Reproducible Research (Data Science Specialization). Peer Assessment 1

Loading the data

```
setwd("C:\\Users\\Jose Manuel\\Copy\\Data Science Specialization\\ReproducibleResearch\\RepData_PeerAsse
unzip("activity.zip")
activity <- read.csv("activity.csv")
```

1. What is the mean total number of steps taken per day?

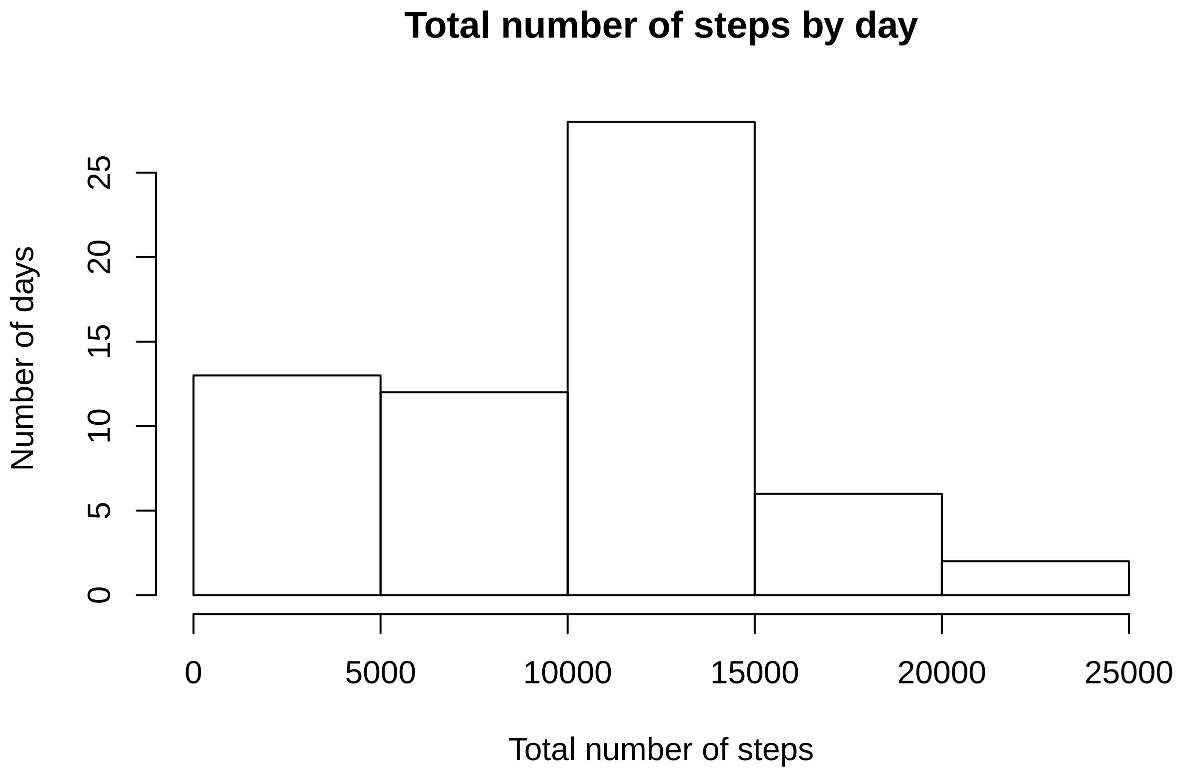
1.1 Total number of steps taken per day

```
Total.steps <- as.data.frame(tapply(activity$steps, activity$date, sum, na.rm=T))
colnames(Total.steps) <- "Total.steps"
Total.steps
```

```
##            Total.steps
## 2012-10-01           0
## 2012-10-02         126
## 2012-10-03       11352
## 2012-10-04       12116
## 2012-10-05       13294
## 2012-10-06       15420
## 2012-10-07       11015
## 2012-10-08           0
## 2012-10-09       12811
## 2012-10-10        9900
## 2012-10-11       10304
## 2012-10-12       17382
## 2012-10-13       12426
## 2012-10-14       15098
## 2012-10-15       10139
## 2012-10-16       15084
## 2012-10-17       13452
## 2012-10-18       10056
## 2012-10-19       11829
## 2012-10-20       10395
## 2012-10-21        8821
## 2012-10-22       13460
## 2012-10-23        8918
## 2012-10-24        8355
## 2012-10-25        2492
## 2012-10-26        6778
## 2012-10-27       10119
## 2012-10-28       11458
## 2012-10-29        5018
## 2012-10-30        9819
## 2012-10-31       15414
## 2012-11-01           0
## 2012-11-02       10600
## 2012-11-03       10571
## 2012-11-04           0
## 2012-11-05       10439
## 2012-11-06        8334
## 2012-11-07       12883
## 2012-11-08        3219
## 2012-11-09           0
## 2012-11-10           0
## 2012-11-11       12608
## 2012-11-12       10765
## 2012-11-13        7336
## 2012-11-14           0
## 2012-11-15          41
## 2012-11-16        5441
## 2012-11-17       14339
## 2012-11-18       15110
## 2012-11-19        8841
## 2012-11-20        4472
## 2012-11-21       12787
## 2012-11-22       20427
## 2012-11-23       21194
## 2012-11-24       14478
## 2012-11-25       11834
## 2012-11-26       11162
## 2012-11-27       13646
## 2012-11-28       10183
## 2012-11-29        7047
## 2012-11-30           0
```

