Course Project 1

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 behavior, 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.

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.4

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Load Data

```r
data_activity<-read.table("./activity.csv",sep=",",header=TRUE)
head(data_activity)
```

```
##   steps       date interval
## 1    NA 2012-10-01        0
## 2    NA 2012-10-01        5
## 3    NA 2012-10-01       10
## 4    NA 2012-10-01       15
## 5    NA 2012-10-01       20
## 6    NA 2012-10-01       25
```

```r
str(data_activity)
```

```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : Factor w/ 61 levels "2012-10-01","2012-10-02",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
```

Set date as date

```r
data_activity$date<-as.Date(data_activity$date,'%Y-%m-%d')
```

Remove NAs

```
index_NA<-is.na(data_activity$steps)
data_activity<-data_activity[!index_NA,]
```

Question 1

What is mean total number of steps taken per day?

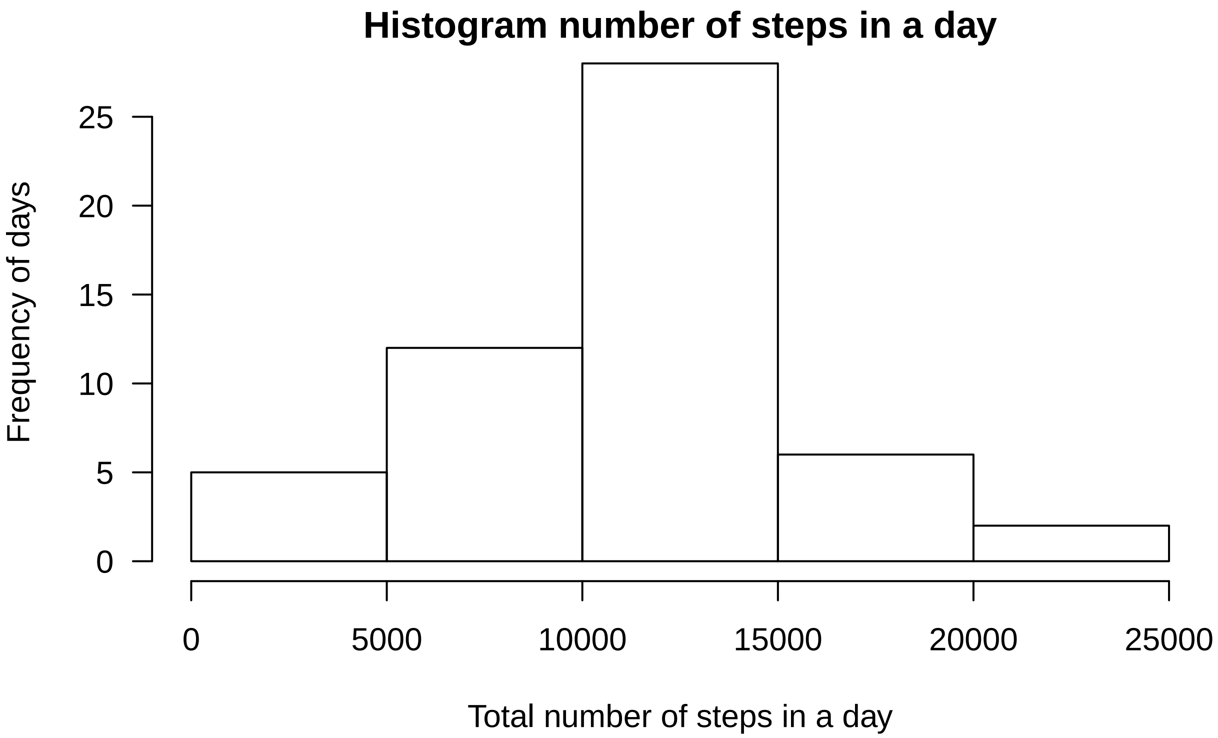
1. Calculate the total number of steps taken per day
2. If you do not understand the difference between a histogram and a barplot, research the difference between them. Make a histogram of the total number of steps taken each day
3. Calculate and report the mean and median of the total number of steps taken per day

Group data by day

```
steps_by_day <- group_by(data_activity,date) %>% summarise(steps_total=sum(steps))
```

Create Histogram

```
par(mar=c(5,4,1,1),las=1)
with(steps_by_day,hist(steps_total, xlab='Total number of steps in a day', ylab='Frequency of days', ma
```



Calculate mean and median

```
summary(steps_by_day)
```

```
##       date             steps_total
##  Min.   :2012-10-02   Min.   :   41
##  1st Qu.:2012-10-16   1st Qu.: 8841
##  Median :2012-10-29   Median :10765
##  Mean   :2012-10-30   Mean   :10766
##  3rd Qu.:2012-11-16   3rd Qu.:13294
##  Max.   :2012-11-29   Max.   :21194
```

Question 2

What is the average daily activity pattern?

