Coursera Reproducible Research Project 1

Selwyn Leonard

23 August 2018

Loading the data

library(ggplot2)  
library(plyr)  
  
##reading/loading the data into R  
activity <- read.csv("activity.csv",header=TRUE,stringsAsFactors = FALSE,strip.white =   
 TRUE,sep = ',')

Processing the activity data

activity$day <- weekdays(as.Date(activity$date))  
activity$DateTime <- as.POSIXct(activity$date, format = "%Y%m%d")  
  
##pulling the data without nas  
clean <- activity[!is.na(activity$steps),]

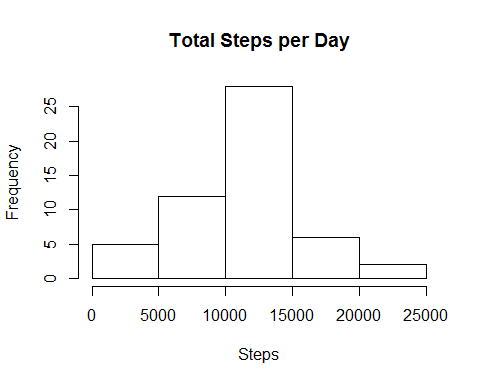
## What is the mean total number of steps taken per day?

Calculate the total number of steps taken per day

sumTable <- aggregate(activity$steps ~ activity$date, FUN=sum)  
colnames(sumTable)<- c("Date", "Steps")

Make a histogram of the total number of steps taken each day

## Creating a histogram of the total steps per day  
hist(sumTable$Steps, breaks = 5, xlab = "Steps", main = "Total Steps per Day")



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

## Mean number of Steps  
as.integer(mean(sumTable$Steps))

## [1] 10766

## Meadian number of Step  
as.integer(median(sumTable$Steps))

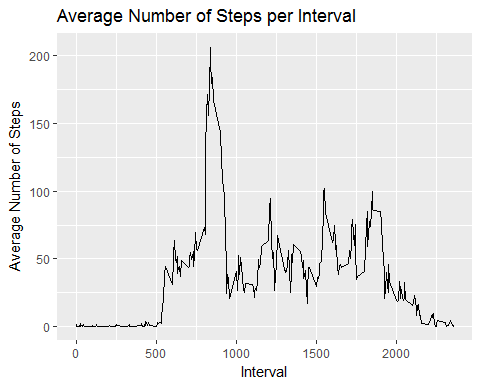
## [1] 10765

The mean number of steps taken each day was 10 766 steps. The median nuber of steps taken each day was 10 765 steps.

## What is the average daily activity pattern

Make a time series plot (i.e. type = “l”) of the 5-minut interval (x-axis) and the average number of steps taken, average across all days (y-axis)

library(plyr)  
library(ggplot2)  
##pulling data without nas  
clean <- activity[!is.na(activity$steps),]  
  
##create average number of steps per interval  
intervalTable <- ddply(clean, .(interval), summarize, Avg = mean(steps))  
  
##Create line plot of average number of steps per interval  
p <- ggplot(intervalTable, aes(x=interval, y=Avg), xlab = "Interval", ylab="Average Number of Steps")  
p + geom\_line()+xlab("Interval")+ylab("Average Number of Steps")+ggtitle("Average Number of Steps per Interval")



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

## Maximum steps by interval  
maxSteps <- max(intervalTable$Avg)  
  
##Which interval contains the maximum average number of steps  
intervalTable[intervalTable$Avg==maxSteps,1]

## [1] 835

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

## Imputing missing values

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

## Number of NAs in the orginal data set  
nrow(activity[is.na(activity$steps),])

## [1] 2304

The total number of rows with steps = ‘NA’ is 2304.

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 the 5-minute interval, etc.

My strategy for filling in NAs will be to subsitute the missing steps with the average 5-minute nterval based on the day of the week.

