Activity Tracker - Reproducible research

Gabe Boer

9 oktober 2017

Header settings saying that code always needs to be shown.

knitr::opts\_chunk$set(echo = TRUE)  
setwd("~/Coursera R folder")

## Loading and preprocessing the data

### 1. Reading the data file

This is an R Markdown document for the Reproducible Research assignment. The first step is to manually manipulate the csv file in notepad. Removing all quotes (""). Didn't find a proper way in R to do it. The second step is reading the data from csv file.

act <- read.csv("activity.csv", sep = ",", header = FALSE, skip = 1, quote = "")  
  
names(act)[1] <- "Steps"  
names(act)[2] <- "Date"  
names(act)[3] <- "Interval"

### 2. Process/Transform the data (if necessary) into a format more suitable for your analysis

The date format is a bit awkward and needs some adjustments. Currently in "yyyy-mm-dd". With lubridate package some adjustments and additional values can be made (weekday) and make interval hours-minutes

#act$Steps <- as.numeric(act$Steps)  
act$Date <- as.Date(act$Date, format = "%Y-%m-%d")

A quick summary of the loaded data will give:

## Steps Date Interval   
## Min. : 0.00 Min. :2012-10-01 Min. : 0.0   
## 1st Qu.: 0.00 1st Qu.:2012-10-16 1st Qu.: 588.8   
## Median : 0.00 Median :2012-10-31 Median :1177.5   
## Mean : 37.38 Mean :2012-10-31 Mean :1177.5   
## 3rd Qu.: 12.00 3rd Qu.:2012-11-15 3rd Qu.:1766.2   
## Max. :806.00 Max. :2012-11-30 Max. :2355.0   
## NA's :2304

## What is mean total number of steps taken per day

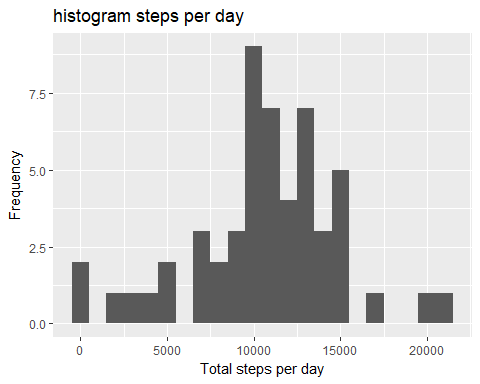
### 1. Number of steps calculation

totalsteps <- aggregate(Steps~Date, act, FUN = sum, na.action = na.omit)  
#head(totalsteps, 10)  
summary(totalsteps)

## Date Steps   
## Min. :2012-10-02 Min. : 41   
## 1st Qu.:2012-10-16 1st Qu.: 8841   
## Median :2012-10-29 Median :10765   
## Mean :2012-10-30 Mean :10766   
## 3rd Qu.:2012-11-16 3rd Qu.:13294   
## Max. :2012-11-29 Max. :21194

### 2. plot in histogram

library(ggplot2)  
qplot(totalsteps$Steps, geom = "histogram",   
 binwidth = 1000,   
   
 xlab = "Total steps per day",   
 ylab = "Frequency",  
 main = "histogram steps per day"  
 )



### 3a. the mean of the total number of steps taken per day

mean(totalsteps$Steps)

## [1] 10766.19

### 3b. the median of the total number of steps taken per day

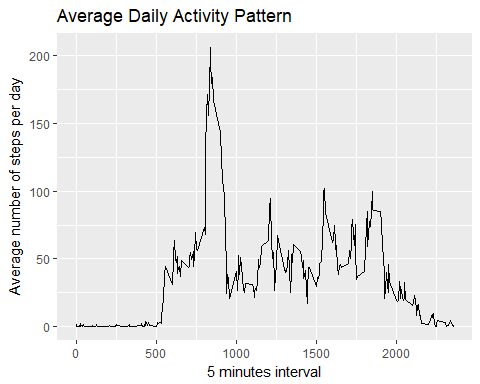
median(totalsteps$Steps)

