Ginorio_Proposal

MGinorio

4/20/2021

Data Preparation

```
library(readr)
library(tidyverse)
library(ggplot2)
library(lubridate)
```

```
water_data <- read_delim("data/UCMR4_All_MA_WY.txt",
    "\t", escape_double = FALSE, locale = locale(encoding = "Latin1"),
    trim_ws = TRUE)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   MRL = col_double(),
##   `AnalyticalResultValue(µg/L)` = col_double()
## )
## i Use `spec()` for the full column specifications.
```

```
names(water_data)
```

```
##  [1] "PWSID"                       "PWSName"
##  [3] "Size"                        "FacilityID"
##  [5] "FacilityName"                "FacilityWaterType"
##  [7] "SamplePointID"               "SamplePointName"
##  [9] "SamplePointType"             "CollectionDate"
## [11] "SampleID"                    "Contaminant"
## [13] "MRL"                         "MethodID"
## [15] "AnalyticalResultsSign"       "AnalyticalResultValue(µg/L)"
## [17] "SampleEventCode"             "MonitoringRequirement"
## [19] "Region"                      "State"
```

```
epa_water <- water_data %>%
  select(PWSName,
         Size,
         FacilityName,
         FacilityWaterType,
```

```
         CollectionDate,
         Contaminant,
         MRL,
         Result = AnalyticalResultsSign,
         Result_value = `AnalyticalResultValue(µg/L)`,
         State)
```

```
epa_water <- epa_water %>%
  mutate(Result = replace(Result, Result == "=", TRUE),
         Date = mdy(CollectionDate),
         CollectionDate = NULL)
```

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

I would like to know, for a specific city and state, the violations and enforcement actions, as well as the definitions, health effects, and sources of contamination for any contaminants.

Cases

What are the cases, and how many are there?

States with contaminant violations in each water treatment facility - 502,043 entries, 10 total columns

Data collection

Describe the method of data collection.

The motivation for this project is to understand our Drinking Water Requirements for States and Public Water Systems by analyzing the data provided by the United States Environmental Protection Agency EPA. When public water systems are found to contain contaminants in amounts exceeding the Maximum Contamination Limit MCL, they are in violation and must take action to restore the quality of their water. Utilizing the Envirofacts REST API, the SDWIS information can be queried and retrieved from the SDWIS database.

Type of study

What type of study is this (observational/experiment)?

Observational

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

EPA REST API National Contaminant Occurrence Database (NCOD)

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

MRL -> Minimum Reporting Level

Independent Variable

You should have two independent variables, one quantitative and one qualitative.

