Assignment 3: Data Exploration

Samantha_White-Murillo

Spring 2024

OVERVIEW

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

Directions

1. Rename this file `<Samantha_White-Murillo>_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.

```
install.packages("readr")
library(readr)

Litter <- read_csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv")

Neonics <- read_csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv")
```

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:Because they are a new class of insecticides that have a similar structure to nicotine, and as being new, it is necesary to identify their effects and efficiency on different type of incects.They have been linked to very enviromental effects as well.

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: Because they can be indicators of soil formation, its phisicochemical components and variations. Also, can generate information related to soil contamination and 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.There is a spatial sampling design that uses Tower plots metodology. To generate the trap deployment. 2. Field Collection data is denerated like ground trams are placed once a year and elevated traps varies by vegetation present at the site. 3. The dry mass values and chemistry subsampling metadata are scheduled to appear on the NEON data portal 45 days after data are collected for each activity.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) #4623 rows, 30 columns
```

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

```
summary(Neonics)
```

```
##    CAS Number        Chemical Name      Chemical Grade
##  Min.   : 58842209   Length:4623        Length:4623
##  1st Qu.:138261413   Class :character   Class :character
##  Median :138261413   Mode  :character   Mode  :character
##  Mean   :147651982
##  3rd Qu.:153719234
##  Max.   :210880925
##  Chemical Analysis Method Chemical Purity    Species Scientific Name
##  Length:4623              Length:4623        Length:4623
```

```
##  Class :character         Class :character   Class :character
##  Mode  :character         Mode  :character   Mode  :character
##
##
##
##  Species Common Name Species Group      Organism Lifestage Organism Age
##  Length:4623         Length:4623        Length:4623        Length:4623
##  Class :character    Class :character   Class :character   Class :character
##  Mode  :character    Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Organism Age Units Exposure Type       Media Type        Test Location
##  Length:4623        Length:4623        Length:4623        Length:4623
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Number of Doses    Conc 1 Type (Author) Conc 1 (Author)
##  Length:4623        Length:4623          Length:4623
##  Class :character   Class :character     Class :character
##  Mode  :character   Mode  :character     Mode  :character
##
##
##
##  Conc 1 Units (Author)    Effect          Effect Measurement   Endpoint
##  Length:4623           Length:4623        Length:4623        Length:4623
##  Class :character      Class :character   Class :character   Class :character
##  Mode  :character      Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Response Site      Observed Duration (Days) Observed Duration Units (Days)
##  Length:4623        Length:4623              Length:4623
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##     Author          Reference Number    Title              Source
##  Length:4623        Min.   :   344   Length:4623        Length:4623
##  Class :character   1st Qu.:108459   Class :character   Class :character
##  Mode  :character   Median :165559   Mode  :character   Mode  :character
##                     Mean   :142189
##                     3rd Qu.:168998
##                     Max.   :180410
##  Publication Year Summary of Additional Parameters
##  Min.   :1982     Length:4623
##  1st Qu.:2005     Class :character
##  Median :2010     Mode  :character
##  Mean   :2008
##  3rd Qu.:2013
##  Max.   :2019
```

```
#    CAS Number        Chemical Name      Chemical Grade     Chemical Analysis Method Chemical Purity
#  Min.   : 58842209   Length:4623        Length:4623        Length:4623              Length:4623
#  1st Qu.:138261413   Class :character   Class :character   Class :character         Class :character
#  Median :138261413   Mode  :character   Mode  :character   Mode  :character         Mode  :character
#  Mean   :147651982
#  3rd Qu.:153719234
#  Max.   :210880925
```

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?

```
mode_value <- as.character(names(Neonics$Effect)[which.max(Neonics$Effect)])
```

```
## Warning in which.max(Neonics$Effect): NAs introduced by coercion
```

```
print(paste("Mode:", mode_value))
```

```
## [1] "Mode: "
```

```
# "Mode: Population"
```

Answer:Because it indicates the extention in terms of individuals that are being affected by the insecticide

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...]

