Lab 1 - Data visualization

Meera Patel

Load Packages

```
install.packages("tidyverse")
library(tidyverse)
```

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

```
library(viridis)
```

Exercise 1

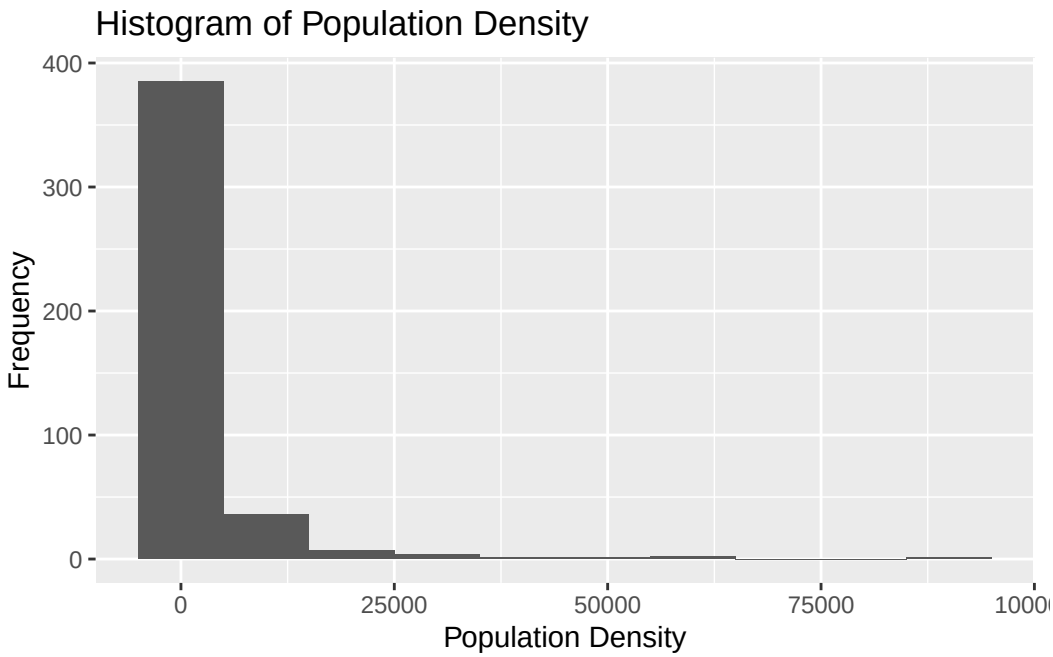
```
#1 label: histogram of population density

glimpse(midwest)
```

```
Rows: 437
Columns: 28
$ PID                 <int> 561, 562, 563, 564, 565, 566, 567, 568, 569, 570,~
$ county              <chr> "ADAMS", "ALEXANDER", "BOND", "BOONE", "BROWN", "~
$ state               <chr> "IL", "IL", "IL", "IL", "IL", "IL", "IL", "IL", "~
$ area                <dbl> 0.052, 0.014, 0.022, 0.017, 0.018, 0.050, 0.017, ~
$ poptotal            <int> 66090, 10626, 14991, 30806, 5836, 35688, 5322, 16~
$ popdensity          <dbl> 1270.9615, 759.0000, 681.4091, 1812.1176, 324.222~
$ popwhite            <int> 63917, 7054, 14477, 29344, 5264, 35157, 5298, 165~
$ popblack            <int> 1702, 3496, 429, 127, 547, 50, 1, 111, 16, 16559,~
$ popamerindian       <int> 98, 19, 35, 46, 14, 65, 8, 30, 8, 331, 51, 26, 17~
$ popasian            <int> 249, 48, 16, 150, 5, 195, 15, 61, 23, 8033, 89, 3~
```

```
$ popother             <int> 124, 9, 34, 1139, 6, 221, 0, 84, 6, 1596, 20, 7, ~
$ percwhite            <dbl> 96.71206, 66.38434, 96.57128, 95.25417, 90.19877,~
$ percblack            <dbl> 2.57527614, 32.90043290, 2.86171703, 0.41225735, ~
$ percamerindan        <dbl> 0.14828264, 0.17880670, 0.23347342, 0.14932156, 0~
$ percasian            <dbl> 0.37675897, 0.45172219, 0.10673071, 0.48691813, 0~
$ percother            <dbl> 0.18762294, 0.08469791, 0.22680275, 3.69733169, 0~
$ popadults            <int> 43298, 6724, 9669, 19272, 3979, 23444, 3583, 1132~
$ perchsd              <dbl> 75.10740, 59.72635, 69.33499, 75.47219, 68.86152,~
$ percollege           <dbl> 19.63139, 11.24331, 17.03382, 17.27895, 14.47600,~
$ percprof             <dbl> 4.355859, 2.870315, 4.488572, 4.197800, 3.367680,~
$ poppovertyknown      <int> 63628, 10529, 14235, 30337, 4815, 35107, 5241, 16~
$ percpovertyknown     <dbl> 96.27478, 99.08714, 94.95697, 98.47757, 82.50514,~
$ percbelowpoverty     <dbl> 13.151443, 32.244278, 12.068844, 7.209019, 13.520~
$ percchildbelowpovert <dbl> 18.011717, 45.826514, 14.036061, 11.179536, 13.02~
$ percadultpoverty     <dbl> 11.009776, 27.385647, 10.852090, 5.536013, 11.143~
$ percelderlypoverty   <dbl> 12.443812, 25.228976, 12.697410, 6.217047, 19.200~
$ inmetro              <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0~
$ category             <chr> "AAR", "LHR", "AAR", "ALU", "AAR", "AAR", "LAR", ~
```

```
ggplot(data = midwest, mapping = aes(x = popdensity)) +
  geom_histogram(binwidth = 10000) +
  labs(x = "Population Density", y = "Frequency", title = "Histogram of Population Density
  scale_color_viridis_d()
```



