Week 2 Project - Reproducible Research

June Kieu

1/6/2020

Import Data

First of all I load the data into R and eliminate **NA** values.

```
setwd("C:/Users/June Kieu/Desktop/Studying R/Week2-Reproducible-Research-Project")
data <- read.csv("activity.csv",header = TRUE,sep = ",")
data1 <- data[!is.na(data$steps),]
```

Total Number of Steps per day

I used **sqldf** package to summarize the number of steps per day and use a histogram to have the visualization of data.

```
require(sqldf)
```

```
## Loading required package: sqldf
```

```
## Warning: package 'sqldf' was built under R version 3.6.2
```

```
## Loading required package: gsubfn
```

```
## Warning: package 'gsubfn' was built under R version 3.6.2
```

```
## Loading required package: proto
```

```
## Warning: package 'proto' was built under R version 3.6.2
```

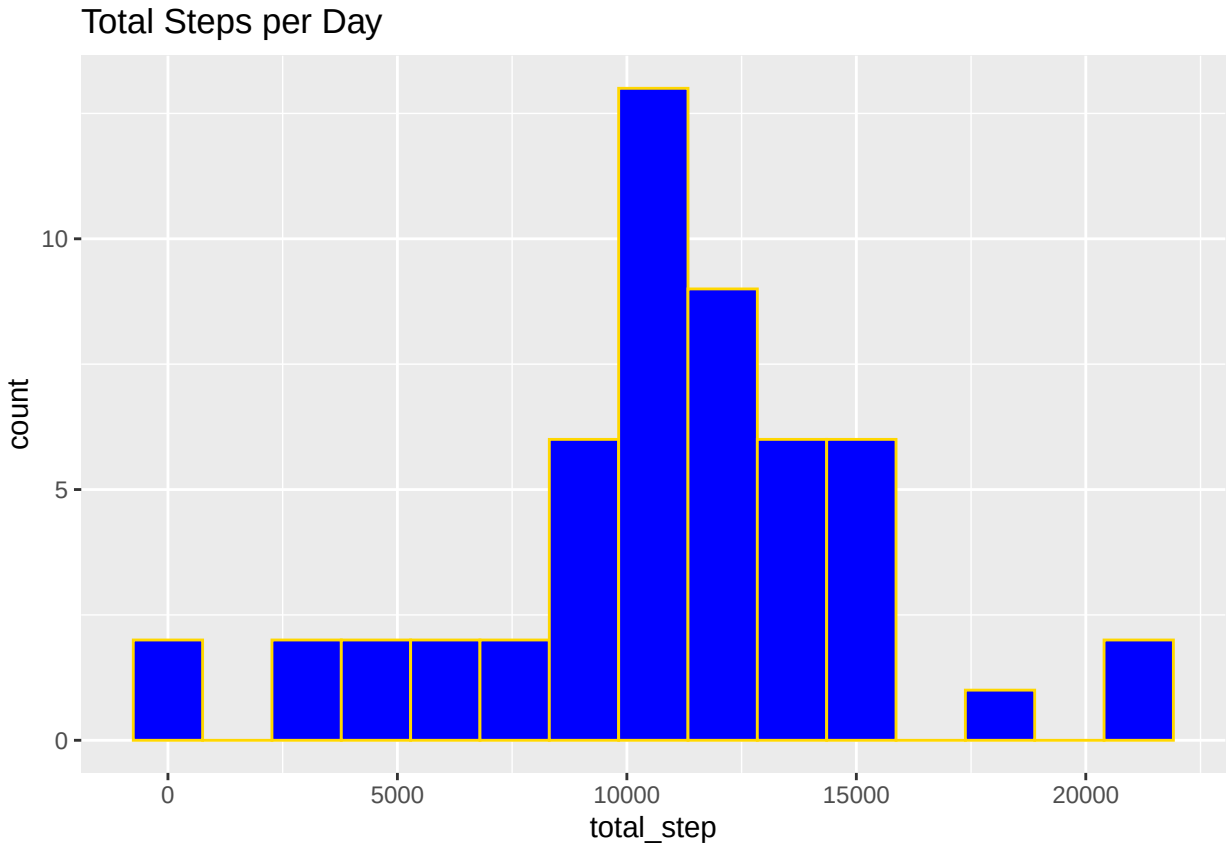
```
## Loading required package: RSQLite
```

```
## Warning: package 'RSQLite' was built under R version 3.6.2
```

```
step_sum <- sqldf("select distinct date,sum(steps) as total_step from data1 group by date")
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
ggplot(data=step_sum,aes(x=total_step)) + geom_histogram(bins = 15,fill="blue",color="gold")+
  ggtitle("Total Steps per Day")
```



