Peer Assessment 1_Reproducible Research

Nizamo

Saturday, September 19, 2015

Reproducible Research - Peer Assessment 1

Loading and preprocessing the data

1. Load the data (i.e. read.csv())

```
file<-"E:/BigDataNizam/Module5/repdata-data-activity/activity.csv"

Data_activity<- read.csv(file, header=TRUE, sep=",")
head(Data_activity)
```

```
##   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
```

```
echo=TRUE
library(lattice)
```

2. Process/transform the data (if necessary) into a format suitable for your analysis

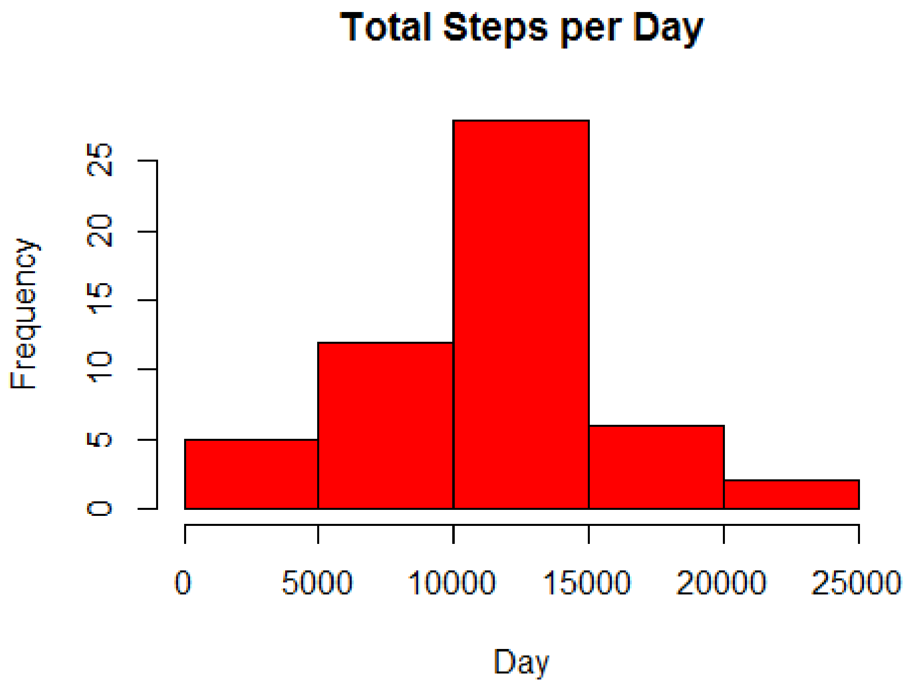
```
Data_activity$date <-as.Date(Data_activity$date,"%Y-%m-%d")
```

What is mean total number of steps taken per day?

```
TotalSteps <-aggregate(steps ~ date, data =Data_activity, sum, na.rm=TRUE)
```

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

```
hist(TotalSteps$steps, main ="Total Steps per Day", xlab="Day", col="red")
```



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

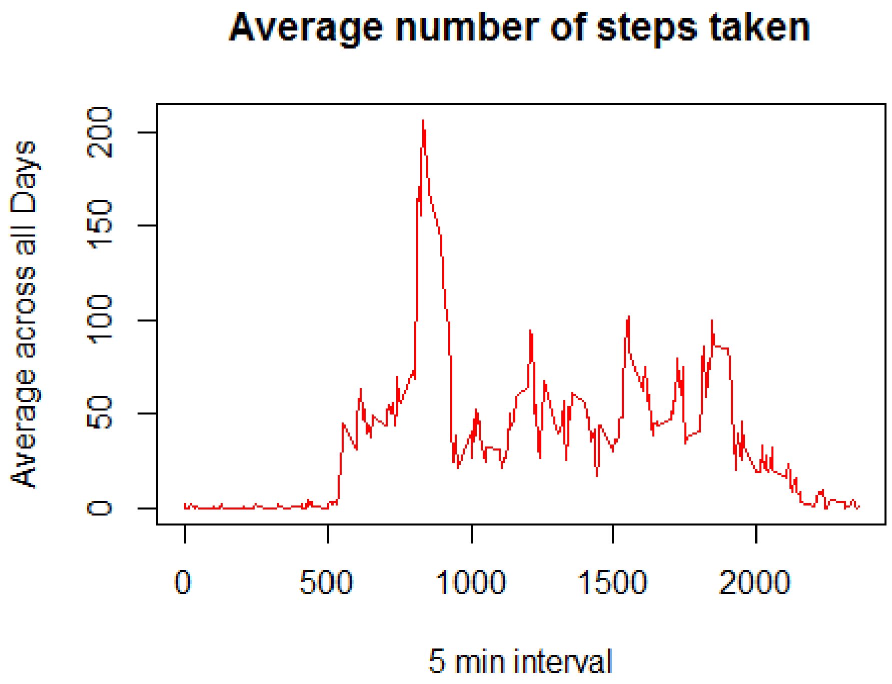
```
mean(TotalSteps$steps)

## [1] 10766.19

median(TotalSteps$steps)

## [1] 10765
```

