Assignment2

Lina Barbosa

5/10/2017

1. Loading Data

Read dataset as tibble

```
gaz_raw <- read_delim("./CA_Features_20170401.zip", delim ="|", col_types = cols()) %>%
  as_tibble()
```

2. Questions 1-4

I created gaz tibble by selecting the listed variables. I changed the names of the variables to lower case, and formatted properly the Date_created and Date_edited varibles as.Date. Also, I changed the Unknown values to NAs and got rid of of the 0 values in the variables longitude and latitude.

```
gaz <- gaz_raw %>%
  select(FEATURE_ID = 1,
         starts_with("FEATURE"),
         STATE_ALPHA,
         COUNTY_NAME,
         ends_with("DEC"),
         ELEV_IN_M,
         MAP_NAME,
         starts_with("DATE")) %>%
  magrittr::set_colnames(value = tolower(colnames(.))) %>%
  mutate(date_created = as.Date(date_created, format = "%m/%d/%y"),
         date_edited = as.Date(date_edited, format = "%m/%d/%y"),
         map_name = ifelse(map_name == "Unknown", NA, map_name)) %>%
  filter(prim_lat_dec != 0, prim_long_dec != 0, state_alpha == "CA")

# This is not part of the assignment (just me having fun)
# check_col_vals <- function(data){
#   sapply(data, function(x){sort(unique(x))})
# }
```

Question 5

Exporting gaz tibble as csv

```
write_delim(gaz, path = "./gaz.csv", delim = "|")
```

Analyze

1. What is the most-frequently-occuring feature name?

```
counts_name <- gaz %>%
  group_by(feature_name) %>%
  summarise(count = n()) %>%
  arrange(count)

# I'm using this -- counts_name$feature_name[counts_name$count == max(counts_name$count)]
# -as an inline code below
```

The most-frequently-occuring feature name Church of Christ.

What is the least-frequently-occuring feature class?

```
counts_class <- gaz %>%
  group_by(feature_class) %>%
  summarise(count = n()) %>%
  arrange(count)
```

The least-frequently-occuring feature class is Isthmus, Sea

3. What is the approximate center point of each county?

```
center_point <- gaz %>%
  group_by(county_name) %>%
  summarise(latitude = mean(prim_lat_dec, na.rm = TRUE),
            longitude = mean(prim_long_dec, na.rm = TRUE),
            longmax = max(prim_long_dec, na.rm = TRUE),
            longmin = min(prim_long_dec, na.rm = TRUE),
            latmax = max(prim_lat_dec, na.rm = TRUE),
            latmin = min(prim_lat_dec, na.rm = TRUE),
            lat = (latmax + latmin)/2,
            long = (longmax + longmin)/2) %>%
  select(-c(longmax, latmax, longmin, latmin))
```

```
ggplot(gaz, aes(x = prim_long_dec, y = prim_lat_dec, color = county_name)) +
  geom_point(pch = ".") +
  geom_point(data = center_point, aes(x = longitude, y = latitude), color = "black") +
  geom_point(data = center_point, aes(x = long, y = lat), color = "darkblue") +
  coord_quickmap() +
  theme_bw() +
  theme(legend.position = "none")
```

```
# FC <- gaz %>%
#   select(feature_class) %>%
#   unique()

# write_csv(FC, "FC.csv")
```

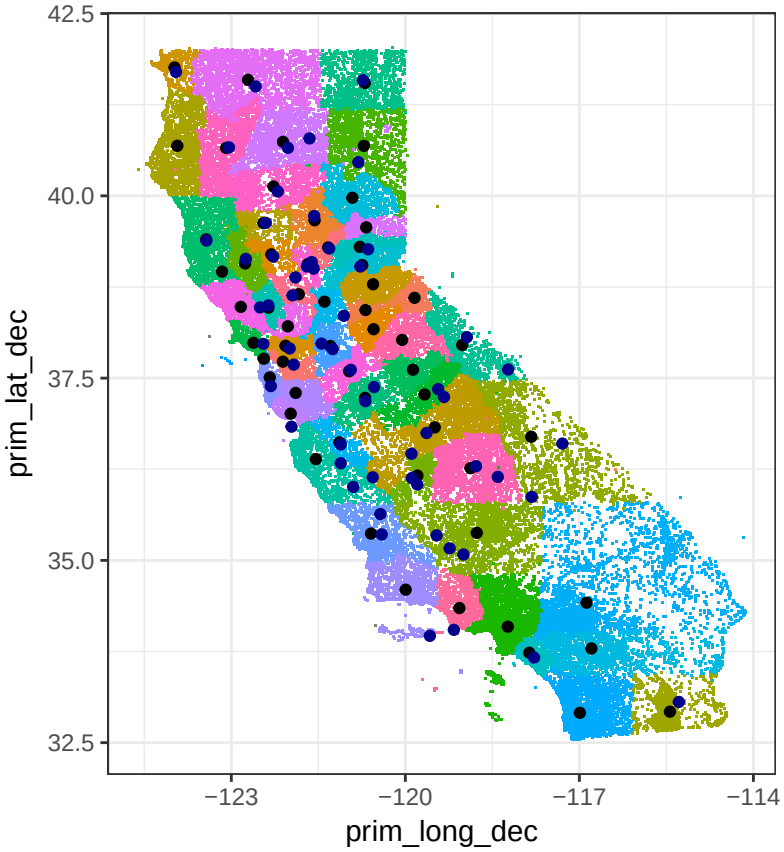

Figure 1: Black points represent the center point calculated using the mean of the latitude and longitud points for each county. Blue points also represent the center point of the bounding box

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

```
FC2 <- read_csv("FC.csv", col_types = cols())

class_made <- gaz %>%
  left_join(FC2) %>%
  group_by(county_name, made_by) %>%
  summarise(count = n()) %>%
  spread(made_by, count) %>%
  mutate(propman = man/(man + nature),
         propnature = 1-propman)
```

```
## Joining, by = "feature_class"
```

```
write_csv(class_made, "Question4.csv")
```

The file "Question4.csv" contains the proportion of the total number of features in each county that are natural and man-made