Assignment 3: Data Exploration

Laura Brockington

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.

```
#checking that my working directory is in the right folder
getwd()
```

```
## [1] "/Users/laura/Desktop/EDA/EDA"
```

```
#installing packages
#install.packages("tidyverse")
#install.packages("lubridate")

#loading in packages
```

```
library("tidyverse") #loading in packages
library("lubridate")
library("ggplot2")

#loading in the two csv files
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",
                    header=TRUE, stringsAsFactors = TRUE)
Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",
                   header=TRUE, stringsAsFactors = TRUE)
```

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: Insects are extremely important in the agriculture industry. They pollinate crops so we can have food and they manage pests that destroy said crops. Therefore, it's important for us to understand how the pesticides, such as neonictinoids, we use affect both beneficial and harmful insects.

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: The amount and type of forest litter and woody debris is an indicator of the health of a forest. It can also help us estimate the amount of carbon held in the forest.

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. One litter trap pair (one elevated trap and one ground trap) is deployed for every 400 m2 plot area, resuling in 1-4 trap pairs per plot 2. Ground traps are sampled once per year 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?

```
#checking the dimensions, 4623 rows with 30 columns
dim(Neonics)
```

```
## [1] 4623   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?

```
#getting a summary of the "Effect" column
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 two most commmon effects studied by far are popluation and mortality. These are likely of interest since they can have a big impact on crop yield.

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...]

```
#determining the six most commonly studied species
sort(summary(Neonics$Species.Common.Name), decreasing=TRUE)
```

```
##                          (Other)                       Honey Bee
##                              670                             667
##                   Parasitic Wasp           Buff Tailed Bumblebee
##                              285                             183
##               Carniolan Honey Bee                      Bumble Bee
##                              152                             140
##                  Italian Honeybee                  Japanese Beetle
##                              113                              94
##                Asian Lady Beetle                  Euonymus Scale
##                               76                              75
##                         Wireworm               European Dark Bee
##                               69                              66
##                Minute Pirate Bug             Asian Citrus Psyllid
##                               62                              60
##                   Parastic Wasp           Colorado Potato Beetle
##                               58                              57
##                  Parasitoid Wasp             Erythrina Gall Wasp
##                               51                              49
##                     Beetle Order       Snout Beetle Family, Weevil
##                               47                              47
##          Sevenspotted Lady Beetle                  True Bug Order
##                               46                              45
##             Buff-tailed Bumblebee                    Aphid Family
##                               39                              38
```

```
##                   Cabbage Looper               Sweetpotato Whitefly
##                               38                                 37
##                    Braconid Wasp                     Cotton Aphid
##                               33                                 33
##                   Predatory Mite              Ladybird Beetle Family
##                               33                                 30
##                       Parasitoid                     Scarab Beetle
##                               30                                 29
##                    Spring Tiphia                      Thrip Order
##                               29                                 29
##              Ground Beetle Family                Rove Beetle Family
##                               27                                 27
##                   Tobacco Aphid                     Chalcid Wasp
##                               27                                 25
##             Convergent Lady Beetle                    Stingless Bee
##                               25                                 25
##                 Spider/Mite Class                Tobacco Flea Beetle
##                               24                                 24
##                  Citrus Leafminer                   Ladybird Beetle
##                               23                                 23
##                        Mason Bee                         Mosquito
##                               22                                 22
##                    Argentine Ant                            Beetle
##                               21                                 21
##          Flatheaded Appletree Borer               Horned Oak Gall Wasp
##                               20                                 20
##                 Leaf Beetle Family                 Potato Leafhopper
##                               20                                 20
##          Tooth-necked Fungus Beetle                     Codling Moth
##                               20                                 19
##           Black-spotted Lady Beetle                     Calico Scale
##                               18                                 18
##                Fairyfly Parasitoid                      Lady Beetle
##                               18                                 18
##             Minute Parasitic Wasps                        Mirid Bug
##                               18                                 18
##                  Mulberry Pyralid                         Silkworm
##                               18                                 18
##                     Vedalia Beetle              Araneoid Spider Order
##                               18                                 17
##                         Bee Order                    Egg Parasitoid
##                               17                                 17
##                      Insect Class           Moth And Butterfly Order
##                               17                                 17
##       Oystershell Scale Parasitoid Hemlock Woolly Adelgid Lady Beetle
##                               17                                 16
##              Hemlock Wooly Adelgid                               Mite
##                               16                                 16
##                       Onion Thrip              Western Flower Thrips
##                               16                                 15
##                      Corn Earworm                 Green Peach Aphid
##                               14                                 14
##                         House Fly                          Ox Beetle
##                               14                                 14
```

```
##                 Red Scale Parasite                 Spined Soldier Bug
##                                 14                                 14
##             Armoured Scale Family                  Diamondback Moth
##                                 13                                 13
##                     Eulophid Wasp                 Monarch Butterfly
##                                 13                                 13
##                     Predatory Bug              Yellow Fever Mosquito
##                                 13                                 13
##               Braconid Parasitoid                    Common Thrip
##                                 12                                 12
##       Eastern Subterranean Termite                           Jassid
##                                 12                                 12
##                         Mite Order                      Pea Aphid
##                                 12                                 12
##                   Pond Wolf Spider          Spotless Ladybird Beetle
##                                 12                                 11
##             Glasshouse Potato Wasp                       Lacewing
##                                 10                                 10
##            Southern House Mosquito           Two Spotted Lady Beetle
##                                 10                                 10
##                         Ant Family                   Apple Maggot
##                                  9                                  9
```

