PA_ggeneral

Gabriel General

31-01-2021

Loading and preprocessing the data

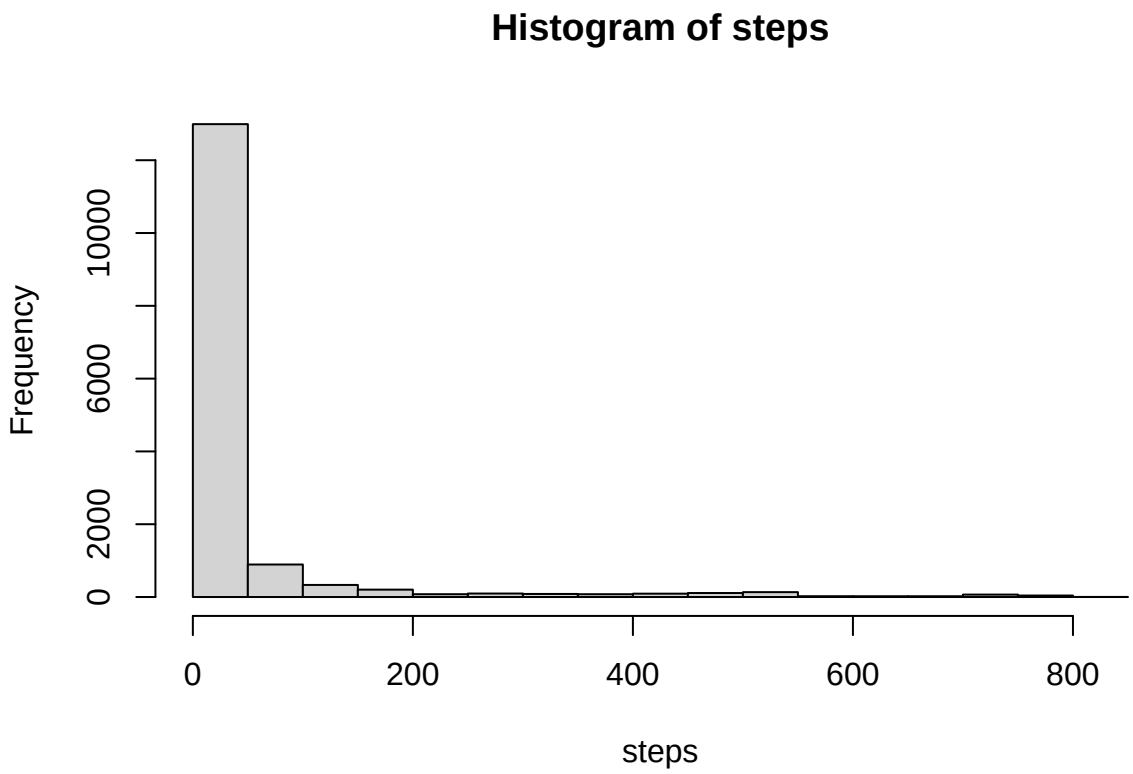
```
DataFileName <- "./activity.zip"
dataDir <- "./data"


if (!file.exists(dataDir)) {
  dir.create(dataDir)
  unzip(zipfile = DataFileName, exdir = dataDir)
}


data <- read.csv("./data/activity.csv", header = TRUE, sep = ",")
data1 <- na.omit(data)


total <- as.numeric(nrow(data))
valid <- as.numeric(nrow(data1))
percent <- trunc((total - valid)*100/total)
steps <- data1$steps
```

There are 13% of missing values ommited.

What is mean total number of steps taken per day?

1. Make a histogram of the total number of steps taken each day

```
hist(steps)
```



2. Calculate and report the mean and median total number of steps taken per day

```
library(xtable)
```

```
## Warning: package 'xtable' was built under R version 4.0.3
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

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

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

