PA1_template - Course Project 1

Greg Ricci

June 10, 2019

Peer-graded Assignment: Course Project 1

Background data information

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.

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 - note :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. Note that the echo = TRUE parameter was added to the code chunk to allow printing of the R code.

The Assigment - Questions and answers

Loading and check data prior to processing

```
# load the data
activity <- read.csv("activity.csv")
# check the data
head(activity)
```

```
##   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
```

```
str(activity)
```

```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : Factor w/ 61 levels "2012-10-01","2012-10-02",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
```

What is mean total number of steps taken per day? (missing values are ignored in the dataset.)

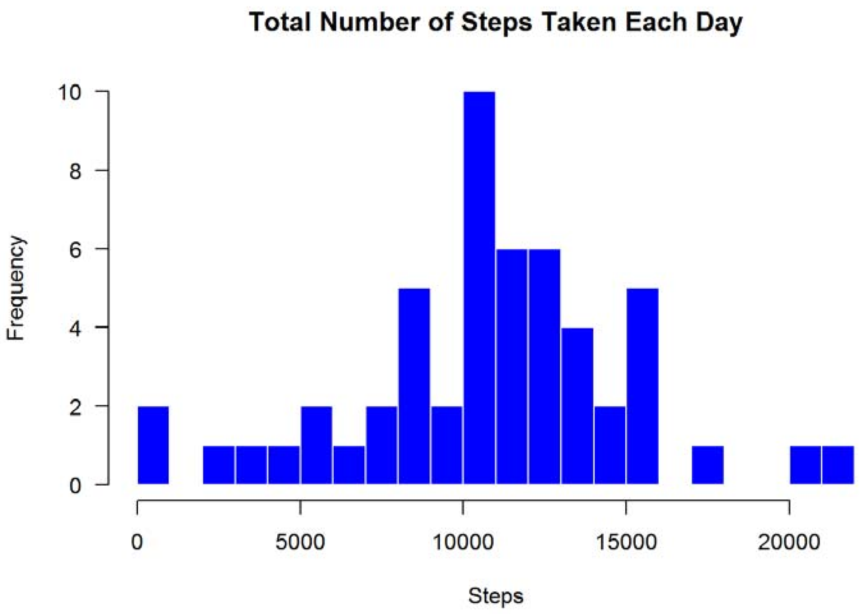
1. Aggregate the total number of steps taken per day

```
# aggragate steps per day
total_step_pd <- aggregate(steps ~ date, data = activity, sum, na.rm = TRUE)
head(total_step_pd)
```

```
##         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
```

2. Display a Histogram of the total number of steps taken each day

```
# Chart total steps
par(mfrow = c(1, 1))
hist(total_step_pd$steps, breaks = 20,
     main = "Total Number of Steps Taken Each Day",
     col = "blue", border = "white", xlab = "Steps", axes = FALSE)
axis(1)
axis(2, las = 1)
```



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

```
# Calculate mean and median
mean(total_step_pd$steps)
```

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

```
median(total_step_pd$steps)
```