```
categorical_summary <- function(column) {
  freq_table <- table(column)
  cat("Frequency Table:\n")
  print(freq_table)
  cat("\nTop 6 most common values:\n")
  print(names(sort(freq_table, decreasing = TRUE)[1:6]))
}

categorical_summary(Neonics$`Species Common Name`)

#"Honey Bee" "Parasitic Wasp" "Buff Tailed Bumblebee" "Carniolan Honey Bee" "Bumble Bee" "Italian Honey
```

Answer:This species paly an important roll in the of plants, also some of them can be a tool for other species poblational regulation.

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?

```r
summary(Neonics$`Conc 1 (Author)`)
```

```
##    Length     Class      Mode
##      4623 character character
```

```r
# Class: character
```

Answer: Because sometimes when importing data, the progrma may sometimes misinterpret the data types. Also the missing data can be interpretated as character for the entire column.

Explore your data graphically (Neonics)

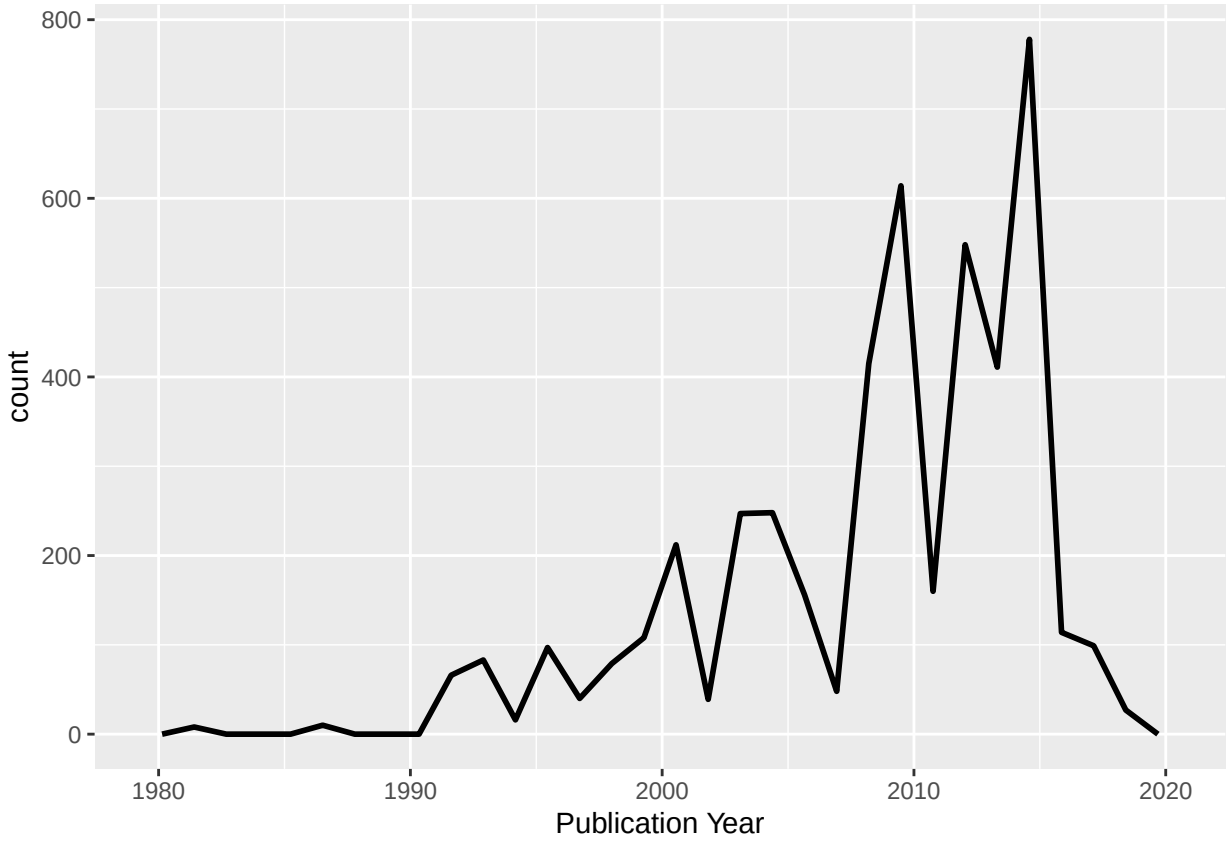
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```r
library(ggplot2)

ggplot(Neonics) +
  geom_freqpoly(aes(x = `Publication Year`), size = 1)
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

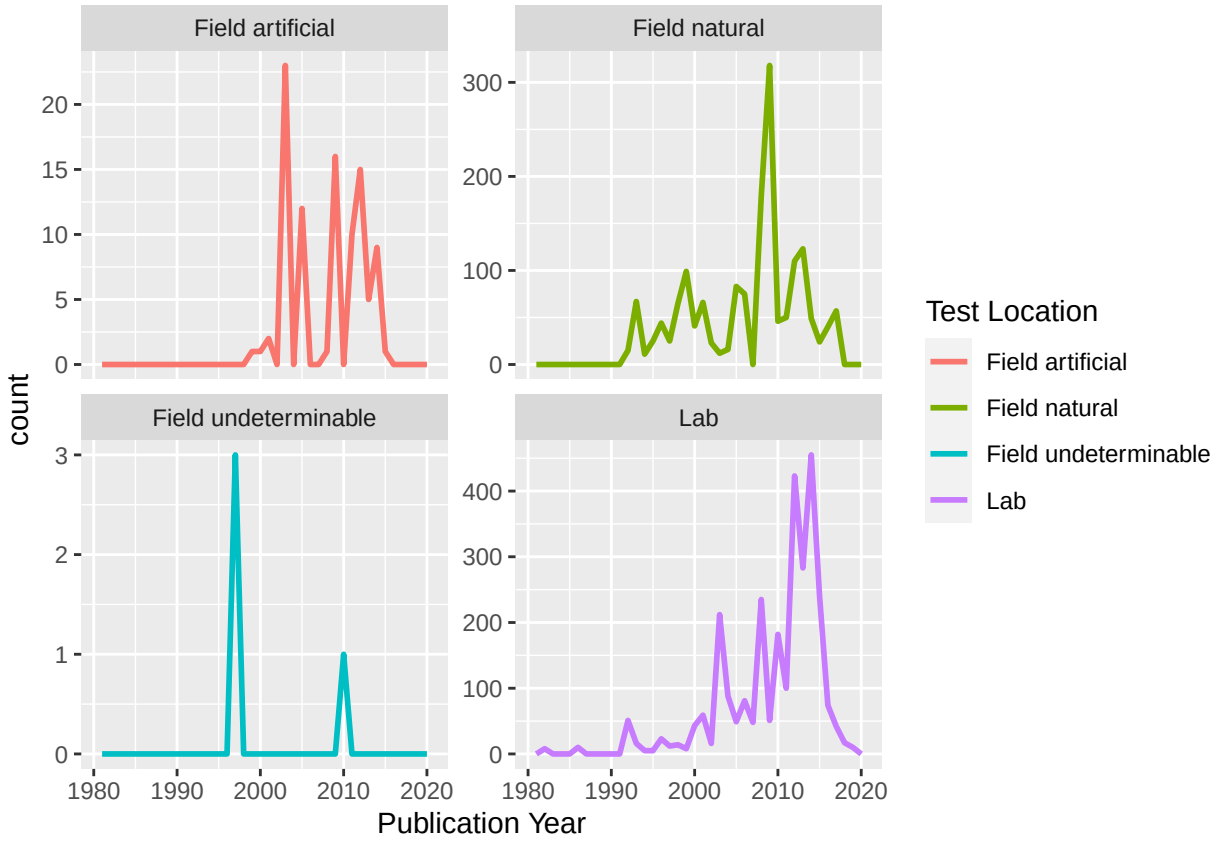
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



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

