Assignment 3: Data Exploration

Enikoe Bihari, 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 1/31/2022.

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.**

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

```
## [1] "C:/Users/eniko/Documents/Duuuuuke/2021-22/Data Analytics/Environmental_Data_Analytics_2022/Assig
```

```
# get and load the packages you need
# install.packages("tidyverse")
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5

## Warning: package 'ggplot2' was built under R version 4.0.5

## Warning: package 'tibble' was built under R version 4.0.5

## Warning: package 'tidyr' was built under R version 4.0.5

## Warning: package 'readr' was built under R version 4.0.5

## Warning: package 'dplyr' was built under R version 4.0.5

## Warning: package 'forcats' was built under R version 4.0.5
```

```
library(ggplot2)

# get the csv files you need
neonics <- read.csv("../Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",
```

```
                      stringsAsFactors = T)
litter <- read.csv("../Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",
                      stringsAsFactors = T)
```

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 ecotoxicologoy of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: There are lots of "good" insects that are harmed by the insecticides used in conventional agriculture. Oftentimes, these chemical don't discriminate between pests and neutral/beneficial insects. For example, bees underpin our entire food system, since they are used to pollinate the crops. In a similar way, many other insects that go unnoticed by us actually form the foundation for entire ecosystems, as they provide critical ecosystem services (serving as food for other animals, pollinators and decomposers for plants). Thus, while insecticide use is often only intended for removing a few insects from and specific plot of land, it can end up harming entire ecological systems.

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 play a myriad of ecological roles. They form the matter that eventually beomes decomposed into soil, they provide food and habitat for many animals, plants, and fungi (among which are the decomposers that make the soil), and they make up an often overlooked but significant component of forest carbon. The size and residence time of this last one can vary tremendously among ecosystems, but in some places, litter, debris, and the soil underneath them can provide a lot of relatively stable carbon storage. Removing litter and debris often increases soil decomposition and erosion rates (which increases CO_2 emissions and soil loss), and it also decreases biodiversity through habitat degradation.

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 is defined as: diameter <2cm and length <50cm; fine woody debris is defined as: diameter <2cm and length >50cm
   2. different sampling frequencies in deciduous forests (1x every 2 weeks) and evergreen forests (1x every 1-2 months)
   3. mass for different functional groups has accuracy of 0.01 g; weights <0.01g indicate presence of functional group but not at a detectable mass

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# get dimensions
dim(neonics)
```

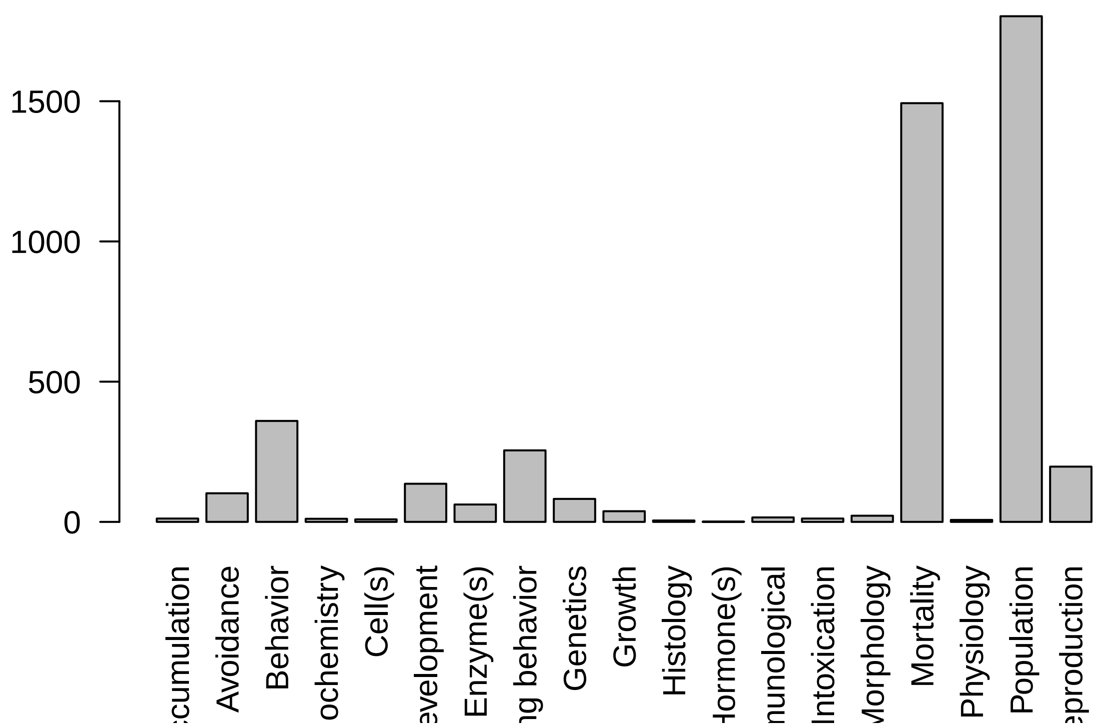
```
## [1] 4623   30
```

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?

