Reproducible Research: Peer Assessment 1

Overview

It is now possible to collect a large amount of data about personal movement using activity monitoring devices. Data was analyzed from one such device which collected data at 5 minute intervals through out the day during the months of October and November, 2012, and include the number of steps taken in 5 minute intervals each day.

Loading and preprocessing the data

The working directory was set, the required packages were uploaded and data downloaded.

```
setwd("C:/Users/Henry/Documents/Rworkingdirectory/Reproducable_Research/Project_1")
##require(utils)
library(dplyr)
library(ggplot2)
library(tidyr)
library(lubridate)
library(xtable)
library(knitr)
library(gridExtra)
url <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip"
fileLocation <- "C:/Users/Henry/Documents/Rworkingdirectory/Reproducable_Research/Project_1/repdata_data
if (!file.exists(fileLocation)){
  download.file(url, fileLocation)
}
extractedFile <- unzip(fileLocation, overwrite=TRUE)
activity <- read.csv(extractedFile)
activity$date <- as.POSIXct(activity$date)
```

What is mean total number of steps taken per day?

1) The total number of steps taken per day was calculated, and is presented below.

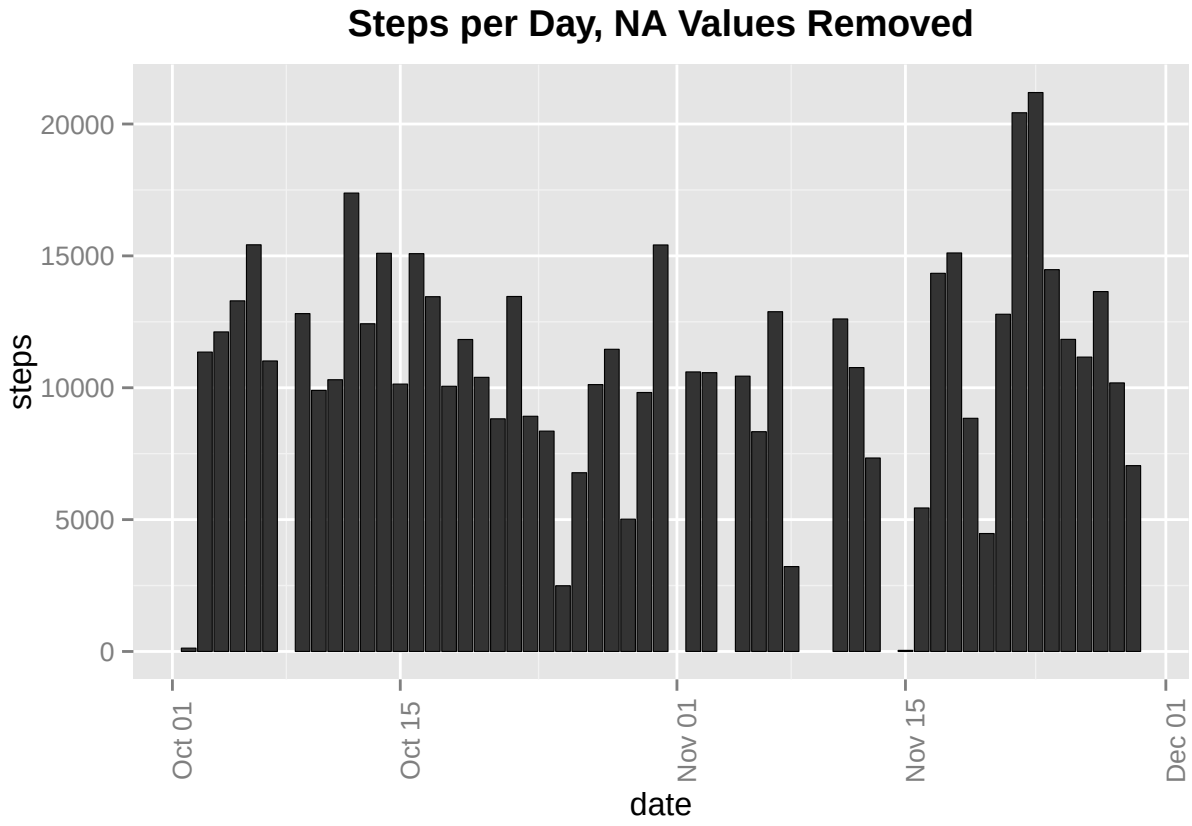
```
sumDailySteps <- select(activity, steps, date) %>%
            na.omit() %>%
            group_by(date) %>%
            summarise(steps = sum(steps))%>%
            as.data.frame()
kable(sumDailySteps)
```