Result_Value -> Quatitative States -> Qualitative Contaminant -> Qualitative

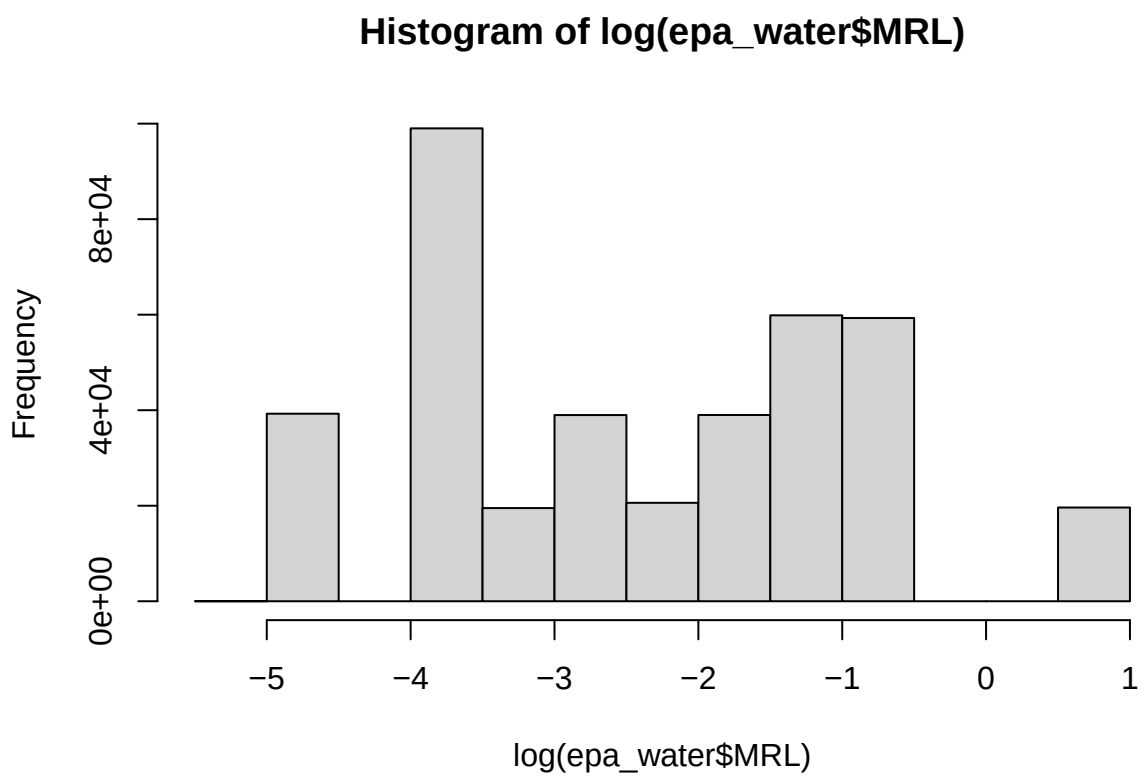
Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

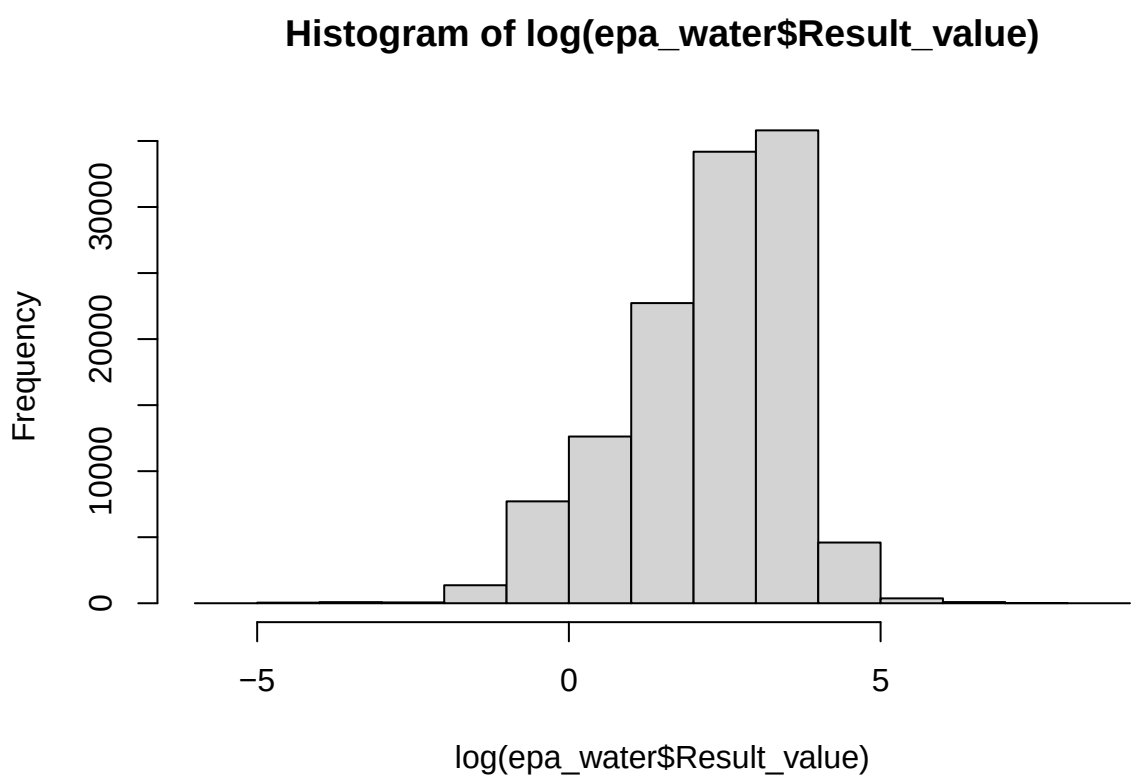
```
summary(epa_water)
```

```
##    PWSName              Size           FacilityName       FacilityWaterType
##  Length:502043      Length:502043      Length:502043      Length:502043
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Contaminant             MRL            Result           Result_value
##  Length:502043      Min.   :0.00     Length:502043      Min.   :   0.0
##  Class :character   1st Qu.:0.03     Class :character   1st Qu.:   3.6
##  Mode  :character   Median :0.09     Mode  :character   Median :  11.3
##                     Mean   :0.25                        Mean   :  18.2
##                     3rd Qu.:0.30                        3rd Qu.:  25.4
##                     Max.   :2.00                        Max.   :3960.0
##                     NA's   :106834                      NA's   :378694
##     State               Date
##  Length:502043      Min.   :2018-01-02
##  Class :character   1st Qu.:2018-11-07
##  Mode  :character   Median :2019-06-10
##                     Mean   :2019-06-03
##                     3rd Qu.:2019-12-16
##                     Max.   :2020-12-08
##
```

