Lab 1 - Data visualization

Ayden Frost

Questions

Part 1

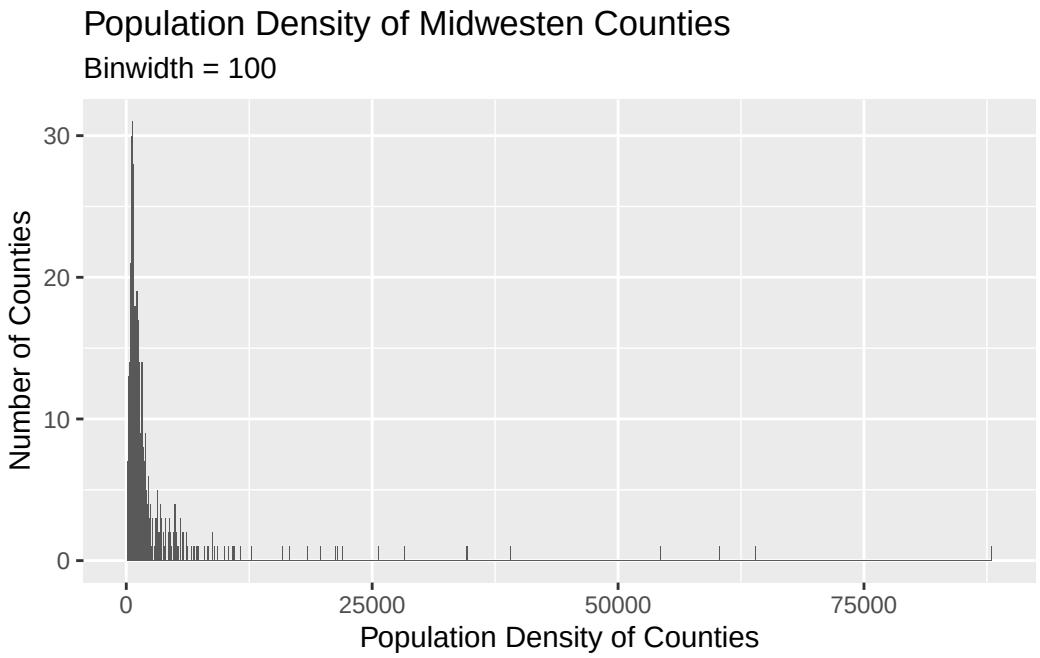
```
library(tidyverse)
```

```
-- 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 beco
```

