PA1_template.Rmd

Aurora González Vidal

01/17/2015

Loading and preprocessing the data

```r
df <- read.csv("activity.csv", sep = ",")
```

We can take a brief look at the data and create a new data frame ommiting NA:

```r
head(df)
```

```
##   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
```

```r
tail(df)
```

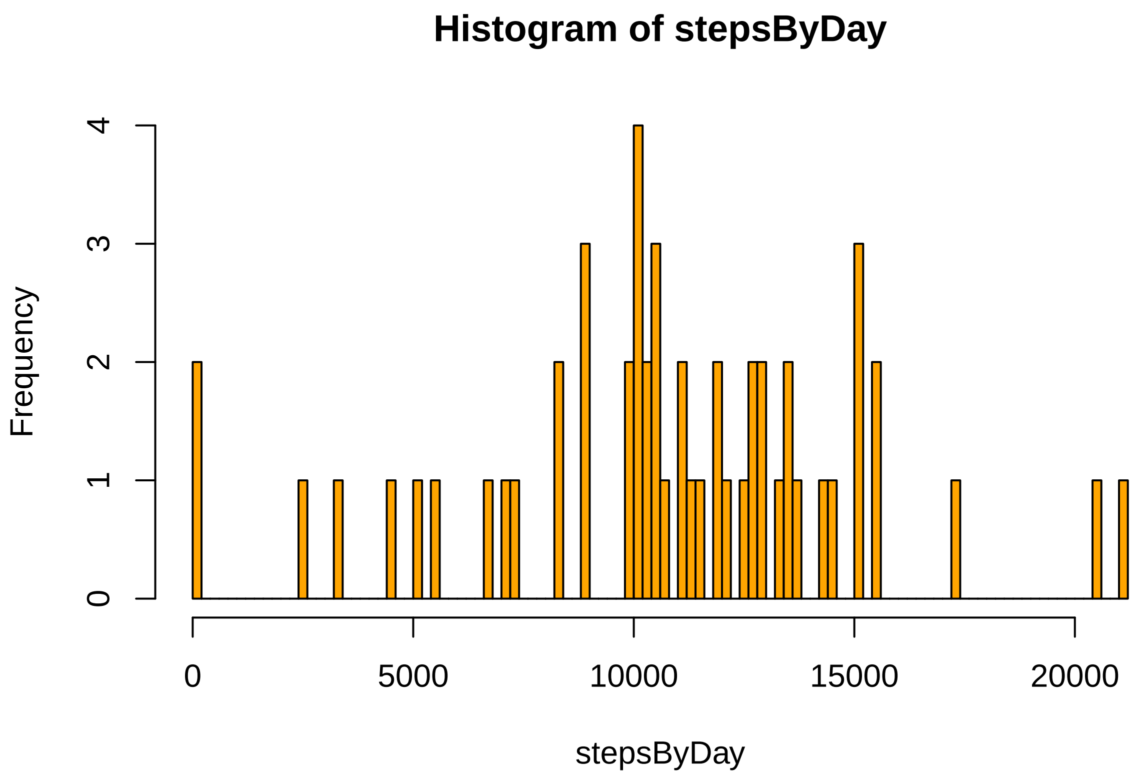
```
##       steps       date interval
## 17563    NA 2012-11-30     2330
## 17564    NA 2012-11-30     2335
## 17565    NA 2012-11-30     2340
## 17566    NA 2012-11-30     2345
## 17567    NA 2012-11-30     2350
## 17568    NA 2012-11-30     2355
```

```r
df1<-na.omit(df)
```

What is mean total number of steps taken per day?

Firt, we make a histogram of the total number of steps taken each day

```r
stepsByDay <- tapply(df1$steps, df1$date, sum, na.rm = T)
hist(stepsByDay, breaks=100, col = "orange")
```



The mean total number of steps taken per day is

```
m <- mean(stepsByDay, na.rm = T )
m
```

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

And the median total number of steps taken per day is

```
md <- median(stepsByDay, na.rm = T)
md
```

