Reproducible Research - Project 1

Mehdi HAMDOUNE

3/11/2021

Instroduction

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 data for this assignment can be downloaded from the course web site:

Dataset: Activity monitoring data [52K]

The variables included in this dataset are:

- steps: Number of steps taking 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.

Loading and preprocessing the data

1- Loading the data

```
setwd("D:/Auto_formation/Coursera_Reproducible_Research/Project 1")

# Importing The LIbrairies
library(tidyr)
library(ggplot2)

# Importing The Dataset
Activity <- read.csv("activity.csv")

head(Activity, 5)
```

```
##   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
```

```
str(Activity, 5)
```

```
## '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 ...
```

2- Preprocessing the data

```
# Format date to Type Date
Activity$date <- as.Date(Activity$date, "%Y-%m-%d")

#Creating weekday variable
weekday <- weekdays(Activity$date)

Activity <- cbind(Activity, weekday)

summary(Activity)
```

```
##      steps             date               interval        weekday
##  Min.   :  0.00   Min.   :2012-10-01   Min.   :   0.0   Length:17568
##  1st Qu.:  0.00   1st Qu.:2012-10-16   1st Qu.: 588.8   Class :character
##  Median :  0.00   Median :2012-10-31   Median :1177.5   Mode  :character
##  Mean   : 37.38   Mean   :2012-10-31   Mean   :1177.5
##  3rd Qu.: 12.00   3rd Qu.:2012-11-15   3rd Qu.:1766.2
##  Max.   :806.00   Max.   :2012-11-30   Max.   :2355.0
##  NA's   :2304
```

What is mean total number of steps taken per day?

1- The total number of steps taken per day

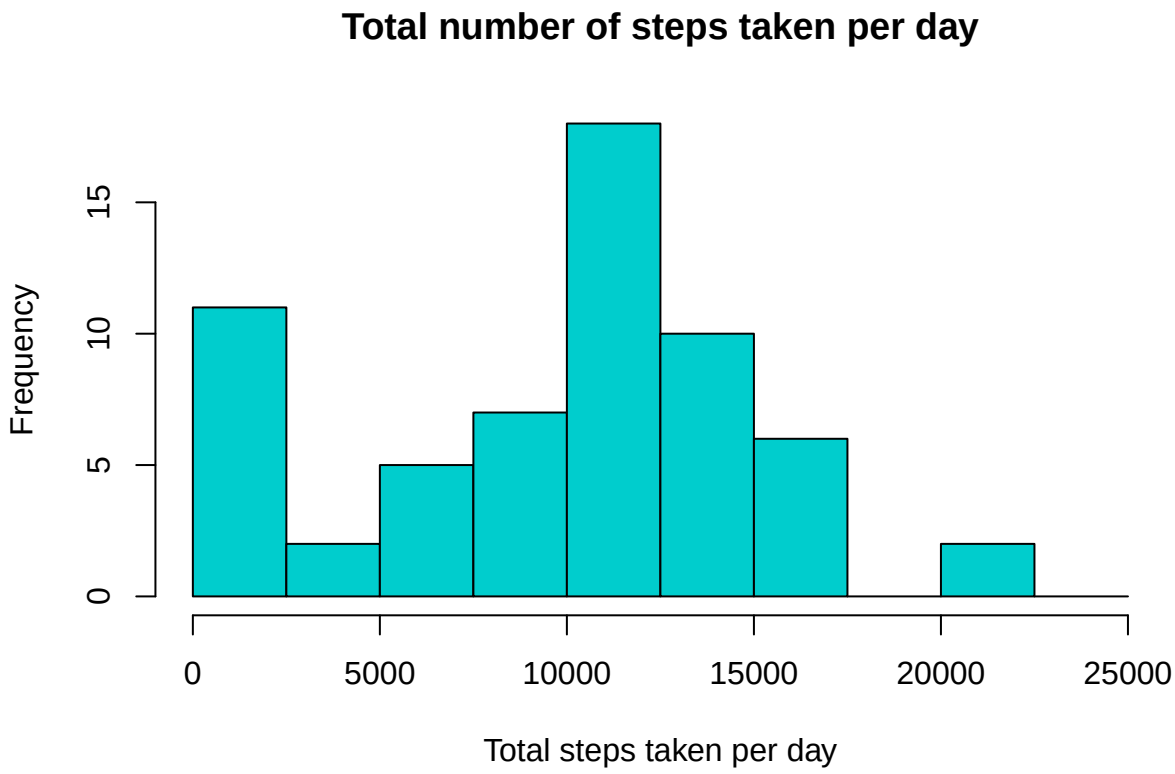
```
#Calculating the total number of steps per day
totalSteps <- with(Activity, aggregate(steps, by = list(date), FUN = sum, na.rm = TRUE))

names(totalSteps)<- c("dates", "steps")


head(totalSteps,7)
```

```
##        dates steps
```

```
## 1 2012-10-01     0
## 2 2012-10-02   126
## 3 2012-10-03 11352
## 4 2012-10-04 12116
## 5 2012-10-05 13294
## 6 2012-10-06 15420
## 7 2012-10-07 11015
```

