01_make_data_.Rmd

Abisai Lujan

2025-01-14

https://github.com/rfordatascience/tidytuesday/blob/main/data/2022/2022-06-14/readme.md

```
drought <- read.csv("./RData/drought.csv")
head(drought)
```

```
##   map_date state_abb valid_start  valid_end stat_fmt drought_lvl area_pct
## 1 20210713        AK  2021-07-13 2021-07-19        2        None    74.35
## 2 20210713        AK  2021-07-13 2021-07-19        2          D0    25.65
## 3 20210713        AK  2021-07-13 2021-07-19        2          D1     0.00
## 4 20210713        AK  2021-07-13 2021-07-19        2          D2     0.00
## 5 20210713        AK  2021-07-13 2021-07-19        2          D3     0.00
## 6 20210713        AK  2021-07-13 2021-07-19        2          D4     0.00
##   area_total pop_pct pop_total
## 1   433133.2   33.91  240644.2
## 2   149435.1   66.09  468985.8
## 3        0.0    0.00       0.0
## 4        0.0    0.00       0.0
## 5        0.0    0.00       0.0
## 6        0.0    0.00       0.0
```

1. Identify response variables

My choice: area_pct (double) Percent of state currently in that drought category

- Examine distributions

First things first! How many zeros are there?

```
sum(drought$area_pct == 0.0)
```

```
## [1] 179290
```

```
sum(drought$area_pct == 0.0)/nrow(drought)
```

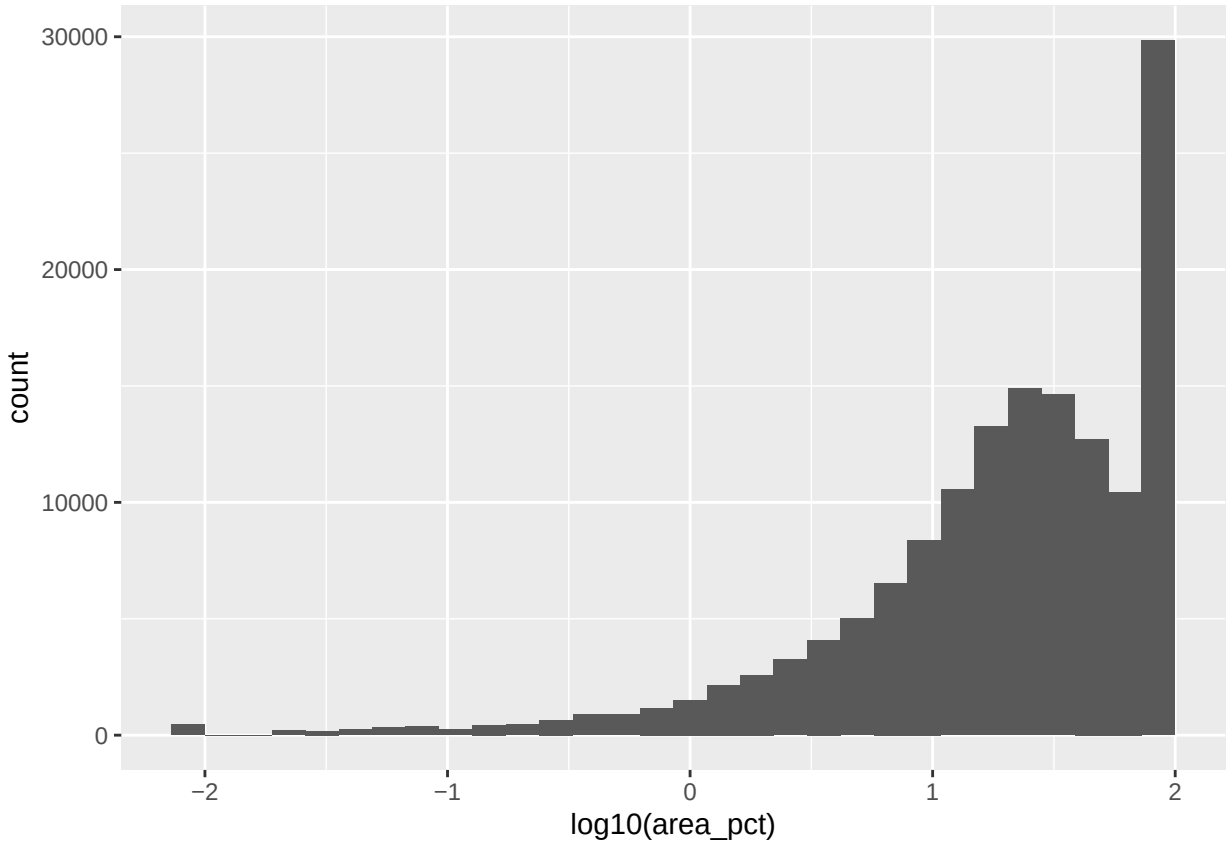
```
## [1] 0.5504286
```

55% of the data have zeros.

