Reproducible Data

Anteneh

July 16, 2015

Loading and preprocessing the data

Downloading the data set to the working directory and make ready for preprocessing the data

```
activity_new <- read.csv("activity.csv", stringsAsFactors=FALSE)
names(activity_new)
```

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

```
str(activity_new)
```

```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : chr  "2012-10-01" "2012-10-01" "2012-10-01" "2012-10-01" ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
```

As we can see from the from the above summary, the dataset that the column containing the dates is not in a properly formatted way so we should adjust it as follow.

```
activity_new$date <- as.POSIXct(activity_new$date, format="%Y-%m-%d")
```

1. What is mean total number of steps taken per day?

And also there are missing values from the column of the steps. Since we have 2304 missing values, we have to exclude these missing values for our data analysis.Total number of missing values(NA’s)

```
sum(is.na(activity_new$steps))
```

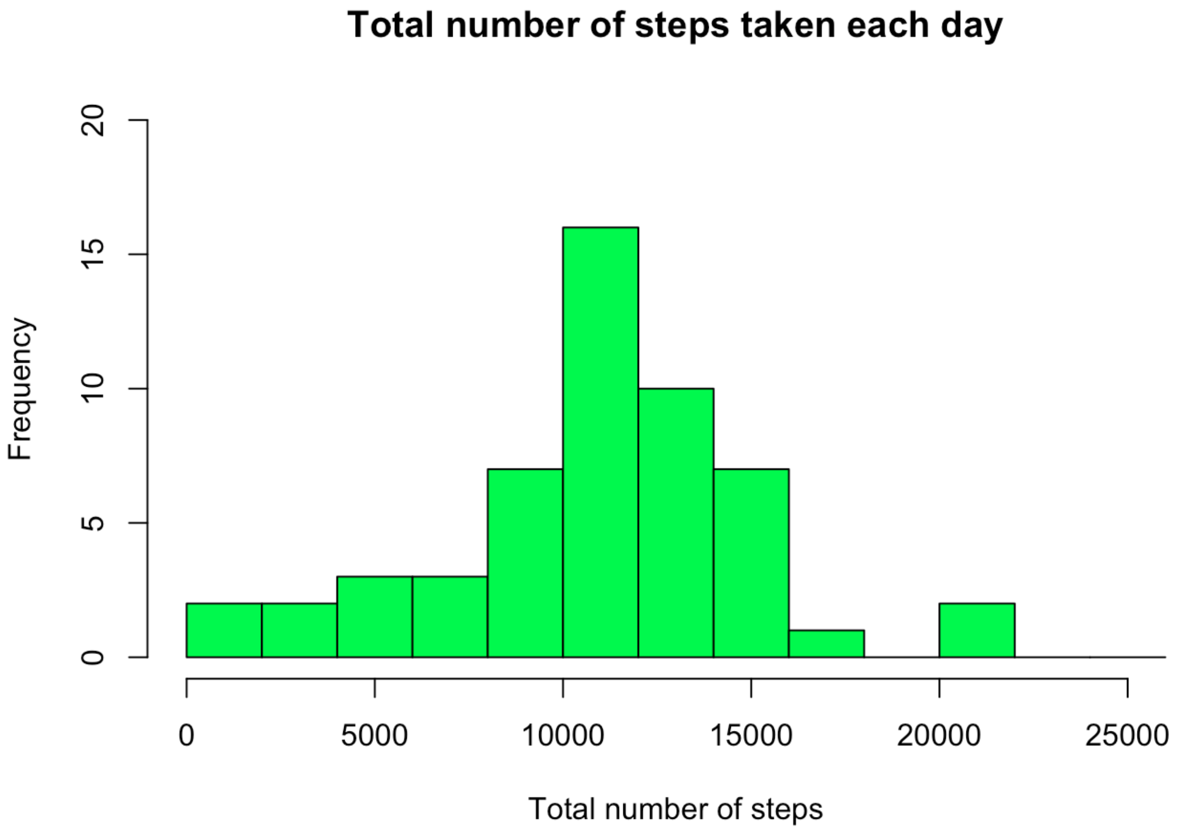
```
## [1] 2304
```

```
activity_noNA<-activity_new[which(!is.na(activity_new$steps)),]
```

```
activity_perday<-tapply(activity_noNA$steps, activity_noNA$date, sum)
```

Using the histogram we can see the mean and the median of number of steps per day.

```
hist(activity_perday,breaks=seq(from=0, to=26000, by=2000),
     col="Green",
     xlab="Total number of steps",
     ylim=c(0, 20),
     main="Total number of steps taken each day")
```



