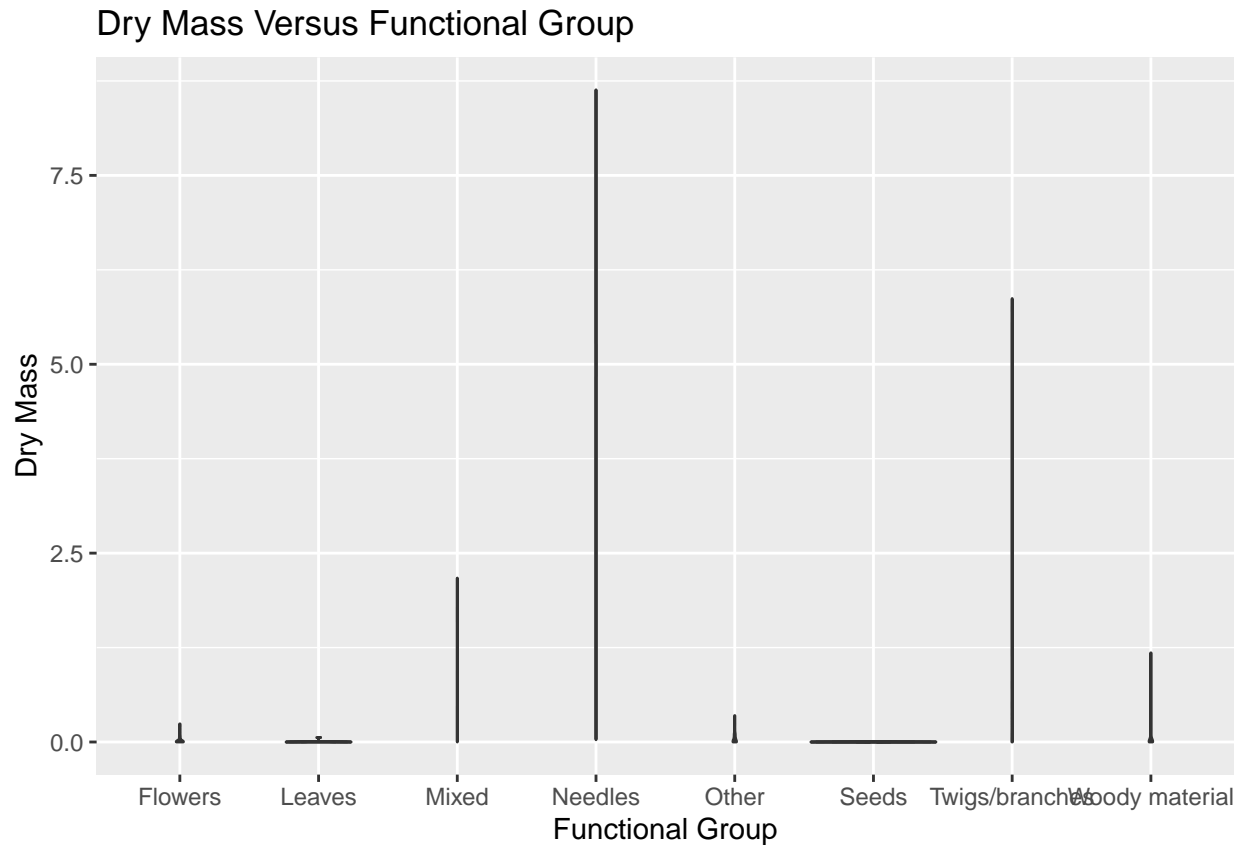


```
#Violin plot of dryMass by functionalGroup
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
    draw_quantiles = c(0.25, 0.5, 0.75)) + labs(y= "Dry Mass", x = "Functional Group") + labs
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```



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

Answer: The data in each category is either very concentrated at a certain value or very distributed (the values are spread out and not repeating), so the width of the violin plot is either very thin or very wild. Thus, we can not extract effective information from the width of the violin plot. The boxplot on the other hand doesn't use the width to show the number of datapoints. It shows the IQR, median, the range and outliers of the data, which gives the audience a better sense of the actual distribution of the data.

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

Answer: Needles, mixed, and twigs/branches.