# Assignment 3: Data Exploration

Wynona Curaming

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.

```
# 1. Set up your working directory
getwd()
```

```
## [1] "/home/guest/ENV 872/EDA_Spring2024"
```

```
# 2. Load packages
library(tidyverse, lubridate)

# 3. Import datasets
Neonics <- read.csv("~/ENV 872/EDA_Spring2024/Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",
```

```
                    stringsAsFactors=TRUE)
Litter <- read.csv("~/ENV 872/EDA_Spring2024/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 may affect the health of key insect pollinators like bees. Decline in bee populations will in turn have significant negative consequences on crop yields, which would then impact food security.

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 role in carbon budgets and nutrient cycling (especially nitrogen, which is accumulated as they decompose). They also contribute to forest biodiversity by creating habitat for some animals like litter beetles.

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. Sampling is executed at terrestrial NEON sites that contain woody vegetation >2m tall. 2. Trap placement within plots may be either targeted or randomized, depending on the vegetation 3. Over time some sampling plots may become impossible to sample, due to disturbance or other local changes

## 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"
```

```r
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 24
##  $ 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 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
##  $ 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
##  $ Conc.1..Author.                 : Factor w/ 1006 levels "<0.0004","<0.025",..: 639 510 813 622 44
##  $ Conc.1.Units..Author.           : Factor w/ 148 levels "%","% v/v","% w/v",..: 132 132 91 91 91 9
##  $ 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, per
##  $ 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
##  $ Reference.Number                : int  107388 107388 103312 103312 103312 103312 103312 103312 10
##  $ Title                           : Factor w/ 458 levels "A Common Pesticide Decreases Foraging Suc
##  $ 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
```

```r
dim(Neonics)
```

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

Answer: This dataset has 4623 observations and 30 variables.

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?

```r
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 5 most common effects studied in descending order are population, mortality, behavior, feeding behavior and reproduction. I think these would be of interest because we'd want to know how the chemical in this pesticide affects insect populations, such as whether there is an increase in mortality, and a difference to behavior (eg. feeding, 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…]

```r
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 studiesd species are the honey bee, parasitic wasp, buff tailed bumblebee, carniolan honey bee, bumble bee and italian honeybee. Bees are of interest over other insects because they are important pollinators, which are crucial for fruit formation in plants. On the other hand, the parasitic wasp is an issue to farmers because they lay their eggs in other insects. Once the larvae hatches, it kills the host insect.

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
class(Neonics$Conc.1..Author.)
```

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

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

