Project 5

Read in the dataset you will be working with:

```
stations <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/mas
ter/data/2022/2022-03-01/stations.csv')

stations
```

```
## # A tibble: 59,927 × 70
##        X     Y OBJECTID FUEL_T…¹ STATI…² STREE…³ INTER…⁴ CITY  STATE ZIP   PLUS4
##    <dbl> <dbl>    <dbl> <chr>    <chr>   <chr>   <chr>   <chr> <chr> <chr> <lgl>
##  1 -86.3  32.4        1 CNG      Spire … 2951 C… <NA>    Mont… AL    36107 NA
##  2 -84.4  33.7        2 CNG      PS Ene… 340 Wh… From I… Atla… GA    30303 NA
##  3 -84.4  33.8        3 CNG      Metrop… 2424 P… <NA>    Atla… GA    30324 NA
##  4 -84.5  33.8        4 CNG      United… 270 Ma… <NA>    Atla… GA    30336 NA
##  5 -95.4  29.8        5 CNG      Clean … 7721A … I-10, … Hous… TX    77007 NA
##  6 -94.4  35.4        6 CNG      Arkans… 2100 S… <NA>    Fort… AR    72903 NA
##  7 -71.0  42.4        7 CNG      Clean … 1000 C… From R… East… MA    02128 NA
##  8 -71.1  42.4        8 CNG      Clean … 16 Rov… Rt 16,… Ever… MA    02149 NA
##  9 -73.9  40.7        9 CNG      Clean … 287 Ma… I-278/… Broo… NY    11211 NA
## 10 -73.9  40.6       10 CNG      Canars… 8424 D… From S… Broo… NY    11236 NA
## # … with 59,917 more rows, 59 more variables: STATION_PHONE <chr>,
## #   STATUS_CODE <chr>, EXPECTED_DATE <chr>, GROUPS_WITH_ACCESS_CODE <chr>,
## #   ACCESS_DAYS_TIME <chr>, CARDS_ACCEPTED <chr>, BD_BLENDS <chr>,
## #   NG_FILL_TYPE_CODE <chr>, NG_PSI <chr>, EV_LEVEL1_EVSE_NUM <dbl>,
## #   EV_LEVEL2_EVSE_NUM <dbl>, EV_DC_FAST_COUNT <dbl>, EV_OTHER_INFO <lgl>,
## #   EV_NETWORK <chr>, EV_NETWORK_WEB <chr>, GEOCODE_STATUS <chr>,
## #   LATITUDE <dbl>, LONGITUDE <dbl>, DATE_LAST_CONFIRMED <chr>, ID <dbl>, …
```

More information about the dataset can be found here: https://github.com/rfordatascience/tidytuesday/tree/master/data/2022/2022-03-01 (https://github.com/rfordatascience/tidytuesday/tree/master/data/2022/2022-03-01) and https://afdc.energy.gov/data_download/alt_fuel_stations_format (https://afdc.energy.gov/data_download/alt_fuel_stations_format)

Question: How does the quantity and type and electric charging stations differ between states, and are these differences related to geographical region?

Introduction: The `stations` dataset contains information from the Alternative Fuels Data Center (AFDC) on almost 60,000 alternative fuel stations from the US, Puerto Rico, and Canada. Each row corresponds to a single station, and the dataset is consistently updated; the version we are working with was last updated on January 3, 2022. There are 70 columns in this dataset, and basic information includes the station's name, fuel type, address, latitude, and longitude. Additional information includes station access details (public vs. private, hours of availability), ownership (federal, state, private, etc.), as well as information relevant to each station's specific fuel type (such as charge level and connector type for electric charging stations).

With so many features in this dataset, many of these columns contain lots of null values, which limits the analysis that can be performed. Even so, this is the most comprehensive dataset available on alternative fuel data stations in North America. The columns relevant to our analysis are:

1. `ID` : Unique numerical identifier for each station
2. `FUEL_TYPE_CODE` : The 3-4 letter code of the station's fuel type ("ELEC" for electric stations)
3. `STATE` : The 2-letter code for the state where the station is located
4. `EV_NETWORK` : The name of the network of the station. Stations without a network are labeled as "Non-Networked".
5. `ACCESS_CODE` : A description of who is allowed to access the station as either "Private" or "Public"
6. `OPEN_DATE` : The date that the station first because available for use (encoded as a string)
7. `EV_LEVEL1_EVSE_NUM` : The number of Level 1 ports available at that charging station
8. `EV_LEVEL2_EVSE_NUM` : The number of Level 2 ports available at that charging station
9. `EV_DC_FAST_COUNT` : The number of DC Fast Charging ports available at that charging station

