Assignment 3: Data Exploration

Sophie Valkenberg

Fall 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 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.

```r
#install(tidyverse)
#install(lubridate)
#install(here)
library(tidyverse)
library(lubridate)
library(here)
here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```r
getwd()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```r
#Here, I left code to install all necessary packages, load them, and also
#double check the working dataset

Neonics <- read.csv(
  file = here('Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv'),
  stringsAsFactors = TRUE)
Litter <- read.csv(
  file = here('Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv'),
  stringsAsFactors = TRUE
)
#using the "here" function, I loaded the two datasets needed for this document
```

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 are a type of insecticide, so therefore knowing the toxicological information, such as mode of action and dose-response for insects to understand the level of impact of neonoicotinoids. They're also not highly specified and known to impact important pollintors. Many states, such as Vermont and New York, have placed increased restrictions on neonictoniods for this reason, including banning neonic-treated seeds.

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: Collecting woody debris is a good way to determine the health of a forest. Woody debris tells us things like diseases that may be impacting the forest, what kinds of animals/insects are living in the forest, and overall forest health.

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 is defined as material that is dropped from the forest canopy and has a butt end diameter <2cm and a length <50 cm; this material is collected in elevated 0.5m2 PVC traps." 2.From the temporal sampling design: ground traps are sampled once per year. 3. "Litter and fine woody debris sampling is executed at terrestrial NEON sites that contain woody vegetation >2m tall."

