Lab 1 - Data visualization

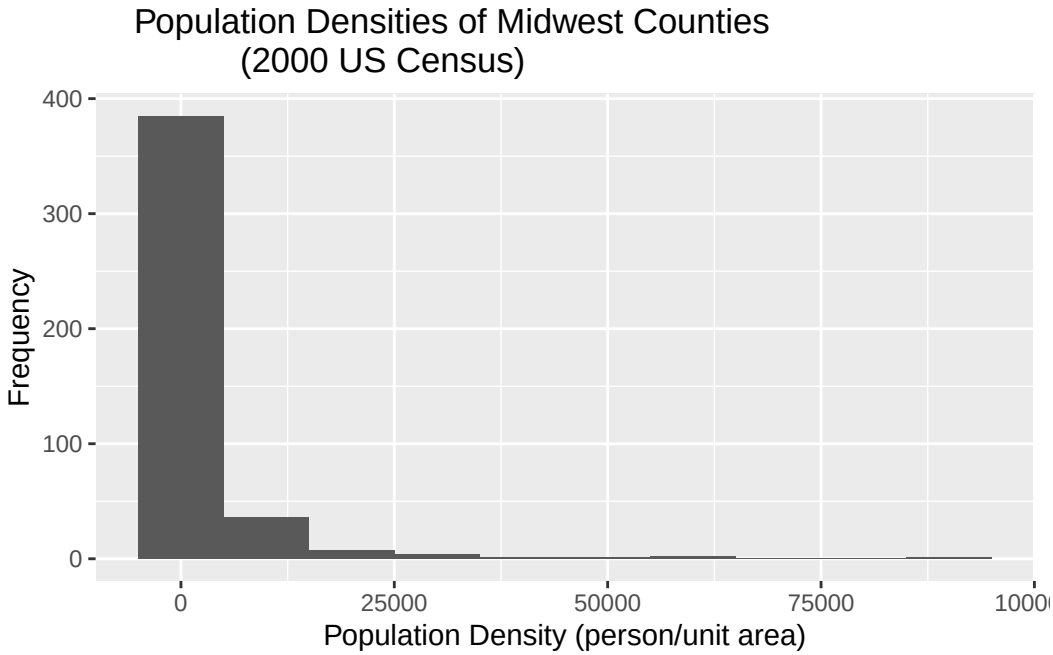
[Ann Chang]

Load Packages

```
library(tidyverse)
library(viridis)
```

Exercise 1

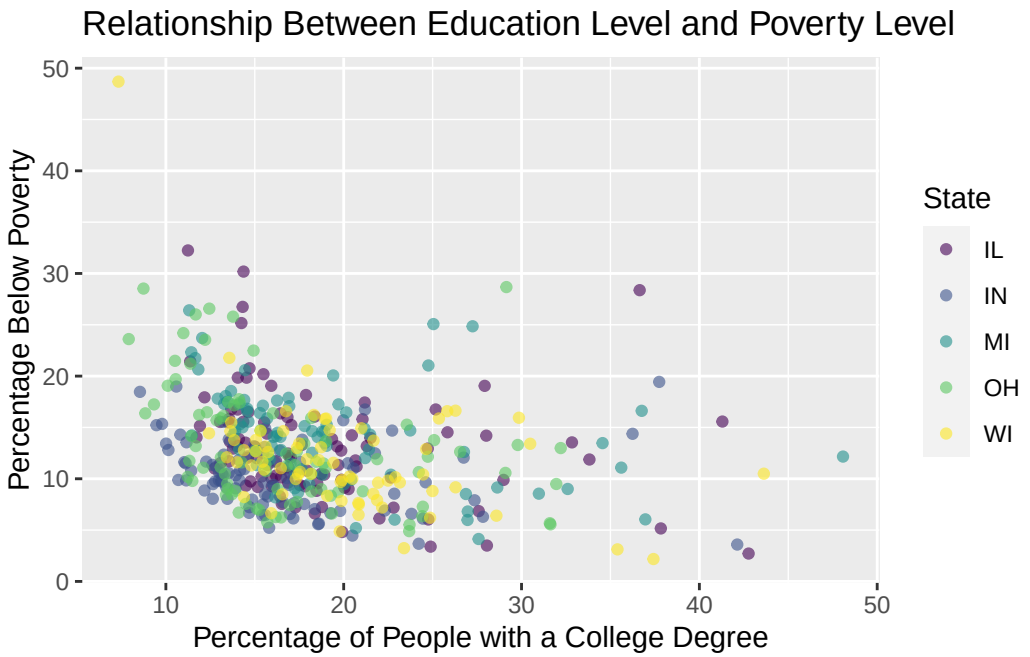
```
ggplot(data = midwest,
       aes(x = popdensity)) +
  geom_histogram(binwidth = 10000) +
  labs(
    x = "Population Density (person/unit area)",
    y = "Frequency",
    title =
      "    Population Densities of Midwest Counties
              (2000 US Census)"
    )
```



