431 Class 04

Wyatt P. Bensken

2021-09-02

Welcome

Today's Agenda

- Review Box plots, Scatterplots, and Loess smooth curves
  - IMS Chapter 5
- Hands on with R to explore data
- Walk through two figures
- Question(s)

Upcoming Due Dates

- **Lab 01 is due Monday, 2021-09-06, at 9PM**
- As always, see our course website for the most recent course updates
  - thomaselove.github.io/431

A Note

- It’s okay if you don’t feel completely comfortable with R and building the visualizations we’ll work on today!
- The goal of today is just to get hands on with R and some data to start building those foundational coding skills
- Be patient with yourself as you learn and don’t be afraid to ask questions

Introduction

Important Visualizations

Box plots

- Summarizes a dataset with 5 statistics, while identifying outliers
  - Median, Interquartile Range (IQR), Range
  - Outliers are generally marked as a point and are generally 1.5x IQR

Scatterplots

- Used to visualize two numerical variables
- Each point is an observation
- Useful in assessing relationship between variables, and the trend

Loess smooth curves

- We can fit a Loess smooth curve to the data, which can help reveal trends in the data which are not well estimated with a straight line.

The data we'll use

- The ggplot2 package contains the `midwest` dataset

```
 [1] "area"                 "category"
 [3] "county"               "inmetro"
 [5] "percadultpoverty"     "percamerindan"
 [7] "percasian"            "percbelowpoverty"
 [9] "percblack"            "percchildbelowpovert"
[11] "percelderlypoverty"   "perchsd"
[13] "percollege"           "percother"
[15] "percpovertyknown"     "percprof"
[17] "percwhite"            "PID"
[19] "popadults"            "popamerindian"
[21] "popasian"             "popblack"
[23] "popdensity"           "popother"
[25] "poppovertyknown"      "poptotal"
[27] "popwhite"             "state"
```

Working with the data

The variables we are interested in

- In this in-class work we are interested in the following variables:
    - `percbelowpoverty`, the percent of people below the poverty line
    - `percollege`, the percent who are college educated
    - `county`, the county name

The tasks we'll accomplish

1. Load and Explore the Data
2. Look at Cuyahoga county (where we are now)
3. Make a boxplot
4. Make a scatterplot
5. Add a Loess smooth to our scatterplot

R Markdown (.Rmd) Template

- There is a .Rmd (R Markdown) template available on today's README and and the Data Downloads page
- This is a template which you should download and save somewhere on your computer for today's activity
  - **Please follow the instructions provided, specific to your operating system, to download the template**
- Note: After Lab 01, all of your Labs will be completed using R Markdown. We have provided templates for Lab 02 and Lab 03

Task 1. Load and Explore the Data

Task 1. Load and Explore the Data

Task 1a. Load the data

- You'll first want to load the tidyverse package we'll need, by running the below code

```r
library(tidyverse)
```

- Next we'll want to load the midwest data into our environment (this isn't necessary, but makes things a bit more intuitive)

```r
midwest <- ggplot2::midwest
```

Task 1b. Learn about the dataset

- By running the below code, we can open up (in our help tab) the documentation for the data

```r
?midwest
```

Task 2. Look at Cuyahoga County

Task 2. Look at Cuyahoga County

- To look at Cuyahoga County we'll need to `filter()` our data to just our observation of interest
  - We'll also `select()` only those two variables we'd like to look at
  - This is a good use of the pipe %>%
- We know from data documentation that `county` is our variable name, and from looking at the data we can see that the county names are all capitalized.
  - It is important to remember that R is case sensitive!

```
midwest %>%
  filter(county == "CUYAHOGA") %>%
  select(county, percbelowpoverty, percollege)
```

