Reproducible Research: Peer Assessment 1

Hatem Jasim Hatem

May 20, 2019

```
library(dplyr)
library(lattice)
```

Loading and preprocessing the data

1. Load the data

```
filename <- "repdata_data_activity.zip"

if (!file.exists(filename)){
  fileURL <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip"
  download.file(fileURL, filename, method="curl")
}

if (!file.exists("repdata_data_activity")) {
  unzip(filename)
}

activity <- read.csv("repdata_data_activity/activity.csv")
```

2. Process/transform the data into a format suitable for your analysis

```
activity$date <- as.Date(activity$date)
activity$day <- weekdays(activity$date)

head(activity)
```

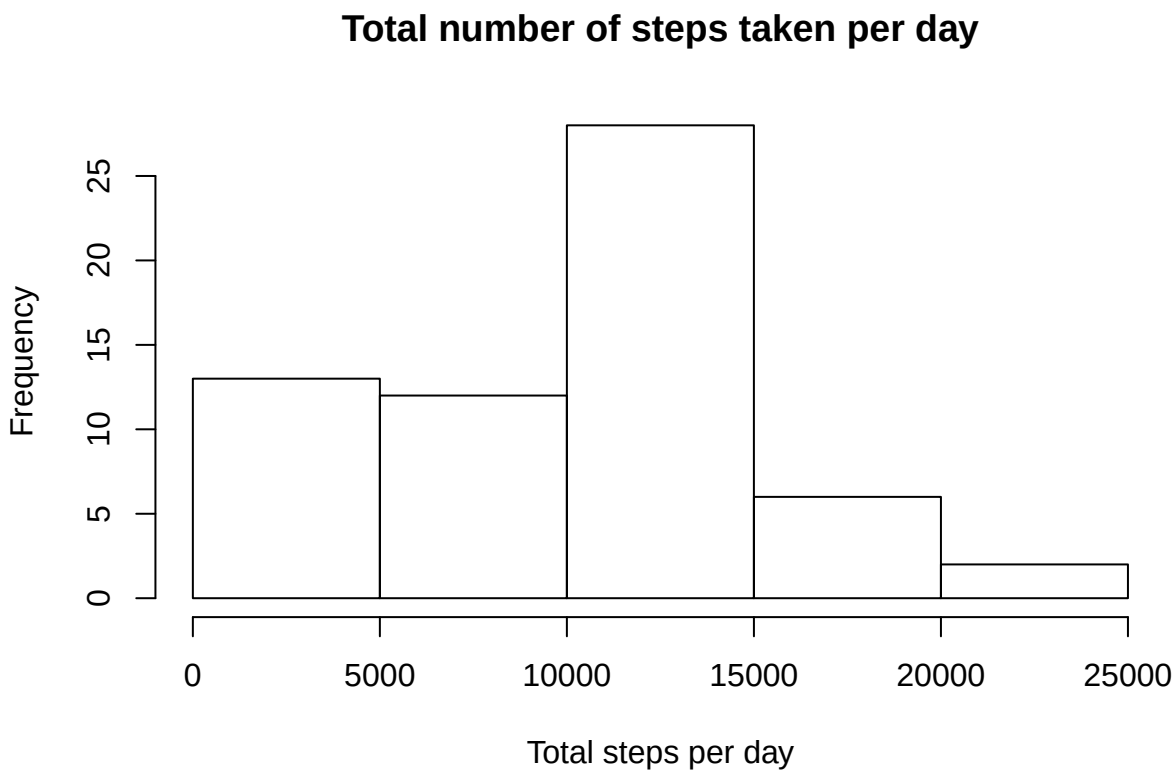
```
##   steps       date interval    day
## 1    NA 2012-10-01        0 Monday
## 2    NA 2012-10-01        5 Monday
## 3    NA 2012-10-01       10 Monday
## 4    NA 2012-10-01       15 Monday
## 5    NA 2012-10-01       20 Monday
## 6    NA 2012-10-01       25 Monday
```

What is mean total number of steps taken per day?

