Exploratory Analysis

Tony Ni, Antonella Basso, Jose Lopez

6/11/2020

Libraries

Goals:

comparing how upgradients look vs downgradients through a graph (make sure the average thing/wrangling to fix it)

pick a chemical of interest

histograms of concentrations (one for upgradient vs downgradient for a single site)

compare the compare the 2 powerplants overall (upgradient at one vs the other)

regression

Reading in Data

```
setwd("~/harvard-summer-biostats")
new_york <- read_csv("data/new_york.csv") #read in data
```

```
## 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(),
##   concentration = col_double()
## )
```

Checking unique/distinct levels

```
#check how many observations there
nrow(new_york)
```

```
## [1] 2964
```

```r
#checking distincts sites
unique(new_york$site)
```

```
## [1] "Dunkirk Generating Station" "Huntley Generating Station"
```

```r
#checking distinct gradient
unique(new_york$gradient)
```

```
## [1] "Upgradient"   "Downgradient"
```

```r
#checking distinct gradient
unique(new_york$measurement.unit)
```

```
## [1] "mg/l"  "pCi/l" "su"
```

```r
#checking disposal area
unique(new_york$disposal.area)
```

```
## [1] "Dunkirk Landfill"    "Huntley Landfill"    "South Settling Pond"
```

```r
#checking type (L = land fuel, SI = surface impoundement, M = mixed/multiunit)
#nelly asks what does type mean???
unique(new_york$type)
```

```
## [1] "L"  "SI"
```

```r
new_york %>%
  group_by(type) %>%
  summarize(n())
```

```
## # A tibble: 2 x 2
##   type  `n()`
##   <chr> <int>
## 1 L      2271
## 2 SI      693
```

```r
#ask if we might want to remove the other measurements for consistency?????
#nelly asks can we find a way to convert these easily!

new_york %>%
  group_by(measurement.unit) %>%
  summarize(n())
```

```
## # A tibble: 3 x 2
##   measurement.unit `n()`
##   <chr>            <int>
## 1 mg/l              2686
## 2 pCi/l              125
## 3 su                 153
```

```r
#checking wells
new_york %>%
  group_by(well.id) %>%
  summarize(n())
```

```
## # A tibble: 17 x 2
##    well.id  `n()`
##    <chr>    <int>
##  1 A-2        175
##  2 BR-12-DG   175
##  3 BR-13-DG   175
##  4 BR-14-UG   175
##  5 BR-20-DG   175
##  6 BR-3-DG    175
##  7 CCR-1      171
##  8 CCR-2      173
##  9 CCR-3      174
## 10 CCR-4      175
## 11 CCR-5      175
## 12 CCR-6      175
## 13 MW-11D     173
## 14 MW-12D     174
## 15 MW-13D     175
## 16 MW-14D     174
## 17 MW-7D      175
```

Wrangling

```r
new_york1 <- new_york %>%
  select(site, disposal.area, type, well.id, gradient, contaminant,
         measurement.unit, concentration) %>%
  mutate(well.id_contaminant = paste0(well.id, "_", contaminant)) %>% #for future use
  rename(c("disposal_area" = "disposal.area", "well_id" = "well.id",
           "unit" = "measurement.unit"))

#fixing 'contaminant' string by removing everything after the comma
new_york1$contaminant=gsub(", total", "", new_york1$contaminant)

#testing
avg_contaminant <- new_york1 %>%
  group_by(well_id, contaminant) %>%
  summarise_each(funs(mean)) %>%
  select(1,2,8) #selecting only numeric columns

#temporarily uniting columns for joining in next step
temp <- avg_contaminant %>%
  unite("well.id_contaminant", well_id, contaminant)

#joining orig dataframe and avg_contaminant dataframe
combined <- left_join(temp, new_york1, by = "well.id_contaminant") %>%
  distinct(well.id_contaminant, .keep_all = TRUE) %>%
  separate(well.id_contaminant, c('well_id', 'contaminant'), sep="_") %>%
  select(1:8)

combined2 <- combined %>% #collapse empty rows
  spread(contaminant, concentration.x) %>%
  group_by(well_id) %>%
  summarise_each(funs(first(.[!is.na(.)]))) %>%
  select(-c(unit))
```

