Assignment 3: Data Exploration

Nikki Egna

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/nikkiegna/Desktop/Classes/Spring 2020/Environmental Data Analytics/Environmental_Data_Ana
```

```
library(tidyverse)
library(ggplot2)

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: Neonicotinoid insecticides are meant to kill insects that are eating agricultural products. It's important to study ecotoxicology on insects to make sure that the insecticides are having their intended consequences and actually killing the insect pests. If they are not killing the intended bugs, their use is unnecessary and should not be utilized. It also is important to study their

effect on other insect populations. For example, Neonicotinoids are thought to have negative impacts on bees, which are not an intended target species. In a controlled setting, it's important to determine the appropriate doses of insecticide necessary to accomplish the intended affect. This ensures that insecticides are applied at effective quantities, and not an unnecessary excess is being released to the environment.

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: It is important to study leaf litter and debris because it is a critical part of biogeochemical cycling in a forest. Decaying lead litter returns essential nutrients, such as Nitrogen, Phosphorus, and Potassium, back into the soil. Thus, a large amount of leaf litter can potentially indicate healthy soils.

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: *Ground sampling occurs once per year* Plots are typically 20x20m or 40x40m *One ground trap and one air trap is deployed for every 400m sq plot area

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

```
# Printed as number of rows, then number of columns
```

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: The common effects are population, mortality, behavior, feeding behavior, reproduction, and development (in order of frequency). These could be of interest because it looks at the different ways that the insecticide may affect insects. For example, whether or not their population size or mortality rates are changing determines if the insecticidde is working. Further, if the insecticide is targeting feeding behavior change, it is important to collect data on the concentration of insecticide that may deter the insects from feeding on a given plant.

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.

```
head(summary(Neonics$Species.Common.Name))
```

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

Answer: The 6 most commonly studied species are all types of bees. This is because bees are not specifically the target of insecticides, but often are affected negatively by their application. Bees are an essential part of the ecosystem, and provide many ecosystem services such as pollination, thus it is critical to understand how insecticides may alter their population.

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

```
str(Neonics$Conc.1..Author.)
```

```
##  Factor w/ 1006 levels "<0.0004","<0.025",..: 639 510 813 622 442 637 500 642 814 784 ...
```

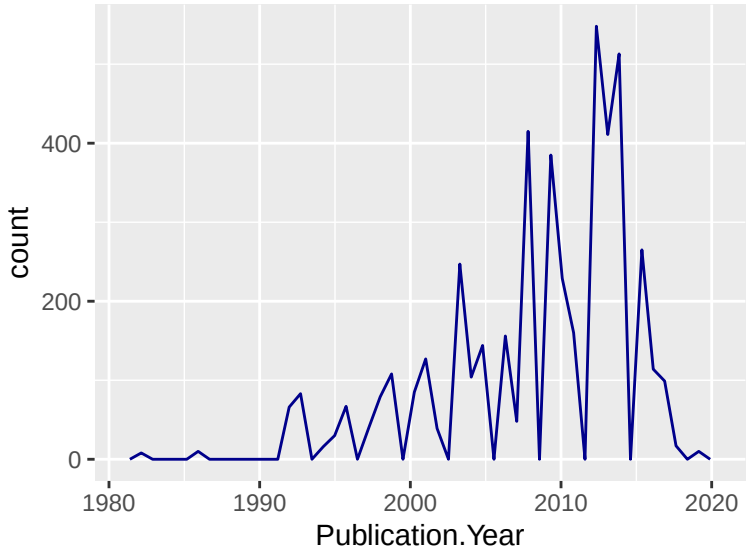
```
#View(Neonics$Conc.1..Author.)
```

Answer: Conc.1..Author. is a factor with 1006 levels. It is not a numeric column because it contains character values within cells with the numbers, for example 1.00/ , <4.00 , NR , or ~10. The presence of characters such as these will prevent R from recognizing this column as numeric.

