Body Fat Data Project

Vanessa Kusi

December 28, 2019

Summary

This project uses the Bodyfat dataset from StatLib. The data includes 15 variables of various body measurements and estimates of the percentage of body fat determined by underwater weighing for 252 men. The goal of this project was to determine the best regression model to predict bodyfat and weight.

```r
library(psych)
```

```
## Warning: package 'psych' was built under R version 3.5.3
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.5.3
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```r
library(car)
```

```
## Warning: package 'car' was built under R version 3.5.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 3.5.3
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

Data Preparation

Load the Data into R and View a Summary of the Data

```r
body_fat <- read.csv("bodyfat.csv",header = TRUE)
```

View Attributes and Dimension of the Data

```r
names(body_fat)
```

```
##  [1] "Density" "bodyfat" "Age"     "Weight"  "Height"  "Neck"    "Chest"
##  [8] "Abdomen" "Hip"     "Thigh"   "Knee"    "Ankle"   "Biceps"  "Forearm"
## [15] "Wrist"
```

```r
dim(body_fat)
```

```
## [1] 252  15
```

There are a total of 252 observations and 15 Attributes in the Dataset. All of them are quantitative. Each one of them is explained below.

Density determined from underwater weighing Percent body fat from Siri's (1956) equation Age (years) Weight (lbs) Height (inches) Neck circumference (cm) Chest circumference (cm) Abdomen 2 circumference (cm) Hip circumference (cm) Thigh circumference (cm) Knee circumference (cm) Ankle circumference (cm) Biceps (extended) circumference (cm) Forearm circumference (cm) Wrist circumference (cm)

Missing Values

Determine if there are any missing values.

```r
sum(is.null(body_fat))
```

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

There are no missing values in the dataset.

View Structure of Dataset

```r
str(body_fat)
```

```
## 'data.frame':    252 obs. of  15 variables:
##  $ Density: num  1.07 1.09 1.04 1.08 1.03 ...
##  $ bodyfat: num  12.3 6.1 25.3 10.4 28.7 20.9 19.2 12.4 4.1 11.7 ...
##  $ Age    : int  23 22 22 26 24 24 26 25 25 23 ...
##  $ Weight : num  154 173 154 185 184 ...
##  $ Height : num  67.8 72.2 66.2 72.2 71.2 ...
##  $ Neck   : num  36.2 38.5 34 37.4 34.4 39 36.4 37.8 38.1 42.1 ...
##  $ Chest  : num  93.1 93.6 95.8 101.8 97.3 ...
##  $ Abdomen: num  85.2 83 87.9 86.4 100 94.4 90.7 88.5 82.5 88.6 ...
##  $ Hip    : num  94.5 98.7 99.2 101.2 101.9 ...
##  $ Thigh  : num  59 58.7 59.6 60.1 63.2 66 58.4 60 62.9 63.1 ...
##  $ Knee   : num  37.3 37.3 38.9 37.3 42.2 42 38.3 39.4 38.3 41.7 ...
##  $ Ankle  : num  21.9 23.4 24 22.8 24 25.6 22.9 23.2 23.8 25 ...
##  $ Biceps : num  32 30.5 28.8 32.4 32.2 35.7 31.9 30.5 35.9 35.6 ...
##  $ Forearm: num  27.4 28.9 25.2 29.4 27.7 30.6 27.8 29 31.1 30 ...
##  $ Wrist  : num  17.1 18.2 16.6 18.2 17.7 18.8 17.7 18.8 18.2 19.2 ...
```

All variables are numeric expect for Age, So converted Age to a numeric variable.

```r
as.numeric(body_fat$Age)
```

```
##   [1] 23 22 22 26 24 24 26 25 25 23 26 27 32 30 35 35 34 32 28 33 28 28 31 32 28
##  [26] 27 34 31 27 29 32 29 27 41 41 49 40 50 46 50 45 44 48 41 39 43 40 39 45 47
##  [51] 47 40 51 49 42 54 58 62 54 61 62 56 54 61 57 55 54 55 54 55 62 55 56 55 61
##  [76] 61 57 69 81 66 67 64 64 70 72 67 72 64 46 48 46 44 47 46 47 53 38 50 46 47
```

```
## [101] 49 48 41 49 43 43 43 52 43 40 43 43 47 42 48 40 48 51 40 44 52 44 40 47 50
## [126] 46 42 43 40 42 49 40 47 50 41 44 39 43 40 49 40 40 52 23 23 24 24 25 25 26
## [151] 26 26 27 27 27 28 28 28 30 31 31 33 33 34 34 35 35 35 35 35 35 35 35 36 36
## [176] 37 37 37 38 39 39 40 40 40 40 40 41 41 41 41 41 42 42 42 42 42 42 42 42 43
## [201] 43 43 43 44 44 44 44 47 47 47 49 49 49 50 50 51 51 51 52 53 54 54 54 55 55
## [226] 55 55 55 56 56 57 57 58 58 60 62 62 63 64 65 65 65 66 67 67 68 69 70 72 72
## [251] 72 74
```

```r
is.numeric(body_fat$Age)
```

```
## [1] TRUE
```

The below summary gives us the Minimum and Maximum values for all the attributes along with other values.

```r
summary(body_fat)
```

```
##     Density         bodyfat           Age            Weight     
##  Min.   :0.995   Min.   : 0.00   Min.   :22.00   Min.   :118.5  
##  1st Qu.:1.041   1st Qu.:12.47   1st Qu.:35.75   1st Qu.:159.0  
##  Median :1.055   Median :19.20   Median :43.00   Median :176.5  
##  Mean   :1.056   Mean   :19.15   Mean   :44.88   Mean   :178.9  
##  3rd Qu.:1.070   3rd Qu.:25.30   3rd Qu.:54.00   3rd Qu.:197.0  
##  Max.   :1.109   Max.   :47.50   Max.   :81.00   Max.   :363.1  
##      Height           Neck           Chest           Abdomen      
##  Min.   :29.50   Min.   :31.10   Min.   : 79.30   Min.   : 69.40  
##  1st Qu.:68.25   1st Qu.:36.40   1st Qu.: 94.35   1st Qu.: 84.58  
##  Median :70.00   Median :38.00   Median : 99.65   Median : 90.95  
##  Mean   :70.15   Mean   :37.99   Mean   :100.82   Mean   : 92.56  
##  3rd Qu.:72.25   3rd Qu.:39.42   3rd Qu.:105.38   3rd Qu.: 99.33  
##  Max.   :77.75   Max.   :51.20   Max.   :136.20   Max.   :148.10  
##       Hip            Thigh            Knee           Ankle           Biceps     
##  Min.   : 85.0   Min.   :47.20   Min.   :33.00   Min.   :19.1   Min.   :24.80  
##  1st Qu.: 95.5   1st Qu.:56.00   1st Qu.:36.98   1st Qu.:22.0   1st Qu.:30.20  
##  Median : 99.3   Median :59.00   Median :38.50   Median :22.8   Median :32.05  
##  Mean   : 99.9   Mean   :59.41   Mean   :38.59   Mean   :23.1   Mean   :32.27  
##  3rd Qu.:103.5   3rd Qu.:62.35   3rd Qu.:39.92   3rd Qu.:24.0   3rd Qu.:34.33  
##  Max.   :147.7   Max.   :87.30   Max.   :49.10   Max.   :33.9   Max.   :45.00  
##     Forearm          Wrist      
##  Min.   :21.00   Min.   :15.80  
##  1st Qu.:27.30   1st Qu.:17.60  
##  Median :28.70   Median :18.30  
##  Mean   :28.66   Mean   :18.23  
##  3rd Qu.:30.00   3rd Qu.:18.80  
##  Max.   :34.90   Max.   :21.40
```

To help get a better understanding of the distrubition of the variables, calculated the standard deviation of each attribute.

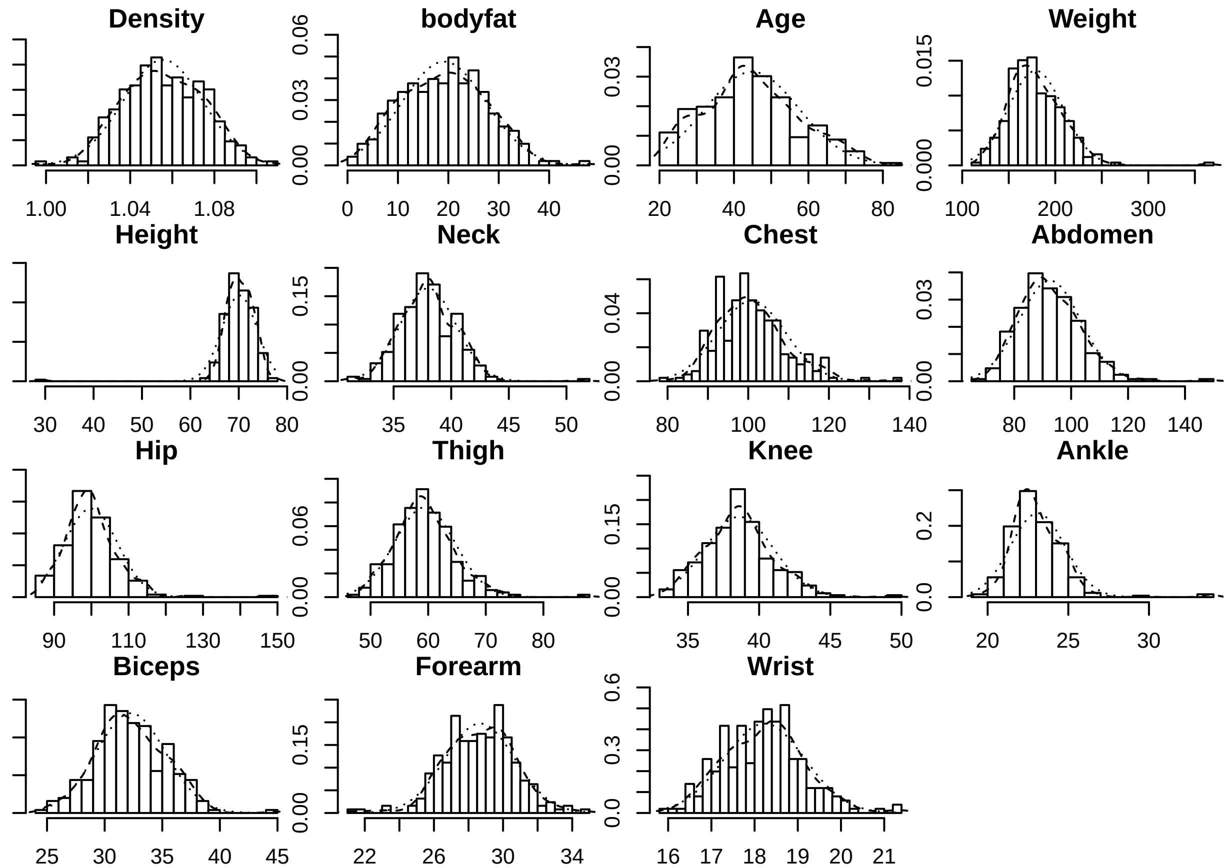
```r
sapply(body_fat, sd)
```