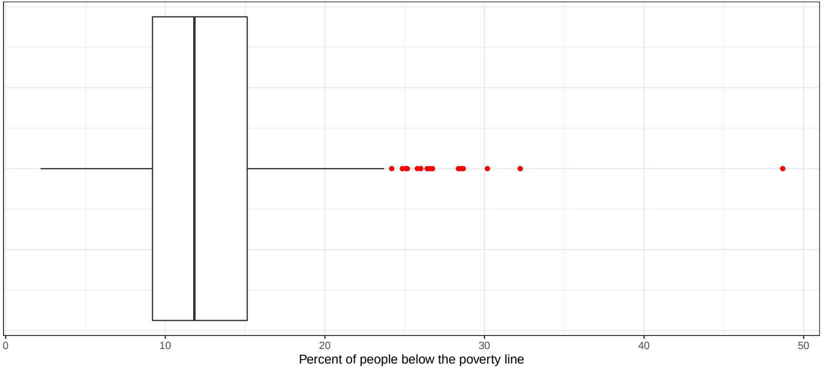
Task 2. Look at Cuyahoga County

```
# A tibble: 1 x 3
  county   percbelowpoverty percollege
  <chr>               <dbl>      <dbl>
1 CUYAHOGA             13.8       25.1
```

Task 3. Make a boxplot

Task 3. Make a boxplot

- Our goal will be to make a boxplot, of `percbelowpoverty` which looks like this



Boxplot of poverty in Midwest counties

These data come from the midwest package in ggplot2

Task 3. Make a boxplot

- We'll work through the code step by step, but the complete code looks like this:

```r
ggplot(data = midwest, aes(x = percbelowpoverty)) +
  geom_boxplot(outlier.color = "red") +
  labs(x = "Percent of people below the poverty line",
       title = "Boxplot of poverty in Midwest counties",
       subtitle = "These data come from the
                      midwest package in ggplot2") +
  theme_bw() +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())
```

Task 3. Make a boxplot

Step 1

- First we'll use ggplot to set our dataset and aesthetics (abbreviated "aes")
    - This code won't run anything, until we add our the geom we would like

```
ggplot(data = midwest, aes(x = percbelowpoverty))
```

Task 3. Make a boxplot

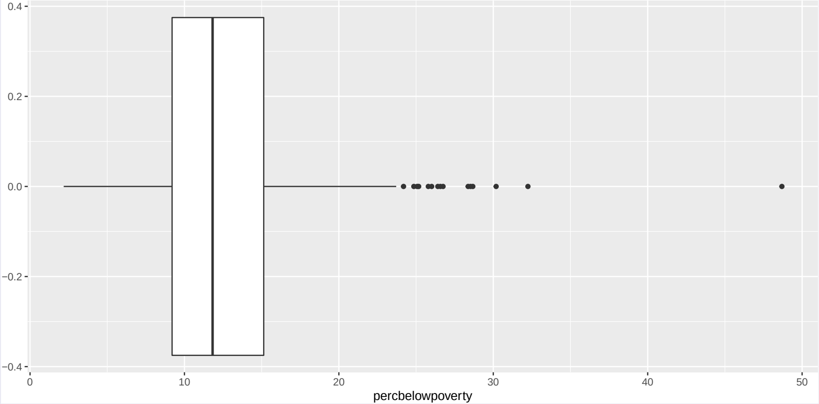
Step 2

- Now we can add (note that we use + here and not the pipe) that we would like the boxplot geom.

```
ggplot(data = midwest, aes(x = percbelowpoverty)) +
  geom_boxplot()
```

