Reproducible Research: Peer Assessment 1

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Loading and preprocessing the data

```
df <- read.csv("activity.csv")
df <- df %>%
        mutate(date = ymd(date)) %>%
        mutate(interval = str_pad(interval, 4, pad="0")) %>%
        mutate(interval = sub("(^\\d\\d)", "\\1:", interval)) %>%
        mutate(interval = parse_date_time(interval, "%H:%M"))
head(df)
```

```
##   steps       date            interval
## 1    NA 2012-10-01 0000-01-01 00:00:00
## 2    NA 2012-10-01 0000-01-01 00:05:00
## 3    NA 2012-10-01 0000-01-01 00:10:00
## 4    NA 2012-10-01 0000-01-01 00:15:00
## 5    NA 2012-10-01 0000-01-01 00:20:00
## 6    NA 2012-10-01 0000-01-01 00:25:00
```

What is mean total number of steps taken per day?

The total number of steps taken per day is as shown in the following chart, for the first five days, at least:

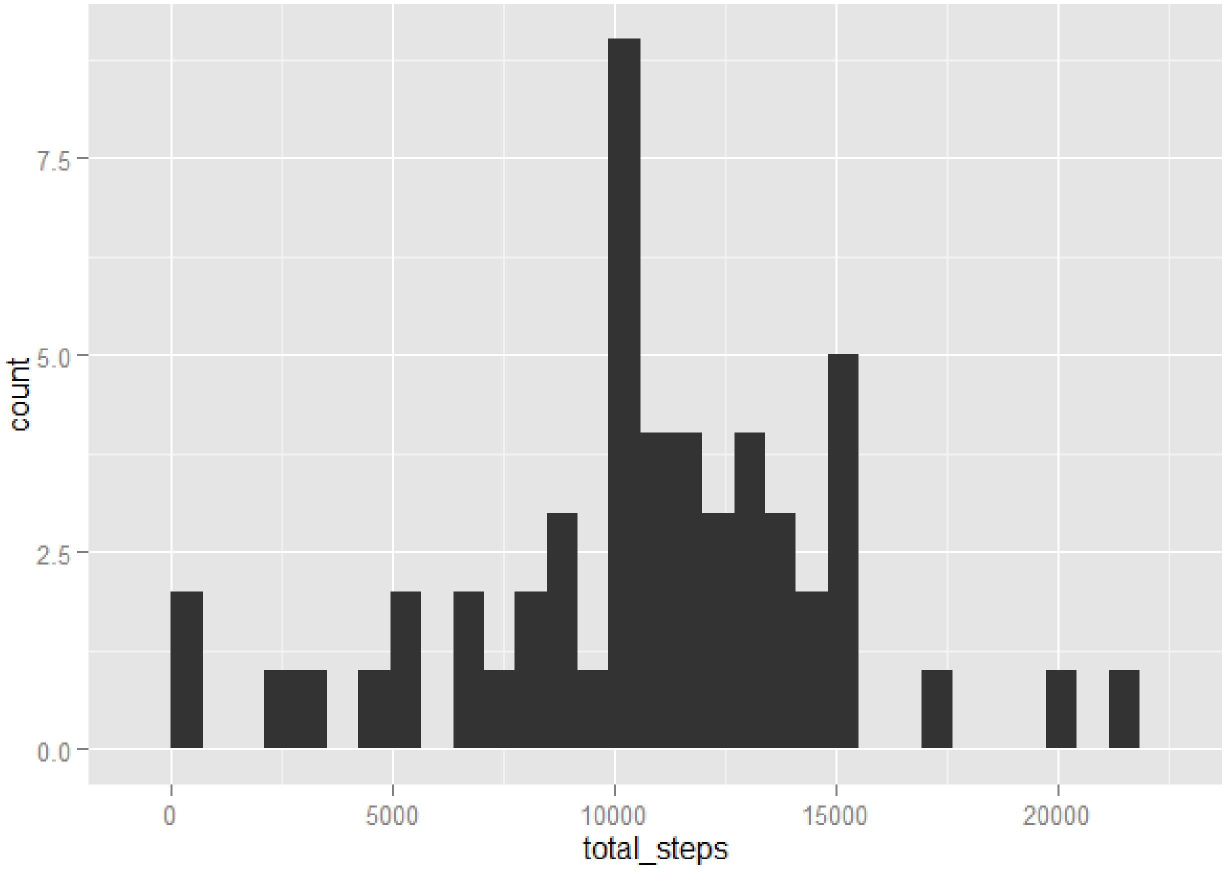
```
dfp <- df %>%
        filter(!is.na(steps)) %>%
        group_by(date) %>%
        summarize(total_steps = sum(steps))
print(head(dfp))
```

```
## Source: local data frame [6 x 2]
##
##         date total_steps
## 1 2012-10-02         126
## 2 2012-10-03       11352
## 3 2012-10-04       12116
## 4 2012-10-05       13294
## 5 2012-10-06       15420
## 6 2012-10-07       11015
```

