Lab 1 - Data visualization

Grace Romero

Load Packages

```
library(tidyverse)
```

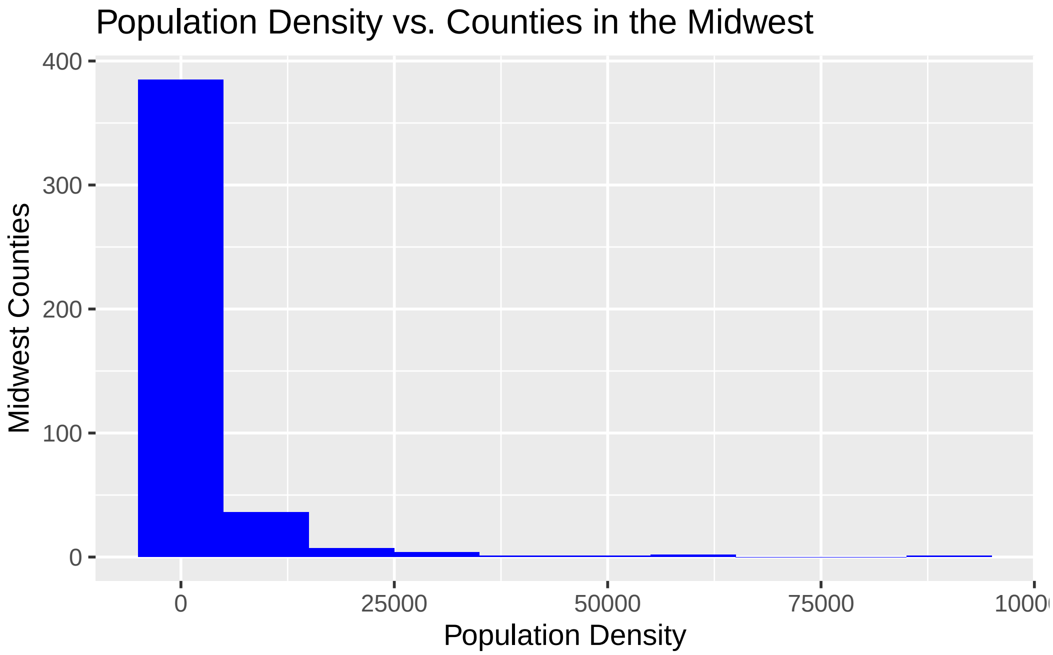
```
Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
had status 1
```

```
library(viridis)
```

Exercise 1

```
#population_density_histogram

ggplot(data = midwest,
       aes(x=popdensity)) +
    geom_histogram(binwidth = 10000, fill = "blue") +
  labs(title= "Population Density vs. Counties in the Midwest",
  x= "Population Density",
  y="Midwest Counties")
```

