Reproducible Research: Peer Assessment 1

* The questions and the following answers of the assignment are listed in the commented sections of the code
* Setting the directory to where the file was downloaded getwd("/../..") setwd("/../..")

# Loading and preprocessing the data  
# Reading the file and transforming the date field to be a factor and then sum steps by date  
library(dplyr)

##   
## Attaching package: 'dplyr'

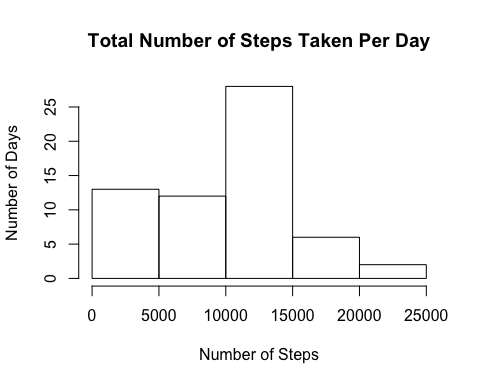
## The following objects are masked from 'package:stats':  
##   
## filter, lag

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

library(lattice)  
act\_data <- read.csv("activity.csv")  
head(act\_data)

## steps date interval  
## 1 NA 10/1/12 0  
## 2 NA 10/1/12 5  
## 3 NA 10/1/12 10  
## 4 NA 10/1/12 15  
## 5 NA 10/1/12 20  
## 6 NA 10/1/12 25

st\_data <- group\_by(act\_data, dt = as.factor(date))  
  
#What is the mean total number of steps taken per day  
  
#Total number of steps taken per day  
steps\_by\_day <- summarise(st\_data, st = sum(steps, na.rm = TRUE))  
  
#Histogram of sum of steps by date  
hist(steps\_by\_day$st, xlab = "Number of Steps", ylab = "Number of Days", main = "Total Number of Steps Taken Per Day")



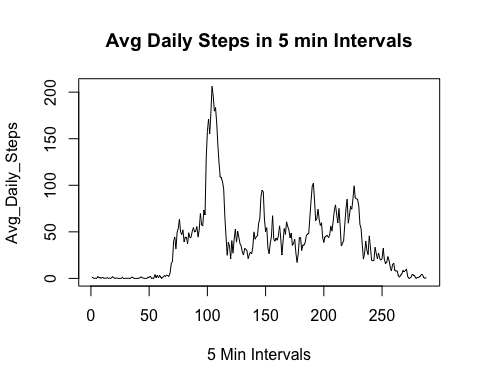
#Used data in steps\_by\_day to get the mean and median  
steps\_sum <- summarise(steps\_by\_day, st\_mean = mean(steps\_by\_day$st), st\_median = median(steps\_by\_day$st))  
steps\_sum

## # A tibble: 1 × 2  
## st\_mean st\_median  
## <dbl> <int>  
## 1 9354.23 10395

#What is the average daily activity pattern  
  
#Transforming the interval field to be a factor and then took the mean of steps by interval which spans multiple dates  
intdata <- group\_by(act\_data, int = as.factor(interval))  
intstdata <- summarise(intdata, Avg\_Daily\_Steps = mean(steps, na.rm = TRUE))  
  
#Converting the mean steps data to time series data  
stseries <- ts(intstdata)  
  
#5-minute interval across all days that contains the maximum number of steps  
stseries\_frame <- as.data.frame(stseries)  
head(stseries\_frame)

## int Avg\_Daily\_Steps  
## 1 1 1.7169811  
## 2 2 0.3396226  
## 3 3 0.1320755  
## 4 4 0.1509434  
## 5 5 0.0754717  
## 6 6 2.0943396

#Plot of interval and Avg Daily Steps  
plot(stseries\_frame, xlab = "5 Min Intervals", main = "Avg Daily Steps in 5 min Intervals", type = "l")



#Getting interval with max avg daily steps  
maxst <- max(stseries\_frame$Avg\_Daily\_Steps)  
max\_num\_steps <- filter(stseries\_frame, Avg\_Daily\_Steps == maxst)  
max\_num\_steps

## int Avg\_Daily\_Steps  
## 1 104 206.1698

#The 104th interval has the max avg daily steps (206.2). This interval corresponds to 830.  
  
