Reproducible Research: Project 1

Esteban Castillo

December 16, 2017

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.

Steps

1. Load the necessary packages to complete the assignment

```
library(ggplot2)
library(scales)
library(Hmisc)
```

2. Load the activity dataset for the assignment

```
if(!file.exists('activity.csv')){
    unzip('repdata%2Fdata%2Factivity.zip')
}
activityData <- read.csv('activity.csv')
```

3. Pre-Process the data into a correct format for the analysis

```
activityData$date<- as.Date(activityData$date)
```

4. Calculate the total number of steps taken per day

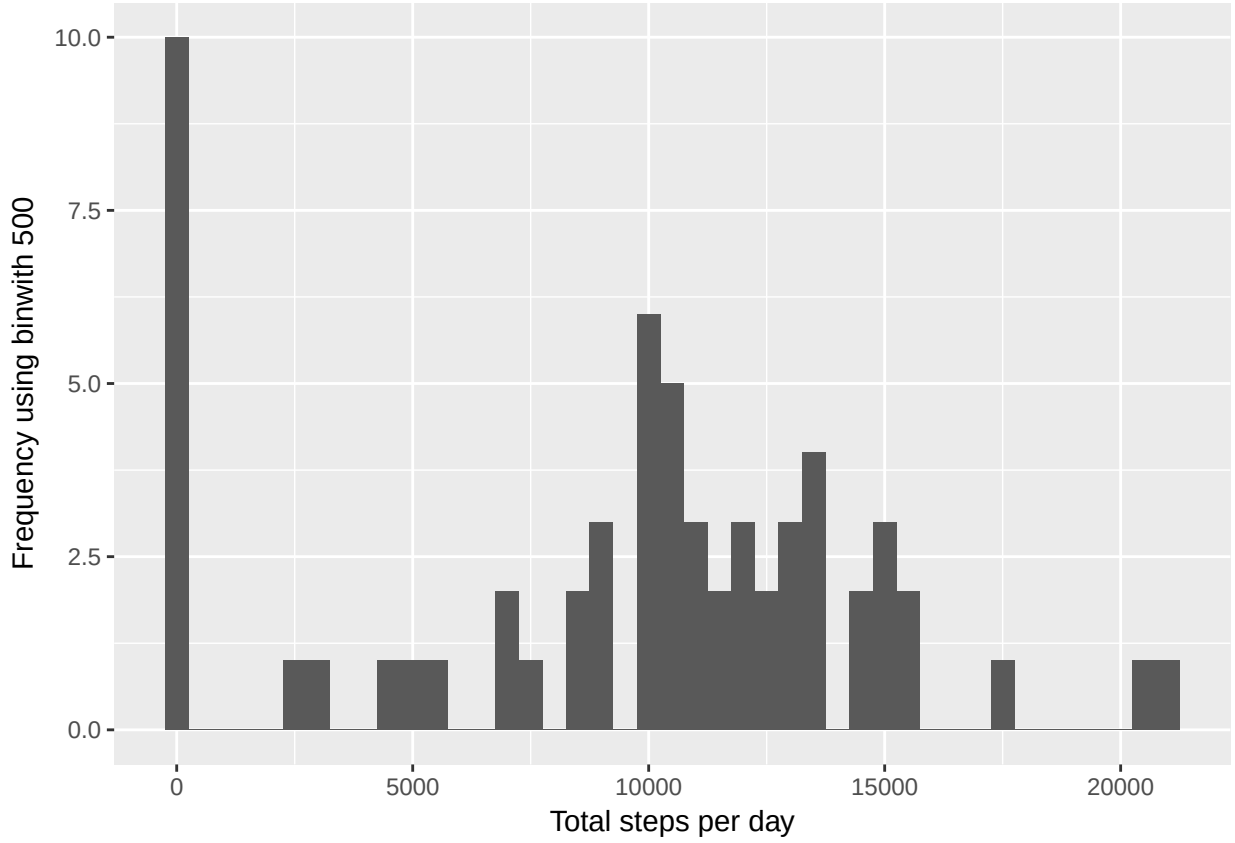
```
stepsByDay <- tapply(activityData$steps, activityData$date, sum, na.rm=TRUE)
stepsByDay
```

```
## 2012-10-01 2012-10-02 2012-10-03 2012-10-04 2012-10-05 2012-10-06
##          0        126      11352      12116      13294      15420
## 2012-10-07 2012-10-08 2012-10-09 2012-10-10 2012-10-11 2012-10-12
##      11015          0      12811       9900      10304      17382
## 2012-10-13 2012-10-14 2012-10-15 2012-10-16 2012-10-17 2012-10-18
##      12426      15098      10139      15084      13452      10056
## 2012-10-19 2012-10-20 2012-10-21 2012-10-22 2012-10-23 2012-10-24
##      11829      10395       8821      13460       8918       8355
## 2012-10-25 2012-10-26 2012-10-27 2012-10-28 2012-10-29 2012-10-30
##       2492       6778      10119      11458       5018       9819
## 2012-10-31 2012-11-01 2012-11-02 2012-11-03 2012-11-04 2012-11-05
```

```
##      15414          0      10600      10571          0      10439
## 2012-11-06 2012-11-07 2012-11-08 2012-11-09 2012-11-10 2012-11-11
##       8334      12883       3219          0          0      12608
## 2012-11-12 2012-11-13 2012-11-14 2012-11-15 2012-11-16 2012-11-17
##      10765       7336          0         41       5441      14339
## 2012-11-18 2012-11-19 2012-11-20 2012-11-21 2012-11-22 2012-11-23
##      15110       8841       4472      12787      20427      21194
## 2012-11-24 2012-11-25 2012-11-26 2012-11-27 2012-11-28 2012-11-29
##      14478      11834      11162      13646      10183       7047
## 2012-11-30
##          0
```