Total Number of Steps Taken per Day's Mean and Median

The mean is 10766; the median is 10765.

```
step_meanmed <- sqldf("select sum(total_step)/count(date) as avg_step,
                      median(total_step) as med_step from step_sum")
step_meanmed$avg_step
```

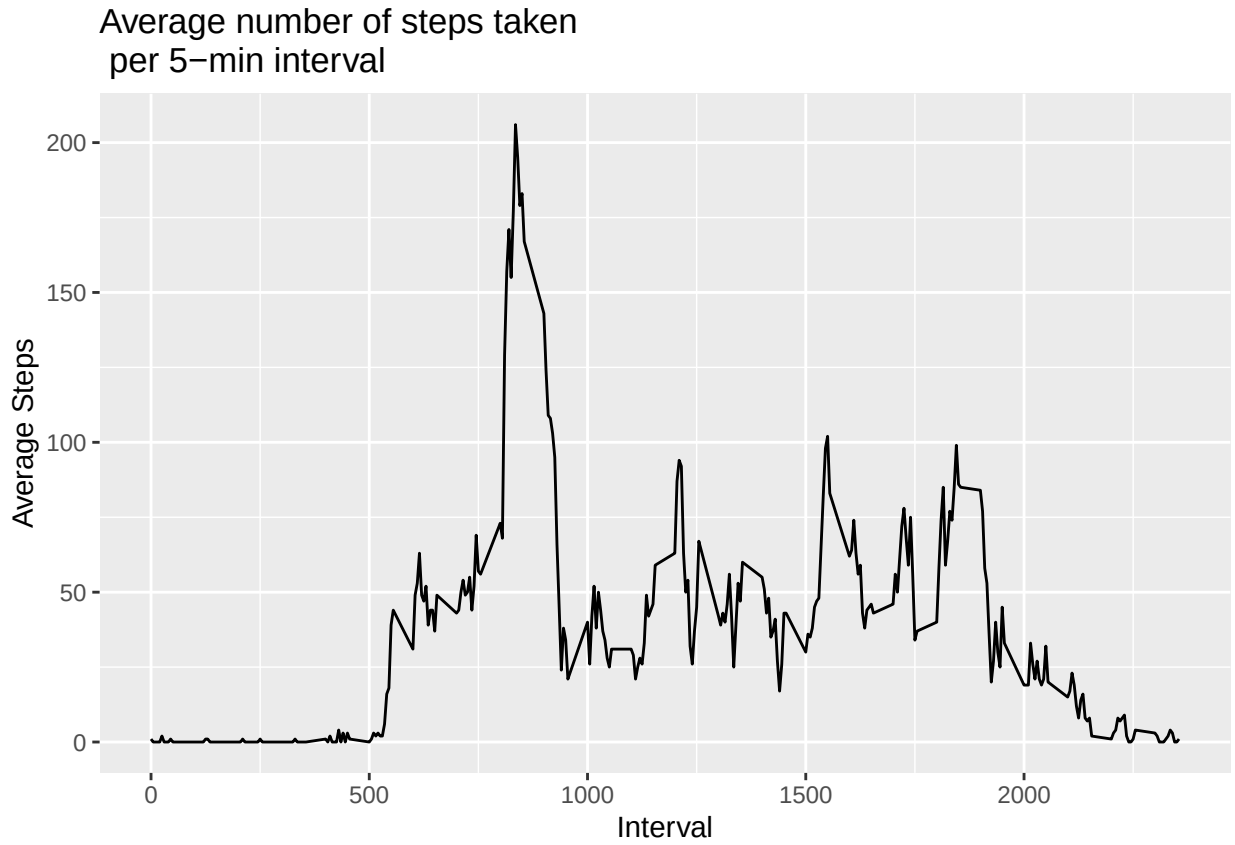
```
## [1] 10766
```

```
step_meanmed$med_step
```

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

Time Series Plot of the Average Number of Steps Taken

```
timeseries <- sqldf("select distinct interval,sum(steps)/count(interval)
                    as five_min_step from data1 group by interval
                    order by interval")
ggplot(timeseries,aes(x=interval,y=five_min_step))+geom_line()+
  labs(title = "Average number of steps taken \n per 5-min interval",
       x="Interval",y="Average Steps")
```