date	steps
2012-10-02	126
2012-10-03	11352
2012-10-04	12116
2012-10-05	13294
2012-10-06	15420
2012-10-07	11015

date	steps
2012-10-09	12811
2012-10-10	9900
2012-10-11	10304
2012-10-12	17382
2012-10-13	12426
2012-10-14	15098
2012-10-15	10139
2012-10-16	15084
2012-10-17	13452
2012-10-18	10056
2012-10-19	11829
2012-10-20	10395
2012-10-21	8821
2012-10-22	13460
2012-10-23	8918
2012-10-24	8355
2012-10-25	2492
2012-10-26	6778
2012-10-27	10119
2012-10-28	11458
2012-10-29	5018
2012-10-30	9819
2012-10-31	15414
2012-11-02	10600
2012-11-03	10571
2012-11-05	10439
2012-11-06	8334
2012-11-07	12883
2012-11-08	3219
2012-11-11	12608
2012-11-12	10765
2012-11-13	7336
2012-11-15	41
2012-11-16	5441
2012-11-17	14339
2012-11-18	15110
2012-11-19	8841
2012-11-20	4472
2012-11-21	12787
2012-11-22	20427
2012-11-23	21194
2012-11-24	14478
2012-11-25	11834
2012-11-26	11162
2012-11-27	13646
2012-11-28	10183
2012-11-29	7047

2) A histogram of total number of steps per day was created.

```
DailyStepsPlot <- ggplot(sumDailySteps, aes(date, steps)) +
        geom_bar(stat="identity", position="dodge", color = "black", size=0.2)+
        ggtitle("Steps per Day, NA Values Removed")+
        theme(plot.title = element_text(face = "bold", vjust = 1.5),
              axis.text.x = element_text(angle = 90))
DailyStepsPlot
```



3) The mean and median number of steps per day was then calculated.

```
MMDailySteps <- select(activity, steps, date) %>%
          na.omit() %>%
          group_by(date) %>%
          summarise(steps = sum(steps))
mean <- as.numeric(summarise(MMDailySteps, steps = mean(steps)))
median <- as.numeric(summarise(MMDailySteps, steps = median(steps)))
MM <- data.frame(mean, median, stringsAsFactors=FALSE)
kable(MM)
```

