Project1 week2

DLO

4/6/2024

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, fig.width = 15, fig.height = 10,
                      fig.keep = 'all' ,fig.path = 'figures\ ', dev = 'png')
```

```
library(ggplot2)
```

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

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

```
my_date <- as.Date("2024-04-07")
weekdays(my_date)
```

```
## [1] "Sunday"
```

```
activity <- cbind(activity, my_date)
```

```
summary(activity)
```

```
##      steps             date                interval         my_date
##  Min.   :  0.00   Min.   :2012-10-01   Min.   :   0.0   Min.   :2024-04-07
##  1st Qu.:  0.00   1st Qu.:2012-10-16   1st Qu.: 588.8   1st Qu.:2024-04-07
##  Median :  0.00   Median :2012-10-31   Median :1177.5   Median :2024-04-07
##  Mean   : 37.38   Mean   :2012-10-31   Mean   :1177.5   Mean   :2024-04-07
##  3rd Qu.: 12.00   3rd Qu.:2012-11-15   3rd Qu.:1766.2   3rd Qu.:2024-04-07
##  Max.   :806.00   Max.   :2012-11-30   Max.   :2355.0   Max.   :2024-04-07
##  NA's   :2304
```

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

```
activityTotalSteps <- with(activity, aggregate(steps, by = list(date), sum, na.rm = TRUE))
```

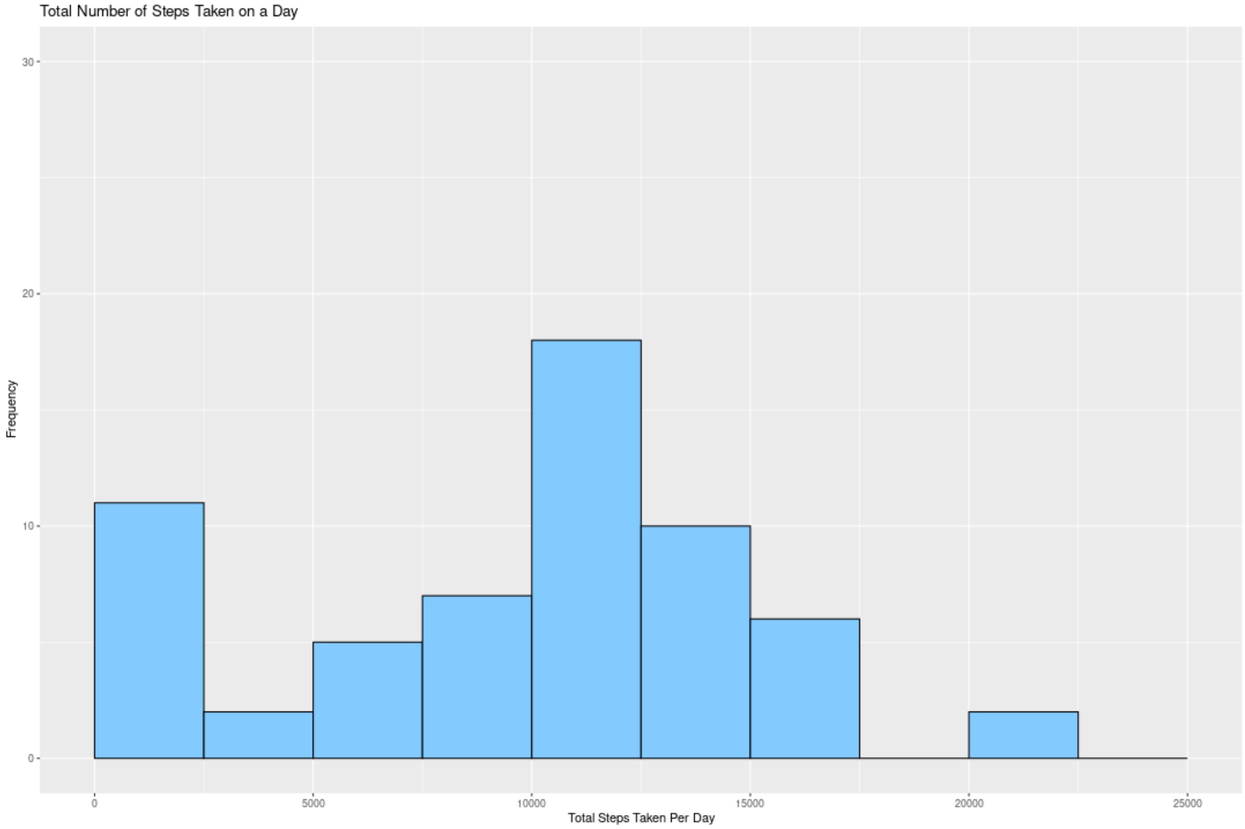
```
names(activityTotalSteps) <- c("Date", "Steps")
```

