R PRACTICE

Week 3

Abstract

In this assignment I will use T test on powerlifting dataset

Mohammad Hossein Movahedi

Movahedi.m@northeastern.edu

First Of all, I import Libraries and the dataset I found the data set on Kaggle. This dataset is a snapshot of the OpenPowerlifting database as of April 2019. OpenPowerlifting is creating a public-domain archive of powerlifting history. (OpenPowerlifting, 2019)

```
 print('Mohammad Hossein Movahedi')

print('Module 3 R practice')

#importing and instaling libraries

install.packages('FSA')

install.packages('magrittr')

install.packages('dplyr')

install.packages('tidyr')

install.packages('plyr')

install.packages('tidyverse')

install.packages('outliers')

install.packages('ggplot2')

install.packages('lubridate')



library(FSA)

library(magrittr)

library(dplyr)

library(tidyr)

library(plyr)

library(tidyverse)

library(scales)

library(lubridate)

library(ggplot2)

library(outliers)



#Importing dataset

data <- read.csv('openpowerlifting.csv')
```

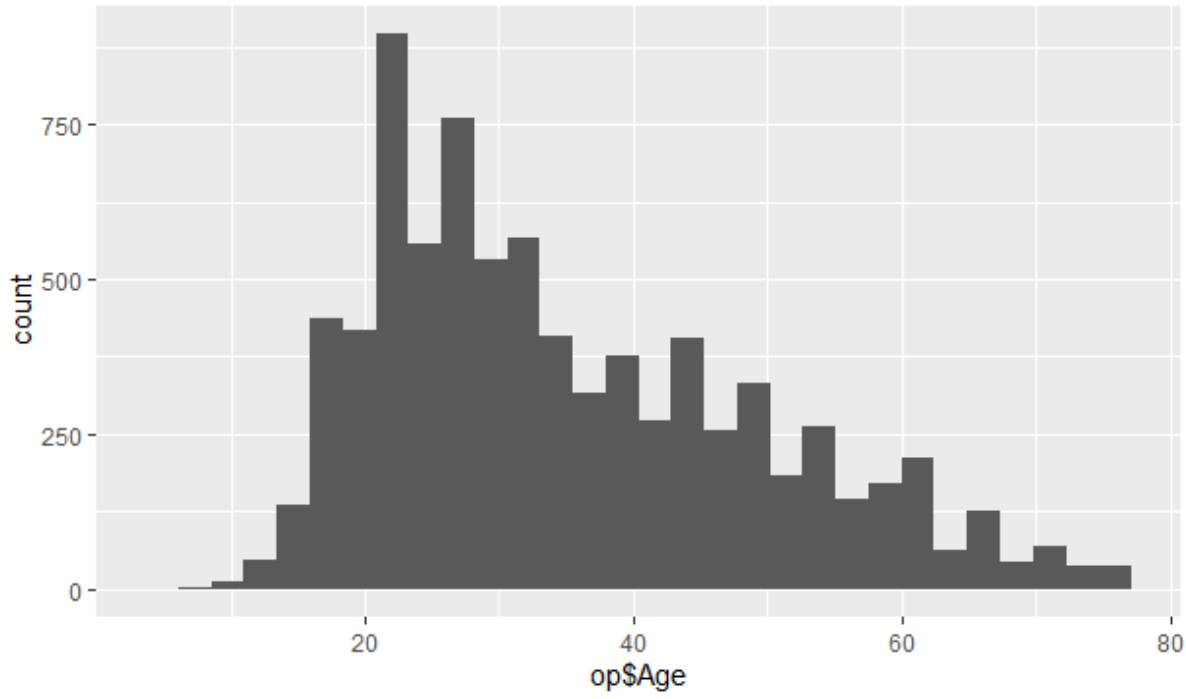
Then because the dataset is extensive, I make a subset of the first 10000 rows