1.2. Histogram of the total number of steps taken each day

```
hist(Total.steps$Total.steps, xlab="Total number of steps", ylab="Number of days",
     main="Total number of steps by day")
```



1.3. Mean and median of the total number of steps taken per day

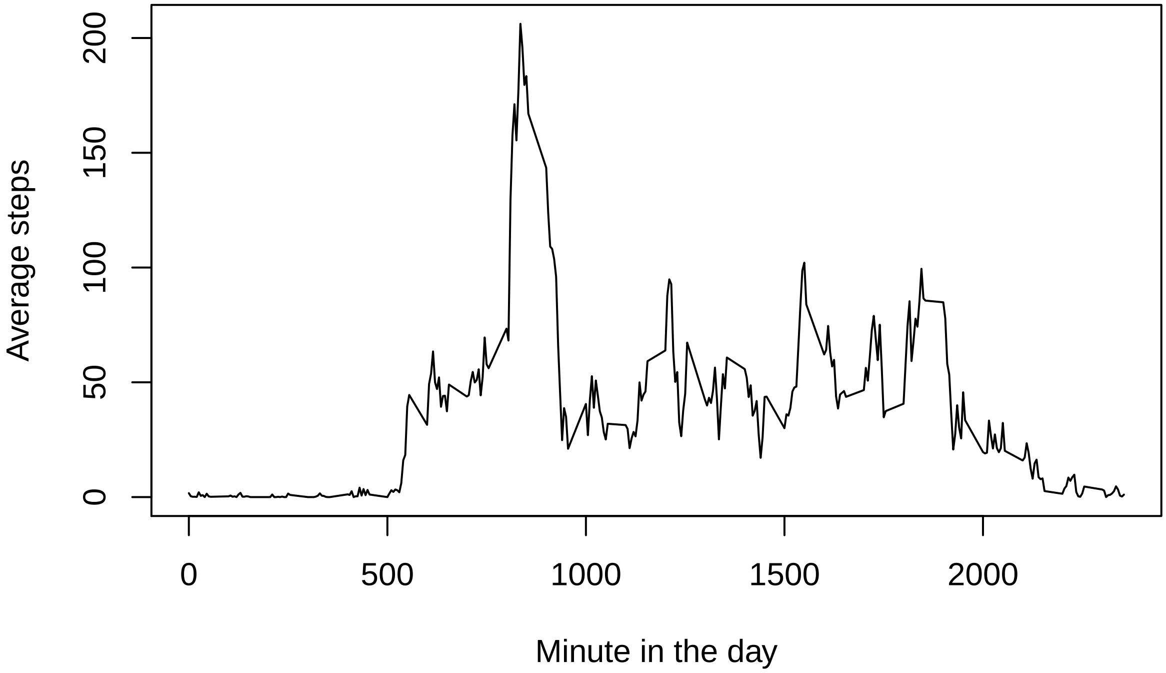
```
Mean <- unlist(tapply(activity$steps, activity$date, mean, na.rm=T))
Median <- unlist(tapply(activity$steps, activity$date, median, na.rm=T))
Summary.steps <- cbind(Mean, Median)
Summary.steps
```

```
##                   Mean Median
## 2012-10-01         NaN     NA
## 2012-10-02  0.4375000      0
## 2012-10-03 39.4166667      0
## 2012-10-04 42.0694444      0
## 2012-10-05 46.1597222      0
## 2012-10-06 53.5416667      0
## 2012-10-07 38.2465278      0
## 2012-10-08         NaN     NA
## 2012-10-09 44.4826389      0
## 2012-10-10 34.3750000      0
## 2012-10-11 35.7777778      0
## 2012-10-12 60.3541667      0
## 2012-10-13 43.1458333      0
## 2012-10-14 52.4236111      0
## 2012-10-15 35.2048611      0
## 2012-10-16 52.3750000      0
## 2012-10-17 46.7083333      0
## 2012-10-18 34.9166667      0
## 2012-10-19 41.0729167      0
## 2012-10-20 36.0937500      0
## 2012-10-21 30.6284722      0
## 2012-10-22 46.7361111      0
## 2012-10-23 30.9652778      0
## 2012-10-24 29.0104167      0
## 2012-10-25  8.6527778      0
## 2012-10-26 23.5347222      0
## 2012-10-27 35.1354167      0
## 2012-10-28 39.7847222      0
## 2012-10-29 17.4236111      0
## 2012-10-30 34.0937500      0
## 2012-10-31 53.5208333      0
## 2012-11-01        NaN     NA
## 2012-11-02 36.8055556      0
## 2012-11-03 36.7048611      0
## 2012-11-04        NaN     NA
## 2012-11-05 36.2465278      0
## 2012-11-06 28.9375000      0
## 2012-11-07 44.7326389      0
## 2012-11-08 11.1770833      0
## 2012-11-09        NaN     NA
## 2012-11-10        NaN     NA
## 2012-11-11 43.7777778      0
## 2012-11-12 37.3784722      0
## 2012-11-13 25.4722222      0
## 2012-11-14        NaN     NA
## 2012-11-15  0.1423611      0
## 2012-11-16 18.8923611      0
## 2012-11-17 49.7881944      0
## 2012-11-18 52.4652778      0
## 2012-11-19 30.6979167      0
## 2012-11-20 15.5277778      0
## 2012-11-21 44.3993056      0
## 2012-11-22 70.9270833      0
## 2012-11-23 73.5902778      0
## 2012-11-24 50.2708333      0
## 2012-11-25 41.0902778      0
## 2012-11-26 38.7569444      0
## 2012-11-27 47.3819444      0
## 2012-11-28 35.3576389      0
## 2012-11-29 24.4687500      0
## 2012-11-30        NaN     NA
```

2. What is the average daily activity pattern?

2.1. Time series plot of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis)

```
steps.interval <- tapply(activity$steps, activity$interval, mean, na.rm=T)
plot(unique(activity$interval), steps.interval, type="l", xlab="Minute in the day",
     ylab="Average steps")
```



2.2. 5-minute interval, on average across all the days in the dataset, containing the maximum number of steps

```
unique(activity$interval)[which.max(steps.interval)]
```

