Final Project Group 2

Donald Rogers, Brannan Kovachev, Jeremy Wright

2022-11-17

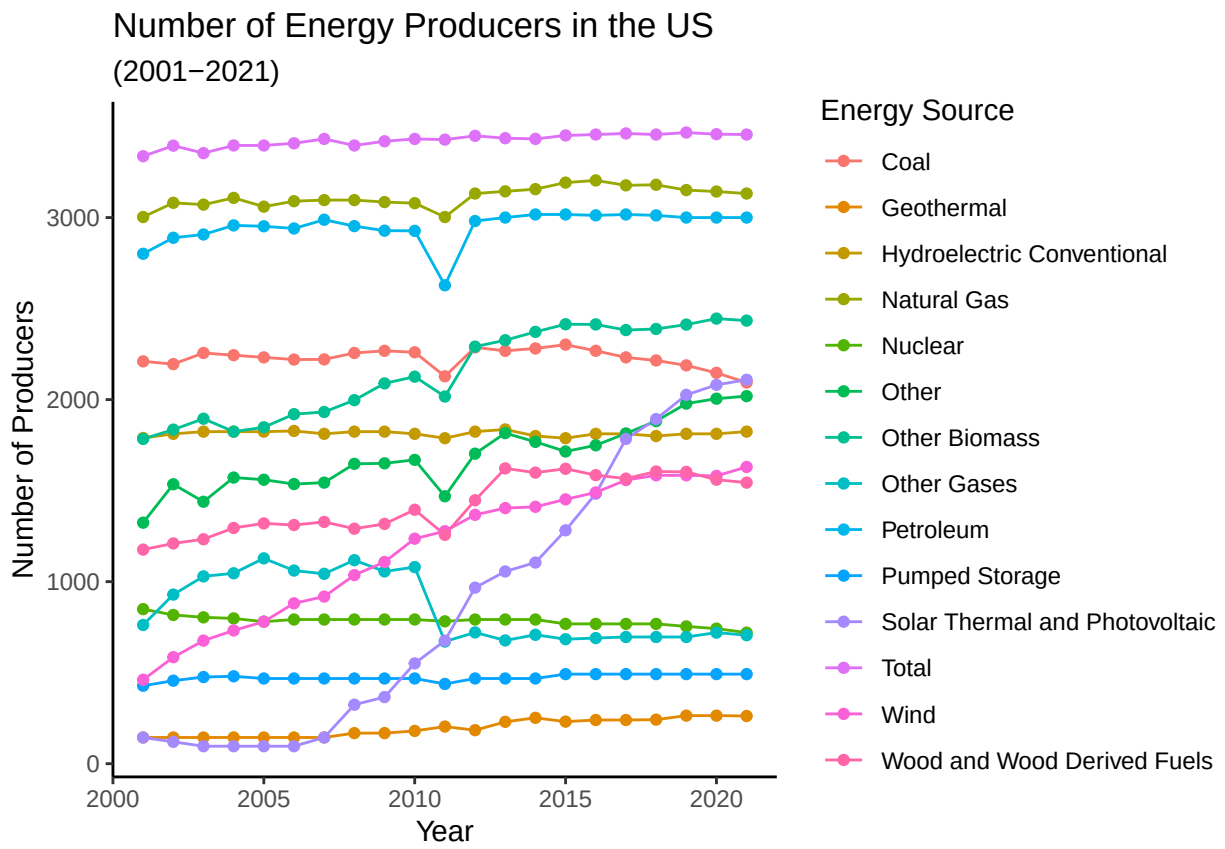
Part 1

Introduction

Our group decided to do a project examining the role that Green Energy plays within the US and how emissions from energy generation can affect the population surrounding them. My focus was on more generalized information, especially looking at the US from a country-wide scale. To start, I first examined how energy was generated in the US.

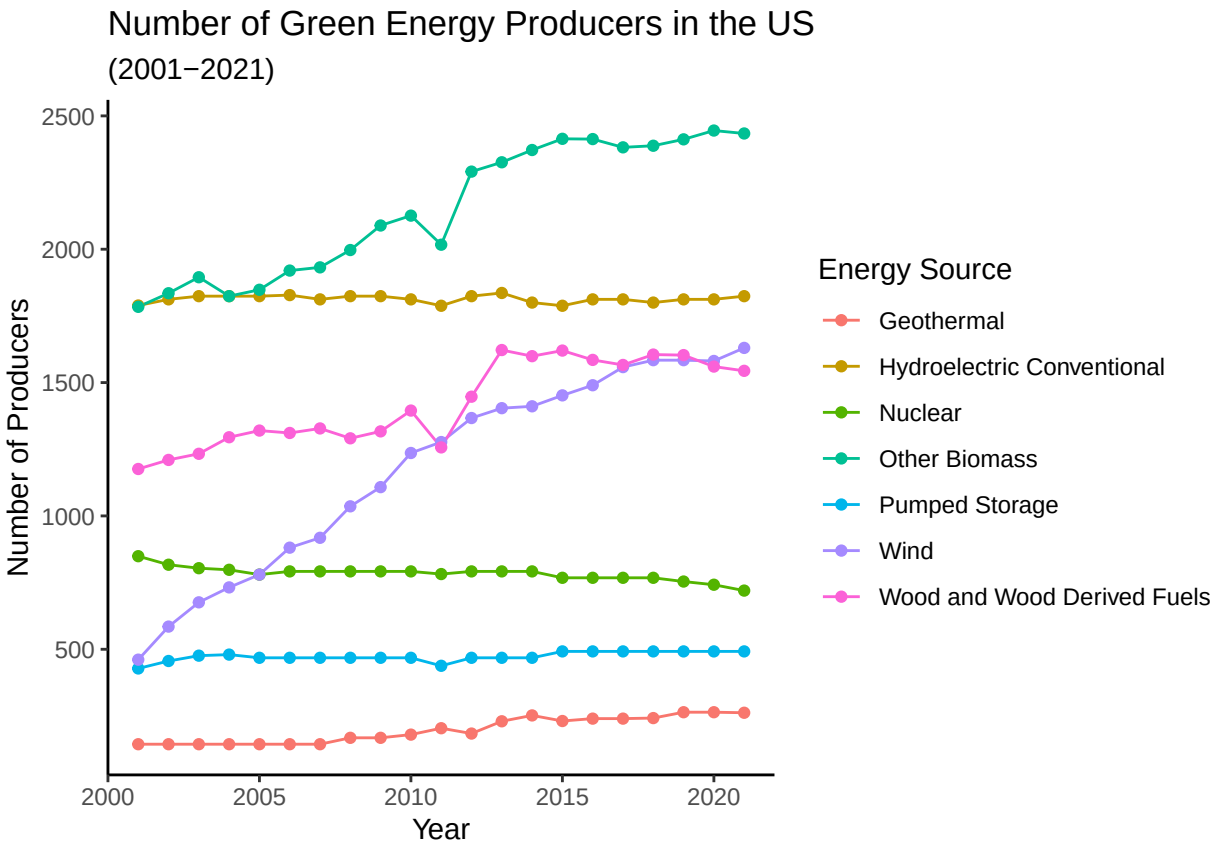
Question 1: What is the historical trend for clean energy since 2001 in the US? First, I generated a line graph detailing the different amounts of energy producers within the US. The dataset had data from 2001-2022, but the 2022 data was incomplete, so the graphs only show up until 2021. This visualization only takes into account the total amount of facilities devoted to each resource, instead of total power generated, but provides a good idea of how our country's energy footprint has changed over the years.

```
# scatterplot with line of data from all years before 2022, colored by source
countyGeneration %>%
  filter(YEAR < 2022) %>%
  group_by(YEAR, `ENERGY SOURCE`) %>%
  summarize(count = n()) %>%
  ggplot(aes(x = YEAR, y = count, group = `ENERGY SOURCE`, color=`ENERGY SOURCE`))+
  geom_line(aes(color = `ENERGY SOURCE`)) +
  geom_point(aes(color = `ENERGY SOURCE`)) +
  labs(title = "Number of Energy Producers in the US", subtitle = "(2001-2021)", y = "Number of Produce
       x = "Year", color = "Energy Source") +
  theme_classic()
```



As can be seen above, energy producers based off of fossil fuels (or "Brown Energy") are still dominant, with natural gas and petroleum topping the charts, but there have been increases to the amount of green energy producers, especially since 2010. Next, let's filter out the brown energy sources so that we can see their trends better.

```
# Same graph but filtering out all sources of brown energy
countyGeneration %>%
  filter(YEAR < 2022) %>%
  filter(`ENERGY SOURCE` %in% c("Geothermal","Hydroelectric Conventional","Nuclear",
                                "Other Biomass", "Pumped Storage","Solar Thermal
                                and Photovoltaic", "Wind", "Wood and Wood Derived Fuels")) %>%
  group_by(YEAR, `ENERGY SOURCE`) %>%
  summarize(count = n()) %>%
  ggplot(aes(x = YEAR, y = count, group = `ENERGY SOURCE`, color=`ENERGY SOURCE`))+
  geom_line(aes(color = `ENERGY SOURCE`)) +
  geom_point(aes(color = `ENERGY SOURCE`)) +
  labs(title = "Number of Green Energy Producers in the US", subtitle = "(2001-2021)",
       y = "Number of Producers", x = "Year", color = "Energy Source") +
  theme_classic()
```



As we see here, most of the different types of green energy have had a large increase since 2001. Solar, Thermal, and Photovoltaic have had the most growth, with Biomass and Wind following close behind. Surprisingly enough, the only form of energy generation that seems to decline is Nuclear. This is likely due to the fact that nuclear energy is seen as dangerous, although many of those concerns have been answered in recent years.

Question 2: What types of energy are most efficient within the United States? Moving to possibly answer the previous question, I next examined which forms of energy were the most efficient, in terms of generating the least amount of pollutants per Megawatt hour of electricity. To provide some contest, an average American household uses about 10 MWh of power per year. The dataset also has emissions separated into three different types. They are Carbon Dioxide, Sulfur Dioxide, and Nitrogen Oxides (NO and NO2). To give a frame of reference for the amount of pollutants generated, driving a car for a year will generate about 4.5 metric tons of CO2. Sulfur Dioxide and Nitrogen Oxides are much more harmful than CO2, and are generated much less frequently than CO2 by consumers. Moving to the visualizations, I first found how many Megawatt hours of energy were generated per state in 2018 (2018 was the most recent year there was data for within the dataset), then how many metric tons of CO2, SO2, and NOx were generated per state in 2018. After joining the two datasets together, the following table was created.

```
# Table showing all states, their emissions by source, and their power generated by source
combined18 %>%
  rename("MWh Generated" = "MWh") %>%
  head(50) %>%
  kbl(caption = "Power Generation and Emissions") %>%
  kable_minimal() %>%
  add_header_above(c(" " = 3, "Gas Emissions (Metric Tons)" = 3))
```

Table 1: Power Generation and Emissions

			Gas Emissions (Metric Tons)		
State	Source	MWh Generated	CO2	SO2	NOx
AK	Coal	1257128	2700198	3040	4426
AK	Hydroelectric Conventional	3328450	NA	NA	NA
AK	Natural Gas	5895805	2774752	12	19076
AK	Other	-6200	NA	NA	NA
AK	Other Biomass	90948	0	38	992
AK	Petroleum	1618558	1304152	2140	13414
AK	Wind	310030	NA	NA	NA
AK	Wood and Wood Derived Fuels	0	NA	NA	NA
AL	Coal	63555040	61108096	24242	30490
AL	Hydroelectric Conventional	22286278	NA	NA	NA
AL	Natural Gas	117600837	51005388	260	19686
AL	Nuclear	78925654	NA	NA	NA
AL	Other	-98	NA	NA	NA
AL	Other Biomass	75344	0	0	908
AL	Other Gases	9608	0	4	742
AL	Petroleum	131407	166042	172	234
AL	Solar Thermal and Photovoltaic	714504	NA	NA	NA
AL	Wood and Wood Derived Fuels	6817414	0	50134	11078
AR	Coal	59992202	58655136	94742	37820
AR	Hydroelectric Conventional	6017550	NA	NA	NA
AR	Natural Gas	41247410	17938780	102	6562
AR	Nuclear	25441636	NA	NA	NA
AR	Other	10174	33264	356	62
AR	Other Biomass	149900	0	26	1536
AR	Petroleum	71034	71726	86	34
AR	Pumped Storage	80490	NA	NA	NA
AR	Solar Thermal and Photovoltaic	406826	NA	NA	NA
AR	Wood and Wood Derived Fuels	2581480	0	23058	6392
AZ	Coal	61489246	61795004	29694	51414
AZ	Hydroelectric Conventional	13964484	NA	NA	NA
AZ	Natural Gas	74336150	31639520	222	23306
AZ	Nuclear	62194518	NA	NA	NA
AZ	Other	-5072	NA	NA	NA
AZ	Other Biomass	74750	0	0	1046
AZ	Petroleum	99892	78600	12	48
AZ	Pumped Storage	-9052	NA	NA	NA
AZ	Solar Thermal and Photovoltaic	10280753	NA	NA	NA
AZ	Wind	1060380	NA	NA	NA
AZ	Wood and Wood Derived Fuels	364240	0	50	196
CA	Coal	562656	2719364	654	1570
CA	Geothermal	23353682	614352	0	0
CA	Hydroelectric Conventional	52661333	NA	NA	NA
CA	Natural Gas	179208948	83144784	378	83738
CA	Nuclear	36427038	NA	NA	NA
CA	Other	1657536	551118	672	1102
CA	Other Biomass	5647697	0	34	39868
CA	Other Gases	2907920	0	16	3778
CA	Petroleum	137756	128118	202	1096
CA	Pumped Storage	-297148	NA	NA	NA
CA	Solar Thermal and Photovoltaic	53970362	NA	NA	NA