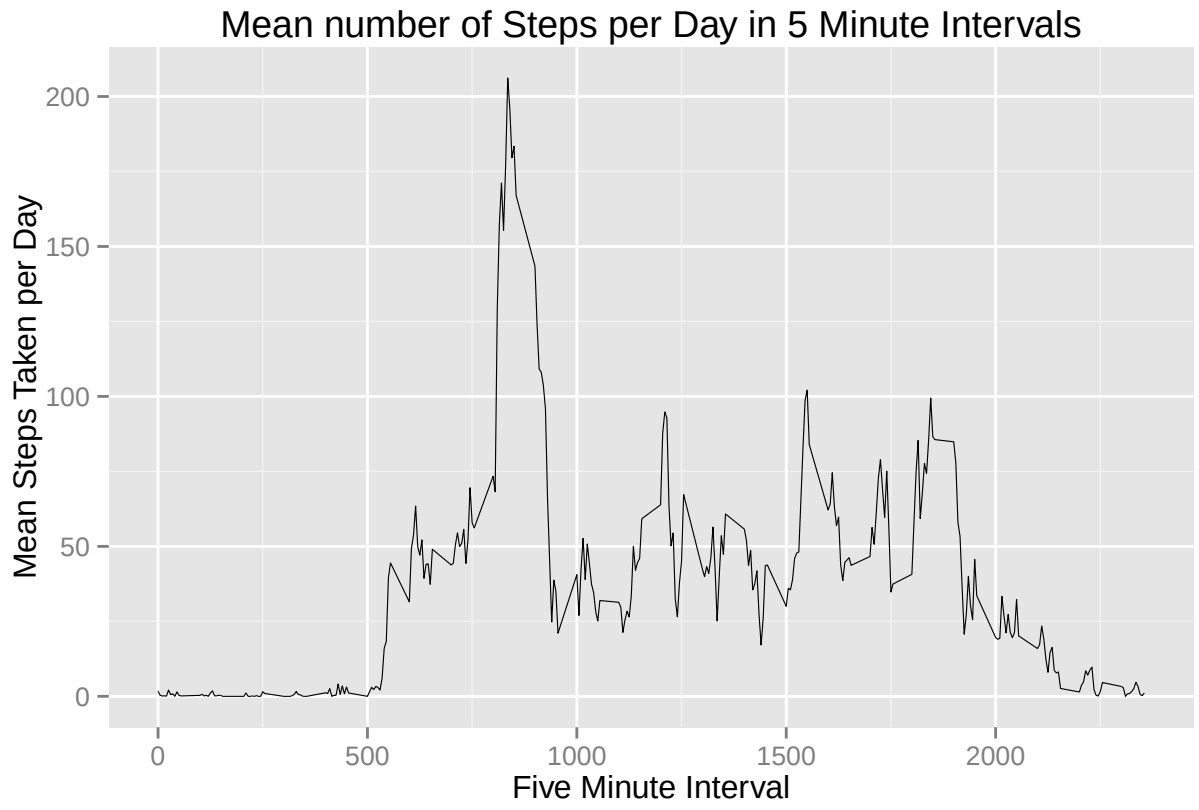
mean	median
10766.19	10765

What is the average daily activity pattern?

1) A time series plot was made of the 5-minute interval and the average number of steps taken per day, averaged accross all days.

```
FiveMinInterval <- select(activity, steps, interval) %>%
          na.omit() %>%
          group_by(interval) %>%
          summarise(steps = mean(steps))%>%
          as.data.frame()
```

```
TimeSeriesPlot1 <- FiveMinInterval %>% ggplot +
          geom_line(aes(interval, steps), stat="identity", position="dodge", colour="black", size=0.2) +
          ggtitle("Mean number of Steps per Day in 5 Minute Intervals") +
          xlab("Five Minute Interval") +
          ylab("Mean Steps Taken per Day")
TimeSeriesPlot1
```



2) The five minute interval with the most number of steps averaged acrross all days evaluate, and is presented with the corresponding maximum average number of steps.

```
FiveMinIntervalMax <- filter(FiveMinInterval, steps == max(steps))
kable(FiveMinIntervalMax)
```

interval	steps
835	206.1698

Imputing missing values

1) The total number of missing values was calculated to be 2304.

```
missingValues <- activity %>%
          complete.cases %>%
        table
unname(missingValues[1])
```

```
## [1] 2304
```

