

## Data Simulation Project Write-Up

Adolescent obesity is a growing problem in the United States. Increasingly sedentary lifestyles and easy access to processed foods have been tagged as major contributors to recent rises in obesity among adolescents. However, there has also been a lot research into other contributing factors. These include parental education, race, and other socioeconomic factors such as family income. Gordon-Larsen et al. (2003) took a look at these factors and saw that “family income and parental education had a limited effect on the disparities in overweight prevalence.” However, “family income and parental education had a limited effect on the disparities in overweight prevalence.”

My data simulation aimed to emulate the distributions of sex, race, and income-level outlined in the research article and analyze some of the same relationships as the researchers. This included:

- % Overweight by Race
  - Stratified by sex
- % Overweight by Income Level
  - Stratified by race
- % Overweight by Parental Education
  - Stratified by race

Based off of the following table from the research article the code below was used to create a sample of the appropriate number of rows for each race, income level, and education level.

Characteristic	Whites	African Americans	Hispanics	Asians
Parental education				
<HS	595 (8.96)	363 (17.66)	925 (41.90)	102 (17.29)
HS/GED	2253 (32.52)	787 (35.19)	566 (24.76)	172 (19.60)
Some College	2072 (29.20)	736 (24.51)	412 (17.70)	185 (17.00)
College Grad/Professional	1904 (25.30)	797 (18.65)	272 (11.59)	421 (41.07)
Family Income				
Mean Income	47,825	27,685	30,579	45,461
\$0 to \$20K	1067 (16.39)	1019 (46.39)	813 (38.17)	113 (15.41)
\$20K to \$40K	2213 (30.50)	938 (33.73)	924 (39.51)	352 (40.05)
\$40K to \$60K	2067 (28.94)	477 (11.63)	337 (13.78)	241 (23.72)
\$60K to \$80K	982 (13.79)	219 (5.32)	119 (5.02)	111 (10.24)
\$80K+	806 (10.37)	142 (2.93)	70 (3.52)	103 (10.58)
<HS, no or some high school; HS/GED, high school/GED diploma.				

# Create column of randomly selected race with probabilities based on the percentages in my research article

```
race <- c('white',sample(c('white','black','hispanic','asian'),size = 9999, replace = TRUE, prob =  
c(.5441,.2131,.1726,.0702)))  
data<- data.frame(race)
```

# Create column of parental education with probabilities based on the percentages from my research article

```
parental_education_white <- sample(c('<HS','HS/GED','Some College','College  
Grad/Professional'),size=(length(data[which(data$race=="white"),])),replace = TRUE,prob =  
c(.0896,.3252,.2920,.2530))
```

```
parental_education_black <- sample(c('<HS','HS/GED','Some College','College  
Grad/Professional'),size=(length(data[which(data$race=="black"),])),replace = TRUE,prob =  
c(.1766,.3519,.2451,.1865))
```

```
parental_education_hispanic <- sample(c('<HS','HS/GED','Some College','College  
Grad/Professional'),size=(length(data[which(data$race=="hispanic"),])),replace = TRUE,prob =  
c(.4190,.2476,.1770,.1159))
```

```
parental_education_asian <- sample(c('<HS','HS/GED','Some College','College  
Grad/Professional'),size=(length(data[which(data$race=="asian"),])),replace = TRUE,prob =  
c(.1729,.1960,.17,.4107))
```

# Add column to data frame

```
data$education[which(data$race == 'white')] <- parental_education_white  
data$education[which(data$race == 'black')] <- parental_education_black  
data$education[which(data$race == 'hispanic')] <- parental_education_hispanic  
data$education[which(data$race == 'asian')] <- parental_education_asian  
data$education <- as.factor(data$education)
```

# Create column of incomes for each race with probabilities based on the percentages from my research article

```
income_level_white <- sample(c('$0-$20K','$20K-$40K','$40K-$60K','$60K-  
$80K','$80K+'),size=(nrow(data[which(data$race=="white"),])),replace = TRUE,prob =  
c(.1639,.3050,.2894,.1379,.1037))
```

```
income_level_black <- sample(c('$0-$20K','$20K-$40K','$40K-$60K','$60K-  
$80K','$80K+'),size=(nrow(data[which(data$race=="black"),])),replace = TRUE,prob =  
c(.4639,.3373,.1163,.0532,.0293))
```

```
income_level_hispanic <- sample(c('$0-$20K','$20K-$40K','$40K-$60K','$60K-  
$80K','$80K+'),size=(nrow(data[which(data$race=="hispanic"),])),replace = TRUE,prob =  
c(.3817,.3951,.1378,.0502,.0352))
```

```
income_level_asian <- sample(c('$0-$20K','$20K-$40K','$40K-$60K','$60K-  
$80K','$80K+'),size=(nrow(data[which(data$race=="asian"),])),replace = TRUE,prob =  
c(.1541,.4005,.2372,.1024,.1058))
```

# Add column to data frame

```
data$income.level[which(data$race == 'white')] <- income_level_white  
data$income.level[which(data$race == 'black')] <- income_level_black  
data$income.level[which(data$race == 'hispanic')] <- income_level_hispanic  
data$income.level[which(data$race == 'asian')] <- income_level_asian  
data$income.level <- as.factor(data$income.level)
```

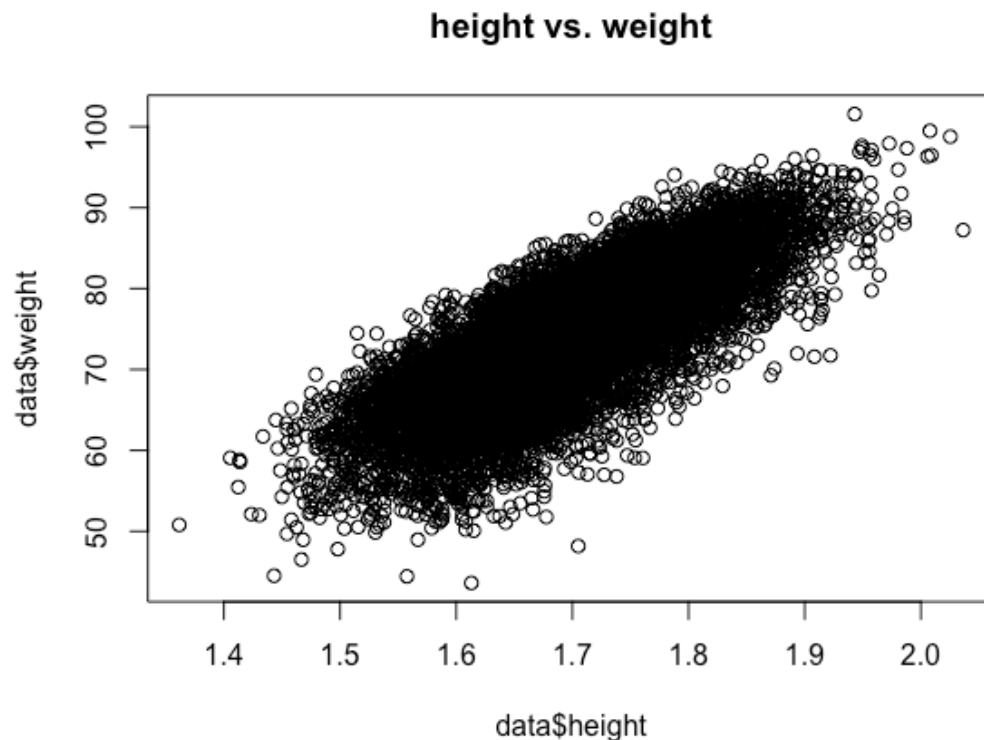
```
# Create column for sex with probabilities based on the percentages from my research article
data$sex <- c('male',sample(c('male','female'),size=(nrow(data)-1),replace = TRUE,prob = c(.535,.465)))
data$sex <- as.factor(data$sex)
```

