Assignment

Loading and preprocessing the data

Unzip and load the file

```
unzip("activity.zip")
activity <- read.csv("activity.csv")
colnames(activity)
```

```
## [1] "steps"    "date"     "interval"
```

Check where the missing values are

```
summary(activity)
```

```
##      steps             date              interval     
##  Min.   :  0.00   Length:17568       Min.   :   0.0  
##  1st Qu.:  0.00   Class :character   1st Qu.: 588.8  
##  Median :  0.00   Mode  :character   Median :1177.5  
##  Mean   : 37.38                      Mean   :1177.5  
##  3rd Qu.: 12.00                      3rd Qu.:1766.2  
##  Max.   :806.00                      Max.   :2355.0  
##  NA's   :2304
```

Transform the date into date format.

```
activity$date <- as.Date(activity$date, format="%Y-%m-%d")
class(activity$date)
```

```
## [1] "Date"
```

What is mean total number of steps taken per day?

Load libraries to use

```
library("ggplot2")
library("magrittr")
library("dplyr")
```

```
## 
## 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
```

Total steps taken per day

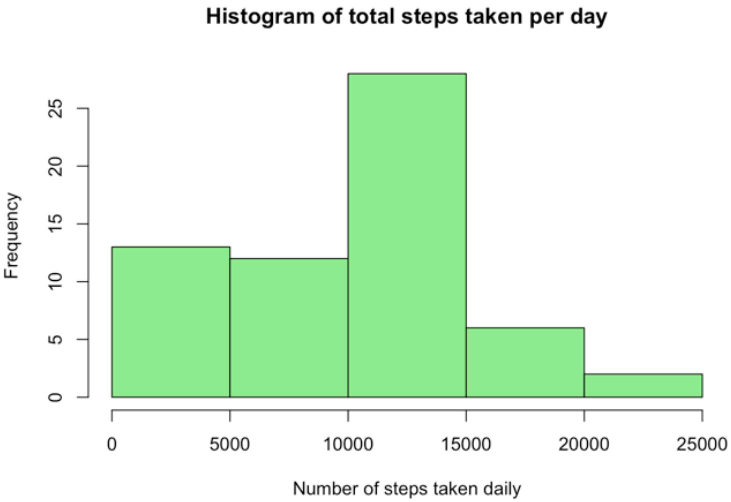
```
sum_ac <- data.frame(total.steps=with(activity, tapply(steps,date,sum,na.rm=TRUE)))
sum_ac
```

```
##            total.steps
## 2012-10-01           0
## 2012-10-02         126
## 2012-10-03       11352
## 2012-10-04       12116
## 2012-10-05       13294
## 2012-10-06       15420
## 2012-10-07       11015
## 2012-10-08           0
## 2012-10-09       12811
## 2012-10-10        9900
## 2012-10-11       10304
## 2012-10-12       17382
## 2012-10-13       12426
## 2012-10-14       15098
## 2012-10-15       10139
## 2012-10-16       15084
## 2012-10-17       13452
## 2012-10-18       10056
## 2012-10-19       11829
## 2012-10-20       10395
## 2012-10-21        8821
## 2012-10-22       13460
## 2012-10-23        8918
## 2012-10-24        8355
## 2012-10-25        2492
## 2012-10-26        6778
## 2012-10-27       10119
## 2012-10-28       11458
## 2012-10-29        5018
## 2012-10-30        9819
## 2012-10-31       15414
## 2012-11-01           0
## 2012-11-02       10600
## 2012-11-03       10571
## 2012-11-04           0
## 2012-11-05       10439
## 2012-11-06        8334
## 2012-11-07       12883
## 2012-11-08        3219
## 2012-11-09           0
## 2012-11-10           0
## 2012-11-11       12608
## 2012-11-12       10765
## 2012-11-13        7336
## 2012-11-14           0
```

```
## 2012-11-15          41
## 2012-11-16        5441
## 2012-11-17       14339
## 2012-11-18       15110
## 2012-11-19        8841
## 2012-11-20        4472
## 2012-11-21       12787
## 2012-11-22       20427
## 2012-11-23       21194
## 2012-11-24       14478
## 2012-11-25       11834
## 2012-11-26       11162
## 2012-11-27       13646
## 2012-11-28       10183
## 2012-11-29        7047
## 2012-11-30           0
```

