Bus 4093H Assignment 6

Ramanpreet Chand

November 1, 2016

# The Gender Gap in Mathematics and Reading

## Question 1

### Importing the Dataset

#setwd  
setwd("/Volumes/My Passport/Workspace/R\_Files/gender-gap-mathematics")  
  
#Load packages  
library(data.table)  
library(dplyr)

## -------------------------------------------------------------------------

## data.table + dplyr code now lives in dtplyr.  
## Please library(dtplyr)!

## -------------------------------------------------------------------------

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':  
##   
## between, last

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

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

library(ggplot2)  
library(tidyr)  
  
#Import the dataset  
file <- "ECLSK\_ass6.csv"  
data <- fread(file, data.table = FALSE, na.strings = "NULL")

## Warning in fread(file, data.table = FALSE, na.strings = "NULL"): Bumped  
## column 8 to type character on data row 40, field contains 'NA'. Coercing  
## previously read values in this column from logical, integer or numeric back  
## to character which may not be lossless; e.g., if '00' and '000' occurred  
## before they will now be just '0', and there may be inconsistencies with  
## treatment of ',,' and ',NA,' too (if they occurred in this column before  
## the bump). If this matters please rerun and set 'colClasses' to 'character'  
## for this column. Please note that column type detection uses the first  
## 5 rows, the middle 5 rows and the last 5 rows, so hopefully this message  
## should be very rare. If reporting to datatable-help, please rerun and  
## include the output from verbose=TRUE.

## Warning in fread(file, data.table = FALSE, na.strings = "NULL"): Bumped  
## column 9 to type character on data row 40, field contains 'NA'. Coercing  
## previously read values in this column from logical, integer or numeric back  
## to character which may not be lossless; e.g., if '00' and '000' occurred  
## before they will now be just '0', and there may be inconsistencies with  
## treatment of ',,' and ',NA,' too (if they occurred in this column before  
## the bump). If this matters please rerun and set 'colClasses' to 'character'  
## for this column. Please note that column type detection uses the first  
## 5 rows, the middle 5 rows and the last 5 rows, so hopefully this message  
## should be very rare. If reporting to datatable-help, please rerun and  
## include the output from verbose=TRUE.

## Warning in fread(file, data.table = FALSE, na.strings = "NULL"): Bumped  
## column 7 to type character on data row 308, field contains 'NA'. Coercing  
## previously read values in this column from logical, integer or numeric back  
## to character which may not be lossless; e.g., if '00' and '000' occurred  
## before they will now be just '0', and there may be inconsistencies with  
## treatment of ',,' and ',NA,' too (if they occurred in this column before  
## the bump). If this matters please rerun and set 'colClasses' to 'character'  
## for this column. Please note that column type detection uses the first  
## 5 rows, the middle 5 rows and the last 5 rows, so hopefully this message  
## should be very rare. If reporting to datatable-help, please rerun and  
## include the output from verbose=TRUE.

#Class Declarations  
class(data$C1R4RSCL) <- "numeric"

## Warning in class(data$C1R4RSCL) <- "numeric": NAs introduced by coercion

class(data$C1R4MSCL) <- "numeric"

## Warning in class(data$C1R4MSCL) <- "numeric": NAs introduced by coercion

class(data$C5R4RSCL) <- "numeric"

## Warning in class(data$C5R4RSCL) <- "numeric": NAs introduced by coercion

class(data$C5R4MSCL) <- "numeric"

## Warning in class(data$C5R4MSCL) <- "numeric": NAs introduced by coercion

#Dimensions  
dim(data)

## [1] 21409 15

#Data Frame  
str(data, max.level = 1)

## 'data.frame': 21409 obs. of 15 variables:  
## $ CHILDID : chr "0001001C" "0001002C" "0001003C" "0001004C" ...  
## $ DOBMM : int 3 7 8 6 10 9 9 12 5 2 ...  
## $ DOBDD : int 22 28 18 24 9 20 2 7 1 3 ...  
## $ DOBYY : int 1993 1992 1993 1993 1992 1993 1993 1992 1993 1993 ...  
## $ GENDER : int 2 2 2 1 2 1 1 2 1 2 ...  
## $ RACE : int 1 1 1 1 1 1 1 8 1 1 ...  
## $ R1\_KAGE : chr "69.07" "77.20" "64.10" "65.93" ...  
## $ C1R4RSCL: num 36.6 50.8 40.7 34.7 38.3 ...  
## $ C1R4MSCL: num 39.5 44.4 28.6 26.6 40.9 ...  
## $ C5R4RSCL: num NA 185 NA 152 150 ...  
## $ C5R4MSCL: num NA 125 NA 114 105 ...  
## $ P1HMAFB : chr "26" "24" "31" "22" ...  
## $ WKSESL : chr "1.49" "1.56" "1.61" "1.04" ...  
## $ P1WICMOM: chr "2" "2" "2" "2" ...  
## $ P1WICCHD: chr "2" "2" "2" "2" ...

