Reproducible Research: Peer Assessment 1

Loading and preprocessing the data

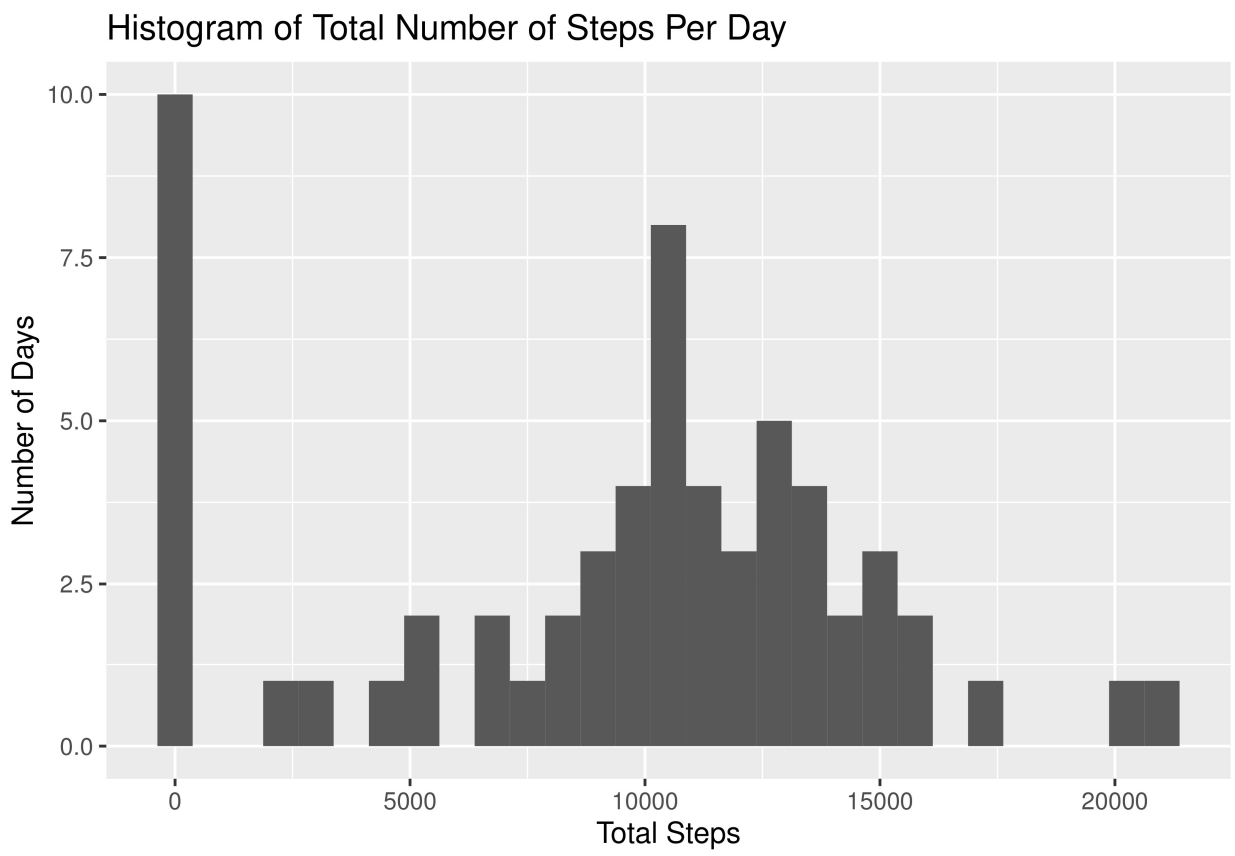
The first step is simply unzipping the file and reading it as a csv. Then, we load in the required package and transform the data

```
unzip(zipfile = "activity.zip")
data <- read.csv("activity.csv")
library(ggplot2)
steps.ds <- tapply(data$steps, data$date, FUN = sum, na.rm = TRUE)
```

What is mean total number of steps taken per day?