What is the average daily activity pattern?

```
time_series<- tapply(Data_activity$steps, Data_activity$interval, mean,
na.rm=TRUE)
```

1. Make a time series plot

```
plot(row.names(time_series), time_series, type = "l", xlab= "5 min interval",
     ylab="Average across all Days", main= "Average number of steps taken",
     col="red")
```



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

```
maximum_interval<- which.max(time_series)
names(maximum_interval)

## [1] "835"
```

Imputing missing values

1. Calculate and report the total number of missing values in the dataset

```
activity_NA <- sum(is.na(Data_activity))
activity_NA

## [1] 2304
```

2. Devise a strategy for filling in all of the missing values in the dataset.

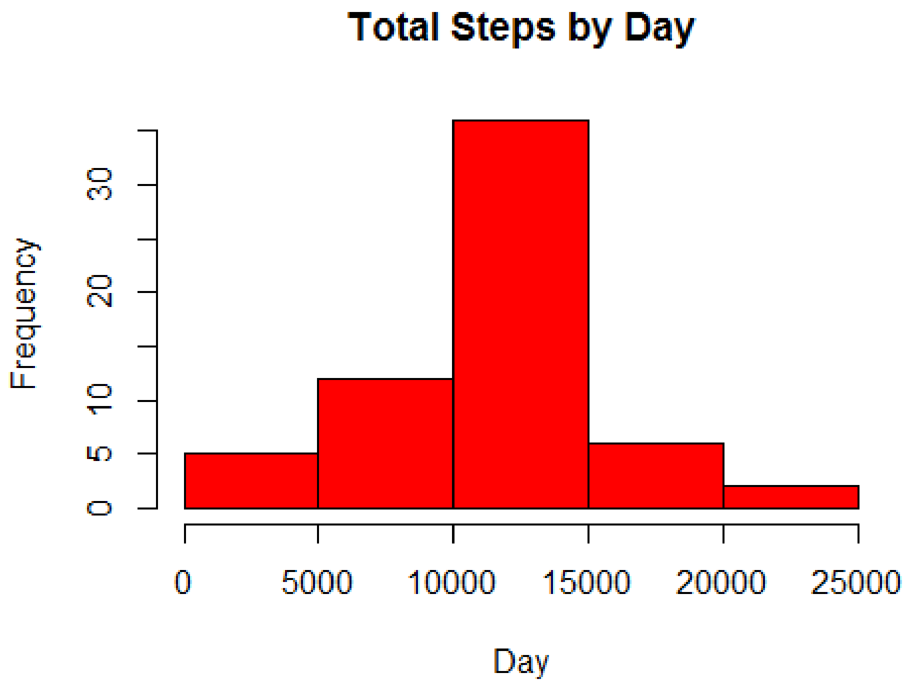
```
Steps_Average<- aggregate(steps ~ interval, data = Data_activity, FUN = mean)
FillNA<- numeric()
for (i in 1 :nrow(Data_activity)) {
  obs<- Data_activity[i,]
  if (is.na(obs$steps)){
    steps <-subset(Steps_Average, interval == obs$interval)$steps
  } else {
    steps <- obs$steps
  }
  FillNA <-c(FillNA, steps)
}
```