```
library(ggplot2)

ggplot(Neonics, aes(x = `Publication Year`, color = `Test Location`)) +
  geom_freqpoly(binwidth = 1, size = 1) +
  facet_wrap(~`Test Location`, scales = "free_y")
```



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

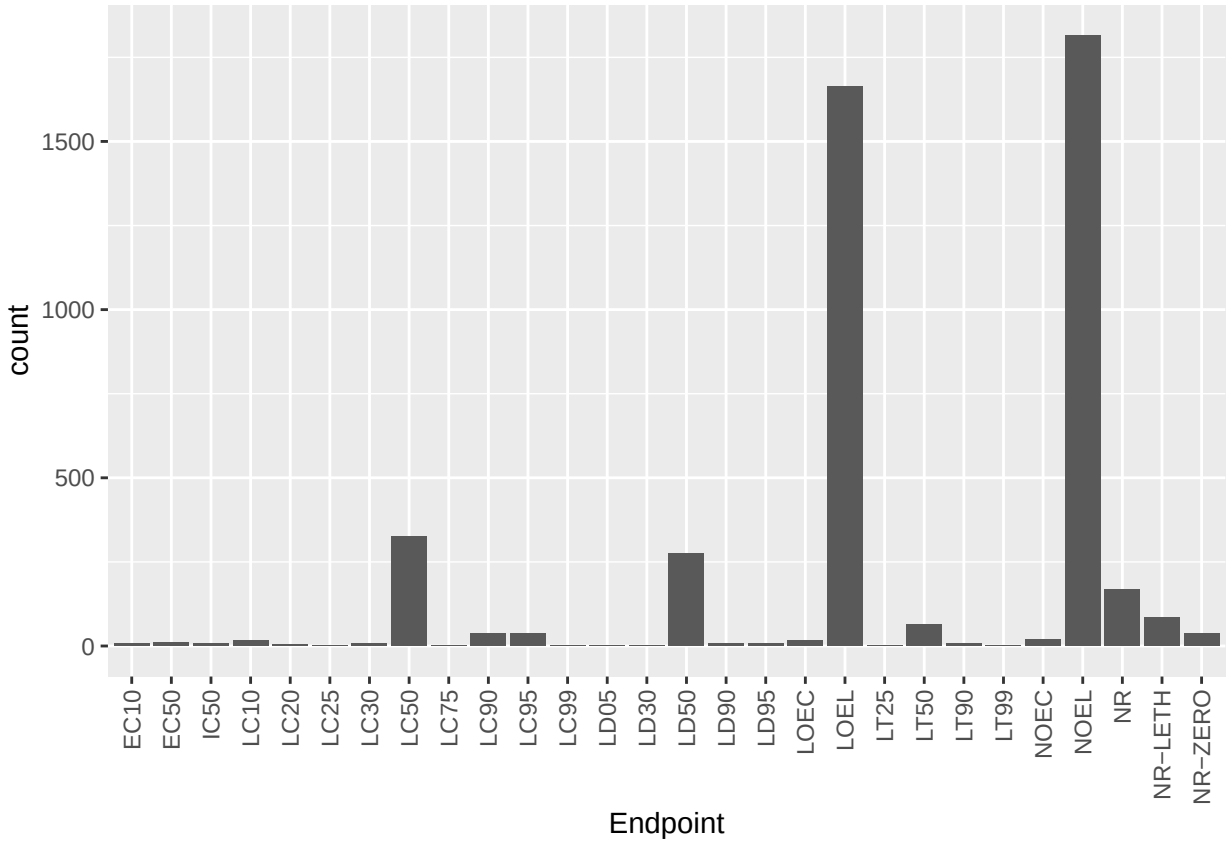
> Answer: the most common test locations are the lab and they have been taken mostly during 2008 and 2017.

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…]

```
library(ggplot2)

ggplot(Neonics) +
  geom_bar(aes(x = Endpoint)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: the two most common end points are NOEL abd LOEl. They are defined as NOEL: No-observable-effect-level, and LOEL: Lowest-observable-effect-leve.

