Assign2

Eunhee Lee

May 11, 2017

### 1-1.Read the gazetteer data as-is into a gaz\_raw tibble a

ca <- read.delim("../wrangle/data/CA\_Features\_20170401.txt", header=TRUE, sep="|",na="")

### 1-2.Copy only the following columns into a gaz tibble

colnames(ca)[1] <- "FEATURE\_ID" # change the first column name  
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

ca <- select(ca,FEATURE\_ID,FEATURE\_NAME, FEATURE\_CLASS, STATE\_ALPHA, COUNTY\_NAME, PRIM\_LAT\_DEC, PRIM\_LONG\_DEC, SOURCE\_LAT\_DEC, SOURCE\_LONG\_DEC, ELEV\_IN\_M, MAP\_NAME, DATE\_CREATED, DATE\_EDITED)

### 1-3.Convert the gaz columns to the appropriate type

#### (Convert any placeholders for unknown data to NA - this process has been done in the first step.)

ca <- as\_tibble(ca)

### 1-4.Delete from gaz rows where:

#### a) the primary latitude or longitude are unknown, b) the feature is not in California

ca <-ca[!(is.na(ca$PRIM\_LAT\_DEC)) | !(is.na(ca$PRIM\_LONG\_DEC)),]  
ca <- filter(ca, STATE\_ALPHA == "CA")  
write.table(ca,"CAdata.csv", sep = "|")

### 2-1 What is the most-frequently-occuring feature name?

Analysis1 <- group\_by(ca,FEATURE\_CLASS)  
AnalyFre <- summarise(Analysis1, count=n())  
leastclass=which.min(AnalyFre$count)  
AnalyFre[leastclass,c("FEATURE\_CLASS","count")]

## # A tibble: 1 × 2  
## FEATURE\_CLASS count  
## <fctr> <int>  
## 1 Isthmus 1

### 2-2.What is the least-frequently-occuring feature class?

Analysis2 <- group\_by(ca,FEATURE\_NAME)  
AnalyFre2 <- summarise(Analysis2, count=n())  
mostname=which.max(AnalyFre2$count)  
AnalyFre2[mostname,c("FEATURE\_NAME","count")]

## # A tibble: 1 × 2  
## FEATURE\_NAME count  
## <fctr> <int>  
## 1 Church of Christ 228

### 2-3.What is the approximate center point of each county?

Analysis3a <- group\_by(ca,COUNTY\_NAME)%>%   
 summarise(Centercounty = mean(PRIM\_LONG\_DEC, na.rm=TRUE))   
colnames(Analysis3a)[2] <- "MEAN\_LONG\_DEC"  
Analysis3b <- group\_by(ca,COUNTY\_NAME)%>%   
 summarise(Centercounty = mean(PRIM\_LAT\_DEC, na.rm=TRUE))   
colnames(Analysis3b)[2] <- "MEAN\_LAT\_DEC"  
Analysis3 <- left\_join(Analysis3a, Analysis3b)

## Joining, by = "COUNTY\_NAME"

Analysis3

## # A tibble: 59 × 3  
## COUNTY\_NAME MEAN\_LONG\_DEC MEAN\_LAT\_DEC  
## <fctr> <dbl> <dbl>  
## 1 Alameda -122.1109 37.72641  
## 2 Alpine -119.4154 38.46444  
## 3 Amador -120.0885 38.24373  
## 4 Butte -120.1446 39.20248  
## 5 Calaveras -120.0967 38.02835  
## 6 Colusa -121.8435 39.04860  
## 7 Contra Costa -122.0017 37.92366  
## 8 Del Norte -123.2160 41.50408  
## 9 El Dorado -119.0383 38.29924  
## 10 Fresno -118.8680 36.63142  
## # ... with 49 more rows

### 2-4.What are the fractions of the total number of features in each county?

Featureclass <-read.table("Featureclass.csv", header=TRUE, sep=",")  
colnames(Featureclass) = c("FEATURE\_CLASS", "CHARACTERISTIC")  
Featureclass <- as\_tibble(Featureclass)  
caF <- merge(ca, Featureclass, by="FEATURE\_CLASS")  
Analysis4a <- group\_by(caF, COUNTY\_NAME)  
Analysis4 <- summarise(Analysis4a,Naturalfraction = nrow(subset(Analysis4a, CHARACTERISTIC == "Natural"))/nrow(Analysis4a)\*100,  
 Manmadefraction = nrow(subset(Analysis4a, CHARACTERISTIC == "Manmade"))/nrow(Analysis4a)\*100)   
  
Analysis4

## # A tibble: 59 × 3  
## COUNTY\_NAME Naturalfraction Manmadefraction  
## <fctr> <dbl> <dbl>  
## 1 Alameda 8.075207 91.92479  
## 2 Alpine 61.058702 38.94130  
## 3 Amador 27.538063 72.46194  
## 4 Butte 34.803616 65.19638  
## 5 Calaveras 31.994848 68.00515  
## 6 Colusa 51.402743 48.59726  
## 7 Contra Costa 16.054436 83.94556  
## 8 Del Norte 54.236120 45.76388  
## 9 El Dorado 41.535142 58.46486  
## 10 Fresno 39.943726 60.05627  
## # ... with 49 more rows