Assignment 3: Data Exploration

Jiyeong Pyo

Fall 2025

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.

```
#Load necessary package
library(tidyverse)
library(lubridate)
library(here)

#check current directory
getwd()
```

```
## [1] "/home/guest/EDE_Fall2025"
```

```r
#Upload two datasets
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

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 is used as pesticide which could be harmful to beneficial insects such as honeybee. Considering the food chain, declining number of useful insects can lead to reduce food resource for entire ecosystem.This is why ecotoxicoloy of neonicotinoids on instects should be covered.

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 foundation of the soil because it is relasing essential nutrients when it is breaken down by micooganism. Plus, since it could become fuel of fire, we need to tracking these data for protecting the forest from enormous fire.

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.Collection Method is varied. Litter used elevated mesh traps and the fine woody devris was gathered by ground traps. 2.Samples for litter were collected up to 13 times a year for tracking seasonal difference, while fine wood debris were collected once per year due to the slow changes. 3. Every five years, sample groups were used for analysis, measuring C, N, 13C, 15N, and lignin.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
dim(Neonics)
```

```
## [1] 4623   30
```

The Nenonics data has 4623 rows and 30 column.

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...]

```
# summary effect column
effect_type <- summary(Neonics$Effect)
# sort sumarry effect column in order of manitude
sort_effect_type <- sort(effect_type,decreasing=TRUE)
sort_effect_type
```

```
##       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
```

Answer:Population effect(1803) is the most commonly studied.I guess this is because tracking the population is easy to quantify and collect the data for analysis.

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...]

```
#summary top six common species name
top_common_name <- summary(Neonics$Species.Common.Name,maxsum=7)
top_common_name
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee
##                   152                   140                   113
##               (Other)
##                  3083
```

Answer:The top six common species are Honey Bee(667),Parasitic Wasp(285), Buff Tailed Bumblebee(183), Carniolan Honey Bee(152), Bumble Bee(140) and Italian Honeybee(113). They are all pollinators, which makes them keystone species in the ecosystem. If their populations decrease, we can assume the ecosystem would face a severe crisis.

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...]

```r
# checking class of `Conc.1..Author.` column
class(Neonics$Conc.1..Author.)
```

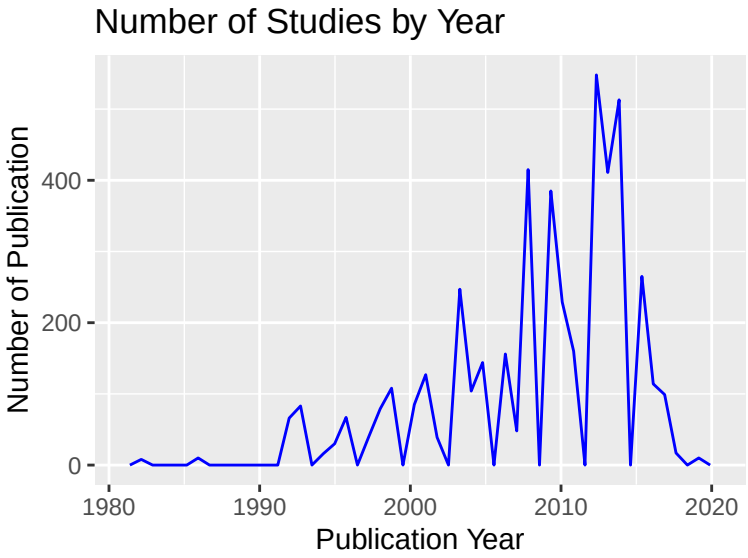
```
## [1] "factor"
```

Answer:'Conc.1..Author.' column is treated as a factor instead of numeric because it contains non-numeric characters like the inequality sign.

