Lab 2 - Data wrangling

Jessica St. Jean

2026-02-10

```
library(tidyverse)
```

Questions

Part 1

```
?midwest
```

Question 1

```
midwest |>
count(state, sort = TRUE)
```

```
# A tibble: 5 x 2
  state     n
  <chr> <int>
1 IL      102
2 IN       92
3 OH       88
4 MI       83
5 WI       72
```

Comment: Illinois has the most amount of counties with 102, and Wisconsin has the least with 72.

Question 2

```
midwest |>
  count(county,state) |>
  count(county, name = "n_states")|>
  filter(n_states == n_distinct(midwest$state))
```

```
# A tibble: 3 x 2
  county   n_states
  <chr>       <int>
1 CRAWFORD        5
2 JACKSON         5
3 MONROE          5
```

Question 3

```
midwest |>
  filter(popdensity > 25000) |>
  select(county,state,popdensity,poptotal,area) |>
  arrange(desc(popdensity))
```

```
# A tibble: 9 x 5
  county    state popdensity poptotal  area
  <chr>     <chr>      <dbl>    <int> <dbl>
1 COOK      IL        88018.  5105067 0.058
2 MILWAUKEE WI        63952.   959275 0.015
3 WAYNE     MI        60334.  2111687 0.035
4 CUYAHOGA  OH        54313.  1412140 0.026
5 DU PAGE   IL        39083.   781666 0.02
6 MARION    IN        34659.   797159 0.023
7 HAMILTON  OH        34649.   866228 0.025
8 FRANKLIN  OH        28278.   961437 0.034
9 MACOMB    MI        25621.   717400 0.028
```

```
midwest |>
  filter(popdensity == max(popdensity)) |>
  select(county,state,popdensity,poptotal,area)
```

```
# A tibble: 1 x 5
  county state popdensity poptotal  area
  <chr>  <chr>      <dbl>    <int> <dbl>
1 COOK   IL        88018.  5105067 0.058
```

Question 4

```
midwest|>
  summarize(
    median(popdensity),
    q1 = quantile(popdensity, 0.25),
    q3 = quantile(popdensity, 0.75)
  )
```

```
# A tibble: 1 x 3
  `median(popdensity)`    q1    q3
                 <dbl> <dbl> <dbl>
1                1156.  622.  2330
```

Comment:

The distribution of population density of counties is unimodal and extremely right-skewed. A typical Midwestern county has population density of 1156.208 people per unit area. The middle 50% of the counties have population densities between 622.4074 to 2330 people per unit area.

Question 5

```
midwest |>
  count(state,inmetro)|>
  group_by(state) |>
  mutate(prop=n/sum(n))
```

```
# A tibble: 10 x 4
# Groups:   state [5]
   state inmetro     n  prop
   <chr>   <int> <int> <dbl>
 1 IL          0    74 0.725
 2 IL          1    28 0.275
 3 IN          0    55 0.598
 4 IN          1    37 0.402
 5 MI          0    58 0.699
 6 MI          1    25 0.301
 7 OH          0    48 0.545
 8 OH          1    40 0.455
 9 WI          0    52 0.722
10 WI          1    20 0.278
```

Question 6

```
midwest |>
  filter(percbelowpoverty >=40,
         percollege <=10) |>
  select(county,state,
         percbelowpoverty,
         percollege)
```

```
# A tibble: 1 x 4
  county    state percbelowpoverty percollege
  <chr>     <chr>            <dbl>      <dbl>
1 MENOMINEE WI                48.7       7.34
```

```
midwest |>
  filter(percbelowpoverty <= 20,
         percollege >= 40) |>
  select(county, state,
         percbelowpoverty,
         percollege)
```

```
# A tibble: 5 x 4
  county    state percbelowpoverty percollege
  <chr>     <chr>            <dbl>      <dbl>
1 CHAMPAIGN IL               15.6        41.3
2 DU PAGE   IL                2.71       42.8
3 HAMILTON  IN                3.59       42.1
4 WASHTENAW MI               12.2        48.1
5 DANE      WI               10.5        43.6
```

