Lab 4b - Wrangling

Sebastian Montesinos

For class Thursday, Feb. 24

Lab Purpose

The text and our examples covered a lot of different wrangling commands. Previously, you also identified what wrangling commands were doing in provided code. Now, it's your turn to generate the necessary code to do the wrangling and make some neat visualizations.

A big part of wrangling data is figuring out what your data needs to look like to generate your desired output. You'll need to tackle that step before you can implement the wrangling itself in R. If you get stuck, I suggest sketching out what your data set needs to look like on paper. What are your observations and what variables do you need? It really does help.

In this lab we will work with the **tidyverse** and **janitor** packages for our wrangling. In addition, the **datasets** package contains a dataset `state` with information on each state such as region that will be useful for us to pull data in from.

You'll work on the lab in the company of classmates to help each other with the challenges of wrangling.

Setting - Exploring Health Expenditure using State-level data

This case study is based on an open case study from the OCS project (Kuo et al. 2019).

Health policy in the United States is complicated, and several forms of health care coverage exist. Various mandates exist and there are many different coverage options available. In this lab, our goal is for us to have a general idea about health care economics in the United States. Our focus will be on health care expenditure, including health care coverage and health care spending, across the United States.

Motivating questions:

- Is there a relationship between health care spending and health care coverage by employers in the United States?
- How does the spending distribution change across geographic regions in the United States?
- Does the relationship between health care coverage and health care spending in the United States change from 2013 to 2014?

Data for this lab come from the Henry J Kaiser Family Foundation (KFF).

- Health Insurance Coverage of the Total Population (2013 – 2016)
- Health Care Expenditures in millions by State of Residence (1991 – 2014)

1 - Understanding the Data

Since our goal is to get a sense of the health expenditure, including health care coverage and health care spending, across states, it would be nice add some information about each state. Namely, the state abbreviation and state region (i.e. north, south, etc). For this we use the various state datasets in the **datasets** R package. Since the package is already loaded, we can refer directly to any of the state datasets (e.g., `state.abb`) even though we don't them loaded in our environment. However, we can make the `state` datasets appear in our environment by running `data(state)`.

```
# Load state datasets into environment
data(state)
```

The state data are split across 7 datasets, all arranged according to alphabetical order of the state names. There are no other variables that can link the datasets together, so we will trust the alphabetical ordering and create our own dataframe from three of the datasets.

```
# Create a data frame with state info
state_info <- data.frame(location = state.name,
                         abbreviation = state.abb,
                         region = state.region)
```

> part a - Run the code below to use `read_csv()` to read in the files containing the healthcare coverage and healthcare spending data. Pay attention to the filepath, making modifications if needed based on your own file organization.

If you copied over the Lab 4b folder in its entirety, the file structure should remain the same. If you pulled individual files over, you may need to adjust. In particular, be sure you pulled the data folder or its files. These files aren't hosted via a web url.

Solution:

```
coverage <- read_csv("data/healthcare_coverage.txt")
spending <- read_csv("data/healthcare_spending.txt")
```

> part b - Now get acquainted with the `coverage` and `spending` datasets. What years are covered in the `coverage` dataset? What years are covered in the `spending` dataset? Are there any mismatches between how R specified the variable type and what you expected the type would be?

(Yes, the answers to these questions are above, but how can you confirm this in the datasets?)

Hint: use `glimpse()`.

