Lab 1 - Data visualization

Siqi Lan

Load Packages

```
library(tidyverse)
```

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

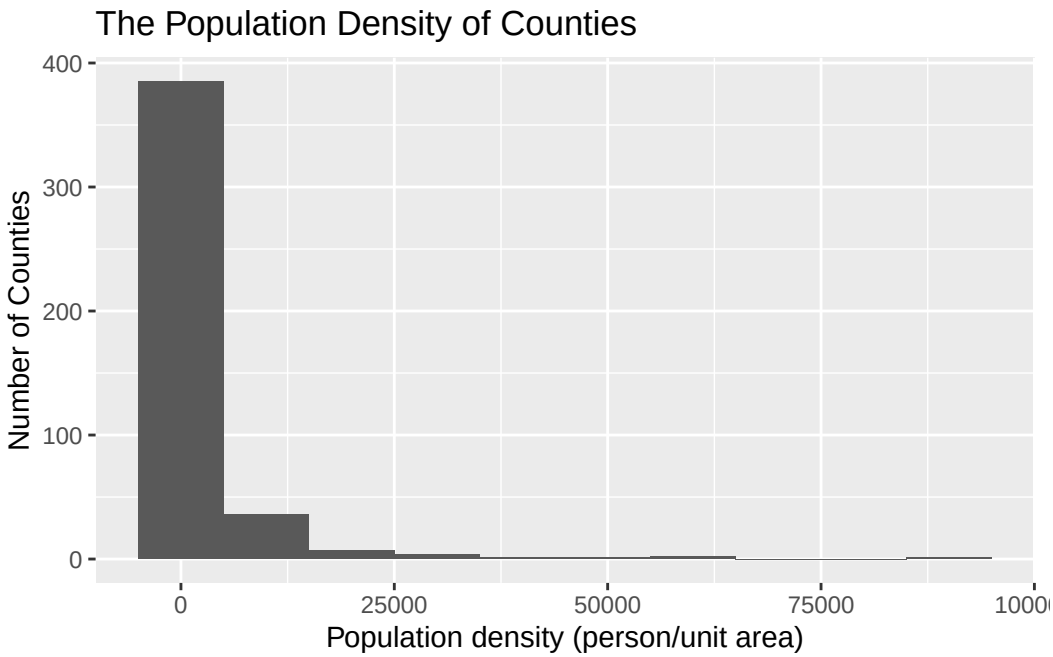
```
library(viridis)
```

```
data(midwest)
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,
       aes(x = popdensity)) +
  geom_histogram(binwidth = 10000) +
  labs(x = "Population density (person/unit area)",
       y = "Number of Counties",
       title = "The Population Density of Counties")
```



1) Describe the shape of the distribution.

   The distribution is right skewed.

2) Does there appear to be any outliers? Briefly explain.

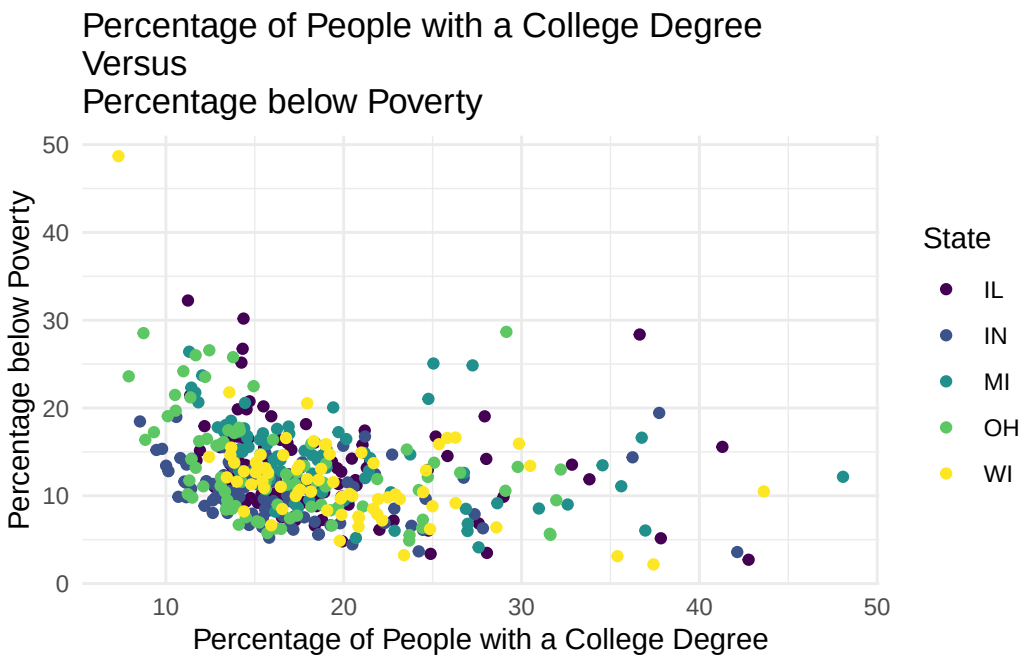
   Yes, from the histgram we can detect that there are some extremely high density around 87500 people per area.