Task 3. Make a boxplot

Step 2



Task 3. Make a boxplot

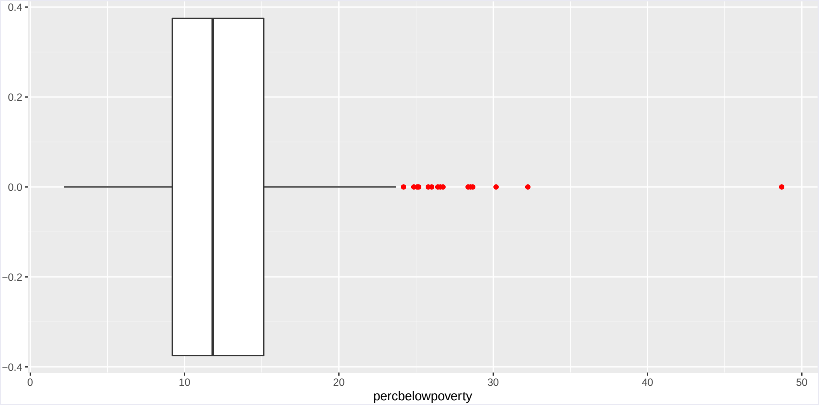
Step 3

- Each geom has a number of options specific to that type of figure, here we'd like to color our outliers red.

```r
ggplot(data = midwest, aes(x = percbelowpoverty)) +
  geom_boxplot(outlier.color = "red")
```

Task 3. Make a boxplot

Step 3



Task 3. Make a boxplot

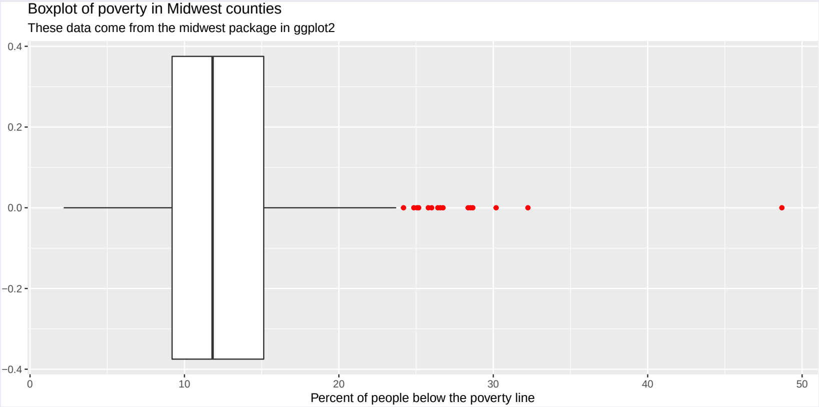
Step 4

- **In this course no figure is complete without appropriate axes labels and titles**
- We can add (again use +) these using the `labs()` statement where we have x, title, and subtitle
    - There are numerous other options such as y, subtitle, and caption, that are available but that we don't use here.

```
ggplot(data = midwest, aes(x = percbelowpoverty)) +
  geom_boxplot(outlier.color = "red") +
  labs(x = "Percent of people below the poverty line",
       title = "Boxplot of poverty in Midwest counties",
       subtitle = "These data come from the
                      midwest package in ggplot2")
```

Task 3. Make a boxplot

Step 4



Task 3. Make a boxplot

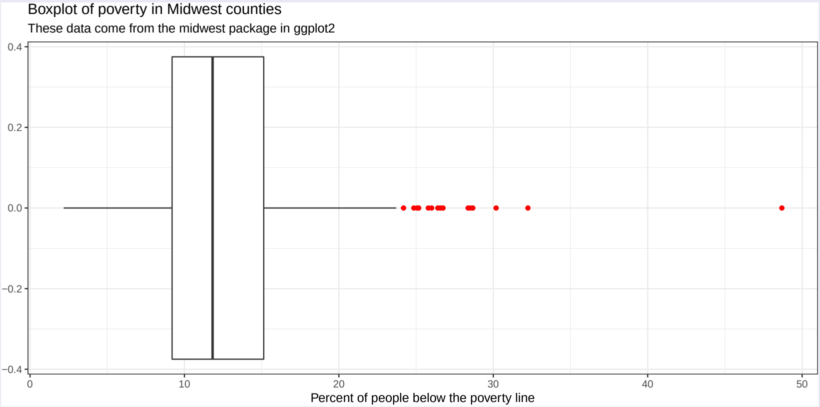
Step 5

- I'd like to get rid of that odd gray background which is, somewhat annoyingly, the default
- We can do this using a theme - `theme_bw()`

```
ggplot(data = midwest, aes(x = percbelowpoverty)) +
  geom_boxplot(outlier.color = "red") +
  labs(x = "Percent of people below the poverty line",
       title = "Boxplot of poverty in Midwest counties",
       subtitle = "These data come from the
                      midwest package in ggplot2") +
  theme_bw()
```

