## Fábio Alves Carvalho

## [carvalho.alves.fabio@gmail.com](mailto:carvalho.alves.fabio@gmail.com)

## Brazil, São Paulo.

echo = TRUE

library(plyr)

## Loading and preprocessing the data

rm(list=ls()) setwd("C:/Git\_Paste/RepData\_PeerAssessment1/") list.files() unzip("activity.zip") data <- read.csv("activity.csv") head(data) head(total.steps.day)

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

### Sum steps by day, create Histogram, and calculate mean and median

library(ggplot2) total.steps.day <- tapply(datadate, sum) #steps\_by\_day <- aggregate(steps ~ date, data, sum) Other way hist(total.steps.day, main = paste("Total Steps Each Day"), col = "blue", xlab = "Number of Steps") #steps.by.day <- aggregate(datadate, FUN = sum) #colnames(steps.by.day)<- c("Date", "Steps") #hist(data$steps, breaks=5, xlab = "Steps", main = "Total Steps per Day") #qplot(total.steps.day, binwidth=1000, xlab="Total Number of Steps by Day")

## Mean of Steps

mean(total.steps.day, na.rm = T)

## Median of Steps

median(total.steps.day, na.rm = T)

### The average number of steps taken each day was 10766 steps.

### The median number of steps taken each day was 10765 steps.

# What is the average daily activity pattern?

averages <- aggregate(x=list(steps=datainterval), 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),]

### The maximum number of steps for a 5-minute interval was 206 steps and the 5-minute interval which had the maximum number of steps was the 835 interval

# Imputing missing values

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

missing.na.data <- is.na(data$steps) table(missing.na.data)

### The total number of rows with steps = 'NA' is 2304.

### 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

### pulling data without nas

clean <- data[!is.na(data$steps),]

#### Create the average number of steps per weekday and interval

avgTable <- ddply(clean, .(interval, date), summarize, Avg = mean(steps))

## Create dataset with all NAs for substitution

nadata<- data[is.na(data$steps),] ## Merge NA data with average weekday interval for substitution newdata<-merge(nadata, avgTable, by=c("interval", "date"))

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

## Reorder the new substituded data in the same format as clean data set

newdata2<- newdata[,c(6,4,1,2,5)] colnames(newdata2)<- c("steps", "date", "interval", "date", "DateTime")

## Merge the NA averages and non NA data together

mergeData <- rbind(clean, newdata2)

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