Assignment 2: Data Ingest

Monique Schafer

5/9/2017

Part 1: Importing and Tidying Data

Loading neccessary packages and importing data

Note: Columns were assigned appropriate type in this step

```
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
```

```
gaz_raw <- read_delim("CA_Features_20170401.txt", delim = "|", col_types = cols(
  FEATURE_ID = col_character(),
  DATE_CREATED = col_date(format = "%m/%d/%Y"),
  DATE_EDITED = col_date(format = "%m/%d/%Y")
))
```

Selecting columns

```
gaz <- select(gaz_raw, FEATURE_ID,FEATURE_NAME, FEATURE_CLASS, STATE_ALPHA, COUNTY_NAME, PRIM_LAT_DEC, 

gaz[, 6:7][gaz[, 6:7] == 0] <- NA

####Removing Unknown Rows and Selecting For California Features Only
gaz <- gaz%>% filter(STATE_ALPHA == "CA") %>%
  filter(!is.na(PRIM_LAT_DEC)) %>%
  filter(!is.na(PRIM_LONG_DEC))

write_csv(gaz,"gaz.csv")
```

Part 2: Analyzing the Gazateer Data

What is the most-frequently-occurring feature name? What is the least-frequently-occuring feature class?

```
feature_class <- count(gaz, vars = FEATURE_CLASS)
print(feature_class)
```

```
## # A tibble: 63 × 2
##       vars     n
```

```
##      <chr> <int>
## 1  Airport  1102
## 2     Arch    79
## 3     Area   284
## 4   Arroyo     2
## 5      Bar   274
## 6    Basin   509
## 7      Bay   419
## 8    Beach   275
## 9    Bench    32
## 10    Bend   108
## # ... with 53 more rows
```

```r
write.csv(feature_class, "features.csv")
```

The most-frequently-occurring feature class is “locale”, the least-frequently-occurring feature class is “isthmus” and “sea”.

What is the approximate center point of each county?

```r
gaz_countygeo <- gaz %>%
  group_by(COUNTY_NAME) %>%
  summarise(county_minlat = min(PRIM_LAT_DEC, na.rm = TRUE),
  county_maxlat = max(PRIM_LAT_DEC, na.rm = TRUE), county_minlong = max(PRIM_LONG_DEC, na.rm = TRUE), c
  )

gaz_countymid <- transmute(gaz_countygeo,
                        COUNTY_NAME = COUNTY_NAME,
                        MID_LAT = (county_maxlat + county_minlat)/2,
                        MID_LONG = (county_maxlong + county_minlong)/2
)

print(gaz_countymid[,1:3], caption = "Latitude and Longitude Midpoints")
```

```
## # A tibble: 59 × 3
##     COUNTY_NAME  MID_LAT  MID_LONG
##           <chr>    <dbl>     <dbl>
## 1       Alameda 37.68525 -121.9243
## 2        Alpine 37.61799 -118.2290
## 3        Amador 38.35542 -121.0613
## 4         Butte 39.72335 -121.5716
## 5     Calaveras 36.46287 -119.8929
## 6        Colusa 39.16739 -122.2780
## 7  Contra Costa 37.90659 -121.9944
## 8     Del Norte 41.69998 -123.9550
## 9     El Dorado 37.97298 -121.4447
## 10       Fresno 36.74745 -119.6338
## # ... with 49 more rows
```

The midpoints for each county are shown in the table above.

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

```r
features_natman <- read_csv("features_natman.csv")
```

```
## Parsed with column specification:
## cols(
##   FEATURE_CLASS = col_character(),
##   Nat_Man = col_character()
## )
```

```r
all_features_natman <- left_join(features_natman, gaz, by = "FEATURE_CLASS")

count_natural <- count(all_features_natman, vars = Nat_Man, by = COUNTY_NAME)
count_natural_tidy <- spread(count_natural, key = vars, value = n)
prop_table_county <- mutate(count_natural_tidy,
                            prop_natural = natural / (manmade+natural),
                            prop_manmade = manmade / (natural+manmade)
)
print(prop_table_county)
```

```
## # A tibble: 59 × 5
##              by manmade natural prop_natural prop_manmade
##           <chr>   <int>   <int>        <dbl>        <dbl>
## 1       Alameda    2436     638    0.2075472    0.7924528
## 2        Alpine     205     356    0.6345811    0.3654189
## 3        Amador     419     184    0.3051410    0.6948590
## 4         Butte     837     517    0.3818316    0.6181684
## 5     Calaveras     698     367    0.3446009    0.6553991
## 6        Colusa     243     280    0.5353728    0.4646272
## 7  Contra Costa    1372     517    0.2736898    0.7263102
## 8     Del Norte     277     369    0.5712074    0.4287926
## 9     El Dorado    1082     878    0.4479592    0.5520408
## 10       Fresno    2250    1747    0.4370778    0.5629222
## # ... with 49 more rows
```