Assignment 3: Data Exploration

Ana Bishop

Spring 2023

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A03_DataExploration.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Assign a useful **name to each code chunk** and include ample **comments** with your code.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.
7. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai.

TIP: If your code extends past the page when knit, tidy your code by manually inserting line breaks.

TIP: If your code fails to knit, check that no `install.packages()` or `View()` commands exist in your code.

Set up your R session

1. Check your working directory, load necessary packages (tidyverse, lubridate), and upload two datasets: the ECOTOX neonicotinoid dataset (ECOTOX_Neonicotinoids_Insects_raw.csv) and the Niwot Ridge NEON dataset for litter and woody debris (NEON_NIWO_Litter_massdata_2018-08_raw.csv). Name these datasets "Neonics" and "Litter", respectively. Be sure to include the subcommand to read strings in as factors.

```
getwd() #checking working directory, confirming that it's ENV872
```

```
## [1] "/Users/anabishop/Documents/Data Analytics/ENV872"
```

```
library(tidyverse)  #loading packages
library(lubridate)
library(ggplot2)

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: One reason we could be interested in these results is so that we know the real effects that these insecticides have on the insects, and can judge how effective they are. Knowing the efficacy of each neonicotinoid could help weigh the cost/benefit of using each type. For example, if one is more environmentally damaging than another, but they both were found to have the same efficacy, this study could help managers choose the less damaging option.

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: These studies are crucial for determining the types of nutrients going back into the soil, as well as the rate at which this re-entry is happening. Investigating what effect the decaying litter and debris has on the surrounding ecosystem is crucial for understanding the nutrient cycling happening within the ecosystem.

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 takes place in locations that have woody vegetation >2m tall. 2. Ground traps are sampled once per year. 3. For sites that have deciduous vegetation or limited access during winter months, litter sampling of elevated traps may be discontinued for up to 6 months during the dormant season.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(neonics)
```

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

```
str(neonics)
```

```
## ’data.frame’:    4623 obs. of  30 variables:
##  $ CAS.Number                      : int  58842209 58842209 58842209 58842209 58842209 58842209 58842
##  $ 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
##  $ 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 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
```

```
dim(litter)
```

```
## [1] 188  19
```

```
str(litter)
```

```
## ’data.frame':    188 obs. of  19 variables:
##  $ uid                    : Factor w/ 188 levels "028eea3d-5c20-4afc-bb7e-a05bab305152",..: 84 96 85
##  $ namedLocation          : Factor w/ 12 levels "NIWO_040.basePlot.ltr",..: 8 8 8 8 8 8 8 8 11 11 ..
##  $ domainID               : Factor w/ 1 level "D13": 1 1 1 1 1 1 1 1 1 1 ...
##  $ siteID                 : Factor w/ 1 level "NIWO": 1 1 1 1 1 1 1 1 1 1 ...
##  $ plotID                 : Factor w/ 12 levels "NIWO_040","NIWO_041",..: 8 8 8 8 8 8 8 8 11 11 ...
##  $ trapID                 : Factor w/ 12 levels "NIWO_040_205",..: 8 8 8 8 8 8 8 8 11 11 ...
##  $ weighDate              : Factor w/ 2 levels "2018-08-06","2018-09-05": 1 1 1 1 1 1 1 1 1 1 ...
##  $ setDate                : Factor w/ 2 levels "2018-07-05","2018-08-02": 1 1 1 1 1 1 1 1 1 1 ...
##  $ collectDate            : Factor w/ 2 levels "2018-08-02","2018-08-30": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ovenStartDate          : Factor w/ 2 levels "2018-08-02T21:00Z",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ ovenEndDate            : Factor w/ 2 levels "2018-08-06T18:02Z",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ fieldSampleID          : Factor w/ 23 levels "NEON.LTR.NIWO040205.20180802",..: 14 14 14 14 14 14
##  $ massSampleID           : Factor w/ 168 levels "NEON.LTR.NIWO040205.20180802.FLR",..: 102 101 103
##  $ samplingProtocolVersion: Factor w/ 1 level "NEON.DOC.001710vE": 1 1 1 1 1 1 1 1 1 1 ...
##  $ functionalGroup        : Factor w/ 8 levels "Flowers","Leaves",..: 7 6 8 1 8 4 5 2 1 8 ...
##  $ dryMass                : num  0.4 0.005 0.04 0.005 0.07 1 0.2 0.005 0.19 1.18 ...
##  $ qaDryMass              : Factor w/ 2 levels "N","Y": 1 1 2 1 1 1 1 1 1 2 ...
##  $ remarks                : logi  NA NA NA NA NA NA ...
##  $ measuredBy             : Factor w/ 2 levels "kstyers@battelleecology.org",..: 1 1 1 1 1 1 1 1 1 1
```

