Peer Assessment 1

The Student

Saturday, July 18, 2015

- Loading and Processing the data

1. Load the data

We will unzip and load the data in file activity.csv. Here we have set the working directory so that the markdowm file is in the same directory as the data file.

```
data <- read.csv(unzip("./activity.zip"), header = TRUE, sep=",")
```

2. Transforming the date column as necessary.

```
data$date <- as.Date(data$date)
```

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

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

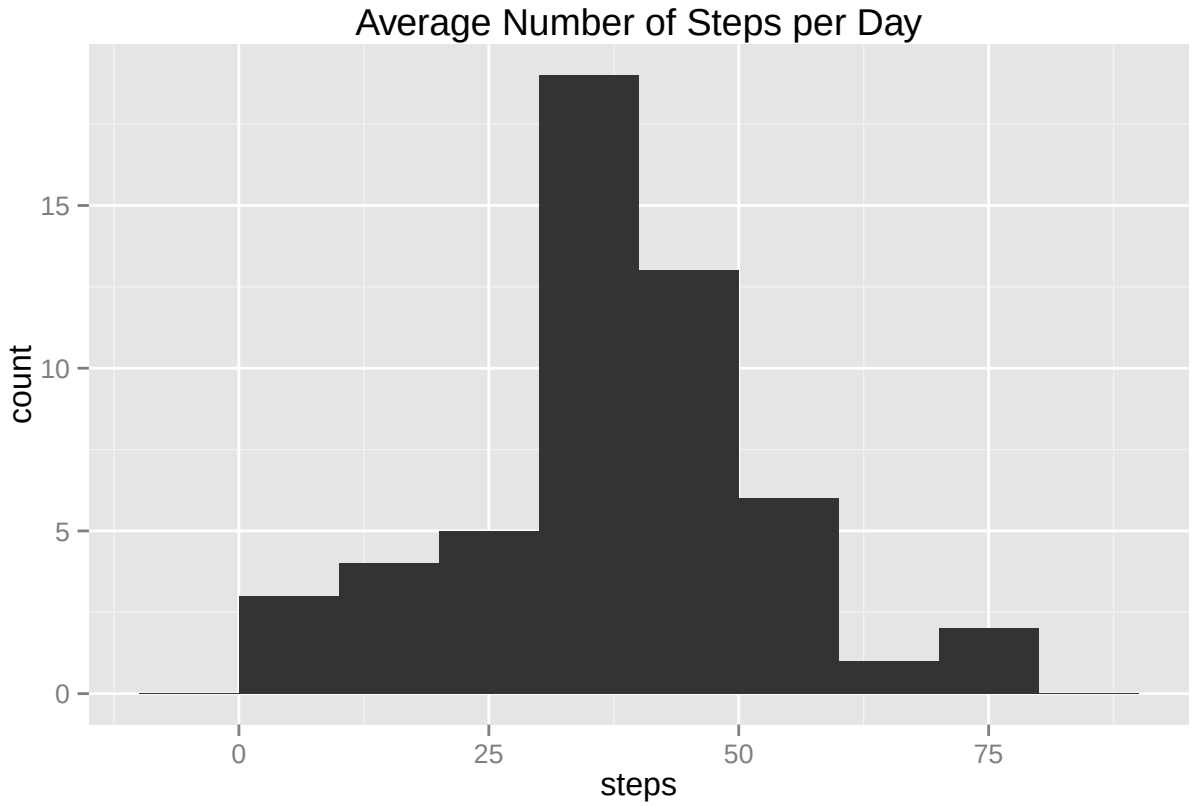
I will use the ggplot2 package here.

We aggregate the number of steps by date and find the average. Then we plot the histogram of the result.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.1.2
```

```
steps_Per_day <- aggregate(data=data,steps~date,FUN=mean);  qplot(steps, data=steps_Per_day, geom="hist
```



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

First the sum of all steps per day (total) will be calculated. Then mean and median wil be computed.

