Assignment 3: Data Exploration

Yin-Chia Yang

Spring 2024

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.

```
getwd()
```

```
## [1] "/home/guest/EDA_Spring2024"
```

```
#setwd("/home/guest/EDA_Spring2024/EDA_forked_Spring2024")
library(tidyverse)
library(lubridate)
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 (also referred to as "neonics") are insecticides derived from nicotine. They act by binding strongly to nicotinic acetylcholine receptors in the central nervous system of insects, causing overstimulation of their nerve cells, paralysis and death. Recent researches have found transmission of neonicotinoids through tripartite food chains—plant to pest to natural enemy—combined with the diversity of nontarget herbivores on treated plants threatens entire food webs by disrupting arthropod communities and interactions. Thus understanding how it is effecting insects is helpful to understand the potential harm to the larger food web.

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:BWe might be interesyed in stuying woody debris because it is an important part of forest and stream ecosystems. Woody derbis plays a role in carbon budgets and nutrient cycling, serves as a source of energy for aquatic ecosystems, provides habitat for terrestrial and aquatic organisms, and contributes to structure and roughness, thereby influencing water flows and sediment transport.

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 and fine woody debris sampling is executed at terrestrial NEON sites that contain woody vegetation on >2m tall. 2.Locations of tower plots are selected randomly within the 90% flux footprint of the primary and secondary airsheds 3. plot edges must be separated by a distance 150% of one edge of the plot.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
colnames(Neonics)
```

```
##  [1] "CAS.Number"                    "Chemical.Name"
##  [3] "Chemical.Grade"                "Chemical.Analysis.Method"
##  [5] "Chemical.Purity"               "Species.Scientific.Name"
##  [7] "Species.Common.Name"           "Species.Group"
##  [9] "Organism.Lifestage"            "Organism.Age"
## [11] "Organism.Age.Units"            "Exposure.Type"
## [13] "Media.Type"                    "Test.Location"
## [15] "Number.of.Doses"               "Conc.1.Type..Author."
```

```
## [17] "Conc.1..Author."                 "Conc.1.Units..Author."
## [19] "Effect"                          "Effect.Measurement"
## [21] "Endpoint"                        "Response.Site"
## [23] "Observed.Duration..Days."        "Observed.Duration.Units..Days."
## [25] "Author"                          "Reference.Number"
## [27] "Title"                           "Source"
## [29] "Publication.Year"                "Summary.of.Additional.Parameters"
```

```
str(Neonics)
```

```
## 'data.frame':    4623 obs. of  30 variables:
##  $ CAS.Number                      : int  58842209 58842209 58842209 58842209 58842209 58842209 5884
##  $ Chemical.Name                   : Factor w/ 9 levels "(1E)-N-[(6-Chloro-3-pyridinyl)methyl]-N-ethy
##  $ 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
##  $ Chemical.Purity                 : Factor w/ 80 levels ">=98",">=99.0",..: 69 69 50 50 50 50 50 50
##  $ Species.Scientific.Name         : Factor w/ 398 levels "Acalolepta vastator",..: 69 69 248 248 248
##  $ Species.Common.Name             : Factor w/ 303 levels "Alfalfa Leafcutter Bee",..: 74 74 142 142
##  $ Species.Group                   : Factor w/ 4 levels "Insects/Spiders",..: 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
##  $ Organism.Age                    : Factor w/ 39 levels "<=24","<=48",..: 39 39 39 39 39 36 39 36 3
##  $ Organism.Age.Units              : Factor w/ 11 levels "Day(s)","Days post-emergence",..: 9 9 4 4 4
##  $ Exposure.Type                   : Factor w/ 24 levels "Choice","Dermal",..: 23 23 11 11 11 11 11
##  $ Media.Type                      : Factor w/ 10 levels "Agar","Artificial soil",..: 7 7 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","' 4-7",..: 30 30 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 "<0.0004","<0.025",..: 639 510 813 622 442
##  $ Conc.1.Units..Author.           : Factor w/ 148 levels "%","% v/v","% w/v",..: 132 132 91 91 91 91
##  $ Effect                          : Factor w/ 19 levels "Accumulation",..: 16 16 16 16 16 16 16 16
##  $ Effect.Measurement              : Factor w/ 155 levels "Abundance","Accuracy of learned task, peri
##  $ 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
##  $ Observed.Duration..Days.        : Factor w/ 361 levels "<.0002","<.0021",..: 145 145 145 145 145
##  $ Observed.Duration.Units..Days.  : Factor w/ 17 levels "Day(s)","Day(s) post-emergence",..: 1 1 1
##  $ Author                          : Factor w/ 433 levels "Abbott,V.A., J.L. Nadeau, H.A. Higo, and M
##  $ Reference.Number                : int  107388 107388 103312 103312 103312 103312 103312 103312 103
##  $ Title                           : Factor w/ 458 levels "A Common Pesticide Decreases Foraging Succ
##  $ Source                          : Factor w/ 456 levels "Acta Hortic.1094:451-456",..: 295 295 296
##  $ Publication.Year                : int  1982 1982 1986 1986 1986 1986 1986 1986 1986 1986 ...
##  $ Summary.of.Additional.Parameters: Factor w/ 943 levels "Purity: \xca NC - NC | Organism Age: \xca
```

```
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 two most common effects are 'Mortality' and 'Population'. These effects might be specifically of interest because they might be considered as more severe impacts.

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. . .]

