Data Science

CS301

The Vaccine Lab

Week 11

Fall 2024

Oliver BONHAM-CARTER

Are you here today?!

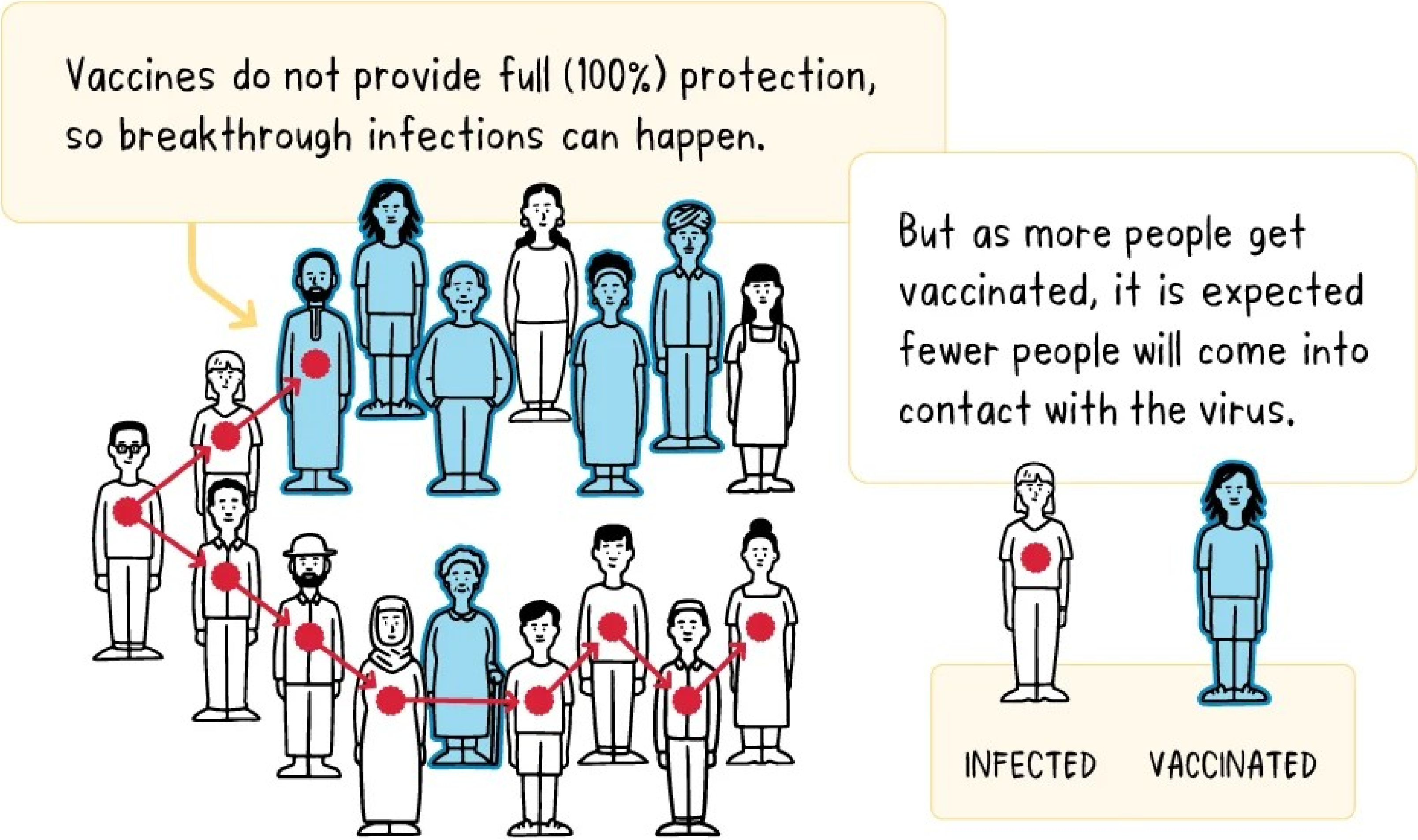
https://forms.gle/iaY7zBmxj8KvsDMa8

Let’s Talk About Lab 5 For A Moment...

- How do you know if something to prevent sickness is working?
- Are the Vaccines working?
  - Are there fewer people with Measles, mumps, Hepatitis B (and other illnesses) as a result of receiving vaccines in 1966?

- History of Vaccines: https://historyofvaccines.org/

Why Vaccines?



World Health Organization:
https://www.who.int/news-room/feature-stories/detail/vaccine-efficacy-effectiveness-and-protection

What Does Research Say?

Comparison of 20th Century Annual Morbidity & Current Morbidity

Disease	20th Century Annual Morbidity*	2010 Reported Cases†	% Decrease
Smallpox	29,005	0	100%
Diphtheria	21,053	0	100%
Pertussis	200,752	21,291	89%
Tetanus	580	8	99%
Polio (paralytic)	16,316	0	100%
Measles	530,217	61	>99%
Mumps	162,344	2,528	98%
Rubella	47,745	6	>99%
CRS	152	0	100%
Haemophilus influenzae (<5 years of age)	20,000 (est.)	270 (16 serotype b and 254 unknown serotype)	99%

Sources:

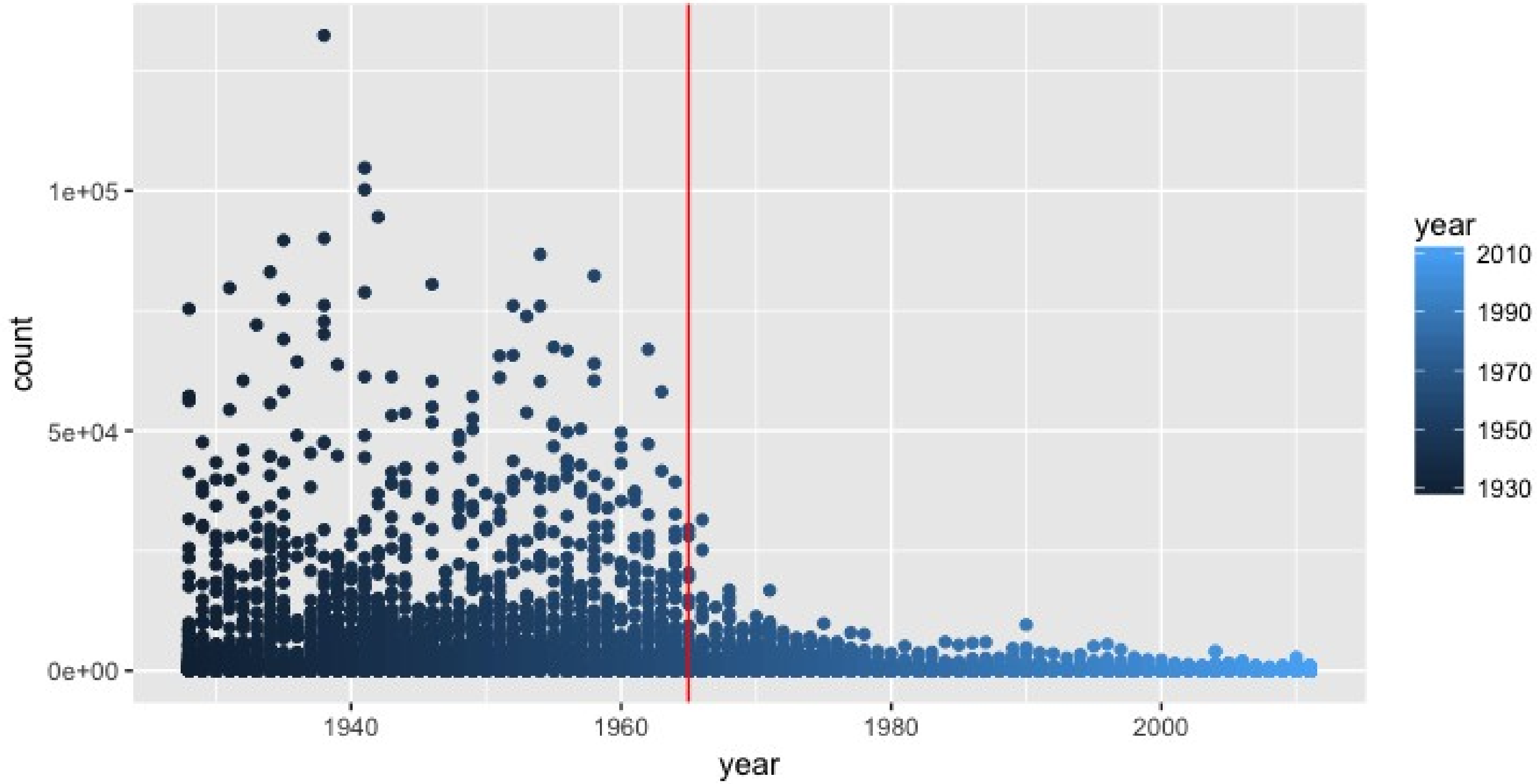
* *JAMA*. 2007;298(18):2155-2163

† CDC. *MMWR* January 7, 2011;59(52);1704-1716. (Provisional *MMWR* week 52 data)

- Vox Article: https://www.vox.com/health-care/2014/10/13/6967317/vaccines-work-this-chart-proves-it

What Does Our Data Say About (All) Vaccines of Data?

```
library(tidyverse)

library(dslabs)

library(dplyr)

ggplot(data = us_contagious_diseases) +  geom_point(mapping = aes(x = year, y = count, color = year)) + geom_vline(xintercept = 1965, color = "red")
```



Cases of Illness

Lab Results

- Use the us contagious disease and dplyr tools to create an object that **stores only the Measles data**, **includes a per 100,000 people rate**, and removes Alaska and Hawaii. **Note that there is a weeks reporting column. Take that into account when computing the rate.**

- # Add the rate column to the data:

  dat_measles_rate <- filter(us_contagious_diseases, disease == "Measles") %>% mutate(rate = (((count*52)/weeks_reporting)) / ((population)/100000))