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

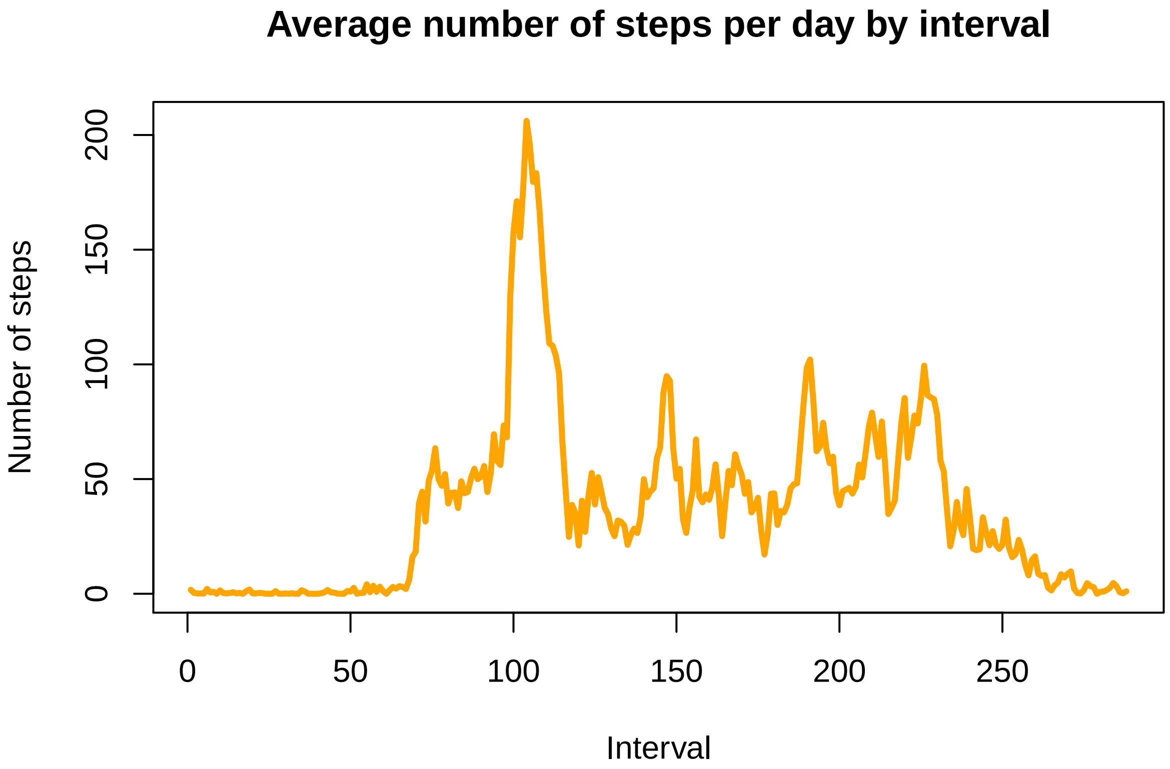
What is the average daily activity pattern?

We compute the average number of steps taken on each interval averaged across all days and save it into the object `stepsByInterval`.

```
stepsByInterval <- tapply(df$steps, df$interval, mean, na.rm =T)
```

And now we make a time series plot

```
plot(stepsByInterval, type="l", xlab="Interval", ylab="Number of steps",
     main="Average number of steps per day by interval", col = "orange", lwd = 3)
```



We find out that the interval which on average across all the days in the dataset contains the maximum number of steps is

```
max_interval <- stepsByInterval[which.max(stepsByInterval)]
max_interval
```

