Lab 1 - Data visualization

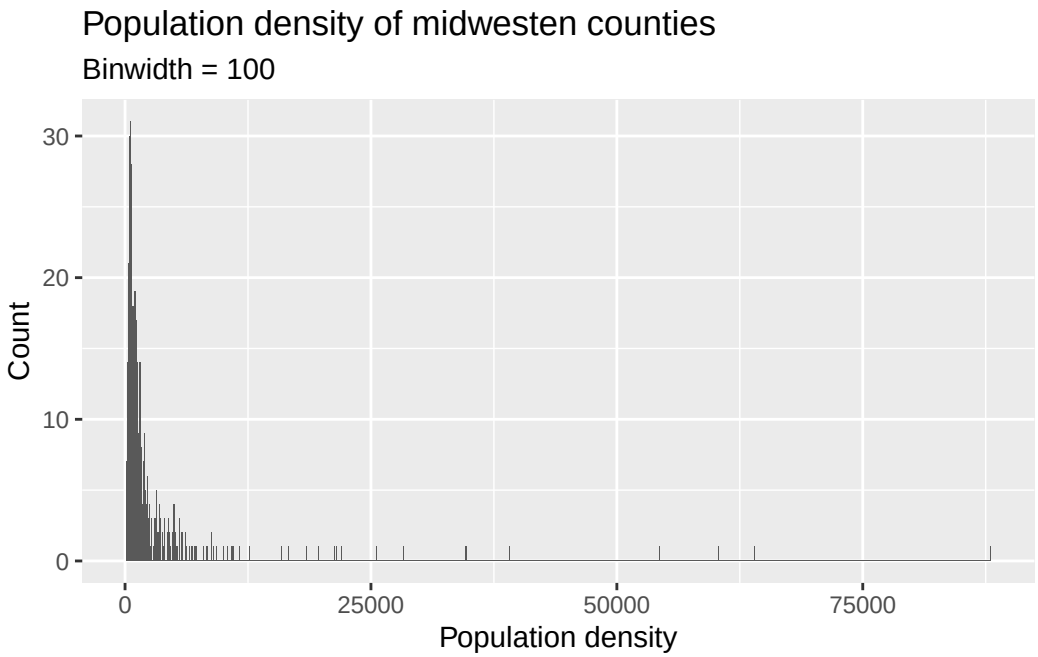
Edmond Niu

```
library(tidyverse)
```

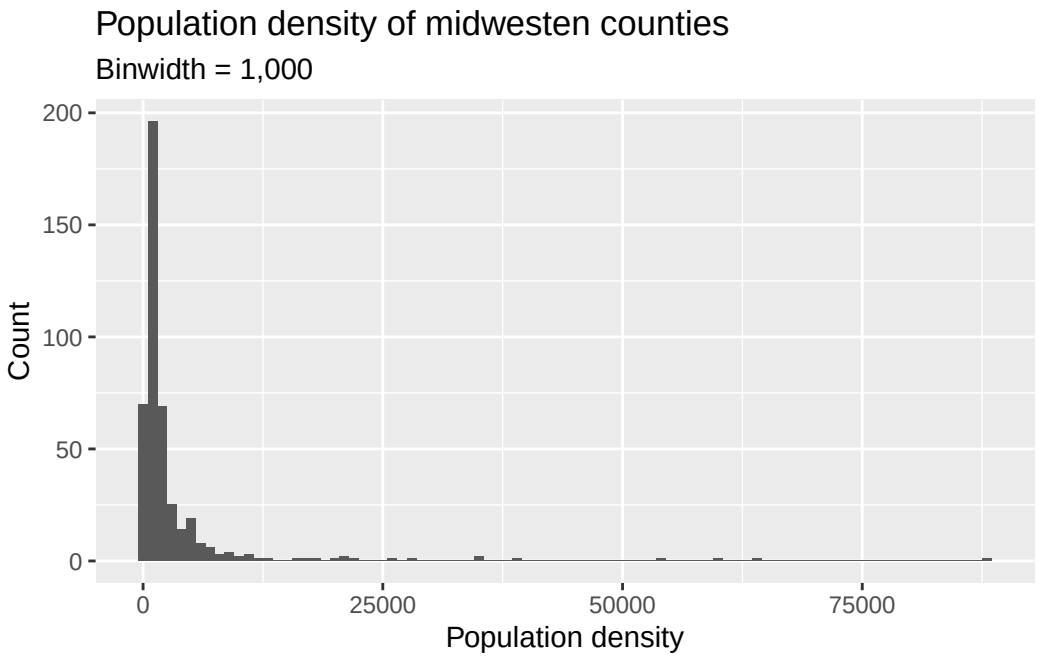
Part 1

Question 1

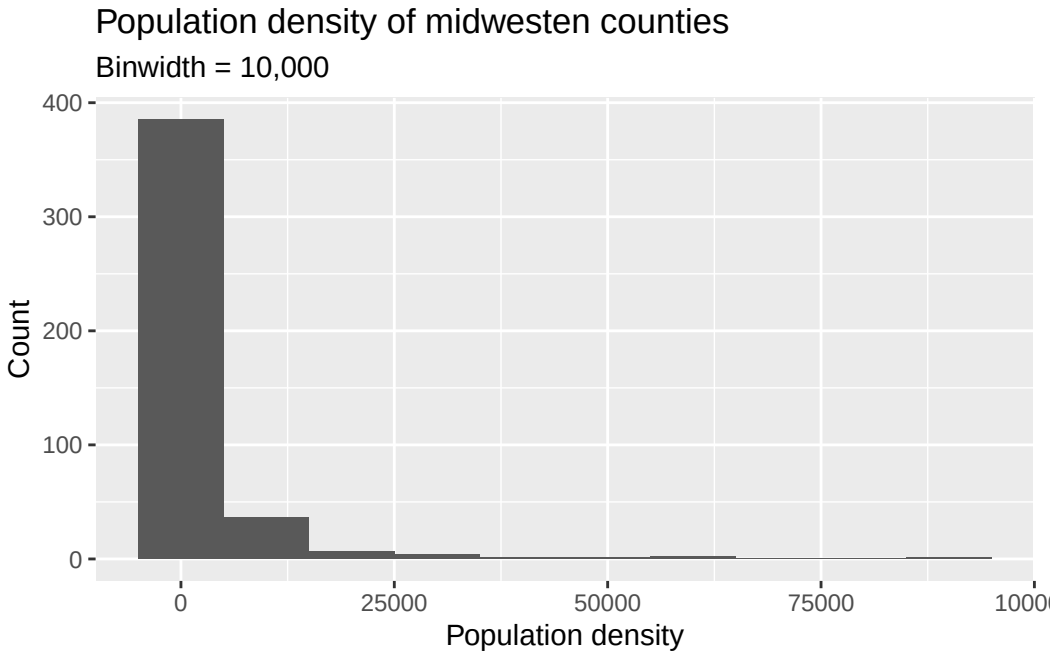
```
ggplot(midwest, aes(x = popdensity)) +
  geom_histogram(binwidth = 100) +
  labs(
    x = "Population density",
    y = "Count",
    title = "Population density of midwesten counties",
    subtitle = "Binwidth = 100"
  )
```



