Reproducible Research Week 2 Course Project 1

Jason Pemberton

28/02/2022

Introduction

It is now possible to collect a large amount of data about personal movement using activity monitoring devices such as a *Fitbit*, *Nike Fuelband*, or *Jawbone Up*. These type of devices are part of the “quantified self” movement – a group of enthusiasts who take measurements about themselves regularly to improve their health, to find patterns in their behaviour, or because they are tech geeks. But these data remain under-utilized both because the raw data are hard to obtain and there is a lack of statistical methods and software for processing and interpreting the data.

This assignment makes use of data from a personal activity monitoring device. This device collects data at 5 minute intervals through out the day. The data consists of two months of data from an anonymous individual collected during the months of October and November, 2012 and include the number of steps taken in 5 minute intervals each day.

The data for this assignment can be downloaded from the course web site:

Dataset: Activity monitoring data (https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip) 52K

The variables included in this dataset are:

steps: Number of steps taking in a 5-minute interval (missing values are coded as NA)

date: The date on which the measurement was taken in YYYY-MM-DD format

interval: Identifier for the 5-minute interval in which measurement was taken

The dataset is stored in a comma-separated-value (CSV) file and there are a total of 17,568 observations in this dataset.

Attach Packages:

```r
library(ggplot2)
library(dplyr)
```

Loading and pre-processing the data.

Download & unzip file for processing. Read and load CSV file into a DataFrame.

```
if(!file.exists("./data")){dir.create("./data")}
fileUrl <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip"
download.file(fileUrl,destfile="./data/activity.zip",method="curl")

unzip(zipfile="./data/activity.zip",exdir="./data")
activity <- read.csv("./data/activity.csv")
activity$date <- as.Date(activity$date)
```

ANALYSIS

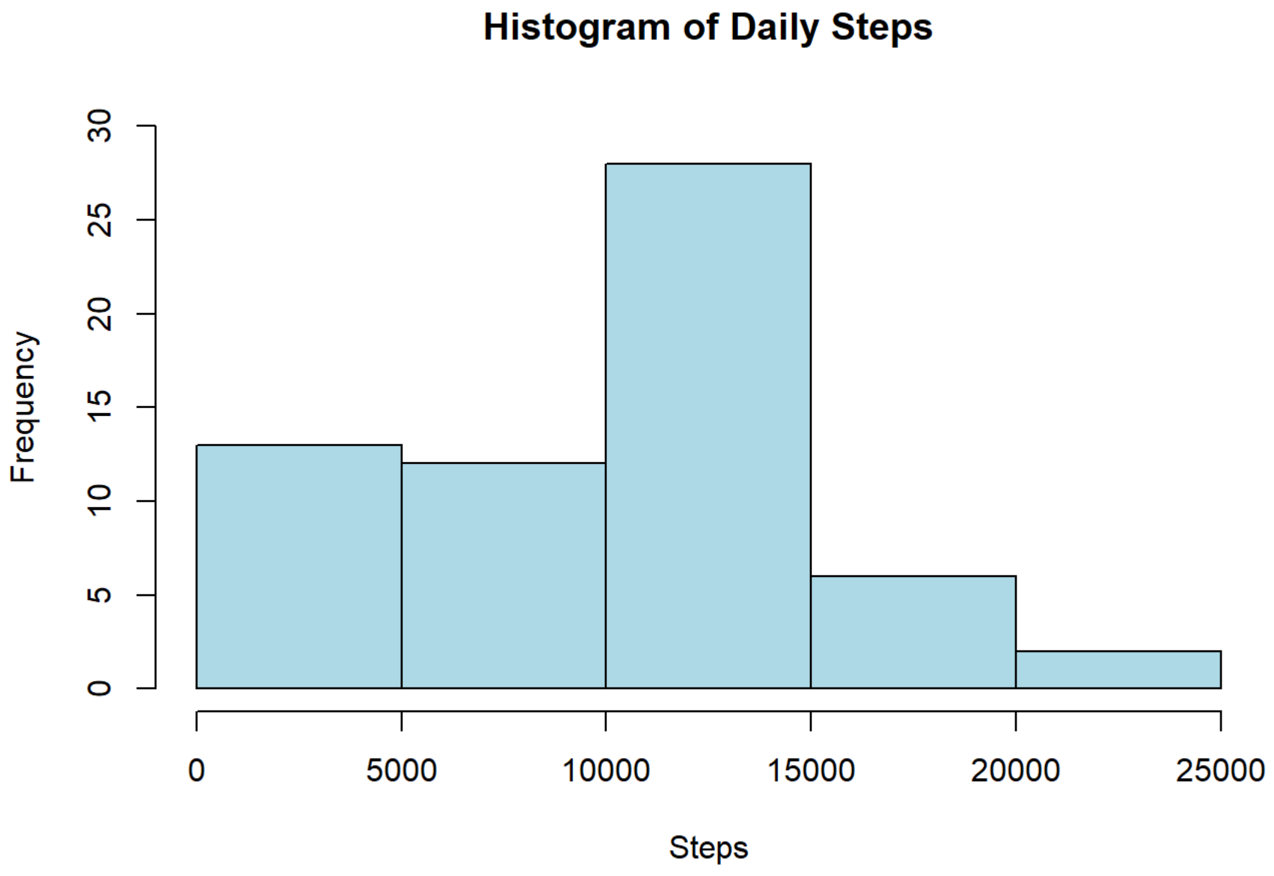
SECTION 1: What is mean total number of steps taken per day?

