Assignment 3: Data Exploration

Pierre Mishra

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] "C:/Users/Peaceful Pierre/Documents/Academics/Spring 2020/Environmental Data Analytics/Environme
```

```
library(tidyverse)
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: Neonicotinoids persist in the environment for long durations. They can harm non-target species and, therefore, can cause unintended ecological effects. We might be interested in their ecotoxicology on insects in order to determine the insects that could be vulnerable to neonicotinoids.

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 might be interested in forest litter and woody debris to study the relationship between climatic varibales such as temperature, humidity etc and forest ecosystems. Decomposition of such litter is also crucial for nutrient cycles and, therefore, the amount of litter and woody debris might be of interest to foresters, ecologists and/or biogeochemists.

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 using elevated and ground traps.

- A single observation of litter data is recorded from a single collection event and a single trap (1 spaital-temporal observation).
- The number and placement (targeted or randomized) of elevated and ground traps deployed depends on physical features of vegetation.
- Frequency of sampling also depends on types of vegetation ranging from once every two-weeks to once a year.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

```
#### dimensions are 4623 rows and 30 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: Some of the most common studied effects are population and mortality of species. These effects might be of interest as they can directly provide evidence related to the negative impacts of the insecticides on environment.

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), 6)
```

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

Answer: Neonicotinoids are absorbed by different parts of plants and can cause toxicity in bees collecting nectar and pollen. Since the population of bees is already decling, it might be a grave concern for ecologists.

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 "~10","~30/","~40/",..: 639 510 813 622 442 637 500 642 814 784 ...
```

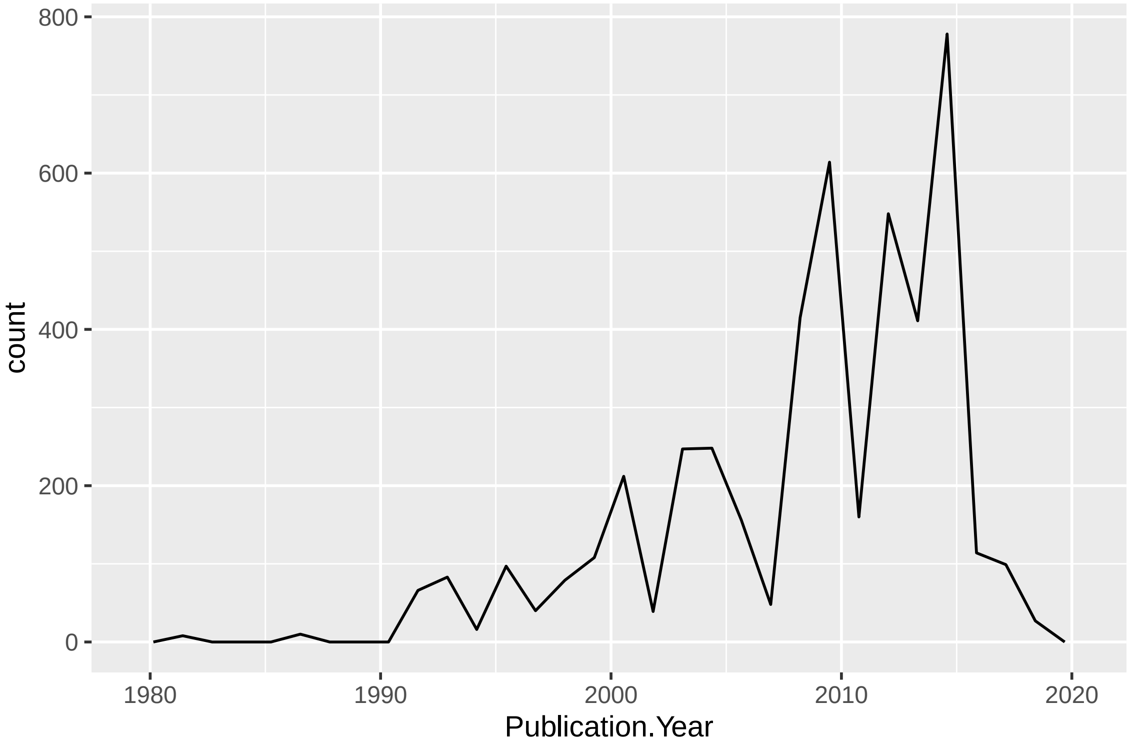
```
#### We could have also used 'class' function
```

Answer: The class of the column is factor. It is not recorded as numeric because some observations have special symbols such as "/", "~" or "<" or even characters like "NR".

Explore your data graphically (Neonics)

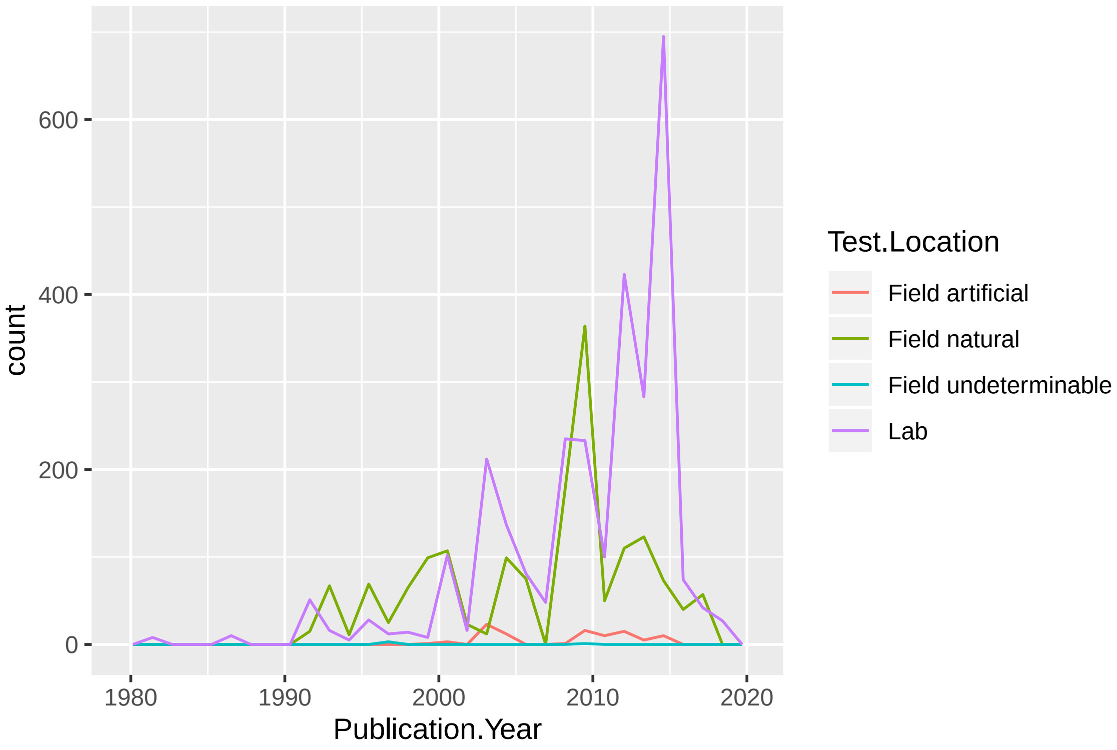
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year), bins = 30)
```



