PA1_template

Rosane Schossler

12/06/2020

Reproducible Research - Wee 2 - Course Project 1

This is the project 1 from week 2 - Cousera - Module 5 (Reproducible Research)

Show current directory

```
        getwd()
```

```
## [1] "C:/Users/rs14351/Desktop/TRACE/Coursera/Modulo 5/week 2/Course Project 1"
```

Configure the directory where the file is

```
setwd("C:/Users/rs14351/Desktop/TRACE/Coursera/Modulo 5/week 2/Course Project 1")
```

Library required

Loading and preprocessing the data

1-Load the data

```
activity<-read.csv("activity.csv")
head(activity,10)
```

```
##    steps       date interval
## 1     NA 2012-10-01        0
## 2     NA 2012-10-01        5
## 3     NA 2012-10-01       10
## 4     NA 2012-10-01       15
## 5     NA 2012-10-01       20
## 6     NA 2012-10-01       25
## 7     NA 2012-10-01       30
## 8     NA 2012-10-01       35
## 9     NA 2012-10-01       40
## 10    NA 2012-10-01       45
```

What is mean total number of steps taken per day?

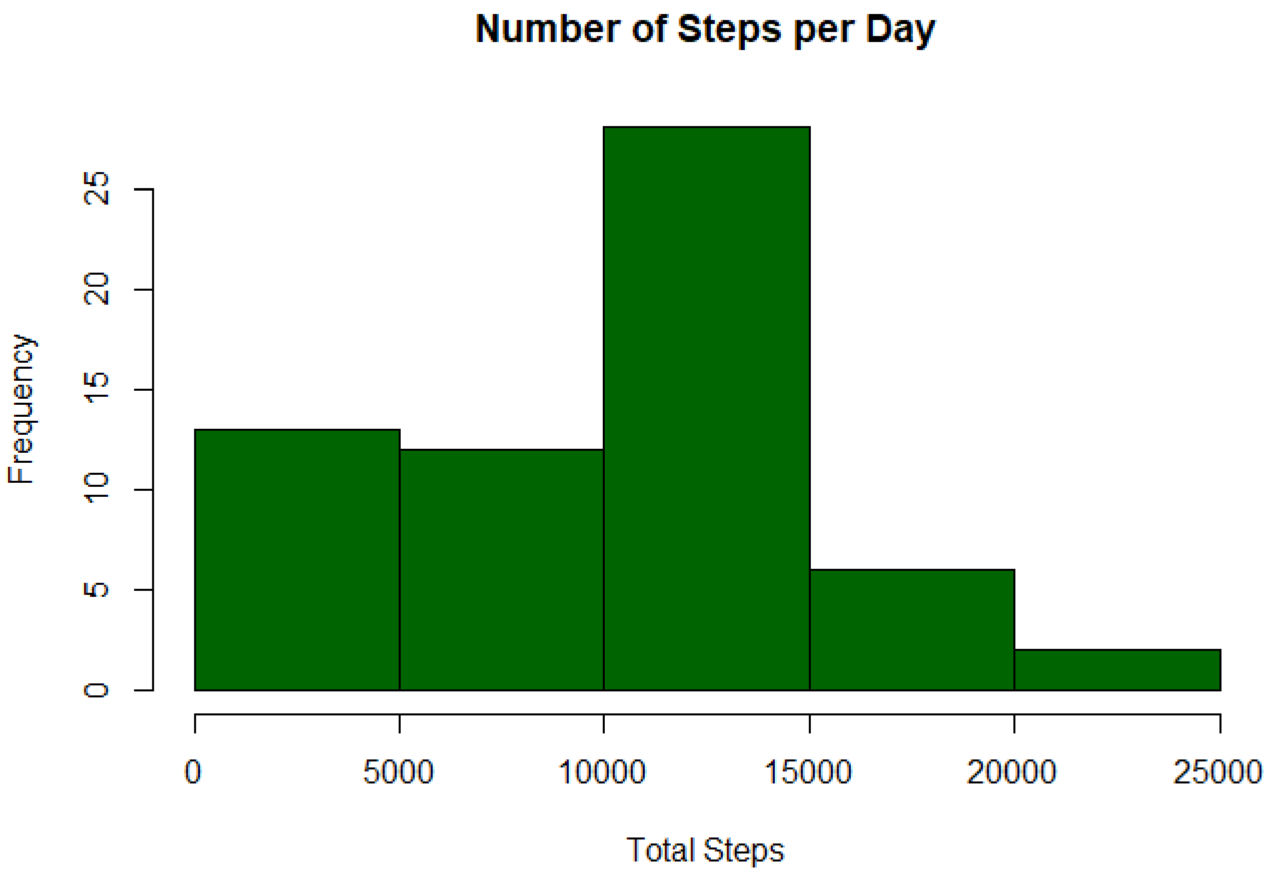
1-Calculate the total number of steps taken per day

```
totalStepsByDay<-aggregate(steps~date, activity, sum)
head(totalStepsByDay,10)
```

```
##          date steps
## 1  2012-10-02   126
## 2  2012-10-03 11352
## 3  2012-10-04 12116
## 4  2012-10-05 13294
## 5  2012-10-06 15420
## 6  2012-10-07 11015
## 7  2012-10-09 12811
## 8  2012-10-10  9900
## 9  2012-10-11 10304
## 10 2012-10-12 17382
```

