Homework 2

Jenny Wang (71401898)

Due October 18, 2019 by 11:59pm

Vaccines have helped save millions of lives. In the 19th century, before herd immunization was achieved through vaccination programs, deaths from infectious diseases, like smallpox and polio, were common. However, today, despite all the scientific evidence for their importance, vaccination programs have become somewhat controversial.

The controversy started with a paper published in 1988 and lead by Andrew Wakefield claiming there was a link between the administration of the measles, mumps and rubella (MMR) vaccine, and the appearance of autism and bowel disease. Despite much science contradicting this finding, sensationalists media reports and fear mongering from conspiracy theorists, led parts of the public to believe that vaccines were harmful. Some parents stopped vaccinating their children. This dangerous practice can be potentially disastrous given that the Center for Disease Control and Prevention (CDC) estimates that vaccinations will prevent more than 21 million hospitalizations and 732,000 deaths among children born in the last 20 years (see Benefits from Immunization during the Vaccines for Children Program Era — United States, 1994-2013, MMWR).

Effective communication of data is a strong antidote to misinformation and fear mongering. In this homework you are going to prepare a report to have ready in case you need to help a family member, friend or acquaintance that is not aware of the positive impact vaccines have had for public health.

The data used for these plots were collected, organized and distributed by the Tycho Project. They include weekly reported counts data for seven diseases from 1928 to 2011, from all fifty states. We include the yearly totals in the `dslabs` package:

```
library(dslabs)
data(us_contagious_diseases)
```

Question 1

Use the `us_contagious_disease` and `dplyr` tools to create an object called `dat` that stores only the Measles data, includes a per 100,000 people rate, and removes Alaska and Hawaii since they only became states in the late 1950s. Note that there is a `weeks_reporting` column. Take that into account when computing the rate.

```
library(dplyr)
```

```
dat <- us_contagious_diseases %>%
       filter(!(state %in% c("Alaska","Hawaii")) & disease=="Measles") %>%
       mutate(rate=(count/weeks_reporting)* 52 / (population/100000)) %>%
       mutate(state=reorder(state, rate))

head(dat)
```

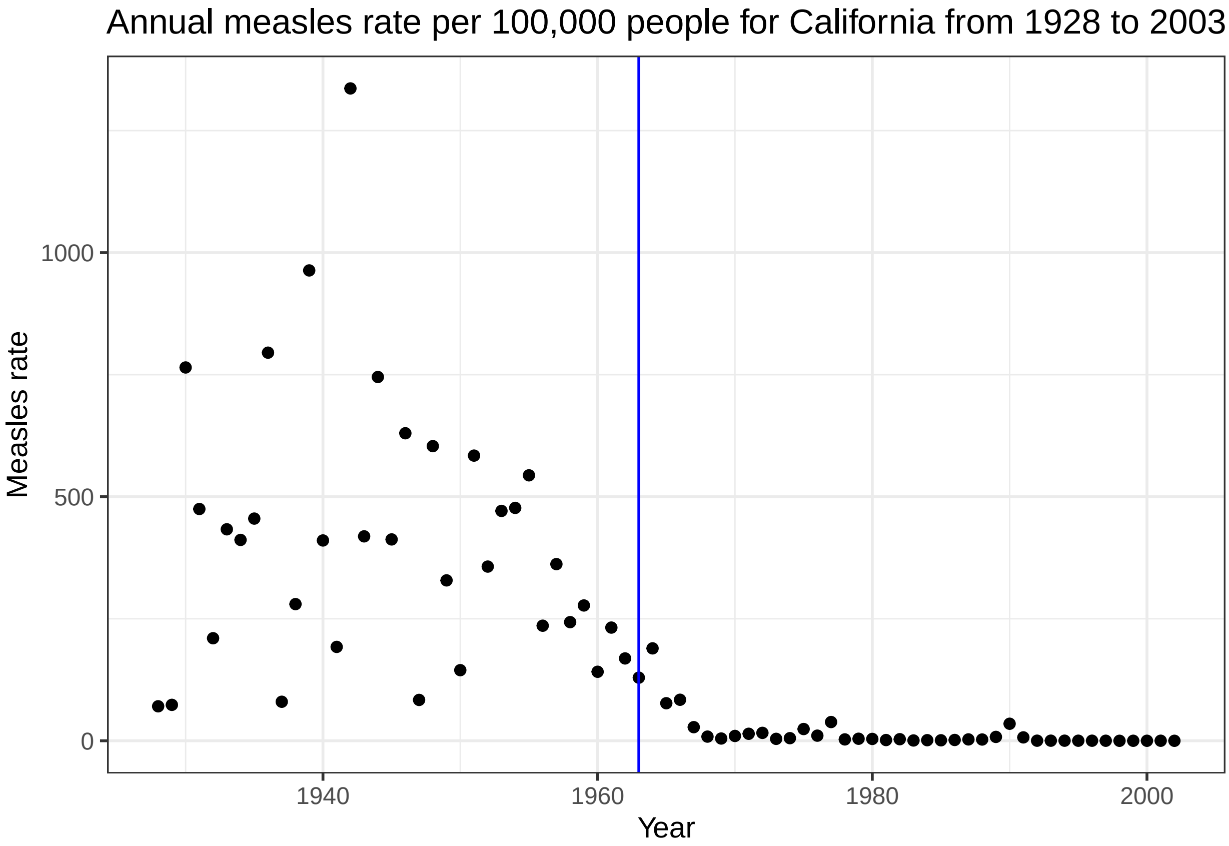
```
##   disease   state year weeks_reporting count population      rate
## 1 Measles Alabama 1928              52  8843    2589923 341.43872
## 2 Measles Alabama 1929              49  2959    2619131 119.89333
## 3 Measles Alabama 1930              52  4156    2646248 157.05255
## 4 Measles Alabama 1931              49  8934    2670818 354.98411
## 5 Measles Alabama 1932              41   270    2693027  12.71577
## 6 Measles Alabama 1933              51  1735    2713243  65.19945
```

Question 2

Plot the Measles disease rate per year for California. Find out when the Measles vaccine was introduced and add a vertical line to the plot to show this year. Note: you should be using `ggplot2` for all plotting.

```
library(ggplot2)
```

```
dat %>%
  filter(state=="California" & !is.na(rate)) %>%
  ggplot() +
  geom_point(aes(x=year, y=rate)) +
  ggtitle("Annual measles rate per 100,000 people for California from 1928 to 2003") +
  xlab("Year") + ylab("Measles rate") +
  geom_vline(xintercept=1963, col="blue") +
  theme_bw() +
  theme(plot.title=element_text(hjust=0.5))
```

