Course Project 1

Lingna Chai

July 10, 2016

Overview

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

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

Dataset: Activity monitoring data https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip (https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip)

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.

Questions

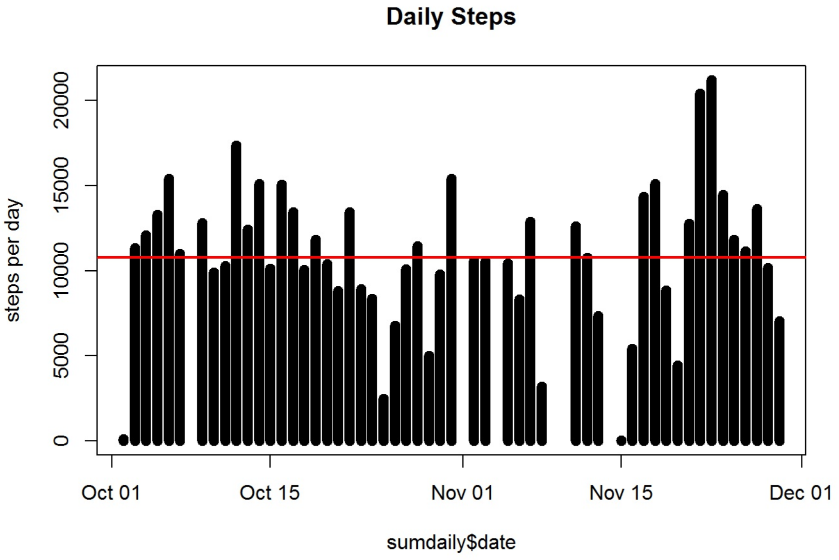
Reading in the dataset and/or processing the data

```
dataset <- read.csv("activity.csv");
dataset$day<- weekdays(as.Date(dataset$date))
dataset$date <- as.Date(dataset$date)
str(dataset)
```

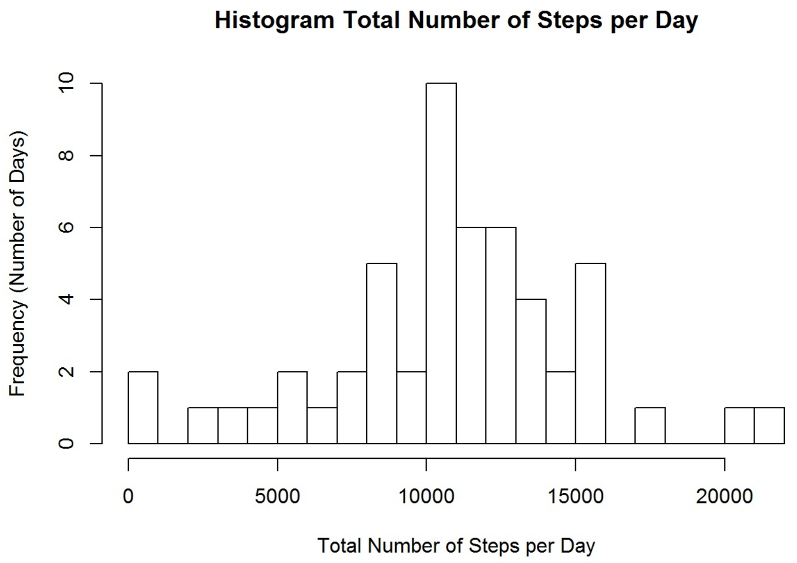
```
## 'data.frame':    17568 obs. of  4 variables:
##  $ steps   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : Date, format: "2012-10-01" "2012-10-01" ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
##  $ day     : chr  "Monday" "Monday" "Monday" "Monday" ...
```

Histogram of the total number of steps taken each day

```
sumdaily <- aggregate(dataset$steps ~ dataset$date,FUN=sum)
colnames(sumdaily)<-c("date","steps")
plot(sumdaily$date,sumdaily$steps,type="h", main= "Daily Steps",ylab ="steps per day",lwd=8)
abline(h=mean(sumdaily$steps,na.rm=TRUE),col="red",lwd=2)
```



```
hist(sumdaily$steps,
     main = "Histogram Total Number of Steps per Day",
     xlab = "Total Number of Steps per Day",
     ylab = "Frequency (Number of Days)",
     breaks=20)
```



Mean and median number of steps taken each day

```
median(sumdaily$steps,na.rm=TRUE)
```