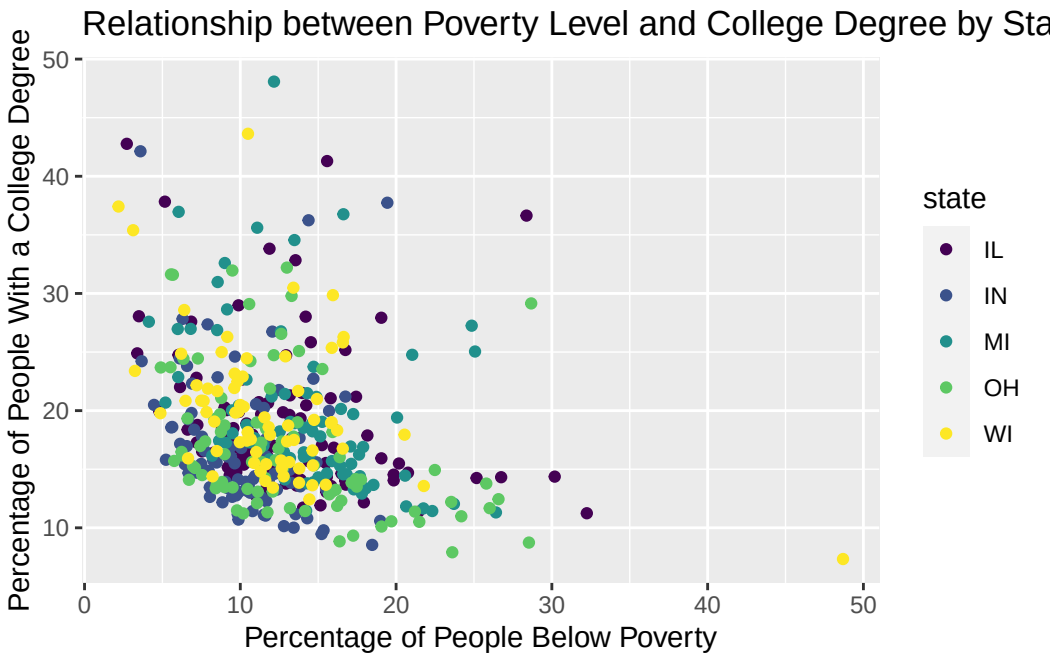
The distribution appears to be unimodal and strongly right-skewed.

There appears to be an outlier with a population density between 75000 and 100000. The data value(s) between 50000 and 75000 population density could also potentially be outliers.

