PS2

Frances Aponte

Data visualization problem set

Measles

1. Load the dslabs package and figure out what is in the us_contagious_diseases dataset. Create a data frame, call it avg, that has a column for year, and a rate column containing the cases of Measles per 10,000 people per year in the US. Because we start in 1928, exclude Alaska and Hawaii. Make sure to take into account the number of weeks reporting each year. If a week was not report, it should not be included in the calculation of the rate.

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----------------------- tidyverse 2.0.0 --
v dplyr     1.1.1     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.1     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(dslabs)
```

```
#List of variables
colnames(us_contagious_diseases)
```

```
[1] "disease"         "state"           "year"            "weeks_reporting"
[5] "count"           "population"
```

```
# List of Diseases
unique(us_contagious_diseases$disease)
```

```
[1] Hepatitis A Measles     Mumps       Pertussis   Polio       Rubella
[7] Smallpox
Levels: Hepatitis A Measles Mumps Pertussis Polio Rubella Smallpox
```

```
#filter the rows for measles, remove alaska and hawaii, weeks reporting greater than 0 is

avg<- us_contagious_diseases |> filter(disease == "Measles" & !state %in% c("Alaska","Hawa
  group_by(year) |>
  summarize(rate = sum(count/weeks_reporting*52, na.rm = TRUE)/sum(population)*10000)

avg
```

```
# A tibble: 75 x 2
    year  rate
   <dbl> <dbl>
 1  1928  43.2
 2  1929  29.6
 3  1930  33.3
 4  1931  37.5
 5  1932  32.5
 6  1933  31.1
 7  1934  60.9
 8  1935  58.1
 9  1936  23.6
10  1937  24.6
# i 65 more rows
```

```
#Year
min(avg$year)
```

```
[1] 1928
```

```r
max(avg$year)
```

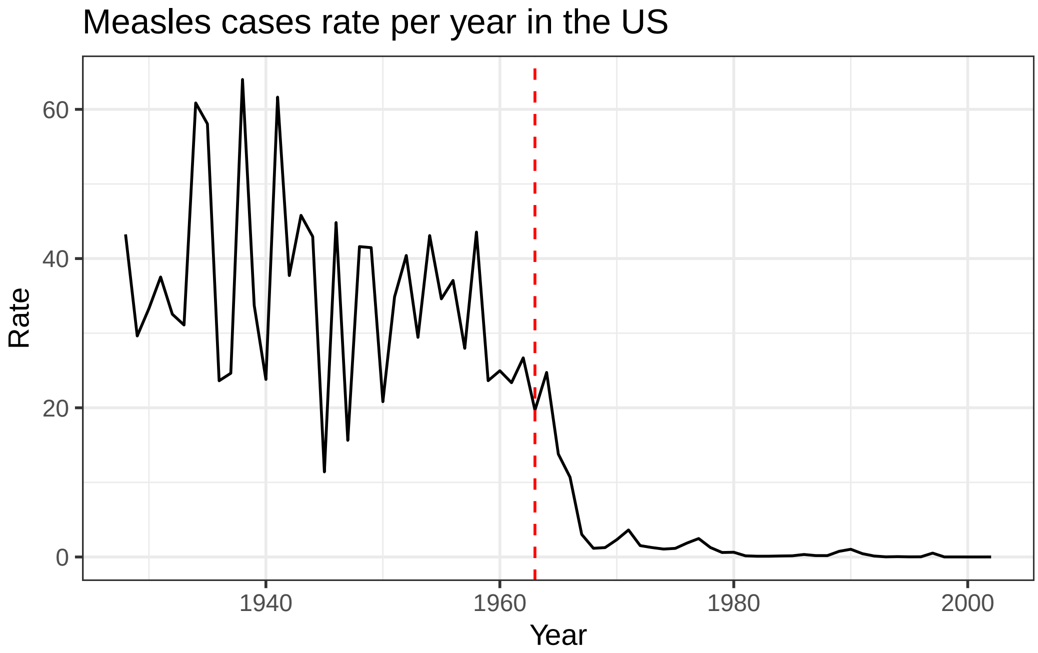
```
[1] 2002
```

```r
# we are dividing by weeks reporting to create a rate with the counts because not all stat

#table(avg$disease)
```

2. Use the data frame avg to make a trend plot showing the cases rate for Measles per year. Add a vertical line showing the year the Measles vaccines was introduced.