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)
# It is a Date

unique(Litter$collectDate)
#"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`?

```
unique(Litter)
```

```
## # A tibble: 188 x 19
##    uid          namedLocation domainID siteID plotID trapID weighDate  setDate
##    <chr>        <chr>         <chr>    <chr>  <chr>  <chr>  <date>     <date>
##  1 7f065fec-b~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-06 2018-07-05
```

```
##  2 88df210b-1~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-06 2018-07-05
##  3 7f3c549c-1~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-06 2018-07-05
##  4 97806ab5-4~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-06 2018-07-05
##  5 9d7c89f5-8~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-06 2018-07-05
##  6 6ca7a3e8-4~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-06 2018-07-05
##  7 a0f02718-2~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-06 2018-07-05
##  8 500eb7f8-1~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-06 2018-07-05
##  9 aa0ce5fb-6~ NIWO_064.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-06 2018-07-05
## 10 a588a308-b~ NIWO_064.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-06 2018-07-05
## # i 178 more rows
## # i 11 more variables: collectDate <date>, ovenStartDate <dttm>,
## #   ovenEndDate <dttm>, fieldSampleID <chr>, massSampleID <chr>,
## #   samplingProtocolVersion <chr>, functionalGroup <chr>, dryMass <dbl>,
## #   qaDryMass <chr>, remarks <lgl>, measuredBy <chr>
```

```
# A tibble: 188 × 19
```

```
summary(Litter)
```

```
##      uid             namedLocation        domainID            siteID
##  Length:188         Length:188         Length:188         Length:188
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     plotID             trapID            weighDate
##  Length:188         Length:188         Min.   :2018-08-06
##  Class :character   Class :character   1st Qu.:2018-08-06
##  Mode  :character   Mode  :character   Median :2018-09-05
##                                        Mean   :2018-08-21
##                                        3rd Qu.:2018-09-05
##                                        Max.   :2018-09-05
##     setDate            collectDate         ovenStartDate
##  Min.   :2018-07-05   Min.   :2018-08-02   Min.   :2018-08-02 21:00:00.00
##  1st Qu.:2018-07-05   1st Qu.:2018-08-02   1st Qu.:2018-08-02 21:00:00.00
##  Median :2018-08-02   Median :2018-08-30   Median :2018-08-30 22:30:00.00
##  Mean   :2018-07-19   Mean   :2018-08-16   Mean   :2018-08-17 08:29:50.43
##  3rd Qu.:2018-08-02   3rd Qu.:2018-08-30   3rd Qu.:2018-08-30 22:30:00.00
##  Max.   :2018-08-02   Max.   :2018-08-30   Max.   :2018-08-30 22:30:00.00
##   ovenEndDate                     fieldSampleID      massSampleID
##  Min.   :2018-08-06 18:02:00.00   Length:188         Length:188
##  1st Qu.:2018-08-06 18:02:00.00   Class :character   Class :character
##  Median :2018-09-05 19:30:00.00   Mode  :character   Mode  :character
##  Mean   :2018-08-22 06:16:45.96
##  3rd Qu.:2018-09-05 19:30:00.00
##  Max.   :2018-09-05 19:30:00.00
##  samplingProtocolVersion functionalGroup       dryMass        qaDryMass
##  Length:188              Length:188         Min.   :0.0000   Length:188
##  Class :character        Class :character   1st Qu.:0.0000   Class :character
##  Mode  :character        Mode  :character   Median :0.0050   Mode  :character
##                                             Mean   :0.6115
##                                             3rd Qu.:0.3200
##                                             Max.   :8.6300
```

```
##  remarks         measuredBy
##  Mode:logical   Length:188
##  NA's:188       Class :character
##                 Mode  :character
##
##
##
```

```
#Length:188
```

Answer: In `unique` function the entire summary of the columns is not displayded (Use `print(n = ...)` to see more rows) vs. the summarey function. in addition the information is not stadisticaly shown in like in summary.

