Econ 106

Lecture 17

slides derived from:

https://r4ds.had.co.nz/graphics-for-communication.html

Reminders

- Lab #5 is due Sunday, 11:59pm (best 4 out of 5 count towards your grade)
- Research Milestone #3 can be turned in until 11:59pm tonight (late penalty)
- Final project is due Sunday 11:59pm

https://pollev.com/vsovero

Grading: Written Report

Things I will be evaluating:

- Does your writeup follow the structure of the outline?
- Do you have a strong hypothesis/objective?
- Are your visualizations clearly connected to the objective?
- Can you interpret your visualizations correctly?
- Can you provide a reason for the observed trends in your visualizations?
- Can you connect the findings to your objective?
- **Does your research project reflect independent thought and originality?**

Student Example 1: Police Stops in Oakland

Objectives & Goals

#1

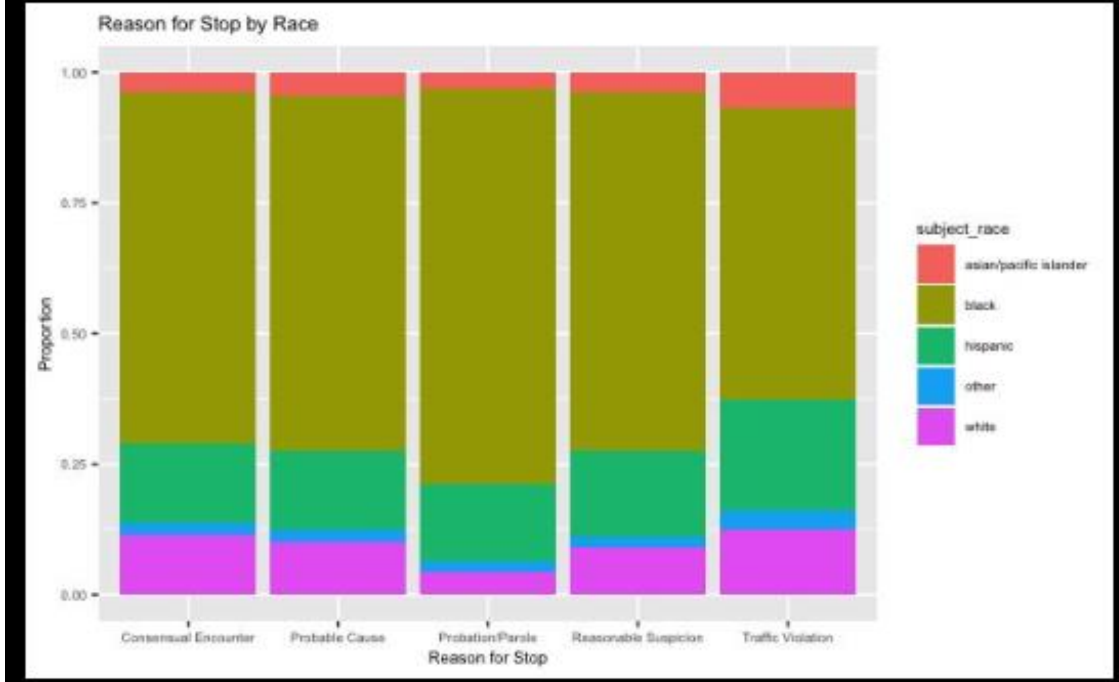
I want to analyze any potential correlations between race and the reason an individual was stopped. I also want to consider if police stops are dependent on race by checking if the subjects in the data set proportionately represent the racial composition of Oakland.

#2

With faceting, I want to show any discrepancies between the way police officers interact with subjects of different races during a stop. I will do this using the logical variables from the data set.

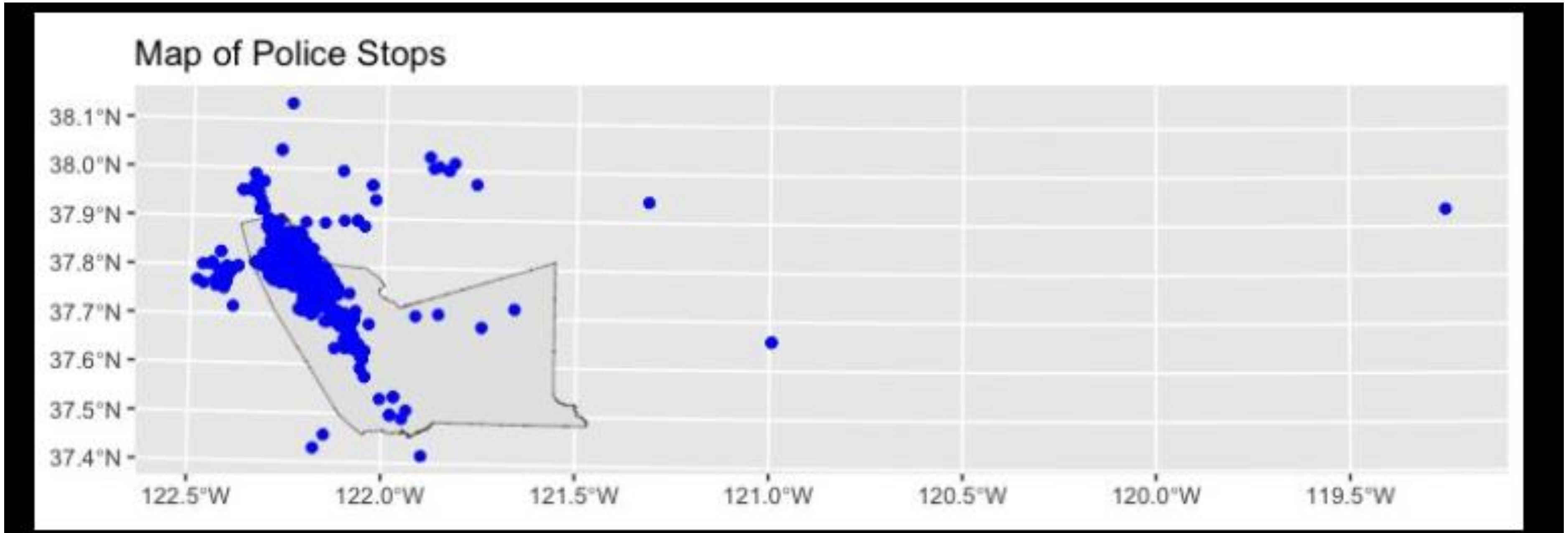
#3

For spatial data, I want to see if there are any clusters or any areas where police stops occur the most. I will then compare these clusters to the racial composition of nearby neighborhoods.



Reason for Stop Broken Down by Race

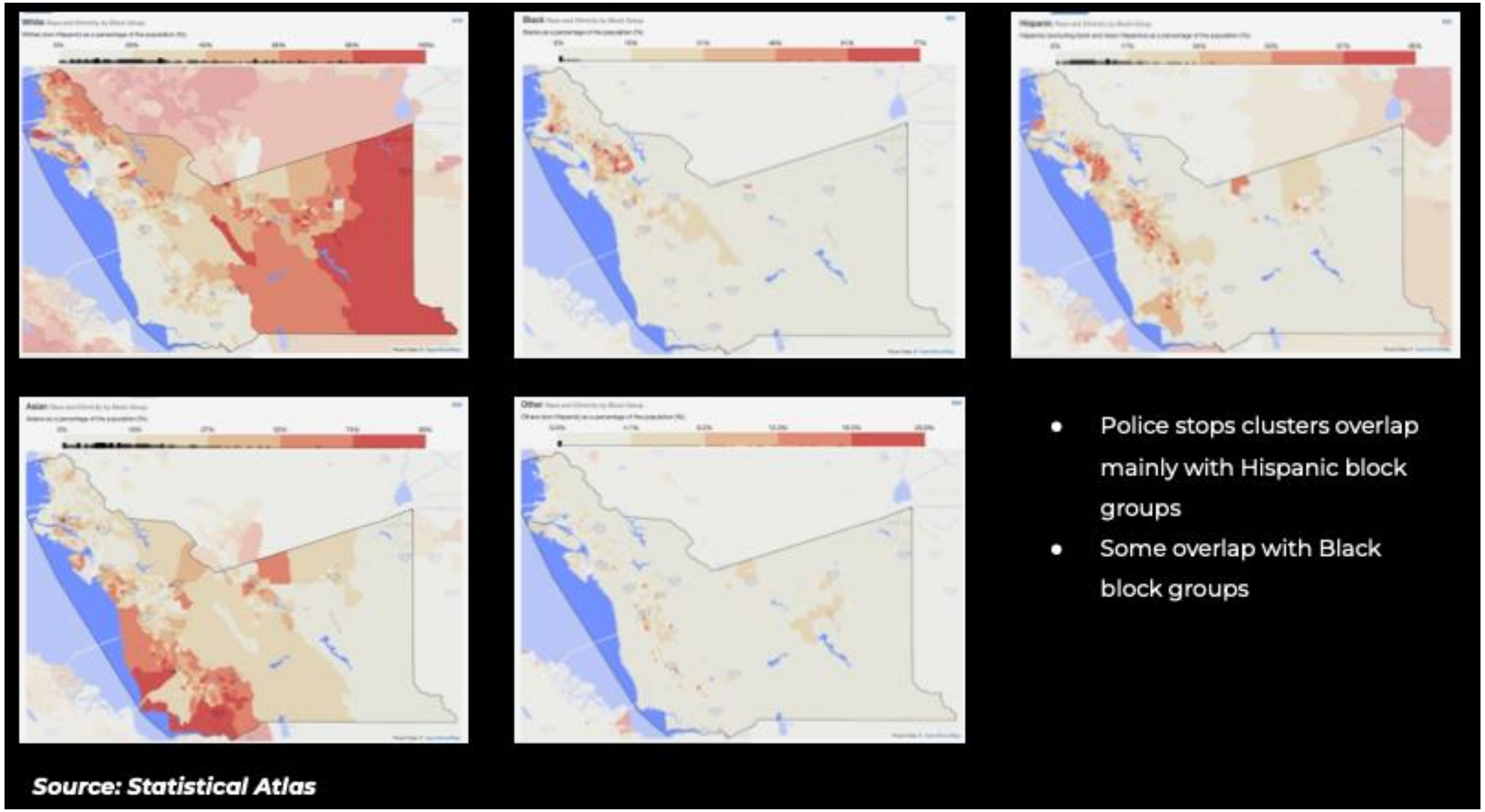
- Black subjects were the biggest proportion for each reason
- Probation/parole was the reason with the highest proportion of Black subjects
- Racial breakdown for probation/parole and traffic violation differs slightly from the rest of the reasons



Location of Each Police Stop Over Alameda County

- Few stops outside of the county
- Large cluster near the top left of the projection
- Smaller cluster to the west of the county border

Student compiled additional information to provide context for the results.

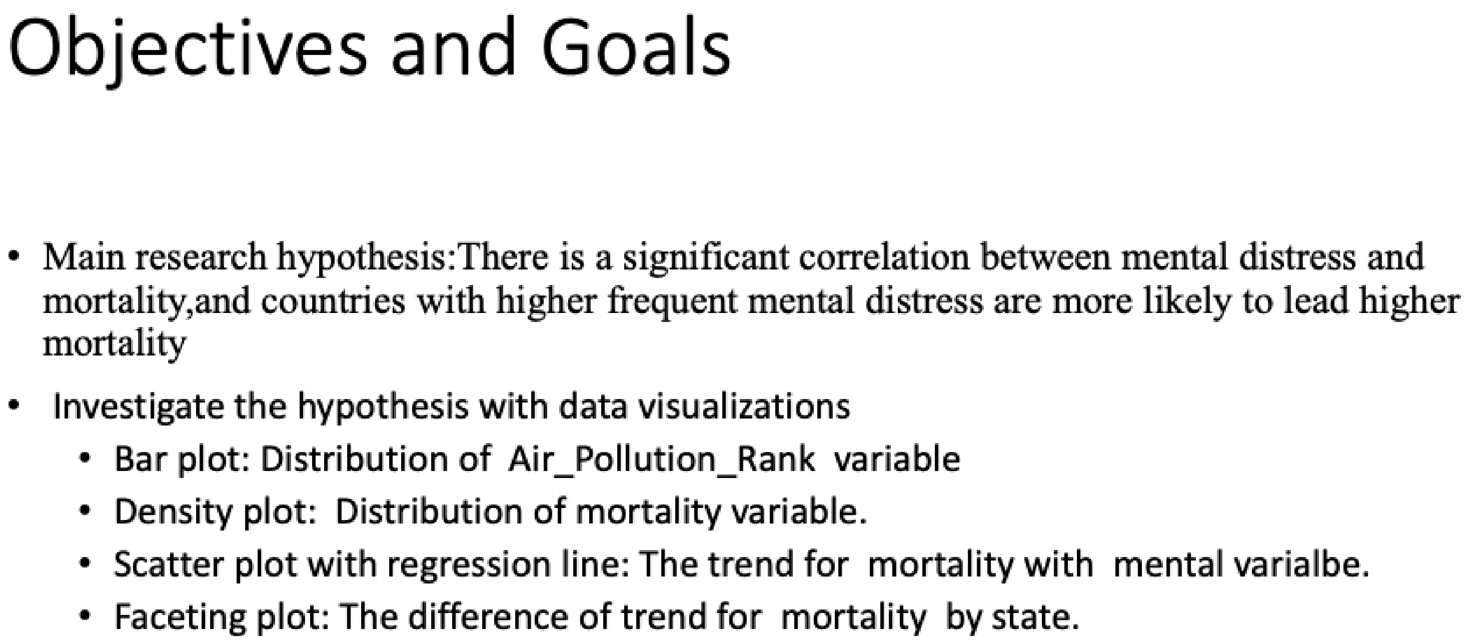


- Police stops clusters overlap mainly with Hispanic block groups
- Some overlap with Black block groups

