Problem Set 2

Jacob Hood

March 20th, 2021

Prerequisite

Load multiple packages to your environment using the following code:

```
knitr::opts_chunk$set(echo = TRUE, results = "asis")
library(pacman)
p_load(tidyverse, foreign, corrplot, stargazer, coefplot, effects, kableExtra, scales)
```

Part 1: The Replication Project

1. Read the paper closely and respond to the following questions:

(a) What are the authors' research questions?

RQ: To what extent has the decrease in the racial pay gap over the past four decades been influenced by the different economic sources and trajectories of men and women?

(b) What is the gap in the literature that the authors aim to fill? How does their analysis advance the literature?

The available literature on racial disparities in earnings largely only focuses on men, obscuring the interaction between race and gender and the complex patterns they produce. The authors are focused on addressing this gap through a longitudinal study of gender differences in racial earnings disparities from 1970 to 2010.They propose a new theoretical framework to establish reasonable expectations for wage disparities between different races/genders. Importantly, the authors try to create a framework that can account for decreases/change in the racial pay gap over time.

(c) What is the population that they are making inferences about? Be specific and make sure to identify the geographical region that they are focusing on, the time period, demographic characteristics, and so on.

The population is all black and white workers (excluding Hispanic/other races) in the United States for every ten years from 1970 to 2010, ages 25 to 59.

(d) Do they have data on all individuals in the population? If they don't, how do they solve this?

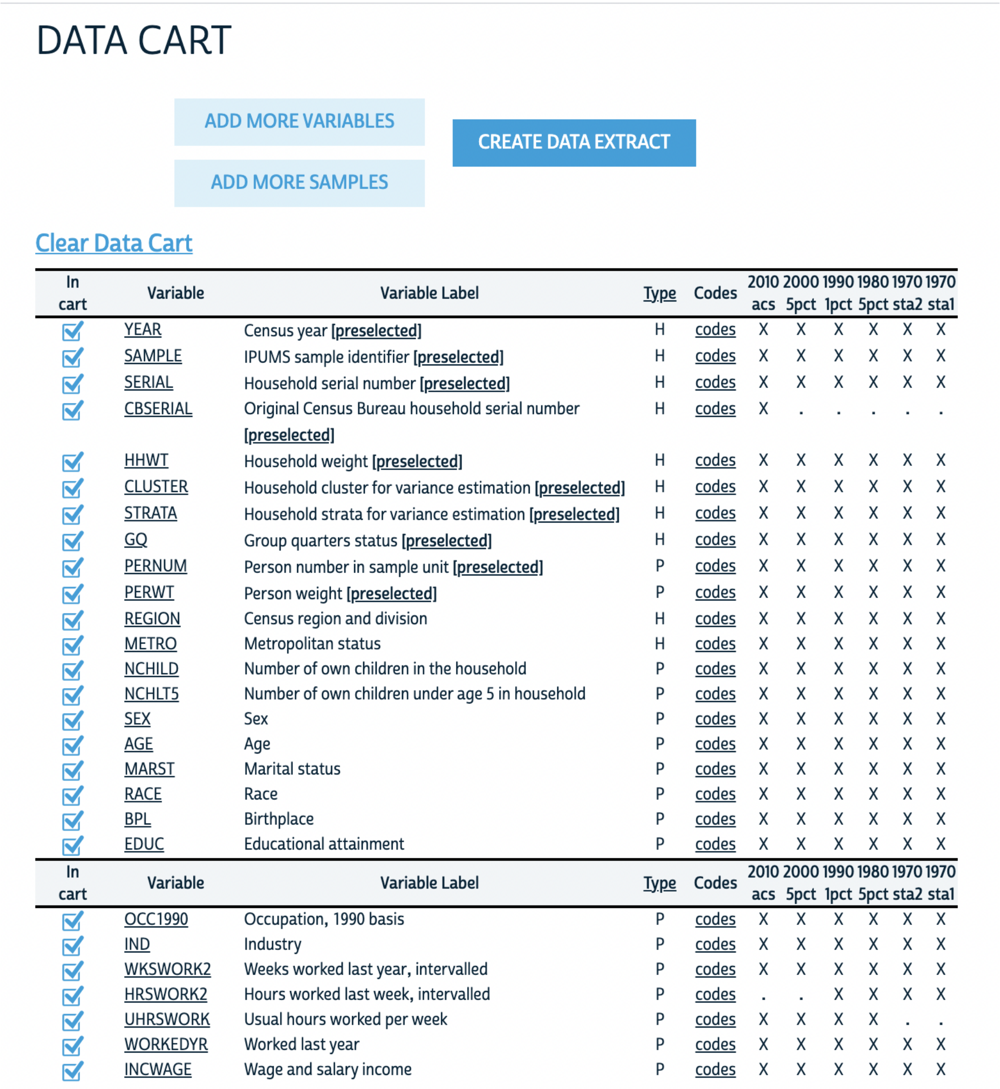
No they do not have data on all individuals in the population. They have to use sample data collected by national surveys. Specifically they use a 5% census sample for 1980 and 2000, a 1% census sample for 1970 and 1990, and American Community Survey data for 2010.

(e) Did the authors collect the data themselves? If they did, describe their sampling procedure. If they didn't, identify and describe the data source and discuss the sampling procedure that was used to collect this data.

No they did not. The authors used secondary data obtained through IPUMS.They use different sources to compile their data set: 1% and 5% random US census samples as well as data from the ACS.

2. As you read the text, make a list of all variables or characteristics of the population that the authors mention throughout the paper (e.g., gender, age, wages, occupation, . . .). Submit a list with all the variables and characteristics that you have identified in the text. - gender (sex)
- race (excluding all but blacks and whites)
- age (25-59)
- level of education
- potential work experience and it squared (age - yrs of schooling -6)^6
- weekly working hours (and weekly working hours logged)
- weekly wage (and weekly wage logged)
- marital status
- nativity status
- number of children
- presence of a child under 5 (=1)
- sector, working in public service (=1)
- region
- metropolitan area (=1)
- occupation (1990 codes)

3. Based on your answers to prior questions, select the samples and variables that you think you will need to replicate the paper in your IPUMS account. Submit a screenshot of the page where you can see the samples and variables that you have selected.



In	Variable	Variable Label	Type	Codes	2010	2000	1990	1980	1970	1970
cart					acs	5pct	1pct	5pct	sta2	sta1
☑	YEAR	Census year [preselected]	H	codes	X	X	X	X	X	X
☑	SAMPLE	IPUMS sample identifier [preselected]	H	codes	X	X	X	X	X	X
☑	SERIAL	Household serial number [preselected]	H	codes	X	X	X	X	X	X
☑	CBSERIAL	Original Census Bureau household serial number [preselected]	H	codes	X	.	.	.	.	.
☑	HHWT	Household weight [preselected]	H	codes	X	X	X	X	X	X
☑	CLUSTER	Household cluster for variance estimation [preselected]	H	codes	X	X	X	X	X	X
☑	STRATA	Household strata for variance estimation [preselected]	H	codes	X	X	X	X	X	X
☑	GQ	Group quarters status [preselected]	H	codes	X	X	X	X	X	X
☑	PERNUM	Person number in sample unit [preselected]	P	codes	X	X	X	X	X	X
☑	PERWT	Person weight [preselected]	P	codes	X	X	X	X	X	X
☑	REGION	Census region and division	H	codes	X	X	X	X	X	X
☑	METRO	Metropolitan status	H	codes	X	X	X	X	X	X
☑	NCHILD	Number of own children in the household	P	codes	X	X	X	X	X	X
☑	NCHLT5	Number of own children under age 5 in household	P	codes	X	X	X	X	X	X
☑	SEX	Sex	P	codes	X	X	X	X	X	X
☑	AGE	Age	P	codes	X	X	X	X	X	X
☑	MARST	Marital status	P	codes	X	X	X	X	X	X
☑	RACE	Race	P	codes	X	X	X	X	X	X
☑	BPL	Birthplace	P	codes	X	X	X	X	X	X
☑	EDUC	Educational attainment	P	codes	X	X	X	X	X	X
In	Variable	Variable Label	Type	Codes	2010	2000	1990	1980	1970	1970
cart					acs	5pct	1pct	5pct	sta2	sta1
☑	OCC1990	Occupation, 1990 basis	P	codes	X	X	X	X	X	X
☑	IND	Industry	P	codes	X	X	X	X	X	X
☑	WKSWORK2	Weeks worked last year, intervalled	P	codes	X	X	X	X	X	X
☑	HRSWORK2	Hours worked last week, intervalled	P	codes	.	.	X	X	X	X
☑	UHRSWORK	Usual hours worked per week	P	codes	X	X	X	X	.	.
☑	WORKEDYR	Worked last year	P	codes	X	X	X	X	X	X
☑	INCWAGE	Wage and salary income	P	codes	X	X	X	X	X	X

Part 2: Regression

Import the dataset `sat_math.dta` to your R environment and examine the effect of IQ and other variables on SAT math score. *Hint:* use `read.dta()`

```r
# loading data into environment
sat_data <- read.dta("sat_math.dta")
```

Variable Name	Variable Detail
`sat_math`	SAT Math Score
`female`	The Female Dummy (Male = 0)
`black/other`	Two Racial Dummies (White as the Reference Group)
`meduy`	Mother's Years of Schooling
`feduy`	Father's Years of Schooling
`hours`	Average Weekly Study Hours
`IQ`	IQ Score (0 to 100)

1. **Report descriptive statistics:**

   (a) Create a table that reports descriptive statistics (you should at least report the means) of all the variables grouped by gender

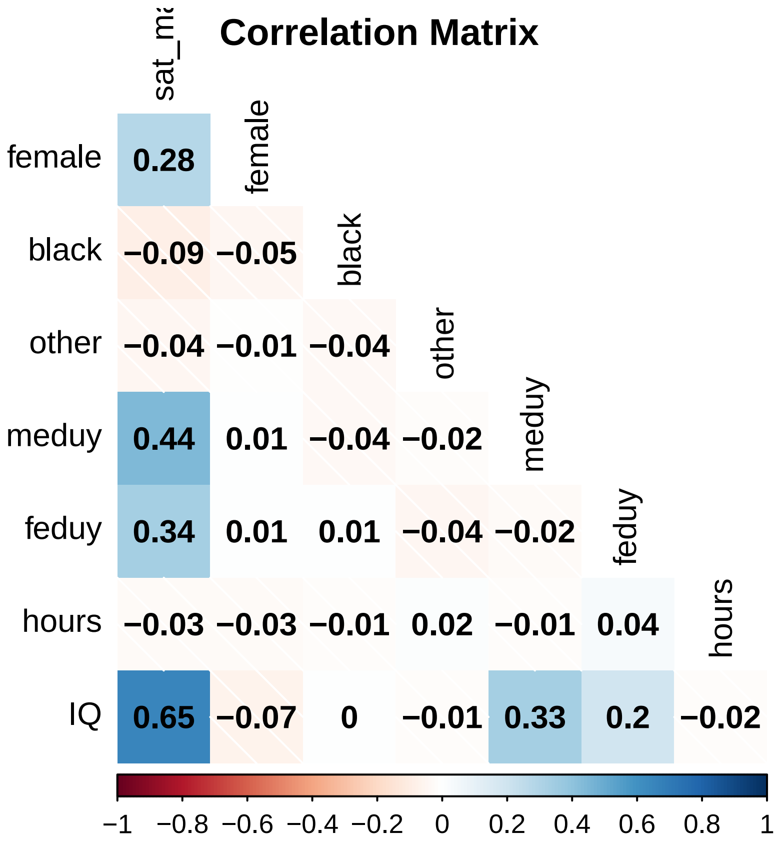
```
# Grouping data by gender (in this case, before recoding, grouping by female
# so male=0 and female=1) and producing descriptive statistics
sat_data <- sat_data %>% group_by(female)
```

```
sat_data %>% summarise(mean_satmath = mean(sat_math),
            mean_meduy = mean(meduy),
            mean_feduy = mean(feduy),
            mean_hours = mean(hours),
            mean_IQ = mean(IQ)) %>% kbl("latex") %>% kable_classic_2(full_width = F, position = "left")
```

female	mean_satmath	mean_meduy	mean_feduy	mean_hours	mean_IQ
0	501.5631	11.97228	11.75050	40.06733	50.53334
1	548.3405	12.02626	11.80808	39.74141	48.75590

   (b) Create a correlation matrix and display it

```
## Set use = "complete.obs" to ignore observations with NAs
M <- cor(sat_data, use = "complete.obs")

# Save the matrix to a dataframe, then use `corrplot` to visualize
corrplot(M,
         method = "shade",
         cl.lim = c(-1,1),
         is.corr = F,
         type = "lower",
         title = "
         Correlation Matrix",
         addCoef.col = "black",
         diag = F,
         tl.col = "black",
         )
```



