Reproducible Research - Project 1

Sabyasachi Maiti

October 23, 2017

Week 2 Assignment

Introduction

This assignment makes use of data from a personal activity monitoring device. This device collects data at 5 minute intervals throughout 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.

Source Data Url: https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip (https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip)

Source Data downloaded: 10/23/2017

This document presents the results from Project Assignment 1 in the Coursera course Reproducible Research, written in a single R markdown document that can be processed by knitr and transformed into an HTML file.

Preprocessing

Loading required libraries.

```
library(sqldf)
```

```
## Warning: package 'sqldf' was built under R version 3.4.2
```

```
## Loading required package: gsubfn
```

```
## Warning: package 'gsubfn' was built under R version 3.4.2
```

```
## Loading required package: proto
```

```
## Warning: package 'proto' was built under R version 3.4.2
```

```
## Loading required package: RSQLite
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

Loading data from source

Code for reading the dataset and converting the date column to proper date format.

```
activity <- read.csv("activity.csv", header = TRUE, sep=",")
activity[,2]<-as.Date(activity$date)
str(activity)
```

```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : Date, format: "2012-10-01" "2012-10-01" ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
```

Plotting total number of steps taken each day

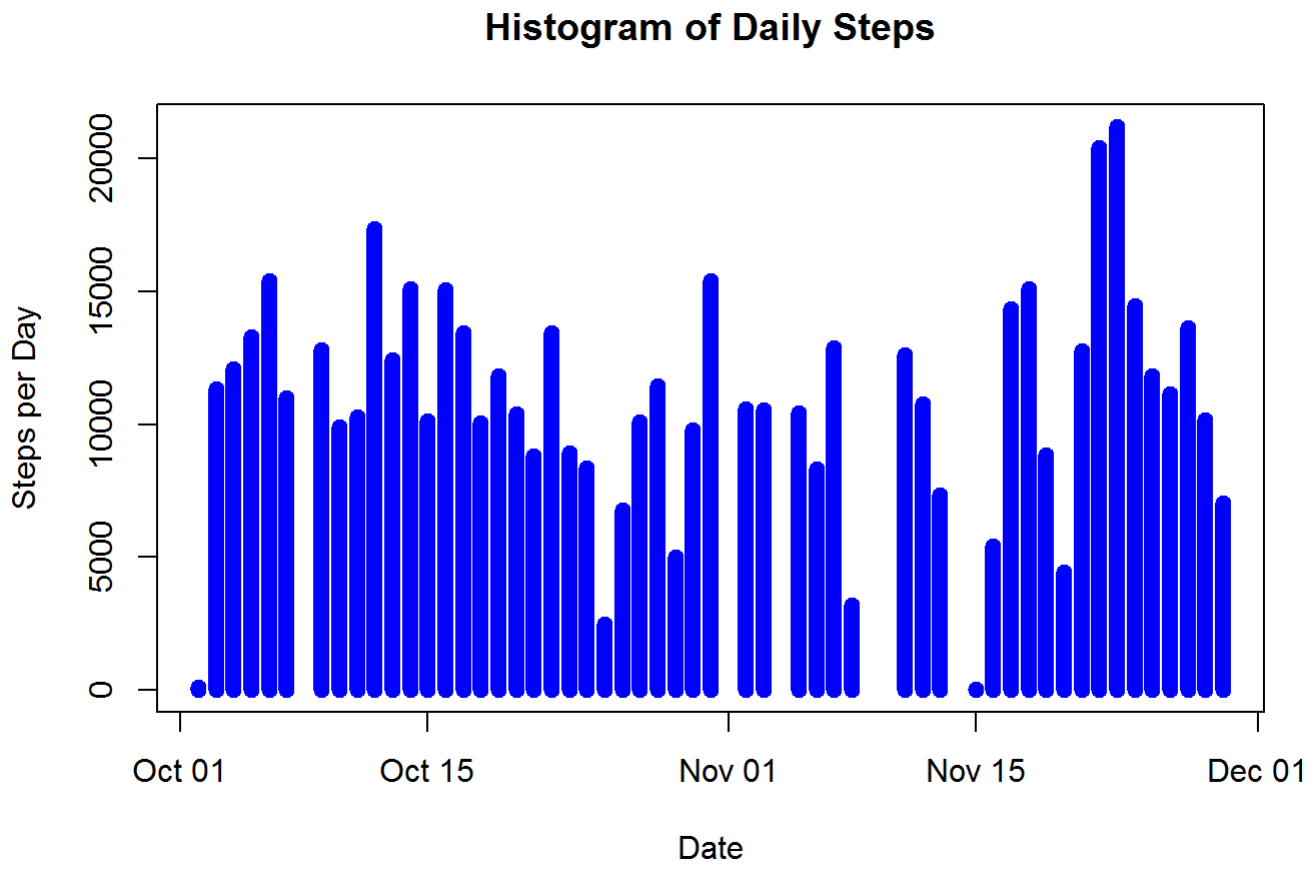
Deriving a data frame with total steps taken on each date, ignoring the rows where steps = NA.

