week 2 rmd

GAURAV

6/28/2020

```r
library(data.table)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: ’dplyr’

## The following objects are masked from ’package:data.table’:
##
##     between, first, last

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

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

```r
file=read.csv("activity.csv",sep = ",",header = T)

file1<- file%>%filter(steps!="NA")
file1$steps<- as.numeric(file1$steps)
file1$interval<- as.factor(file1$interval)
#lets view the summary of file without Na values
summary(file1)
```

```
##      steps                date          interval
##  Min.   :  0.00   2012-10-02:  288   0      :   53
##  1st Qu.:  0.00   2012-10-03:  288   5      :   53
##  Median :  0.00   2012-10-04:  288   10     :   53
##  Mean   : 37.38   2012-10-05:  288   15     :   53
##  3rd Qu.: 12.00   2012-10-06:  288   20     :   53
##  Max.   :806.00   2012-10-07:  288   25     :   53
##                   (Other)   :13536   (Other):14946
```

```r
spd1 <- aggregate(steps~date,file1,FUN = sum)
head(spd1)
```

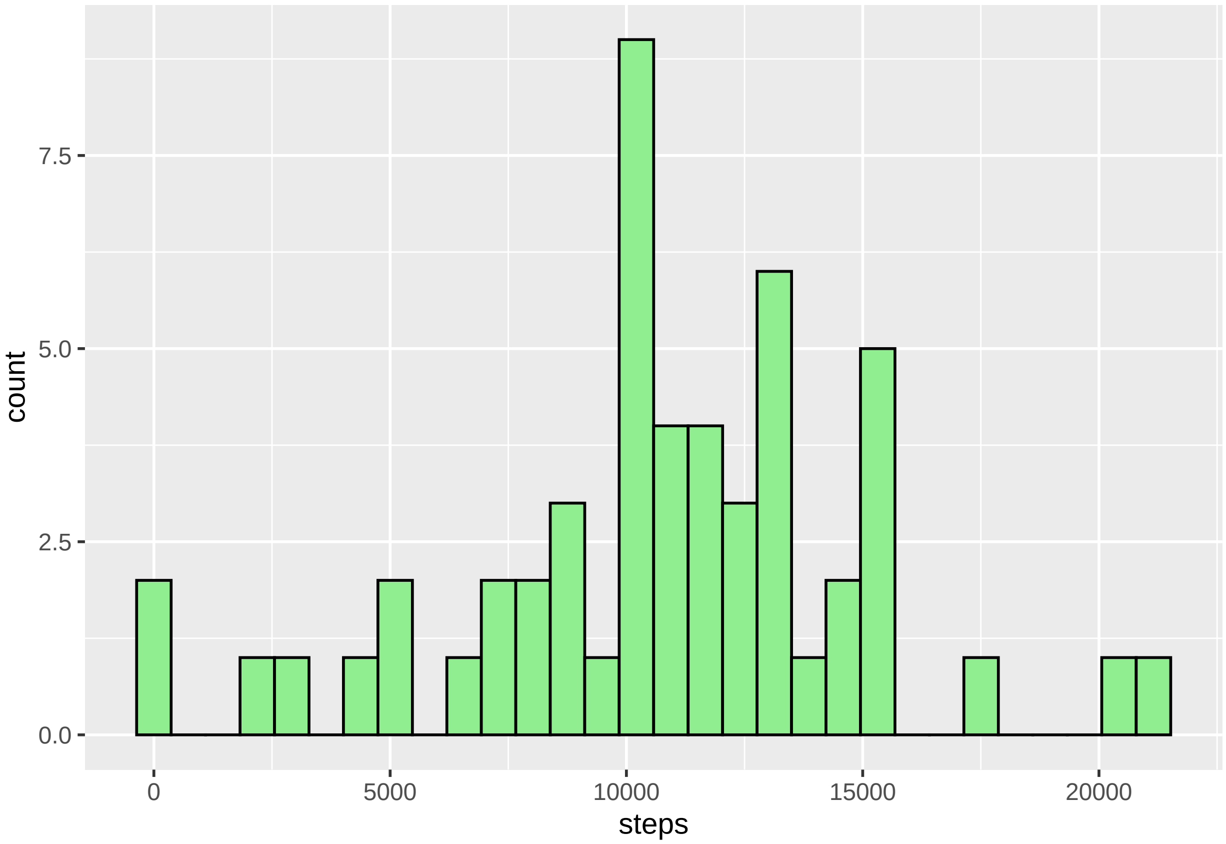
```
##         date steps
## 1 2012-10-02   126
## 2 2012-10-03 11352
## 3 2012-10-04 12116
## 4 2012-10-05 13294
## 5 2012-10-06 15420
## 6 2012-10-07 11015
```

