Environmental Data Analytics Coding Challenge #1: Data Exploration

The following code explores the EPA ECOTOX database entries for neonicotinoid mortality.

The code contains numerous mistakes and errors, which you are tasked with fixing.

Instructions listed in comments throughout the script.

Setup —-

```
getwd()
```

```
## [1] "/Users/laurie/Desktop/Envtl_Data_Analytics/MuzzyGitFile/Lessons"
```

```
# it will automatically got to where the file is; use relative file path: project wd
```

```
library("tidyverse")
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
## Warning: package 'tibble' was built under R version 3.4.3
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
ecotox.neonic <- read.csv("../Data/Raw/ECOTOX_Neonicotinoids_Mortality_raw.csv")
# to go back into folder structure, use ../ then tab (to go forward into folder, use ./ then tab)
```

Basic Data Summaries —-

head(ecotox.neonic,) summary(ecotox.neonic%Chemical.Name) summary(ecotox.neonic%Pub..Year)

```
head(ecotox.neonic, 5) #tell it how many lines to show
```

```
##     CAS.No. Chemical.Name      Species.Name Common.Name    Effect
## 1 138261413  Imidacloprid   Cloeon dipterum      Mayfly Mortality
## 2 111988499   Thiacloprid Gammarus fossarum        Scud Mortality
## 3 138261413  Imidacloprid   Cloeon dipterum      Mayfly Mortality
## 4 138261413  Imidacloprid    Caenis horaria      Mayfly Mortality
## 5 111988499   Thiacloprid   Cloeon dipterum      Mayfly Mortality
##   Measurement Endpoint Dur..Std.        Conc..Type Conc..Mean..Std.
## 1   Mortality     LC10        28       Formulation         0.000041
## 2   Mortality  NR-ZERO         7 Active ingredient         0.000070
```

```
## 3   Mortality     LC50        28       Formulation         0.000195
## 4   Mortality     LC10        28       Formulation         0.000235
## 5   Mortality     LC10        21 Active ingredient         0.000240
##   Conc..Units..Std. Pub..Year
## 1           AI mg/L      2013
## 2           AI mg/L      2017
## 3           AI mg/L      2013
## 4           AI mg/L      2013
## 5           AI mg/L      2016
##
## 1                                                                                Roessink,I., L.B
## 2                                  Englert,D., J.P. Zubrod, M. Link, S. Mertins, R. Schulz, and M. Bu
## 3                                                                                Roessink,I., L.B
## 4                                                                                Roessink,I., L.B
## 5 Van den Brink,P.J., J.M. Van Smeden, R.S. Bekele, W. Dierick, D.M. De Gelder, M. Noteboom, and I.
```

```
summary(ecotox.neonic$Chemical.Name) #have to use $ not %
```

```
##  Acetamiprid Clothianidin  Dinotefuran Imidacloprid Imidaclothiz
##          136           74           59          695            9
##   Nitenpyram   Nithiazine  Thiacloprid Thiamethoxam
##           21           22          106          161
```

```
summary(ecotox.neonic$Pub..Year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1982    2004    2011    2008    2014    2018
```

Fix formatting of column names (spaces originally present were turned to periods upon import)

colnames(ecotox.neonic)[8:12] <- c(Duration, Conc.Type, Conc.Mean, Conc.Units, Pub.Year)

```
#lets look at dimensions first
dim(ecotox.neonic) #1283observations  13variables
```

```
## [1] 1283   13
```

```
str(ecotox.neonic) #whoa its a dataframe
```