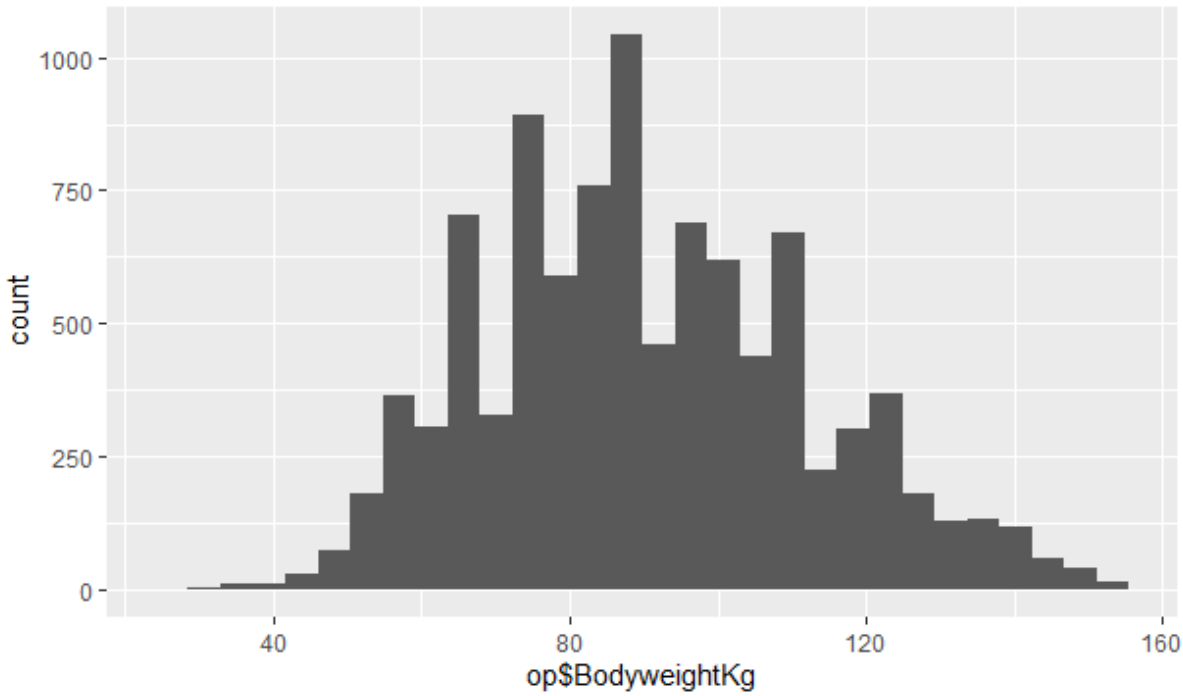
```
#because the dataset is too large, I choose the first 10000 data to work with

data <-head(data,n = 10000)
```

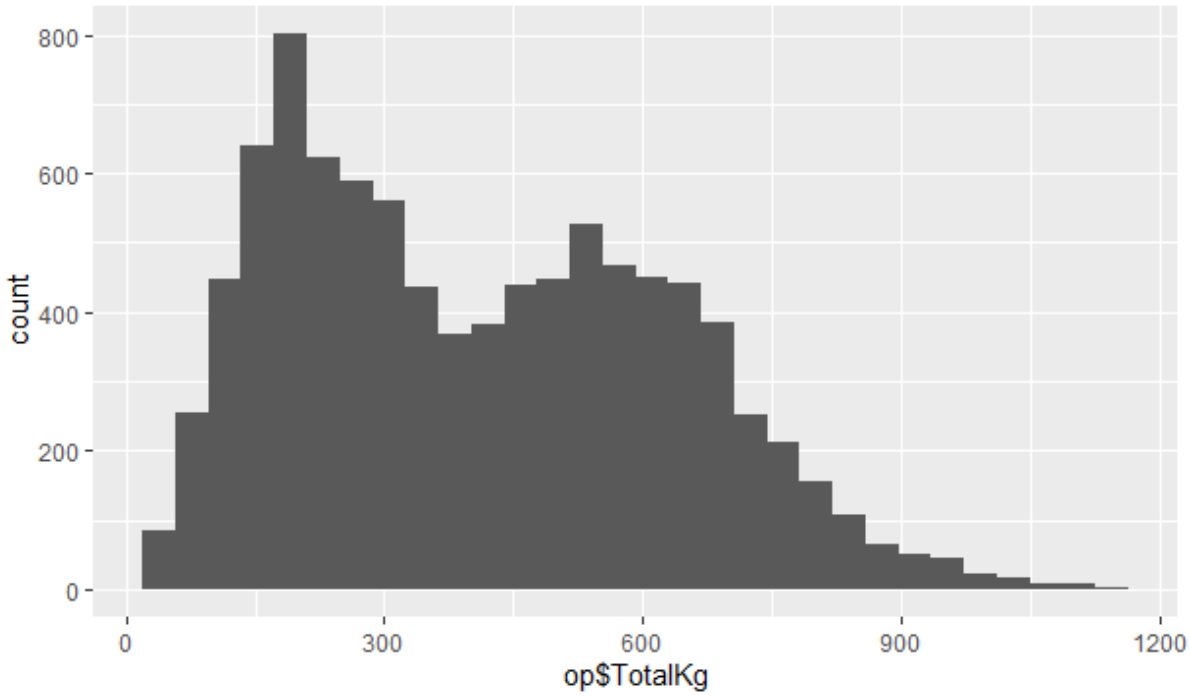
Then I clear duplicates and check the data to see If it looks semi normal or not

```
# first of all I delete duplicate rows

op <- data[!duplicated(data), ]

#first we cheak to see if our data looks normal or not

ggplot(op, aes(x=op$Age)) + geom_histogram()

ggplot(op, aes(x=op$BodyweightKg)) + geom_histogram()

ggplot(op, aes(x=op$TotalKg)) + geom_histogram()
```

