Does a Wage Gap Exist between The Genders and Races?

Bahirah Adewunmi

The data analysis report will assess if there is a statistical difference in income between races and genders, and which factors have a statistically significant affect on this difference.

The purpose of this report is demonstrate my capabilities in ingesting, cleaning, and analyzing data. It's not a report intended for decision owners. For example, I would supress my data cleaning code, and cut a significant amount of analysis completed to only focus on models and tests that have insightful results.

I will first explore the data, form a hypothesis, and test this hypothesis.

Assumptions

I'm assuming that dataset is consistently accurate and any outliers were accurately recorded. For example, I assume that in this dataset there are no records of anyone who is currently 14 years old is coded as a parent at age 30, or someone coded as female has a value coded for the male response to how often they are sad. I'm also assuming that no multicollinearity, heteroscedasticity, and autocorrelation exists within the data. However I will test for multicollinearity later in the data.

Let's Start

First, let's upload the data and assign the columns to more descriptive names. The altered dataset will be renamed to "renamed.data".

```
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.3.3
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
library(stats)
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.3.3
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.3.3
```

```
library(RColorBrewer)
```

```
## Warning: package 'RColorBrewer' was built under R version 3.3.2
```

```
options(scipen=99999)
untouched.data <- read.table("C:/Users/IBM_ADMIN/Desktop/nlsy97_income.csv", header=TRUE, sep=",")
#Rename Columns and copy data to new frame name
renamed.data <- rename(untouched.data, c("E8043100"="incar.totnum","E8043200"="incar.age.first","E80434
```

Also, I'll create a list to provide color to my graphics.

```
par(mar = c(0, 4, 0, 0))
cbPalette <- c("#FFF7FB","#ECE7F2","#D0D1E6","#A6BDDB", "#74A9CF","#3690C0","#0570B0","#034E7B")
cbPaletteContrast <- c("#FFF7EC","#FEE8C8","#FDD49E", "#FDBB84","#FC8D59","#EF6548","#D7301F","#990000")
```

Code Cited: library(RColorBrewer) from "Cheat Sheets for Plotting Symbols and Color Palettes" - http://vis.supstat.com/2013/04/plotting-symbols-and-color-palettes/

Checking the Data for Missing and Invalid Values...

that are not useful in hypothesis testing, correlation tests, and regression analysis.

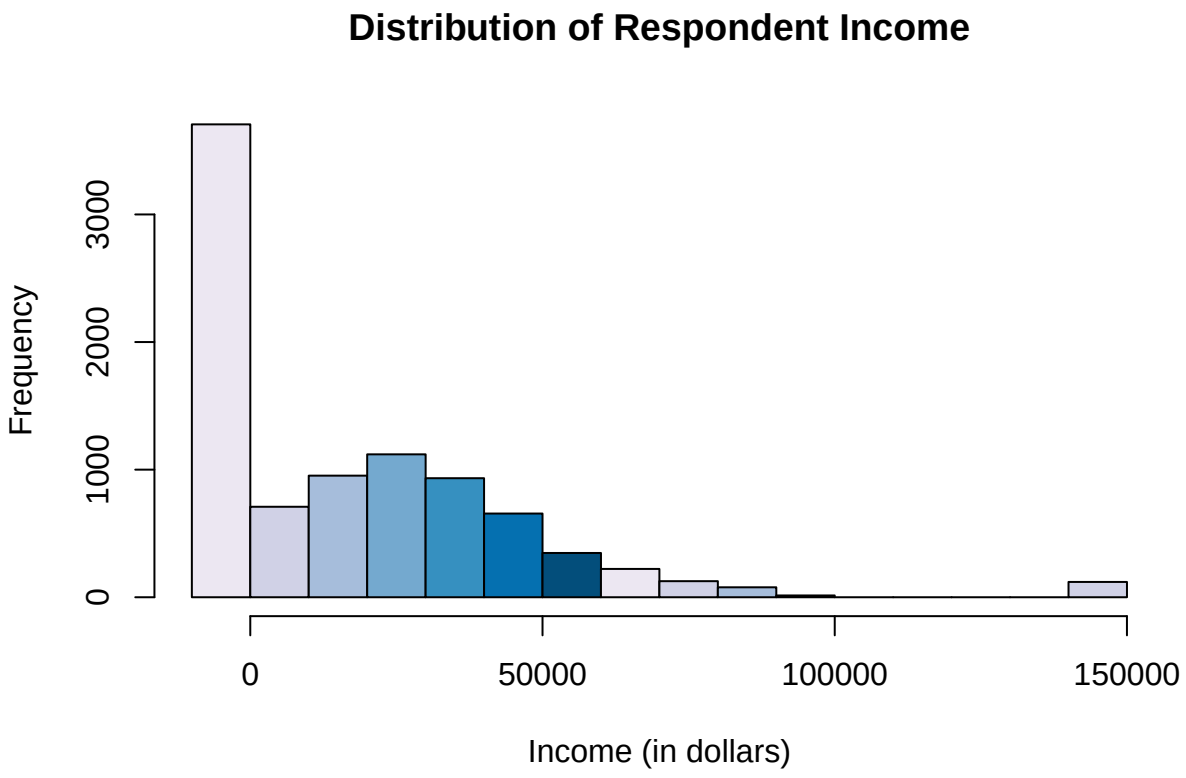
The respondents' [and spouses'] income reported last year (income.lastYear) are the potential variables I will test against other variables in my analysis. It will be the basis of answering whether there's a statistically significant difference in income between the genders. The gender of the respondents are represented by the "gender.youth" variable. These three variables are important, so any NA or missing entries in these variables will be listwise dropped.

Let's check the key variables of my analysis (income and gender) to see if there are any missing values.

```
summary(renamed.data$income.lastYear)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      -5      -4   12000   20160   33420  146000
```

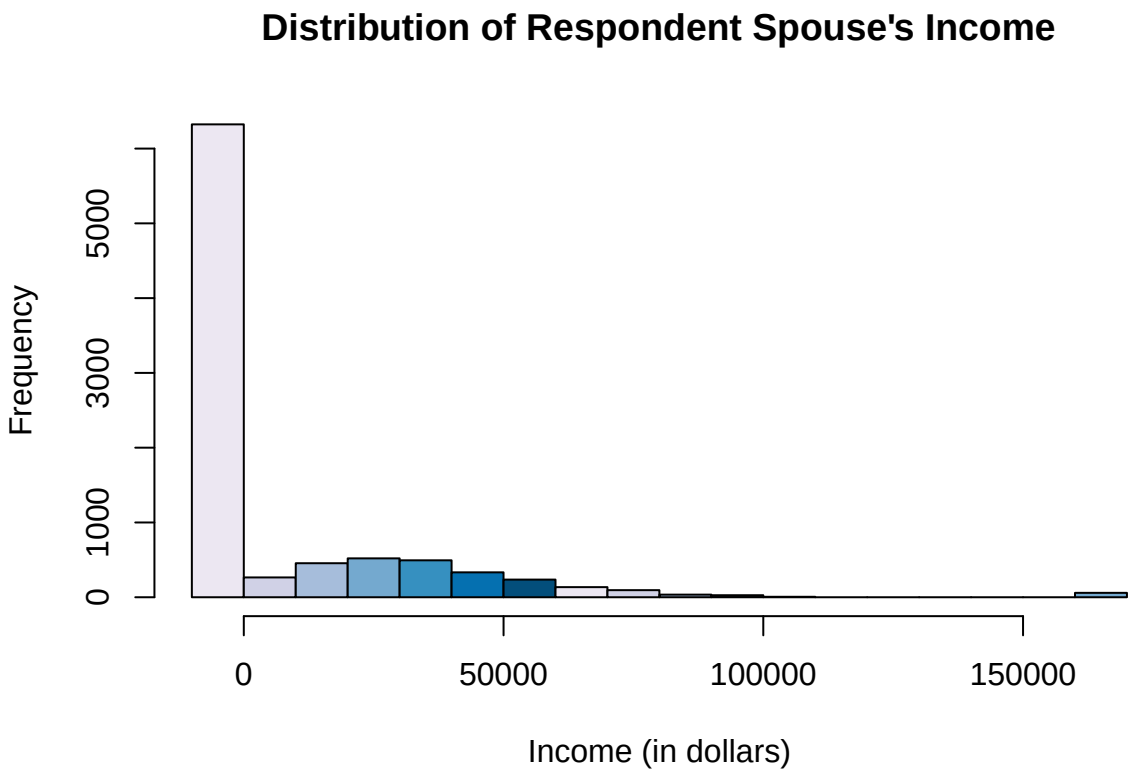
```
hist(renamed.data$income.lastYear,
     main= "Distribution of Respondent Income",
     xlab="Income (in dollars)",
     col = cbPalette[2:8])
```



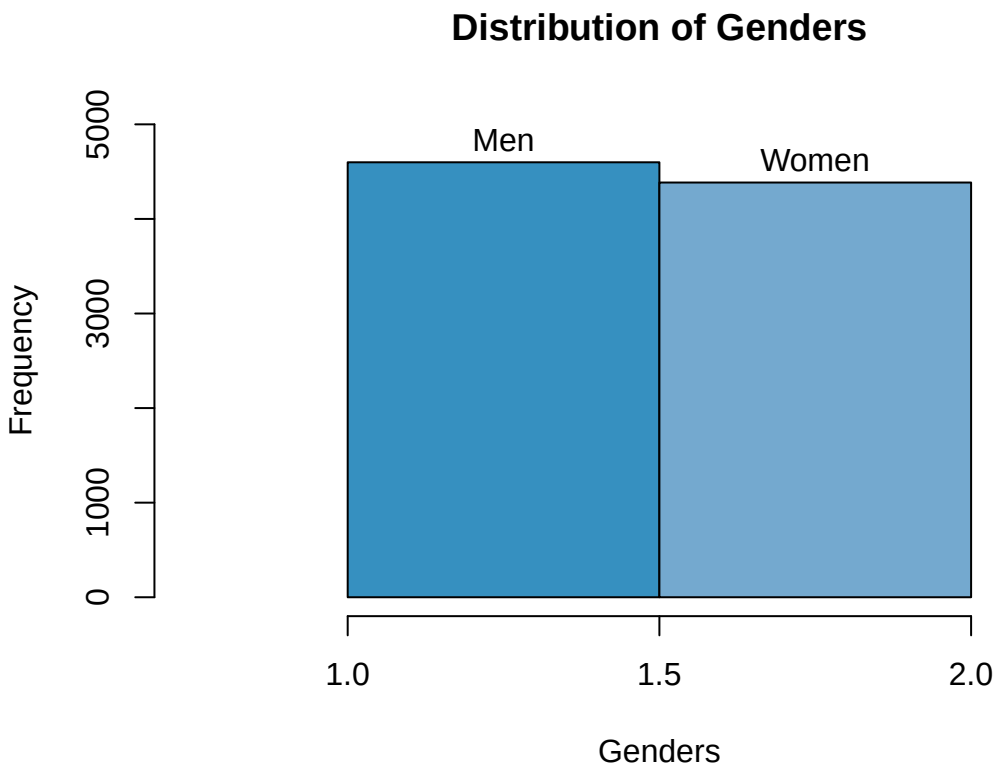

```
summary(renamed.data$income.spouselastYear)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      -5      -4      -4   11290   15000  163800
```

```
hist(renamed.data$income.spouselastYear,
     main= "Distribution of Respondent Spouse's Income",
     xlab="Income (in dollars)",
     col = cbPalette[2:8])
```




```
hist.default(renamed.data$gender.Youth,
             breaks = 2,
     main= "Distribution of Genders",
     xlab="Genders",
     xlim=c(.75,2.25),
     ylim=c(0,5000),
     labels = c("Men","Women"),
     col=c(cbPalette[6],cbPalette[5]))
```



Fortunately, there are no NA values amongst the income, race, and gender variables, but there are entries with different levels of missingness. There are represented by negative numbers. Approximately half of the respondent's income data and 75% spouse's income data was coded as valid skips or non-interview. Using these negative responses aren't helpful to any test that's susceptible to leverage or lift bias.

3682 entries, 43% of the respondent's income data ($income.lastYear) entries can be coded to NA. 6308 entries, more than 75% of the spouse's income entries (income.spouselastYear) can be coded to NA. Since the spouse information has lost nearly all of its data, it's lost its usefulness to this analysis. I won't be using this variable to test my question and will only use the respondent's income level to answer the question.

I could replace and impute a value based on the average income across the respective income column, but I do not have enough information about the data collection to justify that this will not alter any potential statistically significant differences between the genders, harm the precision of confidence intervals, or produce any biased parameter estimates. I'll most likely remove these rows and halve my sample size for further analysis.

Let's check if my selected demographic ("race", "birthYear.Youth","marital.status"), wealth ("debtAge20"","household.net.worth.Parent") education ("highestDegreePrior2011","gradesinHighSchool","highDiploma.By20","coll and crime involvement variables"daysUse.hardDrugs2011","incar.totnum","daysUse.Use.hardDrugs.2011"") for any missing values or invalid values. I chose these variables because I assumed they would have significant influence on an individual's ability to attain, change, and maintain an individual's source of income.