Approach: Our first step is to clean the dataset. We will only use the 9 columns above, and since we are only looking at electric charging stations (the vast majority of the dataset), we will filter for only stations that are electric (`FUEL_TYPE_CODE == "ELEC"`), as well as for stations whose state information is not null. For the 3 columns that count the number of ports, zeroes are treated as null, so we will replace these null values with 0. Additionally, the `EV_NETWORK` column will be broken up into 2 columns: `Non_Networked` (value of 1 if `EV_NETWORK == "Non-Networked"` and 0 otherwise) and `Networked` (the name of the network if it is not null and not "Non-Networked", null otherwise). The final two column additions are `Public` , which encodes public stations as 1 (`ACCESS_CODE == "Public"`) and 0 otherwise, and `OPEN_DATE` , which is a numerical encoding of the date the station became available.

With this cleaned dataset, we will then use hierarchical clustering to group states with similar quantities and types of charging stations, using a dendrogram to visualize how these states are clustered. We can then use this dendogram to pick the appropriate number of clusters and plot a map of the United States that colors each state by its cluster. We must do 2 final pieces of data transformation to make this analysis possible: 1) group our cleaned stations dataset by state, taking the sums of the 3 port count and `Non_Networked` columns, distinct count of different charging networks (`Networked`), and the means of the `Public` and `OPEN_DATE` columns, and 2) joining the `US_states` dataset to this aggregation, which contains geocoded information on each state so that we can plot a map of the United States. These visuals allow us to see how the states are grouped, and we can then determine if geography plays a significant role in this grouping.

Analysis:

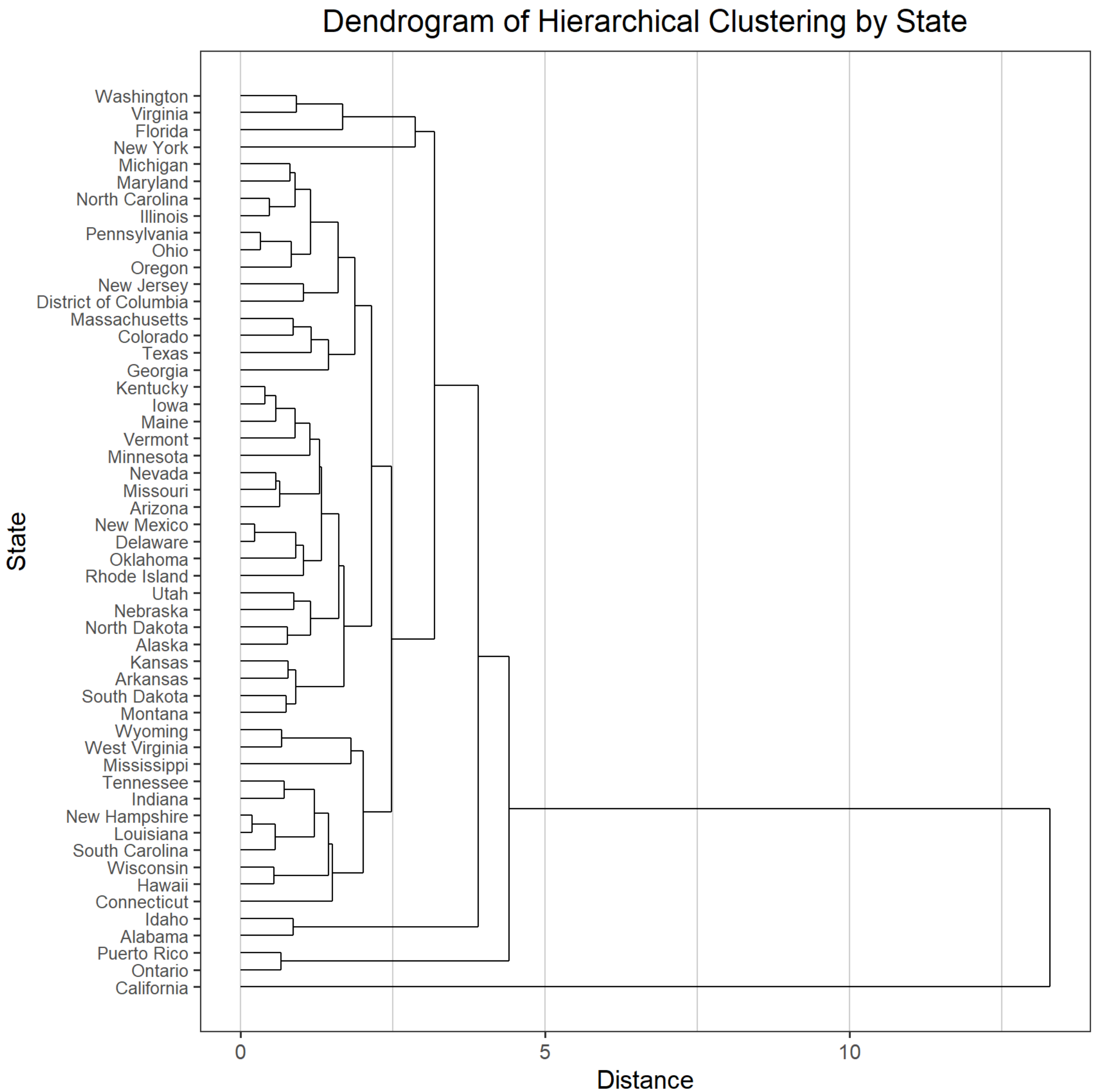
First, we will create the cleaned version of the `stations` dataset, group it by state, then join it to `US_states`:

```
keep_columns = c("ID", "FUEL_TYPE_CODE", "STATE", "EV_NETWORK", "ACCESS_CODE", "OPEN_DATE",
                 "EV_LEVEL1_EVSE_NUM", "EV_LEVEL2_EVSE_NUM", "EV_DC_FAST_COUNT")

# Transforming the raw "stations" dataset
electric_stations <- stations %>%
  filter(
    FUEL_TYPE_CODE == "ELEC",   # Filtering only for electric stations with state info
    !is.na(STATE)
  ) %>%
  select(all_of(keep_columns)) %>% # Keeping only the columns listed above
  mutate(                       # Adding the columns outlined in the Approach section
    EV_LEVEL1_EVSE_NUM = ifelse(is.na(EV_LEVEL1_EVSE_NUM), 0, EV_LEVEL1_EVSE_NUM),
    EV_LEVEL2_EVSE_NUM = ifelse(is.na(EV_LEVEL2_EVSE_NUM), 0, EV_LEVEL2_EVSE_NUM),
    EV_DC_FAST_COUNT = ifelse(is.na(EV_DC_FAST_COUNT), 0, EV_DC_FAST_COUNT),
    Non_Networked = ifelse(EV_NETWORK == "Non-Networked", 1, 0),
    Networked = ifelse(EV_NETWORK != "Non-Networked", EV_NETWORK, NA_character_),
    Public = ifelse(ACCESS_CODE == "public", 1, 0),
    OPEN_DATE = as.numeric(as.Date(OPEN_DATE)),
  )

US_states <- readRDS(url("https://wilkelab.org/SDS375/datasets/US_states.rds"))

# Grouping the cleaned dataset by state then adding geocoded information from "US_states".
stations_grouped <- electric_stations %>%
  group_by(STATE) %>% # Grouping by State
  summarize(          # Summarizing the columns as outlined in the Approach section
    .groups = "keep",
    EV_LEVEL1_EVSE_NUM = sum(EV_LEVEL1_EVSE_NUM, na.rm=TRUE),
    EV_LEVEL2_EVSE_NUM = sum(EV_LEVEL2_EVSE_NUM, na.rm=TRUE),
    EV_DC_FAST_COUNT = sum(EV_DC_FAST_COUNT, na.rm=TRUE),
    Non_Networked = sum(Non_Networked, na.rm=TRUE),
    Networked = n_distinct(Networked, na.rm=TRUE),
    Public = mean(Public, na.rm = TRUE),
    OPEN_DATE = mean(OPEN_DATE, na.rm=TRUE)
  ) %>%
  ungroup() %>%
  # Joining the "US_states" dataset to our grouped "stations" dataset
  full_join(US_states, by = c("STATE" = "state_code")) %>%
  mutate(
    state_name = case_when(
      STATE == "PR" ~ "Puerto Rico", # Matching the state name
      STATE == "ON" ~ "Ontario",       # for Ontario and Puerto Rico
      TRUE ~ name
    )
  )

stations_grouped
```

```
## # A tibble: 53 × 12
##    STATE EV_LEVEL1_…¹ EV_LE…² EV_DC…³ Non_N…⁴ Netwo…⁵ Public OPEN_…⁶ GEOID name
##    <chr>        <dbl>   <dbl>   <dbl>   <dbl>   <int>  <dbl>   <dbl> <chr> <chr>
##  1 AK               3      75      17      41       6  0.943  18315. 02    Alas…
##  2 AL              35     529     118     122       7  0.715  17533. 01    Alab…
##  3 AR               5     396      66      54       6  0.897  17754. 05    Arka…
##  4 AZ              10    1827     424     150      11  0.949  18016. 04    Ariz…
##  5 CA             645   33758    6852    1498      16  0.944  18282. 06    Cali…
##  6 CO              90    3330     588     250      12  0.929  18181. 08    Colo…
##  7 CT              76    1038     316     276      11  0.872  17305. 09    Conn…
##  8 DC              43     743      41      45      10  0.842  18032. 11    Dist…
##  9 DE               5     218      91      23       7  0.925  18137. 10    Dela…
## 10 FL             370    5235    1221     415      12  0.915  17970. 12    Flor…
## # … with 43 more rows, 2 more variables: geometry <MULTIPOLYGON>,
## #   state_name <chr>, and abbreviated variable names ¹EV_LEVEL1_EVSE_NUM,
## #   ²EV_LEVEL2_EVSE_NUM, ³EV_DC_FAST_COUNT, ⁴Non_Networked, ⁵Networked,
## #   ⁶OPEN_DATE
```