Explore your data graphically (Neonics)

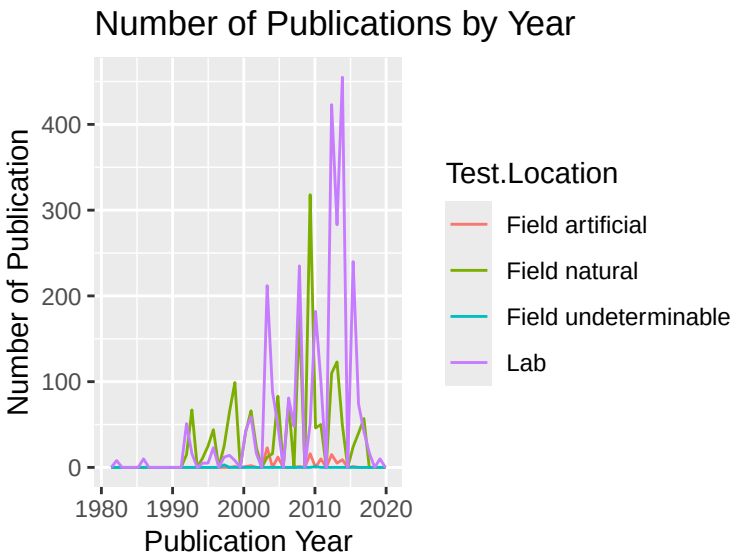
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```r
# Graph number of studies conducted by publication year
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), color = 'blue', bins = 50) +
  scale_x_continuous() +
  labs(title="Number of Studies by Year", x="Publication Year",y="Number of Publication")
```



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

```r
# Graph number of studies conducted by publication year with different colors to indicate different Tes
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 50) +
  scale_x_continuous() +
  labs(title="Number of Publications by Year", x="Publication Year",y="Number of Publication")
```



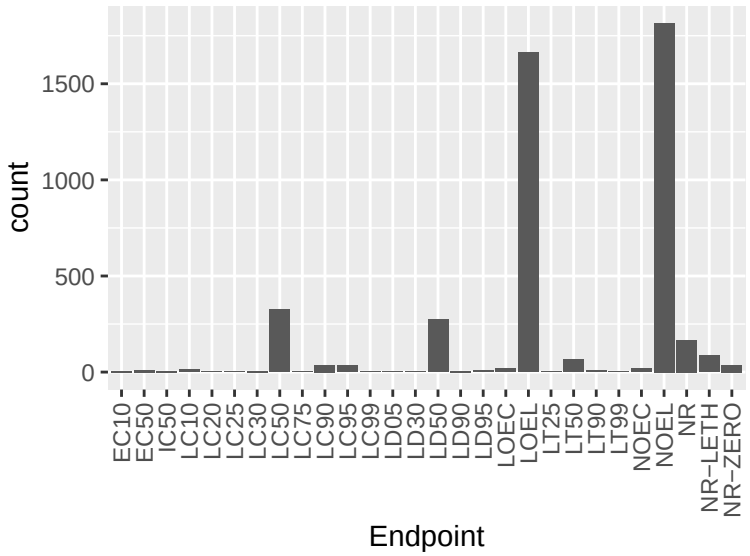
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: Lab studies were the most frequent in overall, showing a dramatic spike to over 400 publications around 2012. Field natural studies were also prominent, especially before 2000 and with significant peaks during the 2010s. In the recent year, studies in both locations decreased.

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...]

```
#Create a bar graph of Endpoint counts
ggplot(Neonics,aes(x = Endpoint)) +
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: LOEL and NOEL are the two most common end points. LOEL(Lowest Observed Effects Residue) means the lowest residue concentration producing effects that were significantly different from responses of controls according to author's reported statistical test.(U.S. Environmental Protection Agency, 2019) NOEL(No-observable-effect-level)is the highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test (NOEAL/NOEC).(U.S. Environmental Protection Agency, 2019)

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.

```
#Check class of collectDate
class(Litter$collectDate)
```

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

```
#Convert factor to date object
collectDate <- ymd(Litter$collectDate)

#Check class of collectDate
class(collectDate)
```

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

```
#Determine the date of the sampling in August 2018
unique(collectDate)
```

