Lab 1 - Data visualization

Jing Liu

Load Packages

```
library(tidyverse)
```

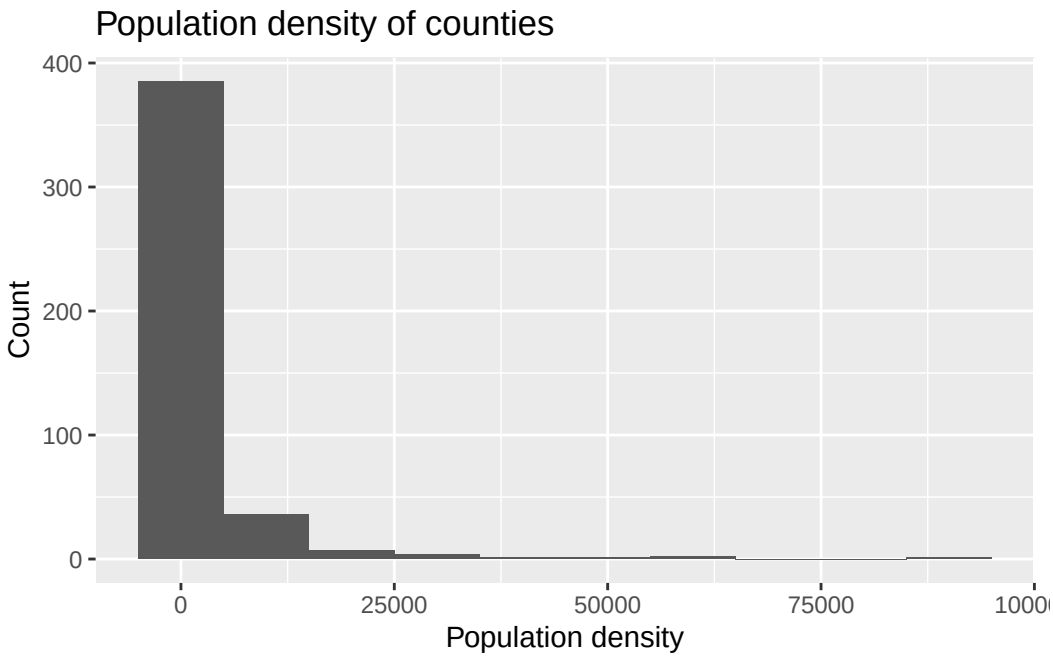
```
Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
had status 1
```

```
library(viridis)
```

Exercise 1

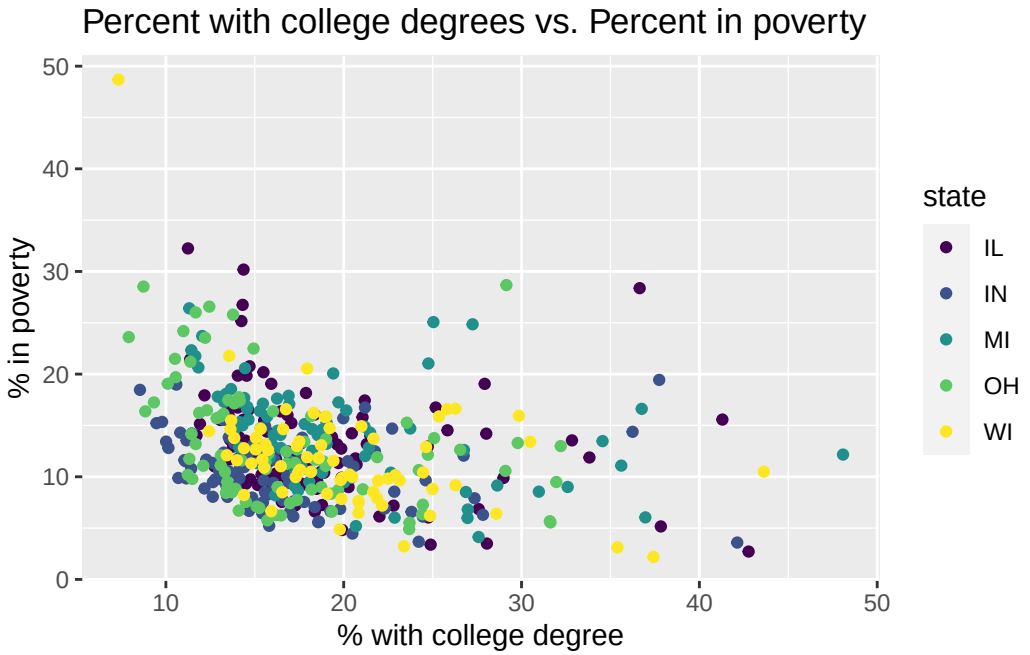
The distribution shown below is unimodal and heavily right skewed. There seems to be a couple of outliers at around 60000 and 90000 on the x-axis.

```
ggplot(midwest, aes(x = popdensity)) +
  geom_histogram(binwidth = 10000) +
  labs(x = "Population density",
       y = "Count",
       title = "Population density of counties")
```



