Assignment 3: Data Exploration

Ayden Schirmacher

Spring 2025

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 Canvas.

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. Load necessary packages (tidyverse, lubridate, here), check your current working directory 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.

```
#loading packages (already in system, now just making sure they're on)
install.packages("tidyverse")
install.packages("lubridate")
install.packages("here")
library(tidyverse)
library(lubridate)
library(here)

#reading neonics and litter files, "here" is set to the ENV repository, specify that we are looking in
neonics<-read.csv(
```

```
  file = here('Data','Raw','ECOTOX_Neonicotinoids_Insects_raw.csv'),
  stringsAsFactors = T)

litter<-read.csv(
  file = here('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 ecotoxicology of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: the ecotoxicology of neonicotinoids is important because, as pesticides, they impact a whole host of ecosystems. Knowing the ecotox of the product helps to understand what species the product might target and how this will impact pollinator and pest systems. Furthermore, we can understand better how the runoff from pesticide use will impact surrounding environments and resources, as well as how plants and animals may become resistant to the product.

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 large role in soil health and bottom canopy ecosystems, especially in forest ecosystems. Understanding this process will help to understand nurtrient cycling and soil chemistry, and how these impact fungal and microbio communities. There are also large implications for carbon sequestration as the organic matter decomposes into the ground and for forest management (i.e. fire suppression, forest health, etc.)

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. According to the manual, "Mass data for each collection event are measured separately for the following functional groups, to an accuracy of 0.01 grams. Weights < 0.01g are reported and may indicate presence of a given functional group, identified in the sorting process, but not present at detectable masses" (p.3) 2. Spatial sampling design varies based on height of vegetation and percent of vegetation that is classified as "woody". 3. Temporal sampling design varies between sites classified as deciduous vs. evergreen. More frequent sampling was conducted at deciduous sites.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

   Answer: There are 4,623 observations of 30 variables

```r
str(neonics) #shows number of rows and colums in dataset using dplyr package
```

```
## '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 1 ..
##  $ Organism.Lifestage              : Factor w/ 20 levels "Adult","Cocoon",..: 1 1 19 19 19 1 19 1 1 1
##  $ Organism.Age                    : Factor w/ 39 levels "<=24","<=48",..: 39 39 39 39 39 36 39 36 36
##  $ 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 1
##  $ Media.Type                      : Factor w/ 10 levels "Agar","Artificial soil",..: 7 7 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","' 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 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 1
##  $ 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 1
##  $ Observed.Duration..Days.        : Factor w/ 361 levels "<.0002","<.0021",..: 145 145 145 145 145 1
##  $ Observed.Duration.Units..Days.  : Factor w/ 17 levels "Day(s)","Day(s) post-emergence",..: 1 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
```

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? [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```r
sort(summary(neonics$Effect)) #sort the number of observations in the neonics effects column
```

```
##       Hormone(s)        Histology       Physiology          Cell(s)
##                1                5                7                9
##     Biochemistry     Accumulation     Intoxication    Immunological
##               11               12               12               16
##       Morphology           Growth       Enzyme(s)         Genetics
##               22               38               62               82
##        Avoidance      Development     Reproduction Feeding behavior
##              102              136              197              255
##         Behavior        Mortality       Population
##              360             1493             1803
```

Answer: The most common effects studied in the neonics dataset were population with 1,803 observations; mortality with 1,493 observations; and behavior with 360 observations. These

are likely the most commonly studied because they are important baselines for effects studies – you cannot know true effects without knowing the overall population health and stability. Furthermore, it is likely that these are easier to study than say, intoxication.

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: Explore the help on the `summary()` function, in particular the `maxsum` argument...]

   Answer: The most common observation was "other". However, the six most common specific species are 1. Honey Bee, 2. Parasitic Wasp; 3. Buff Tailed Bumblebee; 4. Carniolan Honey Bee; 5. Bumble Bee; and 6. Italian Honeybee. All of these are types of bees or wasps belonging to the Hymenoptera order, which includes both bees and wasps. All play crucial roles in pollination. Because of this role in pollination, knowing how pesticides and insecticides impact these organisms is very important – without them, breeding and producing agricultural products would be much, much harder.

```
sort(summary(neonics$Species.Common.Name)) #sort the number of observations in the neonics 'species com
```

```
##                         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
```

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? [Tip: Viewing the dataframe may be helpful...]