Source: Statistical Atlas

Student Example 2: Mortality and Mental Health

Clear objective, but it is unclear how the visualizations are tied to the objective

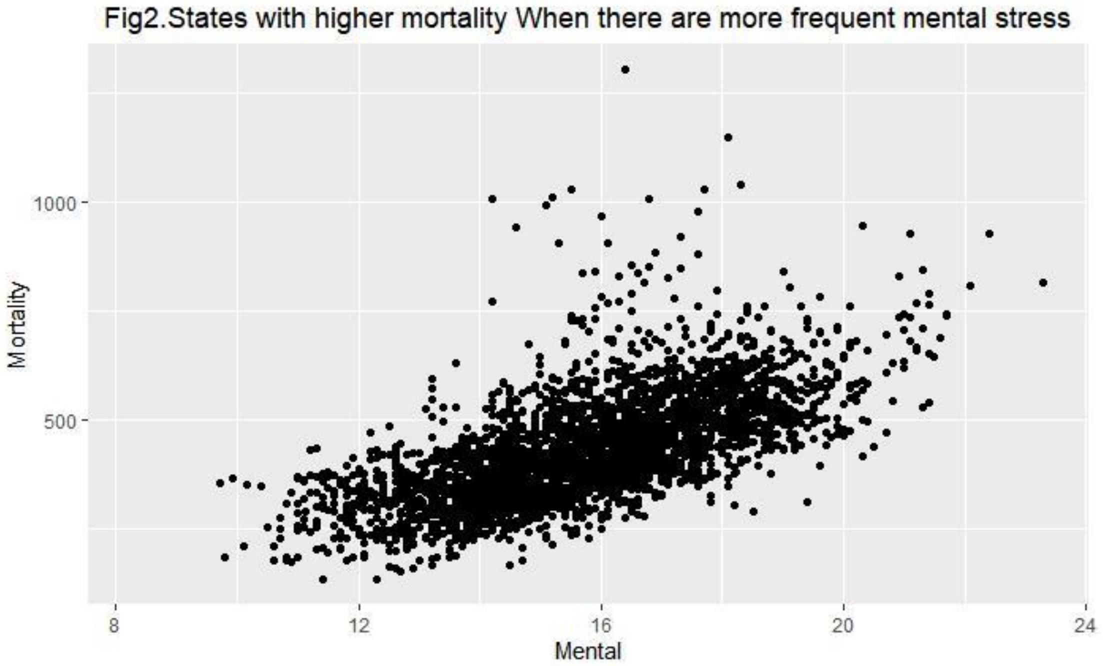


Objectives and Goals

- Main research hypothesis:There is a significant correlation between mental distress and mortality,and countries with higher frequent mental distress are more likely to lead higher mortality
- Investigate the hypothesis with data visualizations
  - Bar plot: Distribution of Air_Pollution_Rank variable
  - Density plot: Distribution of mortality variable.
  - Scatter plot with regression line: The trend for mortality with mental varialbe.
  - Faceting plot: The difference of trend for mortality by state.

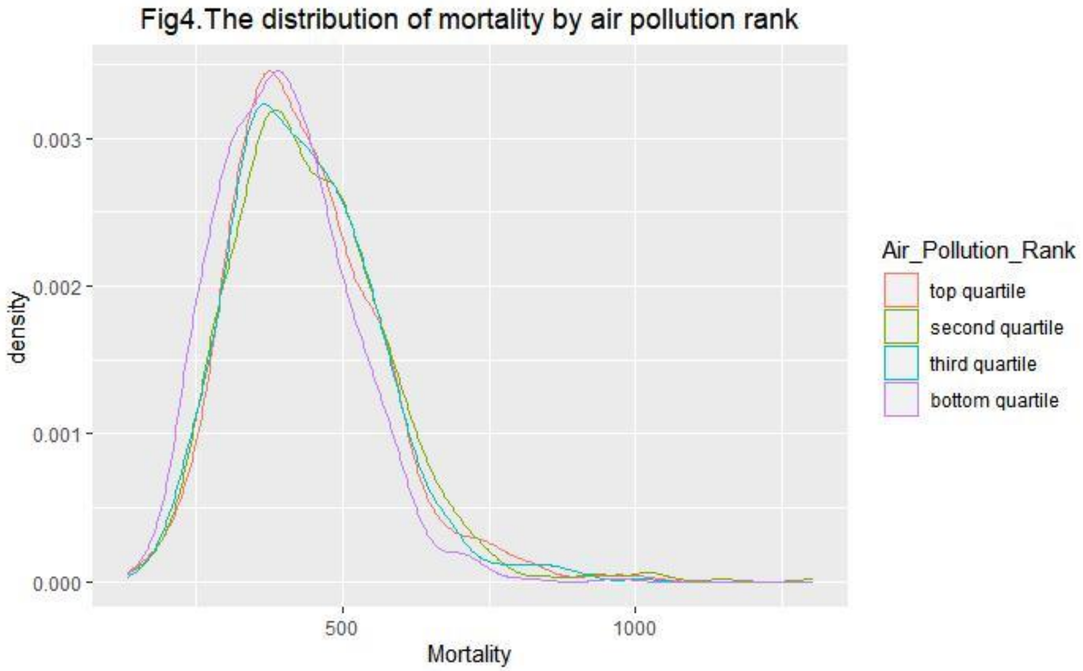
Scatter plot

Most strongly tied to objective (relationship between mental health and mortality)



Density Plot

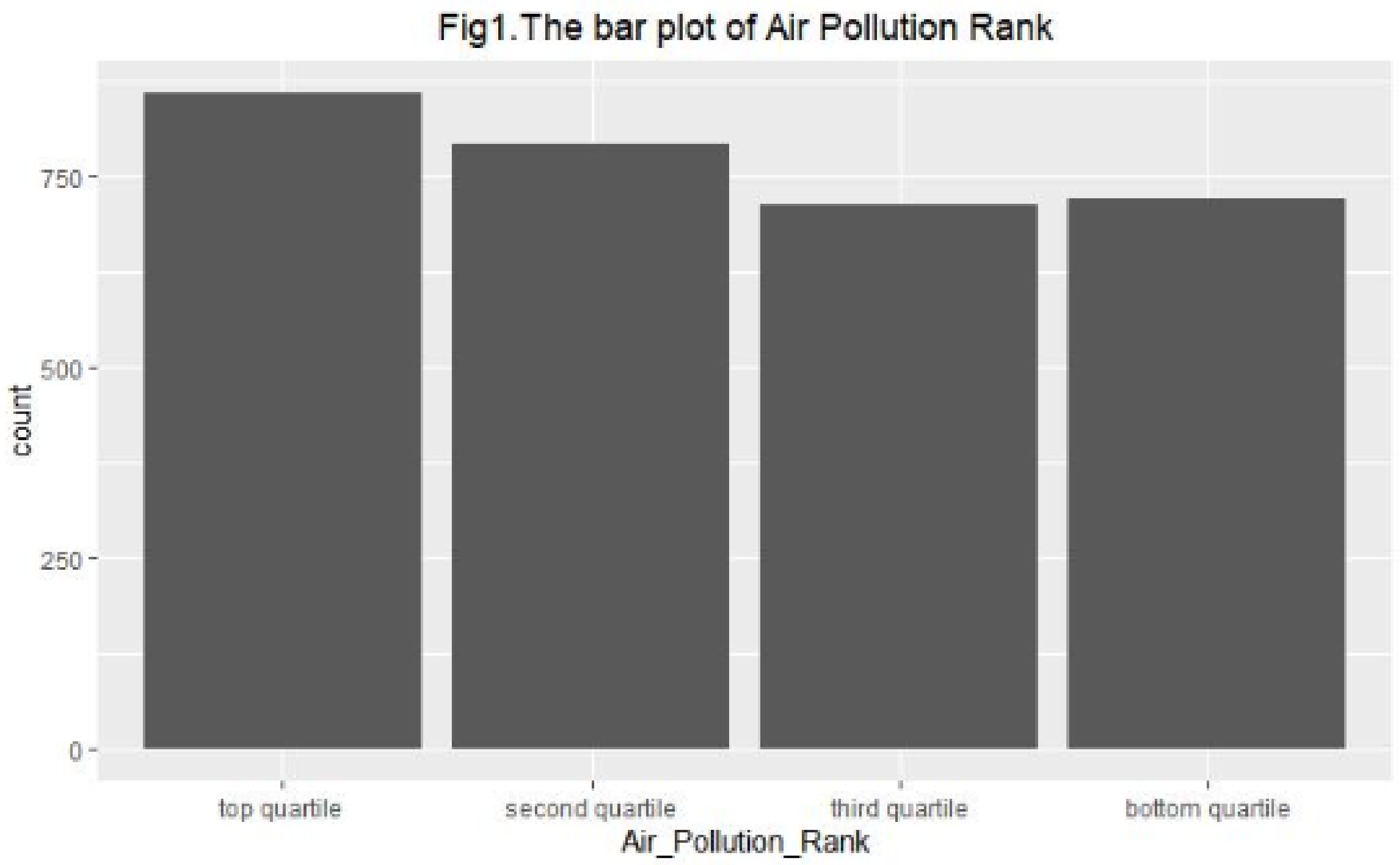
no clear connection to objective (mental health and mortality)



Bar Plot

Not clear how this graph helps investigate the objective (number of counties by air pollution quartile)

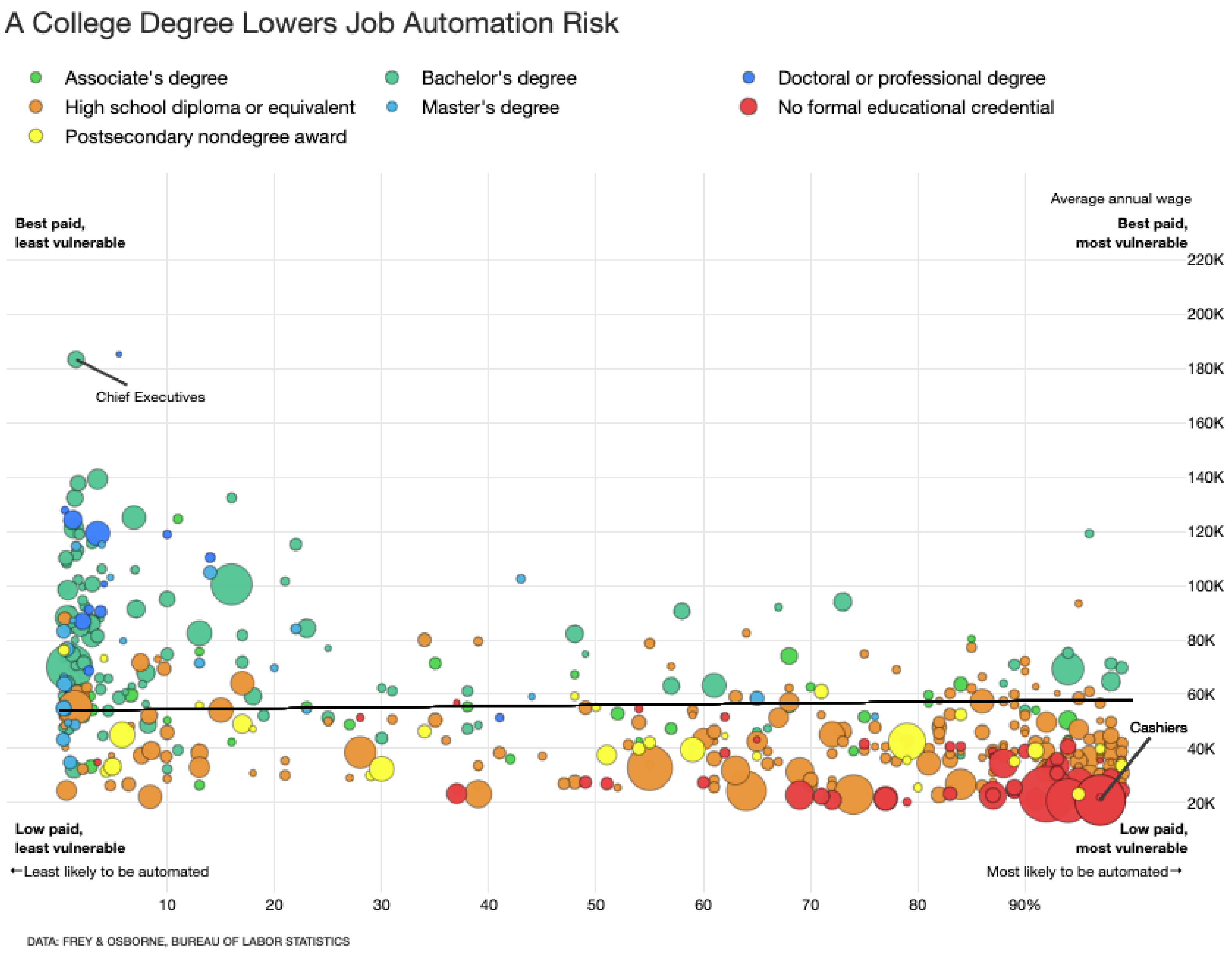
Only has one variable



Outline

- Exploratory vs Explanatory Plots
- Visualization Best Practices
- Customizing your plots

