Lab 5: Wrangle and tidy

Solutions

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)
```

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))
```

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.

```
spending1314 <- spending %>%
  select(Location,
         starts_with(c("2013", "2014")))

coverage1314 <- coverage %>%
  select(Location,
         starts_with(c("2013", "2014")))
```

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.

```
state_coverage1314 <- coverage1314 %>%
  filter(!(Location %in% c("United States", "District of Columbia")))

state_spending1314 <- spending1314 %>%
  filter(!(Location %in% c("United States", "District of Columbia")))
```

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?

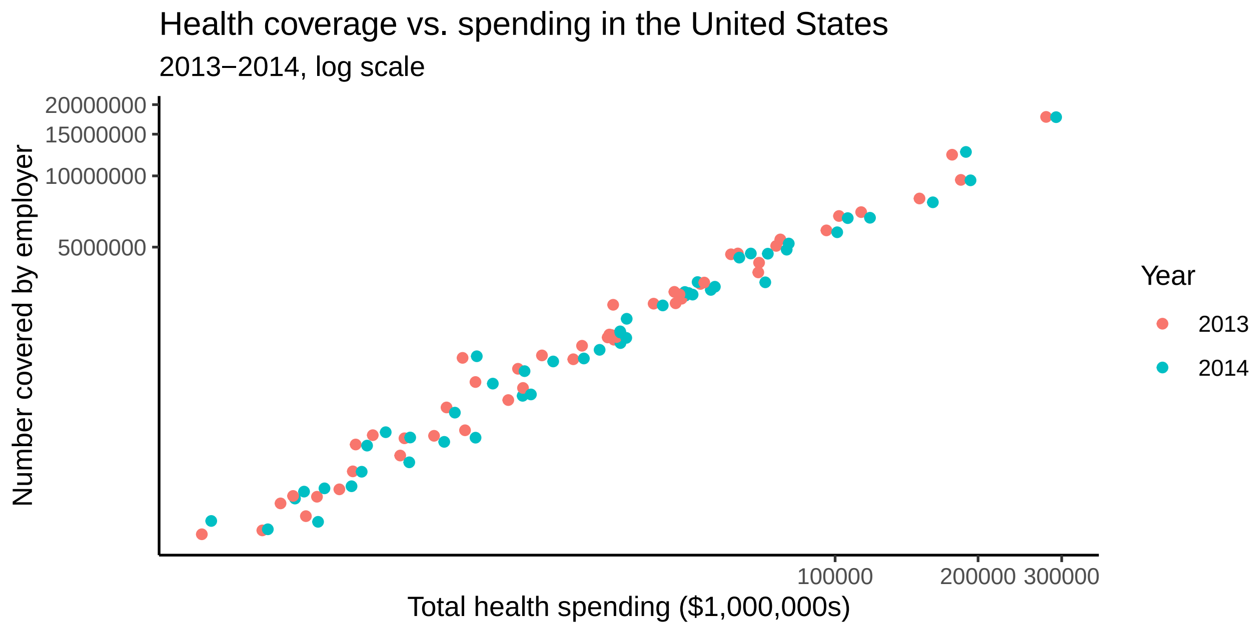
We'll need `log(spending)`, `log(employee coverage)`, and `year` as variables in the dataset in order to create this scatterplot in ggplot. There will be a separate observation/row for each state and year.

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

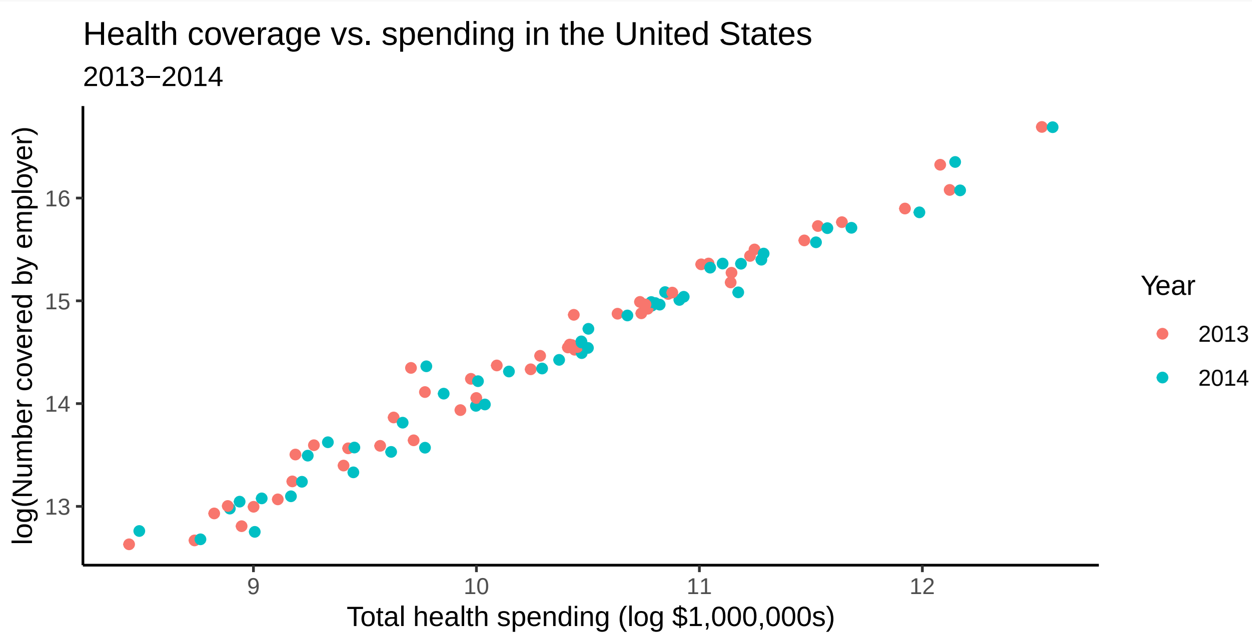
We'll need to merge the spending and coverage datasets, then put the merged dataset into long form such that year is a variable. Then we'll need to break the data back out wide so that we have a variable for employer coverage and another for healthcare spending, in addition to the combined *year* variable.

