Lab 1 - Data visualization

Deandra Rasheesa Maheswari

Questions

Part 1

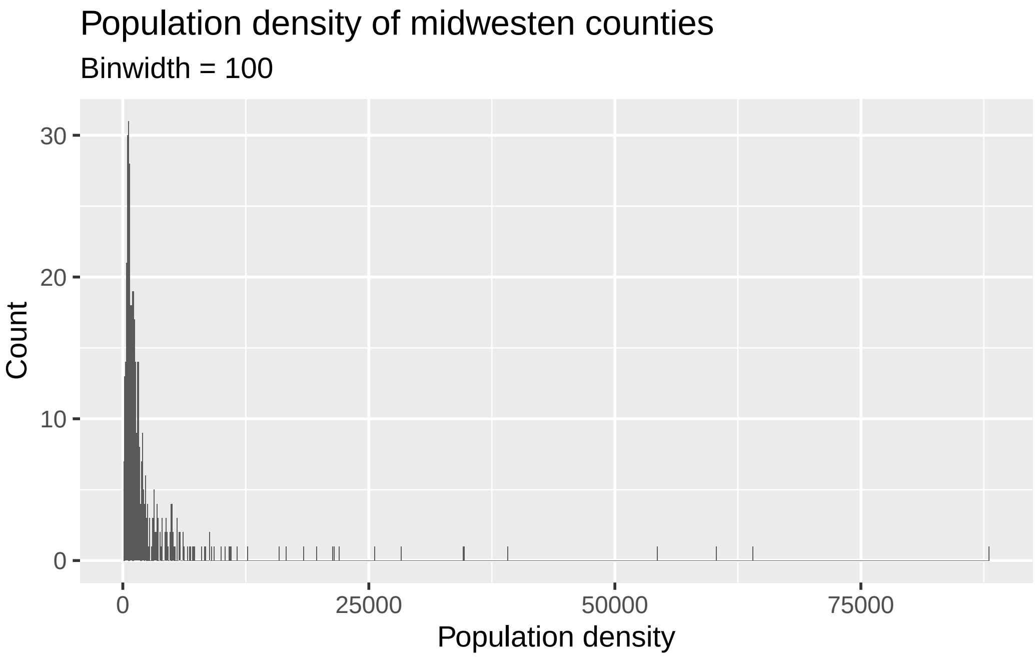
```r
library(tidyverse)
```

```
-- 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 beco
```