Answer: The six most commonly studied species other than "other" are Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, Italian Honey Bee. These are all common pollinators and are therefore of particilar importance to the agriculture industry.

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?

```
#finding the class of the Conc.1..Author column
class(Neonics$Conc.1..Author.)
```

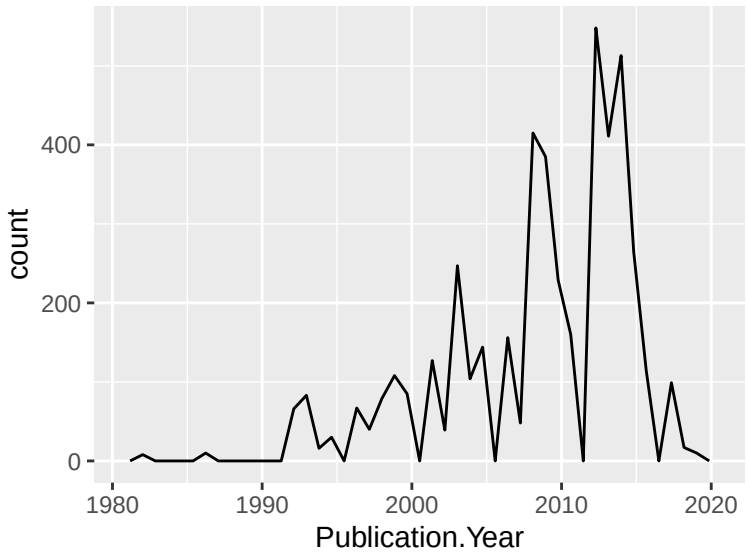
```
## [1] "factor"
```

Answer: The class of the 'Conc.1..Author.' column is factor. It is not numeric because it includes categories that are not numerical, such as NR, therefore is is designated as categorical data or a factor.

Explore your data graphically (Neonics)

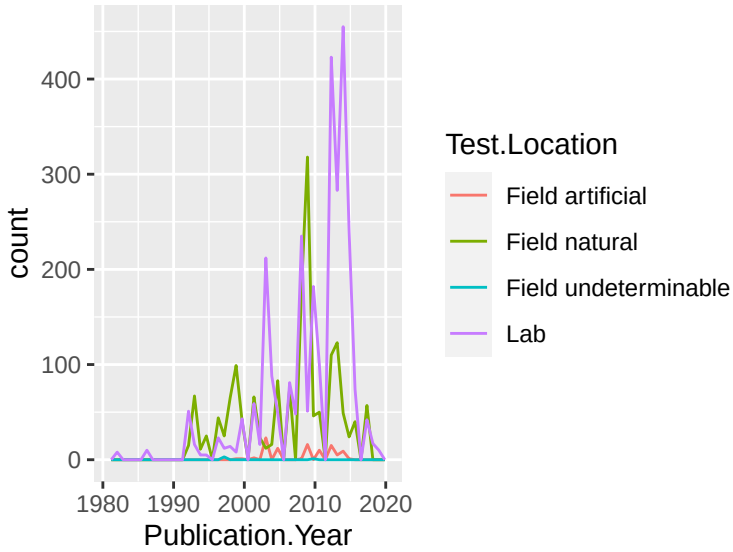
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
#generating a plot of the number of studies conducted by publication year
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins=45)
```



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