a. Calculate the total number of steps taken per day

```
stepsPerDay <- activity %>%
        group_by(date) %>%
        summarize(sumsteps = sum(steps, na.rm = TRUE))
```

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

```
hist(stepsPerDay$sumsteps, main = "Histogram of Daily Steps",
     col="lightblue", xlab="Steps", ylim = c(0,30))
```



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

```
meanPreNA <- round(mean(stepsPerDay$sumsteps))
medianPreNA <- round(median(stepsPerDay$sumsteps))

print(paste("The mean is: ", meanPreNA))
```

```
## [1] "The mean is:  9354"
```

```
print(paste("The median is: ", medianPreNA))
```

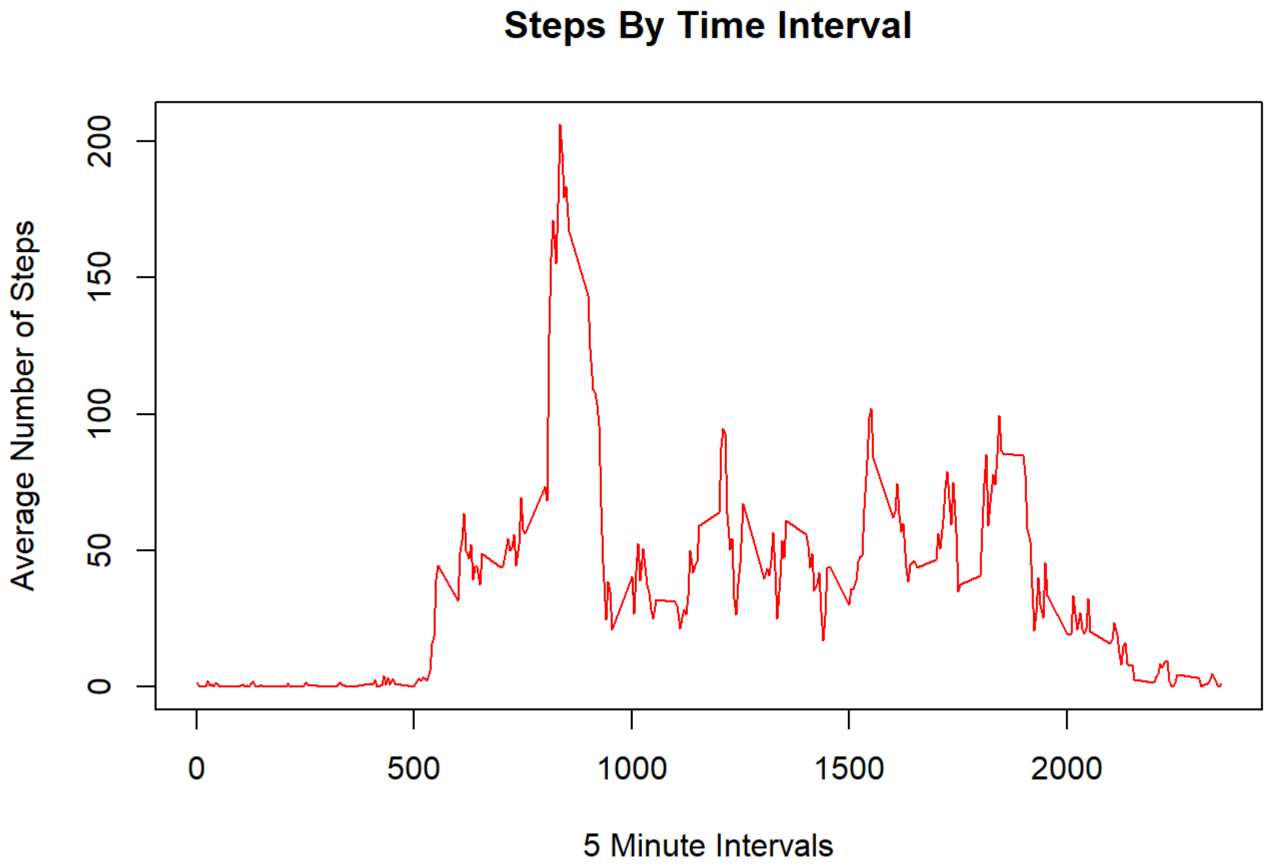
```
## [1] "The median is:  10395"
```

SECTION 2: What is the average daily activity pattern?

a. 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)

```
stepsPerInterval <- activity %>%
        group_by(interval) %>%
        summarize(meansteps = mean(steps, na.rm = TRUE))

plot(stepsPerInterval$meansteps ~ stepsPerInterval$interval,
     col="red", type="l", xlab = "5 Minute Intervals", ylab = "Average Number of Steps",
     main = "Steps By Time Interval")
```



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

```
print(paste("5-Minute Interval containing the most steps on average: ",stepsPerInterval$interval
