Data Analytics
CS390
The Vaccine Lab

Week 12: 9 November 2021

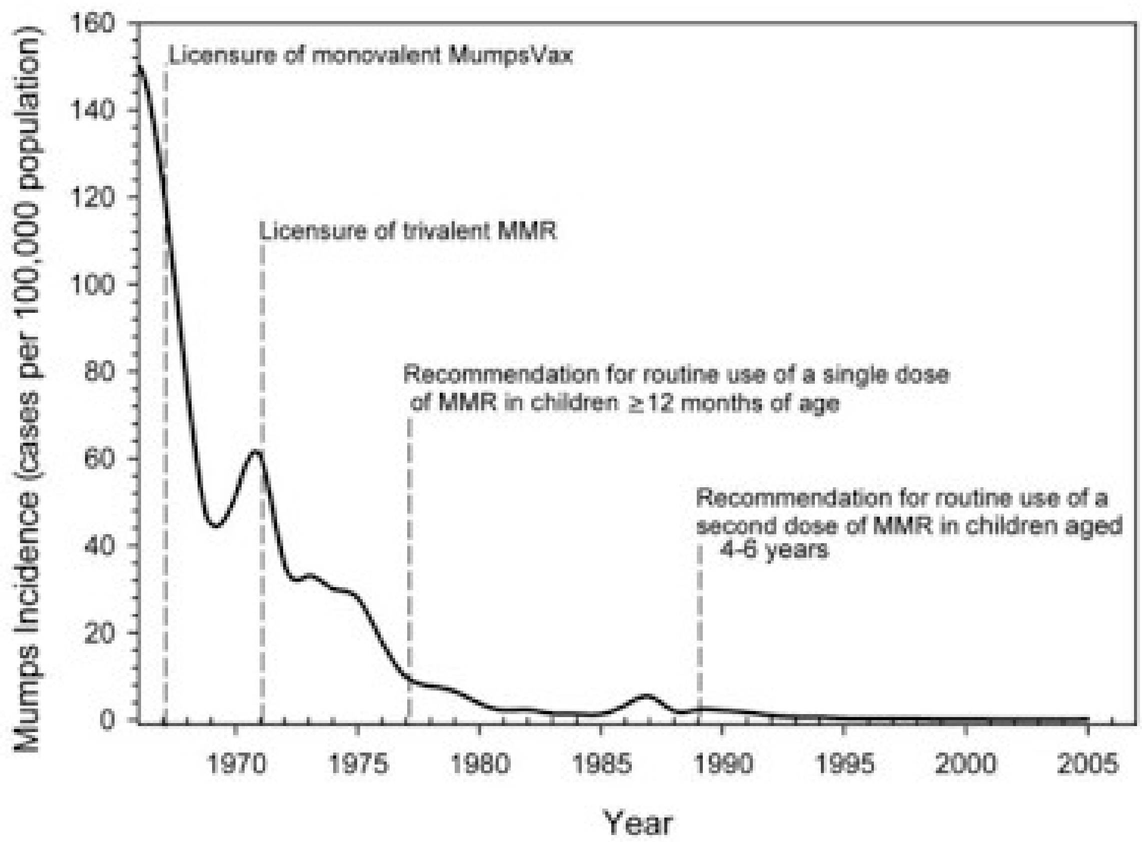
Oliver BONHAM-CARTER

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:

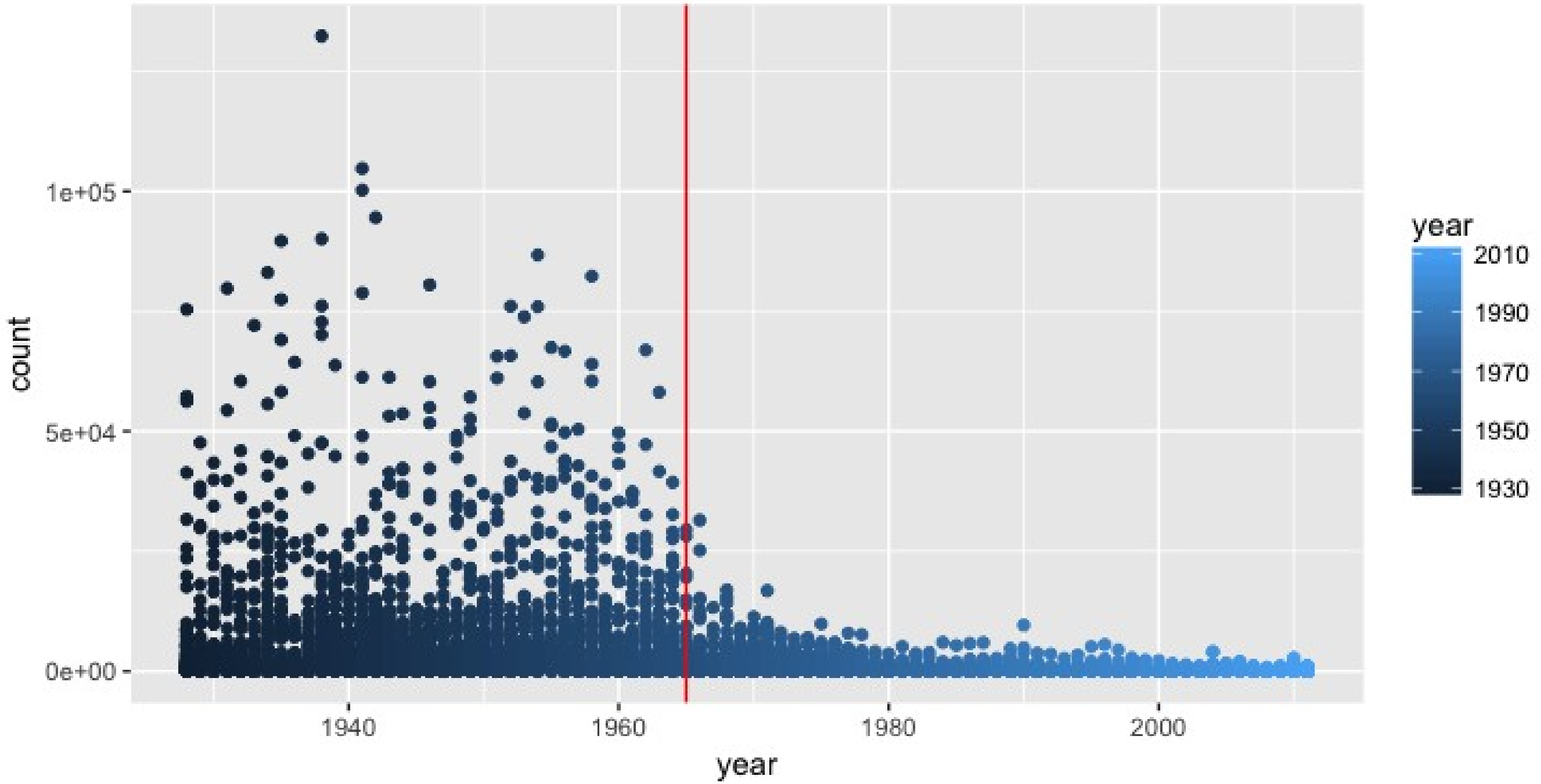
* *JAMA*. 2007;298(18):2155-2163

† 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) / (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")

