Reproducible Research Assignment 1

Geons

Sunday, January 10, 2016

Preparation :

```r
knitr::opts_chunk$set(fig.width=12, fig.height=8, fig.path='figure/',
                      echo=FALSE, warning=FALSE, message=FALSE)
# Clean up before start / tabula rasa (optional)
rm(list=ls())

# Load needed libraries - can be done later as well
library(knitr)
library(plyr)
library(ggplot2)
library(lattice)

# check the workdir and reset if needed
getwd()

## [1] "E:/reproducible_research/assignment_1"

setwd("e:/reproducible_research/assignment_1")
```

Loading and preprocessing the data

We will load the data to initiate processing ##### 1. Unzip if necessary and read the data

2. Reformat the Date column & eliminate NA's

What is mean total number of steps taken per day?

For this part of the assignment, you can ignore the missing values in the dataset.

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

Histogram of Total Number of Steps Taken Each Day - NA's included

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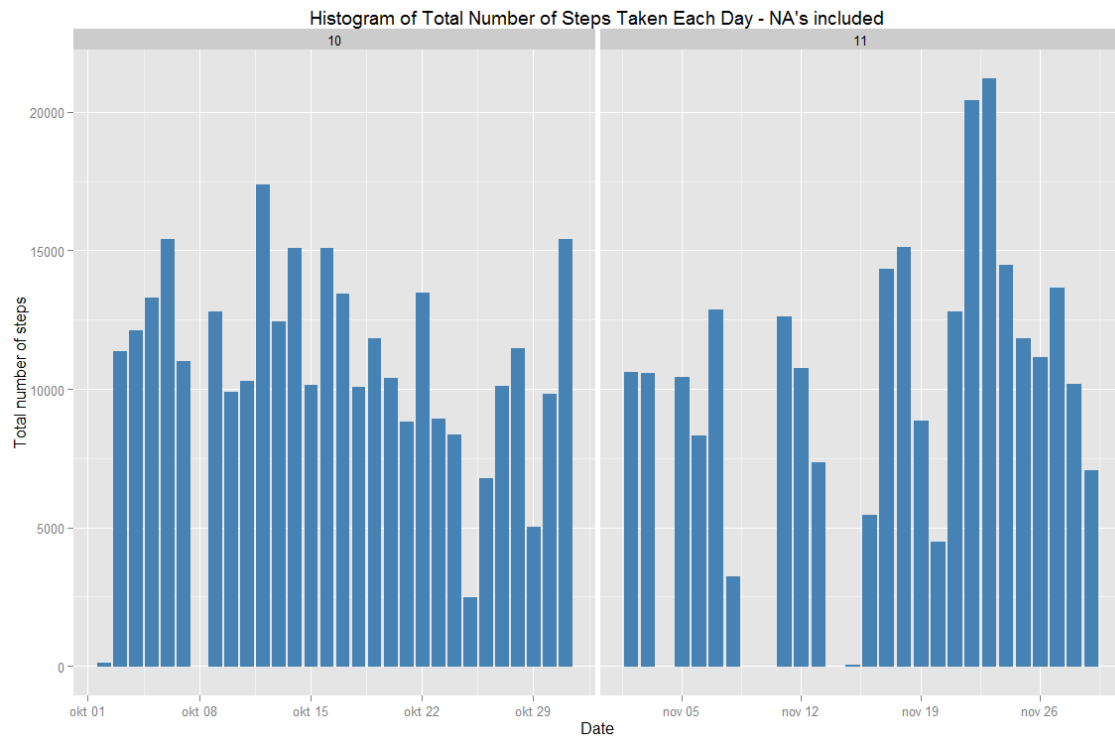
2. Calculate and report the mean and median total number of steps taken per day

