

# peerass1.Rmd

YI HAN

Sunday, April 12, 2015

## Loading and preprocessing the data

```
#setwd("C:/Users/vannem.han/OneDrive/2_Coursera/JHU_Repro_Res/PeerAssess/1")
OriginalData <- read.csv(file="activity.csv", head=TRUE, sep=",")
```

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

```
Mydata <- OriginalData
Date <- Mydata[,2]
DateTable <- table(Date)

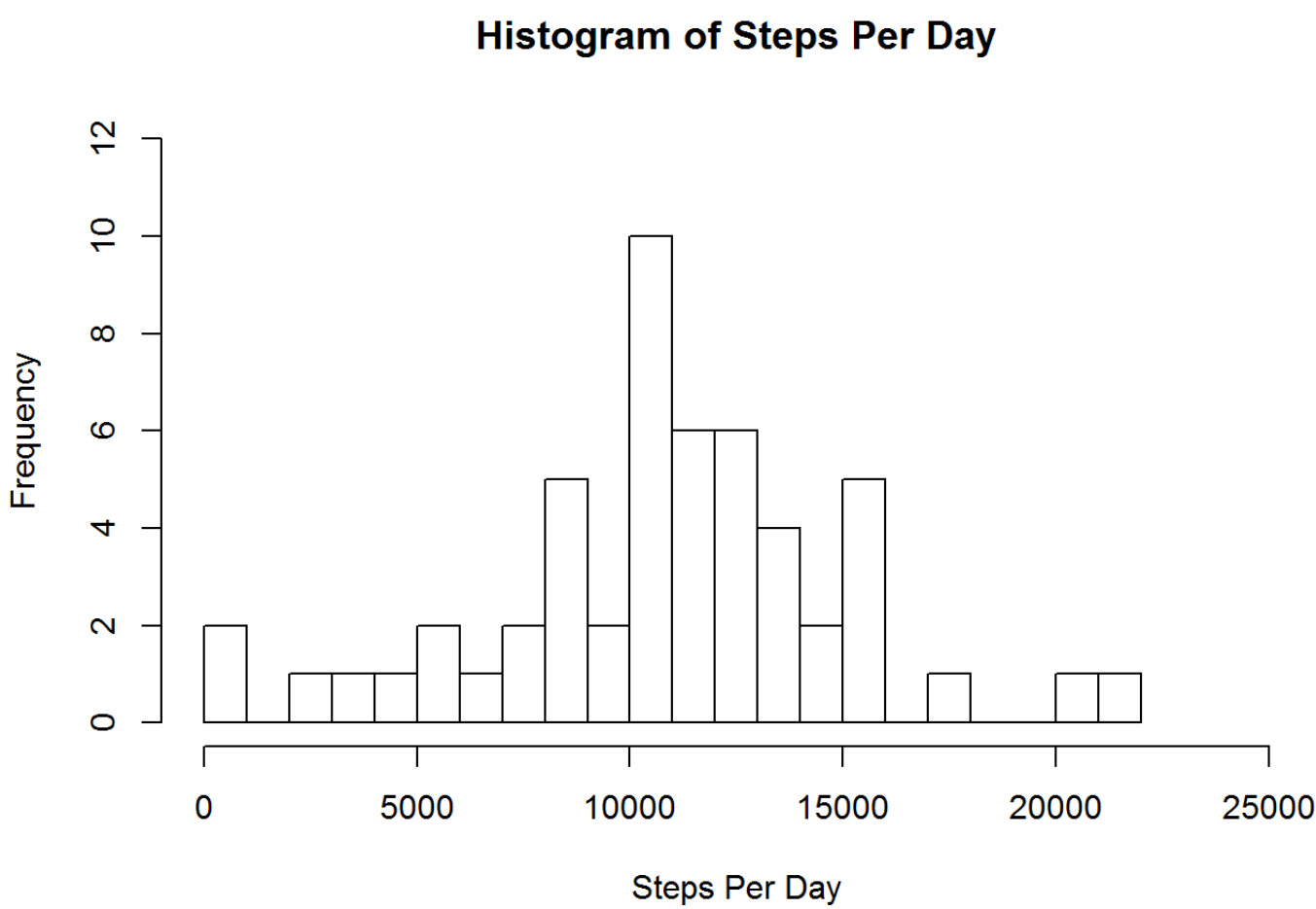
NumIntervalPerDay <- DateTable[1] #each day has 288 intervals
NumDate <- length(DateTable) #It has 61 days

StepsPerDay <- rep(0, NumDate)

for(i in 1:NumDate)
  StepsPerDay[i] = sum(Mydata[(1+NumIntervalPerDay*(i-1)): (NumIntervalPerDay+NumIntervalPerDay*(i-1)),1])

StepsPerDay = StepsPerDay[!is.na(StepsPerDay)]

hist(StepsPerDay, breaks=30, xlab="Steps Per Day", main="Histogram of Steps Per Day", xlim=c(0, 25000), ylim=c(0, 12))
```



```
mean(StepsPerDay)
```