2. Create scatter plots:

- Besides the key dependent variable (DV) `sat_math`, choose one numeric independent variable (IV) that seems to have a meaningful relation to the DV based on the correlation matrix you created, and then create the following plots:

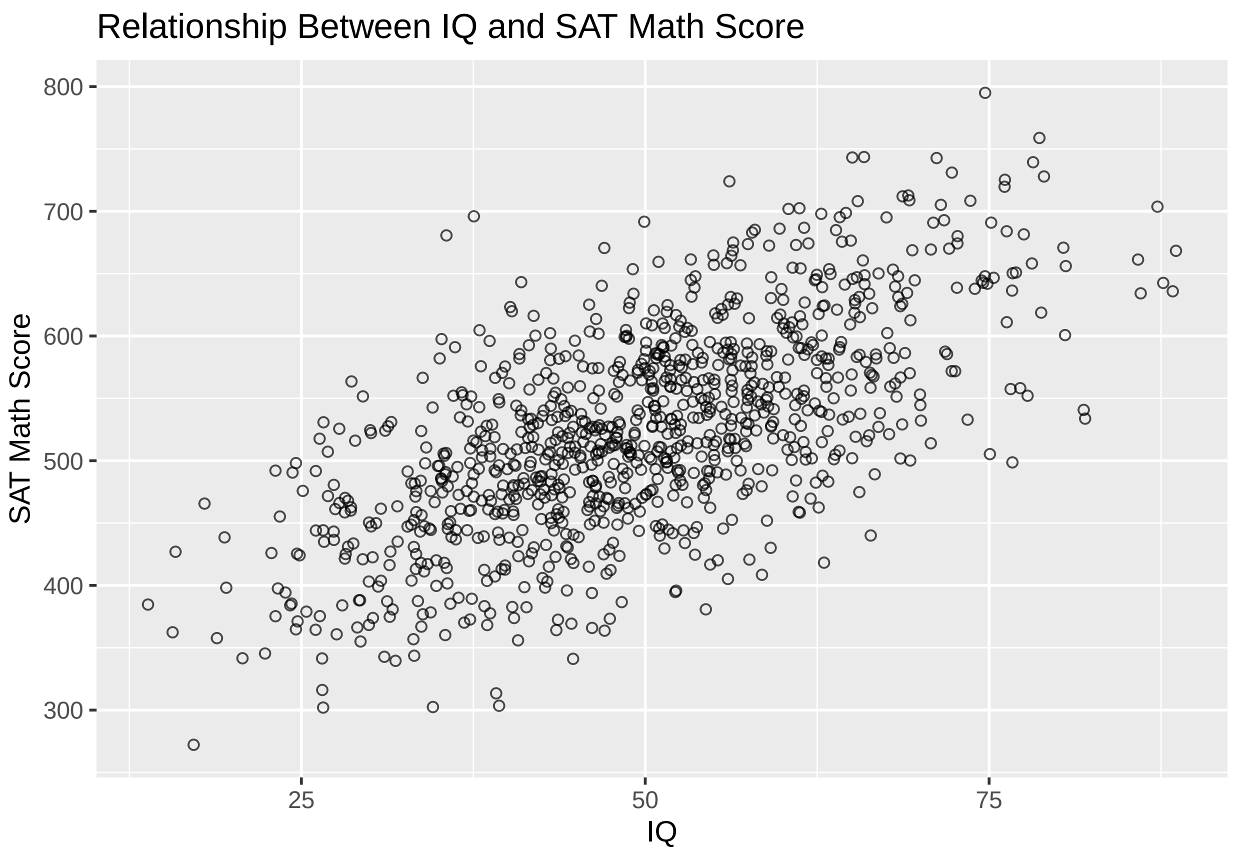
(a) A scatter plot of the DV and IV

(b) A scatter plot of the DV and IV with a fitted linear regression line

(c) A scatter plot of the DV and IV, and each observation is color coded by gender

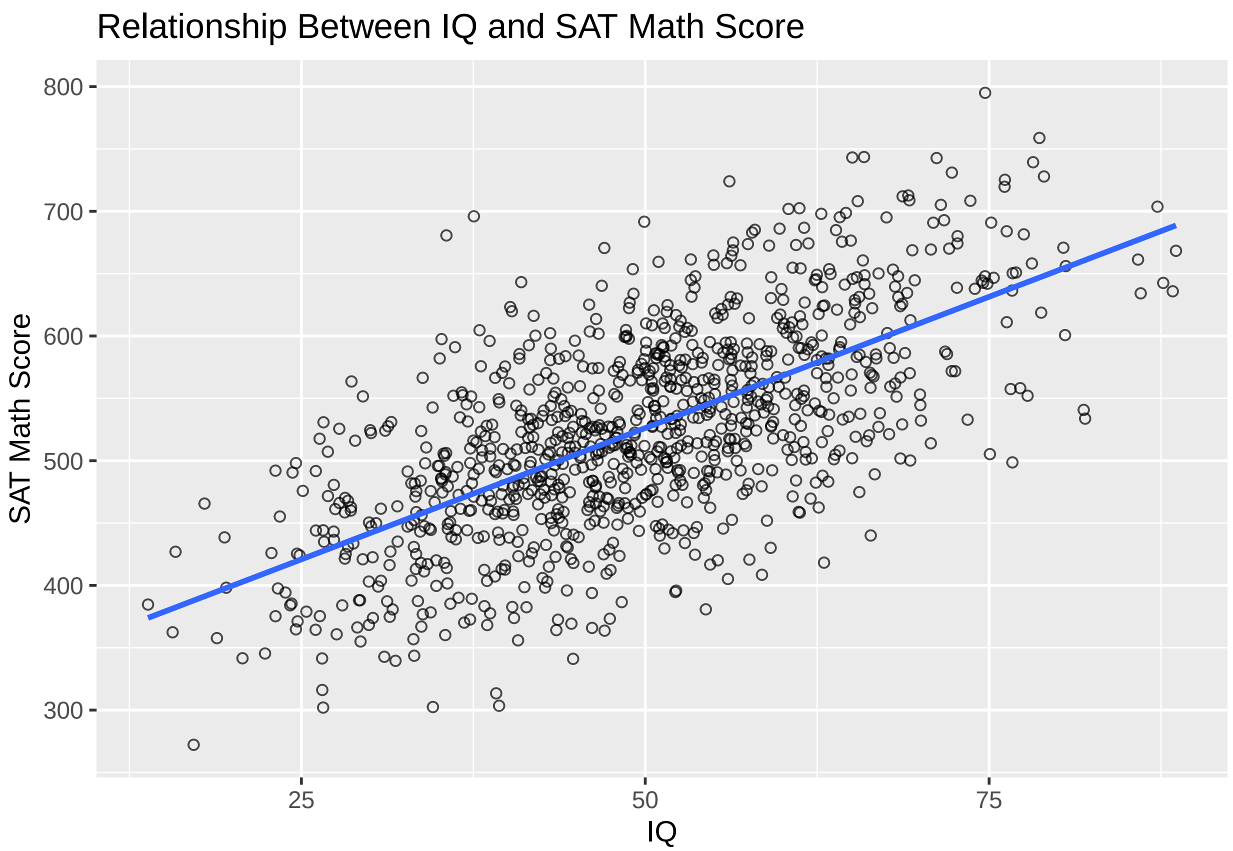
(d) On top of plot (c), fit a linear regression line for each gender group, the lines should also be color coded

```
#   (a) A scatter plot of the DV and IV, IQ
sat_data %>%
  ggplot(aes(x = sat_data$IQ, y = sat_math)) +
  geom_point(shape = 1, alpha = 0.7) +
  labs(title = "Relationship Between IQ and SAT Math Score",
       x = "IQ",
       y = "SAT Math Score")
```

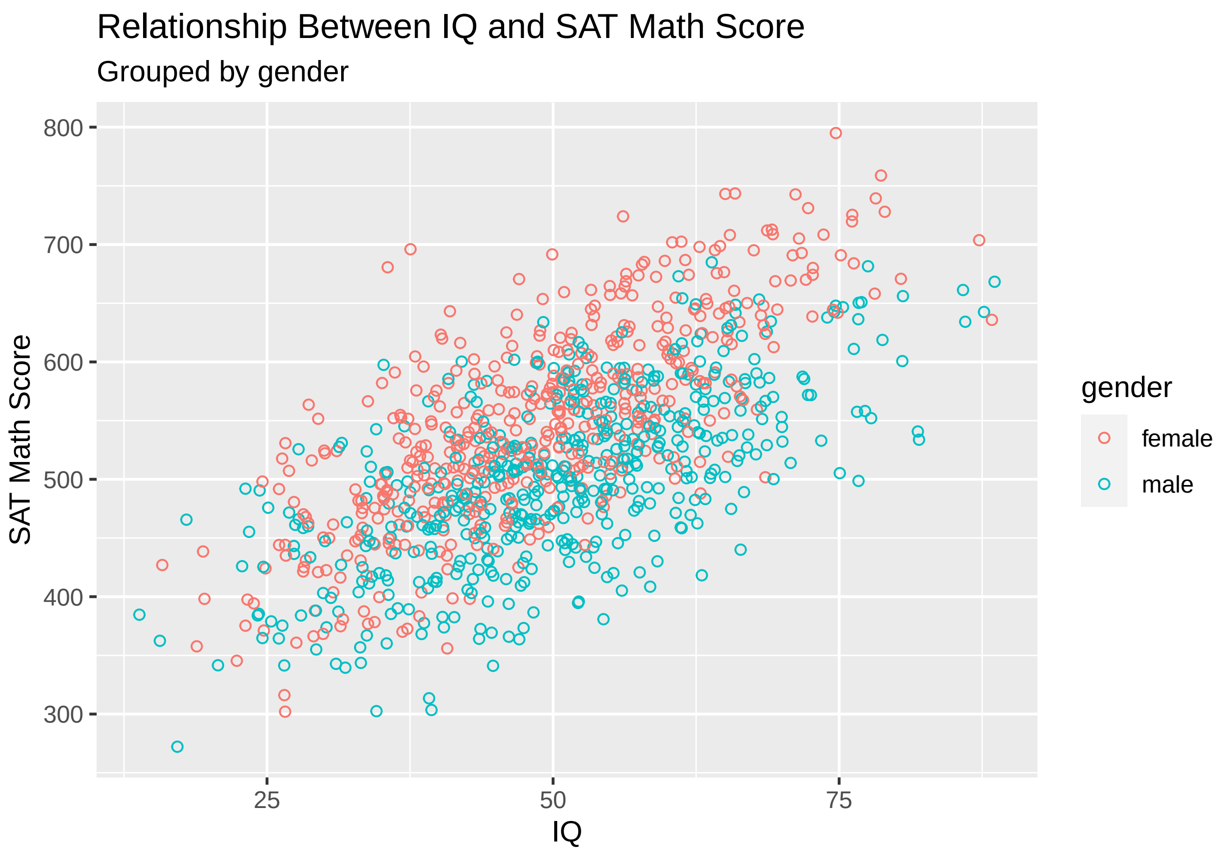


```
#  (b) A scatter plot of the DV and IV with a fitted linear regression line
sat_data %>%
  ggplot(aes(x = IQ, y = sat_math)) +
  geom_point(shape = 1, alpha = 0.7) +
  geom_smooth(method = "lm", se = F) +
  labs(title = "Relationship Between IQ and SAT Math Score",
       x = "IQ",
       y = "SAT Math Score")
```

