Assignment: Course Project 1

Andrea Eoli

2022-09-14

Loading and preprocessing the data

```
df <- tibble(read.csv(unzip("activity.zip")))
str(df)
```

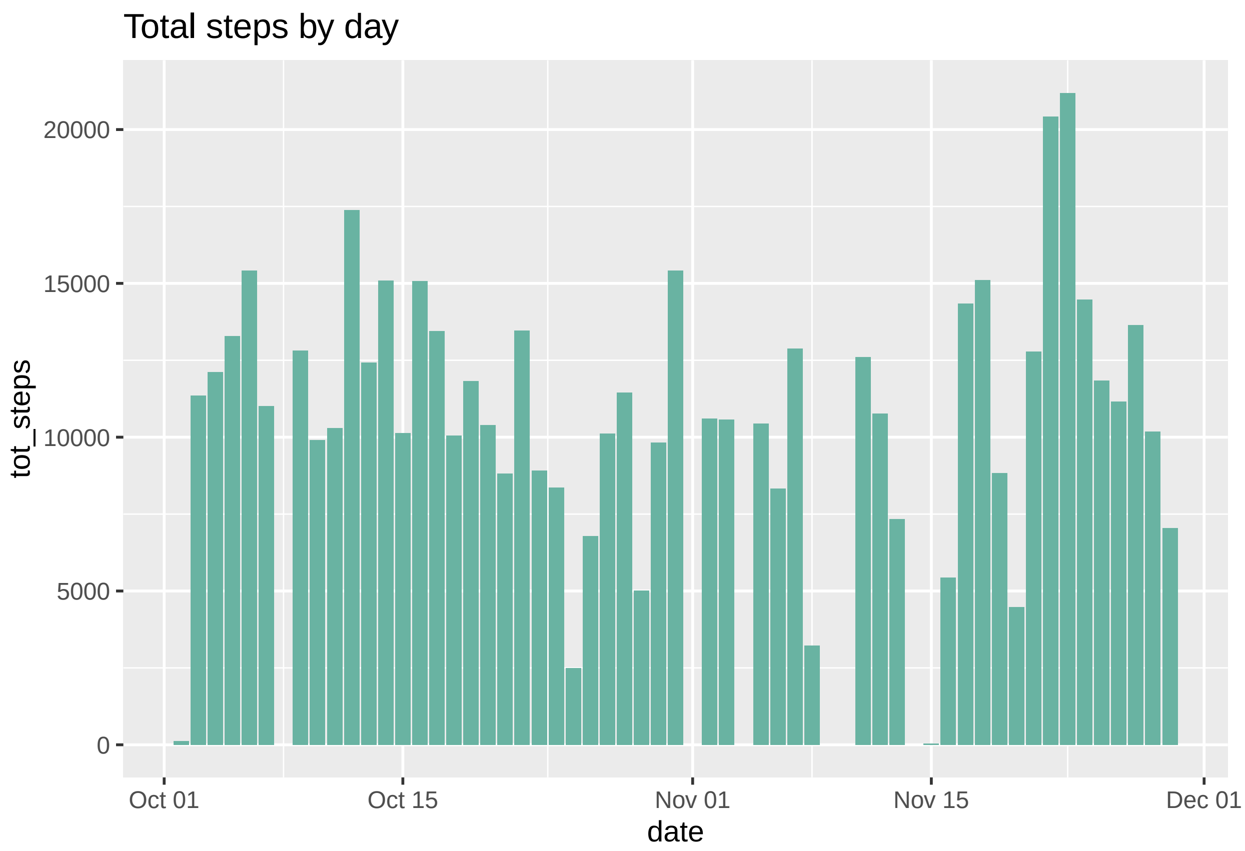
```
## tibble [17,568 x 3] (S3: tbl_df/tbl/data.frame)
##  $ steps   : int [1:17568] NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : chr [1:17568] "2012-10-01" "2012-10-01" "2012-10-01" "2012-10-01" ...
##  $ interval: int [1:17568] 0 5 10 15 20 25 30 35 40 45 ...
```

```
df$date <- as.Date.character(df$date)
```

