project1_activity

weimf1990

12/06/2021

Loading and preprocessing the data

```
#setting the directory
setwd("C:/Users/user/Desktop/Data Science/project/Course 5")

#loading the data
activity <- read.csv("./repdata_data_activity/activity.csv")

#processing the data
activity$day <- weekdays(as.Date(activity$date))
activity$DateTime<- as.POSIXct(activity$date, format="%Y-%m-%d")

##clean the data without na
data <- activity[!is.na(activity$steps),]
```

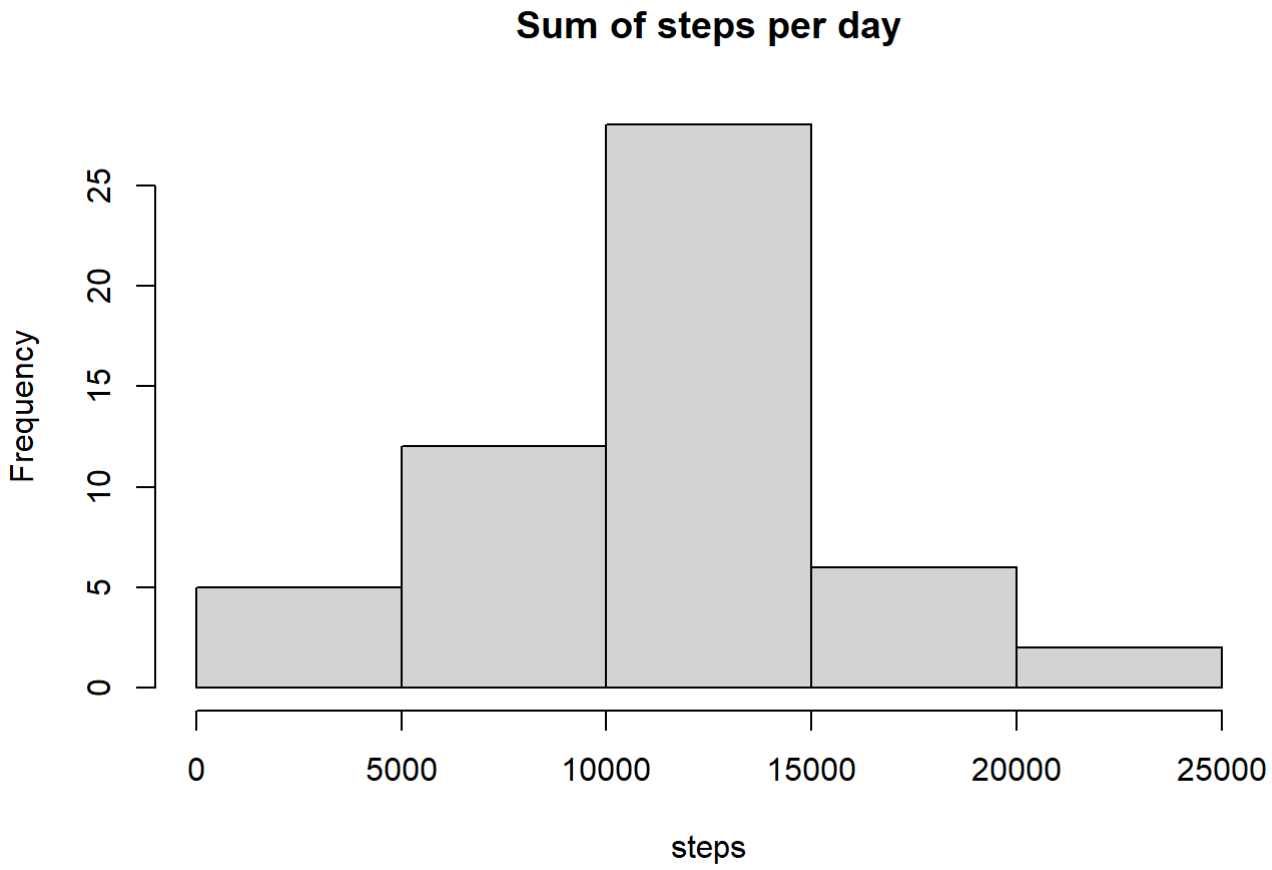
What is mean total number of steps taken per day?