[which.max(stepsPerInterval$meansteps)]))
```

```
## [1] "5-Minute Interval containing the most steps on average:  835"
```

```
print(paste("Average steps for that interval: ",round(max(stepsPerInterval$meansteps))))
```

```
## [1] "Average steps for that interval:  206"
```

SECTION 3: Return missing values

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

```
print(paste("The total number of rows with NA is: ",sum(is.na(activity$steps))))
```

```
## [1] "The total number of rows with NA is:  2304"
```

b. Devise a strategy for filling in all of the missing values in the dataset. (c) Create a new dataset that is equal to the original dataset but with the missing data filled in.

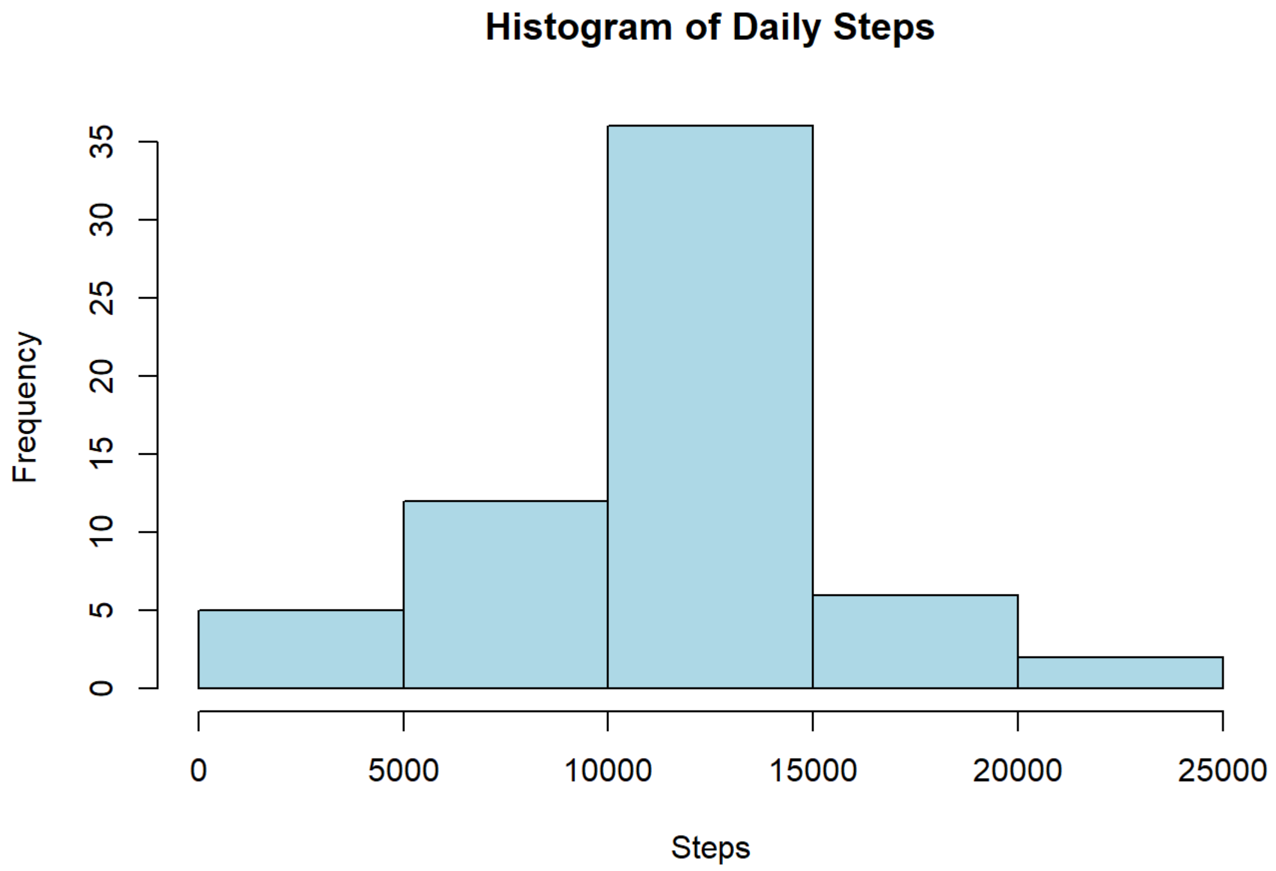
Strategy to solve for missing NA values: The average for the associated interval will be used. The average was built in an earlier step: First, loop through all records of a copy of the 'activity' data. Then, look for records containing NA values. Transform the 'steps' value based on matching the interval in the 'stepsPerInterval' data frame created in a prior step.

```
activityNoNA <- activity
for (i in 1:nrow(activity)){
        if(is.na(activity$steps[i])){
                activityNoNA$steps[i]<- stepsPerInterval$meansteps[activityNoNA$interval[i] == s
tepsPerInterval$interval]
        }
}
```