```
summary(spd1)
```

```
##          date        steps
##  2012-10-02: 1   Min.   :   41
##  2012-10-03: 1   1st Qu.: 8841
##  2012-10-04: 1   Median :10765
##  2012-10-05: 1   Mean   :10766
##  2012-10-06: 1   3rd Qu.:13294
##  2012-10-07: 1   Max.   :21194
##  (Other)   :47
```

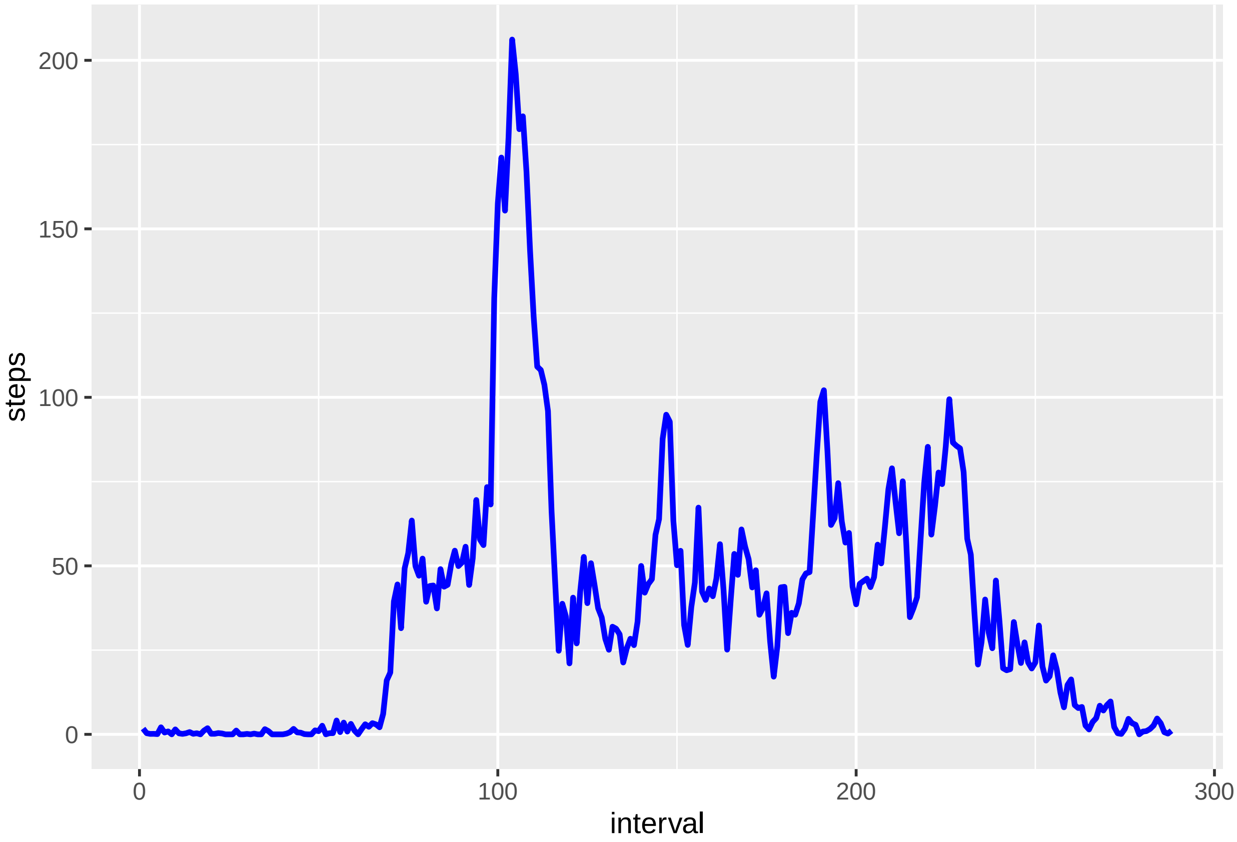
```
histogram<-ggplot(spd1,aes(x=steps))+geom_histogram(fill="lightgreen",col="black")
histogram
```

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



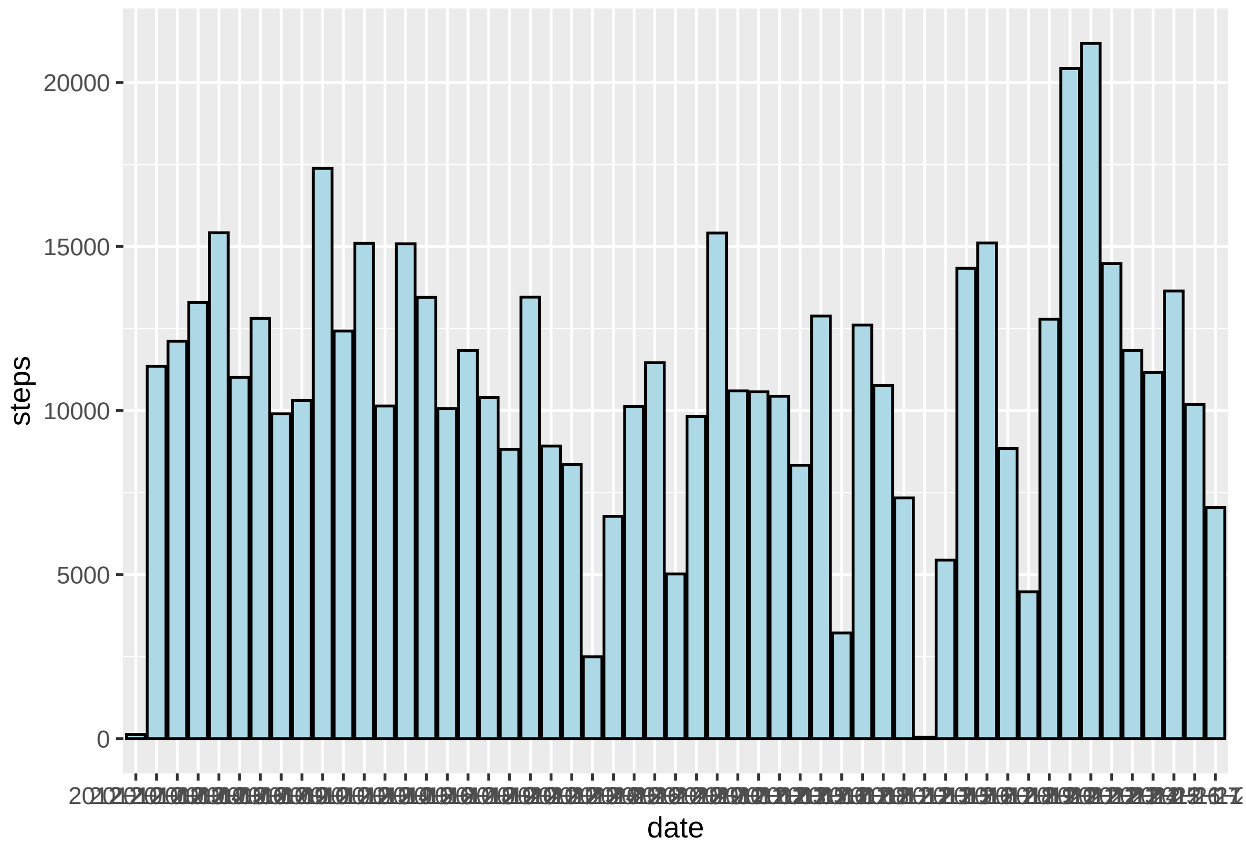
```
#formating the date in this dataset

