Assignment 3: Data Exploration

Max Hermanson

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.

```
getwd()
```

```
## [1] "/Users/mothership/Desktop/EDA_21/Environmental_Data_Analytics_2021_2/Assignments"
```

```
setwd("/Users/mothership/Desktop/EDA_21/Environmental_Data_Analytics_2021_2/Assignments")
getwd()
```

```
## [1] "/Users/mothership/Desktop/EDA_21/Environmental_Data_Analytics_2021_2/Assignments"
```

```
neonic.df <- read.csv("../Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE)
litter.df <- read.csv("../Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv")

library(ggplot2)
```

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: research has shown that neonicotinoid pesticides, which are widely used on farms in the US, may adversely affect pollinators and other important insects. Such effects could have wide-reaching economic, social, and biological implications, and should thus be studied.

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: litter and woody debris can be a good determinant of forest productivity, local soil fertility, and nutrient cycling, and is an important source of habitat for a variety of organisms.

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: * Tower and ground traps collect the litter/debris * These traps are randomly placed within 400m^2 plots. Randomness is maximized by using different randomness placement strategies for different forest cover types
   * The ground traps are sampled once every year, whereas the tower (air) traps are sampled much more frequently, but this frequency depends on site type.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(neonic.df)
```

```
## [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(neonic.df$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
```

```
head(neonic.df)
```

```
##   CAS.Number                                  Chemical.Name
## 1   58842209 Tetrahydro-2-(nitromethylene)-2H-1,3-thiazine
## 2   58842209 Tetrahydro-2-(nitromethylene)-2H-1,3-thiazine
## 3   58842209 Tetrahydro-2-(nitromethylene)-2H-1,3-thiazine
## 4   58842209 Tetrahydro-2-(nitromethylene)-2H-1,3-thiazine
## 5   58842209 Tetrahydro-2-(nitromethylene)-2H-1,3-thiazine
## 6   58842209 Tetrahydro-2-(nitromethylene)-2H-1,3-thiazine
##                                        Chemical.Grade
## 1 Technical grade, technical product, technical formulation
## 2 Technical grade, technical product, technical formulation
## 3 Technical grade, technical product, technical formulation
## 4 Technical grade, technical product, technical formulation
## 5 Technical grade, technical product, technical formulation
```

```
## 6 Technical grade, technical product, technical formulation
##   Chemical.Analysis.Method Chemical.Purity Species.Scientific.Name
## 1               Unmeasured              99  Araecerus fasciculatus
## 2               Unmeasured              99  Araecerus fasciculatus
## 3               Unmeasured              95         Musca domestica
## 4               Unmeasured              95         Musca domestica
## 5               Unmeasured              95         Musca domestica
## 6               Unmeasured              95         Musca domestica
##   Species.Common.Name   Species.Group Organism.Lifestage Organism.Age
## 1  Coffee Bean Weevil Insects/Spiders              Adult           NR
## 2  Coffee Bean Weevil Insects/Spiders              Adult           NR
## 3           House Fly Insects/Spiders              Young           NR
## 4           House Fly Insects/Spiders              Young           NR
## 5           House Fly Insects/Spiders              Young           NR
## 6           House Fly Insects/Spiders              Adult            9
##   Organism.Age.Units    Exposure.Type   Media.Type Test.Location
## 1       Not reported Topical, general No substrate           Lab
## 2       Not reported Topical, general No substrate           Lab
## 3            Hour(s)             Food Filter paper           Lab
## 4            Hour(s)             Food Filter paper           Lab
## 5            Hour(s)             Food Filter paper           Lab
## 6             Day(s)             Food Filter paper           Lab
##   Number.of.Doses Conc.1.Type..Author. Conc.1..Author. Conc.1.Units..Author.
## 1              NR    Active ingredient            27.2             ug/g bdwt
## 2              NR    Active ingredient            19.7             ug/g bdwt
## 3              11    Active ingredient              47                 mg/L
## 4              11    Active ingredient              25                 mg/L
## 5              11    Active ingredient              13                 mg/L
## 6              11    Active ingredient             268                 mg/L
##      Effect Effect.Measurement Endpoint Response.Site Observed.Duration..Days.
## 1 Mortality          Mortality     LD50  Not reported                        1
## 2 Mortality          Mortality     LD50  Not reported                        1
## 3 Mortality          Mortality     LC50  Not reported                        1
## 4 Mortality          Mortality     LC50  Not reported                        1
## 5 Mortality          Mortality     LC50  Not reported                        1
## 6 Mortality          Mortality     LC50  Not reported                        1
##   Observed.Duration.Units..Days.
## 1                        Day(s)
## 2                        Day(s)
## 3                        Day(s)
## 4                        Day(s)
## 5                        Day(s)
## 6                        Day(s)
##                                                  Author
## 1                                  Childers,C.C., and H.N. Nigg
## 2                                  Childers,C.C., and H.N. Nigg
## 3 Johnston,A.M., J. Lohr, J. Moes, K.R. Solomon, and E.R. Zaborski
## 4 Johnston,A.M., J. Lohr, J. Moes, K.R. Solomon, and E.R. Zaborski
## 5 Johnston,A.M., J. Lohr, J. Moes, K.R. Solomon, and E.R. Zaborski
## 6 Johnston,A.M., J. Lohr, J. Moes, K.R. Solomon, and E.R. Zaborski
##   Reference.Number
## 1           107388
## 2           107388
## 3           103312
```