Histogram of the total steps taken each day

```
hist(sum_ac$total.steps, xlab="Number of steps taken daily",main="Histogram of total steps taken per day",col="lightgreen")
```



Mean and median of the steps taken daily

```
mean_ac <- data.frame(mean=with(activity, tapply(steps,date,mean,na.rm=TRUE)), median=with(activity, tapply(steps,date,median,na.rm=TRUE)))
mean_ac <- slice(mean_ac,-grep("NaN",mean_ac$mean))
mean_ac
```

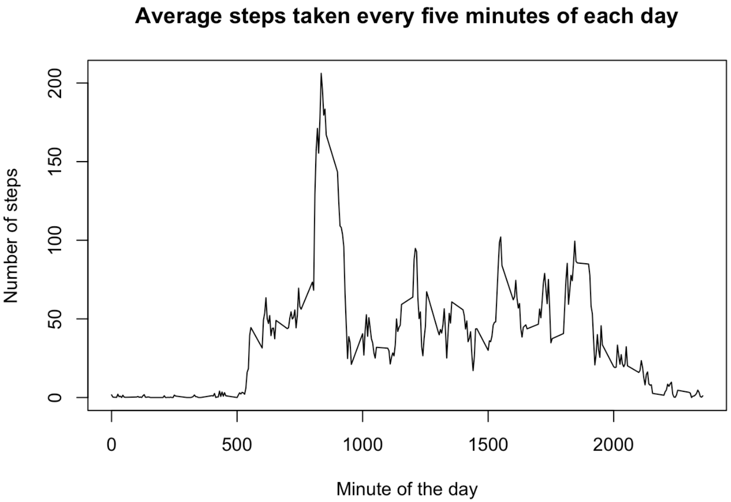
```
##                  mean median
## 2012-10-02  0.4375000      0
## 2012-10-03 39.4166667      0
## 2012-10-04 42.0694444      0
## 2012-10-05 46.1597222      0
## 2012-10-06 53.5416667      0
## 2012-10-07 38.2465278      0
## 2012-10-09 44.4826389      0
## 2012-10-10 34.3750000      0
## 2012-10-11 35.7777778      0
## 2012-10-12 60.3541667      0
## 2012-10-13 43.1458333      0
## 2012-10-14 52.4236111      0
## 2012-10-15 35.2048611      0
## 2012-10-16 52.3750000      0
## 2012-10-17 46.7083333      0
## 2012-10-18 34.9166667      0
## 2012-10-19 41.0729167      0
## 2012-10-20 36.0937500      0
## 2012-10-21 30.6284722      0
## 2012-10-22 46.7361111      0
## 2012-10-23 30.9652778      0
## 2012-10-24 29.0104167      0
## 2012-10-25  8.6527778      0
## 2012-10-26 23.5347222      0
## 2012-10-27 35.1354167      0
## 2012-10-28 39.7847222      0
## 2012-10-29 17.4236111      0
## 2012-10-30 34.0937500      0
## 2012-10-31 53.5208333      0
## 2012-11-02 36.8055556      0
## 2012-11-03 36.7048611      0
## 2012-11-05 36.2465278      0
## 2012-11-06 28.9375000      0
## 2012-11-07 44.7326389      0
## 2012-11-08 11.1770833      0
## 2012-11-11 43.7777778      0
## 2012-11-12 37.3784722      0
## 2012-11-13 25.4722222      0
## 2012-11-15  0.1423611      0
## 2012-11-16 18.8923611      0
## 2012-11-17 49.7881944      0
## 2012-11-18 52.4652778      0
## 2012-11-19 30.6979167      0
## 2012-11-20 15.5277778      0
## 2012-11-21 44.3993056      0
## 2012-11-22 70.9270833      0
## 2012-11-23 73.5902778      0
## 2012-11-24 50.2708333      0
## 2012-11-25 41.0902778      0
## 2012-11-26 38.7569444      0
## 2012-11-27 47.3819444      0
## 2012-11-28 35.3576389      0
## 2012-11-29 24.4687500      0
```

The median for all is 0 and the mean varies along the day.

