

Assignment 3: Data Exploration

Hanbin Lyu

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 sub-command to read strings in as factors.

```
library(tidyverse)
#load package
library(lubridate)
#load package
library(here)
#load package
Neonics = read.csv("../Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv")
#upload file
Litter = read.csv("../Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv")
#upload file
```

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: we should study the ecotoxicology of neonicotinoids on insects because most pollinators are insects, which are essential for the reproduction of many plants, including crops that humans rely on for food. The long-lasting presence of neonicotinoids means they can harm pollinators even if applied months before blooming (Understanding Neonicotinoids, n.d.).

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 are crucial to forest and stream ecosystems as it influences carbon budgets and nutrient cycling, serves as an energy source for aquatic systems, provides habitat for both terrestrial and aquatic species, and adds structure that affects water flow and sediment movement (Scheungrab et al., 2000).

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: litter and woody debris are sampled in the NEON network by collecting dry weights from litter traps according to specific plant functional types. The data collected from these sampling events are initially considered raw data (Level 0) and are then processed and quality checked to create a quality-controlled product (Level 1). This process involves automated quality assurance and control procedures, and detailed metadata is provided alongside the data for publication through the NEON data portal (Jones & Flagg, n.d.). 1."In sites with >50% aerial cover of woody vegetation >2m in height, placement of litter traps is random, using a randomized list of grid cell locations. In sites with <50% woody vegetation cover or patchy vegetation, trap placement is targeted under qualifying vegetation (Jones & Flagg, n.d.)." 2." Ground traps are sampled once per year, while elevated traps are sampled frequently (1x every 2 weeks) in deciduous forest sites during senescence, and infrequently (1x every 1-2 months) at evergreen sites (Jones & Flagg, n.d.)." 3."Woody vegetation cover, as measured by NEON's Airborne Observation Platform and/or vegetation structure protocols, may be used to scale up litterfall production from point measurements (Jones & Flagg, n.d.)."

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
view(as.data.frame(Neonics))
view(as.data.frame(Litter))
#turn two data sets into data frames that is clearer to check
```

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...]

```
common_effects = table(Neonics$Effect)
#make a table for occurrence number of each type of effect
sort(common_effects, decreasing = TRUE)
```

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

```
#rearrange the table in descending order
```

Answer: the most common effect is population. Studying the effects of neonicotinoids on population, mortality, reproduction, feeding behavior, and development is vital because these factors influence insect survival and biodiversity, which are essential for pollination and food security. Understanding these impacts helps identify potential disruptions in ecosystems that affect species interactions and ecosystem functions.

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...]

```
species = table(Neonics$Species.Common.Name)
#make a table for the occurrence number of each species's common name
sort(species, decreasing = TRUE)
```

```
##
##      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 Woolly 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		Asiatic Honey Bee
##	9		9
##	Eulophid Parasitoid		Lacewing Family
##	9		9
##	Mealybug Destroyer		Alfalfa Leafcutter Bee
##	9		8
##	Bee		Bumblebee
##	8		8
##	Chilean Predatory Mite		Dwarf Honey Bee
##	8		8
##	Neotropical Stingless Bee		Parasitic Wasp Family
##	8		8
##	Spiralling Whitefly		Beetle Mite Family
##	8		7
##	Chinch Bug		Macedonian Honey Bee
##	7		7
##	Moth		Potato Tuberworm
##	7		7
##	Russian Wheat Aphid		Soldier Beetle
##	7		7
##	Southern One-Year Canegrub		Tarnished Plant Bug
##	7		7
##	Ambrosia Beetle		Aphid Wasp

##	6	6
##	Black Vine Weevil	Childers Canegrub
##	6	6
##	Coconut Leaf Beetle	Eleven-spotted Ladybird Beetle
##	6	6
##	Encyrtid Wasp	European Red Mite
##	6	6
##	Fall Armyworm	Fruit Fly
##	6	6
##	Hover Fly	Oblique Banded Leaf Roller
##	6	6
##	Obscure Mealybug	Oribatid Mite Suborder
##	6	6
##	Pistachio Psyllid	Redbay Ambrosia Beetle
##	6	6
##	Silverleaf Whitefly	Soybean Aphid
##	6	6
##	Subterranean Termite	Thrip
##	6	6
##	Two-Spotted Spider Mite	Apple Aphid
##	6	5
##	Brown Planthopper	Earwig
##	5	5
##	Green June Beetle	Hornfaced Bee
##	5	5
##	Long Horned Beetle Family	Plum Curculio
##	5	5
##	Rove Beetle	San Jose Scale
##	5	5
##	Scelionid Wasp	Speckled Cutworm Moth
##	5	5
##	Thrip Family	Ant
##	5	4
##	Cabbage Seedpod Weevil	Common Green Lacewing
##	4	4
##	Eucalyptus Gall Wasp	European Apple Sawfly
##	4	4
##	European Honey Bee	European Tarnished Plant Bug
##	4	4
##	Garden Symphylan	Linyphiid Spider
##	4	4
##	Onion Maggot	Oriental Beetle
##	4	4
##	Parsnip Seed Wasp	Pea And Bean Weevil
##	4	4
##	Pear Sucker	Red Imported Fire Ant
##	4	4
##	Striped Cucumber Beetle	Sugarcane Beetle
##	4	4
##	Wasp	Wolf Spider Family
##	4	4
##	Yellow-faced Bumblebee	Ambrosia Bark Beetle
##	4	3
##	Asian Ambrosia Beetle	Beetle Family

