exploratory

Tony Ni

8/30/2020

```r
#Libraries
library(mosaic)
library(tidyverse)
library(usmap)
```

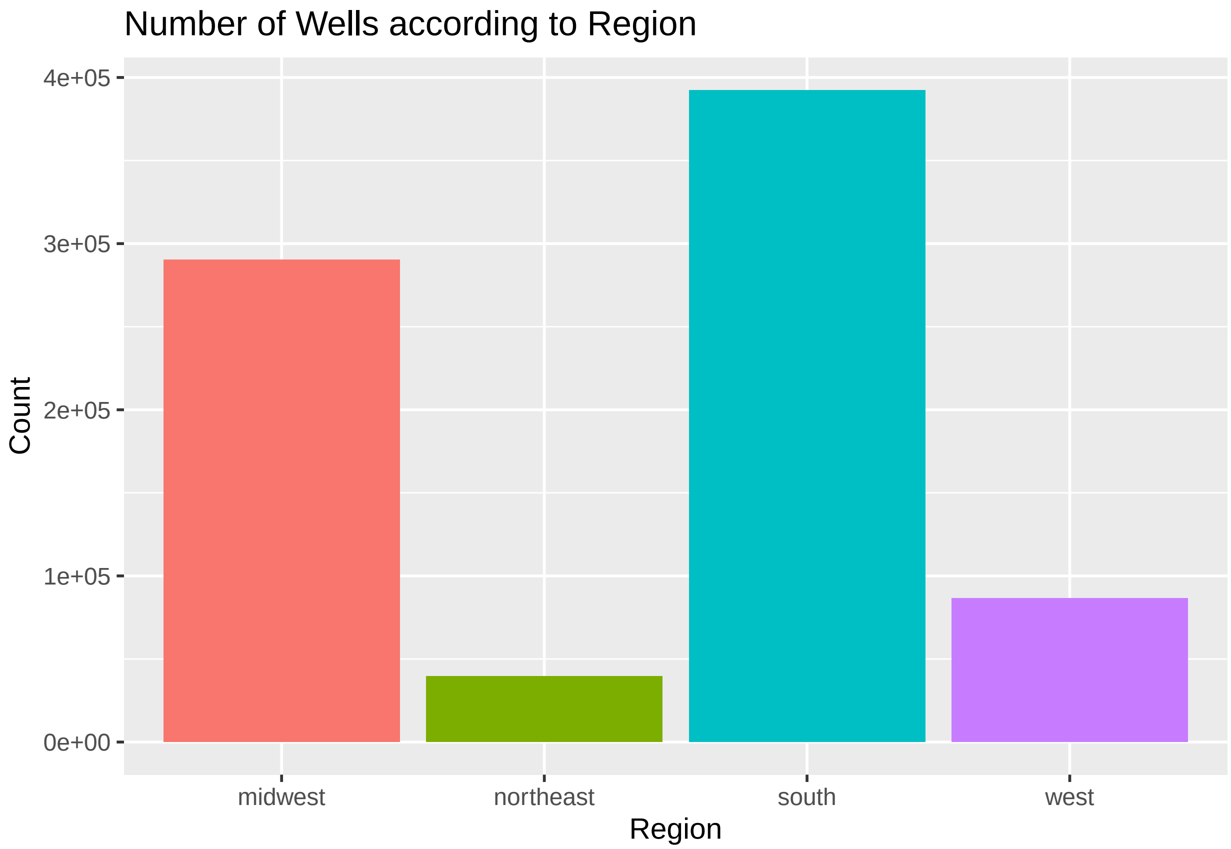
```r
#importing in full dataset
import_df <- read_csv("data/chemical_data.csv")
```

```
## Parsed with column specification:
## cols(
##   state = col_character(),
##   site = col_character(),
##   disposal.area = col_character(),
##   type = col_character(),
##   well.id = col_character(),
##   gradient = col_character(),
##   samp.date = col_character(),
##   contaminant = col_character(),
##   measurement.unit = col_character(),
##   below.detection = col_character(),
##   concentration = col_double(),
##   qualifier = col_character(),
##   link = col_character()
## )
```

```r
#breaking apart into different datasets for each region
northeast <- import_df %>%
  filter(state %in% c("ME", "NH", "VT", "NY", "PA", "NJ", "MD",
                      "MA", "DE", "RI", "CT")) %>%
  mutate(region = "northeast")

midwest <- import_df %>%
  filter(state %in% c("OH", "IN", "MI", "IL", "WI", "MN", "IA",
                      "MO", "ND", "SD", "NE", "KS"))%>%
  mutate(region = "midwest")

west <- import_df %>%
  filter(state %in% c("WA", "MT", "OR", "ID", "WY", "CA", "NV",
                      "UT", "CO", "AZ", "NM", "AK", "HI")) %>%
  mutate(region = "west")

south <- import_df %>%
  filter(state %in% c("WV", "VA", "KY", "TN", "NC", "SC", "GA",
                      "FL", "MS", "AL", "LA", "AR", "OK", "TX", "PR")) %>%
  mutate(region = "south")

#rejoin them back together for future ref. if needed
full <- list(northeast, midwest, west, south) %>%
  reduce(full_join)
```

```
## Joining, by = c("state", "site", "disposal.area", "type", "well.id", "gradient", "samp.date", "contar
## Joining, by = c("state", "site", "disposal.area", "type", "well.id", "gradient", "samp.date", "contar
## Joining, by = c("state", "site", "disposal.area", "type", "well.id", "gradient", "samp.date", "contar
```

```
ggplot(full, aes(x = region)) +
  geom_bar(aes(fill = region), show.legend = FALSE) +
  ggtitle("Number of Wells according to Region") +
  xlab("Region") +
  ylab("Count")
```



```
midwest_n <- midwest %>%
  group_by(state) %>%
  summarize(n = n()) %>%
  arrange(desc(n))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
northeast_n <- northeast %>%
  group_by(state) %>%
  summarize(n = n()) %>%
  arrange(desc(n))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
south_n <- south %>%
  group_by(state) %>%
  summarize(n = n()) %>%
  arrange(desc(n))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
west_n <- west %>%
  group_by(state) %>%
  summarize(n = n()) %>%
  arrange(desc(n))
```

