Lab 1 - Data visualization

Jessica St. Jean

Questions

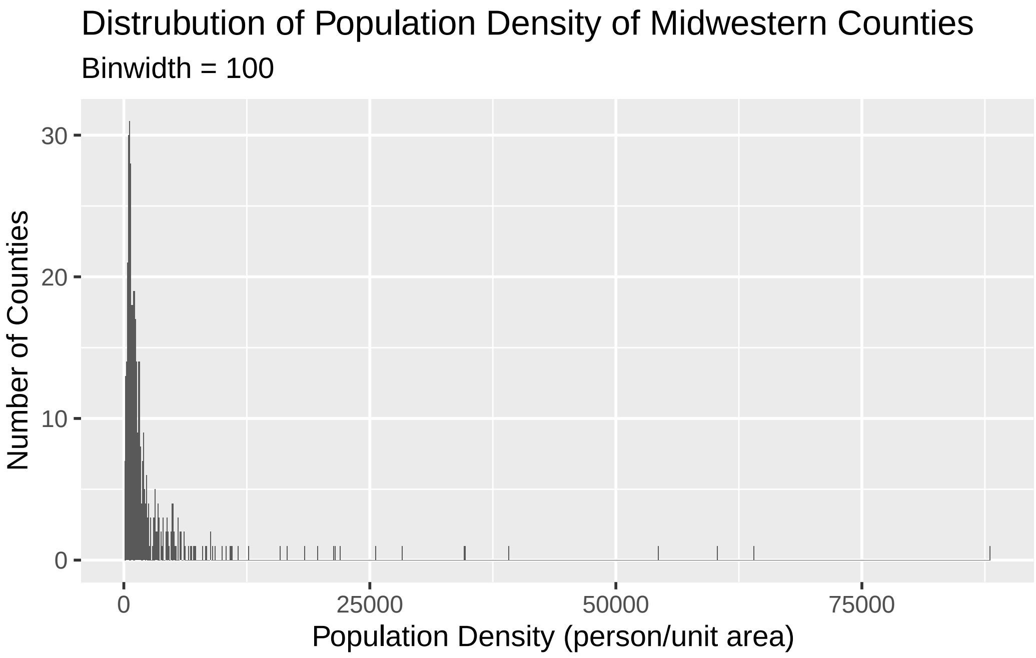
Part 1

```
library(tidyverse)
```

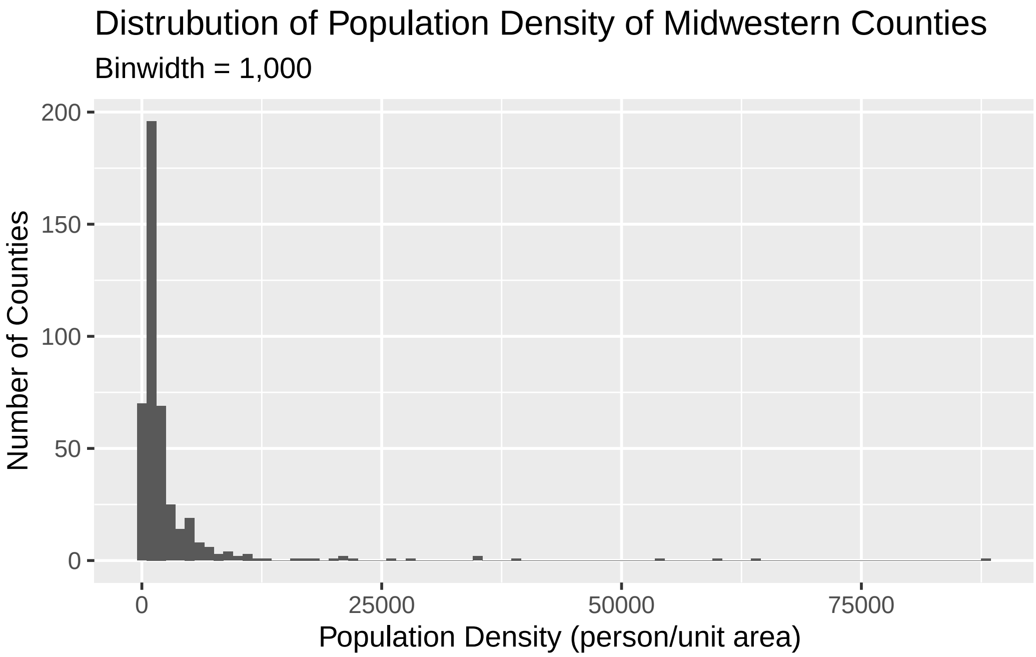
```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.6
v forcats   1.0.1     v stringr   1.6.0
v ggplot2   4.0.1     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.2
v purrr     1.2.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becon
```