We will now cluster the states, using Euclidean distance to calculate the distances in between each state and the UPGMA ("Unweighted Pair Group Method with Arithmetic Mean") clustering methodology for our hierarchical clustering, then plot the resulting dendrogram:

```
# Calculating the Euclidean distance in between each state
dist_out <- stations_grouped %>%
  select(-c("STATE","GEOID","name","geometry")) %>%
  column_to_rownames(var = "state_name") %>%
  scale() %>%
  dist(method = "euclidean")

# Clustering each state using the UPGMA clustering method
hc_out <- hclust(dist_out, method = "average")

# Plotting the dendrogram of the clustering analysis
ggdendrogram(hc_out, rotate=TRUE) +
  # Formatting
  labs(
    title = "Dendrogram of Hierarchical Clustering by State",
    x = "State",
    y = "Distance"
  ) +
  theme_bw(15) +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.y = element_text(size = 10),
    panel.grid.major.x = element_line(color = "gray80", size= 0.5),
    panel.grid.minor.x = element_line(color = "gray80", size= 0.5),
    panel.grid.major.y = element_blank(),
    panel.grid.minor.y = element_blank(),
  )
```


Dendrogram of Hierarchical Clustering by State
State
Washington
Virginia
Florida
New York
Michigan
Maryland
North Carolina
Illinois
Pennsylvania
Ohio
Oregon
New Jersey
District of Columbia
Massachusetts
Colorado
Texas
Georgia
Kentucky
Iowa
Maine
Vermont
Minnesota
Nevada
Missouri
Arizona
New Mexico
Delaware
Oklahoma
Rhode Island
Utah
Nebraska
North Dakota
Alaska
Kansas
Arkansas
South Dakota
Montana
Wyoming
West Virginia
Mississippi
Tennessee
Indiana
New Hampshire
Louisiana
South Carolina
Wisconsin
Hawaii
Connecticut
Idaho
Alabama
Puerto Rico
Ontario
California
0
5
10
Distance

