Assignment Reproducible research

rep\_data<-read.table("rep\_data.csv",header = TRUE,sep=",")  
rep\_data$date<-as.Date(rep\_data$date,format("%m/%d/%Y"))

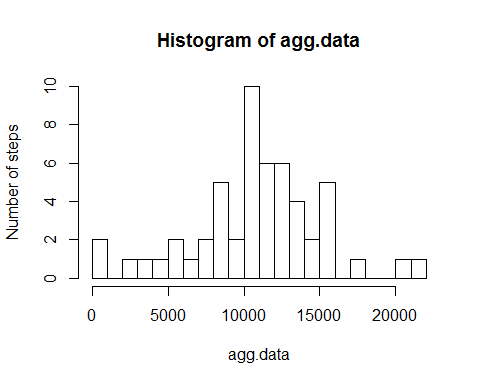
### Mean number of steps

We have to first clean the data and then aggregate for dates, this can be done using the function tapply

clean.data<-rep\_data[!is.na(rep\_data$steps),]  
agg.data<-tapply(clean.data[,1],clean.data[,2],sum)  
agg.data<-agg.data[!is.na(agg.data)]  
head(agg.data)

## 2012-10-02 2012-10-03 2012-10-04 2012-10-05 2012-10-06 2012-10-07   
## 126 11352 12116 13294 15420 11015

hist(agg.data,breaks=20,ylab="Number of steps")



mean(agg.data)

## [1] 10766.19

median(agg.data)

## [1] 10765

Figure 1 gives the histogram of the number of steps. We see that the mean is 10766.19 and the median is 10765 steps.

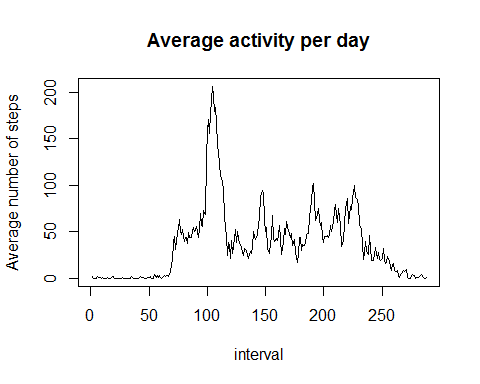
Another option would have been to use the by function, but this doesn't give an array.

agg.data.by<-by(clean.data[,1],clean.data$date,sum)  
head(agg.data.by[!is.na(agg.data.by)])

## clean.data$date  
## 2012-10-02 2012-10-03 2012-10-04 2012-10-05 2012-10-06 2012-10-07   
## 126 11352 12116 13294 15420 11015

### Average daily activity pattern

agg.data2<-tapply(clean.data[,1],clean.data[,3],mean)  
agg.data2<-agg.data2[!is.na(agg.data2)]  
plot(agg.data2,type="l",main="Average activity per day",xlab="interval",ylab="Average number of steps")



match(max(agg.data2),agg.data2)

## [1] 104

max(agg.data2)

## [1] 206.1698

As can be seen above, the maximum value is found for interval 104, the value is 206.1698. Figure 2 gives the time series of the average activity per day.

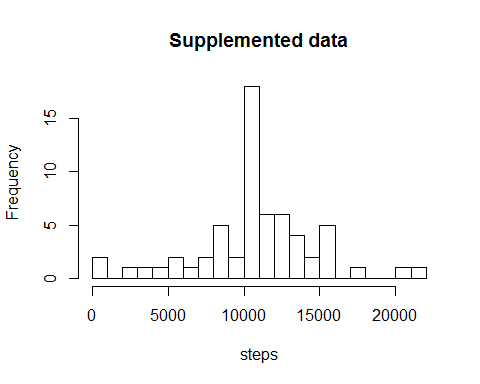
### Missing values

Let's first look, how many missing values there are. The function summary gives us that there are 2304 NA's in steps:

length(which(is.na(rep\_data$steps)))

## [1] 2304

NA\_index<-which(is.na(rep\_data$steps))  
  
rep\_data2<-rep\_data  
  
rep\_data2$steps[NA\_index] <- agg.data2[as.character(rep\_data2$interval[NA\_index])]  
agg.data.supplemented<-tapply(rep\_data2$steps,rep\_data2$date,sum)  
hist(agg.data.supplemented,20,xlab="steps",main="Supplemented data")



mean(agg.data.supplemented)

## [1] 10766.19

median(agg.data.supplemented)

## [1] 10766.19

mean(agg.data.supplemented)-mean(agg.data)

## [1] 0

We see that the mean and median are equal to 10766.19 steps. The difference between the supplemented and mean of the incomplete data is zero. Figure 3 gives a histogram of the supplemented data.

Filling in the missing values doesn't change the mean number of steps of the entire dataset as can be seen above.

### Comparison of weekdays and weekends

#Weekday selection  
dayVector<-weekdays(rep\_data2$date)  
fullVector=1:length(dayVector)  
weekendVector<-which(dayVector %in% c("Saturday","Sunday"))  
weekVector<-setdiff(fullVector,weekendVector)  
  
#Aggregation and plots  
rep\_data.week<-rep\_data2[weekVector,]  
rep\_data.weekend<-rep\_data2[weekendVector,]  
agg.data.week<-tapply(rep\_data.week[,1],rep\_data.week[,3],mean)  
agg.data.weekend<-tapply(rep\_data.weekend[,1],rep\_data.weekend[,3],mean)  
  
xweek<-as.numeric(names(agg.data.week))  
xweekend<-as.numeric(names(agg.data.weekend))  
  
par(mfrow=c(2,1))  
plot(xweek,agg.data.week,type="l",main="Activity week",xlab="interval",ylab="steps",col='red')  
plot(xweekend,agg.data.weekend,type="l",main="Activity weekends",xlab="interval",ylab="steps",col='blue')

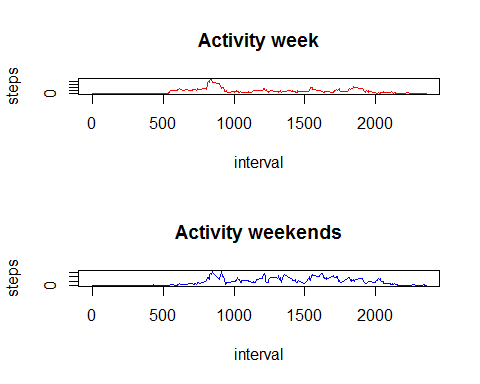


Figure 4 gives the time series of the weekly and weekend data.