2-Histogram of the total number of steps taken each day

```
steps<-with(activity,tapply(steps,date,sum,na.rm=TRUE))
hist(steps,col = "darkgreen",xlab = "Total Steps",ylab = "Frequency",main = "Number of Steps per Day")
```



3-Calculate and report the mean and median of the total number of steps taken per day

```
##Mean

print(meansteps <- mean(steps))
```

```
## [1] 9354.23
```

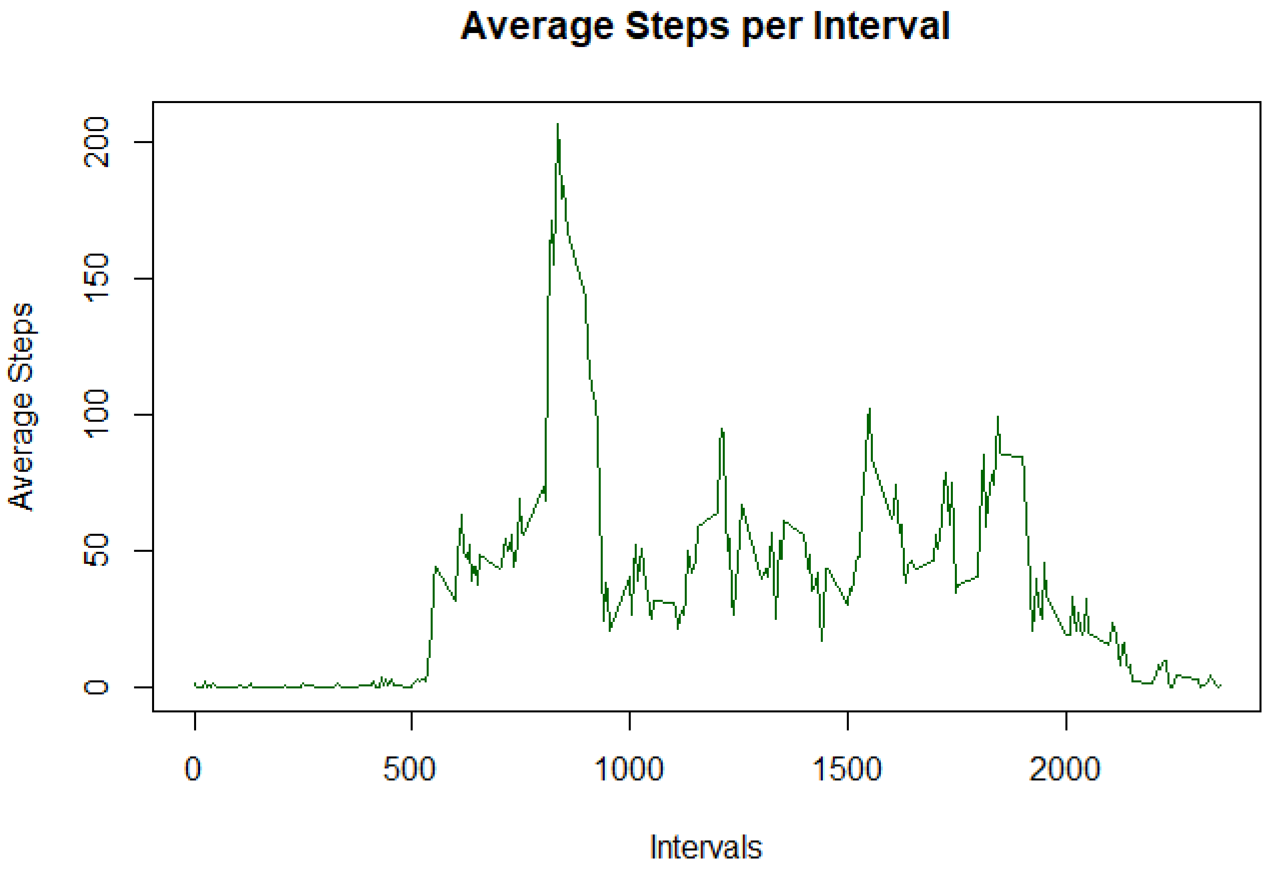
```
##Median

print(mediansteps <- median(steps))
```

```
## [1] 10395
```

What is the average daily activity pattern?

