Lab 1 - Data visualization

Jessica St. Jean

Questions

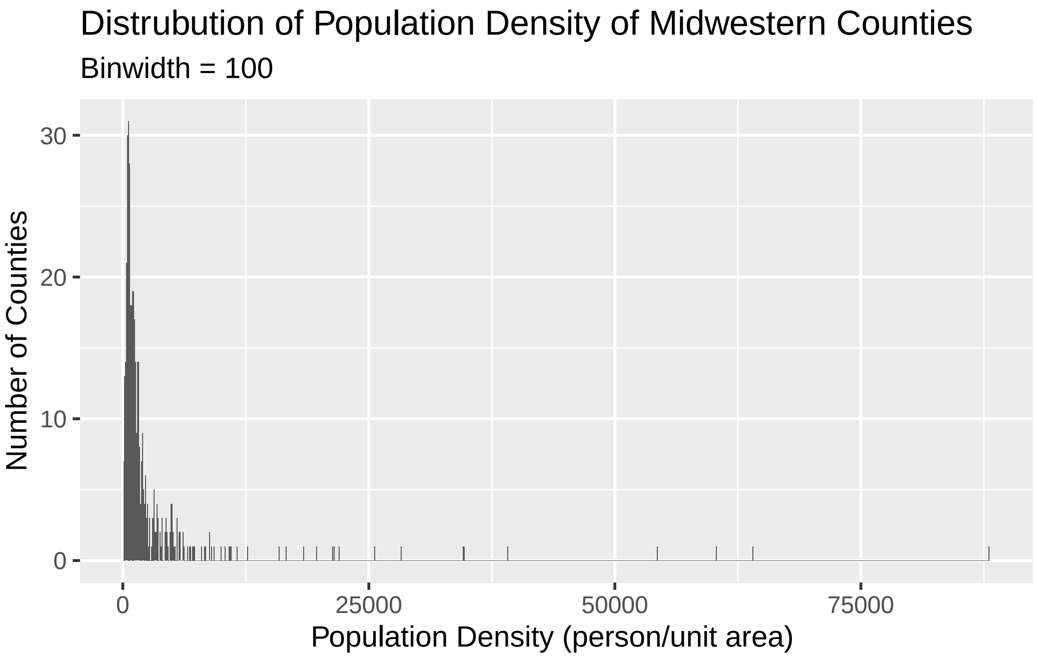
Part 1

```
library(tidyverse)
```

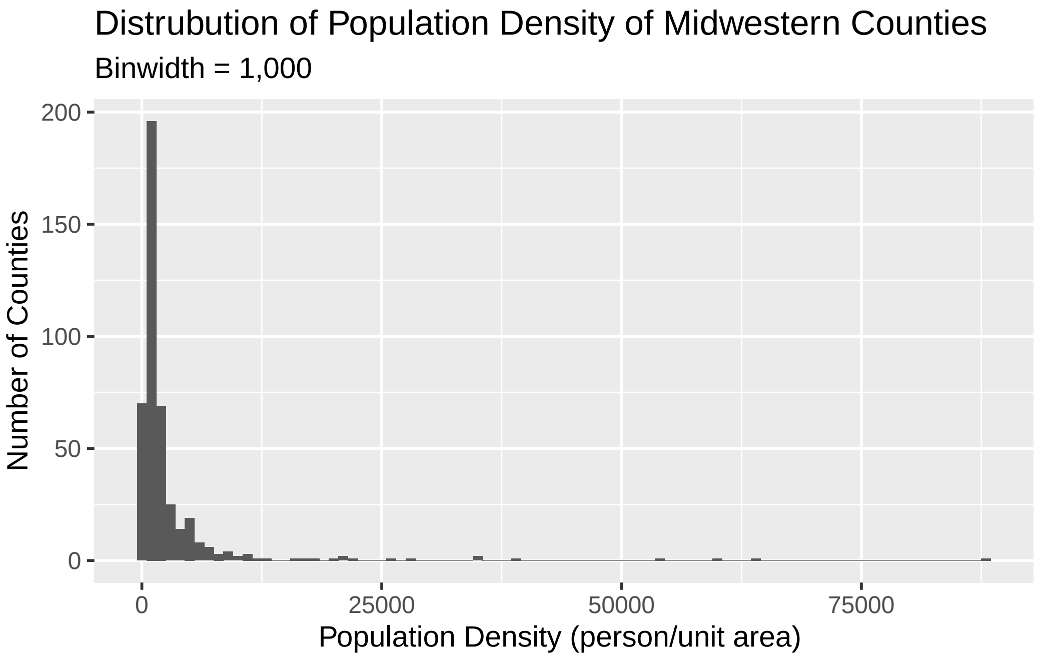
```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.6
v forcats   1.0.1     v stringr   1.6.0
v ggplot2   4.0.1     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.2
v purrr     1.2.0
-- Conflicts ------------------------------------------ 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 becom
```