   Answer: The 'Conc.1.Author' column is a factor class. This is because the column contains some non-numeric values (i.e. "NR/"), so the R software automatically reads the column as non-numeric in order to make sense of it.

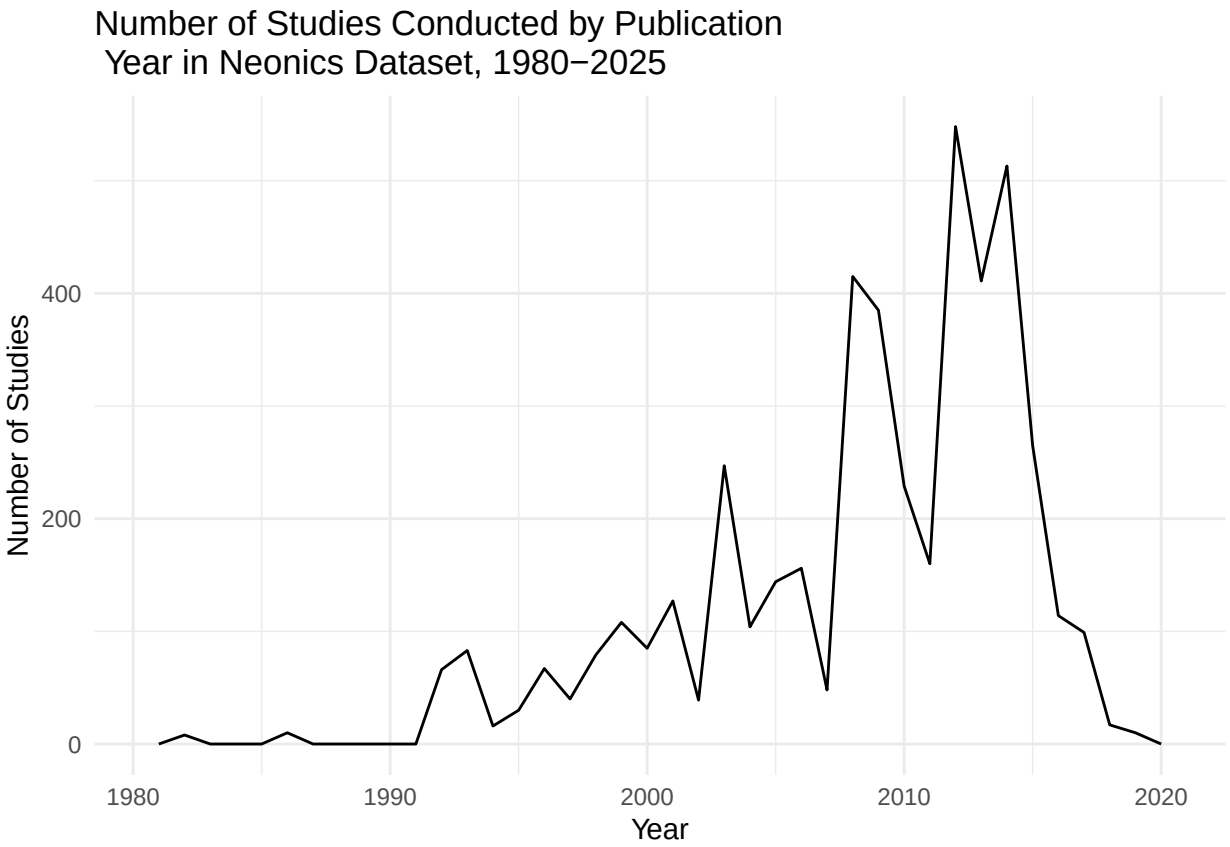
```
class(neonics$Conc.1..Author.) #find class of column
```

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

Explore your data graphically (Neonics)

