PA1_template

Anshul Shah

12/12/2019

1. Calculate the total number of steps taken per day

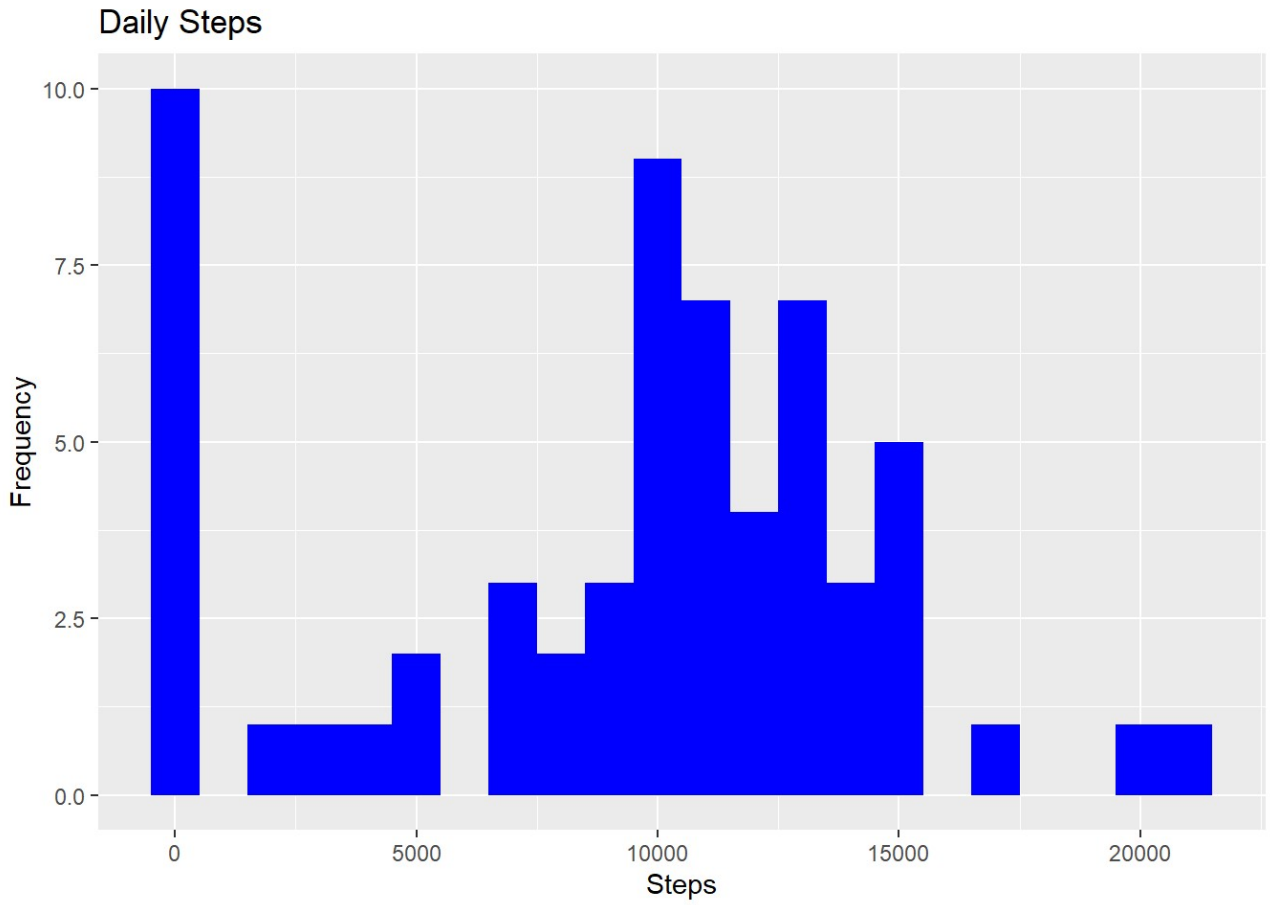
```
library(data.table)
library(ggplot2)
activityDT <- data.table::fread(input = "activity.csv")

Total_Steps <- activityDT[, c(lapply(.SD, sum, na.rm = TRUE)), .SDcols = c("steps"), b
y = .(date)]

head(Total_Steps, 10)
```

```
##           date 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
##  8: 2012-10-08     0
##  9: 2012-10-09 12811
## 10: 2012-10-10  9900
```

2. If you do not understand the difference between a histogram and a barplot, research the difference between them. Make a histogram of the total number of steps taken each day.



