Lab 1 - Data visualization

Emanuel Lopez Gallegos

Load Packages

```
library(tidyverse)
```

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

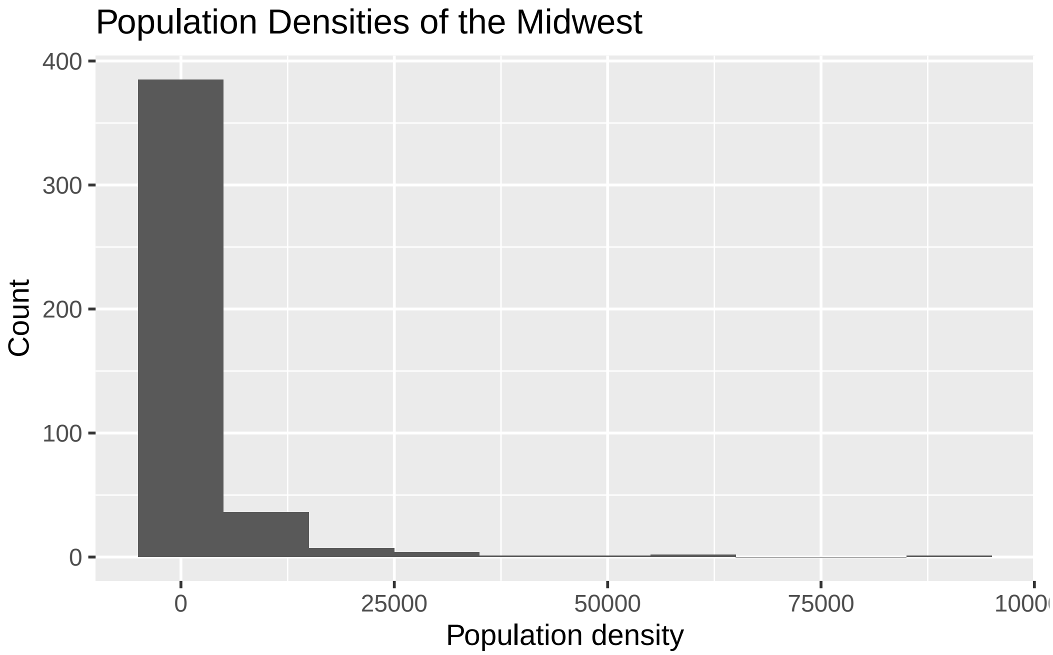
```
library(viridis)
```

```
view(midwest)
```

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.)

```
ggplot(data = midwest,
       mapping = aes(x = popdensity)) +
  geom_histogram(alpha = 1, binwidth = 10000) +
  labs(
    x = "Population density",
    y = "Count",
    title = "Population Densities of the Midwest"
  )
```