Summary of Demographic Variables

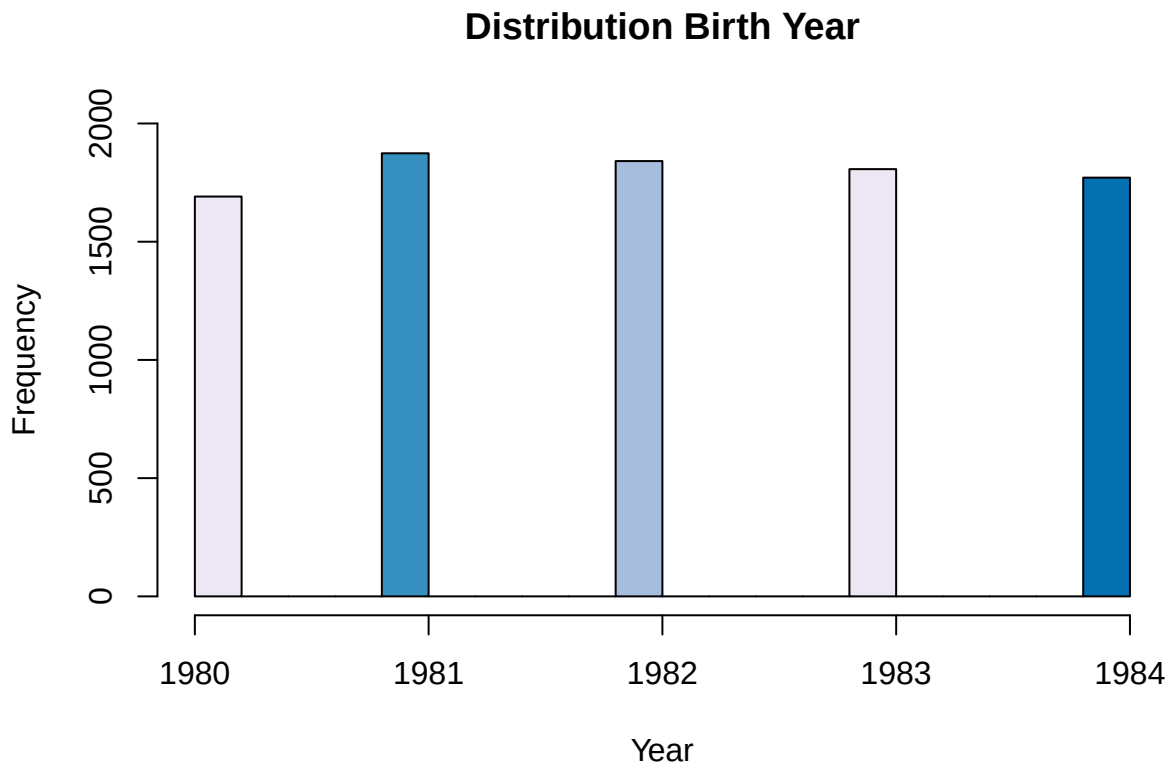
```
summary(renamed.data$race)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   4.000   2.788   4.000   4.000
```

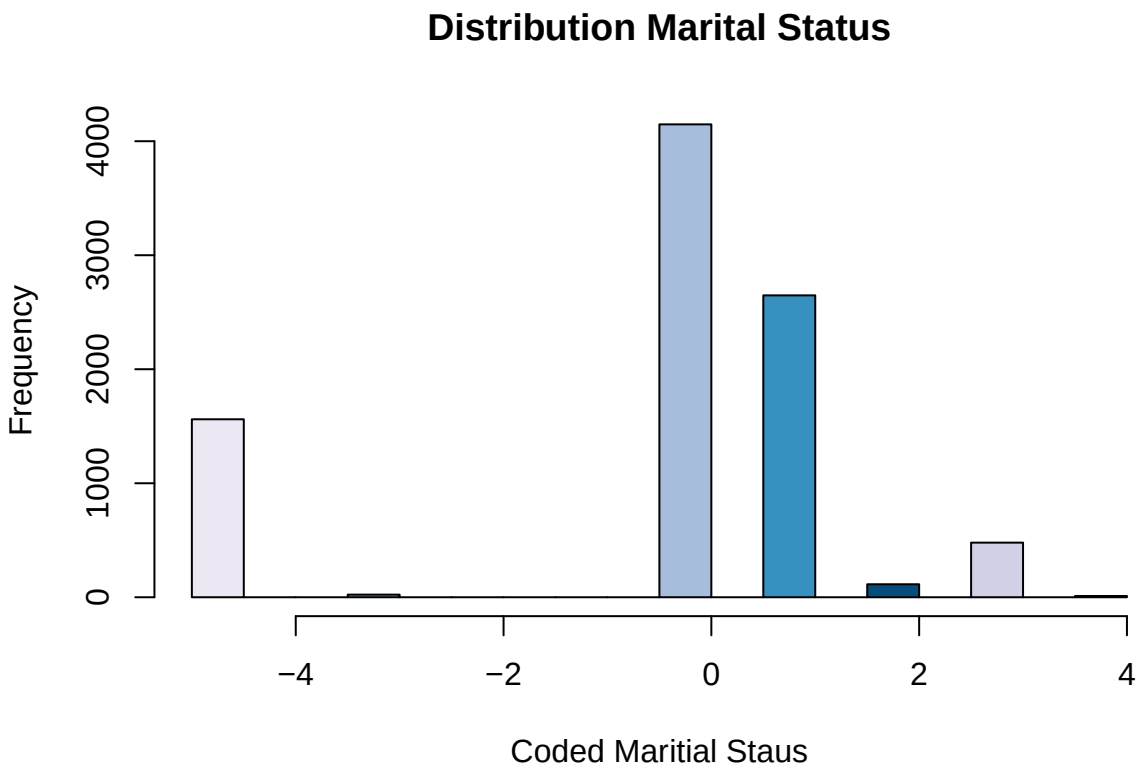
```r
summary(renamed.data$birthYear.Youth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1980    1981    1982    1982    1983    1984
```

```r
hist(renamed.data$birthYear.Youth,
     main= "Distribution Birth Year",
     xlab="Year",
     ylim=c(0,2000),
     col = cbPalette[2:8])
```




```r
hist(renamed.data$marital.status,
     main= "Distribution Marital Status",
     xlab="Coded Maritial Staus",
     col = cbPalette[2:8])
```



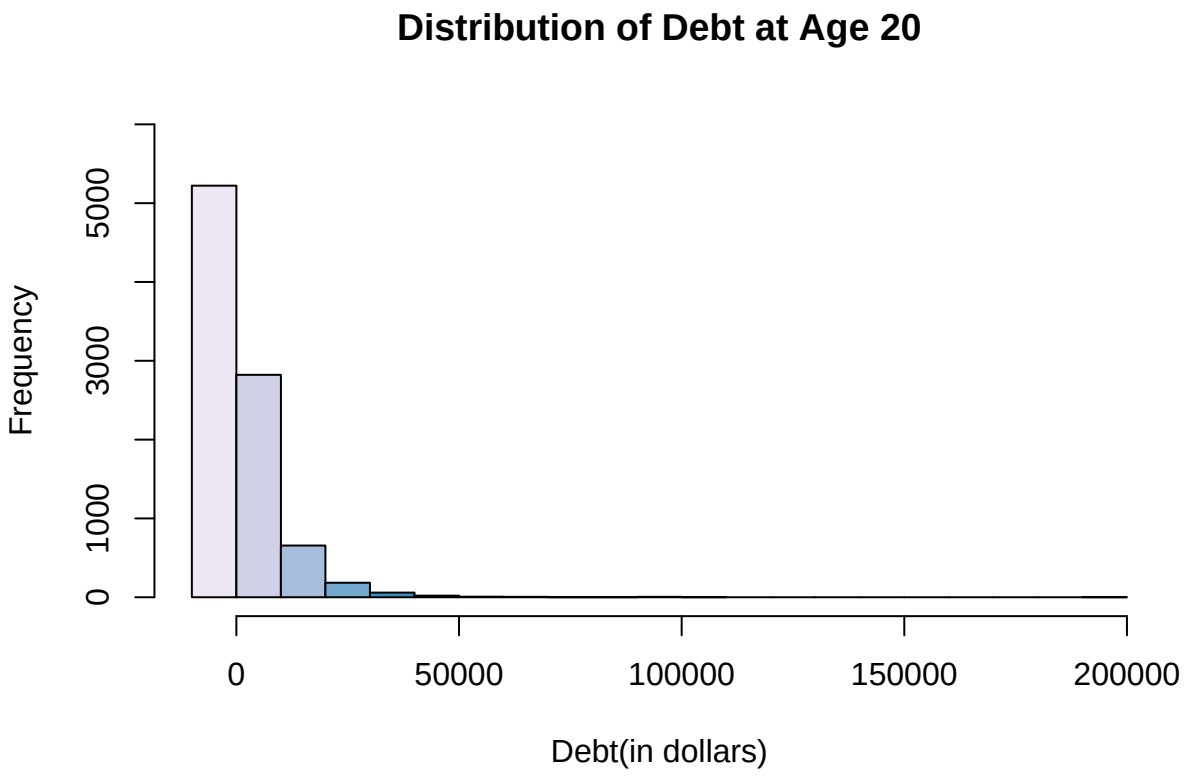
The main takeaway from this summary is that there are no missing values(negative values) in out race (race) and birth year (birthYear.Youth) variables. There's a small portion of missing values in the marital status variable.

Summary of Wealth Variables

```
summary(renamed.data$debtAge20)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      -4       0       0    3164    2800  200000
```

```
hist(renamed.data$debtAge20,
     main= "Distribution of Debt at Age 20",
     xlab="Debt(in dollars)",
     ylim=c(0,6000),
     col = cbPalette[2:8])
```




```
summary(renamed.data$household.net.worth.Parent)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -935300      -3   10500   66450   75260  600000
```

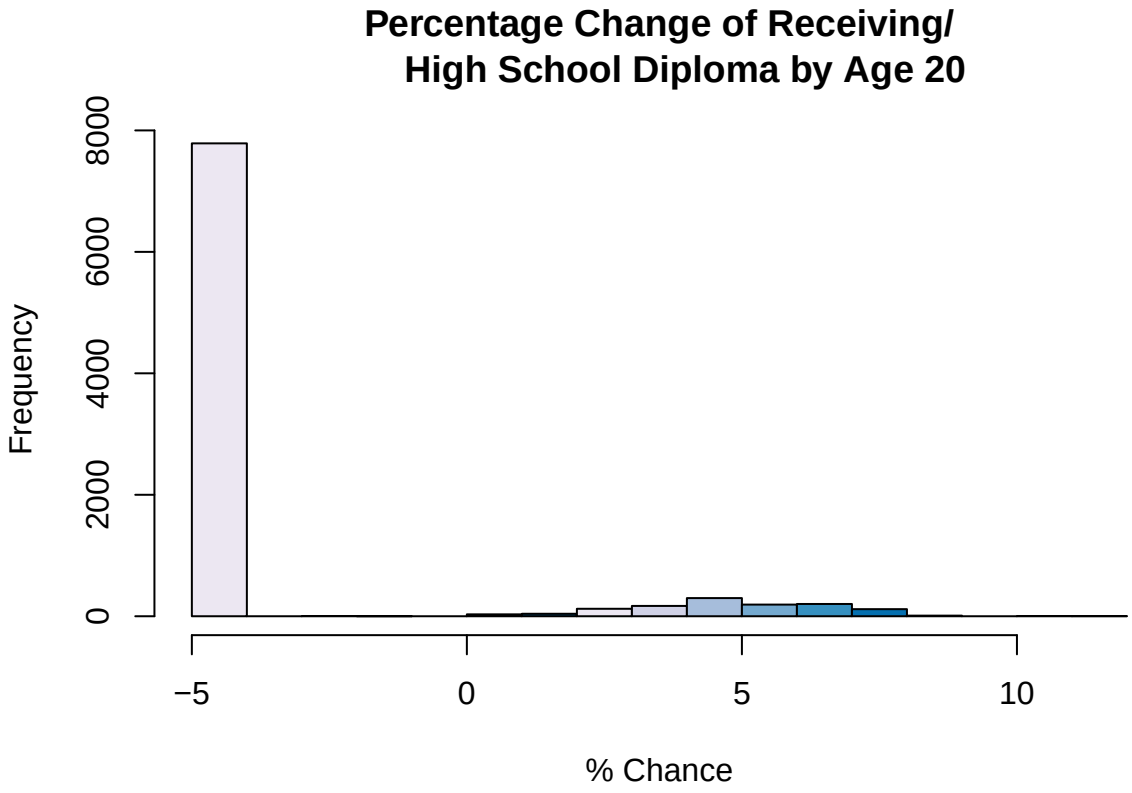
My "debt amount at age 20" and "household net worth according to the parent" is skewed right and a large amount of the data is coded to a missing value. I can infer from this information that this dataset survey imitates the wealth and debt distribution in the United States: a large middle income class, a moderate low income class, and even smaller high income group.

