Grade: A-

EDUC 423A/SOC 302A: Assignment 1

caity McGinley

1/28/2021

# Honor Code Statement

We strongly encourage students to form study groups and students may discuss and work on assignments in groups. We expect that each student understands their own submission. As such, students must write their submissions independently and clearly disclose the names of all other students who were part of their study group. Additionally, lifting code or solutions directly from the internet (e.g., Google, GitHub, Stack Overflow) is a violation of the [Stanford Honor Code](https://communitystandards.stanford.edu/policies-and-guidance/honor-code). We take academic honesty and Honor Code violations extremely seriously and expect the same of students. If you have questions about what may or may not constitute an Honor Code violation, please reach out the teaching team.

**Email:** [caitym26@stanford.edu](mailto:caitym26@stanford.edu) **SUID:** 06211809 **Study Group:** NA I acknowledge and agree to abide by the Honor Code.

**Signed:** Caity McGinley

setwd(“C:/Users/cmcgi/Downloads/Soc 302A\_Lab1”)

#loading data HSL <- read\_csv(“EDUC423A\_HSL.csv”)

HSLfemale)

*# Include all code required to load packages and import data here.*  
*# The echo=TRUE flag ensures that the code will appear in your submission.*  
**library**(tidyverse)

## -- Attaching packages ----------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.3 v dplyr 1.0.2  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## Warning: package 'dplyr' was built under R version 4.0.3

## Warning: package 'forcats' was built under R version 4.0.3

## -- Conflicts -------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

**library**(ggplot2)  
**library**(RColorBrewer)

## Warning: package 'RColorBrewer' was built under R version 4.0.3

**library**(psych)

## Warning: package 'psych' was built under R version 4.0.3

##   
## Attaching package: 'psych'

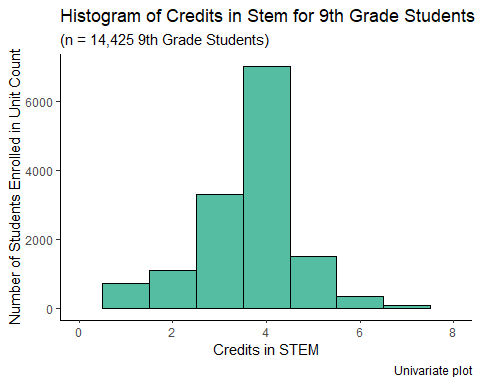
## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

**setwd**("C:/Users/cmcgi/Downloads/Soc 302A\_Lab1")  
  
*#loading data*  
HSL <- **read\_csv**("EDUC423A\_HSL.csv")

# Univariate Plot

*# Include all code required to generate your visualization here.*  
*# An example of a univariate plot would be a histogram.*  
  
*#For coloring purposes and legend readability*  
HSL**$**sex <- **as.character**(HSL**$**female)  
hslnew <- HSL **%>%**  
**mutate**(  
sex = **case\_when**(sex **==** "0" **~** "male",  
sex **==** "1" **~** "female"))  
  
*# ColorBlind friendly palette with black:*  
*# source: http://jfly.iam.u-tokyo.ac.jp/color/*  
cbp2 <- **c**("#CC79A7", "#0072B2", "#F0E442", "#E69F00", "#56B4E9AA",  
 "#D55E00AA", "#009E73AA", "#000000")  
  
  
*#Univariate plot: Distribution of stem credits among students*  
**ggplot**(hslnew, **aes**(x=credits\_stem)) **+**   
 **geom\_histogram**(color = "#000000", fill = "#009E73AA", binwidth = 1 ) **+**  
 **theme\_classic**() **+**  
 **scale\_fill\_manual**(values = cbp2) **+**  
 **xlim**(0, 8) **+**  
 **labs**(title="Histogram of Credits in Stem for 9th Grade Students", subtitle = "(n = 14,425 9th Grade Students)", x="Credits in STEM", y = "Number of Students Enrolled in Unit Count", caption = "Univariate plot")

## Warning: Removed 2 rows containing missing values (geom\_bar).



This is a really clean histogram! It wasn’t required, but I like that you did report some summary stats for the variable you looked at, too.

