Assignment 3: Data Exploration

Samriddha Ghosh

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
```

```
## [1] "D:/DUKE/EDA-Spring2023"
```

```
#the packages tidyverse and lubridate have been installed in the console
using the command:  install.packages{"tidyverse"} and
install.packages{"lubridate"}
```

```
library("tidyverse")
library("lubridate")
neon <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",
stringsAsFactors = TRUE)
str(neon)

## 'data.frame':    188 obs. of  19 variables:
##  $ uid                   : Factor w/ 188 levels "028eea3d-5c20-4afc-bb7e-
a05bab305152",..: 84 96 85 107 112 72 116 49 124 119 ...
##  $ 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 14 14 20 20 ...
##  $ massSampleID          : Factor w/ 168 levels
"NEON.LTR.NIWO040205.20180802.FLR",..: 102 101 103 97 103 99 100 98 139 145
...
##  $ 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 ...

ecotoxico <-
read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",stringsAsFactors=
```

```
TRUE)
str(ecotoxico)

## 'data.frame':    4623 obs. of  30 variables:
##  $ CAS.Number                      : int  58842209 58842209 58842209
58842209 58842209 58842209 58842209 58842209 58842209 58842209 58842209 ...
##  $ Chemical.Name                   : Factor w/ 9 levels "(1E)-N-[(6-
Chloro-3-pyridinyl)methyl]-N'-cyano-N-methylethanimidamide",..: 9 9 9 9 9 9 9
9 9 9 ...
##  $ 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 4 ...
##  $ Chemical.Purity                 : Factor w/ 80 levels
">=98",">=99.0",..: 69 69 50 50 50 50 50 50 50 50 ...
##  $ Species.Scientific.Name         : Factor w/ 398 levels "Acalolepta
vastator",..: 69 69 248 248 248 248 248 248 248 248 ...
##  $ Species.Common.Name             : Factor w/ 303 levels "Alfalfa
Leafcutter Bee",..: 74 74 142 142 142 142 142 142 142 142 ...
##  $ 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 19 ...
##  $ Organism.Age                    : Factor w/ 39 levels
"~10","~24","~7",..: 39 39 39 39 39 36 39 36 36 39 ...
##  $ Organism.Age.Units              : Factor w/ 11 levels "Day(s)","Days
post-emergence",..: 9 9 4 4 4 1 4 1 1 4 ...
##  $ Exposure.Type                   : Factor w/ 24 levels
"Choice","Dermal",..: 23 23 11 11 11 11 11 11 11 11 ...
##  $ Media.Type                      : Factor w/ 10 levels
"Agar","Artificial soil",..: 7 7 3 3 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 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
"~10","~30/","~40/",..: 639 510 813 622 442 637 500 642 814 784 ...
##  $ Conc.1.Units..Author.           : Factor w/ 148 levels "%","% v/v","%
w/v",..: 132 132 91 91 91 91 91 91 91 91 ...
##  $ Effect                          : Factor w/ 19 levels
"Accumulation",..: 16 16 16 16 16 16 16 16 16 16 ...
##  $ Effect.Measurement              : Factor w/ 155 levels
"Abundance","Accuracy of learned task, performance",..: 87 87 87 87 87 87 87
87 87 87 ...
##  $ 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 14 14 14 ...
```