2- Make a histogram of the total number of steps taken each day

```
hist(totalSteps$steps, main = "Total number of steps taken per day", xlab =
       "Total steps taken per day", col ="cyan3",breaks = seq(0,25000, by = 2500))
```



3- Calculate and report the mean and median of the total number of steps taken per day

```
mean(totalSteps$steps)
```

The mean

```
## [1] 9354.23
```

```
median(totalSteps$steps)
```

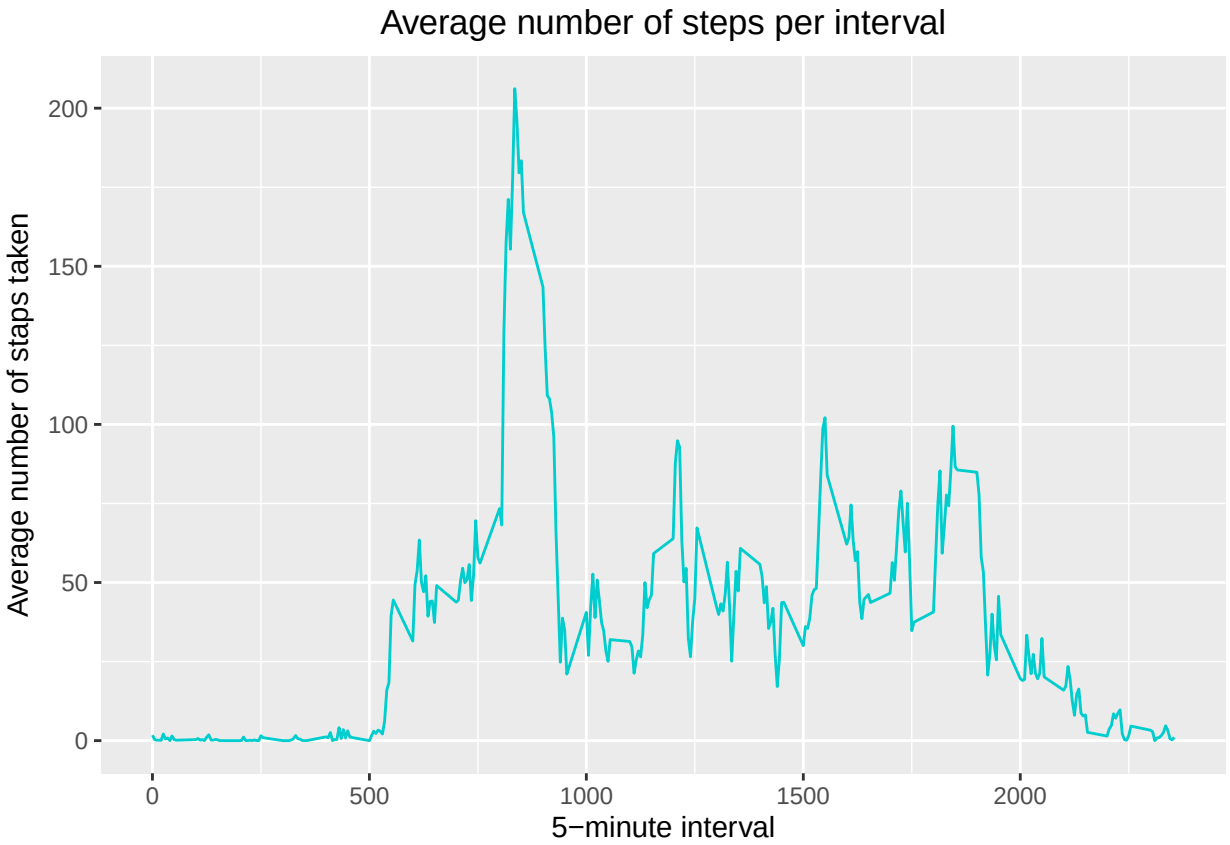
The median

```
## [1] 10395
```

What is the average daily activity pattern?

