Assignment 3: Data Exploration

David Amanfu, Section #001 Tuesday

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Exploration.

Directions

1. Change "Student Name, Section #" on line 3 (above) with your name and section number.
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., "FirstLast_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. **Be sure to add the `stringsAsFactors = TRUE` parameter to the function when reading in the CSV files.**

```
library(dplyr)
library(ggplot2)
library(tidyverse)
library(readr)
getwd()
```

```
## [1] "/Users/davidamanfu/Desktop/Duke MPP/Environ Data /Environmental_Data_Analytics_2022/Assignments
```

```
setwd("~/Desktop/Duke MPP/Environ Data /Environmental_Data_Analytics_2022/")
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:

For pest management! And on the flipside for preservation of certain pollinator species that are important to ecosystems.

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:

From this

website

(https://www.nrs.fs.fed.us/data/lcms/niwot/): "Several ongoing studies are conducted at NIWOT involving global change, and specifically climate change and nitrogen deposition. Interactions between climate and ecosystems with complex topography which generate unique source and sink environments for water and nutrients, are examined at these high elevation sites in the Colorado Front Range."

This gives us some insight as to what's going on with the dataset.

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:

They're sampled in different ways. Ground samples are checked annually. Elevated traps are checked more frequently in deciduous areas (biweekly/fortnightly) versus evergreen areas (bimonthly).

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 "Effect" column, determine the most common effects that are studied. Why might these effects specifically be of interest?

```
summary(Neonics$Effect, maxsum = 10)
```

```
##       Population        Mortality         Behavior Feeding behavior
##             1803             1493              360              255
##     Reproduction      Development        Avoidance         Genetics
##              197              136              102               82
##        Enzyme(s)          (Other)
##               62              133
```

Answer: Population, and mortality, along with reproduction and development are all good measures for how impactful a substance will be on a certain species population.

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, maxsum=10)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee
##                   152                   140                   113
##       Japanese Beetle     Asian Lady Beetle        Euonymus Scale
##                    94                    76                    75
##               (Other)
##                  2838
```

Answer: Bees are so important to our ecosystems, that feels like pretty ommon knowledge. And Asian Lady Beetles are ladybugs!! Also an important pollinator.

