Data Analytics
CS301
Relational Data

Fall 2020

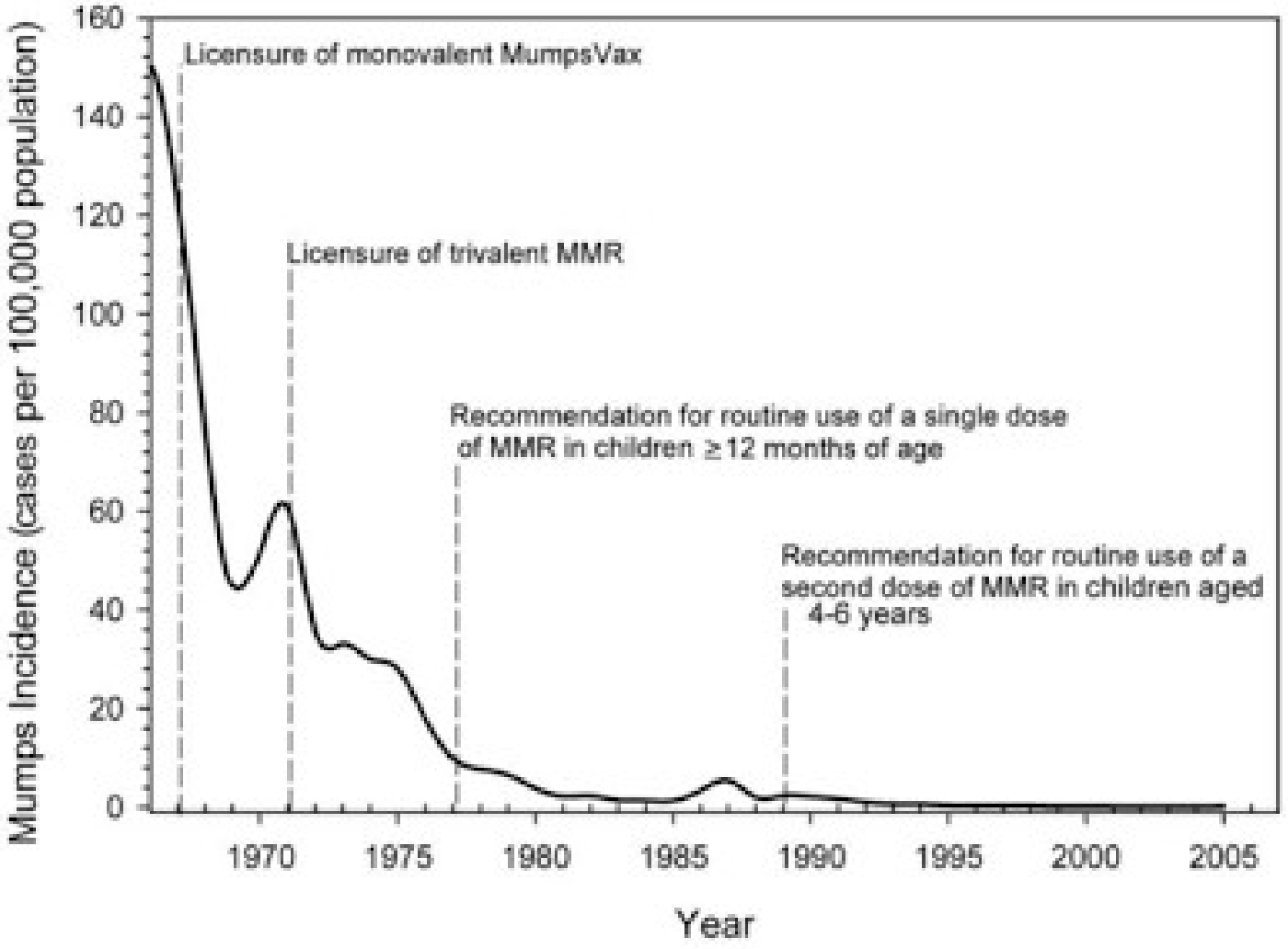
Oliver BONHAM-CARTER

Let's Talk About the Vaccine Lab For A Moment...

- How do you know if therapies are 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://www.historyofvaccines.org/timeline

When to Use Vaccines?



Blog: http://ruleof6ix.fieldofscience.com/2011/10/vaccines-can-you-predict-how-well.html

Do Vaccines Work?

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:

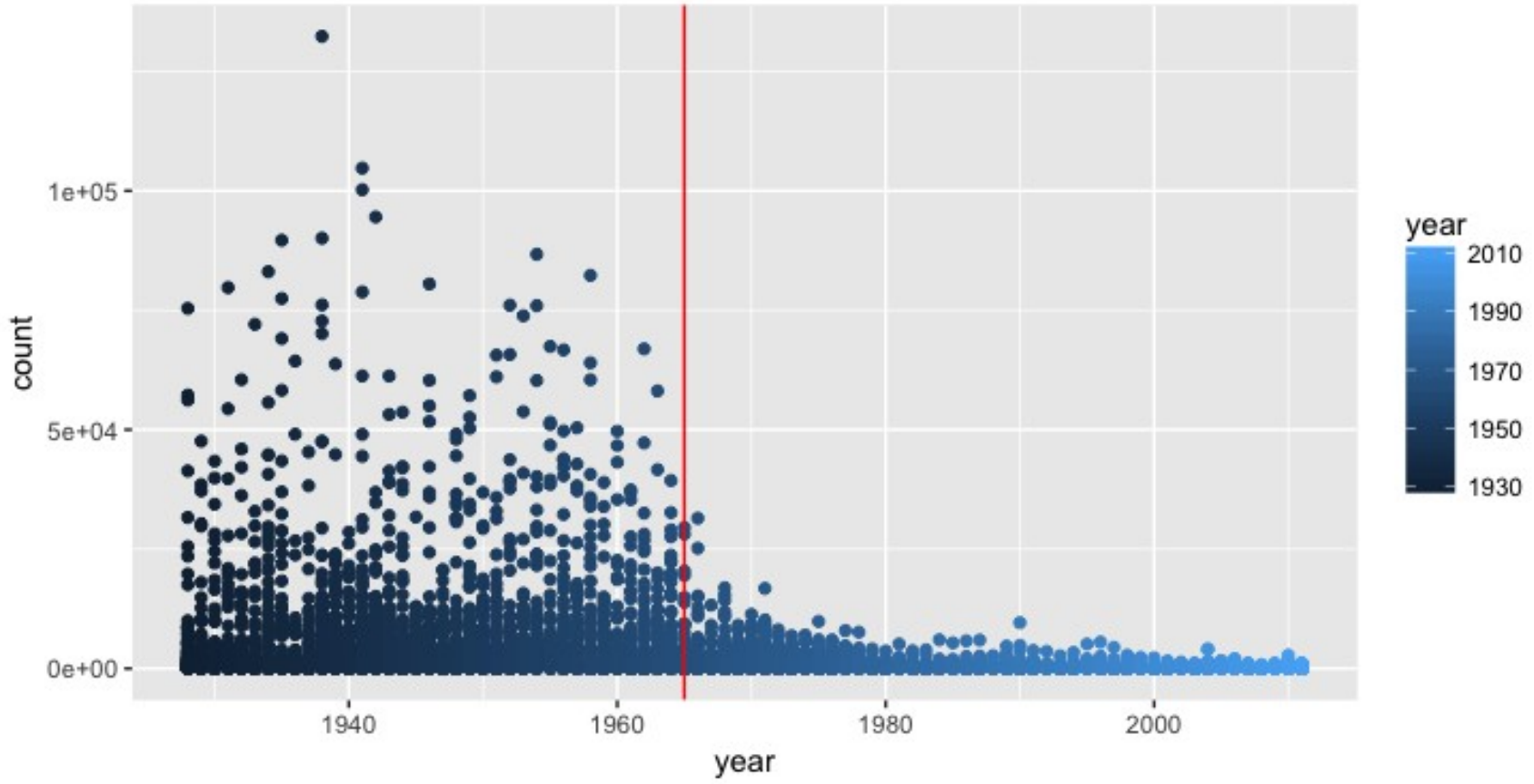
* *JAMA*. 2007;298(18):2155-2163

† 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

- #1) 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/(population / 100000) / (52 / weeks_reporting))

  # 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

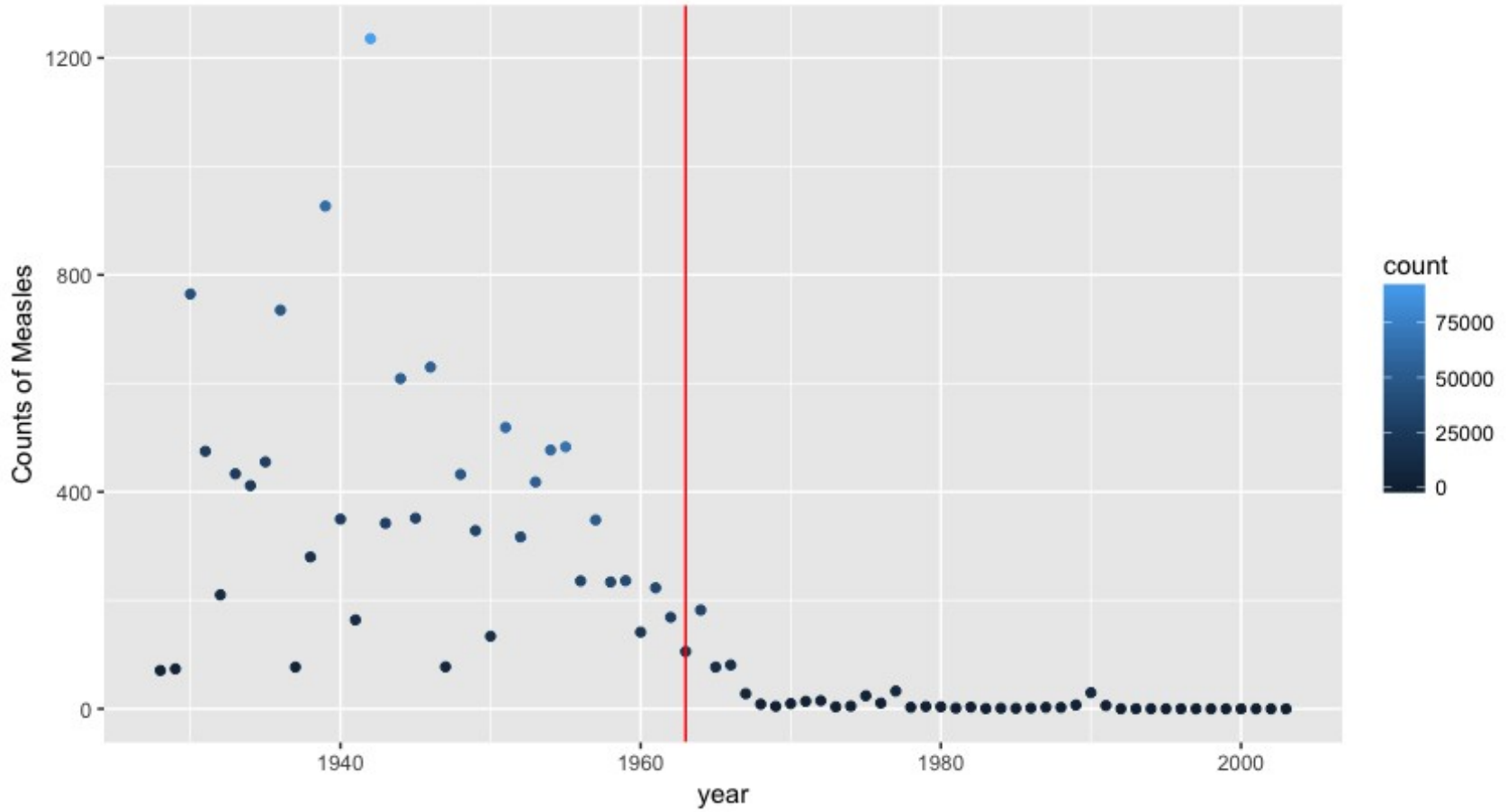
  dat_caliFocus <- filter(dat_measles_rate_lessTwoStates, state == "California")