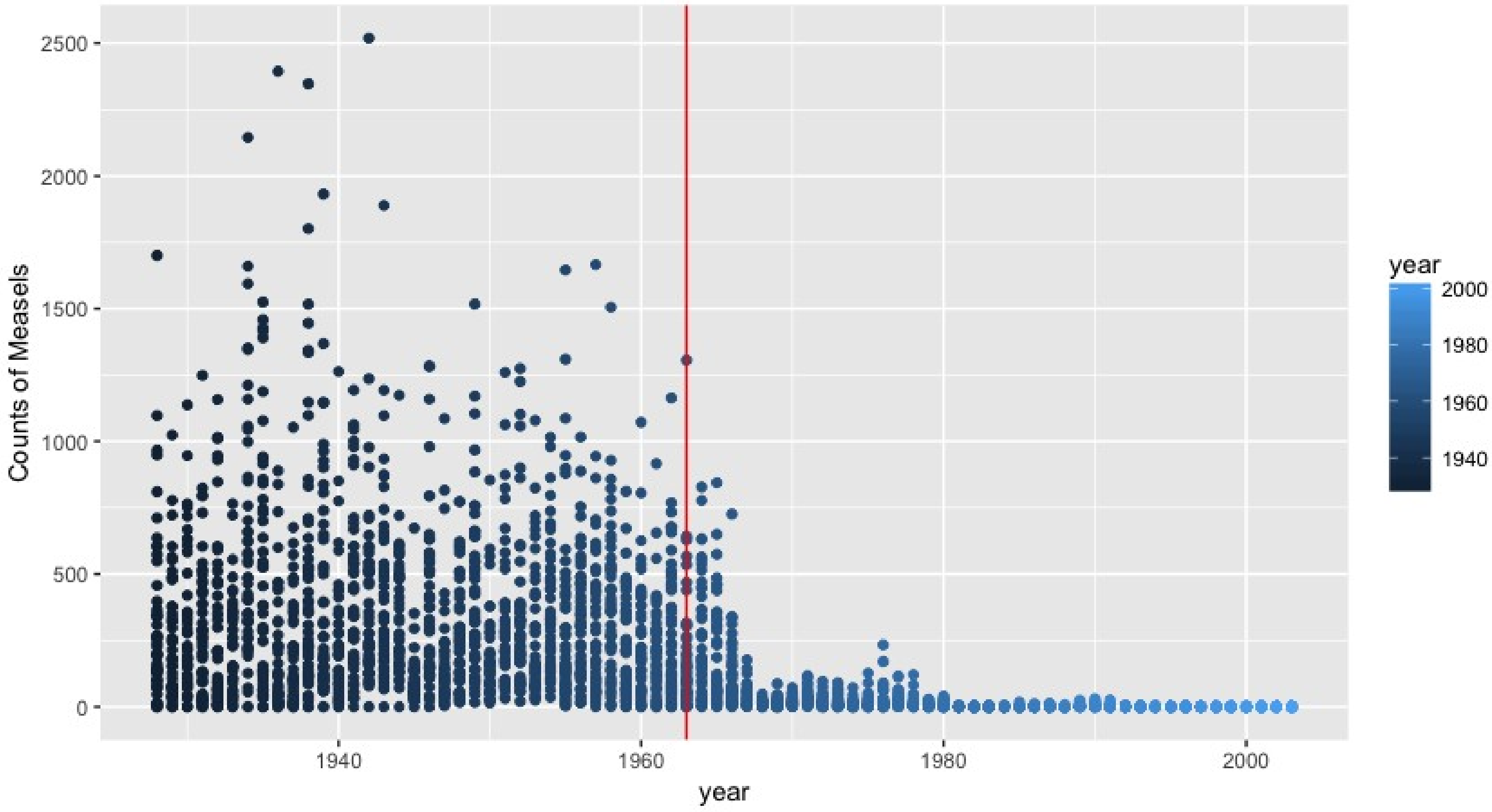
- # Note: the *rate* is one of several possible calculations...

Trim Out Two States

```
# Remove the two states (Alaska and Hawaii)

dat_measles_rate_lessTwoStates <- filter(dat_measles_rate, state != "Alaska", state != "Hawaii")


View(dat_measles_rate_lessTwoStates)


# Plot the results across 48 states

ggplot(data = dat_measles_rate_lessTwoStates, mapping = aes(x = year, y = rate, color = year)) + geom_point() + geom_vline(xintercept = 1963, color = "red") + labs(y = "Counts of Measels")
```

Plot Across 48 States

```
ggplot(data = dat_measles_rate_lessTwoStates, mapping = aes(x = year, y = rate, color = year)) + geom_point() + geom_vline(xintercept = 1963, color = "red") + labs(y = "Counts of Measels")
```

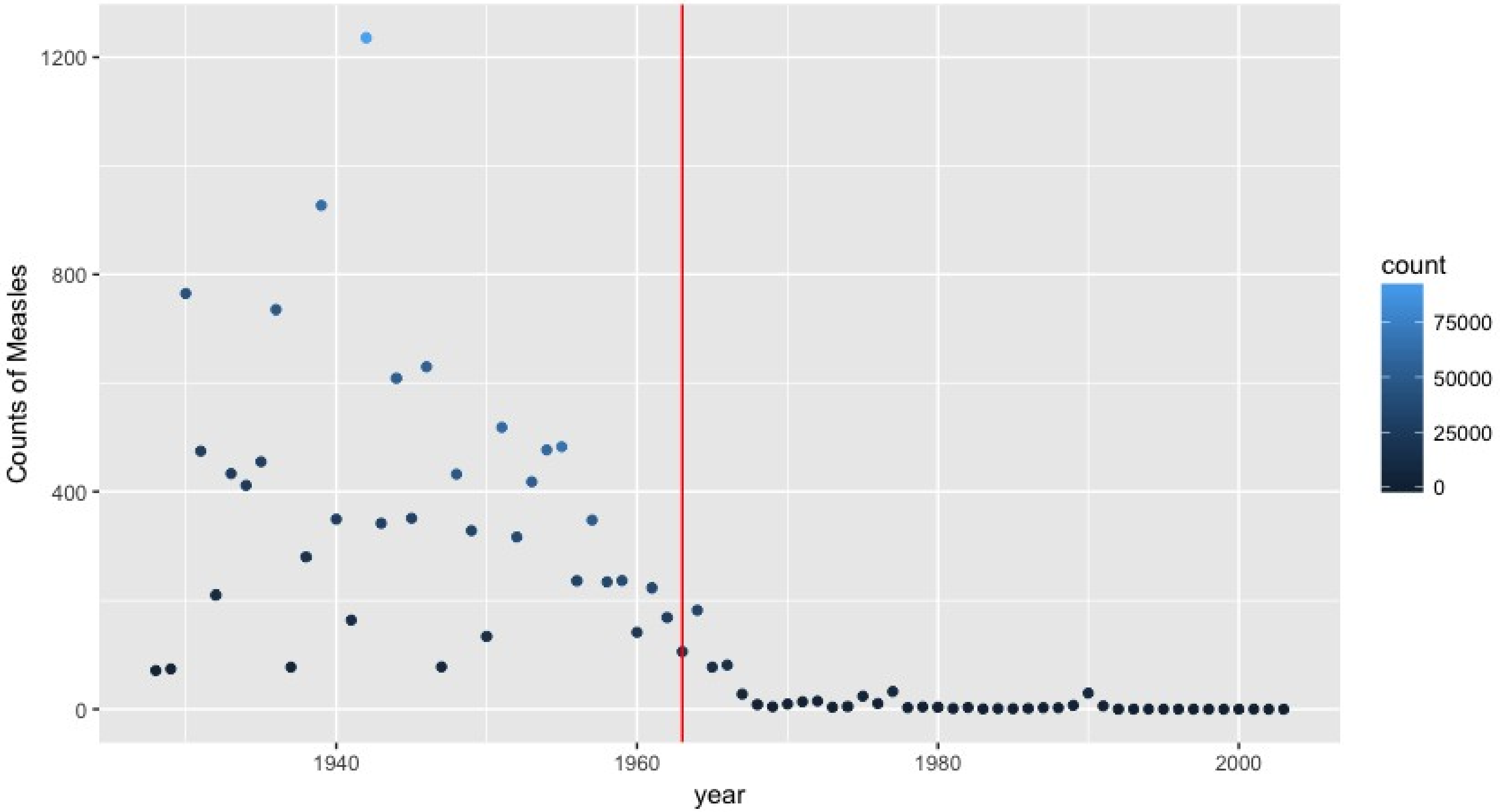


Focus On California

```
# Create table to focus on California

dat_caliFocus <- filter(dat_measles_rate_lessTwoStates, state == "California")

View(dat_caliFocus)

ggplot(data = dat_caliFocus, mapping = aes(x = year, y = rate, color = count)) + geom_point() + geom_vline(xintercept = 1963, color = "red") + labs(y = "Counts of Measles")
```

Data From California, Only

```
ggplot(data = dat_caliFocus, mapping = aes(x = year, y = rate, color = count)) + geom_point() + geom_vline(xintercept = 1963, color = "red") + labs(y = "Counts of Measles")
```



