Lab 1 - Data visualization

Siqi Lan

Load Packages

```
library(tidyverse)
```

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

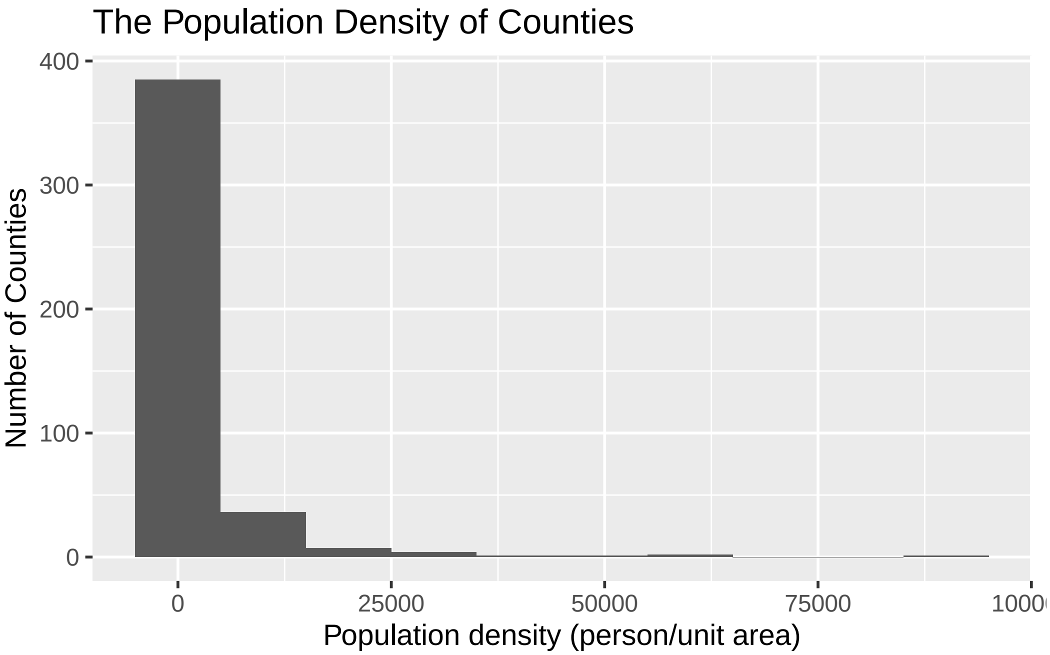
```
library(viridis)
```

```
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.

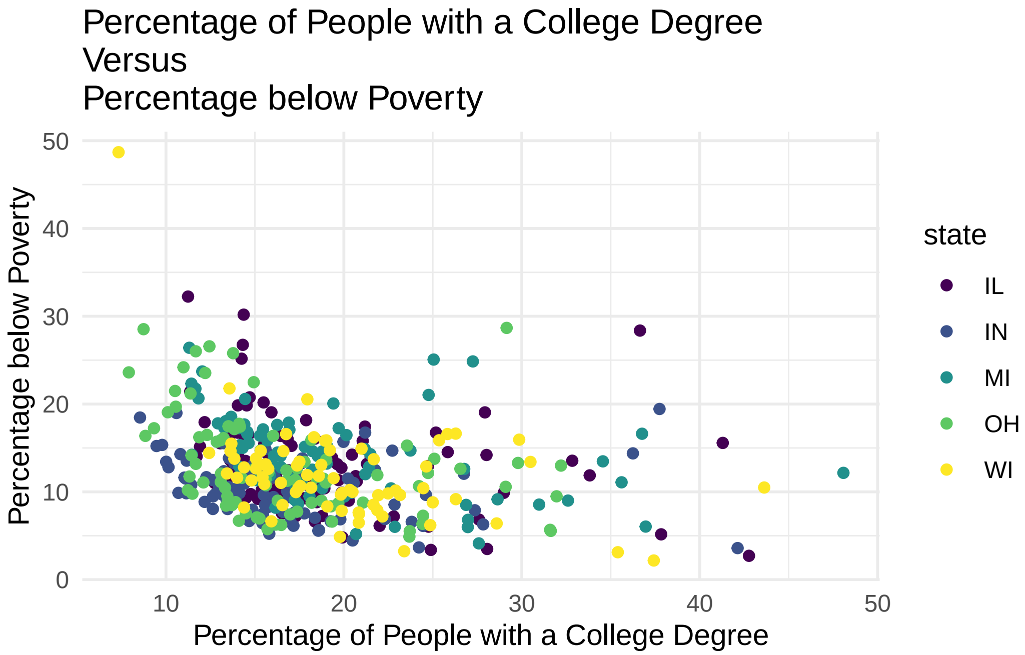
2) Does there appear to be any outliers? Briefly explain.

   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.)

