Lab 1 - Data visualization

Jack Morgenstein

Load Packages

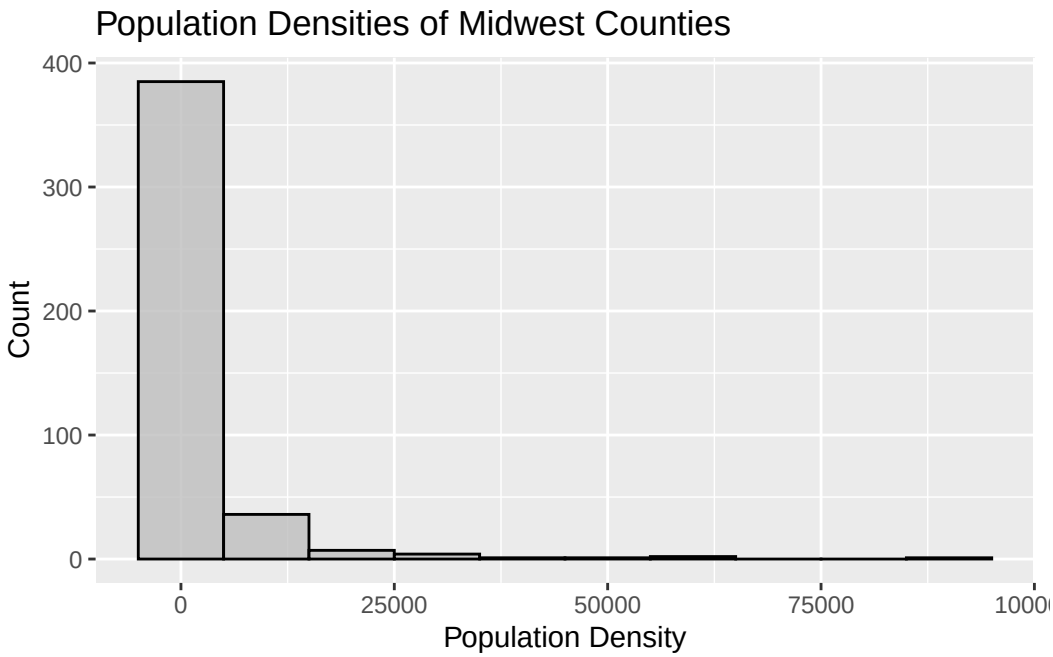
```
library(tidyverse)
```

```
Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
had status 1
```

```
library(viridis)
```

Exercise 1

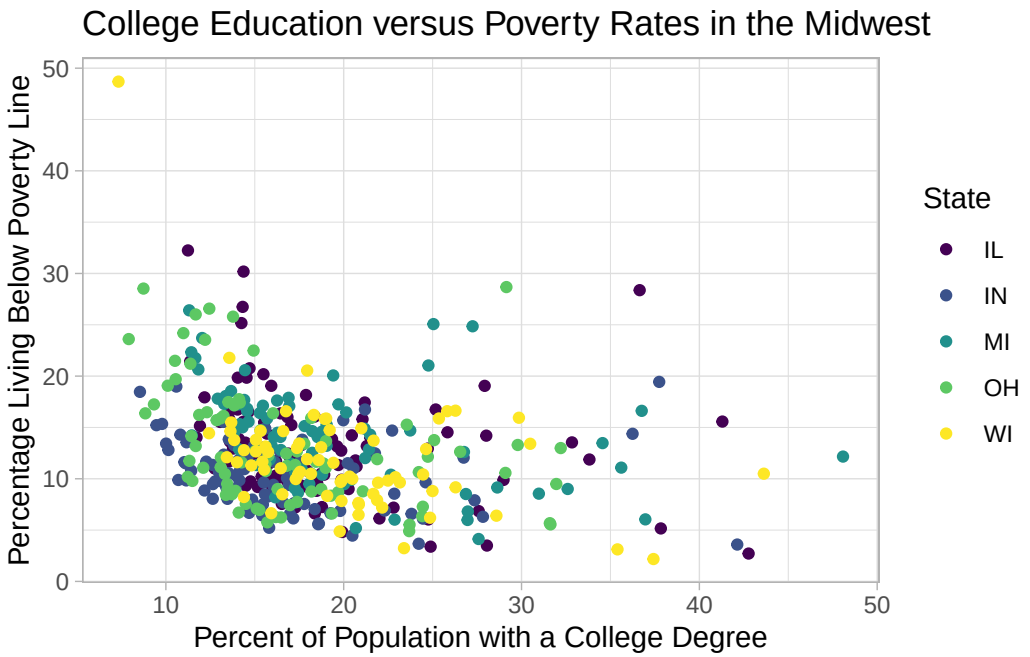
```
ggplot(midwest, mapping = aes(x = popdensity)) +
  geom_histogram(binwidth = 10000, alpha = 0.8, color = "black",
                 fill = "grey") +
  labs(x = "Population Density", y = "Count", title =
         "Population Densities of Midwest Counties")
```



