DATA 606 Data Project Proposal

Alyssa Gurkas

Background

The U.S. Department of Energy prepares an annual Electric Power Report that includes information about energy production, sales, consumption of fossil fuels, environmental data, and other topics related to energy. Additionally, the U.S. Department of Commerce tracks state annual summary statistics which include GDP by state from years 1998-2023.

Project Description

This project will explore if states with a higher GDP are more likely to release CO2 into the atmosphere than states with a lower GDP.

Data Preparation

```
library(readxl)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 4.4.2
```

```
##
## Attaching package: 'Hmisc'
##
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
##
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```r
library(plotly)
```

```
## Warning: package 'plotly' was built under R version 4.4.2
```

```
##
## Attaching package: 'plotly'
##
## The following object is masked from 'package:Hmisc':
##
##     subplot
##
## The following object is masked from 'package:ggplot2':
##
##     last_plot
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following object is masked from 'package:graphics':
##
##     layout
```

```r
emissions <- read_excel("emission_annual.xlsx")
stategdp <- read_excel("stategdp_summary.xlsx")
```

```r
emissions_v2 <- emissions %>%
  filter(`Year` >=1998,
         `Producer Type` == "Total Electric Power Industry",
         `Energy Source` == "All Sources",
         `State` != "US-TOTAL")
```

```r
stategdp_v2 <- stategdp %>%
  select(-GeoFips) %>%
  pivot_longer(!GeoName, names_to = "year", values_to = "gdp") %>%
  filter(`GeoName` != "United States") %>%
  mutate(`State`= state.abb[match(`GeoName`,state.name)]) %>%
  select(-GeoName) %>%
  drop_na()

stategdp_v2$year <- as.numeric(stategdp_v2$year)

stategdp_v3 <- stategdp_v2 %>%
  rename("Year" = "year",
         "GDP" = "gdp")
```

```r
state_gdp_emissions <- left_join(emissions_v2, stategdp_v3, by = c('State','Year'))

state_gdp_emissions <- state_gdp_emissions |>
  drop_na() |>
  rename("CO2" = "CO2\r\n(Metric Tons)")
```

Research question

Are states with higher GDPs more likely to produce energy that emits CO2?

Cases

Each case represents a states annual energy usage and GDP for the respective year. There are 1326 rows in the data set and eight columns: 1. Year 2. State 3. Producer Type 4. Energy Source 5. CO2 (Metric Tons) 6. SO2 (Metric Tons) 7. NOx (Metric Tons) 8. GDP

Data collection

This project uses data from the Department of Energy's (DOE) Annual Electric Power Report and the U.S. Department of Commerce (DOC) State Annual Summary Statistics.

The Annual Electric Power Report is prepared by the Office of Energy Production, Conversion, and Delivery (EPCD), within the U.S. Energy Information Administration in DOE. Data in the report is provided directly by respondents into DOE's information systems. For more information about the data collection process, please see the *Data Quality and Submission* section of the Electric Power Annual Report: eia.gov/electricity/annual/pdf/epa.pdf.

The State Annual Summary Statistics provided by DOC are estimated using two data sources: (1) wages and salaries data from the Bureau of Labor Statistics and (2) value-added, receipts, and payroll data from the Census Bureau's economic censuses. These data sources are then used to estimate the State GDP following the estimation methodology outlined in the DOC's Gross Domestic Product by State Estimation Methodology. For more information please see the methodology: bea.gov/sites/default/files/methodologies/0417_GDP_by_State_Methodology.pdf.

Type of study

This is an observational study.

Data Sources

1. DOC/BEA's GDP and Personal Income by State
2. DOE/EIA's Electric Power Industry Esimated Emissions by State

Response

The response variable is state GDP and is numerical.

Explanatory

The explanatory variable is CO2 Emissions from Energy Production and is numerical.

