assembly

Group 2

7/28/2021

```r
####annual aqi####
data2020.yr <- fread("data/annual_aqi_by_county_2020.csv", stringsAsFactors = TRUE)
data2019.yr <- fread("data/annual_aqi_by_county_2019.csv", stringsAsFactors = TRUE)
data2018.yr <- fread("data/annual_aqi_by_county_2018.csv", stringsAsFactors = TRUE)
data2017.yr <- fread("data/annual_aqi_by_county_2017.csv", stringsAsFactors = TRUE)
data2016.yr <- fread("data/annual_aqi_by_county_2016.csv", stringsAsFactors = TRUE)
data2015.yr <- fread("data/annual_aqi_by_county_2015.csv", stringsAsFactors = TRUE)
data2014.yr <- fread("data/annual_aqi_by_county_2014.csv", stringsAsFactors = TRUE)
data2013.yr <- fread("data/annual_aqi_by_county_2013.csv", stringsAsFactors = TRUE)
data2012.yr <- fread("data/annual_aqi_by_county_2012.csv", stringsAsFactors = TRUE)
data2011.yr <- fread("data/annual_aqi_by_county_2011.csv", stringsAsFactors = TRUE)
data2010.yr <- fread("data/annual_aqi_by_county_2010.csv", stringsAsFactors = TRUE)

####daily aqi####
data2020.day <- fread("data/daily_aqi_by_county_2020.csv", stringsAsFactors = TRUE)
data2019.day <- fread("data/daily_aqi_by_county_2019.csv", stringsAsFactors = TRUE)
data2018.day <- fread("data/daily_aqi_by_county_2018.csv", stringsAsFactors = TRUE)
data2017.day <- fread("data/daily_aqi_by_county_2017.csv", stringsAsFactors = TRUE)
data2016.day <- fread("data/daily_aqi_by_county_2016.csv", stringsAsFactors = TRUE)
data2015.day <- fread("data/daily_aqi_by_county_2015.csv", stringsAsFactors = TRUE)
data2014.day <- fread("data/daily_aqi_by_county_2014.csv", stringsAsFactors = TRUE)
data2013.day <- fread("data/daily_aqi_by_county_2013.csv", stringsAsFactors = TRUE)
data2012.day <- fread("data/daily_aqi_by_county_2012.csv", stringsAsFactors = TRUE)
data2011.day <- fread("data/daily_aqi_by_county_2011.csv", stringsAsFactors = TRUE)
data2010.day <- fread("data/daily_aqi_by_county_2010.csv", stringsAsFactors = TRUE)
data2009.day <- fread("data/daily_aqi_by_county_2009.csv", stringsAsFactors = TRUE)
data2008.day <- fread("data/daily_aqi_by_county_2008.csv", stringsAsFactors = TRUE)
data2007.day <- fread("data/daily_aqi_by_county_2007.csv", stringsAsFactors = TRUE)
