# Analysis of Activity Monitoring Data

*Connie Wang*

*October 10, 2017*

## Introduction

(From Course Project 1, Coursera, JHU Data Science Course 5: Reproducible Research, Week 2)

"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.

The variables included in this dataset are:

- **steps:** Number of steps taken 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."

## Objective

This analysis answers the questions: 1) What is the mean total number of steps taken per day? 2) What is the daily activity pattern? 3) What impact does missing data have on our results? 4) Are there differences in activity patterns between weekdays and weekends?

## Analysis

### Aim 1: What is the mean total number of steps taken per day?

First, let's take a look at our data.

```
data <- read.csv('activity.csv')
summary(data)
```

```
##      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
```

```
tail(data)
```

```
##       steps       date interval
## 17563    NA 2012-11-30     2330
## 17564    NA 2012-11-30     2335
## 17565    NA 2012-11-30     2340
## 17566    NA 2012-11-30     2345
## 17567    NA 2012-11-30     2350
## 17568    NA 2012-11-30     2355
```

```
data <- tbl_df(data) # Convert to tibble for tidyverse processing
```

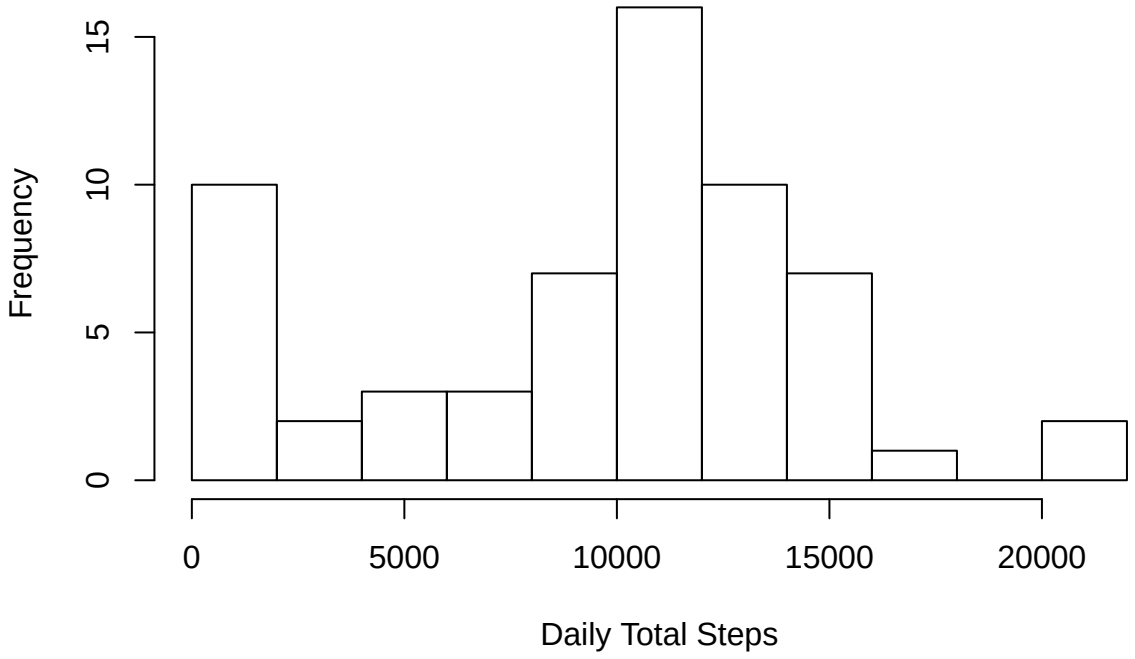
We can see that our data encompasses Oct 1, 2012 to Nov 30, 2012, with 288 entries per day, taken at 5 minute intervals. We can also see that many of the values in the "steps" column (which gives us number of steps taken in the given five minute interval) are missing – 2304 of the entries are coded as NA. We will deal with that later in our analysis.