  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

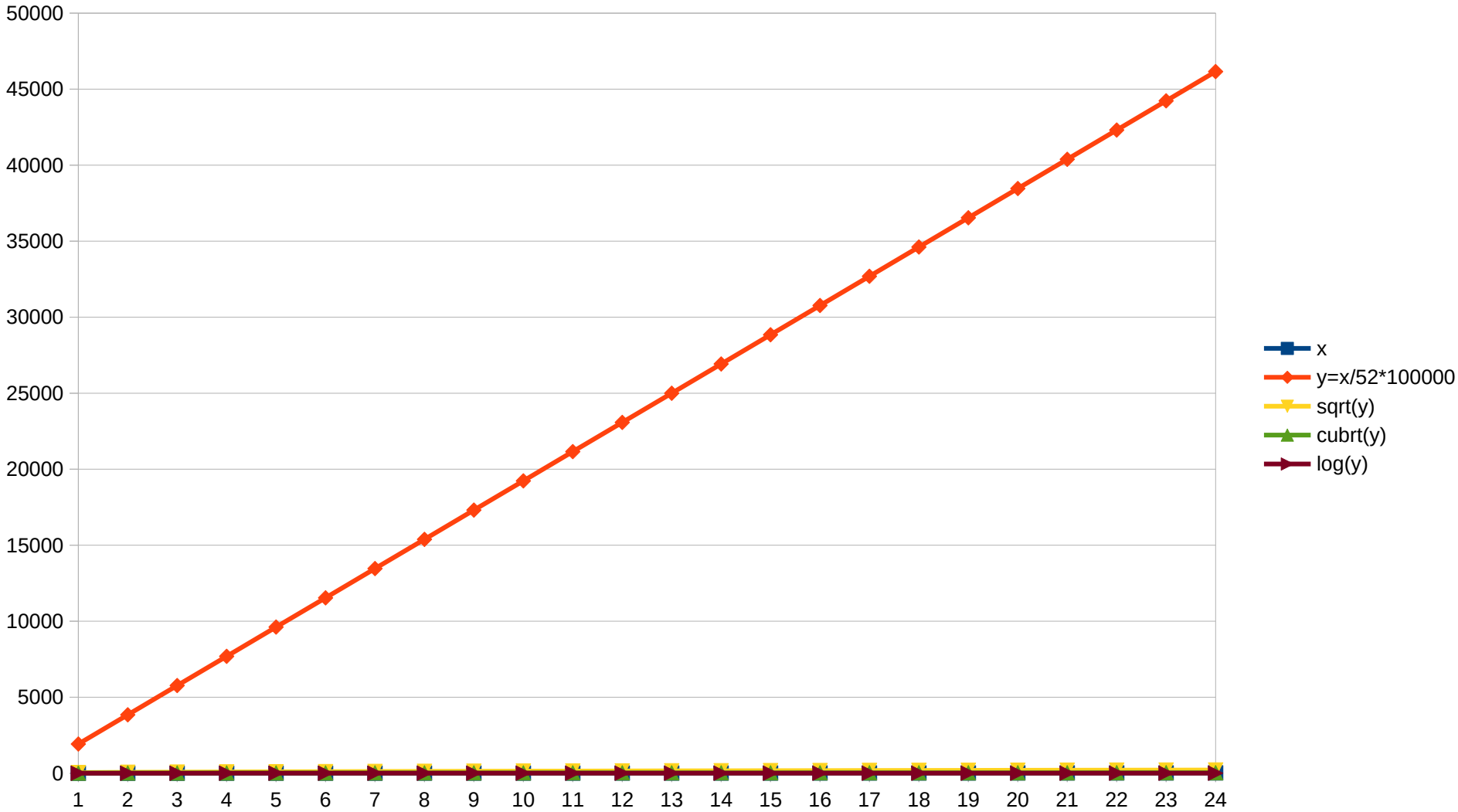
- ggplot(data = dat_caliFocus, mapping = aes(x = year, y = rate, color = count)) + geom_point() + geom_vline(xintercept = 1963, color = "red") + labs(y = "Counts of Measles")



Transformations Help to Fit the Data

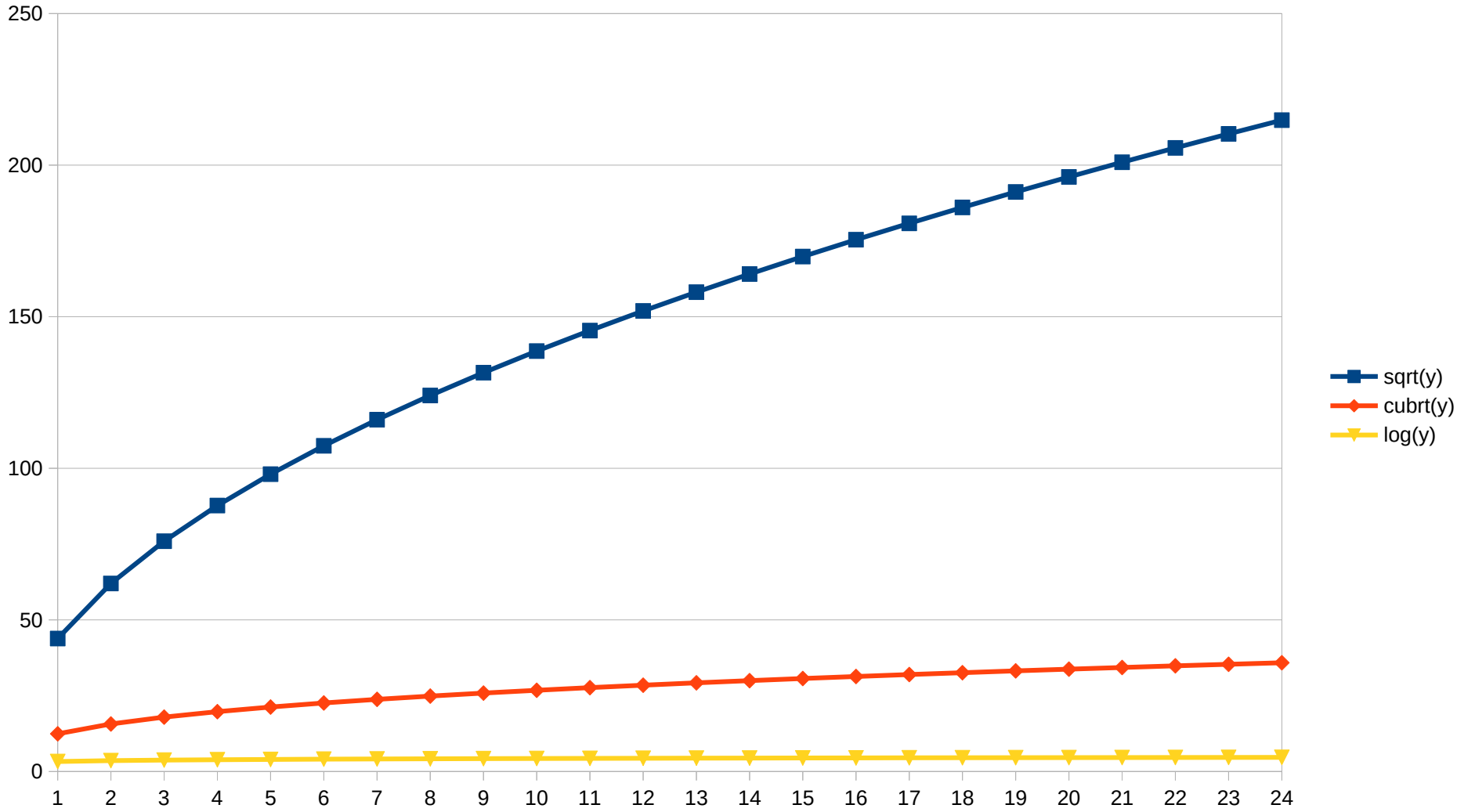
- Messy Data is hard to work with…
- The square root, x to x^(1/2) = sqrt(x), is a transformation with a moderate effect on distribution shape.
- Weaker than the logarithm and the cube root transformations
- Used for reducing right skewness
- Has the advantage that it can be applied to zero values

