Reproducible Research: Peer Assessment 1

Kwasi Abrefa-Kodom

June 16, 2016

Introduction

It is now possible to collect a large amount of data about personal movement using activity monitoring devices such as a Fitbit, Nike Fuelband, or Jawbone Up. These type of devices are part of the "quantified self" movement -- a group of enthusiasts who take measurements about themselves regularly to improve their health, to find patterns in their behavior, or because they are tech geeks. But these data remain under-utilized both because the raw data are hard to obtain and there is a lack of statistical methods and software for processing and interpreting the data.

This assignment makes use of data from a personal activity monitoring device. This device collects data at 5 minute intervals through out the day. The data consists of two months of data from an anonymous individual collected during the months of October and November, 2012 and include the number of steps taken in 5 minute intervals each day.

Assignment

This assignment will be described in multiple parts. You will need to write a report that answers the questions detailed below. Ultimately, you will need to complete the entire assignment in a single R markdown document that can be processed by knitr and be transformed into an HTML file.

Throughout your report make sure you always include the code that you used to generate the output you present. When writing code chunks in the R markdown document, always use echo = TRUE so that someone else will be able to read the code. This assignment will be evaluated via peer assessment so it is essential that your peer evaluators be able to review the code for your analysis.

For the plotting aspects of this assignment, feel free to use any plotting system in R (i.e., base, lattice, ggplot2)

Fork/clone the GitHub repository created for this assignment. You will submit this assignment by pushing your completed files into your forked repository on GitHub. The assignment submission will consist of the URL to your GitHub repository and the SHA-1 commit ID for your repository state.

NOTE: The GitHub repository also contains the dataset for the assignment so you do not have to download the data separately.

Loading and preprocessing the data

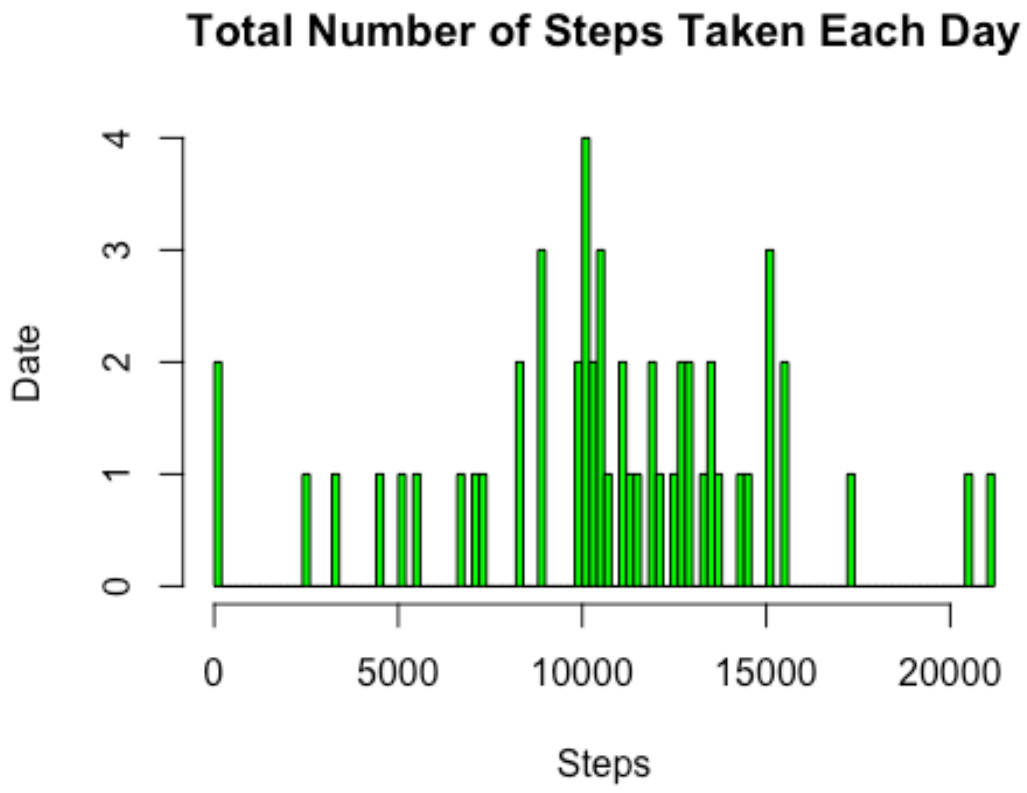
Show any code that is needed to

```
# 1. Load the data(i.e. read.csv())

actData <- read.csv("./datasciencecoursera/activity.csv")
# actData

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

transform_ActData <- na.omit(actData)
#transform_ActData
```

What is mean total number of steps taken per day?

