Lab 3

Christy Lei

Math 241, Week 4

```
# Put all necessary libraries here
library(tidyverse)
library(ggplot2)
library(ggthemes)
```

Due: Thursday, February 27th at 8:30am

Goals of this lab

1. Practice using GitHub.
2. Practice wrangling data.

Data Notes:

- For Problem 2, we will continue to dig into the SE Portland crash data but will use two datasets:
  - `CRASH`: crash level data
  - `PARTIC`: participant level data

```
# Crash level dataset
crash <- read_csv("/home/courses/math241s20/Data/pdx_crash_2018_CRASH.csv")

# Participant level dataset
partic <- read_csv("/home/courses/math241s20/Data/pdx_crash_2018_PARTIC.csv")
```

- For Problem 3, we will look at chronic illness data from the CDC along with the regional mapping for each state.

```
# CDC data
CDC <- read_csv("/home/courses/math241s20/Data/CDC2.csv")

# Regional data
USregions <- read_csv("/home/courses/math241s20/Data/USregions.csv")
```

- For Problem 4, we will use polling data from FiveThirtyEight.com.

```
# Note I only want us to focus on a subset of the variables
polls <- read_csv("/home/courses/math241s20/Data/generic_topline.csv") %>%
  select(subgroup, modeldate, dem_estimate, rep_estimate)
```

Problems

Problem 1: Git Control

In this problem, we will practice interacting with GitHub on the site directly and from the RStudio Server. Do this practice on **your repo**, not your group's Project 1 repo, so that the graders can check your progress with Git.

a. Let's practice creating and closing **Issues**. In a nutshell, **Issues** let us keep track of our work. Within your repo on GitHub.com, create an Issue entitled "Complete Lab 3". Once Lab 3 is done, close the **Issue**. (If you want to learn more about the functionalities of Issues, check out this page.)

b. Edit the ReadMe of your repo to include your name and a quick summary of the purpose of the repo. You can edit from within GitHub directly or on the server. If you edit on the server, make sure to push your changes to GitHub.

c. Upload both your Lab 3 .Rmd and .pdf to your repo on GitHub.

Problem 2: `dplyr` madness

Each part of this problem will require you to wrangle the data and then do one or both of the following:

- Display the wrangled data frame. To ensure it displays the whole data frame, you can pipe `as.data.frame()` at the end of the wrangling.
- Answer a question(s).

Some parts will require you to do a data join but won't tell you that.

a. Produce a table that provides the frequency of the different collision types, ordered from most to least common. What type is most common? What type is least common?

```
crash %>%
  count(COLLIS_TYP_CD) %>%
  arrange(desc(n)) %>%
  as.data.frame()
```

```
##    COLLIS_TYP_CD   n
## 1              3 671
## 2              6 365
## 3              1 241
## 4              5  89
## 5              0  86
## 6              9  51
## 7              4  17
## 8              2  16
## 9              -  12
## 10             7  10
## 11             8   6
## 12             &   3
```

Rear-End (type 3) is the most common collision type and Miscellaneous (type &) is the least common one.

b. For the three most common collision types, create a table that contains:
   - The frequencies of each collision type and weather condition combination.
   - The proportion of each collision type by weather condition.

Arrange the table by weather and within type, most to least common collision type.

```
crash %>%
  count(COLLIS_TYP_CD, WTHR_COND_CD) %>%
  group_by(WTHR_COND_CD) %>%
  mutate(prop = prop.table(n)) %>%
  arrange(desc(n)) %>%
  as.data.frame()
```

```
##    COLLIS_TYP_CD WTHR_COND_CD   n        prop
## 1              3            1 549 0.447797716
## 2              6            1 290 0.236541599
## 3              1            1 188 0.153344209
## 4              5            1  73 0.059543230
## 5              3            3  71 0.381720430
## 6              0            1  50 0.040783034
## 7              6            3  44 0.236559140
## 8              3            2  29 0.308510638
## 9              1            3  28 0.150537634
## 10             9            1  28 0.022838499
## 11             3            0  20 0.434782609
## 12             6            2  20 0.212765957
## 13             0            3  17 0.091397849
## 14             0            2  16 0.170212766
## 15             4            1  15 0.012234910
## 16             1            2  13 0.138297872
## 17             2            1  12 0.009787928
## 18             9            3  11 0.059139785
## 19             -            1   8 0.006525285
## 20             6            0   8 0.173913043
## 21             7            1   8 0.006525285
## 22             9            2   8 0.085106383
## 23             1            0   6 0.130434783
## 24             5            3   6 0.032258065
## 25             5            2   5 0.053191489
## 26             5            0   4 0.086956522
## 27             -            3   3 0.016129032
## 28             &            1   3 0.002446982
## 29             1            6   3 0.333333333
## 30             9            0   3 0.065217391
## 31             0            0   2 0.043478261
## 32             1            5   2 0.666666667
## 33             4            3   2 0.010752688
## 34             6            6   2 0.222222222
## 35             8            1   2 0.001631321
## 36             8            2   2 0.021276596
## 37             8            3   2 0.010752688
## 38             -            0   1 0.021739130
## 39             0            8   1 0.500000000
## 40             1            4   1 1.000000000
## 41             2            0   1 0.021739130
## 42             2            2   1 0.010638298
## 43             2            3   1 0.005376344
## 44             2            6   1 0.111111111
## 45             3            6   1 0.111111111
## 46             3            8   1 0.500000000
```

```
## 47             5            6   1 0.111111111
## 48             6            5   1 0.333333333
## 49             7            0   1 0.021739130
## 50             7            3   1 0.005376344
## 51             9            6   1 0.111111111
```

