Reproducible Research Course Project 1

Ryan Kramer

7/6/2017

Data For the Analysis

The data can be found on the course web site:

Dataset: Activity Monitoring Data

The variables included are:
date: The date, in YYYY-MM-DD format, the measurement was taken.
interval: Identifier for the five-minute interval the measurement was taken.
steps: Number of steps in the five-minute interval. Missing measurements are coded as NA.
The dataset is a .csv file with 17,568 observations.

Loading the Data

```
setwd("/Users/RyanKramer/Desktop")
library(ggplot2)
library(plyr)
activity <- read.csv("activity.csv")
cleanData <- activity[!is.na(activity$steps),]
```

Processing the Data

```
activity$day <- weekdays(as.Date(activity$date))
activity$DateTime <- as.POSIXct(activity$date, format = "%Y-%m-%d")
cleanData$day<- weekdays(as.Date(cleanData$date))
cleanData$DateTime <- as.POSIXct(cleanData$date, format = "%Y-%m-%d")
```

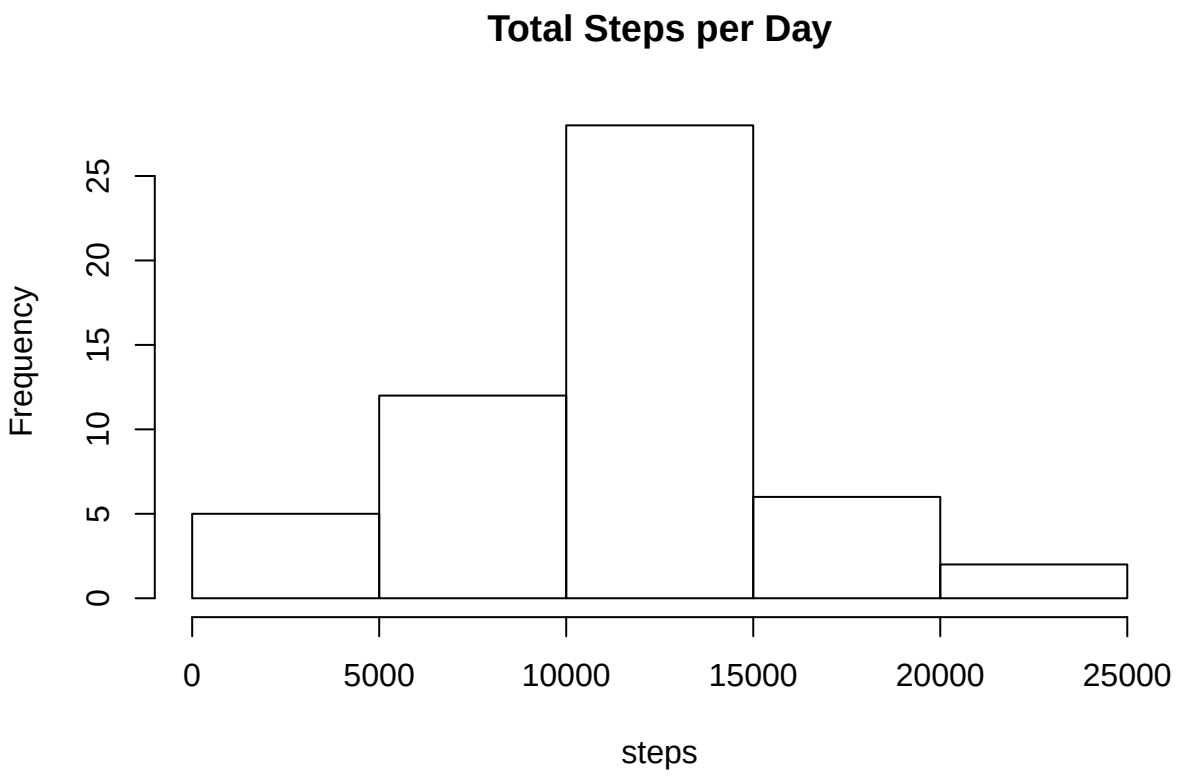
What is the mean total number of steps taken per day?

Total number of steps taken per day:

```
sumSteps <- aggregate(activity$steps ~ activity$date, FUN = sum, )
colnames(sumSteps) <- c("date", "steps")
```

Histogram of number of steps taken each day.

```
hist(sumSteps$steps, xlab="steps", main = "Total Steps per Day", breaks = 5)
```



Calculations for median and mean of number of steps taken per day.

```
##Mean of Steps
as.integer(mean(sumSteps$steps))
```

```
## [1] 10766
```

```
##Median of Steps
as.integer(median(sumSteps$steps))
```

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

The mean number of steps taken each day is 10,766 steps. The median number of steps taken each day is 10,765 steps.

