Income Variability Across Personal Factors

Armando Marquez am8245

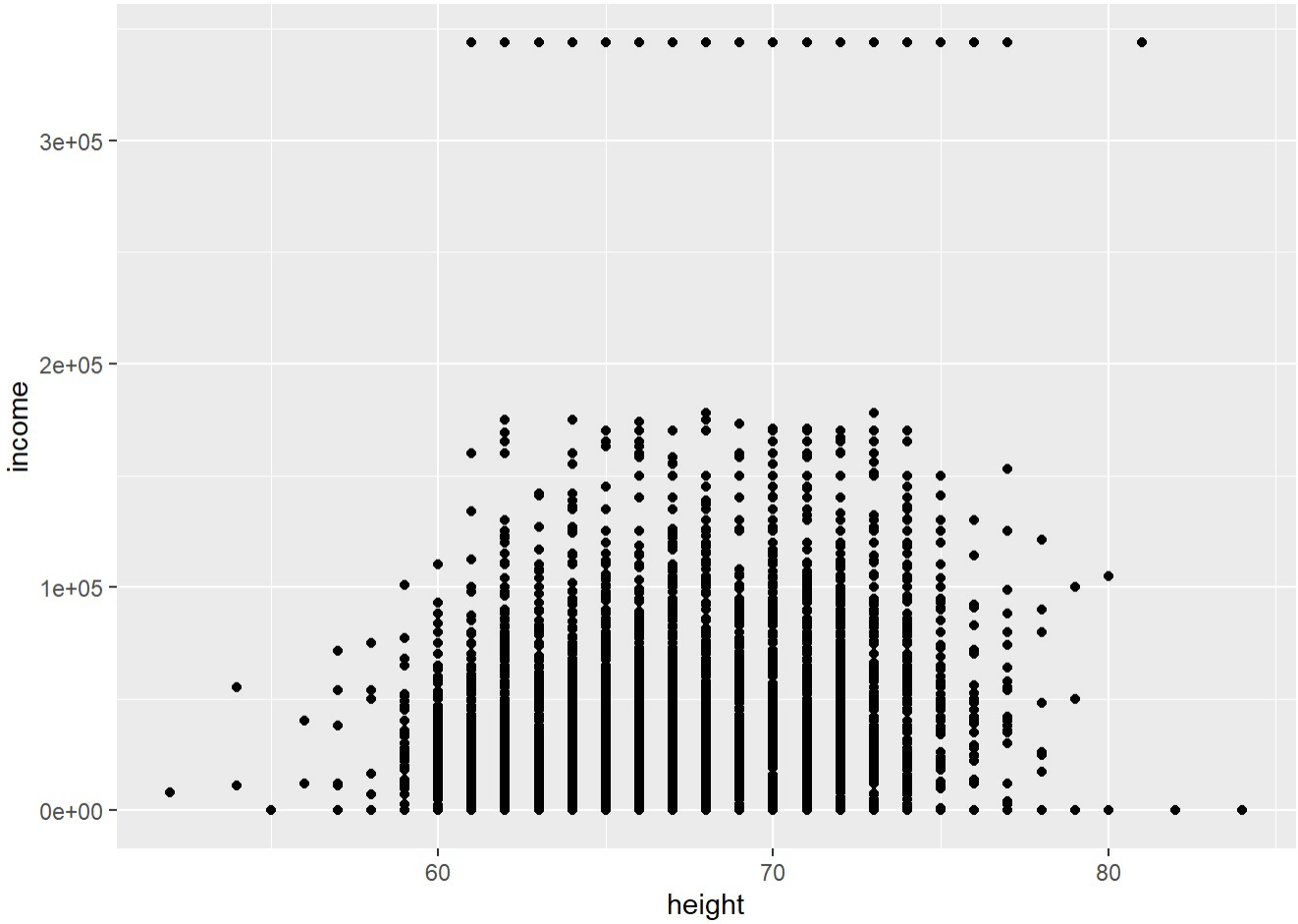
Introduction

Income variability is affected by numerous factors during periods of a persons life. The dataset found in R will be used to understand if income can be related to a persons height, weight, age, marital status, sex, level of education, or afqt test score. The data was already tidy so no steps were taken to further tidy the dataset. Marital status and sex were transformed into numeric variables for further analyzation. N/A answers were also omitted from the datset in order to use complete data for each respondant. I expect to find that level of education, sex, afqt test, and age are all indicators of a persons level of income. I am unsure, but will find out through this analysis if height, weight, marital status, or afqt test score are indicators of a persons level of income. I am most curious to find out if income level is related to height because I have never thought of the two in correlation with each other.

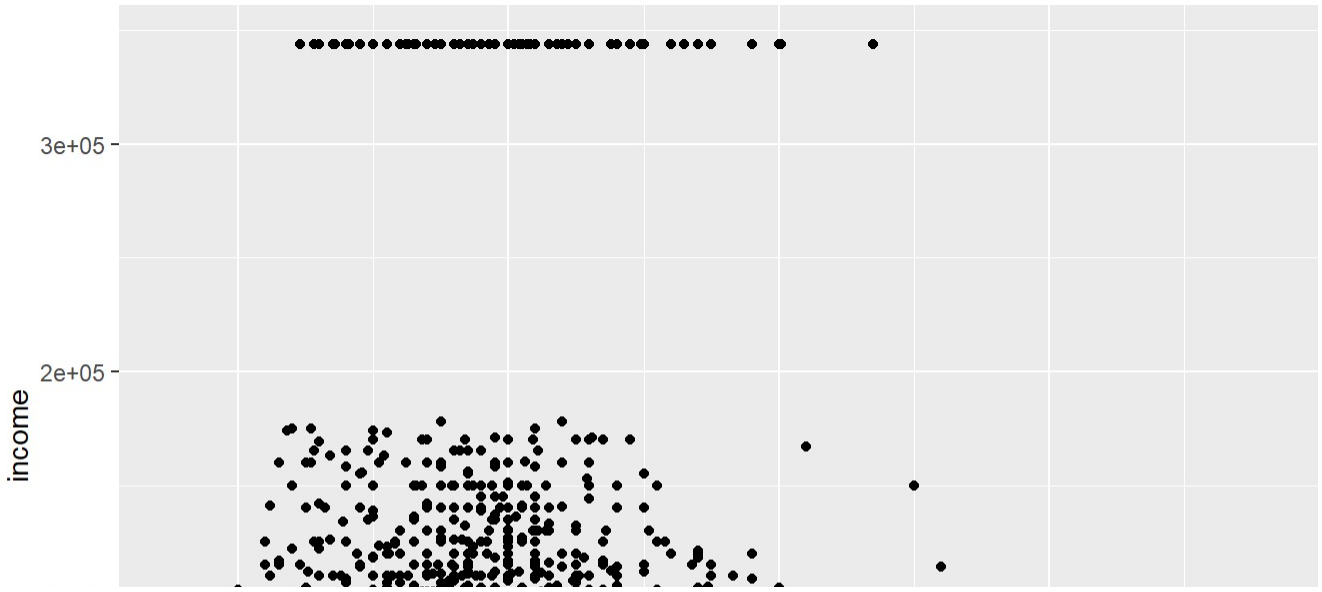
```
#install packages
#install.packages("sandwich")
#install.packages("lmtest")
#Setting libraries
library(modelr)
library(tidyverse)
library(psych)
library(sandwich)
library(lmtest)


#Importing data into a dataset
height <- heights

#Remove any N/As from the dataset
height <- na.omit(height)

# Create a binary variable coded as 0 and 1 for sex
height <- height %>%
  mutate(sex_num = ifelse(sex == "female", 1, 0))

# Create a binary variable coded as 1 through 4 for marital
height <- height %>%
  mutate(marital_num = as.numeric(marital))
```

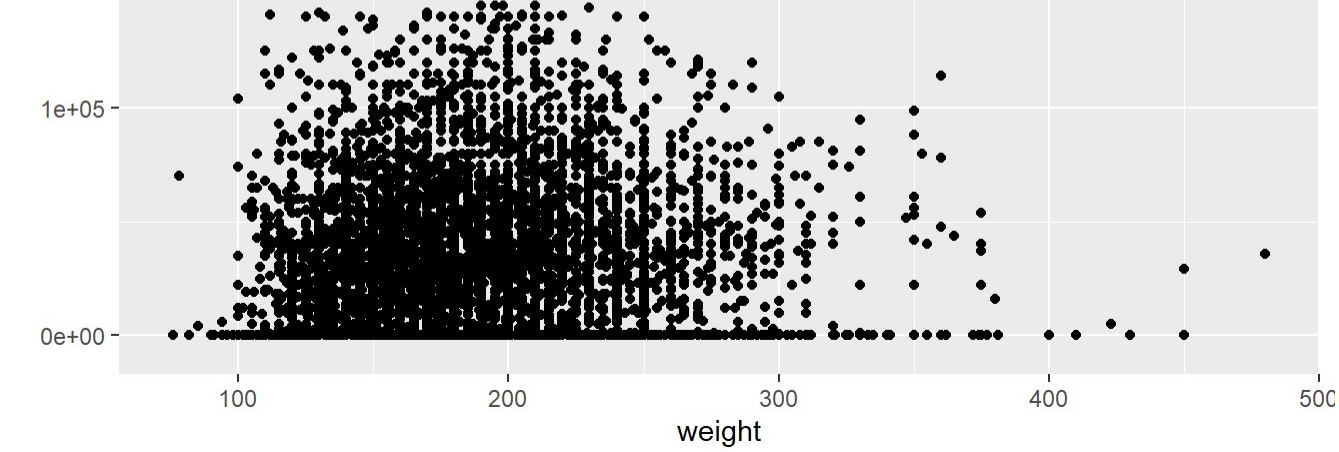
EDA

```
# Visualize the relationship between income and height
ggplot(height, aes(x = height, y = income)) +
  geom_point()
```