file1$date<- as.Date(file1$date,"%Y-%m-%d")
spinterval <- aggregate(steps ~ interval, data = file1, FUN = mean)
spinterval$interval<- as.integer(spinterval$interval)
#lets us see the interval for maximum steps
max_interval<- spinterval[which.max(spinterval$steps),]
#plot for 5 min interval using histogram
time_series<-ggplot(spinterval,aes(x=interval,y=steps))+geom_line(size=1,col="blue")
time_series
```



```
#difference btw hist and bar graphs

barplot<-ggplot(spd1,aes(x=date,y=steps))+geom_bar(stat="identity",col="black",fill="lightblue")
barplot
```



```
#now lets see the total number of missing values in orginal dataset
nomissingValue<- sum(is.na(file$steps))
nomissingValue
```

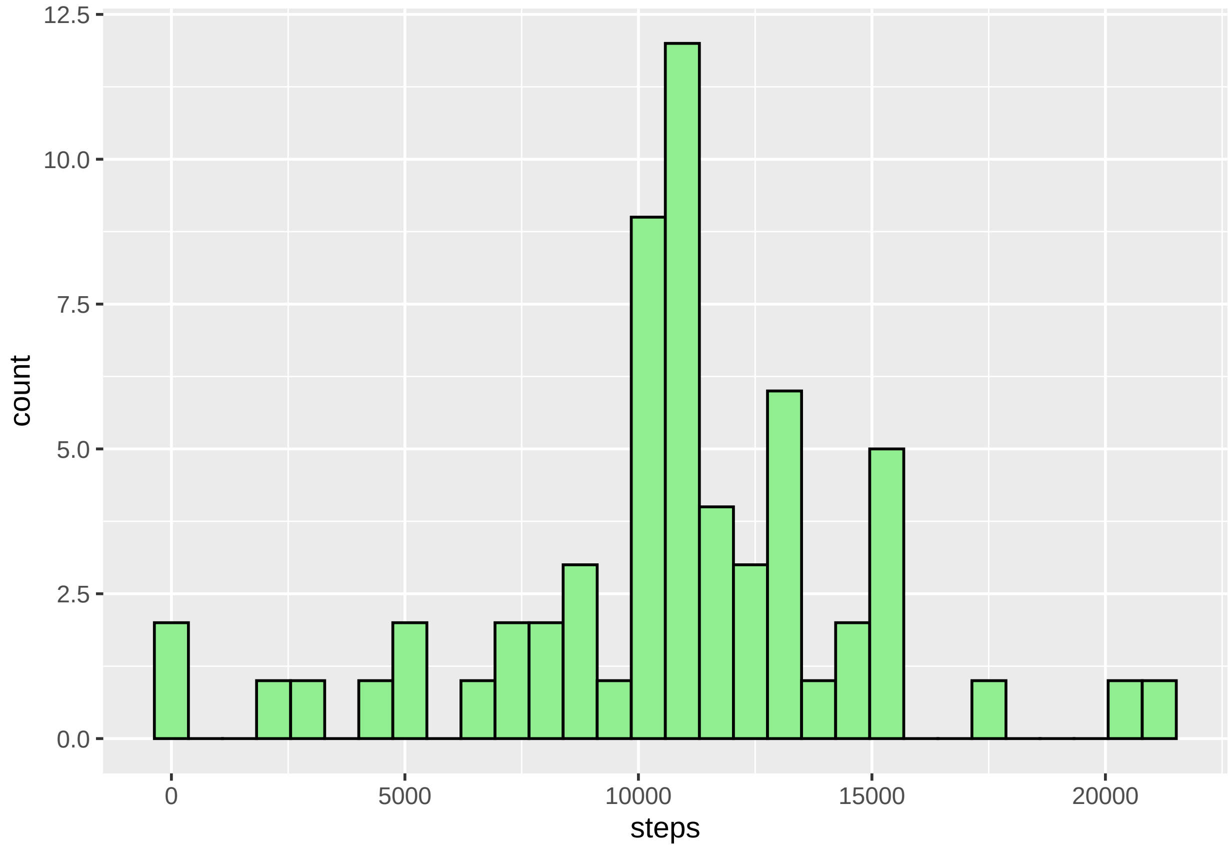
```
## [1] 2304
```