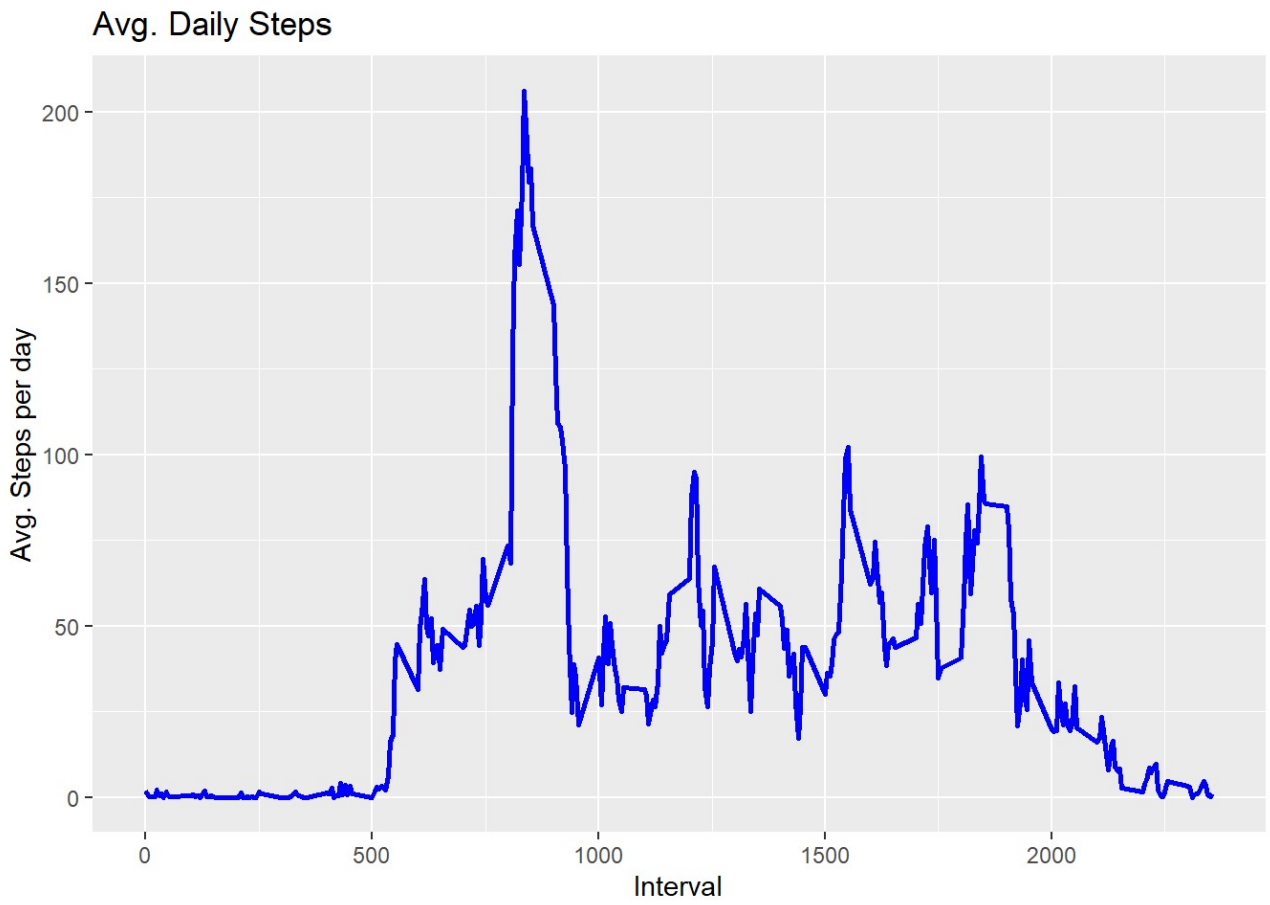
3. Calculate the total number of steps taken per day

```
Total_Steps[, .(Mean_Steps = mean(steps), Median_Steps = median(steps))]
```

```
##    Mean_Steps Median_Steps
## 1:    9354.23        10395
```