data2006.day <- fread("data/daily_aqi_by_county_2006.csv", stringsAsFactors = TRUE)
data2005.day <- fread("data/daily_aqi_by_county_2005.csv", stringsAsFactors = TRUE)
data2004.day <- fread("data/daily_aqi_by_county_2004.csv", stringsAsFactors = TRUE)
data2003.day <- fread("data/daily_aqi_by_county_2003.csv", stringsAsFactors = TRUE)
data2002.day <- fread("data/daily_aqi_by_county_2002.csv", stringsAsFactors = TRUE)
data2001.day <- fread("data/daily_aqi_by_county_2001.csv", stringsAsFactors = TRUE)
data2000.day <- fread("data/daily_aqi_by_county_2000.csv", stringsAsFactors = TRUE)

####yearly greenhouse gas####
airquality_ozone<- read.csv("data/OzoneNational.csv", header=T)
airquality_nitrogen<- read.csv("data/Nitrogen_DioxideNational.csv", header=T)
airquality_sulfur<-read.csv("data/Sulfur_DioxideNational.csv", header=T)
airquality_cmonoxide<- read.csv("data/Carbon_MonoxideNational.csv", header=T)
airquality_lead<- read.csv("data/LeadNational.csv", header=T)
```

```
raw_data.day <- rbind(
  data2020.day, data2019.day, data2018.day, data2017.day, data2016.day, data2015.day, data2014.day, data
  data2012.day, data2011.day, data2010.day, data2009.day, data2008.day, data2007.day, data2006.day, data
  data2004.day,data2003.day, data2002.day, data2001.day, data2000.day
)

raw_data.day <- mutate(raw_data.day,
                       date_formatted = as.Date(Date),
                       month = as.Date(format(date_formatted, "%Y-%m-01"))
)

raw_data.day <- rename(raw_data.day,
                       state = `State Name`,
                       county = `county Name`,
                       state.code = `State Code`,
                       county.code = `County Code`,
                       defining.parameter = `Defining Parameter`,
                       defining.site = `Defining Site`,
                       num.rep.sites = `Number of Sites Reporting`
)

counties_by_month <- group_by(raw_data.day, county, month) %>% summarise(county.mean = mean(AQI))
```

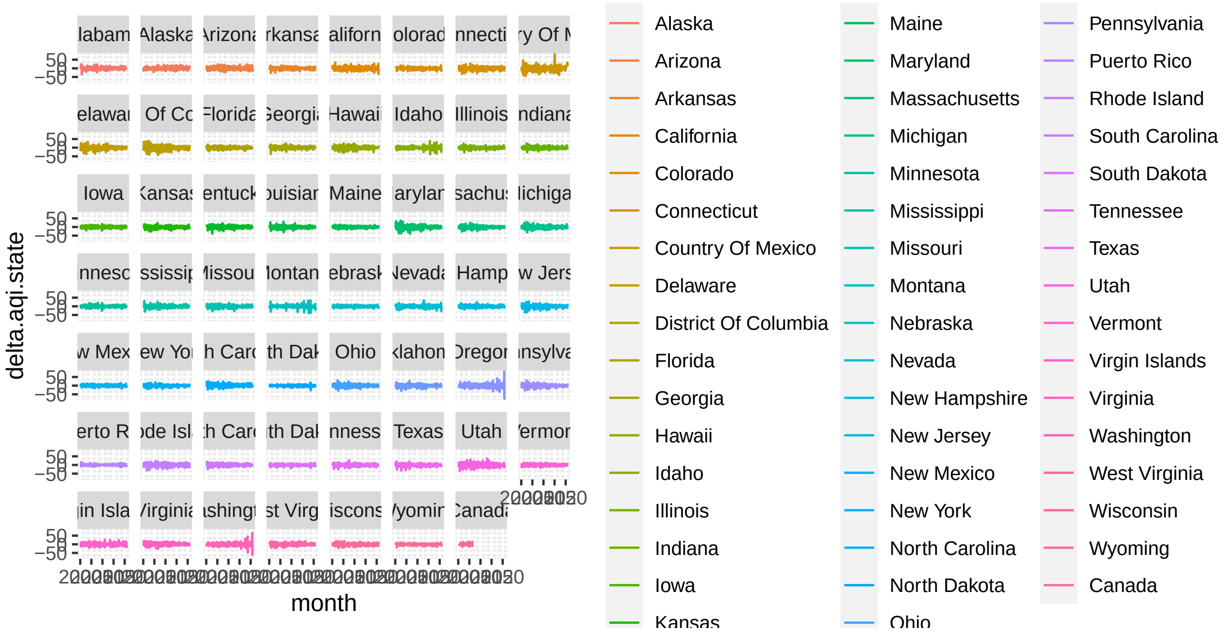
```
## ‘summarise()‘ has grouped output by ’county’. You can override using the ‘.groups‘ argument.
```

```
states_by_month <- group_by(raw_data.day, state, month) %>% summarise(state.mean = mean(AQI))
```

