Assignment 3: Data Exploration

Vicky Jia

## OVERVIEW

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

## Directions

1. Change “Student Name” on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., “Salk\_A03\_DataExploration.Rmd”) prior to submission.

The completed exercise is due on Tuesday, January 28 at 1:00 pm.

## Set up your R session

1. Check your working directory, load necessary packages (tidyverse), 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.

getwd()

## [1] "/Users/Victoria/Environmental\_Data\_Analytics\_2020/Assignments"

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.5.2

## Warning: package 'ggplot2' was built under R version 3.5.2

## Warning: package 'tibble' was built under R version 3.5.2

## Warning: package 'tidyr' was built under R version 3.5.2

## Warning: package 'purrr' was built under R version 3.5.2

## Warning: package 'dplyr' was built under R version 3.5.2

## Warning: package 'stringr' was built under R version 3.5.2

## Warning: package 'forcats' was built under R version 3.5.2

Neonics <- read.csv("../Data/Raw/ECOTOX\_Neonicotinoids\_Insects\_raw.csv")  
Litter <- read.csv("../Data/Raw/NEON\_NIWO\_Litter\_massdata\_2018-08\_raw.csv")

## 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 ecotoxicologoy of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

Answer: According to the scientific research, neonicotinoids are widedly applied into agricultural industry due to its relatively low risk for nontarget organisms and the environment and high-target specificity to insects. In order to efficiently increase toxicity to insects while minimizing the toxicity to human beings, it is necessary to study ecotoxicology of neonicotinoids as different neonicotinoid compounds have been selected based on their specific subtypes of nicotinic receptors that occur in insects. This dataset includes data about chemicals, neonicotinoids, and species groups, which are major information to understand the major components of the neonicotinoids and then to study the effects of neonicotinoids on insects as well as to assess if their are appropriate to use for agriculculture.

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: According to the existed studies, oody debris can greatly contribute to the detrital biomass in forests, freshwaters, and coastal marine areas. Dead trees and woody debris remained in woodlands can recycle nutrients trapped in the wood and provide food and habitat for a wide range of organisms, and then improve biodiversity. Understanding the locations, weight, periods, and types of litter and woody debris allows scientists evaluate the appropriate amount of debris in the area to improve the ecosystem.

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: *Spatial sampling - different land cover types of tower airsheds have the different targeted litter sample plot sizes. For example, forested tower airsheds target the sampling in 20 40m x 40m plots, while the low-statured vegetation over the tower airsheds targets the lier sampling in 4 40m x 40m tower plots plus 26 20m x 20m plots. Trap placement within plots are also varied based on vegetation types.* Temporal sampling - Target sampling frequency for elevated traps are differenet based on vegation types. Specifically, the frequent sampling takes 1x every 2 weeks in deciduous forest sites during senescence, while infrequent year-round sampling takes 1x every 1-2 months at evergreen sites. \*For this product, no single site exceed 3,440 data product instances in a given calendar year as each ground trap may have up to two func􏰁onal groups and is sampled once per year

## Obtain basic summaries of your data (Neonics)

1. What are the dimensions of the dataset?

dim(Neonics)

## [1] 4623 30

1. Using the summary function, determine the most common effects that are studied. Why might these effects specifically be of interest?

summary(Neonics$Effect)

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

Answer: The most common effect is Population. The population indicates the change of amount of effected insects after using neonicotinoids. This is a significant criteria to assess if the use of neonicotinoids will critically destroy the ecosystem balance.

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.

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

Answer: Six most commonly studied speices in the dataset are Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. Bees are widely studied because neonicotinoids are commonly used on farms and in urban landscapes. Such insecticides can be absorbed by plants and then present in pollen and nectar, which could be toxic to bees.

1. Concentrations are always a numeric value. What is the class of Conc.1..Author. in the dataset, and why is it not numeric?

class(Neonics$Conc.1..Author.)

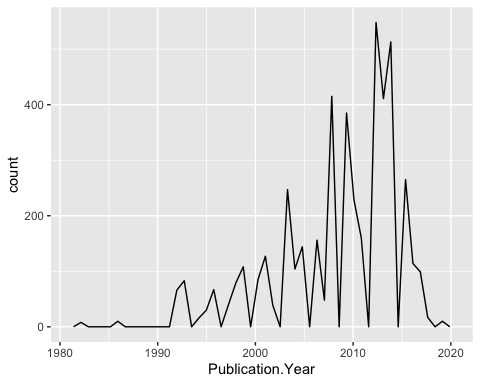
## [1] "factor"

Answer: The class of Conc.1..Author is factor in the dataset. It is not numeric because not all values for this concentration column are integers. It also includes characteristics like NR, estimated values like 4/, and other values. Factors are built on top of integer vectors making the value differently from regular integer vectors and defining the set of allowed values.

