Lab 5: Wrangle and tidy

Dhyey Mavani

September 14, 2021

About this lab

These exercises are designed to give you practice wrangling and tidying data, both from a *processes* perspective (what needs to be done to the data before I can run my fancy analysis and/or make my cool visualization?) and an *R implementation* perspective (how do I implement those steps specifically in R?).

In this lab we will work with the **tidyverse**, **datasets**, and **janitor** packages. The **datasets** package contains a dataset `state` with information on each state such as region.

These have been loaded for you in the `setup` code chunk, but scroll up to verify and load any packages that I may have missed (and of course install any packages that are not yet on your machine).

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, including both federal government-led health care policy, and private insurance company. Before making any inference about the relationship between health condition and health policy, it is important for us to have a general idea about health care economics in the United States. Thus, we are interested in getting a sense of the health 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?

The data

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)

Part 1 **Understanding the data**

1.1 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)
```

1.2 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)
state_info
```

```
         location abbreviation        region
1         Alabama           AL         South
2          Alaska           AK          West
3         Arizona           AZ          West
4        Arkansas           AR         South
5      California           CA          West
6        Colorado           CO          West
7     Connecticut           CT     Northeast
8        Delaware           DE         South
9         Florida           FL         South
10        Georgia           GA         South
11         Hawaii           HI          West
12          Idaho           ID          West
13       Illinois           IL North Central
14        Indiana           IN North Central
15           Iowa           IA North Central
16         Kansas           KS North Central
17       Kentucky           KY         South
18      Louisiana           LA         South
19          Maine           ME     Northeast
20       Maryland           MD         South
21  Massachusetts           MA     Northeast
22       Michigan           MI North Central
```

```
23      Minnesota           MN North Central
24    Mississippi           MS         South
25       Missouri           MO North Central
26        Montana           MT          West
27       Nebraska           NE North Central
28         Nevada           NV          West
29  New Hampshire           NH     Northeast
30     New Jersey           NJ     Northeast
31     New Mexico           NM          West
32       New York           NY     Northeast
33 North Carolina           NC         South
34   North Dakota           ND North Central
35           Ohio           OH North Central
36       Oklahoma           OK         South
37         Oregon           OR          West
38   Pennsylvania           PA     Northeast
39   Rhode Island           RI     Northeast
40 South Carolina           SC         South
41   South Dakota           SD North Central
42      Tennessee           TN         South
43          Texas           TX         South
44           Utah           UT          West
45        Vermont           VT     Northeast
46       Virginia           VA         South
47     Washington           WA          West
48  West Virginia           WV         South
49      Wisconsin           WI North Central
50        Wyoming           WY          West
```

1.3 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.

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

1.4 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? (Yes, the answers to these questions are above, but how can you confirm this in the datasets?) Are there any mismatches between how R specified the variable type and what you expected the type would be?

Using the `glimpse()` function, we can see the `coverage` dataset covers the years 2013 to 2016 and the `spending` dataset covers the years 1991 to 2014. We also see that the variables ending in "Other Public" are being treated as characters.

```
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~
```

1.5 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.

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

```
# A tibble: 52 x 29
   Location  `2013__Employer` `2013__Non-Grou~ `2013__Medicaid` `2013__Medicare`
   <chr>                <dbl>            <dbl>            <dbl>            <dbl>
 1 United S~        155696900         13816000         54919100         40876300
 2 Alabama            2126500           174200           869700           783000
 3 Alaska              364900            24000            95000            55200
 4 Arizona            2883800           170800          1346100           842000
 5 Arkansas           1128800           155600           600800           515200
 6 Californ~         17747300          1986400          8344800          3828500
 7 Colorado           2852500           426300           697300           549700
 8 Connecti~          2030500           126800           532000           475300
 9 Delaware            473700            25100           192700           141300
10 District~           324300            30400           174900            59900
# ... with 42 more rows, and 24 more variables: 2013__Other Public <dbl>,
#   2013__Uninsured <dbl>, 2013__Total <dbl>, 2014__Employer <dbl>,
#   2014__Non-Group <dbl>, 2014__Medicaid <dbl>, 2014__Medicare <dbl>,
#   2014__Other Public <dbl>, 2014__Uninsured <dbl>, 2014__Total <dbl>,
#   2015__Employer <dbl>, 2015__Non-Group <dbl>, 2015__Medicaid <dbl>,
#   2015__Medicare <dbl>, 2015__Other Public <dbl>, 2015__Uninsured <dbl>,
#   2015__Total <dbl>, 2016__Employer <dbl>, 2016__Non-Group <dbl>, ...
```

