Assignment 3: Data Exploration

Nagarajan Vaidya Subramanian

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.

```
setwd("X:/ENV 872 Environmental Data Analytics/Git_codes/EDA-Spring2023/Assignments") getwd()
```

```
## [1] "X:/ENV 872 Environmental Data Analytics/Git_codes/EDA-Spring2023/Assignments"
```

```r
# install.packages('tidyverse') #command to install the two packages which
# needs to be run only once.
install.packages('lubridate') library(tidyverse)
library(lubridate)

# reading in the data sets for litter and woody debris and saving them to
# variables called neonics and litter
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)
str(neonics)  #checking the data structure to ensure that the strings have
been read in as factors

## 'data.frame':    4623 obs. of  30 variables:
##  $ CAS.Number                      : int  58842209 58842209 58842209
58842209 58842209 58842209 58842209 58842209 58842209 58842209 ...
##  $ Chemical.Name                   : Factor w/ 9 levels "(1E)-N-[(6-
Chloro-3-pyridinyl)methyl]-N'-cyano-N-methylethanimidamide",..: 9 9 9 9 9 9 9
9 9 9 ...
##  $ Chemical.Grade                  : Factor w/ 9 levels "Analytical
grade",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ Chemical.Analysis.Method        : Factor w/ 5 levels "Measured","Not
coded",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Chemical.Purity                 : Factor w/ 80 levels
">=98",">=99.0",..: 69 69 50 50 50 50 50 50 50 50 ...
##  $ Species.Scientific.Name         : Factor w/ 398 levels "Acalolepta
vastator",..: 69 69 248 248 248 248 248 248 248 248 ...
##  $ Species.Common.Name             : Factor w/ 303 levels
"Alfalfa Leafcutter Bee",..: 74 74 142 142 142 142 142 142 142 142
... ##  $ Species.Group                   : Factor w/ 4 levels
"Insects/Spiders",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Organism.Lifestage              : Factor w/ 20 levels
"Adult","Cocoon",..: 1 1 19 19 19 1 19 1 1 19 ...
##  $ Organism.Age                    : Factor w/ 39 levels
"~10","~24","~7",..: 39 39 39 39 39 36 39 36 36 39 ...
##  $ Organism.Age.Units              : Factor w/ 11 levels "Day(s)","Days
post-emergence",..: 9 9 4 4 4 1 4 1 1 4 ...
##  $ Exposure.Type                   : Factor w/ 24
levels "Choice","Dermal",..: 23 23 11 11 11 11 11 11 11 11
... ##  $ Media.Type                      : Factor w/ 10
levels "Agar","Artificial soil",..: 7 7 3 3 3 3 3 3 3 3
...
##  $ Test.Location                   : Factor w/ 4 levels "Field
artificial",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Number.of.Doses                 : Factor w/ 30 levels "' 4-5","'
47",..: 30 30 18 18 18 18 18 18 18 18 ...
##  $ Conc.1.Type..Author.            : Factor w/ 3 levels "Active
ingredient",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Conc.1..Author.                 : Factor w/ 1006 levels
```

```
"~10","~30/","~40/",..: 639 510 813 622 442 637 500 642 814 784 ...
```

```
##  $ Conc.1.Units..Author.           : Factor w/ 148 levels "%","% v/v","% w/v",..: 132 132 91 91 91 91 91 91 91 91 ...
##  $ Effect                          : Factor w/ 19 levels "Accumulation",..: 16 16 16 16 16 16 16 16 16 16 ...
##  $ Effect.Measurement              : Factor w/ 155 levels "Abundance","Accuracy of learned task, performance",..: 87 87 87 87 87 87 87 87 87 87 ...
##  $ Endpoint                        : Factor w/ 28 levels "EC10","EC50",..: 15 15 8 8 8 8 8 8 8 8 ...
##  $ Response.Site                   : Factor w/ 19 levels "Abdomen","Brain",..: 14 14 14 14 14 14 14 14 14 14 ...
##  $ Observed.Duration..Days.        : Factor w/ 361 levels "~.1458","~10",..: 145 145 145 145 145 145 145 145 145 145 ...
##  $ Observed.Duration.Units..Days.  : Factor w/ 17 levels "Day(s)","Day(s) post-emergence",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Author                          : Factor w/ 433 levels "Abbott,V.A., J.L. Nadeau, H.A. Higo, and M.L. Winston",..: 66 66 181 181 181 181 181 181 181 181 ...
##  $ Reference.Number                : int  107388 107388 103312 103312 103312 103312 103312 103312 103312 103312 ...
##  $ Title                           : Factor w/ 458 levels "A Common Pesticide Decreases Foraging Success and Survival in Honey Bees",..: 91 91 450 450 450 450 450 450 450 450 ...
##  $ Source                          : Factor w/ 456 levels "Acta Hortic.1094:451-456",..: 295 295 296 296 296 296 296 296 296 296 ... ##  $ Publication.Year                : int  1982 1982 1986 1986 1986 1986 1986 1986 1986 1986 ...
##  $ Summary.of.Additional.Parameters: Factor w/ 943 levels "Purity: \xca NR - NR | Organism Age: \xca NR - NR Not reported | Conc 1 (Author): \xca Formulation NR/ - NR/ % "| __truncated__,..: 797 796 795 794 860 859 858 857 864 871 ... str(litter)

## 'data.frame':    188 obs. of  19 variables:
##  $ uid                     : Factor w/ 188 levels "028eea3d-5c20-4afc-bb7ea05bab305152",..: 84 96 85 107 112 72 116 49 124 119 ... ##  $ namedLocation           : Factor w/ 12 levels "NIWO_040.basePlot.ltr",..: 8 8 8 8 8 8 8 8 11 11 ...
##  $ domainID                : Factor w/ 1 level "D13": 1 1 1 1 1 1 1 1 1 1 ...
##  $ siteID                  : Factor w/ 1 level "NIWO": 1 1 1 1 1 1 1 1 1 1 ...
##  $ plotID                  : Factor w/ 12 levels "NIWO_040","NIWO_041",..: 8 8 8 8 8 8 8 8 11 11 ...
##  $ trapID                  : Factor w/ 12 levels "NIWO_040_205",..: 8 8 8 8 8 8 8 8 11 11 ...
##  $ weighDate               : Factor w/ 2 levels "2018-08-06","2018-09-05": 1 1 1 1 1 1 1 1 1 1 ...
##  $ setDate                 : Factor w/ 2 levels "2018-07-05","2018-08-02":
```

```
1 1 1 1 1 1 1 1 1 1 ...
##  $ collectDate            : Factor w/ 2 levels "2018-08-02","2018-08-30":
1 1 1 1 1 1 1 1 1 1 ...
##  $ ovenStartDate          : Factor w/ 2 levels "2018-08-02T21:00Z",..: 1 1
1 1 1 1 1 1 1 1 ...
##  $ ovenEndDate            : Factor w/ 2 levels "2018-08-06T18:02Z",..: 1 1
1 1 1 1 1 1 1 1 ...
##  $ fieldSampleID          : Factor w/ 23 levels
"NEON.LTR.NIWO040205.20180802",..: 14 14 14 14 14 14 14 14 20 20 ...
##  $ massSampleID           : Factor w/ 168 levels
"NEON.LTR.NIWO040205.20180802.FLR",..: 102 101 103 97 103 99 100 98 139 145
...
##  $ samplingProtocolVersion: Factor w/ 1 level "NEON.DOC.001710vE": 1 1 1 1
1 1 1 1 1 1 ...
##  $ functionalGroup        : Factor w/ 8 levels "Flowers","Leaves",..: 7 6
8 1 8 4 5 2 1 8 ...
##  $ dryMass                : num  0.4 0.005 0.04 0.005 0.07 1 0.2 0.005
0.19 1.18 ...
##  $ qaDryMass              : Factor w/ 2 levels "N","Y": 1 1 2 1 1 1 1 1 1
2 ...
##  $ remarks                : logi  NA NA NA NA NA NA ... ##  $ measuredBy
: Factor w/ 2 levels
"kstyers@battelleecology.org",..: 1 1 1 1 1 1 1 1 1 1 ...
```

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 class of chemicals that are used as insecticides in farms and urban areas. These chemicals are absorbed by plants and get accumulated in their pollen and nectar. These insecticides are not targetted at a specific species; rather they act against a broad range of insects. Bees included. Further, they remain active for many years and can get washed away by rain or irrigation, polluting downstream water bodies.