Question 1

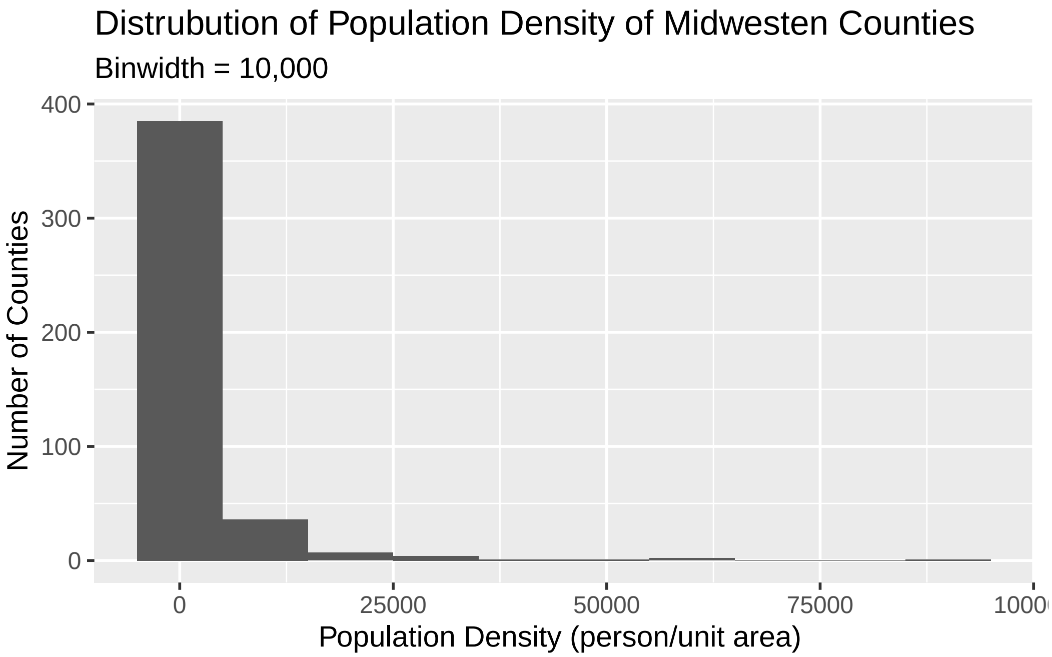
```
ggplot(midwest, aes(x=popdensity))+
  geom_histogram(binwidth = 100) +
  labs(
   x = "Population Density (person/unit area) ",
    y = "Number of Counties",
    title = "Distrubution of Population Density of Midwestern Counties",
    subtitle = "Binwidth = 100"
  )
```



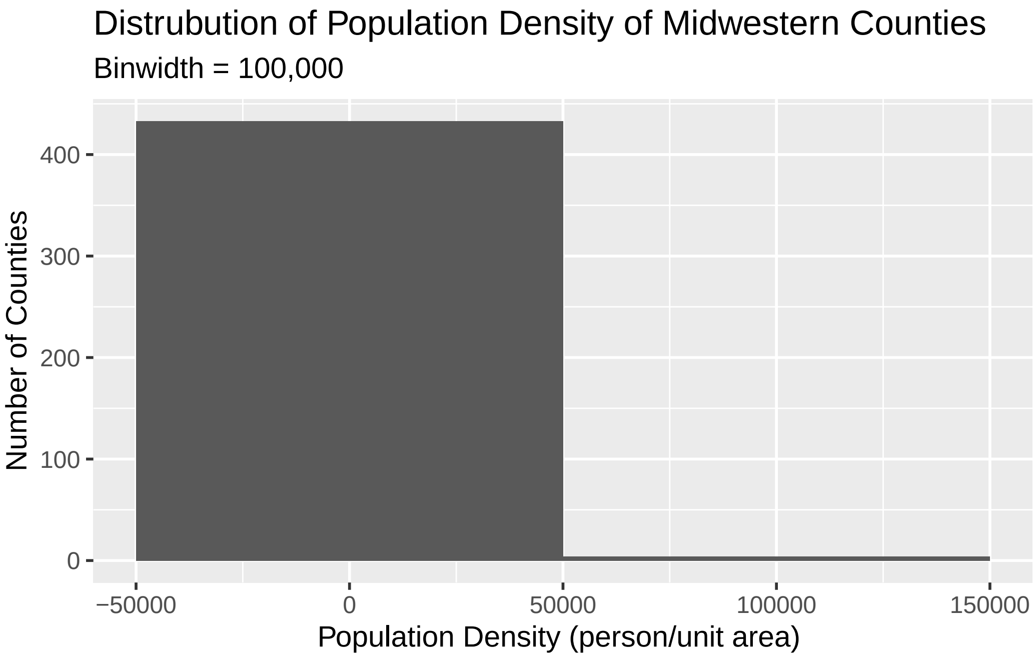
```
ggplot(midwest, aes(x=popdensity))+
  geom_histogram(binwidth = 1000) +
  labs(
     x = "Population Density (person/unit area) ",
    y = "Number of Counties",
    title = "Distrubution of Population Density of Midwestern Counties",
    subtitle = "Binwidth = 1,000"
  )
```



```
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"
  )
```