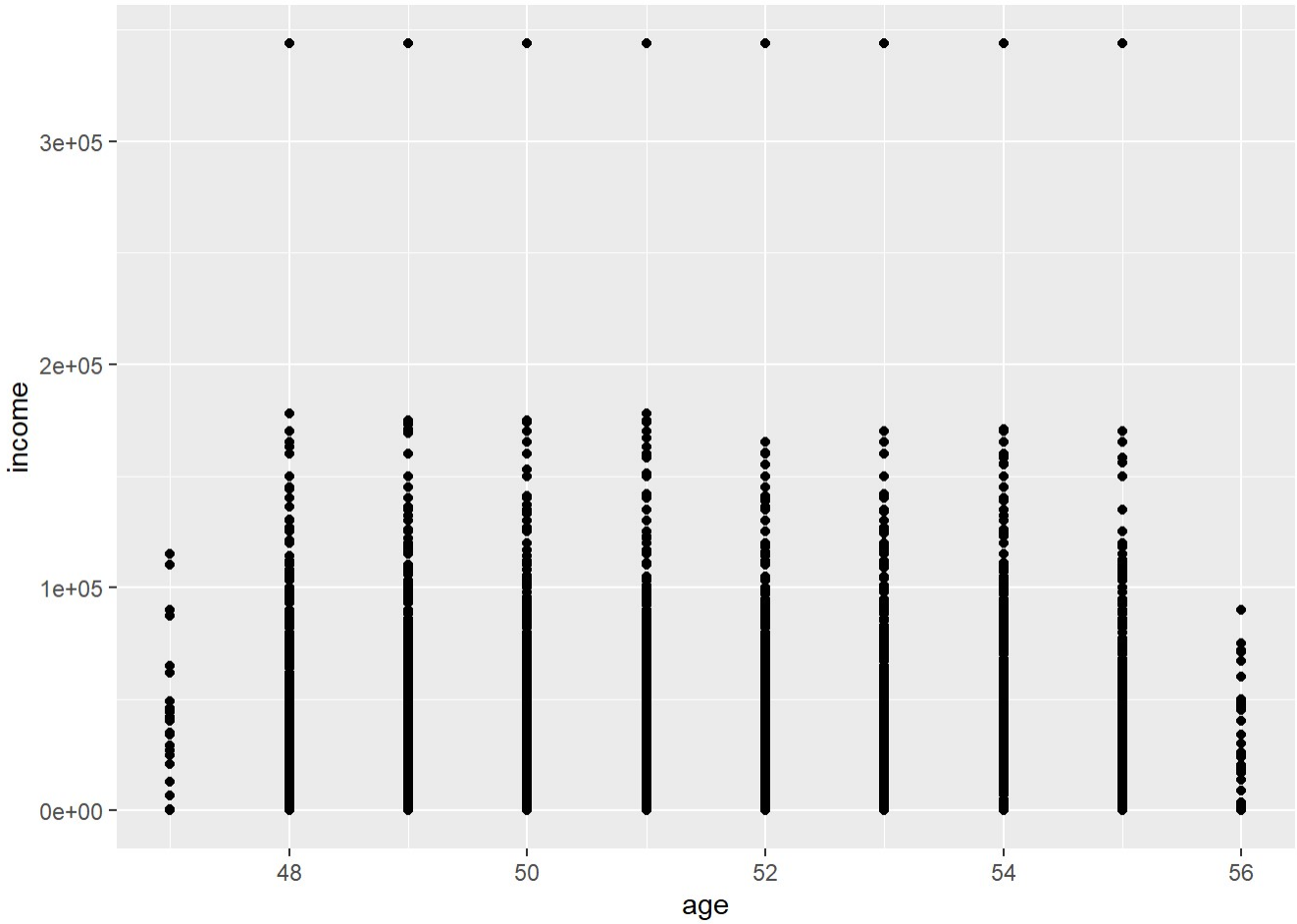



```
# Visualize the relationship between income and weight
ggplot(height, aes(x = weight, y = income)) +
  geom_point()
```

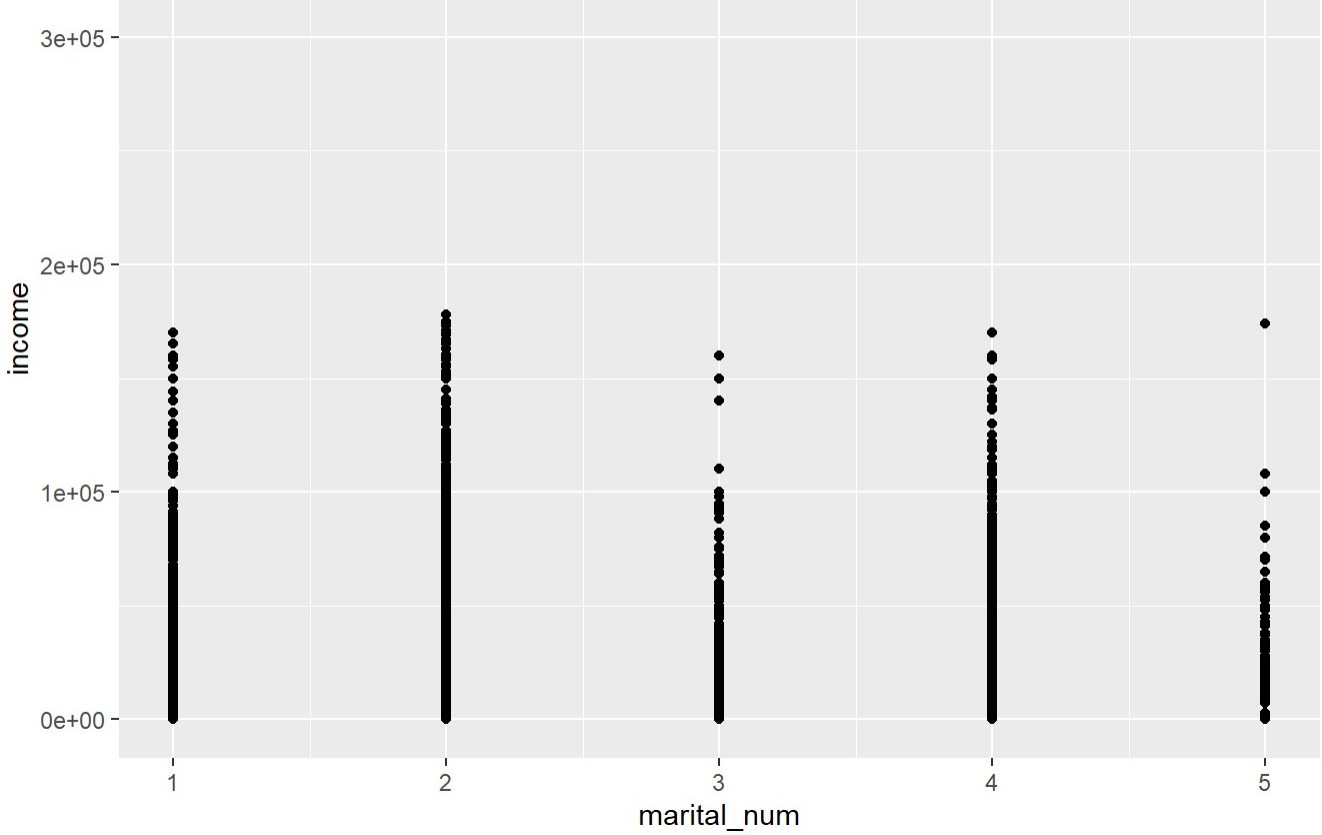




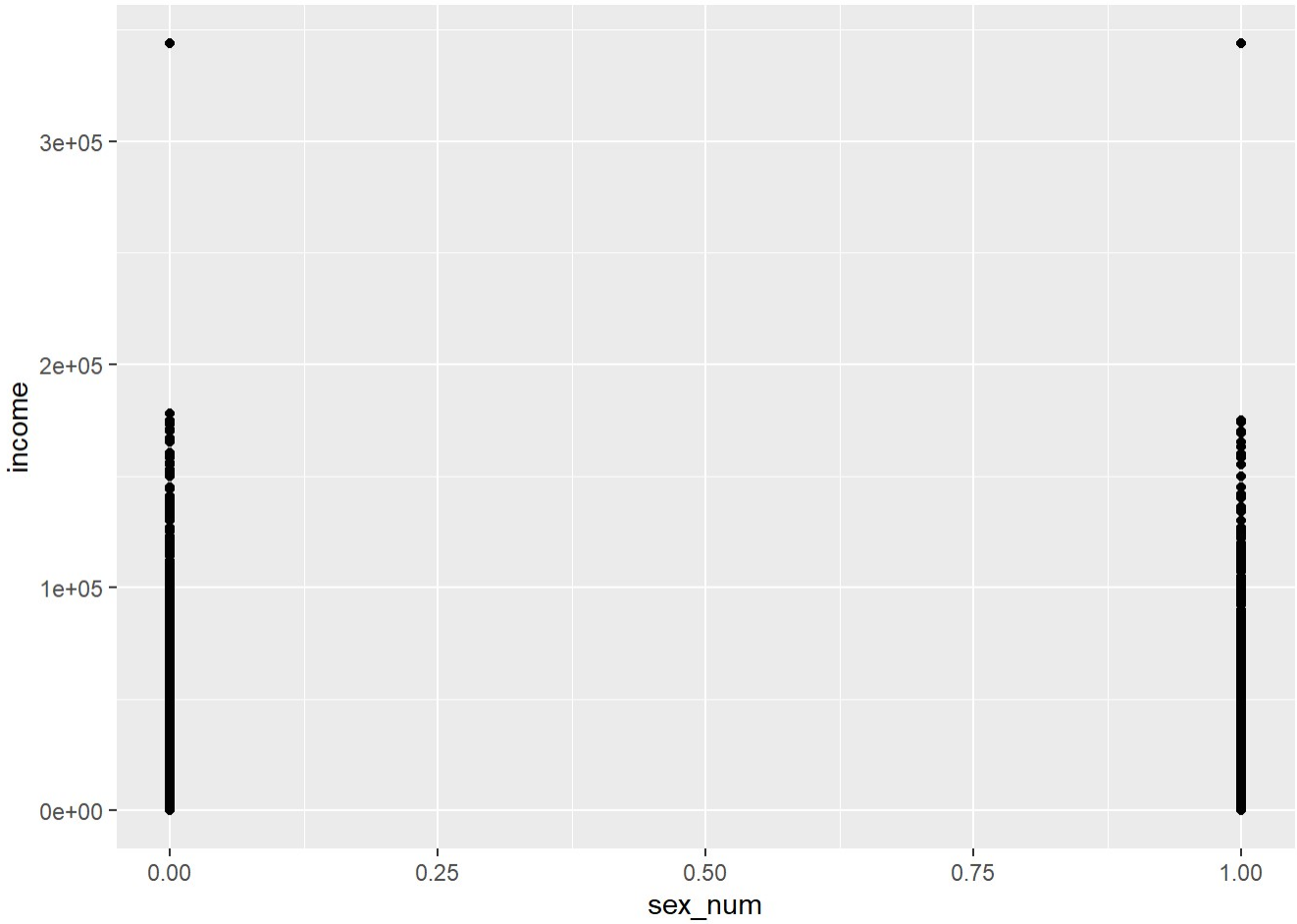

```
# Visualize the relationship between income and age
ggplot(height, aes(x = age, y = income)) +
  geom_point()
```