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 comprised primarily of twigs, branches, leaves and other organic matter from plants. They play an important role in soil formation (by serving as detritus for organisms to feed on) and in the nutrien cycle in the ecosystem. when litter and woody debris is deposited on the ground, it essentially traps (sequesters) carbon from the air in the ground. Studying litter and woody debris gives us an estimate of the role that forests play in sequestering carbon at the scale of ecosystems. This can have a significant impact on the carbon cycle of the world as a whole. Further, the rate of deposition of debris can be affected by disease and pests, so studying that serves as an indicator of the overall health of the forest.

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. As part of the NEON network, litter and woody debris is sampled spatially and temporally.

2. Salient features:

a. In sites with forested tower airsheds, the litter sampling takes place in 20x 40m x 40m plots. In sites with low-statured vegetation over the tower airsheds, litter sampling is done in 4x 40m x 40m tower plots along with 26x 20m x 20m plots. This is done to accommodate co-located soil sampling

b. In places with deciduous vegetation, the sampling in elevated litter traps may be paused for the duration of winter

c. Therefore, The target sampling frequency for elevated traps varies by vegetation present at the site.

Obtain basic summaries of your data (Neonics)

5. What are the 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?

```
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 being studied are Population, Mortality, Behaviour, Feeding behaviour, Reproduction

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…]

```
head(sort(summary(neonics$Species.Common.Name), decreasing = TRUE), 7)
```

```
##               (Other)             Honey Bee        Parasitic Wasp
##                   670                   667                   285
## Buff Tailed Bumblebee   Carniolan Honey Bee            Bumble Bee
##                   183                   152                   140
##      Italian Honeybee
##                   113
```

Answer: Among the six species ofinsects most commonly studied, five of them are different species of bees. This focus on bees is likey because bees are one of the main vectors for natural pollination in ecosystems. When insecticides like Neonicotinoids are used, their adverse effect on bees can cascade through the whole ecosystem and affect all flora that depend on bees as pollination vectors.

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

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

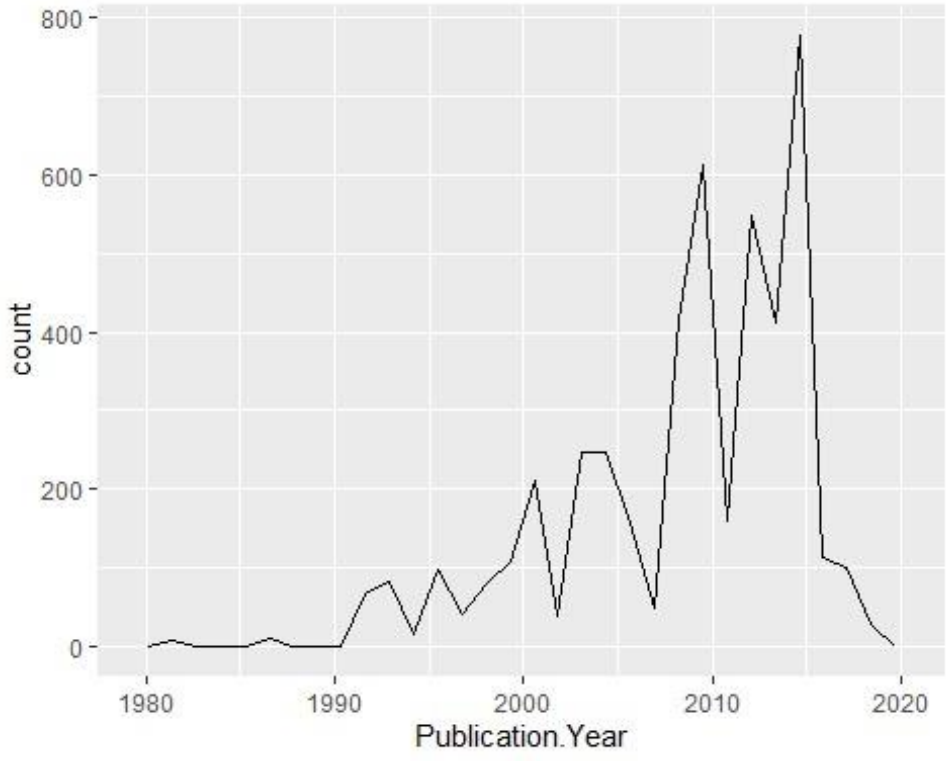
Answer: the class of the `Conc.1..Author.` column as seen by R is `class(neonics$Conc.1..Author.)`. I guess this is because it was read as a string by the read.csv function, and converted to factors by the subcommand stringsAsFactors.

Explore your data graphically (Neonics)

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

