# Week 3 homework assignment

Anubha Nagar

16th September 2021

## DSI-EDA

## Professor Michael Shepherd

#This homework is designed to get you to review the in-class notes + r code as well as work on your own code.

## Part 1: Midwest Data

Recall our use of the `midwest` data from week3-day1.

### 1. In your own words, what does the function in this line of "week3-day1.Rmd" do?

```
#g1 <- g1 + scale_y_continuous(breaks=seq(0, 1000000, 200000), labels = function(x){paste0(x/1000, 'K')
```

Answer 1:

- Y axis is a continuous variable and hence we use scale_y_continuous
- The breaks refers to the y axis limits and the increment value. So in this example "breaks=seq(0, 1000000, 200000" means that the y axis starts from 0 and goes till 1000000 with increments of 200000. Each increment is plotted with a marker.
- labels is used to make thousands in terms of 'K' for example 1000 is written as 1K and so on. So each y label is changed in terms of K.

### 2. Starting with one variable: During week3-day1, we learned about how to make a scatterplot in ggplot using `midwest` data. This was a useful illustration for how to (1) make a guess at a bivariate relationship in the data and (2) explore it using a scatterplot. But ultimately the graphic wasn't that interesting. Sometimes we need to take a step back and simply plot one variable at a time.

Explore the relationship of population totals by state. Include a clear title, and change the xlab and ylab to be easy to read words (labels), try using geom_col for this. Interestingly you *could* force the outcome using geom_histogram() but typically we want to use histograms for a singular variable.

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.0     v forcats 0.5.1
## v purrr   0.3.4
```

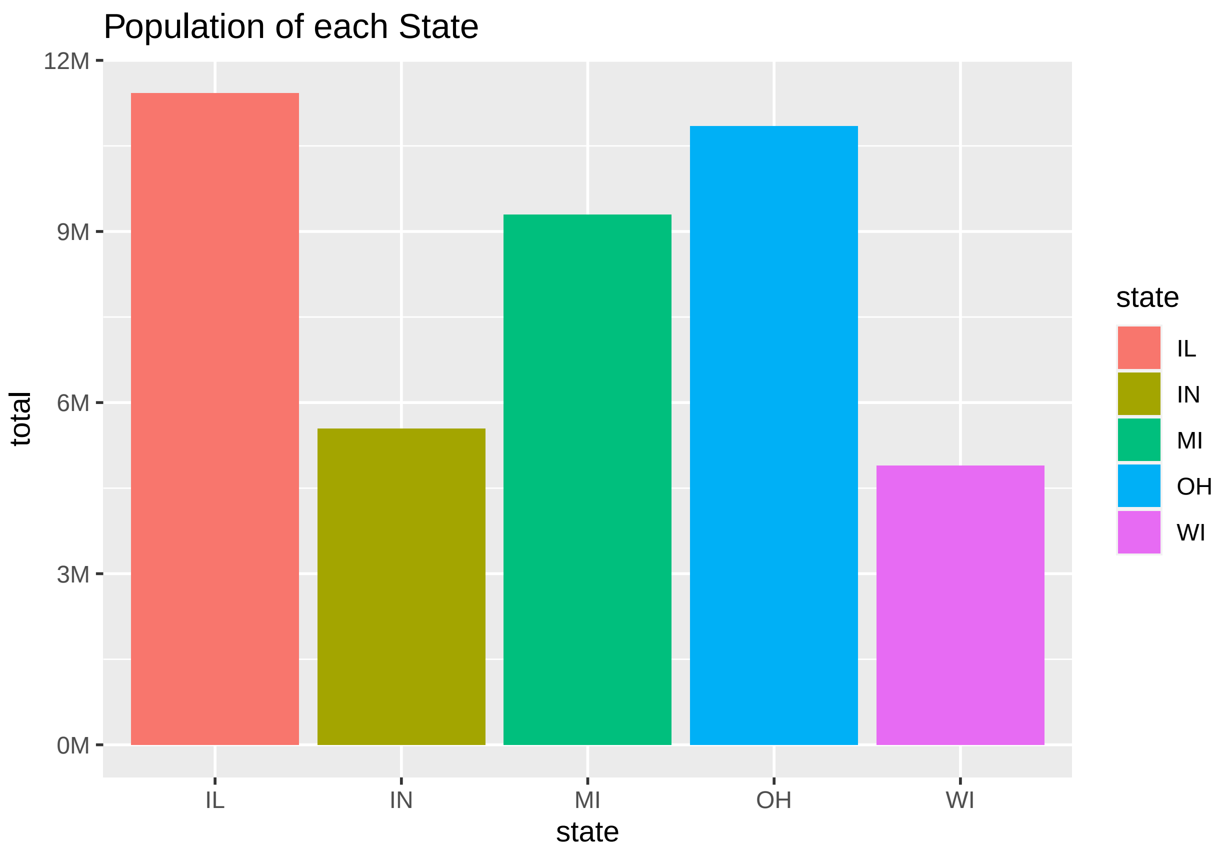
```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#Demographic information of midwest counties
# https://ggplot2.tidyverse.org/reference/midwest.html
data("midwest", package = "ggplot2")