Relevant summary statistics

```
describe(state_gdp_emissions)
```

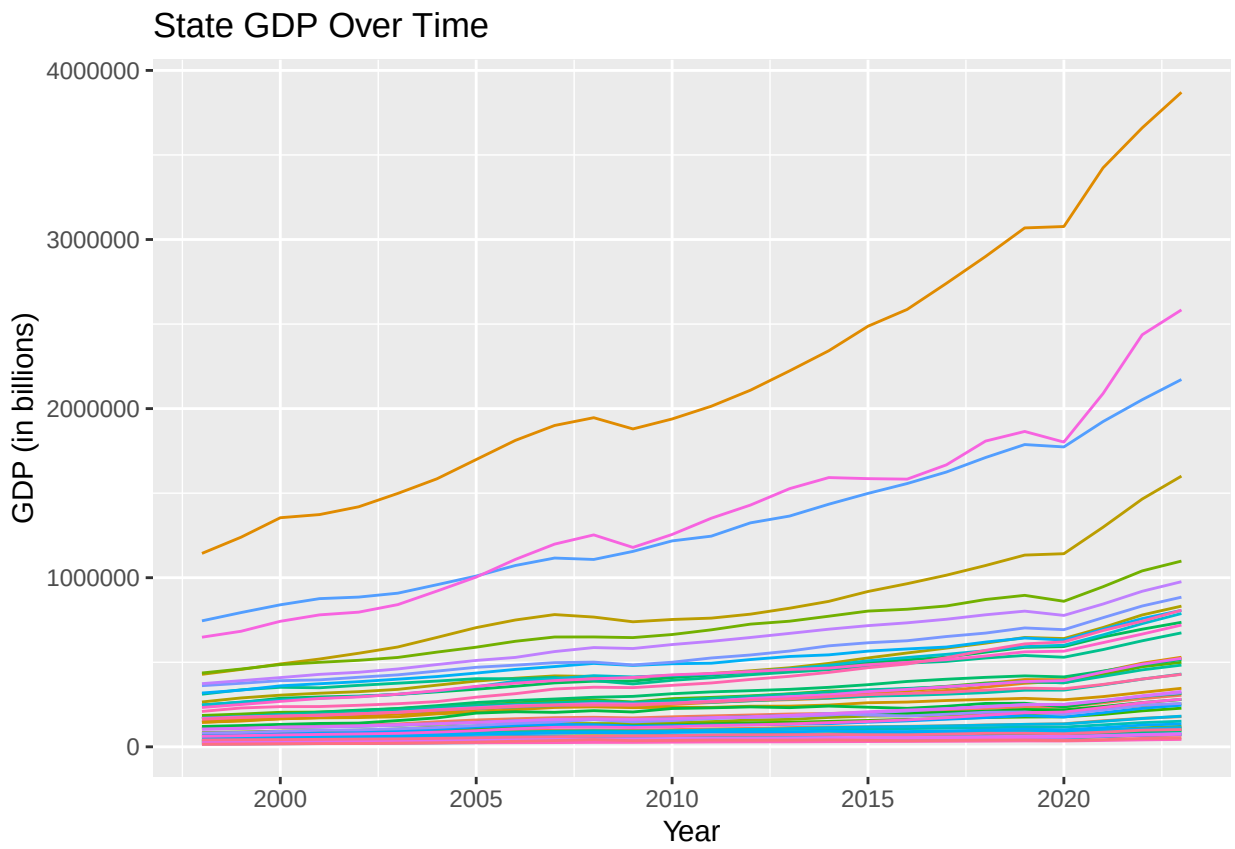
```
## state_gdp_emissions
##
##  8  Variables      1300  Observations
## ---------------------------------------------------------------------------
## Year
##        n  missing distinct     Info     Mean  pMedian      Gmd      .05
##     1300        0       26    0.999     2010     2010    8.661     1999
##      .10      .25      .50      .75      .90      .95
##     2000     2004     2010     2017     2021     2022
##
## lowest : 1998 1999 2000 2001 2002, highest: 2019 2020 2021 2022 2023
## ---------------------------------------------------------------------------
## State
##        n  missing distinct
##     1300        0       50
##
## lowest : AK AL AR AZ CA, highest: VT WA WI WV WY
## ---------------------------------------------------------------------------
## Producer Type
##                              n                        missing
##                           1300                              0
##                       distinct                          value
##                              1 Total Electric Power Industry
##
## Value      Total Electric Power Industry
## Frequency                           1300
## Proportion                             1
## ---------------------------------------------------------------------------
## Energy Source
##           n     missing    distinct       value
##        1300           0           1 All Sources
##
## Value      All Sources
## Frequency         1300
## Proportion           1
## ---------------------------------------------------------------------------
## CO2
##         n   missing  distinct      Info      Mean   pMedian       Gmd       .05
##      1300         0      1300         1  43277808  37130691  41370454   2410656
##       .10       .25       .50       .75       .90       .95
##   3610522  13537718  33019610  57791802  89115625 119709509
##
## lowest :      6583      6733      7098      8016      8209
## highest: 259415102 260213902 261332154 264816156 267464092
## ---------------------------------------------------------------------------
## SO2
## (Metric Tons)
##        n  missing distinct     Info     Mean  pMedian      Gmd      .05
##     1300        0     1283        1   121374    65259   165607    703.6
##      .10      .25      .50      .75      .90      .95
##   2653.7  11865.0  43995.0 125364.8 357329.4 558374.6
##
## lowest :      28      32      34      36      37
## highest: 1095719 1140670 1152407 1222226 1321325
```

```
## ------------------------------------------------------------------------------
## NOx
## (Metric Tons)
##        n  missing distinct     Info     Mean  pMedian      Gmd      .05
##     1300        0     1293        1    60188    46308    64043     2270
##      .10      .25      .50      .75      .90      .95
##     6041    16029    39792    77150   145981   203583
##
## lowest :    409    419    457    484    487, highest: 403364 483133 509777 510931 531361
## ------------------------------------------------------------------------------
## GDP
##        n  missing distinct     Info     Mean  pMedian      Gmd      .05
##     1300        0     1300        1   321806   233110   356429    32552
##      .10      .25      .50      .75      .90      .95
##    43450    74347   189134   389481   703112  1107879
##
## lowest : 14833.2 15685.7 16197.7 16842.7 17083.7
## highest: 3068630 3076760 3423960 3660420 3870380
## ------------------------------------------------------------------------------
```