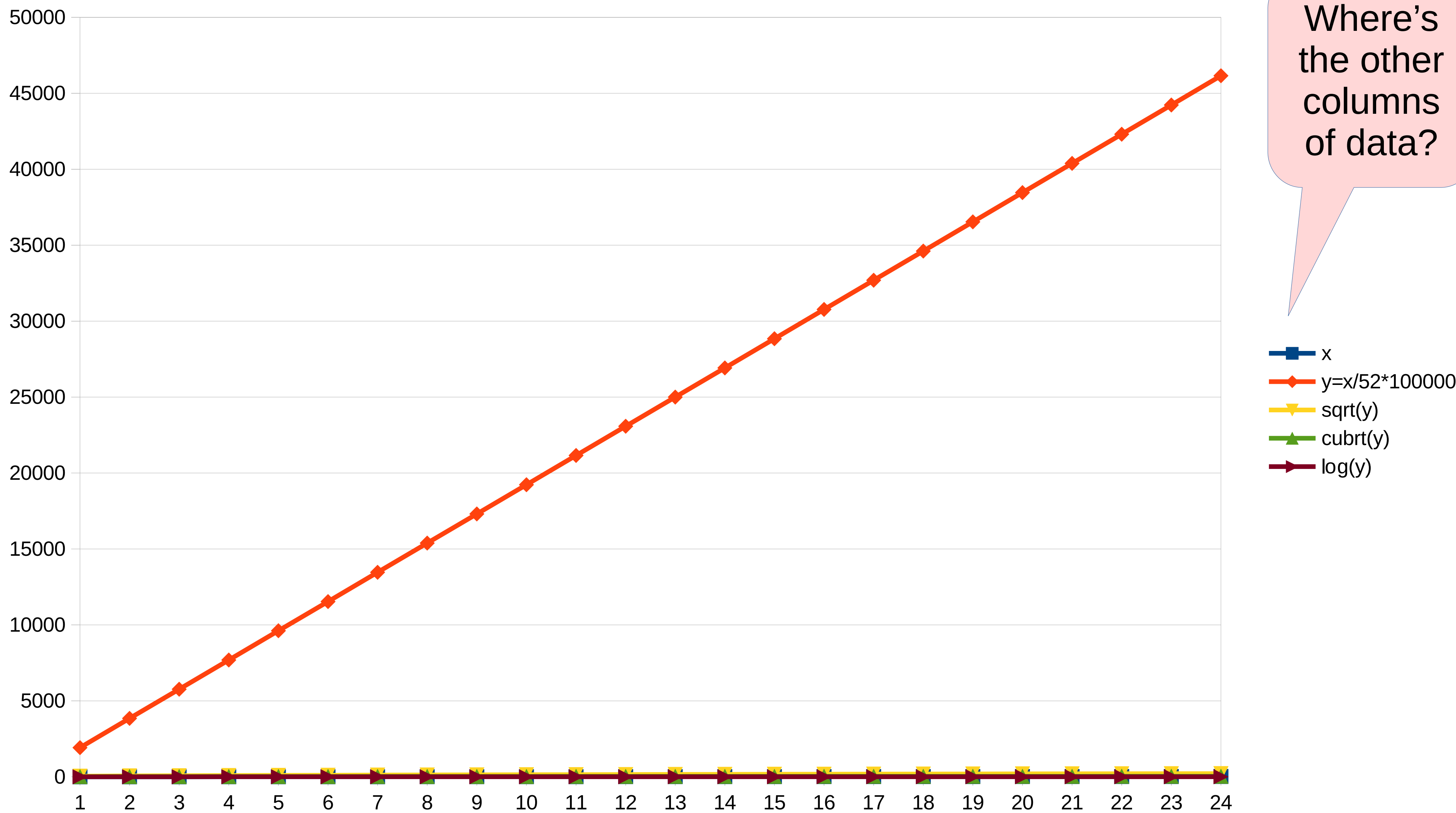
Discussion Points

- *Vaccines administered in 1964 cause a decline in measles cases.*
- *California was used as a sample study: less data but still a good representation of the rest of the dataset*
- *The decades following the vaccines of 1964 show that new threats appear. Vaccines must be updated.*

Transformations Helped Visualize

- *Allows us to visually compare data by changing shape*
- The square root, x to x^(1/2) = sqrt(x), is a transformation with a moderate effect on distribution shape.
- Weaker than the logarithm and the cube root transformations
- Used for reducing right skewness
- Has the advantage that it can be applied to zero values

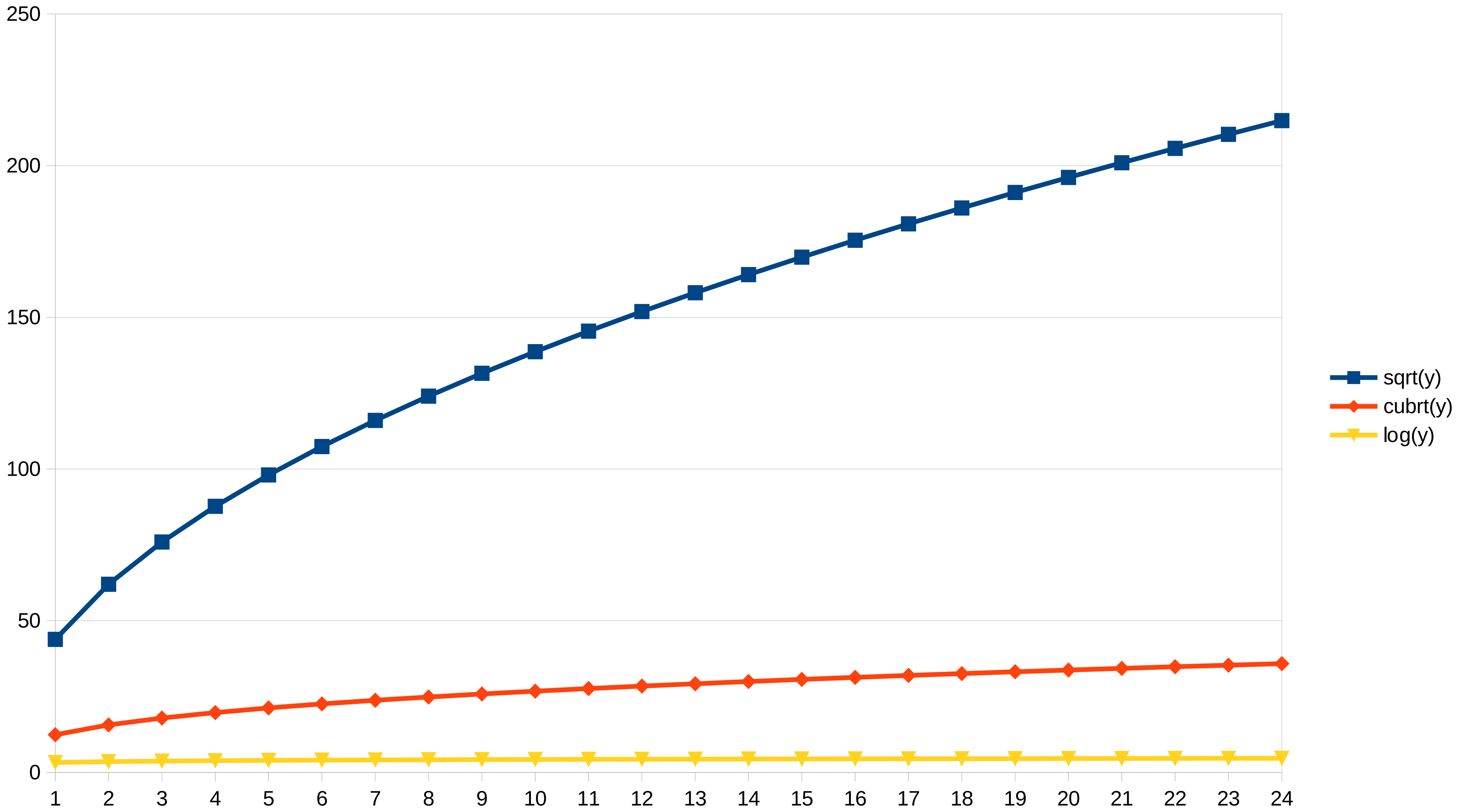
Effects of Transformations on Variables



Where's the other columns of data?

x	y=x/52*10000	sqrt(y)	cubrt(y)	log(y)
1	1923.076923	43.85290097	12.43556587	3.283996656
2	3846.153846	62.01736729	15.6678312	3.585026652
3	5769.230769	75.95545253	17.93518953	3.761117911
4	7692.307692	87.70580193	19.74023034	3.886056648

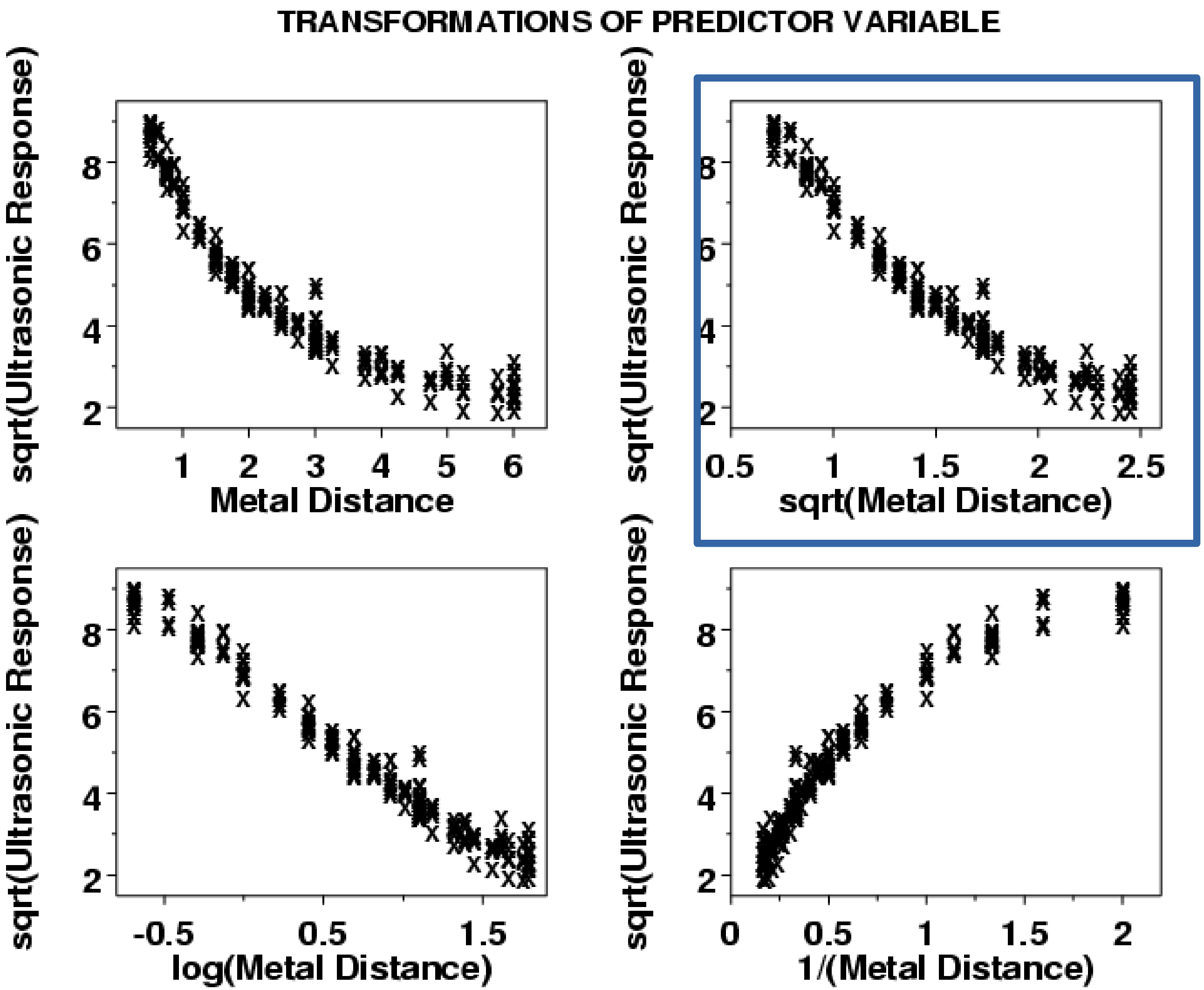
Effects of Transformations on Vars Zoom-in



