Assignment1

Joyce Fang

August 31, 2017

1.Loading and preparing data

Raw data comes from activity monitoring data

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.1
```

```
##
## 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
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.1
```

```
library(lattice)
library(knitr)
##download and unzip the dataset:
filename <-"Activity Monitoring Data.zip"
if(!file.exists(filename)){
  fileUrl <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip"
  download.file(fileUrl, filename)
}
if(!file.exists("Activity Monitoring Data.zip")){
  unzip(filename)
}
monitor <- read.csv("activity.csv", header = TRUE, na="NA")
```

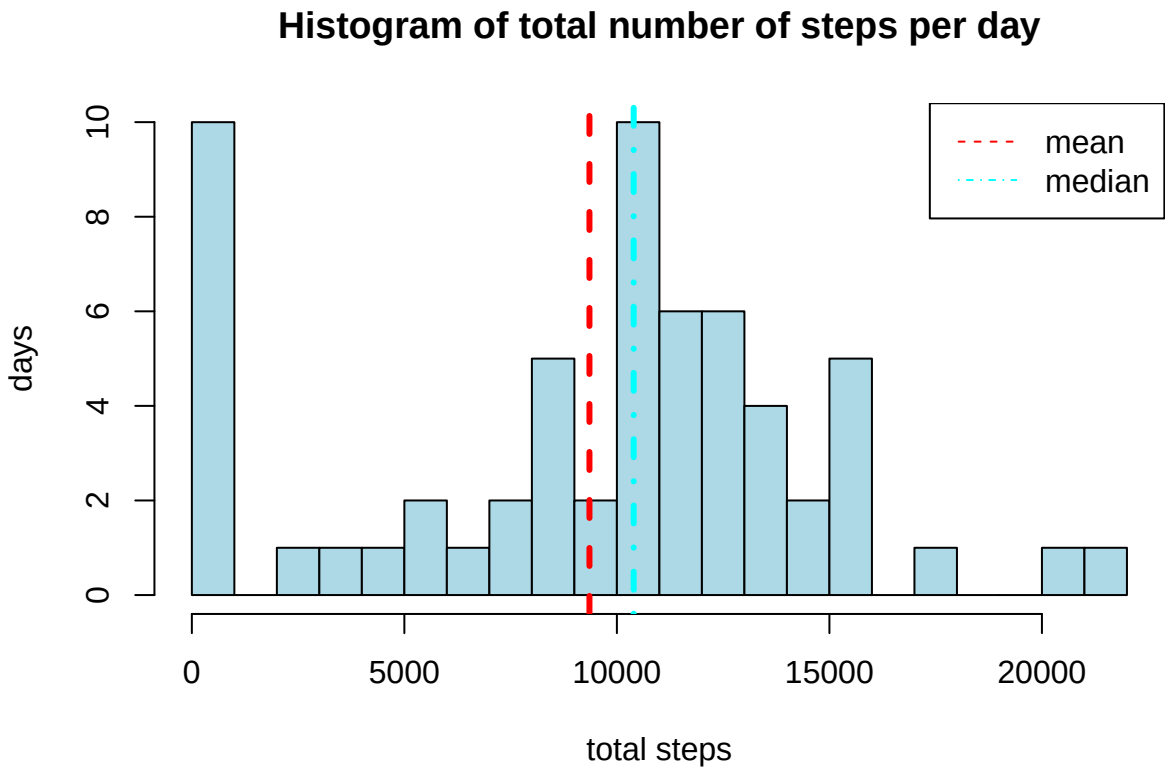
2.What is mean total number of steps taken per day?

First preparing the data by grouping by date to get their total daily steps. Then draw the histogram, adding the vertical lines of mean and median. Finally report the mean and median of the total number of steps taken per day.

```
str(monitor)
```

```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : Factor w/ 61 levels "2012-10-01","2012-10-02",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
```

```
monitor_sum <- aggregate(x=monitor$steps, by=list(monitor$date), FUN = sum, na.rm=TRUE)
names(monitor_sum)<-c("date","total_steps")
#plot the histogram
hist(monitor_sum$total_steps, main="Histogram of total number of steps per day", xlab="total steps", yla
abline(v = mean(monitor_sum$total_steps, na.rm=TRUE),lwd=3, lty=2, col = 2)
abline(v=median(monitor_sum$total_steps, na.rm=TRUE), lwd=3, lty=4, col= 5)
legend("topright", c("mean", "median"), lty =c(2,4), col=c(2,5))
```