For this part of the assignment, you can ignore the missing values in the dataset.

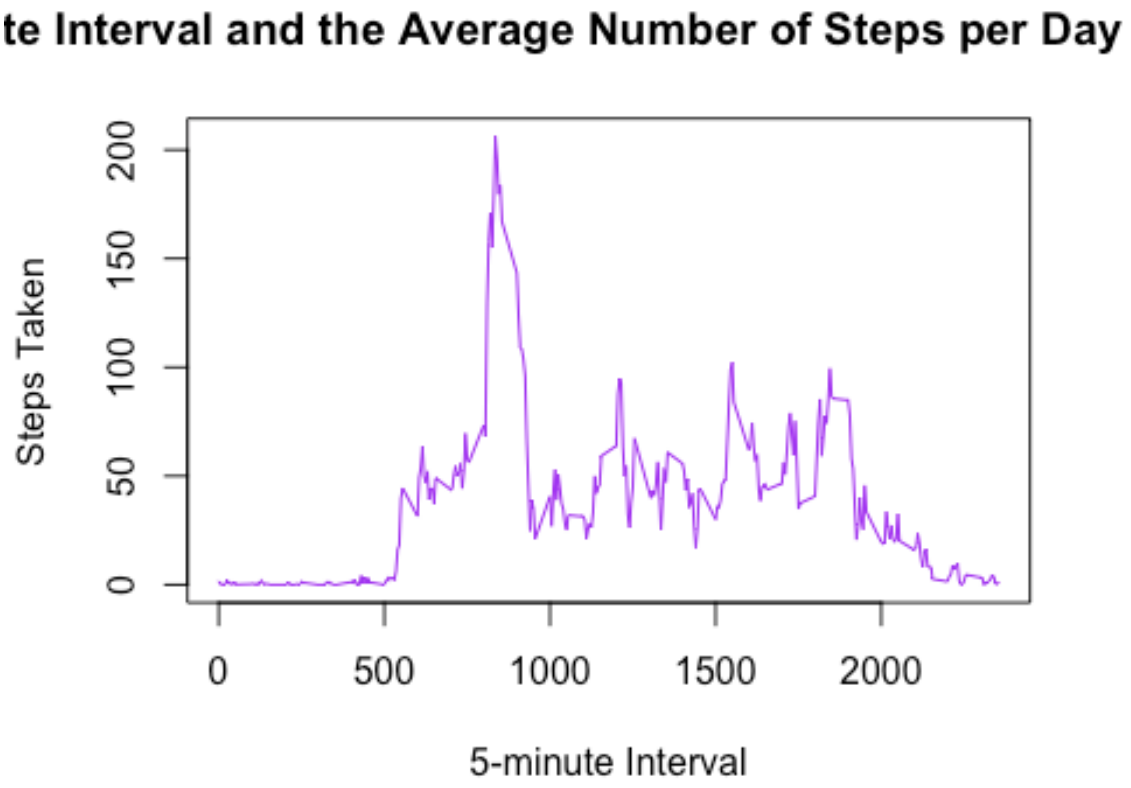
```
# 1. Make a histogram of the total number of steps taken each day

stepsDate <- aggregate(steps ~ date, transform_ActData, sum)
x <- as.numeric(stepsDate$steps)
hist(x, main = "Total Number of Steps Taken Each Day", breaks = 100, col =
"green", xlab = "Steps", ylab = "Date")
```



```
# 2. Calculate and report the mean and median total number of steps taken per
day

# Report of the mean of the total number of steps taken per day is
mean_StepsTaken <- mean(stepsDate$steps)
mean_StepsTaken

## [1] 10766.19

# Report of the median of the total number of steps taken per day is
median_StepsTaken <- median(stepsDate$steps)
median_StepsTaken

## [1] 10765
```