```
ggplot(data = neonics, mapping = aes(x = Publication.Year)) + geom_freqpoly()
```

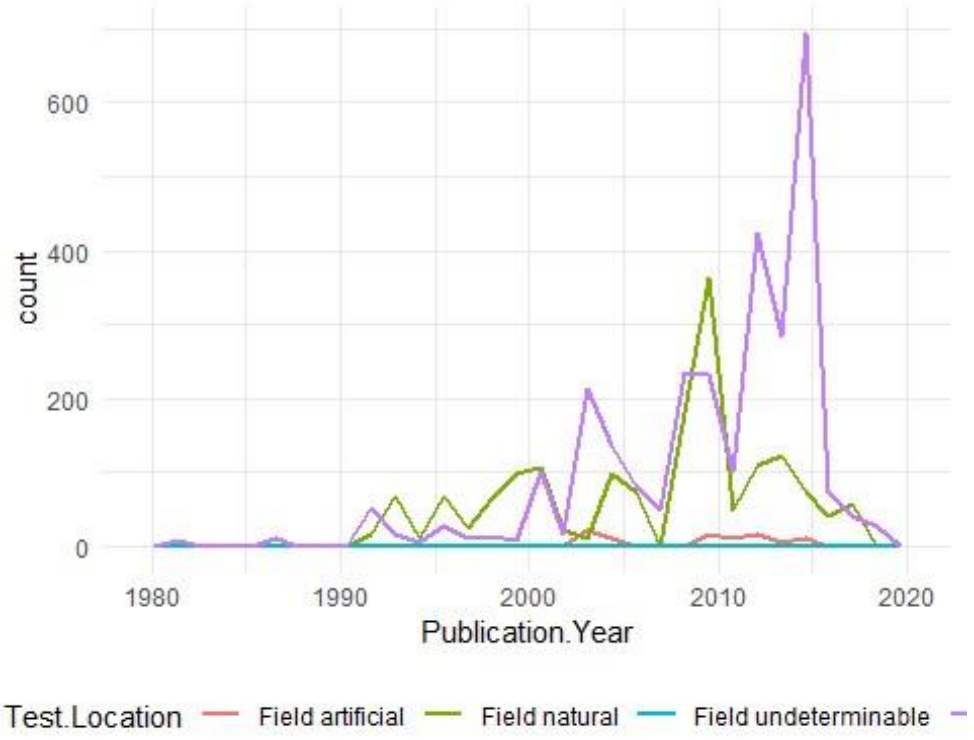
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
ggplot(data = neonics, mapping = aes(x = Publication.Year, colour =
Test.Location)) +
    geom_freqpoly(size = 1) + theme_minimal() + theme(legend.position =
"bottom")
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.

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



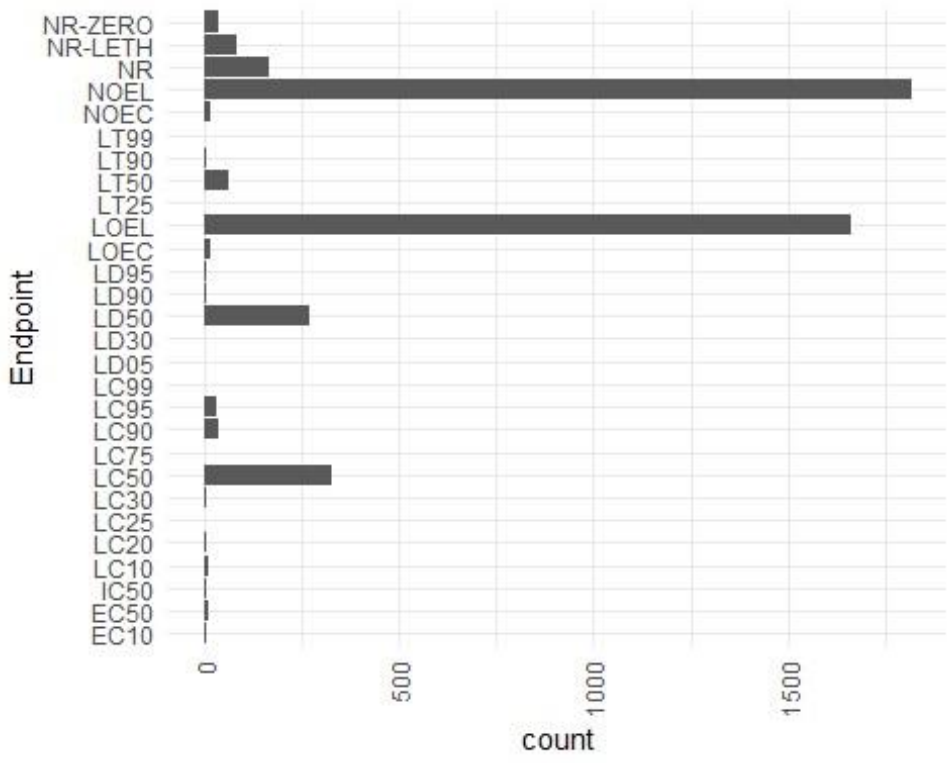
Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: The most common test location across the whole time range seems to be 'Lab' followed by 'Field natural'. This has varied over time, with 'Field natural' locations being more common between 1990-2000 and 'Lab' locations being more common from 2000-present.

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…]