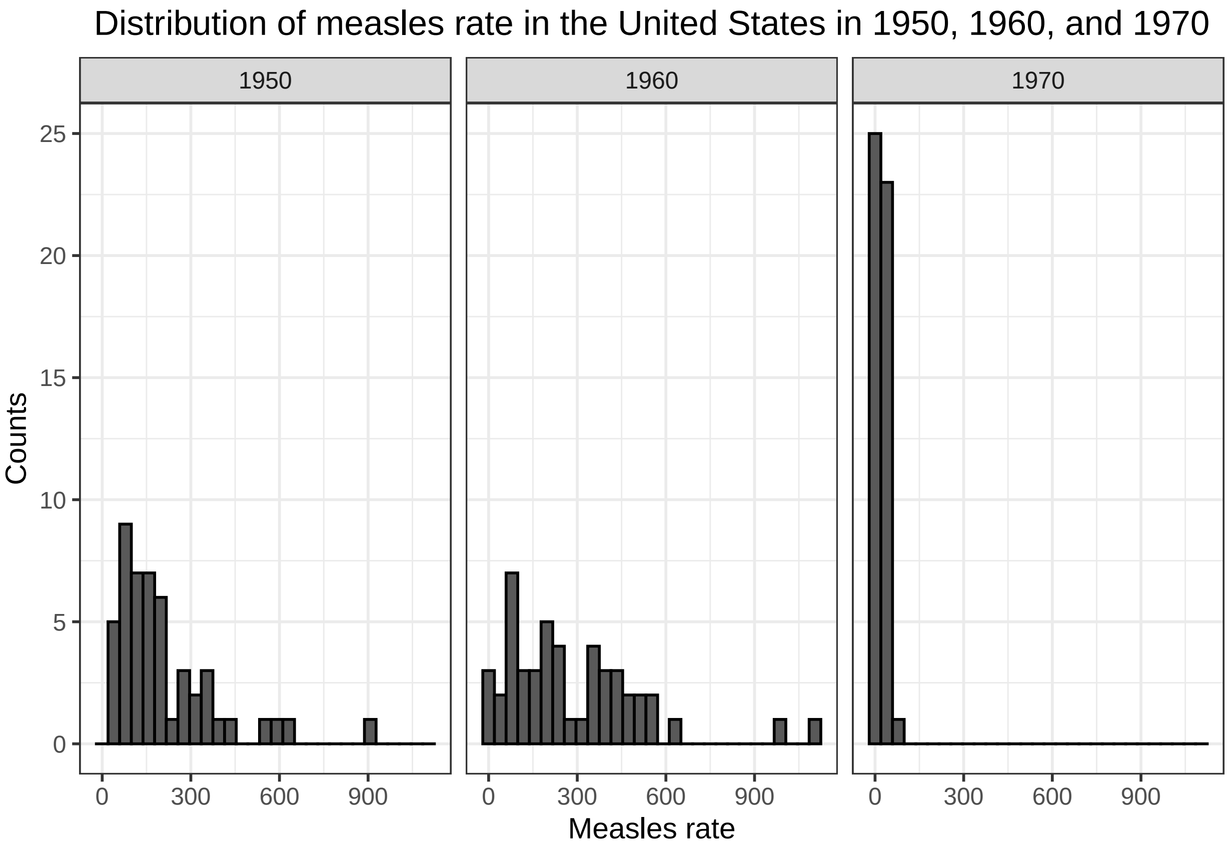


From the plot, it is very clear that the measles rate was on a rapid and consistent decline once the measles vaccine was introduced in 1963. However, it is difficult to visualize how different the individual rates are from the 1980s to the 2000s, as they are all very close to zero. Data transformation can help (see square root transformation in Question 4).

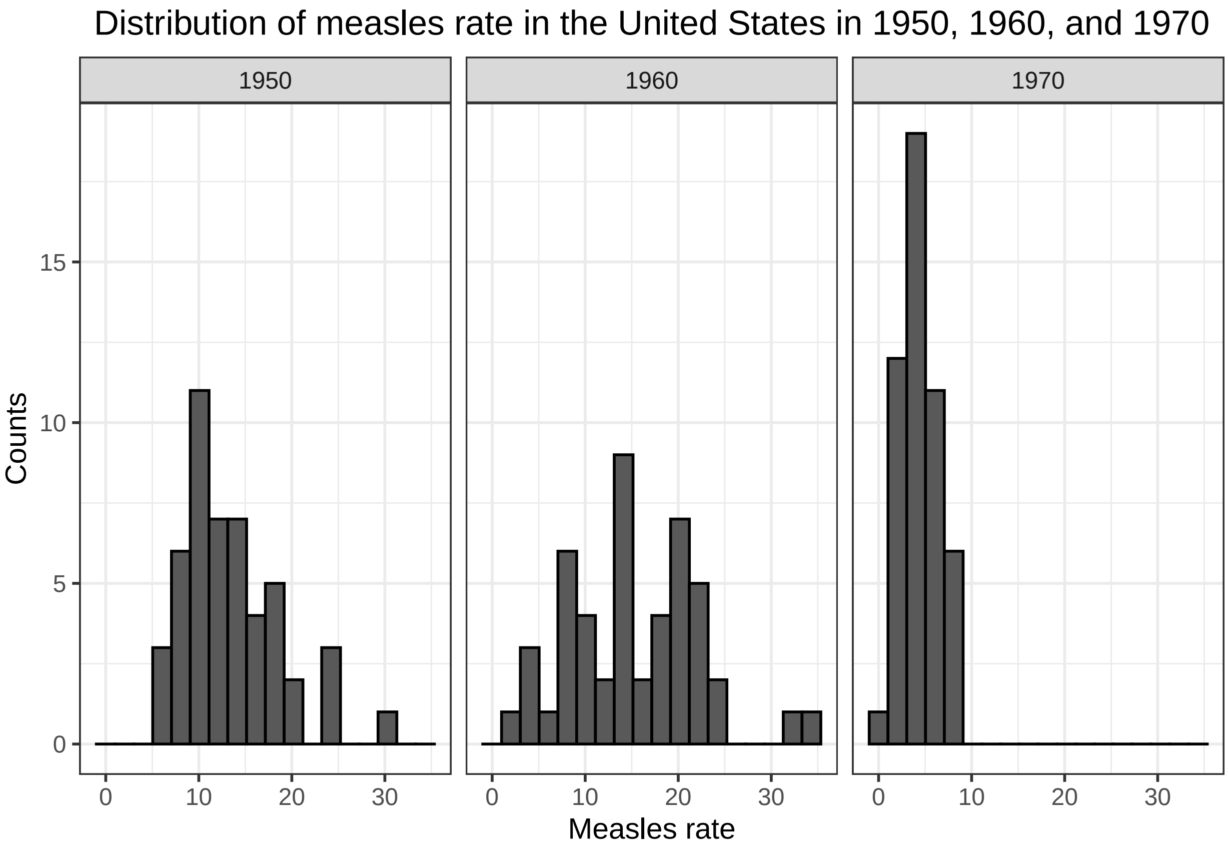
Question 3

Note these rates start off as counts. For larger counts we can expect more variability. There are statistical explanations for this which we don't discuss here, but transforming the data might help stabilize the variability such that it is closer across levels. For 1950, 1960, and 1970, plot the histogram of the data across states with and without the square root transformation. Which seems to have more similar variability across years? Make sure to pick binwidths that result in informative plots.

```
bw <- 2*IQR(dat$rate, na.rm=TRUE) / length(dat$rate)^(1/3) # Set binwidth
dat %>%
  filter(year %in% c(1950, 1960, 1970)) %>%
  ggplot(aes(rate)) +
  facet_grid(. ~ year) +
  geom_histogram(binwidth=bw, color="black") +
  ggtitle("Distribution of measles rate in the United States in 1950, 1960, and 1970") +
  xlab("Measles rate") + ylab("Counts") +
  theme_bw() +
  theme(plot.title=element_text(hjust=0.5))
```



```
bw <- 2*IQR(sqrt(dat$rate), na.rm=TRUE) / length(dat$rate)^(1/3)
dat %>%
  filter(year %in% c(1950, 1960, 1970)) %>%
  ggplot(aes(sqrt(rate))) +
  facet_grid(. ~ year) +
  geom_histogram(binwidth=bw, color="black") +
  ggtitle("Distribution of measles rate in the United States in 1950, 1960, and 1970") +
  xlab("Measles rate") + ylab("Counts") +
  theme_bw() +
  theme(plot.title=element_text(hjust=0.5))
```

