Final Project

Michael Morales

December 2, 2018

library(arules)

## Loading required package: Matrix

##   
## Attaching package: 'arules'

## The following objects are masked from 'package:base':  
##   
## abbreviate, write

library(ggplot2)  
library(arulesViz)

## Loading required package: grid

library(kernlab)

##   
## Attaching package: 'kernlab'

## The following object is masked from 'package:ggplot2':  
##   
## alpha

## The following object is masked from 'package:arules':  
##   
## size

library(e1071)  
library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':  
##   
## margin

Reading in the csv file

plData <- read.csv(file="C:/PLTrimmed2.csv", header = TRUE)

Subsetting, correlations, plots

wilks.bw <- subset(plData, select = c("Wilks", "BodyweightKg")) #subset of Wilks and BodyweightKg columns  
summary(wilks.bw)

## Wilks BodyweightKg   
## Min. : 28.23 Min. : 23.90   
## 1st Qu.:295.14 1st Qu.: 68.60   
## Median :343.60 Median : 82.20   
## Mean :346.92 Mean : 85.54   
## 3rd Qu.:393.51 3rd Qu.: 99.20   
## Max. :779.38 Max. :242.40

wilks.yr <- subset(plData, select = c("Wilks", "Year")) #subset of Wilks and Year columns  
summary(wilks.yr)

## Wilks Year   
## Min. : 28.23 Min. :1989   
## 1st Qu.:295.14 1st Qu.:2014   
## Median :343.60 Median :2016   
## Mean :346.92 Mean :2014   
## 3rd Qu.:393.51 3rd Qu.:2017   
## Max. :779.38 Max. :2018

bench.squat <- subset(plData, select = c("BestBenchKg", "BestSquatKg")) #subset of bench and squat weight columns  
summary(bench.squat)

## BestBenchKg BestSquatKg   
## Min. : 6.8 Min. :-175.0   
## 1st Qu.: 75.0 1st Qu.: 127.5   
## Median :111.1 Median : 172.5   
## Mean :114.2 Mean : 176.9   
## 3rd Qu.:145.0 3rd Qu.: 217.5   
## Max. :455.9 Max. : 573.8

bench.dead <- subset(plData, select = c("BestBenchKg", "BestDeadliftKg")) #subset of bench and deadlift weight columns  
summary(bench.dead)

## BestBenchKg BestDeadliftKg   
## Min. : 6.8 Min. : 2.27   
## 1st Qu.: 75.0 1st Qu.:149.69   
## Median :111.1 Median :195.00   
## Mean :114.2 Mean :194.93   
## 3rd Qu.:145.0 3rd Qu.:237.50   
## Max. :455.9 Max. :440.00

#finding correlations between the values in the subset columns  
cor(wilks.bw)

## Wilks BodyweightKg  
## Wilks 1.0000000 0.1843692  
## BodyweightKg 0.1843692 1.0000000

cor(wilks.yr)

## Wilks Year  
## Wilks 1.0000000 -0.1028035  
## Year -0.1028035 1.0000000

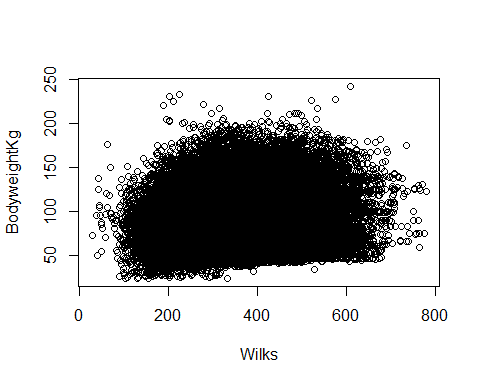
cor(bench.squat)

## BestBenchKg BestSquatKg  
## BestBenchKg 1.0000000 0.8944746  
## BestSquatKg 0.8944746 1.0000000

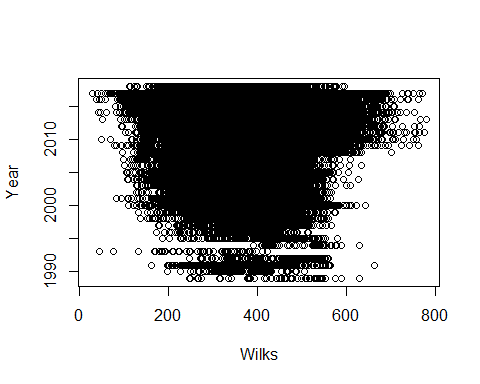
cor(bench.dead)