```
## ‘summarise()‘ has grouped output by ’state’. You can override using the ‘.groups‘ argument.
```

```
states_by_month <- group_by(states_by_month, state) %>% mutate(delta.aqi.state = state.mean - lag(state

raw_data.day <- merge(raw_data.day, counties_by_month, by=c("county", "month"))
raw_data.day <- merge(raw_data.day, states_by_month, by=c("state", "month"))
```

```
ggplot(states_by_month, aes(x = month, y = delta.aqi.state, color = state)) +
  geom_line() +
  facet_wrap( ~ state)
```

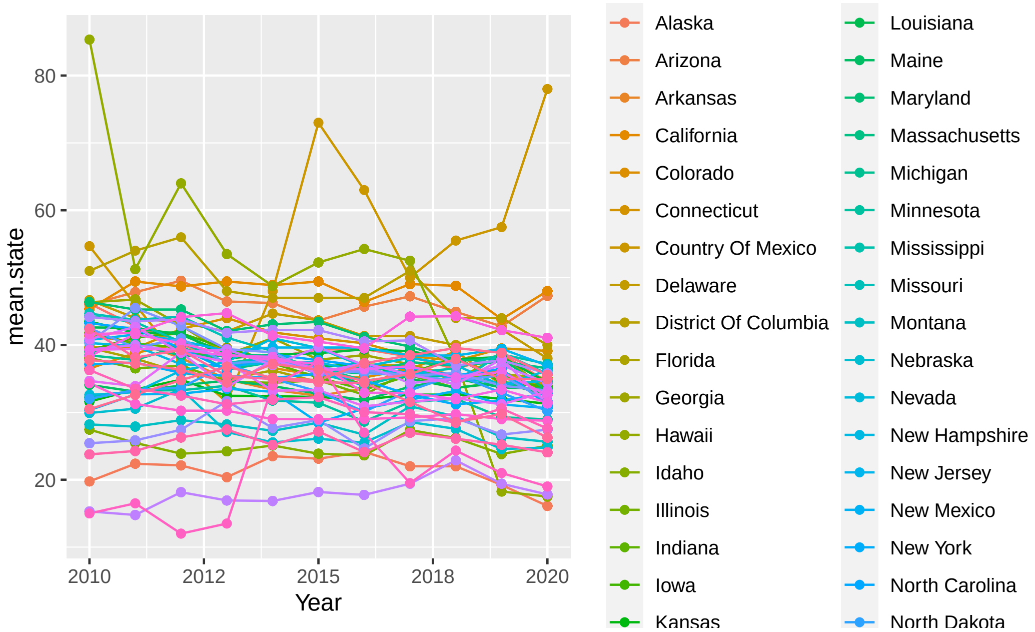


```r
raw_data.yr <- rbind(data2020.yr, data2019.yr, data2018.yr, data2017.yr, data2016.yr,
                     data2015.yr, data2014.yr, data2013.yr, data2012.yr, data2011.yr, data2010.yr)

raw_data.yr <- raw_data.yr %>% rename(
  num.days = `Days with AQI`,
  good.days = `Good Days`,
  mod.days = `Moderate Days`,
  sens.group.days = `Unhealthy for Sensitive Groups Days`,
  unhlthy.days = `Unhealthy Days`,
  very.unhlthy.days = `Very Unhealthy Days`,
  haz.days = `Hazardous Days`,
  max.aqi = `Max AQI`,
  med.aqi = `Median AQI`,
  state = State
)

raw_data.yr <- raw_data.yr %>%
  mutate(good.pct = good.days/num.days*100,
         mod.pct = mod.days/num.days*100,
         sens.group.pct = sens.group.days/num.days*100,
         unhlthy.pct = unhlthy.days/num.days*100,
         very.unhlthy.pct = very.unhlthy.days/num.days*100,
         haz.pct = haz.days/num.days*100)

mean.state.df <- raw_data.yr %>% group_by(state, Year) %>% summarize(mean.state = mean(med.aqi),
                                                                     peak.state = max(max.aqi))
```