#tidytuesday



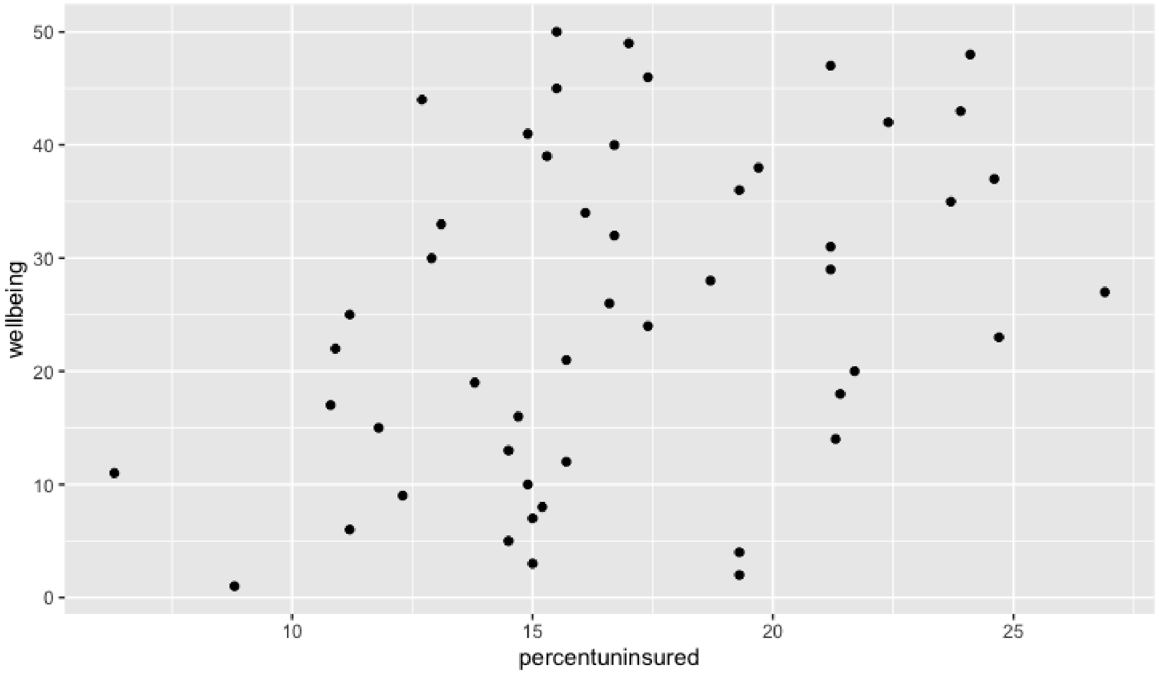
https://connorrothschild.github.io/v2/post/tidy-tuesday-replication/

Exploratory vs Explanatory Plots

- As you move further into your data analysis, you will shift from making **exploratory plots** to **explanatory plots**.
- **Exploratory Plots:**
  - data displays to help you better understand and discover hidden patterns in the data you’re working with.
- **Explanatory Plots**:
  - data displays that aim to **communicate insights to others**.

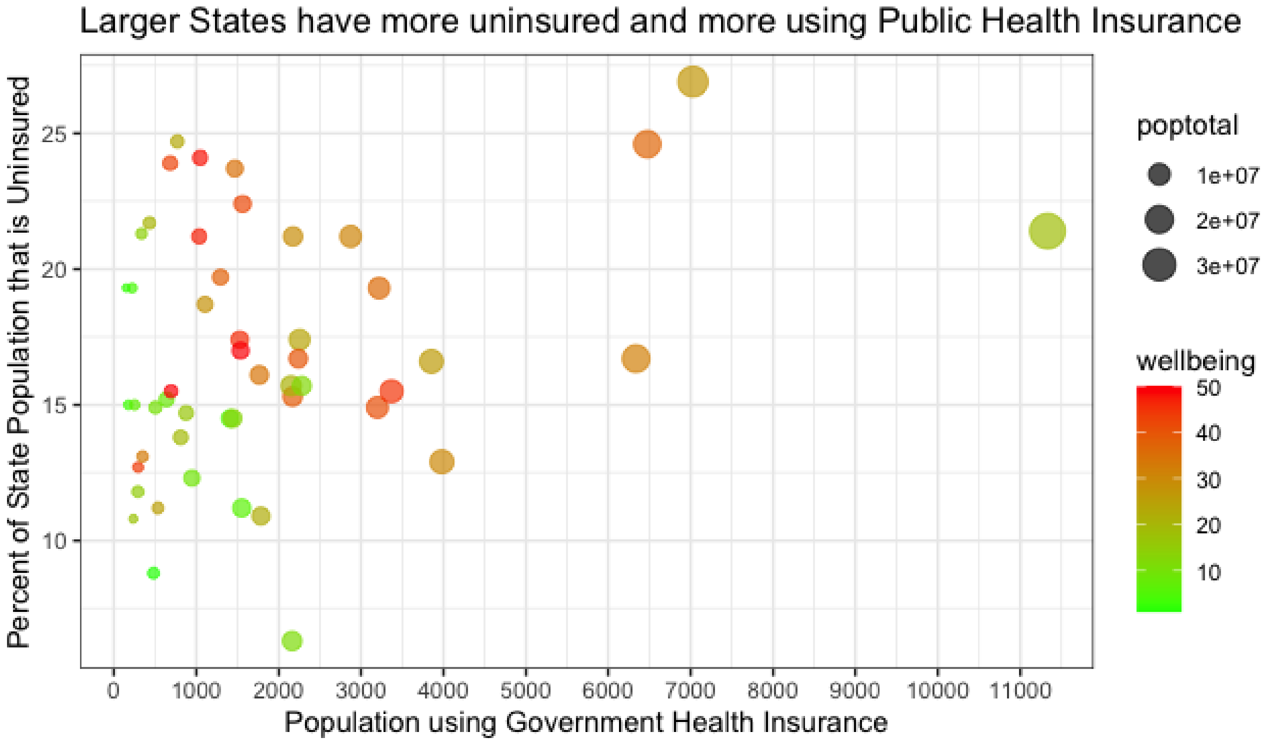
Exploratory Plot Example

- When exploring your data, you will make a lot of plots
- They won’t have a lot of formatting/labeling
- These plots are for “internal use”- they help you understand your data



Explanatory Plot Example

- These are plots for "external use"- communicating your findings to others
- Things to check:
  - the axis labels should all be clear
  - the labels should all be large enough to read
  - the colors should all be carefully chosen

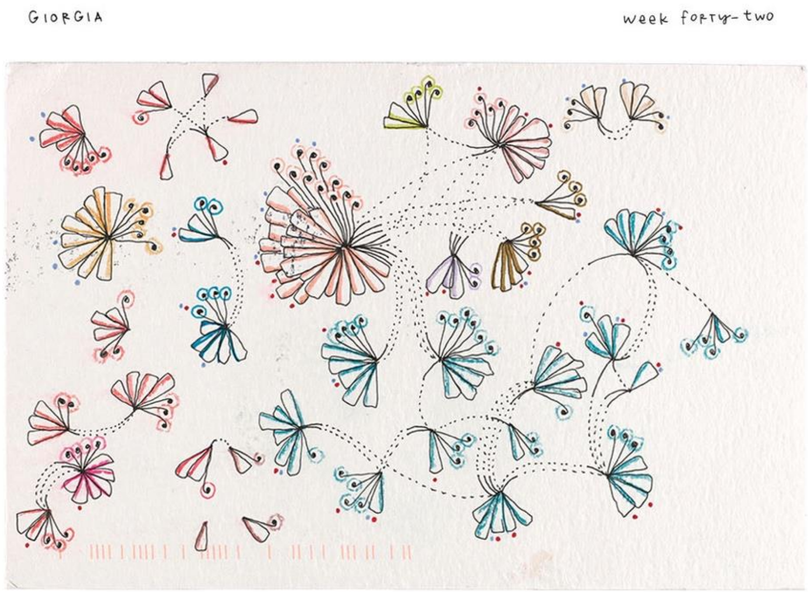


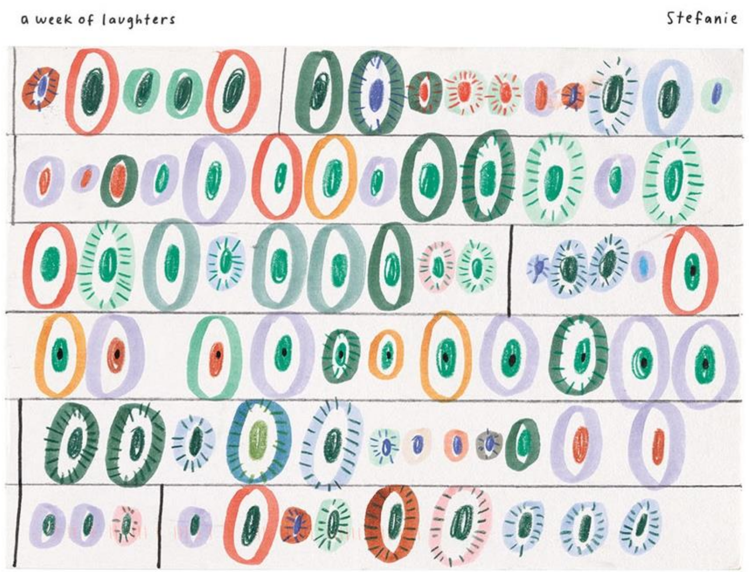
Communicating Your Findings with Plots

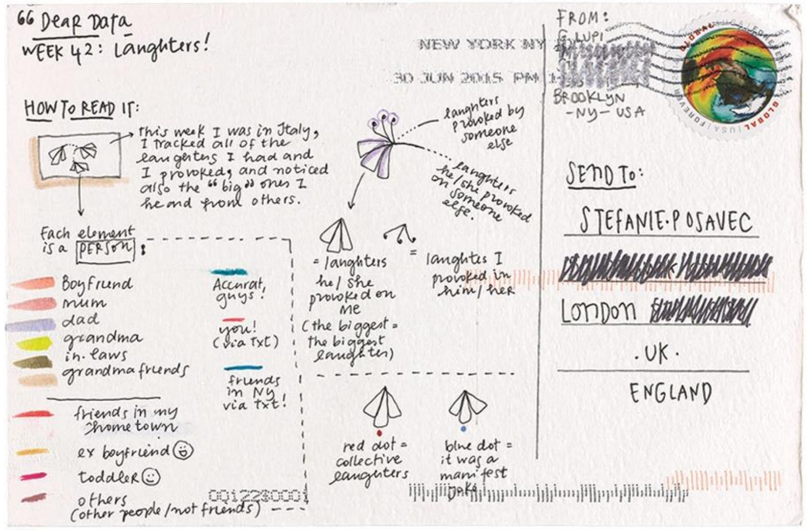
- Ask yourself: What is the **central message** you are trying to communicate?
- Decide, then build your plot around that message.
- Make that message as easy to see as you can.
- **Remove the clutter** -- get rid of any features of the visualization that do not contribute to the central message.

Data Visualization as Art

http://www.dear-data.com/theproject







66 Dear Data
WEEK 42: Laughters!

NEW YORK NY
30 JUN 2015 PM

HOW TO READ IT:

this week I was in Italy, I tracked all of the laughters I had and I provoked, and noticed also the "big" ones I heard from others.

Each element is a PERSON:

- Boyfriend
- mum
- dad
- grandma
- in-laws
- grandma friends
- friends in my home town
- ex boyfriend
- toddler
- others (other people / not friends)
- Accurat, guys!
- you! (via txt)
- friends in NY via txt!

laughters provoked by someone else

laughters he/she provoked on someone else.

= laughters he/she provoked on me (the biggest = the biggest laughter)

= laughtes I provoked in him/her

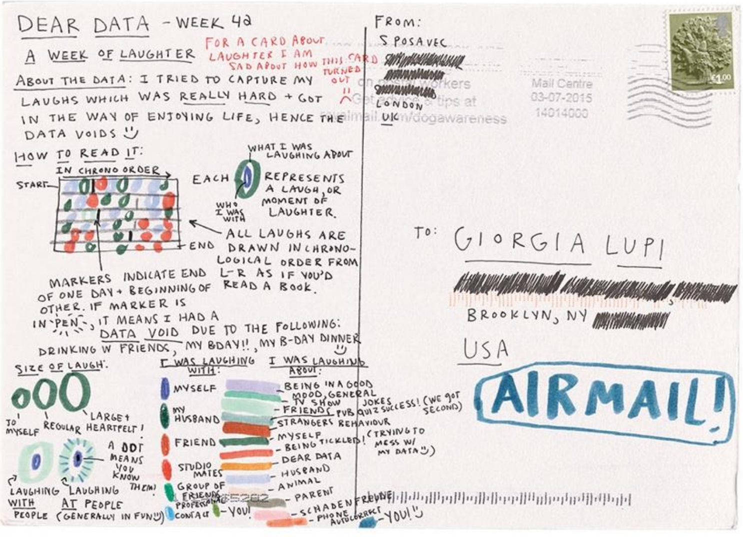
red dot = collective laughters

blue dot = it was a manifest joke

When data-gathering gets in the way: Giorgia laughs and then has to report it. So, is she fully enjoying her cheerful moments as they pop up?

