Project 3

Sookja Kang, sk26949

This is the dataset used in this project:

```
measles <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2
```

Link to the dataset: *https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-02-25*

Part 1

Question:

what is the relationship between schools' MMR vaccination rate and overall vaccination rate for each school type in California?

Introduction:

I am using the `Measles` dataset that contains 66,113 schools across the US states. In this dataset, each row represents an individual school and each column provides a school's information and its vaccine data from 2017 - 2019 (a total of 16 columns: index ID, school's state, school academic year, school name, school type[public, private, charter], city, county, district, enrollment, MMR vaccination rate, overall vaccination rate, percentage of students exempted from vaccination for religious reasons, percentage of students exempted from vaccination for medical reasons, percentage of students exempted from vaccination for personal reasons, and two unknown variables[no description on the website]). I am interested in schools in California so I used 'filter()' function to make a subset of `measles_c` to answer the part 1 question.

To understand the relationship between schools' MMR vaccination rate and overall vaccination rate for each school type in California, I am going to work with *mmr* and *overall* columns.

1. mmr: each school's Measles, Mumps, and Rubella (MMR) vaccination rate
2. overall: each school's overall vaccination rate

Approach:

My approach is to understand the relationships between MMR vaccination rate and overall vaccination rate for each school type in California. First, I am going to make a table to present two linear regression models of overall vaccination rate against MMR vaccination rate for public and private school types in California. Next, I am going to use separated scatter plots to visualize the overall vaccination rate against the MMR vaccination rate for each school type. These plots will include regression lines.

1. nest(): to nest data by the `type` column
2. mutate(): to make a new column (total_number) using the n column that created from count()
3. map(): to fit linear models to two school types (public & private)
4. lm(): to fit a linear model using data of overall vaccination rate against MMR vaccination rate
5. unnest(): to unnest output from glance
6. glue(): to place R^2 and P variables into a text string
7. sifnif(): to round to 3 significant digits for R^2 and P-value
8. select(): to select variables: `type, overall, mmr, and values`
9. ggplot(): to make a scatter plot using geom_point()
10. geom_text(): to insert labels of R^2 and P-value in the scatter plots

11. geomsmooth(method = "lm"): to plot a linear regression on each facet of school types
12. facet_wrap():to create scatter plot facets for public and private schools

Analysis:

```
measles_c <- measles %>%
  filter(state == "California") %>%
  filter(!is.na(mmr))
head(measles_c)
```

```
## # A tibble: 6 x 16
##   index state year  name  type  city  county district enroll   mmr overall xrel
##   <dbl> <chr> <chr> <chr> <chr> <chr> <chr>  <lgl>     <dbl> <dbl>   <dbl> <lgl>
## 1     1 Cali~ 2018~ Abby~ Publ~ Teme~ River~ NA          137    99      96 NA
## 2     2 Cali~ 2018~ Abra~ Publ~ Sant~ Orange NA          135    99      99 NA
## 3     2 Cali~ 2018~ Abra~ Publ~ Sant~ Orange NA          135    99      99 NA
## 4     2 Cali~ 2018~ Abra~ Publ~ Sant~ Orange NA          135    99      99 NA
## 5     2 Cali~ 2018~ Abra~ Publ~ Sant~ Orange NA          135    99      99 NA
## 6     2 Cali~ 2018~ Abra~ Publ~ Sant~ Orange NA          135    99      99 NA
## # ... with 4 more variables: xmed <dbl>, xper <dbl>, lat <dbl>, lng <dbl>
```

```
measles_c_fit <- measles_c %>%
  nest(data = -type) %>%
  mutate(fit = map(data, ~lm(overall ~ mmr, data = .x)),
         glance_out = map(fit, glance)) %>%
  unnest(cols = glance_out)
measles_c_fit
```

```
## # A tibble: 2 x 15
##   type  data  fit   r.squared adj.r.squared sigma statistic p.value    df
##   <chr> <lis> <lis>     <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>
## 1 Publ~ <tib~ <lm>      0.986         0.986  2.30   969491.       0     1
## 2 Priv~ <tib~ <lm>      0.993         0.993  3.80   418360.       0     1
## # ... with 6 more variables: logLik <dbl>, AIC <dbl>, BIC <dbl>,
## #   deviance <dbl>, df.residual <int>, nobs <int>
```