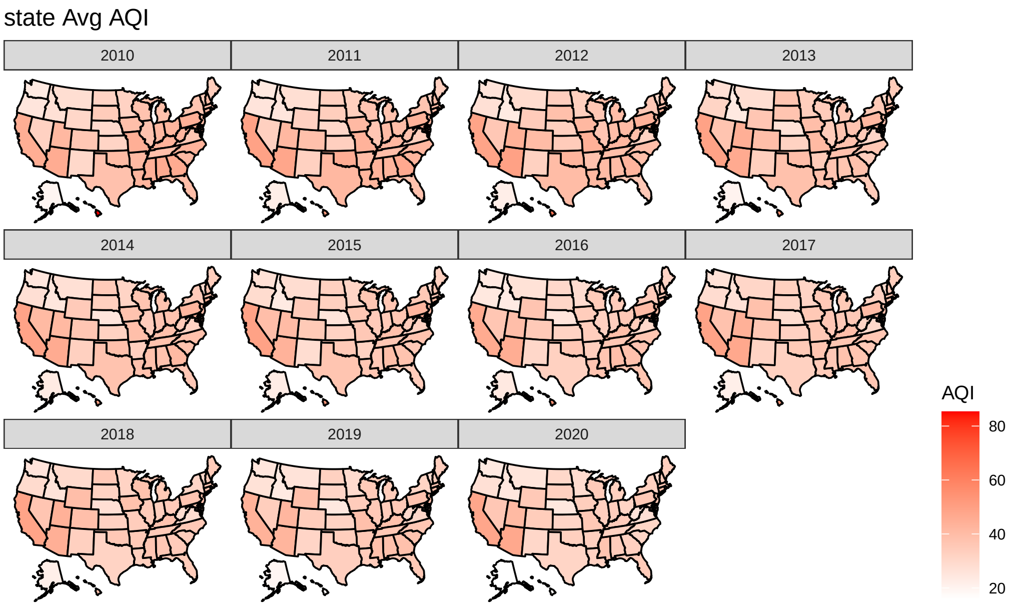
```
## `summarise()` has grouped output by 'state'. You can override using the `.groups` argument.
```

```r
raw_data.yr <- merge(raw_data.yr, mean.state.df, by=c("state", "Year"))
```

```
ggplot(raw_data.yr, aes(x = Year, y = mean.state, color = state)) +
  geom_line() +
  geom_point()
```



```
plot_usmap(regions = "state",
           #regions = "counties", for county level summary
           data = mean.state.df,
           values = "mean.state", exclude = c("District of Columbia", "Country of Mexico",
                                              "Virgin Islands", "Puerto Rico")
           , color = "black") +
  scale_fill_gradient(
    low = "white", high = "red",
    name = "AQI",
    label = scales::comma) +
  labs(title = "state Avg AQI") +
  theme(legend.position = "right") +
  facet_wrap(~ Year)
```



```
# EDA
airquality_nitrogen <- mutate(airquality_nitrogen,
                              nitrogen_concentration = airquality_nitrogen$X10th.Percentile+airquality_n
airquality_ozone <- mutate(airquality_ozone,
                           ozone_concentration = airquality_ozone$X10th.Percentile+airquality_ozone$X90t
airquality_sulfur <- mutate(airquality_sulfur,
                            sulfur_concentration = airquality_sulfur$X10th.Percentile+airquality_sulfur$
airquality_lead <- mutate(airquality_lead,
                          lead_concentration = airquality_lead$X10th.Percentile+airquality_lead$X90th.Pe
airquality_cmonoxide <- mutate(airquality_cmonoxide,
                               cmonoxide_concentration = airquality_cmonoxide$X10th.Percentile+airqualit

#data format
str(airquality_ozone)
```

```
## 'data.frame':    31 obs. of  6 variables:
##  $ Year                : int  1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 ...
##  $ Mean                : num  0.0877 0.0882 0.0837 0.0848 0.0851 ...
##  $ Number.of.Trend.Sites: int  394 394 394 394 394 394 394 394 394 394 ...
##  $ X10th.Percentile     : num  0.068 0.067 0.068 0.065 0.069 0.071 0.07 0.068 0.071 0.07 ...
##  $ X90th.Percentile     : num  0.108 0.11 0.098 0.103 0.099 0.109 0.1 0.102 0.107 0.104 ...
##  $ ozone_concentration  : num  0.176 0.177 0.166 0.168 0.168 0.18 0.17 0.17 0.178 0.174 ...
```

```
str(airquality_nitrogen)
```

```
## 'data.frame':    41 obs. of  6 variables:
##  $ Year                  : int  1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 ...
##  $ Mean                  : num  112.9 109.9 110.1 99.5 93.1 ...
##  $ Number.of.Trend.Sites : int  20 20 20 20 20 20 20 20 20 20 ...
##  $ X10th.Percentile      : num  66.5 66.5 66.5 59.5 59.8 ...
##  $ X90th.Percentile      : num  190 190 190 165 145 ...
##  $ nitrogen_concentration: num  256 256 256 224 205 ...
```

```
str(airquality_sulfur)
```

```
## ’data.frame’:    41 obs. of  6 variables:
##  $ Year                 : int  1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 ...
##  $ Mean                 : num  162 154 141 154 142 ...
##  $ Number.of.Trend.Sites: int  32 32 32 32 32 32 32 32 32 32 ...
##  $ X10th.Percentile     : num  50 50 40 37 30 40 30 25 30 30 ...
##  $ X90th.Percentile     : num  271 251 250 321 280 250 258 242 220 261 ...
##  $ sulfur_concentration : num  321 301 290 358 310 290 288 267 250 291 ...
```

```
str(airquality_cmonoxide)
```

