Project 1

lucy

April 5, 2016

Loading & Preprocessing Data

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.4
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.2.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
url <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip"
download.file(url, destfile="./data.zip", method="curl")
unzip("data.zip")
activity <- read.csv("activity.csv",quote="\"")
activity$date<-as.POSIXct(activity$date, "%Y-%M-%D")
```

```
## Warning in strptime(xx, f <- "%Y-%m-%d %H:%M:%OS", tz = tz): unknown
## timezone '%Y-%M-%D'
```

```
## Warning in as.POSIXct.POSIXlt(x): unknown timezone '%Y-%M-%D'
```

```
## Warning in strptime(xx, f <- "%Y/%m/%d %H:%M:%OS", tz = tz): unknown
## timezone '%Y-%M-%D'
```

```
## Warning in as.POSIXct.POSIXlt(x): unknown timezone '%Y-%M-%D'
```

```
## Warning in strptime(xx, f <- "%Y-%m-%d %H:%M", tz = tz): unknown timezone
## '%Y-%M-%D'
```

```
## Warning in as.POSIXct.POSIXlt(x): unknown timezone '%Y-%M-%D'
```

```
## Warning in strptime(xx, f <- "%Y/%m/%d %H:%M", tz = tz): unknown timezone
## '%Y-%M-%D'
```

```
## Warning in as.POSIXct.POSIXlt(x): unknown timezone '%Y-%M-%D'
```

```
## Warning in strptime(xx, f <- "%Y-%m-%d", tz = tz): unknown timezone '%Y-%M-
## %D'
```

```
## Warning in as.POSIXct.POSIXlt(x): unknown timezone '%Y-%M-%D'
```

```
## Warning in strptime(x, f, tz = tz): unknown timezone '%Y-%M-%D'
```

```
## Warning in as.POSIXct.POSIXlt(as.POSIXlt(x, tz, ...), tz, ...): unknown
## timezone '%Y-%M-%D'
```

```
activity$steps<-as.numeric(activity$steps)
head(activity)
```

```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone '%Y-%M-%D'
```

```
##   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
```

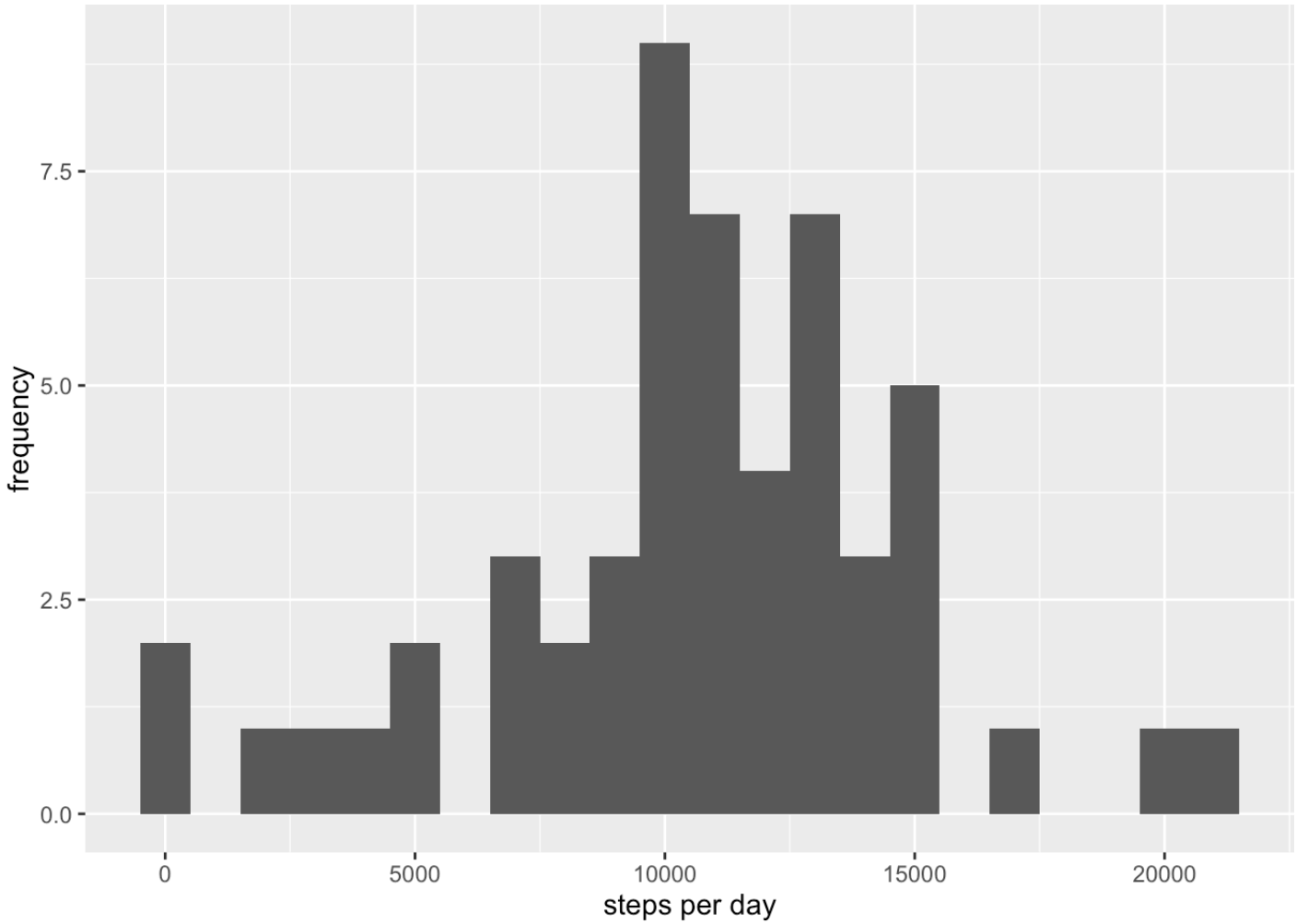
##What is mean total number of steps taken per day?