The previous table showed only the first 50 observations and with over 500 observations in total, there's not much purpose to putting all of them in one table. So, since we are focusing on the entire US, let's look at the totals for the entire US to get a good idea of what energy source is the most efficient.

```
# Table filtered to just totals from US
combined18 %>%
  filter(State == "US") %>%
  rename("MWh Generated" = "MWh", "Energy Source" = "Source") %>%
  kbl(caption = "US Total Power Generation and Emissions") %>%
  kable_minimal() %>%
  kable_styling(latex_options = "hold_position") %>%
  add_header_above(c(" " = 3, "Gas Emissions (Metric Tons)" = 3))
```

Table 2: US Total Power Generation and Emissions

			Gas Emissions (Metric Tons)		
State	Energy Source	MWh Generated	CO2	SO2	NOx
US	Total,Coal	2291924383	2340396682	2405118	1566792
US	Total,Geothermal	31934268	840078	0	0
US	Total,Hydroelectric Conventional	585047978	NA	NA	NA
US	Total,Natural Gas	2937453247	1324805112	7664	789368
US	Total,Nuclear	1614168954	NA	NA	NA
US	Total,Other	25946409	28154278	39750	70790
US	Total,Other Biomass	41791530	0	2828	253370
US	Total,Other Gases	26925497	0	1446	52768
US	Total,Petroleum	50451278	54496090	149992	82530
US	Total,Pumped Storage	-11809078	NA	NA	NA
US	Total,Solar Thermal and Photovoltaic	127650645	NA	NA	NA
US	Total,Wind	545299578	NA	NA	NA
US	Total,Wood and Wood Derived Fuels	82010618	0	536056	154764

As the table shows, there are many different types of energy that don't have any emissions recorded for those types of power plants. This is to be expected, as many of these sources don't generate any emissions since they don't require anything to be burnt to generate power. Nuclear, while appearing to generate emissions, only generates steam, since the nuclear reactions in the core of the plant are just used to heat water and don't make any gases by themselves. So, to answer the question of which energy generation is the most efficient (in terms of amount of emissions), we have our answer: Hydroelectric, Nuclear, Solar, Thermal, Photovoltaic, and Wind. I didn't include Pumped Storage, as the table shows that we actually lose power in total when using pumped storage. To see which forms of energy that create emissions are most efficient, all of the sources with no emissions were dropped for the next table.

```
# Further filtered to not include sources with no emissions in all three categories
combined18 %>%
  na.omit() %>%
  filter(State == "US") %>%
  rename("MWh Generated" = "MWh", "Energy Source" = "Source") %>%
  kbl(caption = "US Total Power Generation and Emissions") %>%
  kable_minimal() %>%
  kable_styling(latex_options = "HOLD_position") %>%
  add_header_above(c(" " = 3, "Gas Emissions (Metric Tons)" = 3))
```

Table 3: US Total Power Generation and Emissions

State	Energy Source	MWh Generated	Gas Emissions (Metric Tons)		
			CO2	SO2	NOx
US	Total,Coal	2291924383	2340396682	2405118	1566792
US	Total,Geothermal	31934268	840078	0	0
US	Total,Natural Gas	2937453247	1324805112	7664	789368
US	Total,Other	25946409	28154278	39750	70790
US	Total,Other Biomass	41791530	0	2828	253370
US	Total,Other Gases	26925497	0	1446	52768
US	Total,Petroleum	50451278	54496090	149992	82530
US	Total,Wood and Wood Derived Fuels	82010618	0	536056	154764

As we can see from the remaining fuels, most of the emissions created in the US come from Brown Energy, such as Coal and Natural Gas, which also account for a majority of the energy created. Before doing any calculations, we can see that Natural Gas is much more efficient than Coal when it comes to emissions, as there were more MWh generated by Natural Gas sources yet less emissions in every category. Some forms of Green Energy (such as Wood, Biomass, and Geothermal) are still found in this table since they produce some forms of emissions, but not in all three categories.

```
# Puts CO2 calculations in table and drops any value that generates 0 tons/mwh
emissionsMWh %>%
  filter(State == "US") %>%
  select(State, Source, `Tons CO2/MWh`) %>%
  rename("Energy Source" = "Source") %>%
  filter(`Tons CO2/MWh` > 0) %>%
  arrange(`Tons CO2/MWh`) %>%
  kbl(caption = "Carbon Dioxide Generation per Megawatt Hour") %>%
  kable_minimal() %>%
  kable_styling(latex_options = "hold_position")
```

Table 4: Carbon Dioxide Generation per Megawatt Hour

State	Energy Source	Tons CO2/MWh
US	Total,Geothermal	0.0263065
US	Total,Natural Gas	0.4510047
US	Total,Coal	1.0211492
US	Total,Petroleum	1.0801726
US	Total,Other	1.0850934

```
# Same as above but with SO2
emissionsMWh %>%
  filter(State == "US") %>%
  select(State, Source, `Tons SO2/MWh`) %>%
  rename("Energy Source" = "Source") %>%
  filter(`Tons SO2/MWh` > 0) %>%
  arrange(`Tons SO2/MWh`) %>%
  kbl(caption = "Sulfur Dioxide Generation per Megawatt Hour") %>%
  kable_minimal() %>%
  kable_styling(latex_options = "hold_position")
```

Table 5: Sulfur Dioxide Generation per Megawatt Hour

State	Energy Source	Tons SO2/MWh
US	Total,Natural Gas	0.0000026
US	Total,Other Gases	0.0000537
US	Total,Other Biomass	0.0000677
US	Total,Coal	0.0010494
US	Total,Other	0.0015320
US	Total,Petroleum	0.0029730
US	Total,Wood and Wood Derived Fuels	0.0065364

```
# Same as above but with NOx
emissionsMWh %>%
  filter(State == "US") %>%
  select(State, Source, `Tons NOx/MWh`) %>%
  rename("Energy Source" = "Source") %>%
  filter(`Tons NOx/MWh` > 0) %>%
  arrange(`Tons NOx/MWh`) %>%
  kbl(caption = "Nitrous Oxides (NO and NO2) Generation per Megawatt Hour") %>%
  kable_minimal() %>%
  kable_styling(latex_options = "hold_position")
```

Table 6: Nitrous Oxides (NO and NO2) Generation per Megawatt Hour

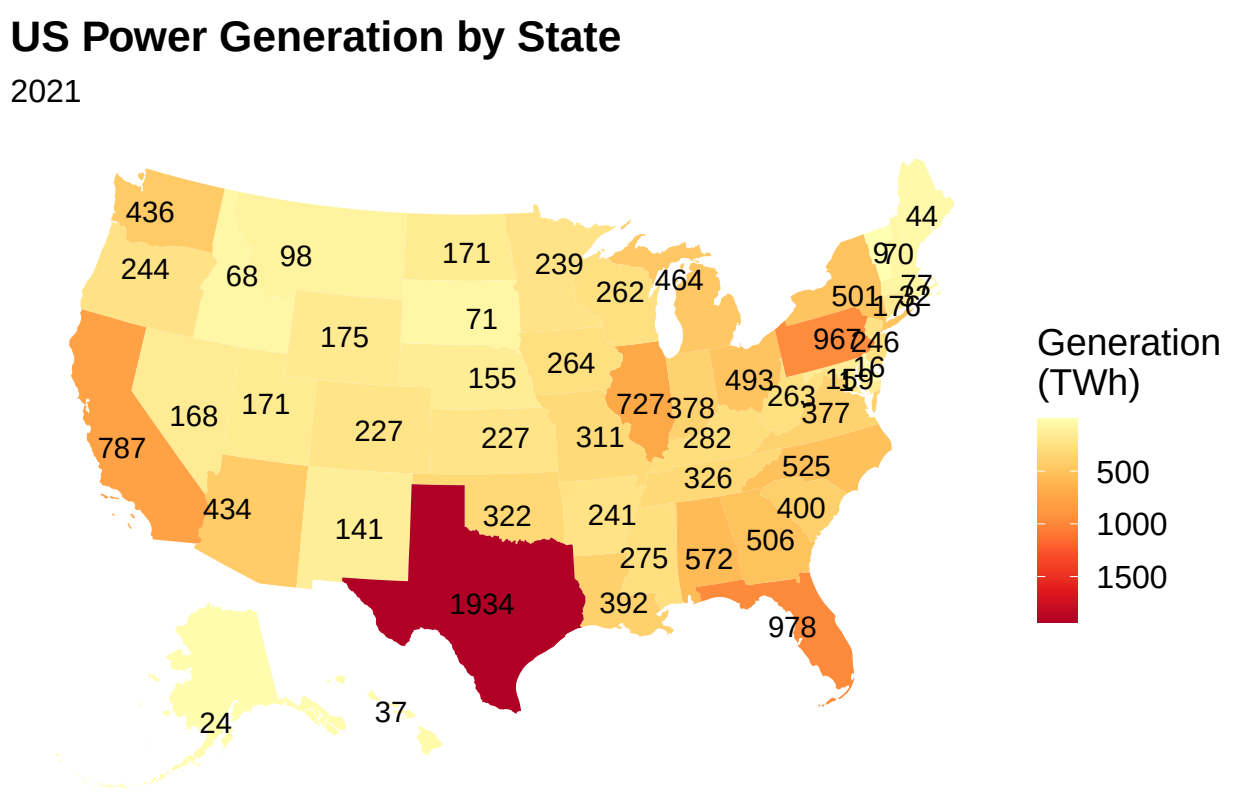
State	Energy Source	Tons NOx/MWh
US	Total,Natural Gas	0.0002687
US	Total,Coal	0.0006836
US	Total,Petroleum	0.0016358
US	Total,Wood and Wood Derived Fuels	0.0018871
US	Total,Other Gases	0.0019598
US	Total,Other	0.0027283
US	Total,Other Biomass	0.0060627

As can be seen by these three tables, many different energy sources are more and less efficient for the different types of emissions. As mentioned before, some of the Green Energy sources only generate one or two types of emissions, meaning they won't appear on each table. As we can see from the tables, Natural Gas is the most efficient Brown Energy source that we have, as it has the least amount of SO2 and NOx emissions per MWh, and is only behind Geothermal energy on the CO2 table. Geothermal is considered to be a nonrenewable source of energy, although it does make emissions, which makes Natural Gas the most efficient Brown Energy on all three charts. When referencing the line graph under Question 1, we see that Coal and Petroleum had the most Energy Producers within the US, but when comparing them to Natural Gas, we can see that Natural Gas is anywhere from twice as efficient (Tons CO2/MWh) to 20 times as efficient (Tons SO2/MWh) as its main competitors.

Question 3: Which US States generate the most energy? This question was relatively straightforward, as I wrangled one of our datasets to group the MWh generated by state in 2021 (most recent year with complete data). I put the data into a Choropleth Map, which can be seen below.

```
# Centroid for putting total generation on choropleth
centroid <- aggregate(data = state_gen_map, cbind(x, y) ~ gen, FUN = mean)
```

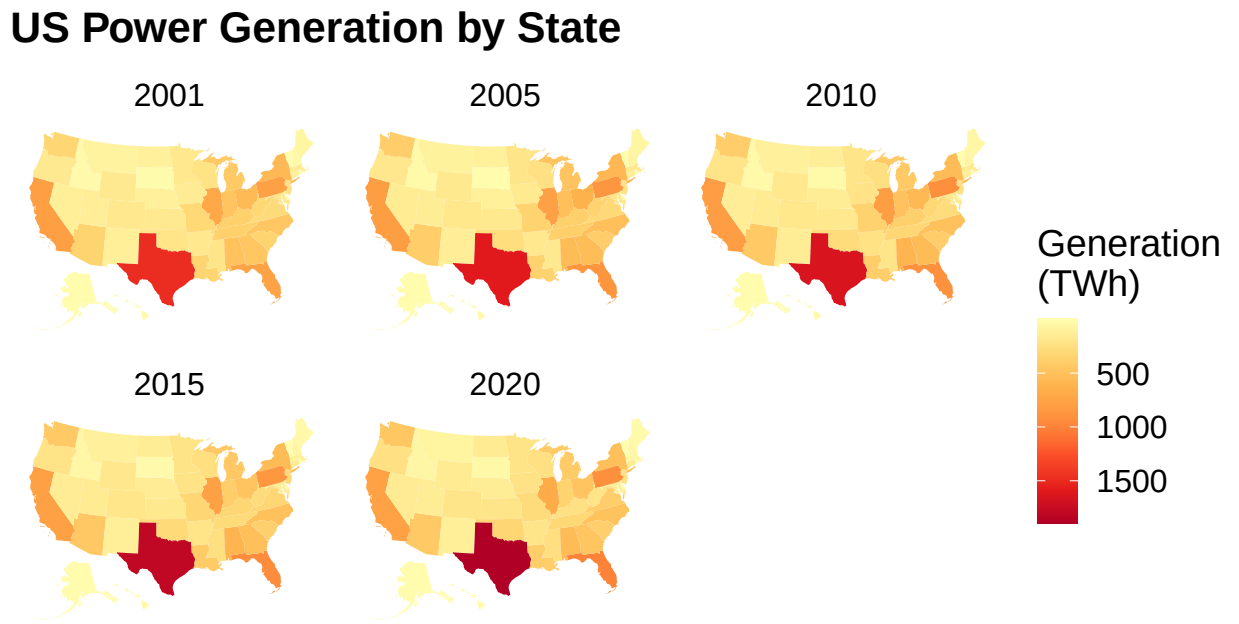
```
# Choropleth of US as heat map of generation
state_gen_map %>%
  ggplot(mapping = aes(x = x, y = y, fill = gen)) +
  geom_polygon(mapping = aes(group = group)) +
  geom_text(data = centroid, mapping = aes(x = x, y = y, label =sprintf("%0.0f", gen))) +
  scale_fill_distiller(palette = "YlOrRd", trans = "reverse") +
  coord_equal() +
  theme_map() +
  labels(title = "US Power Generation by State", subtitle = "2021",
       x = element_blank(), y = element_blank(), fill = "Generation\n(TWh)")
```