```
## 4           103312
## 5           103312
## 6           103312
##
## 1                                                                                  Contact Toxicity of Insecti
## 2                                                                                  Contact Toxicity of Insecti
## 3 Toxicity of Synergized and Unsynergized Nitromethylene Heterocycle Insecticide (SD 35651) to Suscep
## 4 Toxicity of Synergized and Unsynergized Nitromethylene Heterocycle Insecticide (SD 35651) to Suscep
## 5 Toxicity of Synergized and Unsynergized Nitromethylene Heterocycle Insecticide (SD 35651) to Suscep
## 6 Toxicity of Synergized and Unsynergized Nitromethylene Heterocycle Insecticide (SD 35651) to Suscep
##                              Source Publication.Year
## 1   J. Econ. Entomol.75(3): 556-559             1982
## 2   J. Econ. Entomol.75(3): 556-559             1982
## 3 J. Econ. Entomol.79(6): 1439-1442             1986
## 4 J. Econ. Entomol.79(6): 1439-1442             1986
## 5 J. Econ. Entomol.79(6): 1439-1442             1986
## 6 J. Econ. Entomol.79(6): 1439-1442             1986
##
## 1 Purity: \xca NR - NR | Organism Age: \xca NR - NR Not reported | Conc 1 (Author): \xca Active ingre
## 2 Purity: \xca NR - NR | Organism Age: \xca NR - NR Not reported | Conc 1 (Author): \xca Active ingre
## 3               Purity: \xca NR - NR | Organism Age: \xca 24 - 48 Hour(s) | Conc 1 (Author): \xca Ac
## 4               Purity: \xca NR - NR | Organism Age: \xca 24 - 48 Hour(s) | Conc 1 (Author): \xca Ac
## 5                Purity: \xca NR - NR | Organism Age: \xca 24 - 48 Hour(s) | Conc 1 (Author): \xca A
## 6              Purity: \xca NR - NR | Organism Age: \xca NR - NR Day(s) | Conc 1 (Author): \xca Acti
```

Answer: Mortality, Population, and Behavior have the highest numbers of recorded observations. Mortality and population are both metrics of abundance for insect populations of interest, which can be good indicators of fitness. If a strong correlation exists between neonic use and population changes, this could be very informative in neonic policy. Insights into behavior could be both an indirect measure of fitness and potential changes in pollination activity, which could be extremely important given how reliant many foodwebs are on pollination.

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(neonic.df$Species.Common.Name)
```

```
##                          Honey Bee                     Parasitic Wasp
##                                667                                285
##              Buff Tailed Bumblebee                Carniolan Honey Bee
##                                183                                152
##                         Bumble Bee                   Italian Honeybee
##                                140                                113
##                    Japanese Beetle                  Asian Lady Beetle
##                                 94                                 76
##                    Euonymus Scale                           Wireworm
##                                 75                                 69
##                  European Dark Bee                  Minute Pirate Bug
##                                 66                                 62
##               Asian Citrus Psyllid                      Parastic Wasp
##                                 60                                 58
##             Colorado Potato Beetle                    Parasitoid Wasp
##                                 57                                 51
##                Erythrina Gall Wasp                       Beetle Order
##                                 49                                 47
```

```
##        Snout Beetle Family, Weevil          Sevenspotted Lady Beetle
##                                 47                                46
##                     True Bug Order            Buff-tailed Bumblebee
##                                 45                                39
##                       Aphid Family                  Cabbage Looper
##                                 38                                38
##                Sweetpotato Whitefly                   Braconid Wasp
##                                 37                                33
##                      Cotton Aphid                  Predatory Mite
##                                 33                                33
##              Ladybird Beetle Family                      Parasitoid
##                                 30                                30
##                      Scarab Beetle                   Spring Tiphia
##                                 29                                29
##                       Thrip Order              Ground Beetle Family
##                                 29                                27
##                 Rove Beetle Family                  Tobacco Aphid
##                                 27                                27
##                      Chalcid Wasp          Convergent Lady Beetle
##                                 25                                25
##                     Stingless Bee               Spider/Mite Class
##                                 25                                24
##                Tobacco Flea Beetle                 Citrus Leafminer
##                                 24                                23
##                    Ladybird Beetle                      Mason Bee
##                                 23                                22
##                          Mosquito                  Argentine Ant
##                                 22                                21
##                             Beetle        Flatheaded Appletree Borer
##                                 21                                20
##                Horned Oak Gall Wasp               Leaf Beetle Family
##                                 20                                20
##                   Potato Leafhopper        Tooth-necked Fungus Beetle
##                                 20                                20
##                       Codling Moth         Black-spotted Lady Beetle
##                                 19                                18
##                       Calico Scale              Fairyfly Parasitoid
##                                 18                                18
##                        Lady Beetle           Minute Parasitic Wasps
##                                 18                                18
##                          Mirid Bug                Mulberry Pyralid
##                                 18                                18
##                           Silkworm                  Vedalia Beetle
##                                 18                                18
##               Araneoid Spider Order                       Bee Order
##                                 17                                17
##                     Egg Parasitoid                     Insect Class
##                                 17                                17
##            Moth And Butterfly Order      Oystershell Scale Parasitoid
##                                 17                                17
## Hemlock Woolly Adelgid Lady Beetle             Hemlock Wooly Adelgid
##                                 16                                16
##                               Mite                      Onion Thrip
##                                 16                                16
```

```
##              Western Flower Thrips                         Corn Earworm
##                                 15                                   14
##                  Green Peach Aphid                           House Fly
##                                 14                                   14
##                           Ox Beetle                  Red Scale Parasite
##                                 14                                   14
##                  Spined Soldier Bug                Armoured Scale Family
##                                 14                                   13
##                   Diamondback Moth                       Eulophid Wasp
##                                 13                                   13
##                  Monarch Butterfly                       Predatory Bug
##                                 13                                   13
##              Yellow Fever Mosquito                  Braconid Parasitoid
##                                 13                                   12
##                        Common Thrip         Eastern Subterranean Termite
##                                 12                                   12
##                              Jassid                           Mite Order
##                                 12                                   12
##                           Pea Aphid                   Pond Wolf Spider
##                                 12                                   12
##           Spotless Ladybird Beetle               Glasshouse Potato Wasp
##                                 11                                   10
##                            Lacewing              Southern House Mosquito
##                                 10                                   10
##             Two Spotted Lady Beetle                        Ant Family
##                                 10                                    9
##                         Apple Maggot                          (Other)
##                                  9                                  670
```