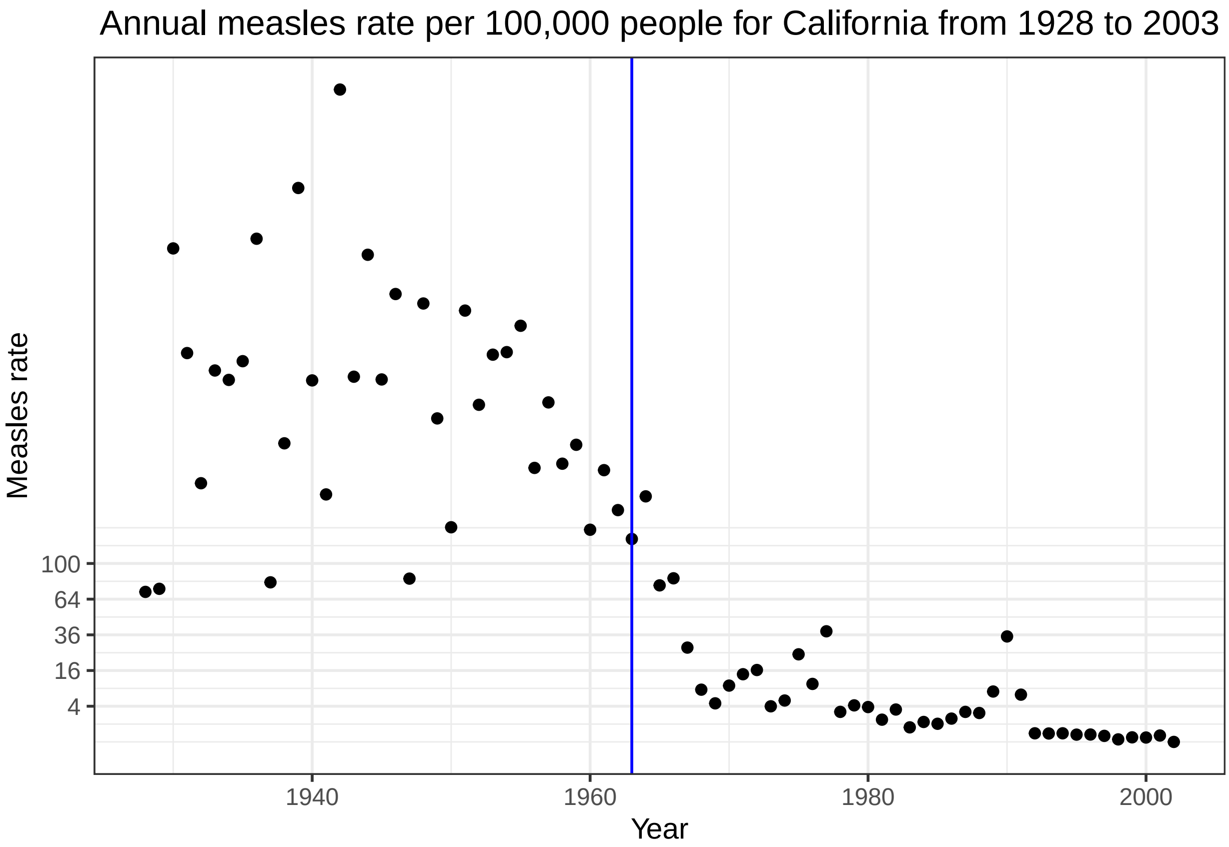


The square root transformed counts seem to have more similar variability across years. From the plots, the square root transformation works well on reducing the variability of the data, as well as reducing the right skewness. The values for measles rate are relatively small, so the square root transformation works fairly well. However, one may also consider logarithmic transformations, which could have a greater effect on the distribution. One advantage of the square root transformation is that it can be applied to zero values, whereas in a logarithmic transformation, `log(0)` is undefined. From the plots (both transformed and untransformed), we can see that the measles rate declined in 1970 compared to 1960 and before. This is of course associated with the introduction of the measles vaccine.

Question 4

Plot the Measles disease rate per year for California. Use the square root transformation. Make sure that the numbers $4, 16, 36, \ldots, 100$ appear on the y-axis. Find out when the Measles vaccine was introduced and add a vertical line to the plot to show this year.

```
dat %>%
  filter(state=="California" & !is.na(rate)) %>%
  ggplot() +
  geom_point(aes(x=year, y=rate)) +
  scale_y_continuous(trans="sqrt", breaks=c(4, 16, 36, 64, 100)) +
  ggtitle("Annual measles rate per 100,000 people for California from 1928 to 2003") +
  xlab("Year") + ylab("Measles rate") +
  geom_vline(xintercept=1963, col="blue") +
  theme_bw() +
  theme(plot.title=element_text(hjust=0.5))
```

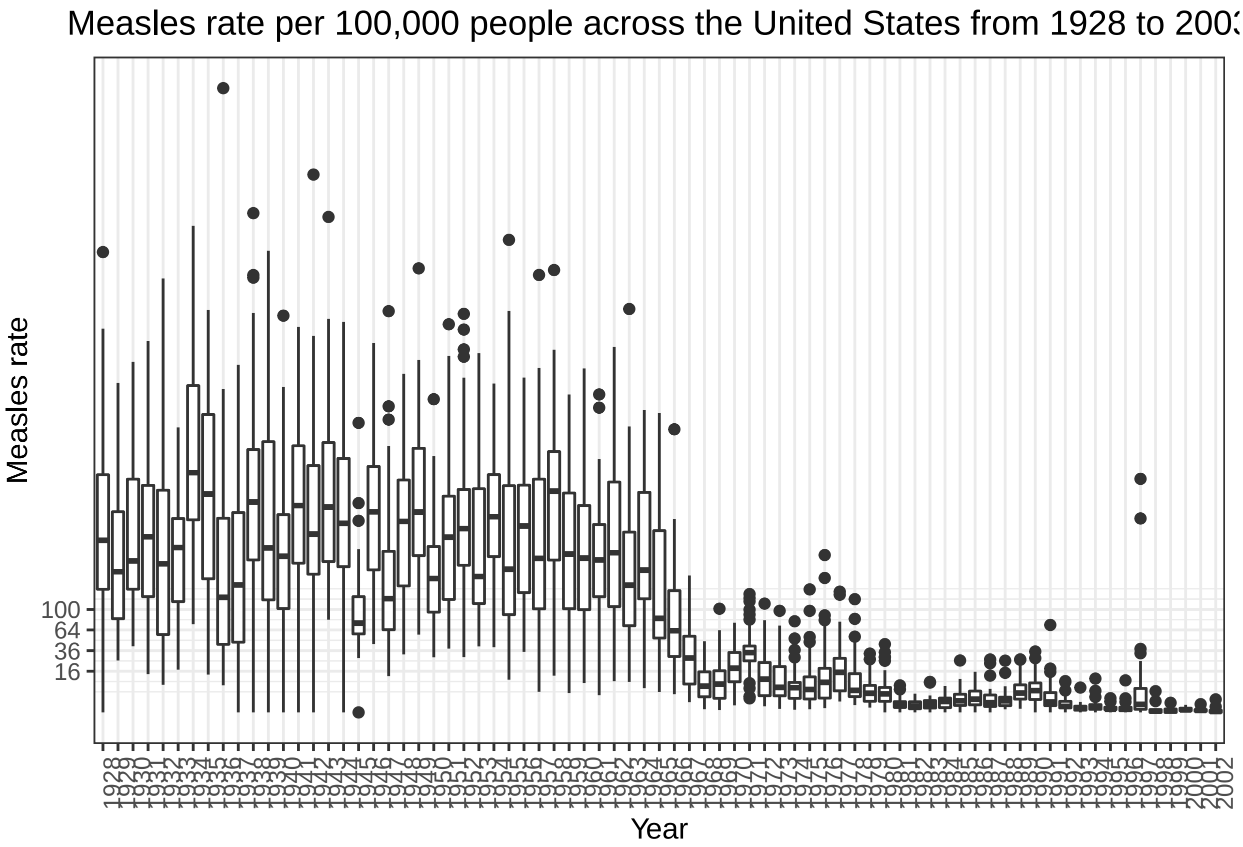


From the plot, it is very clear that the measles rate was on a rapid and consistent decline once the measles vaccine was introduced in 1963. The square root transformation makes it easier to visualize the individual rates from the 1980s to the 2000s.