The data distribution is strongly right skewed. A vast majority of counties have population densities between 0 and 10,000 while a much smaller number of counties have any higher population densities. There appear to be two outlier bins. A small number of counties have population densities between 55 and 65 thousand people as well as between 85 and 95 thousand people.

Exercise 2

```
ggplot(midwest, mapping = aes(x = percollege, y = percbelowpoverty,
                              color = state)) +
  geom_point() +
  labs(x = "Percent of Population with a College Degree",
       y = "Percentage Living Below Poverty Line",
       title = "College Education versus Poverty Rates in the Midwest",
       color = "State") +
  theme_light() +
  scale_color_viridis_d(option = "D")
```



Exercise 3

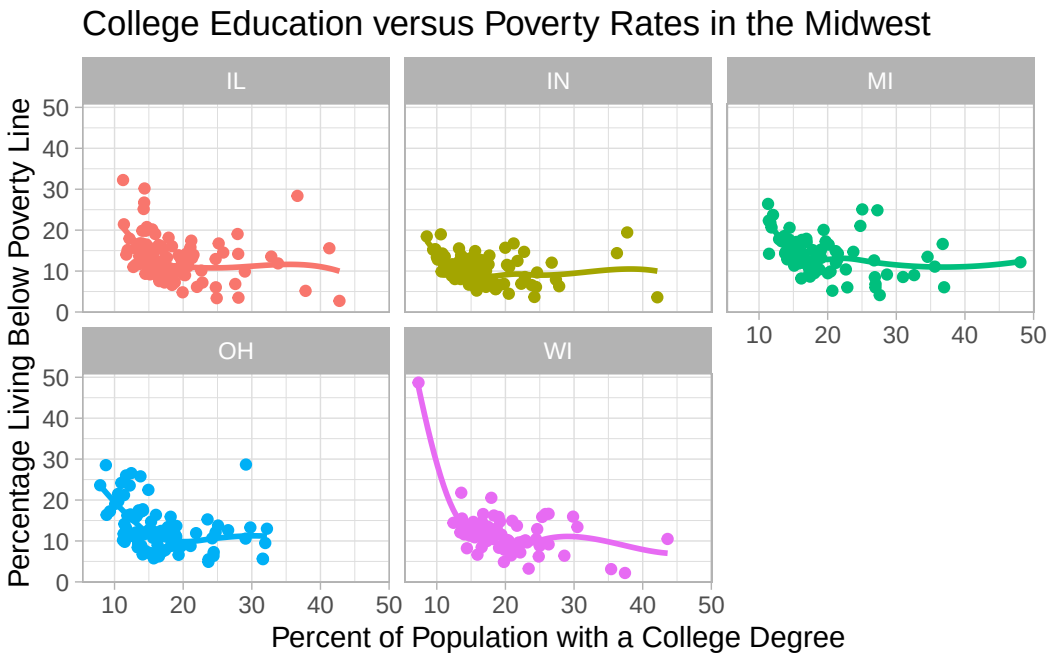
A vast majority of counties across all five Midwest states have college education rates between 10 and 25 percent and poverty rates between 5 and 20 percent. The states have a few key differences. Wisconsin has far less counties with college education rates lower than 15 percent than the other four observed states. Indiana has lower poverty rates on average than the other states. Ohio has overall the highest poverty and lowest college education rates. Illinois is somewhat similar to Ohio except that it has a number of counties with higher college education percentages.

