title: Peer assessment 1 Reproducible Research Week 2 author: Madelon den Boeft date: March 13, 2018 output: md\_document ==============================================

1. Load the data activity

activity\_dataset <- read.csv("activity.csv")

1. What is the mean total number of steps taken per day? ignore missing values?

Make a histogram of the total number of steps taken each day

total\_steps\_bydate<- aggregate(steps ~ date, activity\_dataset, sum)  
hist(total\_steps\_bydate$steps, main = paste("Total number of steps each day"), col="blue", xlab="Total number of steps", ylab= "Number of days")



Calculate and report the mean and median total number of steps taken each day

mean\_steps <- mean(total\_steps\_bydate$steps)  
median\_steps <- median(total\_steps\_bydate$steps)  
sprintf("Mean total number of steps taken per day = %.2f", mean\_steps)

## [1] "Mean total number of steps taken per day = 10766.19"

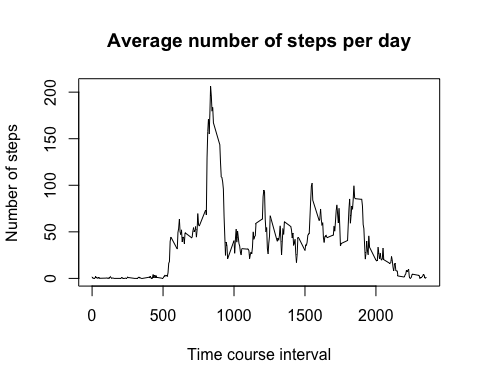
sprintf("Median total number of steps taken per day = %.2f", median\_steps)

## [1] "Median total number of steps taken per day = 10765.00"

1. What is the average daily activity pattern?

Make a time series plot of the 5-minute interval and average number of steps taken averaged across all days

steps\_by\_interval <- aggregate(steps ~ interval, activity\_dataset, mean)  
plot(steps\_by\_interval$interval, steps\_by\_interval$steps, type="l", xlab="Time course interval", ylab="Number of steps",main="Average number of steps per day")



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

max\_interval <- steps\_by\_interval[which.max(steps\_by\_interval$steps),1]  
sprintf("5-minute interval with maximum number of steps = %.2f", max\_interval)

## [1] "5-minute interval with maximum number of steps = 835.00"

1. Imputing missing values

Calculate and report the total number of missing values in the dataset

total\_missing\_data <- sum(!complete.cases(activity\_dataset))  
sprintf("Total number of missing values = %.2f", total\_missing\_data)

## [1] "Total number of missing values = 2304.00"

Devise a strategy for filling in all the missing values in the dataset Create a new dataset that is equal to the original one but with the missing data filled in

imputed\_data <- transform(activity\_dataset, steps = ifelse(is.na(activity\_dataset$steps), steps\_by\_interval$steps[match(activity\_dataset$interval, steps\_by\_interval$interval)], activity\_dataset$steps))

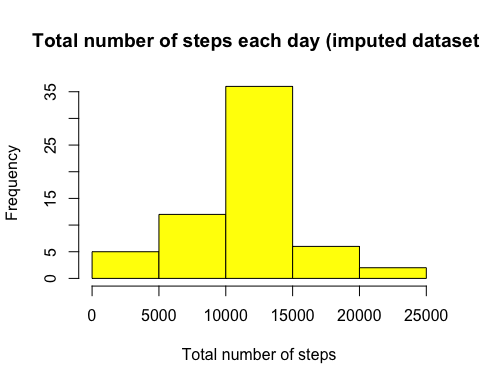
Test variable missing data to see if all the missings are gone / indeed imputed

test\_missing\_data <- sum(!complete.cases(imputed\_data))

Test result positive! Continue with the imputed\_data dataset…

Make a histogram of the total number of steps each day

total\_steps\_eachday\_imputed<- aggregate(steps ~ date, imputed\_data, sum)  
hist(total\_steps\_eachday\_imputed$steps, main = paste("Total number of steps each day (imputed dataset)"), col="yellow", xlab="Total number of steps")



Calculate and report the mean and median total number of steps each day

mean\_steps\_imputed <- mean(total\_steps\_eachday\_imputed$steps)  
median\_steps\_imputed <- median(total\_steps\_eachday\_imputed$steps)  
sprintf("Mean total number of steps taken per day (imputed)= %.2f", mean\_steps\_imputed)

## [1] "Mean total number of steps taken per day (imputed)= 10766.19"

sprintf("Median total number of steps taken per day (imputed) = %.2f", median\_steps\_imputed)

## [1] "Median total number of steps taken per day (imputed) = 10766.19"

Do these values differ?

sprintf("The difference in mean between the original and imputed dataset is %.2f ",mean\_steps\_imputed-mean\_steps)

## [1] "The difference in mean between the original and imputed dataset is 0.00 "

sprintf("The difference in median between the original and imputed dataset is %.2f ",median\_steps\_imputed-median\_steps)

## [1] "The difference in median between the original and imputed dataset is 1.19 "

What is the impact of imputing the missing values on the estimates of the total daily number of steps?

total <- sum(total\_steps\_bydate$steps)  
total\_imputed <- sum(total\_steps\_eachday\_imputed$steps)  
sprintf("The difference in total number of steps between original and imputed dataset = %.2f", total\_imputed-total)

## [1] "The difference in total number of steps between original and imputed dataset = 86129.51"

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

Create a factor variable with two levels “weekday” and “weekend”

weekend\_days <- c("zaterdag", "zondag")  
week\_days <- c("maandag", "dinsdag", "woensdag", "donderdag", "vrijdag")  
  
imputed\_data$dow = as.factor(ifelse(is.element(weekdays(as.Date(imputed\_data$date)),weekend\_days), "Weekend", "Weekday"))  
imputed\_steps\_by\_interval <- aggregate(steps ~ interval + dow, imputed\_data, mean)

Make a panel plot containinga time series plot of the 5-minute interval and the average number of steps taken, averaged across all weekday days or weekend days

library(lattice)  
xyplot(imputed\_steps\_by\_interval$steps ~ imputed\_steps\_by\_interval$interval|imputed\_steps\_by\_interval$dow, main="Average Steps per Day by Interval",xlab="Interval", ylab="Steps",layout=c(1,2), type="l")