Task 3. Make a boxplot

Step 5



Task 3. Make a boxplot

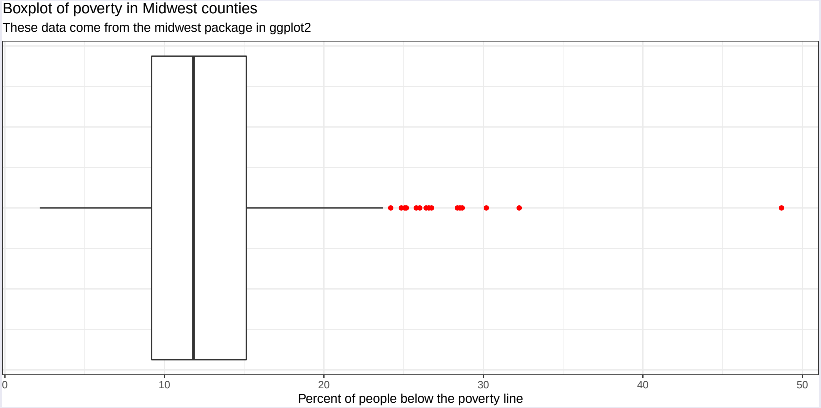
Step 6 - the final figure

- Finally, in this boxplot the y-axis text and tick marks are not informative or helpful, so we'd like to remove them
- The theme() command has a whole host of options, but here we'll just use 2.

```
ggplot(data = midwest, aes(x = percbelowpoverty)) +
  geom_boxplot(outlier.color = "red") +
  labs(x = "Percent of people below the poverty line",
       title = "Boxplot of poverty in Midwest counties",
       subtitle = "These data come from the
                      midwest package in ggplot2") +
  theme_bw() +
  theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank())
```

Task 3. Make a boxplot

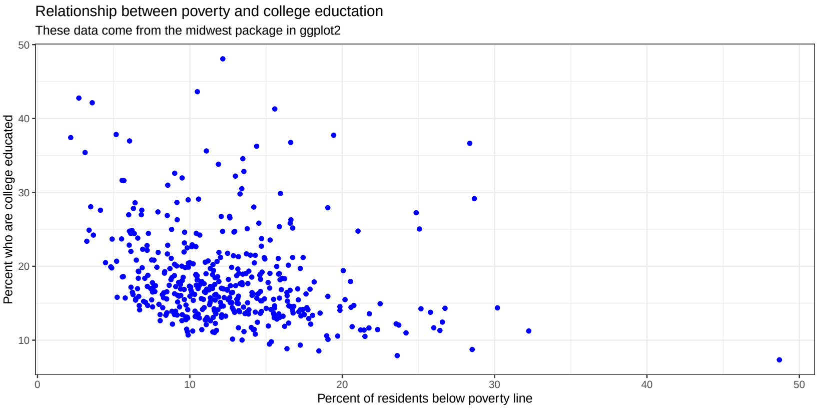
Step 6 - the final figure



Task 4. Make a scatterplot

Task 4. Make a scatterplot

- Now we'd like to make a scatterplot with `percbelowpoverty` on our x-axis and `percollege` on our y-xis



Task 4. Make a scatterplot

Step 1

- Each figure we'll make in R using the `ggplot2` package will share substantial syntax, including the first step
- Here, however we assign not just an x aesthetic but a y aesthetic as well

```
ggplot(data = midwest, aes(x = percbelowpoverty,
                           y = percollege)) +
```