FROM:
G. LUPI
[redacted]
BROOKLYN
-NY- USA

SEND TO:
STEFANIE.POSAVEC
[redacted]
London [redacted]
.UK.
ENGLAND



DEAR DATA - WEEK 42

A WEEK OF LAUGHTER

FOR A CARD ABOUT LAUGHTER I AM SAD ABOUT HOW THIS CARD TURNED OUT :(

ABOUT THE DATA: I TRIED TO CAPTURE MY LAUGHS WHICH WAS REALLY HARD + GOT IN THE WAY OF ENJOYING LIFE, HENCE THE DATA VOIDS :)

HOW TO READ IT:

IN CHRONO ORDER
START
END

EACH REPRESENTS A LAUGH, OR MOMENT OF LAUGHTER.

WHAT I WAS LAUGHING ABOUT
WHO I WAS WITH

ALL LAUGHS ARE DRAWN IN CHRONOLOGICAL ORDER FROM L-R AS IF YOU'D READ A BOOK.

MARKERS INDICATE END OF ONE DAY + BEGINNING OF OTHER. IF MARKER IS IN "PEN", IT MEANS I HAD A DATA VOID DUE TO THE FOLLOWING: DRINKING W FRIENDS, MY BDAY!!, MY B-DAY DINNER :)

SIZE OF LAUGH:
TO MYSELF
REGULAR
LARGE + HEARTFELT!

A DOT MEANS YOU KNOW THEM!

LAUGHING WITH PEOPLE
LAUGHING AT PEOPLE (GENERALLY IN FUN!)

I WAS LAUGHING WITH:
- MYSELF
- MY HUSBAND
- FRIEND
- STUDIO MATES
- GROUP OF FRIENDS
- PROFESSIONAL CONTACT
- YOU!

I WAS LAUGHING ABOUT:
- BEING IN A GOOD MOOD, GENERAL
- TV SHOW
- JOKES
- FRIENDS
- PUB QUIZ SUCCESS! (WE GOT SECOND)
- STRANGERS BEHAVIOUR
- MYSELF
- BEING TICKLED! (TRYING TO MESS W/ MY DATA :))
- DEAR DATA
- HUSBAND
- ANIMAL
- PARENT
- SCHADENFREUDE
- PHONE AUTOCORRECT
- YOU! :)

FROM:
S POSAVEC
[redacted]
LONDON
UK

Mail Centre
03-07-2015
14014000

on postal workers
Get advice & tips at
royalmail.com/dogawareness

TO: GIORGIA LUPI
[redacted]
BROOKLYN, NY [redacted]
USA

AIRMAIL!

Stefanie was banned by her husband from gathering data at her birthday dinner in a posh restaurant (for obvious reasons). Hence the "data void".

Data Example for today

- We’re going to use data from the cspp package (https://github.com/IPPSR/cspp)

```
cspp_data <- get_cspp_data(vars=c("percentuninsured",
"wellbeing", "sdce", "doctorsPerCapita","higrenew",
"popgovhealthins", "popnohealthins", "popprivhealthins",
"hmdindex", "health_pro" ),
years = seq(2010,2010))
```

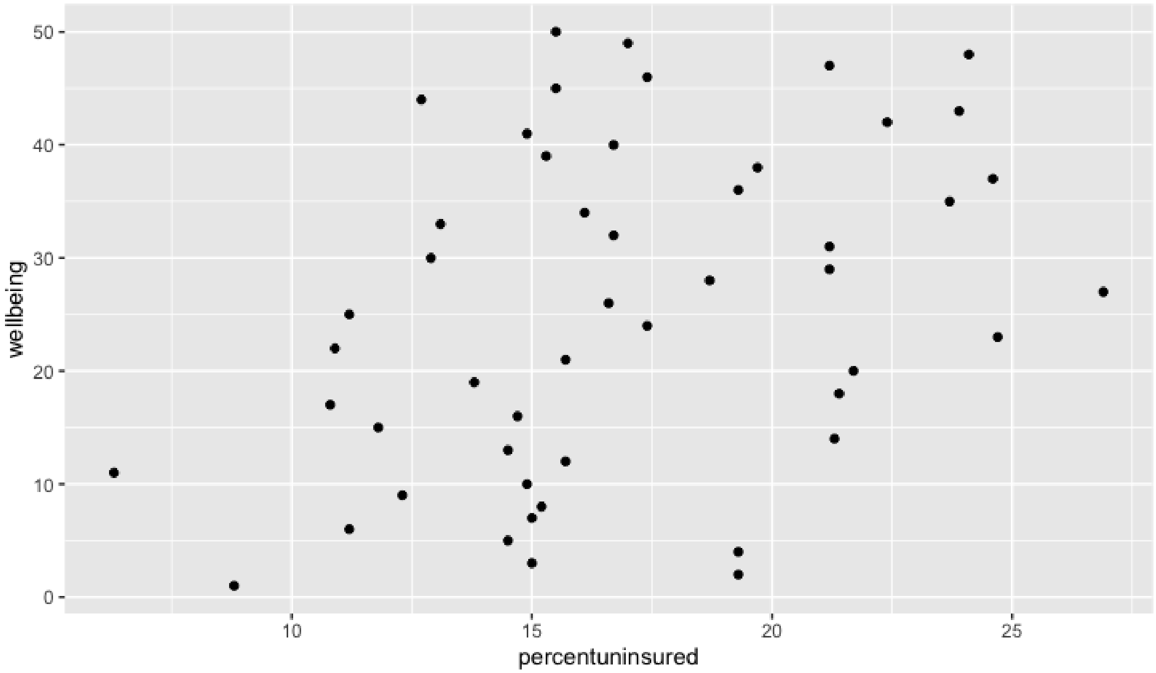
Plot Adjustments

1. Labels
   a) title
   b) x and y axis
   c) annotations
2. Scales
   a) x and y axis
   b) color

Example

- Let’s see what we can do to improve this scatter plot of a state’s well being ranking against the percent of the state population that is uninsured

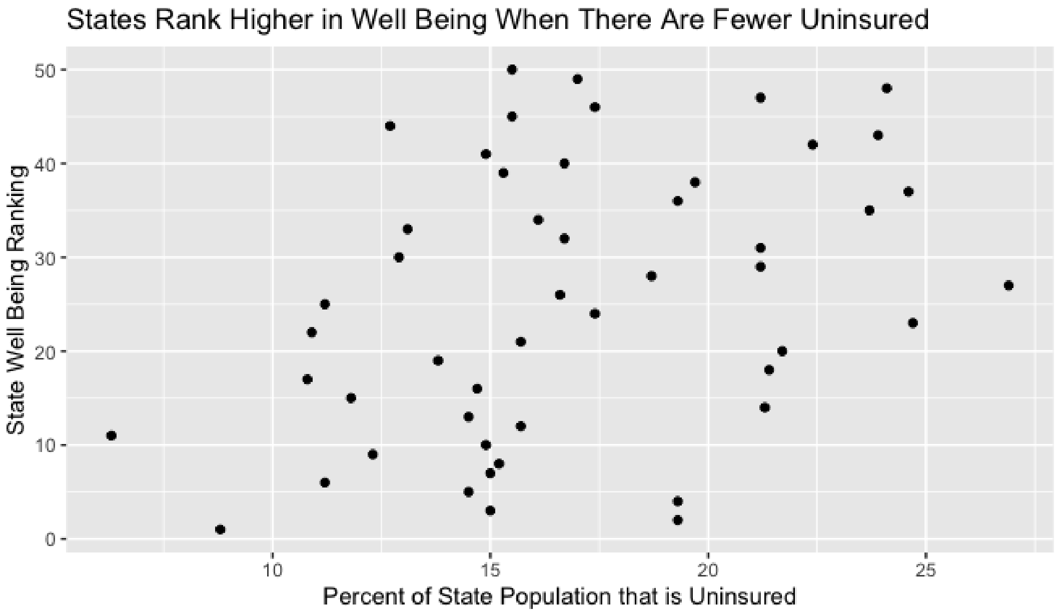
```
ggplot(data=cspp_data,
      mapping=aes(x=percentuninsured, y=wellbeing))+
 geom_point()
```



Labels

- **labs**() - specify labels
- Arguments:
  - **title**: plot title
  - **x**: x axis label
  - **y**: y axis label

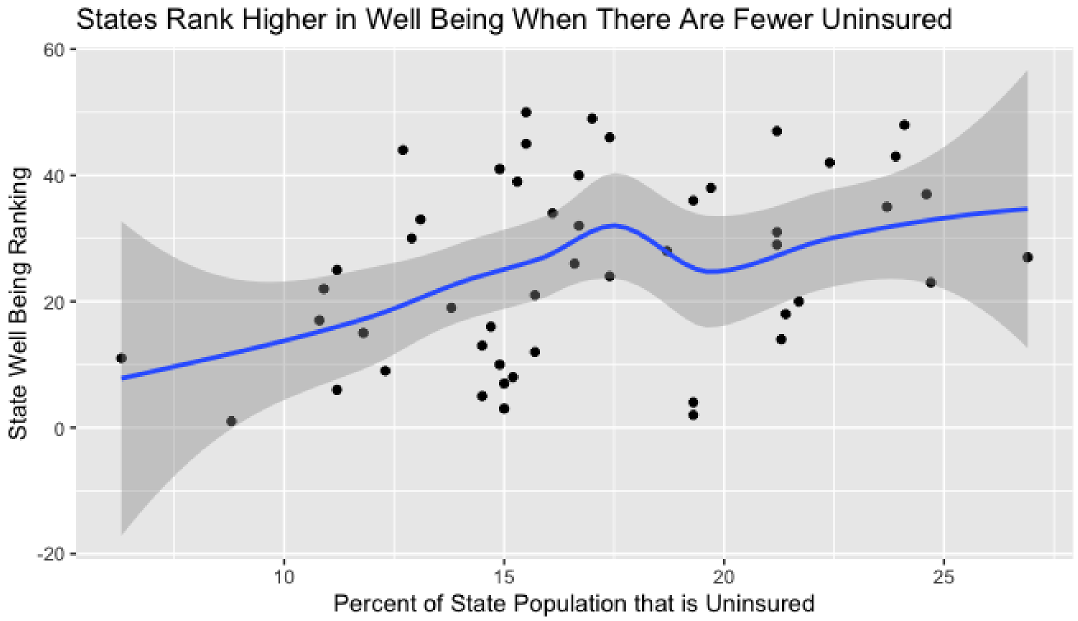
```
ggplot(data=cspp_data,
      mapping=aes(x=percentuninsured, y=wellbeing))+
  geom_point()+
labs(title = 'States Rank Higher in Well Being When There Are Fewer Uninsured',
        x = 'Percent of State Population that is Uninsured',
        y = 'State Well Being Ranking' )
```



Add line of best fit

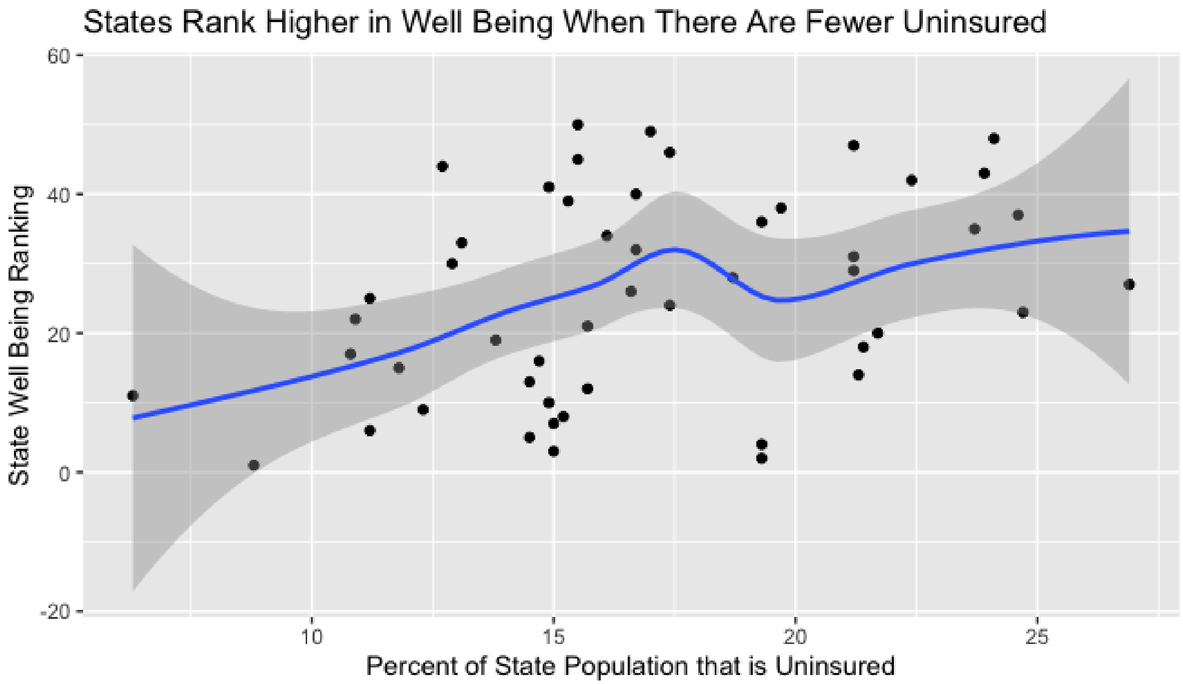
- We can use geom_smooth() to add a line of best fit (expected value of y for every value of x)

```
ggplot(data=cspp_data,
     mapping=aes(x=percentuninsured, y=wellbeing))+
  geom_point()+
labs(title = 'States Rank Higher in Well Being When There Are
Fewer Uninsured',
       x = 'Percent of State Population that is Uninsured',
       y = 'State Well Being Ranking' )+
geom_smooth()
```