```
##  $ Observed.Duration..Days.        : Factor w/ 361 levels
"~.1458","~10",..: 145 145 145 145 145 145 145 145 145 145 ...
##  $ Observed.Duration.Units..Days.  : Factor w/ 17 levels "Day(s)","Day(s)
post-emergence",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Author                          : Factor w/ 433 levels "Abbott,V.A.,
J.L. Nadeau, H.A. Higo, and M.L. Winston",..: 66 66 181 181 181 181 181 181
181 181 ...
##  $ Reference.Number                : int  107388 107388 103312 103312
103312 103312 103312 103312 103312 103312 ...
##  $ Title                           : Factor w/ 458 levels "A Common
Pesticide Decreases Foraging Success and Survival in Honey Bees",..: 91 91
450 450 450 450 450 450 450 450 ...
##  $ Source                          : Factor w/ 456 levels "Acta
Hortic.1094:451-456",..: 295 295 296 296 296 296 296 296 296 296 ...
##  $ Publication.Year                : int  1982 1982 1986 1986 1986 1986
1986 1986 1986 1986 ...
##  $ Summary.of.Additional.Parameters: Factor w/ 943 levels "Purity: \xca NR
- NR | Organism Age: \xca NR - NR Not reported | Conc 1 (Author): \xca
Formulation NR/ - NR/ % "| __truncated__,..: 797 796 795 794 860 859 858 857
864 871 ...
```

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 one of the most often used insecticides in agriculture, and because of its effects on insects, especially beneficial insects like pollinators, there may be interest in the ecotoxicology of neonicotinoids on insects. Since insects are so important to ecosystem function, any harm done to their numbers may have long-lasting effects on biodiversity. Numerous insects, including honeybees, bumblebees, and other pollinators are hazardous to neonicotinoids, according to research. Since both domesticated plants and wild plants depend on these insects for pollination, their extinction could result in lower yields and a decrease in the variety of plants. In addition to harming beneficial insects, neonicotinoids may also unintentionally harm non-target species like birds, aquatic insects, and other creatures that consume insects. Ecosystems may suffer long-term consequences as a result of neonicotinoids' environmental persistence and capacity to pollute soil and water. In order to improve agricultural operations and safeguard the environment, it is crucial to study the ecotoxicology of neonicotinoids on insects. For this, the dataset from the ECOTOX Knowledgebase of the Environmental Protection Agency is quite helpful.

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: 1. Understanding Forest Ecosystem Functioning: Litter and woody debris are an integral part of the forest ecosystem, and studying them can provide important insights into the functioning of the ecosystem. For example, litter and woody debris play a role in soil formation and nutrient cycling, and provide habitat for a variety of plant and animal species. 2. Carbon Sequestration: Forests play an important role in removing carbon dioxide from the atmosphere, and litter and woody debris are an important part of the carbon cycle. By studying the amount and type of litter and woody debris that accumulates in a forest, researchers can get a better understanding of the role forests play in carbon sequestration. 3. Monitoring Forest Health: Changes in the amount and type of litter and woody debris in a forest can be an indicator of forest health. For example, a decrease in the amount of litter and woody debris can indicate that a forest is under stress from factors such as disease, insect infestation, or climate change. These are just a few of the reasons why litter and woody debris are an important area of study in forest ecology. By studying litter and woody debris, researchers can gain a better understanding of the functioning of forest ecosystems, the role forests play in the global carbon cycle, and the impacts of land use changes on forest ecosystems.

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: A.Spatial Sampling and Temporal Sampling are used to sample litter and woody debris as part of the NEON network

B. Salient information about the sampling methods: 1.In sites with forested tower airsheds, the litter sampling is targeted to take place in 20 40m x 40m plots. In sites with low-statured vegetation over the tower airsheds, litter sampling is targeted to take place in 4 40m x 40m tower plots (to accommodate co-located soil sampling) plus 26 20m x 20m plots. 2.Target sampling frequency for elevated traps varies by vegetation present at the site. 3.At sites with 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?

```
#Showing the dimensions of the Data-set named neon
dim(neon)

## [1] 188  19

nrow(neon)   #specifying the row numbers
```

```
## [1] 188

ncol(neon)   #specifying the column numbers

## [1] 19

#Showing the dimensions of the Data-set named ecotoxico
dim(ecotoxico)

## [1] 4623   30

nrow(ecotoxico)  #specifying the row numbers

## [1] 4623

ncol(ecotoxico)  #specifying the column numbers

## [1] 30
```

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

```
column_vector<-ecotoxico$Effect
summary(column_vector)

##     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 is Mortality, followed by Population. Since insects are so important to ecosystem function, any harm done to their numbers may have long-lasting effects on biodiversity. Hence, mortality of these chemicals and their ability to cause population changes in insects are observed.

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…]

```
common_name_vector<-ecotoxico$Species.Common.Name
summary(common_name_vector)

##                          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: Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee and Italian Honeybee are the most commonly studied specied in the dataset. Common characteristic: All of these species play important roles in the

ecosystem, either as pollinators of crops and wildflowers or as predators and parasites of other insects. Reason for studying: These bees being most effective pollinators, have greater impact on the ecosystem if harmed by insecticides, Hence, are they are most commonly studied compared to the other insects.

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(ecotoxico$Conc.1..Author.)
```

