Clusters California Houses

Emanuel Becerra Soto

February 11, 2020

Clusters in California Houses

For a new R Chunck

Ctrl + Alt + I

```
# Importing, Function and Global ----------------------------------------

library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggmap)
```

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.
## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
library(ggthemes)

# Using Ale Key
register_google(key='AIzaSyAs5nX1K2RGYQOVph_fnG1nEoF_L_Wgj2o')

# Loading the data ------------------------------------------------------

file <- 'cal_houses.csv'
cal_houses <- read_csv(file)
```

```
## Warning: Missing column names filled in: 'X1' [1]
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   MedInc = col_double(),
##   HouseAge = col_double(),
##   AveRooms = col_double(),
##   AveBedrms = col_double(),
##   Population = col_double(),
##   AveOccup = col_double(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   MedPrice = col_double(),
```

```
##   Cluster = col_double()
## )
```

```
# Removing 1st col that has no meaning
cal_houses <- cal_houses %>%
  select(-(X1))

cal_houses$Cluster <-
  factor(cal_houses$Cluster, levels = 0:3)
```

```
# Getting the maps ------------------------------------------------------

# San Francisco ---------------------------------------------------------

# san_francisco_map <- get_map('San Francisco, CA', source='stamen',
#                              zoom=10, maptype='terrain')
# saveRDS(san_francisco_map, 'san_francisco_map.Rds')
san_francisco_map  <- readRDS('san_francisco_map.Rds')

p_sf <- ggmap(san_francisco_map)+
  geom_point(data=cal_houses,
             aes(x=Longitude,
                 y=Latitude,
                 color=Cluster),
             size=2, alpha=0.2, shape=15)+
  scale_colour_brewer(palette='Set1')+
  theme_fivethirtyeight()
```

```
# Los Angeles -----------------------------------------------------------

# los_angeles_map <- get_map('Los Angeles, CA', source='stamen',
#                            zoom=10, maptype='terrain')
# saveRDS(los_angeles_map, 'los_angeles_map.Rds')
los_angeles_map  <- readRDS('los_angeles_map.Rds')

p_la <- ggmap(los_angeles_map)+
  geom_point(data=cal_houses,
             aes(x=Longitude,
                 y=Latitude,
                 color=Cluster),
             size=2, alpha=0.2, shape=15)+
  scale_colour_brewer(palette='Set1')+
  theme_fivethirtyeight()
```

```
# Plotting California ---------------------------------------------------

# california_map <- get_map('California, USA', source='stamen',
#                           zoom = 6, maptype='terrain')
# saveRDS(california_map, 'california_map.Rds')
california_map  <- readRDS('california_map.Rds')

p_cal <- ggmap(california_map)+
  geom_point(data=cal_houses,
             aes(x=Longitude,
                 y=Latitude,
```

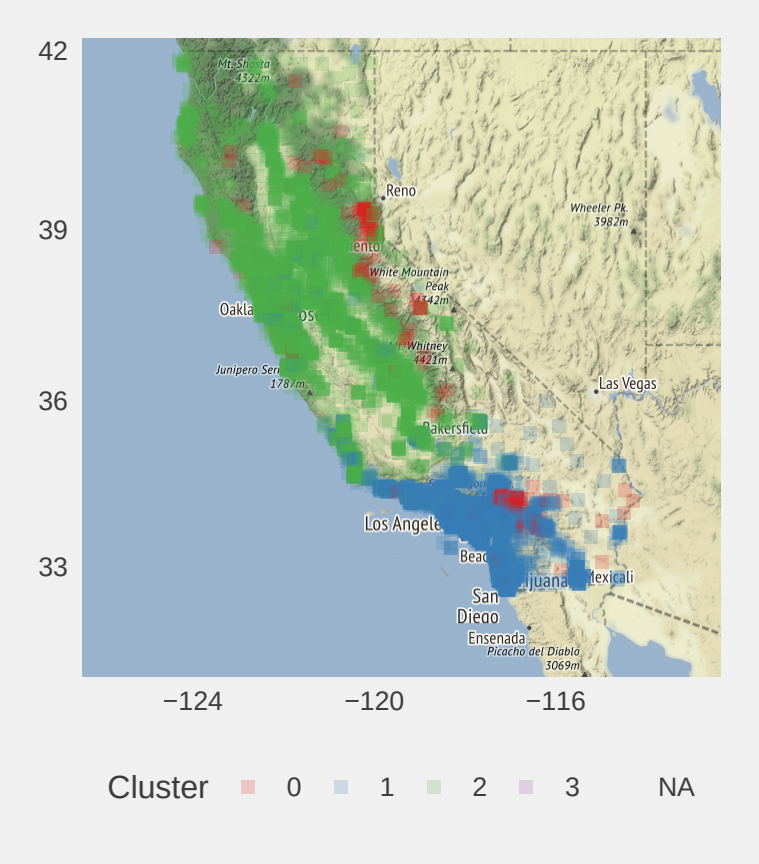
```
                color=Cluster),
            size=2, alpha=0.2, shape=15)+
  scale_colour_brewer(palette='Set1')+
  theme_fivethirtyeight()
```

