Fit_Project_MachineLearning

Nancy

2024-08-31

1. Loading and preprocessing the data:

```
download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip", "activity_data.zip
unzip("activity_data.zip")
act <- read.csv("activity.csv", stringsAsFactors = F)
```

Install ggplot2 library

Convert the dates to date format

```
act$date <- as.Date(act$date)
```

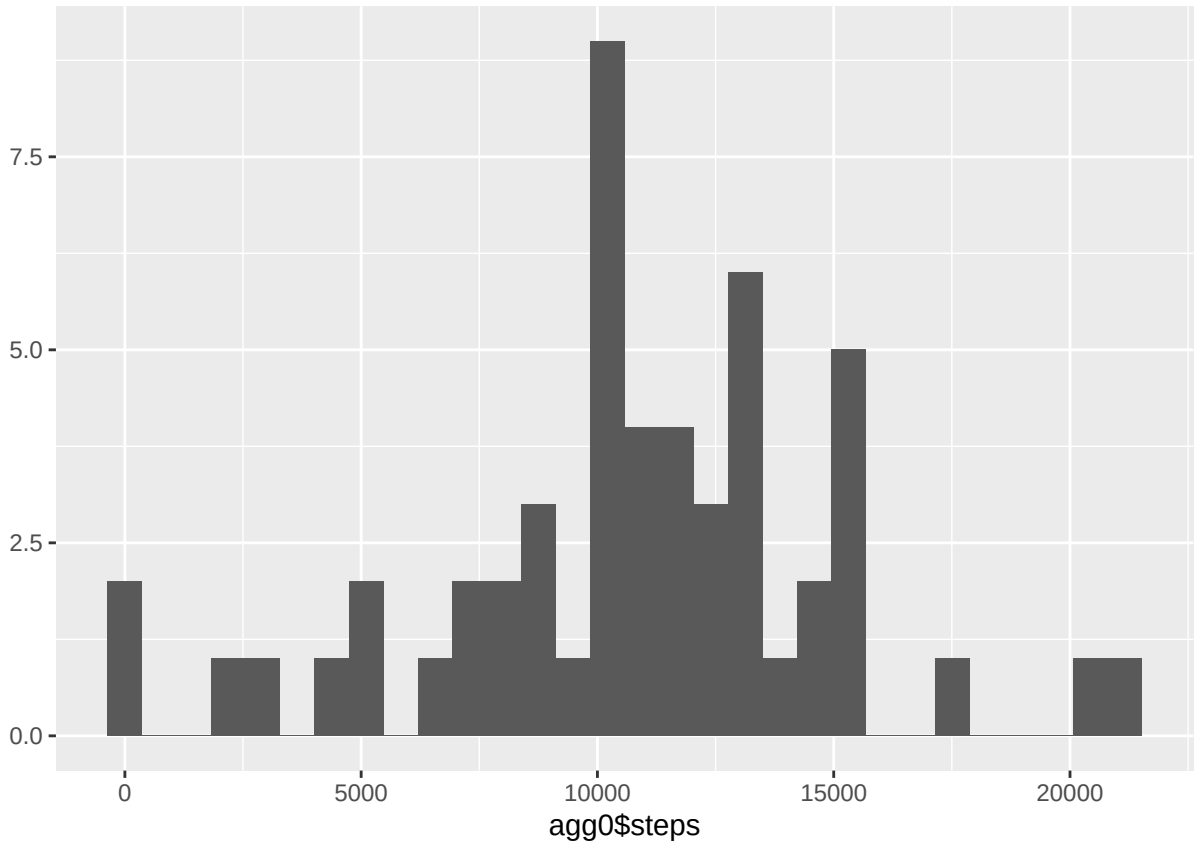
2. What is mean total number of steps taken per day?

```
agg0 <- aggregate(steps ~ date, FUN = sum, data = act)
qplot(agg0$steps)
```

Histogram:

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

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



mean and median: **Mean** of steps taken each day:

```
(agg1 <- aggregate(steps ~ date, FUN = mean, data = act))
```

```
##          date      steps
## 1  2012-10-02  0.4375000
## 2  2012-10-03 39.4166667
## 3  2012-10-04 42.0694444
## 4  2012-10-05 46.1597222
## 5  2012-10-06 53.5416667
## 6  2012-10-07 38.2465278
## 7  2012-10-09 44.4826389
## 8  2012-10-10 34.3750000
## 9  2012-10-11 35.7777778
## 10 2012-10-12 60.3541667
## 11 2012-10-13 43.1458333
## 12 2012-10-14 52.4236111
## 13 2012-10-15 35.2048611
## 14 2012-10-16 52.3750000
## 15 2012-10-17 46.7083333
## 16 2012-10-18 34.9166667
## 17 2012-10-19 41.0729167
## 18 2012-10-20 36.0937500
## 19 2012-10-21 30.6284722
## 20 2012-10-22 46.7361111
## 21 2012-10-23 30.9652778
## 22 2012-10-24 29.0104167
```

```
## 23 2012-10-25  8.6527778
## 24 2012-10-26 23.5347222
## 25 2012-10-27 35.1354167
## 26 2012-10-28 39.7847222
## 27 2012-10-29 17.4236111
## 28 2012-10-30 34.0937500
## 29 2012-10-31 53.5208333
## 30 2012-11-02 36.8055556
## 31 2012-11-03 36.7048611
## 32 2012-11-05 36.2465278
## 33 2012-11-06 28.9375000
## 34 2012-11-07 44.7326389
## 35 2012-11-08 11.1770833
## 36 2012-11-11 43.7777778
## 37 2012-11-12 37.3784722
## 38 2012-11-13 25.4722222
## 39 2012-11-15  0.1423611
## 40 2012-11-16 18.8923611
## 41 2012-11-17 49.7881944
## 42 2012-11-18 52.4652778
## 43 2012-11-19 30.6979167
## 44 2012-11-20 15.5277778
## 45 2012-11-21 44.3993056
## 46 2012-11-22 70.9270833
## 47 2012-11-23 73.5902778
## 48 2012-11-24 50.2708333
## 49 2012-11-25 41.0902778
## 50 2012-11-26 38.7569444
## 51 2012-11-27 47.3819444
## 52 2012-11-28 35.3576389
## 53 2012-11-29 24.4687500
```

Median number of steps taken each day:

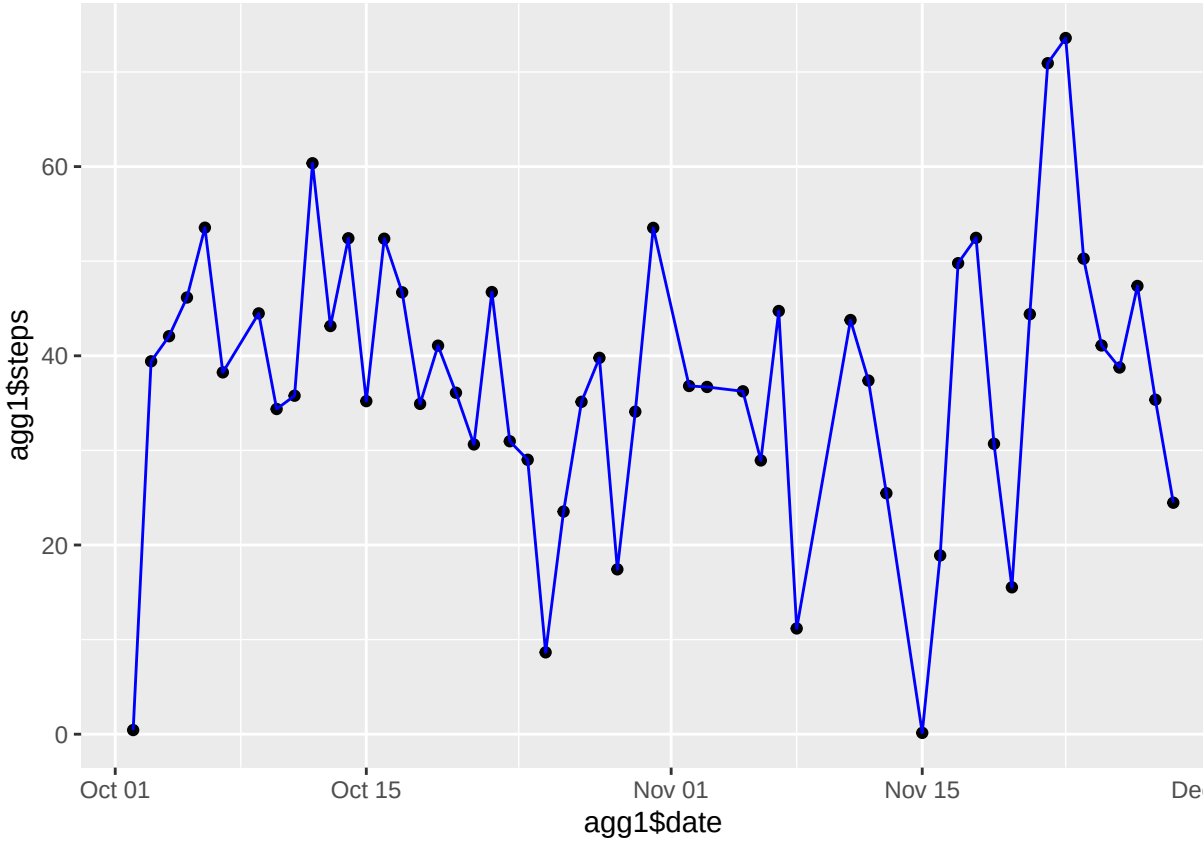
```
(agg2 <- aggregate(steps ~ date, FUN = median, data = act))
```

```
##          date steps
## 1  2012-10-02     0
## 2  2012-10-03     0
## 3  2012-10-04     0
## 4  2012-10-05     0
## 5  2012-10-06     0
## 6  2012-10-07     0
## 7  2012-10-09     0
## 8  2012-10-10     0
## 9  2012-10-11     0
## 10 2012-10-12     0
## 11 2012-10-13     0
## 12 2012-10-14     0
## 13 2012-10-15     0
## 14 2012-10-16     0
## 15 2012-10-17     0
## 16 2012-10-18     0
## 17 2012-10-19     0
## 18 2012-10-20     0
```

```
## 19 2012-10-21     0
## 20 2012-10-22     0
## 21 2012-10-23     0
## 22 2012-10-24     0
## 23 2012-10-25     0
## 24 2012-10-26     0
## 25 2012-10-27     0
## 26 2012-10-28     0
## 27 2012-10-29     0
## 28 2012-10-30     0
## 29 2012-10-31     0
## 30 2012-11-02     0
## 31 2012-11-03     0
## 32 2012-11-05     0
## 33 2012-11-06     0
## 34 2012-11-07     0
## 35 2012-11-08     0
## 36 2012-11-11     0
## 37 2012-11-12     0
## 38 2012-11-13     0
## 39 2012-11-15     0
## 40 2012-11-16     0
## 41 2012-11-17     0
## 42 2012-11-18     0
## 43 2012-11-19     0
## 44 2012-11-20     0
## 45 2012-11-21     0
## 46 2012-11-22     0
## 47 2012-11-23     0
## 48 2012-11-24     0
## 49 2012-11-25     0
## 50 2012-11-26     0
## 51 2012-11-27     0
## 52 2012-11-28     0
## 53 2012-11-29     0
```

3. What is the average daily activity pattern?

```
qplot(agg1$date, agg1$steps) +
    geom_line(aes(x = agg1$date, y = agg1$steps), colour = "blue")
```



Time series plot:

```
loc <- which(act$steps == max(na.omit(act$steps)))
act[loc,]
```

The 5-minute interval contains the max number of steps:

```
##       steps       date interval
## 16492   806 2012-11-27      615
```