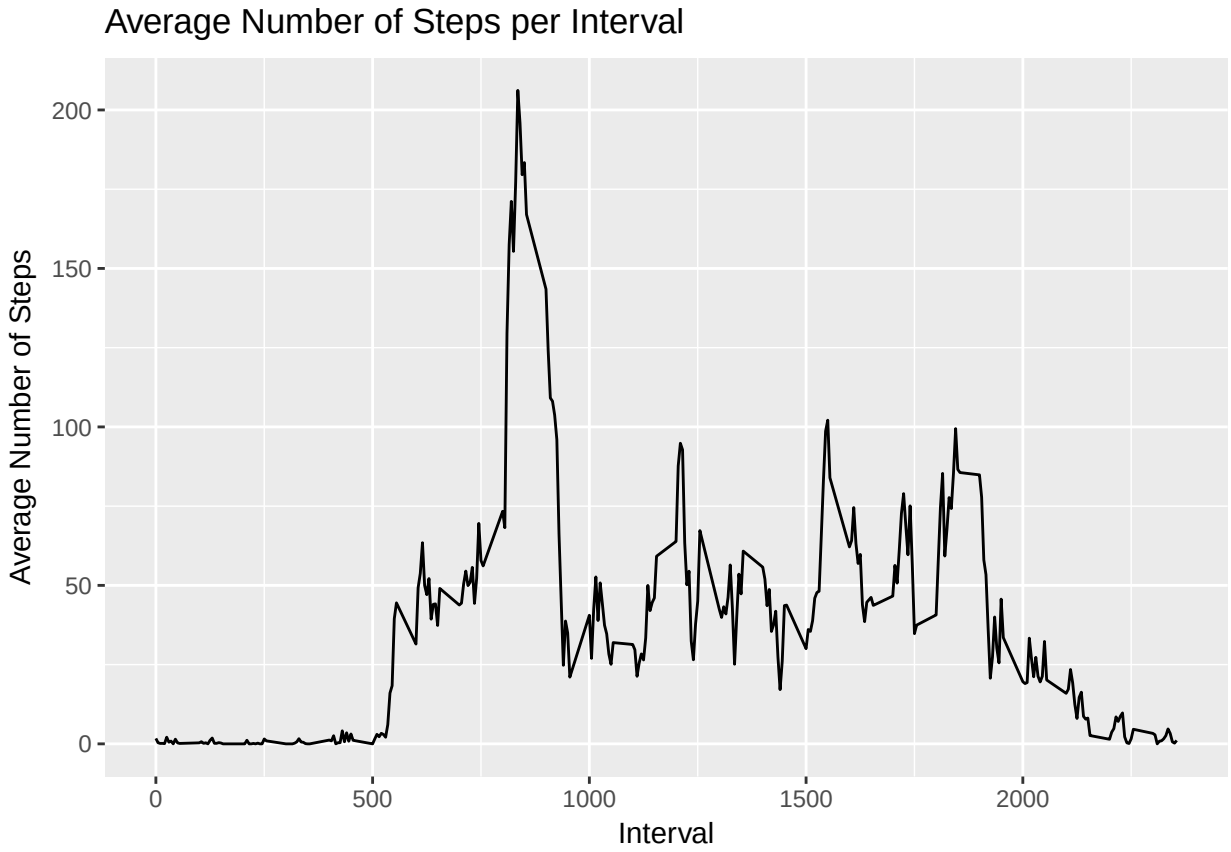
What is the average daily activity pattern?

Time series plot of the five-minute interval and the average number of steps taken, averaged across all days.

```
library(plyr)
library(ggplot2)

#Average number of steps per interval
intervalData <- ddply(cleanData, .(interval), summarize, Avg = mean(steps))

#Line plot of average number of steps per interval
p <- ggplot(intervalData, aes(x=interval, y=Avg), xlab = "Interval", ylab = "Average Number of Steps")
p + geom_line() + xlab("Interval") + ylab("Average Number of Steps")+ ggtitle("Average Number of Steps
```



Which five-minute interval contains the maximum number of steps?

```
##Max step by interval
maxSteps <- max(intervalData$Avg)
##Which interval contains max average number of steps?
intervalData[intervalData$Avg==maxSteps, 1]
```

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

The max number of steps for a five-minute interval is 206 steps. The five-minute interval with the max number of steps was the 835 interval.