## Explore your data graphically (Neonics)

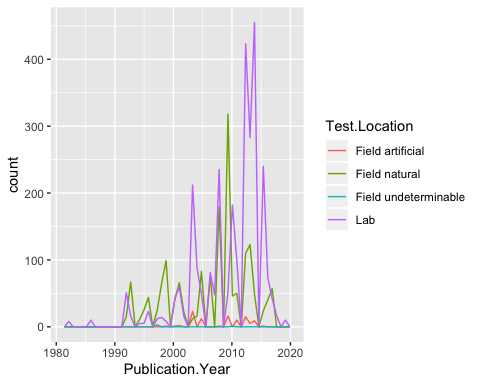
1. Using geom\_freqpoly, generate a plot of the number of studies conducted by publication year.

ggplot(Neonics) +  
 geom\_freqpoly(aes(x = Publication.Year), bins = 50)



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

ggplot(Neonics) +  
 geom\_freqpoly(aes(x = Publication.Year, color = Test.Location ), bins = 50)

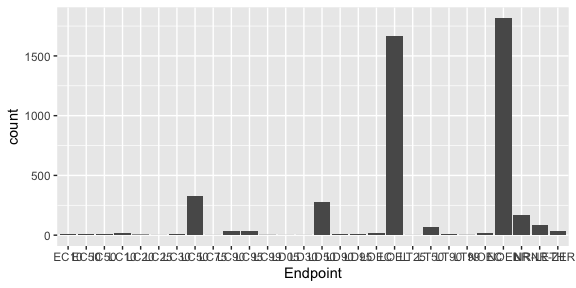


Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: This graph shows that the frequency of using different test locations over time, and also illustrates the trend of test locations changing over time. Lab and Field natural are most common test locations. However, before 2000, field natural was more commonly used as a test location; in early 2000s, lab was more common to use; then late 2000s to 2010, field natural was more commonly used again. Since 2010, lab has been the most common test location for publications.

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.

ggplot(Neonics, aes(x = Endpoint)) +  
 geom\_bar()



Answer: Two most common end points are NOEL and LOEL. NOEL is No-observable-effect-level; it means highest dose (concentration) producing effects not significantly different from responses of controls according to author’s reported statistical test (NOEAL/NOEC). LOEL is Lowest-observable-effect-level; it means lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC).

## 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.

class(Litter$collectDate)

## [1] "factor"

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

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

1. Using the unique function, determine how many plots were sampled at Niwot Ridge. How is the information obtained from unique different from that obtained from summary?

unique(Litter$plotID)

## [1] NIWO\_061 NIWO\_064 NIWO\_067 NIWO\_040 NIWO\_041 NIWO\_063 NIWO\_047  
## [8] NIWO\_051 NIWO\_058 NIWO\_046 NIWO\_062 NIWO\_057  
## 12 Levels: NIWO\_040 NIWO\_041 NIWO\_046 NIWO\_047 NIWO\_051 ... NIWO\_067

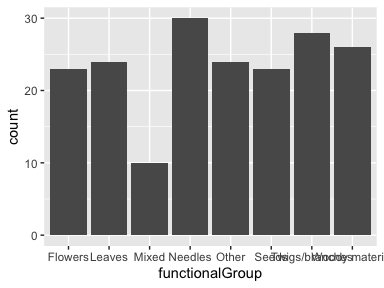
summary(Litter$plotID)

## 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: 12 plots were sampled at Niwot Ridge. ‘unique’ only returns plot ID and returns the number of return variables indicated by the ‘[]’ value in front of each line; ‘summary’ also returns plot ID, and it can tell the number of sites for each plot, but cannot indicate the total number of returned variables.

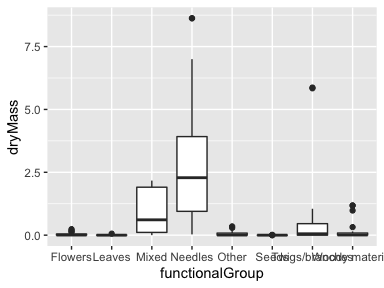
1. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

ggplot(Litter, aes(x = functionalGroup)) +  
 geom\_bar()

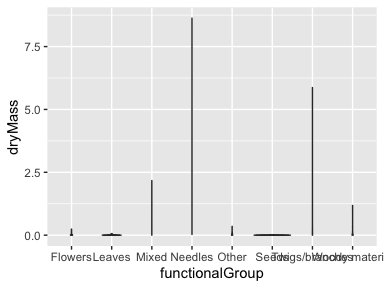


1. Using geom\_boxplot and geom\_violin, create a boxplot and a violin plot of dryMass by functionalGroup.

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



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



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

Answer: In this dataset, dry mass values for some litter types are concentrated at a lower range, but for others they are very dispersed. It is hard to set all function groups at the same scale in the visualization. For the situation of the diverse ranges for different groups, boxplot which could illustrate the outliers, median, and quantiles is more effective at visualization.

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

Answer: Needles