Reproducible Research Course Project - Part I

Arthi Murugesan

March 6,2016

Contents

Introduction

This report is the first peer assignment for the reproducible research. The exploratory analysis of daily steps taken is provided as part of the report, along with making the code reproducible.

The activity data, is as the data that was downloaded from https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip, as of 6th March 2016. The dataset consists of steps taken per day, on a 5 minute interval. In total, there are 17,568 observations available.

Loading and preprocessing the data

The data was loaded after removing the old environment variables present in R.

```
#Remove all variables in the R environment - to start fresh
rm(list=ls(all=TRUE))

#Load all the activity data

activity <- read.csv('activity.csv', header = T)
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
```

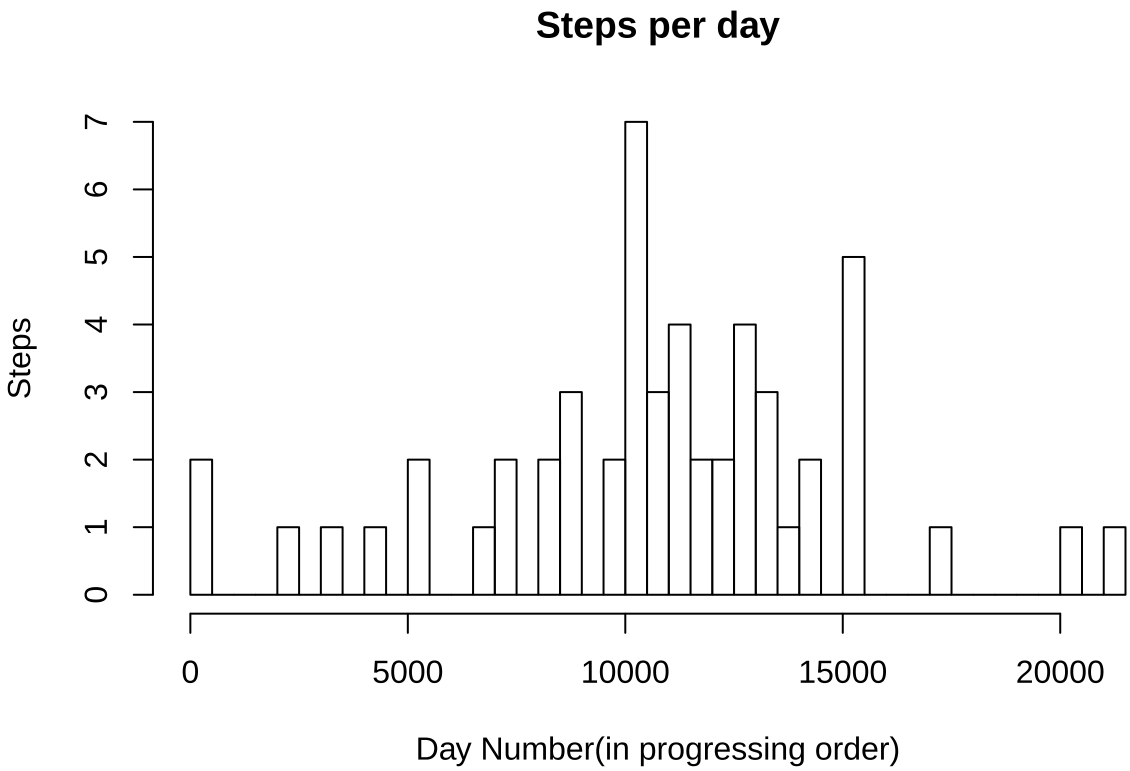
```
summary(activity)
```

```
##      steps                  date          interval
##  Min.   :  0.00   2012-10-01:  288   Min.   :   0.0
##  1st Qu.:  0.00   2012-10-02:  288   1st Qu.: 588.8
##  Median :  0.00   2012-10-03:  288   Median :1177.5
##  Mean   : 37.38   2012-10-04:  288   Mean   :1177.5
##  3rd Qu.: 12.00   2012-10-05:  288   3rd Qu.:1766.2
##  Max.   :806.00   2012-10-06:  288   Max.   :2355.0
##  NA's   :2304     (Other)   :15840
```

```
#Preprocess data to remove any data with NULL
steps_per_day<-aggregate(steps~date,data=activity,sum,na.rm=TRUE)
steps_per_interval <-aggregate(steps~interval,data=activity,mean,na.rm=TRUE)
```

What is mean total number of steps taken per day?

```
hist(steps_per_day$steps,breaks = 75,main="Steps per day",xlab="Day Number(in progressing order)", ylab
```



```
mean_steps<-round(mean(steps_per_day$steps), 2)
median_steps<-round(median(steps_per_day$steps), 2)
```

The mean number of steps taken per day is 1.07662, and the median is 1.076510. As we can see, the mean and median are not the same.

What is the average daily activity pattern?