Exercise 4

```
ggplot(midwest, mapping = aes(x = percollege, y = percbelowpoverty,
                              color = state)) +
  geom_point() +
  labs(x = "Percent of Population with a College Degree",
       y = "Percentage Living Below Poverty Line",
       title = "College Education versus Poverty Rates in the Midwest") +
  theme_light() +
  theme(legend.position = "none") +
```

```
    facet_wrap("state") +
    geom_smooth(se = FALSE)
```

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




```
    scale_color_viridis_d(option = "D")
```

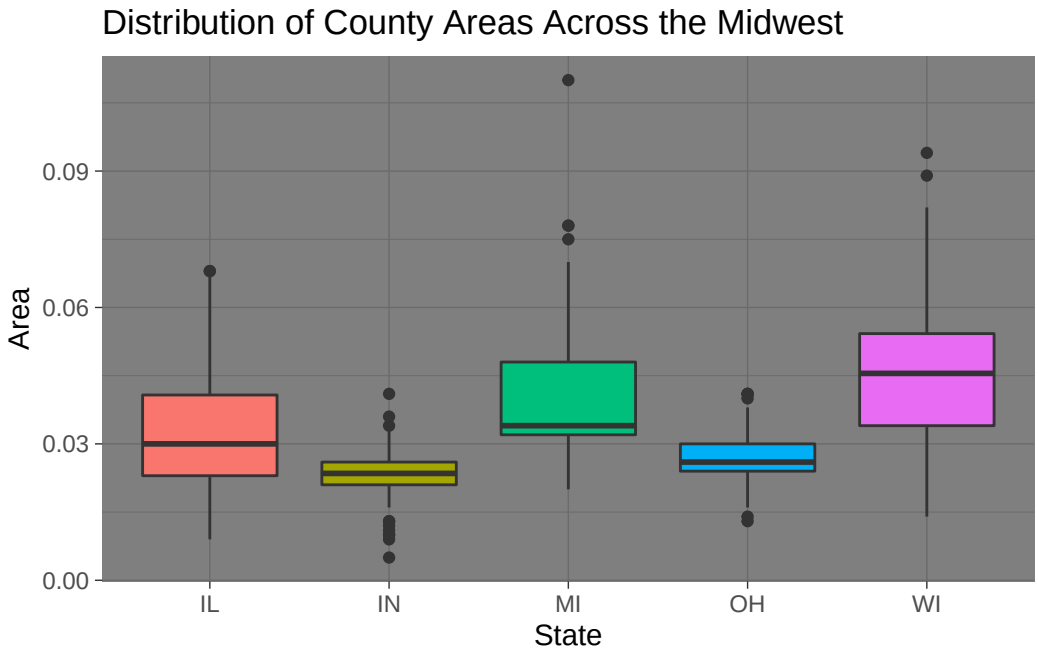
```
<ggproto object: Class ScaleDiscrete, Scale, gg>
    aesthetics: colour
    axis_order: function
    break_info: function
    break_positions: function
    breaks: waiver
    call: call
    clone: function
    dimension: function
    drop: TRUE
    expand: waiver
    get_breaks: function
    get_breaks_minor: function
```

```
get_labels: function
get_limits: function
guide: legend
is_discrete: function
is_empty: function
labels: waiver
limits: NULL
make_sec_title: function
make_title: function
map: function
map_df: function
n.breaks.cache: NULL
na.translate: TRUE
na.value: NA
name: waiver
palette: function
palette.cache: NULL
position: left
range: <ggproto object: Class RangeDiscrete, Range, gg>
    range: NULL
    reset: function
    train: function
    super:  <ggproto object: Class RangeDiscrete, Range, gg>
rescale: function
reset: function
scale_name: viridis_d
train: function
train_df: function
transform: function
transform_df: function
super:  <ggproto object: Class ScaleDiscrete, Scale, gg>
```

This plot is preferable to the plot in exercise 2. This plot makes it easier to analyze the data from each individual state. Additionally, the fit lines give a good rough feeling of how college education rates correlation to poverty rates overall in each state.

