Reproducible Research: Peer Assessment 1

## Loading R Packages

library(plyr)  
library(chron)  
require(ggplot2)

## Loading required package: ggplot2

require(scales)

## Loading required package: scales

library(ggplot2)

## Loading and preprocessing the data

*Assume the data file of activity.csv has been saved in a proper directory.*

d=read.csv(file="C:/Temp/repdata\_data\_activity/activity.csv",sep= ",", h=T)

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

#### 1.1. Calculate the total number of steps taken per day

#d  
#summary(d)  
#nrow(d)  
  
#d1 <- aggregate(d[, 1], list(d$date), sum)  
  
  
d\_1 <- ddply(d, ~date, summarise, total\_steps\_per\_day=sum (steps) , rm.na = TRUE )  
d\_1

## date total\_steps\_per\_day rm.na  
## 1 2012-10-01 NA TRUE  
## 2 2012-10-02 126 TRUE  
## 3 2012-10-03 11352 TRUE  
## 4 2012-10-04 12116 TRUE  
## 5 2012-10-05 13294 TRUE  
## 6 2012-10-06 15420 TRUE  
## 7 2012-10-07 11015 TRUE  
## 8 2012-10-08 NA TRUE  
## 9 2012-10-09 12811 TRUE  
## 10 2012-10-10 9900 TRUE  
## 11 2012-10-11 10304 TRUE  
## 12 2012-10-12 17382 TRUE  
## 13 2012-10-13 12426 TRUE  
## 14 2012-10-14 15098 TRUE  
## 15 2012-10-15 10139 TRUE  
## 16 2012-10-16 15084 TRUE  
## 17 2012-10-17 13452 TRUE  
## 18 2012-10-18 10056 TRUE  
## 19 2012-10-19 11829 TRUE  
## 20 2012-10-20 10395 TRUE  
## 21 2012-10-21 8821 TRUE  
## 22 2012-10-22 13460 TRUE  
## 23 2012-10-23 8918 TRUE  
## 24 2012-10-24 8355 TRUE  
## 25 2012-10-25 2492 TRUE  
## 26 2012-10-26 6778 TRUE  
## 27 2012-10-27 10119 TRUE  
## 28 2012-10-28 11458 TRUE  
## 29 2012-10-29 5018 TRUE  
## 30 2012-10-30 9819 TRUE  
## 31 2012-10-31 15414 TRUE  
## 32 2012-11-01 NA TRUE  
## 33 2012-11-02 10600 TRUE  
## 34 2012-11-03 10571 TRUE  
## 35 2012-11-04 NA TRUE  
## 36 2012-11-05 10439 TRUE  
## 37 2012-11-06 8334 TRUE  
## 38 2012-11-07 12883 TRUE  
## 39 2012-11-08 3219 TRUE  
## 40 2012-11-09 NA TRUE  
## 41 2012-11-10 NA TRUE  
## 42 2012-11-11 12608 TRUE  
## 43 2012-11-12 10765 TRUE  
## 44 2012-11-13 7336 TRUE  
## 45 2012-11-14 NA TRUE  
## 46 2012-11-15 41 TRUE  
## 47 2012-11-16 5441 TRUE  
## 48 2012-11-17 14339 TRUE  
## 49 2012-11-18 15110 TRUE  
## 50 2012-11-19 8841 TRUE  
## 51 2012-11-20 4472 TRUE  
## 52 2012-11-21 12787 TRUE  
## 53 2012-11-22 20427 TRUE  
## 54 2012-11-23 21194 TRUE  
## 55 2012-11-24 14478 TRUE  
## 56 2012-11-25 11834 TRUE  
## 57 2012-11-26 11162 TRUE  
## 58 2012-11-27 13646 TRUE  
## 59 2012-11-28 10183 TRUE  
## 60 2012-11-29 7047 TRUE  
## 61 2012-11-30 NA TRUE

#### 1.2. Histogram of the total number of steps taken each day

jpeg("C:/Temp/repdata\_data\_activity/figures/histogram\_steps\_per\_day.jpeg");  
ggplot(d\_1, aes(x = total\_steps\_per\_day)) +  
 geom\_histogram(col="black", fill = "dark blue", binwidth = 1000) +  
 labs(title = "Histogram of steps per day", x = "Steps per day", y = "Frequency")