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 <- state_coverage1314 %>%
  inner_join(state_spending1314, 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() %>%
  mutate(log_total_spending = log(total_health_spending),
         log_employer = log(employer))

# log scale
ggplot(data = healthcare, aes(x = total_health_spending, y = employer,
                              color = year)) +
  coord_trans(x = "log", y = "log") +
  geom_point() +
  labs(x = "Total health spending ($1,000,000s)",
       y = "Number covered by employer",
       color = "Year",
       title = "Health coverage vs. spending in the United States",
       subtitle = "2013-2014, log scale")
```




```
# log transformation
ggplot(data = healthcare, aes(x = log_total_spending, y = log_employer,
                              color = year)) +
  geom_point() +
  labs(x = "Total health spending (log $1,000,000s)",
       y = "log(Number covered by employer)",
       color = "Year",
       title = "Health coverage vs. spending in the United States",
       subtitle = "2013-2014")
```

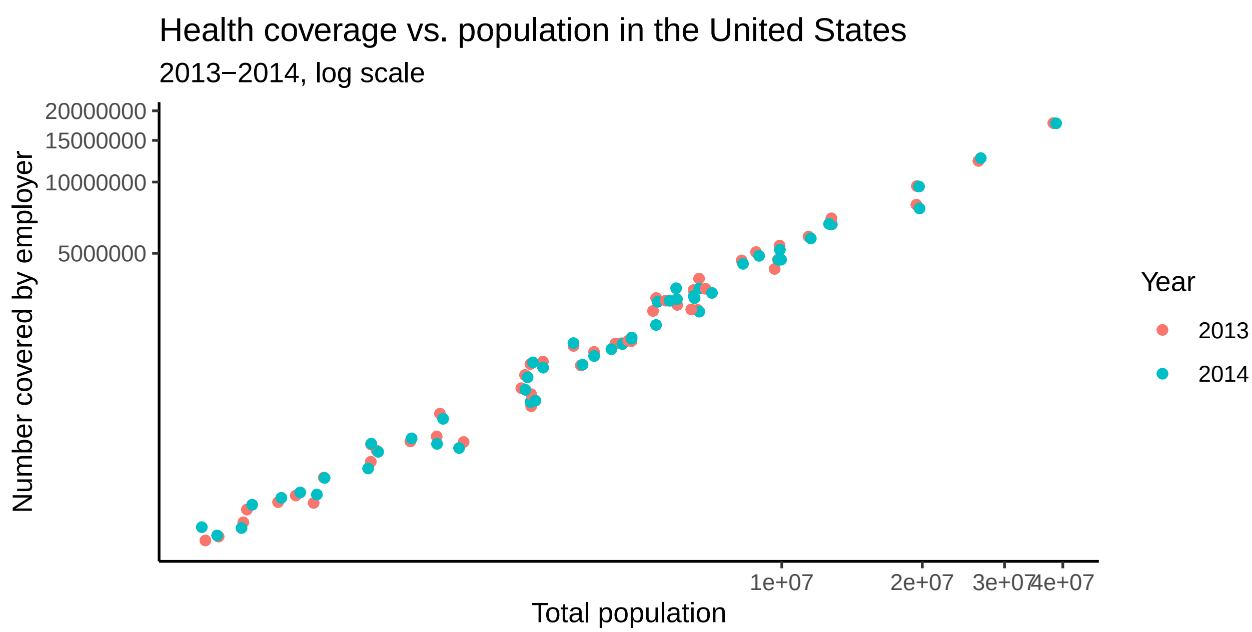


Part 3 **Adjusting for population size** We see there is a strong relationship between healthcare spending and coverage within each year. However, we might suspect that health care coverage and spending are each strongly related to population size. In the `coverage` dataset, the "total" coverage category 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?

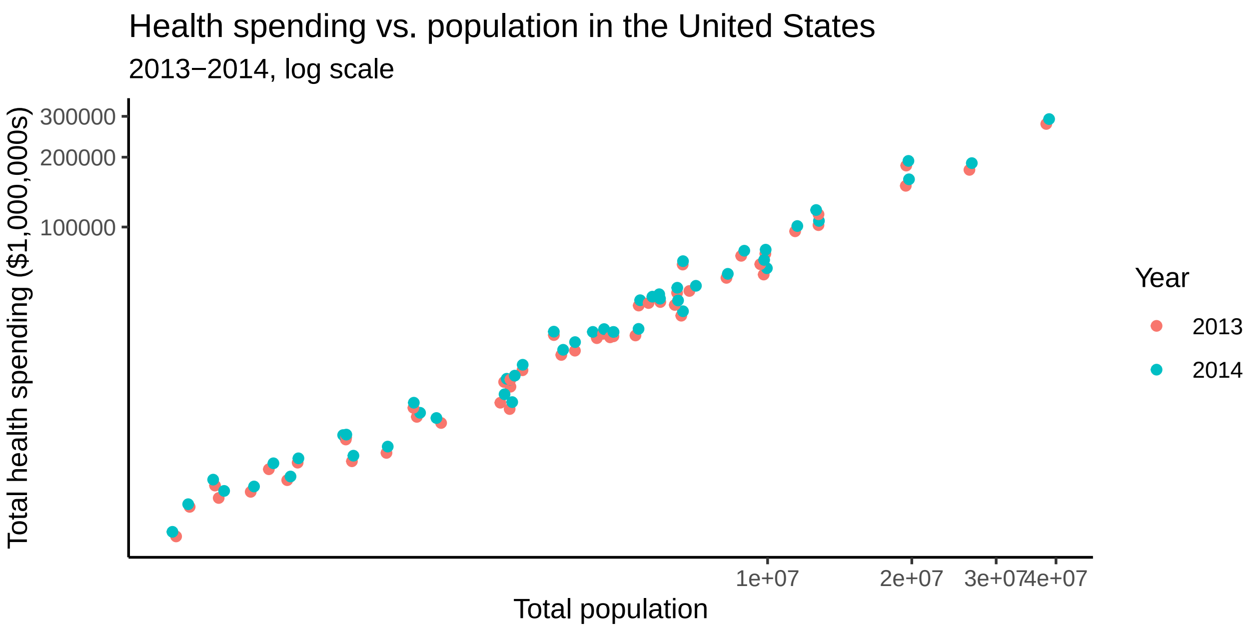
We see strong, positive, linear relationships between population size and both healthcare spending and healthcare coverage by employers (all on log scales).

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

ggplot(data = healthcare, aes(x = total_population, y = employer,
                              color = year)) +
  coord_trans(x = "log", y = "log") +
  geom_point() +
  labs(x = "Total population",
       y = "Number covered by employer",
       color = "Year",
       title = "Health coverage vs. population in the United States",
       subtitle = "2013-2014, log scale")
```




```
ggplot(data = healthcare, aes(x = total_population, y = total_health_spending,
                              color = year)) +
  coord_trans(x = "log", y = "log") +
  geom_point() +
  labs(x = "Total population",
```

```
       y = "Total health spending ($1,000,000s)",
       color = "Year",
       title = "Health spending vs. population in the United States",
       subtitle = "2013-2014, log scale")
```