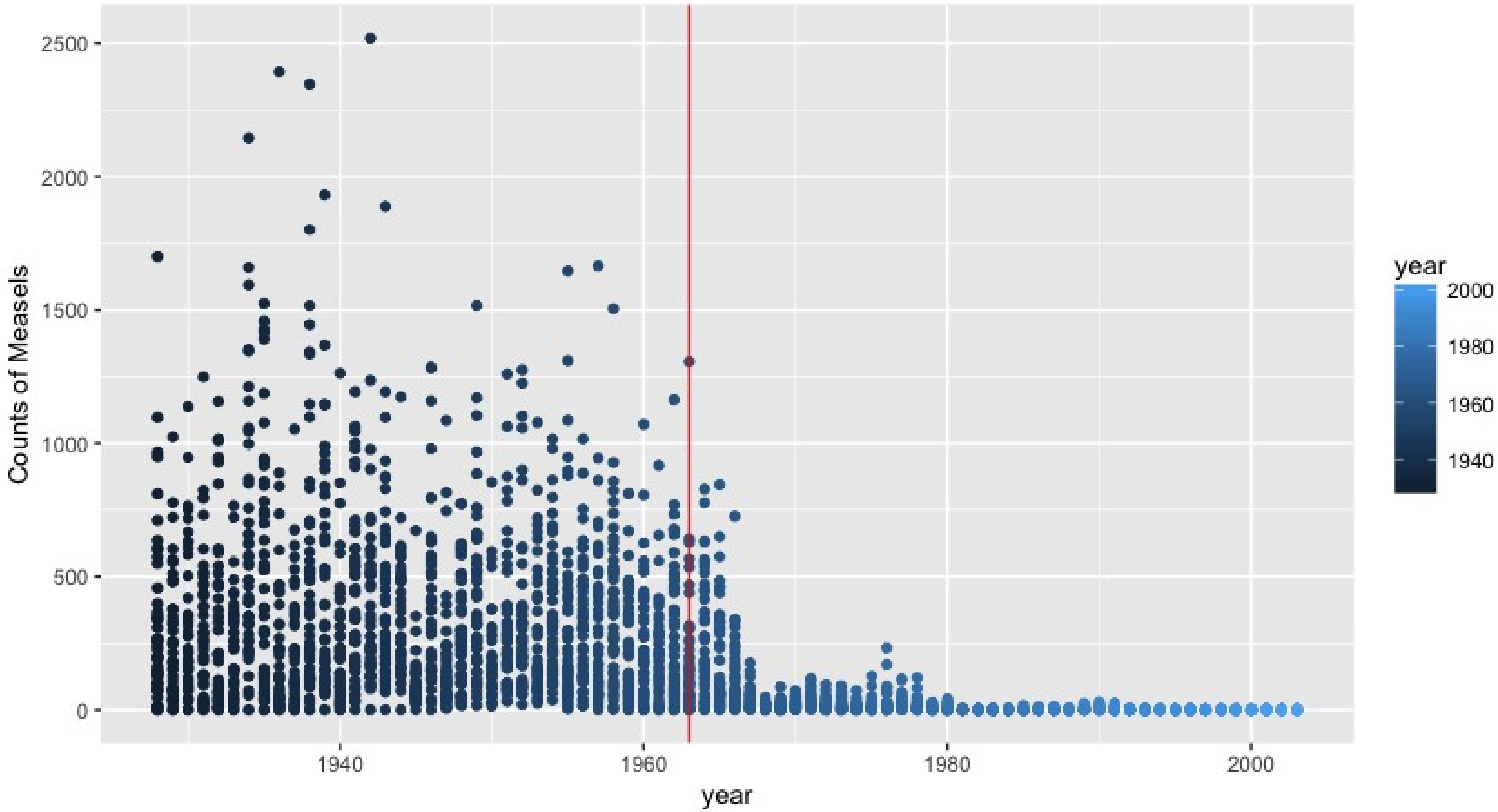
  View(dat_measles_rate_lessTwoStates)

- # 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

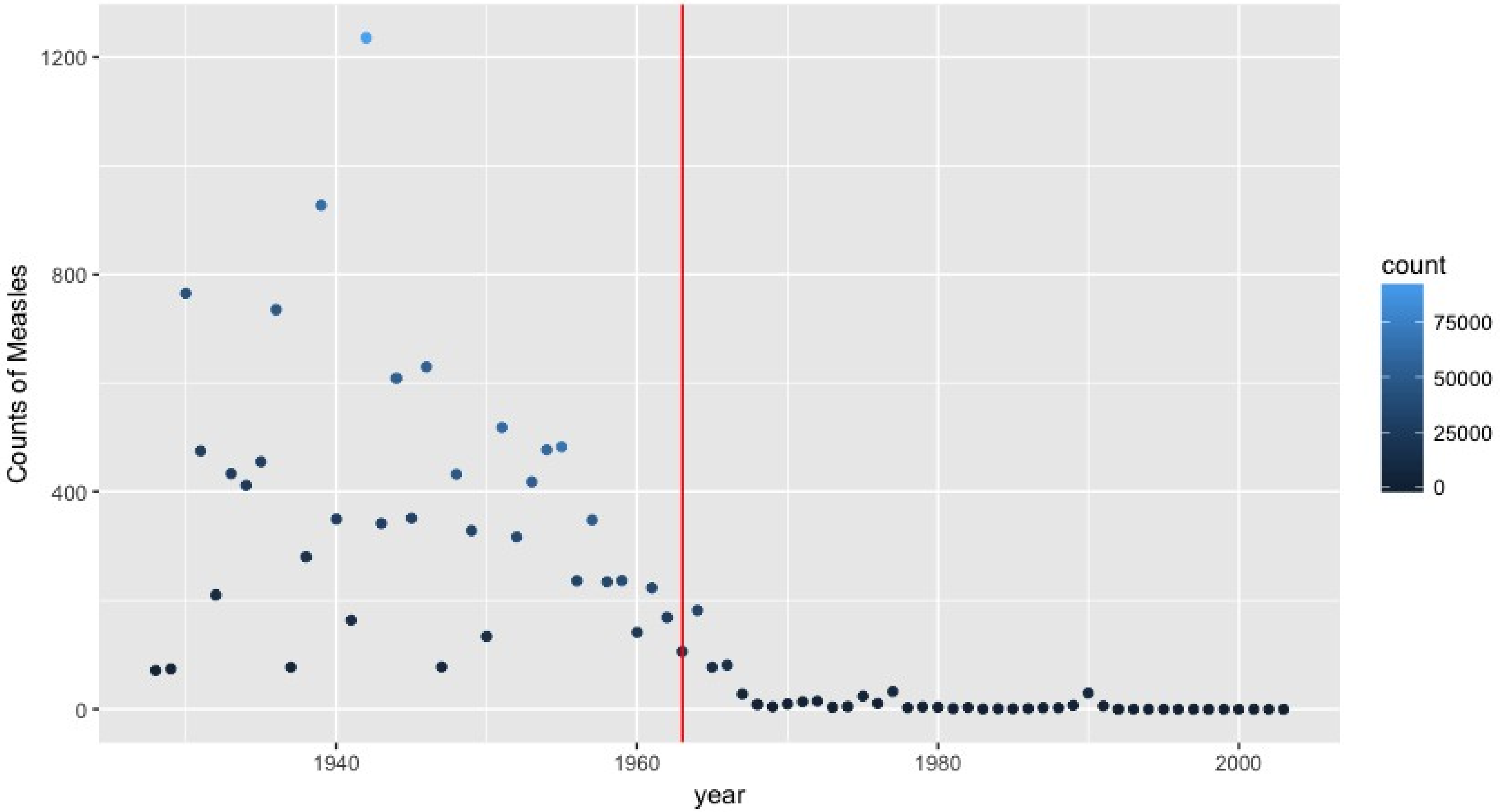
  dat_caliFocus <- filter(dat_measles_rate_lessTwoStates, state == "California")