The neonics data frame has 4623 rows and 30 columns, and the litter dataset has 188 rows and 19 columns.

6. Using the `summary` function on the “Effect” column, determine the most common effects that are studied. Why might these effects specifically be of interest?

```
summary(neonics$Effect)
```

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

Answer: The most common effects studied are population, mortality, behavior, and feeding behavior. Again, these variables are likely of the most interest in order to determine the efficacy of each insecticide. Those variables all help determine how the insects react to the chemicals.

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…]

```
summary(sort(neonics$Species.Common.Name))
```

```
##                          Honey Bee                     Parasitic Wasp
##                                667                                285
##              Buff Tailed Bumblebee                Carniolan Honey Bee
##                                183                                152
##                         Bumble Bee                   Italian Honeybee
##                                140                                113
##                    Japanese Beetle                  Asian Lady Beetle
##                                 94                                 76
##                     Euonymus Scale                           Wireworm
##                                 75                                 69
##                  European Dark Bee                  Minute Pirate Bug
##                                 66                                 62
##               Asian Citrus Psyllid                      Parastic Wasp
##                                 60                                 58
##             Colorado Potato Beetle                    Parasitoid Wasp
##                                 57                                 51
##                Erythrina Gall Wasp                       Beetle Order
##                                 49                                 47
##        Snout Beetle Family, Weevil           Sevenspotted Lady Beetle
##                                 47                                 46
##                     True Bug Order              Buff-tailed Bumblebee
##                                 45                                 39
##                       Aphid Family                     Cabbage Looper
##                                 38                                 38
##               Sweetpotato Whitefly                      Braconid Wasp
##                                 37                                 33
##                       Cotton Aphid                     Predatory Mite
##                                 33                                 33
##             Ladybird Beetle Family                         Parasitoid
##                                 30                                 30
##                      Scarab Beetle                      Spring Tiphia
##                                 29                                 29
##                        Thrip Order               Ground Beetle Family
##                                 29                                 27
##                 Rove Beetle Family                      Tobacco Aphid
##                                 27                                 27
##                       Chalcid Wasp             Convergent Lady Beetle
##                                 25                                 25
##                      Stingless Bee                  Spider/Mite Class
##                                 25                                 24
##                Tobacco Flea Beetle                   Citrus Leafminer
##                                 24                                 23
##                    Ladybird Beetle                          Mason Bee
##                                 23                                 22
##                           Mosquito                      Argentine Ant
##                                 22                                 21
##                             Beetle         Flatheaded Appletree Borer
##                                 21                                 20
##               Horned Oak Gall Wasp                 Leaf Beetle Family
##                                 20                                 20
##                  Potato Leafhopper         Tooth-necked Fungus Beetle
##                                 20                                 20
##                       Codling Moth          Black-spotted Lady Beetle
##                                 19                                 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
##              Araneoid Spider Order                          Bee Order
##                                 17                                 17
##                     Egg Parasitoid                       Insect Class
##                                 17                                 17
##           Moth And Butterfly Order       Oystershell Scale Parasitoid
##                                 17                                 17
## Hemlock Woolly Adelgid Lady Beetle              Hemlock Wooly Adelgid
##                                 16                                 16
##                               Mite                        Onion Thrip
##                                 16                                 16
##              Western Flower Thrips                       Corn Earworm
##                                 15                                 14
##                  Green Peach Aphid                          House Fly
##                                 14                                 14
##                          Ox Beetle                 Red Scale Parasite
##                       14                               14
##        Spined Soldier Bug            Armoured Scale Family
##                       14                               13
##          Diamondback Moth                    Eulophid Wasp
##                       13                               13
##         Monarch Butterfly                    Predatory Bug
##                       13                               13
##     Yellow Fever Mosquito              Braconid Parasitoid
##                       13                               12
##             Common Thrip      Eastern Subterranean Termite
##                       12                               12
##                   Jassid                       Mite Order
##                       12                               12
##                Pea Aphid                  Pond Wolf Spider
##                       12                               12
##   Spotless Ladybird Beetle           Glasshouse Potato Wasp
##                       11                               10
##                 Lacewing           Southern House Mosquito
##                       10                               10
##    Two Spotted Lady Beetle                        Ant Family
##                       10                                9
##             Apple Maggot                           (Other)
##                        9                              670
```

