Lab 1 - Data visualization

Fardeen Chowdhuruy

Questions

Part 1

Data Setup

```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 becor
```

```r
install.packages("ggthemes")
```

```
Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.5'
(as 'lib' is unspecified)
```

```r
library(ggthemes)
library(scales)
```

```
Attaching package: 'scales'

The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor
```

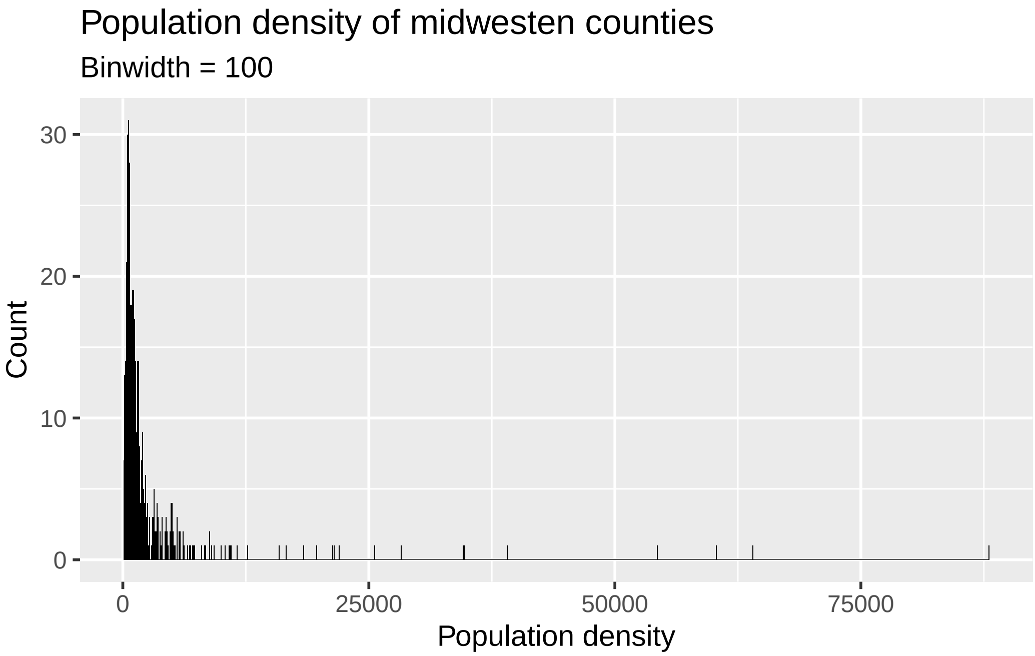
```
midwest
```

```
# A tibble: 437 x 28
     PID county  state  area poptotal popdensity popwhite popblack popamerindian
   <int> <chr>   <chr> <dbl>    <int>      <dbl>    <int>    <int>         <int>
 1   561 ADAMS   IL    0.052    66090      1271.    63917     1702            98
 2   562 ALEXAN~ IL    0.014    10626       759      7054     3496            19
 3   563 BOND    IL    0.022    14991       681.    14477      429            35
 4   564 BOONE   IL    0.017    30806      1812.    29344      127            46
 5   565 BROWN   IL    0.018     5836       324.     5264      547            14
 6   566 BUREAU  IL    0.05     35688       714.    35157       50            65
 7   567 CALHOUN IL    0.017     5322       313.     5298        1             8
 8   568 CARROLL IL    0.027    16805       622.    16519      111            30
 9   569 CASS    IL    0.024    13437       560.    13384       16             8
10   570 CHAMPA~ IL    0.058   173025      2983.   146506    16559           331
# i 427 more rows
# i 19 more variables: popasian <int>, popother <int>, percwhite <dbl>,
#   percblack <dbl>, percamerindan <dbl>, percasian <dbl>, percother <dbl>,
#   popadults <int>, perchsd <dbl>, percollege <dbl>, percprof <dbl>,
#   poppovertyknown <int>, percpovertyknown <dbl>, percbelowpoverty <dbl>,
#   percchildbelowpovert <dbl>, percadultpoverty <dbl>,
#   percelderlypoverty <dbl>, inmetro <int>, category <chr>
```