*#Exploring Data*   
**mean**(hslnew**$**credits\_stem)

## [1] 3.625442

**median**(hslnew**$**credits\_stem)

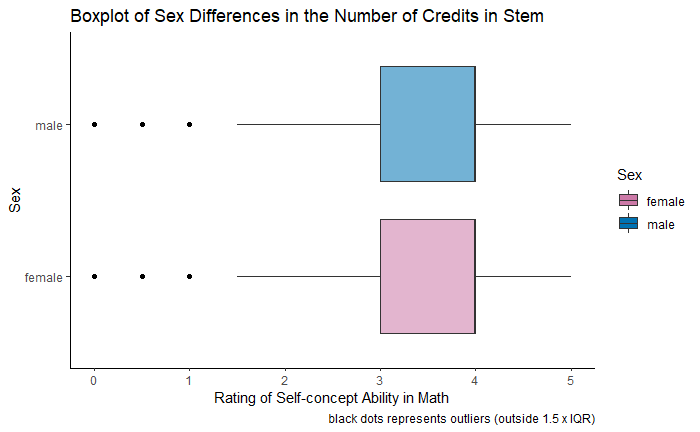
## [1] 4

**skew**(hslnew**$**credits\_stem)

## [1] -0.6645387

# Bivariate Plot

*# Include all code required to generate your visualization here.*  
  
*#Bivariate plot: categorical sex vs continuous Self-concept ability in math variable*  
**ggplot**(hslnew, **aes** (x=credits\_stem,y=sex, fill= sex, alpha =.1)) **+**  
 **geom\_boxplot**(outlier.color = "#000000") **+**  
 **scale\_alpha**(guide = 'none') **+**  
 **labs**(fill = "Sex") **+**  
 **labs**(title="Boxplot of Sex Differences in the Number of Credits in Stem", x="Rating of Self-concept Ability in Math", y = "Sex", caption = "black dots represents outliers (outside 1.5 x IQR)") **+**  
 **theme\_classic**() **+**  
 **xlim**(0, 5) **+**   
 **scale\_fill\_manual**(values = cbp2)

## Warning: Removed 589 rows containing non-finite values (stat\_boxplot).

Very clear boxplot, as well. I suspect you could get pushback for using such normative color scheme, but everything is really clear - especially that you were explicit about the convention used to classify points as outliers.

# Comparative Bivariate Plot

*# Include all code required to generate your visualization here.*  
  
*# Filtering by sex to find bivariate relationships by sex*  
 male <- hslnew **%>%**  
 **filter** (sex **==** "male"   
 )  
female <- hslnew **%>%**  
 **filter** (sex **==** "female"   
 )  
*#Running linear regressions*  
lin\_male <- **lm**(credits\_stem **~** self\_concept\_ability\_math, data=male, na.rm=TRUE)

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'na.rm' will be disregarded

Regressions, by default, do rowwise deletion of any observations that contain missing data in one of the variables in the regression. They are not at all robust to missing data - which is why imputation is such an important technique!

**print**(lin\_male)

##   
## Call:  
## lm(formula = credits\_stem ~ self\_concept\_ability\_math, data = male,   
## na.rm = TRUE)  
##   
## Coefficients:  
## (Intercept) self\_concept\_ability\_math   
## 4.2660 -0.3485

**summary**(lin\_male)

##   
## Call:  
## lm(formula = credits\_stem ~ self\_concept\_ability\_math, data = male,   
## na.rm = TRUE)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.9175 -0.5689 0.2568 0.5825 5.1281   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.26600 0.04463 95.58 <2e-16 \*\*\*  
## self\_concept\_ability\_math -0.34853 0.02149 -16.22 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.196 on 7159 degrees of freedom  
## Multiple R-squared: 0.03544, Adjusted R-squared: 0.03531   
## F-statistic: 263.1 on 1 and 7159 DF, p-value: < 2.2e-16

lin\_female <- **lm**(credits\_stem **~** self\_concept\_ability\_math, data=female, na.rm=TRUE)