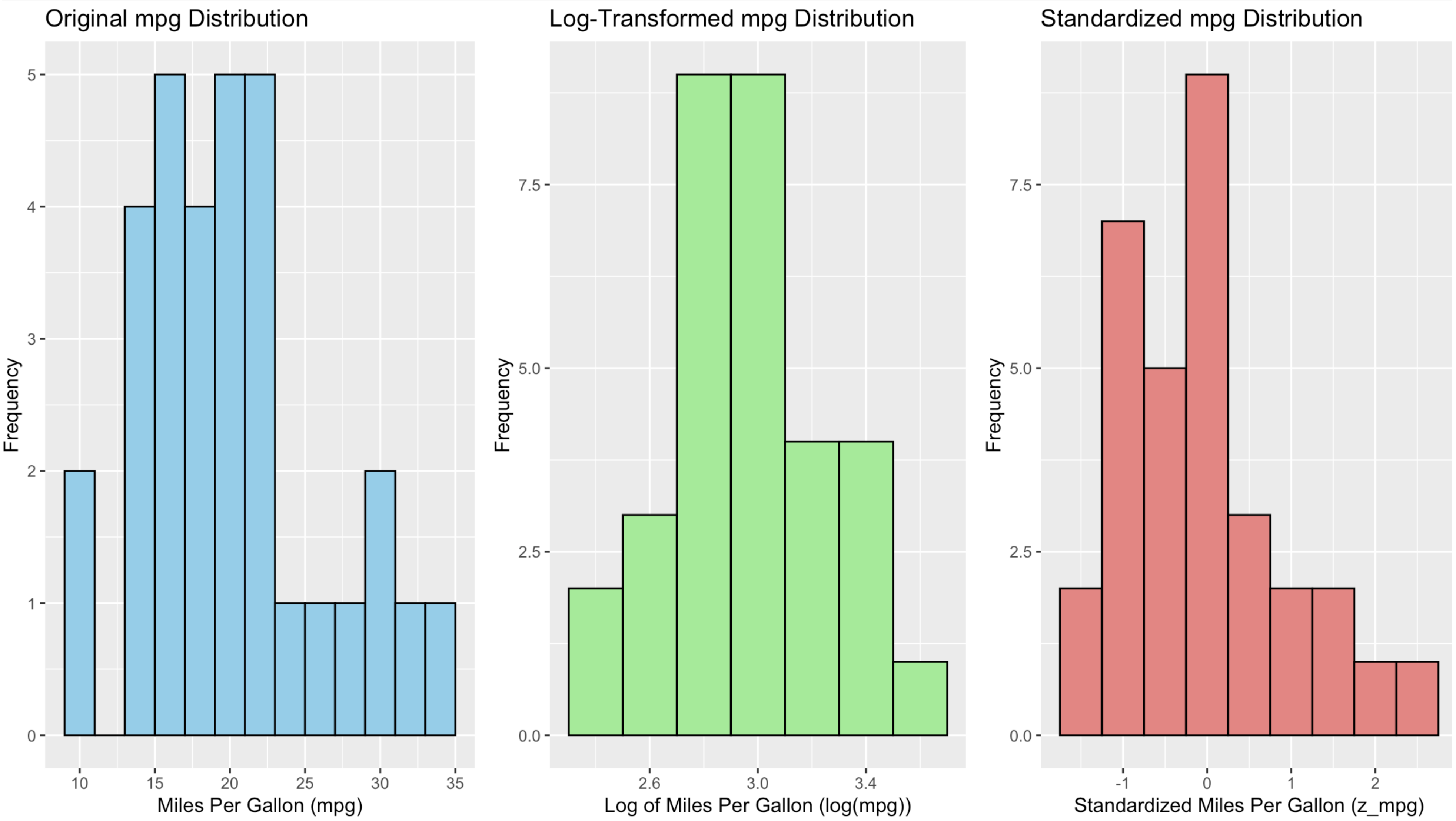
Common Transformations

- Reduce the Y into a smaller space to see trends.
- Places all points on a similar playing ground
- P ← (x,y)
- Trans(p) ←

  (x, sqrt(y))



More On Transformations



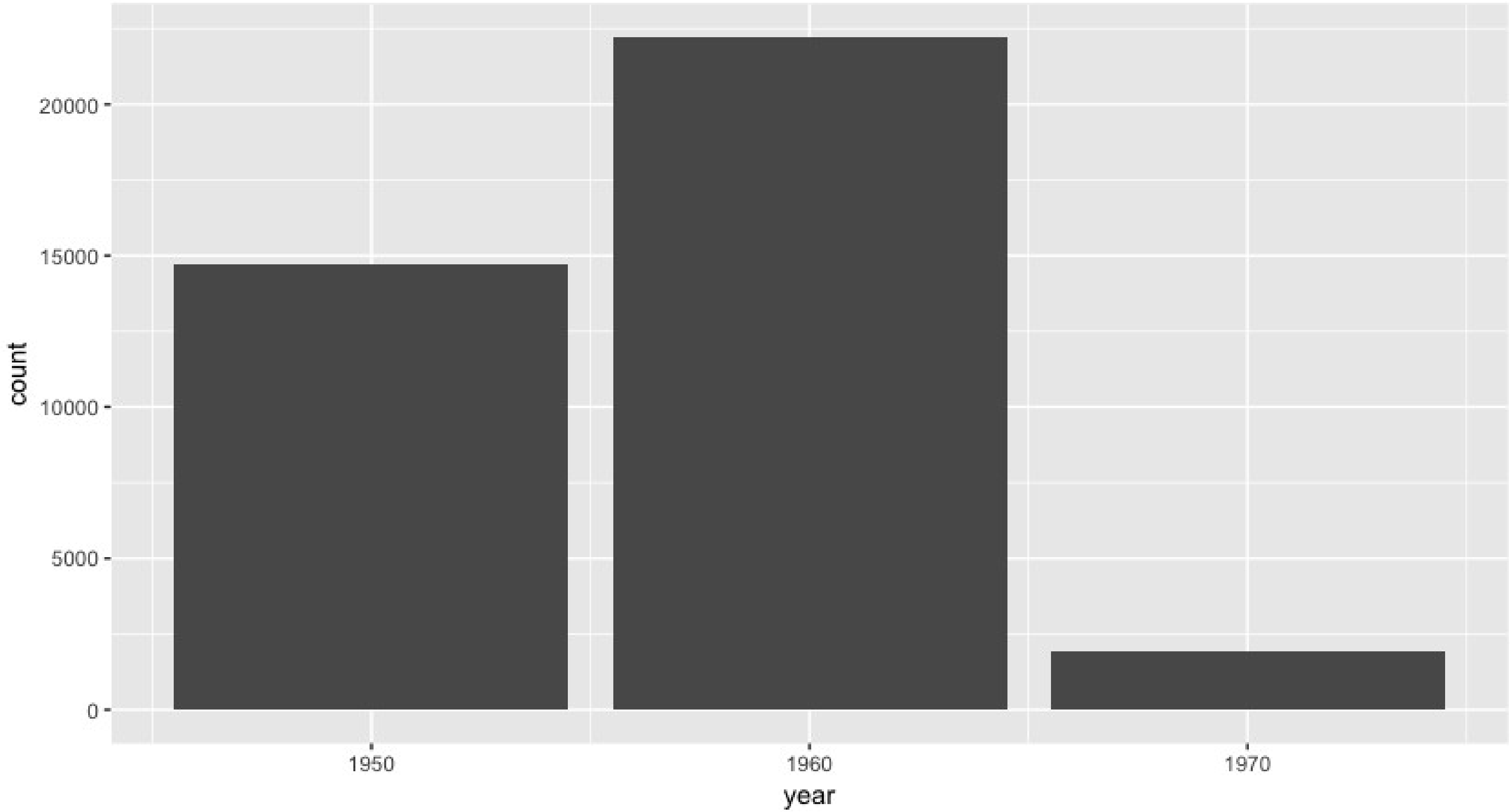
The 1950’s, 1960’s and 1970’s Without Transformation

plot three bars to see what happened in the 1950’s, 1960’s and 1970’s.

```
ggplot(data = dat_caliFocus %>% filter(year == 1950 | year == 1960 | year == 1970)) + geom_bar(mapping = aes(x = year, y = count), stat = "identity")
```