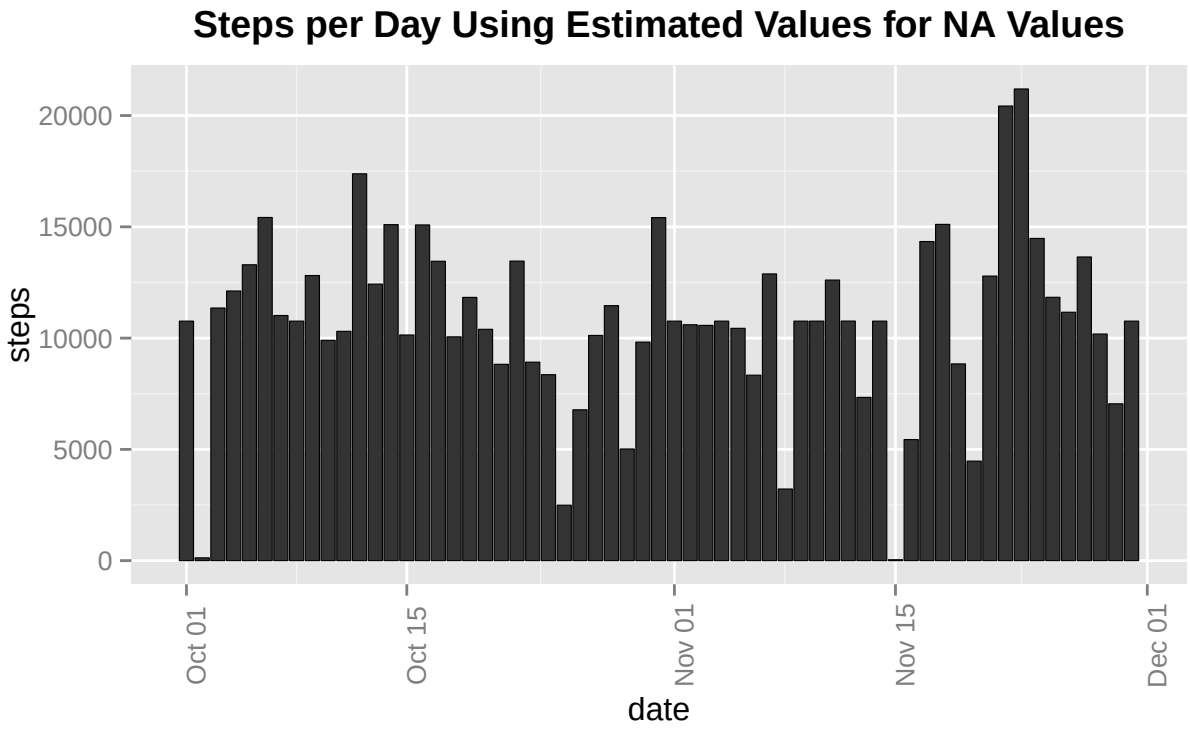
2) (Includes part 3)The missing data was filled in using the mean of the corresponding interval.

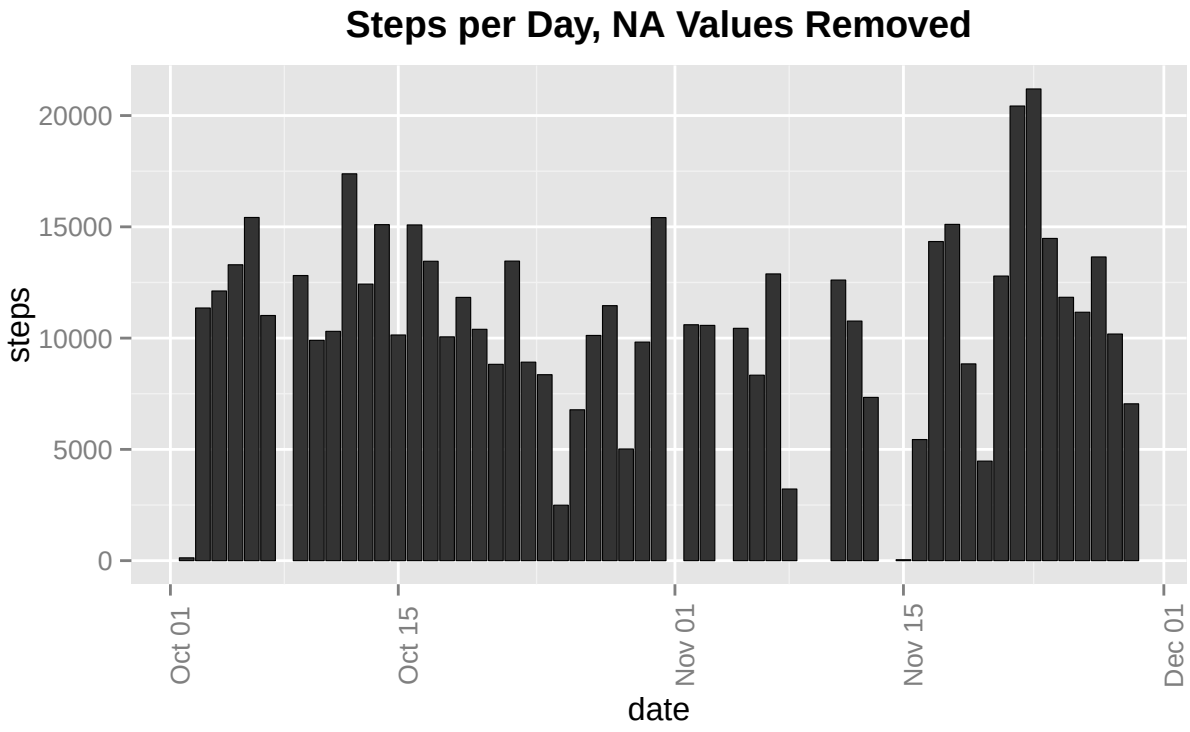
```
activityAdjustedNA <- activity %>%
          left_join(FiveMinInterval %>% select(interval, steps.average = steps), by = "interval") %>%
          mutate(steps.average = ifelse(is.na(steps), round(steps.average), steps)) %>%
          select(interval, date, steps.average) %>%
          rename(steps = steps.average)
```

