# Water Quality - Fall Research

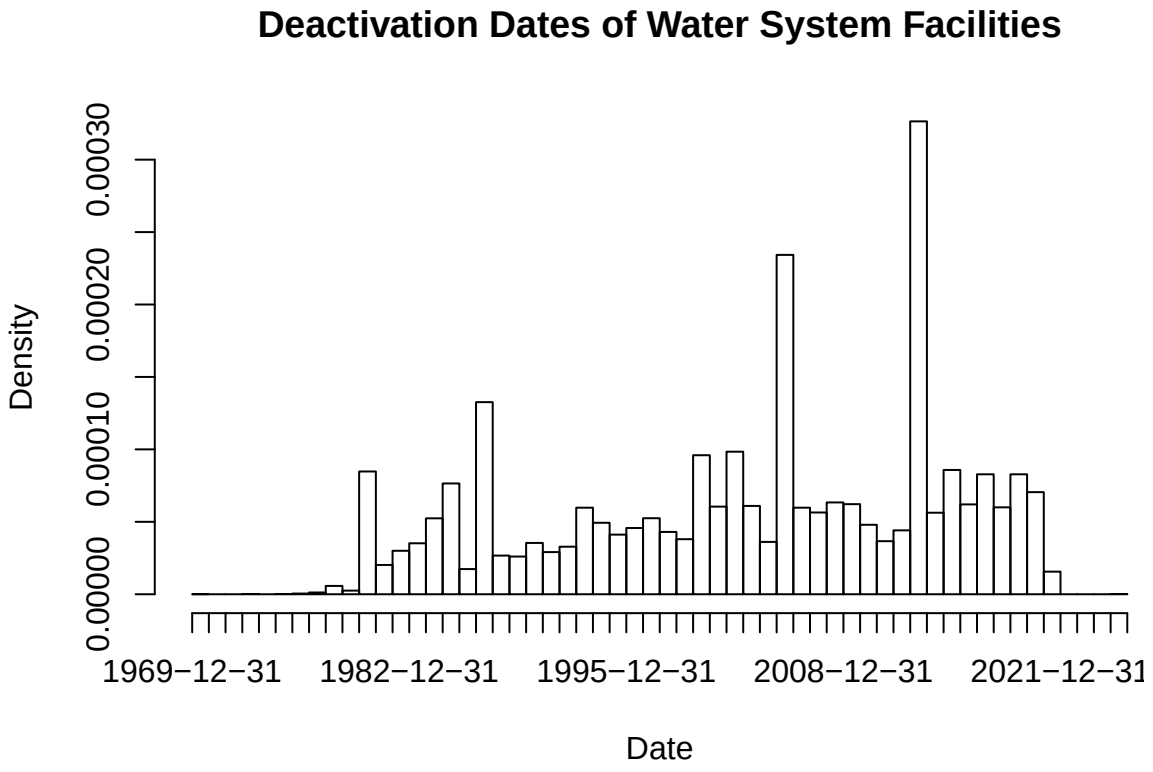
## EPA API call and Exploration

```
library(httr)
library(jsonlite)

# getting one PWS
epa_data <- GET("https://data.epa.gov/efservice/WATER_SYSTEM_FACILITY/PWSID/TX0610262/CSV")
epa_data$status_code
epa_data <- content(epa_data)

# getting all PWS facilities in Texas
epa_data <- GET("https://data.epa.gov/efservice/WATER_SYSTEM_FACILITY/PRIMACY_AGENCY_CODE/TX/ROWS/0:100
epa_data$status_code
epa_data <- content(epa_data)

epa_data2 <- GET("https://data.epa.gov/efservice/WATER_SYSTEM_FACILITY/PRIMACY_AGENCY_CODE/TX/ROWS/10000
epa_data2$status_code
epa_data2 <- content(epa_data2)

epa_facility_data <- rbind(epa_data, epa_data2)

write.csv(epa_facility_data, "epa_facility_data.csv")
```

```
epa_facility_data <- read.csv("~/Downloads/epa_facility_data.csv")


length(unique(epa_facility_data$WATER_SYSTEM_FACILITY.PWSID)) # almost 16k total individual PWS
```

```
## [1] 15770
```

```
hist(as.Date(epa_facility_data$WATER_SYSTEM_FACILITY.FACILITY_DEACTIVATION_DATE, format = "%d-%b-%y"),
     main = "Deactivation Dates of Water System Facilities",
     xlab = "Date")


hist(as.Date(epa_facility_data$WATER_SYSTEM_FACILITY.FACILITY_DEACTIVATION_DATE, format = "%d-%b-%y"),
     main = "Deactivation Dates of Water System Facilities",
     xlab = "Date")
```