- The distribution of the data seem to be right-skewed.
- There seem to be some outliers of relatively large population densities, such as the incidences beyond 50000 people/unit area.

Exercise 2

```
ggplot(data = midwest,
       aes(x = percollege,
           y = percbelowpoverty,
           color = state)) +
  geom_point(alpha = 0.6) +
  labs(
    x = "Percentage of People with a College Degree",
    y = "Percentage Below Poverty",
    title = "Relationship Between Education Level and Poverty Level",
    color = "State"
  ) +
  scale_color_viridis_d()
```



Exercise 3

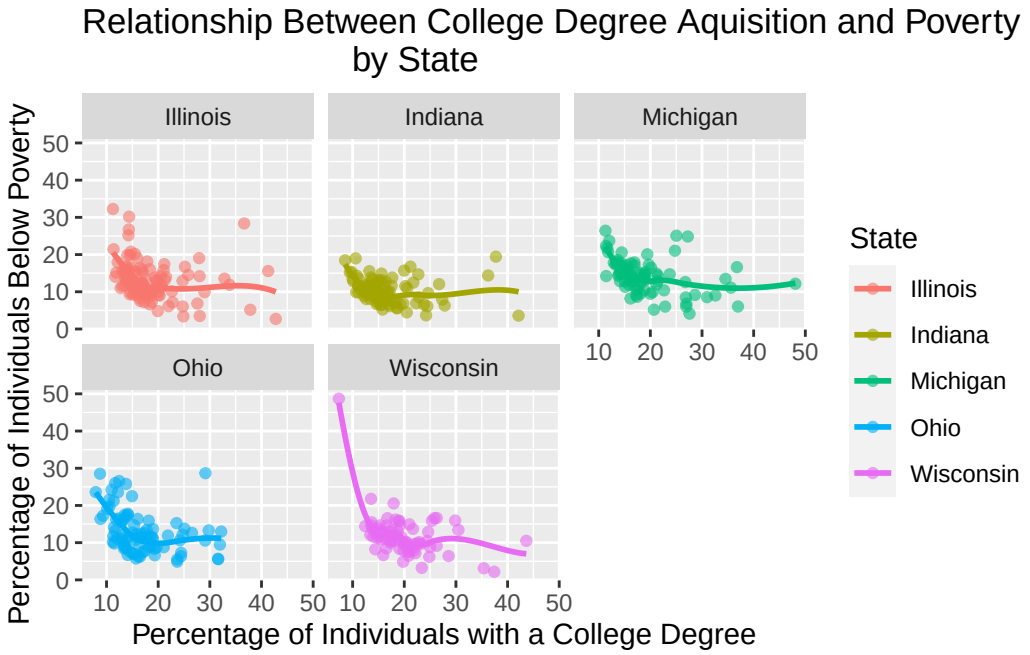
There seems to be a negative correlation between the percentage of people below the poverty line and the percentage of people with a college degree. Illinois seems to have relatively more outlying counties where the percentage of below poverty individuals is larger than what would be expected from the percentage of people with a college degree. This is also seen in some counties in Ohio. The counties in Wisconsin are relatively clustered closely together, with no counties' percentage of individuals below poverty exceeding above 23%.