```
# Plotting Outliers ------------------------------------------------------

condition <- cal_houses$Cluster == 1 | cal_houses$Cluster == 3
outliers_houses <- cal_houses[condition,]
# View(outliers_houses)

p_cal_out <- ggmap(california_map)+
  geom_point(data=outliers_houses,
             aes(x=Longitude,
                 y=Latitude,
                 color=Cluster), size=3, shape=15)+
  scale_colour_brewer(palette='Set1', drop=FALSE)+
  theme_fivethirtyeight()
```

```
p_cal_points <- ggmap(california_map)+
  geom_point(data=cal_houses,
             aes(x=Longitude,
                 y=Latitude,
                 color=Cluster),
             size=2)+
  scale_colour_brewer(palette='Set1')+
  theme_fivethirtyeight()
```

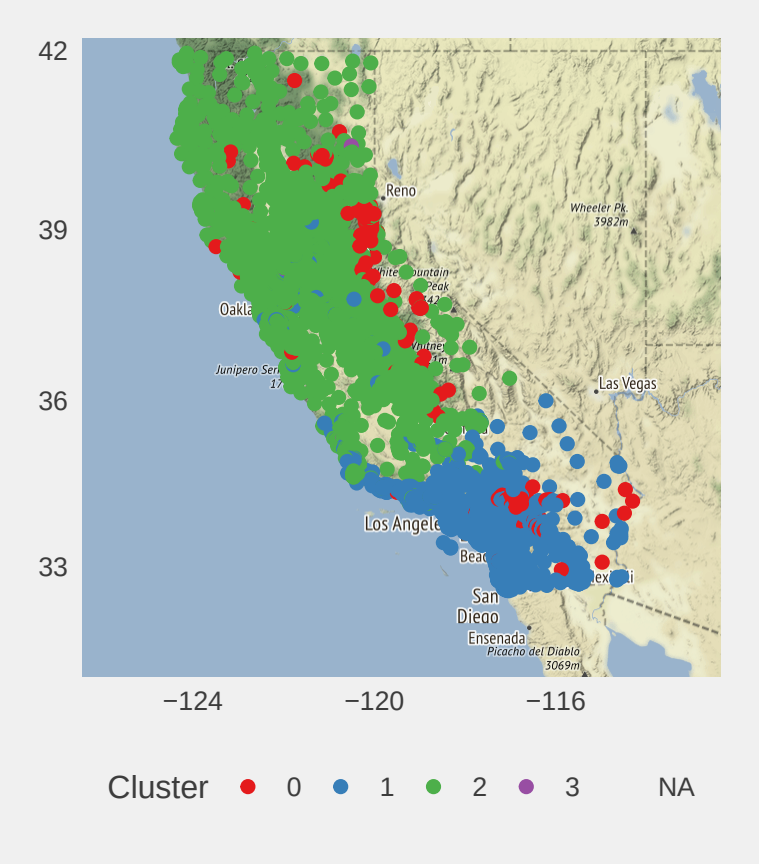
```
p_cal
```

