RepData_PeerAssessment1

Somaya AlGabry

13/02/2021

This is my first project assignment for Reproducible Research course by Johns Hopkins University on Coursera.

For more details about the data please have a look on README file. You can also download the data from here

First I am setting echo = True to make sure the code is always included throughout the report.

```
knitr::opts_chunk$set(echo = TRUE)
```

load packages:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

Loading and preprocessing the data:

Show any code that is needed to

1.Load the data (i.e. read.csv())
2.Process/transform the data (if necessary) into a format suitable for your analysis

```r
# read data:
if (!file.exists("activity.csv")) {
  unzip("activity.zip")
}
data <- read.csv("activity.csv")

# explore data:
head(data)
```

```
##   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
```

```r
str(data)
```

```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : chr  "2012-10-01" "2012-10-01" "2012-10-01" "2012-10-01" ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
```

```r
# change the format of date to date format:
data$date <- as.Date(data$date)
```

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.

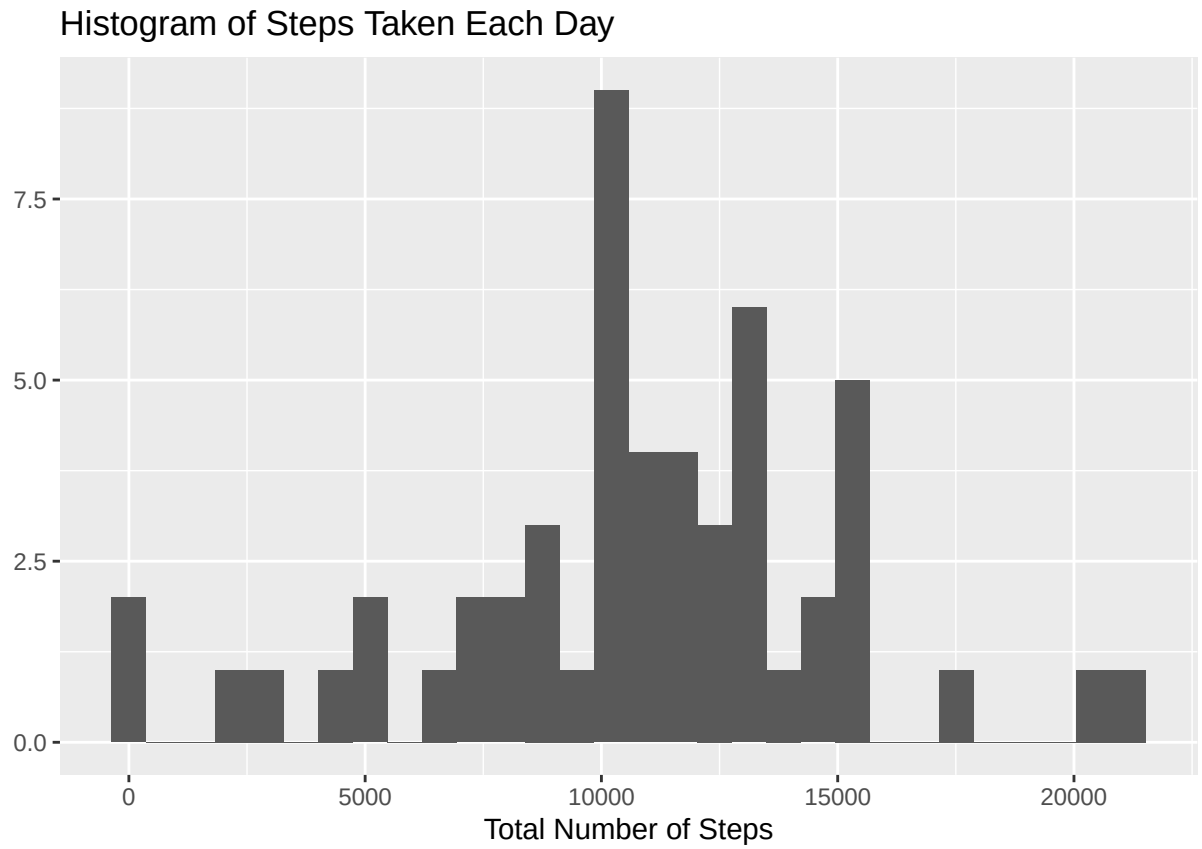
Calculate the total number of steps taken per day If you do not understand the difference between a histogram and a barplot, research the difference between them. Make a histogram of the total number of steps taken each day Calculate and report the mean and median of the total number of steps taken per day

```r
# total number of steps each day:
data2 <- data %>% filter(!is.na(steps)) %>%
  group_by(date) %>%
  summarise(stepsum= sum(steps))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
# histogram:
qplot(data2$stepsum, geom = "histogram", xlab = "Total Number of Steps", main = "Histogram of Steps Tak
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#mean and median:
mean(data2$stepsum)
```

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

```
median(data2$stepsum)
```

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

mean of the total number of steps taken per day = 10766.19 steps/day median of the total number of steps taken per day = 10765 steps/day

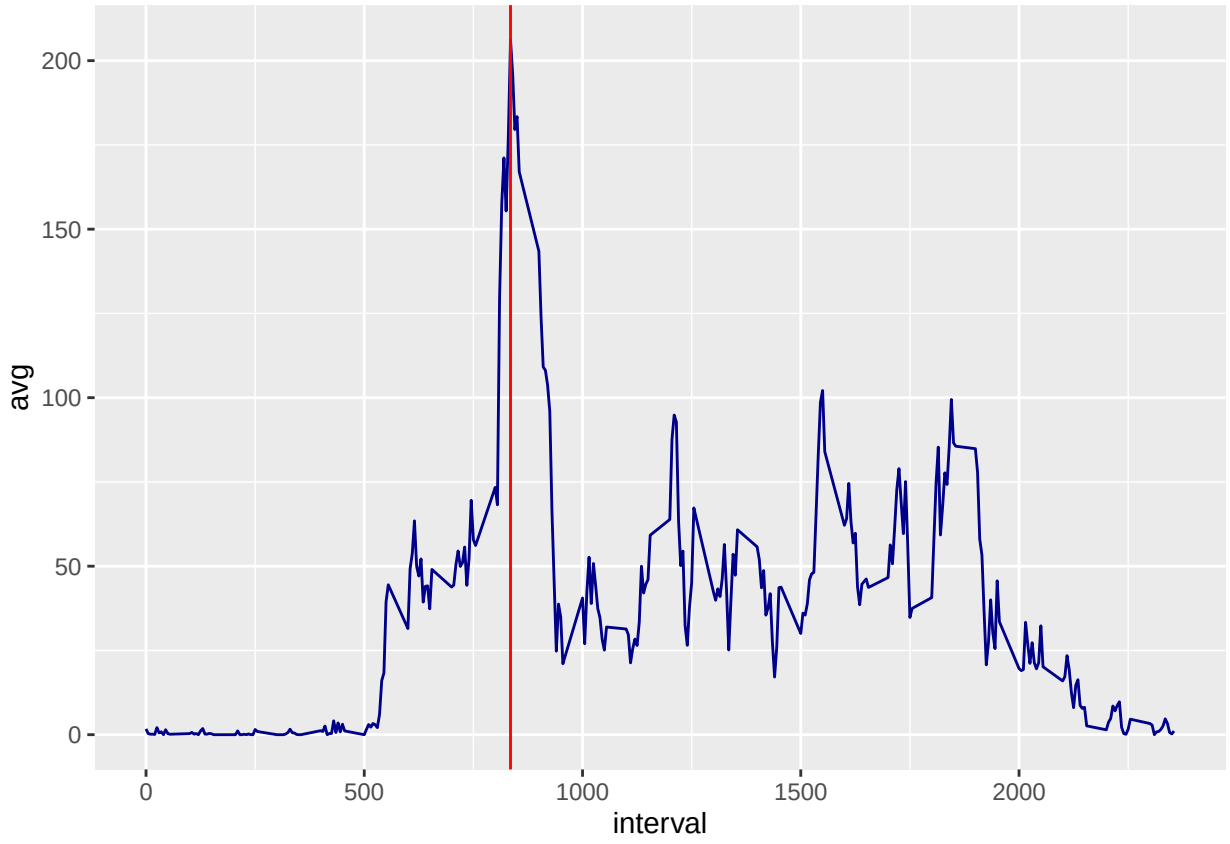
What is the average daily activity pattern?

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

```
data3 <- data %>% filter(!is.na(steps)) %>%
  group_by(interval) %>%
  summarise(avg = mean(steps))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
ggplot(data3, aes( x = interval, y = avg)) +
  geom_line(col = "darkblue") +
  geom_vline(xintercept = 835, col = "red")
```




```
data3[which.max(data3$avg),]
```

```
## # A tibble: 1 x 2
##   interval   avg
##      <int> <dbl>
## 1      835  206.
```

interval 835 contains the maximum number of steps which is 206.1698 steps.

