Reproducable Research Proj 1

Elizabeth Ellis

August 11, 2020

Loading in Data

First, our data needs to be loaded into R. The data can be found in the github repository in the activity.zip file. The data will be saved in the R object 'data'.

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

Next, we need to clean the data up. The dates need to be converted to the proper date formats. Weekdays are then calculated and added to the dataset

```
activity$date <- as.POSIXct(activity$date, "%Y-%m-%d", tz = "EST")
weekday <- weekdays(activity$date)
activity <- cbind(activity,weekday)
```

Finally, let's load in the necessary R packages!

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

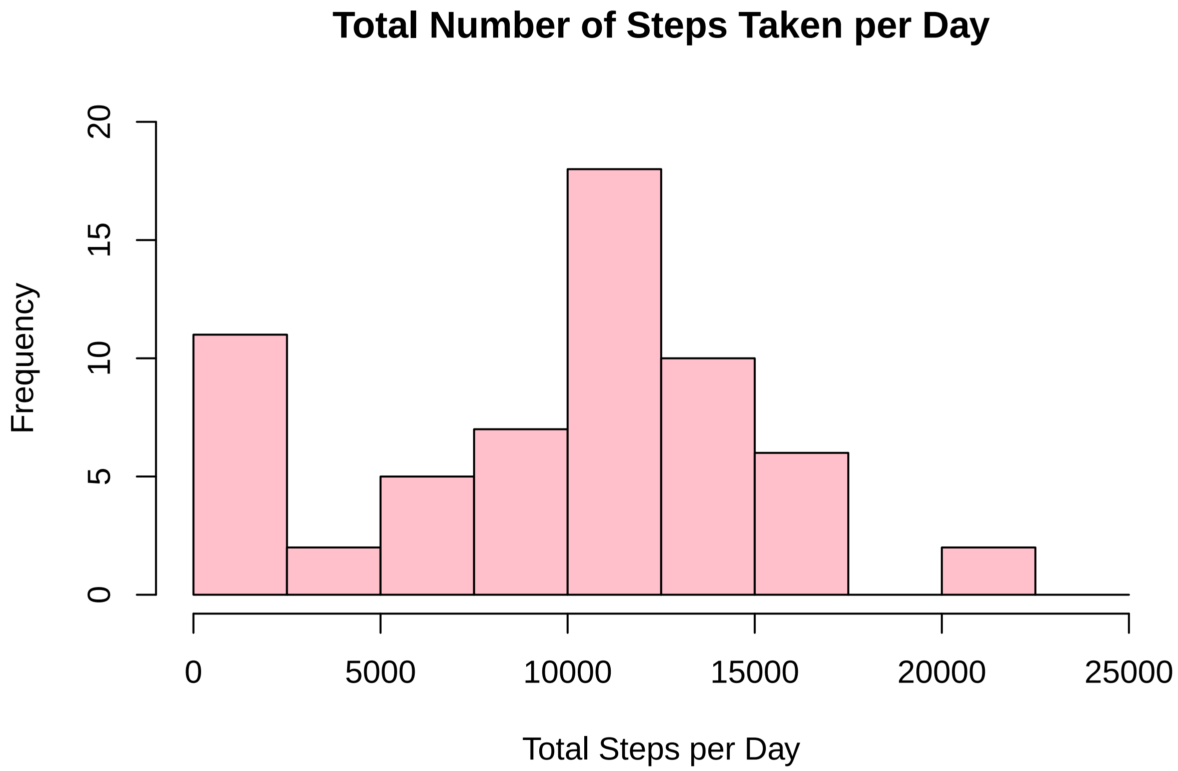
What is the mean total number of steps taken per day?