Task 4. Make a scatterplot

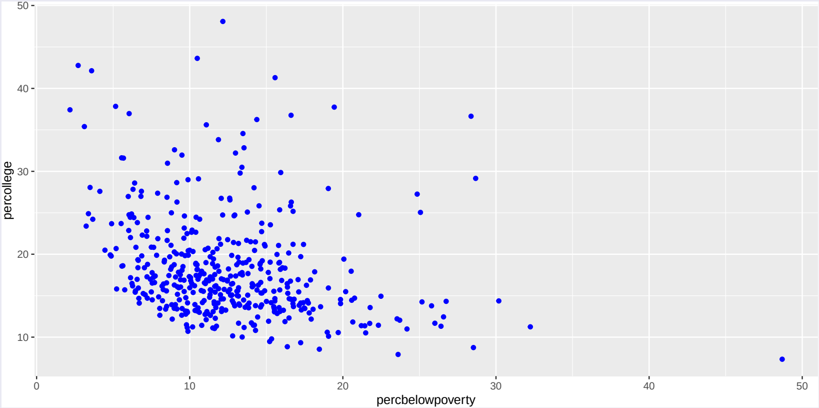
Step 2

- In this example, we want to add a point geom, which makes a scatterplot when we have properly assigned our x and y variables in the aesthetic
- We can again take advantage of the options within our geom to make the points a specific color.

```
ggplot(data = midwest, aes(x = percbelowpoverty,
                           y = percollege)) +
  geom_point(color = "blue")
```

Task 4. Make a scatterplot

Step 2



Task 4. Make a scatterplot

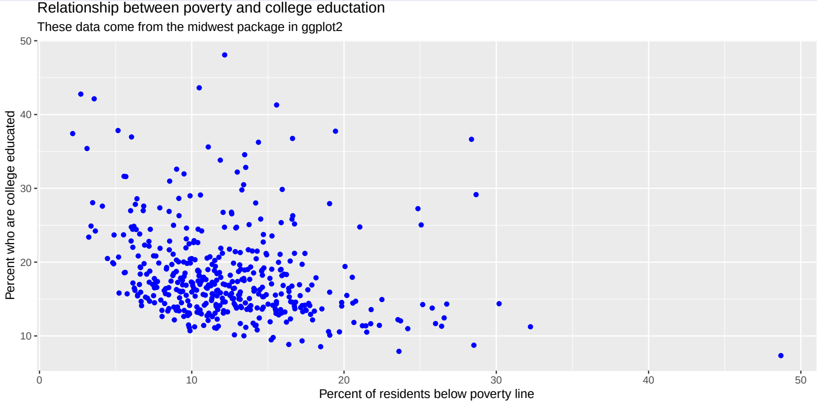
Step 3

- As in our boxplot, we **must** add appropriate axis legends

```
ggplot(data = midwest, aes(x = percbelowpoverty,
                           y = percollege)) +
  geom_point(color = "blue") +
  labs(x = "Percent of residents below poverty line",
       y = "Percent who are college educated",
       title = "Relationship between poverty
                  and college eductation",
       subtitle = "These data come from the
                      midwest package in ggplot2")
```