Imputing Missing Values

Calculate and report the total number of missing values in the dataset

```
nrow(activity[is.na(activity$steps),])
```

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

The total number of rows with steps = NA is 2304.

One strategy for filling in NAs is to substitute the missing steps with the average five-minute interval based on the day of the week.

```
avgTable <- ddply(cleanData, .(interval, day), summarize, Avg= mean(steps))

##Create data with all NAs for substitution
```

```
naTable <- activity[is.na(activity$steps),]
newTable <- merge(naTable, avgTable, by = c("interval", "day"))
```

Create new dataset that is the original dataset with the missing data filled in.

```
newData <- newTable[, c(6,4,1,2,5)]
colnames(newData) <- c("steps", "date", "interval", "day", "DateTime")
mergeData <- rbind(cleanData, newData)
```

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?

```
sumSteps2 <- aggregate(mergeData$steps ~ mergeData$date, FUN = sum)
colnames(sumSteps2) <- c("date", "steps")

## Mean steps with no more NA data.
as.integer(mean(sumSteps2$steps))
```

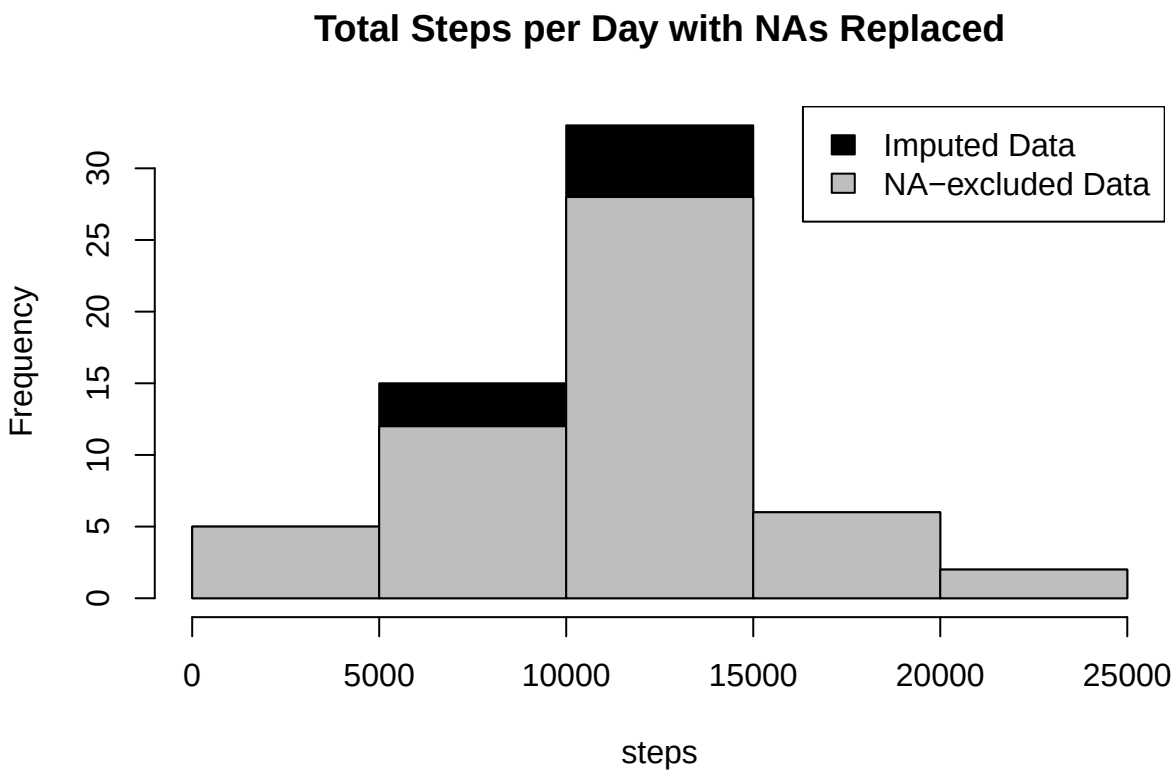
```
## [1] 10821
```

```
##Median steps with no more NA data.
as.integer(median(sumSteps2$steps))
```

```
## [1] 11015
```

The mean steps taken per day increased by 55 steps. The median steps taken per day increased by 250 steps.

```
##Histogram showing total steps per day.
hist(sumSteps2$steps, breaks = 5, xlab = "steps", main = "Total Steps per Day with NAs Replaced", col =
hist(sumSteps$steps, breaks = 5, xlab = "steps", main = "Total Steps per day with NAs Replaced", col =
legend("topright", c("Imputed Data", "NA-excluded Data"), fill = c("black", "grey"))
```



The overall shape of the distribution did not change.