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'na.rm' will be disregarded

**print**(lin\_female)

##   
## Call:  
## lm(formula = credits\_stem ~ self\_concept\_ability\_math, data = female,   
## na.rm = TRUE)  
##   
## Coefficients:  
## (Intercept) self\_concept\_ability\_math   
## 4.2960 -0.2961

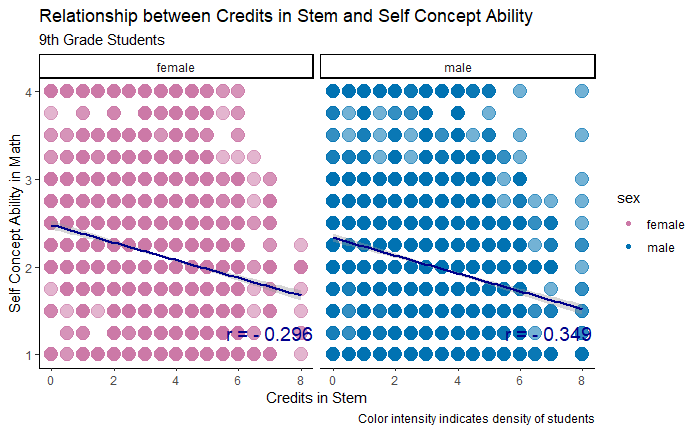
**summary**(lin\_female)

##   
## Call:  
## lm(formula = credits\_stem ~ self\_concept\_ability\_math, data = female,   
## na.rm = TRUE)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.9998 -0.5557 0.2684 0.5002 4.3703   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.29596 0.04387 97.92 <2e-16 \*\*\*  
## self\_concept\_ability\_math -0.29612 0.01984 -14.93 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.116 on 7262 degrees of freedom  
## Multiple R-squared: 0.02977, Adjusted R-squared: 0.02964   
## F-statistic: 222.8 on 1 and 7262 DF, p-value: < 2.2e-16

*# Creating Data Frame to Hold r values and positions on soon to be made graph*  
dat\_text <- **data.frame**(  
 label = **c**("r = - 0.296", "r = - 0.349 "),  
 sex = **c**("female", "male"),  
 x = **c**(7, 7),  
 y = **c**(1.25, 1.25)  
)  
Cool touch!!  
*#Creating Labels*  
comp.labels <- **labs**(x = "Credits in Stem", y = "Self Concept Ability in Math",  
 title = "Relationship between Credits in Stem and Self Concept Ability",  
 caption = "Color intensity indicates density of students")  
*#Creating Graph*  
comp.bivariate <- **ggplot**(hslnew, **aes**(x = credits\_stem, y = self\_concept\_ability\_math)) **+**  
 **geom\_point**(**aes**(color = sex, position = "dodge", alpha=10, size=1)) **+**  
 **scale\_alpha**(guide = 'none') **+**  
 **scale\_size**(guide = 'none') **+**  
 **scale\_color\_manual**(values = cbp2) **+**  
 **geom\_smooth**(method = "lm", color = "darkblue") **+**  
 comp.labels **+** **labs**(subtitle = "9th Grade Students") **+**  
 **facet\_wrap**(**~** sex) **+**  
 **theme\_classic**() **+**   
 **ylim** (1,4) **+**  
 **xlim**(0,8)

## Warning: Ignoring unknown aesthetics: position

comp.bivariate **+** **geom\_text**(  
 data = dat\_text,  
 color = "darkblue",   
 size = 5,  
 mapping = **aes**(x = x, y = y, label = label)  
)

## `geom\_smooth()` using formula 'y ~ x'

I generally like this plot, too - I can see why you switched to the hexbin scatter, as well. I’m a little unclear why you ran the regressions (not that it was a bad thing). I generally like this plot - though I think you could have got a little more mileage by scaling the size of the dots with density instead of the alpha. Additionally, position=’dodge’ doesn’t do anything for scatter plots. I like the creativity you showed here and there’s a lot of good information in it!

