Assignment 3: Data Exploration

Ardath Dixon

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 <>.

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.

```
setwd("/Users/ardat/OneDrive/Documents/DataAnalytics/Environmental_Data_Analytics_2021/")
getwd()
```

```
## [1] "C:/Users/ardat/OneDrive/Documents/DataAnalytics/Environmental_Data_Analytics_2021"
```

```
library(tidyverse)
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE)
Litter <- read.csv("./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 ecotoxicologoy of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: We might be interested in the exotoxicology of neonicotinoids on insects to see if/how the chemicals used in agriculture affect organisms' health. It would be of interest to learn if these chemicals have repercussions on the overall insect population and therefore on the overall ecosystem. We also would be interested to see how these insecticides biologically impact the insects, and therefore if there is a potential danger that would translate to the agricultural crops and therefore to human health when the crops are consumed. Thirdly, some crops rely on insect pollination, so analysis of the chemicals' effects on a variety of species is an area of interest to study.

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: We would be interested to study the litter and woody debris in forests to assess what natural mechanisms influence trees' growth. Specifically, we would be interested to see what composes the litter, and if its presence has a positive influence on growth. This would be relevant for learning how to improve forest growth patterns, then seeing if there was a way to replicate the biological benefits of forest litter for human-influenced areas such as planted tree farms or reforestation efforts.

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: Litter and woody debris is sampled in the NEON network using litter traps. Elevated PVC litter traps are 0.5 square meter mesh baskets elevated about 80cm above the ground. Ground traps are also 0.5 square meters large. Sites are selected if they contain vegetation >2 meters tall and are located in a tower plot. Some forests have frequent sampling (e.g. every two weeks in deciduous forests during senescence) and others are more infrequent (e.g. once every 1-2 months for evergreens). *

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

6. Using the summary function on the "Effects" column, 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 Mortality are the most commonly studied effects. These are likely because alive or dead and count of the organisms are easier to examine for the samples. Also, when analyzing insects' reactions to chemicals, quantity of living insects is the simplest and most general way to quanitify the chemicals' effects. Meanwhile, characteristics such as biochemistry or enzymes require more in depth research done for a particular sample and give more nuanced, specific conclusions.

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.Scientific.Name,7)
```

```
##               Apis mellifera             Bombus terrestris 
##                          667                           183 
##   Apis mellifera ssp. carnica              Bombus impatiens 
##                          152                           140 
## Apis mellifera ssp. ligustica             Popillia japonica 
##                          113                            94 
##                      (Other) 
##                         3274
```

Answer: The six most commonly studied species are the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. These all bees (plus one wasp with similarly striped appearance). Honeybees and bumblebees feed on nectar and therefore often serve as pollinators for several plant species. They might be of interest over other insects because they have an impact on plants' distribution and overall growth trends.

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"
```

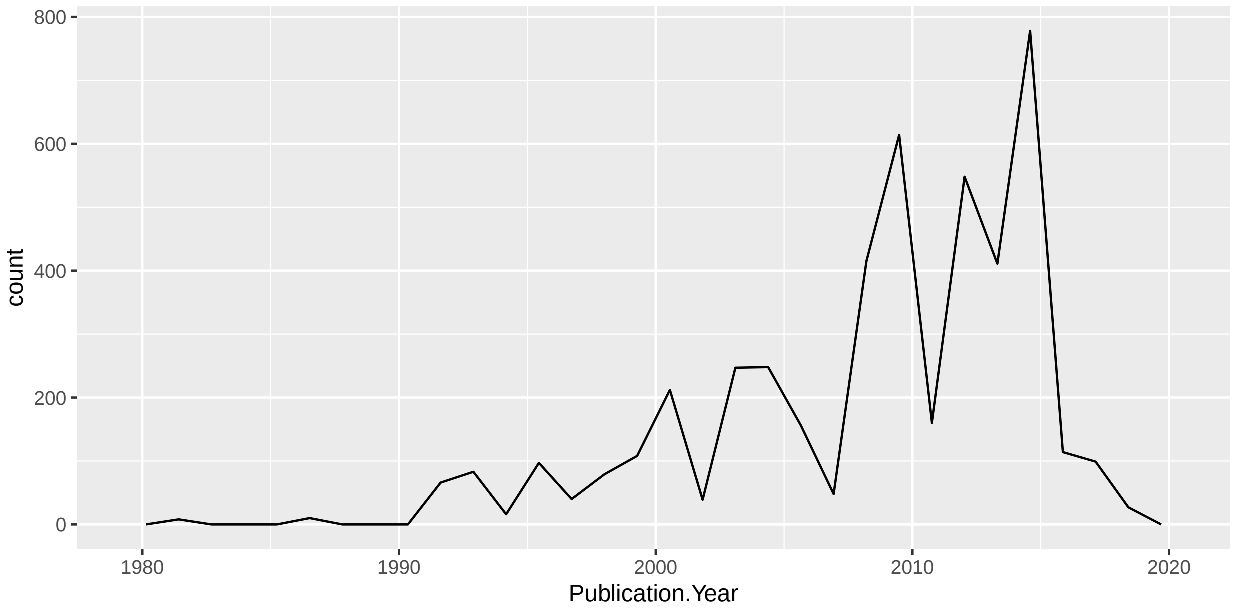
Answer: This class is factor not numeric because there are some entries with "NR/" for not recorded, "~10" for approximately 10, and other values with / included in the cell. There are digits other than purely number values in that column of data, hence R's interpretation as a factor variable. It was originally a character variable, but since the csv was imported with the stringsAsFactors command and this class includes characters (aka a string), Conc.1..Author. is now read as a factor.

Explore your data graphically (Neonics)

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