What is the average daily activity pattern?

```
# 1. Make 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 days (y-
axis)

stepsInterval <- aggregate(steps ~ interval, transform_ActData, mean)

plot(stepsInterval$interval, stepsInterval$steps, main = "5-Minute Interval
and the Average Number of Steps per Day by Interval", col = "purple", type =
"l", xlab = "5-minute Interval", ylab = "Steps Taken")
```



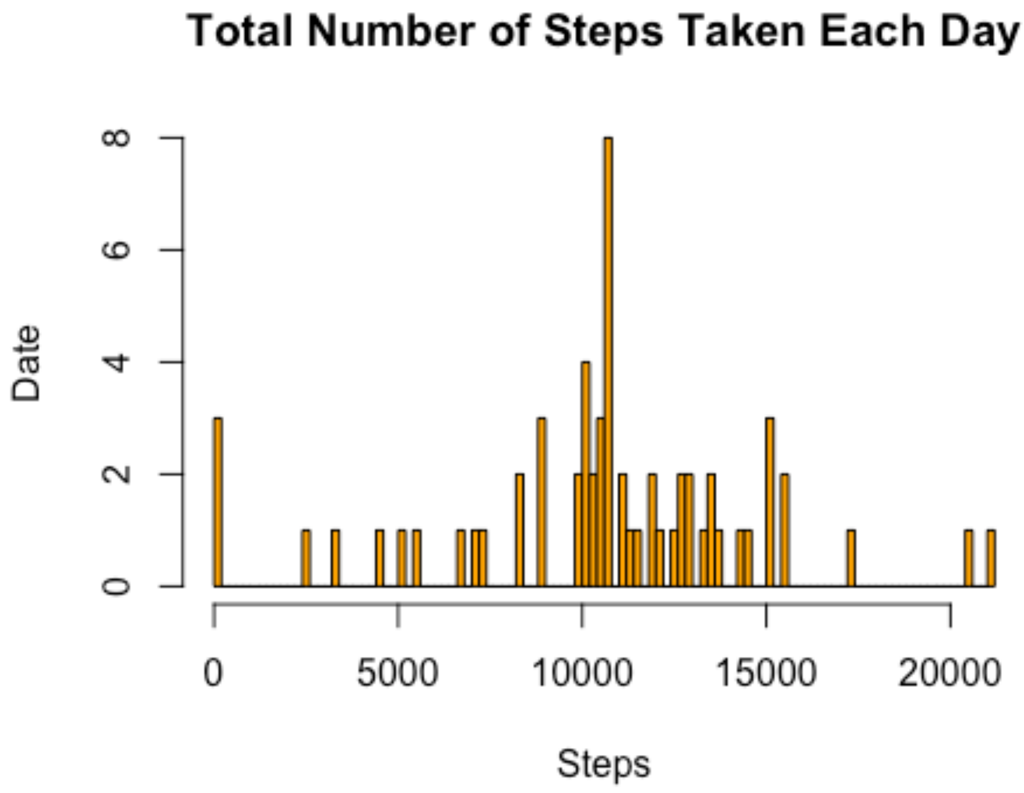
```
# 2. Which 5-minute interval, on average across all the days in the dataset,
contains the maximum number of steps?

highInterval <- stepsInterval[which.max(stepsInterval$steps),1]
highInterval

## [1] 835
```

Imputing missing values

Note that there are a number of days/intervals where there are missing values (coded as NA). The presence of missing days may introduce bias into some calculations or summaries of the data.

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

summary(is.na(actData$steps))

##    Mode   FALSE    TRUE    NA's
## logical   15264    2304       0
```

```
# 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.

filledData <- transform(actData, steps = ifelse(is.na(actData$steps),
stepsInterval$steps[match(actData$interval, stepsInterval$interval)],
actData$steps))

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

filledData[as.character(filledData$date) == "2012-10-01", 1] <- 0

# 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?

stepsDate1 <- aggregate(steps ~ date, filledData, sum)
x1 <- stepsDate1$steps
hist(x1, main = "Total Number of Steps Taken Each Day", breaks = 100, col =
"orange", xlab = "Steps", ylab = "Date")
```