Exercise 5

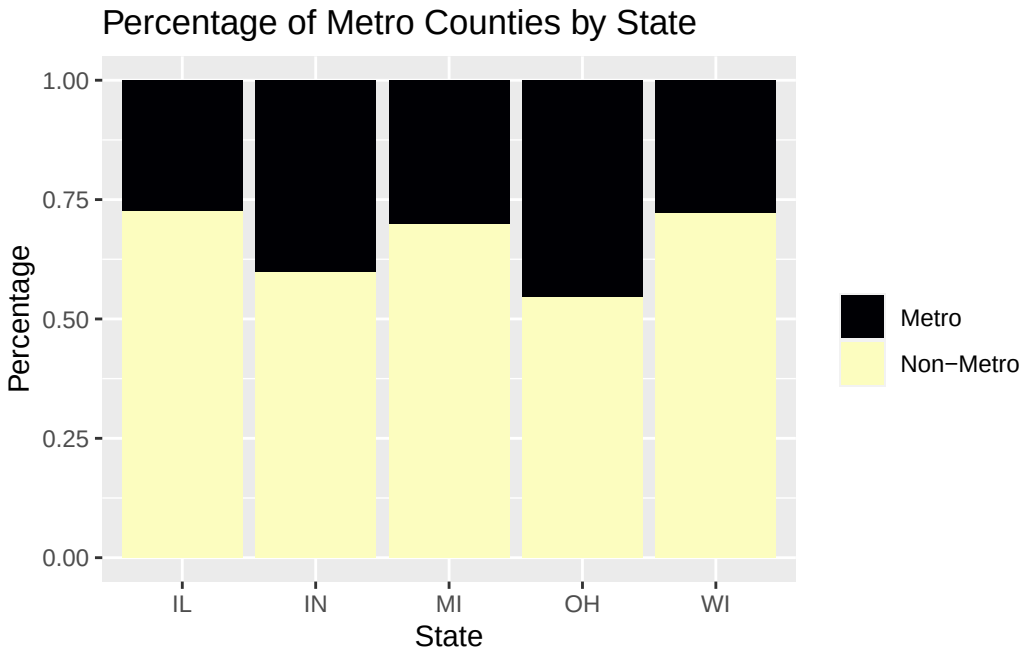
```
ggplot(midwest, mapping = aes(x = state, y = area, fill = state)) +
  geom_boxplot() +
  theme_dark() +
  labs(x = "State", y = "Area",
       title = "Distribution of County Areas Across the Midwest") +
  theme(legend.position = "none")
```



These plots reveal that on average Indiana and Ohio tend to have the smallest counties while Wisconsin and Michigan tend to have the largest counties of the observed states. Additionally, Ohio and Indiana tend to have many counties of very similar sizes whereas the other three states have much greater variance in the size of their counties. Michigan has the largest singular county as observed by the outlier data point near the top of the graph.

Exercise 6

```
midwest <- midwest |>
  mutate(metro = if_else(inmetro == 1, "Metro", "Non-Metro"))
ggplot(midwest, aes(fill=metro, y = 1, x=state)) +
    geom_bar(position="fill", stat="identity") +
  scale_fill_viridis_d(option = "A") +
  labs(x = "State", y = "Percentage",
       title = "Percentage of Metro Counties by State", fill = "")
```



It appears Ohio and Indiana have the highest percentage of metropolitan counties, both above 37.5%. The other three states all have just above 25% metropolitan counties. This correlates to the county area distribution of Ohio and Indiana being similar and the county area distribution of the other three states being similar. This would make it appear as though non-metropolitan counties have on average a larger area which correlates with what is true in reality.

Exercise 7

```
ggplot(midwest, mapping = aes(x = percollege, y = popdensity,
                              color = percbelowpoverty)) +
  geom_point(size = 2, alpha = 0.5) +
  facet_wrap(~state) +
  labs(x = "% college educated", y = "Population density (person / unit area)",
       title = "Do people with college degrees tend to live in denser areas?",
       color = "% below\npoverty line") +
  theme_minimal()
```