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

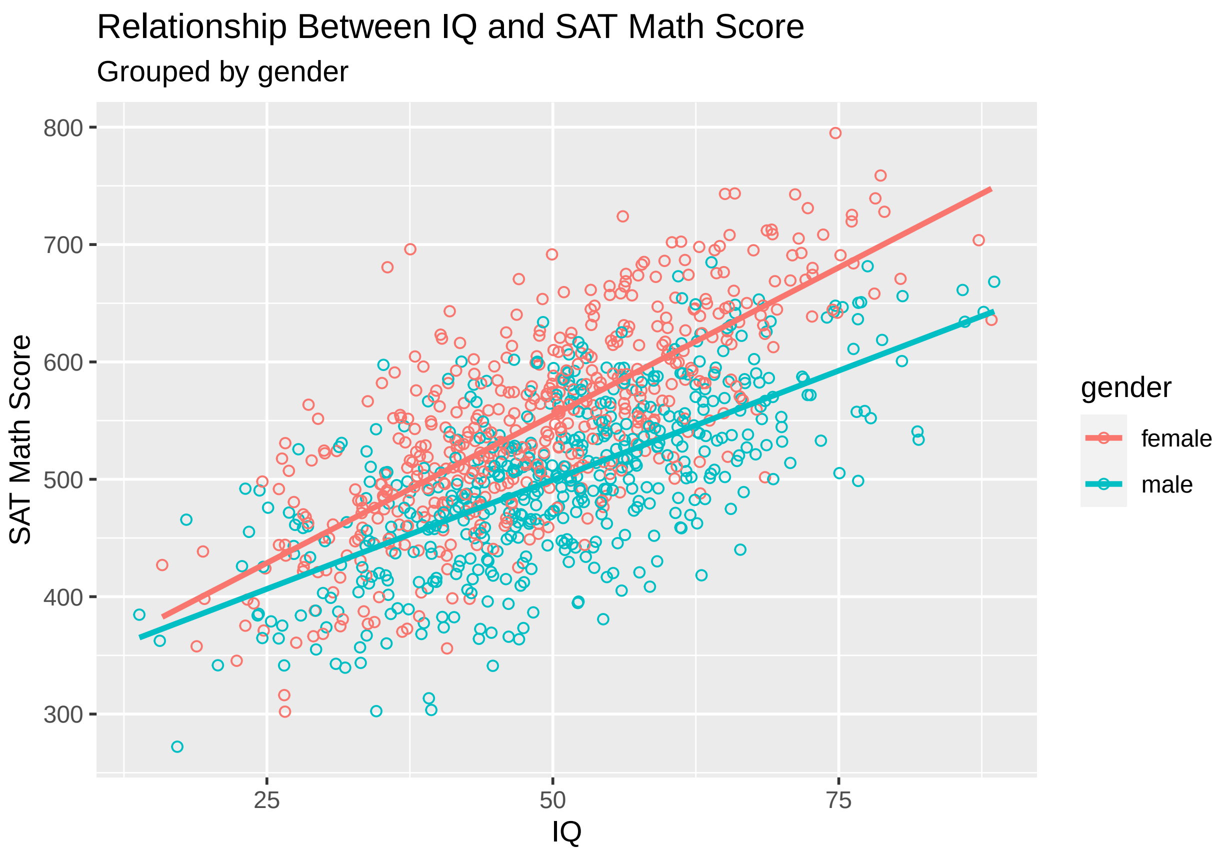


```
# (c) A scatter plot of the DV and IV, and each observation is color coded by gender
# Create new variable 'gender'
sat_data <- sat_data %>%
  mutate(gender = ifelse(female == 1, "female", "male"))

sat_data %>% as_tibble() %>% ggplot(aes(x = sat_data$IQ, y = sat_math, color = gender)) +
  geom_point(shape = 1) +
  labs(title = "Relationship Between IQ and SAT Math Score",
       x = "IQ",
       y = "SAT Math Score",
       subtitle = "Grouped by gender")
```



```
# (d) On top of plot (c), fit a linear regression line for each gender group, the lines should also be
sat_data %>% as_tibble() %>%
  ggplot(aes(x = IQ, y = sat_math, color = gender)) +
  geom_point(shape = 1) +
  geom_smooth(method = "lm", se = F) +
  labs(title = "Relationship Between IQ and SAT Math Score",
       x = "IQ",
       y = "SAT Math Score",
       subtitle = "Grouped by gender")
```

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



3. Additional exploratory data analysis:

(a) What are your preliminary findings/reflections on the data based on the descriptive statistics, the correlation matrix, and the scatter plots?

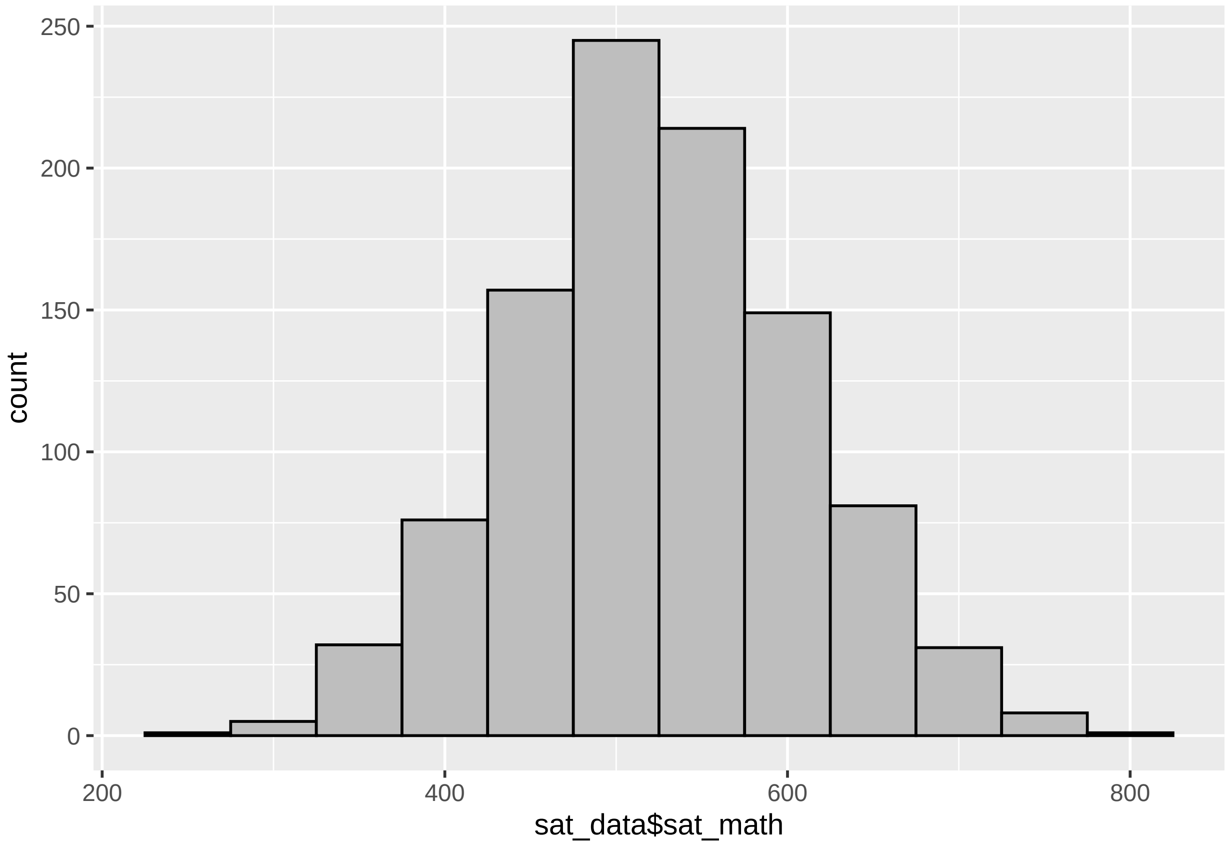
Based on the descriptive statistics, we know that females have a higher average SAT math score than males. On average females score 47 points higher than males. Also, all together students spend an average of 39 hours studying per week and males and females spend roughly the same amount of time studying each week. In terms of IQ, males on average scored 2 points higher. We also know the data is somewhat skewed to the right (median < mean), which means that less people are making higher SAT math scores than if the data were normally distributed. Based on the correlation plot, IQ has the strongest positive correlation with SAT math scores (0.65). Parents' years of schooling is also strongly correlated with higher SAT math scores, more so than with the other explanatory variables. For number of hours studied per week, there is a very minor negative relationship between average hours studied per week and SAT math scores (-0.03).This is confirmed by scatterplots.We can also see interaction between IQ and gender, and parent's education and gender.For instance, we can see that the effect of IQ seems to be more significant for females than males.

(b) What other exploratory data analysis will be useful for you to better understand the data before modeling? Please implement some additional exploratory data analysis and discuss your preliminary findings

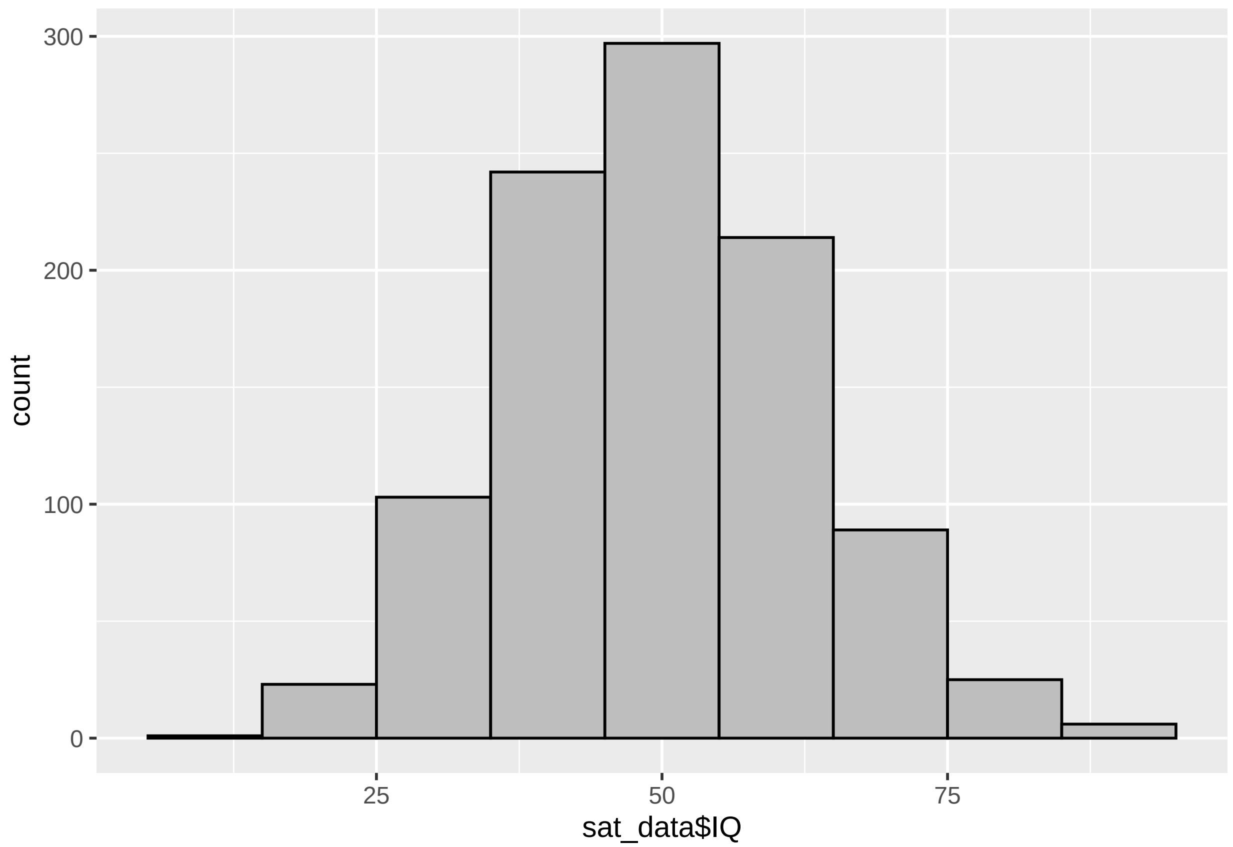
First we can create histograms of each variable to understand their distributions. Other exploratory data analyses before modeling include plotting the relationship between SAT

math scores and our other predictor variables to visually understand their relationships.The data can also be grouped by race in order to look at racial differences in SAT math scores at different levels of predictors.

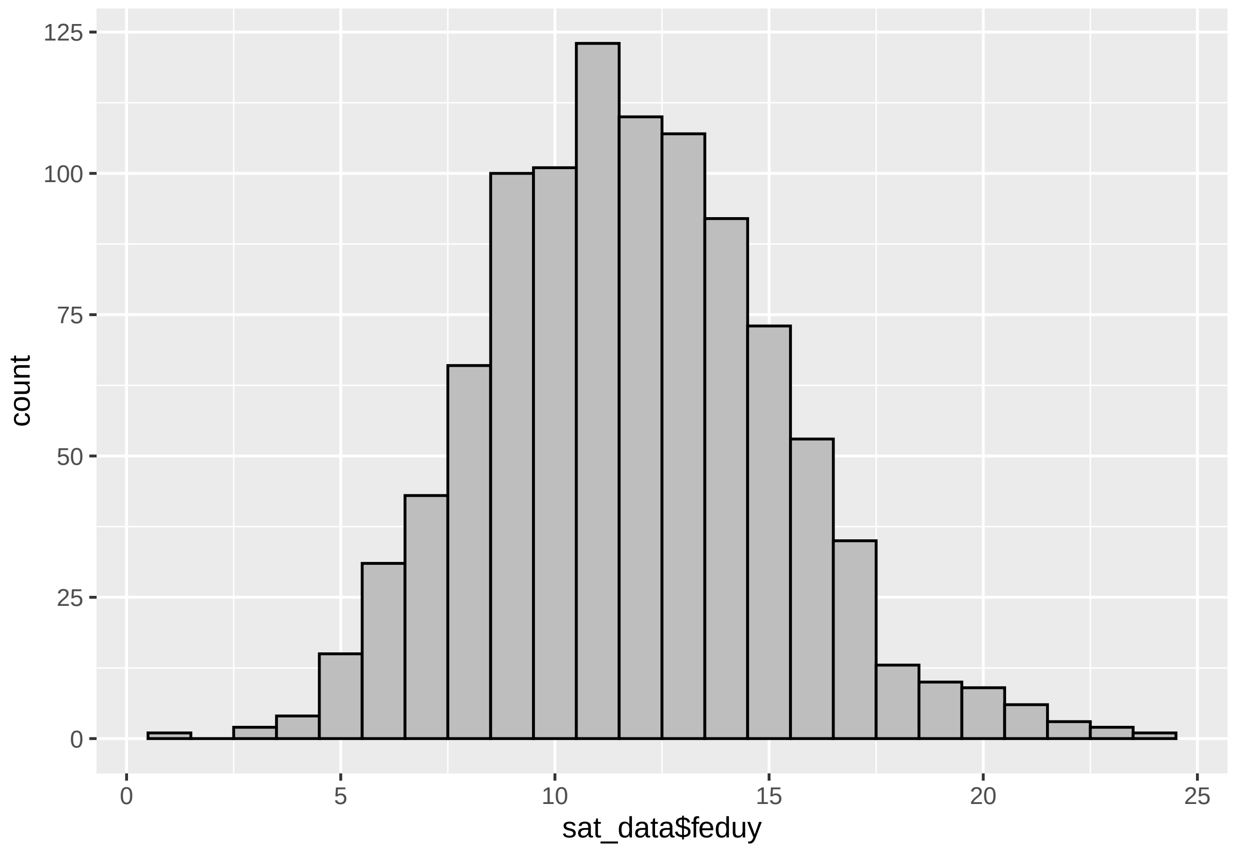
```
# Creating histograms of variables

# Plotting dependent variable
sat_data %>% as_tibble() %>%
  ggplot(aes(x = sat_data$sat_math)) +
  geom_histogram(binwidth = 50, fill = "grey", color = "black")
```