1-Plot of the 5-minute interval and the average number of steps taken, averaged across all days

```
avg_steps<-with(activity,tapply(steps,interval,mean,na.rm=TRUE))
intervals<-unique(activity$interval)
new<-data.frame(cbind(avg_steps,intervals))
plot(new$intervals,new$avg_steps, col="darkgreen",type = "l",xlab = "Intervals",
     ylab = "Average Steps",main = "Average Steps per Interval"
     )
```



2-Which 5-minute interval, on average across all the days in the dataset, contains the maximum number of steps?

```
index<-which.max(new$avg_steps)
max<-new[index,2]
```

Imputing missing values

1-Calculate and report the total number of missing values in the dataset (i.e. the total number of rows with NAs)

```
print(sum(is.na(activity$steps)))
```

```
## [1] 2304
```

2-Devise a strategy for filling in all of the missing values in the dataset. The strategy does not need to be sophisticated. For example, you could use the mean/median for that day, or the mean for that 5-minute interval, etc.

```
replacewithmean <- function(x) replace(x, is.na(x), mean(x, na.rm = TRUE))
```

3-Create a new dataset that is equal to the original dataset but with the missing data filled in.

```
meandata <- activity%>% group_by(interval) %>% mutate(steps= replacewithmean(steps))
head(meandata,10)
```

```
## # A tibble: 10 x 3
## # Groups:   interval [10]
##     steps date       interval
##     <dbl> <fct>         <int>
##  1 1.72   2012-10-01        0
##  2 0.340  2012-10-01        5
##  3 0.132  2012-10-01       10
##  4 0.151  2012-10-01       15
##  5 0.0755 2012-10-01       20
##  6 2.09   2012-10-01       25
##  7 0.528  2012-10-01       30
##  8 0.868  2012-10-01       35
##  9 0      2012-10-01       40
## 10 1.47   2012-10-01       45
```

4-Make a histogram of the total number of steps taken each day and Calculate and report the mean and median total number of steps taken per day. Do these values differ from the estimates from the first part of the assignment? What is the impact of imputing missing data on the estimates of the total daily number of steps?

