Lab 1 - Data visualization

Jessica St. Jean

Questions

Part 1

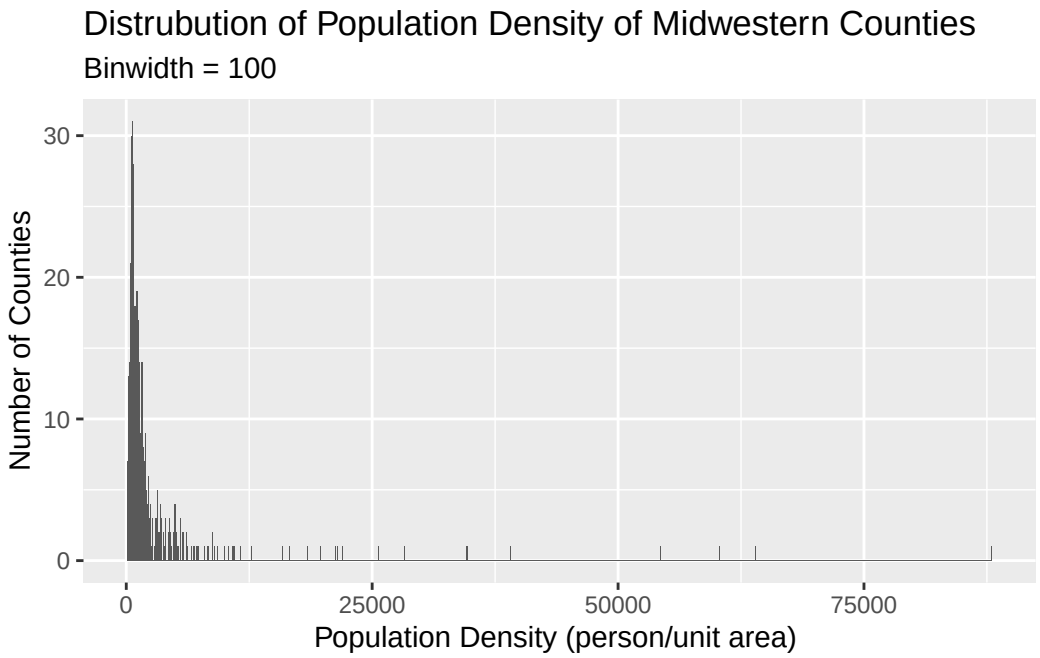
```
library(tidyverse)
```

```
-- 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 beco
```

