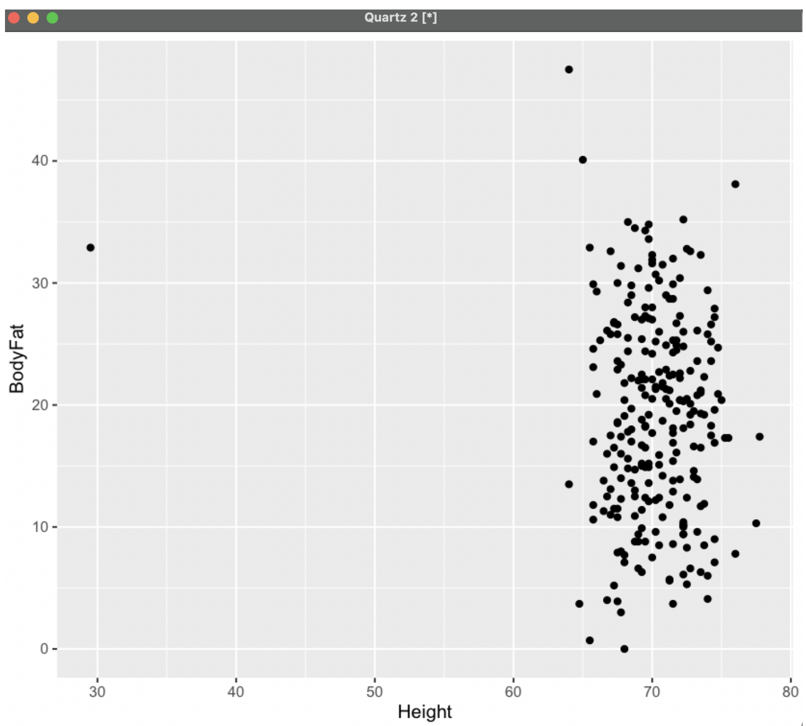
Homework 5

Prepare your answers as a **single PDF file**.
Group work: You may work in groups of 1-3. Include all group member names in the PDF file. You may work with students in both sections (375-01, -02). Only one person in the group should submit to Canvas.
Due: check on Canvas.

Body fat percentage refers to the relative proportions of body weight in terms of lean body mass (muscle, bone, internal organs, and connective tissue) and body fat. The most accurate means of estimating body fat percentage are cumbersome and require specialized equipment. Instead, we can estimate body fat percentage from other measurements.

The bodyfat.csv file in the Datasets module on Canvas contains 13 measurements from subjects (all men) along with their body fat percentage[1]. Read the file and answer the following questions.

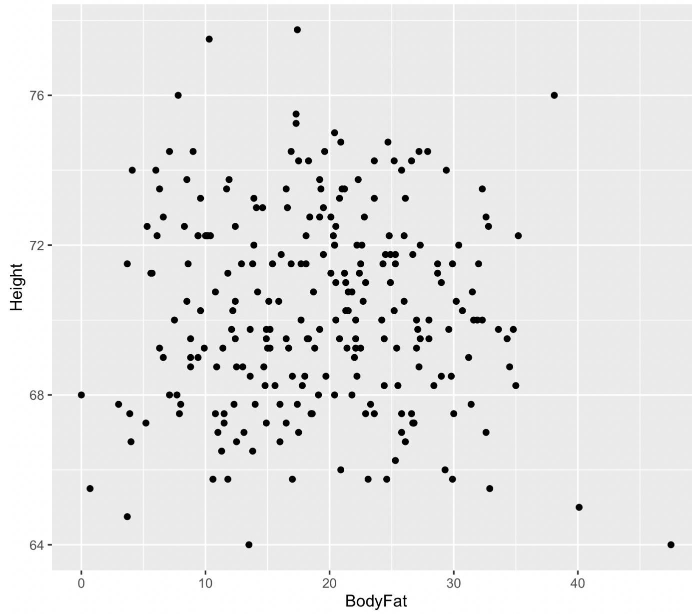
a. Plot `BodyFat` vs. `Height` (code, plot) Which is the dependent variable? Which is the independent variable?
    i. BodyFat is the dependent variable and height is the independent variable. **ggplot(data=bodyfat) + geom_point(mapping=aes(x = Height, y = BodyFat))**



[1] https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset?resource=download

b. There is one obvious outlier in the Height column. Remove the corresponding row from the data and plot again. (Show: code to remove the row, plot). This will be the data used for the following questions. Confirm that the mean Height is now 70.31076.

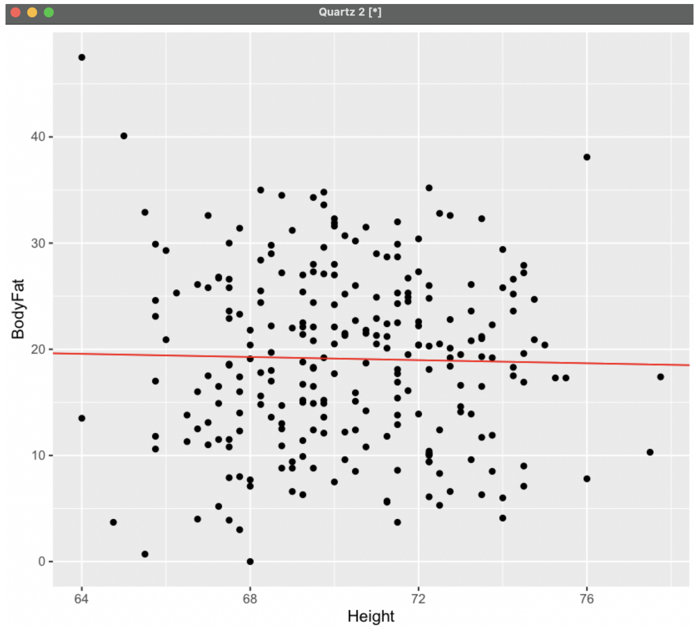
bodyfat2 <- bodyfat %>% filter(Height > min(Height))
ggplot(bodyfat2) + geom_point(mapping = aes(x = BodyFat, y = Height))
bodyfat %>% filter(Height > min(Height)) %>% summarise(mean(Height))



c. Create a linear model of `BodyFat` vs. `Height`. (code, output of summary(model))

m <- lm(formula = BodyFat ~ Height, data = bodyfat2)
mycf <- coef(m)
ggplot(data=bodyfat2) + geom_point(mapping = aes(x=Height, y=BodyFat)) + geom_abline(slope = mycf[2], intercept = mycf[1], color="red")

```
> summary(m)

Call:
lm(formula = BodyFat ~ Height, data = bodyfat2)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.268  -6.697   0.286   6.162  27.933 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  24.3412    14.2206   1.712   0.0882 .
Height       -0.0746     0.2021  -0.369   0.7124  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.355 on 249 degrees of freedom
Multiple R-squared:  0.0005468, Adjusted R-squared:  -0.003467 
F-statistic: 0.1362 on 1 and 249 DF,  p-value: 0.7124
```



i. What is the R2 value?
0.0005468

ii. Is this a “good” model? Why or why not?
No, because the R2 value is very low. This means there is a lot of variance in the data.

iii. What is the linear equation relating BodyFat and Height according to this model?
lm(formula = BodyFat ~ Height, data = bodyfat2)

d. Create a linear model of `BodyFat` vs. **`Weight`**. (code, output of summary(model))
m <- lm(formula = BodyFat ~ Weight, data = bodyfat2)
mycf <- coef(m)

```
> summary(m)

Call:
lm(formula = BodyFat ~ Weight, data = bodyfat2)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.7382  -4.7052   0.0973   4.9305  21.4419 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -11.88891    2.57914   -4.61 6.45e-06 ***
Weight        0.17327    0.01423   12.17  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.616 on 249 degrees of freedom
Multiple R-squared:  0.3731,	Adjusted R-squared:  0.3706 
F-statistic: 148.2 on 1 and 249 DF,  p-value: < 2.2e-16
```

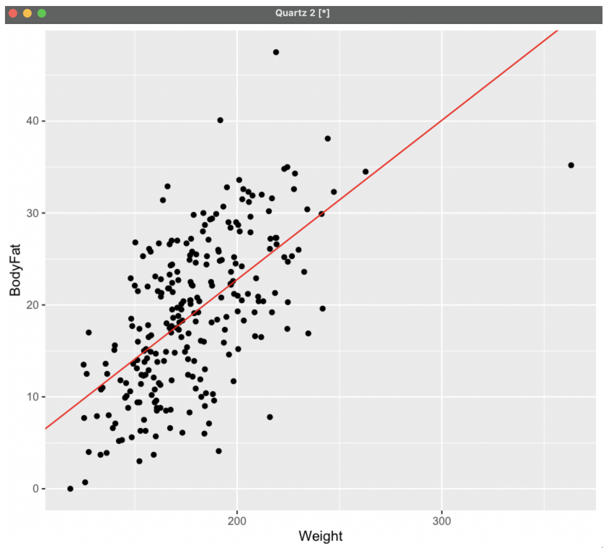
i. What is the R2 value?
0.3731

ii. Is this a better model than that based on Height? Why or why not?

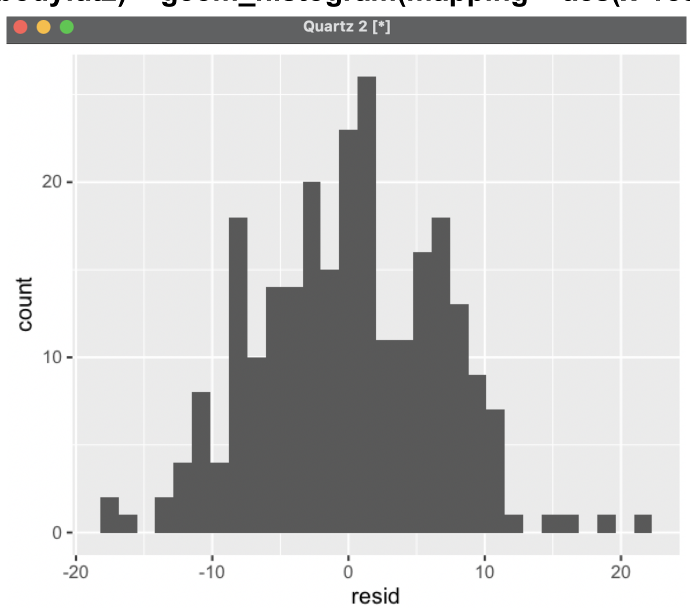
Yes because there was much less data variance as shown by the R-squared value.

iii. What is the linear equation relating BodyFat and Weight according to this model?
lm(formula = BodyFat ~ Weight, data = bodyfat2)

iv. Plot BodyFat vs. Weight and overlay the best fit line. Use a different color for the line. (plot, code)
ggplot(data=bodyfat2) + geom_point(mapping = aes(x=Weight, y=BodyFat)) + geom_abline(slope = mycf[2], intercept = mycf[1], color="red")



v. Plot the histogram of residuals (plot, code). Does this show an approximately normal distribution?
bodyfat2 <- bodyfat2 %>% add_residuals(m)
ggplot(data=bodyfat2) + geom_histogram(mapping = aes(x=resid))



The histogram has a relatively normal distribution.

vi. From the model, predict the BodyFat for two persons: Person A weighs 175 lbs, Person B weighs 250 lbs. Include the 99% **confidence** intervals for the predictions. In which prediction (for Person A or B), are you more confident? Why?

```
> predict(m, predx, interval = "confidence", level = 0.99)
       fit      lwr      upr
1 18.43402 17.34082 19.52722
2 31.42956 28.58522 34.27390
```

A because the predicted Body Fat % is more closely related to actual results for a person who weighs ~175lbs at 70” height.

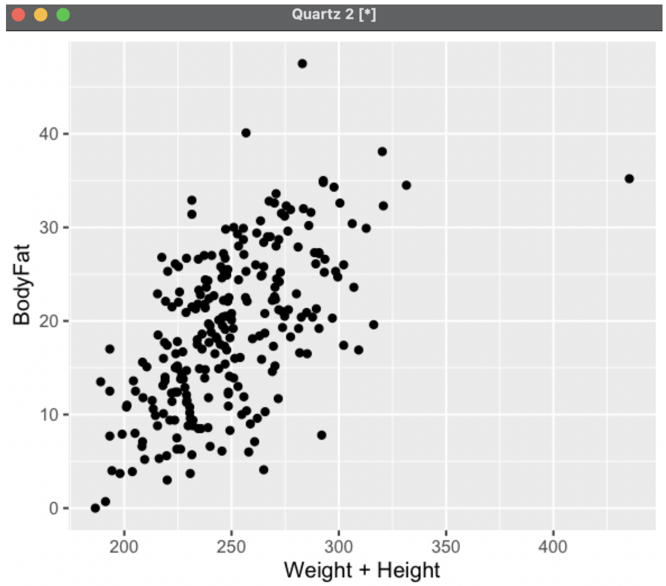
e. Create a linear model of `BodyFat` vs. **`Weight and Height`**. (code, output of summary(model))

```
> summary(m)

Call:
lm(formula = BodyFat ~ Weight + Height, data = bodyfat2)

Residuals:
     Min       1Q   Median       3Q      Max 
-24.0328  -3.6411   0.0281   4.3236  13.2125 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 72.52439   10.42582   6.956 3.09e-11 ***
Weight       0.23195    0.01446  16.037  < 2e-16 ***
Height      -1.34979    0.16265  -8.299 6.81e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.865 on 248 degrees of freedom
Multiple R-squared:  0.5094,	Adjusted R-squared:  0.5054 
F-statistic: 128.7 on 2 and 248 DF,  p-value: < 2.2e-16
```

i. What is the R2 value?
0.5094

ii. Is this a better model than that based only on Weight or Height? Why or why not?
Yes because there was less data variance when it was based on weight and height than when it was based on just one factor or the other. This is proven by the R-squared value.

iii. What is the linear equation relating BodyFat, Weight, and Height according to this model?
BodyFat = 72.52439 + 0.23195 x Weight + (-1.34979) x Height



iv. From the model, predict the BodyFat for two persons: Person A weighs 175 lbs, Person B weighs 250 lbs. Both persons have height=70". Include the 99% confidence intervals for the predictions. In which prediction (for Person A or B), are you more confident? Why?
I am more confident in the prediction for person A because the predicted Body Fat % is more closely related to actual results for a person who weighs ~175lbs at 70" height.

```
       fit      lwr      upr
1 18.62932 17.65829 19.60036
2 36.02524 33.12275 38.92772
```

f. Add a new transformed variable **BMI** = **Weight/Height^2** to the dataset. Create a linear model of `BodyFat` vs. **`BMI`**.

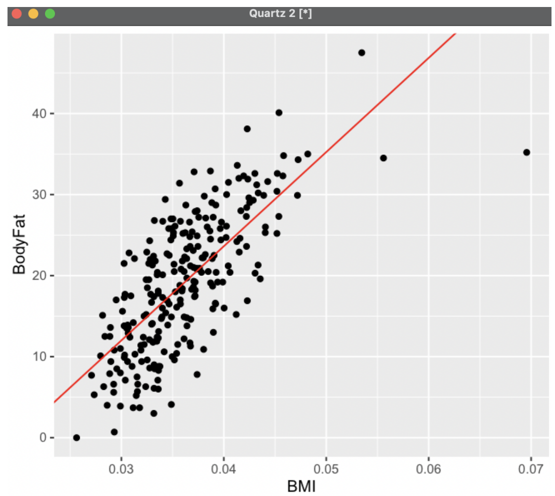
i. Give R code, output of summary(model)
bodyfat3 <- bodyfat2 %>% mutate(BMI = (Weight / Height^2))
m <- lm(formula = BodyFat ~ BMI, data = bodyfat3)
mycf <- coef(m)
summary(m)

```
>  summary(m)

Call:
lm(formula = BodyFat ~ BMI, data = bodyfat3)

Residuals:
     Min       1Q   Median       3Q      Max 
-22.7769  -3.7061   0.1652   4.1546  12.8061 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -22.859      2.553  -8.955   <2e-16 ***
BMI         1161.973     69.977  16.605   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.757 on 249 degrees of freedom
Multiple R-squared:  0.5255,	Adjusted R-squared:  0.5236 
F-statistic: 275.7 on 1 and 249 DF,  p-value: < 2.2e-16
```

ii. Is this a better model than the previous models? Why or why not?
Yes because there was less data variance when it was based on BMI than when it was based on Weight and Height or one or the other. This is proven by the R-squared value.