```r
avg |>
  ggplot(aes(year,rate)) +
  geom_line() +
  geom_vline(xintercept = 1963, linetype = "dashed", color = "red") +
  labs(x="Year", y="Rate", title="Measles cases rate per year in the US") +
  theme_bw()
```

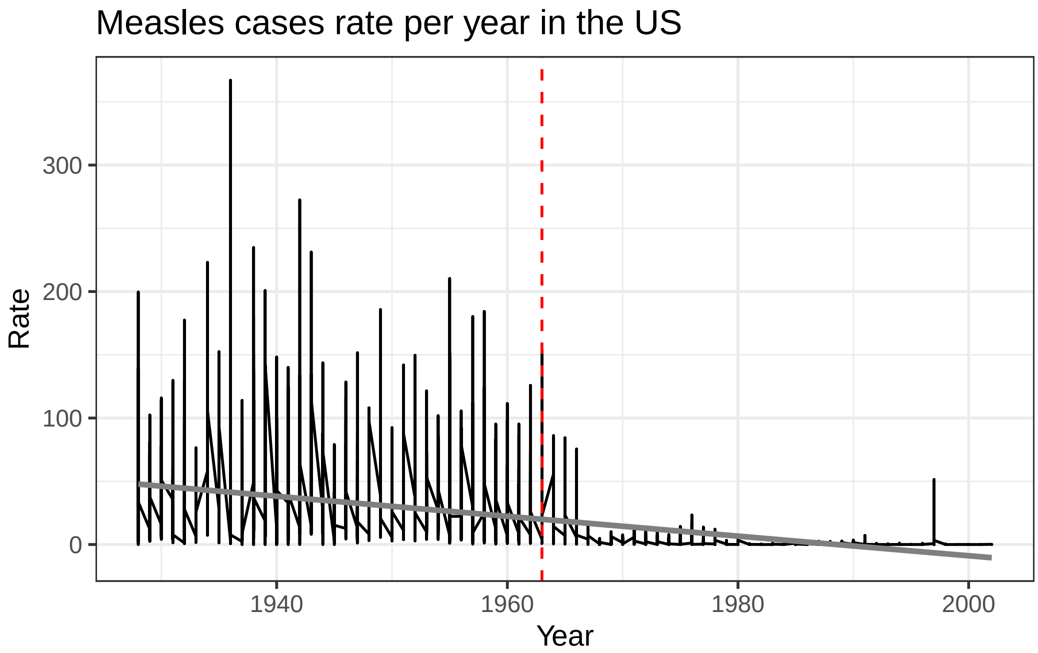


3. Add a grey trend line for each state to the plot above. Use a transformation that keeps the high rates from dominating the figure.

```
avg<- us_contagious_diseases |> filter(disease == "Measles" & !state %in% c("Alaska","Hawa
  group_by(year,state) |>
  summarize(rate = sum(count/weeks_reporting*52, na.rm = TRUE)/sum(population)*10000)
```

```
`summarise()` has grouped output by 'year'. You can override using the
`.groups` argument.
```