The next step in replicating the data was creating height and weight columns. I retrieved the mean and standard deviation for height and weight for men and women at 16 years old (as this was the mean age of the population in the article) from CDC census data.<sup>1</sup> Using this data I calculated the mean BMI for the two groups (mean weight/mean height<sup>2</sup>). These two columns needed to be correlated so I used the following code to create a vector of heights and then a correlated vector of weights for each sex.

```
male_heights <- rnorm(male_n, mean=male_height_mean, sd=male_height_sd)
male_weights <- .6 * mean_male_bmi * (male_heights^2) + .4* (male_weight_mean + rnorm(male_n,
sd=male_weight_sd,xi=-1.5))
```

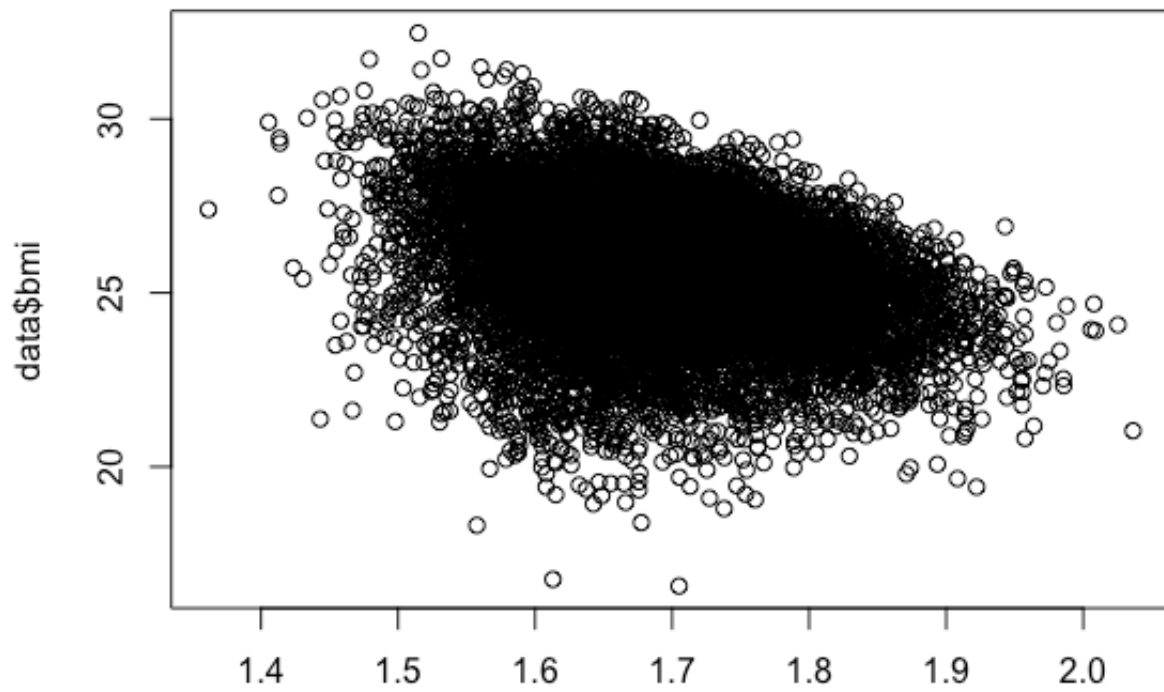
BMI was then calculated using the formula described above and added as a column to the data frame. Based on that column an “obese” column was created with 0 = not obese (BMI < 25) and 1 = obese (BMI >= 25). To ensure that the data was properly created the following plots were created to see the relationships between height and weight, height and BMI, and weight and BMI.

### **Height, Weight, BMI Scatter Plots**

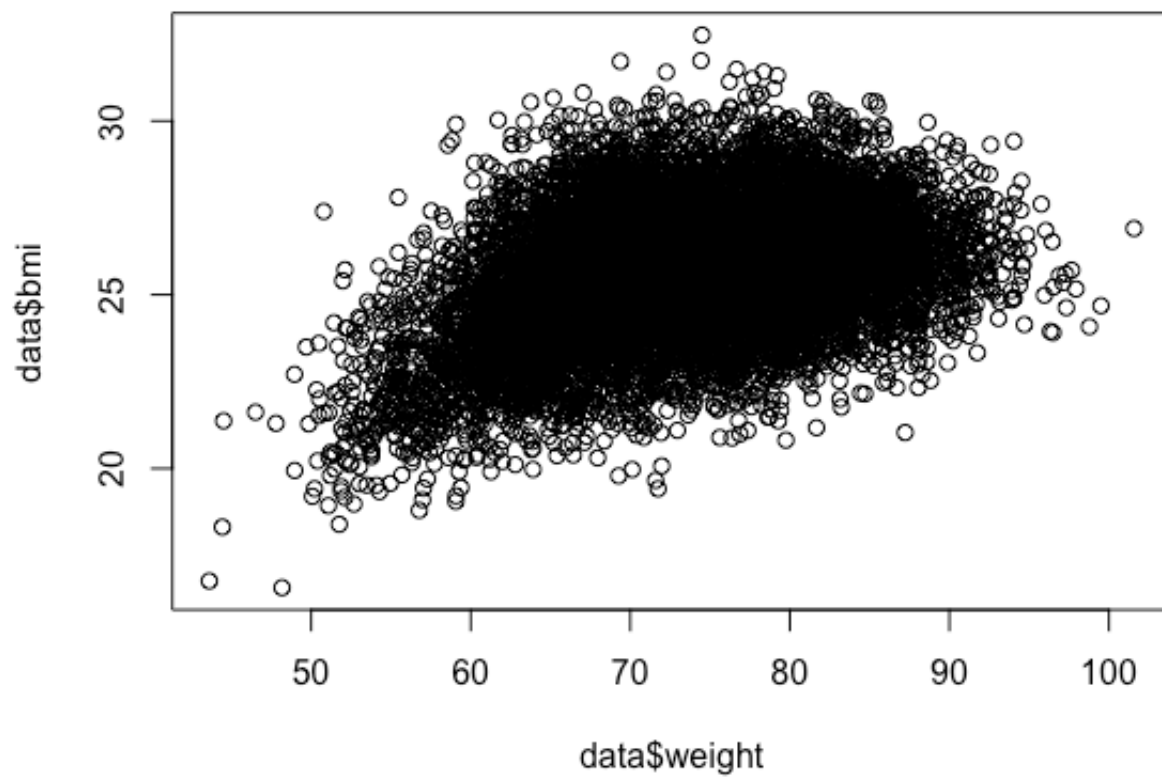


<sup>1</sup> <https://www.cdc.gov/growthcharts/data/set1clinical/cj411021.pdf> and <https://www.cdc.gov/growthcharts/data/set1clinical/cj411022.pdf>