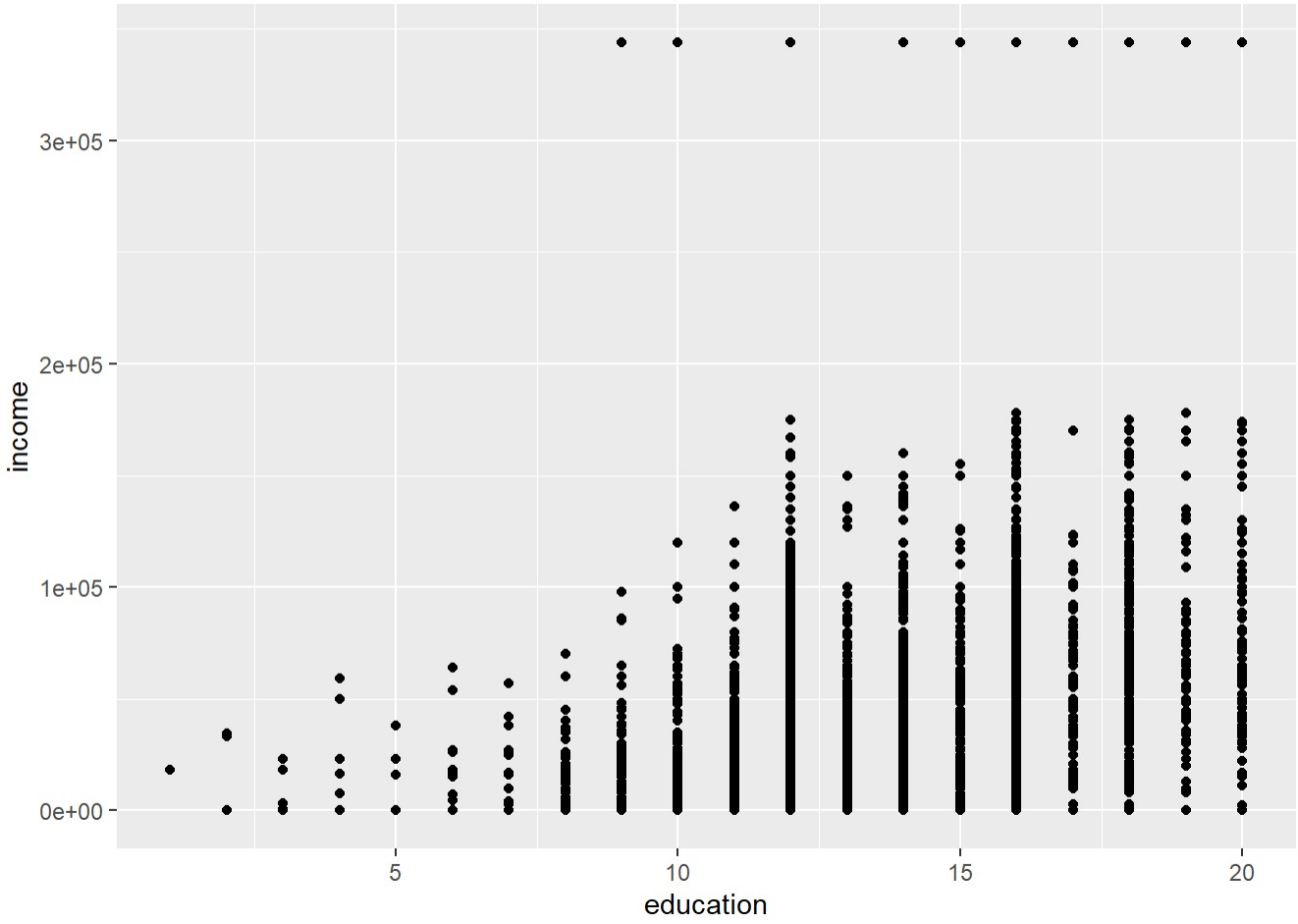

```
# Visualize the relationship between income and marital status
ggplot(height, aes(x = marital_num, y = income)) +
  geom_point()
```



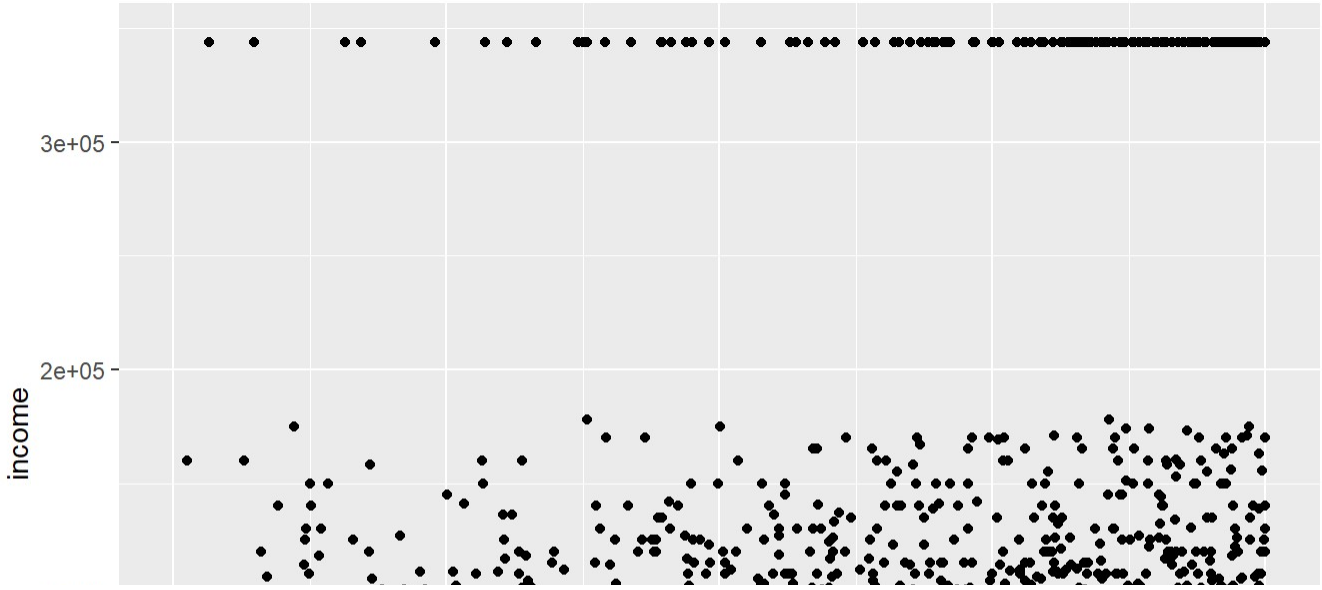

```
# Visualize the relationship between income and sex
ggplot(height, aes(x = sex_num, y = income)) +
  geom_point()
```

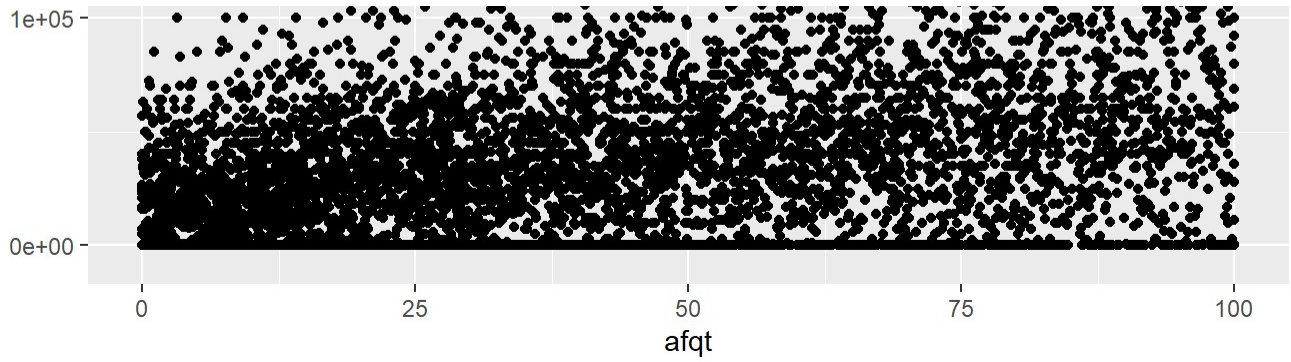


```
# Visualize the relationship between income and education
ggplot(height, aes(x = education, y = income)) +
  geom_point()
```




```
# Visualize the relationship between income and afqt
ggplot(height, aes(x = afqt, y = income)) +
  geom_point()
```



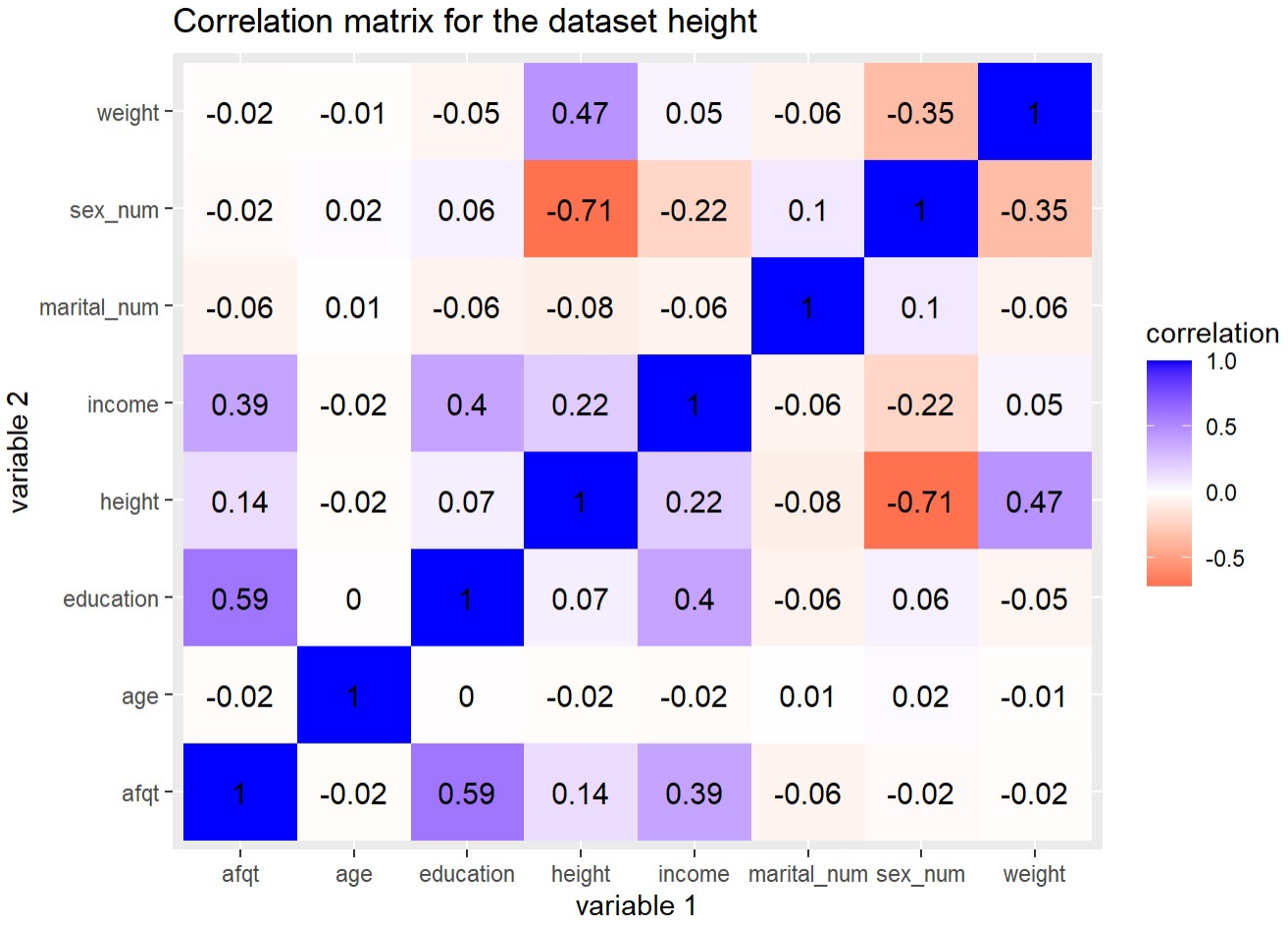



```
#Find the correlation between numeric variables
height_num <- height %>%
  select_if(is.numeric)
cor(height_num, use = "pairwise.complete.obs")
```

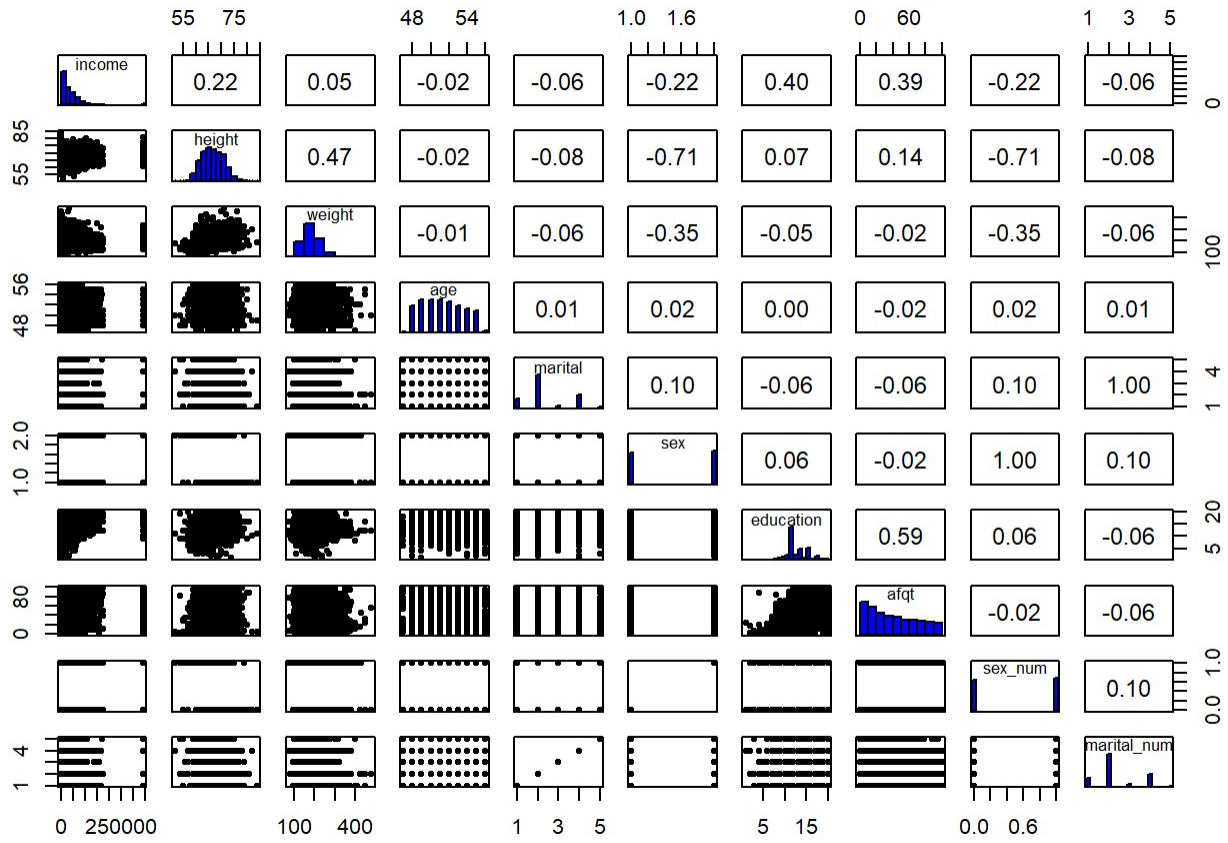
```
##                  income      height      weight          age    education
## income       1.00000000  0.21765001  0.05165439 -0.022207505  0.395366329
## height       0.21765001  1.00000000  0.46738699 -0.017951217  0.065709755
## weight       0.05165439  0.46738699  1.00000000 -0.011208878 -0.045353200
## age         -0.02220750 -0.01795122 -0.01120888  1.000000000 -0.002904316
## education    0.39536633  0.06570976 -0.04535320 -0.002904316  1.000000000
## afqt         0.39277088  0.14330694 -0.01502948 -0.017658098  0.591999252
## sex_num     -0.21544789 -0.71193342 -0.35051869  0.023768891  0.059389457
## marital_num -0.05717119 -0.08374923 -0.06131210  0.008661660 -0.061814773
##                    afqt     sex_num marital_num
## income       0.39277088 -0.21544789 -0.05717119
## height       0.14330694 -0.71193342 -0.08374923
## weight      -0.01502948 -0.35051869 -0.06131210
## age         -0.01765810  0.02376889  0.00866166
## education    0.59199925  0.05938946 -0.06181477
## afqt         1.00000000 -0.02471357 -0.05824574
## sex_num     -0.02471357  1.00000000  0.09623519
## marital_num -0.05824574  0.09623519  1.00000000
```

Average heighted people seemed to earn the most amount of money. The distribution seemed normally distributed. Slimmer people also seemed to earn more money, unfortunately. Income did not seem too related to age except at the ends with younger and older people. Anyone in between did not seem to make a difference. Separated and widowed people seemed to have the least amount of money. Sex did not play a big factor in this data. An increase in education and AFQT score increased income.

```
#Build a heatmap of the correlations between numeric variables
cor(height_num, use = "pairwise.complete.obs") %>%
  # Save as a data frame
  as.data.frame %>%
  # Convert row names to an explicit variable
  rownames_to_column %>%
  # Pivot so that all correlations appear in the same column
  pivot_longer(-1, names_to = "other_var", values_to = "correlation") %>%
  ggplot(aes(rowname, other_var, fill=correlation)) +
  # Heatmap with geom_tile
  geom_tile() +
  # Change the scale to make the middle appear neutral
  scale_fill_gradient2(low="red",mid="white",high="blue") +
  # Overlay values
  geom_text(aes(label = round(correlation,2)), color = "black", size = 4) +
  # Give title and labels
  labs(title = "Correlation matrix for the dataset height", x = "variable 1", y = "va
riable 2")
```



```
#Visualize correlation for all numeric variables with univariate and bivariate graphs
pairs.panels(height,
             method = "pearson",
             hist.col = "blue",
             smooth = FALSE, density = FALSE, ellipses = FALSE)
```