```
ggplot(data = neonics, mapping = aes(y = Endpoint)) + geom_bar() +
theme_minimal() +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



Answer: The two most common end points in this dataset are "NOEL" and "LOEL". NOEL represents the no-observable-effect-level which is the highest dose that produces effects that are not significantly different from a control. LOEL represents lowest-observable-effect-level which is the lowest dose that produces a significantly different effect compared to a control group.

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.

```
class(litter$collectDate)

## [1] "factor"

litter$collectDate <- ymd(litter$collectDate)  #from opening the data
frame we see that the dates are in yyyy-mm-dd format. So I am using the
ymd() function in the lubridate library to convert the dates from their
current format (which is 'factor') to yyyy-mm-dd class(litter$collectDate)

## [1] "Date"

unique(litter$collectDate)

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

Litter was sampled on two dates: 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?

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

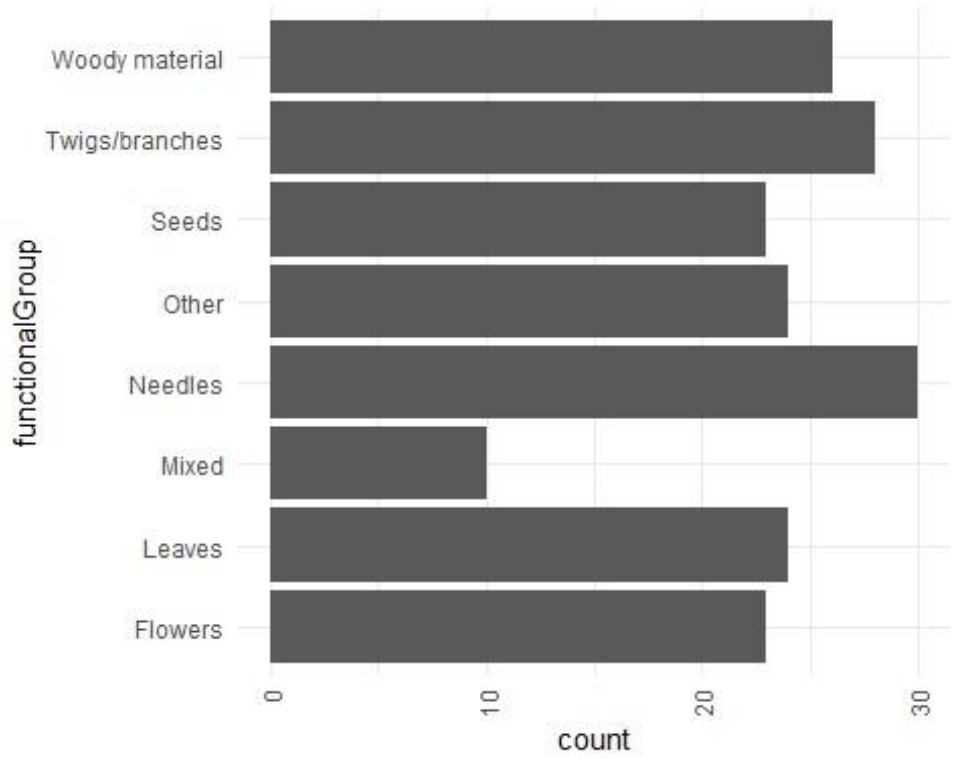
```
## NIWO_040.basePlot.ltr NIWO_041.basePlot.ltr NIWO_046.basePlot.ltr
##                    20                    19                    18
## NIWO_047.basePlot.ltr NIWO_051.basePlot.ltr NIWO_057.basePlot.ltr
##                    15                    14                     8
## NIWO_058.basePlot.ltr 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 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
```

Answer: the summary function shows the different plots that were sampled at Niwot Ridge along with the frequency of samples from each. Whereas the unique function returns only the different plots.

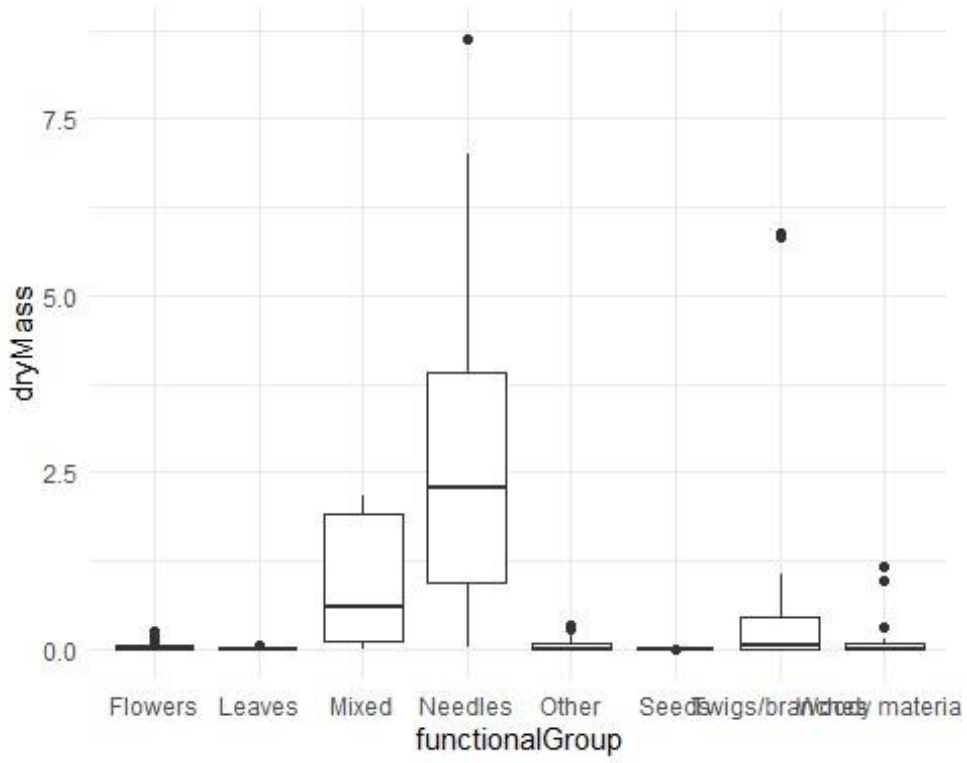
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(data = litter, mapping = aes(y = functionalGroup)) + geom_bar() +
theme_minimal() +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```