1.6 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.

```r
coverage_intersection <- coverage %>%
  select(Location, starts_with(c("2013", "2014")))
coverage_intersection
```

```
# A tibble: 52 x 15
   Location  `2013__Employer` `2013__Non-Grou~ `2013__Medicaid` `2013__Medicare`
   <chr>                <dbl>            <dbl>            <dbl>            <dbl>
 1 United S~        155696900         13816000         54919100         40876300
 2 Alabama            2126500           174200           869700           783000
 3 Alaska              364900            24000            95000            55200
 4 Arizona            2883800           170800          1346100           842000
 5 Arkansas           1128800           155600           600800           515200
 6 Californ~         17747300          1986400          8344800          3828500
 7 Colorado           2852500           426300           697300           549700
 8 Connecti~          2030500           126800           532000           475300
 9 Delaware            473700            25100           192700           141300
10 District~          324300            30400           174900            59900
# ... with 42 more rows, and 10 more variables: 2013__Other Public <dbl>,
#   2013__Uninsured <dbl>, 2013__Total <dbl>, 2014__Employer <dbl>,
#   2014__Non-Group <dbl>, 2014__Medicaid <dbl>, 2014__Medicare <dbl>,
#   2014__Other Public <dbl>, 2014__Uninsured <dbl>, 2014__Total <dbl>
```

```r
spending_intersection <- spending %>%
  select(Location, starts_with(c("2013", "2014")))
spending_intersection
```

```
# A tibble: 52 x 3
   Location             `2013__Total Health Spending` `2014__Total Health Spend~
   <chr>                                        <dbl>                      <dbl>
 1 United States                              2435624                    2562824
 2 Alabama                                      33788                      35263
 3 Alaska                                        7684                       8151
 4 Arizona                                      41481                      43356
 5 Arkansas                                     20500                      21980
 6 California                                  278168                     291989
 7 Colorado                                     34090                      36398
 8 Connecticut                                  34223                      35413
 9 Delaware                                      9038                       9587
10 District of Columbia                          7443                       7871
# ... with 42 more rows
```

1.7 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.

```r
coverage_intersection <- coverage_intersection %>%
  filter(Location != c("United States", "District of Columbia"))
coverage_intersection
```

```
# A tibble: 50 x 15
   Location    `2013__Employer` `2013__Non-Grou~ `2013__Medicaid` `2013__Medicare`
   <chr>                  <dbl>            <dbl>            <dbl>            <dbl>
 1 Alabama              2126500           174200           869700           783000
 2 Alaska                364900            24000            95000            55200
 3 Arizona              2883800           170800          1346100           842000
 4 Arkansas             1128800           155600           600800           515200
 5 California          17747300          1986400          8344800          3828500
 6 Colorado             2852500           426300           697300           549700
 7 Connecticut          2030500           126800           532000           475300
 8 Delaware              473700            25100           192700           141300
 9 Florida              8023400           968200          3190900          3108800
10 Georgia              4700500           401600          1503000          1280400
# ... with 40 more rows, and 10 more variables: 2013__Other Public <dbl>,
#   2013__Uninsured <dbl>, 2013__Total <dbl>, 2014__Employer <dbl>,
#   2014__Non-Group <dbl>, 2014__Medicaid <dbl>, 2014__Medicare <dbl>,
#   2014__Other Public <dbl>, 2014__Uninsured <dbl>, 2014__Total <dbl>
```

```
spending_intersection <- spending_intersection %>%
  filter(Location != c("United States", "District of Columbia"))
spending_intersection
```

```
# A tibble: 50 x 3
   Location    `2013__Total Health Spending` `2014__Total Health Spending`
   <chr>                               <dbl>                         <dbl>
 1 Alabama                             33788                         35263
 2 Alaska                               7684                          8151
 3 Arizona                             41481                         43356
 4 Arkansas                            20500                         21980
 5 California                         278168                        291989
 6 Colorado                            34090                         36398
 7 Connecticut                         34223                         35413
 8 Delaware                             9038                          9587
 9 Florida                            150547                        160624
10 Georgia                             62399                         66447
# ... with 40 more rows
```

Part 2 **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.

