Reproducible Analysis on Activity Monitoring Data

Jiachang (Ernest) Xu

6/20/2017

Section 1: Loading and preprocessing the data

First of all, we load the raw data, and take a look at the top 6 rows of the raw data.

```
## loading data from activity.csv
if(!exists("activity.raw")) {
    activity.raw <- read.csv("./activity.csv")
}
head(activity.raw)
```

```
##   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
```

We can immediately find some missing values in the "steps" column. Therefore, we have the need to process the raw data to make it analytic data. Steps of data cleanning include: (1) removing NA values in all three columns, (2) reformatting the "date" column to datetime objects, and (3) converting "interval" column into "interval.index" (a.k.a. the i-th 5-minute interval). After the data cleaning is done, we can take a quick look at the valid data frame.

```
## removing missing values
activity.valid <- activity.raw[!is.na(activity.raw$steps) & !is.na(activity.raw$date) & !is.na(activity

## reformatting date object
activity.valid$date <- as.Date(activity.valid$date)

## converting interval into interval.index
colnames(activity.valid)[3] <- c("interval.index")
activity.valid$interval.index <- activity.valid$interval.index / 5

## glancing at activity.valid
head(activity.valid)
```

```
##     steps       date interval.index
## 289     0 2012-10-02              0
## 290     0 2012-10-02              1
## 291     0 2012-10-02              2
## 292     0 2012-10-02              3
## 293     0 2012-10-02              4
## 294     0 2012-10-02              5
```

Section 2: What is mean total number of steps taken per day?

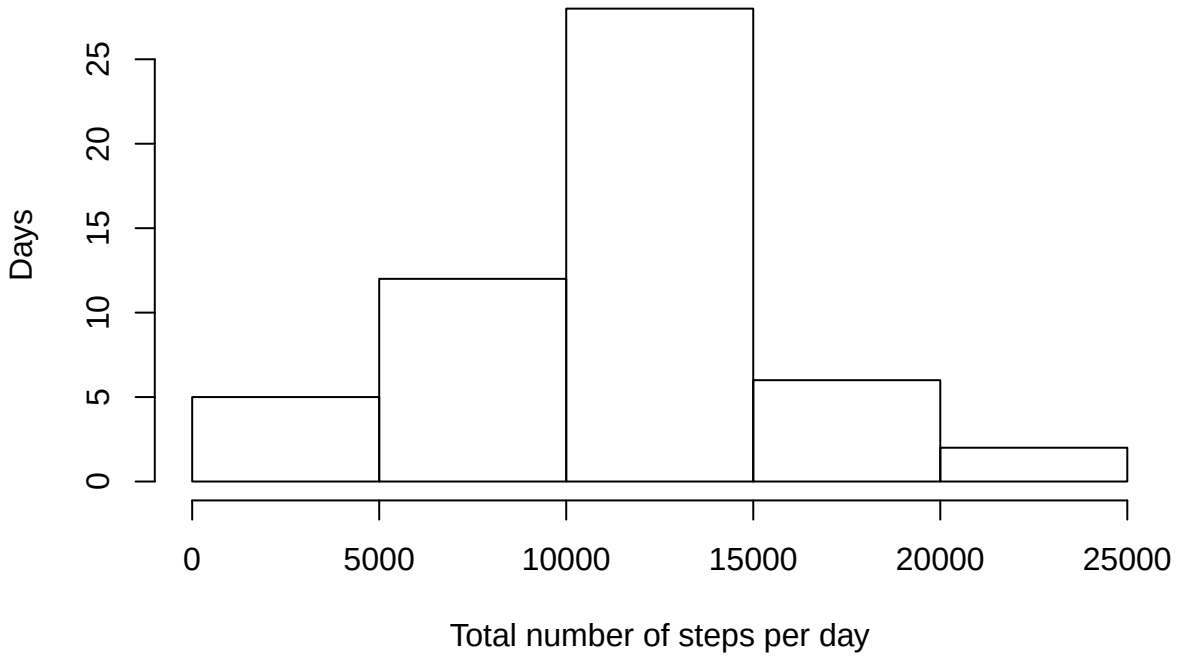
We use the aggregate() function to find the total number of step taken per day.

```
## finding sum of steps per day
activity.dailysteps <- aggregate(steps ~ date, activity.valid, sum)
```

We can plot a histogarm of total number of steps per day from the aggregated data from above.

```
## plotting histogram of the total number of steps per day
hist(activity.dailysteps$steps, xlab = "Total number of steps per day", ylab = "Days", main = "Figure 1
```

Figure 1: Total Number of Steps per Day



Then, we can easily calculate the avarage from the aggregated daily activity data.

```
mean(activity.dailysteps$steps)
```

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

Section 3: What is the average daily activity pattern?