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

```
T.S.P.D<-activity%>% group_by(date)%>%
    summarise(steps= sum(steps, na.rm = TRUE))
```

```
hist(T.S.P.D$steps,
     main = "Total number of steps taken per day",
     xlab = "Total steps per day")
```



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

```
mean(T.S.P.D$steps)
```

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

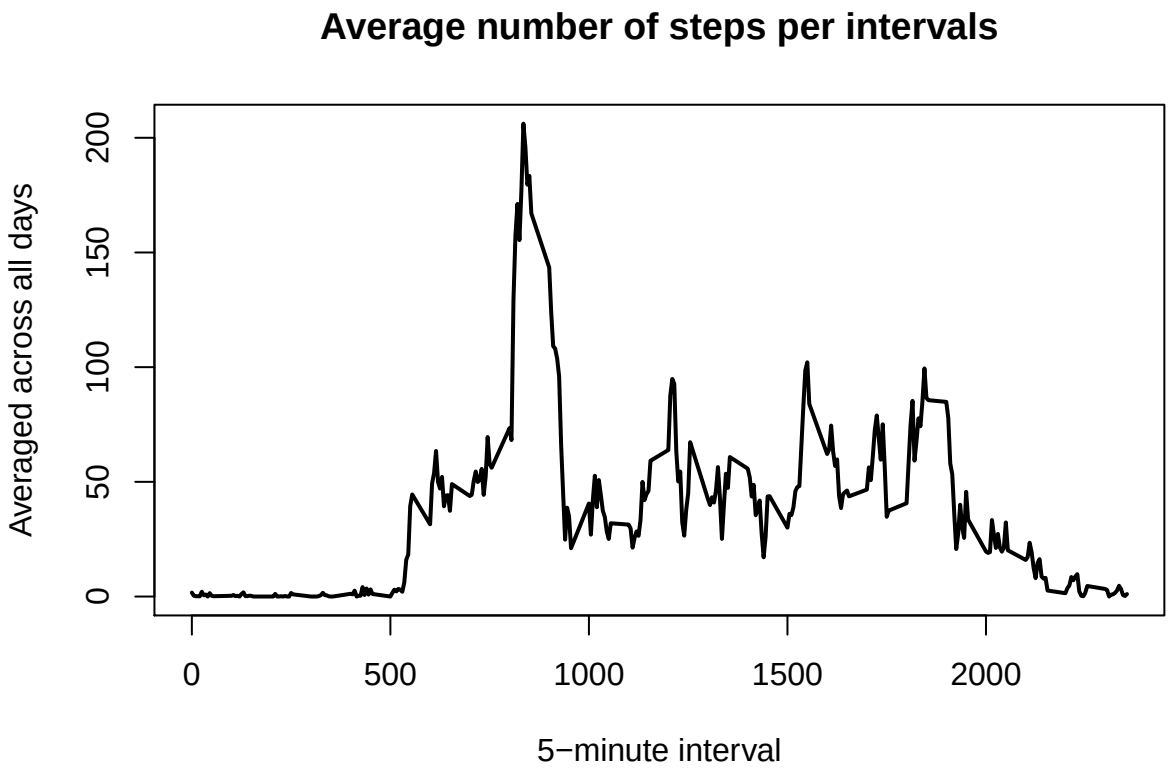
```
median(T.S.P.D$steps)
```

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

What is the average daily activity pattern?

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

```r
A.D.A <- activity %>% group_by(interval)%>%
    summarise(mean= mean(steps, na.rm = TRUE))


plot(A.D.A$interval, A.D.A$mean,
     type = "l", lwd = 2, xlab="5-minute interval",
     ylab="Averaged across all days",
     main="Average number of steps per intervals")
```



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

```r
activity %>% group_by(interval) %>%
  summarize(step_mean = mean(steps, na.rm = TRUE)) %>%
  filter(step_mean == max(step_mean))
```