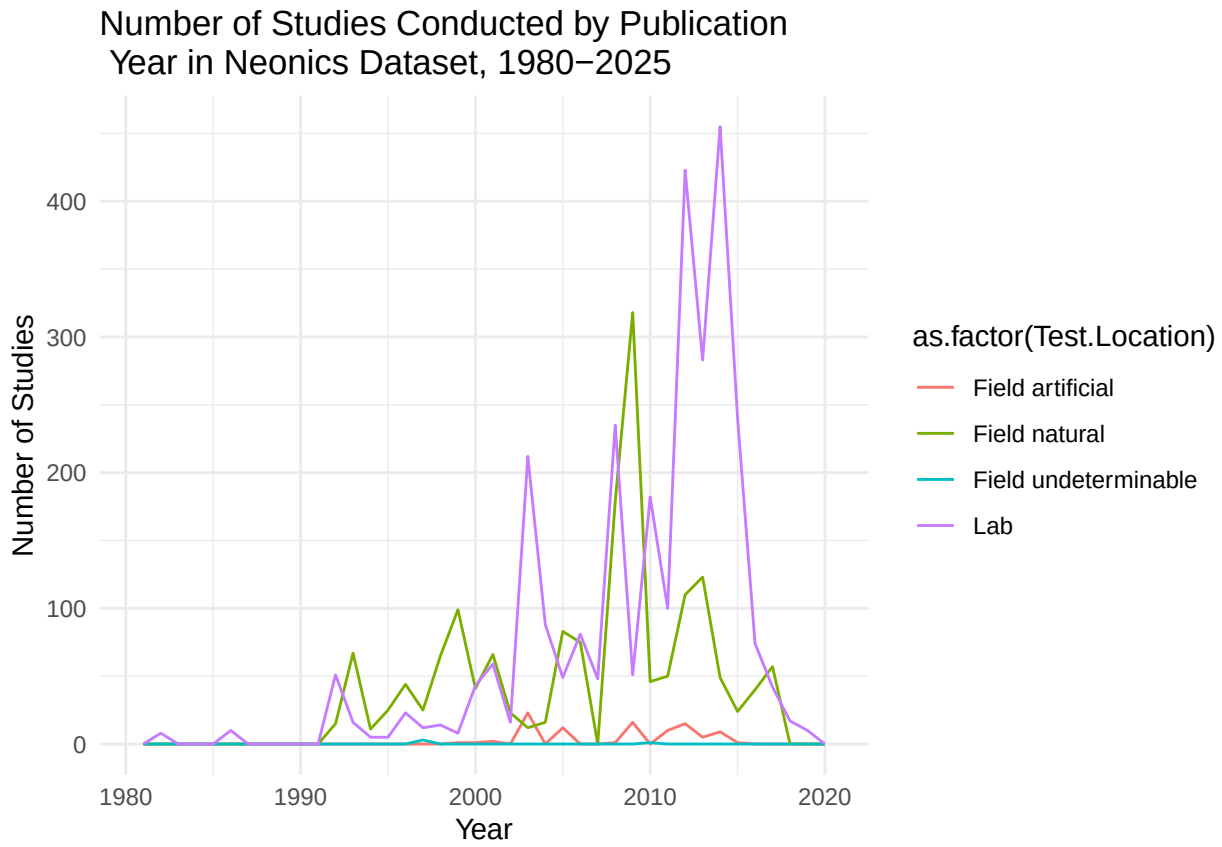
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(neonics, aes(x = Publication.Year)) + #establish publication year (time) as independent variable
  geom_freqpoly(binwidth = 1) + #frequency plot code
  labs(title = "Number of Studies Conducted by Publication\n Year in Neonics Dataset, 1980-2025",
       x = "Year",
       y = "Number of Studies") +
  theme_minimal() #make professional
```



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 = as.factor(Test.Location))) + #color the graph by loca
  geom_freqpoly(binwidth = 1) + #size of lines
  labs(title = "Number of Studies Conducted by Publication\n Year in Neonics Dataset, 1980-2025",
       x = "Year",
       y = "Number of Studies") +
  theme_minimal()
```