Looking at the map, the answer to the question becomes obvious. Texas, by almost double, generates the most power in the entire US. The EIA reports that most of this huge power draw is based off of their industrial sector, accounting for over half of their total power generation. This begs the question, has Texas always been the most power hungry state? Going back to 2001, and jumping 5 years each time, the following choropleths are generated.

```
# heatmap of generation wrapped by year
gap5yr_map %>%
  ggplot(mapping = aes(x = x, y = y, fill = gen)) +
  geom_polygon(mapping = aes(group = group)) +
  scale_fill_distiller(palette = "YlOrRd", trans = "reverse") +
  coord_equal() +
  theme_map() +
  labs(title = "US Power Generation by State",
```

```
      x = element_blank(), y = element_blank(), fill = "Generation\n(TWh)") +
facet_wrap(~YEAR)
```



After looking at these choropleths, we can see that Texas has been far ahead of every state in terms of power generation for the last 20 years, and has only been increasing the amount of power generated in each subsequent year. Looking at the map, most states seem to be generating more power as we progress through the 21st Century, but the general ranking of highest producers seems to be consistent, with Texas at the top, and California, Florida, Illinois, and Pennsylvania being right behind. The top of the Northeast (above New York) seems to generate the least amount of power per state, with all of the states falling behind the Southeast and Midwest, but the West, save california, Arizona, and Washington doesn't seem to be much higher.

Conclusion

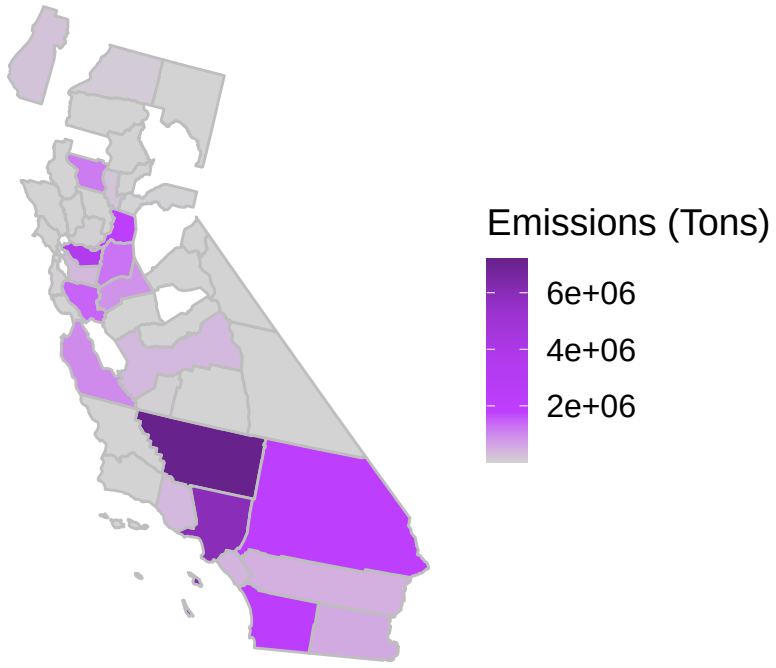
Overall, we can see a few key things. First, the US seems to be building more and more Green Energy sources, but still isn't replacing a large amount of Brown Energy sources, instead supplementing them with Green Energy. Second, we can see that every Green Energy source is infinitely more efficient than any Brown Energy source, but of the different types of Brown Energy, Natural Gas causes far and away the least amount of emissions per MWh. Third, and finally, we see that Texas is the most power-hungry state in the US, and has been for the entire 21st century. We also saw that power generation has been increasing across the board in the last two decades, but all states seem to be increasing at a similar rate. Moving forward, we should look to push more and more Green Energy sources to the US so that we can start to decrease the amount of Brown Energy producers, which would also cut emissions. If we can't push Green Energy everywhere though, we should make sure to recommend Natural Gas as the next best option to reduce the amount of emissions in the US and then save our planet for generations to come.

Part 2

Introduction As a whole, my team is looking at various trends and relationships with regards to supplying electric power in the United States. Rather than looking at the U.S. in general, my part of the project is about zooming into one of the largest states, California, and looking at the emissions produced in the pursuit of supplying electricity. First, I will visualize the emissions produced by each county. Using this wrangled data, I will then examine the relationship between a county's change in population and its quantity of emissions. Finally, I will compare the Demographics of various counties based on how bad their emissions are in order to determine if any groups are disproportionately suffering the effects of emissions.

Question 1: What is the distribution of emissions due to the supply of electric power across California counties? We will begin by creating a heat map of the emissions of each county across California. However, we should be specific about what type of emissions we care about. I am interested in the emissions as related to the supply of electric power. While power plants certainly play a hand in this, they are not the only source of such emissions. As such, I have used a data set which contains site locations, their emissions, and a 'NAICS Code' among a few other variables. Most notably, this NAICS Code will be how we identify locations related to supplying electric power. Without going into too much detail, a NAICS code is self assigned by a company based on what industry they are a part of. Each code is built using sets of digits where more digits are added when more specificity is required. This allows us to be as general or specific as we like in terms of which industry(s) and sector(s) we target. For my purposes, I am using code '2211' which includes all "Electric Power Generation, Transmission, and Distribution" company sites. This allows me to capture the full breadth of emissions from sites in California related to supplying electric power. Here is the choropleth of such site's emissions in California:

```
#Get total emissions in a County
caEmissionsByCounty <- NEI2017 %>%
  filter(state=="CA",`pollutant type(s)`!="nan", str_starts(as.character(.$`naics code`),"2211")) %>%
  select(`fips code`,county,`pollutant type(s)`,`total emissions`,`emissions uom`,`naics code`) %>%
  mutate(`total emissions`=case_when(as.character(`emissions uom`)=="LB" ~ `total emissions`*.0005,
                                     TRUE ~ `total emissions`),
         `fips code` = paste0('0',as.character(`fips code`))) %>%
  select(-`emissions uom`) %>%
  group_by(county,`fips code`) %>%
  summarise(`county_total_emissions`=sum(`total emissions`)) %>%
  rename(id = `fips code`)

#Join with map data
counties <- left_join(county_map,caEmissionsByCounty,by="id") %>% na.omit()

#Plot Choropleth of California as a Heat map of Emissions
counties %>%
  ggplot(mapping = aes(x=long,y=lat,group=group,fill=county_total_emissions)) +
  geom_polygon(color="gray") +
  coord_equal() +
  theme_map() +
  scale_fill_gradientn(colors=c("lightgray","darkorchid1","darkorchid2","darkorchid3","darkorchid4"),
                       name="Emissions (Tons)") +
  theme(legend.position = "right") +
  labs(title="Site Emissions in California Counties", subtitle = "Electricity Power Generation, Transmis
```

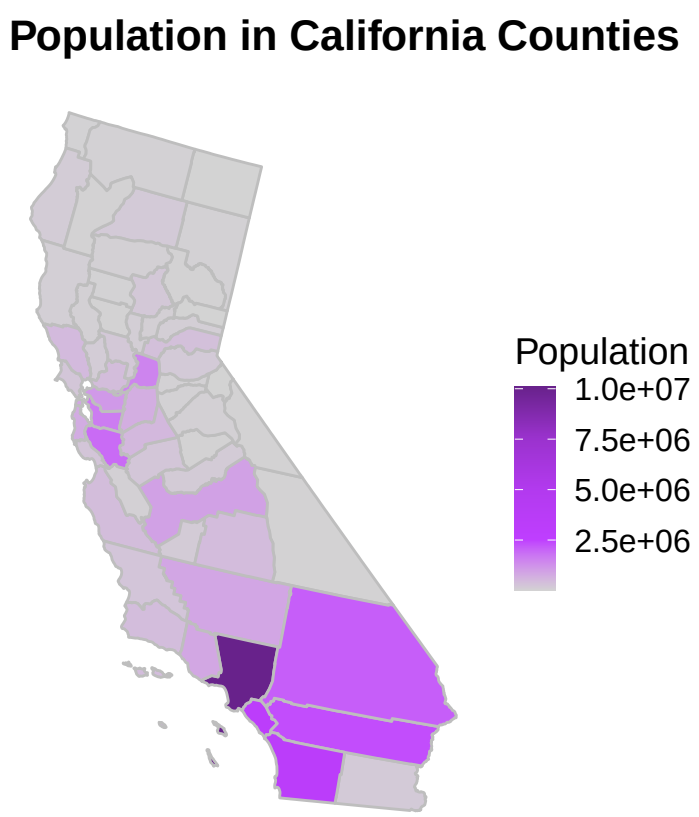


The emissions in the state seem to be congregated around two nodes. The northern node includes the Eastern Bay Area and some of the northern Central Counties while the southern node is generally encompassed by Southern Counties, most notably Kern and Los Angeles. Other than those clusters, the remaining counties have much fewer emissions with a number not even appearing as they lack emissions entirely.

Based on a hunch and global trends, let's take a look at the heatmap of the population in each of California's counties:

```
#Get total population in a County
CaCountyPopulation <- acs2017 %>%
  filter(State == "California") %>%
  rename(id=CountyId) %>%
  select(-State) %>%
  select(id, TotalPop, County) %>%
  mutate(County = str_remove(County," County"),
         id = paste0('0',as.character(id))) %>%
  rename(county = County)

#Join with map data
countiesPop <- left_join(county_map,CaCountyPopulation,by="id") %>% na.omit()

#Plot Choropleth of California as a Heat map of Population
countiesPop %>%
  ggplot(mapping = aes(x=long,y=lat,group=group,fill=TotalPop)) +
  geom_polygon(color="gray") +
  coord_equal() +
  theme_map() +
```

```
  scale_fill_gradientn(colors=c("lightgray","darkorchid1","darkorchid2","darkorchid3","darkorchid4"),
                       name="Population") +
  theme(legend.position = "right") +
  labs(title="Population in California Counties")
```

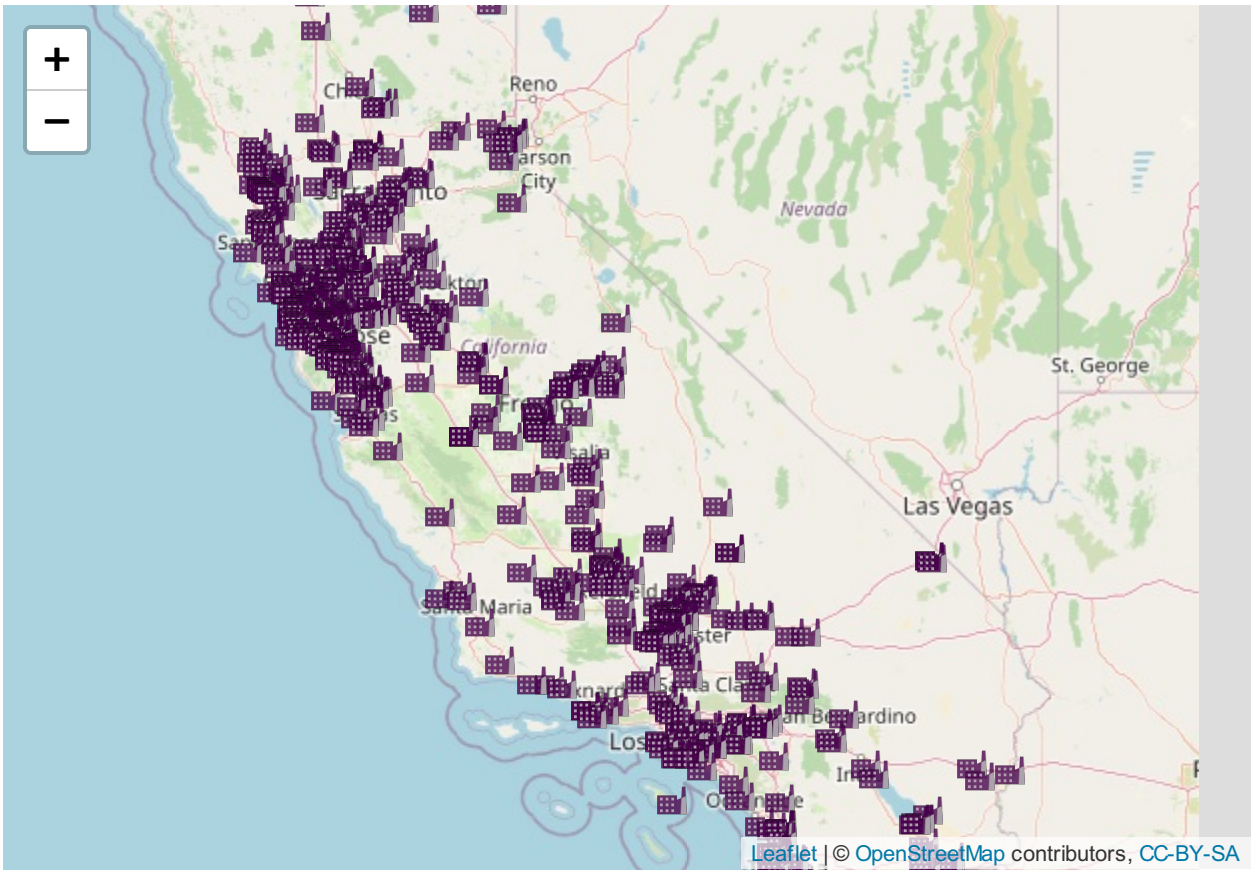


Indeed, it's relatively close to what one might suspect. The two nodes seem similar with one near the Bay area and one Near Los Angeles. However, the actual relative magnitude of the particular counties does look different. For example, when we were looking at the emissions, Kern was the darkest shaded county in the southern node. However, in creating a heat map of population, we see that Los Angeles has the most individuals within its borders. This leads us to conclude that there is certainly a positive association between emissions and raw population but not 1:1 correlation.