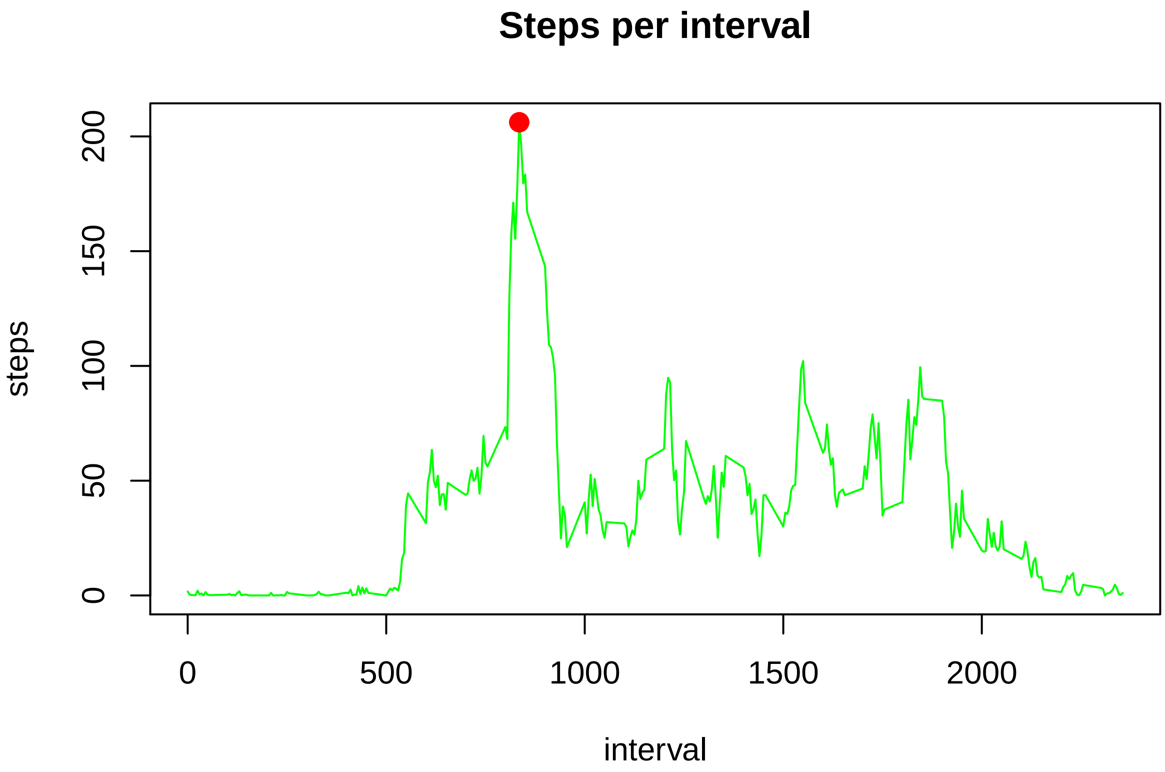
```
plot(steps~interval,data=steps_per_interval,type="l",col='green',main='Steps per interval')

#Find Interval That Has The Maximum Avg Steps
max_steps <- steps_per_interval[which.max(steps_per_interval$steps),]

max_steps
```

```
##     interval    steps
## 104      835 206.1698
```

```
#Collect Cooridinates of The Max Interval For Graphing
points(max_steps$interval,  max_steps$steps, col = 'red', lwd = 3, pch = 10)
```



Here is the plot for the average daily activity pattern. As noticed, the average activity peaks at the interval 835 with the average number of steps taken being 206.1698113.

