Reproducible Research Peer Assessment 1

Colleen

3/16/2022

Loading and preprocessing the data

1.Load the data

```
activity <- read.csv("activity.csv")
head(activity)
```

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

```
str(activity)
```

```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : chr  "2012-10-01" "2012-10-01" "2012-10-01" "2012-10-01" ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
```

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

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

What is the mean total number of steps taken per day?

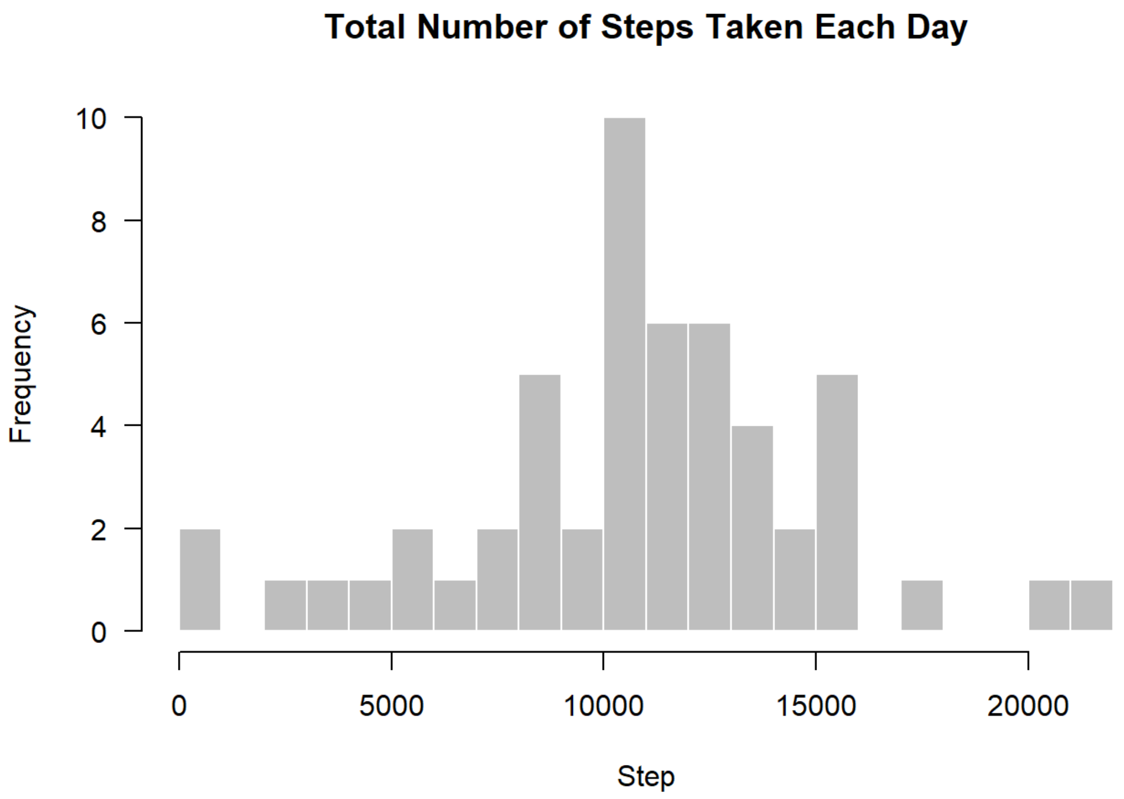
1.calculate the total number of steps taken per day

```
total_step <- aggregate(steps ~ date, data = activity, sum, na.rm = TRUE)
head(total_step)
```

```
##         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.Create a Histogram of the total number of steps taken per day.