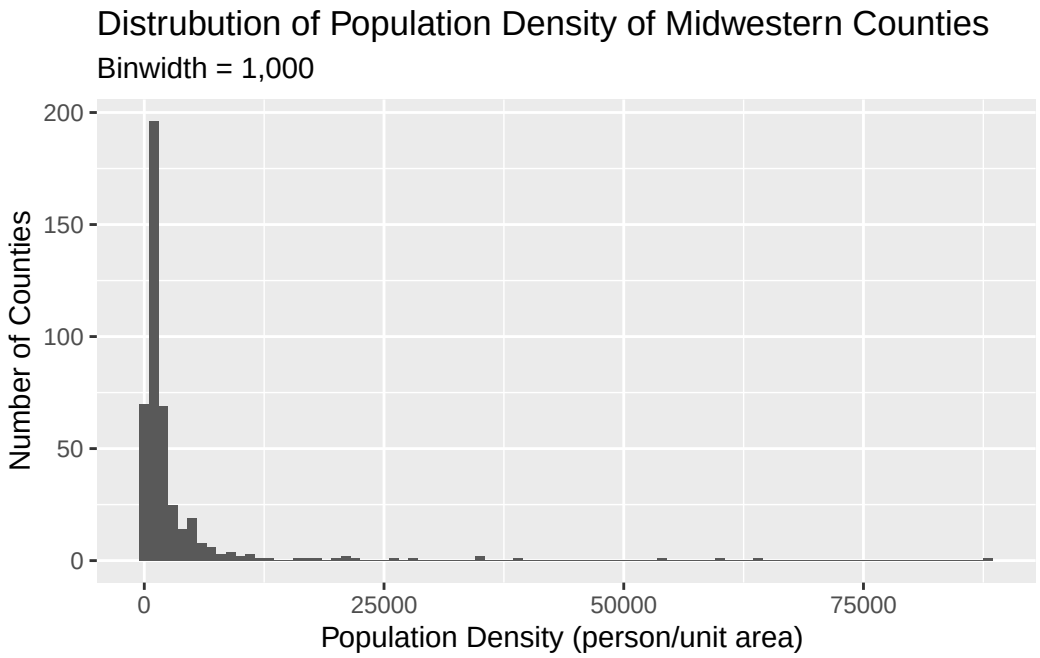
```
?midwest
```

Question 1

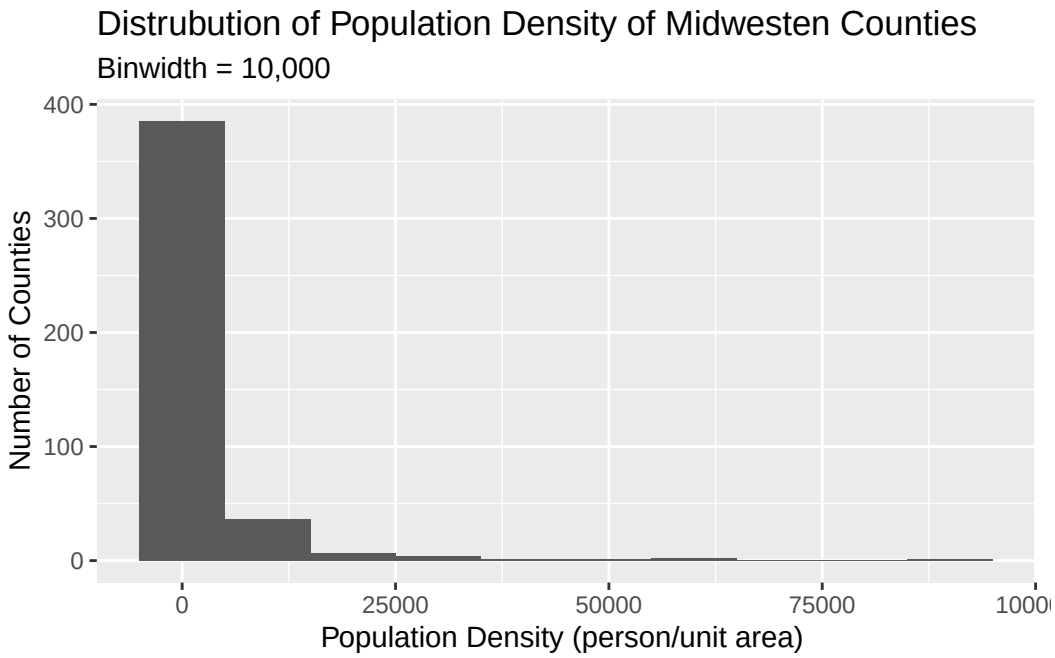
```
ggplot(midwest, aes(x=popdensity))+
  geom_histogram(binwidth = 100) +
  labs(
   x = "Population Density (person/unit area) ",
    y = "Number of Counties",
    title = "Distrubution of Population Density of Midwestern Counties",
    subtitle = "Binwidth = 100"
  )
```