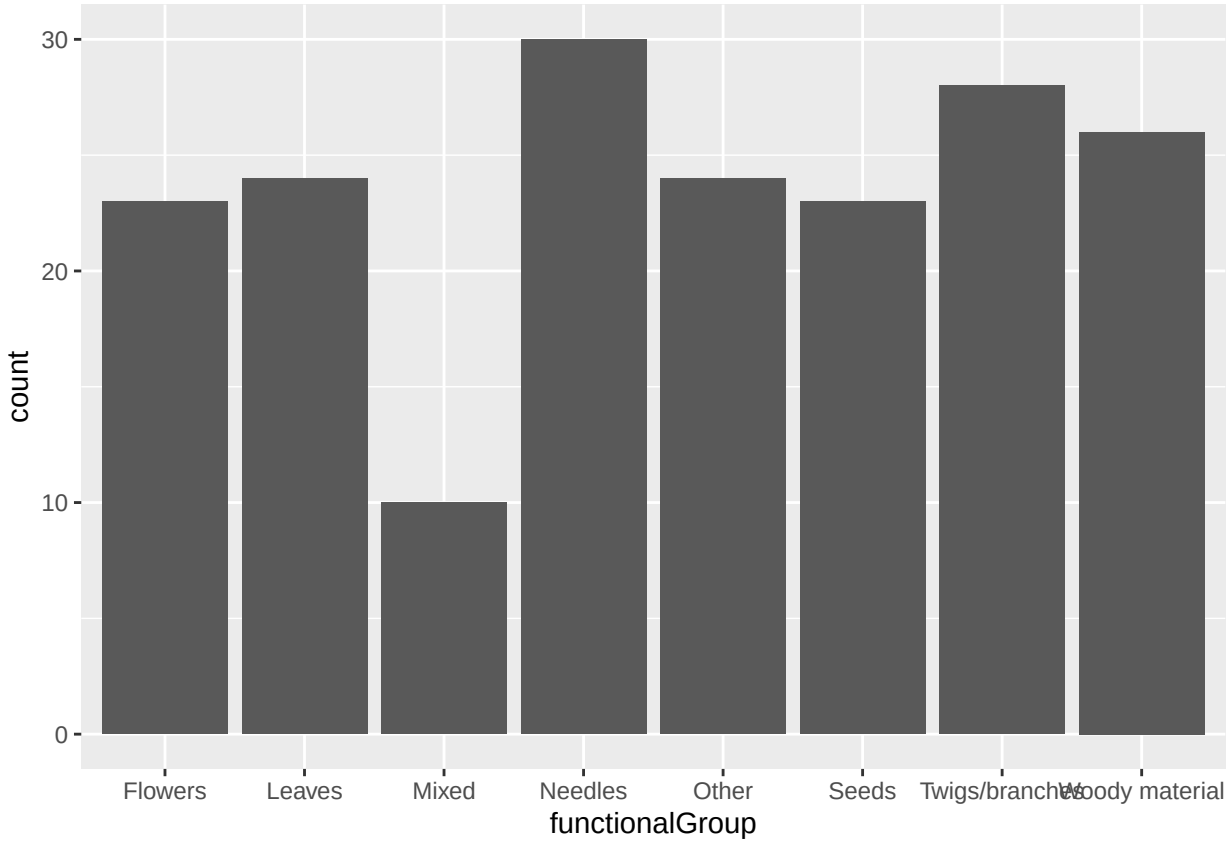
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
library(ggplot2)

