

# Lab 1 - Data visualization

Alex LaTrenta

## Load Packages

```
library(tidyverse)
```

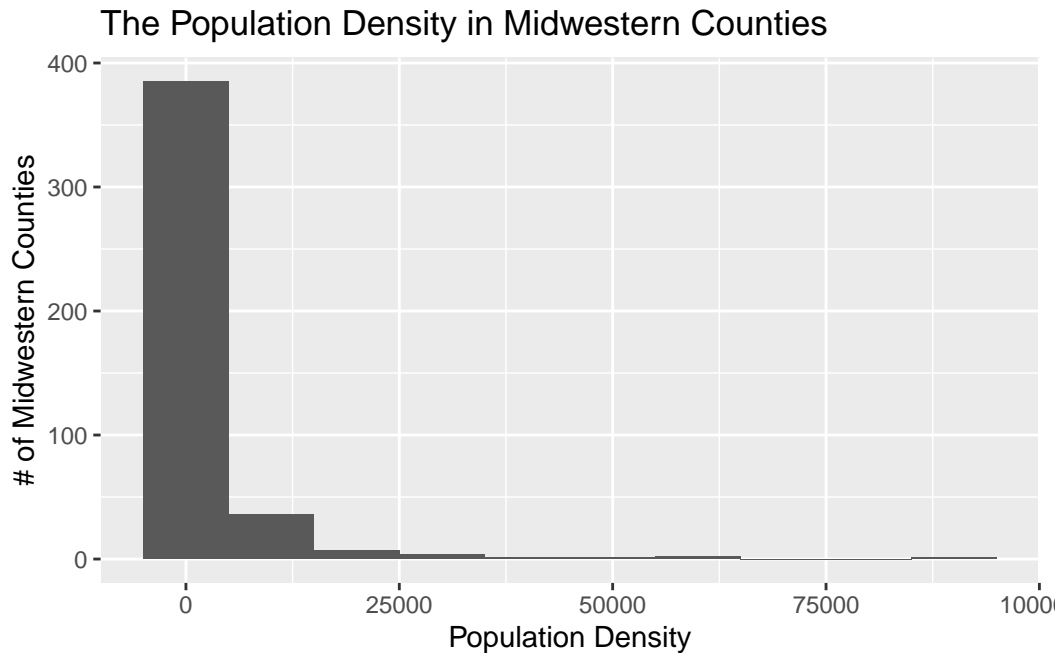
```
Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
had status 1
```

## Exercise 1

(Type your answer to Exercise 1 here. Add code chunks as needed. Don't forget to label your code chunk. Do not use spaces in code chunk labels.)

Problem 1:

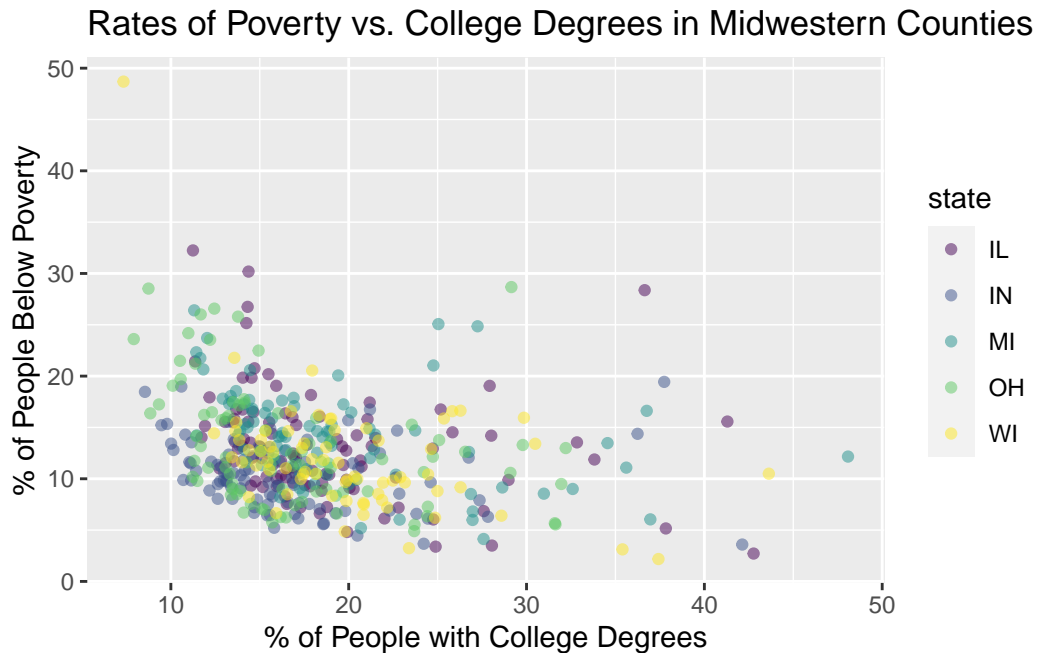
```
ggplot(data = midwest,
       aes(x = popdensity)) +
geom_histogram(binwidth = 10000) +
labs(x = "Population Density",
     y = "# of Midwestern Counties",
     title = "The Population Density in Midwestern Counties")
```