##	3	3
##	Birch Leafminer	Black Twig Borer
##	3	3
##	Braconid Parasitoid Wasp	California Red Scale
##	3	3
##	Crucifer Flea Beetle	Cutworm
##	3	3
##	Delphacid Planthopper	Egyptian Cotton Leafworm
##	3	3
##	Encyrtid Parasitoid	Fly/Mosquito/Midge Order
##	3	3
##	Formosan Subterranean Termite	Fruit-tree Pinhole Borer
##	3	3
##	Green Rice Leafhopper	Ground Beetle
##	3	3
##	Ichneumonid Wasp	Large-Jawed Orb Weaver Family
##	3	3
##	Leaf Cutting Ant	Mediterranean Fruit Fly
##	3	3
##	Minute Flour Bug	Mite Family
##	3	3
##	Moth Family	Negatoria Canegrub
##	3	3
##	Sap Beetle Family	Scale Insect Order
##	3	3
##	Scarab Beetle Family	Sheet-Web Weaver Family
##	3	3
##	Spider	Sugarcane Grub
##	3	3
##	Tenebrionid Beetle	Alfalfa Plant Bug
##	3	2
##	Alkali Bee	Aphid
##	2	2
##	Assassin Bug	Azalea Lace Bug
##	2	2
##	Banana Aphid	Brown Scale
##	2	2
##	Brown Stinkbug	Budworm
##	2	2
##	Cabbage Aphid	Cabbage White
##	2	2
##	Cardamom Thrip	Carrot Weevil
##	2	2
##	Celer Crab Spider	Centipede Class
##	2	2
##	Citricola Scale	Clouded Plant Bug
##	2	2
##	Coffee Bean Weevil	Cotton Fleahopper
##	2	2
##	Egyptian Alfalfa Weevil	Engraver Beetle
##	2	2
##	Fig Longicorn Beetle	Glassy-winged Sharpshooter
##	2	2
##	Hawthorn Lace Bug	Hister Beetle Family

##		2		2
##	Jumping Spider Family		Lined Click Beetle	
##		2		2
##	Maple Spider Mite		Meshweaver Spider	
##		2		2
##	Minute Pirate Bug Family		Predaceous Fly	
##		2		2
##	Pygmy Mangold Beetle		Rose Sawfly	
##		2		2
##	Serpentine Leafminer		Spider Mite Destroyer	
##		2		2
##	Spotted Tentiform Leafminer		Stink Bug	
##		2		2
##	Tawny Mole Cricket		Tick/Chigger/Mite Order	
##		2		2
##	Turf Running-spider		Turnip Aphid	
##		2		2
##	Western Bigeyed Bug		Western Damsel Bug	
##		2		2
##	Western Plant Bug		White Apple Leafhopper Nymph	
##		2		2
##	White-backed Planthopper		Whitemarked Fleahopper	
##		2		2
##	Antlike Flower Beetle		Banded Soft-winged Flower Beetle	
##		1		1
##	Banded Sunflower Moth		Bee Family	
##		1		1
##	Beet Armyworm		Black Citrus Aphid	
##		1		1
##	Blue Alfalfa Aphid		Cabbage Root Fly	
##		1		1
##	Cactus Lady Beetle		Citrus Red Mite	
##		1		1
##	Cottony Cushion Sale		Crapemyrtle Aphid	
##		1		1
##	Damselbug Family		Ectoparasitoid Wasp	
##		1		1
##	English Grain Aphid		Fairyfly	
##		1		1
##	Flea Beetle		Gall Midge	
##		1		1
##	Grasshopper/Cricket/Locust Order		Greenhouse Whitefly	
##		1		1
##	Grey Sunflower Seed Weevil		Harvestman Spider Order	
##		1		1
##	Hawthorn Leaf Miner		Longtailed Fruit Fly Parasite	
##		1		1
##	Minute Lady Beetles		Painted Maple Aphid	
##		1		1
##	Pepper Weevil		Pine False Webworm	
##		1		1
##	Plant Bug		Pollen Beetle	
##		1		1
##	Predacious Mite		Predator Bug	

