Reproducible Research - Course Project 1

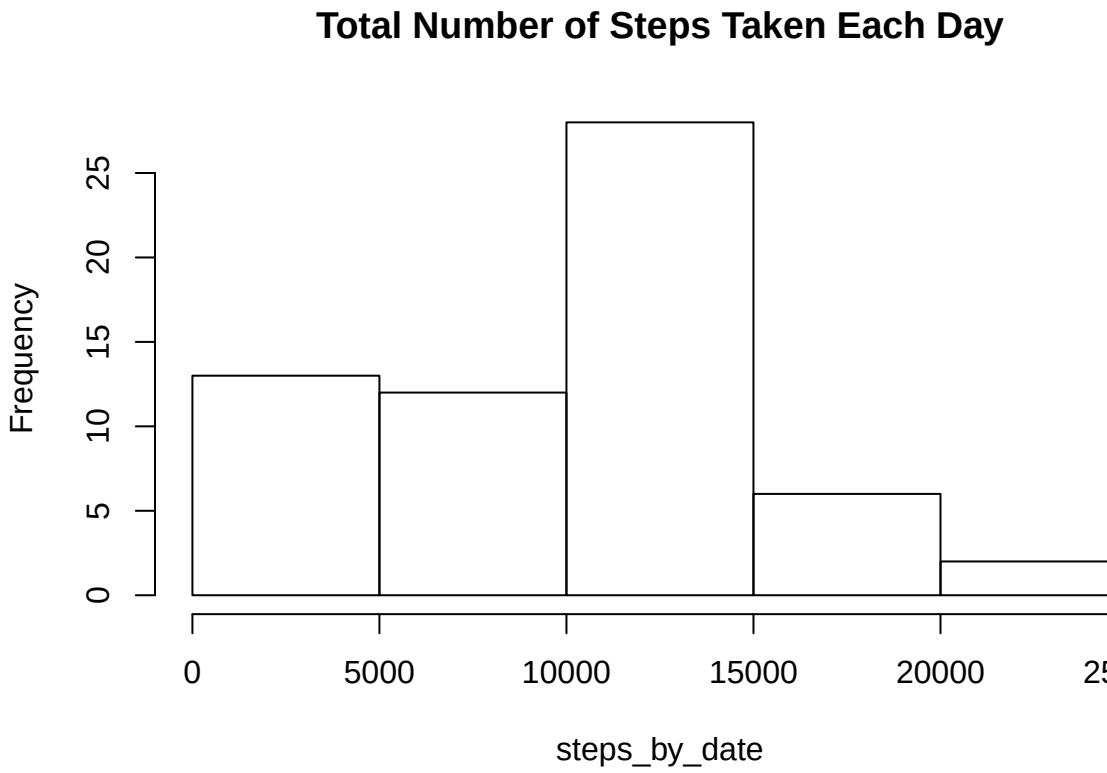
Step 1 : Loading the data

```
setwd("c:/users/vijay/documents/rep_research/data")
activity = read.csv("activity.csv")
```

Step 2 : What is the mean total number of steps taken per day?

Histogram of the total number of steps taken each day

```
#use tapply to sum the number of steps by date
steps_by_date = as.numeric(tapply(activity$steps, activity$date, sum, na.rm = T))
#plot a histogram
hist(steps_by_date, main = "Total Number of Steps Taken Each Day")
```



to create the histogram-1.pdf

Mean and Median of the Total Number of Steps Taken Each Day

```
#mean of the number of steps taken each day
mean(steps_by_date)
```

```
## [1] 9354.23
```