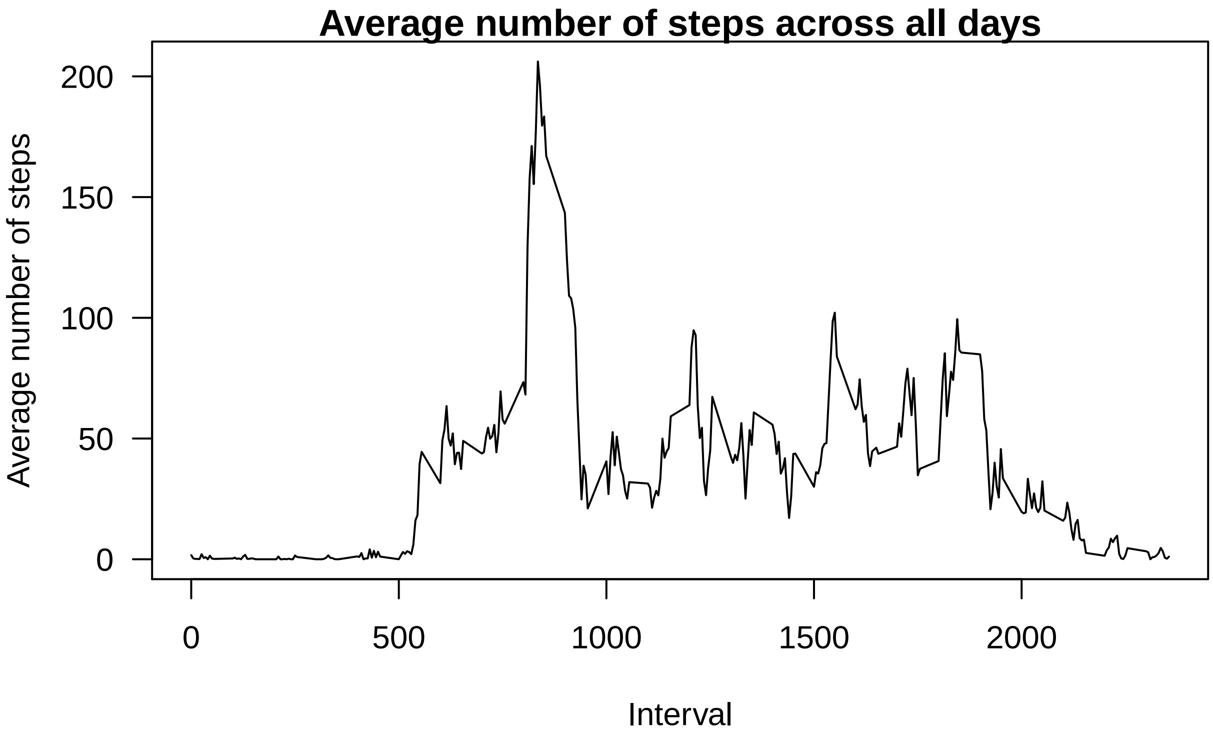
1. Make a time series plot (i.e. type = "l") of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis)
2. Which 5-minute interval, on average across all the days in the dataset, contains the maximum number of steps?

```
steps_by_interval <- group_by(data_activity,interval) %>% summarise(steps_mean=mean(steps))
head(steps_by_interval)
```

```
## # A tibble: 6 x 2
##   interval steps_mean
##      <int>      <dbl>
## 1        0     1.72
## 2        5     0.340
## 3       10     0.132
## 4       15     0.151
## 5       20     0.0755
## 6       25     2.09
```

Plot average number accross all steps per 5 minute intervals

```
par(mar=c(5,4,1,1),las=1)
with(steps_by_interval,plot(interval,steps_mean,type='l',xlab='Interval',ylab='Average number of steps'
```



Find 5-minute interval with maximum number of steps

```
max_steps_interval <- which.max(steps_by_interval$steps_mean)
result<-steps_by_interval[max_steps_interval,]
steps_average_max<-result$steps_mean
interval_max<-result$interval
```

The 5-minute interval 835, on average accross all the days in the dataset, has the maximum number of steps: 206.1698113

Question 3

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?

