Submission

2022-09-28

Code for reading in the dataset and/or processing the data

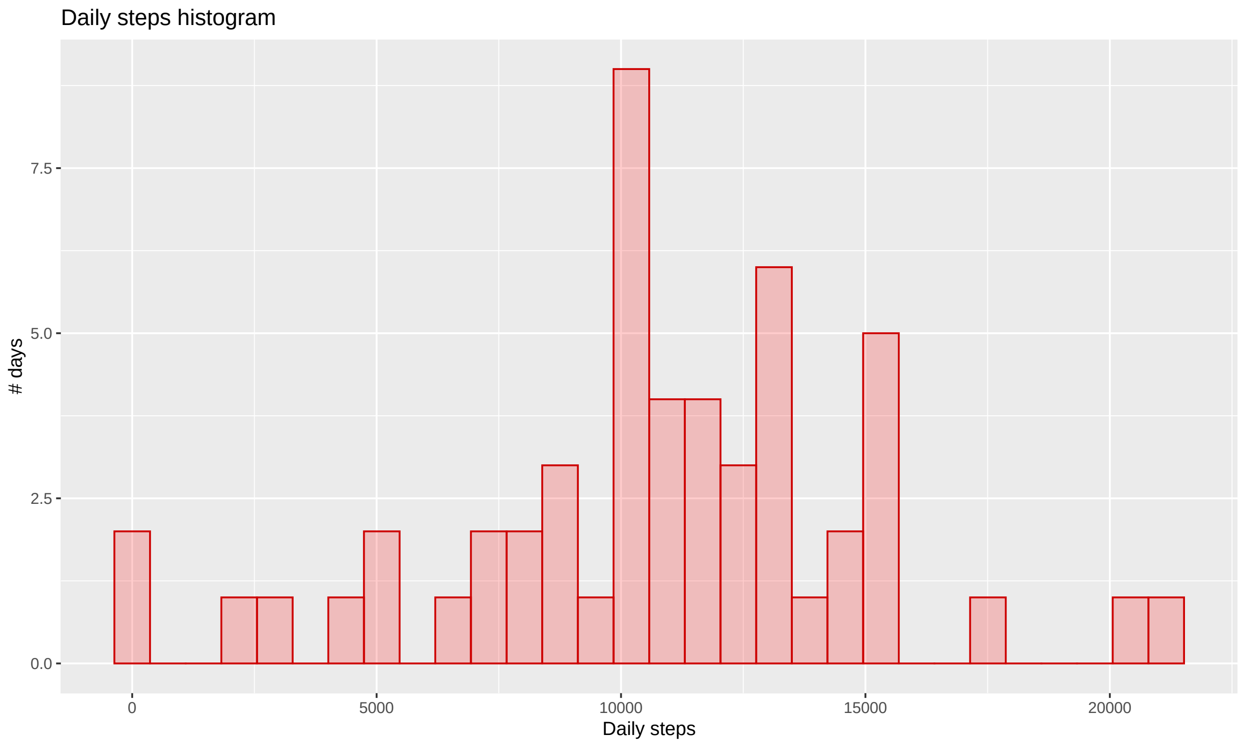
```
df <- read.csv("activity.csv")
```

Histogram of the total number of steps taken each day

```
library(ggplot2)
library(dplyr)

tdf <- df %>%
    filter(!is.na(steps)) %>%
    group_by(date) %>%
    summarise(t_steps = sum(steps))

his <- ggplot(tdf, aes(x = t_steps)) + geom_histogram(color = "red3", fill = "red",
    alpha = 0.2) + labs(title = "Daily steps histogram", y = "# days", x = "Daily steps")
his
```



Mean and median number of steps taken each day

```
tdf %>%
    select(t_steps) %>%
    summarise(mean = mean(t_steps), median = median(t_steps))
```

```
## # A tibble: 1 x 2
##      mean median
##     <dbl>  <int>
## 1 10766.  10765
```

