Project 1, Reproducable Research

Determine the working directory and load the data

```
setwd("C:/Users/rober/Documents/R/Activity Monitoring")
activity<- read.csv("activity.csv", header=TRUE)
df <- data.frame(activity)
```

Question 1:

a) What is mean total number of steps taken per day?

b) Calculate the total number of steps taken per day.

c) Make a histogram of the total number of steps taken each day.

d) Calculate and report the mean and median of the total number of steps taken per day.

b) The script below calculates the total number of steps taken each day.

```
steps<- aggregate(steps~date, data=df, FUN=sum, na.rm=TRUE)
```

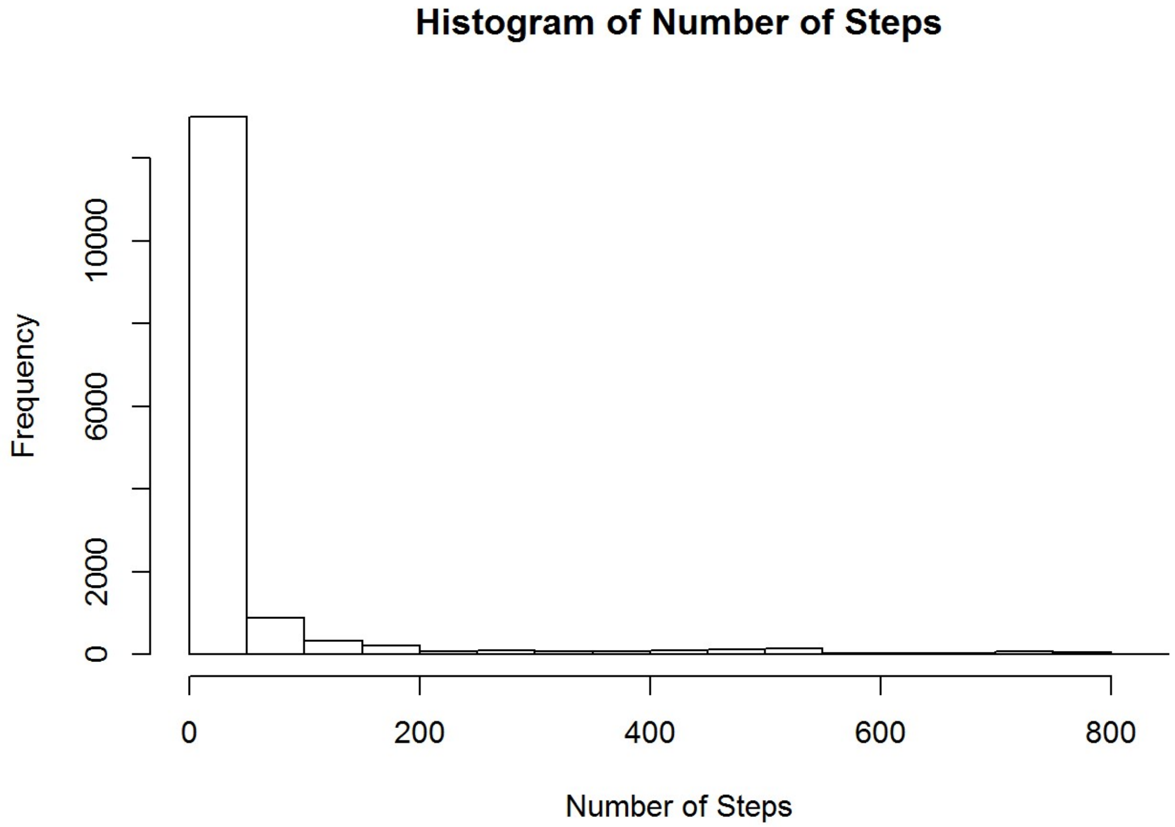
a/d) The mean number of steps is 10,766 and the median number of steps is 10,765.

```
summary(steps)
```

```
##          date        steps      
##  2012-10-02: 1   Min.   :   41  
##  2012-10-03: 1   1st Qu.: 8841  
##  2012-10-04: 1   Median :10765  
##  2012-10-05: 1   Mean   :10766  
##  2012-10-06: 1   3rd Qu.:13294  
##  2012-10-07: 1   Max.   :21194  
##  (Other)   :47
```

c) See histogram for number of steps.

```
hist(activity$steps, xlab= "Number of Steps", ylab="Frequency", main="Histogram
of Number of Steps")
```



Question 2

a) What is the average daily activity pattern?