Using the fread() function of the data.tables library, I imported the **Early Childhood Longitudinal Study (ECLS)**, while classifing each score in reading and math as a **numeric** data type. A numeric data type, unlike the integer data type, will store an exact value instead of rounded integer.

## Question No.2

data <- data %>% filter(C1R4RSCL > 0, C1R4MSCL > 0, C5R4RSCL > 0, C5R4MSCL > 0, GENDER > 0, RACE > 0, P1HMAFB > 0)

I've eliminated the observations that contain negative values by filtering the dataset's values to be greater than 0.

## Question No.3

data <- data %>% mutate(GENDER = plyr::mapvalues(data$GENDER, from=c(1,2), to=c("Male", "Female")),  
 RACE = plyr::mapvalues(data$RACE, from=c(1,2,3,4,5,6,7,8),   
 to=c("White", "Non-His Black or Afn Amn, Non-His", "His, Race Sepific",  
 "His, Race Non-Specific", "Asian", "Native Hawaiian, other PI",  
 "Amn Indian or ALSK Native", "More than one Race, Non-His")))

I've redefined the **Gender** and **Race** categories using the mapvalues() function from the plyr package.

#Data Reshape:  
dataS <- gather(data, Score\_Type, Score, C1R4RSCL:C5R4MSCL)  
head(dataS,5)

## CHILDID DOBMM DOBDD DOBYY GENDER RACE R1\_KAGE P1HMAFB WKSESL P1WICMOM  
## 1 0001002C 7 28 1992 Female White 77.20 24 1.56 2  
## 2 0001004C 6 24 1993 Male White 65.93 22 1.04 2  
## 3 0001005C 10 9 1992 Female White 74.53 25 1.99 2  
## 4 0001006C 9 20 1993 Male White 63.00 32 1.55 2  
## 5 0001007C 9 2 1993 Male White 63.60 36 0.62 2  
## P1WICCHD Score\_Type Score  
## 1 2 C1R4RSCL 50.82  
## 2 2 C1R4RSCL 34.72  
## 3 2 C1R4RSCL 38.28  
## 4 2 C1R4RSCL 32.57  
## 5 2 C1R4RSCL 31.98

I've reshaped my data using the **gather()** function of the **tidyr** package. Now, I can access the scores of 1st grade and 5th grade, reading and mathematics from one column.

## Question No.4

data.AvgScore <- dataS %>% group\_by(GENDER, Score\_Type) %>% summarize(avgScore = round(mean(Score, na.rm=TRUE),2)) %>% data.table()  
data.AvgScore <- spread(data.AvgScore, Score\_Type, avgScore)  
data.AvgScore

## GENDER C1R4MSCL C1R4RSCL C5R4MSCL C5R4RSCL  
## 1: Female 27.04 36.51 98.62 131.89  
## 2: Male 27.34 35.22 103.49 127.10

As the Harvard Paper stated, male and female 1st graders on average start with similar reading and mathematical abilities. However, by the 5th grade, male and female students diverge in Reading and Math. Females seem to fair on average better in Reading compared to Boys who fair on average better in Math.

## Question No.5

data.AvgScore.Race <- dataS %>% group\_by(RACE, Score\_Type) %>% summarize(avgScore = round(mean(Score, na.rm=TRUE),2)) %>% data.table()  
data.AvgScore.Race <- spread(data.AvgScore.Race, Score\_Type, avgScore)  
data.AvgScore.Race

## RACE C1R4MSCL C1R4RSCL C5R4MSCL C5R4RSCL  
## 1: Amn Indian or ALSK Native 22.90 30.57 85.88 106.97  
## 2: Asian 29.54 39.60 108.54 134.57  
## 3: His, Race Non-Specific 22.80 31.59 94.13 119.28  
## 4: His, Race Sepific 24.86 33.78 96.94 126.47  
## 5: More than one Race, Non-His 26.56 35.75 100.37 129.07  
## 6: Native Hawaiian, other PI 23.66 32.72 93.34 117.18  
## 7: Non-His Black or Afn Amn, Non-His 22.78 32.85 83.83 111.65  
## 8: White 28.83 37.04 105.78 134.99

Among the 8 races included in the dataset, 1st grade and 5th grade **Asians** have the highest average in Reading and Math. The weakest Race is **American Indian or Alaska Native** in 1st and 5th grade who scored on average less than the other races in Reading and Math.