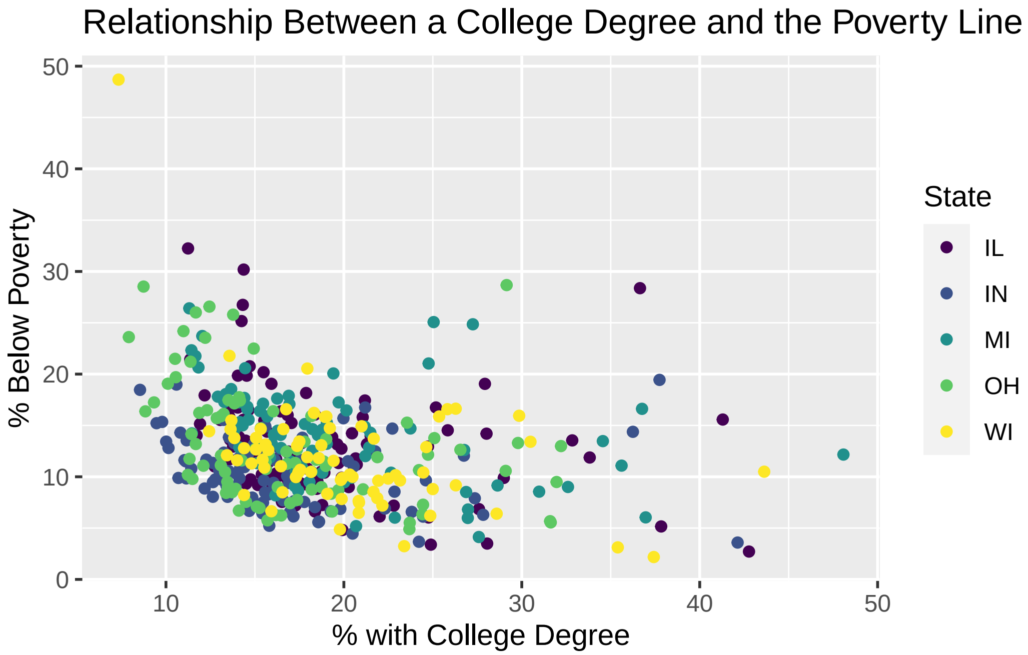


The shape of this is histogram is skewed right. Additionally, there do seem to be outliers because on the tail of the plot, there is a small blue bar indicating that there are data points between x = 75000 and x = 10000.

Exercise 2

```
#college_degree_vs_poverty_scatterplot

ggplot(data = midwest,
       aes(x=percollege, y = percbelowpoverty, color = state)) +
    geom_point() +
  labs (title="Relationship Between a College Degree and the Poverty Line",
  x= "% with College Degree",
  y="% Below Poverty",
  color = "State") + scale_colour_viridis_d()
```



Exercise 3

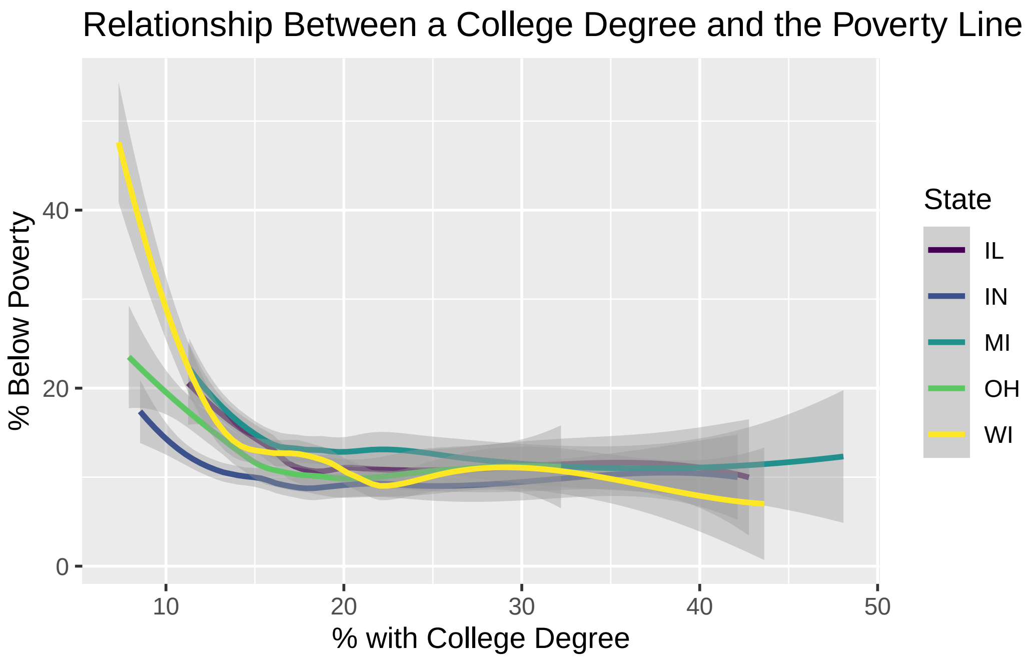
In the data plot from Exercise 2, I observe that as the percentage of people with a college degree increases, the percentage of people below poverty decreases. It seems that in all of the states, these variables are associated. I don't think this relationship is linear because there is a lot of curvature in the trend of the data. It seems that the data from all of the states have similar tendencies, but there are a few distributional differences. For example, the points representative of people in Wisconsin begin higher on the y-axis than the points representative of people in the other three states.

