PA1: Reproducible Research

hakanh

Saturday, February 14, 2015

## 1. LOADING AND PREPROCESSING THE DATA

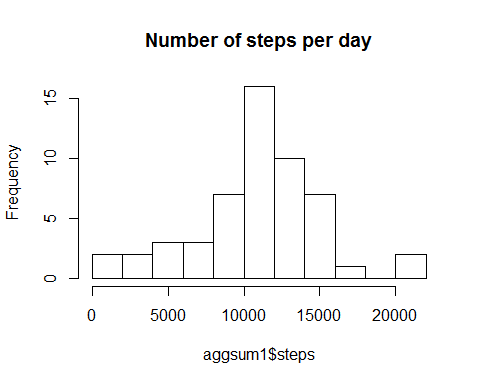
The data is read from the working directory into RStudio, and this data is called "data1".

setwd("C:/Users/Hakan/RepData")  
data1<-read.csv("reproduce.csv")

## 2. WHAT IS THE TOTAL NUMBER OF STEPS TAKEN PER DAY?

The total number of steps per day is caluclated and stored in "aggsum1". This is plotted in a histogam.

aggsum1<-aggregate(steps~date, data=data1,FUN="sum")  
histsum<-hist(aggsum1$steps, breaks=10, main="Number of steps per day")



The mean and median values of steps per day are calculated. They turn out to be 10766 and 10765.

meanvalue<-mean(aggsum1$steps)  
cat("The mean value of steps taken per day is", round(meanvalue, 0))

## The mean value of steps taken per day is 10766

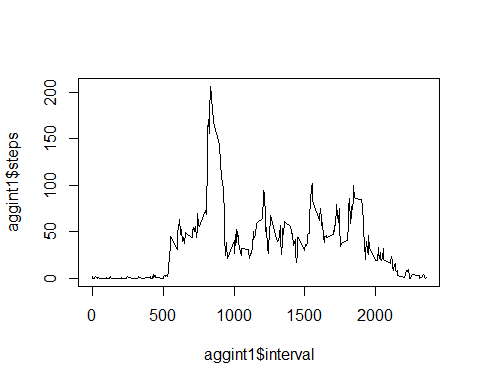
medianvalue<-median(aggsum1$steps)  
cat("The median value of steps taken per day is", round(medianvalue, 0))

## The median value of steps taken per day is 10765

## 3. WHAT IS THE AVERAGE DAILY ACTIVITY PATTERN?

The mean values of steps for each interval is calulated and put in "aggint1". (The number of steps for each interval is summarized for all days and divided with the number of days.) This is plotted in a time series plot.

aggint1<-aggregate(steps~interval, data=data1, FUN="mean")  
plot(aggint1$interval, aggint1$steps, type="l")



The interval with the highest number of steps is found by sorting "aggint1" on mean value of steps and checking the interval number on the top line. It turns out to be the interval at 8.35.

maxsteps<-aggint1[order(-aggint1$steps),]  
cat("The maximum number of steps in taken in the interval", maxsteps[1,1])

## The maximum number of steps in taken in the interval 835

## 4. IMPUTING MISSING VALUES

The number of line with NA is calculated by subtracing the the number of lines with a number equal to 0 or bigger from the total number of lines in "data1". It turns out be 2304 lines.

number<-subset(data1, steps>=0)  
NAnumber<-nrow(data1)-nrow(number)  
cat("The number of NA in the steps column is ", NAnumber)

## The number of NA in the steps column is 2304

The method chose to replace NA is to put in the value from the line above. As the value on the first line is NA, this value is initiated as 0. The number of replacements turns out to be 2304, according to the caculation above. The new dataset is called "data2".

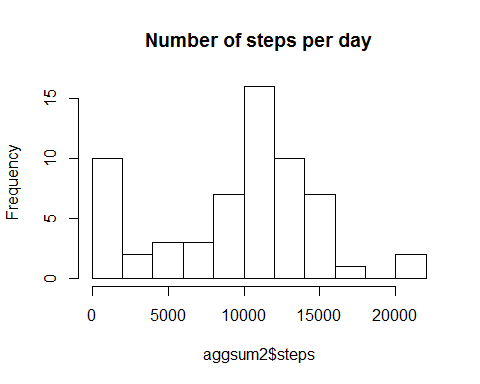
data2<-data1  
n<-1  
data2[1, 1]<-0  
for(i in 1:nrow(data2)){  
 if (is.na(data2[i, 1])){  
 data2[i,1]<-data2[i-1,1]  
 n<-n+1  
 }  
}  
cat("The number of replacement is " , n)

## The number of replacement is 2304

The histogram is drawn again, this time for the dataset with the NA replaced, and mean and average of steps is calculated again. It is obvious that the method I chose resulted in a higher number of days with no steps taken. It looks as if there are ten days with only NA that are not days with no steps.

This does have an effect on the mean values that goes down to 9354 from 10766. The median is of course not that sensitive and is only reduced to 10395 from 10765.

aggsum2<-aggregate(steps~date, data=data2,FUN="sum")  
histsum<-hist(aggsum2$steps, breaks=10, main="Number of steps per day")



meanvalue<-mean(aggsum2$steps)  
cat("The mean value of steps taken per day is", round(meanvalue, 0))

## The mean value of steps taken per day is 9354

medianvalue<-median(aggsum2$steps)  
cat("The median value of steps taken per day is", round(medianvalue, 0))

## The median value of steps taken per day is 10395