Explore your data graphically (Neonics)

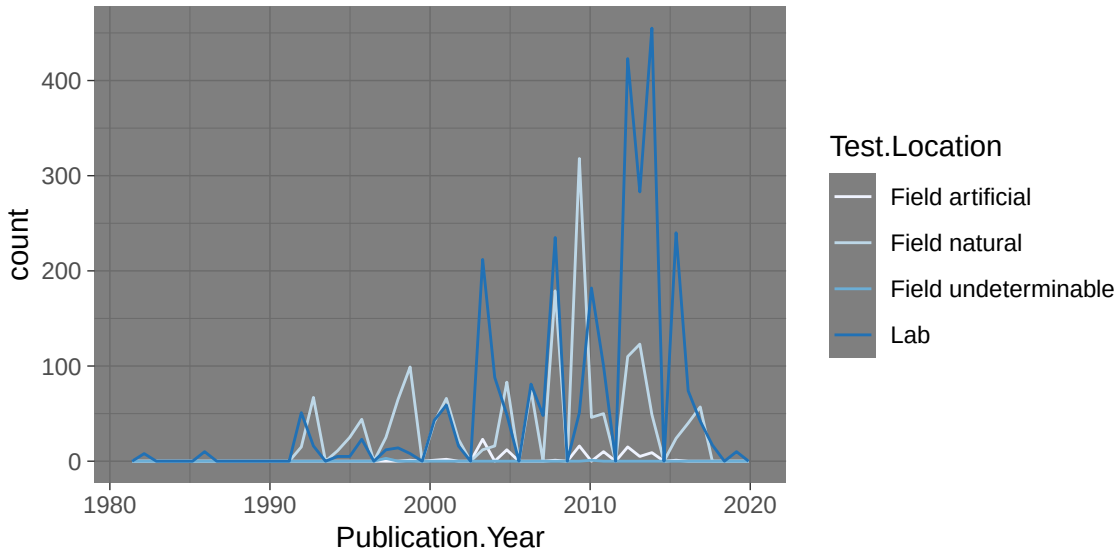
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
pub.year.graph<-ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50, col="darkblue")

plot(pub.year.graph)
```



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

```
pub.year.graph.color<-ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 50) +
  theme(legend.position = "right")+ scale_color_brewer(palette = "Blues")+ theme_dark()