```
ggplot(Neonics)+
  geom_freqpoly(aes(x = Publication.Year))
```

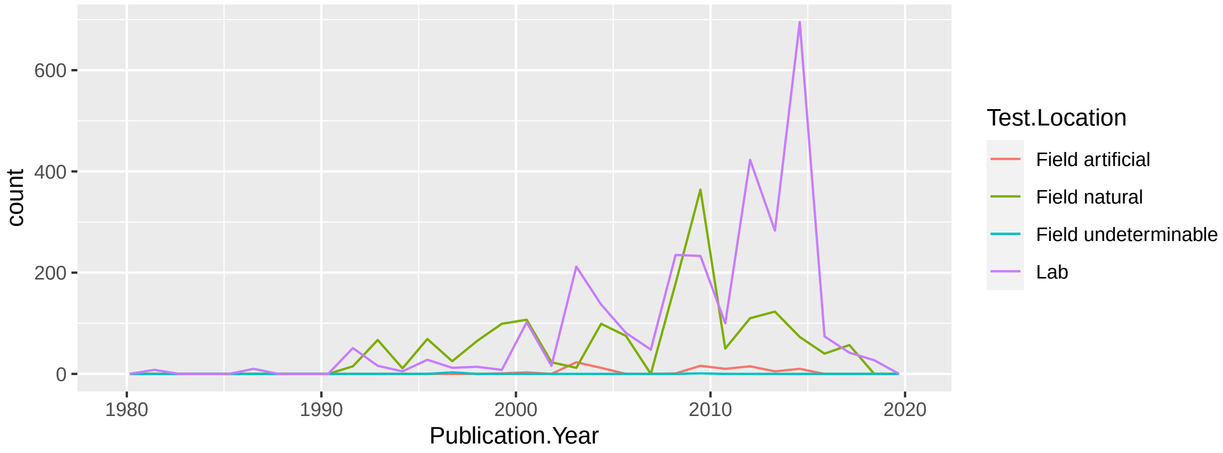
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



10. 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))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



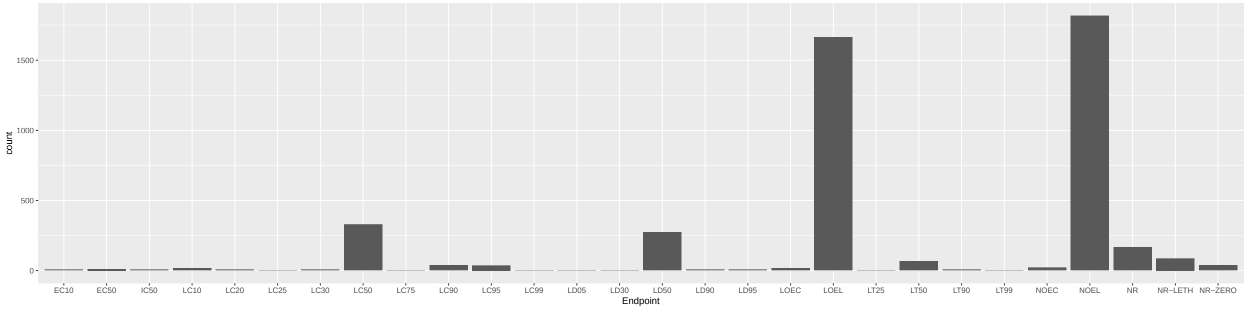
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test location is in the lab, and the second most common is field natural. In the beginning, lab was slightly above field natural, but field natural was the most common test location between ~1992 and ~2002. Lab became the most common location from ~2002 t0 2008, when field natural again became the most common location until 2011. Lab surpassed to be the most common between 2011 and ~2017. The widest gap between the two is between 2011 and 2016, when lab far surpassed field natural test locations. Other sections of the graph show a much closer relationship between the two, following similar overall trends to that of the total tests made.

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()
```