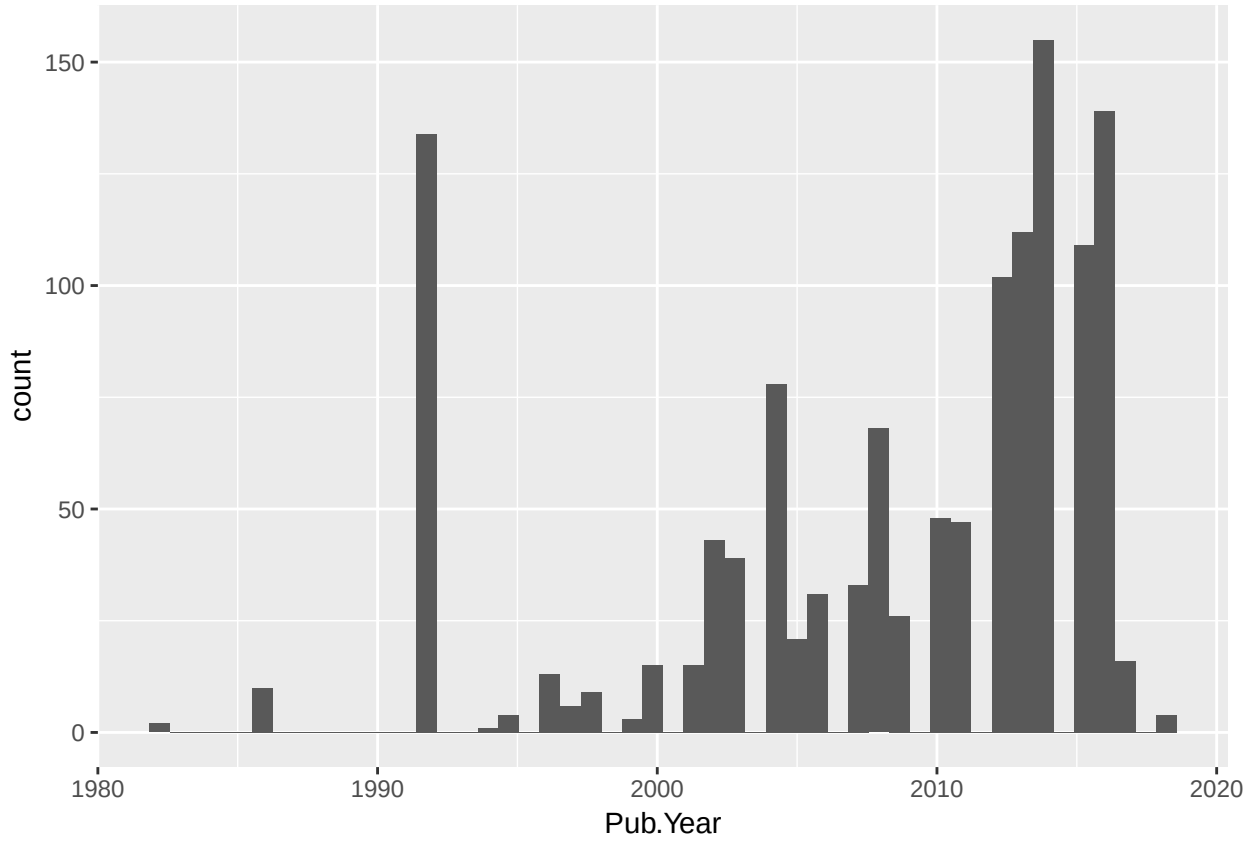
```
## 'data.frame':    1283 obs. of  13 variables:
##  $ CAS.No.          : int  138261413 111988499 138261413 138261413 111988499 111988499 111988499 111
##  $ Chemical.Name    : Factor w/ 9 levels "Acetamiprid",..: 4 8 4 4 8 8 8 8 4 4 ...
##  $ Species.Name     : Factor w/ 172 levels "Acipenser transmontanus",..: 54 86 54 43 54 54 54 54 43 5
##  $ Common.Name      : Factor w/ 124 levels "Alderfly","Alfalfa Plant Bug",..: 68 97 68 68 68 68 68 68
##  $ Effect           : Factor w/ 1 level "Mortality": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Measurement      : Factor w/ 1 level "Mortality": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Endpoint         : Factor w/ 23 levels "EC10","EC50",..: 5 23 9 5 5 5 5 9 9 20 ...
##  $ Dur..Std.        : num  28 7 28 28 21 28 14 28 28 4 ...
##  $ Conc..Type       : Factor w/ 3 levels "Active ingredient",..: 2 1 2 2 1 1 1 1 2 1 ...
##  $ Conc..Mean..Std. : num  0.000041 0.00007 0.000195 0.000235 0.00024 0.00027 0.0003 0.0003 0.000316
##  $ Conc..Units..Std.: Factor w/ 16 levels "AI mg/kg bdwt",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Pub..Year        : int  2013 2017 2013 2013 2016 2016 2016 2016 2013 1992 ...
##  $ Citation         : Factor w/ 198 levels "Aaen,S.M., L.A. Hamre, and T.E. Horsberg. A Screening of
```

```
colnames(ecotox.neonic) <- c("CAS.n", "ChemicalName", "SpeciesName",
                             "CommonName", "Effect", "Measurement",
                             "Endpoint", "Duration", "Conc.Type",
                             "Conc.Mean", "Conc.Units", "Pub.Year",
                             "Citation")
#rename columns, use quotes ; BE CAREFUL, the structure didn't make sense after this command
#(what if I just wanna show SOME of the columns?)
```

