Contents

1 Reproducible Research: Peer Assessment 1

by Melinda Higgins dated 07/09/2014

1.1 Loading and preprocessing the data

```
# The following code assumes a relative path where the data are located in a subdirectory called "/data,
activityData <- read.csv("data/activity.csv", header=TRUE)
```

1.1.1 Summary of the dataset

The data contain 3 columns of data which are " steps, date, interval “.

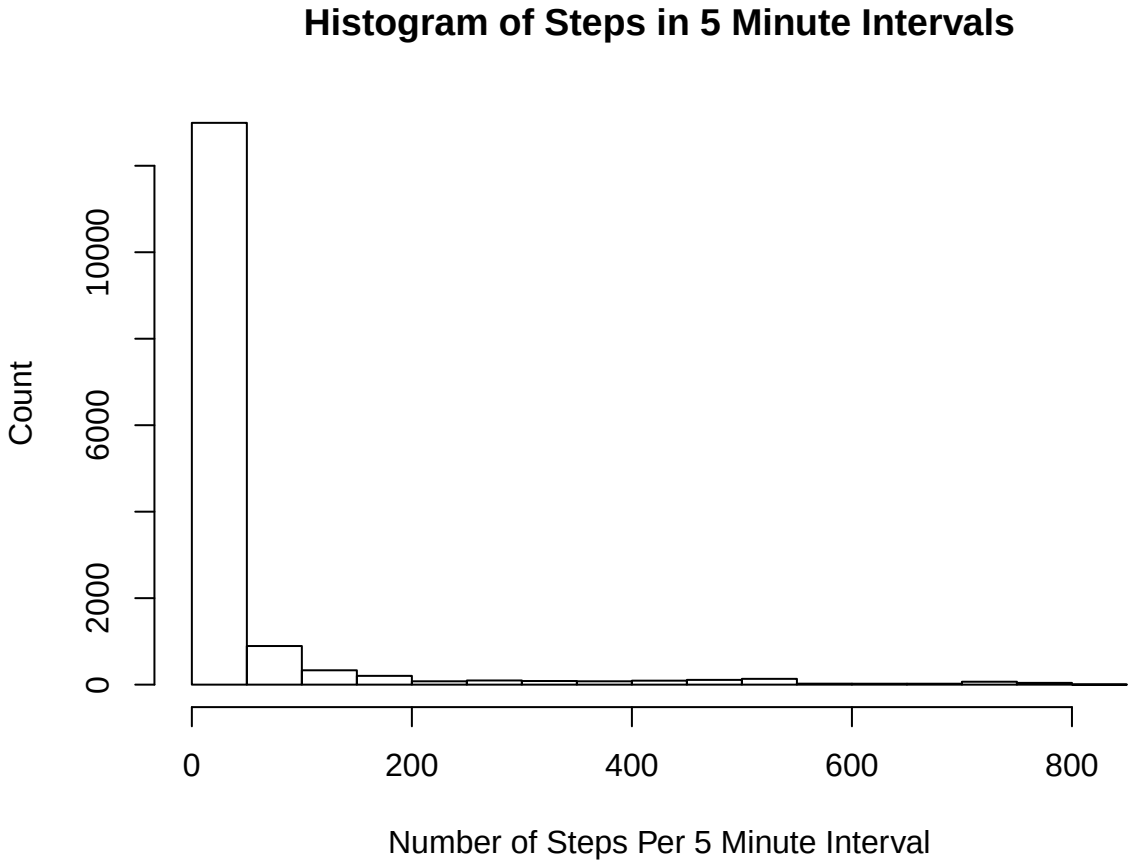
There are 17568 rows of data covering 61 days. Each day the number of steps were recorded every 5 minutes during 24 hours for 288 total intervals per day.