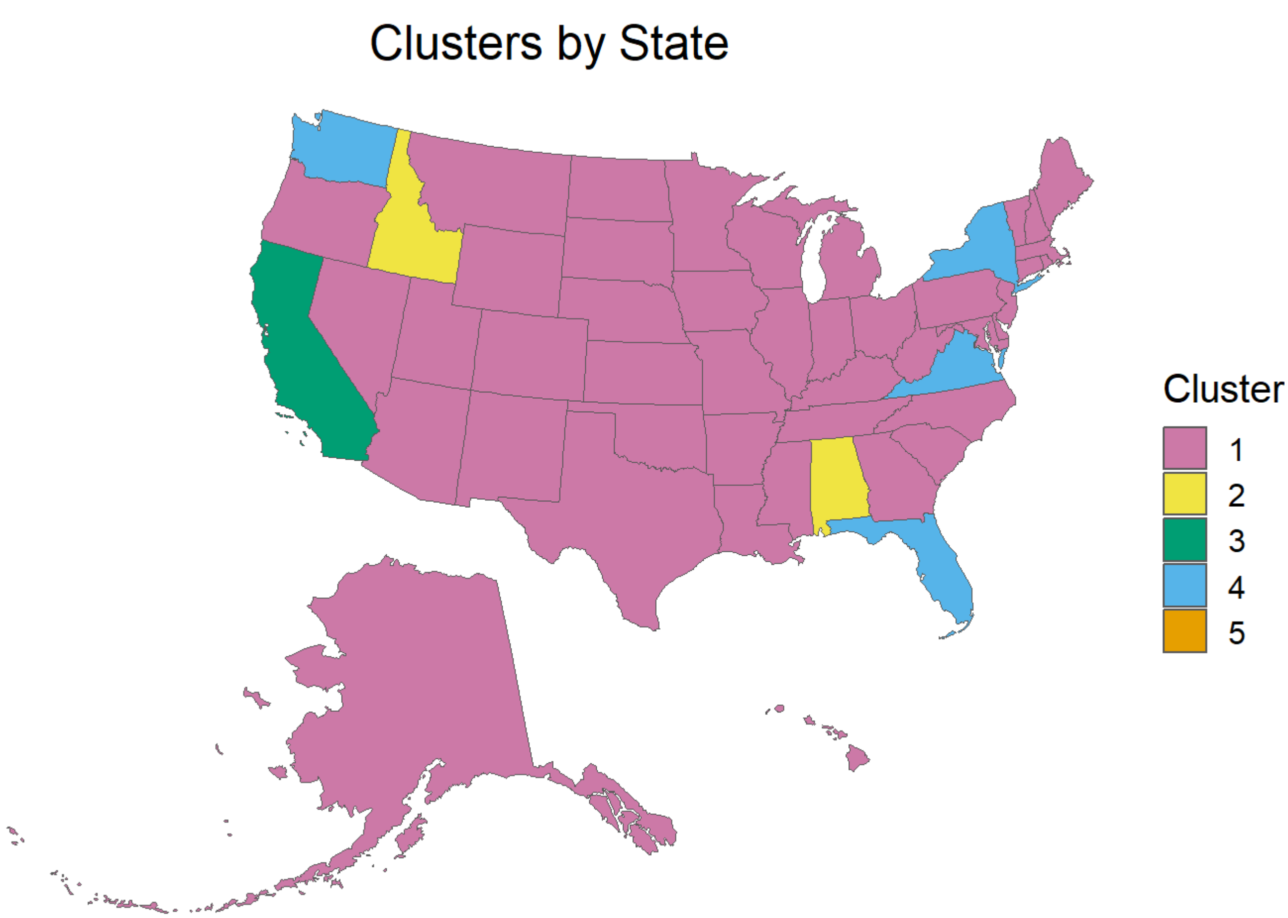
From this dendrogram, we have decided to split the states into 5 clusters. To get an idea of the properties of these clusters, we will provide a table of the mean values of each of the columns used to cluster the states:

```
# Cutting the dataset into 5 clusters
cluster <- cutree(hc_out, k = 5)

# Adding the cluster label to our "stations_grouped" dataset
stations_clustered <- stations_grouped %>%
  left_join(
    tibble(
      state_name = names(cluster),
      cluster = factor(cluster)
    ),
    by = "state_name"
  )

# Grouping "stations_clustered" by cluster then taking the mean of each of its columns
cluster_summary <- stations_clustered %>%
  group_by(cluster) %>%
  summarize(
    .groups = "keep",
    Number_of_States = n(), # Adding the number of states in each cluster
    EV_LEVEL1_EVSE_NUM = mean(EV_LEVEL1_EVSE_NUM, na.rm=TRUE),
    EV_LEVEL2_EVSE_NUM = mean(EV_LEVEL2_EVSE_NUM, na.rm=TRUE),
    EV_DC_FAST_COUNT = mean(EV_DC_FAST_COUNT, na.rm=TRUE),
    Non_Networked = mean(Non_Networked, na.rm=TRUE),
    Networked = mean(Networked, na.rm=TRUE),
    Public = mean(Public, na.rm = TRUE),
    OPEN_DATE = mean(OPEN_DATE, na.rm=TRUE)
  ) %>%
  # Recoding "OPEN_DATE" as a date value
  mutate(OPEN_DATE = as.Date(OPEN_DATE, origin = "1970-01-01")) %>%
  ungroup()

cluster_summary
```

```
## # A tibble: 5 × 9
##   cluster Number_of_…¹ EV_LE…² EV_LE…³ EV_DC…⁴ Non_N…⁵ Netwo…⁶ Public OPEN_DATE
##   <fct>          <int>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>  <dbl> <date>
## 1 1                 44    41.3   1165.    261.    124.    8.84  0.916 2018-10-31
## 2 2                  2    22.5    392      99      92     7.5   0.732 2018-02-28
## 3 3                  1   645    33758    6852    1498    16     0.944 2020-01-20
## 4 4                  4   218     4430.    890     340.   13     0.908 2019-04-10
## 5 5                  2     0        7       1       0     1.5   1     2021-06-01
## # … with abbreviated variable names ¹Number_of_States, ²EV_LEVEL1_EVSE_NUM,
## #   ³EV_LEVEL2_EVSE_NUM, ⁴EV_DC_FAST_COUNT, ⁵Non_Networked, ⁶Networked
```