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

13. Using the `unique` function, list the different plotIDs 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 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
```

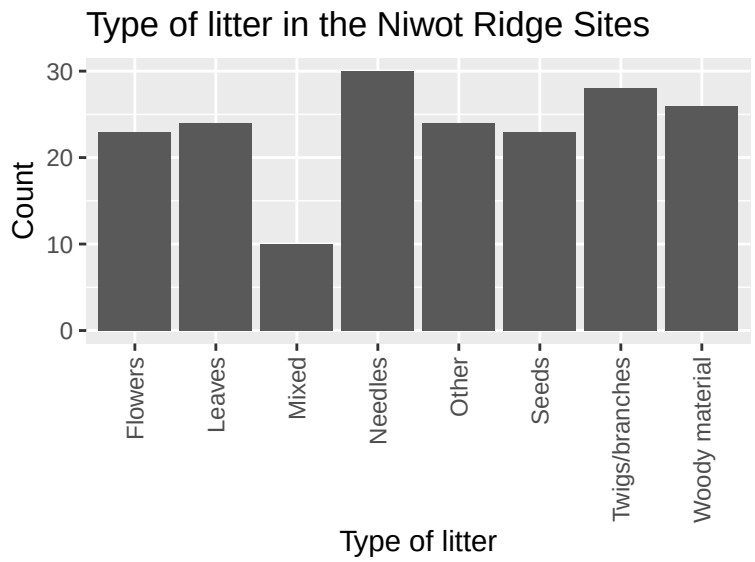
```
summary(Litter$plotID)
```

```
## 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:Unique function only returns the category of plotIDs while summary returns list of the category with the counts of occurence.

