ESM 262Assignment 2

Rucha Thakar

May 11, 2017

PART 1

Setting working directory

```
setwd("C:/boxsync/rthakar/Courses/Spring2017/ESM262/EnvInformatics/Assignment2")

library (tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library (dplyr)
library (gdata)
```

```
## gdata: Unable to locate valid perl interpreter
## gdata:
## gdata: read.xls() will be unable to read Excel XLS and XLSX files
## gdata: unless the 'perl=' argument is used to specify the location
## gdata: of a valid perl intrpreter.
## gdata:
## gdata: (To avoid display of this message in the future, please
## gdata: ensure perl is installed and available on the executable
## gdata: search path.)
```

```
## gdata: Unable to load perl libaries needed by read.xls()
## gdata: to support 'XLX' (Excel 97-2004) files.
```

```
##
```

```
## gdata: Unable to load perl libaries needed by read.xls()
## gdata: to support 'XLSX' (Excel 2007+) files.
```

```
##
```

```
## gdata: Run the function 'installXLSXsupport()'
## gdata: to automatically download and install the perl
## gdata: libaries needed to support Excel XLS and XLSX formats.
```

```
##
## Attaching package: 'gdata'
```

```
## The following objects are masked from 'package:dplyr':
##
```

```
##     combine, first, last
## The following object is masked from 'package:purrr':
##
##     keep
## The following object is masked from 'package:stats':
##
##     nobs
## The following object is masked from 'package:utils':
##
##     object.size
## The following object is masked from 'package:base':
##
##     startsWith
```

```
library (lubridate)
```

```
##
## Attaching package: 'lubridate'
## The following object is masked from 'package:base':
##
##     date
```

```
library (pander)
```

Reading in the data as-is

```
gaz_raw <- read.delim("C:/boxsync/rthakar/Courses/Spring2017/ESM262/EnvInformatics/Assignment2/CA_Featu
```

Selecting required columns and converting data frame to tibble

```
gaz <- gaz_raw %>% select (ï..FEATURE_ID,FEATURE_NAME,FEATURE_CLASS,STATE_ALPHA, COUNTY_NAME, PRIM_LAT_


colnames(gaz) <- c("featureID","feature_name","feature_class", "state_alpha", "county_name", "primary_la

gaz <- as.tibble(gaz)
```

Change class to appropriate types

```
gaz$primary_latitude <- as.numeric(gaz$primary_latitude)
gaz$primary_longitude <- as.numeric(gaz$primary_longitude)
gaz$source_latitude <- as.numeric(gaz$source_latitude)
gaz$source_longitude <- as.numeric(gaz$source_longitude)
gaz$elevation <- as.numeric(gaz$elevation)
gaz$date_created <- mdy (gaz$date_created)
gaz$date_edited <- mdy (gaz$date_edited)
```

According to https://geonames.usgs.gov/domestic/states_fileformat.htm, Records showing "Unknown" and zeros for the latitude and longitude DMS and decimal fields, respectively, indicate that the coordinates of the feature are unknown. They are recorded in the database as zeros to satisfy the format requirements of a numerical data type. They are not errors and do not reference the actual geographic coordinates at 0 latitude, 0 longitude.

```
gaz <- gaz %>% filter (primary_longitude != 0 | primary_latitude != 0)
```

```
#gaz <- unknownToNA(gaz, unknown = c("NA",""))
```

In addition, drop NA values

```
gaz <- gaz %>%
  drop_na(primary_latitude) %>%
  drop_na(primary_longitude)
```

Removed all features that do not belong to the state of California, USA

```
California <- gaz %>% filter(state_alpha == "CA")
```

Writing final file to the disk as final_file_part1.csv using "|" as separator

```
write.table(California, file="California_Data.csv", sep = "|")
```

PART 2

most-frequently-occuring feature name in California

```
(California %>% count (feature_name)%>% filter (n == max(n)))
```

```
## # A tibble: 1 × 2
##       feature_name     n
##              <chr> <int>
## 1 Church of Christ   228
```

least-frequently-occuring feature class in California

```
(California %>% count (feature_class)%>% filter (n == min(n)))
```

```
## # A tibble: 2 × 2
##   feature_class     n
##           <chr> <int>
## 1       Isthmus     1
## 2           Sea     1
```

approximate center point of each county

Removing empty county names

```
California <- California %>% filter (county_name != "")

County_Group <- group_by(California, county_name)
County_Mean <- summarize (County_Group, mean (primary_longitude), mean (primary_latitude))
```

fractions of the total number of features in each county that are natural vs. man-made

63 feature classes are categorized between natural and man-made according to the classification described in https://geonames.usgs.gov/apex/f?p=gnispq:8:0:::::

```
manmade <- c("Airport", "Bridge", "Building", "Canal", "Cemetery", "Census", "Church", "Civil", "Crossi
California$feature <- c(0)
California$n <- seq (1:nrow(California))

California$feature <- ifelse(California$feature_class %in% manmade, "Manmade", "Natural")
```