```
dev.copy(png, "Plot1.png")
```

```
## png
##   3
```

```
dev.off()
```

```
## pdf
##   2
```

```
print(mean(monitor_sum$total_steps, na.rm=TRUE))
```

```
## [1] 9354.23
```

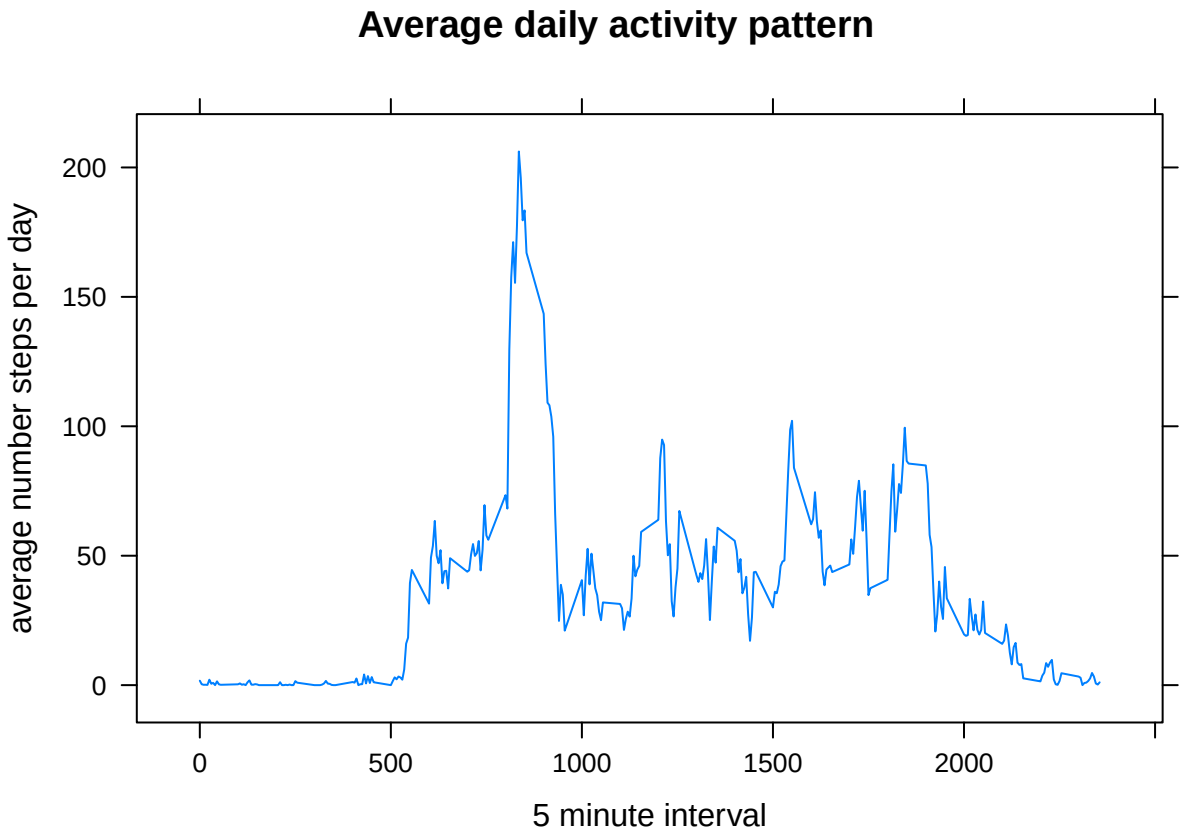
```
print(median(monitor_sum$total_steps, na.rm= TRUE))
```

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

3.What is the average daily activity pattern?

Make a time series plot(with type="l") of the 5-minute interval(x-axis) and total number of steps taken, averaged across all the day(y-axis),finally get the interval with the largest number of steps.

```
##time seires problem
monitor_interval <- aggregate(x=list(steps=monitor$steps), by = list(interval = monitor$interval), FUN=
xyplot(monitor_interval$steps~monitor_interval$interval,  type="l", xlab="5 minute interval", ylab="ave
```




```
#theme(axis.text.x = element_text(angle = 45,hjust = 1))
max <- monitor_interval[which.max(monitor_interval$steps), ]
dev.copy(png, "Plot2.png")
```

```
## png
##   3
```

```
dev.off()
```

```
## pdf
##   2
```

```
max
```

```
##     interval    steps
```

```
## 104      835 206.1698
```

4.Imputing missing values

1.check how many NA are in steps and interval

```
monitor_subna <- monitor
na_row <- is.na(monitor$steps)
na_in <- is.na(monitor$interval)
print(sum(na_row))
```

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