Question 5

Now, this is just California. Does the pattern hold for other states? Use boxplots to get an idea of the distribution of rates for each year, and see if the pattern holds across states.

```r
# Find the national mean measles rate before and after the vaccine was introduced
dat %>%
  filter(!is.na(rate)) %>%
  ggplot(aes(x=as.factor(year), y=rate)) +
  geom_boxplot() +
  scale_y_continuous(trans="sqrt", breaks=c(4, 16, 36, 64, 100)) +
  ggtitle("Measles rate per 100,000 people across the United States from 1928 to 2003") +
  xlab("Year") + ylab("Measles rate") +
  geom_vline(xintercept=1963, color="blue") +
  theme_bw() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

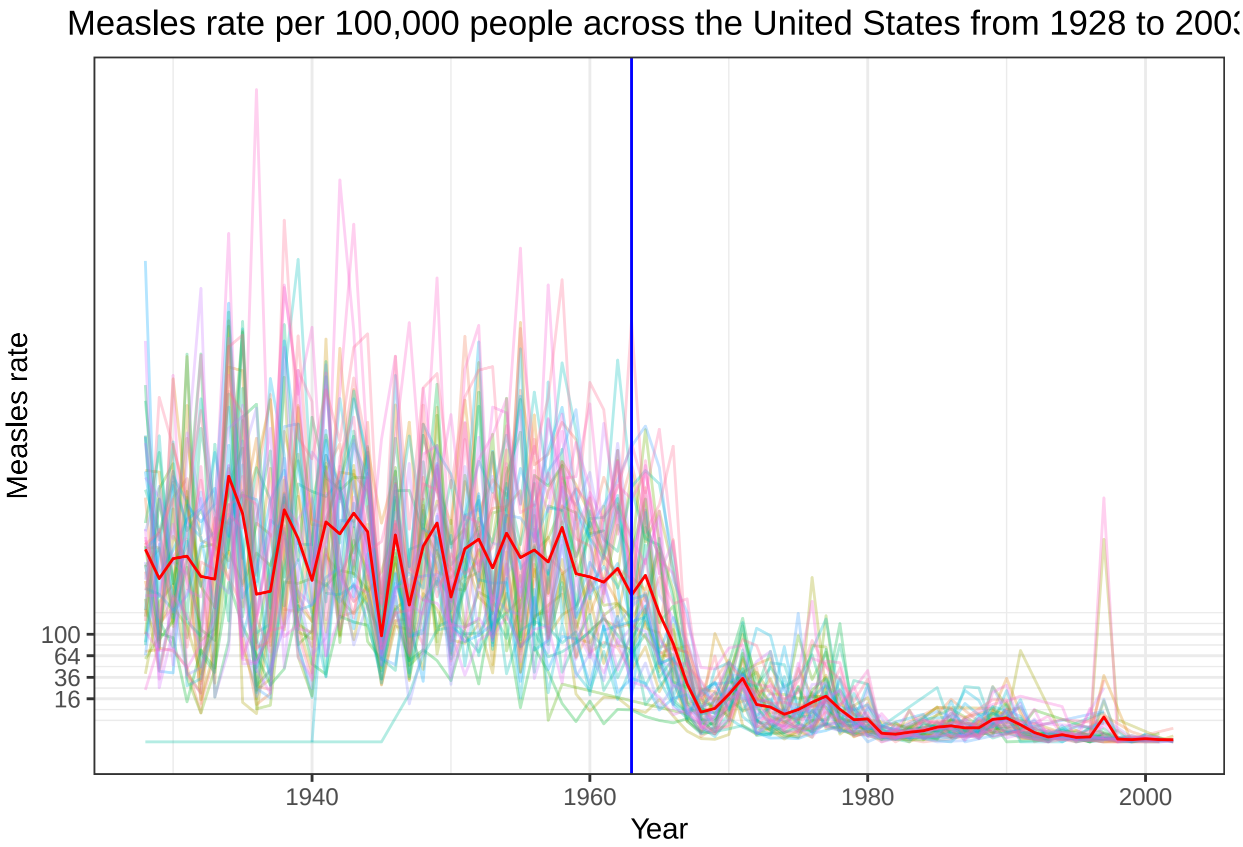


The pattern in measles rate holds for the other states throughout the years: starting with higher rates before 1963 and lower rates after. The decline in the prevalence from the boxplot is also markedly sharp shortly after 1963, indicating the effectiveness of the vaccine.

Question 6

One problem with the boxplot is that it does not let us see state-specific trends. Make a plot showing the trends for all states. Add the US average to the plot. Hint: Note there are missing values in the data.

```r
# For trends, use a line plot
# Increase the transparency for easier visualization
dat %>%
  filter(!is.na(rate)) %>%
  ggplot(aes(x=year, y=rate, color=state)) +
  geom_line(alpha=0.3) +
  scale_y_continuous(trans="sqrt", breaks=c(4, 16, 36, 64, 100)) +
  ggtitle("Measles rate per 100,000 people across the United States from 1928 to 2003") +
  xlab("Year") + ylab("Measles rate") +
  stat_summary(fun.y=mean, geom="line", color="red") + # US average
  geom_vline(xintercept=1963, color="blue") +
  theme_bw() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(legend.position="none")
```



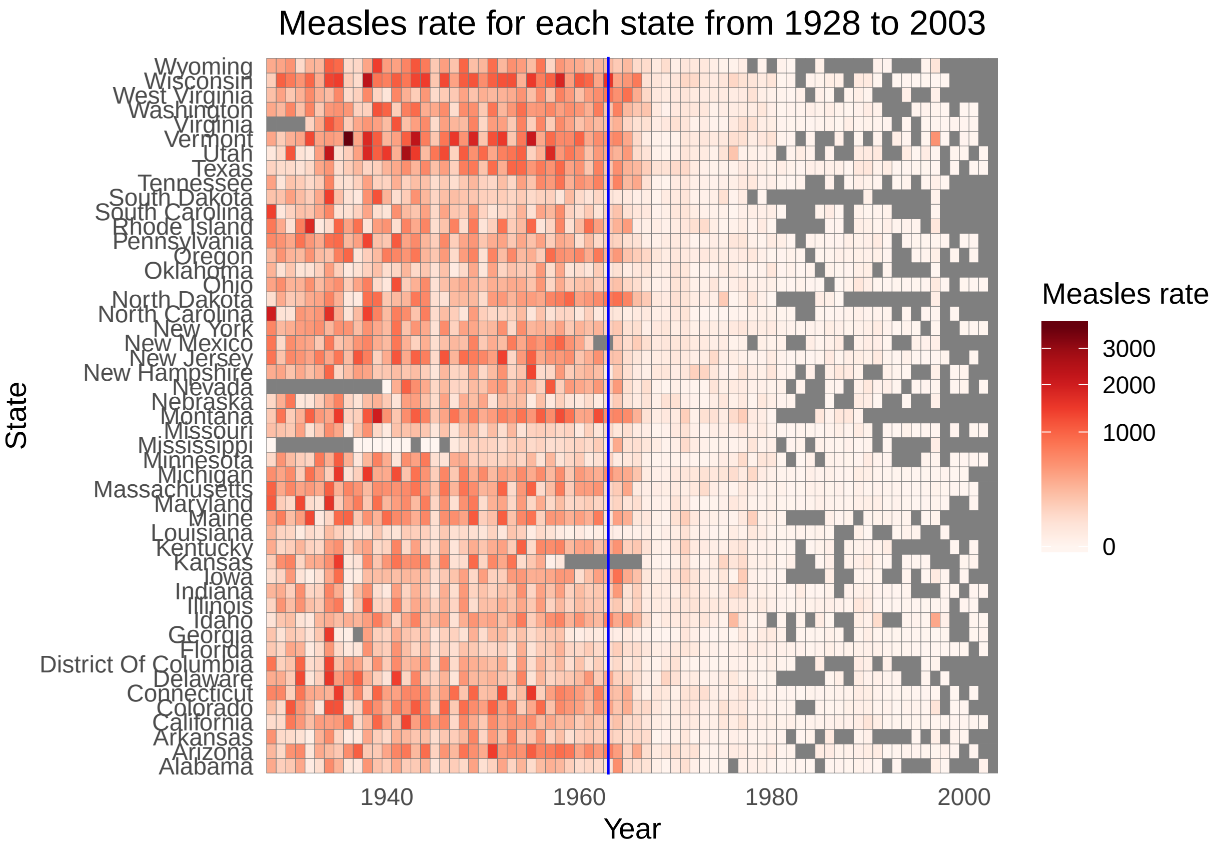
Although it is difficult to see the trend of each state in this plot, it is clear that the measles rate experienced a sharp decline shortly following the introduction of the vaccine. The US average measles rate throughout the years serves as a fairly accurate representation of the trends in the different states, as the number of states that had a measles rate above the average is approximately the same as those below.

Question 7

One problem with the plot above is that we can't distinguish states from each other. There are just too many. We have three variables to show: year, state and rate. If we use the two dimensions to show year and state then we need something other than vertical or horizontal position to show the rates. Try using color. Hint: Use the the geometry `geom_tile` to tile the plot with colors representing disease rates.

