Reproducible week 2

Anamaria Ahuis

11 Februar 2017

Reading the data from the provided source

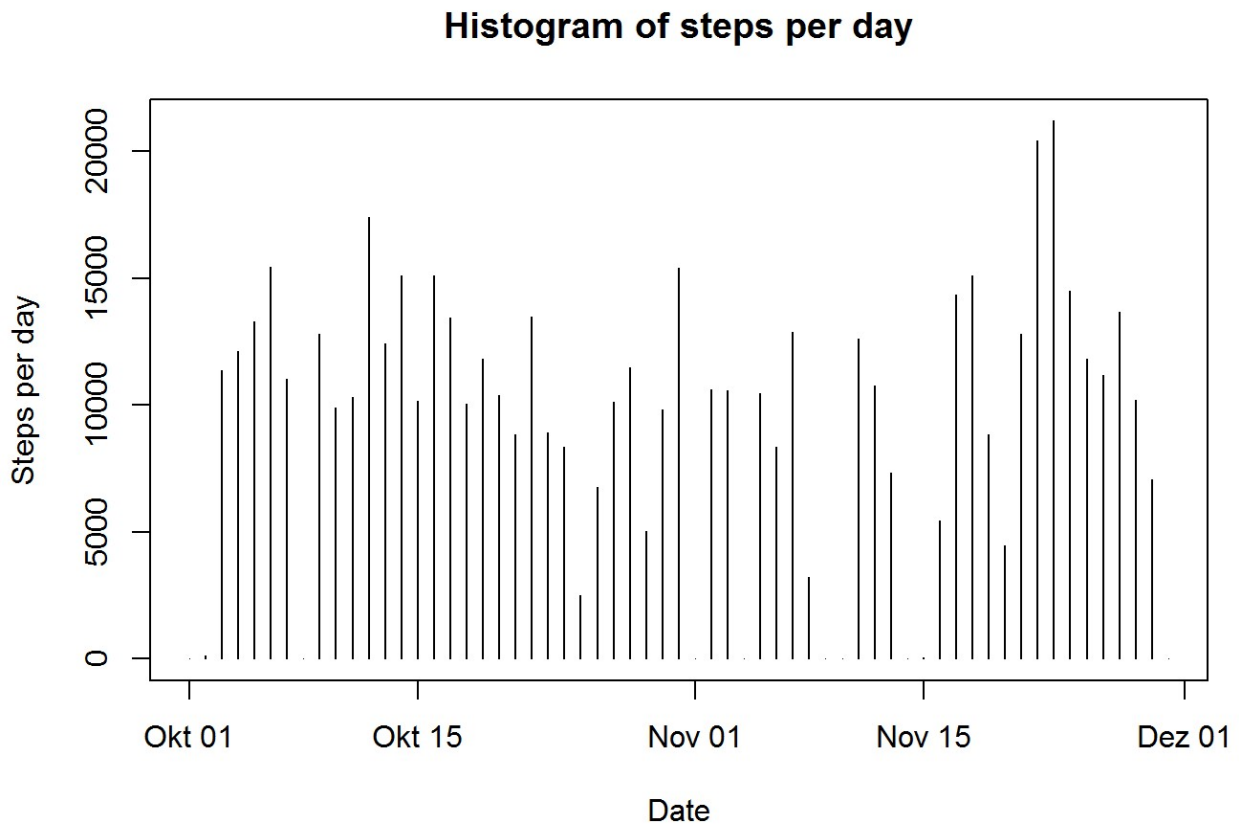
```
df <- read.csv("activity.csv")
df$date <- as.Date(df$date, format="%Y-%m-%d")
dim(df)
```

```
## [1] 17568     3
```

After reading the data from the file in the dataframe “df”, we notice that this has 17568 observations (rows) and 3 variables (columns).

Histogram of the total number of steps taken each day

```
#compute the number of steps per day
stepspd <- aggregate(df$steps, list(df$date), sum, na.rm=TRUE)
#change the names of the aggregated data frame to something meaningful
colnames(stepspd) <- c("day", "steps")
#plot the histogram
plot(stepspd$day, stepspd$steps, type ="h", xlab="Date", ylab="Steps per day",
main="Histogram of steps per day")
```



By looking at the histogram we notice that there are some “dips” in the data, corresponding to very few steps during those days. For the time being, there is so far no explanation for this.