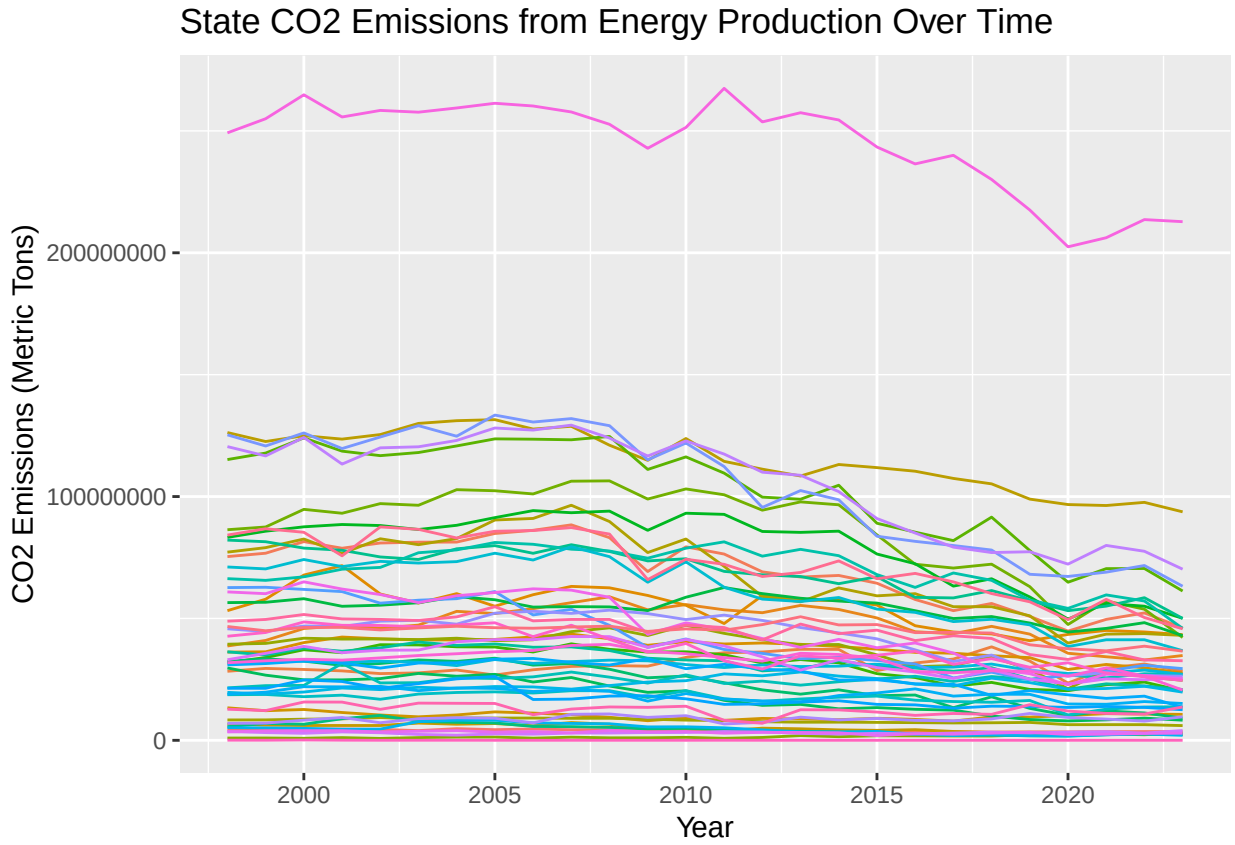
```
#Findings: DC may need to be excluded from this evaluation due to missing data.

options(scipen=999)

ggplot(state_gdp_emissions, aes(x=Year, y=GDP, color=State))+
  geom_line()+
  labs(title ="State GDP Over Time",
       x="Year",
       y= "GDP (in billions)")+
  theme(legend.position="none")
```




```
ggplot(state_gdp_emissions, aes(x=Year, y=`CO2`, color=State))+
  geom_line()+
  labs(title ="State CO2 Emissions from Energy Production Over Time",
       x="Year",
       y= "CO2 Emissions (Metric Tons)")+
   theme(legend.position="none")
```