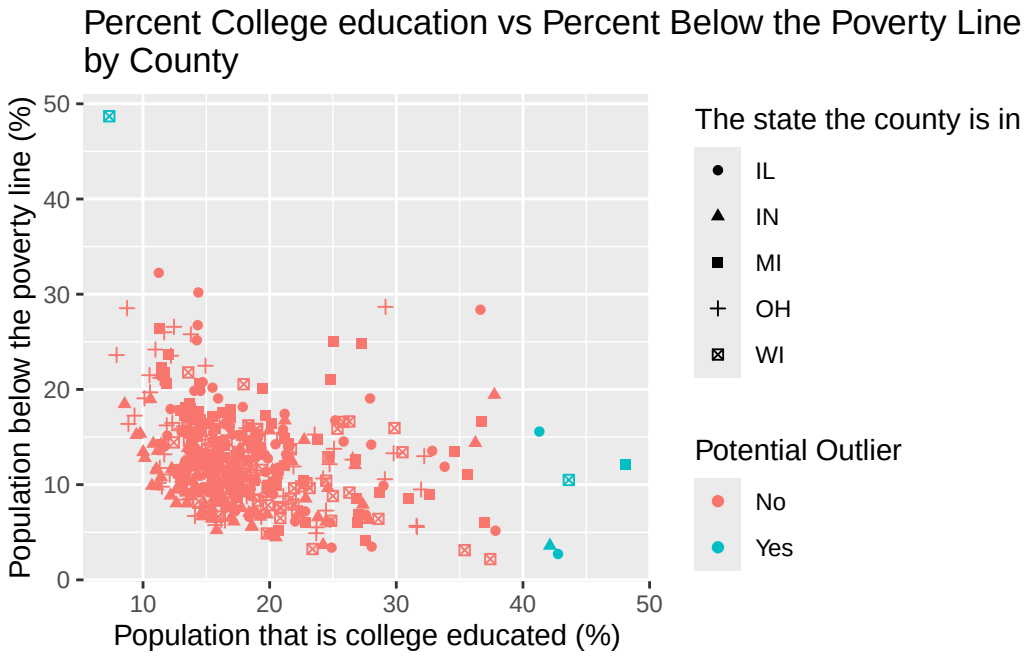
```
midwest |>
  filter(
    (percbelowpoverty >= 40 & percollege <= 10) |
  (percbelowpoverty <=20 & percollege >= 40)
) |>
  select(county, state,
         percbelowpoverty,
         percollege)
```

```
# A tibble: 6 x 4
  county    state percbelowpoverty percollege
  <chr>     <chr>            <dbl>      <dbl>
1 CHAMPAIGN IL               15.6        41.3
2 DU PAGE   IL                2.71       42.8
3 HAMILTON  IN                3.59       42.1
4 WASHTENAW MI               12.2        48.1
5 DANE      WI               10.5        43.6
6 MENOMINEE WI               48.7         7.34
```

```
midwest |>
   mutate(
     potential_outlier = if_else(
       (percbelowpoverty >= 40 & percollege <= 10) |
         (percbelowpoverty <=20 & percollege >= 40),
       "Yes",
       "No"
     )
     )|>
  select(county, state,
         percbelowpoverty,
         percollege,
         potential_outlier)|>
  arrange(potential_outlier)
```

```
# A tibble: 437 x 5
   county    state percbelowpoverty percollege potential_outlier
   <chr>     <chr>            <dbl>      <dbl> <chr>
 1 ADAMS     IL               13.2        19.6 No
 2 ALEXANDER IL               32.2        11.2 No
 3 BOND      IL               12.1        17.0 No
 4 BOONE     IL                7.21       17.3 No
 5 BROWN     IL               13.5        14.5 No
 6 BUREAU    IL               10.4        18.9 No
 7 CALHOUN   IL               15.1        11.9 No
 8 CARROLL   IL               11.7        16.2 No
 9 CASS      IL               13.9        14.1 No
10 CHRISTIAN IL               11.7        13.6 No
# i 427 more rows
```

```r
midwest |>
   mutate(
     potential_outlier = if_else(
       (percbelowpoverty >= 40 & percollege <= 10) |
         (percbelowpoverty <=20 & percollege >= 40),
       "Yes",
       "No"
     )
     )|>
ggplot(aes(x=percollege, y=percbelowpoverty, colour = potential_outlier, shape = state)) +
geom_point() +
  labs(x= "Population that is college educated (%)",
    y= "Population below the poverty line (%)",
    colour = "Potential Outlier",
    shape = "The state the county is in",
    title= "Percent College education vs Percent Below the Poverty Line \nby County")
```



Question 7

Question 8

Part 2

Question 9