  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

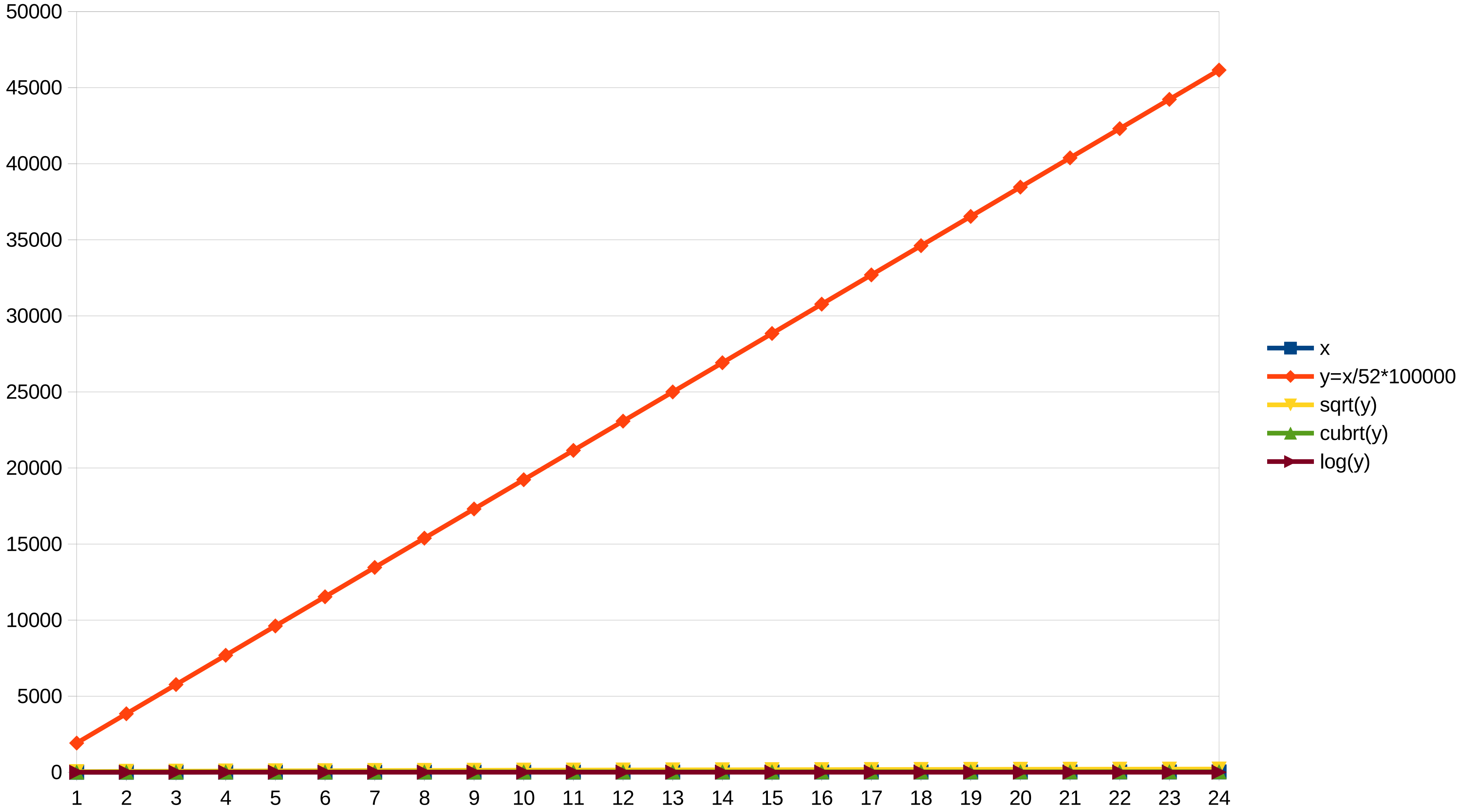
- ggplot(data = dat_caliFocus, mapping = aes(x = year, y = rate, color = count)) + geom_point() + geom_vline(xintercept = 1963, color = "red") + labs(y = "Counts of Measles")



