Reproducible Research Project 1

Jaime Jordan

August 27, 2016

# Reproducible Research Project 1

## Jaime Jordan

Loading and preprocessing the data:

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.2.5

##   
## 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

Attaching package: 'dplyr'

setwd("C:/Users/Jaime/Desktop/Coursera/ReproducibleResearch")  
activitydata<- read.csv("activity.csv")  
head(activitydata)

## steps date interval  
## 1 NA 2012-10-01 0  
## 2 NA 2012-10-01 5  
## 3 NA 2012-10-01 10  
## 4 NA 2012-10-01 15  
## 5 NA 2012-10-01 20  
## 6 NA 2012-10-01 25

dim(activitydata)

## [1] 17568 3

glimpse(activitydata)

## Observations: 17,568  
## Variables: 3  
## $ steps (int) NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...  
## $ date (fctr) 2012-10-01, 2012-10-01, 2012-10-01, 2012-10-01, 2012...  
## $ interval (int) 0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 100, 10...

summary(activitydata)

## steps date interval   
## Min. : 0.00 2012-10-01: 288 Min. : 0.0   
## 1st Qu.: 0.00 2012-10-02: 288 1st Qu.: 588.8   
## Median : 0.00 2012-10-03: 288 Median :1177.5   
## Mean : 37.38 2012-10-04: 288 Mean :1177.5   
## 3rd Qu.: 12.00 2012-10-05: 288 3rd Qu.:1766.2   
## Max. :806.00 2012-10-06: 288 Max. :2355.0   
## NA's :2304 (Other) :15840

activitydata$date<- as.Date(activitydata$date) ##convert chr to date

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

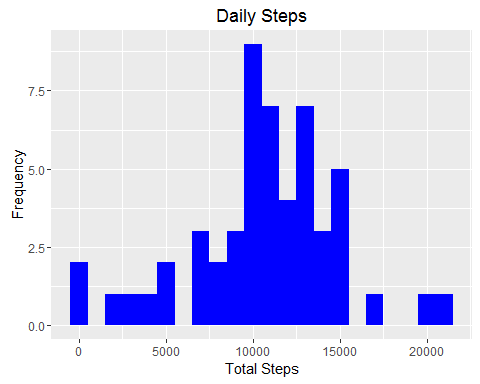
###Calculate the total number of steps taken per day:  
Total\_Steps<-activitydata%>%  
 group\_by(date)%>%  
 filter(!is.na(steps))%>%  
 summarize(total\_steps = sum(steps, na.rm=TRUE))  
Total\_Steps

## Source: local data frame [53 x 2]  
##   
## date total\_steps  
## (date) (int)  
## 1 2012-10-02 126  
## 2 2012-10-03 11352  
## 3 2012-10-04 12116  
## 4 2012-10-05 13294  
## 5 2012-10-06 15420  
## 6 2012-10-07 11015  
## 7 2012-10-09 12811  
## 8 2012-10-10 9900  
## 9 2012-10-11 10304  
## 10 2012-10-12 17382  
## .. ... ...

##Plot using ggplot  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.2.5

ggplot(Total\_Steps, aes(x=total\_steps)) +  
 geom\_histogram(fill = "blue", binwidth = 1000) +  
 labs(title = "Daily Steps", x= "Total Steps", y = "Frequency")



##Calculate and report the mean and median of the total number of steps taken per day  
Mean\_Steps<- mean(Total\_Steps$total\_steps, na.rm=TRUE)  
Mean\_Steps