4.1) A histogram of the total number of steps taken each day was created, and is presented below alongside the histogram showing steps per day when NA values were removed.

```
DailyStepsAdjusted <- select(activityAdjustedNA, steps, date)
DailyStepsAdjusted <- DailyStepsAdjusted %>%
          na.omit() %>%
          group_by(date) %>%
          summarise(steps = sum(steps))

AdjustedDailyStepsPlot <- ggplot(DailyStepsAdjusted, aes(date, steps)) +
        geom_bar(stat="identity", position="dodge", color = "black", size=0.2)+
        ggtitle("Steps per Day Using Estimated Values for NA Values")+
        theme(plot.title = element_text(face = "bold", vjust = 1.5),
              axis.text.x = element_text(angle = 90))
plots2 <- grid.arrange(AdjustedDailyStepsPlot, DailyStepsPlot, ncol=1)
```





4.2) The Mean and median total number of steps taken per day with NA values replaced reported alongside those figures when the NA values were removed from

```
adjustedMean <- select(activityAdjustedNA, steps, date) %>%
          group_by(date) %>%
```

```
          summarise(steps = sum(steps)) %>%
          summarise(mean = mean(steps))
adjustedMedian <- select(activityAdjustedNA, steps, date) %>%
          group_by(date) %>%
          summarise(steps = sum(steps)) %>%
          summarise(median = median(steps))
AveMedianComp <- data.frame(mean, median, adjustedMean, adjustedMedian, stringsAsFactors=FALSE)
names(AveMedianComp) <- c("Mean NA Removed", "Median NA Removed", "Mean NA Estimated", "Median NA Estima
kable(AveMedianComp)
```

Mean NA Removed	Median NA Removed	Mean NA Estimated	Median NA Estimated
10766.19	10765	10765.64	10762

There is very little diffrence between the values of the Mean and Median when the NA values are removed compared to when the NA values are estimated using the mean value for the interval.

Are there differences in activity patterns between weekdays and weekends?

The dataset using the estimated vales in place of missing values will be used in this analysis. The means and medians of the datasets with removed and estimated values are very close in value, and the distribution of the data set with estimated values is less irratically variable than the dataset with NA values removed. This provides some evidence the dataset with estimated values provides a more accurate reflection of the subject's true movements.