Transformations Help to Fit the Data

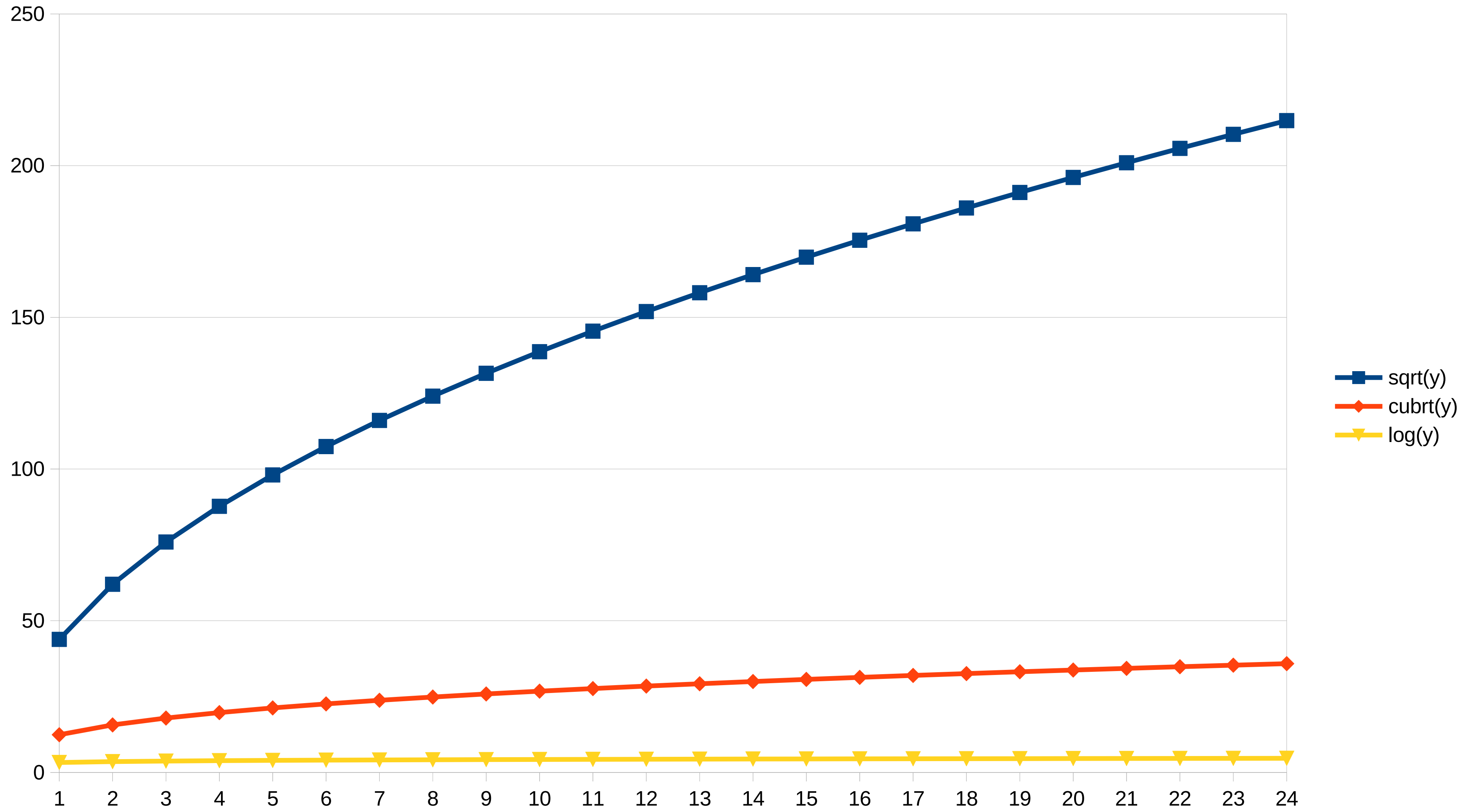
- The square root, x to x^(1/2) = sqrt(x), is a transformation with a moderate effect on distribution shape.
- Weaker than the logarithm and the cube root transformations
- Used for reducing right skewness
- Has the advantage that it can be applied to zero values

Effects of Transformations on Vars



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

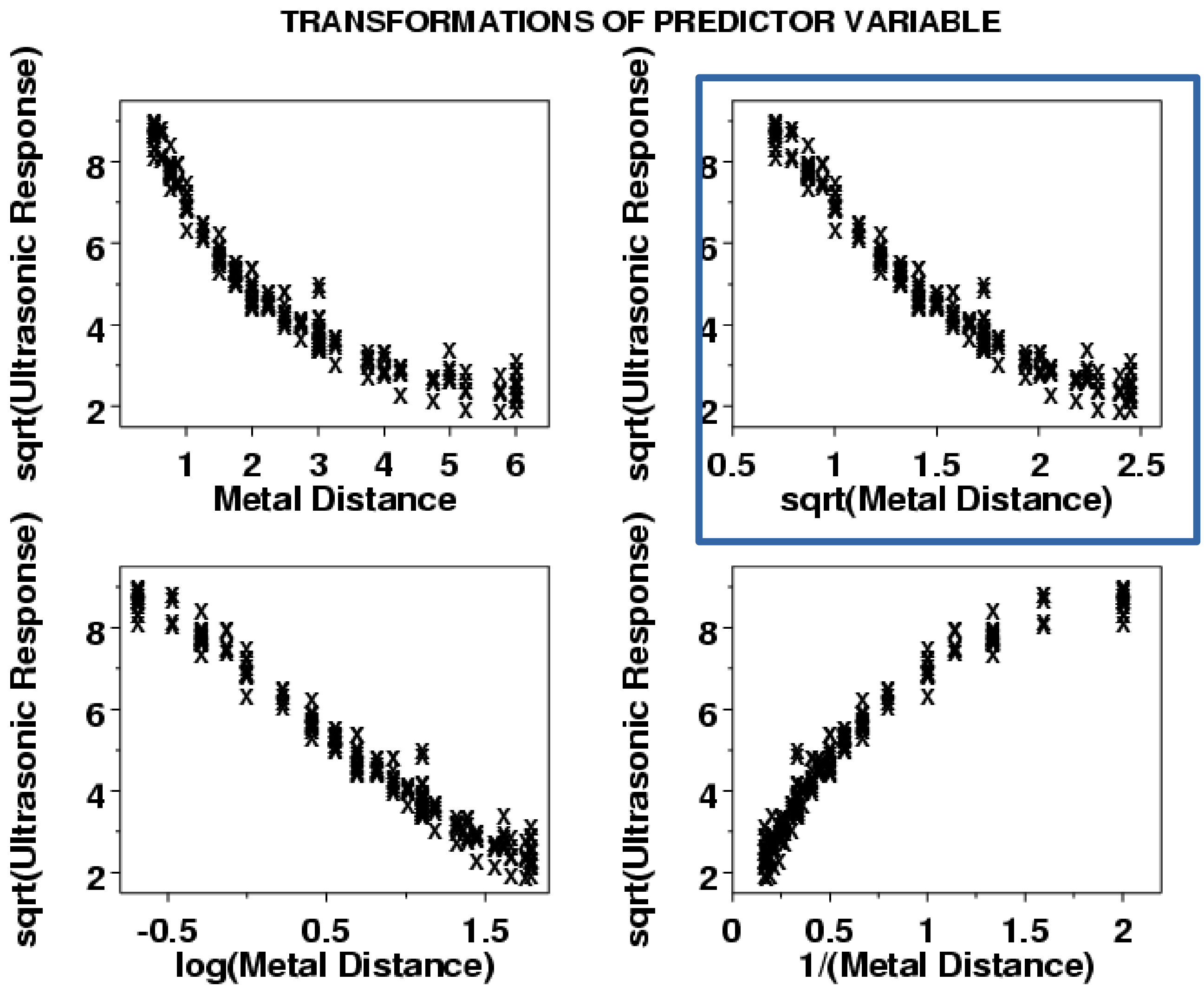
Effects of Transformations on Vars Zoom-in