```
## Warning: Removed 1032 rows containing missing values (geom_point).
```



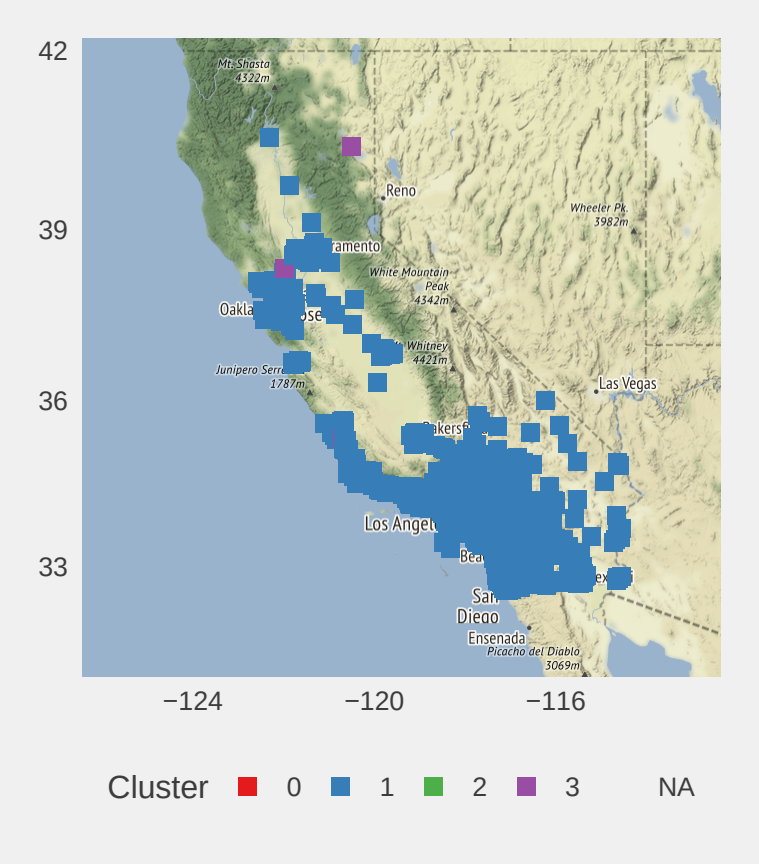

```
p_cal_points
```

```
## Warning: Removed 1032 rows containing missing values (geom_point).
```



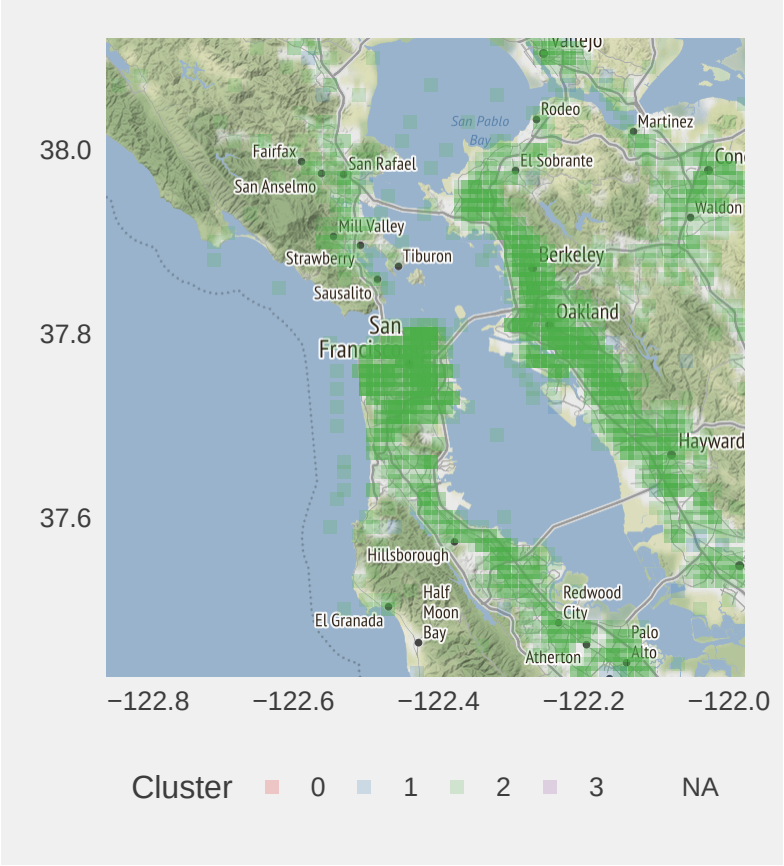