Answer: The two most common end points are NOEL and LOEL. NOEL stands for No Observable Effect Level, meaning that effects were not significantly different from the controls when given the highest doses. LOEL stands for Lowest Observable Effect Level, meaning the effects were significantly different from the controls when given the lowest dose levels.

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"
```

```
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`?

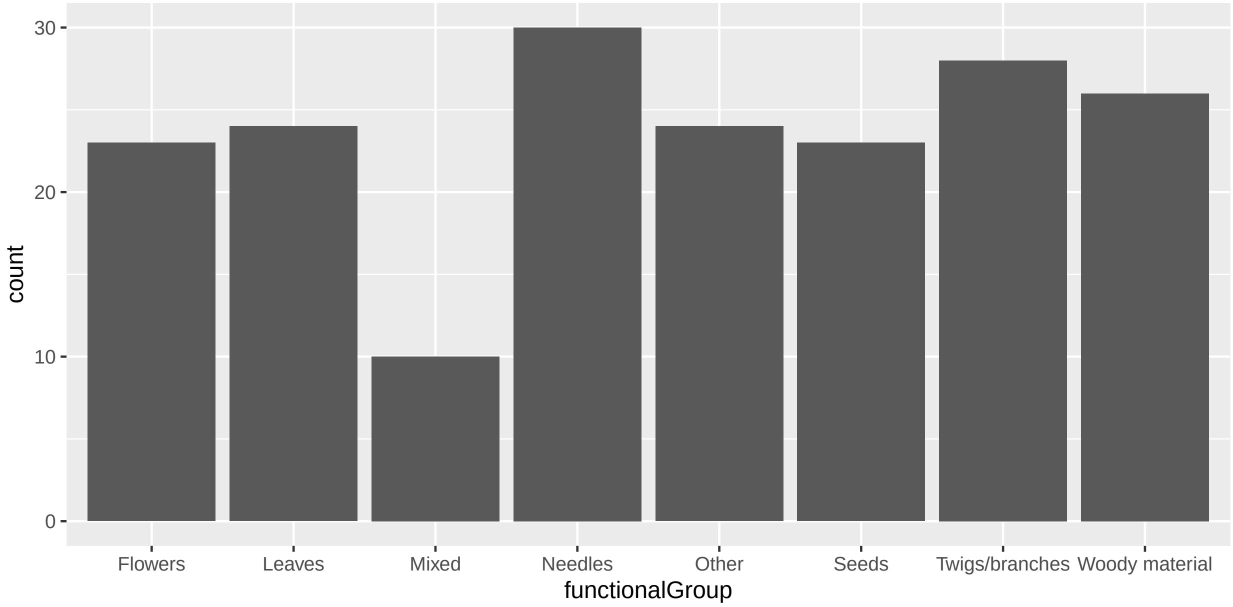
```
nlevels(unique(Litter$plotID))
```

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

Answer: Unique is different than Summary because Summary reports the quanitity of entries with each value along with the value names, while Unique solely reports the names. nlevels(unique()) gives solely the total of different value names for a given column.