Define function to get the average of the interval to use for substitution of NA values

```
get_average_interval <- function(interval_input,interval_dataset){
      for (i in 1:nrow(interval_dataset)){
            if (interval_dataset$interval[i] == interval_input){
                  steps_average<-interval_dataset$steps_mean[i]
                  return(steps_average)
            }
      }
}
```

Read the data again

```
data_activity_nas<-read.table("./activity.csv",sep=",",header=TRUE)
head(data_activity_nas)
```

```
##   steps       date interval
## 1    NA 2012-10-01        0
## 2    NA 2012-10-01        5
## 3    NA 2012-10-01       10
## 4    NA 2012-10-01       15
## 5    NA 2012-10-01       20
## 6    NA 2012-10-01       25
```

```
str(data_activity_nas)
```

```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : Factor w/ 61 levels "2012-10-01","2012-10-02",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
```

Set date as date

```
data_activity_nas$date<-as.Date(data_activity_nas$date,'%Y-%m-%d')
```

Number of rows with NA

```
number_nas <- sum(is.na(data_activity_nas$steps))
```

The number of rows with NAs is 2304

Substitute NAs for the mean of the interval

```
for (i in 1:nrow(data_activity_nas)) {
      if (is.na(data_activity_nas$steps[i])){
            average_of_interval <-get_average_interval(data_activity_nas$interval[i],steps_by_interval)
            data_activity_nas$steps[i]<-average_of_interval
      }
}
```

Assign a new name

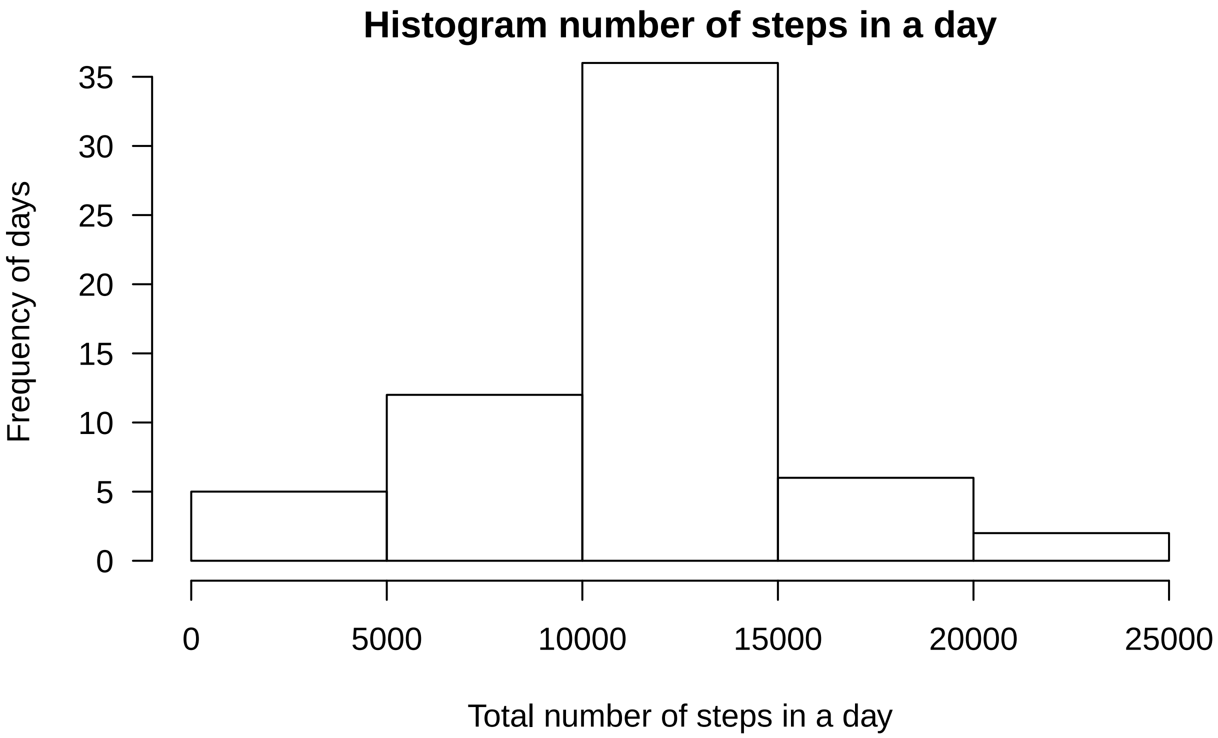
```
data_activity1 <- data_activity_nas
```

Group by date

```
steps_by_day1 <- group_by(data_activity1,date) %>% summarise(steps_total=sum(steps))
```

Create Histogram

```
par(mar=c(5,4,1,1),las=1)
with(steps_by_day1,hist(steps_total, xlab='Total number of steps in a day', ylab='Frequency of days', ma
```



Calculate mean and median

```
summary(steps_by_day1)
```

```
##       date             steps_total   
##  Min.   :2012-10-01   Min.   :   41  
##  1st Qu.:2012-10-16   1st Qu.: 9819  
##  Median :2012-10-31   Median :10766  
##  Mean   :2012-10-31   Mean   :10766  
##  3rd Qu.:2012-11-15   3rd Qu.:12811  
##  Max.   :2012-11-30   Max.   :21194
```