```
total_daily_steps<-aggregate(steps~date,data=data,FUN=sum)
cat("The first and last five total number of steps taken daily are respectively:","\n")
```

```
## The first and last five total number of steps taken daily are respectively:
```

```
head(total_daily_steps)
```

```
##         date steps
## 1 2012-10-02   126
## 2 2012-10-03 11352
## 3 2012-10-04 12116
## 4 2012-10-05 13294
## 5 2012-10-06 15420
## 6 2012-10-07 11015
```

```
tail(total_daily_steps)
```

```
##          date steps
## 48 2012-11-24 14478
```

```
## 49 2012-11-25 11834
## 50 2012-11-26 11162
## 51 2012-11-27 13646
## 52 2012-11-28 10183
## 53 2012-11-29  7047
```

```
mean_total <- mean(total_daily_steps$steps)
median_total <- median(total_daily_steps$steps)

cat("Mean total number of steps: ", mean_total)
```

```
## Mean total number of steps:  10766
```

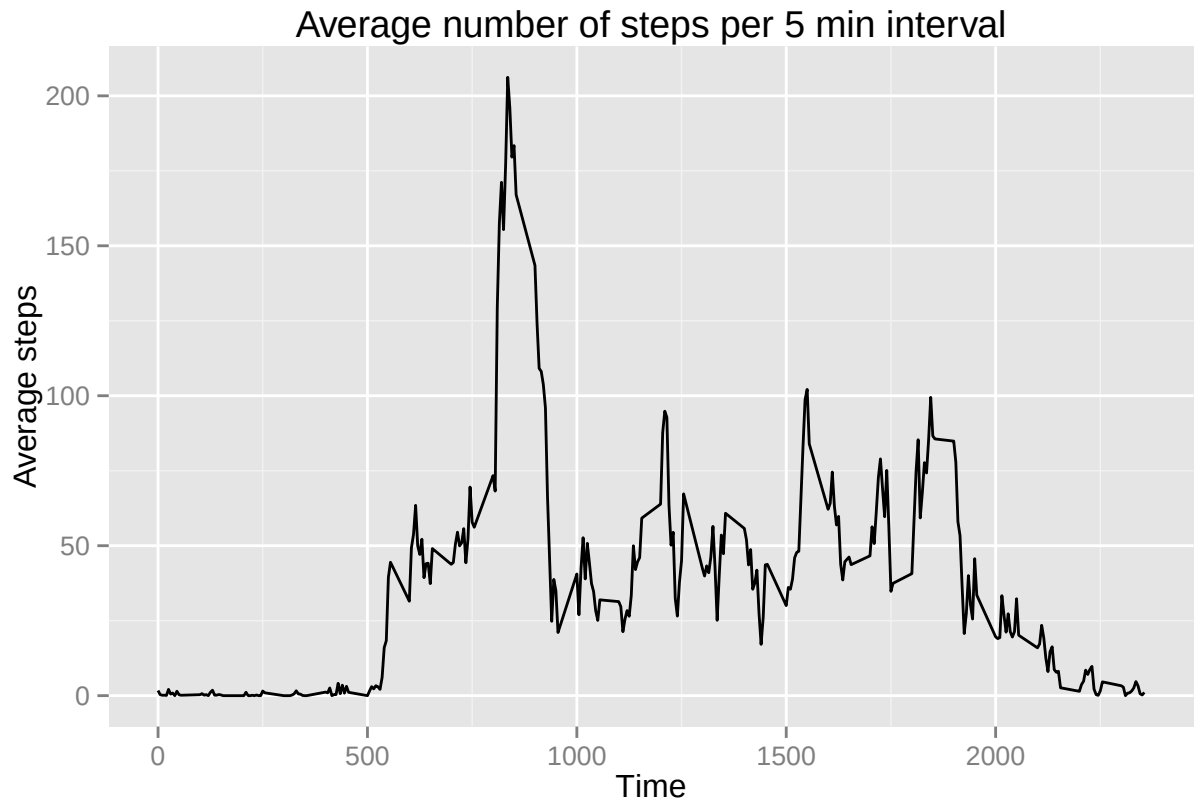
```
cat("Median total number of steps: ", median_total)
```

```
## Median total number of steps:  10765
```

- What is the average daily activity pattern?

1. Make a time series plot of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis).

```
timeseries <- aggregate(steps~interval, data=data, FUN=mean)
qplot(interval, steps, data=timeseries, geom="line", main="Average number of steps per 5 min interval",
```



2.

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

```
maximum_position = which.max(timeseries$steps)
message(c("The maximum number of steps is in the interval ", maximum_position, ", which correspond to "
```

```
## The maximum number of steps is in the interval 104, which correspond to 835 min.
```

- 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).

The NAS are available only in the steps. So we count these.