Exercise 4

```
state.labs <- c("Illinois", "Indiana", "Michigan", "Ohio", "Wisconsin")
names(state.labs) <- c("IL", "IN", "MI", "OH", "WI")

ggplot(data = midwest,
       aes(x = percollege,
           y = percbelowpoverty,
           color = state)
       ) +
  geom_point(alpha = 0.6) +
```

```
facet_wrap(~state,
           labeller = labeller(state = state.labs)
           ) +
geom_smooth(se = FALSE) +
scale_color_viridis_d() +
labs(
  x = "Percentage of Individuals with a College Degree",
  y = "Percentage of Individuals Below Poverty ",
  title = "Relationship Between College Degree Aquisition and Poverty Level,
                          by State",
  color = "State"
  ) +
scale_color_discrete(labels = state.labs)
```

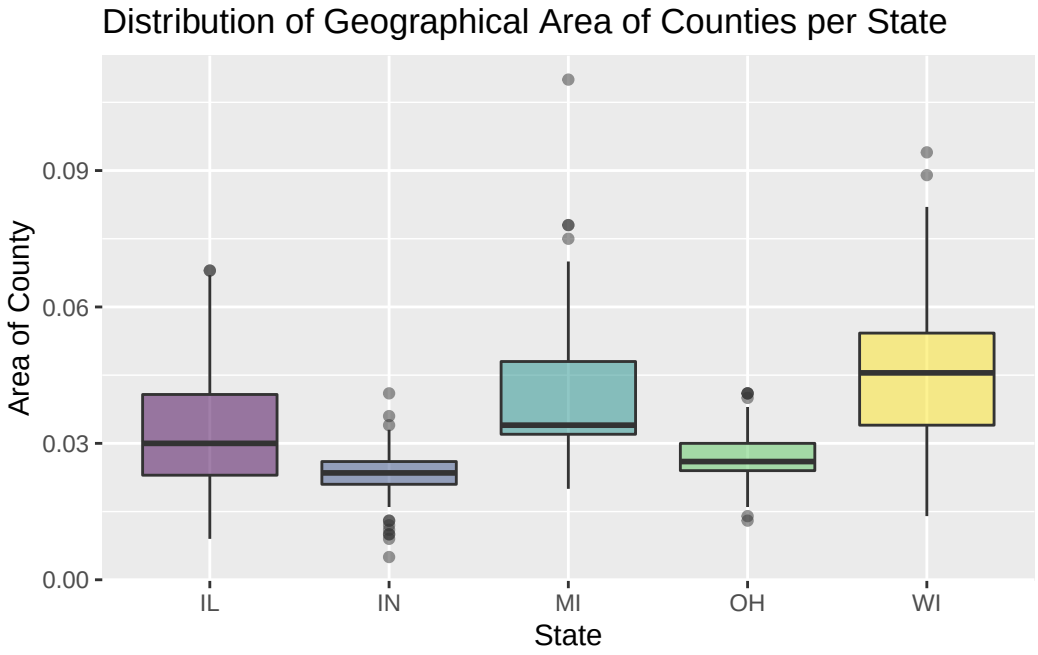


The faceted scatterplot is much more preferable than the original scatterplot from Exercise 2 because it is much easier to see the relationship between these two percentages for each individual state. This relationship for each state was much harder to see in the plot from Exercise 2 because the points were too crowded together to be able to separate the states, even with the differential coloring.

