RR_week2

01/06/2021

Course Project 1

What is mean total number of steps taken per day?

total number of steps taken per day

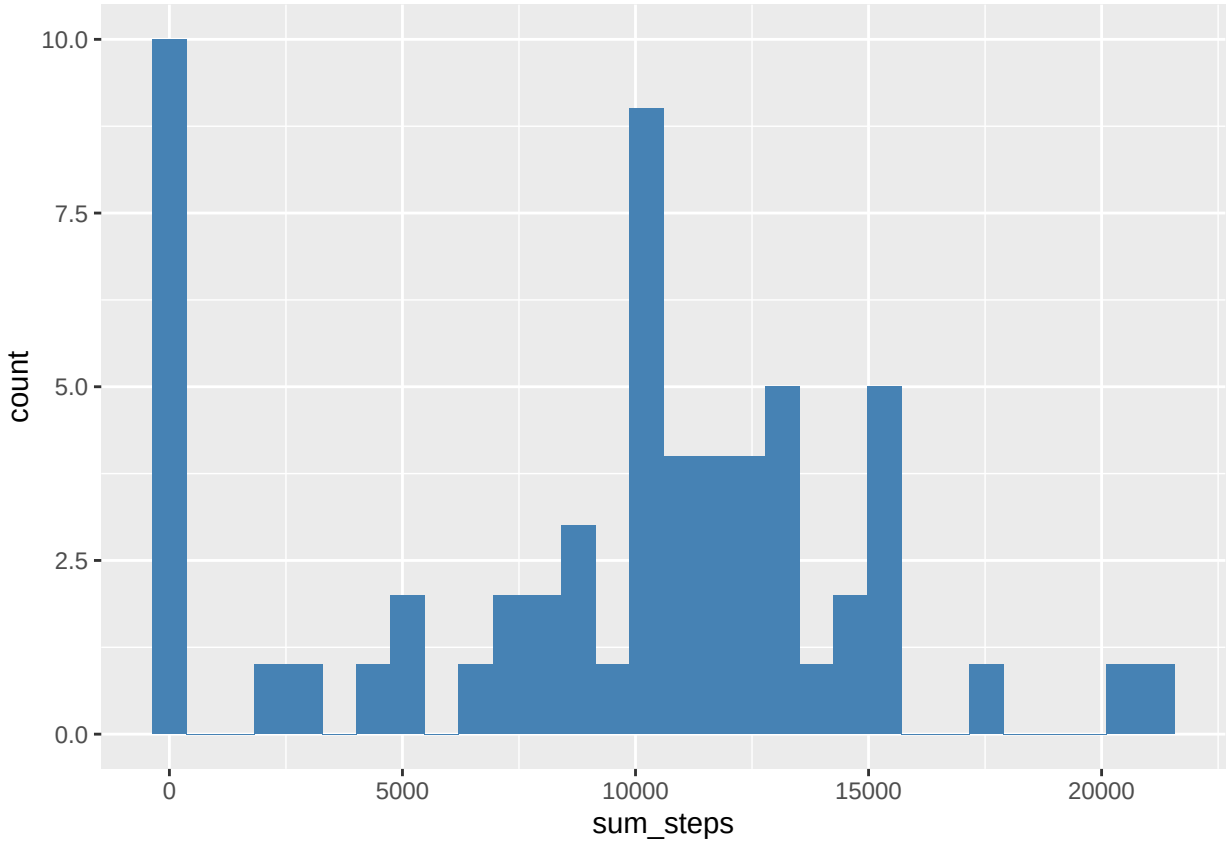
```
x<- sum(data$steps, na.rm = TRUE)
x
```

```
## [1] 570608
```

```
data_day<- data%>%
  group_by(date)%>%
  summarise(sum_steps = sum(steps, na.rm = T))
data_day$sum_steps<- as.numeric(data_day$sum_steps)

ggplot(data_day, aes(x = sum_steps )) + geom_histogram( fill = 'steel blue')
```

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



mean and median of the total number of steps taken per day

```
data%>%
  summarize(median = median(steps, na.rm = T), mean = mean(steps, na.rm = T))
```