To calculate the total number of steps taken per day, we will simply ignore any missing values and sum the total number of steps taken each day. We will plot this data in a histogram.

```
# First get daily step count totals
daily <- data %>%
  group_by(date) %>%
  summarize(total = sum(steps, na.rm=T))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.1
```

```
# Now plot daily total step counts as a histogram with 15 bins
hist(daily$total, breaks = 15, xlab = 'Daily Total Steps', main = NULL)
```



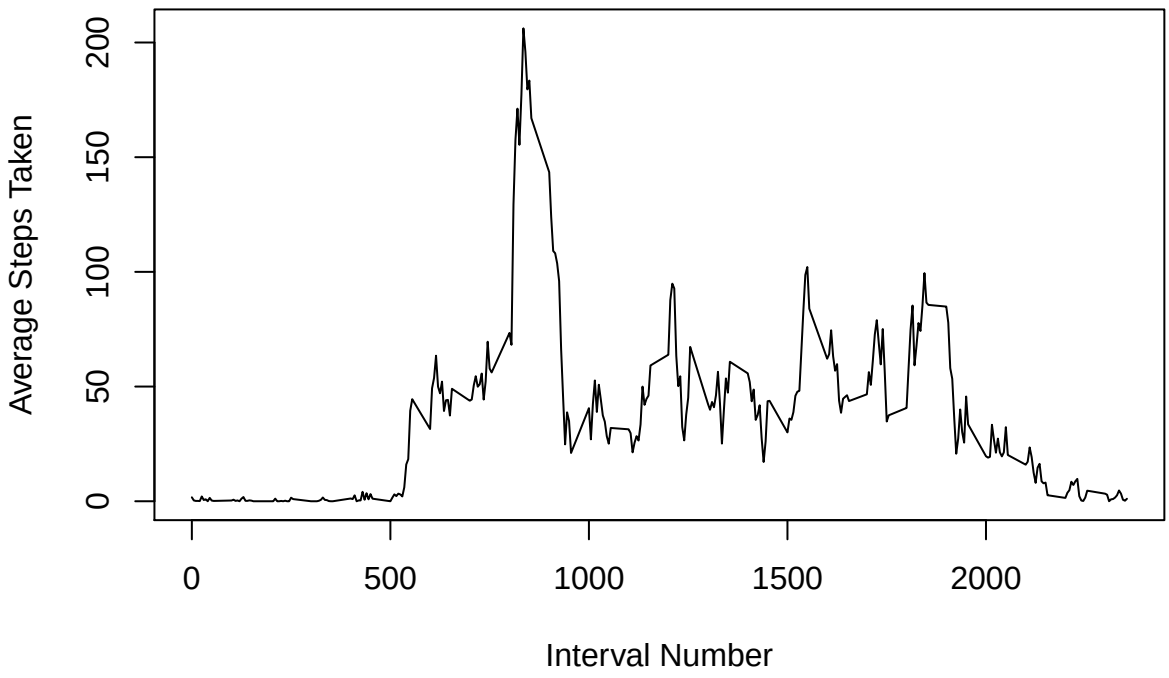

```
# Calculate the mean and median of the total number of steps taken per day
mean <- mean(daily$total)
median <- median(daily$total)
```

On **average**, 9354.2295082 steps were taken per day. The **median** number of daily steps was 10395.

### Aim 2: What is the average daily activity pattern?

Now we will look at the daily activity pattern – how many steps generally occur at each time of day? To investigate this question, we will plot the number of steps taken, averaged across all days, by the time of day (given by the index of the 5-minute interval during which the data was collected).

```
# First get averages across all days for each 5-minute interval
intervals <- data %>%
  group_by(interval) %>%
  summarize(average = mean(steps, na.rm=T))

# Plot this data as a line graph
plot(x = intervals$interval, y = intervals$average, type = 'l',
     xlab = "Interval Number", ylab = "Average Steps Taken")
```




```
# Get interval of maximum number of steps, and approximate what time of day it is by taking the interva
max <- intervals$interval[which.max(intervals$average)]
time <- (max/2355)*24
```