What is mean total number of steps taken per day?

```
q1 <- df %>% drop_na() %>% group_by(date) %>% summarize(tot_steps = sum(steps))

ggplot(q1, aes(x=date, y=tot_steps)) +
  geom_bar(stat = "identity", fill = "#69b3a2") + ggtitle("Total steps by day")
```



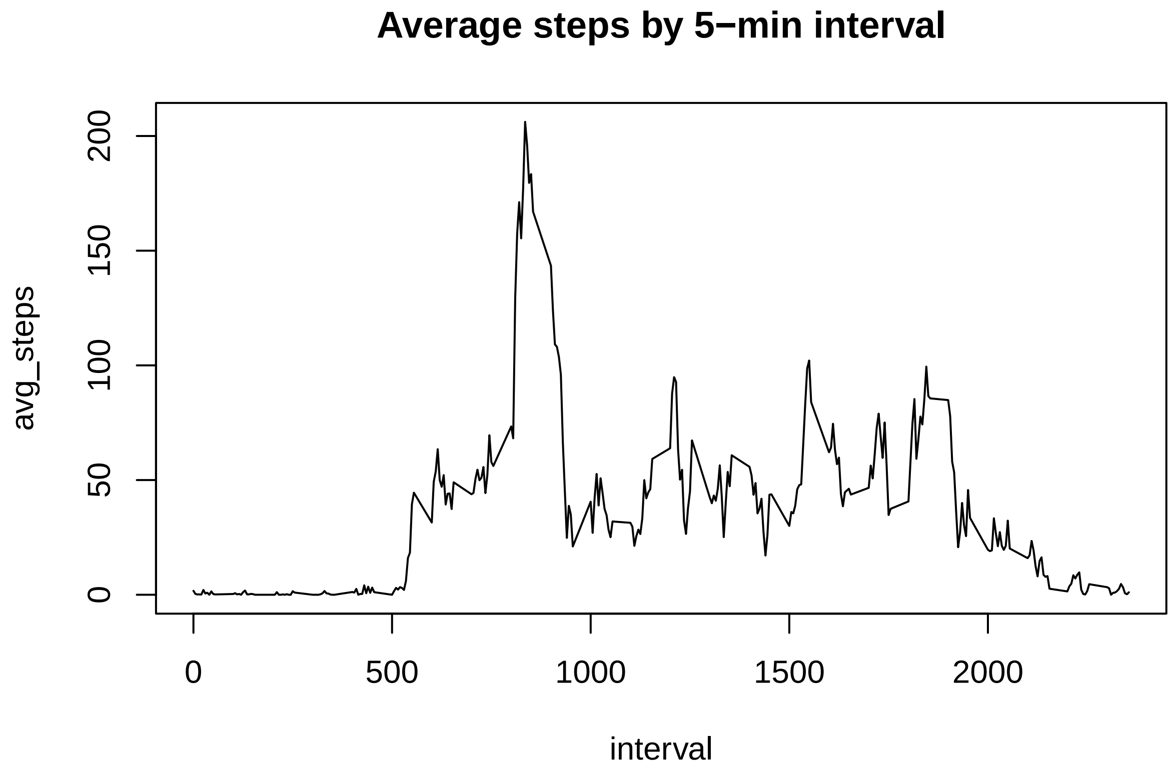
```
mean <- mean(q1$tot_steps)
median <- median(q1$tot_steps)
```

The mean of the total steps taken per day is 10766, while the median is 10765.

What is the average daily activity pattern?

```
q2 <- df %>% drop_na() %>% group_by(interval) %>% summarize(avg_steps = mean(steps))

plot(q2, type = "l", main = "Average steps by 5-min interval")
```



```
max <- as.numeric(q2[q2$avg_steps == max(q2$avg_steps),"interval"])
```

The 5-minute interval, on average across all the days in the dataset, that contains the maximum number of steps is 835.

