Victoria Larson: Sprint Review

Analyzing, Understanding, & Summarizing Disease Data

devleague

3 data sets
20 minutes
3 questions

An ambitious goal…

Disease data!

Subsets.. Subsets… Subsets…

Question 1: What was the total Measles Count for Hawaii vs. the United States in 1960.

1a. Total Measles Count for Hawaii

Measles for all in 1960

```
Measles_1960 <- subset(Year_1960,disease=="Measles")
```

```
MeaslesHI_1960 <- subset(Measles_1960,state=="Hawaii")
HIMeaslesCount <-MeaslesHI_1960$count
```

Answer: 5322

1b. Measles average (52 weeks) for entire USA in 1960

```
USMeasles_avg <- mean(Measles_1960$count)
```

Answer 8474.922

Question 2: Which state had the highest count (in 52 weeks) for each disease.

The First Trial… and error

```
#filtering out just Small pox  from the weeks reporting 52
Smlpx52wks <- subset(Week52_reporting, disease=="Smallpox")
#doing which.max tells you which line the code is on.
which.max(Smlpx52wks$count)
#ANSWER!
SmallpoxHigh <- Smlpx52wks[163,]
SmallpoxHigh
```

```
> SmallpoxHigh
       disease   state year weeks_reporting count population
17911 Smallpox Indiana 1930              52  5239    3238503
```

Then Ben found me.

Diseases %>% subset(weeks_reporting == 52) %>% group_by(disease) %>% summarise(max(count))

```
# A tibble: 7 x 2
  disease      `max(count)`
  <fct>               <dbl>
1 Hepatitis A         10821
2 Measles            132342
3 Mumps                9867
4 Pertussis           22013
5 Polio               22013
6 Rubella              8384
7 Smallpox             5239
```

The ddply way:

```
DiseaseHigh <- ddply(Week52_reporting, 'disease', function(x) x[x$count==max(x$count),])
```

	disease	state	year	weeks_reporting	count	population
1	Hepatitis A	California	1968	52	10821	19219725
2	Measles	Pennsylvania	1938	52	132342	9851738
3	Mumps	Michigan	1975	52	9867	9156979
4	Pertussis	New York	1939	52	22013	13406915
5	Polio	New York	1939	52	22013	13406915
6	Rubella	California	1971	52	8384	20300216
7	Smallpox	Indiana	1930	52	5239	3238503

Question 2B: Compare the Year/State count to the Year/USA Count. What is the individual states overall percentage in relation to the nation.

```
total <- merge(DiseaseHigh, Week52_reporting, by=c("disease", "year"))
```

```
> total
```

	disease	year	state.x	weeks_reporting.x	count.x	population.x	state.y	weeks_reporting.y	count.y	population.y
1	Hepatitis A	1968	California	52	10821	19219725	Tennessee	52	1034	3821245
2	Hepatitis A	1968	California	52	10821	19219725	Iowa	52	492	2806962
3	Hepatitis A	1968	California	52	10821	19219725	Connecticut	52	502	2964628
4	Hepatitis A	1968	California	52	10821	19219725	Missouri	52	852	4613078
5	Hepatitis A	1968	California	52	10821	19219725	California	52	10821	19219725
6	Hepatitis A	1968	California	52	10821	19219725	Mississippi	52	458	2186210
7	Hepatitis A	1968	California	52	10821	19219725	Kansas	52	410	2237660
8	Hepatitis A	1968	California	52	10821	19219725	Michigan	52	2185	8714199
9	Hepatitis A	1968	California	52	10821	19219725	Utah	52	259	1012163
10	Hepatitis A	1968	California	52	10821	19219725	Virginia	52	577	4510316
11	Hepatitis A	1968	California	52	10821	19219725	New York	52	3728	18128492
12	Hepatitis A	1968	California	52	10821	19219725	Massachusetts	52	1023	5613418

Behold….

