Peer Assessment 1

Jorge Salvador Paredes Merino

Sunday, October 19, 2014

Loading and preprocessing the data

Show any code that is needed to:

1. Load the data (i.e. read.csv())
2. Process/transform the data (if necessary) into a format suitable for your analysis

```
library(ggplot2) # use in plot
library(lattice) # use in histogram
library(dplyr)   # use in preprocessing
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Get csv dataset from the zip file
data = read.csv(unz("activity.zip", "activity.csv"), header=TRUE,
                colClasses = c("integer","Date","integer"))
```

What is mean total number of steps taken per day?

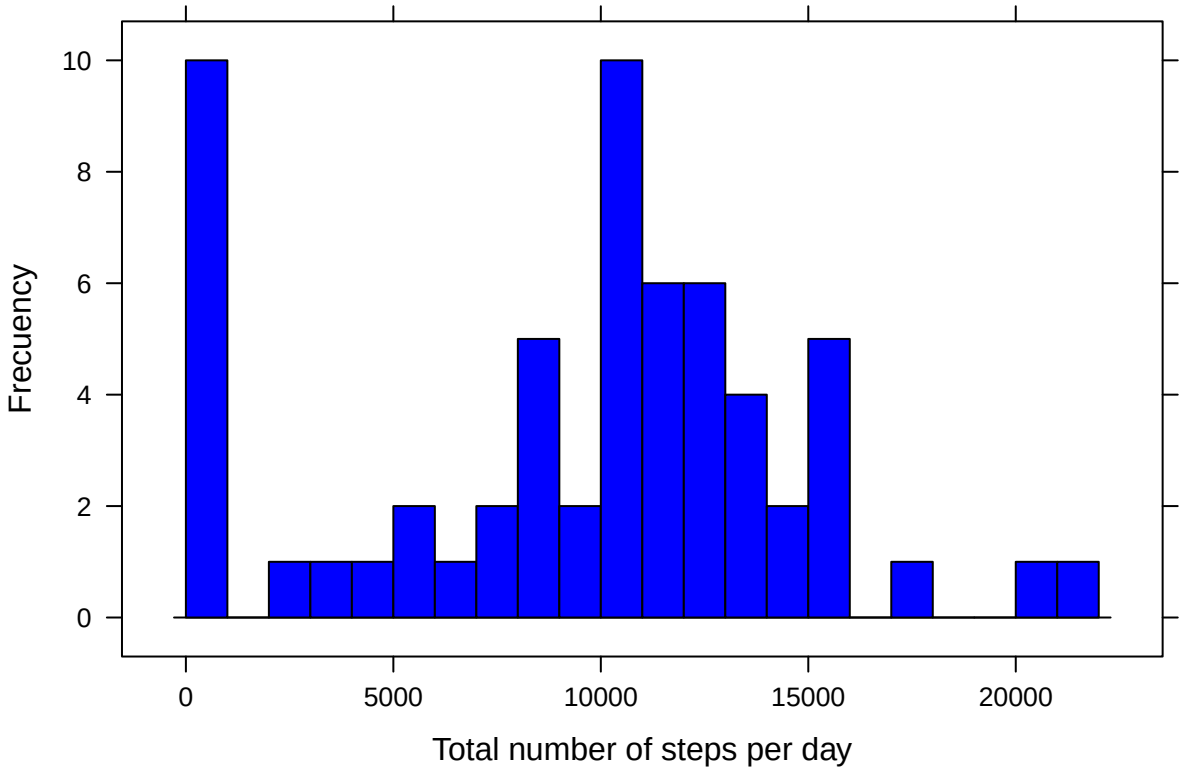
For this part of the assignment, you can ignore the missing values in the dataset.