Add line of best fit

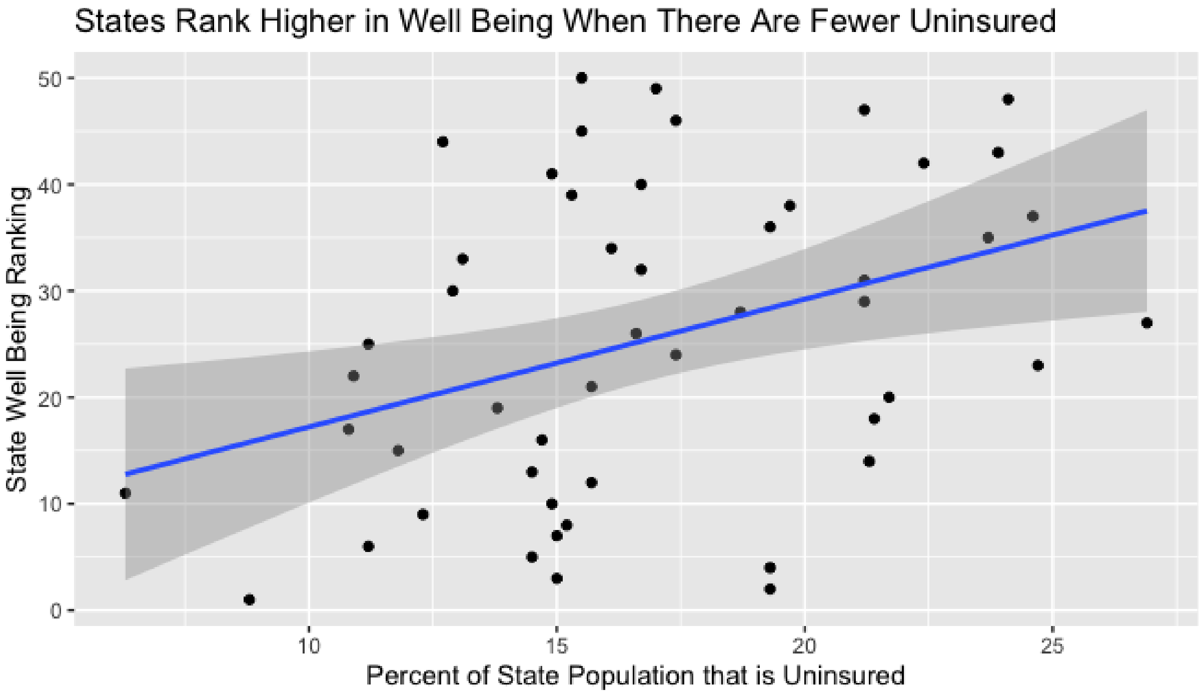
- Blue line: estimated line of best fit
- **Interpretation**:
  - mostly negative relationship between percent uninsured and ranking
  - seems mostly linear
- Dark grey: 95% CI for estimated line of best fit
- **Interpretation:** the relationship is not precisely estimated for the very low and high values of percent uninsured (wide bands)



Make it linear

- estimate a straight line with the method argument

```
ggplot(data=cspp_data,
      mapping=aes(x=percentuninsured, y=wellbeing))+
  geom_point()+
labs(title = 'States Rank Higher in Well Being When There Are
Fewer Uninsured',
         x = 'Percent of State Population that is Uninsured',
         y = 'State Well Being Ranking' )+
geom_smooth(method= "lm")
```



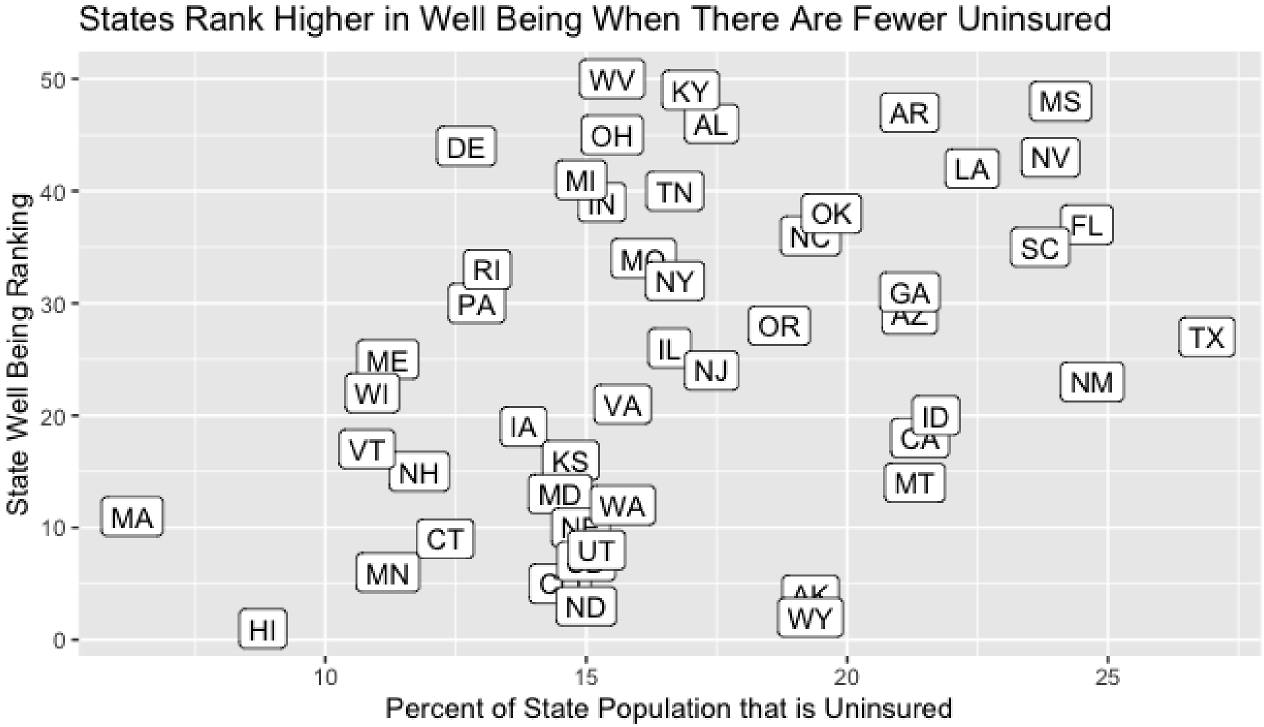
Class Exercise

- Create a scatter plot with doctors per capita on the x-axis and well being ranking (1-best, 50-worst) on the y-axis
- add a line of best fit (try linear and nonlinear)

https://pollev.com/vsovero

Annotations

- In addition to labelling major components of your plot, it's often useful to label individual observations or groups of observations.



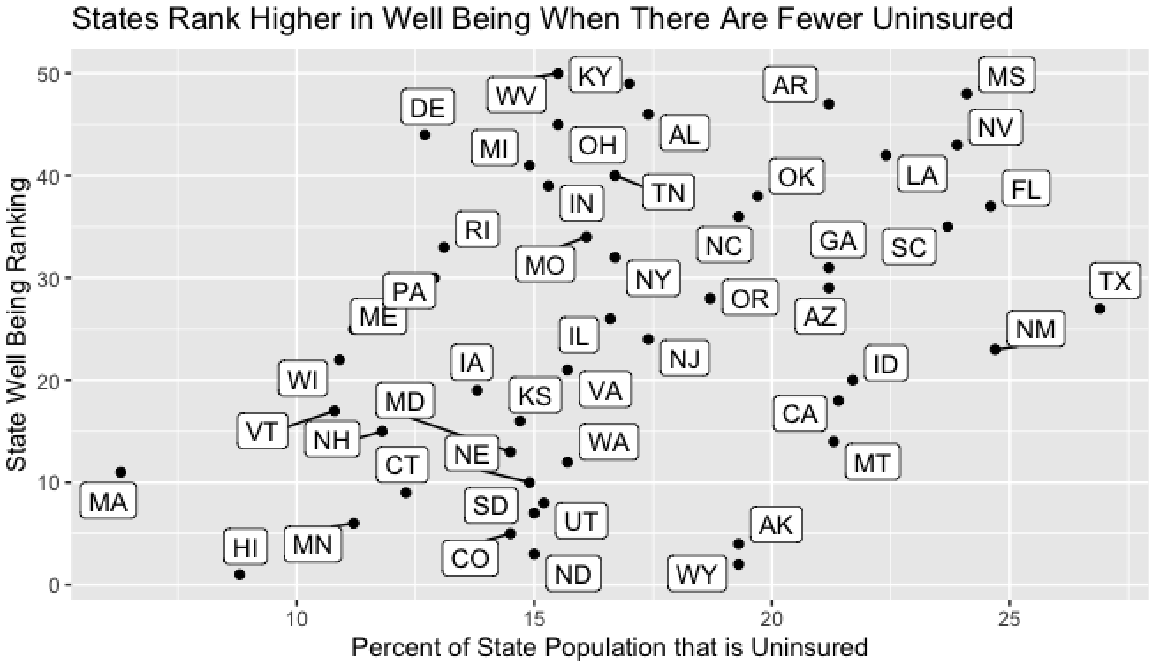
Annotations

- **geom_label**() – add annotations to a geom
- Arguments:
    - **label**
- Remember to use aes() when referencing variable names

```
ggplot(data=cspp_data,
       mapping=aes(x=percentuninsured, y=wellbeing))+
 geom_point()+
labs(title = 'States Rank Higher in Well Being When There Are Fewer Uninsured',
        x = 'Percent of State Population that is Uninsured',
        y = 'State Well Being Ranking' )+
geom_label(aes(label=st )
```

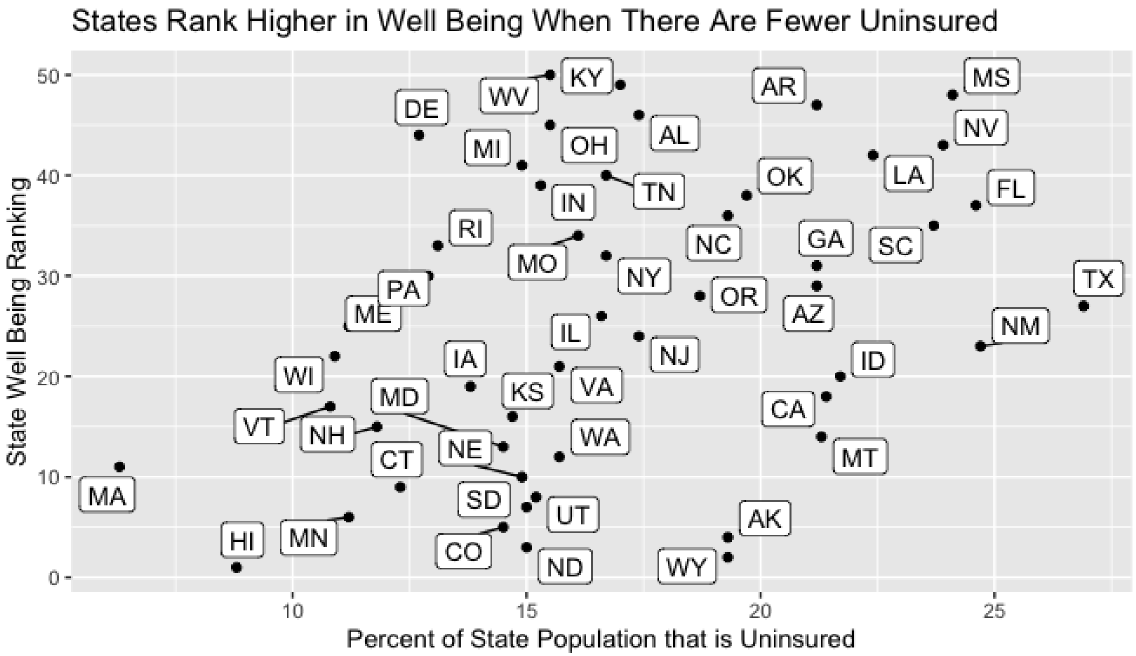
Annotations

- The annotations can get cluttered if they are too close to one another
- We can shift the labels away from the points using geom_label_repel() from the ggrepel package