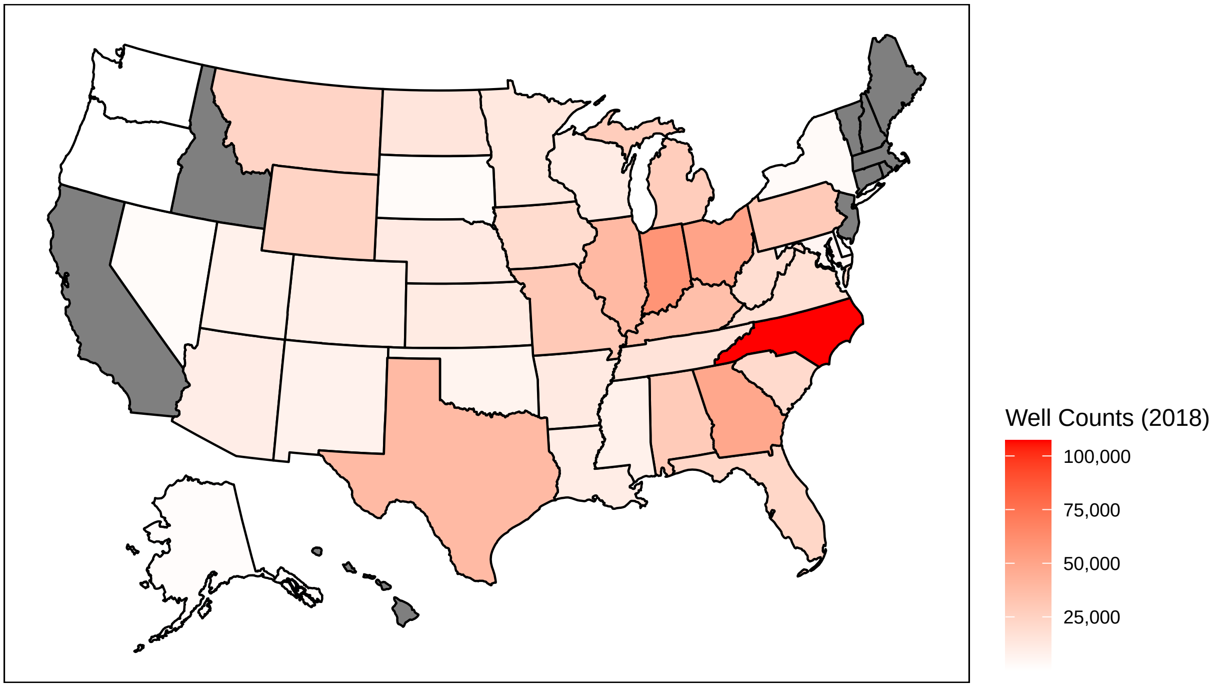
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
states_n <- rbind(midwest_n, northeast_n, south_n, west_n)
```

idea: make colored map based on how many wells are in each state

```
state_name <- state.name
state_abb <- state.abb
states_map <- map_data("state")

plot_usmap(data = states_n, values = "n", regions = "states") +
  scale_fill_continuous(low = "white", high = "red",
                        name = "Well Counts (2018)",
                        label = scales::comma) +
  theme(legend.position = "right",
        panel.background = element_rect(color = "black",
                                        fill = "white")) +
  ggtitle("Count of Groundwater Wells across U.S. States")
```

Count of Groundwater Wells across U.S. States



North Carolina has a significant number of wells amongst all states (over 100,000) compared to the next highest which is Indiana with around 58,000.

Let’s focus in on North Carolina only for now!