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))

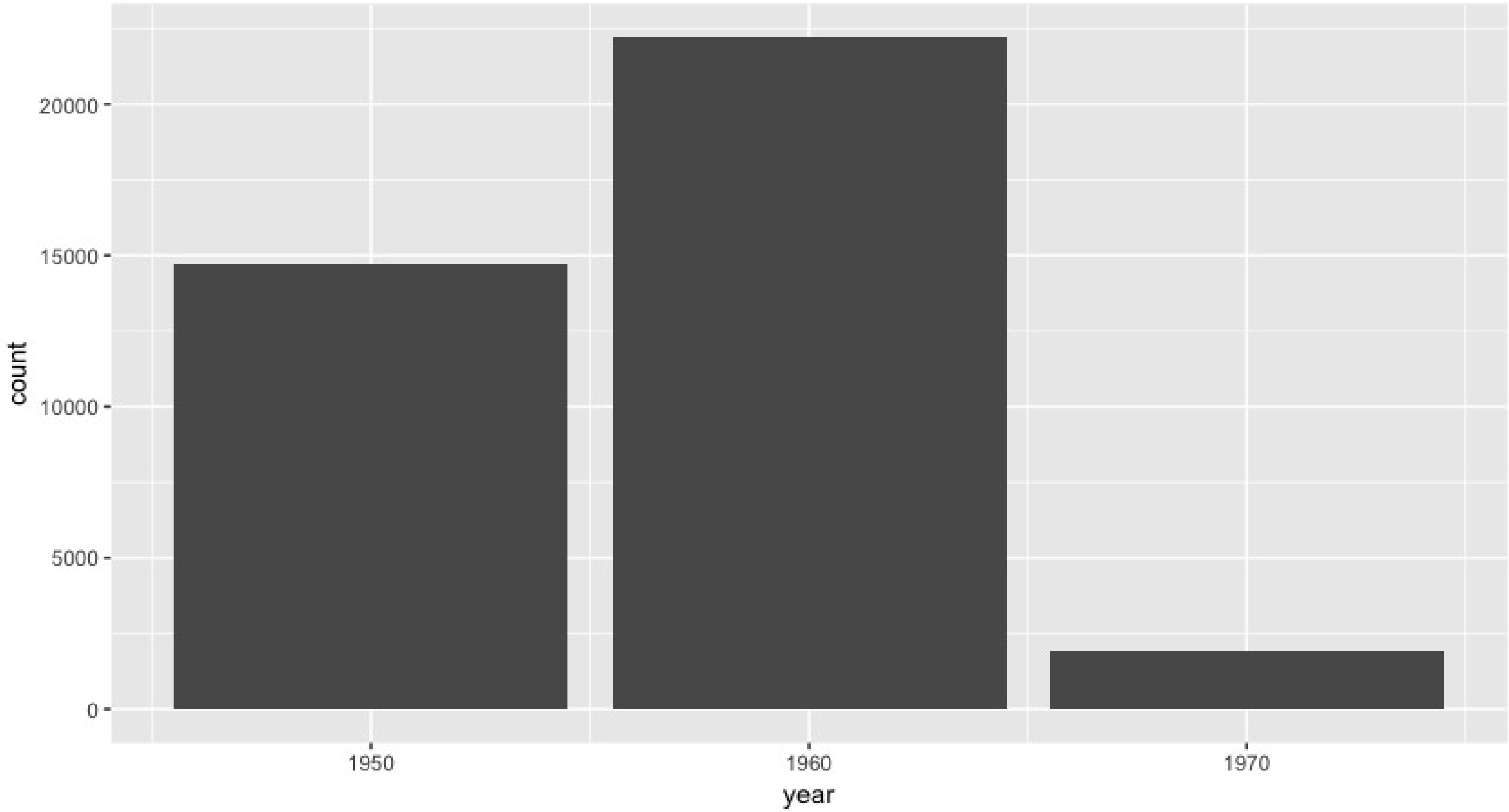


The 1950’s, 1960’s and 1970’s Without Transformation

- #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")

