Mini Project 2 : Data wrangling and visualization

Contents

Administrative Notes

- **Peer graded Presentations**: Day 3, PM Sessions
- **Learning Goal**: To make informative, appropriate, and compelling data graphics after reshaping the data

The Details

You will work with a partner or two to write a short blog post that contains at least one data graphic. Your goal is to tell us something interesting using a well-crafted, thoughtfully-prepared data graphic. One data graphic should suffice, but you may include more if you choose.

Your blog post should be short. We envision an introductory paragraph that explains your findings and provides some context to your data, the data graphic, and then a caption-like paragraph providing more detail about what to look for in the data graphic and how to interpret it. That is it. You will not earn more points by including more words or data graphics. What we are looking for is something that is insightful and well-crafted. As always, you should spend some time thinking about context, scale, color, etc.

Here are some examples of articles that are similar in spirit to yours. Most of these are much longer than yours will be, and may contain multiple graphics, but the idea is similar: use a good data graphic to tell us something we don't already know.

- How to Tell Someone's Age When All You Know Is Her Name
- A Better Way To Find The Best Flights And Avoid The Worst Airports
- NYC Taxis and Uber
- Data on people who went to ER for wall-punching

The Data

We will be using the data from the Henry J Kaiser Family Foundation (KFF)

1. Health Insurance Coverage of the Total Population - Includes years 2013-2016
2. Health Care Expenditures by State of Residence (in millions) Includes years 1991-2014

Data is already in 1_Data folder.

The Questions

1. Is there a relationship between healthcare coverage and healthcare spending in the United States?
2. How does the spending distribution change across geographic regions in the United States?
3. Does the relationship between healthcare coverage and healthcare spending in the United States change from 2013 to 2014?

Build your solution with any tool you like. This could be a single graph, multiple graphs organized in your preferred layout or even an interactive dashboard if you prefer. Remember that your primary goal is effectiveness.

What to do

1. As a group:
    - Describe your data briefly in a textual manner.
    - Do not forget to mention how you dealt with missing data (if any).
    - Try to think of one or two research questions from the data. For the healthcare data, you may combine the coverage and spending for relevant years and formulate some questions you think would be worth exploring.
    - Clean or reshape the data by all the necessary transformations to get a final dataset that you will use to generate the diagrams. Think of graphs or charts that will throw some light in the direction of the research questions you posed.
    - If you are using data that is included in a separate file (like an Excel or `.csv` file), submit this too. If you do not, we will not be able to *reproduce and replicate* your results.
    - Try to sum up your findings in a short paragraph in the end.
2. Make sure to avoid the normal Data Viz mistakes
3. You can review Rmarkdown syntax here:

- SWC: Rmarkdown overview

Presentation

1. You will share with me your Project via RStudio cloud and you will present it during the camp
2. Others are to provide feedback on projects.

Step 0: Load necessary libraries and datasets

Load libraries that we use and the two healthcare datasets specified for the project. We spent some time on the website reading about what's included in the datasets.

```
library(tidyverse)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang

## -- Attaching packages ------------------------------------------------------------------------- tidy

## v ggplot2 3.1.1     v purrr   0.3.2
## v tibble  2.1.1     v dplyr   0.8.1
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

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

```
library(dplyr)
library(ggplot2)
library(modelr)
library(broom)
```

```
##
## Attaching package: 'broom'
```

```
## The following object is masked from 'package:modelr':
##
##     bootstrap
```

```
spend <- read.csv("../1_Dataset/hcare_spend.csv")
cov <- read.csv("../1_Dataset/hcare_cov.csv")
```

Step 1: Tidy data into more accessible form

Using the tidyverse functions that we've learned, clean the two datasets so that data analysis wolud be more efficient.

```
spending <- spend %>%
  gather("year", "total_spending", -1) %>%
  mutate(year = as.numeric(substr(year, 2, 5)), total_spending = as.numeric(total_spending)) %>%
  rename(state = Location)

coverage <- cov %>%
  mutate_all(as.character) %>%
  gather("category", "amount", -1) %>%
  mutate(amount = as.numeric(ifelse(amount == "N/A", 0, amount)),
         year = as.numeric(str_sub(category, 2, 5)),
         category = str_sub(category, 8)) %>%
  rename(state = Location) %>%
  select(state, year, category, amount) %>%
  spread(category, amount) %>%
  rename(employer = Employer, medicaid = Medicaid, medicare = Medicare, non_group = Non.Group, other_pul
```

#Step 2: Join two tables Inner join spending and coverage tables on two variables: state, year. The only overlap were for the years 2013 and 2014. We computed three new variables: *gov_coverage: number of people whose insurance is government-funded (medicare, medicaid, other_public)* gov_cov_percent: gov_coverage as a percent of the total population (both insured and uninsured) *spending_per_gov_cap: per capita government spending on healthcare (only includes people whose insured is government funded) "United States" is a case in data, so we filter it out.