Question 1

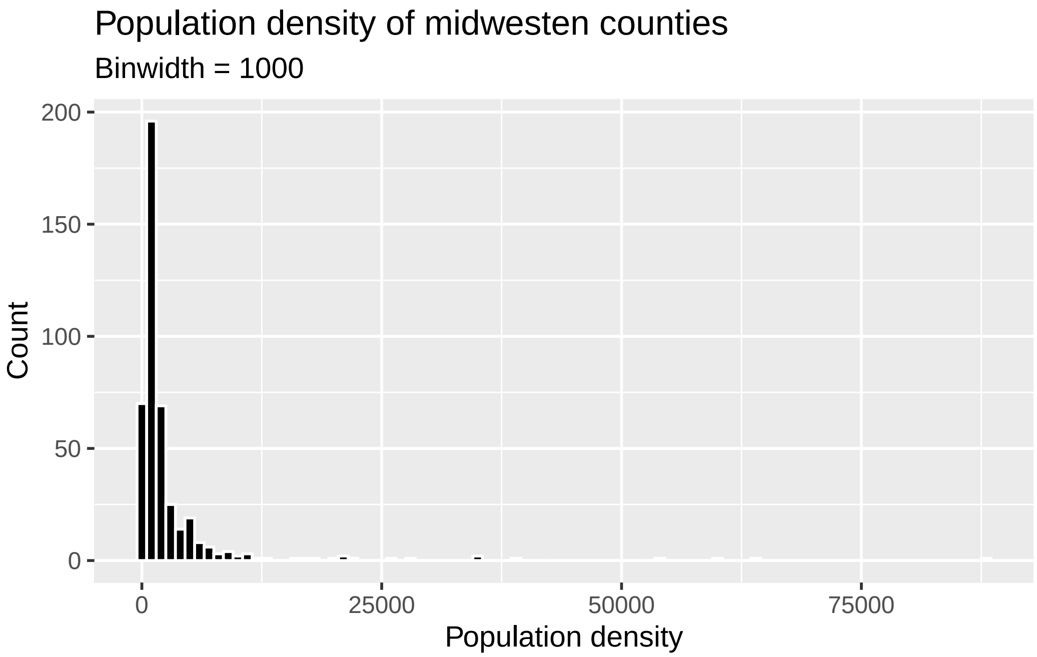
Binwidth = 100

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

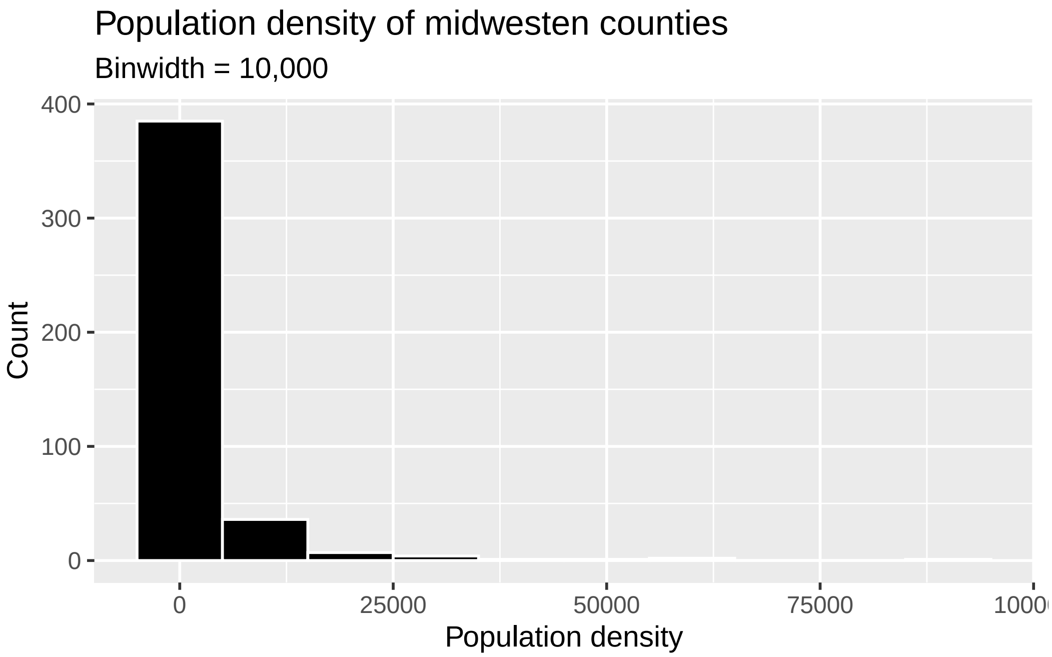


Binwidth = 1000

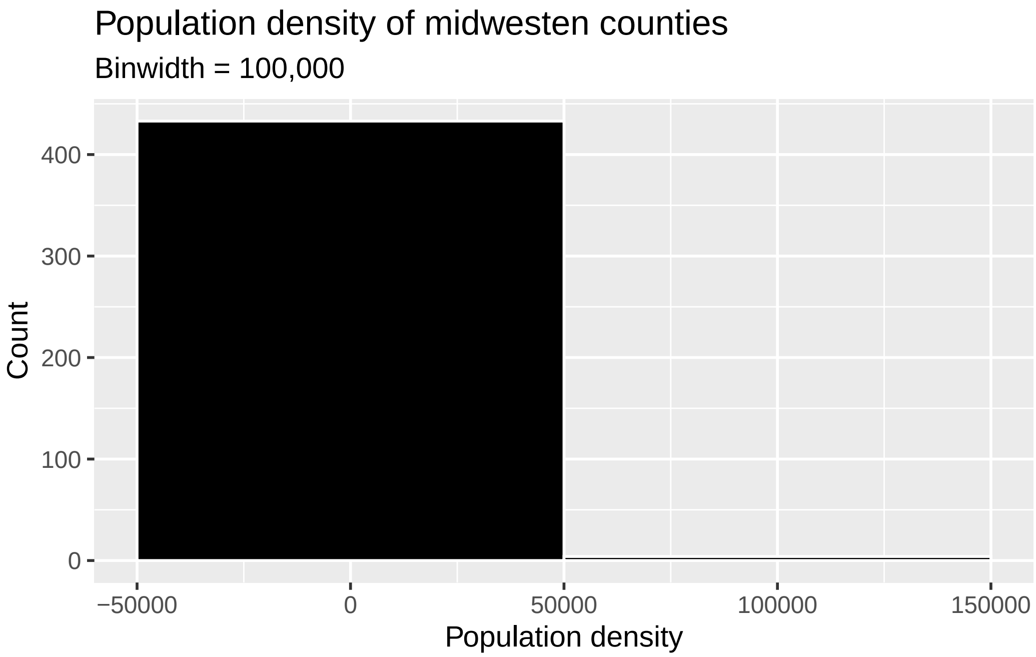
```
ggplot(midwest, aes(x = popdensity)) +
  geom_histogram(binwidth = 1000, fill= "black", colour= "white") +
  labs(
    x = "Population density",
    y = "Count",
    title = "Population density of midwesten counties",
    subtitle = "Binwidth = 1000"
  )
```



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



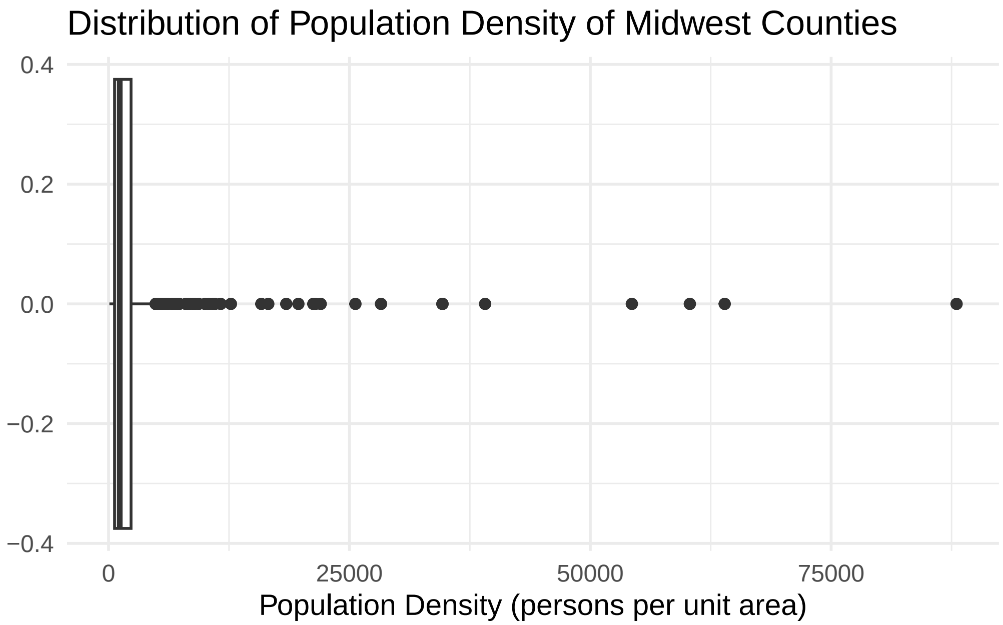
```r
ggplot(midwest, aes(x = popdensity)) +
  geom_histogram(binwidth = 100000,fill= "black", colour= "white") +
  labs(
    x = "Population density",
    y = "Count",
    title = "Population density of midwesten counties",
    subtitle = "Binwidth = 100,000"
  )
```



I would say Binwidth = 1000 is a great option as It included most of the data in a one side and in a ogranised pattern as if you want to check certain number of pupulation densitity and how many are there in midwest 1000 would be easier than 10,000 or 100 whereas 100000 does not show any represenation of data based on the quanity variable. It shows the shape of the diagram and the skewness of the data. it aslo shows poetential outliers so showing better describing the analysis.

