Data Analytics
CS301
Relational Data

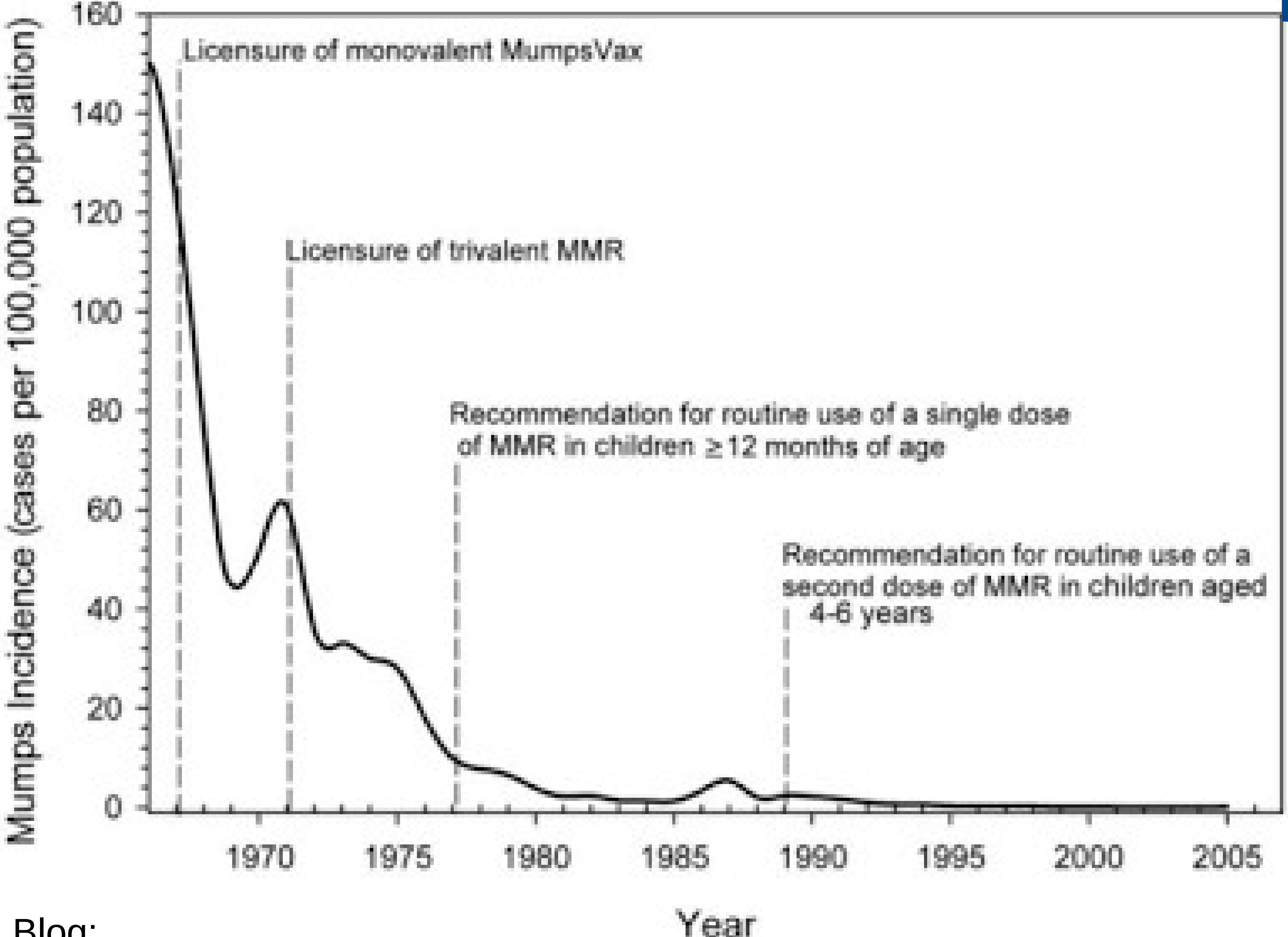
Fall 2018
Oliver Bonham-Carter

Let’s Talk About Lab 4 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://www.historyofvaccines.org/timeline

What Do Others Say About Vaccines?



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

What Do Others Say About Vaccines?

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:

