Lab 1 - Data visualization

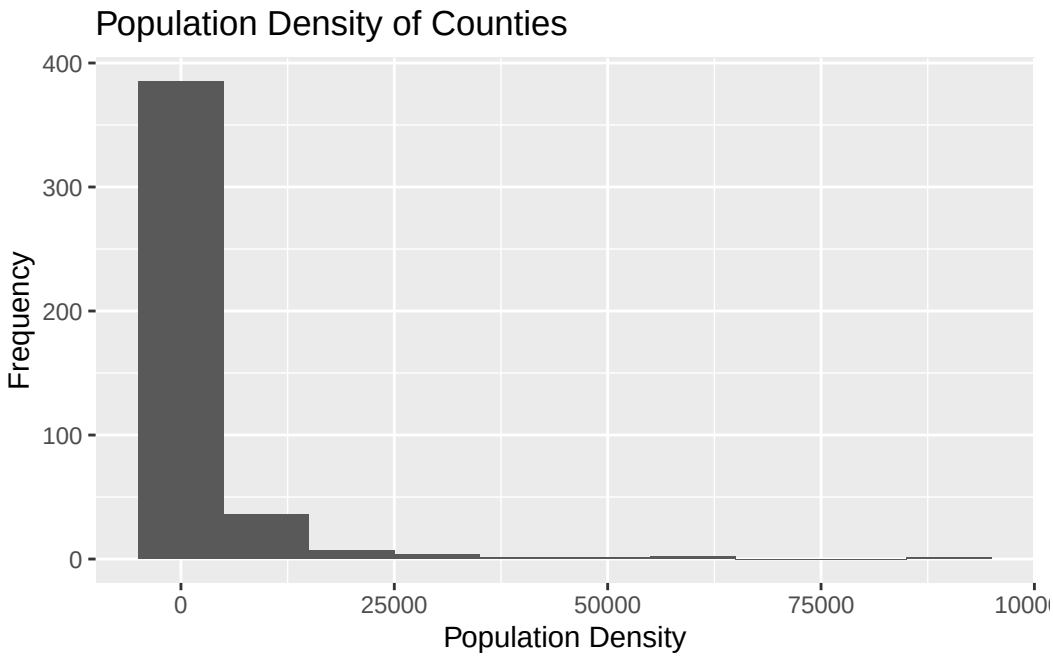
Robertson Waweru

Load Packages

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

Exercise 1

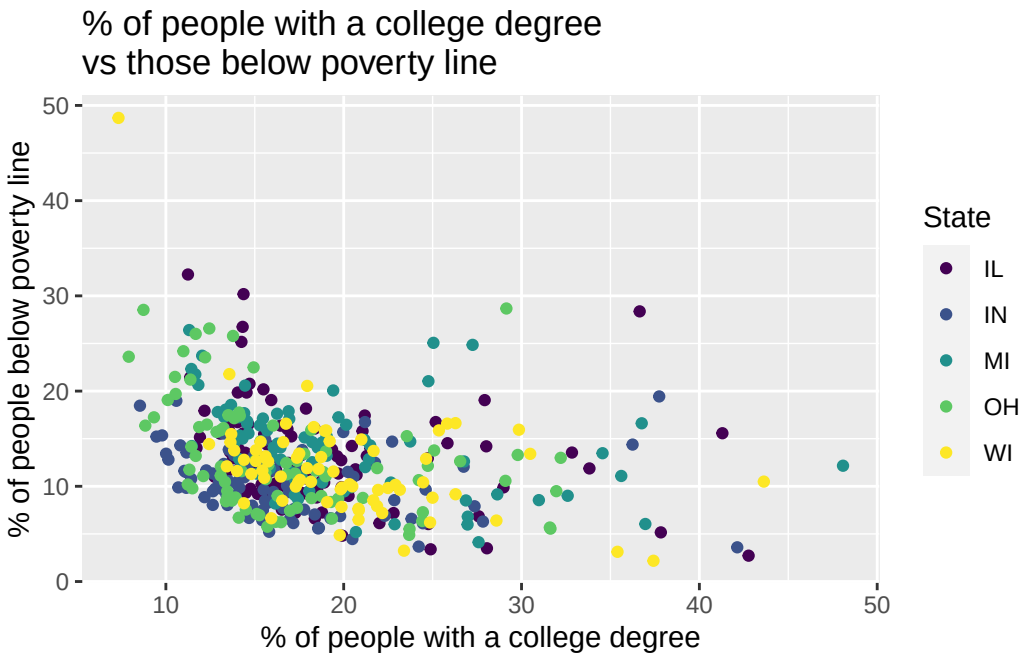
```
ggplot(midwest,
       aes(x = popdensity)) +
  geom_histogram(binwidth = 10000) +
  labs(
    x = "Population Density",
    y = "Frequency",
    title = "Population Density of Counties"
  )
```



Describe the shape of the distribution. The distribution is right skewed and unimodal. Does there appear to be are any outliers? Briefly explain. There are a few counties that have a larger population density than the rest, the most prominent being Cook County in Illinois with 88018.3966. This can be explained by the fact that Cook County consists of 134 municipalities including the City of Chicago.