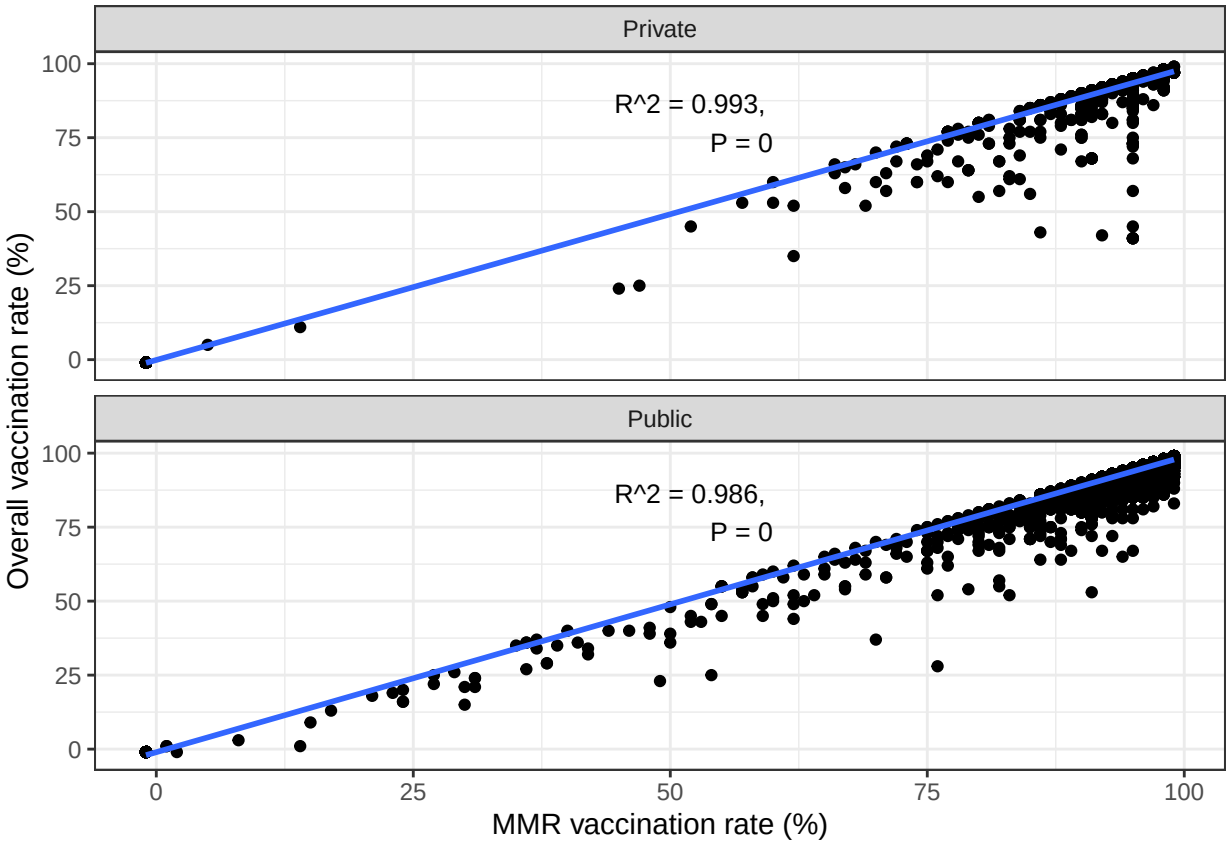
```
label_data <- measles_c_fit %>%
  mutate(overall = 80,
         mmr = 60,
         values = glue("R^2 = {signif(r.squared, 3)},
                       P = {signif(p.value, 3)}")) %>%
  select(type, overall, mmr,  values)

label_data
```

```
## # A tibble: 2 x 4
##   type    overall   mmr values
##   <chr>     <dbl> <dbl> <glue>
## 1 Public       80    60 "R^2 = 0.986, \nP = 0"
## 2 Private      80    60 "R^2 = 0.993, \nP = 0"
```

```
measles_c %>%
  ggplot(aes(mmr, overall)) +
  geom_point() +
  geom_text(data = label_data, aes(label = values),
    size = 10/.pt, hjust = 1) +
  geom_smooth(method = "lm", se = FALSE) +
```

```
  facet_wrap(vars(type), ncol=1) +
  scale_x_continuous(name = "MMR vaccination rate (%)") +
  scale_y_continuous(name = "Overall vaccination rate (%)") +
  theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Discussion:

Both the regression summary table and plots with linear regression lines show that there is a statistically significant and strong positive relationship between MMR vaccination rate and overall vaccine rate in both public and private schools in California. 99.3% of the variance for private schools and 98.6% variance for public schools in overall vaccination rates are explained by MMR vaccination rates in this regression model. As a result, the MMR vaccination rate is highly and positively correlated with the overall vaccine rate at both types of schools in California.

Part 2

Question:

What are the top 5 States that contain high numbers of schools reporting vaccine relevant information? How do the proportions of the top 5 States change between private and public schools?

Introduction:

I am going to use the same dataset of `measles` (including 66,113 schools with 16 columns) that I used for the part 1 question. To answer this part 2 question, I am going to use the following two variables. 1. state: school's state 2. type: three different school types (Public, Private, Charter).

Approach: *Your approach here.*