Effects of Transformations on Variables



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

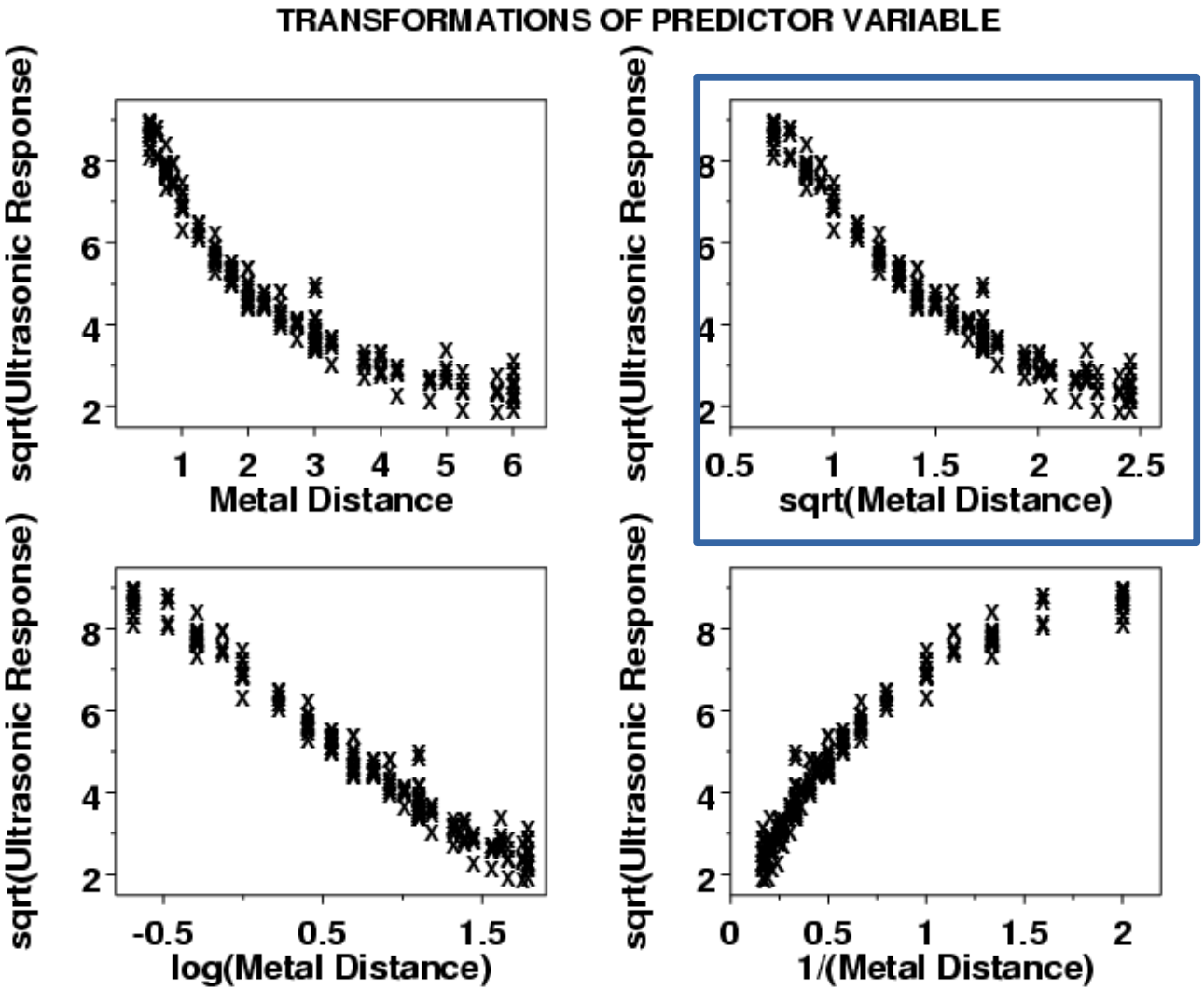
Effects of Transformations on Variables, Zoomed-in



Transformations Help to Fit the Data (example)

- 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))

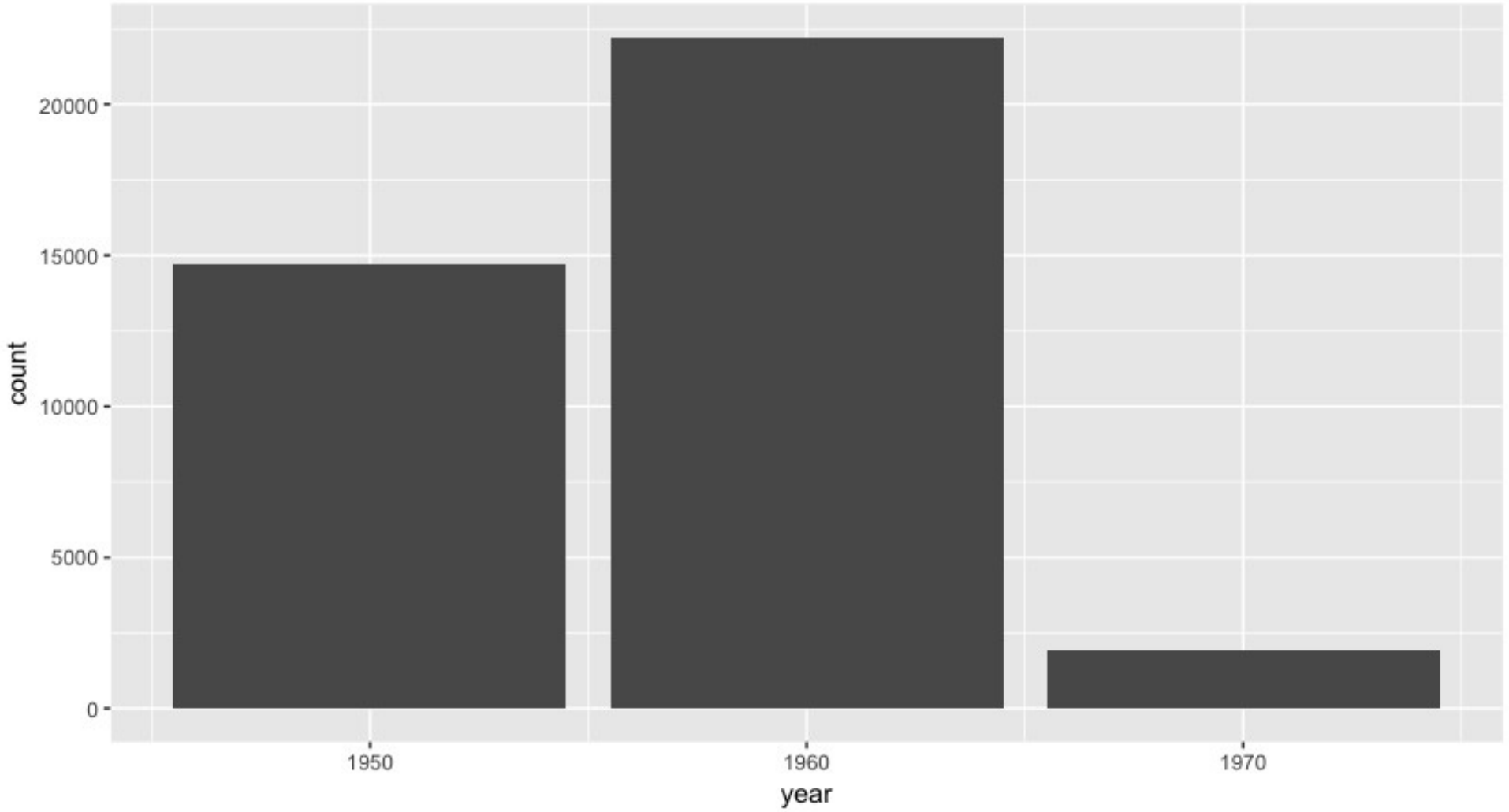


The 1950's, 1960's and 1970's Without Transformation

```
# Plot three bars to see cases 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")
```