```
totalStepsdf <- data.frame(activityTotalSteps)
```

Plotting a histogram using ggplot2

```
g <- ggplot(totalStepsdf, aes(x = Steps)) +
  geom_histogram(breaks = seq(0, 25000, by = 2500), fill = "#83CAFF", col = "black") +
  ylim(0, 30) +
  xlab("Total Steps Taken Per Day") +
  ylab("Frequency") +
  ggtitle("Total Number of Steps Taken on a Day")
```

```
print (g)
```



```
mean(activityTotalSteps$Steps)
```

```
## [1] 9354.23
```

Mean 9354.23

```
median(activityTotalSteps$Steps)
```

```
## [1] 10395
```

Median 10395

What is the average daily activity pattern?

average number of steps taken, averaged across all days by 5-min intervals.

```
averageDailyActivity <- aggregate(activity$steps, by = list(activity$interval),
                                  FUN = mean, na.rm = TRUE)
```

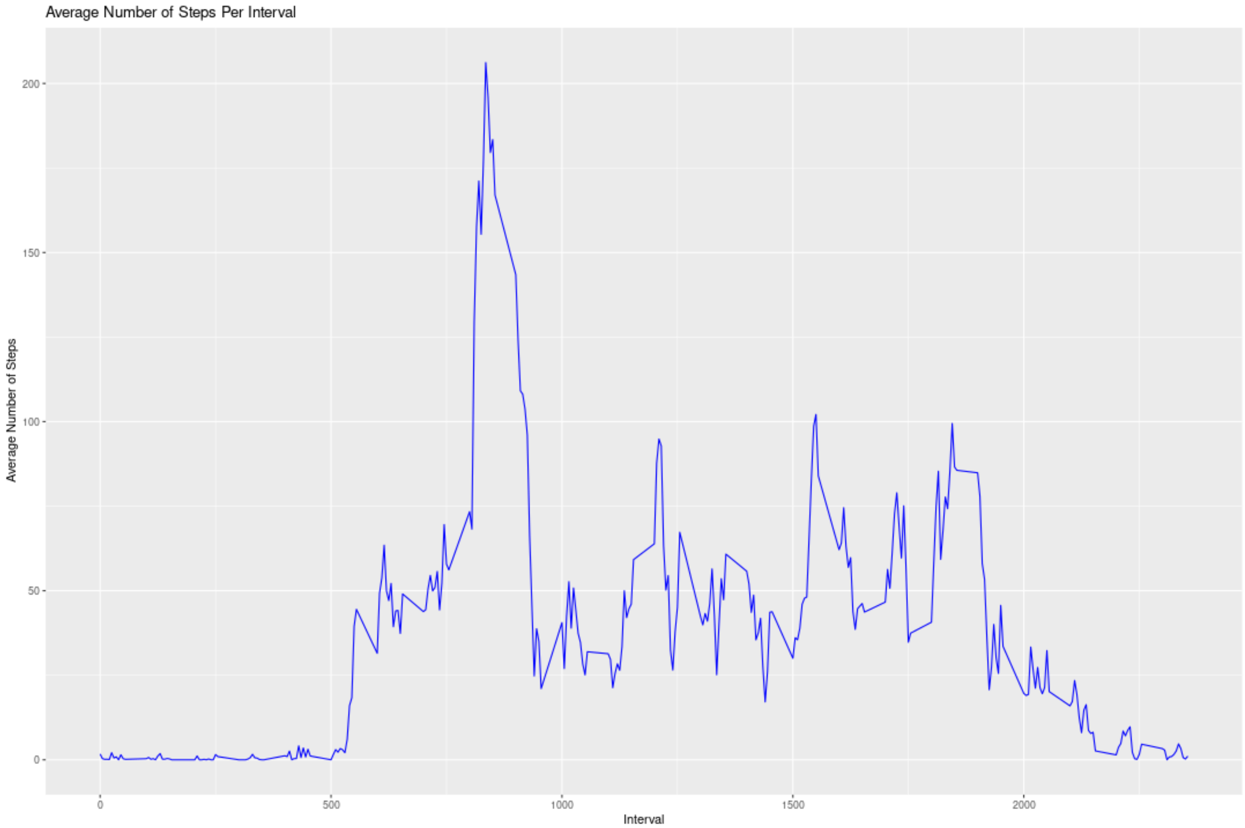
```
names(averageDailyActivity) <- c("Interval", "Mean")
```

Converting the data set into a dataframe

