Reproducible Research: Analysis of daily activity

Jagrut

07/08/2020

### Synopsis

This data analysis makes use of the activity data of an user over two months. The analysis is focused on finding insights like average steps taken each day, change in activity patterns between weekdays and weekends, etc.

### Loading and preprocessing the data

data <- read.csv("activity.csv")  
data$date <- as.Date(data$date)

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

library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

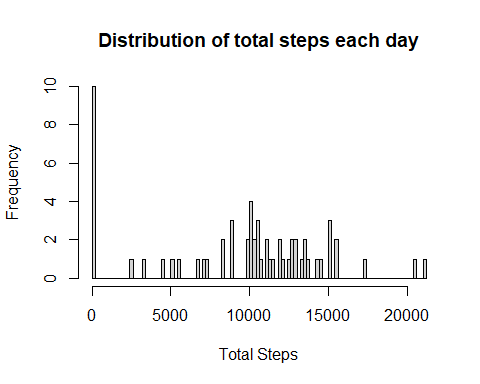
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

#Calculate total steps taken by date  
Total.steps <- data %>% group\_by(date) %>%   
 summarise(Total.steps = sum(steps, na.rm = T))

## `summarise()` ungrouping output (override with `.groups` argument)

hist(Total.steps$Total.steps, breaks = 100,   
 main = "Distribution of total steps each day",   
 xlab = "Total Steps")



mean(Total.steps$Total.steps, na.rm = T)

## [1] 9354.23

median(Total.steps$Total.steps, na.rm = T)

## [1] 10395

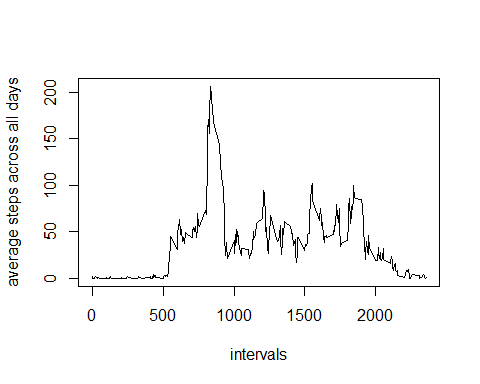
The mean and median are 9354.23 and 10395 respectively.

### What is the average daily activity pattern?

act.arranged = arrange(data, interval)  
average.steps = act.arranged %>% group\_by(interval) %>%   
 summarise(mean(steps, na.rm = T))

## `summarise()` ungrouping output (override with `.groups` argument)

names(average.steps) <- c("intervals", "ave.steps")  
plot(average.steps, type = "l", ylab = "average steps across all days")



# interval for maximum number of steps taken  
max.id <- which(average.steps$ave.steps ==   
 max(average.steps$ave.steps))  
#Calculate the 5-min interval that gives maximum average steps  
average.steps[max.id, ]$intervals

## [1] 835

The plot shows that the average steps daily follows an uneven distribution.It is also found that the peak activity is found at 835th interval in a day, on an average. ### Imputing missing values

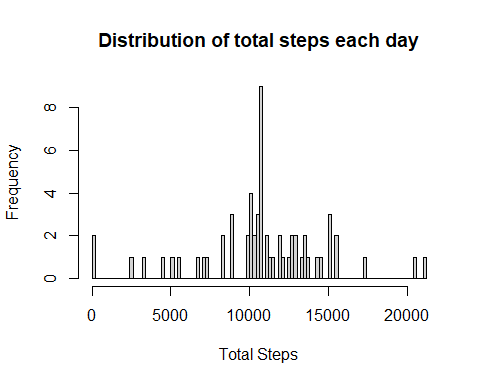
# number of rows containing NAs  
nrow(data) - sum(complete.cases(data))

## [1] 2304

#data frame containing NAs and one that does not  
data.na <- data[!complete.cases(data),]  
data.free <- data[complete.cases(data), ]  
na.id = which(!complete.cases(data) == T)  
#Filling in the NAs with mean for that 5-min time interval  
for(i in na.id){  
 ave.temp = c()  
 ave.temp = average.steps$ave.steps[average.steps$intervals == data$interval[i]]  
 data[i,1] = ave.temp  
}  
Total.steps <- data %>% group\_by(date) %>%   
 summarise(Total.steps = sum(steps, na.rm = T))

## `summarise()` ungrouping output (override with `.groups` argument)

hist(Total.steps$Total.steps, breaks = 100,   
 main = "Distribution of total steps each day",   
 xlab = "Total Steps")



mean(Total.steps$Total.steps, na.rm = T)

## [1] 10766.19

median(Total.steps$Total.steps, na.rm = T)

## [1] 10766.19

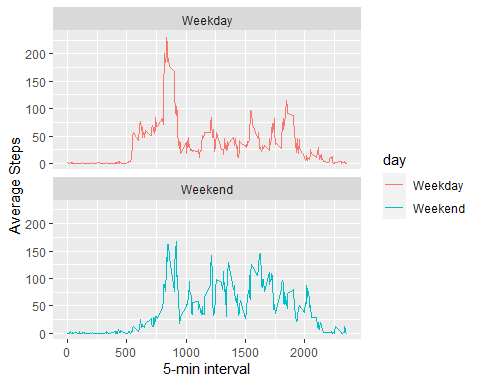
There are various ways to replace missing values. Here, they are replaced with the average steps for the relevant 5-min time interval. The little skewness observed before is gone now. The distribution is more like a normal distribution. The mean and median comes out to be the same when we replace the missing values in this manner.

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

data$day <- weekdays(data$date)  
for(i in 1:nrow(data)){  
 if(data$day[i] == "Saturday" | data$day[i] == "Sunday"){  
 data$day[i] <- "Weekend"  
 }  
 else{  
 data$day[i] <- "Weekday"  
 }  
}  
data.sub = data %>% group\_by(interval, day) %>% summarise(avg.steps = mean(steps))

## `summarise()` regrouping output by 'interval' (override with `.groups` argument)

ggplot(data = data.sub, aes(x = interval, y = avg.steps)) +  
 geom\_line (aes(col = day)) + facet\_wrap(~ day, ncol = 1) +  
 ylab("Average Steps") + xlab("5-min interval")



The plot clearly shows that there are differences in the average steps taken during weekdays and weekends. The average steps taken between 500 and 750th intervals are greater for weekdays than weekends. This may suggest that the user is getting ready in the morning to go to work on weekdays. After the usual surge of steps around 830th interval in either case, more steps are taken on weekends than weekdays. This may suggest that the user is at the gym during the weekend, owing to being more available on weekends.