```
##    0.37/      10/      NR/       NR        1     1023    0.40/       2/
##      208      127      108       94       82       80       69       63
##       10   0.053/      100      50/     0.5/     0.03    0.05/     0.45
##       62       59       56       51       45       44       43       43
##     0.1/    0.45/     1.0/    2.27/       50    0.125     500/      0.5
##       42       40       40       40       36       33       33       32
##   0.048/    0.15/       1/       48    25.0/      12/    0.027      2.4
##       30       30       30       30       28       27       26       26
##     0.2/    0.56/     100/        3    0.01/    1000/       3/    0.336
##       25       24       23       23       22       22       22       21
##     1.5/     0.05      1.5    2.60/    20.0/        6    6.80/    62.5/
##       21       20       20       20       20       20       20       20
##    0.005     0.4/    0.18/     0.3/     1000       40 0.00355/      0.1
##       18       18       17       17       17       17       16       16
##      0.4     150/      300      80/    0.053     0.24     0.28     125/
##       16       16       16       16       15       15       15       15
##        9   0.0001  0.0004/   0.084/     0.15      0.6    12.5/   144.0/
```

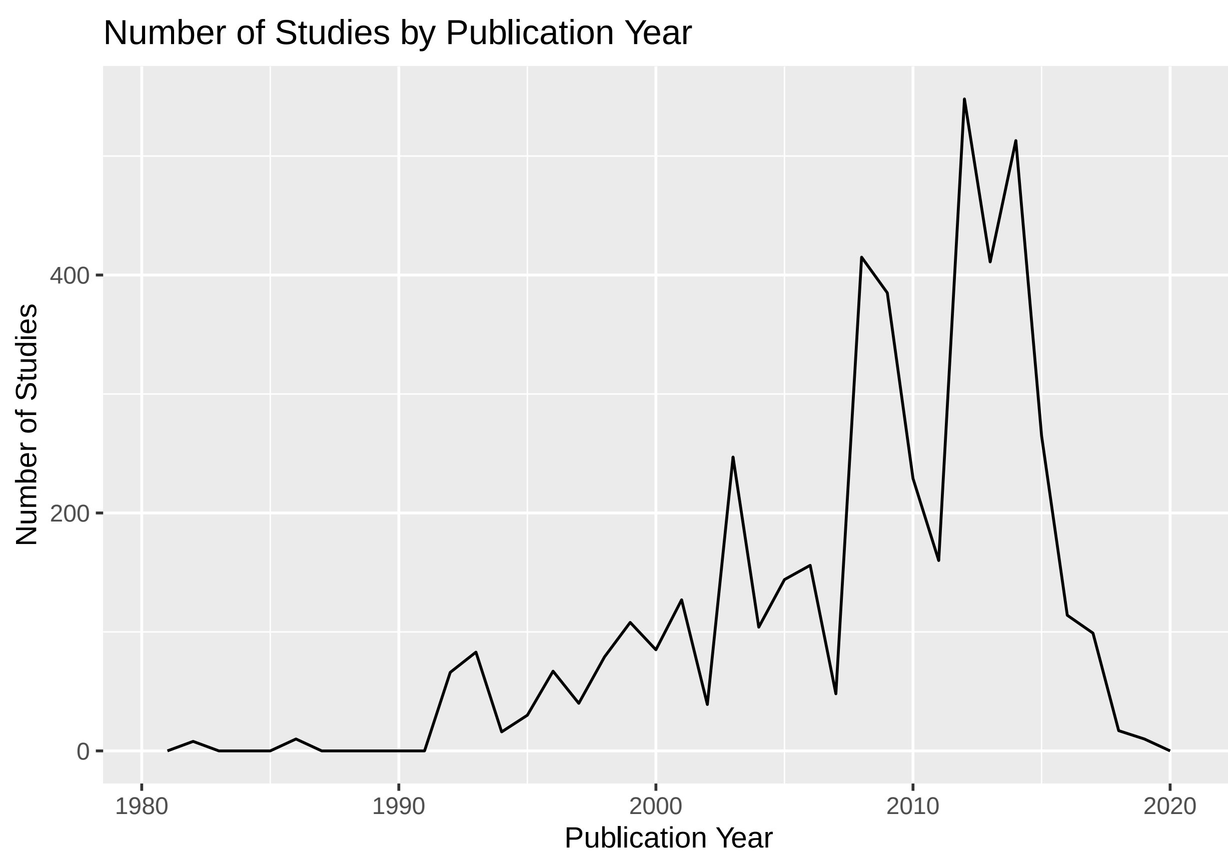
```
