ReproducibleProj1

Xin Dai

March 5, 2021

Loading and preprocessing the data

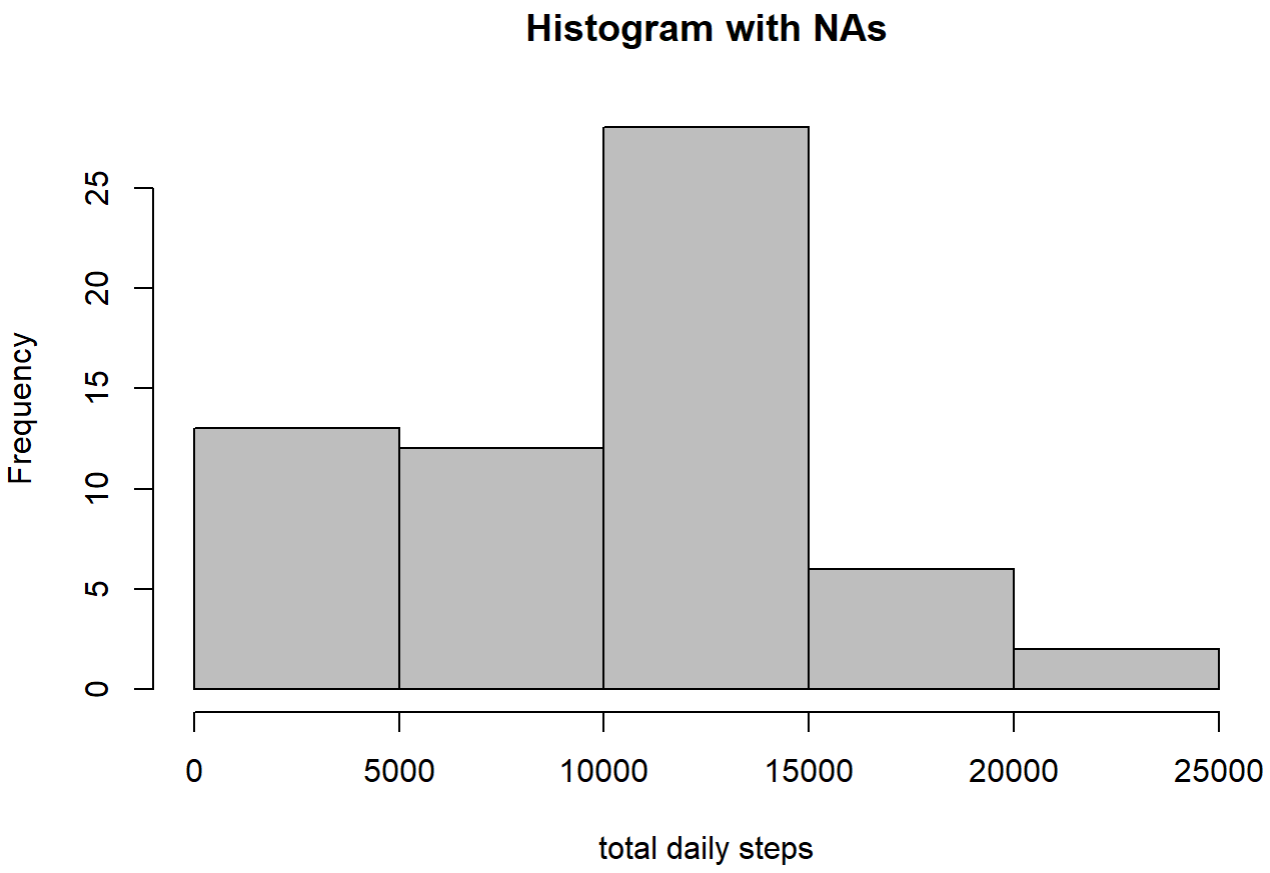
```
dat.loc <- "D:\\Coursera\\ReproducibleResearch_RPeng\\Projects\\proj1"
dat <- read.csv(paste(dat.loc, "activity.csv", sep = "\\"), sep = ",")
str(dat)
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : Factor w/ 61 levels "2012-10-01","2012-10-02",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
dat$date2 <- as.Date(dat$date, "%Y-%m-%d")
```

