PA1_template

Felix Garcia A.

15/6/2020

Course Project 1

Assignment

This assignment will be described in multiple parts. You will need to write a report that answers the questions detailed below. Ultimately, you will need to complete the entire assignment in a single R markdown document that can be processed by knitr and be transformed into an HTML file.

Throughout your report make sure you always include the code that you used to generate the output you present. When writing code chunks in the R markdown document, always use echo = TRUE so that someone else will be able to read the code. This assignment will be evaluated via peer assessment so it is essential that your peer evaluators be able to review the code for your analysis.

For the plotting aspects of this assignment, feel free to use any plotting system in R (i.e., base, lattice, ggplot2)

Fork/clone the GitHub repository created for this assignment. You will submit this assignment by pushing your completed files into your forked repository on GitHub. The assignment submission will consist of the URL to your GitHub repository and the SHA-1 commit ID for your repository

Loading the required data

First we need to load the required packages and the data base from your directory:

```
library(ggplot2)
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
```

```
if (!file.exists('activity.csv')) {
  unzip(zipfile = "activity.zip")
}
activityData <- read.csv(file="activity.csv", header=TRUE)
```

Question 1: What is mean total number of steps taken per day?

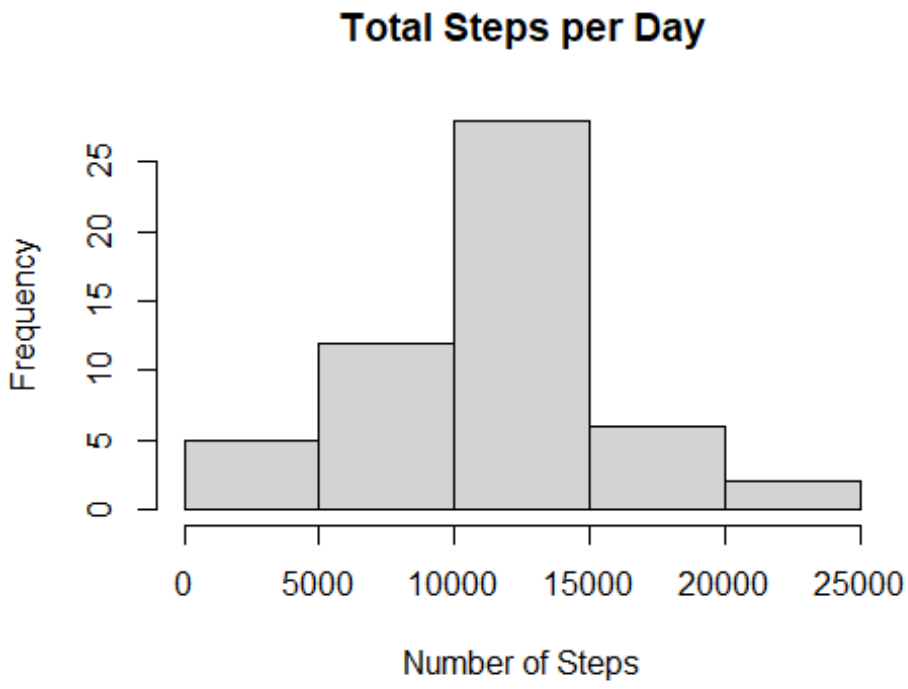
For this question we need to calculate the total steps taken per day.

```
totalSteps <- aggregate(steps ~ date, activityData, FUN=sum)
head(totalSteps)
```

```
##         date steps
## 1 2012-10-02   126
## 2 2012-10-03 11352
## 3 2012-10-04 12116
## 4 2012-10-05 13294
## 5 2012-10-06 15420
## 6 2012-10-07 11015
```

Then we plot a histogram to see its distribution.

```
hist(totalSteps$steps,
     main = "Total Steps per Day",
     xlab = "Number of Steps")
```



Finally we calculate and report the mean and median of total steps taken per day.

```
meanSteps <- mean(totalSteps$steps, na.rm = TRUE)
medSteps <- median(totalSteps$steps, na.rm = TRUE)
meanSteps
```