c. Create a column for whether or not a crash happened on a weekday or on the weekend and then create a data frame that explores if the distribution of collision types varies by whether or not the crash happened during the week or the weekend.

```
weekday = c(1,2,3,4,5)
weekend = c(6,7)
crash$weekday_true =
  ifelse(crash$CRASH_WK_DAY_CD %in% weekday, "Yes",
  ifelse(crash$CRASH_WK_DAY_CD %in% weekend, "No", NA))

crash_weekday <- crash %>%
  count(COLLIS_TYP_CD, weekday_true) %>%
  group_by(weekday_true) %>%
  mutate(prop = prop.table(n)) %>%
  arrange(desc(n))

crash_weekday
```

```
## # A tibble: 23 x 4
## # Groups:   weekday_true [2]
##    COLLIS_TYP_CD weekday_true     n   prop
##    <chr>         <chr>        <int>  <dbl>
##  1 3             Yes            461 0.419
##  2 6             Yes            256 0.233
##  3 3             No             210 0.450
##  4 1             Yes            172 0.156
##  5 6             No             109 0.233
##  6 1             No              69 0.148
##  7 0             Yes             61 0.0555
##  8 5             Yes             60 0.0545
##  9 9             Yes             40 0.0364
## 10 5             No              29 0.0621
## # ... with 13 more rows
```

It seems that the most common collision type (Type 3) happened mostly on weekdays.

d. First determine what proportion of crashes involve pedestrians. Then, for each driver license status, determine what proportion of crashes involve pedestrians. What driver license status has the highest rate of crashes that involve pedestrians?

```
#proportion of crashes that involve pedestrians
crash %>%
  count(pedestrians_involved = (CRASH_TYP_SHORT_DESC == "PED")) %>%
  mutate(prop = prop.table(n))
```

```
## # A tibble: 2 x 3
##   pedestrians_involved     n   prop
##   <lgl>                <int>  <dbl>
## 1 FALSE                 1481 0.945
## 2 TRUE                    86 0.0549
```

```r
#proportion of crashes that involve pedestrians for each driver license status
crash_and_partic <- left_join (crash, partic, by = c("CRASH_ID" = "CRASH_ID"))
crash_and_partic %>%
  group_by(DRVR_LIC_STAT_CD) %>%
  count(pedestrians_involved = (CRASH_TYP_SHORT_DESC == "PED")) %>%
  mutate(prop = prop.table(n)) %>%
  filter(pedestrians_involved == TRUE) %>%
  arrange(desc(n))
```

```
## # A tibble: 5 x 4
## # Groups:   DRVR_LIC_STAT_CD [5]
##   DRVR_LIC_STAT_CD pedestrians_involved     n    prop
##              <dbl> <lgl>                <int>   <dbl>
## 1                1 TRUE                    72 0.0290
## 2                3 TRUE                     7 0.121
## 3               NA TRUE                     7 0.00873
## 4                2 TRUE                     6 0.0193
## 5                9 TRUE                     2 0.00769
```

5.49% of all the crashs involve pedestrians. Driver license type 1 (Valid Oregon license or permit) has the highest rate of crashs (2.9%) that involve pedestrians.

e. Create a data frame that contains the age of drivers and collision type. (Don't print it.) Complete the following:
   - Find the average and median age of drivers.
   - Find the average and median age of drivers by collision type.
   - Create a graph of driver ages.
   - Create a graph of driver ages by collision type.

Draw some conclusions.

```r
collision_type_age <- crash_and_partic %>%
  select(COLLIS_TYP_CD, AGE_VAL)

# Find the average and median age of drivers
collision_type_age %>%
  summarize(mean_age = mean(as.numeric(AGE_VAL)),
            median_age = median(as.numeric(AGE_VAL)))
```