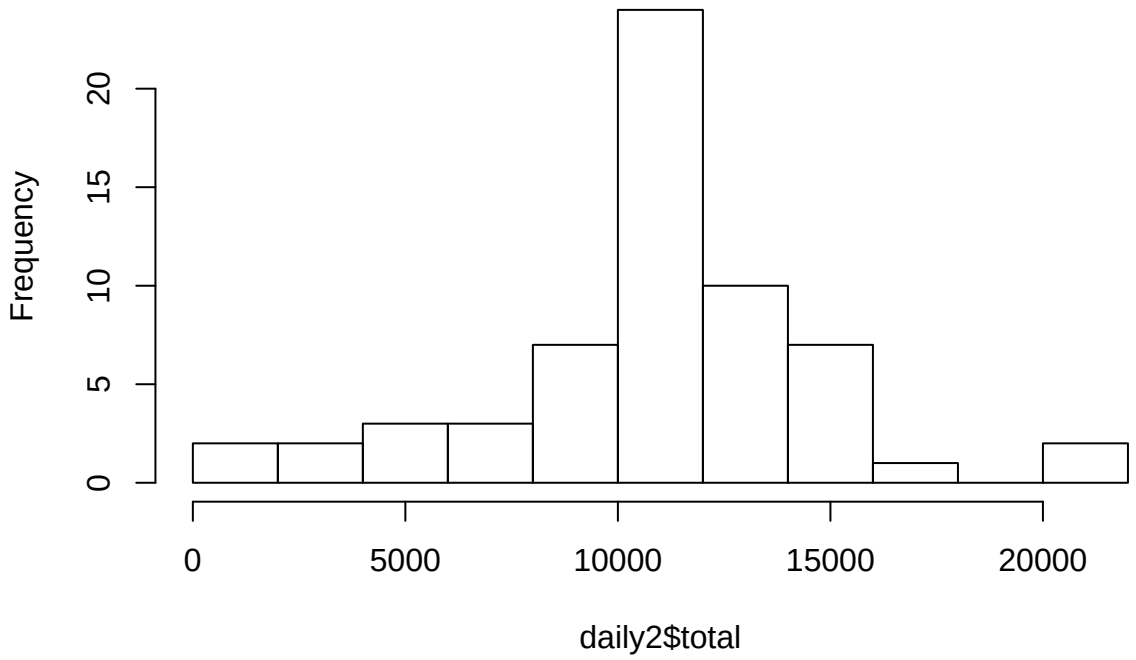
On average, the five-minute interval 835 contains the **maximum number of steps**. This occurs approximately 8.5095541 hours after midnight.

### Aim 3: What impact does missing data have on our results?

To answer this question, we will first impute the missing values in the "steps" column of our data (coded as NA). For simplicity, we will use the mean for the associated 5-minute interval as an estimate of the missing value.

```
# Fill in missing values with mean of 5-minute interval
dataComplete <- data %>%
  group_by(interval) %>%
  mutate(average = mean(steps, na.rm=T))
dataComplete$steps[is.na(dataComplete$steps)] <-
  dataComplete$average[is.na(dataComplete$steps)]

# Get daily step count totals
daily2 <- dataComplete %>%
  group_by(date) %>%
  summarize(total = sum(steps, na.rm=T))

# Now plot daily total step counts as a histogram with 15 bins
```

```r
hist(daily2$total, breaks = 15, main = NULL)
```