1. Calculate the total number of steps taken per day

```
activity_total_steps <- with(activity, aggregate(steps, by = list(date), FUN = sum, na.rm = TRUE))
names(activity_total_steps) <- c("date", "steps")
```

2. Create a histogram of the total number of steps taken per day

```
hist(activity_total_steps$steps, main = "Total Number of Steps Taken per Day", xlab = "Total Steps per 
```



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

```
mean1 <- mean(activity_total_steps$steps)
median1 <- median(activity_total_steps$steps)
paste("The mean number of steps taken per a day is", mean1)
```

```
## [1] "The mean number of steps taken per a day is 9354.22950819672"
```

```
paste("The median number of steps taken per a day is", median1)
```

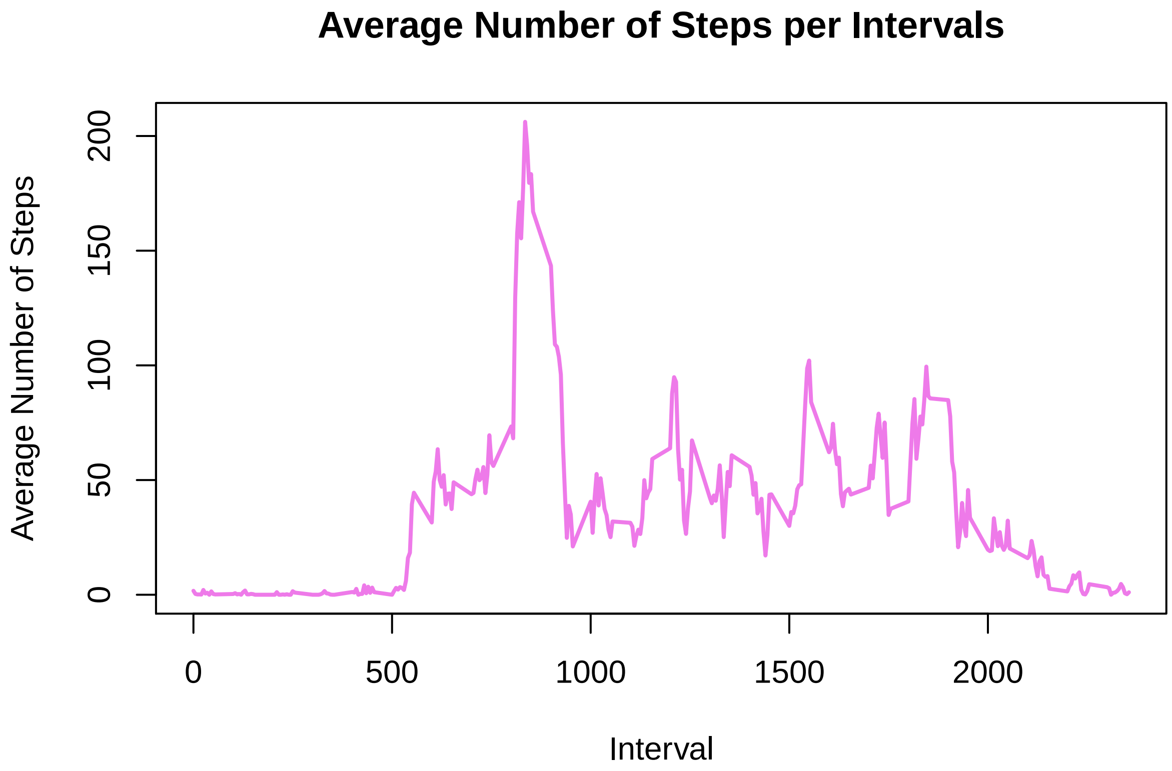
```
## [1] "The median number of steps taken per a day is 10395"
```

What is the average daily activity pattern?

1. Make a time series plot (i.e. `type = "l"`type = “l”) of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis)

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

```
plot(average_daily_activity$interval, average_daily_activity$mean, type = "l", col="orchid2", lwd = 2,
```



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

```
maxint <- average_daily_activity[which.max(average_daily_activity$mean), ]$interval
paste("Interval number", maxint, "has the maximum number of steps.")
```

```
## [1] "Interval number 835 has the maximum number of steps."
```

Imputing Missing Values

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

```
misssteps <- sum(is.na(activity$steps))
paste("The number of missing values is", misssteps)
```

```
## [1] "The number of missing values is 2304"
```