**height vs. bmi**



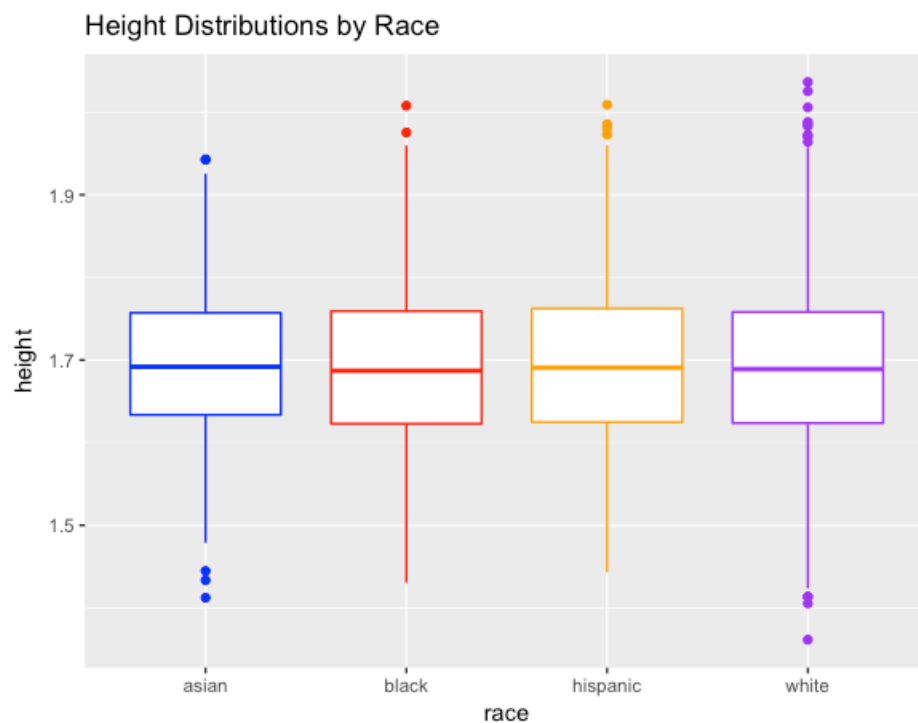
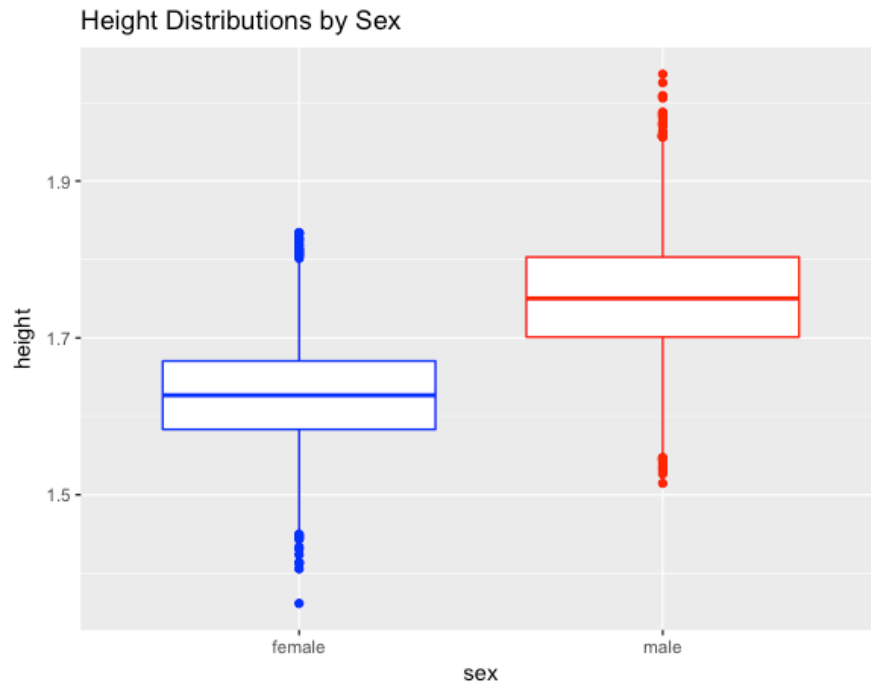
**weight vs. bmi**



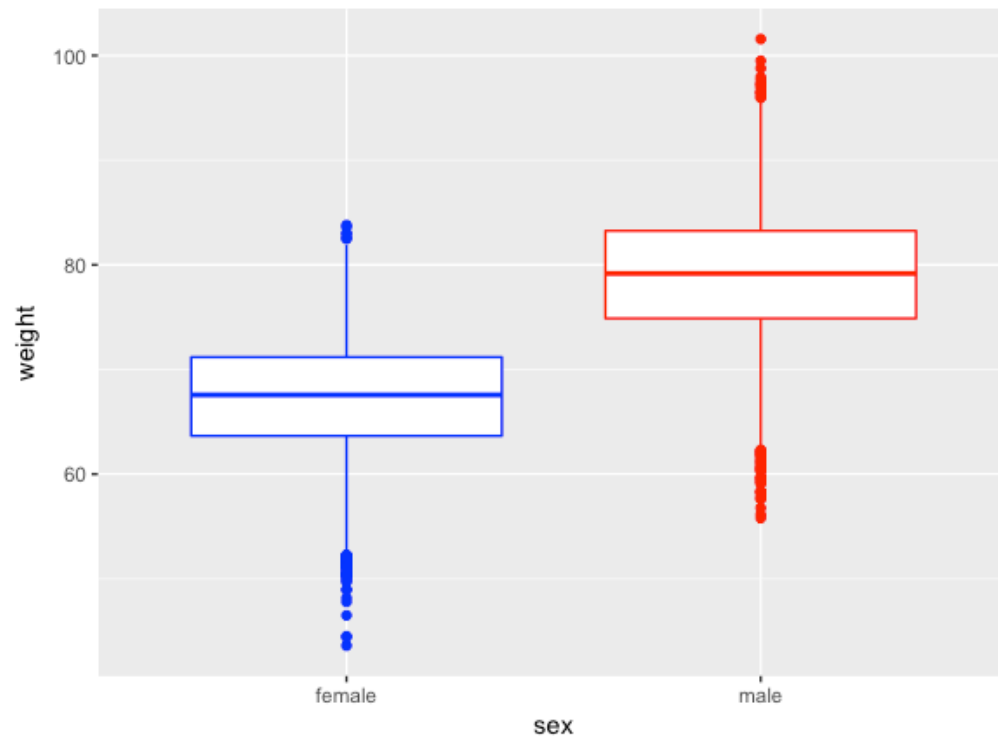
Based on the plots it was clear to see that the correlated variables had been created successfully. Higher heights correlated with higher weights and higher weights somewhat correlated with higher BMIs.

