Title : Peer Assessment1 R Markdown file

r1

Saturday, March 14, 2015

# This is an R Markdown document. For the Peer Assessment 1

# This document has both code and results

library(knitr)

## Warning: package 'knitr' was built under R version 3.1.3

library(data.table)

## Warning: package 'data.table' was built under R version 3.1.2

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.1.2

library(plyr)

## Warning: package 'plyr' was built under R version 3.1.2

opts\_chunk$set(echo = TRUE, results = 'hold')

# Loading and preprocessing the data

# cleaning the data

## Demostrating the sample of the data

activitydata <- read.csv('activity.csv', header = TRUE, sep = ",",  
 colClasses=c("numeric", "character", "numeric"))  
activitydata$date <- as.Date(activitydata$date, format = "%Y-%m-%d")  
activitydata$interval <- as.factor(activitydata$interval)  
head(activitydata)

## 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

# The mean total number of steps taken per day

## 1. the total number of steps taken per day without NA

### We format the data into a string and produce a mean to show the top few rows

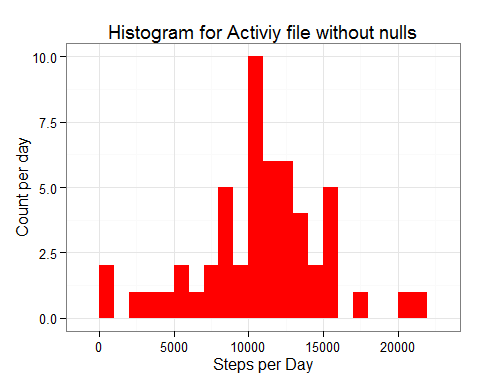
str(activitydata)  
steps\_per\_day <- aggregate(steps ~ date, activitydata, sum, na.rm=TRUE)  
colnames(steps\_per\_day) <- c("date","steps")  
head(steps\_per\_day)

## 'data.frame': 17568 obs. of 3 variables:  
## $ steps : num NA NA NA NA NA NA NA NA NA NA ...  
## $ date : Date, format: "2012-10-01" "2012-10-01" ...  
## $ interval: Factor w/ 288 levels "0","5","10","15",..: 1 2 3 4 5 6 7 8 9 10 ...  
## date steps  
## 1 2012-10-02 126  
## 2 2012-10-03 11352  
## 3 2012-10-04 12116  
## 4 2012-10-05 13294  
## 5 2012-10-06 15420  
## 6 2012-10-07 11015

## 2. histogram of the total number of steps taken each day

### We are producing a ggplot for the activity file

ggplot(steps\_per\_day, aes(x = steps)) +   
 geom\_histogram(fill = "red", binwidth = 1000) + theme\_bw() +  
 labs(title="Histogram for Activiy file without nulls",   
 x = "Steps per Day", y = "Count per day")



## 3. Calculate and report the mean and median of the total number of steps taken per day

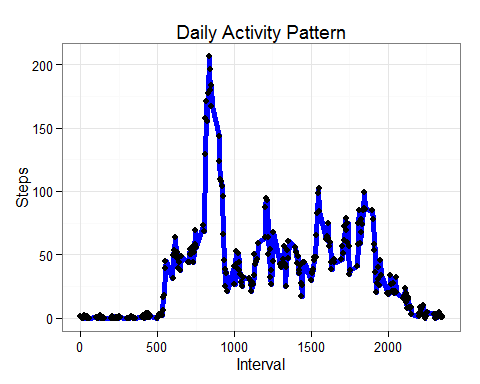
mean\_of\_steps <- mean(steps\_per\_day$steps, na.rm=TRUE)  
median\_of\_steps <- median(steps\_per\_day$steps, na.rm=TRUE)  
  
mean\_of\_steps  
median\_of\_steps

## [1] 10766.19  
## [1] 10765

# the average daily activity pattern

## 1. Making a time series line chart plot using geom\_line of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis)

steps\_interval <- aggregate(activitydata$steps, by = list(interval = activitydata$interval), FUN=mean, na.rm=TRUE)  
steps\_interval$interval <-   
 as.integer(levels(steps\_interval$interval)[steps\_interval$interval])  
colnames(steps\_interval) <- c("interval", "steps")  
ggplot(steps\_interval, aes(x=interval, y=steps)) + theme\_bw() +   
 geom\_line(color="blue", size=2) + geom\_jitter() +  
 labs(title="Daily Activity Pattern", x="Interval", y="Steps")



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

steps\_int <-steps\_interval[which.max(   
 steps\_interval$steps),]

The step with the maximum step intervals is 835, 206.1698113

# Imputing missing values

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

missing\_steps <- sum(is.na(activitydata$steps))

The total missing steps are 2304

## 2. Devising a strategy for filling in all of the missing values in the dataset.

na\_fill <- function(data, pervalue) {  
 na\_index <- which(is.na(data$steps))  
 na\_replace <- unlist(lapply(na\_index, FUN=function(idx){  
 interval = data[idx,]$interval  
 pervalue[pervalue$interval == interval,]$steps  
 }))  
 fill\_steps <- data$steps  
 fill\_steps[na\_index] <- na\_replace  
 fill\_steps  
}  
  
activity\_fill <- data.frame(   
 steps = na\_fill(activitydata, steps\_interval),   
 date = activitydata$date,   
 interval = activitydata$interval)  
str(activity\_fill)

## 'data.frame': 17568 obs. of 3 variables:  
## $ steps : num 1.717 0.3396 0.1321 0.1509 0.0755 ...  
## $ date : Date, format: "2012-10-01" "2012-10-01" ...  
## $ interval: Factor w/ 288 levels "0","5","10","15",..: 1 2 3 4 5 6 7 8 9 10 ...

The total missing steps are 0

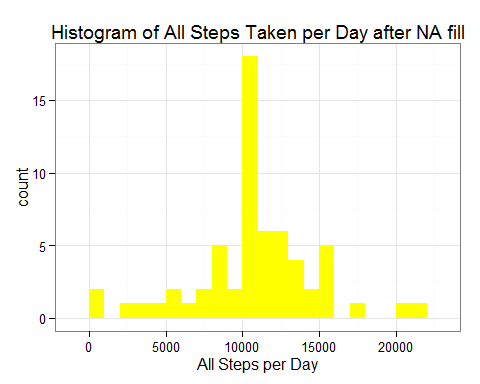
## 3.Creating a new dataset that is equal to the original dataset but with the missing data filled in.

all\_steps\_per\_day <- aggregate(steps ~ date, activity\_fill, sum)  
colnames(all\_steps\_per\_day) <- c("date","steps")  
  
head(all\_steps\_per\_day)

## date steps  
## 1 2012-10-01 10766.19  
## 2 2012-10-02 126.00  
## 3 2012-10-03 11352.00  
## 4 2012-10-04 12116.00  
## 5 2012-10-05 13294.00  
## 6 2012-10-06 15420.00

## 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

ggplot(all\_steps\_per\_day, aes(x = steps)) +   
 geom\_histogram(fill = "yellow", binwidth = 1000) + theme\_bw() +  
 labs(title="Histogram of All Steps Taken per Day after NA fill",   
 x = "All Steps per Day", y = "count")

 ## Missing value analysis:

All\_mean\_steps <- mean(all\_steps\_per\_day$steps)  
All\_median\_steps <- median(all\_steps\_per\_day$steps)  
  
All\_mean\_steps  
All\_median\_steps

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

There is a very small differnce between mean\_of\_steps and All\_mean\_step. There is a very small differnce between median\_of\_steps and All\_median\_steps.