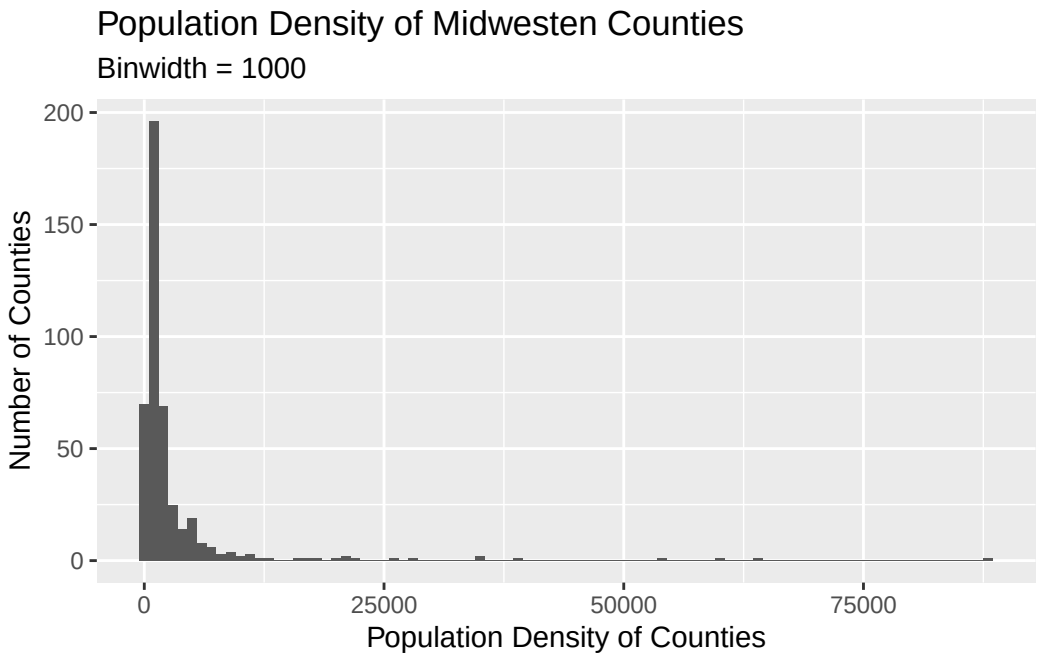
```
data("midwest")
```

Question 1

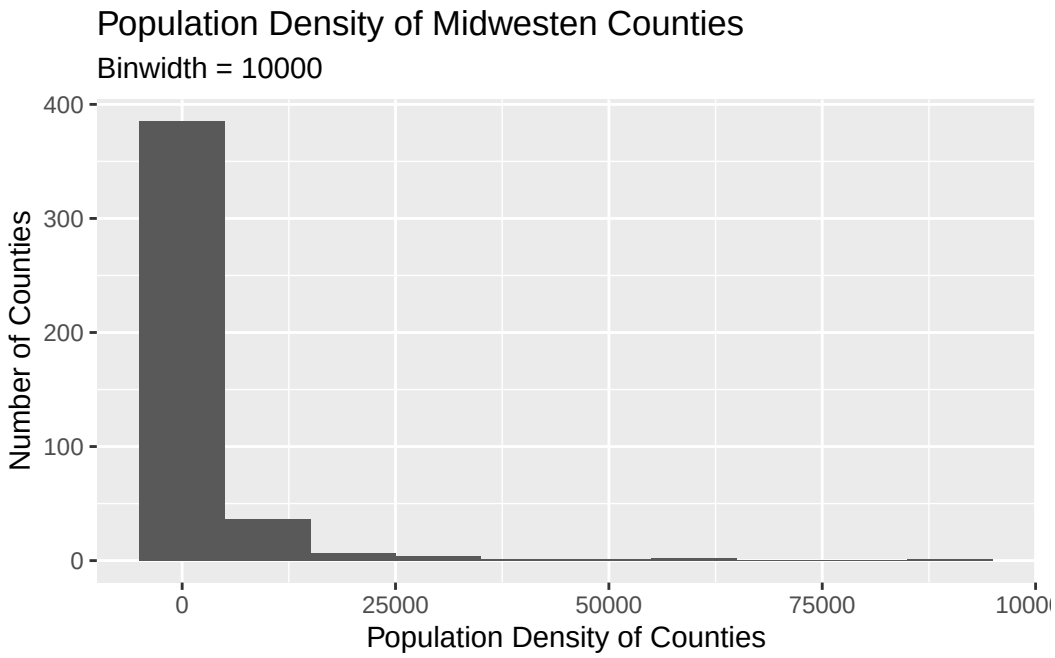
```
ggplot(midwest, aes(x = popdensity)) +
  geom_histogram(binwidth = 100) +
  labs(
    x = "Population Density of Counties",
    y = "Number of Counties",
    title = "Population Density of Midwesten Counties",
    subtitle = "Binwidth = 100"
  )
```