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
summary(Neonics)
```

```
##    CAS.Number       
##  Min.   : 58842209  
##  1st Qu.:138261413  
##  Median :138261413  
##  Mean   :147651982  
##  3rd Qu.:153719234  
##  Max.   :210880925  
##                     
##                                                                                  Chemical.Name
##  (2E)-1-[(6-Chloro-3-pyridinyl)methyl]-N-nitro-2-imidazolidinimine                     :2658  
##  3-[(2-Chloro-5-thiazolyl)methyl]tetrahydro-5-methyl-N-nitro-4H-1,3,5-oxadiazin-4-imine: 686  
##  [C(E)]-N-[(2-Chloro-5-thiazolyl)methyl]-N’-methyl-N’’-nitroguanidine                  : 452  
##  (1E)-N-[(6-Chloro-3-pyridinyl)methyl]-N’-cyano-N-methylethanimidamide                 : 420  
##  N’’-Methyl-N-nitro-N’-[(tetrahydro-3-furanyl)methyl]guanidine                         : 218  
##  [N(Z)]-N-[3-[(6-Chloro-3-pyridinyl)methyl]-2-thiazolidinylidene]cyanamide             : 128  
##  (Other)                                                                               :  61  
##                                                     Chemical.Grade
##  Not reported                                             :3989   
##  Technical grade, technical product, technical formulation: 422   
##  Pestanal grade                                           :  93   
##  Not coded                                                :  53   
##  Commercial grade                                         :  27   
##  Analytical grade                                         :  15   
##  (Other)                                                  :  24   
##                                                   Chemical.Analysis.Method
##  Measured                                                    : 230        
##  Not coded                                                   :  51        
##  Not reported                                                :   5        
##  Unmeasured                                                  :4321        
##  Unmeasured values (some measured values reported in article):  16        
##                                                                           
##                                                                           
##  Chemical.Purity                  Species.Scientific.Name
##  NR     :2502    Apis mellifera               : 667      
##  25     : 244    Bombus terrestris            : 183      
##  50     : 200    Apis mellifera ssp. carnica  : 152      
##  20     : 189    Bombus impatiens             : 140      
##  70     : 112    Apis mellifera ssp. ligustica: 113      
##  75     :  89    Popillia japonica            :  94      
##  (Other):1287    (Other)                      :3274      
##             Species.Common.Name
##  Honey Bee            : 667    
##  Parasitic Wasp       : 285    
##  Buff Tailed Bumblebee: 183    
##  Carniolan Honey Bee  : 152    
##  Bumble Bee           : 140    
##  Italian Honeybee     : 113    
##  (Other)              :3083
##                                                       Species.Group
##  Insects/Spiders                                              :3569
##  Insects/Spiders; Standard Test Species                       :  27
##  Insects/Spiders; Standard Test Species; U.S. Invasive Species: 667
##  Insects/Spiders; U.S. Invasive Species                       : 360
##
##
##
##     Organism.Lifestage  Organism.Age             Organism.Age.Units
##  Not reported:2271     NR     :3851   Not reported        :3515
##  Adult       :1222     2      : 111   Day(s)              : 327
##  Larva       : 437     3      : 105   Instar              : 255
##  Multiple    : 285     <24    :  81   Hour(s)             : 241
##  Egg         : 128     4      :  81   Hours post-emergence:  99
##  Pupa        :  69     1      :  59   Year(s)             :  64
##  (Other)     : 211     (Other): 335   (Other)             : 122
##                     Exposure.Type         Media.Type
##  Environmental, unspecified:1599   No substrate:2934
##  Food                      :1124   Not reported: 663
##  Spray                     : 393   Natural soil: 393
##  Topical, general          : 254   Litter      : 264
##  Ground granular           : 249   Filter paper: 230
##  Hand spray                : 210   Not coded   :  51
##  (Other)                   : 794   (Other)     :  88
##               Test.Location  Number.of.Doses        Conc.1.Type..Author.
##  Field artificial    :  96   2      :2441    Active ingredient:3161
##  Field natural       :1663   3      : 499    Formulation      :1420
##  Field undeterminable:   4   5      : 314    Not coded        :  42
##  Lab                 :2860   6      : 230
##                              4      : 221
##                              NR     : 217
##                              (Other): 701
##  Conc.1..Author. Conc.1.Units..Author.              Effect
##  0.37/  : 208    AI kg/ha  : 575       Population      :1803
##  10/    : 127    AI mg/L   : 298       Mortality       :1493
##  NR/    : 108    AI lb/acre: 277       Behavior        : 360
##  NR     :  94    AI g/ha   : 241       Feeding behavior: 255
##  1      :  82    ng/org    : 231       Reproduction    : 197
##  1023   :  80    ppm       : 180       Development     : 136
##  (Other):3924    (Other)   :2821       (Other)         : 379
##               Effect.Measurement    Endpoint                   Response.Site
##  Abundance             :1699     NOEL   :1816   Not reported          :4349
##  Mortality             :1294     LOEL   :1664   Midgut or midgut gland:  63
##  Survival              : 133     LC50   : 327   Not coded             :  51
##  Progeny counts/numbers: 120     LD50   : 274   Whole organism        :  41
##  Food consumption      : 103     NR     : 167   Hypopharyngeal gland  :  27
##  Emergence             :  98     NR-LETH:  86   Head                  :  23
##  (Other)               :1176     (Other): 289   (Other)               :  69
##  Observed.Duration..Days.       Observed.Duration.Units..Days.
##  1      : 713             Day(s)               :4394
##  2      : 383             Emergence            :  70
##  NR     : 355             Growing season       :  48
##  7      : 207             Day(s) post-hatch    :  20
##  3      : 183             Day(s) post-emergence:  17
##  0.0417 : 133             Tiller stage         :  15
##  (Other):2649             (Other)              :  59
##                                                                             Author
##  Peck,D.C.                                                                    : 208
##  Frank,S.D.                                                                   : 100
##  El Hassani,A.K., M. Dacher, V. Gary, M. Lambin, M. Gauthier, and C. Armengaud:  96
##  Williamson,S.M., S.J. Willis, and G.A. Wright                                :  93
##  Laurino,D., A. Manino, A. Patetta, and M. Porporato                          :  88
##  Scholer,J., and V. Krischik                                                  :  82
##  (Other)                                                                      :3956
##  Reference.Number
##  Min.   :   344
##  1st Qu.:108459
##  Median :165559
##  Mean   :142189
##  3rd Qu.:168998
##  Max.   :180410
##
##
##  Long-Term Effects of Imidacloprid on the Abundance of Surface- and Soil-Active Nontarget Fauna in Tu
##  Reduced Risk Insecticides to Control Scale Insects and Protect Natural Enemies in the Production and
##  Effects of Sublethal Doses of Acetamiprid and Thiamethoxam on the Behavior of the Honeybee (Apis mel
##  Exposure to Neonicotinoids Influences the Motor Function of Adult Worker Honeybees
##  Toxicity of Neonicotinoid Insecticides on Different Honey Bee Genotypes
##  Chronic Exposure of Imidacloprid and Clothianidin Reduce Queen Survival, Foraging, and Nectar Storin
##  (Other)
##                                              Source     Publication.Year
##  Agric. For. Entomol.11(4): 405-419           : 200   Min.   :1982
##  Environ. Entomol.41(2): 377-386              : 100   1st Qu.:2005
##  Arch. Environ. Contam. Toxicol.54(4): 653-661:  96   Median :2010
##  Ecotoxicology23:1409-1418                    :  93   Mean   :2008
##  Bull. Insectol.66(1): 119-126                :  88   3rd Qu.:2013
##  PLoS One9(3): 14 p.                          :  82   Max.   :2019
##  (Other)                                      :3964
##  Summary.of.Additional.Parameters
##  Purity: \xca NR - NR | Organism Age: \xca NR - NR Not reported | Conc 1 (Author): \xca Active ingre
##  Purity: \xca NR - NR | Organism Age: \xca NR - NR Not reported | Conc 1 (Author): \xca Active ingre
##  Purity: \xca NR - NR | Organism Age: \xca NR - NR Not reported | Conc 1 (Author): \xca Active ingre
##  Purity: \xca NR - NR | Organism Age: \xca NR - NR Not reported | Conc 1 (Author): \xca Active ingre
##  Purity: \xca NR - NR | Organism Age: \xca NR - NR Not reported | Conc 1 (Author): \xca Active ingre
##  Purity: \xca NR - NR | Organism Age: \xca NR - NR Not reported | Conc 1 (Author): \xca Formulation
##  (Other)
```

