Reproducible Research Project

Rupesh Patel

25 October 2018

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.

Questions to be answered:

-What is mean total number of steps taken per day? -What is the average daily activity pattern? -Imputing missing values -Are there differences in activity patterns between weekdays and weekends?

Setting global option to turn warnings off

```
knitr::opts_chunk$set(warning=FALSE)
```

Loading Packages and preprocessing the data

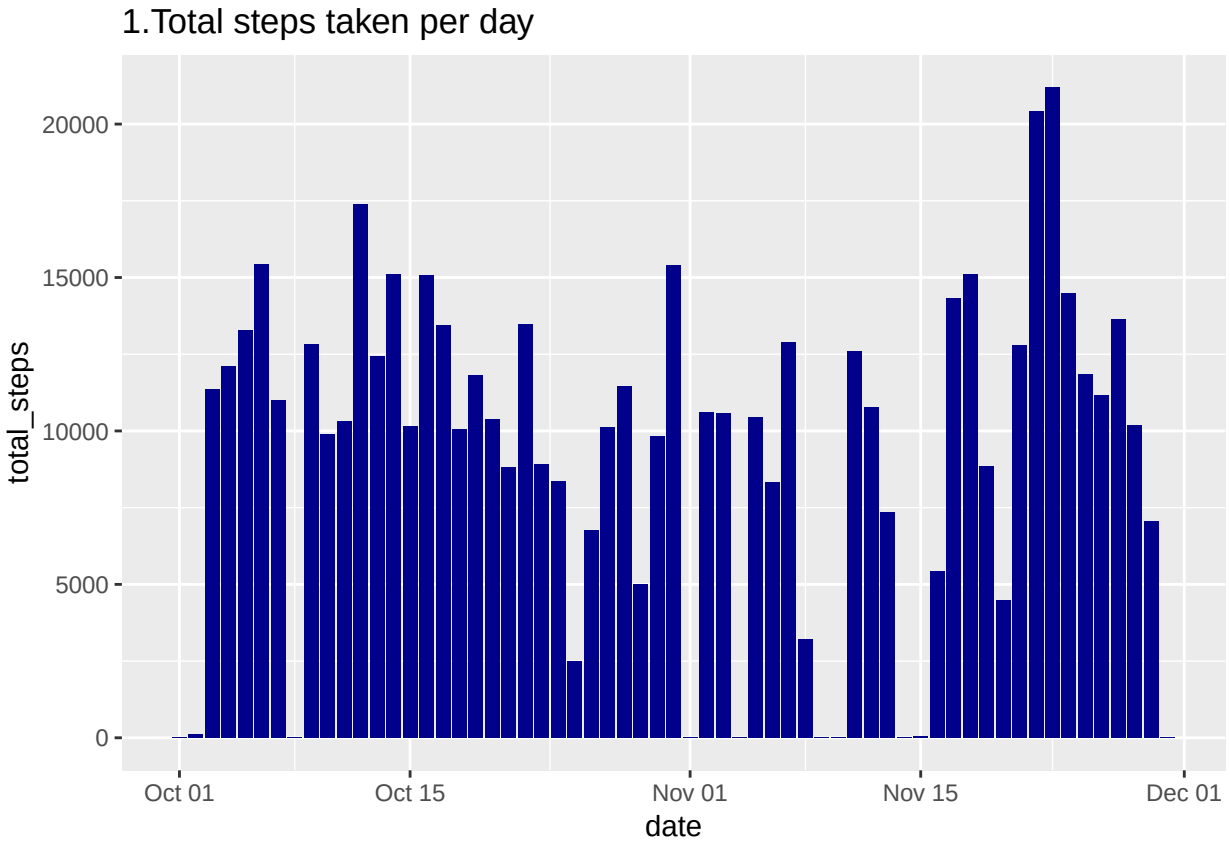
```
library(ggplot2)
library(lubridate)
library(dplyr)
library(gridExtra)
```

```
active <- read.csv("E:/R_Studio/activity.csv")

active$date = ymd(active$date)

summary(active)
```

```
##      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
```

1.What is mean total number of steps taken per day?

```
act1 <- active %>%
        group_by(date) %>%
        summarise(total_steps = sum(steps,na.rm = TRUE))

plot1 <- ggplot(act1,aes(x = date,y =  total_steps)) + geom_col(fill = "darkblue") + ggtitle("1.Total s
print(plot1)
```



Here is the mean of the total number of steps taken per day

```
mean(act1$total_steps,na.rm = TRUE)
```