Question 1

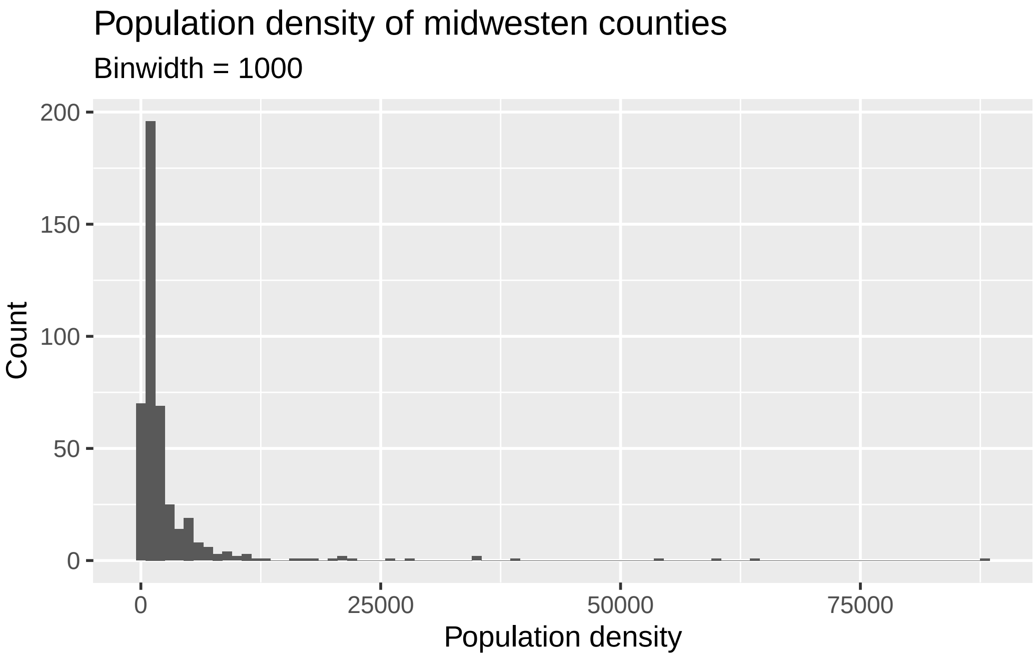
Binwidth = 100

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



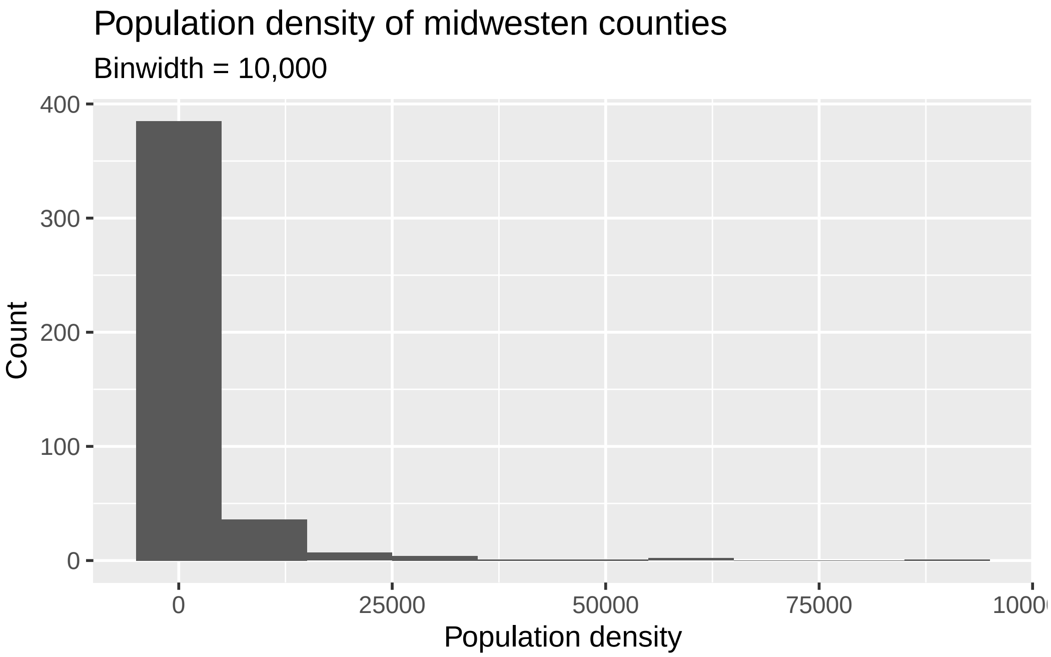
Binwidth = 1000

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



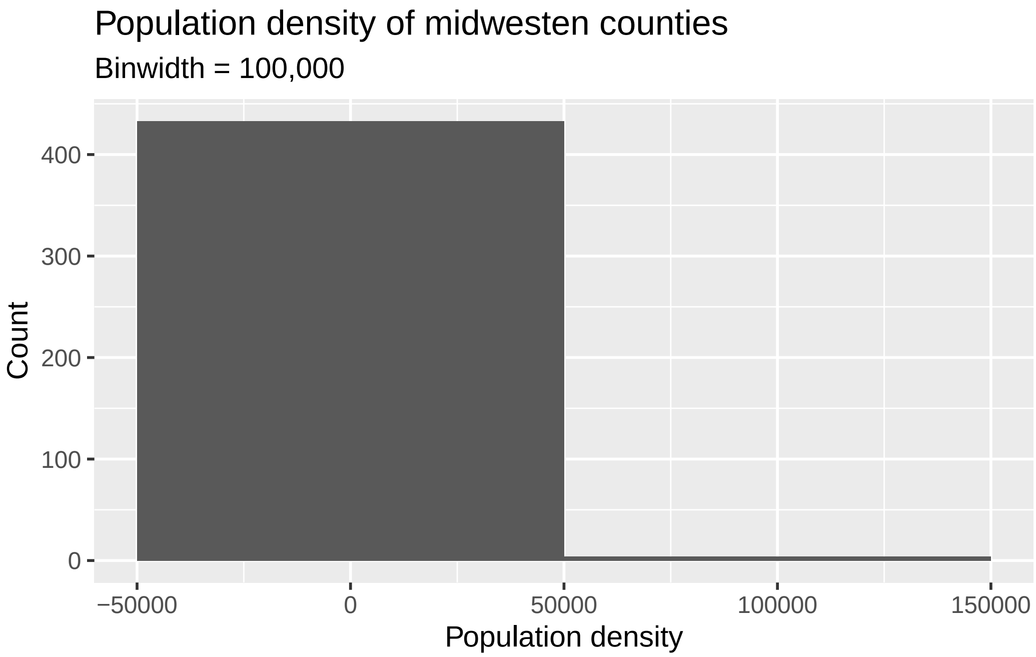
Binwidth = 10,000

```r
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"
  )
```