### Cleaning and Merging Intakes and TCEQ Data

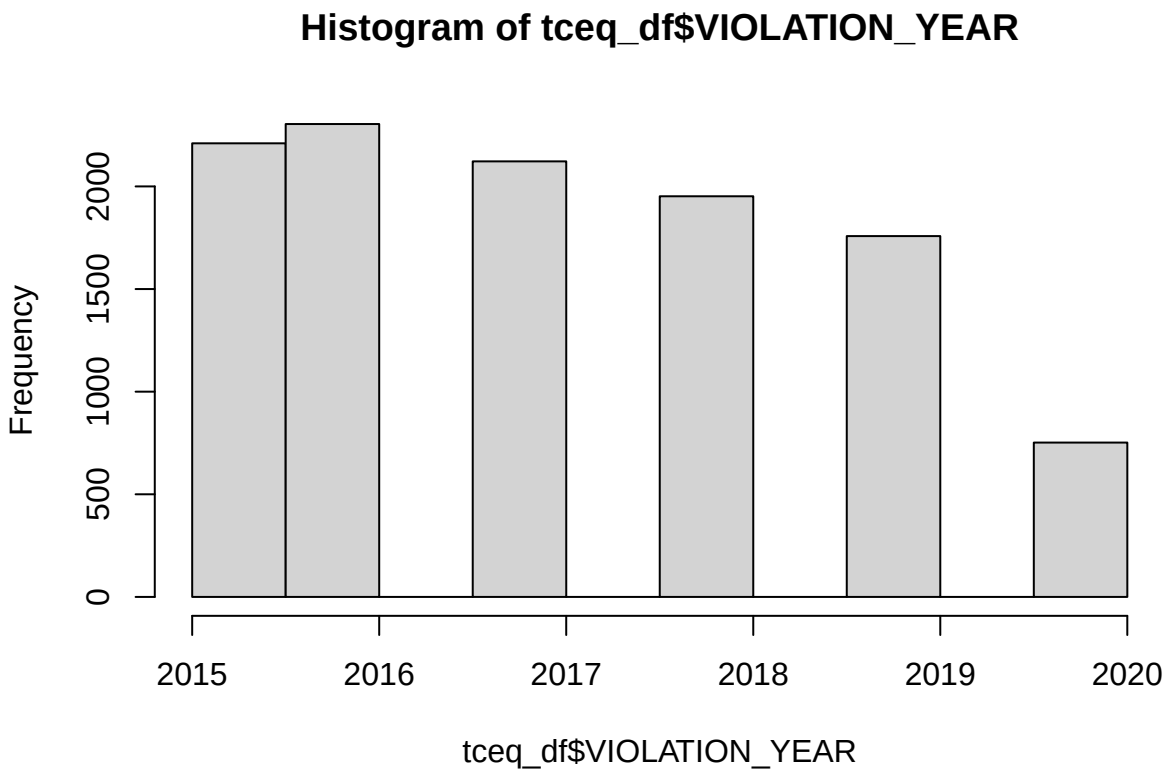
```
library(magrittr)
library(ggplot2)
intakes <- read.csv("~/Downloads/violation_intake_utilities.csv")
tceq_df <- read.csv("~/Downloads/PIR54466_Health_Based_Violations.csv")

mean(unique(tceq_df$WSID) %in% unique(intakes$WSID))
```

```
## [1] 0
```

```
# these are not thee  same for some reason
# three spaces after every set of numbers? for some reason
# head(tceq_df$WSID)
gsub("[ \t]{3,}", "", "TX0010001   ") # practice line to gsub to fix merge
```

```
## [1] "TX0010001"
```

```
# cleaning
# fixing extra spaces behind WSID
tceq_df$WSID <- gsub("[ \t]{3,}", "", tceq_df$WSID)
tceq_df$BEGIN_DT <- as.Date(tceq_df$BEGIN_DT, format ="%m/%d/%Y")
tceq_df$END_DT <- as.Date(tceq_df$END_DT, format ="%m/%d/%Y")

# data exploration
#YEARS <- unique(tceq_df$VIOLATION_YEAR)
hist(tceq_df$VIOLATION_YEAR)
```




```
#intakes and tceq data merged
intakes_tceq <- merge(tceq_df, intakes, by.x="WSID", by.y="WSID", all.y=TRUE)
# taking out commas and making numeric
intakes_tceq$POPULATION <- as.numeric(gsub(",", "", intakes_tceq$POPULATION))

write.csv(intakes_tceq, "~/Downloads/merged_intakes_tceq_oct_18.csv")
```

### Basic Linear Regression & Plot Analysis

```
out <- lm(Number.of.Violations ~ Number.of.Facilities +
            Number.of.Site.Visits + PopulationServed.Count +
            LAT_DD + LONG_DD + fips + BEGIN_DT + END_DT + POPULATION +
            VIOLATION_YEAR+ Primary.Source + HORZ_ACC,
          data = intakes_tceq)
summary(out)
```

```
##
## Call:
## lm(formula = Number.of.Violations ~ Number.of.Facilities + Number.of.Site.Visits +
##     PopulationServed.Count + LAT_DD + LONG_DD + fips + BEGIN_DT +
##     END_DT + POPULATION + VIOLATION_YEAR + Primary.Source + HORZ_ACC,
##     data = intakes_tceq)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -160.701  -40.440   -1.664   36.570  190.536
##
## Coefficients:
##                                       Estimate Std. Error  t value Pr(>|t|)
```