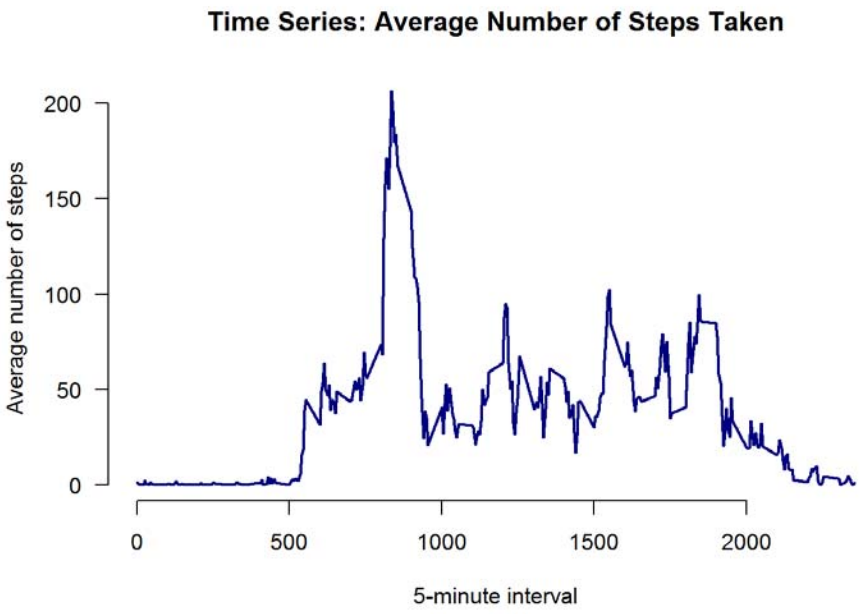
```
## [1] 10765
```

What is the average daily activity pattern?

1. Make a time series plot (i.e. type = “l”type=“l”) of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis).

```
# Calculate average steps
avg_step <- aggregate(steps ~ interval, data = activity, mean, na.rm = TRUE)
```

```
# Chart Time Series: Average Number of Steps Taken
plot(avg_step$interval, avg_step$steps, type = "l", lwd = 2, col = "navy",
     main = "Time Series: Average Number of Steps Taken", axes = FALSE,
     xlab = "5-minute interval", ylab = "Average number of steps")
axis(1)
axis(2, las = 1)
```



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

```
# Calculate which 5-minute intervalcontains the maximum number of steps
avg_step$interval[which.max(avg_step$steps)]
```

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

Imputing missing values

Note that there are a number of days/intervals where there are missing values (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)

```
# Calculate total number of missing values in the dataset
# or dim(activity[activity$steps == "NA", ])[1]
sum(is.na(activity))
```

```
## [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.

```
# Use the mean of 5-minute interval to fill in the values of the missing values.
```

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

```
# Create a new dataset no NAs

imp <- activity # new dataset called imp
for (i in avg_step$interval) {
    imp[imp$interval == i & is.na(imp$steps), ]$steps <-
        avg_step$steps[avg_step$interval == i]
}
head(imp)
```

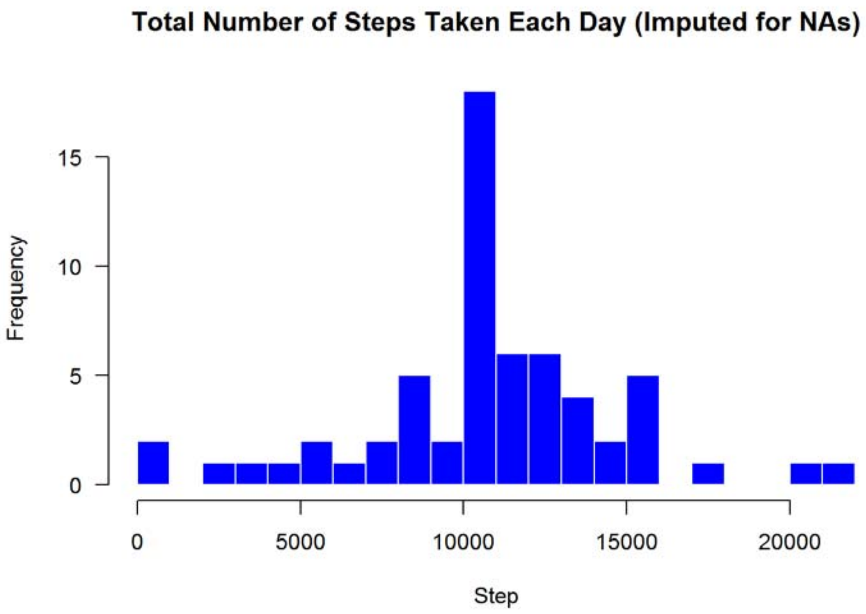
```
##       steps       date interval
## 1 1.7169811 2012-10-01        0
## 2 0.3396226 2012-10-01        5
## 3 0.1320755 2012-10-01       10
## 4 0.1509434 2012-10-01       15
## 5 0.0754717 2012-10-01       20
## 6 2.0943396 2012-10-01       25
```