What is the average daily activity pattern?

Average steps taken every 5-minutes of each day

```
five_min_steps <- data.frame(minute=unique(activity$interval),ave.steps=with(activity,
tapply(steps,interval,mean,na.rm=TRUE)),row.names=1:length(unique(activity$interval)))
with(five_min_steps,plot(minute,ave.steps,type="l",main="Average steps taken every five
minutes of each day",ylab="Number of steps",xlab="Minute of the day"))
```



```
max_interval <- grep(max(five_min_steps$ave.steps),five_min_steps$ave.steps)
```

The 5-minutes interval with the maximum averaged number of steps is the 104 interval of the day.

Imputing missing values

Total number of rows with NAs

```
sum(is.na(activity$steps))
```

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

Create data frame with NAs in "steps" filled with mean for each 5-minutes interval Find the dates in which there are NAs

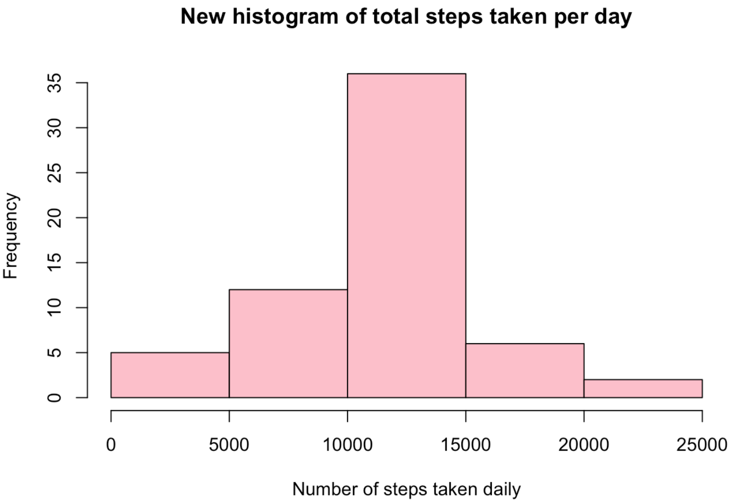
```
date_na <- rownames(sum_ac)[grep("^0",sum_ac$total.steps)] %>% as.Date()
```

Fill the 5-minutes intervals of these dates with the mean for each interval

```
for(i in 1:length(date_na)){
    x <- which(activity$date == date_na[i])
    for(j in 1:length(x)){
      for(k in 1:nrow(five_min_steps)){
        if(j == k){
          activity[x[j],1] <- five_min_steps$ave.steps[k]
        }
      }
    }
}
```

Histogram of the total number of steps taken each day with the new filled-in data frame

```
sum_ac <- data.frame(total.steps=with(activity, tapply(steps,date,sum,na.rm=TRUE)))
hist(sum_ac$total.steps, xlab="Number of steps taken daily",main="New histogram of total steps taken per day",col="pink")
```



Imputing the missing data using the mean of the five minutes intervals, does impact the results, particularly for the frecuency of 0-5000 steps per day with the value below the half of the original data, and the frecuency of 10000-15000 steps that originally was around 25, being now of around 35.