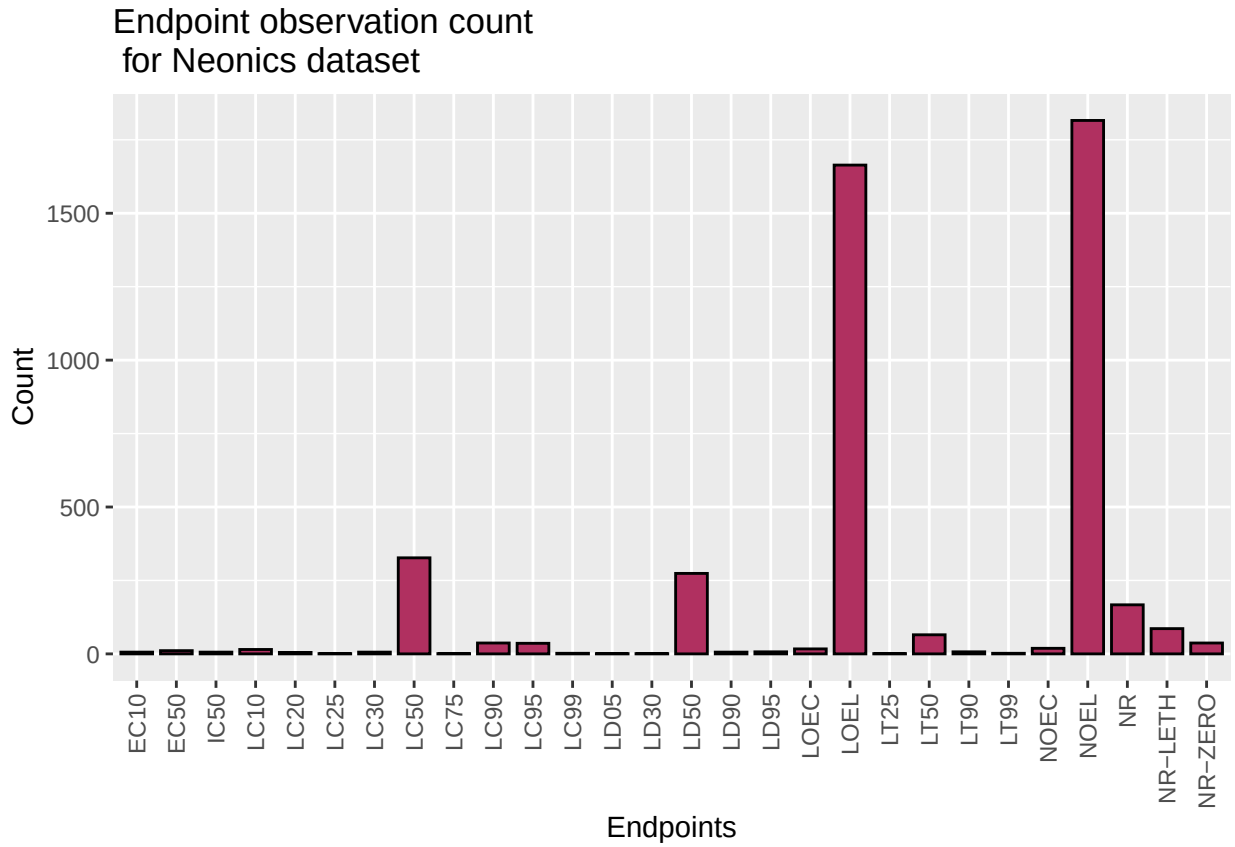
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The graph shows that the number of studies was largest between 2010 and 2015, with the larges spike being in ~2014 and a very steep dropoff after that year. By using color differentiation, we can see that the most common test location is in natural fields between ~1992 and ~2000, before being overtaken by testing locations in labs between ~2000 and ~2020. However, there was a brief period from ~2008 to ~2010 where natural field sampling was the predominant location. Articficial fields and undeterminable fields present very small proportions of the overall test location data.

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))+ #endpoint as independent variable
  geom_bar(width=0.8, fill="maroon", color="black")+ #fill bars with color to make pretty
         labs(x="Endpoints",
              y="Count",
              title="Endpoint observation count\n for Neonics dataset")+
         theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) #rotate and align x-axis l
```



Answer: The two most common endpoints are LOEL and NOEL. According to the ECOTOX_CodeAppendix, these represent the information listed below. Both are a part of the Terrestrial database usage (p. 721-722) - LOEL: Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC) - 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)

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.

    Answer: in August 2018, data was collected on August 2nd and August 30

```
class(litter$collectDate) #check class of collectDate, its a factor
```

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

```
litter$collectDate<-as.Date(litter$collectDate) #change to Date class
class(litter$collectDate) #confirm change
```

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

```r
unique(litter$collectDate, 2018-08) #check dates in august 2018 that data was collected
```

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

13. Using the `unique` function, determine how many different plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

    Answer: The information obtained from unique counts only how many different plot IDs there are and tells what the names of those unique plots are. Summary expands on this information by telling how many times a study was taken from those unique sites.