```
# find the most common effects used in the study
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
```

```
barplot(summary(neonics$Effect), las = 2)
```



Answer: Population appears the most, followed by Mortality. Population is a good response variable to look at because it is ultimately dictated by of a bunch of other smaller factors interacting and compounding on one another (e.g. the other variables in this data set, like development, reproduction, mortality, growth, physiology, biochemistry, etc.). It is a good overall indicator of whether something is harming or benefitting a species, because it embodies the effects of all the other factors. Ultimately, it is the variable that first drew attention to the problem and is the easiest to measure, so it is the key rallying point for biologists in terms of pesticide use. Mortality makes up half the equation of whether a population grows, shrinks, or remains stable, so it is also a good metric of overall wellbeing of a species (and might be easier to measure than total population). It is also the result of a bunch of other interacting factors, and it is the ultimate outcomes we want to avoid.

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.

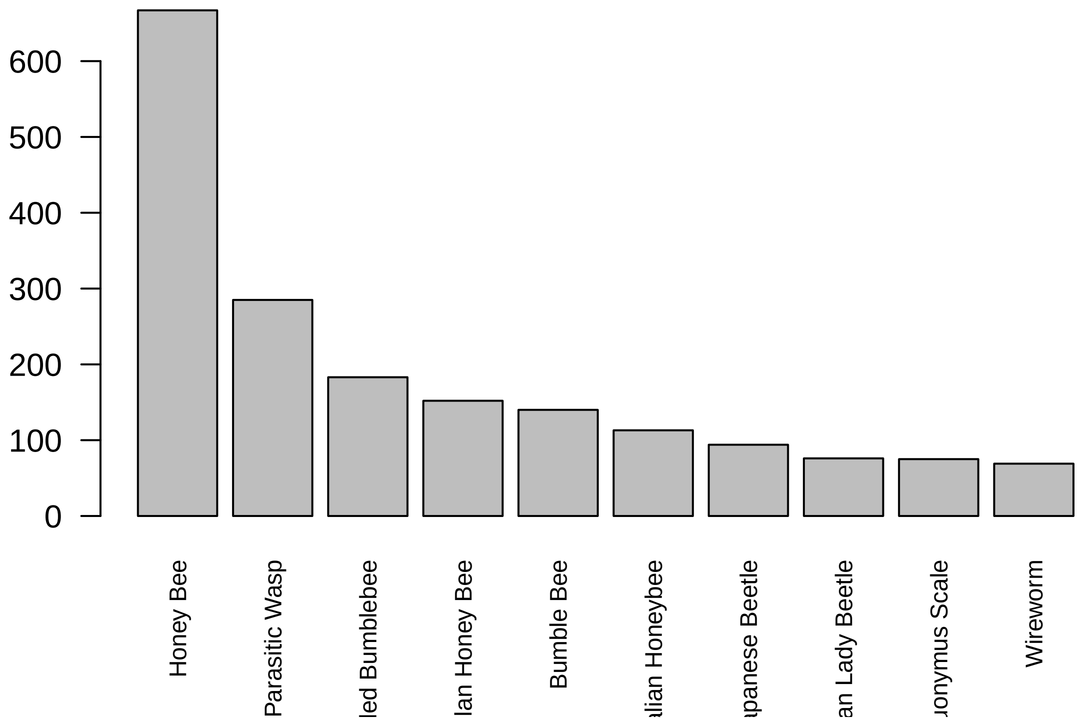
```r
# find the most common species used in the study
summary(neonics$Species.Common.Name)
```