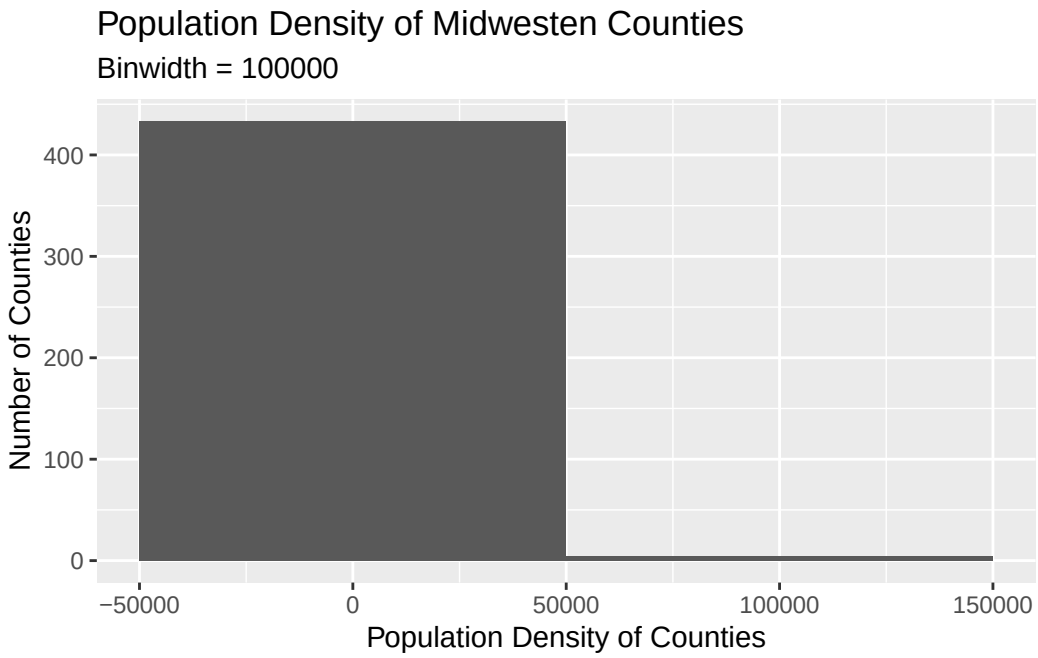

```
ggplot(midwest, aes(x = popdensity)) +
  geom_histogram(binwidth = 1000) +
  labs(
    x = "Population Density of Counties",
    y = "Number of Counties",
    title = "Population Density of Midwesten Counties",
    subtitle = "Binwidth = 1000"
  )
```




```
ggplot(midwest, aes(x = popdensity)) +
  geom_histogram(binwidth = 10000) +
  labs(
    x = "Population Density of Counties",
    y = "Number of Counties",
    title = "Population Density of Midwesten Counties",
    subtitle = "Binwidth = 10000"
  )
```



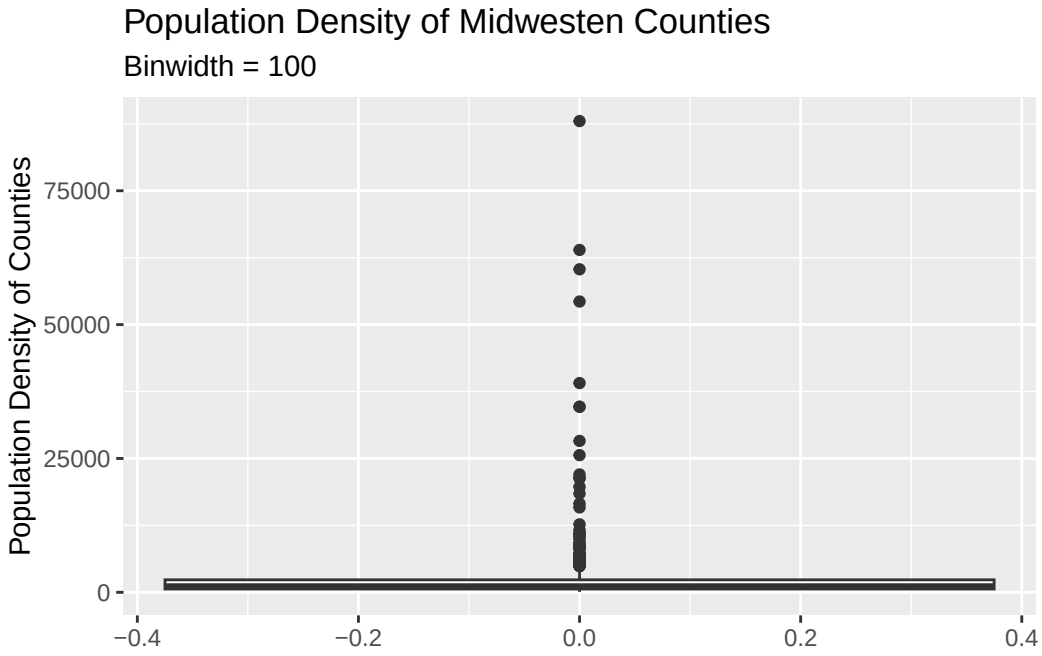

```
ggplot(midwest, aes(x = popdensity)) +
  geom_histogram(binwidth = 100000) +
  labs(
    x = "Population Density of Counties",
    y = "Number of Counties",
    title = "Population Density of Midwesten Counties",
    subtitle = "Binwidth = 100000"
  )
```



The 1000 Binwidth histogram would be ideal as it visualizes the data in a presentable way that is easy to interpret when compared to the other 3 histograms.