```
dim(Neonics)
```

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

```
#The summary command gave a summary of all information in the dataset
#The dim command gave only the dimensions of the dataset
```

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…]

```
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
```

```
#This command specified to only the Effect column summarized all of the
#information found in that column
```

Answer: Population seems to be the most commonly studied effect, followed by mortality. These specific effects are likely of interest because it shows the impact of the neonicotinoid being studied on the overall insect population and the insect's life, allowing the researcher to understand the implications of using certain doses of these insecticides.

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…]

```
summary(Neonics$Species.Common.Name, maxsum=7)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee
##                   152                   140                   113
##               (Other)
##                  3083
```

```
#This command gave a summary of specifically common names of the
#top 7 studied insects (top 6 and then a count of "other")
```

Answer: Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, Italian Honeybee All of these insects are in the order Hymenopetra and are all pollinators. It makes sense that they are of interest over other insects since pollinators are extrememly important and also at high risk of being impacted by neonicotinoids.

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…]

```
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
```

```r
#These commands show first the class of this column, then a summary of the
#values in the column. This is helpful in figuring out why the class for this
#column is "Factor" and not "numeric" like concentrations often are
```

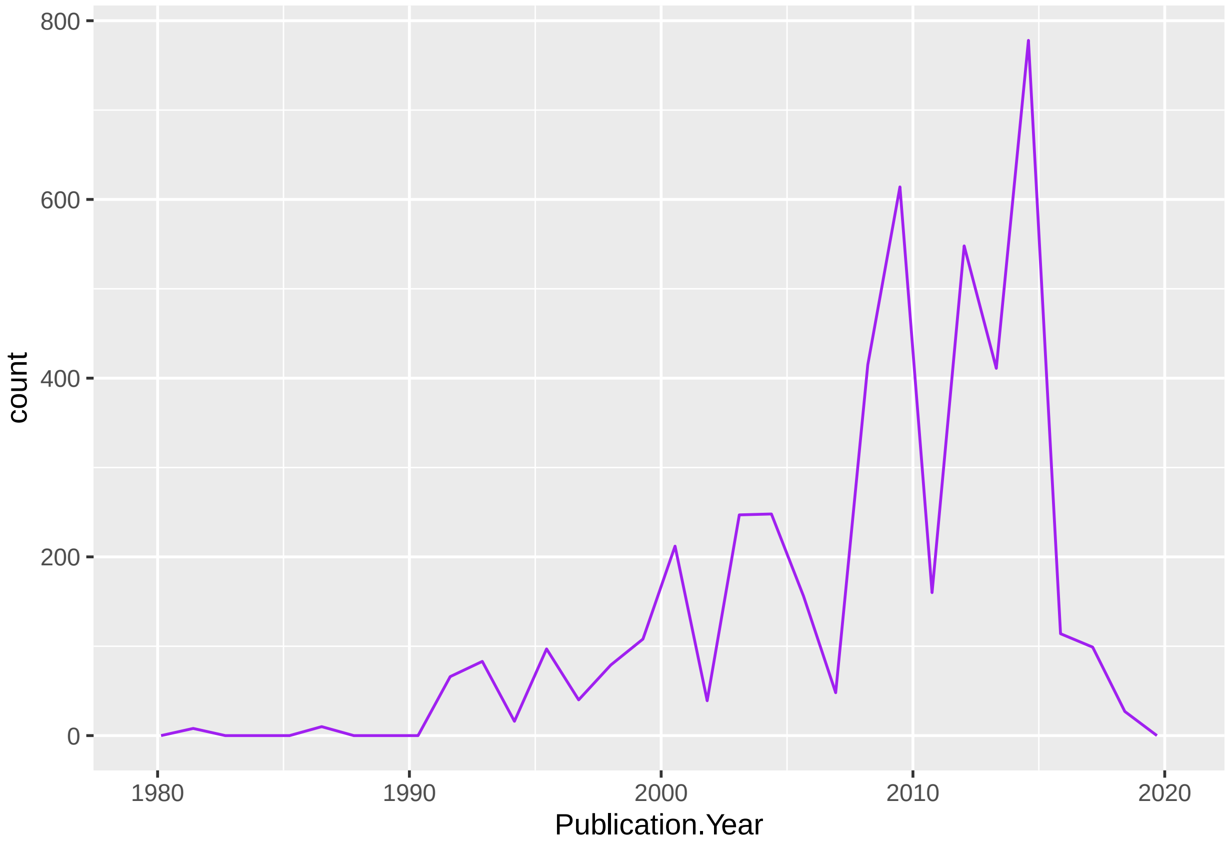
Answer: This is a factor likely because there are non-numeric characters in the some of the cells. Some of the cells had "/" after the number.

Explore your data graphically (Neonics)

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