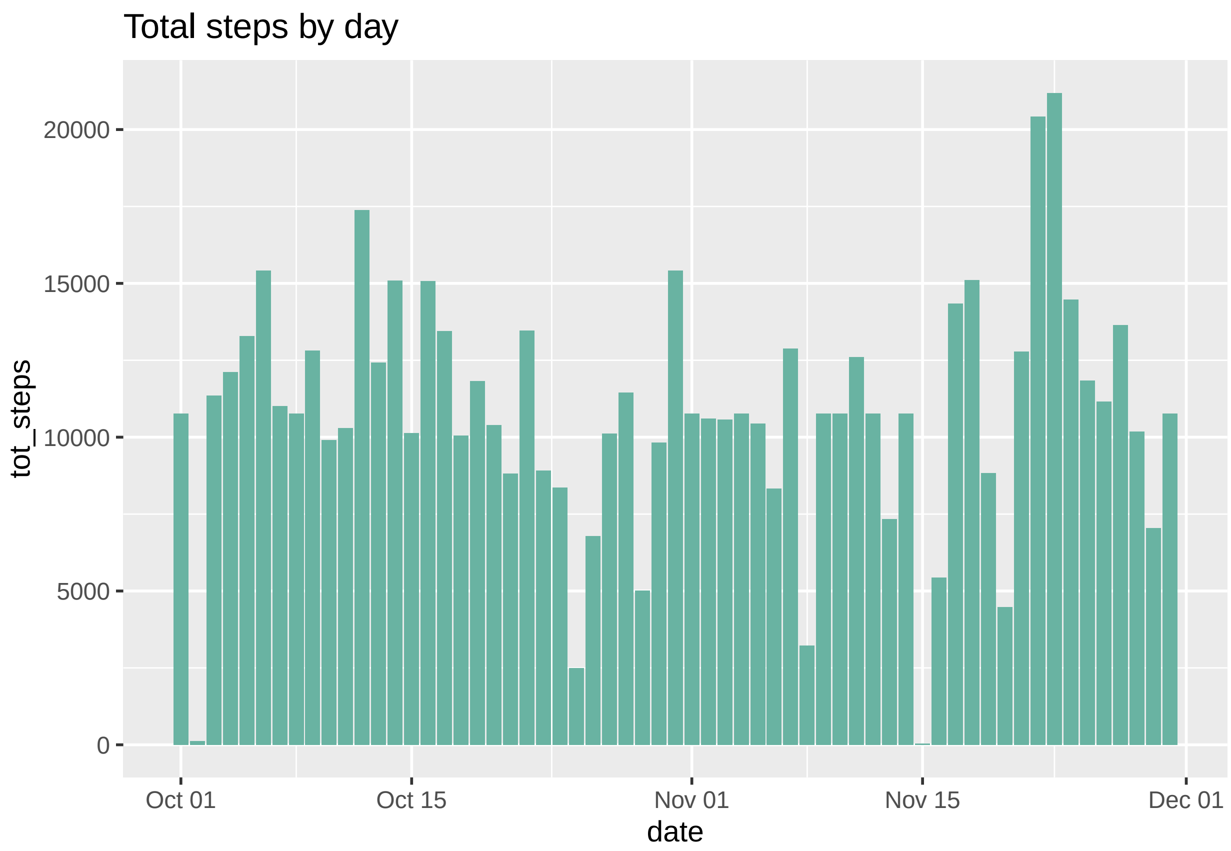
Imputing missing values

```
# Count NAs
NAs <- sum(is.na(df))

# Impute NAs with AVG(interval)
q3 <- df
q3$steps <- ave(q3$steps, q3$interval, FUN=function(x)
  ifelse(is.na(x), mean(x, na.rm = TRUE), x))

# Plot
q3_grouped <- q3 %>% group_by(date) %>% summarize(tot_steps = sum(steps))

q3_grouped %>%
  ggplot(aes(x=date, y=tot_steps)) +
  geom_bar(stat = "identity", fill = "#69b3a2") + ggtitle("Total steps by day")
```



```
mean2 <- mean(q3_grouped$tot_steps)
median2 <- median(q3_grouped$tot_steps)

if (mean == mean2 & median == median2) {q3_comp <- "both are identical to the previously calculated valu
} else{q3_comp <- "there are some differences compared to the previous values"}
```

The total number of missing values in the dataset is 2304. The new mean is 10766 and the new median is 10766, there are some differences compared to the previous values. When comparing the imputed dataset with the previous one, the main difference is that there are 8 new observations/rows.