Question 2

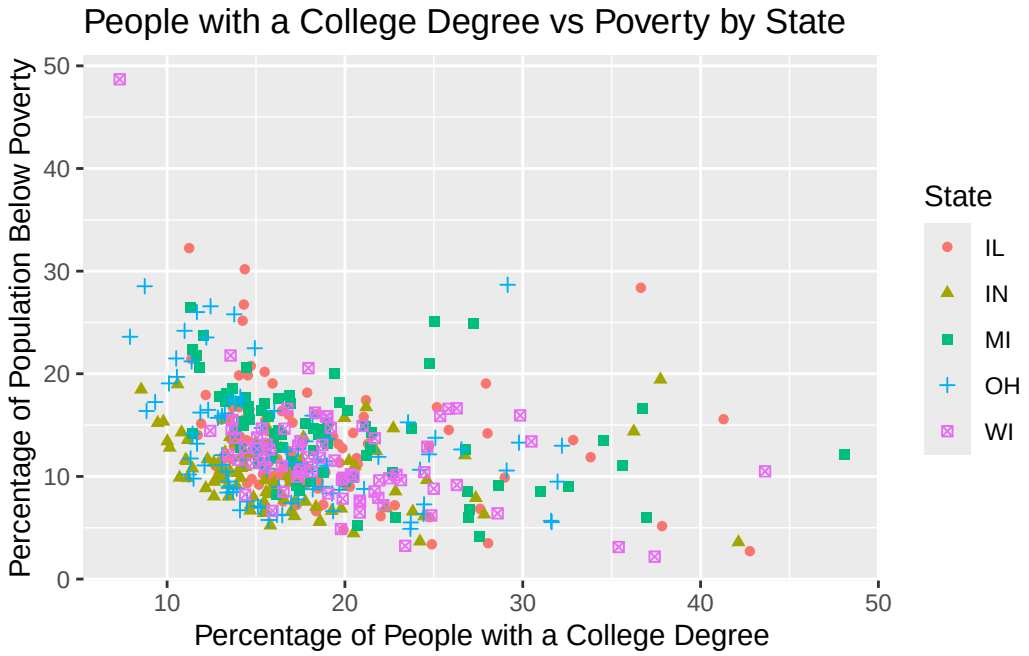
```
ggplot(midwest, aes(y = popdensity)) +
  geom_boxplot() +
  labs(
      y = "Population Density of Counties",
    title = "Population Density of Midwesten Counties",
    subtitle = "Binwidth = 100"
  )
```



The distribution of population sizes amongst counties displayed in both the histogram and boxplot show that most counties tend to be under 12,500 individuals per unit area, with only a small handful of counties breaking the threshold. One county that stands out is Cook county, as it has a population density of roughly 88,000. The most likely reason behind this could be the existence of a city or larger community within the county.

Question 3

```
ggplot(midwest, aes(y = percbelowpoverty, x = percollege, color = state, shape = state)) +
  geom_point() +
  labs(
    color = "State",
    shape = "State",
    x = "Percentage of People with a College Degree",
    y = "Percentage of Population Below Poverty",
    title = "People with a College Degree vs Poverty by State"
  )
```

