Assignment 3: Data Exploration

Asreeta Ushasri

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.

## Install Packages and Datasets

#install.packages("tidyverse")  
#install.packages("lubridate")  
#install.packages("here")  
  
library(tidyverse)  
library(lubridate)  
library(here)  
  
getwd()

## [1] "/home/guest/AU\_EDE\_Fall2024"

here()

## [1] "/home/guest/AU\_EDE\_Fall2024"

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)

## Learn about your system

1. 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: According to the Natural Resources Defense Council, neonic instectides indiscriminantly poison pests and ecologically important insects. For example, bees, butterflies and other pollinators have been killed by neonics in recent years. Neonics impact the nervous system of insects and impede immune systems. Neonics also remain in the soil for years after agricultural application. Studying the ecotoxicology of neonicotinoids on insects is crucial for determining the lasting impact of these strong pesticides, and whether researchers can protect important pollinators from the long-term harm of neonics (Lindwall, 2022).

1. 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: The USDA Forest Service describes woody litter as debris from fallen or decaying trees and plant life in forests (Scheungrab et al., 2000). This material is an important part of the carbon cycle and nutrient recycling in ecosystems. The quality of the woody debris can impact the quality of future soil in these ecosystems. Studying woody debris can help researchers determine if there are any harmful chemicals seeping into protected forests, or if there are any nutrient deficiencies in the soil system (Scheungrab et al., 2000).

1. 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: The NEON\_Litterfall\_UserGuide is located in the Metadata folder, under Data.

1. The litter debris samples are collected from elevated traps, and the fine wood debris samples are collected from ground traps (Jones, K. & Flagg, C., 2017, p. 3).
2. Elevated traps are sampled once every 1-2 months in evergreen forests, or once every two weeks in deciduous forests (Jones, K. & Flagg, C., 2017, p. 5).
3. Ground traps are sampled once a year (Jones, K. & Flagg, C., 2017, p. 5).

## Obtain basic summaries of your data (Neonics)

1. What are the dimensions of the dataset?

The Neonics dataset has 4623 rows and 30 columns. The Litter dataset has 188 rows and 19 columns.

## View Datasets

dim(Neonics)

## [1] 4623 30

#view(Neonics)  
  
  
dim(Litter)

## [1] 188 19

#view(Litter)

1. 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…]

## Effect Column in Neonics Dataset

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

sort(summary(Neonics$Effect))

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

summary(Litter$Effect)

## Length Class Mode   
## 0 NULL NULL

Answer: The population and mortality of insects are the most common effects studied in this dataset. Neonic pesticides are particularly concerning because they kill insects widely and indiscriminately, including pollinators (Lindwall, 2022). The mortality rate and population decrease of butterflies, bees, and ecologically important insects is one of the largest concerns with neonic pesticides. Therefore, studying population and mortality rates with respect to neonic insecticides is useful and critical for scientists.

1. 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 Common Name in Neonics Dataset

summary(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

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

Answer: The six most common species of insects in this data set include the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumblee Bee, and Italian Honeybee (excluding the other category, which has the largest observations). All of these species are a type of bee or wasp, which are important pollinator insects. Bees alone are responsible for pollinating a major sector of the world’s food supply (Bush, 2020). Bees and wasps are also currently facing an unprecedented population decline (Bush, 2020).

1. 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…]

## Concentration Author in Neonics Dataset

class(Neonics$Conc.1..Author)

## [1] "factor"

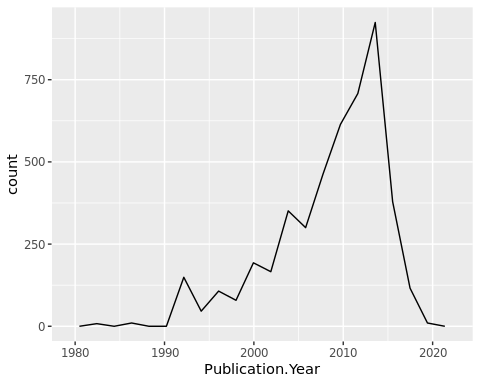
Answer: The Conc.1.Author class is a factor. This is not a numeric class because many values in the column have approximation symbols “~” or backslashes at the end of the digits “/”. Some rows also have NR as their value for this column.