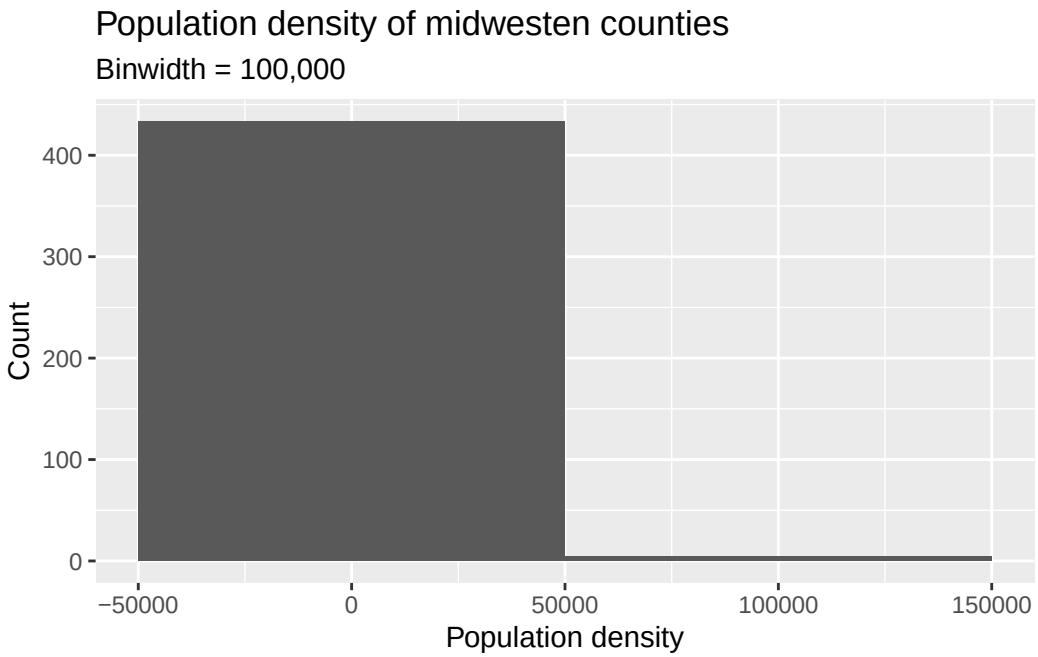

```
ggplot(midwest, aes(x = popdensity)) +
  geom_histogram(binwidth = 1000) +
  labs(
    x = "Population density",
    y = "Count",
    title = "Population density of midwesten counties",
    subtitle = "Binwidth = 1,000"
  )
```




```
ggplot(midwest, aes(x = popdensity)) +
  geom_histogram(binwidth = 10000) +
  labs(
    x = "Population density",
    y = "Count",
    title = "Population density of midwesten counties",
    subtitle = "Binwidth = 10,000"
  )
```



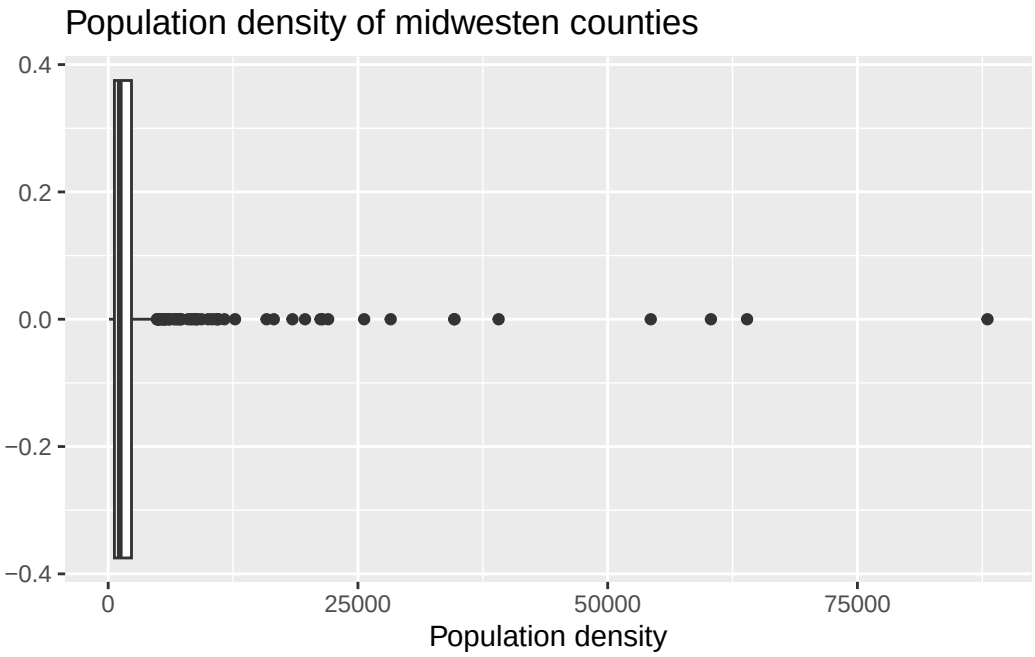

```
ggplot(midwest, aes(x = popdensity)) +
  geom_histogram(binwidth = 100000) +
  labs(
    x = "Population density",
    y = "Count",
    title = "Population density of midwesten counties",
    subtitle = "Binwidth = 100,000"
  )
```



The plot with binwidth 1,000 is the best representation of the data because you are able to see the trend of how the number of counties varies based on population density. With the other bins, they become to big to be able to observe the trend accurately between the two variables or too small to notice the difference.

