Unzip the folder and read in the data in the activity.csv file.

**if** (!file.exists('activity.csv')) {

unzip(zipfile = "activity.zip")

}

activityData <- read.csv(file="activity.csv", header=TRUE)

What is mean total number of steps taken per day?

*# Calculate the total steps taken per day*

totalSteps <- aggregate(steps ~ date, activityData, FUN=sum)

*# Make a histogram of the total number of steps taken per day*

hist(totalSteps$steps,

main = "Total Steps per Day",

xlab = "Number of Steps")

*# Calculate and report the mean and median of total steps taken per day*

meanSteps <- mean(totalSteps$steps, na.rm = TRUE)

medSteps <- median(totalSteps$steps, na.rm = TRUE)

Mean Number of Steps Taken per Day = 1.076618910^{4}

Median Number of Steps Taken per Day = 10765

What is the average daily activity pattern?

*# Make a time-series plot of the 5-minute interval and the average number of*

*# steps taken, averaged acoss all days.*

**library**(ggplot2)

meanStepsByInt <- aggregate(steps ~ interval, activityData, mean)

ggplot(data = meanStepsByInt, aes(x = interval, y = steps)) +

geom\_line() +

ggtitle("Average Daily Activity Pattern") +

xlab("5-minute Interval") +

ylab("Average Number of Steps") +

theme(plot.title = element\_text(hjust = 0.5))

*# Which 5-minute interval across all days contain the maximum number of steps*

maxInt <- meanStepsByInt[which.max(meanStepsByInt$steps),]

Imputing Missing Values

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

missingVals <- is.na(activityData$steps)

*# Devise a strategy for filling in all of the missing values*

There are 17568 missing values. I will replace these missing values with the 5-day average of that respective interval.

*# Create a new dataset that is equal to the original dataset but with*

*# the missing data filled in.*

imp\_activityData <- transform(activityData,

steps = ifelse(is.na(activityData$steps),

meanStepsByInt$steps[match(activityData$interval,

meanStepsByInt$interval)],

activityData$steps))

*# Make a histogram of the total number of steps taken each day and*

*# and report the mean and median.*

impStepsByInt <- aggregate(steps ~ date, imp\_activityData, FUN=sum)

hist(impStepsByInt$steps,

main = "Imputed Number of Steps Per Day",

xlab = "Number of Steps")

impMeanSteps <- mean(impStepsByInt$steps, na.rm = TRUE)

impMedSteps <- median(impStepsByInt$steps, na.rm = TRUE)

diffMean = impMeanSteps - meanSteps

diffMed = impMedSteps - medSteps

diffTotal = sum(impStepsByInt$steps) - sum(totalSteps$steps)

There is a difference of 0 in the mean steps of the two dataset. There is a difference of -1.076381110^{4} in the median steps of the two dataset. There is a difference of 8.612950910^{4} in the total steps of the two dataset.

Are there differences in activity patterns between weekdays and weekends?

*# Create a new factor variable in the dataset with two levels - "weekend" and "weekday"*

DayType <- **function**(date) {

day <- weekdays(date)

**if** (day %in% c('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday'))

**return** ("weekeday")

**else** **if** (day %in% c('Saturday', 'Sunday'))

**return** ("weekend")

**else**

**stop** ("Invalid Date Format.")

}

imp\_activityData$date <- as.Date(imp\_activityData$date)

imp\_activityData$day <- sapply(imp\_activityData$date, FUN = DayType)

*# Make a panel plot containnig a time-series plot of the 5-minute interval*

*# and the average number of steps taken across all weekdays or weekends*

meanStepsByDay <- aggregate(steps ~ interval + day, imp\_activityData, mean)

ggplot(data = meanStepsByDay, aes(x = interval, y = steps)) +

geom\_line() +

facet\_grid(day ~ .) +

ggtitle("Average Daily Activity Pattern") +

xlab("5-minute Interval") +

ylab("Average Number of Steps") +

theme(plot.title = element\_text(hjust = 0.5))