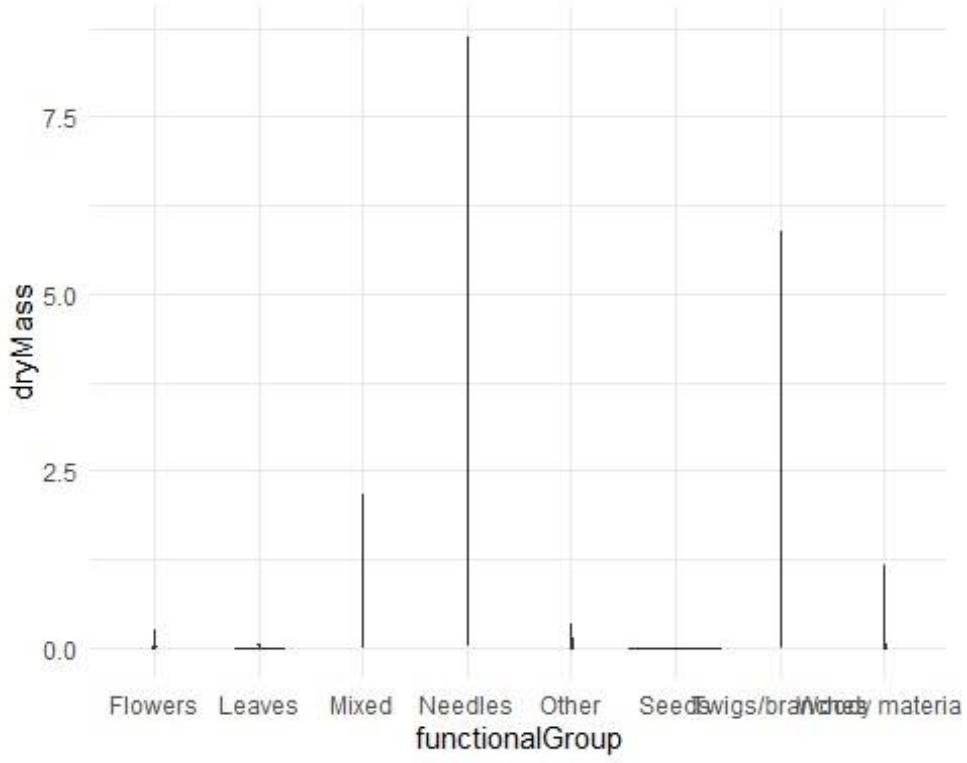


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(data = litter, mapping = aes(y = dryMass, x = functionalGroup)) +
geom_boxplot() +     theme_minimal()
```




```
ggplot(data = litter, mapping = aes(y = dryMass, x = functionalGroup)) +
geom_violin() +
    theme_minimal()
```



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

Answer: the boxplot provides more information in the same graphic that helps us understand the data better. Specifically, the boxes show the median and interquantile range, the whiskers show the spread of the data, and the points show outliers. Whereas the violin plot only shows the range of the data in each category which is hard to interpret without the context provided by the boxplot.

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

Answer: The highest biomass at these sites comes from needles, then 'mixed' followed by twigs and branches. This can be infered from the height of the whiskers in the boxplot.