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), bins = 30)
```

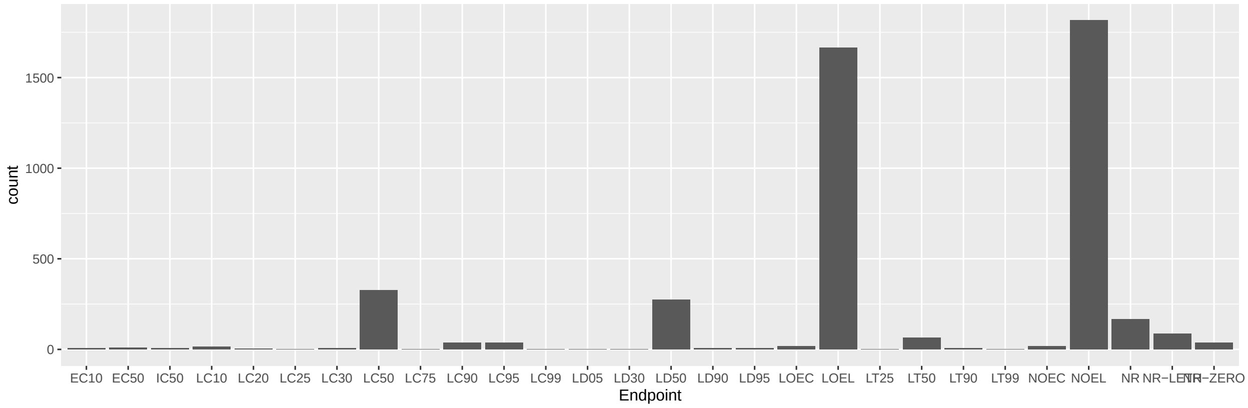


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test location is lab. However, between 2007 to 2010 there were a higher number of experiments in natural fields.

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.

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



> Answer: The two most common end points are LOEL and NOEL. LOEL is defined as lowest-observable-effect-level which means that lowest dose insecticides produced significant effects that were different from controls. NOEL is defined as No-observable-effect-level which means that highest dose of insecticides produced effects that were not significantly different from controls.

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.

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

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