Binwdith = 100,000

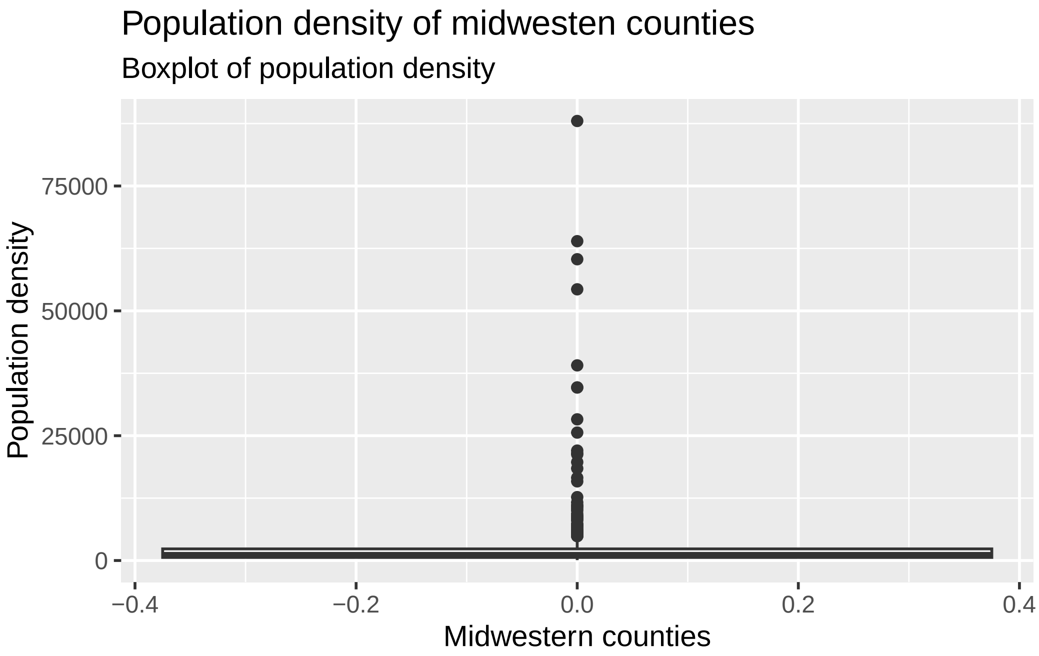
```
ggplot(midwest, aes(x = popdensity)) +
  geom_histogram(binwidth = 100000) +
  labs(
    x = "Population density",
    y = "Count",
    title = "Population density of midwesten counties",
    subtitle = "Binwidth = 100,000"
  )
```



A graph with binwidth 1,000 is the most appropriate because it shows the overall distribution clearly without losing important detail, not too noisy nor oversmooth.