Solution: Spending covers 1991 to 2014, while coverage covers 2013 to 2016. There are some N/A values in the coverage dataset, where there only should be doubles.

```
glimpse(coverage)
```

```
Rows: 52
Columns: 29
$ Location             <chr> "United States", "Alabama", "Alaska", "Arizona", ~
$ `2013__Employer`     <dbl> 155696900, 2126500, 364900, 2883800, 1128800, 177~
$ `2013__Non-Group`    <dbl> 13816000, 174200, 24000, 170800, 155600, 1986400,~
```

```
$ `2013__Medicaid`     <dbl> 54919100, 869700, 95000, 1346100, 600800, 8344800~
$ `2013__Medicare`     <dbl> 40876300, 783000, 55200, 842000, 515200, 3828500,~
$ `2013__Other Public` <chr> "6295400", "85600", "60600", "N/A", "67600", "675~
$ `2013__Uninsured`    <dbl> 41795100, 724800, 102200, 1223000, 436800, 559410~
$ `2013__Total`        <dbl> 313401200, 4763900, 702000, 6603100, 2904800, 381~
$ `2014__Employer`     <dbl> 154347500, 2202800, 345300, 2835200, 1176500, 177~
$ `2014__Non-Group`    <dbl> 19313000, 288900, 26800, 333500, 231700, 2778800,~
$ `2014__Medicaid`     <dbl> 61650400, 891900, 130100, 1639400, 639200, 961880~
$ `2014__Medicare`     <dbl> 41896500, 718400, 55300, 911100, 479400, 4049000,~
$ `2014__Other Public` <chr> "5985000", "143900", "37300", "N/A", "82000", "63~
$ `2014__Uninsured`    <dbl> 32967500, 522200, 100800, 827100, 287200, 3916700~
$ `2014__Total`        <dbl> 316159900, 4768000, 695700, 6657200, 2896000, 387~
$ `2015__Employer`     <dbl> 155965800, 2218000, 355700, 2766500, 1293700, 177~
$ `2015__Non-Group`    <dbl> 21816500, 291500, 22300, 278400, 200200, 3444200,~
$ `2015__Medicaid`     <dbl> 62384500, 911400, 128100, 1711500, 641400, 101381~
$ `2015__Medicare`     <dbl> 43308400, 719100, 60900, 949000, 484500, 4080100,~
$ `2015__Other Public` <chr> "6422300", "174600", "47700", "189300", "63700", ~
$ `2015__Uninsured`    <dbl> 28965900, 519400, 90500, 844800, 268400, 2980600,~
$ `2015__Total`        <dbl> 318868500, 4833900, 705300, 6739500, 2953000, 391~
$ `2016__Employer`     <dbl> 157381500, 2263800, 324400, 3010700, 1290900, 181~
$ `2016__Non-Group`    <dbl> 21884400, 262400, 20300, 377000, 252900, 3195400,~
$ `2016__Medicaid`     <dbl> 62303400, 997000, 145400, 1468400, 618600, 985380~
$ `2016__Medicare`     <dbl> 44550200, 761200, 68200, 1028000, 490000, 4436000~
$ `2016__Other Public` <chr> "6192200", "128800", "55600", "172500", "67500", ~
$ `2016__Uninsured`    <dbl> 28051900, 420800, 96900, 833700, 225500, 3030800,~
$ `2016__Total`        <dbl> 320372000, 4834100, 710800, 6890200, 2945300, 391~
```

```
glimpse(spending)
```

```
Rows: 52
Columns: 25
$ Location                       <chr> "United States", "Alabama", "Alaska", "A~
$ `1991__Total Health Spending`  <dbl> 675896, 10393, 1458, 9269, 5632, 81438, ~
$ `1992__Total Health Spending`  <dbl> 731455, 11284, 1558, 9815, 6022, 87949, ~
$ `1993__Total Health Spending`  <dbl> 778684, 12028, 1661, 10655, 6397, 91963,~
$ `1994__Total Health Spending`  <dbl> 820172, 12742, 1728, 11364, 6810, 94245,~
$ `1995__Total Health Spending`  <dbl> 869578, 13590, 1879, 12042, 7343, 96870,~
$ `1996__Total Health Spending`  <dbl> 917540, 14450, 2076, 12850, 7817, 100215~
$ `1997__Total Health Spending`  <dbl> 969531, 15462, 2240, 13418, 8393, 103681~
$ `1998__Total Health Spending`  <dbl> 1026103, 15860, 2386, 14465, 8814, 11122~
$ `1999__Total Health Spending`  <dbl> 1086280, 16451, 2569, 15550, 9407, 11603~
$ `2000__Total Health Spending`  <dbl> 1162035, 17504, 2867, 16646, 10009, 1216~
$ `2001__Total Health Spending`  <dbl> 1261944, 18619, 3276, 18129, 10846, 1323~
$ `2002__Total Health Spending`  <dbl> 1367628, 20209, 3642, 20390, 11797, 1438~
$ `2003__Total Health Spending`  <dbl> 1477697, 22491, 3955, 22464, 12578, 1582~
$ `2004__Total Health Spending`  <dbl> 1587994, 23797, 4256, 24795, 13470, 1700~
$ `2005__Total Health Spending`  <dbl> 1696222, 25338, 4765, 28190, 14611, 1829~
$ `2006__Total Health Spending`  <dbl> 1804672, 26638, 5048, 30766, 15431, 1944~
$ `2007__Total Health Spending`  <dbl> 1918820, 27700, 5426, 33366, 16426, 2093~
$ `2008__Total Health Spending`  <dbl> 2010690, 28765, 5807, 35547, 17246, 2210~
$ `2009__Total Health Spending`  <dbl> 2114221, 30095, 6112, 37258, 18071, 2295~
$ `2010__Total Health Spending`  <dbl> 2194625, 30728, 6519, 38620, 18735, 2419~
$ `2011__Total Health Spending`  <dbl> 2272582, 31398, 6928, 39295, 19356, 2538~
$ `2012__Total Health Spending`  <dbl> 2365948, 32848, 7406, 40495, 20076, 2667~
```

```
$ `2013__Total Health Spending` <dbl> 2435624, 33788, 7684, 41481, 20500, 2781~
$ `2014__Total Health Spending` <dbl> 2562824, 35263, 8151, 43356, 21980, 2919~
```