Answer: honey bee, parasitic wasp, buff tailed bumblebee, carniolan honey bee, bumble bee, italian honeybee. Five out of six of these are bees/pollinators. They are likely of interest because they are major insect pollinators, which may make them most susceptible to neonic exposure upon pollinating agricultural crops. It is possible the parasitic wasp populations are correlated with bee populations due to a parasitic relationship.

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(neonic.df$Conc.1..Author.)
```

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

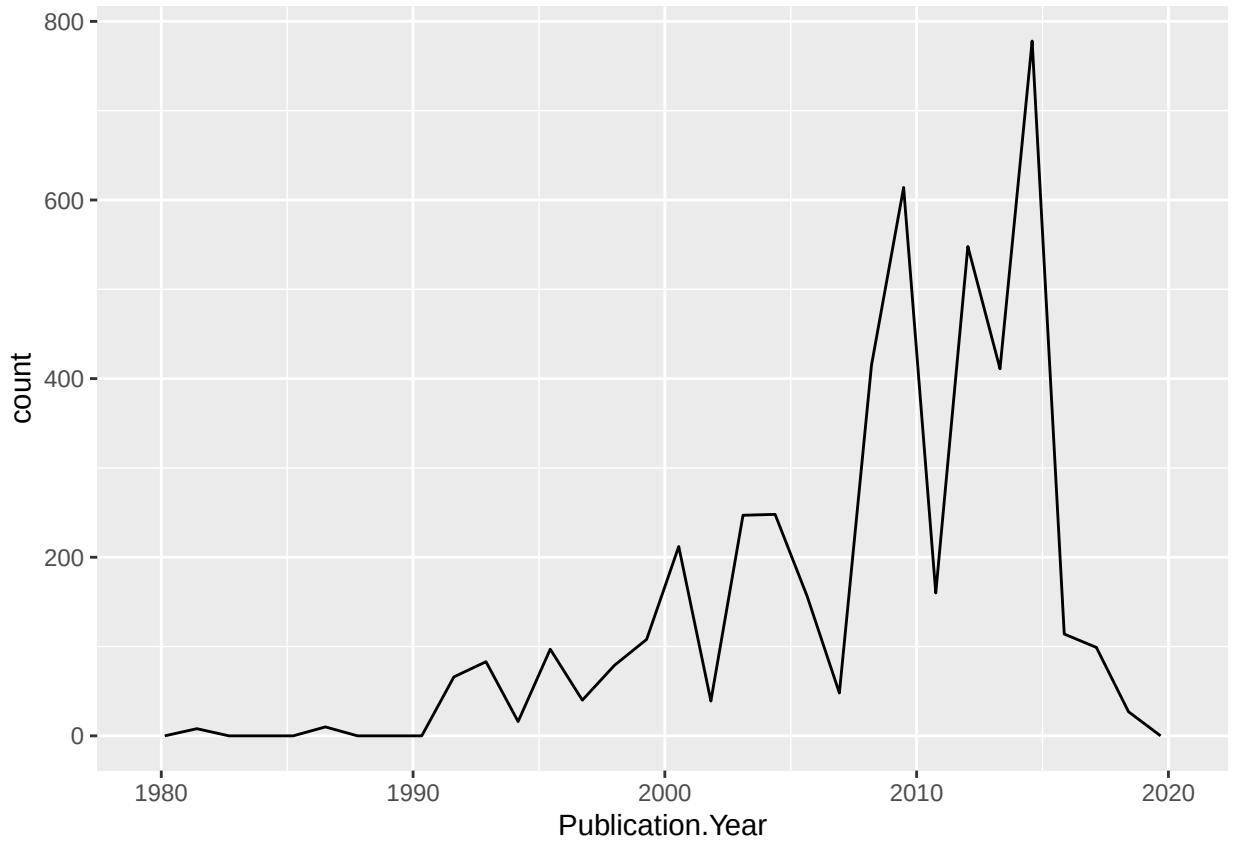
Answer: conc.1..Author is in the factor data format. Perhaps because dosage concentrations were kept constant for each type of neonic studied, and were thus stored as categorical data instead of numeric data.

Explore your data graphically (Neonics)

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

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

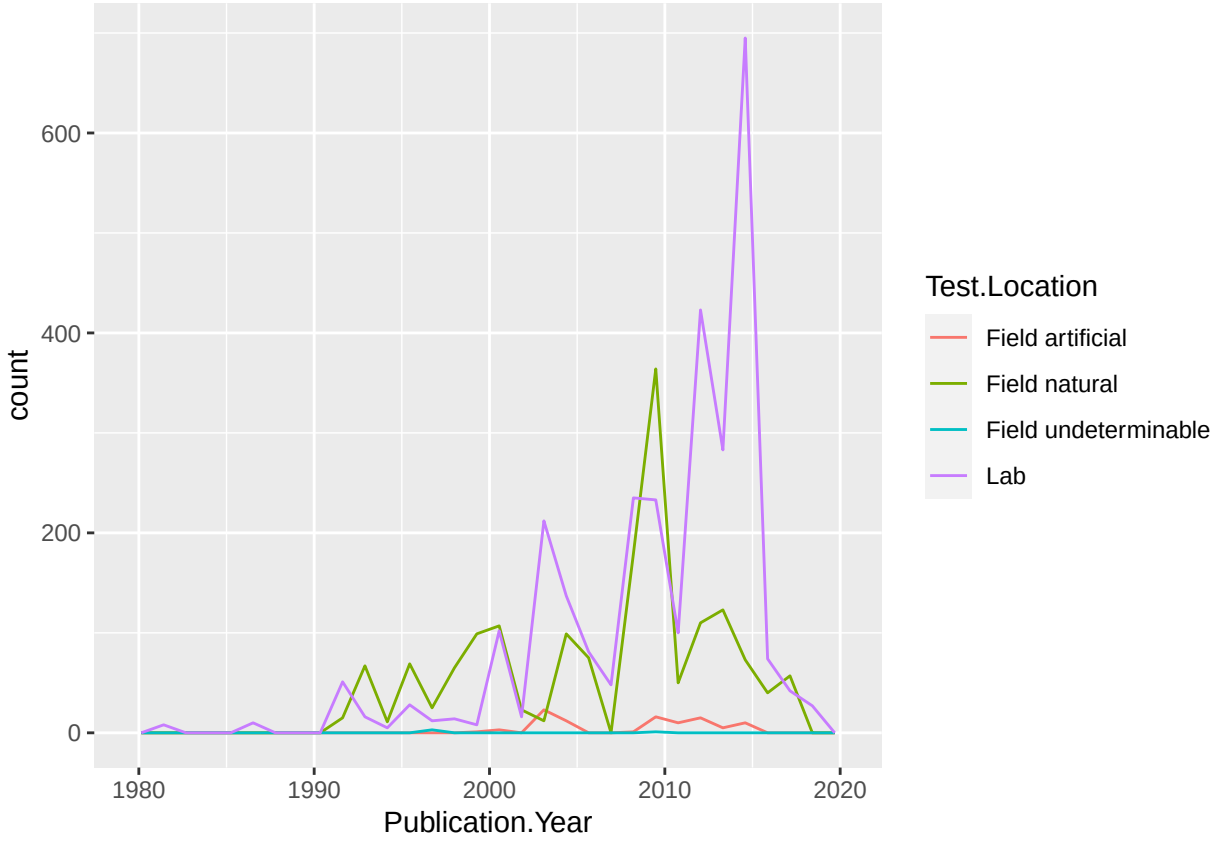
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
ggplot(neonic.df)+
  geom_freqpoly(aes(x=Publication.Year, color = Test.Location))
```