The graph is very right-skewed with most counties having a population density significantly lower than 25,000 in the low thousands. There are some outliers with medium (~60,000) and large (~85,000) population densities which appear as thin bars on the graph's x axis.

## Exercise 2

```
ggplot(data = midwest,
       aes(x = percollege, y = percbelowpoverty, color = state)) +
geom_point(alpha = .5) +
scale_color_viridis_d() +
labs(x = "% of People with College Degrees",
     y = "% of People Below Poverty",
     title = "Rates of Poverty vs. College Degrees in Midwestern Counties")
```



### Exercise 3

Describe what you observe in the plot from the previous exercise. In your description, include similarities and differences in the patterns across states.

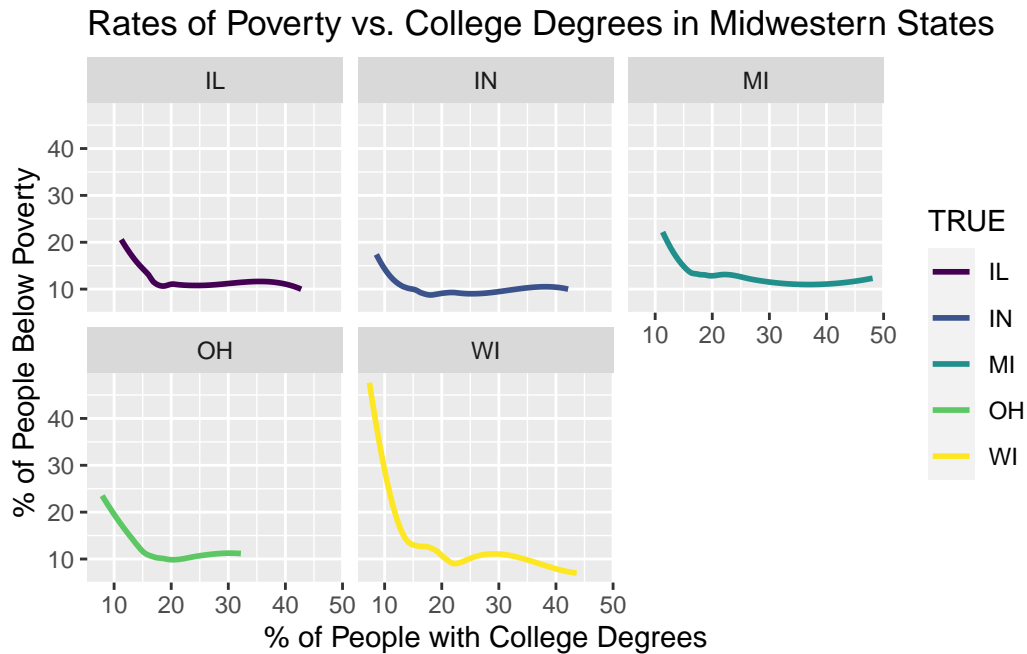
Across most midwestern states, there is a negative linear relationship between the % of people under the poverty line and the % of people with college degrees. This implies that low rates of college degrees may be associated with higher levels of poverty. All states have a large amount of counties with the % College Degrees and % poverty between 10 to 20. OH and MI tend to have outlier counties with low % college degrees and high % poverty despite much variation. WI and IN tend to have more counties with high % college degrees and low % poverty. IL has outlier countries with both of the described characteristics.