## [1] 10766.19

Median\_Steps<- median(Total\_Steps$total\_steps, na.rm=TRUE)  
Median\_Steps

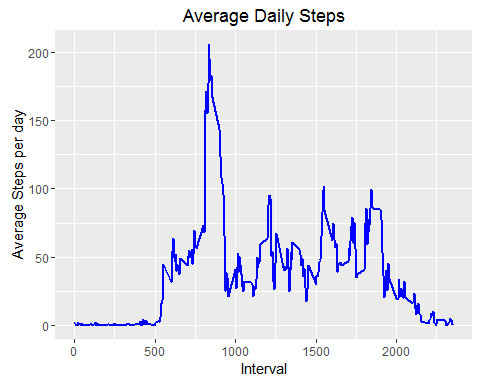
## [1] 10765

## What is the average daily activity pattern?

## Calculate the average steps  
Interval<- activitydata%>%  
 group\_by(interval)%>%  
 filter(!is.na(steps))%>%  
 summarize(avg\_steps = mean(steps, na.rm=TRUE))  
Interval

## Source: local data frame [288 x 2]  
##   
## interval avg\_steps  
## (int) (dbl)  
## 1 0 1.7169811  
## 2 5 0.3396226  
## 3 10 0.1320755  
## 4 15 0.1509434  
## 5 20 0.0754717  
## 6 25 2.0943396  
## 7 30 0.5283019  
## 8 35 0.8679245  
## 9 40 0.0000000  
## 10 45 1.4716981  
## .. ... ...

##Plot average steps  
ggplot(Interval, aes(x=interval , y=avg\_steps)) +  
 geom\_line(color="blue", size=1) +  
 labs(title = "Average Daily Steps", x = "Interval", y = "Average Steps per day")



###Which 5-minute interval, on average across all the days in the dataset, contains the maximum number of steps?  
Interval[which.max(Interval$avg\_steps),]

## Source: local data frame [1 x 2]  
##   
## interval avg\_steps  
## (int) (dbl)  
## 1 835 206.1698

## Imputing Missing Values

##Calculate total number of missing values in the dataset  
sum(is.na(activitydata$steps))

## [1] 2304

##Imputing missing values using mean for each day and create a new dataset that is equal to the original dataset but with the missing data filled in.  
activitydata2<- activitydata  
nas<- is.na(activitydata2$steps)  
avg\_interval<- tapply(activitydata2$steps, activitydata2$interval, mean, na.rm=TRUE, simplify = TRUE)  
activitydata2$steps[nas] <- avg\_interval[as.character(activitydata2$interval[nas])]  
names(activitydata2)

## [1] "steps" "date" "interval"

##Check if no missing value is appearing  
sum(is.na(activitydata2))

## [1] 0

##Reorder columns (for better understanding of the data)  
activitydata2 <- activitydata2[, c("date", "interval", "steps")]  
head(activitydata2)

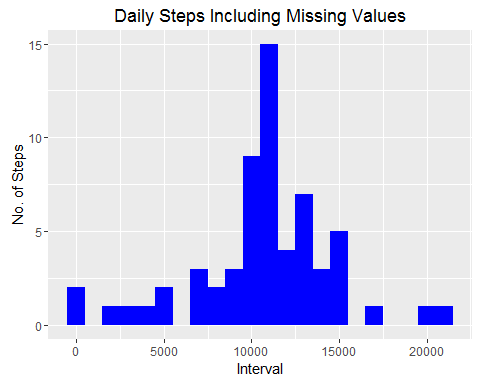
## date interval steps  
## 1 2012-10-01 0 1.7169811  
## 2 2012-10-01 5 0.3396226  
## 3 2012-10-01 10 0.1320755  
## 4 2012-10-01 15 0.1509434  
## 5 2012-10-01 20 0.0754717  
## 6 2012-10-01 25 2.0943396

## Create histogram of the total number of steps taken each day and calculate mean and median total number of steps taken per day.