## Warning: Removed 8 rows containing non-finite values (stat\_bin).

dev.off()

## png   
## 2

#### 1.3 The mean and median of the total number of steps taken per day

d\_2 <- ddply(d, ~date, summarise, mean\_steps=mean(steps), median\_steps= median(steps) , rm.na = TRUE )  
  
d\_2

## date mean\_steps median\_steps rm.na  
## 1 2012-10-01 NA NA TRUE  
## 2 2012-10-02 0.4375000 0 TRUE  
## 3 2012-10-03 39.4166667 0 TRUE  
## 4 2012-10-04 42.0694444 0 TRUE  
## 5 2012-10-05 46.1597222 0 TRUE  
## 6 2012-10-06 53.5416667 0 TRUE  
## 7 2012-10-07 38.2465278 0 TRUE  
## 8 2012-10-08 NA NA TRUE  
## 9 2012-10-09 44.4826389 0 TRUE  
## 10 2012-10-10 34.3750000 0 TRUE  
## 11 2012-10-11 35.7777778 0 TRUE  
## 12 2012-10-12 60.3541667 0 TRUE  
## 13 2012-10-13 43.1458333 0 TRUE  
## 14 2012-10-14 52.4236111 0 TRUE  
## 15 2012-10-15 35.2048611 0 TRUE  
## 16 2012-10-16 52.3750000 0 TRUE  
## 17 2012-10-17 46.7083333 0 TRUE  
## 18 2012-10-18 34.9166667 0 TRUE  
## 19 2012-10-19 41.0729167 0 TRUE  
## 20 2012-10-20 36.0937500 0 TRUE  
## 21 2012-10-21 30.6284722 0 TRUE  
## 22 2012-10-22 46.7361111 0 TRUE  
## 23 2012-10-23 30.9652778 0 TRUE  
## 24 2012-10-24 29.0104167 0 TRUE  
## 25 2012-10-25 8.6527778 0 TRUE  
## 26 2012-10-26 23.5347222 0 TRUE  
## 27 2012-10-27 35.1354167 0 TRUE  
## 28 2012-10-28 39.7847222 0 TRUE  
## 29 2012-10-29 17.4236111 0 TRUE  
## 30 2012-10-30 34.0937500 0 TRUE  
## 31 2012-10-31 53.5208333 0 TRUE  
## 32 2012-11-01 NA NA TRUE  
## 33 2012-11-02 36.8055556 0 TRUE  
## 34 2012-11-03 36.7048611 0 TRUE  
## 35 2012-11-04 NA NA TRUE  
## 36 2012-11-05 36.2465278 0 TRUE  
## 37 2012-11-06 28.9375000 0 TRUE  
## 38 2012-11-07 44.7326389 0 TRUE  
## 39 2012-11-08 11.1770833 0 TRUE  
## 40 2012-11-09 NA NA TRUE  
## 41 2012-11-10 NA NA TRUE  
## 42 2012-11-11 43.7777778 0 TRUE  
## 43 2012-11-12 37.3784722 0 TRUE  
## 44 2012-11-13 25.4722222 0 TRUE  
## 45 2012-11-14 NA NA TRUE  
## 46 2012-11-15 0.1423611 0 TRUE  
## 47 2012-11-16 18.8923611 0 TRUE  
## 48 2012-11-17 49.7881944 0 TRUE  
## 49 2012-11-18 52.4652778 0 TRUE  
## 50 2012-11-19 30.6979167 0 TRUE  
## 51 2012-11-20 15.5277778 0 TRUE  
## 52 2012-11-21 44.3993056 0 TRUE  
## 53 2012-11-22 70.9270833 0 TRUE  
## 54 2012-11-23 73.5902778 0 TRUE  
## 55 2012-11-24 50.2708333 0 TRUE  
## 56 2012-11-25 41.0902778 0 TRUE  
## 57 2012-11-26 38.7569444 0 TRUE  
## 58 2012-11-27 47.3819444 0 TRUE  
## 59 2012-11-28 35.3576389 0 TRUE  
## 60 2012-11-29 24.4687500 0 TRUE  
## 61 2012-11-30 NA NA TRUE