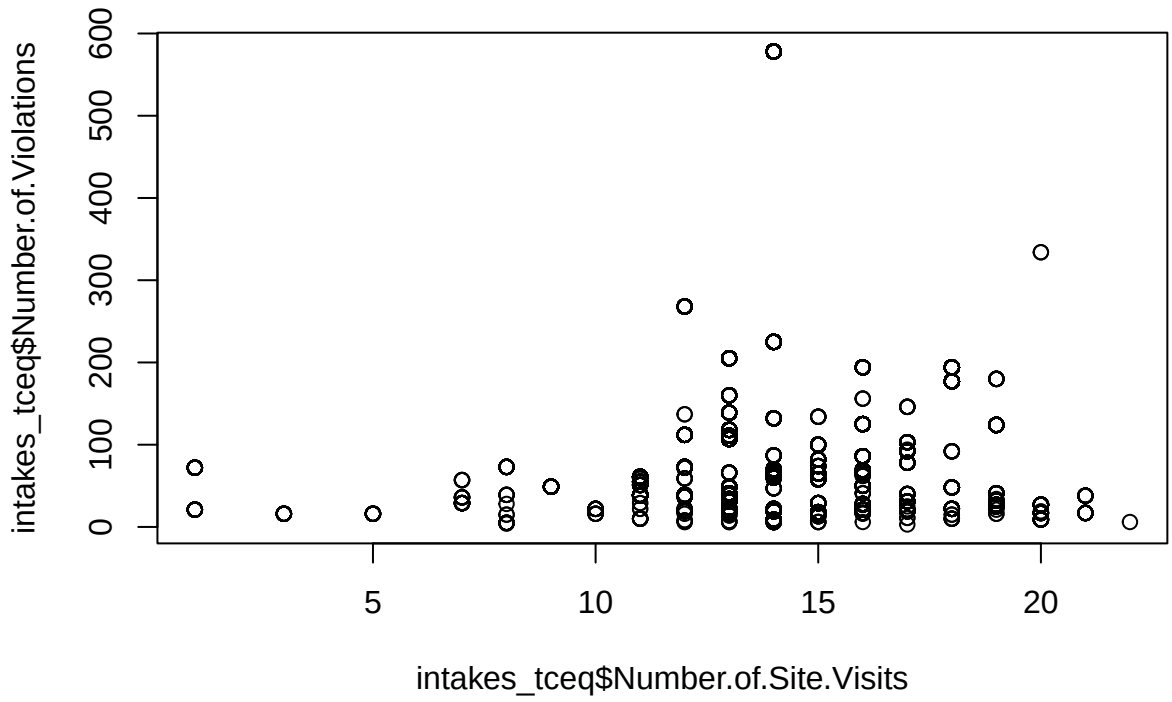
```
## (Intercept)                              1.793e+04  5.054e+03    3.548 0.000391
## Number.of.Facilities                     6.501e-02  3.567e-02    1.823 0.068409
## Number.of.Site.Visits                   -1.226e+00  3.108e-01   -3.944 8.10e-05
## PopulationServed.Count                   3.281e-02  6.127e-03    5.356 8.81e-08
## LAT_DD                                  -4.756e+00  4.545e-01  -10.464  < 2e-16
## LONG_DD                                  8.821e-01  4.578e-01    1.927 0.054075
## fips                                    -9.256e-02  5.562e-03  -16.641  < 2e-16
## BEGIN_DT                                 2.908e-01  2.486e-02   11.701  < 2e-16
## END_DT                                  -2.695e-01  2.386e-02  -11.296  < 2e-16
## POPULATION                              -3.300e-02  6.126e-03   -5.387 7.41e-08
## VIOLATION_YEAR                          -6.433e+00  2.565e+00   -2.508 0.012160
## Primary.SourceGround water purchased    -4.839e+02  6.826e+00  -70.884  < 2e-16
## Primary.SourceSurface water             -4.763e+02  3.166e+00 -150.453  < 2e-16
## Primary.SourceSurface water purchased   -4.432e+02  6.330e+00  -70.024  < 2e-16
## HORZ_ACC                                 4.255e-03  2.604e-04   16.338  < 2e-16
##
## (Intercept)                             ***
## Number.of.Facilities                    .
## Number.of.Site.Visits                   ***
## PopulationServed.Count                  ***
## LAT_DD                                  ***
## LONG_DD                                 .
## fips                                    ***
## BEGIN_DT                                ***
## END_DT                                  ***
## POPULATION                              ***
## VIOLATION_YEAR                          *
## Primary.SourceGround water purchased    ***
## Primary.SourceSurface water             ***
## Primary.SourceSurface water purchased   ***
## HORZ_ACC                                ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.73 on 6556 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.8273, Adjusted R-squared:  0.8269
## F-statistic:  2243 on 14 and 6556 DF,  p-value: < 2.2e-16
```

```
out <- lm(Number.of.Violations ~ VIOLATION_YEAR, data = intakes_tceq)
summary(out)
```