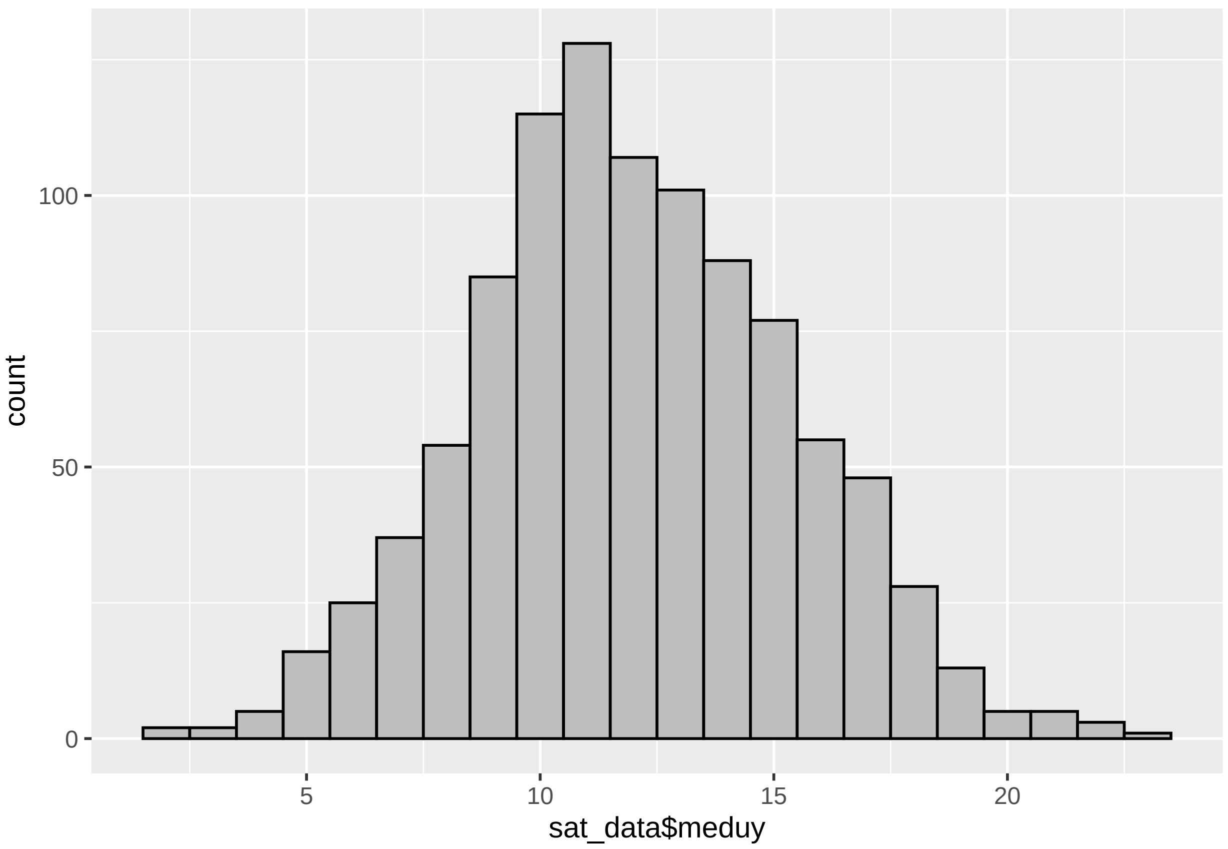
```
# Plotting independent variables
sat_data %>% as_tibble() %>%
  ggplot(aes(x = sat_data$IQ)) +
  geom_histogram(binwidth = 10, fill = "grey", color = "black")
```



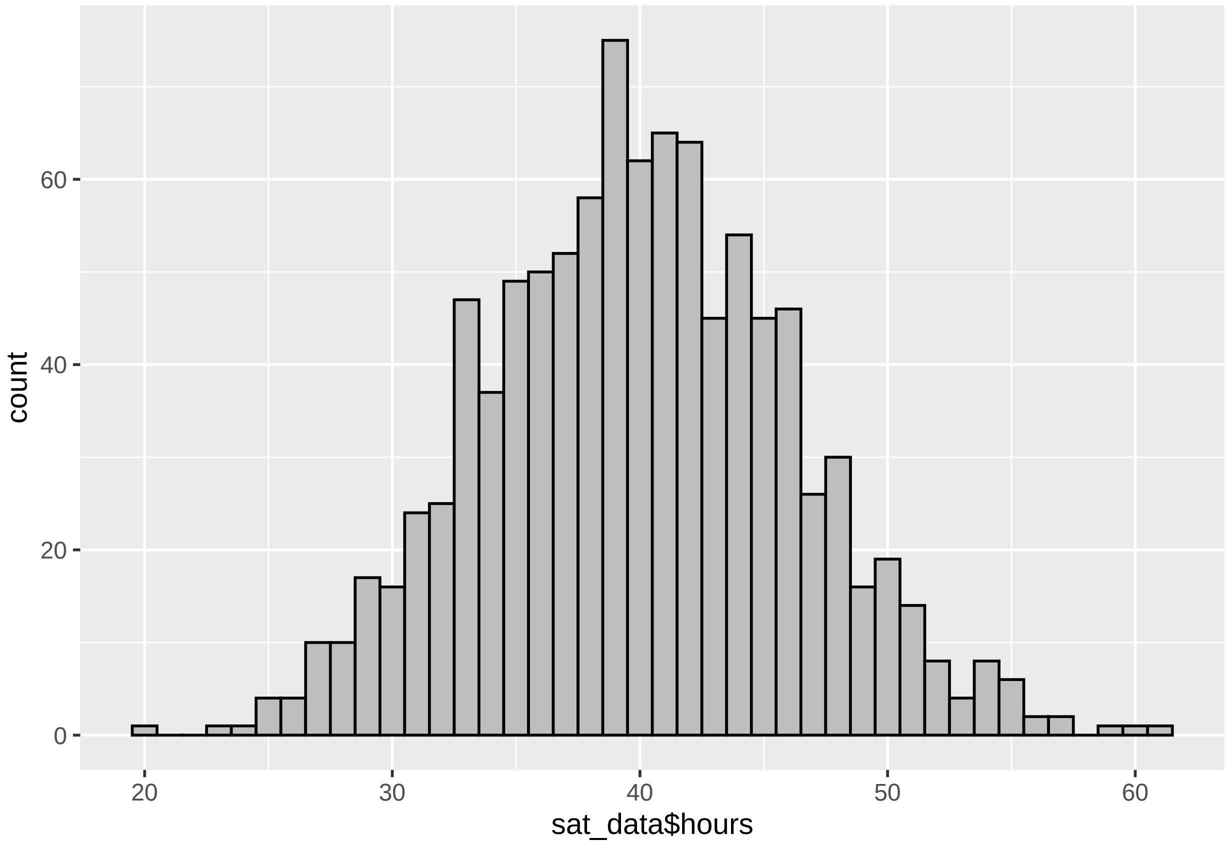
```
sat_data %>% as_tibble() %>%
  ggplot(aes(x = sat_data$feduy)) +
  geom_histogram(binwidth = 1, fill = "grey", color = "black")
```



```
sat_data %>% as_tibble() %>%
  ggplot(aes(x = sat_data$meduy)) +
  geom_histogram(binwidth = 1, fill = "grey", color = "black")
```

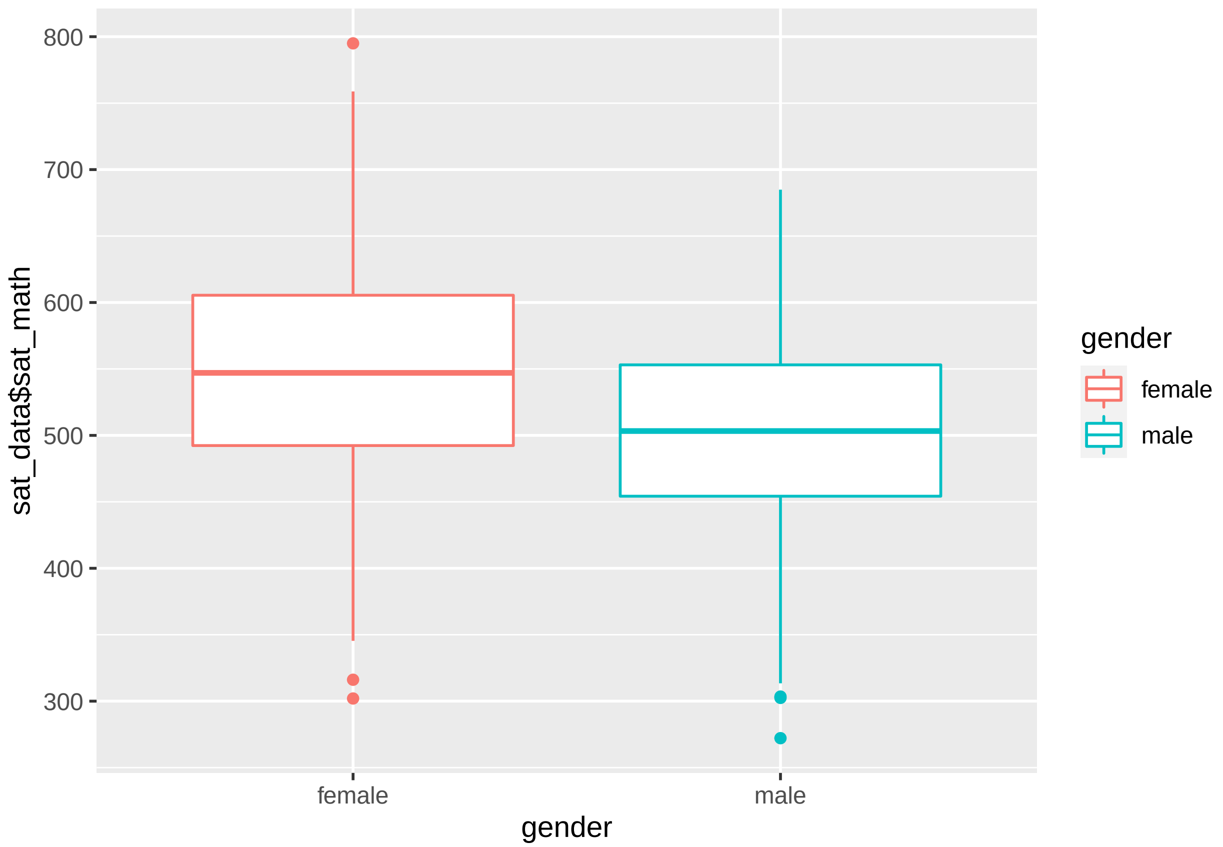


```
sat_data %>% as_tibble() %>%
  ggplot(aes(x = sat_data$hours)) +
  geom_histogram(binwidth = 1, fill = "grey", color = "black")
```



We can see that the distribution of these variables are all aproximately normal and it doesn't look like we need to drop any outliers that are skewing our data.