Exercise 4

```
#college_degree_vs_poverty_smooth_curve_fit

ggplot(data = midwest,
       aes(x=percollege, y = percbelowpoverty, color = state, se = FALSE)) +
    geom_smooth() +
  labs (title= "Relationship Between a College Degree and the Poverty Line",
  x= "% with College Degree",
  y="% Below Poverty",
  color = "State") + scale_colour_viridis_d()
```

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

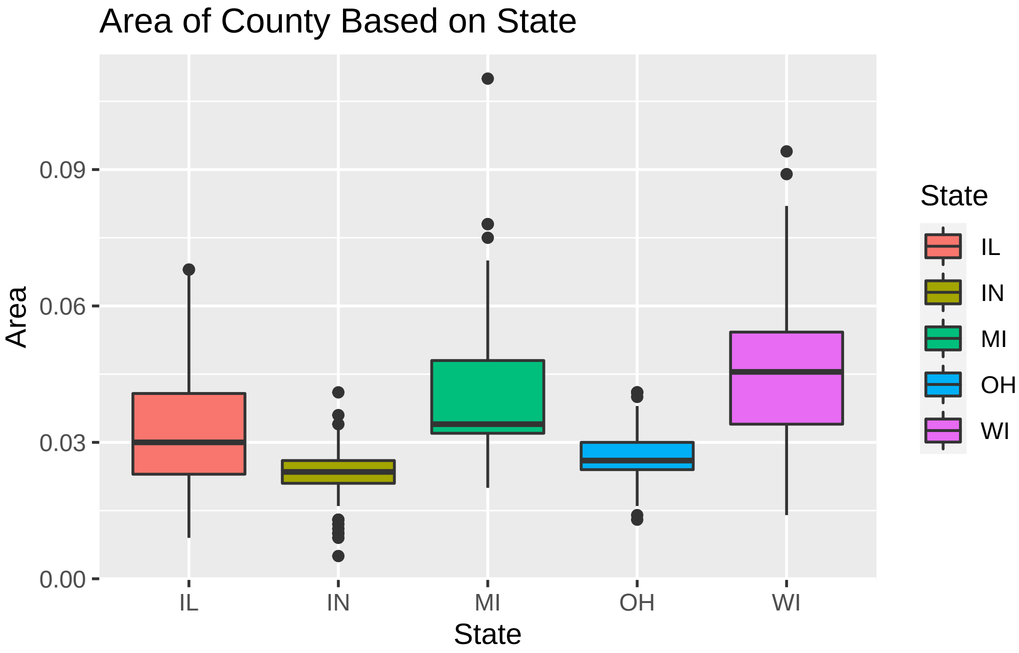


I prefer this plot to the plot from Exercise 3. This is because I can more clearly see the general tendencies of the data in this plot than in the other one. I can also more clearly identify the differences between the states with this plot because it is easier to compare four lines than a large number of dots.

Exercise 5

```
#area_of_county_based_on_state_boxplots

ggplot(data=midwest,
  aes(x= state, y = area, fill = state )) +
  geom_boxplot() +
  labs (title="Area of County Based on State",
  x= "State",
  y="Area", fill = "State")
```



From this plot, I can observe the average size of the counties in each of the states in this study. I can also determine how varied the size of the counties in each state is by looking it the size of the box. The closer together the top and bottom boundaries, the less variety there is. For example, in Indiana, the counties are more similar in size than in Wisconsin.

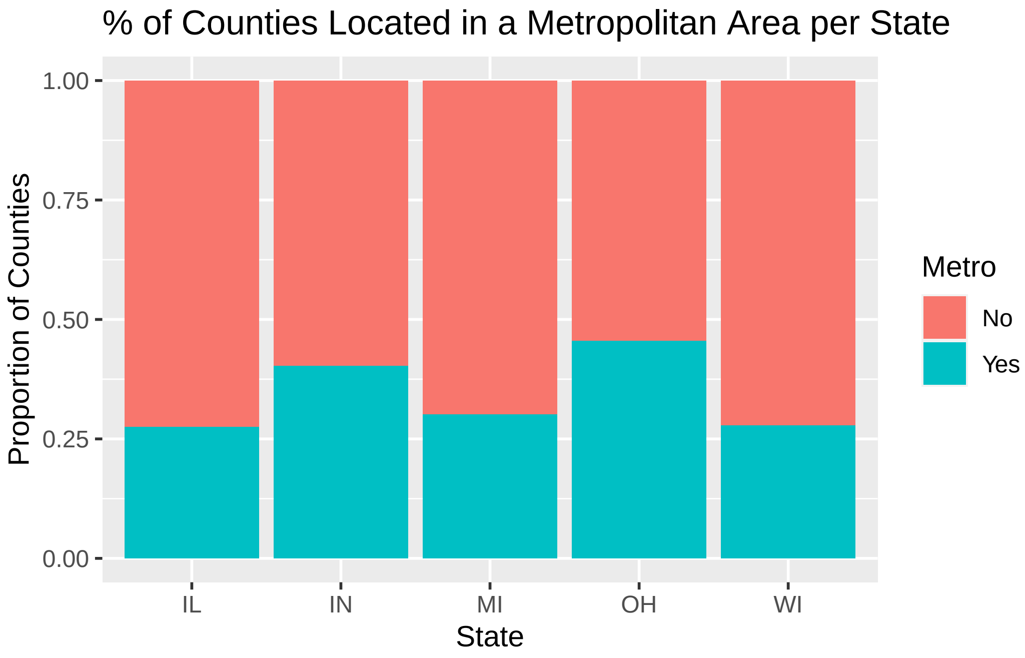
The state with the largest county is Michigan. The dot which extends off of the Michigan plot is an outlier, representing the maximum value of the data. Because it is almost equal to y=0.12, greater than any of the other outliers of the other boxplots, I know that this area is the greatest of all the counties in this entire data set.