```
sum(is.na(imp))
```

```
## [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?

```
# Aggegate the total number of steps taken each day
total_step_imp <- aggregate(steps ~ date, data = imp, sum, na.rm = TRUE)
## Create Chart
hist(total_step_imp$steps, breaks = 20,
     main = "Total Number of Steps Taken Each Day (Imputed for NAs)",
     col = "blue", border = "white", xlab = "Step", axes = FALSE)
axis(1)
axis(2, las = 1)
```



```
# report the mean and median total number of steps taken per day
mean(total_step_imp$steps)
```

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

```
median(total_step_imp$steps)
```

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

4. Second part = 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?

- The mean is the same as the mean from the first part of the assignment, but the median is not. When Imputing the missing data points the calculation is using the average of the 5-minute interval results in more data points which are equal to the mean and therefore there is a smaller variation of the distribution.

Are there differences in activity patterns between weekdays and weekends?

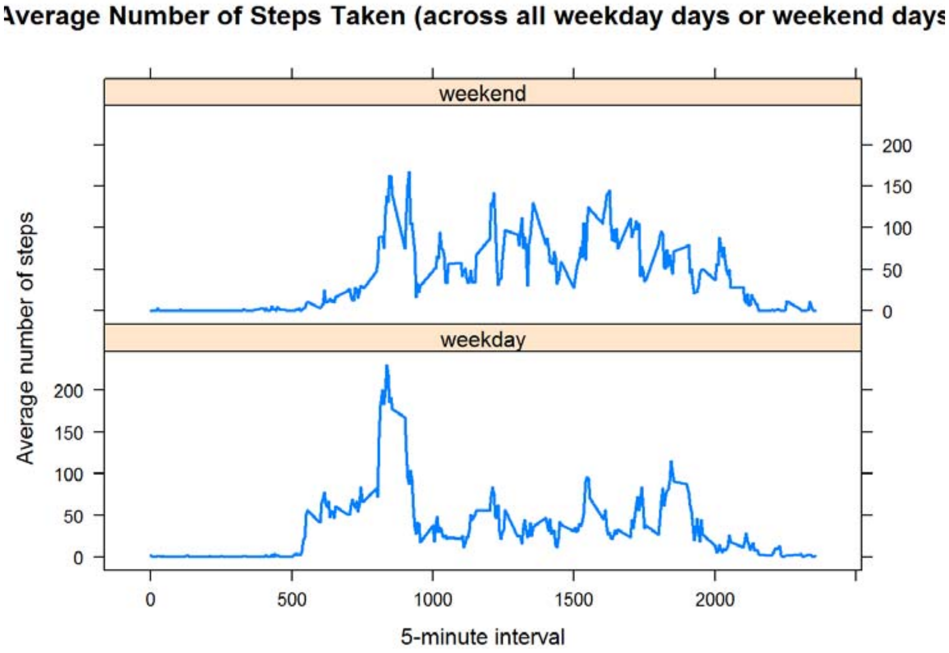
For this part the weekdays()weekdays() function may be of some help here. Use the dataset with the filled-in missing values for this part.

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.

```
# Create a new factor variable in the dataset with two levels weekday and weekend
# convert date to a date() class variable
imp$date <- as.Date(strptime(imp$date, format="%Y-%m-%d"))
imp$day <- weekdays(imp$date)
imp$week <- ""
imp[imp$day == "Saturday" | imp$day == "Sunday", ]$week <- "weekend"
imp[!(imp$day == "Saturday" | imp$day == "Sunday"), ]$week <- "weekday"
imp$week <- factor(imp$week)
```

2. Make a panel plot containing a time series plot (i.e. type = “l”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.

```
# panel plot containing a time series plot i.e. type = "l" of the 5-minute interval
avg_step_imp <- aggregate(steps ~ interval + week, data = imp, mean)
library(lattice)
xyplot(steps ~ interval | week, data = avg_step_imp, type = "l", lwd = 2,
       layout = c(1, 2),
       xlab = "5-minute interval",
       ylab = "Average number of steps",
       main = " Average Number of Steps Taken (across all weekday days or weekend days)")
```


Average Number of Steps Taken (across all weekday days or weekend days
weekend
weekday
Average number of steps
5-minute interval