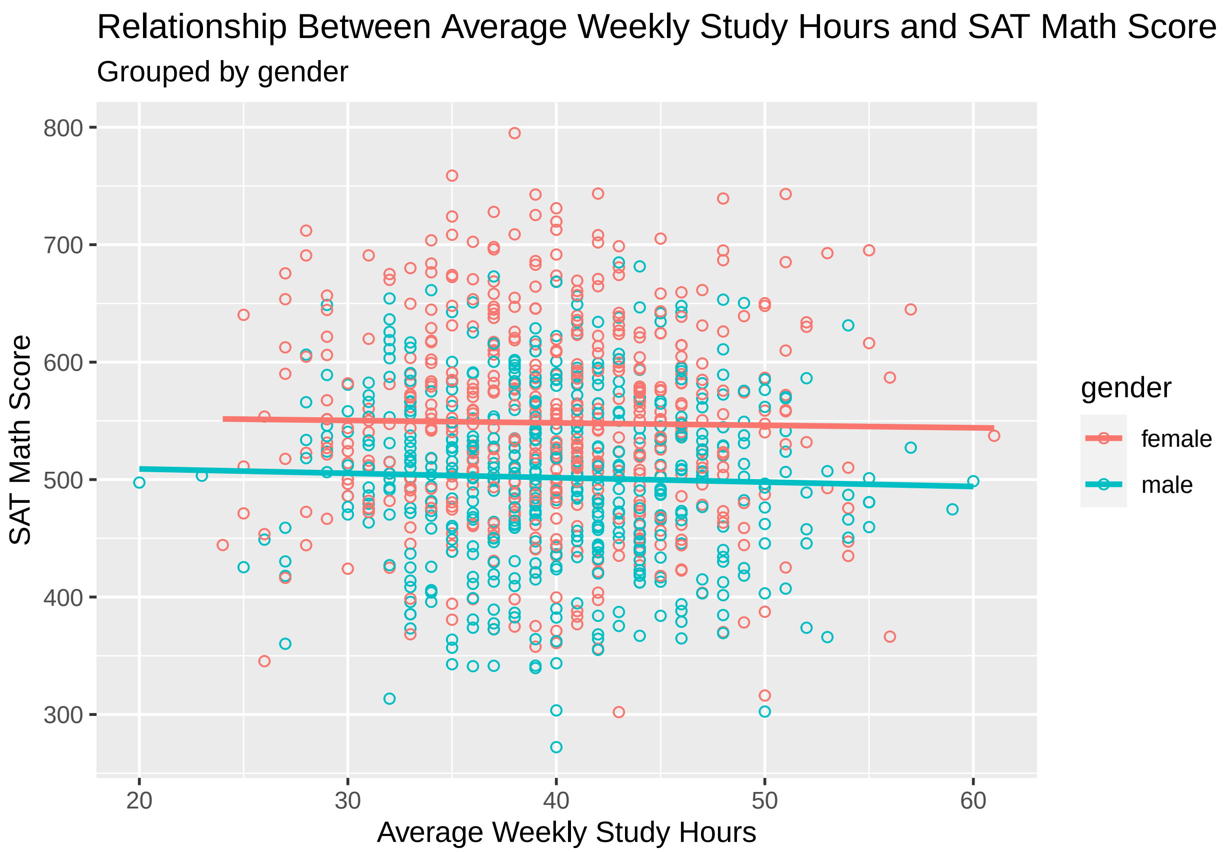
```
# Boxplot of scores by gender to compare means
sat_data %>%
  ggplot() +
  geom_boxplot(aes(x = gender, y = sat_data$sat_math, color = gender))
```



As we saw in the descriptive statistics table, the mean math score for females is higher than males by almost 50 points.

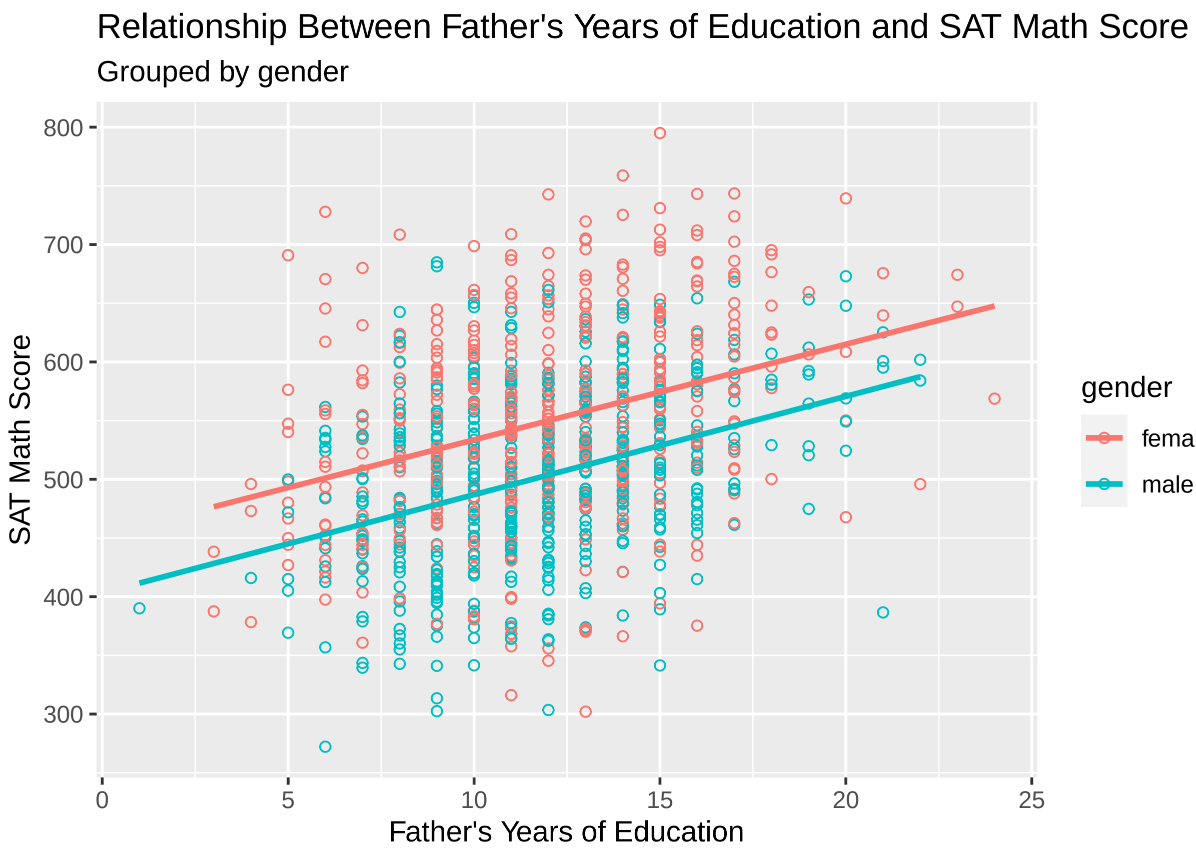
```
# Creating scatterplots with other predictors

# Looking at hours, still grouped by gender
sat_data %>% as_tibble() %>%
  ggplot(aes(x = hours, y = sat_math, color = gender)) +
  geom_point(shape = 1) +
  geom_smooth(method = "lm", se = F) +
  labs(title = "Relationship Between Average Weekly Study Hours and SAT Math Score",
       x = "Average Weekly Study Hours",
       y = "SAT Math Score",
       subtitle = "Grouped by gender")
```

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



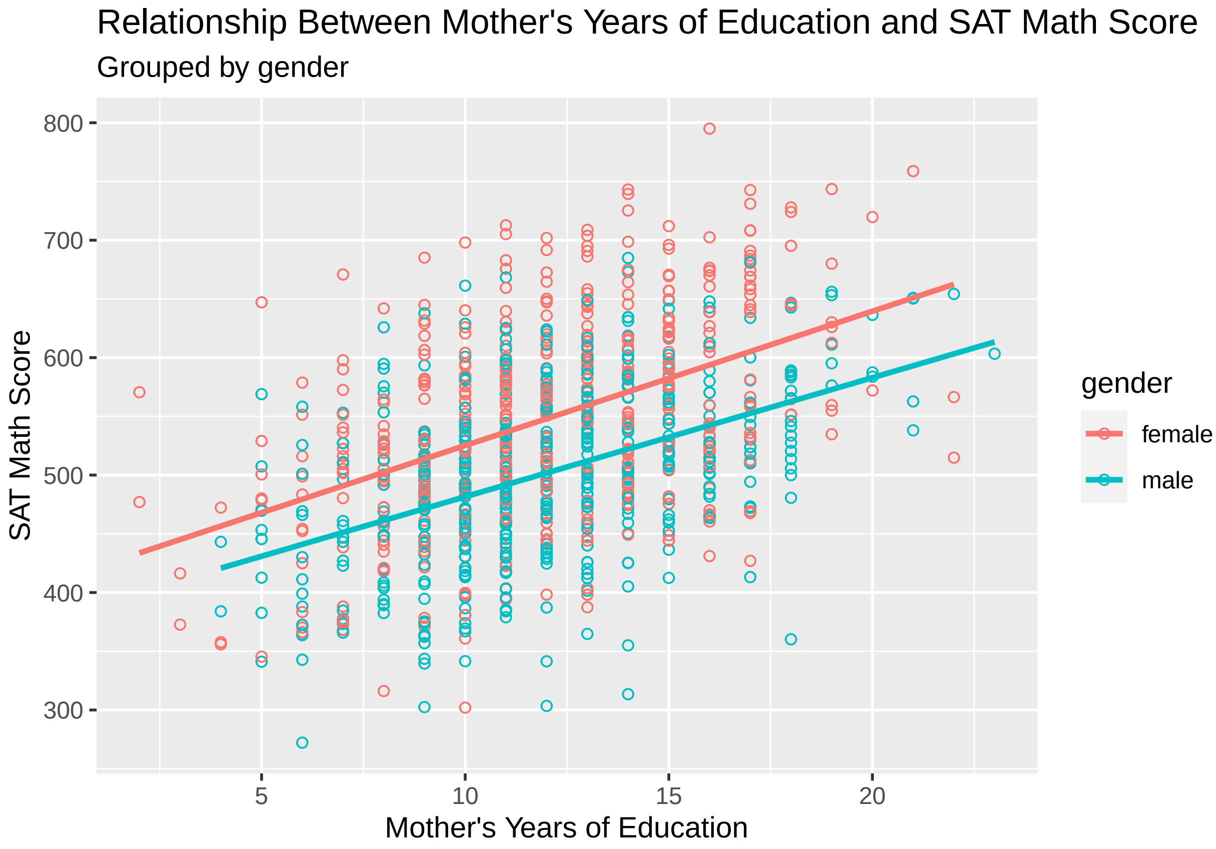
```
# Looking at father's years of education, still grouped by gender
sat_data %>% as_tibble() %>%
  ggplot(aes(x = feduy, y = sat_math, color = gender)) +
  geom_point(shape = 1) +
  geom_smooth(method = "lm", se = F) +
  labs(title = "Relationship Between Father's Years of Education and SAT Math Score",
       x = "Father's Years of Education",
       y = "SAT Math Score",
       subtitle = "Grouped by gender")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```
# Looking at mother's years of education, still grouped by gender
sat_data %>% as_tibble() %>%
  ggplot(aes(x = meduy, y = sat_math, color = gender)) +
  geom_point(shape = 1) +
  geom_smooth(method = "lm", se = F) +
  labs(title = "Relationship Between Mother's Years of Education and SAT Math Score",
       x = "Mother's Years of Education",
       y = "SAT Math Score",
       subtitle = "Grouped by gender")
```

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



We can see that the the predictors most strongly correlated with SAT math score are IQ and parent's education.

```r
# Grouping data by race
sat_data <- sat_data %>% group_by(black, other)
```