Next I wanted to see what the distributions for height, weight and BMI looked like across sex, race, and income level. The following plots display this information:

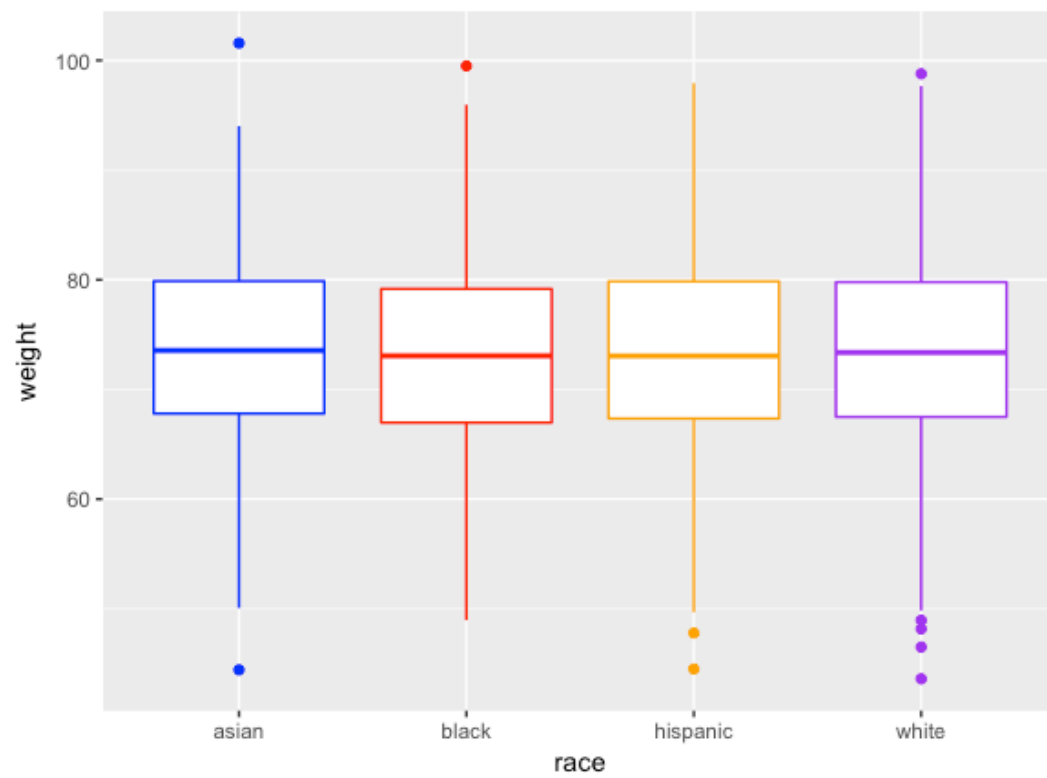
### **Population Characteristics Distributions**



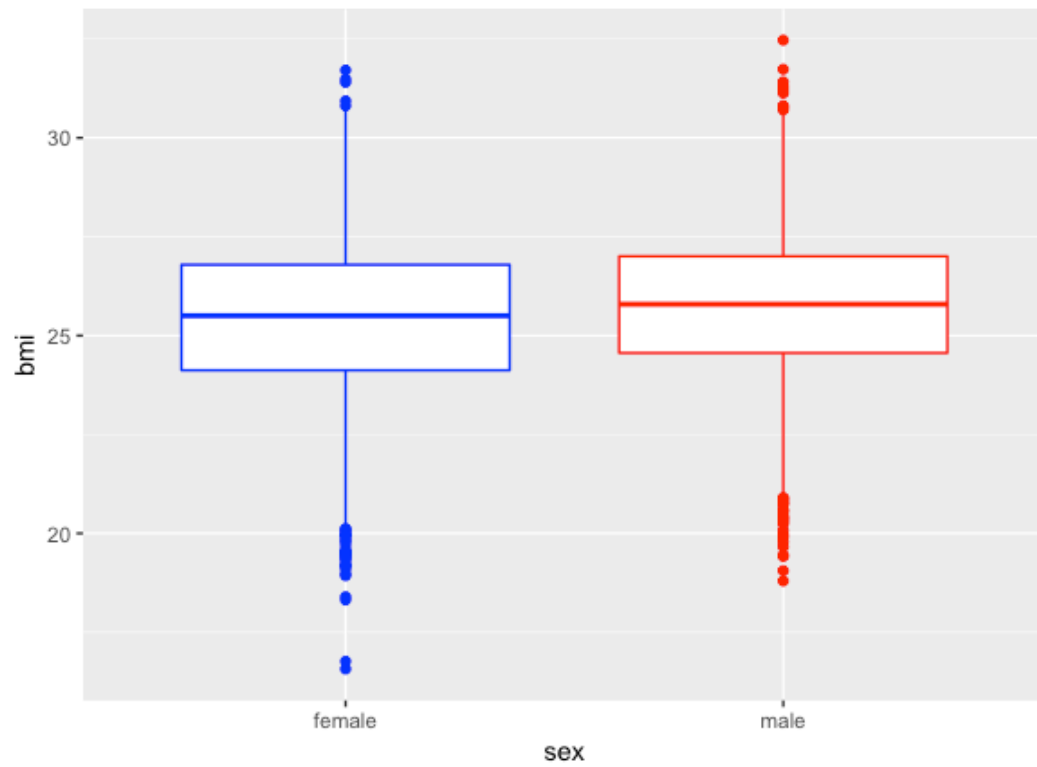
Weight Distributions by Sex



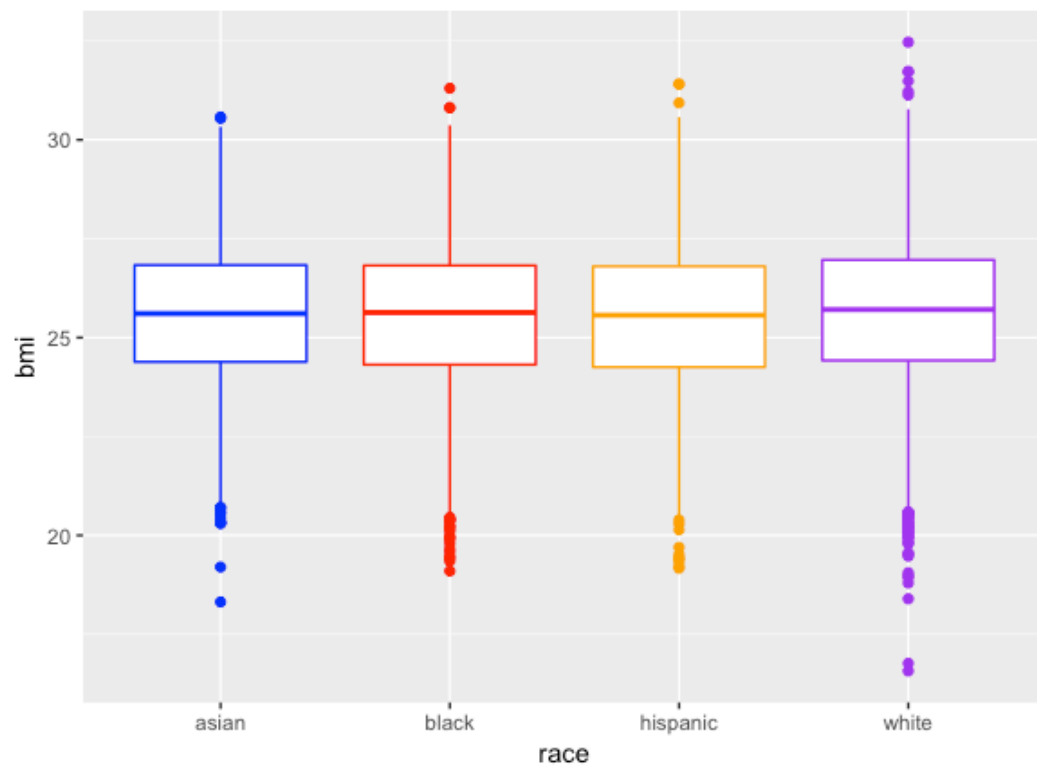
Weight Distributions by Race



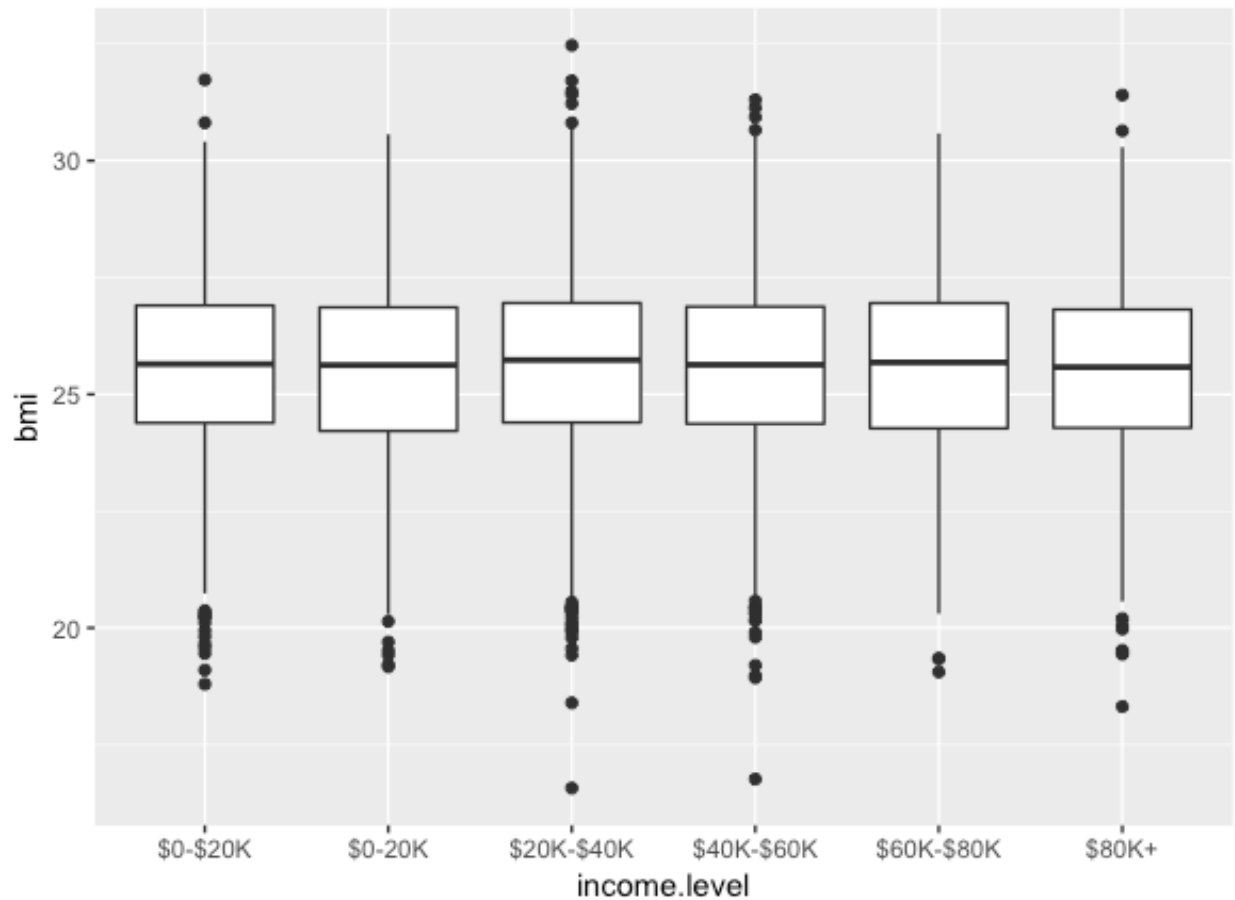
BMI Distributions by Sex



BMI Distributions by Race



BMI Distributions by Income Level



A t-test was then performed to see if there were statistically significant differences in mean BMI across sexes.

```
> t.test(data_male$bmi,data_female$bmi)

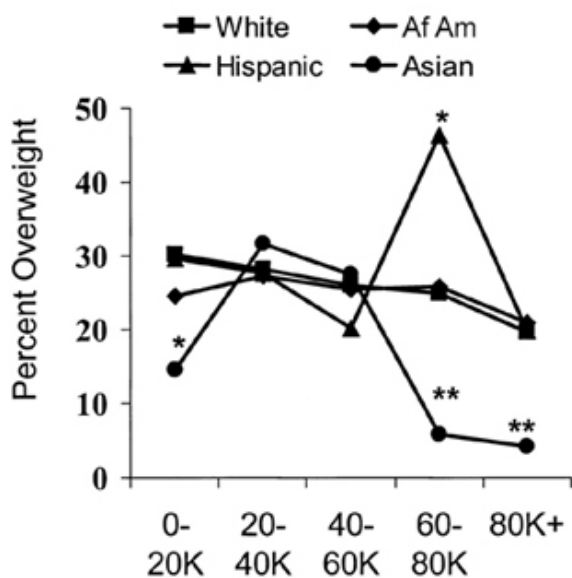
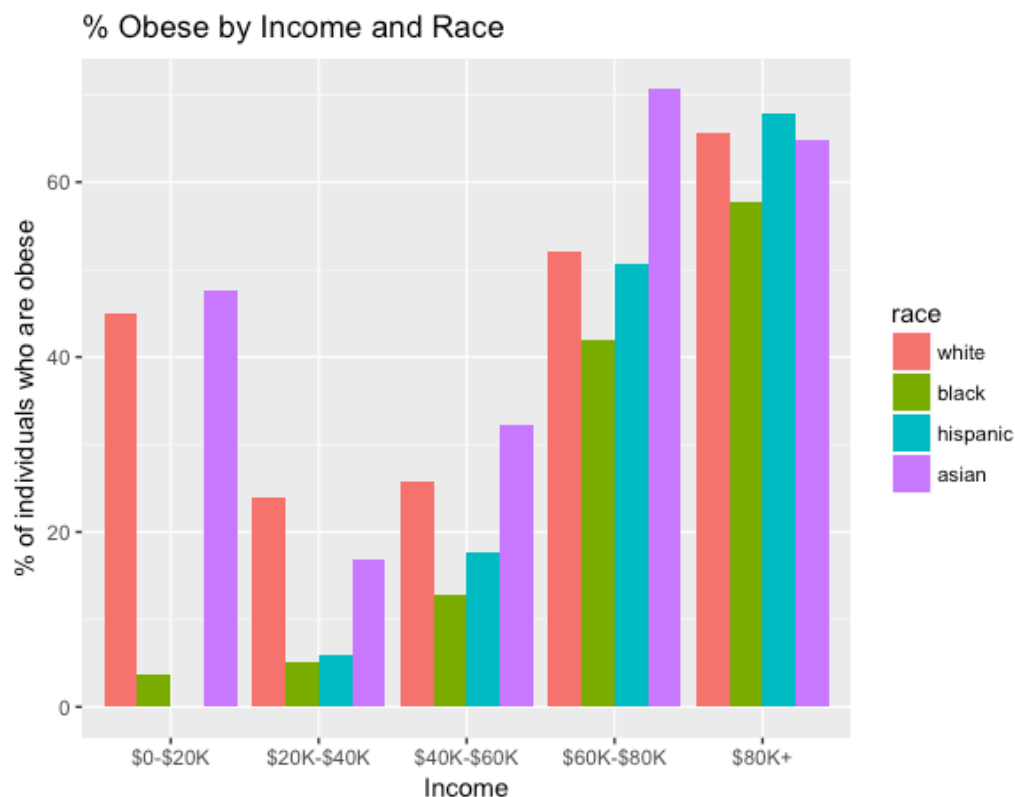
Welch Two Sample t-test

data: data_male$bmi and data_female$bmi
t = 9.1516, df = 9574, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2750047 0.4249254
sample estimates:
mean of x mean of y
25.75852 25.40855
```

The results of this t-test show that there is a statistically significant difference in the mean BMIs of each sex ( $p = 2.2e-16$ ,  $CI = (0.275, 0.424)$ ).



As there are not equivalent numbers of rows for each race the percent of each that were obese needed to be calculated in order to perform chi-squared tests and to plot that data in a manner that would reflect the true information. After creating data frames for each race, income level, and education level. The first variable I looked at was the % of people overweight by income level and race. This was also done in the article.