midwest
```

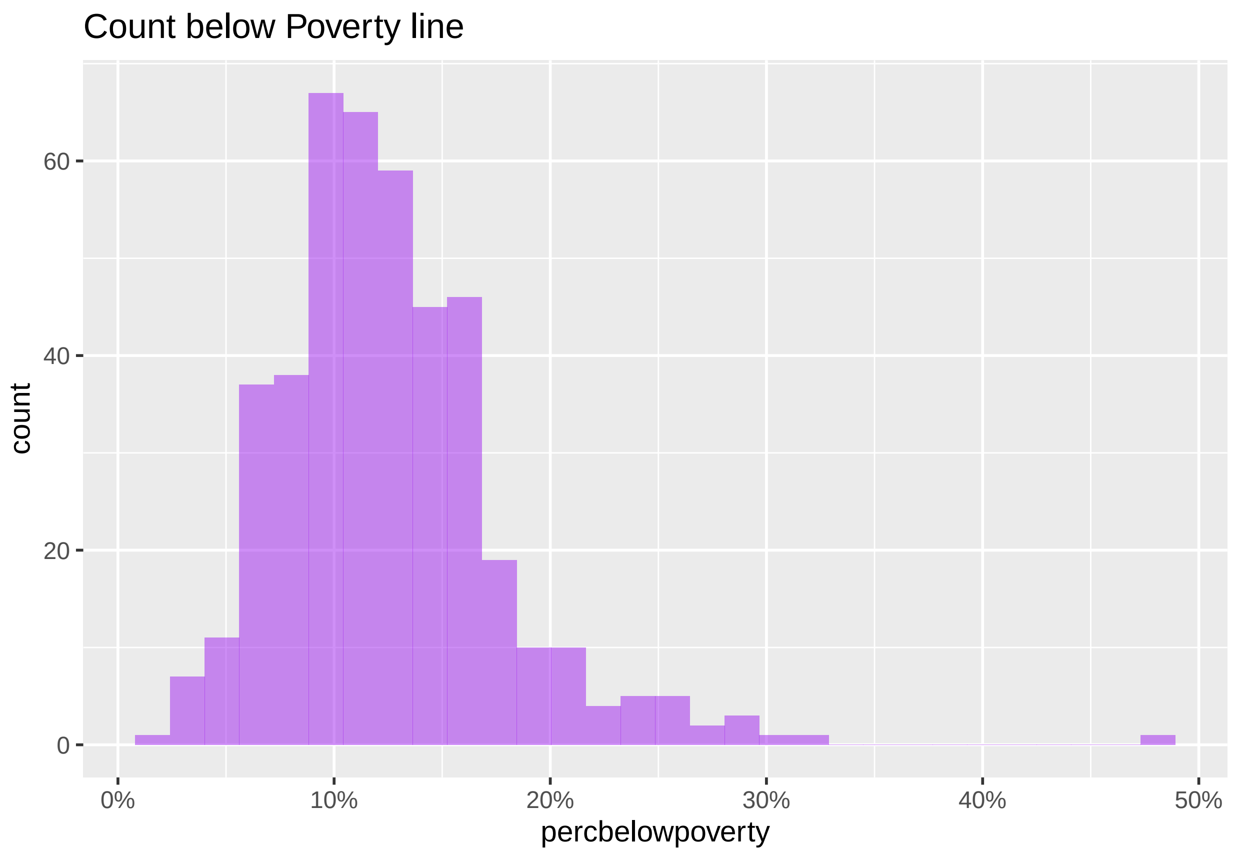
```
## # A tibble: 437 x 28
##      PID county    state  area poptotal popdensity popwhite popblack popamerindian
##    <int> <chr>     <chr> <dbl>    <int>      <dbl>    <int>    <int>         <int>
##  1   561 ADAMS     IL    0.052    66090      1271.    63917     1702            98
##  2   562 ALEXANDER IL    0.014    10626       759      7054     3496            19
##  3   563 BOND      IL    0.022    14991       681.    14477      429            35
##  4   564 BOONE     IL    0.017    30806      1812.    29344      127            46
##  5   565 BROWN     IL    0.018     5836       324.     5264      547            14
##  6   566 BUREAU    IL    0.05     35688       714.    35157       50            65
##  7   567 CALHOUN   IL    0.017     5322       313.     5298        1             8
##  8   568 CARROLL   IL    0.027    16805       622.    16519      111            30
##  9   569 CASS      IL    0.024    13437       560.    13384       16             8
## 10   570 CHAMPAIGN IL    0.058   173025      2983.   146506    16559           331
## # ... with 427 more rows, and 19 more variables: popasian <int>,
## #   popother <int>, percwhite <dbl>, percblack <dbl>, percamerindan <dbl>,
## #   percasian <dbl>, percother <dbl>, popadults <int>, perchsd <dbl>,
## #   percollege <dbl>, percprof <dbl>, poppovertyknown <int>,
## #   percpovertyknown <dbl>, percbelowpoverty <dbl>, percchildbelowpovert <dbl>,
## #   percadultpoverty <dbl>, percelderlypoverty <dbl>, inmetro <int>,
## #   category <chr>
```