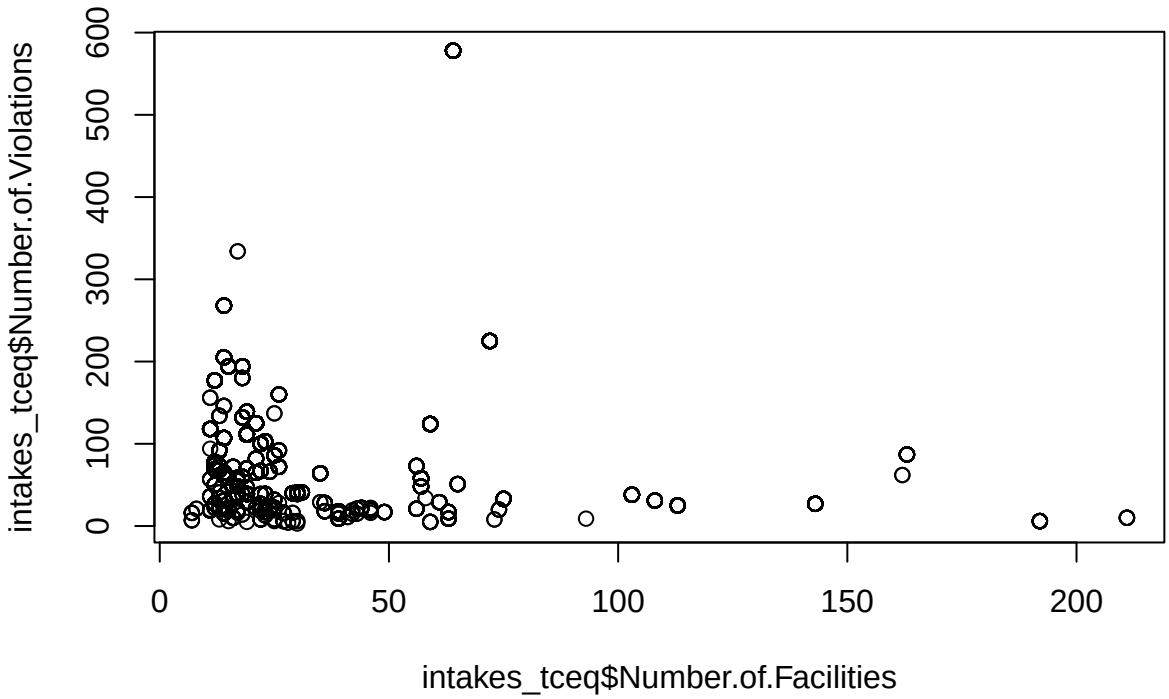
```
##
## Call:
## lm(formula = Number.of.Violations ~ VIOLATION_YEAR, data = intakes_tceq)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -148.34  -92.60  -38.60   51.03  447.40
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -12336.260   2309.978  -5.340 9.58e-08 ***
## VIOLATION_YEAR      6.187      1.145   5.402 6.81e-08 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143.1 on 6584 degrees of freedom
## Multiple R-squared:  0.004413,   Adjusted R-squared:  0.004262
## F-statistic: 29.19 on 1 and 6584 DF,  p-value: 6.806e-08
```

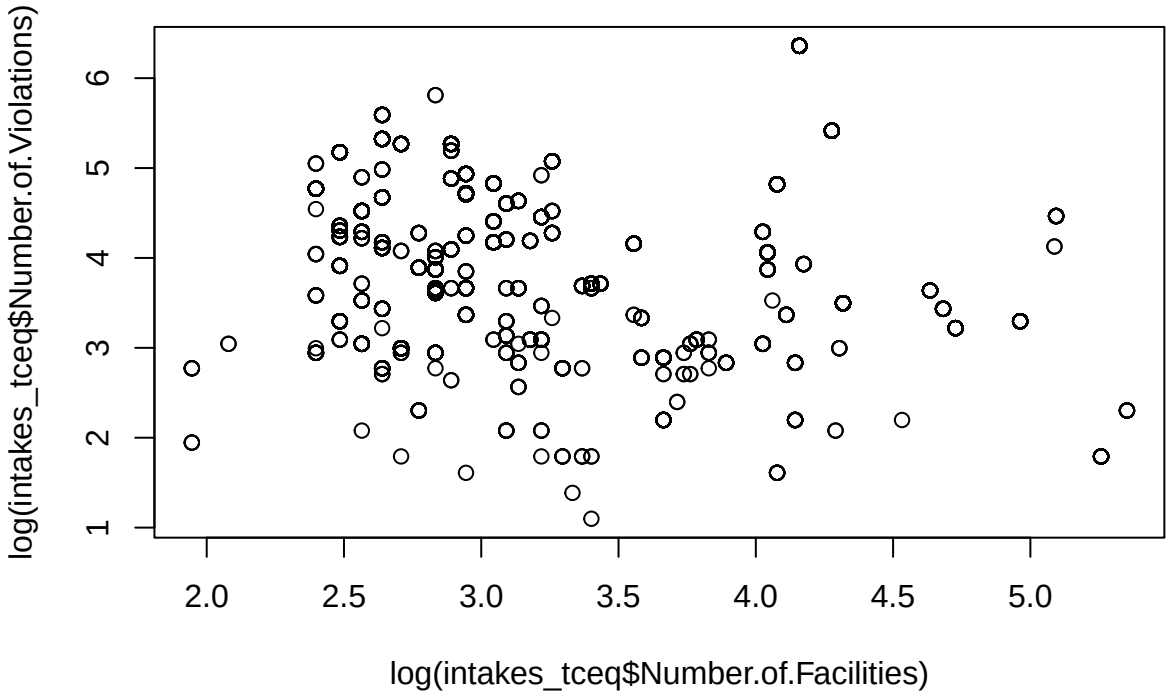
```
# thought "logical" class meant binary -> actually all NA values
#out <- glm(Number.of.Violations ~ OWNR_DES, data = intakes_tceq)


plot(intakes_tceq$Number.of.Site.Visits, intakes_tceq$Number.of.Violations)
```




```
plot(intakes_tceq$Number.of.Facilities, intakes_tceq$Number.of.Violations)
```




```
plot(log(intakes_tceq$Number.of.Facilities), log(intakes_tceq$Number.of.Violations))
```