Imputing missing values

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

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

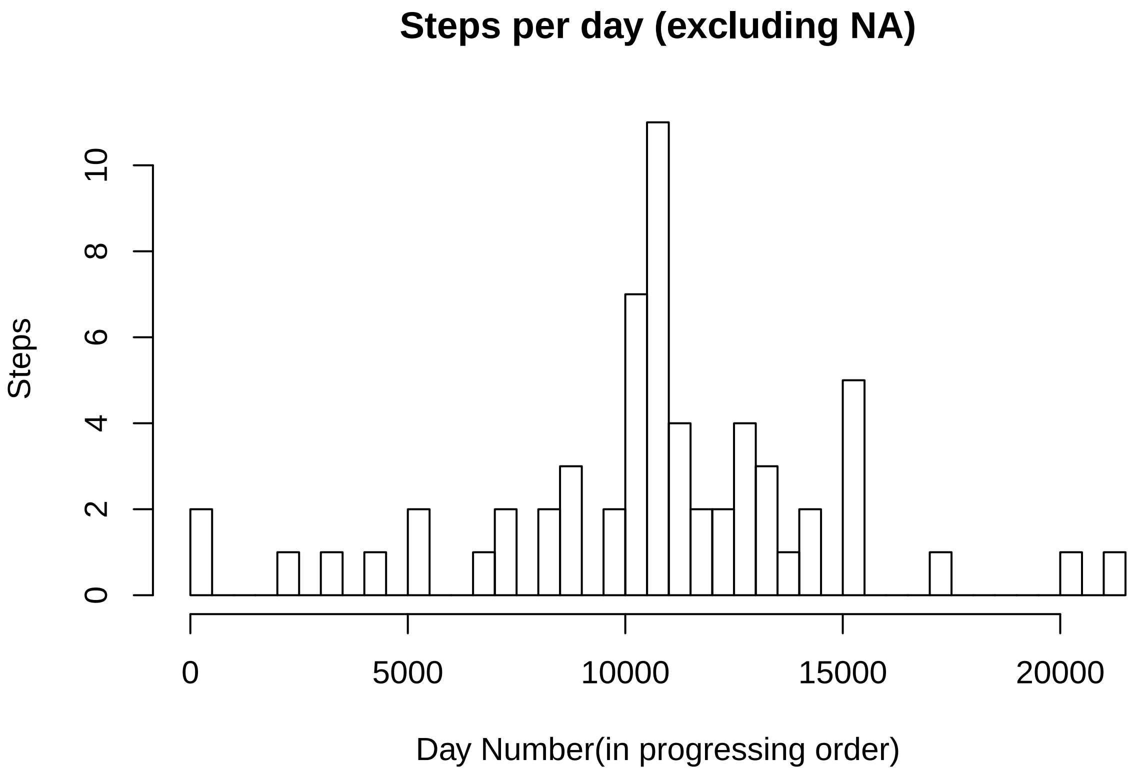
```
steps_per_interval$avg_steps<-steps_per_interval$steps
steps_per_interval$steps <- NULL

activity_final <- merge(activity, steps_per_interval, by="interval")
activity_final <- activity_final[order(activity_final$date),]

activity_final$steps[is.na(activity_final$steps)] <- activity_final$avg_steps[is.na(activity_final$step


steps_per_day_corrected<-aggregate(steps~date,data=activity_final,sum,na.rm=TRUE)

hist(steps_per_day_corrected$steps,breaks = 75,main="Steps per day (excluding NA)",xlab="Day Number(in
```



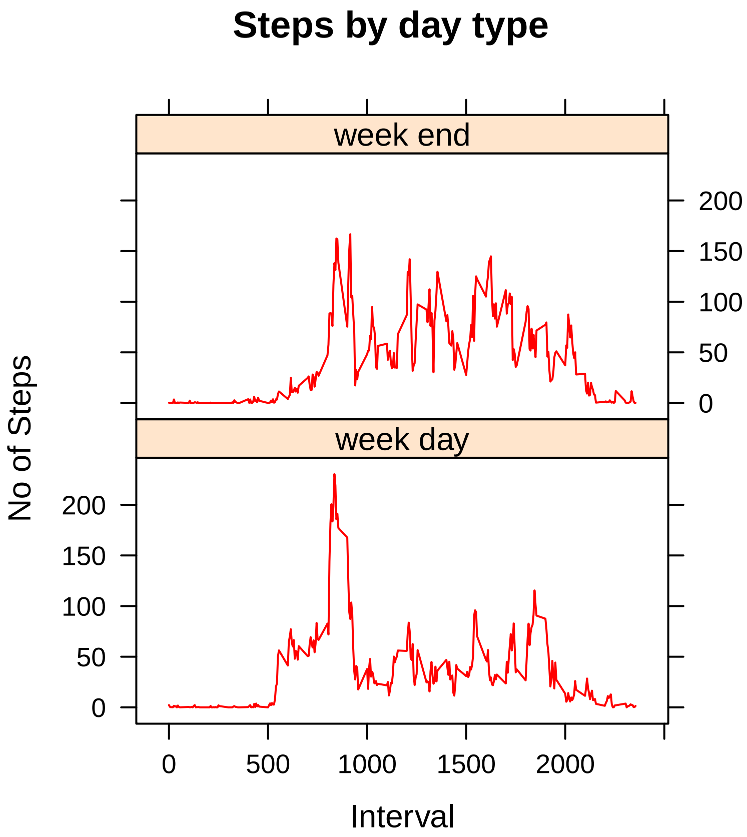
```
mean_steps<-round(mean(steps_per_day_corrected$steps), 2)
median_steps<-round(median(steps_per_day_corrected$steps), 2)
```

There are total of 2304 missing steps. The steps are replaced by the average steps taken for that interval in general. Once this is corrected, in the final dataset we noticed the mean number of steps taken are 1.07662, and the median number of steps taken are 1.07662. This shows the mean and the median being the same.

Are there differences in activity patterns between weekdays and weekends?

We notice the activities peaked during the early part of the day on weekdays. While in the case of weekends, the activities are almost average along the day but gradually reducing from earlier in the day to latter part of the day.

```
activity_final$day_type<-as.factor(ifelse(as.POSIXlt(as.Date(activity_final$date))$wday%%6==0,"week end
steps_per_interval_day_type <- aggregate(steps~interval+day_type,activity_final,mean)
xyplot(steps~interval|day_type,data=steps_per_interval_day_type,aspect=1/2,type="l",col='red',ylab='No
```


Steps by day type
week end
week day
No of Steps
Interval