Summary of Education Variables

```
summary(renamed.data$highestDegreePrior2011)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -5.000   0.000   2.000   1.028   3.000   7.000
```

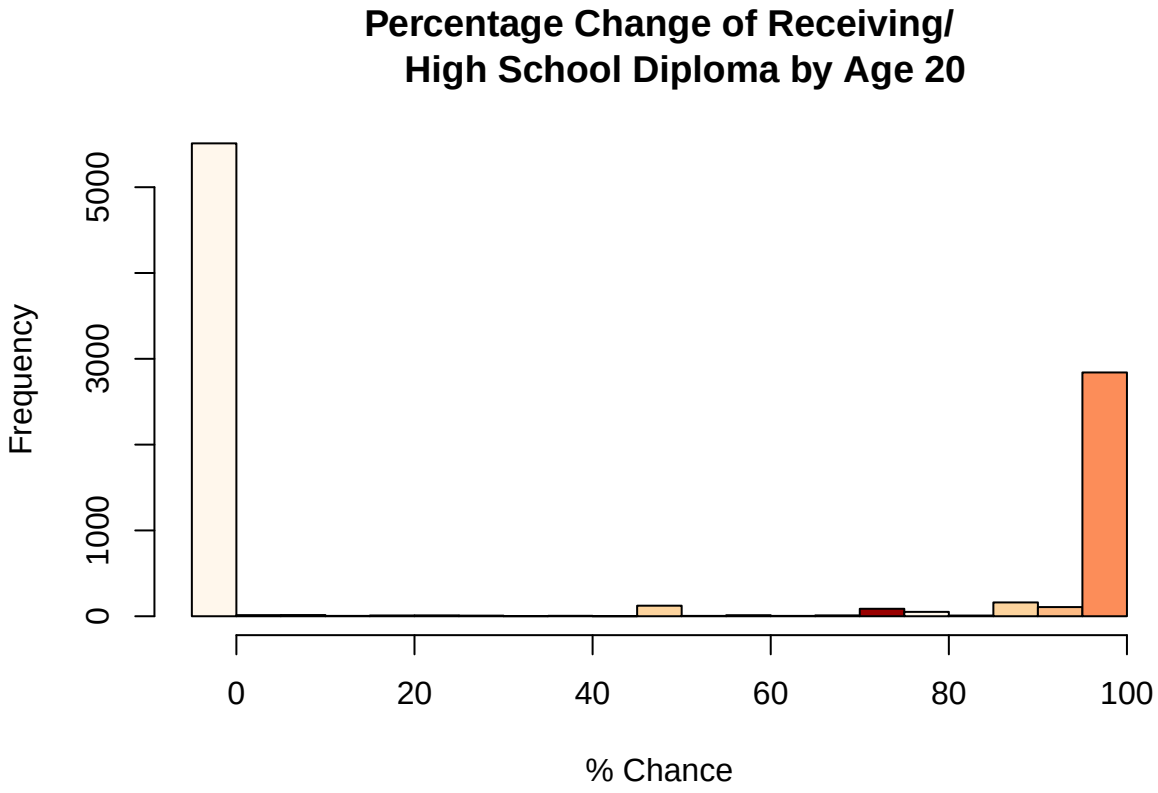
```
hist(renamed.data$gradesinHighSchool,
     main = "Percentage Change of Receiving/
     High School Diploma by Age 20",
     xlab = "% Chance",
     col = cbPalette[2:8])
```




```
summary(renamed.data$highDiploma.By20)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -4.00   -4.00   -4.00   34.21  100.00  100.00
```

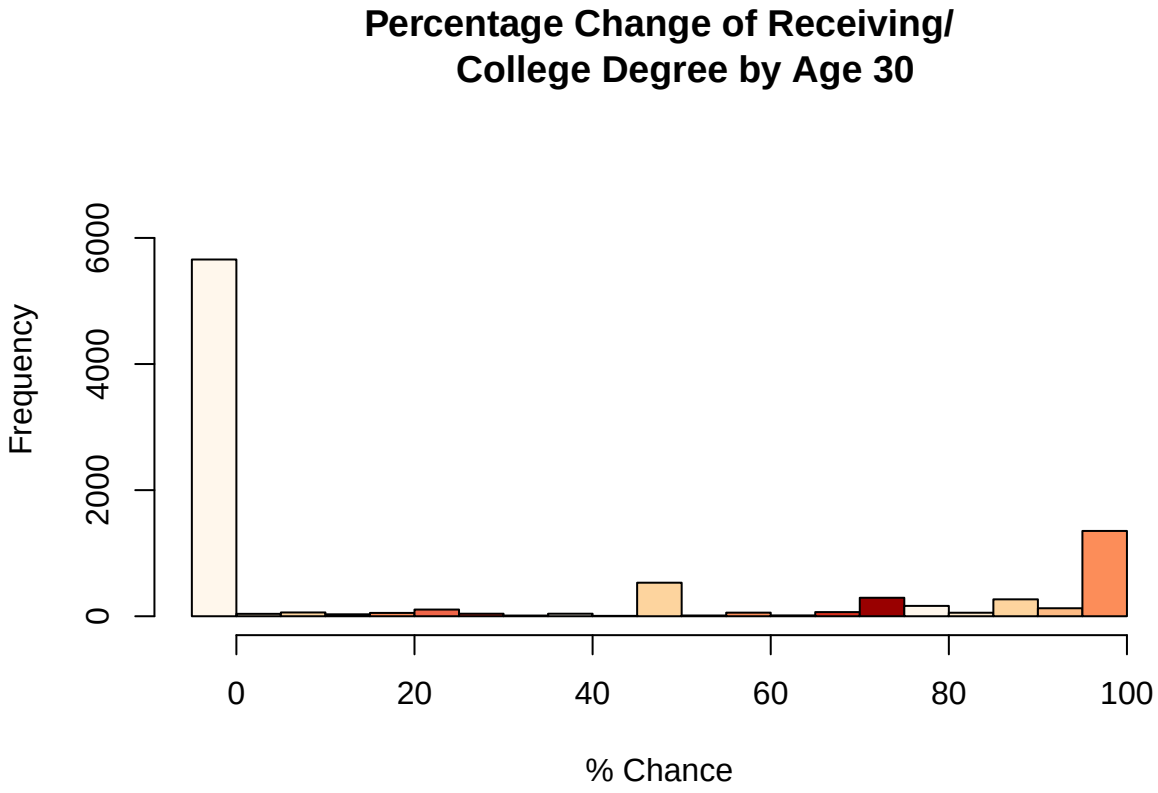
```
hist(renamed.data$highDiploma.By20,
     main = "Percentage Change of Receiving/
     High School Diploma by Age 20",
     xlab = "% Chance",
     col = cbPaletteContrast)
```




```
summary(renamed.data$collegeDegree.By30)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -4.00   -4.00   -4.00   26.02   75.00  100.00
```

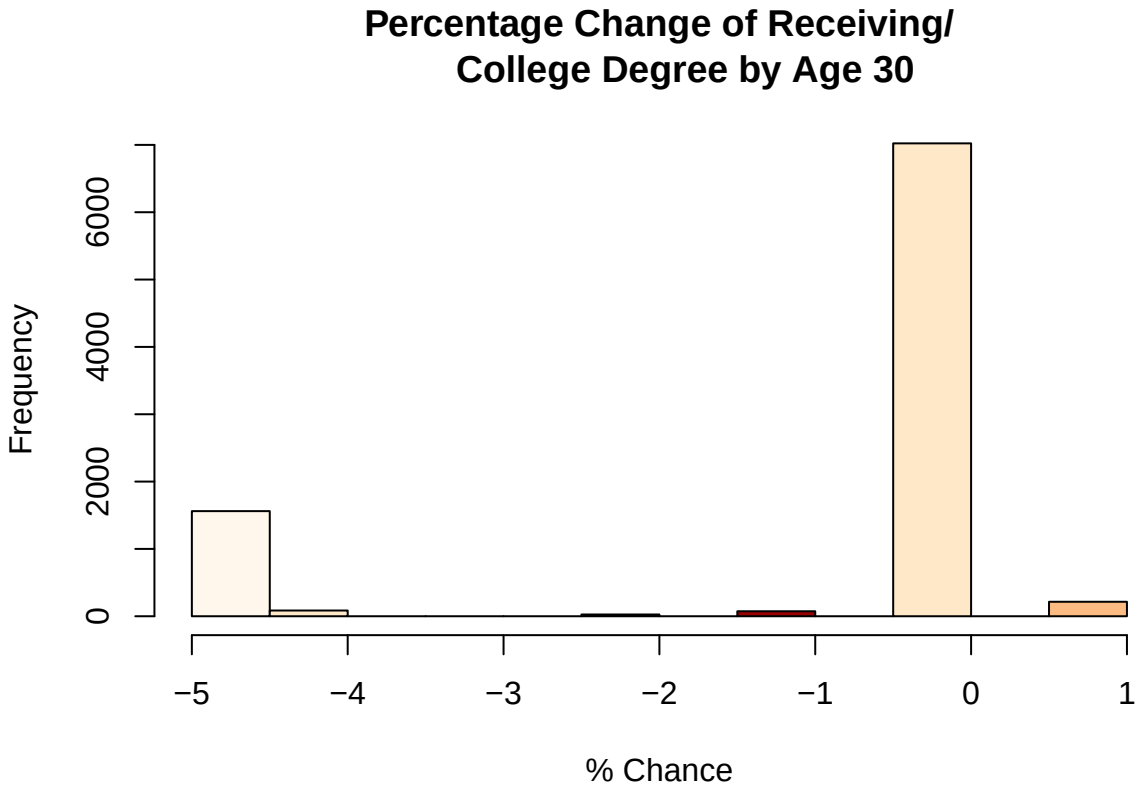
```
hist(renamed.data$collegeDegree.By30,
     main = "Percentage Change of Receiving/
     College Degree by Age 30",
     xlab = "% Chance",
     ylim = c(0,7500),
     col = cbPaletteContrast)
```



I'm not going to use "high school diploma by age 20" variable in my model or explore it any further. More than half of the data is coded as a missing value, and there are no helpful alternative values that can be imputed or replaced by the missing values.

Summary of Drug & Incarceration Variables

```
hist(renamed.data$daysUse.hardDrugs2011,
     main = "Percentage Change of Receiving/
     College Degree by Age 30",
     xlab = "% Chance",
     col = cbPaletteContrast)
```




```
summary(renamed.data$incar.totnum)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.1653  0.0000 12.0000
```

Luckily only a small percentage of the "Used hard drugs since the last interview" and none of the "total number of incarcerations" variables have missing values. We'll code all of the missing values to NA.

Cleaning the Data

Since the analysis is testing whether there is a statistically significant difference between income across genders, it's not helpful to include refusal, don't know, valid skip, and non-interview entries in this variable. It's out of scope to test whether there are correlations in missing income information and other variables like age or gender.

Therefore I will transform missing values in the respondent income variable to NA, then use na.omit to do listwise deletion.

I'll convert all other missing values in the other variables to NA after removing all entries with missing income information.

Income level data

```
cleanIncome.data <-renamed.data
cleanIncome.data <-transform(renamed.data, income.lastYear = mapvalues(income.lastYear, c(-5,-4,-3,-2,-
```

```
## The following `from` values were not present in `x`: -3
```

```
summary(cleanIncome.data$income.lastYear)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0   18000   30000   34160   45000  146000    3682
```

```
newdata <-na.omit(cleanIncome.data)
```

After cleaning the respondent income variable and done a listwise deletion, I've removed 3682 rows of data from the dataset. We now have 5302 rows in the dataset.

Wealth & Job Variables

```
newdata <-transform(newdata, debtAge20 = mapvalues(debtAge20, c(-5,-4,-3,-2,-1), c(NA,NA,NA,NA,NA)),
                    household.net.worth.Parent = mapvalues(household.net.worth.Parent, c(-3,-4),c(NA,NA
```

```
## The following `from` values were not present in `x`: -5, -2, -1
```

The accepted values of the debt variables is only above zero. We'll remove the negative entries, which are the missing values. However we will accept the negative values in the "household net worthaccording to parent" variable. This reflects that a person may have negative equity in large assets or an amount of debt that eclipses positive value assets.

Drug Variables

```
newdata <-transform(newdata,
                    daysUse.hardDrugs2011 = as.factor(mapvalues(daysUse.hardDrugs2011, c(-1,-2,-4,-5,1,
```

```
## The following `from` values were not present in `x`: -5
```

For the binary "if the respondent has used hard drugs since the last interview"" (days.hardDrugs2011) variable, I will be grouping the refuse and valid skip responses together. The non-interview and don't know values will be coded to NA.

Demographic Variables

Let's change the values of my important factors, race and gender, to accurately represent their values. we'll need to do this to make sure R treats these variables accurately. For example: there's no "1" or "2" race. 1,2,3, and 4 represent white, black, other race.

```
newdata <- transform(newdata,
                          race = as.factor(mapvalues(race, c(1,2,3,4), c("black","Hispanic","mixed","no
             gender.Youth = as.factor(mapvalues(gender.Youth,
                              c(1,2), c("Men", "Women"))),
                              marital.status = as.factor(mapvalues(marital.status, c(0,1,2,3,4,-3,-5),
```

```
## The following `from` values were not present in `x`: -5
```

Let's do this process again for the other variables that I assume will have some affect in detecting the significant difference between incomes across genders.