##		1		1
##	Pseudocentipede Class		Pteromalid Wasp Family	
##		1		1
##	Red Sunflower Seed Weevil		Rice Leaf Folder Moth	
##		1		1
##	Rose Grain Aphid		Scale Picnic Beetle	
##		1		1
##	Shiny Spider Beetle		Southern Army Worm	
##		1		1
##	Spirea Aphid		Spotted Sunflower Stem Weevil	
##		1		1
##	Strawberry Blossom Weevil		Sunflower Midge	
##		1		1
##	Sunflower Moth		Ten-spot Ladybird Beetle	
##		1		1
##	Tobacco Thrip		Twicestabbed Lady Beetle	
##		1		1
##	Wasp Family		Weevil	
##		1		1
##	Yellow Mealworm Beetle			
##		1		

#rearrange the table in descending order

Answer: the six most commonly studied species are Honey Bee, Parasitic Wasp, Buff Tailed Bumble Bee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. People are more interested in bees because they can all directly impact the reproduction of flowering plants and the overall health of ecosystems. Their activities support biodiversity and help maintain the balance of various habitats.

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] "character"
```

#check the class of `Conc.1..Author.`

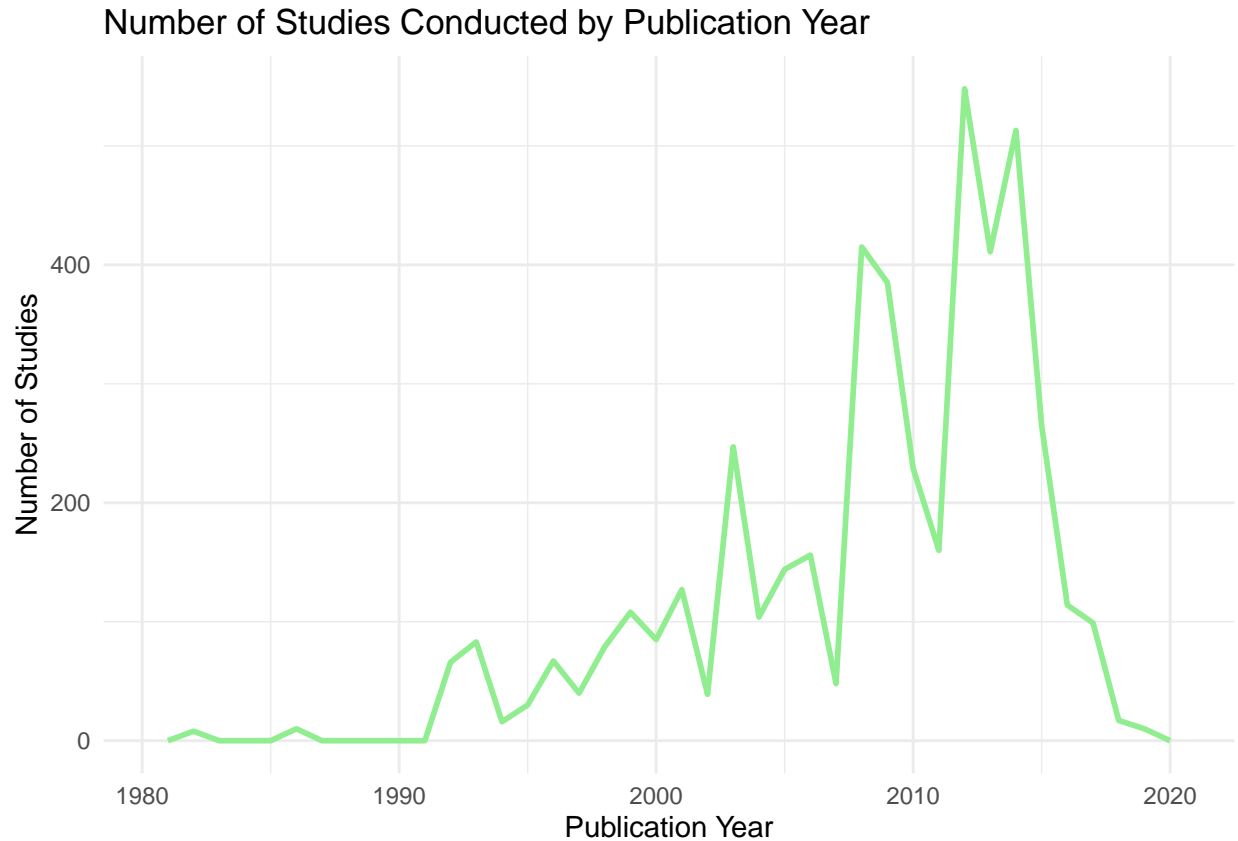
Answer: the class of ‘Conc.1..Author.’ is character. It is not numeric value because it contains some letters and symbols like “/” and “NR”.

Explore your data graphically (Neonics)

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

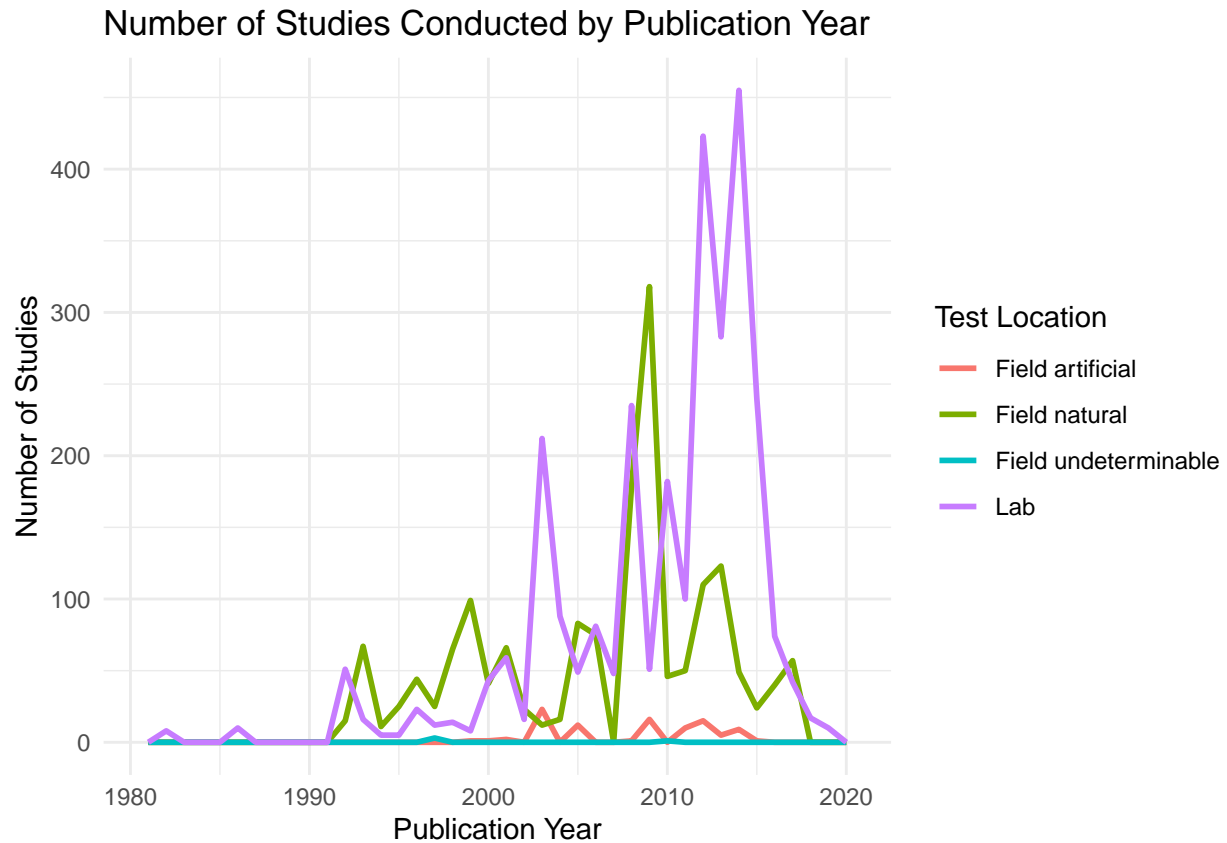
```
library(ggplot2)
#load the package
ggplot(data = Neonics, aes(x = Publication.Year)) + #select the data
  geom_freqpoly(binwidth = 1, color = "palegreen2", linewidth = 1) + #determine details
```

```
labs(title = "Number of Studies Conducted by Publication Year",
     x = "Publication Year",
     y = "Number of Studies") + #name the plot
theme_minimal() #choose plot theme
```



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

