Reproducible Research - Course Project 1

Wayne Chan

November 23, 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.

Data Source

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.

Loading and preprocessing the data

First, I will create a folder for this course storage and also treat as the working folder for R programming. Then, it download the zip file through the internet and store in the specified folder which created before. Finally, I unzip and read into R program with table named "activity".

```
# Create Project folder
if (!dir.exists("./reproducible-research")) dir.create("reproducible-re
search")
# Change working folder
if (dir.exists("reproducible-research")) setwd("./reproducible-research
")

# Download & upzip source file
if(!file.exists("activity.csv")) {
    ZipFile <- "./repdata.zip"
    if(!file.exists(ZipFile)) {
        ZipURL <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata
%2Factivity.zip"
        download.file(url=ZipURL,destfile=ZipFile,method="auto")
    }
    unzip(ZipFile)
    if (file.exists("activity.csv")) file.remove("./repdata.zip")
}

# Read csv into R
activity <- read.csv("./activity.csv", header=TRUE, sep=",")
```

Find out the dimension of table "activity".

```
dim(activity)

## [1] 17568     3
```

Check the column names oftable "activity".

```
names(activity)

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

Scan the head & tail records oftable "activity".

```
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

tail(activity)

##       steps       date interval
## 17563    NA 2012-11-30     2330
## 17564    NA 2012-11-30     2335
## 17565    NA 2012-11-30     2340
## 17566    NA 2012-11-30     2345
## 17567    NA 2012-11-30     2350
## 17568    NA 2012-11-30     2355
```

Display the Structure of table "activity".

```
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 ...
```

After previous information, we know that the table "activity" has 17568 records, each record has 3 variables that two integer variables "steps" and "interval" and one factor variable "date". There are many "na" values in variable "steps" which would be include in my further study.

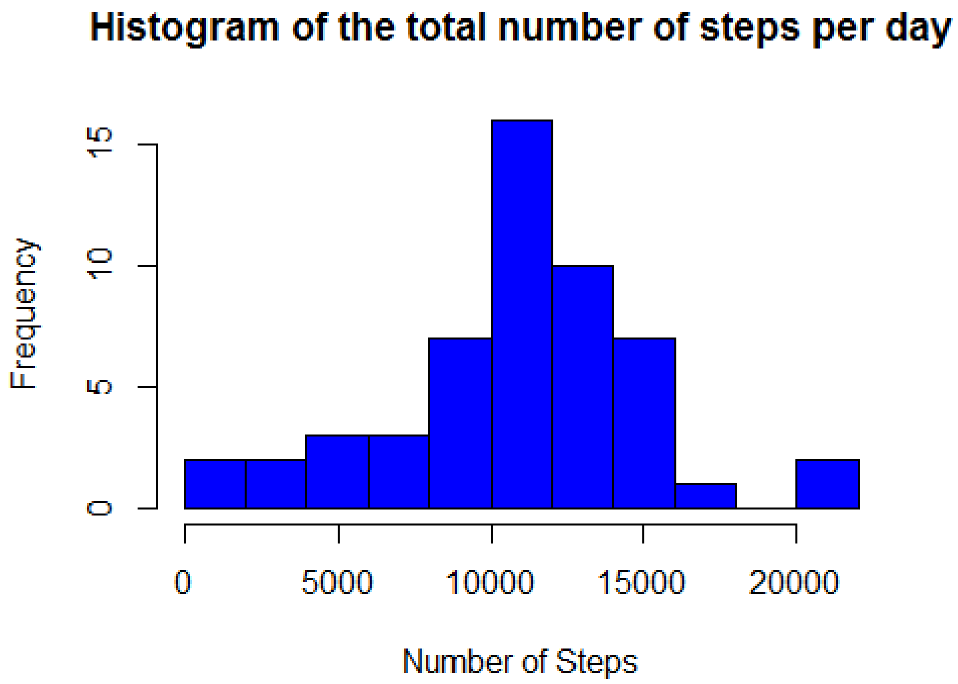
Remove those incomplete records with "na" values and format the variable "date" as date format;

```
activity$date <- as.Date(activity$date, "%Y-%m-%d")
```

What is mean total number of steps taken per day?

Plot the histogram of the total number of steps taken each day.

```
SumStepByDate <- aggregate(steps ~ date, data=activity, sum, na.action
= na.omit)

hist(SumStepByDate$steps, breaks=10, col="blue", xlab="Number of Steps",
 main= "Histogram of the total number of steps per day")
```



Find the mean value of the total number of steps taken each day.