```
par(mfrow = c(1, 1))
hist(total_step$steps, breaks = 20,
     main = "Total Number of Steps Taken Each Day",
     col = "grey", border = "white", xlab = "Step", axes = FALSE)
axis(1)
axis(2, las = 1)
```



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

```
mean(total_step$steps)
```

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

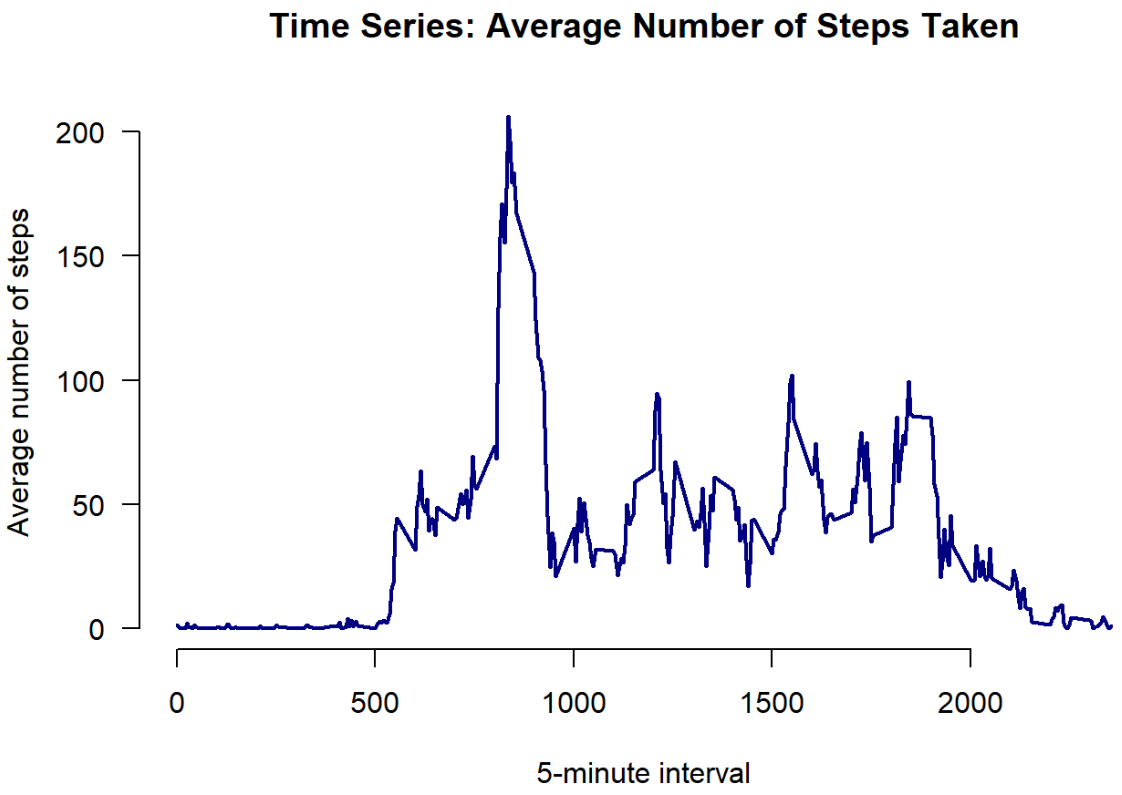
```
median(total_step$steps)
```

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

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).

```
avg_step <- aggregate(steps ~ interval, data = activity, mean, na.rm = TRUE)
plot(avg_step$interval, avg_step$steps, type = "l", lwd = 2, col = "navy",
     main = "Time Series: Average Number of Steps Taken", axes = FALSE,
     xlab = "5-minute interval", ylab = "Average number of steps")
axis(1)
axis(2, las = 1)
```



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

```
avg_step$interval[which.max(avg_step$steps)]
```

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

The 835-th 5-minute interval contains the maximum number of steps.

Inputing missing values

1.Calculate and report the total number of missing values in the dataset (ie, the total number of rows with na’s).