Education Variables

```
newdata <- transform(newdata,
                     highestDegreePrior2011 = as.factor(mapvalues(highestDegreePrior2011, c(0,1,2,3,4,5
                     collegeDegree.By30= mapvalues(collegeDegree.By30, c(-1,-2,-3,-4,-5), c(NA,NA,NA,NA
```

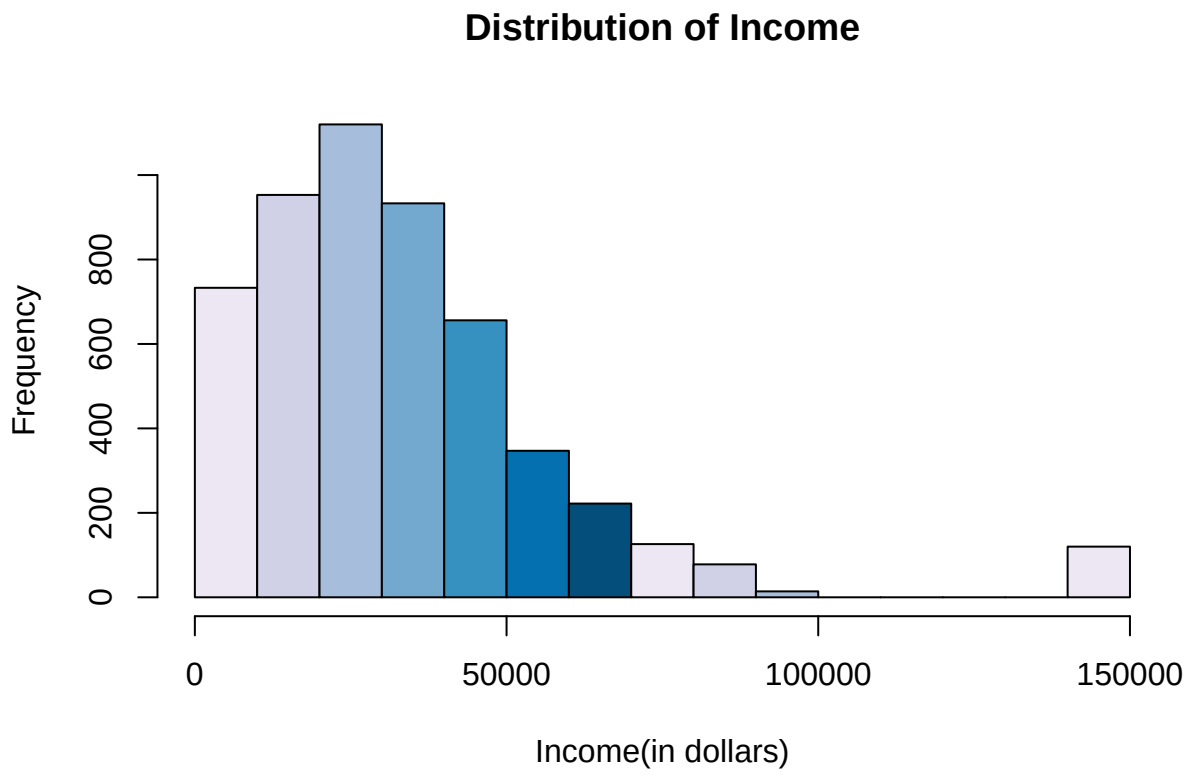
```
## The following `from` values were not present in `x`: -5
## The following `from` values were not present in `x`: -5
```

The missing values of "highest degree prior to 2011" will be recoded as NAs. All of the missing values in the "percentage chance of having a college degree by 30" (collegeDegree.By30) variable will be grouped in NA. I

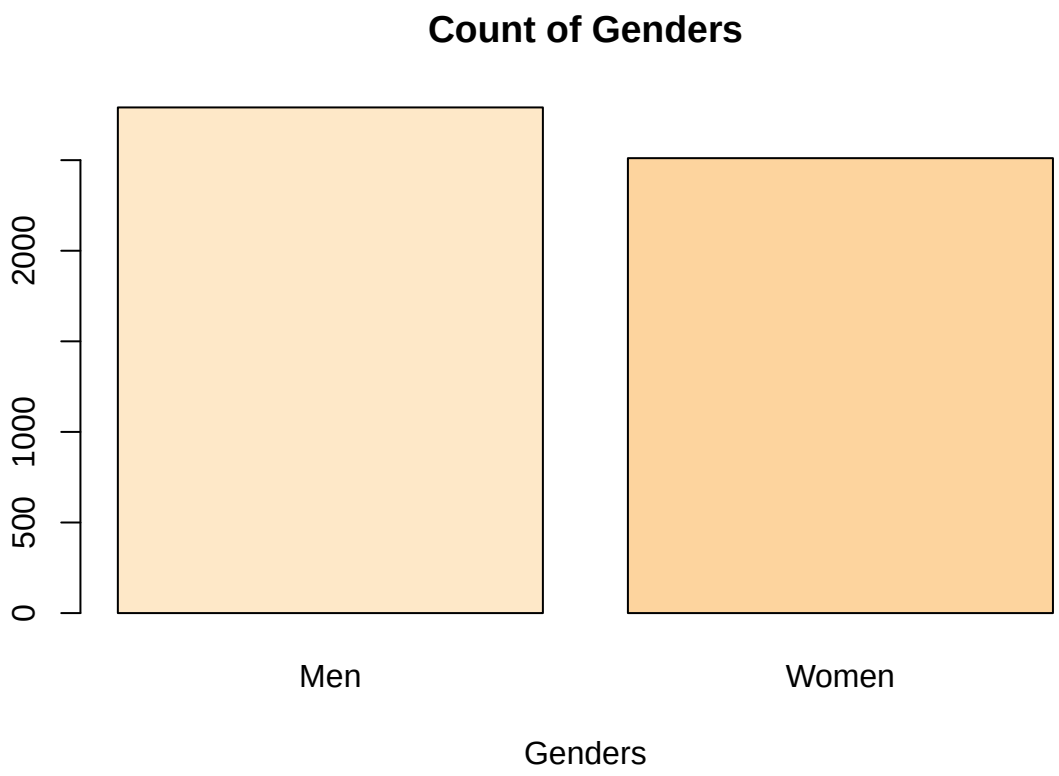
don't have enough information to make an assumption as to why a large portion of the sample decided to skip this question. Therefore I can't code this response as anything other than NA.

Here's the new distribution of my income and gender after cleaning the data.

```
hist(newdata$income.lastYear,
     main= "Distribution of Income",
     xlab="Income(in dollars)",
     col = cbPalette[2:8])
```




```
plot(newdata$gender.Youth,
     main= "Count of Genders",
     xlab="Genders",
     col = cbPaletteContrast[2:8])
```



The income data is skewed to the right with high outliers. Most likely the last bin on the right includes topcoded income values.

Comparing Averages

Table showing mean income across genders

```
gender.incomeTable <- with(newdata, aggregate(income.lastYear, by =list(gender.Youth), FUN=mean))
gender.income.SdTable <- with(newdata, aggregate(income.lastYear, by =list(gender.Youth), FUN=sd))
Men.mean.income <- round(gender.income.SdTable[[1,2]], digits = 2)
Women.mean.income <-round(gender.income.SdTable[[2,2]], digits = 2)
Men.sd.income <- round(gender.income.SdTable[[1,2]], digits = 2)
Women.sd.income <-round(gender.income.SdTable[[2,2]], digits = 2)
```

```
gender.incomeTable <- kable(with(newdata, aggregate(income.lastYear, by =list(gender.Youth), FUN=mean),
gender.income.SdTable <- kable(with(newdata, aggregate(income.lastYear, by =list(gender.Youth), FUN=sd)
```

A comparison of the means of income across genders shows that Men have a mean income of $27911.22 and Women have a mean income of $22335.51, a difference of $5575.71.

Therefore my hypothesis is Men's income have a statistically significantly different income than Women's*

The standard deviation of Women's income($22335.51) is less than Men's($27911.22), showing that Women's income's have less variation.

Is the trend in average income between Men and Women the same across race groups?

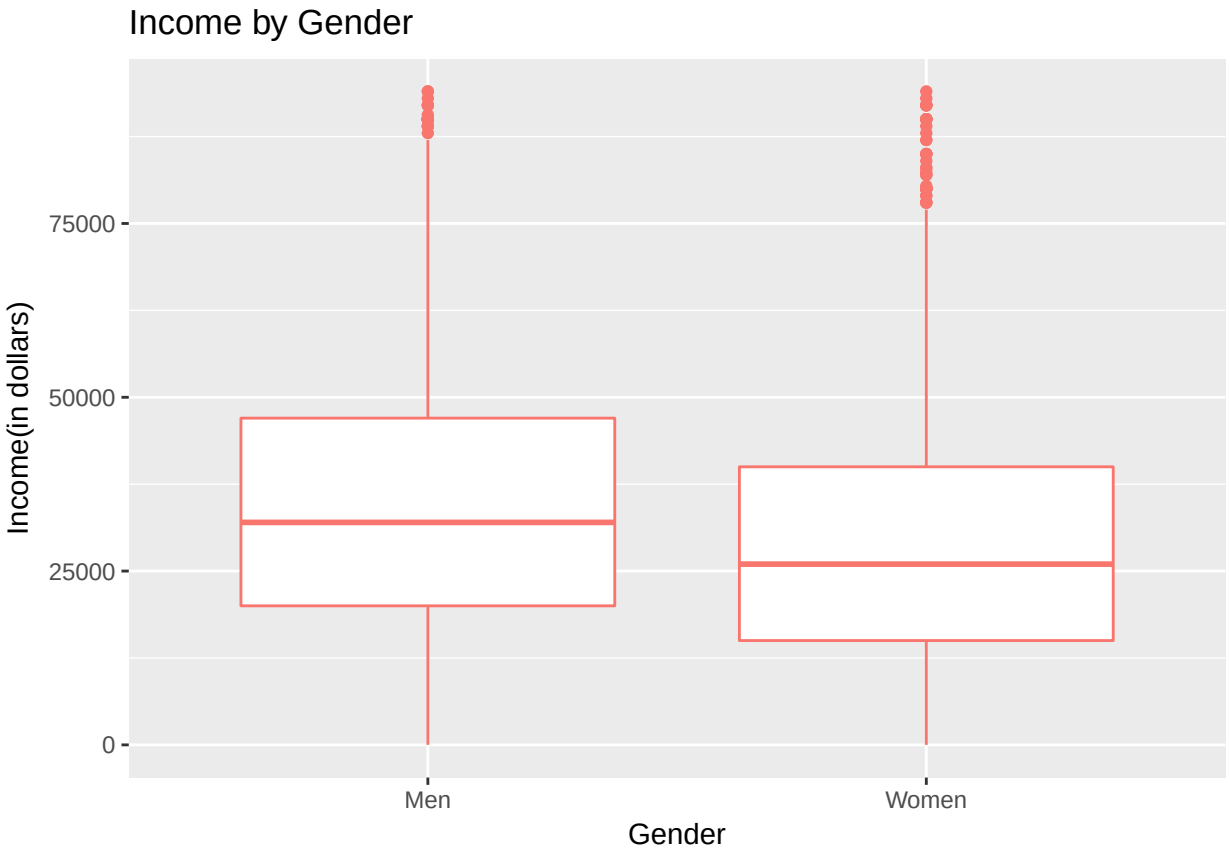
Exploring the Data

```
kable(with(newdata, tapply(income.lastYear, INDEX =list(gender.Youth,race), FUN=mean), format="markdown
```

	black	Hispanic	mixed	non-black
Men	29743.99	35193.49	45866.80	41824.27
Women	25885.15	26762.86	30814.29	33159.35

Looking at the means of income across genders and race, all mens' mean incomes, regardless of race, are higher than Womens'.

```
income.sub <-subset(newdata, subset = income.lastYear<146000)
gender.box <- ggplot(income.sub, aes(x=gender.Youth, y=income.lastYear, color=cbPaletteContrast[1]))
gender.box+geom_boxplot()+theme(legend.position="none")+
  ggtitle("Income by Gender")+
      xlab("Gender")+
  ylab("Income(in dollars)")
```

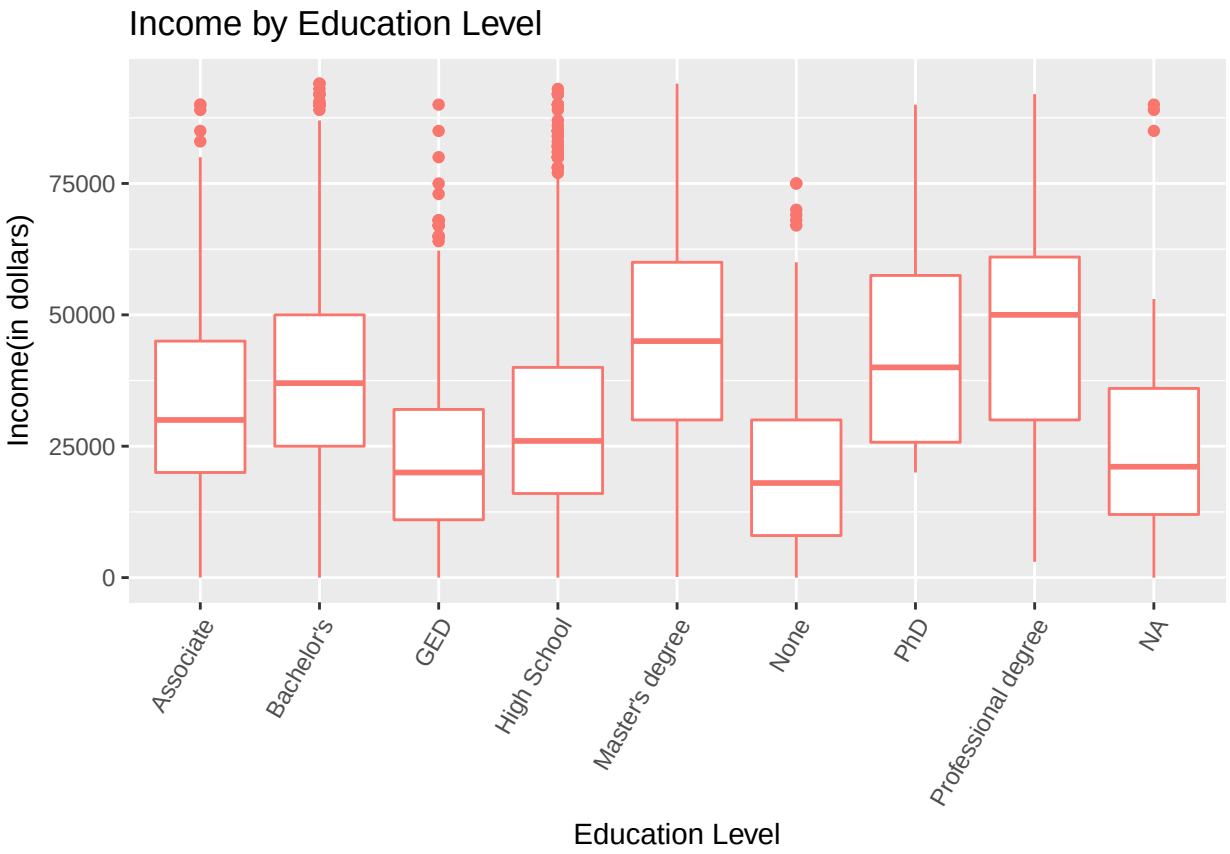


Note: I've removed the extreme high outliers (multiple values of $146,000) of income.