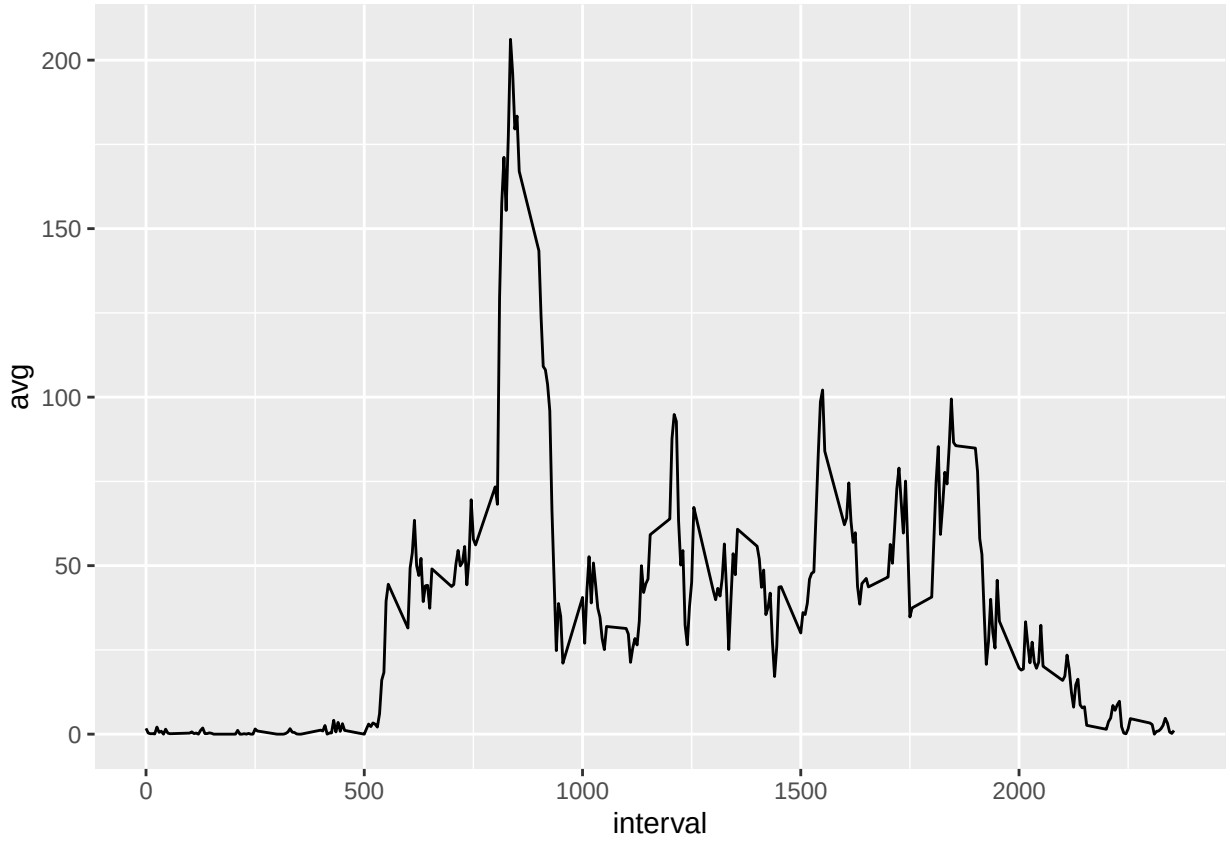
```
##   median    mean
## 1      0 37.3826
```

What is the average daily activity pattern?

```
data$date <- as.Date(data$date, format= "%Y-%m-%d")
```

```
data2<- data%>%
  group_by(interval)%>%
  summarise(avg = mean(steps, na.rm = T))

ggplot(data2, aes(interval, avg))+geom_line()
```



5 minute interval who contains the maximum number of steps

```
data2%>%
  slice(which.max(avg))
```

```
## # A tibble: 1 x 2
##   interval   avg
##      <int> <dbl>
## 1      835  206.
```

Imputing missing values

sum of missing data

```
sum(!complete.cases(data))
```

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

Filling the missing data thanks to the mice package.

```
act<- data.frame(data)
temp<- mice(act, m=5, method = "pmm", seed = 509)
```