```
StepsPerDay <- sqldf("select date, sum(steps) steps from activity where steps <> 'NA' group by d
ate")
str(StepsPerDay)
```

```
## 'data.frame':    53 obs. of  2 variables:
##  $ date : Date, format: "2012-10-02" "2012-10-03" ...
##  $ steps: int  126 11352 12116 13294 15420 11015 12811 9900 10304 17382 ...
```

Histogram of the total number of steps taken each day.

```
with(StepsPerDay,plot(date,steps,type="h", main="Histogram of Daily Steps", xlab="Date", ylab="S
teps per Day", col="blue", lwd=8))
```



Mean and median number of steps taken each day.

```
paste("Mean Steps per Day =", mean(StepsPerDay$steps, na.rm=TRUE))
```

```
## [1] "Mean Steps per Day = 10766.1886792453"
```

```
paste("Median Steps per Day =", median(StepsPerDay$steps, na.rm=TRUE))
```

```
## [1] "Median Steps per Day = 10765"
```

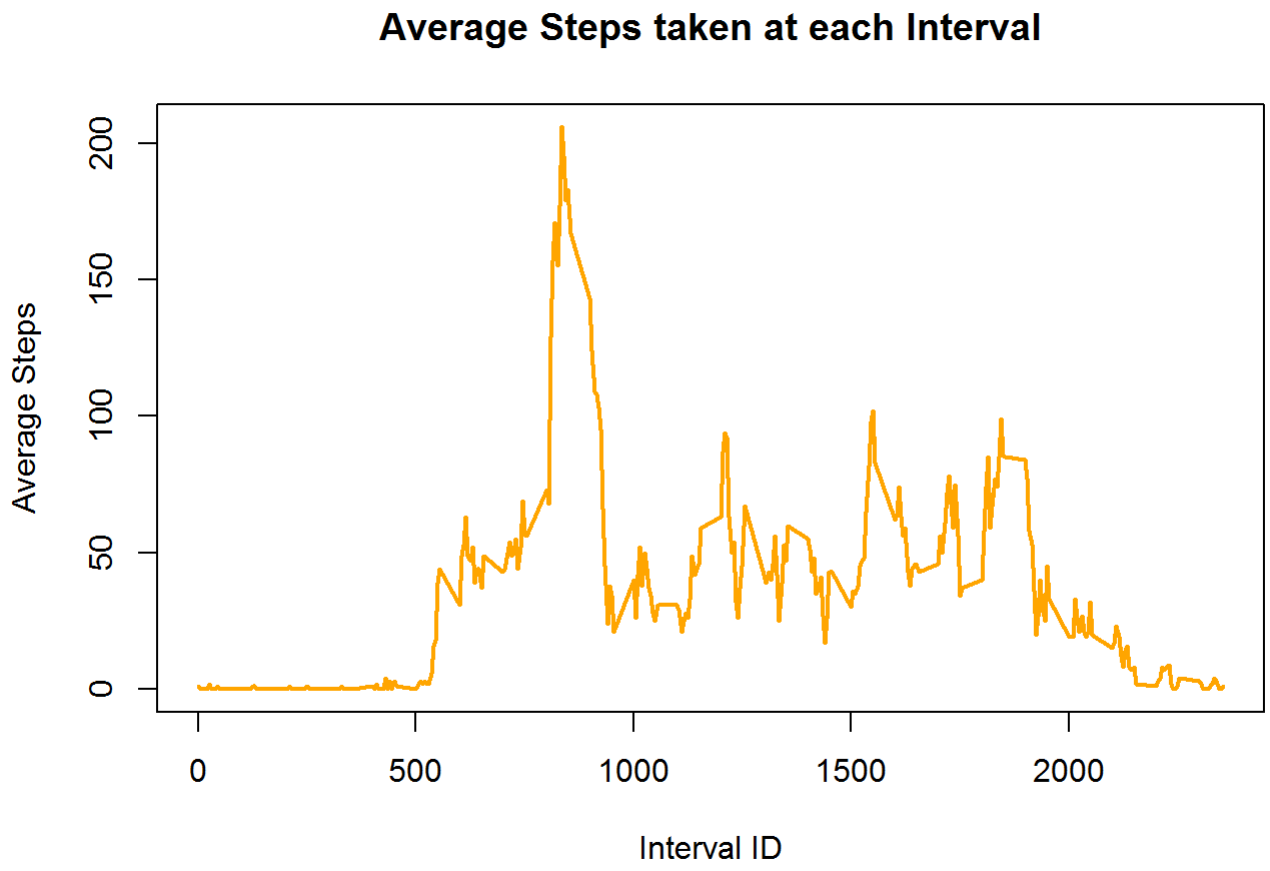
Plotting the average number of steps taken

Deriving a data frame with average steps taken per 5 minute interval, ignoring the rows where steps = NA.