## [1] 10765

## What is the average daily activity pattern?

### 1. Make a time series plot of the 5-min interval (x) and the avg number of steps (y)

min5 <- aggregate(Steps~Interval, data = act, FUN = mean, na.rm = TRUE)  
  
ggplot(  
 data = min5,  
 aes(  
 x = Interval,  
 y = Steps)) +  
 geom\_line() +  
 xlab("5 minutes interval") +  
 ylab("Average number of steps per day") +  
 ggtitle("Average Daily Activity Pattern")



### 2. which 5 min interval contains max number of steps?

max <- min5$Interval[which.max(min5$Steps)]  
max

## [1] 835

## Imputing missing values

The number of rows with NA values was allready available from the summary on the data set. ### 1. calculate and report the total number of missing values

summary(act)

## Steps Date Interval   
## Min. : 0.00 Min. :2012-10-01 Min. : 0.0   
## 1st Qu.: 0.00 1st Qu.:2012-10-16 1st Qu.: 588.8   
## Median : 0.00 Median :2012-10-31 Median :1177.5   
## Mean : 37.38 Mean :2012-10-31 Mean :1177.5   
## 3rd Qu.: 12.00 3rd Qu.:2012-11-15 3rd Qu.:1766.2   
## Max. :806.00 Max. :2012-11-30 Max. :2355.0   
## NA's :2304

Showing that in 'dimension Steps' there are 2.304 rows with value NA.

### 2. Devise strategy for filling the missing values in data set

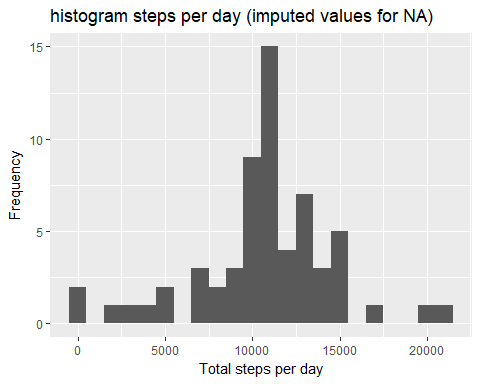
The 2.304 lines will get the average value of the interval (across days). ### 3. create new dataset including values for missing values (6)

# make copy of original df  
act2 <- act  
  
# NA records wegschrijven  
act2na <- is.na(act2$Steps)  
  
# determine average of interval without na records  
intavg <- tapply(act2$Steps, act2$Interval, mean, na.rm=TRUE, simplify = TRUE)  
  
# add values of average to the na records  
act2$Steps[act2na] <- intavg[as.character(act2$Interval[act2na])]  
head(act2, 10)

## Steps Date Interval  
## 1 1.7169811 2012-10-01 0  
## 2 0.3396226 2012-10-01 5  
## 3 0.1320755 2012-10-01 10  
## 4 0.1509434 2012-10-01 15  
## 5 0.0754717 2012-10-01 20  
## 6 2.0943396 2012-10-01 25  
## 7 0.5283019 2012-10-01 30  
## 8 0.8679245 2012-10-01 35  
## 9 0.0000000 2012-10-01 40  
## 10 1.4716981 2012-10-01 45

### 4. Make histogram (7)

totstepsim <- aggregate(Steps~Date, act2, FUN = sum)  
qplot(totstepsim$Steps, geom = "histogram",   
 binwidth = 1000,   
   
 xlab = "Total steps per day",   
 ylab = "Frequency",  
 main = "histogram steps per day (imputed values for NA)"  
 )



## Are there differences in activity patterns between weekdays and weekends?

### 1. Create a new factor variable in dataset

# df with imputed values is starting point, make copy  
act3 <- act2  
  
# create weekdays column using function weekdays()  
act3$wkd <- weekdays(act3$Date)  
  
# create column where is stated weekday or weekend using ifelse(test, yes, no)  
act3$work <- ifelse(act3$wkd == "zaterdag" | act3$wkd == "zondag", "weekend", "weekday")

### 2. Make a panel plot containing time series plot (8)

min52 <- aggregate(Steps~Interval + work, data = act3, FUN = mean)  
  
ggplot(  
 data = min52,  
 aes( x = Interval,  
 y = Steps,  
 color = work)) +  
 #facet\_grid(work~.) +  
 facet\_wrap(~work, ncol =1, nrow = 2) +  
 geom\_line() +   
 labs(title="Average number of daily steps",   
 x="Intervals",  
 y="Number of steps")

