Reproducible Research: Peer Assessment 1

Sylvia Seow

September 2015

Pre load all the required package

```
## load all the required library
library(plyr)
library(ggplot2)
```

Loading and preprocessing the data

The data used for processing will be used from the working directory, if the csv file is not found in the working directory, it will be unzip from the zip file included in the working directory

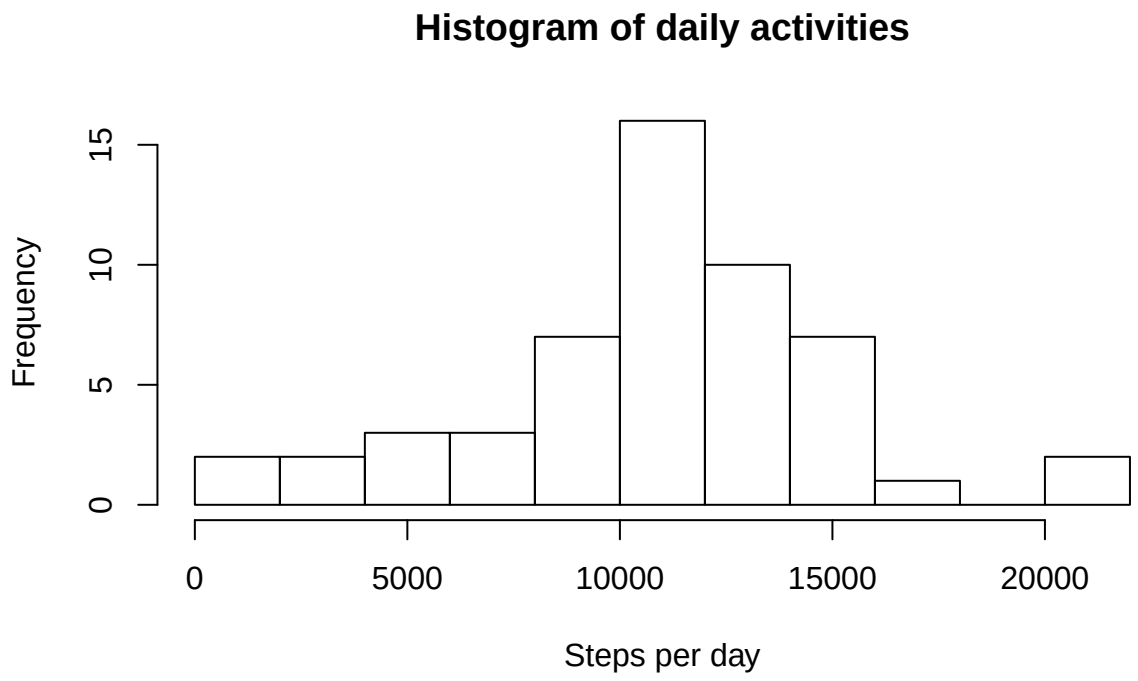
```
if (!file.exists("activity.csv"))
  unzip("activity.zip")

data <-read.csv("activity.csv", header= TRUE, sep=",",
                colClasses = c("integer","character","integer"))
## format the date into yyyy-mm-dd
data$date <-as.Date(data$date,"%Y-%m-%d")
```

the variable of the dataset included are . steps - number of steps taken in 5 min interval (NA defined as missing value) . date - the date of the data record in YYYY-MM-DD format . interval - the identifier of the 5-minute interval when the measurement is taken

What is mean total number of steps taken per day?