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"
```

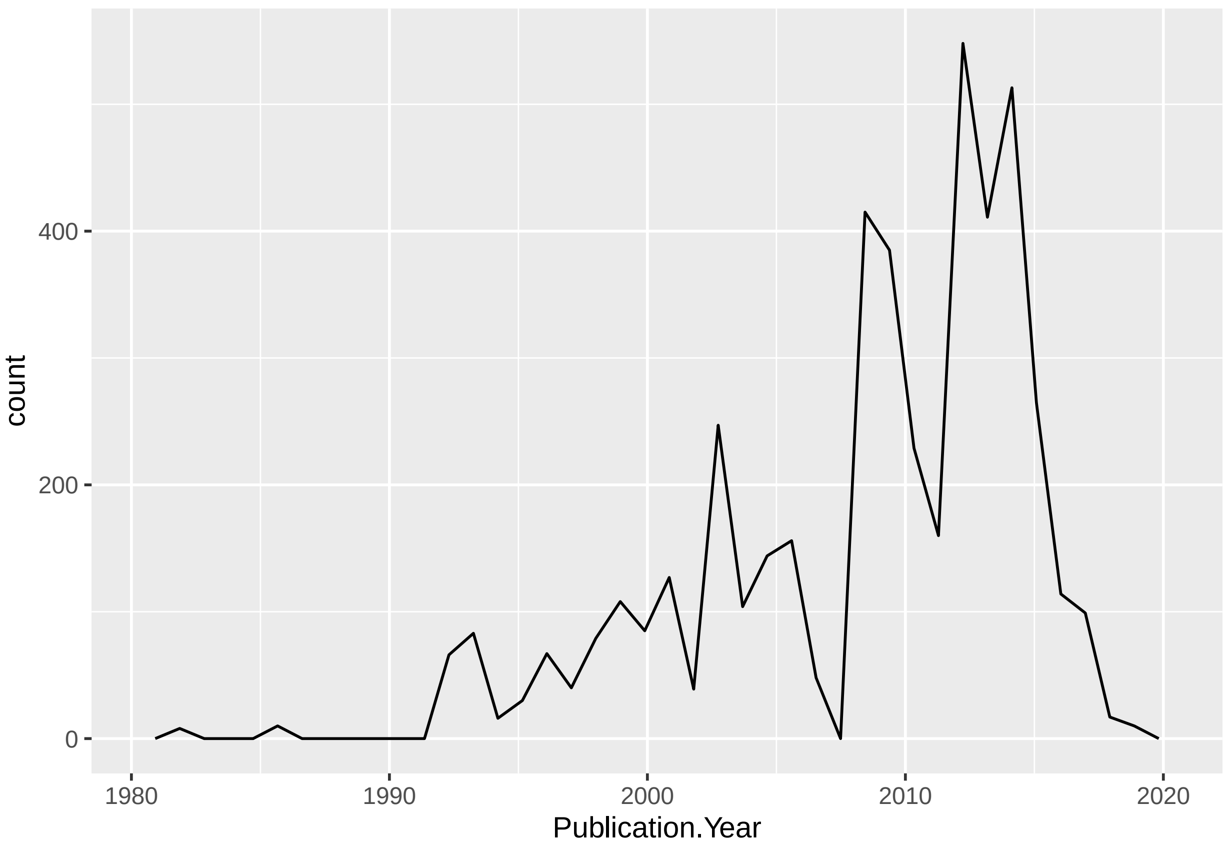
Answer:

It's a factor because we imported the whole CSV with strings as factors, which means that all of the numeric values were loaded in as strings rather than numerics, and thus converted.

Explore your data graphically (Neonics)

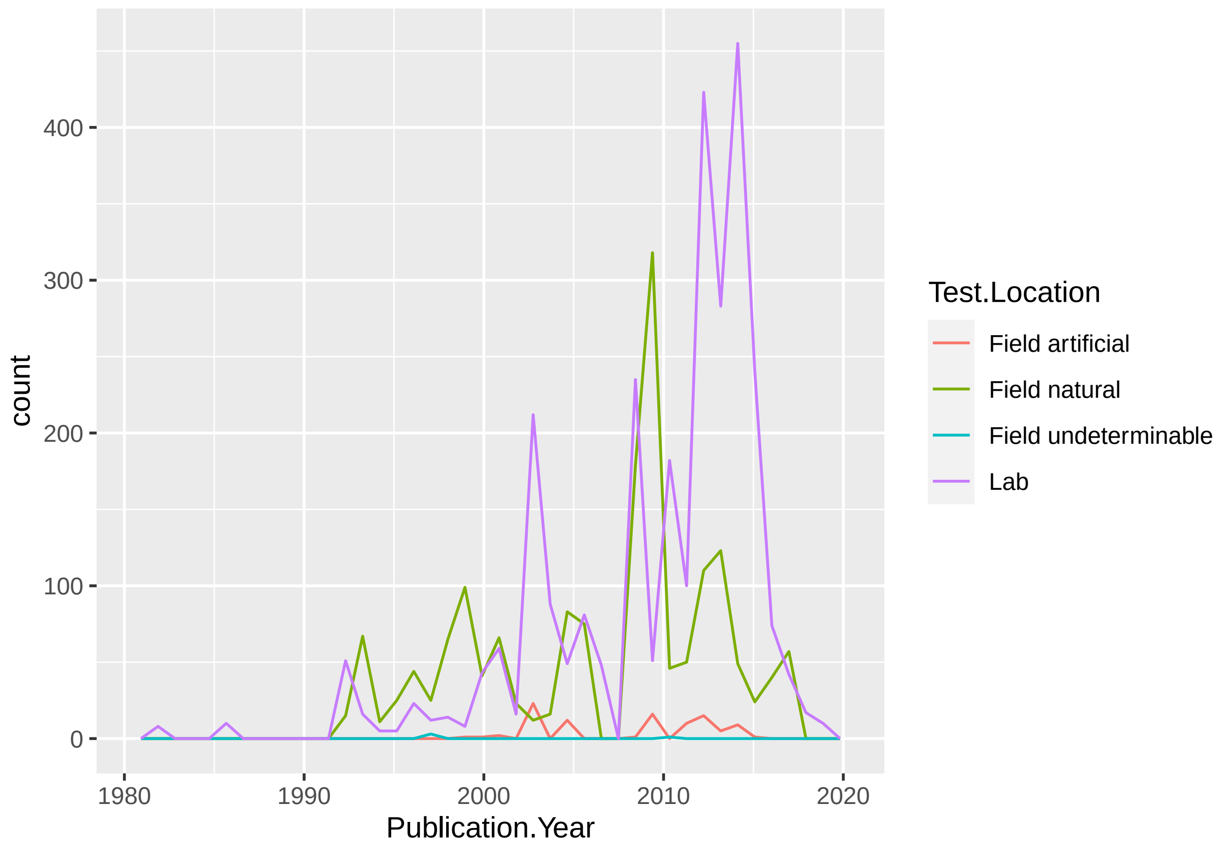
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics)+
  geom_freqpoly(aes(x=Publication.Year),bins=40)
```