Imputing missing values

Note that there are a number of days/intervals where there are missing values (coded as NA). 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)

```r
sum(is.na(data))
```

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

The total number of missing values are 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.
3. Create a new dataset that is equal to the original dataset but with the missing data filled in.

```r
data4 <- data
data4$steps[is.na(data4$steps)] <- data3$avg
sum(is.na(data4))
```

```
## [1] 0
```

There are no missing values in the new dataset. They have been replaced by the average of the 5-minute-interval.

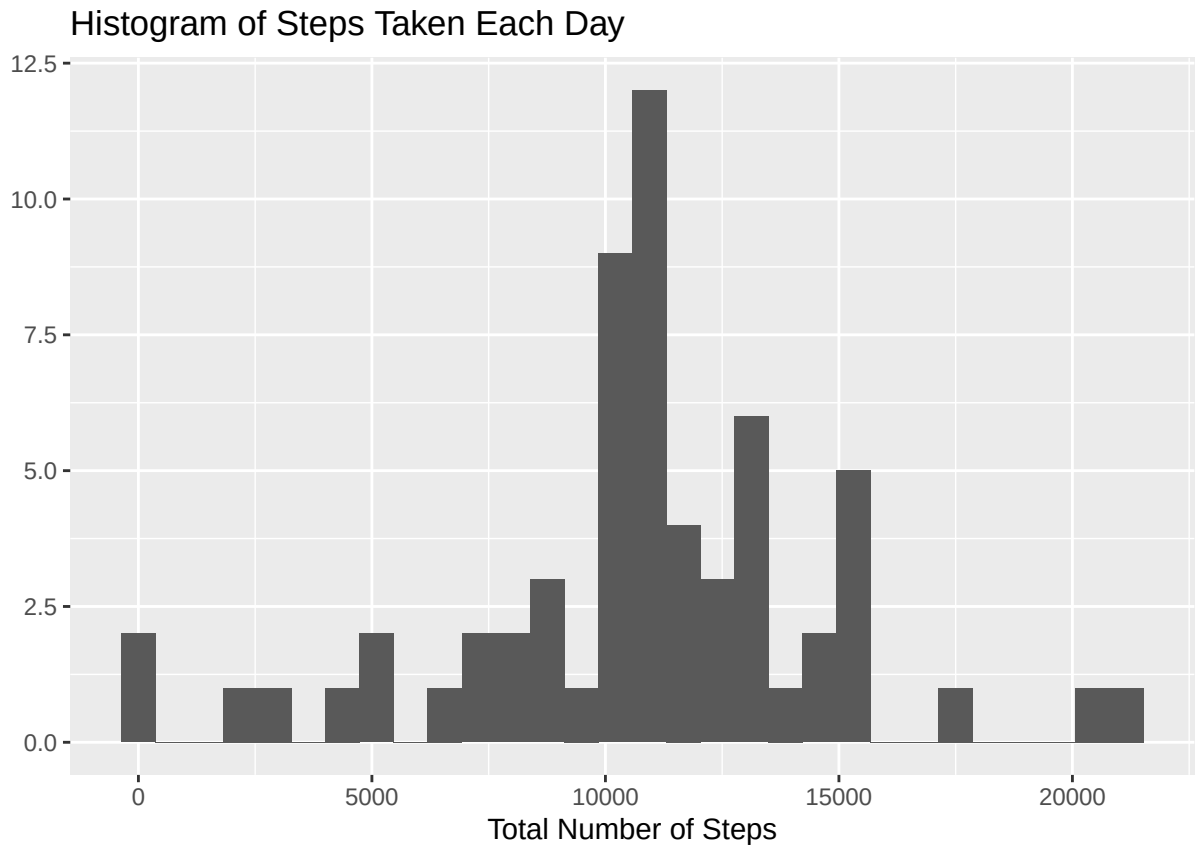
4. Make a histogram of the total number of steps taken each day and Calculate and report the mean and median total number of steps taken per day. Do these values differ from the estimates from the first part of the assignment? What is the impact of imputing missing data on the estimates of the total daily number of steps?

```r
# total number of steps each day:
data5 <- data4 %>% filter(!is.na(steps)) %>%
  group_by(date) %>%
  summarise(stepsum= sum(steps))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
# histogram:
qplot(data5$stepsum, geom = "histogram", xlab = "Total Number of Steps", main = "Histogram of Steps Tak
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# mean and median:
mean(data5$stepsum)
```

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

```
median(data5$stepsum)
```

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

Both the mean and median are 10766.19 steps/day. The mean does not change after replacing missing values but the median changed and became equal to the mean.

Are there differences in activity patterns between weekdays and weekends?

For this part the weekdays() function may be of some help here. Use the dataset with the filled-in missing values for this part.