```
## ’data.frame’:    41 obs. of  6 variables:
##  $ Year                   : int  1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 ...
##  $ Mean                   : num  9.46 9.21 8.59 9.11 8.34 ...
##  $ Number.of.Trend.Sites  : int  36 36 36 36 36 36 36 36 36 36 ...
##  $ X10th.Percentile       : num  4.6 3.9 4.8 4.5 4.4 3.9 4.5 4.5 4.1 3.6 ...
##  $ X90th.Percentile       : num  16.8 14.6 13.9 16.2 13.7 12.3 12.6 10.2 10.5 10.4 ...
##  $ cmonoxide_concentration: num  21.4 18.5 18.7 20.7 18.1 16.2 17.1 14.7 14.6 14 ...
```

```
str(airquality_lead)
```

```
## ’data.frame’:    11 obs. of  6 variables:
##  $ Year                 : int  2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 ...
##  $ Mean                 : num  0.2272 0.2277 0.2013 0.1378 0.0776 ...
##  $ Number.of.Trend.Sites: int  83 83 83 83 83 83 83 83 83 83 ...
##  $ X10th.Percentile     : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0 0 ...
##  $ X90th.Percentile     : num  0.64 0.66 0.51 0.43 0.17 0.11 0.12 0.11 0.11 0.09 ...
##  $ lead_concentration   : num  0.65 0.67 0.52 0.44 0.18 0.12 0.13 0.12 0.11 0.09 ...
```

```
#quick summary. missing values may be shown
summary(airquality_ozone)
```

```
##       Year           Mean        Number.of.Trend.Sites X10th.Percentile
##  Min.   :1990   Min.   :0.0652   Min.   :394           Min.   :0.0540  
##  1st Qu.:1998   1st Qu.:0.0694   1st Qu.:394           1st Qu.:0.0600  
##  Median :2005   Median :0.0787   Median :394           Median :0.0650  
##  Mean   :2005   Mean   :0.0780   Mean   :394           Mean   :0.0645  
##  3rd Qu.:2012   3rd Qu.:0.0849   3rd Qu.:394           3rd Qu.:0.0680  
##  Max.   :2020   Max.   :0.0896   Max.   :394           Max.   :0.0710  
##  X90th.Percentile ozone_concentration
##  Min.   :0.075    Min.   :0.131      
##  1st Qu.:0.080    1st Qu.:0.140      
##  Median :0.092    Median :0.158      
##  Mean   :0.092    Mean   :0.156      
##  3rd Qu.:0.101    3rd Qu.:0.169      
##  Max.   :0.110    Max.   :0.180      
```

```
summary(airquality_nitrogen)
```

```
##       Year           Mean       Number.of.Trend.Sites X10th.Percentile
##  Min.   :1980   Min.   : 40.5   Min.   :20            Min.   :29.4    
##  1st Qu.:1990   1st Qu.: 48.3   1st Qu.:20            1st Qu.:37.0    
##  Median :2000   Median : 66.6   Median :20            Median :45.0    
##  Mean   :2000   Mean   : 71.3   Mean   :20            Mean   :46.5    
##  3rd Qu.:2010   3rd Qu.: 94.4   3rd Qu.:20            3rd Qu.:56.0    
##  Max.   :2020   Max.   :112.9   Max.   :20            Max.   :66.5    
##  X90th.Percentile nitrogen_concentration
##  Min.   : 55      Min.   : 84.3         
##  1st Qu.: 65      1st Qu.:103.6         
##  Median :112      Median :159.5         
##  Mean   :116      Mean   :162.1         
##  3rd Qu.:165      3rd Qu.:222.5         
##  Max.   :190      Max.   :256.5
```