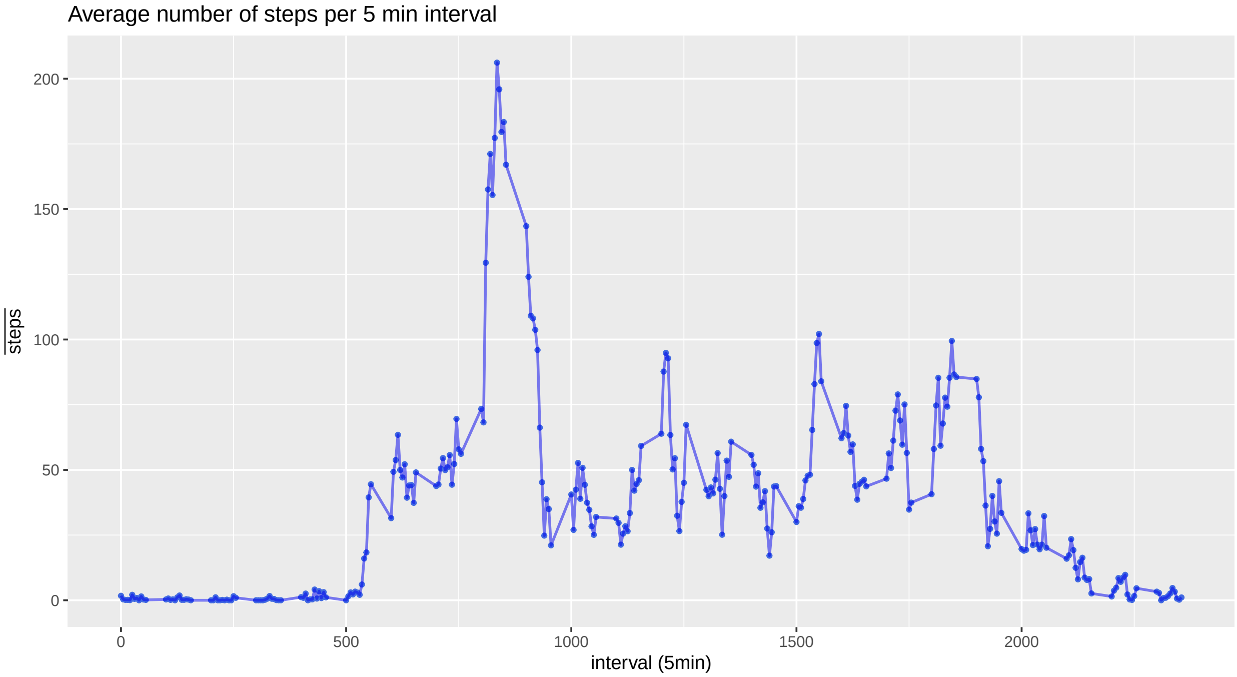
Time series plot of the average number of steps taken

```
library(latex2exp)

ts_df <- df %>%
    filter(!is.na(steps)) %>%
    group_by(interval) %>%
    summarise(avg_steps = mean(steps)) %>%
    mutate(pretty_interval = paste(as.character(interval%/%100),
        "h", as.character(interval%%100), sep = ""))

ts <- ggplot(ts_df, aes(x = interval, y = avg_steps)) +
    geom_point(size = 1, color = "royalblue", fill = "blue") +
    geom_line(size = 0.75, alpha = 0.5, color = "blue2") +
    labs(title = "Average number of steps per 5 min interval",
        x = "interval (5min)") + ylab(TeX("$\\bar{steps}$"))

ts
```



## The 5-minute interval that, on average, contains the maximum number of steps

```
ts_df %>%
    top_n(1, avg_steps) %>%
    select(pretty_interval)
```

```
## # A tibble: 1 x 1
##   pretty_interval
##   <chr>
## 1 8h35
```

Code to describe and show a strategy for imputing missing data

```
print(paste("# rows with missing values:", as.character(df %>%
    filter(is.na(steps)) %>%
    nrow())))
```

