Reproducible Research: Peer Assessment 1

Loading and preprocessing the data

At first, we load the packages needed to clean-up the data, namely `lubridate` and `dplyr`.

```
library(dplyr)
library(lubridate)

path_to_data <- file.path('data', 'activity.csv')
measured_data <- tbl_df(read.csv(path_to_data))
analytic_data <- measured_data %>%
    mutate(date = ymd(as.character(date))) %>%
    mutate(interval = new_period(hour = interval %/% 100, minute = interval %% 100))
```

The following is a preview of the pre-processed data.

```
head(analytic_data)
```

```
## Source: local data frame [6 x 3]
##
##   steps       date interval
##   (int)     (time)    (dbl)
## 1    NA 2012-10-01       0S
## 2    NA 2012-10-01    5M 0S
## 3    NA 2012-10-01   10M 0S
## 4    NA 2012-10-01   15M 0S
## 5    NA 2012-10-01   20M 0S
## 6    NA 2012-10-01   25M 0S
```

What is mean total number of steps taken per day?

Firstly, we are going to calculate daily total steps, and store it in `daily_total_steps` for further use.

```
daily_total_steps <- analytic_data %>%
    group_by(date) %>%
    summarise(total_steps = sum(steps, na.rm = TRUE))

head(daily_total_steps)
```

```
## Source: local data frame [6 x 2]
##
##         date total_steps
##       (time)       (int)
## 1 2012-10-01           0
## 2 2012-10-02         126
## 3 2012-10-03       11352
## 4 2012-10-04       12116
## 5 2012-10-05       13294
## 6 2012-10-06       15420
```

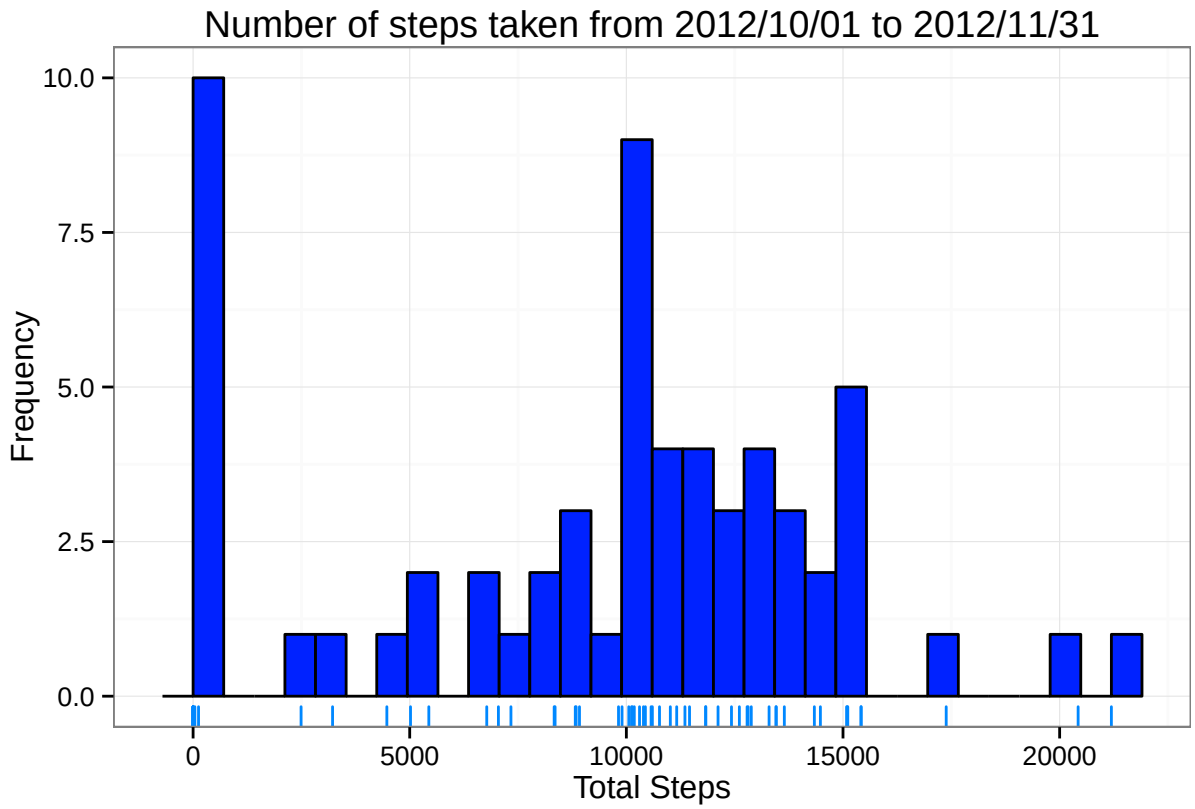
Next, we are going to take a look on the distribution of the data - some of the data are missing, so it would be of interest to take a look on the skewed distribution and how we are going to fix it.

```
library(ggplot2)

colPalette <- rainbow(45)

g <- ggplot(daily_total_steps, aes(total_steps))
g <- g + geom_histogram(col = 'black', fill = colPalette[30]) # colors with blue
g <- g + labs(x = 'Total Steps', y = 'Frequency')
g <- g + labs(title = 'Number of steps taken from 2012/10/01 to 2012/11/31')
g <- g + geom_rug(col = colPalette[27])
g <- g + theme_bw(base_size = 12, base_family = 'sans')

print(g)
```



As we can see, the missing values highly skew the total number of steps taken daily - 10 days with near-to-zero steps are recorded. We are going to discuss what we are going to do with the missing value later - for now, let's see the effect of such outliers with summary statistics of the data; more specifically, median and mean.

```
daily_total_steps_summary <- summary(daily_total_steps$total_steps)
median_total_steps <- daily_total_steps_summary['Median']
mean_total_steps <- daily_total_steps_summary['Mean']

print(daily_total_steps_summary[3:4])
```

```
## Median   Mean
##  10400   9354
```

We recorded median of 10400 and mean of 9354.

What is the average daily activity pattern?

Inputing missing values

Are there differences in activity patterns between weekdays and weekends?