```
library(RColorBrewer)
```

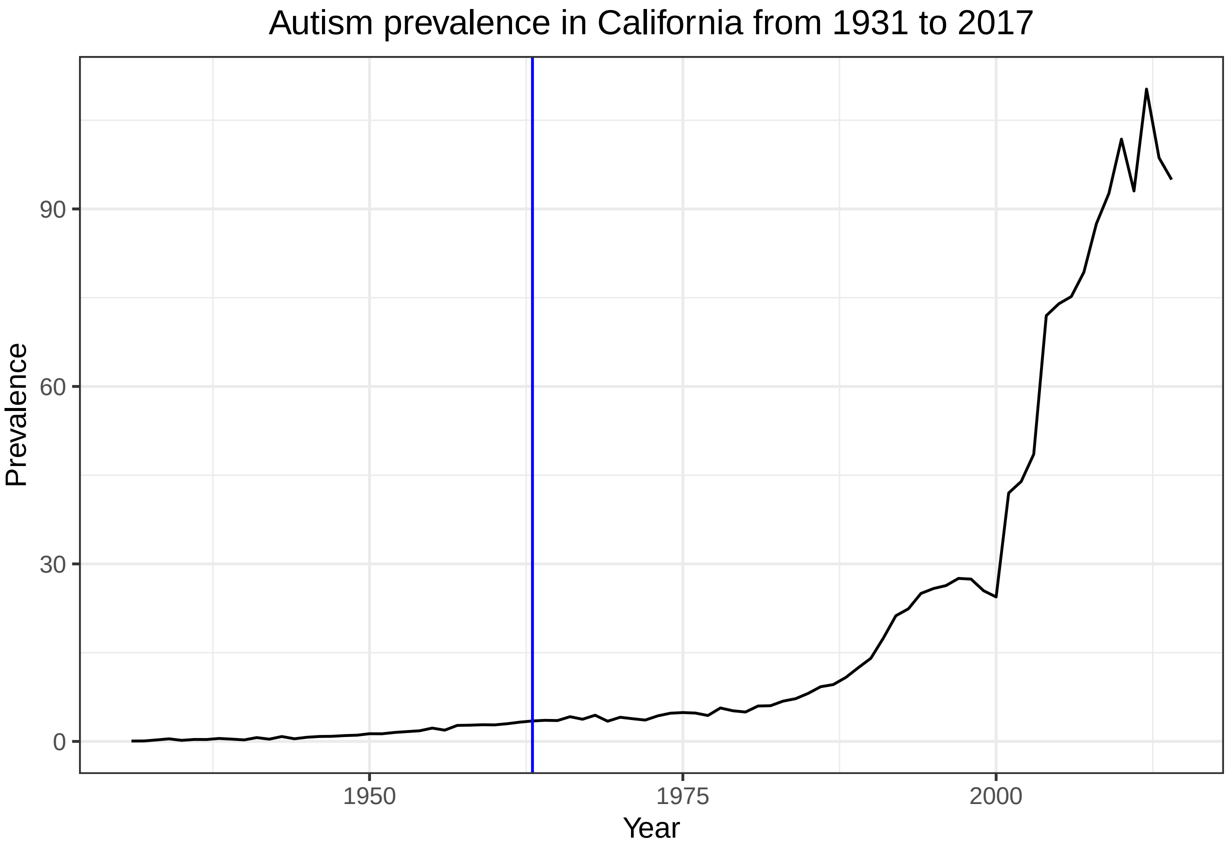
```
dat %>%
  ggplot(aes(year, state,  fill = rate)) +
  geom_tile(color="grey50") +
  scale_x_continuous(expand=c(0,0)) +
  scale_fill_gradientn(colors=brewer.pal(9, "Reds"), trans="sqrt") +
  geom_vline(xintercept=1963, col="blue") +
  theme_minimal() +
  theme(panel.grid=element_blank()) +
  ggtitle("Measles rate for each state from 1928 to 2003") +
  xlab("Year") + ylab("State") +
  theme(plot.title=element_text(hjust=0.5)) +
  labs(fill="Measles rate")
```



8. The plots above provide strong evidence showing the benefits of vaccines: as vaccines were introduced, disease rates were reduced. But did autism increase? Find yearly reported autism rates data and provide a plot that shows if it has increased and if the increase coincides with the introduction of vaccines.

```
library(reshape2)
```

```
autism_ca <- read.csv("autism_ca.csv",
                      header=TRUE, stringsAsFactors=FALSE)
prev_avail <- "^X[0-9]{4}\\.1$"
prev <- grep(prev_avail, names(autism_ca)) # Match and find columns with prevalence data

autism_ca <- autism_ca %>%
  select(X, prev)
names(autism_ca) <- c("birth_year", seq(1997, 2006), 2014, 2016, 2017)

autism_ca <- apply(autism_ca, 2, as.numeric) %>%
  as.data.frame() # Ignore non-numeric values and convert back to a data frame

autism_ca <- autism_ca %>%
  mutate(prevalence=apply(autism_ca[, -1], 1, mean, na.rm=TRUE))
# Mean prevalence of each birth year across reported years

