Reproducible Research: Peer Assessment 1

Wei

February 14, 2019

## Loading libraries and data

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.5.2

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

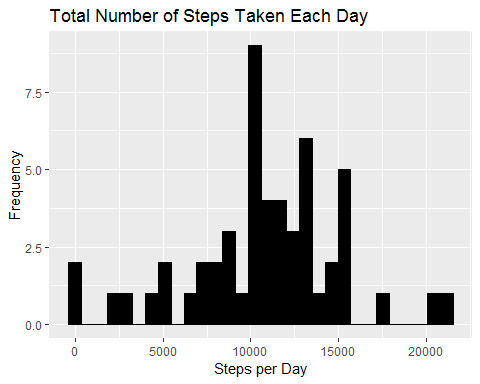
setwd("E:/Wei/coursera/certificate2 data science/course5/")  
 data <- read.csv("activity.csv")

## Q1: What is mean total number of steps taken per day?

total\_steps <- data %>% group\_by(date) %>% summarise(Steps = sum(steps))  
 ggplot(data = total\_steps, mapping = aes(x = Steps)) +   
 geom\_histogram(fill = "black", colour = "black") + scale\_x\_continuous("Steps per Day") +   
 scale\_y\_continuous("Frequency") + ggtitle("Total Number of Steps Taken Each Day")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

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



mean(total\_steps$Steps, na.rm=TRUE)

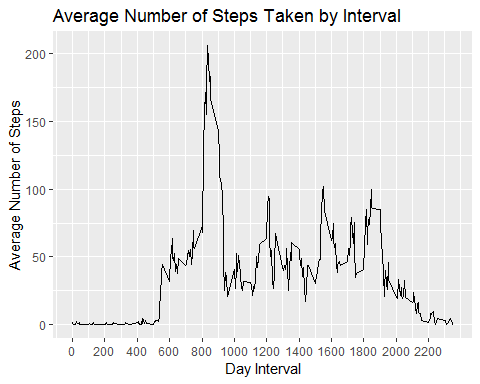
## [1] 10766.19

median(total\_steps$Steps, na.rm=TRUE)

## [1] 10765

## Q2: What is the average daily activity pattern?

actInterval <- data %>% group\_by(interval) %>% summarise(meanSteps = mean(steps, na.rm = TRUE))  
 ggplot(data = actInterval, mapping = aes(x = interval, y = meanSteps)) +   
 geom\_line() + scale\_x\_continuous("Day Interval", breaks = seq(min(actInterval$interval), max(actInterval$interval), 200)) +  
 scale\_y\_continuous("Average Number of Steps") + ggtitle("Average Number of Steps Taken by Interval")



m=which.max(actInterval$meanSteps)  
 actInterval$interval[m]

## [1] 835

Interval 835 contains the maximum number of step among the 5-minute intervals

## Q3: Imputing missing values

missing=subset(data,is.na(steps)==TRUE)  
 nrow(missing)

## [1] 2304

The dataset contains 2304 rows with missing values (coded as NA).

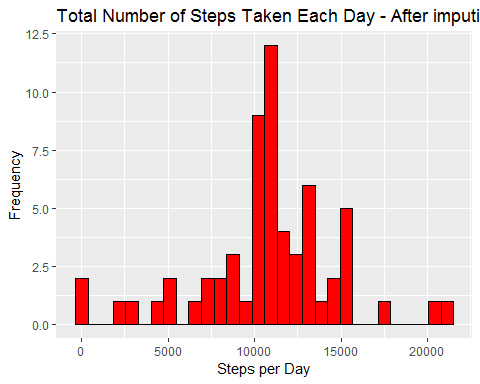
All missing values are filled with mean value for that 5-minute interval.

# First merge the original data with the mean by interval  
 actDat <- data %>% left\_join(actInterval, by = "interval")  
 # Next create a new column by replacing the missing data with the mean  
 actDat$fillSteps <- ifelse(is.na(actDat$steps), actDat$meanSteps, actDat$steps)  
 # Next drop the "steps" column and the "meanSteps" column, rename the "fillSteps" column to "steps"  
 data2 <- subset(actDat,select=c(fillSteps,date,interval))  
 colnames(data2) <- c("steps", "date", "interval")

Let’s make a histogram of the total steps taken each day and calculate the mean and median.

total\_steps2 <- data2 %>% group\_by(date) %>% summarise(Steps = sum(steps))  
 ggplot(data = total\_steps2, mapping = aes(x = Steps)) + geom\_histogram(fill = "red", colour = "black") +   
 scale\_x\_continuous("Steps per Day") + scale\_y\_continuous("Frequency") +   
 ggtitle("Total Number of Steps Taken Each Day - After imputing the Missing Values")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



mean(total\_steps2$Steps, na.rm=TRUE)

## [1] 10766.19

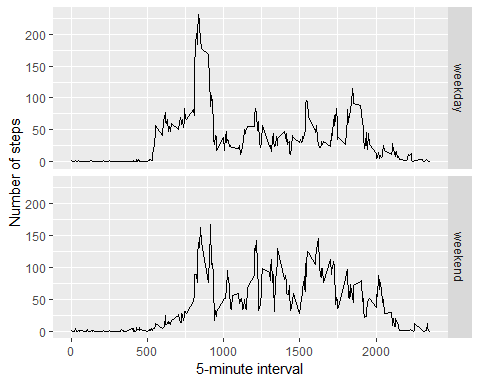
median(total\_steps2$Steps, na.rm=TRUE)

## [1] 10766.19

The mean and median are almost the same. Before imputation the mean is slightly higher than the median (10766.19 Vs 10765). The peak of the histogram shifts slightly to the left, and the histogram becomes more symmetric.

## Q4. Are there differences in activity patterns between weekdays and weekends?

weekday\_weekend <- function(date) {  
 day <- weekdays(date)  
 if (day %in% c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday"))  
 return("weekday")  
 else if (day %in% c("Saturday", "Sunday"))  
 return("weekend")  
 else  
 stop("invalid date")  
 }  
 data2$date <- as.Date(data2$date)  
 data2$day <- sapply(data2$date, FUN=weekday\_weekend)  
 averages <- aggregate(steps ~ interval + day, data=data2, mean)  
 ggplot(averages, aes(interval, steps)) + geom\_line() + facet\_grid(day ~ .) + xlab("5-minute interval") + ylab("Number of steps")



Yes, it seems there are two major differences between weekdays and weekends. (1) People tend to wake up later on weekends, so the steps don’t grow as fast as on weekdays for the early intervals. (2) The steps of the intervals after the first peak is bigger on weekends than weekdays. Because on weekends people can do a lot of acitivities such as workout, sports or shipping, but on weekdays people can spend more time in the office, so the steps are lower.