```
##     Density     bodyfat         Age      Weight      Height        Neck 
##  0.01903143  8.36874041 12.60203972 29.38915989  3.66285579  2.43091323 
##       Chest     Abdomen         Hip       Thigh        Knee       Ankle 
##  8.43047553 10.78307680  7.16405767  5.24995203  2.41180459  1.69489340 
##      Biceps     Forearm       Wrist 
##  3.02127375  2.02069117  0.93358493
```

Distribution and Outliers

Determine Distribution of the Attributes.

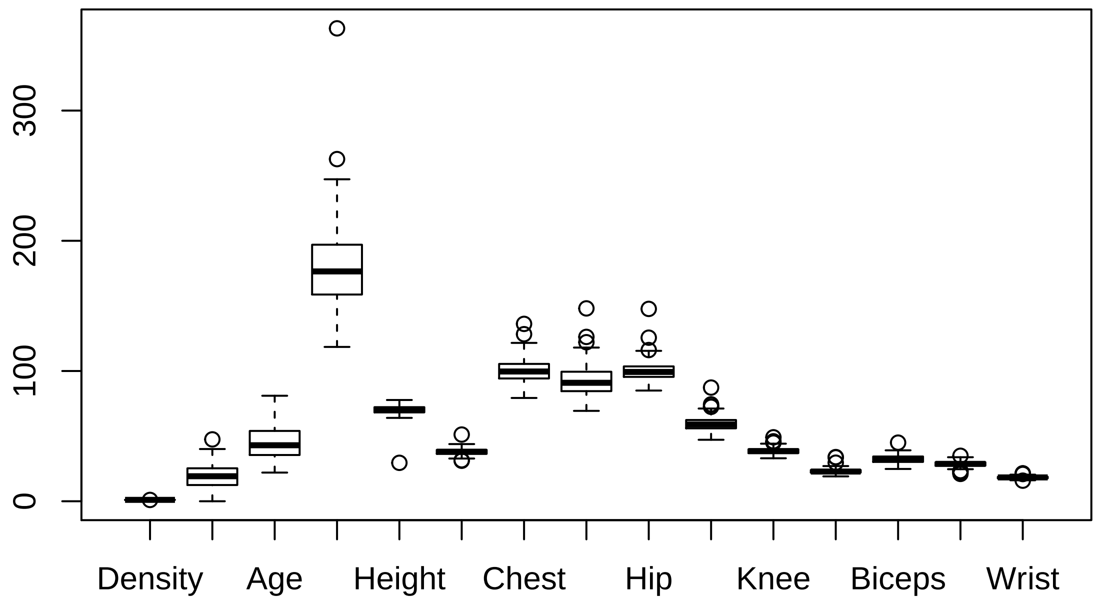
```r
library(psych)
multi.hist(body_fat)
```



All the attributtes expect for Weight, Height, Hip, Neck, Abdomen and Ankle appear normally distributed. Need tolook closer at these 6 attributes to determine why they are not normally distributed.

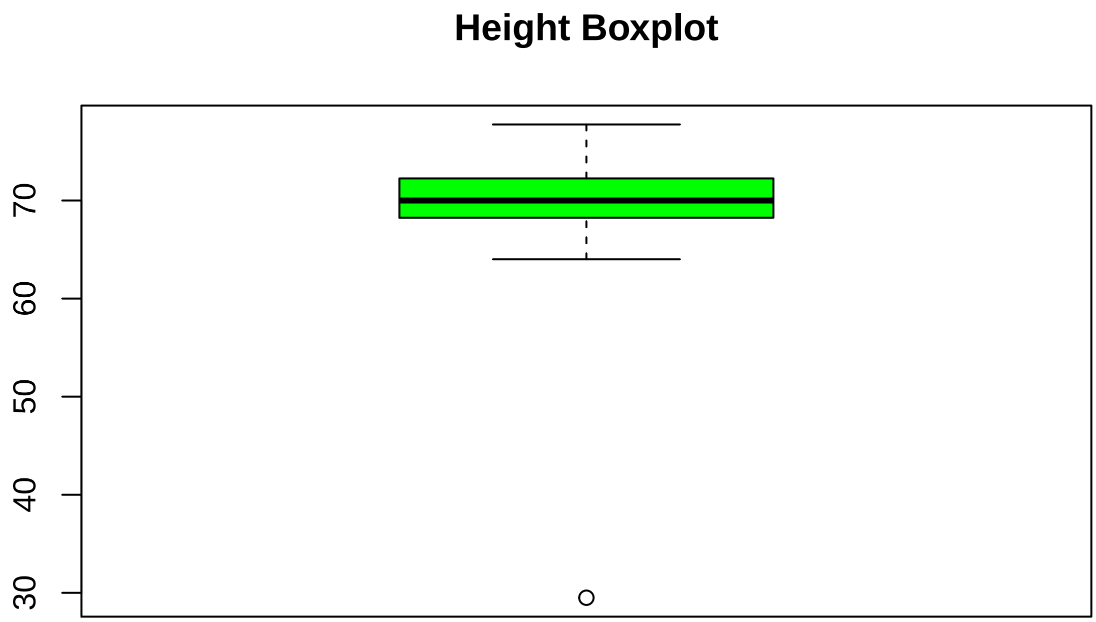
Determine Any Outlier Values. To further investigate attributes created boxplots to get a better understanding of the distribution.

```r
boxplot(body_fat)
```



Height Outlier

```
boxplot(body_fat$Height, main = "Height Boxplot", col = "green")
```




```
boxplot.stats(body_fat$Height)
```

```
## $stats
## [1] 64.00 68.25 70.00 72.25 77.75
##
## $n
## [1] 252
##
## $conf
## [1] 69.60188 70.39812
##
## $out
## [1] 29.5
```

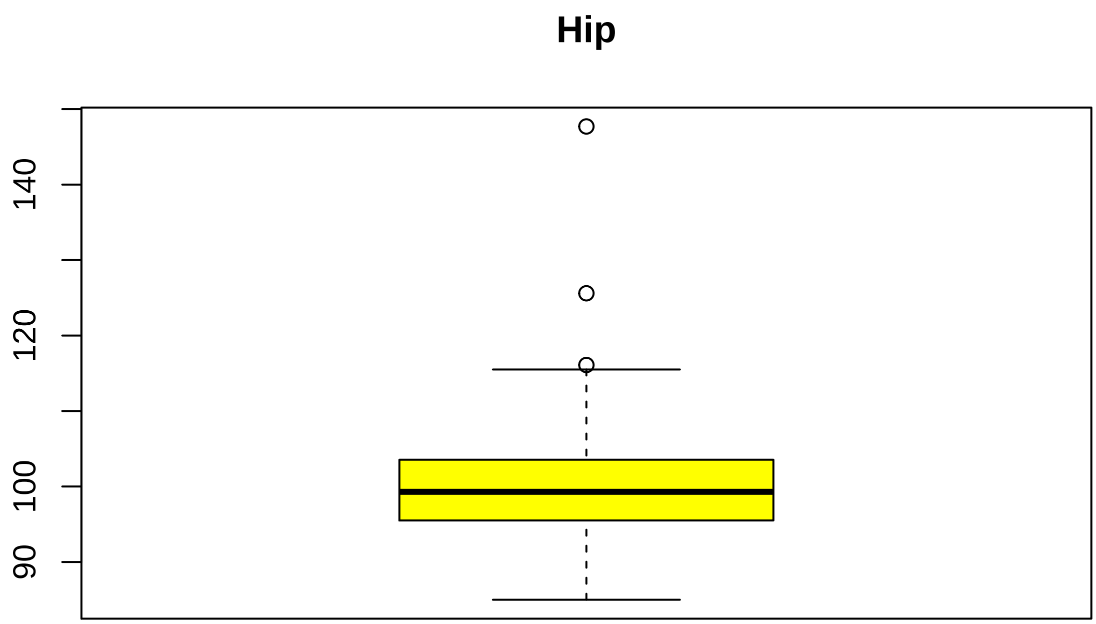
```
subset(body_fat,body_fat$Height == 29.50)
```

```
##    Density bodyfat Age Weight Height Neck Chest Abdomen   Hip Thigh Knee Ankle
## 42   1.025    32.9  44    205   29.5 36.6   106   104.3 115.5  70.6 42.5  23.7
##    Biceps Forearm Wrist
## 42   33.6    28.7  17.4
```

Outlier identified was a height of 29.5 inches. Looking for this observation in the dataset age associated with this height is 44 and bodyfat recorded was 32.9%. Seems like there may be an error because it seems very unlikely that any person especially an older male with high bodyfat (in the obese range) would be very short. So this record can be removed.

Hip Outliers

```r
boxplot(body_fat$Hip, main = "Hip", col = "yellow")
```




```r
boxplot.stats(body_fat$Hip)
```

```
## $stats
## [1]  85.00  95.50  99.30 103.55 115.50
##
## $n
## [1] 252
##
## $conf
## [1]  98.49878 100.10122
##
## $out
## [1] 116.1 147.7 125.6
```

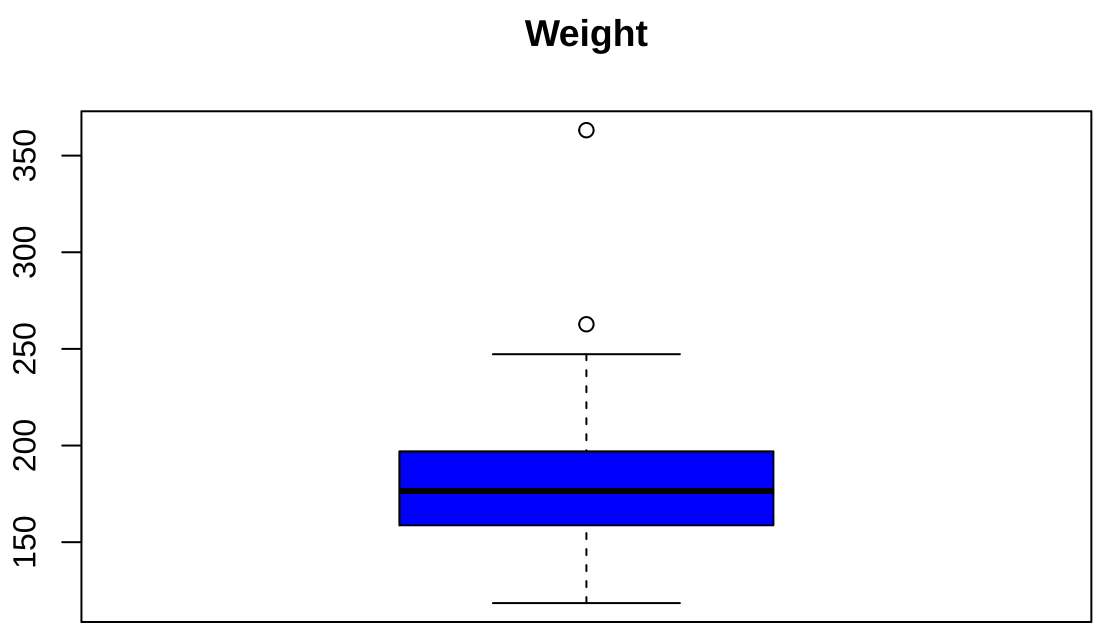
```r
hip_outliers<- boxplot(body_fat$Hip, plot = FALSE)$out
body_fat[which(body_fat$Hip %in% hip_outliers),]
```

```
##    Density bodyfat Age Weight Height Neck Chest Abdomen   Hip Thigh Knee Ankle
## 35  1.0263    32.3  41 247.25  73.50 42.1 117.0   115.6 116.1  71.2 43.3  26.3
## 39  1.0202    35.2  46 363.15  72.25 51.2 136.2   148.1 147.7  87.3 49.1  29.6
## 41  1.0217    34.5  45 262.75  68.75 43.2 128.3   126.2 125.6  72.5 39.6  26.6
##    Biceps Forearm Wrist
## 35   37.3    31.7  19.7
## 39   45.0    29.0  21.4
## 41   36.4    32.7  21.4
```

Found 3 outliers for the Hip attribute. Further investigation finds that these 3 records have high bodyfat percentages (labelled as obese) which would explain the large/abnormal hip size.

Weight Outliers

```
boxplot(body_fat$Weight, main = "Weight", col = "blue")
```




```
boxplot.stats(body_fat$Weight)
```

```
## $stats
## [1] 118.50 158.75 176.50 197.00 247.25
##
## $n
## [1] 252
##
## $conf
## [1] 172.693 180.307
##
## $out
## [1] 363.15 262.75
```

```
weight_outliers<- boxplot(body_fat$Weight, plot = FALSE)$out
body_fat[which(body_fat$Weight %in% weight_outliers),]
```

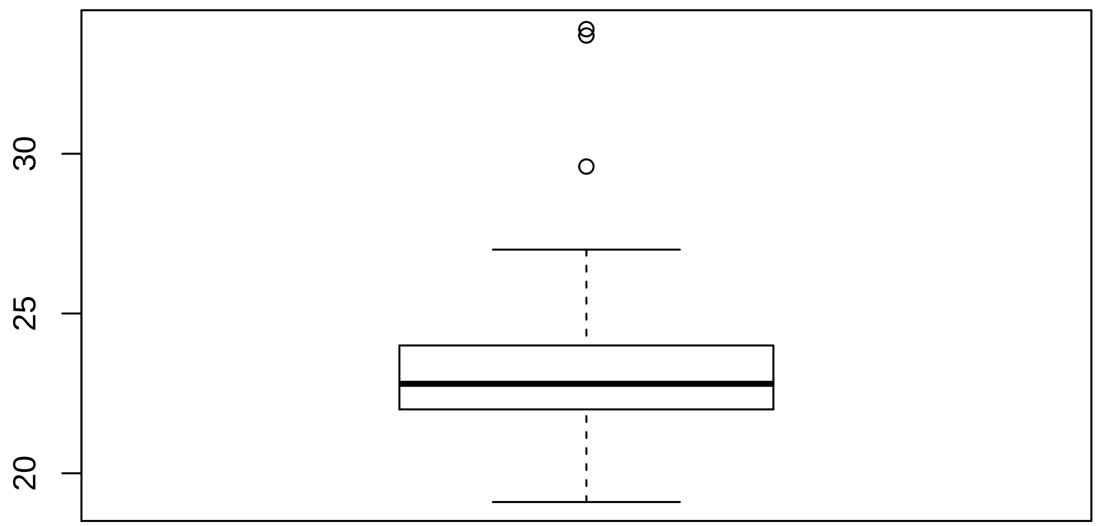
```
##    Density bodyfat Age Weight Height Neck Chest Abdomen   Hip Thigh Knee Ankle
## 39  1.0202    35.2  46 363.15  72.25 51.2 136.2   148.1 147.7  87.3 49.1  29.6
## 41  1.0217    34.5  45 262.75  68.75 43.2 128.3   126.2 125.6  72.5 39.6  26.6
##    Biceps Forearm Wrist
```