```
health <- spending %>%
  inner_join(coverage, by = c("state", "year")) %>%
  mutate(gov_coverage = medicare + medicaid + other_public,
         gov_cov_percent = gov_coverage/(total_coverage + uninsured),
         spending_per_gov_cap = total_spending/gov_coverage)
```

```
## Warning: Column `state` joining factor and character vector, coercing into
## character vector
```

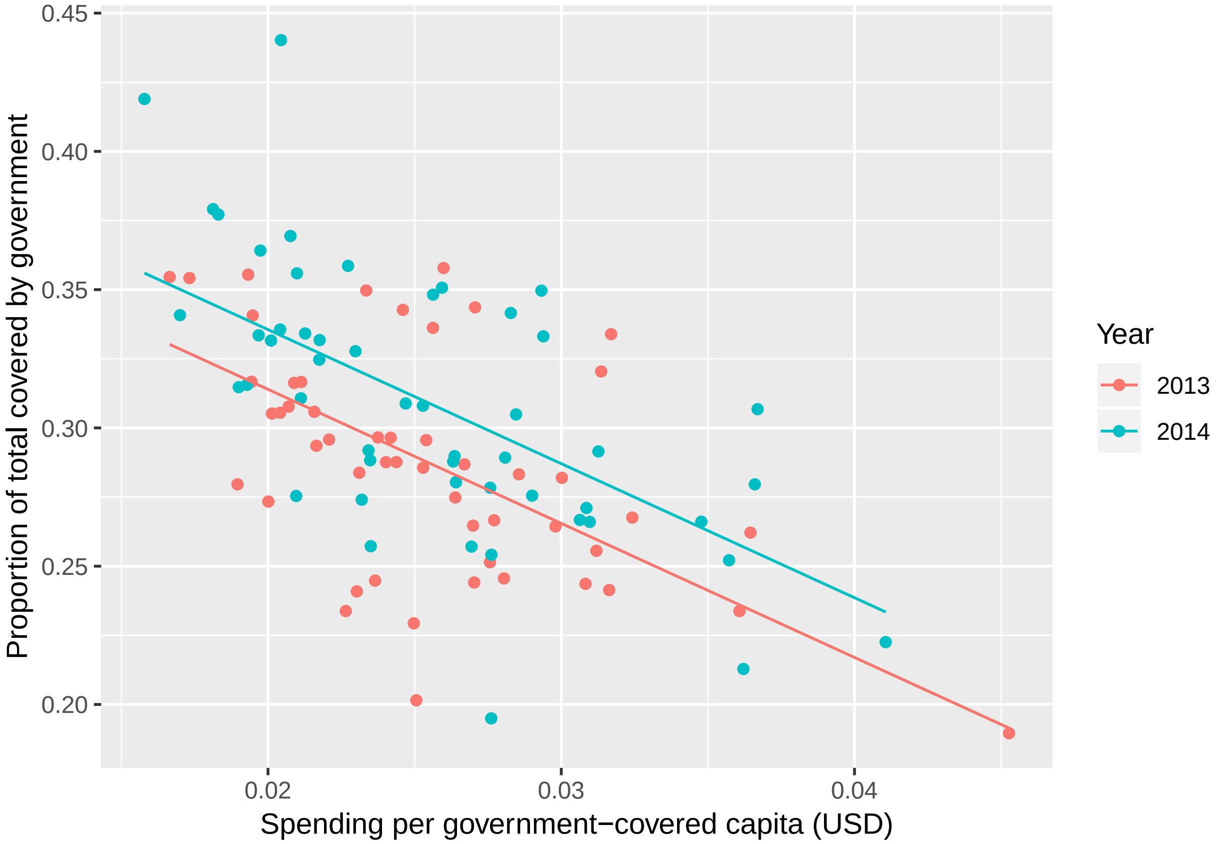
```
health$year <- factor(health$year)