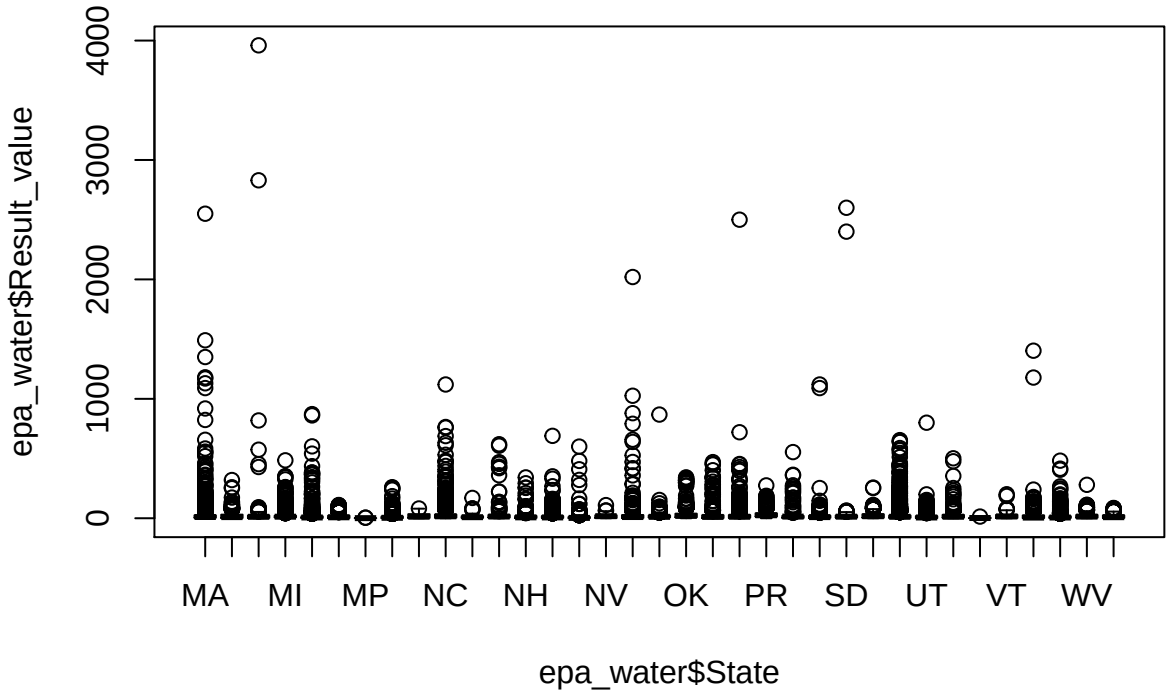
```
hist(log(epa_water$MRL))
```




```
hist(log(epa_water$Result_value))
```