```
## 39   45.0    29.0  21.4
## 41   36.4    32.7  21.4
```

Two of these outliers are classified as outliers for the Hip attribute. I removed the 2 outlier attributes for Weight that are related to outliers for Hip and left the other Hip outlier (labelled 116.1) because it just falls outside the maximum value for Hip (115.50) and I don't want to remove too many Obese bodyfat records as it could impact the regression model.

Ankle Outliers

```
boxplot(body_fat$Ankle)
```




```
boxplot.stats(body_fat$Ankle)
```

```
## $stats
## [1] 19.1 22.0 22.8 24.0 27.0
##
## $n
## [1] 252
##
## $conf
## [1] 22.60094 22.99906
##
## $out
## [1] 33.9 29.6 33.7
```

```
ankle_outliers<- boxplot(body_fat$Ankle, plot = FALSE)$out
body_fat[which(body_fat$Ankle %in% ankle_outliers),]
```

```
##    Density bodyfat Age Weight Height Neck Chest Abdomen   Hip Thigh Knee Ankle
## 31  1.0716    11.9  32 182.00  73.75 38.7 100.5    88.7  99.8  57.5 38.7  33.9
## 39  1.0202    35.2  46 363.15  72.25 51.2 136.2   148.1 147.7  87.3 49.1  29.6
## 86  1.0386    26.6  67 167.00  67.50 36.5  98.9    89.7  96.2  54.7 37.8  33.7
##    Biceps Forearm Wrist
## 31   32.5    27.7  18.4
## 39   45.0    29.0  21.4
## 86   32.4    27.7  18.2
```

3 outliers where detected for the Ankle attribute. the first one has a Ankle measurement of 33.9 cm with a Bodyfat measurement of 11.9. Based on the American Council on Exercise Bodyfat table this would follow under Athletes.

```r
subset(body_fat, bodyfat >= 6 & bodyfat <= 14, select = c(bodyfat, Ankle))
```

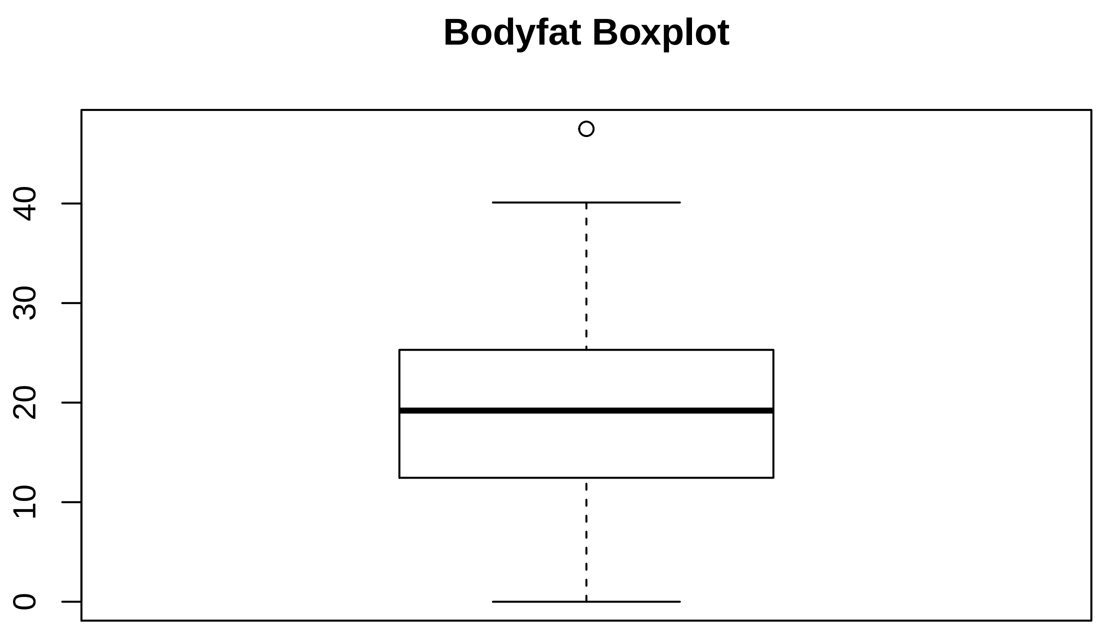
```
##     bodyfat Ankle
## 1      12.3  21.9
## 2       6.1  23.4
## 4      10.4  22.8
## 8      12.4  23.2
## 10     11.7  25.0
## 11      7.1  25.2
## 12      7.8  25.9
## 25     14.0  22.9
## 27      7.9  21.4
## 30      8.8  22.6
## 31     11.9  33.9
## 33     11.8  24.5
## 45      7.7  21.0
## 46     13.9  23.4
## 47     10.8  22.5
## 49     13.6  20.6
## 51     10.2  22.4
## 52      6.6  21.0
## 53      8.0  21.4
## 54      6.3  22.6
## 68     13.8  21.5
## 69      6.3  22.4
## 70     12.9  21.6
## 72      8.8  21.6
## 73      8.5  23.1
## 74     13.5  19.1
## 75     11.8  20.9
## 77      8.8  24.2
## 89      8.3  23.8
## 93      8.5  22.1
## 95      9.0  24.6
## 97      9.6  22.9
## 98     11.3  23.3
## 118    13.9  23.3
## 125    13.8  23.5
## 144     9.4  22.7
## 145    10.3  23.2
## 151     9.4  20.4
## 153    10.1  23.8
```

```
## 158    10.0  25.0
## 159    12.5  21.8
## 161     9.4  21.0
## 163    13.0  23.5
## 176     9.9  22.2
## 177    13.1  22.0
## 183    11.5  21.8
## 184    12.1  22.7
## 186     8.6  22.6
## 191    11.4  21.8
## 199     6.6  22.5
## 201    12.2  23.8
## 204     6.0  24.0
## 209     9.6  22.5
## 210    10.8  22.6
## 211     7.1  21.7
## 217    13.6  21.5
## 218     7.5  22.6
## 221    12.4  23.9
## 223    11.5  23.4
## 225    10.9  21.8
## 226    12.5  19.7
## 231    10.6  21.3
## 239    12.4  22.3
## 248    11.0  21.5
```

Investigating bodyfat in the Athletes range (6% -14%) and comparing it to Ankle measurement. This record of 33.9 cm in Ankle is an outlier for this category. The second outlier (29.6 cm) is an outlier for weight so removing it from weight will so remove it here. I have also decided to remove the 3rd outlier since it falls far outside the maximum value.

Bodyfat Outliers

```r
boxplot(body_fat$bodyfat, main = "Bodyfat Boxplot")
```




```r
boxplot.stats(body_fat$bodyfat)
```

```
## $stats
## [1]  0.00 12.45 19.20 25.30 40.10
##
## $n
## [1] 252
##
## $conf
## [1] 17.92103 20.47897
##
## $out
## [1] 47.5
```

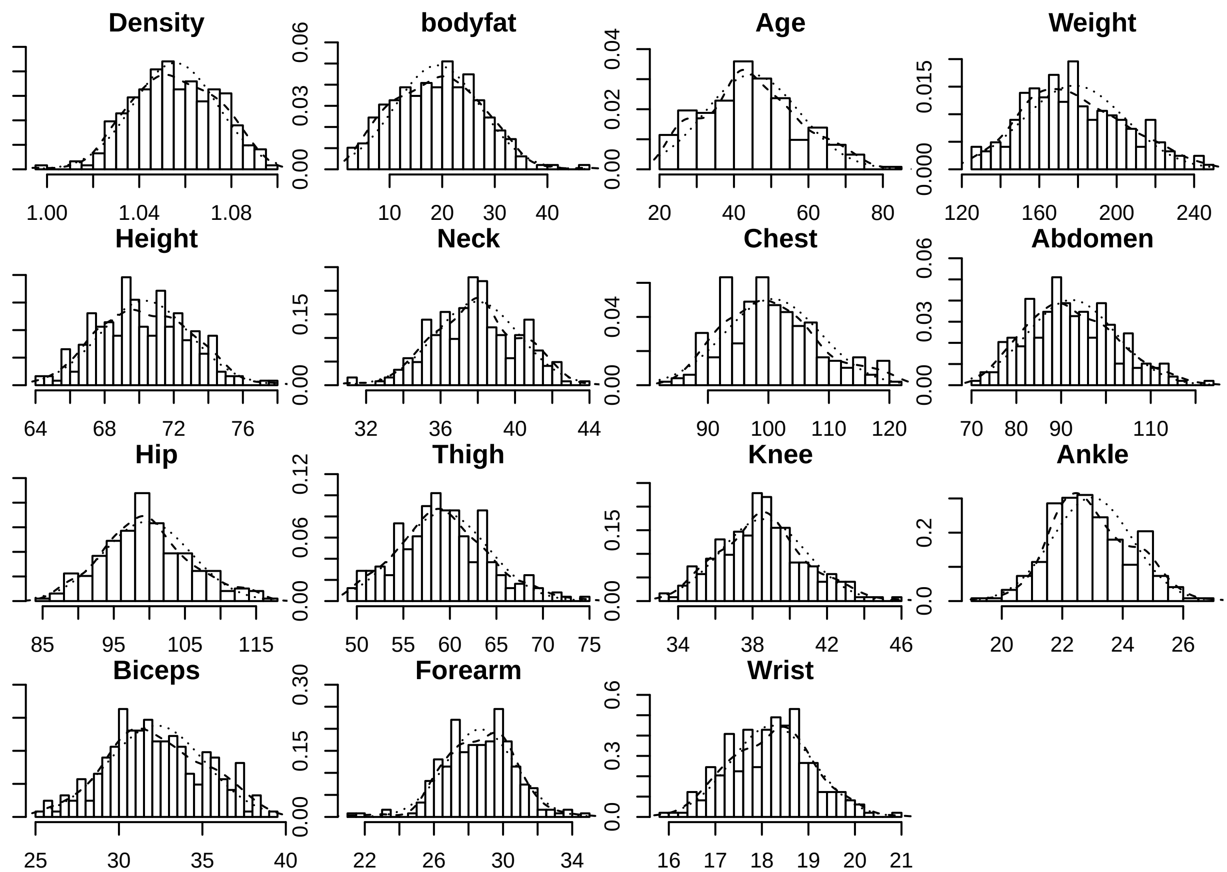
```r
subset(body_fat, body_fat$bodyfat < 2.00)
```

```
##     Density bodyfat Age Weight Height Neck Chest Abdomen  Hip Thigh Knee Ankle
## 172  1.0983     0.7  35 125.75   65.5 34.0  90.8    75.0 89.2  50.0 34.8  22.0
## 182  1.1089     0.0  40 118.50   68.0 33.8  79.3    69.4 85.0  47.2 33.5  20.2
##     Biceps Forearm Wrist
## 172   24.8    25.9  16.9
## 182   27.7    24.6  16.5
```

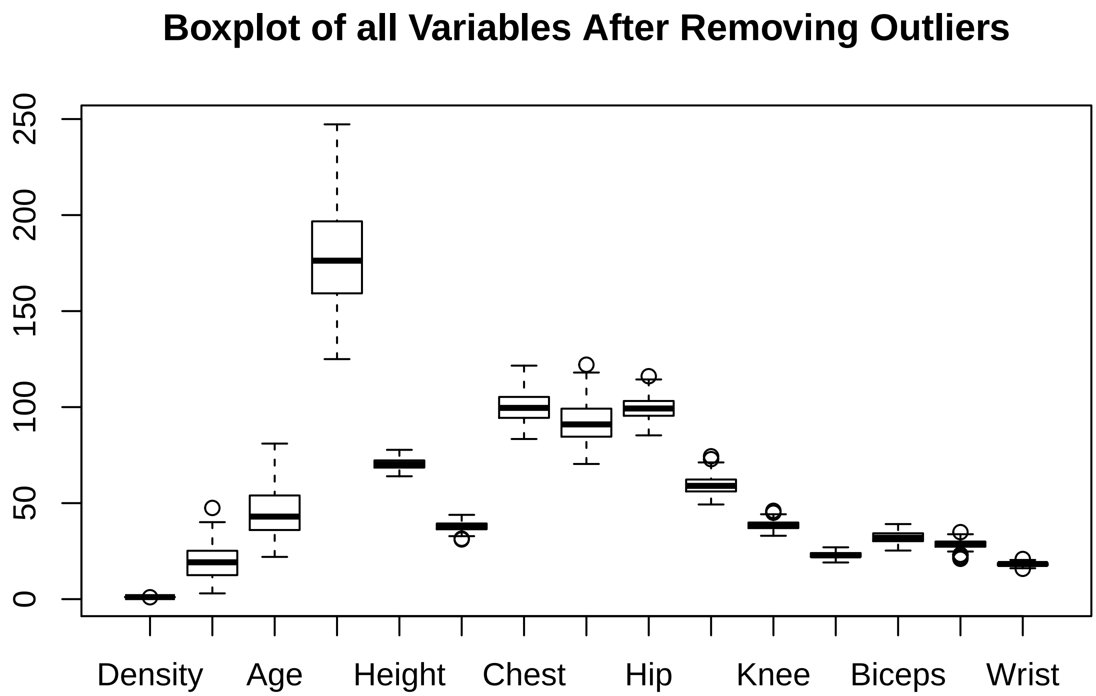
Creating a boxplot for bodyfat to determine the outliers only shows one outlier above the maximum value. However looking at the summary from earlier the minimum record for bodyfat is 0. Having 0 bodyfat is very extreme and possibility fatal and could be an error. According to the American Council on Exercise the minimum or essential level of fat in Men is 2-5%. So I will remove this record and other possible records that have less than 2% bodyfat.

Removing Outliers

```
bodyfat_2 <- body_fat[-c(42,39,41,31, 86, 172, 182),]
multi.hist(bodyfat_2)
```




```
boxplot(bodyfat_2, main = "Boxplot of all Variables After Removing Outliers")
```



I decided to remove the outliers that didn't make sense and could possibly be an error as I wanted to model the data better and predict future data better. Based on the Boxplot there are still some outliers. However I don't want to remove all the outliers because I don't want to remove what could be valuable information. Now the outliers are closer to the maximum values so there are not as extreme. The data is also now more normally distributed.

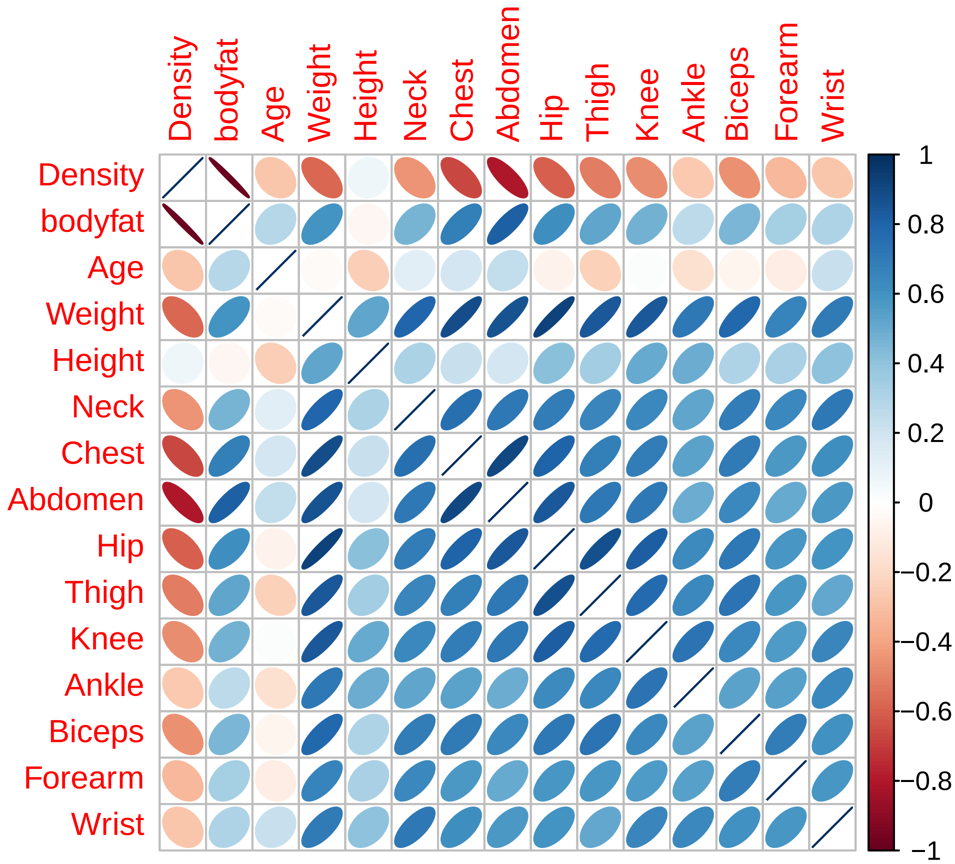
Correlation Between Attributes

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.5.3
```

```
## corrplot 0.84 loaded
```

```
cor_table<-cor(bodyfat_2)
corrplot(cor_table, method = 'ellipse')
```



I created a correlation matrix to determine the correlation between attributes. Most of the attributes in the dataset have moderate to high correlation with each other. The highest correlation exists between Density and Bodyfat which makes sense since Density is used to calculate Bodyfat. There is a strong negative relationship between them (-0.987), so decresing density increases bodyfat and vice versa. The weakest correlation exists between Age and Weight. ###Training and Test Sets

```
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 3.5.3
```

```
set.seed(15)
sample_df <- sample.split(bodyfat_2$bodyfat,SplitRatio = 0.70)
train_df <- subset(bodyfat_2, sample_df == TRUE)
test_df <- subset (bodyfat_2, sample_df == FALSE)
```

Since this dataset has a lot of variables that are highly correlated some of these variables can be removed when building the model. To determine which variables to eliminate, I used the findCorrelation in the Caret package to find highly correlated attributes. I used a cutoff of 0.60 to remove attributes with an absolute correlation of 0.60 or higher. I Chose 0.60 as the cutoff because generally a correlation of 0.60 or more represents a high correlation.

```
library(caret)
highly_correlated <- findCorrelation(cor_table,cutoff = 0.6)
highly_correlated
```

```
##  [1]  4  9  8  7 10 11  6 13 15  2
```

```
#or only use training set
findCorrelation(cor(train_df), cutoff = 0.6)
```

```
## [1]  4  9  8  7 10 11  6 13 15  2
```

The results show the indexes of the variables with the largest mean absolute correlation meaning they are highly correlated with the other attributes. These are Weight, Hip, Abdomen, Chest, Thigh, Knee, Neck, Bicpes and Bodyfat.

Modelling

Bodyfat Model With All Attributes

```
model1 <- lm(bodyfat~.,train_df)
summary(model1)
```

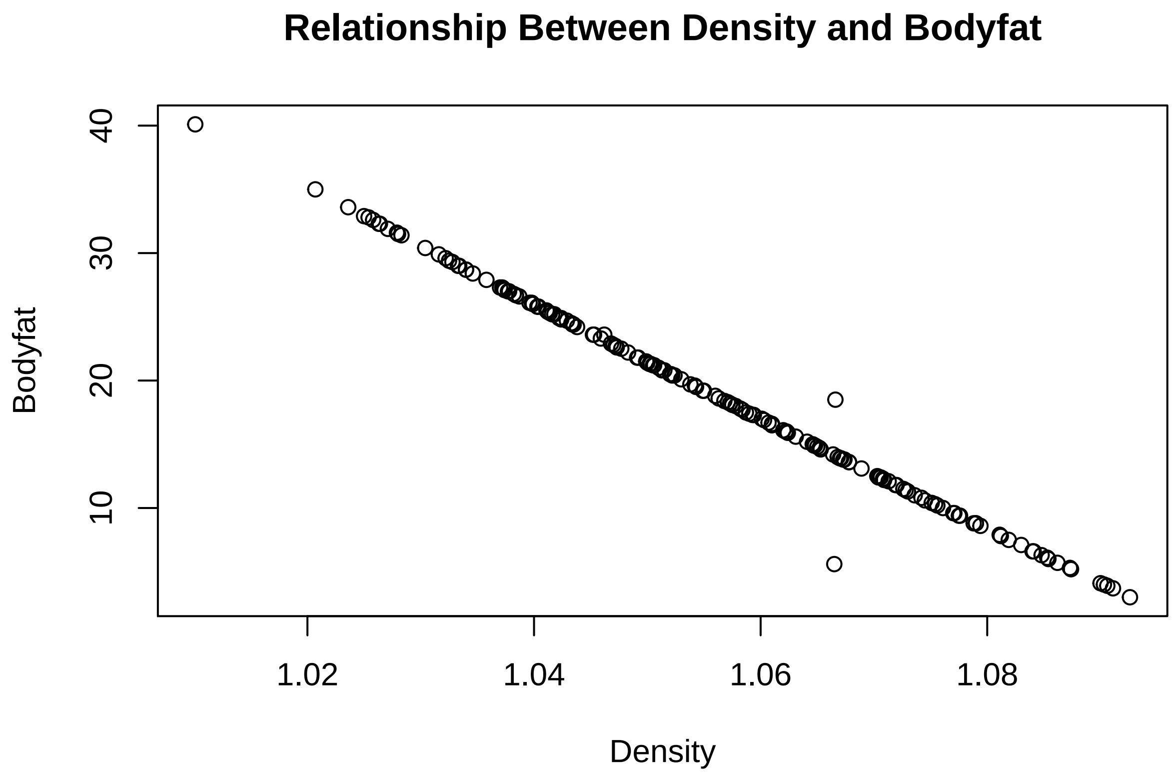
```
##
## Call:
## lm(formula = bodyfat ~ ., data = train_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2020 -0.1158 -0.0188  0.1571  4.2501
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.798e+02  8.602e+00  55.780   <2e-16 ***
## Density     -4.364e+02  6.320e+00 -69.047   <2e-16 ***
## Age          1.046e-02  7.501e-03   1.394    0.165
## Weight       2.029e-03  1.447e-02   0.140    0.889
## Height      -4.258e-02  4.234e-02  -1.006    0.316
## Neck         4.044e-02  5.159e-02   0.784    0.434
## Chest       -5.301e-03  2.360e-02  -0.225    0.823
## Abdomen     -3.903e-04  2.387e-02  -0.016    0.987
## Hip          3.136e-02  3.191e-02   0.983    0.327
## Thigh        9.793e-03  3.538e-02   0.277    0.782
## Knee        -5.945e-02  5.774e-02  -1.030    0.305
## Ankle        4.227e-02  8.237e-02   0.513    0.609
## Biceps      -1.745e-03  3.881e-02  -0.045    0.964
## Forearm      1.430e-02  4.571e-02   0.313    0.755
## Wrist       -9.473e-02  1.289e-01  -0.735    0.464
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7689 on 156 degrees of freedom
## Multiple R-squared:  0.9915, Adjusted R-squared:  0.9907
## F-statistic:  1299 on 14 and 156 DF,  p-value: < 2.2e-16
```

Using 0.05 as our significance level, Density is the only significant variable for predicting bodyfat.

Bodyfat Model Removing Only Density

Since Density is used to calculate bodyfat it makes sense that it is very significant in predicting bodyfat. The plot below shows that the relationship between Density and Bodyfat is almost perfectly linear with a correlation of -0.995.

```
plot(train_df$Density, train_df$bodyfat, main = "Relationship Between Density and Bodyfat", xlab = "Dens
```




```
cor(train_df$Density, train_df$bodyfat)
```

```
## [1] -0.9955198
```

The regression results with only Density and Bodyfat are also significant.

```
model2 <- lm(bodyfat~.,train_df[c(1,2)])
summary(model2)
```

```
##
## Call:
## lm(formula = bodyfat ~ ., data = train_df[c(1, 2)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5937 -0.0758 -0.0256  0.0788  4.3509
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  488.811      3.434   142.4   <2e-16 ***
## Density     -445.023      3.251  -136.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7574 on 169 degrees of freedom
## Multiple R-squared:  0.9911, Adjusted R-squared:  0.991
## F-statistic: 1.873e+04 on 1 and 169 DF,  p-value: < 2.2e-16
```

However because the methods to measure density are not easily accessible to the average person. I tried creating a regression model without it.

```
model3 <- lm(bodyfat~.,train_df[-c(1)])
summary(model3)
```

```
##
## Call:
## lm(formula = bodyfat ~ ., data = train_df[-c(1)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6951  -2.6643  -0.1379   2.8783   8.8044
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.28106   28.42361   0.010  0.99212
## Age          0.07387    0.04169   1.772  0.07835 .
## Weight      -0.01290    0.08100  -0.159  0.87365
## Height      -0.28147    0.23631  -1.191  0.23541
## Neck         0.11596    0.28885   0.401  0.68863
## Chest       -0.19266    0.13126  -1.468  0.14416
## Abdomen      0.89453    0.11224   7.970 3.06e-13 ***
## Hip         -0.05840    0.17852  -0.327  0.74399
## Thigh        0.03377    0.19813   0.170  0.86486
## Knee        -0.18097    0.32317  -0.560  0.57629
## Ankle       -0.02458    0.46123  -0.053  0.95757
## Biceps       0.27766    0.21612   1.285  0.20078
## Forearm      0.33425    0.25467   1.312  0.19128
## Wrist       -2.03779    0.70454  -2.892  0.00437 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.306 on 157 degrees of freedom
## Multiple R-squared:  0.7316, Adjusted R-squared:  0.7093
## F-statistic: 32.91 on 13 and 157 DF,  p-value: < 2.2e-16
```

Without Density the Adjusted R-Squared does reduce but is still a high value at 70.93%. Abdomen and Wrist are now significant variables at a significance level of 0.05 in predicting Bodyfat.

```
#install.packages("car")
#model with just Age and Abdomen
model4 <-lm(bodyfat~.,train_df[c(2,8,15)])
summary(model4)
```

```
##
## Call:
## lm(formula = bodyfat ~ ., data = train_df[c(2, 8, 15)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1864  -3.2902  -0.3682   3.0520  11.1757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14.95305    7.39585  -2.022   0.0448 *
```

```
## Abdomen       0.79653    0.04388  18.151  < 2e-16 ***
## Wrist        -2.16476    0.48625  -4.452 1.55e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.459 on 168 degrees of freedom
## Multiple R-squared:  0.692,  Adjusted R-squared:  0.6884
## F-statistic: 188.8 on 2 and 168 DF,  p-value: < 2.2e-16
```

```
library(car)
vif(model4)
```

```
##  Abdomen    Wrist
## 1.460069 1.460069
```

The values for variance inflation factor (VIF) for Adbomen and Wrist are both below 5. Therefore collinearity is not a problem between these 2 variables and both can be left in the model.

The performance of the model between only Abdomen and Wrist is low compared to the model with only Density. In model2, with Density removed Age also showed up as a significant variable but at an alpha of 10%. Building a model with Age, Adbomen and Wrist is shown below.

```
model5<-lm(bodyfat~.,train_df[c(2,3,8,15)])
summary(model5)
```

```
##
## Call:
## lm(formula = bodyfat ~ ., data = train_df[c(2, 3, 8, 15)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4060  -2.9484  -0.4819   2.9145  10.4532
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.96122    7.21277  -1.797  0.07415 .
## Age           0.09129    0.02767   3.299  0.00119 **
## Abdomen       0.77078    0.04336  17.778  < 2e-16 ***
## Wrist        -2.36906    0.47658  -4.971 1.64e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.333 on 167 degrees of freedom
## Multiple R-squared:  0.7109, Adjusted R-squared:  0.7057
## F-statistic: 136.9 on 3 and 167 DF,  p-value: < 2.2e-16
```

```
vif(model5)
```

```
##      Age  Abdomen    Wrist
## 1.113323 1.508995 1.485144
```

Alone, all 3 variables show up as significant with an alpha of 5%. The Adjusted R^2 also increased compared to the model with only Abdomen and wrist. Looking at VIF the values are also below 5 so there is no problem with multicollinearity. Thus this will be the final model for predicting bodyfat.

Models with Weight

```r
modelA<- lm(Weight~., train_df)
summary(modelA)
```

```
##
## Call:
## lm(formula = Weight ~ ., data = train_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.7795  -2.2659   0.0455   2.8031   9.4868
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -358.31164  215.97681  -1.659 0.099119 .
## Density       33.64710  196.45673   0.171 0.864234
## bodyfat        0.06215    0.44305   0.140 0.888627
## Age           -0.09207    0.04111  -2.240 0.026537 *
## Height         1.96075    0.17496  11.207  < 2e-16 ***
## Neck           0.84824    0.27790   3.052 0.002670 **
## Chest          0.83481    0.11220   7.440 6.31e-12 ***
## Abdomen        0.57513    0.12380   4.645 7.17e-06 ***
## Hip            0.79841    0.16517   4.834 3.18e-06 ***
## Thigh          0.38257    0.19344   1.978 0.049724 *
## Knee           0.37925    0.31915   1.188 0.236528
## Ankle          1.49576    0.44022   3.398 0.000862 ***
## Biceps         0.56728    0.20989   2.703 0.007640 **
## Forearm        0.36619    0.25135   1.457 0.147158
## Wrist          0.91780    0.71090   1.291 0.198601
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.255 on 156 degrees of freedom
## Multiple R-squared:  0.9727, Adjusted R-squared:  0.9703
## F-statistic: 397.6 on 14 and 156 DF,  p-value: < 2.2e-16
```

Using all attributes to predict weight also provides significant results. The variables: Age, Height, Neck, Chest, Abdomen, Hip, Thigh, Ankle, Biceps are all significant variables at the 0.05 significance level.

```r
modelB<-lm(Weight~., train_df[c(3,4,5,6,7,8,9,10,12,13)])
summary(modelB)
```

```
##
## Call:
## lm(formula = Weight ~ ., data = train_df[c(3, 4, 5, 6, 7, 8,
##     9, 10, 12, 13)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2618  -2.1574   0.0315   2.6038  10.3653
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -320.41378   11.16471 -28.699  < 2e-16 ***
```

```
## Age           -0.06596    0.03736  -1.766 0.079355 .  
## Height         2.06138    0.16581  12.432  < 2e-16 ***
## Neck           1.04352    0.25745   4.053 7.85e-05 ***
## Chest          0.84531    0.11112   7.607 2.21e-12 ***
## Abdomen        0.55360    0.10130   5.465 1.73e-07 ***
## Hip            0.87527    0.16073   5.445 1.90e-07 ***
## Thigh          0.37886    0.18331   2.067 0.040351 *  
## Ankle          2.07718    0.36741   5.654 6.99e-08 ***
## Biceps         0.72525    0.19112   3.795 0.000209 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.273 on 161 degrees of freedom
## Multiple R-squared:  0.9716, Adjusted R-squared:   0.97 
## F-statistic: 612.5 on 9 and 161 DF,  p-value: < 2.2e-16
```

Age is no longer significant at 0.05 so removed.

```
modelC<-lm(Weight~., train_df[c(4,5,6,7,8,9,10,12,13)])
summary(modelC)
```

```
## 
## Call:
## lm(formula = Weight ~ ., data = train_df[c(4, 5, 6, 7, 8, 9, 
##     10, 12, 13)])
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -21.4686  -2.2897  -0.1404   2.6625  10.2870 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -327.89186   10.39746 -31.536  < 2e-16 ***
## Height         2.11403    0.16417  12.877  < 2e-16 ***
## Neck           0.96111    0.25484   3.771 0.000227 ***
## Chest          0.84809    0.11183   7.584 2.47e-12 ***
## Abdomen        0.47667    0.09204   5.179 6.55e-07 ***
## Hip            0.91799    0.15994   5.740 4.55e-08 ***
## Thigh          0.50250    0.17051   2.947 0.003682 ** 
## Ankle          2.04185    0.36926   5.530 1.26e-07 ***
## Biceps         0.72436    0.19236   3.766 0.000232 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.301 on 162 degrees of freedom
## Multiple R-squared:  0.9711, Adjusted R-squared:  0.9696 
## F-statistic: 679.8 on 8 and 162 DF,  p-value: < 2.2e-16
```