Mean and median of steps taken every 5-minutes of each day, using the new filled-in data frame

```
ac <- data.frame(mean=with(activity, tapply(steps,date,mean)), median=with(activity, tapply(steps,date,median)))
```

```
summary(mean_ac)
```

```
##       mean              median 
##  Min.   : 0.1424   Min.   :0  
##  1st Qu.:30.6979   1st Qu.:0  
##  Median :37.3785   Median :0  
##  Mean   :37.3826   Mean   :0  
##  3rd Qu.:46.1597   3rd Qu.:0  
##  Max.   :73.5903   Max.   :0
```

```
summary(ac)
```

```
##       mean              median      
##  Min.   : 0.1424   Min.   : 0.000  
##  1st Qu.:34.0938   1st Qu.: 0.000  
##  Median :37.3826   Median : 0.000  
##  Mean   :37.3826   Mean   : 4.474  
##  3rd Qu.:44.4826   3rd Qu.: 0.000  
##  Max.   :73.5903   Max.   :34.113
```

Imputing the missing data using the mean of the five minutes intervals, did not impact much the results for the mean of total steps taken daily, just the first quantile is slightly higher now and the third quantile, is slightly lower. However, the median changed for the dates where the data was imputed and it remained 0 for the rest.

Are there differences in activity patterns between weekdays and weekends?

Establish which day of the week each date was

```
activity$weekdays <- weekdays(activity$date)
activity$weekdays <- ifelse(activity$weekdays %in% c("Saturday","Sunday"),sub(pattern="
(.*)",replacement="weekend",activity$weekdays),sub(pattern="(.*)",replacement="weekday"
,activity$weekdays)) %>% as.factor()
str(activity$weekdays)
```

```
##  Factor w/ 2 levels "weekday","weekend": 1 1 1 1 1 1 1 1 1 1 ...
```

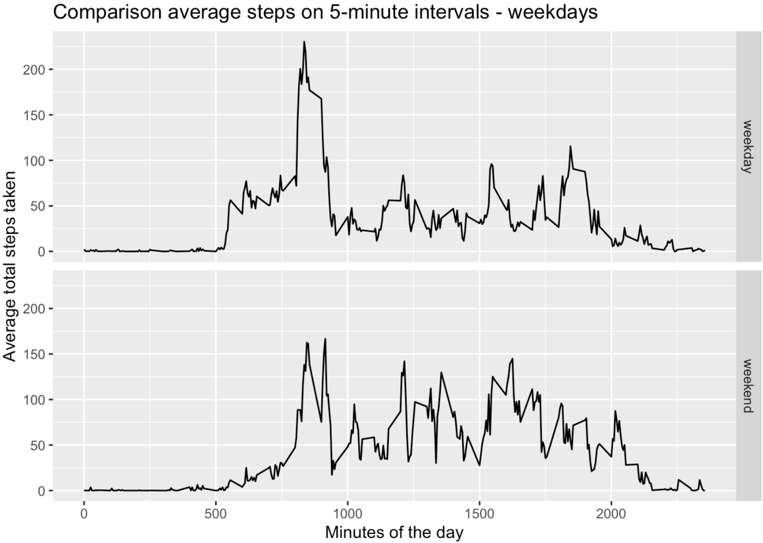
Construct the table of the 5-minute interval and the average number of steps taken, averaged across all weekday days or weekend days.

```
interval_weekend <- subset(activity,subset=activity$weekdays == "weekend")
interval_weekend <- data.frame(weekdays= "weekend",minute=unique(interval_weekend$inter
val),ave.steps=with(interval_weekend, tapply(steps,interval,mean,na.rm=TRUE)),row.names
=1:length(unique(interval_weekend$interval)))

interval_weekday <- subset(activity,subset=activity$weekdays == "weekday")
interval_weekday <- data.frame(weekdays="weekday",minute=unique(interval_weekday$interv
al),ave.steps=with(interval_weekday, tapply(steps,interval,mean,na.rm=TRUE)),row.names=
1:length(unique(interval_weekday$interval))) %>% rbind(.,interval_weekend)
```

Plot the resulting table

```
qplot(minute,ave.steps,data=interval_weekday,
      geom="line",
      main="Comparison average steps on 5-minute intervals - weekdays",
      xlab="Minutes of the day",
      ylab="Average total steps taken",
      facets=weekdays~.)
```



Comparison average steps on 5-minute intervals - weekdays