The mean and median of the total number of steps taken per a day without NA's are 10766.19 and 10765, respectively.

```
mean(activity_perday)
```

```
## [1] 10766.19
```

```
median(activity_perday)
```

```
## [1] 10765
```

2. What is the average daily activity pattern?

```
mean_daily_activity<-tapply(activity_noNA$steps, activity_noNA$interval,mean)
```

check

```
mean_data <- aggregate(activity_noNA$steps, by=list(activity_noNA$interval),
                       FUN=mean,na.rm=TRUE)
```

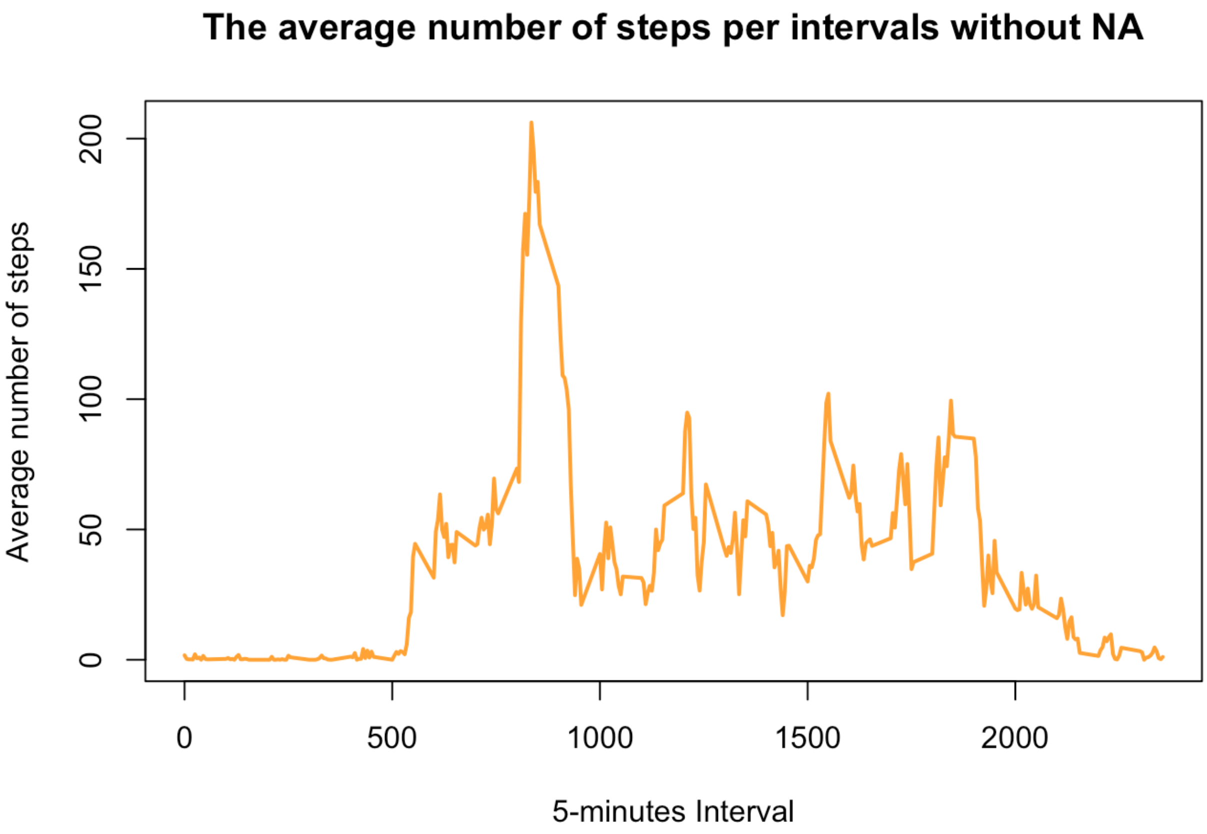
```
names(mean_data) <- c("interval", "mean")
```

```
head(mean_data)
```

```
##   interval      mean
## 1        0 1.7169811
## 2        5 0.3396226
## 3       10 0.1320755
## 4       15 0.1509434
## 5       20 0.0754717
## 6       25 2.0943396
```

Compute the time series plot

```
plot(mean_data$interval, mean_data$mean, type="l",col="orange",
     lwd=2,xlab="5-minutes Interval", ylab="Average number of steps",
     main="The average number of steps per intervals without NA")
```



The position of the maximum mean

```
mean_daily_activity[mean_daily_activity==max(mean_daily_activity)]
```

```
##      835
## 206.1698
```

The maximum average number of steps throughout the days is 835 with 206.1698 steps.