```
totalstep<-aggregate(activity$steps, na.rm=T, by=list(activity$date), FUN=sum)
```

```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone '%Y-%M-%D'

## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone '%Y-%M-%D'
```

```
totalstep<-subset(totalstep, totalstep$x!="0")
g<-ggplot(totalstep, aes(x=x))
g+geom_histogram(binwidth=1000)+xlab("steps per day")+ylab("frequency")
```




```
mean_steps <- mean(totalstep$x, na.rm = TRUE)
median_steps<-median(totalstep$x, na.rm=TRUE)
mean_steps
```

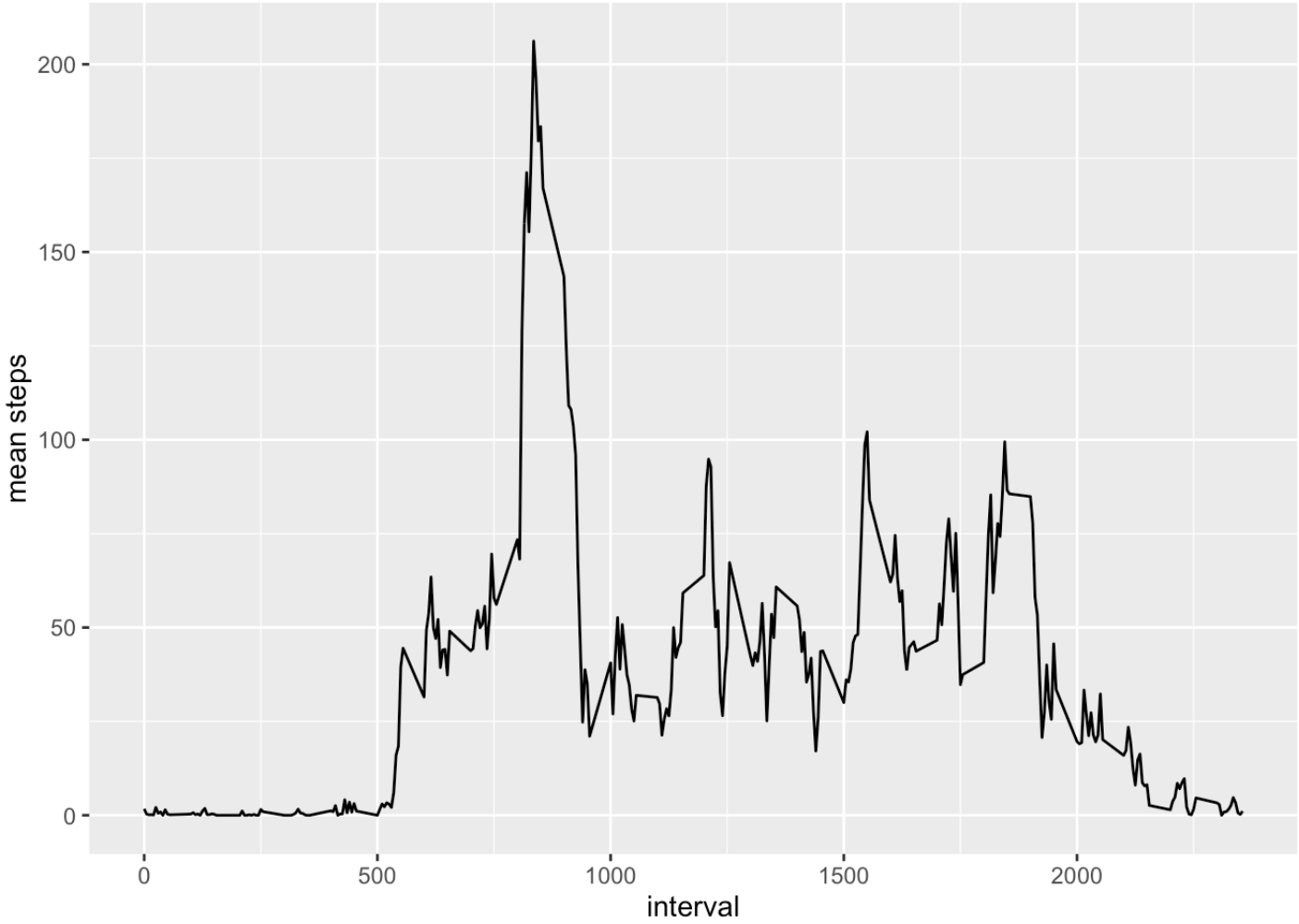
```
## [1] 10766.19
```