```
summary(airquality_sulfur)
```

```
##       Year           Mean       Number.of.Trend.Sites X10th.Percentile
##  Min.   :1980   Min.   : 10.3   Min.   :32            Min.   : 3.0    
##  1st Qu.:1990   1st Qu.: 40.9   1st Qu.:32            1st Qu.:10.9    
##  Median :2000   Median : 82.0   Median :32            Median :26.0    
##  Mean   :2000   Mean   : 82.0   Mean   :32            Mean   :23.2    
##  3rd Qu.:2010   3rd Qu.:118.4   3rd Qu.:32            3rd Qu.:31.0    
##  Max.   :2020   Max.   :162.4   Max.   :32            Max.   :50.0    
##  X90th.Percentile sulfur_concentration
##  Min.   : 16      Min.   : 19         
##  1st Qu.: 84      1st Qu.: 95         
##  Median :151      Median :180         
##  Mean   :150      Mean   :173         
##  3rd Qu.:218      3rd Qu.:248         
##  Max.   :321      Max.   :358
```

```
summary(airquality_cmonoxide)
```

```
##       Year           Mean      Number.of.Trend.Sites X10th.Percentile
##  Min.   :1980   Min.   :1.27   Min.   :36            Min.   :0.70    
##  1st Qu.:1990   1st Qu.:1.76   1st Qu.:36            1st Qu.:0.90    
##  Median :2000   Median :3.79   Median :36            Median :2.00    
##  Mean   :2000   Mean   :4.31   Mean   :36            Mean   :2.33    
##  3rd Qu.:2010   3rd Qu.:6.56   3rd Qu.:36            3rd Qu.:3.60    
##  Max.   :2020   Max.   :9.46   Max.   :36            Max.   :4.80    
##  X90th.Percentile cmonoxide_concentration
##  Min.   : 1.80    Min.   : 2.50          
##  1st Qu.: 2.60    1st Qu.: 3.40          
##  Median : 5.40    Median : 7.40          
##  Mean   : 6.61    Mean   : 8.95          
##  3rd Qu.: 9.80    3rd Qu.:13.60          
##  Max.   :16.80    Max.   :21.40
```

```
summary(airquality_cmonoxide)
```

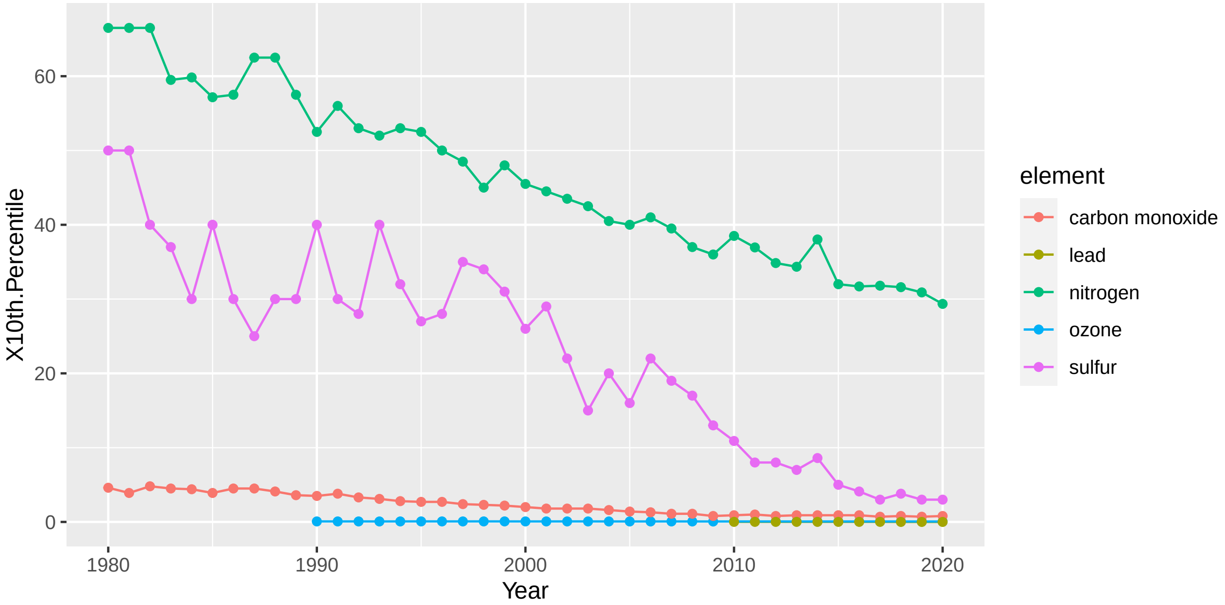
```
##       Year           Mean      Number.of.Trend.Sites X10th.Percentile
##  Min.   :1980   Min.   :1.27   Min.   :36             Min.   :0.70
##  1st Qu.:1990   1st Qu.:1.76   1st Qu.:36             1st Qu.:0.90
##  Median :2000   Median :3.79   Median :36             Median :2.00
##  Mean   :2000   Mean   :4.31   Mean   :36             Mean   :2.33
##  3rd Qu.:2010   3rd Qu.:6.56   3rd Qu.:36             3rd Qu.:3.60
##  Max.   :2020   Max.   :9.46   Max.   :36             Max.   :4.80
##  X90th.Percentile cmonoxide_concentration
##  Min.   : 1.80    Min.   : 2.50
##  1st Qu.: 2.60    1st Qu.: 3.40
##  Median : 5.40    Median : 7.40
##  Mean   : 6.61    Mean   : 8.95
##  3rd Qu.: 9.80    3rd Qu.:13.60
##  Max.   :16.80    Max.   :21.40
```

```
ozone.fit <- lm(Year~ozone_concentration, data =  airquality_ozone)