## BestBenchKg BestDeadliftKg  
## BestBenchKg 1.0000000 0.8604488  
## BestDeadliftKg 0.8604488 1.0000000

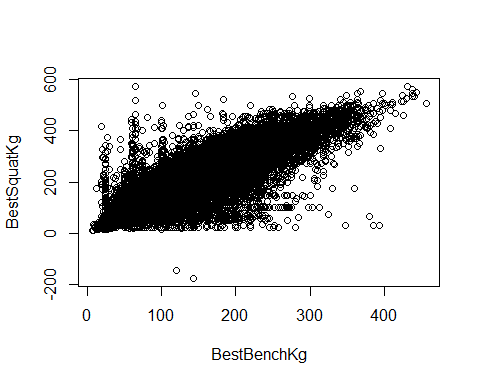
#plotting the subsets  
plot(wilks.bw)



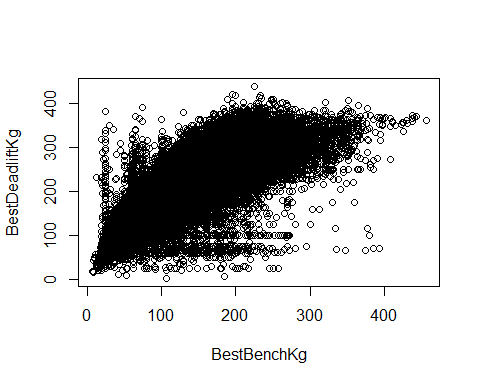
plot(wilks.yr)



plot(bench.squat)



plot(bench.dead)



Fitting regression models

#Fitting regressions by the same variables as the correlation subsets  
wilksBW.mod <- lm(Wilks ~ BodyweightKg, data = plData)  
summary(wilksBW.mod)

##   
## Call:  
## lm(formula = Wilks ~ BodyweightKg, data = plData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -341.43 -50.22 -2.65 45.82 435.85   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.934e+02 5.574e-01 526.31 <2e-16 \*\*\*  
## BodyweightKg 6.259e-01 6.301e-03 99.33 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 75.36 on 280403 degrees of freedom  
## Multiple R-squared: 0.03399, Adjusted R-squared: 0.03399   
## F-statistic: 9867 on 1 and 280403 DF, p-value: < 2.2e-16

wilksYr.mod <- lm(Wilks ~ Year, data = plData)  
summary(wilksYr.mod)

##   
## Call:  
## lm(formula = Wilks ~ Year, data = plData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -343.48 -51.19 -3.13 46.48 429.66   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4203.80748 70.47450 59.65 <2e-16 \*\*\*  
## Year -1.91460 0.03498 -54.73 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 76.27 on 280403 degrees of freedom  
## Multiple R-squared: 0.01057, Adjusted R-squared: 0.01057   
## F-statistic: 2995 on 1 and 280403 DF, p-value: < 2.2e-16

benchSq.mod <- lm(BestBenchKg ~ BestSquatKg, data = plData)  
summary(benchSq.mod)

##   
## Call:  
## lm(formula = BestBenchKg ~ BestSquatKg, data = plData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -306.58 -12.78 -2.03 10.59 374.55   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.4752988 0.1163832 -12.68 <2e-16 \*\*\*  
## BestSquatKg 0.6540766 0.0006174 1059.34 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 21.27 on 280403 degrees of freedom  
## Multiple R-squared: 0.8001, Adjusted R-squared: 0.8001   
## F-statistic: 1.122e+06 on 1 and 280403 DF, p-value: < 2.2e-16

benchDL.mod <- lm(BestBenchKg ~ BestDeadliftKg, data = plData)  
summary(benchDL.mod)

##   
## Call:  
## lm(formula = BestBenchKg ~ BestDeadliftKg, data = plData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -221.73 -14.40 -2.77 10.48 366.16   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -23.46127 0.16065 -146.0 <2e-16 \*\*\*  
## BestDeadliftKg 0.70639 0.00079 894.2 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 24.24 on 280403 degrees of freedom  
## Multiple R-squared: 0.7404, Adjusted R-squared: 0.7404   
## F-statistic: 7.996e+05 on 1 and 280403 DF, p-value: < 2.2e-16

Creating training and testing datasets

randIndex <- sample (1:dim(plData)[1]) #making a randomized index to avoid systematic bias  
cutPoint2\_3 <- floor(2\*dim(plData)[1]/3) #calculating 2/3 cut point based on number of rows  
trainPLdata <- plData[randIndex[1:cutPoint2\_3],] #randomized training data of 2/3 of the dataset  
testPLdata <- plData[randIndex[(cutPoint2\_3+1):dim(plData)[1]],] #randomized test data of /3 of the set