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

Imputing missing values

The total number of missing values in the dataset is

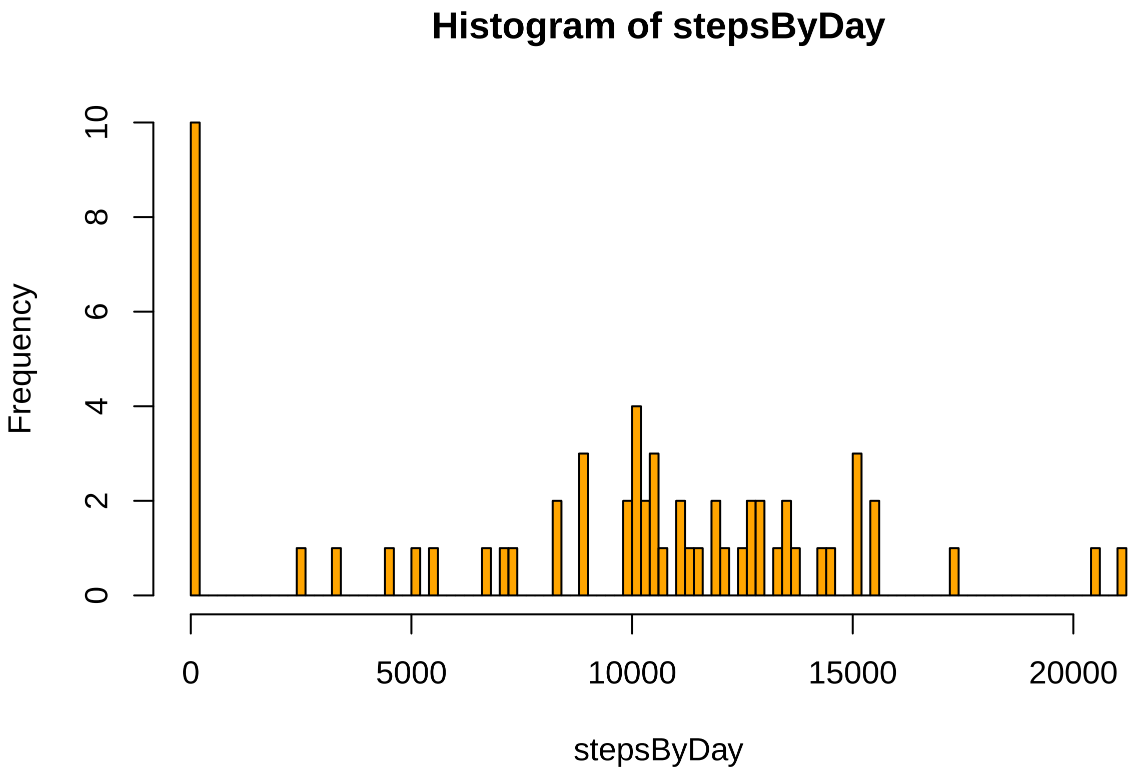
```
NAnumber <- sum(!complete.cases(df))
NAnumber
```

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

We substitute each missing value for the mean of steps of the interval that the missing value belongs and save it in a new data set named `df2`.

```
df2<-df
for (i in 1:length(df2)){
 if(is.na(df2$steps[i])){ #when we find a missing value
   df2$steps[i] <- mean(df2$steps[df$interval==df2$interval[i]], na.rm = T)
   #we substutite it by
}
}
```

```
stepsByDay <- tapply(df2$steps, df2$date, sum, na.rm = T)
hist(stepsByDay, breaks=100, col = "orange")
```



```
mean(stepsByDay, na.rm = T )
```

```
## [1] 9354.265
```

```
median(stepsByDay, na.rm = T)
```

```
## [1] 10395
```

Are there differences in activity patterns between weekdays and weekends?

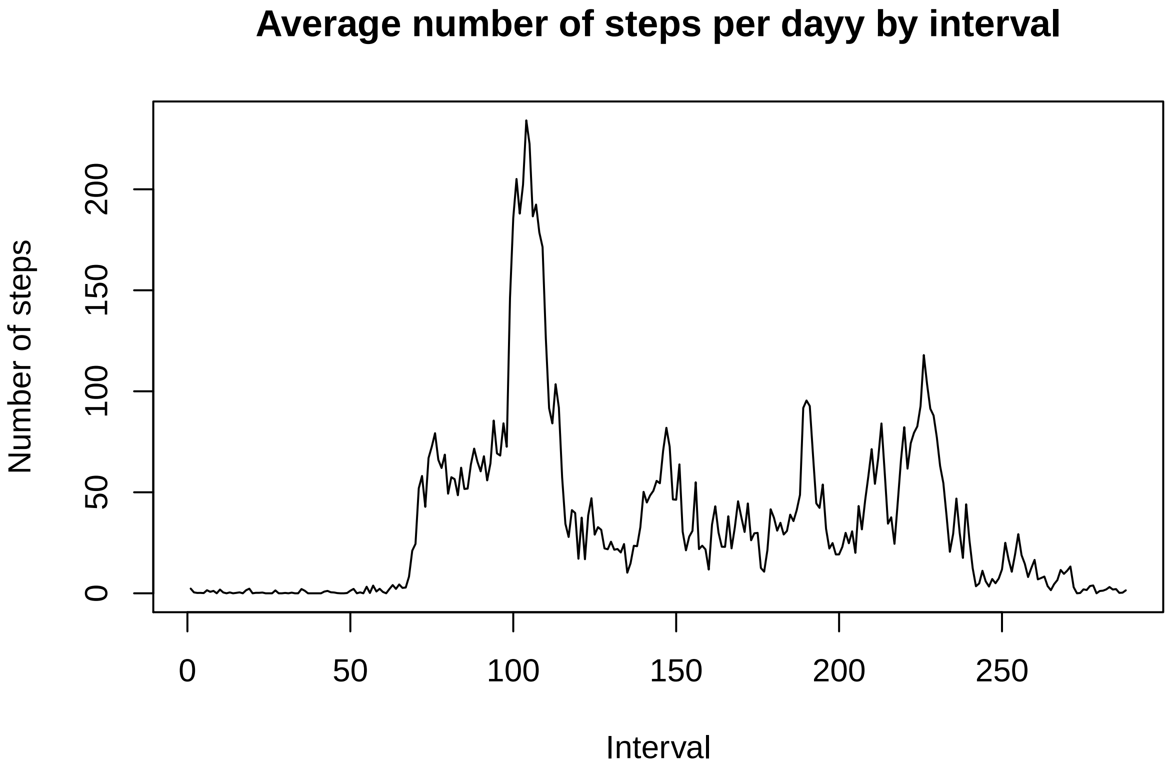
```
weekdays <- c("Monday", "Tuesday", "Wednesday", "Thursday",
              "Friday")
df2$nu = as.factor(ifelse(is.element(weekdays(as.Date(df2$date)),weekdays), "Weekday", "Weekend"))




df3 <- df2[df2$nu == "Weekday",]
df4 <-df2[df2$nu == "Weekend",]
```