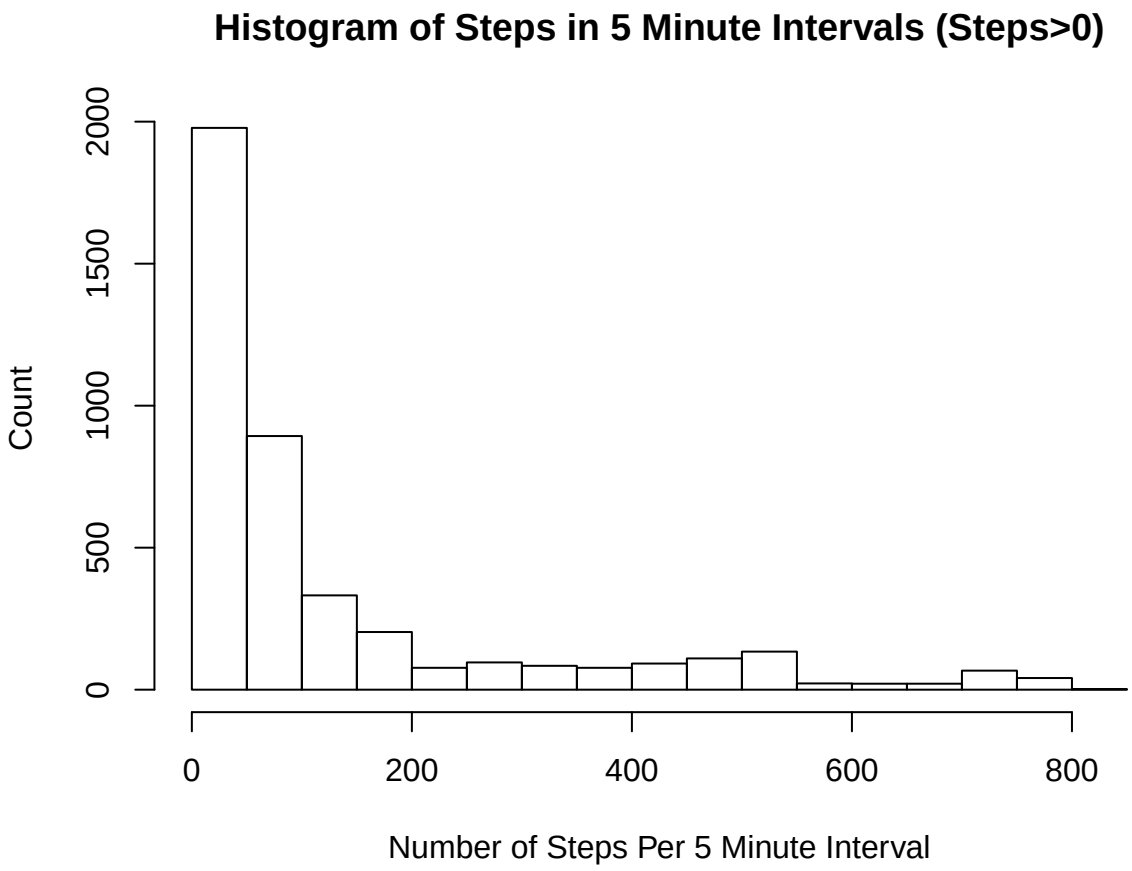
1.1.2 Missing data, Zeros, and Potential Extreme Values (Potential Outliers)

- There are 2304 NA’s (missing values);
- There are 11014 zeros (0’s);
- and 4250 entries with non-zero steps ranging from 1 to 806 steps in the 5 minute intervals.

Some of these high values seem extreme - possibly too many steps recorded in a 5 minute interval. For example, the maximum number of steps of 806 in 5 minutes implies making 2.6867 steps every second.

1.1.3 Histogram of the Number of Steps Per 5 minute Interval (ignoring missing data)





1.2 What is mean total number of steps taken per day?

Calculate the overall mean and the mean for each day. Also run for the medians per day and overall - with an without the "imputation" method tried below...