2.1 First, sketch what the scatterplot should look like on paper or Google jamboard or some other 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?

The x-axis must be spending, and y-axis must be coverage by employers (statewise)

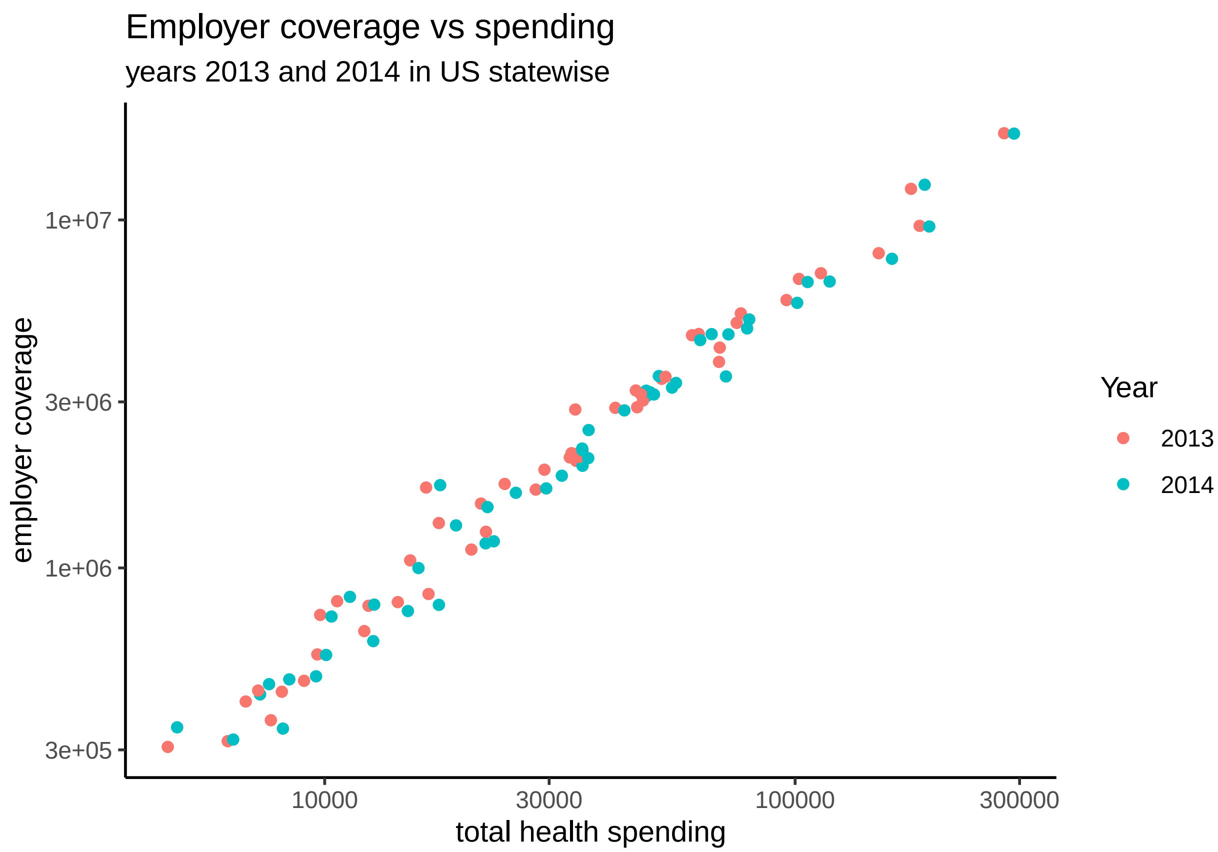
2.2 What are some of the steps that will need to be taken to get the data in that form?

We need to select employers columns of 2013 and 2014.