5. Make a histogram of the total number of steps taken each day

```
qplot(stepsByDay,
      xlab='Total steps per day',
      ylab='Frequency using binwith 500',
      binwidth=500)
```



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

```
stepsByDayMean <- mean(stepsByDay)
stepsByDayMean
```

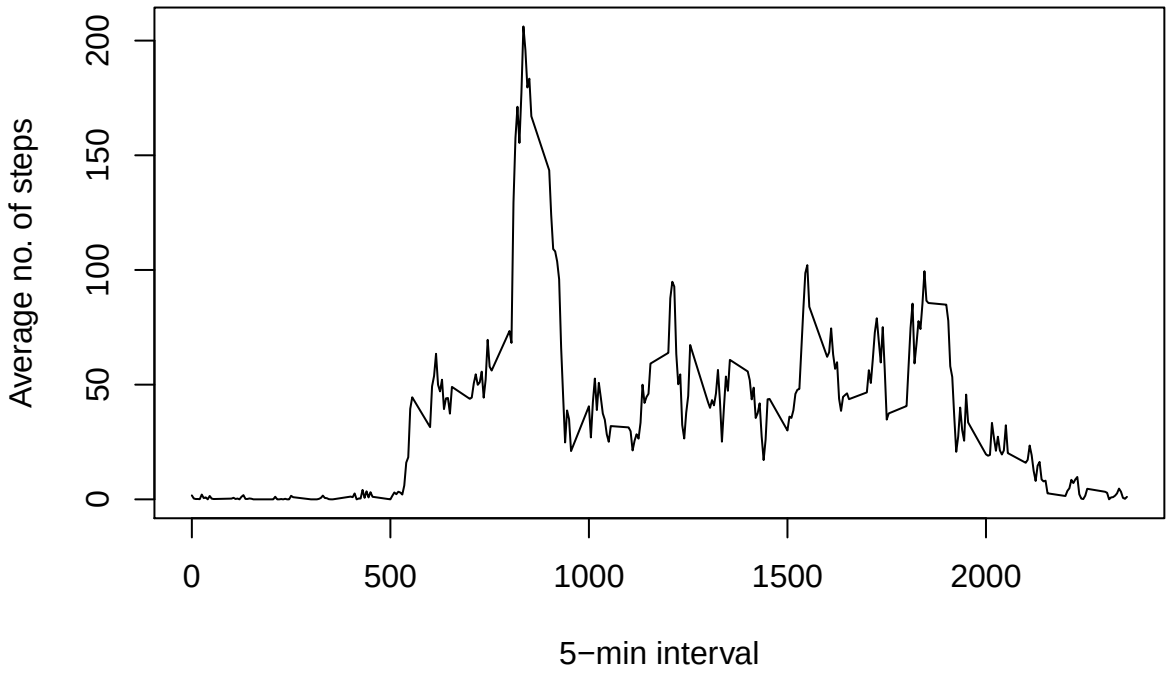
```
## [1] 9354.23
```

```
stepsByDayMedian <- median(stepsByDay)
stepsByDayMedian
```

```
## [1] 10395
```

7. What is the average daily activity pattern?

```
avg<- tapply(activityData$steps, activityData$interval, mean, na.rm=TRUE)
plot(names(avg), avg, xlab="5-min interval", type="l", ylab="Average no. of steps")
```




```
maxavg<- max(avg)
maxavg #5-minute interval
```

```
## [1] 206.1698
```

```
maxinterval<- as.numeric(names(avg)[which(avg==max(avg))])
maxinterval # Max Average Value
```

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

8. Imputing missing values

- Calculate and report the total number of missing values in the dataset

```
numMissingValues <- length(which(is.na(activityData$steps)))
numMissingValues #Number of missing values
```