```
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 Wooly 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: The six most commonly studied species are 'Honey Bee', 'Parasitic Wasp', 'Buff Tailed Bumblebee', 'Carniolan Honey Bee', 'Bumble Bee', and ' Italian Honeybee'. They are all from a Bee/Wasp family, and they are all pollinators, which might be the research insterest of how the neonicotinoids impacts the food web.

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"
```

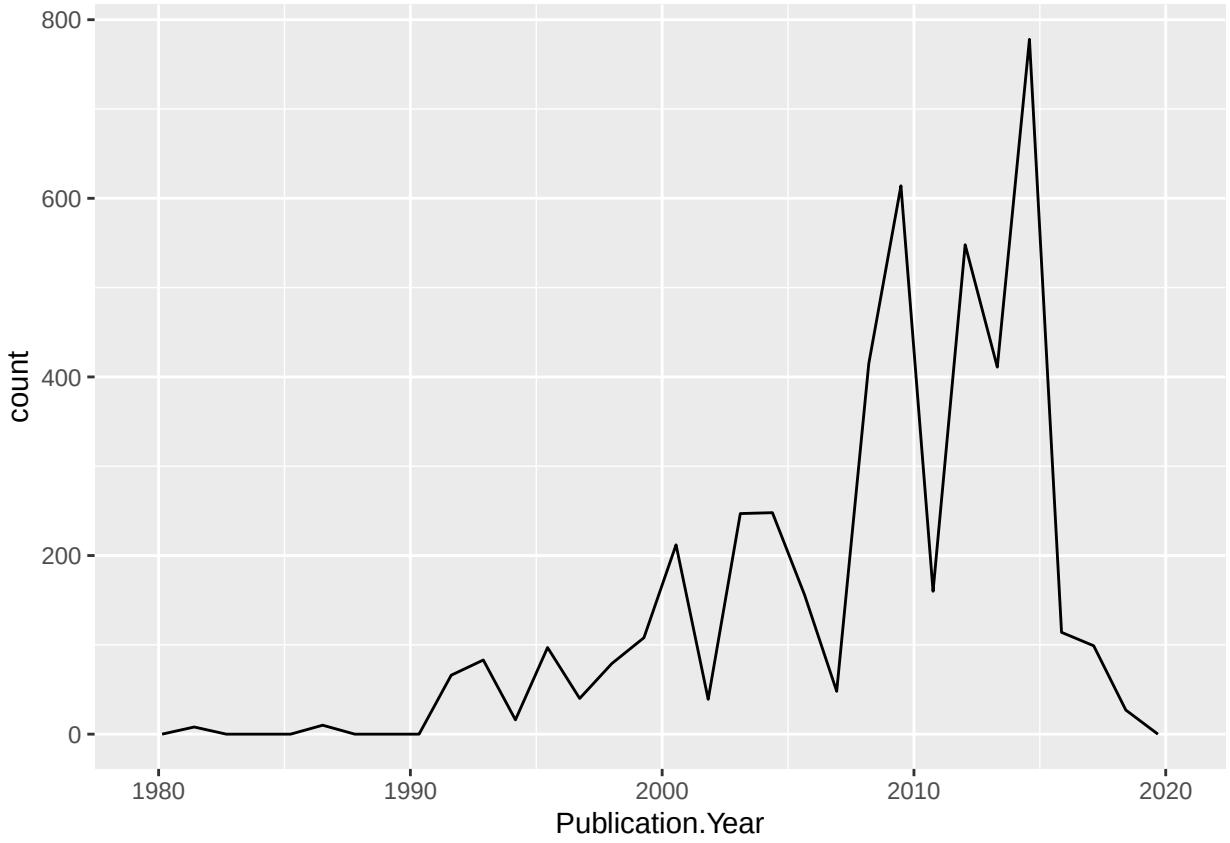
Answer: It is "character", because it was imported with a .csv file.

Explore your data graphically (Neonics)

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

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

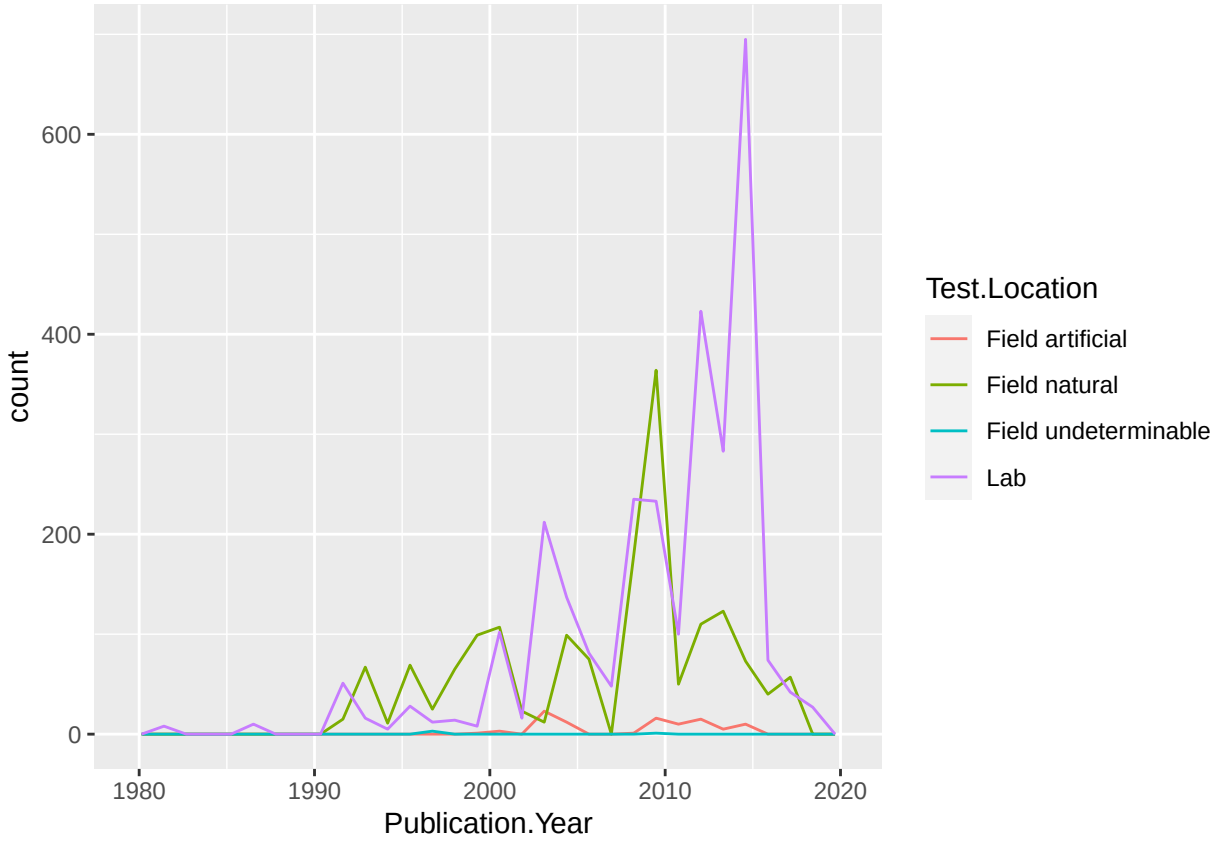
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
ggplot(Neonics, aes(x = Publication.Year, color = Test.Location)) +
geom_freqpoly()
```