Now, we want to visaulize the data and find the mean and median.

```
qplot(steps.ds, binwidth = 750, xlab = "Total Steps", ylab = "Number of Days", main = "Histogram of Tota
```



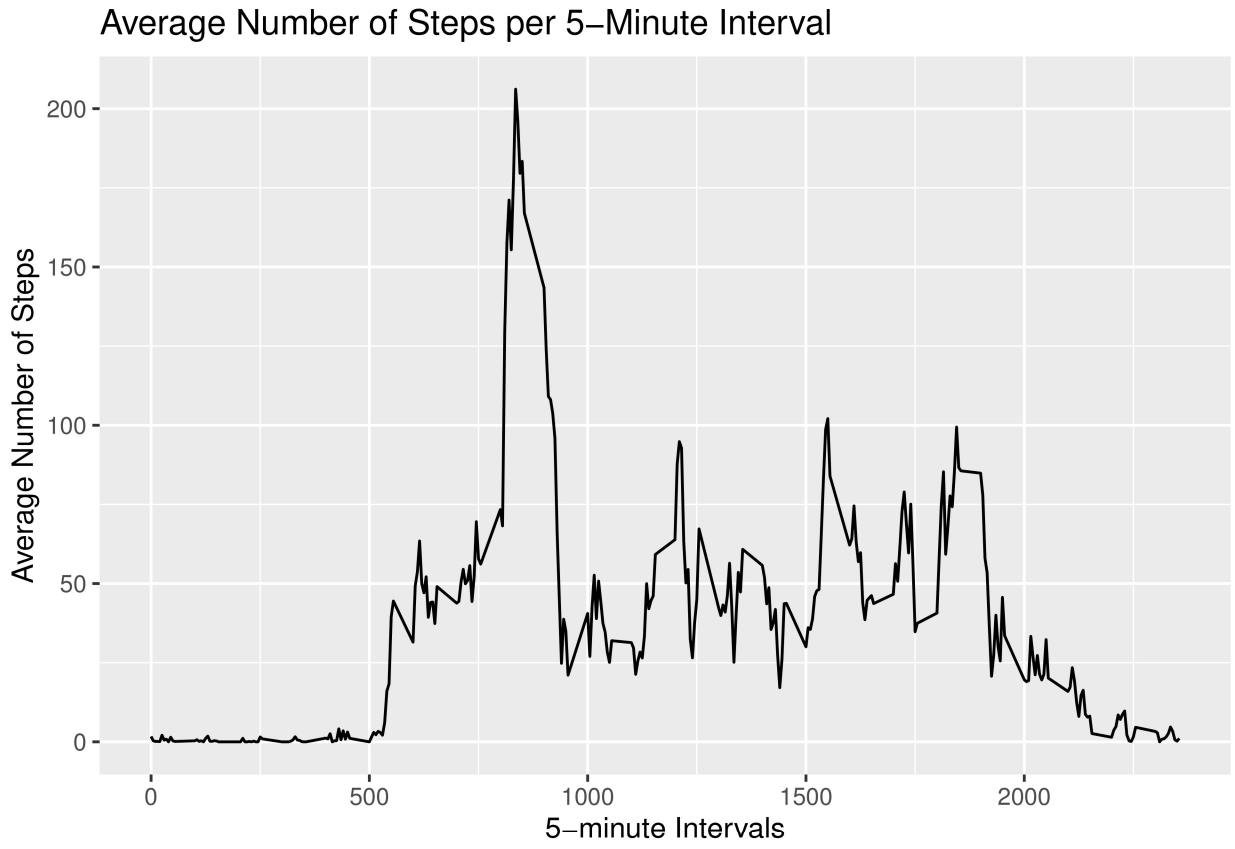
```
daymean <- mean(steps.ds, na.rm = TRUE)
daymedian <- median(steps.ds, na.rm = TRUE)
```

Therefore, our mean is 9354.2295082 and our median is 10395.

What is the average daily activity pattern?

Next, we want to figure out which 5-minute interval, on average across all the days in the dataset, contains the maximum number of steps.

```
data.avg <- aggregate(x = list(steps = data$steps), by = list(interval = data$interval), FUN = mean, na
ggplot(data = data.avg, aes(x = interval, y = steps)) +
      geom_line() +
      xlab("5-minute Intervals") +
      ylab("Average Number of Steps") +
      ggtitle("Average Number of Steps per 5-Minute Interval")
```




```
max <- data.avg[which.max(data.avg$steps), ]
table(max)
```

```
##         steps
## interval 206.169811320755
##      835                 1
```

Therefore, the 5-minute interval with the most number of steps was interval 835 with 206 steps.
Imputing missing values The next step is finding the number of missing values (represented by NA in the data), replace them with the average number of steps per day and create a new histogram with the filled data to see if anything has changed.

```
na.total <- is.na(data$steps)
table(na.total)
```