Question 2

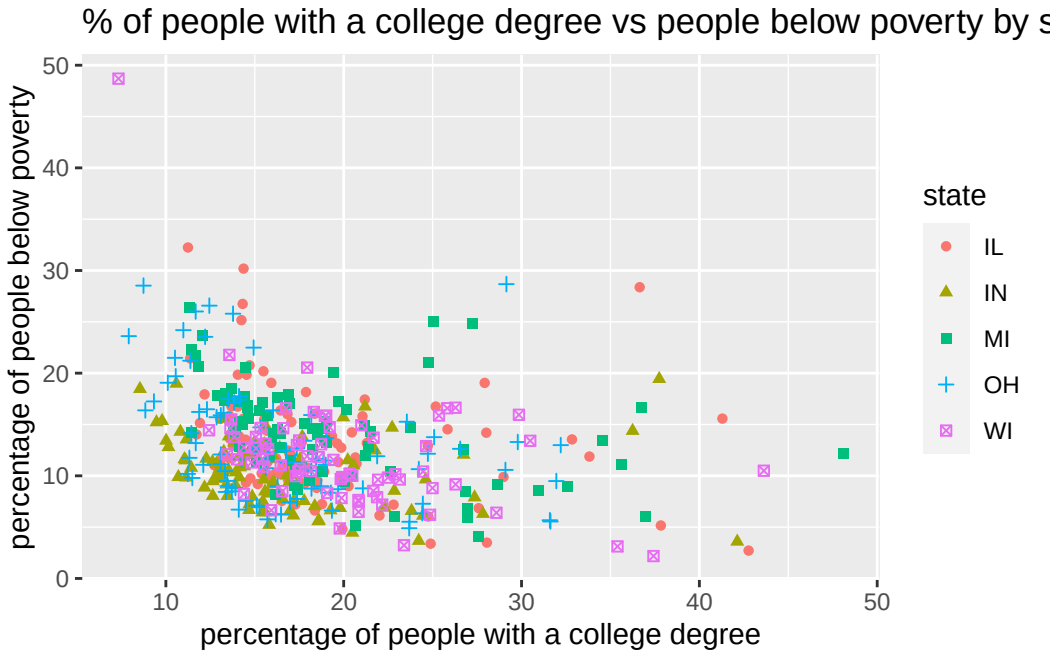
```
ggplot(midwest, aes(x = popdensity)) +
  geom_boxplot() +
  labs(
    x = "Population density",
    title = "Population density of midwesten counties",
  )
```



The distribution of population density of counties is skewed towards the lower population densities. Most counties in the Midwest have a population density close to 0. There are some outliers. One of them is Wayne, MI. This makes sense because Wayne county has a city (Wayne) in Michigan that causes the population density to spike.

Question 3

```
ggplot(midwest, aes(x=percollege, y=percbelowpoverty, color = state, shape = state)) +
    geom_point() +
    labs(
      x = "percentage of people with a college degree",
      y = "percentage of people below poverty",
      title = "% of people with a college degree vs people below poverty by state"
    )
```



