Reproducible Research Project 1

Ruomeng Cui

August 14, 2015

This is the project document for the first project in reporducible research.

Loading and preprocessing the data

We first load the data.

```
activity <- read.csv('./activity.csv')
```

Mean and Median of steps per day

We then compute the mean and median of steps that are taken per day. As we can see, the mean is 10766.19, and the median is 10765.

```
dailySteps <- aggregate(steps~date, data = activity, sum)
mean(dailySteps$steps)
```

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

```
median(dailySteps$steps)
```

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

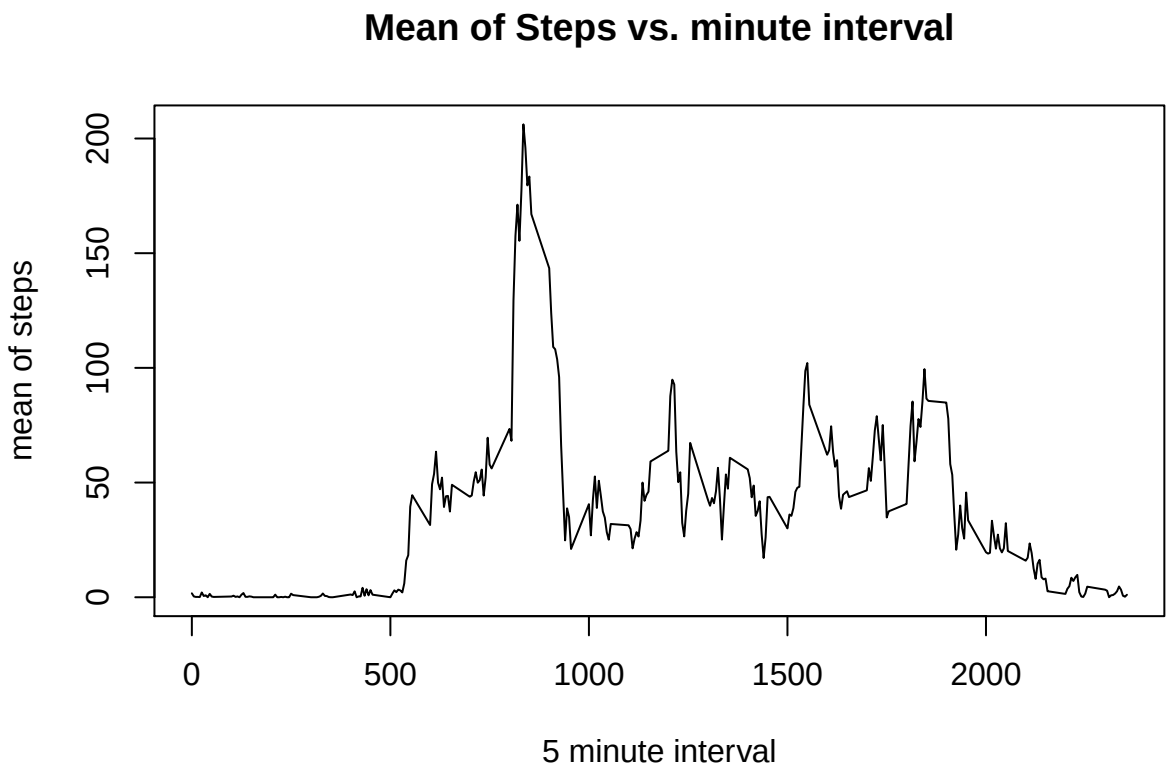
Daily activity pattern

We make a time series plot of 5-minute interval and the average number of steps across all days. It is easy to see that the interval has highest average number of steps stars at 835 minutes.

```
intervalSteps <- aggregate(steps~interval, data = activity, mean)
intervalSteps$interval[which.max(intervalSteps$step)]
```

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

```
plot(intervalSteps$interval, intervalSteps$steps, type = 'l',
     xlab = '5 minute interval', ylab = 'mean of steps',
     main = 'Mean of Steps vs. minute interval')
```



Imputing missing values

The total number of missing values in the data set is 2304.

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

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

Strategy: we can use the mean in a particular time interval to impute the missing values for that time interval. We then use this strategy to imputate the data

```
activity2 <- activity
impute_steps = NULL
for (i in activity2$interval[is.na(activity2$steps)]) {
  impute_steps <-
    c(impute_steps, intervalSteps$steps[intervalSteps$interval == i])
}
activity2$steps[is.na(activity$steps)] <- impute_steps
```

We then make a histogram of the total number of steps taken each data. The median total number of steps per day is 10766.19 and the mean is 10766.19. The mean is the same (since we impute based on the mean), while the median of the new data set is higher than the median in the first part since we imputate some missing values.