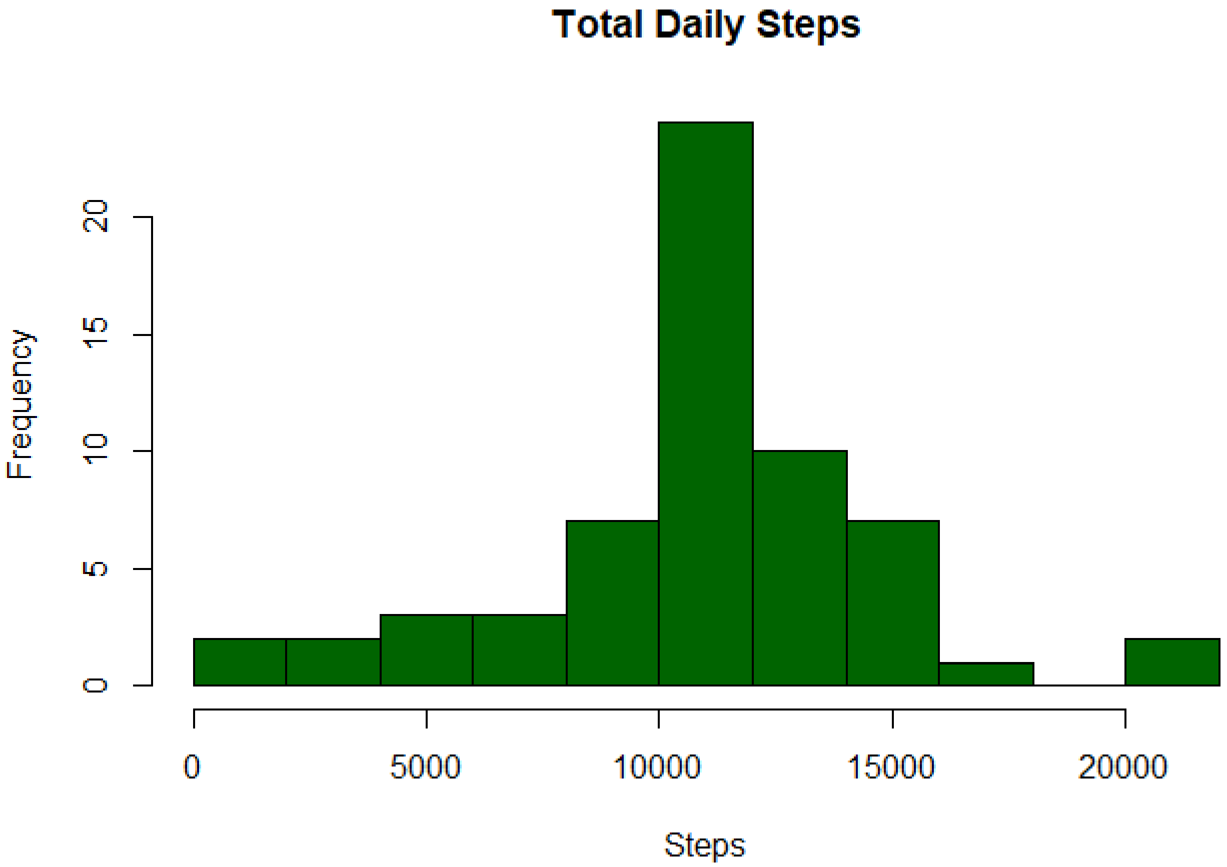
```
FullSummedDataByDay <- aggregate(meandata$steps, by=list(meandata$date), sum)

names(FullSummedDataByDay)[1] ="date"
names(FullSummedDataByDay)[2] ="totalsteps"
head(FullSummedDataByDay,15)
```

```
##          date totalsteps
## 1  2012-10-01   10766.19
## 2  2012-10-02     126.00
## 3  2012-10-03   11352.00
## 4  2012-10-04   12116.00
## 5  2012-10-05   13294.00
## 6  2012-10-06   15420.00
## 7  2012-10-07   11015.00
## 8  2012-10-08   10766.19
## 9  2012-10-09   12811.00
## 10 2012-10-10    9900.00
## 11 2012-10-11   10304.00
## 12 2012-10-12   17382.00
## 13 2012-10-13   12426.00
## 14 2012-10-14   15098.00
## 15 2012-10-15   10139.00
```

```
summary(FullSummedDataByDay)
```

```
##          date      totalsteps
##  2012-10-01: 1   Min.   :   41
##  2012-10-02: 1   1st Qu.: 9819
##  2012-10-03: 1   Median :10766
##  2012-10-04: 1   Mean   :10766
##  2012-10-05: 1   3rd Qu.:12811
##  2012-10-06: 1   Max.   :21194
##  (Other)   :55
```

```
hist(FullSummedDataByDay$totalsteps,col = "darkgreen",xlab = "Steps", ylab = "Frequency", main = "Total Daily Steps", breaks = 10)
```



```
print(mean_steps_2<-mean(FullSummedDataByDay$totalsteps))
```

```
## [1] 10766.19
```

```
print(median_steps_2<-median(FullSummedDataByDay$totalsteps))
```

```
## [1] 10766.19
```

There is no difference in mean before and after imputing

Are there differences in activity patterns between weekdays and weekends?