# Now we plot the trends
autism_ca %>%
  filter(!is.na(prevalence)) %>%
  ggplot(aes(x=birth_year, y=prevalence)) +
  geom_line() +
  ggtitle("Autism prevalence in California from 1931 to 2017") +
  xlab("Year") + ylab("Prevalence") +
  geom_vline(xintercept=1963, color="blue") +
  theme_bw() +
  theme(plot.title=element_text(hjust=0.5))
```



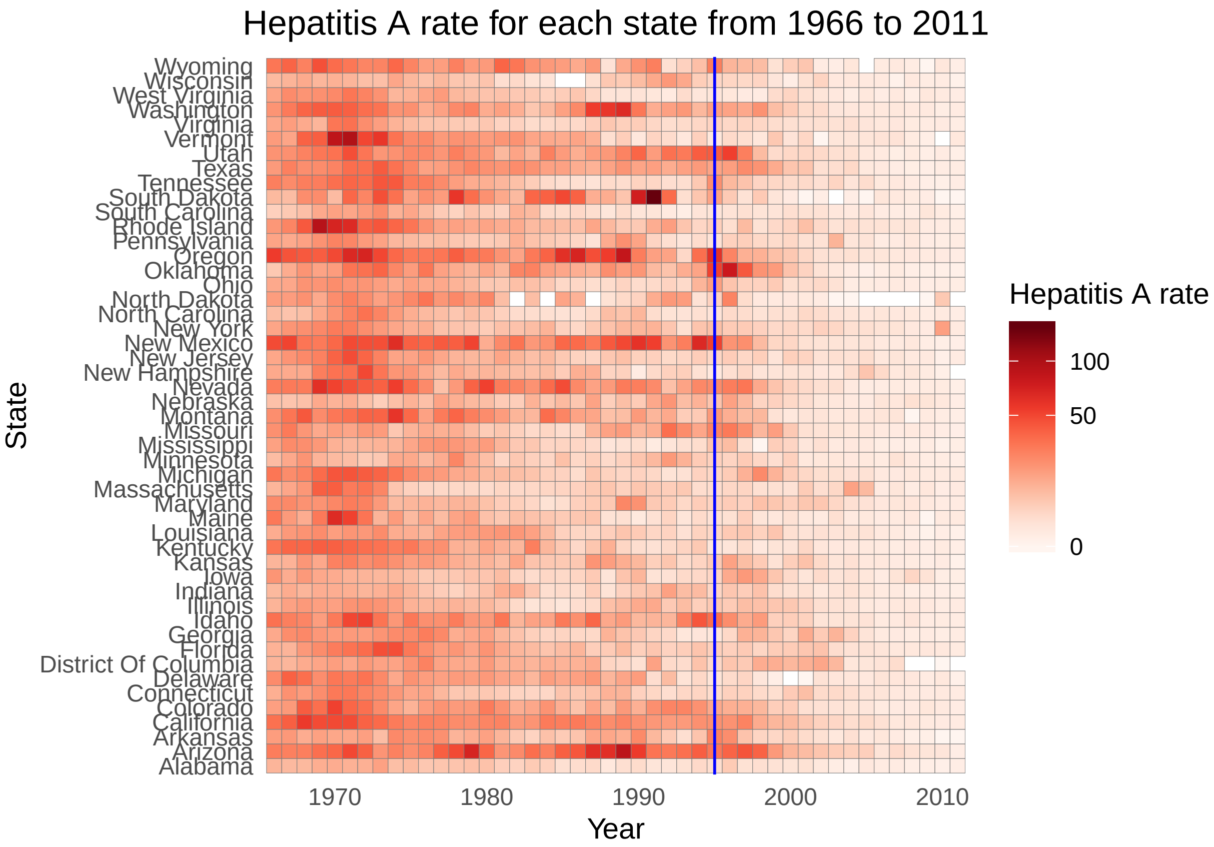
The plot represents the prevalence of autism collected on children born in 1931 to 2017 in California, as reported by Nevison *et al.* (2018). The raw data used come from Supplementary File S1 from the original paper. For each birth year, the average of the autism rate was calculated for each reported year.

From the plot, we can see that the autism prevalence in California experienced an increase during the time period between 1931 and 2017, with particularly sharp rates starting around 1975 and continuing throughout the 2000s. The measles vaccine was introduced in 1963, and it is tempting to conclude that this - in addition to other vaccination programs - was directly associated, if not caused the rise in autism prevalence. However, if we consult additional sources around autism, we can see that there has been changes to the definition of autism over the years. From Evans (2013): "changes in diagnostic methods from the 1960s to the 1980s meant that autism came to be associated with 'profound mental retardation and other developmental or physical disorders' thereby *increasing the number of children who were considered to display autistic traits.