Reproducible research

## basic setting

##   
## 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

# We first remove NAs from the data, then change structure of date from factors to date.

data1 <- data1\_raw[!is.na(data1\_raw$steps), ]  
data1$date <- as.Date(data1$date)

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

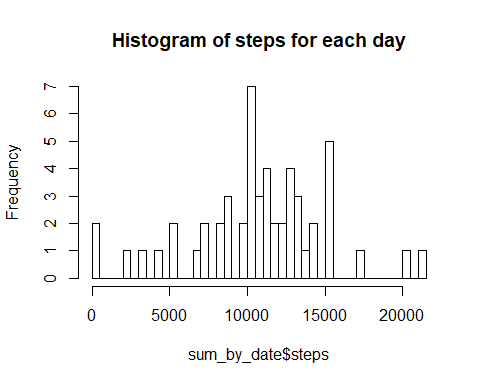
# 1. Calculate the total number of steps taken per day

data1\_grouped\_date <- group\_by(data1,date)  
sum\_by\_date <- summarize\_all(data1\_grouped\_date, sum)  
head(sum\_by\_date)

## # A tibble: 6 × 3  
## date steps interval  
## <date> <int> <int>  
## 1 2012-10-02 126 339120  
## 2 2012-10-03 11352 339120  
## 3 2012-10-04 12116 339120  
## 4 2012-10-05 13294 339120  
## 5 2012-10-06 15420 339120  
## 6 2012-10-07 11015 339120

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

hist(sum\_by\_date$steps, breaks = 52, main = "Histogram of steps for each day")



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

mean(sum\_by\_date$steps)

## [1] 10766.19

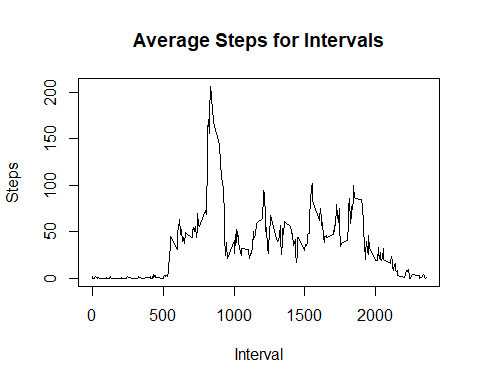
median(sum\_by\_date$steps)

## [1] 10765

# What is the average daily activity pattern?

# 1. Make a time series plot (i.e. type = "l") of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis)

data1\_grouped\_interval <- group\_by(data1, interval)  
data1\_grouped\_date$date <- 0  
mean\_by\_interval <- summarize(data1\_grouped\_interval, mean(steps))  
 colnames(mean\_by\_interval) <- c("interval", "steps")  
plot(mean\_by\_interval$steps~ mean\_by\_interval$interval, type="l",ylab = "Steps", xlab= "Interval", main="Average Steps for Intervals" )



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

head(mean\_by\_interval[rev(order(mean\_by\_interval$steps)),][1,])

## # A tibble: 1 × 2  
## interval steps  
## <int> <dbl>  
## 1 835 206.1698

# Imputing missing values

# Note that there are a number of 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.

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

## steps date interval   
## Mode :logical Mode :logical Mode :logical   
## FALSE:15264 FALSE:17568 FALSE:17568   
## TRUE :2304 NA's :0 NA's :0   
## NA's :0

# 2.Devise a strategy for filling in all of the missing values in the dataset. The strategy does not need to be sophisticated. For example, you could use the mean/median for that day, or the mean for that 5-minute interval, etc.

# We will fill empty rows with average for the 5 minute intervals. Selected\_intervals will indicate which interval has empty rows.

selected\_intervals <- (data1\_raw$interval[is.na(data1\_raw)] %/% 100 ) \* 12 + (data1\_raw$interval[is.na(data1\_raw)] %% 100 ) / 5 + 1

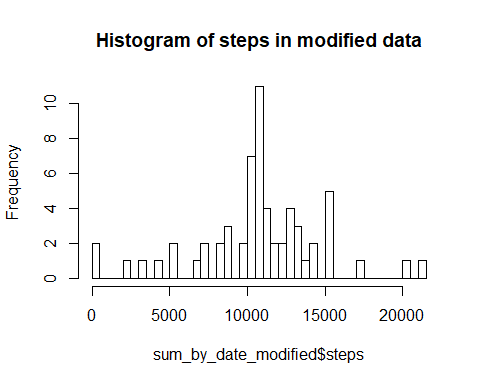
# 3.Create a new dataset that is equal to the original dataset but with the missing data filled in.

data1\_modified <- data1\_raw  
data1\_modified$steps <- replace(data1\_raw$steps, is.na(data1\_raw$steps), mean\_by\_interval$steps[selected\_intervals])

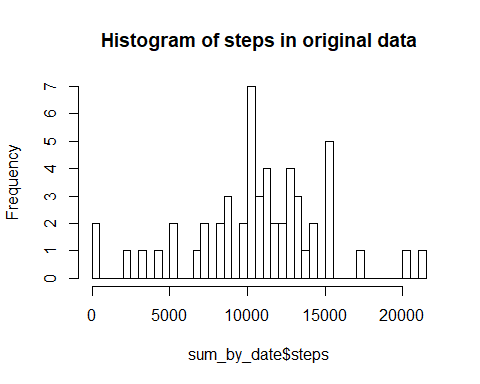
# 4.Make a histogram of the total number of steps taken each day and Calculate and report the mean and median total number of steps taken per day. Do these values differ from the estimates from the first part of the assignment? What is the impact of imputing missing data on the estimates of the total daily number of steps?

# After replacing empty rows with average for the corresponding interval, they look alike, but the spike at the medium soared.

data1\_modified\_grouped <- group\_by(data1\_modified, date)  
sum\_by\_date\_modified <- summarize\_all(data1\_modified\_grouped, sum)  
sum\_by\_date\_modified$date <- as.Date(sum\_by\_date\_modified$date)  
  
hist(sum\_by\_date\_modified$steps, breaks = 40, main = "Histogram of steps in modified data")



hist(sum\_by\_date$steps, breaks = 40, main = "Histogram of steps in original data")

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

# 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.

data1$weekday <- weekdays(data1$date)  
data1$weekday[data1$weekday != "Saturday" & data1$weekday != "Sunday"] <- "weekday"  
data1$weekday[data1$weekday == "Saturday" | data1$weekday == "Sunday"] <- "weekend"

# 2. #Make a panel plot containing a time series plot (i.e. type = "l") of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all weekday days or weekend days (y-axis). See the README file in the GitHub repository to see an example of what this plot should look like using simulated data.