The maxium average steps taken per 5 minute interval is 206; in interval 835.

```
timeseries[which(timeseries$five_min_step==max(timeseries$five_min_step)),]
```

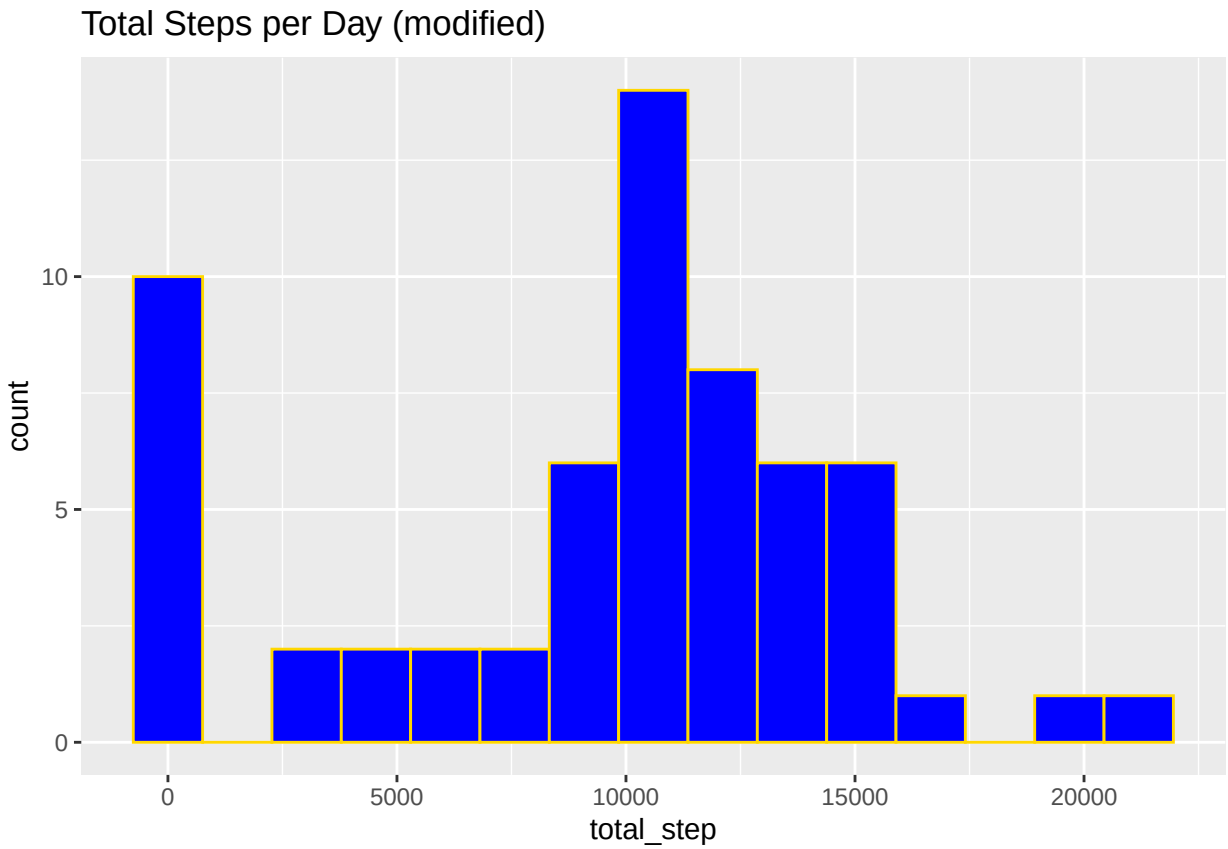
```
##     interval five_min_step
## 104      835           206
```

Dealing with NA value

I would replace NA values in Steps with median of Steps. Replacing NAs with mean of step would give NAs value of roughly 10000 steps. I believe that is too many steps.