Back to our conversation about vaccines

The 1950's, 1960's and 1970's Without Transformation

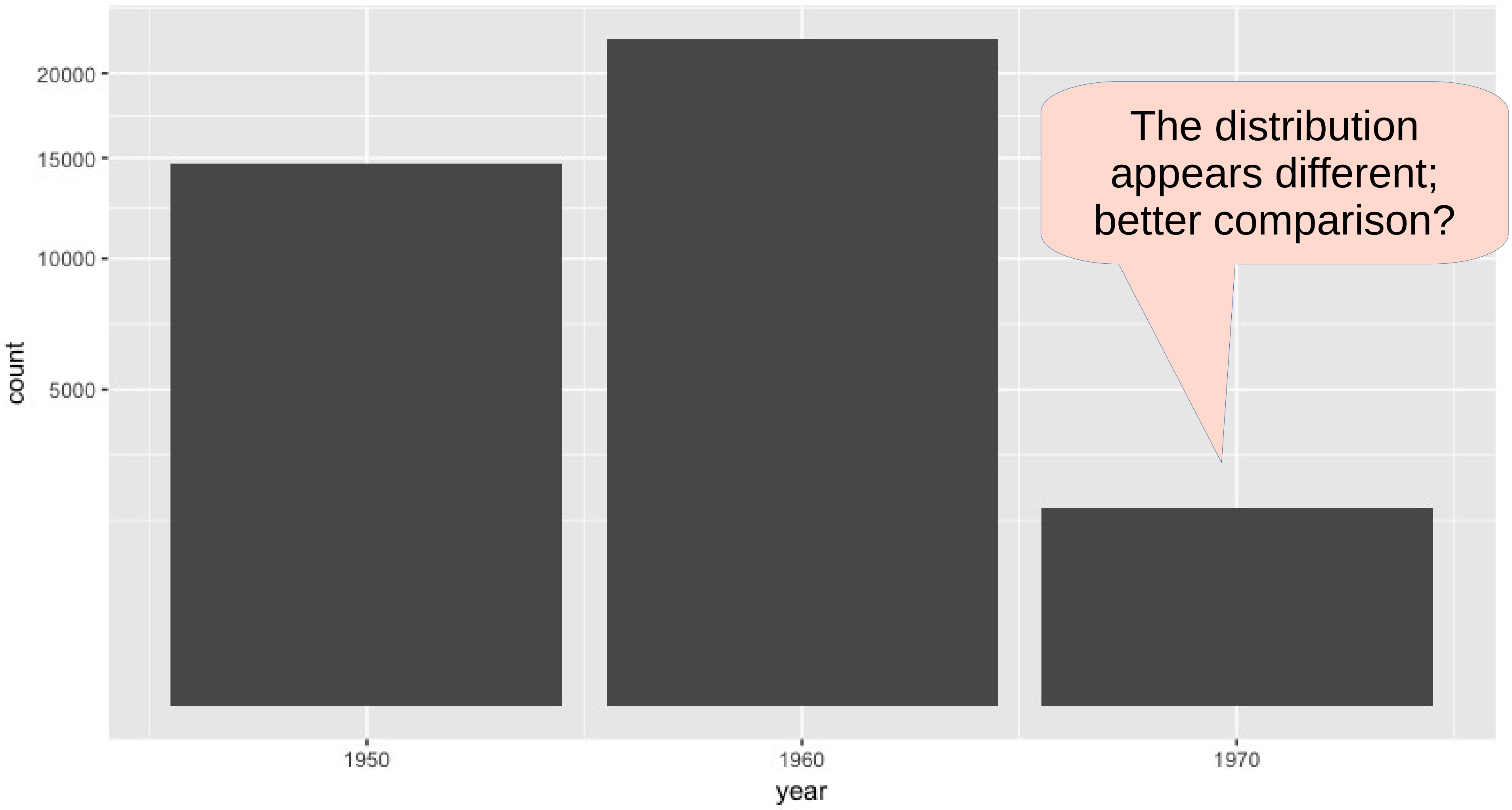


The 1950's, 1960's and 1970's With Sqrt() Transformation

#plot three bars to see what happened in the 1950's, 1960's and 1970's.

```
ggplot(data = dat_caliFocus %>% filter(year == 1950 | year == 1960 | year == 1970)) + geom_bar(mapping = aes(x = year, y = sqrt(count)), stat = "identity")
```

The 1950's, 1960's and 1970's
With Sqrt() Transformation



The 1950’s, 1960’s and 1970’s
Without Transformation

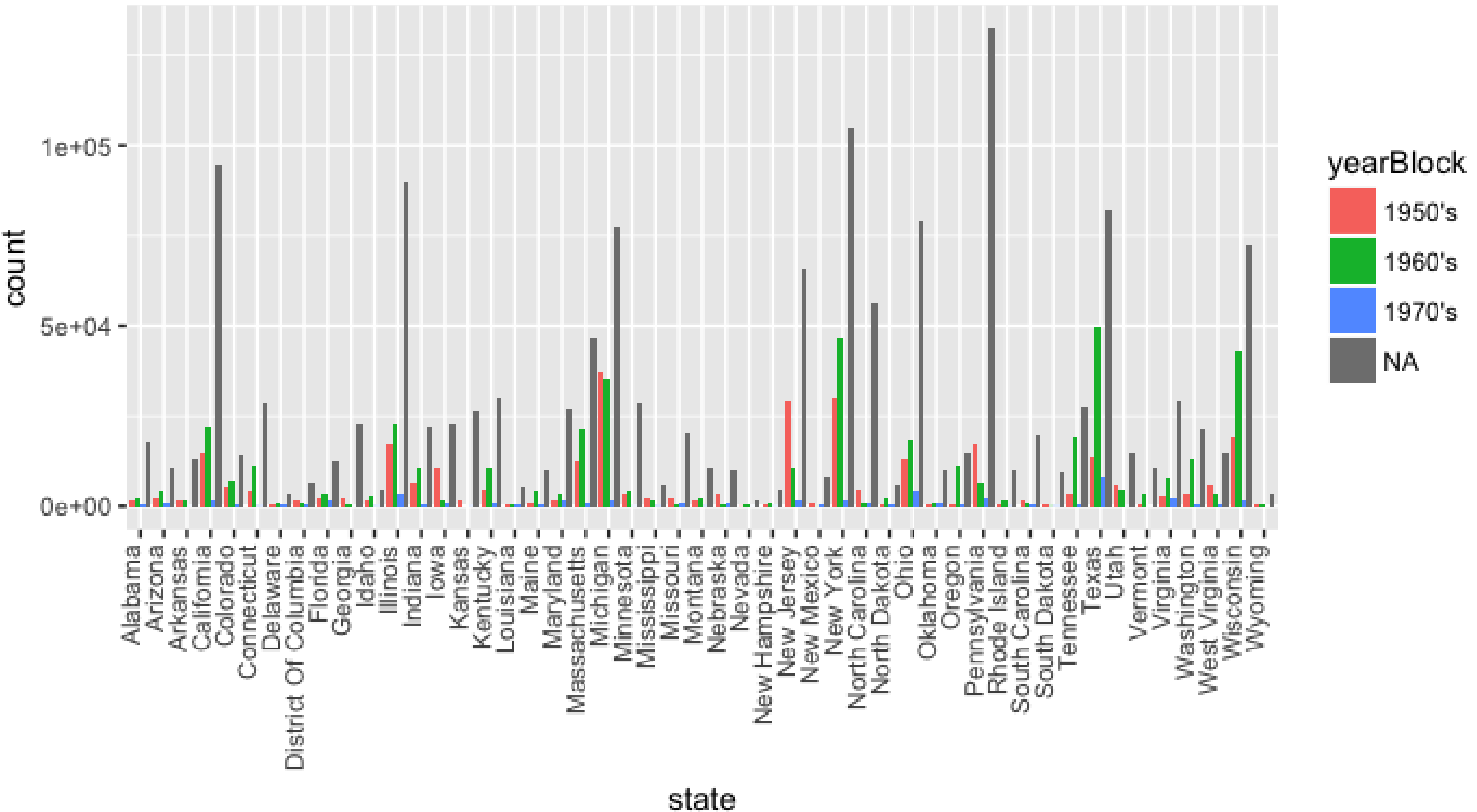
#create some “block”, containers to hold the data for each year.

```
dat_measles_rate_lessTwoStates$yearBlock[dat_measles_rate_lessTwoStates$year == 1950] <- "1950’s"

dat_measles_rate_lessTwoStates$yearBlock[dat_measles_rate_lessTwoStates$year == 1960] <- "1960’s"

dat_measles_rate_lessTwoStates$yearBlock[dat_measles_rate_lessTwoStates$year == 1970] <- "1970’s"
```

#Without transformation, Multi-bar per state,

```
ggplot(data = dat_measles_rate_lessTwoStates) + geom_bar(mapping = aes(x = state, y = count, fill = yearBlock), position = "dodge", stat = "identity") + theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust=-0.01))
```