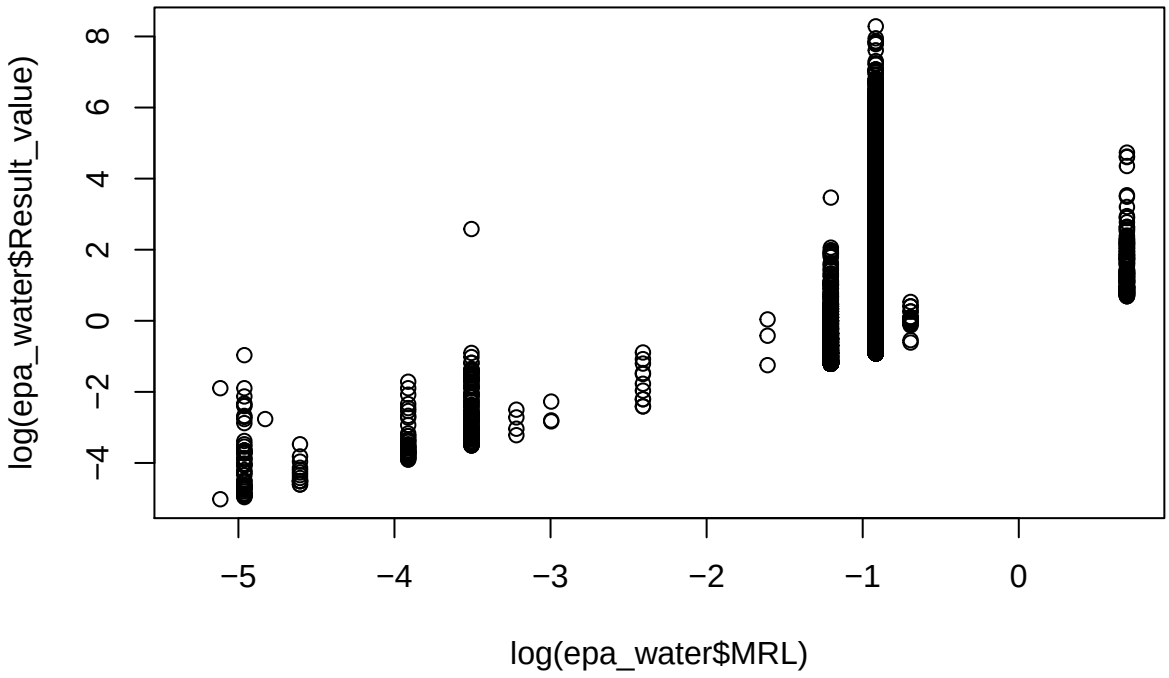



```
boxplot(epa_water$Result_value ~ epa_water$State)
```