From this box plot which compares income against gender, the inner quartile range of Men's income has a higher range than Women. There's a visual difference between the genders, but is there a statistically significant difference?

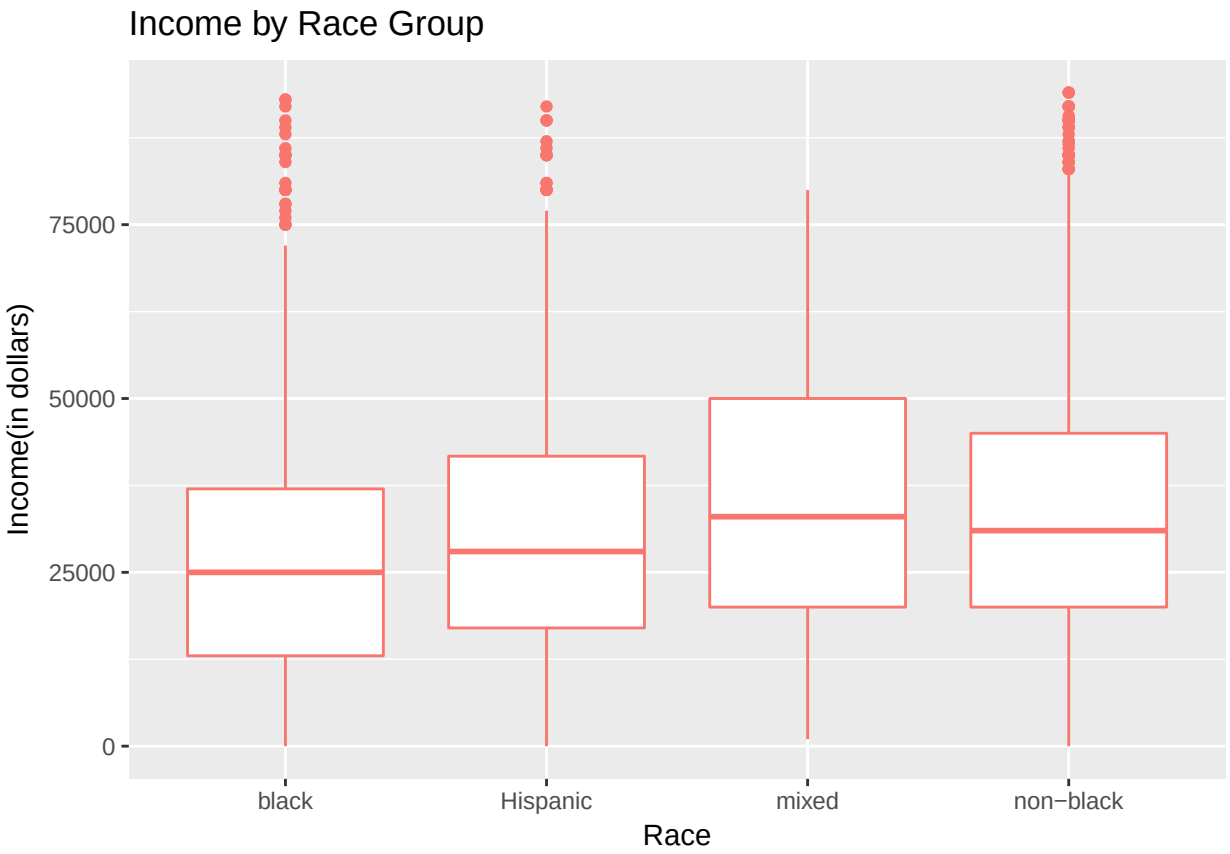
Let's look at the distribution of income against other variables.

```
education.box <- ggplot(income.sub, aes(x=highestDegreePrior2011, y=income.lastYear, color=cbPalette[1]))
education.box+geom_boxplot()+theme(legend.position="none")+
  ggtitle("Income by Education Level")+
      xlab("Education Level")+
  ylab("Income(in dollars)")+ theme(axis.text.x = element_text(angle = 60, vjust = 1, hjust = 1))
```



As one could assume, respondents with higher degrees (Bachelor's, Master's, Professional, and PhD) have a higher median and set of quartiles for income than lower degrees (Associate, GED, None, and High School Diploma). If Men and Women have a disproportionate amount of higher degrees, this may also be contributing to any disparity across the genders.

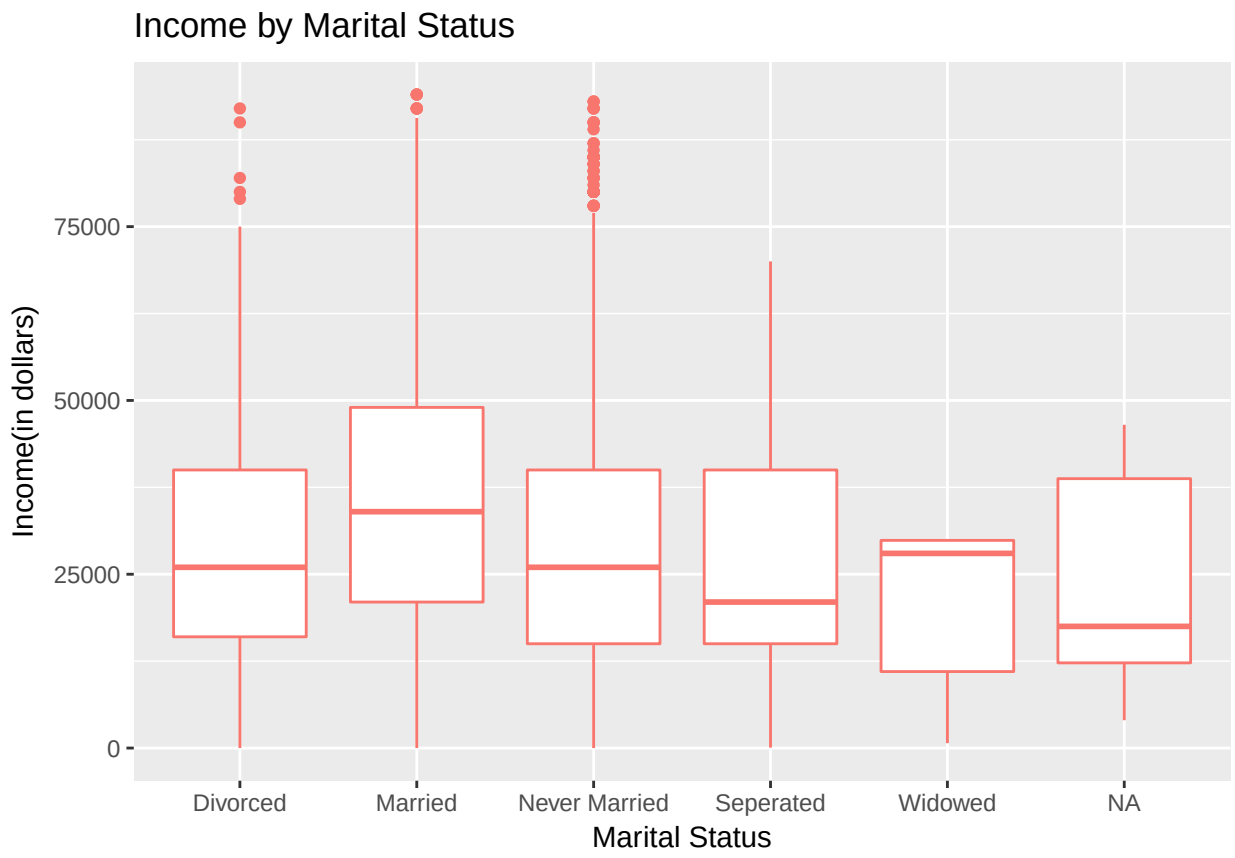
```
race.box <- ggplot(income.sub, aes(x=race, y=income.lastYear, color=cbPalette[4]))
race.box+geom_boxplot()+theme(legend.position="none")+
  ggtitle("Income by Race Group")+
      xlab("Race")+
  ylab("Income(in dollars)")
```



Note: I've removed the extreme high outliers (multiple values of $146,000) of income.

We can see the same trend in income across race groups. Race will be another variable to analyze whether it effects income levels.

```
married.box <- ggplot(income.sub, aes(x=marital.status, y=income.lastYear, color=cbPalette[5]))
married.box+geom_boxplot()+theme(legend.position="none")+
  ggtitle("Income by Marital Status")+
      xlab("Marital Status")+
  ylab("Income(in dollars)")
```



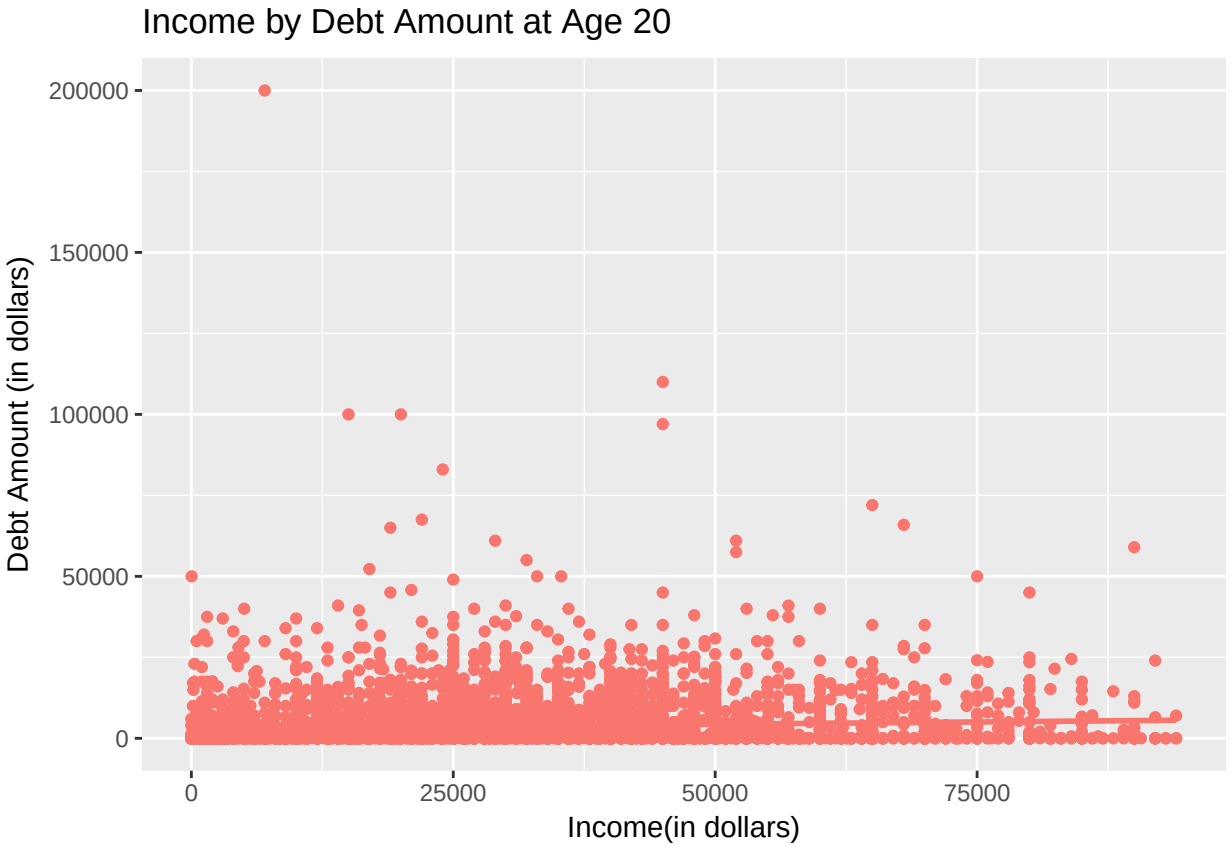
Note: I've removed the extreme high outliers (multiple values of $146,000) of income.

The Widowed and NA respondents have highly skewed income. In the case of the Widowed respondents, the median income is almost the same as the income at the 75% percentile. Over all, Married, Divorced, and Never Married have a higher and more normally distributed income than the other martial statuses.

```
income.debt <- ggplot(income.sub, aes(x=income.lastYear, y= debtAge20, color=cbPalette[2]))
income.debt +geom_point()+theme(legend.position="none")+
  ggtitle("Income by Debt Amount at Age 20")+
  geom_smooth(method = "lm") +
  xlab("Income(in dollars)") +
  ylab("Debt Amount (in dollars)")
```

```
## Warning: Removed 121 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 121 rows containing missing values (geom_point).
```



Note: This scatter plot removed 127 rows were NA values were entered in the "debt at age 20" variable and I've removed the income variable's high outliers (multiple values of $146,000).

From the more concentrated grouping at the axis of this scatterplot, I can infer that those with relatively lower amounts of income have higher amounts of debt. However, the (hard to see) regression line is flat. There may be no relationship between debt and income.

Linear Regression Model

I will construct a model based on the most important variable (gender) and progressively add the other variables we've explored to weed out any interacting variables and minimize omitted variable bias.