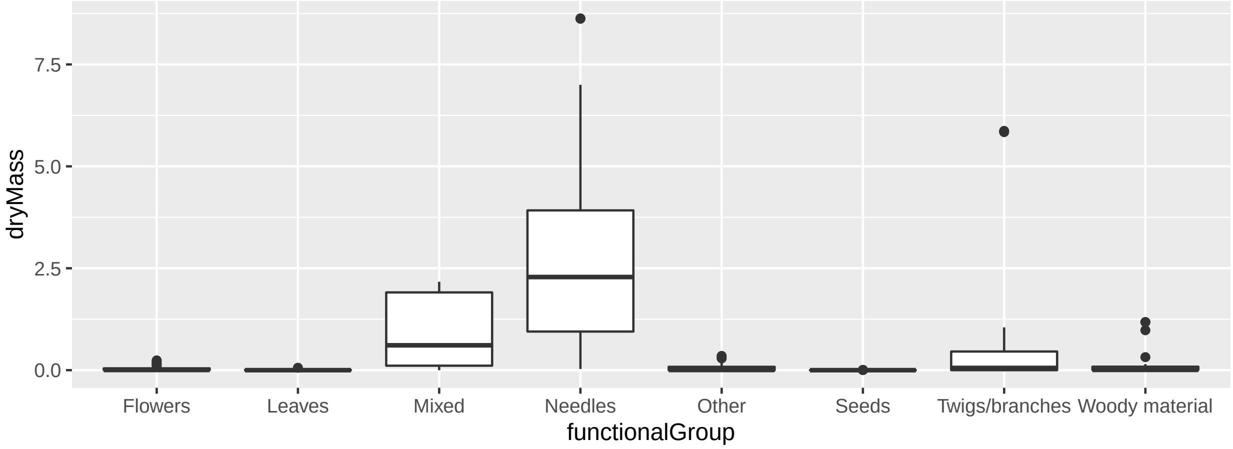
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

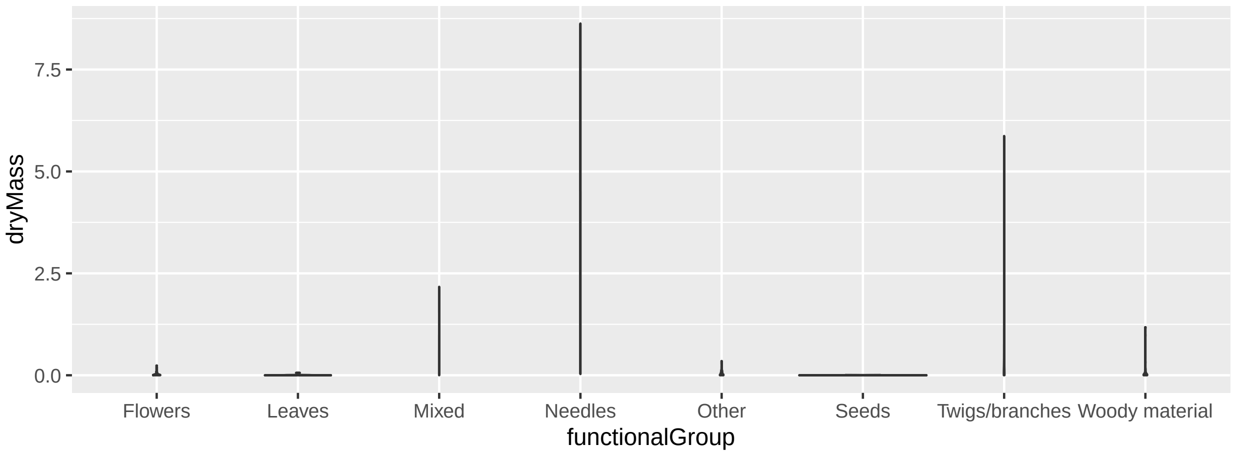


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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



```
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 gives a more effective visualization in this case because it clearly shows the median dryMass for every category. For the violin plot, specific values are only shown if there is a certain concentration of them. The violin plot solely shows the range and distribution trends; while the boxplot shows the range, the median, and the overall distribution with more detail (quantiles, outliers, etc.).

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

> Answer: Sites with a litter of needles tends to have the highest biomass. Mixed litter has the second highest biomass (despite twigs/branches having one recording with high dryMass, its median and quantiles are still far below those of mixed litter).