Mean and median number of steps taken each day

```
print(paste0("Mean number of steps per day: ", mean(stepspd$steps)))
```

```
## [1] "Mean number of steps per day: 9354.22950819672"
```

```
print(paste0("Median number of steps per day: ", median(stepspd$steps)))
```

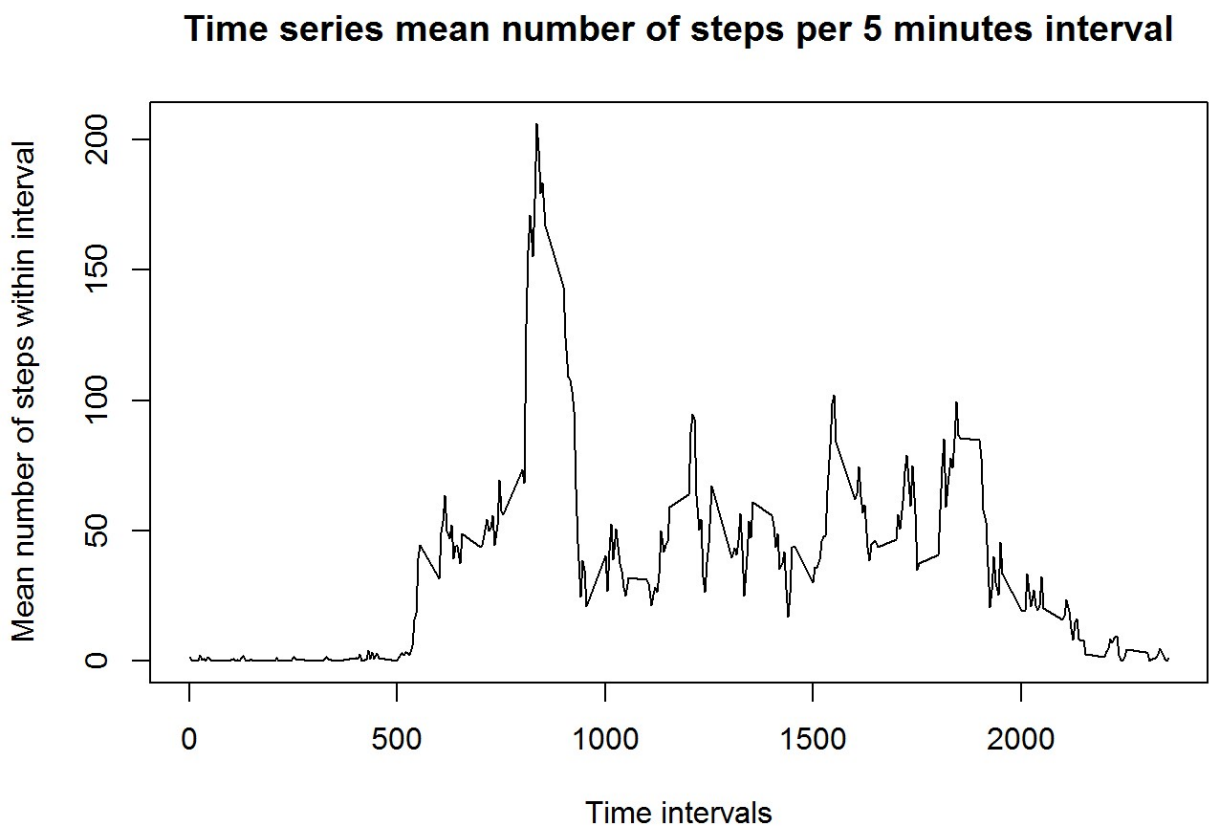
```
## [1] "Median number of steps per day: 10395"
```

[1] “Mean number of steps per day: 9354.22950819672” [1] “Median number of steps per day: 10395”

Time series plot of the average number of steps taken

Compute the average number of steps taken, averaged across all days (y-axis) on each interval

```
stepspi <- aggregate(df$steps, list(df$interval), mean, na.rm=TRUE)
colnames(stepspi) <- c("intv", "steps")
plot(stepspi$intv, stepspi$steps, type="l", xlab="Time intervals", ylab="Mean n
umber of steps within interval", main="Time series mean number of steps per 5 m
inutes interval")
```



A peak in the data is visible, as well as variations during the day. From the variations we could deduct that this person is moving more at regular intervals, e.g. like walking every couple of hours, and that there is a peak in the activity, perhaps the time when the person (more often) exercises intesively. Let's take a look when does this person (on average) makes sport:

```
print(paste0("The interval with greatest steps average is: ", stepspi[stepspi$s
teps==max(stepspi$steps),]$intv))
```

```
## [1] "The interval with greatest steps average is: 835"
```

[1] "The interval with greatest steps average is: 835"

We notice that the interval 835 roughly corresponds to the peak in the plot above.

Strategy for imputing missing data

We will first investigate the total number of missing values in the dataset:

```
print(paste0("There are ",sum(is.na(df$steps))," NA values in the dataset."))
```

```
## [1] "There are 2304 NA values in the dataset."
```

[1] “There are 2304 NA values in the dataset.”

