Reproducible Research: Assignment 1

Contents

Coursera - Reproducible Research - Assignment 1 **1**

Coursera - Reproducible Research - Assignment 1

Author: Alexander Barrantes Herrera

Date: April 25, 2020

Download and unzip the dataset

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

Loading and preprocessing the data

```
act <- read.csv("activity.csv")
act$date <- as.Date(act$date, format = "%Y-%m-%d")
```

What is mean total number of steps taken per day?

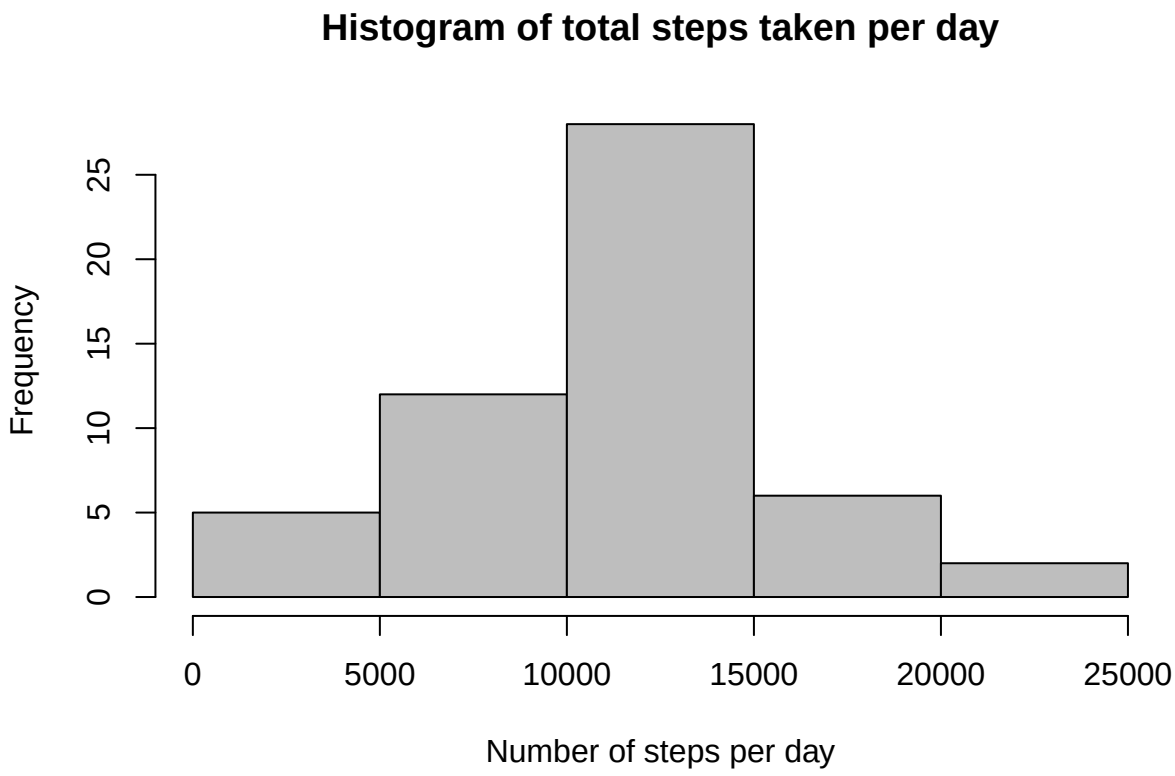
1. Calculate the total number of steps taken per day

```
steps_per_day <- aggregate(steps ~ date, data = act, FUN = "sum", na.rm = TRUE)
head(steps_per_day)
```

```
##         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
```

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

```
hist(steps_per_day$steps, xlab = "Number of steps per day", main ="Histogram of total steps taken per da
```



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

```
mean <- format(mean(steps_per_day$steps), 2)
mean
```

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

```
median <- format(median(steps_per_day$steps), 2)
median
```