plot(pub.year.graph.color)
```

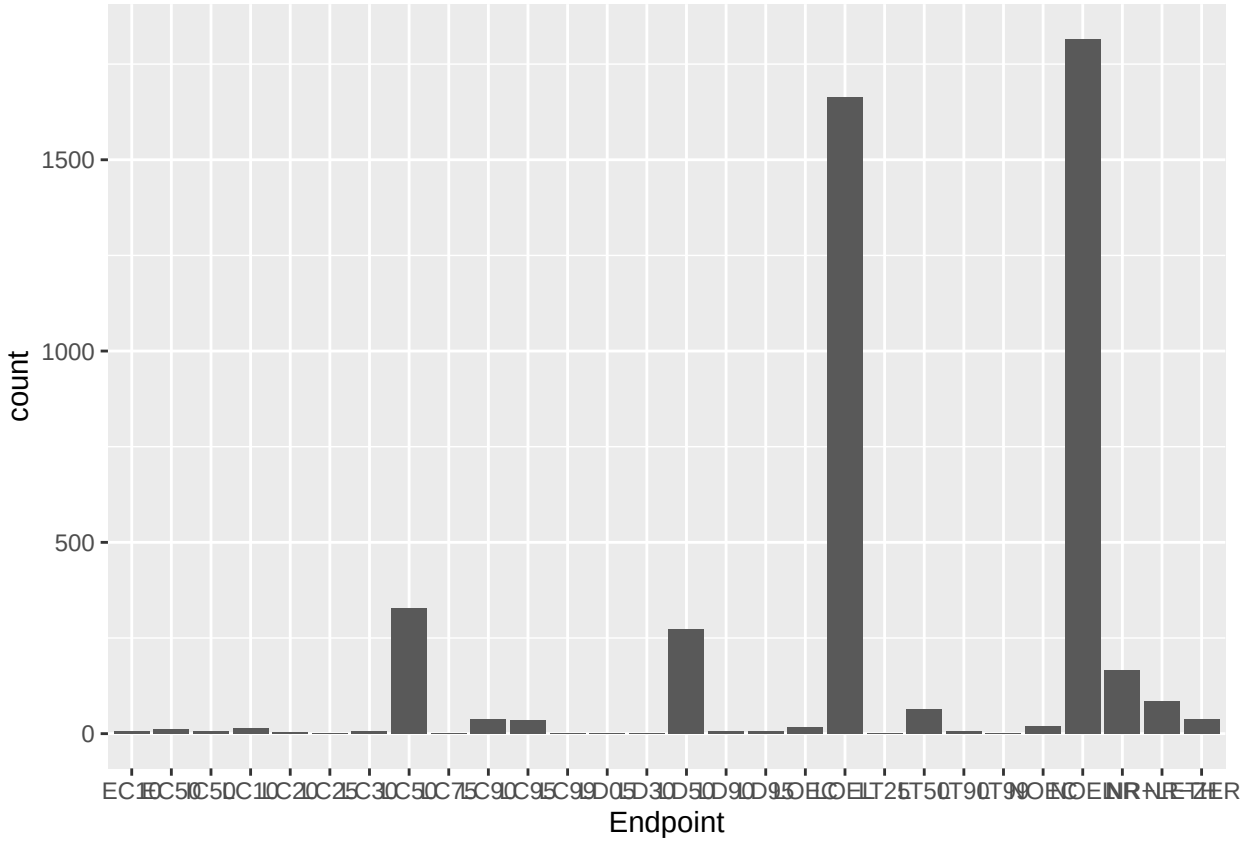


Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: The most common test locations are in the lab and the natural field. There is a great deal of variation over time, shown by the spikes and pitfalls in the graph.

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.

```
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar()
```




```
table(Neonics$Endpoint)
```

```
##
##     EC10     EC50     IC50     LC10     LC20     LC25     LC30     LC50     LC75
##        6       11        6       15        5        1        6      327        1
##     LC90     LC95     LC99     LD05     LD30     LD50     LD90     LD95     LOEC
##       37       36        2        1        1      274        6        7       17
##     LOEL     LT25     LT50     LT90     LT99     NOEC     NOEL       NR NR-LETH
##     1664        1       65        7        2       19     1816      167       86
## NR-ZERO
##       37
```

Answer: LOEL and NOEL are the two most common endpoints. LOEL is defined in the ECOTOXCodeAppendix as "Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC)" and NOEL is defined as "No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test (NOEAL/NOEC)"

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"
```

```
unique(Litter$collectDate)
```

```
## [1] 2018-08-02 2018-08-30
## Levels: 2018-08-02 2018-08-30
```

```
#August 2nd and August 30th

Litter$collectDate <- as.Date(Litter$collectDate)
class(Litter$collectDate)
```

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

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
```

```
length(unique(Litter$plotID))
```

```
## [1] 12
```

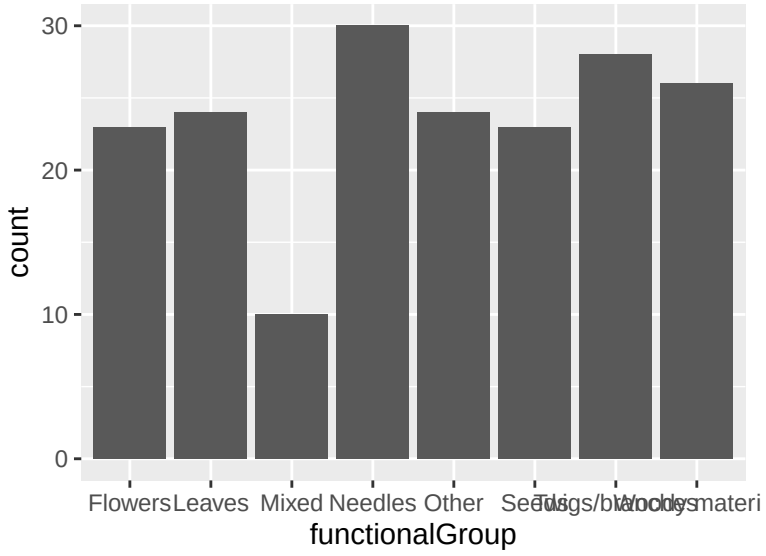
```
length(summary(Litter$plotID))
```