```
##                        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 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
barplot(summary(neonics$Species.Common.Name)[1:10], las = 2, cex.name = .75)
```



Answer:
1. Honey Bee
2. Parasitic Wasp
3. Buff Tailed Bumblebee
4. Carniolian Honeybee
5. Bumble Bee
6. Italian Honeybee

All of these are in the Hymenoptera family, which are key agricultural pollinators. Much of our agricultural system depends solely on pollination from these guys, and they have become a product in and of themselves in the industry. Farmers hire beekeepers - who actually travel the country for this - to park their hives near their fields to pollinate their crops. Thus, bees have a tremendous economic value riding on them. I imagine the wasps and other more obscure bees might have been used as indiciator species or proxies for their more important relatives (maybe they were cheaper or easier to experiment on).

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

```r
# investigate the data column
class(neonics$Conc.1..Author.)
```

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

```r
head(neonics$Conc.1..Author.)
```

```
## [1] 27.2 19.7 47   25   13   268
## 1006 Levels: ~10 ~30/ ~40/ ~41 <0.0004 <0.025 <0.088 <0.5 <1.5 <10/ ... NR/
```

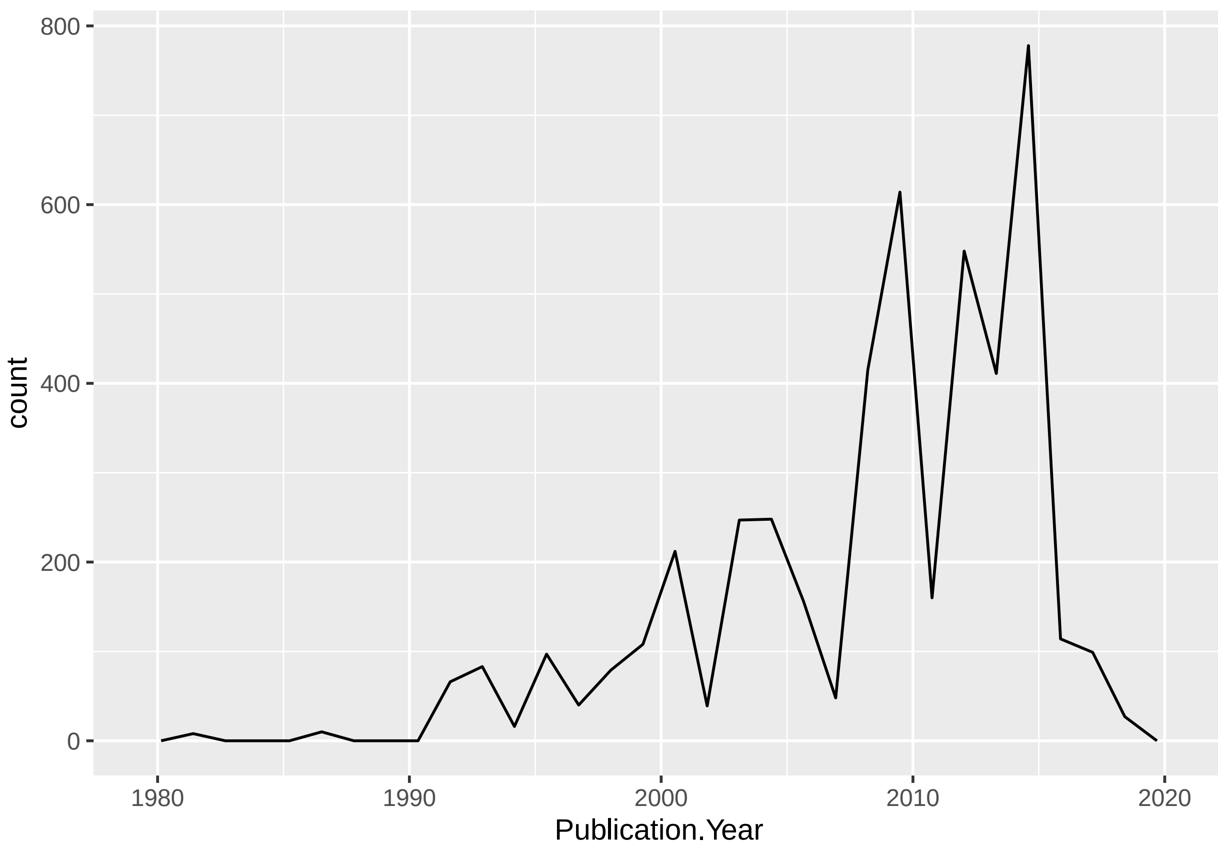
> **Answer:** It appears that there are some symbols like ~, <, and / in the column, which makes R default to reading them as factors rather than numbers when generating the data frame.

Explore your data graphically (Neonics)

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