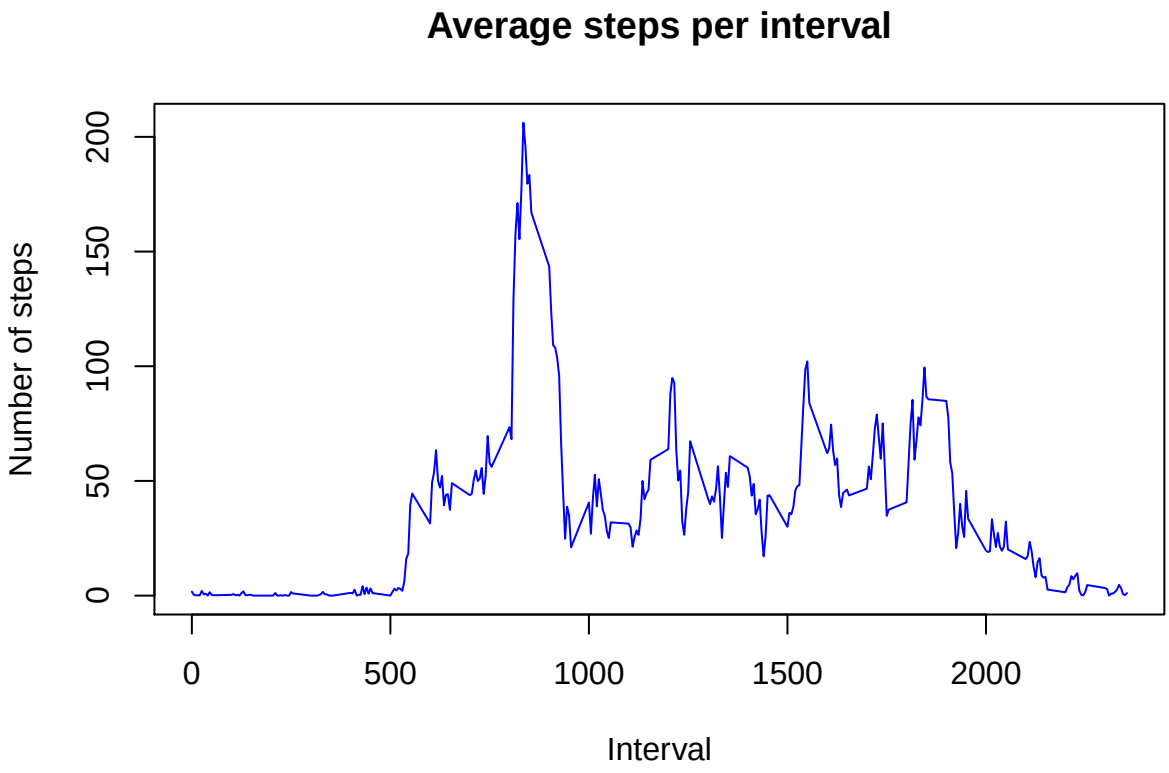
```
## [1] "10765"
```

- Mean: **10766.19**
- Median: **10765**

What is the average daily activity pattern?

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

```
int <- aggregate(steps ~ interval, data=act, FUN = "mean", na.rm=TRUE)
with(int, plot(interval, steps, type = "l", xlab = "Interval", ylab = "Number of steps", main = "Averag
```



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

```
index_for_max <- which(int$steps==max(int$steps))
max_steps <- int$steps[index_for_max]
max_interval <- int$interval[index_for_max]
```

- Highest average number of steps per day: **206.1698113**
- Interval with the highest average number of steps: **835**

Imputing missing values with the imputeTS package (package needs to be installed first)

1. Calculate and report the total number of missing values in the dataset (i.e. the total number of rows with NAs)

```
tot_na <- sum(is.na(act$steps))
tot_na
```

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

- Total number of rows with NA's: **2304**

2. Devise a strategy for filling in all of the missing values in the dataset. 3. Create a new dataset that is equal to the original dataset but with the missing data filled in.

```
##Load, and if necessary install package "imputeTS": install.packages("imputeTS")
library(imputeTS)
```

```
## Warning: package 'imputeTS' was built under R version 3.6.3
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
## Use the function na.mean() to replace the missing values with the mean value across all intervals
act_imputed <- na.mean(act)
```