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=40)
```

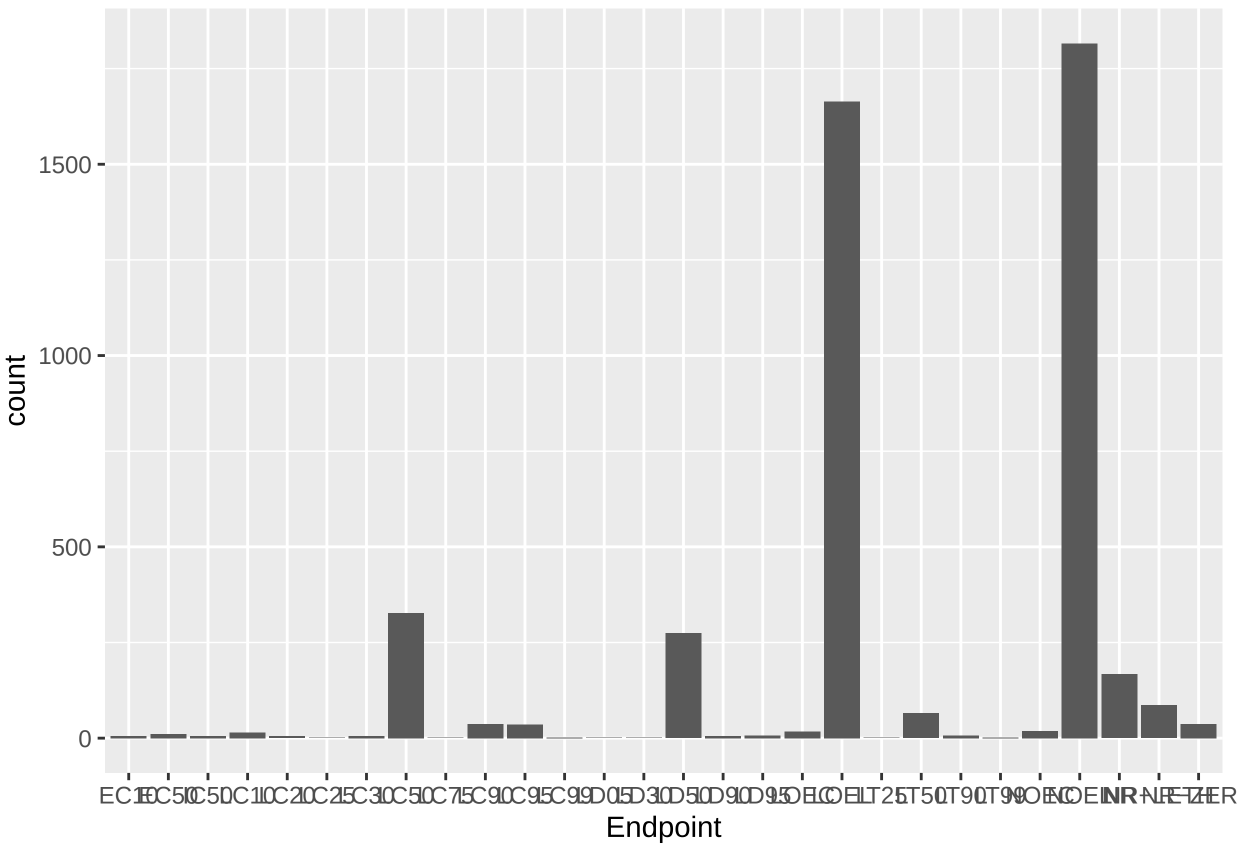


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer:
>
> The most common test locations are labs and natural field test sites. They seem to flip flop in popularity relative to each other. But generally, labs seem to be the premier test location.

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)+
  geom_bar(aes(x=Endpoint))
```



Answer:

NOEL and LOEL, both for Terrestrial

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 for

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"
```

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`?

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.

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

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

Answer:

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

Answer: