Reproducable Research Project 1

fvon

Monday, March 09, 2015

Introduction

It is now possible to collect a large amount of data about personal movement. 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.

Data

The data for this assignment can be downloaded from the course web site:

Dataset: Activity monitoring data [52K] The variables included in this dataset are:

steps: Number of steps taking in a 5-minute interval (missing values are coded as NA)

date: The date on which the measurement was taken in YYYY-MM-DD format

interval: Identifier for the 5-minute interval in which measurement was taken

The dataset is stored in a comma-separated-value (CSV) file and there are a total of 17,568 observations in this dataset.

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.

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

Load the data (i.e. read.csv())

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.1.2
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
```

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

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.1.2
```

```
library(lattice)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.1.2
```

```
data_Steps <- read.csv("./Data/activity.csv", header=T, sep=',', na.strings="?",
                       colClasses = c("numeric", "character", "numeric"),
                       check.names=F, stringsAsFactors=F, comment.char="",
                       quote='\"')

names(data_Steps)
```

```
## [1] "steps"    "date"     "interval"
```

```
head(data_Steps)
```

```
##   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
```

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

```
All_Steps <- aggregate(steps ~ date, data = data_Steps, sum, na.rm = TRUE)
```

What is mean total number of steps taken per day?

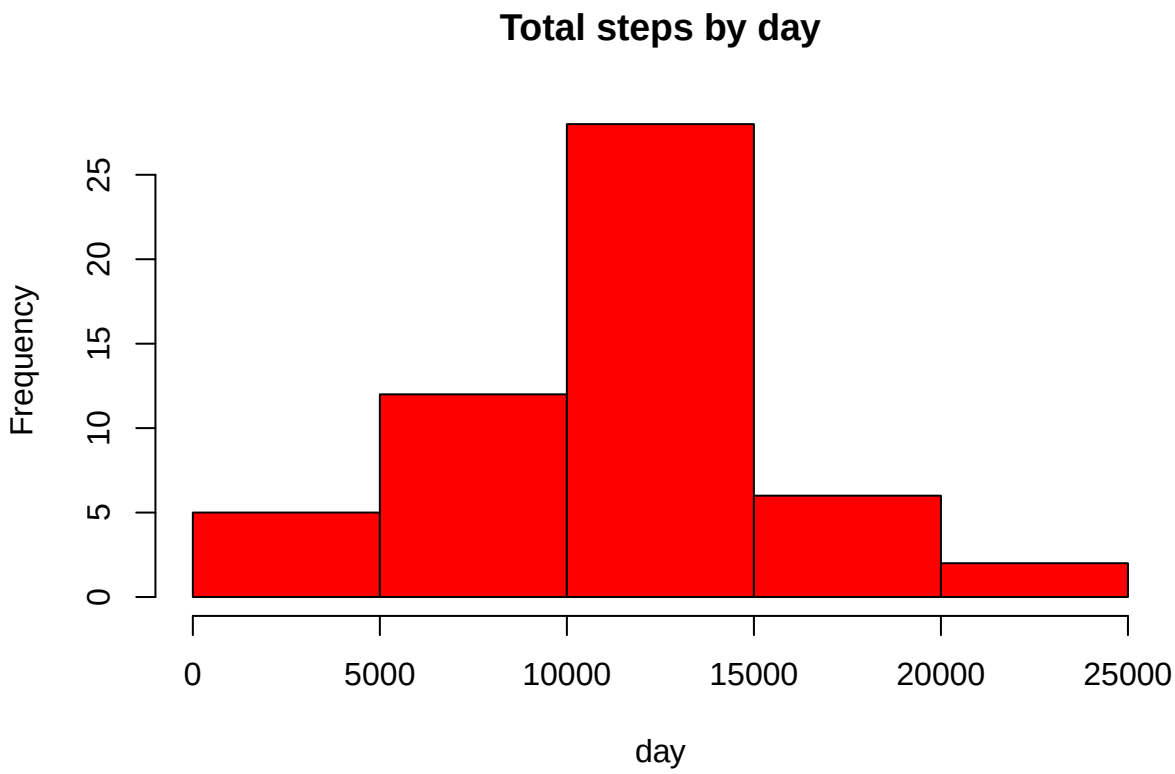
```
mean(All_Steps$steps, trim = 0, na.rm = TRUE)
```

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

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

Calculate the total number of steps taken per day

```
hist(All_Steps$steps, main = "Total steps by day", xlab = "day", col = "red")
```



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

```
mean(All_Steps$steps, trim = 0, na.rm = FALSE)
```

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

```
median(All_Steps$steps, na.rm = FALSE)
```

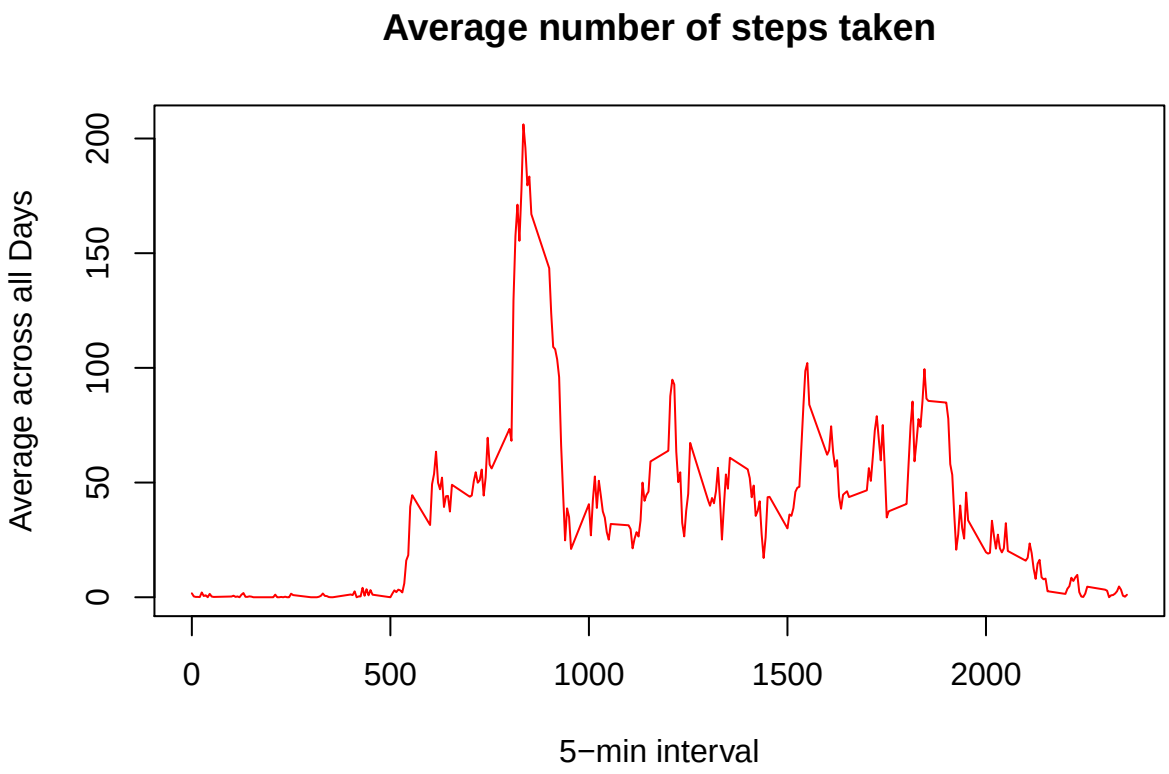
```
## [1] 10765
```

What is the average daily activity pattern?

```
time_series <- tapply(data_Steps$steps, data_Steps$interval, mean, na.rm = TRUE)
```

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)

```
plot(row.names(time_series), time_series, type = "l", xlab = "5-min interval",
    ylab = "Average across all Days", main = "Average number of steps taken",
    col = "red")
```



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

```
max_interval <- which.max(time_series)
names(max_interval)
```

```
## [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.

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