Question 1

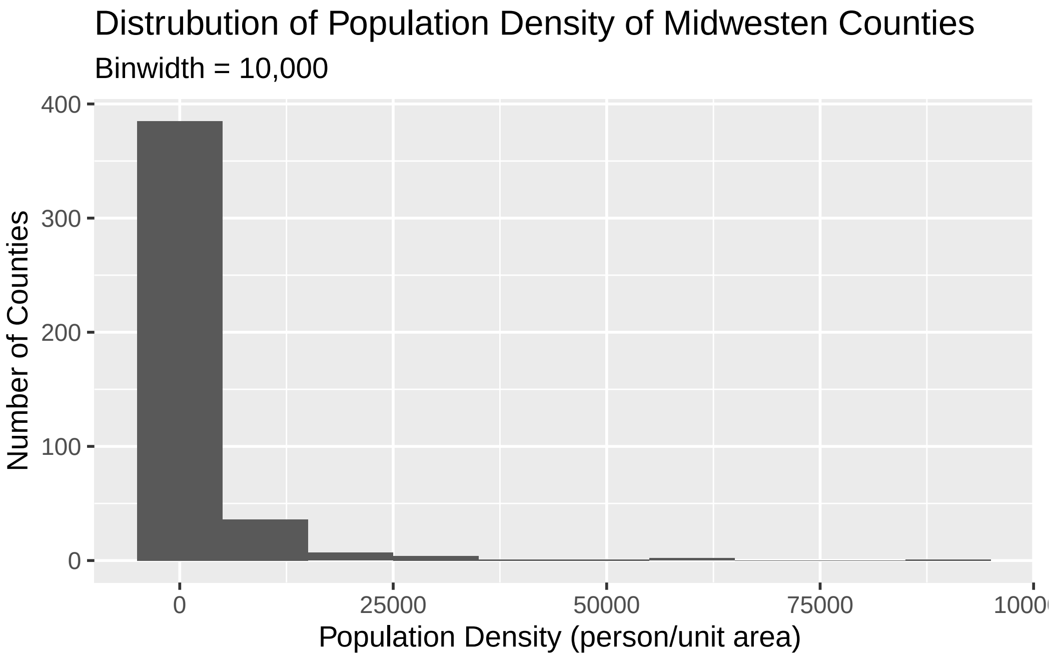
```
ggplot(midwest, aes(x=popdensity))+
  geom_histogram(binwidth = 100) +
  labs(
   x = "Population Density (person/unit area) ",
    y = "Number of Counties",
    title = "Distrubution of Population Density of Midwestern Counties",
    subtitle = "Binwidth = 100"
  )
```