```
print(sum(na_in))
```

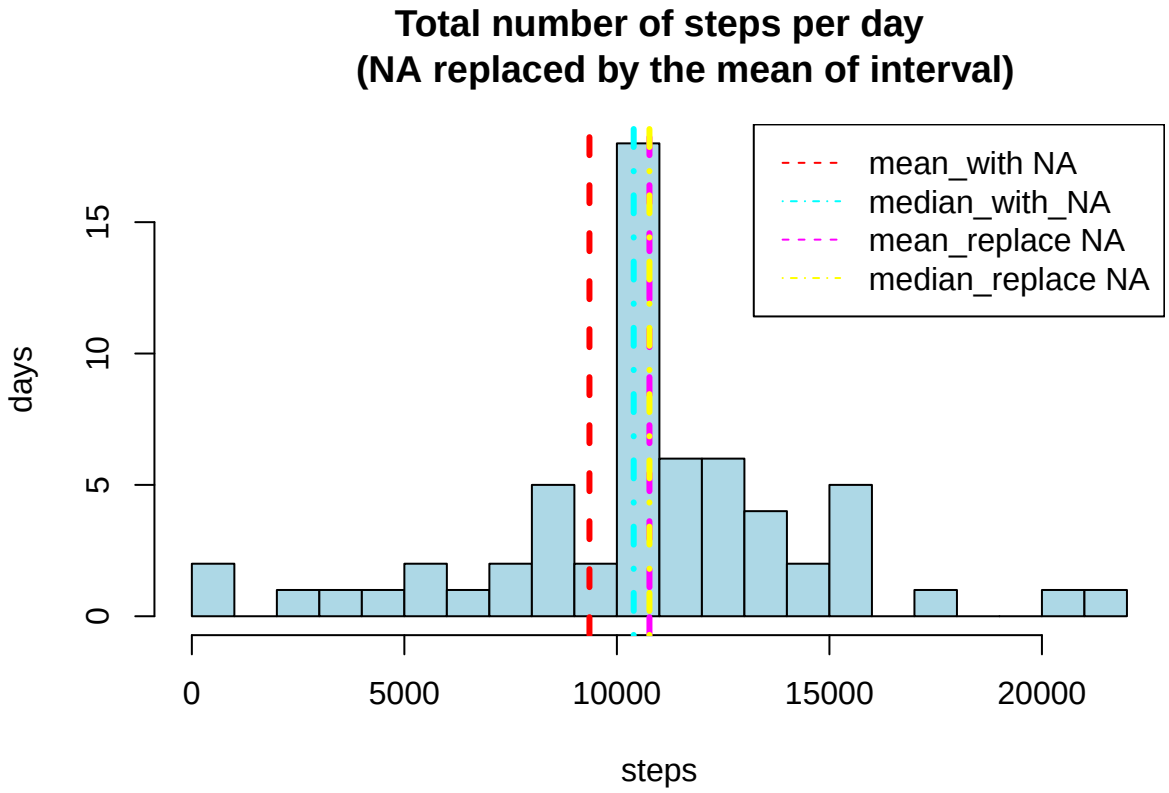
```
## [1] 0
```

there is no NA in interval 2.Replace NA with the mean of 5-minute interval to a new data set:monitor_subna

```
average_interval <- tapply(monitor$steps, monitor$interval, mean, na.rm=TRUE, simplify = TRUE)
monitor_subna$steps[na_row] <- average_interval[as.character(monitor_subna$interval[na_row])]
#get the average # of steps in 5-minute intervals, the method is the same as question1
subna_sum <- aggregate(x=monitor_subna$steps, by=list(monitor_subna$date), FUN = sum)
names(subna_sum) <- c("Date","steps")
```

4.make histogram of total number of steps and compare with the means and medians with/without NA replacement

```
hist(subna_sum$steps, breaks = 30, main="Total number of steps per day
     (NA replaced by the mean of interval)", xlab="steps", ylab="days", col = "lightblue")
abline(v = mean(subna_sum$steps),lwd=3, lty=2, col = 6)
abline(v=median(subna_sum$steps, na.rm=TRUE), lwd=3, lty=4, col= 7)
#add mean/median of steps without NA replace
abline(v = mean(monitor_sum$total_steps, na.rm=TRUE),lwd=3, lty=2, col = 2)
abline(v=median(monitor_sum$total_steps, na.rm=TRUE), lwd=3, lty=4, col= 5)
legend("topright", c("mean_with NA", "median_with_NA","mean_replace NA", "median_replace NA"), lty =c(2
```