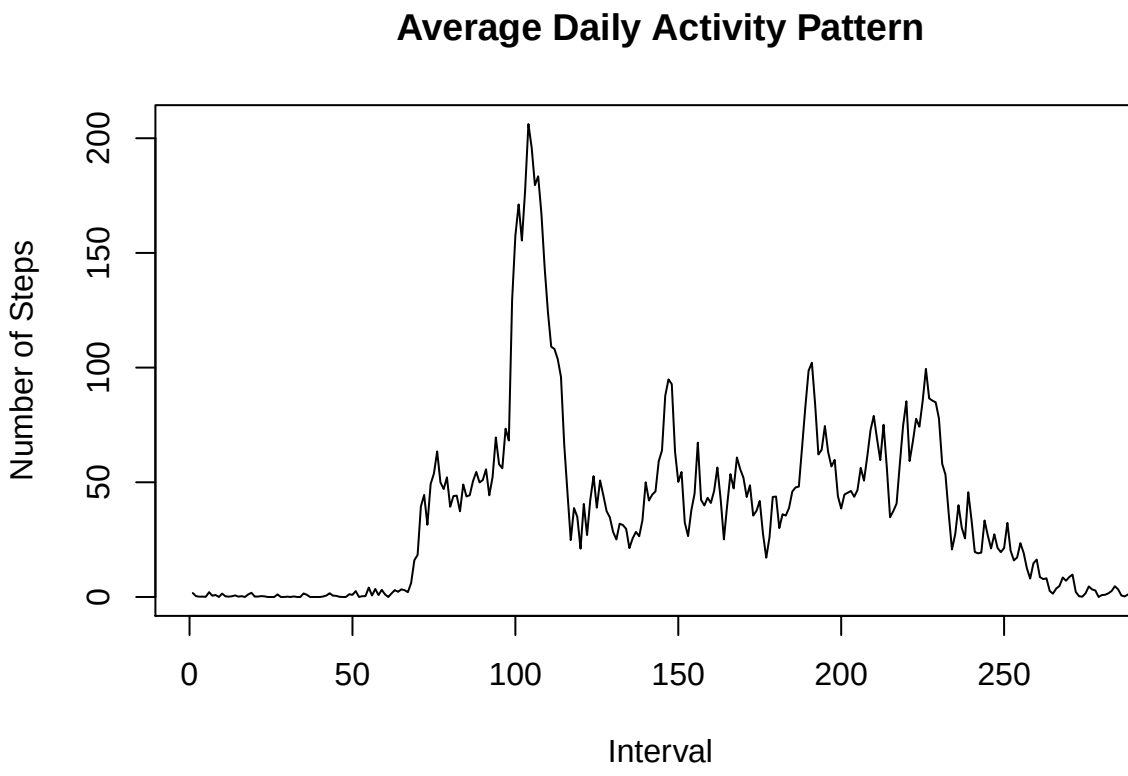
```
#median of the number of steps taken each day
median(steps_by_date)
```

```
## [1] 10395
```

Step 3 : What is the average daily activity pattern?

Time-Series plot of daily activity pattern

```
adap = as.numeric(tapply(activity$steps, activity$interval, mean, na.rm = T))
plot(adap, type="l", xlab = "Interval", ylab = "Number of Steps", main = "Average Daily Activity Patter
```



daily activity pattern-1.pdf

5-minute interval with the maximum number of steps

```
which.max(adap)
```

```
## [1] 104
```

The 104th 5-minute interval (approximately 9 am) has the most number of steps.

Step 4 : Imputing Missing Values

Total Number of rows with missing values (NAs)

```
table(is.na(activity))
```

```
##
## FALSE  TRUE
## 50400  2304
```

There are 2,304 rows with missing values in the "activity"" data frame (all from the "steps" field)

Create dataset with missing values filled by imputed values

```
#use "mice" package
library(mice)
```

```
## Loading required package: Rcpp
## Loading required package: lattice
## mice 2.22 2014-06-10
```

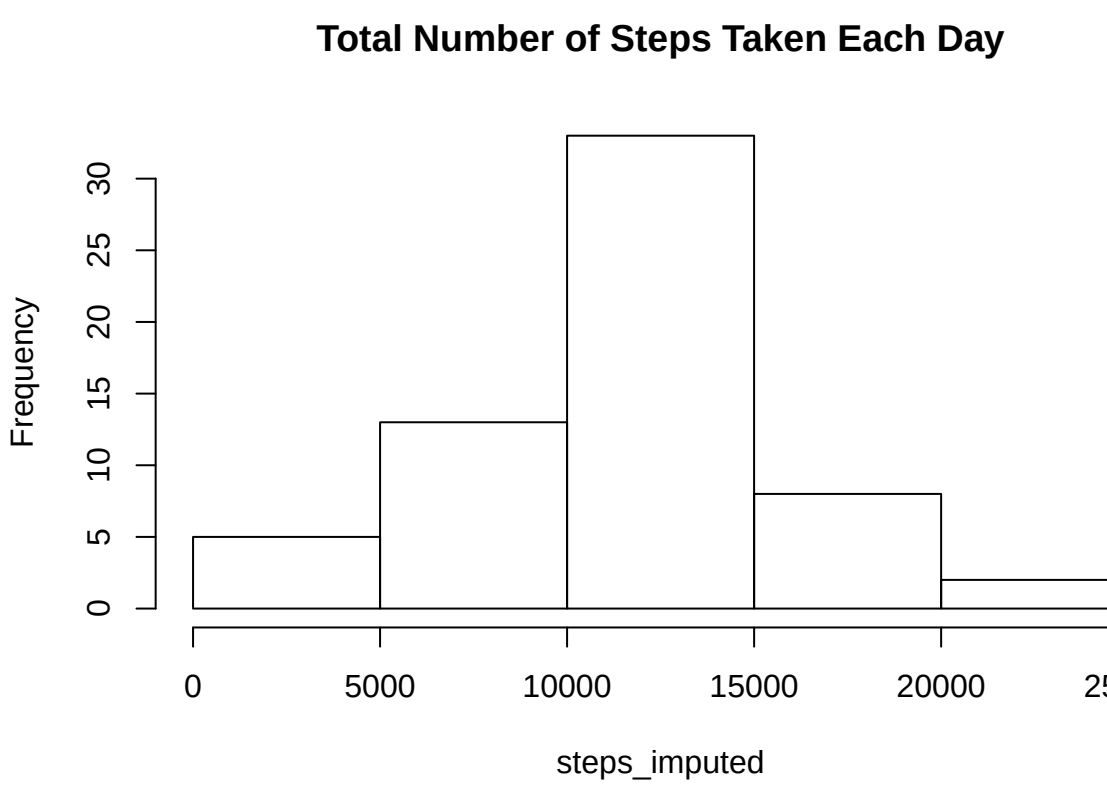
```
#set the seed so as to make it reproducible
set.seed(144)
#do the imputation to create the new dataset("activity2") with the missing data filled in
activity2 = complete(mice(activity))
```

```
##
##  iter imp variable
##   1   1  steps
##   1   2  steps
##   1   3  steps
##   1   4  steps
##   1   5  steps
##   2   1  steps
##   2   2  steps
##   2   3  steps
##   2   4  steps
##   2   5  steps
##   3   1  steps
```

```
##   3   2  steps
##   3   3  steps
##   3   4  steps
##   3   5  steps
##   4   1  steps
##   4   2  steps
##   4   3  steps
##   4   4  steps
##   4   5  steps
##   5   1  steps
##   5   2  steps
##   5   3  steps
##   5   4  steps
##   5   5  steps
```

```
#use tapply to sum the number of steps by date
steps_imputed = as.numeric(tapply(activity2$steps, activity$date, sum, na.rm = T))
#plot a histogram
hist(steps_imputed, main = "Total Number of Steps Taken Each Day")
```

Histogram of the total number of steps taken each day (including imputed data) to create the imputed values histogram-1.pdf



Mean and Median of the Total Number of Steps Taken Each Day

```
#mean of the number of steps taken each day
mean(steps_imputed)
```

```
## [1] 11124.85
```

```
#median of the number of steps taken each day
median(steps_imputed)
```

```
## [1] 11352
```

The histogram is almost the same. Mean and median values, expectedly increase, as additional values have been added.