Exercise 6

```
#percentage_of_counties_in_metro_area_bar_plot

midwest <- midwest |>
  mutate(metro = if_else(inmetro == 1, "Yes", "No"))

ggplot(data=midwest,
      aes(x = state, fill = metro)) +
      geom_bar(position="fill") +
  labs (title="% of Counties Located in a Metropolitan Area per State",
        x= "State",
        y="Proportion of Counties", fill = "Metro")
```

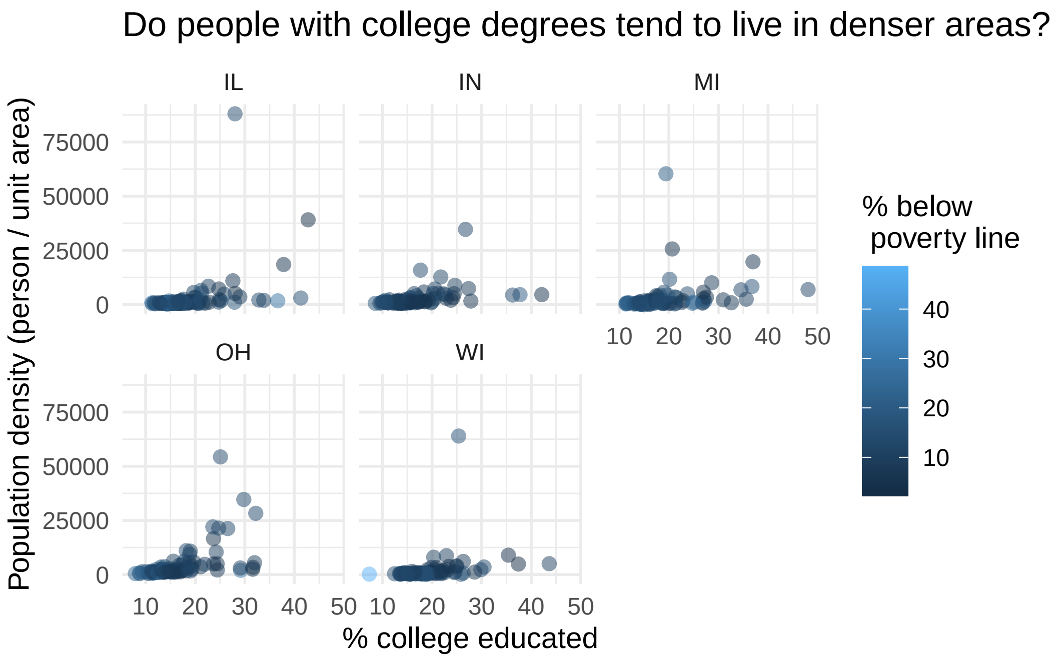


From this plot, I notice that most counties are not located in a metropolitan area in each state in the data set. I also notice that the bars fill the whole column from y = 0 to y = 1. If I removed position = “fill” from the code, however, they do not and the values along the y-axis change. I’m not sure why there’s a difference in doing that.

Exercise 7

```
#recreated_plot

ggplot(data=midwest,
       aes(x = percollege, y = popdensity, color = percbelowpoverty)) +
        geom_point(size = 2, alpha = 0.5, ) +
  labs (title="Do people with college degrees tend to live in denser areas?",
        x= "% college educated",
        y="Population density (person / unit area)",
        color = "% below \n poverty line") +
  facet_wrap("state") + theme_minimal()
```


Do people with college degrees tend to live in denser areas?
IL
IN
MI
OH
WI
Population density (person / unit area)
% college educated
% below
poverty line