2.3 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. Then, create the scatterplot!

```
healthcare <- coverage_intersection %>%
  select(Location, "2013__Employer",
         "2014__Employer",
         "2013__Total",
         "2014__Total") %>%
  inner_join(spending_intersection, by = "Location") %>%
  pivot_longer(cols = -Location,
               names_sep = "__",
               names_to = c("year", "category"),
               values_to = "amount") %>%
  pivot_wider(names_from = "category",
              values_from = "amount") %>%
  clean_names()

ggplot(data = healthcare,
       mapping = aes(x = total_health_spending,
                     y = employer,
                     color = year)) +
  geom_point() +
  scale_x_log10()+
  scale_y_log10()+
  labs(title = "Employer coverage vs spending",
```

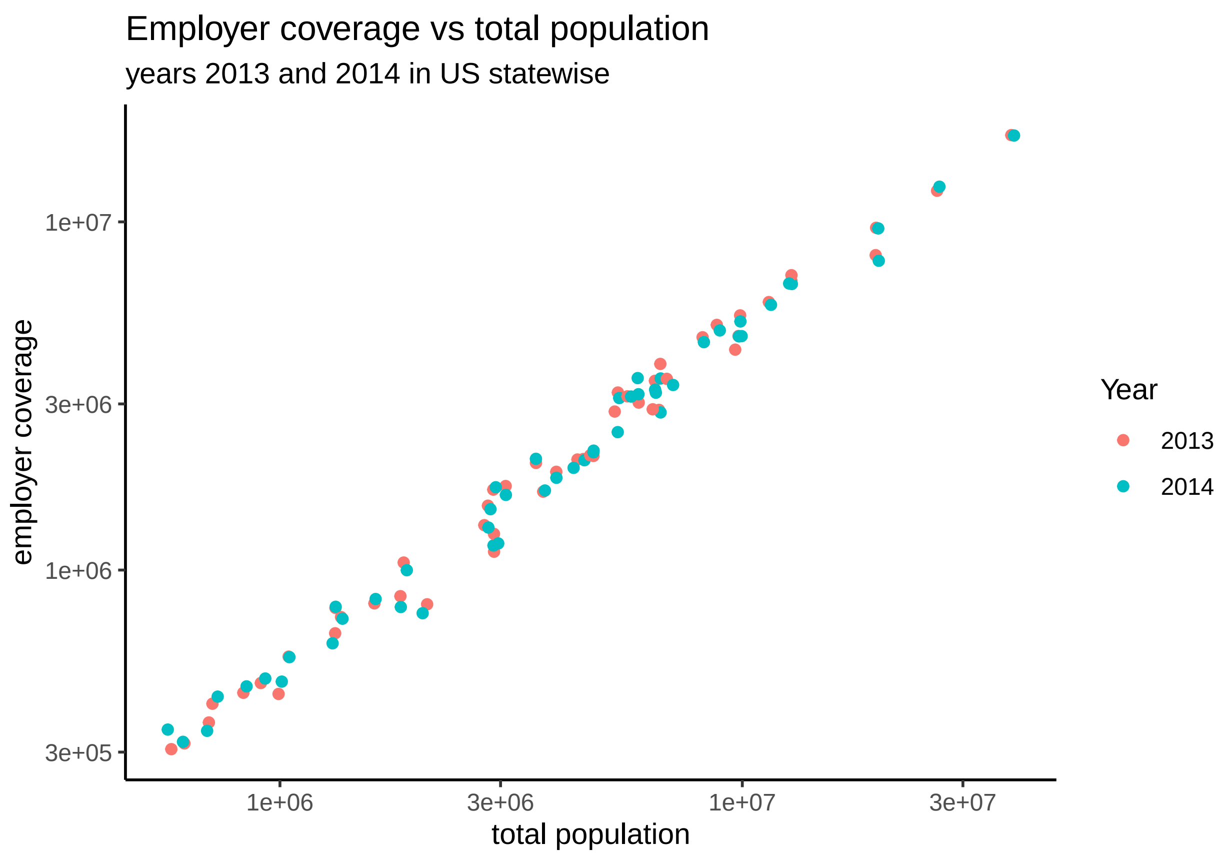
```
       subtitle = "years 2013 and 2014 in US statewise",
       y = "employer coverage",
       x = "total health spending",
       color = "Year")
```



Part 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!