```
median_steps
```

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

What is the average daily activity pattern?

```
activity_n<-subset(activity, activity$steps!="NA")
steps_ave<-aggregate(activity_n$steps, by=list(activity_n$interval), FUN=mean)
g<-ggplot(steps_ave, aes(x=Group.1, y=x))
g+geom_line()+xlab("interval")+ylab("mean steps")
```



##Imputing missing values

```
activity.replaceNA<- activity %>% group_by(interval)  %>% mutate(steps= ifelse(is.na(
steps), mean(steps, na.rm=TRUE), steps))
head(activity.replaceNA)
```

```
## Source: local data frame [6 x 3]
## Groups: interval [6]
```

```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone '%Y-%M-%D'

## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone '%Y-%M-%D'
```

```
##       steps       date interval
##       (dbl)     (time)    (int)
## 1 1.7169811 2012-10-01        0
## 2 0.3396226 2012-10-01        5
## 3 0.1320755 2012-10-01       10
## 4 0.1509434 2012-10-01       15
## 5 0.0754717 2012-10-01       20
## 6 2.0943396 2012-10-01       25
```

```
totalstep<-aggregate(activity.replaceNA$steps, na.rm=T, by=list(activity.replaceNA$da
te), FUN=sum)
```

```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone '%Y-%M-%D'

## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone '%Y-%M-%D'
```

```
head(totalstep)
```

```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone '%Y-%M-%D'
```

```
##      Group.1        x
## 1 2012-10-01 10766.19
## 2 2012-10-02   126.00
## 3 2012-10-03 11352.00
## 4 2012-10-04 12116.00
## 5 2012-10-05 13294.00
## 6 2012-10-06 15420.00
```

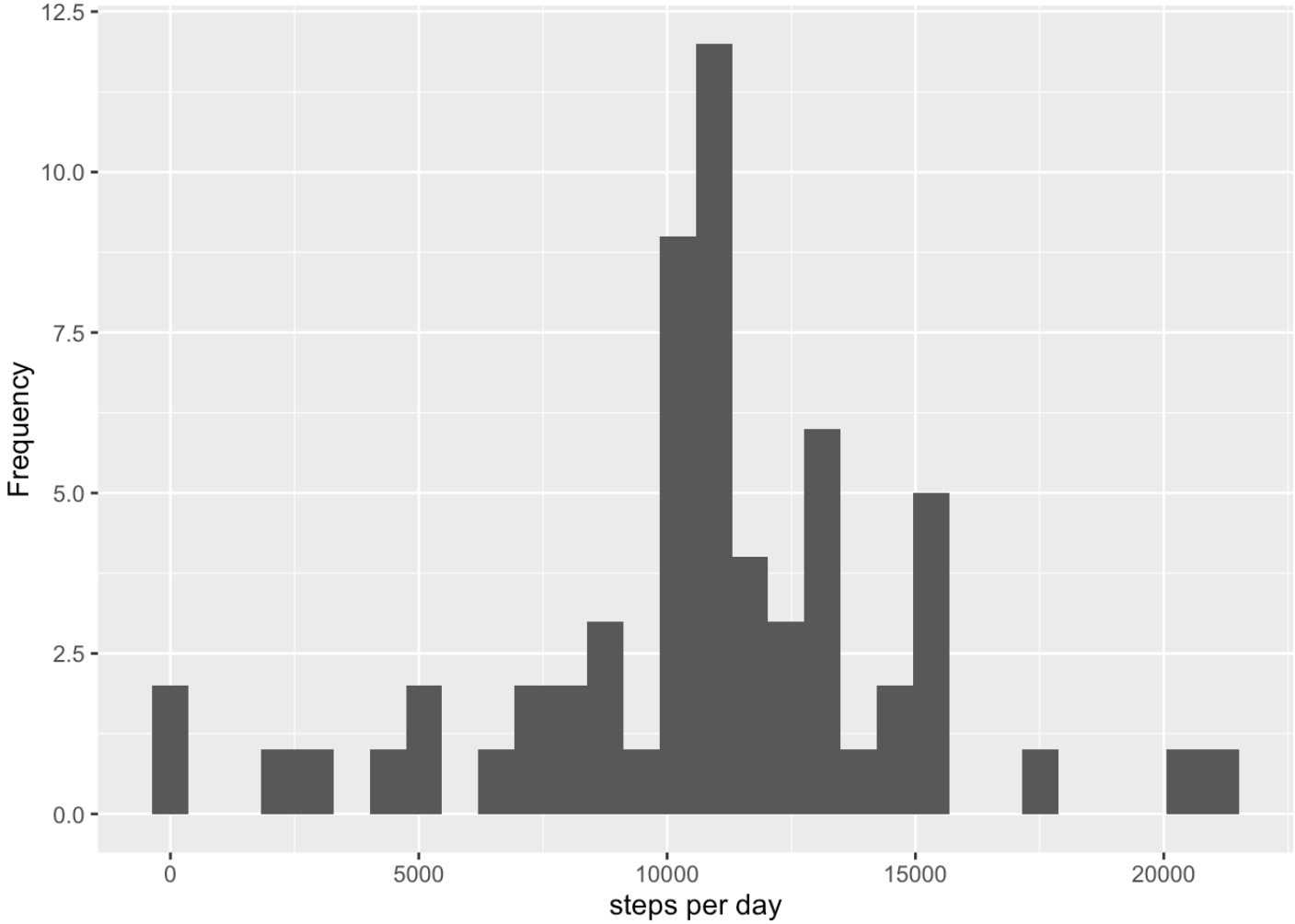
```
sum(is.na(totalstep$steps))
```

```
## Warning in is.na(totalstep$steps): is.na() applied to non-(list or vector)
## of type 'NULL'
```

```
## [1] 0
```

```
g<-ggplot(totalstep, aes(x))
g+geom_histogram()+xlab("steps per day")+ylab("Frequency")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
mean(totalstep$x)
```

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

```
median(totalstep$x)
```

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

##The mean number of steps remains the same since we inserted the prior mean for the missing steps. The median becomes equal to the mean

Are there differences in activity patterns between weekdays and weekends?