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

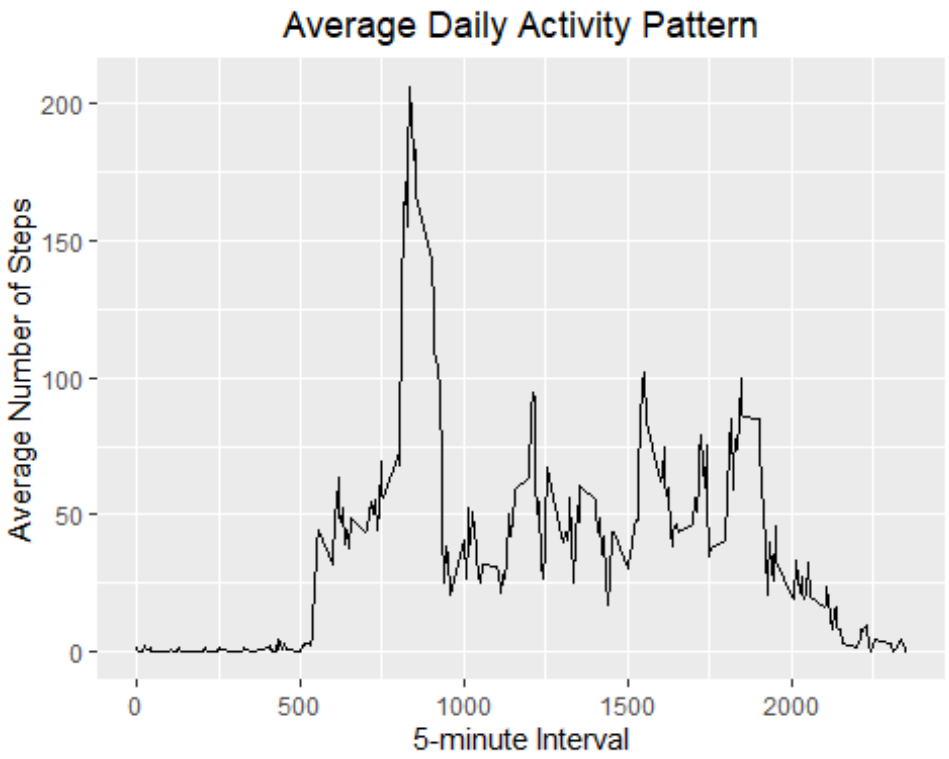
```
medSteps
```

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

Question 2: What is the average daily activity pattern?

First we need to make a time-series plot of the 5-minute interval and secondly the average number of steps taken, averaged across all days.

```
meanStepsByInt <- aggregate(steps ~ interval, activityData, mean)
meanStepsByInt %>% ggplot(aes(x = interval, y = steps)) +
  geom_line() +
  ggtitle("Average Daily Activity Pattern") +
  xlab("5-minute Interval") +
  ylab("Average Number of Steps") +
  theme(plot.title = element_text(hjust = 0.5))
```



Also we calculate in which 5-minute interval across all days contain the maximum number of steps

```
maxInt <- meanStepsByInt[which.max(meanStepsByInt$steps),]
maxInt
```

```
##     interval    steps
## 104      835 206.1698
```

Quesyion 3: Imputing Missing Values

For this question we need to calculate and report the total number of missing values in the dataset.

```
missingVals <- is.na(activityData$steps)
sum(missingVals)

## [1] 2304
```

So we Devise a strategy for filling in all of the missing values. Existing missing values we will replace these missing values with the 5-day average of that respective interval.

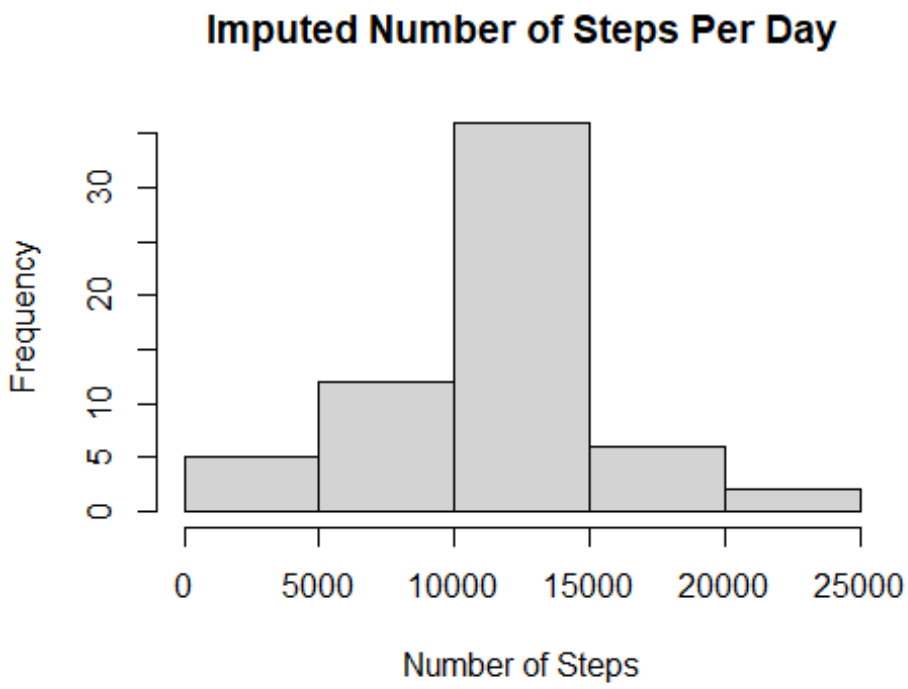
- Creating a new dataset that is equal to the original dataset but with the missing data filled in.

```
imp_activityData <- transform(activityData,
                    steps = ifelse(is.na(activityData$steps),
                    meanStepsByInt$steps[match(activityData$interval,
                    meanStepsByInt$interval)],
                    activityData$steps))
head(imp_activityData)

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

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

