1.1 - Biodynamic Data Exploration

Julian Barg

January 20, 2019

1. Load data

```
library(tidyverse)
library(readr)

biodynamic <- read_csv("downloads/biodynamic_history.csv")
glimpse(biodynamic)
```

```
## Observations: 1,252
## Variables: 15
## $ acreage            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ address            <chr> "555 College Avenue\nAngwin, CA 94508", "Vi...
## $ business           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ crops              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ date               <date> 2014-10-09, 2014-10-09, 2014-10-09, 2014-1...
## $ email              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ link               <chr> "https://web.archive.org/web/20141009050729...
## $ name               <chr> "ADAMVS", "Alquimia Agricola", "Ambassador ...
## $ phone              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ processed_products <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ profile            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ state              <chr> "CA", NA, "IL", "CA", "CA", "CA", "NY", "CA...
## $ vineyard           <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ website            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ winery             <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
```

2. First cleaning

2.1 Keep address, date, name, type & acreage

```
biodynamic <- select(biodynamic, name, date, state, address, vineyard, winery, acreage)

glimpse(biodynamic)
```

```
## Observations: 1,252
## Variables: 7
## $ name     <chr> "ADAMVS", "Alquimia Agricola", "Ambassador Organics",...
## $ date     <date> 2014-10-09, 2014-10-09, 2014-10-09, 2014-10-09, 2014...
## $ state    <chr> "CA", NA, "IL", "CA", "CA", "CA", "NY", "CA", "NV", "...
## $ address  <chr> "555 College Avenue\nAngwin, CA 94508", "Villa Nueva\...
## $ vineyard <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ winery   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ acreage  <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
```

2.2 Clean acreage

```
parse_numbers <- "\\d*,?\\d+\\.?\\d*"
biodynamic$acreage <- str_extract(biodynamic$acreage, parse_numbers)
biodynamic$acreage <- gsub(",", "", biodynamic$acreage)
biodynamic$acreage <- as.numeric(biodynamic$acreage)
head(biodynamic$acreage[!is.na(biodynamic$acreage)])
```

```
## [1]  26.5 162.0  32.0  82.0  42.0  26.5
```

Further visual inspection of the table confirmed that all entries have been correctly identified.

2.3 Keep only entries for CA

Sort table for visual inspection

```
biodynamic <- biodynamic[order(biodynamic$name, biodynamic$date), ]
```

Visual inspection of the table confirmed that state and address match for all entries, and that no organization has changed address.

```
biodynamic <- biodynamic[which(biodynamic$state == "CA"), ]
```

3. Fill gaps

We take advantage of the fact that TRUE corresponds to 1 to apply vineyard/winery to all observations for one organization.

```
biodynamic <- biodynamic %>%
  group_by(name) %>%
  mutate(acreage = max(acreage, na.rm = TRUE)) %>%
  mutate(winery = max(winery, na.rm = TRUE)) %>%
  mutate(vineyard = max(vineyard, na.rm = TRUE))
```

We cause values of -Inf which we will replace with NA.

```
library(naniar)

biodynamic <- biodynamic %>%
  replace_with_na(list(vineyard = -Inf, winery = -Inf, acreage = -Inf))
biodynamic$vineyard <- as.logical(biodynamic$vineyard)
biodynamic$winery <- as.logical(biodynamic$winery)
```

4. Keep only vineyards/wineries

```
biodynamic <- biodynamic[which(biodynamic$vineyard == TRUE | biodynamic$winery == TRUE), ]
head(biodynamic)
```

```
## # A tibble: 6 x 7
## # Groups:   name [1]
##   name   date       state address                  vineyard winery acreage
##   <chr>  <date>     <chr> <chr>                    <lgl>    <lgl>    <dbl>
## 1 ADAMVS 2014-10-09 CA    "555 College Avenue\nAn~ TRUE     NA          NA
```

```
## 2 ADAMVS 2014-10-09 CA    "555 College Avenue\nAn~ TRUE     NA          NA
## 3 ADAMVS 2015-01-05 CA    "555 College Avenue\nAn~ TRUE     NA          NA
## 4 ADAMVS 2015-05-12 CA    "555 College Avenue\nAn~ TRUE     NA          NA
## 5 ADAMVS 2015-10-21 CA    "501 White Cottage Road~ TRUE     NA          NA
## 6 ADAMVS 2015-10-23 CA    "501 White Cottage Road~ TRUE     NA          NA
```

5. Downsample to year

First, we resample to year. Visual inspection shows no conflicts (except spelling) within companies for any of the entries. We keep the last entry for each company-year observation.

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
biodynamic <- biodynamic %>%
  mutate(year = year(date)) %>%
  group_by(name, year) %>%
  mutate(last = last(date)) %>%
  filter(date == last) %>%
  ungroup() %>%
  select(-c(year, last))

head(biodynamic, 5)
```

```
## # A tibble: 5 x 7
##   name   date       state address                  vineyard winery acreage
##   <chr>  <date>     <chr> <chr>                    <lgl>    <lgl>    <dbl>
## 1 ADAMVS 2014-10-09 CA    "555 College Avenue\nAn~ TRUE     NA          NA
## 2 ADAMVS 2014-10-09 CA    "555 College Avenue\nAn~ TRUE     NA          NA
## 3 ADAMVS 2015-11-21 CA    "501 White Cottage Road~ TRUE     NA          NA
## 4 ADAMVS 2016-12-22 CA    "501 White Cottage Road~ TRUE     NA          NA
## 5 ADAMVS 2017-12-14 CA    "501 White Cottage Road~ TRUE     NA          NA
```

There are some duplicate entries (same day). The entries differ by nothing,or may have additional entry for "United States" in the address. We therefore apply unique to the dataframe without the address column. We also truncate day & month, and rename date column to year.

```
biodynamic <- biodynamic[!duplicated(biodynamic[, -5]), ]
biodynamic$date <- year(biodynamic$date)
biodynamic <- rename(biodynamic, year=date)
head(biodynamic)
```

```
## # A tibble: 6 x 7
##   name    year state address                  vineyard winery acreage
##   <chr>  <dbl> <chr> <chr>                    <lgl>    <lgl>    <dbl>
## 1 ADAMVS  2014 CA    "555 College Avenue\nAngw~ TRUE     NA          NA
## 2 ADAMVS  2015 CA    "501 White Cottage Road N~ TRUE     NA          NA
## 3 ADAMVS  2016 CA    "501 White Cottage Road N~ TRUE     NA          NA
## 4 ADAMVS  2017 CA    "501 White Cottage Road N~ TRUE     NA          NA
## 5 ADAMVS  2019 CA    "501 White Cottage Road N~ TRUE     NA          NA
```

```
## 6 AmByth E~  2014 CA    "510 Sequoia Lane\nTemple~ TRUE     TRUE        42
```

6. Assume continuous existence

```
for (n in biodynamic$name){
  from = min(biodynamic[biodynamic$name == n, ]$year)
  to = max(biodynamic[biodynamic$name == n, ]$year)
  for (y in from:to){
    if (!(y %in% biodynamic[biodynamic$name == n, ]$year)){
      biodynamic <- plyr::rbind.fill(biodynamic, data.frame(name = n, year = y))
    }
  }
}
```

Generate quick map that shows progression (GIF)