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

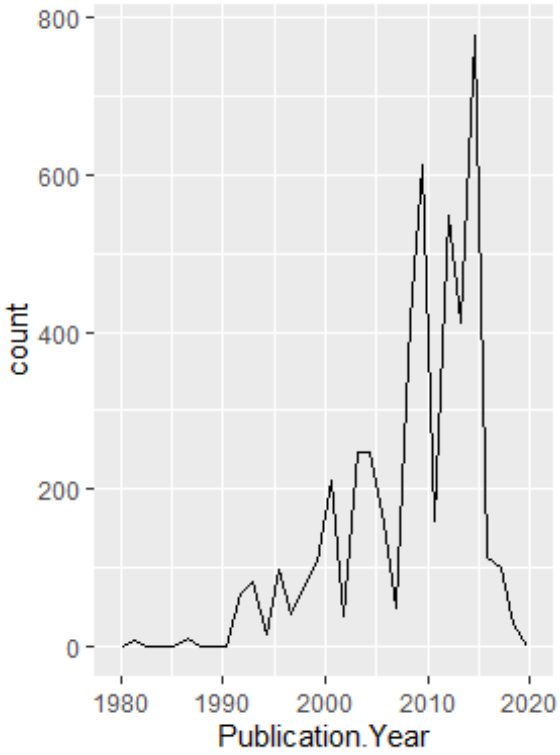
Answer:In our assignment, the argument stringsAsFactors = TRUE is used in the read.csv() function to specify whether character vectors in the data frame should be converted to factors. By default, stringsAsFactors = TRUE, which means that character vectors in the data frame will be automatically converted to factors. Therefore, when we use the class() to return the class type, factor is returned instead of numeric.

Explore your data graphically (Neonics)

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

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

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

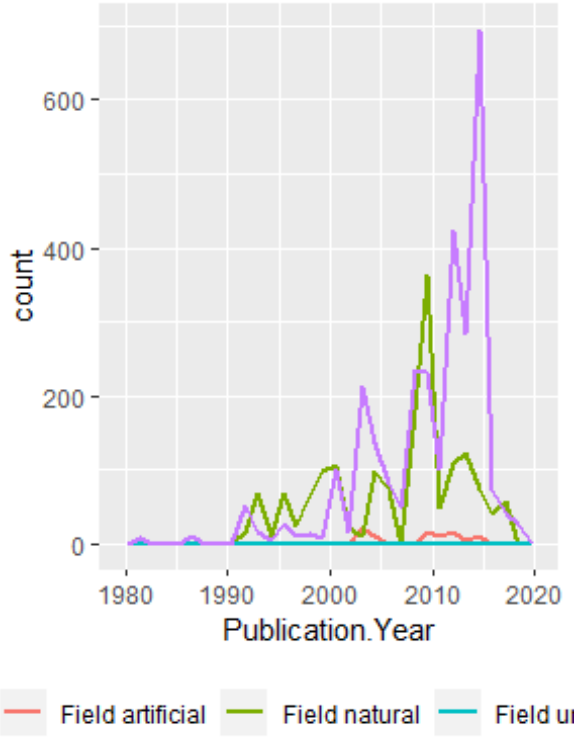


10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
ggplot(data = ecotoxico, mapping = aes(x = Publication.Year, colour =
Test.Location)) +
  geom_freqpoly(size=1) +
  theme(legend.position = "bottom")

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.

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



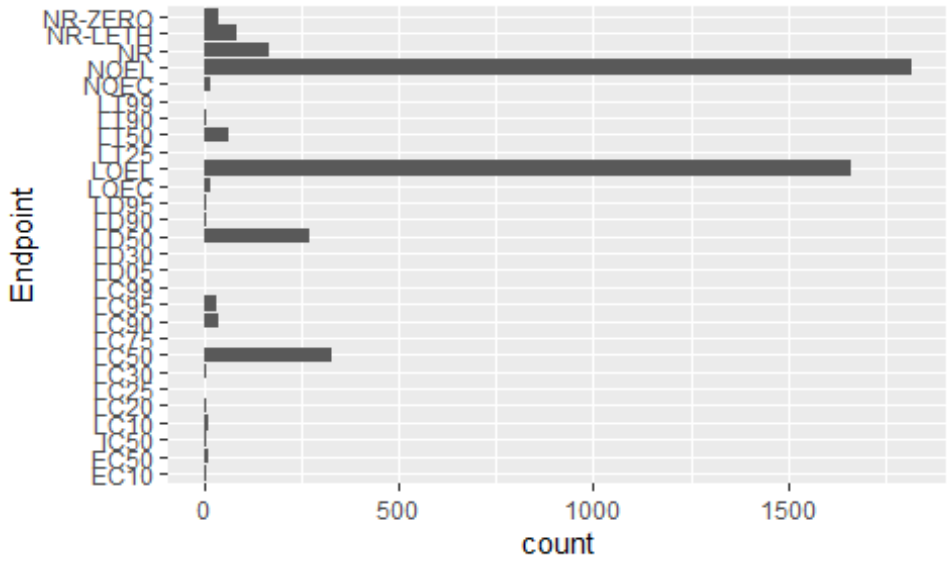
Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: Lab and Field nature. Yes, with time lab tests keep increasing by several manifolds compared to any other test locations.

