Anaylsis of AMD

Niall Graham

09 October 2017

Loading and preprocessing the data

Set Working Directory

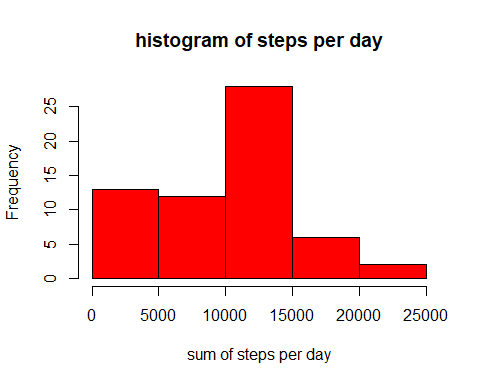
echo = TRUE  
activity <- NULL  
activity <- read.csv("activity.csv", header = T, sep = ",")

Variables created during by this code are set to NULL

echo = TRUE  
df\_summary <- NULL  
su2 <- NULL  
su <- NULL  
mn\_int <- NULL  
activity2 <- NULL  
mean\_su2 <- NULL  
median\_su2 <- NULL  
activity2\_weekend <- NULL  
activity2\_weekday <- NULL  
mean\_activity2\_weekday <- NULL  
mean\_activity2\_weekend <- NULL

1. Total Number of Steps Taken Per Day

echo = TRUE  
su <- tapply(activity$steps, activity$date, sum, na.rm=T)  
echo = TRUE  
hist(su, xlab = "sum of steps per day", main = "histogram of steps per day",col="red")



Mean Total Number of Steps

echo = TRUE  
mean\_su <- round(mean(su))  
median\_su <- round(median(su))  
print(c("The mean is",mean\_su))

## [1] "The mean is" "9354"

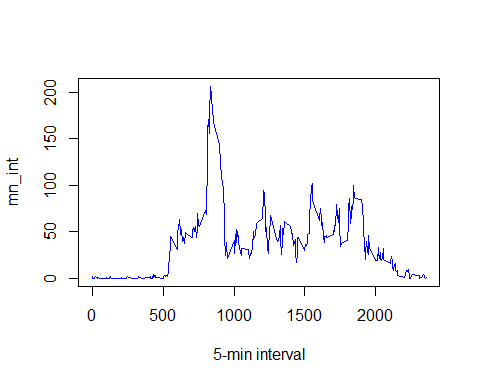
Median Total Number of Steps

print(c("The median is",median\_su))

## [1] "The median is" "10395"

1. Average Daily Activity Pattern

echo = TRUE  
mn\_int <- tapply(activity$steps, activity$interval, mean, na.rm=T)  
plot(mn\_int ~ unique(activity$interval), type="l", xlab = "5-min interval", col="blue")



2.b. 5- Minute Intervals

echo = TRUE  
mn\_int[which.max(mn\_int)]

## 835   
## 206.1698

3.Imputing missing values

echo = TRUE  
table(is.na(activity) == TRUE)

##   
## FALSE TRUE   
## 50400 2304

summary(activity)

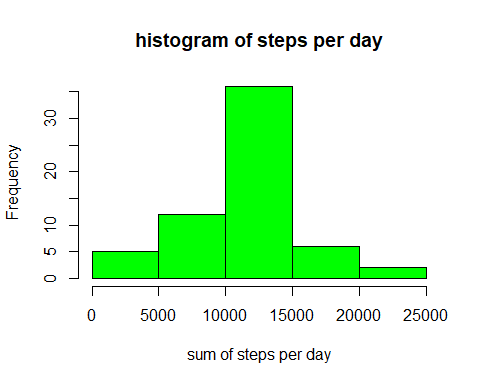
## steps date interval   
## Min. : 0.00 2012-10-01: 288 Min. : 0.0   
## 1st Qu.: 0.00 2012-10-02: 288 1st Qu.: 588.8   
## Median : 0.00 2012-10-03: 288 Median :1177.5   
## Mean : 37.38 2012-10-04: 288 Mean :1177.5   
## 3rd Qu.: 12.00 2012-10-05: 288 3rd Qu.:1766.2   
## Max. :806.00 2012-10-06: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

1. Filling in all of the Missing Values in the Dataset

echo = TRUE  
activity2 <- activity # creation of the dataset that will have no more NAs  
for (i in 1:nrow(activity)){  
 if(is.na(activity$steps[i])){  
 activity2$steps[i]<- mn\_int[[as.character(activity[i, "interval"])]]  
 }  
}

Histogram of steps per day

echo = TRUE  
su2 <- tapply(activity2$steps, activity2$date, sum, na.rm=T)  
hist(su2, xlab = "sum of steps per day", main = "histogram of steps per day", col="green")



Remove Na's

mean\_su2 <- round(mean(su2))  
median\_su2 <- round(median(su2))

The Mean:

echo = TRUE  
print(c("The mean is",mean\_su2))

## [1] "The mean is" "10766"

The Median:

print(c("The median is",median\_su2))

## [1] "The median is" "10766"

Comparison of the new values with the old values

echo = TRUE  
df\_summary <- rbind(df\_summary, data.frame(mean = c(mean\_su, mean\_su2), median = c(median\_su, median\_su2)))  
rownames(df\_summary) <- c("with NA's", "without NA's")  
print(df\_summary)

## mean median  
## with NA's 9354 10395  
## without NA's 10766 10766

echo = TRUE  
summary(activity2)

## 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.: 27.00 2012-10-05: 288 3rd Qu.:1766.2   
## Max. :806.00 2012-10-06: 288 Max. :2355.0   
## (Other) :15840

Confirmed there are no more NA's in the steps variable.

4.Differences in Activity Patterns between Weekdays and Weekends:

New Colmun added to dataframe containing the factors weekday days & weekend days

echo = TRUE  
activity2$weekday <- c("weekday")  
activity2[weekdays(as.Date(activity2[, 2])) %in% c("Saturday", "Sunday", "samedi", "dimanche", "saturday", "sunday", "Samedi", "Dimanche"), ][4] <- c("weekend")  
table(activity2$weekday == "weekend")

##   
## FALSE TRUE   
## 12960 4608

activity2$weekday <- factor(activity2$weekday)

In order to visualize the difference between weekends and days of the week, a new dataframe is created to be usable by the lattice package.

echo = TRUE  
activity2\_weekend <- subset(activity2, activity2$weekday == "weekend")  
activity2\_weekday <- subset(activity2, activity2$weekday == "weekday")  
  
mean\_activity2\_weekday <- tapply(activity2\_weekday$steps, activity2\_weekday$interval, mean)  
mean\_activity2\_weekend <- tapply(activity2\_weekend$steps, activity2\_weekend$interval, mean)

echo = TRUE  
library(lattice)  
df\_weekday <- NULL  
df\_weekend <- NULL  
df\_final <- NULL  
df\_weekday <- data.frame(interval = unique(activity2\_weekday$interval), avg = as.numeric(mean\_activity2\_weekday), day = rep("weekday", length(mean\_activity2\_weekday)))  
df\_weekend <- data.frame(interval = unique(activity2\_weekend$interval), avg = as.numeric(mean\_activity2\_weekend), day = rep("weekend", length(mean\_activity2\_weekend)))  
df\_final <- rbind(df\_weekday, df\_weekend)  
  
xyplot(avg ~ interval | day, data = df\_final, layout = c(1, 2),   
 type = "l", ylab = "Number of steps")