Question 2

```r
ggplot(midwest, aes(y = popdensity)) +
  geom_boxplot() +
  labs(
    x = "Midwestern counties",
    y = "Population density",
    title = "Population density of midwesten counties",
    subtitle = "Boxplot of population density"
  )
```

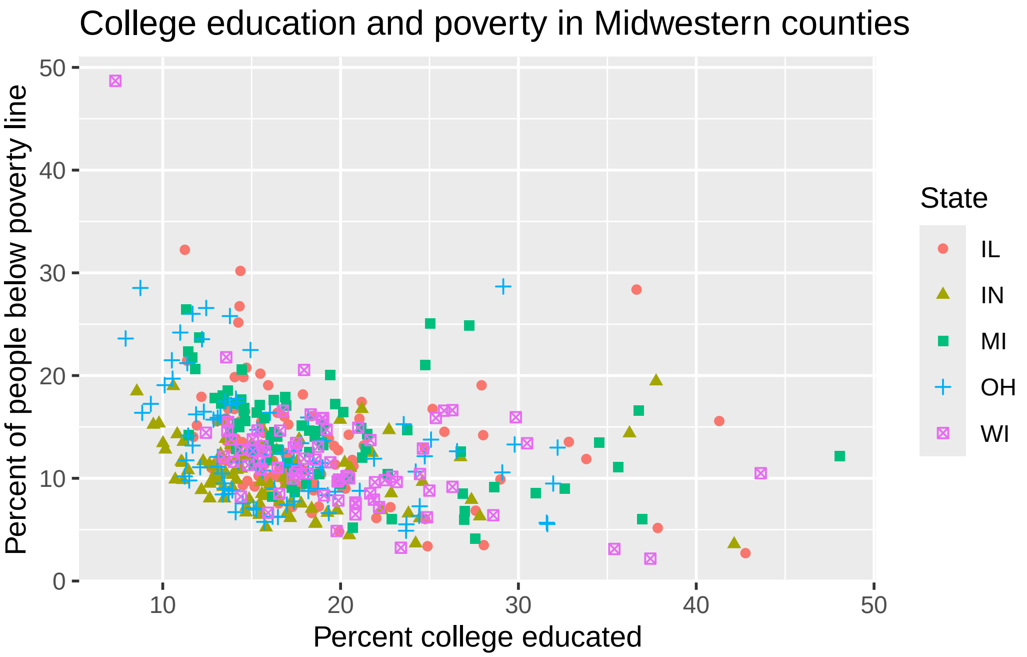


```r
view(midwest)
```

The distribution of population density among Midwestern counties is highly right-skewed, which means most counties having relatively low population density. One clear outlier is Cook County, IL, which has much higher population density than most other counties.

Question 3

```r
ggplot(midwest, aes(x = percollege , y = percbelowpoverty,
                    color = state, shape = state)) +
  geom_point() +
  labs(
    x = "Percent college educated",
    y = "Percent of people below poverty line",
    color = "State",
    shape = "State",
    title = "College education and poverty in Midwestern counties",
  )
```



Overall, there is a negative relationship between percentage of people with a college degree and the percentage of people living below the poverty line. Counties with higher levels of college education tend to have lower poverty rates, while counties with lower levels of college education tend to have higher poverty rates.