```r
unique(litter$plotID) #using unique function
```

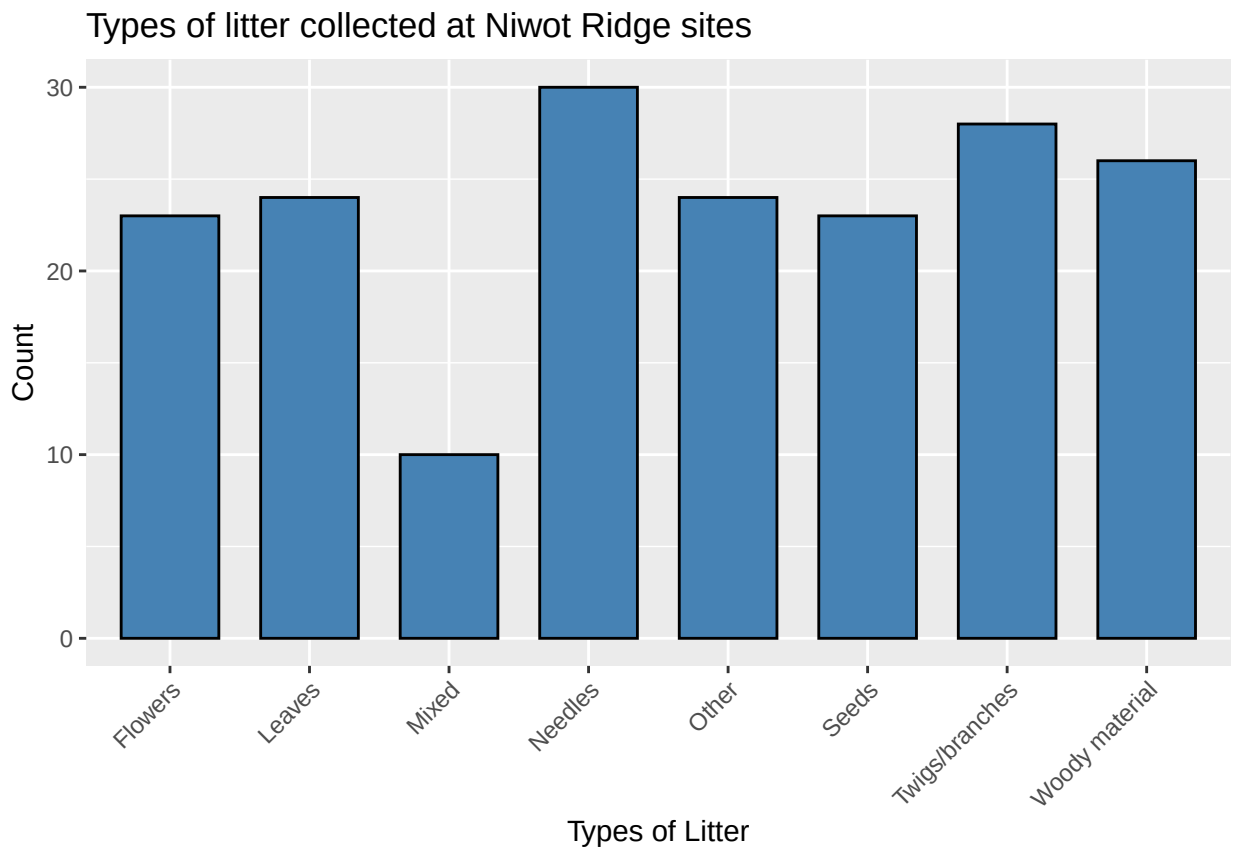
```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

```r
summary(litter$plotID) #using summary function for more detail
```

```
## 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
```