Histogram of the total number of steps taken each day

```
## taking the sum of steps evey day
sumsteps <- aggregate(data$steps ~ data$date, FUN=sum, )
colnames(sumsteps)<- c("Date", "Steps")
```

Histogram of the total number of steps taken each day

```
hist(sumsteps$Steps, xlab="steps", main = "Sum of steps per day")
```



Mean and median number of steps taken each day

```
mean_steps <- mean(sumsteps$Steps)
mean_steps
```

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

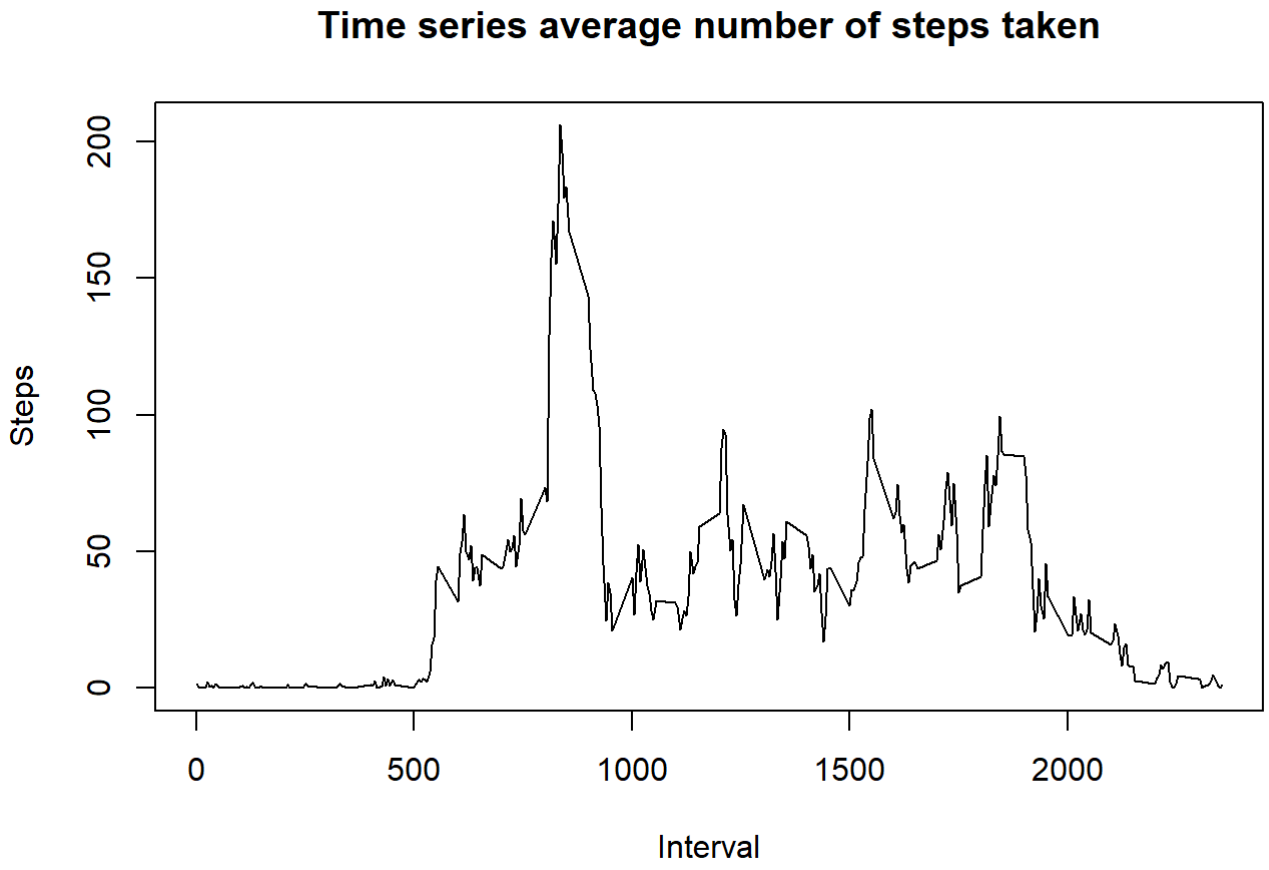
```
median_steps <- median(sumsteps$Steps)
median_steps
```

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

The mean and median number of steps taken each day is 10,766 and 10,765, respectively.