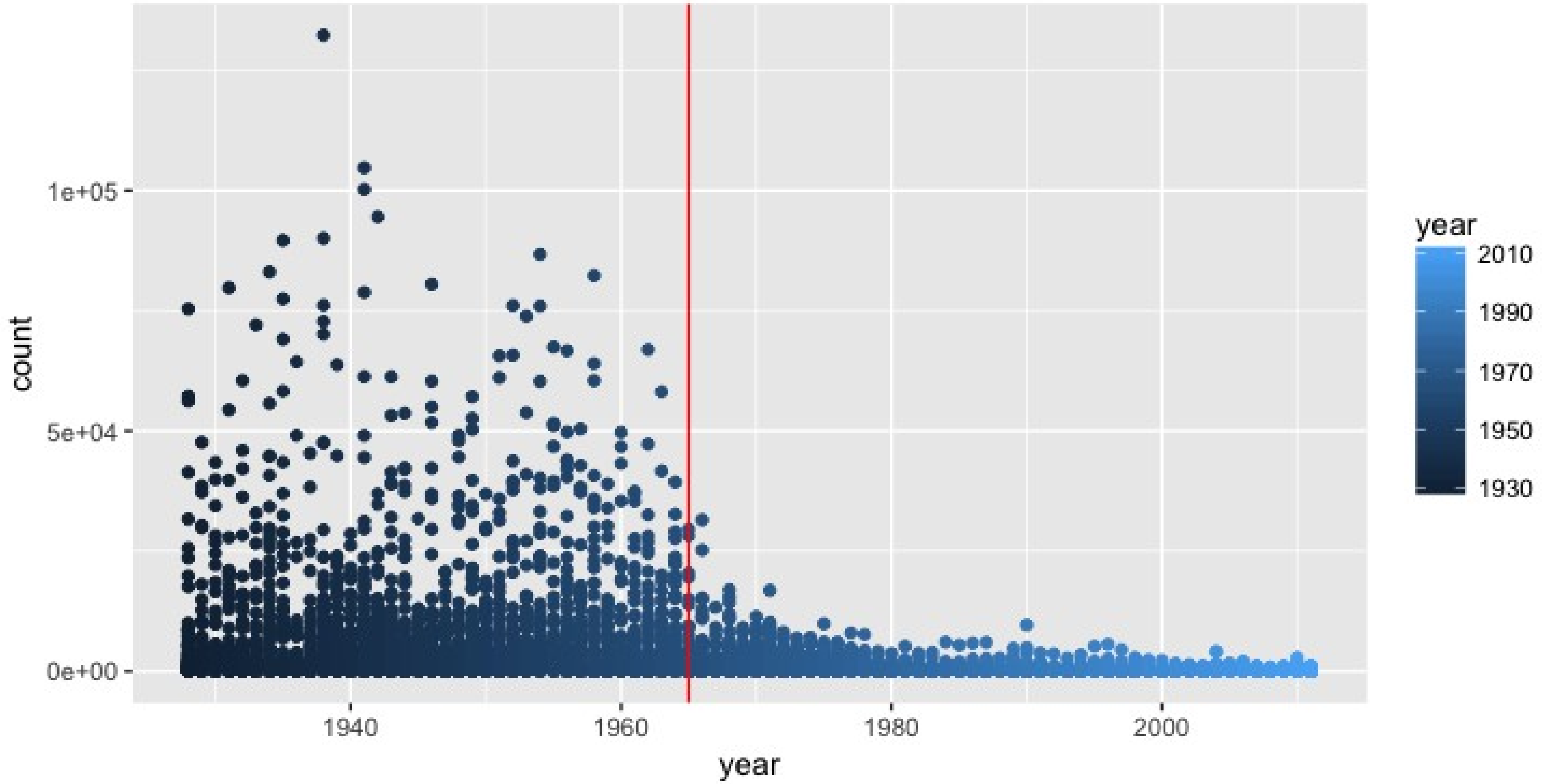
* *JAMA*. 2007;298(18):2155-2163

† 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 * (weeks_reporting/52))

# Note: the rate could be one of several
possible calculations to work with the data.
```

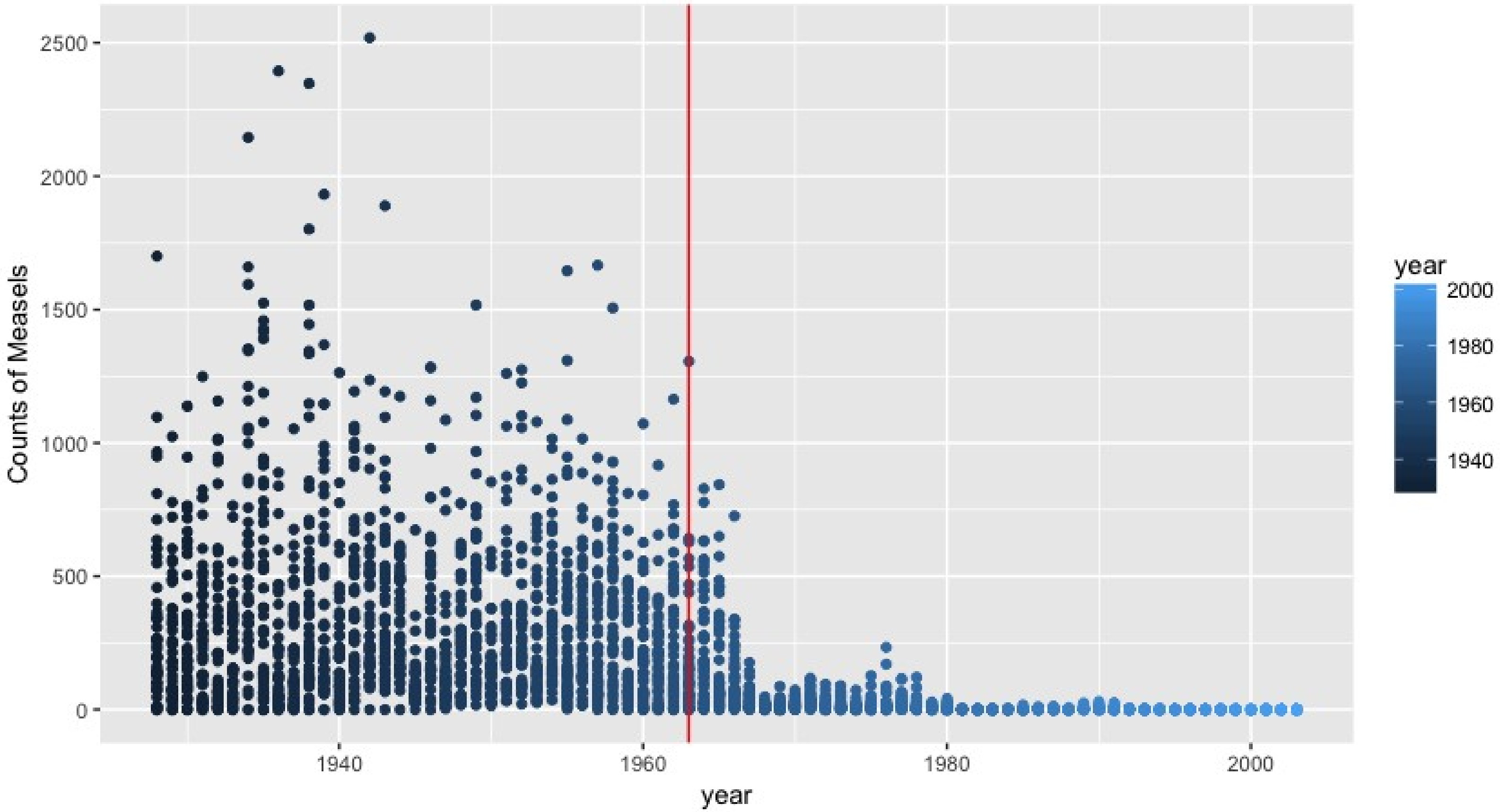
Trim Out Data of Two States: Alaska and Hawaii