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

```
median(StepsPerDay)
```

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

- 2. The above pic is the histogram of the total number of steps taken each day.
- 3. The mean and median of the otal number of steps taken per day is 10766.19 and 10765.

## What is the average daily activity pattern?

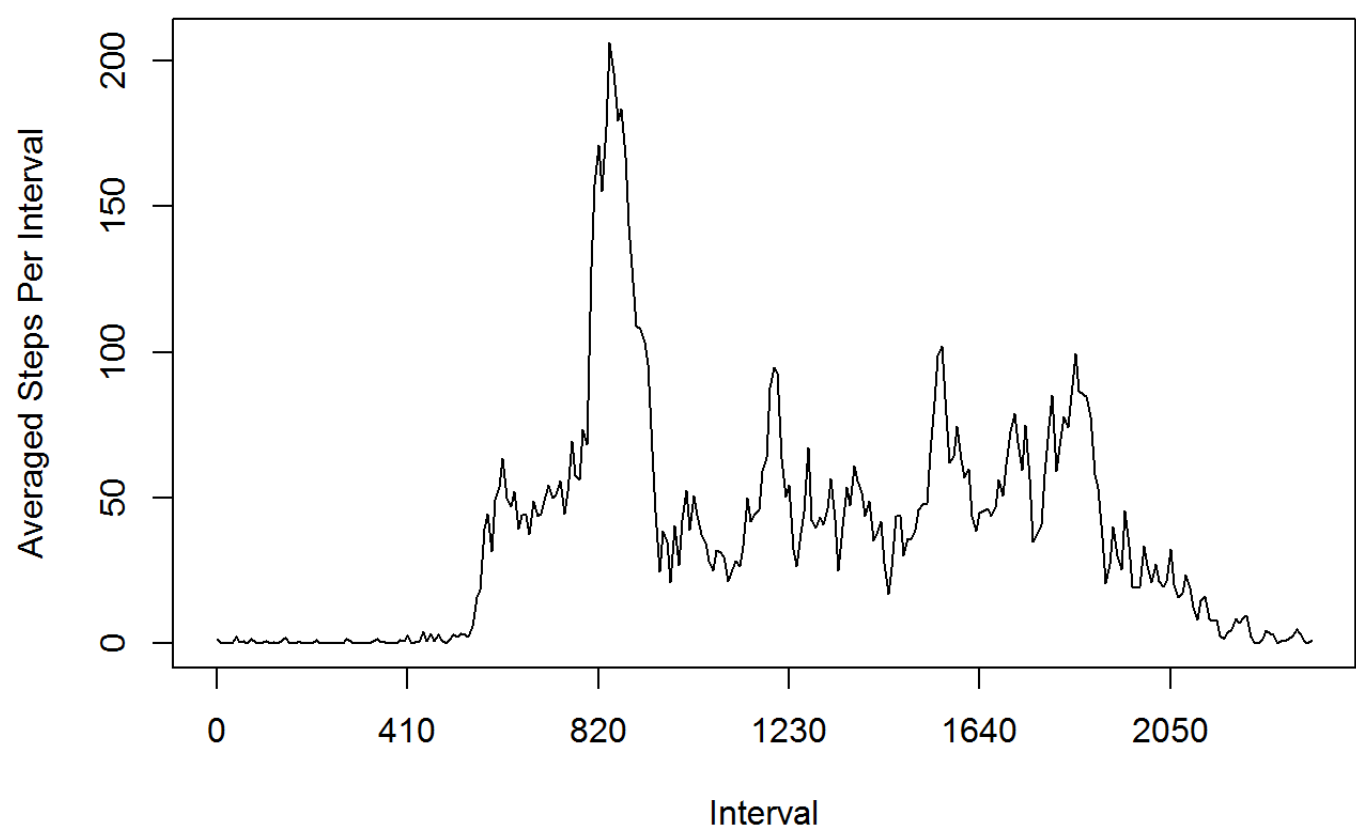
```
StepsPerInterval <- rep(0, NumIntervalPerDay)

for(i in 1:NumIntervalPerDay)
  StepsPerInterval[i] = mean(Mydata[seq(from=i, to=i+NumIntervalPerDay*(NumDate-1), by=NumIntervalPerDay),1], na.rm=TRUE)

Interval <- Mydata[,3]
IntervalTable <- table(Interval)
IntervalNames <- names(IntervalTable) #interval names like 0 5 10 ... for plot

StepsPerInterval <- ts(StepsPerInterval)
plot(StepsPerInterval,xaxt='n', main="Time series of Averaged Steps Per Interval", ylab="Averaged Steps Per Interval", xlab="Interval")
axis(1, at=seq(1,NumIntervalPerDay,50),labels=IntervalNames[seq(1,NumIntervalPerDay,50)])
```