## Create the average number of steps per week and interval  
avgTable <- ddply(clean, .(interval, day), summarize, Avg = mean(steps))  
  
## Create dataset with all NAs for substitution  
nadata<- activity[is.na(activity$steps),]  
## Merge NA data with average weekday interval for substitution  
newdata<-merge(nadata, avgTable, by=c("interval", "day"))

## 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 smae format as clean data set  
newdata2 <- newdata[,c(6,4,1,2,5)]  
colnames(newdata2) <- c("steps", "date", "interval", "day", "DateTime")  
  
## Merge he NA averages and non NA data together  
mergeData <- rbind(clean, newdata2)

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 missing data on the estimates of the total daily number of steps?

## Create sum of steps per date to compare with step 1  
sumTable2 <- aggregate(mergeData$steps ~ mergeData$date, FUN=sum, )  
colnames(sumTable2)<- c("Date", "Steps")  
  
## Mean of Steps with NA data taken care of  
as.integer(mean(sumTable2$Steps))

## [1] 10821

## Median of Steps with NA data taken care of  
as.integer(median(sumTable2$Steps))

## [1] 11015

## Creating the histogram of total steps per day, categorized by data set to show impact  
hist(sumTable2$Steps, breakes=5, xlab = "Steps", main = "Total Steps per Day with NAs Fixed", col = "Black")

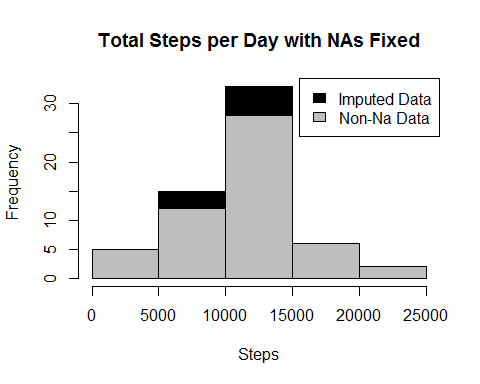
## Warning in plot.window(xlim, ylim, "", ...): "breakes" is not a graphical  
## parameter

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):  
## "breakes" is not a graphical parameter

## Warning in axis(1, ...): "breakes" is not a graphical parameter

## Warning in axis(2, ...): "breakes" is not a graphical parameter

hist(sumTable$Steps, breaks = 5, xlab = "Steps", main = "Total Steps per Day with NAs Fixed", col = "Grey", add=T)  
legend("topright", c("Imputed Data", "Non-Na Data"), fill = c("black", "grey"))



The new mean of the imputed data is 10 821 steps to the old mean of 10 766 steps. That creates a difference of 55 steps on average per day. The new median of the imputed data is 11 015 steps compared to the old median of the 10 765 steps. That creates a difference of 250 steps for the median. However, the overall shape of the distribution has not changed.

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

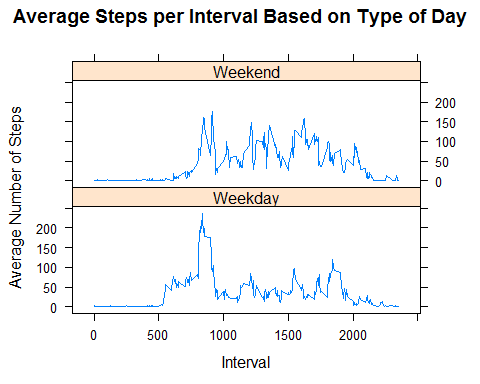
Create a new factor variable in the dataset with two levels - “weekday” and “weekend” including whether a given date is a weekday or weekend day.

## Create new category based on the days of the week  
mergeData$DayCategory <- ifelse(mergeData$day %in% c("Saturday", "Sunday"), "Weekend", "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 and weekend days (y-axis).

library(lattice)

## Summarize data by interval and type of day  
intervalTable2 <- ddply(mergeData, .(interval, DayCategory), summarize, Avg = mean(steps))  
  
##Plot data in a panel plot  
xyplot(Avg~interval|DayCategory, data=intervalTable2, type="l", layout = c(1,2),  
 main="Average Steps per Interval Based on Type of Day",   
 ylab="Average Number of Steps", xlab="Interval")



Yes, the step activity trends are different based on whether the day occurs on a weekend or not. This may be due to people having an increased opportunity for activity beyond normal work hours for those who work during the week.