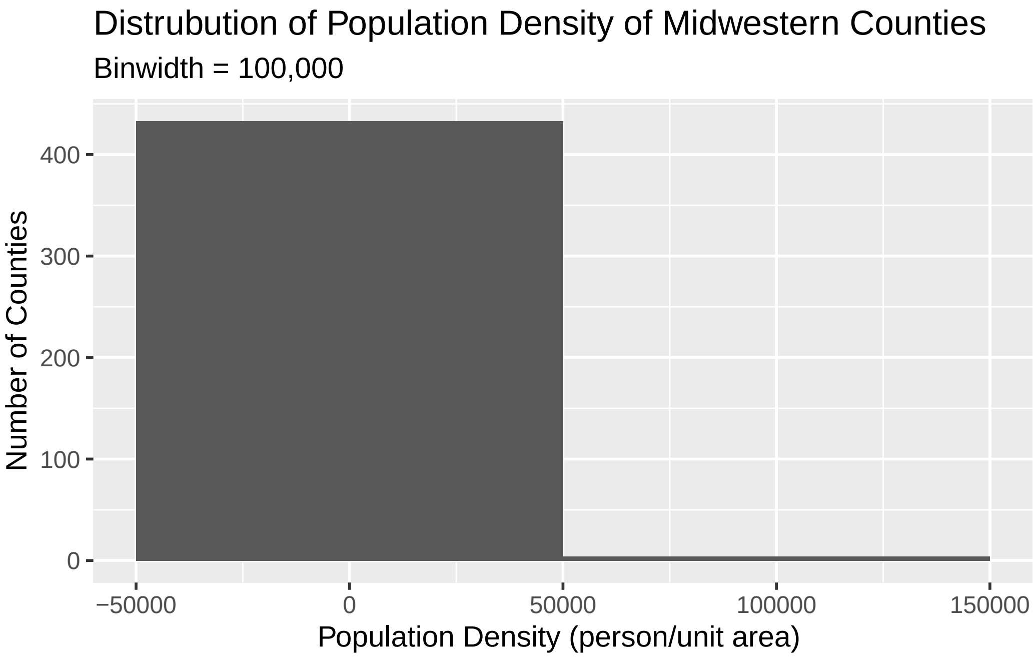
```
ggplot(midwest, aes(x=popdensity))+
  geom_histogram(binwidth = 1000) +
  labs(
     x = "Population Density (person/unit area) ",
    y = "Number of Counties",
    title = "Distrubution of Population Density of Midwestern Counties",
    subtitle = "Binwidth = 1,000"
  )
```



```
ggplot(midwest, aes(x=popdensity))+
  geom_histogram(binwidth = 10000) +
  labs(
    x = "Population Density (person/unit area) ",
    y = "Number of Counties",
    title = "Distrubution of Population Density of Midwesten Counties",
    subtitle = "Binwidth = 10,000"
  )
```



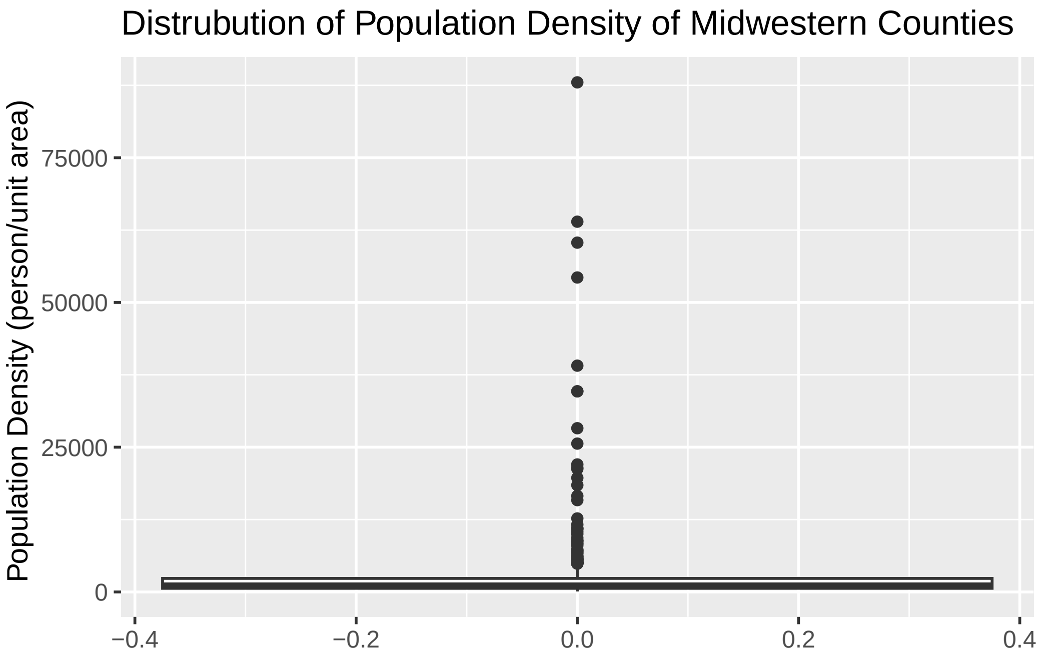
```
ggplot(midwest, aes(x=popdensity))+
  geom_histogram(binwidth = 100000) +
  labs(
    x = "Population Density (person/unit area) ",
    y = "Number of Counties",
    title = "Distrubution of Population Density of Midwestern Counties",
    subtitle = "Binwidth = 100,000"
  )
```



Comment: I think that the most appropriate binwidth is 1,000 because it most clearly illustrates the data, the binwidth of 100 makes it difficult to read because the bins are too small and ambiguous in what number they represent between the labels on the x-axis. The 10,000 and 100,000 binwidths are inappropriate because they group too many population densities together in the 0-25,000 population density range, the 100,000 bandwidth especially. The 10,000 and 100,000 binwidth plots also seem to suggest that there are population densities below 0 which is not possible. Therefore 1,000 is the most well suited because it have enough separation between population densities is not extremely small and does not make there appear to be data below 0.