The 1950's, 1960's and 1970's

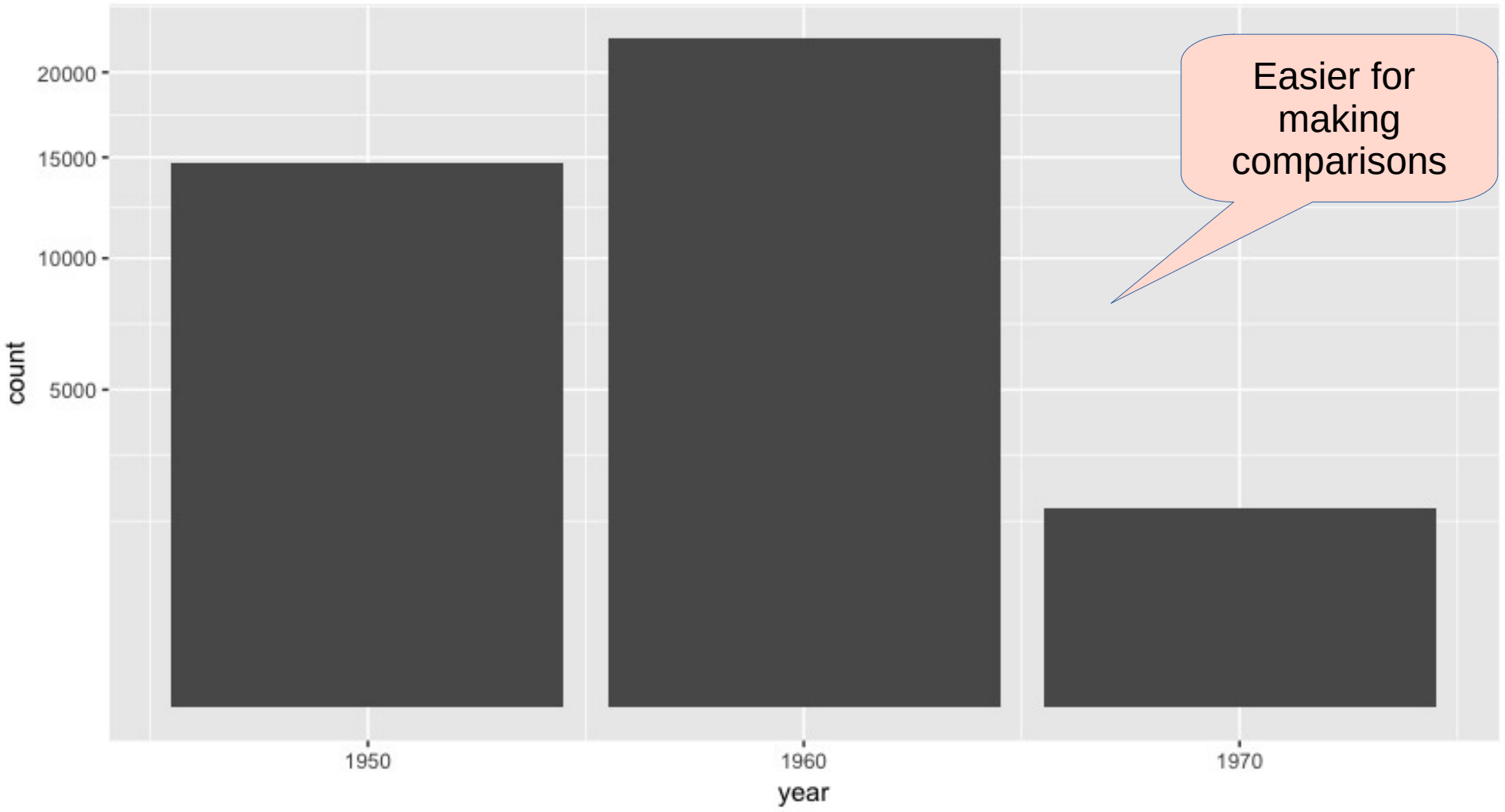
Without Transformation



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

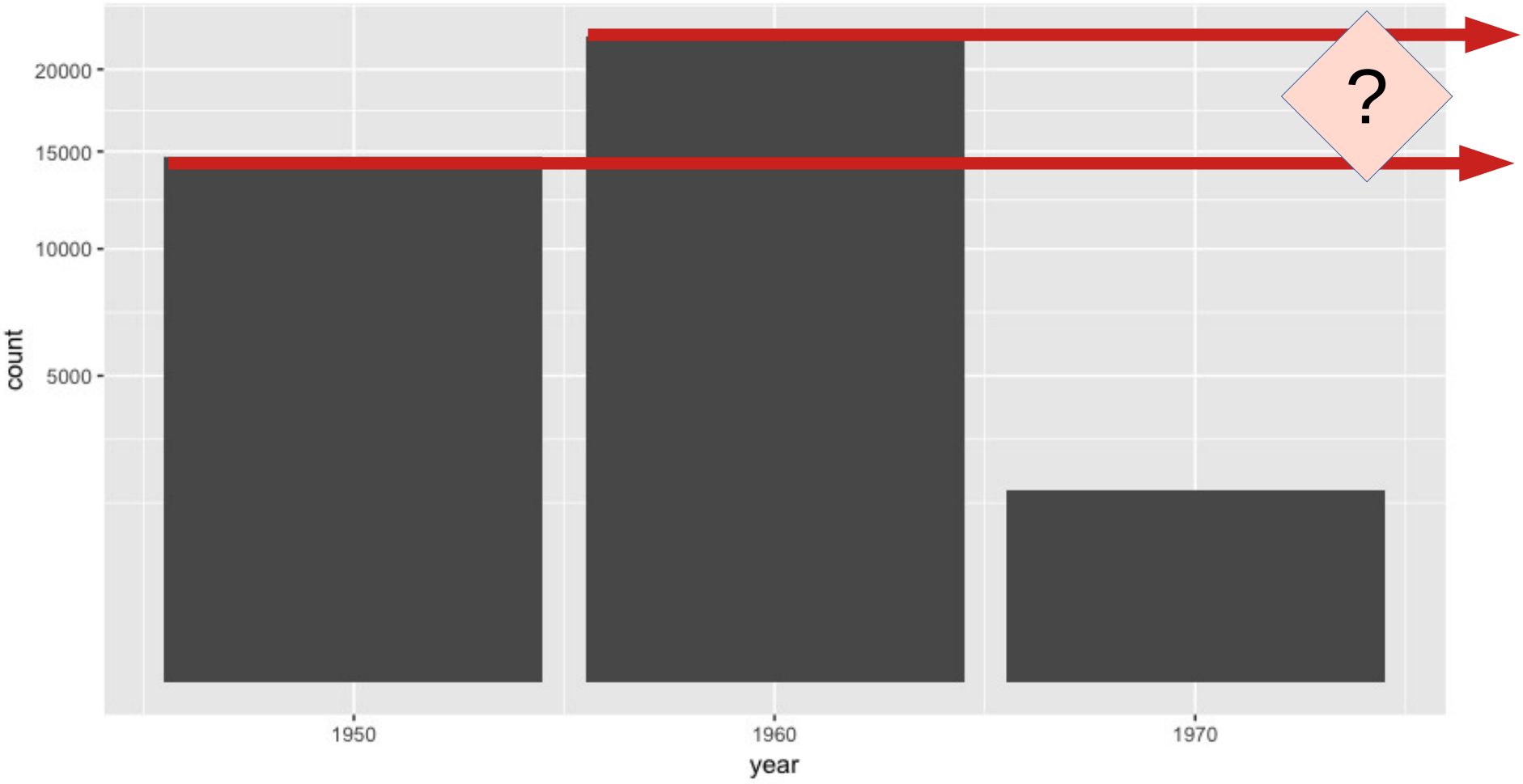
```
# Plot three bars to see cases in the 1950's, 1960's and 1970's.

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



Why the Rise?

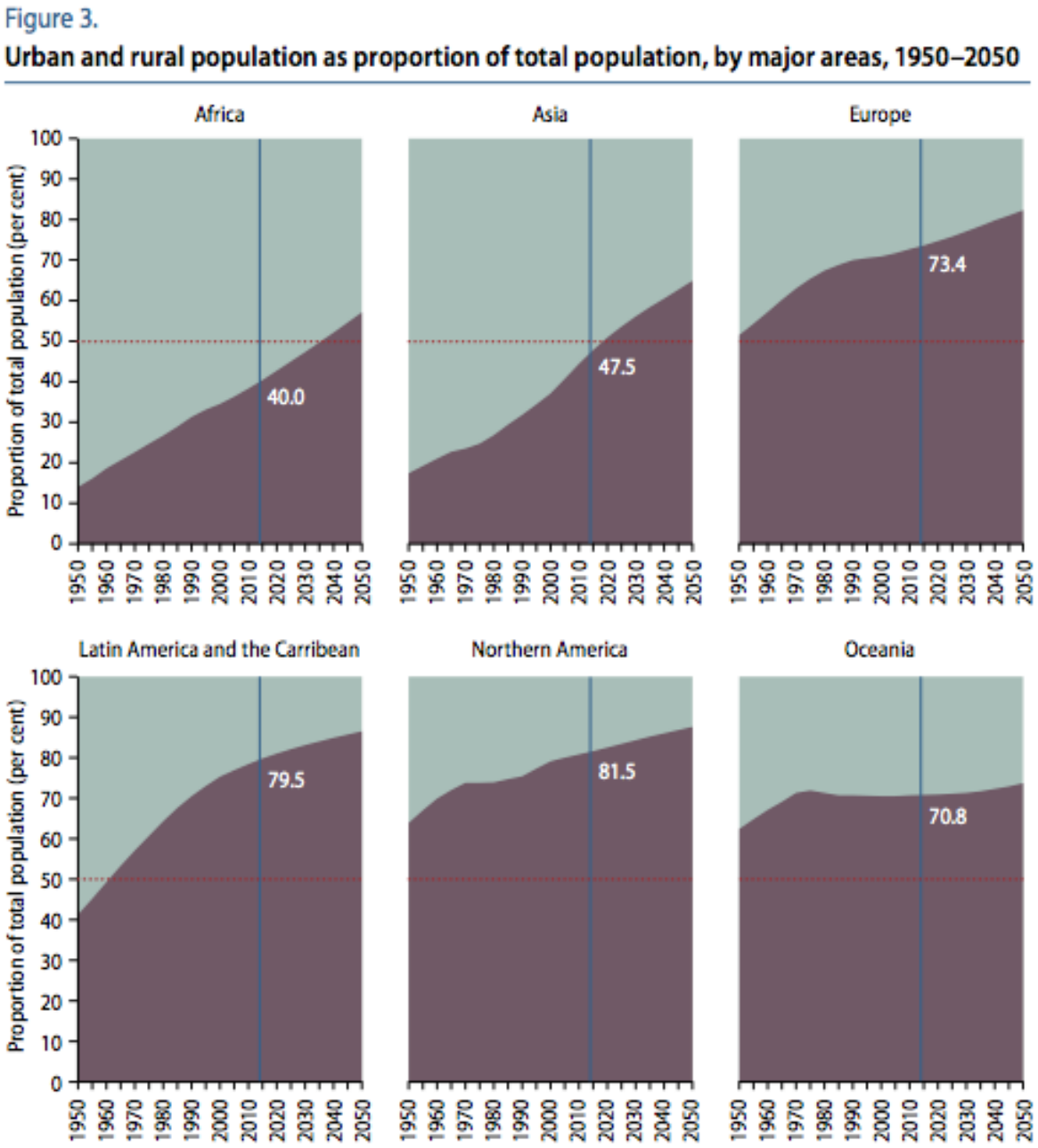
Can we explain why the cases in the 1960's were more than the 1950's?



Urban Versus Rural

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

- **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

Rural Living

- People began to migrate to populated and urban areas (cities) leading up to the 1960's.
- Close contact with others made it easy for ailments to transfer.

Histograms of the Cases in California by Year