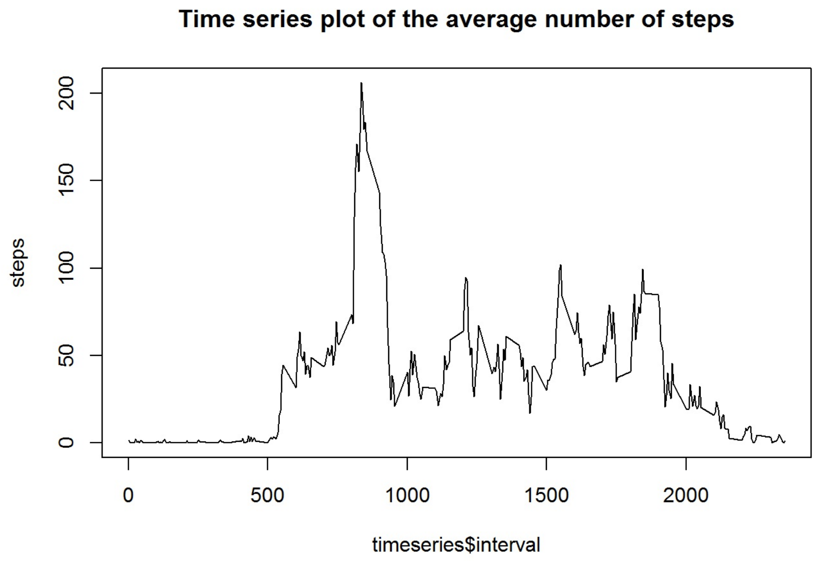
```
## [1] 10765
```

```
mean(sumdaily$steps,na.rm=TRUE)
```

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

Time series plot of the average number of steps taken

```
library(ggplot2)
library(plyr)
narmdataset <- dataset[!is.na(dataset$steps),]
timeseries <- ddply(narmdataset, .(interval),summarise, avg.steps = mean(steps))
plot(timeseries$interval,timeseries$avg.steps,type ="l",main= "Time series plot of the average number of steps"
,ylab ="steps")
```



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

```
max <- narmdataset[narmdataset$steps==max(narmdataset$steps),3]
```

Code to describe and show a strategy for imputing missing data

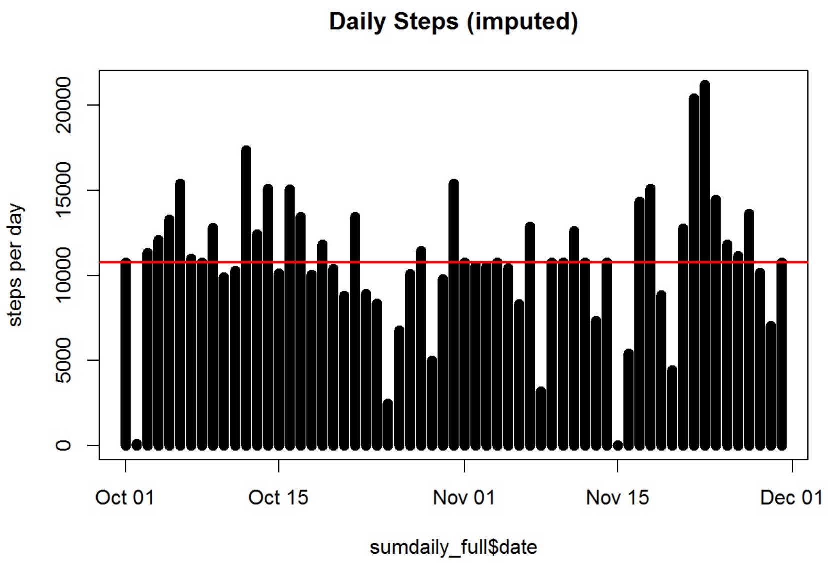
```
avgday<- ddply(narmdataset,.(interval,day),summarize, Avg = mean(steps))
```