* In other words, this broadening of definition contributed to an increase in autism prevalence beginning in the 1960s. Although this review focused on British children, the DSM definition is also applied to their American conterparts. Therefore, it is valid to make the same conclusion on US autism rates. Additionally, the smallpox vaccine was invented in the 1700s, and we do not see elevated autism rates between the start of data

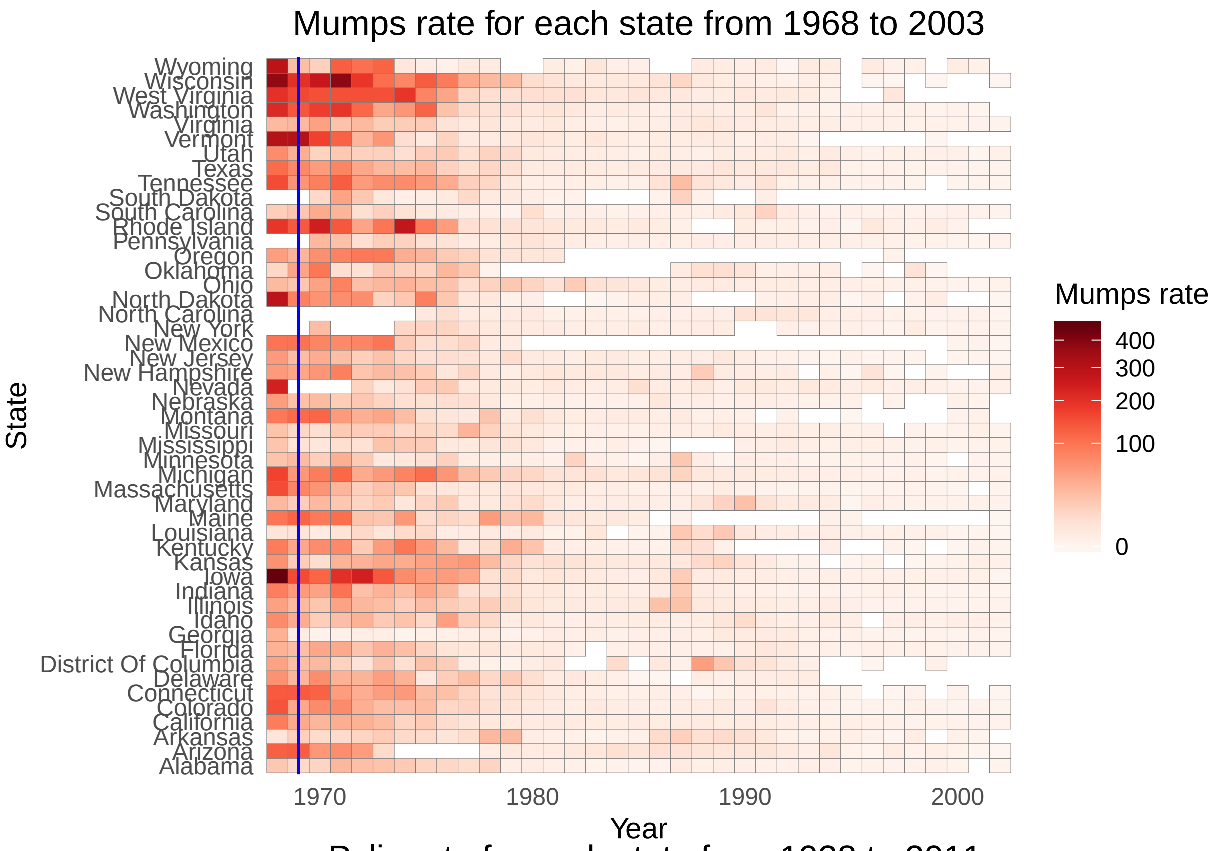
collection and the 1970s. This evidence counters the claim that vaccines have a role to play in the rise of autism. Compared to the argument that introduction of vaccines was the cause of more autism cases, the expansion of diagnosis criteria is probably a more suitable explanation.

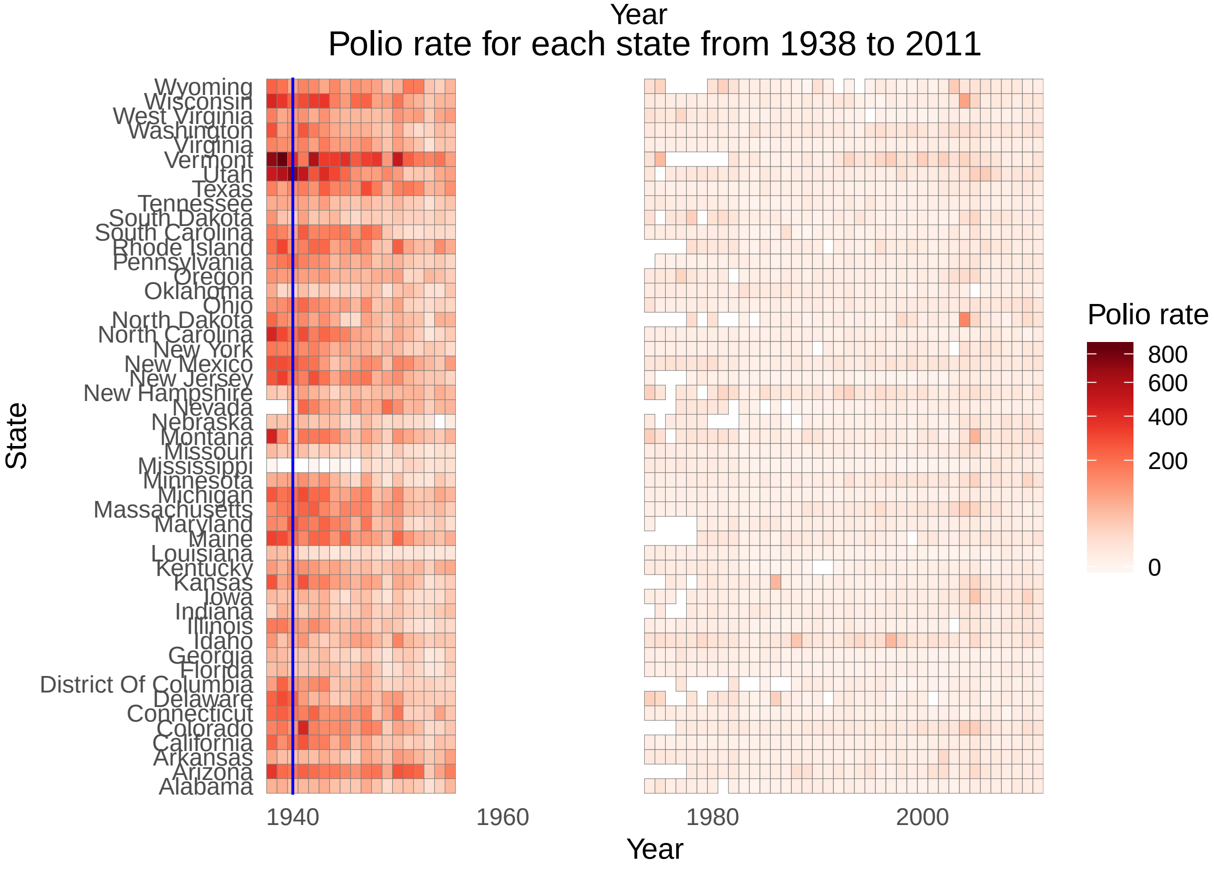
9. Use data exploration to determine if other diseases (besides Measles) have enough data to explore the effects of vaccines. Prepare a report (minimum 1 paragraph, maximum 3 paragraphs) with as many plots as you think are necessary to provide a case for the benefit of vaccines. Note that there was a data entry mistake and the data for Polio and Pertussis are exactly the same.

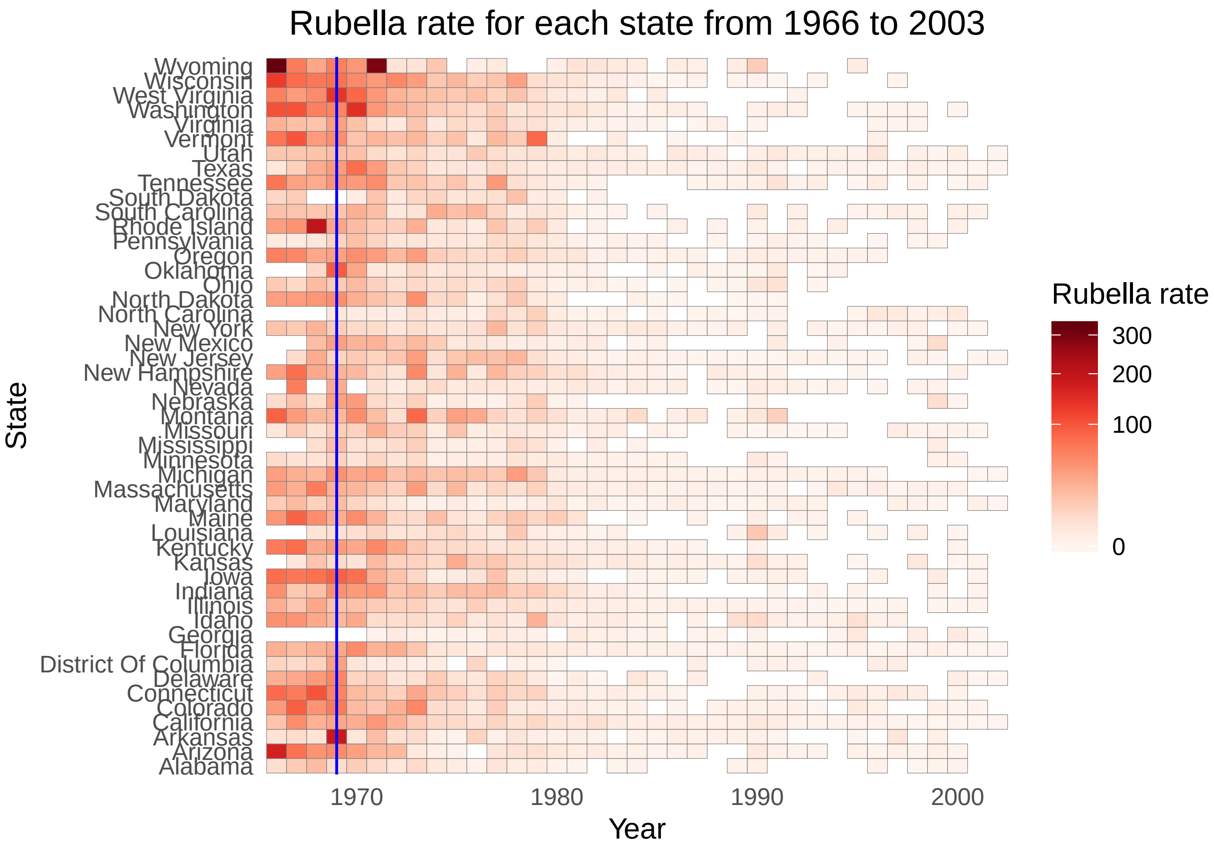
```r
all_diseases <- us_contagious_diseases %>%
  filter(!(state %in% c("Alaska","Hawaii"))) %>%
  mutate(rate=(count/weeks_reporting)* 52 / (population/100000)) %>%
  mutate(state=reorder(state, rate))