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

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

```
# Ignoring the missing values in the dataset.
steps_by_day = data %>% group_by(date) %>% summarize(Steps = sum(steps, na.rm=TRUE))

# Histogram of the total number of steps taken each day
histogram(steps_by_day$Steps, breaks = 20,
          xlab = "Total number of steps per day",
          ylab = "Frecuency",
          col = "blue", type = "count")
```




```
# Mean and median of the total number of steps taken per day
mean(steps_by_day$Steps, na.rm = TRUE)    # mean required
```

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

```
median(steps_by_day$Steps, na.rm = TRUE) # median required
```

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

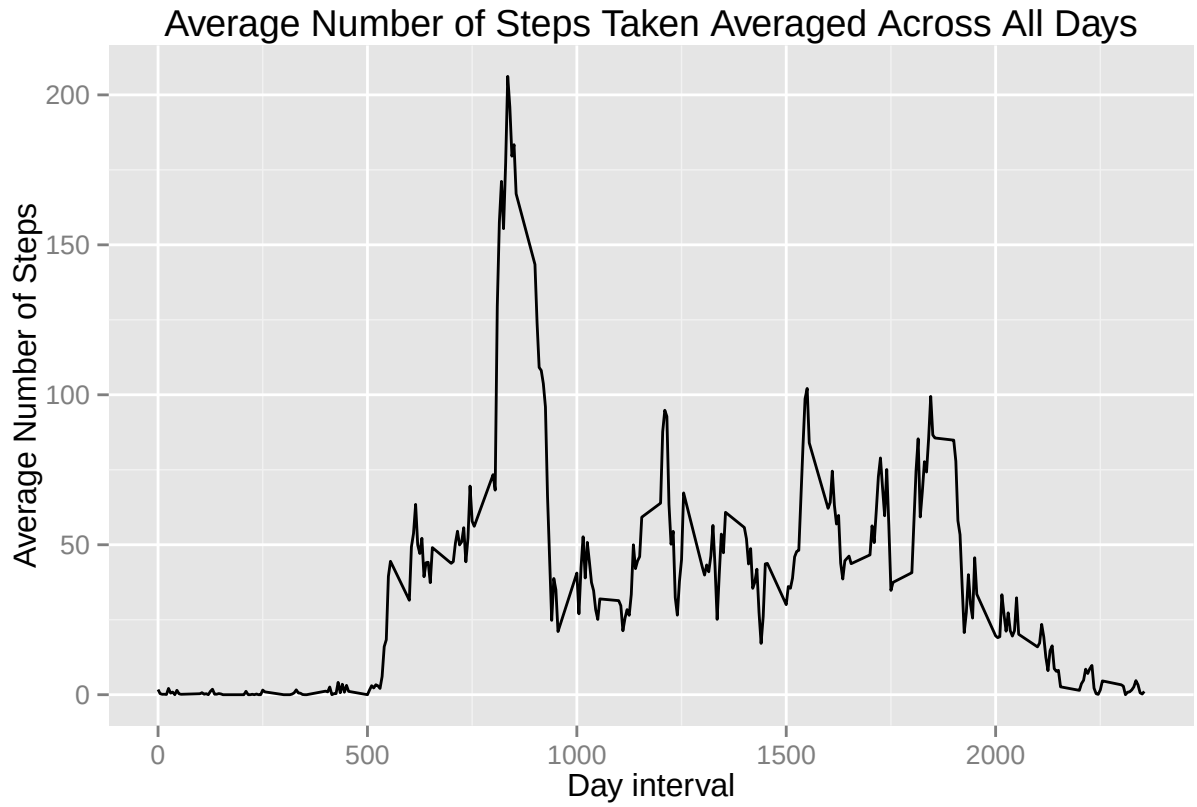
Mean and median are not equals by they are relative close.

What is the average daily activity pattern?

1. 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)
2. Which 5-minute interval, on average across all the days in the dataset, contains the maximum number of steps?

```
# Calculate average steps for each of 5-minute interval during a 24-hour period
avg_daily_activity = data %>% group_by(interval) %>%
  summarize(meanSteps = mean(steps, na.rm = TRUE))
```

```
qplot(x=interval, y=meanSteps, data = avg_daily_activity,  geom = "line",
      xlab="Day interval",
      ylab="Average Number of Steps",
      main="Average Number of Steps Taken Averaged Across All Days"
)
```



The maximum average number of steps is 206.17 which belongs to interval 835

Imputing missing values

Note that there are a number of days/intervals where there are missing values (coded as NA). The presence of missing days may introduce bias into some calculations or summaries of the data.