Exercise 2

```
#l label: scatterplot

ggplot(midwest, aes(x = percbelowpoverty, y = percollege, color = state)) +
  geom_point() +
  scale_color_viridis_d() +
  labs(x = "Percentage of People Below Poverty", y = "Percentage of People With a College
```



Exercise 3

It appears that for most states in this dataset, there is a general weak negative correlation between the percentage of people below poverty and the percentage of people with a college degree. This means that in general, as the percentage of people below poverty increases, the percentage of people with college degrees increases. This correlation appears to be stronger in Wisconsin and Ohio and weaker in the other Midwestern states in this dataset. There is a strong outlier in Wisconsin, at which there are approximately 48% of people below poverty and around 5% of people with a college degree. In Illinois, there appears to be very little relationship between the percentage of people with a college degree and the percentage of people below the poverty line, whereas in Wisconsin, there appears to be a stronger relationship between the 2 variables.

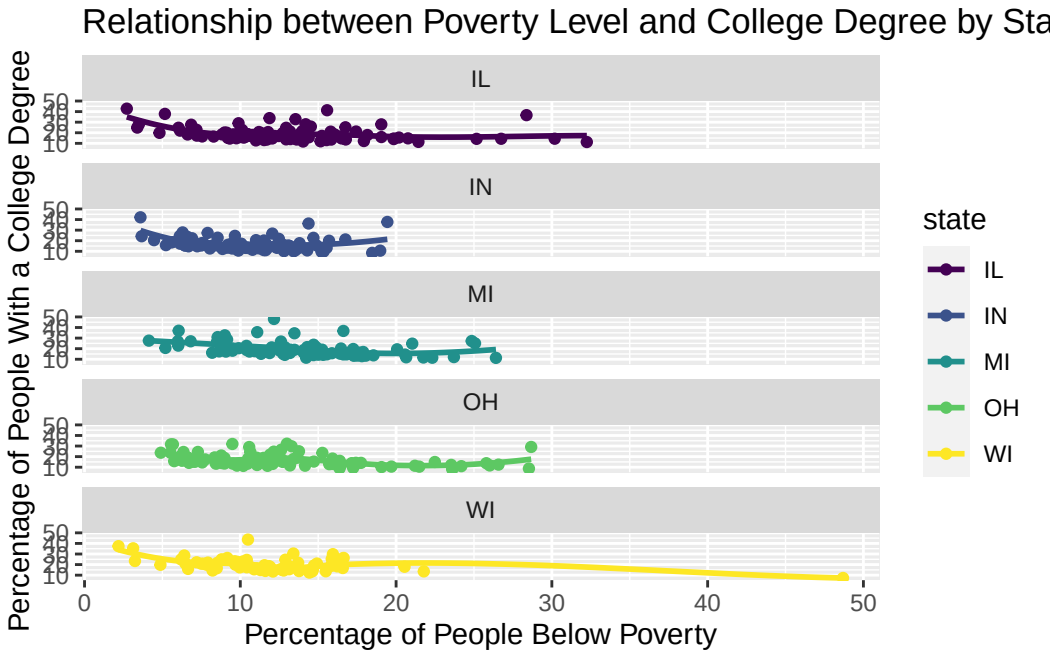
Exercise 4

```
#1 label: scatterplot by state

ggplot(midwest, aes(x = percbelowpoverty, y = percollege, color = state)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  scale_color_viridis_d() +
```

```
    labs(x = "Percentage of People Below Poverty", y = "Percentage of People With a College
    facet_wrap(~ state, nrow = 5)
```