```
meanPerDay <- aggregate(steps ~ date, data1, mean)
names(meanPerDay)[2] <- "meanSteps"
medianPerDay <- aggregate(steps ~ date, data1, median)
names(medianPerDay)[2] <- "medianSteps"
summarySteps <- merge(meanPerDay,medianPerDay, all = TRUE)
tblMeanPerDay <- xtable(summarySteps)
print(tblMeanPerDay,type = "html")
```

date

meanSteps

medianSteps

1

2012-10-02

0.44

0.00

2

2012-10-03

39.42

0.00

3

2012-10-04

42.07

0.00

4

2012-10-05

46.16

0.00

5

2012-10-06

53.54

0.00

6

2012-10-07

38.25

0.00

7

2012-10-09

44.48

0.00

8
2012-10-10
34.38
0.00
9
2012-10-11
35.78
0.00
10
2012-10-12
60.35
0.00
11
2012-10-13
43.15
0.00
12
2012-10-14
52.42
0.00
13
2012-10-15
35.20
0.00
14
2012-10-16
52.38
0.00
15
2012-10-17
46.71
0.00
16
2012-10-18
34.92
0.00

17

2012-10-19

41.07

0.00

18

2012-10-20

36.09

0.00

19

2012-10-21

30.63

0.00

20

2012-10-22

46.74

0.00

21

2012-10-23

30.97

0.00

22

2012-10-24

29.01

0.00

23

2012-10-25

8.65

0.00

24

2012-10-26

23.53

0.00

25

2012-10-27

35.14

0.00

26

2012-10-28

39.78

0.00

27

2012-10-29

17.42

0.00

28

2012-10-30

34.09

0.00

29

2012-10-31

53.52

0.00

30

2012-11-02

36.81

0.00

31

2012-11-03

36.70

0.00

32

2012-11-05

36.25

0.00

33

2012-11-06

28.94

0.00

34

2012-11-07

44.73

0.00

35
2012-11-08
11.18
0.00
36
2012-11-11
43.78
0.00
37
2012-11-12
37.38
0.00
38
2012-11-13
25.47
0.00
39
2012-11-15
0.14
0.00
40
2012-11-16
18.89
0.00
41
2012-11-17
49.79
0.00
42
2012-11-18
52.47
0.00
43
2012-11-19
30.70
0.00

44
2012-11-20
15.53
0.00
45
2012-11-21
44.40
0.00
46
2012-11-22
70.93
0.00
47
2012-11-23
73.59
0.00
48
2012-11-24
50.27
0.00
49
2012-11-25
41.09
0.00
50
2012-11-26
38.76
0.00
51
2012-11-27
47.38
0.00
52
2012-11-28
35.36
0.00

53

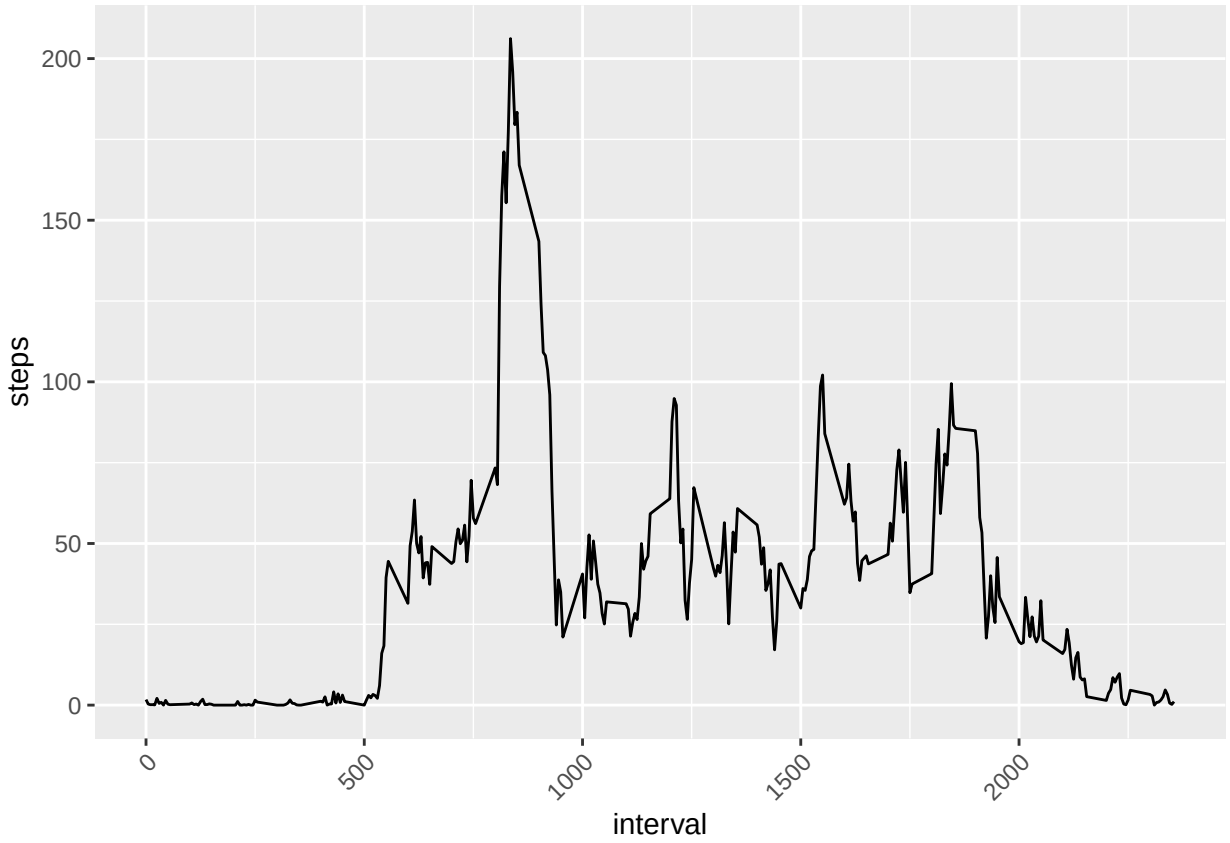
2012-11-29

24.47

0.00

What is the average daily activity pattern?

```
library(ggplot2)
library(dplyr)


meanStepInterval <- aggregate(steps ~ interval, data1, mean)
library(ggplot2)

ggplot(meanStepInterval, aes(interval, steps)) +
  geom_line() +
  theme(axis.text.x = element_text(angle=45, hjust=1, vjust = 1))
```




```
maxStepInterval <- meanStepInterval[which.max(meanStepInterval$steps),]
maxInterval <- maxStepInterval$interval
maxStep <- maxStepInterval$steps
```

2. Which 5-minute interval, on average across all the days in the dataset, contains the maximum number of steps?

A: The 5-minute interval that contains the maximum number of steps is 835 with an average of 206.1698113 steps

Imputing missing values

Imputing missing values Note that there are a number of days/intervals where there are missing values (coded as NA). The presence of missing days may introduce bias into some calculations or summaries of the data.

1. Calculate and report the total number of missing values in the dataset (i.e. the total number of rows with NAs)

```
total <- as.numeric(nrow(data))
valid <- as.numeric(nrow(data1))
missingValues <- total - valid
percent <- trunc((total - valid)*100/total)
steps <- data1$steps
```

There are 2304 missing values, a 13% of total values.

2. Devise a strategy for filling in all of the missing values in the dataset. The strategy does not need to be sophisticated. For example, you could use the mean/median for that day, or the mean for that 5-minute interval, etc.

```
# total <- as.numeric(nrow(data))
# valid <- as.numeric(nrow(data1))
# missingValues <- total - valid
# percent <- trunc((total - valid)*100/total)
# steps <- data1$steps




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