```r
ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year), color = "purple")
```

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

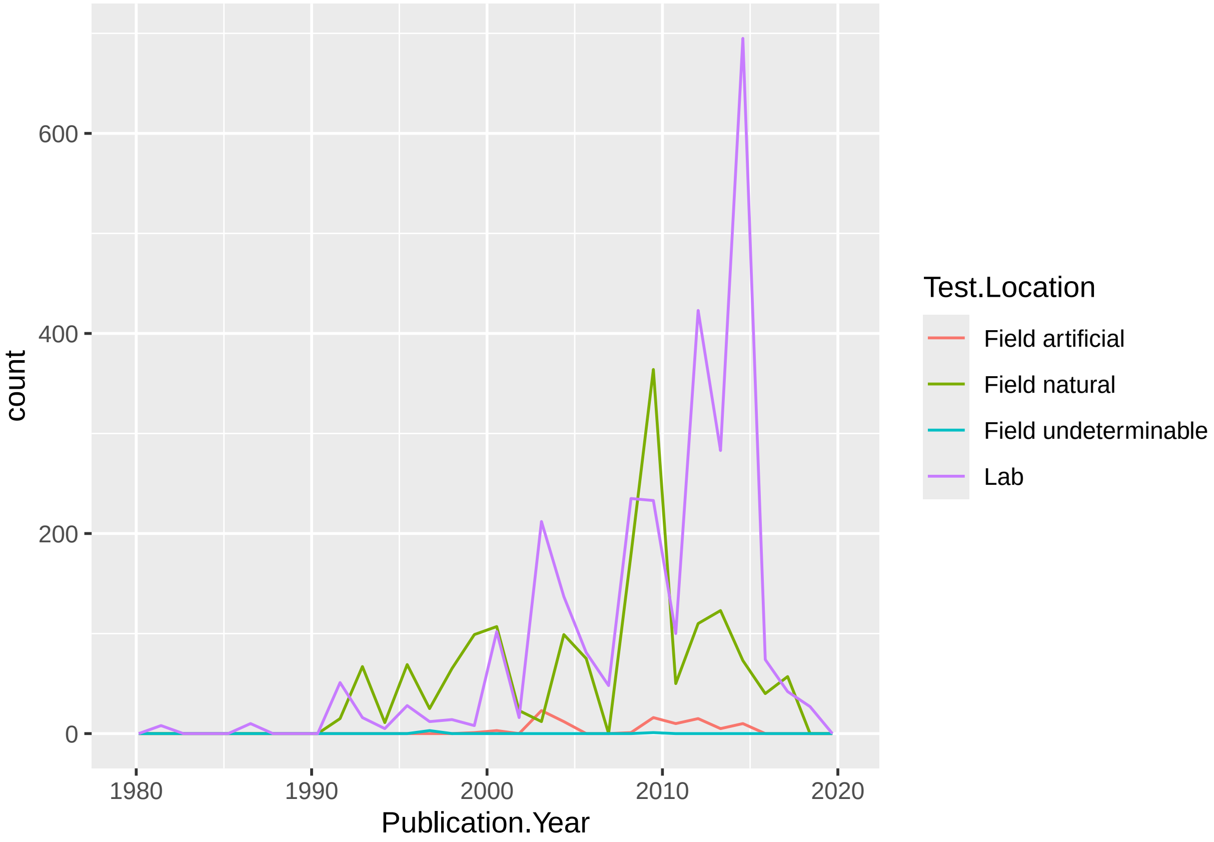