geom_bar(aes(Litter$functionalGroup))
```

```
## mapping: x = ~Litter$functionalGroup
## geom_bar: just = 0.5, width = NULL, na.rm = FALSE, orientation = NA
## stat_count: width = NULL, na.rm = FALSE, orientation = NA
## position_stack
```

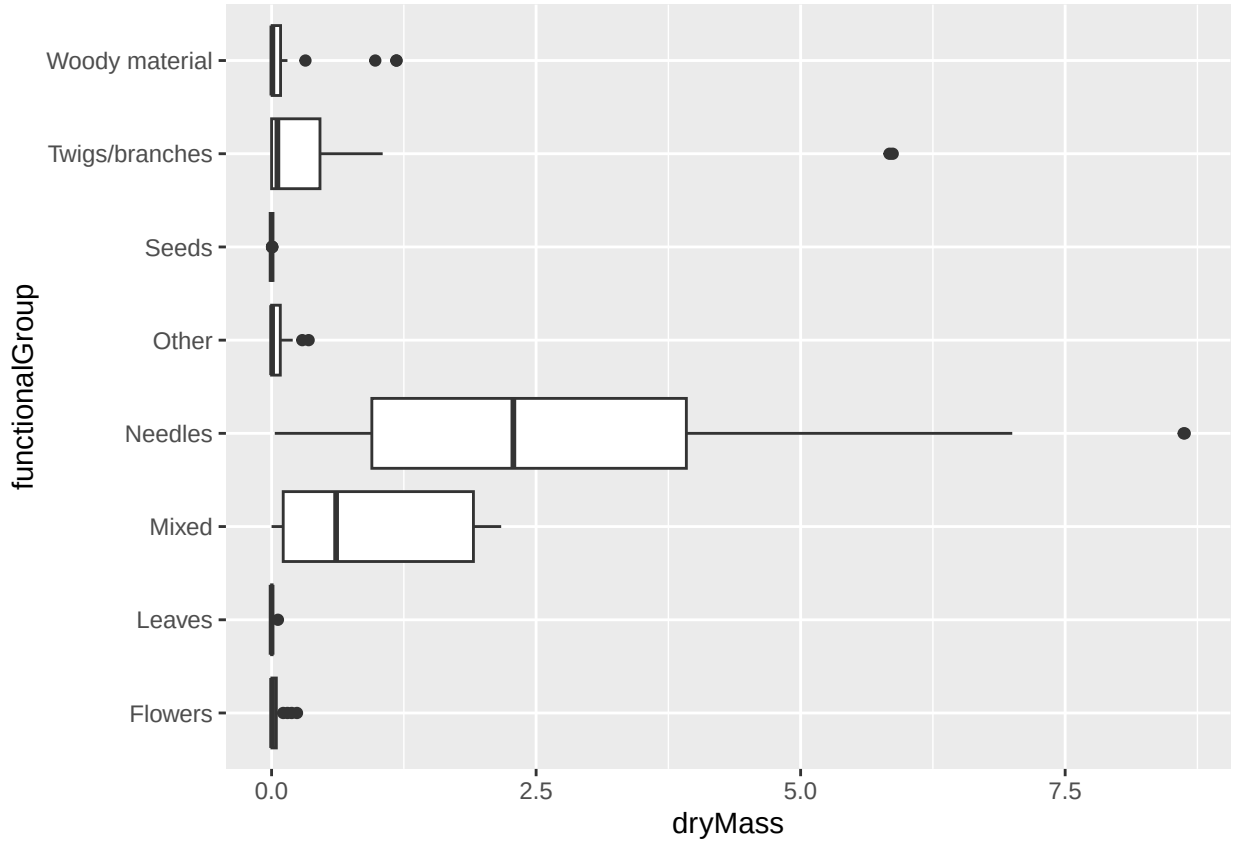
```
ggplot(Litter) +
  geom_bar(aes(x = functionalGroup))
```



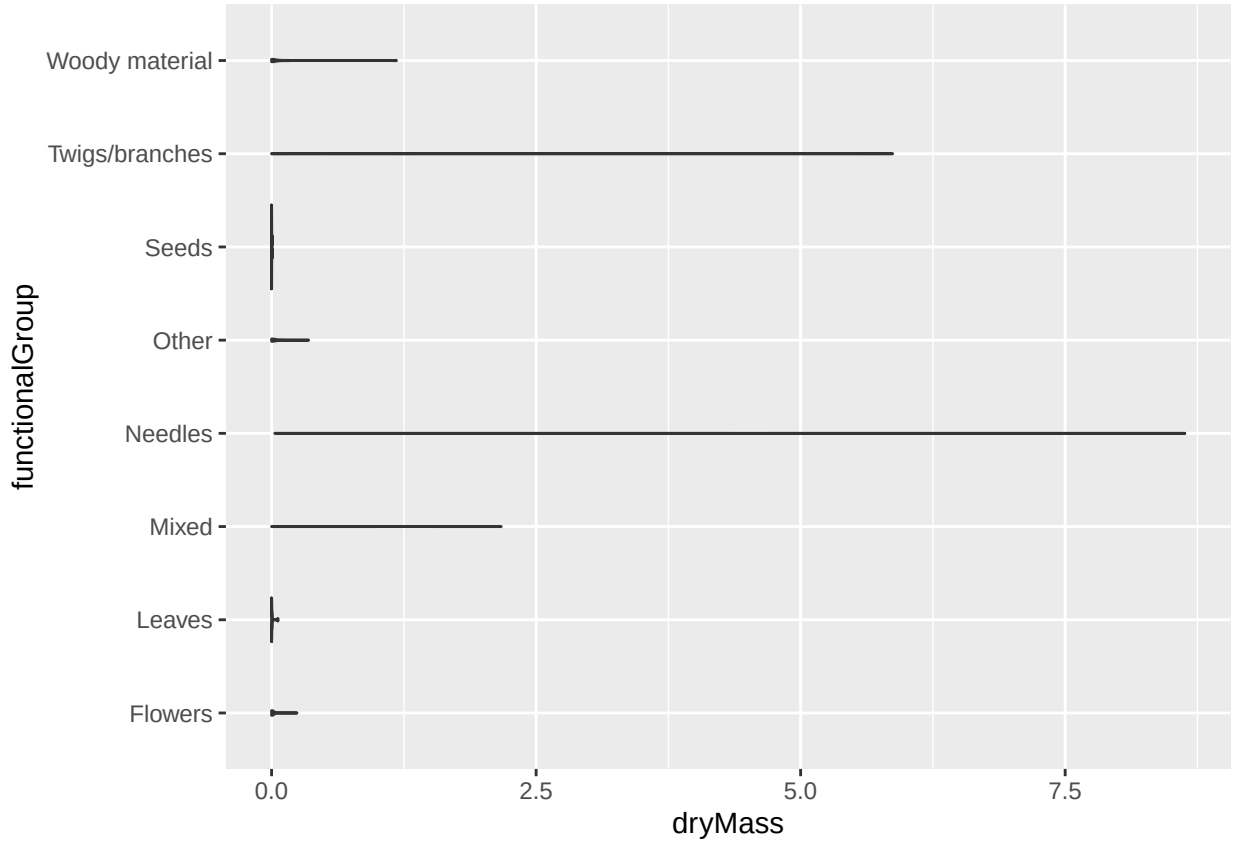
15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
library(ggplot2)

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




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



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

Answer:Because it shows stadistics of the data. Allowing a better interpretation of it.

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

Answer:The needles are the ones with more dry bymass followed by twigs/branches.