Answer: The six most commonly studied species are the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. These insects are all pollinators, and it's likely that they have the highest interest because the researchers need to preserve these populations while causing mortality to other unwanted species. Therefore, these species need to be heavily studied in order to determine how to keep them coming back despite insecticide being applied.

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.) # factor
```

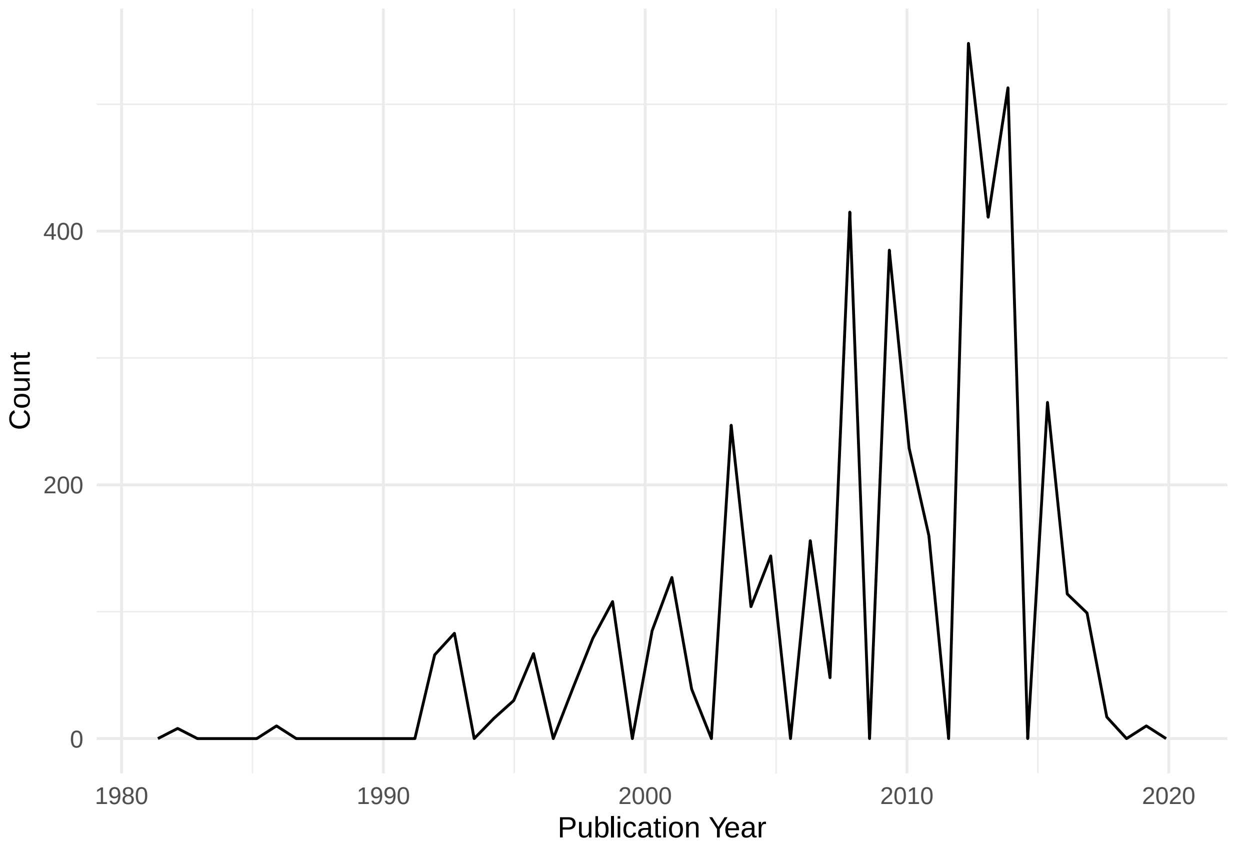
```
## [1] "factor"
```

Answer: The class is a factor. Factors are typically used to classify categorical variables, meaning that this column was input as a categorical variable instead of a continuous one.

Explore your data graphically (Neonics)

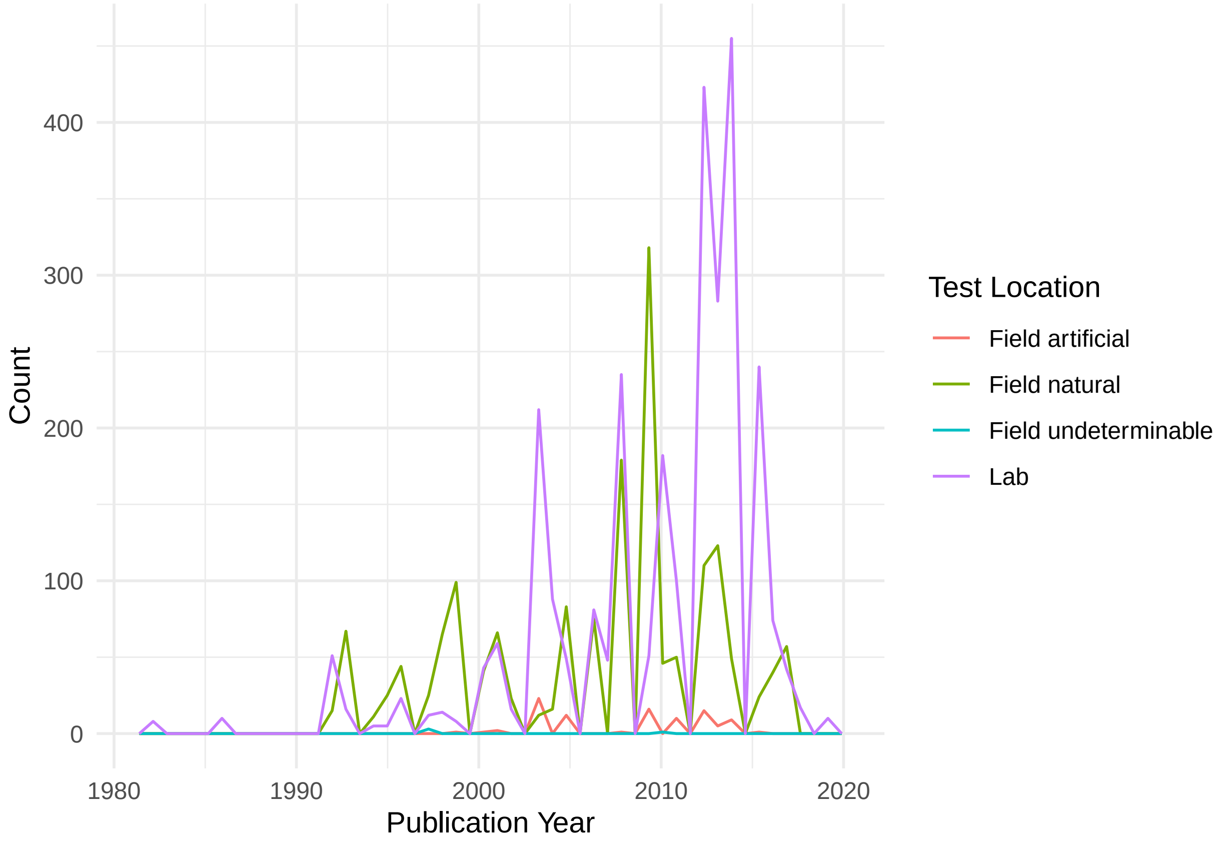
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50) +
  labs(x = "Publication Year", y = "Count") +
  theme_minimal()
```