Exercise 5

```
state.labs <- c("Illinois", "Indiana", "Michigan", "Ohio", "Wisconsin")
names(state.labs) <- c("IL", "IN", "MI", "OH", "WI")

ggplot(data = midwest,
       aes(x = state,
           y = area,
           fill = state)
       ) +
  geom_boxplot(
    alpha = 0.5,
    show.legend = FALSE
    ) +
  scale_fill_discrete(
    labels = state.labs
    ) +
  labs(
    x = "State",
    y = "Area of County",
    title = "Distribution of Geographical Area of Counties per State",
    fill = "State",
    ) +
  scale_fill_viridis_d()
```



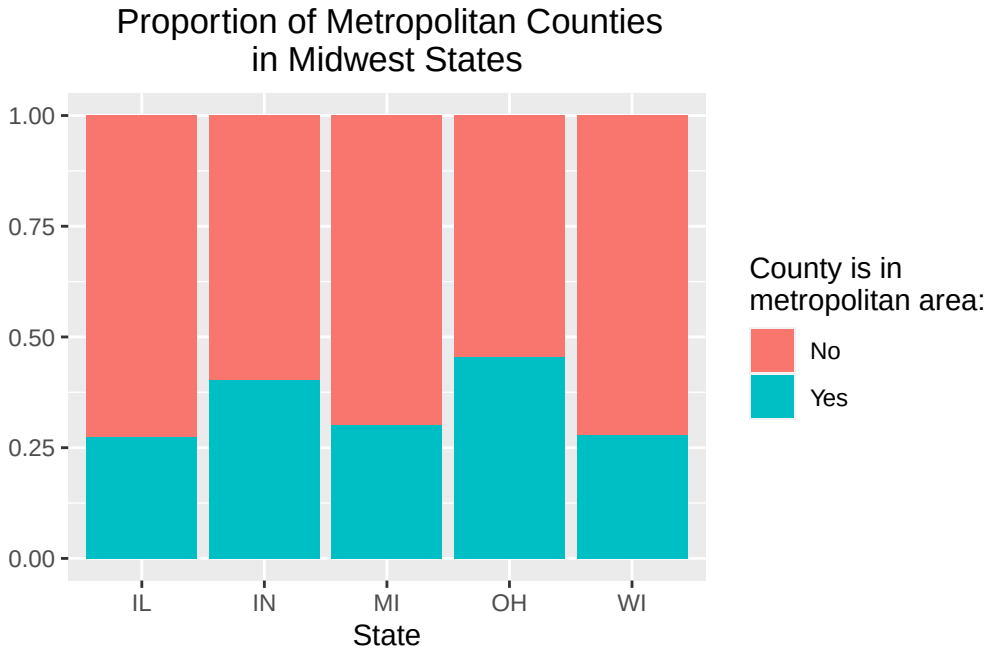
- The states of Wisconsin, Michigan, and Illinois have counties that are on the larger side relative to the average county size of Indiana and Ohio. The spread of the data for county geographical area is more widespread for the states of Illinois, Michigan, and Wisconsin in relation to Indiana and Ohio as well.
- MI has the county with the largest geographical area. This is identified by the outlier that is represented by a point that lies higher than the others in relation to the x-axis.

Exercise 6

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

ggplot(data = midwest) +
  geom_bar(
    mapping = aes(x = state, fill = metro),
    position = "fill"
    ) +
  labs(
    x = "State",
    y = " ",
    fill = "County is in \nmetropolitan area:",
```

```
    title = "     Proportion of Metropolitan Counties
                   in Midwest States"
    )
```



According to the plot, it can be seen that most counties in all the states presented are located outside of metropolitan areas. Of the states observed in this dataset, Ohio has the highest proportion of counties that are located in metropolitan areas.

Exercise 7

```
ggplot(data = midwest,
       aes(x = percollege,
           y = popdensity,
           color = percbelowpoverty
       )) +
  geom_point(alpha = 0.5, size = 2) +
  facet_wrap(~state) +
  theme_minimal() +
  labs(
    x = "% college educated",
```

```
  y = "Population density (person / unit area)",
  title = "Do people with college degrees tend to live in denser areas?",
  color = "% below poverty line"
)
```