The 1950’s, 1960’s and 1970’s
Without Transformation

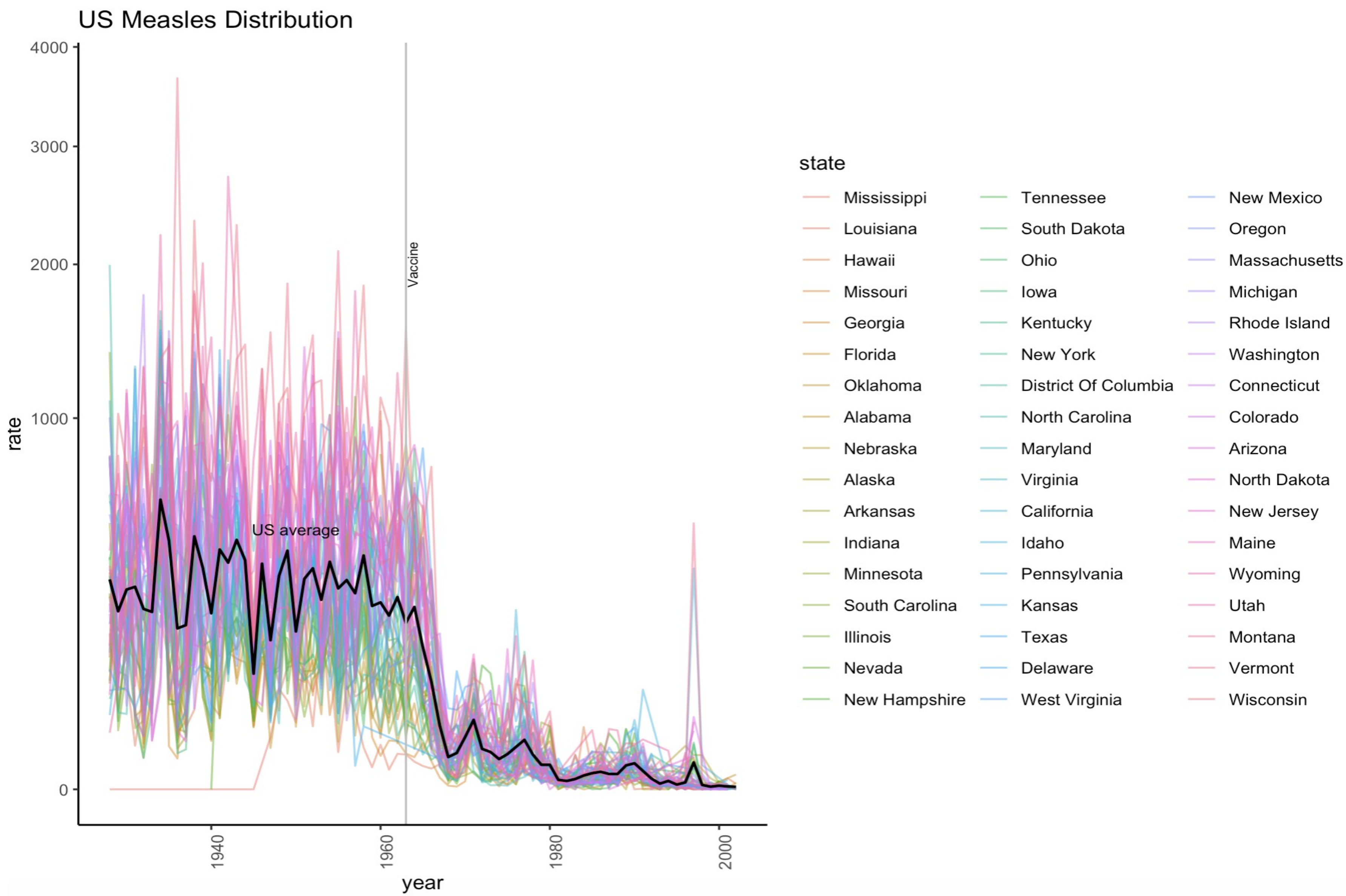


Comparing States;Two plots

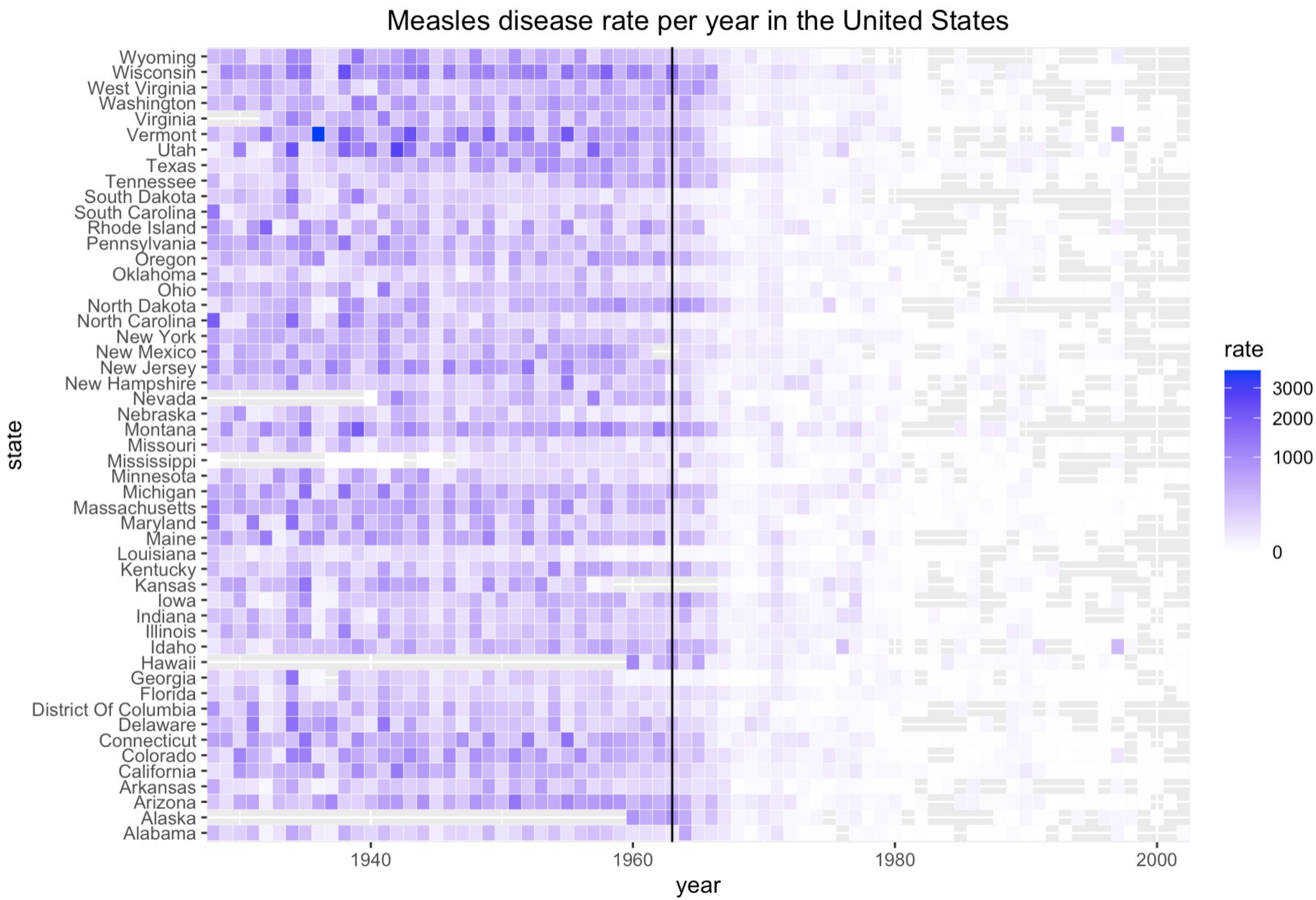
No transformation

```
dat_measles_rate %>%

  filter(!is.na(rate)) %>% mutate(state = reorder(state, rate, FUN = mean)) %>%
ggplot(aes(year,rate, color = state)) + geom_line(alpha = 0.5) + theme_classic() +
theme(axis.text.x = element_text(angle = 90, hjust = 1)) + scale_y_sqrt() +
stat_summary(fun.y=mean, geom="line",lwd=0.7,col="black") + annotate("text", x =
1950, y = 490, label = "US average", size = 3) + ggtitle("US Measles Distribution") +
annotate("text", x = 1963, y= 2000, angle = 90, label = paste("paste(Vaccine)", collapse
= "_"), vjust = 1.2, parse = TRUE, size = 2.5) + geom_vline(xintercept = 1963, col =
"grey")



dat_measles_rate%>% filter(!is.na(rate)) %>%

  ggplot(aes(year, state)) + geom_tile(aes(fill = rate), color = "white") +
scale_fill_gradient(low = "white", high = "blue", trans = "sqrt") +
scale_x_continuous(expand = c(0,0)) + ggtitle("Measles disease rate per year in the
United States") + theme(plot.title = element_text(hjust = 0.5)) + geom_vline(xintercept =
1963, col = "black")
```