Exercise 2

```
ggplot(midwest,
       aes(x = percollege, y = percbelowpoverty, color = state)) +
  geom_point() +
  labs(
    x = "% of people with a college degree",
    y = "% of people below poverty line",
    title = "% of people with a college degree \nvs those below poverty line",
    color = "State"
  ) +
  scale_color_viridis_d()
```



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. In general, counties with higher percentages of people with a college degree have marginally lower percentages of people below the poverty line. Differences: The counties in Wisconsin show significant decline in the percentage of people below poverty line with increase in the percentage of people with a college degree. Indiana seems to have most of its counties with a lower percentage of people below the poverty line while having a relatively low percentage of people with a college degree with counties that a have a higher percentage of people with a college degree having a higher percentage of people below poverty line.

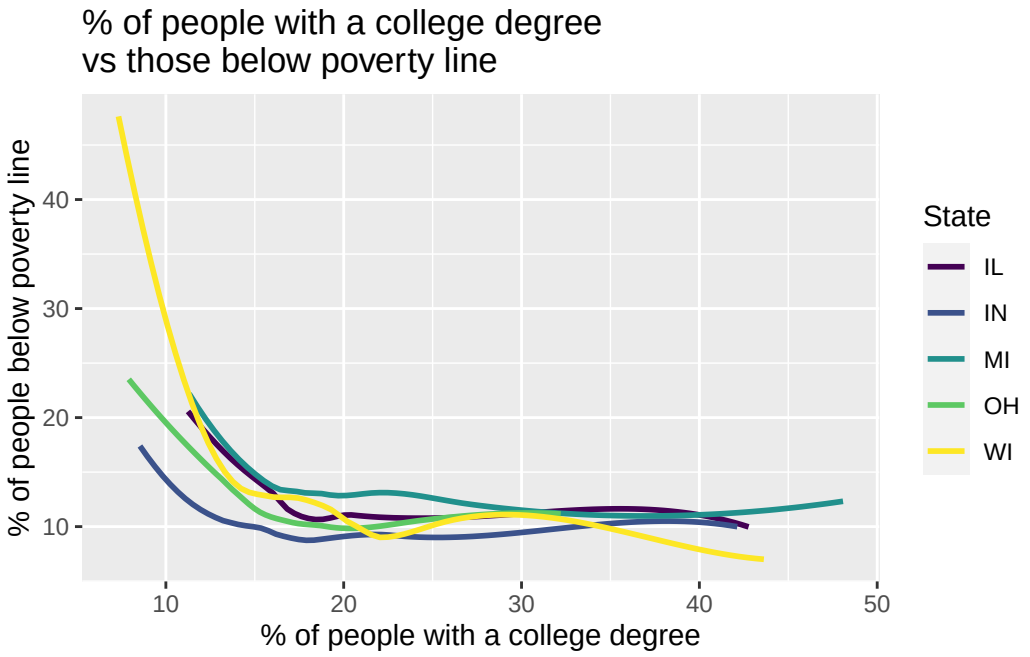
Exercise 4

Now, let's examine the relationship between the same two variables, using a separate plot for each state. Label the axes and give the plot a title. Use geom_smooth with the argument se = FALSE to add a smooth curve fit to the data.

```
ggplot(midwest,
       aes(x = percollege, y = percbelowpoverty, color = state)) +
  geom_smooth(se = FALSE) +
```

```
    labs(
      x = "% of people with a college degree",
      y = "% of people below poverty line",
      title = "% of people with a college degree \nvs those below poverty line",
      color = "State"
    ) +
    scale_color_viridis_d()
```

