Indentifying Contaminated Wells

Tony Ni

7/26/2020

Libraries

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------- t

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.2     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------- tidyver
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Reading in Data

```
df <- read_csv("data/long_illinois.csv") #read in data
```

```
## Parsed with column specification:
## cols(
##   well_id = col_character(),
##   site = col_character(),
##   disposal_area = col_character(),
##   type = col_character(),
##   gradient = col_character(),
##   contaminant = col_character(),
##   concentration = col_double()
## )
```

Introduction

We are seeking to identify the contaminated wells manually (with filters).

More Wrangling

```r
#creating vector of contaminants
contaminant <- unique(df$contaminant)

#creating vector of threshold values for each contaminant
threshold <- c(6/1000, 10/1000, 2, 4/1000, 3, 5/1000, NA, NA, 100/1000,
               6/1000, NA, 15/1000, 40/1000, 2/1000, 40/1000, NA, 5,
               50/1000, 500, 2/1000, NA)

#combining names and values into a df
contam_t <- cbind(contaminant, threshold) %>%
  na.omit()
```

```r
#creating function to obtain all observations with values above threshold
#for upgradient  (kind of janky)

#it would be REALLY useful it the function could detect if values
#were repeated a lot (so that we can exclude the values where
#the repeated values/limited by device observations are excluded)

getOverThreshold <- function(df){
  datalist = list()
  for(i in 1:nrow(contam_t)){ #for each contaminant i
    df1 <- filter(df, gradient == "Upgradient")
    df2 <- filter(df1, contaminant == contam_t[i])
    data <- filter(df2, concentration > contam_t[nrow(contam_t) + i])
    datalist[[i]] <- data
  }
  toReturn <- do.call(rbind, datalist)

  return(toReturn)
}
```

```r
#using the function on data
overthreshold_df <- getOverThreshold(df) #df with all observations over threshold value
glimpse(overthreshold_df)
```

```
## Rows: 181
## Columns: 7
## $ well_id       <chr> "AP-4", "AP-5", "GO1D", "GO2D", "GO3D", "HAMW-31", "M...
## $ site          <chr> "Dallman Power Generating Station", "Dallman Power Ge...
## $ disposal_area <chr> "Dallman Ash Pond, Lakeside Ash Pond", "Dallman Ash P...
## $ type          <chr> "SI", "SI", "SI", "SI", "L", "SI", "L", "L", "L", "SI...
## $ gradient      <chr> "Upgradient", "Upgradient", "Upgradient", "Upgradient...
## $ contaminant   <chr> "Antimony", "Antimony", "Antimony", "Antimony", "Anti...
## $ concentration <dbl> 21.8812500, 21.8812500, 0.5005000, 0.5005000, 1.00000...
```