Manova

```
# Perform MANOVA with 5 response variables listed in cbind()
manova_height <- manova(cbind(income,height,weight,age,education,afqt) ~ marital, dat
a = height)

# OUtput of MANOVA
summary(manova_height)
```

```
##              Df   Pillai approx F num Df den Df    Pr(>F)
## marital       4 0.094484   26.765     24  26552 < 2.2e-16 ***
## Residuals 6640
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# MANOVA is significant to perform a one-way ANOVA for each variable
summary.aov(manova_height)
```

```
##  Response income :
##               Df     Sum Sq    Mean Sq F value    Pr(>F)
## marital        4 8.7272e+11 2.1818e+11   70.29 < 2.2e-16 ***
## Residuals   6640 2.0611e+13 3.1040e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response height :
##               Df Sum Sq Mean Sq F value    Pr(>F)
## marital        4   1081  270.20  16.356 2.435e-13 ***
## Residuals   6640 109691   16.52
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response weight :
##               Df   Sum Sq Mean Sq F value    Pr(>F)
## marital        4    59344 14836.0  7.5217 4.842e-06 ***
## Residuals   6640 13096988  1972.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response age :
##               Df Sum Sq Mean Sq F value    Pr(>F)
## marital        4    104  25.966  5.2297 0.0003331 ***
## Residuals   6640  32968   4.965
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response education :
##               Df Sum Sq Mean Sq F value    Pr(>F)
## marital        4   1434  358.47  54.509 < 2.2e-16 ***
## Residuals   6640  43667    6.58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response afqt :
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## marital        4  386525   96631  122.94 < 2.2e-16 ***
## Residuals   6640 5219270     786
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# ANOVA is significant to perform post-hoc analysis for each
# For income
pairwise.t.test(height$income,height$marital, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  height$income and height$marital
##
##           single  married separated divorced
## married   < 2e-16 -       -         -
## separated 0.07547 < 2e-16 -         -
## divorced  0.00420 < 2e-16 0.00016   -
## widowed   0.62080 5.6e-09 0.48921   0.06308
##
## P value adjustment method: none
```

```
# For height
pairwise.t.test(height$height,height$marital, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  height$height and height$marital
##
##           single  married separated divorced
## married   0.88616 -       -         -
## separated 0.00132 0.00057 -         -
## divorced  0.00014 1.5e-06 0.44621   -
## widowed   3.8e-09 9.7e-10 0.00133   2.6e-05
##
## P value adjustment method: none
```

```
# For weight
pairwise.t.test(height$weight,height$marital, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  height$weight and height$marital
##
##           single  married separated divorced
## married   0.0003  -       -         -
## separated 0.0065  0.4514  -         -
## divorced  8.4e-08 0.0037  0.4284    -
## widowed   0.0223  0.3855  0.7625    0.8332
##
## P value adjustment method: none
```

```
# For age
pairwise.t.test(height$age,height$marital, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  height$age and height$marital
##
##           single  married separated divorced
## married   0.03933 -       -         -
## separated 0.71712 0.37853 -         -
## divorced  0.80320 0.04470 0.83548   -
## widowed   4.9e-05 0.00071 0.00070   5.9e-05
##
## P value adjustment method: none
```

```
#For education
pairwise.t.test(height$education,height$marital, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  height$education and height$marital
##
##           single  married separated divorced
## married   < 2e-16 -       -         -
## separated 0.0021  < 2e-16 -         -
## divorced  0.1149  < 2e-16 2.1e-05   -
## widowed   0.2174  3.5e-08 0.3876    0.0453
##
## P value adjustment method: none
```

```
# For AFQT
pairwise.t.test(height$afqt,height$marital, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  height$afqt and height$marital
##
##           single  married separated divorced
## married   < 2e-16 -       -         -
## separated 8.6e-05 < 2e-16 -         -
## divorced  7.0e-08 < 2e-16 1.4e-14   -
## widowed   0.16    < 2e-16 0.21      6.7e-05
##
## P value adjustment method: none
```

66 tests were performed. The probability of at least one type one error is 0.968. The significance level was adjusted for multiple comparisons (Bonferroni a = 0.0007).

A one-way MANOVA was conducted to determine the effect of Marital Status (Single, Married, Separated, Divorced) on five dependent variables (Income, Height, Weight, Age, Education, and AFQT). Significant differences were found among the four Marital Statuses for at least one of the dependent variables(Pillali's trace = 0.093, pseudo P(24, 26552) = 26.765 , p<.0001.

Univariate ANOVAs for each dependent variable were conducted as a follow-up tests to the MANOVA, were also significant for Income (F(4,6640) = 70.29, p<.0001), Height (F(4,6640) = 16.36, p<.0001), Weight (F(4,6640) = 7.52, p<.0001), Age (F(4,6640) = 5.23, p<.0001), Education (F(4,6640) = 54.51, p<.0001), and AFQT (F(4,6640) = 122.94, p<.0001).

Post hoc analysis was performed conducting pairwise comparisons to determine which Marital Status differed in Income, Height, Weight, Age, Education, AFQT.For income, single and separated, single and divorced, single and widowed, separated and widowed, and divorced and separated all did not have significant differences. For height, single and married, single and separated, separated and divorced, and separated and widowed did not have significant differences. For weight, only sinlge and divorced had a significant difference. For age, single and widowed, separated and widowed, and divorced and widowed had significant differences. For education, Single and separated, single and divorced, single and widowed, separated and widowed, and divorced and widowed did not have significant differences. For AFQT, only single and widowed, and separated and widowed did not have significant differences.

Assumptions for were most likely all met except for possibility of outliers and the homogeneity of within-group covariances when viewing the results of the tests. The data had random independent observations, with multivariate responses of the numeric variables.

Randomization Test

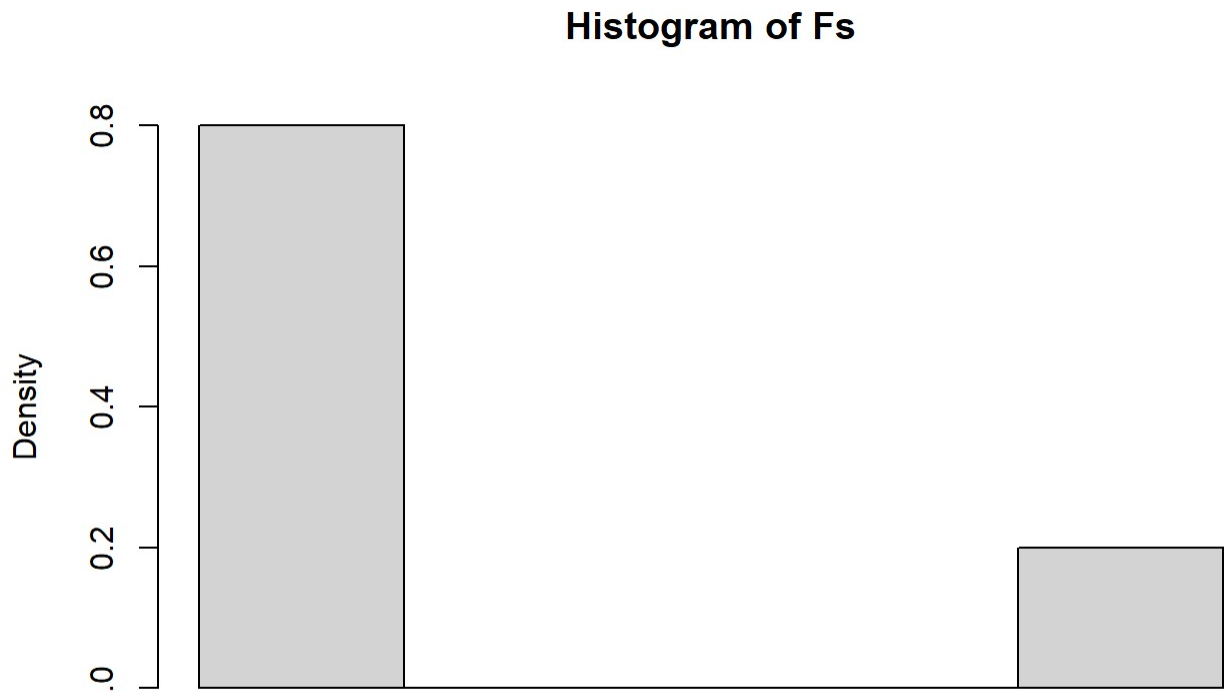
```
# Run ANOVA to compare the income by different heights to get F statistic
summary(aov(income ~ height, data = height))
```

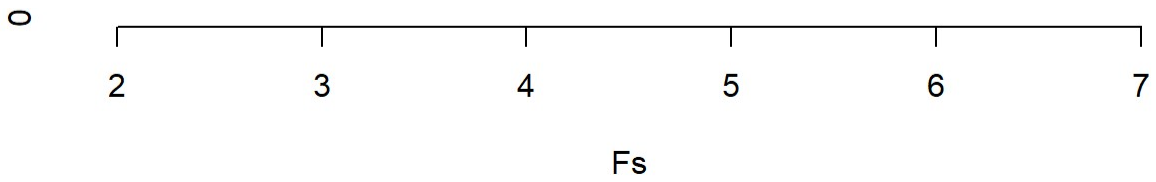
```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## height         1 1.018e+12 1.018e+12   330.3 <2e-16 ***
## Residuals   6643 2.047e+13 3.081e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H0: There is no significant difference between income and height

HA: There is significant difference between income and height

```r
# Observed F-statistic, running anova
obs_F <- 330.3

# Randomization test (using replicate)
Fs <- replicate(5,{
  # Randomly permute the response variable across income
  new <- height %>%
    mutate(income = sample(income))
  # Compute variation within groups
  SSW <- new %>%
    group_by(height) %>%
    summarize(SSW = sum((income - mean(income))^2)) %>%
    summarize(sum(SSW)) %>%
    pull
  # Compute variation between groups
  SSB <- new %>%
    mutate(mean = mean(income)) %>%
    group_by(height) %>%
    mutate(groupmean = mean(income)) %>%
    summarize(SSB = sum((mean - groupmean)^2)) %>%
    summarize(sum(SSB)) %>%
    pull
  # Compute the F-statistic (ratio of MSB and MSW)
  # df for SSB is 10 groups - 1 = 9
  # df for SSW is 6645 Observations - 10 groups = 6635
  (SSB/9)/(SSW/6635)
})

# Represent the distribution of the F-statistics for each randomized sample
hist(Fs, prob=T); abline(v = obs_F, col="red",add=T)
```






```
# Calculate the proportion of F statistic that are greater than the observed F-statis
tic
mean(Fs > obs_F)
```

