Reproducible Research: Peer Assessment 1

## Loading Data

library(dplyr)

##   
## Attaching package: 'dplyr'

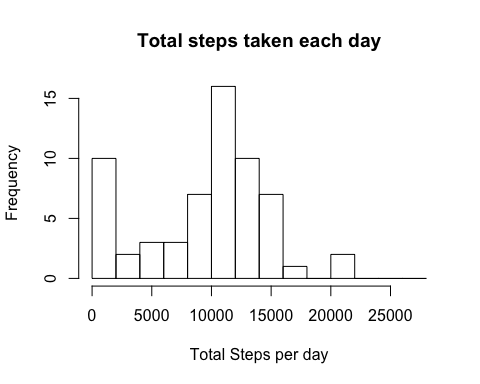
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

activity\_data <- read.csv("/Users/fezekilemdluli/Coursera/reproducible-research/activity.csv")  
activity\_data$date<-as.Date(activity\_data$date,tz="")

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

total\_steps\_per\_day <- with(activity\_data, aggregate(steps,list(date),FUN = sum,na.rm=TRUE))  
  
colnames(total\_steps\_per\_day)<- c("date", "steps")  
  
hist(total\_steps\_per\_day$steps, xlab = "Total Steps per day", main = "Total steps taken each day",breaks = seq(0,28000, by=2000))

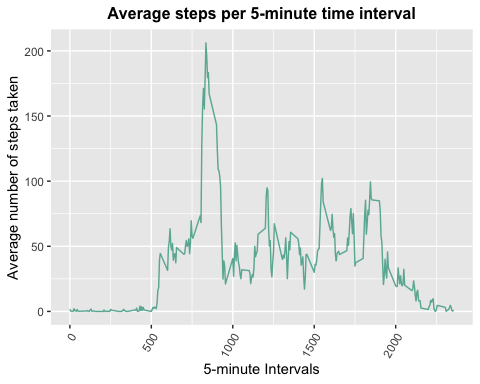


mean\_steps <- mean(total\_steps\_per\_day$steps, na.rm= TRUE)  
  
median\_steps <- median(total\_steps\_per\_day$steps, na.rm= TRUE)

* The Mean is 9354.2295082
* The Median is 10395

## What is the average daily activity pattern?

library(ggplot2)  
  
activity\_subset<- aggregate(steps ~ interval, data = activity\_data, FUN = mean, na.action = na.omit)  
colnames(activity\_subset) <- c("interval", "steps")   
  
  
p <-ggplot(activity\_subset, aes(activity\_subset$interval, activity\_subset$steps))  
p + geom\_line(color='#69b3a2') +  
 xlab ('5-minute Intervals')+  
 ylab('Average number of steps taken')+  
 theme(axis.text.x=element\_text(angle=60, hjust=1))+  
 theme(plot.title = element\_text(face = "bold", size = 12, hjust = 0.5)) +  
 ggtitle('Average steps per 5-minute time interval')



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

max\_step\_interval <-activity\_subset[activity\_subset$steps == max(activity\_subset$steps), ]$interval  
  
max\_step\_interval

## [1] 835

## 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\_values <- sum(is.na(activity\_data$steps)==TRUE)

* The the total number of missing values in the dataset is 2304

2.Strategy for filling in all of the missing values in the dataset. Missing values will be replaced by the mean for the day.

activity\_data$complete <- ifelse(is.na(activity\_data$steps),round(activity\_subset$steps[match(activity\_data$interval,activity\_subset$interval)],0),activity\_data$steps)

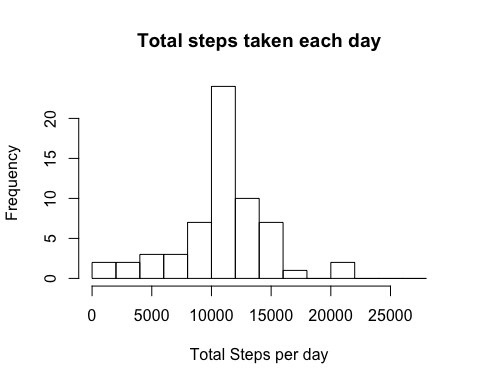
Create a new dataset that is equal to the original but with missing values filled in.

new\_dataset <-data.frame(steps=activity\_data$complete,date=activity\_data$date, interval = activity\_data$interval)  
head(new\_dataset,10)

## steps date interval  
## 1 2 2012-10-01 0  
## 2 0 2012-10-01 5  
## 3 0 2012-10-01 10  
## 4 0 2012-10-01 15  
## 5 0 2012-10-01 20  
## 6 2 2012-10-01 25  
## 7 1 2012-10-01 30  
## 8 1 2012-10-01 35  
## 9 0 2012-10-01 40  
## 10 1 2012-10-01 45

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?

complete\_steps\_per\_day <- with(new\_dataset, aggregate(steps,list(date),FUN = sum,na.rm=TRUE))  
#complete\_steps\_per\_day <- aggregate(new\_dataset$steps,list(new\_dataset$date),FUN = sum)  
  
colnames(complete\_steps\_per\_day)<- c("date", "steps")  
  
hist(complete\_steps\_per\_day$steps, xlab = "Total Steps per day", main = "Total steps taken each day",breaks = seq(0,28000, by=2000))



Calculate Mean and Median using completed data

imp\_mean<-mean(complete\_steps\_per\_day$steps)  
imp\_mean

## [1] 10765.64

imp\_median<- median(complete\_steps\_per\_day$steps)  
imp\_median

## [1] 10762

Comarison of Mean and Median

* Original mean is 9354.2295082 while imputed mean is 10765.64`
* Original median is 10395 while imputed mean is 10762

Both the Mean and Median have slightly increased after filling the missing values.

## 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

day\_of\_the\_week <-weekdays(new\_dataset$date, abbreviate = FALSE)  
new\_dataset<-cbind(new\_dataset,day\_of\_the\_week)  
  
day\_factor <-ifelse(new\_dataset$day\_of\_the\_week == 'Saturday'| new\_dataset$day\_of\_the\_week=="Sunday", "Weekend", "Weekday")  
  
new\_dataset<-cbind(new\_dataset,day\_factor)  
  
  
head(new\_dataset)

## steps date interval day\_of\_the\_week day\_factor  
## 1 2 2012-10-01 0 Monday Weekday  
## 2 0 2012-10-01 5 Monday Weekday  
## 3 0 2012-10-01 10 Monday Weekday  
## 4 0 2012-10-01 15 Monday Weekday  
## 5 0 2012-10-01 20 Monday Weekday  
## 6 2 2012-10-01 25 Monday Weekday

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).

activity\_subset2<- aggregate(steps ~ interval + day\_factor, data = new\_dataset, FUN = mean, na.action = na.omit)  
   
p <-ggplot(activity\_subset2, aes(activity\_subset2$interval/100, activity\_subset2$steps))  
p + geom\_line(color='#69b3a2') +  
 xlab ('5-minute Intervals')+  
 ylab('Average number of steps taken')+  
 theme(axis.text.x=element\_text(angle=60, hjust=1))+  
 theme(plot.title = element\_text(face = "bold", size = 12, hjust = 0.5)) +  
 ggtitle('Average steps per 5-minute time interval')+  
 facet\_grid(day\_factor ~ .)