# table(all_diseases$disease)

all_diseases %>%
  filter(disease=="Hepatitis A" & !is.na(rate)) %>%
  ggplot(aes(year, state,  fill = rate)) +
  geom_tile(color="grey50") +
  scale_x_continuous(expand=c(0,0)) +
  scale_fill_gradientn(colors=brewer.pal(9, "Reds"), trans="sqrt") +
  geom_vline(xintercept=1995, col="blue") +
  theme_minimal() +
  theme(panel.grid=element_blank()) +
  ggtitle("Hepatitis A rate for each state from 1966 to 2011") +
  xlab("Year") + ylab("State") +
  theme(plot.title=element_text(hjust=0.5)) +
  labs(fill="Hepatitis A rate")
```



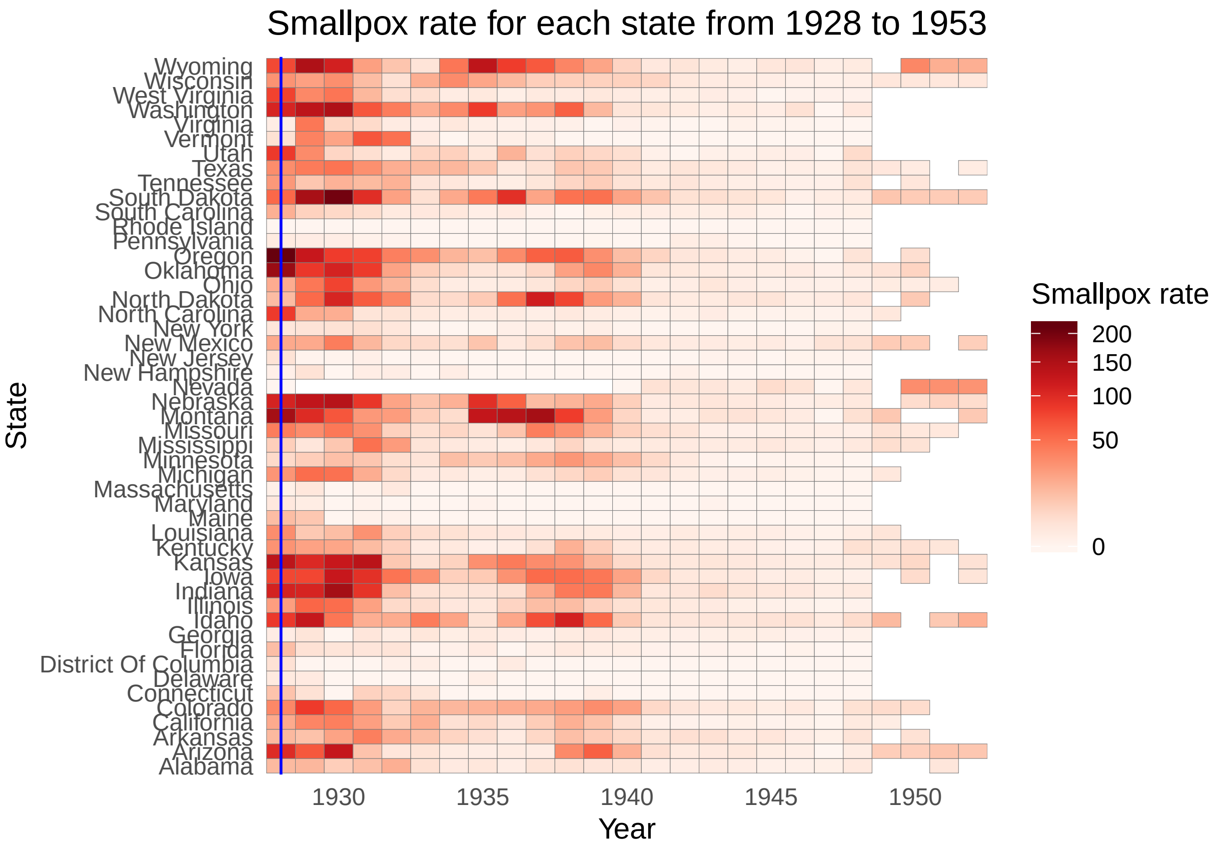
```
# Code is silenced to save space - no major changes, except "measles"/"hepatitis A" were
# replaced with the other disease names
```


Mumps rate for each state from 1968 to 2003
State
Wyoming
Wisconsin
West Virginia
Washington
Virginia
Vermont
Utah
Texas
Tennessee
South Dakota
South Carolina
Rhode Island
Pennsylvania
Oregon
Oklahoma
Ohio
North Dakota
North Carolina
New York
New Mexico
New Jersey
New Hampshire
Nevada
Nebraska
Montana
Missouri
Mississippi
Minnesota
Michigan
Massachusetts
Maryland
Maine
Louisiana
Kentucky
Kansas
Iowa
Indiana
Illinois
Idaho
Georgia
Florida
District Of Columbia
Delaware
Connecticut
Colorado
California
Arkansas
Arizona
Alabama
1970
1980
1990
2000
Year
Mumps rate
400
300
200
100
0


Polio rate for each state from 1938 to 2011
State
Wyoming
Wisconsin
West Virginia
Washington
Virginia
Vermont
Utah
Texas
Tennessee
South Dakota
South Carolina
Rhode Island
Pennsylvania
Oregon
Oklahoma
Ohio
North Dakota
North Carolina
New York
New Mexico
New Jersey
New Hampshire
Nevada
Nebraska
Montana
Missouri
Mississippi
Minnesota
Michigan
Massachusetts
Maryland
Maine
Louisiana
Kentucky
Kansas
Iowa
Indiana
Illinois
Idaho
Georgia
Florida
District Of Columbia
Delaware
Connecticut
Colorado
California
Arkansas
Arizona
Alabama
1940
1960
1980
2000
Year
Polio rate
800
600
400
200
0



```r
# The smallpox vaccine was first invented by Edward Jenner in 1796. We do not have data
# dating so far back, but we can still construct the tile plot to visualize trends
# in the prevalence of smallpox, with the blue line starting at 1928 (beginning of data
# collection). Also, smallpox was eradicated in 1980, a milestone in human history

all_diseases %>%
  filter(disease=="Smallpox" & !is.na(rate)) %>%
  ggplot(aes(year, state,  fill = rate)) +
  geom_tile(color="grey50") +
  scale_x_continuous(expand=c(0,0)) +
  scale_fill_gradientn(colors=brewer.pal(9, "Reds"), trans="sqrt") +
  geom_vline(xintercept=1928, col="blue") +
  theme_minimal() +
  theme(panel.grid=element_blank()) +
  ggtitle("Smallpox rate for each state from 1928 to 1953") +
  xlab("Year") + ylab("State") +
  theme(plot.title=element_text(hjust=0.5)) +
  labs(fill="Smallpox rate")
```



This dataset contains detailed records of disease prevalence of hepatitis A, measles, mumps, pertussis/polio, rubella, and smallpox. A tile plot was constructed for each disease with the introduction year of its respective vaccine highlighted in blue. From every single one of these plots, we can clearly see a rapid and stark decrease in the disease rate after each vaccine was introduced. Children's lives have been saved numberless times from what was once deadly by quick and simple injections that have proven to be effective throughout the years. The eradication of smallpox marks a pivotal point in medicine and overall human civilization as it shows the promise vaccines hold. Additionally, according to the WHO, polio case numbers have decreased by more than 99% since efforts were implemented to motivate vaccination against the disease. This trend repeats itself for all the diseases we have so far looked at. We have demonstrated evidence that the rise in autism rates was a result of changes in the definition of the disorder rather than the introduction of vaccines. The benefit of vaccines is easily comprehended upon exploring this dataset. As such, we as public health professionals should continue to encourage every effort put into combating infectious diseases by vaccinating eligible children, as well as to invest in resources that drive the development of new, effective vaccines.