---

1. Make 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 days (y-axis)



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

```
IntervalDT[steps == max(steps), .(max_interval = interval)]
```

```
##    max_interval
## 1:          835
```

---

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

```
activityDT[is.na(steps), .N ]
```

```
## [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, you could use the mean/median for that day, or the mean for that 5-minute interval, etc.

```
activityDT[is.na(steps), "steps"] <- round(activityDT[, c(lapply(.SD, mean, na.rm = TR
UE)), .SDcols = c("steps")])
```

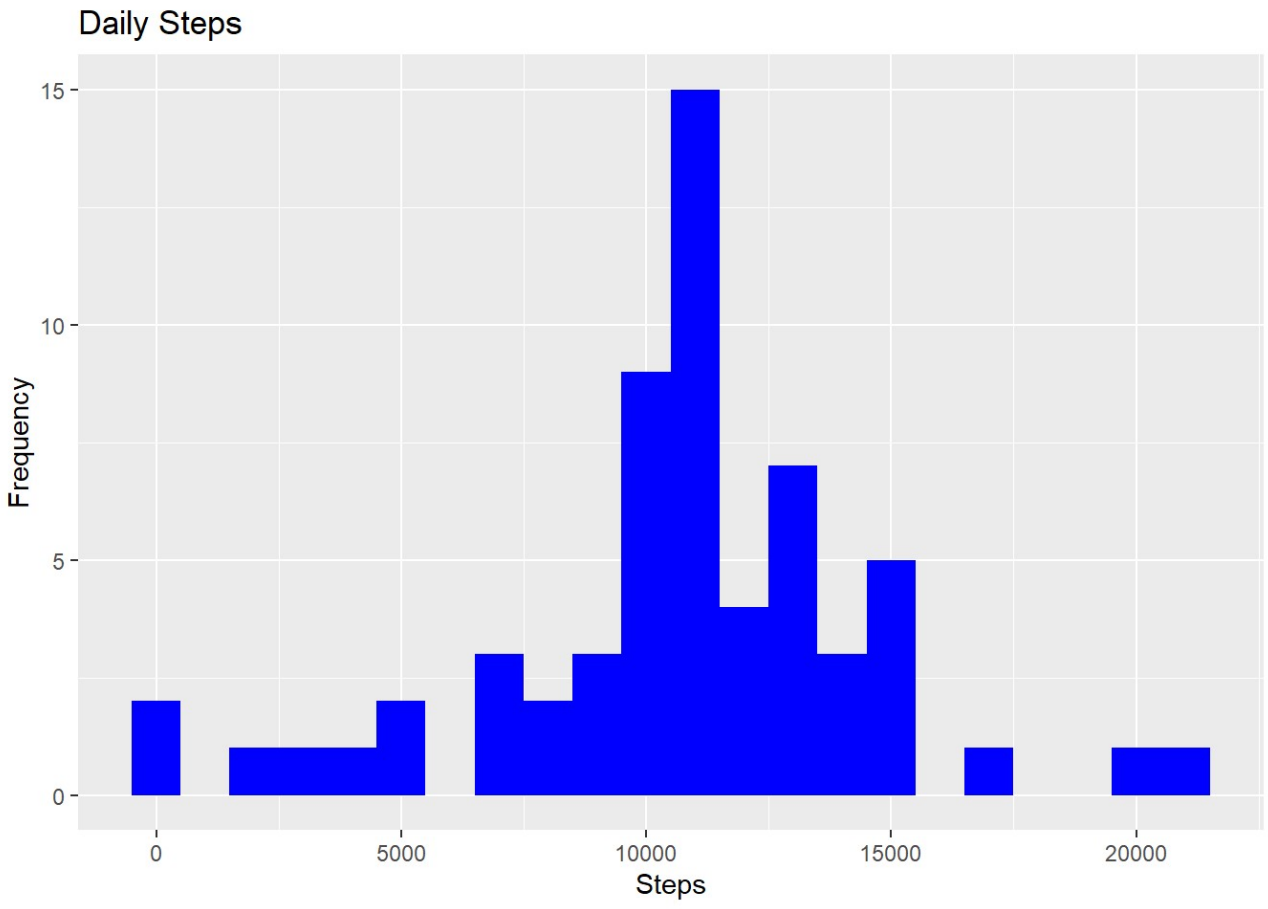
4. Make a histogram of the total number of steps taken each day and calculate and report the mean and median total number of steps taken per day. Do these values differ from the estimates from the first part of the assignment? What is the impact of imputing missing data on the estimates of the total daily number of steps?

