Lab 1 - Data visualization

Hailey Jang

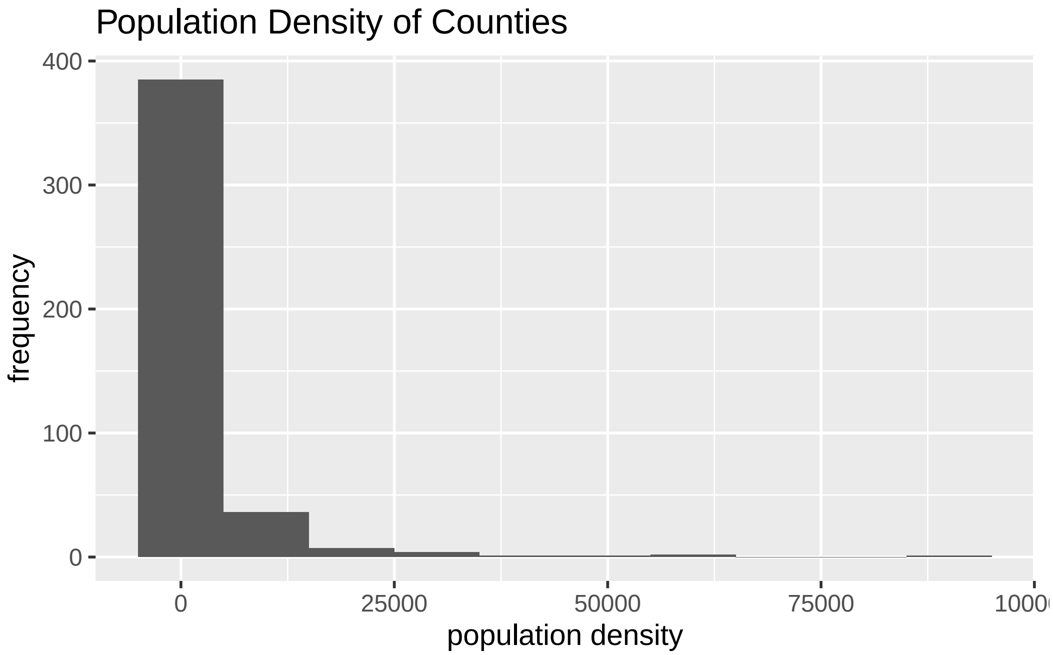
Load Packages

```
library(tidyverse)
library(viridis)
```

Exercise 1

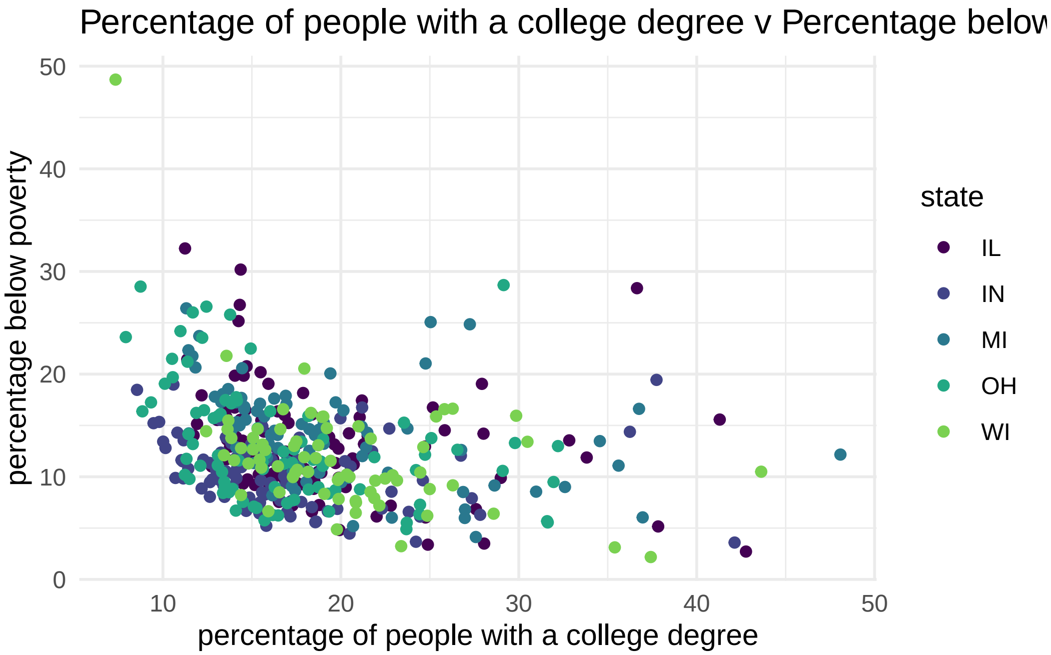
(The shape of the distribution is right-skewed. There seem to be outliers at population density of around 9000.)

```
ggplot(midwest, mapping=aes(x=popdensity))+
  geom_histogram(binwidth=10000)+
  labs(
    x="population density",
    y= "frequency",
    title="Population Density of Counties"
  )
```