1- Make a time series plot of the 5-minute interval and the average number of steps taken, averaged across all days

```
stapes_int <- aggregate(steps ~ interval, Activity, mean, na.rm = TRUE)

#Plotting of the 5-minute interval and the average number of steps
ggplot(stapes_int, aes(interval, steps, col = "darkslategray1")) +
  geom_line(color='cyan3') +
  ggtitle("Average number of steps per interval") +
  ylab(expression("Average number of staps taken")) +
  xlab("5-minute interval") +
  theme(plot.title = element_text(hjust = 0.5))
```



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

```
stapes_int[which.max(stapes_int$steps),]$interval
```

```
## [1] 835
```

Imputing missing values

1- Calculate and report the total number of missing values in the dataset

```
sum(is.na(Activity$steps))
```

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

2- Fill the missing data (the mean for that 5-minute interval)

```
fill <- stapes_int$steps[match(Activity$interval,stapes_int$interval)]
```

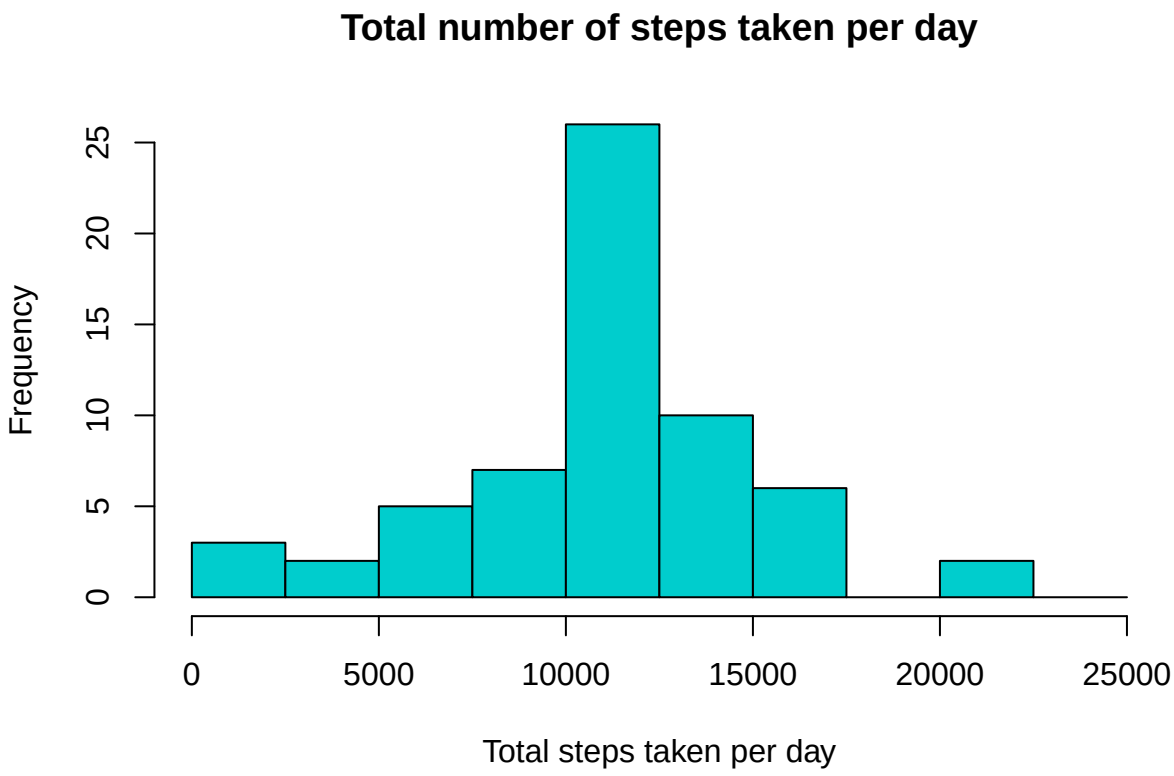
3- Create a new dataset that is equal to the original dataset but with the missing data filled in

```
activity.clean <- transform(Activity, steps = ifelse(is.na(Activity$steps),
                            yes=fill, no = Activity$steps))

head(activity.clean, 5)
```

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

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.

```
totalSteps.clean <- with(activity.clean, aggregate(steps, by = list(date), FUN = sum, na.rm = TRUE))

names(totalSteps.clean)<- c("dates", "steps")

hist(totalSteps.clean$steps, main = "Total number of steps taken per day", xlab =
      "Total steps taken per day", col ="cyan3",breaks = seq(0,25000, by = 2500))
```




```
mean(totalSteps.clean$steps)
```

The mean

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

```
median(totalSteps.clean$steps)
```

The median

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

Are there differences in activity patterns between weekdays and weekends?