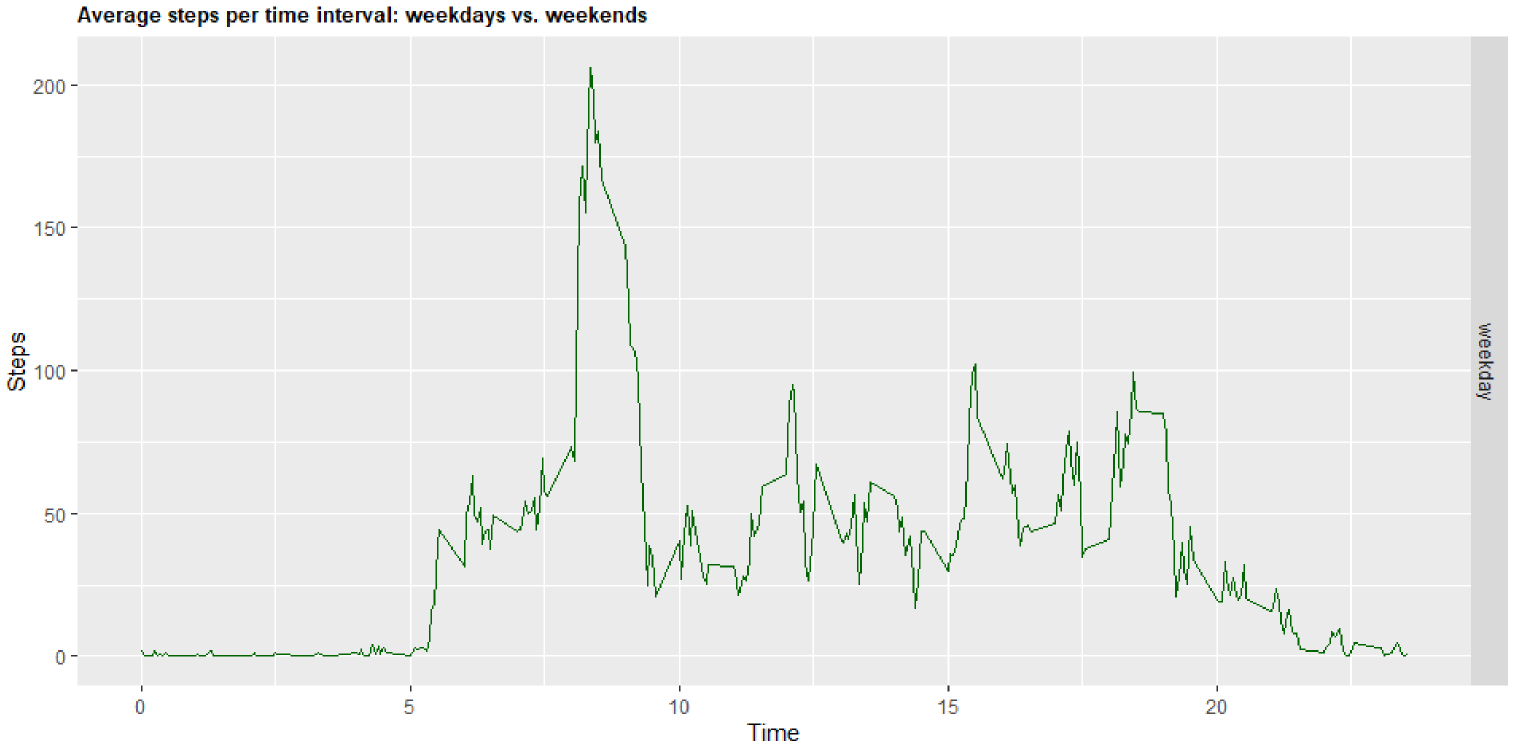
1-Create a new factor variable in the dataset with two levels – “weekday” and “weekend” indicating whether a given date is a weekday or weekend day.

```
meandata$RealDate <- as.Date(meandata$date, format = "%Y-%m-%d")
meandata$weekday <- weekdays(meandata$RealDate)
meandata$DayType <- ifelse(meandata$weekday=='Saturday' | meandata$weekday=='Sunday', 'weekend','weekday')
head(meandata, n=10)
```

```
## # A tibble: 10 x 6
## # Groups:   interval [10]
##     steps date       interval RealDate   weekday       DayType
##     <dbl> <fct>         <int> <date>     <chr>         <chr>  
##  1 1.72   2012-10-01        0 2012-10-01 segunda-feira weekday
##  2 0.340  2012-10-01        5 2012-10-01 segunda-feira weekday
##  3 0.132  2012-10-01       10 2012-10-01 segunda-feira weekday
##  4 0.151  2012-10-01       15 2012-10-01 segunda-feira weekday
##  5 0.0755 2012-10-01       20 2012-10-01 segunda-feira weekday
##  6 2.09   2012-10-01       25 2012-10-01 segunda-feira weekday
##  7 0.528  2012-10-01       30 2012-10-01 segunda-feira weekday
##  8 0.868  2012-10-01       35 2012-10-01 segunda-feira weekday
##  9 0      2012-10-01       40 2012-10-01 segunda-feira weekday
## 10 1.47   2012-10-01       45 2012-10-01 segunda-feira weekday
```

2-Make a panel plot containing a time series plot (i.e. type = “l”) of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all weekday days or weekend days (y-axis). See the README file in the GitHub repository to see an example of what this plot should look like using simulated data.

```
FullSummedDataByDay <- aggregate(steps~interval+DayType,data=meandata,FUN=mean,na.action=na.omit)
FullSummedDataByDay$time <- FullSummedDataByDay$interval/100
j <- ggplot(FullSummedDataByDay, aes(time, steps))
j+geom_line(col="darkgreen")+ggtitle("Average steps per time interval: weekdays vs. weekends")+xlab("Ti
me")+ylab("Steps")+theme(plot.title = element_text(face="bold", size=10))+facet_grid(DayType ~ .)
```


Average steps per time interval: weekdays vs. weekends
Steps
Time
weekday