Plot histogram of counts of publication years

ggplot(ecotox.neonic, aes(x = Pub.Year)) geom_histogram()

```
ggplot(ecotox.neonic) + #make sure it knows to continue
  geom_histogram(aes(x = Pub.Year), bins = 50) #the histogram has to have an x-axis
```

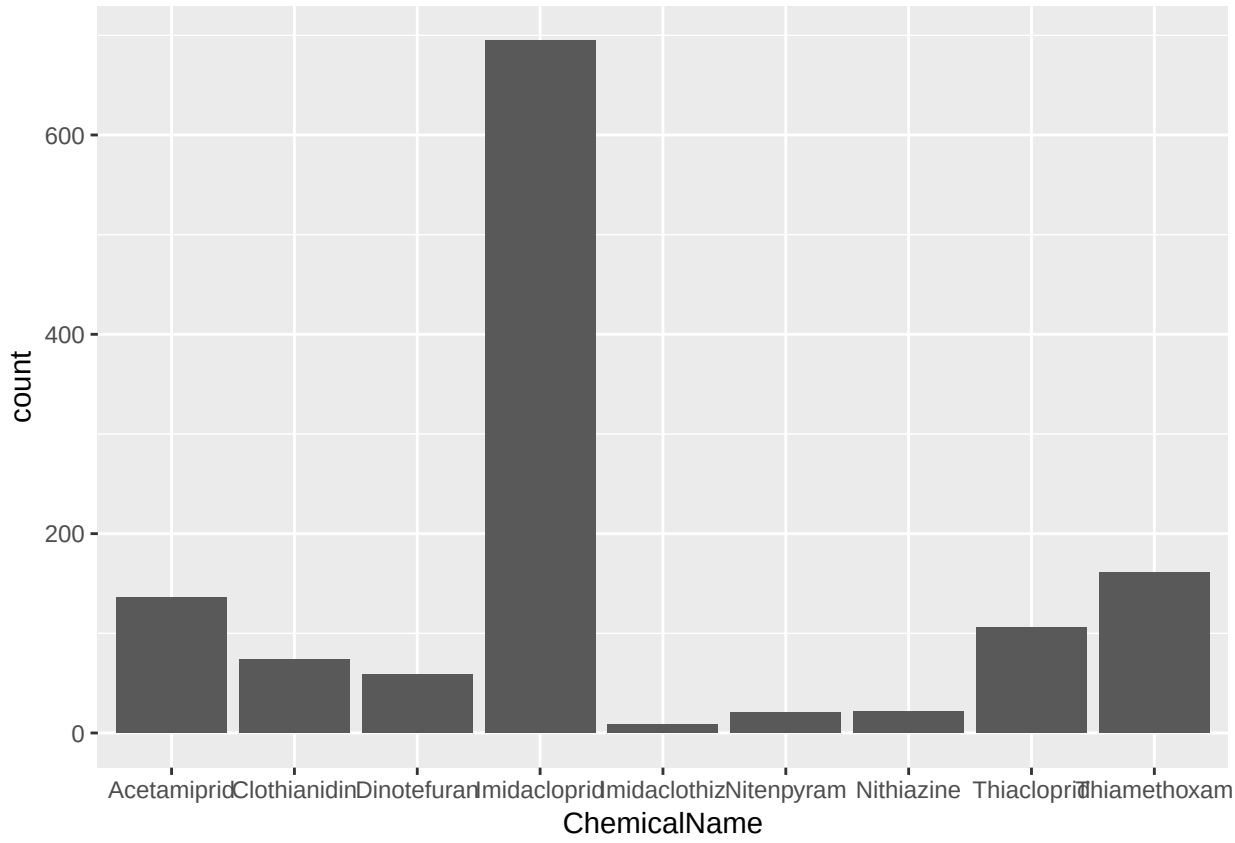


Plot histogram of counts chemical names

hint: what is the class of Chemical.Name? There are two options for a solution.