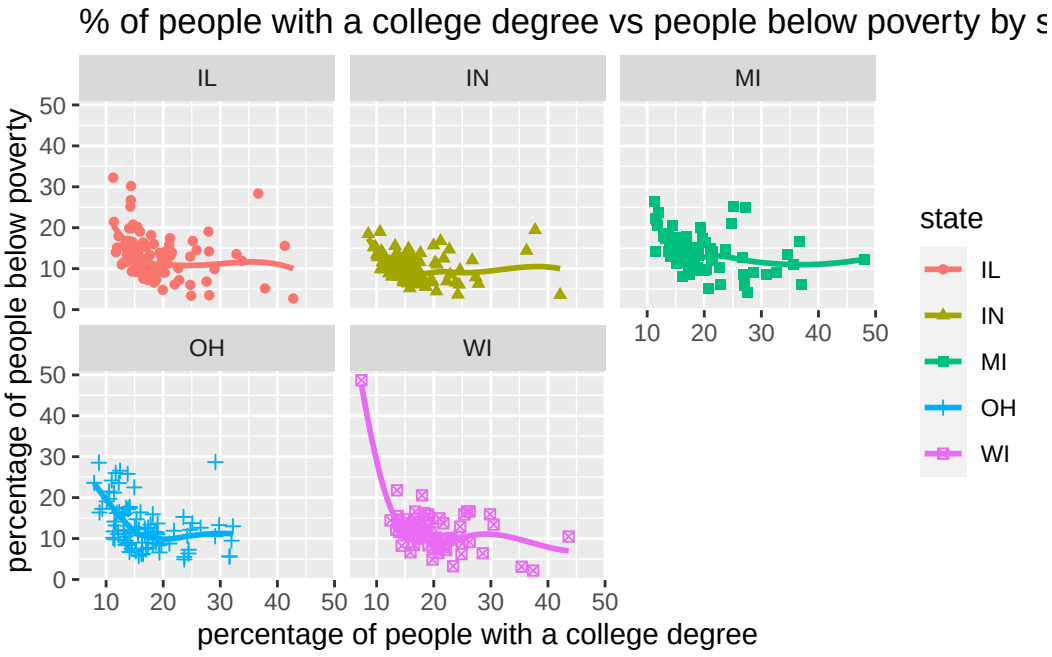
The overall relationship between percentage of people with a college degree and percentage below poverty in Midwestern states is that the lower the percentage of people with a college degree, the higher the percentage of people below poverty is (it is an inverse relationship). One county that is a clear outlier is MENOMINEE, WI. This is because in MENOMINEE, yes, only 7% of people have a college degree, but a shockingly high 48% of people are below poverty, which is more than we would expect.

I can identify that this relationship varies across states due to the color AND shape coding which helps me visualize how each state's counties (i.e., WI = purple squares) reflect or do not reflect this relationship between these two variables.

Question 4

```
ggplot(midwest, aes(x=percollege, y=percbelowpoverty, color = state, shape = state)) +
    geom_point() +
    geom_smooth(se = FALSE) +
    labs(
      x = "percentage of people with a college degree",
      y = "percentage of people below poverty",
      title = "% of people with a college degree vs people below poverty by state"
    ) +
  facet_wrap(~ state)
```

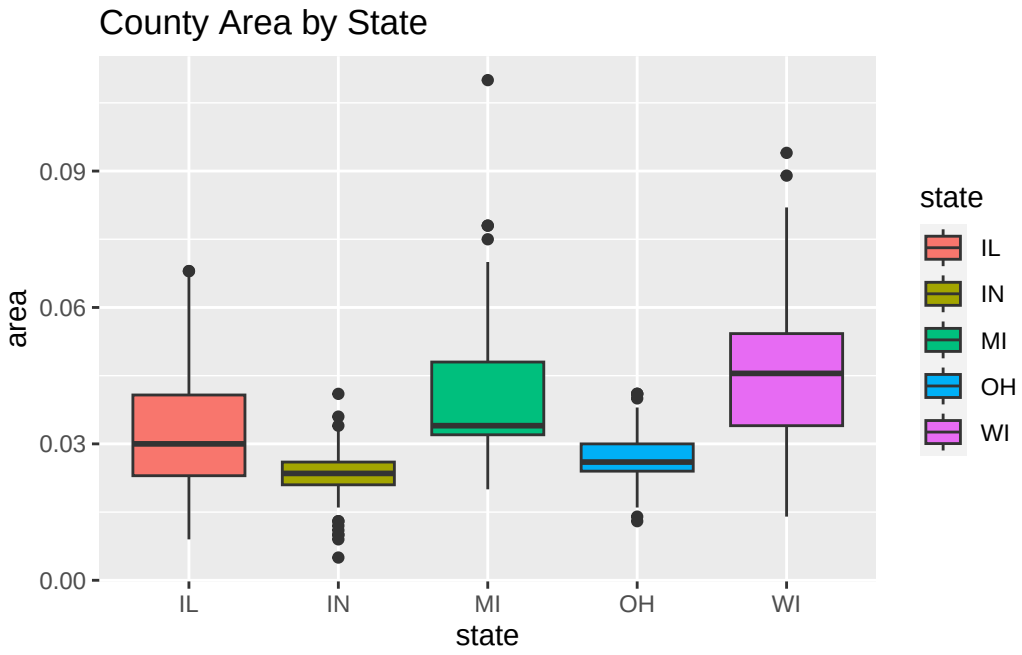
```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



I like this plot more than the plot in question 3 because it makes the data more easily understandable and differentiable by state. I can analyze trends between states more easily.

Question 5