```
StepsPerInt <- sqldf("select interval, avg(steps) steps from activity where steps <> 'NA' group
 by interval")
```

Time series plot of the average number of steps taken per 5 minute interval.

```
with(StepsPerInt,plot(interval, steps, type="l", main="Average Steps taken at each Interval", xl
ab="Interval ID", ylab="Average Steps", col="orange", lwd=2))
```



The 5-minute interval that, on average, contains the maximum number of steps.

```
paste("Interval with max value =", StepsPerInt$interval[which(StepsPerInt$steps == max(StepsPerI
nt$steps))])
```

```
## [1] "Interval with max value = 835"
```

Replacing missing values

Checking number of rows in activity data set with NA rows.

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

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

The missing steps will be replaced by the average steps for the particular interval number.

Copy activity data to a new dataset actNoNA.

```
actNoNA <- activity
```

Create new dataset actWithAvg by merging actNoNA with StepsPerInt using interval column, thereby adding the average steps per interval as a new column steps.aspi.

```
actWithAvg = merge(actNoNA, StepsPerInt, by="interval", suffixes=c(".act", ".aspi"))
```

Identify the NA rows in actNoNA.

```
naIndex = which(is.na(actNoNA$steps))
```

Replace the NA rows in actNoNA with average steps from actWithAvg.

```
actNoNA[naIndex,"steps"] = actWithAvg[naIndex,"steps.aspi"]
```

Checking that there are no missing values.

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