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



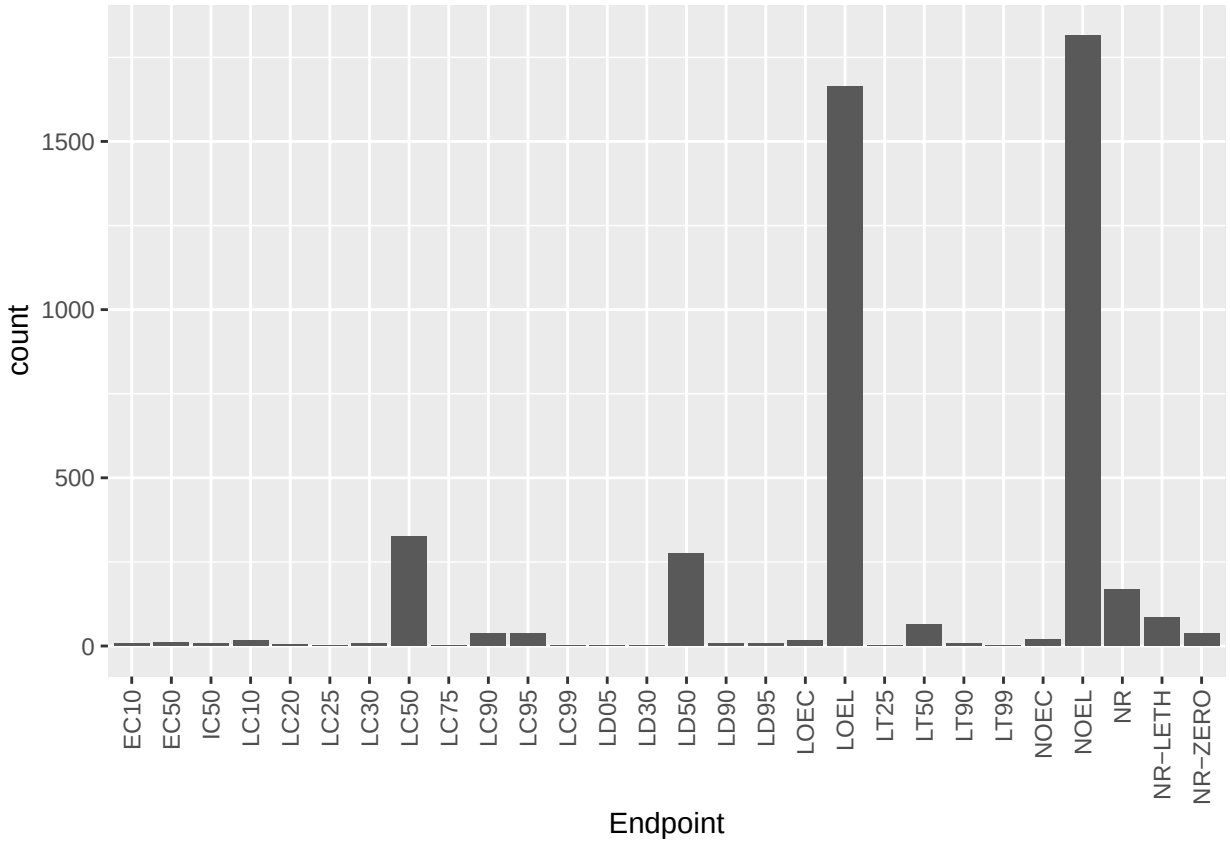
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer:The two most common test locations are 'Lab' and 'Feild natural'. Numbers of est conducted in 'Feild natural' exceeded test conducted in 'Lab' during 1992 to 2000, 2008 to 2011, and around the year of 2017. For the rest of the time period, 'Lab' remains the most common test location.