The histogram below show the frequency of the number of steps taken per day . All missing value in the dataset is ignored.

```
dlyActivity <- ddply(data, "date", summarise, steps=sum(steps))
hist(dlyActivity$steps, breaks=10, main ="Histogram of daily activities",
     xlab="Steps per day")
```



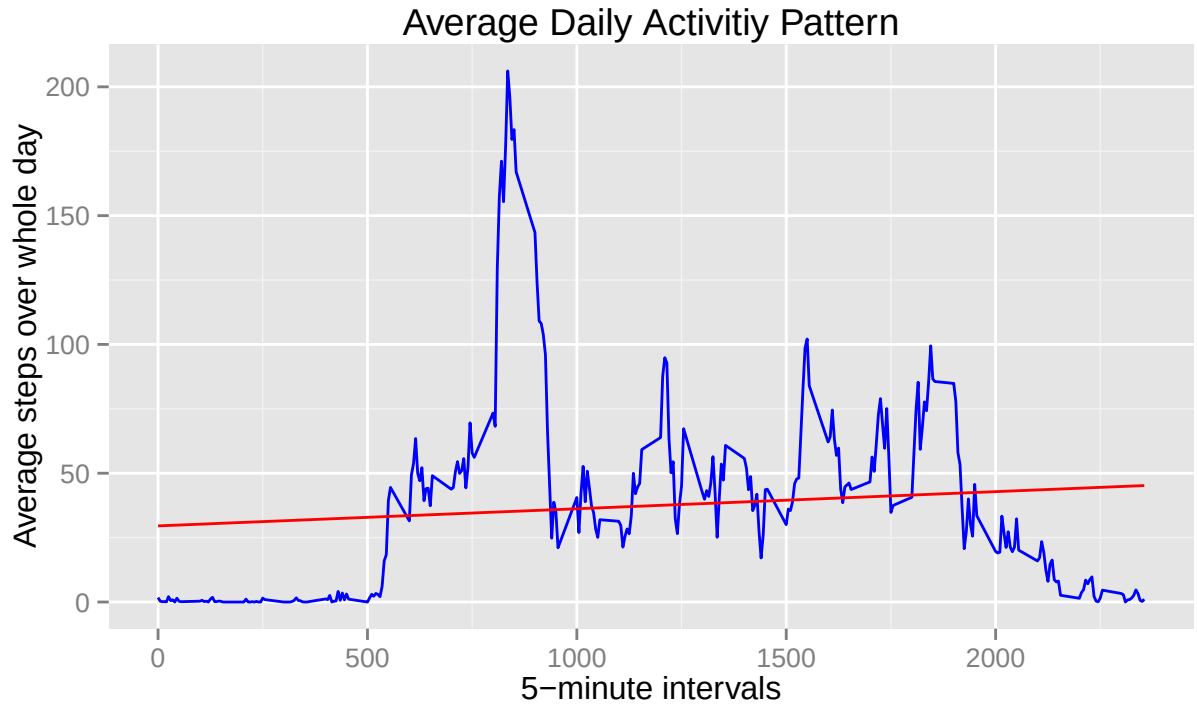

```
avg <- mean(dlyActivity$steps, na.rm = TRUE)
mdn <- median(dlyActivity$steps, na.rm = TRUE)
```

mean is 10766.1886792453 , median is 10765

What is the average daily activity pattern?

We will next plot the average of steps per interval to detect the average daily activity pattern for the given dataset. In this case, again the missing value are ignored too.

```
intervalActivity <- ddply(data,"interval", summarise, steps=avg  <- mean(steps,na.rm=TRUE))

g <- ggplot(intervalActivity, aes(interval,steps))
g <- g + geom_line(colour = "blue")
g <- g + geom_smooth(method="lm", se= FALSE, col="red", aes(group=1))
g <- g + labs(x="5-minute intervals") + labs(y= "Average steps over whole day")
g <- g + labs(title="Average Daily Activitiy Pattern")
print (g)
```




```
maxint <-intervalActivity[intervalActivity$steps==max(intervalActivity$steps),1]
```