Comparing States
No Transformation



Comparing States
No Transformation

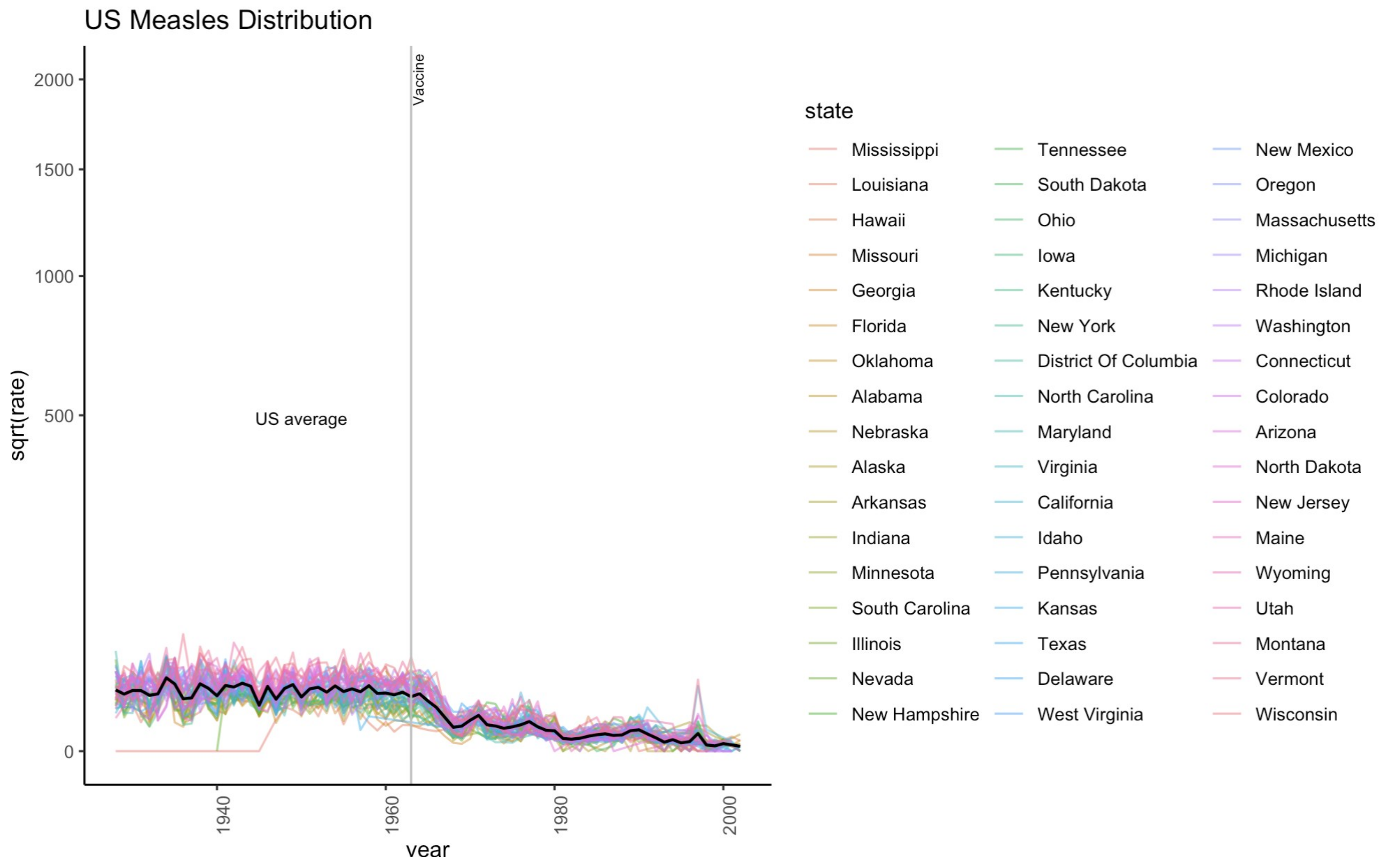


Comparing States
With Transformation

```
dat_measles_rate %>%
  filter(!is.na(rate)) %>% mutate(state = reorder(state, rate, FUN =
mean)) %>%
  ggplot(aes(year,sqrt(rate), color = state)) + geom_line(alpha = 0.5) +
  theme_classic() + theme(axis.text.x = element_text(angle = 90, hjust
= 1)) +
  scale_y_sqrt() + stat_summary(fun.y=mean,
geom="line",lwd=0.7,col="black") +
  annotate("text", x = 1950, y = 490, label = "US average", size = 3) +
  ggtitle("US Measles Distribution") +
  annotate("text", x = 1963, y= 2000, angle = 90, label =
paste("paste(Vaccine)", collapse = "_"), vjust = 1.2, parse = TRUE, size
= 2.5) + geom_vline(xintercept = 1963, col = "grey")
```

Does this help compare?

Comparing States
With Transformation

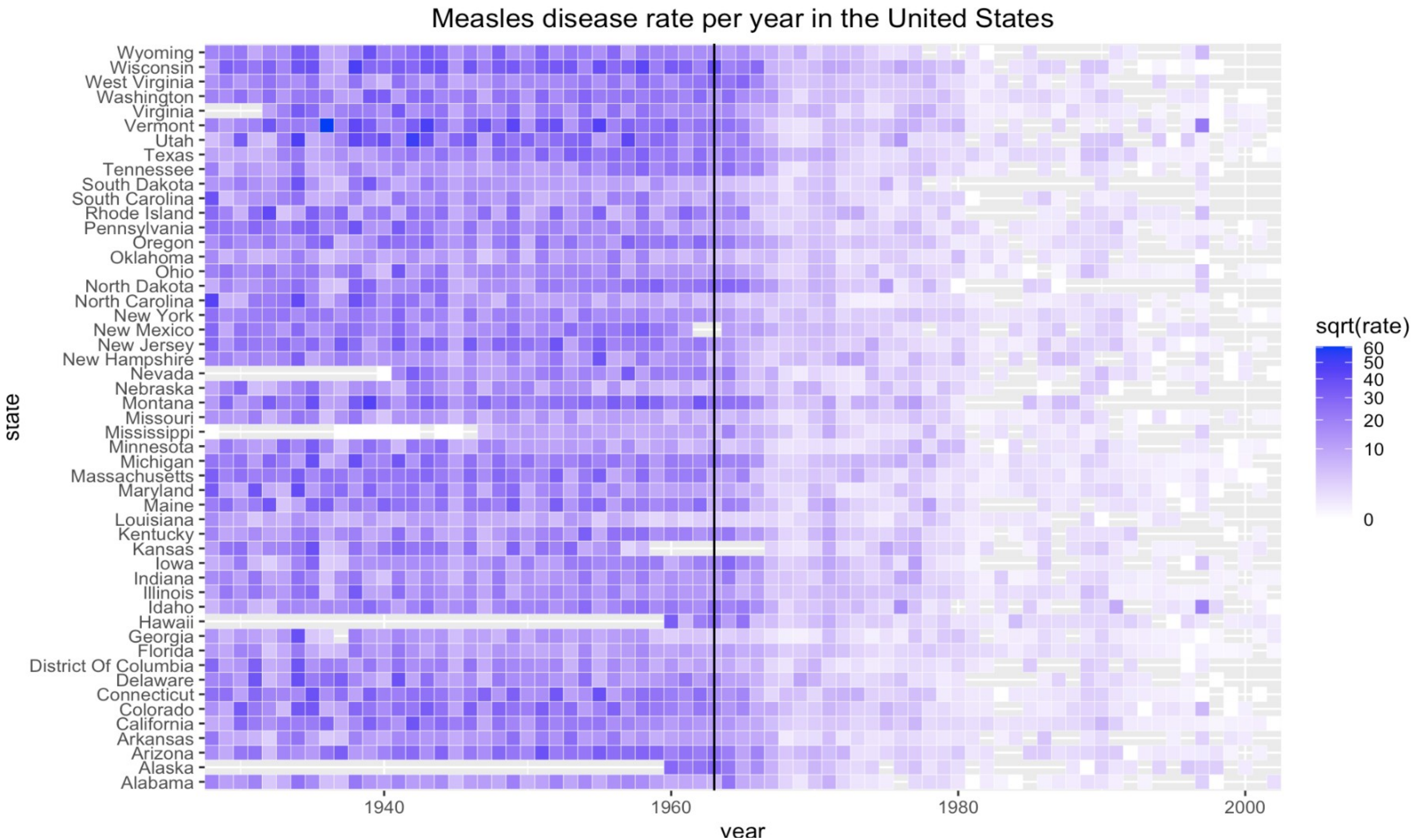


Comparing States
With Transformation

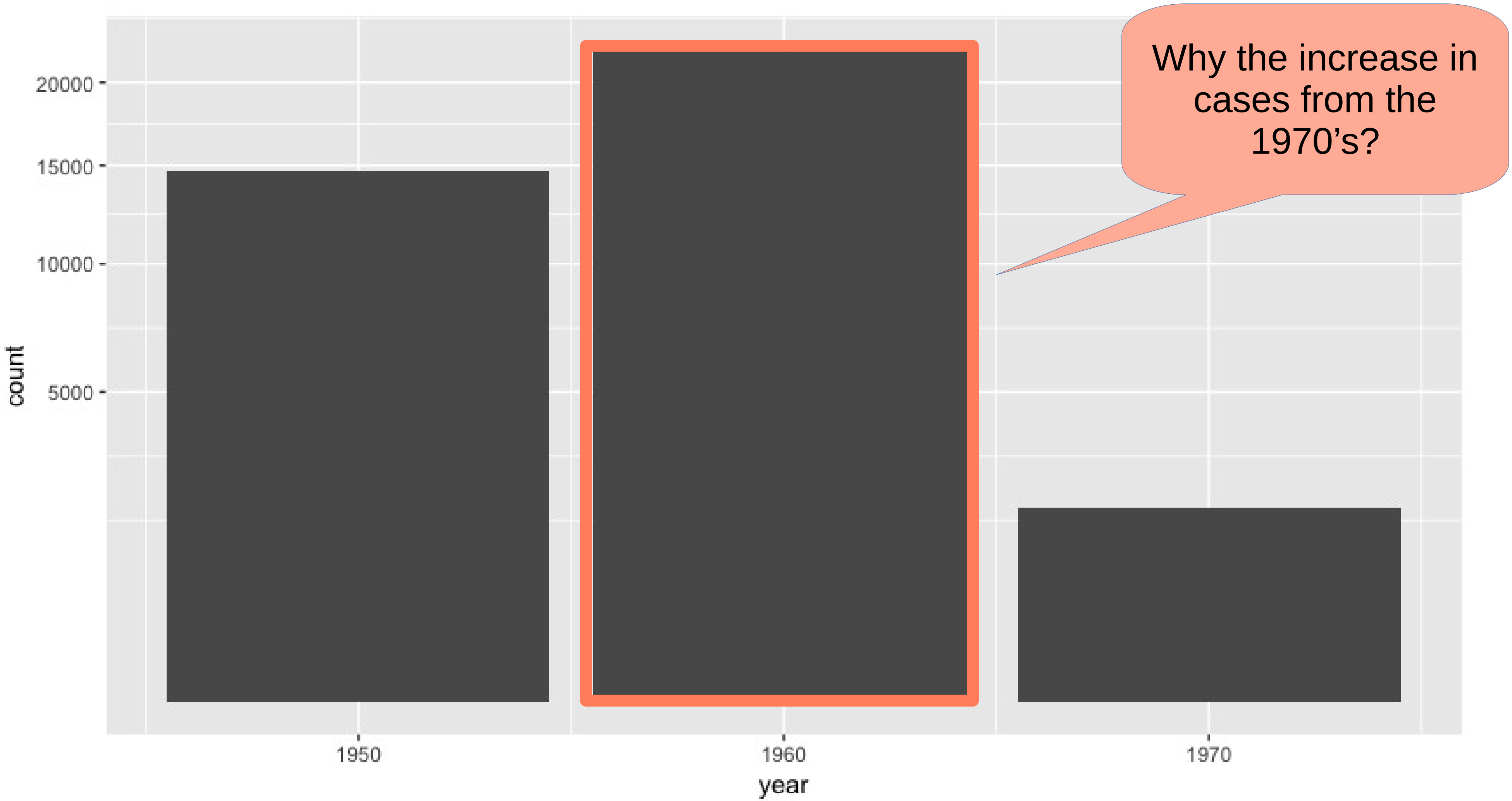
```
# with transform
dat_measles_rate%>% filter(!is.na(rate)) %>%
  ggplot(aes(year, state)) + geom_tile(aes(fill = sqrt(rate)), color = "white") + scale_fill_gradient(low = "white", high = "blue", trans = "sqrt") + scale_x_continuous(expand = c(0,0)) + ggtitle("Measles disease rate per year in the United States") + theme(plot.title = element_text(hjust = 0.5)) + geom_vline(xintercept = 1963, col = "black")
```

Does this help compare?