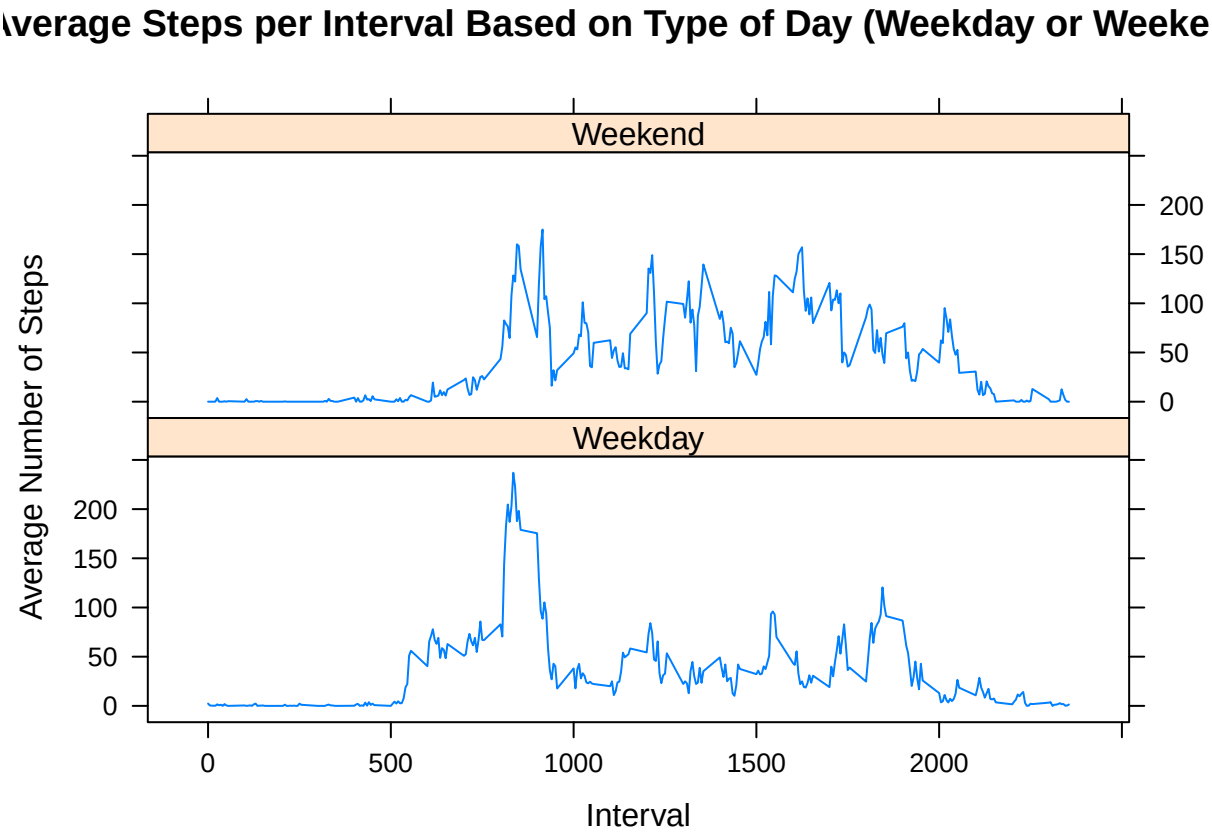
Are there differences in activity patterns between weekdays and weekends?

Create factor variable in dataset with two levels- weekdays and weekends. Indicates whether a date is a weekday or weekend day.

```
mergeData$DayLevel <- ifelse(mergeData$day %in% c("Saturday", "Sunday"), "Weekend", "Weekday")
```

Make a panel plot containing a time series plot of the five-minute interval and the average number of steps taken, averaged across all weekdays or weekend days.

```
library(lattice)
intervalData2<- ddply(mergeData, .(interval, DayLevel), summarize, Avg = mean(steps))
xyplot(Avg~interval|DayLevel, data = intervalData2, type = "l", layout = c(1,2), main = "Average Steps
```



The activity trends are different based on whether the date is a weekend or not.