Week 4 HW

Zai Rutter

11/20/2021

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v stringr 1.4.0
## v tidyr   1.2.0     v forcats 0.5.1
## v readr   2.1.2

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

```
library(readr)
library(readxl)

setwd("~/Documents/Upenn/Data 210/Week 4/homework")
```

Question 1

In the chart created, there is generally a higher percent of college completion by adults when they are not in poverty. Hoewver, for children in poverty, there is less of a difference in college completion in relation to thier percent in poverty. Still, the less percent of poverty generally equates to higher college completion percents.

```r
# A
ed<- read_csv("education_long.csv")
```

```
## Rows: 795 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Name, Year
## dbl (1): College_Completion_Rate
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
ed.w<- read_csv("education_long.csv")
```

```
## Rows: 795 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Name, Year
## dbl (1): College_Completion_Rate
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

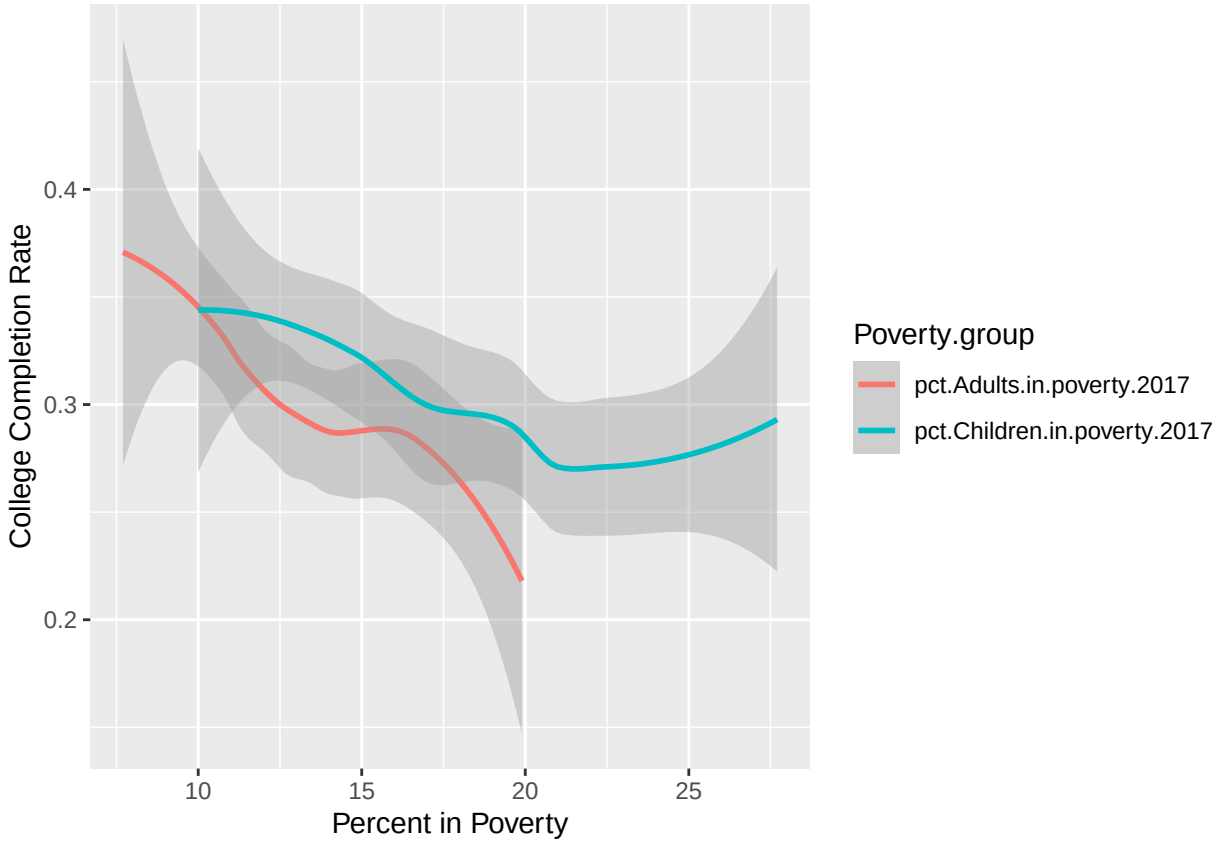
```r
# B
ed.w <- pivot_wider(ed.w, names_from = Year, values_from = College_Completion_Rate)