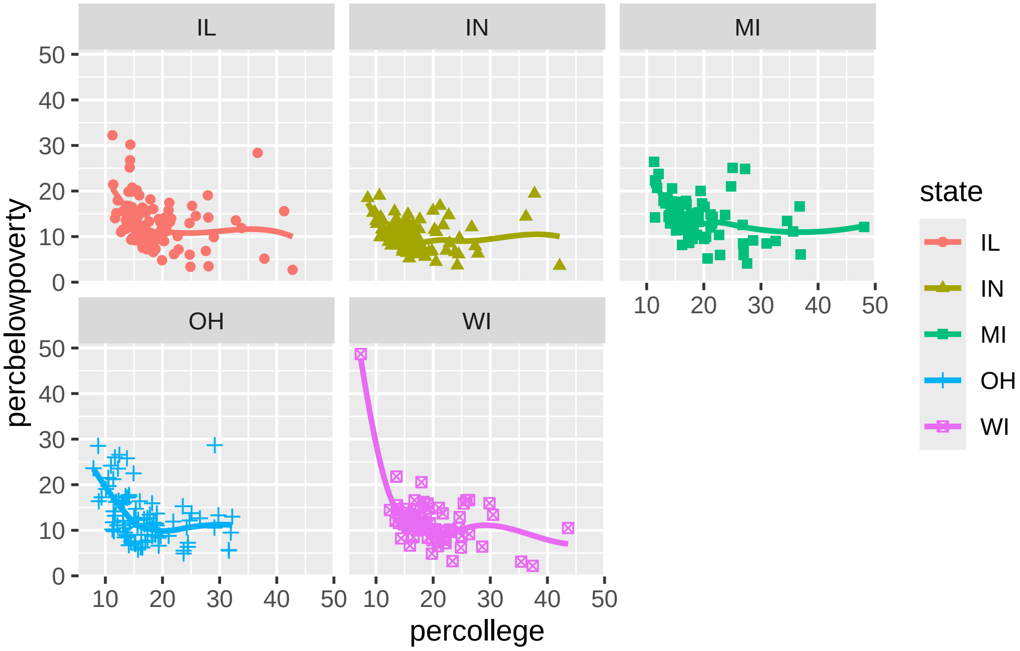
```
view(midwest)
```

From the scatter plot above, WI with square and purple point is the outlier. Based on the data, it can be seen that Minominee, WI has 48% of people below poverty with 7% of people college educated.

Question 4

```
ggplot(midwest, aes(x = percollege , y = percbelowpoverty,
                    color = state, shape = state)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  facet_wrap( ~ state)
```

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

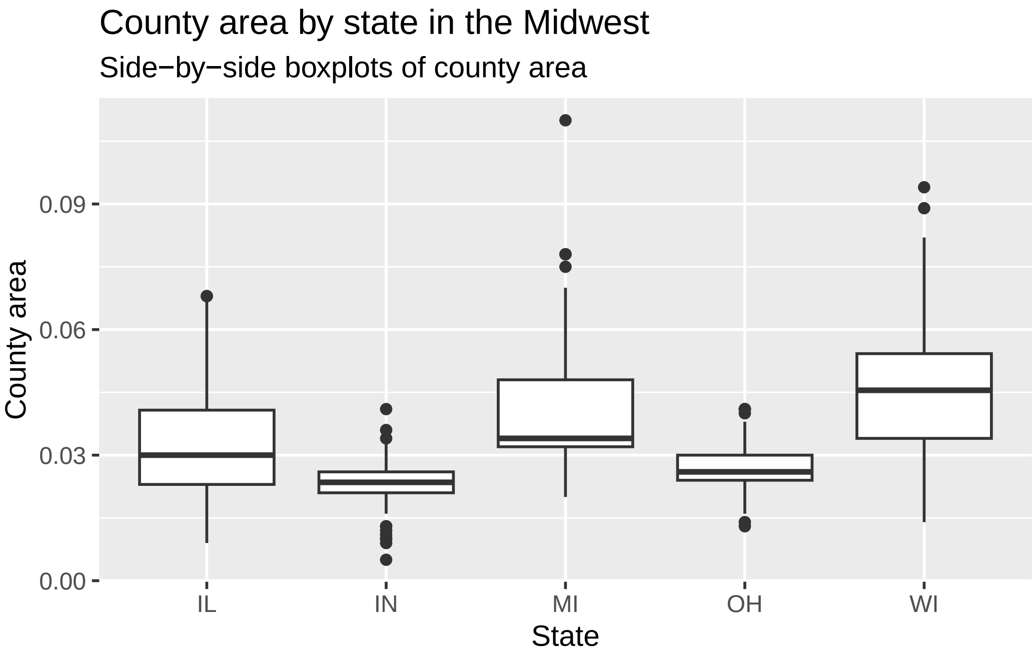


```
  labs(
    x = "Percent college educated",
    y = "Percent of people below poverty line",
    shape = "State",
    color = "State",
    title = "College education and poverty in Midwestern counties"
  )
```

```
<ggplot2::labels> List of 5
 $ x      : chr "Percent college educated"
 $ y      : chr "Percent of people below poverty line"
 $ shape  : chr "State"
 $ colour: chr "State"
 $ title  : chr "College education and poverty in Midwestern counties"
```

Question 5

```
ggplot(midwest, aes(x = state, y = area)) +
  geom_boxplot() +
  labs(
    x = "State",
    y = "County area",
    title = "County area by state in the Midwest",
    subtitle = "Side-by-side boxplots of county area"
  )
```



Typical county sizes vary across Midwestern states. Wisconsin and Michigan tend to have larger counties on average, as indicated by higher median county areas. However, Indiana and Ohio generally have smaller counties with lower median values.