My approach is to 1) identify the top 5 states with high numbers of schools reporting vaccine information and 2) understand the proportion changes of top 5 States by different school types. To answer the first question, horizon bars (geom_bar()) will be used to visualize amounts of each state's school number. Additionally, I will use facet_wrap() to facet by two different school types. This will make it easy to compare states' different numbers across the different school types. Then, I will make a pie chart to visually compare the top 5 states' proportion changes across the school types. This pie chart highlights each slice's proportion (presenting each state) in a whole circle. During the data wrangling, `NA` cells from the type variable will be removed using the filter function. Also, among the three different school types, `Charter` will be excluded since it is not relevant to the question.

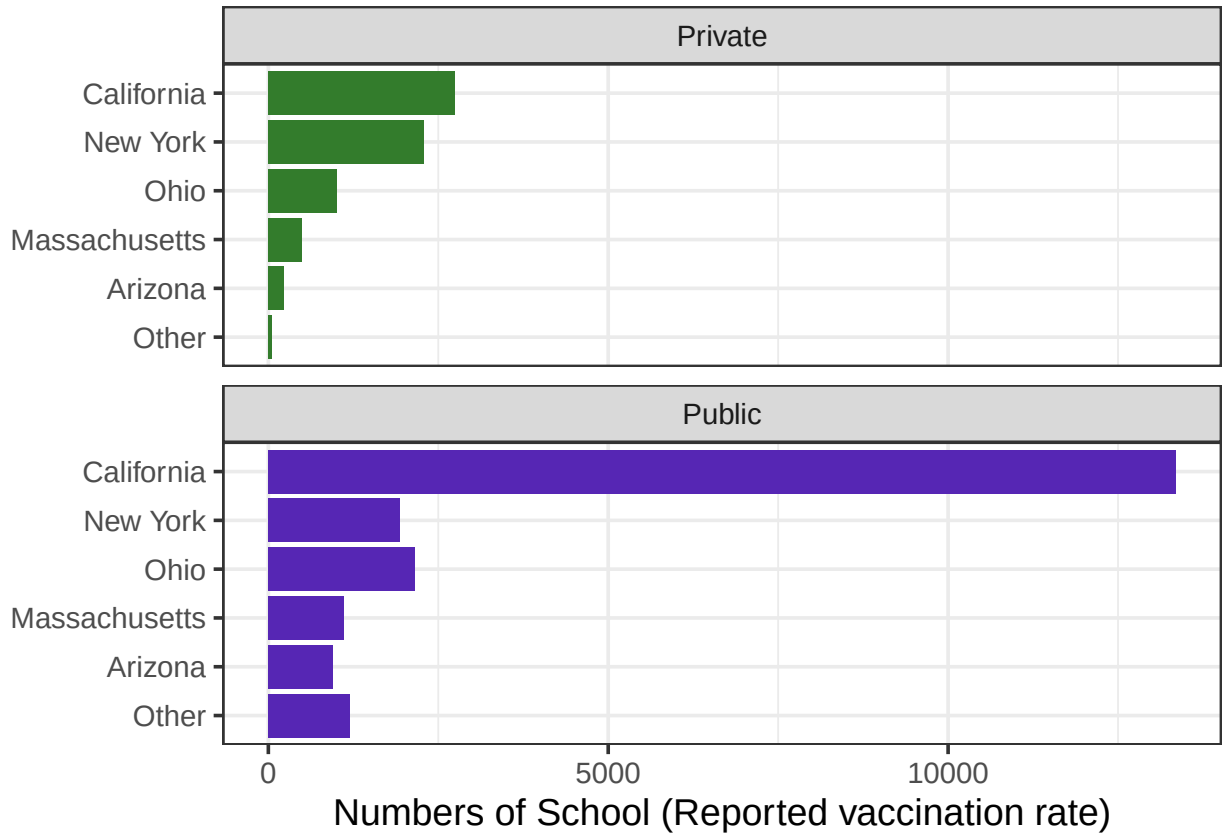
To present horizontal bars: mutate(): make new columns fct_lump_n(): to keep 5 most frequent states and lump all others into "Other" fct_infreq(): to reorder based on frequency of `state` from most to least fct_rev(): to reverse the order of `state` geom_bar(): to count before plotting bars

To make pie charts, the following functions will be used:

1. filter():
    a. to extract only top 5 states ("California", "New York", "Ohio", "Massachusetts", "Arizona")
    b. to remove "NA" values in the type column
2. count(): tocount numbers for each subcategory of `state` and `type`
3. mutate(): make a new column of `total_number` using the n column
4. arrange() and -desc(): to sort the total_number by ascending count
5. fct_reorder(): to reorder the `state` column by the total_number
6. group_by(): to group by the `type` column
7. mutate(): to make new columns
    a. the end_angle, start_angle, mid_angle for each pie slice
    b. horizontal and vertical justifications for outer labels
8. ggplot(): to plot the pie_data
9. geom_arc_bar(): to specify the exact location of the pie center in the x-y plane
10. geom_text(): to insert and locate values for pie slices
11. coord_fixed(): to ensure that the pie is round
12. facet_wrap(): to create pie chart facets for each school type
13. theme_void(): to remove the x-y plane

Analysis:

```
measles_f <- measles %>%
  filter(!is.na(type))

measles_f$type <- factor(measles_f$type,
                         levels = c("Private", "Public"))

measles_f %>%
  filter(type != "NA") %>%
  mutate(state = fct_lump_n(fct_infreq(state), 5)) %>%
  ggplot(aes(y = fct_rev(state), fill = type)) +
  geom_bar(show.legend = FALSE) +
  scale_x_continuous(name = "Numbers of School (Reported vaccination rate)") +
  scale_y_discrete(name = NULL) +
  scale_fill_manual(values = c('Private' = "#337C2C", 'Public' = "#5626B4")) +
  facet_wrap(vars(type), ncol = 1) +
  theme_bw(14)
```



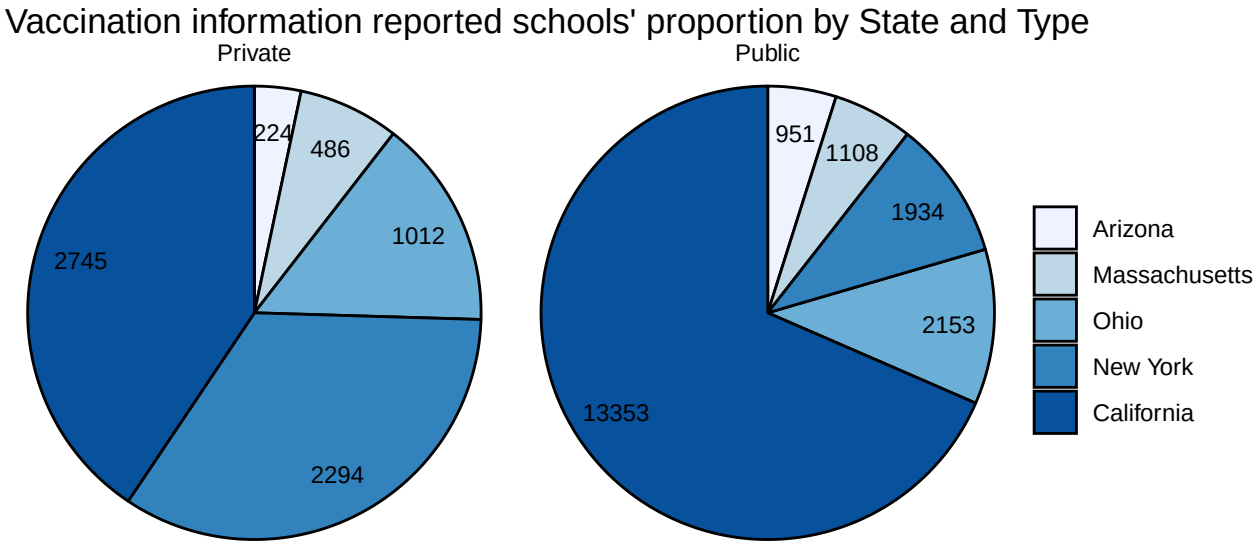