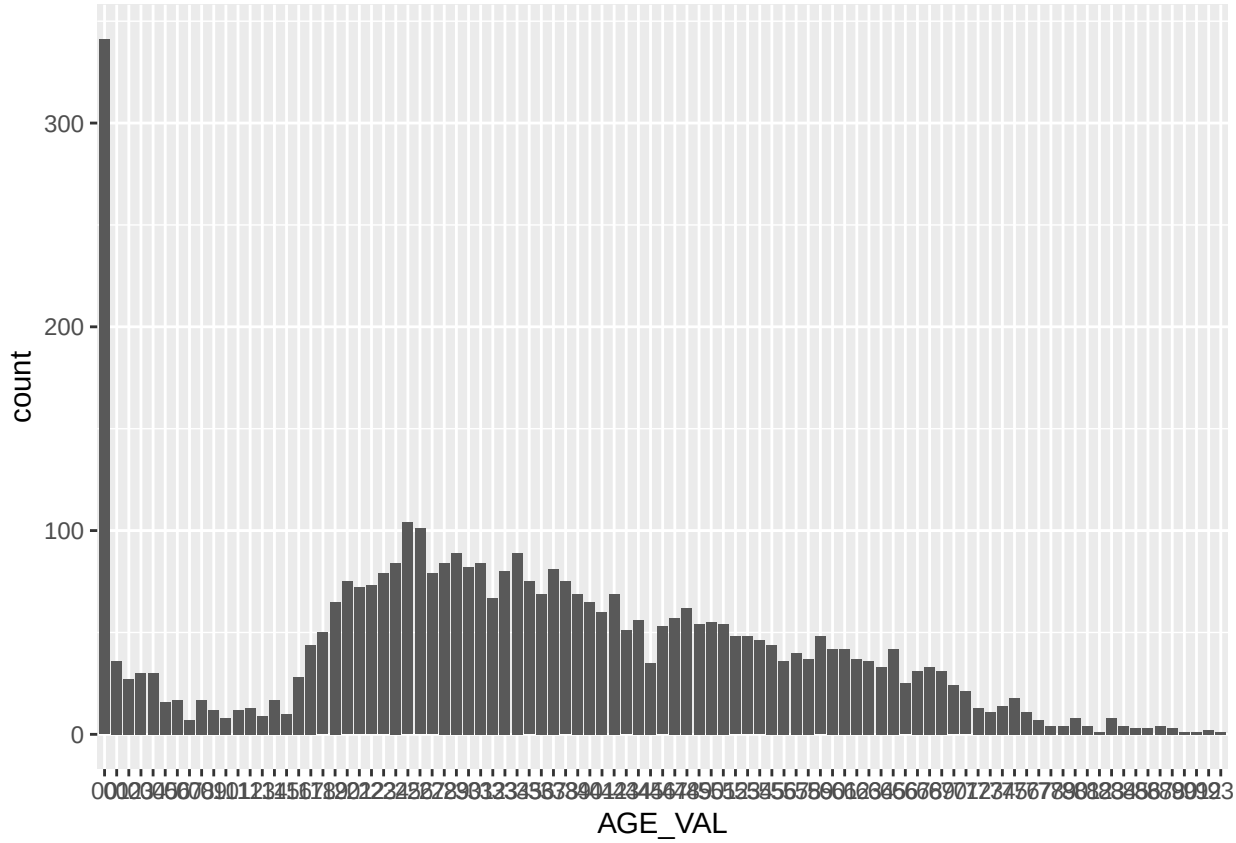
```
## # A tibble: 1 x 2
##   mean_age median_age
##      <dbl>      <dbl>
## 1     34.9         34
```

```r
# Find the average and median age of drivers by collision type
collision_type_age %>%
  group_by(COLLIS_TYP_CD) %>%
  summarize(mean_age = mean(as.numeric(AGE_VAL)),
            median_age = median(as.numeric(AGE_VAL)))
```