Finally, we will plot a map of the United States with each state colored by its cluster:

```
stations_clustered %>%
  ggplot() +
  geom_sf(aes(geometry=geometry, fill=cluster)) +
  # Formatting
  scale_fill_manual(values = c("#CC79A7","#F0E442","#009E73","#56B4E9","#E69F00")) +
  theme_void(15) +
  labs(
    title = "Clusters by State",
    fill = "Cluster"
  ) +
  theme(plot.title = element_text(hjust = 0.5)) +
  coord_sf()
```



Discussion: The states are broken up into 5 clusters: 5) Ontario (Canada) and Puerto Rico; 4) Florida, Washington, Virginia, and New York; 3) California; 2) Alabama and Idaho; and 1) everything else. The main factor in grouping the states seems to be the total number of stations, which is reflected in each cluster's average values of `EV_LEVEL1_EVSE_NUM` , `EV_LEVEL2_EVSE_NUM` , `EV_DC_FAST_COUNT` , `Non_Networked` , and `Networked` as seen in the `cluster_summary` table. From largest to smallest, the clusters are ordered by these values as clusters 3, 4, 1, 2, and 5. This ordering is almost exactly the same if we look at the average percentage of stations that are public (the `Public` column), except cluster 5 is 100% public and clusters 4 and 1 are swapped. One additional observation is that the average `OPEN_DATE` of the stations in each cluster seems to be correlated with the total number of stations; clusters with more stations tend to have younger stations on average (with the exception of cluster 5). This suggests that the states with the most stations have added the majority of their stations more recently (as recent as 2019 and 2020).

The map that shades each state by cluster does not show a strong relationship between cluster and geographic location; clusters 2 and 4 specifically are spread throughout the entire United States. It should also be noted that cluster 5 (Ontario and Puerto Rico) are not reflected on this map. Factors driving the type and quantity of stations may include the population of each state and whether or not that state includes large metropolitan areas, but perhaps the most important factor is state legislation. California in particular is very electric vehicle friendly, pledging to stop the sale of all new gasoline vehicles by 2035, and this is reflected by it being in its own cluster. Other states like New York have also put forth strong efforts to make the shift from gasoline to electric. The data suggests that Alabama and Idaho (cluster 2) have fallen behind in this regard, and the sample sizes for Ontario and Puerto Rico (cluster 5) are too small to draw any conclusions from.

Overall, the quantity of stations, not the type, seems to be the biggest factor in what makes the electric vehicle charging infrastructure similar among states. Geography is less important than strong legislation in building out charging stations to support EV adoption. The world is making the shift to sustainable energy, so looking to the states that are leading this charge in America will teach us how and where to build charging stations in a way that will make this shift as easy as possible.