```
#Pie Chart#
measles_f_s <- filter(measles_f,
                      state %in% c("California", "New York", "Ohio", "Massachusetts", "Arizona")) %>%
  filter(type != "NA")

measle_data <- measles_f_s %>%
  count(state, type) %>%
  mutate(total_number = n) %>%
  arrange(-desc(total_number)) %>%
  mutate(state = fct_reorder(state, total_number))

measle_data
```

```
## # A tibble: 10 x 4
##    state         type        n total_number
##    <fct>         <fct>   <int>        <int>
##  1 Arizona       Private   224          224
##  2 Massachusetts Private   486          486
##  3 Arizona       Public    951          951
##  4 Ohio          Private  1012         1012
##  5 Massachusetts Public   1108         1108
##  6 New York      Public   1934         1934
##  7 Ohio          Public   2153         2153
##  8 New York      Private  2294         2294
##  9 California    Private  2745         2745
## 10 California    Public  13353        13353
```

```r
pie_data <- measle_data %>%
  group_by(type) %>%
  mutate(end_angle = 2*pi*cumsum(n)/sum(n),
         start_angle = lag(end_angle, default = 0),
         mid_angle = 0.5*(start_angle + end_angle),
         hjust = ifelse(mid_angle > pi, 1, 0),
         vjust = ifelse(mid_angle < pi/2 | mid_angle > 3*pi/2, 0, 1))
pie_data
```

```
## # A tibble: 10 x 9
## # Groups:   type [2]
##    state    type      n total_number end_angle start_angle mid_angle hjust vjust
##    <fct>    <fct> <int>        <int>     <dbl>       <dbl>     <dbl> <dbl> <dbl>
##  1 Arizona  Priv~   224          224     0.208       0         0.104     0     0
##  2 Massach~ Priv~   486          486     0.660       0.208     0.434     0     0
##  3 Arizona  Publ~   951          951     0.306       0         0.153     0     0
##  4 Ohio     Priv~  1012         1012     1.60        0.660     1.13      0     0
##  5 Massach~ Publ~  1108         1108     0.663       0.306     0.485     0     0
##  6 New York Publ~  1934         1934     1.29        0.663     0.975     0     0
##  7 Ohio     Publ~  2153         2153     1.98        1.29      1.63      0     1
##  8 New York Priv~  2294         2294     3.73        1.60      2.67      0     1
##  9 Califor~ Priv~  2745         2745     6.28        3.73      5.01      1     0
## 10 Califor~ Publ~ 13353        13353     6.28        1.98      4.13      1     1
```

```r
ggplot(pie_data, aes(x0 = 0, y0 = 0, r0 = 0, r = 1,
                     start = start_angle, end = end_angle,
                     fill = state)) +
  geom_arc_bar() +
  geom_text(size = 3,
            aes(x = 0.8 * sin(mid_angle),
                y = 0.8 * cos(mid_angle),
                label = total_number) )+
  coord_fixed() +
  facet_wrap(~type) +
  theme_void() +
  scale_fill_brewer(name = NULL) +
  ggtitle("Vaccination information reported schools' proportion by State and Type")
```



Discussion:

California, New York, Ohio, Massachusetts, and Arizona are the states with the highest numbers of schools reporting vaccine information for both private and public schools. The orders of the top 5 states stay similar in the private and public schools except for New York and Ohio. California has the highest numbers of schools in both private and public schools. From these horizontal bar charts, I can visually compare the proportion changes of the top 5 states by school type.

The pie charts also present the proportion changes of the top 5 states across the school types. Comparing to the bar charts, the pie charts show the same results of proportion changes. It is easy to understand how much proportion of each slice (state) is in a whole circle (school type). Also, these pie charts allow comparing the proportions across the two different school types. However, as you see in my pie charts, these pie charts do not accurately show the raw number differences across the two different school types. Similar sizes of pie slices across the school types show huge differences in terms of raw numbers. Using pie charts to compare proportion changes with a huge number differences across the different groups might not accurately present the real differences. In this case, horizontal bar charts can be suitable to show differences across groups.