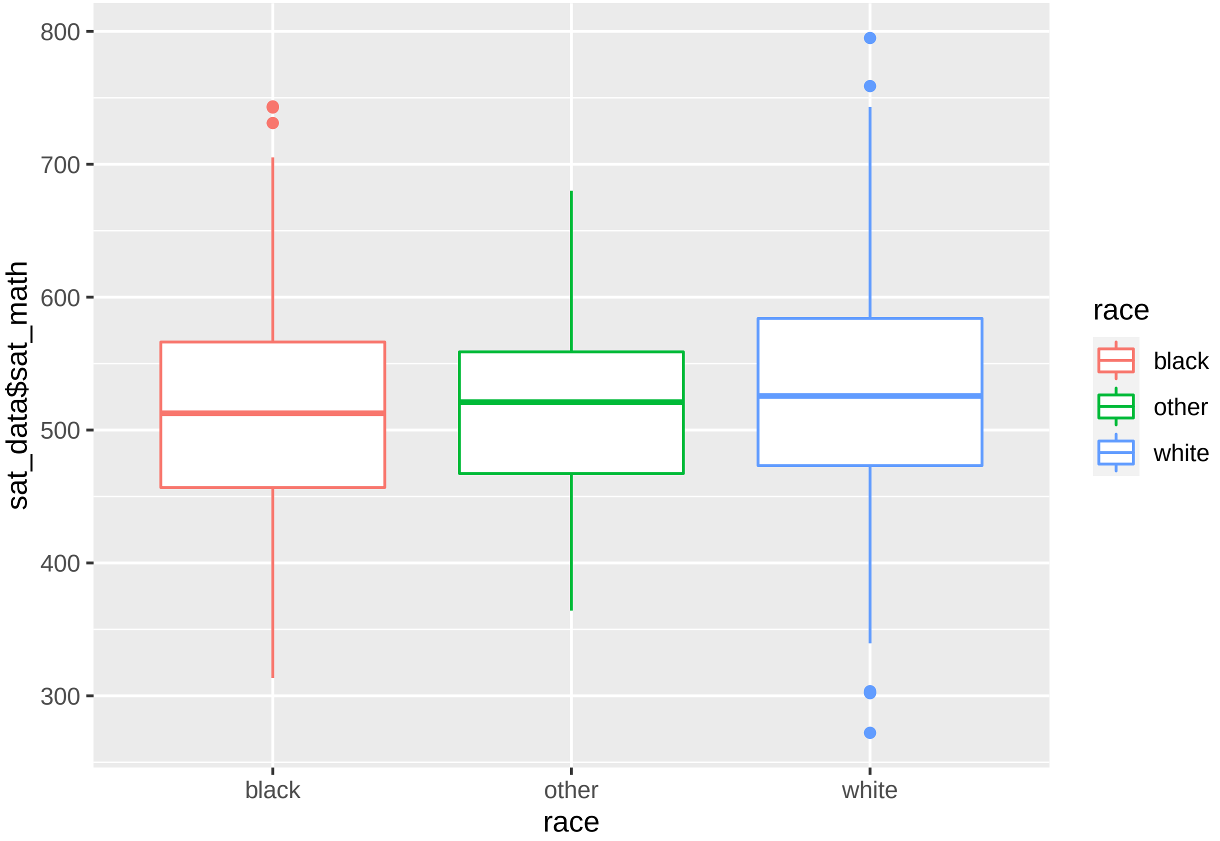
```r
# Creating descriptive stats table
sat_data %>% summarise(mean_satmath = mean(sat_math),
            mean_meduy = mean(meduy),
            mean_feduy = mean(feduy),
            mean_hours = mean(hours),
            mean_IQ = mean(IQ)) %>% kbl("latex") %>% kable_classic_2(full_width = F, position = "left")
```

```
## 'summarise()' has grouped output by 'black'. You can override using the '.groups' argument.
```

black	other	mean_satmath	mean_meduy	mean_feduy	mean_hours	mean_IQ
0	0	529.8538	12.07821	11.83520	39.97067	49.59184
0	1	516.2339	11.89024	11.17073	39.79268	49.89340
1	0	510.9515	11.79144	11.80214	39.50267	50.10527
1	1	497.5668	11.40000	12.13333	42.46667	45.65400

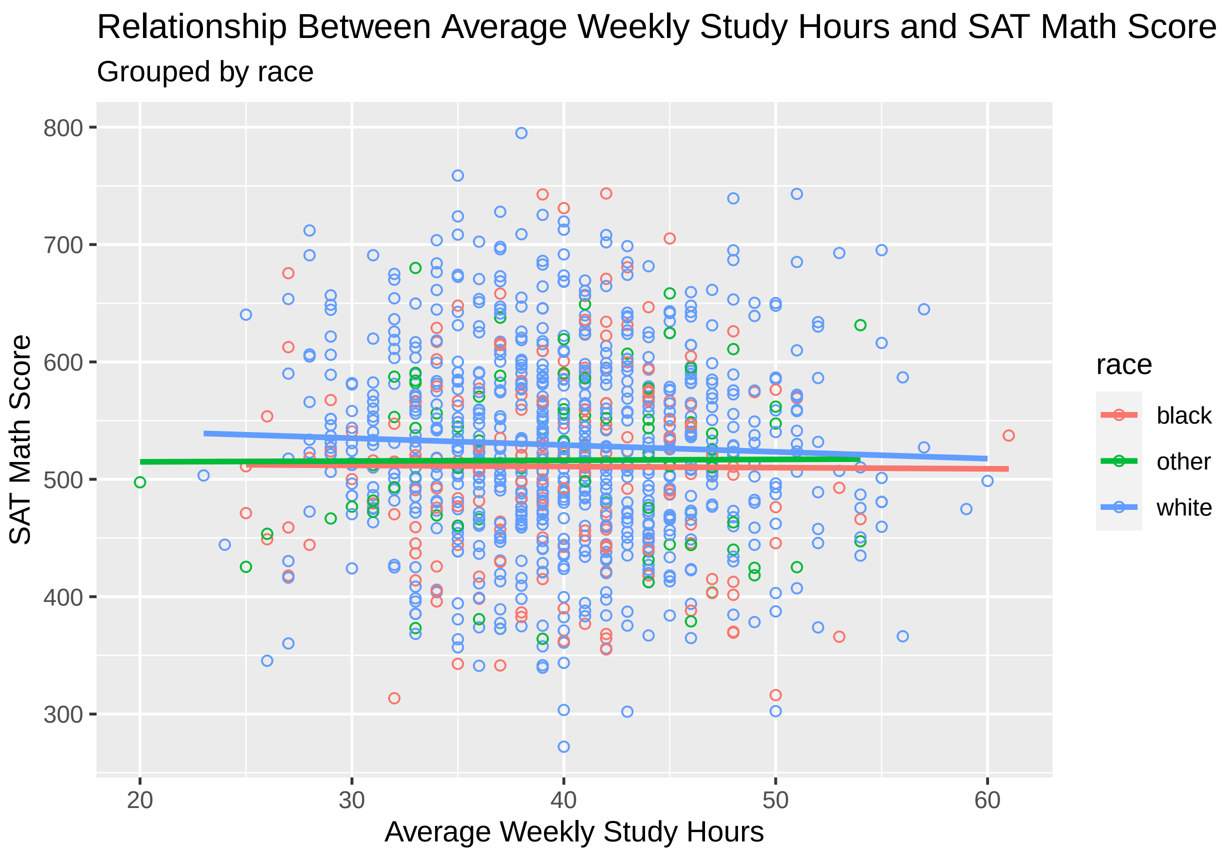
Grouped by race we can see that white people have the highest mean SAT math score and the few people who are both black and other have the lowest. In general, we can can see that non-white people on average score lower than white people.

```
# Running similar comparisons with new race variable
sat_data <- sat_data %>% mutate(race = ifelse(black == 1 & other == 0, "black",
                                              ifelse(black == 0 & other ==1, "other", "white")))

# Creating more plots comparing races

# Boxplot of scores across races, comparing means
sat_data %>%
  ggplot() +
  geom_boxplot(aes(x = race, y = sat_data$sat_math, color = race))
```



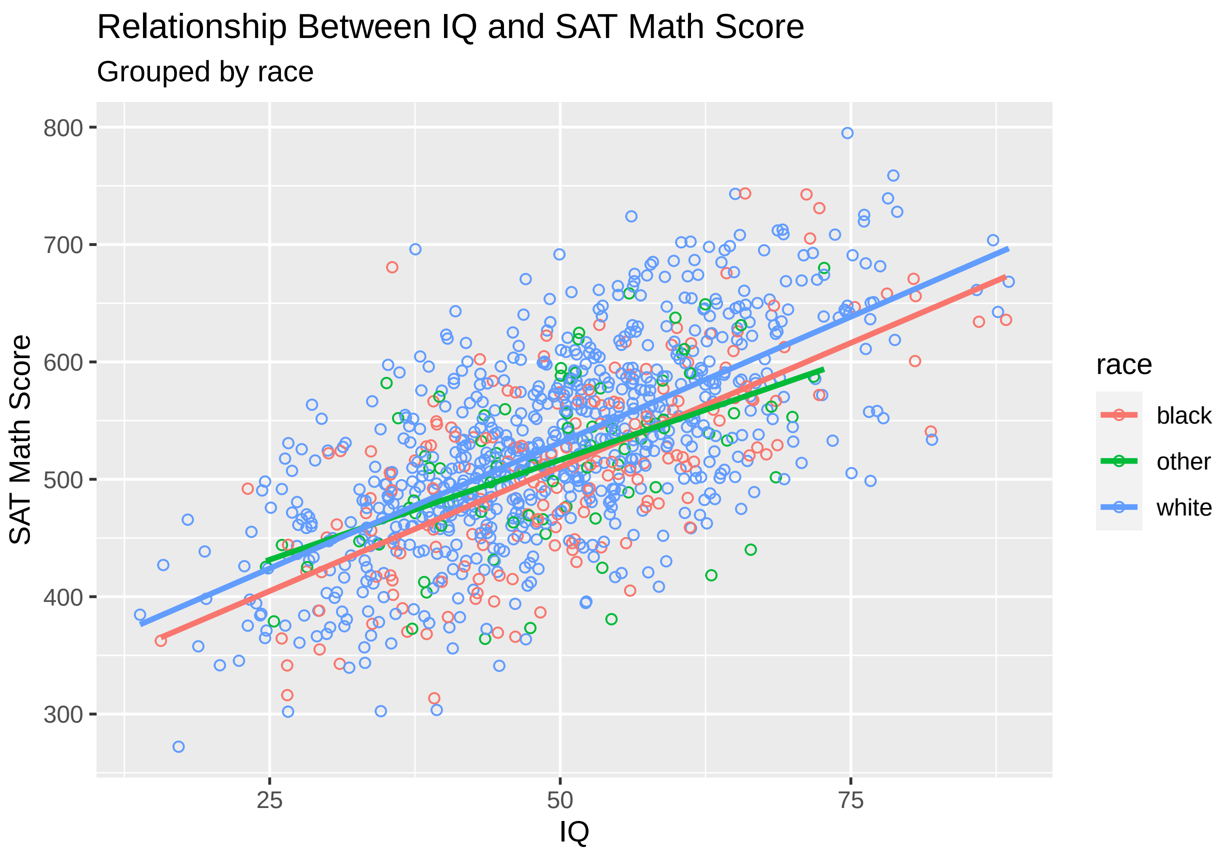
```
# Looking at hours studied, grouped by race
sat_data %>% as_tibble() %>%
  ggplot(aes(x = hours, y = sat_math, color = race)) +
  geom_point(shape = 1) +
  geom_smooth(method = "lm", se = F) +
  labs(title = "Relationship Between Average Weekly Study Hours and SAT Math Score",
       x = "Average Weekly Study Hours",
       y = "SAT Math Score",
       subtitle = "Grouped by race")
```