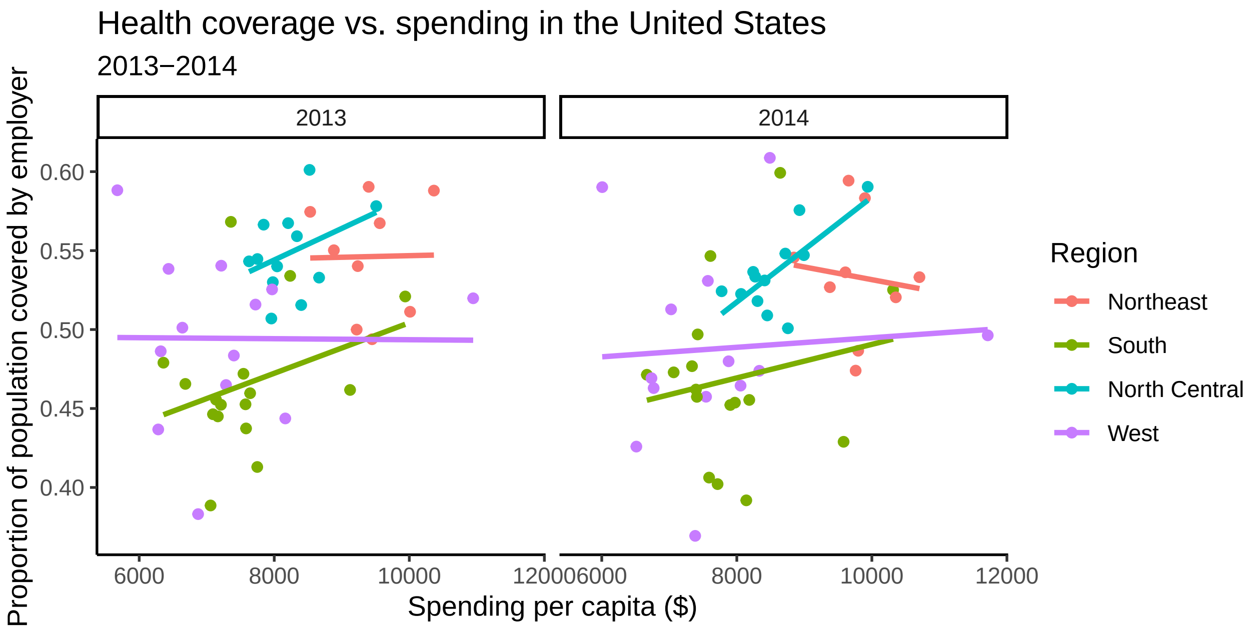
3.2 To account for total population, create a scatterplot of spending per capita versus proportion with employer coverage. This time, *color by region* and *facet by year* (think about what additional steps you need to take to make this happen!). The total spending column is reported in millions (`1e6`). 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.

Taking all states together, there appears to be a weak-moderate, positive, linear association between spending per capita and proportion with employer coverage both in 2013 and 2014. Within the North Central and South regions of the US, as spending per capita increases, the proportion with employer coverage increases. Within the Northeast and the West, there doesn't appear to be much association between spending per capita and proportion with employer coverage.

```
healthcare <- healthcare %>%
  mutate(spending_per_capita = total_health_spending * 1e6 / total_population,
         proportion_covered = employer / total_population) %>%
  left_join(state_info, by = "location")
```

```
ggplot(data = healthcare, mapping = aes(x = spending_per_capita, y = proportion_covered,
                                        color = region)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(~ year) +
   labs(x = "Spending per capita ($)",
        y = "Proportion of population covered by employer",
        color = "Region",
```

```
       title = "Health coverage vs. spending in the United States",
       subtitle = "2013-2014")
```



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?

Alaska, the most per capita on health care in 2013 and 2014 (over $10,000 per capita). Utah, Arizona, Nevada, and Georgia spent the least per capita on health care in 2013 and 2014 (less than $6,800 per capita).

```
# Find only the max and min
healthcare %>%
  select(location, year, spending_per_capita) %>%
  group_by(year) %>%
  filter(spending_per_capita == max(spending_per_capita) |
           spending_per_capita == min(spending_per_capita))
```

```
# A tibble: 4 x 3
# Groups:   year [2]
  location year  spending_per_capita
  <chr>    <chr>               <dbl>
1 Alaska   2013               10946.
2 Alaska   2014               11716.
3 Utah     2013                5677.
4 Utah     2014                6007.
```

```
# Using pivot and arrange to see more states
spending_wide <- healthcare %>%
  select(location, year, spending_per_capita) %>%
  # Add prefix since variable names shouldn't start with numbers
  pivot_wider(names_from = "year", names_prefix = "year",
              values_from = "spending_per_capita")

spending_wide %>%
  arrange(year2013) %>%
  head()
```

```
# A tibble: 6 x 3
  location year2013 year2014
  <chr>       <dbl>    <dbl>
1 Utah        5677.    6007.
2 Arizona     6282.    6513.
3 Nevada      6319.    6737.
4 Georgia     6359.    6668.
5 Colorado    6435.    6769.
6 Idaho       6639.    7027.
```

```
spending_wide %>%
  arrange(year2014) %>%
  head()
```

```
# A tibble: 6 x 3
  location year2013 year2014
  <chr>       <dbl>    <dbl>
1 Utah        5677.    6007.
2 Arizona     6282.    6513.
3 Georgia     6359.    6668.
4 Nevada      6319.    6737.
5 Colorado    6435.    6769.
```

```
## 6 Idaho        6639.    7027.
```

```r
spending_wide %>%
  arrange(desc(year2013)) %>%
  head()
```

```
## # A tibble: 6 x 3
##   location      year2013 year2014
##   <chr>            <dbl>    <dbl>
## 1 Alaska          10946.   11716.
## 2 Massachusetts   10364.   10705.
## 3 Vermont         10011.   10355.
## 4 Delaware         9940.   10314.
## 5 Connecticut      9562.    9898.
## 6 North Dakota     9510.    9939.
```

```r
spending_wide %>%
  arrange(desc(year2014)) %>%
  head()
```

```
## # A tibble: 6 x 3
##   location      year2013 year2014
##   <chr>            <dbl>    <dbl>
## 1 Alaska          10946.   11716.
## 2 Massachusetts   10364.   10705.
## 3 Vermont         10011.   10355.
## 4 Delaware         9940.   10314.
## 5 North Dakota     9510.    9939.
## 6 Connecticut      9562.    9898.
```

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.)