3. Create a new dataset that is equal to the original dataset but with the missing data filled in.

```
new_DataActivity <- Data_activity
new_DataActivity$steps <- FillNA
```

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

```
TotalSteps2 <- aggregate(steps ~ date, data = new_DataActivity, sum, na.rm=
TRUE)
hist(TotalSteps2$steps, main="Total Steps by Day", xlab= "Day", col="red")
```



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

```
mean(TotalSteps2$steps)

## [1] 10766.19

median(TotalSteps2$steps)

## [1] 10766.19
```

State the impact of imputing missing data to the estimates of the total daily number of steps?

The mean is still the same but the median have a little bit increase.

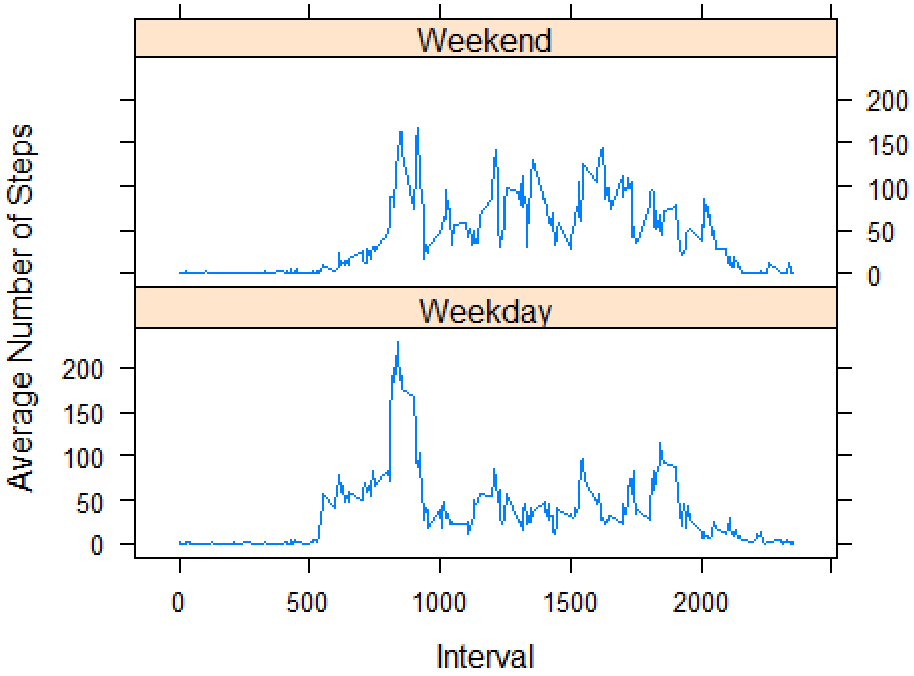
Are there differences in activity patterns between weekdays and weekends?

1. Create a new factor variable in the dataset with two levels -- "weekday" and "weekend" indicating whether a given date is a weekday or weekend day.

```
day <- weekdays(new_DataActivity$date)
daylevel <- vector()
for (i in 1: nrow(new_DataActivity)) {
  if (day[i] == "Saturday"){
    daylevel[i] <- "Weekend"
  } else if (day[i] == "Sunday") {
    daylevel[i] <-"Weekend"
  } else {
    daylevel[i]= "Weekday"
  }
}
new_DataActivity$daylevel <- daylevel
new_DataActivity$daylevel <- factor(new_DataActivity$daylevel)

StepsByDay <- aggregate (steps ~ interval + daylevel, data =
new_DataActivity, mean)
names(StepsByDay)<- c("interval", "daylevel", "steps")
```

2. Make a panel plot containing a time series plot

```
xyplot (steps ~ interval|daylevel, StepsByDay, type = "l",layout =c(1,2),
        xlab ="Interval", ylab ="Average Number of Steps")
```



From the time series plot, we can conclude that on weekend, there are more movement activities happened compare on weekday.

touch PA1_template