```
#Here, I created a plot showing the number of studies conducted organized by
#publication year
```

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))
```

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



```
#I reused the same command from the previous R chunk, but added a line of code
#that indicates that the plot should be differentiated by color using the
#Test Locations as the coloration source
```

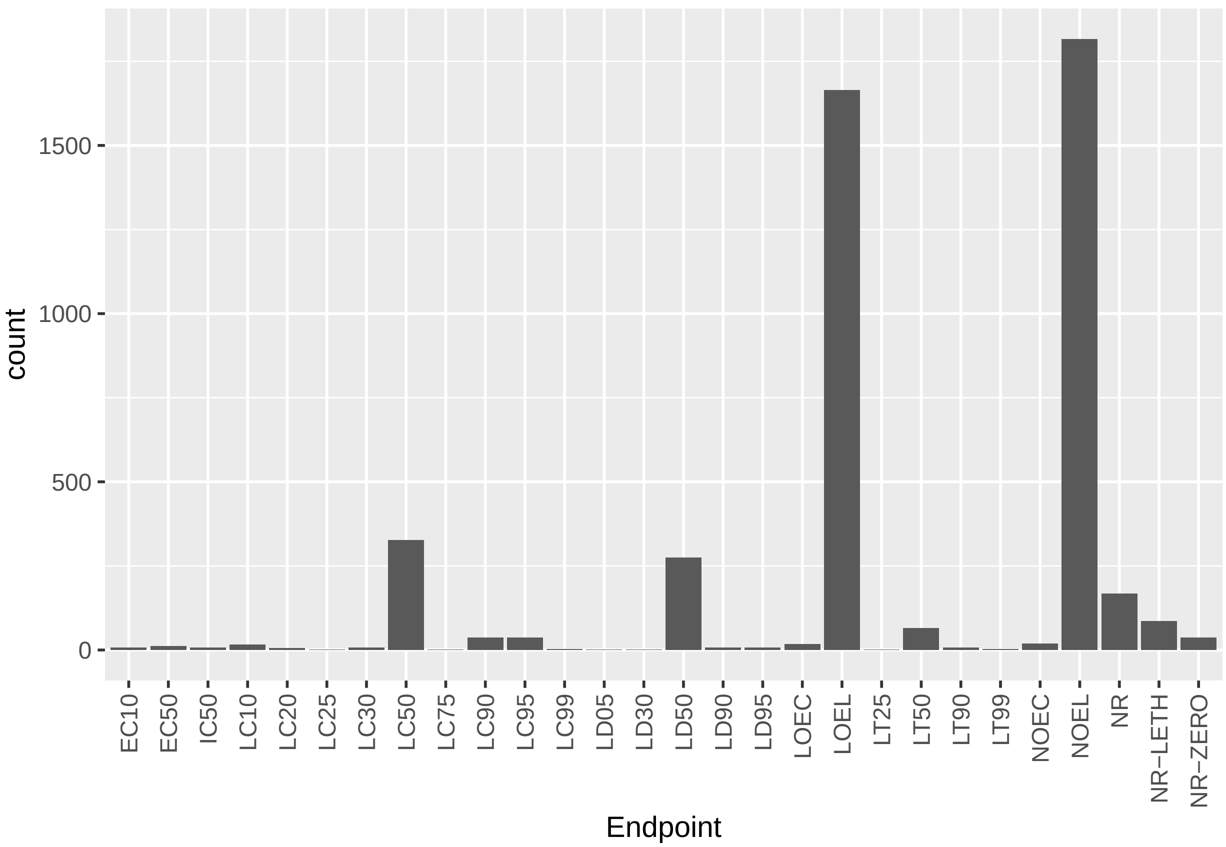
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: Lab is most common test location and natural field follows as the second most common. Natural field and lab were relatively equally common for the most part until lab locations spiked in the 2010s. Natural field locations began appearing in the 1990s and before then the lab was the most common.