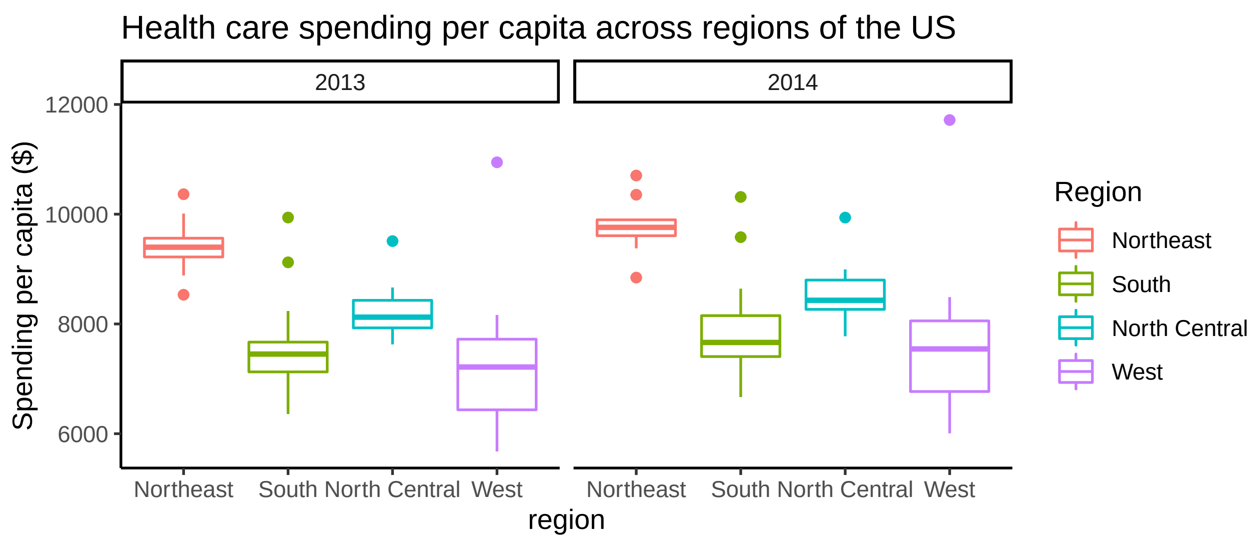
```r
# Summary statistics
healthcare %>%
  group_by(year, region) %>%
  summarise(n = n(),
            mean = mean(spending_per_capita),
            sd = sd(spending_per_capita),
            median = median(spending_per_capita))
```

```
## # A tibble: 8 x 6
## # Groups:   year [2]
##   year  region            n  mean    sd median
##   <chr> <fct>         <int> <dbl> <dbl>  <dbl>
## 1 2013  Northeast         9 9406.  549.  9399.
## 2 2013  South            16 7591.  886.  7451.
## 3 2013  North Central    12 8236.  511.  8124.
## 4 2013  West             13 7302. 1313.  7215.
## 5 2014  Northeast         9 9778.  534.  9760.
## 6 2014  South            16 7934.  924.  7663.
```

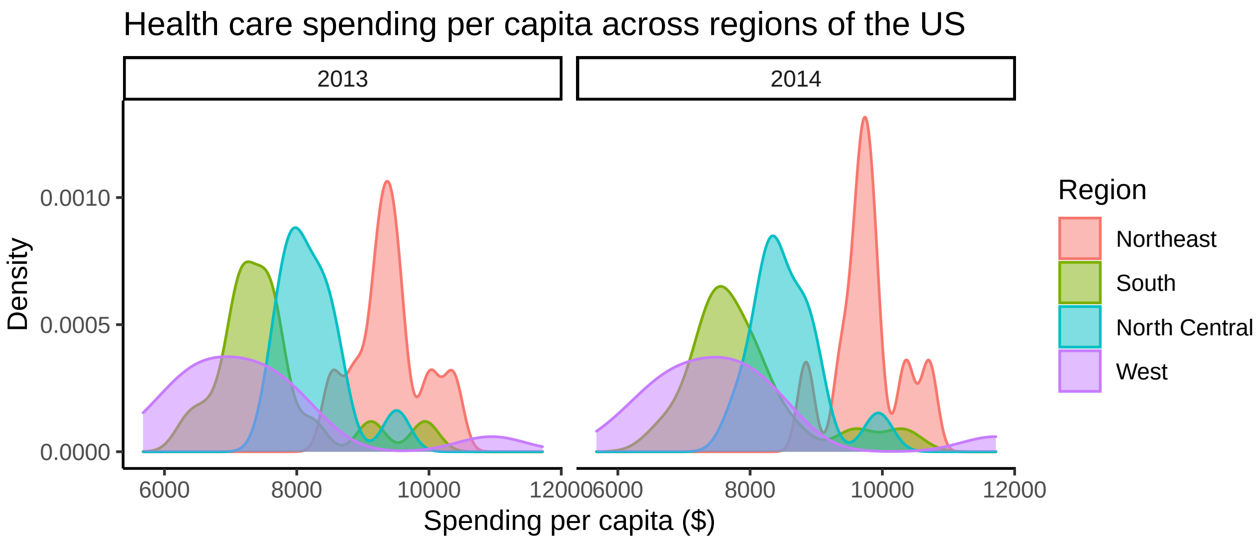
```
7 2014  North Central     12 8572.  558.  8431.
8 2014  West              13 7694. 1414.  7545.
```

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.

```
# Boxplots
ggplot(data = healthcare, mapping = aes(x = region, y = spending_per_capita,
                                        color = region)) +
    geom_boxplot() +
    facet_wrap(~ year) +
    labs(y = "Spending per capita ($)",
         color = "Region",
         title = "Health care spending per capita across regions of the US")
```