Time series of Averaged Steps Per Interval



```
#max(StepsPerInterval)
#which.max(StepsPerInterval) #This-th 5 minute interval
IntervalNames[which.max(StepsPerInterval)] #This 5 minute interval
```

```
## [1] "835"
```

1. The above pic is the time series plot of the 5-minute interval and the average number of steps taken, averaged across all days.
2. The 835-th 5-minute interval, on average across all the days, contains the maximum number of steps.

# Imputing missing values

```
Mydata <- OriginalData
sum(is.na(Mydata[,1])) #the total number of missing values
```

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

```
Naposition = which(is.na(Mydata[,1]))

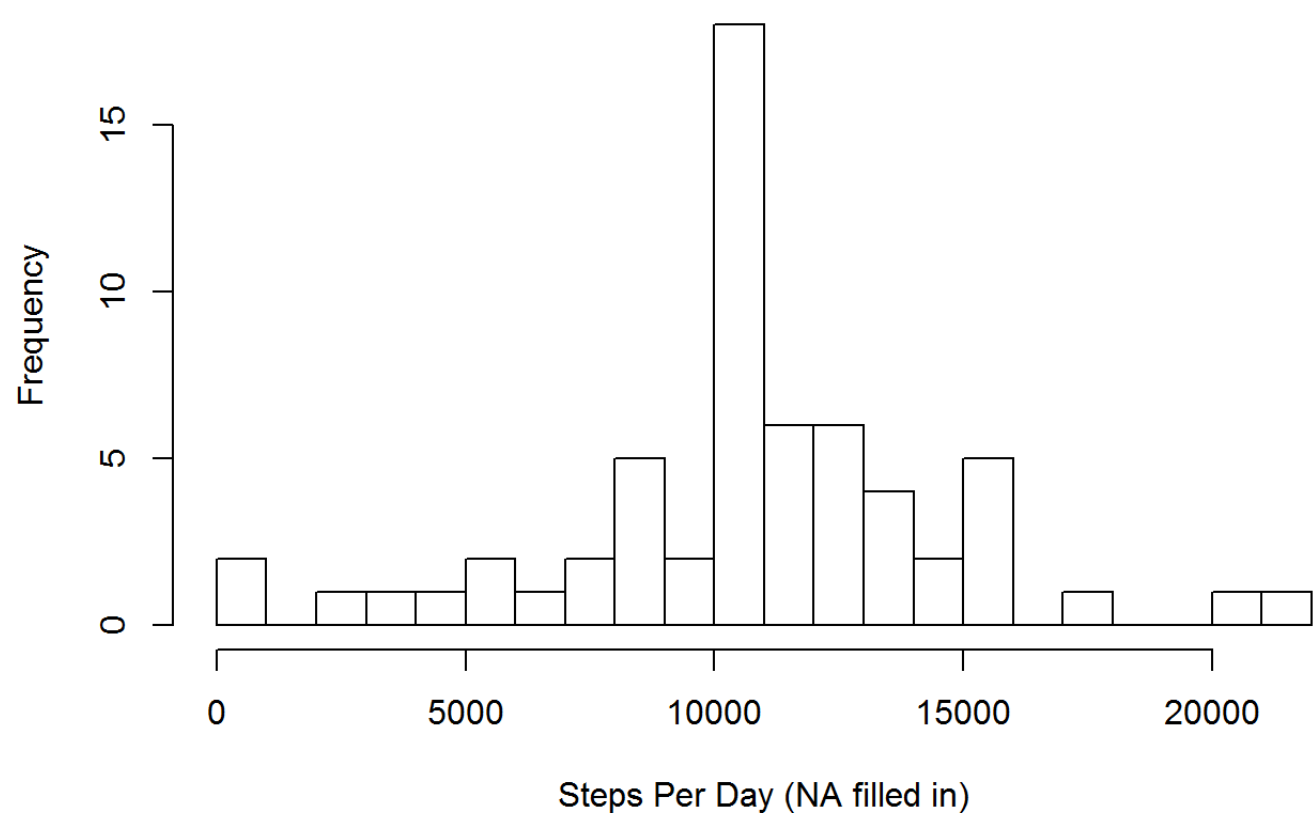
for(i in Naposition)
  Mydata[i,1] = StepsPerInterval[which(IntervalNames==Mydata[i,3])] #My strategy: the mean for that 5-minute interval

StepsPerDay_fillin <- rep(0, NumDate)

for(i in 1:NumDate)
  StepsPerDay_fillin[i] = sum(Mydata[(1+NumIntervalPerDay*(i-1)): (NumIntervalPerDay+NumIntervalPerDay*(i-1)),1])

hist(StepsPerDay_fillin, breaks=30, xlab="Steps Per Day (NA filled in)", main="Histogram of Steps Per Day (NA filled in)")
```

Histogram of Steps Per Day (NA filled in)



```
mean(StepsPerDay_fillin)
```

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

```
median(StepsPerDay_fillin)
```

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

- 1. 2304 is the total number of missing values in the dataset.
- 2. I use the mean for that 5-minute interval for filling in all of the missing values in the dataset.
- 3. The mean and median of the NA-filled dataset is 10766.19 and 10766.19. The mean value is the same from the previou one, while the median value changes. This is because I use the mean for that 5-minute interval to fill in the NA.