Question 2

```
ggplot(midwest, aes(x = popdensity)) +
  geom_boxplot() +
  labs(
    title = "Distribution of Population Density of Midwest Counties",
    y = "",
    x = "Population Density (persons per unit area)
    "
  ) +
  theme_minimal() #got it from google for better look for scatterplot diagram
```



The data analysis from question 1 and 2 shows that the pop densisty of counties of midwest area is mostly around 0 to 15,000 with median below 6,000. However, it shows there are possible outlier of counties showing some counties with better reesources leading to higher population densisty and as population densisty depend on lot of facots among counties there are many outlies but the data is mostly pushed toward left. Hence, from the data I feel median does not provide much information with having so many outliers and mean would be benefical.

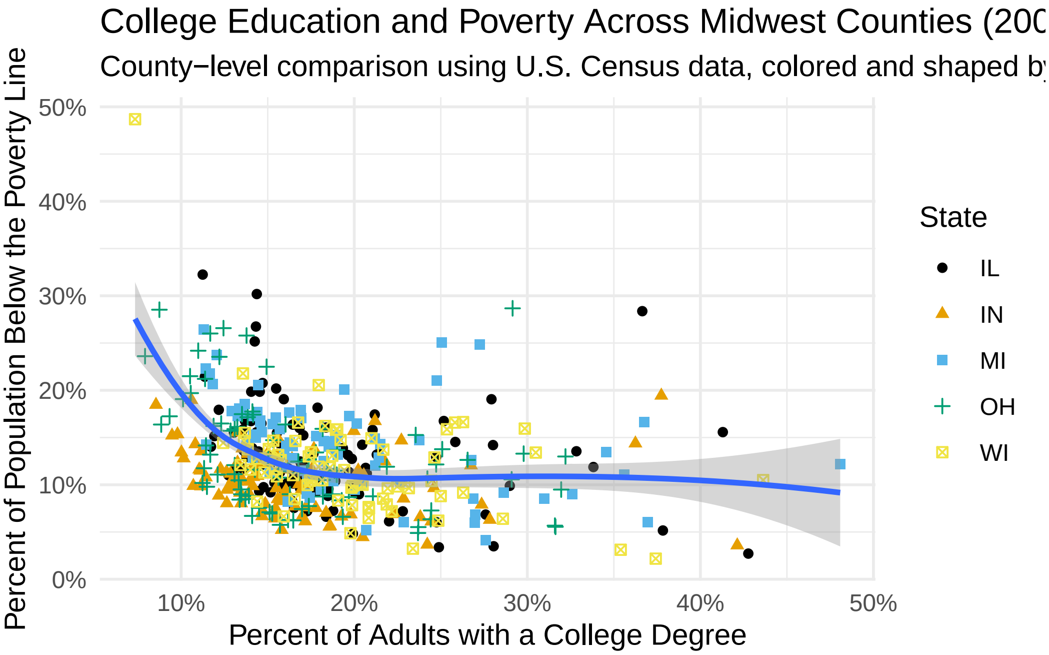
Overall, the population densisty of midwest counties is has low variability withmost data pushed on left with may otliers and showing the vaiarbiliry with the average spread.

Cook county is one of the outlier with highest pupulation density and far from the median.

Question 3

```r
ggplot(midwest, aes(x = percollege, y = percbelowpoverty))+
  geom_point(aes(color = state, shape= state ))+
  geom_smooth()+
  labs(
    title = 'College Education and Poverty Across Midwest Counties (2000)',
    subtitle = "County-level comparison using U.S. Census data, colored and shaped by state"
    x= "Percent of Adults with a College Degree",
    y= "Percent of Population Below the Poverty Line",
    color = "State",
    shape= "State"
  )+
  scale_x_continuous(labels = scales::label_percent(scale = 1)) +
  scale_y_continuous(labels = scales::label_percent(scale = 1))+  #scale x make its look bett

  scale_color_colorblind()+
  theme_minimal() #once again it makes it look better
```