```
cat("There are", sum(is.na(data$steps)),"missing values.")
```

```
## There are 2304 missing values.
```

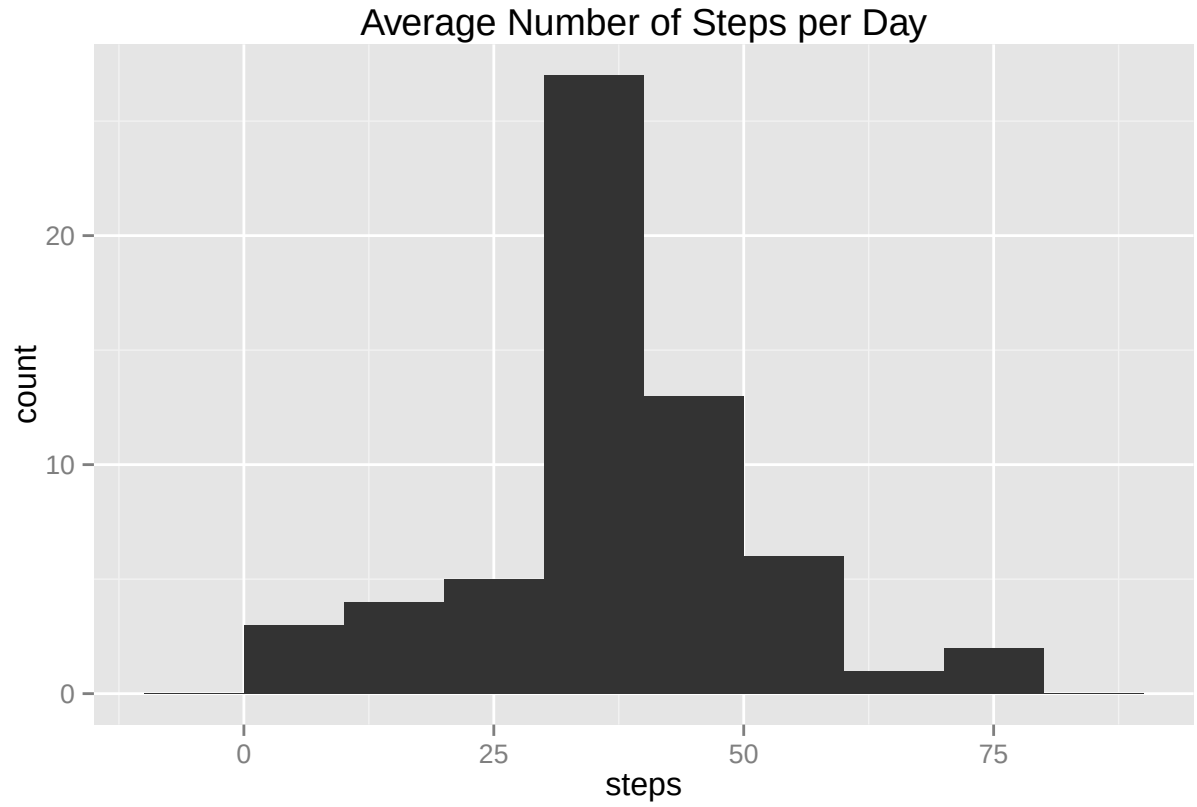
2. Devise a strategy for filling in all of the missing values in the dataset.
3. Create a new dataset that is equal to the original dataset.

I will fill the missing value with the average number of steps in that interval, which was calculated in a previous exercise and is stored in the variable timeseries. The following for loop does just that.

```
newdata <- data
ind <- 1:nrow(data)
for (i in ind)
{
  if (is.na(data$steps[i]))
  {
    newdata$steps[i] <- timeseries$steps[which(timeseries$interval==data$interval[i])]
  }
}
```

4. 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.

