Reproducible Research: Peer Assessment 1

# Load required packages  
library(knitr)  
library(dplyr)

##   
## 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

library(ggplot2)  
knitr::opts\_chunk$set(echo = TRUE)  
options(scipen=999)

## Loading and preprocessing the data

1. Load the data (i.e. read.csv())
2. Process/transform the data (if necessary) into a format suitable for your analysis

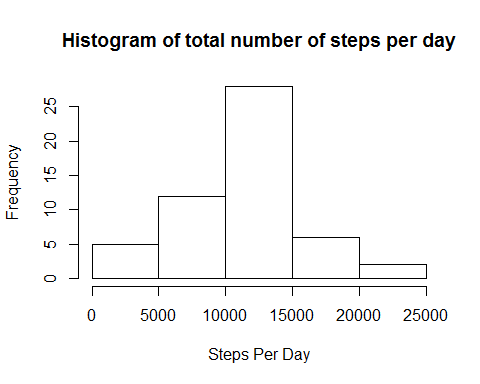
# set the destination file path  
f <- file.path(getwd(), "activity.zip")  
# downoad file to given destination  
download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip",f)  
# unzip acitivity data file  
uzf <- unzip(f)  
# load data  
raw\_data <- read.csv(uzf)  
# omit NAs  
data <- na.omit(raw\_data)

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

1. Calculate the total number of steps taken per day
2. If you do not understand the difference between a histogram and a barplot, research the difference between them. Make a histogram of the total number of steps taken each day

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

steps\_per\_day <- tapply(data$steps, data$date, sum, na.rm=TRUE)  
## plot historgram  
hist(steps\_per\_day, main="Histogram of total number of steps per day", xlab="Steps Per Day", border=1)



## Mean of the total steps takes per day?  
mean(steps\_per\_day, na.rm=TRUE)

## [1] 10766.19

median(steps\_per\_day, na.rm=TRUE)

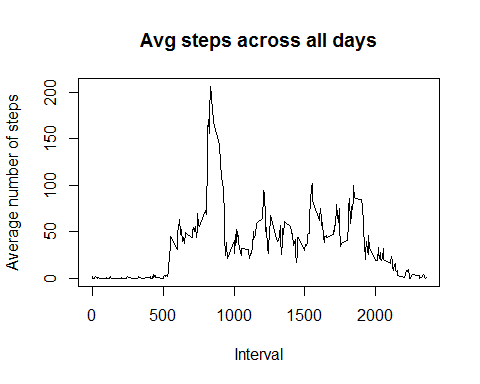
## [1] 10765

**Mean of the total number of steps per day is 10766.19 ,and median is 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)

interval\_steps <- aggregate(steps~interval, data, mean)  
plot(interval\_steps$interval, interval\_steps$steps, col=1,main="Avg steps across all days", xlab="Interval", ylab="Average number of steps", type='l')



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

# find row id of max avg number of steps in the interval  
max\_row\_id = which.max(interval\_steps$steps)  
## find the interval for the given max avg number of steps in the interval  
interval\_steps[max\_row\_id,]

## interval steps  
## 104 835 206.1698

**The interval 835 has the maximum average value of steps 206.17**

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

sum(is.na(raw\_data))

## [1] 2304

**Total number of rows with NA's is 2304**

1. 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

**My strategy is to fill the missing values with the mean for that 5 minute interval**

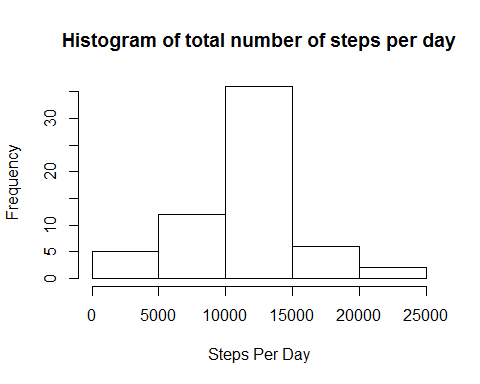
1. Create a new dataset that is equal to the original dataset but with the missing data filled in.

## Imputing missing values

fill\_data <- raw\_data  
interval\_steps <- aggregate(steps~interval, data, mean)  
for ( i in 1:nrow(raw\_data)) {   
 if(is.na(raw\_data$steps[i])) {   
 row\_id <- which(interval\_steps$interval==raw\_data$interval[i])  
 fill\_data$steps[i] <- interval\_steps$steps[row\_id]  
 }  
 }  
fill\_steps\_per\_day <- tapply(fill\_data$steps, fill\_data$date, sum, na.rm=TRUE)

1. 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?

hist(fill\_steps\_per\_day, main="Histogram of total number of steps per day", xlab="Steps Per Day", border=1)



# Mean of the total steps takes per day after imputing missing values  
mean(fill\_steps\_per\_day)

## [1] 10766.19

median(fill\_steps\_per\_day)

## [1] 10766.19

**Mean values remain same, but there is a slight difference in medians**

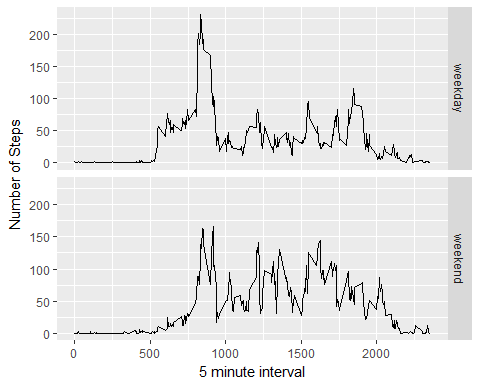
## Are there differences in activity patterns between weekdays and weekends?

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.

fill\_data['day\_type'] <- weekdays(as.Date(fill\_data$date))  
fill\_data$day\_type[fill\_data$day\_type %in% c('Saturday','Sunday')] <- "weekend"  
fill\_data$day\_type[fill\_data$day\_type != "weekend"] <- "weekday"

1. 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.

# convert day type from character to factor  
fill\_data$day\_type <- as.factor(fill\_data$day\_type)  
steps\_by\_day\_type <- aggregate(steps~interval+day\_type, fill\_data, mean )  
ggplot(steps\_by\_day\_type, aes(interval, steps))+ geom\_line()+facet\_grid(day\_type~.)+xlab("5 minute interval")+ ylab("Number of Steps")



**Relatively more foot steps in the range of 1000-2000 over the weekends**