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

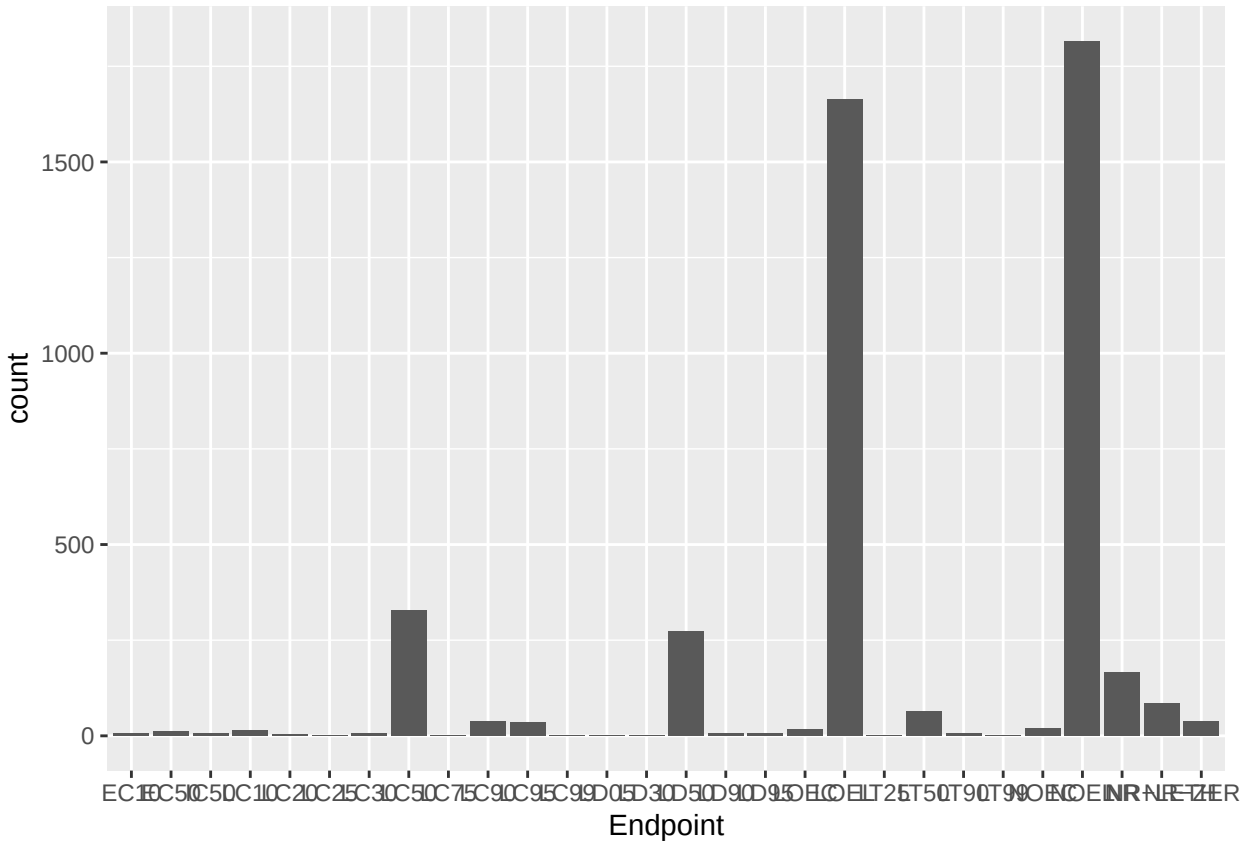


Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: The most common test locations are in the lab and field. Field studies seem to have dominated the number of studies in the 90's, but lab studies began to take over from 2001 onwards. It appears research on neonics heavily declined between 2015 and 2016.

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(neonic.df)+
  geom_bar(aes(x=Endpoint))
```



Answer: LOEL and NOEL are the two most common endpoints.

LOEL endpoints are defined as: "Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC)" - metadata pdf NOEL endpoints are 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.df$collectDate)
```

```
## [1] "character"
```

```
litter.df$collectDate[1:10]
```

```
##  [1] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [6] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
```

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

```
##   [1] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##   [6] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [11] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [16] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [21] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
```

```
##  [26] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [31] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [36] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [41] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [46] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [51] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [56] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [61] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [66] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [71] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [76] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [81] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [86] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [91] "2018-08-02" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
##  [96] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [101] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [106] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [111] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [116] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [121] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [126] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [131] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [136] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [141] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [146] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [151] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [156] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [161] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [166] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [171] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [176] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [181] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [186] "2018-08-30" "2018-08-30" "2018-08-30"
```

```
class(litter.df$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`?