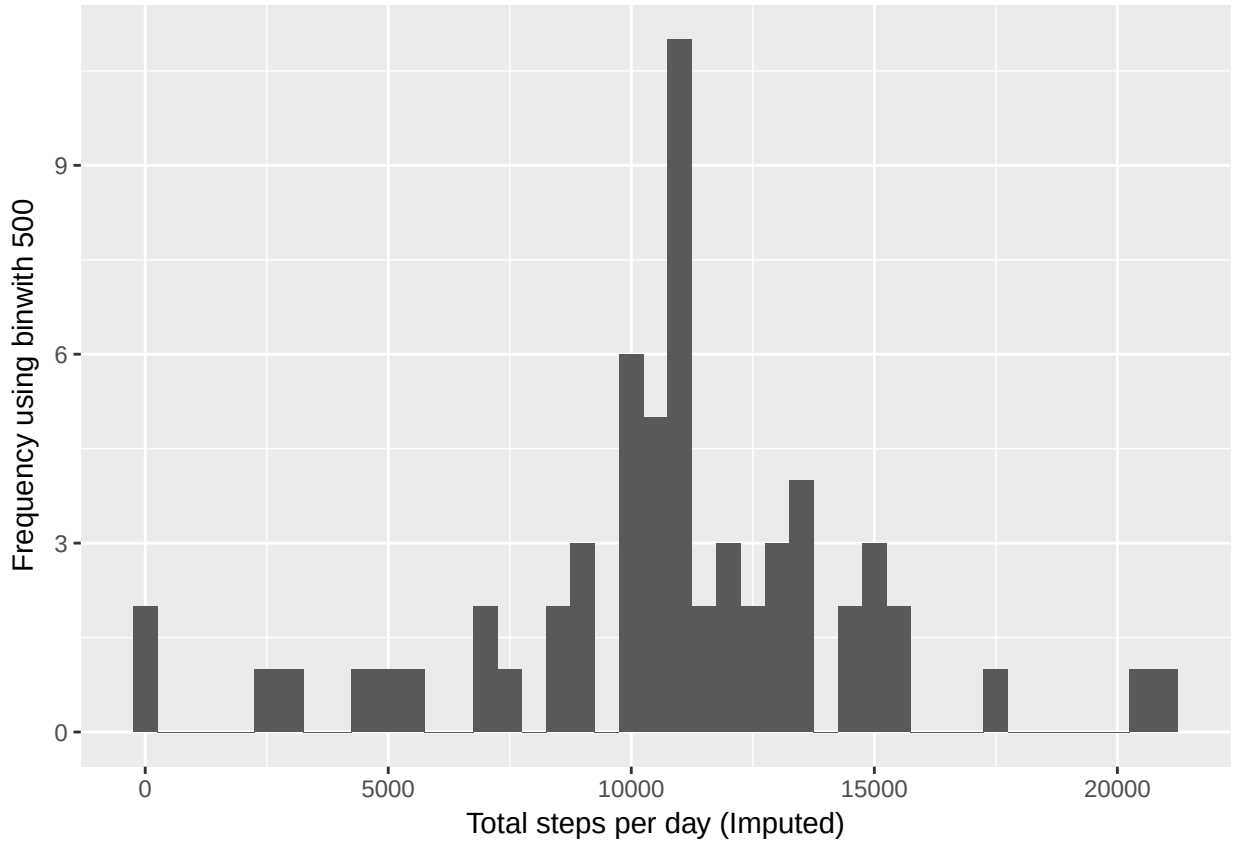
```
## [1] 2304
```

- Devise a strategy for filling in all of the missing values in the dataset: Create a new dataset that is equal to the original dataset but with the missing data filled i

```
activityDataImputed <- activityData
activityDataImputed$steps <- impute(activityData$steps, fun=mean)
```

- Make a histogram of the total number of steps taken each day

```
stepsByDayImputed <- tapply(activityDataImputed$steps, activityDataImputed$date, sum)
qplot(stepsByDayImputed, xlab='Total steps per day (Imputed)',
      ylab='Frequency using binwith 500', binwidth=500)
```



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

```
stepsByDayMeanImputed <- mean(stepsByDayImputed)
stepsByDayMeanImputed
```

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

```
stepsByDayMedianImputed <- median(stepsByDayImputed)
stepsByDayMedianImputed
```

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

9. Are there differences in activity patterns between weekdays and weekends?

- 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.

```
activityDataImputed$dateType <-  ifelse(as.POSIXlt(activityDataImputed$date)$wday %in% c(0,6),
                                        'weekend', 'weekday')
```

- Make a panel plot containing a time series plot

```
averagedActivityDataImputed <- aggregate(steps ~ interval + dateType,
                                         data=activityDataImputed, mean)
ggplot(averagedActivityDataImputed, aes(interval, steps)) +
```

```
    geom_line() +
    facet_grid(dateType ~ .) +
    xlab("5-minute interval") +
    ylab("avarage number of steps")
```