3. Filling the missing values

Filling the missing values by the mean/median. From above we have 2304 missing valuse but we don’t know the position of NA’s. Find the position and create a vector of means/medians.

```
#Finding the position of the NA's
```

```
NA_position <- which(is.na(activity_new$steps))
```

```
#Creating a vector
```

```
Missing_mean_vector<- rep(mean(activity_new$steps, na.rm=TRUE), times=length(NA
_position))
```

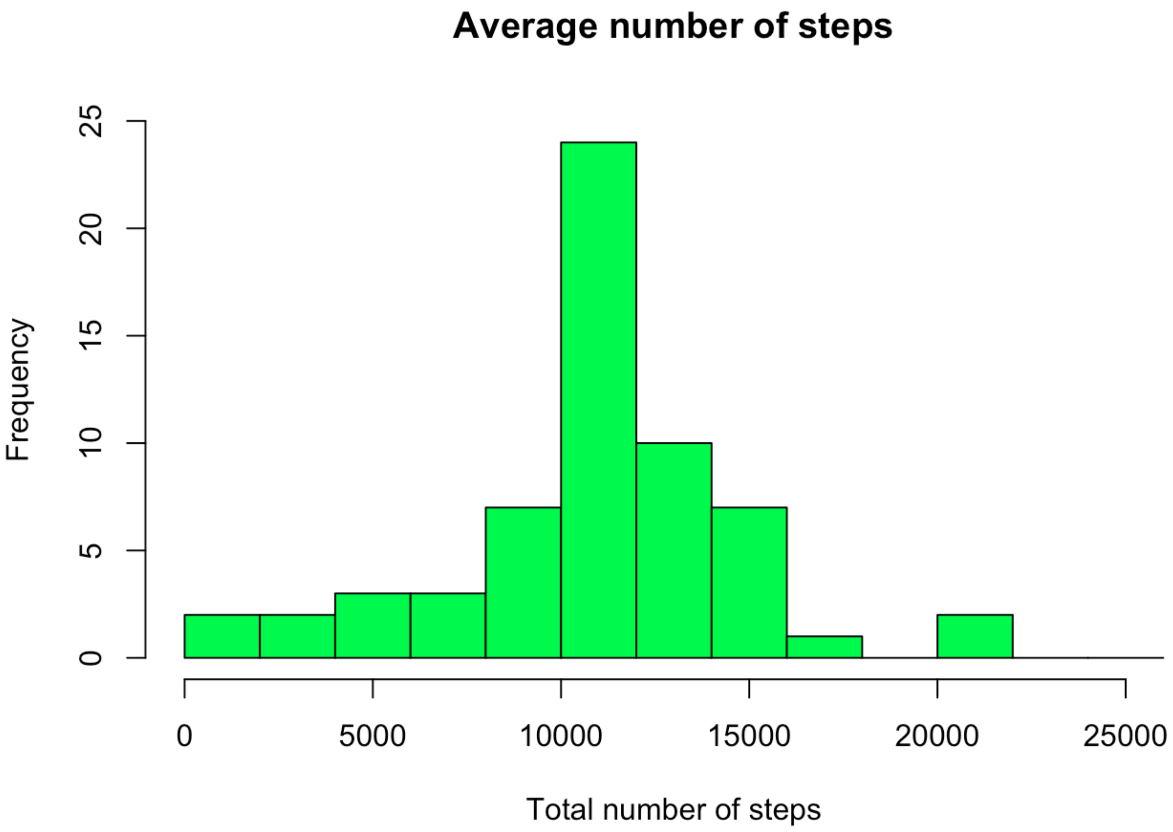
```
#Replacing NA's by mean
```

```
activity_new[NA_position, "steps"] <- Missing_mean_vector
```

Plot using Histogram for new activity data set where NA’s are replaced by mean.

```
activity_with_NA<-tapply(activity_new$steps, activity_new$date, sum)
```

```
hist(activity_with_NA,breaks=seq(from=0, to=26000, by=2000),col="Green",
     xlab="Total number of steps",ylim=c(0, 25),main="Average number of steps")
```



```
mean(activity_with_NA)
```

```
## [1] 10766.19
```

```
median(activity_with_NA)
```

```
## [1] 10766.19
```

The mean and median of the total number of steps taken per a day with NA’s are 10766 and 10766, respectively. The impact of the missing values on the mean is 0.19 where as the median is 1.0.