```
#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")
```

Code shown on previous slide



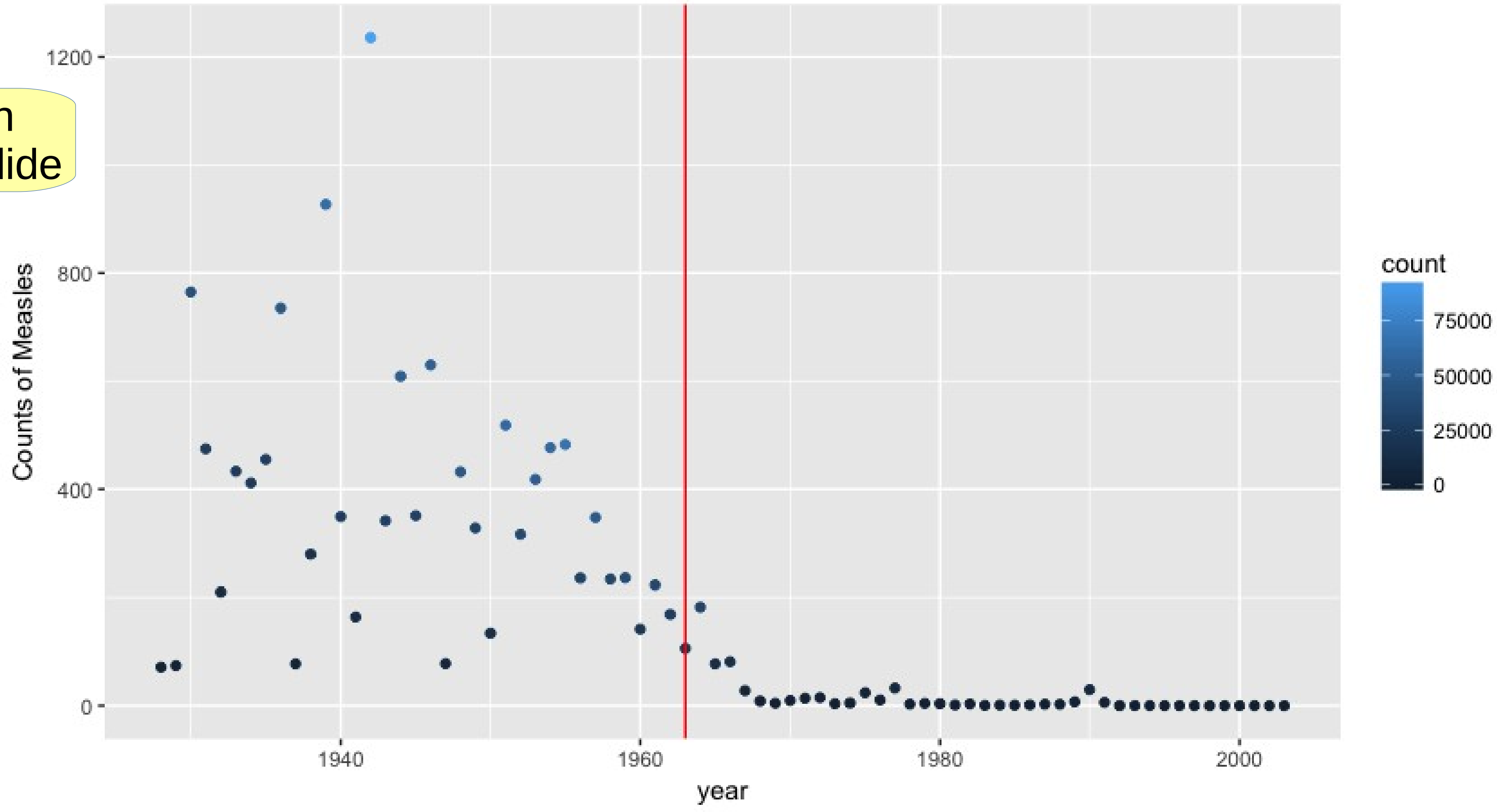
Focus On California