```
Total_Steps <- activityDT[, c(lapply(.SD, sum, na.rm = TRUE)), .SDcols = c("steps"), b
y = .(date)]
```

```
Total_Steps[, .(Mean_Steps = mean(steps), Median_Steps = median(steps))]
```

```
##    Mean_Steps Median_Steps
## 1:   10751.74        10656
```

```
library(ggplot2)
ggplot(Total_Steps, aes(x = steps)) +
    geom_histogram(fill = "blue", binwidth = 1000) +
    labs(title = "Daily Steps", x = "Steps", y = "Frequency")
```



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.

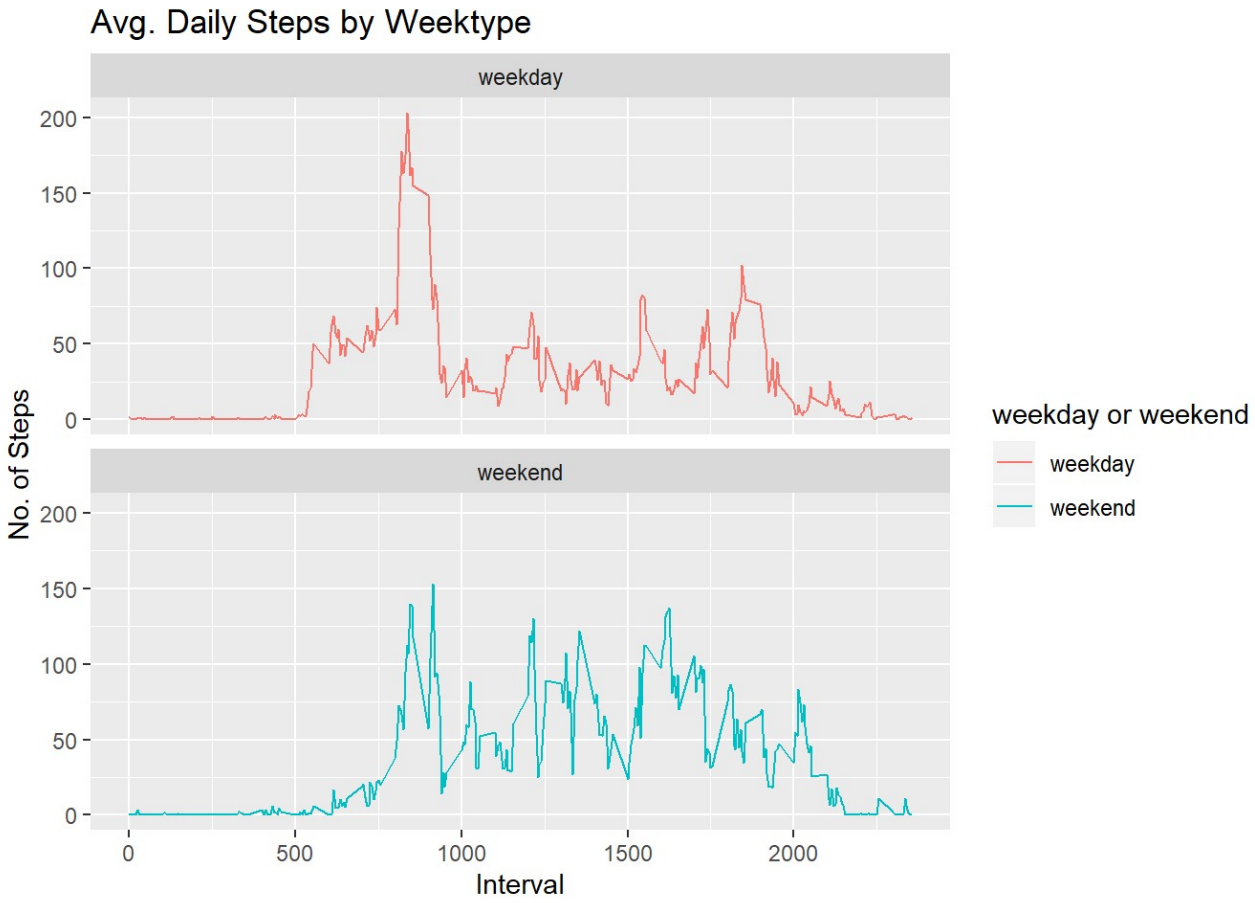
```
activityDT <- data.table::fread(input = "activity.csv")
activityDT[, date := as.POSIXct(date, format = "%Y-%m-%d")]
activityDT[, `Day of Week`:= weekdays(x = date)]
activityDT[grepl(pattern = "Monday|Tuesday|Wednesday|Thursday|Friday", x = `Day of Wee
k`), "weekday or weekend"] <- "weekday"
activityDT[grepl(pattern = "Saturday|Sunday", x = `Day of Week`), "weekday or weeken
d"] <- "weekend"
activityDT[, `weekday or weekend` := as.factor(`weekday or weekend`)]
head(activityDT, 10)
```

```
##     steps       date interval Day of Week weekday or weekend
##  1:    NA 2012-10-01        0      Monday            weekday
##  2:    NA 2012-10-01        5      Monday            weekday
##  3:    NA 2012-10-01       10      