#Plot 1 : site visits vs violations - actually a very very interesting plot! with almost a normal, unimodal distribution with a mean at around 15 and a standard deviation of around 3 (can calculate these, these are just estimates for now)

# Plot 2: facilities vs violations - does not appear to be highly correlated, high majority of violations are 100 or less and 50 or less facilities

# Plot 3: was interested in the log transformed plot, which seems to confirm that there is not really much of a shape in this set of data, although there might be a type of clustering before and after 4 (would be interesting to look into)

### Graphing from New Merge

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

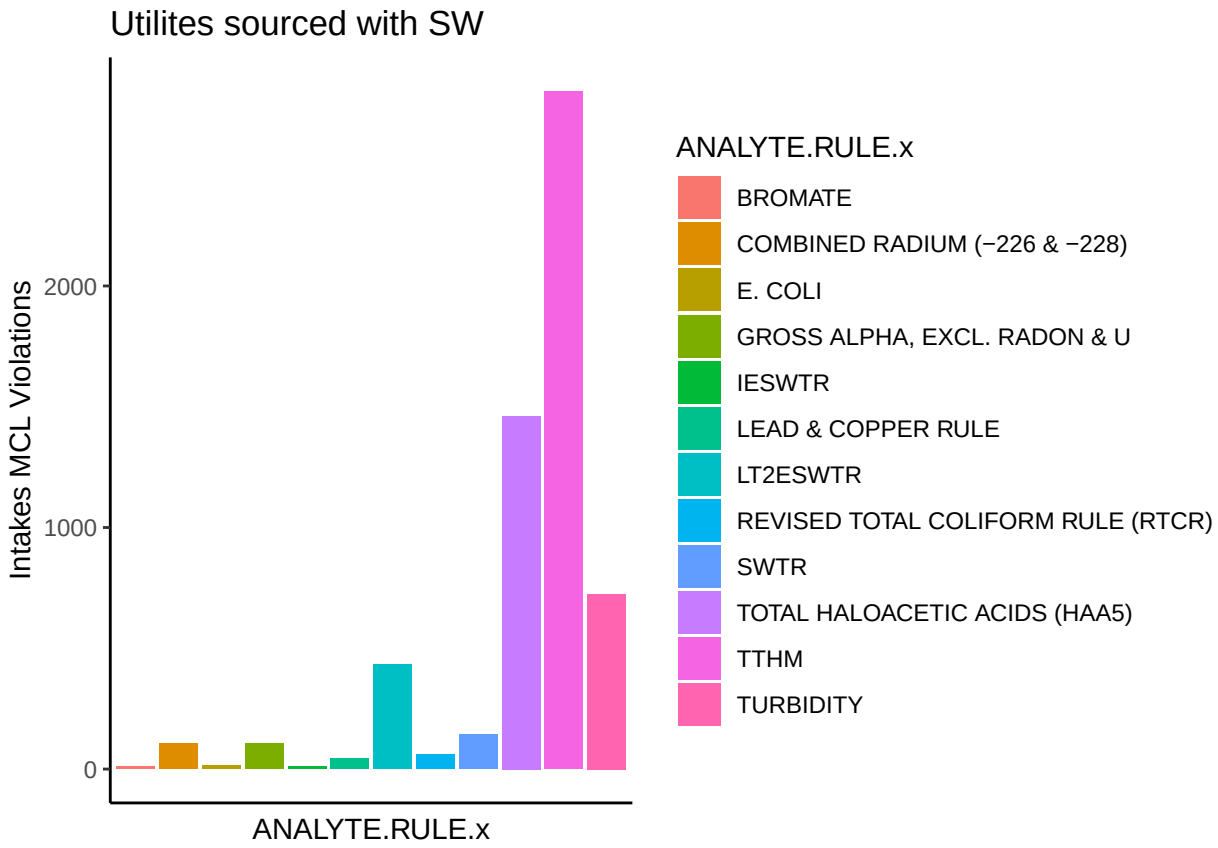
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(gridExtra)
```

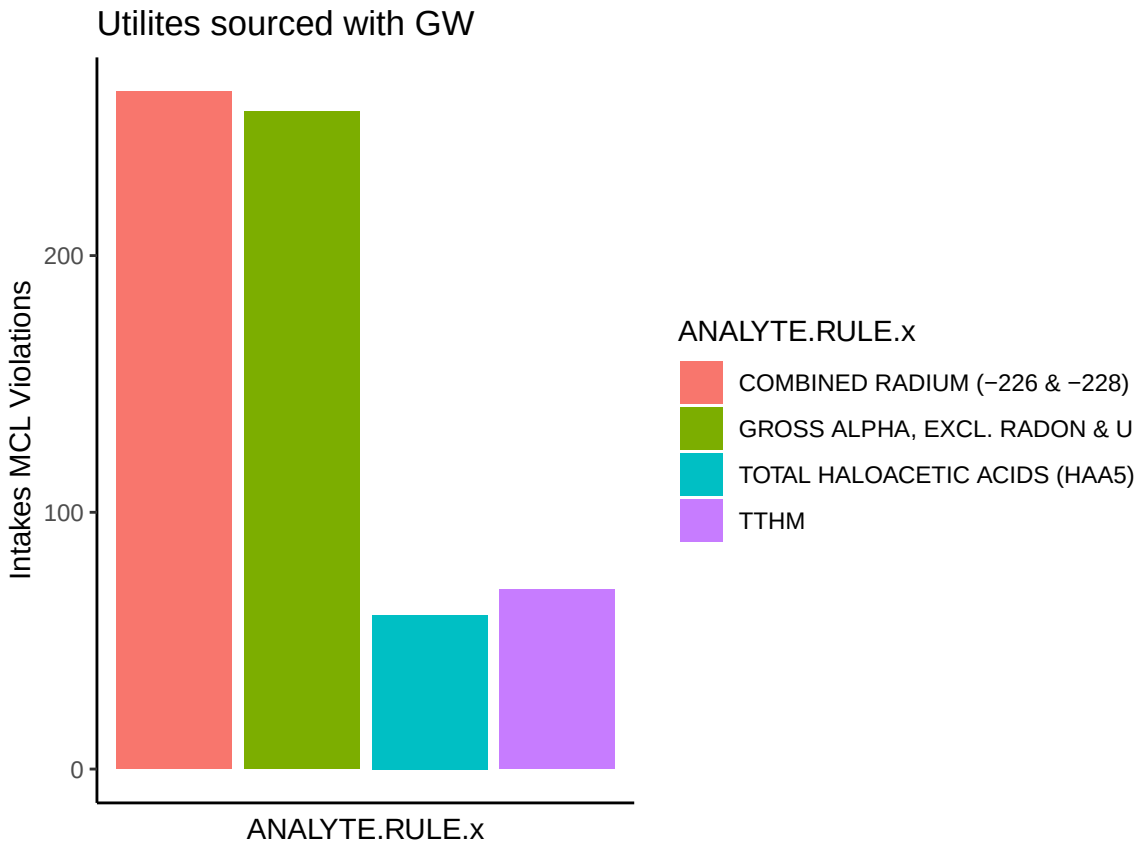
```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