```r
#### The class is a factor and not date.
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
class(Litter$collectDate)
```

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

```r
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`?

```r
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
```

```r
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
```

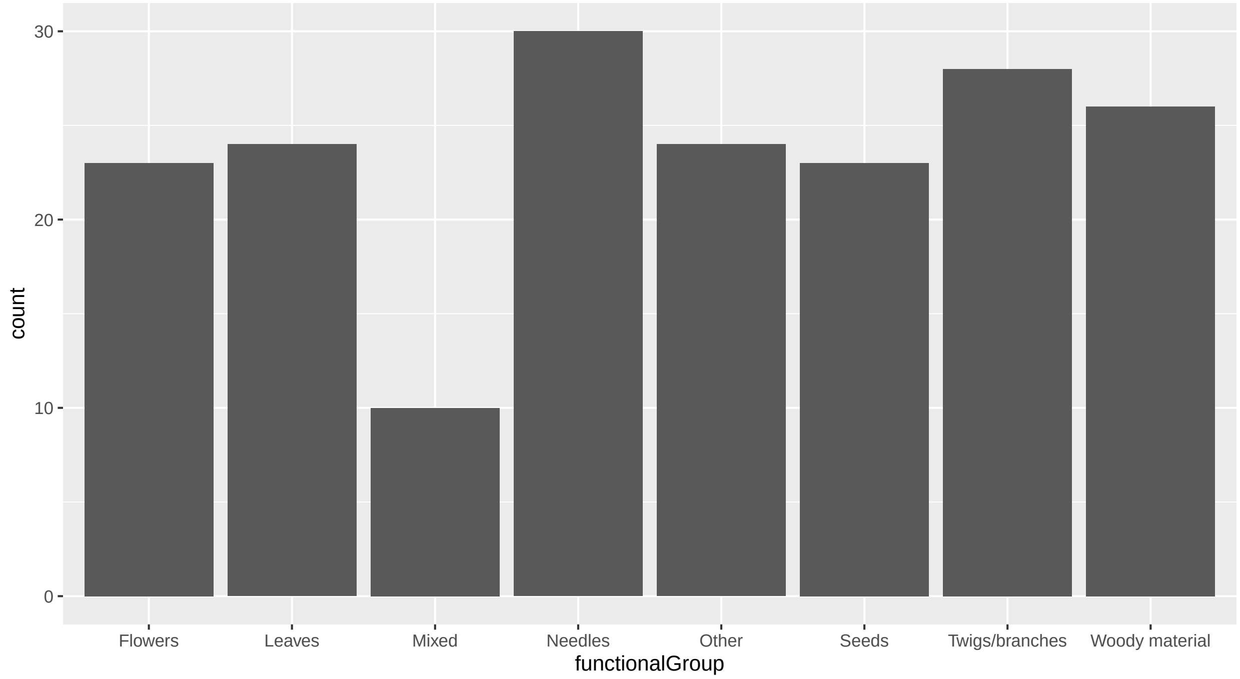
```r
length(unique(Litter$plotID))
```

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

> Answer:If I use the ‘unique’ function I have to count each unique value manually to obtain the number of plots sampled. The ‘summary’ function provides a count of how many times each unique plot is recorded in our data. A better way count the number of plots is by using “length” function along with “unique” function.

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.