```
# for missing value ,replace them with mean of median values but for now i am using mean value
mean(spinterval$steps)
```

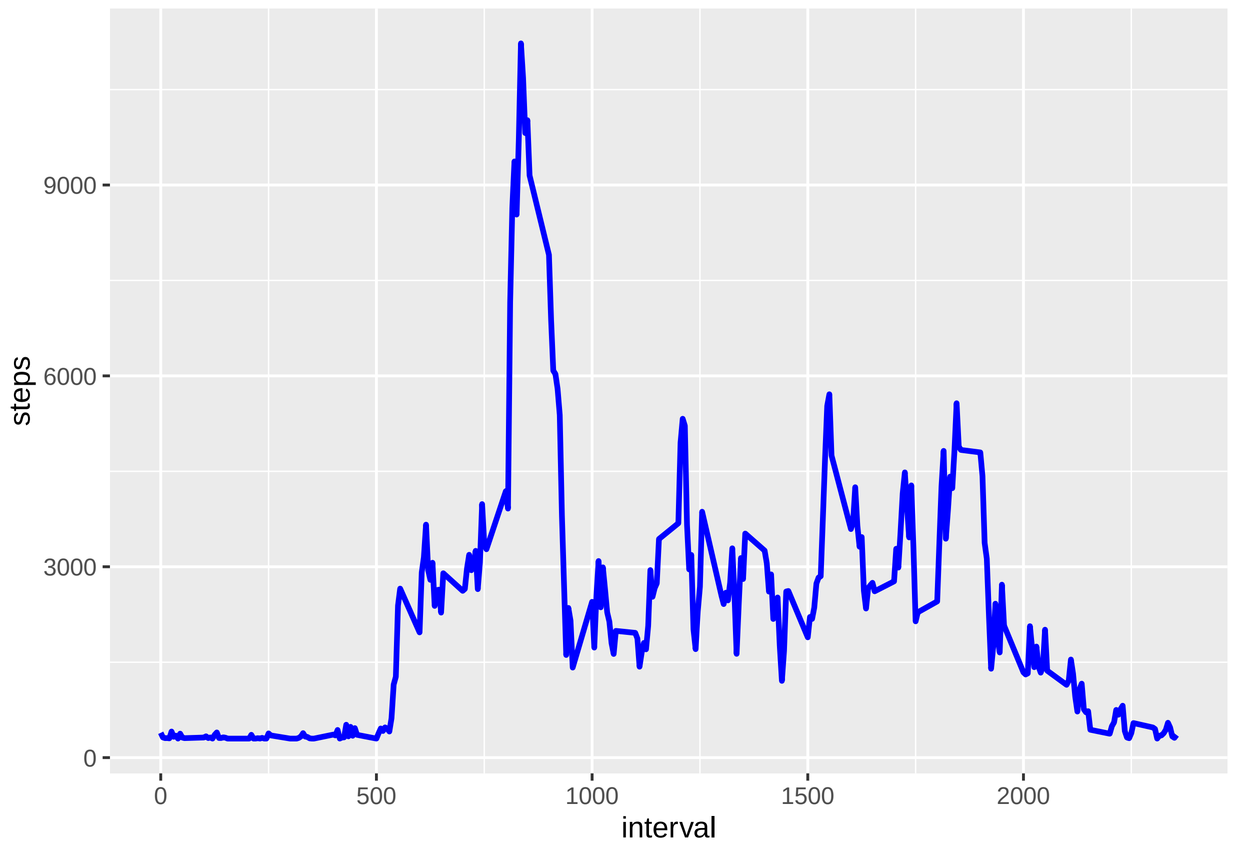
```
## [1] 37.3826
```