Question 2

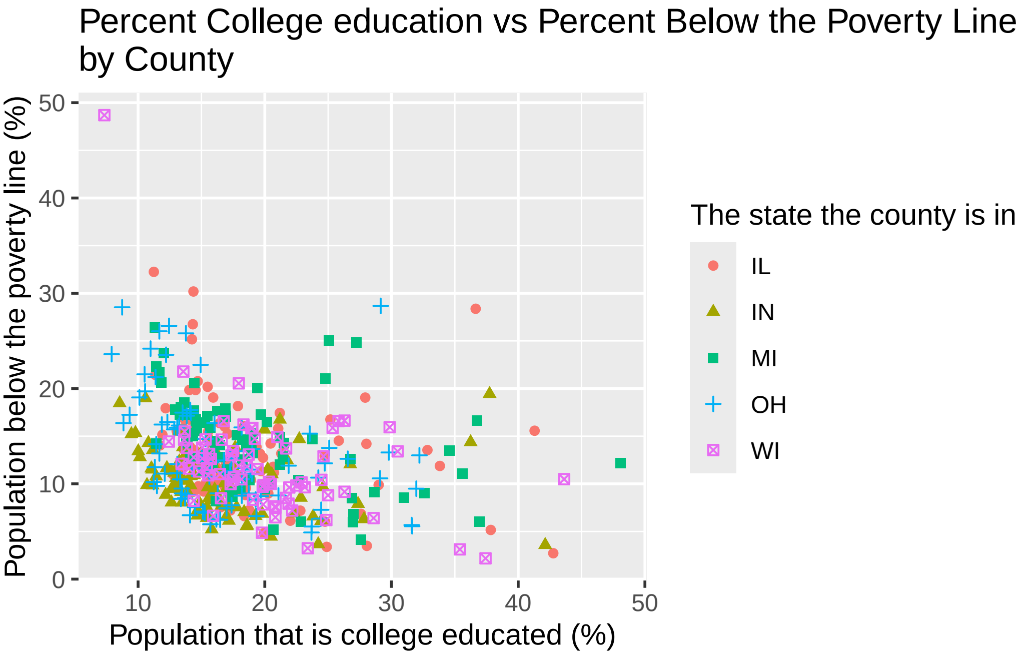
```r
ggplot(midwest, aes(y = popdensity)) +
  geom_boxplot() +
  labs(
    y = "Population Density (person/unit area) ",
    title = "Distrubution of Population Density of Midwestern Counties",)
```



Comment: The distribution of the population density seems to be almost entirely within the lower range (very close to 0) which is also reflected in the previous plots. There are some outliers including the most extreme outlier, Cook County in Illinois. It is hard for me to determine if it makes sense to me that this specific county is an outlier because I am unfamiliar with the counties and their populations within the United States. However, I would assume that this county encompasses a large city or a very dense urban area while the majority of counties in the midwest may be in rural areas and have lower population densities. After looking it up it seems like this county includes the city of Chicago which makes sense and confirms my assumptions.

Question 3

```
ggplot(midwest, aes(x=percollege, y=percbelowpoverty, colour = state, shape = state)) +
geom_point() +
  labs(x= "Population that is college educated (%)",
    y= "Population below the poverty line (%)",
    colour = "The state the county is in",
    shape = "The state the county is in",
    title= "Percent College education vs Percent Below the Poverty Line \nby County")
```