##        15        14        14        14        14        14        14        14
##      350/     40.0/       48/        56       84/     0.17/       125        14
##        14        14        14        14        14        13        13        13
##        16        17    0.047/     0.25/     0.28/     1.28/     1.81/       112
##        13        13        12        12        12        12        12        12
##       150      2.5/        25       60/       75/     0.02/    0.025/      0.29
##        12        12        12        12        12        11        11        11
##     37.5/        4/         5   (Other)
##        11        11        11      1817
```

> Answer: It is a factor. It is not numeric because some of the values have characters like "/" and "NR". We may need to do some data cleaning first, consulting domain-specific guidance. The NEON_Litterfall_UserGuide.pdf document doesn't seem to have information on this.

### Explore your data graphically (Neonics)

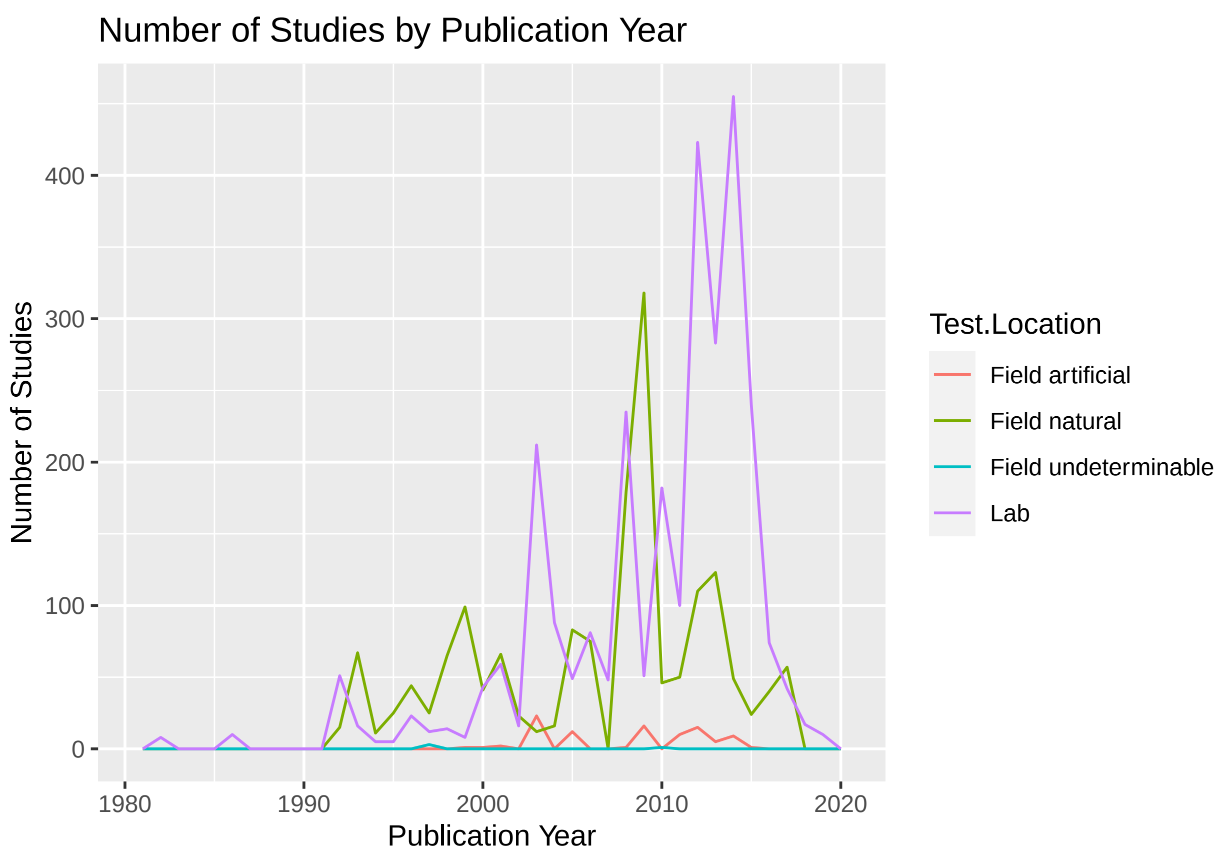
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```r
studies_by_pub_year<-ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year), binwidth=1) +
  labs(title = "Number of Studies by Publication Year",
       x = "Publication Year",
       y = "Number of Studies")
studies_by_pub_year
```