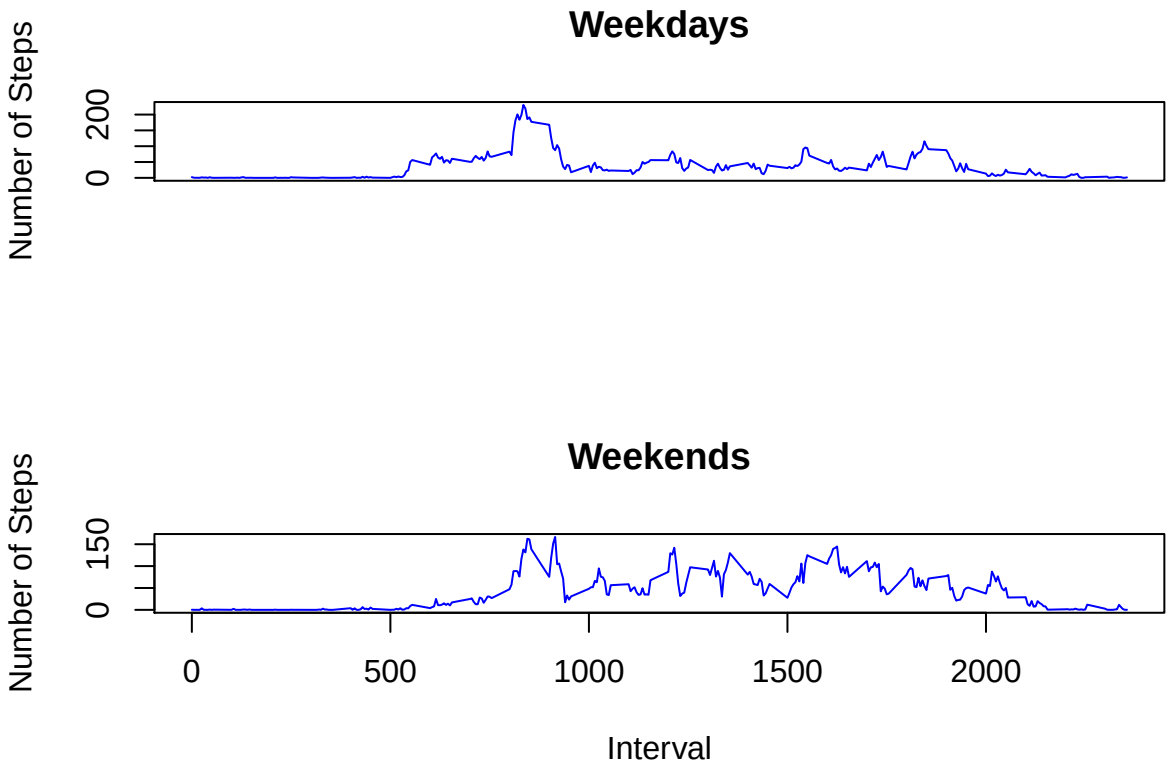

```r
# Calculate the mean and median of the total number of steps taken per day
mean2 <- mean(daily$total)
median2 <- median(daily$total)
```

On average, 9354.2295082 steps were taken per day. The median number of daily steps was 10395. This can be compared to 9354.2295082 and 10395 from aim 1. We see that imputing the missing data in this way did not change our mean and median total daily number of steps at all.

**Aim 4: Are there differences in activity patterns between weekdays and weekends?**

```r
# Add new column in dataset specifying whether weekend or weekday
dataComplete$date <- as.Date(as.character(dataComplete$date))
dataComplete <- dataComplete %>%
  ungroup %>%
  mutate(day = weekdays(dataComplete$date))

# Turn column containing days of the week to factor weekday/weekend
dataComplete$day <- replace(dataComplete$day, grep("Saturday|Sunday", dataComplete$day), "Weekend")
dataComplete$day <- replace(dataComplete$day, grep("Monday|Tuesday|Wednesday|Thursday|Friday|Saturday|Su
dataComplete$day <- as.factor(dataComplete$day)

# Create panel plot with time series plot of the average ssteps taken by 5-minute interval, averaged ac
```

```r
par(mfrow=c(2,1))

# First get averages for weekday days
weekdays <- dataComplete %>%
  filter(day == 'Weekday') %>%
  group_by(interval) %>%
  summarize(average = mean(steps))

# Plot this data as a line graph
plot(x = weekdays$interval, y = weekdays$average, type = 'l', col = "blue", xlab = '', ylab = 'Number o

# Then do the same for weekend days
weekends <- dataComplete %>%
  filter(day == 'Weekend') %>%
  group_by(interval) %>%
  summarize(average = mean(steps))
plot(x = weekends$interval, y = weekends$average, type = 'l', col = "blue", xlab = 'Interval', ylab = '
```




Users appear to be active throughout the day on weekends, but sedentary most of the day on weekdays. This makes sense, as many people have sitting jobs, but are perhaps more active on weekends (or at least running around on errands)