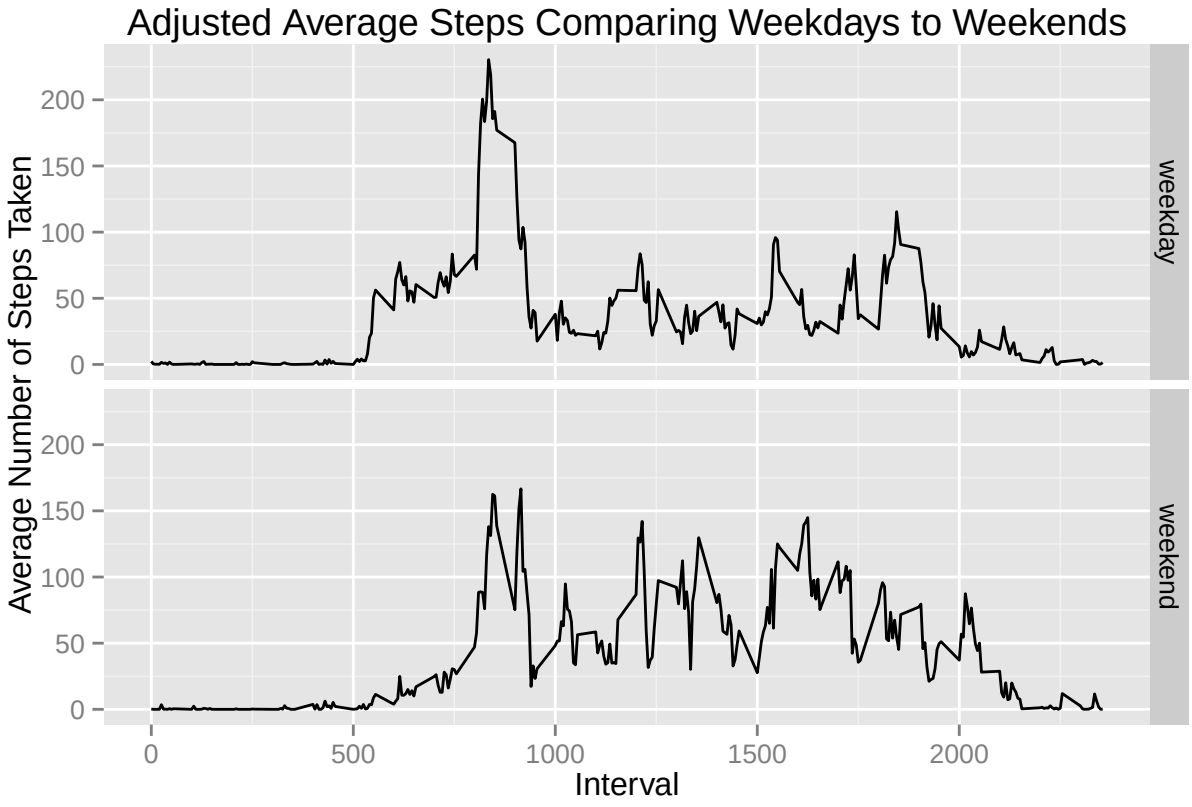
1) A new factor variable in the dataset was created with two levels, weekday and weekend, indicating whether a given date is a weekday or weekend day.

```
dataByDay <- activityAdjustedNA %>%
            mutate(day = factor(wday(date) %in%
            c(1, 7), labels = c("weekday", "weekend")))
```

2) A panel plot was made containing a time series plot of the 5-minute interval and the average number of steps taken, averaged across all weekday days or weekend days.

```
meandataByDay <- dataByDay %>%
            select(steps, interval, day) %>%
            group_by(interval, day) %>%
            summarise(steps = mean(steps))

meandataByDay %>%
            ggplot +
            geom_line(aes(interval, steps)) +
            facet_grid(day ~ .) +
            ggtitle("Adjusted Average Steps Comparing Weekdays to Weekends")+
            xlab("Interval") +
            ylab("Average Number of Steps Taken")
```


Adjusted Average Steps Comparing Weekdays to Weekends
Average Number of Steps Taken
200
150
100
50
0
weekday
weekend
0
500
1000
1500
2000
Interval