#mean (d1$x, na.rm = TRUE )  
#median (d1$x, na.rm = TRUE )

## What is the average daily activity pattern?

#### 2.1. 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)

jpeg("C:/Temp/repdata\_data\_activity/figures/Time series plot.jpeg");  
ggplot(data=d\_3, aes(x= interval, y=mean\_steps, group =1)) + geom\_line(color= "dark blue")  
  
dev.off()

## png   
## 2

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

d\_3$mx <- max(d\_3$mean\_steps, na.rm= true)  
  
newdata <- unique( subset (d\_3, d\_3$mean\_steps == d\_3$mx) )  
newdata

## interval mean\_steps rm.na mx  
## 104 835 206.1698 TRUE 206.1698

At the inverval of **835** contains the maximum number of steps.

## Imputing missing values

#### 3.1. Calculate and report the total number of missing values in the dataset (i.e. the total number of rows with NAs)

n <- sum(is.na(d$steps) | is.na(d$date) | is.na(d$interval))  
n

## [1] 2304

There are 2304 missing values in the dataset

#### 3.2. filling in all of the missing values in the dataset, replace missing value with mean steps of the day

#### &

#### 3.3. Create a new dataset that is equal to the original dataset but with the missing data filled in.

#n1 <- sum(is.na(d$steps))  
#n1  
rplc<- merge (d, d\_2, by = "date")  
rplc$steps[is.na(rplc$steps)] <- rplc$mean\_steps

## Warning in rplc$steps[is.na(rplc$steps)] <- rplc$mean\_steps: number of  
## items to replace is not a multiple of replacement length

n2<-sum(is.na(rplc$steps))  
n2

## [1] 576

The missing data is replaced with the means steps of the day. The dataset with missing data filled in is created. Now there are only 576 missing data points, used to be 2304 missing data points.

#### 3.4. 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?

rplc1 <- ddply(rplc, ~date, summarise, total\_steps=sum(steps))  
  
as.numeric(rplc1$total\_steps)

## [1] NA 126 11352 12116 13294 15420 11015 126 12811 9900 10304  
## [12] 17382 12426 15098 10139 15084 13452 10056 11829 10395 8821 13460  
## [23] 8918 8355 2492 6778 10119 11458 5018 9819 15414 11352 10600  
## [34] 10571 12116 10439 8334 12883 3219 13294 15420 12608 10765 7336  
## [45] 11015 41 5441 14339 15110 8841 4472 12787 20427 21194 14478  
## [56] 11834 11162 13646 10183 7047 NA

jpeg("C:/Temp/repdata\_data\_activity/figures/Histogram\_w\_imputing\_missing.jpeg");  
ggplot(rplc1 , aes(x = total\_steps)) +  
 geom\_histogram(col="black", fill = "dark blue", binwidth = 1000) +  
 labs(title = "Histogram of steps per day", x = "Steps per day", y = "Frequency")

## Warning: Removed 2 rows containing non-finite values (stat\_bin).

dev.off()

## png   
## 2

The dataset filled with mean value now constains only 2 missing data points, while the original data contains 8.

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

#### 4.1 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.

#library(chron)  
  
rplc$wk = ifelse (chron::is.weekend(rplc$date), "weekend", "weekday")

#### 4.2. A panel plot containing a time series plot of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all weekday days or weekend days (y-axis).

The following graph shows the activity pattern for the weekdays and weekend. Comparing with the weekend, people tend to be more active in the early time of a day; while people have about the same level of activity in the later time (late afternoon and evening) as in the morning of a day.

jpeg("C:/Temp/repdata\_data\_activity/figures/Time\_series\_Weekday\_weekend.jpeg");  
  
ggplot(rplc\_nona, aes(x= interval,y=mean\_s\_w\_I, color = wk, group=1)) +geom\_line() +  
 facet\_grid(factor(wk) ~ .) +  
 xlab("Interval") +  
 ylab( "Steps") +  
 theme(axis.title.y = element\_text(face='bold',size=14,color='black', vjust=1),  
 axis.text.x = element\_text(face='bold',size=14,color='black'),  
 legend.title=element\_blank())  
  
dev.off()

## png   
## 2