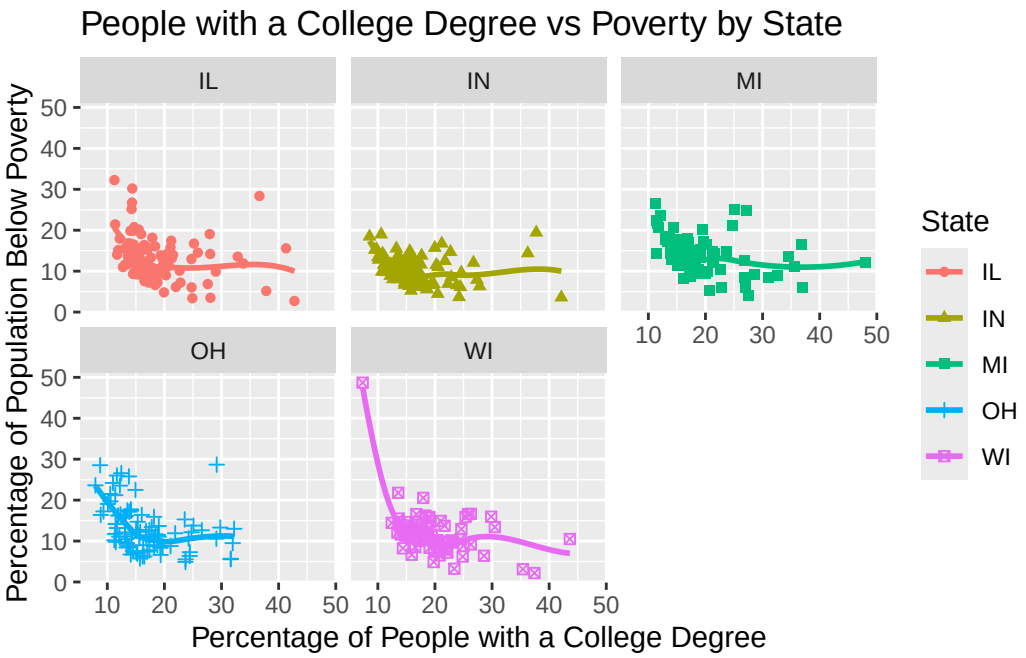


The scatterplot shows that as populations become more educated and have degrees, they tend to see fewer people in poverty when compared to less educated populations. The one county that stood out greatly is MENOMINEE county where nearly 50% of the population sits below the poverty line while only roughly 7% have a degree. It would be very difficult to determine the state trends as the data within the scatterplot is very congested and hard to read if trying to make inferences on the states.

Question 4

```
ggplot(midwest, aes(y = percbelowpoverty, x = percollege, color = state, shape = state)) +
  geom_point() +
  facet_wrap(~ state) +
  geom_smooth(se = FALSE) +
  labs(
    color = "State",
    shape = "State",
    x = "Percentage of People with a College Degree",
    y = "Percentage of Population Below Poverty",
    title = "People with a College Degree vs Poverty by State"
  )
```