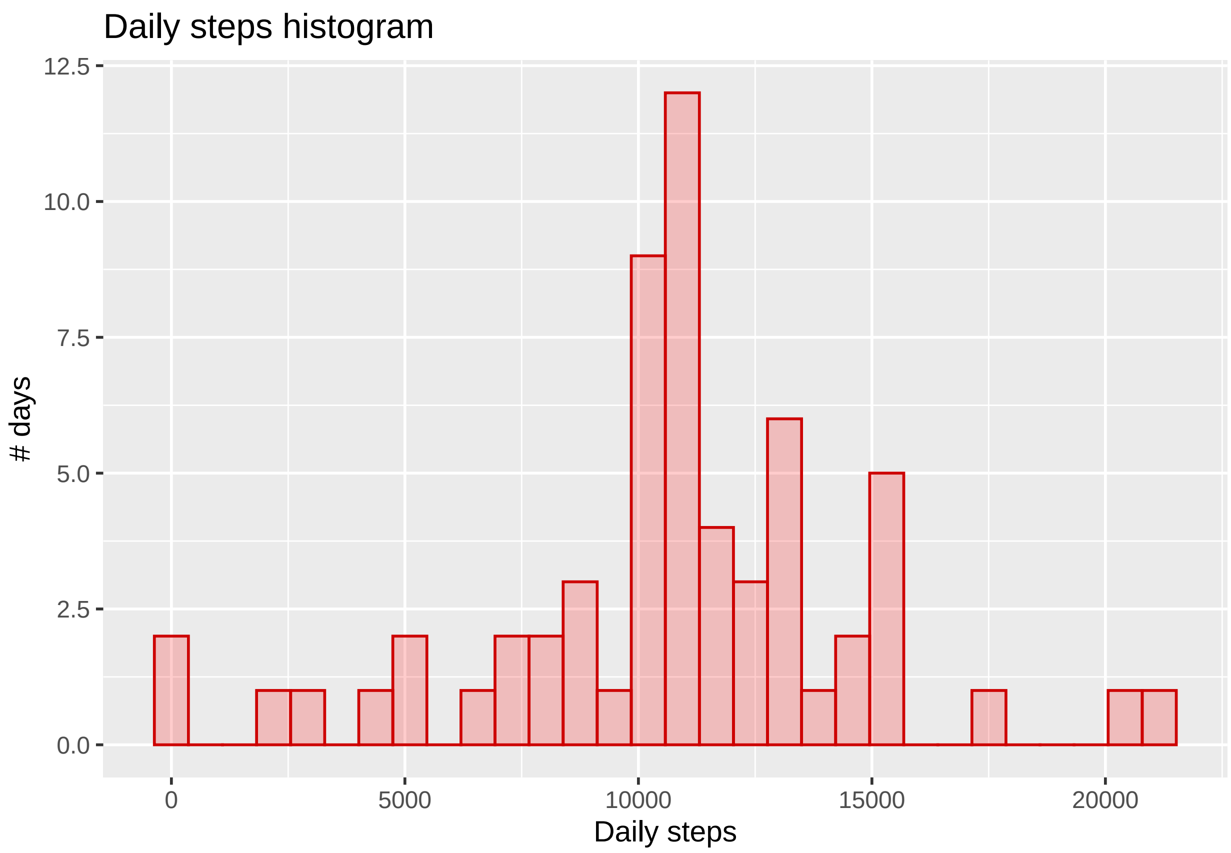
```
## [1] "# rows with missing values: 2304"
```

```
# Imputation using the mean value of the 5min interval
df <- df %>%
    left_join(ts_df, by = "interval") %>%
    mutate(up_steps = if_else(is.na(steps), avg_steps, as.numeric(steps)))
```

Histogram of the total number of steps taken each day after missing values are imputed

```
# Histogram of the updated dataset
up_tdf <- df %>%
    group_by(date) %>%
    summarise(t_steps = sum(up_steps))

his <- ggplot(up_tdf, aes(x = t_steps)) + geom_histogram(color = "red3", fill = "red",
    alpha = 0.2) + labs(title = "Daily steps histogram", y = "# days", x = "Daily steps")
his
```



```
# Mean/Median total steps / day after imputation
up_tdf %>%
    select(t_steps) %>%
    summarise(mean = mean(t_steps), median = median(t_steps))
```

```
## # A tibble: 1 x 2
##     mean median
##    <dbl>  <dbl>
## 1 10766. 10766.
```

The mean did not change, however imputing missing values with the mean value of each 5min interval moved the median in the direction of the former, thereby explaining the similar values.

Panel plot comparing the average number of steps taken per 5-minute interval across weekdays and weekends