Since we know that may of the variables are highly correlated performed a test to help with multilinearity.

```
vif(modelC)
```

```
##   Height     Neck    Chest  Abdomen      Hip    Thigh    Ankle   Biceps 
## 1.559589 2.505664 6.667641 6.901654 8.377654 5.226100 1.995582 2.646672
```

Chest, Abdomen, Hip and Thigh all exceed VIF values of 5, a sign a problematic amount of collinearity (James et al. 2014). These variables will be removed one at a time starting with Hip, the variable with the largest amount of VIF to test how the VIF and the model changes as they are removed.

```
#without Hip
vif(lm(Weight~., train_df[c(4,5,6,7,8,10,12,13)]))
```

```
##   Height     Neck    Chest  Abdomen    Thigh    Ankle   Biceps
## 1.383371 2.426788 6.528247 5.543541 3.145463 1.986598 2.646382
```

```
#without Chest and Hip
vif(lm(Weight~., train_df[c(4,5,6,8,10,12,13)]))
```

```
##   Height     Neck  Abdomen    Thigh    Ankle   Biceps
## 1.380285 2.221516 2.425849 3.095134 1.964767 2.397131
```

```
summary(lm(Weight~., train_df[c(4,5,6,8,10,12,13)]))
```

```
##
## Call:
## lm(formula = Weight ~ ., data = train_df[c(4, 5, 6, 8, 10, 12,
##     13)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.1282  -2.9437   0.4301   3.5843  14.4815
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -307.66865   13.01809 -23.634  < 2e-16 ***
## Height         2.36868    0.19716  12.014  < 2e-16 ***
## Neck           1.32181    0.30631   4.315 2.75e-05 ***
## Abdomen        1.23703    0.06966  17.759  < 2e-16 ***
## Thigh          0.97773    0.16751   5.837 2.77e-08 ***
## Ankle          2.51245    0.46771   5.372 2.63e-07 ***
## Biceps         1.21474    0.23369   5.198 5.93e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.491 on 164 degrees of freedom
## Multiple R-squared:  0.9523, Adjusted R-squared:  0.9505
## F-statistic: 545.5 on 6 and 164 DF,  p-value: < 2.2e-16
```

```
#without Hip, Abdomen and Chest
vif(lm(Weight~., train_df[c(4,5,6,10,12,13)]))
```

```
##   Height     Neck    Thigh    Ankle   Biceps
## 1.323061 1.801864 2.735195 1.960123 2.378579
```

```
summary(lm(Weight~., train_df[c(4,5,6,10,12,13)]))
```

```
##
## Call:
## lm(formula = Weight ~ ., data = train_df[c(4, 5, 6, 10, 12, 13)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.250  -5.960  -0.851   5.676  37.377
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -314.3958    22.1796 -14.175  < 2e-16 ***
## Height         1.6558     0.3290   5.033 1.25e-06 ***
## Neck           3.6860     0.4702   7.839 5.32e-13 ***
## Thigh          1.9921     0.2684   7.422 5.80e-12 ***
## Ankle          2.9163     0.7963   3.662 0.000336 ***
## Biceps         1.5798     0.3968   3.982 0.000102 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.359 on 165 degrees of freedom
## Multiple R-squared:  0.8605, Adjusted R-squared:  0.8563
## F-statistic: 203.6 on 5 and 165 DF,  p-value: < 2.2e-16
```

```r
#without Hip, Abdomen, Chest and Thigh
vif(lm(Weight~., train_df[c(4,5,6,12,13)]))
```

```
##   Height     Neck    Ankle   Biceps
## 1.317963 1.715229 1.712109 1.860545
```

```r
summary(lm(Weight~., train_df[c(4,5,6,12,13)]))
```

```
##
## Call:
## lm(formula = Weight ~ ., data = train_df[c(4, 5, 6, 12, 13)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.888  -6.184  -1.313   6.132  36.501
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -328.6005    25.4436 -12.915  < 2e-16 ***
## Height         1.8074     0.3781   4.780 3.84e-06 ***
## Neck           4.4513     0.5282   8.427 1.63e-14 ***
## Ankle          5.0186     0.8569   5.857 2.47e-08 ***
## Biceps         2.9542     0.4041   7.311 1.07e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.78 on 166 degrees of freedom
## Multiple R-squared:  0.8139, Adjusted R-squared:  0.8095
## F-statistic: 181.6 on 4 and 166 DF,  p-value: < 2.2e-16
```

The model without Chest and Hip removes multicollinearity and produces similar results as the model with these two variables included (modelC). Which is a good sign because it leads to a simpler model without compromising the model accuracy.

```r
modelD<-lm(Weight~., train_df[c(4,5,6,8,10,12,13)])
summary(modelD)
```

```
##
## Call:
## lm(formula = Weight ~ ., data = train_df[c(4, 5, 6, 8, 10, 12,
##     13)])
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -25.1282  -2.9437   0.4301   3.5843  14.4815
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -307.66865   13.01809 -23.634  < 2e-16 ***
## Height         2.36868    0.19716  12.014  < 2e-16 ***
## Neck           1.32181    0.30631   4.315 2.75e-05 ***
## Abdomen        1.23703    0.06966  17.759  < 2e-16 ***
## Thigh          0.97773    0.16751   5.837 2.77e-08 ***
## Ankle          2.51245    0.46771   5.372 2.63e-07 ***
## Biceps         1.21474    0.23369   5.198 5.93e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.491 on 164 degrees of freedom
## Multiple R-squared:  0.9523, Adjusted R-squared:  0.9505
## F-statistic: 545.5 on 6 and 164 DF,  p-value: < 2.2e-16
```

```
vif(modelD)
```

```
##   Height     Neck  Abdomen    Thigh    Ankle   Biceps
## 1.380285 2.221516 2.425849 3.095134 1.964767 2.397131
```

Prediction

Predicting Bodyfat using the Test Set

```
bodyfat_prediction<-predict(model5, test_df)
#creating new column for predicted values
test_df$predicted_bodyfat <-bodyfat_prediction
head(test_df[c(2,16)], n=10)
```

```
##    bodyfat predicted_bodyfat
## 3     25.3          17.47216
## 6     20.9          17.45286
## 7     19.2          17.38952
## 10    11.7          11.94343
## 15    22.1          21.42001
## 16    20.9          21.72499
## 18    22.9          19.38804
## 20    16.5          24.39763
## 21    19.1          19.92180
## 22    15.2          18.60347
```

Predicting Weight Using the Test Set

```
weight_prediction <-predict(modelD, test_df)
#creating new column for predicted values
test_df$predicted_weight <-weight_prediction
head(test_df[c(4,17)], n=10)
```

```
##    Weight predicted_weight
## 3  154.00         156.4893
## 6  210.25         209.9320
## 7  181.00         171.2445
## 10 198.25         199.4296
```

```
## 15 187.75         199.0912
## 16 162.75         165.5683
## 18 209.25         209.8837
## 20 211.75         213.3240
## 21 179.00         180.8061
## 22 200.50         198.2373
```

In order to determine how effective my models are in predicting, I need to calculate and compared the RMSE.

```
error_bodyfat<- bodyfat_prediction -test_df$bodyfat
#RMSE
sqrt(mean(error_bodyfat^2))
```

```
## [1] 4.239244
```

```
error_weight <- weight_prediction - test_df$Weight
#RMSE
sqrt(mean(error_weight^2))
```

```
## [1] 5.614205
```

In order to compare the RMSE for the Test and Training Sets need to calculate RMSE for the training sets based on the residuals from the models.

```
mse_bodyfat <- mean(residuals(model5)^2)
mse_bodyfat
```

```
## [1] 18.33481
```

```
rmse_bodyfat <- sqrt(mse_bodyfat)
rmse_bodyfat
```

```
## [1] 4.281917
```

```
mse_weight <- mean(residuals(modelD)^2)
mse_weight
```

```
## [1] 28.91291
```

```
rmse_weight <- sqrt(mse_weight)
rmse_weight
```

```
## [1] 5.377073
```

The RMSE in the test sets is slightly higher than the training sets. This slight difference between the two is an indicator of a good model. There is no overfitting or underfitting.