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

```
ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year, color=Test.Location), binwidth=1)+
  labs(title = "Number of Studies by Publication Year",
       x = "Publication Year",
       y = "Number of Studies")
```



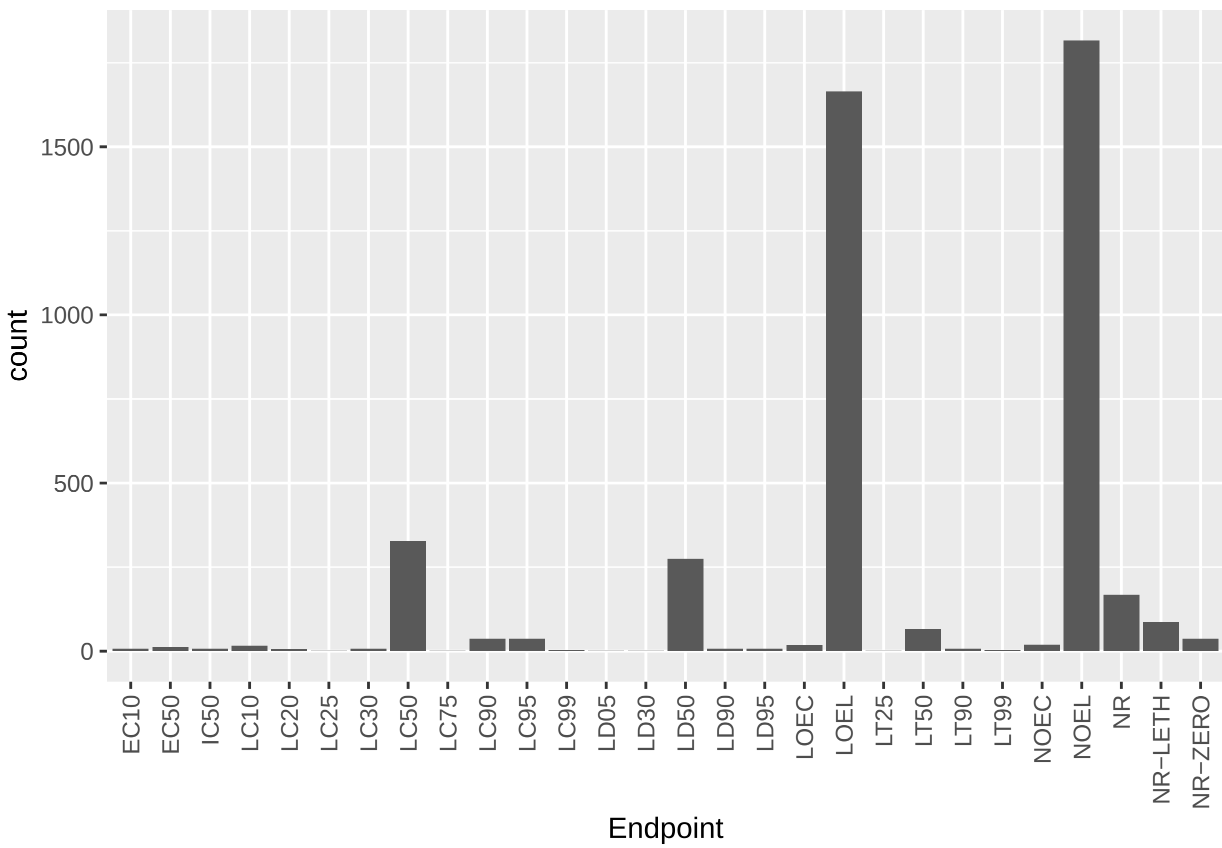

Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer:Overall, most common test location is on the lab, followed by the field (natural).From the mid-1990s to 2000 there were more studies based on the field (natural). From 2009 onwards (except for 2017), there were more lab studies than field studies. This probably shows a shift in disciplinary norms about what kind of test location is preferable.

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




> Answer: The two most common end points are NOEL and LOEL. NOEL refers to "No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls". LOEL refers to "Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls".

## 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.

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

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

```r
summary(Litter$collectDate)
```

```
## 2018-08-02 2018-08-30
##         91         97
```

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

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

```r
unique(dates_converted)
```

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

Answer: The litter was sampled on August 2, 2018 and August 30, 2018.

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
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
```

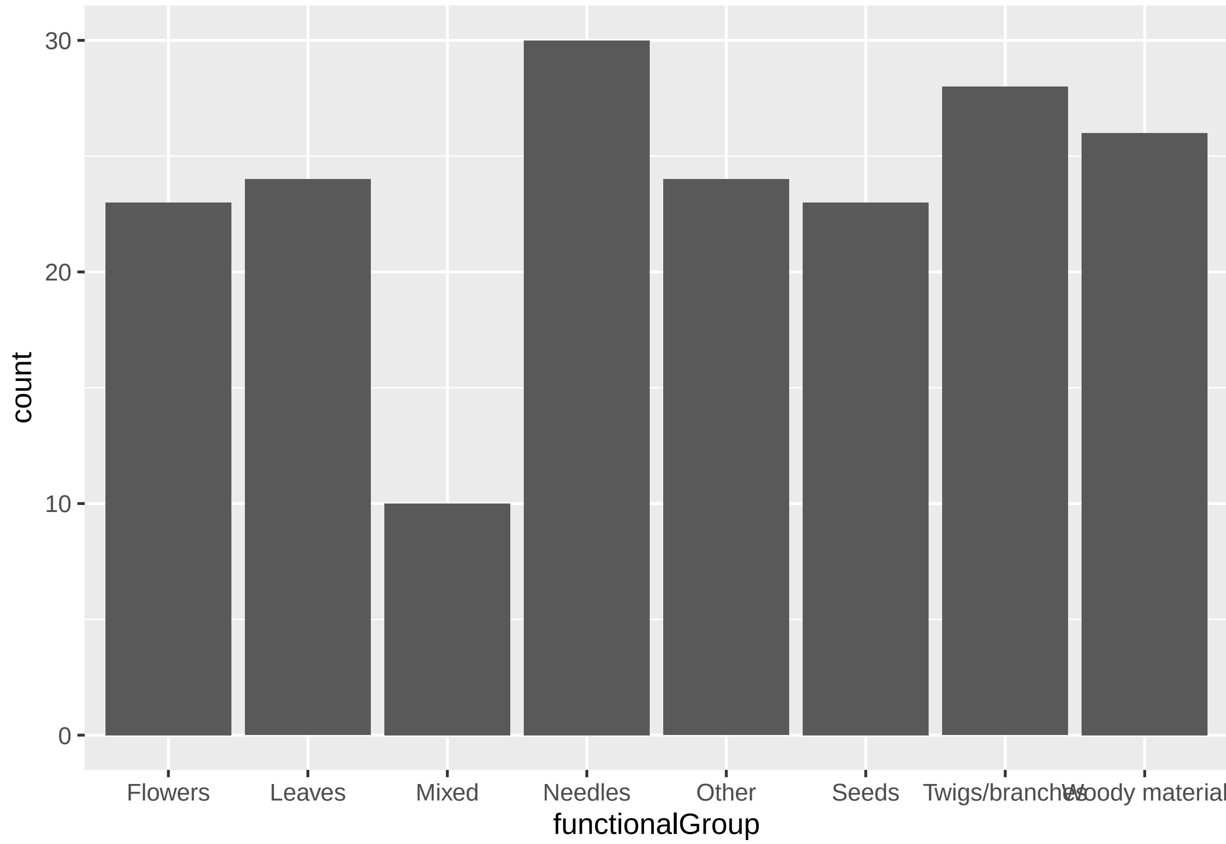
```r
unique(Litter$namedLocation)
```