There were clearly some differences in the articles results and mine. The original research showed that there are some spikes in obesity rates. For example, Hispanics had a large spike at the 60-80K income level.. However, my data shows that there were much higher levels of obesity at higher income levels and a steady increase as income increases. Also, Asians and Whites had very high percents at low incomes, which differs greatly from the articles data. I performed a chi-squared test of the percents differed over income levels.

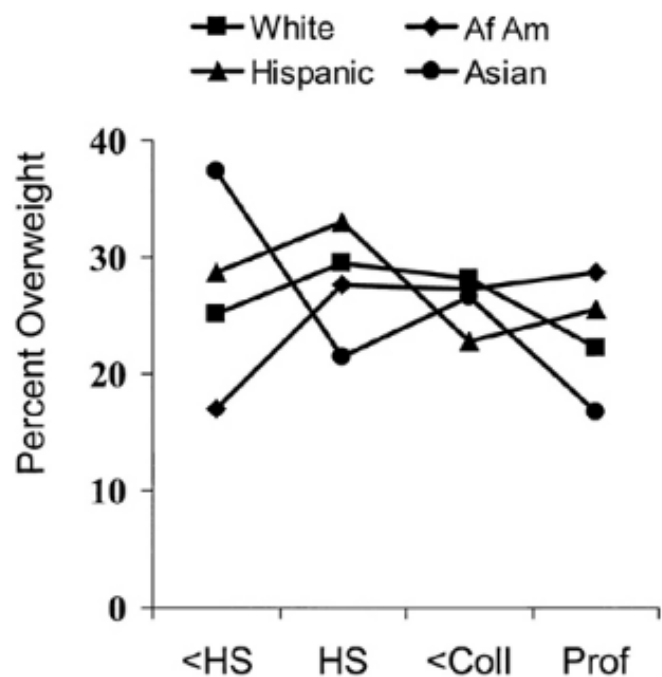
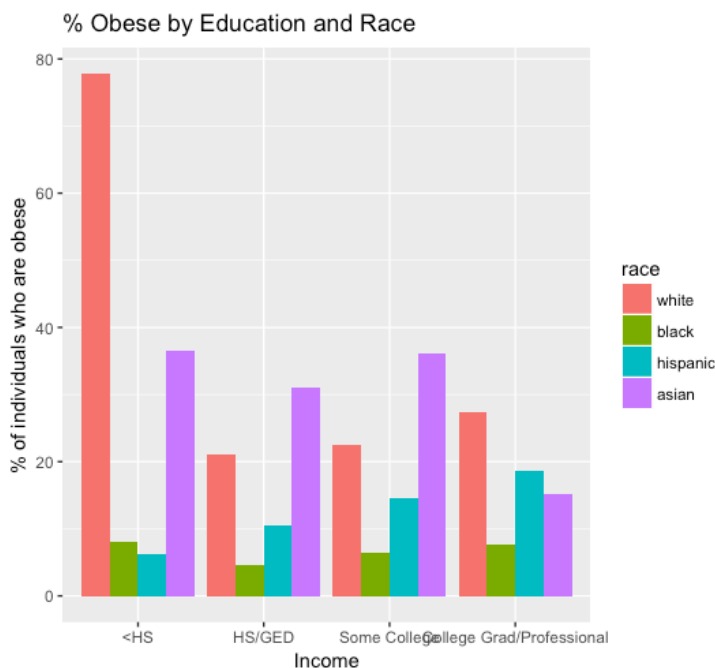
```
> chisq.test((perc_ow_race_income))
```

Pearson's Chi-squared test

```
data: (perc_ow_race_income)
X-squared = 47.848, df = 12, p-value = 3.321e-06
```

These results show that there is a statistically significant difference the percents of obesity across all of the income levels and races If more time was available it would be interesting to investigate the specific pairwise differences across races at each income level using.