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



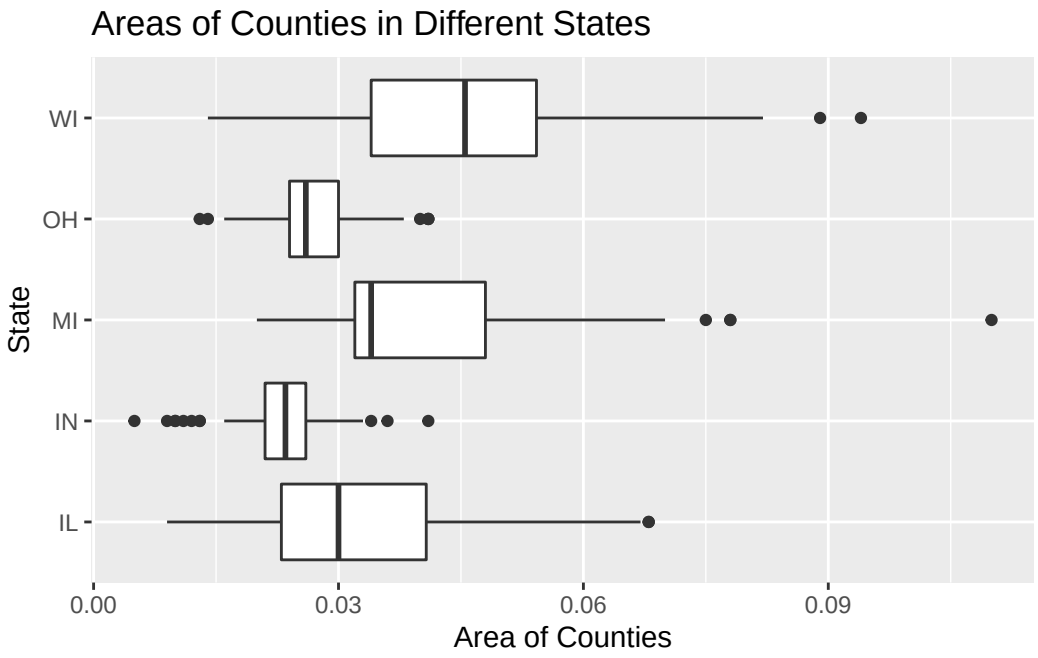
I prefer this plot rather than the one from Exercise 2 because I'm more clearly able to see the overall trend and relationship between the percentage of people below poverty and the percentage of people with a college degree for each individual state. The line provides a better idea of the overall trend in the percentage of people with college degrees as the percentage of people below poverty increases. This plot also separates each scatterplot by state, which is helpful because sometimes, the datapoints for each state were covered up by other datapoints from different states in the plot from Exercise 2, which made it more difficult to visualize the trend for each state, since many datapoints appeared to be "missing."

Exercise 5

```
#1 label: boxplot of area

ggplot(midwest, aes(x = area, y = state)) +
  geom_boxplot() +
```

```
  labs(x = "Area of Counties", y = "State", title = "Areas of Counties in Different States
```



It appears that Wisconsin has the largest median county size, followed by Michigan, then Illinois, then Ohio, and finally Indiana. Each dataset has at least one datapoint that does not fall within the first and third quartiles. Wisconsin appears to have the largest spread (variability) in county size, while Indiana appears to have the smallest spread. Michigan and Illinois appear to be slightly right-skewed in the area of their counties. Michigan appears to have an outlier, with a very large county.

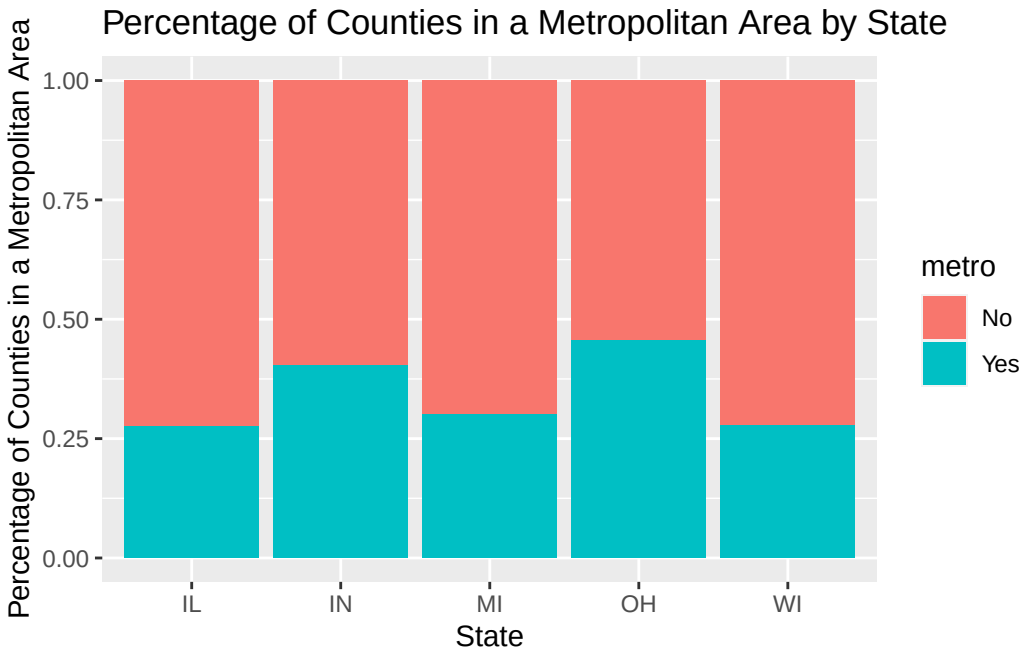
Michigan appears to have the single largest county from the states in this dataset. This is because Michigan has a datapoint that is the furthest up on the x-axis–that is, there is a county in Michigan that has a higher area than any of the other counties in the other states. This single datapoint (and potential outlier) implies that Michigan has the largest county out of the 5 Midwestern states in this dataset.