Exercise 2

(Type your answer to Exercise 2 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,
       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 \nVersus
Percentage below Poverty",
       color = "State") +
  scale_color_viridis_d(option = "D", end = 1) +
```

```
theme_minimal()
```



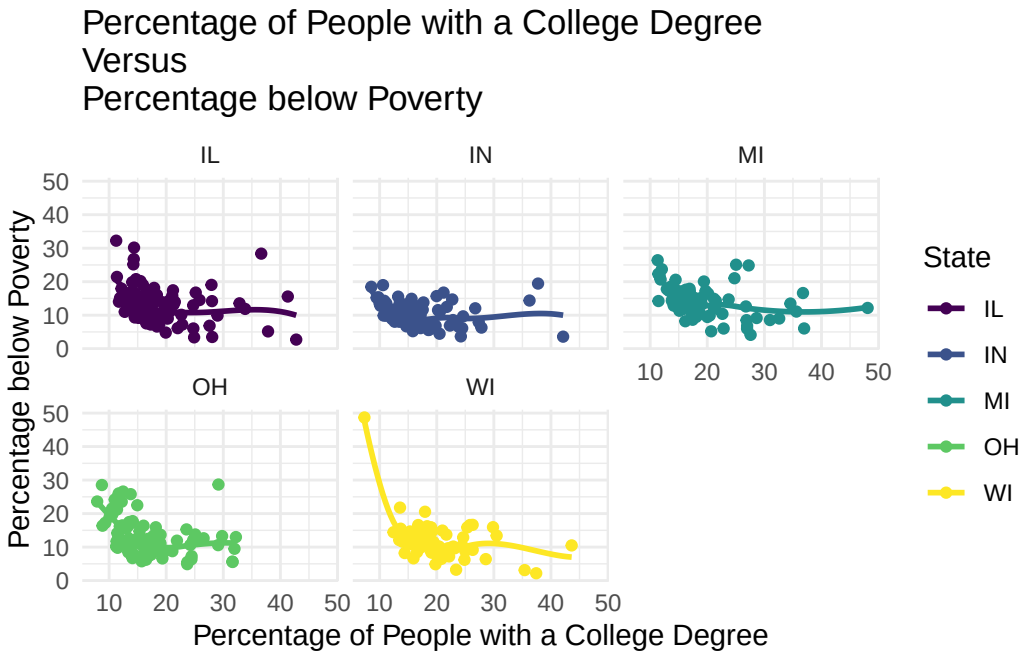
Exercise 3

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

The scattered points are telling the story that there is a negative relationship between the percentage of people with a college degree and their percentage below poverty. The similarity across the states is that all the states share the same negative relationship mentioned before and these points are located where the percentage of people with a college degree is from 10% to 20% and the percentage below poverty is from 5% to 18%. And the differences between the states are that: overall WI has higher percentage of people with a college degree and lower percentage below poverty; IL and MI have many scattered points which is quite far from the curve most points converged to.

Exercise 4

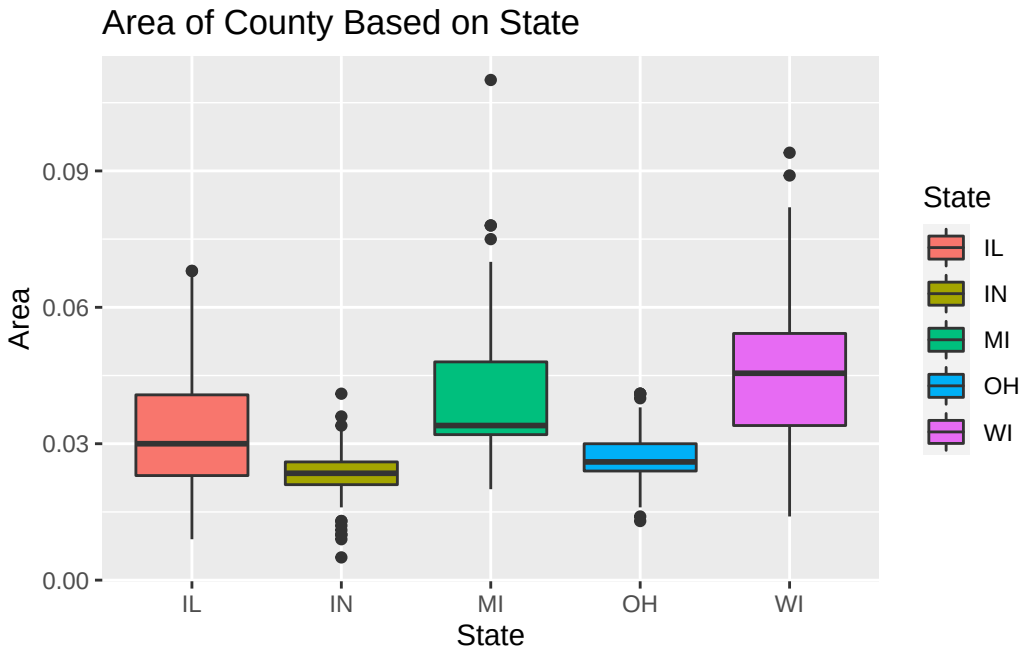
```
ggplot(data = midwest,
       aes(x = percollege, y = percbelowpoverty, color = state)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  facet_wrap(~state) +
  labs(x = "Percentage of People with a College Degree",
       y = "Percentage below Poverty",
       title = "Percentage of People with a College Degree \nVersus
Percentage below Poverty",
       color = "State") +
  scale_color_viridis_d(option = "D", end = 1) +
  theme_minimal()
```



Compared with the plot in Ex.2, I prefer the plot in Ex.4 since it can convey a much clearer relationship of these two variables without some influence from outliers. Also, With subplots of each state, more specifically negative relationships of these states can be told by audiences.

Exercise 5