```r
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") +
  scale_color_viridis_d(option = "D", end = 1) +
  theme_minimal()
```



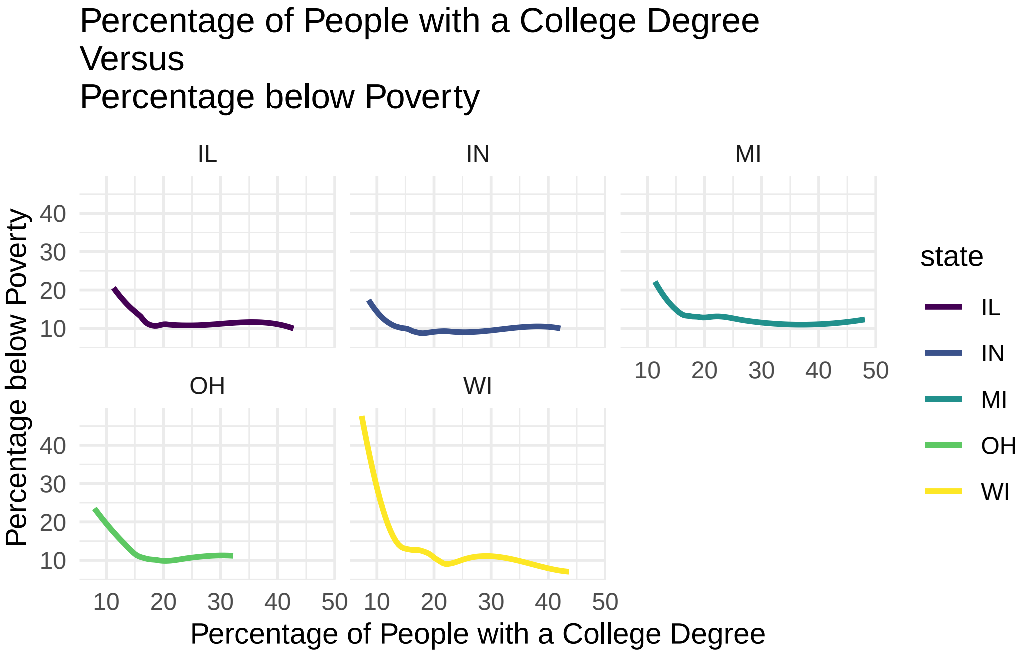
Exercise 3

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

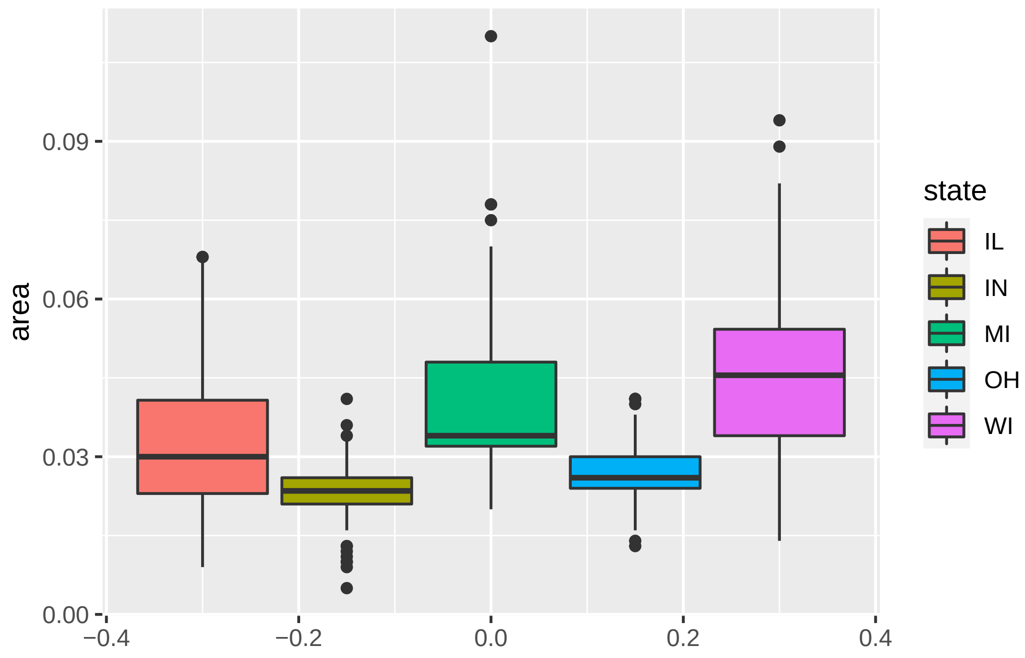
```
ggplot(data = midwest,
       aes(x = percollege, y = percbelowpoverty, color = state)) +
  geom_smooth(se = FALSE) +
  facet_wrap(~state) +
  labs(x = "Percentage of People with a College Degree",
       y = "Percentage below Poverty",
       title = "Percentage of People with a College Degree \nVersus
Percentage below Poverty") +
  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 clearer relationship of these two variables without the influence from outliers. Also, With subplots of each state, more specifically negative relationships of these states can be told by audiences.