#C
PovertyReport <- read_excel("PovertyReport.xlsx",
                            sheet = "PovertyReport", col_types = c("text",
                                                                   "skip", "skip", "skip", "numeric",
                                                                   "skip", "numeric", "numeric", "numeri
                                                                   "numeric", "numeric"), skip = 5)
```

```
## New names:
## * Percent -> Percent...2
## * `Lower Bound` -> `Lower Bound...3`
## * `Upper Bound` -> `Upper Bound...4`
## * Percent -> Percent...5
## * `Lower Bound` -> `Lower Bound...6`
## * ...
```

```r
# D
ed.w <- ed.w %>%
  select(Name,`2013-2017_Total`,`2013-2017_Urban`,`2013-2017_Rural`)

# E
PovertyReport.w <- PovertyReport
PovertyReport.w <- PovertyReport.w[-c(53,54),]

PovertyReport.w <- rename(PovertyReport.w,
                          pct.Adults.in.poverty.2017=Percent...2,
```

```
                          LowerBound.Adults.in.poverty.2017='Lower Bound...3',
                          UpperBound.Adults.in.poverty.2017='Upper Bound...4',
                          pct.Children.in.poverty.2017='Percent...5',
                          LowerBound.Child.in.poverty.2017='Lower Bound...6',
                          UpperBound.Child.in.poverty.2017='Upper Bound...7')

Poverty.education <- merge(x = PovertyReport.w,
                y = ed.w,
                by = "Name")

# F
long.poverty.education <- pivot_longer(Poverty.education,
             cols=c(pct.Adults.in.poverty.2017, pct.Children.in.poverty.2017),
             names_to = "Poverty.group",
               values_to = "value"  )

long.poverty.education$total<-long.poverty.education$`2013-2017_Total`

ggplot(long.poverty.education, mapping=aes(x= value, y= total, color=Poverty.group )) +
  geom_smooth() +
  labs(y="College Completion Rate", x = "Percent in Poverty")
```

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



Question 2

The results from the graph created show that the less percent of black and hispanic students the higher the reading score. For asians, the highest CR score rested around 75% percent of studen body make up. For whites, the highest reading score rested around 10 - 55 percent.

```
## A
Snapshot6_12 <- read_csv("2006_-_2012_School_Demographics_and_Accountability_Snapshot.csv") # he calls
```

```
## Rows: 10075 Columns: 38
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): DBN, Name, fl_percent
## dbl (35): schoolyear, frl_percent, total_enrollment, prek, k, grade1, grade2...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