```
#generating a plot of the number of studies conducted by publication year
#colored by test location
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins=45)
```



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

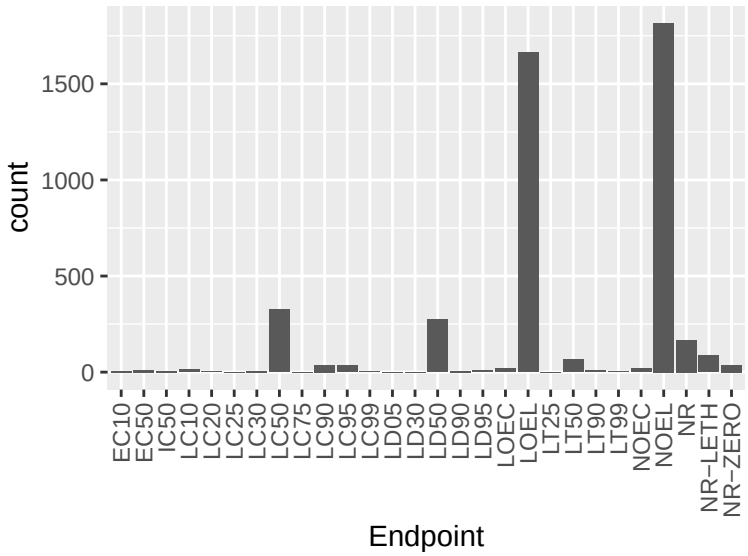
Answer: The most common test locations are in the lab, however between about 1990 and 2000 and 2007 and 2010, field natural were, on average, more common test locations.

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...]

```
#generating a bar graph of the number of each endpoint type.
ggplot(Neonics) +
  geom_histogram(aes(x = Endpoint), stat="count") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

```
## Warning in geom_histogram(aes(x = Endpoint), stat = "count"): Ignoring unknown
## parameters: `binwidth`, `bins`, and `pad`
```



Answer: The two most common endpoints are LOEL and NOEL. LOEL is lowest observable effect level, or the lowest dose producing effects that were significantly different from responses of controls. NOEL is no observable effect level, or the highest dose producing effects not significantly different from responses of controls according to author's reported statistical test.

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.

```
#finding the class of the collectDate column
class(Litter$collectDate)
```

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

```
#changing the class from factor to date
Litter$collectDate <- ymd(Litter$collectDate)
class(Litter$collectDate)
```

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

```
#determining which dates litter was sampled in Aug 2018
unique(Litter$collectDate)
```

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

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`?

```
#determing how many plots were sampled from Niwot Ridge
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)
```