```
avg |>
  ggplot(aes(year, rate)) +
  geom_line() +
  geom_vline(xintercept = 1963, linetype = "dashed", color = "red") +
  geom_smooth(method = "lm", formula = y ~ log(x), se = FALSE, color = "grey50") +
  labs(x="Year", y="Rate", title="Measles cases rate per year in the US") +
  theme_bw()
```



4. In the plot above we can't tell which state is which curve. Using color would be challenging as it is hard if not impossible to find 48 colors humans can distinguish. To make a plot where you can compare states knowing which is which, use one of the axis for state and the other for year, and then use hue or intensity as a visual cue for rates. Use a sqrt transformation to avoid the higher rates taking up all the color scale. Use grey

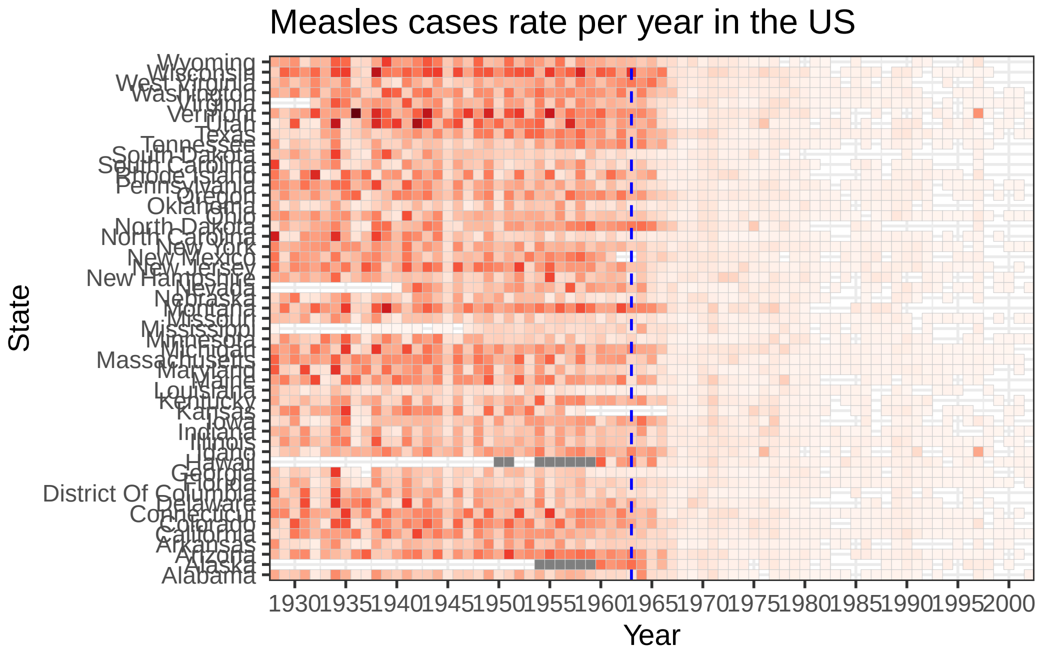
to denote missing data. Order the states based on their highest peak. You can include Hawaii and Alaska.

```
## use this color pallete
reds <- RColorBrewer::brewer.pal(9, "Reds")

#Data organized by state (higher rates to lower rates)
avg <-
  us_contagious_diseases |>
  filter(disease == "Measles" & weeks_reporting > 0) |>
  group_by(year, state) |>
  summarize(rate = sum(count/weeks_reporting*52, na.rm = TRUE)/sum(population)*10000)
```

```
`summarise()` has grouped output by 'year'. You can override using the
`.groups` argument.
```

```
avg |>
  ggplot(aes(year, state, fill=sqrt(rate))) +
  geom_tile(color= "grey80")+
  geom_vline(xintercept = 1963, linetype = "dashed", color = "blue") +
  scale_fill_gradientn(colors=reds)+
  labs(x="Year", y="State", title="Measles cases rate per year in the US") +
  theme_bw() +
  theme(legend.position = "none") +
  scale_x_continuous(breaks = seq(1930, 2002, by = 5),expand=c(0,0))
```



5. Incorporate one or more of the figures you just created to write a 2-3 page report, using quarto, describing the evidence these data show about vaccines in controlling disease. Upload your report and code to a GitHub repository.