1.3 What is the average daily activity pattern?

Look at the raw data - scatterplot by time of day (the increments) - look at overall and by day of the week - also look at a table summarizing these data given the n, mean, median, maybe Q1, Q3, min and max - over all 61 days by time of day and again by day of week. make some plots to go along with...

```
library(chron)
library(lattice)

activitydate <- chron(as.character(activityData$date),format=c(dates="y-m-d"),out.format=c("day months y

activityData2 <- activityData
activityData2$date2 <- activitydate
activityData2$weekday <- weekdays(activityData2$date2)

library(xtable)
```

```
testdataf <- activityData2[1:30,]
testdataf.xtable <- xtable(activityData2[1:30,])
print(testdataf.xtable, floating=FALSE)
```

```
## Warning: class of 'x' was discarded
```

```
## % latex table generated in R 3.1.0 by xtable 1.7-3 package
## % Sat Jul 12 11:52:05 2014
## \begin{tabular}{rrlrrl}
##   \hline
##  & steps & date & interval & date2 & weekday \\
##   \hline
## 1 &  & 2012-10-01 &   0 & 15614.00 & Mon \\
##   2 &  & 2012-10-01 &   5 & 15614.00 & Mon \\
##   3 &  & 2012-10-01 &  10 & 15614.00 & Mon \\
##   4 &  & 2012-10-01 &  15 & 15614.00 & Mon \\
##   5 &  & 2012-10-01 &  20 & 15614.00 & Mon \\
##   6 &  & 2012-10-01 &  25 & 15614.00 & Mon \\
##   7 &  & 2012-10-01 &  30 & 15614.00 & Mon \\
##   8 &  & 2012-10-01 &  35 & 15614.00 & Mon \\
##   9 &  & 2012-10-01 &  40 & 15614.00 & Mon \\
##   10 &  & 2012-10-01 &  45 & 15614.00 & Mon \\
##   11 &  & 2012-10-01 &  50 & 15614.00 & Mon \\
##   12 &  & 2012-10-01 &  55 & 15614.00 & Mon \\
##   13 &  & 2012-10-01 & 100 & 15614.00 & Mon \\
##   14 &  & 2012-10-01 & 105 & 15614.00 & Mon \\
##   15 &  & 2012-10-01 & 110 & 15614.00 & Mon \\
##   16 &  & 2012-10-01 & 115 & 15614.00 & Mon \\
##   17 &  & 2012-10-01 & 120 & 15614.00 & Mon \\
##   18 &  & 2012-10-01 & 125 & 15614.00 & Mon \\
##   19 &  & 2012-10-01 & 130 & 15614.00 & Mon \\
##   20 &  & 2012-10-01 & 135 & 15614.00 & Mon \\
##   21 &  & 2012-10-01 & 140 & 15614.00 & Mon \\
##   22 &  & 2012-10-01 & 145 & 15614.00 & Mon \\
##   23 &  & 2012-10-01 & 150 & 15614.00 & Mon \\
##   24 &  & 2012-10-01 & 155 & 15614.00 & Mon \\
##   25 &  & 2012-10-01 & 200 & 15614.00 & Mon \\
##   26 &  & 2012-10-01 & 205 & 15614.00 & Mon \\
##   27 &  & 2012-10-01 & 210 & 15614.00 & Mon \\
##   28 &  & 2012-10-01 & 215 & 15614.00 & Mon \\
##   29 &  & 2012-10-01 & 220 & 15614.00 & Mon \\
##   30 &  & 2012-10-01 & 225 & 15614.00 & Mon \\
##    \hline
## \end{tabular}
```