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



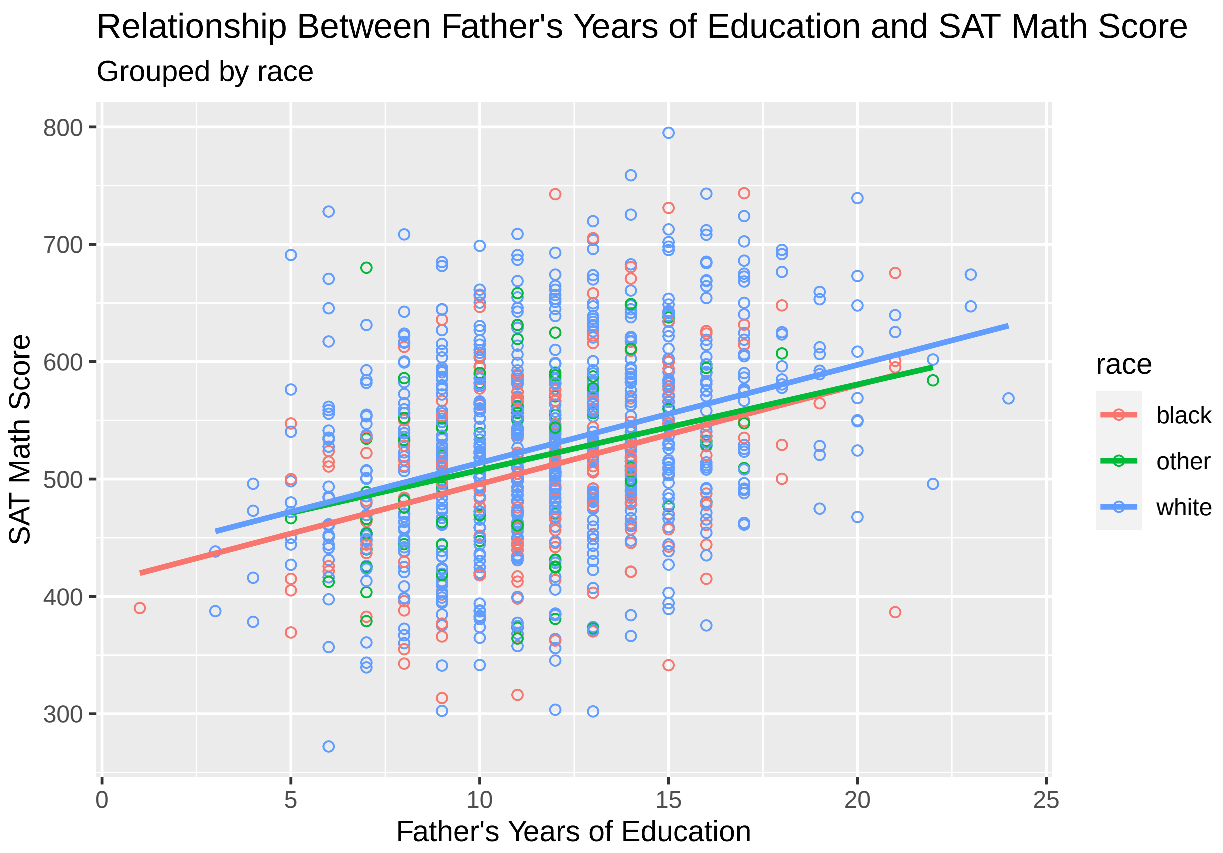
```
# Looking at IQ, still grouped by race
sat_data %>% as_tibble() %>%
  ggplot(aes(x = IQ, y = sat_math, color = race)) +
  geom_point(shape = 1) +
  geom_smooth(method = "lm", se = F) +
  labs(title = "Relationship Between IQ and SAT Math Score",
       x = "IQ",
       y = "SAT Math Score",
       subtitle = "Grouped by race")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



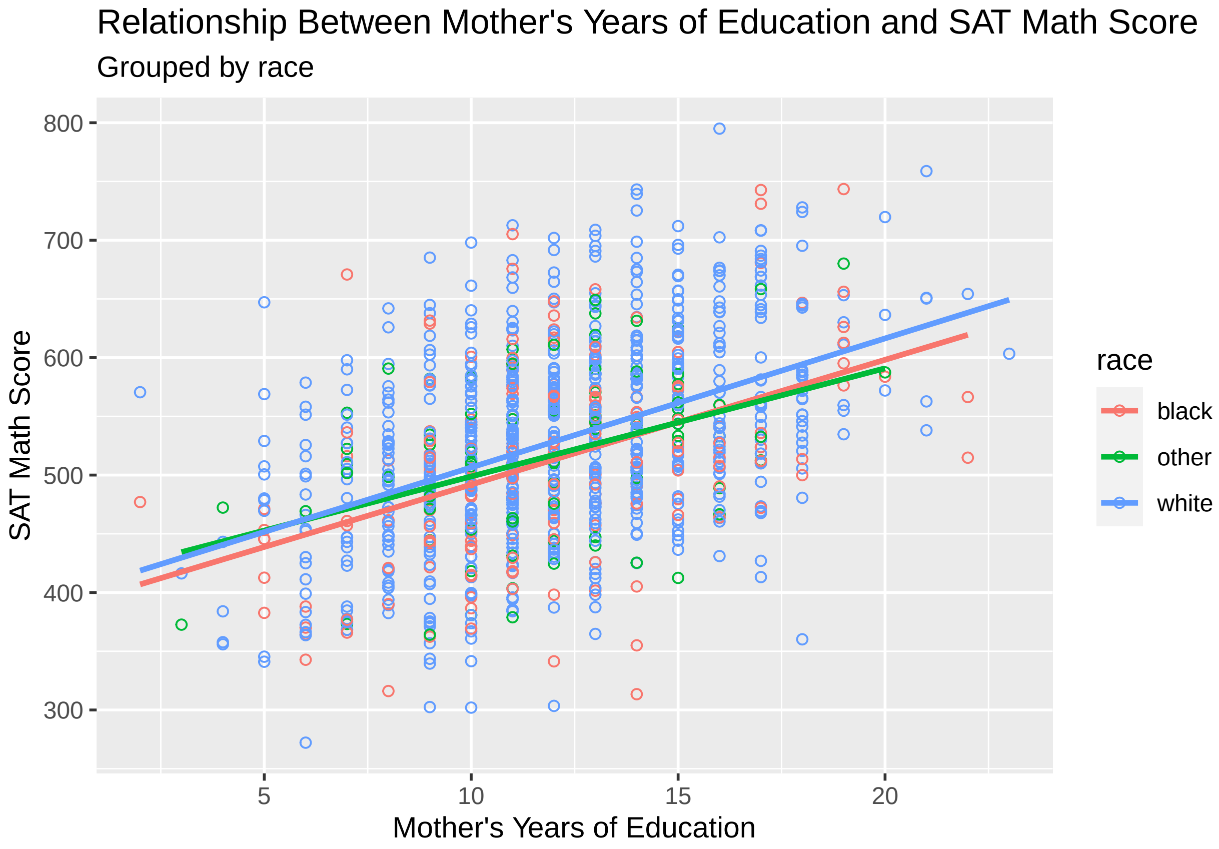
```
# Looking at father's years of education, still grouped by race
sat_data %>% as_tibble() %>%
  ggplot(aes(x = feduy, y = sat_math, color = race)) +
  geom_point(shape = 1) +
  geom_smooth(method = "lm", se = F) +
  labs(title = "Relationship Between Father's Years of Education and SAT Math Score",
       x = "Father's Years of Education",
       y = "SAT Math Score",
       subtitle = "Grouped by race")
```

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



```
# Looking at mother's years of education, still grouped by race
sat_data %>% as_tibble() %>%
  ggplot(aes(x = meduy, y = sat_math, color = race)) +
  geom_point(shape = 1) +
  geom_smooth(method = "lm", se = F) +
  labs(title = "Relationship Between Mother's Years of Education and SAT Math Score",
       x = "Mother's Years of Education",
       y = "SAT Math Score",
       subtitle = "Grouped by race")
```

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



As with when the data were grouped by gender, IQ has the most significant relationship with SAT math score, and study hours has the lowest. We also observe some interaction between IQ and race, and between IQ and parent's education.

4. Nested models:

- Build five nested models that use `sat_math` as the DV and report regression results in a table using `stargazer()` from the `stargazer` package.

(a) Model 1: Baseline (only add "IQ" as the independent variable)

(b) Model 2: Model 1 + Demographic Characteristics

(c) Model 3: Model 2 + Parental Education

(d) Model 4: Model 3 + Weekly Study Hours

(e) Model 5: Model 4 + An Interaction Between IQ and the Female Dummy

```r
# Creating the five models
m1 <- lm(sat_math ~ IQ, data = sat_data)
m2 <- lm(sat_math ~IQ + female + black + other, data = sat_data)
m3 <- lm(sat_math ~IQ + female + black + other + feduy + meduy, data = sat_data)
m4 <- lm(sat_math ~IQ + female + black + other + feduy + meduy + hours, data = sat_data)
m5 <- lm(sat_math ~IQ + female + black + other + feduy + meduy + hours + IQ*female, data = sat_data)
```

```r
stargazer(m1, m2, m3, m4, m5, type = "latex", omit.stat=c("f", "ser"))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Mar 19, 2021 - 16:04:44

Table 2:

	Dependent variable:				
	sat_math				
	(1)	(2)	(3)	(4)	(5)
IQ	4.211***	4.356***	3.487***	3.484***	2.893***
	(0.154)	(0.138)	(0.132)	(0.132)	(0.175)
female		53.831***	51.693***	51.597***	−9.552
		(3.565)	(3.142)	(3.143)	(12.357)
black		−16.130***	−14.455***	−14.519***	−15.324***
		(4.432)	(3.906)	(3.906)	(3.861)
other		−10.809*	−6.918	−6.825	−6.152
		(6.005)	(5.294)	(5.295)	(5.231)
feduy			5.725***	5.747***	5.703***
			(0.474)	(0.475)	(0.469)
meduy			6.434***	6.434***	6.379***
			(0.491)	(0.491)	(0.485)
hours				−0.264	−0.255
				(0.251)	(0.248)
IQ:female					1.232***
					(0.241)
Constant	315.605***	286.110***	184.937***	195.423***	226.197***
	(7.906)	(7.519)	(8.921)	(13.391)	(14.530)
Observations	1,000	1,000	1,000	1,000	1,000
R^2	0.428	0.543	0.646	0.647	0.656
Adjusted R^2	0.427	0.541	0.644	0.644	0.653
Note:					*p<0.1; **p<0.05; ***p<0.01

5. For the result of Model 1:

(a) What are the hypotheses that you are testing in this model with your t-values in the `(Intercept)` and `IQ` row of the modeling results?

Our null hypothesis for IQ is, holding all other variables constant, $\beta_{\mathbf{IQ}} = 0$. Our alternative hypothesis is $\beta_{\mathbf{IQ}}$!= 0. Our prediction equation is $\hat{\mathbf{Y}} = 4.211\mathbf{X} + 315.605$.

Our null hypothesis for Intercept is, holding all other variables constant, $\beta_0 = 0$. Our alternative hypothesis is β_0 != 0. Our prediction equation is $\hat{Y}$ = 4.211X + 315.605.

(b) Create a 95% confidence interval for the parameter β_{IQ} based on Model 1 result.

```
#Creating confidence interval
upperbound <- 4.211 + (1.96*0.154)
lowerbound <- 4.211 - (1.96*0.154)
CI <- c(lowerbound,upperbound) %>% print()
```

```
[1] 3.90916 4.51284
```

6. Interpret regression coefficients:

(a) How does the coefficient of "IQ" change across models? What could be the possible reason(s) for such changes?

The coefficient of IQ changes from 4.211, to 4.356, to 3.487, to 3.484, and finally to 2.893. The increase from m1 to m2 suggests that adding these minority demographic predictors (female/black/other) increases the explanatory power of the partial effects of IQ by itself on SAT math score.The decreases of the coefficient from m2 to m5 suggests the partial effects of IQ by itself on SAT math scores decreases in power when other significant predictors are added to the model, such as parent's years of education, hour spent studying, and interaction between gender and IQ.

(b) Interpret the coefficient of "black" in Model 4.

The coefficient of -14.519 for "black" in model 4 means that black people (the comparison group) on average score 14.519 points lower than white people (the reference group), holding all else constant.

(c) Interpret the coefficient of "meduy" in Model 4.

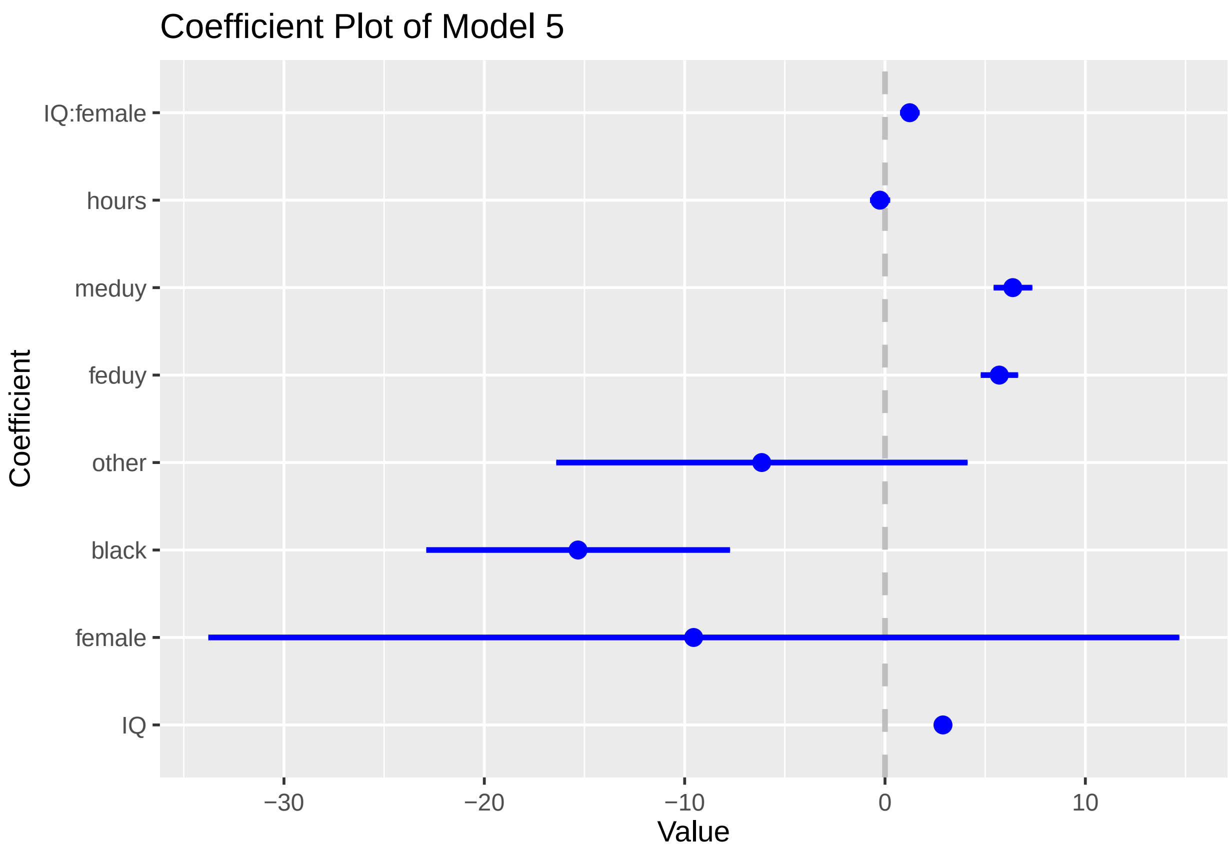
The coefficient of 6.434. for "meduy" in model 4 means that for each additional year of a person's mother's education, their SAT math score increases by 6.434, holding all else constant.