```
## [1] 0
```

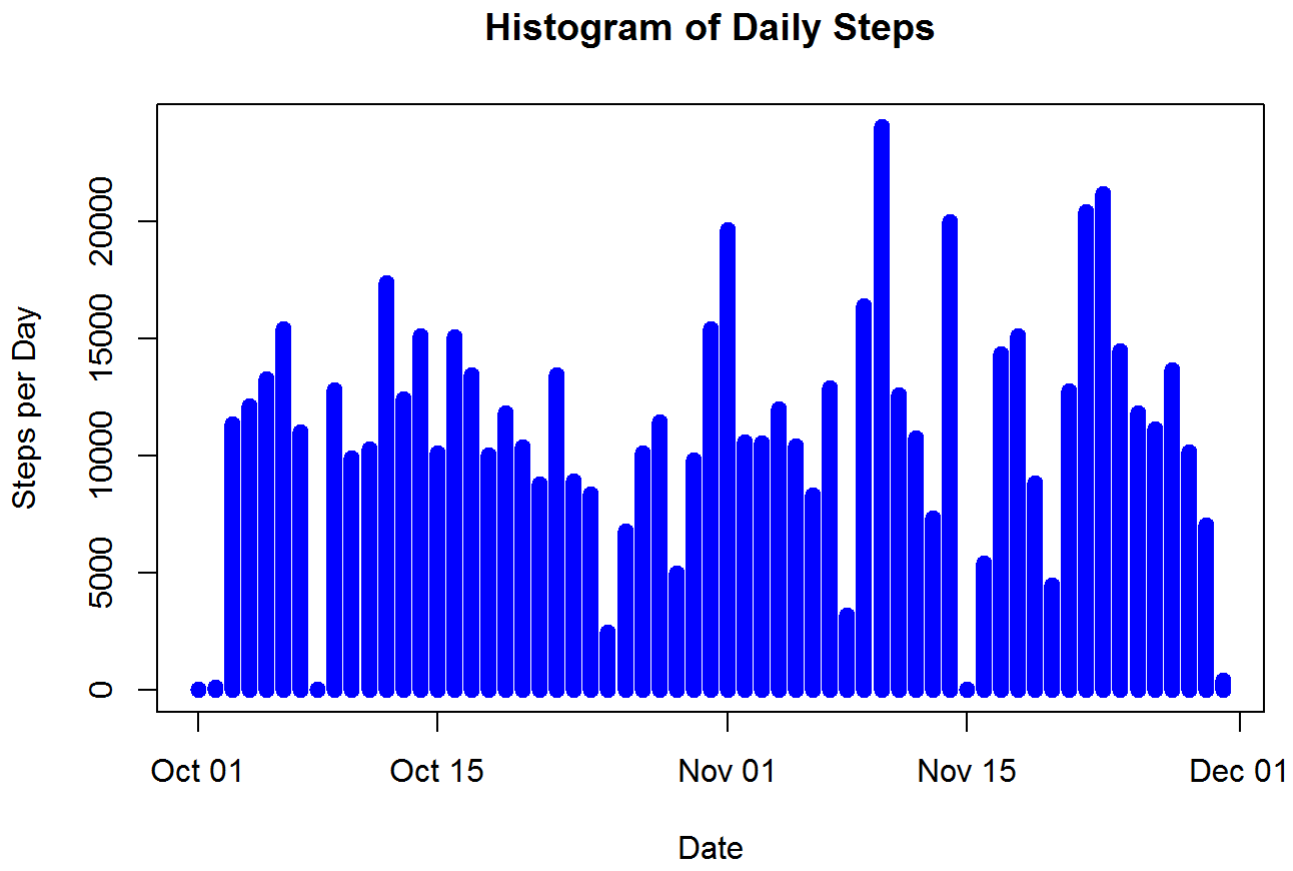
Plotting total number of steps taken each day after missing values are imputed

Deriving a data frame with total steps taken on each date using the updated dataset actNoNA

```
UpdStepsPerDay <- sqldf("select date, sum(steps) steps from actNoNA group by date")
```

Histogram of the total number of steps taken each day.

```
with(UpdStepsPerDay,plot(date,steps,type="h", main="Histogram of Daily Steps", xlab="Date",
ylab="Steps per Day", col="blue", lwd=8))
```



Updated mean and median number of steps taken each day.

```
paste("Mean Steps per Day =", mean(UpdStepsPerDay$steps, na.rm=TRUE))
```

```
## [1] "Mean Steps per Day = 10871.9836065574"
```

```
paste("Median Steps per Day =", median(UpdStepsPerDay$steps, na.rm=TRUE))
```

```
## [1] "Median Steps per Day = 11015"
```

Compare the average number of steps taken per 5-minute interval across weekdays and weekends

Add a new column to the actNoNA dataset to identify whether a day is a weekday or weekend. Then convert the column to factor

```
act_mod <- mutate(actNoNA, weektype = ifelse(weekdays(actNoNA$date) == "Saturday" | weekdays(act
NoNA$date) == "Sunday", "weekend", "weekday"))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.2
```