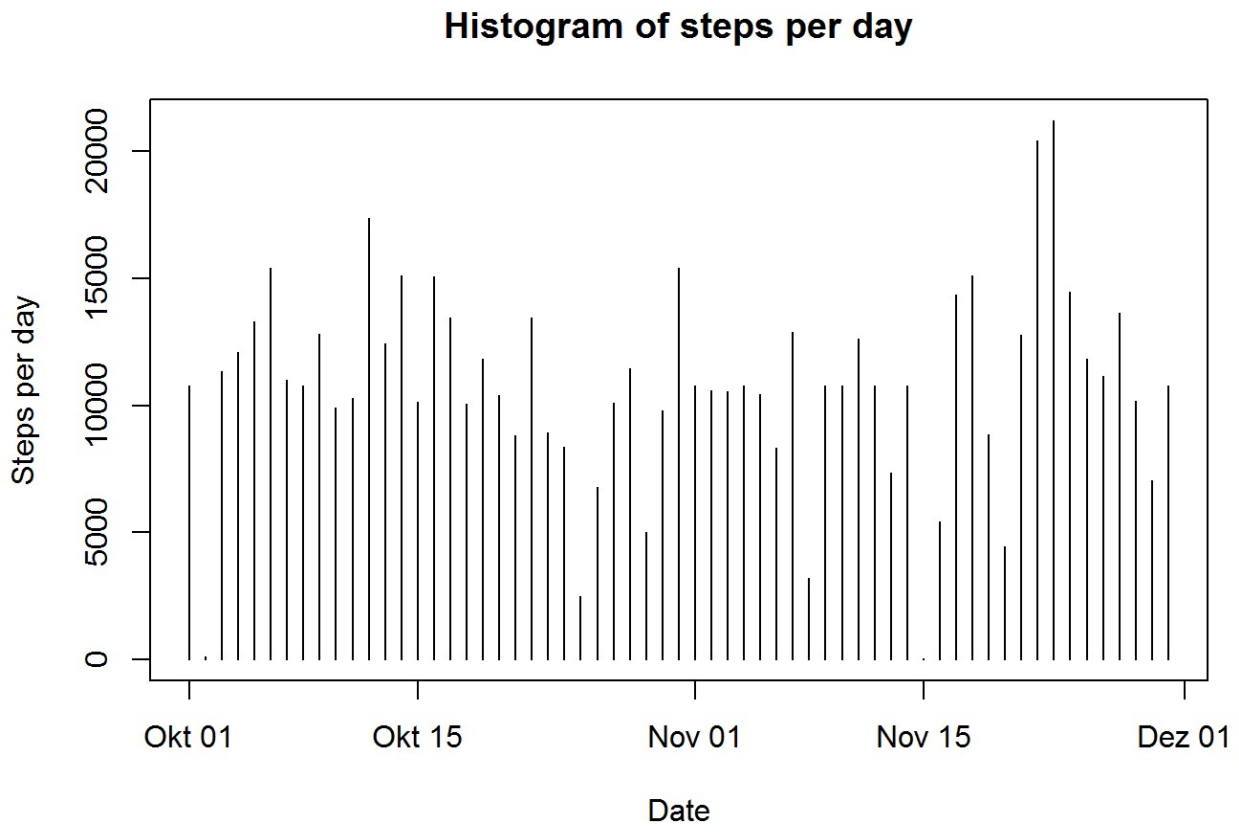
We will fill the missing values in the dataset with the mean for the corresponding interval:

```
dforig <- df
for (i in 1:dim(df)[1]) {
    if (is.na(df[i,]$steps)) df[i,]$steps <- stepspi[stepspi$intv==df[i,]$inter
val,]$steps
}
dfnew <- df
df <- dforig
rm("dforig")
```

The original data set, containing NAs, is in the df dataframe, and the dfnew dataframe contains the new data set with no NA values for the steps variable.

Histogram of the total number of steps taken each day

```
stepspd2 <- aggregate(dfnew$steps, list(dfnew$date), sum, na.rm=TRUE)
colnames(stepspd2) <- c("day", "steps")
plot(stepspd2$day, stepspd2$steps, type ="h", xlab="Date", ylab="Steps per da
y", main="Histogram of steps per day")
```



We observe that there is generally more uniform activity - suggesting perhaps that there were lots of NA values for some days (what could cause that?) - but that there still are some days with few steps.

The mean and median total number of steps taken per day:

```
print(paste0("Mean number of steps per day, no NA: ", mean(stepspd2$steps)))
```

```
## [1] "Mean number of steps per day, no NA: 10766.1886792453"
```

```
print(paste0("Median number of steps per day, no NA: ", median(stepspd2$step
s)))
```

```
## [1] "Median number of steps per day, no NA: 10766.1886792453"
```

[1] “Mean number of steps per day, no NA: 10766.1886792453” [1] “Median number of steps per day, no NA: 10766.1886792453”

The mean is with almost 2000 extra steps bigger than the values computed with NA present. The NA values have a rather limited effect upon the median values.