```
# Create table to focus on California
dat_caliFocus <- filter(dat_measles_rate_lessTwoStates, state == "California")

dat_caliFocus$yearBlock[dat_caliFocus$year >= 1920 & dat_caliFocus$year <= 1929] <- "1920’s"
dat_caliFocus$yearBlock[dat_caliFocus$year >= 1930 & dat_caliFocus$year <= 1939] <- "1930’s"
dat_caliFocus$yearBlock[dat_caliFocus$year >= 1940 & dat_caliFocus$year <= 1949] <- "1940’s"
dat_caliFocus$yearBlock[dat_caliFocus$year >= 1950 & dat_caliFocus$year <= 1959] <- "1950’s"
dat_caliFocus$yearBlock[dat_caliFocus$year >= 1960 & dat_caliFocus$year <= 1969] <- "1960’s"
dat_caliFocus$yearBlock[dat_caliFocus$year >= 1970 & dat_caliFocus$year <= 1979] <- "1970’s"
dat_caliFocus$yearBlock[dat_caliFocus$year >= 1980 ] <- "1980's onward"
```

Set up the histogram blocks (of cases)

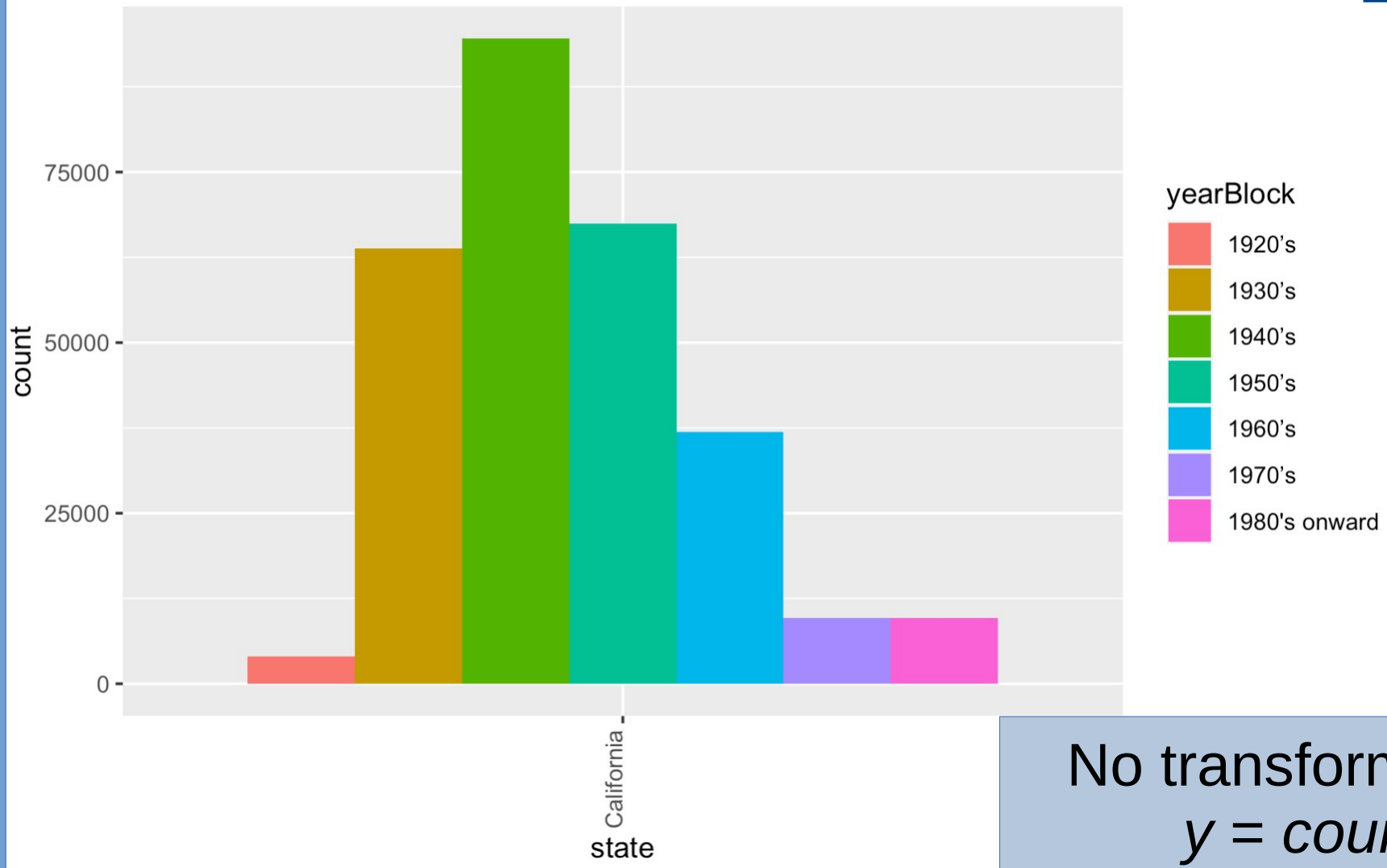
Histograms of the Cases in California by Year

No transformation
y = count

```
ggplot(data = dat_caliFocus ) +
  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))
```

Plot the Blocks (of cases)

Histograms of the Cases in California by Year



No transformation
y = count

Plot the Blocks (of cases)

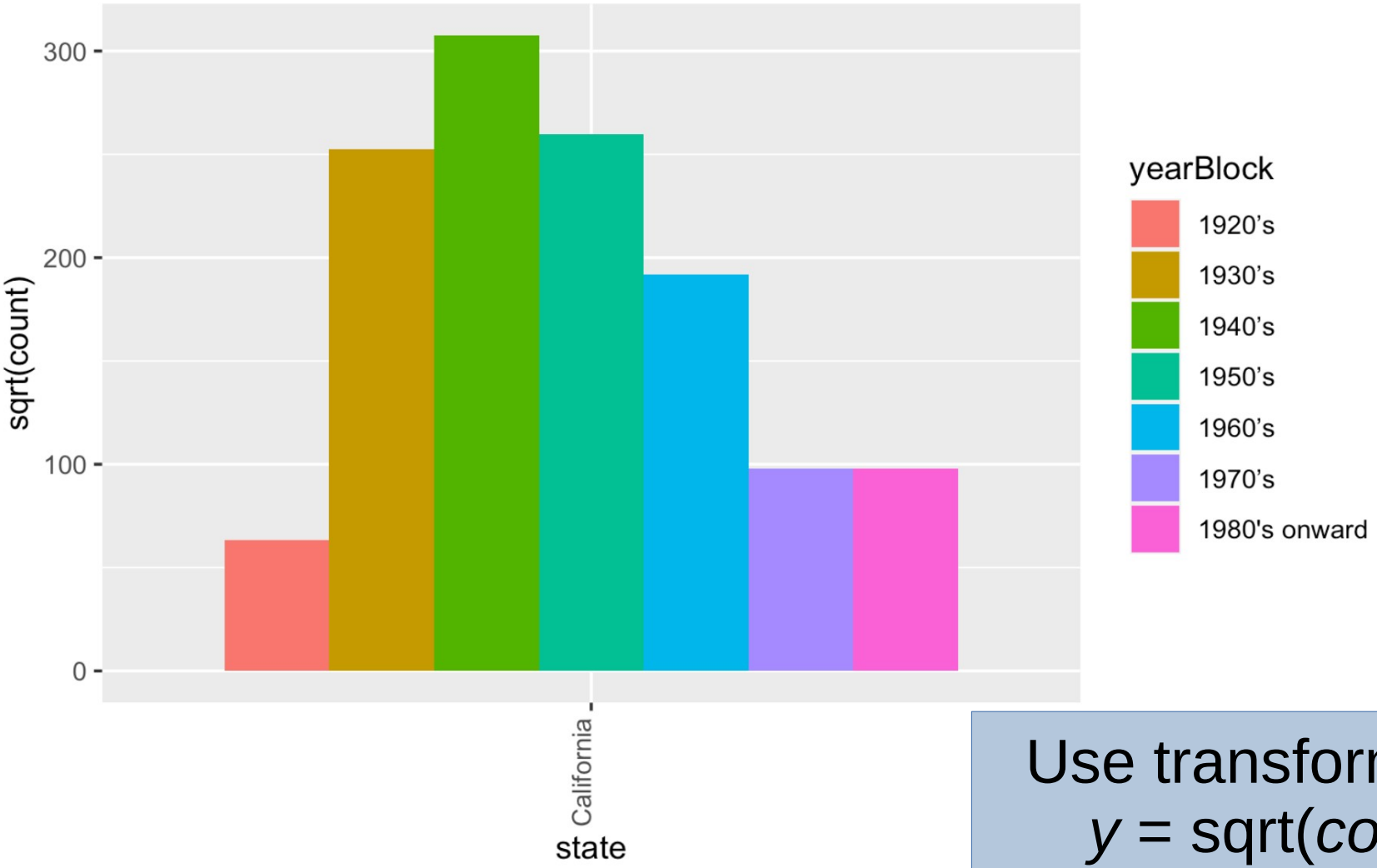
Histograms of the Cases in California by Year

Use transformation
y = sqrt(*count*)

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

Plot the Blocks (of cases)

Histograms of the Cases in California by Year



Use transformation
y = sqrt(*count*)

Plot the Blocks (of cases)

Histograms of the Cases in States by Year