Total\_Steps2<- activitydata2%>%  
 group\_by(date)%>%  
 summarize(total\_steps = sum(steps, na.rm=TRUE))  
Total\_Steps2

## Source: local data frame [61 x 2]  
##   
## date total\_steps  
## (date) (dbl)  
## 1 2012-10-01 10766.19  
## 2 2012-10-02 126.00  
## 3 2012-10-03 11352.00  
## 4 2012-10-04 12116.00  
## 5 2012-10-05 13294.00  
## 6 2012-10-06 15420.00  
## 7 2012-10-07 11015.00  
## 8 2012-10-08 10766.19  
## 9 2012-10-09 12811.00  
## 10 2012-10-10 9900.00  
## .. ... ...

ggplot(Total\_Steps2, aes(x = total\_steps)) +  
 geom\_histogram(fill = "blue", binwidth = 1000) +  
 labs(title = "Daily Steps Including Missing Values", x = "Interval", y = "No. of Steps")



##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?  
Mean\_Steps2<- mean(Total\_Steps2$total\_steps, na.rm=TRUE)  
Mean\_Steps2

## [1] 10766.19

Median\_Steps2<- median(Total\_Steps2$total\_steps, na.rm=TRUE)  
Median\_Steps2

## [1] 10766.19

##Answer: The impact of imputing missing data with the average number of steps in the same 5-minute interval is that both the mean and the median are the same: 10766.19

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

head(activitydata2)

## date interval steps  
## 1 2012-10-01 0 1.7169811  
## 2 2012-10-01 5 0.3396226  
## 3 2012-10-01 10 0.1320755  
## 4 2012-10-01 15 0.1509434  
## 5 2012-10-01 20 0.0754717  
## 6 2012-10-01 25 2.0943396

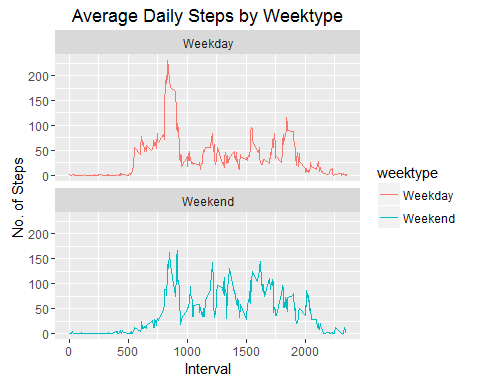
##Create new variable called WeekType for Weekday and Weekend  
activitydata2<- activitydata2%>%  
 mutate(weektype= ifelse(weekdays(activitydata2$date)=="Saturday" | weekdays(activitydata2$date)=="Sunday", "Weekend", "Weekday"))  
  
head(activitydata2)

## date interval steps weektype  
## 1 2012-10-01 0 1.7169811 Weekday  
## 2 2012-10-01 5 0.3396226 Weekday  
## 3 2012-10-01 10 0.1320755 Weekday  
## 4 2012-10-01 15 0.1509434 Weekday  
## 5 2012-10-01 20 0.0754717 Weekday  
## 6 2012-10-01 25 2.0943396 Weekday

##Plotting  
Interval2<-activitydata2%>%  
 group\_by(interval, weektype)%>%  
 summarize(avg\_steps2 = mean(steps, na.rm=TRUE))  
head(Interval2)

## Source: local data frame [6 x 3]  
## Groups: interval [3]  
##   
## interval weektype avg\_steps2  
## (int) (chr) (dbl)  
## 1 0 Weekday 2.25115304  
## 2 0 Weekend 0.21462264  
## 3 5 Weekday 0.44528302  
## 4 5 Weekend 0.04245283  
## 5 10 Weekday 0.17316562  
## 6 10 Weekend 0.01650943

plot<- ggplot(Interval2, aes(x = interval , y=avg\_steps2, color=weektype)) +  
 geom\_line() +  
 labs(title = "Average Daily Steps by Weektype", x = "Interval", y = "No. of Steps") +  
 facet\_wrap(~weektype, ncol = 1, nrow=2)  
print(plot)



##Answer: During the weekday, the test object is more active earlier in the day but the object is more active throughout the weekends. This is most likely due to the object working during the weekdays.