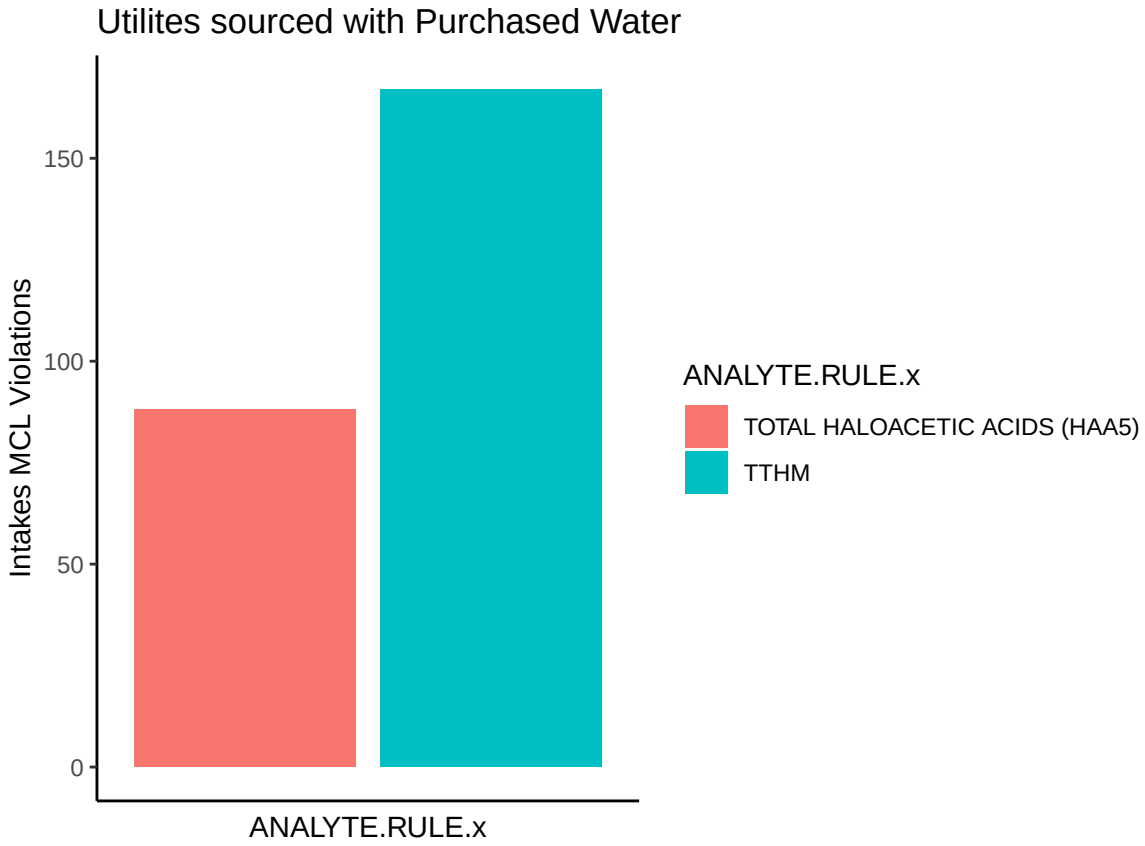
```
#intakes_tceq$Primary.Source <- factor(intakes_tceq$Primary.Source, levels = c("Ground Water", "Ground

# ggplot of violations based on New Intakes Data
# surface water
intakes_tceq %>% filter(Primary.Source=="Surface water"|
                          Primary.Source=="Surface water purchased")%>%
  group_by(`ANALYTE.RULE.x`) %>% count() %>%  filter(n>10)%>%
  ggplot(aes(x=`ANALYTE.RULE.x`, y=n, fill=`ANALYTE.RULE.x`))+
  geom_bar(stat="identity")+ labs(y="Intakes MCL Violations",
                                  title="Utilites sourced with SW")+
  theme_classic() + theme(axis.ticks.x = element_blank(),
                          axis.text.x = element_blank())
```



