# Reproducible Research: Peer Assessment 1

In this file, I present the answers for the peer assessment 1 from the *Reproducible Research* course @ Coursera.

## Loading required libraries

The first step consists on charge the required libraries to process the data.

* **ggplot2**: A plotting system for R, based on the grammar of graphics, which tries to take the good parts of base and lattice graphics and none of the bad.
* **knitr**: A general-purpose tool for dynamic report generation in R (Literate Statistical Programming).
* **impute**: Imputation for microarray data (currently KNN only)
* **markdown**: R Markdown is an authoring format that enables easy creation of dynamic documents, presentations, and reports from R.
* **xtable**: Coerce data to LaTeX and HTML tables.

#install.packages("ggplot2")  
library(ggplot2)  
library(knitr)  
library(impute)  
library(markdown)  
library(xtable)

## Loading and preprocessing the data

In order to perform the analysis, I loaded all the variables included in the dataset.

* **steps**: Number of steps taking in a 5-minute interval (missing values are coded as NA)
* **date**: The date on which the measurement was taken in YYYY-MM-DD format
* **interval**: Identifier for the 5-minute interval in which measurement was taken

The dataset is stored in a comma-separated-value (CSV) file and there are a total of 17,568 observations in this dataset.

The following code chunk loads the data in R.

acdat <- read.csv(unz("activity.zip", "activity.csv"),  
 header = TRUE,  
 colClasses = c("numeric","POSIXct","numeric"))

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

In this section, an histogram present the distribution of the number of steps taken across all days.

stepsperday <- with(acdat,tapply(steps,date,sum,na.rm=T))  
binw <- (range(stepsperday,na.rm = T)[2]-range(stepsperday,na.rm = T)[1])/30  
qplot(stepsperday, binwidth =binw, xlab= "Steps per day", ylab = "Frequency")

plot of chunk steps per day

It is important to notice that there exists a high frecuency of zeros on the histogram due to the lack of data (missing values) in some days. Those missing values will be imputed and the new histogram is presented in two sections below.

In order to provide a more detail view of the data, the following table present the mean and median number of steps taken each day (empty boxes indicates missing data):

step\_mean <- with(acdat,tapply(steps,date,mean,na.rm=T))  
step\_median <- with(acdat,tapply(steps,date,median,na.rm=T))  
mm <- data.frame(step\_mean, step\_median)  
xt <- xtable(mm)  
print(xt, type= "html")

step\_mean

step\_median

2012-10-01

2012-10-02

0.44

0.00

2012-10-03

39.42

0.00

2012-10-04

42.07

0.00

2012-10-05

46.16

0.00

2012-10-06

53.54

0.00

2012-10-07

38.25

0.00

2012-10-08

2012-10-09

44.48

0.00

2012-10-10

34.38

0.00

2012-10-11

35.78

0.00

2012-10-12

60.35

0.00

2012-10-13

43.15

0.00

2012-10-14

52.42

0.00

2012-10-15

35.20

0.00

2012-10-16

52.38

0.00

2012-10-17

46.71

0.00

2012-10-18

34.92

0.00

2012-10-19

41.07

0.00

2012-10-20

36.09

0.00

2012-10-21

30.63

0.00

2012-10-22

46.74

0.00

2012-10-23

30.97

0.00

2012-10-24

29.01

0.00

2012-10-25

8.65

0.00

2012-10-26

23.53

0.00

2012-10-27

35.14

0.00

2012-10-28

39.78

0.00

2012-10-29

17.42

0.00

2012-10-30

34.09

0.00

2012-10-31

53.52

0.00

2012-11-01

2012-11-02

36.81

0.00

2012-11-03

36.70

0.00

2012-11-04

2012-11-05

36.25

0.00

2012-11-06

28.94

0.00

2012-11-07

44.73

0.00

2012-11-08

11.18

0.00

2012-11-09

2012-11-10

2012-11-11

43.78

0.00

2012-11-12

37.38

0.00

2012-11-13

25.47

0.00

2012-11-14

2012-11-15

0.14

0.00

2012-11-16

18.89

0.00

2012-11-17

49.79

0.00

2012-11-18

52.47

0.00

2012-11-19

30.70

0.00

2012-11-20

15.53

0.00

2012-11-21

44.40

0.00

2012-11-22

70.93

0.00

2012-11-23

73.59

0.00

2012-11-24

50.27

0.00

2012-11-25

41.09

0.00

2012-11-26

38.76

0.00

2012-11-27

47.38

0.00

2012-11-28

35.36

0.00

2012-11-29

24.47

0.00

2012-11-30

means <- mean(stepsperday)  
medians <- median(stepsperday)

Finally, I compute the mean of the daily steps, 9354.23 and its median, 10395.00.

## What is the average daily activity pattern?

In this section, the reader can observe the average dail activity pattern for the studied subject.

acdat$datetime <- acdat$date+acdat$interval\*60  
interAvg <- with(acdat, tapply(steps, interval,FUN = mean, na.rm =T))   
interdf <- data.frame(interval = as.numeric(names(interAvg)),interAvg)  
ggplot(data = interdf, aes(interval, interAvg)) + geom\_line(color="steelblue",size=1)+ ylab("Number of steps") + xlab("Date/time")

plot of chunk time series plot

maxd <- which.max(interdf$interAvg)  
hour <- floor(interdf$interval[maxd]/60)  
minute <- (interdf$interval[maxd]/60 - floor(interdf$interval[maxd]/60))\*60

The plot shows that the 5-minute interval 835 of the day (i.e. 13:55) contains the maximum number of steps (206.17) on average.