```
file2<-file
file2$steps[which(is.na(file2$steps))]=mean(spinterval$steps)
spd2<- aggregate(steps~date,data=file2,FUN=sum)
spinterval2<-aggregate(steps~interval,data=file2,FUN=sum)
ggplot(spd2,aes(x=steps))+geom_histogram(fill="lightgreen",col="black")
```

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



```
ggplot(spinterval2,aes(x=interval,y=steps))+geom_line(size=1,col="blue")
```



```
#now we will look into the difference of summarisies of all files
summary(file2)
```

```
##      steps                date          interval     
##  Min.   :  0.00   2012-10-01:  288   Min.   :   0.0  
##  1st Qu.:  0.00   2012-10-02:  288   1st Qu.: 588.8  
##  Median :  0.00   2012-10-03:  288   Median :1177.5  
##  Mean   : 37.38   2012-10-04:  288   Mean   :1177.5  
##  3rd Qu.: 37.38   2012-10-05:  288   3rd Qu.:1766.2  
##  Max.   :806.00   2012-10-06:  288   Max.   :2355.0  
##                   (Other)   :15840
```

```
summary(file1)
```

```
##      steps             date               interval    
##  Min.   :  0.00   Min.   :2012-10-02   0      :   53  
##  1st Qu.:  0.00   1st Qu.:2012-10-16   5      :   53  
##  Median :  0.00   Median :2012-10-29   10     :   53  
##  Mean   : 37.38   Mean   :2012-10-30   15     :   53  
##  3rd Qu.: 12.00   3rd Qu.:2012-11-16   20     :   53  
##  Max.   :806.00   Max.   :2012-11-29   25     :   53  
##                                        (Other):14946
```

```
summary(file)
```

```
##      steps                  date          interval     
##  Min.   :  0.00   2012-10-01:  288   Min.   :   0.0  
##  1st Qu.:  0.00   2012-10-02:  288   1st Qu.: 588.8  
##  Median :  0.00   2012-10-03:  288   Median :1177.5  
##  Mean   : 37.38   2012-10-04:  288   Mean   :1177.5  
##  3rd Qu.: 12.00   2012-10-05:  288   3rd Qu.:1766.2  
##  Max.   :806.00   2012-10-06:  288   Max.   :2355.0  
##  NA's   :2304     (Other)   :15840                   
```