```
# 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")
```

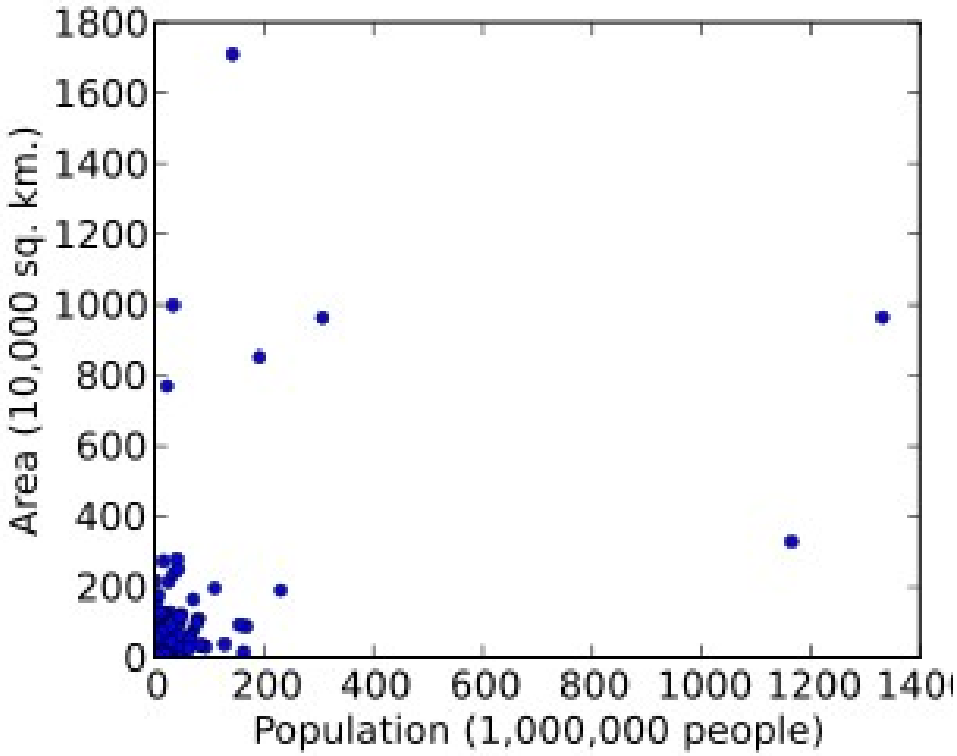
Data From California, Only

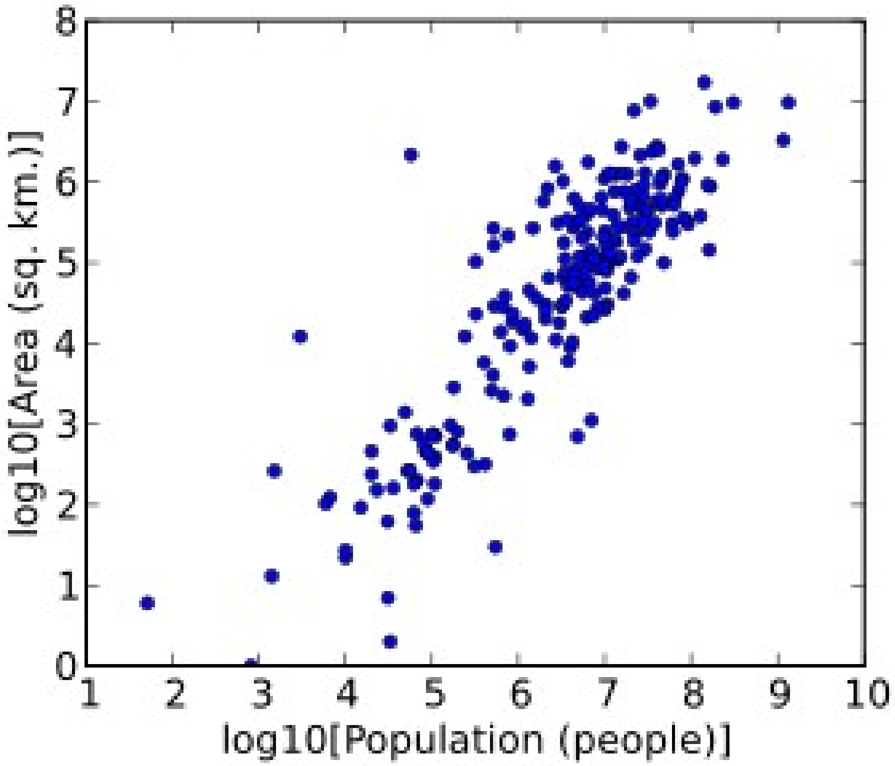
```
ggplot(data = dat_caliFocus, mapping = aes(x = year, y
= rate, color = count)) + geom_point() +
geom_vline(xintercept = 1963, color = "red")
+ labs(y = "Counts of Measles")
```

Code shown
on previous slide



Transformations
Help to Fit the Data





Not transformed

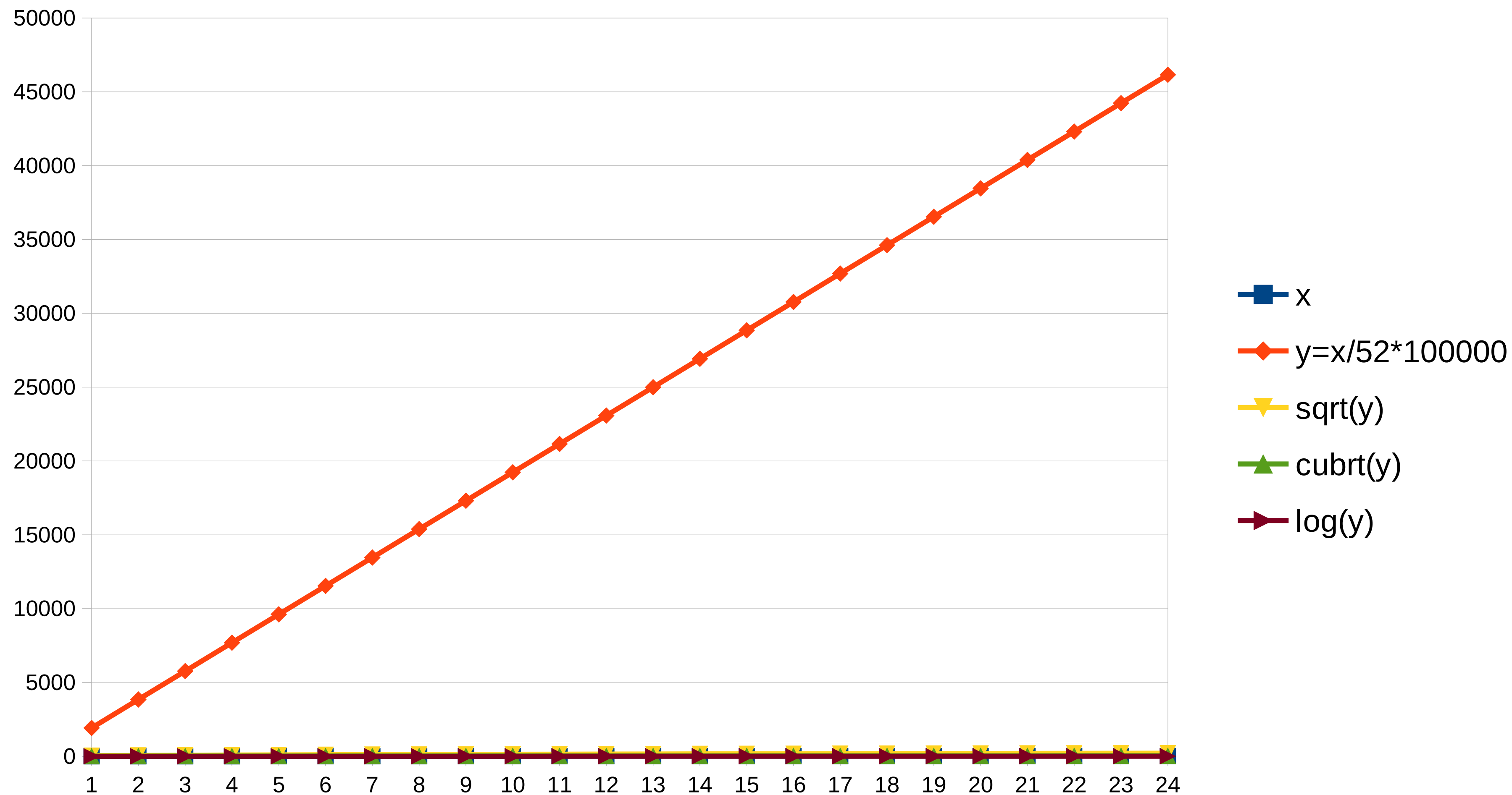
Transformed (using logs)