```
# make the plot using ggplot
ggplot(neonics) +
  geom_freqpoly(aes(x = Publication.Year))
```

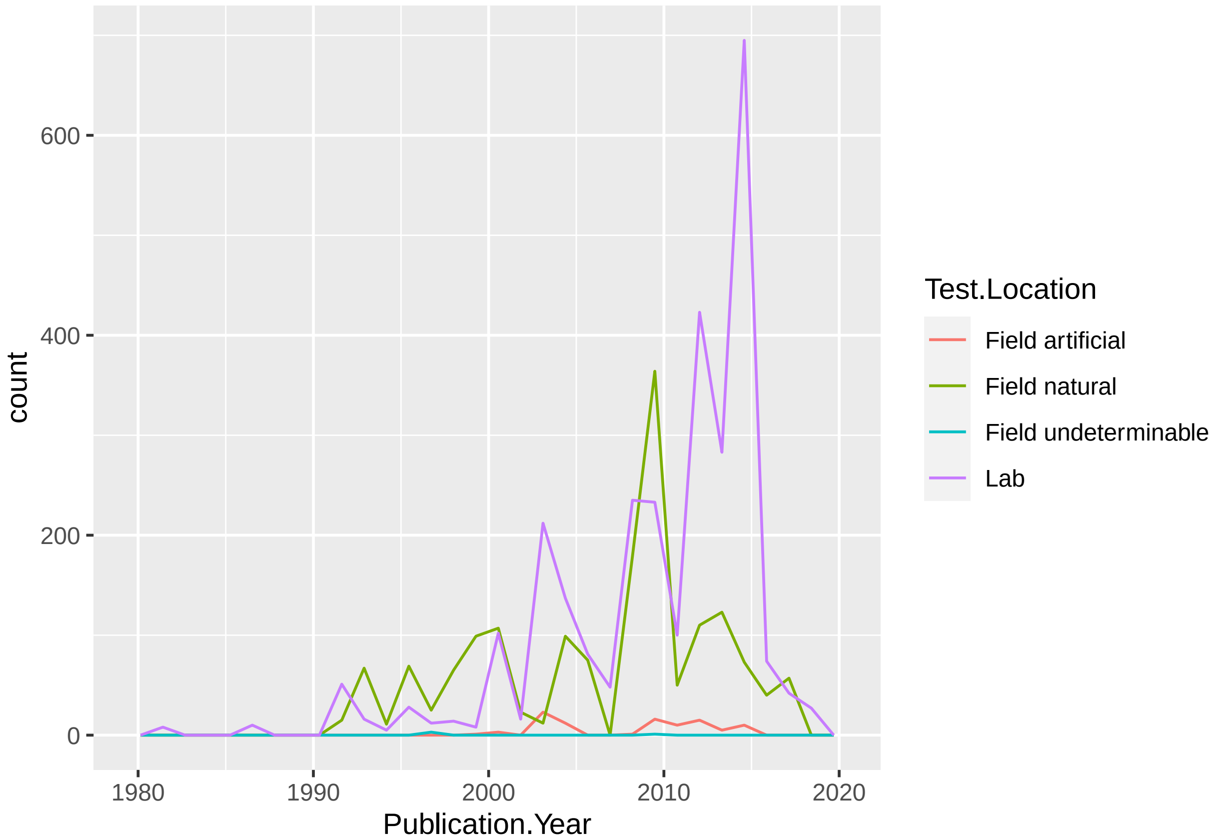
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors. Interpret this graph. What are the most common test locations, and do they differ over time?