1. Calculate and report the total number of missing values in the dataset (i.e. the total number of rows with NAs)
2. Devise a strategy for filling in all of the missing values in the dataset. The strategy does not need to be sophisticated. For example, you could use the mean/median for that day, or the mean for that 5-minute interval, etc.
3. Create a new dataset that is equal to the original dataset but with the missing data filled in.
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. 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?

```r
# Total number of missing values
sum(is.na(data$steps))
```

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

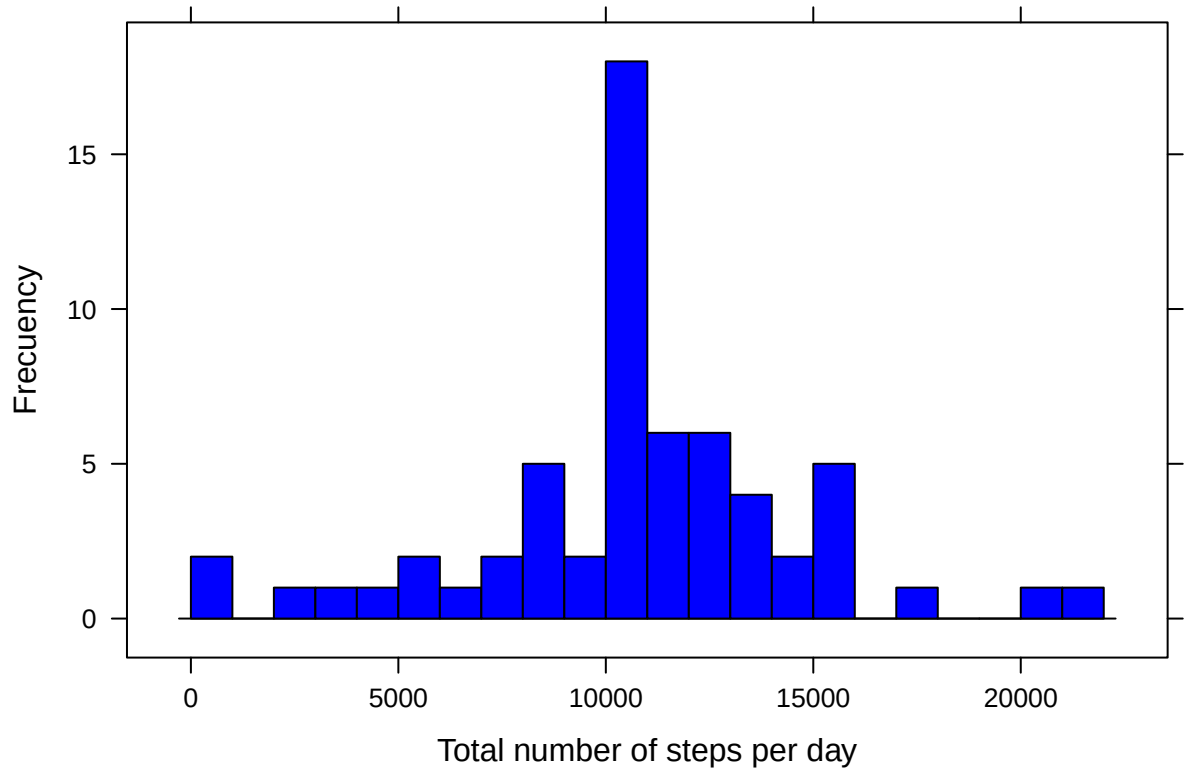
This quantity of missing values represents 13.11% of the total measurements. In this case, NAs will be filled with the mean of intervals through all days.

```r
# Strategy: Impute missing values with means
aux = data %>% left_join(avg_daily_activity, by = "interval")
aux$fillSteps = ifelse(is.na(aux$steps), aux$meanSteps, aux$steps)
aux$steps = NULL; aux$meanSteps = NULL

colnames(aux) = c("interval", "date", "steps")

# Steps by day without missing values
steps_by_day_2 = aux %>% group_by(date) %>% summarize(Steps = sum(steps))

# Histogram
histogram(steps_by_day_2$Steps, breaks = 20,
          xlab = "Total number of steps per day",
          ylab = "Frecuency",
          col = "blue", type = "count")
```



```r
# Mean and median
mean(steps_by_day_2$Steps, na.rm = TRUE)    # mean required
```

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

```r
median(steps_by_day_2$Steps, na.rm = TRUE) # median required
```

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

Owing to imputatation of many missing values with the means, we got identicals mean and median required. Also, the distribution in the middle doest change a lot with respect of the first histogram. The difference is noted in lowest values of steps per day, probably these chunks contains more missing values.