```
midwest %>%
  group_by(state)%>%
  summarize(total= sum(poptotal))%>%
  ggplot(aes(x = state, y = total))+ geom_col(aes(fill = state))+ ggtitle('Population of each State') +
```



### 3. Make a histogram for the percent of people below poverty

```
ggplot(midwest, aes(x = percbelowpoverty)) +
  geom_histogram(bins = 30, fill = "purple", alpha = 0.5)+labs(title = "Count below Poverty line")+scale
```



### Part 3

In class we worked on Nashville schools data. Print one `best` graphic from the Nashville schools data and write one paragraph about the graphic. If you are using the graphic your group made, try to improve it. As an added challenge for those who want one, create a completely different graphic. What did you learn? Why is this interesting?

```
# packages you need libraried for today
library(dplyr)
library(ggplot2)
metro_nash_schools <- read.csv("metro-nash-schools.csv")
head(metro_nash_schools)
```

```
##   School.Year      School.Level School.ID                   School.Name
## 1       18-19           Charter       743      Valor Flagship Academy
## 2       18-19     Middle School       545              Madison Middle
## 3       18-19       High School       450             Hume-Fogg High
## 4       18-19 Elementary School       575 Thomas A. Edison Elementary
## 5       18-19 Elementary School       185  Carter-Lawrence Elementary
## 6       18-19       High School       290        East Nashville School
##   State.School.ID Zip.Code Grade.PreK.3yrs Grade.PreK.4yrs Grade.K Grade.1
## 1            8045    37211              NA              NA      NA      NA
## 2             622    37115              NA              NA      NA      NA
## 3             355    37203              NA              NA      NA      NA
## 4             208    37013               4              31     135     156
## 5             670    37203              NA              17      44      44
## 6             203    37206              NA              NA      NA      NA
##   Grade.2 Grade.3 Grade.4 Grade.5 Grade.6 Grade.7 Grade.8 Grade.9 Grade.10
## 1      NA      NA      NA     120     120     116     133     223       NA
## 2      NA      NA      NA     156     131     123     144      NA       NA
## 3      NA      NA      NA      NA      NA      NA      NA     222      228
## 4     155     144     164      NA      NA      NA      NA      NA       NA
## 5      58      52      63      NA      NA      NA      NA      NA       NA
## 6      NA      NA      NA      NA      NA      NA      NA     172      180
##   Grade.11 Grade.12 American.Indian.or.Alaska.Native Asian
## 1       NA       NA                               NA    45
## 2       NA       NA                               