## Explore your data graphically (Neonics)

1. Using geom\_freqpoly, generate a plot of the number of studies conducted by publication year.

## Neonics Sample Frequency by Publication Year

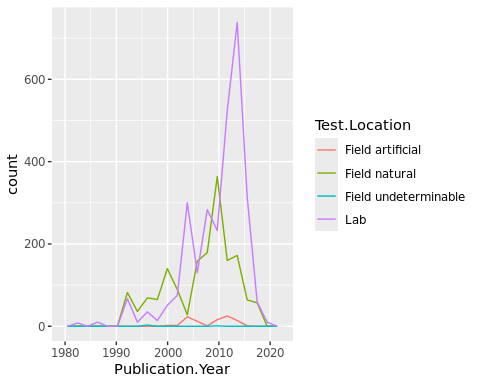
ggplot(Neonics) + geom\_freqpoly(aes(x = Publication.Year), bins = 20)



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

## Neonics Sample Frequency by Test Location

ggplot(Neonics) + geom\_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 20)



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

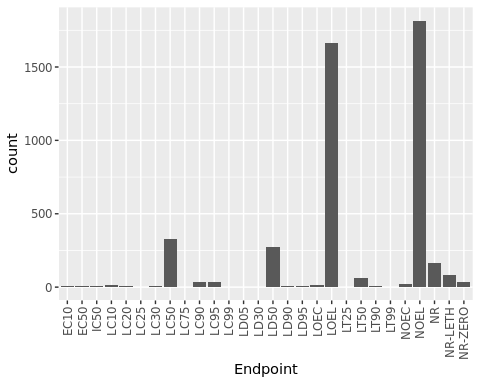
Answer: The most common test location is the lab, followed by the natural field. However, this trend differs over time. From 1990 - 2000, the natural field was the most common test location, and the lab was the second most popular test location. Starting in 2000, the lab became the most common test location, whereas the natural field became the second most popular test location. From 1980 - 2020, the artificial field and the undeterminable field were the least common test locations.

1. 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…]

## Count of Endpoints in Neonics Dataset, Bar Chart

ggplot(data = Neonics, aes(x = Endpoint)) +  
 geom\_bar() + theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))



Answer: The two most common endpoints are NOEL and LOEL. NOEL is defined as the “No-observable-effect-level” in the terrestrial database (GDIT, 2019, p. 723). LOEL is defined as the “Lowest-observable-effect-level” in the terrestrial database (GDIT, 2019, p. 722).

## Explore your data (Litter)

1. 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.

The litter was sampled on August 2, 2018 and August 30, 2018.

##Litter Sample Collection Date

class(Litter$collectDate)

## [1] "factor"

Litter$collectDate <- as.Date(Litter$collectDate, format = '%Y-%m-%d')  
  
Aug2018 <- unique(Litter$collectDate, 2018-08)  
  
Aug2018

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

1. 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?

## Litter Sample Plot Locations

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)

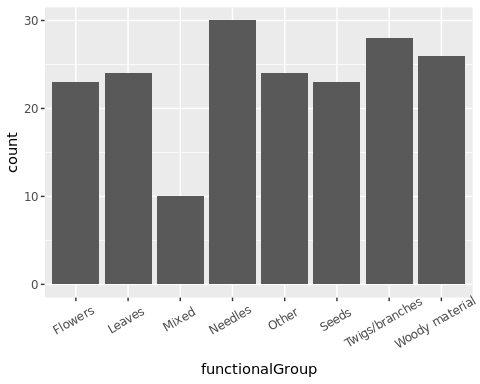
## NIWO\_040 NIWO\_041 NIWO\_046 NIWO\_047 NIWO\_051 NIWO\_057 NIWO\_058 NIWO\_061   
## 20 19 18 15 14 8 16 17   
## NIWO\_062 NIWO\_063 NIWO\_064 NIWO\_067   
## 14 14 16 17

Answer: The unique function provides the categories for the plot sites, showcasing twelve unique plot IDs. The summary function includes a count of how many samples were obtained at each of the twelve plot sites.