```
Niwot_ridge_plots <- unique(litter.df$plotID)
Niwot_ridge_summary<-summary(litter.df$plotID)
#litter.df$siteID[1:100]
length(Niwot_ridge_plots)
```

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

```
Niwot_ridge_summary
```

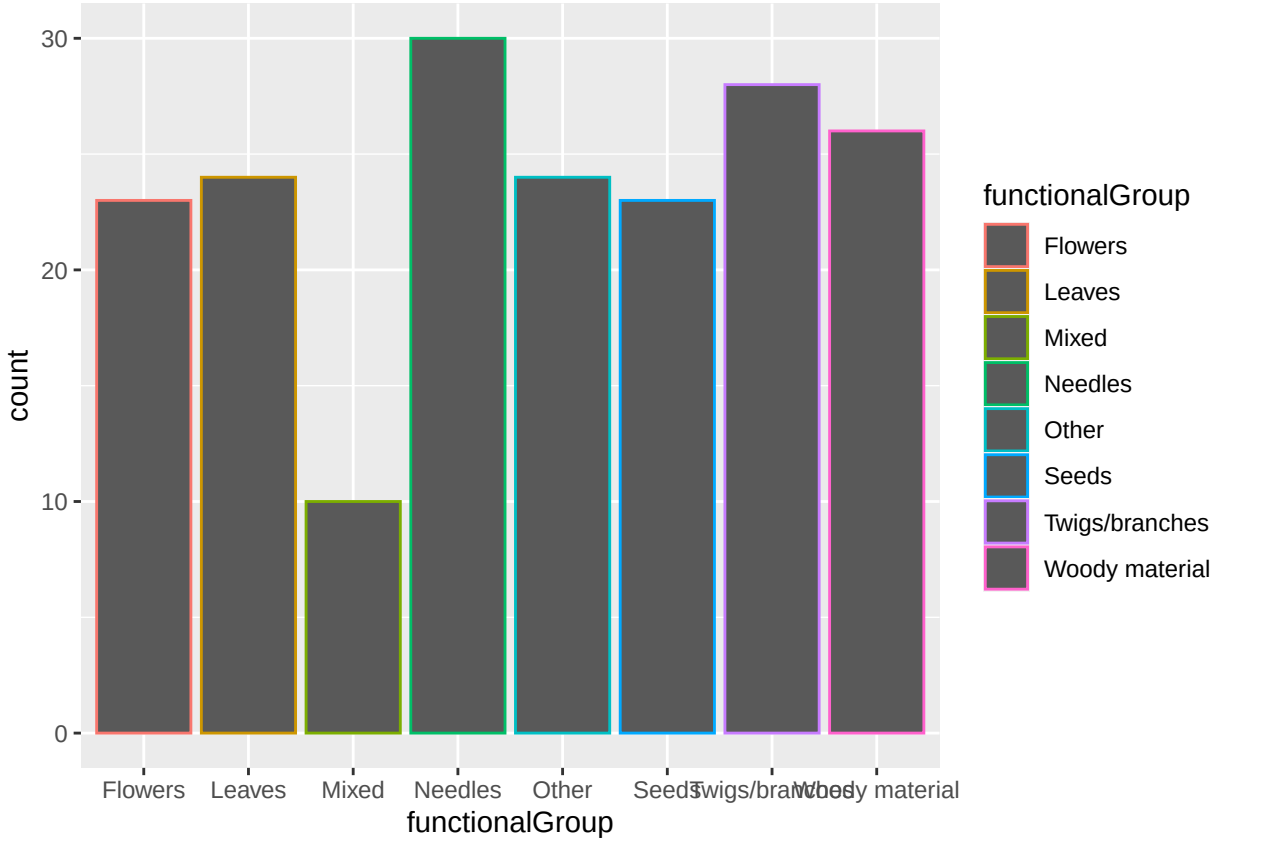
```
##    Length     Class      Mode
##       188 character character
```

Answer: 12 plots were sampled at Niwot Ridge. The 'unique' function delineates subgroups within variables, whereas 'provides' general information about each variable as a whole (length, class, mode).

14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the

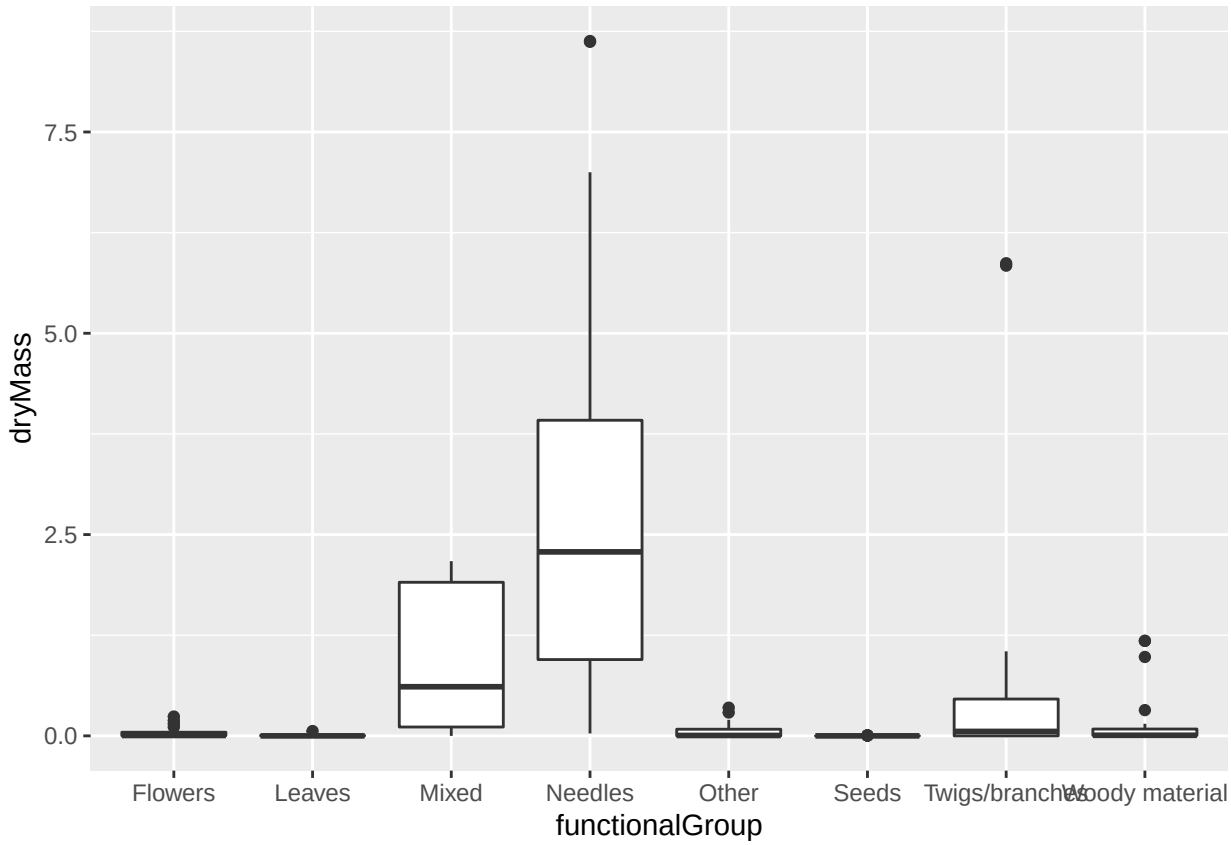
Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(litter.df)+
  geom_bar(aes(x=functionalGroup, color=functionalGroup))
```