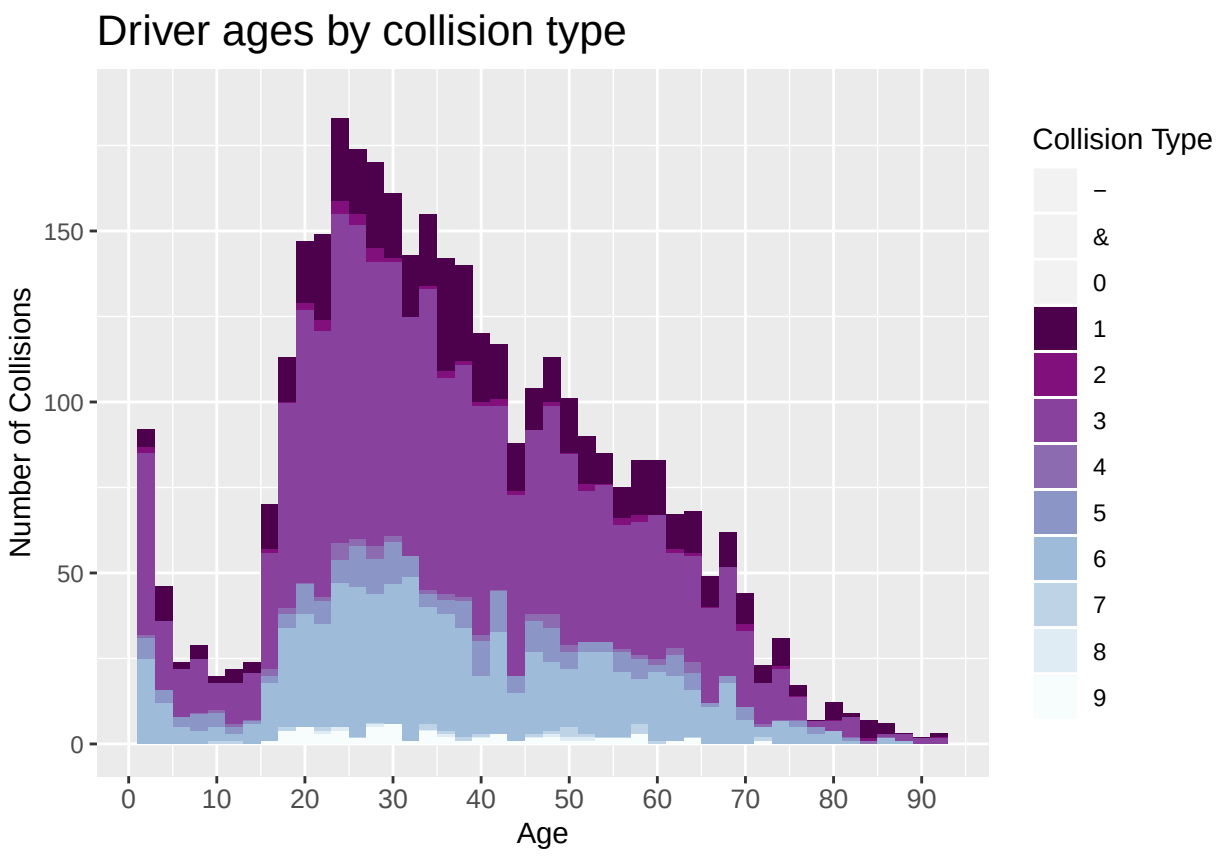
```
## # A tibble: 12 x 3
##    COLLIS_TYP_CD mean_age median_age
##    <chr>            <dbl>      <dbl>
##  1 -                 36.0         37
##  2 &                 49.1         40
##  3 0                 44.7         44
##  4 1                 36.6         36
```

```
##  5 2                 35.8         29
##  6 3                 34.0         33
##  7 4                 35.8         33
##  8 5                 33.3         33
##  9 6                 34.7         33
## 10 7                 35.6         39
## 11 8                 37.4         36
## 12 9                 34.8         31
```

```r
# Create a graph of driver ages
collision_type_age %>%
  ggplot(aes(x = AGE_VAL)) +
  geom_bar()
```




```r
# Create a graph of driver ages by collision type
collision_type_age %>%
  filter(AGE_VAL != "00") %>% #filtering out the unknown age
  ggplot(aes(x = as.numeric(AGE_VAL), fill = COLLIS_TYP_CD)) +
  geom_histogram(binwidth = 2) +
  scale_fill_brewer(palette = 3, direction = -1) +
  scale_x_continuous(breaks = c(0,10,20,30,40,50,60,70,80,90)) +
  theme(plot.title = element_text(size = 16)) +
  theme(axis.text.x = element_text(size = 10)) +
  labs(title = "Driver ages by collision type",
       x = "Age", y = "Number of Collisions", fill = "Collision Type")
```



Problem 3: Chronically Messy Data

a. Turning to the CDC data, let's get a handle of what is represented there. For 2016 (use `YearStart`), how many distinct topics were tracked?

```
CDC %>%
  filter(YearStart == 2016) %>%
  count(Topic) %>%
  nrow()
```

```
## [1] 16
```

16 distinct topics were tracked.

b. Let's study influenza vaccination patterns! Create a dataset that contains the age adjusted prevalence of the "Influenza vaccination among noninstitutionalized adults aged >= 18 years" for Oregon and the US from 2010 to 2016.

```
influenza_vacc_prevalance <- CDC %>%
  filter(YearStart %in% c(2010, 2011, 2012, 2013, 2014, 2015, 2016),
         Question == "Influenza vaccination among noninstitutionalized adults aged >= 18 years",
         DataValueType == "Age-adjusted Prevalence") %>%
  select(YearStart, LocationAbbr, DataValue, DataValueType, Question) %>%
  arrange(YearStart) %>%
  pivot_wider(names_from = DataValueType,
              values_from = DataValue) %>%
  as.data.frame()
```