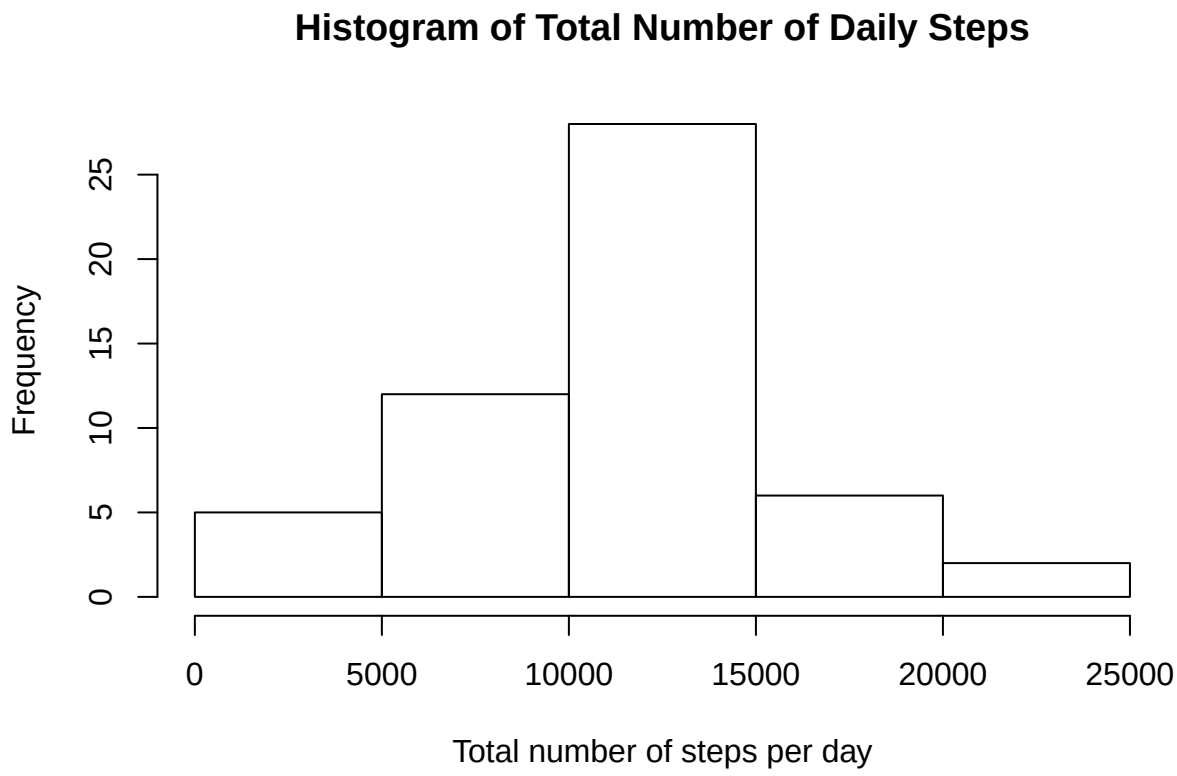
```r
dailySteps2 <- aggregate(steps~date, data = activity2, sum)
mean(dailySteps2$steps)
```

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

```r
median(dailySteps2$steps)
```

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

```r
hist(dailySteps$steps, xlab = 'Total number of steps per day',
     main = 'Histogram of Total Number of Daily Steps')
```



Difference in weekdays and weekends

We generate two plots to indicate the average steps per time interval for both weekdays and weekends. As we can see, during weekdays, people have a much higher peak in the morning and have lower activities udring the day. While during the weekends, people have lower peak in the morninig, and consistent number of steps throughout the day.

```r
activity2$weekdate <- factor(as.numeric(weekdays(as.Date(activity2$date))
                             %in% c('Saturday', 'Sunday')),
                             levels = c(0, 1), label = c('weekday', 'weekend'))
intervalSteps2 <- aggregate(steps~interval+weekdate, data = activity2, mean)
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.1.2
```

```
ggplot(aes(x=interval, y=steps), data = intervalSteps2) +
  geom_line() + facet_grid(.~weekdate) +
  ggtitle('Average Steps per Time Interval for Weekdays and Weekends')
```