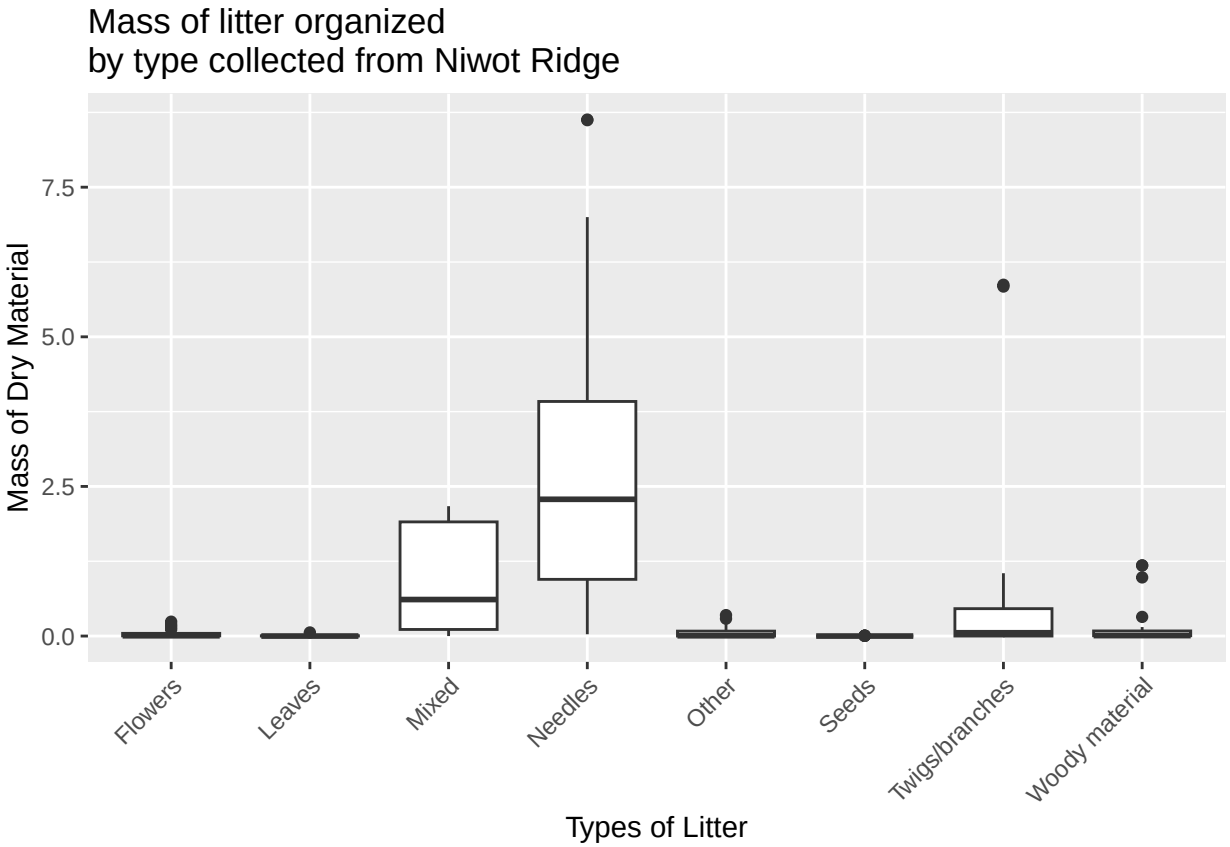
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
ggplot(litter, aes(x=functionalGroup))+ #types of litter as independent variable
  geom_bar(width=0.7, fill="steelblue", color="black")+ #color for bins
  labs(x="Types of Litter",
       y="Count",
       title="Types of litter collected at Niwot Ridge sites")+
  theme(axis.text.x = element_text(angle = 45, hjust=1)) #rotate and align x-axis labels
```

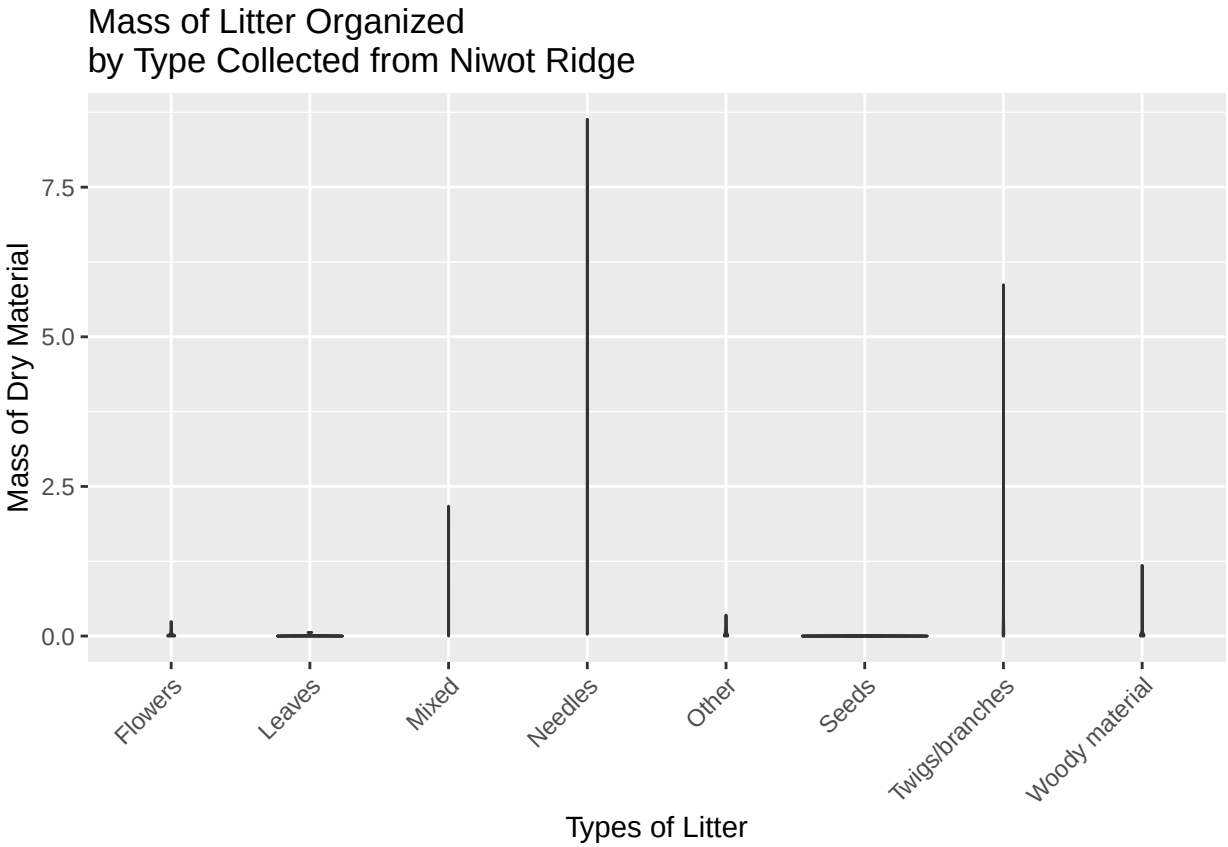


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(litter, aes(x=functionalGroup, y=dryMass))+ #use dryMass as dependent variable
  geom_boxplot(width=0.7)+ #make boxplot
  labs(x="Types of Litter",
       y="Mass of Dry Material",
       title="Mass of litter organized \nby type collected from Niwot Ridge")+
  theme(axis.text.x = element_text(angle = 45, hjust=1)) #adjust titles for legibility
```




```
ggplot(litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass), draw_quantiles=c(0.25, 0.5, 0.75))+ #use dryMass a
  labs(x = "Types of Litter",
       y = "Mass of Dry Material",
       title = "Mass of Litter Organized \nby Type Collected from Niwot Ridge")+
    theme(axis.text.x = element_text(angle = 45, hjust=1)) #adjust titles for legibility
```



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

> Answer: The violin plot is not showing any useful information. This could be due to the fact that the distribution of litter types was fairly evenly distributed, so the density of observations is not extremely variable.

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

> Answer: Needles is the highest, followed by mixed with the most consistent high biomass, though twigs/branches has a high outlier.