Transformations Help to Fit the Data

- The square root, ***`x to x^(1/2) = sqrt(x)`***, is a transformation with a moderate effect on distribution shape.
- This approach is weaker than the logarithm and the cube root transformations in its ability to influence the distribution shape.
- Used for reducing right skewness
- Has the advantage that it can be applied to zero values.
- Commonly applied to counted data, especially if the values are mostly rather small

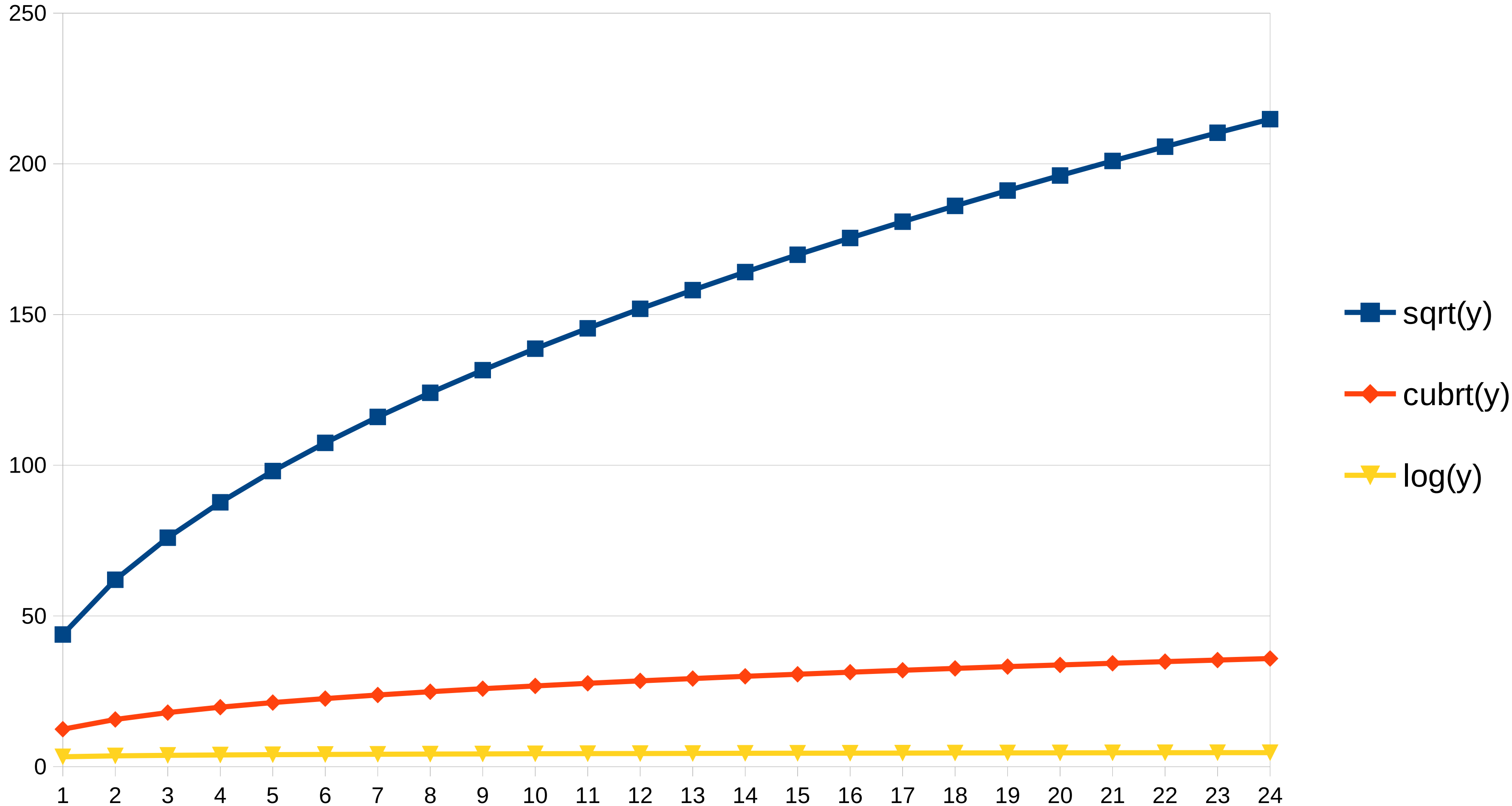