Summary Statistics

```
new_york %>%
  group_by(contaminant) %>%
  summarize()
```

```
## # A tibble: 21 x 1
##    contaminant
##    <chr>
##  1 Antimony, total
##  2 Arsenic, total
##  3 Barium, total
##  4 Beryllium, total
##  5 Boron, total
##  6 Cadmium, total
##  7 Calcium, total
##  8 Chloride
##  9 Chromium, total
## 10 Cobalt, total
## # ... with 11 more rows
```

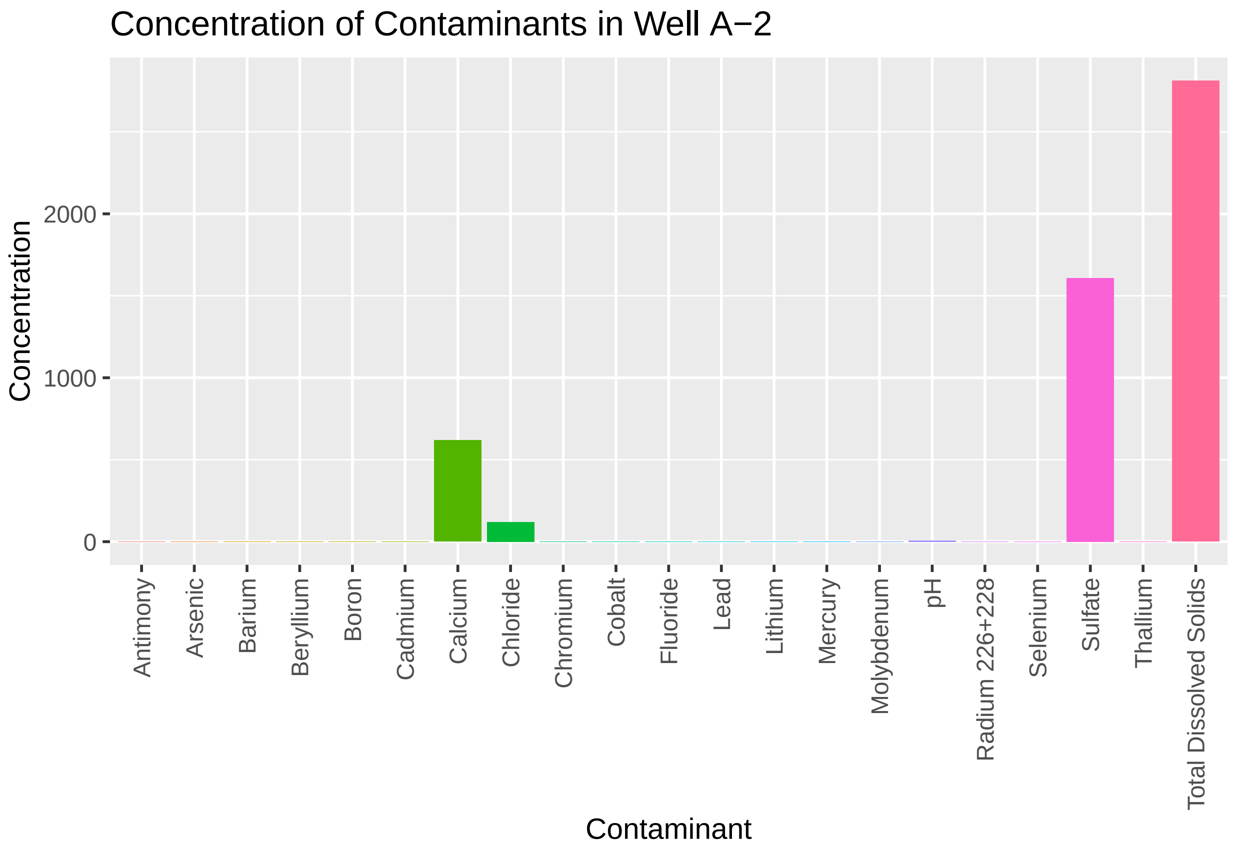
```
new_york %>%
  group_by(contaminant) %>%
  summarize(mean_concentration = mean(concentration))
```

```
## # A tibble: 21 x 2
##    contaminant      mean_concentration
##    <chr>                         <dbl>
##  1 Antimony, total             0.0451
##  2 Arsenic, total              0.00929
##  3 Barium, total               0.146
##  4 Beryllium, total            0.00441
##  5 Boron, total                1.29
##  6 Cadmium, total              0.005
##  7 Calcium, total            339.
##  8 Chloride                   59.7
##  9 Chromium, total             0.00649
## 10 Cobalt, total               0.05
## # ... with 11 more rows
```