What is mean total number of steps taken per day?

```
##load "tidyverse" package
library(tidyverse)
```

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

```
sum1 <- dat %>% group_by(date) %>%
              summarize(daily.step = round(sum(steps, na.rm = TRUE), digits = 0))
hist(sum1$daily.step, col = "gray", main = "Histogram with NAs", xlab = "total daily steps")
```



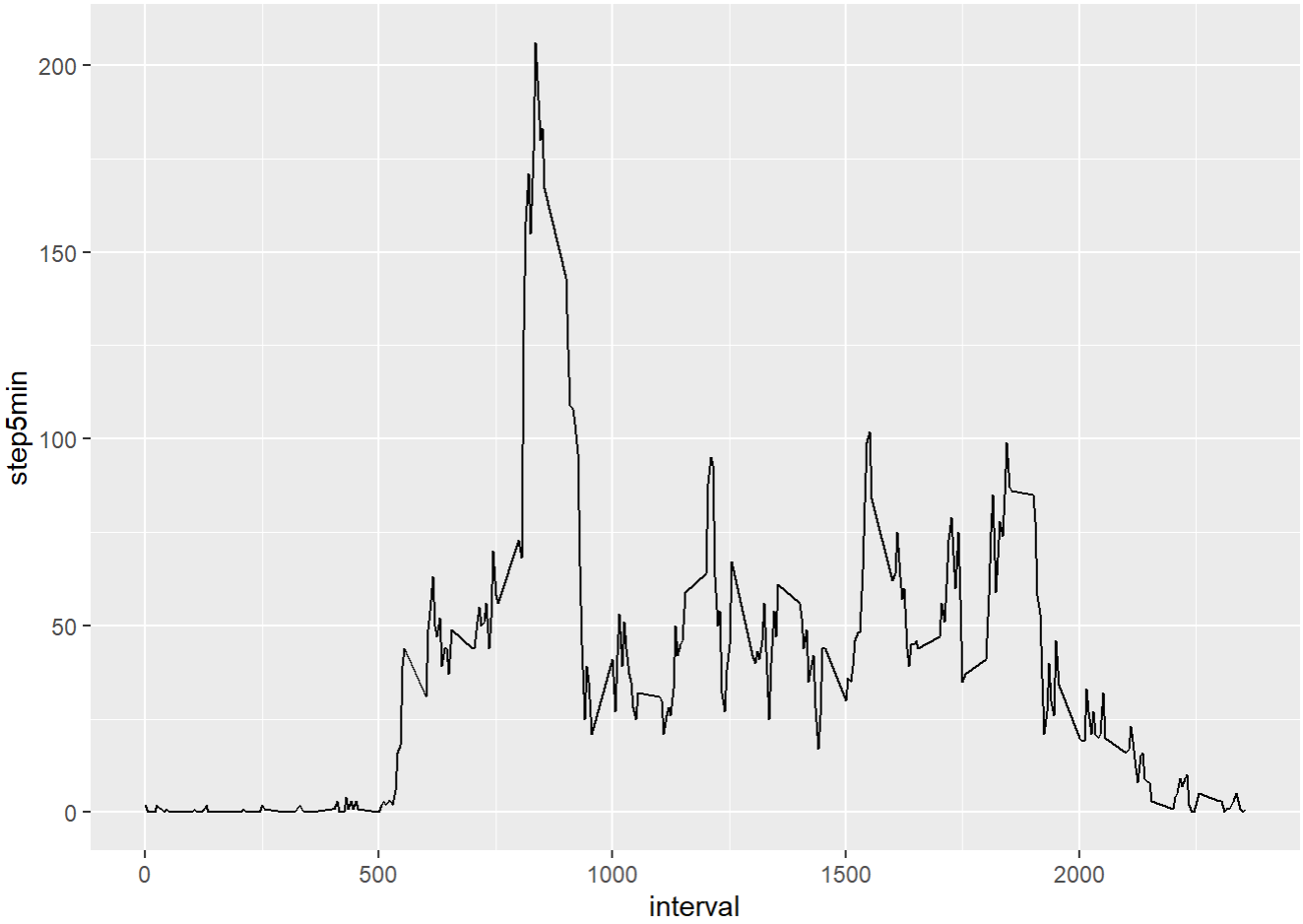
2. Calculate and report the mean and median total number of steps taken per day

```
(summarize(sum1, daily.mean = round(mean(daily.step, na.rm = TRUE), digits = 0),
                 daily.median = round(median(daily.step, na.rm = TRUE), digits = 0)))
## # A tibble: 1 x 2
##   daily.mean daily.median
##        <dbl>        <dbl>
## 1       9354        10395
```