(d) Interpret the coefficient of the interaction effect between IQ and the Female Dummy in Model 5.

The coefficient of 1.232 for the interaction term in model 5 means that the effect of IQ on SAT math scores is greater for females (the comparison group) than males (the reference group), with a 1.232 point difference.

```


# Using coefplot
coefplot(m5,
         innerCI = 1.96,
         outerCI = 1.96,
         intercept = F,
         title = "Coefficient Plot of Model 5")
```

7. Create a coefficient plot for Model 5 with appropriate title and labels.

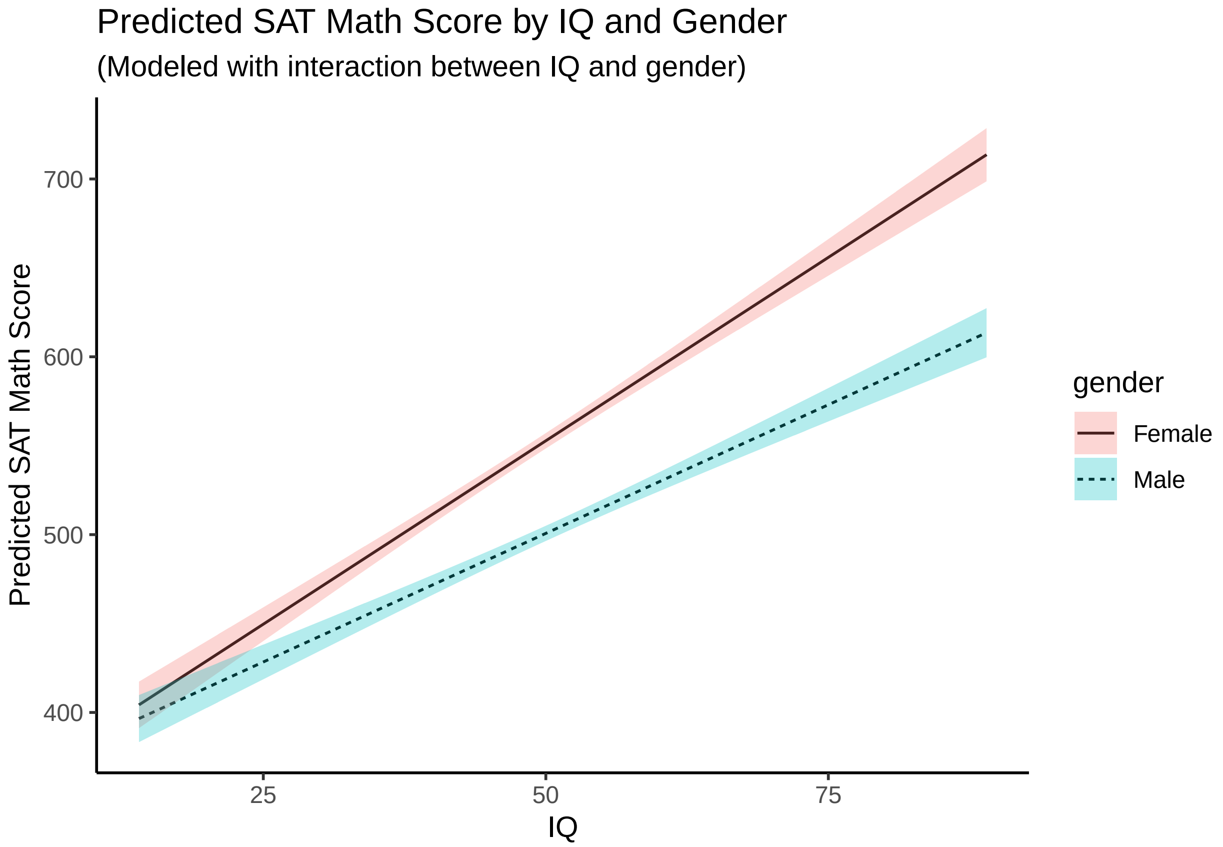


```
pred_IV <- tibble(IQ = rep(14:89, 2)) %>%         #create df with key IV
  mutate(female = c(rep(0, 76), rep(1, 76)),    # fix other variables at mean
         hours = mean(sat_data$hours, na.rm = T),
         meduy = mean(sat_data$meduy, na.rm = T),
         feduy = mean(sat_data$feduy, na.rm = T),
         black = mean(sat_data$black, na.rm = T),
         other = mean(sat_data$other, na.rm = T))
```

```
# use `predict` to predict the Y
# the model you are using, the df you use, setting CI
predicted_math <- predict(m5, pred_IV, interval = "confidence", level = 0.95)

# bind the columns
pred_result <- cbind(pred_IV, predicted_math)
```

```
# Plot
pred_result %>%
  mutate(gender = ifelse(female == 0, "Male", "Female")) %>%# Convert dummy to character variable
  ggplot(aes(x = IQ, y = fit)) +
  geom_line(aes(linetype = gender)) +   # group linetype by gender
  geom_ribbon(aes(ymin = lwr, ymax = upr, fill = gender), alpha = 0.3) + # add 95% CI
  theme_classic() +
  labs(x = "IQ",
       y = "Predicted SAT Math Score") +
  ggtitle("Predicted SAT Math Score by IQ and Gender",
          subtitle = "(Modeled with interaction between IQ and gender)")
```

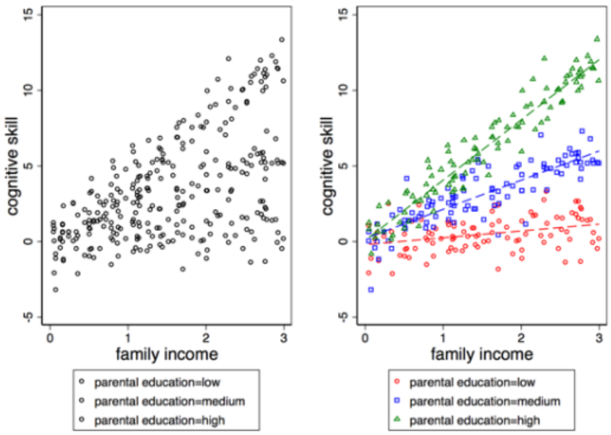
8. On the basis of Model 5, by holding other variables at their means, create a figure demonstrating the predicted SAT math score by gender and IQ levels (with confidence interval)



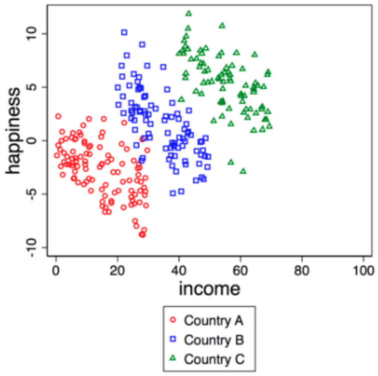
Part 3 (Bonus) Data Simulation

Simulation is a fun and effective way to learn about statistical inference. You will get a better understanding of how each population parameter affects the shape of the distribution.

Now that we have learned about how to identify interactions from a given sample, you can try simulate a data whose true data generating process involves interaction between two variables. For example, you can try to reproduce a similar scatter plot we saw in class (the right panel) by simulating a data whose variables have such associations:



Or, you can try to reproduce a scatter plot that demonstrates the Simpson's Paradox:

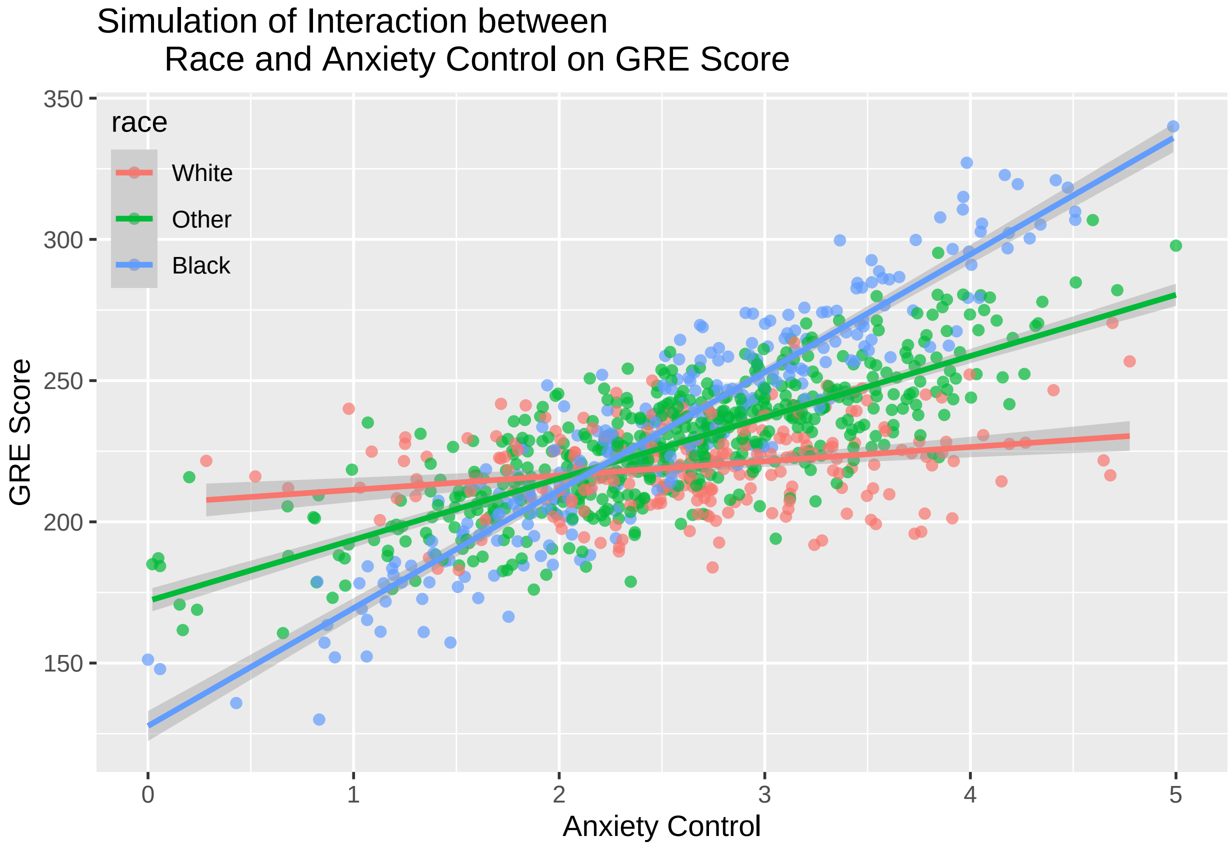


Note: Your output does not need to replicate the exact layout of the example graphs. You will get extra credit as long as you generate a similar graph that illustrates the relationship (either a positive or negative interaction, or the Simpson's Paradox) clearly. Remember to use `set.seed()` for any random process.

```r
# This simulation shows the effects of race and a made-up measurement called "anxiety control"
# (measured from 0-5) on GRE scores (130-340). The simulation shows that, for black and other
# races, the effects of self control are more significant for their GRE scores.

n <- 1000

# Simulating IV variables, race and anxiety control, and DV variable GRE Score
set.seed(1234)
race <- rbinom(n, 2, 0.5)
anxiety_control <- rnorm(n)
GRE <- 30 + 6*race + 2*anxiety_control + (6*race)*(2*anxiety_control) + rnorm(n, 0, 10)

# labeling three different race values
race = factor(race, labels = c("White", "Other", "Black"))

# rescaling variables
GRE = rescale(GRE, to = c(130, 340))
anxiety_control = rescale(anxiety_control, to = c(0, 5))
```

```
# creating data frame with variables
data <- data.frame(
  race,
  GRE,
  anxiety_control
)

# Creating the scatterplot
data %>% ggplot(aes(anxiety_control, GRE)) +
  geom_point(aes(col = race), alpha = 0.7) +
  geom_smooth(aes(col = race), method = 'lm') +
  labs(title = "Simulation of Interaction between
       Race and Anxiety Control on GRE Score", x = "Anxiety Control", y = "GRE Score") +
  theme(legend.background = element_rect(fill = "transparent"),
        legend.justification = c(0, 1),
        legend.position = c(0, 1))
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```


Simulation of Interaction between
Race and Anxiety Control on GRE Score
race
White
Other
Black
GRE Score
Anxiety Control