Reproducible research - Activity analysis

Susie X

Data input and preprocessing:

```
Data=read.csv("activity.csv",header=T)
dim(Data)
```

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

```
Data$date=as.Date(Data$date)
str(Data)
```

```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : Date, format: "2012-10-01" "2012-10-01" ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
```

```
length(unique(Data$interval))
```

```
## [1] 288
```

What is mean total number of steps taken per day?

Total number of steps each day with incomplete cases removed:

```
library(dplyr)
```

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

```
Data_comp= Data[complete.cases(Data),]
SumData_comp = Data_comp %>% group_by(date) %>% summarize(daysumsteps=sum(steps))
```

Mean of sum steps each day with missing data removed:

```
mean(SumData_comp$daysumsteps)
```

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

What is the average daily activity pattern?

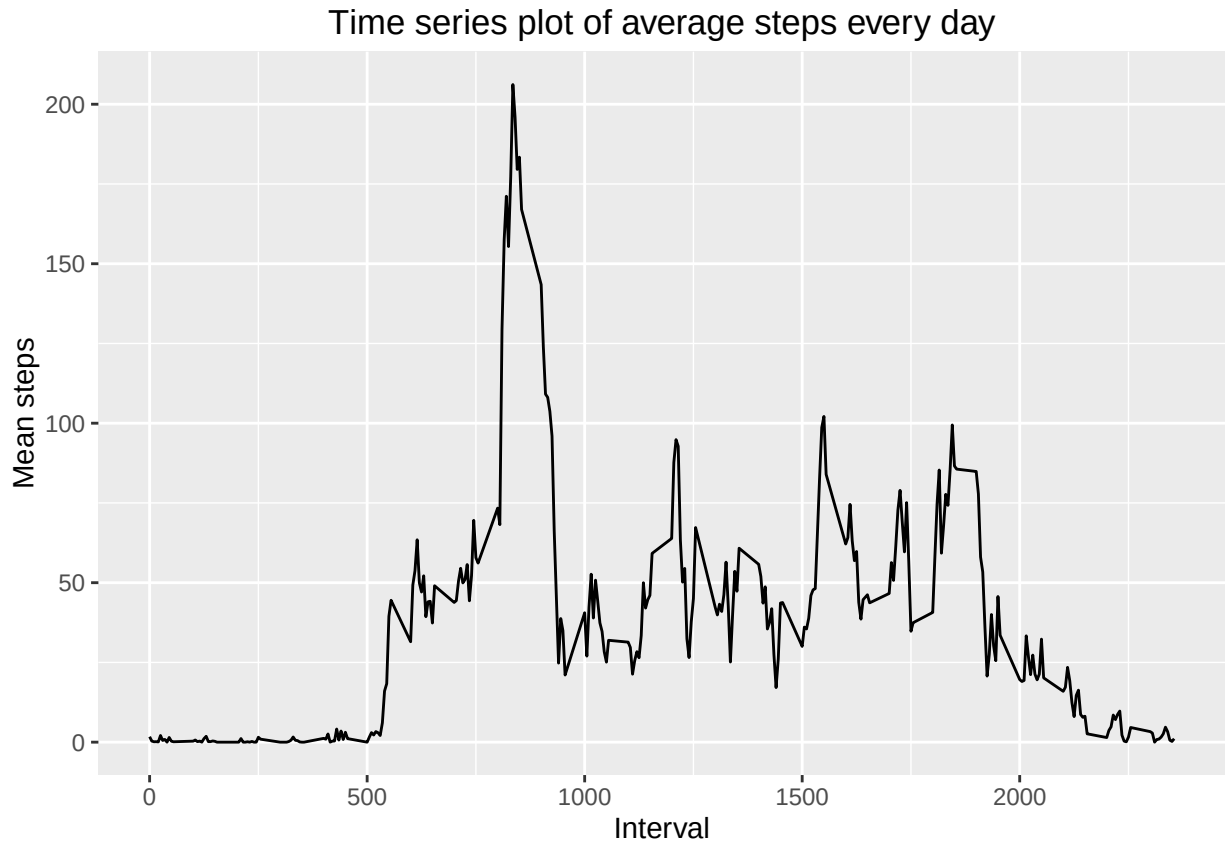
```
AverData_comp = Data_comp %>% group_by(interval) %>% summarize(aver = mean(steps))
```

Interval with highest average steps:

```
AverData_comp$interval[which(AverData_comp$aver == max(AverData_comp$aver))]
```

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

```
library(ggplot2)
ggplot(AverData_comp, aes(interval, aver)) + geom_line() + xlab("Interval") + ylab("Mean steps") + ggti
```



Imputing missing values:

Number of missing values in dataset:

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

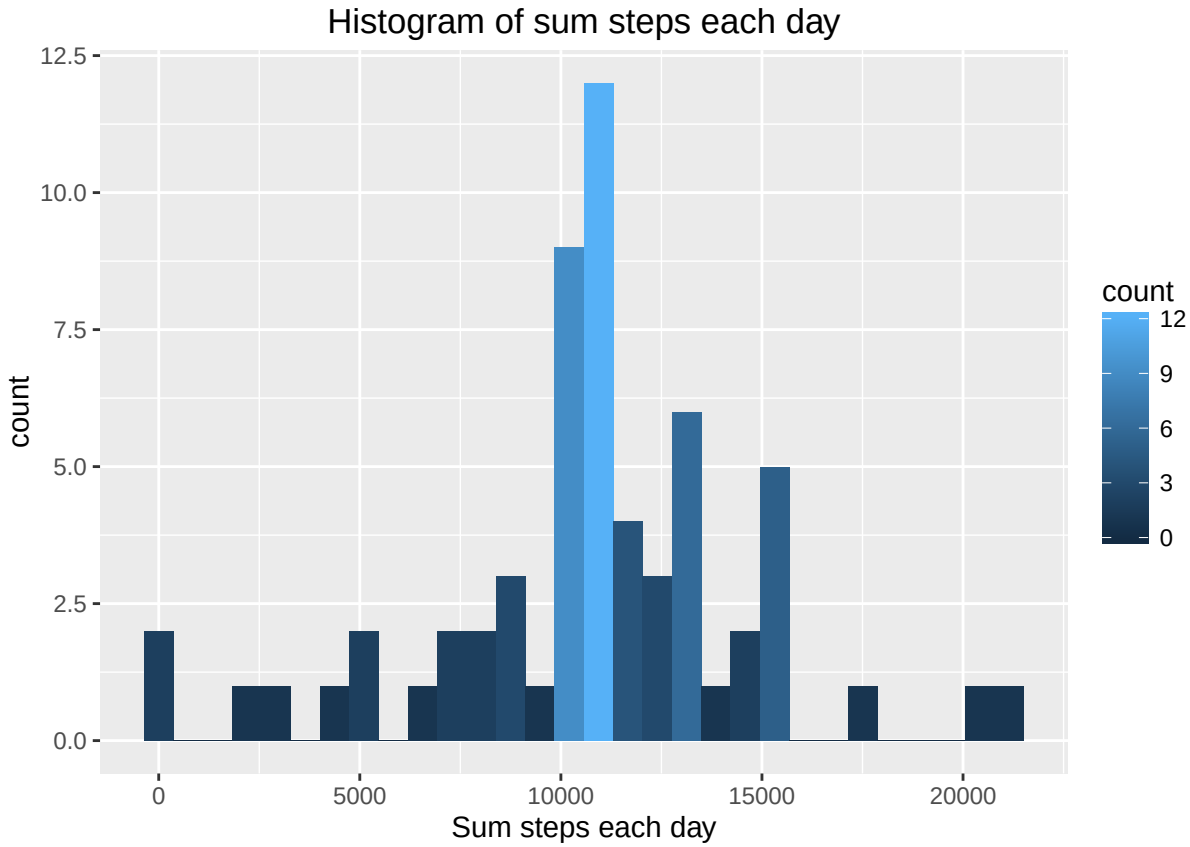
```
## [1] 2304
```

Use average steps in the interval to fill in missing data:

```
Datanew= Data
for (i in 1: 17568) {
if(is.na(Datanew[i,1])){
      Datanew[i,1] = AverData_comp$aver[which(AverData_comp$interval==Datanew[i,3])]
}
}
sum(is.na(Datanew$steps))
```

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

```
SumData_new = Datanew %>% group_by(date) %>% summarize(daysumsteps=sum(steps))
qplot(daysumsteps,data=SumData_new,geom="histogram")+ geom_histogram(aes(fill = ..count..))+xlab("Sum s
```



Mean and median of sum steps each day:

```
mean(SumData_new$daysumsteps)
```

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

```
median(SumData_new$daysumsteps)
```

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

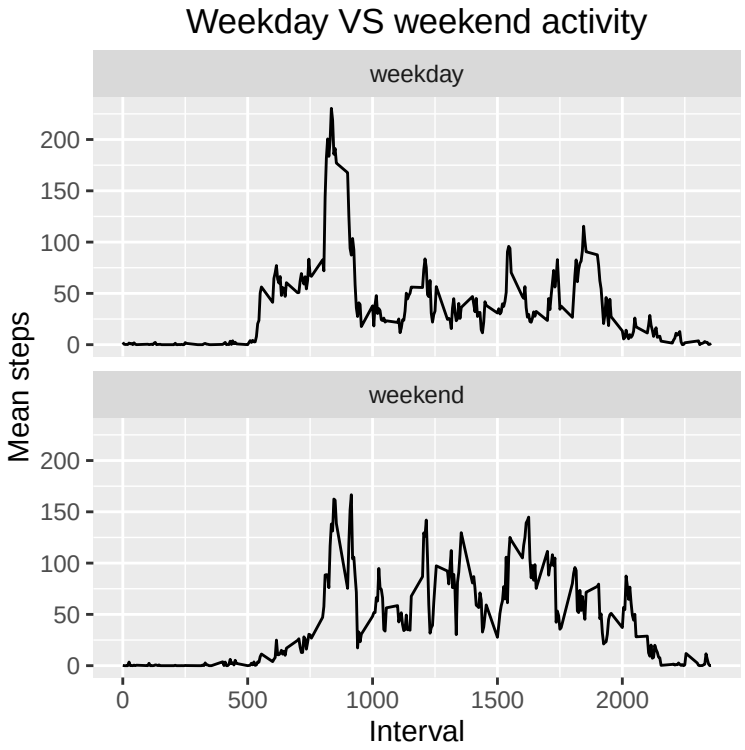
It does not affect the mean because the result in first step is from dataset with missing value removed and in this step, the missing value was filled by mean steps of the same interval.

Are there differences in activity patterns between weekdays and weekends?

```
Datanew = Datanew %>% mutate(day = weekdays(date)) %>% mutate(wd = ifelse(day %in% c("Monday","Tuesday"

library(data.table)
New = data.table(Datanew)
New[,mean := mean(steps),by=list(wd,interval)]

ggplot(New, aes(interval, mean)) + geom_line() + facet_wrap(~wd,ncol=1) + xlab("Interval") + ylab("Mean
```



Both weekdays and weekend have peak activity around interval near 800, but there are more activities after the peak in weekends. Also there are slightly more activities before the peak in weekdays.