#Imputing Missing Values  
  
#Total number of rows with NAs  
narows <- subset(act\_data, is.na(act\_data$steps))  
head(narows)

## steps date interval  
## 1 NA 10/1/12 0  
## 2 NA 10/1/12 5  
## 3 NA 10/1/12 10  
## 4 NA 10/1/12 15  
## 5 NA 10/1/12 20  
## 6 NA 10/1/12 25

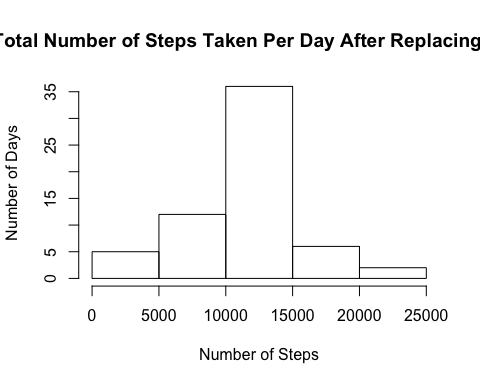
count(narows)

## # A tibble: 1 × 1  
## n  
## <int>  
## 1 2304

#Imputing the data and replacing the NAs with means by interval  
narows1 <- which(is.na(act\_data$steps))  
mean\_steps <- rep(mean(act\_data$steps, na.rm = TRUE), times = length(narows1))  
act\_data[narows1, "steps"] <- mean\_steps  
rm(mean\_steps, narows1)  
  
#Modified data set  
head(act\_data)

## steps date interval  
## 1 37.3826 10/1/12 0  
## 2 37.3826 10/1/12 5  
## 3 37.3826 10/1/12 10  
## 4 37.3826 10/1/12 15  
## 5 37.3826 10/1/12 20  
## 6 37.3826 10/1/12 25

#New data set and histograme of new data  
st\_data\_new <- group\_by(act\_data, dt\_new = as.factor(date))  
steps\_by\_day\_new <- summarise(st\_data\_new, st\_new = sum(steps))  
hist(steps\_by\_day\_new$st\_new, xlab = "Number of Steps", ylab = "Number of Days", main = "Total Number of Steps Taken Per Day After Replacing NAs")



#Mean and median of the new data set  
steps\_sum\_new <- summarise(steps\_by\_day\_new, st\_mean\_new = mean(steps\_by\_day\_new$st\_new, na.rm = TRUE), st\_median\_new = median(steps\_by\_day\_new$st\_new, na.rm = TRUE))  
steps\_sum\_new

## # A tibble: 1 × 2  
## st\_mean\_new st\_median\_new  
## <dbl> <dbl>  
## 1 10766.19 10766.19

# Yes the values differ. The impact is that the means and medians are larger now as more data has been added  
  
#Are there differences in activity patterns between weekdays and weekends?  
  
#The new data set with filled in values is act\_data which has been modified in the steps above  
#Weekdays creation  
wd\_data <- data.frame(date = as.Date(act\_data$date, format = "%m/%d/%y"), steps = act\_data$steps, interval = act\_data$interval)  
wd\_data1 <- data.frame(date = wd\_data$date, steps = wd\_data$steps, interval = wd\_data$interval, wkday = weekdays(wd\_data$date))  
wd\_data1$wkday <- ifelse(weekdays(wd\_data1$date) %in% c("Saturday", "Sunday"), "weekend", "weekday")  
wkd\_agg <- aggregate(steps ~ interval + wkday, data = wd\_data1, mean)  
  
#Time series plot by weekend and weekdays  
xyplot(steps ~ interval | wkday, data = wkd\_agg, type = "l", lwd = 2, layout = c(1,2), xlab = "Number of 5-min Intervals", ylab = "Avg Number of Steps", main = "Avg Num of Steps during weekdays and weekends")