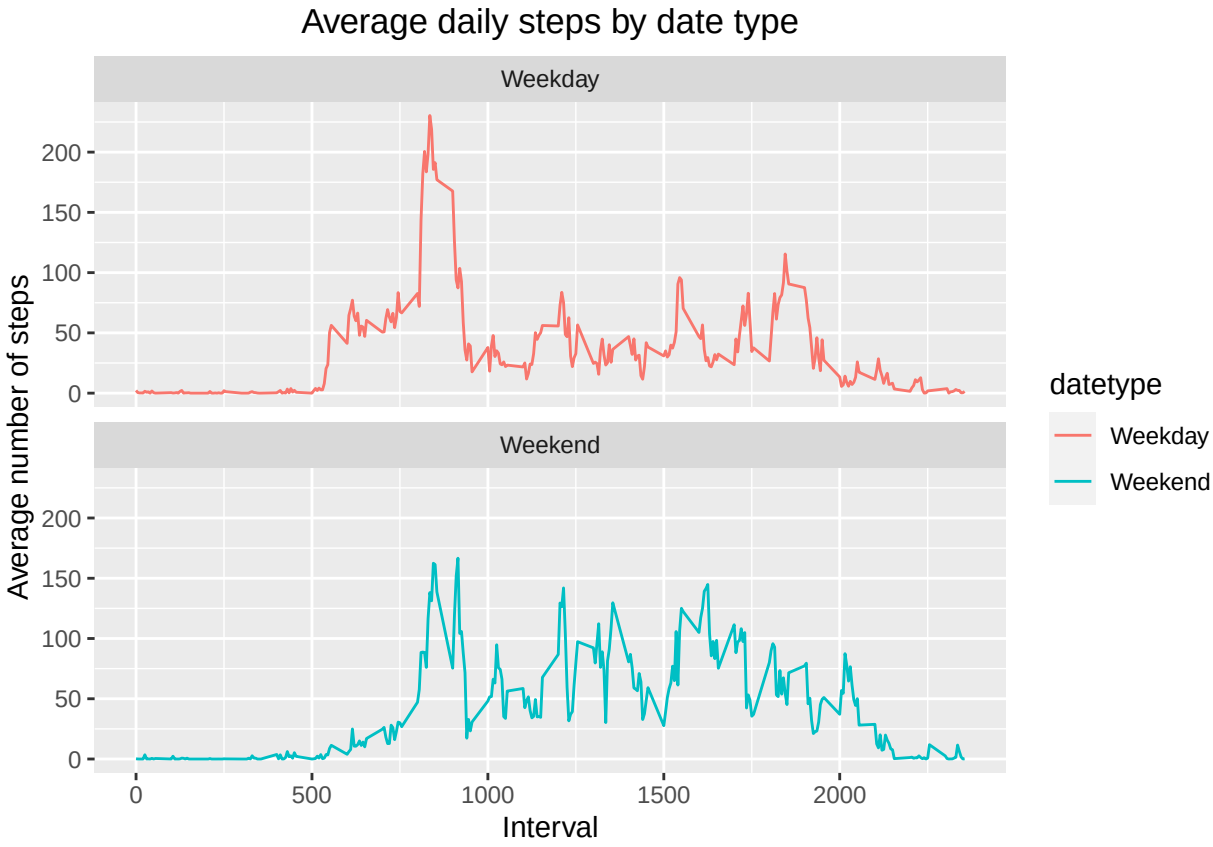
1- Create a new factor variable in the dataset with two levels – "weekday" and "weekend" indicating whether a given date is a weekday or weekend day

```
activity.clean$datetype <- sapply(activity.clean$date, function(x) {
        if (weekdays(x) == "Saturday" | weekdays(x) =="Sunday")
                {y <- "Weekend"}
        else
                {y <- "Weekday"}
        y
        })

head(activity.clean, 5)
```

```
##       steps       date interval weekday datetype
## 1 1.7169811 2012-10-01        0  Monday  Weekday
## 2 0.3396226 2012-10-01        5  Monday  Weekday
## 3 0.1320755 2012-10-01       10  Monday  Weekday
## 4 0.1509434 2012-10-01       15  Monday  Weekday
## 5 0.0754717 2012-10-01       20  Monday  Weekday
```

2- Make a panel plot containing a time series plot of the 5-minute interval and the average number of steps taken, averaged across all weekday days or weekend days

```
activity.datetype <- aggregate(steps~interval+datetype, activity.clean, mean)

ggplot(activity.datetype, aes(x = interval, y = steps, color = datetype))+
  geom_line() +
  labs(title = "Average daily steps by date type", x ="Interval", y = "Average number of steps") +
  facet_wrap(~datetype, ncol = 1, nrow = 2) +
  theme(plot.title = element_text(hjust = 0.5))
```


Average daily steps by date type
Weekday
Weekend
Average number of steps
0
50
100
150
200
0
500
1000
1500
2000
Interval
datetype
Weekday
Weekend