1. 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.

## Litter Count by Functional Groups, Bar Chart

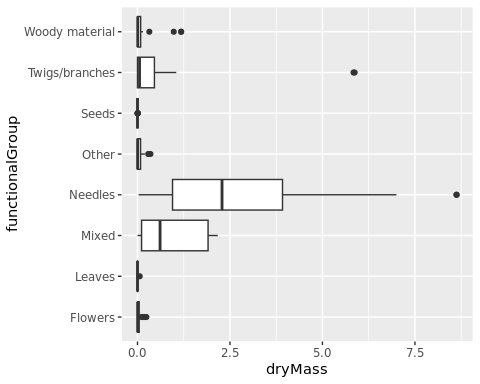
ggplot(data = Litter, aes(x = functionalGroup)) +  
 geom\_bar() + theme(axis.text.x = element\_text(angle = 30, vjust = 0.8, hjust=0.6))



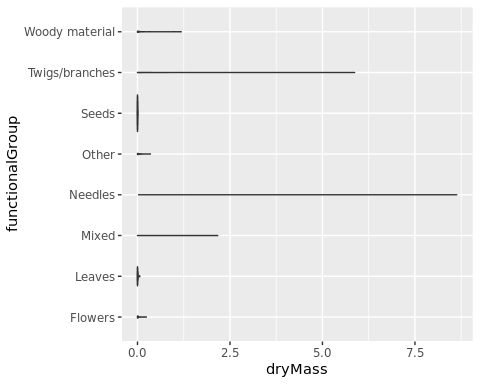
1. Using geom\_boxplot and geom\_violin, create a boxplot and a violin plot of dryMass by functionalGroup.

## Litter Dry Mass by Functional Group, Box Plot and Violin Plot

ggplot(Litter) +  
 geom\_boxplot(aes(x = dryMass, y = functionalGroup))



ggplot(Litter) +  
 geom\_violin(aes(x = dryMass, y = functionalGroup),   
 draw\_quantiles = c(0.25, 0.5, 0.75))



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

Answer: The violin plots include the dryMass outliers for each functional group. Therefore, the violin plots appear as a straight line, with no visual indication of the quartile values. The boxplot is more effective in this case because it separates the outliers from the quartile summaries of each category.

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

Answer: The needles functional group has the highest median and third quartile value for dry mass, indicating that this group tends to have the highest biomass. The mixed functional group tends to have the second highest biomass, shown with the second highest median. Although the twigs/branches functional group has an outlier with a biomass above 5.0, the median and thrid quartile values are below the mixed functional group. In general, the twigs/branches has a lower biomass than the mixed group, which makes the twigs/branches the third highest biomass category. The other categories are fairly similar in having a median biomass value under 1.0.

References

Lindwall, C. (2022). Neonicotinoids 101: The Effects on Humans and Bees. National Resources Defense Council. <https://www.nrdc.org/stories/neonicotinoids-101-effects-humans-and-bees>

Scheungrab, Donna B.; Trettin, Carl C.; Lea, Russ; Jurgensen, Martin F. 2000. Woody debris. In: Gen. Tech. Rep. SRS-38. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. p. 47-48.

Bush, G. (2020) How you can keep bees from becoming endangered. Ohio State Impact. <https://www.osu.edu/impact/research-and-innovation/bee-population>

General Dynamics Information Technology (GDIT). (2019). ECOTOXicology Knowledgebase System, Ecotox Code Appendix.

Jones, K. & Flagg, C. (2017). NEON User Guide to Litterfall and Fine Woody Debris Sampling (NEON.DP1.10033). Revision A.