Assignment 3: Data Exploration

Jared Wang

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.

Package used in this exercise

```
library(ggthemes)
library(ggplot2)
library(tidyverse)
library(gridExtra)
```

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()
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

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

   Answer: not all insects in the ecosystem are target species. On the contrary, many insects are important components of the system. Therefore, we need to understand effects of pesticides on insects to better estimate their adverse effects.

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: these materials are important because they provide crucial habitats for organisms and are indespensable for keeping the microclimate in forests.

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: * All samples were collected in sites with 2 m tall woody vegetation. * Litter were collected with elevated 0.5 m2 traps. * Woody debris were collected with 3 m * 0.5 m ground traps.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

   Answer: there are 30 variables and 4623 observations.

6. 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: Population and density are the most commonly studied effects. It is likely because they are the easiest effects to observe and tells the most about changes in ecological functions of the species.

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.

```
summary(Neonics$Species.Common.Name) #summary results hidden to save space
summary(Neonics$Species.Group)
```

   Answer: the six most commonly studied species are honey bee, parasitic wasp, buff tailed bumblebee, carniolan honey bee, bumble bee, and italian honeybee. Decline in bee population around the globe has been a growing concerning issue. Prevalence of pesticides might be a significant cause of their population decline.

8. 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: It is a factor. There are undesired symbols such as "/" and "~" in the data.

Explore your data graphically (Neonics)

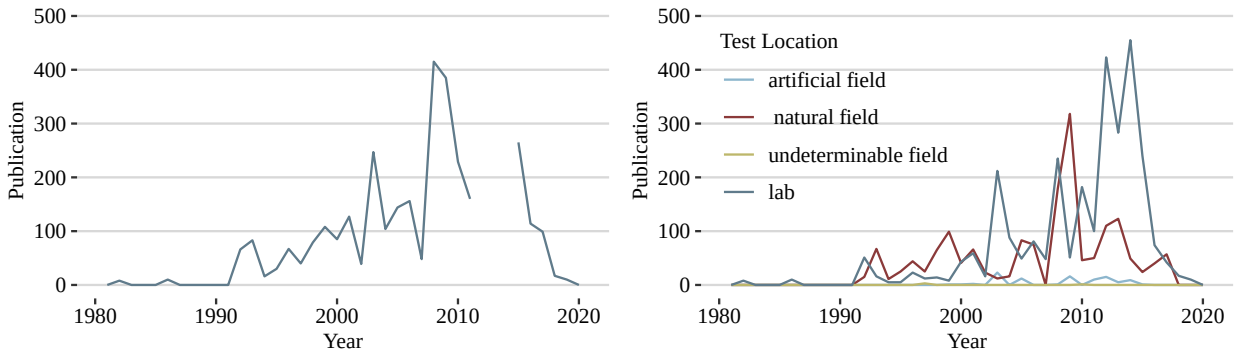
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

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

```
#publication per year
freq.pub <- ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), binwidth = 1, color = "lightskyblue4") +
  labs(y= "Publication", x = "Year") +
  scale_y_continuous(limits = c(0, 500)) +
  theme_hc() +
  theme(axis.title = element_text(family = "serif", size = (10)),
        axis.text = element_text(family = "serif", size = (10), color = "black"))

#publication per year, by location
freq.pub.loc <- ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), binwidth = 1) +
  labs(y= "Publication", x = "Year", color = "Test Location") +
  scale_y_continuous(limits = c(0, 500)) +
  theme_hc() +
  scale_color_manual(labels = c("artificial field", " natural field",
                                "undeterminable field", "lab"),
                     values = c("lightskyblue3", "indianred4",
                                "darkkhaki", "lightskyblue4")) +
  theme(axis.title = element_text(family = "serif", size = (10)),
        axis.text = element_text(family = "serif", size = (10), color = "black"),
        legend.position = c(0.25, 0.6),
        legend.title = element_text(size = 10, family = "serif"),
        legend.text = element_text(size = 10, family = "serif"),
        legend.key = element_rect(color = NA, fill = NA),
        legend.background = element_rect(color = NA, fill = NA))

grid.arrange(freq.pub, freq.pub.loc, ncol = 2)
```

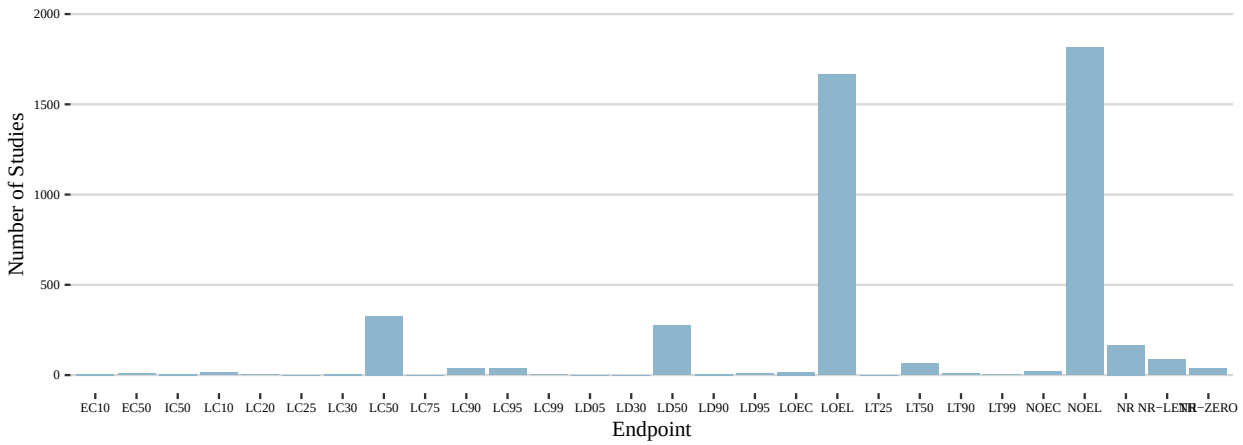


Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: Lab and natural field are the most common locations. They peak at generally the similar time, but not exactly match each other.