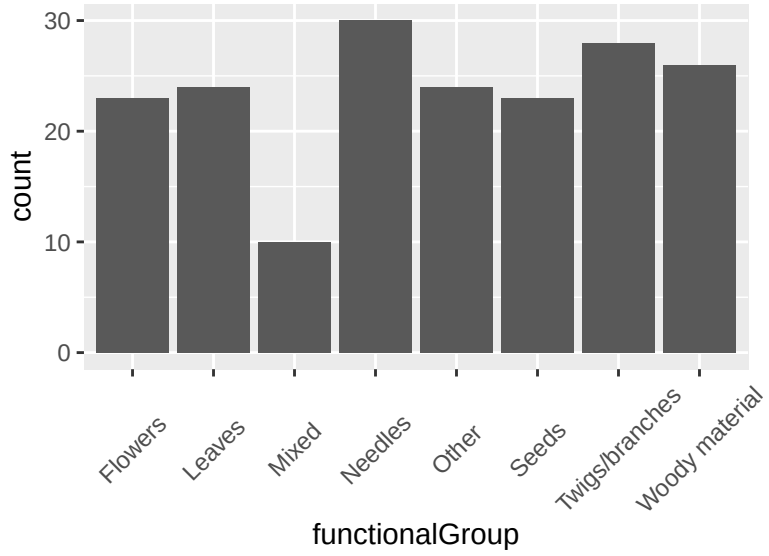
```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

Answer: 12 plots were sampled at the Niwot Ridge. The 'unique' function tell us how many different values were under plotID, essentially giving us the number of plots. The 'summary' function tells us the number of rows in each plotID.

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.

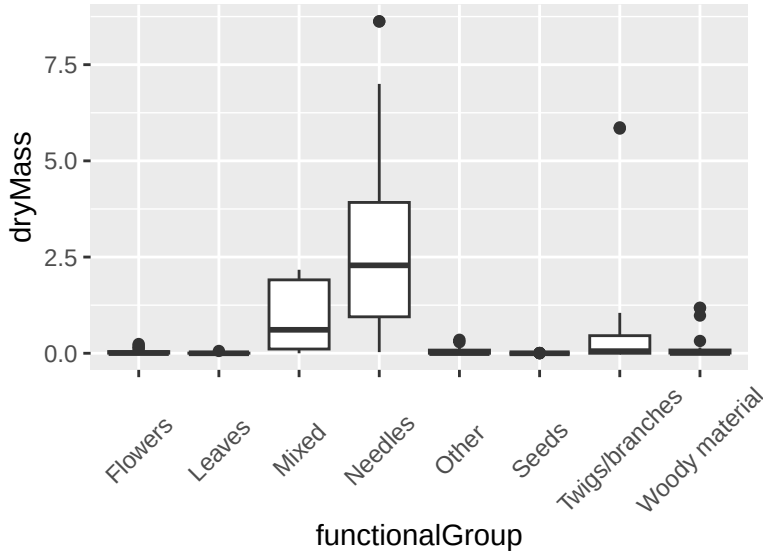
```
#creating a bar graph of count of each functional group
ggplot(Litter) +
  geom_histogram(aes(x = functionalGroup), stat="count") +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5))
```

```
## Warning in geom_histogram(aes(x = functionalGroup), stat = "count"): Ignoring
## unknown parameters: `binwidth`, `bins`, and `pad`
```

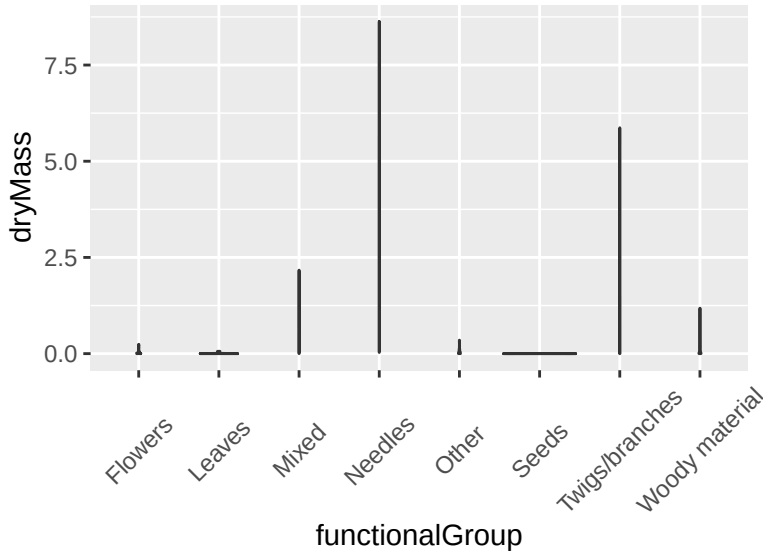


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#creating a boxplot of dry mass by functional group
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass)) +
                theme(axis.text.x = element_text(angle = 45, vjust = 0.5))
```




```
#creating a violin plot of drymass by functional group
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75)) +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5))
```



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

Answer: The boxplot is a more effective visualization option in this case because the densities of each drymass amount for each functional group are all quite low, so the violin plot is not

showing us much information. The boxplot on the other hand gives us the median, range, and some variability.

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

Answer: Needles and mixed litter tend to have the highest dry biomass at the sites.