Loading and preprocessing the data Load the data Process/transform the data (if necessary) into a format suitable for your analysis

# download file from web  
download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip", destfile = "activity.zip", mode="wb")  
# unzip data and read   
unzip("activity.zip")  
stepdata <- read.csv("activity.csv", header = TRUE)  
head(stepdata)

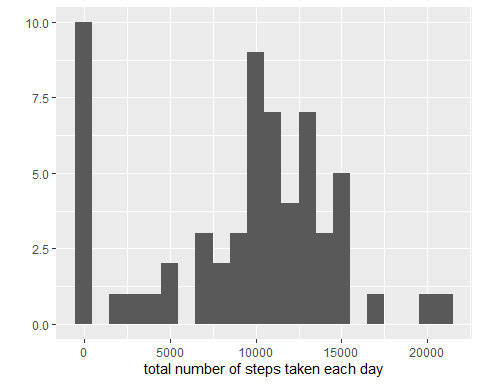
## steps date interval  
## 1 NA 2012-10-01 0  
## 2 NA 2012-10-01 5  
## 3 NA 2012-10-01 10  
## 4 NA 2012-10-01 15  
## 5 NA 2012-10-01 20  
## 6 NA 2012-10-01 25

## mean total number of steps taken per day.

library(ggplot2)  
total.steps <- tapply(stepdata$steps, stepdata$date, FUN=sum, na.rm=TRUE)

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

qplot(total.steps, binwidth=1000, xlab="total number of steps taken each day")



##### 2. Mean and median total number of steps taken per day

mean(total.steps, na.rm=TRUE)

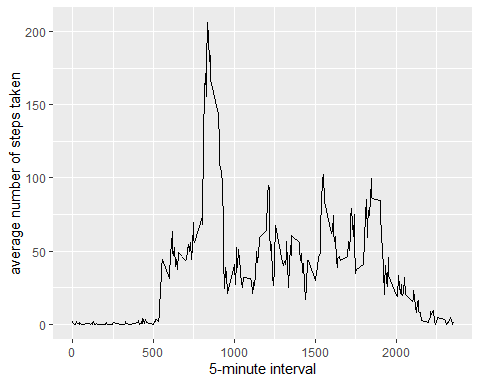
## [1] 9354.23

median(total.steps, na.rm=TRUE)

## [1] 10395

## What is the average daily activity pattern?

averages <- aggregate(x=list(steps=stepdata$steps), by=list(interval=stepdata$interval),  
 FUN=mean, na.rm=TRUE)  
ggplot(data=averages, aes(x=interval, y=steps)) +  
 geom\_line() +  
 xlab("5-minute interval") +  
 ylab("average number of steps taken")



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

averages[which.max(averages$steps),]

## interval steps  
## 104 835 206.1698

## Imputing missing values

There are many days/intervals where there are missing values (coded as NA). The presence of missing days may introduce bias into some calculations or summaries of the data.

missing <- is.na(stepdata$steps)  
# How many missing  
table(missing)

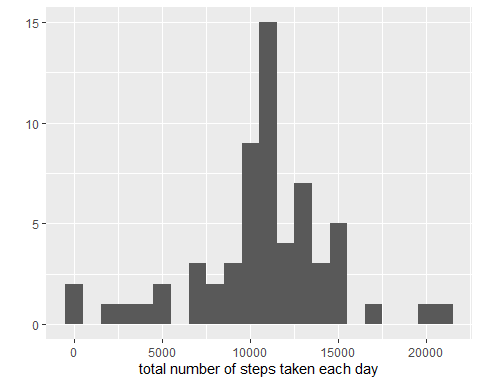
## missing  
## FALSE TRUE   
## 15264 2304

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

# Replace each missing value with the mean value of its 5-minute interval  
fill.value <- function(steps, interval) {  
 filled <- NA  
 if (!is.na(steps))  
 filled <- c(steps)  
 else  
 filled <- (averages[averages$interval==interval, "steps"])  
 return(filled)  
}  
filled.data <- stepdata  
filled.data$steps <- mapply(fill.value, filled.data$steps, filled.data$interval)

Now, using the filled data set, let’s make a histogram of the total number of steps taken each day and calculate the mean and median total number of steps.

total.steps <- tapply(filled.data$steps, filled.data$date, FUN=sum)  
qplot(total.steps, binwidth=1000, xlab="total number of steps taken each day")



mean(total.steps)

## [1] 10766.19

median(total.steps)

## [1] 10766.19

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

First, let’s find the day of the week for each measurement in the dataset. In this part, we use the dataset with the filled-in values.

weekday.or.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")  
}  
filled.data$date <- as.Date(filled.data$date)  
filled.data$day <- sapply(filled.data$date, FUN=weekday.or.weekend)

Now, let’s make a panel plot containing plots of average number of steps taken on weekdays and weekends.

averages <- aggregate(steps ~ interval + day, data=filled.data, mean)  
ggplot(averages, aes(interval, steps)) + geom\_line() + facet\_grid(day ~ .) +  
 xlab("5-minute interval") + ylab("Number of steps")