part c - The previous question was intentionally leading—you should have identified some mismatched variable types in the `coverage` dataset. This happened because missing numeric values were recorded as text ("N/A") instead of left empty. Run the code below to fix this problem. Then, in your own words, describe what the various commands (na_if, mutate, across) did.

```
coverage <- coverage %>%
  na_if("N/A") %>%
  mutate(across(.cols = ends_with("Public"),
                .fns = as.numeric))
```

Solution: R isn't recognizing the n/a as NA in the actual dataset, so na_if finds N/As and replaces them with NAs. Mutate goes over every column ending with public and .cols tells mutate what column to work over, telling them to change the variable type to numeric.

part d - If we're interested in the relationship between spending and coverage, we'll only be able to use observations that have information on both. That is, we won't be using data from years for which we only have spending information or only have coverage information. Remove any variables we won't be using from `coverage` and `spending`.

Hint: the `starts_with()` function from the **tidyselect** package (already loaded) could help with efficiency here. Also, sometimes "removing" many means "keeping" a few.

Solution:

```
relcoverage <- coverage %>%
  select(Location |starts_with('2013') | starts_with('2014'))
relspending <- spending %>%
  select(Location |starts_with('2013') | starts_with('2014'))
```

part e - There are 50 states in the United States but 52 observations in the `coverage` and `spending` datasets. The two "bonus" cases contain information about the US as a whole and Washington DC. Remove these observations from both datasets.

Solution:

```
state_relcoverage <- relcoverage %>%
  filter(Location != "United States" & Location != "District of Columbia")
state_relspending <- relspending %>%
   filter(Location != "United States" & Location != "District of Columbia")
```

2 - Spending and Coverage Relationship

Is there a relationship between healthcare spending and healthcare coverage by employers in the United States?

We'll want to create a scatterplot with `log(spending)` on the x-axis and `log(employer coverage)` on the y-axis, with the points colored by year. (*Why logs?* Both these variables are right-skewed and have large outliers; feel free to check out their histograms and/or look at the un-logged scatterplot if you'd like, as well.) This is a simple enough scatterplot, but we'll need to do a bit of data tidying before the data are in an appropriate format to create the plot.

> part a - First, sketch what the scatterplot should look like on paper or your tablet or some app (what are the axes? what does each point represent?). What does your dataset need to look like in order to create the scatterplot in `ggplot()`? What will each observation (row) in the dataset represent? What variables (columns) do you need?

Solution: We will need a table with columns for year, spending, and employee coverage, with an observation for each state and year.

> part b - What are some of the steps that will need to be taken to get the data in that form?

Remember you can reference what the data sets look like currently in multiple ways.

Solution: We will need to combine the spending and coverage sets. Then, we will need to turn year into a variable and ensure that we have a variable for coverage and spending: this will require two pivots

> part c - Now implement those steps in R, tidying the dataset for plotting. After the final step, use the `clean_names()` function from the **janitor** package to clean the variable names.