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

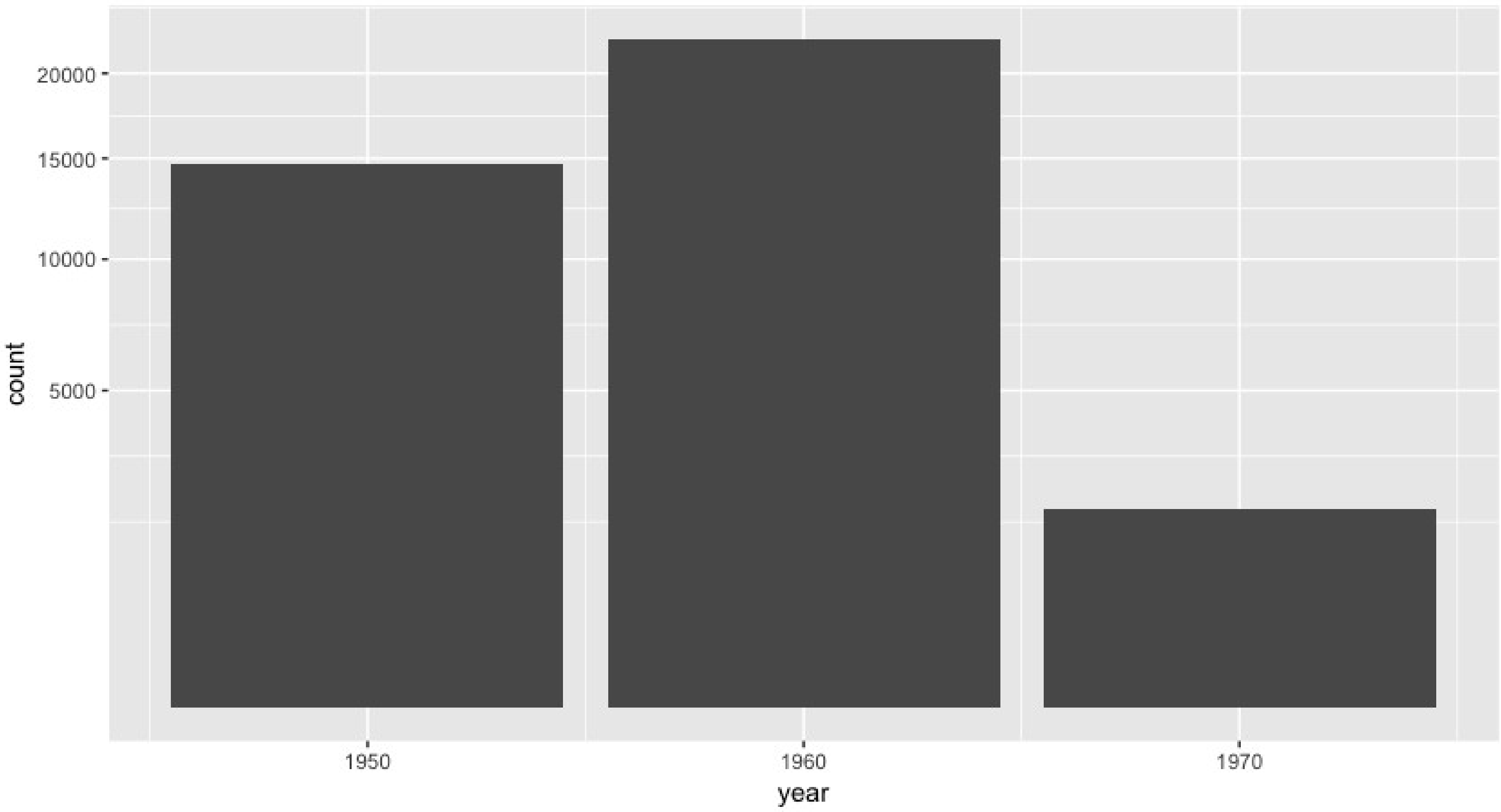


The 1950's, 1960's and 1970's With Sqrt Transformation

- #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

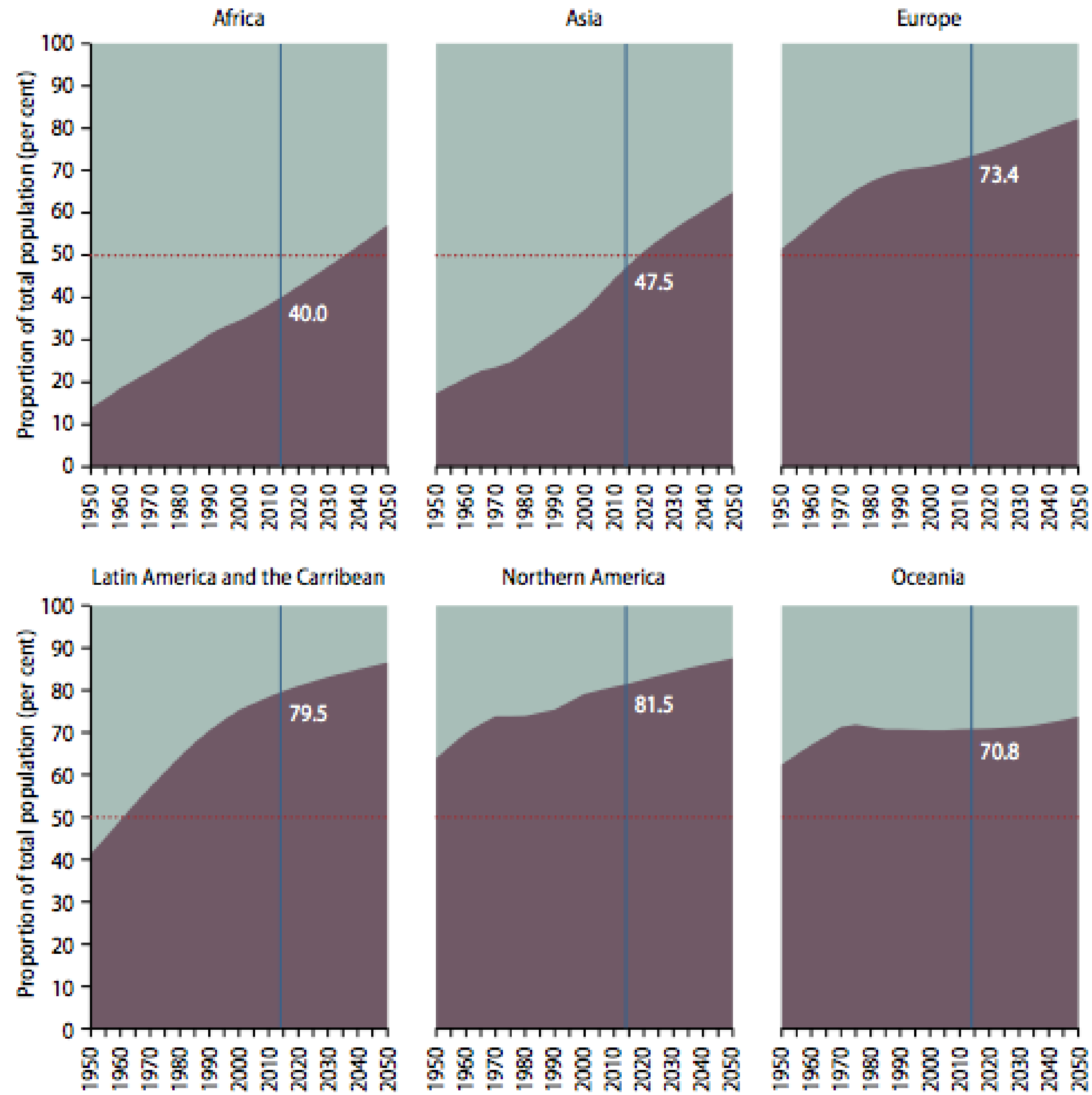


Urban Versus Rural

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**: 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