Comparison of weekdays and weekends' values

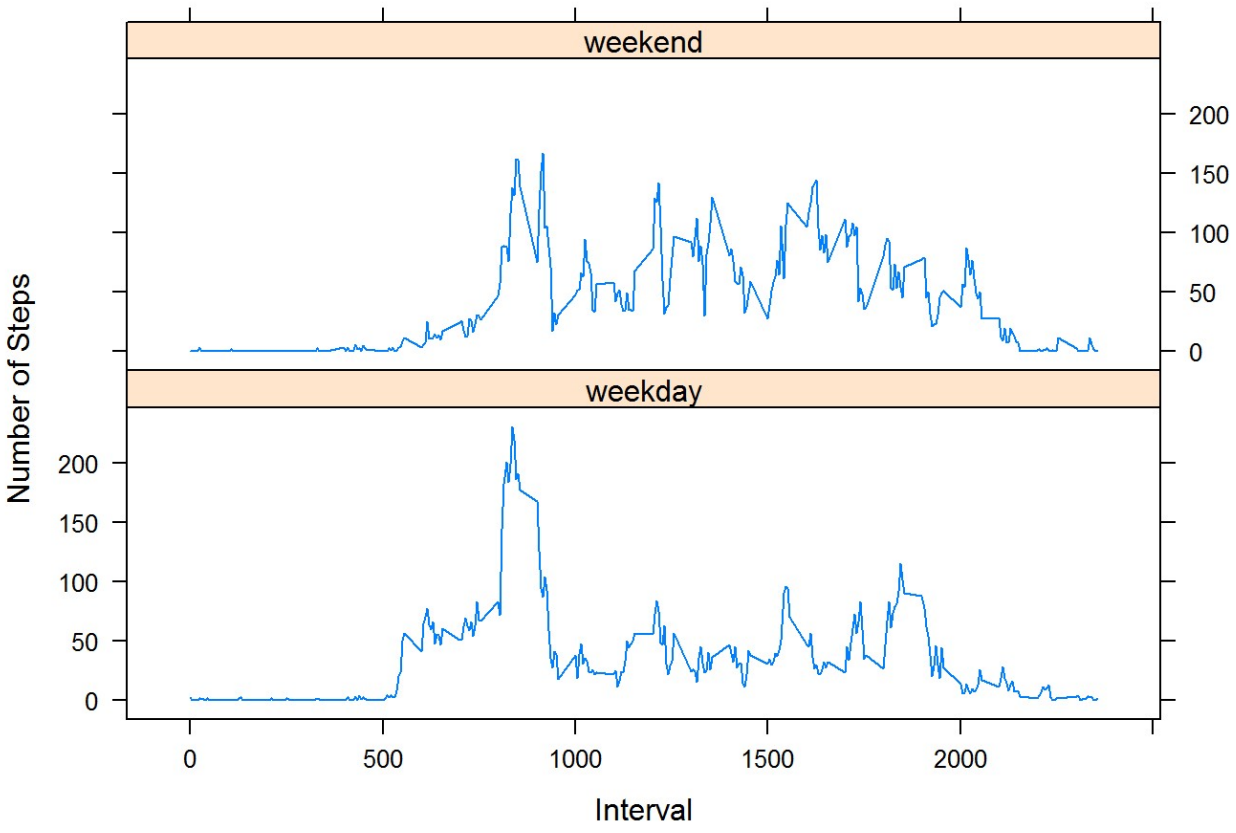
We first create a new factor variable in the dataset with two levels - "weekday" and "weekend" indicating whether a given date is a weekday or weekend day.

```
dfnew$pos <- "weekday"
dfnew[weekdays(dfnew$date) %in% c("Samstag","Sonntag"),]$pos <- "weekend"
dfnew$pos <- as.factor(dfnew$pos)
```

Panel plot containing a time series plot (i.e. type = "l") of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all weekday days or weekend days (y-axis).

```
library(lattice)
stepspp <- aggregate(dfnew$steps, list(dfnew$pos, dfnew$interval), mean)
colnames(stepspp) <- c("Dow","Interval","avgsteps")

xyplot(data=stepspp,avgsteps~Interval|factor(Dow),
       type='l',layout=c(1,2),
       xlab='Interval',ylab='Number of Steps')
```



By examining these plots, we can tell that this person is generally more active during the week, perhaps walking to, from or at work. The peak of activity is during both weekend and weekdays at the same time in the day. The person is generally more active after this intensity peak, for the rest of the day.