Exercise 2

```
ggplot(midwest, aes(x = percollege, y = percbelowpoverty, color = state)) +
  geom_point() +
  scale_color_viridis_d() +
  labs(x="% with college degree",
       y="% in poverty",
       title="Percent with college degrees vs. Percent in poverty")
```



Exercise 3

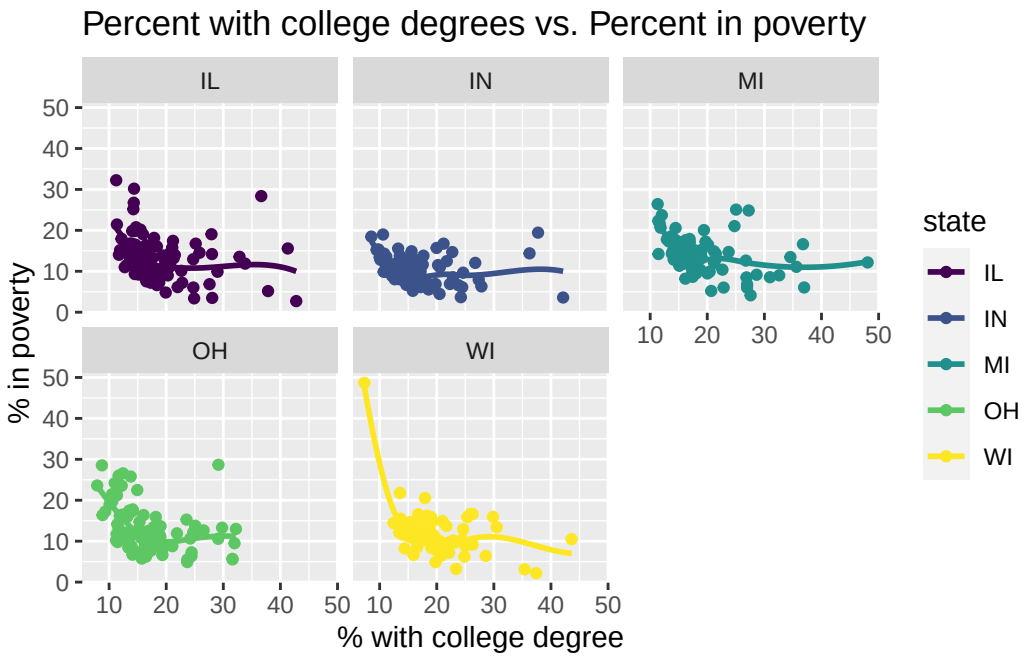
In all the states collectively, there seems to be a rough inverse relationship between percentage of people with college degrees and percentage of people living in poverty. Illinois follows this trend the least, with multiple counties having both high percentages of people with college degrees and high percentages of people living in poverty. Indiana seems to have the lowest percentages of people with college degrees and percentages of people in poverty. Wisconsin has the most constant trend, with the percentage of people with college degrees widely varying (although more right-skewed) and the percentage of people living in poverty being relatively constant among different counties. There is one major outlier for Wisconsin, however, that has an unusually high percentage of people in poverty low percentage of people with college degrees.