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

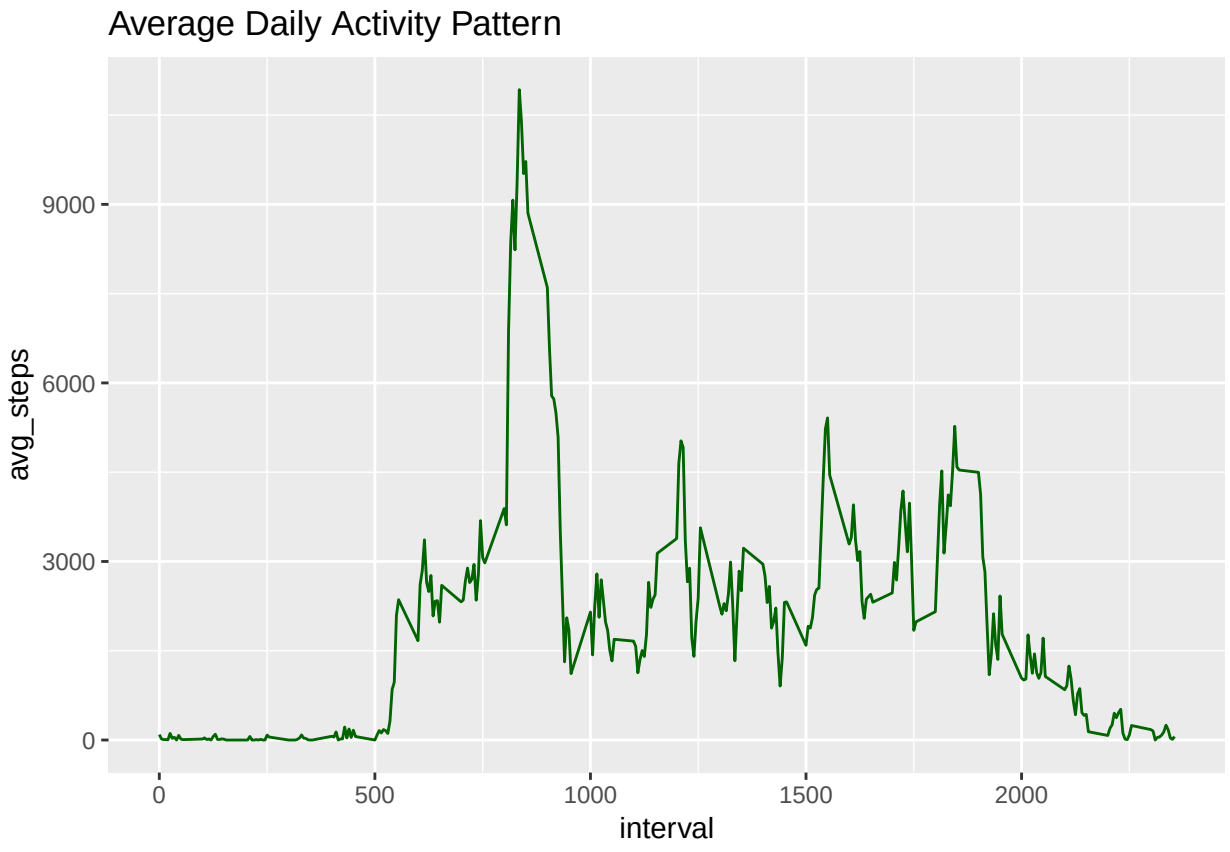
Here is the median of the total number of steps taken per day:

```
median(act1$total_steps,na.rm = TRUE)
```

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

2.What is the average daily activity pattern?

```
act2 <- active %>%
        group_by(interval) %>%
        summarise(avg_steps = sum(steps,na.rm = TRUE))

plot2 <- ggplot(act2,aes(x = interval,y = avg_steps)) + geom_line(colour = "darkgreen") + ggtitle("Aver

print(plot2)
```



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

```
act2$interval[act2$avg_steps == max(act2$avg_steps,na.rm = TRUE)]
```

```
## [1] 835
```

3. Imputing missing values

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.