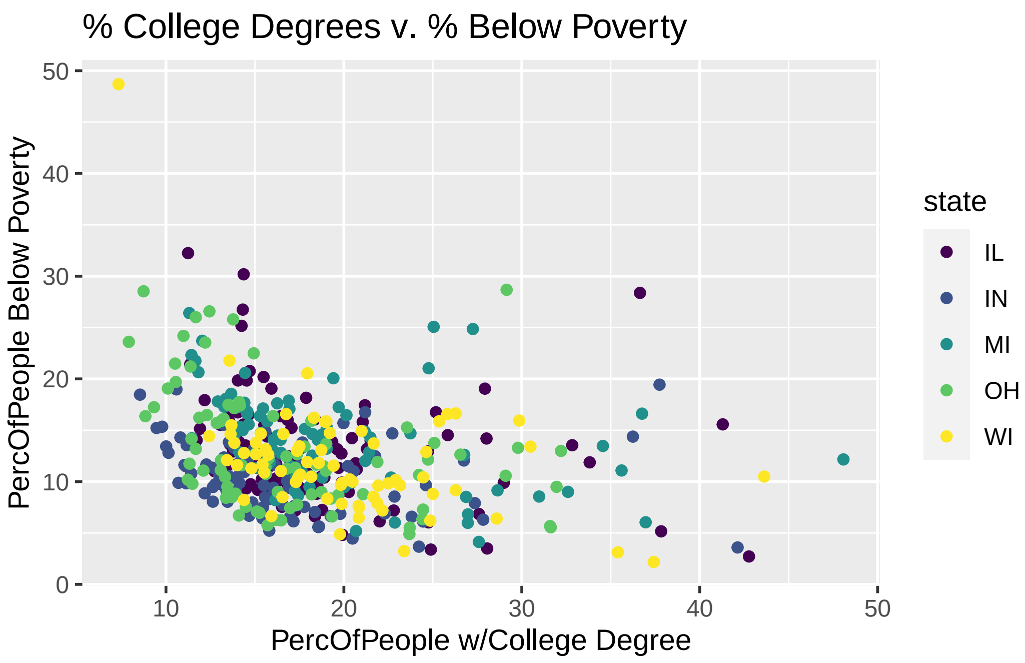


The distribution is skewed to the right.

Above, we can see some data in between the bin of 75000 and 100000, our outliers, which is far from most of the data sitting near 0.

Exercise 2

```
ggplot(data = midwest,
       mapping = aes(x = percollege, y = percbelowpoverty, color = state)) +
  scale_color_viridis_d() +
  geom_point() +
  labs(
    x = "PercOfPeople w/College Degree",
    y = "PercOfPeople Below Poverty",
    title = "% College Degrees v. % Below Poverty "
  )
```



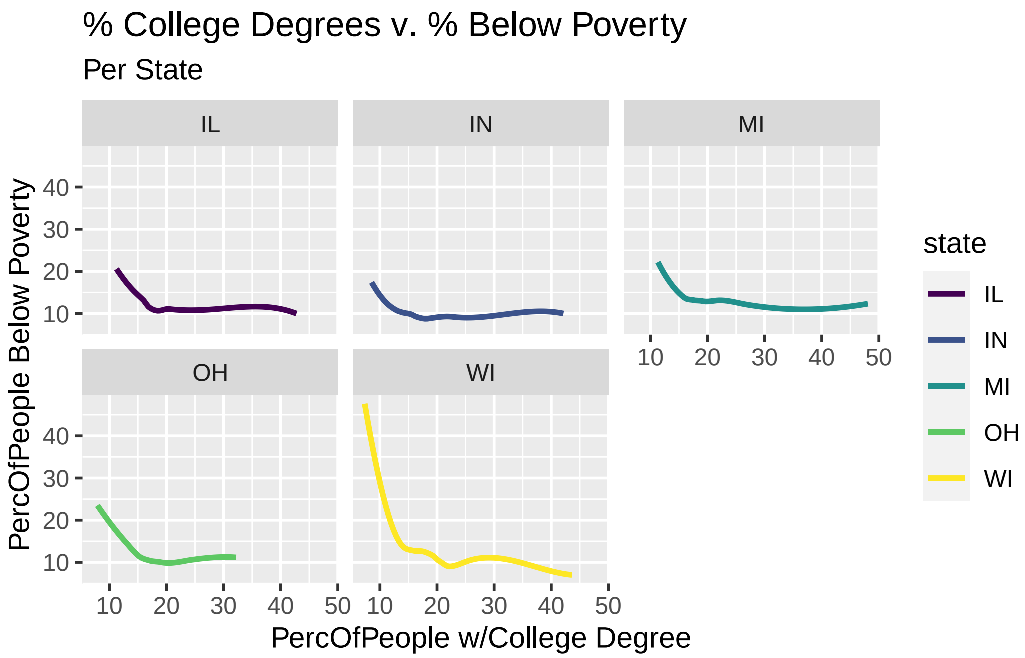
Exercise 3

What I observe is that there is a negative association, but not too extreme. Almost all of the states have a great distribution of counties where the percent of college degrees is between 10-30 percent and have a percentage of people below the poverty line between 5-20 percent. I do think that IL has the most notable curve that expresses the relationship where less college degrees means more people under the poverty line.