The interval that contains the maximum number of steps in the dataset is **835**

Imputing missing values

Total number of missing case

```
nrow(data[!complete.cases(data),])
```

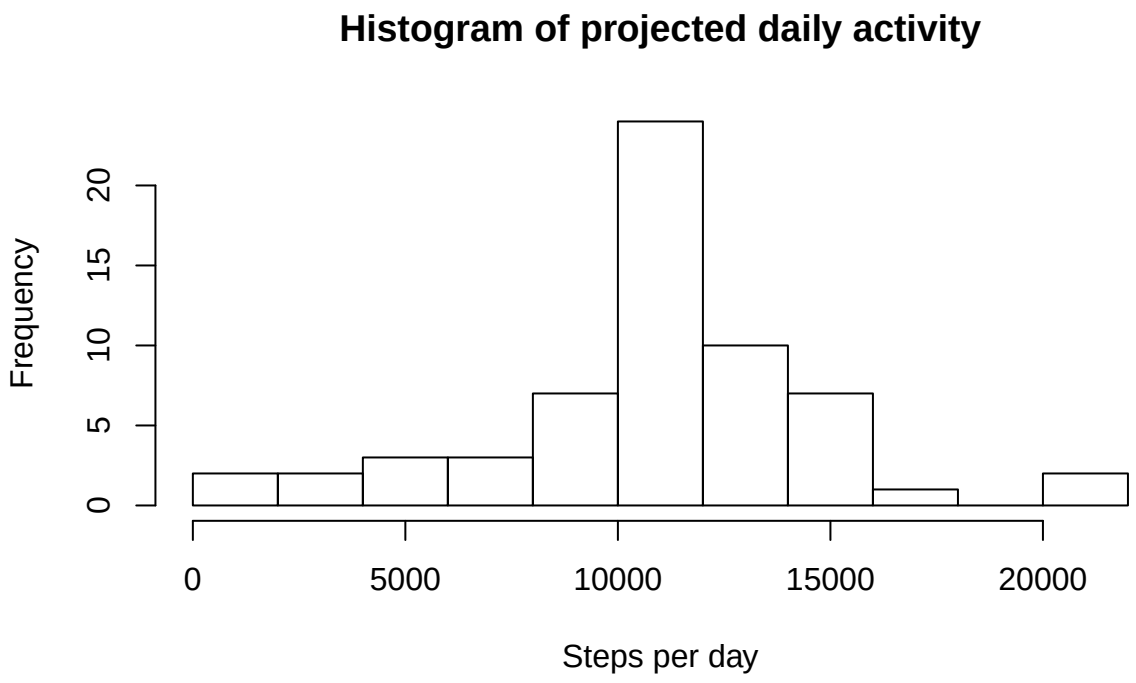
```
## [1] 2304
```

In here we will derived the impute missing data strategy as if the data is missing, we will impute the data with average value of the interval

```
impute.value <- function(steps,interval)
{
  imputev <-NA
  if (!is.na(steps))
      imputev <- c(steps)
  else
      imputev <- intervalActivity[intervalActivity$interval==interval,"steps"]
  return(imputev)
}

projected.data <-data
projected.data$steps <-mapply(impute.value,projected.data$steps, projected.data$interval)
projected.data.sum <- ddply(projected.data,"date",summarise, steps=sum(steps))
```

Using the "projected" dataset, let make a histogram of total number of steps taken each day and calculate the mean and median total number of steps.

```
hist(projected.data.sum$steps, breaks=10, main="Histogram of projected daily activity",
     xlab= "Steps per day")
```




```
projected.mean = mean(projected.data.sum$steps)
projected.median =median(projected.data.sum$steps)
```

After imputing missing value, we can summarised that mean for projected as 10766.19, and median projected as 10766.19