Monday            weekday
##  4:    NA 2012-10-01       15      Monday            weekday
##  5:    NA 2012-10-01       20      Monday            weekday
##  6:    NA 2012-10-01       25      Monday            weekday
##  7:    NA 2012-10-01       30      Monday            weekday
##  8:    NA 2012-10-01       35      Monday            weekday
##  9:    NA 2012-10-01       40      Monday            weekday
## 10:    NA 2012-10-01       45      Monday            weekday
```

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). See the README file in the GitHub repository to see an example of what this plot should look like using simulated data.

```
activityDT[is.na(steps), "steps"] <- activityDT[, c(lapply(.SD, median, na.rm = TRU
E)), .SDcols = c("steps")]
IntervalDT <- activityDT[, c(lapply(.SD, mean, na.rm = TRUE)), .SDcols = c("steps"), b
y = .(interval, `weekday or weekend`)]

ggplot(IntervalDT , aes(x = interval , y = steps, color=`weekday or weekend`)) + geom_
line() + labs(title = "Avg. Daily Steps by Weektype", x = "Interval", y = "No. of Step
s") + facet_wrap(~`weekday or weekend` , ncol = 1, nrow=2)
```


Avg. Daily Steps by Weektype
weekday
weekend
No. of Steps
200
150
100
50
0
0
500
1000
1500
2000
Interval
weekday or weekend
weekday
weekend