# Reproducible Research Course Project 1

# Introduction

This assignment makes use of data from a personal activity monitoring device. This device collects data at 5 minute intervals through out the day. The data consists of two months of data from an anonymous individual collected during the months of October and November, 2012 and include the number of steps taken in 5 minute intervals each day.

This document provides the step-by-step analysis with both the code and data. The data is sourced from [Activity Monitoring Data](https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip).

## Analysis

The analysis is comprised of the following steps:  
- Loading and Preprocessing the Data  
- Calculating and Plotting the Mean Total Number of Steps  
- Plotting the Average Daily Activity Pattern  
- Imputing Missing Values  
- Plotting the Differences in Activity between Weekdays and Weekends

### Loading and Preprocessing the Data

The data is loaded and preprocessed using the following R code:

library(dplyr)

##   
## 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

library(lattice)  
###############################################################################################################  
# Loading and preprocessing the data  
###############################################################################################################  
##setwd("C:/Users/leedo002/Documents/Education/Data Science by JHU/Course 5 - Reproducible Research/Week 2/Course Project")  
##activity<-read.csv("activity.csv")   
activity<-read.csv(unz("activity.zip", "activity.csv"))

### Calculating and Plotting the Mean Total Number of Steps

The first analytical step of this analysis is to answer the question "*What is mean total number of steps taken per day?*"

This step performs the following while ignoring missing values:  
1. Calculate the total number of steps taken per day  
2. Plot a histogram of the total number of steps taken each day  
3. Calculate and report the mean and median of the total number of steps taken per day

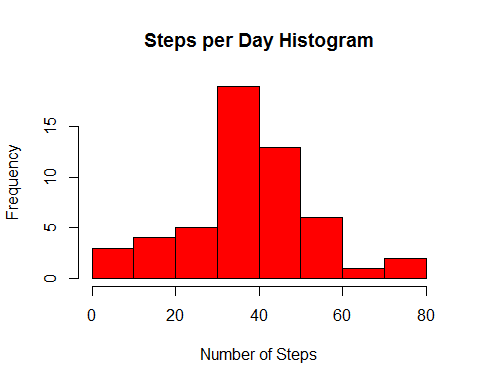
The total number of steps taken per day was calculated using the following R code:

TotNumSteps<- sum(activity$steps,na.rm=TRUE)

The result of this calculation is 570608.

The histogram of the total number of steps taken each day was plotted using the following R code:

AggActNoNA <- aggregate(steps ~ date, data=activity, mean, na.rm=TRUE)  
hist(AggActNoNA$steps  
 , col="red"  
 , main="Steps per Day Histogram"  
 , xlab="Number of Steps"  
 )



The mean and median of the total number of steps take per day were calculated using the following R code:

StepsPerDayMeanNoNA<-mean(AggActNoNA$steps)  
StepsPerDayMedianNoNA<-median(AggActNoNA$steps)

The mean calculated was 37.3825996  
and the median calculated was 37.3784722.

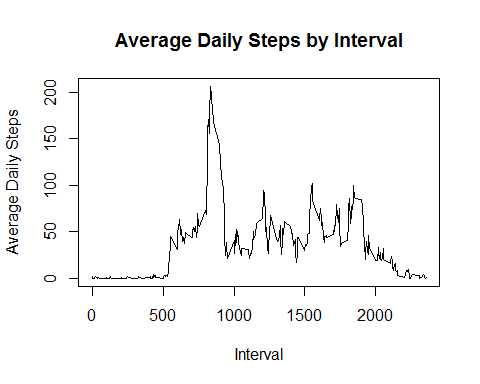
### Plotting the Average Daily Activity Pattern

The first analytical step of this analysis is to answer the question "*What is the average daily activity pattern?*"

This step performs the following:  
1. 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)  
2.Which 5-minute interval, on average across all the days in the dataset, contains the maximum number of steps?

The time series plot of the average number of steps taken each day by interval was plotted using the following R code:

AggStepsByIntervalNoNA <- aggregate(steps ~ interval, data=activity, mean, na.rm=TRUE)  
plot( AggStepsByIntervalNoNA$interval  
 , AggStepsByIntervalNoNA$steps  
 , type="l"  
 , xlab="Interval"  
 , ylab="Average Daily Steps"  
 , main="Average Daily Steps by Interval"  
 )



The interval with the maximum of the total number of steps taken each day was calculated using the following R code:

IntervalWithMostSteps<-AggStepsByIntervalNoNA[AggStepsByIntervalNoNA$steps==max(AggStepsByIntervalNoNA$steps),]$interval

The result of the calculation was 835.

### Imputing Missing Values

The third analytical step of this analysis is to answer the question "*What is the impact on the total daily number of steps when imputing missing values?*"

This step performs the following: 1.Calculate and report the total number of missing values in the dataset (i.e. the total number of rows with NAs)  
2.Devise a strategy for filling in all of the missing values in the dataset. The strategy does not need to be sophisticated. For example, you could use the mean/median for that day, or the mean for that 5-minute interval, etc.  
3.Create a new dataset that is equal to the original dataset but with the missing data filled in.  
4.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?