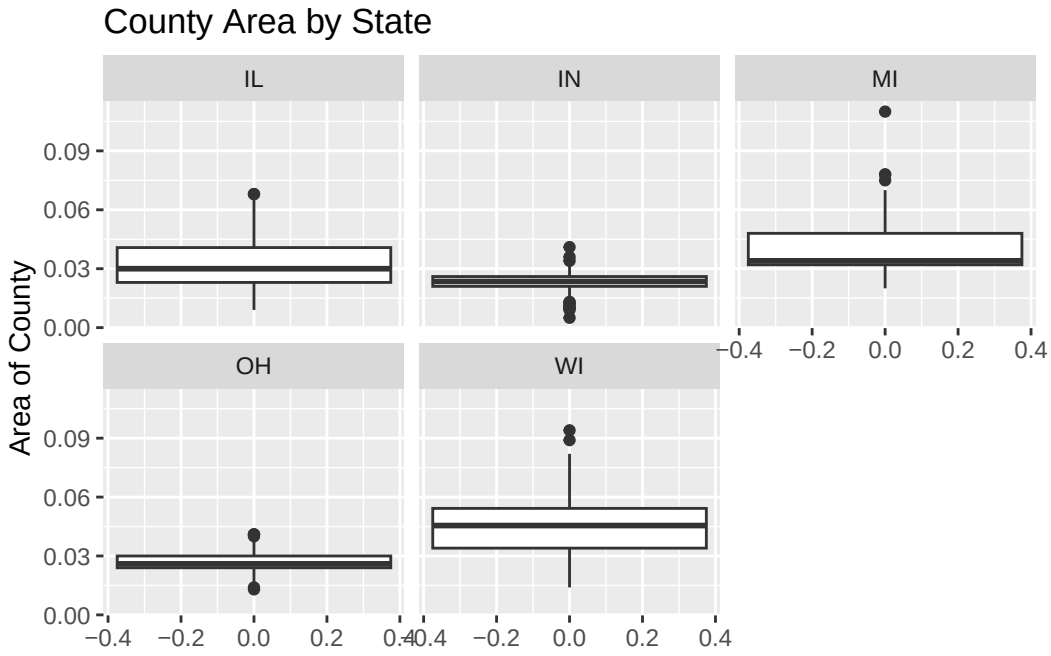
```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



This new plot definitely makes analyzing the data by state much easier when compared to the heavily congested scatterplot in question 3. The data is much more clear and concise in the facet wrapped scatter plot.

Question 5

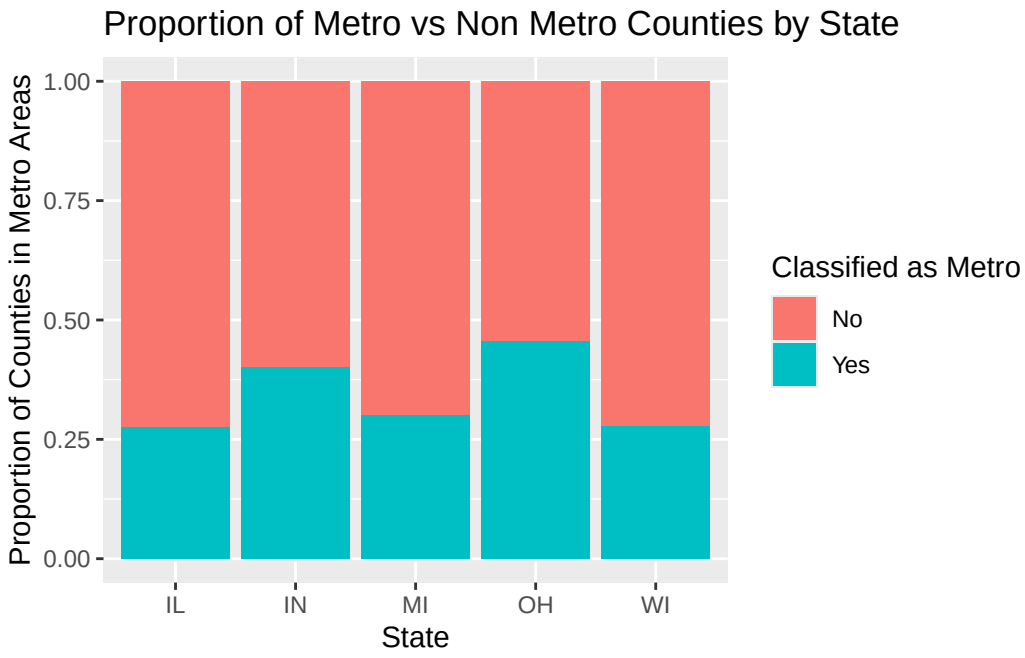
```
ggplot(midwest, aes(y = area)) +
  geom_boxplot() +
  facet_wrap(~ state) +
  labs(
      y = "Area of County",
    title = "County Area by State"
  )
```



Ohio and Indiana have noticeably smaller county sizes when compared to the other 3 states which sit in relatively similar ranges and have similar median county areas. Ohio and Indiana have lower median county areas as well as lower overall ranges for said areas. MARQUETTE county Michigan has the highest county area.

Question 6

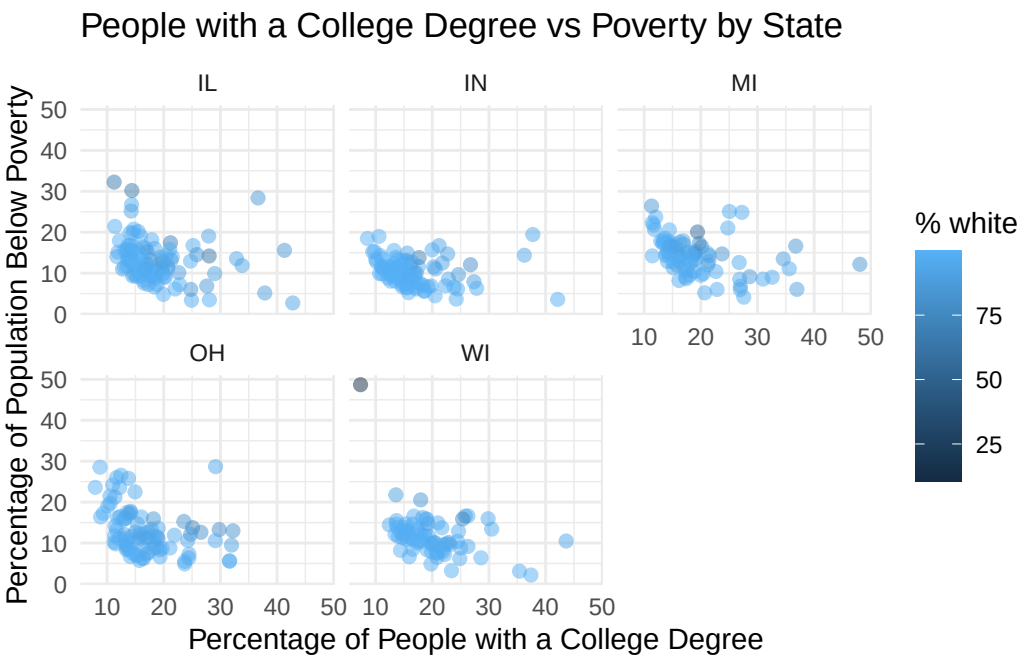
```
midwest <- midwest |>
  mutate(metro = if_else(inmetro == 1, "Yes", "No"))
ggplot(midwest, aes(x = state, , fill = metro)) +
  geom_bar(position = "fill") +
  labs(
    x = "State",
    y = "Proportion of Counties in Metro Areas",
    fill = "Classified as Metro",
    title = "Proportion of Metro vs Non Metro Counties by State"
  )
```



Indiana and Ohio both tend to have more counties that are considered metro in comparison to the other 3 states. Both Indiana and Ohio seem to have roughly 10-15% more counties that are classified as metro.

Question 7