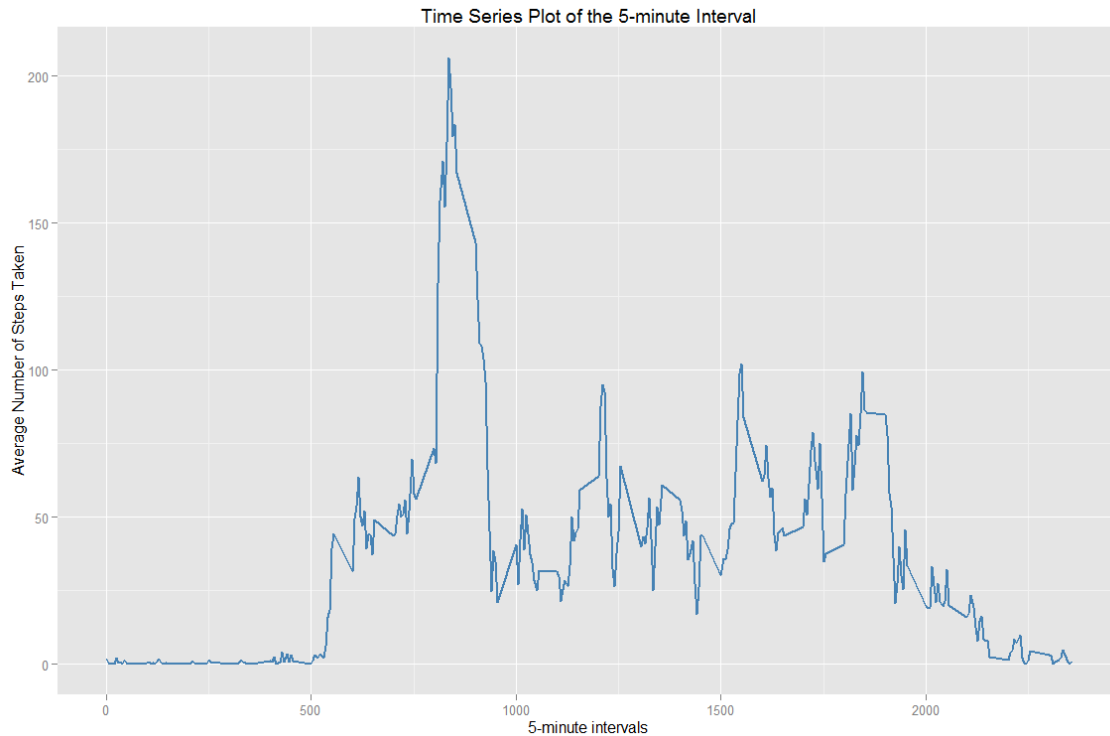The mean total number of steps taken each day is :

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

Median total number of steps taken per day:

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

What is the average daily activity pattern?

1. Make a time series plot (i.e. type = "l") of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis)

2. Which 5-minute interval, on average across all the days in the dataset, contains the maximum number of steps?

The 5-minute interval that contains on average the most steps is :

```
##     interval meanOfSteps
## 104      835    206.1698
```

Imputing missing values

The presence of missing days may introduce bias into some calculations or summaries of the data. ##### 1. Calculate and report the total number of missing values in the dataset (i.e. the total number of rows with NAs) The total number of rows with NAs:

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

2. Devise a strategy for filling in all of the missing values in the dataset. The strategy does not need to be sophisticated.

For example, you could use the mean/median for that day, or the mean for that 5-minute interval, etc.

I chose to use the median for that 5-minute interval to fill each NA value in the steps column. So first i need to create an overview of the median steps

3. Create a new dataset that is equal to the original dataset but with the missing data filled in.

4. Make a histogram of the total number of steps taken each day and Calculate and report the mean and median total number of steps taken per day.