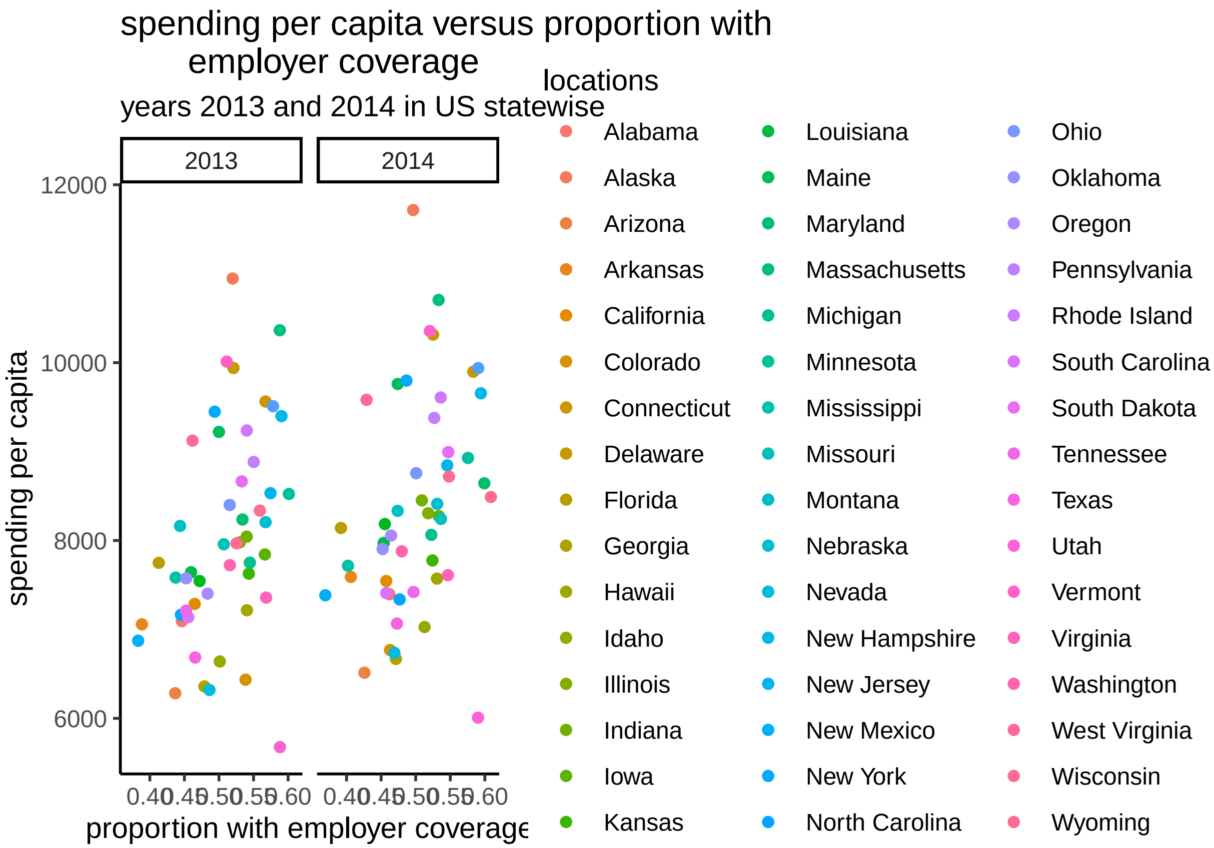
3.1 Using the dataset you created in Part 2, rename the `total` column to `total_population` to make the variable name more informative. Create a scatterplot of employer coverage versus population size (*note*: "plot blank vs blank" means "plot y variable vs x variable"). Then create a second scatterplot of healthcare spending vs. population size. What do you notice?

```r
healthcare <- coverage_intersection %>%
  select(Location, "2013__Employer",
         "2014__Employer",
         "2013__Total",
         "2014__Total") %>%
  inner_join(spending_intersection, by = "Location") %>%
  pivot_longer(cols = -Location,
               names_sep = "__",
               names_to = c("year", "category"),
               values_to = "amount") %>%
  pivot_wider(names_from = "category",
              values_from = "amount") %>%
  clean_names()

healthcare <- healthcare %>%
  rename(total_population = total)

ggplot(data = healthcare) +
  geom_point(mapping = aes(x = total_population,
                     y = employer,
                     color = year)) +
  scale_x_log10()+
  scale_y_log10()+
  labs(title = "Employer coverage vs total population",
       subtitle = "years 2013 and 2014 in US statewise",
       y = "employer coverage",
       x = "total population",
       color = "Year")
```