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

3. Imputing missing values

3.1. Total number of missing values in the dataset

```
sum(is.na(activity))
```

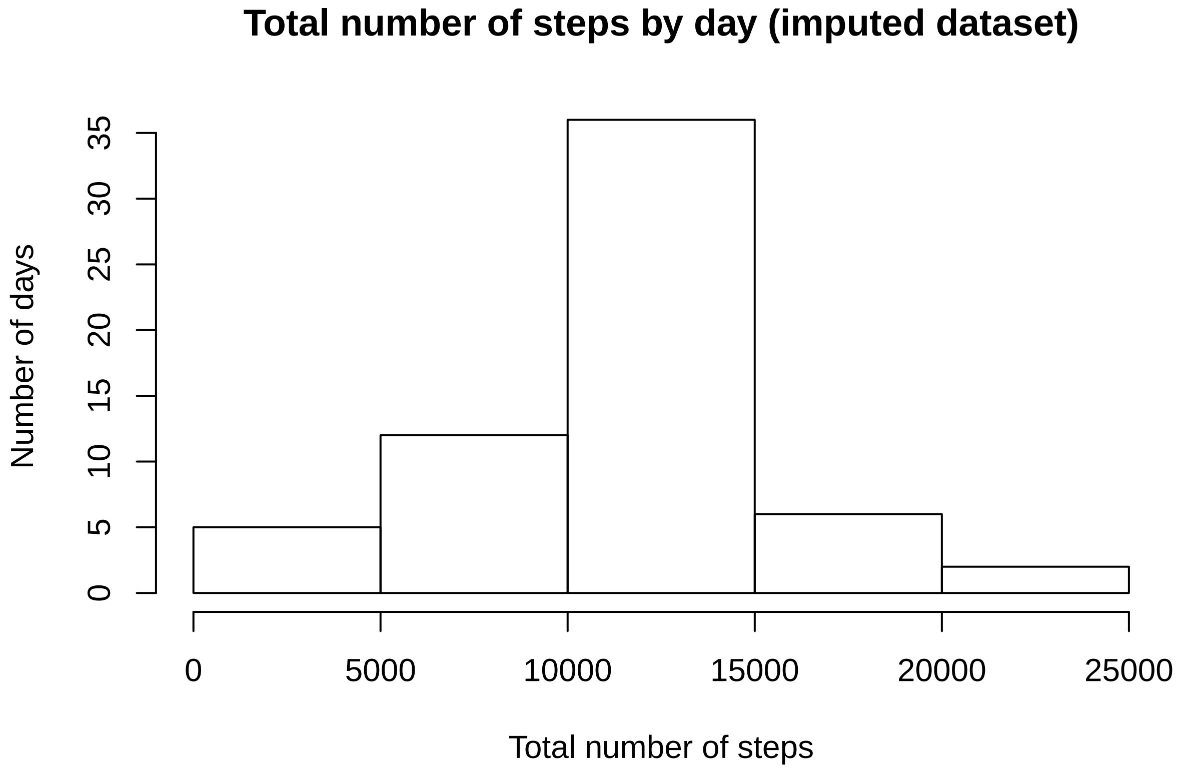
```
## [1] 2304
```

3.2 and 3.3. Filling in all of the missing values in the dataset using the mean of the corresponding 5-minute interval. A new dataset (activity.new) that is equal to the original dataset but with the missing data filled is created.

```r
steps.interval <- rep(steps.interval, length(unique(activity$date)))
activity.new <- activity
activity.new$steps[is.na(activity$steps)] <- steps.interval[is.na(activity$steps)]
```

3.4a. Histogram of the total number of steps taken each day in the imputed dataset

```r
Total.steps.new <- as.data.frame(tapply(activity.new$steps, activity$date, sum, na.rm=T))
hist(Total.steps.new[,1], xlab="Total number of steps", ylab="Number of days",
     main="Total number of steps by day (imputed dataset)")
```



3.4b. Mean and median of the total number of steps taken per day (imputed dataset).