# Discussion

*Using your visualizations, address the research question. Be sure to discuss the design decisions you made in creating your visualizations and how they impact the appropriateness of your visualizations for answering the research question. Please refer to the evaluation criteria in the assignment description on Canvas.*

*You have examined a suitable: univariate plot a bivariate plot and a comparative bivariate plot to answer the main research question You have convincingly argued for your visualizations. Type graph: why is it suitable? Colors Size Symbols Scales Labels*

***Design and Code***

***Overall***

I chose a color-blind friendly palette for all of my graphs, except one. I think universal design is important and accommodating for differences is extremely important. For my own graph where I did not incorporate the color-blind friendly palette, I made sure the default colors were not red and green. For my own research question, I maintained the color scheme from my previous boxplot, but added new colors from the brewer palette for my fill in order to have some contrast. I alpha-ed the color values for the comparative bivariate graph as well in order to give more depth to the graph.

Additionally, I provided appropriate titles to each graph, accompanied by subtitles and legends with captions as necessary. Additionally, I plotted labels for the r correlations for the comparative bivariate graphs for easy interpretation of the relationship between the two variables. Your titles were really clear and informative, which I thought was a great touch and nod toward universal design.  
For my own research question, I maintained the color scheme, but added new colors from the brewer palette for my fill in order to have some contrast.

I manipulated both axis to reflect the range of the continuous variables. This was to eliminate unnecessary spacing.

***Univariate Plot:***

My univariate plot is a histogram of the data. I chose a histogram to get a better idea of the frequency distribution of students taking credits in STEM. Since this is a univariate plot, I did not look at the sex variable. The histogram shows that most students take around 4 credits (aka classes) in STEM. The histogram also shows me the ~ number of students taking each number of credits as well as the skewness of the data set. The credits in stem has a slight negative skew of -0.6645387, with a mean of 3.6 and a median of 4.

I decided to keep the bins = 1 because the continuous scale wasn’t very large. Making it larger lost integrity of the measure and making it smaller made it difficult to see (.1 was too small).

Since I was only plotting one variable, I wanted to use a color that would not be used in my bivariate analysis–that’s why I chose color blind friendly green!

***Bivariate Plot:***

For my bivariate plot, I wanted to see the dispersion of the data so I decided to use a box plot to evaluate sex differences in the rating of self-concept ability. I find box plots useful when looking at both categorical variables, such as sex, and continuous variables like the self-concept ability in math. Box plots show the median values, the IQR (upper and lower quartile), the max and min, as well as relative skewness. As you can see, males and females generally have the same median value for their self-reported concept ability in math. However, the male group has a slight negative skew, meaning that the mean and median will be less than the mode. On the contrary, the female group has a slight positive skew, meaning the mean and median will be greater than the mode. Box plots also visualize outliers. Males have 3 outliers compared to females with 2.

**Color -** The difference in color signifies the two different groups, males and females. Within the color-blind friendly palette, I decided to go with a version of blue and pink. While I do believe that these colors have been grossly “genderfied” they do clearly split the grouping of male and female in a way that is easily interpretable to a general audience. Within the box plot, I manipulated the alpha for aesthetic purposes. Great to see you acknowledge this, but you are totally correct - it made the plot immediately readable and provided clear contrast.

I chose a color-blind friendly palette for all of my graphs. The difference in color signifies the two different groups, males and females. Within the color-blind friendly palette, I decided to go with a version of blue and pink. While I do believe that these colors have been grossly “genderfied” they do clearly split the grouping of male and female in a way that is easily interpretable to a general audience. Within the box plot, I manipulated the alpha for aesthetic purposes, but within the other plots, I did not as it made the points turn a strange color when they overlapped (pink + blue makes purple) and making the colors solid made the graph look cleaner. I think you could have condensed the writing here to avoid copy/pasting the bit about the gendering of your color choices.

***Comparative Bivariate Plot:***

For this graph, I used a bivariate scatter plot to display the relationship between self-concept ability in math and the number of credits in stem for the different sexes. I ran a regression on the response of self-concept ability in math to the number of credits taken in STEM. I was able to plot the label right to next the lm line. Additionally, for the comparative graph, I used facet\_wrap to put the two sexes next to each other for easy viewing purposes.