Histogram of Total Number of Steps Taken Each Day (no missing data)
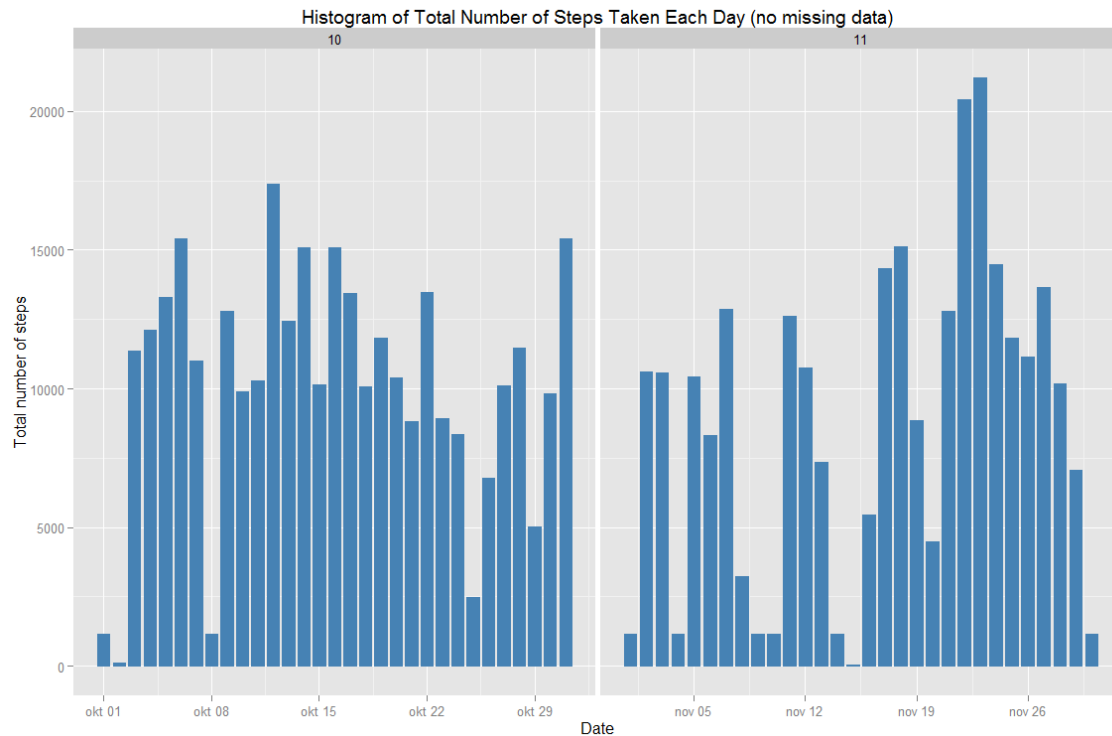
 * Do these values differ from the estimates from the first part of the assignment?

The Mean and median total number of steps taken per day are :

Compare them with the two before imputing missing data:

```
# I started this code block with {r, echo=TRUE, results='show',
warning=FALSE, message=TRUE}
# I would put echo = FALSE if it weren't for the assignment ... but hey it's
an exercise!
message("Substracting the real mean (", round(realMean), ") from the fake
mean (", round(fakeMean), ") gives : ", round(fakeMean - realMean) , " !)")

## Substracting the real mean (10766) from the fake mean (9504) gives : -1262
!)

message("Substracting the real median (", realMedian, ") from the fake median
(", fakeMedian, ") gives : ", fakeMedian - realMedian , " !)")

## Substracting the real median (10765) from the fake median (10395) gives :
-370 !)
```

- What is the impact of imputing missing data on the estimates of the total daily number of steps?