We use the aggregate() function to find the average number of step taken per 5-minute interval across the monitoring timeline from the valid data.

```
## find average steps per 5-minute interval
activity.pattern <- aggregate(steps ~ interval.index, activity.valid, mean)
```

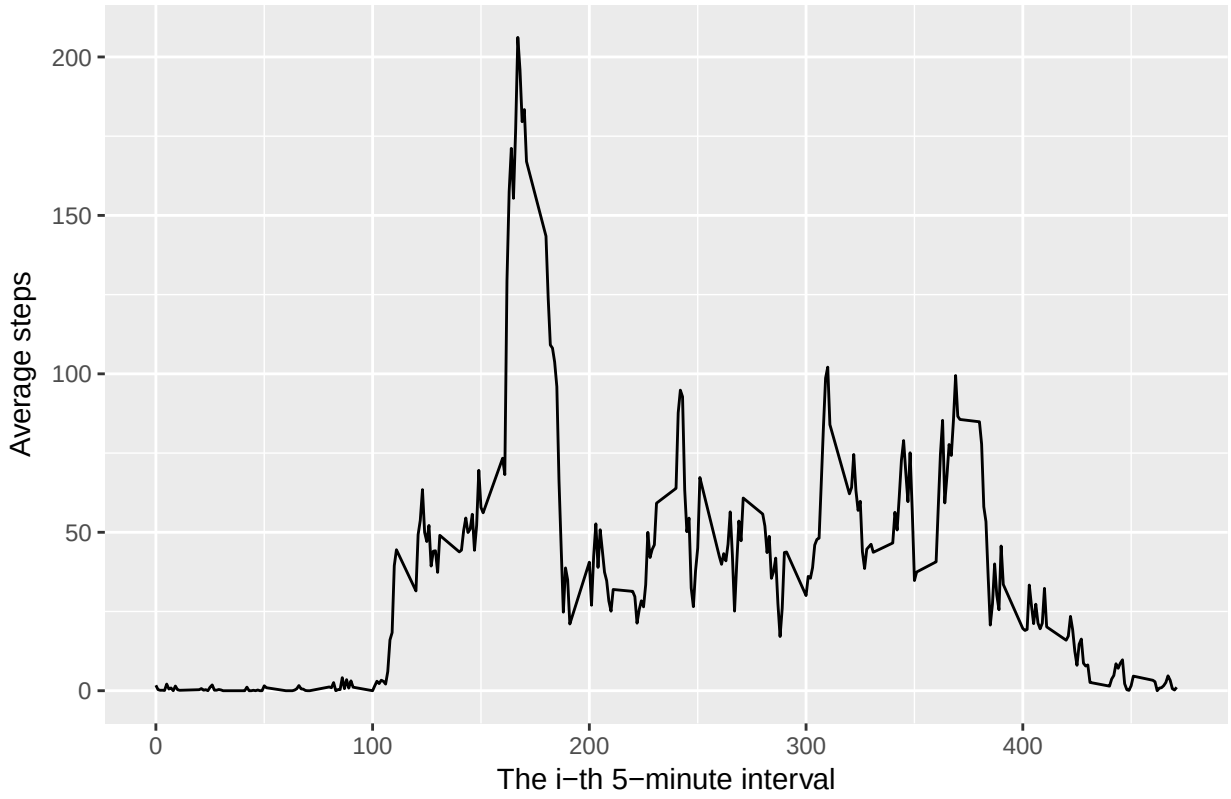
We use ggplot2 package to plot a line graph of average number of step taken per 5-minute interval from the aggregated interval activity data.

```
## plotting daily activity pattern
require(ggplot2)
```

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

```
ggplot(activity.pattern, aes(interval.index, steps)) + geom_line() + xlab("The i-th 5-minute interval")
```

Figure 2: Average Daily Activity Pattern



Section 4: Imputing missing values

Please refer back to Section 1 for details. The histogram of total number of steps per day is already displayed in Section 2.

Are there differences in activity patterns between weekdays and weekends?

First, we need to subset the valid data frame into weekdays and weekends data frame.

```
## subsetting weekdays and weekends
activity.weekdays <- activity.valid[weekdays(activity.valid$date)!="Sunday" & weekdays(activity.valid$da
activity.weekends <- activity.valid[weekdays(activity.valid$date)=="Sunday" | weekdays(activity.valid$da
head(activity.weekdays)
```

```
##     steps       date interval.index
## 289     0 2012-10-02              0
## 290     0 2012-10-02              1
## 291     0 2012-10-02              2
## 292     0 2012-10-02              3
## 293     0 2012-10-02              4
## 294     0 2012-10-02              5
```

```
head(activity.weekends)
```

```
##      steps       date interval.index
## 1441     0 2012-10-06              0
```

```
## 1442     0 2012-10-06             1
## 1443     0 2012-10-06             2
## 1444     0 2012-10-06             3
## 1445     0 2012-10-06             4
## 1446     0 2012-10-06             5
```