The revised histogram is below:

```
require(sqldf)
step_sum1 <- sqldf("select distinct date,sum(steps1) as total_step from data group by date")
require(ggplot2)
ggplot(data=step_sum1,aes(x=total_step)) + geom_histogram(bins = 15,fill="blue",color="gold")+
  ggtitle("Total Steps per Day (modified)")
```



The revised mean and median are:

The revised mean is 9354 and the median is 10395.

```
step_meanmed1 <- sqldf("select sum(total_step)/count(date) as avg_step,
                       median(total_step) as med_step from step_sum1")
step_meanmed1$avg_step
```

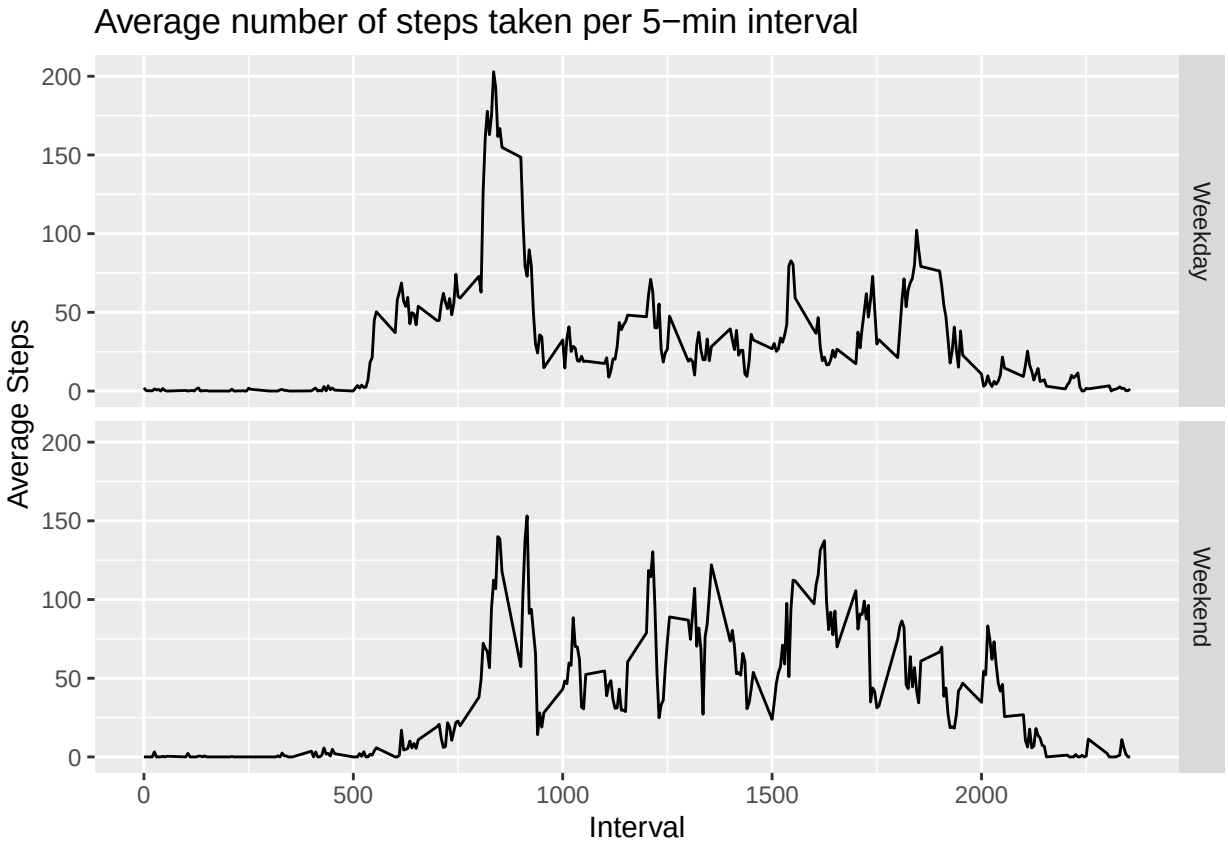
```
## [1] 9354.23
```

```
step_meanmed1$med_step
```

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

Comparison between Weekday and Weekend

```
data$date <- as.Date(data$date)
data$DOW <- weekdays(data$date)
data$WDWE <- ifelse(data$DOW=="Saturday" | data$DOW=="Sunday","Weekend","Weekday")
WDWE <- sqldf("select distinct WDWE,interval,sum(steps1)/count(interval) as five_min_steps
              from data group by WDWE,interval")
ggplot(WDWE,aes(x=interval,y=five_min_steps))+geom_line()+ facet_grid(WDWE~.)+
  labs(title = "Average number of steps taken per 5-min interval",
       x="Interval",y="Average Steps")
```


Average number of steps taken per 5−min interval
Average Steps
Interval
Weekday
Weekend
0
50
100
150
200
0
500
1000
1500
2000