```
##  [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: Summary shows the number of observations for each plot Unique doesn't show the number of observations for each plot but only lists out the types of plots and number of types of plots.

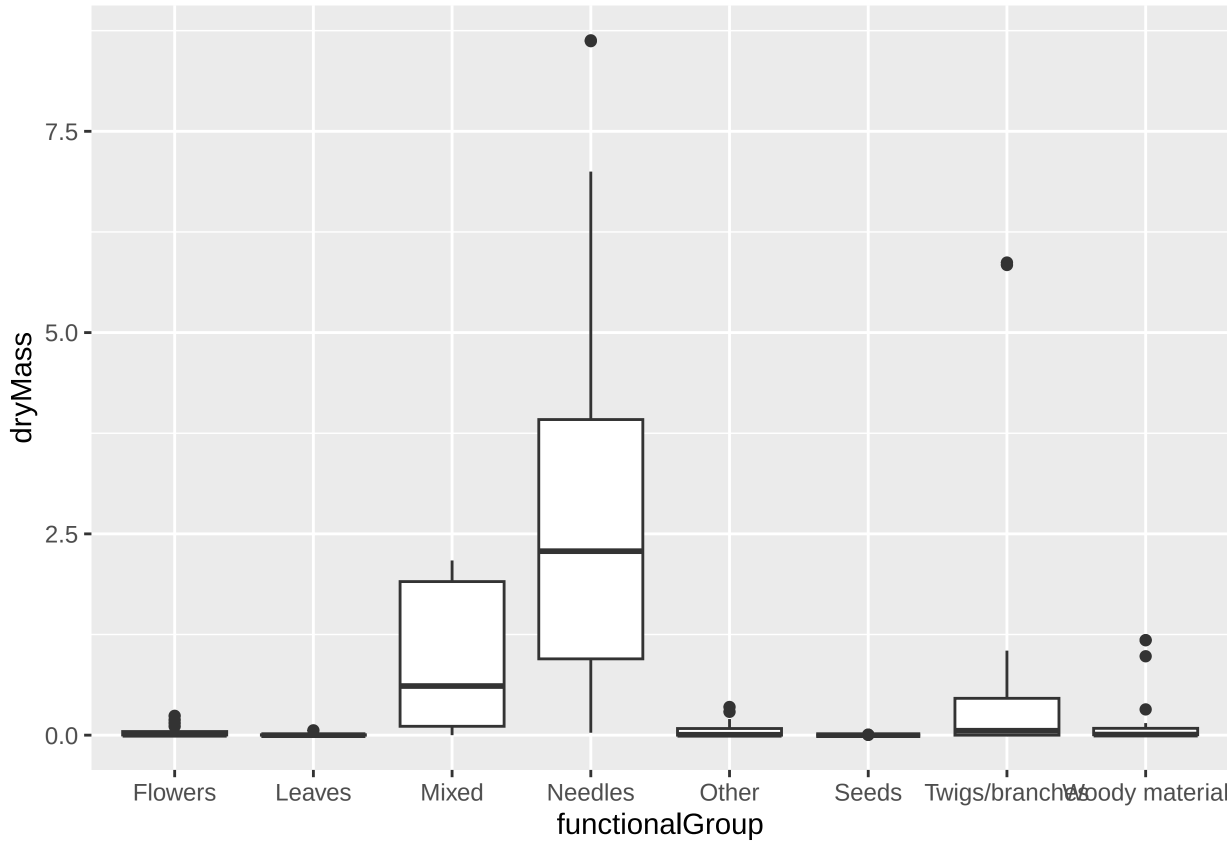
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