Exercise 2

```
ggplot(midwest, mapping=aes(x=percollege, y=percbelowpoverty, color=state))+
  geom_point()+
  labs(
    x="percentage of people with a college degree",
    y= "percentage below poverty",
    title="Percentage of people with a college degree v Percentage below poverty by State"
  )+
  scale_color_viridis_d(option="D", end=0.8)+
  theme_minimal()
```



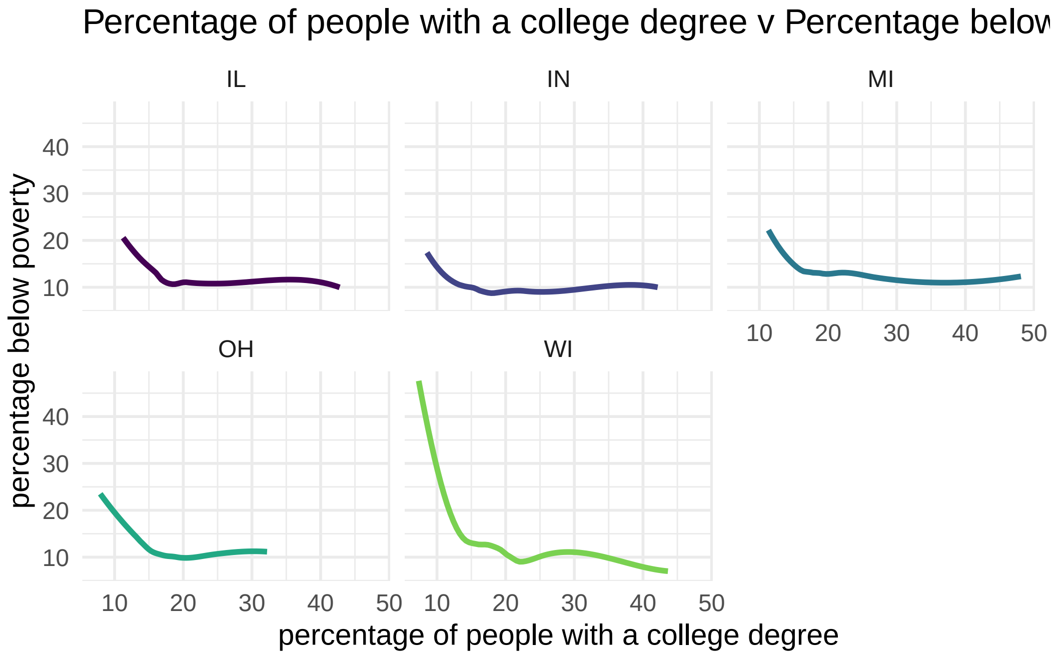
Exercise 3

In the plot from exercise 2, I observe that there are several more points clustered near the lower left, indicating lower percentage of people with a college degree tend to also be in the lower range of percentage of people living below poverty. Similarities is patterns among states are that most of the states follow the trend I mentioned in the previous sentence. Differences in patterns among states are that Illinois tends to have data points are higher in percentage of having a college degree while Ohio in particular has data points that are slightly lower in percentage of having a college degree.

Exercise 4

```
ggplot(midwest, mapping=aes(x=percollege, y=percbelowpoverty, color=state))+
  geom_smooth(se=FALSE, show.legend=FALSE)+
  labs(
    x="percentage of people with a college degree",
    y= "percentage below poverty",
    title="Percentage of people with a college degree v Percentage below poverty"
  )+
  facet_wrap(~state)+
  scale_color_viridis_d(option="D", end=0.8)+
  theme_minimal()
```