A histogram showing this data is as follows:

```
qplot(data = dfp, x = total_steps)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



As can be seen in the summary table below, the average number of steps taken per day is 10766 and the median is 10765.

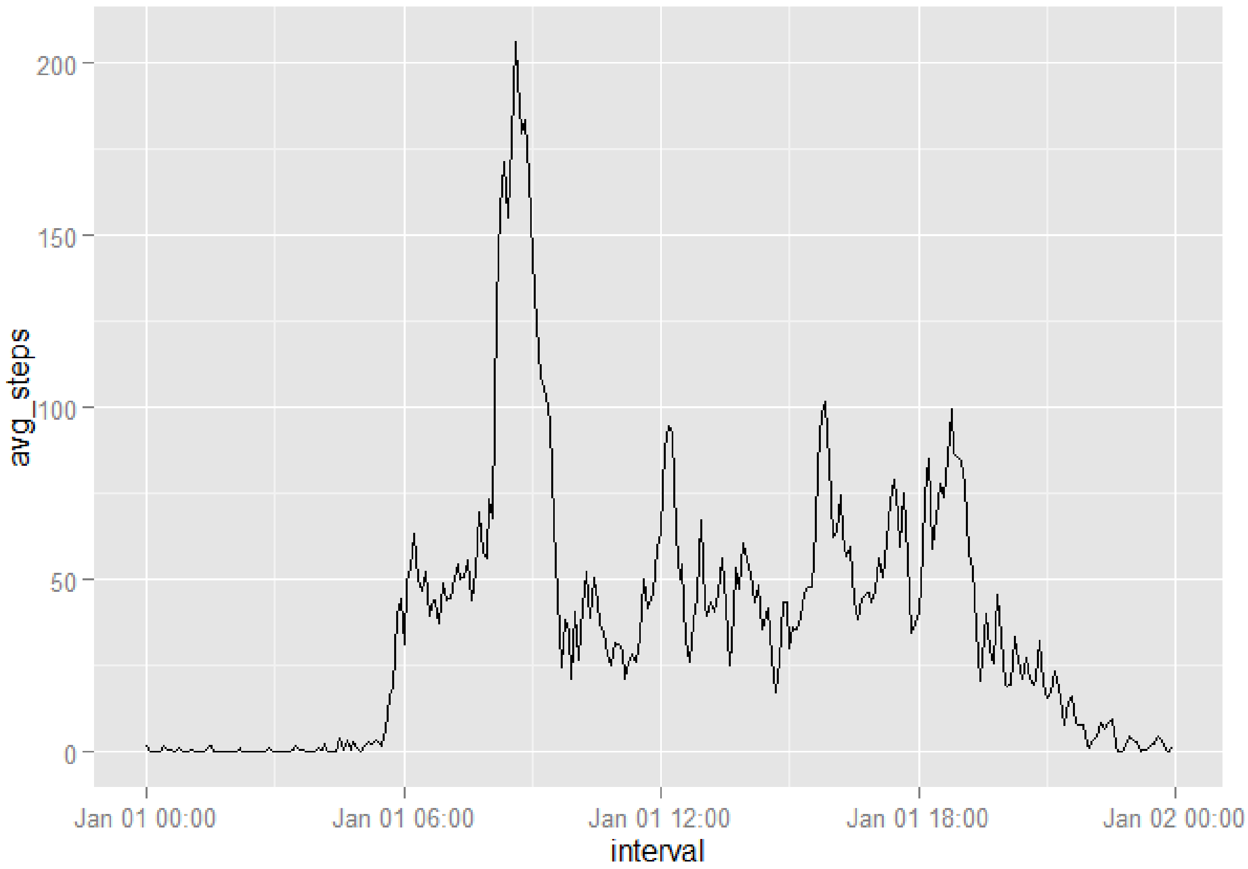
```
summary(dfp)
```