```
#create vertical side-by-side boxplots
ggplot(midwest, aes(x=state, y=area, fill=state)) +
  geom_boxplot() +
  ggtitle('County Area by State')
```

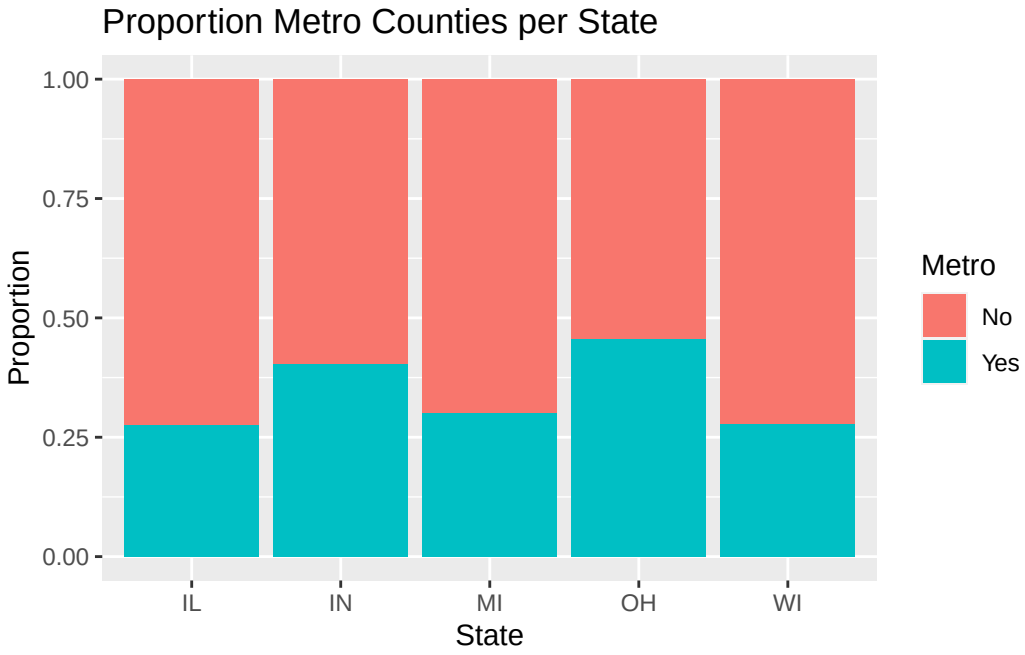


Most county area sizes are pretty similar across states (with medians around 0.03,0.04). However, county sizes vary across states as well. IN has the lowest county area sizes, followed by OH, IL, MI, and then WI (who has the largest county area sizes), in that order.

The state that has the single largest county is MI. The name of this county is MARQUETTE, MI.

Question 6

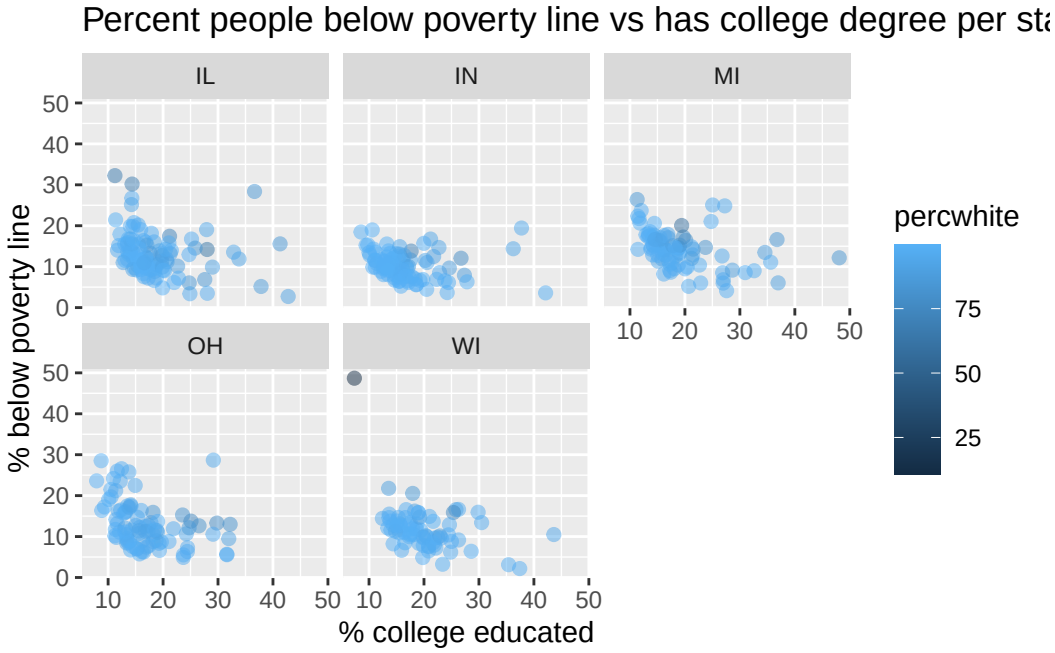
```
midwest <- midwest |>
  mutate(metro = if_else(inmetro == 1, "Yes", "No"))