Are there differences in activity patterns between weekdays and weekends?

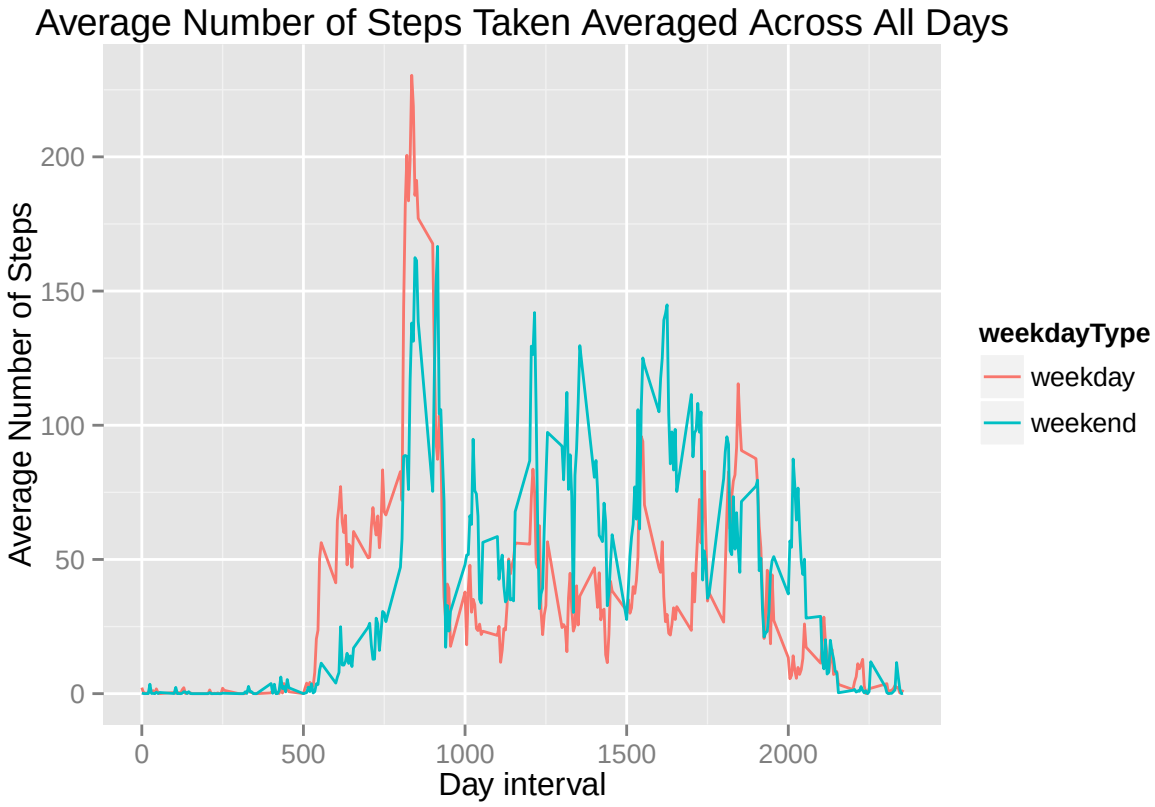
```r
Sys.setlocale("LC_TIME", "English") # To ensure english weekdays
```

```
## [1] "English_United States.1252"
```

```r
aux$weekdayType = ifelse( weekdays(aux$date) %in% c("Saturday", "Sunday"), "weekend", "weekday")

data_intervales = aux %>% group_by(interval, weekdayType) %>%
  summarize(meanSteps = mean(steps, na.rm = TRUE))

qplot(x=interval, y=meanSteps, data = data_intervales,  colour = weekdayType, geom = "line",
      xlab="Day interval",
      ylab="Average Number of Steps",
      main="Average Number of Steps Taken Averaged Across All Days"
)
```



With respect of last graphic, behaviours are different. Weekend has less steps than weekday in the first hours maybe by the need to go early to work. Furthermore, we can think that people tend to sleep more hours in the weekend.