```
sum(is.na(activity))
```

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

There are 2304 missing values in the dataset.

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. -Using the mean of 5 minute interval to fill in the missing values

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

```
imp <- activity
for (i in avg_step$interval) {
    imp[imp$interval == i & is.na(imp$steps), ]$steps <-
        avg_step$steps[avg_step$interval == i]
}
head(imp)
```

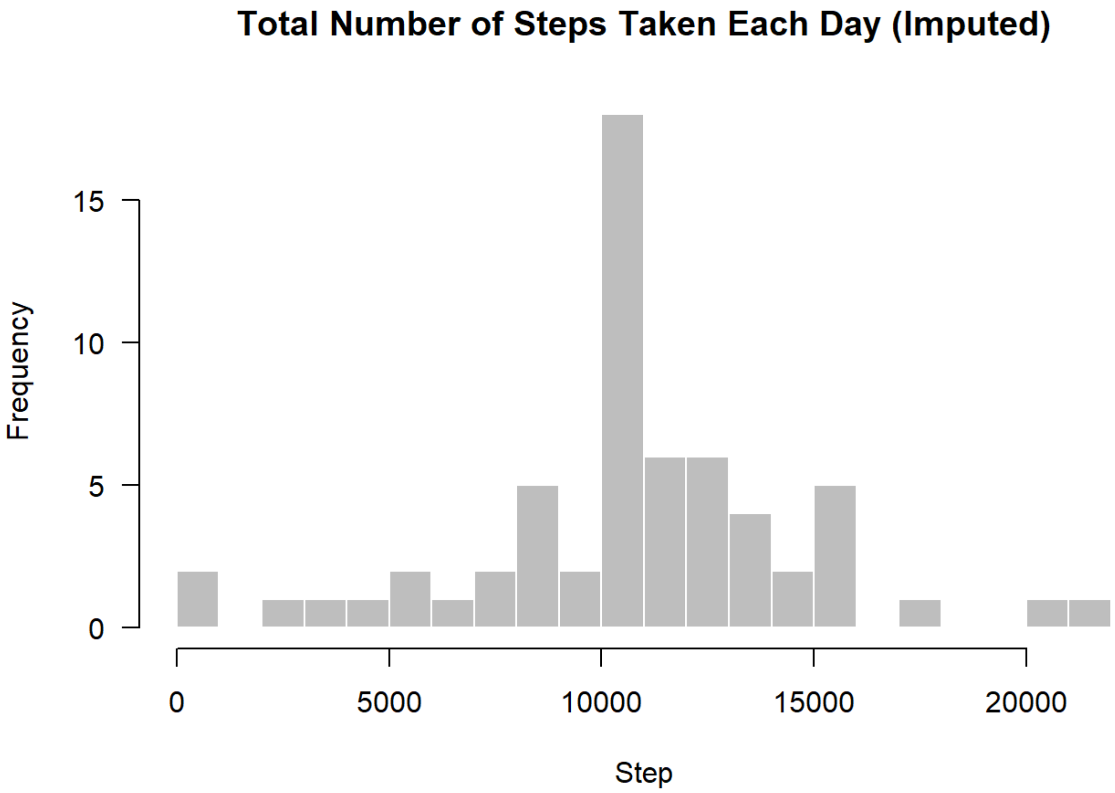
```
##       steps       date interval
## 1 1.7169811 2012-10-01        0
## 2 0.3396226 2012-10-01        5
## 3 0.1320755 2012-10-01       10
## 4 0.1509434 2012-10-01       15
## 5 0.0754717 2012-10-01       20
## 6 2.0943396 2012-10-01       25
```

```
sum(is.na(imp))
```

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

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_step_imp <- aggregate(steps ~ date, data = imp, sum, na.rm = TRUE)
hist(total_step_imp$steps, breaks = 20,
     main = "Total Number of Steps Taken Each Day (Imputed)",
     col = "grey", border = "white", xlab = "Step", axes = FALSE)
axis(1)
axis(2, las = 1)
```



```
mean(total_step_imp$steps)
```

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

```
median(total_step_imp$steps)
```

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

After inputting the missing values, the mean is the same as the first part of the assignment. This is because I used an average of the 5 minute interval for the missing values, resulting in more data that is equal to the mean.

Are there differences in activity patterns between weekdays and weekends?