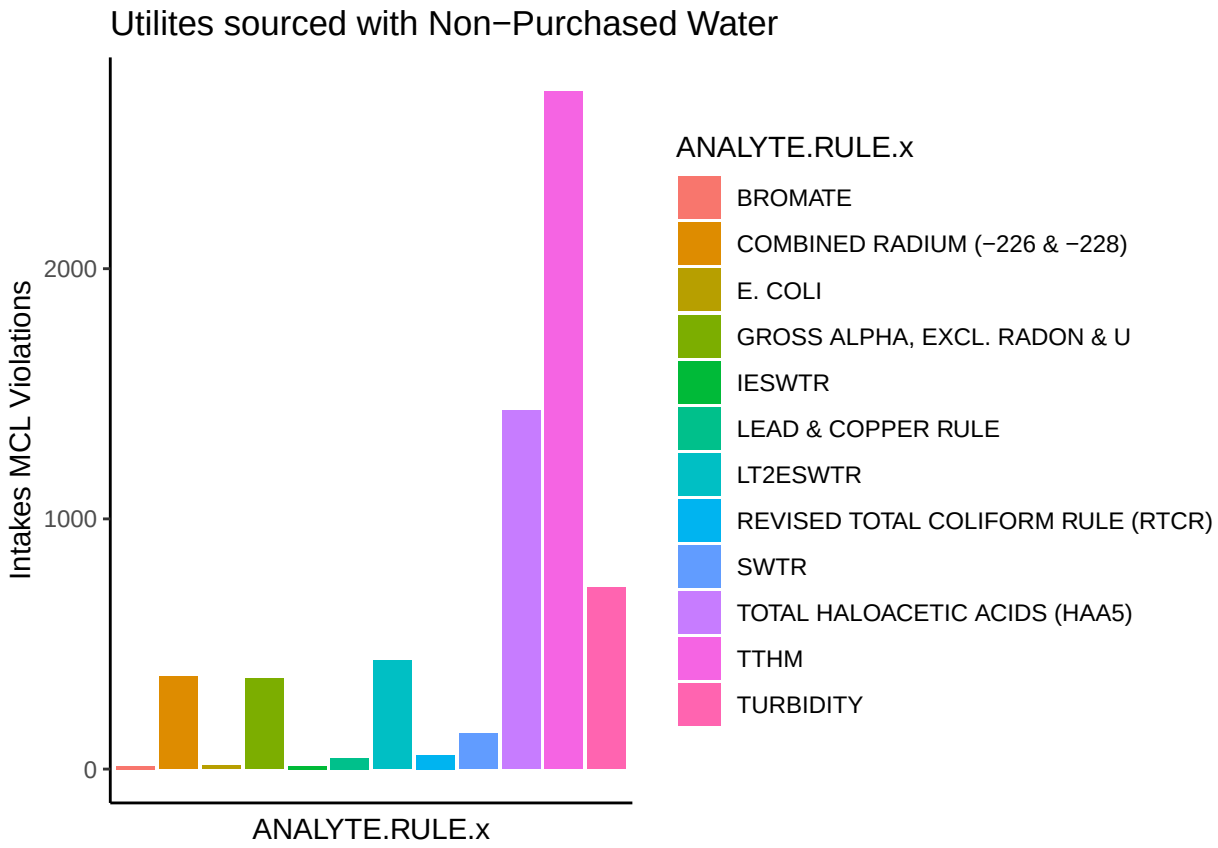

```
#ground water
intakes_tceq %>% filter(Primary.Source=="Ground water"|
                        Primary.Source=="Ground water purchased")%>%
  group_by(`ANALYTE.RULE.x`) %>% count() %>%  filter(n>10)%>%
  ggplot(aes(x=`ANALYTE.RULE.x`, y=n, fill=`ANALYTE.RULE.x`))+geom_bar(stat="identity")+
  labs(y="Intakes MCL Violations", title="Utilites sourced with GW")+theme_classic() +
  theme(axis.ticks.x = element_blank(),axis.text.x = element_blank())
```