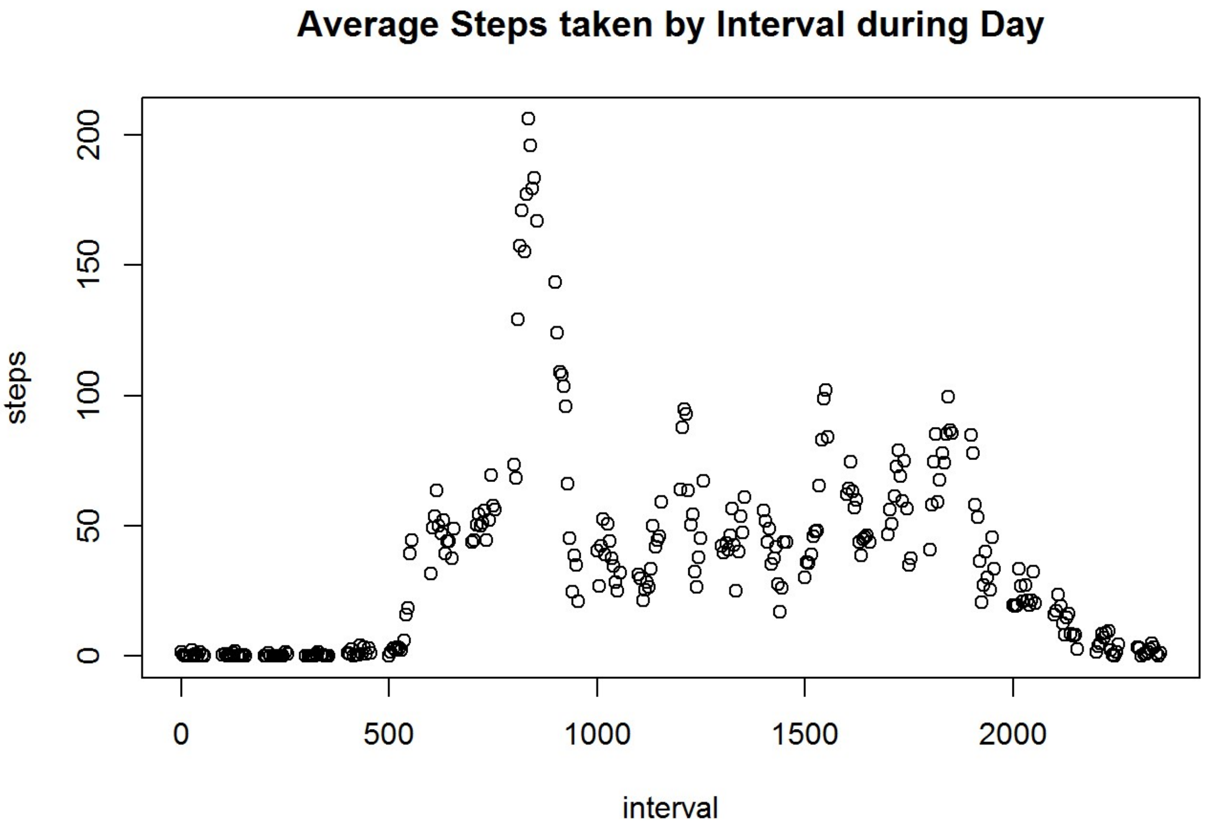
b) Make a time series plot (i.e. type = “l”) of the 5-minute interval (x-axis) and the average number of steps taken, averaged ## across all days (y-axis)

c) Which 5-minute interval, on average across all the days in the dataset, contains the maximum number of steps?

d) What is the average daily activity pattern?

b) time series plot

```
x<- df$interval
y<- df$steps
meanbyinterval <- aggregate(steps~interval, df, mean, na.rm=TRUE)
plot(meanbyinterval, main="Average Steps taken by Interval during Day")
```



```
max(meanbyinterval$steps)
```

```
## [1] 206.1698
```

c) the maximum of the average number of steps taken is 206. Based on the meaninterval data table this corresponds to the interval #835.

d) the average daily activity pattern is unimodal with a peak occuring at the interval 835, ranging between 50 and 100 between the #intervals 500 and 2000, and approximately zero on either side of those intervals.