3.2 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!). 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`. Based on this figure, write a brief paragraph describing the relationship between health care spending and coverage in the US.

```
ggplot(data = healthcare) +
  geom_point(mapping = aes(x = (employer/total_population),
                     y = (total_health_spending/
                            total_population * 1000000),
                     color = location)) +
  facet_grid(cols = vars(year))+
  labs(title = "spending per capita versus proportion with
       employer coverage",
       subtitle = "years 2013 and 2014 in US statewise",
       y = "spending per capita",
       x = "proportion with employer coverage",
       color = "locations")
```




```
healthcare
```

```
# A tibble: 100 x 5
   location   year  employer total_population total_health_spending
   <chr>      <chr>    <dbl>            <dbl>                 <dbl>
 1 Alabama    2013   2126500          4763900                 33788
 2 Alabama    2014   2202800          4768000                 35263
 3 Alaska     2013    364900           702000                  7684
 4 Alaska     2014    345300           695700                  8151
 5 Arizona    2013   2883800          6603100                 41481
 6 Arizona    2014   2835200          6657200                 43356
 7 Arkansas   2013   1128800          2904800                 20500
 8 Arkansas   2014   1176500          2896000                 21980
 9 California 2013  17747300         38176400                278168
10 California 2014  17703700         38701300                291989
# ... with 90 more rows
```

Part 4 **How does spending vary by state and region?**

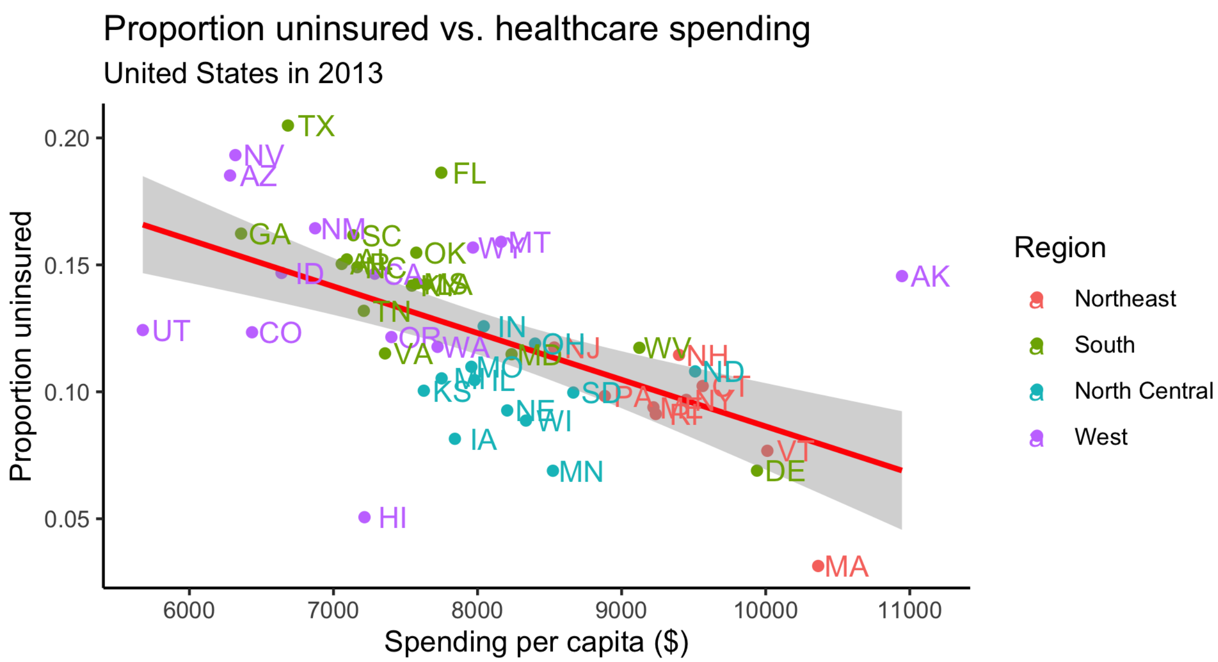
4.1 Which US state spent the most per capita on health care in 2013? 2014? The least in each year?