Exercise 6

```
#l label: metro plot

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

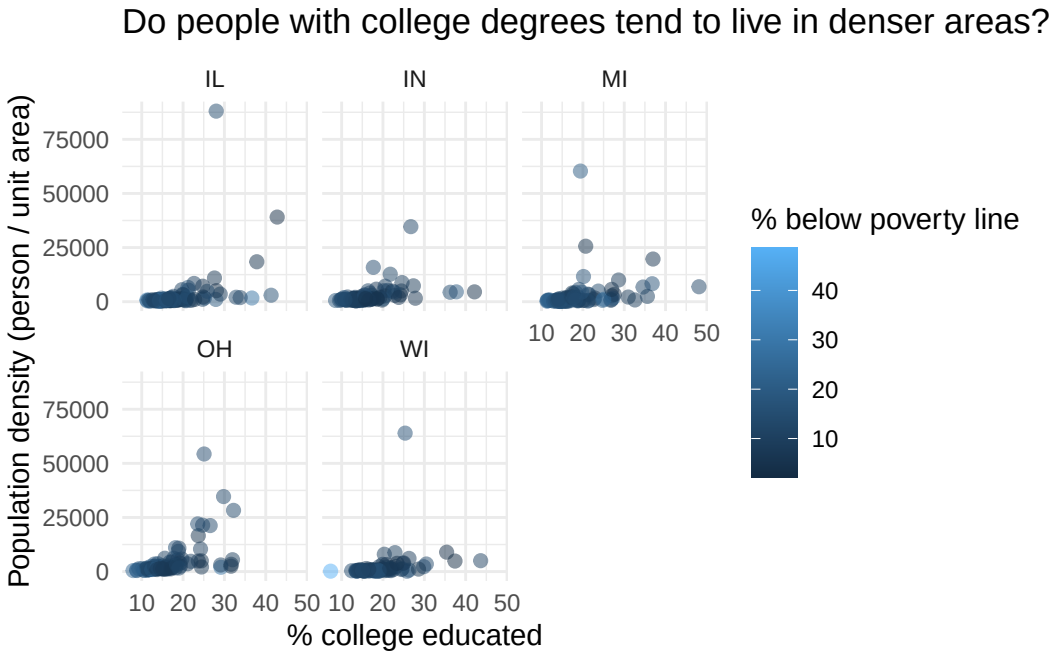
```r
ggplot(midwest, aes(x = state, fill = metro)) +
  geom_bar(position = "fill") +
  labs(x = "State", y = "Percentage of Counties in a Metropolitan Area", title = "Percenta
```



Exercise 7

```r
#1 label: problem solving

ggplot(midwest, aes(x = percollege, y = popdensity, color = percbelowpoverty)) +
  geom_point(alpha = 0.5, size = 2) +
  labs(x = "% college educated", y = "Population density (person / unit area)", title = "D
  facet_wrap(~ state) +
  theme_minimal()
```


Do people with college degrees tend to live in denser areas?
IL
IN
MI
OH
WI
Population density (person / unit area)
75000
50000
25000
0
10 20 30 40 50
% college educated
% below poverty line
40
30
20
10