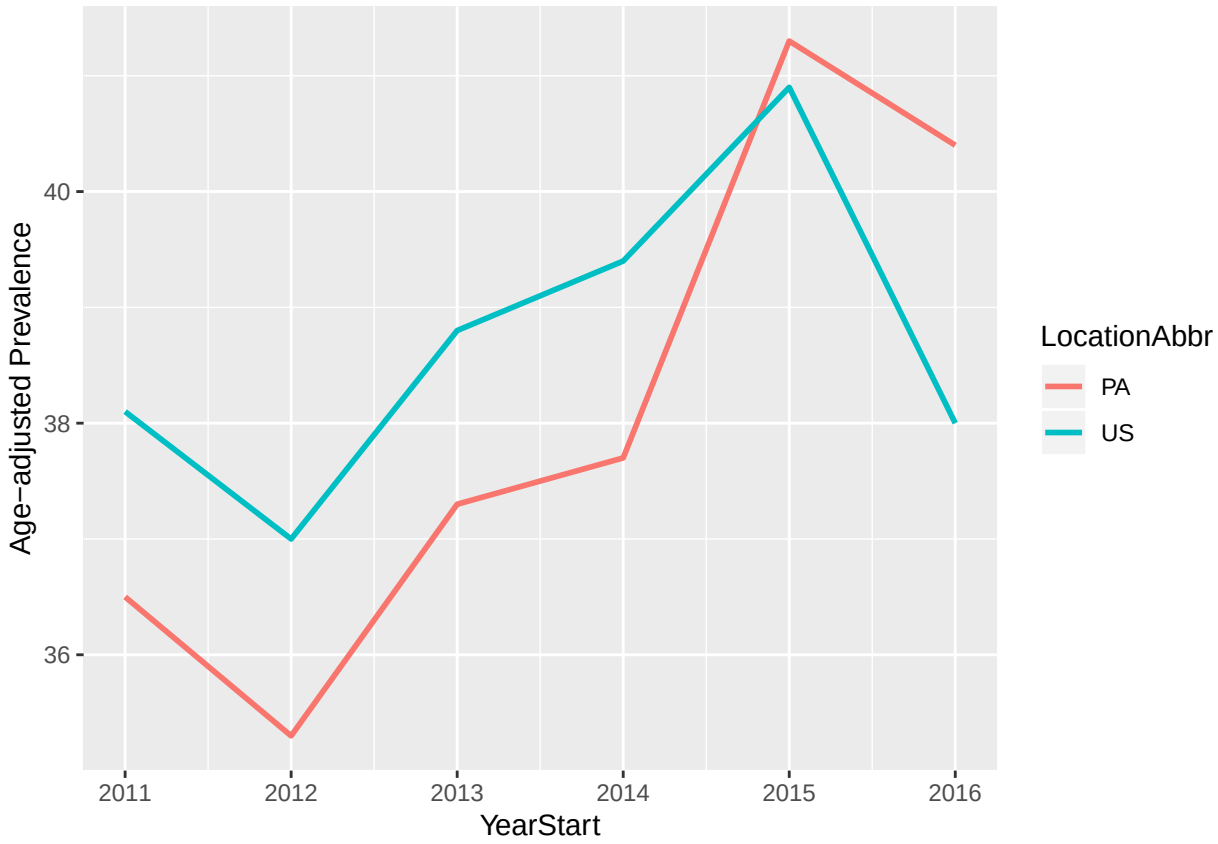
```
influenza_vacc_oregon_us <- influenza_vacc_prevalance %>%
  filter(LocationAbbr%in% c("OR", "US"))

influenza_vacc_oregon_us
```

```
##    YearStart LocationAbbr
## 1       2011           OR
## 2       2011           US
## 3       2012           OR
## 4       2012           US
## 5       2013           OR
## 6       2013           US
## 7       2014           OR
## 8       2014           US
## 9       2015           OR
## 10      2015           US
## 11      2016           US
## 12      2016           OR
##                                                                  Question
## 1  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 2  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 3  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 4  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 5  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 6  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 7  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 8  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 9  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 10 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 11 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 12 Influenza vaccination among noninstitutionalized adults aged >= 18 years
##    Age-adjusted Prevalence
## 1                      33.0
## 2                      38.1
## 3                      33.0
## 4                      37.0
## 5                      33.6
## 6                      38.8
## 7                      35.4
## 8                      39.4
## 9                      38.2
## 10                     40.9
## 11                     38.0
## 12                     34.4
```

c. Create a graph comparing the immunization rates of Pennsylvania and the US. Comment on the observed trends in your graph

```
influenza_vacc_prevalance %>%
  filter(LocationAbbr %in% c("PA", "US")) %>%
  ggplot(aes(x = YearStart, y = `Age-adjusted Prevalence`,
             color = LocationAbbr)) +
  geom_line(size = 1)
```



For both Pennsylvania and the US, there is a steady increase in immunization rates from 2011 to 2015 (with the exception of a slight drop in 2012) and there was a huge decrease in immunization rates in 2016.

In general, the immunization rate in Pennslyvania is lower than that in the US with the exception of year 2015 and 2016.

d. Let's see how immunization rates vary by region of the country. Join the regional dataset to our CDC dataset so that we have a column signifying the region of the country.

```
CDC <- left_join (CDC, USregions, by = c("LocationDesc" = "State"))
```

e. Why are there NAs in the region column of the new dataset?

A left join in question (d) returns all rows from CDC, and all columns from CDC and USregions. However, some location names used in the CDC dataset (55 locations for the LocationDesc variable) are not present in the dataset USregions (50 locations for the State variable), so this creates NAs for the variable Region in the new dataset.