```
ggplot(data = Neonics, aes(x = Publication.Year, color = Test.Location)) + #select data
geom_freqpoly(binwidth = 1, linewidth = 1) + #determine details
labs(title = "Number of Studies Conducted by Publication Year",
     x = "Publication Year",
     y = "Number of Studies") + #name the plot
theme_minimal() + #choose plot theme
scale_color_discrete(name = "Test Location") #name the legend
```



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

Answer: in general, the lab is the most common test location. Before the year 2000, studies conducted in the field were more numerous than those done in the lab. Around 2014, studies conducted in the lab reached their peak. “Field undeterminable” is the least used test location, appearing almost as a line overlapping with the x-axis.

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.

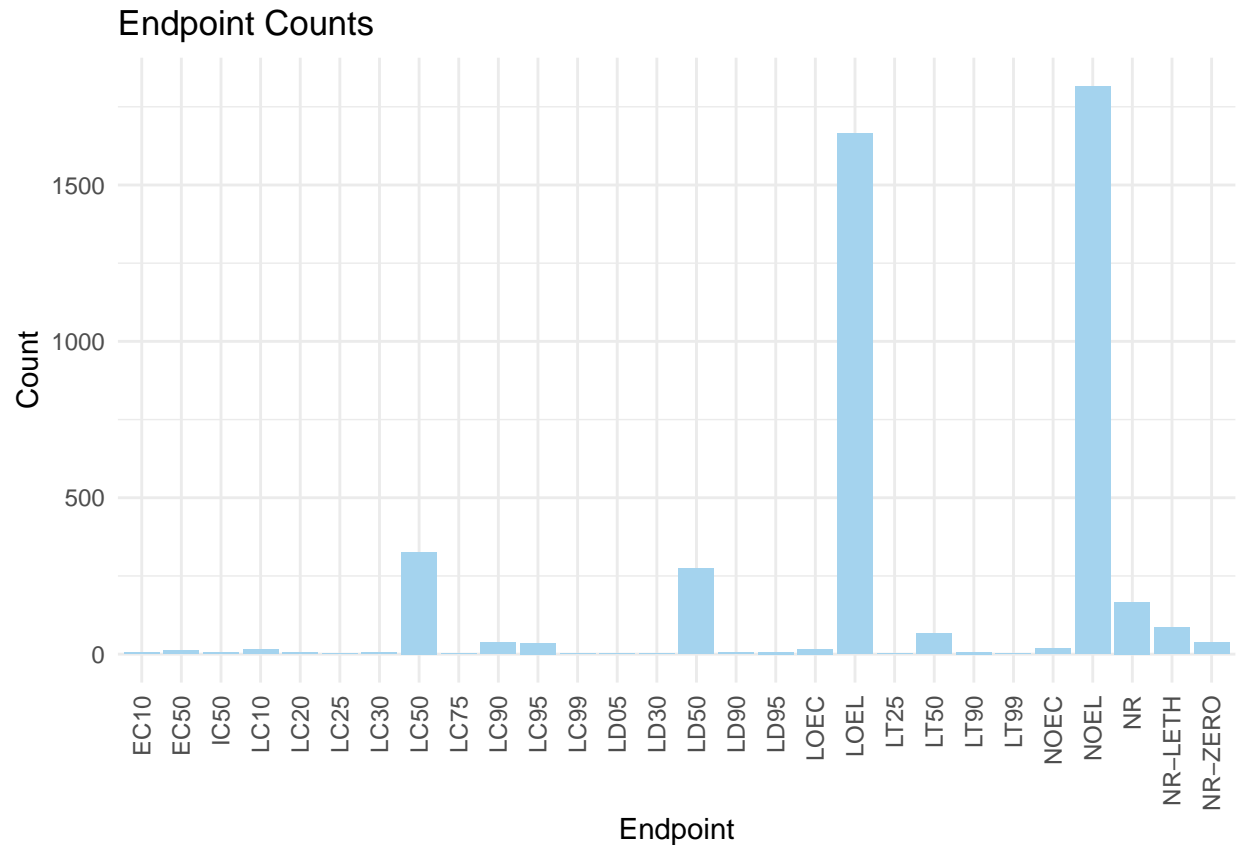
[**TIP:** Add `theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))` to the end of your plot command to rotate and align the X-axis labels...]