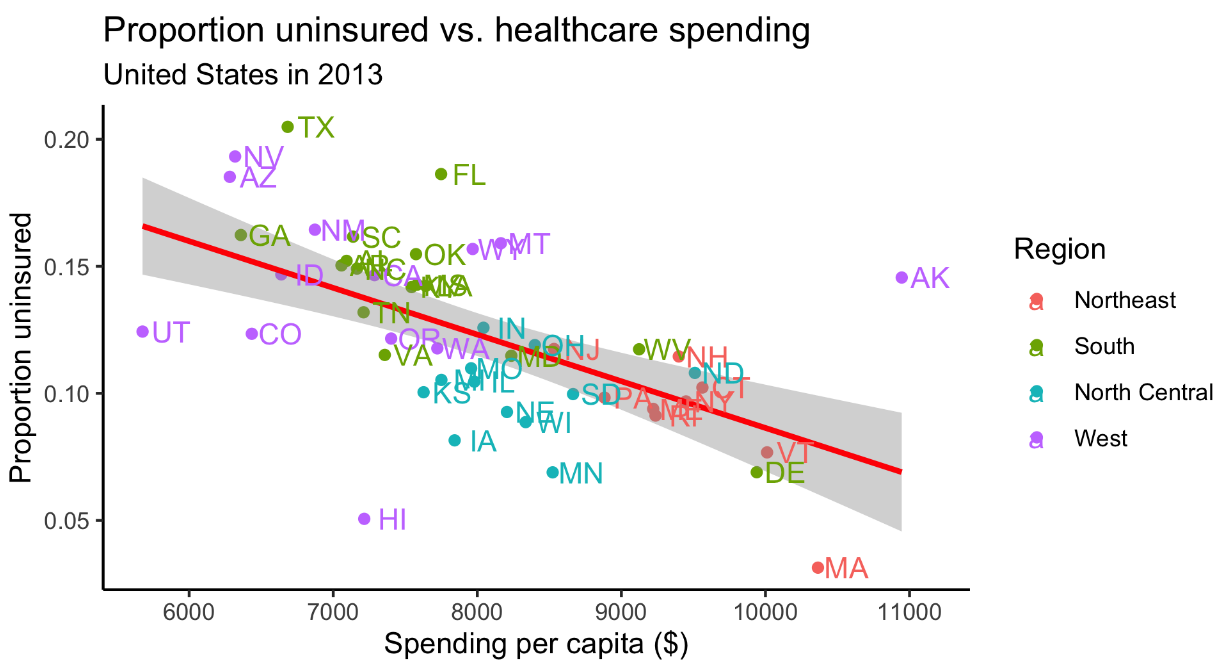



```
# Density plots
ggplot(data = healthcare, mapping = aes(x = spending_per_capita,
                                        color = region, fill = region)) +
    geom_density(alpha = 0.5) +
    facet_wrap(~ year) +
    labs(x = "Spending per capita ($)",
         y = "Density",
         color = "Region",
         fill = "Region",
         title = "Health care spending per capita across regions of the US")
```

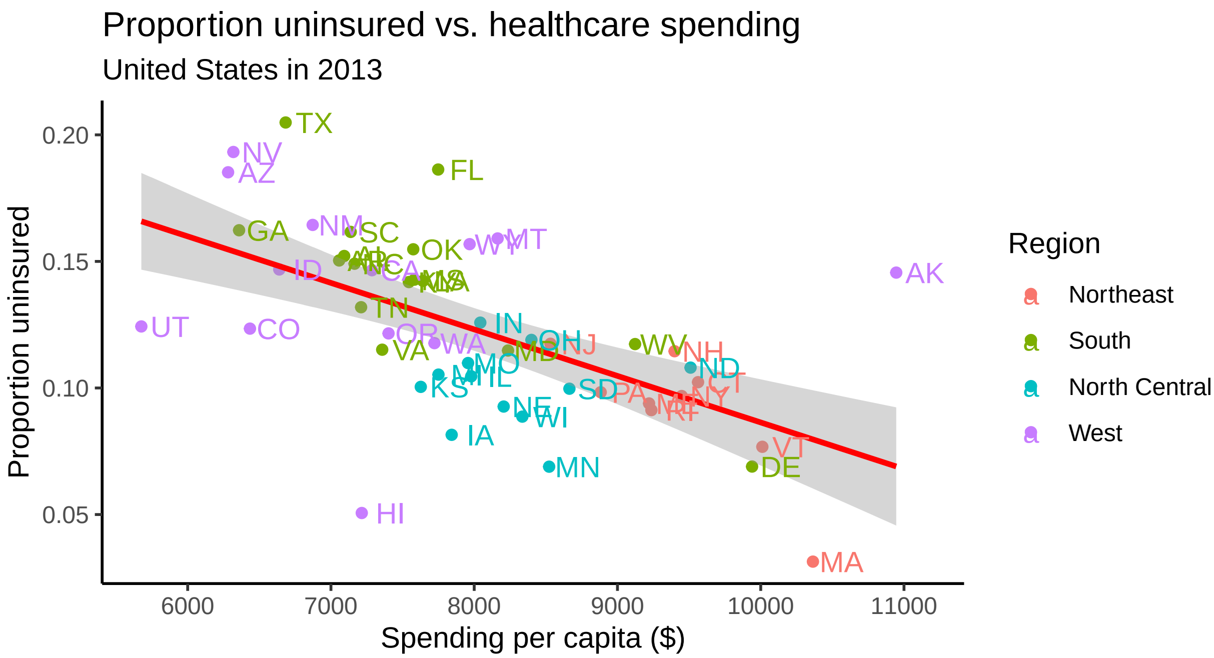

Health care spending per capita across regions of the US
2013
2014
Density
0.0010
0.0005
0.0000
6000
8000
10000
12000
Spending per capita ($)
Region
Northeast
South
North Central
West

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.