f. Create a dataset that contains the age adjusted influenza immunization rates in 2016 for each state in the country and sort it by highest immunization to lowest. Which state has the highest immunization?

```
immunization_by_state <- CDC %>%
  filter(YearStart == 2016,
         Question == "Influenza vaccination among noninstitutionalized adults aged >= 18 years",
         DataValueType == "Age-adjusted Prevalence") %>%
  select(YearStart, LocationAbbr, DataValue, DataValueType, Question, Region) %>%
  pivot_wider(names_from = DataValueType,
              values_from = DataValue) %>%
  arrange(desc(`Age-adjusted Prevalence`)) %>%
  as.data.frame()
```

```
immunization_by_state
```

```
##    YearStart LocationAbbr
## 1       2016           SD
## 2       2016           RI
## 3       2016           IA
## 4       2016           NE
## 5       2016           NC
## 6       2016           MN
## 7       2016           CO
## 8       2016           MD
## 9       2016           VA
## 10      2016           CT
## 11      2016           WV
## 12      2016           ND
## 13      2016           MO
## 14      2016           MA
## 15      2016           PA
## 16      2016           NH
## 17      2016           VT
## 18      2016           DE
## 19      2016           OK
## 20      2016           WA
## 21      2016           NM
## 22      2016           NY
## 23      2016           ME
## 24      2016           HI
## 25      2016           US
## 26      2016           DC
## 27      2016           NJ
## 28      2016           AR
## 29      2016           MT
## 30      2016           TX
## 31      2016           OH
## 32      2016           KY
## 33      2016           UT
## 34      2016           KS
## 35      2016           IN
## 36      2016           MS
## 37      2016           SC
## 38      2016           CA
## 39      2016           IL
## 40      2016           WI
## 41      2016           AL
## 42      2016           MI
## 43      2016           GA
## 44      2016           TN
## 45      2016           AK
## 46      2016           OR
## 47      2016           AZ
## 48      2016           WY
## 49      2016           ID
## 50      2016           LA
## 51      2016           FL
```

```
## 52      2016           NV
## 53      2016           GU
## 54      2016           PR
## 55      2016           VI
##                                                                    Question
## 1  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 2  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 3  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 4  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 5  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 6  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 7  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 8  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 9  Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 10 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 11 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 12 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 13 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 14 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 15 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 16 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 17 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 18 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 19 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 20 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 21 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 22 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 23 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 24 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 25 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 26 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 27 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 28 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 29 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 30 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 31 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 32 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 33 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 34 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 35 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 36 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 37 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 38 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 39 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 40 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 41 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 42 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 43 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 44 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 45 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 46 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 47 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 48 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 49 Influenza vaccination among noninstitutionalized adults aged >= 18 years
```

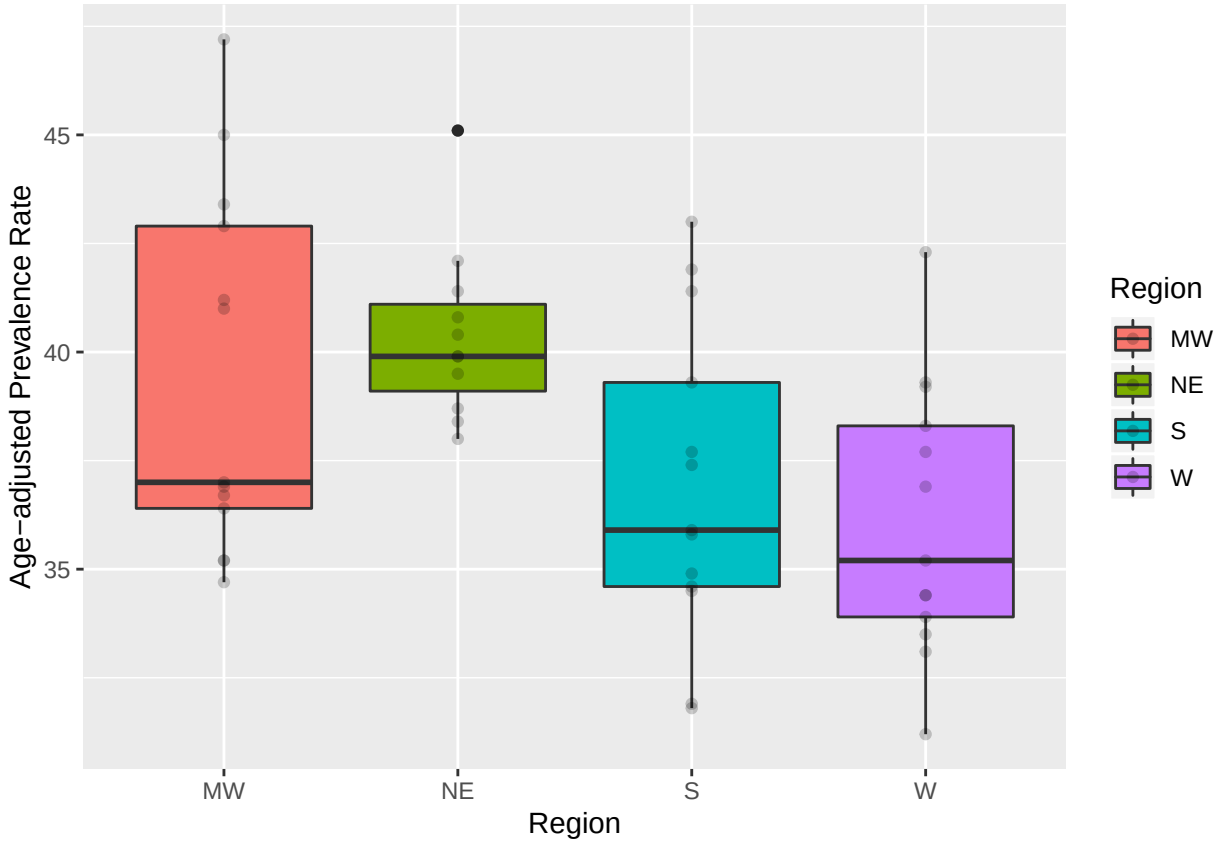
```
## 50 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 51 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 52 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 53 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 54 Influenza vaccination among noninstitutionalized adults aged >= 18 years
## 55 Influenza vaccination among noninstitutionalized adults aged >= 18 years
##    Region Age-adjusted Prevalence
## 1      MW                    47.2
## 2      NE                    45.1
## 3      MW                    45.0
## 4      MW                    43.4
## 5       S                    43.0
## 6      MW                    42.9
## 7       W                    42.3
## 8      NE                    42.1
## 9       S                    41.9
## 10     NE                    41.4
## 11      S                    41.4
## 12     MW                    41.2
## 13     MW                    41.0
## 14     NE                    40.8
## 15     NE                    40.4
## 16     NE                    39.9
## 17     NE                    39.9
## 18     NE                    39.5
## 19      S                    39.3
## 20      W                    39.3
## 21      W                    39.2
## 22     NE                    38.7
## 23     NE                    38.4
## 24      W                    38.3
## 25   <NA>                    38.0
## 26   <NA>                    38.0
## 27     NE                    38.0
## 28      S                    37.7
## 29      W                    37.7
## 30      S                    37.4
## 31     MW                    37.0
## 32     MW                    36.9
## 33      W                    36.9
## 34     MW                    36.7
## 35     MW                    36.4
## 36      S                    35.9
## 37      S                    35.8
## 38      W                    35.2
## 39     MW                    35.2
## 40     MW                    35.2
## 41      S                    34.9
## 42     MW                    34.7
## 43      S                    34.6
## 44      S                    34.5
## 45      W                    34.4
## 46      W                    34.4
## 47      W                    33.9
```

```
## 48      W                    33.5
## 49      W                    33.1
## 50      S                    31.9
## 51      S                    31.8
## 52      W                    31.2
## 53   <NA>                    30.9
## 54   <NA>                    23.9
## 55   <NA>                    14.9
```