```
library(dplyr)
endpoint = Neonics %>% count(Endpoint)
#count the Endpoint
endpoint
```

##	Endpoint	n
## 1	EC10	6
## 2	EC50	11
## 3	IC50	6
## 4	LC10	15
## 5	LC20	5
## 6	LC25	1

```
## 7      LC30      6
## 8      LC50    327
## 9      LC75      1
## 10     LC90     37
## 11     LC95     36
## 12     LC99      2
## 13     LD05      1
## 14     LD30      1
## 15     LD50    274
## 16     LD90      6
## 17     LD95      7
## 18     LOEC     17
## 19     LOEL 1664
## 20     LT25      1
## 21     LT50     65
## 22     LT90      7
## 23     LT99      2
## 24     NOEC     19
## 25     NOEL 1816
## 26          NR   167
## 27  NR-LETH     86
## 28  NR-ZERO     37
```

```
ggplot(data = endpoint, aes(x = Endpoint, y = n)) + #select data
  geom_bar(stat = "identity", fill = "lightskyblue2") + #determine details
  labs(title = "Endpoint Counts",
        x = "Endpoint",
        y = "Count") + #name the plot
  theme_minimal() + #choose plot theme
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



#rotate and align the X-axis labels

Answer: the two most common endpoint are LOEL and NOEL. “Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC). 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) (ECOTOX_CodeAppendix).

Explore your data (Litter)

- 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] "character"
```

#check the class of collectDate

```
collect_date = as.Date(Litter$collectDate, format = "%Y-%m-%d")
```

#change the class to date

```
class(collect_date)
```

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

```
#check the class again
August_2018 = unique(collect_date[format(collect_date, "%Y-%m") == "2018-08"])
#determine which dates litter was sampled in August 2018
August_2018
```

```
## [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`?

```
unique(Litter$siteID)
```

```
## [1] "NIWO"
```

```
#determine how many different plots were sampled at Niwot Ridge
summary(Litter$siteID)
```

```
##      Length      Class      Mode
##      188 character character
```

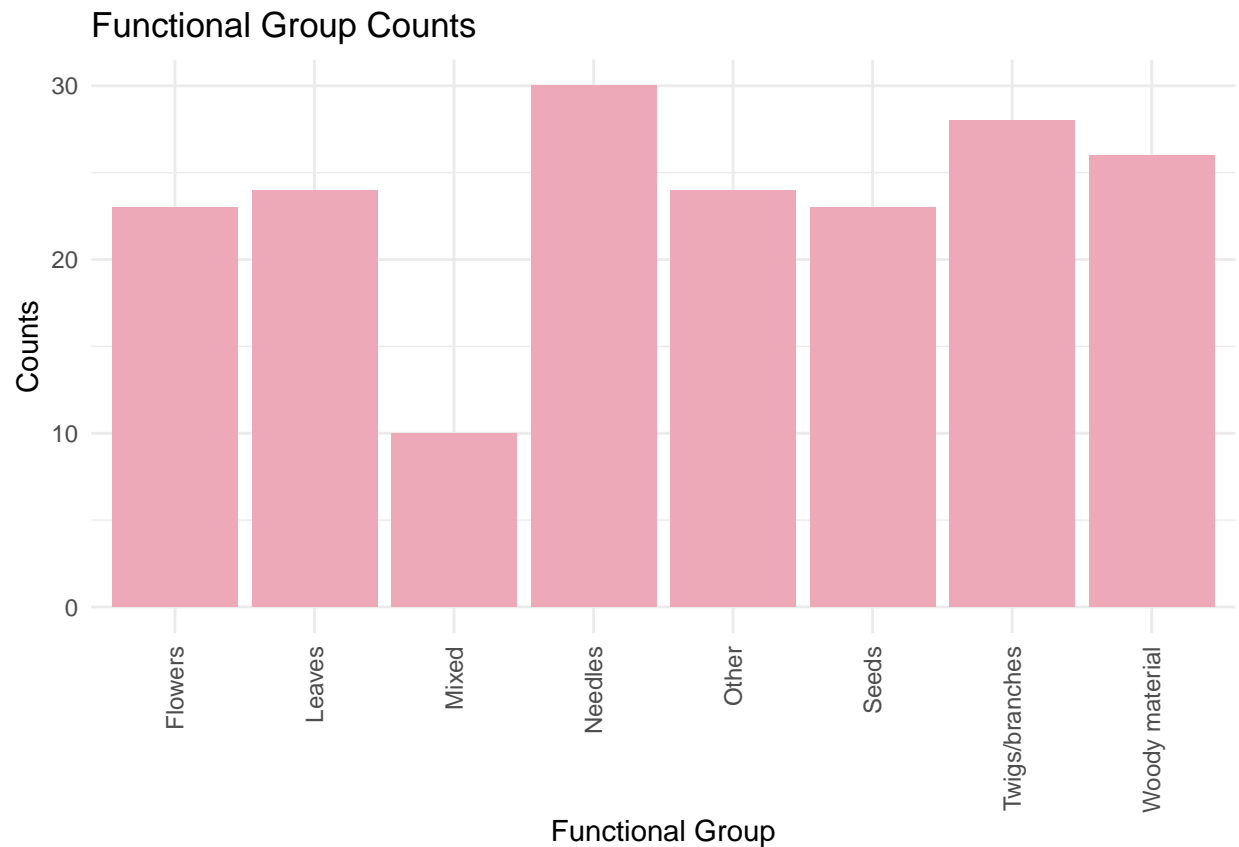
Answer: the result from ‘unique’ function is “NIWO”, which means all the plots were sampled at Niwot Ridge. We know there are 188 plots, so 188 different plots were sampled at NIWO. The result from ‘summary’ function shows the length, class, and mode.

