project data regions take 2

Abby Durrant

3/16/2022

```
library(knitr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(moderndive)
```

```
library(tidyverse)
df_epa <- read_csv("https://reed-statistics.github.io/math141s22-wells-website/projects/epa-emissions/gl
```

```
## Rows: 8301 Columns: 38
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (9): Facility.Name, City, State, Zip.Code, Address, County, Category, I...
## dbl (29): Facility.Id, FRS.Id, Latitude, Longitude, Primary.NAICS.Code, Emis...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

For this research question, start investigating the question by: Producing useful summaries of the variables and their relationships. Graphing each variable and the relationships between variables. Completing any useful data wrangling.

In an Rmd file, write a two page summary that: States your research question and some initial answers/findings related to the questions Introduces the data and addresses what/who the data represent (for your variables of interest) Presents at least three summary statistics and discusses what they suggest about the data. Presents at least three data visualizations and discusses what they suggest about the data. Includes your R code.

Research Question Summary

Research Question and variables

A research questions we asked is whether or not the region the facility is in has any correlation to their emissions in the most recent years. To do this, we had to begin by defining what a "region" is. We found the United States Census Bureau has defined four distinct regions that each state falls into, with no overlap (ex. Maryland is in the south region and now the south and the northeast). Using the "mutate" function, we

created a new variable that categorizes the facilities using the given "State" column. Because there is not another variable we are looking at besides the categorical region and qualitative emissions, a histogram was the best way to look at the data. The 2018 emissions numbers were used to show the distribution of facilities within a region by their emissions numbers and how many facilities the given region has. The average amount of emissions per facility in a given region was also calculated.

Initial findings

Several things were found in this first glance at the data. First, the region with the most observations, thus facilities, is the South, with 2129 observations. The average amount of emissions for the south is also higher than the other 3 regions.

Looking at the graphs, we can see a heavy right skew, meaning there are far more "smaller" facilities contributing a smaller amount o emissions with a few bigger plants that contribute a bigger amount of emissions.

This comparison of the the four specific regions makes a few points very clear. The south takes the cake with the largest emissions average from 2018, over the second place (the Midwest) by about 80,000.

Adding regionality to the data using the most recent region specifications from the Census Bureau.