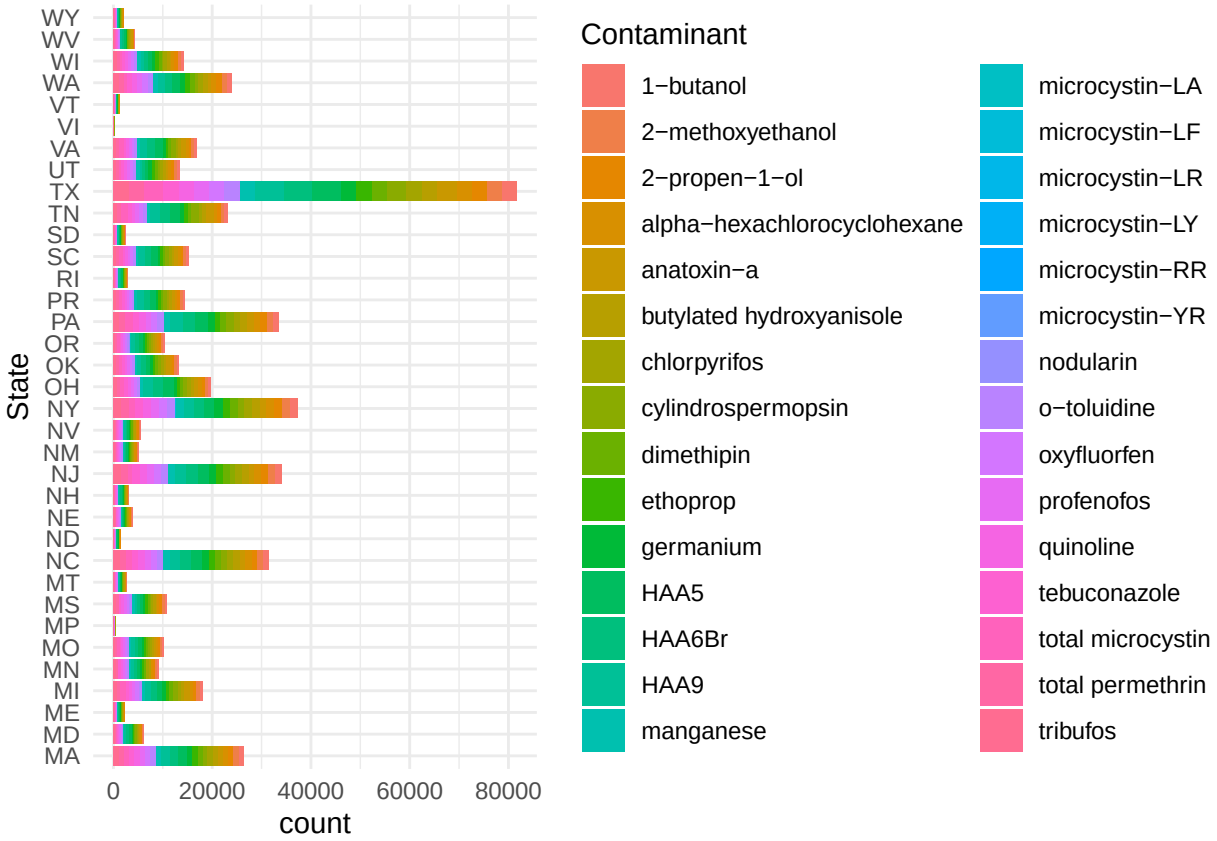



```
# MRL independent Variable (x)
# Result_Value dependent variable (y)

plot(log(epa_water$Result_value) ~ log(epa_water$MRL))
```




```
ggplot(epa_water) +
 aes(x = State, fill = Contaminant) +
 geom_bar() +
 scale_fill_hue(direction = 1) +
 coord_flip() +
 theme_minimal()
```


State
WY
WV
WI
WA
VT
VI
VA
UT
TX
TN
SD
SC
RI
PR
PA
OR
OK
OH
NY
NV
NM
NJ
NH
NE
ND
NC
MT
MS
MP
MO
MN
MI
ME
MD
MA
0
20000
40000
60000
80000
count
Contaminant
1−butanol
2−methoxyethanol
2−propen−1−ol
alpha−hexachlorocyclohexane
anatoxin−a
butylated hydroxyanisole
chlorpyrifos
cylindrospermopsin
dimethipin
ethoprop
germanium
HAA5
HAA6Br
HAA9
manganese
microcystin−LA
microcystin−LF
microcystin−LR
microcystin−LY
microcystin−RR
microcystin−YR
nodularin
o−toluidine
oxyfluorfen
profenofos
quinoline
tebuconazole
total microcystin
total permethrin
tribufos