```
MeanStep <- mean(SumStepByDate$steps, na.rm = TRUE)
print(MeanStep)

## [1] 10766.19
```

Find the median value of the total number of steps taken each day.

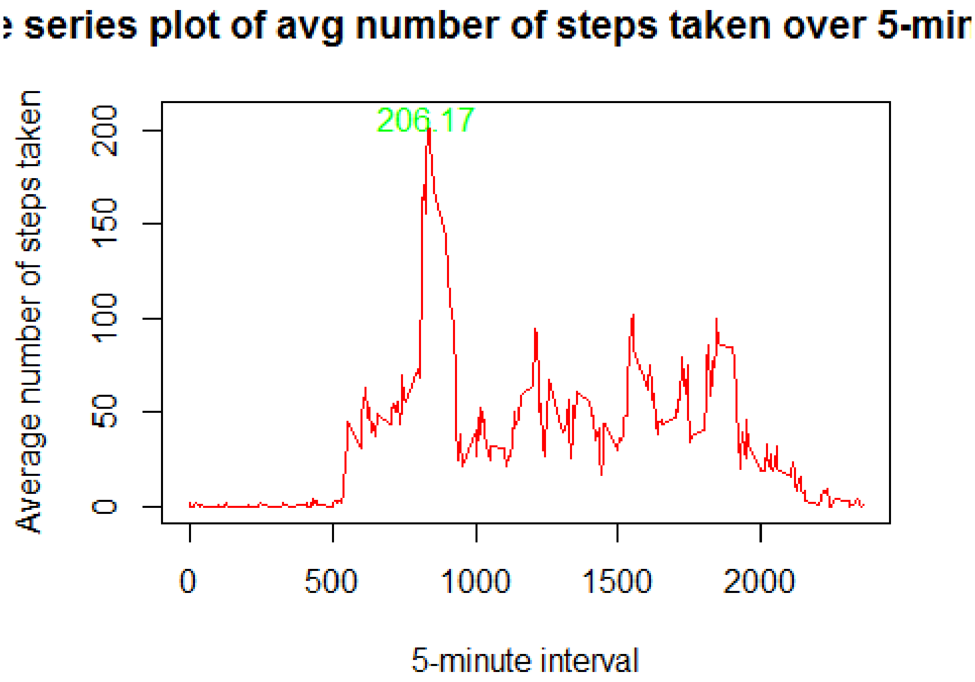
```
MedianStep <- median(SumStepByDate$steps, na.rm = TRUE)
print(MedianStep)

## [1] 10765
```

What is the average daily activity pattern?

Ploat a time series line chart forthe average number of steps taken

```
MeanByTime <- aggregate(activity$steps ~ interval, data=activity, mean)
names(MeanByTime) <- c("interval","MeanStep")

MaxPonit <- MeanByTime[which.max(MeanByTime$MeanStep),]

plot(MeanByTime$MeanStep ~ as.numeric(MeanByTime$interval), type="l", x
lab = "5-minute interval", ylab="Average number of steps taken", col="r
ed", main="Time series plot of avg number of steps taken over 5-min int
erval")
```

```
text(x=MaxPonit$interval, y=MaxPonit$MeanStep,labels=round(MaxPonit$Mea
nStep,2), pch=11, col="green")
```



Imputing missing values

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

## [1] 2304

sum(is.na(activity$date))

## [1] 0

sum(is.na(activity$interval))

## [1] 0
```

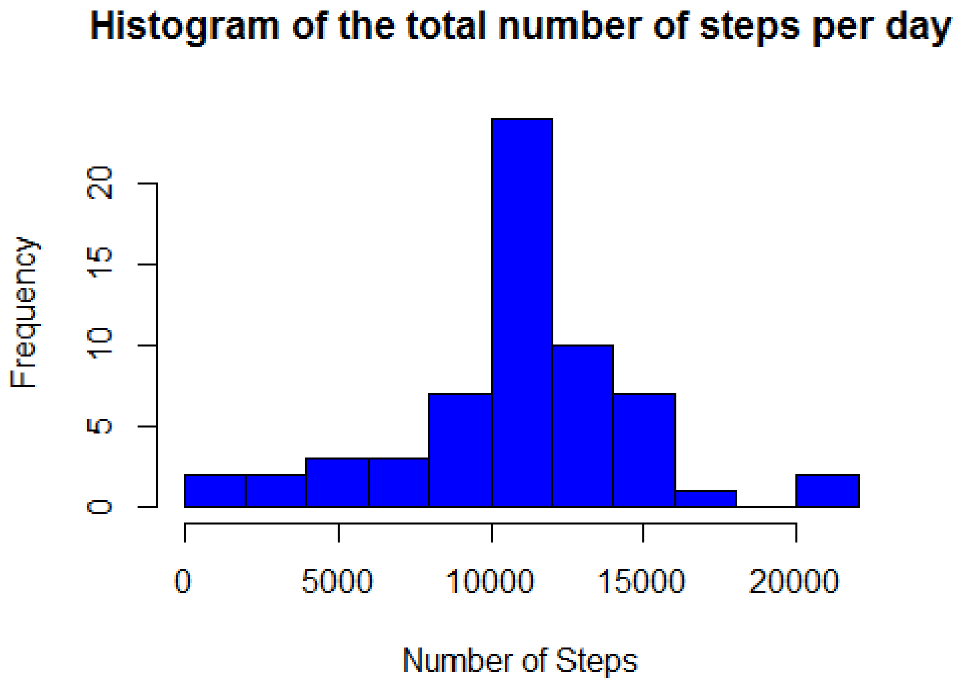
The total number of missing values in the dataset is 2304 and all missing values in variable "steps".