```
averageActivitydf <- data.frame(averageDailyActivity)
```

Plotting

```
da <- ggplot(averageActivitydf, mapping = aes(Interval, Mean)) +
  geom_line(col = "blue") +
  xlab("Interval") +
  ylab("Average Number of Steps") +
  ggtitle("Average Number of Steps Per Interval")

print(da)
```



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

```
averageDailyActivity[which.max(averageDailyActivity$Mean), ]$Interval
```

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

835

Imputing Missing Values

#calculating NAs

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

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

#2304 NAs

#filling in all of the missing values in the dataset. Match mean of daily activity with NAs

```
imputedSteps <- averageDailyActivity$Mean[match(activity$interval, averageDailyActivity$Interval)]
```

#Create a new dataset including NAs

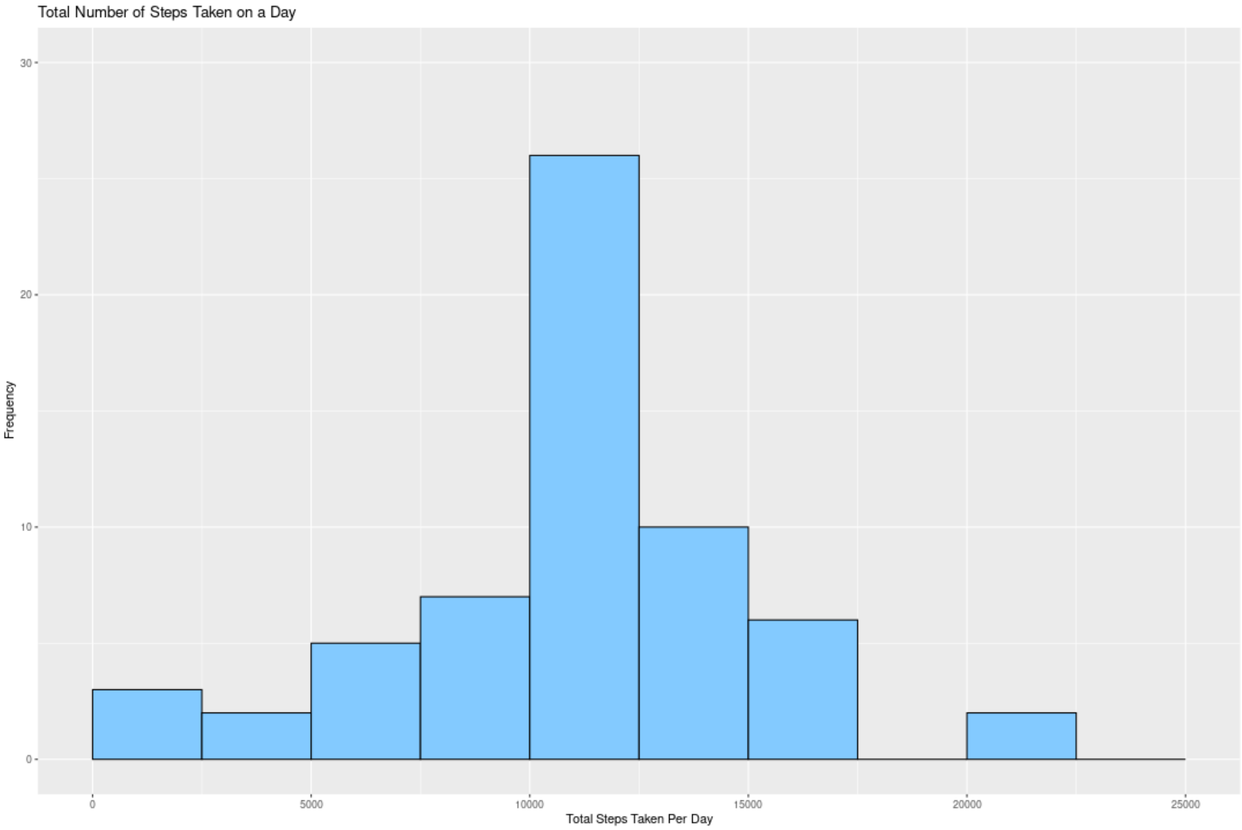
```
activityImputed <- transform(activity,
                             steps = ifelse(is.na(activity$steps), yes = imputedSteps, no = activity$st





totalActivityImputed <- aggregate(steps ~ date, activityImputed, sum)





names(totalActivityImputed) <- c("date", "dailySteps")






sum(is.na(totalActivityImputed$dailySteps))
```

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

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?