**Color –** I wanted continuity from my box plot comparing gender so I continued the color scheme, just added alpha to show how that data was distributed amongst the cohort of students. By adding alpha, there is a greater understanding of where the relationship is focalized for most students.

**Size-** I also increased the size so that the geom\_points were easier to see as well as the alpha.

**CONCLUSIONS**

**So, are motivational beliefs about math are related to STEM course-taking and if there is a difference between male and female students?**

**TLDR:** Females generally rank themselves higher in self-concept ability than males while taking the same number of credits as males.

This is align with research and meta-analysis that tends to find that girls outperform boys in most areas of education. Girls average higher grades (Buchmann et al. 2008), high school graduation rates (Snyder and Dillow 2012), and college enrollment rates (Buchmann and DiPrete 2006). But this does not translate to the workplace or even the actions or perceptions of parents and teachers. Research consistently finds that gender status beliefs—or cultural expectations about traits girls and boys possess—associate boys with increased competency and social esteem (Thébaud and Charles 2018). From kindergarten through college, students perceive boys as more intelligent (Bian, Leslie, and Cimpian 2017). Parents often perceive their sons as having higher IQs than their daughters (Furnham, Reeves, and Budhani 2002), and teachers are more likely to identify boys as gifted (Petersen 2013). Nice touch tying this back to theory.

**References**

Bian, Lin, Sarah-Jane Leslie, and Andrei Cimpian. 2017. “Gender Stereotypes about Intellectual Ability Emerge Early and Influence Children’s Interests.” Science 355(6323):389–91.

Buchmann, Claudia, and Thomas A. DiPrete. 2006. “The Growing Female Advantage in College Completion: The Role of Family Background and Academic Achievement.” American Sociological Review 71(4):515–41.

Furnham, Adrian, Emma Reeves, and Salima Budhani. 2002. “Parents Think Their Sons Are Brighter Than Their Daughters: Sex Differences in Parental SelfEstimations and Estimations of Their Children’s Multiple Intelligences.” The Journal of Genetic Psychology 163(1):24–39.

Petersen, Jennifer. 2013. “Gender Differences in Identification of Gifted Youth and in Gifted Program Participation: A Meta-Analysis.” Contemporary Educational Psychology 38(4):342–48.

Snyder, Thomas D., and Sally A. Dillow. 2012. “Digest of Education Statistics 2011.” National Center for Education Statistics. Washington, DC: U.S. Department of Education.

Thébaud, Sarah, and Maria Charles. 2018. “Segregation, Stereotypes, and STEM.” Social Sciences 7(111):1–18. Thorne, Barrie. 1993. Gender Play: Girls and Boys in School. New Brunswick, NJ: Rutgers University Press.

# Your own question

Now explore the other variables in this dataset and develop a question. Answer your question with data visualization.

**RQ: How does race and mother’s education level influence a female student’s self concept ability in math?**

Be careful with the framing of your question - instead of implying “influence” (which itself implies specifical directionality to a causal relationship), ask if the variables are associated with each other.

**Conclusion: Overall Female and Male Comparison:**

In the groups with mother’s that did not have an MA or PhD, females rank themselves higher in math concept ability than males for all races. However, students with mother’s with a PhD or MA rank their self-concept ability the equally, regardless of the student’s sex. Mother’s education can be used as an indicator across race to predict the student’s self-concept ability in math.

***Analysis***

Group One: Mother’s with high school education or less

Female and Male Comparison by Race

Black: The median value for black students looks the same, but the upper quartile for females in larger than that of males, meaning that the max for females is higher than males. However, there’s more outliers for males.

Hispanic: The median value for female Hispanic students is higher than that of Hispanic males with mother’s of the same education background. Additionally, the lower quartile for males is lower than that of females, meaning that the min for females is higher than the males. They both have the same upper outliers, but there is a low outlier for females for the rating of self concept ability in math.

Other: The median value for Other students looks the same, but the lower quartile for males is lower than that of females, meaning that the min for females is higher than the males. They both have the around same number of upper outliers, but there is one more for females for the rating of self concept ability in math.

White: The median value for White female students is higher than that of White males with mother’s of the same education background. Additionally, the lower quartile for males is lower than that of females, meaning that the min for females is higher than the males. There’s more outliers for females all around for both upper and lower measures.