Variability in county size also differs by state. Wisconsin and Michigan show greater variability, shown by taller boxes and longer whiskers. While Indiana and Ohio have more compact box plot which means the county sizes in these states are less variable.

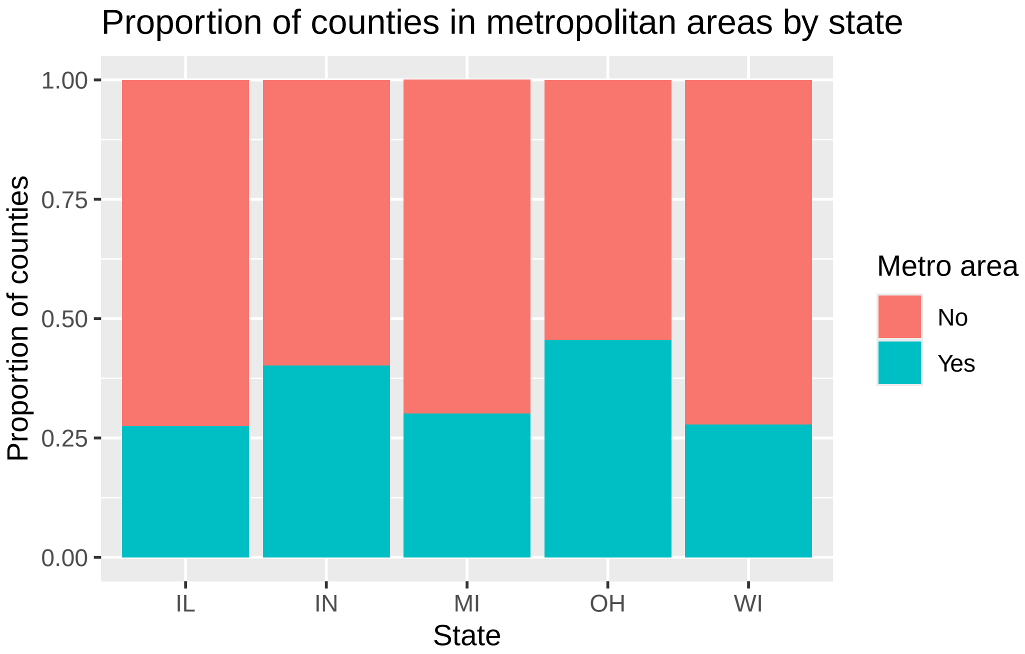
```
view(midwest)
```

The single largest county by area is Marquette, Michigan with 0.110. Furthermore, the top 10 of county areas are dominated by Wisconsin and Michigan and this result is reasonable given Wisconsin and Michigan are the largest states.