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



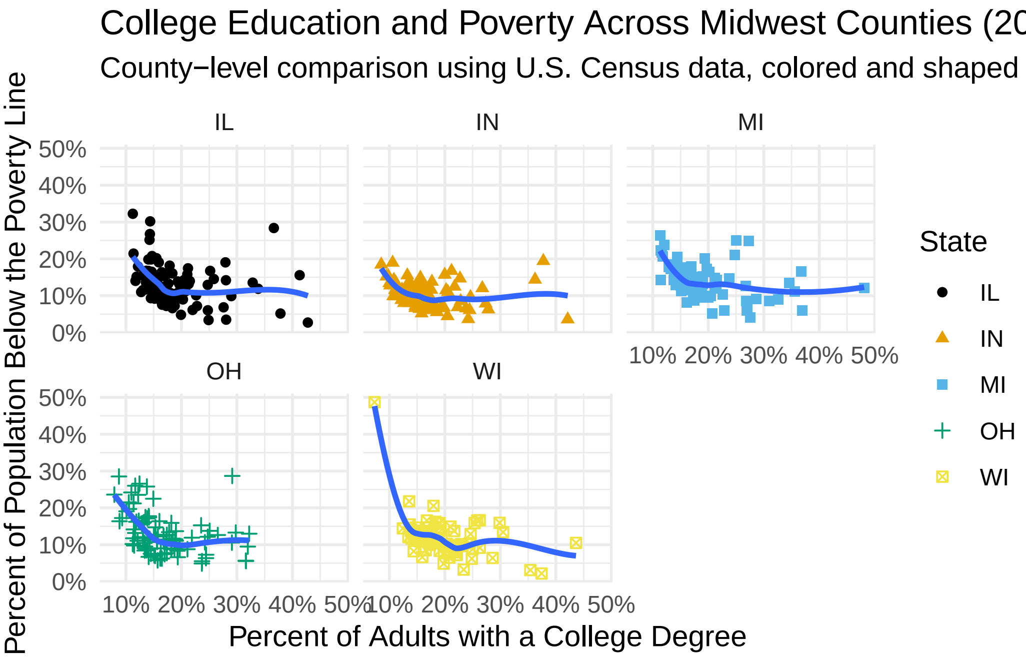
#According it to the scatter plot - Percent of Adulls with a college degree and population below poverty has somewhat a negative relationship in midwestern states. It indicates that with a rise in college degree among adult the population below poverty line falls. It shows a reationship not a causation. But looking into best with line, the co-relation is strong hence a full conclusion cannot be made.

#One of the outlier is JACKSON county. with high pupulation of college degree but still a comparative high poverty level.

Question 4

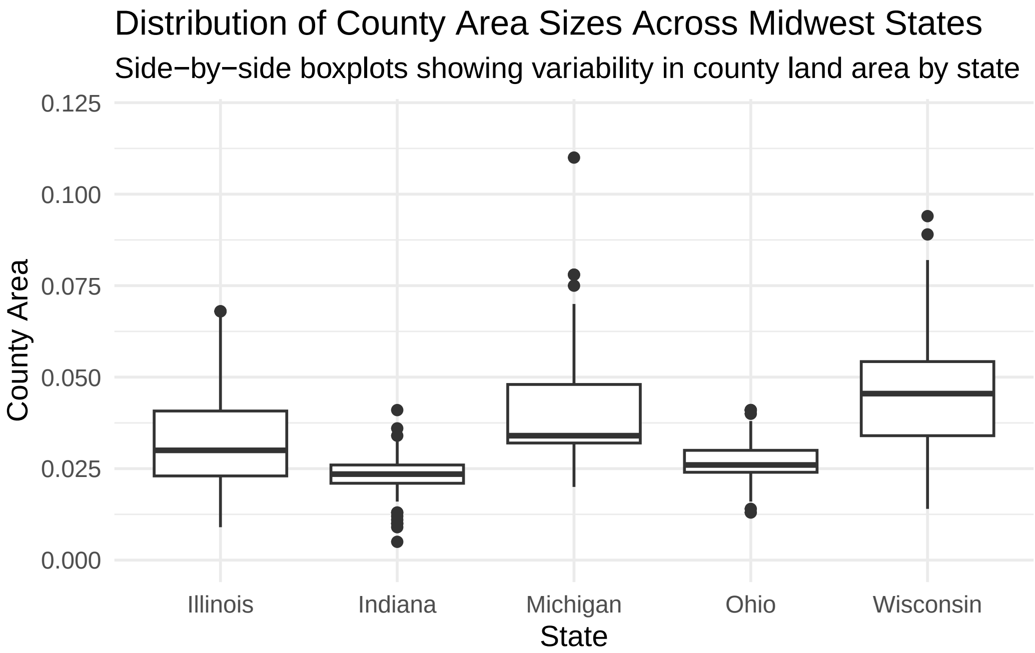
```
ggplot(midwest, aes(x = percollege, y = percbelowpoverty))+
  geom_point(aes(color = state, shape= state ))+
  geom_smooth( se = FALSE)+ #I did not use linear method because i feel it should be exponen
  labs(
    title = 'College Education and Poverty Across Midwest Counties (2000)',
    subtitle = "County-level comparison using U.S. Census data, colored and shaped by state"
    x= "Percent of Adults with a College Degree",
    y= "Percent of Population Below the Poverty Line",
    color = "State",
    shape= "State"
  )+
  scale_x_continuous(labels = scales::label_percent(scale = 1)) +
  scale_y_continuous(labels = scales::label_percent(scale = 1))+  #scale x make its look bet
  facet_wrap(~state)+

  scale_color_colorblind()+
  theme_minimal() #once again it makes it look better
```