What is the average daily activity pattern?

```
Steps_interval <- tapply(data$steps, data$interval, FUN=mean)

plot(as.numeric(names(Steps_interval)),
     Steps_interval,
     xlab = "Interval",
     ylab = "Steps",
     main = "Time series average number of steps taken",
     type = "l")
```



The 5-minute interval that, on average, contains the maximum number of steps

```
#dataframe
Steps_interval_table <- data.frame(Steps_interval)
Steps_interval_table <- cbind(rownames(Steps_interval_table), Steps_interval_table)

names(Steps_interval_table)[1] <- "interval"

##Maximum number of steps by interval
maxsteps <- max(Steps_interval_table$Steps_interval)
maxsteps
```

```
## [1] 206.1698
```

```
##Which interval contains the maximum average number of steps
Steps_interval_table[Steps_interval_table$Steps_interval == maxsteps,1]
```

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

The maximum number of steps for a 5-minute interval is 206 steps.

The 5-minute interval which has the maximum number of steps is the 835 interval.

Imputing missing values

```
##total number of rows with NAs
nrow(activity[is.na(activity$steps),])
```

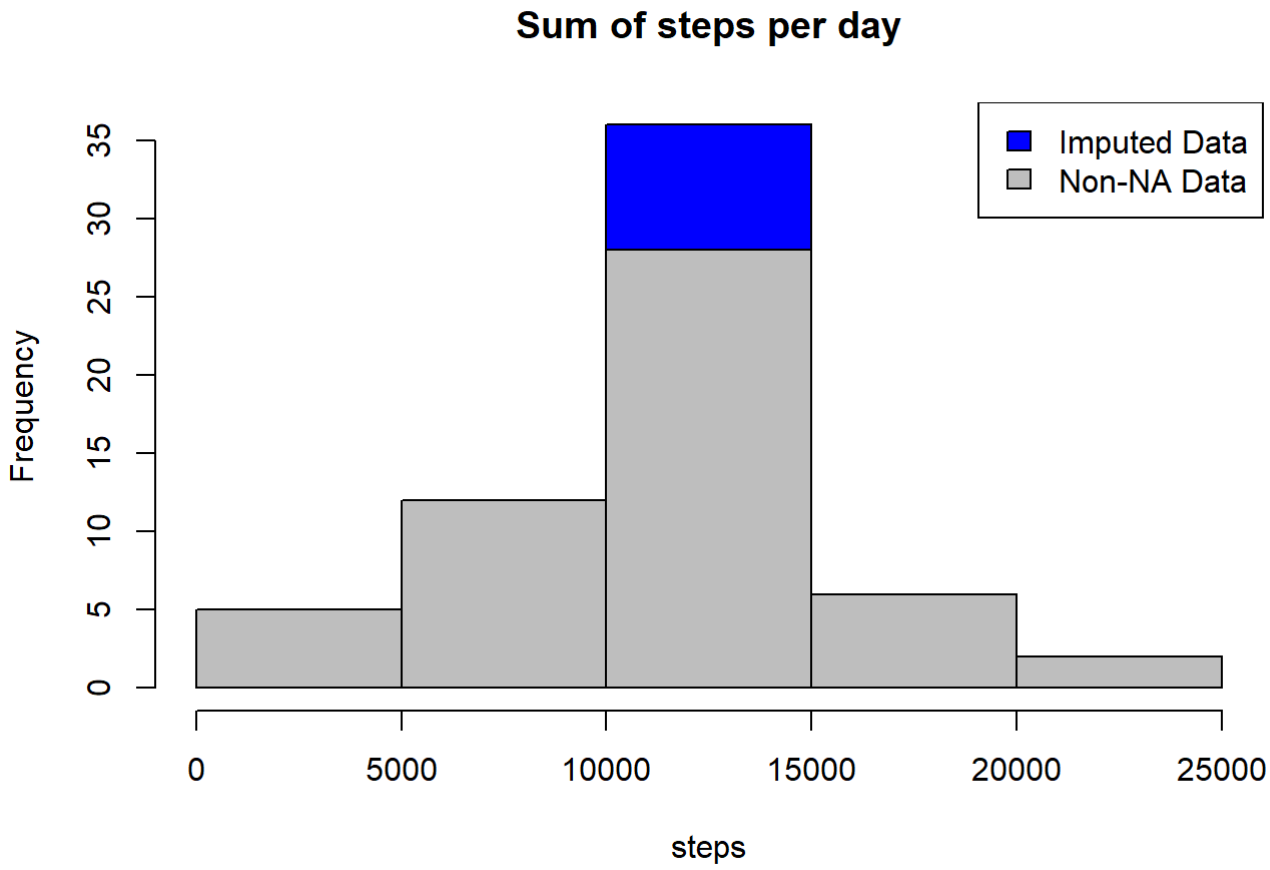
```
## [1] 2304
```

There are 2,304 missing values in the data. I choose to impute these missing data using the mean number of steps by the particular interval across dates.

```
## Create dataset with all NAs for substitution
nadata <- activity[is.na(activity$steps),]

## Merge NA data with Steps_interval _table, the mean values generated in previous steps
newdata <-merge(nadata, Steps_interval_table, by="interval")