```
## [1] 0
```

The F-statistic was 330.3 from the original sample. This value appeard 0 times in the sample randomized. The F-statistic observed from the original sample is very different from the distribution of the F-Statistic if the data was mixed up. It was very unlikely to observe alone, so we reject the null hypothesis. Height is a continuous variable.

Linear Regression

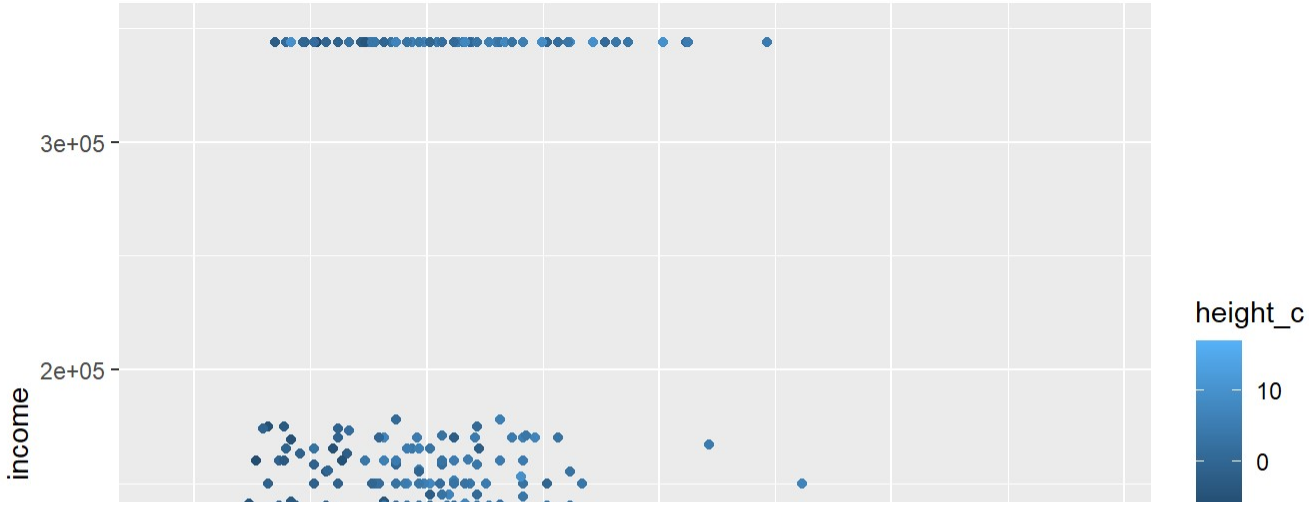
```
# Fit a multiple linear regression model with both predictors
fit <- lm(income ~ height + weight*height, data = height)
summary(fit)
```

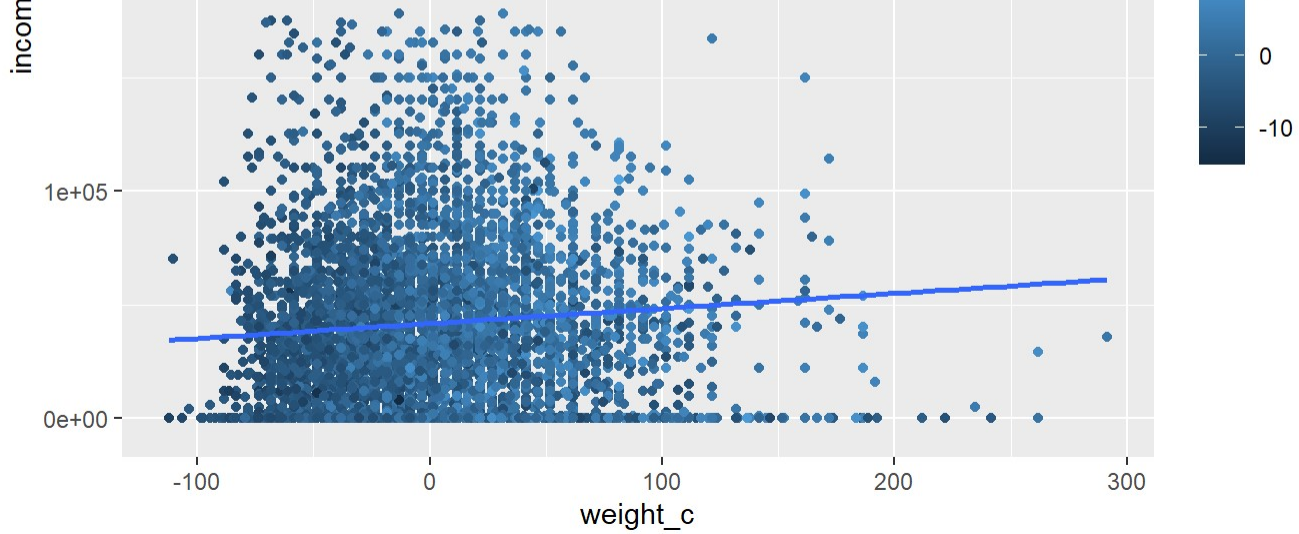
```
##
## Call:
## lm(formula = income ~ height + weight * height, data = height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -100812  -31099  -11073   14835  322415
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.164e+05  4.652e+04  -4.652 3.36e-06 ***
## height         4.079e+03  7.000e+02   5.827 5.90e-09 ***
## weight         1.393e+02  2.369e+02   0.588    0.557
## height:weight -3.286e+00  3.510e+00  -0.936    0.349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55420 on 6641 degrees of freedom
## Multiple R-squared:  0.0507, Adjusted R-squared:  0.05028
## F-statistic: 118.2 on 3 and 6641 DF,  p-value: < 2.2e-16
```

```r
# Center the data around the means
height$height_c <- height$height - mean(height$height)

# Center the data around the means
height$weight_c <- height$weight - mean(height$weight)

# Include an interaction term in the regression model with centered predictors
fit_c <- lm(income ~ height_c + weight_c * height_c, data = height)
summary(fit_c)
```

```
##
## Call:
## lm(formula = income ~ height_c + weight_c * height_c, data = height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -100812  -31099  -11073   14835  322415
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       42098.024    742.278  56.715  < 2e-16 ***
## height_c           3460.376    188.797  18.329  < 2e-16 ***
## weight_c            -81.274     17.294  -4.700 2.66e-06 ***
## height_c:weight_c    -3.286      3.510  -0.936    0.349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55420 on 6641 degrees of freedom
## Multiple R-squared:  0.0507, Adjusted R-squared:  0.05028
## F-statistic: 118.2 on 3 and 6641 DF,  p-value: < 2.2e-16
```

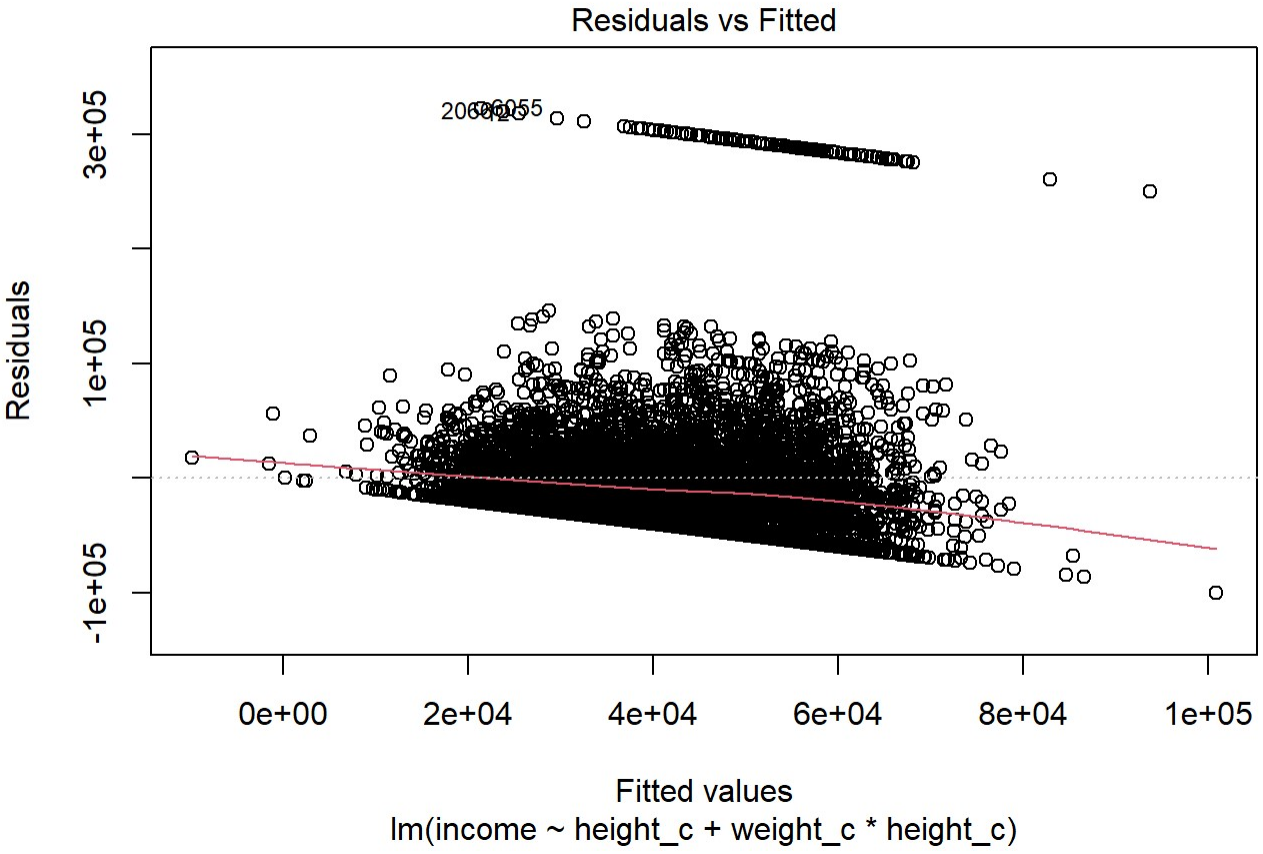
```r
# Visualize the relationships between the three variables
ggplot(height, aes(x = weight_c, y = income, color = height_c)) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE)
```



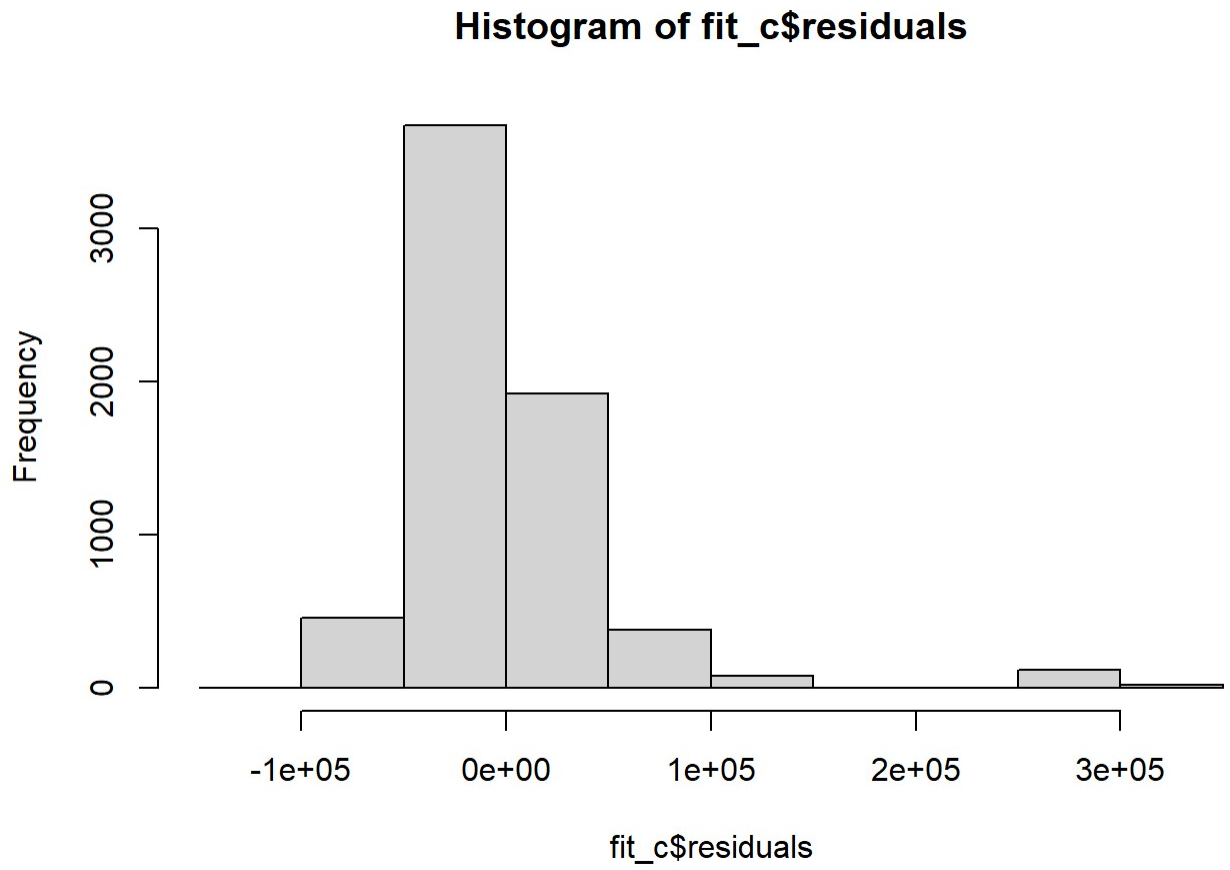


Controlling for weight, there is a significant effect of height on income. For every one unit increase in height, income increases 3460.38 dollars on average (t = 56.715, df = 6641, p < .001). Controlling for height, there is a significant effect of weight on income. For every one unit increase in weight, income decreases -81.274 dollars on average (t = 18.329, df = 6641, p < .001).After controlling for weight and height, there is no difference in income(t = -0.936, df = 6641, p = 0.349). This only explains about 5 percent of our variation.

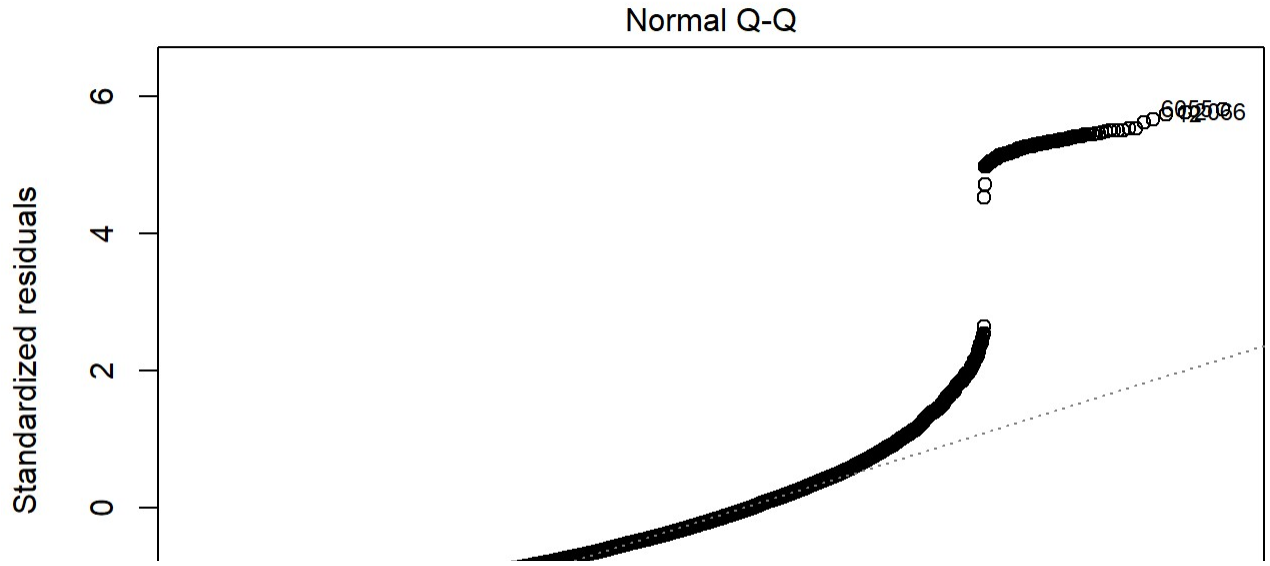
```
# Check for assumtions
# Residuals against fitted values plot to check for any problematic patterns (nonline
ar, equal variance)
plot(fit_c, which = 1)
```

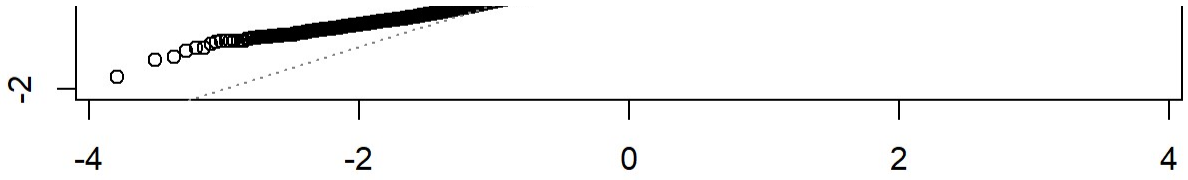