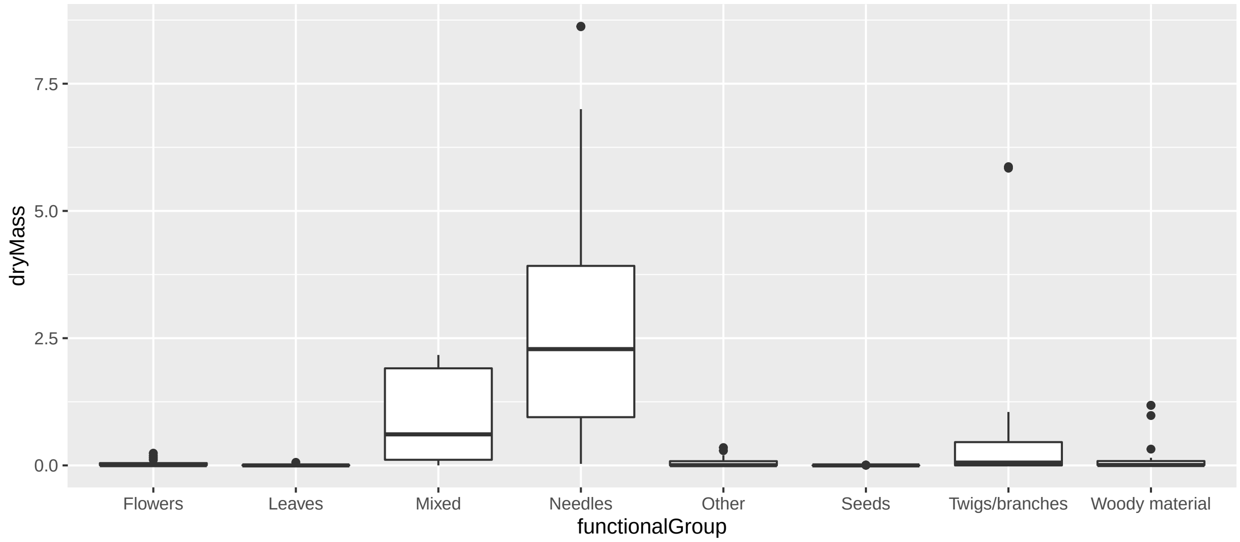
```r
ggplot(Litter) + geom_bar(aes(x = functionalGroup))
```



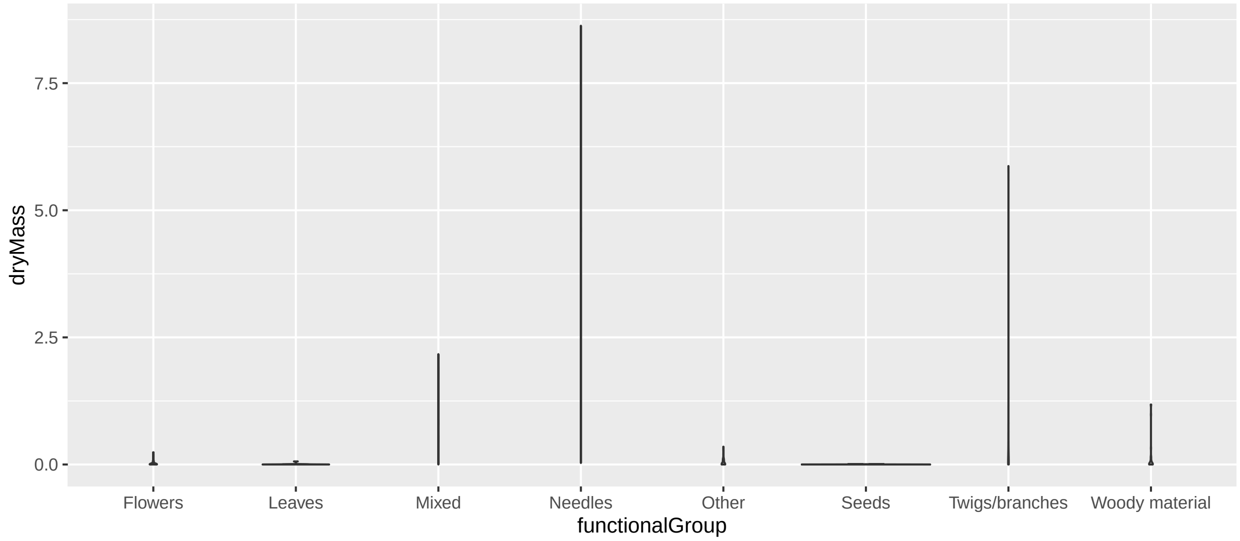
```
#### Each type of litter is collected at the Niwot Ridge sites including flowers, leaves,
#### needles, seeds, twigs/branches, mixed and others.
```

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), scale = 'area')
```



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

> Answer: Because in our case the kernel density of dryMass values is not enough to see any tangible differences in shapes that can tell us more information about our distribution. Instead all we see is a thin line. Therefore, we can just go ahead and use boxplots.

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

Answer: Needles