Exercise 4

```
ggplot(data = midwest,
       mapping = aes(x = percollege, y = percbelowpoverty, color = state)) +
  scale_color_viridis_d() +
  geom_smooth(se = FALSE) +
  labs(
    x = "PercOfPeople w/College Degree",
    y = "PercOfPeople Below Poverty",
    title = "% College Degrees v. % Below Poverty",
    subtitle = "Per State"
  ) +
  facet_wrap(~ state)
```

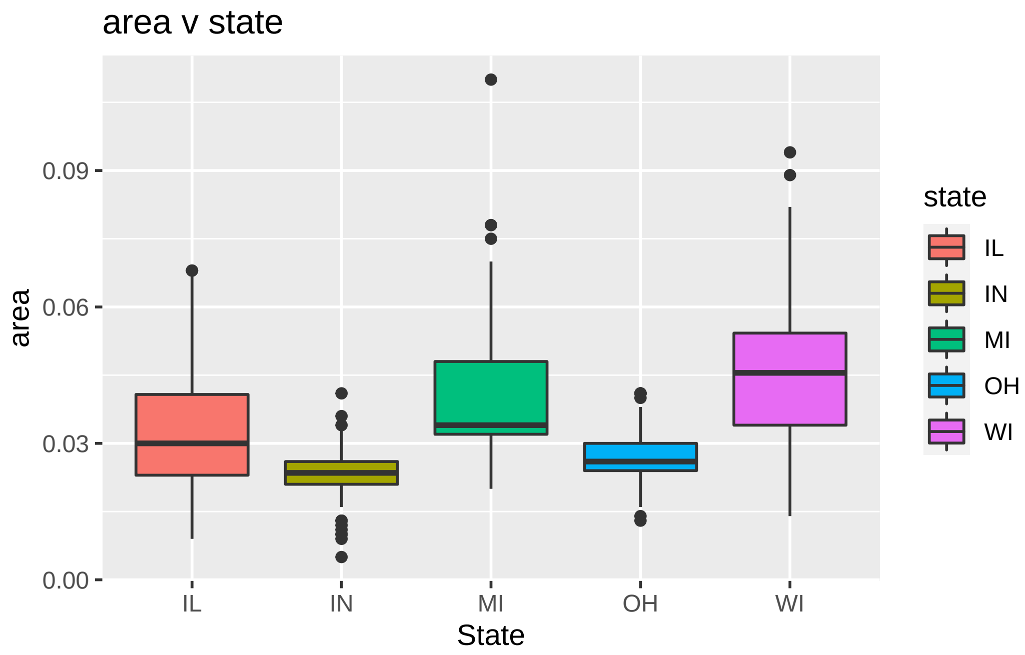
```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



I prefer the geom_smooth plot a lot better since it is showing the average of all the dots in a line. It is easier to read but it is harder to find outliers since it doesn't show you the specific points.

Exercise 5