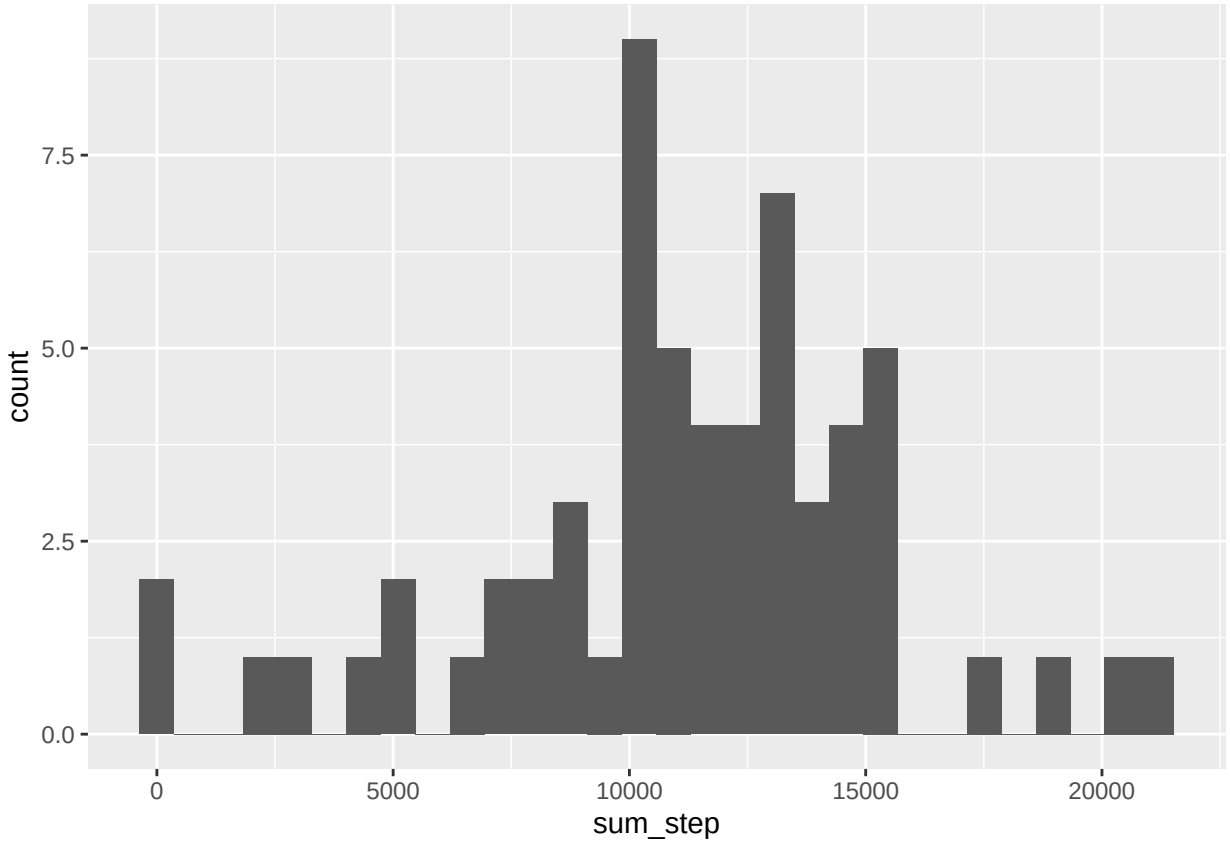
```
##
##  iter imp variable
##   1   1  steps
##   1   2  steps
##   1   3  steps
##   1   4  steps
##   1   5  steps
##   2   1  steps
##   2   2  steps
##   2   3  steps
##   2   4  steps
##   2   5  steps
##   3   1  steps
##   3   2  steps
##   3   3  steps
##   3   4  steps
##   3   5  steps
##   4   1  steps
##   4   2  steps
##   4   3  steps
##   4   4  steps
##   4   5  steps
##   5   1  steps
##   5   2  steps
##   5   3  steps
##   5   4  steps
##   5   5  steps
```

```
completed<- complete(temp, 1)
```

histogram with the completed dataset.

```
comp<- data.frame(completed)
completed2<- completed%>%
  group_by(date)%>%
  summarise(sum_step =  sum(steps))
ggplot(data = completed2, aes( x = sum_step))+geom_histogram()
```

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



calculation of mean and median

```
completed2%>%
  summarise(mean= mean(sum_step), median = median(sum_step))
```

```
## # A tibble: 1 x 2
##     mean median
##    <dbl>  <int>
## 1 11203.  11352
```

the mean and the median are now way higher than before.

#differences in activity patterns between weekdays and weekends

```
#create new column with the different day of the week
completed<- completed%>%
  mutate(days = weekdays(date))

#create another column and put either week day or week end depending on the day.

completed<- completed%>%
  mutate(week_end = if_else(days == c("lundi","mardi","mercredi","jeudi","vendredi"), "week_day", "week
```

```
## Warning in days == c("lundi", "mardi", "mercredi", "jeudi",
## "vendredi"): la taille d'un objet plus long n'est pas
## multiple de la taille d'un objet plus court
```

plot

```
plot<- completed%>%
  group_by(interval, week_end)%>%
  summarise(avg = mean(steps, na.rm = T))
```

```
## `summarise()` has grouped output by 'interval'. You can override using the `.groups` argument.
```

```
ggplot(plot, aes(interval, avg))+geom_line()+facet_wrap("week_end")
```