nitrogen.fit <- lm(Year~ nitrogen_concentration, data =  airquality_nitrogen)
sulfur.fit <- lm(Year~ sulfur_concentration, data =  airquality_sulfur)
cmonoxide.fit <- lm(Year~ cmonoxide_concentration, data =  airquality_cmonoxide)
lead.fit <- lm(Year~ lead_concentration, data =  airquality_lead)

ozone.temp <- select(airquality_ozone, c("Year", "X90th.Percentile", "X10th.Percentile"))
nitrogen.temp <- select(airquality_nitrogen, c("Year", "X90th.Percentile", "X10th.Percentile"))
sulfur.temp <- select(airquality_sulfur, c("Year", "X90th.Percentile", "X10th.Percentile"))
cmonoxide.temp <- select(airquality_cmonoxide, c("Year", "X90th.Percentile", "X10th.Percentile"))
lead.temp <- select(airquality_lead, c("Year", "X90th.Percentile", "X10th.Percentile"))
ozone.temp <- mutate(ozone.temp, element = "ozone")
nitrogen.temp <- mutate(nitrogen.temp, element = "nitrogen")
sulfur.temp <- mutate(sulfur.temp, element = "sulfur")
cmonoxide.temp <- mutate(cmonoxide.temp, element = "carbon monoxide")
lead.temp <- mutate(lead.temp, element = "lead")

ppm <- rbind(ozone.temp, nitrogen.temp, sulfur.temp, cmonoxide.temp, lead.temp)
ppm <- ppm[,c(1, 4, 3, 2)]
ppm <- group_by(ppm, Year, element, X10th.Percentile)
summarise(ppm)
```

```
## ‘summarise()‘ has grouped output by ’Year’, ’element’. You can override using the ‘.groups‘ argument
```

```
## # A tibble: 165 x 3
## # Groups:   Year, element [165]
##     Year element          X10th.Percentile
##    <int> <chr>                      <dbl>
##  1  1980 carbon monoxide              4.6
##  2  1980 nitrogen                    66.5
##  3  1980 sulfur                      50
##  4  1981 carbon monoxide              3.9
##  5  1981 nitrogen                    66.5
##  6  1981 sulfur                      50
##  7  1982 carbon monoxide              4.8
##  8  1982 nitrogen                    66.5
##  9  1982 sulfur                      40
## 10  1983 carbon monoxide              4.5
## # ... with 155 more rows
```

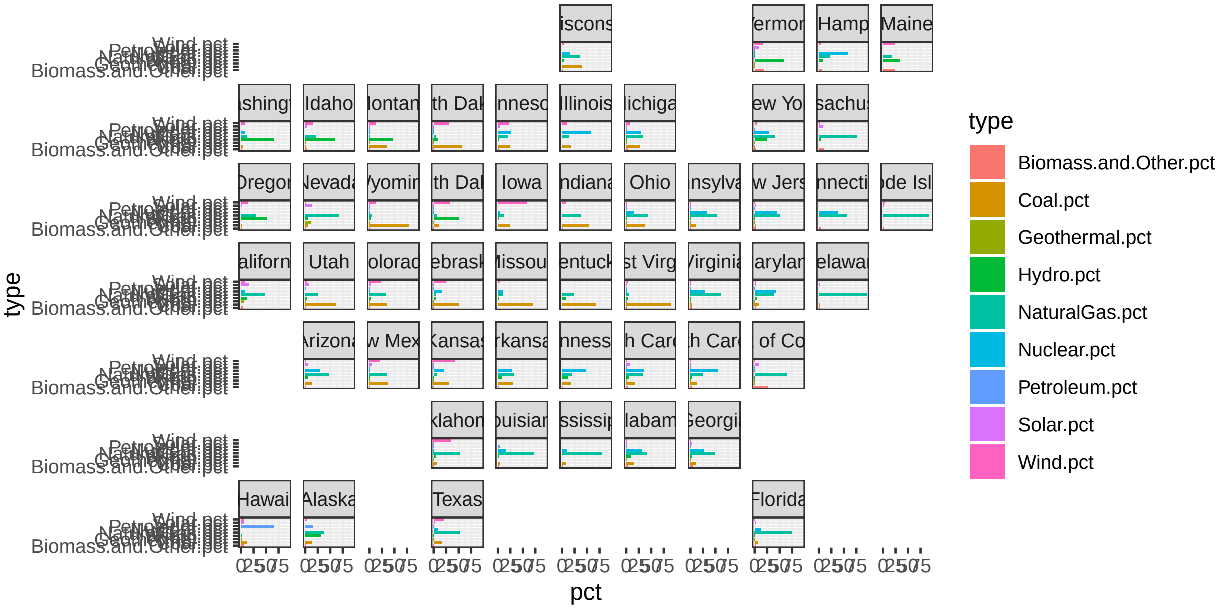
```
ppm.plot <- ggplot(ppm, aes(x = Year, y = X10th.Percentile, color = element, group = element)) +
  geom_line() +
  geom_point()
ppm.plot
```



```
###individual elmnts plots####
# #Ozone concentration by year
# fit1 <- lm(Year~ozone_concentration, data = airquality_ozone)
# ggplot(airquality_ozone , aes(x = Year, y = ozone_concentration, color = ozone_concentration)) +
#   ggtitle(" Ozone Concentration by Year") +
#   geom_line() +
#   geom_point()
#
# #Nitrogen concentration by year
# fit1 <- lm(airquality_nitrogen$Year~ nitrogen_concentration, data =  airquality_nitrogen)
# ggplot(airquality_nitrogen , aes(x = Year, y = nitrogen_concentration, color = nitrogen_concentration
#   