```
# split up the plots by test location
ggplot(neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location))
```

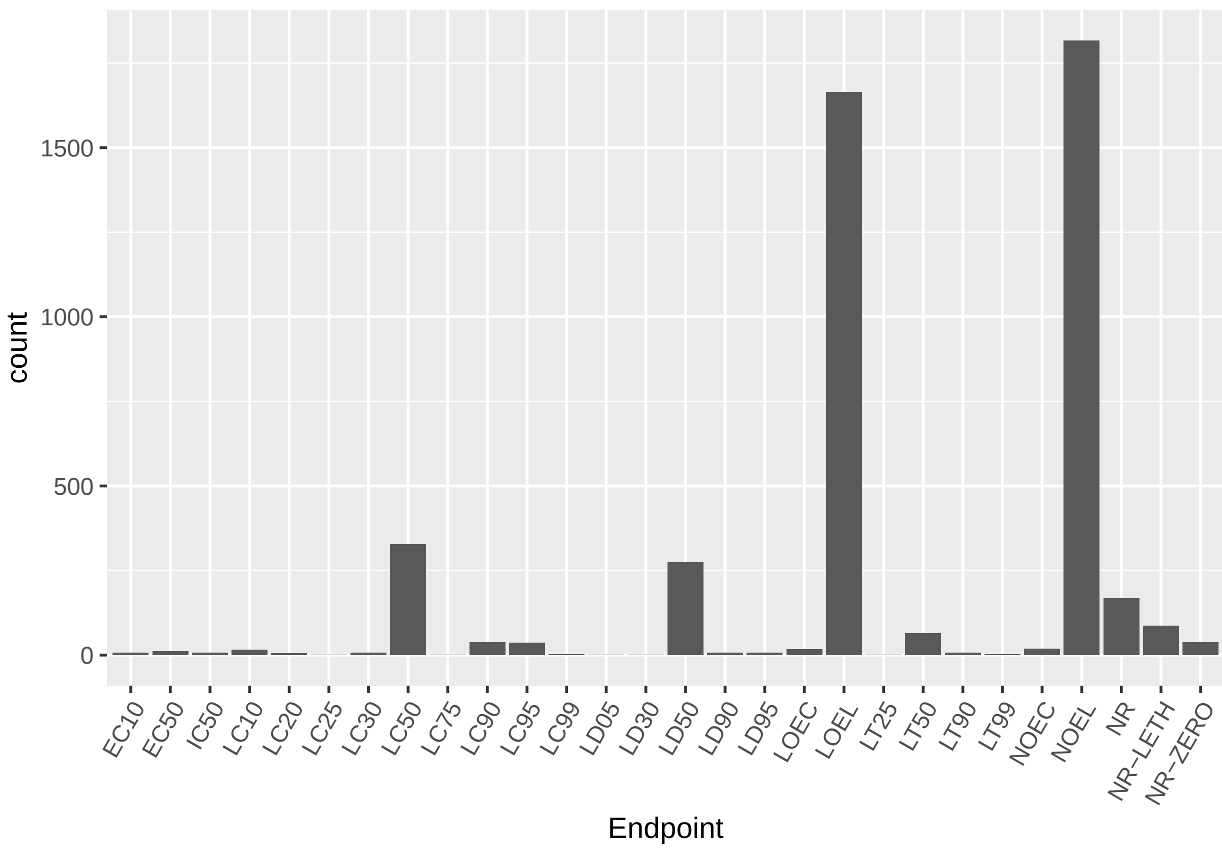
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



Answer: Lab tests dominate, expecially recently (after 2010) when they drastically outnumbered other test locations. Natural field tests are second, and there are a few periods of time when these outnumbered lab tests (1990-2000 and just befor 2010). In the 80s, only lab tests were done, and the other types of tests only started getting popular after 1990. Artificial field tests are by far the least common, and it appears that there are some year in which none were conducted at all.

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.

```
# make the plot using ggplot
ggplot(neonics) +
  geom_bar(aes(x = Endpoint)) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



Answer:
1. 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)"
2. LOEL: "Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC)"

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.

```
# find the class of the collection date
class(litter$collectDate)
```

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

```
head(litter$collectDate)
```

```
## [1] 2018-08-02 2018-08-02 2018-08-02 2018-08-02 2018-08-02 2018-08-02
## Levels: 2018-08-02 2018-08-30
```

```
# make into a date and check
litter$collectDate <- as.Date(litter$collectDate, format = "%Y-%m-%d")
class(litter$collectDate)
```

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

```r
# find unique dates
unique(litter$collectDate)
```