```
dev.copy(png, "Plot3.png")
```

```
## png
##   3
```

```
dev.off()
```

```
## pdf
##   2
```

The means of total steps per day do no change before/after replacement of NA, but the median increased to the value of mean.

5.Are there differences in activity patterns between weekdays and weekends?

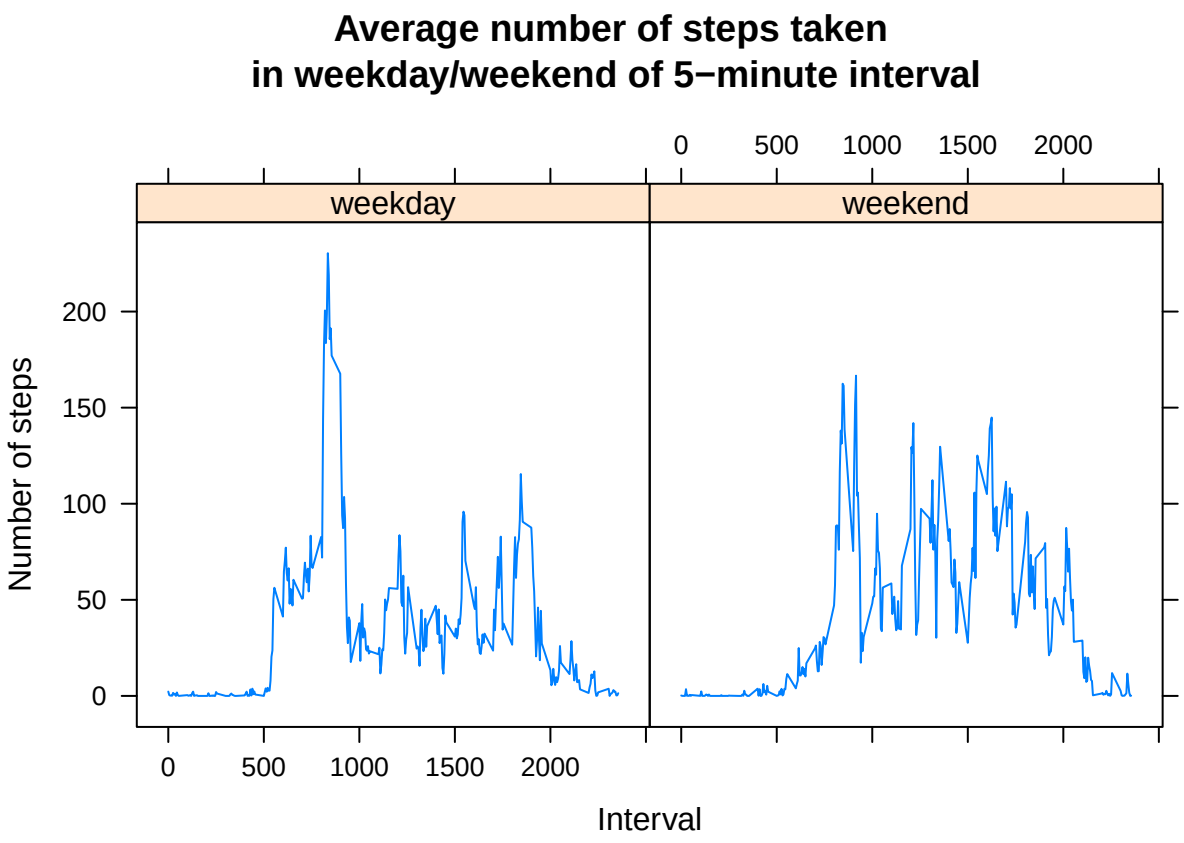
1.create new variables in the dataset with"weekday","weekend".

```
monitor_subna$weekday <- weekdays(as.Date(monitor_subna$date))
m_f <- c("Monday","Tuesday","Wednesday","Thursday","Friday")
s_s <- c("Saturday","Sunday")
monitor_subna$weekday[monitor_subna$weekday%in%m_f]<- "weekday"
monitor_subna$weekday[monitor_subna$weekday%in%s_s]<- "weekend"
```

2.make a panel plot containing a time series plot of 5-minute interval and the avg number of steps taken.

```
#first change the week type to factor
monitor_subna$weekday <- as.factor(monitor_subna$weekday)
subna_interval_sum <- aggregate(steps ~ interval + weekday, data = monitor_subna, mean)
```

```
xyplot(steps~interval|weekday, data=subna_interval_sum, type="l", xlab="Interval", ylab="Number of steps
in weekday/weekend of 5-minute interval")
```




```
dev.copy(png, "Plot4.png")
```

```
## png
##   3
```

```
dev.off()
```

```
## pdf
##   2
```