```
##       date                         total_steps
##  Min.   :2012-10-02 00:00:00   Min.   :   41
##  1st Qu.:2012-10-16 00:00:00   1st Qu.: 8841
##  Median :2012-10-29 00:00:00   Median :10765
##  Mean   :2012-10-30 17:12:27   Mean   :10766
##  3rd Qu.:2012-11-16 00:00:00   3rd Qu.:13294
##  Max.   :2012-11-29 00:00:00   Max.   :21194
```

What is the average daily activity pattern?

The time series plot of the average number of steps taken by time interval across the number of days is as follows:

```
dfs <- df %>%
        filter(!is.na(steps)) %>%
        group_by(interval) %>%
        summarize(avg_steps=mean(steps))
qplot(data=dfs, x=interval, y=avg_steps, geom="line")
```



The time interval containing the maximum number of steps is 0000-01-01 08:35:00.

Imputing missing values

The number of missing values in the dataset is 2304.

To impute the missing values, I will apply the mean for the time of day across all days.

```
dfn <- left_join(df, dfs, by=c("interval"="interval")) %>%
        do(mutate(., steps = ifelse(is.na(steps), avg_steps, steps))) %>%
        select(-avg_steps)
```

See here for usage of “do” in mutate (https://github.com/hadley/dplyr/issues/464)

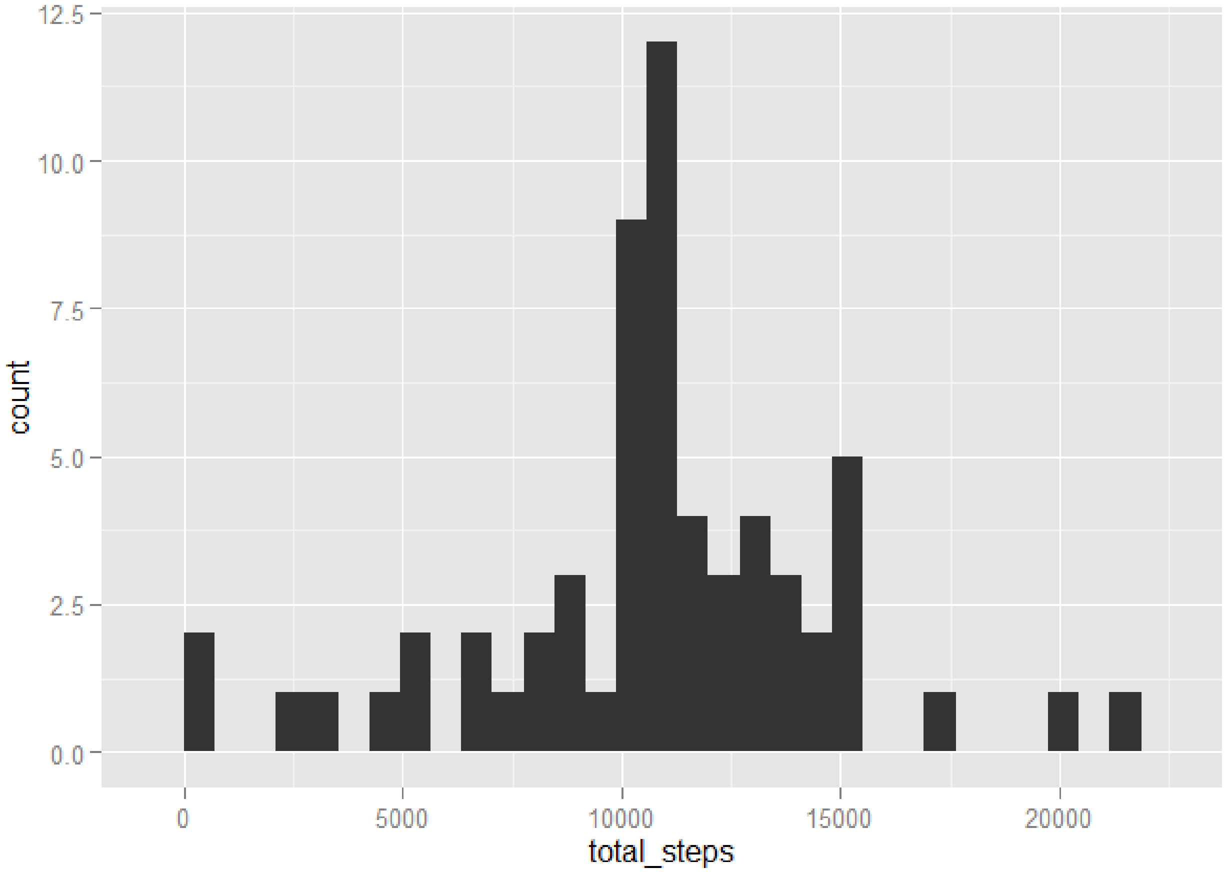
```
dfp2 <- dfn %>%
        group_by(date) %>%
        summarize(total_steps = sum(steps))
print(head(dfp2))
```