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

```
sum(is.na(active))
```

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

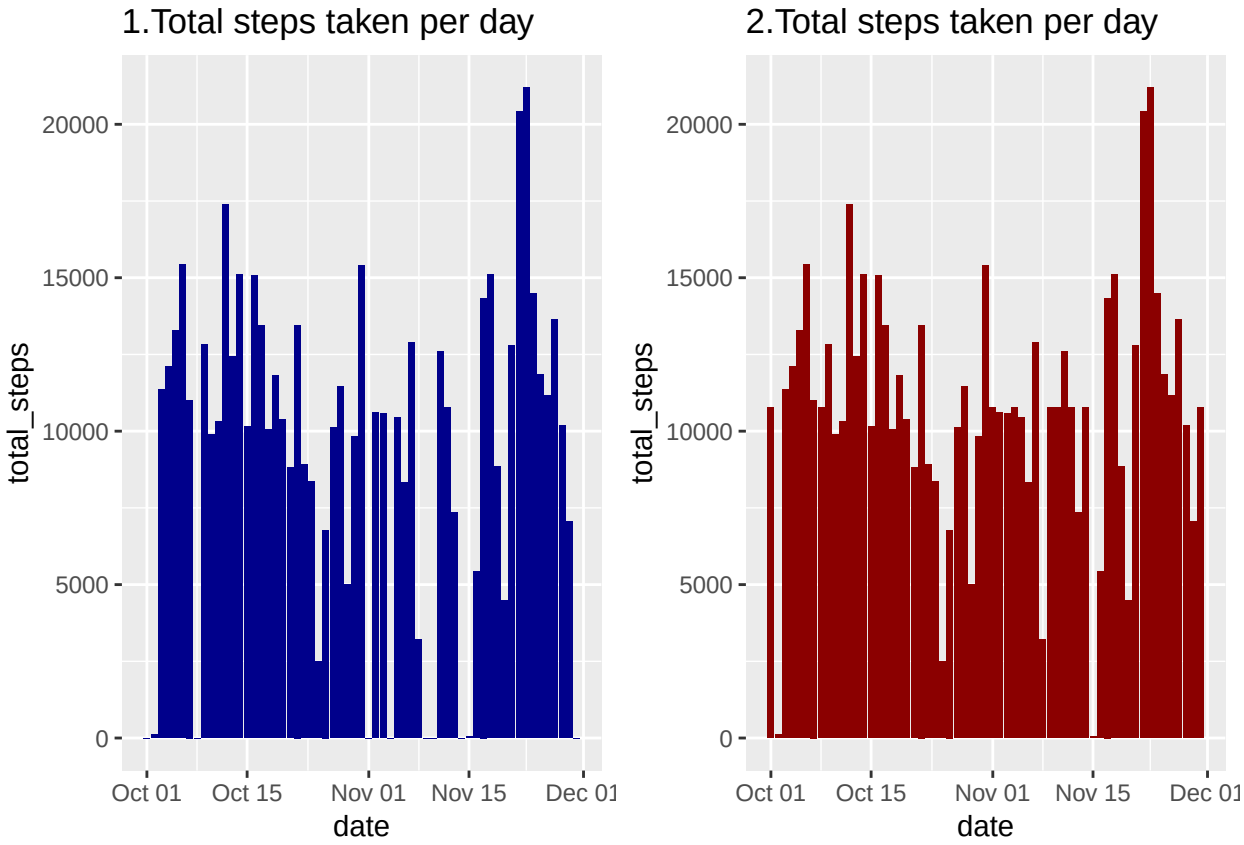
Devise a strategy for filling in all of the missing values in the dataset. I am going to substitute each NA with a fixed value. I set the fixed value equivalent to the overall mean of the variable activity$steps.

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

```
act3 <- active

act3$steps[is.na(act3$steps)] <- mean(act3$steps,na.rm = TRUE)
```

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?

```
act3$date = ymd(act3$date)
act4 <- act3 %>%
        group_by(date) %>%
        summarise(total_steps = sum(steps))


plot3 <- ggplot(act4,aes(x = date,y =  total_steps)) + geom_col(fill = "darkred") + ggtitle("2.Total st
grid.arrange(plot1, plot3, ncol=2)
```



4. Are there differences in activity patterns between weekdays and weekends?

Filtering Data based on Weekday and Weekend

```
act3$weekday <- factor(format(act3$date, "%A"))

levels(act3$weekday) <- list(weekday = c("Monday", "Tuesday",
                                         "Wednesday", "Thursday",
                                         "Friday"), weekend =
                                       c("Saturday", "Sunday"))
act5 <- act3 %>%
        filter(weekday == "weekday") %>%
        group_by(interval) %>%
        summarise(avg_steps = sum(steps))


act6 <- act3 %>%
        filter(weekday == "weekend") %>%
        group_by(interval) %>%
        summarise(avg_steps = sum(steps))
```

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).

```
plot4 <- ggplot(act5,aes(x = interval,y = avg_steps)) + geom_line(colour = "darkgreen") +
        ggtitle("Weekday")
```

```
plot5 <- ggplot(act6,aes(x = interval,y = avg_steps)) + geom_line(colour = "darkred") +
        ggtitle("Weekend")

grid.arrange(plot4, plot5, nrow=2)
```