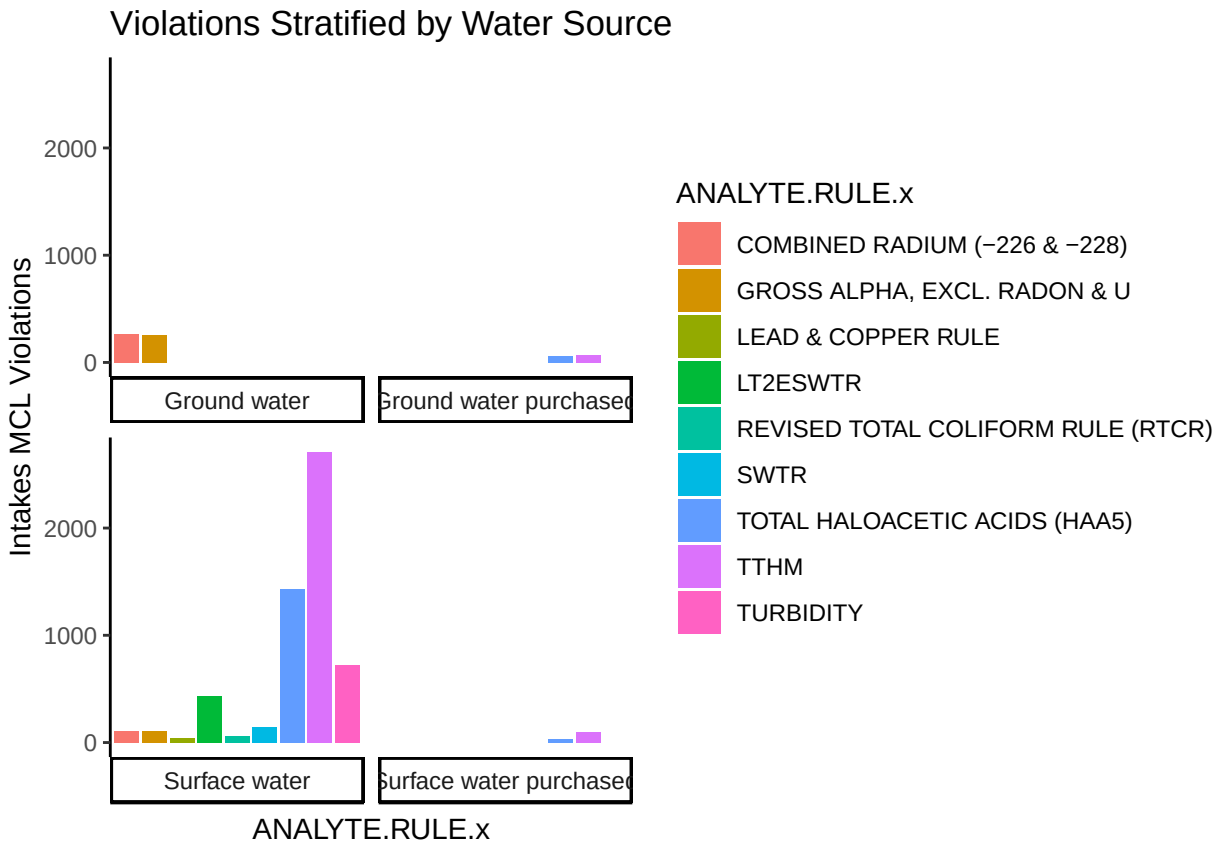

```
# purchased
intakes_tceq %>%
  filter(Primary.Source=="Ground water purchased"| Primary.Source=="Surface water purchased")%>%group_by
  filter(n>10)%>% ggplot(aes(x=`ANALYTE.RULE.x`, y=n, fill=`ANALYTE.RULE.x`))+
  geom_bar(stat="identity")+
  labs(y="Intakes MCL Violations", title="Utilites sourced with Purchased Water")+
  theme_classic() +
  theme(axis.ticks.x = element_blank(), axis.text.x = element_blank())
```




```
# non purchased (does this always mean locally sited?)
intakes_tceq %>% filter(Primary.Source=="Ground water"| Primary.Source=="Surface water")%>%
  group_by(`ANALYTE.RULE.x`) %>% count() %>%  filter(n>10)%>%
  ggplot(aes(x=`ANALYTE.RULE.x`, y=n, fill=`ANALYTE.RULE.x`))+
  geom_bar(stat="identity")+
  labs(y="Intakes MCL Violations", title="Utilites sourced with Non-Purchased Water")+
  theme_classic() + theme(axis.ticks.x = element_blank(),
                          axis.text.x = element_blank())
```




```
# all together!
intakes_tceq %>% group_by(`ANALYTE.RULE.x`, Primary.Source) %>%
  count() %>%  filter(n>25)%>% ggplot(aes(x=`ANALYTE.RULE.x`, y=n, fill=`ANALYTE.RULE.x`))+
  geom_bar(stat="identity")+
  labs(y="Intakes MCL Violations", title="Violations Stratified by Water Source")+
  theme_classic() + theme(axis.ticks.x = element_blank(),axis.text.x = element_blank()) +
  facet_wrap(. ~ Primary.Source, strip.position = "bottom")
```




From the final graph, we can see that the non-purchased surface water has a much much higher set of violations, especially ones that are very harmful and uncommmon like Lead and Coliform. Ground water relatively has a much lower amount of TTHM and Turbidity, but it has violations in Radion and Gross Alpha at around double the amount of surface water. I also found it interesteing that TTHM and HAA5 are the violations for both of the purchased categories, leading me to believe that the purchased water might need to be excluded or at least flagged in subsequent analysis because it must come from very different sources than the non-purchased water.