## Question No.6

data.quantile <- data %>% group\_by(GENDER) %>% summarize(C1Read10 = quantile(C1R4RSCL, .1, na.rm = TRUE),  
 C1Read90 = quantile(C1R4RSCL, .9, na.rm = TRUE),  
 C1Math10 = quantile(C1R4MSCL, .1, na.rm = TRUE),  
 C1Math90 = quantile(C1R4MSCL, .9, na.rm = TRUE),  
 C5Read10 = quantile(C5R4RSCL, .1, na.rm = TRUE),  
 C5Read90 = quantile(C5R4RSCL, .9, na.rm = TRUE),  
 C5Math10 = quantile(C5R4MSCL, .1, na.rm = TRUE),  
 C5Math90 = quantile(C5R4MSCL, .9, na.rm = TRUE)) %>% data.table()  
data.quantile

## GENDER C1Read10 C1Read90 C1Math10 C1Math90 C5Read10 C5Read90 C5Math10  
## 1: Female 26.19 47.160 17.58 38.610 96.38 165.855 66.675  
## 2: Male 25.68 45.914 17.07 39.894 89.26 163.720 68.202  
## C5Math90  
## 1: 129.600  
## 2: 135.294

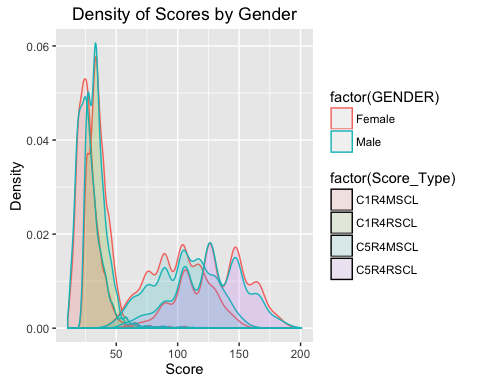
In 1st grade, 10% of male and female students score around 25~26 points and 90% tend to score around 45~47 points in reading. In 5th grade, the variance increases where 10% of males are scoring 89.26 points while 10% of females of scoring 96.38 points, and where 90% of males are scoring 163.720 points while 90% of females of scoring 165.855 points. A similar change in variance is reflected in mathematics scores. However, male students score better by 5th grade, according to the quantiles.

## Question No.7

data.Score <- dataS %>% select(GENDER, Score\_Type, Score) %>% group\_by(GENDER) %>% data.frame()  
head(data.Score, 5)

## GENDER Score\_Type Score  
## 1 Female C1R4RSCL 50.82  
## 2 Male C1R4RSCL 34.72  
## 3 Female C1R4RSCL 38.28  
## 4 Male C1R4RSCL 32.57  
## 5 Male C1R4RSCL 31.98

data.dens <- ggplot(data = data.Score, aes(x=Score, ..density..)) +  
 labs(x="Score", y="Density", title="Density of Scores by Gender") +  
 geom\_density(aes(color=factor(GENDER), fill=factor(Score\_Type)), alpha=0.1)  
data.dens



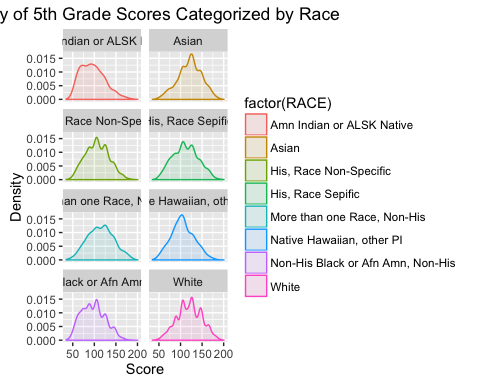
In 1st grade, male and female students were are par with their performances in reading and mathematics. However, by 5th grade males students performed better than female students in reading, while female students performed better than male in mathematics.