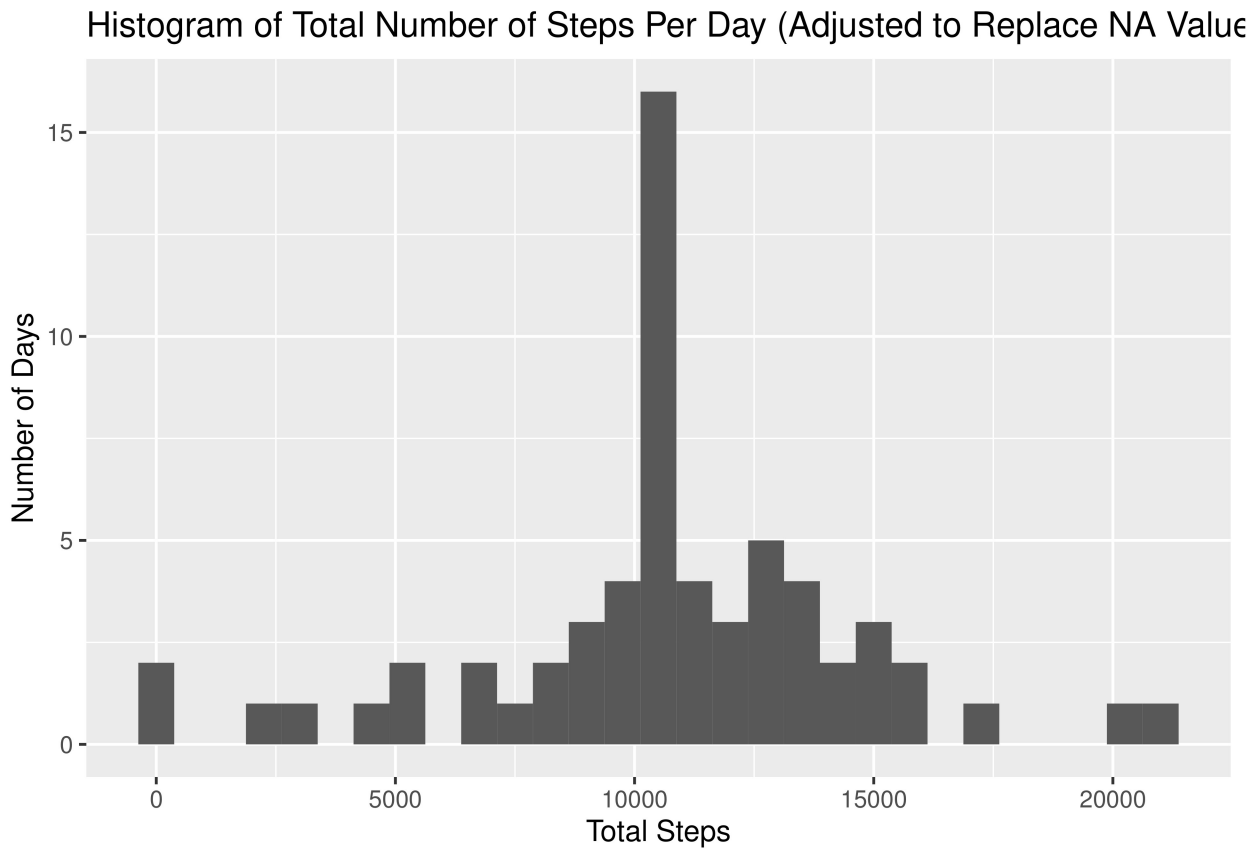
```
## na.total
## FALSE  TRUE
## 15264  2304
```

```
fill.na <- function(steps, interval) {
  fill <- NA
  if (!is.na(steps))
    fill <- c(steps) else fill <- (data.avg[data.avg$interval == interval, "steps"])
    return(fill)
}
fill.data <- data
fill.data$steps <- mapply(fill.na, fill.data$steps, fill.data$interval)

fill.natotal <- is.na(fill.data$steps)
table(fill.natotal)
```

```
## fill.natotal
## FALSE
## 17568
```

```
fillsteps.ds <- tapply(fill.data$steps, fill.data$date, FUN = sum)
qplot(fillsteps.ds, binwidth = 750, xlab = "Total Steps", ylab = "Number of Days", main = "Histogram of
```




```
fillmean <- mean(fillsteps.ds)
fillmedian <- median(fillsteps.ds)
```