```
# Histogram of residuals
hist(fit_c$residuals)
```




```
# Q-Q plot to check for normality of the residuals
plot(fit_c, which = 2)
```





Theoretical Quantiles
lm(income ~ height_c + weight_c * height_c)

```
# Uncorrected Standard Errors
summary(fit_c)$coef
```

```
##                       Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)       42098.024447 742.278145 56.7146220 0.000000e+00
## height_c           3460.375749 188.797142 18.3285389 3.061048e-73
## weight_c            -81.273764  17.293512 -4.6996679 2.658375e-06
## height_c:weight_c    -3.285764   3.510224 -0.9360554 3.492787e-01
```

```
# Robust Standard Errors
coeftest(fit_c, vcov = vcovHC(fit_c))
```

```
##
## t test of coefficients:
##
##                     Estimate Std. Error t value  Pr(>|t|)
## (Intercept)       42098.0244   758.0963 55.5312 < 2.2e-16 ***
## height_c           3460.3757   208.2839 16.6137 < 2.2e-16 ***
## weight_c            -81.2738    15.7152 -5.1717 2.388e-07 ***
## height_c:weight_c    -3.2858     3.2626 -1.0071    0.3139
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All tests passed for linearity, normality, and homoscedasticity.

The estimates all stayed the same. All of the standard errors changed slightly to account for assumptions that were not met. If we were trying to build a confidence interval, it would be wider for height and weight, but smaller for their interaction. This would help us to lessen the possibility of the rejection of the null hypothesis. The p value also changed for all of the categories.

```
# Use the function replicate to repeat the process
samp_SEs <- replicate(5000, {
  # Bootstrap your data (resample observations)
  boot_data <- sample_frac(height, replace = TRUE)
  # Fit regression model
  fitboot <- lm(income ~ height_c + height_c*weight_c, data = boot_data)
  # Save the coefficients
  coef(fitboot)
})

# Estimated SEs
samp_SEs %>%
  # Transpose the obtained matrices
  t %>%
  # Consider the matrix as a data frame
  as.data.frame %>%
  # Compute the standard error (standard deviation of the sampling distribution)
  summarize_all(sd)
```

```
##   (Intercept) height_c weight_c height_c:weight_c
## 1    761.5573 207.3635 15.76739          3.310414
```

```
# We can also consider a confidence interval for the estimates
samp_SEs %>%
  # Transpose the obtained matrices
  t %>%
  # Consider the matrix as a data frame
  as.data.frame %>%
  # Pivot longer to group by and summarize each coefficient
  pivot_longer(everything(), names_to = "estimates", values_to = "value") %>%
  group_by(estimates) %>%
  summarize(lower = quantile(value,.025), upper = quantile(value,.975))
```

```
## # A tibble: 4 x 3
##   estimates            lower    upper
## * <chr>                <dbl>    <dbl>
## 1 (Intercept)       40638.   43610.
## 2 height_c           3048.    3870.
## 3 height_c:weight_c    -9.59     3.32
## 4 weight_c           -112.     -49.5
```

```
# Compare to original fit
confint(fit_c, level = 0.95)
```

```
##                          2.5 %       97.5 %
## (Intercept)       40642.92082 43553.128079
## height_c           3090.27270  3830.478800
## weight_c           -115.17460   -47.372924
## height_c:weight_c   -10.16693     3.595403
```

Comparisons were made to the bootstrap sample to the original model. Standard deviation increased for the lower height that was mean centered and decreased for the upper limit. Weight mean centered had the lower boundary increased and the upper boundary decreased. The interaction between the two had the lower boundary decreased and the upper boundary increased. This change was made for assumptions not having been met. Losing assumptions means we would lose precision in the estimation.