First is a model with income (topcoded values removed) across genders.

```
basis.lm <- lm(income.lastYear ~ as.factor(gender.Youth), data=newdata)
summary(basis.lm)
```

```
##
## Call:
## lm(formula = income.lastYear ~ as.factor(gender.Youth), data = newdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37912 -15998  -4912  10088 116004
##
## Coefficients:
```

```
##                               Estimate Std. Error t value
## (Intercept)                    37911.6      481.2   78.78
## as.factor(gender.Youth)Women  -7913.7      699.3  -11.32
##                                          Pr(>|t|)
## (Intercept)                  <0.0000000000000002 ***
## as.factor(gender.Youth)Women <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25420 on 5300 degrees of freedom
## Multiple R-squared:  0.02359,    Adjusted R-squared:  0.02341
## F-statistic: 128.1 on 1 and 5300 DF,  p-value: < 0.00000000000000022
```

Men were used as a base line in this model. This model's coefficient for gender shows that given a person is a woman, they would make on average $7,914 less than a man. Gender is also a highly significant predictor of income. Even at the very low test threshold of .001 (compared to the standard level of .05), the difference in income across genders is significant. Note that the standard error for the gender variable is only $699.30. This signifies that my coefficient is a pretty strong estimate.

Let's look at income across gender and race.

```
race.lm <- lm(income.lastYear ~ as.factor(gender.Youth)+as.factor(race), data=newdata)
summary(race.lm)
```

```
##
## Call:
## lm(formula = income.lastYear ~ as.factor(gender.Youth) + as.factor(race),
##     data = newdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41329 -16329  -3802   9921 121923
##
## Coefficients:
##                               Estimate Std. Error t value
## (Intercept)                    31668.6      809.3  39.133
## as.factor(gender.Youth)Women  -7589.2      690.7 -10.987
## as.factor(race)Hispanic        3133.5     1037.9   3.019
## as.factor(race)mixed          11125.0     3587.7   3.101
## as.factor(race)non-black       9660.1      863.4  11.188
##                                          Pr(>|t|)
## (Intercept)                  < 0.0000000000000002 ***
## as.factor(gender.Youth)Women < 0.0000000000000002 ***
## as.factor(race)Hispanic                   0.00255 **
## as.factor(race)mixed                      0.00194 **
## as.factor(race)non-black     < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25090 on 5297 degrees of freedom
## Multiple R-squared:  0.04996,    Adjusted R-squared:  0.04924
## F-statistic: 69.64 on 4 and 5297 DF,  p-value: < 0.00000000000000022
```

The summary of this model shows that race didn't affect the significance level of gender in the model. Futhuremore, the listed p-values show that all levels of race have a statistically significant difference in average income against the baselines. In this model, men and black were used as the baseline. We can read this as: if

everything is held constant, if the respondent is Hispanic, mixed race, or non-black they earned earn more than black respondents.

My multiple and adjusted R-square has gotten larger, which signifies that this model is worse fit than only including gender in the model.

Let's look at an income model that includes gender, race, and marital status.

```
married.lm <- lm(income.lastYear ~ as.factor(gender.Youth) + as.factor(race)+as.factor(marital.status),
summary(married.lm)
```

```
##
## Call:
## lm(formula = income.lastYear ~ as.factor(gender.Youth) + as.factor(race) +
##     as.factor(marital.status), data = newdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -45422 -16089  -4245   9755 124225
##
## Coefficients:
##                                       Estimate Std. Error t value
## (Intercept)                            29491.5     1592.7  18.517
## as.factor(gender.Youth)Women           -7714.9      687.3 -11.225
## as.factor(race)Hispanic                 2155.7     1037.3   2.078
## as.factor(race)mixed                   10451.4     3557.7   2.938
## as.factor(race)non-black                8231.5      871.2   9.449
## as.factor(marital.status)Married        7699.0     1470.2   5.237
## as.factor(marital.status)Never Married   597.5     1447.9   0.413
## as.factor(marital.status)Seperated     -3343.9     3214.5  -1.040
## as.factor(marital.status)Widowed       -7987.4     9495.8  -0.841
##                                                  Pr(>|t|)
## (Intercept)                           < 0.0000000000000002 ***
## as.factor(gender.Youth)Women          < 0.0000000000000002 ***
## as.factor(race)Hispanic                             0.03774 *
## as.factor(race)mixed                                0.00332 **
## as.factor(race)non-black              < 0.0000000000000002 ***
## as.factor(marital.status)Married                 0.00000017 ***
## as.factor(marital.status)Never Married              0.67989
## as.factor(marital.status)Seperated                  0.29827
## as.factor(marital.status)Widowed                    0.40030
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24850 on 5283 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.06889,    Adjusted R-squared:  0.06748
## F-statistic: 48.86 on 8 and 5283 DF,  p-value: < 0.00000000000000022
```

The Married marital status is the only level of the marital status variable that's significant. Holding all else constant and against the baseline, which is black men that have a marital status of NA, married respondents earn $5.24 more. Also, the Hispanic race level was lowered to the standard .05 level of significance.

Separated and Widowed respondents have negative coefficents, which means these respondents earn around $1 less than the other marital groups. This is interesting considering I'm analyzing an income, not household net worthvariable. Maybe there are underlying psychological reasons as to why being separated and widowed correlates to an individual earning less? Whatever the case, the relationship between these particular marital

groups and income is not statistically significant.

my multiple and adjusted R-square has even larger, which signifies that this model is an even worse fit than only including gender and race in the model.

What contributes to the Income Gap?

Gender doesn't lose its significance or change the polarization of the coefficient with the inclusion of marital status, but it lessens a bit. It would be interesting to explore whether marital status has any interaction with gender. Maybe being a Women and married will mitigate the income gap with men?

```
married.interact.lm <- lm(income.lastYear ~ as.factor(gender.Youth)*as.factor(marital.status) + as.facto
summary(married.interact.lm)
```

```
##
## Call:
## lm(formula = income.lastYear ~ as.factor(gender.Youth) * as.factor(marital.status) +
##     as.factor(race), data = newdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -48252 -16193  -4200  10190 125280
##
## Coefficients:
##                                                                  Estimate
## (Intercept)                                                       30206.6
## as.factor(gender.Youth)Women                                      -9484.7
## as.factor(marital.status)Married                                   9426.5
## as.factor(marital.status)Never Married                            -2401.8
## as.factor(marital.status)Seperated                                -4256.5
## as.factor(marital.status)Widowed                                  -1716.6
## as.factor(race)Hispanic                                            2529.9
## as.factor(race)mixed                                              11273.8
## as.factor(race)non-black                                           8618.7
## as.factor(gender.Youth)Women:as.factor(marital.status)Married     -3957.3
## as.factor(gender.Youth)Women:as.factor(marital.status)Never Married  6254.6
## as.factor(gender.Youth)Women:as.factor(marital.status)Seperated    1556.2
## as.factor(gender.Youth)Women:as.factor(marital.status)Widowed    -10836.2
##                                                                  Std. Error
## (Intercept)                                                          2230.0
## as.factor(gender.Youth)Women                                         2755.9
## as.factor(marital.status)Married                                     2246.9
## as.factor(marital.status)Never Married                               2206.9
## as.factor(marital.status)Seperated                                   4925.6
## as.factor(marital.status)Widowed                                    14450.0
## as.factor(race)Hispanic                                              1034.2
## as.factor(race)mixed                                                 3546.4
## as.factor(race)non-black                                              869.2
## as.factor(gender.Youth)Women:as.factor(marital.status)Married        2964.5
## as.factor(gender.Youth)Women:as.factor(marital.status)Never Married  2911.3
## as.factor(gender.Youth)Women:as.factor(marital.status)Seperated      6482.6
## as.factor(gender.Youth)Women:as.factor(marital.status)Widowed       19103.4
##                                                                  t value
## (Intercept)                                                       13.545
## as.factor(gender.Youth)Women                                      -3.442
## as.factor(marital.status)Married                                   4.195
```

```
## as.factor(marital.status)Never Married                                  -1.088
## as.factor(marital.status)Seperated                                      -0.864
## as.factor(marital.status)Widowed                                        -0.119
## as.factor(race)Hispanic                                                  2.446
## as.factor(race)mixed                                                     3.179
## as.factor(race)non-black                                                 9.916
## as.factor(gender.Youth)Women:as.factor(marital.status)Married           -1.335
## as.factor(gender.Youth)Women:as.factor(marital.status)Never Married      2.148
## as.factor(gender.Youth)Women:as.factor(marital.status)Seperated          0.240
## as.factor(gender.Youth)Women:as.factor(marital.status)Widowed           -0.567
##                                                                             Pr(>|t|)
## (Intercept)                                                        < 0.0000000000000002
## as.factor(gender.Youth)Women                                                0.000583
## as.factor(marital.status)Married                                           0.0000277
## as.factor(marital.status)Never Married                                      0.276506
## as.factor(marital.status)Seperated                                          0.387540
## as.factor(marital.status)Widowed                                            0.905444
## as.factor(race)Hispanic                                                     0.014470
## as.factor(race)mixed                                                        0.001487
## as.factor(race)non-black                                           < 0.0000000000000002
## as.factor(gender.Youth)Women:as.factor(marital.status)Married               0.181978
## as.factor(gender.Youth)Women:as.factor(marital.status)Never Married         0.031729
## as.factor(gender.Youth)Women:as.factor(marital.status)Seperated             0.810293
## as.factor(gender.Youth)Women:as.factor(marital.status)Widowed               0.570576
##
## (Intercept)                                                        ***
## as.factor(gender.Youth)Women                                       ***
## as.factor(marital.status)Married                                   ***
## as.factor(marital.status)Never Married
## as.factor(marital.status)Seperated
## as.factor(marital.status)Widowed
## as.factor(race)Hispanic                                            *
## as.factor(race)mixed                                               **
## as.factor(race)non-black                                           ***
## as.factor(gender.Youth)Women:as.factor(marital.status)Married
## as.factor(gender.Youth)Women:as.factor(marital.status)Never Married *
## as.factor(gender.Youth)Women:as.factor(marital.status)Seperated
## as.factor(gender.Youth)Women:as.factor(marital.status)Widowed
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24740 on 5279 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.07781,    Adjusted R-squared:  0.07571
## F-statistic: 37.12 on 12 and 5279 DF,  p-value: < 0.00000000000000022
```

Never married Women is the only interaction level that has a statistical significance. Unlike the main effect of the Women level, never married Women have a positive relationship compared to the Men baseline. Holding all things constant, this group earns $6,254.60 more than black men with NA marital status.

Another interaction to consider is race and gender. I saw during the exploratory analysis that the income gap existed for Men and Women across all race groups. Let's test if it's a statistically significant exacerbate of the income gap.

```r
race.interact.lm <- lm(income.lastYear ~ as.factor(gender.Youth)*as.factor(race), data=newdata)
summary(race.interact.lm)
```

```
##
## Call:
## lm(formula = income.lastYear ~ as.factor(gender.Youth) * as.factor(race),
##     data = newdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41824 -16191  -3885   9841 120117
##
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                         29744       1044
## as.factor(gender.Youth)Women                        -3859       1453
## as.factor(race)Hispanic                              5450       1454
## as.factor(race)mixed                                16123       4695
## as.factor(race)non-black                            12080       1220
## as.factor(gender.Youth)Women:as.factor(race)Hispanic    -4572       2076
## as.factor(gender.Youth)Women:as.factor(race)mixed      -11194       7280
## as.factor(gender.Youth)Women:as.factor(race)non-black   -4806       1726
##                                                  t value
## (Intercept)                                       28.499
## as.factor(gender.Youth)Women                      -2.656
## as.factor(race)Hispanic                            3.749
## as.factor(race)mixed                               3.434
## as.factor(race)non-black                           9.898
## as.factor(gender.Youth)Women:as.factor(race)Hispanic  -2.202
## as.factor(gender.Youth)Women:as.factor(race)mixed     -1.538
## as.factor(gender.Youth)Women:as.factor(race)non-black  -2.784
##                                                            Pr(>|t|)
## (Intercept)                                       < 0.0000000000000002
## as.factor(gender.Youth)Women                                 0.007937
## as.factor(race)Hispanic                                      0.000179
## as.factor(race)mixed                                         0.000599
## as.factor(race)non-black                          < 0.0000000000000002
## as.factor(gender.Youth)Women:as.factor(race)Hispanic          0.027731
## as.factor(gender.Youth)Women:as.factor(race)mixed             0.124180
## as.factor(gender.Youth)Women:as.factor(race)non-black         0.005389
##                                                  
## (Intercept)                                      ***
## as.factor(gender.Youth)Women                     **
## as.factor(race)Hispanic                          ***
## as.factor(race)mixed                             ***
## as.factor(race)non-black                         ***
## as.factor(gender.Youth)Women:as.factor(race)Hispanic  *
## as.factor(gender.Youth)Women:as.factor(race)mixed
## as.factor(gender.Youth)Women:as.factor(race)non-black **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25070 on 5294 degrees of freedom
## Multiple R-squared:  0.05163,    Adjusted R-squared:  0.05038
```

```
## F-statistic: 41.17 on 7 and 5294 DF,  p-value: < 0.00000000000000022
```