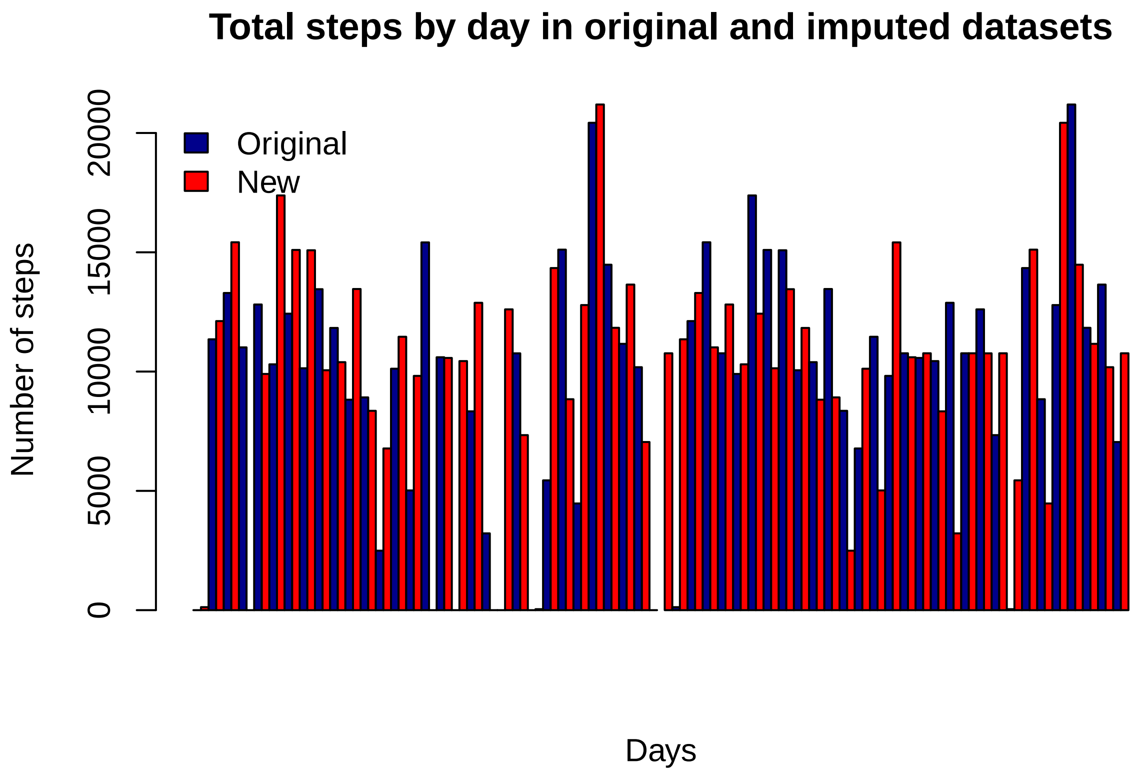
```r
Mean <- unlist(tapply(activity.new$steps, activity$date, mean, na.rm=T))
Median <- unlist(tapply(activity.new$steps, activity$date, median, na.rm=T))
Summary.steps <- cbind(Mean, Median)
Summary.steps
```

```
##                  Mean   Median
## 2012-10-01 37.3825996 34.11321
## 2012-10-02  0.4375000  0.00000
## 2012-10-03 39.4166667  0.00000
## 2012-10-04 42.0694444  0.00000
## 2012-10-05 46.1597222  0.00000
## 2012-10-06 53.5416667  0.00000
## 2012-10-07 38.2465278  0.00000
## 2012-10-08 37.3825996 34.11321
## 2012-10-09 44.4826389  0.00000
## 2012-10-10 34.3750000  0.00000
## 2012-10-11 35.7777778  0.00000
## 2012-10-12 60.3541667  0.00000
## 2012-10-13 43.1458333  0.00000
## 2012-10-14 52.4236111  0.00000
## 2012-10-15 35.2048611  0.00000
## 2012-10-16 52.3750000  0.00000
## 2012-10-17 46.7083333  0.00000
## 2012-10-18 34.9166667  0.00000
## 2012-10-19 41.0729167  0.00000
## 2012-10-20 36.0937500  0.00000
## 2012-10-21 30.6284722  0.00000
## 2012-10-22 46.7361111  0.00000
## 2012-10-23 30.9652778  0.00000
## 2012-10-24 29.0104167  0.00000
## 2012-10-25  8.6527778  0.00000
## 2012-10-26 23.5347222  0.00000
## 2012-10-27 