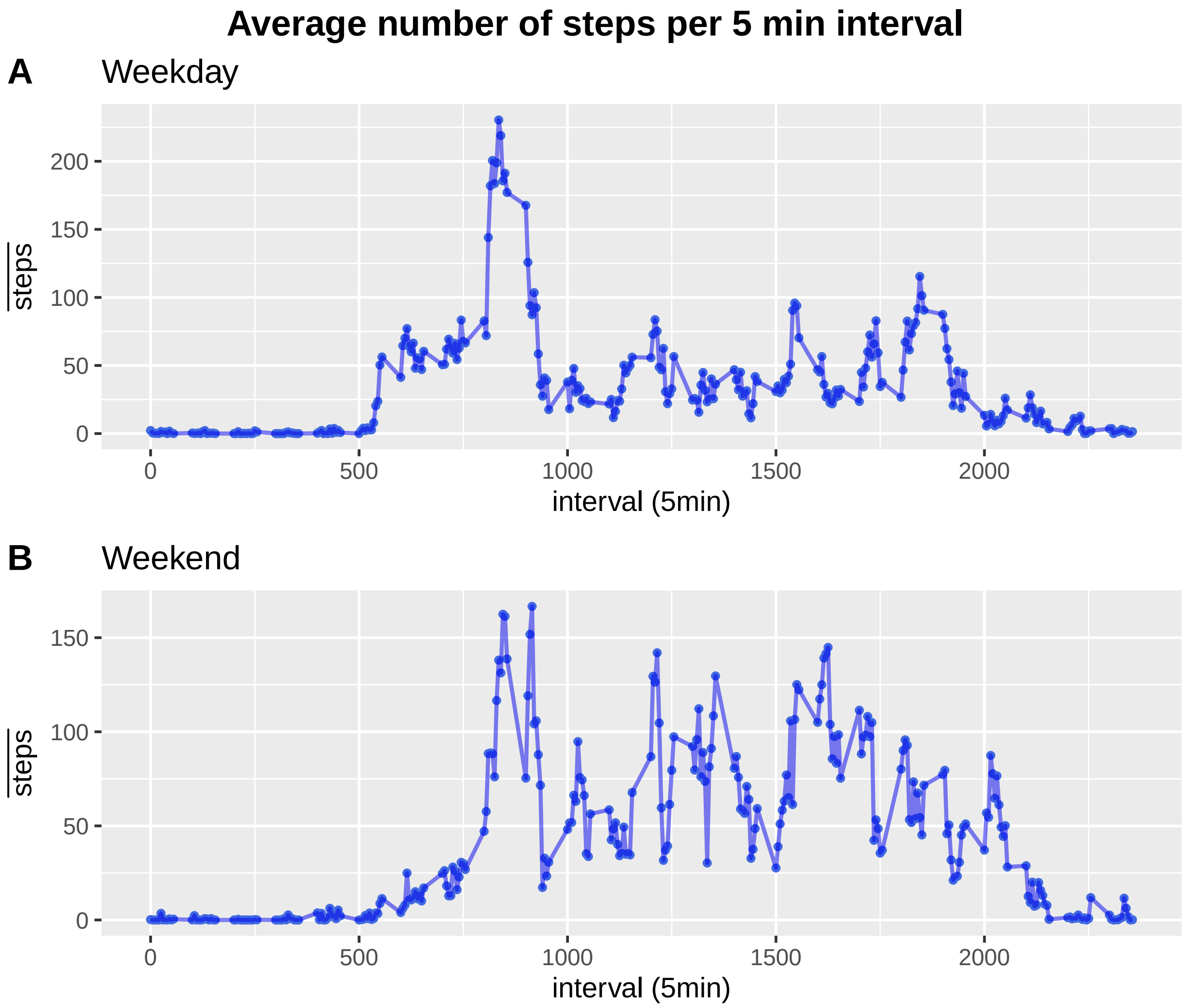
```
library(ggpubr)

up_ts_df <- df %>%
    mutate(isweekend = weekdays(as.Date(date)) %in%
        c("Saturday", c("Sunday"))) %>%
    group_by(interval, isweekend) %>%
    summarise(avg_steps = mean(up_steps))
```

```
wendts <- ggplot(up_ts_df %>%
    filter(isweekend == TRUE), aes(x = interval, y = avg_steps)) +
    geom_point(size = 1, color = "royalblue", fill = "blue") +
    geom_line(size = 0.75, alpha = 0.5, color = "blue2") +
    labs(title = "Weekend", x = "interval (5min)") +
    ylab(TeX("$\\bar{steps}$"))
wdayts <- ggplot(up_ts_df %>%
    filter(isweekend == FALSE), aes(x = interval, y = avg_steps)) +
    geom_point(size = 1, color = "royalblue", fill = "blue") +
    geom_line(size = 0.75, alpha = 0.5, color = "blue2") +
    labs(title = "Weekday", x = "interval (5min)") +
    ylab(TeX("$\\bar{steps}$"))


p <- ggarrange(wdayts, wendts, ncol = 1, nrow = 2,
    labels = c("A", "B"), common.legend = TRUE)

annotate_figure(p, top = text_grob("Average number of steps per 5 min interval",
    face = "bold", size = 14))
```


Average number of steps per 5 min interval
A
Weekday
$\overline{steps}$
interval (5min)
B
Weekend
$\overline{steps}$
interval (5min)