ggplot(ecotox.neonic, x = Chemical.Name) + geom_histogram()

```r
ggplot(ecotox.neonic, aes(x = ChemicalName)) +
  geom_bar() #make sure it's the name that actually is
```



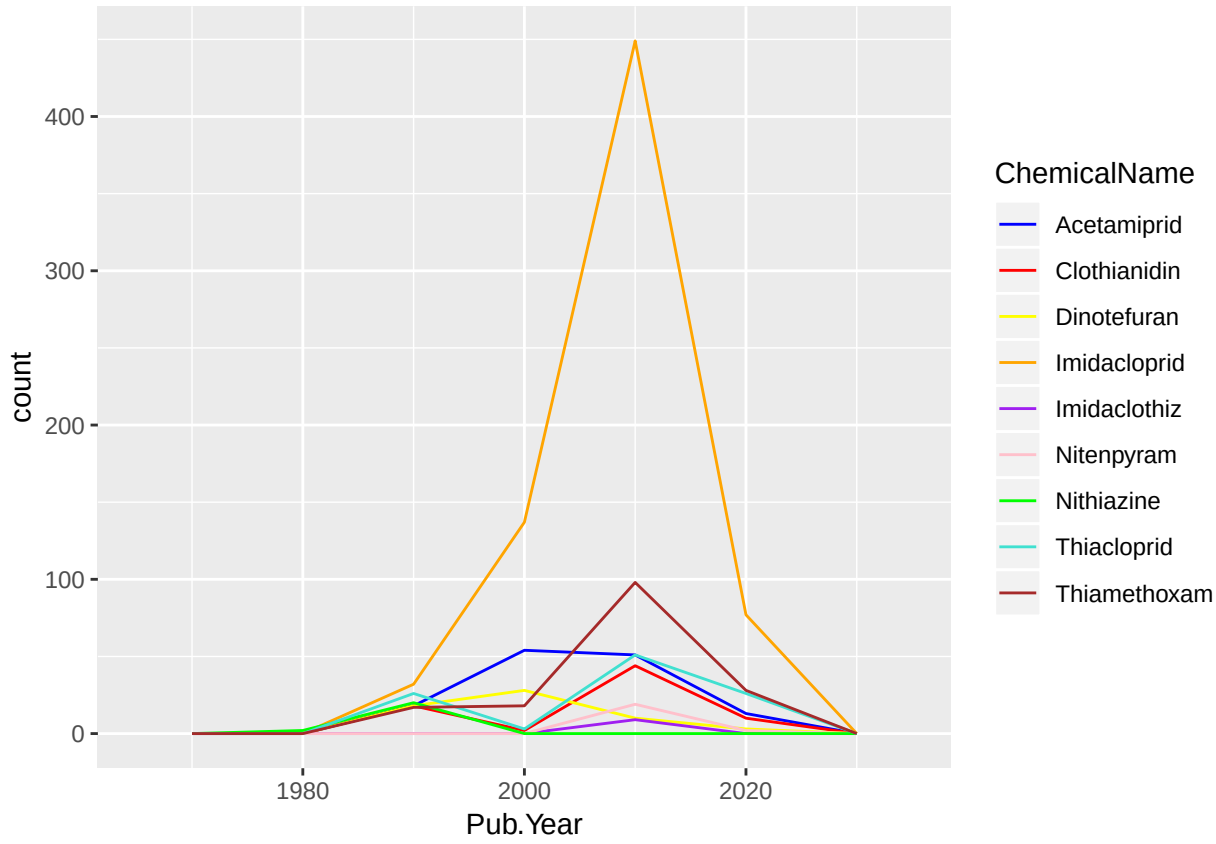

```r
#histogram wouldn't work with discrete variable, that deosn't make sense at all; use geom_bar
```

Plot frequency polygon of publication years divided by chemical name

Define colors as something other than ggplot default

ggplot(ecotox.neonic) + geom_freqpoly(aes(x = Pub.Year, color = Chemical.Name) + theme(legend.position = "right")

```r
ggplot(ecotox.neonic) +
  geom_freqpoly(aes(x = Pub.Year, color = ChemicalName), binwidth = 10) +
  scale_color_manual(values = c("blue", "red", "yellow", "orange", "purple", "pink", "green", "turquois
```




```
  theme(legend.position = "right")
```

```
## List of 1
##  $ legend.position: chr "right"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```