Visualizations

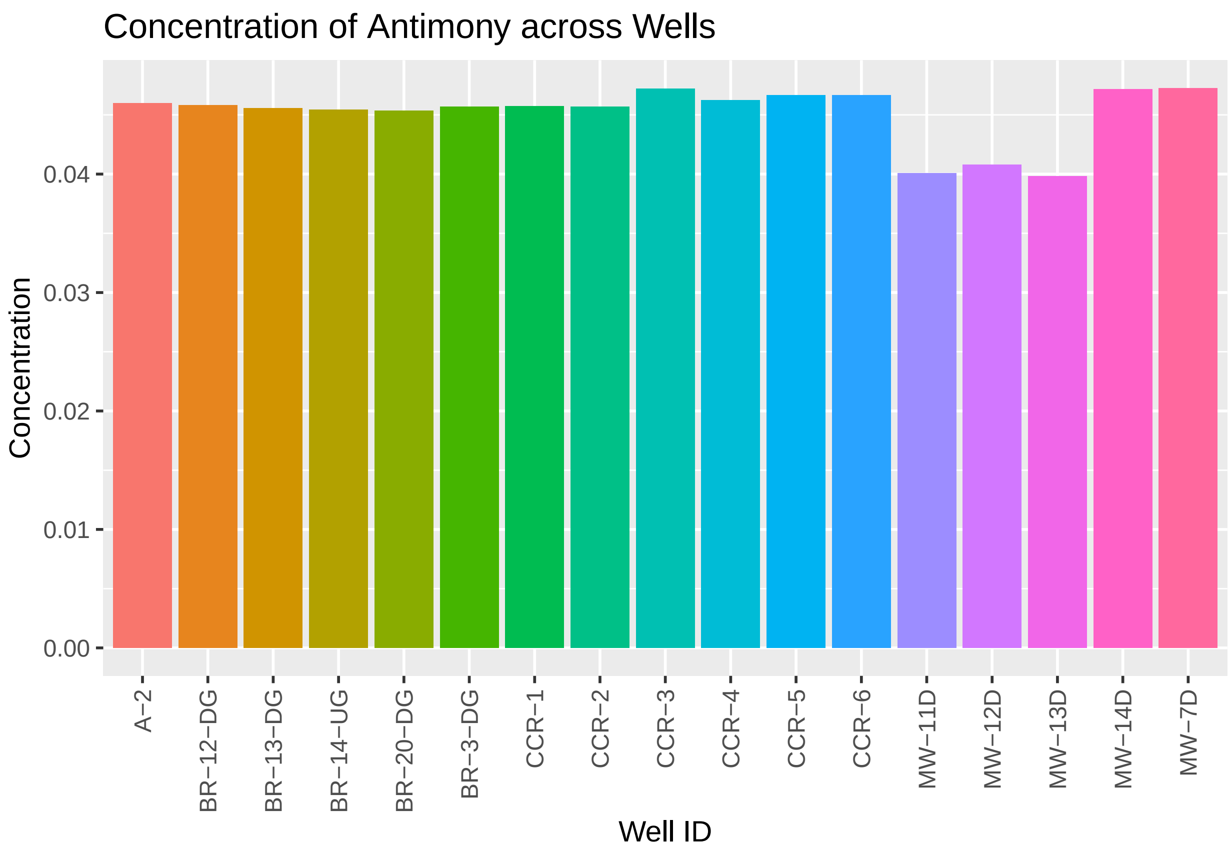
```
#looking at one specific well and the contaminant concentrations within it

ggplot(data = combined %>%
         filter(well_id == "A-2"), aes(x = contaminant, y = concentration.x,
                                        fill = contaminant)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  xlab("Contaminant") +
  ylab("Concentration") +
  ggtitle("Concentration of Contaminants in Well A-2") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



```
#looking at one contaminant for all wells (counts)

ggplot(data = combined %>%
         filter(contaminant == "Antimony"),
       aes(x = well_id, y = concentration.x, fill = well_id)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  xlab("Well ID") +
  ylab("Concentration") +
  ggtitle("Concentration of Antimony across Wells") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```


Concentration of Antimony across Wells
Concentration
0.00
0.01
0.02
0.03
0.04
A−2
BR−12−DG
BR−13−DG
BR−14−UG
BR−20−DG
BR−3−DG
CCR−1
CCR−2
CCR−3
CCR−4
CCR−5
CCR−6
MW−11D
MW−12D
MW−13D
MW−14D
MW−7D
Well ID