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(ecotoxico, aes(y = Endpoint)) +
  geom_bar()
```



Answer: From the graph it is visible that NOEL = 1816; LOEL = 1664 are the two most common end points. 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.

```
class(neon$collectDate) # returns factor hence need to be changed to date format

## [1] "factor"

class(neon$collectDate)

## [1] "factor"

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

## [1] "Date"
```

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`?

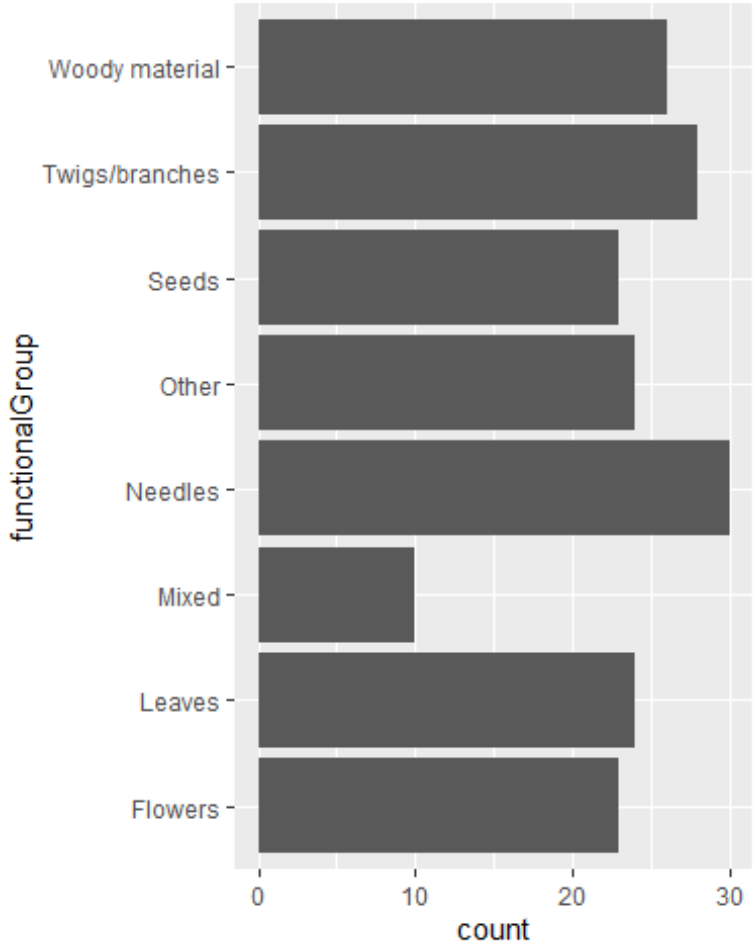
```
unique(neon$plotID)
```