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

```r
# find unique locations
unique(litter$namedLocation)
```

```
##  [1] NIWO_061.basePlot.ltr NIWO_064.basePlot.ltr NIWO_067.basePlot.ltr
##  [4] NIWO_040.basePlot.ltr NIWO_041.basePlot.ltr NIWO_063.basePlot.ltr
##  [7] NIWO_047.basePlot.ltr NIWO_051.basePlot.ltr NIWO_058.basePlot.ltr
## [10] NIWO_046.basePlot.ltr NIWO_062.basePlot.ltr NIWO_057.basePlot.ltr
## 12 Levels: NIWO_040.basePlot.ltr ... NIWO_067.basePlot.ltr
```

```r
summary(litter$namedLocation)
```

```
## NIWO_040.basePlot.ltr NIWO_041.basePlot.ltr NIWO_046.basePlot.ltr
##                    20                    19                    18
## NIWO_047.basePlot.ltr NIWO_051.basePlot.ltr NIWO_057.basePlot.ltr
##                    15                    14                     8
## NIWO_058.basePlot.ltr NIWO_061.basePlot.ltr NIWO_062.basePlot.ltr
##                    16                    17                    14
## NIWO_063.basePlot.ltr NIWO_064.basePlot.ltr NIWO_067.basePlot.ltr
##                    14                    16                    17
```

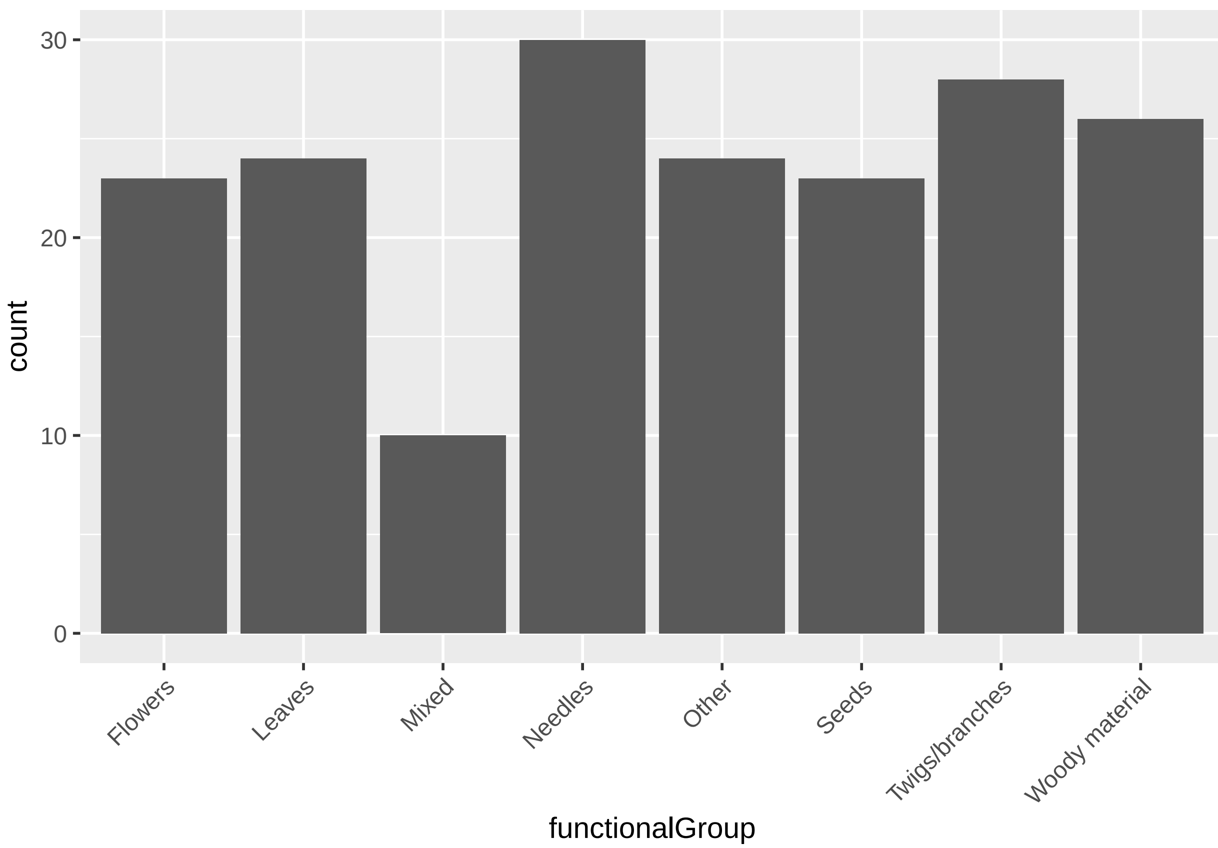
```r
length(summary(litter$namedLocation))
```

```
## [1] 12
```