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



What I prefer among question 3 and question depends on what i am trying to get from the data. IF i want to compare solely among state than I would choose question 4 but if i want to see the relationship in general i would prefer question 3. However, the graph in question 4 looks more organised to me. ### Question 5

```r
ggplot(midwest, aes(x= state, y= area))+
  geom_boxplot()+
  labs(
    title = "Distribution of County Area Sizes Across Midwest States",
    subtitle = "Side-by-side boxplots showing variability in county land area by state",
    x = "State",
    y= "County Area"
  )+
  scale_x_discrete(
    labels = c(
      IL = "Illinois",
      IN = "Indiana",
      MI = "Michigan",
      OH = "Ohio",
      WI = "Wisconsin"
    )
  )+

  #i wanted to put full name of states in this plot to make the graph look easier and a smal

  scale_y_continuous(
    limit = c(0, 0.12),

  )+
  #I use this fucntion taking help from chat gpt to show the lable for the higes areaof coun
  theme_minimal()
```



Michigan and wisconsisn has a higer variability in county sizes with higher interquartlie range and bigger whisker with several high area outliers in indicaitng a spread between high and low county area/

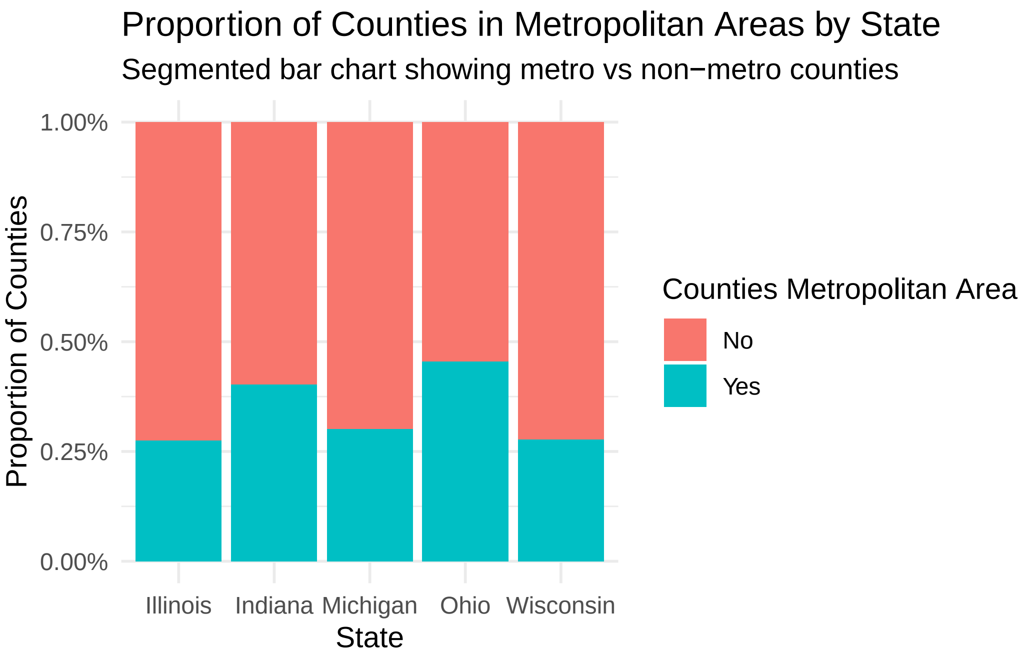
For Indiana and Ohia, it shows the size of country more consistent with more compact data and small interquartile range

For Illinoise I feel it has moderate variability with bigger interquartile range then Indiana and Ohi. but smaller than michigan and wisconsin. Also compared to michigan and wisconsisn their outlier is withinin the largest county showning less extreme variability.

Michigan has the single largest county. Name of the country is MARQUETTE.