Related to this question, though mostly pursued out of curiosity, I was interested in seeing if there was a noticeable clustering of sites in these two nodes or if the emissions were perhaps being produced by a few, very substantial sites. In order to visualize this question, I created the following Leaflet:

```
#Create Leaflet DataFrame of locations of sites related to supplying electricity
energySiteLocationsCA <- NEI2017 %>%
  filter(state=="CA",`pollutant type(s)`!="nan", str_starts(as.character(.$`naics code`),"2211")) %>%
  select(`site name`,`naics description`,`site latitude`, `site longitude`,
         address,city,`zip code`,`postal abbreviation`)%>%
  distinct()

#Create Leaflet Labels using site information
energySiteLabels <- sprintf("<b>%s</b></br>
                            %s</br>
```

```
                          %s</br>
                          %s, %s %s",
                  energySiteLocationsCA$`site name`,
                  energySiteLocationsCA$`naics description`,
                  energySiteLocationsCA$address,
                  energySiteLocationsCA$city,
                  energySiteLocationsCA$`postal abbreviation`,
                  energySiteLocationsCA$`zip code`) %>%
  lapply(htmltools::HTML)

#Create Leaflet Map
mapStates = map("state", fill = TRUE, plot = FALSE)

#Create Custom Icon
greenLeafIcon <- makeIcon(
  iconUrl = "img/factoryIcon2.png",
  iconWidth = 17, iconHeight = 17)

#Create Leaflet
energySiteLocationsCA %>%
  leaflet(options=leafletOptions(zoomSnap=0.1)) %>%
  setView(lng=-120.5049115302131, lat=37.53869072160772, zoom=5.8) %>%
  addTiles() %>%
  addMarkers(~`site longitude`, ~`site latitude`, popup=energySiteLabels,label=energySiteLabels,icon = g
```



As we can see, there does seem to be many sites clustered in these regions. It is not just the act of a few

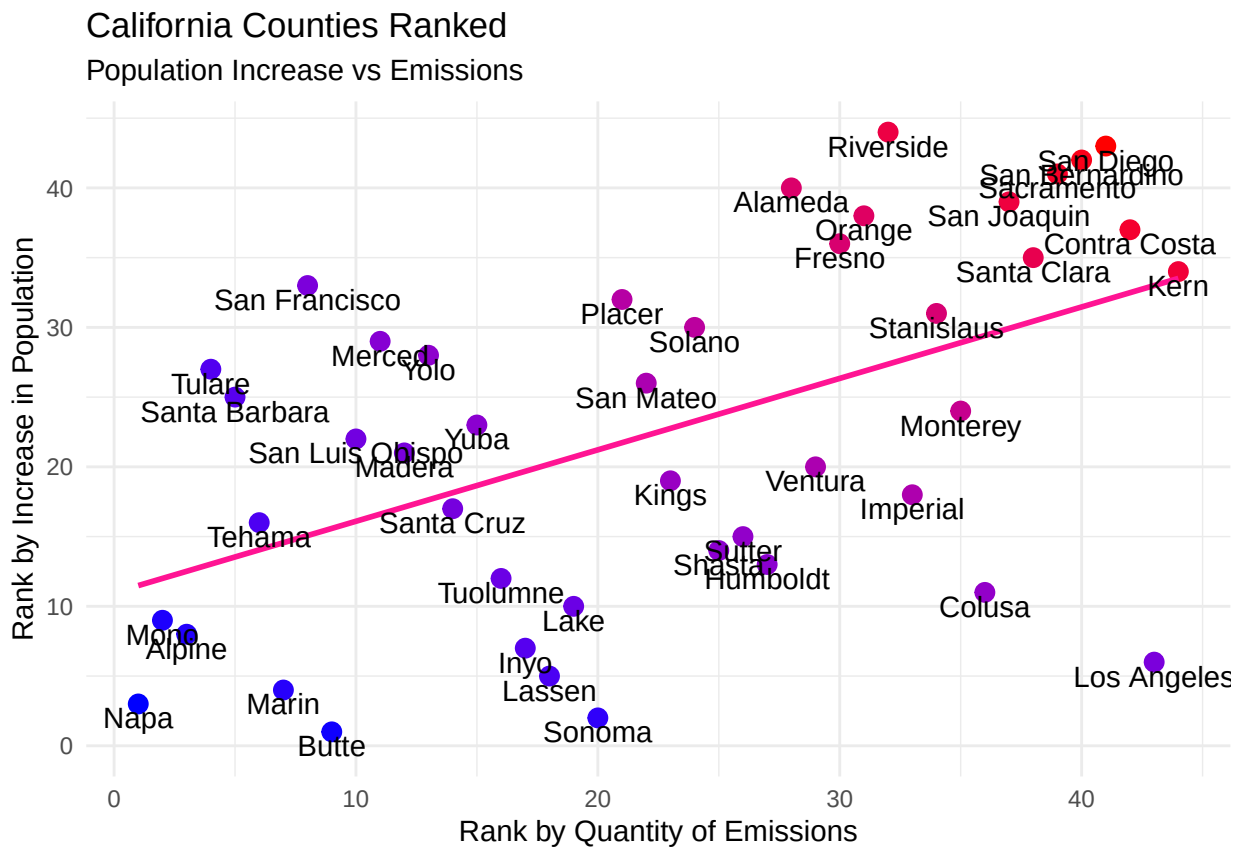
very bad emitters that cause this concentration of emissions. You can even hover over each location to see that they are individually named and independent sites.

Question 2: What is the trend between population change and a county's emissions? I was interested in determining whether counties with large amounts of emissions had a relatively faster decreasing population. The detail about 'relatively faster' is important as I am not looking at if their population decreased **at all**. No, instead, I want to know whether it is decreasing (or perhaps increasing) faster than other counties in California. We need to properly make this comparison as otherwise we might correlate a statewide decrease in population with emissions when in fact there are a host of economic reasons California has had for in increasing number of citizens emigrating from the state. Ultimately, My hypothesis was that counties with more emissions would have a population increasing more slowly or entirely decreasing when compared to those with less emissions.

In order to determine this, I got each county's rank in terms of quantity of emissions and in terms of average population change over the past five years. I then plotted these two ranks and drew a best fit line to see the direction of correlation:

```
#Create a Dataframe of the average Population Change in each California County for the past 5 years
caCountyPopulationChange <- populationData %>%
  filter(STNAME=="California",COUNTY!= "000") %>%
  select(STATE,CTYNAME,COUNTY,NPOPCHG_2015,NPOPCHG_2016,NPOPCHG_2017,NPOPCHG_2018,NPOPCHG_2019) %>%
  mutate(fips = paste0(.$STATE,.$COUNTY),
         fiveYearAvgPopChange = (NPOPCHG_2015+NPOPCHG_2016+NPOPCHG_2017+NPOPCHG_2018+NPOPCHG_2019)/5,
         CTYNAME=str_remove(CTYNAME," County")) %>%
  select(fips, fiveYearAvgPopChange,CTYNAME) %>%
  rename(id = fips,name=CTYNAME)

#Join the Population Change dataframe with the County Emissions dataframe from question 1
emissionsPopChangeByCounty <- left_join(caCountyPopulationChange, caEmissionsByCounty, by="id") %>%
  select(-county) %>%
  na.omit()

#Sort by population change
emissionsPopChangeByCounty <- emissionsPopChangeByCounty %>%
  arrange(fiveYearAvgPopChange)

#Add rank in terms of population change
emissionsPopChangeByCounty <- emissionsPopChangeByCounty %>% add_column(rankOfPopChange = NA)
for (i in 1:length(emissionsPopChangeByCounty$fiveYearAvgPopChange)){
  emissionsPopChangeByCounty$rankOfPopChange[i] = i
}

#Sort by emissions
emissionsPopChangeByCounty <- emissionsPopChangeByCounty %>%
  arrange(county_total_emissions)

#Add rank in terms of emissions
emissionsPopChangeByCounty <- emissionsPopChangeByCounty %>% add_column(rankOfEmissions = NA)
for (i in 1:length(emissionsPopChangeByCounty$county_total_emissions)){
  emissionsPopChangeByCounty$rankOfEmissions[i] = i
}

#Plot the Rank of a County in terms of it's Population Change vs its Rank in terms of Quantity of Emiss
emissionsPopChangeByCounty %>%
```

```
ggplot() +
geom_point(aes(x=rankOfEmissions,y=rankOfPopChange,
               color=rankOfPopChange*rankOfEmissions),size=3) +
geom_smooth(aes(x=rankOfEmissions,y=rankOfPopChange), method = "lm", se = FALSE, color="deeppink") +
scale_colour_gradient(low="blue",high="red") +
theme_minimal() +
annotate('text', emissionsPopChangeByCounty$rankOfEmissions, emissionsPopChangeByCounty$rankOfPopChang
xlab(label="Rank by Quantity of Emissions") +
ylab(label="Rank by Increase in Population") +
labs(title="California Counties Ranked", subtitle = "Population Increase vs Emissions") +
guides(color="none",size="none")
```



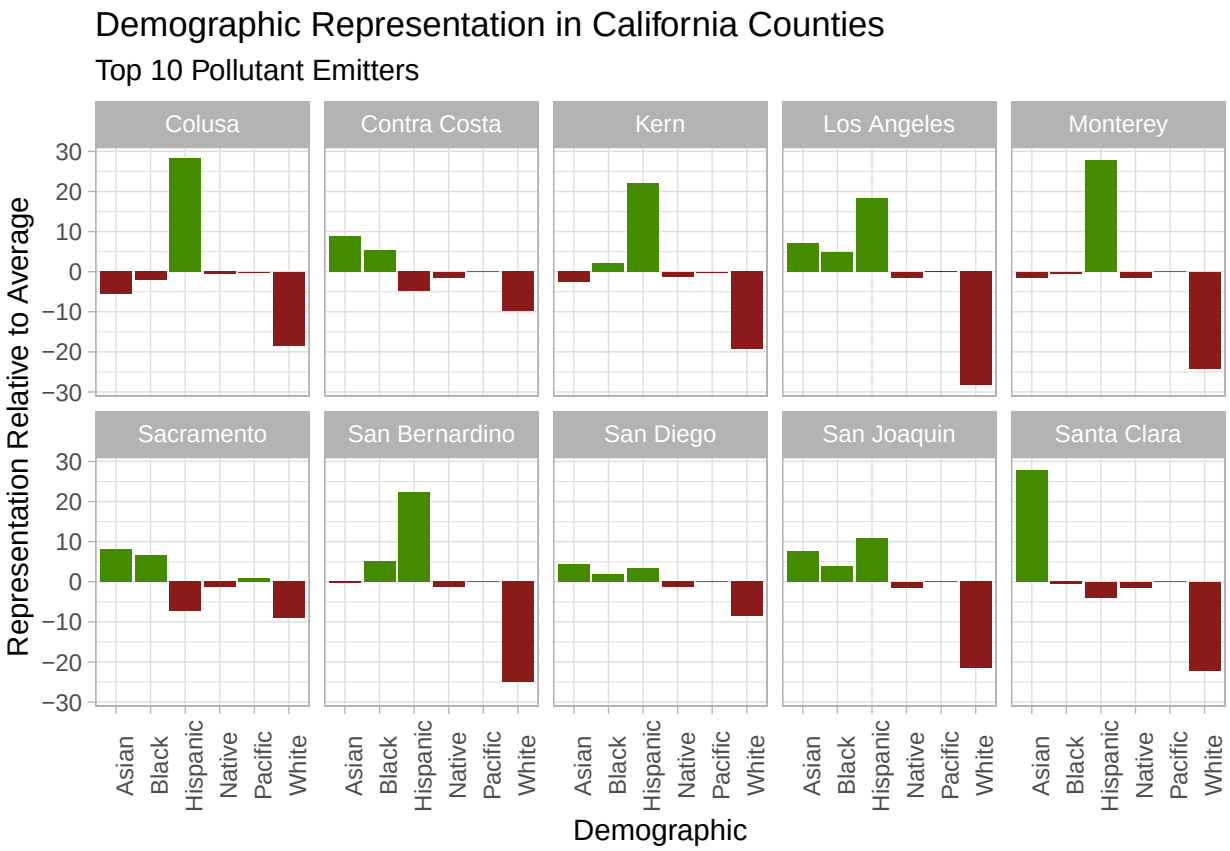
As evident by the chart, it turns out that the greater emissions you have, the more positive your population change is. This makes sense as it is a normal trend of population clusters around the world. Simply, if you have more people coming in, you'll need more energy for them and thus will create more emissions when supplying that energy. It seems that the pollution in and around these regions is not bad enough yet to convince individuals to move elsewhere.