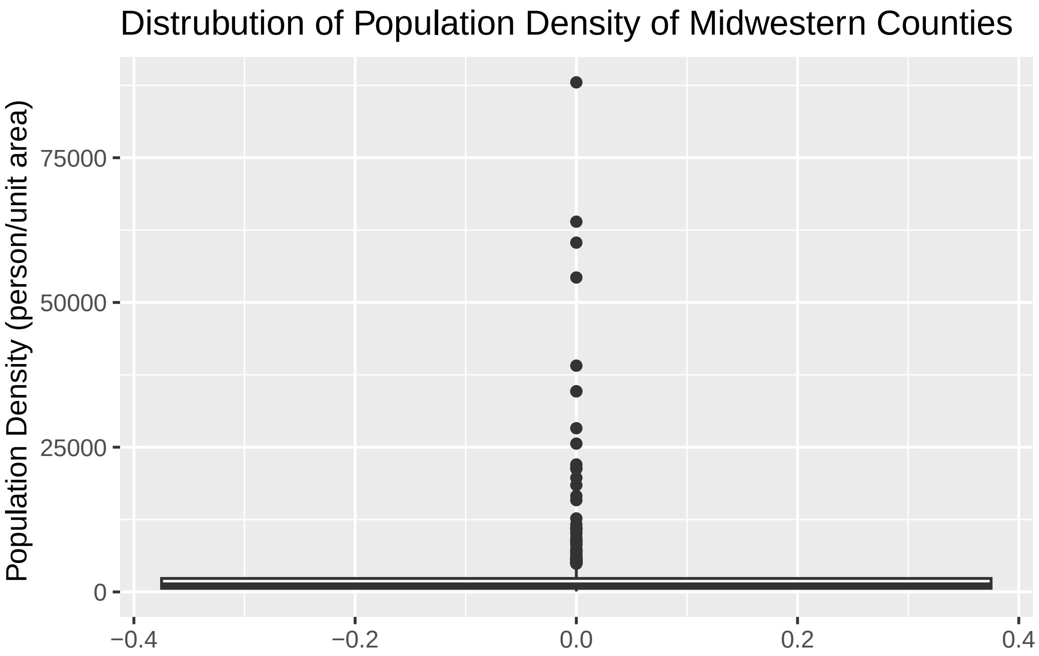
```
ggplot(midwest, aes(x=popdensity))+
  geom_histogram(binwidth = 100000) +
  labs(
    x = "Population Density (person/unit area) ",
    y = "Number of Counties",
    title = "Distrubution of Population Density of Midwestern Counties",
    subtitle = "Binwidth = 100,000"
  )
```



Comment: I think that the most appropirate 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 binwidth plot also seems 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 seperation between population densities is not extremely small and does not make there appear to be data below 0.