> **Answer:** 12 unique plots; `unique`shows just the unique values in the column, while `summary` shows the unique values and the number of times they occur.

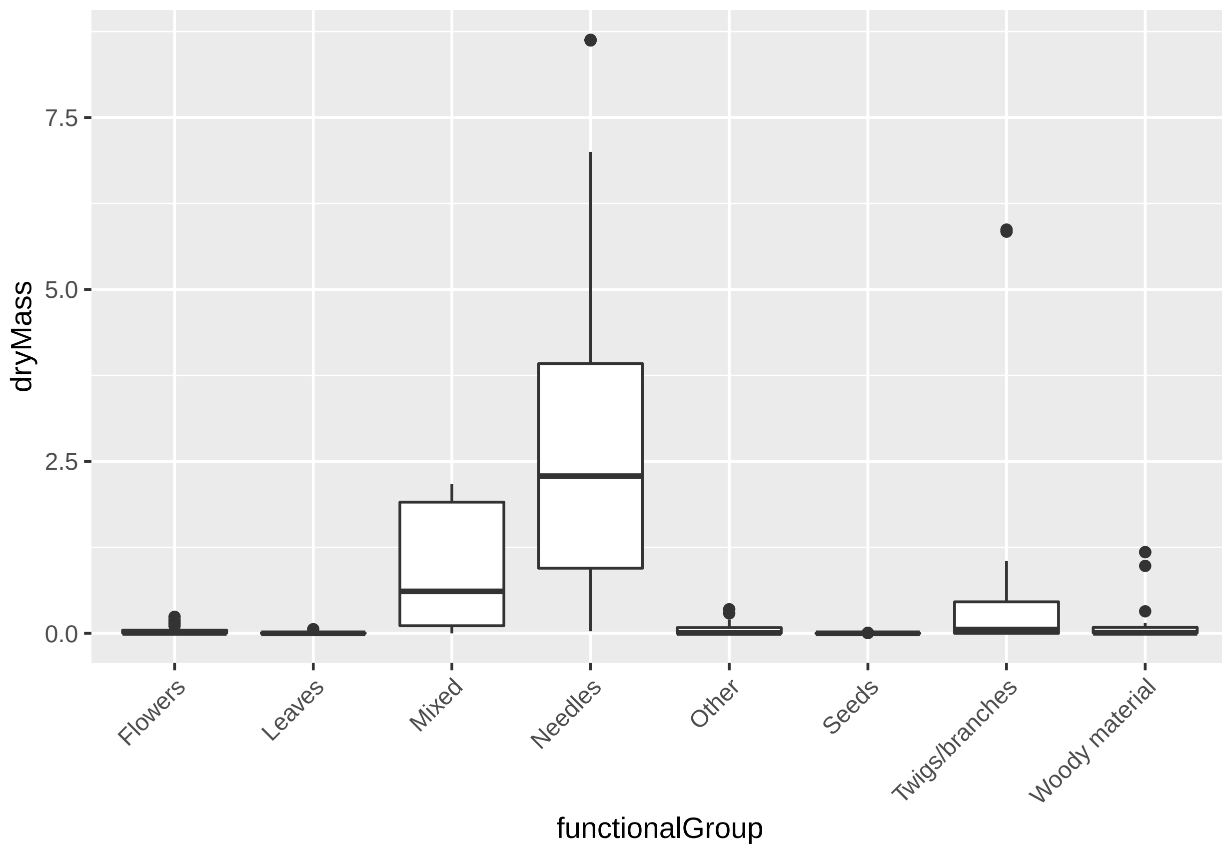
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
# make bar graph
ggplot(litter) +
  geom_bar(aes(x = functionalGroup)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