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

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

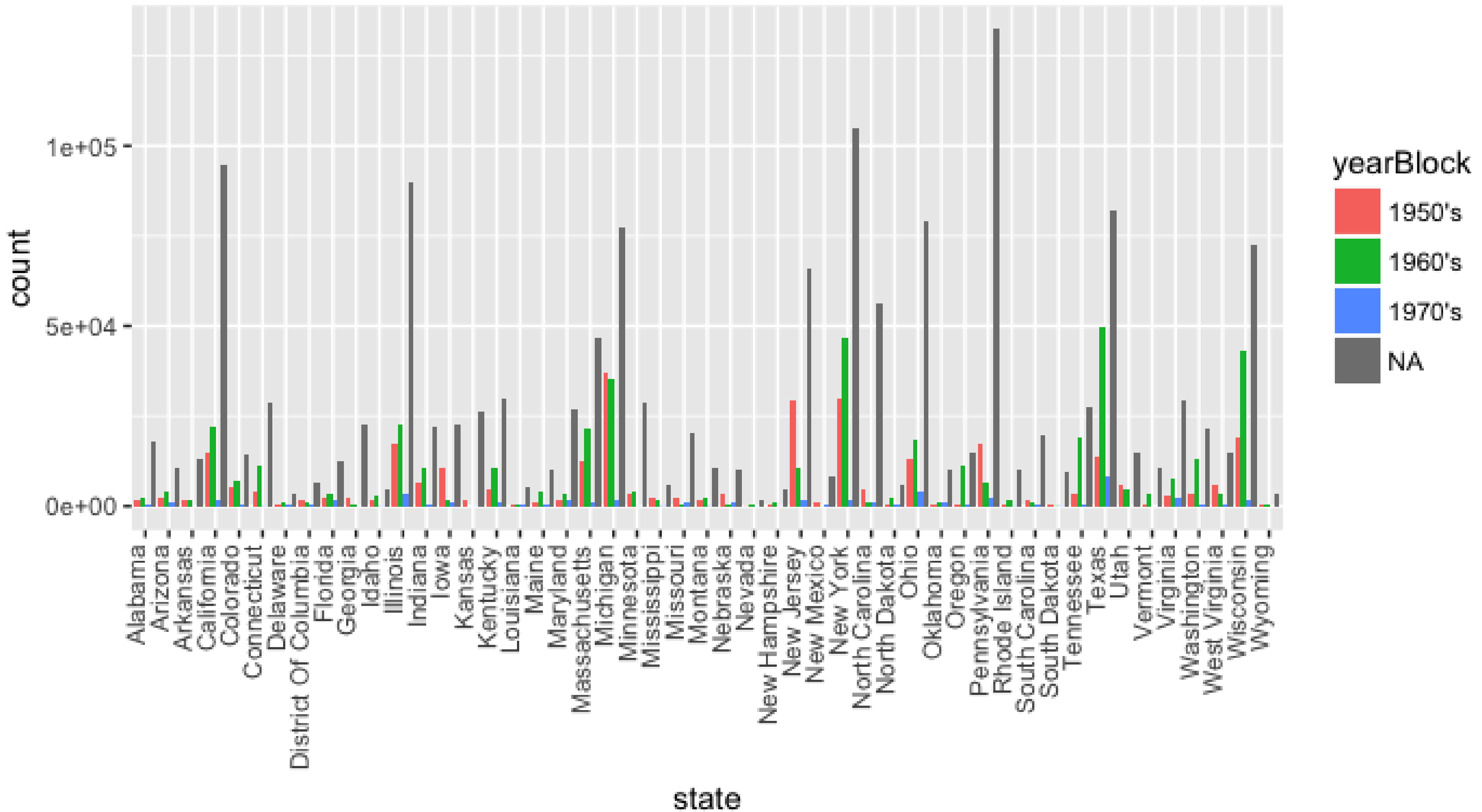
```
dat_measles_rate_lessTwoStates$yearBlock[dat_measles_rate_lessTwoStates$year == 1950] <- "1950’s"

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,

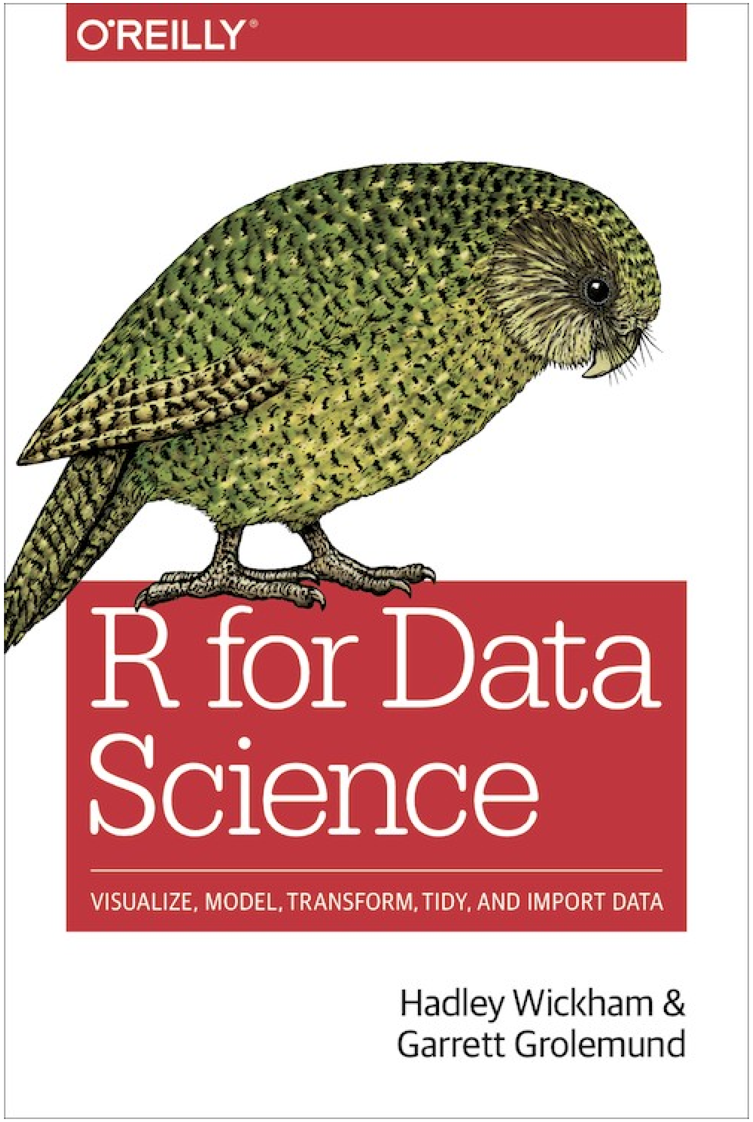
```
ggplot(data = dat_measles_rate_lessTwoStates) + geom_bar(mapping = aes(x = state, y = count, fill = yearBlock), position = "dodge", stat = "identity") + theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust=-0.01))
```

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

- 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))



Where in the Web?
Where in the Book?



- Note the chapter differences!
- Book:
  - $\boxed{\text{Chap 10}}$
- Web:
  - $\boxed{\text{Chap 13}}$
- Relational Data

Relational Databases

- A database table is similar to those that we have been using already.

attributes (or columns)

ID	*name*	*dept_name*	*salary*
10101	Srinivasan	Comp. Sci.	65000
12121	Wu	Finance	90000
15151	Mozart	Music	40000
22222	Einstein	Physics	95000
32343	El Said	History	60000
33456	Gold	Physics	87000
45565	Katz	Comp. Sci.	75000
58583	Califieri	History	62000
76543	Singh	Finance	80000
76766	Crick	Biology	72000
83821	Brandt	Comp. Sci.	92000
98345	Kim	Elec. Eng.	80000

tuples (or rows)

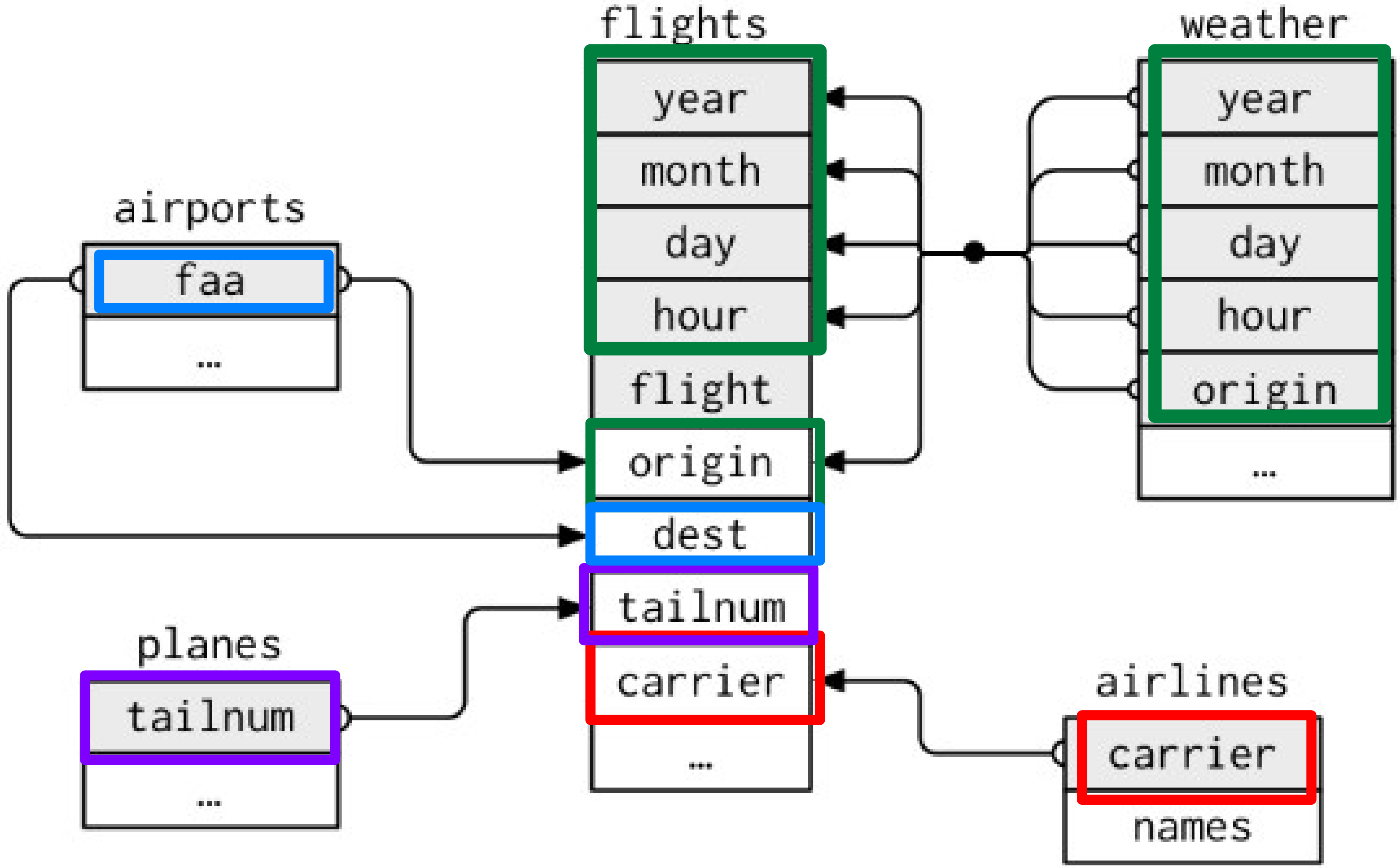
Let’s Look at Some Tables

- library(tidyverse)
- library(nycflights13)
- #show built-in tables