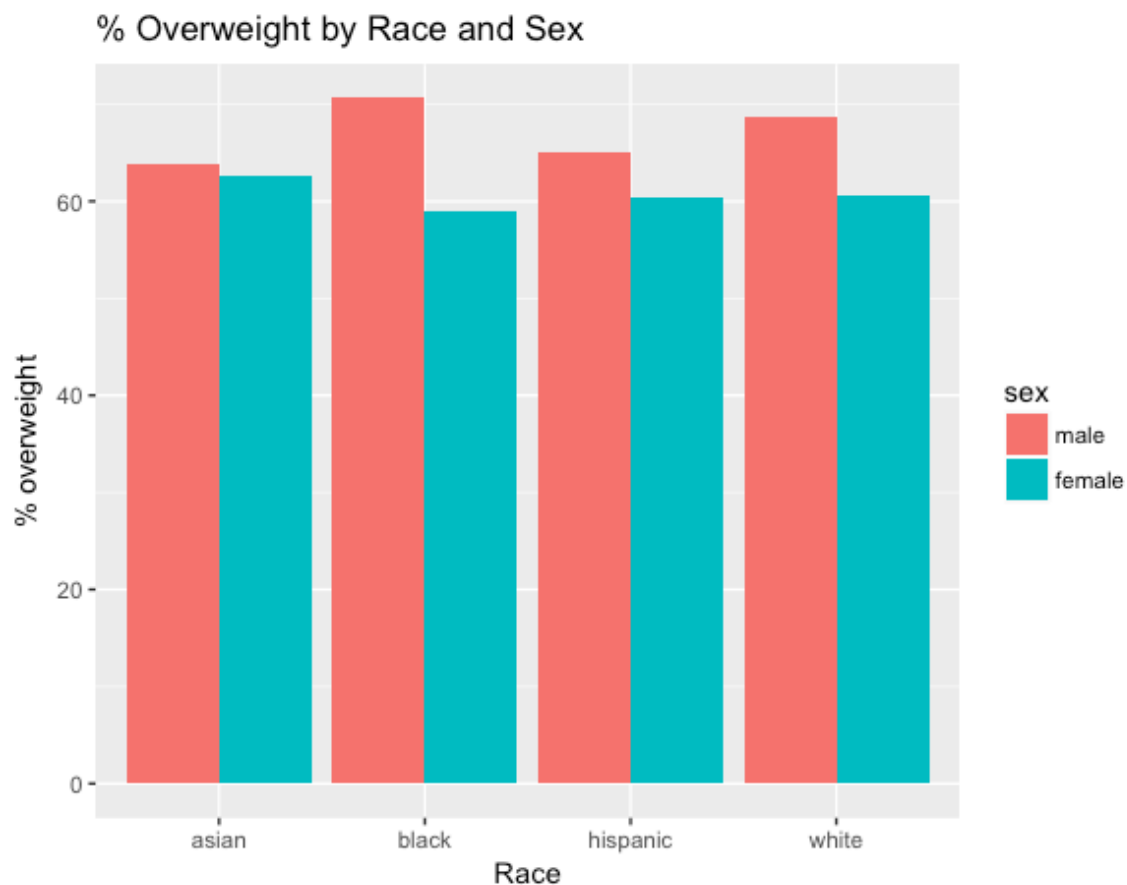
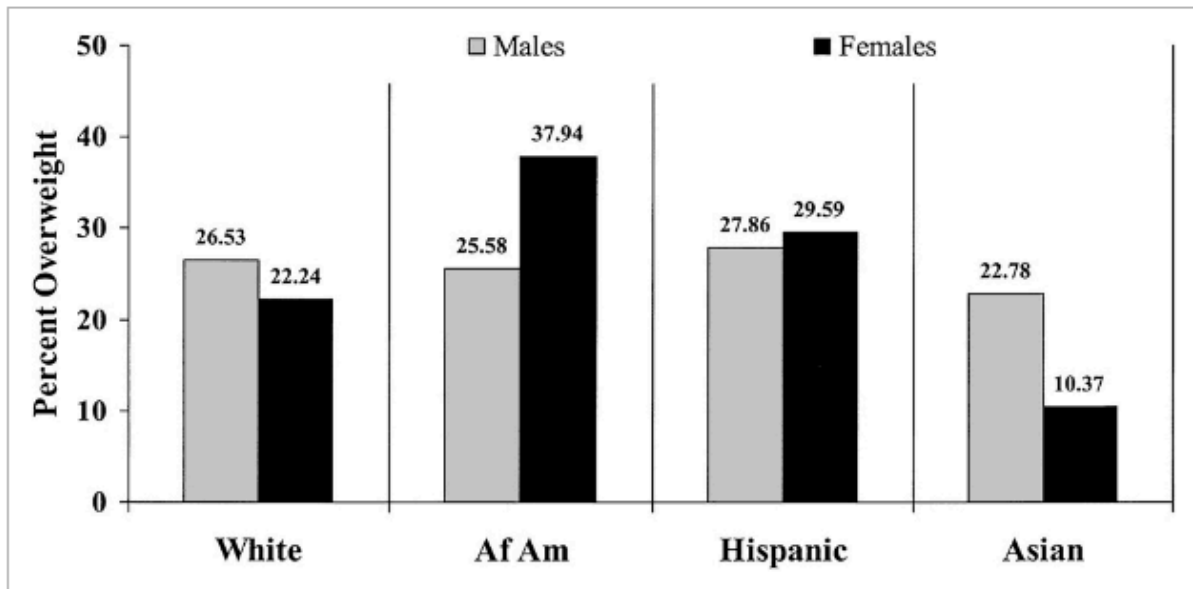
Parental education was the other categorical variable that the researchers looked at to see if there was an effect on adolescent obesity. Below are the graphs



The main differences seen between the two plots is the percentage of whites whose parents had <HS education is much higher than the original data. There also seems to be a

higher percentage of Asians who are obese at higher levels of parental education than in the article.