4.2 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. Write one paragraph summarizing a comparison of 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.)

The distributions look similar in 2013 and 2014. Spending per capita is highest and has the least variability in the Northeast (median around $9,500 and IQR around $300). The most variability in spending is in the West, which has a median of around $7,300, an IQR around $1,200, and a large outlier at over $10,000. The North Central and South regions have medians around $8,100 and $7,500, respectively and both have outliers on the high side.

Part 5 **Does the relationship between healthcare spending and the proportion of uninsured in the United States change from 2013 to 2014?**

5.1 Re-create the plot below for 2013. *Hint*: use `nudge_x` and/or `nudge_y` in the `geom_text()` layer.




```
# Add proportion uninsured to dataset
# healthcare <- healthcare %>%
#   mutate(proportion_uninsured = uninsured / total_population)
#
# healthcare %>%
#   filter(year == "2013") %>%
#   ggplot(aes(x = spending_per_capita, y = proportion_uninsured,
#              color = region)) +
#   geom_point() +
#   geom_smooth(method = "lm", col = "red") +
#   geom_text(aes(label = abbreviation), nudge_x = 200) +
#   labs(x = "Spending per capita ($)",
#        y = "Proportion uninsured",
#        color = "Region",
#        title = "Proportion uninsured vs. healthcare spending",
#        subtitle = "United States in 2013")
```

5.2 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?

The relationship between healthcare spending and the proportion uninsured looks pretty similar from 2013 to 2014.

```
# healthcare %>%
#   filter(year == "2014") %>%
#   ggplot(aes(x = spending_per_capita, y = proportion_uninsured,
#              color = region)) +
#   geom_point() +
#   geom_smooth(method = "lm", col = "red") +
#   geom_text(aes(label = abbreviation), nudge_x = 200) +
#   labs(x = "Spending per capita ($)",
#        y = "Proportion uninsured",
#        color = "Region",
#        title = "Proportion uninsured vs. healthcare spending",
#        subtitle = "United States in 2014")
```

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

```
# healthcare %>%
#   ggplot(aes(x = spending_per_capita, y = proportion_uninsured,
#              color = region)) +
#   geom_point() +
#   geom_smooth(method = "lm", col = "red") +
#   geom_text(aes(label = abbreviation), nudge_x = 200) +
#   facet_wrap(~ year, nrow = 2) +
#   labs(x = "Spending per capita ($)",
#        y = "Proportion uninsured",
#        color = "Region",
#        title = "Proportion uninsured vs. healthcare spending",
#        subtitle = "United States, 2013-2014")
```

5.4 Lastly, plot the points for both years on the same plot, this time colored by year instead of region. Make sure to specify the `group` aesthetic for year as well to get two lines. Which of these three visualizations do you find most helpful for comparing the relationship between 2013 and 2014? Why?

I think the last one where both years are on the same figure (colored by year instead of region) is most helpful for comparing the relationship between 2013 and 2014. We lose the visual cue of region, but it's easiest for me to see the change in the relationship when the points and lines are on the same plot.

```
# healthcare %>%
#   ggplot(aes(x = spending_per_capita, y = proportion_uninsured,
```

```
#               color = year, group = year)) +
#   geom_point() +
#   geom_smooth(method = "lm") +
#   geom_text(aes(label = abbreviation), nudge_x = 200) +
#   labs(x = "Spending per capita ($)",
#        y = "Proportion uninsured",
#        color = "Year",
#        title = "Proportion uninsured vs. healthcare spending",
#        subtitle = "United States, 2013-2014")
```

Part 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
- change the font of the text on the y-axis

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