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

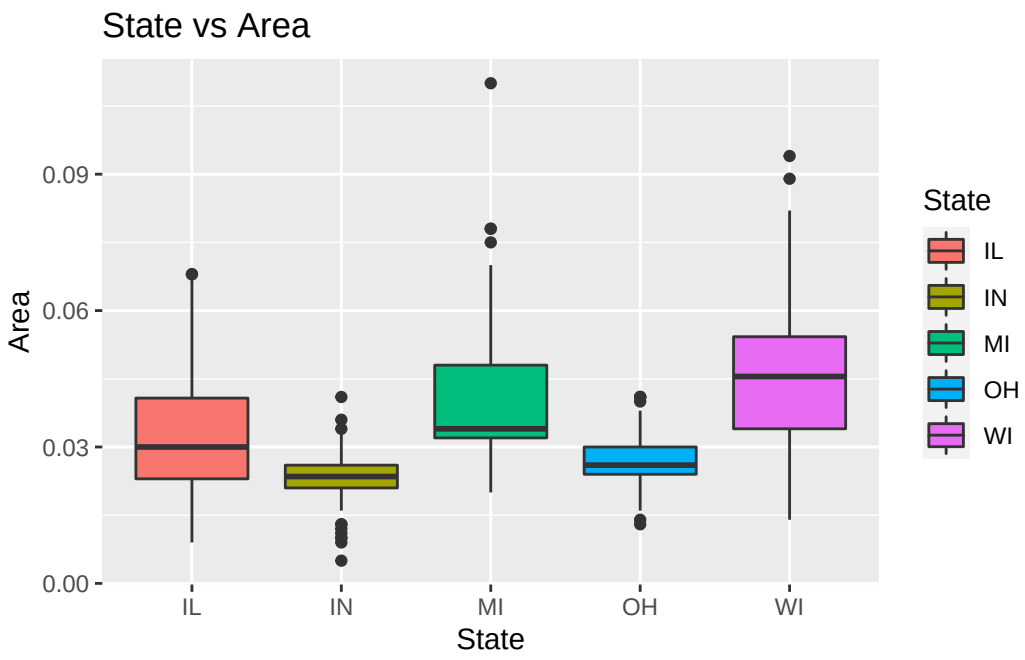



```
    # can also work as scale_color_viridis(discrete = TRUE)
```

Which plot do you prefer - this plot or the plot in Ex 2? Briefly explain your choice. I prefer the curve fit plot because it shows the trend of the data for each state as the percentage of people with a college degree increase with is much more useful in drawing conclusions than a scatterplot which maps to each data point which makes it harder to identify trends and draw meaningful conclusions.

Exercise 5

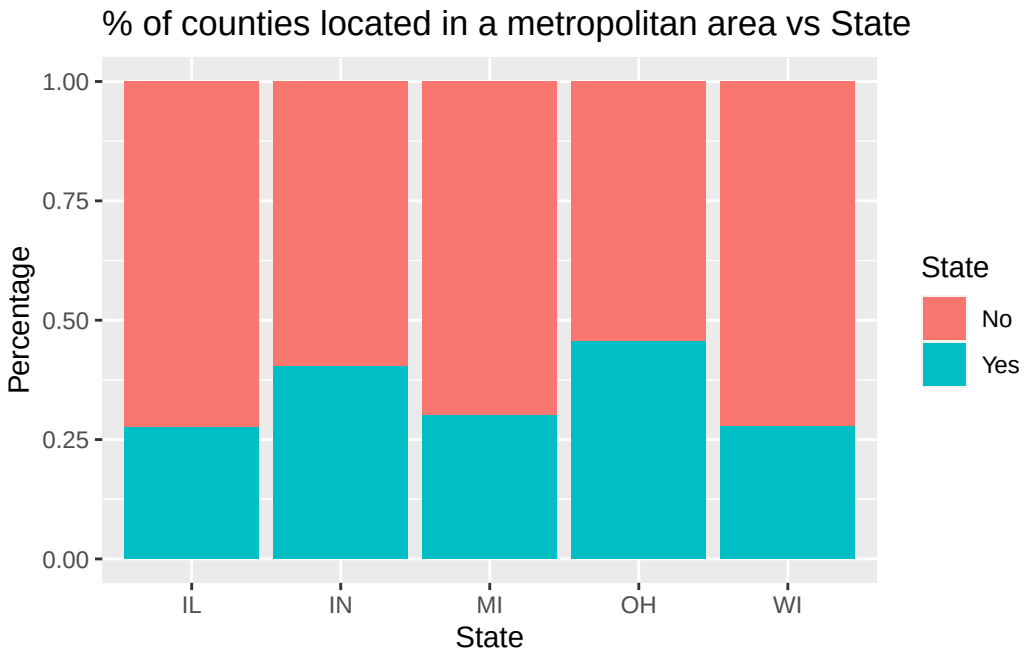
```
ggplot(midwest,
       aes(x = state, y = area, fill = state)) +
  geom_boxplot() +
  labs(
    x = "State",
    y = "Area",
    title = "State vs Area",
    fill = "State"
  )
```



Describe what you observe from the plot. Most of Indiana's and Ohio's counties are smaller compared to the rest of the states. Wisconsin and Michigan have larger counties with Michigan having the largest county. Illinois has counties with moderate areas relative to the other states. Which state has the single largest county? Michigan How do you know based on the plot? Michigan's box plot has an outlier with a larger area than the other outliers.

Exercise 6

```
midwest <- midwest |>
  mutate(metro = if_else(inmetro == 1, "Yes", "No"))
ggplot(midwest,
       aes(x = state, fill = metro)) +
  geom_bar(position = "fill") +
labs(
    x = "State",
    y = "Percentage",
    title = "% of counties located in a metropolitan area vs State",
    fill = "State"
  )
```



What do you notice from the plot? In general, majority of the counties in all of the states are not in metropolitan areas. Ohio and Indiana have a higher percentage of their counties in metropolitan areas compared to the rest of the states with Ohio having the highest percentage of its counties in a metropolitan area.

Exercise 7

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