```
library(dplyr)
activitynew<-activity.replaceNA
weekdays <- c("Monday", "Tuesday", "Wednesday", "Thursday",
              "Friday")
activitynew$weekdate = as.factor(ifelse(is.element(weekdays(as.Date(activitynew$date)
),weekdays), "Weekday", "Weekend"))
head(activitynew)
```

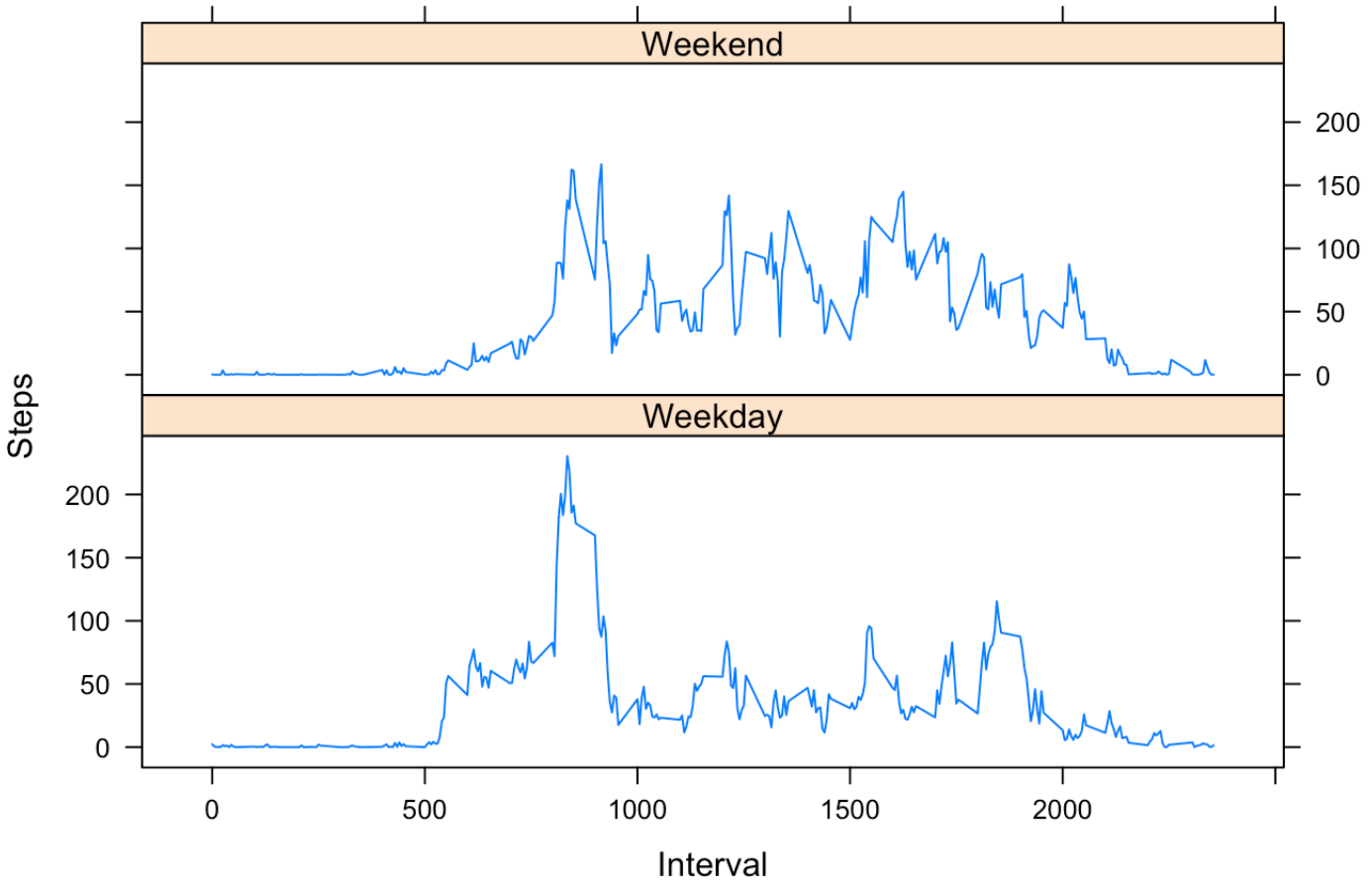
```
## Source: local data frame [6 x 4]
## Groups: interval [6]
```

```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone '%Y-%M-%D'

## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone '%Y-%M-%D'
```

```
##       steps       date interval weekdate
##       (dbl)     (time)    (int)   (fctr)
## 1 1.7169811 2012-10-01        0  Weekday
## 2 0.3396226 2012-10-01        5  Weekday
## 3 0.1320755 2012-10-01       10  Weekday
## 4 0.1509434 2012-10-01       15  Weekday
## 5 0.0754717 2012-10-01       20  Weekday
## 6 2.0943396 2012-10-01       25  Weekday
```

```
library(lattice)
stepsbyinterval<- aggregate(steps ~ interval + weekdate, activitynew, mean)
xyplot(stepsbyinterval$steps ~ stepsbyinterval$interval|stepsbyinterval$weekdate, mai
n="Average steps per day by interval",xlab="Interval", ylab="Steps",layout=c(1,2), ty
pe="l")
```


Average steps per day by interval
Weekend
Weekday
Steps
Interval
0
50
100
150
200
0
500
1000
1500
2000