```
totalImputedStepsdf <- data.frame(totalActivityImputed)




t <- ggplot(totalImputedStepsdf, aes(x = dailySteps)) +
  geom_histogram(breaks = seq(0, 25000, by = 2500), fill = "#83CAFF", col = "black") +
  ylim(0, 30) +
  xlab("Total Steps Taken Per Day") +
  ylab("Frequency") +
  ggtitle("Total Number of Steps Taken on a Day")
```

```r
print(t)
```



#The mean of the total number of steps taken per day

```r
mean(totalActivityImputed$dailySteps)
```

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

#10766.19

The median of the total number of steps taken per day

```r
median(totalActivityImputed$dailySteps)
```

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

#10766.19

Are there differences in activity patterns between weekdays and weekends?

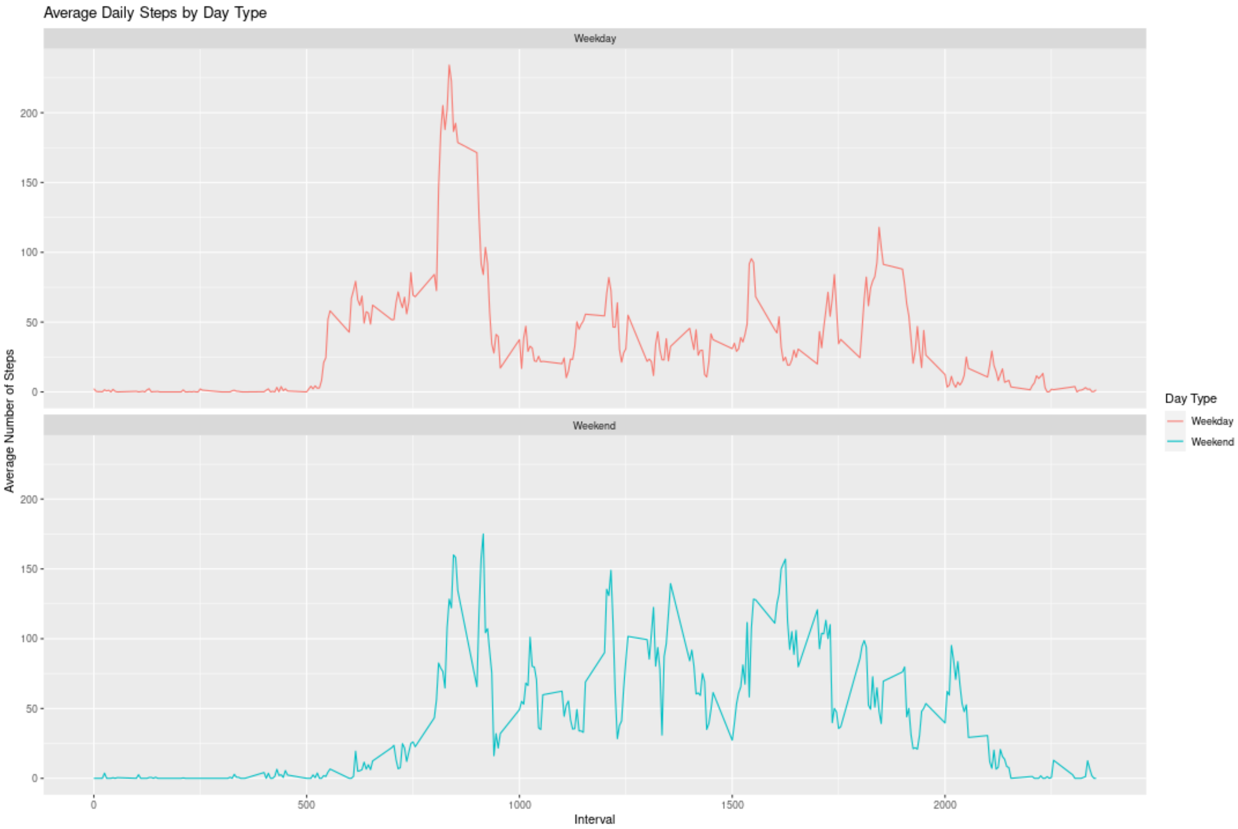
```r
activity$date <- as.Date(strptime(activity$date, format="%Y-%m-%d"))
```

Creating a function that distinguises weekdays from weekends

```
activity$dayType <- sapply(activity$date, function(x) {
  if(weekdays(x) == "Saturday" | weekdays(x) == "Sunday")
  {y <- "Weekend"}
  else {y <- "Weekday"}
  y
})
```

#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.

```
activityByDay <-  aggregate(steps ~ interval + dayType, activity, mean, na.rm = TRUE)




dayPlot <-  ggplot(activityByDay, aes(x = interval , y = steps, color = dayType)) +
  geom_line() + ggtitle("Average Daily Steps by Day Type") +
  xlab("Interval") +
  ylab("Average Number of Steps") +
  facet_wrap(~dayType, ncol = 1, nrow=2) +
  scale_color_discrete(name = "Day Type")





print(dayPlot)
```


Average Daily Steps by Day Type
Weekday
Weekend
Average Number of Steps
0
50
100
150
200
Interval
0
500
1000
1500
2000
Day Type
Weekday
Weekend