### Exercise 4

```
ggplot(data = midwest,
       aes(x = percollege, y = percbelowpoverty, color = state)) +
  geom_smooth(alpha = .5, se = FALSE) +
  scale_color_viridis_d(TRUE) +
  facet_wrap("state") +
  labs(x = "% of People with College Degrees",
```

```
y = "% of People Below Poverty",
title = "Rates of Poverty vs. College Degrees in Midwestern States")
```

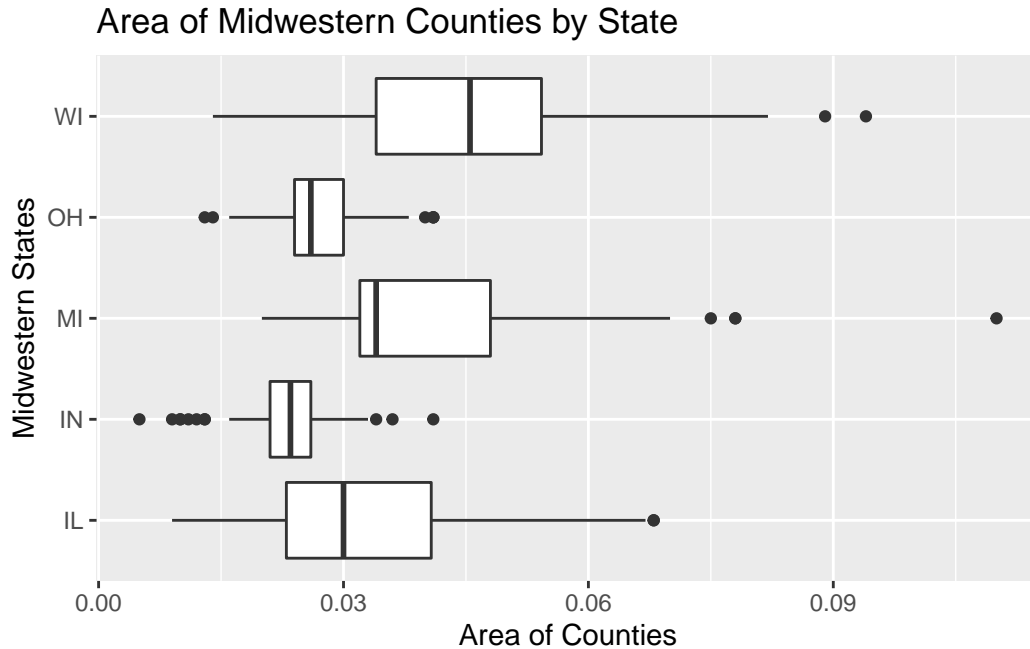
``geom_smooth()`` using method = 'loess' and formula 'y ~ x'



I prefer this plot more because it is easier to visualize the differences between the states when they are separated in a grid. Furthermore, the linear format better summarizes the exact linear trends between the two variables which is visibly clearer than a mass of dots.

## Exercise 5

```
ggplot(data = midwest,
       aes(x = area, y = state)) +
geom_boxplot() +
labs(x = "Area of Counties",
     y = "Midwestern States",
     title = "Area of Midwestern Counties by State")
```



Some states do tend to have counties that are geographically larger than others. WI has the counties with the largest area based on the large median (around .045) and the largest Interquartile Range followed by MI, IL/OH, and then IN with the smallest county area median around .025 and smallest IQR. MI has the single largest county due to the outlier dot above an area of ~.1 which is the larger than any other outlier in this grid.

## Exercise 6

```
midwest <- midwest |>
  mutate(metro = if_else(inmetro == 1, "Yes", "No"))
ggplot(data = midwest,
  aes(x = state, fill = metro)) +
  geom_bar() +
  labs(x = "Midwestern States",
    y = "Counties in a Metro Area",
    title = "How many counties are in a metro area in Midwestern states?")
```

### How many counties are in a metro area in Midwestern states?



Some states do have more counties which are considered metro areas such as OH (with around 50% metro counties) in contrast to WI or IL which have lower proportions of metro counties.

### Exercise 7