```
act_mod$weektype <- as.factor(act_mod$weektype)
str(act_mod)
```

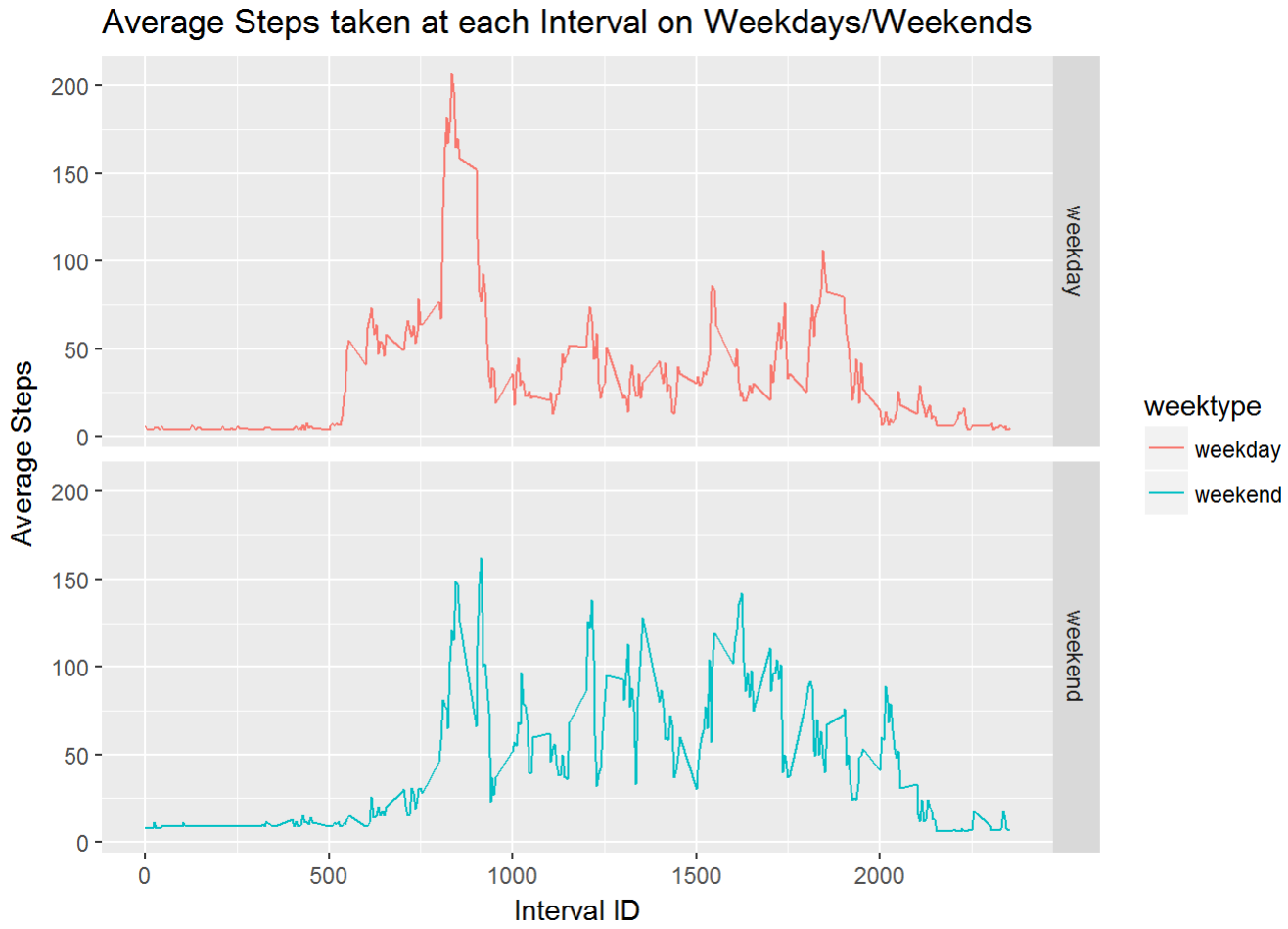
```
## 'data.frame':    17568 obs. of  4 variables:
##  $ steps   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ date    : Date, format: "2012-10-01" "2012-10-01" ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
##  $ weektype: Factor w/ 2 levels "weekday","weekend": 1 1 1 1 1 1 1 1 1 1 ...
```

Deriving a data frame with average steps taken per 5 minute interval by weektype

```
StepsPerIntWeek <- sqldf("select interval, weektype, avg(steps) steps from act_mod group by inte
rval, weektype")
```

Panel plot of the weekday and weekend data

```
qplot(interval,steps,data = StepsPerIntWeek, geom = "line", facets = weektype~., color=weektype,
 main="Average Steps taken at each Interval on Weekdays/Weekends", xlab="Interval ID", ylab="Ave
rage Steps")
```



From the two plots it seems that the test subject is more active throughout the weekends compared to weekdays.