Exercise 4

```
ggplot(midwest, aes(x = percollege, y = percbelowpoverty, color = state)) +
  geom_point() +
  facet_wrap(vars(state)) +
  scale_color_viridis_d() +
  labs(x = "% with college degree",
```

```
        y = "% in poverty",
        title = "Percent with college degrees vs. Percent in poverty") +
    geom_smooth(se = FALSE)
```

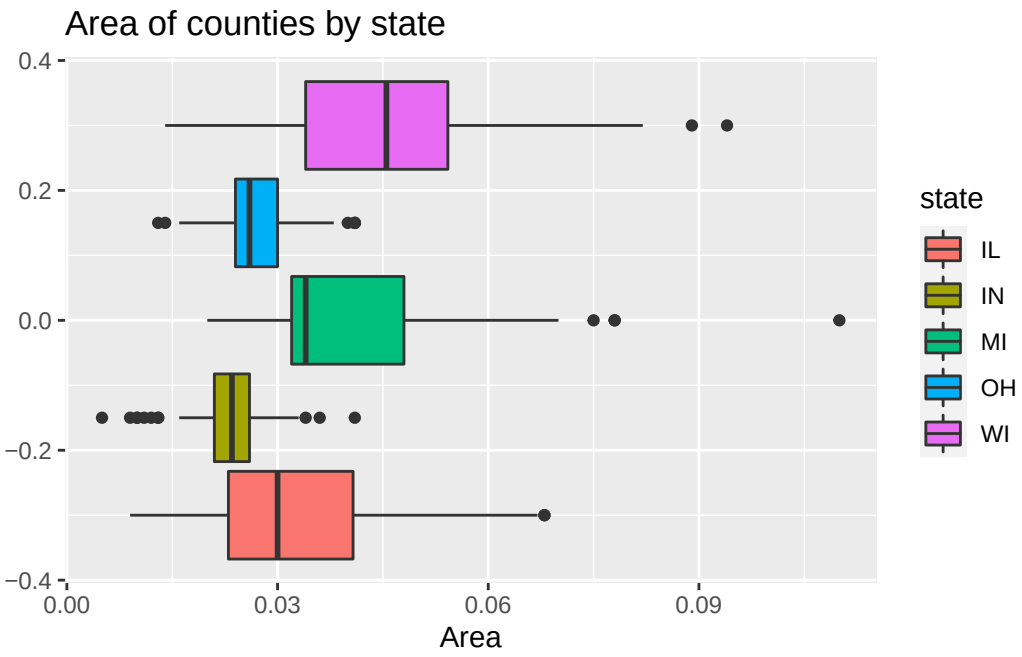
```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



I personally prefer this plot over the one in Exercise 2. In Exercise 2, all the states were on the same graph, which made it hard to distinguish different points that were overlapping each other. This is especially problematic for states whose colors were very similar to each other. It also it made it hard to analyze trends of individual states, because other unrelated points would obstruct the view of points of interest. In addition, Exercise 4 has helpful trend lines that makes it easier to analyze the trends of each state.

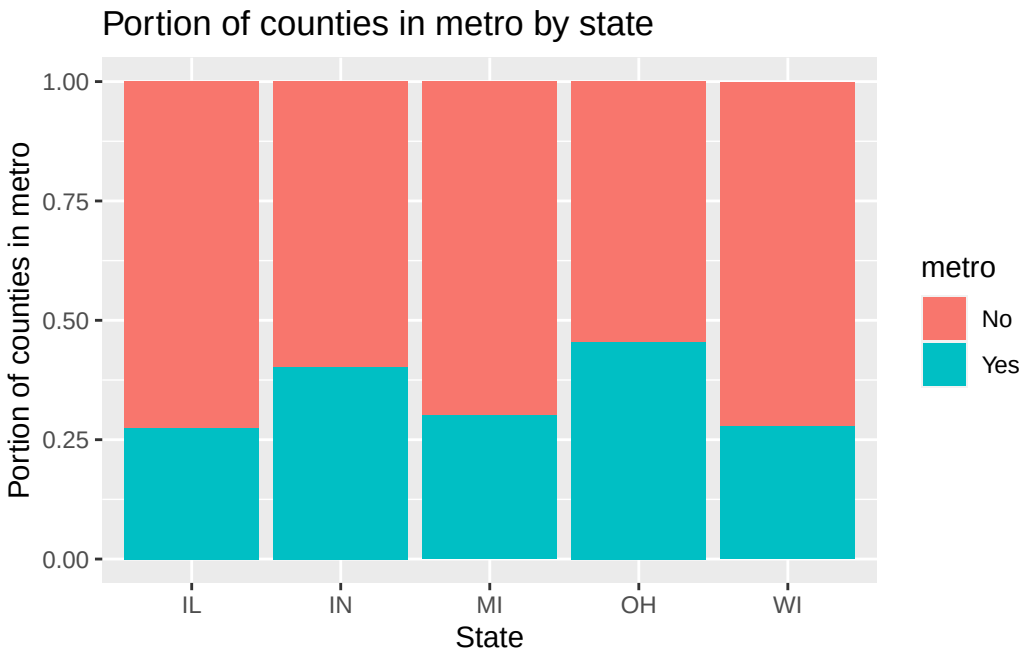
Exercise 5