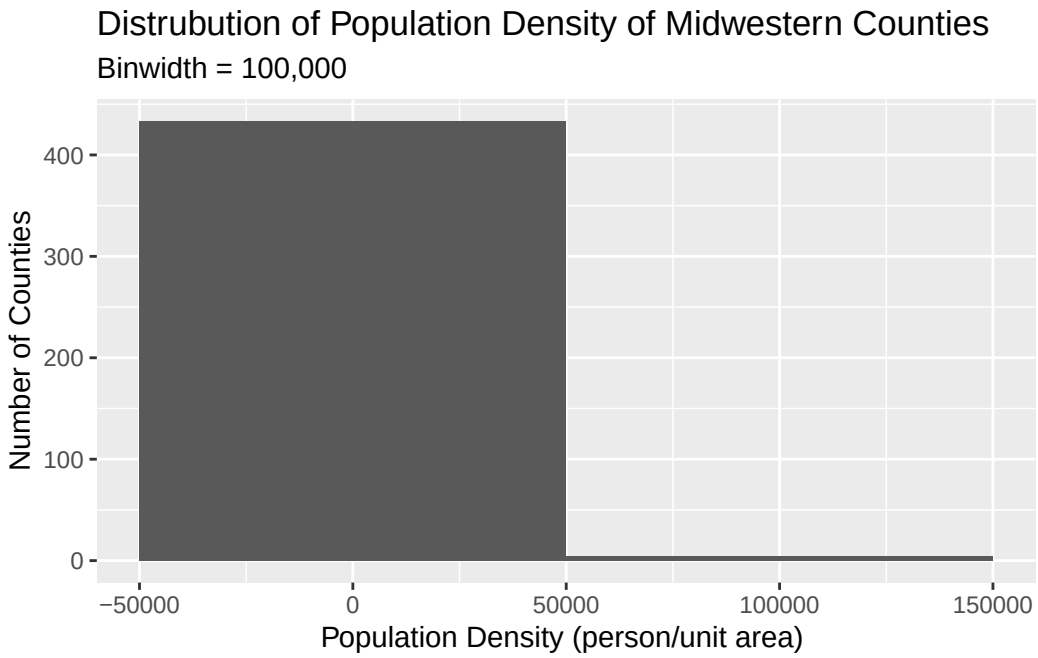

```
ggplot(midwest, aes(x=popdensity))+
  geom_histogram(binwidth = 1000) +
  labs(
     x = "Population Density (person/unit area) ",
    y = "Number of Counties",
    title = "Distrubution of Population Density of Midwestern Counties",
    subtitle = "Binwidth = 1,000"
  )
```




```
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"
  )
```



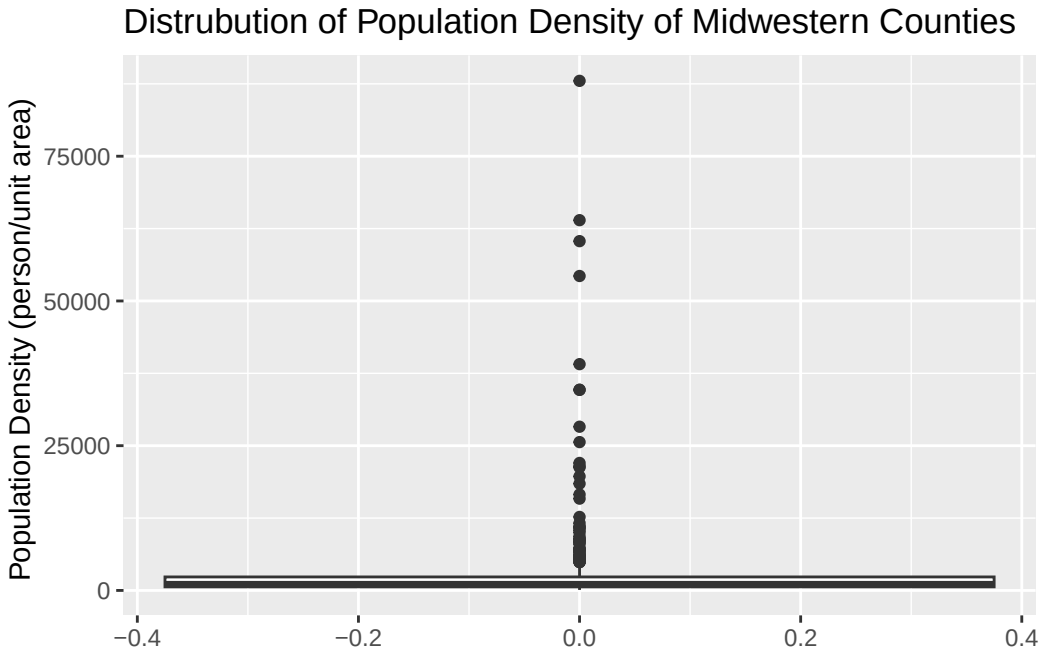

```
ggplot(midwest, aes(x=popdensity))+
  geom_histogram(binwidth = 100000) +
  labs(
    x = "Population Density (person/unit area) ",
    y = "Number of Counties",
    title = "Distrubution of Population Density of Midwestern Counties",
    subtitle = "Binwidth = 100,000"
  )
```



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, Mecklenburg County, NC. 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 of the United States

Question 3

Question 4

Question 5

Question 6

Question 7

Part 2

Enough about the Midwest!

Question 8

Question 9