Comparing States
With Transformation



Going back… Did the 1960's show an increase in cases?

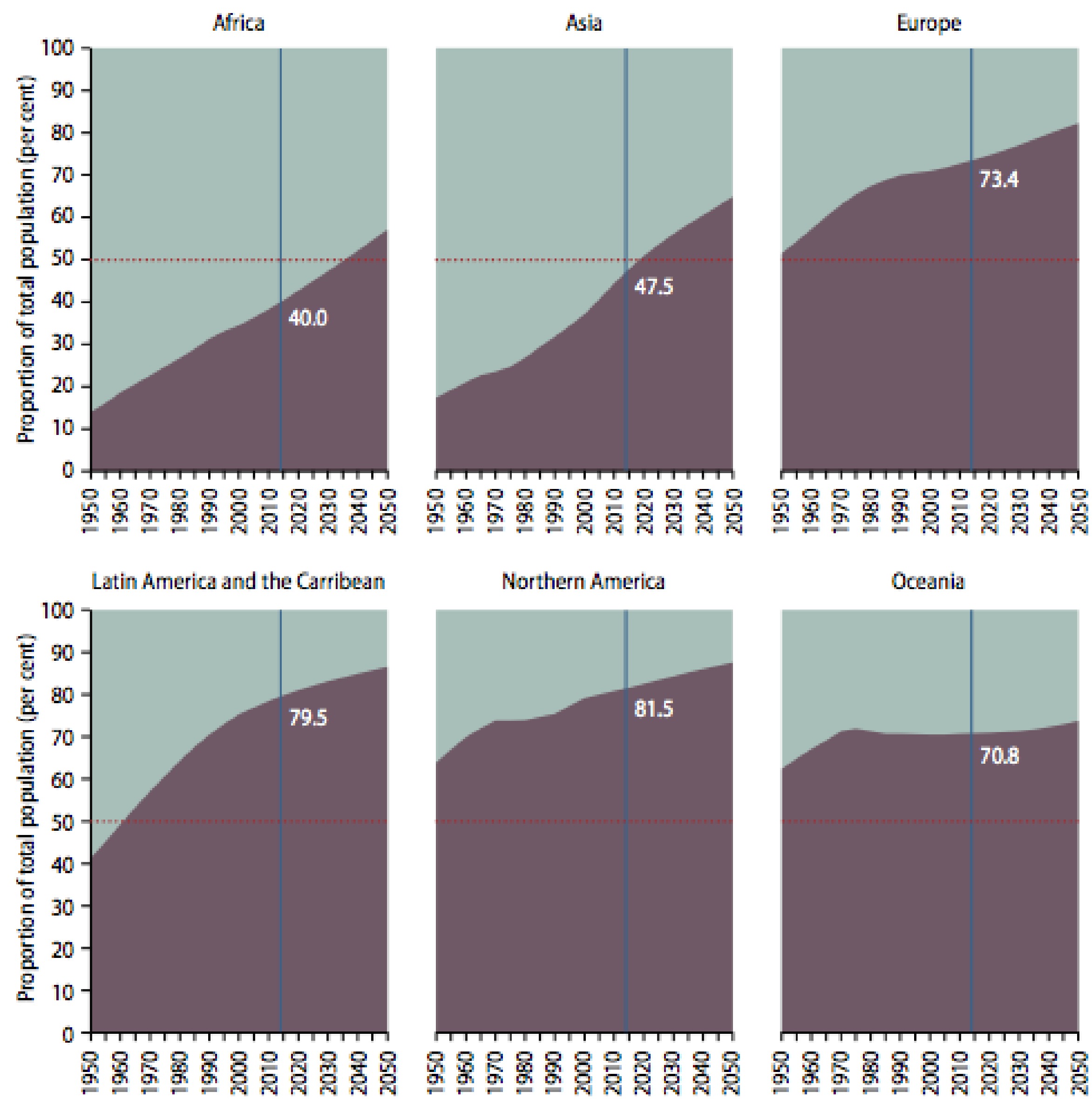


Urban Versus Rural

Urbanization has occurred in all major areas, yet Africa and Asia remain mostly rural

Figure 3.

Urban and rural population as proportion of total population, by major areas, 1950–2050



Urban population

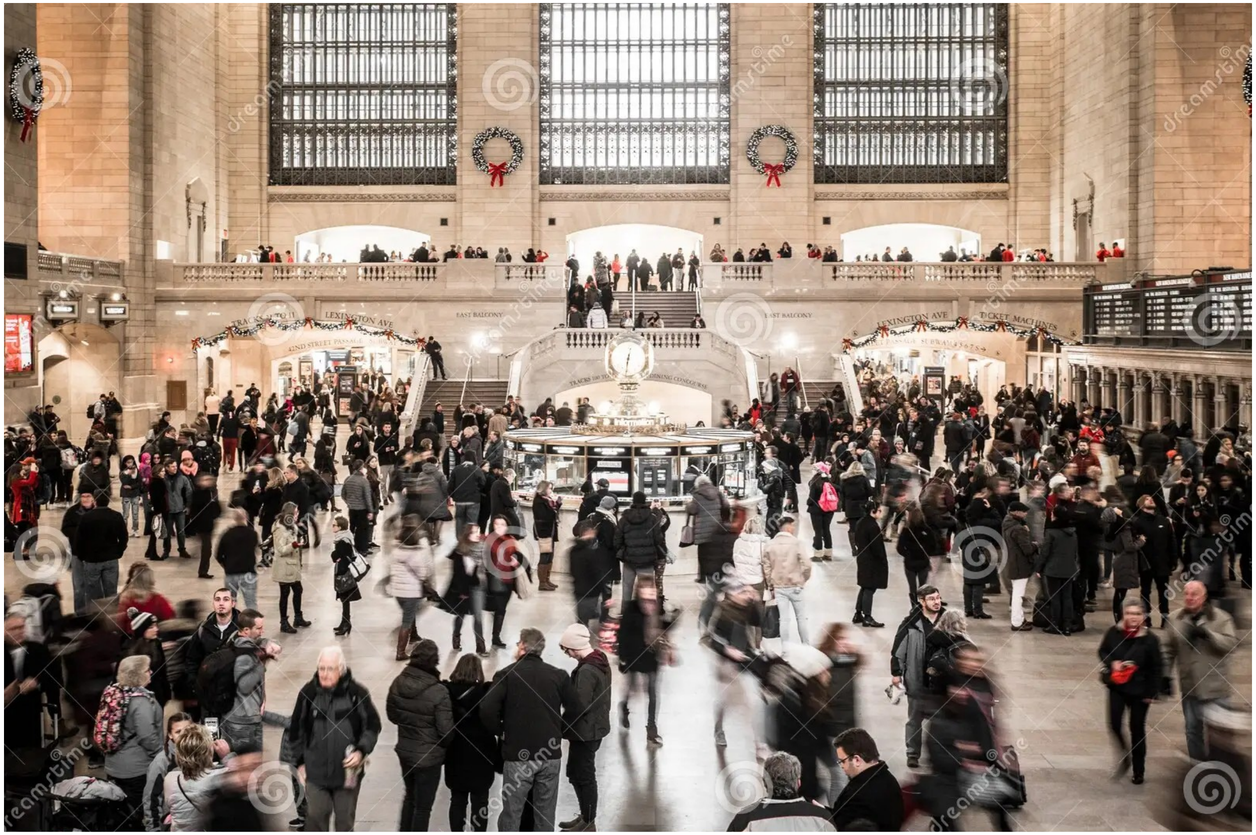
Rural population

- **Urban**: City dwelling
- **Rural**: Country dwelling
- Vaccinations:
  - Were there fewer people available from whom to contract viruses?
- Less opportunity to see others?

https://esa.un.org/unpd/wup/publications/files/wup2014-highlights.Pdf

More People, More Contact

Could a crowding of cities in the 1960's create a better place for measles to spread?



As good as this explanation seems, we still need to verify it by facts.

Concluding Ideas
Why is this *also* important?

Open and Transparent: By making the data about vaccines available to the general public, anyone can study the data to resolve questions of their own.

Building Further: This promotes further transparency, credibility, and encourages further exploration of the topic.

Implications: The potential implications of the findings on vaccine policy, public health, or vaccine development may be further explored, with new data released to the public to increase trust.

Concluding Ideas
Why is this *also* important?

Furthering Trust: Similar studies may be convinced to also release their data to facilitate investigators to study it and develop deeper understandings of health-related research.

Positive impact: The stakeholders of vaccine education and development would be able to help healthcare providers, patients, and members of communities make informed decisions about vaccine usage.