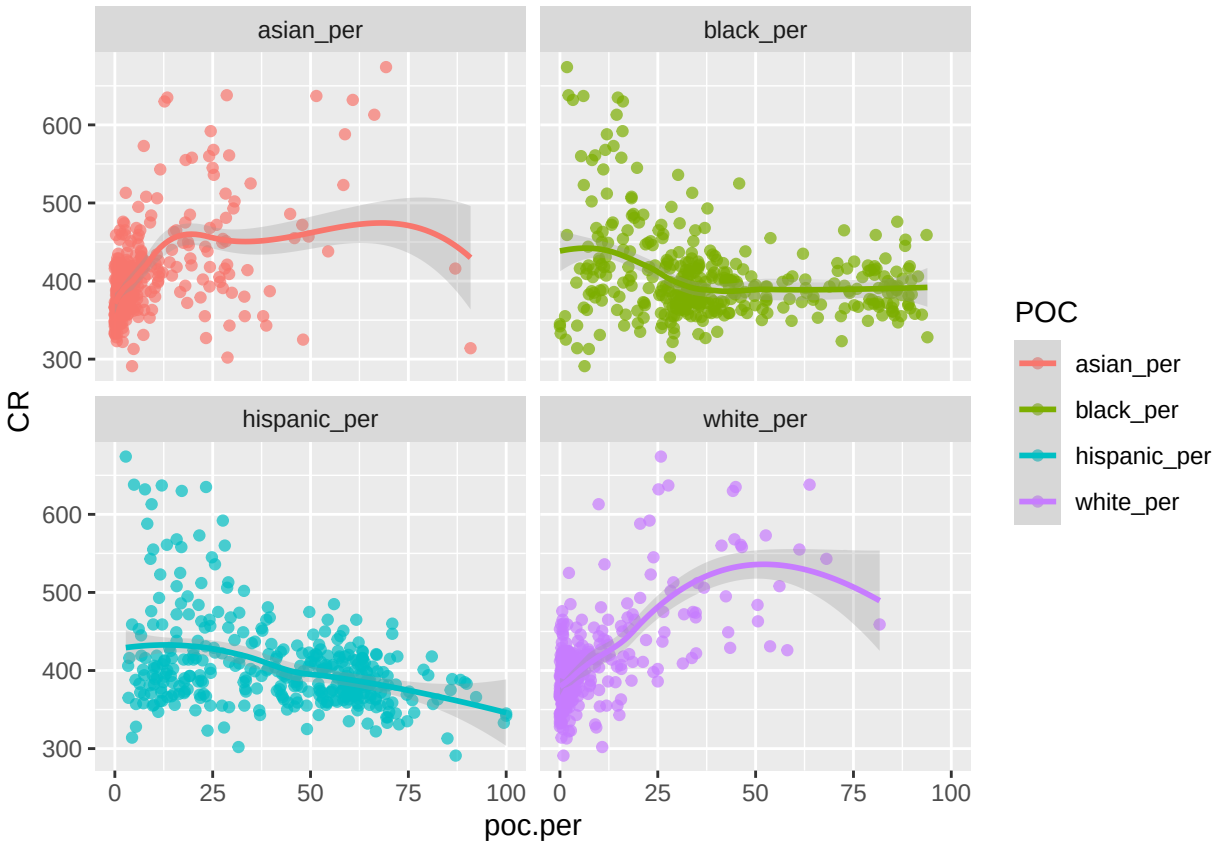
```
SAT_2010 <- read_csv("SAT__College_Board__2010_School_Level_Results.csv")
```

```
## Rows: 460 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): DBN, School Name
## dbl (4): Number of Test Takers, Critical Reading Mean, Mathematics Mean, Wri...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
### B
# For this problem, we want to merge the 2010 SAT dataset with the the NYC school
# profiles for that year.
# First, subset the NYC school profile so that it only includes data from the
# 2009/2010 school year.

Snapshot6_12.w <- subset(Snapshot6_12, schoolyear == 20092010)

### B
# If you look at the dimensions of the data, it seems that there are substantially
# more schools included in the profile2010 data. Look through both datasets
# carefully to see why this might be
# (Hint: search for areas where there are missing values)

## This is because the profile includes some elementary schools and such where as the SAT data does not

### C
# Remove the values that do not belong.
Snapshot6_12.w$dummy <- Snapshot6_12.w$DBN %in% SAT_2010$DBN

Snapshot6_12.w <- subset(Snapshot6_12.w, dummy == T)
```

```r
## D
# Merge the datasets. Which variable is the most appropriate to use as a unique identifier?

SAT_Snapshot <- merge(x = SAT_2010,
      y = Snapshot6_12.w,
      by = c("DBN"),
      all.x = T)

### E
## Now, let's explore this data a bit.
# Say we want to see how the racial demographics of a school impact the school's
# average critical reading score on the SAT.
# Plot these two variables for each race on four different graphs and discuss findings.

SAT_Snapshot.w <- pivot_longer( SAT_Snapshot,
                                cols = c(asian_per,black_per,hispanic_per,white_per),
                                names_to = "POC",
                                values_to = "poc.per")
SAT_Snapshot.w$CR <- SAT_Snapshot.w$`Critical Reading Mean`


SAT_Snapshot.w %>%
  group_by(POC) %>%
  ggplot(SAT_Snapshot.w, mapping = aes(x=poc.per, y=CR, color=POC)) +
  geom_point(alpha =.7) +
  facet_wrap(~ POC) +
  geom_smooth(alpha=.3)
```

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

## Warning: Removed 356 rows containing non-finite values (stat_smooth).

## Warning: Removed 356 rows containing missing values (geom_point).
```


asian_per
black_per
hispanic_per
white_per
CR
600
500
400
300
0
25
50
75
100
poc.per
POC
asian_per
black_per
hispanic_per
white_per