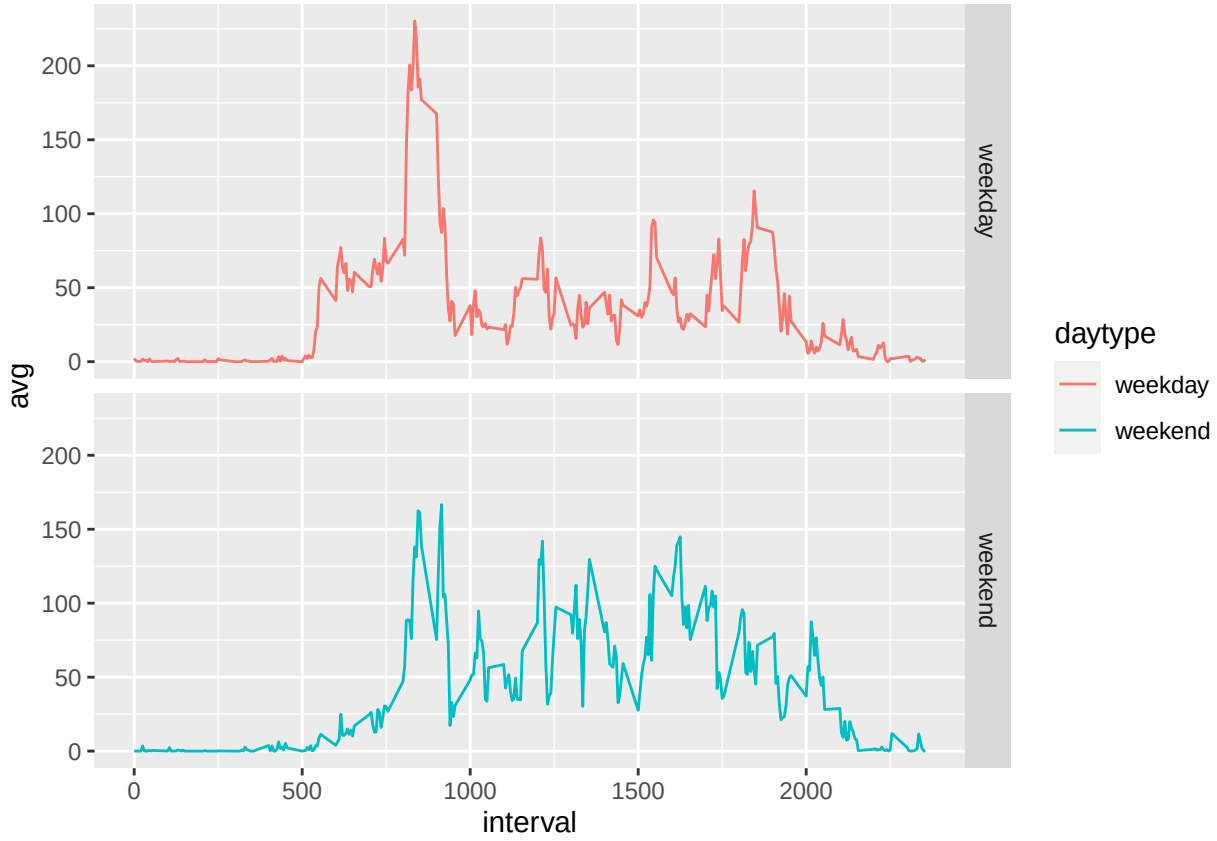
Create a new factor variable in the dataset with two levels – "weekday" and "weekend" indicating whether a given date is a weekday or weekend day. Make a panel plot containing a time series plot (i.e. type = "l") of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all weekday days or weekend days (y-axis).

```r
# add daytype:
day <- weekdays(data4$date)
daytype <-  ifelse(day == "Saturday" | day == "Sunday", "weekend", "weekday")
daytype <- as.factor(daytype)
data6 <- cbind(data4, daytype)
head(data6)
```

```
##       steps       date interval daytype
## 1 1.7169811 2012-10-01        0 weekday
## 2 0.3396226 2012-10-01        5 weekday
## 3 0.1320755 2012-10-01       10 weekday
## 4 0.1509434 2012-10-01       15 weekday
## 5 0.0754717 2012-10-01       20 weekday
## 6 2.0943396 2012-10-01       25 weekday
```

```r
# time-series plot:
data7 <- data6 %>%
  group_by(interval, daytype) %>%
  mutate(avg = mean(steps))

ggplot(data7, aes( x = interval, y = avg, color = daytype)) +
  geom_line() +
  facet_grid(daytype~.)
```



From the plot, the activity in the early hours of the day is higher in the weekdays, but in the late hours it is slightely higher in the weekends.