```
ggplot(midwest, aes(x = state, y=area, fill=state)) +
  geom_boxplot() +
  scale_color_viridis_d() +
  labs(
    x = "State",
    y = "area",
    title = "area v state",
  )
```



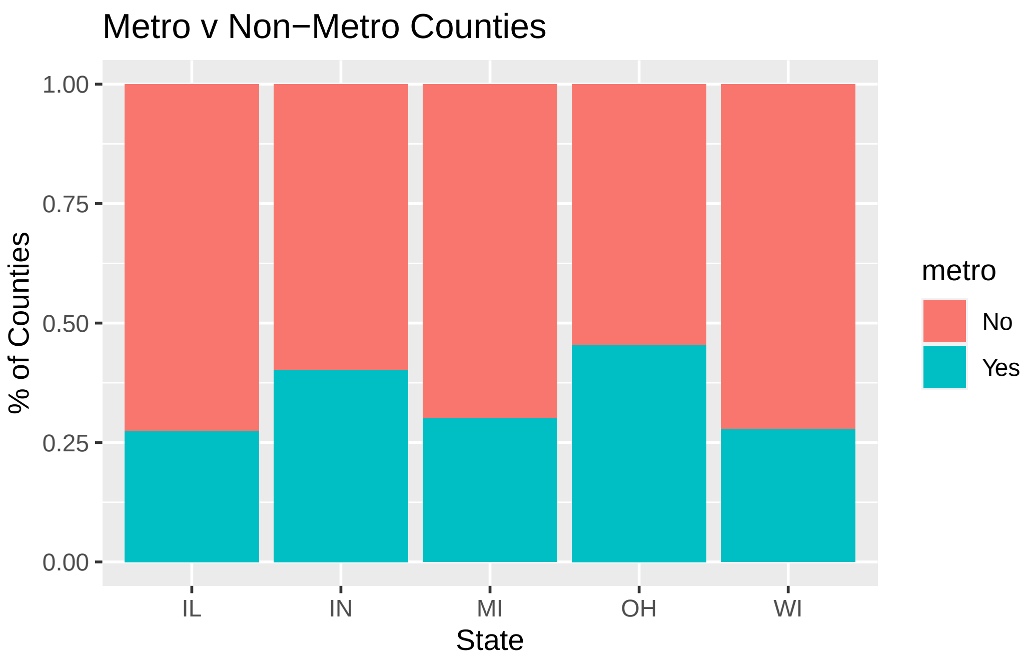
We can observe here that WI has a higher average of area in counties compared to the other states. All the other four states share a pretty similar mean but do have varying Quartiles and whiskers.

The state that has the largest single county is MI. We can tell by its placement far above the others, acting as an outlier.

Exercise 6

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

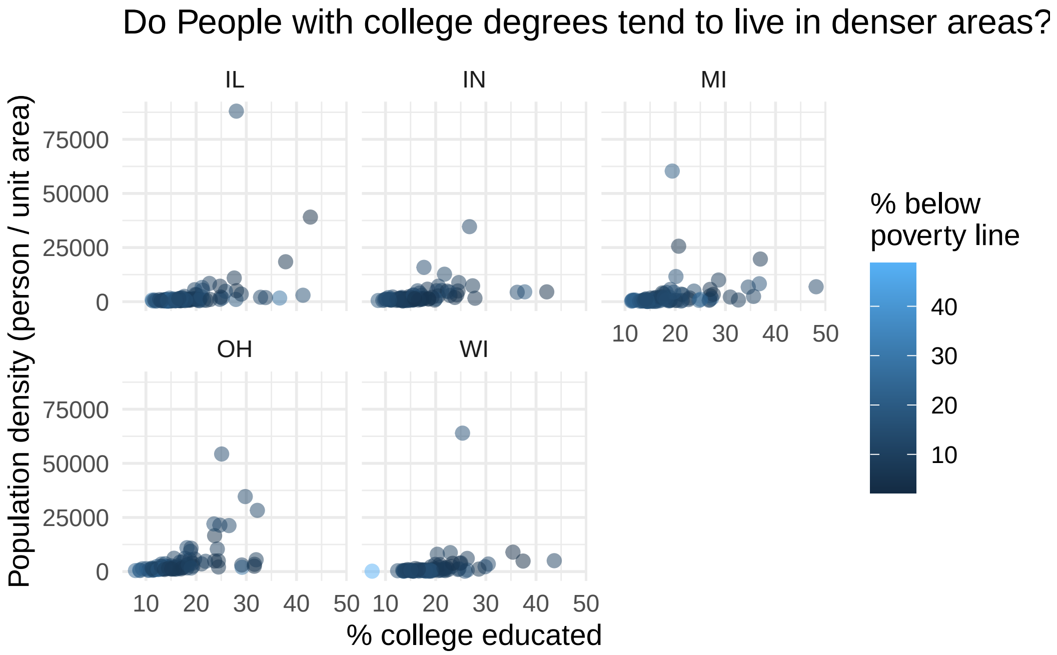
```
ggplot(midwest, aes(x = state, fill = metro)) +
  geom_bar(position = "fill") +
  scale_color_viridis_d() +
  labs(
    x = "State",
    y = "% of Counties",
    title = "Metro v Non-Metro Counties",
  )
```



I noticed that IN and OH have almost 50% of their counties considered “metropolitan”.

Exercise 7

```
ggplot(midwest,
       mapping = aes(x = percollege, y = popdensity, color = percbelowpoverty)) +
  geom_point(size = 2, alpha = 0.5) +
  labs(
    x = "% college educated",
    y = "Population density (person / unit area)",
    title = "Do People with college degrees tend to live in denser areas?",
    color = "% below \npoverty 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)
75000
50000
25000
0
10
20
30
40
50
% college educated
% below poverty line
40
30
20
10