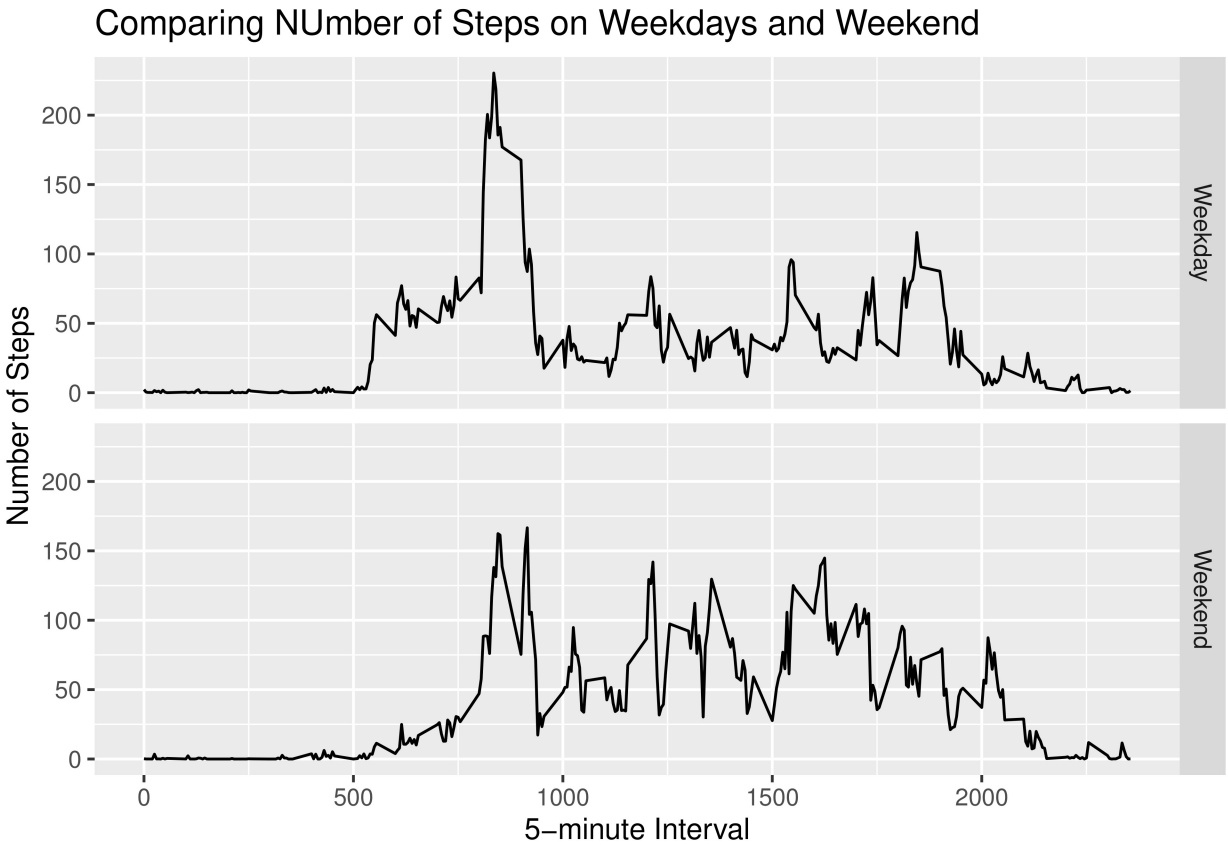
Therefore, our mean is 1.0766189×10^4 and our median is 1.0766189×10^4. Both ended up being higher than the original predictions without the adjustment.Each value replaced increased the total daily number of steps.

Are there differences in activity patterns between weekdays and weekends?

Lastly, we want to check activity patterns in weekdays and weekends. The first step is to set the dates as weekdays and then classify them as either weekdays or weekend. Then, we find the means and make a time series plot comparing the two.

```
fill.data$weekdays <- weekdays(as.Date(fill.data$date))
fill.data$daytype <- ifelse(fill.data$weekdays %in% c("Saturday", "Sunday"), "Weekend", "Weekday")

weekday.avg <- aggregate(steps ~ interval + daytype, data = fill.data, mean)
ggplot(weekday.avg, aes(interval, steps)) + geom_line() + facet_grid(daytype ~ .) +
  xlab("5-minute Interval") +
  ylab("Number of Steps") +
  ggtitle("Comparing NUmber of Steps on Weekdays and Weekend")
```


Comparing NUmber of Steps on Weekdays and Weekend
Number of Steps
0
50
100
150
200
Weekday
Weekend
0
500
1000
1500
2000
5−minute Interval