Question 3

a) Calculate and report the total number of missing values in the dataset (i.e. the total number of rows with NAs)

b) Create a new dataset that is equal to the original dataset but with the missing data filled in.

c) Make a histogram of the total number of steps taken each day and Calculate and report the mean and median total number of steps ##taken per day. Do these values differ from the estimates from the first part of the assignment? What is the impact of imputing ##missing data on the estimates of the total daily number of steps?

```
sum(is.na(activity$steps))
```

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

```
sum(is.na(activity$date))
```

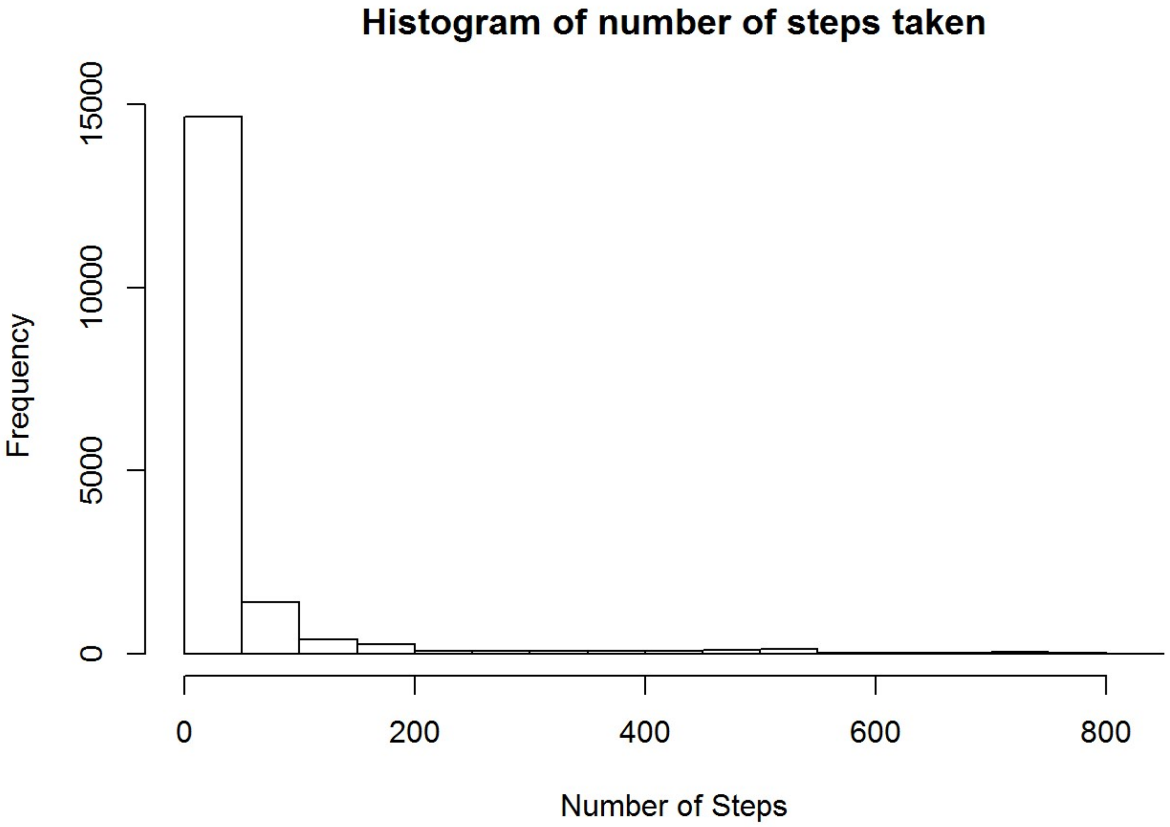
```
## [1] 0
```

```
sum(is.na(activity$interval))
```

```
## [1] 0
```

```
activity_replacena <- ifelse(is.na(activity$steps) == TRUE, meanbyinterval$steps[activity$interval %in% meanbyinterval$interval], activity$steps)

incomplete <- sum(!complete.cases(activity))
imputed_data <- transform(activity, steps = ifelse(is.na(activity$steps), meanbyinterval$steps[match(activity$interval, meanbyinterval$interval)], activity$steps))
hist(imputed_data$steps, xlab="Number of Steps", main="Histogram of number of steps taken")
```



```
summary(imputed_data)
```

```
##      steps                 date          interval
##  Min.   :  0.00   2012-10-01:  288   Min.   :   0.0
##  1st Qu.:  0.00   2012-10-02:  288   1st Qu.: 588.8
##  Median :  0.00   2012-10-03:  288   Median :1177.5
##  Mean   : 37.38   2012-10-04:  288   Mean   :1177.5
##  3rd Qu.: 27.00   2012-10-05:  288   3rd Qu.:1766.2
##  Max.   :806.00   2012-10-06:  288   Max.   :2355.0
##                   (Other)   :15840
```

a) The total number of “NA” in the Steps column is 2,304. There are no NA’s in the date or interval columns.

b) See below for the histogram.

c) Once the imputed data have been added the mean and median values stay the same as the original activity data.

Question 4:

Are there differences in activity patterns between weekdays and weekends?