```
View(airlines)
View(airports)
View(planes)
View(weather)
```

Relational Databases

- The data of these built-in tables is “connected” in the sty



Relational Databases

- **Primary Keys:** Unique identifier for each row of the table.
  - Ex: planes$tailnum
- **Foreign Keys:** Unique identifier for row in another table.
  - Ex: flights$tailnum
  - Is a foreign key since it exists in the flights table and matches a flight to a unique plane.

Plural or Singular Elements?

```
# If something is unique: there is only one of it. Here each tailnum entry is unique

planes %>% count(tailnum)

# Try setting up a test to see if there are any more than one of an entry (necessary to be a primary key)

planes %>% count(tailnum) %>% filter(n > 1)

# A key could be a combination of things

weather %>% count(year, month, day, hour, origin) %>% filter(n > 1)

flights %>% count(year, month, day, flight) %>% filter(n > 1)
```

Note: the *n* > 1 (a tally) comes from the count() function

Keys From
Singular Elements

- **Baby-name Data**
- First: **install.packages("babynames")**
- **library(babynames) and tidyverse too!**
- Then find the primary keys in,

  babynames:babynames

- **Baseball data:**
- First: **install.packages("Lahman")**
- **library(Lahman)**
- Then find the primary keys in,

  Lahman::Batting

Possible Solutions: Find Some Keys!

- **Baby-name Data**

```
babynames::babynames %>% count(name, year, sex) %>% filter(n >1)
```

- **Baseball data:**

```
Lahman::Batting %>% group_by(playerID, yearID, stint) %>% filter(n() > 1) %>%  nrow()
```

Simple Analysis with BabyNames

```
library(babynames)
library(tidyverse)
# Try this: combine and count common name, year, sex details,
how many are there for each name?
babynames::babynames %>% count(name, year, sex) %>%
filter(n >1)


bn <- babynames::babynames %>% select(name)
# find names beginning with O
bn[grep("^O", bn$name),]

# how many Olivers are there?
count(bn[grep("^Oliver", bn$name),])
```