```
steps_NA <- sum(is.na(data_Steps$steps))
steps_NA
```

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

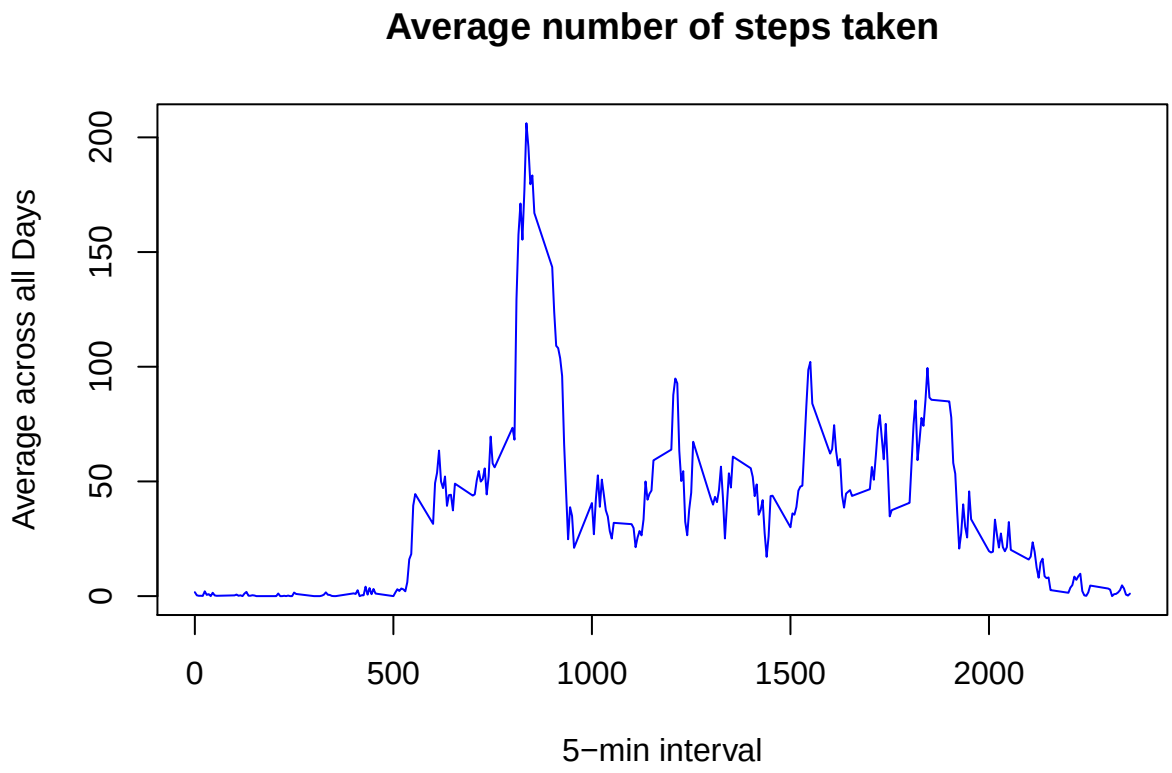
Devise a strategy for filling in all of the missing values in the dataset. The strategy does not need to be sophisticated. For example, you could use the mean/median for that day, or the mean for that 5-minute interval, etc.

```
data_Steps$steps <- ifelse(is.na(data_Steps$steps),mean(data_Steps$steps, trim = 0, na.rm = TRUE),data_S
```

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

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?

```
plot(row.names(time_series), time_series, type = "l", xlab = "5-min interval",
    ylab = "Average across all Days", main = "Average number of steps taken",
    col = "blue")
```



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.

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.

```
data_Steps$date <- as.POSIXlt(data_Steps$date)

data_Steps$day<-weekdays(data_Steps$date)

data_Steps$D_E <- ifelse((data_Steps$day=="Sunday" | data_Steps$day=="Saturday"),"Weekend", "Weekday")
```

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.

```
xyplot(steps ~ interval | D_E, data_Steps, type = "l", layout = c(1, 2),
    xlab = "Interval", ylab = "Number of steps")
```