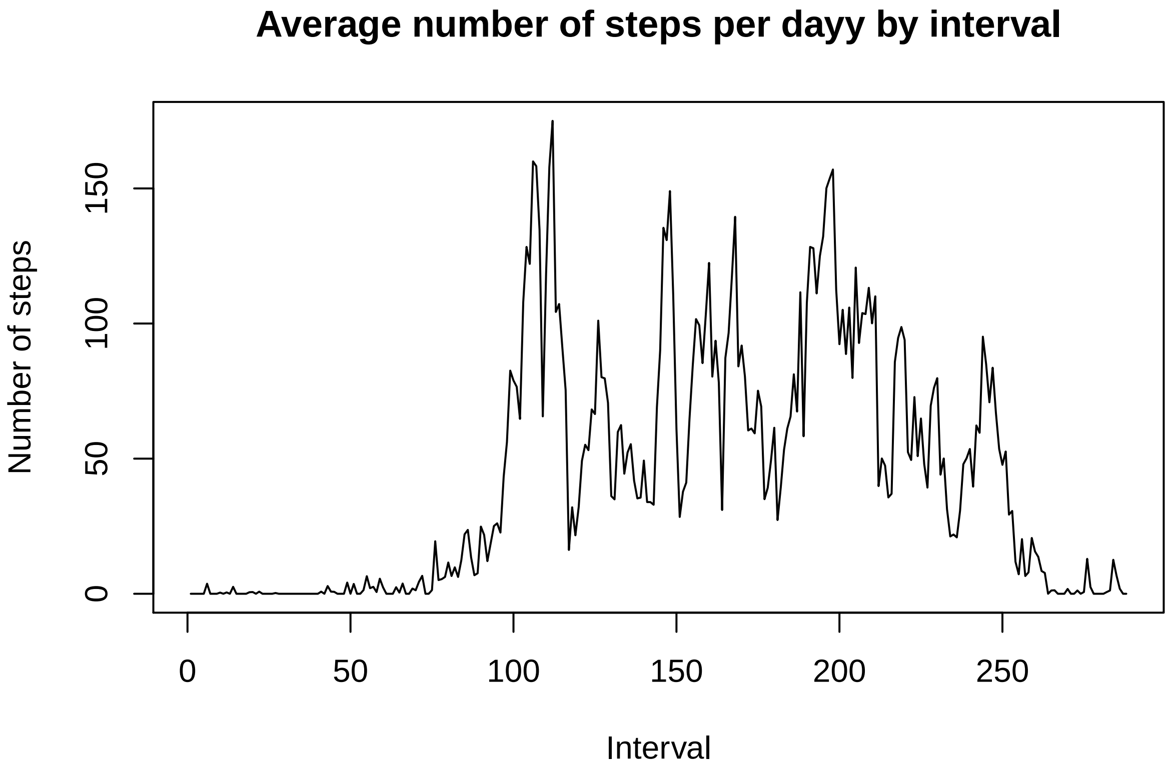
```
stepsByInterval3 <- tapply(df3$steps, df3$interval, mean, na.rm =T)

plot(stepsByInterval3, type="l", xlab="Interval", ylab="Number of steps",main="Average number of steps p
```



```
stepsByInterval4 <- tapply(df4$steps, df4$interval, mean, na.rm =T)

plot(stepsByInterval4, type="l", xlab="Interval", ylab="Number of steps",main="Average number of steps p
```


Average number of steps per dayy by interval
Number of steps
0
50
100
150
0
50
100
150
200
250
Interval