35.1354167  0.00000
## 2012-10-28 39.7847222  0.00000
## 2012-10-29 17.4236111  0.00000
## 2012-10-30 34.0937500  0.00000
## 2012-10-31 53.5208333  0.00000
## 2012-11-01 37.3825996 34.11321
## 2012-11-02 36.8055556  0.00000
## 2012-11-03 36.7048611  0.00000
## 2012-11-04 37.3825996 34.11321
## 2012-11-05 36.2465278  0.00000
## 2012-11-06 28.9375000  0.00000
## 2012-11-07 44.7326389  0.00000
## 2012-11-08 11.1770833  0.00000
## 2012-11-09 37.3825996 34.11321
## 2012-11-10 37.3825996 34.11321
## 2012-11-11 43.7777778  0.00000
## 2012-11-12 37.3784722  0.00000
## 2012-11-13 25.4722222  0.00000
## 2012-11-14 37.3825996 34.11321
## 2012-11-15  0.1423611  0.00000
## 2012-11-16 18.8923611  0.00000
## 2012-11-17 49.7881944  0.00000
## 2012-11-18 52.4652778  0.00000
## 2012-11-19 30.6979167  0.00000
## 2012-11-20 15.5277778  0.00000
## 2012-11-21 44.3993056  0.00000
## 2012-11-22 70.9270833  0.00000
## 2012-11-23 73.5902778  0.00000
## 2012-11-24 50.2708333  0.00000
## 2012-11-25 41.0902778  0.00000
## 2012-11-26 38.7569444  0.00000
## 2012-11-27 47.3819444  0.00000
## 2012-11-28 35.3576389  0.00000
## 2012-11-29 24.4687500  0.00000
## 2012-11-30 37.3825996 34.11321
```