Disease_percent = total %>% group_by(disease) %>% summarise(first(count.x), sum(count.y), first(state.x), first(year)) %>% mutate(percent.of.disease=`first(count.x)`/`sum(count.y)`*100)

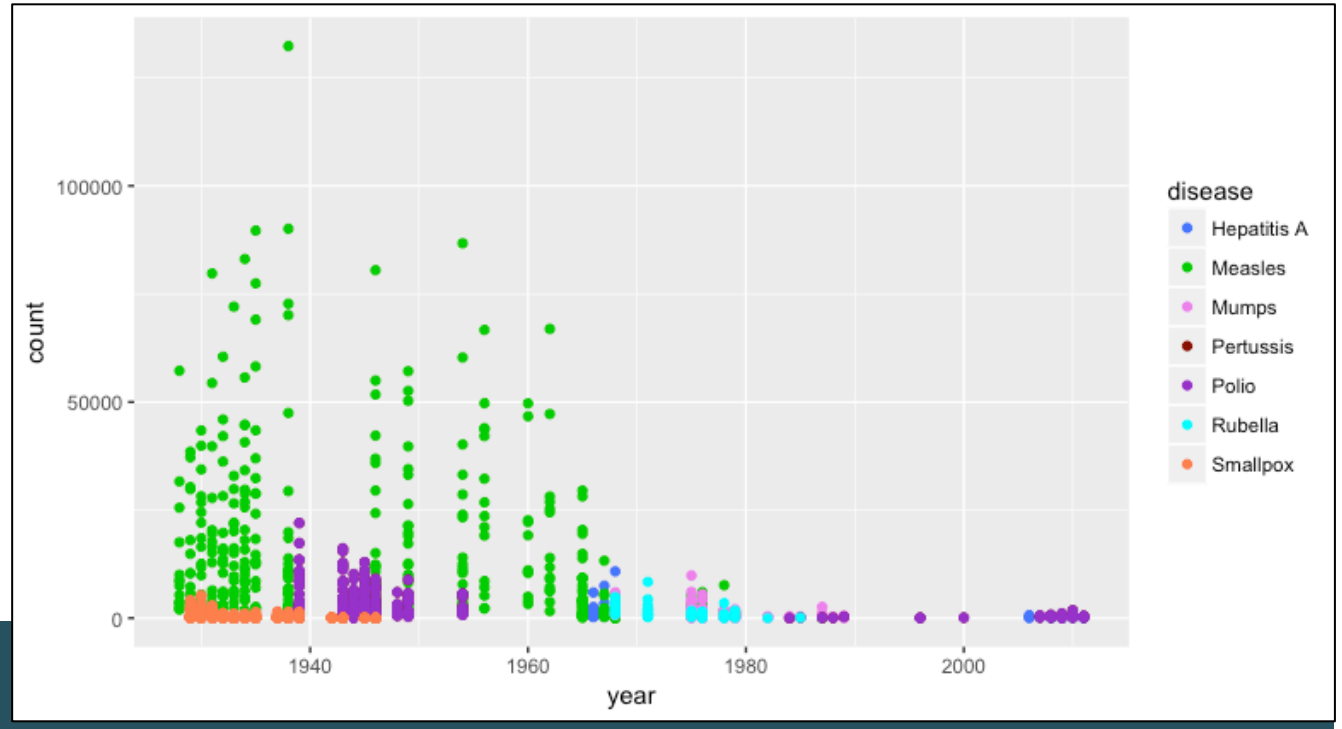
```
# A tibble: 7 x 6
  disease     `first(count.x)` `sum(count.y)` `first(state.x)` `first(year)` percent.of.disea…
  <fct>                  <dbl>          <dbl> <fct>                  <dbl>             <dbl>
1 Hepatitis A            10821          29804 California               1968              36.3
2 Measles               132342         620916 Pennsylvania             1938              21.3
3 Mumps                   9867          49271 Michigan                 1975              20.0
4 Pertussis              22013         185222 New York                 1939              11.9
5 Polio                  22013         185222 New York                 1939              11.9
6 Rubella                 8384          32045 California               1971              26.2
7 Smallpox                5239          44027 Indiana                  1930              11.9
```

Fun with GGPLOT

```
Week52plot <- ggplot(Week52_reporting, aes(x=year, y=count, color=disease)) + geom_point()

Week52plot <- (Week52plot + scale_color_manual(values = c("royalblue1", "green3", "violet",
"red4", "darkorchid", "cyan1","coral")))

options(scipen = 10)
```



The End

devleague