4. Imputing missing values:

In this case **mean imputation** is used

```
avg <- mean(na.omit(act$steps))
avg <- floor(avg)   # round down
```

assign avg to all NA in a new dataset:

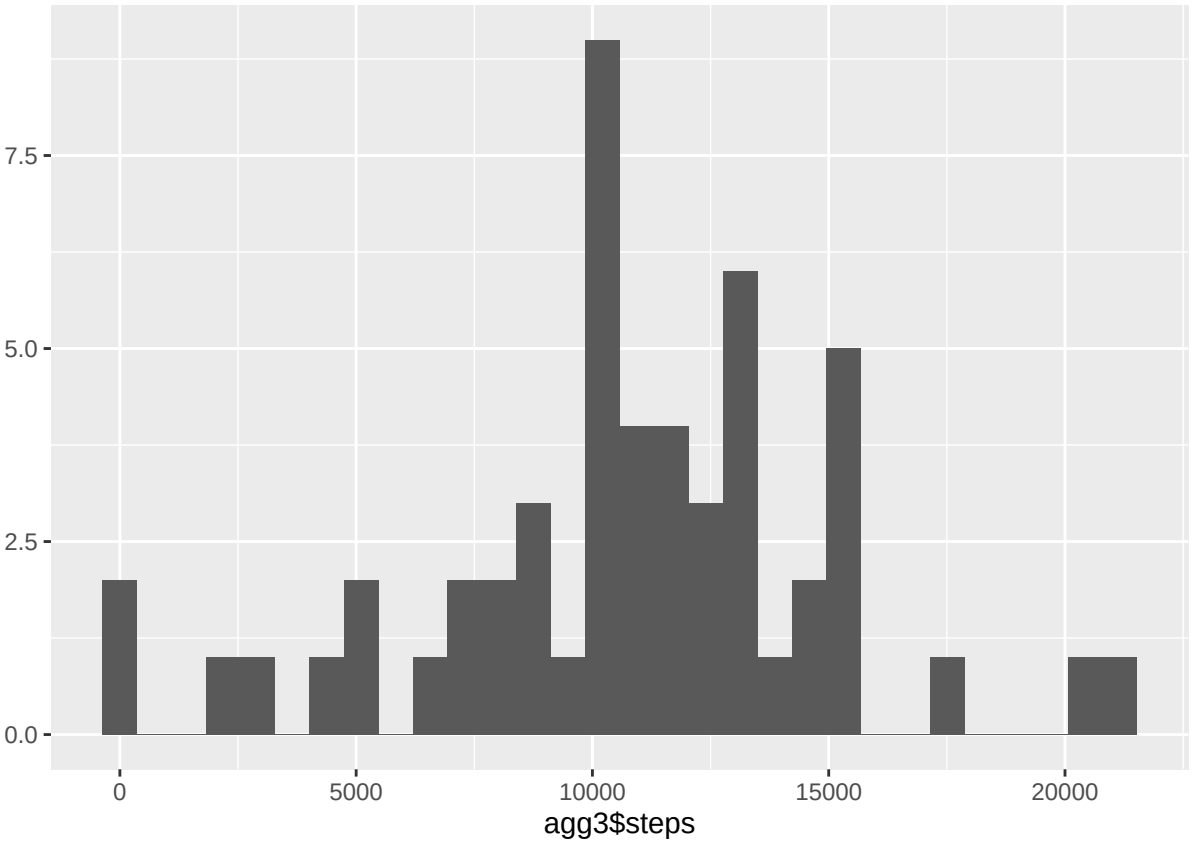
```
imputed <- act

for (i in 1:length(imputed$steps)) {
    if (is.na(imputed$steps[i])) {
        imputed$steps[i] <- avg
    }
}
```

```
agg3 <- aggregate(steps ~ date, FUN = sum, data = act)
qplot(agg3$steps)
```

Histogram after missing values are imputed:

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



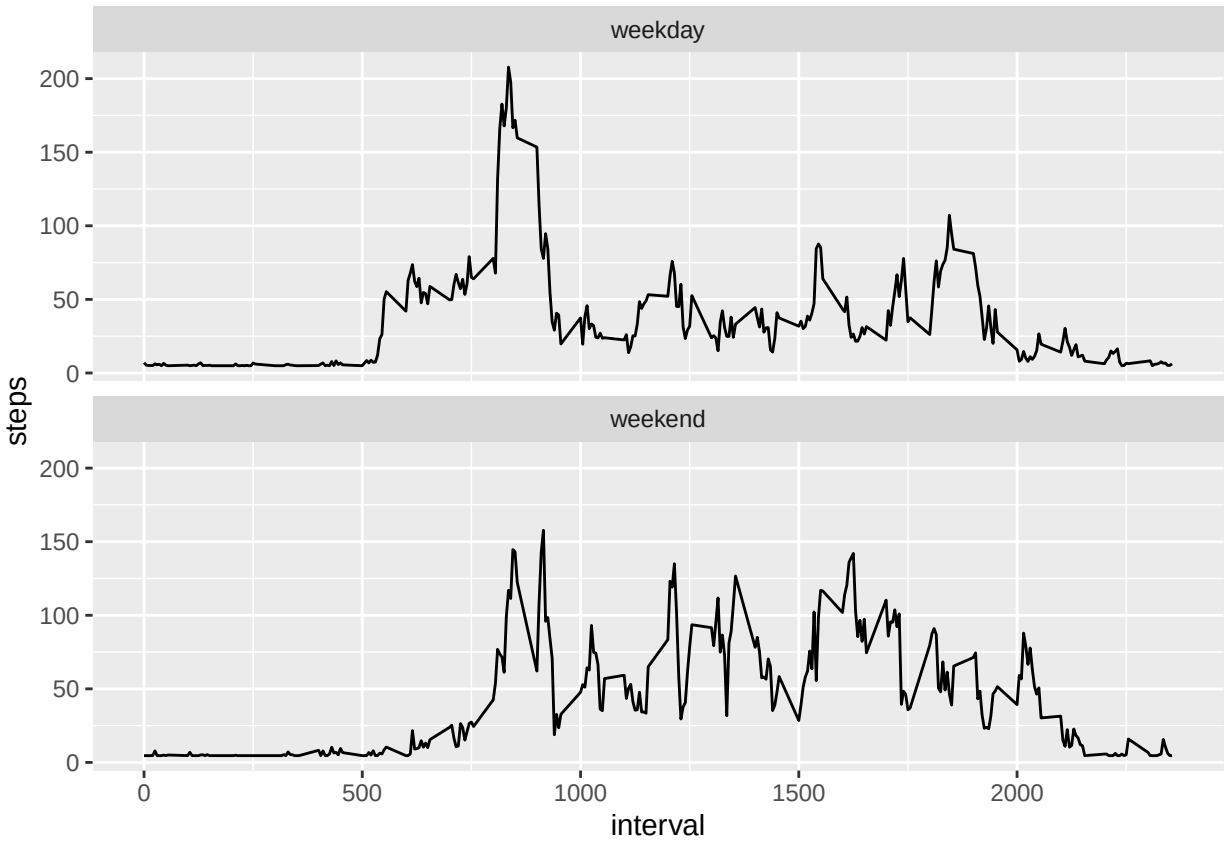
5. Differences in activity patterns between weekdays and weekends:

Panel plot:

```
loc <- which(weekdays(imputed$date) %in% c("Saturday", "Sunday"))
weekday <- imputed[-loc,]
weekend <- imputed[loc,]
agg4 <- aggregate(steps ~ interval, data = weekday, FUN = mean)
agg5 <- aggregate(steps ~ interval, data = weekend, FUN = mean)

agg4$w <- rep("weekday", length(agg4[,1]))
agg5$w <- rep("weekend", length(agg5[,1]))
agg4 <- rbind(agg4, agg5)

ggplot(data = agg4, mapping = aes(x = interval, y = steps)) +
    geom_line() + facet_wrap(~w, nrow = 2)
```


weekday
weekend
steps
interval
0
50
100
150
200
0
500
1000
1500
2000