2. Devise a strategy for filling in all of the missing values in the dataset. (Replace missing values with mean score)

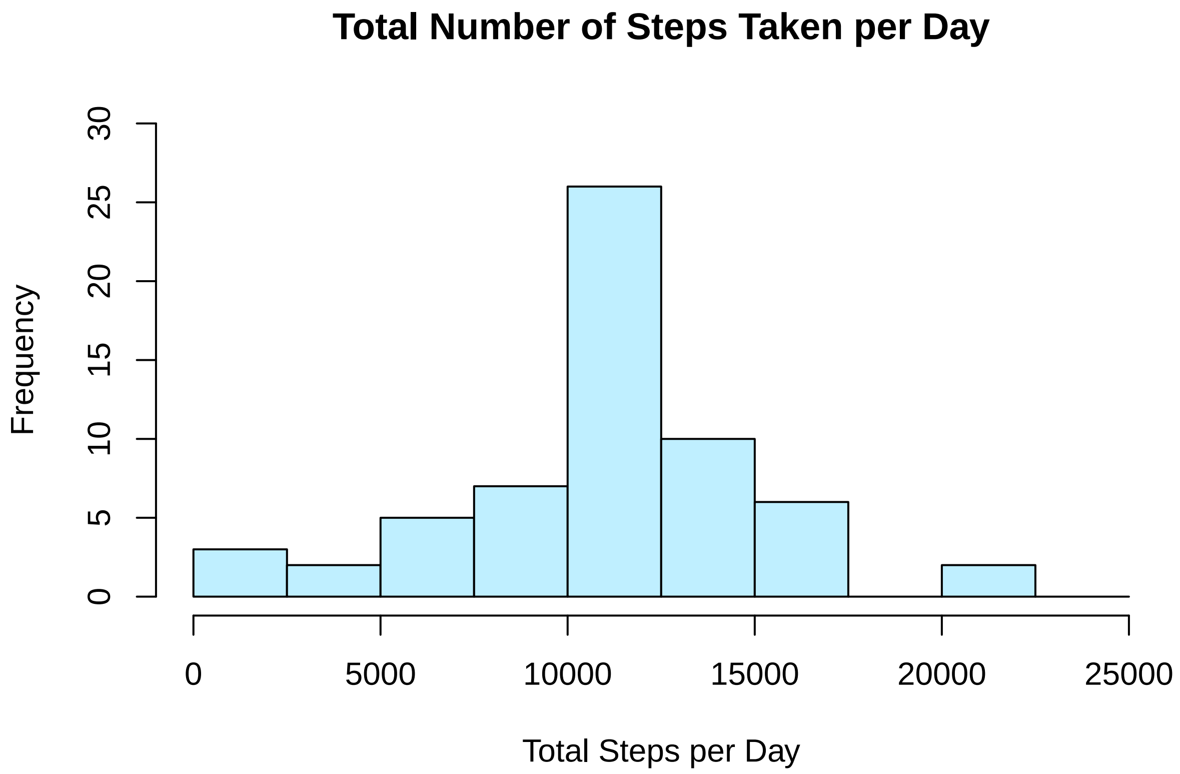
```
imputed_steps <- average_daily_activity$mean[match(activity$interval, average_daily_activity$interval)]
```

3. Create a new dataset with the missing values replaced by the mean

```
activity_imputed <- transform(activity, steps = ifelse(is.na(activity$steps), yes = imputed_steps, no =
total_steps_imputed <- aggregate(steps ~ date, activity_imputed, sum)
names(total_steps_imputed) <- c("date", "daily_steps")
```

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

```r
hist(total_steps_imputed$daily_steps, main = "Total Number of Steps Taken per Day", xlab = "Total Steps
```



5. Calculate the new mean and median

```r
mean2 <- mean(total_steps_imputed$daily_steps)
median2 <- median(total_steps_imputed$daily_steps)
paste("The new mean number of steps taken per a day is", mean2)
```

```
## [1] "The new mean number of steps taken per a day is 10766.1886792453"
```

```r
paste("The new median number of steps taken per a day is", median2)
```

```
## [1] "The new median number of steps taken per a day is 10766.1886792453"
```

6. How are the histogram, mean, and median different from the original dataset?

    - The central bin on the histogram is larger and the mean and median are both higher
    - The mean and median are now the same number.