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

```
stepsPerDay <- activityNoNA %>%
        group_by(date) %>%
        summarize(sumsteps = sum(steps, na.rm = TRUE))

hist(stepsPerDay$sumsteps, main = "Histogram of Daily Steps",
     col="lightblue", xlab="Steps")
```



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

```
meanPostNA <- round(mean(stepsPerDay$sumsteps), digits = 2)
medianPostNA <- round(median(stepsPerDay$sumsteps), digits = 2)

print(paste("The mean is: ", mean(meanPostNA)))
```

```
## [1] "The mean is:  10766.19"
```

```
print(paste("The median is: ", median(medianPostNA)))
```

```
## [1] "The median is:  10766.19"
```

```
NACompare <- data.frame(mean = c(meanPreNA,meanPostNA),median = c(medianPreNA,medianPostNA))
rownames(NACompare) <- c("Pre NA Transformation", "Post NA Transformation")
print(NACompare)
```

```
##                            mean   median
## Pre NA Transformation   9354.00 10395.00
## Post NA Transformation 10766.19 10766.19
```

When you include missing values for all included records you see an increase in both the mean and median. The mean increases from 9354.23 to 10766.19.Note that NA values in the first part of the project were ignored (na.rm = TRUE). Once averages were applied to the missing values the overall mean increased.

SECTION 4: Are there differences in activity patterns between weekdays and weekends?

a. 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.

```
activityDoW <- activityNoNA
activityDoW$date <- as.Date(activityDoW$date)
activityDoW$day <- ifelse(weekdays(activityDoW$date) %in% c("Saturday", "Sunday"), "weekend", "w
eekday")
activityDoW$day <- as.factor(activityDoW$day)
```

b. 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).