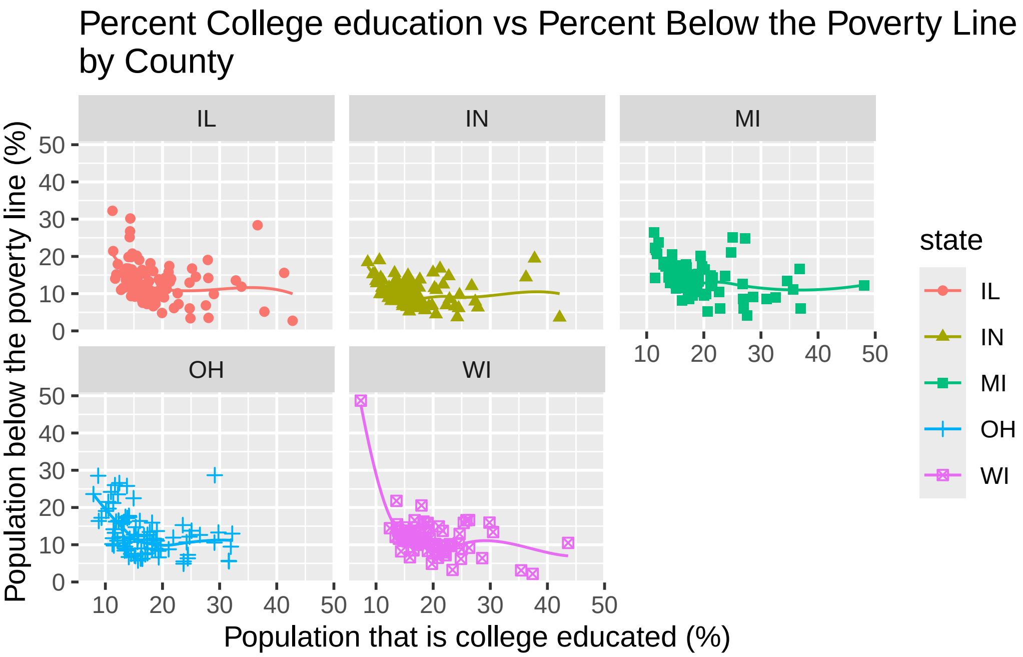


Comment: The general relationship between the percent of the population below the poverty line and the percent of the population that is college educated is that when the percent of college education goes up the percent of the population below the poverty rate goes down. An outlier for this general trend can be seen in the county of Jackson in Illinois, where 36.64% of the population is college educated but 28.37% of the population is below the poverty line. This does not follow the general trend of the rest of the counties with other counties with percent college educated between 30-40 % with poverty rates averaging

Question 4

```
ggplot(midwest, aes(x=percollege, y=percbelowpoverty, colour = state, shape = state)) +
geom_point() +
  geom_smooth(se= FALSE, linewidth = 0.5) +
  facet_wrap(~state) +
  labs(x= "Population that is college educated (%)",
    y= "Population below the poverty line (%)",
    title= "Percent College education vs Percent Below the Poverty Line \nby County")
```

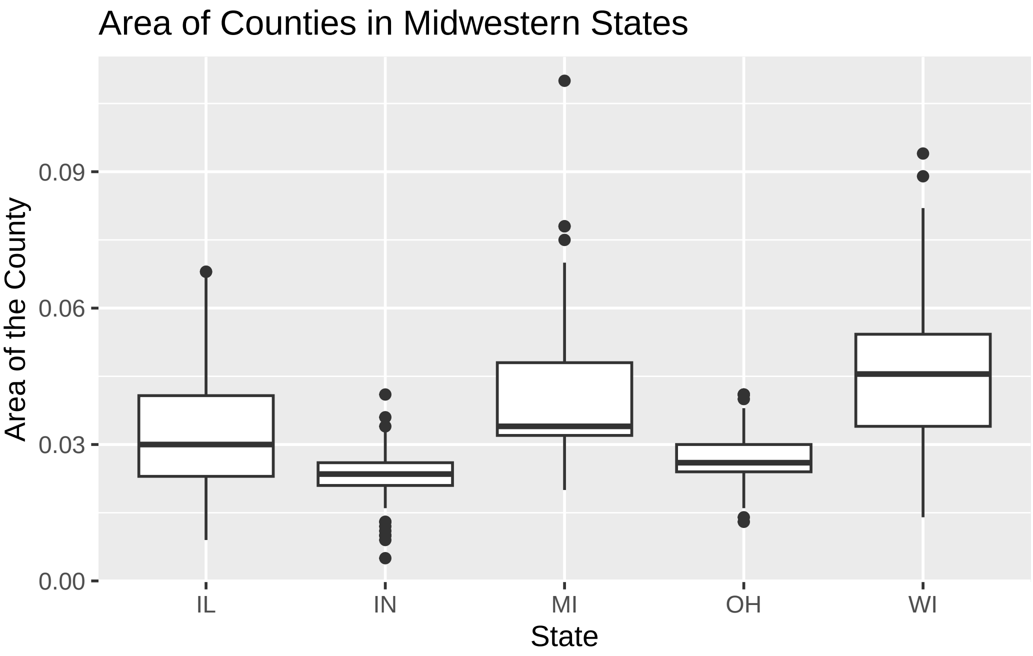
```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Comment: I prefer this plot because it is easier to distinguish the trends of each state as the plot from question 3 had a lot of points all clumped together. this plot still has this somewhat but it lessens it and makes it clearer. The line also helps to illustrate the trend which I prefer.

Question 5