```
impStepsByInt <- aggregate(steps ~ date, imp_activityData, FUN=sum)
hist(impStepsByInt$steps,
     main = "Imputed Number of Steps Per Day",
     xlab = "Number of Steps")
```




```
impMeanSteps <- mean(impStepsByInt$steps, na.rm = TRUE)
impMedSteps <- median(impStepsByInt$steps, na.rm = TRUE)
diffMean = impMeanSteps - meanSteps
diffMed = impMedSteps - medSteps
diffTotal = sum(impStepsByInt$steps) - sum(totalSteps$steps)
diffMean; diffMed; diffTotal
```

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

```
## [1] 1.188679
```

```
## [1] 86129.51
```

There is a difference of 0 in the mean steps of the two dataset. There is a difference of 1.188679 in the median steps of the two dataset. There is a difference of 86129.51 in the total steps of the two dataset.

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

At first we need to create a new factor variable in the dataset with two levels - "weekend" and "weekday".

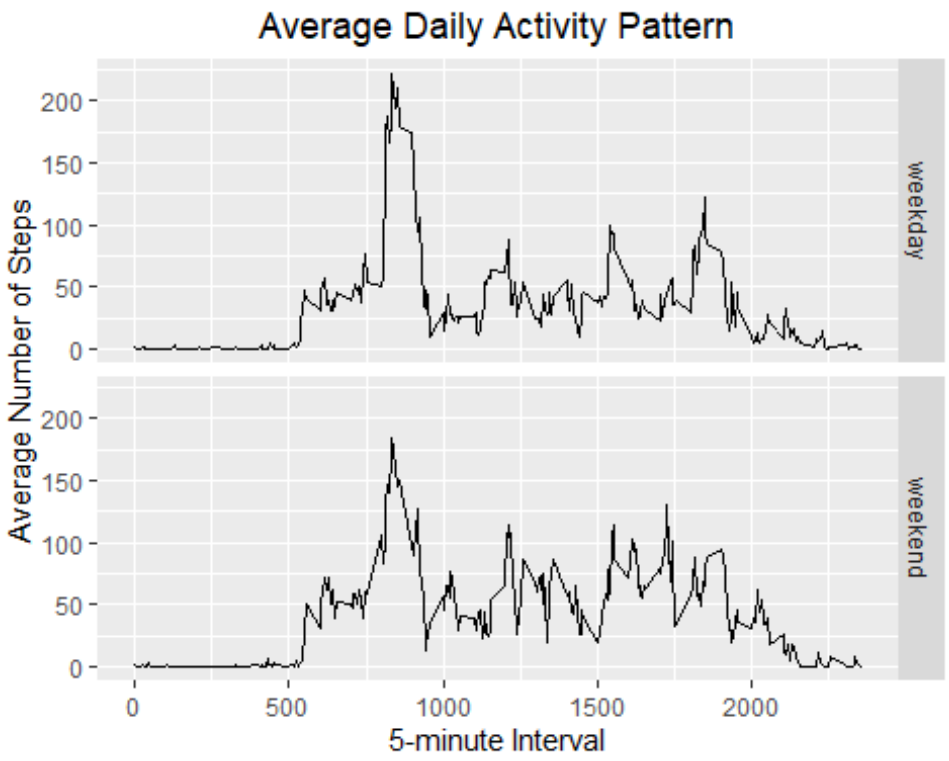
```
imp_activityData$date <- as.Date(imp_activityData$date)
imp_activityData <-imp_activityData %>%
  mutate(day =
           ifelse(weekdays(date) %in% c('lunes', 'martes', 'miercoles',
```

```
'jueves', 'viernes','Monday', 'Tuesday', 'Wednesday', 'Thursday',
'Friday'),
                  "weekday","weekend"))
head(imp_activityData)
```

```
##       steps       date interval     day
## 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
```

Finally we make a panel plot containing a time-series plot of the 5-minute interval and the average number of steps taken across all weekdays or weekends.

```
meanStepsByDay <- aggregate(steps ~ interval + day, imp_activityData,
mean)
ggplot(data = meanStepsByDay, aes(x = interval, y = steps)) +
  geom_line() +
  facet_grid(day ~ .) +
  ggtitle("Average Daily Activity Pattern") +
  xlab("5-minute Interval") +
  ylab("Average Number of Steps") +
  theme(plot.title = element_text(hjust = 0.5))
```



So, we can say that exists a difference in measures but not in the patters.