## Imputing missing values

As the reader may have seen in previous sections, the orginal data contains some missing values: 2304.

In order to perform an exploratory analysis, the missing values were imputed using the impute.knn() function from the impute package. This function imputes the missing valules using nearest neighbor averaging.

The R code used is he following.

acdat2 <- acdat   
acdat2$steps<- (impute.knn(  
 as.matrix(cbind(acdat2$steps,  
 acdat2$interval,  
 as.numeric(acdat2$date))))$data)[,1]

The following histogram exhibits the data distribution with the imputed data. It worth to notice that this graph doesn't has an excess of zeros produced by the missing values.

stepsperday2 <- with(acdat2,tapply(steps,date,sum,na.rm=T))  
binw2 <- (range(stepsperday2,na.rm = T)[2]-range(stepsperday2,na.rm = T)[1])/30  
qplot(stepsperday2, binwidth =binw2, xlab= "Steps per day", ylab = "Frequency")

plot of chunk imputation plot

The following table presents the steps mean and median per day with the imputed data.

step\_mean <- with(acdat2,tapply(steps,date,mean,na.rm=T))  
step\_median <- with(acdat2,tapply(steps,date,median,na.rm=T))  
mm <- data.frame(step\_mean, step\_median)  
xt <- xtable(mm)  
print(xt, type= "html")

step\_mean

step\_median

2012-10-01

37.38

37.38

2012-10-02

0.44

0.00

2012-10-03

39.42

0.00

2012-10-04

42.07

0.00

2012-10-05

46.16

0.00

2012-10-06

53.54

0.00

2012-10-07

38.25

0.00

2012-10-08

37.38

37.38

2012-10-09

44.48

0.00

2012-10-10

34.38

0.00

2012-10-11

35.78

0.00

2012-10-12

60.35

0.00

2012-10-13

43.15

0.00

2012-10-14

52.42

0.00

2012-10-15

35.20

0.00

2012-10-16

52.38

0.00

2012-10-17

46.71

0.00

2012-10-18

34.92

0.00

2012-10-19

41.07

0.00

2012-10-20

36.09

0.00

2012-10-21

30.63

0.00

2012-10-22

46.74

0.00

2012-10-23

30.97

0.00

2012-10-24

29.01

0.00

2012-10-25

8.65

0.00

2012-10-26

23.53

0.00

2012-10-27

35.14

0.00

2012-10-28

39.78

0.00

2012-10-29

17.42

0.00

2012-10-30

34.09

0.00

2012-10-31

53.52

0.00

2012-11-01

37.38

37.38

2012-11-02

36.81

0.00

2012-11-03

36.70

0.00

2012-11-04

37.38

37.38

2012-11-05

36.25

0.00

2012-11-06

28.94

0.00

2012-11-07

44.73

0.00

2012-11-08

11.18

0.00

2012-11-09

37.38

37.38

2012-11-10

37.38

37.38

2012-11-11

43.78

0.00

2012-11-12

37.38

0.00

2012-11-13

25.47

0.00

2012-11-14

37.38

37.38

2012-11-15

0.14

0.00

2012-11-16

18.89

0.00

2012-11-17

49.79

0.00

2012-11-18

52.47

0.00

2012-11-19

30.70

0.00

2012-11-20

15.53

0.00

2012-11-21

44.40

0.00

2012-11-22

70.93

0.00

2012-11-23

73.59

0.00

2012-11-24

50.27

0.00

2012-11-25

41.09

0.00

2012-11-26

38.76

0.00

2012-11-27

47.38

0.00

2012-11-28

35.36

0.00

2012-11-29

24.47

0.00

2012-11-30

37.38

37.38

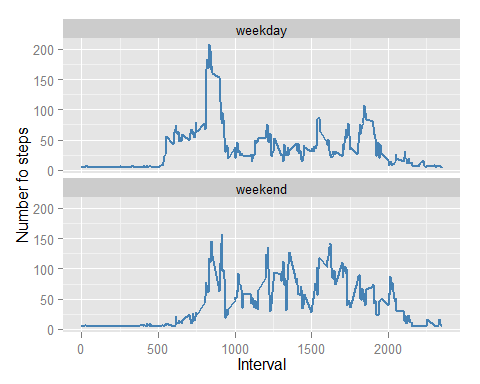
means2 <- mean(stepsperday2)  
medians2 <- median(stepsperday2)

Therefore, we can compute a more precise steps' mean and median across the days: 10766.19 and 10766.19, respectively.

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

The activity patters differ between weekends and midweek. This can be observe in the following plot.

weekend <- c(weekdays(as.Date(c("19072014","20072014"),format="%d%m%Y")))  
acdat2$weekday <- ifelse(weekdays(acdat2$date) %in% weekend, "weekend", "weekday")  
acdat2$weekday <- as.factor(acdat2$weekday)  
interAvg2 <- aggregate(acdat2$steps,by = list(acdat2$weekday,acdat2$interval),mean,na.rm=T)  
ggplot(interAvg2, aes(Group.2,x)) + geom\_line(color = "steelblue", size =1) + facet\_wrap(~ Group.1,2) + xlab(label = "Interval") + ylab("Number fo steps")



In the weekday plot, the reader can observe a peak of activity on the morning but after certain hour that activity become more stable until the end of the day. On the other hand, the weekend plot shows how the activity is stable along the day without a marked peak.