http://fmwww.bc.edu/repec/bocode/t/transint.html

Effects of Transformations on Values



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
5	9615.384615	98.05806757	21.26451851	3.982966661
6	11538.46154	107.4172311	22.59692282	4.062147907

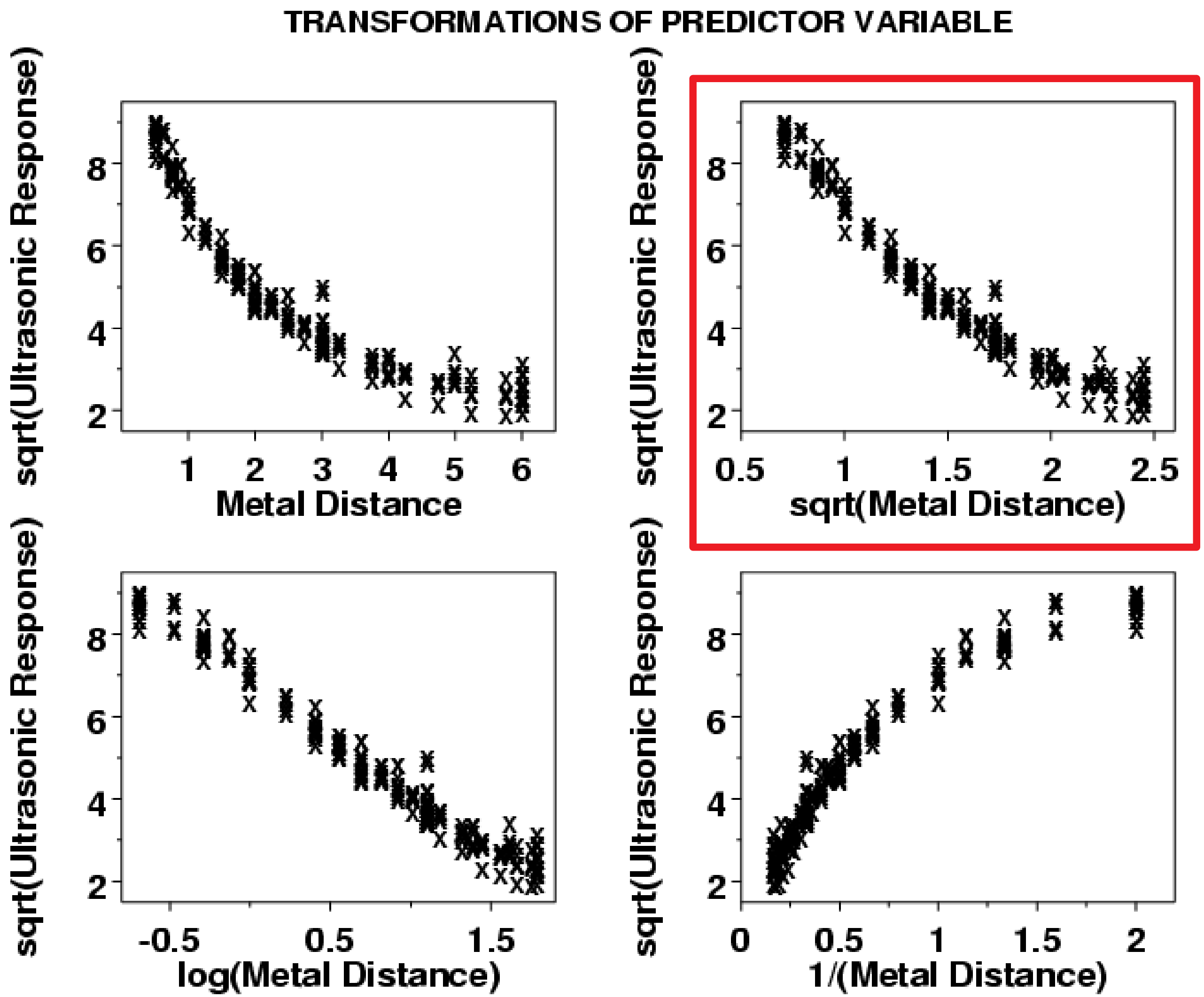
Effects of Transformations on Values Zoom-in



Transformations Help to Fit the Data

- 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))

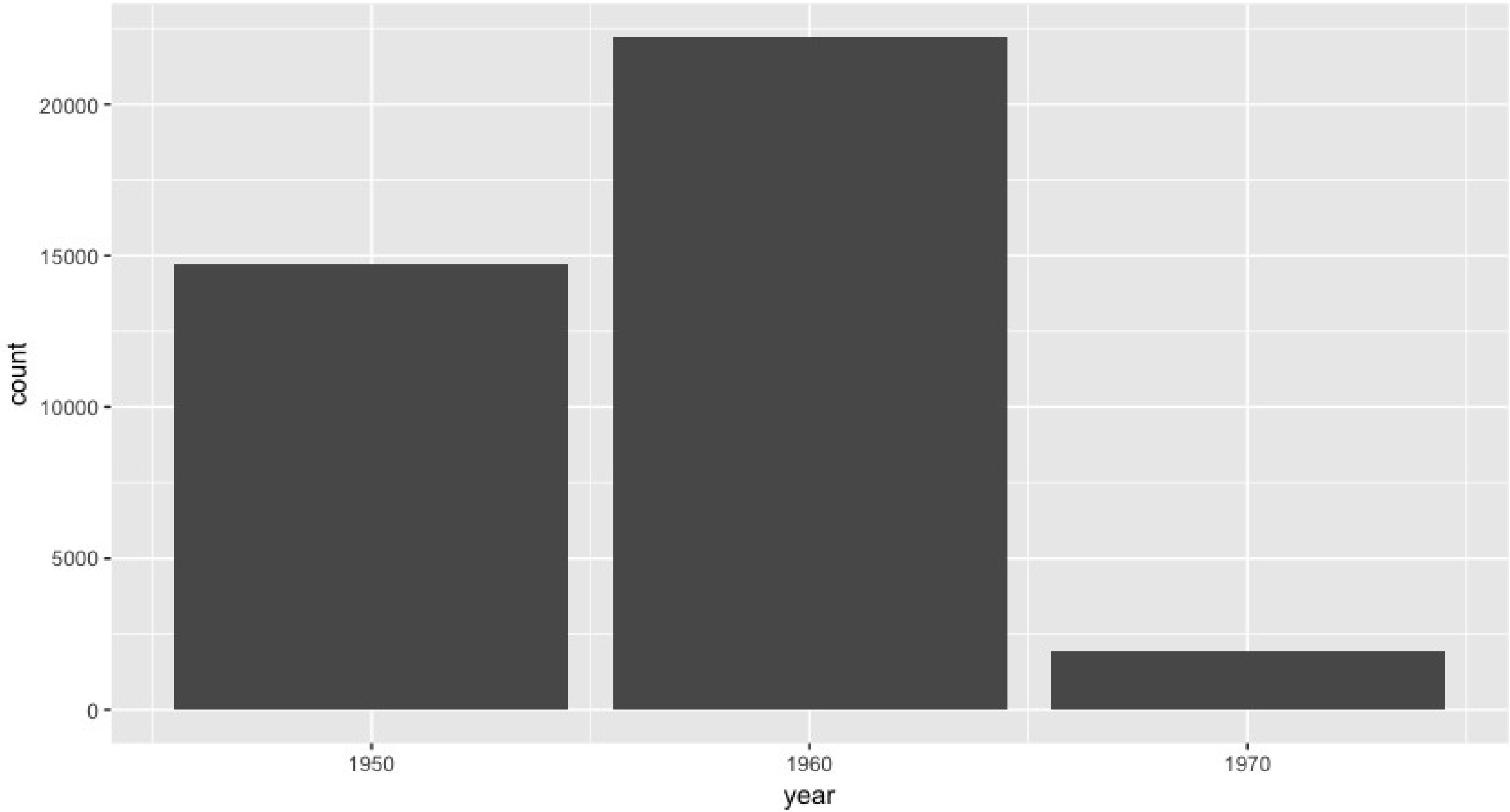


The 1950’s, 1960’s and 1970’s
Without Transformation