```
ggplot(midwest, aes(y = percbelowpoverty, x = percollege, color = percwhite)) +
  geom_point(size = 2, alpha = 0.5) +
  facet_wrap(~ state) +
  labs(
    color = "% white",
    x = "Percentage of People with a College Degree",
    y = "Percentage of Population Below Poverty",
    title = "People with a College Degree vs Poverty by State"
  ) +
  theme_minimal()
```



One county that is a clear outlier is MENOMINEE. In this councty 89 percent of the population is Indian American and another 10 percent is white, while the remaining population is neither asian or black but instead a mix of other races.

Part 2

Enough about the Midwest!

Question 8

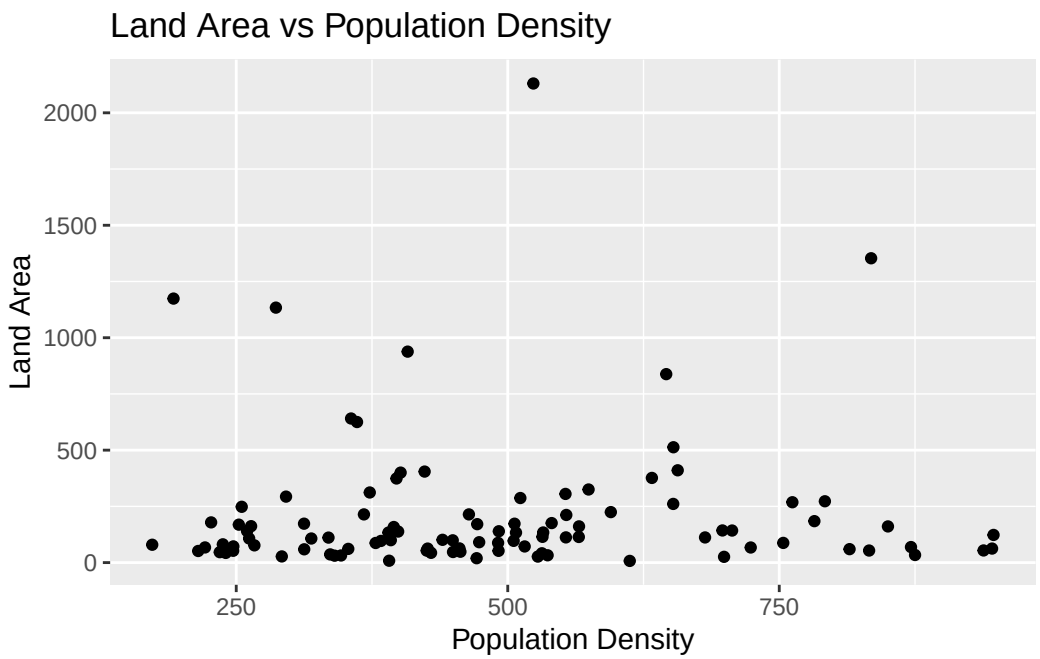
I would estimate that as land areas increase, population densities will decrease. Thus i would guess that they have a negative relationship.

```
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.
```

```r
ggplot(nc_county, aes(y = density, x = land_area_mi2)) +
  geom_point() +
  labs(
    x = "Population Density",
    y = "Land Area",
    title = "Land Area vs Population Density"
  )
```



The scatterplot shows that there is no clear relationship between the two variables, thus my guess was incorrect.

Question 9