```
#in the 3rd quad the values has changed from 12.00 to 37.38(mean) and no other effect
#mean and median for factor dates is also the same
mean(spd2$steps)
```

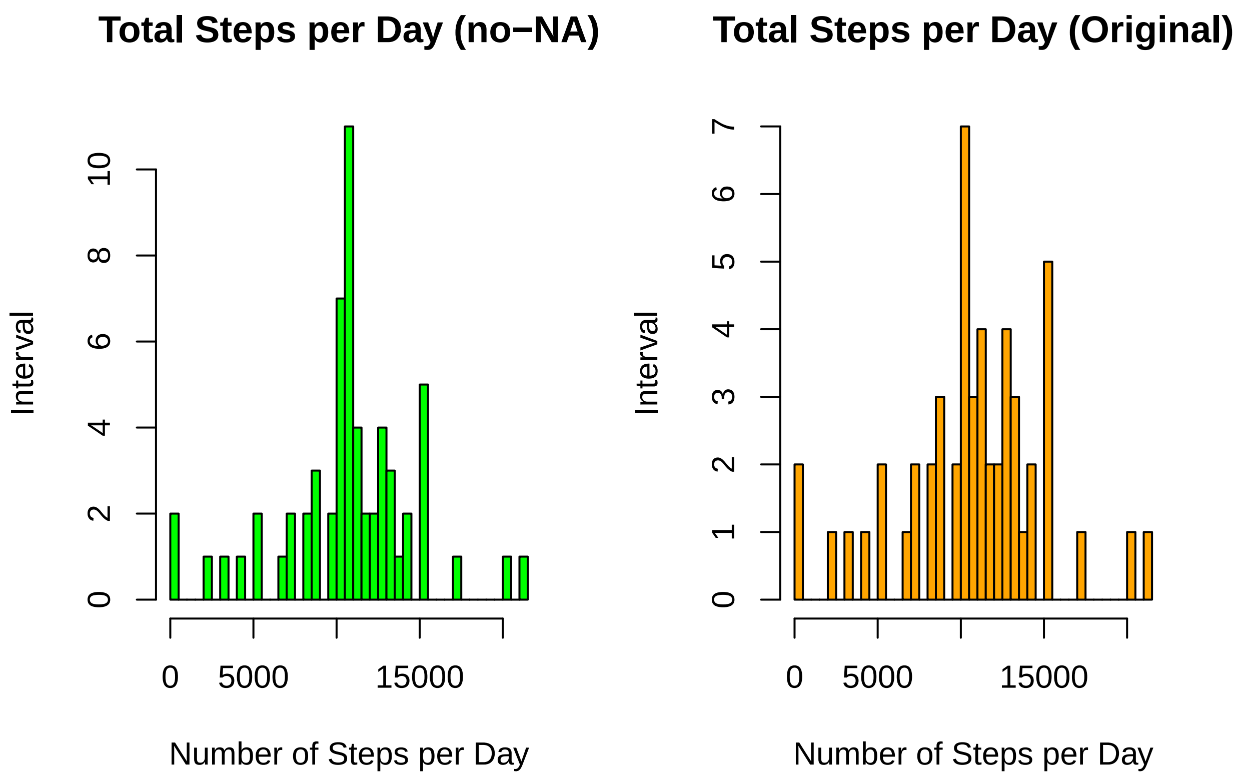
```
## [1] 10766.19
```

```
median(spd2$steps)
```

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

```
#plot btw thw both datasets (with no NA's ,with Na's)
par(mfrow=c(1,2))