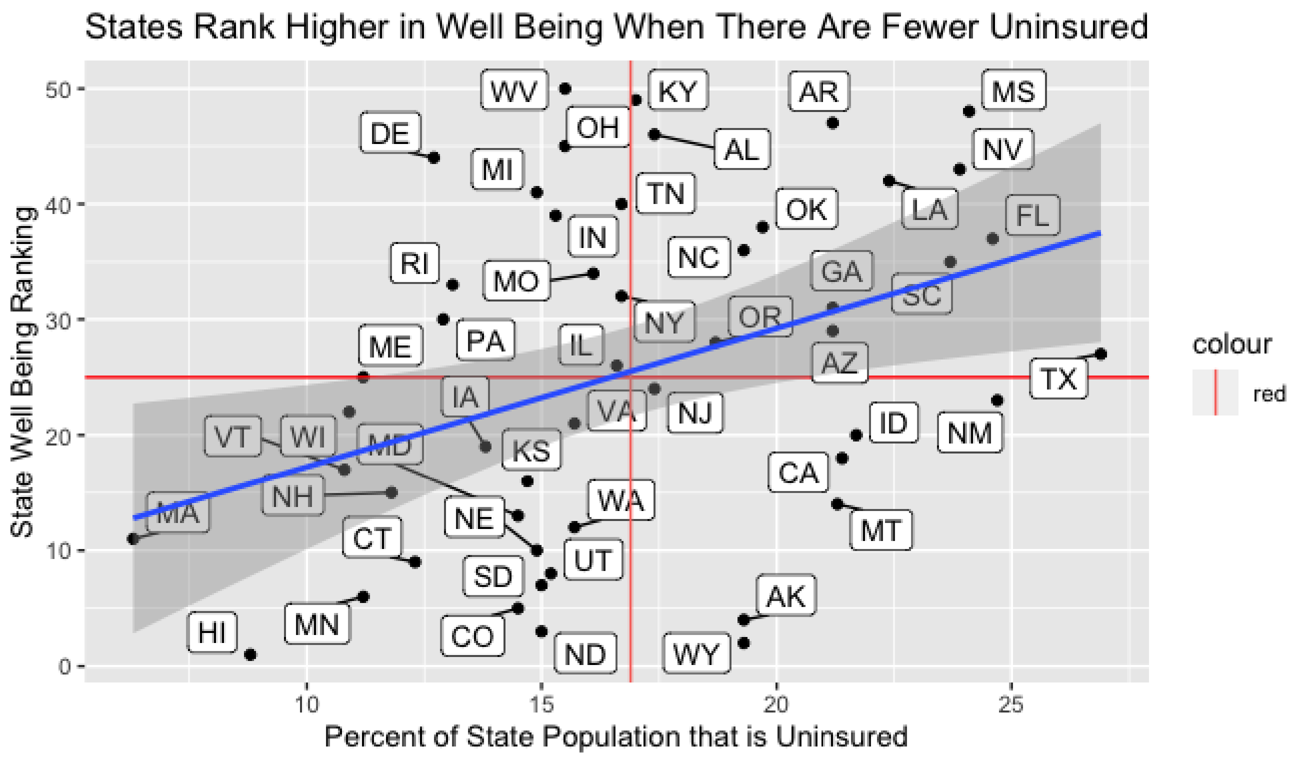
ggrepel package

```
ggplot(data=cspp_data,
      mapping=aes(x=percentuninsured, y=wellbeing))+
  geom_point()+
labs(title = 'States Rank Higher in Well Being When
There Are Fewer Uninsured',
        x = 'Percent of State Population that is Uninsured',
        y = 'State Well Being Ranking' )+
geom_hline(yintercept=25, color='red' )+
geom_label_repel(aes(label=st)
```



Quick PSA: don't overdo it

- You can definitely put too much on a graph
- This looks like a hot mess



Class Exercise

- Create a scatter plot with doctors per capita on the x-axis and well being ranking (1-best, 50-worst) on the y-axis
- label the states using ggrepel

https://pollev.com/vsovero

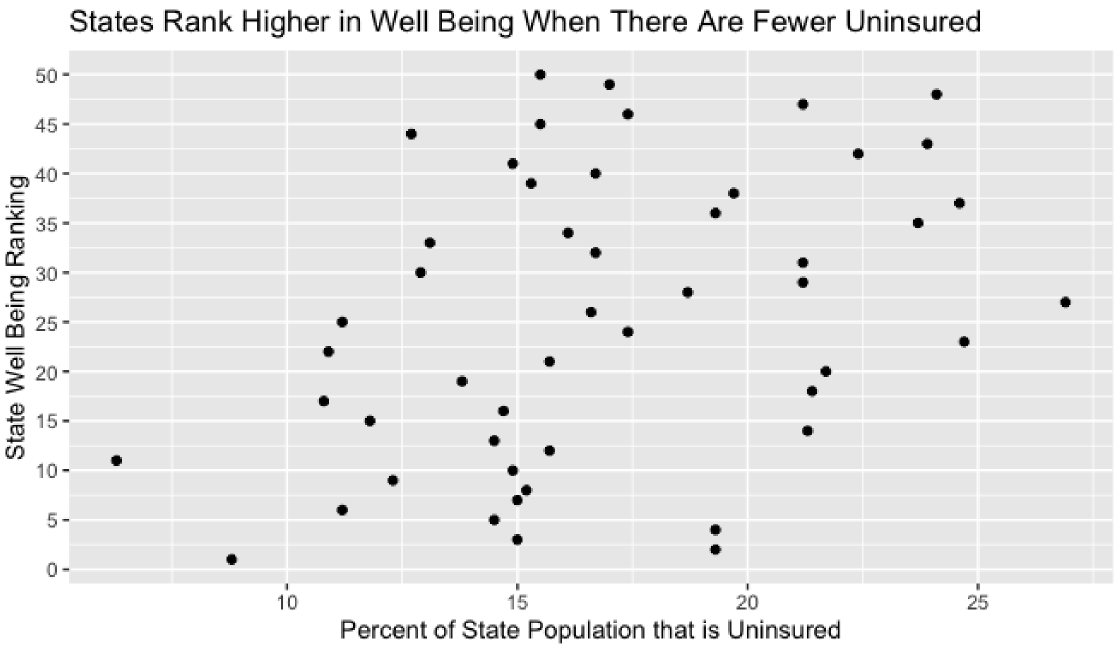
Scales

- Scales control how your data is mapped on your plot
- Some common adjustments:
  - axis ticks and labels
  - colors

Continuous Scales

- The scale of your plot for continuous variables can be controlled using:
  - scale_x_continuous()
  - scale_y_continuous()

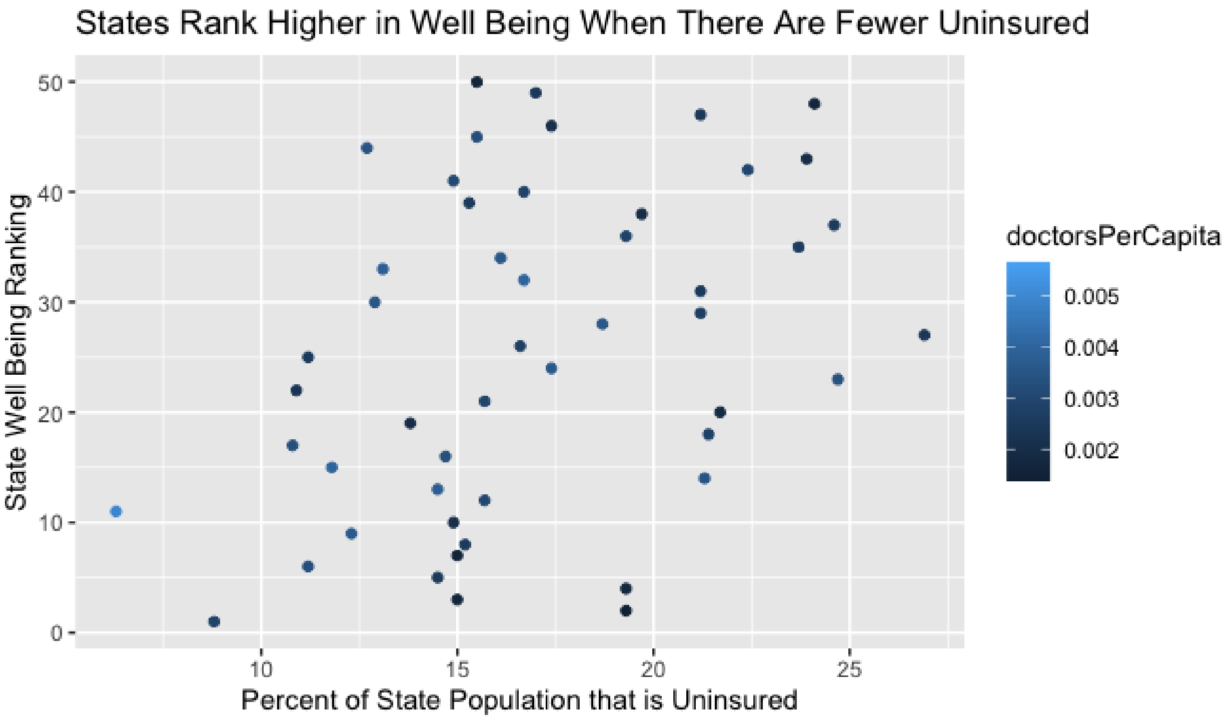
```
ggplot(data=cspp_data,
     mapping=aes(x=percentuninsured, y=wellbeing))+
  geom_point()+
labs(title = 'States Rank Higher in Well Being When There Are
Fewer Uninsured',
       x = 'Percent of State Population that is Uninsured',
       y = 'State Well Being Ranking' )+
scale_y_continuous(breaks=seq(0,50, by=5))
```



Color Scales

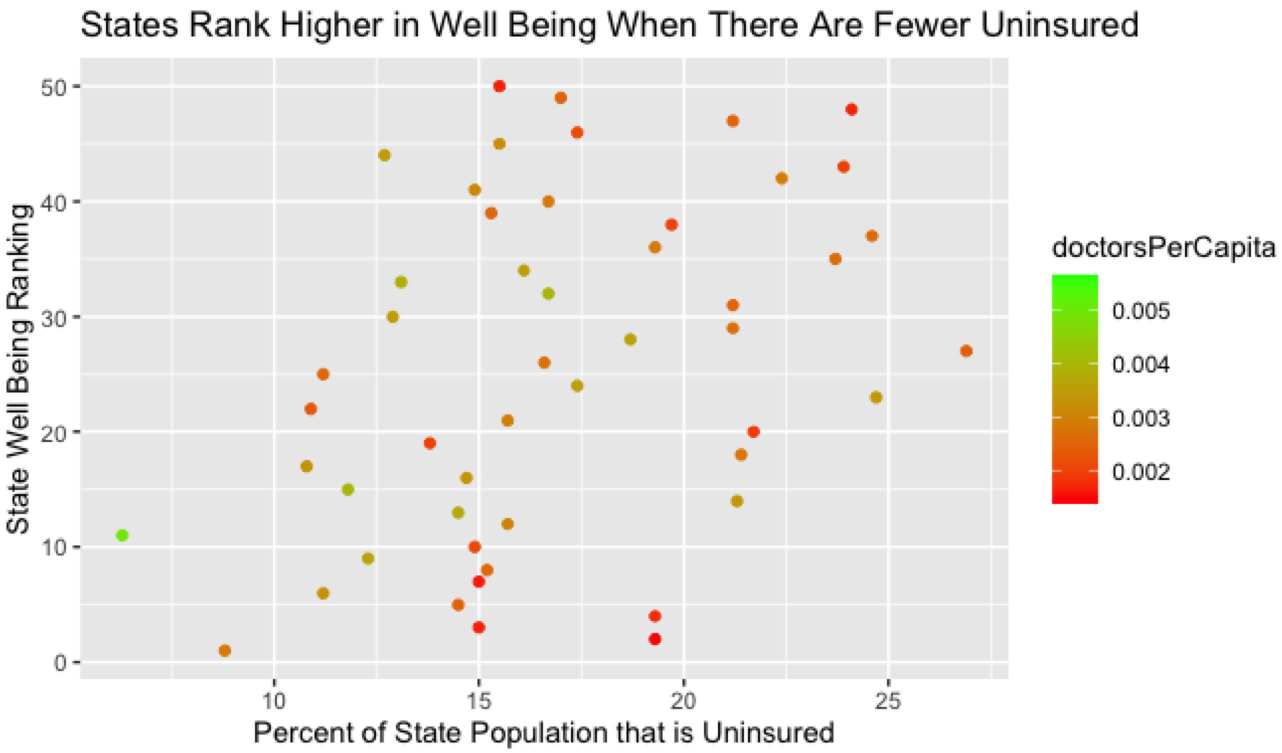
- ggplot will automatically pick a sequential color scale when we map a quantitative variable to color

```
ggplot(data=cspp_data,
      mapping=aes(x=percentuninsured, y=wellbeing))+
  geom_point(aes(color=doctorsPerCapita)+
labs(title = 'States Rank Higher in Well Being When There Are Fewer Uninsured',
         x = 'Percent of State Population that is Uninsured',
         y = 'State Well Being Ranking' )
```