We use the aggregate() function to find the average number of step taken per 5-minute interval for weekdays and weekends.

```
## aggregating by weekdays/weekends
activity.weekdays.pattern <- aggregate(steps ~ interval.index, activity.weekdays, mean)
activity.weekends.pattern <- aggregate(steps ~ interval.index, activity.weekends, mean)
```

Then, we combine two data frames into one.

```
activity.pattern.2 <- cbind(activity.weekdays.pattern, activity.weekends.pattern$steps)
colnames(activity.pattern.2)[2:3] <- c("weekdays.steps", "weekends.steps")
head(activity.pattern.2)
```

```
##   interval.index weekdays.steps weekends.steps
## 1              0      2.3333333       0.000000
## 2              1      0.4615385       0.000000
## 3              2      0.1794872       0.000000
## 4              3      0.2051282       0.000000
## 5              4      0.1025641       0.000000
## 6              5      1.5128205       3.714286
```

Finally, we use ggplot2 package to plot two lines of average number of step taken per 5-minute interval from the aggregated interval activity data, representing weekdays and weekends patterns.

```
## plotting average daily pattern by weekdays/weekends
require(ggplot2)
ggplot(activity.pattern.2, aes(interval.index)) +
    geom_line(aes(y = weekdays.steps, colour = "weekdays.steps")) +
    geom_line(aes(y = weekends.steps, colour = "weekends.steps")) +
    xlab("The i-th 5-minute interval") + ylab("Average steps") +
    ggtitle("Figure 3: Average Daily Activity Pattern by Weekdays/Weekends")
```

Figure 3: Average Daily Activity Pattern by Weekdays/Weekends

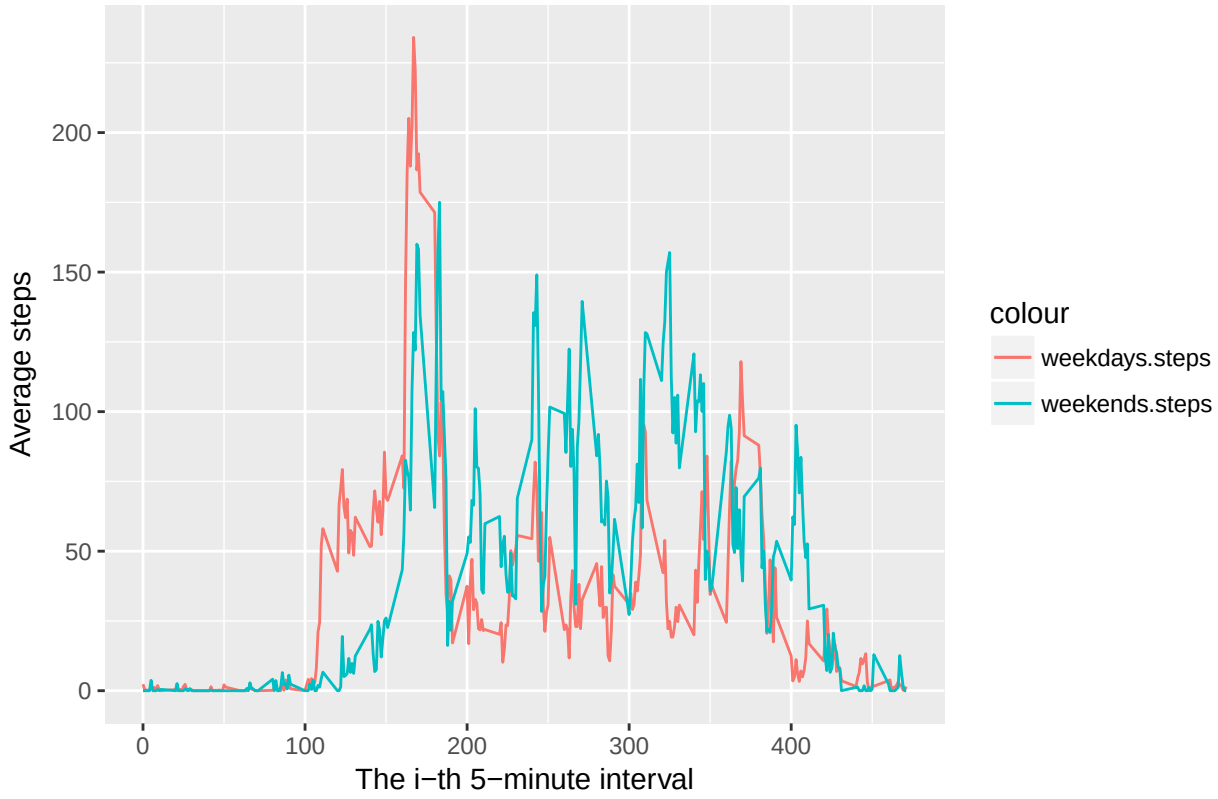


Figure 3 shows that during weekdays, this person tends to have more activities during morning rush hours, and in the evening. However, this person has much less fluctuation during daytime on weekends.