Question 2

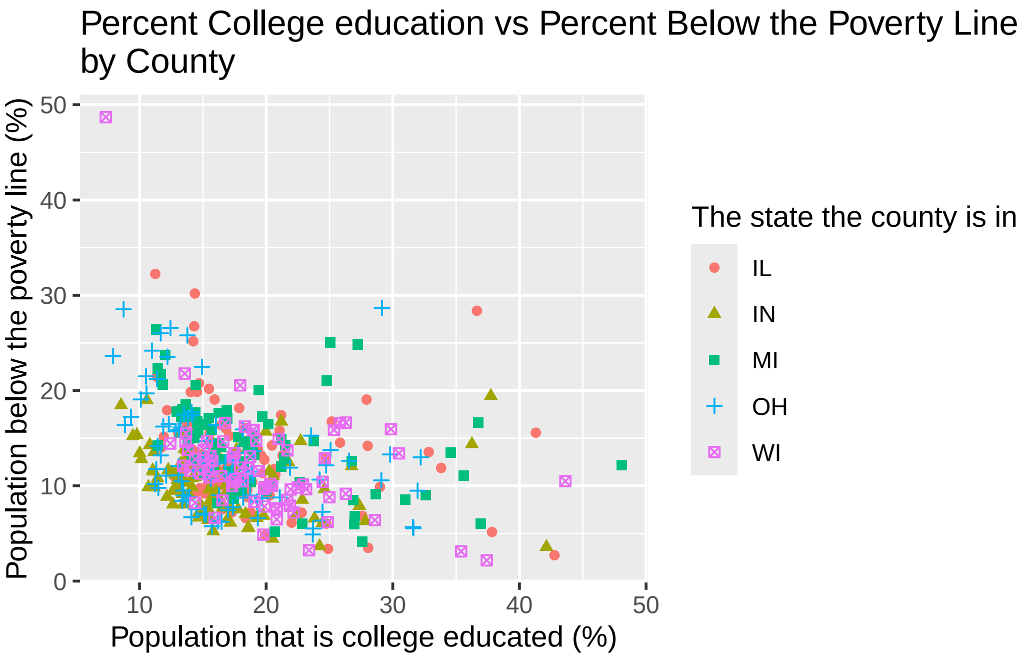
```
ggplot(midwest, aes(y = popdensity)) +
  geom_boxplot() +
  labs(
    y = "Population Density (person/unit area) ",
    title = "Distrubution of Population Density of Midwestern Counties",)
```



Comment: The distribution of the population density seems to be largely within the lower range. There are some outliers including the most extreme outlier, Cook County is Illinois. It is hard for me to determine if it makes sense to me that this 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 dense urban area while the majority of counties in the midwest would 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")
```

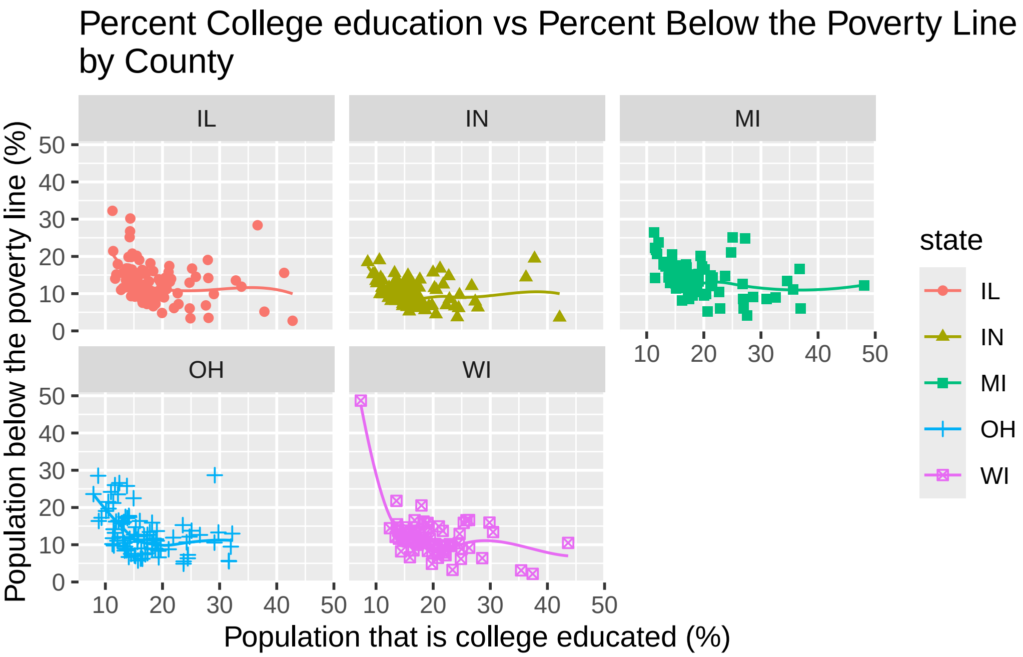


Comment: The general relationship between the percent of the population below the poverty line and the percent of the population that is college educated is that when the percent of college education goes up the percent of the population below the poverty rate goes down. An outlier for this general trend 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 and 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.

Question 5

Question 6

Question 7

Part 2

Enough about the Midwest!

Question 8

Question 9