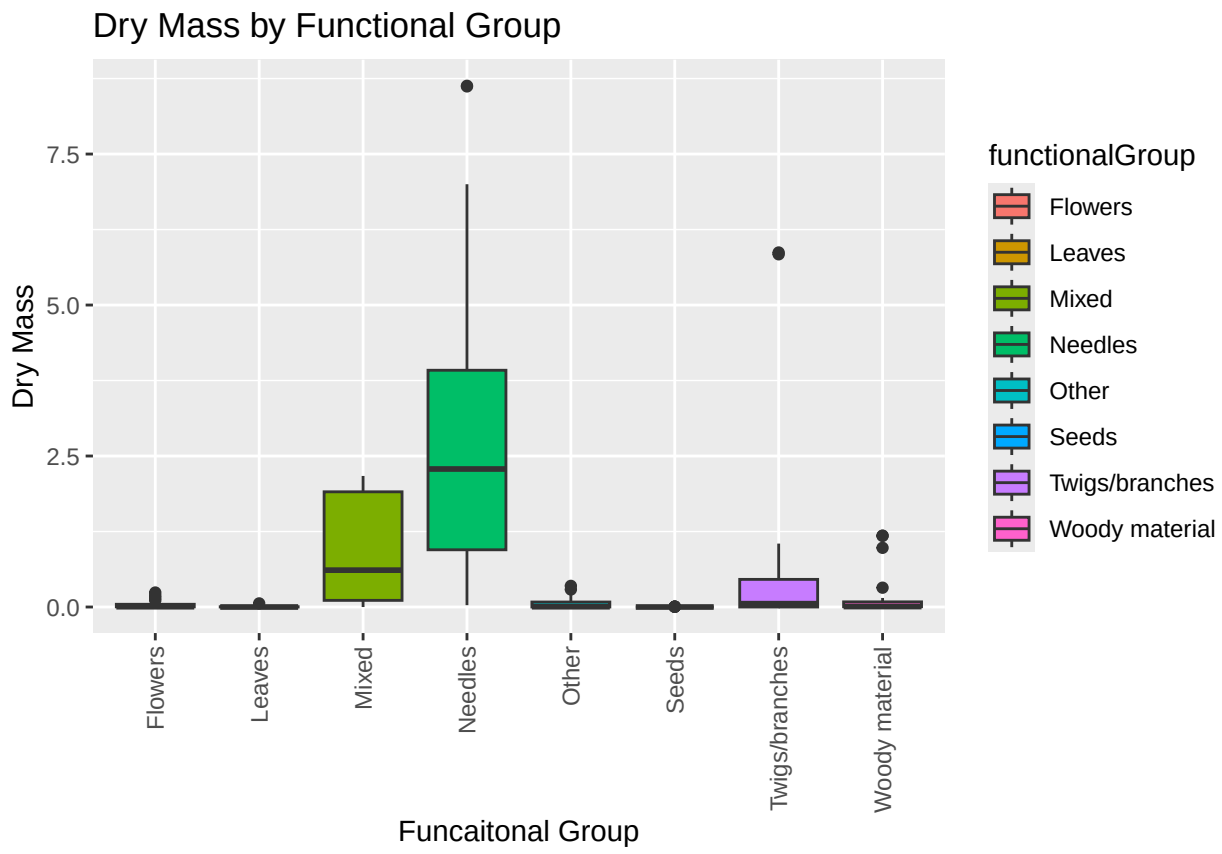
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
ggplot(Litter,aes(x = functionalGroup)) +
  geom_bar()+
 theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
 labs(title="Type of litter in the Niwot Ridge Sites", x="Type of litter",y="Count")
```

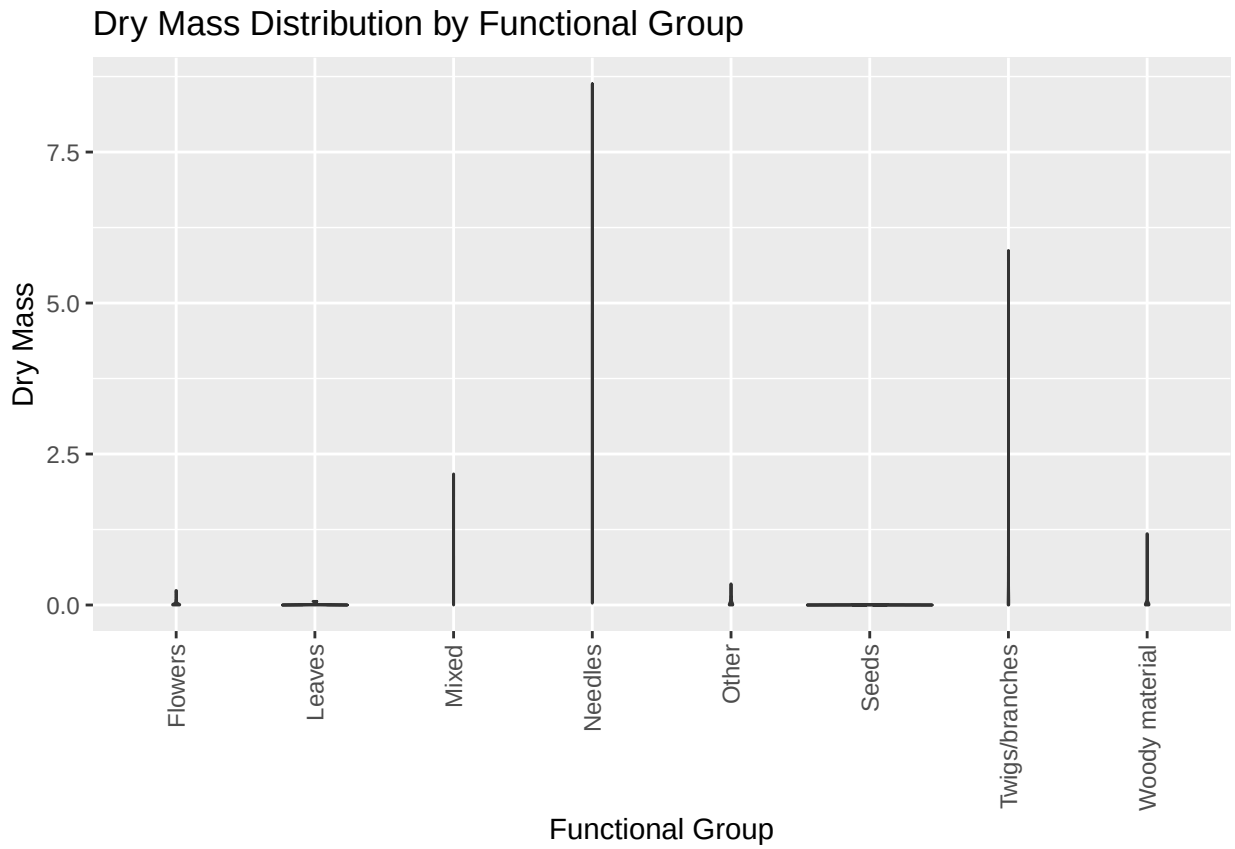


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```r
# Create Box plot
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass, fill = functionalGroup)) +
  labs(title = "Dry Mass by Functional Group",
       x = "Funcaitonal Group",
       y = "Dry Mass") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```




```
#Create Violin plot
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass)) +
  labs(title = "Dry Mass Distribution by Functional Group",
       x = "Functional Group",
       y = "Dry Mass") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



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

Answer:A box plot is better than a violin plot for this data, since the data has significant outliers and lacks spread, which is difficult to visualize effectively with a violin plot.

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

Answer:Needles biomass has the highest biomass at these sites.