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



Answer:The two most common end points are 'LOEL' and 'NOEL'. When the author of the information identifies a value as the "highest tested concentration having no statistically significant adverse effect", the reviewer should code this as a NOEL/NOEC; the "lowest tested concentration having a statistically significant adverse effect" is coded as a LOEL/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.

```
class(Litter$collectDate)
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
class(Litter$collectDate)
```

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

```
unique(Litter$collectDate, incomparables = FALSE)
```

```
## [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`?

```
unique(Litter$plotID, incomparables = FALSE)
```

```
##  [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
```

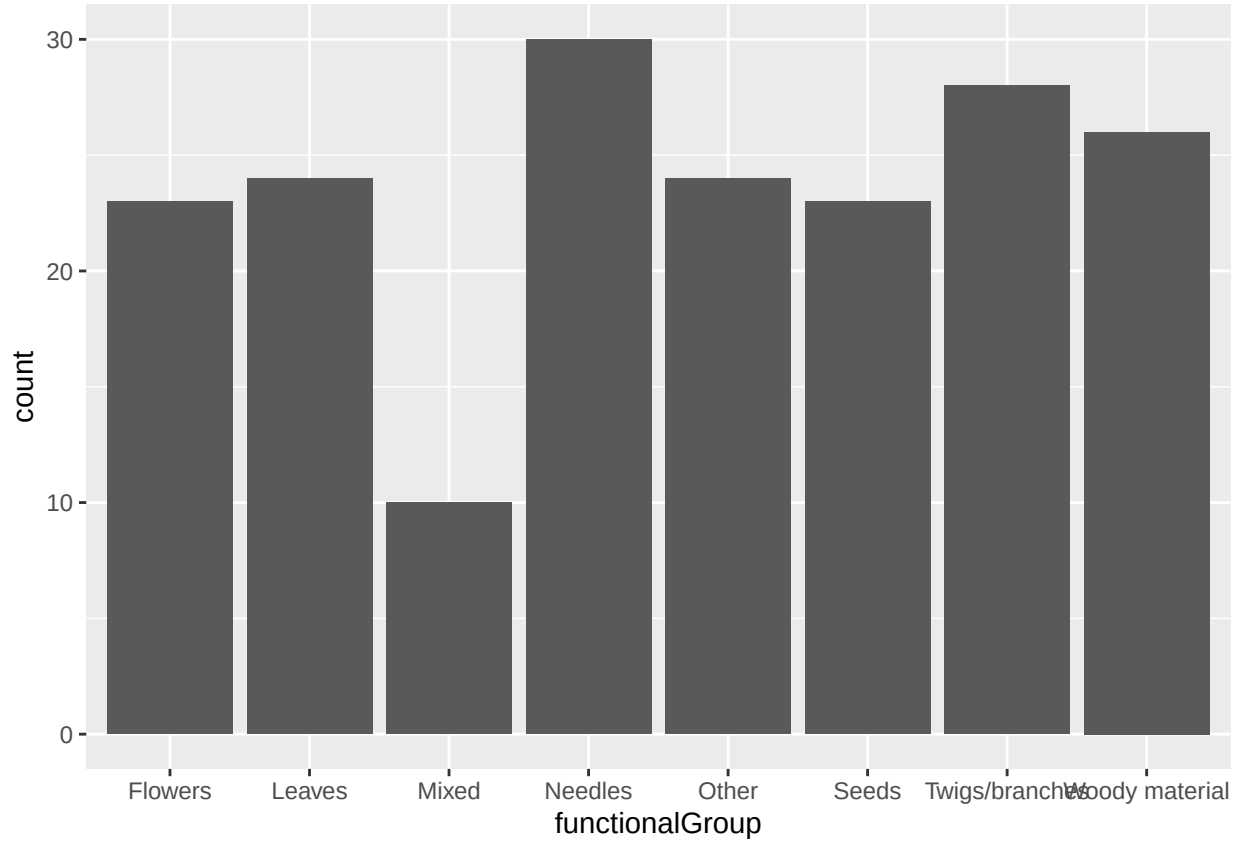
```
summary(Litter$plotID)
```