Question 3: What are the demographics in the counties with the most emissions? Finally, I was interested in determining if the emissions from power generation were disproportionately affecting underrepresented demographic groups. In order to determine this, I started by making a list of the top 10 counties which produce the most emissions. I also got data about the demographics of all the counties in California. Using this demographic information, I found the average percent representation of each demographic across all of California. This "average percent" value is important as I used it to find the percentage point difference between a demographic across all of California and its specific representation in

each of the top 10 emitting counties. Rephrasing this statement for clarification; I found the difference of a demographic's percent representation statewide and its representation in the top 10 emitting counties. I was then able to create a column chart of this "difference from Average Representation" value and each of the Demographics, faceting it by county:

```
#Get the top 10 counties with the worst emissions
top10EmittingCounties <- caEmissionsByCounty %>%
  ungroup() %>%
  arrange(desc(county_total_emissions)) %>%
  mutate(id=as.double(id)) %>%
  select(-county) %>%
  head(10)

#Get the Demographics of California Counties
CaCountyDemographics <- acs2017 %>%
  filter(State == "California") %>%
  rename(id=CountyId) %>%
  select(-State) %>%
  select(id, Hispanic, White, Black, Native, Asian, Pacific, County) %>%
  mutate(County = str_remove(County," County")) %>%
  rename(county = County)

#Get the average percentage of each racial demographic in California
meanHispanic <- mean(CaCountyDemographics$Hispanic)
meanWhite <- mean(CaCountyDemographics$White)
meanBlack <- mean(CaCountyDemographics$Black)
meanNative <- mean(CaCountyDemographics$Native)
meanAsian <- mean(CaCountyDemographics$Asian)
meanPacific <- mean(CaCountyDemographics$Pacific)

#Get demographics of top10EmittingCounties by joining dataframes,
#and then Get each Demographic's percent difference from the average in each county
top10EmittingCountiesDemos <- left_join(top10EmittingCounties, CaCountyDemographics) %>%
  pivot_longer(Hispanic:Pacific,names_to = "demographic",values_to = "percent") %>%
  mutate(diffFromAvgRepresentation = case_when(demographic == "Hispanic" ~ percent - meanHispanic,
                                               demographic == "White" ~ percent - meanWhite,
                                               demographic == "Black" ~ percent - meanBlack,
                                               demographic == "Native" ~ percent - meanNative,
                                               demographic == "Asian" ~ percent - meanAsian,
                                               demographic == "Pacific" ~ percent - meanPacific),
         diffFromAvgRepresentation = round(diffFromAvgRepresentation,2),
         sign = sign(diffFromAvgRepresentation))

#Create a Column chart of the difference of each demographic's percent difference from the average
#in the counties with the worst emissions
top10EmittingCountiesDemos %>%
  ggplot(aes(x=demographic,y=diffFromAvgRepresentation,fill=factor(sign))) +
  geom_col() +
  theme_light() +
  theme(axis.text.x = element_text(angle = 90)) +
  xlab(label="Demographic") +
  ylab(label="Representation Relative to Average") +
  labs(title="Demographic Representation in California Counties",
       subtitle = "Top 10 Pollutant Emitters",
```

```
      fill="Demographic") +
  scale_fill_manual(values = c("firebrick4", "chartreuse4")) +
  guides(fill="none") +
  facet_wrap(~ county, nrow = 2)
```

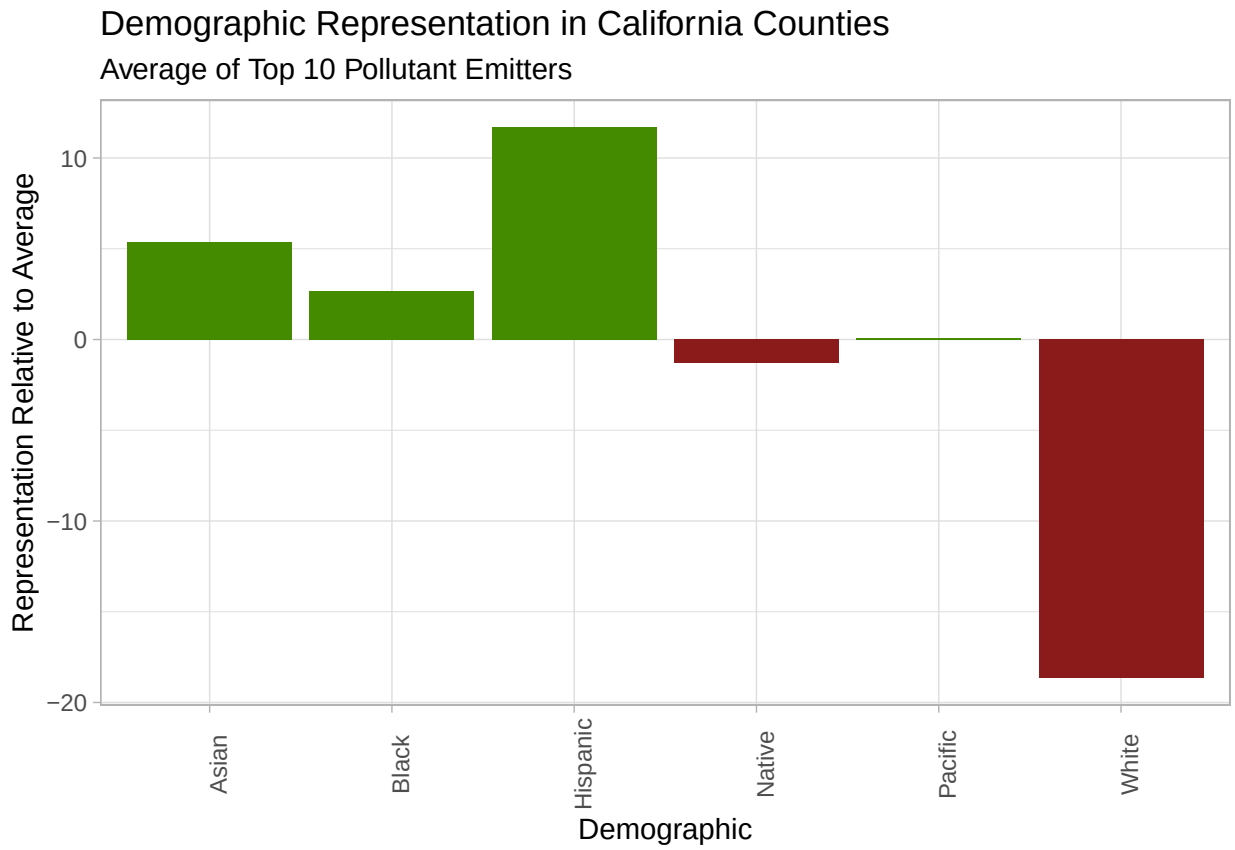


What we see above is that any time a Demographic's percent representation in a county is less than the state average, the value is negative and its bar is colored red. Of course, when its percent average in that county is greater than the state average, it is positive and colored green. Immediately, what stands out is that the white Demographic in the counties with the worst emissions always has a smaller representation than its state average. The Native and Pacific Demographics are already so small such that their relative difference from the mean is very minor, but it appears that they are less than their mean state representation as well. The remaining three demographics in this list do *on occasion* score below their state average, but they tend to stay very near it when they do. Generally though, I think it is fair to say that the Asian, Black, and Hispanic demographics are disproportionately represented in the counties with the worst emissions. In fact, the few times one of those three crosses into the red, either one or both of the other two are excessively over-represented in that county.

Out of curiosity about confirming the above mentioned trends, I took the average of each Demographics representation in these 10 counties and found the difference from the state wide mean in order to produce the following chart:

```
#Get average Demographic representation of the top 10 emitting counties
top10EmittingCountiesDemosAverage <- top10EmittingCountiesDemos %>%
  group_by(demographic) %>%
  summarize(avgDemoRep = mean(diffFromAvgRepresentation)) %>%
  mutate(sign = sign(avgDemoRep))
```

```r
#Graph the demographics of the above average
top10EmittingCountiesDemosAverage %>%
  ggplot(aes(x=demographic,y=avgDemoRep,fill=factor(sign))) +
  geom_col() +
  theme_light() +
  theme(axis.text.x = element_text(angle = 90)) +
  xlab(label="Demographic") +
  ylab(label="Representation Relative to Average") +
  labs(title="Demographic Representation in California Counties",
       subtitle = "Average of Top 10 Pollutant Emitters",
       fill="Demographic") +
  scale_fill_manual(values = c("firebrick4", "chartreuse4")) +
  guides(fill="none")
```

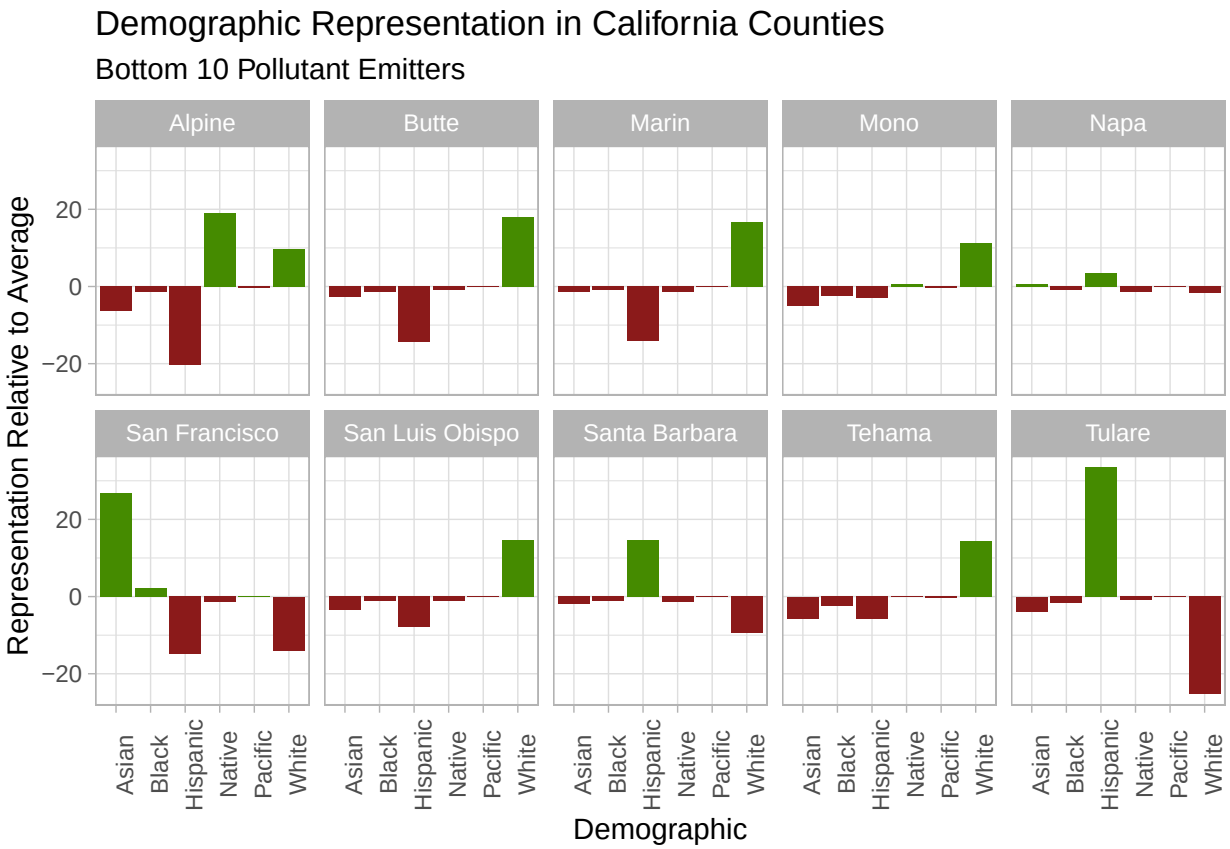


Though it is a much smaller, simpler visualization, I think this bar chart makes the truth self-evident. Demographics which are underrepresented statewide have to suffer the worst emission pollution across California's counties.