```
df_epa_region <- df_epa %>%
  mutate(
    Region = case_when(
      State %in% c("ME", "NH", "VT", "MA", "CT", "RI",
                   "NY", "PA", "NJ") ~ "Northeast",
      State %in% c("MD", "DE", "WV", "DC", "VA", "NC", "SC", "GA", "FL",
                   "KY", "TN", "MS", "AL",
                   "OK", "AR", "LA", "TX") ~ "South",
      State %in% c("WI", "MI", "IL", "IN", "OH",
                   "ND", "MN", "SD", "IA", "NE", "MO", "KS") ~ "Midwest",
      State %in% c("MT", "ID", "WY", "NV", "UT", "CO", "AZ", "NM",
                   "WA", "OR", "CA", "AK", "HI") ~ "West")) %>%
  drop_na()
```

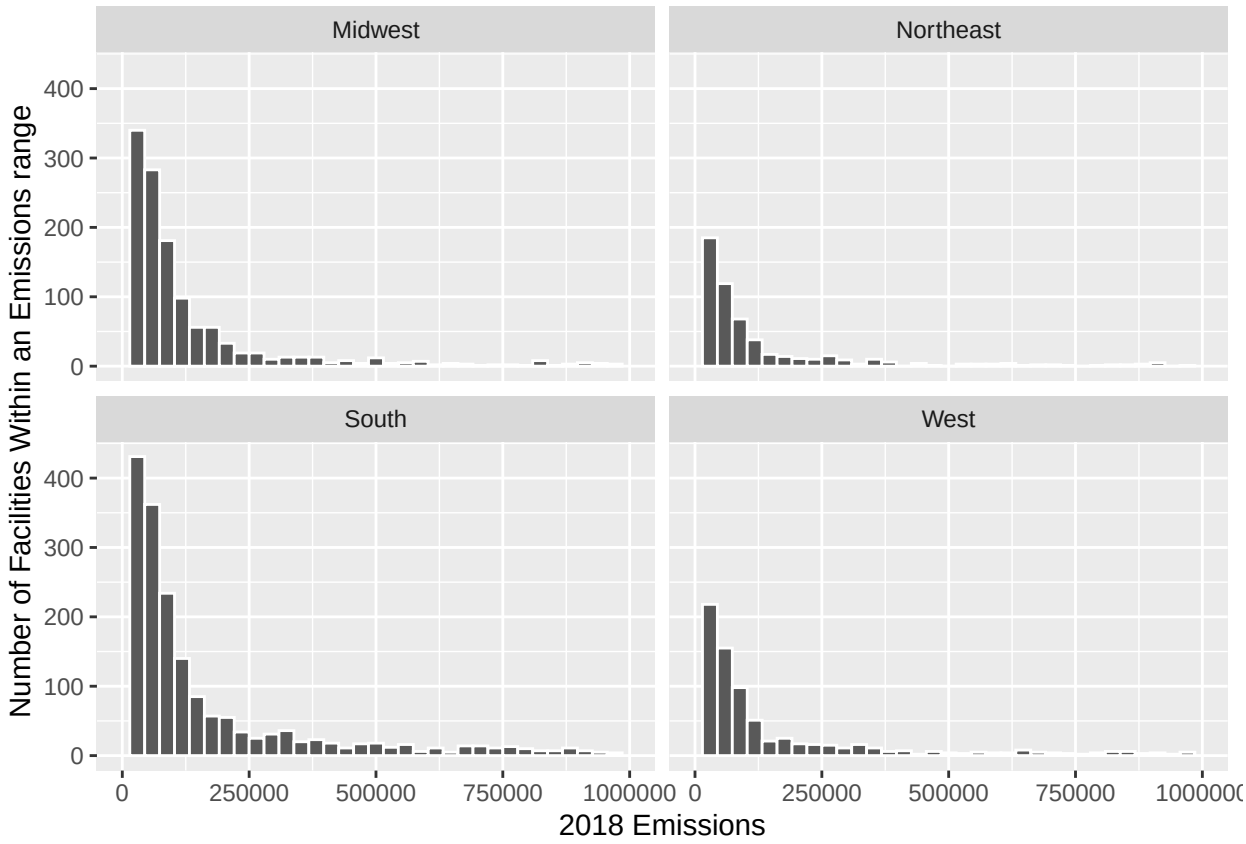
This data represents the given regions within the United States that various states can fall into

Graphing the regions apart to compare their specific facilities and their emissions

```
ggplot(df_epa_region, aes(x = Emissions.2018)) + geom_histogram(bins = 35, color = "white") + facet_wra
)) + labs(x = "2018 Emissions", y = "Number of Facilities Within an Emissions range")
```

```
## Warning: Removed 556 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 8 rows containing missing values (geom_bar).
```



Seperate the four regions to compare their distributions and their average emissions to compare emissions amoung the regions.

```
epa_midwest <- df_epa_region %>%
  filter(Region == "Midwest")
epa_south <- df_epa_region %>%
  filter(Region == "South")
epa_northeast <- df_epa_region %>%
  filter(Region == "Northeast")
epa_west <- df_epa_region %>%
  filter(Region == "West")
```

```
mean(epa_midwest$Emissions.2018)
```

```
## [1] 513763.1
```

```
mean(epa_northeast$Emissions.2018)
```

```
## [1] 278132.8
```

```
mean(epa_south$Emissions.2018)
```

```
## [1] 586854
```

```
mean(epa_west$Emissions.2018)
```

```
## [1] 402412.3
```

```
sd(epa_midwest$Emissions.2018)
```

```
## [1] 1591370
```

```
sd(epa_northeast$Emissions.2018)
```

```
## [1] 807806.5
```

```
sd(epa_south$Emissions.2018)
```

```
## [1] 1529476
```

```
sd(epa_west$Emissions.2018)
```

```
## [1] 1161303
```

Combined Emissions

This is to visualize all of the combined emissions based on regions.

```
df_epa_region <- df_epa_region %>% mutate(Cumulative_Emissions = if_else(is.na(Emissions.2018), 0, Emiss
```

```
ggplot(df_epa_region, aes(x = Cumulative_Emissions, y = Region)) + geom_boxplot(alpha = .5) + labs(x = "
```