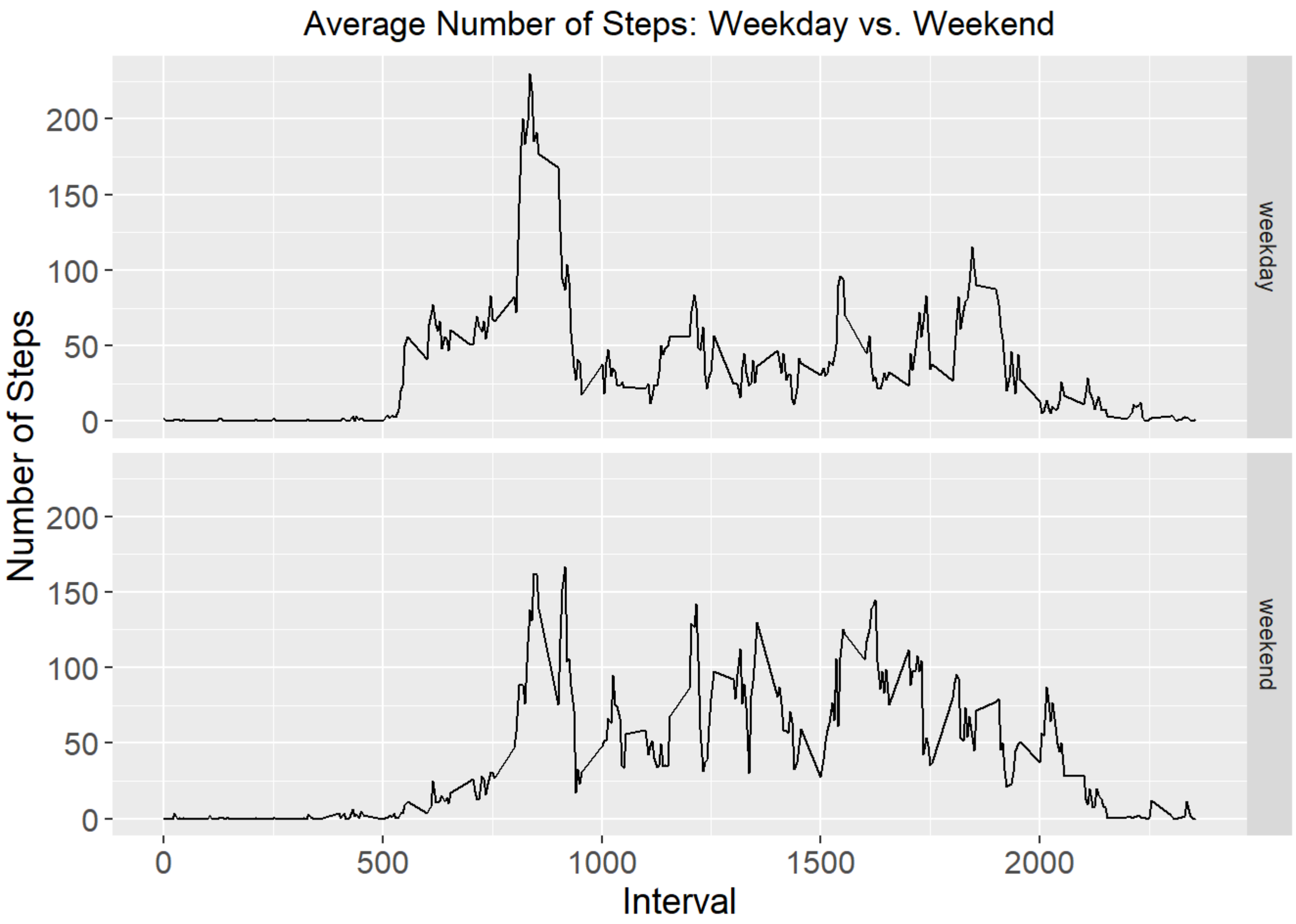
```
activityWeekday <- filter(activityDoW, activityDoW$day == "weekday")
activityWeekend <- filter(activityDoW, activityDoW$day == "weekend")

activityWeekday <- activityWeekday %>%
        group_by(interval) %>%
        summarize(steps = mean(steps))
activityWeekday$day <- "weekday"

activityWeekend <- activityWeekend %>%
        group_by(interval) %>%
        summarize(steps = mean(steps))
activityWeekend$day <- "weekend"

wkdayWkend <- rbind(activityWeekday, activityWeekend)
wkdayWkend$day <- as.factor(wkdayWkend$day)


g <- ggplot (wkdayWkend, aes (interval, steps))
g + geom_line() + facet_grid (day~.) +
        theme(axis.text = element_text(size = 12),axis.title = element_text(size = 14)) +
        labs(y = "Number of Steps") + labs(x = "Interval") +
        ggtitle("Average Number of Steps: Weekday vs. Weekend") +
        theme(plot.title = element_text(hjust = 0.5))
```



The visualizations shows slight differences in the step patterns throughout the average daily intervals. Weekdays show a large spike in early morning which could coincide with people walking to work/school or transit stations. While step counts on weekends are more consistent throughout the day.