## Question No.8

data.Race <- data %>% select(C5R4RSCL, C5R4MSCL, RACE) %>% group\_by(RACE) %>% data.frame()  
data.Race <- gather(data.Race, RACE, Score, C5R4RSCL:C5R4MSCL)  
head(data.Race,5)

## RACE RACE Score  
## 1 White C5R4RSCL 185.22  
## 2 White C5R4RSCL 151.64  
## 3 White C5R4RSCL 150.18  
## 4 White C5R4RSCL 153.59  
## 5 White C5R4RSCL 106.11

data.RaceDens <- ggplot(data = data.Race, aes(x=Score, ..density..)) +  
 labs(x="Score", y="Density", title="Density of 5th Grade Scores Categorized by Race") +  
 geom\_density(aes(color=factor(RACE), fill=factor(RACE)), alpha=0.1) +  
 facet\_wrap(~RACE, ncol=2)  
  
data.RaceDens



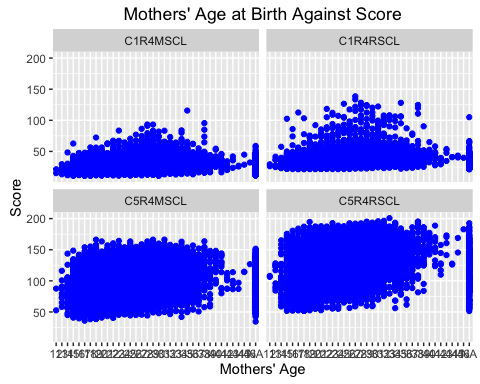
Among the 8 races, Asian 5th graders have a nearly normalized score set in reading and mathematics, compared to American Indian or Alaskan natives or Non-Hispanic Blacks or African Americans that are non-Hispanic 5th graders whose scores are more skewed to the right.

## Question No.9

data.Mothers <- dataS %>% select(P1HMAFB, Score\_Type, Score) %>% group\_by(P1HMAFB) %>% data.frame()  
head(data.Mothers,5)

## P1HMAFB Score\_Type Score  
## 1 24 C1R4RSCL 50.82  
## 2 22 C1R4RSCL 34.72  
## 3 25 C1R4RSCL 38.28  
## 4 32 C1R4RSCL 32.57  
## 5 36 C1R4RSCL 31.98

data.MothersPlot <- ggplot(na.omit(data.Mothers), aes(x=P1HMAFB, y=Score)) +   
 labs(x = "Mothers' Age", y = "Score", title = "Mothers' Age at Birth Against Score") +  
 geom\_point(color="blue") +  
 geom\_smooth(method=lm, color="orange", se = FALSE) +  
 facet\_wrap(~Score\_Type)  
  
data.MothersPlot

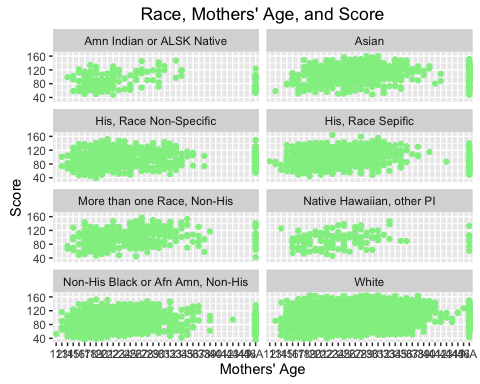


There seems to be an almost normalized distribution of childrens' scores in reading and mathetimatics in the 1st grade depending on the mothers age.

However, by the 5th grade, score levels increase in variance. Particularly in mathematics, 5th graders seems to perform better than in reading.

## Question No.10

data.MothersByRacePlot <- na.omit(data) %>% ggplot(aes(x=P1HMAFB, y=C5R4MSCL)) +  
 geom\_point(color="lightgreen") +  
 labs(x="Mothers' Age", y="Score", title = "Race, Mothers' Age, and Score") +  
 geom\_smooth(method=lm, color="black", se=FALSE) +  
 facet\_wrap(~RACE, ncol = 2)  
  
data.MothersByRacePlot



Among the 8 races included in the Harvard study, White mothers have the greatest variance in terms of age and their child's mathematics score in 5th grade. It seems almost as likely for a mother in her early 20s to give birth to a child who would score well in comparison to a mother in her late 30s.

However, American Indian or Alaskan Native mothers along with Native Hawaiian and other mothers tend to have their first child when they are younger, less than 30 years. According to the scatter plots, these children tend to perform less well compared to children whose mother first bore a child at 30 years or older.