```
## Source: local data frame [6 x 2]
## 
##         date total_steps
## 1 2012-10-01    10766.19
## 2 2012-10-02      126.00
## 3 2012-10-03    11352.00
## 4 2012-10-04    12116.00
## 5 2012-10-05    13294.00
## 6 2012-10-06    15420.00
```

A histogram showing this data is as follows:

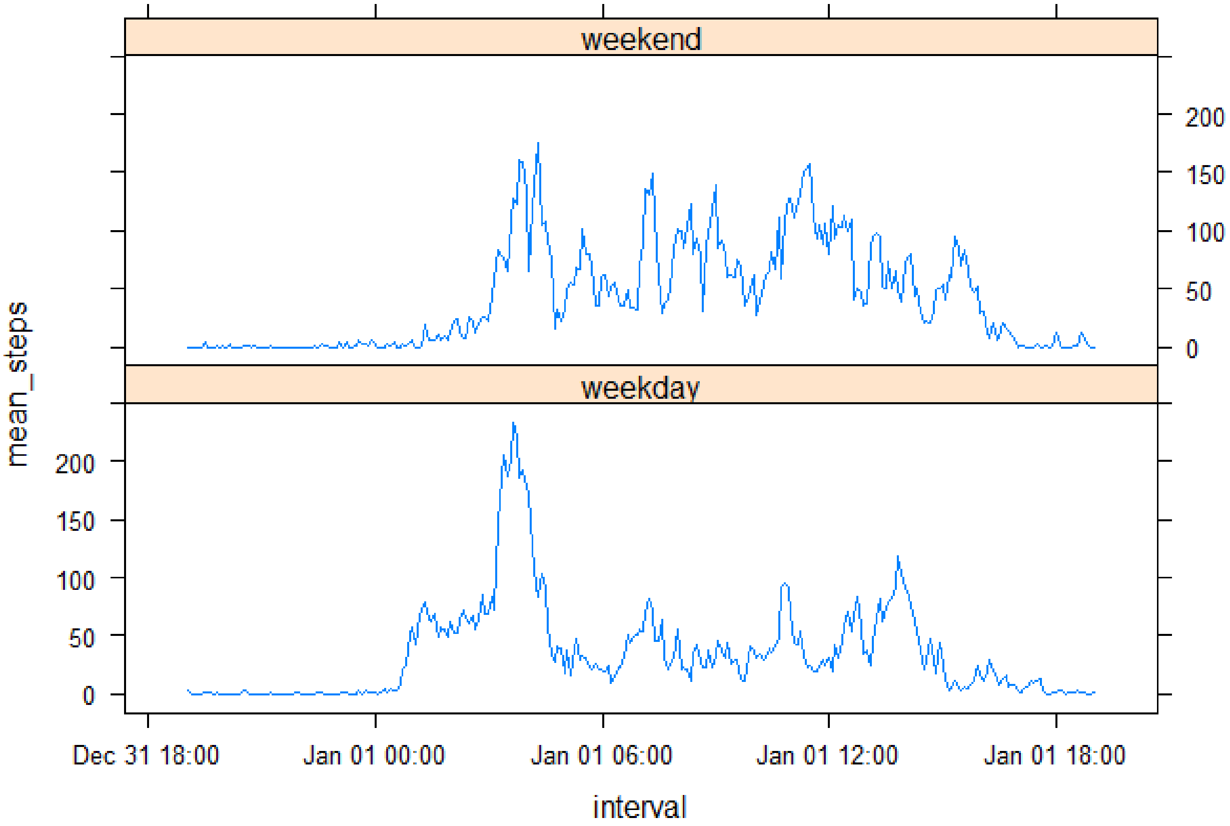
```
qplot(data = dfp2, x = total_steps)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



Are there differences in activity patterns between weekdays and weekends?

```
dfl <- df %>%
        filter(!is.na(steps)) %>%
        mutate(weekday = ifelse(wday(date, label = TRUE) %in% c("Sat", "Sun"),
"weekend", "weekday")) %>%
        group_by(interval, weekday) %>%
        summarize(mean_steps = mean(steps))
xyplot(mean_steps ~ interval | weekday, data=dfl, layout=c(1,2), type="l")
```


weekend
weekday
mean_steps
Dec 31 18:00
Jan 01 00:00
Jan 01 06:00
Jan 01 12:00
Jan 01 18:00
interval