## Remove steps column in newdata & rename Steps_interval to steps (consistent with the colum
n names in data)
newdata <- subset(newdata, select = -c(steps))
names(newdata)[5] <- "steps"

## Merge newdata with the intitial dataset without NA values (data)
mergeData <- rbind(data, newdata)
```

Histogram of the total number of steps taken each day after missing values are imputed

```
## taking the sum of steps evey day
sumsteps_2 <- aggregate(mergeData$steps ~ mergeData$date, FUN=sum, )
colnames(sumsteps_2)<- c("Date", "Steps")

hist(sumsteps_2$Steps, xlab="steps", main = "Sum of steps per day", col = "blue")
hist(sumsteps$Steps, xlab="Steps", main = "Sum of steps per day", col="Grey", add=T)
legend("topright", c("Imputed Data", "Non-NA Data"), fill=c("blue", "grey") )
```



Report the mean and median total number of steps taken per day.

```
imputedmean_steps <- mean(sumsteps_2$Steps)
imputedmean_steps
```

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

```
imputedmedian_steps <- median(sumsteps_2$Steps)
imputedmedian_steps
```

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

```
#Do these values differ from the estimates from the first part of the assignment?
mean_diff <- imputedmean_steps - mean_steps
mean_diff
```

```
## [1] 0
```

```
median_diff <- imputedmedian_steps - median_steps
median_diff
```

```
## [1] 1.188679
```

The mean remains the same as prior to imputation, and the median value increases by 1.18. Based on the figure plotted above, the difference lies in the steps ranging from 10,000 to 15,000.

Are there differences in activity patterns between weekdays and weekends?