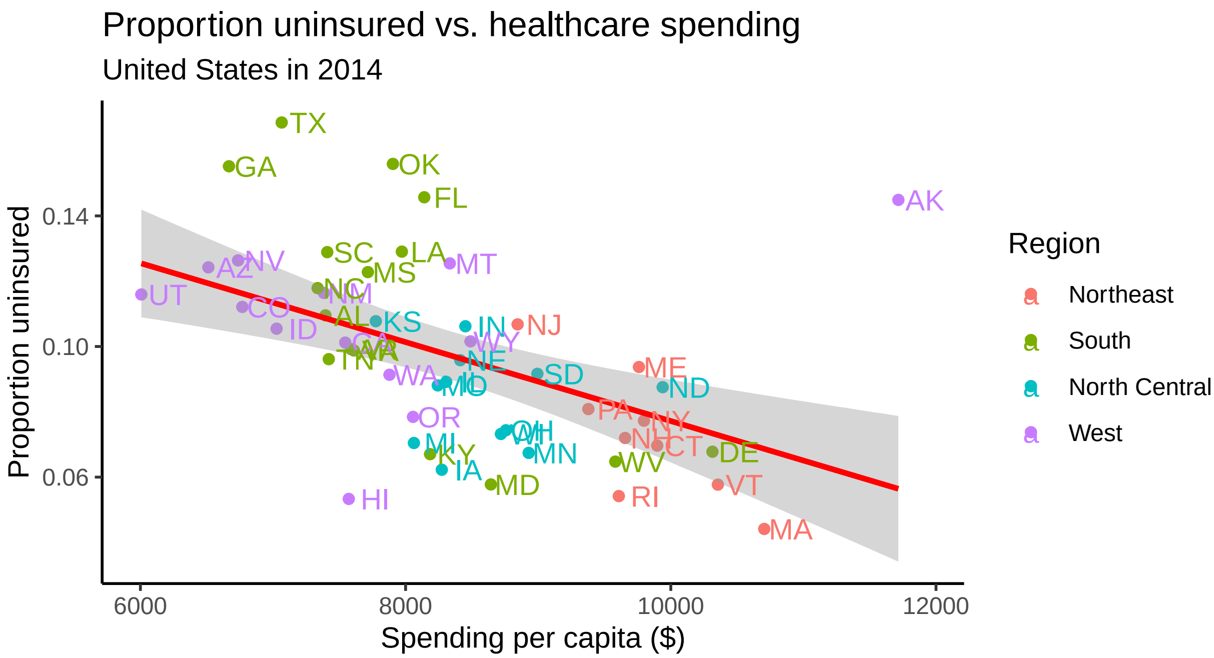

```
# Add proportion uninsured to dataset
healthcare <- healthcare %>%
  mutate(proportion_uninsured = uninsured / total_population)

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?

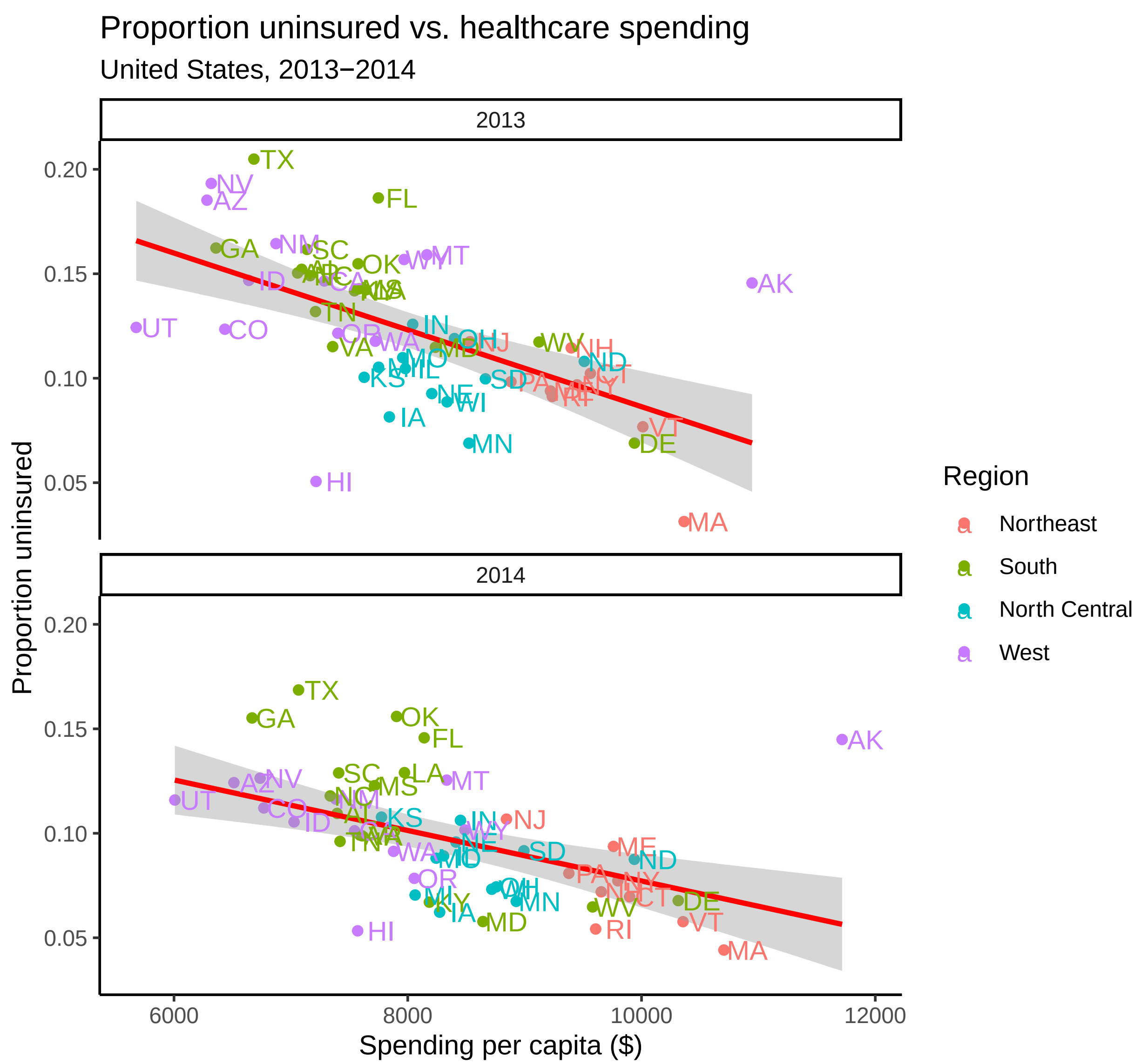
The relationship between healthcare spending and the proportion uninsured looks pretty similar from 2013 to 2014.

```
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")
```

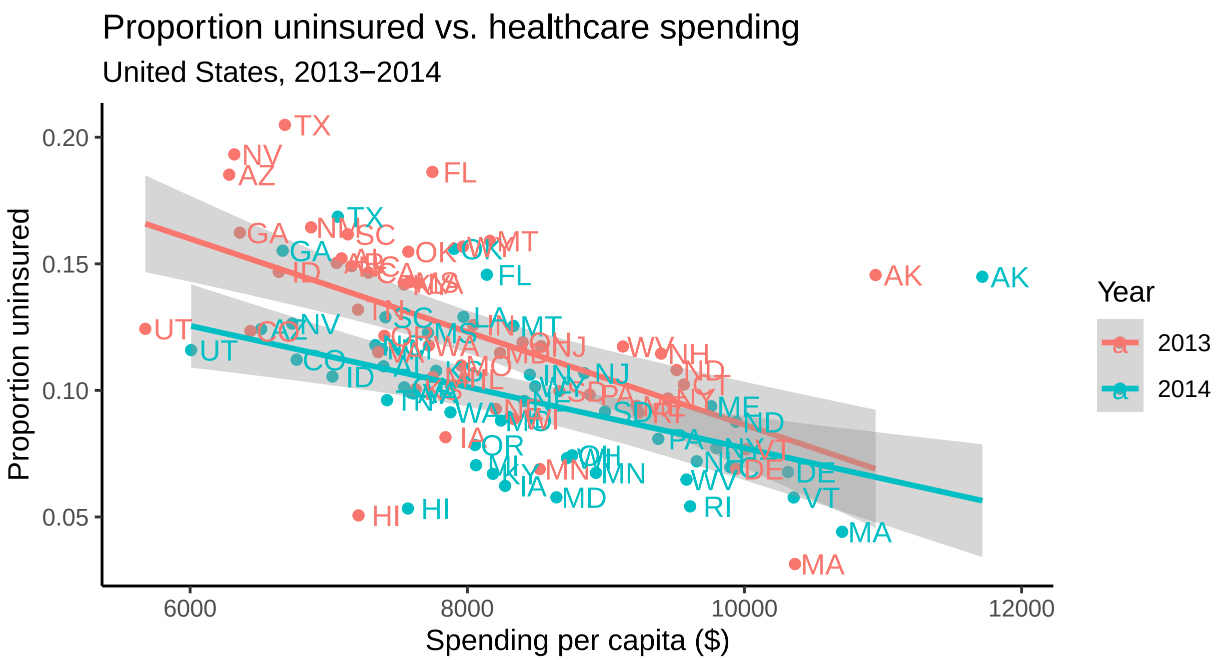


5.4 Lastly, plot the points for both years on the same plot, this time colored by year instead of region. Make sure to specify the `group` aesthetic for year as well to get two lines. Which of these three visualizations do you find most helpful for comparing the relationship between 2013 and 2014? Why?

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 ($)",
```