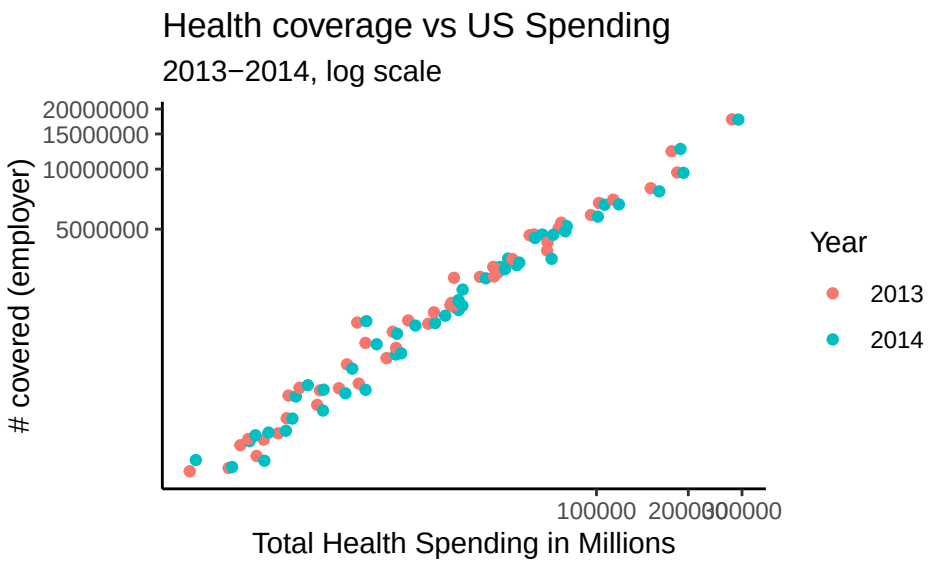
Solution:

```
healthcarejoin <- state_relcoverage %>%
  inner_join(state_relspending, by = "Location")

healthcarestep <- healthcarejoin %>%
  pivot_longer(cols = -Location,
               names_sep = "__",
               names_to = c("year", "category"),
               values_to = "amount")

healthcarewider <- healthcarestep %>%
  pivot_wider(names_from = "category",
              values_from = "amount") %>%
  clean_names()
```

> part d - Now, create the scatterplot! Describe what you see.

Solution:

```
ggplot(data = healthcarewider, aes(x = total_health_spending, y = employer, color = year)) +
  geom_point() +
  coord_trans(x = "log10", y = "log10") +
  labs(x = "Total Health Spending in Millions",
       y = "# covered (employer)",
       color = "Year",
       title = "Health coverage vs US Spending",
       subtitle = "2013-2014, log scale")
```



3 - Adjusting for population size

We see there is a strong relationship between healthcare spending and coverage within each year. However, we might suspect that health care coverage and spending are each strongly related to population size. In the `coverage` dataset, the "total" coverage category is not really a formal type of health care coverage; it actually represents the total number of people in the state in that year. This is useful information!

> part a - Using the dataset you created in Part 2, rename the `total` column to `total_population` to make the variable name more informative. Then create a scatterplot of employer coverage (y) versus population size (x).

Solution:

> part b - Now create a second scatterplot of healthcare spending vs. population size. What do you notice?

Solution:

> part c - To account for total population, create a scatterplot of spending per capita versus proportion with employer coverage. This time, *color by region* and *facet by year* (think about what additional steps you need to take to make this happen!).

Hint: The total spending column is reported in millions (`1e6`). Therefore, to calculate `spending_per_capita` we will need to adjust for this scaling factor to report it on the original scale (just dollars) and then divide by `total_population`.

Solution:

> part d - Based on your last figure, write a few sentences describing the relationship between health care spending and coverage in the US.

Solution:

4 - State and Region

How does spending vary by state and region?

> part a - Which US state spent the most per capita on health care in 2013? 2014? The least in each year?

Solution:

> part b - How does the spending distribution change across geographic region in the US? Create an appropriate figure to visualize the distribution of spending per capita on health care by region.

Solution:

> part c - Write a few sentences comparing the distributions. (Note that you probably will also want to generate summary statistics by region in order to include specific values in your summary paragraph.)