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.

```
functional_group = Litter %>% count(functionalGroup)
#count the functionalGroup
functional_group
```

```
## functionalGroup n
## 1      Flowers 23
## 2      Leaves 24
## 3      Mixed 10
## 4      Needles 30
## 5      Other 24
## 6      Seeds 23
## 7 Twigs/branches 28
## 8 Woody material 26
```

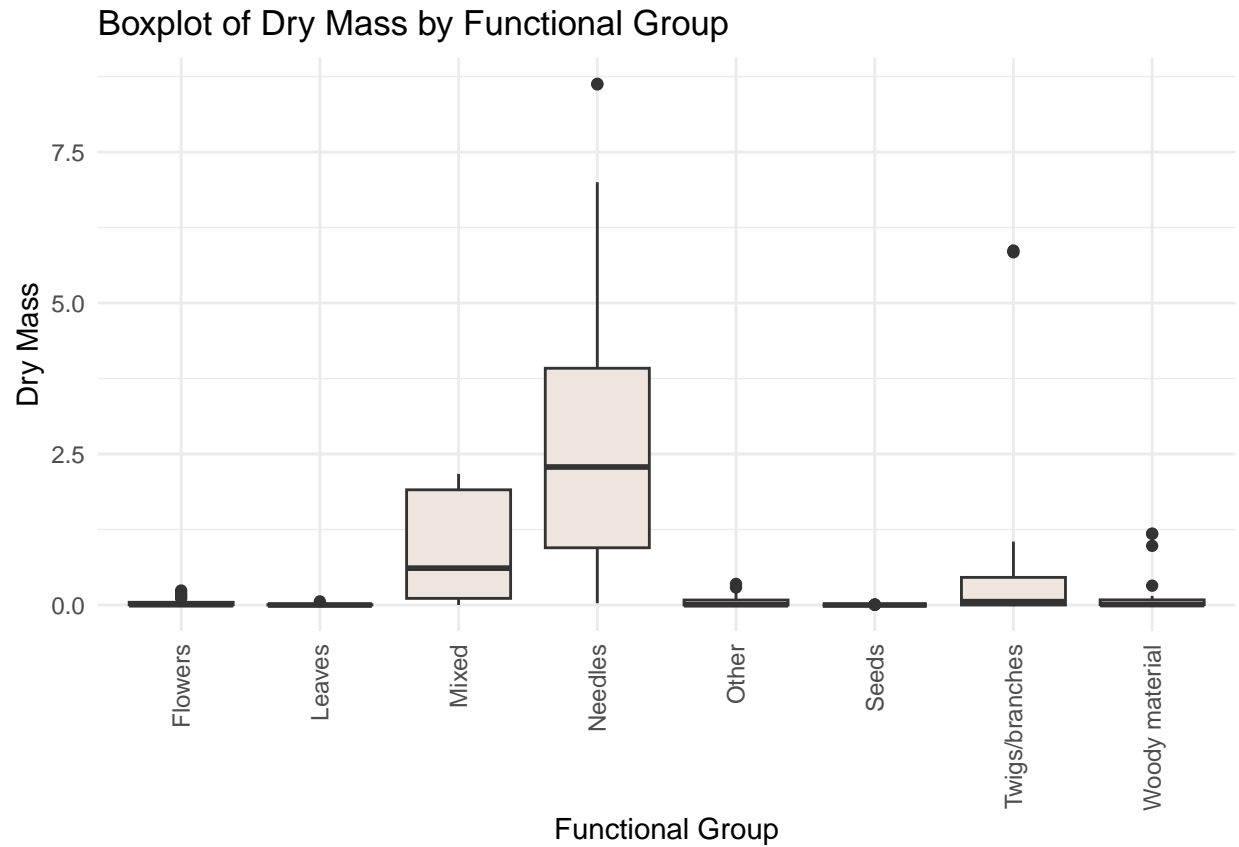
```
ggplot(data = functional_group, aes(x = functionalGroup, y = n)) + #select data
  geom_bar(stat = "identity", fill = "pink2") + #determine details
  labs(title = "Functional Group Counts",
       x = "Functional Group", y = "Counts") + #name the plot
  theme_minimal() + #choose plot theme
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



#rotate and align the X-axis labels

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

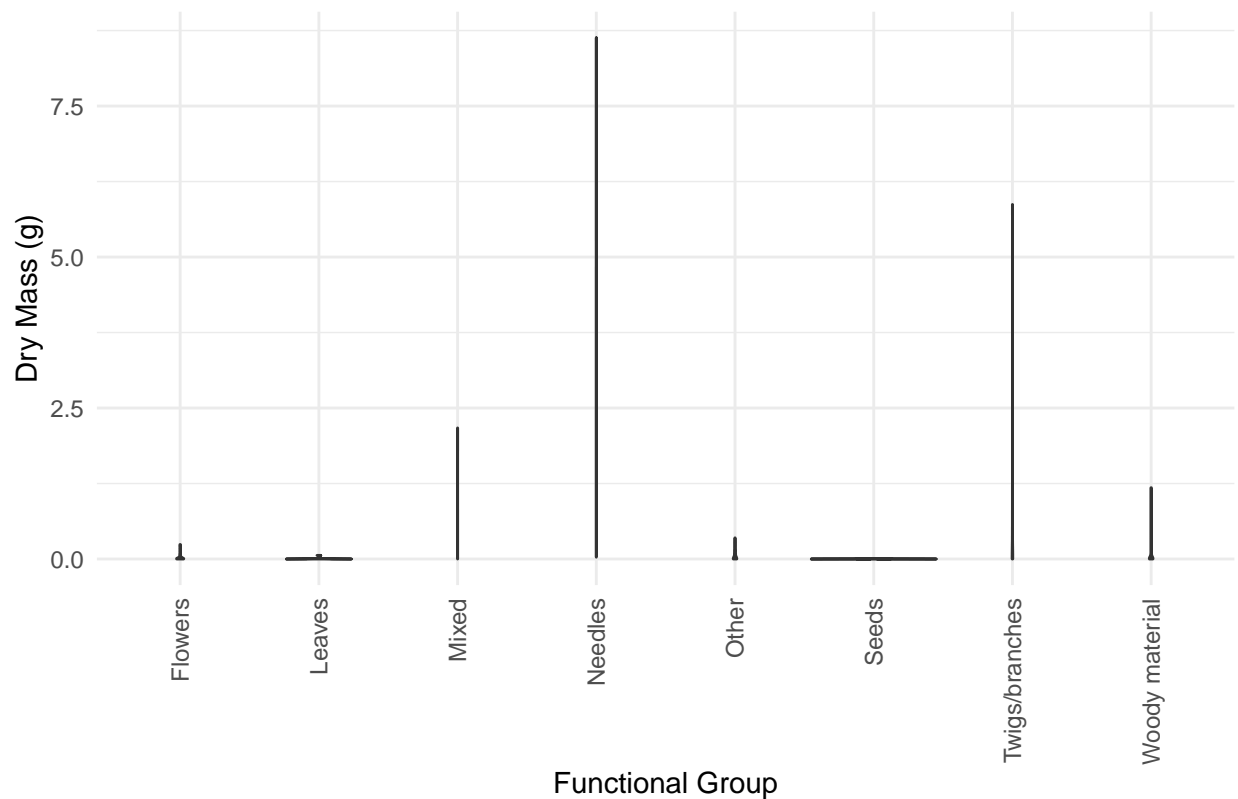
```
ggplot(data = Litter, aes(x = functionalGroup, y = dryMass)) + #select data
  geom_boxplot(fill = "seashell2") + #determine detail
  labs(title = "Boxplot of Dry Mass by Functional Group",
        x = "Functional Group", y = "Dry Mass") + #name the plot
  theme_minimal() + #choose plot theme
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



```
#rotate and align the X-axis labels

ggplot(data = Litter, aes(x = functionalGroup, y = dryMass)) + #select data
  geom_violin(fill = "indianred2") + #determine detail
  labs(title = "Violin Plot of Dry Mass by Functional Group",
        x = "Functional Group", y = "Dry Mass (g)") + #name the plot
  theme_minimal() + #choose plot theme
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```


Violin Plot of Dry Mass by Functional Group



#rotate and align the X-axis labels

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

Answer: in this case, many values are close to zero which means the variation of the data points is small, resulting in vertical lines and not look like the violin at all.

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

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