Week 2- Assignment

Loading and preprocessing the data

Load the data (i.e. read.csv())

```
setwd("/Coursera/Reproducible Research/Week 2- Assignment")
Activity_dat<-read.csv("activity.csv")
```

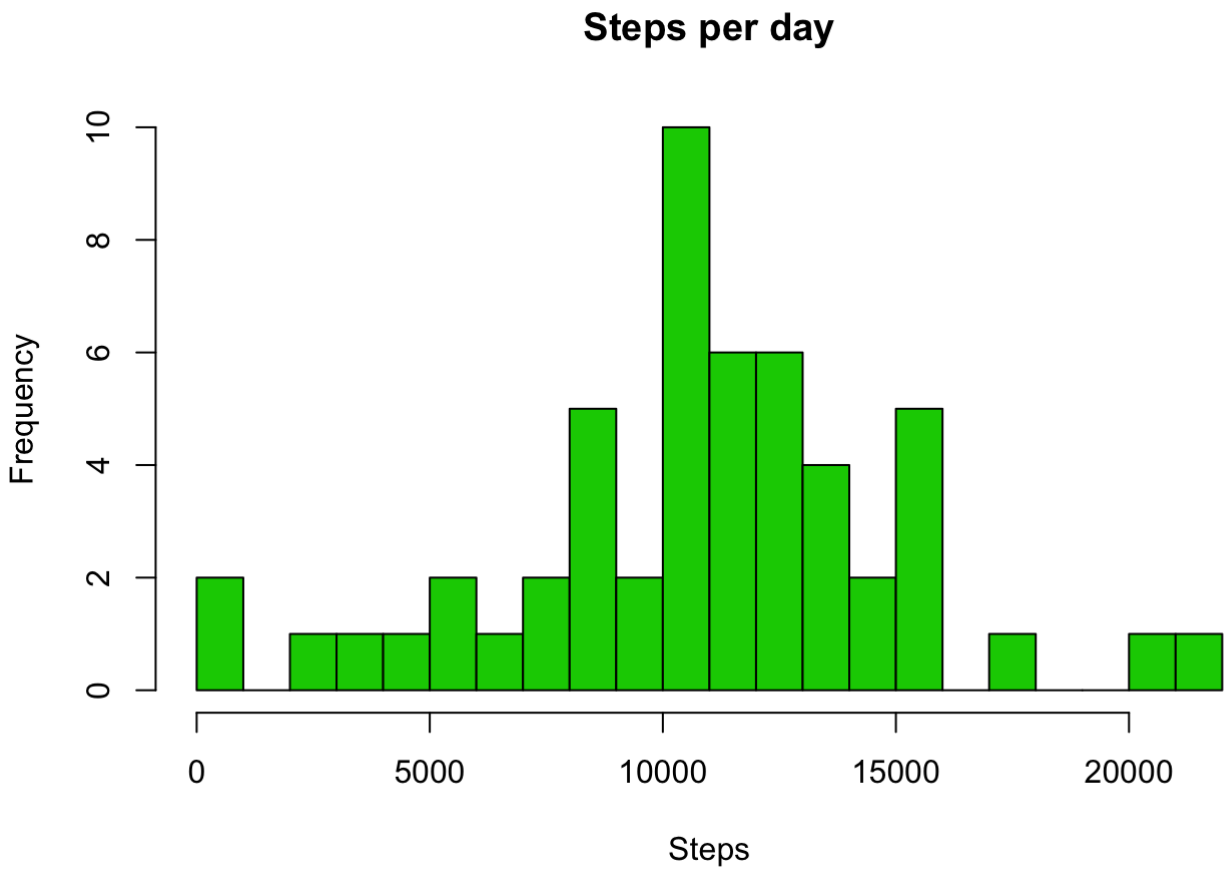
Process/transform the data (if necessary) into a format suitable for your analysis

```
Activity_dat$date<-as.Date(Activity_dat$date)
```

What is mean total number of steps taken per day?