Are there differences in activity patterns between weekdays and weekends?

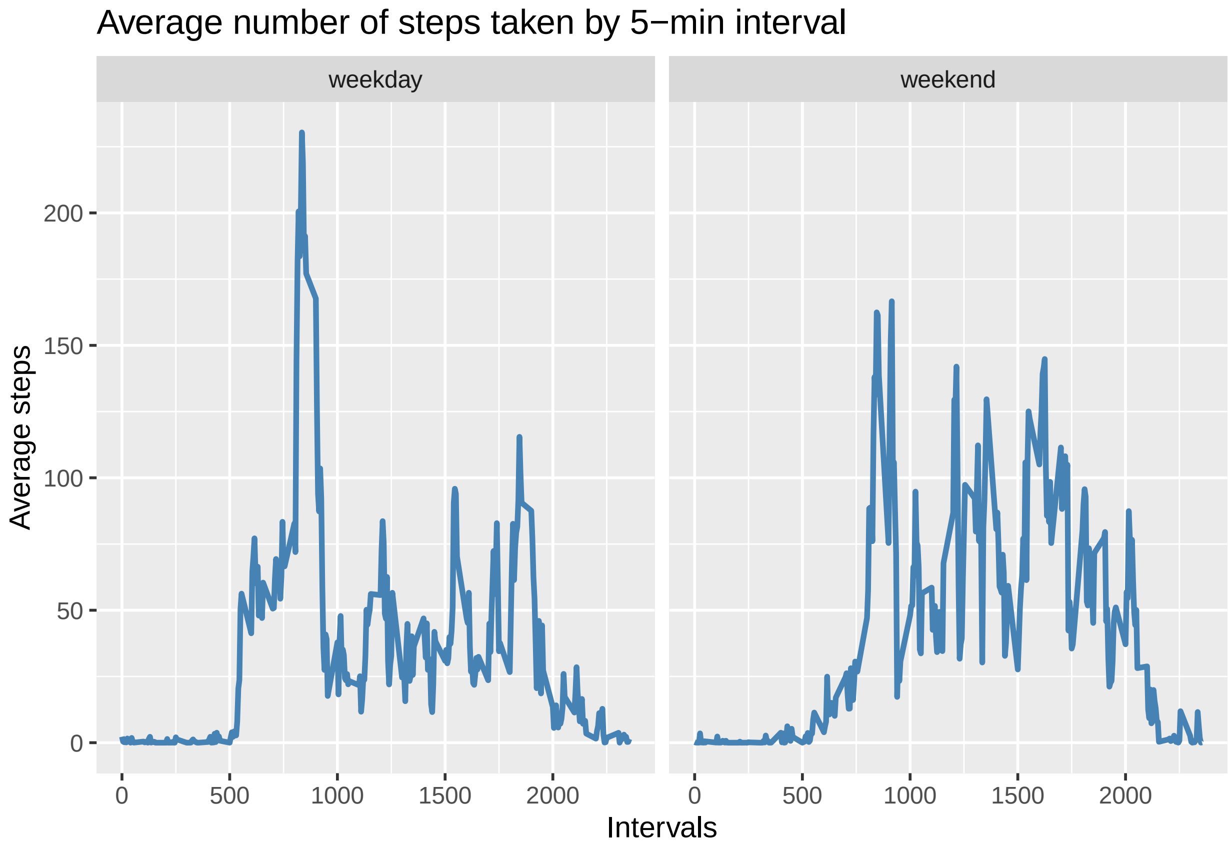
```
q4 <- q3

q4$weekday <- weekdays(q4$date)
q4$type <- factor(ifelse(q4$weekday %in% c("Saturday","Sunday"), "weekend","weekday"))


q4_avg <- q4 %>% group_by(type,interval) %>% summarize(avg_steps = mean(steps))
```

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

```r
ggplot(data = q4_avg, aes(interval, avg_steps)) +
  geom_line(color = "steelblue", size = 1) +
  labs(title = "Average number of steps taken by 5-min interval",
       y = "Average steps", x = "Intervals") +
  facet_wrap(~ type)
```


Average number of steps taken by 5−min interval
weekday
weekend
Average steps
Intervals