For this part the weekdays() function may be of some help here. Use the dataset with the filled-in missing values for this part.

a) Create a new factor variable in the dataset with two levels - “weekday” and “weekend” indicating whether a given date is a ##weekday or weekend day.

b) Make a panel plot containing a time series plot (i.e. type = “l”) of the 5-minute interval (x-axis) and the average number of ## steps taken, averaged across all weekday days or weekend days (y-axis). See the README file in the GitHub repository to see an ##example of what this plot should look like using simulated data.

```
imputed_data_day <- weekdays(as.Date(imputed_data$date))
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.3.2
```

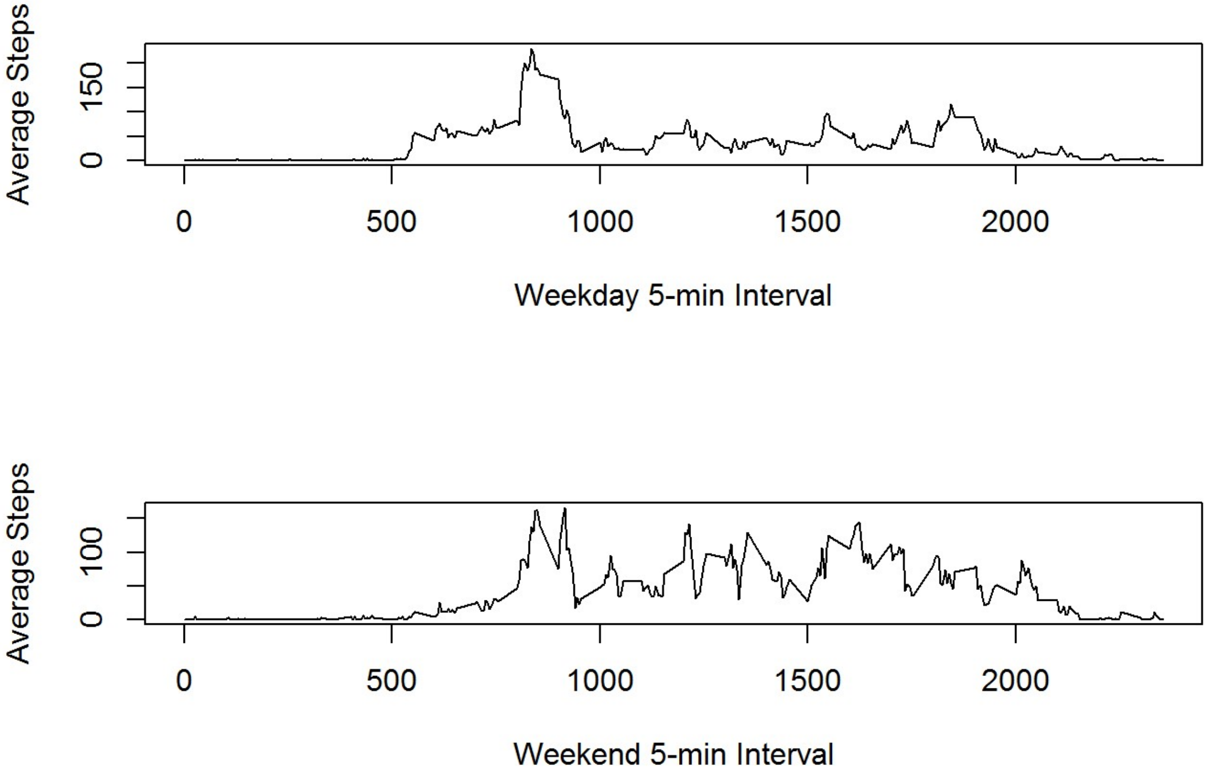
```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
imputed_data$date <- as.Date(imputed_data$date)
weekdays1 <- c('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday')
imputed_data$Day <- factor((weekdays(imputed_data$date) %in% weekdays1), levels=c(FALSE, TRUE), labels=c('weekend', 'weekday'))
colnames(imputed_data) <- c("Steps","Date", "Interval", "Day_Type")

par(mfrow=c(2,1))
weekday <- subset(imputed_data, Day_Type=="weekday")
weekday_mean <- aggregate(Steps ~ Interval, weekday, mean, na.rm=TRUE)
plot(weekday_mean, type="l", xlab="Weekday 5-min Interval", ylab="Average Steps")
weekend <- subset(imputed_data, Day_Type=="weekend")
weekend_mean <- aggregate(Steps ~ Interval, weekend, mean, na.rm=TRUE)
plot(weekend_mean, type="l", xlab="Weekend 5-min Interval", ylab="Average Steps")
```



#a) See above for the new factor variable. #b) See above for the weekday and weekend panel plots.