```
## [1] 12
```

Answer: Summary tells you how many there are of each unique plotID, whereas unique will show you just the different names. However, wrapping both in the length() function will give you the same answer.

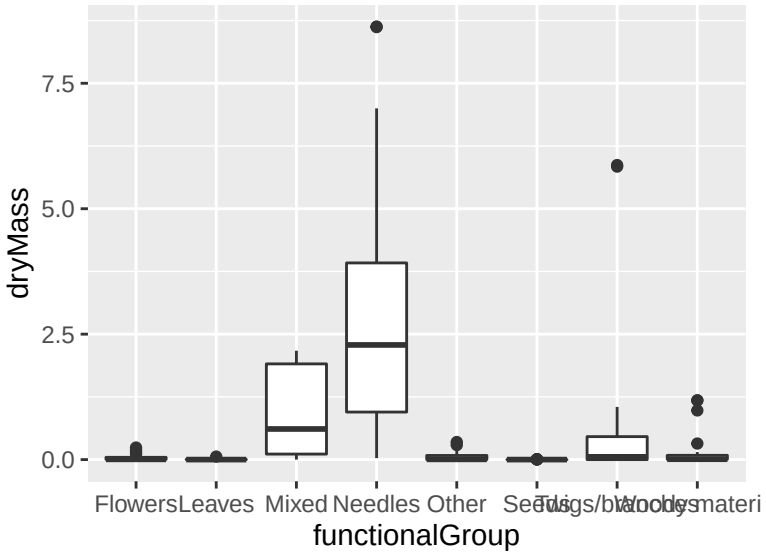
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar()
```

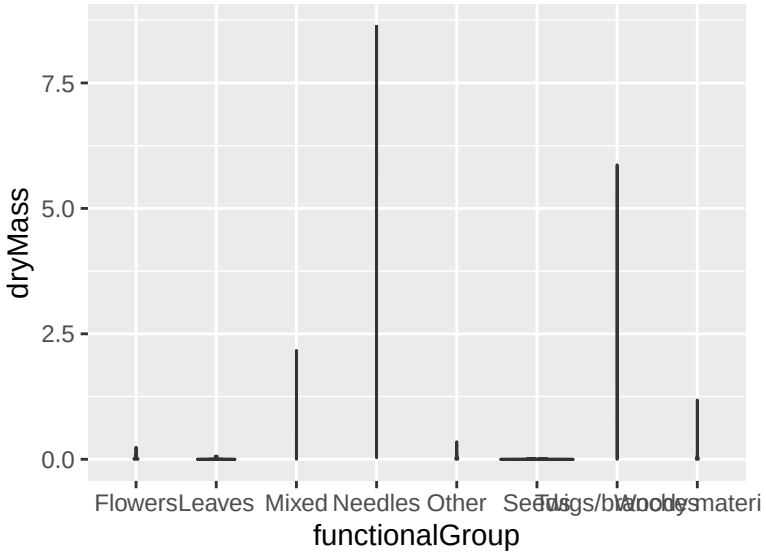


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass, group = cut_width(functionalGroup, 1)))
```




```
#
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass))
```



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

> Answer: The boxplot reveals more information in this case because it displays the median, the outlier, the IQR, and the middle 50% of the data. The violin plot only shows the density distribution, which in this case, does not reveal much about the data due to the relatively uniform density distributions. This violin plot is also thrown off by outliers, as seen by comparing the Twigs/branches functional group between the boxplot and the violin plot.

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

> Answer: Needles and Mixed functional groups tend to have the highest biomass at these sites.