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

```
ggplot(litter.df)+
  geom_boxplot(aes(x=functionalGroup, y=dryMass))
```

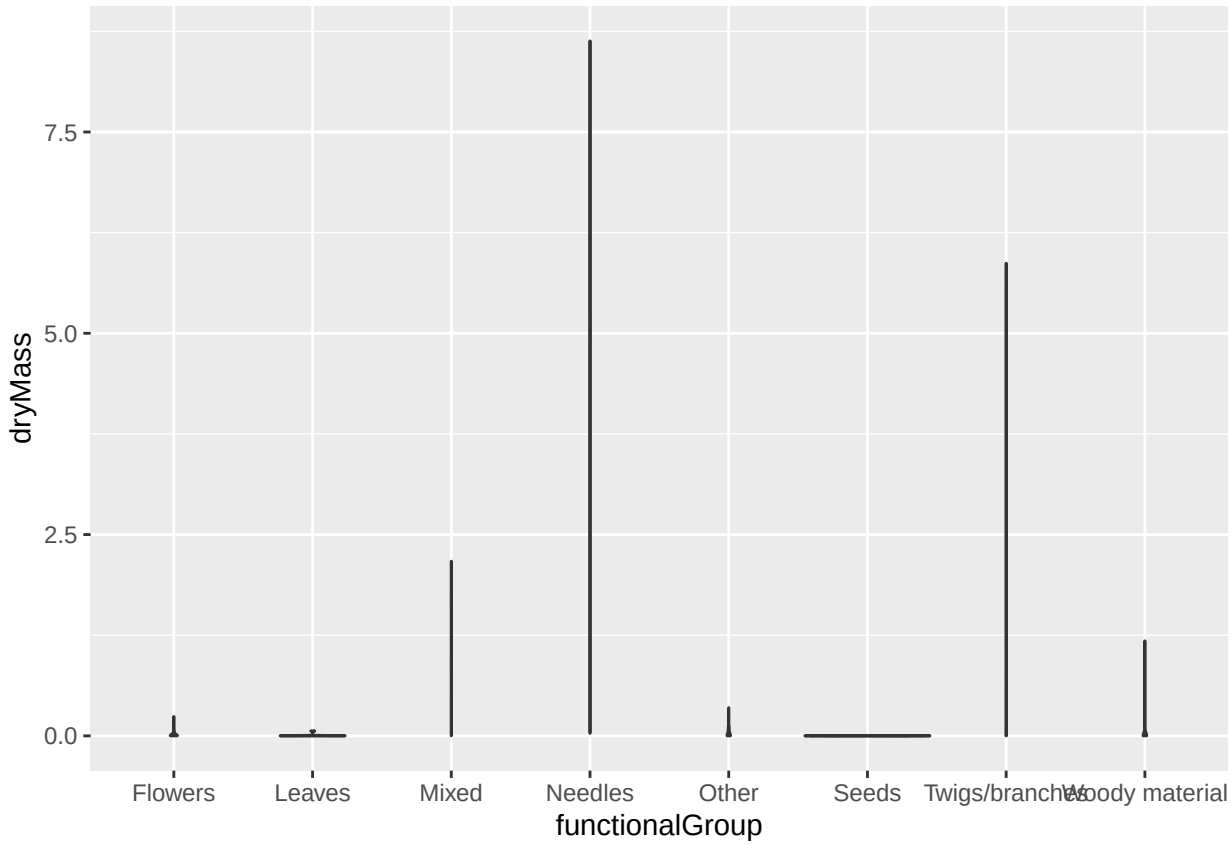



```
ggplot(litter.df)+
  geom_violin(aes(x=functionalGroup, y=dryMass), draw_quantiles = c(0.25,0.5,0.75))
```