```
# Report of the mean of the total number of steps taken per day including the
NAs is
mean_StepsTaken1 <- mean(stepsDate1$steps)
mean_StepsTaken1

## [1] 10589.69

# Report of the median of the total number of steps taken per day including
the NAs is
median_StepsTaken1 <- median(stepsDate1$steps)
median_StepsTaken1

## [1] 10766.19

# Difference in Values
Difference_in_Means <- mean_StepsTaken1 - mean_StepsTaken
Difference_in_Means

## [1] -176.4949

Difference_in_Medians <- median_StepsTaken1 - median_StepsTaken
Difference_in_Medians

## [1] 1.188679
```

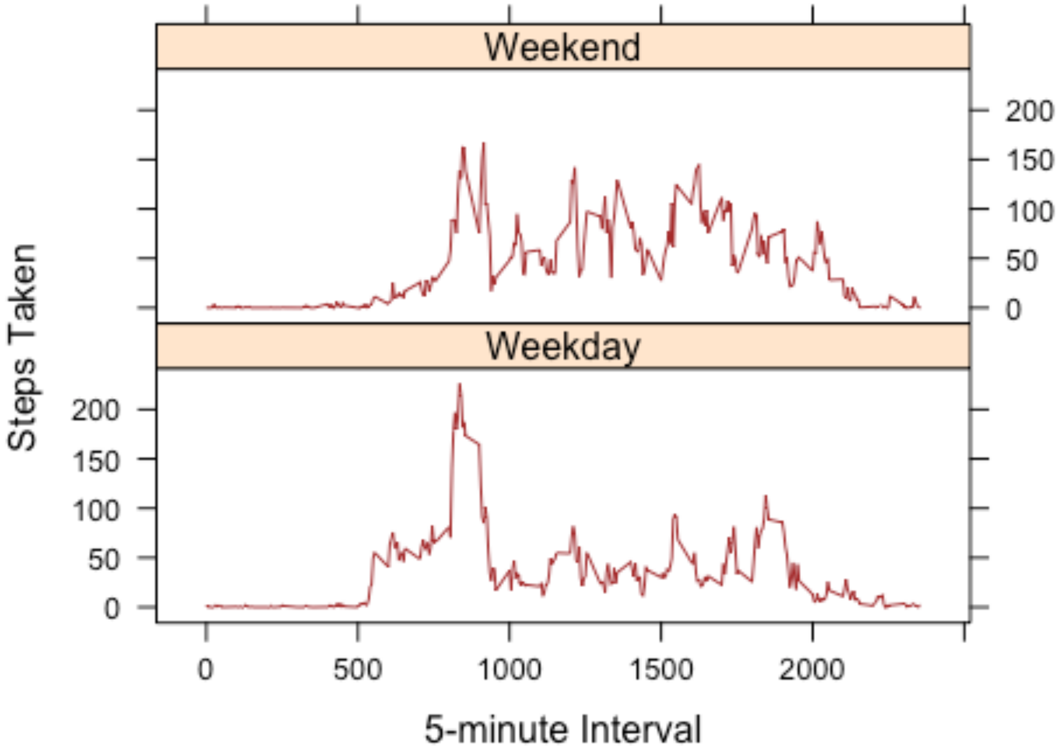
```
# Impact
Impact <- sum(stepsDate1$steps) - sum(stepsDate$steps)
Impact

## [1] 75363.32
```

Are there differences in activity patterns between weekdays and weekends?

For this part the weekdays() function may be of some help here. Use the dataset with the filled-in missing values for this part.

```
# 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.

weekdays <- c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday")
filledData$dow =
as.factor(ifelse(is.element(weekdays(as.Date(filledData$date)),weekdays),
"Weekday", "Weekend"))

# 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). The plot should look
something like the sample panel plot, which was created using simulated data:

library(lattice)

stepsInterval1 <- aggregate(steps ~ interval + dow, filledData, mean)
xyplot(stepsInterval1$steps ~ stepsInterval1$interval|stepsInterval1$dow,
main = "5-Minute Interval and the Average Number of Steps per Day by
Interval", col = "brown", xlab = "5-minute Interval", ylab = "Steps Taken",
layout = c(1, 2), type = "l")
```

Interval and the Average Number of Steps per Day by


Weekend
Weekday
Steps Taken
5-minute Interval