Seeing the relative Demographic representation in each of the top 10 worst polluting counties, I was curious about the same distribution in the 10 counties which pollute the least:

```r
#Get the 10 Counties with the least emissions and their demographics
bottom10EmittingCounties <- caEmissionsByCounty %>%
  ungroup() %>%
  arrange(county_total_emissions) %>%
  mutate(id=as.double(id)) %>%
```

```r
  select(-county) %>%
  head(10)

#Get demographics of bottom10EmittingCounties
bottom10EmittingCountiesDemos <- left_join(bottom10EmittingCounties, CaCountyDemographics) %>%
  pivot_longer(Hispanic:Pacific,names_to = "demographic",values_to = "percent") %>%
  mutate(diffFromAvgRepresentation = case_when(demographic == "Hispanic" ~ percent - meanHispanic,
                                               demographic == "White" ~ percent - meanWhite,
                                               demographic == "Black" ~ percent - meanBlack,
                                               demographic == "Native" ~ percent - meanNative,
                                               demographic == "Asian" ~ percent - meanAsian,
                                               demographic == "Pacific" ~ percent - meanPacific),
         diffFromAvgRepresentation = round(diffFromAvgRepresentation,2),
         sign = sign(diffFromAvgRepresentation))

#Graph the demographics of the 10 least emitting counties
bottom10EmittingCountiesDemos %>%
  ggplot(aes(x=demographic,y=diffFromAvgRepresentation,fill=factor(sign))) +
  geom_col() +
  theme_light() +
  theme(axis.text.x = element_text(angle = 90)) +
  xlab(label="Demographic") +
  ylab(label="Representation Relative to Average") +
  labs(title="Demographic Representation in California Counties",
       subtitle = "Bottom 10 Pollutant Emitters",
       fill="Demographic") +
  scale_fill_manual(values = c("firebrick4", "chartreuse4")) +
  guides(fill="none") +
  facet_wrap(~ county, nrow = 2)
```

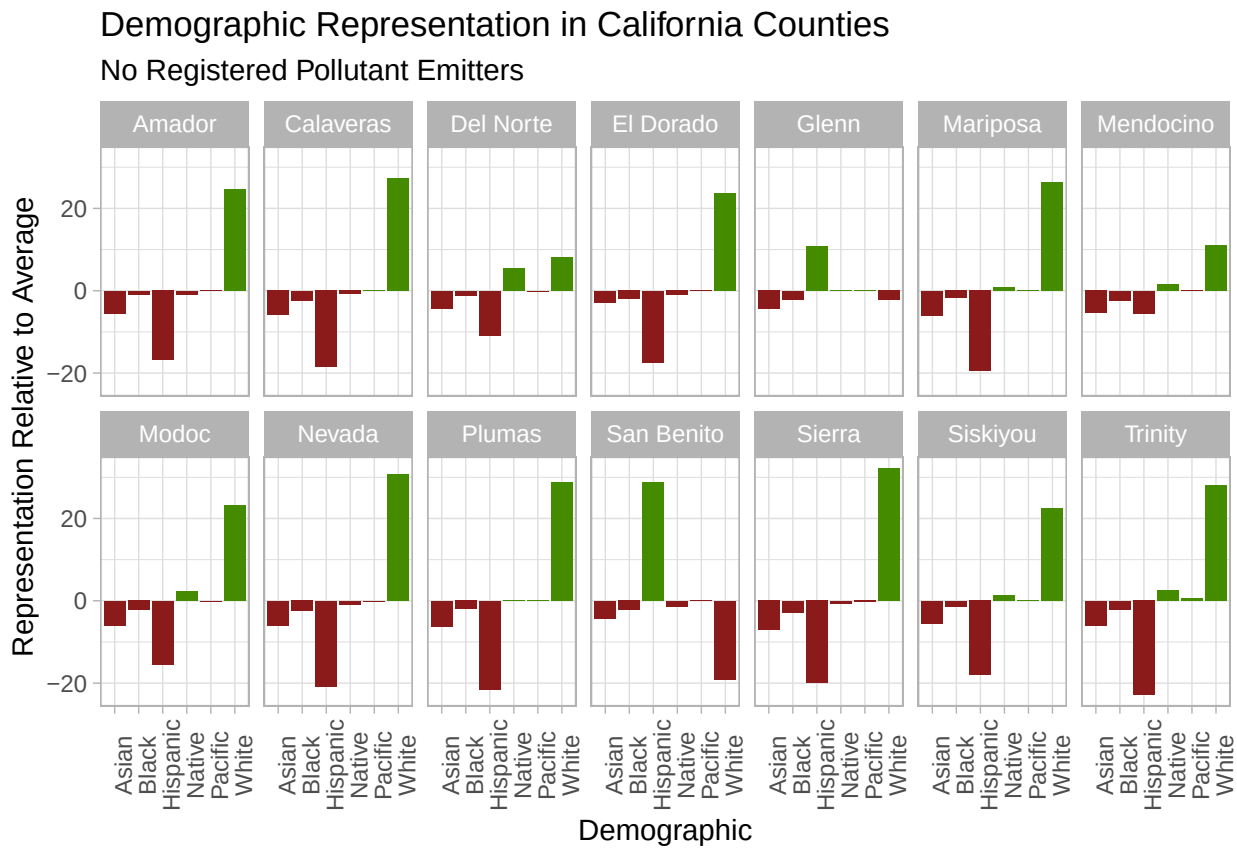


At first glance, this chart seems nearly the inverse of the previous, but personally I was surprised that there were still such large negative measurements for the white demographic. After all, it can't be that every county is less than the state average, right? Right. It turns out that the above visualization is of the counties with the least emission *as long as they still had emissions.*

Instead, lets take a look a the counties with truly the fewest emissions. Those with absolutely none:

```
#We can graph Demographics of counties with NO Emissions

#Get all FIPS in California
caFips <- countyFIPS  %>%
  filter(STATE=="CA") %>%
  select(STCOUNTYFP) %>%
  rename(id = STCOUNTYFP) %>%
  distinct() %>%
  pull(id) %>%
  as.integer()

#Get FIPS of Counties that have Emissions for Energy
caCountyWithEnergyEmissionsFips <- pull(caEmissionsByCounty,id) %>%
  as.integer()

#Get FIPS of Counties that don't have Emissions for Energy
caCountyNoEnergyEmissionsFips <- data.frame(id = setdiff(caFips,caCountyWithEnergyEmissionsFips))


#Get demographics of counties that don't have Emissions for Energy
```

```r
noEmissionsCounties <- left_join(caCountyNoEnergyEmissionsFips, CaCountyDemographics,by="id") %>%
  pivot_longer(Hispanic:Pacific,names_to = "demographic",values_to = "percent") %>%
  mutate(diffFromAvgRepresentation = case_when(demographic == "Hispanic" ~ percent - meanHispanic,
                                               demographic == "White" ~ percent - meanWhite,
                                               demographic == "Black" ~ percent - meanBlack,
                                               demographic == "Native" ~ percent - meanNative,
                                               demographic == "Asian" ~ percent - meanAsian,
                                               demographic == "Pacific" ~ percent - meanPacific),
         diffFromAvgRepresentation = round(diffFromAvgRepresentation,2),
         sign = sign(diffFromAvgRepresentation))

#Generate Barchart of Demographics in Counties with No Emissions
noEmissionsCounties %>%
  ggplot(aes(x=demographic,y=diffFromAvgRepresentation,fill=factor(sign))) +
  geom_col() +
  theme_light() +
  theme(axis.text.x = element_text(angle = 90)) +
  xlab(label="Demographic") +
  ylab(label="Representation Relative to Average") +
  labs(title="Demographic Representation in California Counties",
       subtitle = "No Registered Pollutant Emitters",
       fill="Demographic") +
  scale_fill_manual(values = c("firebrick4", "chartreuse4")) +
  guides(fill="none") +
  facet_wrap(~ county, nrow = 2)
```



County	Asian	Black	Hispanic	Native	Pacific	White
Colusa	1.5	0.9	58.4	1.1	0.1	36.3
Contra Costa	15.8	8.3	25.3	0.2	0.5	44.9
Kern	4.5	5.1	52.2	0.5	0.1	35.4
Los Angeles	14.3	7.9	48.4	0.2	0.2	26.5
Monterey	5.6	2.5	57.9	0.2	0.5	30.6
Sacramento	15.1	9.5	22.8	0.4	1.1	45.7
San Bernardino	6.7	8.0	52.3	0.3	0.3	29.8
San Diego	11.5	4.7	33.4	0.4	0.4	46.2
San Joaquin	14.8	6.7	40.8	0.2	0.5	33.2
Santa Clara	34.9	2.4	26.1	0.2	0.3	32.6

Now we can really see the truth of the matter. Every single county (baring 2 of 14) which don't produce emissions in California is dominated by white representation. The only counties that are not dominate by white representation are dominated by the Hispanic demographic, and this is likely as it is a cultural hub for this group. In fact, other than San Benito and Glenn (the counties dominated by the Hispanic demographic), the three underrepresented groups of Asian, Black, and Hispanic are all entirely in the red. If it wasn't evident before, it should be clear now that certain demographics suffer the pollution of emissions far worse than others.

Finally, as we've been look at the percent difference from the state mean this whole time, I was curious about seeing the raw percent representation of each demographic across these two extremes. Below I have created two tables. The first is of the top 10 counties with the most emissions, and the second is of those counties with no emissions:

```
#Get Data pf Demographics for Counties with the Most Emissions
top10EmittingCountiesDemosTable <- left_join(top10EmittingCounties, CaCountyDemographics) %>%
  select(county, Asian, Black, Hispanic, Native, Pacific, White) %>%
  arrange(county) %>%
  rename(County = county)

#Get Data of Demographics for Counties with No Emissions
noEmissionsCountiesDemosTable <- left_join(caCountyNoEnergyEmissionsFips, CaCountyDemographics) %>%
  select(county, Asian, Black, Hispanic, Native, Pacific, White) %>%
  arrange(county) %>%
  rename(County = county)

#Show Table of Top 10 Emitters
top10EmittingCountiesDemosTable %>% kbl(align="c") %>%
  kable_styling(bootstrap_options="striped",
                full_width=FALSE,
                position="center") %>%
  row_spec(0,color="white",background="#2f4f4f") %>%
  row_spec(1:10,color="#2f4f4f")
```

```
#Show Table of No Emitters
noEmissionsCountiesDemosTable %>%
  kbl(align="c") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "responsive"),
                full_width=FALSE,position="center") %>%
  row_spec(0,color="white",background="#2f4f4f") %>%
  row_spec(1:14, color="#2f4f4f")
```

County	Asian	Black	Hispanic	Native	Pacific	White
Amador	1.5	2.0	13.2	0.7	0.2	79.3
Calaveras	1.2	0.6	11.5	0.9	0.5	82.0
Del Norte	2.7	1.8	19.2	7.2	0.1	62.8
El Dorado	4.2	0.9	12.6	0.6	0.2	78.5
Glenn	2.6	0.8	40.8	1.7	0.3	52.5
Mariposa	1.0	1.2	10.6	2.4	0.3	81.0
Mendocino	1.7	0.6	24.5	3.3	0.2	65.9
Modoc	1.1	0.7	14.6	4.0	0.1	78.0
Nevada	1.1	0.5	9.2	0.7	0.1	85.4
Plumas	0.8	0.9	8.5	1.8	0.3	83.5
San Benito	2.7	0.7	58.9	0.3	0.2	35.6
Sierra	0.0	0.1	10.1	1.0	0.0	87.0
Siskiyou	1.6	1.5	12.0	2.9	0.3	77.2
Trinity	1.1	0.8	7.2	4.3	0.9	82.8

Even without getting into the nuance of difference from the state mean, we can at a glance come to similar conclusions as before. I would even argue that seeing the magnitude of each county's demographic breakdown makes the situation seem even more self-evident.

Conclusion Overall, the conclusions we can come to about California's emissions in relation to electricity supply are relatively standard and unfortunately too common. It produces more emissions near it largest population centers: Los Angeles and the Bay Area. As the population of a region changes more positively, it tends to have greater emissions. Finally, demographics which are underrepresented statewide must suffer the worst of the pollution from emissions. Though it would be very challenging and there are many practical difficulties, it may one day be useful to consider decoupling these trends. California is just one state, but these trends have been spoken about as a worldwide phenomenon. Eventually, populations will suffer too much from pollution. We will need to find a way to limit or remove the contaminates from our largest centers of civilization if we hope to for a better future.

Section 3 - Jeremy

Question 1: What counties in CA generate the most power, and which ones consume the most power?

To answer this question, we are going to use 5 different data sets. The first dataset, powerplants, contains the name, latitude, longitude, fuels, and power generation of many different power plants. However, we are interested in the counties the power plants are located in. Thus, we are going to use a shape file to determine where the power plants intersect with the counties.

The powerplants dataset was sourced from the following website: https://datasets.wri.org/dataset/globalpowerplantdatabase.

The process for finding the FIPS code from the latitude and longitude was sourced from the following website (with modifications for reduced overall code): https://shiandy.com/post/2020/11/02/mapping-lat-long-to-fips/.

The county shape files were sourced from the following website: https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html.

```
# County / Name/ Lat / Long / powerGeneration / primary__fuel / other_fuels / others
powerplants <- read_csv("./data/global_power_plant_database_v_1_3/global_power_plant_database.csv",
```

```
                        show_col_types = FALSE) %>%
  filter(country == "USA")

# GEOID (fips) / state FIPS / county name / geometry
countyCAShape <- st_read("./data/cb_2018_us_county_500k/cb_2018_us_county_500k.shp", quiet = TRUE) %>%
  filter(STATEFP == "06")

# Add a column to powerplants, making a point geometry object.  'remove = false' prevents needing to ma
powerplants_geom <- powerplants %>%
  filter(!is.na(latitude), !is.na(longitude)) %>%
  st_as_sf(coords = c("longitude", "latitude"), remove=FALSE, crs = st_crs(countyCAShape))

# This variable stores which point objects is within which countyCAShape object.
intersected <- st_intersects(powerplants_geom, countyCAShape)

# Retrieve the fips code for all powerplants that match to a CA county, dropping the rest.  Then, add t
# Note: The %/% operator performs integer division.  The %% operator performs the modulus operation.
powerplants_ca_FIPS <- powerplants_geom %>%
  mutate(intersection = as.integer(intersected),
         fips = if_else(is.na(intersection), "",
                        countyCAShape$GEOID[intersection])) %>%
  filter(!is.na(intersection)) %>%
  mutate(fips = as.integer(fips),
         stateFIPS = fips%/%1000,
         countyFIPS = fips%%1000) %>%
  select(!intersection) %>%
  st_set_geometry(NULL)

# View a portion of the wrangled data.
powerplants_ca_FIPS %>%
  select(country:primary_fuel, generation_gwh_2019, fips) %>%
  head(n=5)
```

```
## # A tibble: 5 x 10
##   country country_~1 name  gppd_~2 capac~3 latit~4 longi~5 prima~6 gener~7  fips
##   <chr>   <chr>      <chr> <chr>     <dbl>   <dbl>   <dbl> <chr>     <dbl> <int>
## 1 USA     United St~ 1420~ USA005~     1.3    33.8   -118. Solar      1.78  6037
## 2 USA     United St~ 2081~ USA005~     1.2    36.7   -119. Solar      1.54  6019
## 3 USA     United St~ 2097~ USA005~     1.5    37.4   -121. Solar      2.66  6047
## 4 USA     United St~ 2127~ USA005~     1.3    39.3   -122. Solar      2.41  6101
## 5 USA     United St~ 2555~ USA005~     1      34.0   -118. Solar      1.25  6037
## # ... with abbreviated variable names 1: country_long, 2: gppd_idnr,
## #   3: capacity_mw, 4: latitude, 5: longitude, 6: primary_fuel,
## #   7: generation_gwh_2019
```

Now that we know what counties the power plants are in, we need to be able to map the counties. The third and fourth data sets give us the coordinates for the borders of California and the counties in California. However, the county co do not include the FIPS code though. Thus, we are going to use the county.fips package to find this last piece of information.