```
ggplot(midwest, aes(x=area, fill = state)) +
  geom_boxplot() +
  labs(x="Area", title="Area of counties by state")
```



Exercise 6

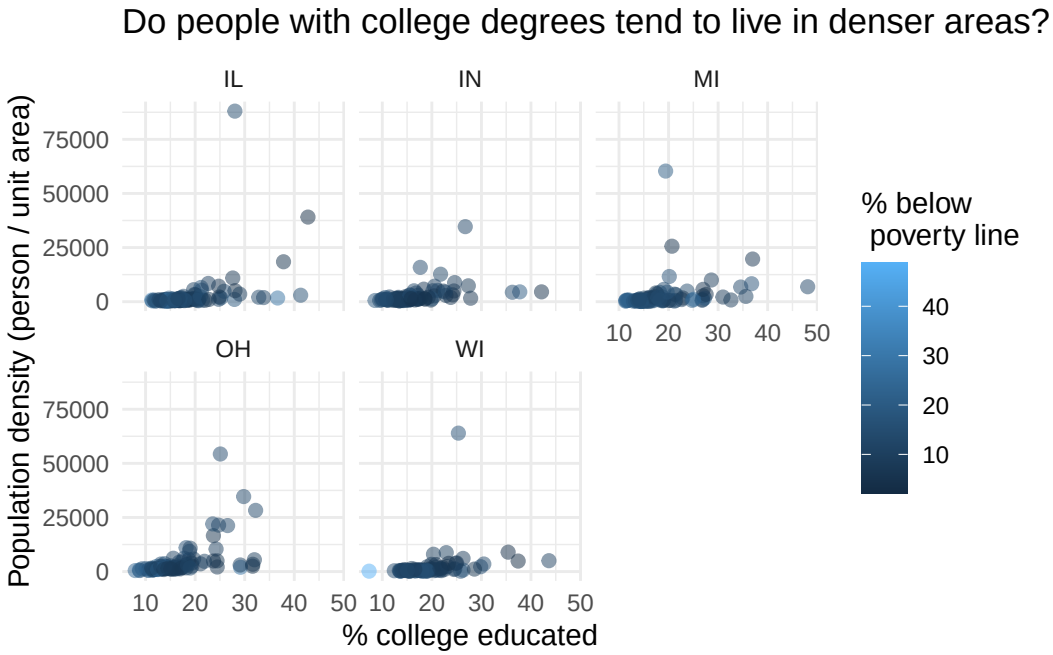
```
midwest <- midwest |>
  mutate(metro = if_else(inmetro == 1, "Yes", "No"))

ggplot(data=midwest, aes(x=state, fill=metro))+
  geom_bar(position="fill")+
  labs(x="State",
       y="Portion of counties in metro",
       title="Portion of counties in metro by state")
```



Every single state in the plot above has more non-metro counties than metro counties. Ohio has the highest percentage of metro counties, and Illinois has the lowest percentage of metro counties.

Exercise 7

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