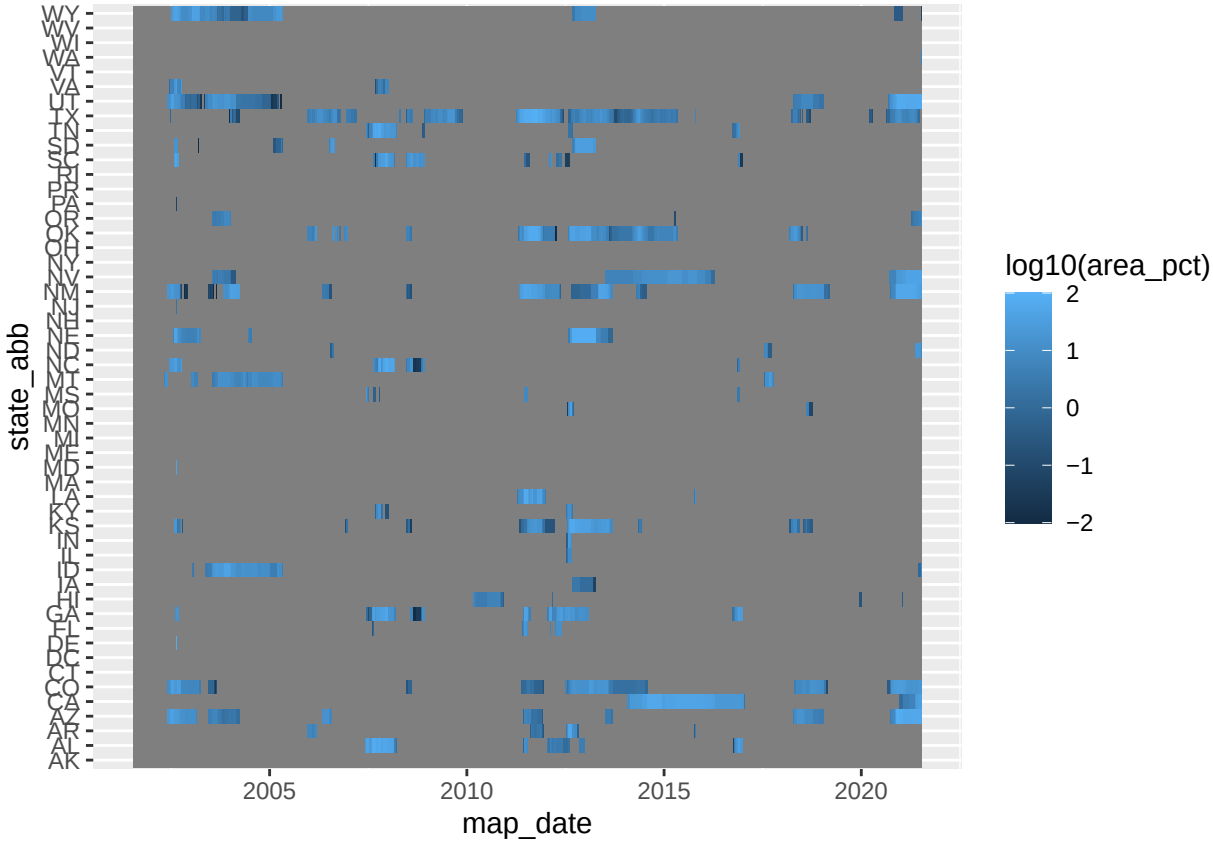
Distribution of non-zero values

```
drought %>%
  filter(area_pct > 0) %>%
  ggplot() +
  geom_histogram(aes(x=log10(area_pct)))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
drought %>%
  mutate(map_date = as.Date(as.character(map_date), format = "%Y%m%d")) %>%
  ggplot()+
  geom_tile(aes(x=map_date, y=state_abb, fill=log10(area_pct)))
```



2. Identify explanatory variables

- Make histograms or bar graphs
- Look for imbalance