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, hjust=1))
```



Answer: LOEL and NOEL are the most common endpoints. LOEL is the lowest observed effect level, meaning that at this concentration/level of toxicant was the lowest dose where an effect was observed. NOEL is the no observed effect level was the highest dose where no effect was observed.

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"
```

```
#Here, I determined the class for the collectDate column which returned as
#a factor

Litter$collectDate <- as.Date(Litter$collectDate, format= '%Y-%m-%d')
#Here, I re-classed the collectDate column to a date instead of a factor
class(Litter$collectDate)
```

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

```r
#Here, I double checked to confirm the new class is "Date"
unique(Litter$collectDate)
```

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

```r
#August 2nd and 30th were the sample dates in August
```

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`?

```r
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
```

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

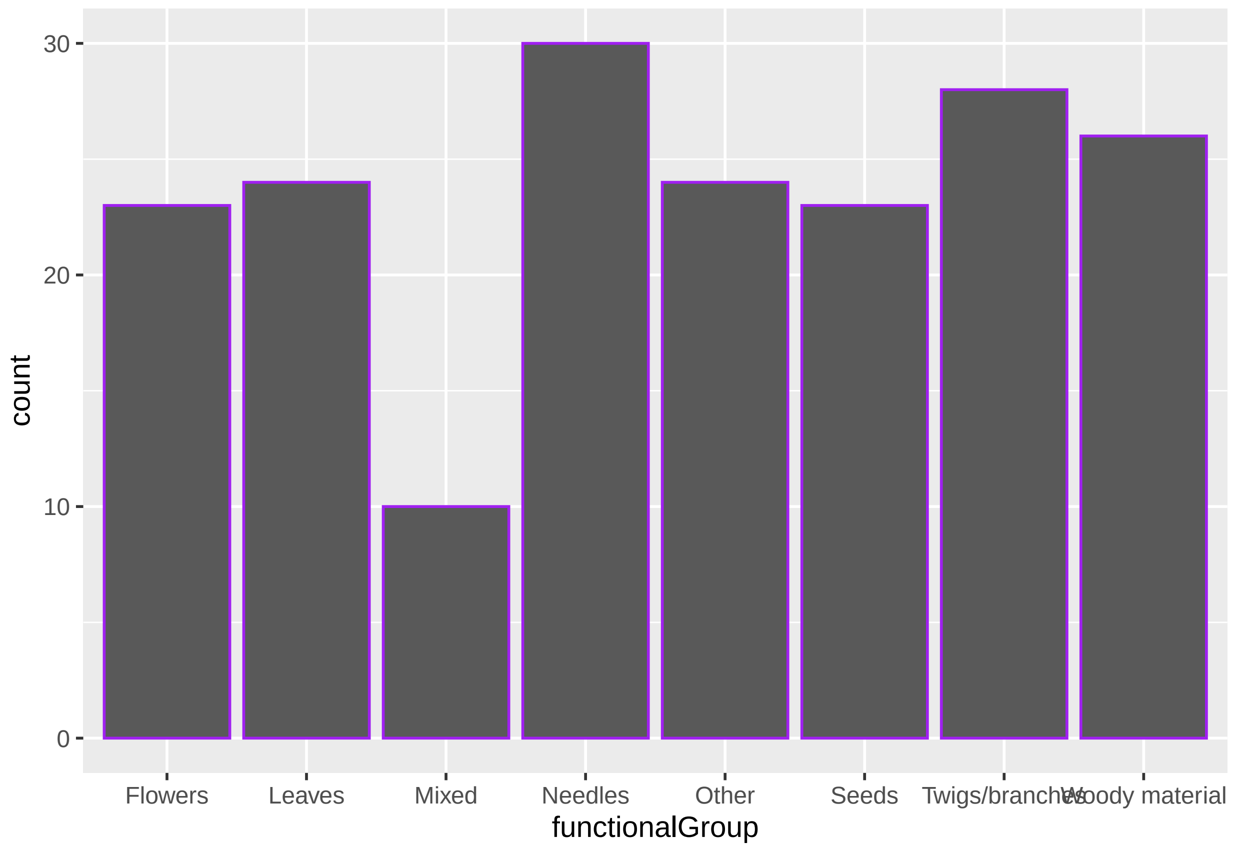
```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