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

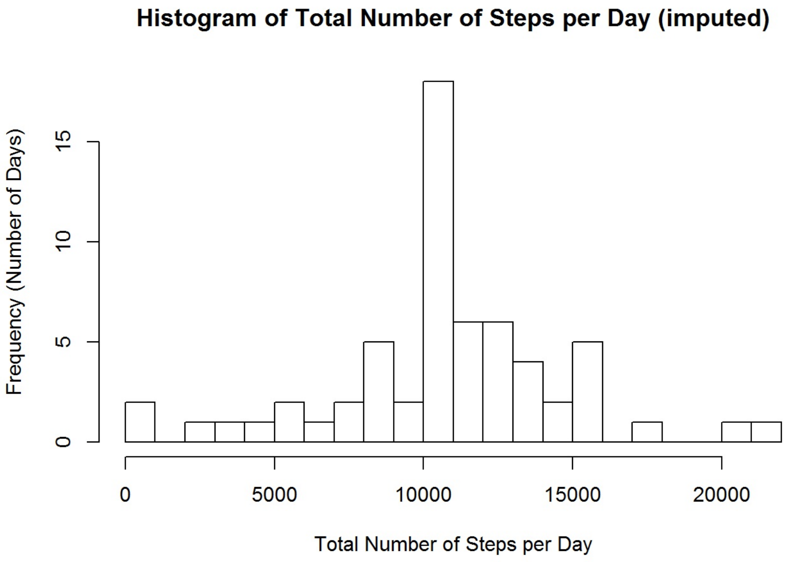
Missing values are imputed using the average steps taken in a identified 5 minutes. After imputing the missing data, the general distribution of the histogram seems unchanged but the distribution is smoother than non-imputed data.

```
library(dplyr)
data.full<- inner_join (dataset,timeseries,by="interval") %>%
            mutate(steps=ifelse(is.na(steps),avg.steps,steps))%>%
            select(date,interval,steps)

sumdaily_full <- aggregate(data.full$steps ~ data.full$date,FUN=sum)
colnames(sumdaily_full)<-c("date","steps")
plot(sumdaily_full$date,sumdaily_full$steps,type="h", main= "Daily Steps (imputed)",ylab ="steps per day",lwd=8
)
abline(h=mean(sumdaily_full$steps,na.rm=TRUE),col="red",lwd=2)
```



```
hist(sumdaily_full$steps,
     main = "Histogram of Total Number of Steps per Day (imputed)",
     xlab = "Total Number of Steps per Day",
     ylab = "Frequency (Number of Days)",
     breaks=20)
```



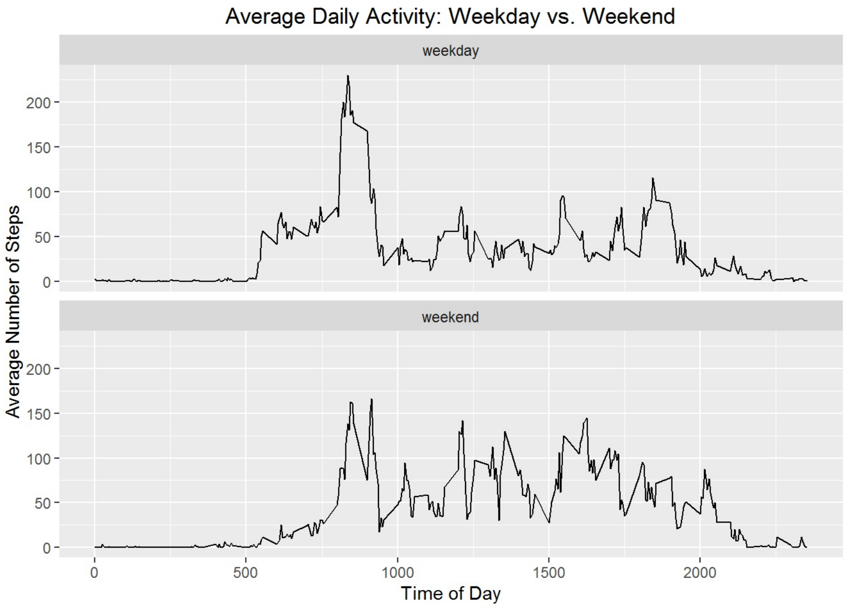
Panel plot comparing the average number of steps taken per 5-minute interval across weekdays and weekends

On average, people tend to perform more activities through out the day compared to a weekday.

```
data.full$weekday <-weekdays(data.full$date)
data.full$is.weekend <- as.factor(ifelse(data.full$weekday %in% c("Saturday","Sunday"), "weekend","weekday"))

activity.pattern <- data.full %>%
                    group_by(is.weekend,interval) %>%
                    summarize(avg.steps=mean(steps))

ggplot(activity.pattern, aes(interval, avg.steps)) +
    geom_line() +
    facet_wrap(~ is.weekend, ncol=1) +
    xlab("Time of Day") +
    scale_y_continuous("Average Number of Steps") +
    ggtitle("Average Daily Activity: Weekday vs. Weekend")
```


Average Daily Activity: Weekday vs. Weekend
weekday
weekend
Average Number of Steps
0
50
100
150
200
0
50
100
150
200
0
500
1000
1500
2000
Time of Day