These results show that there's no significant difference between black men (this model's baseline) and mixed Women in term of income. This interaction term also has a large standard error so the coefficient estimate is not very strong. However there is a significant difference between black men and all other Women groups' incomes. The coefficient for non-black Women show that they make $4,806 less than black men. Note that the standard error for Hispanic Women is almost half of its coefficient. This cautions me to use this coefficient as a good estimate in the difference between Hispanic Women's income and black men's.

```
age.interact.lm <- lm(income.lastYear ~ as.factor(gender.Youth)*as.factor(birthYear.Youth), data=newdata
summary(age.interact.lm)
```

```
##
## Call:
## lm(formula = income.lastYear ~ as.factor(gender.Youth) * as.factor(birthYear.Youth),
##     data = newdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -43594 -16179  -4276  10353 118819
##
## Coefficients:
##                                                          Estimate
## (Intercept)                                                 43594
## as.factor(gender.Youth)Women                               -11810
## as.factor(birthYear.Youth)1981                              -2427
## as.factor(birthYear.Youth)1982                              -5194
## as.factor(birthYear.Youth)1983                              -9318
## as.factor(birthYear.Youth)1984                             -10947
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1981   3180
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1982   4422
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1983   5072
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1984   6346
##                                                          Std. Error
## (Intercept)                                                    1121
## as.factor(gender.Youth)Women                                   1618
## as.factor(birthYear.Youth)1981                                 1533
## as.factor(birthYear.Youth)1982                                 1541
## as.factor(birthYear.Youth)1983                                 1543
## as.factor(birthYear.Youth)1984                                 1542
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1981    2224
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1982    2228
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1983    2228
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1984    2243
##                                                          t value
## (Intercept)                                               38.875
## as.factor(gender.Youth)Women                              -7.300
## as.factor(birthYear.Youth)1981                            -1.583
## as.factor(birthYear.Youth)1982                            -3.370
## as.factor(birthYear.Youth)1983                            -6.038
## as.factor(birthYear.Youth)1984                            -7.099
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1981  1.429
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1982  1.984
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1983  2.276
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1984  2.829
```

```
##                                                                    Pr(>|t|)
## (Intercept)                                                   < 0.0000000000000002
## as.factor(gender.Youth)Women                                     0.000000000000329
## as.factor(birthYear.Youth)1981                                           0.113452
## as.factor(birthYear.Youth)1982                                           0.000757
## as.factor(birthYear.Youth)1983                                   0.000000001669033
## as.factor(birthYear.Youth)1984                                   0.000000000001419
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1981              0.152960
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1982              0.047292
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1983              0.022877
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1984              0.004682
##
## (Intercept)                                                   ***
## as.factor(gender.Youth)Women                                  ***
## as.factor(birthYear.Youth)1981
## as.factor(birthYear.Youth)1982                                ***
## as.factor(birthYear.Youth)1983                                ***
## as.factor(birthYear.Youth)1984                                ***
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1981
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1982 *
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1983 *
## as.factor(gender.Youth)Women:as.factor(birthYear.Youth)1984 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25230 on 5292 degrees of freedom
## Multiple R-squared:  0.0402, Adjusted R-squared:  0.03856
## F-statistic: 24.63 on 9 and 5292 DF,  p-value: < 0.00000000000000022
```

In this model three out of the four years have a statistically different income from men's income, which is the baseline. Interesting enough the incomes for Women in all four interacting terms are positive. For "Womne:as.factor(birthYear.Youth)1982", the coefficients translate if everything is held constant, Women who were born in 1982 make $4,422 more than men born in the same year.

I can conclude that age mitigates the income gap between genders.

I won't run an interaction model on the debt variable. It will return a long and unhelpful list of results

Let's look at a income model that includes gender, race, if the individual has used hard drugs since the last interview, the amount of debt at age 20, the household net worthaccording to the respondent's parent.

```
drug.wealth.marital.lm <- lm(income.lastYear ~ as.factor(gender.Youth) + as.factor(race)+as.factor(daysU
summary(drug.wealth.marital.lm)
```

```
##
## Call:
## lm(formula = income.lastYear ~ as.factor(gender.Youth) + as.factor(race) +
##     as.factor(daysUse.hardDrugs2011) + debtAge20 + household.net.worth.Parent +
##     marital.status, data = newdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59698 -14856  -3261   9350 122479
##
## Coefficients:
##                                          Estimate    Std. Error t value
## (Intercept)                           26307.074016   1760.111475  14.946
```

```
## as.factor(gender.Youth)Women            -7659.493205    768.714362  -9.964
## as.factor(race)Hispanic                  2613.590585   1175.443012   2.223
## as.factor(race)mixed                     6039.495507   3910.064117   1.545
## as.factor(race)non-black                 4637.591432   1026.423302   4.518
## as.factor(daysUse.hardDrugs2011)refuse -14049.428798   4294.338259  -3.272
## as.factor(daysUse.hardDrugs2011)yes     -2992.315042   2397.377876  -1.248
## debtAge20                                   0.149854      0.046311   3.236
## household.net.worth.Parent                  0.037033      0.002849  13.001
## marital.statusMarried                    7861.809372   1603.024811   4.904
## marital.statusNever Married               994.949145   1577.786170   0.631
## marital.statusSeperated                 -2528.503918   3504.594720  -0.721
## marital.statusWidowed                   -5790.100273   9109.059679  -0.636
##                                                    Pr(>|t|)
## (Intercept)                           < 0.0000000000000002 ***
## as.factor(gender.Youth)Women          < 0.0000000000000002 ***
## as.factor(race)Hispanic                              0.02624 *
## as.factor(race)mixed                                 0.12252
## as.factor(race)non-black                        0.000006423 ***
## as.factor(daysUse.hardDrugs2011)refuse               0.00108 **
## as.factor(daysUse.hardDrugs2011)yes                  0.21205
## debtAge20                                            0.00122 **
## household.net.worth.Parent            < 0.0000000000000002 ***
## marital.statusMarried                           0.000000976 ***
## marital.statusNever Married                          0.52834
## marital.statusSeperated                              0.47066
## marital.statusWidowed                                0.52505
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23770 on 3865 degrees of freedom
##   (1424 observations deleted due to missingness)
## Multiple R-squared:  0.1165, Adjusted R-squared:  0.1138
## F-statistic: 42.49 on 12 and 3865 DF,  p-value: < 0.00000000000000022
```

The summary shows that by adding these new factors, all levels of race have a slightly lower statistical significance. All levels of race, but mixed, remains significant. Furthermore, debt at age 20, household net worthaccording to the respondent's parent, and the refuse response of hard drug use were found to be statistically significantly predictors of income at stringent and moderate levels of testing.

Household net worthaccording to the parent and debt at age 20 are significant, but have small scale effects on income. Holding all things constant, with every dollar of debt, income increases by 12 cents. Holding all things constant, with every dollar of parent household net worth, income increases by 3 cents. The relationship between income and debt seems intuitive. If a respondent came from a poor household, the respondent most likely needed to take out debt to pay for an education or a business loan to achieve a higher income. It's easy to assume that individuals who have a high amount of debt and those with low and moderate incomes are highly correlated.

In terms of parent household net worth, this trend makes sense as well. If a respondent was raised in a wealthy household, they most likely have access to opportunities and education that will afford them a higher income. Income and parent household net worthhas a positive, significant, but surprisingly small scale relationship.

Not surprisingly, affirmative and refuse responses for the "if you used hard drugs since the last interview" have a negative coefficient and therefore, negative relationship with income. It would be easy to assume that a portion of respondents that answered this question refused to answer because they didn't want to disclose that they had used drugs. It's hard to imagine why someone who hadn't used drugs would not disclose they didn't, other than a lack of engagement while taking the survey. Holding all things constant, for every person

who refused to answer this question, there was a $.00013 decrease in income. We see the same negative, significant, but slightly larger relationship with income for those that admitted to using drugs.

My multiple and adjusted R-squared measures are nearly the same as the marital status, race, and gender model. So this model does no better or worse in terms of predicting income.

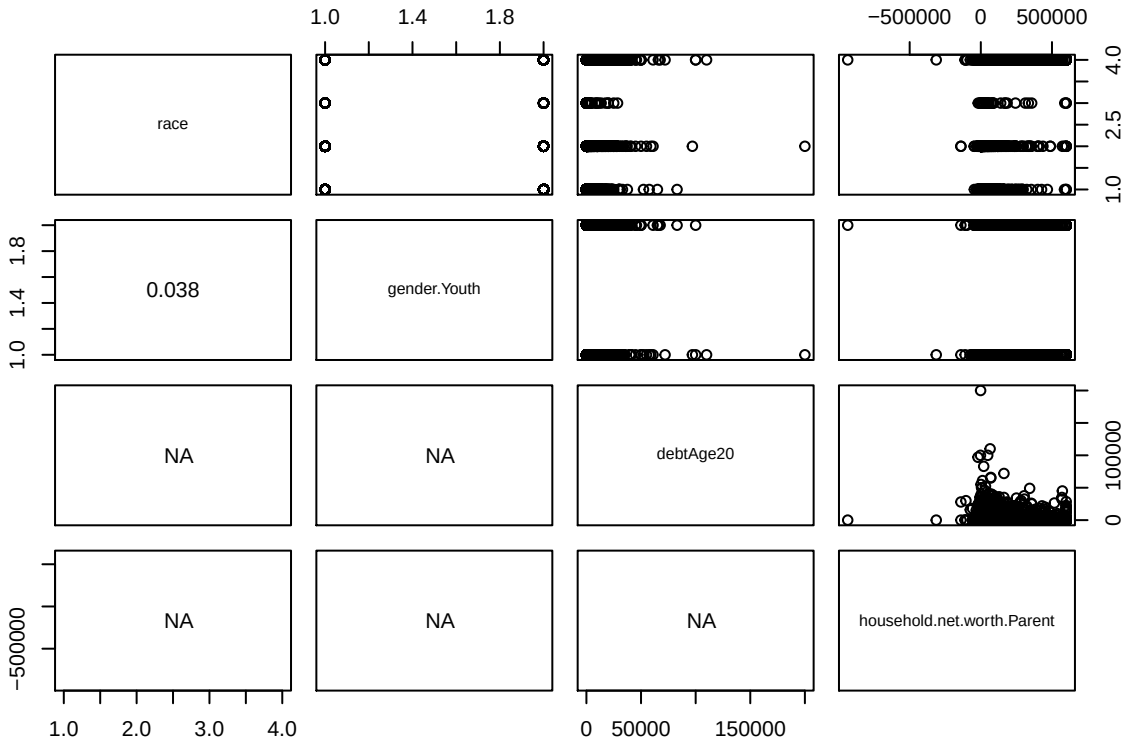
```
age.drug.wealth.lm <- lm(income.lastYear ~ as.factor(gender.Youth) + as.factor(race)+as.factor(daysUse.
summary(age.drug.wealth.lm)
```

```
##
## Call:
## lm(formula = income.lastYear ~ as.factor(gender.Youth) + as.factor(race) +
##     as.factor(daysUse.hardDrugs2011) + debtAge20 + household.net.worth.Parent +
##     as.factor(birthYear.Youth), data = newdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -62417 -14834  -3332   9610 122908
##
## Coefficients:
##                                             Estimate    Std. Error t value
## (Intercept)                                32432.024336   1214.958856  26.694
## as.factor(gender.Youth)Women               -7814.254773    768.216192 -10.172
## as.factor(race)Hispanic                     3476.718068   1167.635462   2.978
## as.factor(race)mixed                        6211.769424   3916.334823   1.586
## as.factor(race)non-black                    6037.109259   1012.536729   5.962
## as.factor(daysUse.hardDrugs2011)refuse    -15882.056067   4304.981917  -3.689
## as.factor(daysUse.hardDrugs2011)yes        -3592.134814   2400.280706  -1.497
## debtAge20                                      0.182875      0.046376   3.943
## household.net.worth.Parent                     0.037248      0.002852  13.061
## as.factor(birthYear.Youth)1981             -1082.659949   1227.998794  -0.882
## as.factor(birthYear.Youth)1982             -3010.675883   1238.772079  -2.430
## as.factor(birthYear.Youth)1983             -6427.367777   1221.307003  -5.263
## as.factor(birthYear.Youth)1984             -7244.294432   1232.641105  -5.877
##                                                       Pr(>|t|)
## (Intercept)                                < 0.0000000000000002 ***
## as.factor(gender.Youth)Women               < 0.0000000000000002 ***
## as.factor(race)Hispanic                                0.002923 **
## as.factor(race)mixed                                   0.112794
## as.factor(race)non-black                         0.00000000271 ***
## as.factor(daysUse.hardDrugs2011)refuse                 0.000228 ***
## as.factor(daysUse.hardDrugs2011)yes                    0.134592
## debtAge20                                        0.00008177080 ***
## household.net.worth.Parent                 < 0.0000000000000002 ***
## as.factor(birthYear.Youth)1981                         0.378023
## as.factor(birthYear.Youth)1982                         0.015128 *
## as.factor(birthYear.Youth)1983                   0.00000014965 ***
## as.factor(birthYear.Youth)1984                   0.00000000453 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23830 on 3875 degrees of freedom
##   (1414 observations deleted due to missingness)
## Multiple R-squared:  0.1105, Adjusted R-squared:  0.1078
## F-statistic: 40.13 on 12 and 3875 DF,  p-value: < 0.00000000000000022
```

Since the Hispanic race level lost its significance in the last model, we should test for collinearity. We want to conclude whether the coefficients the model has produced are imprecise estimates of their effect on income.

Note: When testing for collinearity using "if drugs had been used since the last interview" and "marital status", the matrix returned NA values for these variables.

```
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste0(prefix, txt)
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = pmax(1, cex.cor * r))
}
txt <- c("race","gender.Youth","debtAge20","household.net.worth.Parent")
pairs(newdata[,txt], lower.panel = panel.cor)
```



The values in the lower panel are all less than .30. This is low, therefore I can assume there's no collinearity in the added variables to explain the Hispanic level's change in significance.