Group Two: Mother’s with an AA or BA/BS

Female and Male Comparison by Race

Black: The median value for black students looks the same, but there are more female outliers on both lower and upper measures.

Hispanic: The median value for Hispanic students looks the same, but the males ratings are skewed.

Other: Females students that identify as Other have a higher median rating of self concept ability compared to their male counterparts.

White: White students have the same median rating, but the upper quartile is higher for females and lower quartile is lower for males. There’s more outliers for male scores.

**Group Three: Mother’s with a MA or PhD**

**Female and Male Comparison by Race**

Black: Black students have the same median rating, but the upper quartile is higher for females and lower quartile is lower for males. There’s more outliers for male scores.

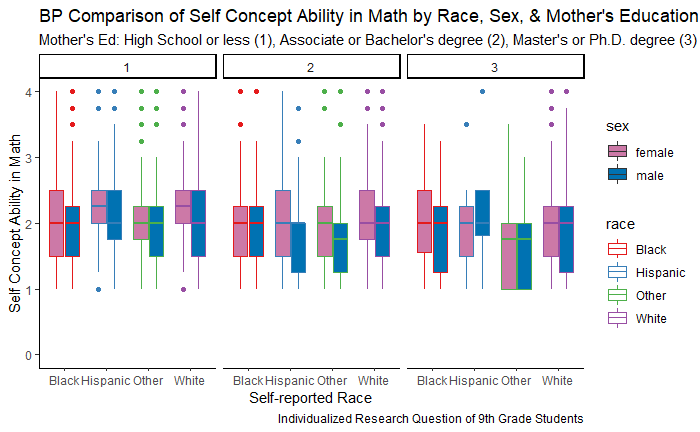
Hispanic: Hispanic students have the same median rating, but the upper quartile is higher for females and lower quartile is lower for females. There’s the same number of outliers for both sexes.

Other: Hispanic students have the same median rating, but females have a higher max score than males.

White: White students have the same median rating, but there is larger variation in male scores. There is higher number of female outliers, hinting that more females rank themselves 2-3 than 4.

*# Include all code required to generate your visualization here.*  
   
*#race. This variable reflects the student's self-identified race if available. If missing, student race is defined by school or parental information. There are five values: White (1), Hispanic (2), Black (3), Asian (4), and (5) Other.*  
  
*#Recording race as character instead of numeric*   
hslnew**$**race <- **as.character**(hslnew**$**race)  
**head**(hslnew**$**race)

## [1] "1" "1" "3" "1" "1" "1"

researchq<- hslnew **%>%**  
**mutate**(  
race = **case\_when**(race **==** "1" **~** "White",  
race **==** "2" **~** "Hispanic",   
race **==** "3" **~** "Black",   
race **==** "4" **~** "Asian",   
race **==** "5" **~** "Other"))  
  
rq.labels <- **labs**(x = "Self-reported Race", y = "Self Concept Ability in Math",  
 title = "Box Plot Comparison of Self Concept Ability in Math by Race, Sex, & Mother's Education ",  
 subtitle = "High School or less (1), Associate or Bachelor's degree (2), Master's or Ph.D. degree (3) ")  
  
**ggplot**(researchq) **+**  
 **geom\_boxplot**(**aes**(x = race, y = self\_concept\_ability\_math, fill = sex, color = race)) **+**  
 **scale\_fill\_manual**(values = cbp2) **+**  
 **scale\_color\_brewer**(palette = "Set1") **+**  
 rq.labels **+** **labs**(caption = "Individualized Research Question") **+**  
 **facet\_wrap**(**~** prnt\_female\_edu) **+**  
 **theme\_classic**() **+**  
 **ylim** (0,4)

Woah, this plot is incredibly dense, but overall really readable. If you turned the alpha way up on the blue/pink fills, it may have allowed the colors you used for your race variable to be more clearly visible. One thing that may be of interest is the how many students fall into each of these categories. Analyzing this many categorical variables simultaneously forces you to consider the intersection of a lot of different classes - and I think your plot summarizes all of that information quite well.