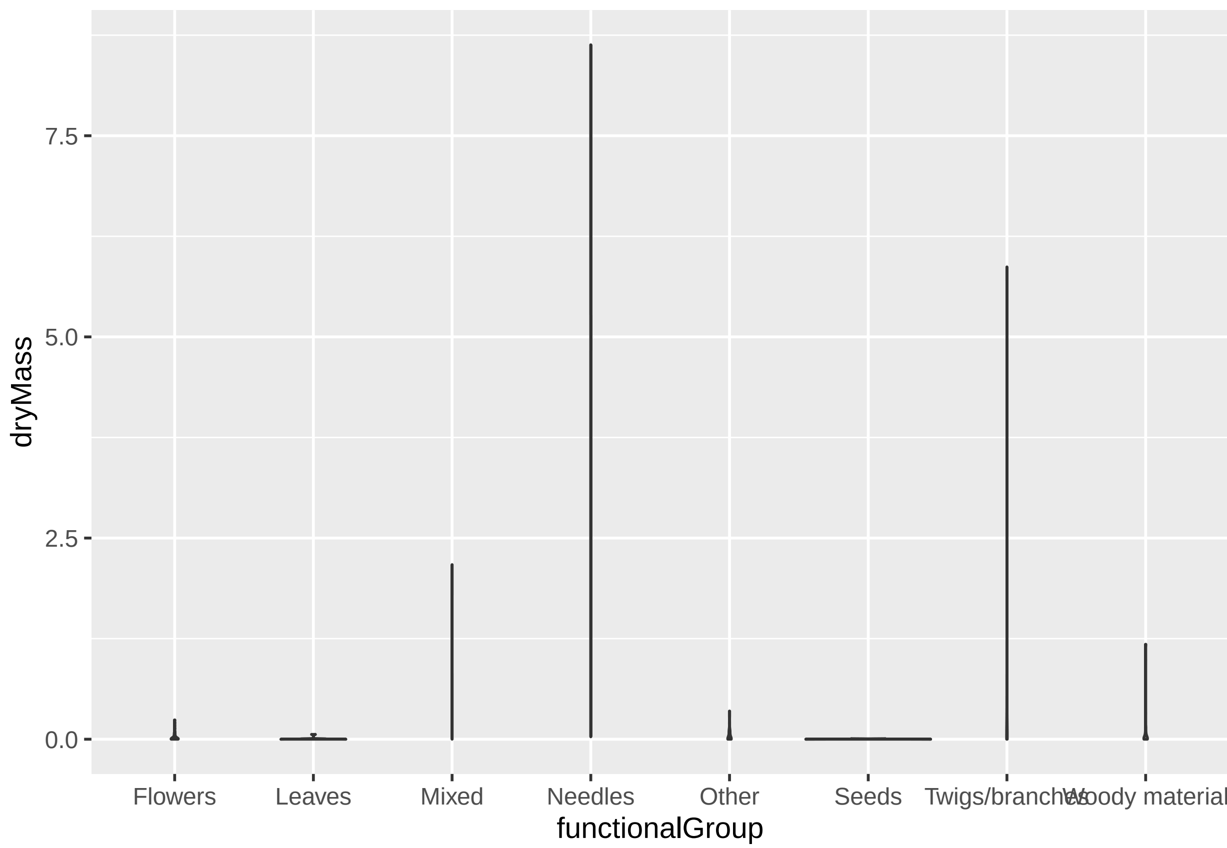



15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
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: There is not much variability in distribution within each functional group of litter. Hence, we can see only straight vertical or horizontal lines in the violin plot.

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

> Answer: Needles tend to have the highest biomass at these sites.