```
ggplot(midwest, aes(x= state, y = area)) +
  geom_boxplot() +
  labs(
    y = "Area of the County",
    x = "State",
    title = "Area of Counties in Midwestern States",)
```

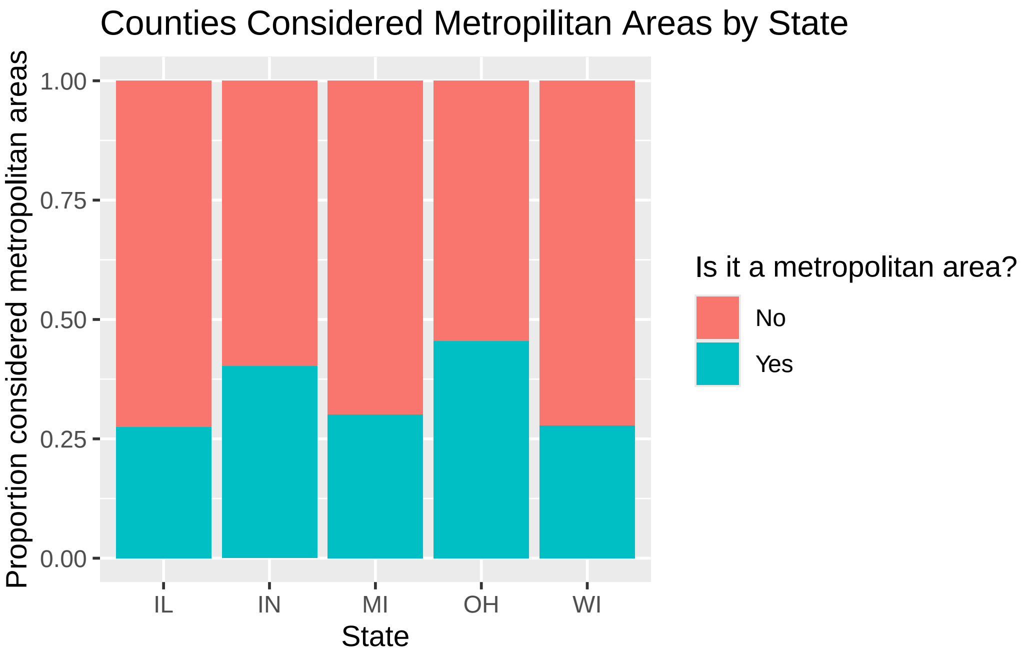


Comment: The size of the counties does not seem to vary a large amount between them however, Ohio and Indiana do seem to have generally smaller counties than Illinois, Michigan, and Wisconsin. There are also larger variations in size within Illinois, Michigan and Wisconsin. The state with the largest county is Michigan, it is Marquette County.

Question 6

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

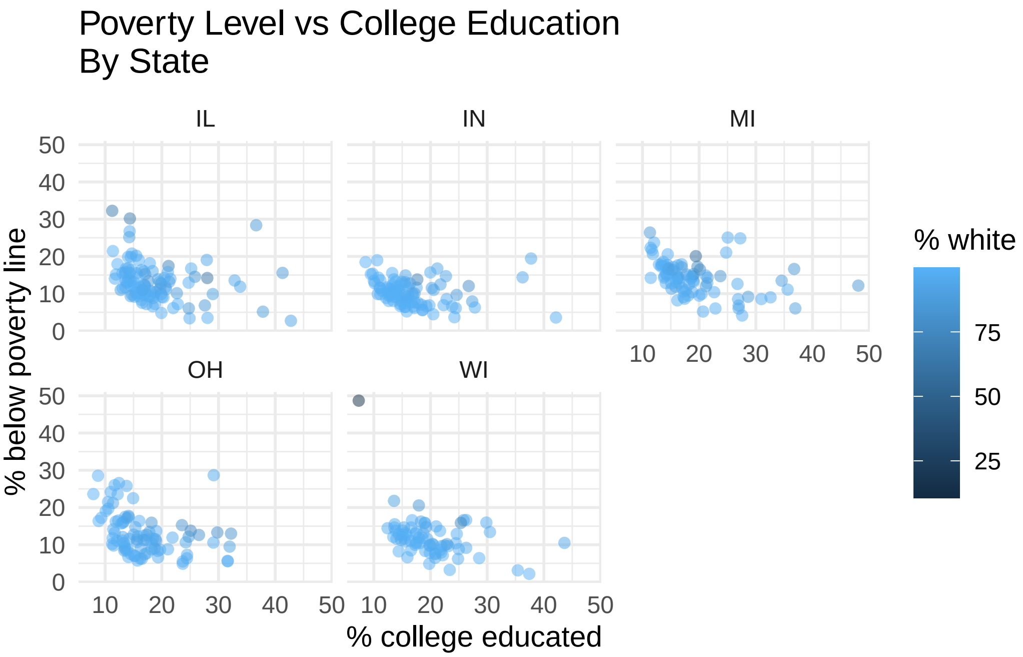
```r
ggplot(midwest, aes(x= state, fill = metro)) +
  geom_bar(position = "fill") +
  labs(
    x= "State",
    y= "Proportion considered metropolitan areas",
    title= "Counties Considered Metropilitan Areas by State",
    fill = "Is it a metropolitan area?"
  )
```



Comment: Comparing between the states the proportions range from the lowest, approximately 26% of counties in Illinois and Wisconsin being counted as metropolitan areas to the highest, approximately 45% in Ohio. Indiana with approximately 38% and Michigan with approximately 27% fall in between. This shows that the amount of counties in Ohio that are considered metropolitan areas is almost double the amount in Illinois, Wisconsin, and Michigan. While it is not quite double we can see that there is a significantly larger amount in Ohio.

Question 7