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), bins = 50) +
  labs(x = "Publication Year", y = "Count", color = "Test Location") +
  theme_minimal()
```



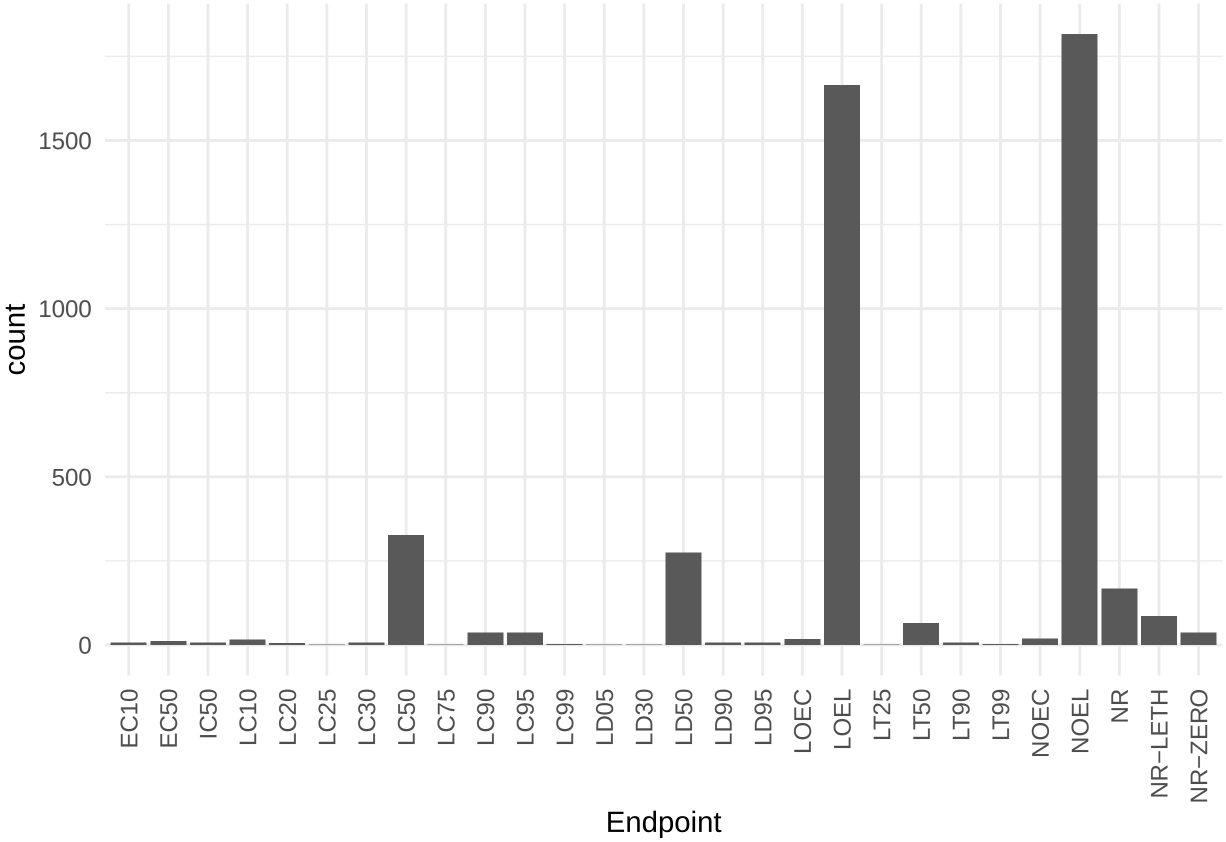
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: Overall, the number of studies conducted by publication year appears to have exponentially increased up until about 2014, when there was a peak. The most common test locations were the lab and the field (natural), while the least common was the field (undeterminable). These results makes sense, and follow what I would have expected.

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_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: The two most common endpoints are NOEL, followed by LOEL. NOEL's database usage is Terrestrial, and it is defined as having "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)". LOEL also has a Terrestrial database usage, and is similar to NOEL, defined as "lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/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) #the class is a factor
```