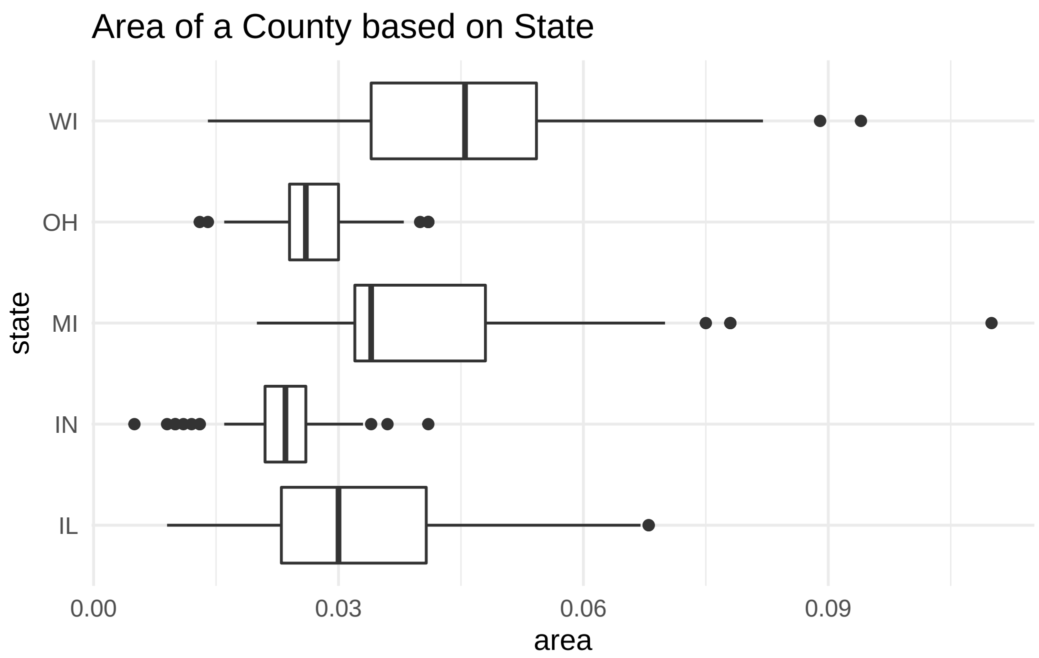
```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



I prefer this plot over the plot in Ex 2 because the first plot was hard to tell where the individual data points were, so it was difficult to compare the patterns. However, for this plot, it is much more clear what the pattern looks like, making it easier to compare them.

Exercise 5

```
ggplot(midwest, mapping=aes(x=area, y=state))+
  geom_boxplot(show.legend=FALSE)+
  labs(
    x="area",
    y= "state",
    title="Area of a County based on State"
  )+
  theme_minimal()
```



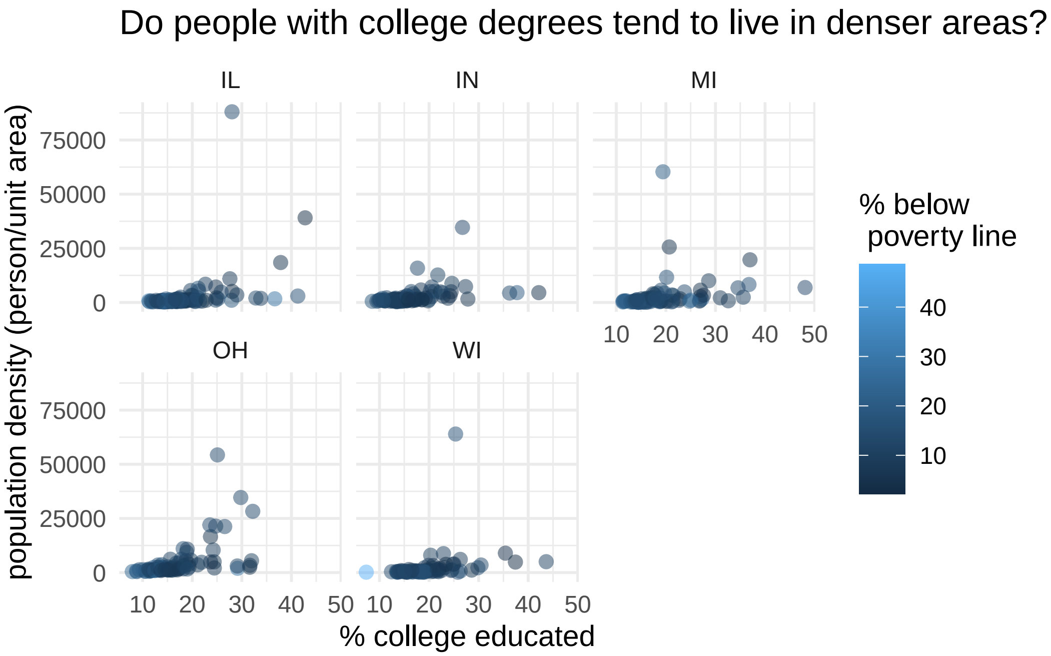
I observe from the plot that WI has the largest IQR in area while IN has the smallest. IN has several outliers compared to the other states. The state with the single largest is MI, and the plot shows it because it is the data point with the highest area.

Exercise 6

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

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 \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