```
ggplot(midwest, aes(x=percollege, y=percbelowpoverty, colour= percwhite,))+
  geom_point(alpha=0.5) +
 facet_wrap(~state) +
  labs(
    x= "% college educated",
    y = "% below poverty line",
    colour = "% white",
    title ="Poverty Level vs College Education\nBy State"
  )+
  theme_minimal()
```



Comment: Menominee county is an outlier in Wisconsin, it has a low percentage of people that are college educated and a high percent of people below the poverty line. This county has an extremely high percentage of Native American residents, at 89.18% with 0% of the population being Black or Asian and a very small percentage of White residents or residents labeled as Other.

Note: I did not include the size as 2 in my code because it made the points look way larger than the points in the reference graph.

Part 2

Enough about the Midwest!

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

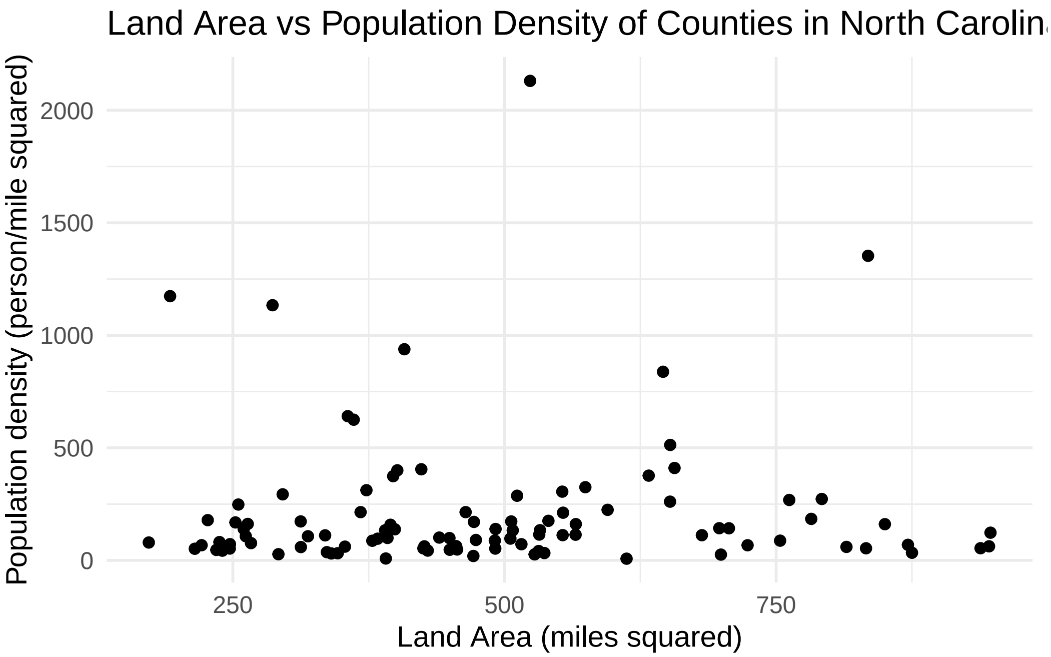
```
Rows: 100 Columns: 7
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): county, state_abb, state_name
dbl (4): land_area_m2, land_area_mi2, population, density