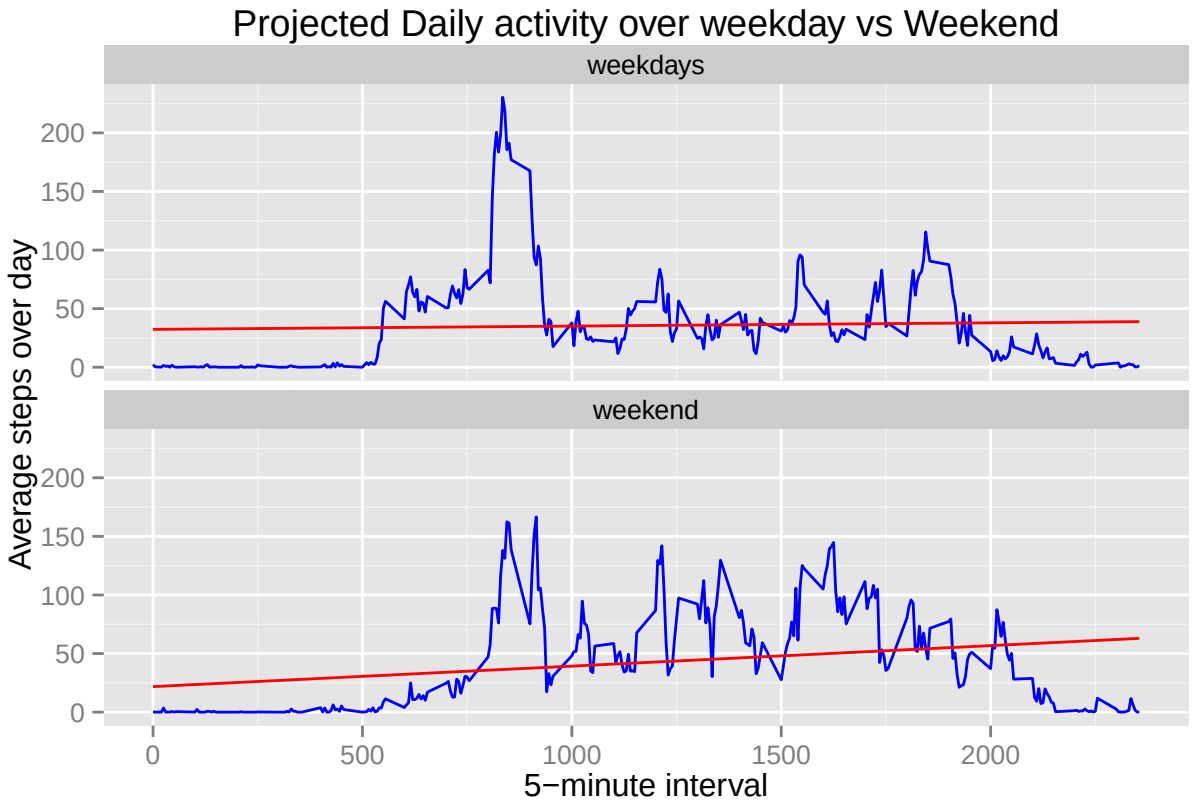
We can summarised that the pattern looks almost the same like the previous histogram, but the major difference is that the number of steps has been increased overall.

Are there differences in activity patterns between weekdays and weekends?

There are differences in activity patterns between weekday and weekend. They are clearly demonstrated as plot beneath.

```
weekend <- weekdays(projected.data$date) %in% c("Saturday","Sunday")
projected.data$DayType[!weekend] <-"weekdays"
projected.data$DayType[weekend] <- "weekend"

week.activity <-ddply(projected.data, c("interval", "DayType"),
                          summarise, steps = mn <-mean(steps, na.rm=TRUE))
```

```
g <- ggplot(week.activity, aes(interval, steps))
g <- g + geom_line(colour="blue")
g <- g + facet_wrap(~DayType, nrow=2)
g <- g + geom_smooth(method="lm", se= FALSE, col="red", aes(group=1))
g <- g + labs(x="5-minute interval") + labs(y="Average steps over day")
g <- g + labs(title= "Projected Daily activity over weekday vs Weekend")
print (g)
```



We can summarised that during the **weekday** , most steps are done during the morning session, and less steps recorded for rest of the day. However on **Weekends**, steps made by subject is more balance and evenly distributed during the days, more steps are taken in the evening, if compare to weekday result.