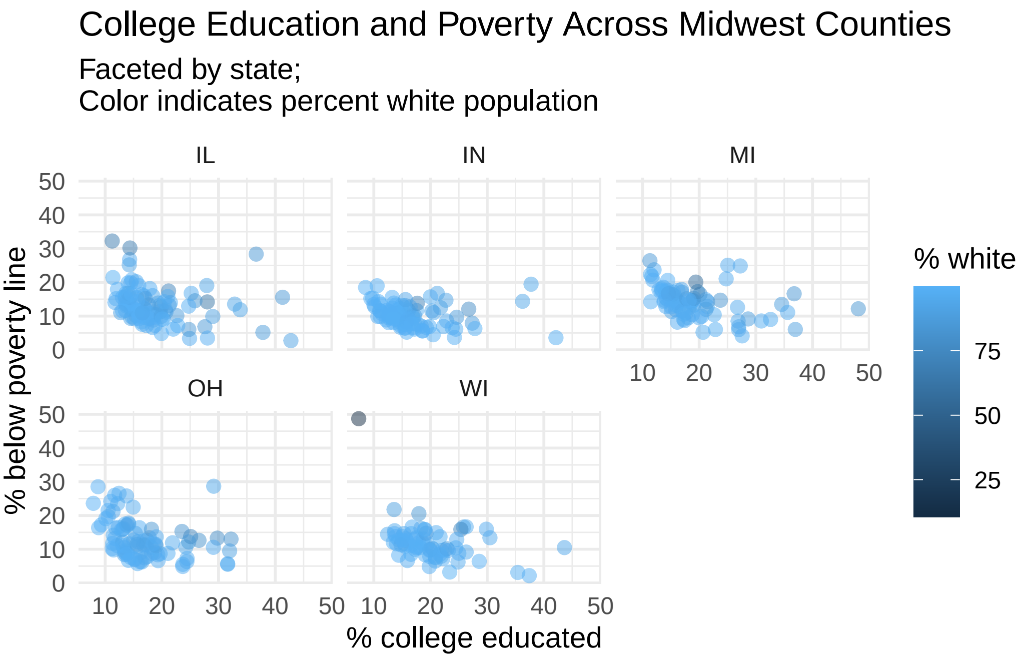
Question 6

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

ggplot(midwest, aes(x = state,fill = metro))+
  geom_bar(position = "fill")+
  scale_y_continuous(labels = scales::label_percent(scale=1))+
  scale_x_discrete(
    labels = c(
      IL = "Illinois",
      IN = "Indiana",
      MI = "Michigan",
      OH = "Ohio",
      WI = "Wisconsin"
    )
  )+
  labs(
    title = "Proportion of Counties in Metropolitan Areas by State",
    subtitle = "Segmented bar chart showing metro vs non-metro counties",
    x = "State",
    y = "Proportion of Counties",
    fill = "Counties Metropolitan Area"
  )+

  theme_minimal()
```



Illinois has slightly more than 25% in metropolital area which is similar with Michigan and wisconsin. However, Indiana has slight her than 37.5% counties in metropolitan area. Ohio has the higes counties in metorpolitan area according to the data with more than 40% but below 50%. ### Question 7

```
ggplot(midwest, aes(x = percollege, y = percbelowpoverty, colour = percwhite))+
  geom_point(size= 2, alpha = 0.5)+
  facet_wrap(~state)+
   labs(
    title = "College Education and Poverty Across Midwest Counties",
    subtitle = "Faceted by state;\nColor indicates percent white population",
    x = "% college educated",
    y = "% below poverty line",
    color = "% white"
  )+
  theme_minimal()
```



MENOMINEE county in Wisconsin state is a outlier with way higher bewlo poverty rate of 48.6911% compared to other counties. In this county 10.7% of thew population are white, but the most of the people in the county are from American Indians race with 89%. Whereas the country does not have any black population but there are 5% of other races. This chose the county majority population is american indians whereas whites and other races are minority.

Part 2

Enough about the Midwest!

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

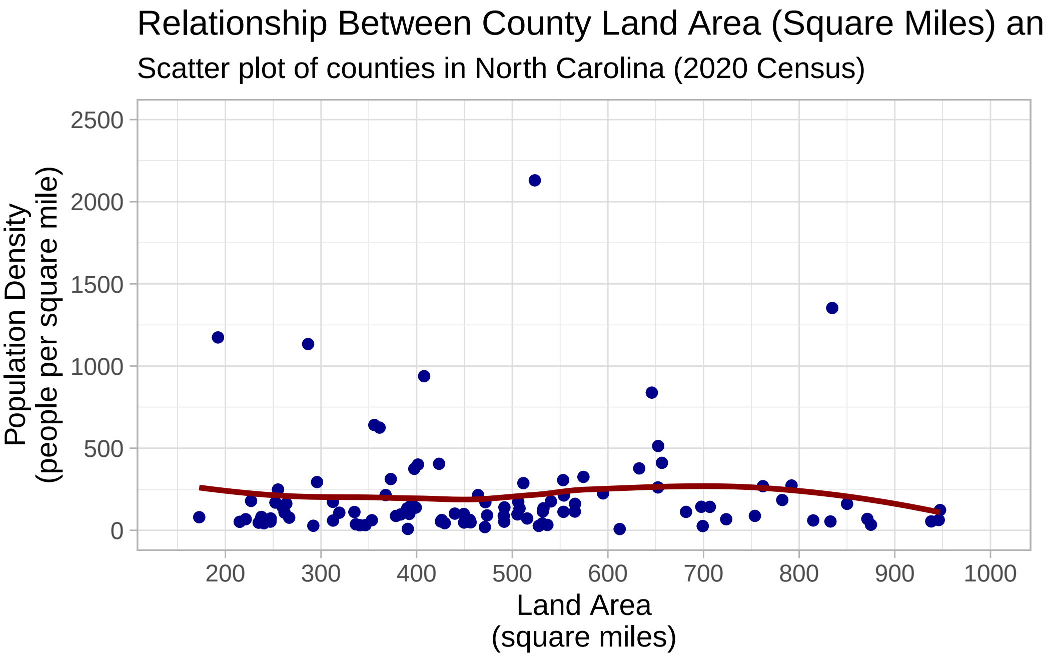
```
Rows: 100 Columns: 7
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): county, state_abb, state_name
dbl (4): land_area_m2, land_area_mi2, population, density