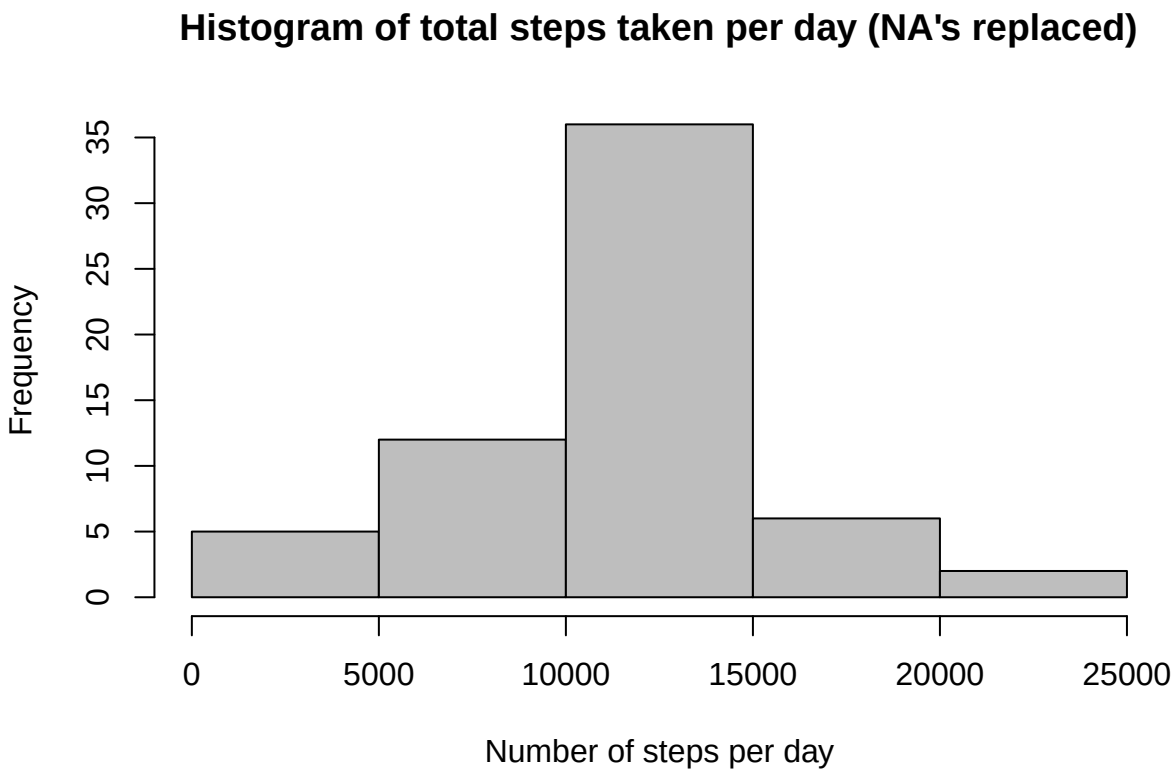
```
## Warning: na.mean will replaced by na_mean.
##     Functionality stays the same.
##     The new function name better fits modern R code style guidelines.
##     Please adjust your code accordingly.
```

```
head(act_imputed)
```

```
##     steps       date interval
## 1 37.3826 2012-10-01        0
## 2 37.3826 2012-10-01        5
## 3 37.3826 2012-10-01       10
## 4 37.3826 2012-10-01       15
## 5 37.3826 2012-10-01       20
## 6 37.3826 2012-10-01       25
```

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?

```
steps_new <- aggregate(steps ~ date, data = act_imputed, FUN = "sum", na.rm = TRUE)
hist(steps_new$steps, xlab = "Number of steps per day", main ="Histogram of total steps taken per day (
```




```
mean_new <- format(mean(steps_new$steps))
mean_new
```

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

```
median_new <- format(median(steps_new$steps), 2)
median_new
```

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

Difference between the original mean/median and the new mean/median (NA's replaced with the mean)

- Original Mean: **10766.19**
- New Mean (NA's replaced): **10766.19**
- Original Median: **10765**
- New Median (NA's replaced): **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.

```r
## note that I used the spanish words for saturday and sunday, as the language settings of my system ar
act_imputed$daytype <- ifelse(weekdays(act_imputed$date)== "domingo" | weekdays(act_imputed$date)== "s?
head(act_imputed)
```

```
##     steps       date interval  daytype
## 1 37.3826 2012-10-01        0 weekdays
## 2 37.3826 2012-10-01        5 weekdays
## 3 37.3826 2012-10-01       10 weekdays
## 4 37.3826 2012-10-01       15 weekdays
## 5 37.3826 2012-10-01       20 weekdays
## 6 37.3826 2012-10-01       25 weekdays
```

2. 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
library(lattice)
steps_imputed <- aggregate(steps ~ interval + daytype, data=act_imputed, FUN = "mean", na.rm=TRUE)
xyplot(steps_imputed$steps ~ steps_imputed$interval | daytype, data = steps_imputed, type="l", ylab = "
```