ggplot(midwest, aes(x = state, fill = metro)) +
  geom_bar(position = "fill") +
  labs(
    x="State",
    y="Proportion",
    title= "Proportion Metro Counties per State",
    fill="Metro"
  )
```



OH and IN have the highest percentage of counties in metropolitan areas (IL, MI, and WI have the lowest percentage of counties in metropolitan areas). OH is around 42% and IN is around 39%, while IL, MI, and WI are around 25-28% in counties that are in metropolitan areas.

Question 7

```
ggplot(midwest, aes(x = percollege, y=percbelowpoverty, color=percwhite)) +
  geom_point(alpha=0.5, size=2) +
  facet_wrap(~state) +
  labs(
    x = "% college educated",
    y = "% below poverty line",
    title = "Percent people below poverty line vs has college degree per state"
  )
```



One county that is a clear outlier in Wisconsin (WI) is MENOMINEE, WI. The population composition of this county is approximately 10% White and 90% American Indian.

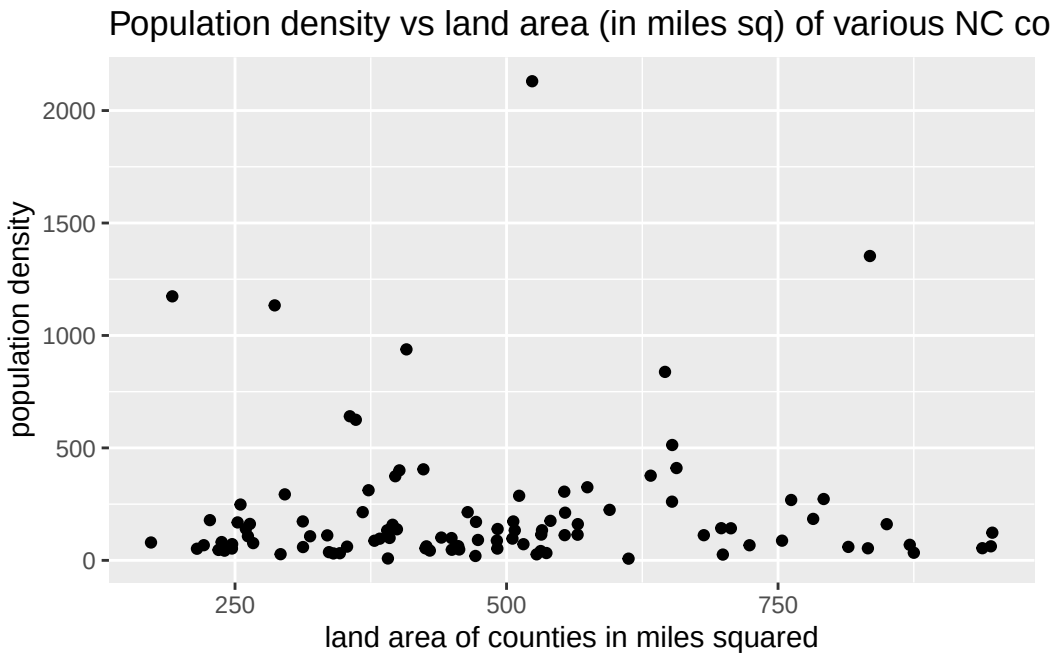
Part 2