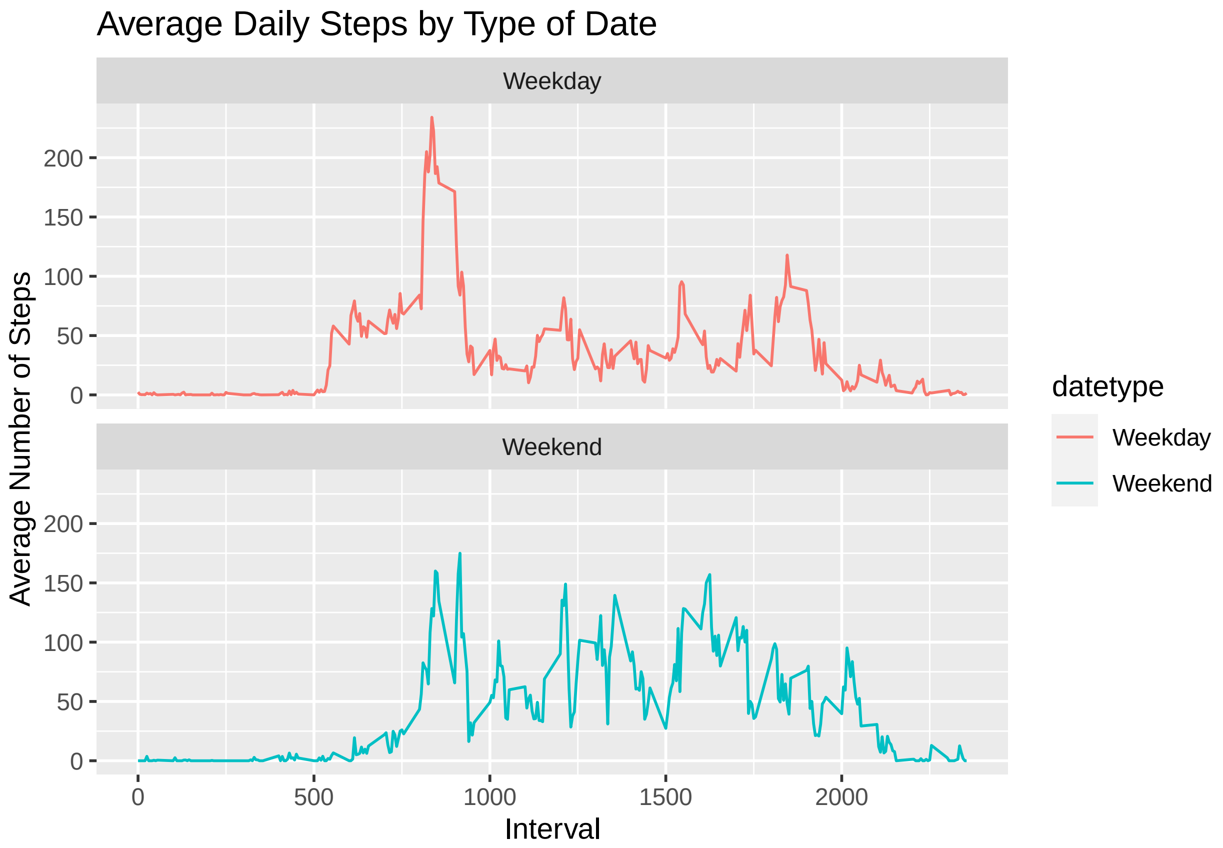
Are there differences in activity patterns between weekdays and weekends?

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.

```r
activity$date <- as.Date(strptime(activity$date, format="%Y-%m-%d"))
activity$datetype <- sapply(activity$date, function(x) {
        if (weekdays(x) == "Saturday" | weekdays(x) == "Sunday")
                {y <- "Weekend"} else
                {y <- "Weekday"}
                y
        })
```

2. Make a panel plot containing a time series plot (i.e. `type = "l"`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).

```
activity_by_date <- with(activity, aggregate(steps~interval + datetype, FUN = mean, na.rm = TRUE))
plot<- ggplot(activity_by_date, aes(x = interval , y = steps, color = datetype)) +
       geom_line() +
       labs(title = "Average Daily Steps by Type of Date", x = "Interval", y = "Average Number of Steps
       facet_wrap(~datetype, ncol = 1, nrow=2)
print(plot)
```


Average Daily Steps by Type of Date
Weekday
Weekend
Average Number of Steps
Interval
datetype
Weekday
Weekend