Question 6

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

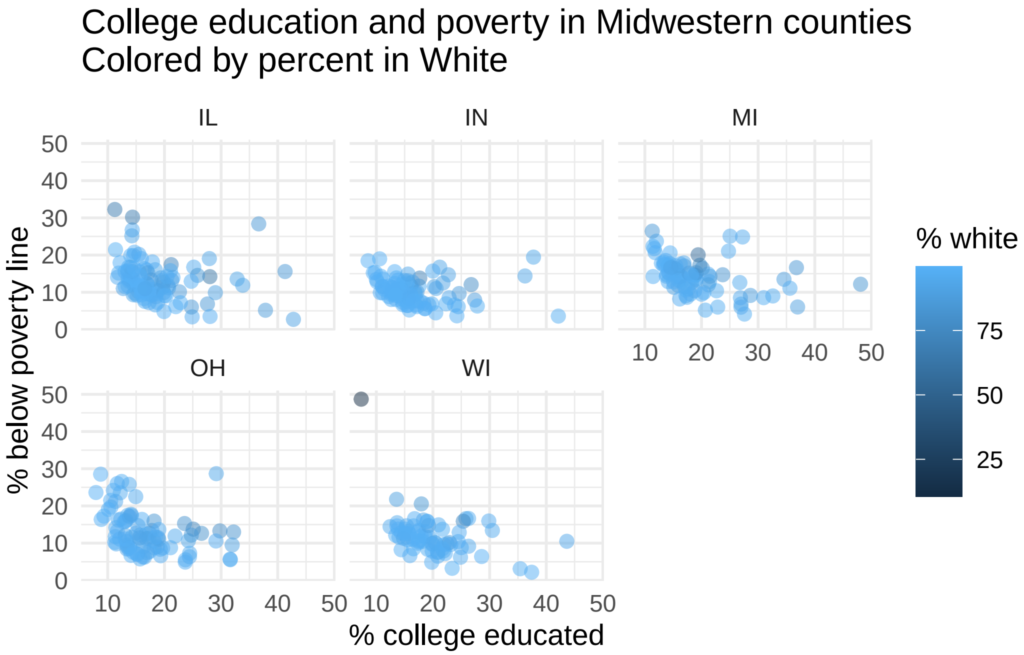
```r
ggplot(midwest, aes(x = state, fill = metro))+
  geom_bar(position = "fill")+
  labs(
    x = "State",
    y = "Proportion of counties",
    fill = "Metro area",
    title = "Proportion of counties in metropolitan areas by state"
  )
```



The share of counties located in metropolitan areas varies across Midwestern states. Ohio has the highest proportion of metro counties, for about 45%. Indiana follows with about 40% of counties classified as metropolitan. Michigan has around 30%, while Illinois and Wisconsin have lower proportions at 25%.

Question 7

```r
ggplot(midwest, aes(x = percollege, y = percbelowpoverty, color = percwhite)) +
  geom_point(size = 2, alpha = 0.5) +
  facet_wrap( ~ state, ncol = 3) +
  labs (
    x = "% college educated",
    y = "% below poverty line",
    color = "% white",
    title = "College education and poverty in Midwestern counties\nColored by percent in Whi
  ) +
  theme_minimal()
```



A clear outlier in Wisconsin is Menominee , Wisconsin. When looking at population composition, this county has a noticeably higher percentage of residents classified as "other races" compared to most counties, indicated with darker point.

Part 2

Enough about the Midwest!

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

```
Rows: 100 Columns: 7
-- Column specification --------------------------------------------------------
Delimiter: ","
```