The resulting graphs







As it can be seen, all of them look normal.

Then I deleted outliers

```
#then we delete outliers in age

boxplot(op$Age)$out

d <- boxplot(op$Age, plot=FALSE)$out

op<- op[-which(op$Age %in% d),]

#then we delete outliers in Bodyweightkg

boxplot(op$BodyweightKg)$out

d <- boxplot(op$BodyweightKg, plot=FALSE)$out

op<- op[-which(op$BodyweightKg %in% d),]

#then we delete outliers in Totol

boxplot(op$TotalKg)$out

d <- boxplot(op$TotalKg, plot=FALSE)$out

op<- op[-which(op$TotalKg %in% d),]
```

Then I subset the data by gender

```
#then we subset the data by sex

mp <- filter(op,op$Sex == "M")

fp <- filter(op,op$Sex == "F")
```

Now we test to see if there is a difference between men's Bodyweight and the average and then we test the difference between the Age of men and average and then we ran the same two tests for women.

```
 # now we test to see if there is diffrence between men Bodyweight and the averageaveW <-
mean(op$BodyweightKg,na.rm = T)

aveA <- mean(op$Age,na.rm = T)

t.test(mp$BodyweightKg, mu = aveW)

t.test(mp$Age, mu = aveA)

aveW <- mean(op$BodyweightKg,na.rm = T)

aveA <- mean(op$Age,na.rm = T)

t.test(fp$BodyweightKg, mu = aveW)

t.test(fp$Age, mu = aveA)
```

```
 > print('males')

[1] "males"

> aveW <- mean(op$BodyweightKg,na.rm = T)

> aveA <- mean(op$Age,na.rm = T)

> t.test(mp$BodyweightKg, mu = aveW)


          One Sample t-test


data:  mp$BodyweightKg

t = 25.792, df = 7344, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 90.26215

95 percent confidence interval:

 95.79461 96.70476

sample estimates:

mean of x

 96.24968


> t.test(mp$Age, mu = aveA)


          One Sample t-test


data:  mp$Age

t = 1.5372, df = 6061, p-value = 0.1243

alternative hypothesis: true mean is not equal to 35.13817

95 percent confidence interval:

 35.05928 35.79027

sample estimates:

mean of x

 35.42478
```

```
 [1] "females"
> aveW <- mean(op$BodyweightKg,na.rm = T)
> aveA <- mean(op$Age,na.rm = T)
> t.test(fp$BodyweightKg, mu = aveW)

          One Sample t-test

data:  fp$BodyweightKg
t = -53.566, df = 2385, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 90.26215
95 percent confidence interval:
 71.18512 72.53256
sample estimates:
mean of x
 71.85884

> t.test(fp$Age, mu = aveA)

          One Sample t-test

data:  fp$Age
t = -2.9429, df = 2016, p-value = 0.003288
alternative hypothesis: true mean is not equal to 35.13817
95 percent confidence interval:
 33.69180 34.84861
sample estimates:
mean of x
  34.2702
```

The results are not very shocking; there is a significant difference between females' average age, body weight, and population. Also, there was a considerable difference between males' body weight and total. However, there is no significant difference between males' age and total age (p-value = 0.12), which is shocking considering that women's average age was 34.2. In contrast, men's average age was 35.4, showing 95% of professional powerlifters are in their mid-30s. This data can indicate that considering the 95 percent confidence interval for men's age is (35.05928, 35.79027), it is almost a year more than women's (33.69180 34.84861). Maybe it is the result of men having more access to powerlifting pieces of equipment, and we can close this age gap by making more powerlifting pieces of equipment designed specifically for women.