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

```
litter$collectDate <- as.Date(litter$collectDate, format = "%Y-%m-%d")
class(litter$collectDate) #the class is now a Date
```

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

```
unique(litter$collectDate)
```

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

Litter was sampled on August 2nd and August 30th, 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`?

```
unique(litter$plotID)
```

```
##  [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
```

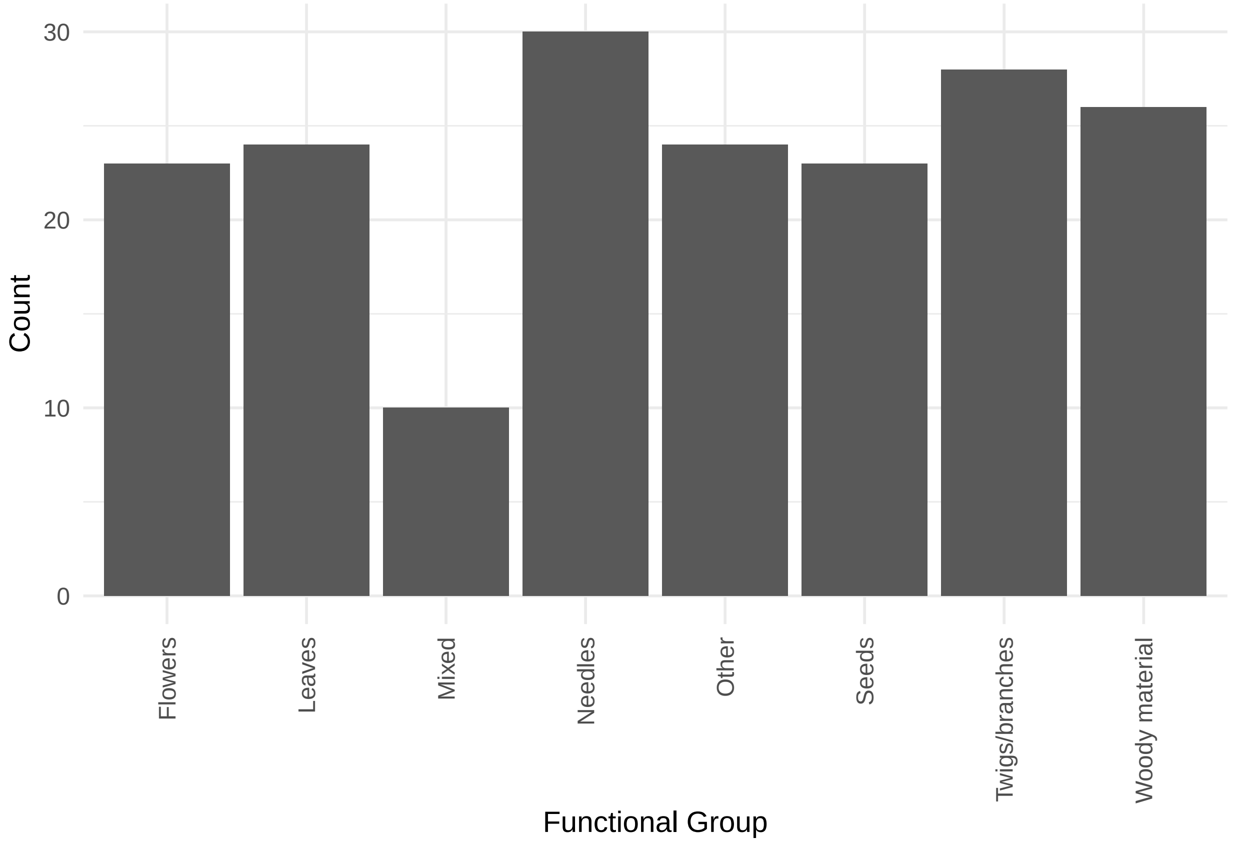
```
summary(litter$plotID)
```

```
## 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: There were 12 plots sampled at Niwot Ridge. The unique function returns every unique entry in the requested column, while summary returns every unique entry in the requested column along with a count of how many times that entry occurred.

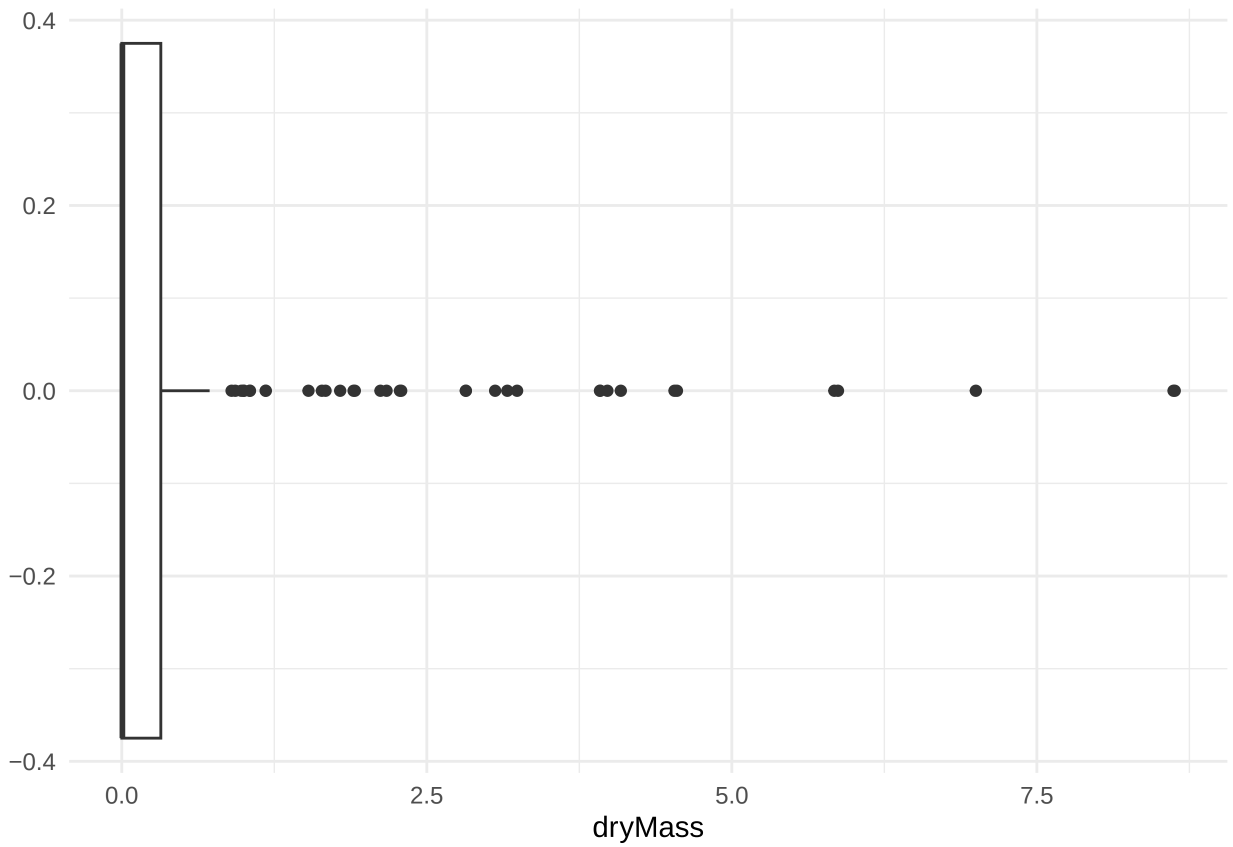
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(litter, aes(x = functionalGroup)) +
  geom_bar() +
  labs(x = "Functional Group", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



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

```
ggplot(litter, aes(x = dryMass)) +
  geom_boxplot() +
  theme_minimal()
```



```
#ggplot(litter, aes(x = dryMass)) +      ##I had to make this command a comment
  #geom_violin()                         so that the document would knit
```

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

> Answer: The boxplot was able to successfully provide statistics on the data, while the violin plot could not be created due to the fact that it is missing the y aesthetic. This likely means that the plot could not accurately calculate the distribution of the data.

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

> Answer: Needles dominate the heaviest biomass rankings, though twigs/branches also had a biomass far higher than many other functional groups.