Solution:

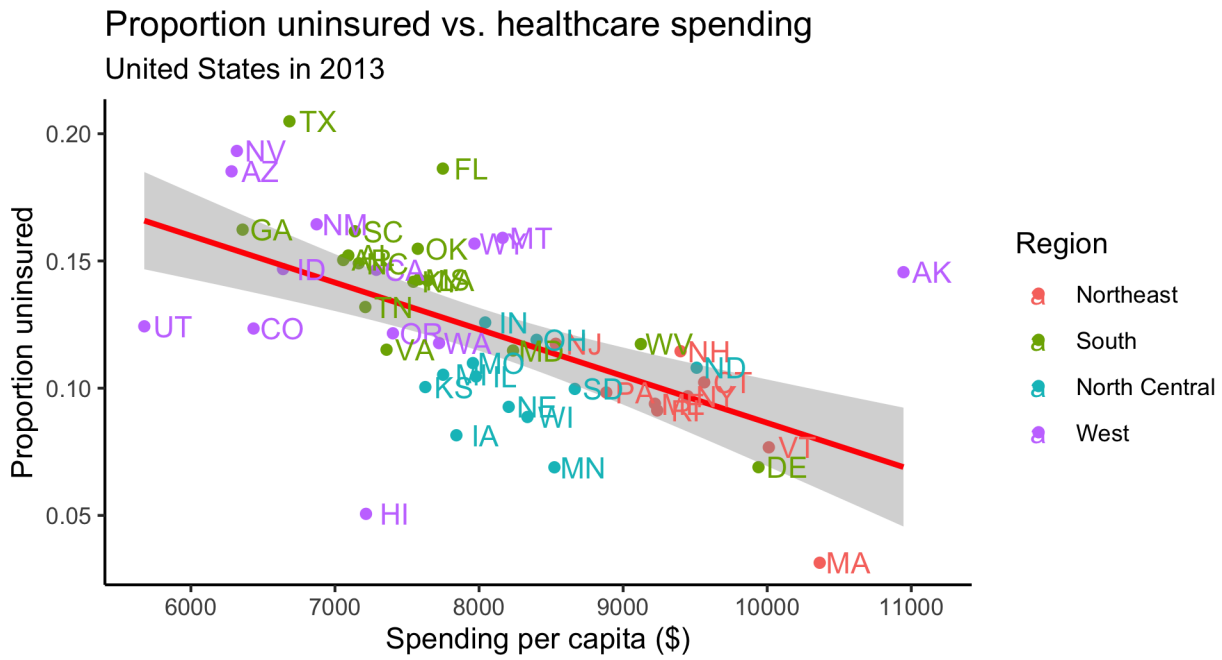
```
# Summary Statistics
```

5 - Spending and Proportion of Uninsured Individuals

Does the relationship between healthcare spending and the proportion of uninsured in the United States change from 2013 to 2014?

part a - Re-create the plot below for 2013.

Hint: use `nudge_x` and/or `nudge_y` in the `geom_text()` layer.



Solution:

part b - Next, create an analogous plot (separately) for 2014. Does the relationship between health care spending and the proportion of uninsured change from 2013 to 2014?

Solution:

part c - Now combine your two plots into one graph, creating one figure that is facetted by year and still colored by region.

Solution:

part d - Lastly, plot the points for both years on the same plot, this time colored by year instead of region. Make sure you get two lines.

Solution:

part e - Which of these three visualizations do you find most helpful for comparing the relationship between 2013 and 2014? Why?

The three visuals are those from parts a and b (together), part c, and part d.

Solution:

6 - Bonus

Done early? Try to figure out how to make these additional updates to the first figure from the last exercise to hone your plotting skills:

- remove the "a" on the points in the legend
- change the background to be grey
- make the numbers on the x-axis larger

You can also try changing fonts for your text, but this is very hard on Windows machines. For Macs, you could visit this site to learn more.

References

Kuo, Pei-Lun and Jager, Leah and Taub, Margaret and Hicks, Stephanie. (2019, February 14). opencasestudies/ocs-healthexpenditure: Exploring Health Expenditure using State-level data in the United States (Version v1.0.0). Zenodo. http://doi.org/10.5281/zenodo.2565307