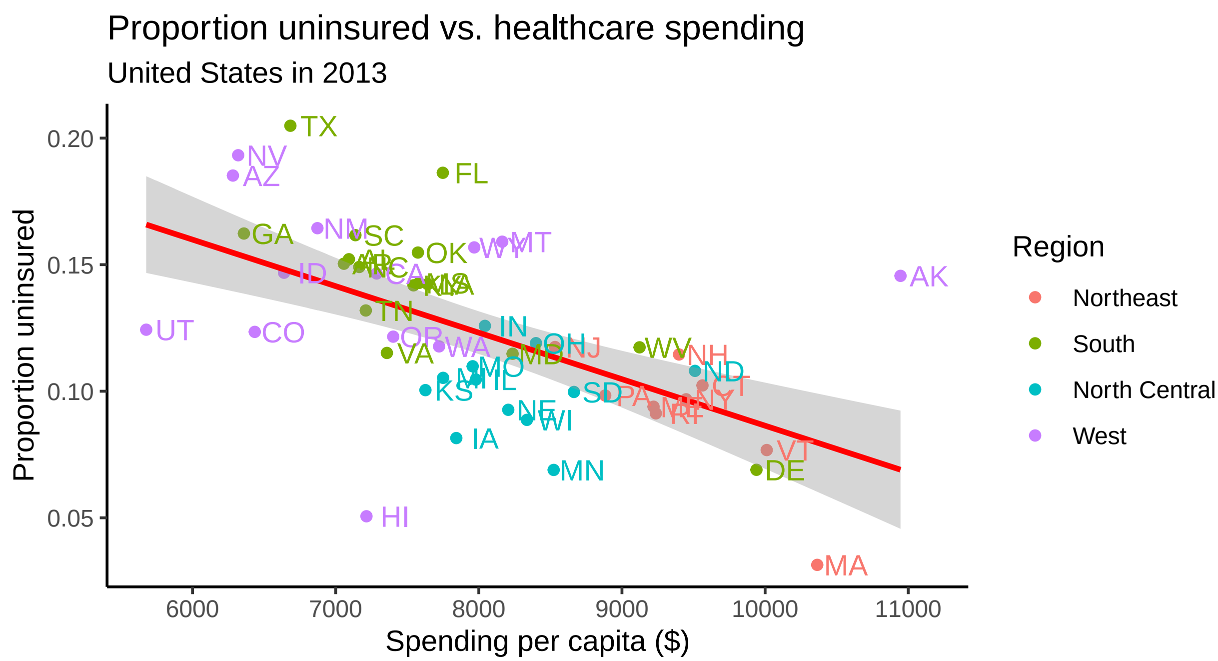
```
       y = "Proportion uninsured",
       color = "Year",
       title = "Proportion uninsured vs. healthcare spending",
       subtitle = "United States, 2013-2014")
```



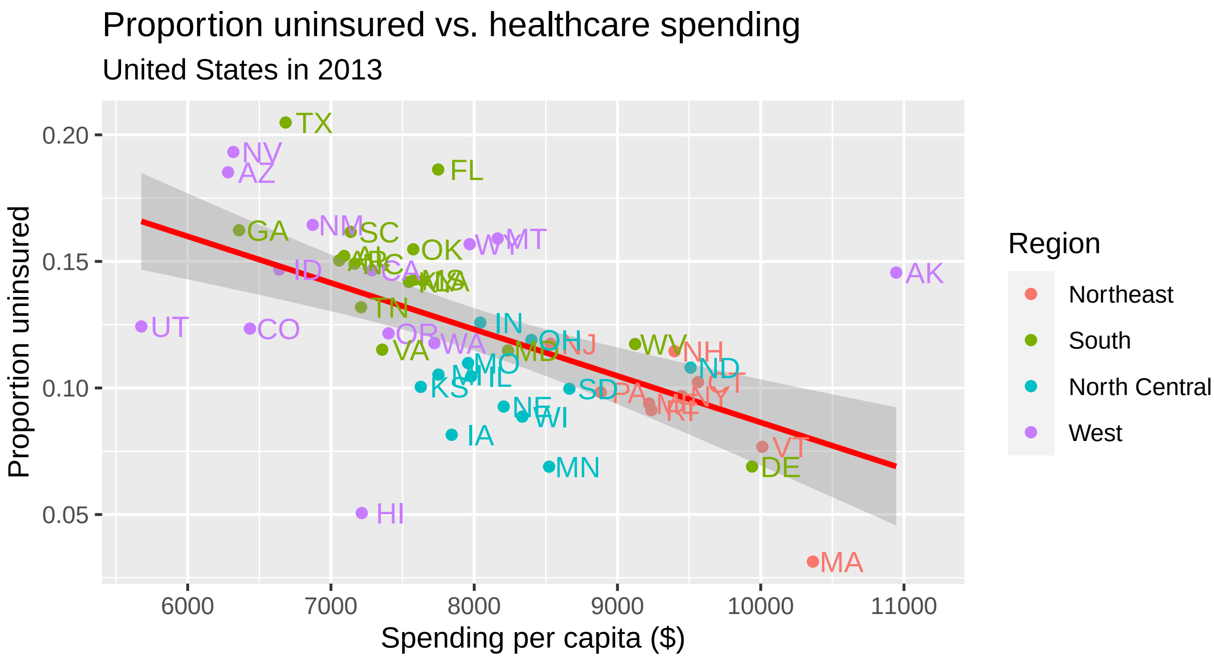
Part 6 **Bonus** Done early? Try to figure out how to make these additional updates to the figure from Problem 5.1 to hone your plotting skills:

- 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

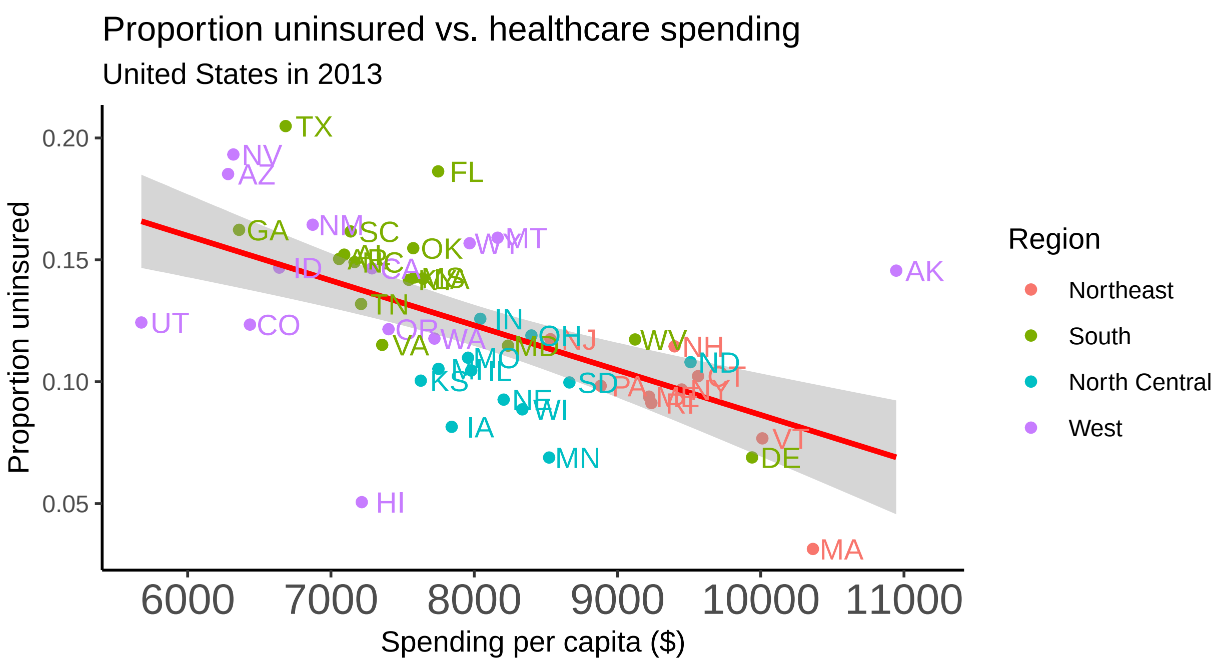
```
base_fig <- healthcare %>%
  filter(year == "2013") %>%
  ggplot(aes(x = spending_per_capita, y = proportion_uninsured,
             color = region)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red") +
  labs(x = "Spending per capita ($)",
       y = "Proportion uninsured",
       color = "Region",
       title = "Proportion uninsured vs. healthcare spending",
       subtitle = "United States in 2013")

# remove the "a" on the points in the legend
# (use show.legend=FALSE option within geom_text)
base_fig +
  geom_text(aes(label = abbreviation), nudge_x = 200, show.legend = FALSE)
```



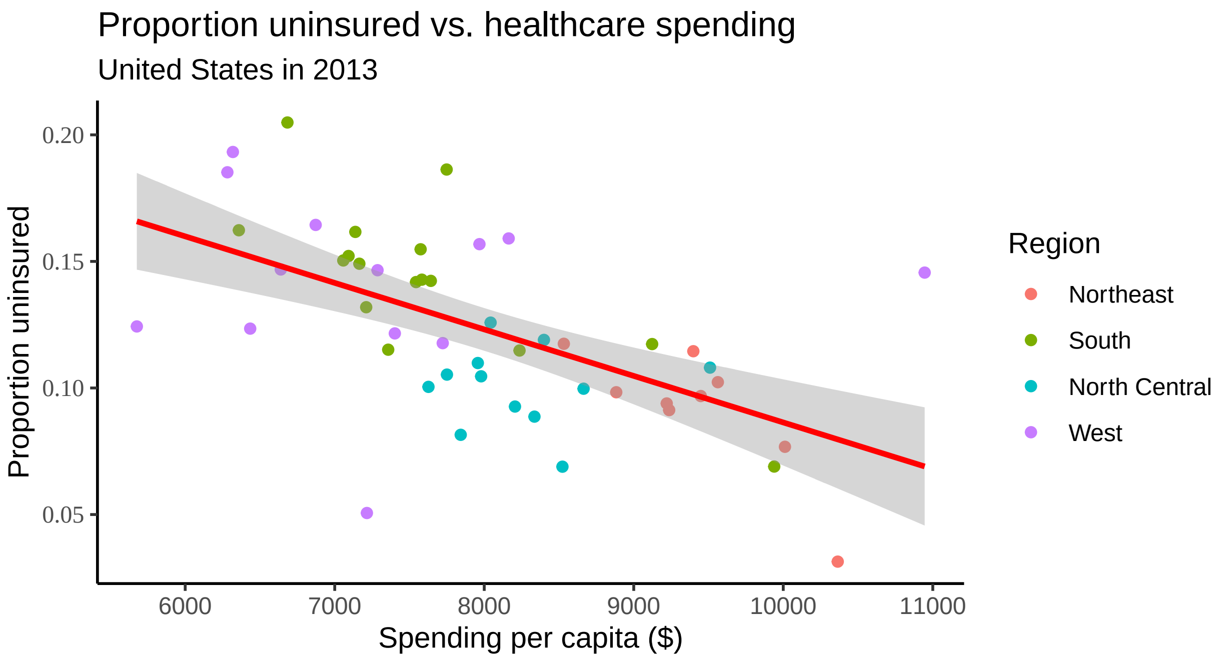

```
# change the background to be grey
base_fig +
  geom_text(aes(label = abbreviation), nudge_x = 200, show.legend = FALSE) +
  theme_grey()
```




```r
# make the numbers on the x-axis larger
base_fig +
  geom_text(aes(label = abbreviation), nudge_x = 200, show.legend = FALSE) +
  theme(axis.text.x = element_text(size = 16))
```




```r
# change the font of the text on the y-axis
# note this throws error on Windows; harder to change fonts on Windows...
# http://www.cookbook-r.com/Graphs/Fonts/
base_fig + theme(axis.text.y = element_text(family = "Times"))
```


Proportion uninsured vs. healthcare spending
United States in 2013
Proportion uninsured
0.20
0.15
0.10
0.05
6000
7000
8000
9000
10000
11000
Spending per capita ($)
Region
Northeast
South
North Central
West

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