The California outline and the county outlines are sourced from the ggplot2 map_data function. The county.fips dataset is from the maps package.

```
# Retrieve the California state border coordinates, for use in later visualizations.
ca_border <- map_data("state") %>%
  filter(region == "california")

# Read in the county outlines, filtering for counties in California.
ca_county_map <- map_data("county") %>%
  filter(region == "california")

# Wrangle county.fips, creating 4 new columns: state name, county name, state FIPS, county FIPS.
# Then, join ca_county_map into the wrangled county.fips.  By being very specific with our "by" argumen
ca_county_map_fips <- county.fips %>%
  separate(col = polyname, into=c("state", "county"), sep = ",") %>%
  filter(state == "california") %>%
  mutate(fips = as.integer(fips),
         stateFIPS = fips%/%1000,
         countyFIPS = fips%%1000) %>%
  left_join(x = ca_county_map,
            y = .,
            by = c("region" = "state",
                   "subregion" = "county"))

# View a portion of the wrangled data.
head(ca_county_map_fips, n=5)
```

```
##        long      lat group order     region subregion fips stateFIPS countyFIPS
## 1 -121.4785 37.48290   157  6965 california   alameda 6001         6          1
## 2 -121.5129 37.48290   157  6966 california   alameda 6001         6          1
## 3 -121.8853 37.48290   157  6967 california   alameda 6001         6          1
## 4 -121.8968 37.46571   157  6968 california   alameda 6001         6          1
## 5 -121.9254 37.45998   157  6969 california   alameda 6001         6          1
```

We have now wrangled most of the data we will be using for the power production half of the question. Next, we are going to compute the answer to our question.

Some of the steps done here are not necessary to answer the question, but they will provide us with interesting data to plot.

```
# Find the total generation of a county by the fuel used.  Note that this drops specific information ab
total_generation_by_county_and_fuel <- powerplants_ca_FIPS %>%
  filter(!is.na(generation_gwh_2019)) %>%
  group_by(fips, primary_fuel) %>%
  summarize(sumGenerationGWH_2019 = sum(generation_gwh_2019))
```

```
## `summarise()` has grouped output by 'fips'. You can override using the
## `.groups` argument.
```

```
# Find the total generation of a county.
total_generation_by_county <- total_generation_by_county_and_fuel %>%
  group_by(fips) %>%
  summarize(sumGenerationGWH_2019 = sum(sumGenerationGWH_2019)) %>%
  mutate(primary_fuel = "Total County")

# Join the two by adding the totals as rows.  Remove any rows that have a generation equal to zero, whe
```

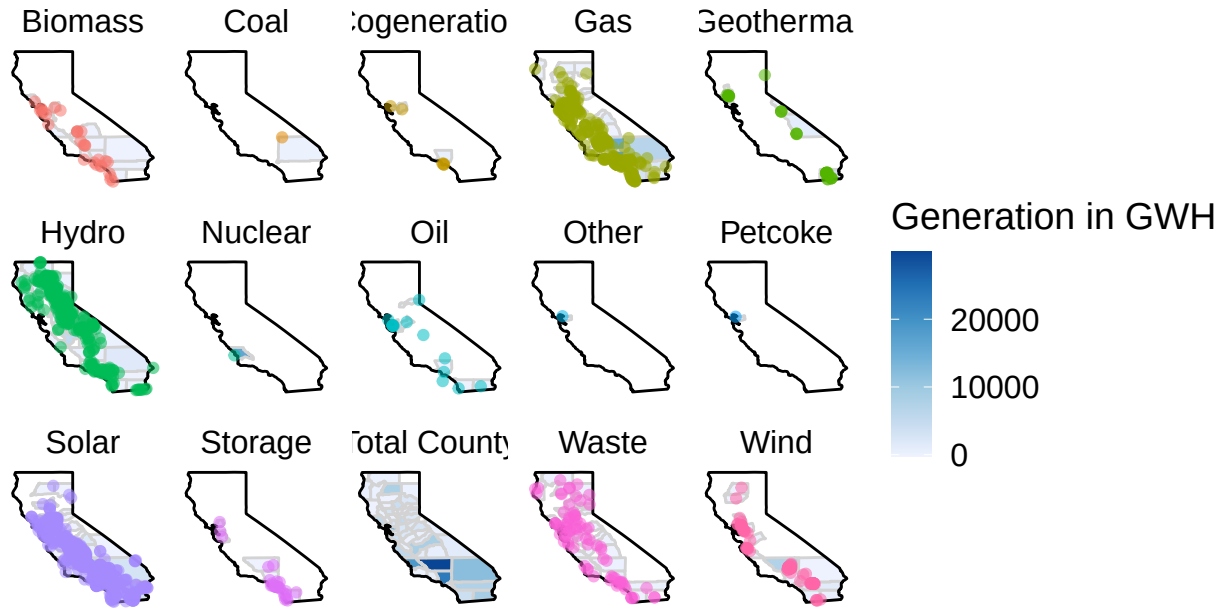
```
total_generation_by_county_and_fuel <- total_generation_by_county_and_fuel %>%
  full_join(total_generation_by_county, by=c("fips", "primary_fuel", "sumGenerationGWH_2019")) %>%
  filter(!near(sumGenerationGWH_2019, 0))

# Lastly, join this information with the coordinates wrangled earlier.
total_generation_map <- ca_county_map_fips %>%
  left_join(total_generation_by_county_and_fuel, by="fips")
```

This is simply a table showing the answer to the question.

```
total_generation_map %>%
  select(region:sumGenerationGWH_2019) %>%
  unique() %>%
  filter(primary_fuel=="Total County") %>%
  slice_max(n=5, order_by=sumGenerationGWH_2019)
```

```
##       region       subregion fips stateFIPS countyFIPS primary_fuel
## 1 california            kern 6029         6         29 Total County
## 2 california     los angeles 6037         6         37 Total County
## 3 california san luis obispo 6079         6         79 Total County
## 4 california    contra costa 6013         6         13 Total County
## 5 california  san bernardino 6071         6         71 Total County
##   sumGenerationGWH_2019
## 1              29959.89
## 2              23378.14
## 3              18690.75
## 4              14164.07
## 5              11968.47
```

```
total_generation_map %>%
  na.omit(sumGenerationGWH_2019) %>%
  ggplot() +
  geom_polygon(mapping = aes(long, lat, group=group, fill=sumGenerationGWH_2019),
               color= "lightgray") +
  geom_polygon(data = ca_border,
               mapping = aes(x=long, y=lat, group=group),
               fill=NA,
               color="black") +
  geom_point(data = powerplants_ca_FIPS,
             mapping = aes(x=longitude, y=latitude, color = primary_fuel, alpha = .5),
             show.legend = FALSE) +
  theme_map() +
  coord_equal() +
  labs(title = "California Power Generation by County", subtitle = "Separated by Fuel Source, Powerplan
  scale_fill_distiller(palette = "Blues", na.value="yellow", trans="reverse") +
  guides(fill = guide_colorbar(title = "Generation in GWH", reverse=TRUE)) +
  facet_wrap(~ primary_fuel, nrow=3)
```




```
total_generation_map %>%
  filter(is.na(primary_fuel)) %>%
  view()
```

Prepare the data related to the consumption of power.

```
# Read in this dataset for the other half of the quesion.
ca_county_power_use <- read_csv("./data/ElectricityByCounty CALIFORNIA Usage in Millions of kWh.csv",
                                show_col_types = FALSE) %>%
  mutate(County = tolower(County))

ca_county_power_use_map <- ca_county_map_fips %>%
  left_join(y = ca_county_power_use, by = c("subregion" = "County"))
```

Plotting power consumption

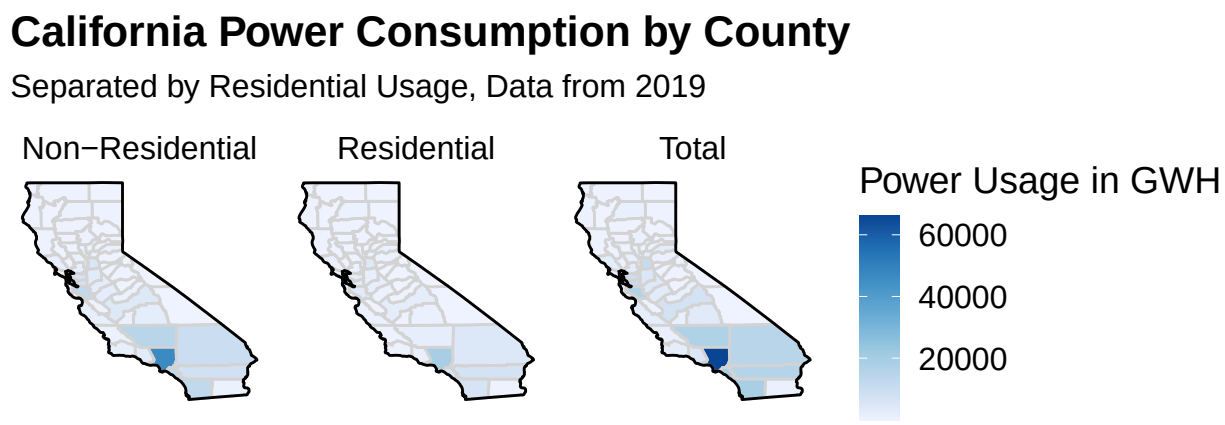
These are the 5 California counties that consume the most power.

```
ca_county_power_use_map %>%
  filter(Sector == "Total") %>%
  select(region:`2019`) %>%
  unique() %>%
  slice_max(order_by = `2019`, n=5)
```

```
##       region   subregion fips stateFIPS countyFIPS Sector     2019
## 1 california los angeles 6037         6         37  Total 66118.67
## 2 california      orange 6059         6         59  Total 19459.51
## 3 california   san diego 6073         6         73  Total 19047.67
## 4 california        kern 6029         6         29  Total 17105.08
## 5 california santa clara 6085         6         85  Total 16664.46
```

The graph of power consumption.

```
ca_county_power_use_map %>%
  ggplot() +
  geom_polygon(mapping = aes(long, lat, group=group, fill=`2019`),
               color= "lightgray") +
  geom_polygon(data = ca_border,
               mapping = aes(x=long, y=lat, group=group),
               fill=NA,
               color="black") +
  theme_map() +
  coord_equal() +
  scale_fill_distiller(palette = "Blues", na.value="yellow", trans="reverse") +
  labs(title = "California Power Consumption by County", subtitle = "Separated by Residential Usage, Da
  guides(fill = guide_colorbar(reverse=TRUE, title = "Power Usage in GWH")) +
  facet_wrap(~ Sector)
```



```
#  geom_point(data = wipPowerPlants, mapping = aes(x=longitude, y=latitude, color = primary_fuel)) +
```

Notes here about storage, non-power generating counties, etc.

```
total_generation_by_county_and_fuel %>%
  filter(primary_fuel == "Storage") %>%
  view()
```

Question 2: In California, what counties have an energy surplus and which counties have an energy deficit?