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Question 8

Comment: I believe it will be a negative relationship where as land area increases the population density decreases.

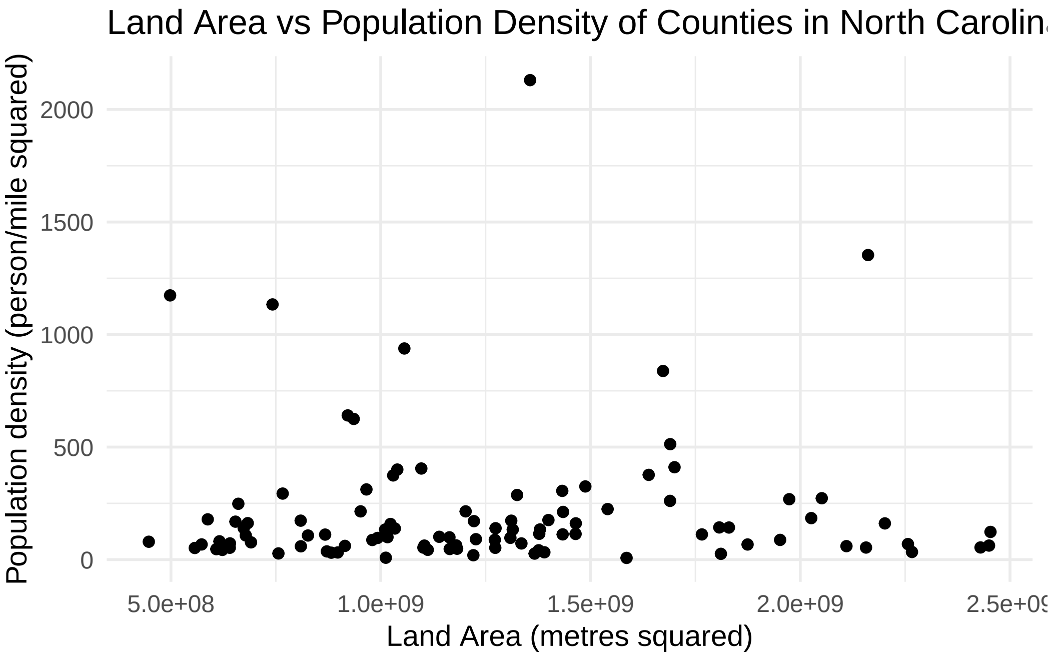
```
ggplot(nc_county, aes(x=land_area_mi2, y=density)) +
geom_point() +
  labs(
    x="Land Area (miles squared)",
    y = "Population density (person/mile squared)",
    title = "Land Area vs Population Density of Counties in North Carolina"
  )+
  theme_minimal()
```



Comment: There doesn’t seem to be a strong relationship which is not what I expected.

Question 9

```
ggplot(nc_county, aes(x=land_area_m2, y=density)) +
geom_point() +
  labs(
    x="Land Area (metres squared)",
    y = "Population density (person/mile squared)",
    title = "Land Area vs Population Density of Counties in North Carolina"
  )+
  theme_minimal()
```



Comment: The relationship (or lack of one) is the same as the plot in question 8, this is due to the points not moving, the population density was calculated in and uses square miles, so those numbers well remain the same. The numbers that change are the x-axis which is now scaled to be in metres squared, so the points do not move but the new x-axis changes the meaning/context.