What is the average daily activity pattern?

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)

```
sum2 <-dat %>% group_by(interval) %>%
              summarize(step5min = round(mean(steps, na.rm = T), digits = 0))
ggplot(sum2, aes(x = interval, y = step5min)) + geom_line()
```



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

```
(filter(sum2, step5min == max(step5min)))
## # A tibble: 1 x 2
##   interval step5min
##      <int>    <dbl>
## 1      835      206
```

The maxium number of steps occured at 8:35 am and the person walked more than 200 steps during the 5-minute interval.

Imputing missing values

There are days/intervals where there are missing values (coded as NA). The presence of missing days may introduce bias into some calculations or summaries of the data.

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

```
sum(is.na(dat$steps))
## [1] 2304
```

2. Devise a strategy for filling in all of the missing values in the dataset. **For this report, the median for that 5-minute interval was used**.

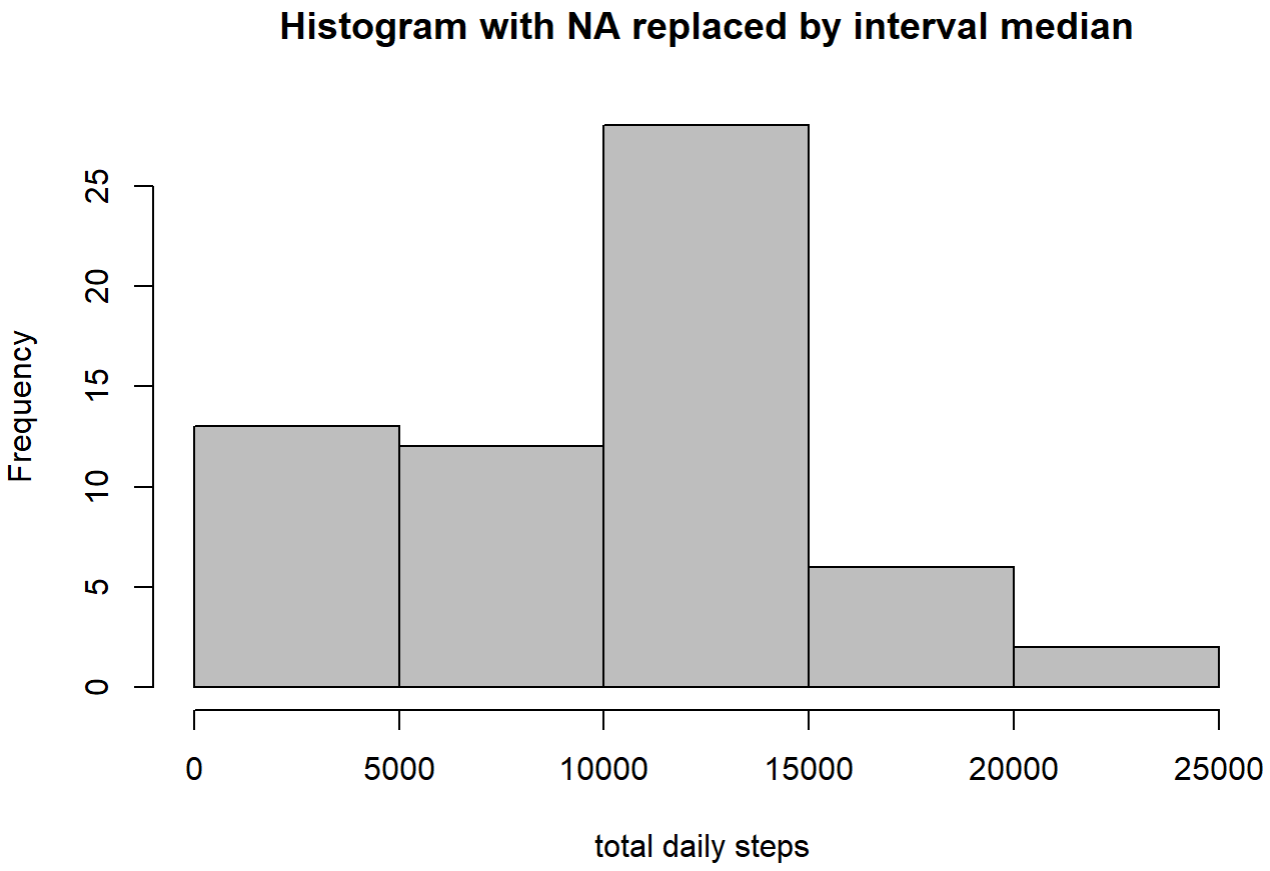
3. Create a new dataset that is equal to the original dataset but with the missing data filled in.

```
dat.noNA <- dat %>% group_by(interval) %>%
                mutate(steps2 = ifelse(is.na(steps), median(steps, na.rm=TRUE), steps)) %>%
                arrange(interval)
```

```
# Example output to show NA values at interval = 800 is replaced by 41, the median at that int
erval.
dat.noNA %>% filter(is.na(steps)) %>% filter(interval == 800)
## # A tibble: 8 x 5
## # Groups:   interval [1]
##   steps date       interval date2      steps2
##   <int> <fct>         <int> <date>      <int>
## 1    NA 2012-10-01      800 2012-10-01     41
## 2    NA 2012-10-08      800 2012-10-08     41
## 3    NA 2012-11-01      800 2012-11-01     41
## 4    NA 2012-11-04      800 2012-11-04     41
## 5    NA 2012-11-09      800 2012-11-09     41
## 6    NA 2012-11-10      800 2012-11-10     41
## 7    NA 2012-11-14      800 2012-11-14     41
## 8    NA 2012-11-30      800 2012-11-30     41
```