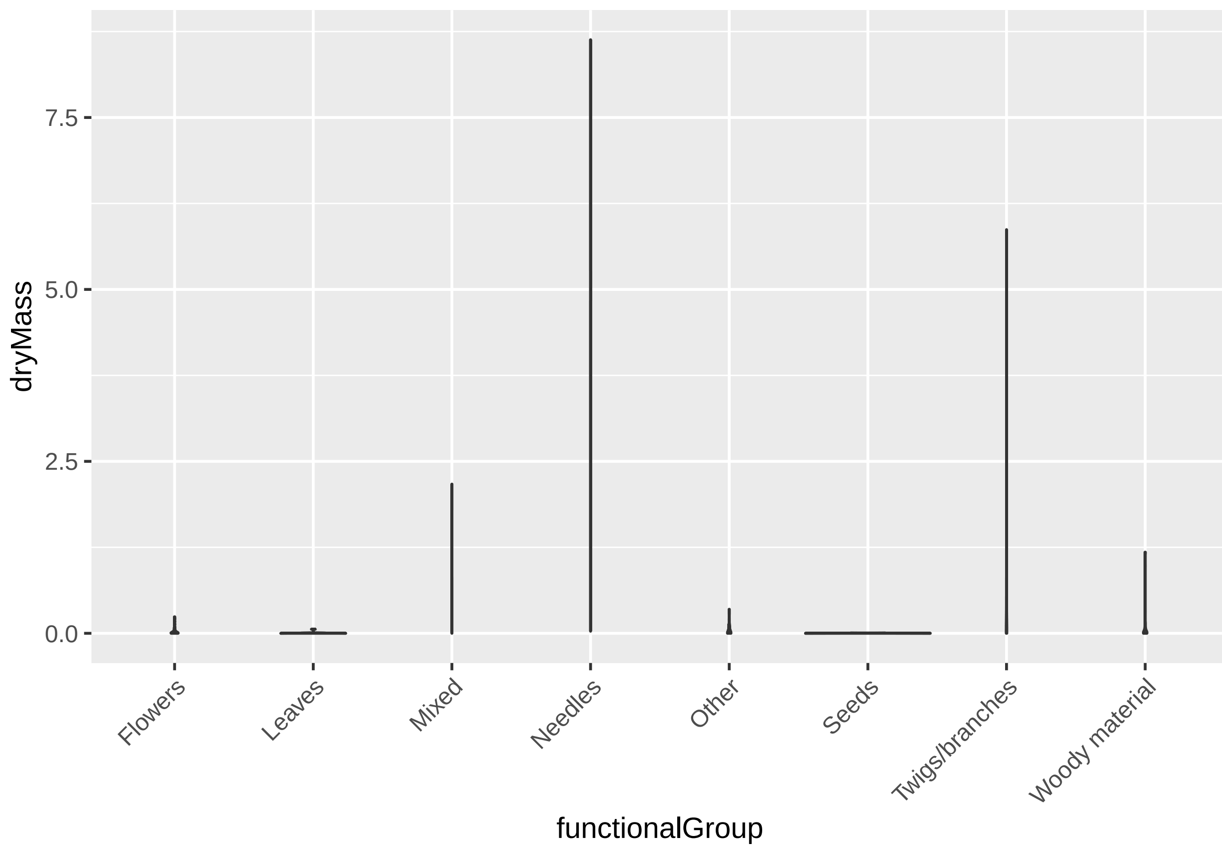


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# make boxplots
ggplot(litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# make violin plots
ggplot(litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



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

> **Answer:** The ranges and nominal values of the groups differ greatly, so when they are put side by side, you can't even see the min/max/IQR of some of them (they are very low and very close together). The minute differences in the groups with very small ranges are a bit more visible in the boxplots than the violin plots. For the groups with very large ranges, it seems that there aren't very many observations and/or they are very evenly distributed across the "bins", because there are no bulges (clusters) in the violin plots at all. In essence, the boxplot only uses 4 "bins" (min-25Q-med-75Q-max), which makes it easier to see how the points are distributed throughout the ranges. Additionally, nearly all the groups have outliers, and these only appear in the boxplots (while in the violin plots, they are just represented by a long line and you can't really see how many outliers there are).

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

> **Answer:** Needles have the highest biomass by far, regardless of what metric you look at. Their entire distribution is higher tham all the other litter types, including their median, upper and lower quartiles, and maximum value.