```
##  [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
```

Answer:12 plots were sampled at Niwot Ridge. In a nutshell, the unique function is used to extract unique values from a vector, while the summary function is used to provide a summary of the values in a vector or data frame.

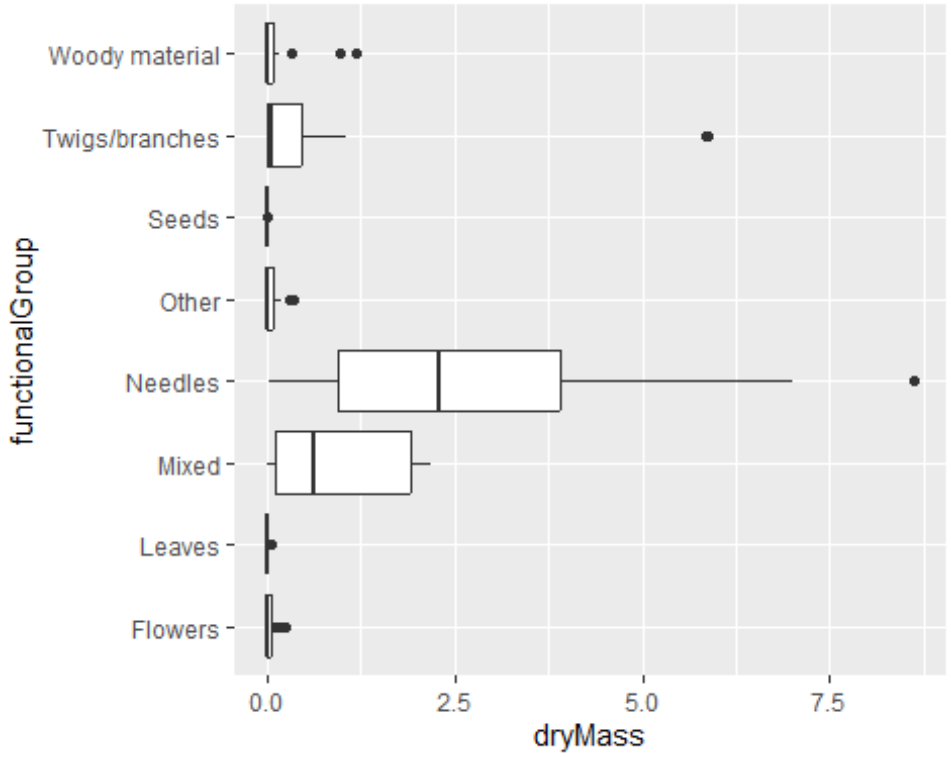
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(neon, aes(y = functionalGroup)) +
  geom_bar()
```

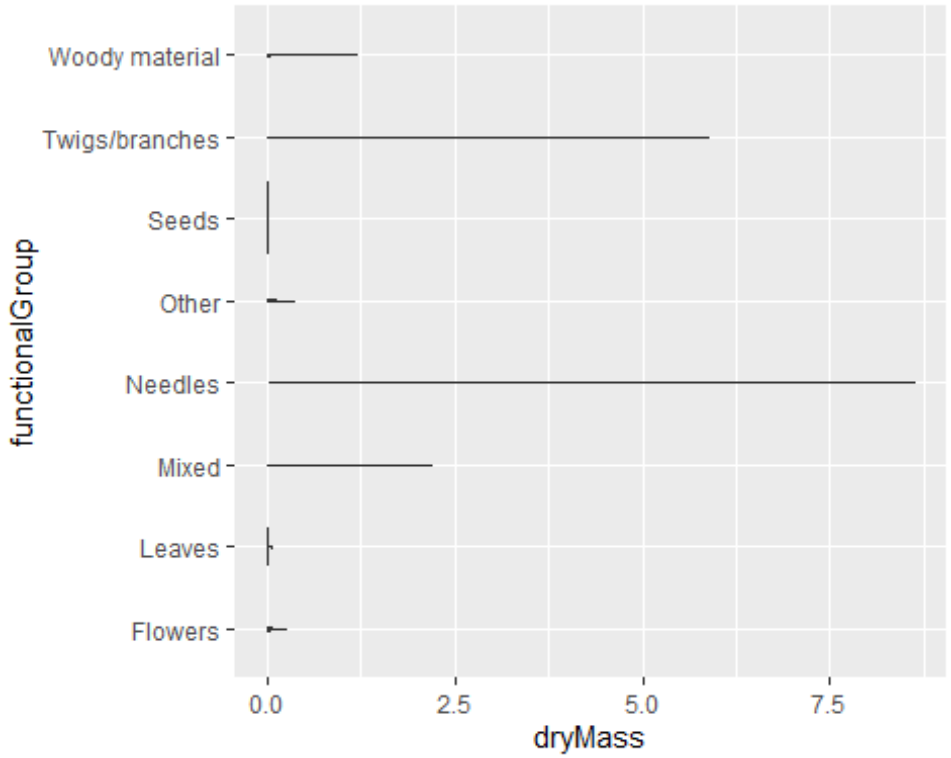


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




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



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

    Answer: A box plot is another way of showing the relationship between a numeric variable and a categorical variable. Compared to the violin plot, the box plot leans more on summarization of the data, primarily just reporting a set of descriptive statistics for the numeric values on each categorical level.In particular, the whiskers depict the distribution of the data, the dots highlight outliers, and the boxes display the median and interquantile range. The boxplot provides context, but the violin plot merely displays the range of data for each category, making it difficult to understand.

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

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