```
ggplot(data = midwest,
       mapping = aes(x = state, y = area, group = state, fill = state)) +
  geom_boxplot(show.legend = T) +
  labs(x = "State",
       y = "Area",
       title = "Area of County Based on State",
       fill = 'State')
```



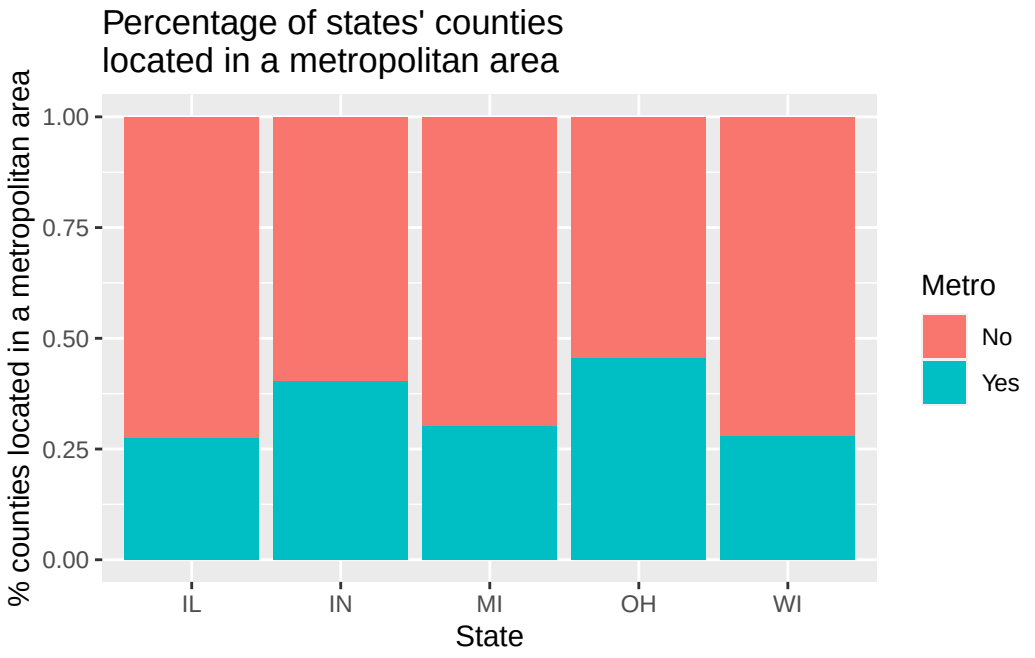
1) From the boxplot, we can easily find out that WI has the mean maximum area and its box is higher than other states'; IN has the mean minimum area and its box is lower and smaller than others', which also has the most outliers recognized by R; MI has the medium area and the highest outlier, but its Q2 (median) is too close to Q1 (25th Percentile) and far from Q3 (75th Percentile); IL has similar Q2 as MI but it has a lower Q3; and OH's box is just a little higher than IN's, with some close outliers. Therefore, WI seems to have counties that tend to be geographically larger than others as its box is higher than others'.

2) MI state seems to have the single largest county as it has the highest point shown in the plot , which is regarded to be a potential outlier by R.

Exercise 6

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

```
ggplot(data = midwest,
       mapping = aes(x = state, fill = metro)) +
  geom_bar(position="fill", show.legend = T) +
  labs(x = "State",
       y = "% counties located in a metropolitan area",
       title = "Percentage of states' counties \nlocated in a metropolitan area",
       fill = 'Metro')
```



From the plot, it is obvious that OH has a highest percentage of their counties located in a metropolitan area and IN has the second highest percentage. For the other three states, their share a similar low percentage of their counties located in a metropolitan area which is hard to tell directly.

Exercise 7

```
ggplot(data = midwest,
       mapping = aes(x = percollege, y = popdensity, color = percbelowpoverty)) +
  geom_point(size = 2, alpha = 0.5, show.legend = T) +
  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 \npoverty line") +
  theme_minimal()
```