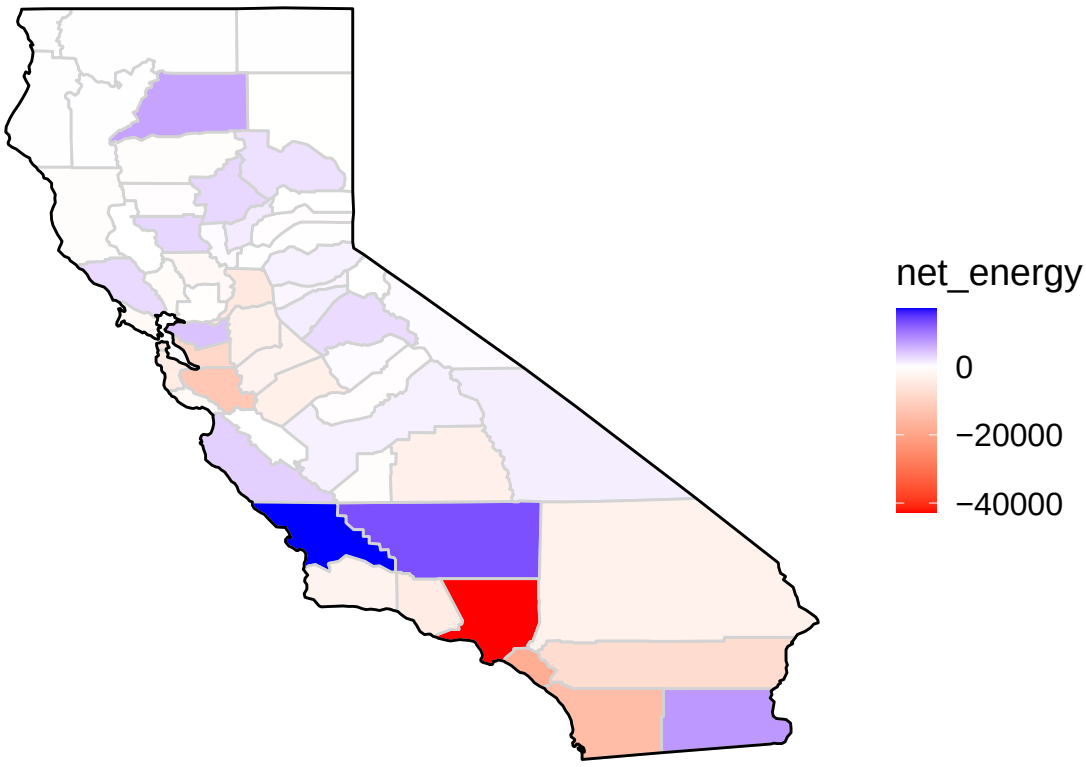
Get the stats.

```
net_consumption <- ca_county_power_use_map %>%
  select(long:`2019`) %>%
  unique() %>%
  filter(Sector == "Total")


net_production <- total_generation_by_county_and_fuel %>%
  filter(primary_fuel == "Total County")

usage_vs_production <- left_join(net_consumption, net_production, by="fips") %>%
  select(!c("Sector", "primary_fuel")) %>%
  rename(net_usage = `2019`,
         net_generation = sumGenerationGWH_2019) %>%
  mutate(net_generation = case_when(is.na(net_generation) ~ 0,
                                    TRUE ~ net_generation),
         net_energy = net_generation - net_usage,
         ratio_prod_v_use = net_generation/net_usage,
         ratio_net_v_use = net_energy/net_usage)
```

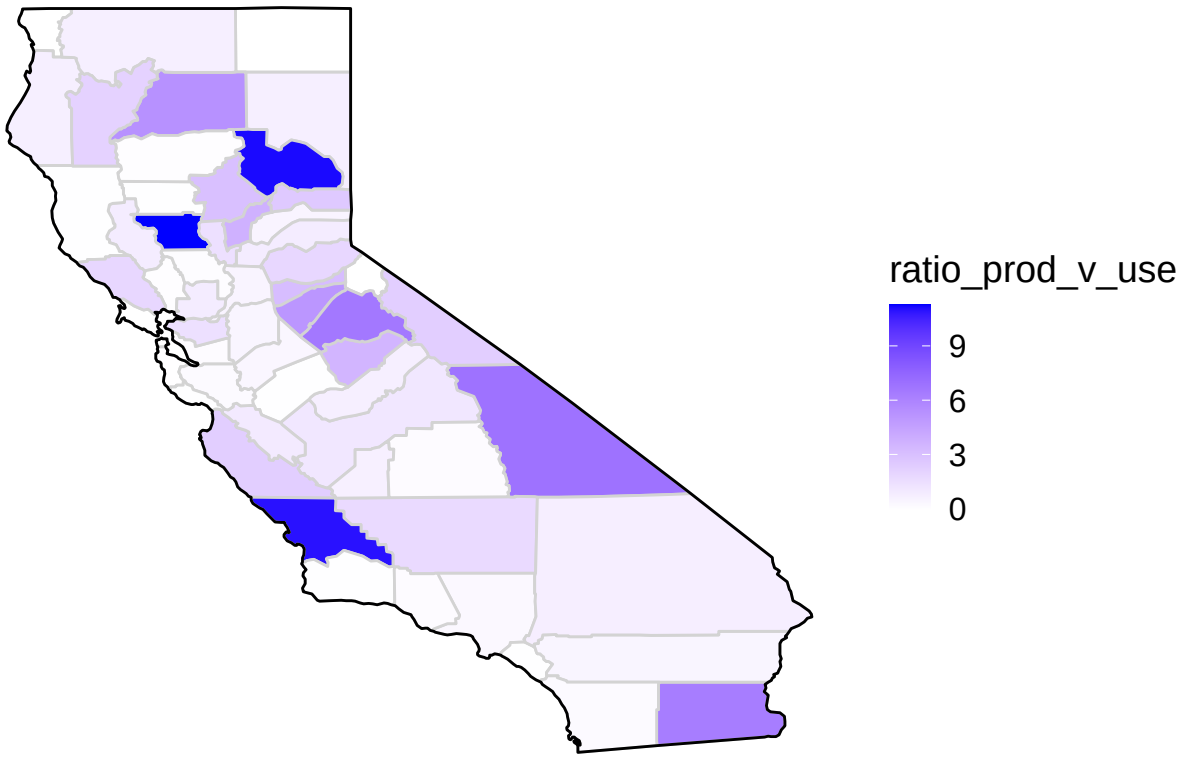
Plot the graphs.

```
zero_as_percentile_1 <- (0 - min(usage_vs_production$net_energy)) / (max(usage_vs_production$net_energy
usage_vs_production %>%
  ggplot() +
  geom_polygon(mapping = aes(x=long,
                             y=lat,
                             group=group,
                             fill=net_energy),
               color="lightgray") +
  geom_polygon(data = ca_border,
               mapping = aes(x=long, y=lat, group=group),
               fill=NA,
               color="black") +
  coord_equal() +
  theme_map() +
  scale_fill_gradientn(colours = c("red", "white", "blue"),
```

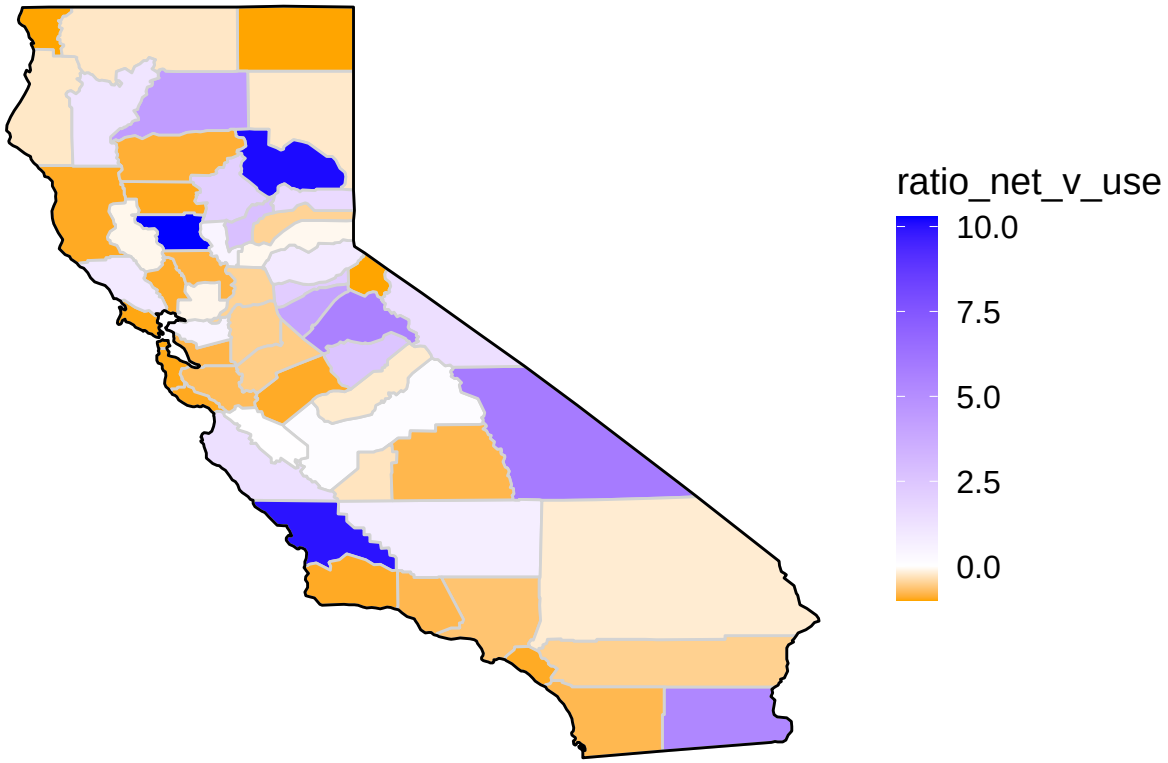
```
                        values = c(0, zero_as_percentile_1, 1)) +
  guides(fill = guide_colorbar(draw.ulim = TRUE))
```




```
usage_vs_production %>%
  ggplot() +
  geom_polygon(mapping = aes(x=long,
                             y=lat,
                             group=group,
                             fill=ratio_prod_v_use),
               color="lightgray") +
  geom_polygon(data = ca_border,
               mapping = aes(x=long, y=lat, group=group),
               fill=NA,
               color="black") +
  coord_equal() +
  theme_map() +
  scale_fill_gradientn(colours = c("white", "blue"),
                       values = c(0, 1)) +
  guides(fill = guide_colorbar())
```




```
zero_as_percentile_3 <- (0 - min(usage_vs_production$ratio_net_v_use)) / (max(usage_vs_production$ratio_
usage_vs_production %>%
  ggplot() +
  geom_polygon(mapping = aes(x=long,
                             y=lat,
                             group=group,
                             fill=ratio_net_v_use),
               color="lightgray") +
  geom_polygon(data = ca_border,
               mapping = aes(x=long, y=lat, group=group),
               fill=NA,
               color="black") +
  coord_equal() +
  theme_map() +
  scale_fill_gradientn(colours = c("orange", "white", "blue"),
                       values = c(0, zero_as_percentile_3, 1)) +
  guides(fill = guide_colorbar(barheight = 10))
```



Output the tables containing the counties with the highest and lowest net_energy values, by both number and ratio of net_energy to power consumption.

```
usage_vs_production %>%
  select(region:ratio_net_v_use) %>%
  unique() %>%
  slice_max(order_by=net_energy, n = 5)
```

```
##       region       subregion fips stateFIPS countyFIPS net_usage net_generation
## 1 california san luis obispo 6079         6         79  1707.386      18690.754
## 2 california            kern 6029         6         29 17105.082      29959.891
## 3 california        imperial 6025         6         25  1415.791       8966.348
## 4 california          shasta 6089         6         89  1535.591       8222.612
## 5 california    contra costa 6013         6         13  9639.409      14164.069
##   net_energy ratio_prod_v_use ratio_net_v_use
## 1  16983.368        10.947002       9.9470021
## 2  12854.809         1.751520       0.7515198
## 3   7550.557         6.333102       5.3331020
## 4   6687.021         5.354689       4.3546891
## 5   4524.660         1.469392       0.4693918
```

```
usage_vs_production %>%
  select(region:ratio_net_v_use) %>%
  unique() %>%
  slice_min(order_by=net_energy, n = 5)
```

```
##       region   subregion fips stateFIPS countyFIPS net_usage net_generation
## 1 california los angeles 6037         6         37  66118.67      23378.143
## 2 california      orange 6059         6         59  19459.51       1235.606
## 3 california   san diego 6073         6         73  19047.67       4121.234
## 4 california santa clara 6085         6         85  16664.46       4147.040
## 5 california     alameda 6001         6          1  10684.09       1879.522
##   net_energy ratio_prod_v_use ratio_net_v_use
## 1 -42740.530       0.35357853      -0.6464215
## 2 -18223.902       0.06349627      -0.9365037
## 3 -14926.440       0.21636415      -0.7836358
## 4 -12517.420       0.24885535      -0.7511447
## 5  -8804.564       0.17591790      -0.8240821
```

```
usage_vs_production %>%
  select(region:ratio_net_v_use) %>%
  unique() %>%
  slice_max(order_by=ratio_net_v_use, n = 5)
```

```
##       region       subregion fips stateFIPS countyFIPS net_usage net_generation
## 1 california          colusa 6011         6         11  285.4851       3216.651
## 2 california          plumas 6063         6         63  210.4883       2339.017
## 3 california san luis obispo 6079         6         79 1707.3856      18690.754
## 4 california            inyo 6027         6         27  208.6063       1429.587
## 5 california        tuolumne 6109         6        109  452.6442       2981.892
##   net_energy ratio_prod_v_use ratio_net_v_use
## 1   2931.166        11.267315       10.267315
## 2   2128.529        11.112337       10.112337
## 3  16983.368        10.947002        9.947002
## 4   1220.981         6.853038        5.853038
## 5   2529.248         6.587717        5.587717
```

```
usage_vs_production %>%
  select(region:ratio_net_v_use) %>%
  unique() %>%
  slice_min(order_by=ratio_net_v_use, n = 5)
```

```
##       region     subregion fips stateFIPS countyFIPS  net_usage net_generation
## 1 california        alpine 6003         6          3   18.90621          0.000
## 2 california     del norte 6015         6         15  210.25367          0.000
## 3 california         modoc 6049         6         49  145.17415          0.000
## 4 california san francisco 6075         6         75 5603.60421        102.411
## 5 california         marin 6041         6         41 1355.34274         29.581
##    net_energy ratio_prod_v_use ratio_net_v_use
## 1   -18.90621       0.00000000      -1.0000000
## 2  -210.25367       0.00000000      -1.0000000
## 3  -145.17415       0.00000000      -1.0000000
## 4 -5501.19321       0.01827592      -0.9817241
## 5 -1325.76175       0.02182548      -0.9781745
```

Question 3: What is the per capita income trend of those in locations where energy is produced versus the income of those where it is consumed?