3.4c. 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?

As can be observed comparing both histrograms, the imputation has a great impact on the distribution of steps. And also on the total daily number of steps, as can be observed in the next figure:

```
barplot(cbind(Total.steps[,1],Total.steps.new[,1]), ylab="Number of steps", xlab="Days",
        main="Total steps by day in original and imputed datasets",  col=c("darkblue","red"),
        beside=TRUE, legend = c("Original", "New"), args.legend = list(x = "topleft", bty = "n"))
```



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

4.1. A new factor variable in the dataset with two levels, “weekday” and “weekend” indicating whether a given date is a weekday or weekend day, is created

```
Sys.setlocale("LC_TIME", "English")
```

```
## [1] "English_United States.1252"
```

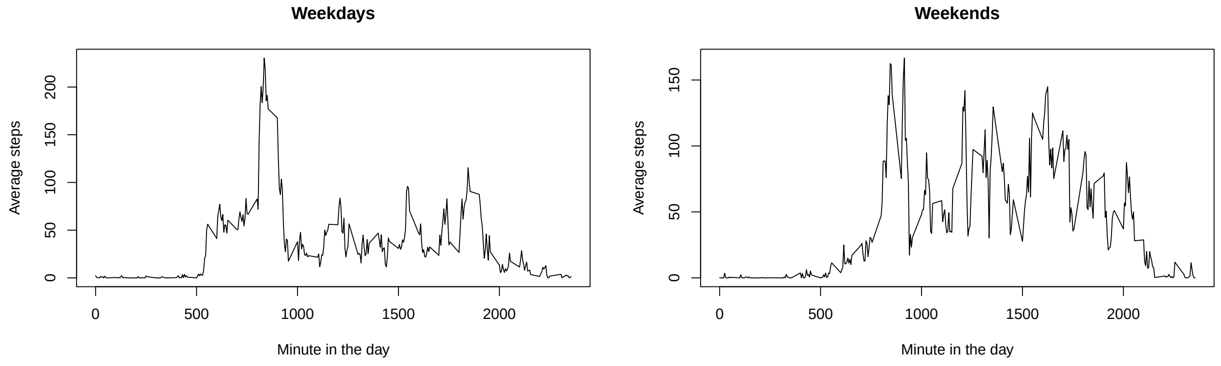
```
activity.new$daytype <- weekdays(as.Date(activity.new$date))
days.type <- function(x) ifelse(x %in% c("Sunday","Saturday"), "weekend", "weekday")
activity.new$daytype <- as.factor(sapply(activity.new$daytype,days.type))
```

4.2. Panel plot containing a time series plot of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all weekday days and weekend days (y-axis).

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

base.temp <- activity.new[activity.new$daytype=="weekday",]
steps.interval <- tapply(base.temp$steps, base.temp$interval, mean)
plot(unique(base.temp$interval), steps.interval, type="l", xlab="Minute in the day",
     ylab="Average steps", main="Weekdays")

base.temp <- activity.new[activity.new$daytype=="weekend",]
steps.interval <- tapply(base.temp$steps, base.temp$interval, mean)
plot(unique(base.temp$interval), steps.interval, type="l", xlab="Minute in the day",
     ylab="Average steps", main="Weekends")
```


Weekdays
Average steps
Minute in the day
Weekends
Average steps
Minute in the day