After replacing the missing data by the median 5-min interval value, the fake mean of total steps taken per day is 1262 steps less than that of the mean of the recorded data; the new (fake) median of total steps taken per day is 370 steps less than that of the real (recorded) median.
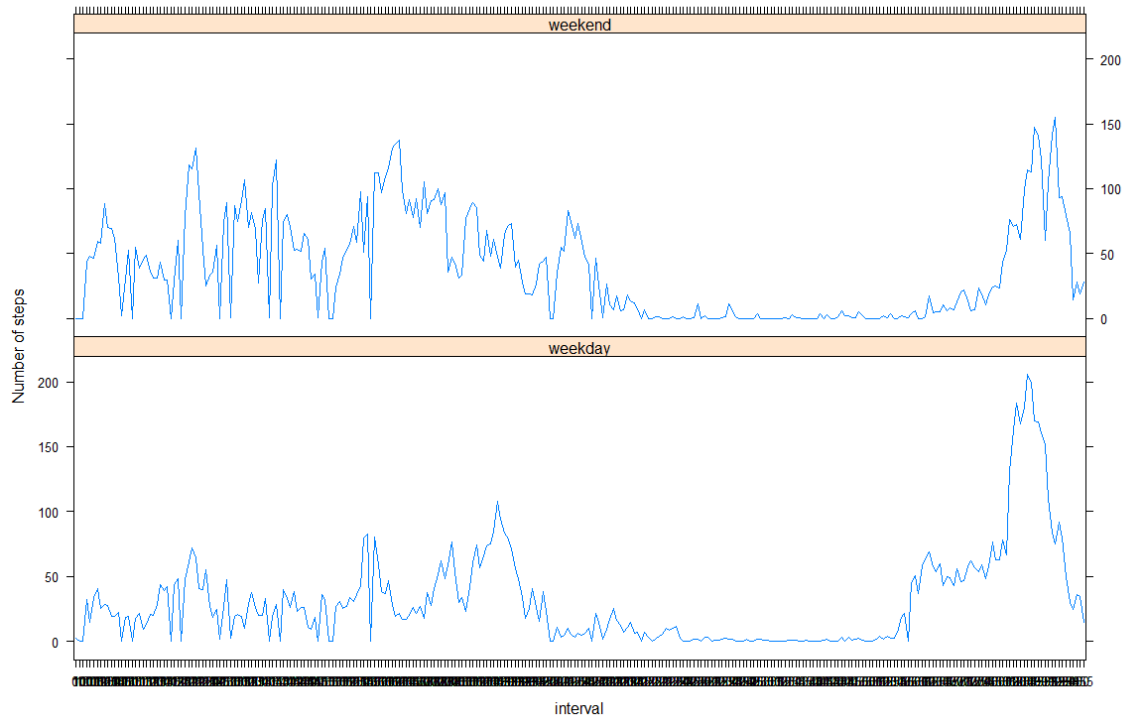
We can also look at the difference in a graph :
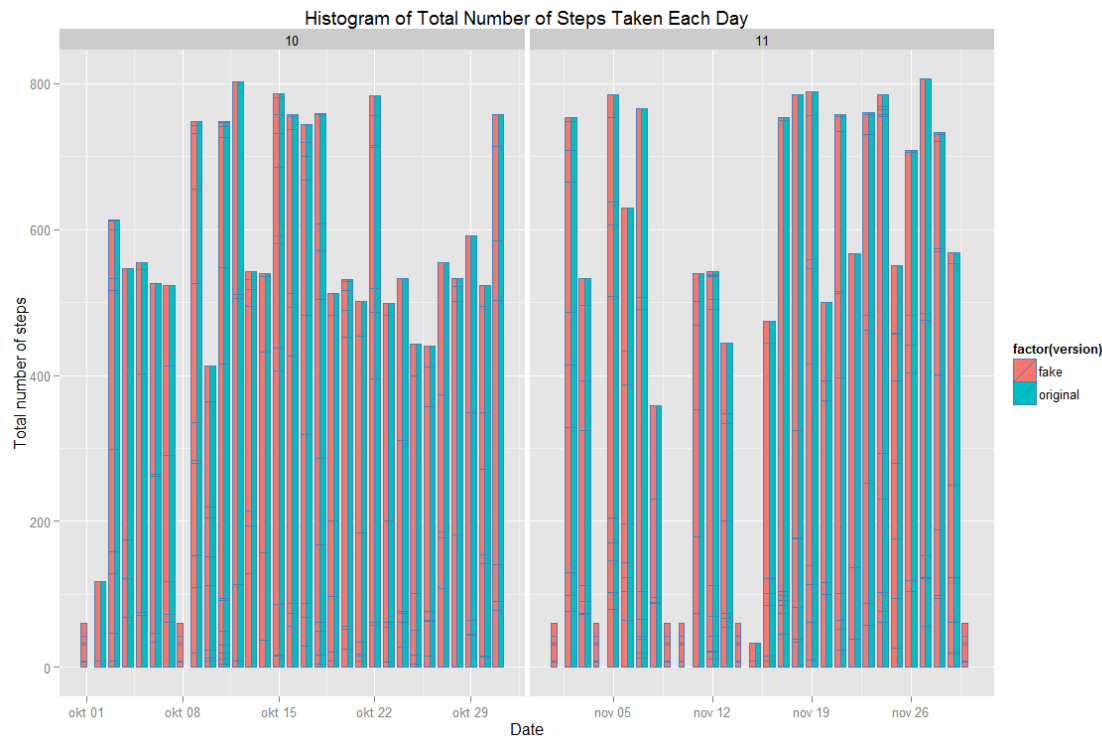
 As such the differences do not look as big. Which goes to show that graphs do not always reveal all (or that I should have chosen another graph for this illustration).

Are there differences in activity patterns between weekdays and weekends?

For this part the dataset with the filled-in missing values was used as instructed.

There seems to be a slight difference between weekdays and weekends. Weekdays are characterised by a longer period of low activity probably corresponding to time spent at work (desk). Also sleep seems to be longer during the weekend (the really flat part of the graph).