```
## 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 at Niwot Ridge (NIWO_40~67). The 'unique' function goes down the dataset and shows the plots that were sampled, in the order of the dataset, in a non-repeating way. The 'summary' function shows the count of each plot, and presents them in a numerical order.

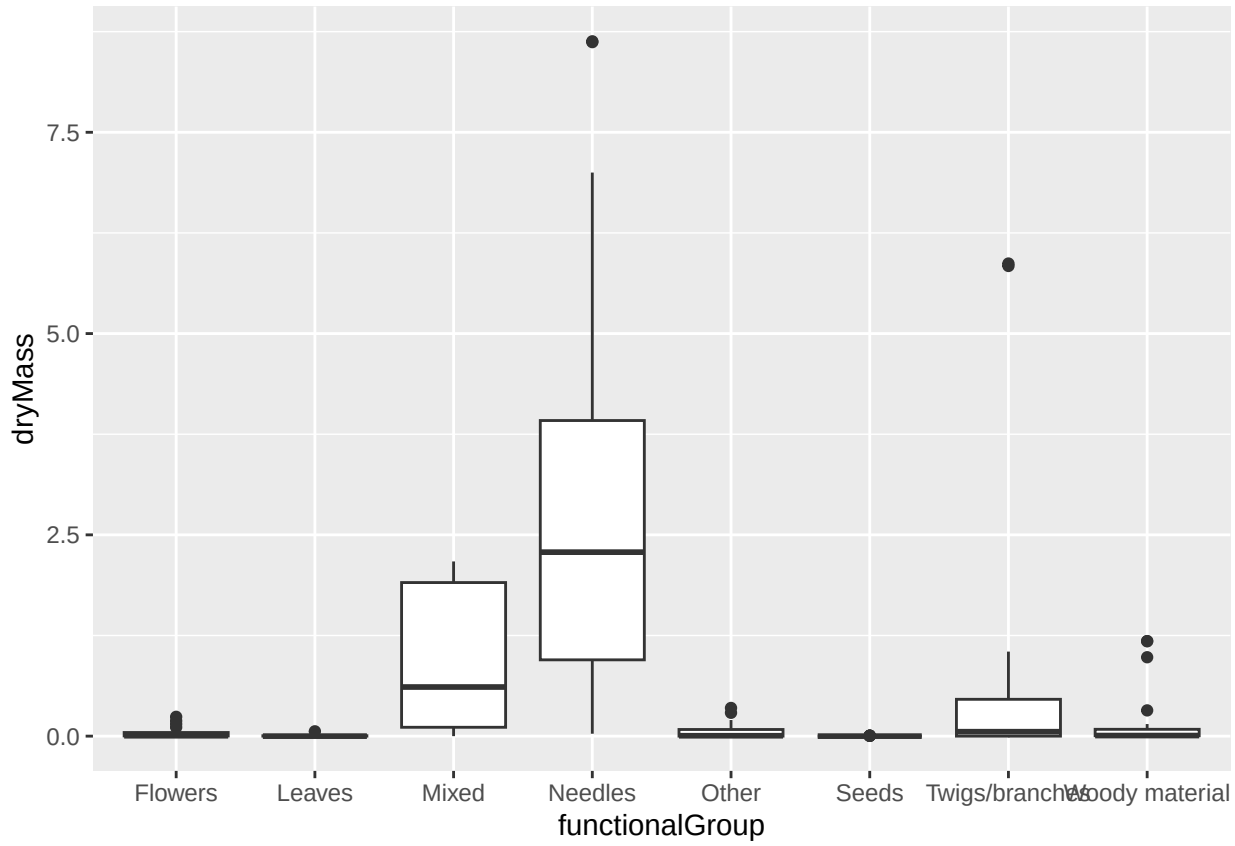
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