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

```
Sys.setlocale("LC_TIME", "English") #set weekdays to be in English language
```

```
## [1] "English_United States.1252"
```

```
DateNames <- names(DateTable)

DayorEnd <- weekdays(as.Date(DateNames))
Seq_end <- c(which(DayorEnd=="Saturday"),which(DayorEnd=="Sunday"))

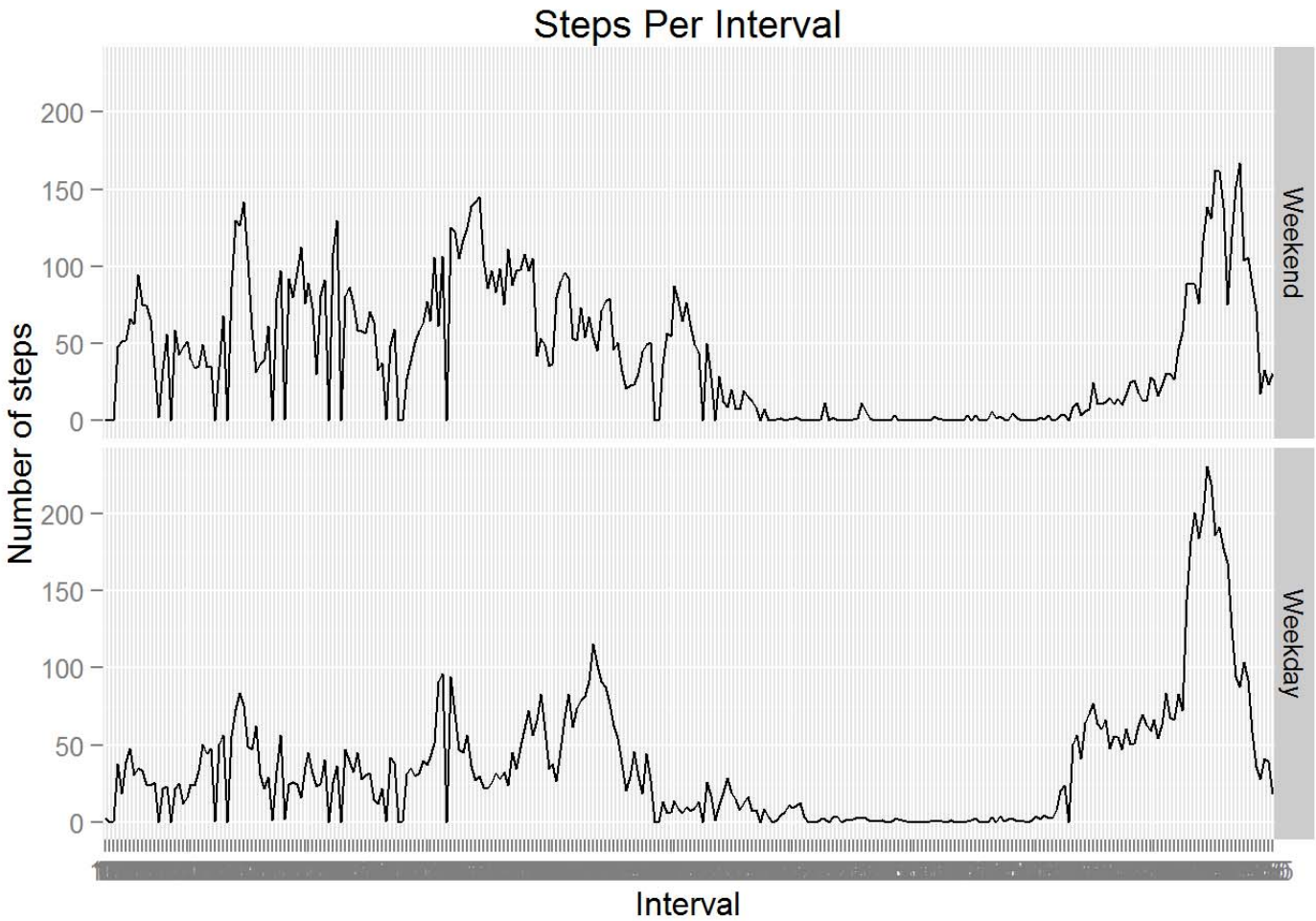
DayorEnd[Seq_end] = "weekend"
DayorEnd[which(DayorEnd!="weekend")] = "weekday"
DayorEnd = factor(DayorEnd)

Weekend <- rep(0, NumIntervalPerDay)
Weekday <- rep(0, NumIntervalPerDay)

for(i in 1:NumIntervalPerDay)
{ DateOneInterval <- seq(from=i, to=i+NumIntervalPerDay*(NumDate-1), by=NumIntervalPerDay)#coresponse to DateNames and DayorEnd.

  Weekend[i] = mean(Mydata[DateOneInterval[Seq_end],1])
  Weekday[i] = mean(Mydata[DateOneInterval[-Seq_end],1])
}

df <- data.frame(IntervalNames, Weekend, Weekday)
library(reshape2)
mm <- melt(df,id.var="IntervalNames")
library(ggplot2)
ggplot(mm, aes(IntervalNames, value, group=1)) +
  #geom_point() +
  geom_line() +
  facet_grid(variable~.) +
  labs(x="Interval",y="Number of steps",title="Steps Per Interval")
```



2. The above pic is the panel plot containing a time series plot of the 5 minutes interval (x-axis) and the average number of steps taken, averaged across all weekday days or weekend days (y-axis).