Panel may not knit properly. Displays NA values across debt and household net worth row's lower panel in html. Doesn't print an error message either. However it displays properly in plot window after running the chunk in the console. Values under race column are: .038, .059,.28. Under gender.Youth column are .034, .0017. Under the debt Age column .034

I'll test if any of my values statistical significance changes significantly if we remove gender.

```
nogender.lm <- lm(income.lastYear ~ as.factor(race)+as.factor(daysUse.hardDrugs2011)+debtAge20+household
summary(nogender.lm)
```

```
##
## Call:
## lm(formula = income.lastYear ~ as.factor(race) + as.factor(daysUse.hardDrugs2011) +
##     debtAge20 + household.net.worth.Parent + as.factor(birthYear.Youth),
##     data = newdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -55136 -15336  -3769   9380 123360
##
## Coefficients:
##                                            Estimate    Std. Error t value
## (Intercept)                                28315.272870   1160.613408  24.397
## as.factor(race)Hispanic                     3876.297120   1182.299324   3.279
## as.factor(race)mixed                        7084.003367   3966.813362   1.786
## as.factor(race)non-black                    6592.257233   1024.342220   6.436
## as.factor(daysUse.hardDrugs2011)refuse -14397.664761   4359.008838  -3.303
## as.factor(daysUse.hardDrugs2011)yes     -2927.311656   2430.899756  -1.204
## debtAge20                                   0.165480      0.046953   3.524
## household.net.worth.Parent                  0.036889      0.002889  12.768
## as.factor(birthYear.Youth)1981           -987.332426   1244.088774  -0.794
## as.factor(birthYear.Youth)1982          -3096.268580   1255.010812  -2.467
## as.factor(birthYear.Youth)1983          -6269.055158   1237.244863  -5.067
## as.factor(birthYear.Youth)1984          -6988.821189   1248.569024  -5.597
##                                                Pr(>|t|)
## (Intercept)                                < 0.0000000000000002 ***
## as.factor(race)Hispanic                                0.001052 **
## as.factor(race)mixed                                   0.074207 .
## as.factor(race)non-black                     0.000000000138 ***
## as.factor(daysUse.hardDrugs2011)refuse                 0.000965 ***
## as.factor(daysUse.hardDrugs2011)yes                    0.228582
## debtAge20                                              0.000429 ***
## household.net.worth.Parent                 < 0.0000000000000002 ***
## as.factor(birthYear.Youth)1981                         0.427466
## as.factor(birthYear.Youth)1982                         0.013663 *
## as.factor(birthYear.Youth)1983               0.000000423097 ***
## as.factor(birthYear.Youth)1984               0.000000023257 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24150 on 3876 degrees of freedom
##   (1414 observations deleted due to missingness)
## Multiple R-squared:  0.08679,    Adjusted R-squared:  0.0842
## F-statistic: 33.49 on 11 and 3876 DF,  p-value: < 0.00000000000000022
```

Overall, not including the gender variable improves the fit of our model a bit (by .02, as shown as the multiple and adjusted R square values). The level of significance for our debt variable and refuse level of our drug variables lowers one level. Also the magnitude of some of our variables levels' coefficients lower, but overall all variables' coefficients' polarity with income remain the same and statistically significant.

```
drug.lm <- lm(income.lastYear ~ gender.Youth+as.factor(daysUse.hardDrugs2011), data=newdata)
summary(drug.lm)
```

```
##
## Call:
## lm(formula = income.lastYear ~ gender.Youth + as.factor(daysUse.hardDrugs2011),
##     data = newdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -38119 -16119  -4119   9902 116405
##
## Coefficients:
##                                       Estimate Std. Error t value
## (Intercept)                            38118.6      488.8  77.985
## gender.YouthWomen                      -8020.7      699.9 -11.460
## as.factor(daysUse.hardDrugs2011)refuse -15593.2     3847.1  -4.053
## as.factor(daysUse.hardDrugs2011)yes      -500.6     2078.5  -0.241
##                                                  Pr(>|t|)
## (Intercept)                           < 0.0000000000000002 ***
## gender.YouthWomen                     < 0.0000000000000002 ***
## as.factor(daysUse.hardDrugs2011)refuse            0.0000512 ***
## as.factor(daysUse.hardDrugs2011)yes                  0.81
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25400 on 5289 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.02666,    Adjusted R-squared:  0.02611
## F-statistic: 48.29 on 3 and 5289 DF,  p-value: < 0.00000000000000022
```

I will drop the drug variable since there's no significant difference between the yes and no level (which is used as the baseline) in the drug variable. This is not a helpful answer as to whether drug use effects income level. These p-values are consistent in previous models and this relatively simple model.

Also the marital status variable in multiple models had only one level, married, that was a significant predictor of income. If all or most levels were significant, this would be a convincing predictor of income. I'll exclude this variable in my final model as well.

```
final.lm <- lm(income.lastYear ~ gender.Youth+ as.factor(race)+debtAge20+household.net.worth.Parent+bir
summary(final.lm)
```

```
##
## Call:
## lm(formula = income.lastYear ~ gender.Youth + as.factor(race) +
##     debtAge20 + household.net.worth.Parent + birthYear.Youth,
##     data = newdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -62197 -14984  -3349   9463 123190
##
## Coefficients:
##                                 Estimate     Std. Error t value
## (Intercept)                3971893.457493  542936.943944   7.316
```

```
## gender.YouthWomen            -7689.807535     767.852353 -10.015
## as.factor(race)Hispanic       3489.037447    1166.271572   2.992
## as.factor(race)mixed          6203.547987    3914.733233   1.585
## as.factor(race)non-black      6070.243996    1011.173476   6.003
## debtAge20                        0.186045       0.046386   4.011
## household.net.worth.Parent       0.037321       0.002855  13.072
## birthYear.Youth              -1989.575444     273.929530  -7.263
##                                          Pr(>|t|)
## (Intercept)                   0.000000000000310 ***
## gender.YouthWomen           < 0.0000000000000002 ***
## as.factor(race)Hispanic                  0.00279 **
## as.factor(race)mixed                     0.11312
## as.factor(race)non-black      0.000000002112206 ***
## debtAge20                     0.000061648838911 ***
## household.net.worth.Parent  < 0.0000000000000002 ***
## birthYear.Youth               0.000000000000455 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23870 on 3886 degrees of freedom
##   (1408 observations deleted due to missingness)
## Multiple R-squared:  0.1065, Adjusted R-squared:  0.1049
## F-statistic:  66.2 on 7 and 3886 DF,  p-value: < 0.00000000000000022
```

Thus I conclude that age, race, and debt are predictors and major influencers of the income gap between men and Women.

Discussion

My confidence is only as good as the data itself. I had to exclude almost half of the original dataset because there are missing values in the income variable. I also wish I knew more about the data collection process so that I could have explored ways to transform the data to preserve other intact variables. Also be be confident that what I have worked with was collected randomly and the other variables I was using were not skewed.

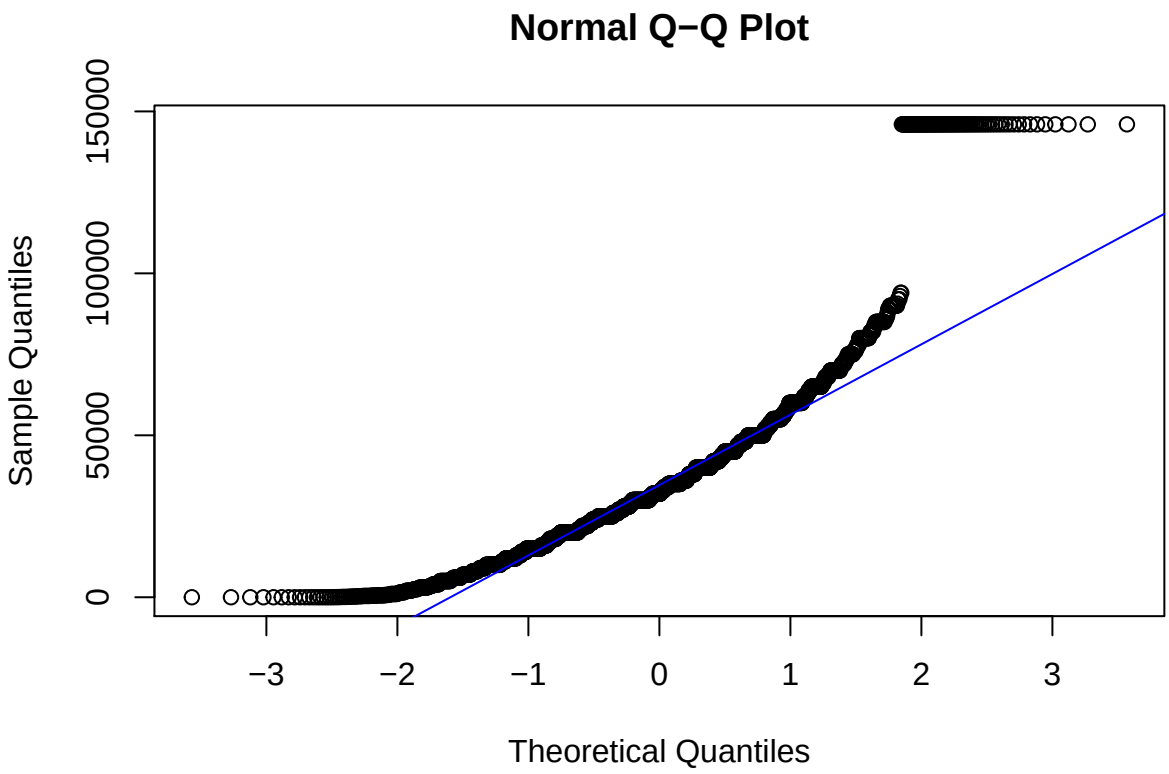
Also, I included the topcoded income values in my models, so its possible that the fit of my model may be highly skewed.

However given the regression models I've built and the exploratory analysis I conducted on the variables, I'm very confident that an income gap exists between Women and Men. I'm less, but still pretty confident that the variables that I've chosen are strong predictors of the income gap between Women and Men. It's very possible I may have not even considered to explore a variable that's a significant predictor of income.

Additional Work Excluded From the Project

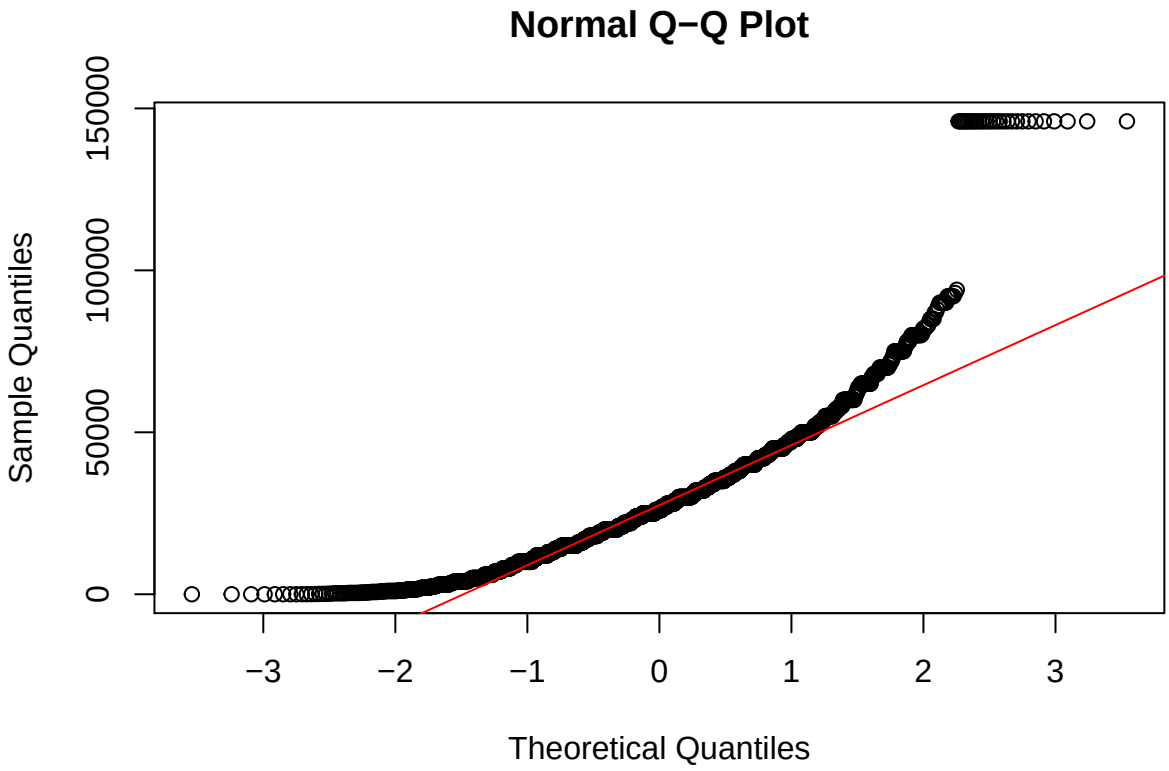
Before thoroughly reading the project prompt, I thought of using hypothesis testing to determine if there's a statistically significant difference between income levels of Men and Women. Before calculating confidence intervals and setting any hypothesis tests, I would need to confirm that the distribution of the incomes across the genders are approximately normally distributed. If they're not, I'd have to use non-parametric test to test if there is a significant income gap between the genders.

```
with(newdata, qqnorm(income.lastYear[gender.Youth=="Men"]))
with(newdata, qqline(income.lastYear[gender.Youth=="Men"], col="blue"))
```



The points that lie off the curving lines of both graphs are the topcoded values, we'll ignore these. However, since most of the line is on the blue line, we can assume that the distribution of Men's income is normally distributed and can be used in a t.test.

```
with(newdata, qqnorm(income.lastYear[gender.Youth=="Women"]))
with(newdata, qqline(income.lastYear[gender.Youth=="Women"], col="red"))
```



Again, the points that lie off the curving lines of both graphs are the topcoded values of the highest incomes amongst Women, we'll ignore these. Since most of the line is on the blue line, we can confidently assume that the distribution of Women's income is normally distributed and can be used in a t.test.