Logistic Regression

```
# Fit a linear regression model
fit_lor <- glm(sex_num ~ marital, data = height, family = "binomial")
summary(fit_lor)
```

```
##
## Call:
## glm(formula = sex_num ~ marital, family = "binomial", data = height)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8591  -1.1751   0.6254   1.1798   1.2724
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.22059    0.06192  -3.562 0.000367 ***
## maritalmarried    0.21506    0.07028   3.060 0.002214 **
## maritalseparated  0.62605    0.12613   4.963 6.93e-07 ***
## maritaldivorced   0.40838    0.08107   5.038 4.71e-07 ***
## maritalwidowed    1.75306    0.22104   7.931 2.17e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9207.1  on 6644  degrees of freedom
## Residual deviance: 9103.8  on 6640  degrees of freedom
## AIC: 9113.8
##
## Number of Fisher Scoring iterations: 3
```

```
# Based on predicted probabilities...
height$prob1 <- predict(fit_lor, type = "response")

# ... we can classify a sex as male or female (apply a cutoff of 0.5)
height$predicted <- ifelse(height$prob1 > .5, "female", "male")

# Confusion matrix: compare true to predicted condition
table(true_condition = height$sex, predicted_condition = height$predicted) %>%
  addmargins
```

```
##               predicted_condition
## true_condition female male  Sum
##         male      833 2400 3233
##         female   1138 2274 3412
##         Sum      1971 4674 6645
```

```
# Accuracy (correctly classified cases)
(1138 + 2400)/6645
```

```
## [1] 0.5324304
```

```
# Sensitivity (true positive rate) females
1138 / 3412
```

```
## [1] 0.3335287
```

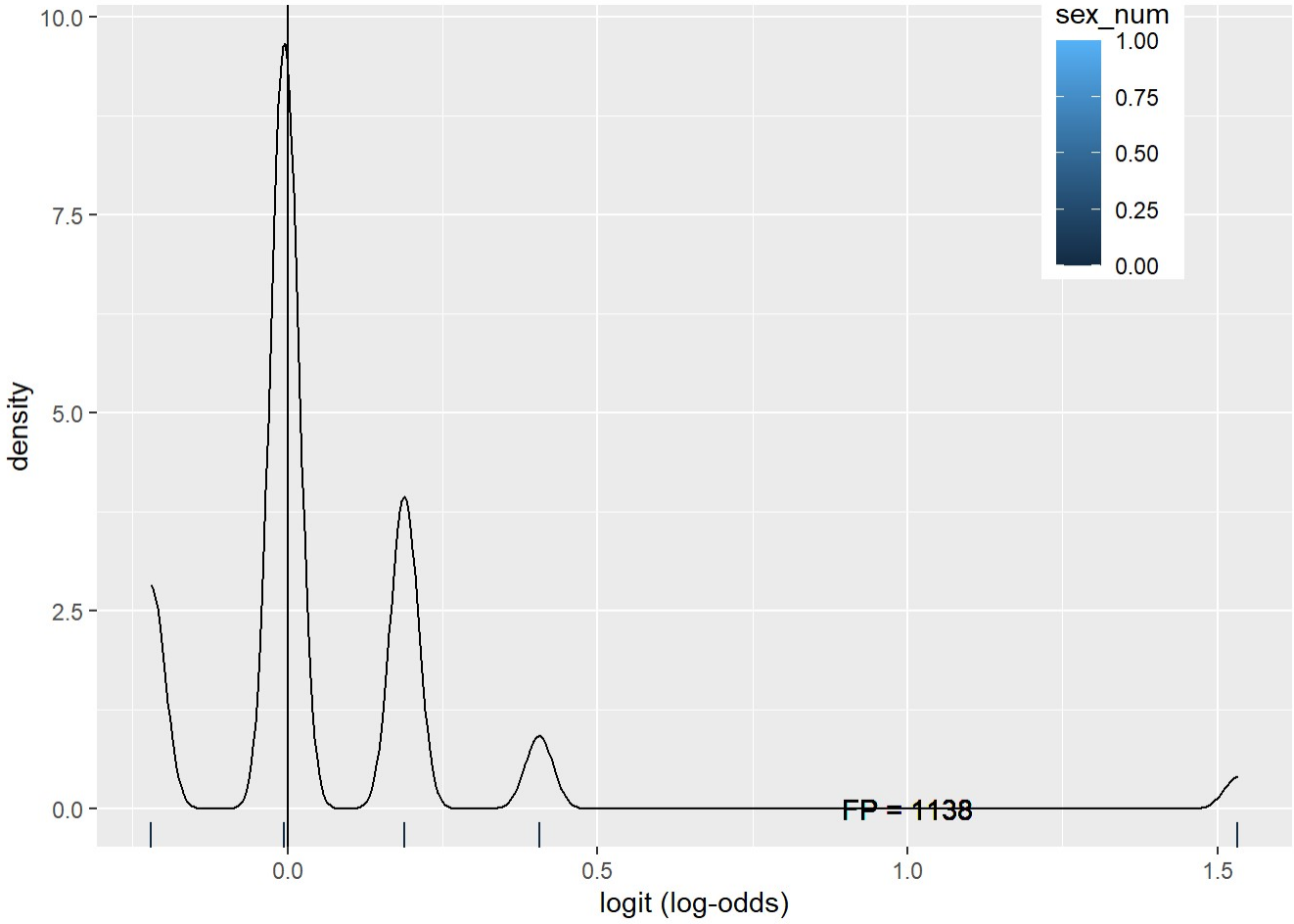
```
# Specificity (true negative rate) males
2400 / 3233
```

```
## [1] 0.7423446
```

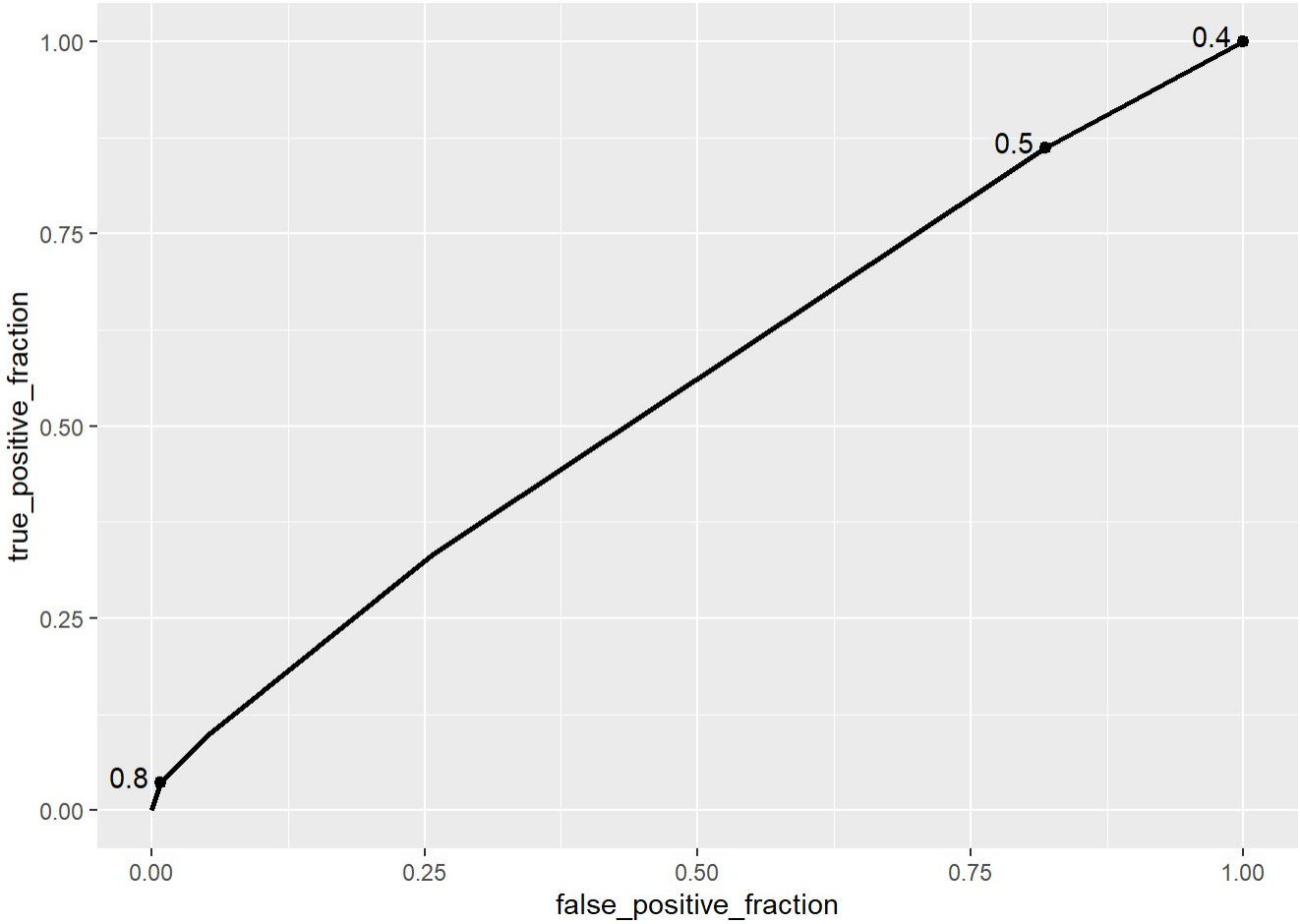
```
# Precision (Positive Predictive Value, PPV)
1138/1971
```

```
## [1] 0.5773719
```

```
# Predicted log odds
height$logit <- predict(fit_lor, type = "link")

# Density plot of log-odds for each outcome
height %>%
  ggplot() +
  geom_density(aes(logit, color = sex_num, fill = sex_num), alpha = .4) +
    geom_rug(aes(logit, color = sex_num)) +
  geom_text(x = -5, y = .07, label = "TN = 2400") +
  geom_text(x = -1.75, y = .008, label = "FN = 2274") +
  geom_text(x = 1, y = .006, label = "FP = 1138") +
  geom_text(x = 5, y = .04, label = "TP = 833") +
  theme(legend.position = c(.85,.85)) +
  geom_vline(xintercept = 0) +
  xlab("logit (log-odds)")
```



```
# Call the library plotROC
library(plotROC)

# Based on predicted probabilities...
height$prob1 <- predict(fit_lor, type = "response")

# Plot ROC depending on values of sex_num and its probabilities displaying some cutof
f values
ROCplot1 <- ggplot(height) +
  geom_roc(aes(d = sex_num, m = prob1), cutoffs.at = list(0.1, 0.5, 0.9))
ROCplot1
```




```
# Calculate the area under the curve still using the library plotROC with function ca
lc_auc
calc_auc(ROCplot1)
```

```
##   PANEL group       AUC
## 1     1    -1 0.5517618
```

All of the categories for the glm were significant. For every male, marital status would increase by about .215 for married, .626 for separated, .408 for divorced, and 1.75 for widowed. It would decrease -.220 for single males.

Accuracy shows the proportion of accurately classified cases that is about .53 for this data. Females sensitivity for proportion of true positive is 0.33 for this data. Male specificity for proportion of true negative is 0.74 for this data. The proportion of positive prediction is given by the precision 0.58

The ROC plot is very close to 45 degrees with a AUC of about .55. This is not a very accurate test that was performed for marital status versus sex.

```
##           sysname           release           version          nodename
##         "Windows"         "10 x64"     "build 19041" "DESKTOP-1SMB59F"
##           machine             login              user    effective_user
##          "x86-64"           "Admin"           "Admin"           "Admin"
```