```
nc_county <- read_csv("data/nc-county.csv")
```

Question 8

The relationship between population density and land area might be negative. The higher the population density, the lower the land area. That's what major cities are. You don't have a really dense population for a large area.

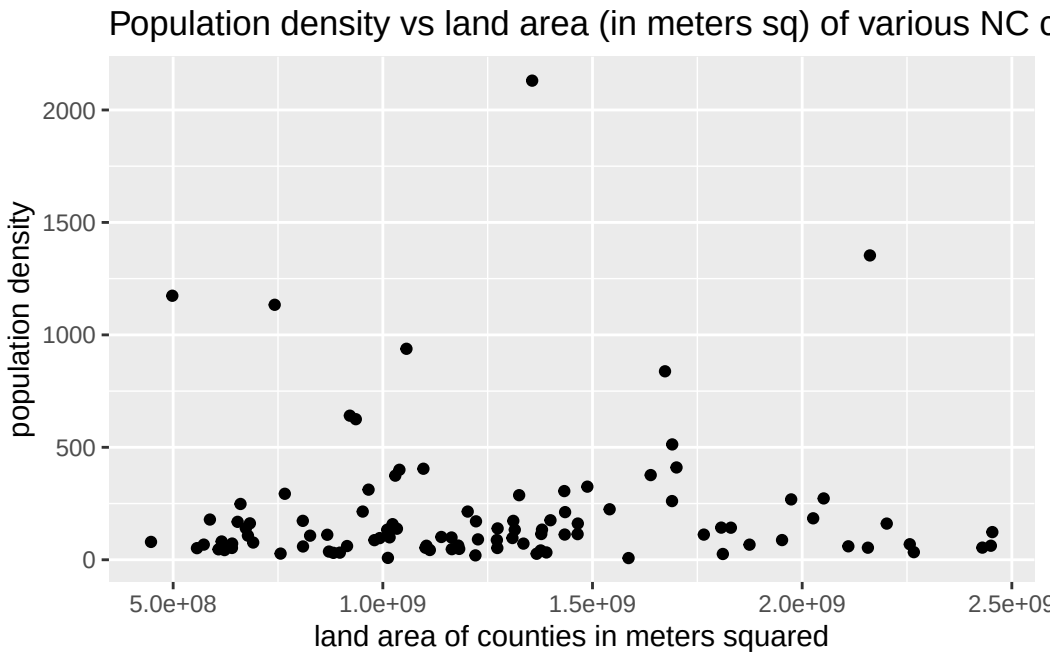
```
ggplot(nc_county, aes(x=land_area_mi2, y=density)) +
  geom_point() +
  labs(
    x= "land area of counties in miles squared",
    y= "population density",
    title= "Population density vs land area (in miles sq) of various NC counties"
  )
```



There seems to be no relationship between population density and land area of NC counties.

Question 9

```
ggplot(nc_county, aes(x=land_area_m2, y=density)) +
  geom_point() +
  labs(
    x= "land area of counties in meters squared",
    y= "population density",
    title= "Population density vs land area (in meters sq) of various NC counties"
  )
```



The relationship observed in this plot (no relationship between the two variables) is the exact same as the relationship observed in the plot for Question 8. Changing the unit in which the land area of counties is measured should not and does not change the relationship between population density and land area.

Question 10

No answer needed here! Just select questions and pages to indicate where your responses are located when you upload your lab PDF to Gradescope and you'll get full points on this question.