ggtitle(" Nitrogen Concentration by Year") +
#   geom_line() +
#   geom_point()
#
# #Sulfur concentration by year
# fit1 <- lm(airquality_sulfur$Year~ sulfur_concentration, data =  airquality_sulfur)
# ggplot(airquality_sulfur , aes(x = Year, y = sulfur_concentration, color = sulfur_concentration)) +
#   ggtitle(" Sulfur Concentration by Year") +
#   geom_line() +
#   geom_point()
#
# #CO1 concentration by year
# fit1 <- lm(airquality_cmonoxide$Year~ cmonoxide_concentration, data =  airquality_cmonoxide)
# ggplot(airquality_cmonoxide , aes(x = Year, y = cmonoxide_concentration, color = cmonoxide_concentrat
#   ggtitle(" CO Concentration by Year") +
#   geom_line() +
#   geom_point()
```

```r
#
# #Lead concentration by year
# fit1 <- lm(airquality_lead$Year~ lead_concentration, data =  airquality_lead)
# ggplot(airquality_lead , aes(x = Year, y = lead_concentration, color = lead_concentration)) +
#   ggtitle(" Lead Concentration by Year") +
#   geom_line() +
#   geom_point()
```

```r
url <- "https://www.nei.org/resources/statistics/state-electricity-generation-fuel-shares"
enrgy_srcs <- read_html(url) %>% html_table()

state_data <- data.frame(enrgy_srcs)
state_data <- state_data %>% rename(
  Nuclear.pct = Nuclear....,
  Coal.pct = Coal....,
  NaturalGas.pct = Natural.Gas....,
  Petroleum.pct = Petroleum....,
  Hydro.pct = Hydro....,
  Geothermal.pct = Geothermal....,
  Solar.pct = Solar...PV....,
  Wind.pct = Wind....,
  Biomass.and.Other.pct = Biomass.and.Other....
)

state_data[state_data$Hydro.pct == "(0.2)", c("Hydro.pct")] <- 0.2
state_data[state_data$Biomass.and.Other.pct == "(0.0)", c("Biomass.and.Other.pct")] <- 0.0
state_data[state_data$State == "Iowa1", c("State")] <- "Iowa"
state_data[state_data$State == "New York2", c("State")] <- "New York"

state_data %<>% mutate_at(c(
  "Nuclear.pct",
  "Coal.pct",
  "NaturalGas.pct",
  "Petroleum.pct",
  "Hydro.pct",
  "Geothermal.pct",
  "Solar.pct",
  "Wind.pct",
  "Biomass.and.Other.pct"),
  as.numeric)

data_long <- gather(state_data, "type", "pct", -State)

stateEnrgyDist <- ggplot(data_long, aes(y = pct, x = type, fill = type)) +
  geom_col() +
  coord_flip() +
  theme_bw() +
  facet_geo(~ State, move_axes = FALSE)
stateEnrgyDist
```


iscons
ermor
Hamp
Maine
ıshingt
Idaho
lontan
th Dak
nnesc
Illinois
lichiga
ew Yo
sachu
Oregor
Nevad
/yomir
th Dal
Iowa
ndiana
Ohio
nsylva
w Jers
nnecti
ɔde Isl
aliforn
Utah
olorad
ebrask
lissou
entuck
st Virg
/irginia
arylar
elawa
rizona
w Mex
Kansa
rkansa
nness
h Caro
h Caro
of Co
klahon
uisian
ssissip
labam
Georgi
Hawai
Alaska
Texas
Florida
Wind.pct
Biomass.and.Other.pct
type
0 25 50 75
pct
type
Biomass.and.Other.pct
Coal.pct
Geothermal.pct
Hydro.pct
NaturalGas.pct
Nuclear.pct
Petroleum.pct
Solar.pct
Wind.pct