hist(spd2$steps,
     main = "Total Steps per Day (no-NA)",
     xlab = "Number of Steps per Day",
     ylab = "Interval",
     col="green",
     breaks=50)
##Histogram with the orginal dataset
hist(spd1$steps,
     main="Total Steps per Day (Original)",
     xlab="Number of Steps per Day",
     ylab = "Interval",
     col="orange",
     breaks=50)
```

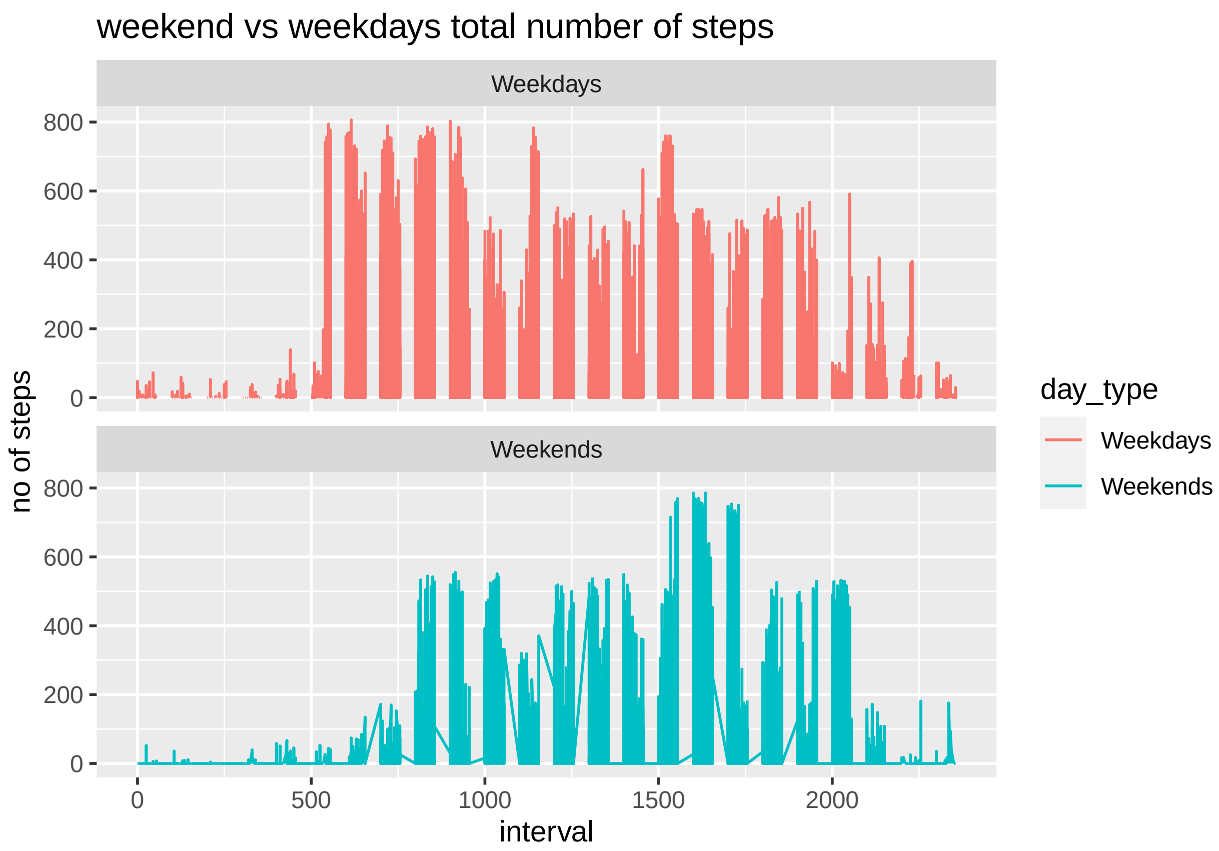


```
#lets assign the dates according to weekends nd weekdays
file3<- file%>%mutate(day_type= ifelse(weekdays(as.Date(file2$date,"%Y-%m-%d")) =="Saturday"
                                       |weekdays(as.Date(file2$date,"%Y-%m-%d")) =="Sunday","Weekends"
#file3 is having new factor column (file contains Na Values)
file3$day_type<- as.factor(file3$day_type)
head(file3)
```

```
##   steps       date interval day_type
## 1    NA 2012-10-01        0 Weekdays
## 2    NA 2012-10-01        5 Weekdays
## 3    NA 2012-10-01       10 Weekdays
## 4    NA 2012-10-01       15 Weekdays
## 5    NA 2012-10-01       20 Weekdays
## 6    NA 2012-10-01       25 Weekdays
```

```
#plotting the 5 min interval for both weekdays and weekends
ggplot(data=file3,aes(x=interval,y=steps,color=day_type))+geom_line()+labs(title = "weekend vs weekdays
```

```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```



```r
knitr::opts_chunk$set(echo = TRUE)
```