```
#plot three bars to see 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 the vaccines lab...

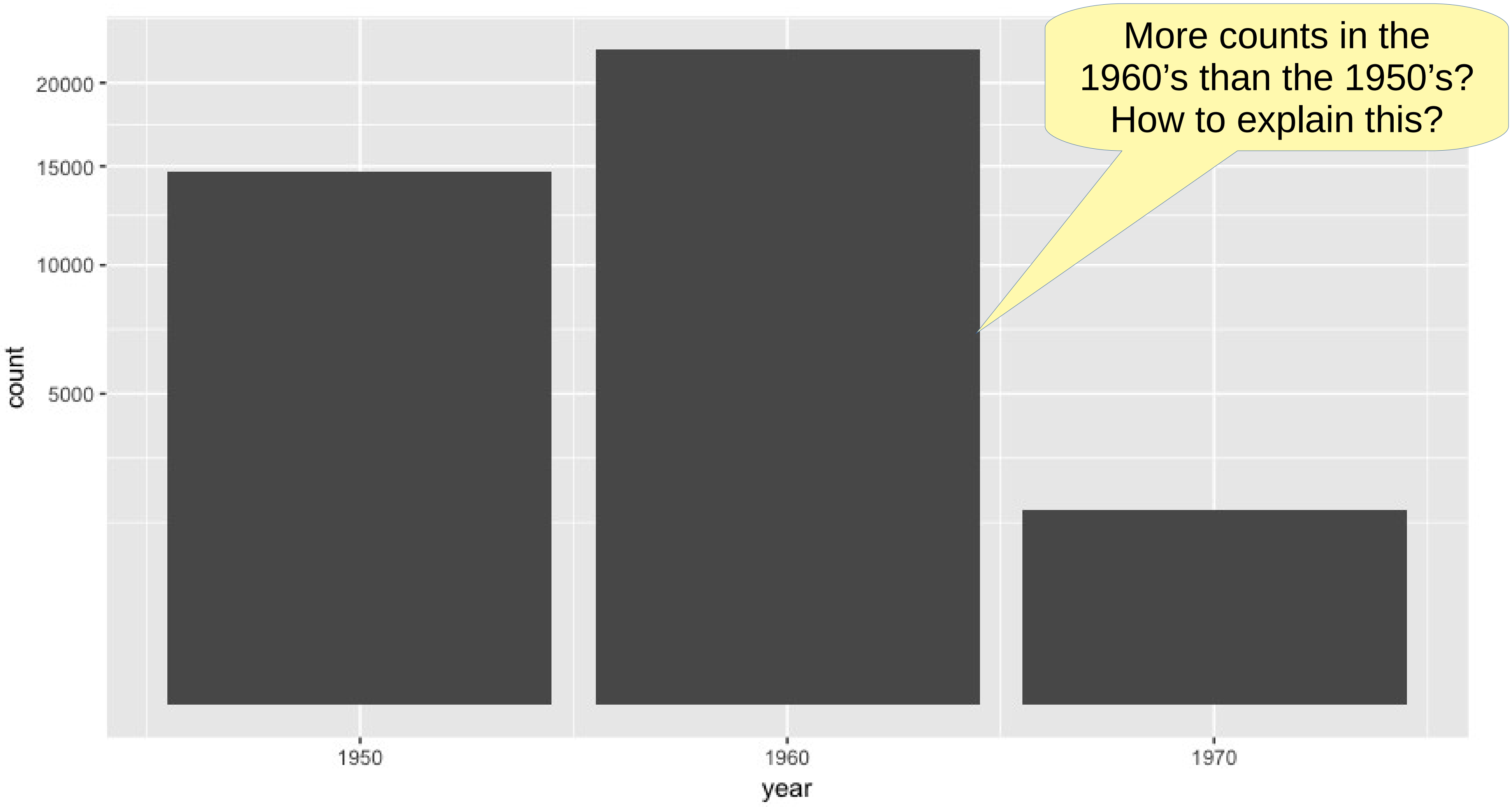
The 1950's, 1960's and 1970's Without Transformation