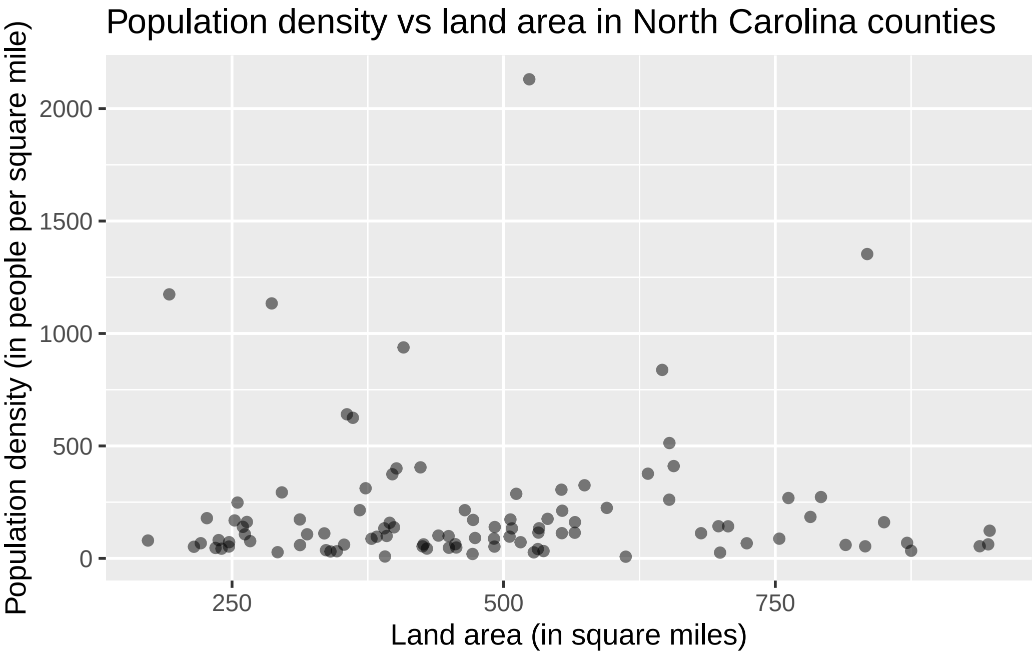
```
chr (3): county, state_abb, state_name
dbl (4): land_area_m2, land_area_mi2, population, density

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

I expect a negative relationship between land area and population density. For instance, land with larger land area tends to be more rural thereby the population density is lower than the smaller land area.

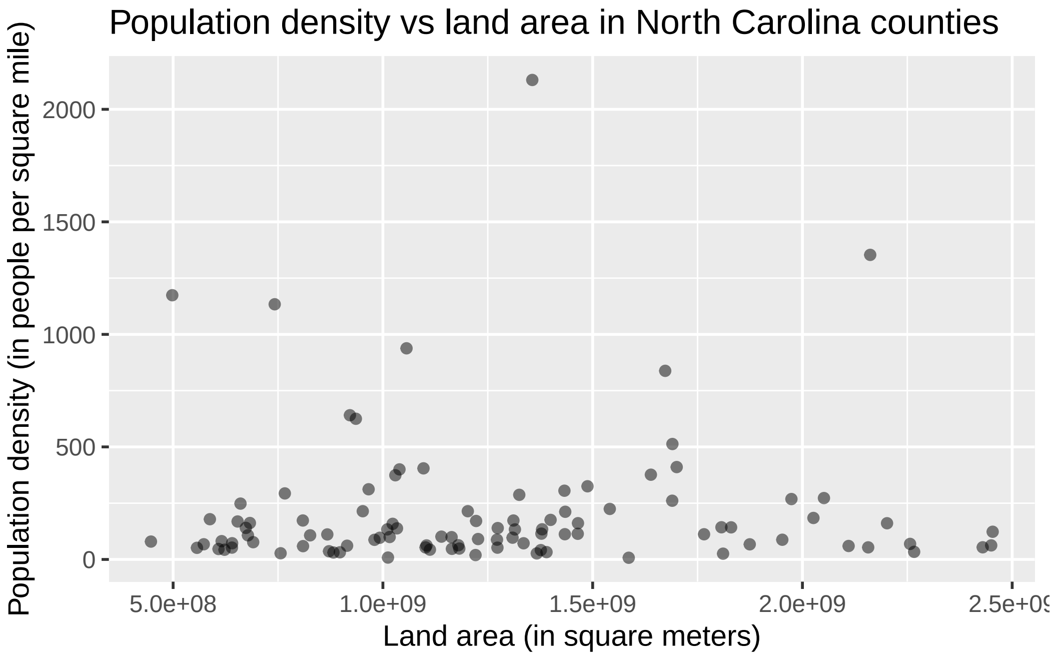
```
ggplot(nc_county, aes(x = land_area_mi2, y = density)) +
  geom_point(alpha = 0.5) +
  labs(
    x = "Land area (in square miles)",
    y = "Population density (in people per square mile)",
    title = "Population density vs land area in North Carolina counties"
  )
```



The plot shows a negative relationship between land area and population density. Smaller counties tend to have higher population densities, while larger counties generally have lower population densities.

Question 9

```r
ggplot(nc_county, aes(x = land_area_m2, y = density)) +
  geom_point(alpha = 0.5) +
  labs(
    x = "Land area (in square meters)",
    y = "Population density (in people per square mile)",
    title = "Population density vs land area in North Carolina counties"
  )
```



The scatter plot using land area in square meters shows the same negative relationship as the plot using land area in square miles. The only difference is the scale of the x-axis, which is much larger when land area is measured in square meters. Changing the units does not change the underlying relationship between land area and population density.