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?

```
sum3 <- dat.noNA %>% group_by(date) %>% summarize(daily.noNA = round(sum(steps2), digits = 0))
hist(sum3$daily.noNA, col = "gray", main = "Histogram with NA replaced by interval median",
     xlab = "total daily steps ")
```




```
sum3 %>% summarize(daily.mean.noNA = round(mean(daily.noNA), digits = 0),
                   daily.median.noNA = round(median(daily.noNA), digits = 0))
## # A tibble: 1 x 2
##   daily.mean.noNA daily.median.noNA
##             <dbl>             <dbl>
## 1            9504             10395
```

When NA is replaced with interval median, the average of daily steps is **higher** than that when NA is removed. The median of daily steps remains the same.

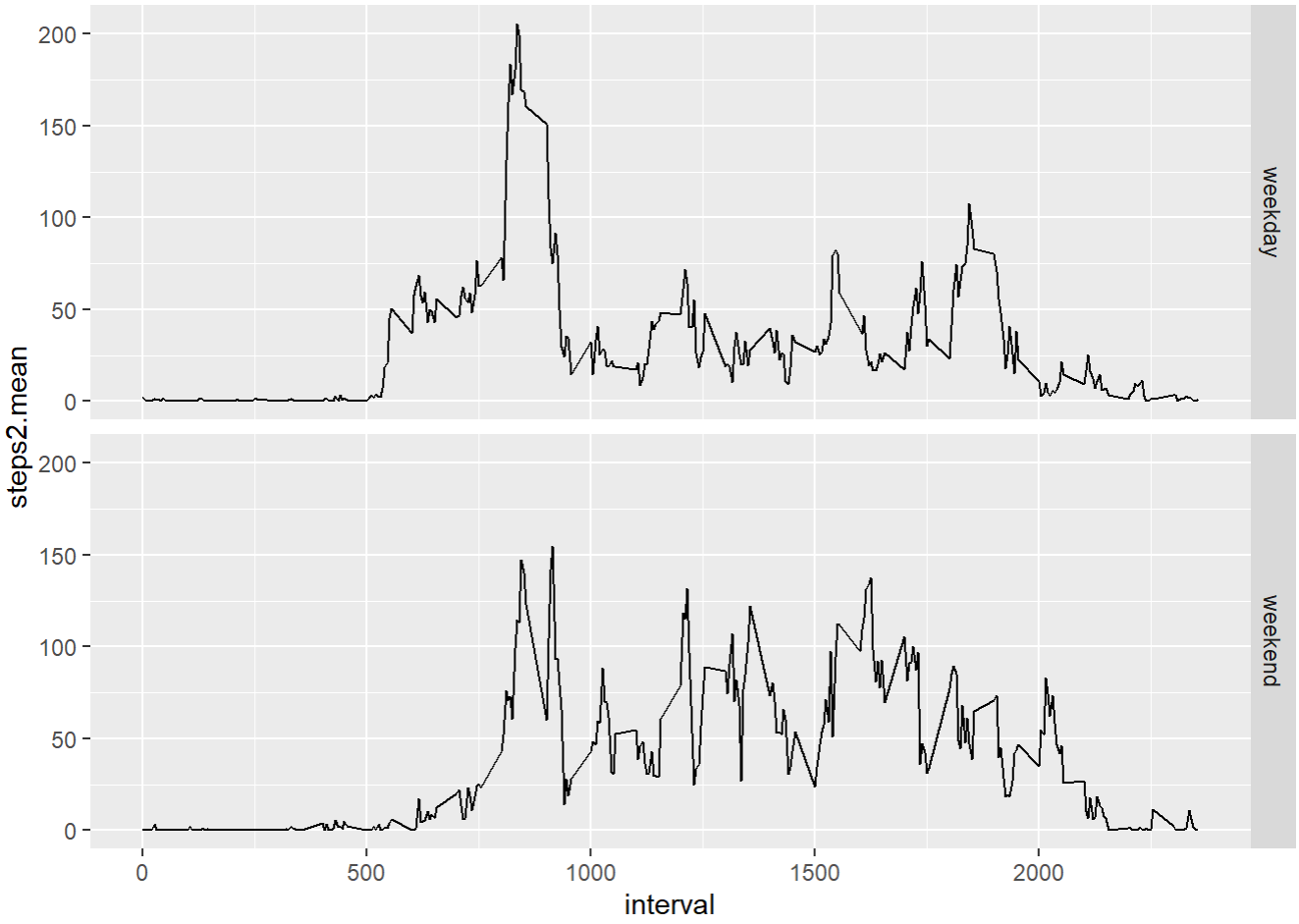
Are there differences in activity patterns between weekdays and weekends?

1. Creat a factor variable with two levels “weekday” and “weekend” indicating whether a given date is a weekday or a weekend day and,

```
dat.noNA$dateweek <- as.factor(weekdays(dat.noNA$date2))
dat.noNA <- dat.noNA %>% mutate(weekdayid = ifelse(dateweek %in% c("Saturday", "Sunday"),
                                                   "weekend", "weekday"))
```

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

```
dat.noNA %>% group_by(interval, weekdayid) %>% summarise(steps2.mean = mean(steps2)) %>%
       ggplot(aes(x = interval, y = steps2.mean)) + geom_line() + facet_grid(weekdayid~.)
```



The 5-minute interval steps during weekdays and weekends **are different**. The person started to walk actively after 5 am during weekdays but at around 8 am during weekends. Although the peak steps occurred between 8-9 am both weekdays and weekends, the person walked more during daytime (9 am - 8 pm) on weekends. Walking activity decreased and person perhaps ready to rest after 8 pm on weekdays but after 9 pm on weekends.