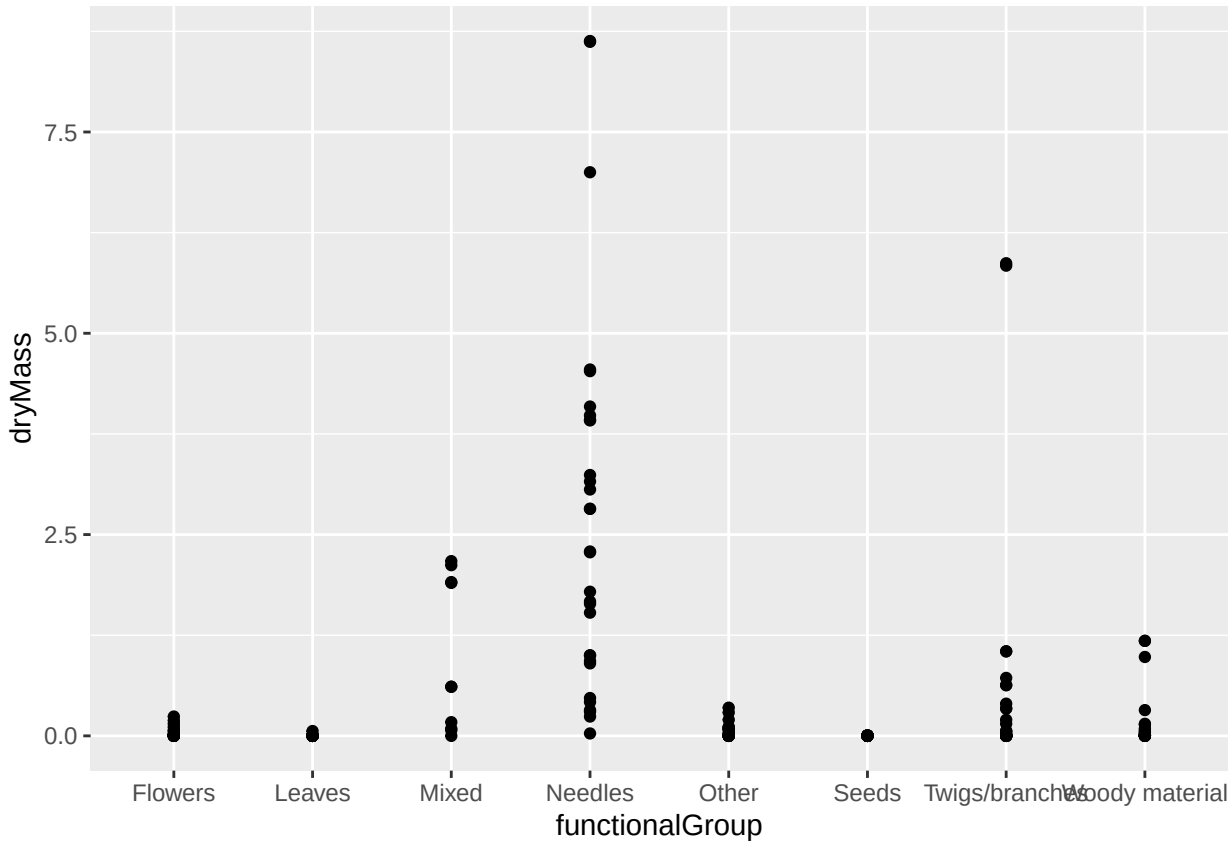
```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```




```
ggplot(litter.df)+
  geom_point(aes(x=functionalGroup, y=dryMass))
```



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

> Answer: The violin plots look extremely stretched, possibly due to a wide range of values and the presence of outliers.

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

> Answer: Mixed litter, Needles, and Twigs/Branches tend to have the highest biomass at these sites. This was determined by examining the boxplots of functionGroup v Litter. A greater proportion of the samples from these three variables weigh significantly more than the other variables, as seen by the interquartile range of their boxplots.