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

Before doing the coding i have guessed that popution density and land area has a negative relationship as population densitty should increase when land area decreases.

```r
ggplot(nc_county, aes(x = land_area_mi2, y= density))+
  geom_point(colour= "darkblue")+
  geom_smooth(se = FALSE, color = "darkred")+
  labs(
    title = "Relationship Between County Land Area (Square Miles) and Population Density",
    subtitle = "Scatter plot of counties in North Carolina (2020 Census)",
    x = "Land Area\n(square miles)",
    y = "Population Density\n(people per square mile)"
  )+
  scale_x_continuous(limits = c(150,1000),
  breaks = seq(100, 1000, by = 100))+ # i wanted my grid to be by 100 till 950 so that it is

  scale_y_continuous(limits = c(0,2500))+

  theme_light() #i like the plot design using theme_light instead of theme_minimal
```

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

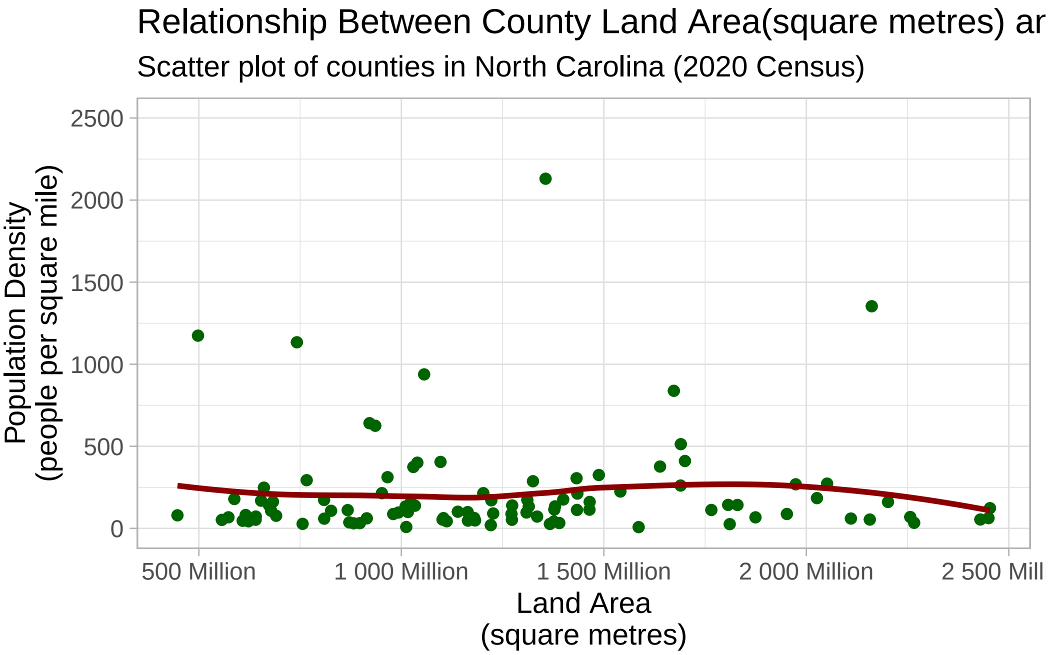


After analysing the data and making a scatter plot, I would so if you just see the best fit line, my guess of a negative relationship was not correct. This is becase even though land area is increasing, most of the county's population density is not in decreasing strongly. insteade the population density remains at the similar level despite county with higher land area. Hence, it can be said that population of a country does not strongly depend on Land Area.

Question 9

```
ggplot(nc_county, aes(x = land_area_m2, y= density))+
  geom_point(colour= "darkgreen")+
  geom_smooth(se = FALSE, color = "darkred")+
  labs(
    title = "Relationship Between County Land Area(square metres) and Population Density",
    subtitle = "Scatter plot of counties in North Carolina (2020 Census)",
    x = "Land Area\n(square metres)",
    y = "Population Density\n(people per square mile)"
  )+
  scale_x_continuous(labels = label_number(scale = 1e-6, suffix = " Million")) +
   scale_y_continuous(limits = c(0,2500))+

  theme_light() #i like the plot design using theme_light instead of theme_minimal
```

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



Relationship Status is same meaning popuulation density (people per square mile)- is has no strong relation with Land Area of square metres. I feel the reason it is showing similar is because in the population density because the density does not only depend on the land area. The number of people living the county(population) matter which did not change in this analysis.