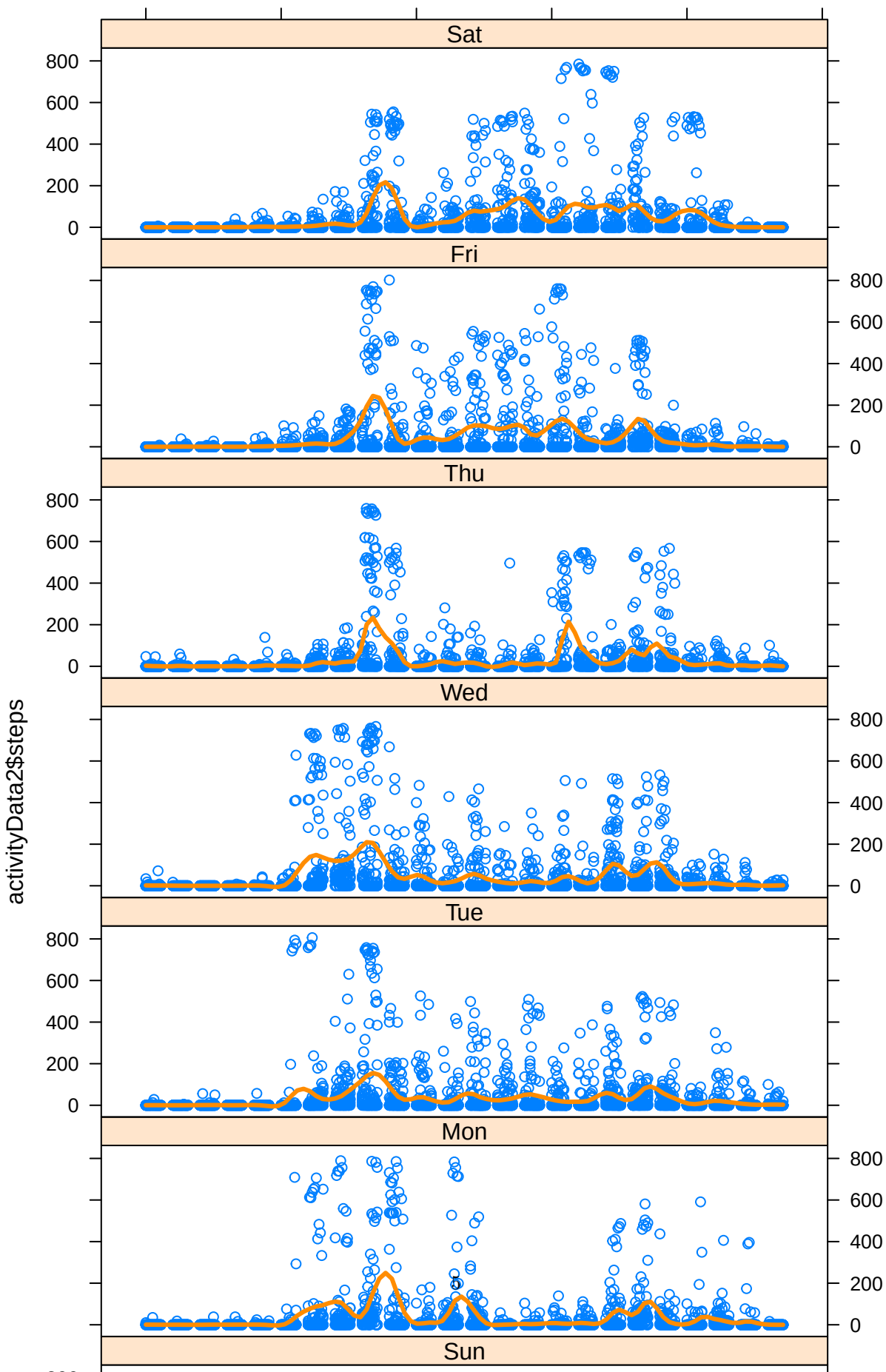
```
xyplot(activityData2$steps ~ activityData2$interval | factor(activityData2$weekday), type=c("p","spline
```

Estimate

Std. Error

t value

Pr(>|t|)


Sat
Fri
Thu
Wed
Tue
Mon
Sun
activityData2$steps
0
200
400
600
800
5

(Intercept)

29.5539

1.7868

16.54

0.0000

activityData2$interval

0.0066

0.0013

5.08

0.0000

1.3.1 Summary Statistics for Number of Steps in 5 Minute Intervals

sumsteps

0

3

12

1

37.4

1

806

1

2304

1

2 ### Summary Statistics for Number of Steps (in 5 min interval)

Statistic	Value
Mean	37.38
Median	0
SD	112
Min	0
Max	806

2.1 Imputing missing values

Need to consider why the data are "missing" (NA's). These are most likely due to inactivity, although zeros were also recorded during other entries. Perhaps the NAs are due to the monitor being turned off or other reason no data was recorded.

If we assume that the NA's are due to inactivity then substituting 0's for NA would be appropriate. If this cannot be confirmed, a smoothing approach could be used taking a simple average of the reading before and after the NA is noted. Although if the reading before and after are both NAs then perhaps a zero is best.

However, the number of steps are highly right-skewed with obvious zero-inflation. This distribution indicates that using the mean to substitute for the NAs is not appropriate. Instead using the median would be better - although these are essentially zero for every day as well - again suggesting that substituting zero's for the NAs would be appropriate.

It is noted that without a proper explanation for the source of the NAs, any method used for substitution will be biased. To what extent the chosen method is biased is unverifiable without further information.

2.2 Are there differences in activity patterns between weekdays and weekends?

Need to figure out what function is needed to extract day of the week from the provided dates. Also need to check the date formatting...