Considering the format of the dataset, I designed a paired test between the Best Bench result and the Best Deadlift result to see if there is a relationship between them

```
# now we test if there is a relation betweenBestBenchKg and BestDeadliftKg

t.test(op$BestSquatKg,op$BestDeadliftKg -16 ,paired = TRUE)
```

the Result is

```
> t.test(op$BestSquatKg,op$BestDeadliftKg -16 ,paired = TRUE)



          Paired t-test



data:  op$BestSquatKg and op$BestDeadliftKg - 16

t = -0.52679, df = 6208, p-value = 0.5984

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -1.1158843  0.6431834

sample estimates:

mean of the differences

             -0.2363505
```

As is can be seen, the T value of the test is -0.52, and the degree of freedom is 6208. Therefore, we can use the Z-score table to calculate the probability of the difference between the Best Bench result and the Best Deadlift result being less than 16. According to the table probability of this is 30 %; thus, we can't reject the null hypothesis.

Bibliography

momova97 (2014). momova97/ALY6010_Movahedi: This is where I will keep my project's R code. [online] GitHub. Available at: https://github.com/momova97/ALY6010_Movahedi [Accessed 15 Mar. 2022].

OpenPowerlifting (2019). Powerlifting Database. [online] Kaggle.com. Available at: https://www.kaggle.com/open-powerlifting/powerlifting-database [Accessed 15 Mar. 2022].

Appendix

```
 print('Mohammad Hossein Movahedi')

print('Module 3 R practice')

#importing and instaling libraries

install.packages('FSA')

install.packages('magrittr')

install.packages('dplyr')

install.packages('tidyr')

install.packages('plyr')

install.packages('tidyverse')

install.packages('outliers')

install.packages('ggplot2')

install.packages('lubridate')



library(FSA)

library(magrittr)

library(dplyr)

library(tidyr)

library(plyr)

library(tidyverse)

library(scales)

library(lubridate)

library(ggplot2)

library(outliers)



#Importing dataset

data <- read.csv('openpowerlifting.csv')

#because the dataset is too large I choose firsta 10000 data to work with

data <-head(data,n = 10000)
```

```
# first of all I delete duplicate rows

op <- data[!duplicated(data), ]

#first we cheak to see if our data looks normal or not

ggplot(op, aes(x=op$Age)) + geom_histogram()

ggplot(op, aes(x=op$BodyweightKg)) + geom_histogram()

ggplot(op, aes(x=op$TotalKg)) + geom_histogram()

#then we delete outliers in age

boxplot(op$Age)$out

d <- boxplot(op$Age, plot=FALSE)$out

op<- op[-which(op$Age %in% d),]

#then we delete outliers in Bodyweightkg

boxplot(op$BodyweightKg)$out

d <- boxplot(op$BodyweightKg, plot=FALSE)$out

op<- op[-which(op$BodyweightKg %in% d),]

#then we delete outliers in Totol

boxplot(op$TotalKg)$out

d <- boxplot(op$TotalKg, plot=FALSE)$out

op<- op[-which(op$TotalKg %in% d),]

#then we subset the data by sex

mp <- filter(op,op$Sex == "M")

fp <- filter(op,op$Sex == "F")



# now we test to see if there is diffrence between men Bodyweight and the average

print('males')

aveW <- mean(op$BodyweightKg,na.rm = T)

aveA <- mean(op$Age,na.rm = T)

t.test(mp$BodyweightKg, mu = aveW)

t.test(mp$Age, mu = aveA)

# now the same test for weman
```

```
print('females')

aveW <- mean(op$BodyweightKg,na.rm = T)

aveA <- mean(op$Age,na.rm = T)

t.test(fp$BodyweightKg, mu = aveW)

t.test(fp$Age, mu = aveA)

# now we test if there is relation betweenBestBenchKg and BestDeadliftKg

t.test(op$BestSquatKg,op$BestDeadliftKg -16 ,paired = TRUE)
```