South Dakota has the highest immunization rate.

g. Construct a graphic of the 2016 influenza immunization rates by region of the country. Don't include locations without a region. Comment on your graphic.

```
immunization_by_state %>%
  drop_na() %>%
  ggplot(aes(x = Region, y = `Age-adjusted Prevalence`, fill = Region)) +
  geom_boxplot() + geom_point(width = 1, alpha = 0.2) +
  labs(y = "Age-adjusted Prevalence Rate")
```



In 2016, NE has the highest median of the influenza immunization rates compared to other regions, and MW has the largest IQR which means the immunization rates are the most spread. FOR S and W, the influenza immunization rates are similar in terms of the average (i.e. median) and variability (i.e. IQR).

Problem 4: Tidying Data Like a Boss

I was amazed by the fact that many of the FiveThirtyEight datasets are actually not in a perfectly *tidy* format. Let's tidy up this dataset related to polling.

a. Why is this data not currently in a tidy format?

Some rows contain mutiple observations.

```
polls
```

```
## # A tibble: 1,529 x 4
##    subgroup  modeldate dem_estimate rep_estimate
##    <chr>     <chr>            <dbl>        <dbl>
##  1 All polls 9/18/2018         48.8         39.8
##  2 All polls 9/17/2018         49.0         39.9
##  3 All polls 9/16/2018         49.0         39.9
##  4 All polls 9/15/2018         49.0         39.9
##  5 All polls 9/14/2018         48.9         39.8
##  6 All polls 9/13/2018         48.8         39.7
##  7 All polls 9/12/2018         48.8         39.6
##  8 All polls 9/11/2018         48.5         39.9
##  9 All polls 9/10/2018         48.4         39.9
## 10 All polls 9/9/2018          48.4         39.9
## # ... with 1,519 more rows
```

b. Create a tidy dataset of the `All polls` subgroup.

```
polls %>% pivot_wider(names_from = subgroup,
              values_from = c(dem_estimate, rep_estimate))
```

```
## # A tibble: 522 x 7
##    modeldate `dem_estimate_A~ dem_estimate_Vo~ dem_estimate_Ad~ `rep_estimate_A~
##    <chr>                <dbl>            <dbl>            <dbl>            <dbl>
##  1 9/18/2018             48.8               NA               NA             39.8
##  2 9/17/2018             49.0               NA               NA             39.9
##  3 9/16/2018             49.0               NA               NA             39.9
##  4 9/15/2018             49.0               NA               NA             39.9
##  5 9/14/2018             48.9               NA               NA             39.8
##  6 9/13/2018             48.8               NA               NA             39.7
##  7 9/12/2018             48.8               NA               NA             39.6
##  8 9/11/2018             48.5               NA               NA             39.9
##  9 9/10/2018             48.4               NA               NA             39.9
## 10 9/9/2018              48.4               NA               NA             39.9
## # ... with 512 more rows, and 2 more variables: rep_estimate_Voters <dbl>,
## #   rep_estimate_Adults <dbl>
```

c. Now let's create a new untidy version of `polls`. Focusing just on the estimates for democrats, create a data frame where each row represents a subgroup (given in column 1) and the rest of the columns are the estimates for democrats by date.

```
untidy_poll <- polls %>%
  select(!rep_estimate) %>%
  pivot_wider(names_from = modeldate,
              values_from = dem_estimate)

untidy_poll %>%
  ggplot(aes())
```

d. Why might someone want to transform the data like we did in part c?