Color Scales

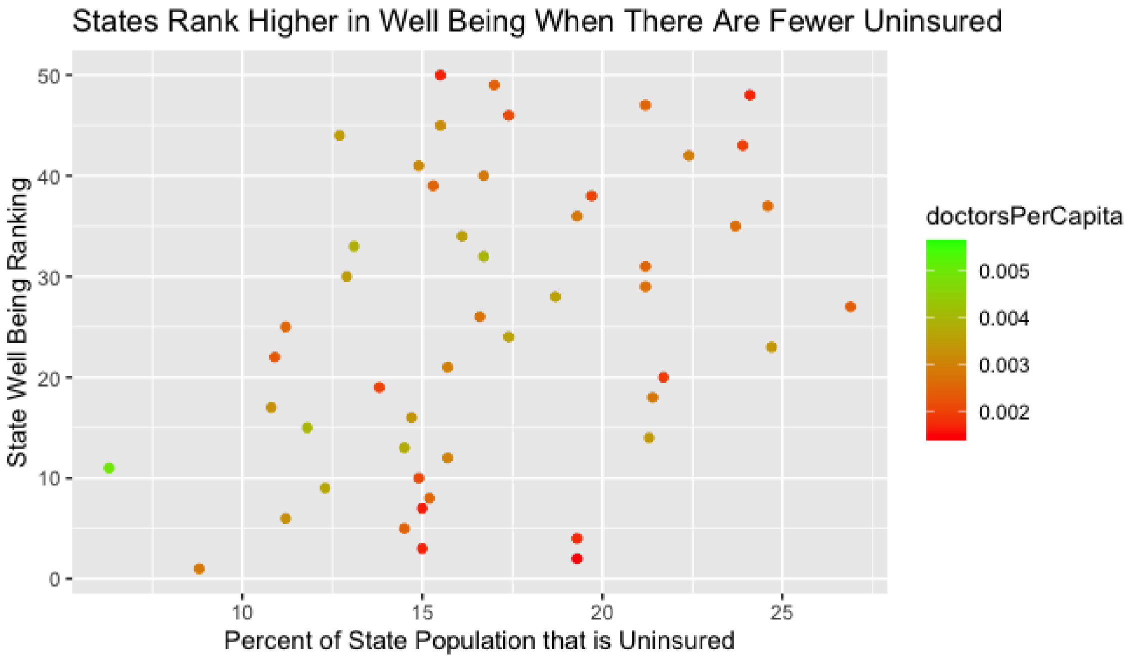
- We can adjust the color scales for quantitative variables using scale_color_gradient()
- You can set the colors on the high and low ends of the scale:
  - red
  - green



Color Scales

- We can adjust the color scale manually by picking the colors in the high and low end
- the red-green color scale allows us to convey whether a number is “good” or “bad”

```
ggplot(data=cspp_data,
     mapping=aes(x=percentuninsured, y=wellbeing))+
  geom_point(aes(color=doctorsPerCapita)+
labs(title = 'States Rank Higher in Well Being When There Are
Fewer Uninsured',
       x = 'Percent of State Population that is Uninsured',
       y = 'State Well Being Ranking' )+
geom_label_repel(aes(label=st )+
scale_color_gradient(low="red",  high="green”)
```



Class Exercise

- Create a scatter plot with doctors per capita on the x-axis and well being on the y axis
- color the points with percent uninsured using a green and red gradient (green low, red high)

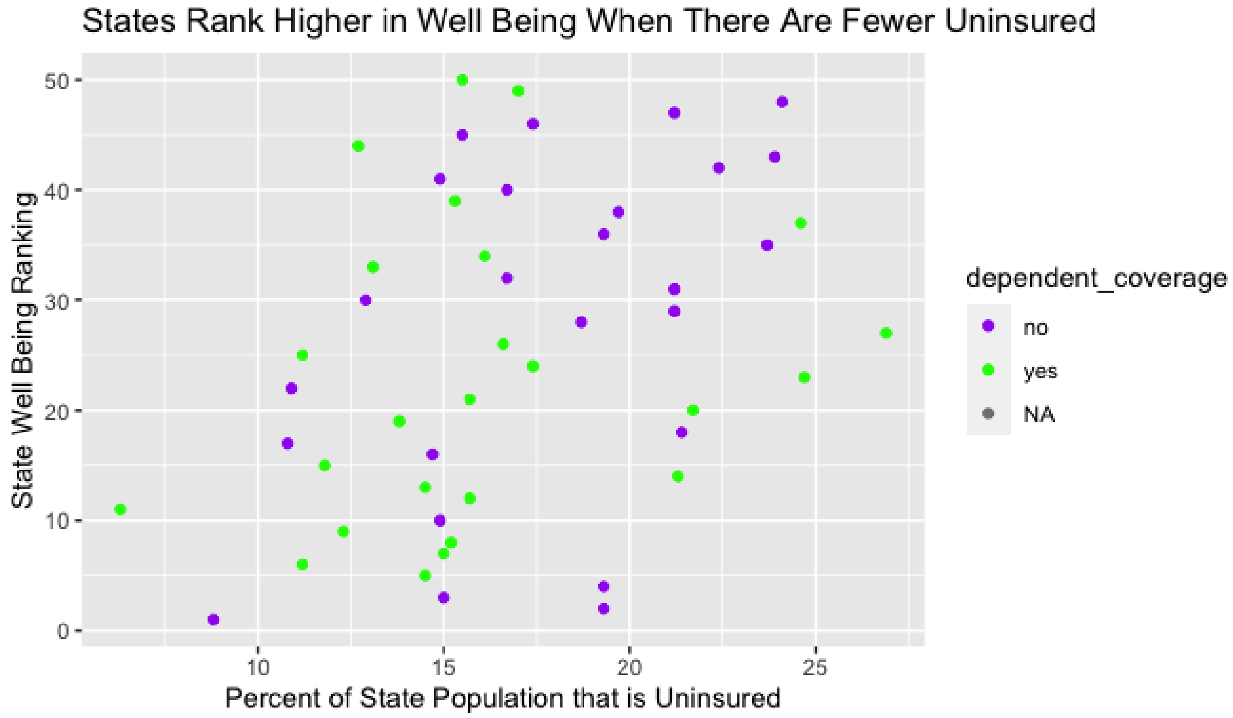
Color Scales for Categorical Variables

- scale_color_manual(): manually create color scale
- scale_color_brewer(): use a ColorBrewer palette

Manually Select Color Scale

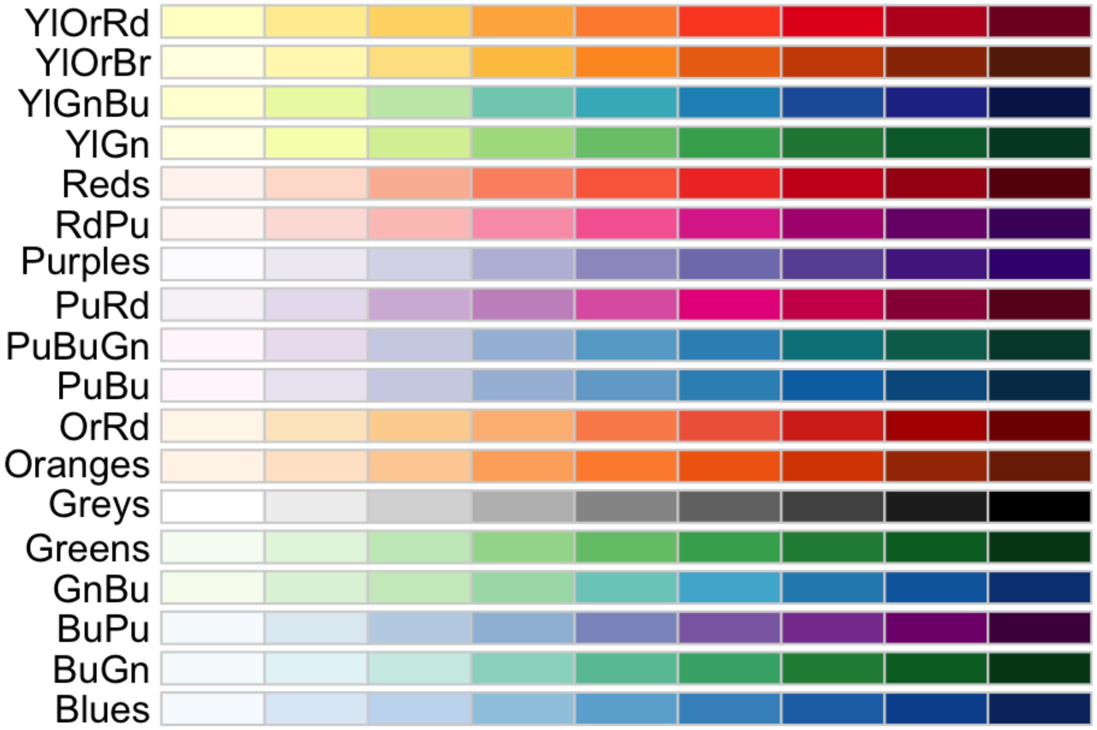
- Now the legend shows the levels of the dependent_coverage factor variable
- It assigns a color to each level

```
ggplot(data=cspp_data,
 aes(x=percentuninsured, y=wellbeing))+
 geom_point(aes( color=dependent_coverage)) +
labs(title = 'States Rank Higher in Well Being When There
Are Fewer Uninsured',
       x = 'Percent of State Population that is Uninsured',
       y = 'State Well Being Ranking' ) +
scale_color_manual(values= c("purple", "green"))
```



ColorBrewer Sequential Color Scales

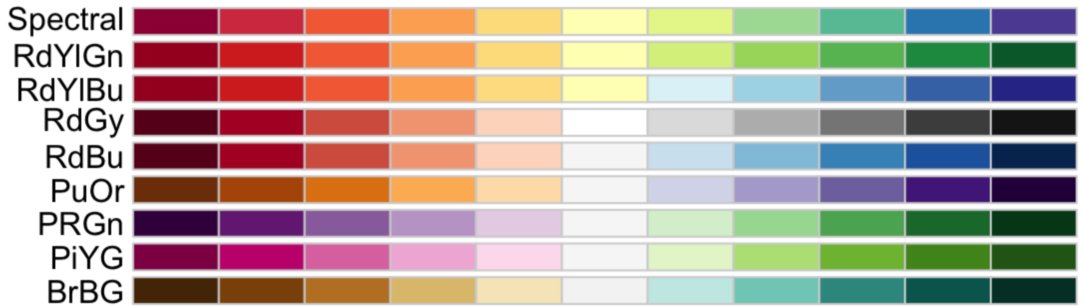
- ColorBrewer provides sets of colors (palettes)
- Sequential palettes are good for ordinal categorical variables
- Educational levels:
  - high school
  - college
  - graduate school



ColorBrewer Diverging Color Scales

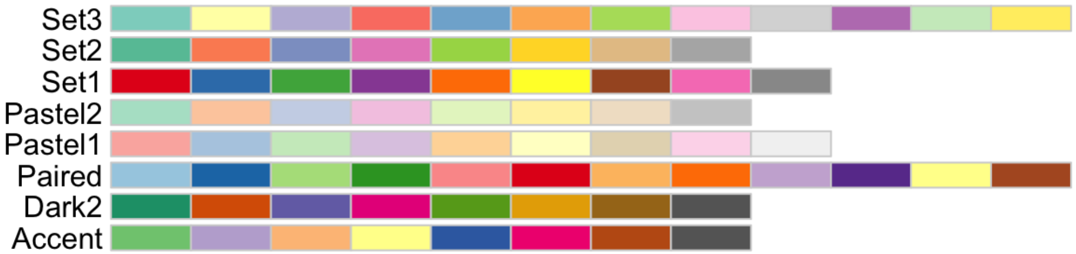
Diverging palettes are good for ordinal categorical variables

- Use this when your values are ordered in two directions relative to a center.
- political affiliation:
  - liberal
  - centrist
  - conservative



ColorBrewer Qualitative Color Scales

- Qualitative (nominal) palettes are good for categorical Variables whose values have no ordering.
- Major:
  - Economics
  - Business
  - Statistics