The graphs above are not very helpful if they are not interactive, therefore, I am providing the plots that are interactive and if you hover you can see summary information. Note, the interactive plots are only accessible in HTML files and within the Rmd.

```
# ggplotly(gdp_plot)
#
# ggplotly(co2_plot)
```

```
state_gdp_emissions |>
summarise(median_GDP = median(GDP),
            mean_GDP = mean(GDP),
            max_GDP = max(GDP),
            iqr_GDP = IQR(GDP),
            sd_GDP = sd(GDP)
  )
```

```
## # A tibble: 1 x 5
##   median_GDP mean_GDP  max_GDP iqr_GDP  sd_GDP
##        <dbl>    <dbl>    <dbl>   <dbl>   <dbl>
## 1    189134.  321806. 3870379. 315134. 427444.
```

```
state_gdp_emissions |>
summarise(median_CO2 = median(CO2),
            mean_CO2 = mean(CO2),
            max_CO2 = max(CO2),
```

```
            iqr_CO2 = IQR(CO2),
            sd_CO2 = sd(CO2)
  )
```

```
## # A tibble: 1 x 5
##   median_CO2  mean_CO2   max_CO2  iqr_CO2    sd_CO2
##        <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
## 1  33019610. 43277808. 267464092 44254084 42624868.
```

```
length(unique(state_gdp_emissions$State))
```

```
## [1] 50
```