- They are interested in comparing estimates of votes from the three subgroups on a particular day
- They just want to scroll horizontally to see if there's a general trend of increasing or decreasing number of estimates.

Problem 5: YOUR TURN!

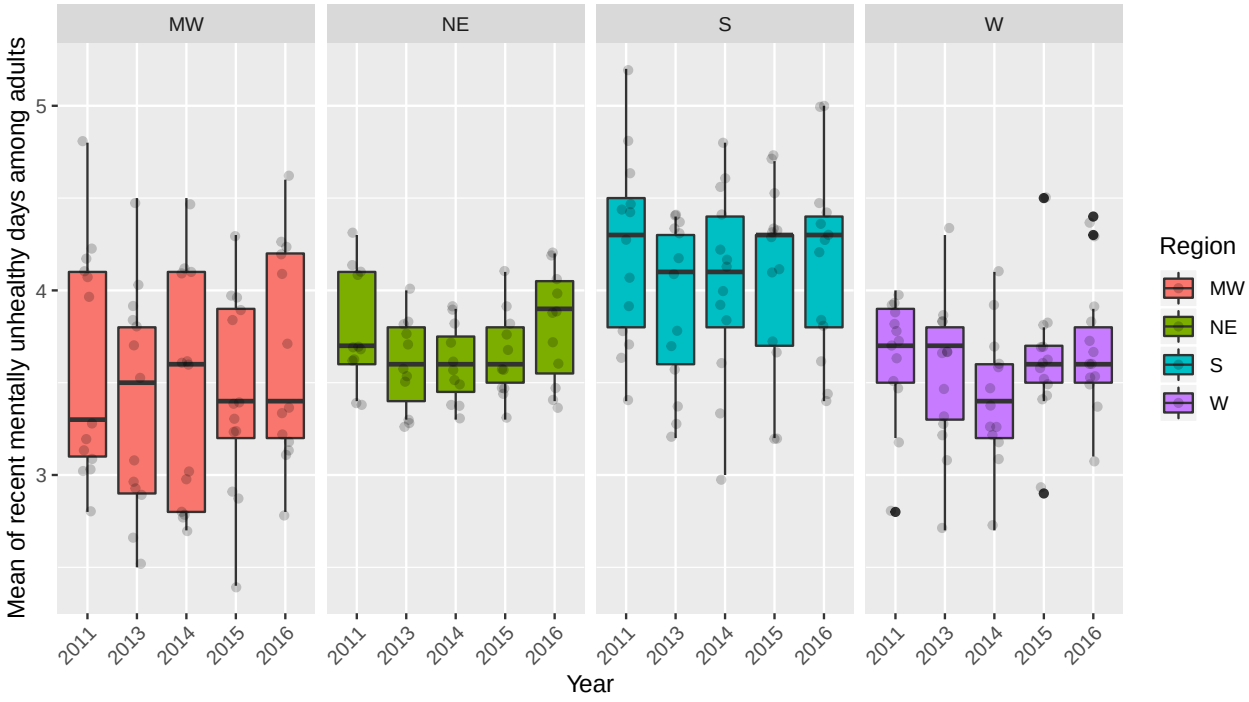
Now it is your turn. Pick one (or multiple) of the datasets used on this lab. Ask a question of the data. Do some data wrangling to produce statistics (use at least two wrangling verbs) and a graphic to answer the question. Then comment on any conclusions you can draw about your question.

- For the CDC dataset: I want to explore more for the question "Recent mentally unhealthy days among adults aged >= 18 years". I'm curious to learn about the trend of the mean number of mentally unhealthy days among adults from year 2010 to 2016, and how that differs in each region.

```
mental_health_region <- CDC %>%
  filter(YearStart %in% c(2010,2011,2012,2013,2014,2015,2016),
         Question == "Recent mentally unhealthy days among adults aged >= 18 years",
         DataValueType == "Mean") %>%
  select(YearStart, DataValue, DataValueType, Question, Region, LocationAbbr) %>%
  pivot_wider(names_from = DataValueType,
              values_from = DataValue) %>%
  arrange(desc(YearStart)) %>%
  as.data.frame()

mental_health_region %>%
  drop_na() %>%
```

```
ggplot(aes(x = factor(YearStart), y = Mean, fill = Region)) +
geom_boxplot() +
geom_jitter(width = 0.1, alpha = 0.2) +
xlab("Year") +
ylab("Mean of recent mentally unhealthy days among adults") +
facet_wrap(~Region, ncol = 4) +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



In S, adults consistanly experience more mentally unhealthy days compared to other regions. For all regions, it seems that there the mean number of mentally unstable days stayed the same throughout the years.