```r
#Here, I determined how many unique plot IDs there are for the Niwot Ridge
#samples. I also used the summary command to determine the differences between
#the information obtained from the unique commans vs. the summary command
```

Answer: The unique function only returns the plot IDs, whereas the the summary function seems to also return the count for each plot ID.

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.

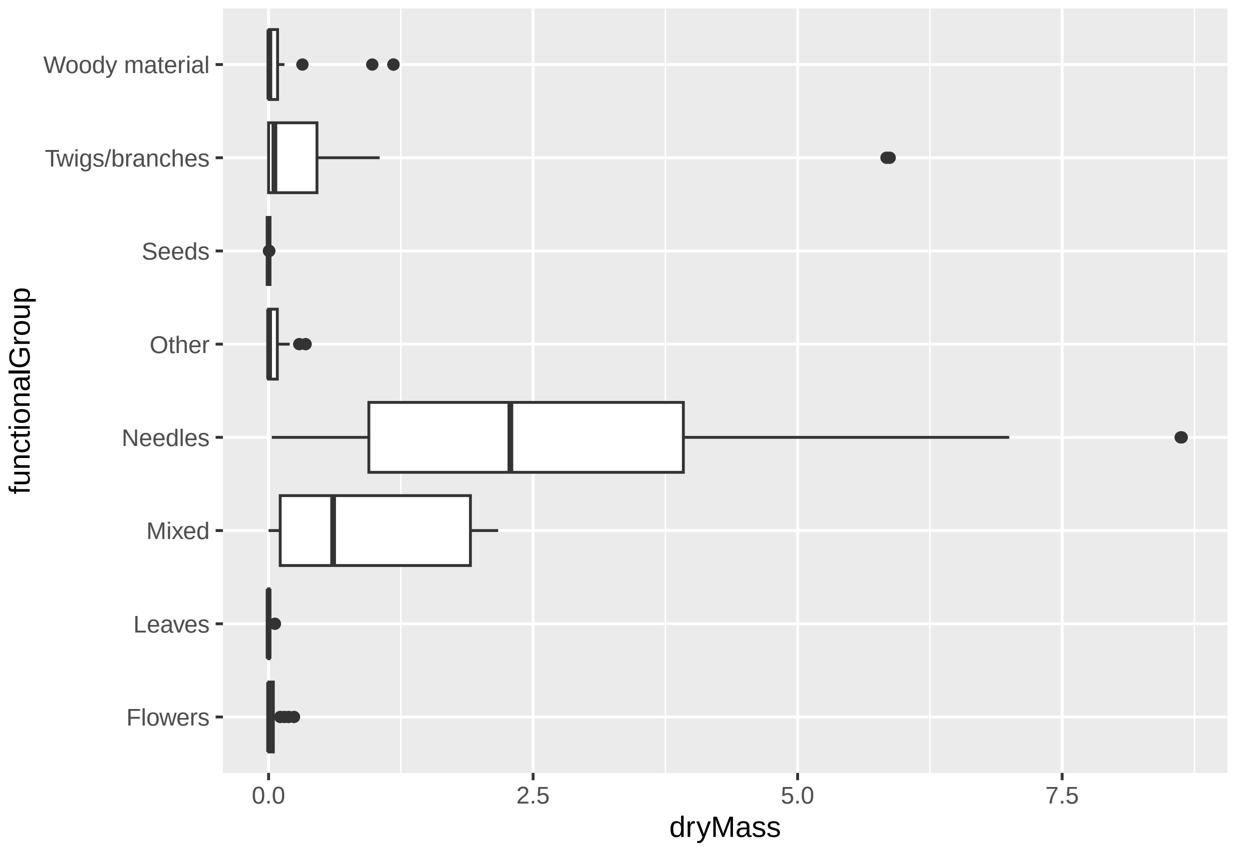
```r
ggplot(Litter)+
  geom_bar(aes(x=functionalGroup), color="purple")
```