Do these values differ from the estimates from the first part of the assignment?

```
summary(steps_by_day)
```

```
##       date             steps_total   
##  Min.   :2012-10-02   Min.   :   41  
##  1st Qu.:2012-10-16   1st Qu.: 8841  
##  Median :2012-10-29   Median :10765  
##  Mean   :2012-10-30   Mean   :10766
##  3rd Qu.:2012-11-16   3rd Qu.:13294
##  Max.   :2012-11-29   Max.   :21194
```

```
summary(steps_by_day1)
```

```
##       date             steps_total
##  Min.   :2012-10-01   Min.   :   41
##  1st Qu.:2012-10-16   1st Qu.: 9819
##  Median :2012-10-31   Median :10766
##  Mean   :2012-10-31   Mean   :10766
##  3rd Qu.:2012-11-15   3rd Qu.:12811
##  Max.   :2012-11-30   Max.   :21194
```

The values differ slightly. Using the 5-min interval averages to substitute the NAs does not alter the results.

Question 4

Are there differences in activity patterns between weekdays and weekends?

1. For this part the weekdays() function may be of some help here. Use the dataset with the filled-in missing values for this part.
2. 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.
3. 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.

```
library(ggplot2)
data_activity1$date<-as.Date(data_activity1$date,'%Y-%m-%d')

data_activity1$day_of_week<-weekdays(data_activity1$date)
data_activity1$typeday[data_activity1$day_of_week %in% c('Monday','Tuesday','Wednesday','Thursday','Fri
data_activity1$typeday[data_activity1$day_of_week %in% c('Saturday','Sunday')] <- 'weekend'
```

New Factor variable

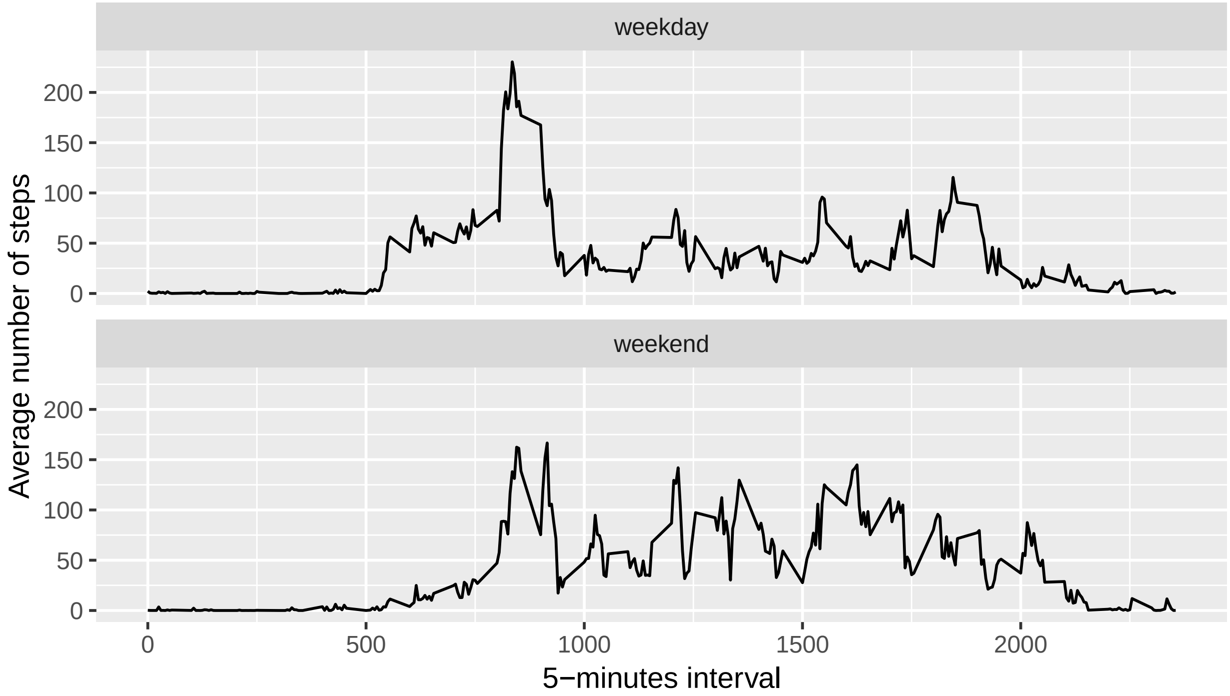
```
data_activity1$typeday<-as.factor(data_activity1$typeday)
```

Average steps by interval

```
data_activity1_interval <- group_by(data_activity1,interval,typeday) %>% summarise(steps_mean=mean(steps
```

Panel Plot containing a time series plot

```
qplot(interval,steps_mean, data = data_activity1_interval,
      geom=c("line"),
      xlab = "5-minutes interval",
      ylab = "Average number of steps",
      main = "") +
      facet_wrap(~ typeday, ncol = 1)
```



The activity differs between weekends and weekdays. During the weekend the activity starts later and people are more active.