```
# Focus on all states

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


dat_measles_rate_lessTwoStates$yearBlock[dat_measles_rate_lessTwoStates$year >= 1920 & dat_measles_rate_lessTwoStates$year <= 1929] <- "1920’s"

dat_measles_rate_lessTwoStates$yearBlock[dat_measles_rate_lessTwoStates$year >= 1930 & dat_measles_rate_lessTwoStates$year <= 1939] <- "1930’s"

dat_measles_rate_lessTwoStates$yearBlock[dat_measles_rate_lessTwoStates$year >= 1940 & dat_measles_rate_lessTwoStates$year <= 1949] <- "1940’s"

dat_measles_rate_lessTwoStates$yearBlock[dat_measles_rate_lessTwoStates$year >= 1950 & dat_measles_rate_lessTwoStates$year <= 1959] <- "1950’s"

dat_measles_rate_lessTwoStates$yearBlock[dat_measles_rate_lessTwoStates$year >= 1960 & dat_measles_rate_lessTwoStates$year <= 1969] <- "1960’s"

dat_measles_rate_lessTwoStates$yearBlock[dat_measles_rate_lessTwoStates$year >= 1970 & dat_measles_rate_lessTwoStates$year <= 1979] <- "1970’s"

dat_measles_rate_lessTwoStates$yearBlock[dat_measles_rate_lessTwoStates$year >= 1980 ] <- "1980's onward"
```

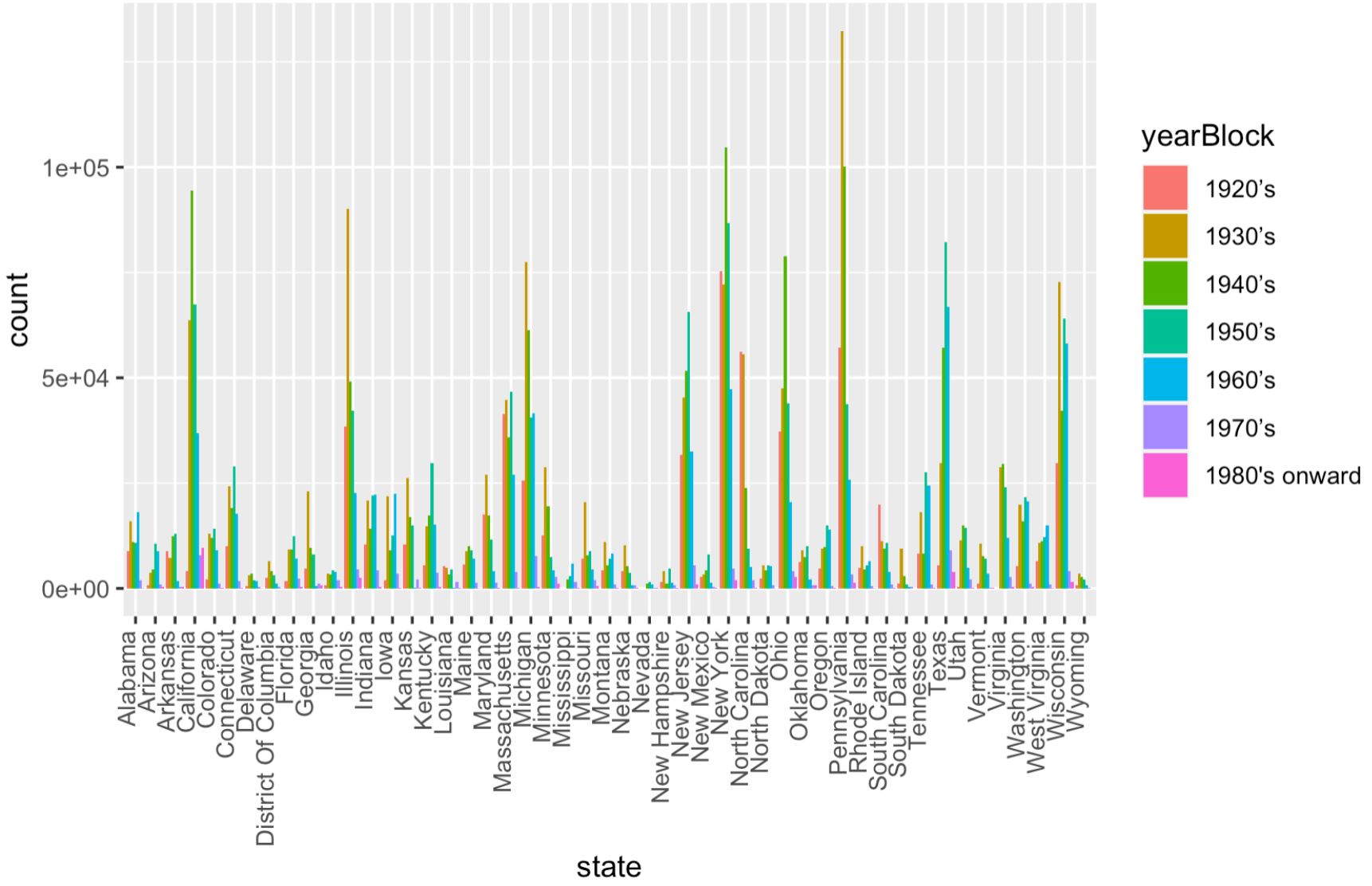
Setup the histogram locks (of cases)

Histograms of the Cases in States by Year

```
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))
```

Plot the Blocks (of cases)

Histograms of the Cases in States by Year



Plot the Blocks (of cases)