```
#Here, I created a bar graph for the Litter dataset with the Functional Groups
#as the X-axis. I colored it purple for fun :)
```

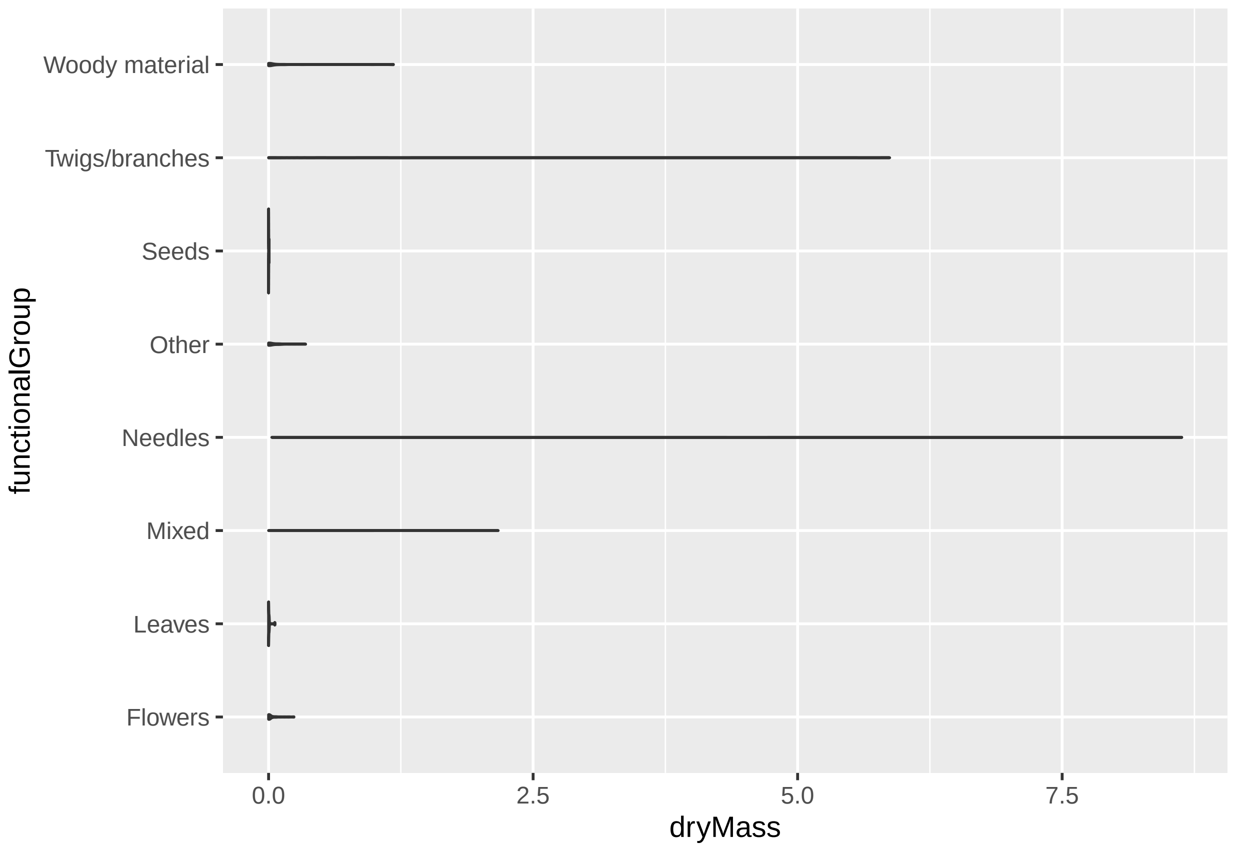
15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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



```
#Here, I created the boxplot comparing "dryMass" versus "functionalgroup"

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



```
#Here, I created the violin plot comparing "dryMass" versus "functionalgroup"
```

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

> Answer:The box plot is a more effective visualization than the violin plot because it more readily shows the detail in the data. The violin plot is difficult to understand because the different “violins” just look like thin, straight lines. Additionally, the violin plot for “twigs/branches” in particular is confusing since it stretches out rather far, but the box plot shows that the mass is actually quite low other than an outlier or two.

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

> Answer: According to the Box plot, it seems that Needles tend to have the highest biomass at these sights, followed by mixed biomass and twigs/branches.