NA     2
## 3      224      209                                4    95
## 4       NA       NA                               NA    19
## 5       NA       NA                                1     2
## 6      190      145                                1     5
##   Black.or.African.American Hispanic.Latino
## 1                       104             131
## 2                       316             166
## 3                       206              57
## 4                       281             214
## 5                       225              19
## 6                       631              18
##   Native.Hawaiian.or.Other.Pacific.Islander White Male Female
## 1                                         2   430  349    363
## 2                                        NA    70  303    251
## 3                                         1   520  338    545
## 4                                         1   274  406    383
## 5                                         1    30  139    139
## 6                                        NA    32  303    384
##   Economically.Disadvantaged Disability Limited.English.Proficiency Latitude
## 1                        230         78                         185 36.07080
## 2                        382         81                         145 36.26389
## 3                         81         36                           3 36.15952
## 4                        377         79                         314 36.06288
## 5                        162         31                          37 36.14365
## 6                        272         52                           3 36.18063
##   Longitude             Mapped.Location
## 1 -86.72549 (36.07080058, -86.72549463)
## 2 -86.71621 (36.26389402, -86.71620849)
## 3 -86.78154 (36.15952461, -86.78153602)
## 4 -86.60464 (36.06288453, -86.60463837)
## 5 -86.78585 (36.14365344, -86.78585349)
## 6 -86.75047 (36.18062644, -86.75047137)
```

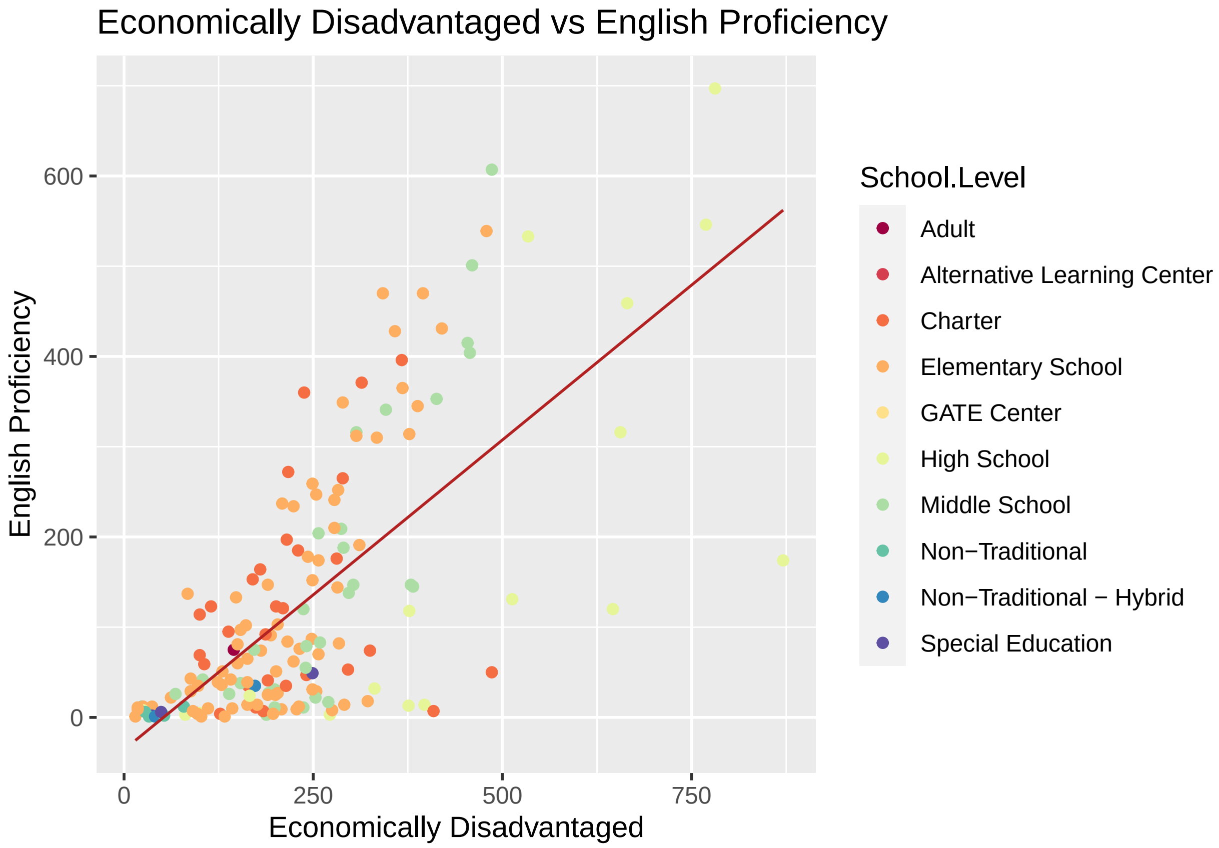
## Is there a positive relationship between the number of economically disadvantaged students and english proficiency in public schools in Nashville?

```r
g1 <- ggplot(metro_nash_schools, aes(x=Economically.Disadvantaged, y=Limited.English.Proficiency)) +
      geom_point(aes(col=School.Level)) + geom_smooth(method='lm', col='firebrick', size=0.5, se = F) +
g1
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 3 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```


Economically Disadvantaged vs English Proficiency
English Proficiency
Economically Disadvantaged
School.Level
Adult
Alternative Learning Center
Charter
Elementary School
GATE Center
High School
Middle School
Non−Traditional
Non−Traditional − Hybrid
Special Education