health_states <- filter(health, state != "United States")
```

#Step 3: Graphing Create linear model of gov_cov_percent as a function of spending_per_gov_cap and the year, plot alongside data points. Found that they are negatively correlated, which we interpreted as: a state that spends more money on healthcare per person covered covers a smaller proportion of its population.

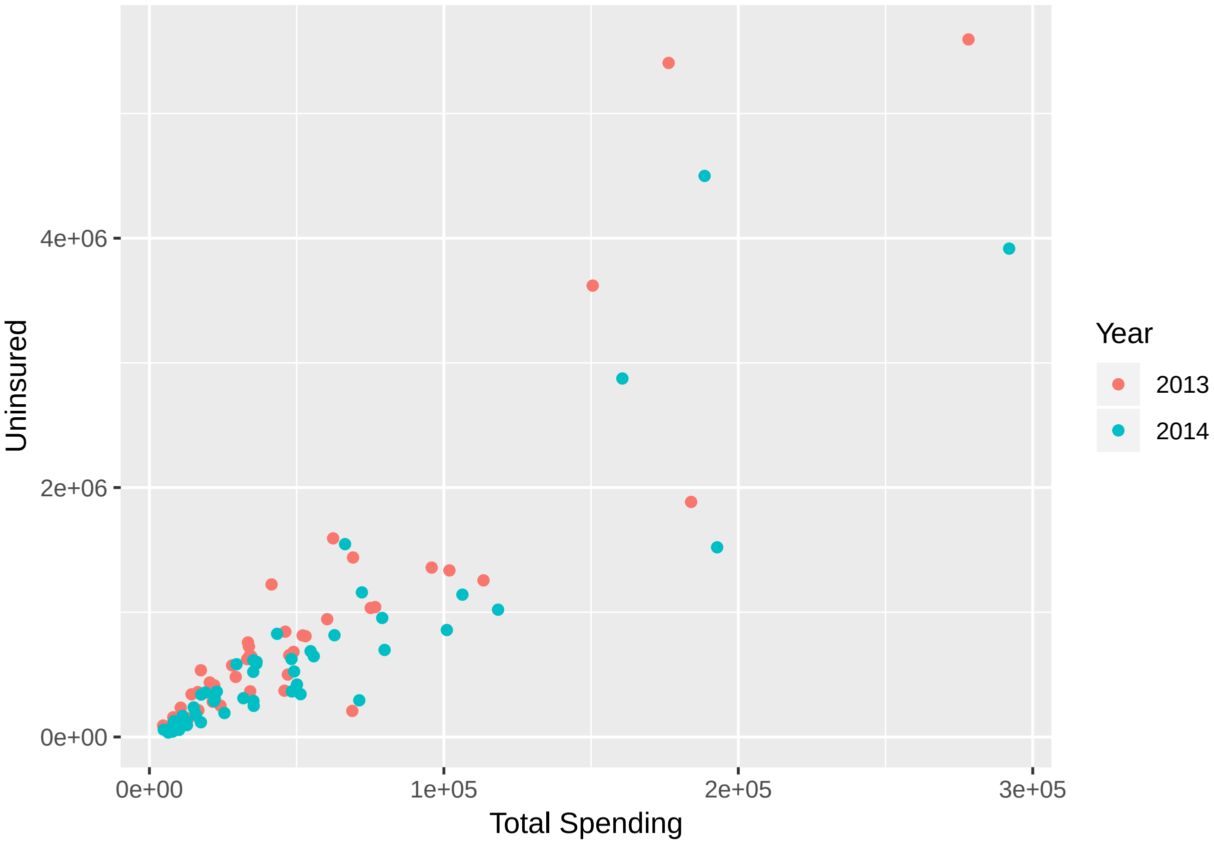
```
mod_h <- lm(gov_cov_percent ~ spending_per_gov_cap + year, data = health_states)

mod_h %>%
  augment(data = health_states) %>%
  ggplot(mapping = aes(x = spending_per_gov_cap, color = year)) +
  geom_point(mapping = aes(y = gov_cov_percent)) +
  geom_line(mapping = aes(y = .fitted)) +
  labs(x = "Spending per government-covered capita (USD)", y = "Proportion of total covered by governmen
```



We had found a positive correlation between total spending and total coverage, and also a positive correlation between total spending and the number of people uninsured. We think this is due to variations in state population. This led us to our analysis based on proportion of population covered as opposed to absolute numbers.

```
filter(health, state != "United States") %>%
  ggplot(mapping = aes(x = total_spending, y = uninsured, color = year)) +
  geom_point() +
  labs(x = "Total Spending", y = "Uninsured", color = "Year")
```


Uninsured
4e+06
2e+06
0e+00
0e+00
1e+05
2e+05
3e+05
Total Spending
Year
● 2013
● 2014