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



Urban Versus Rural

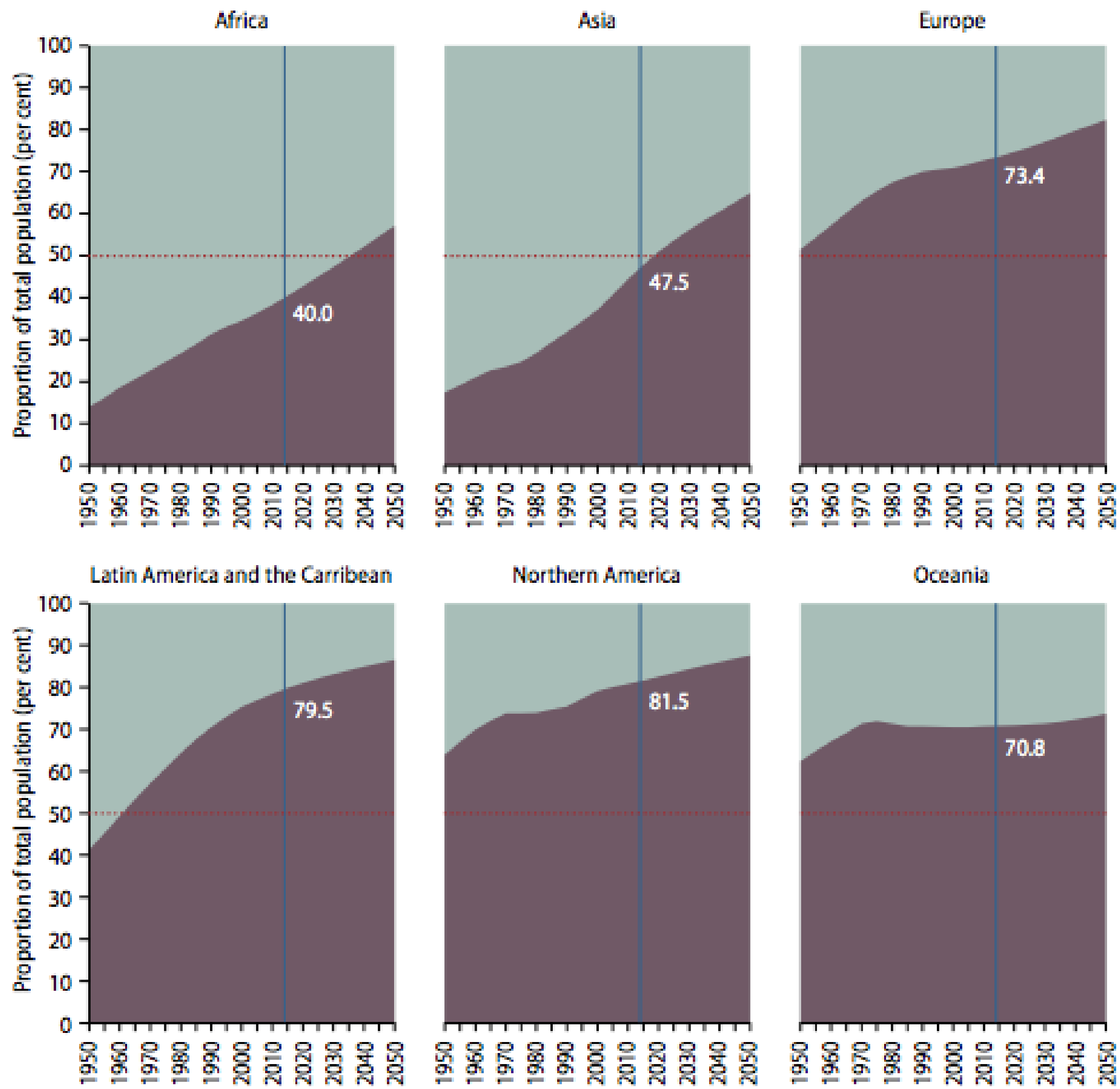
A possible Explanation for the 1950's

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

- **Urban**: City dwelling
- **Rural**: Country dwelling
- Vaccinations in Rural Areas:
  - Were there fewer people available in from whom to contract viruses?
  - Less opportunity to see others in country?
- Country areas: you are breathing your neighbour's breath.

Figure 3.

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



Urban population

Rural population

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

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

```
library(tidyverse)

library(dslabs)

library(dplyr)

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

# Filter out all data except in the years 1950, 1960, and 1970

dat_measles_rate_lessTwoStates <- dat %>% filter(year == 1950 | year == 1960 | year == 1970)

#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

```
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))
```