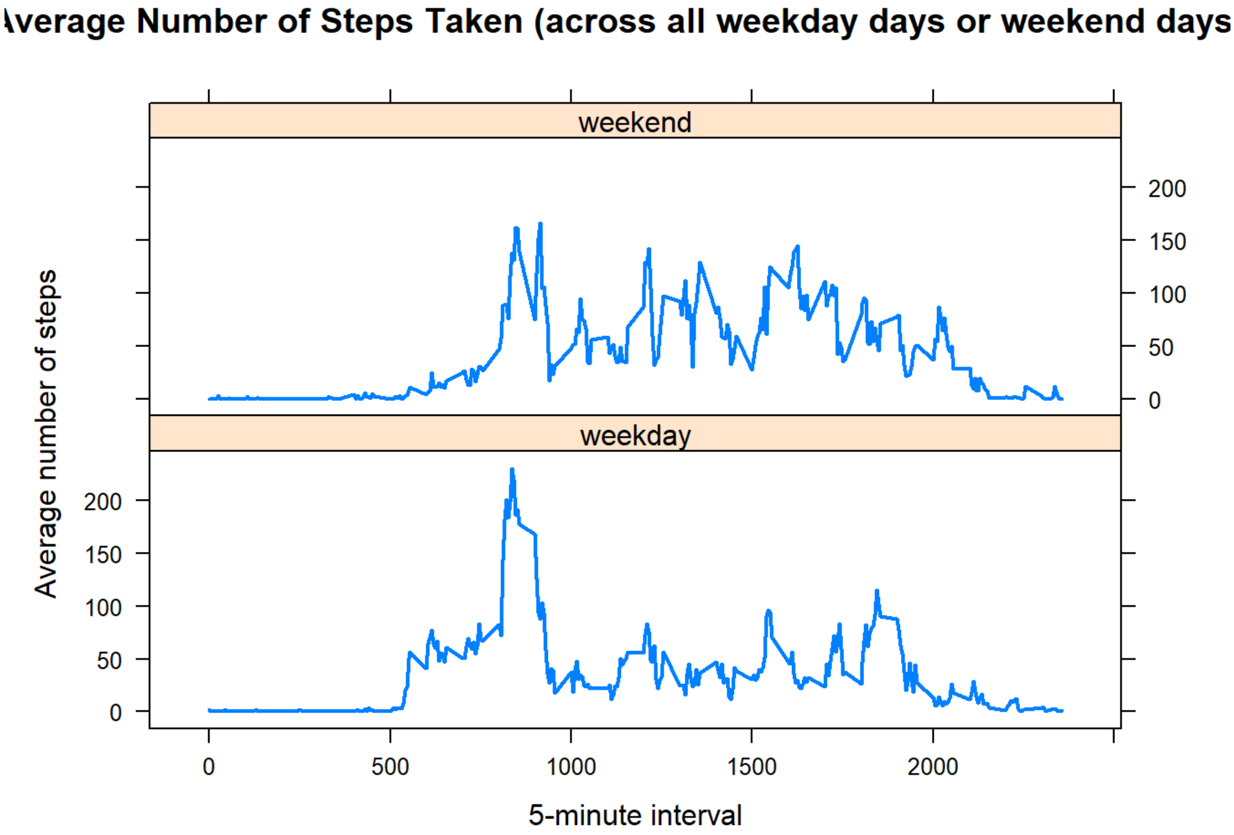
For this part the weekdays() function may be of some help here. Use the dataset with the filled-in missing values for this part.

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.

```
imp$day <- weekdays(imp$date)
imp$week <- ""
imp[imp$day == "Saturday" | imp$day == "Sunday", ]$week <- "weekend"
imp[!(imp$day == "Saturday" | imp$day == "Sunday"), ]$week <- "weekday"
imp$week <- factor(imp$week)
```

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.

```
avg_step_imp <- aggregate(steps ~ interval + week, data = imp, mean)
library(lattice)
xyplot(steps ~ interval | week, data = avg_step_imp, type = "l", lwd = 2,
       layout = c(1, 2),
       xlab = "5-minute interval",
       ylab = "Average number of steps",
       main = "Average Number of Steps Taken (across all weekday days or weekend d
ays)")
```


\verage Number of Steps Taken (across all weekday days or weekend days
weekend
weekday
Average number of steps
0
50
100
150
200
0
500
1000
1500
2000
5-minute interval