```
NC <- south %>%
  filter(state %in% "NC")

#count of sites
NC %>%
  group_by(site) %>%
  summarize(n = n()) %>%
  arrange(desc(n))
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
## # A tibble: 13 x 2
##    site                                    n
##    <chr>                               <int>
##  1 "L.V. Sutton Energy Complex"        15683
##  2 "Belews Creek Steam Station"        13414
##  3 "Cliffside Steam Station"           13362
##  4 "Roxboro Steam Electric Plant"      10320
##  5 "Allen Steam Station"                9938
##  6 "Buck Steam Station"                 9813
##  7 "Dan River Steam Station"            7885
##  8 "Mayo Steam Electric Plant"          7560
##  9 "H.F. Lee Energy Complex"            6120
## 10 "Marshall Steam Station"             6047
## 11 "Asheville Steam Electric Plant"     4115
## 12 "W.H. Weatherspoon Power Plant"      2520
## 13 "Brickhaven No. 2 Mine Tract \"A\""   357
```

There are 13 different “sites” in which the wells can belong to.

```
#count of disposal.area
NC %>%
  group_by(disposal.area) %>%
  summarize(n = n()) %>%
  arrange(desc(n))
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
## # A tibble: 22 x 2
##    disposal.area                                                              n
##    <chr>                                                                  <int>
##  1 Active Ash Basin                                                       15434
##  2 CCP Landfill                                                           11572
##  3 1971 and 1984 Ash Basins                                               10620
##  4 Active Ash Basin, Retired Ash Basin, Retired Ash Basin Landfill         9938
##  5 CCR Multiunit 2 (West Ash Basin, East and West FGD Settling Ponds, FGD~ 6120
##  6 Active Ash Basin and Industrial Landfill No. 1                          6047
##  7 Craig Road Landfill                                                     5156
##  8 Primary Pond (Ash Basin 2), Secondary Pond (Ash Basin 3)                5003
##  9 Additional Primary Pond (Ash Basin 1)                                   4810
## 10 CCR Multiunit 1 (East Ash Pond, Industrial Landfill)                    4200
## # ... with 12 more rows
```

Within each well, there are multiple disposal areas also (total count of 22).

```
#count of type
NC %>%
  group_by(type) %>%
  summarize(n = n()) %>%
  arrange(desc(n))
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
## # A tibble: 3 x 2
##   type      n
##   <chr> <int>
## 1 SI    57084
## 2 M     26309
## 3 L     23741
```

In the case for the NC wells, there are 57,084 SI wells, 26,306 M wells, and 23,741 L wells.

Dangerous Toxins in Coal Ash

Some of the most dangerous contaminants often found in coal ash include: arsenic, lead, mercury, cadmium, chromium, and selenium (https://www.psr.org/wp-content/uploads/2018/05/coal-ash-toxics.pdf)

Could the type of disposal unit affect the concentration (low/medium/high) of these contaminants?

```
NC_subset <- NC %>%
  filter(contaminant %in% c("Arsenic, dissolved", "Arsenic, total",
                            "Lead, total", "Mercury, total",
                            "Cadmium, dissolved", "Cadmium, total",
                            "Chromium, total", "Selenium, Dissolved",
                            "Selenium, Total"))
NC_subset2 <- NC_subset %>%
  group_by(contaminant, type, measurement.unit, below.detection) %>%
  summarize(n = n())
```

```
## `summarise()` regrouping output by 'contaminant', 'type', 'measurement.unit' (override with `.groups
```

```
NC_subset3 <- NC_subset2[-c(27), ] #removing strange sole observation


NC_subset4 <- NC_subset3 %>%
  group_by(contaminant, type, measurement.unit) %>%
  summarize(prop = n/(sum(n))) %>%
  mutate(below.detection = case_when(
    row_number() %% 2 == 1 ~ "<", #odd
    row_number() %% 2 == 0 ~ "NA")) %>% #even
  filter(below.detection %in% "<") %>%
  arrange(desc(prop))
```

```
## `summarise()` regrouping output by 'contaminant', 'type', 'measurement.unit' (override with `.groups
```

```
knitr::kable(NC_subset4)
```

contaminant	type	measurement.unit	prop	below.detection
Mercury, total	M	ug/l	0.6861314	<
Cadmium, total	M	ug/l	0.5842788	<
Mercury, total	L	ug/l	0.4650767	<
Lead, total	M	ug/l	0.4505673	<
Cadmium, total	L	ug/l	0.4327087	<
Lead, total	L	ug/l	0.3637138	<
Mercury, total	SI	ug/l	0.3383978	<
Cadmium, total	SI	ug/l	0.3173343	<
Lead, total	SI	ug/l	0.2924724	<
Arsenic, total	M	ug/l	0.1742301	<
Arsenic, total	L	ug/l	0.1533220	<
Chromium, total	M	ug/l	0.1296596	<
Arsenic, total	SI	ug/l	0.1201657	<
Chromium, total	L	ug/l	0.1081772	<
Chromium, total	SI	ug/l	0.0949586	<