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

Imputing missing values

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

```r
sum(is.na(activity$steps))
```

```
## [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, I use the mean for that day.

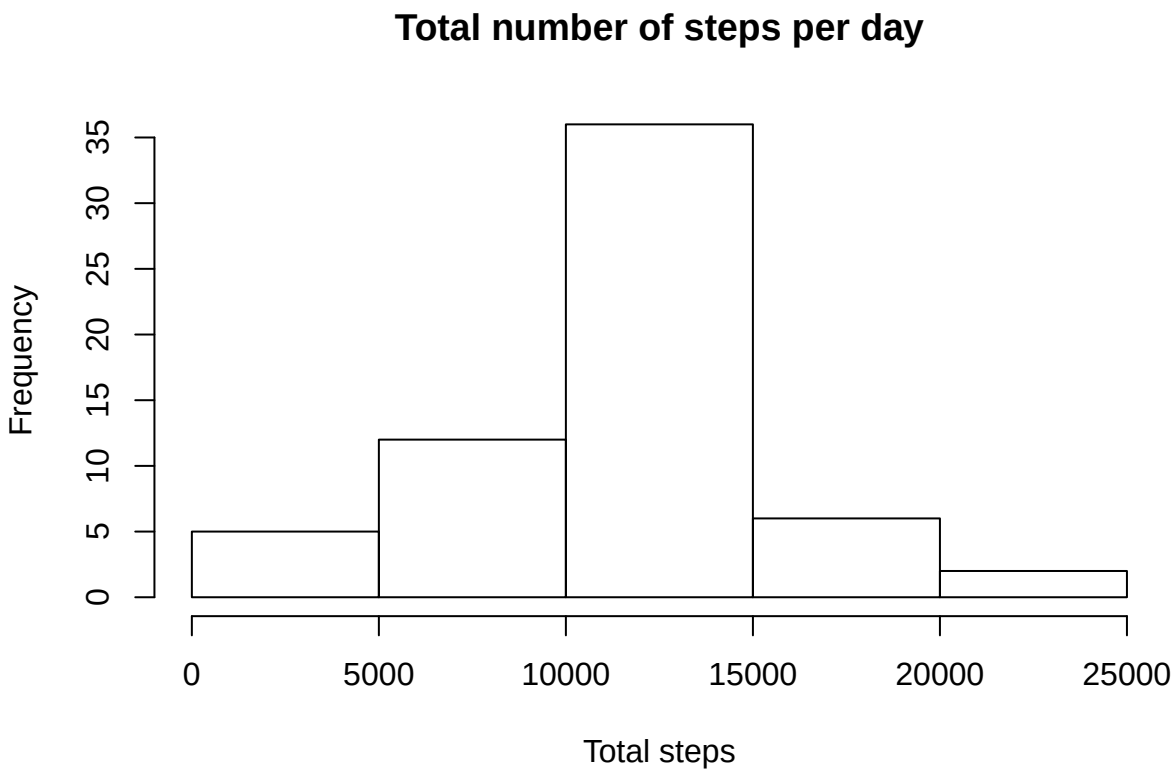
```r
mean_missing <- mean(activity$steps, na.rm = TRUE)

activity_replace <- activity %>%
            mutate(replace_steps  = ifelse(is.na(steps),
                                           mean_missing, steps))
```

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

```r
activity_new<-activity_replace%>% group_by(date)%>%
    summarise(steps= sum(replace_steps, na.rm = TRUE))
```

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
hist(activity_new$steps,
     xlab = "Total steps",
     main = "Total number of steps per day")
```




```
mean(activity_new$steps)
```

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

```
median(activity_new$steps)
```

```
## [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_replace$DayType <- ifelse(activity_replace$day
                                   %in% c("Saturday", "Sunday"),
                                   "Weekend", "Weekday")

A.p.I.a.Dt <- activity_replace%>% group_by(interval, DayType)%>%
                              summarize(mean = mean(steps, na.rm = TRUE))
```

2. Make a panel plot containing a time series plot (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). The plot should look something like the following, which was created using simulated data:

```
xyplot(mean~interval|DayType, data=A.p.I.a.Dt, type="l",
       layout = c(1,2),
       main="Average Steps per Interval Based on Type of Day",
       ylab="Average Number of Steps", xlab="5-minute interval")
```