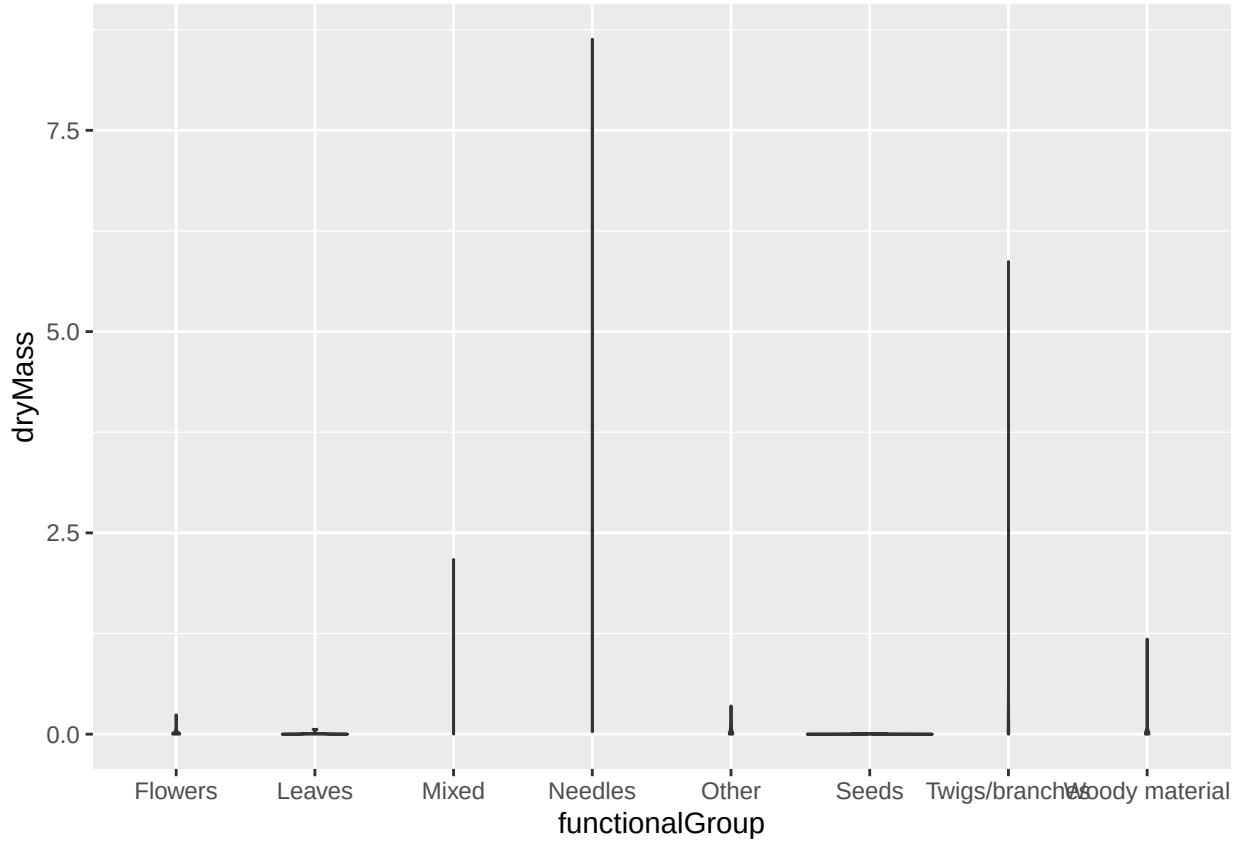


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




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



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

Answer:Yes, because it is able to demonstrate more information on the distribution of the data.

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

Answer: The 'Needles' tends to have the highest boimass.