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.

```
#number of studies of different endpoints
ggplot(Neonics) +
  geom_bar(aes(x = Endpoint), fill = "lightskyblue3") +
  labs(y= "Number of Studies", x = "Endpoint") +
  scale_y_continuous(limits = c(0, 2000)) +
  theme_hc() +
  theme(axis.title = element_text(family = "serif", size = (10)),
        axis.text = element_text(family = "serif", size = (6), color = "black"))
```



Answer: LOEL and NOEL are the two most commonly used endpoints. LOEL is the lowest observed effect level and NOEL is the no observed effect level.

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.

```
class(Litter$collectDate)
```

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

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

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

```
collect.date <- unique(Litter$collectDate)
collect.date
```

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

Answer: litter was sampled in 2nd and 30th of Augsut.

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`?

```
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
```

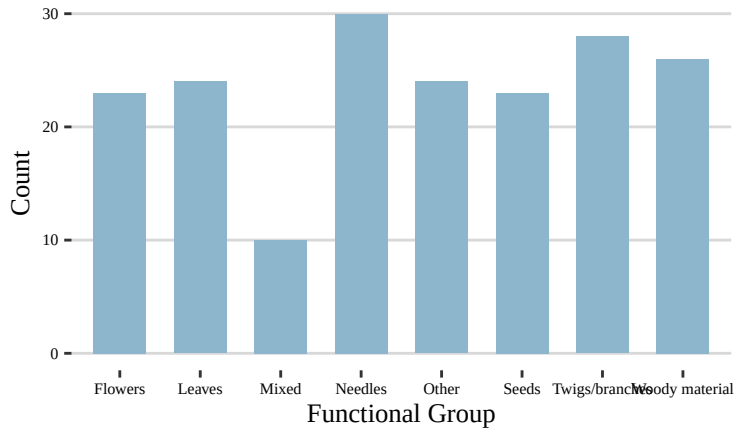
```
summary(Litter$plotID)
```

```
## 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. the "unique" function only provides names of variables, while the "summary" function also tells us the number of observations for each level.

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.

```
# sample size of each functional group - bar plot
ggplot(Litter) +
  geom_bar(aes(x = functionalGroup), width = 0.65, fill = "lightskyblue3") +
  labs(y= "Count", x = "Functional Group") +
  scale_y_continuous(limits = c(0, 30)) +
  theme_hc() +
  theme(axis.title = element_text(family = "serif", size = (10)),
        axis.text = element_text(family = "serif", size = (6), color = "black"))
```



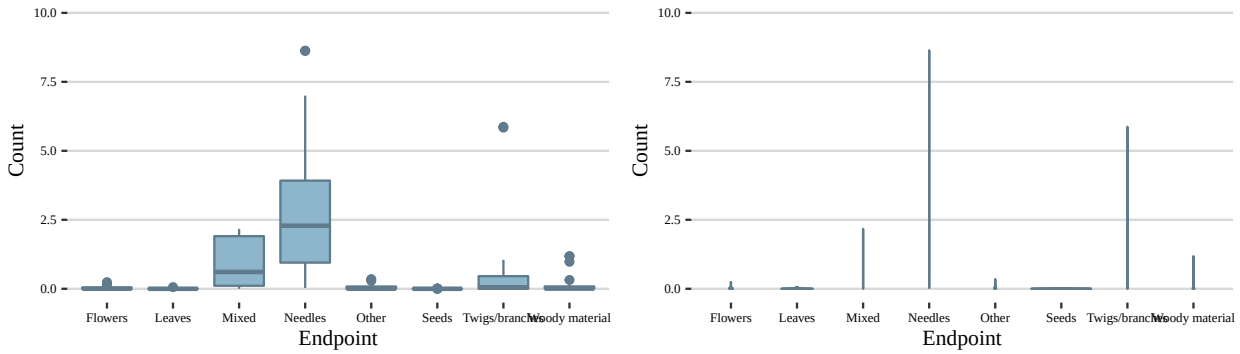
15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#dry mass by functional group - boxplot
box.funcgrp <- ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass),
               fill = "lightskyblue3", color = "lightskyblue4") +
  labs(y= "Count", x = "Endpoint") +
  scale_y_continuous(limits = c(0, 10)) +
  theme_hc() +
  theme(axis.title = element_text(family = "serif", size = (10)),
        axis.text = element_text(family = "serif", size = (6), color = "black"))

#dry mass by functional group - violin plot
vio.funcgrp <- ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75),
              fill = "lightskyblue3", color = "lightskyblue4") +
  labs(y= "Count", x = "Endpoint") +
```

```
  scale_y_continuous(limits = c(0, 10)) +
  theme_hc() +
  theme(axis.title = element_text(family = "serif", size = (10)),
        axis.text = element_text(family = "serif", size = (6), color = "black"))

grid.arrange(box.funcgrp, vio.funcgrp, ncol = 2)
```



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

Answer: The boxplot tells us more information about distribution of data.

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

Answer: Needles tend to have the highest biomass in sampled sites.