The total number of missing values in the dataset was calculated using the following R code:

TotMissingValues<-sum(is.na(activity$steps))

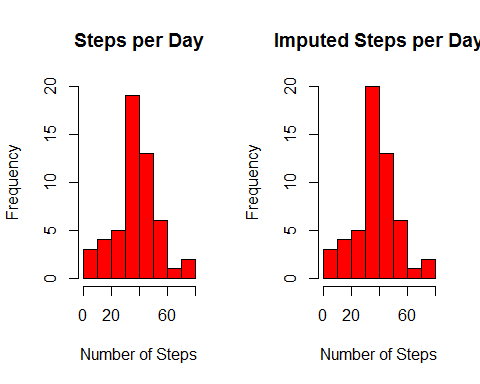
The result of the calculation was 2304.

All of the missing values were populated by using the mean of the corresponding interval across all days. The following R code was used to impute the values and populate each missing value:

activityImputed<-activity  
naTrue<-as.numeric(rownames(activityImputed[is.na(activityImputed$steps),]))  
activityImputed[naTrue,]$steps <- AggStepsByIntervalNoNA[AggStepsByIntervalNoNA$interval==  
 activityImputed[naTrue,]$interval,]$steps

A histogram was plotted for the total number of steps taken each day using the following R code:

AggActImputed <- aggregate(steps ~ date, data=activityImputed, mean)  
par(mfrow=c(1,2))  
  
#### last histogram  
hist(AggActNoNA$steps  
 , col="red"  
 , main="Steps per Day"  
 , xlab="Number of Steps"  
 , ylim=c(0,20)  
)  
  
### this histogram  
hist(AggActImputed$steps  
 , col="red"  
 , main="Imputed Steps per Day"  
 , xlab="Number of Steps"  
)



And finally the mean and median of the total number of steps taken per day were calculated using the following R code:

StepsPerDayMeanImputed<-mean(AggActImputed$steps)  
StepsPerDayMedianImputed<-median(AggActImputed$steps)  
  
MeanDiffPct <- ((StepsPerDayMeanImputed - StepsPerDayMeanNoNA)/StepsPerDayMeanNoNA)\*100  
MedianDiffPct <- ((StepsPerDayMedianImputed - StepsPerDayMedianNoNA)/StepsPerDayMedianNoNA)\*100

The value of the mean was 37.3825996  
and the percent difference compared to the mean with missing values removed was 0 .  
The value of the median was 37.3805359  
and the percent difference compared to the mean with missing values removed was 0.005521 .

### Plotting the Differences in Activity between Weekdays and Weekends

The final analytical step of this analysis is to answer the question "*Are there differences in activity patterns between weekdays and weekends?*"

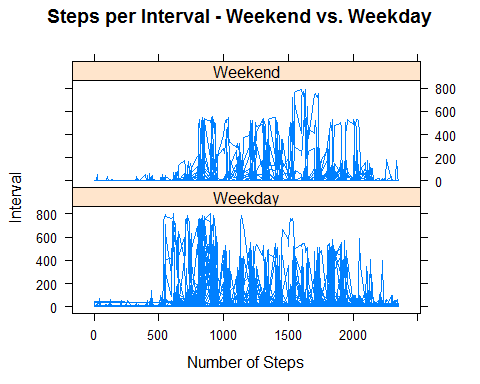
This step performs the following: 1.Create a new factor variable in the dataset with two levels - "weekday" and "weekend" indicating whether a given date is a weekday or weekend day.  
2.Make a panel plot containing a time series plot (i.e. type = "l") of the 5-minute interval (x-axis) and the average 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.

The new factor variable for the type of day was created using the following R code:

actImpDay <- transform( activityImputed  
 , DayNum=as.numeric(strftime(activityImputed$date,"%u"))  
 , DayType=as.character("")   
 , stringsAsFactors=FALSE  
 )  
  
actImpDay[row.names(actImpDay[actImpDay$DayNum<6,]),]$DayType<-"Weekday"  
actImpDay[row.names(actImpDay[actImpDay$DayNum>5,]),]$DayType<-"Weekend"  
  
actImpDayTyp <- transform( actImpDay  
 , date=actImpDay$date, interval=actImpDay$interval, steps=actImpDay$steps  
 , DayType=as.factor(actImpDay$DayType)  
 )

A panel time series plot of the average number of steps taken across weekday days vs. weekend days by 5-minute interval was plotted using the following R code:

xyplot(steps~interval | DayType  
 , actImpDayTyp  
 , type="l"  
 , layout=c(1,2)  
 , xlab="Number of Steps"  
 , ylab="Interval"  
 , main="Steps per Interval - Weekend vs. Weekday"  
 )



# End of Document