```
Asteps_day <- aggregate(steps~date, data=newdata,FUN=mean)
qplot(steps, data=Asteps_day, geom="histogram", binwidth=10, main="Average Number of Steps per Day")
```




```
Daily_total_steps <- aggregate(steps~date, data=newdata,FUN=sum)
mean_total <- mean(Daily_total_steps$steps)
median_total <- median(Daily_total_steps$steps)
cat("Mean total number of steps: ", mean_total)
```

```
## Mean total number of steps:  10766
```

```
cat("Median total number of steps: ", median_total)
```

```
## Median total number of steps:  10766
```

The means are the same in both cases of imputed and missing values. The medians are different though. The new median is higher by 1 and is now equal to the mean.

- Are there differences in activity patterns between weekdays and weekends?

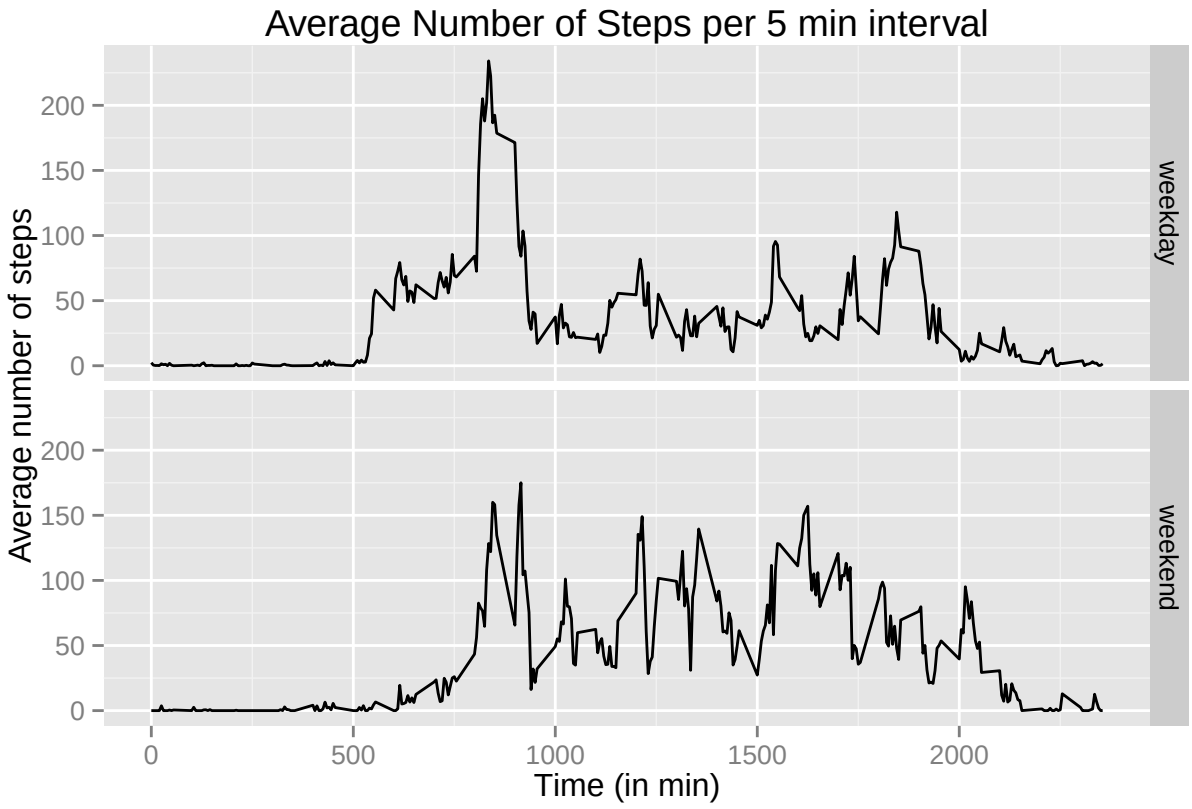
1. Create a new factor variable in the dataset with two levels - “weekday” and “weekend” indicating whether a given date is a weekday or weekend day.

```
daytype <-""
weekend <- c("Saturday", "Sunday")
for (i in ind)
{
```

```
  if (sum(weekdays(data$date[i]) == weekend)>0)
  {
    daytype[i] = "weekend"
  }
  else
    daytype[i] = "weekday"
}
data <- cbind(data, daytype)
```

2. 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.

```
weekday_dat<-aggregate(steps~interval, data=data, subset=daytype=="weekday", FUN=mean)
weekend_dat<-aggregate(steps~interval, data=data, subset=daytype=="weekend", FUN=mean)
weekday_dat$daytype <- "weekday"
weekend_dat$daytype <- "weekend"
qplot(interval, steps, data=rbind(weekend_dat, weekday_dat), geom="line", facets=daytype~., main="Avera
```



The individual seems to start activity at a later time interval in the weekend. It also ends later. In the mid-day intervals, the individual has an increased activity in the weekend as compared to the weekdays.