iii. What is the equation relating BodyFat, Weight, and Height according to this model? Is this a linear or nonlinear equation?
The equation relating BodyFat, Weight, and Height together is BMI = (Weight / Height^2). This is a linear equation.

iv. Plot `BodyFat` vs. `BMI` and overlay the best fit model as a straight line. (code, plot)
ggplot(data=bodyfat3) + geom_point(mapping = aes(x=BMI, y=BodyFat)) + geom_abline(slope = mycf[2], intercept = mycf[1], color="red")



v. From the model, predict the BodyFat for two persons: Person A weighs 175 lbs, Person B weighs 250 lbs. Both persons have height=70”. Include the 99% confidence intervals for the predictions.
predx <- data.frame(Weight = c(175, 250), Height = c(70, 70), BMI = c(175/(70^2),250/(70^2)))
predict(m, predx, interval = "confidence", level = 0.99)

vi. Body Mass Index (BMI) is actually defined as a person’s weight in kilograms divided by the square of height in meters[2] but your data has Weight in pounds and Height in inches. Thus, the correct BMI transformation should have been BMI = (Weight/2.20)/(Height*0.0254)^2. Would using this correct BMI transformation result in a different model from what was calculated? Why or why not?
Yes there is a different slope because every value needs to be converted to kg and meters which results in 704.55.

---

[2] https://www.cdc.gov/healthyweight/assessing/bmi/index.html

g. Add a new categorical variable (factor) **AgeGroup** to the dataset. AgeGroup should have three values: “Young” for Age<40, “Middle” for Age between 40 and 60, and “Older” for Age>60.

   i. Show R code that adds the AgeGroup variable. This can be done with mutate and the cut() function like so: `cut (Age, breaks = c(-Inf,40,60,Inf), labels = c("Young", "Middle", "Older")`[Code]
   bodyfat4 <- bodyfat3 %>% mutate(ageGroup=cut(Age, breaks = c(-Inf,40,60,Inf), labels = c("Young", "Middle", "Older")))

   ii. Create a linear model of `BodyFat` vs. **`BMI and AgeGroup`**.[Code, output of summary(model)]
   m <- lm(formula = BodyFat ~ BMI + ageGroup, data = bodyfat4)
   summary(m)

```
> summary(m)

Call:
lm(formula = BodyFat ~ BMI + ageGroup, data = bodyfat4)

Residuals:
     Min       1Q   Median       3Q      Max 
-21.4537  -3.9137  -0.1361   3.7127  12.0269 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -22.8344     2.4552  -9.301  < 2e-16 ***
BMI            1105.0576    67.8315  16.291  < 2e-16 ***
ageGroupMiddle    2.6113     0.7607   3.433    7e-04 ***
ageGroupOlder     5.3074     1.1075   4.792 2.85e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.502 on 247 degrees of freedom
Multiple R-squared:   0.57,	Adjusted R-squared:  0.5648 
F-statistic: 109.2 on 3 and 247 DF,  p-value: < 2.2e-16
```

   iii. How many dummy (i.e., 0-1) variables were created in the model?
   2

   iv. Is this a better model than the previous models? Why or why not?
   Yes because there was less data variance when the model included age group than when it was based on Weight and Height or one or the other. This is proven by the R-squared value being greater than previous models.

   v. What are the set of equations relating BodyFat, BMI, and AgeGroup according to this model?
   if Age < 40
   BodyFat = -22.8344+1105.0576*BMI
   if Age > 40 and Age < 60
   BodyFat = -22.8344+1105.0576*BMI+2.6113*1
   if Age > 60
   BodyFat = -22.8344+1105.0576*BMI+5.3074*1

   vi. Plot `BodyFat` vs. `BMI` and overlay the model predictions (Hint: add a new column with predictions and plot the predictions using geom_line. You should see multiple lines, one for each value of the discrete variable). [Code, plot]
   bodyfat4 <- bodyfat4 %>% add_predictions(m)

```
ggplot(data = bodyfat4) +
  geom_point(mapping = aes(y = BodyFat, x = BMI, color = ageGroup)) +
  geom_line(mapping = aes(x = BMI, y = pred, color = ageGroup))
```