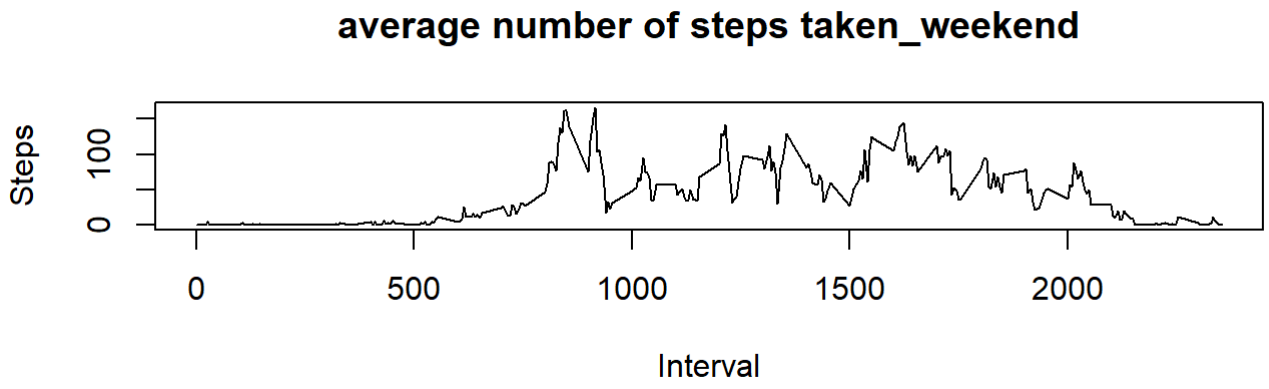
```
## Create a new factor variable in the dataset with two levels – “weekday” and “weekend” indi
cating whether a given date is a weekday or weekend day.
mergeData$Daylevels <- ifelse(mergeData$day %in% c("星期六", "星期日"), "Weekend", "Weekday")

# Calculate average steps for weekends
weekends <- subset(mergeData, Daylevels == "Weekend")
Steps_weekend <- tapply(weekends$steps, weekends$interval, FUN=mean)

# Calculate average steps for weekdays
weekdays <- subset(mergeData, Daylevels == "Weekday")
Steps_weekday <- tapply(weekdays$steps, weekdays$interval, FUN=mean)

summary(Steps_weekend)
```

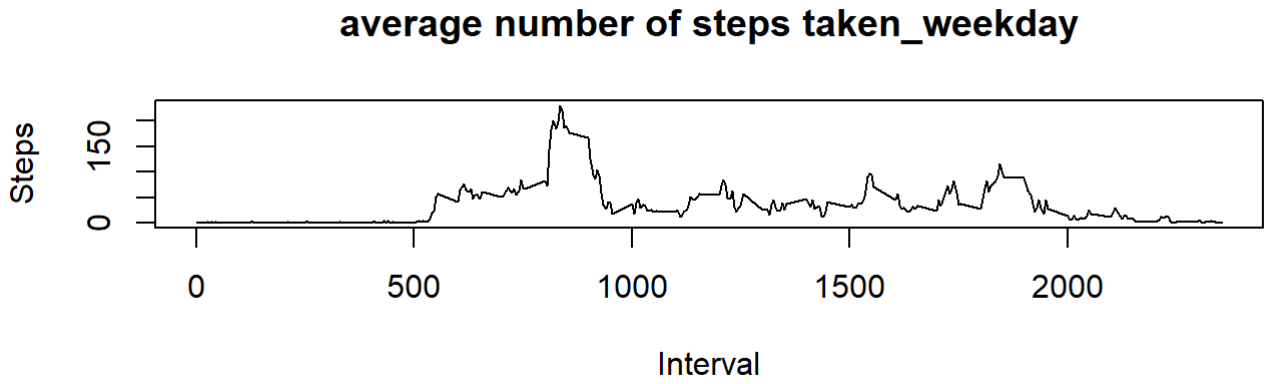
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.241  32.340  42.366  74.654 166.639
```

```
summary(Steps_weekday)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.247  25.803  35.611  50.854 230.378
```

```
#Make a panel plot containing a time series plot (type = "l") of the 5-minute interval (x-axi
s) and the average number of steps taken, averaged across all weekday days or weekend days (y
-axis).
# Set a 2 panel plot
par(mfrow=c(2,1))

plot(as.numeric(names(Steps_weekend)),
     Steps_weekend,
     xlab = "Interval",
     ylab = "Steps",
     main = "average number of steps taken_weekend",
     type = "l")

plot(as.numeric(names(Steps_weekday)),
     Steps_weekday,
     xlab = "Interval",
     ylab = "Steps",
     main = "average number of steps taken_weekday",
     type = "l")
```





The step activity trends are different between the weekend and weekday. Specifically, people seem to have more opportunity in activities on weekends.