Exercise 5

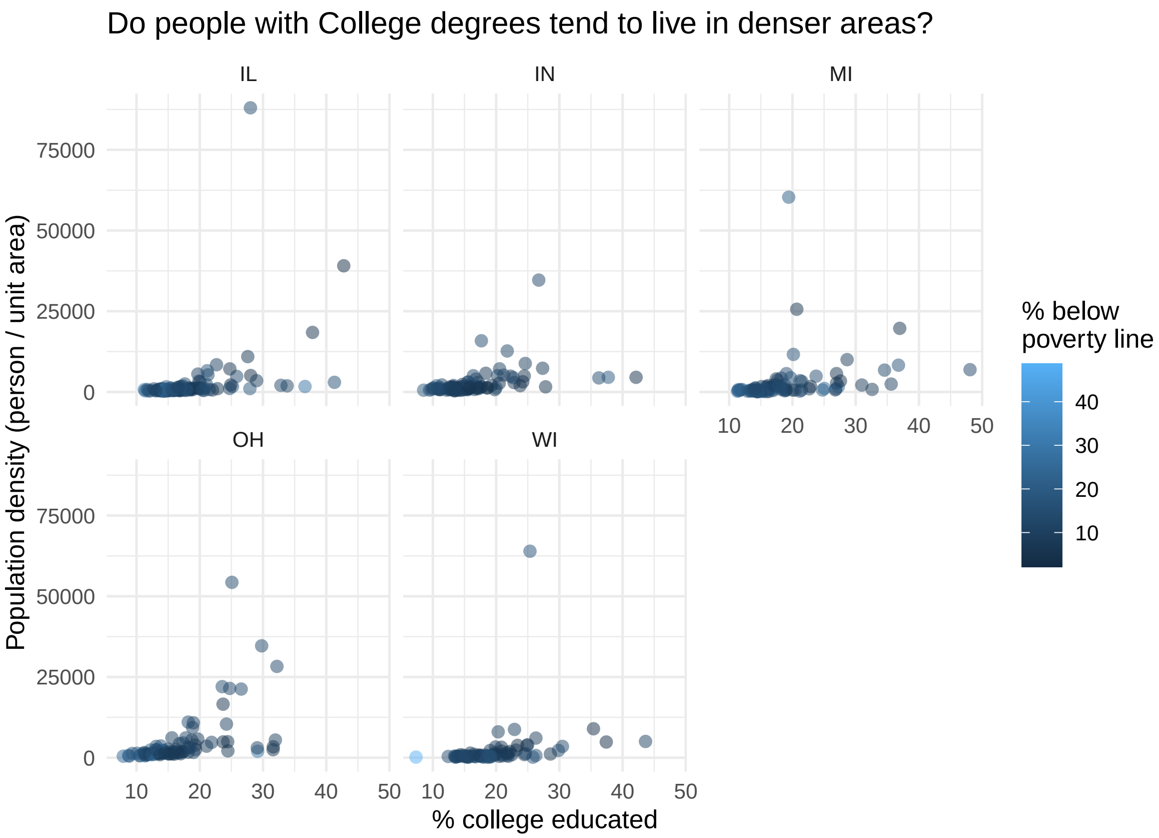
```
ggplot(data = midwest,
       mapping = aes(y = area, group = state, fill = state)) +
  geom_boxplot(show.legend = T)
```



Exercise 6

Exercise 7

```r
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()
```


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