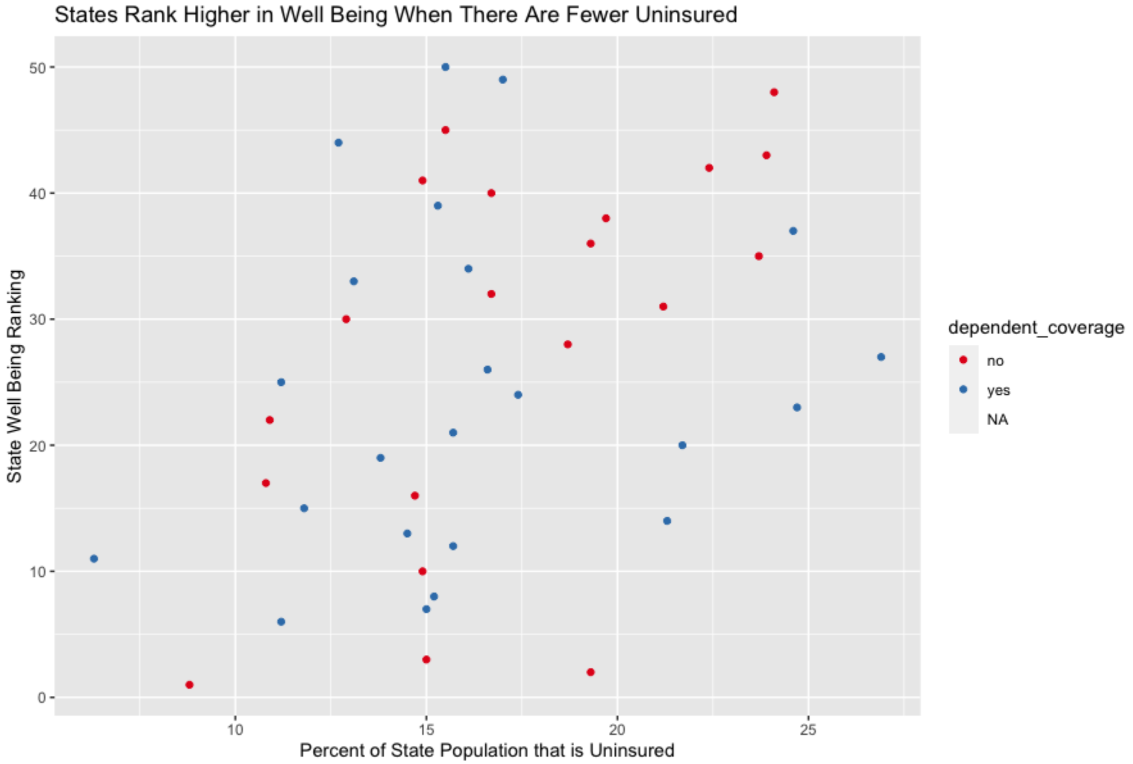


https://pollev.com/vsovero

Palette Example

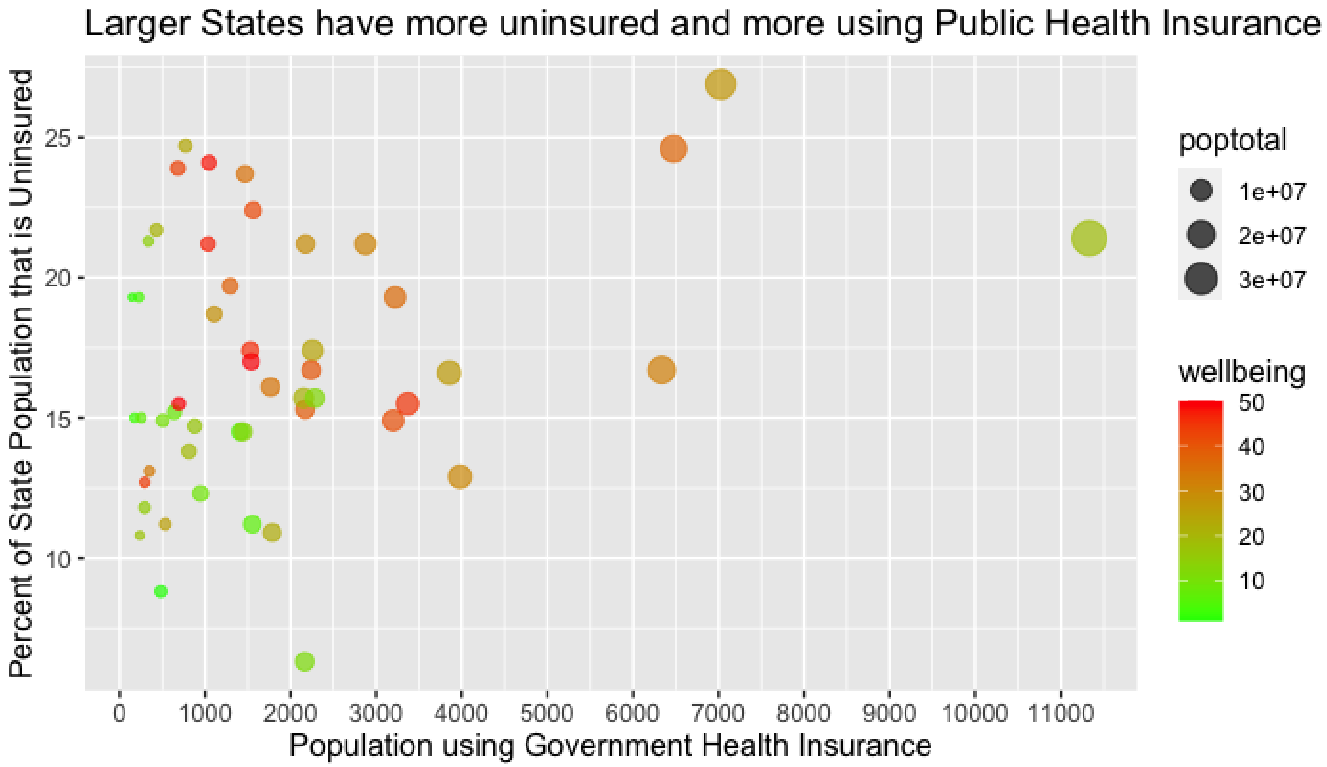
- When we select a palette, ggplot will take colors in the order in which they are listed on the palette
- first color in Set1 is red, the second color is blue

```
ggplot(data=cspp_data,
 aes(x=percentuninsured, y=wellbeing))+
 geom_point(aes( color=dependent_coverage)) +
labs(title = 'States Rank Higher in Well Being When There Are Fewer Uninsured',
       x = 'Percent of State Population that is Uninsured',
       y = 'State Well Being Ranking' ) +
scale_color_brewer(palette= "Set1")
```



Use Themes

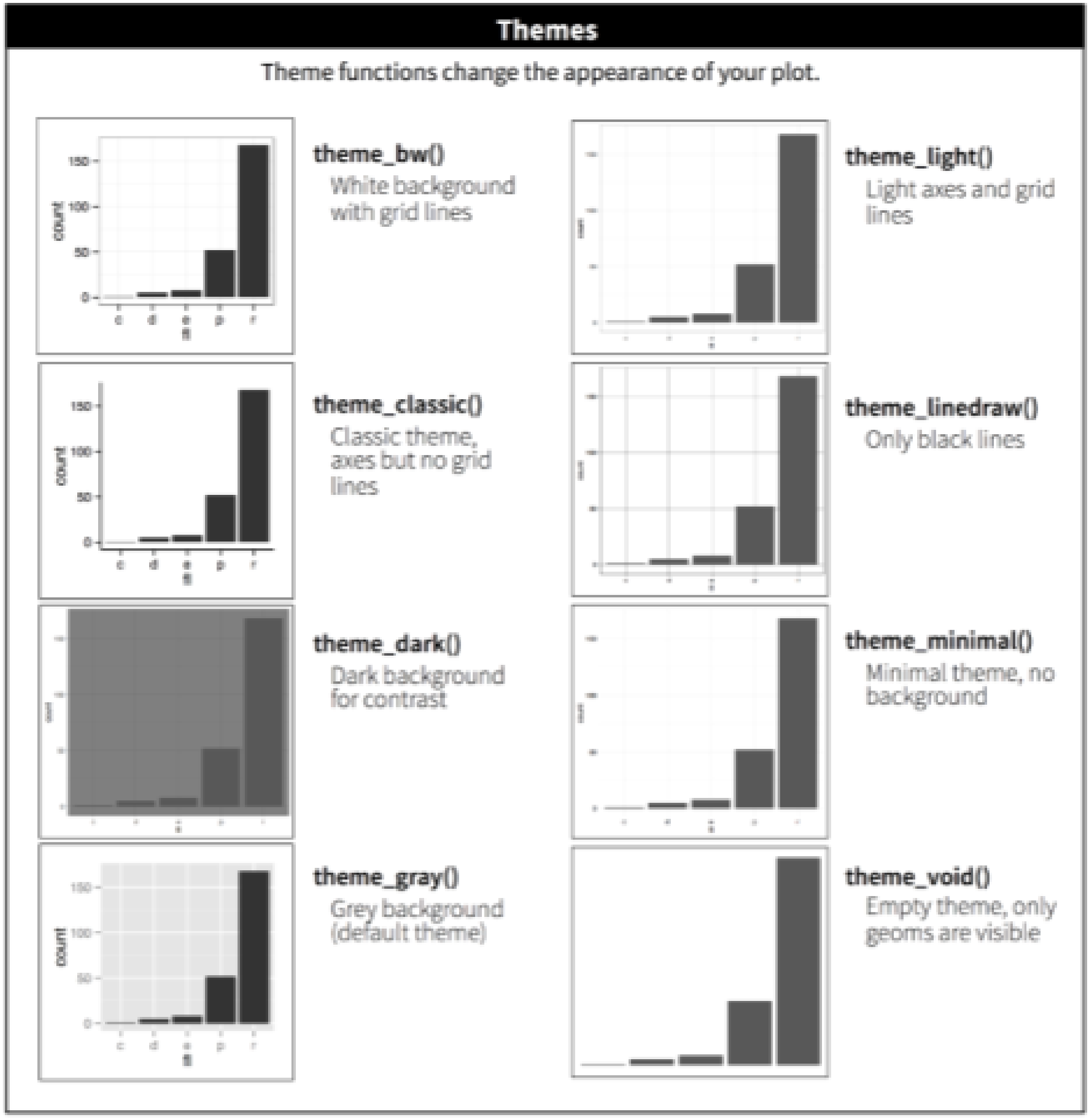
- Themes control the non-data settings of the plot
- **theme**() allows you to make adjustments to:
  - font size
  - legend position



Use Themes

- You can also use themes to change the grid settings:
  - Ex: **theme_bw**()
- Apply theme using **+** operator

https://ggplot2.tidyverse.org/reference/ggtheme.html



Themes

Theme functions change the appearance of your plot.

theme_bw() White background with grid lines

theme_light() Light axes and grid lines

theme_classic() Classic theme, axes but no grid lines

theme_linedraw() Only black lines

theme_dark() Dark background for contrast

theme_minimal() Minimal theme, no background

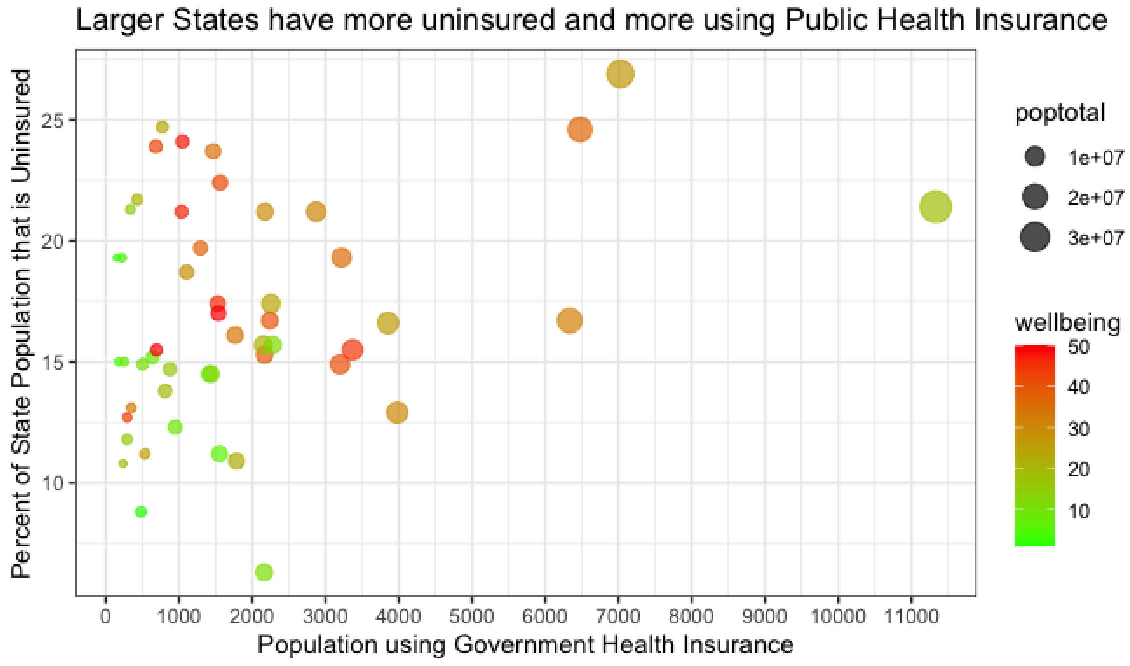
theme_gray() Grey background (default theme)

theme_void() Empty theme, only geoms are visible

Figure 28.3: The eight themes built-in to ggplot2.

Use Themes

```
 ggplot(data=cspp_data,
  aes(x=popgovhealthins, y=percentuninsured))+
  geom_point(aes( color=wellbeing, size=poptotal),
 alpha=.7) +
labs(title = 'Larger States have more uninsured and
more using Public Health Insurance',
        x = 'Population using Government Health
Insurance',
        y = Percent of State Population that is
Uninsured' ) +
scale_color_gradient(low= "green", high= "red", )+
theme(text=element_text(size=12,  family="Arial")) +
theme_bw()
```


Larger States have more uninsured and more using Public Health Insurance
Percent of State Population that is Uninsured
Population using Government Health Insurance
poptotal
1e+07
2e+07
3e+07
wellbeing