The original research also investigated the difference in percent of those overweight in each race by sex. The plots for my data and the article's data are shown below:



The graphs for Whites, and Hispanics for my data and the articles are all somewhat similar. However, it is clear to see that in the original research black females had a much higher percent of overweight adolescents than black males. The original research also showed that Asian males had a much greater percent of overweight adolescents than Asian females.

I performed a chi-squared test to see if % of people overweight differed significantly across races. The results showed that the groups were not statistically different

```
> chisq.test(select(overweight_perc, -sex))
```

Pearson's Chi-squared test

```
data: select(overweight_perc, -sex)
X-squared = 0.4725, df = 3, p-value = 0.9249
```

One thing that the researchers did not look at (or at least did not report on) was the differences in percent of people that were obese across sexes. The following performed a t-test for the mean percent of people that were overweight for all races

### **t-test for % of People Overweight Across Sexes**

```
> t.test(ow_transpose$male, ow_transpose$female)
```

Welch Two Sample t-test

```
data: ow_transpose$male and ow_transpose$female
t = 3.6054, df = 4.3906, p-value = 0.01932
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.625225 11.061735
sample estimates:
mean of x mean of y
67.02502 60.68154
```

These results show that there is a statistically different mean percentage of people that are overweight when compared across sex ( $p = 0.01932$ ,  $CI = (1.625, 11.062)$ ).

Since obesity can be classified as a logical variable I decided to try some logistic regression models to see if there was a strong enough correlation to perform an SVM or Clustering machine-learning model. The following logistic regression models (along with a linear model to predict BMI) were created.

## Logistic and Linear Models

```
Call:
glm(formula = obese ~ . - height - weight, family = binomial,
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.01296   0.00000   0.00000   0.00000   0.01425

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -84551.950  242180.968  -0.349   0.727
raceblack        -9.893    270.332  -0.037   0.971
racehispanic     7.731    132.084   0.059   0.953
racewhite        2.959    247.365   0.012   0.990
educationCollege Grad/Professional -6.918    267.116  -0.026   0.979
educationHS/GED  -5.828    155.595  -0.037   0.970
educationSome College -0.428    724.885  -0.001   1.000
income.level$20K-$40K  3.406    133.638   0.025   0.980
income.level$40K-$60K -1.540    147.153  -0.010   0.992
income.level$60K-$80K 14.548    239.769   0.061   0.952
income.level$80K+   -2.137    1334.944  -0.002   0.999
sexmale         -5.231    134.008  -0.039   0.969
bmi             3382.232   9688.189   0.349   0.727

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1.3125e+04 on 9999 degrees of freedom
Residual deviance: 9.2612e-04 on 9987 degrees of freedom
AIC: 26.001

Number of Fisher Scoring iterations: 25
```

```
> glm_race <- glm(obese~race,data = data,family = binomial)
> summary(glm_race)
```

```
Call:
glm(formula = obese ~ race, family = binomial, data = data)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4433  -1.3906   0.9330   0.9499   0.9858

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.56185    0.07832   7.174 7.28e-13 ***
raceblack     -0.07332    0.09043  -0.811   0.418
racehispanic  -0.09293    0.09219  -1.008   0.313
racewhite      0.04448    0.08330   0.534   0.593
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 13125 on 9999 degrees of freedom
Residual deviance: 13116 on 9996 degrees of freedom
AIC: 13124
```

Number of Fisher Scoring iterations: 4

. |

```
glm(formula = obese ~ income.level, family = binomial, data = data)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.4440	-1.4011	0.9324	0.9597	0.9738

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.53635	0.04573	11.730	<2e-16 ***
income.level\$0-20K	-0.03180	0.09141	-0.348	0.728
income.level\$20K-\$40K	0.07146	0.05873	1.217	0.224
income.level\$40K-\$60K	-0.02433	0.06328	-0.385	0.701
income.level\$60K-\$80K	0.05238	0.08060	0.650	0.516
income.level\$80K+	-0.03656	0.08529	-0.429	0.668

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 13125 on 9999 degrees of freedom  
Residual deviance: 13120 on 9994 degrees of freedom  
AIC: 13132

Number of Fisher Scoring iterations: 4

Call:

```
glm(formula = obese ~ education, family = binomial, data = data)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.4315	-1.4114	0.9431	0.9602	0.9656

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.52097	0.04888	10.657	<2e-16 ***
educationCollege Grad/Professional	0.05890	0.06485	0.908	0.364
educationHS/GED	0.04763	0.06135	0.776	0.438
educationSome College	0.01395	0.06326	0.221	0.825

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 13125 on 9999 degrees of freedom  
Residual deviance: 13124 on 9996 degrees of freedom  
AIC: 13132

Number of Fisher Scoring iterations: 4

```

> lm_all <- lm(bmi~.-obese-height-weight,data = data)
> summary(lm_all)

Call:
lm(formula = bmi ~ . - obese - height - weight, data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.9691 -1.2360  0.0713  1.3087  6.6106

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    25.40676    0.09633  263.737  <2e-16 ***
raceblack      -0.03738    0.08496   -0.440    0.660
racehispanic   -0.08241    0.09355   -0.881    0.378
racewhite       0.07495    0.07683    0.976    0.329
educationCollege Grad/Professional -0.01101    0.06201   -0.178    0.859
educationHS/GED -0.02061    0.05839   -0.353    0.724
educationSome College -0.02744    0.06068   -0.452    0.651
income.level$0-20K    0.04061    0.10576    0.384    0.701
income.level$20K-$40K  0.05812    0.05643    1.030    0.303
income.level$40K-$60K -0.03279    0.06108   -0.537    0.591
income.level$60K-$80K -0.02560    0.07604   -0.337    0.736
income.level$80K+    -0.12331    0.08115   -1.520    0.129
sexmale         0.34701    0.03805    9.119  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.898 on 9987 degrees of freedom
Multiple R-squared:  0.009798, Adjusted R-squared:  0.008608
F-statistic: 8.235 on 12 and 9987 DF, p-value: 1.152e-15

```

The model using all of the variables as well as models using each of the possible predictor variables (excluding height and weight as these are directly used to calculate BMI) did not produce any significant predictive p-values. I expected this from my data, as there was not much variation when looking at the distributions of BMIs across each of the predictor variables. Therefore I decided to forgo trying to fit a SVM or clustering machine learning model to the data. The linear model's extremely small R-Squared value (0.008608), the residual vs. fitted plot, and Q-Q plot showed that a linear model was not a good fit in trying to predict BMI of an individual.

Overall I consider my simulation a success. As I was able to recreate the proper distributions of sex, race, income level, and parental education. However, since the original data set was created from a random sample of the entire adolescent population of the United States - and then I tried to recreate it using what sample parameters were available - it should be expected that there would be a good amount of variation in the original and my simulated version. If more variables had been collected during the original research such as the distribution of ages, parent ages, and activity levels it may have been possible to glean more predictive value from the analysis of the data.