Using the average number of steps taken every 5-minute interval fill in the missing step values in the orignal dataset and save in a new dataset.

```
NewActivity <- activity

for (i in 1:nrow(NewActivity)) {
      if (is.na(NewActivity$steps[i])) {
            key <- NewActivity$interval[i]
            NewActivity$steps[i] <- as.integer(MeanByTime[which(MeanByT
ime$interval==key), 2])
      }
}

sum(is.na(NewActivity$steps))

## [1] 0
```

Plot the histogram of total steps taken by day.

```
NewSumStepByDate <- aggregate(steps ~ date, data=NewActivity, sum, na.a
ction = na.omit)

hist(NewSumStepByDate$steps, breaks=10, col="blue", xlab="Number of Ste
ps", main= "Histogram of the total number of steps per day")
```



Find the new mean value of the total number of steps taken each day.

```
NewMeanStep <- mean(NewSumStepByDate$steps, na.rm = TRUE)
print(NewMeanStep)

## [1] 10749.77
```

Find the new median value of the total number of steps taken each day.

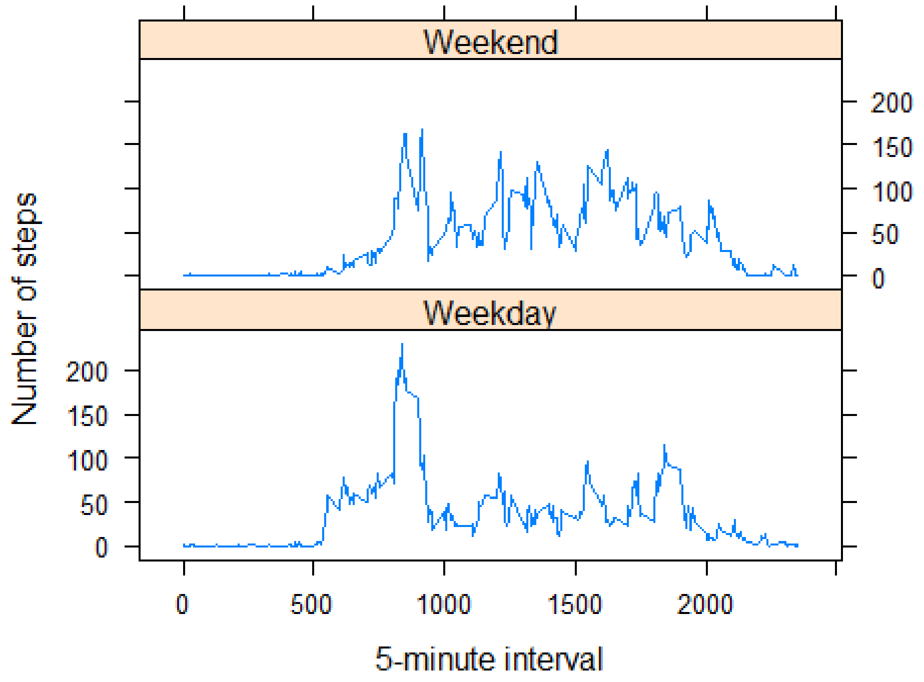
```
NewMedianStep <- median(NewSumStepByDate$steps, na.rm = TRUE)
print(NewMedianStep)
```

```
## [1] 10641
```

```
Weekend <- NewActivity[(weekdays(NewActivity$date) %in% c("Saturday", "Sunday")),]
Weekday <- NewActivity[!(weekdays(NewActivity$date) %in% c("Saturday", "Sunday")),]
```

Are there differences in activity patterns between weekdays and weekends?

```
WeekendMBT <- aggregate(Weekend$steps ~ interval, data=Weekend, mean)
names(WeekendMBT) <- c("interval","MeanStep")
WeekendMBT$Day <- "Weekend"
WeekdayMBT <- aggregate(Weekday$steps ~ interval, data=Weekday, mean)
names(WeekdayMBT) <- c("interval","MeanStep")
WeekdayMBT$Day <- "Weekday"

Both_MBT <- rbind(WeekendMBT, WeekdayMBT)

library(lattice)
xyplot(MeanStep ~ interval | Day, data=Both_MBT, layout=c(1, 2), type="l", xlab="5-minute interval", ylab="Number of steps")
```



It can be observed that the average number of steps taken in weekday was more fluctuated than weeked.