```
p_cal_out
```

```
## Warning: Removed 1032 rows containing missing values (geom_point).
```



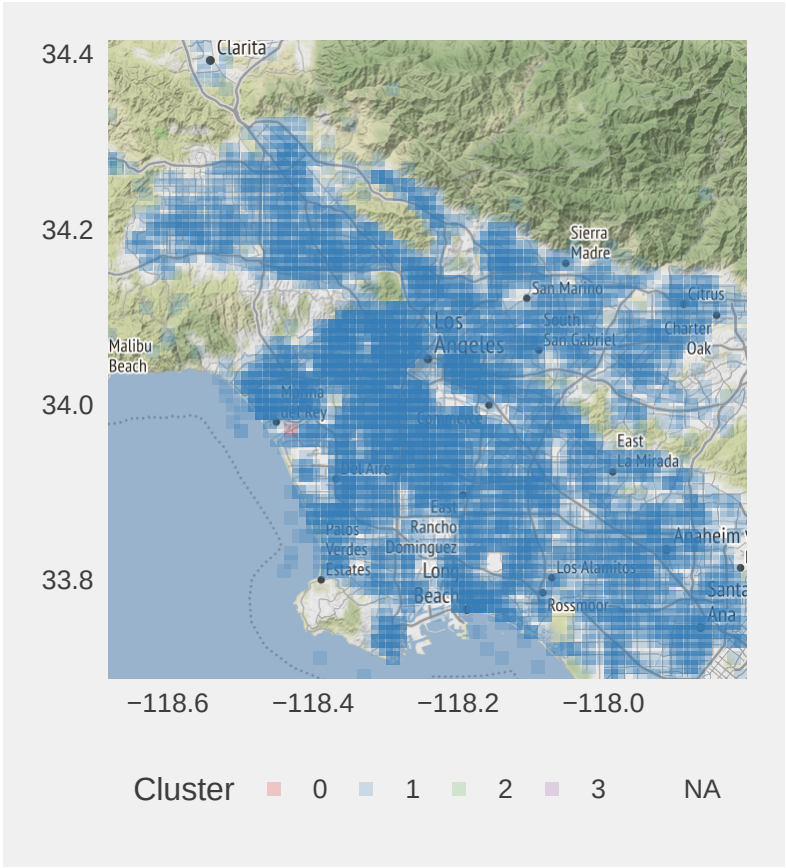

```
p_sf
```

```
## Warning: Removed 18101 rows containing missing values (geom_point).
```




```
p_la
```

```
## Warning: Removed 14501 rows containing missing values (geom_point).
```


34.4
34.2
34.0
33.8
Clarita
Sierra Madre
San Marino
Citrus
Charter Oak
South San Gabriel
Malibu Beach
East La Mirada
Rancho Dominguez
Los Alamitos
Rossmoor
−118.6
−118.4
−118.2
−118.0
Cluster
0
1
2
3
NA