Task 4. Make a scatterplot

Step 3



Task 4. Make a scatterplot

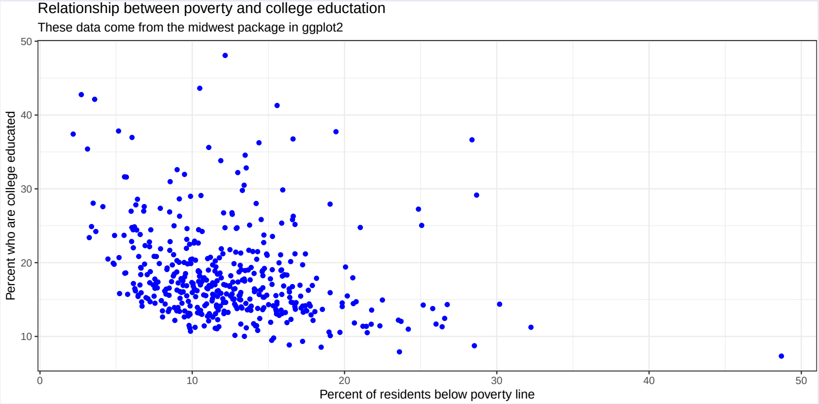
Step 4 - the final figure

- Finally, we'd like to again use our theme_bw() to get rid of that gray background

```
ggplot(data = midwest, aes(x = percbelowpoverty,
                           y = percollege)) +
  geom_point(color = "blue") +
  labs(x = "Percent of residents below poverty line",
       y = "Percent who are college educated",
       title = "Relationship between poverty
                 and college eductation",
       subtitle = "These data come from the
                   midwest package in ggplot2") +
  theme_bw()
```

Task 4. Make a scatterplot

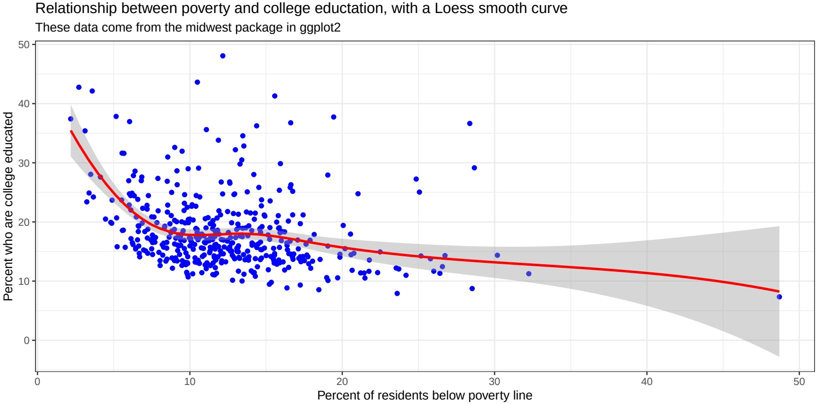
Step 4 - the final figure



Task 5. Add a Loess smooth

Task 5. Add a Loess smooth

- We'd now like to add a Loess smooth curve to our scatterplot to examine what type of relationship is fit with a smooth line.



Relationship between poverty and college eductation, with a Loess smooth curve

These data come from the midwest package in ggplot2

Task 5. Add a Loess smooth

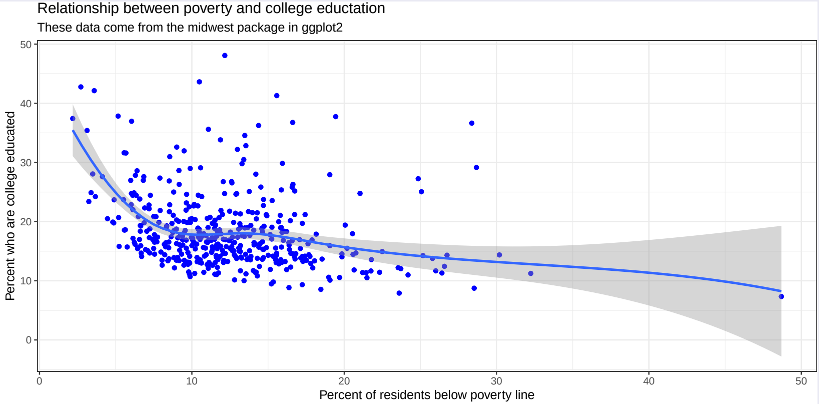
Step 1

- One of the most powerful parts of ggplot and R is the ability to layer geoms
- We'll want to make sure to specify that we want a Loess curve in our method

```
ggplot(data = midwest, aes(x = percbelowpoverty,
                           y = percollege)) +
  geom_point(color = "blue") +
  geom_smooth(method = "loess", formula = y ~ x) +
  labs(x = "Percent of residents below poverty line",
       y = "Percent who are college educated",
       title = "Relationship between poverty
                  and college eductation",
       subtitle = "These data come from the
                      midwest package in ggplot2") +
  theme_bw()
```

Task 5. Add a Loess smooth

Step 1



Task 5. Add a Loess smooth

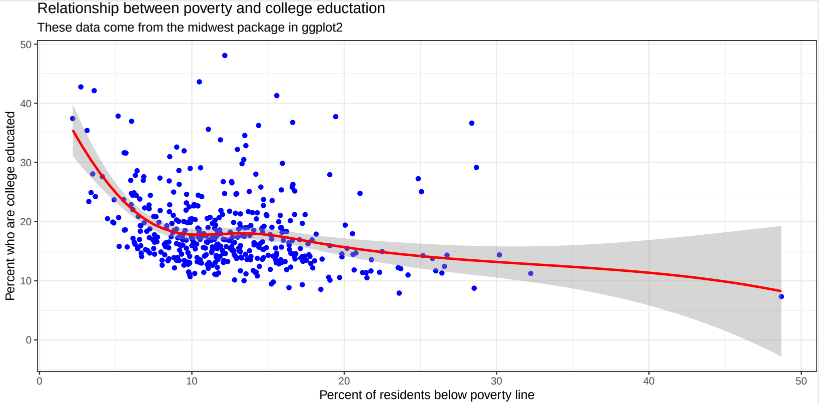
Step 2

- We can easily change the color of our Loess curve, to better differentiate it from the points.

```
ggplot(data = midwest, aes(x = percbelowpoverty,
                           y = percollege)) +
  geom_point(color = "blue") +
  geom_smooth(method = "loess", formula = y ~ x,
                                color = "red") +
  labs(x = "Percent of residents below poverty line",
       y = "Percent who are college educated",
       title = "Relationship between poverty
                  and college eductation",
       subtitle = "These data come from the
                      midwest package in ggplot2") +
  theme_bw()
```

Task 5. Add a Loess smooth

Step 2



Task 5. Add a Loess smooth

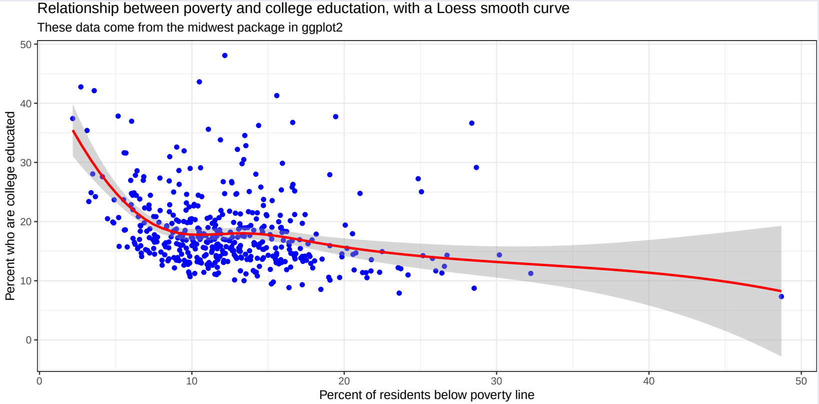
Step 3

- Finally, we'll want to update our title to reflect the new figure

```
ggplot(data = midwest, aes(x = percbelowpoverty,
                           y = percollege)) +
  geom_point(color = "blue") +
  geom_smooth(method = "loess", formula = y ~ x,
                                color = "red") +
  labs(x = "Percent of residents below poverty line",
       y = "Percent who are college educated",
       title = "Relationship between poverty
                  and college eductation,
                  with a Loess smooth curve",
       subtitle = "These data come from the
                      midwest package in ggplot2") +
  theme_bw()
```

Task 5. Add a Loess smooth

Step 3



Knit the file

- At the top of the R Markdown you should see a small button that says “Knit”. Click this.
  - This turns your R Markdown into an HTML file
  - Again, this will be how you will complete Lab 02 - Lab 07

Questions and Discussion