Make a histogram of the total number of steps taken each day

```
Agg_Activity<- data.frame(aggregate(x=Activity_dat$steps,by=list(Date=Activity_dat$date)
, FUN=sum))
Agg_Activity$x<-as.numeric(Agg_Activity$x)
hist(Agg_Activity$x,xlab="Steps",ylab="Frequency",main="Steps per day",col=3,breaks=25)
```



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

```
Agg_Activity<-na.omit(Agg_Activity)
mean(Agg_Activity$x)
```

```
## [1] 10766.19
```

```
median(Agg_Activity$x)
```

```
## [1] 10765
```

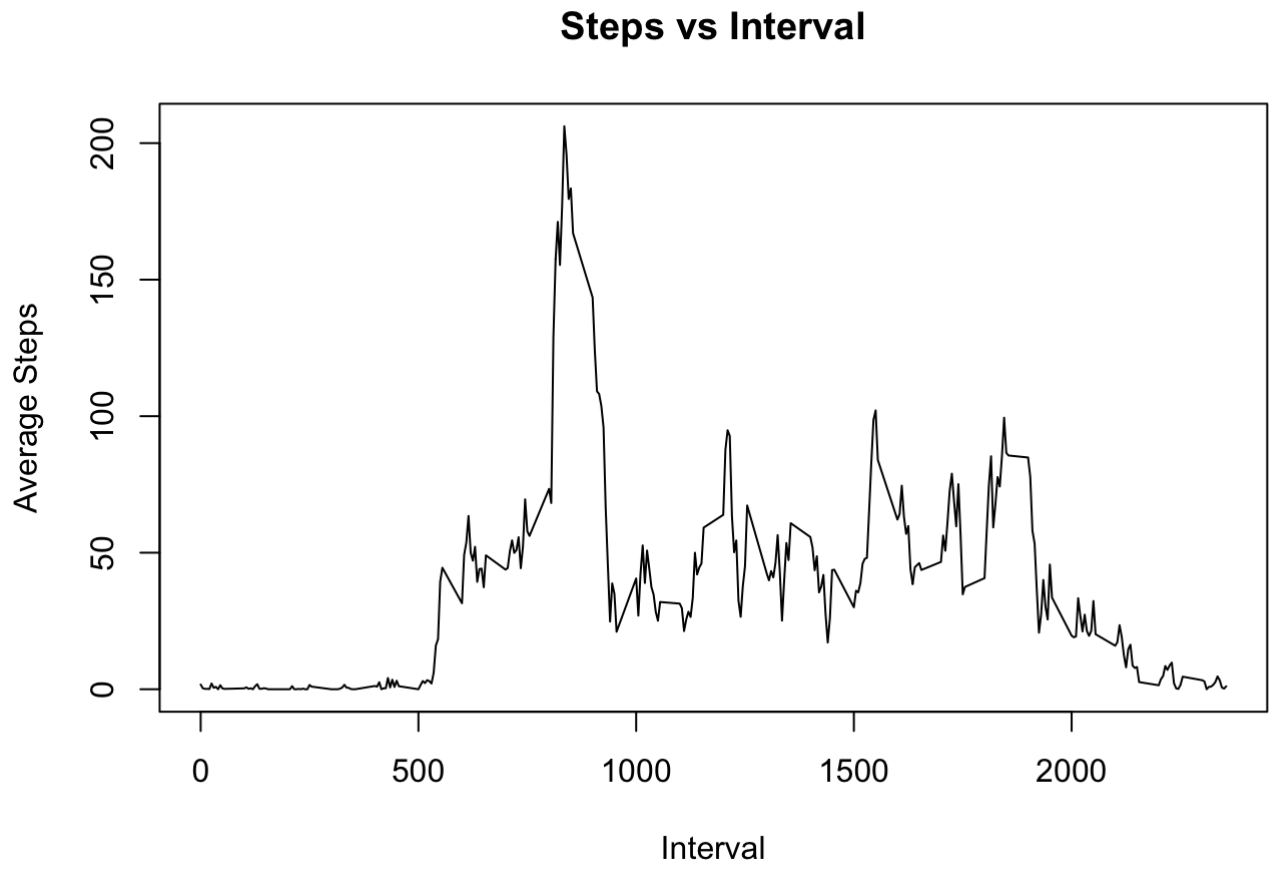
What is the average daily activity pattern?

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)

```
Activity_dat_omit<-na.omit(Activity_dat)
```

Aggregating the data

```
Agg_Activity<- data.frame(aggregate(x=Activity_dat_omit$steps,by=list(interval=Activity_
dat_omit$interval), FUN=mean))
plot(Agg_Activity$interval,Agg_Activity$x,type='l',xlab="Interval",ylab="Average Steps",
main="Steps vs Interval")
```



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

```
Agg_Activity[which.max(Agg_Activity$x),]
```

```
##     interval        x
## 104      835 206.1698
```

Imputing missing values

Note that there are a number of days/intervals where there are missing values (coded as NA). The presence of missing days may introduce bias into some calculations or summaries of the data.

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

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

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

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.

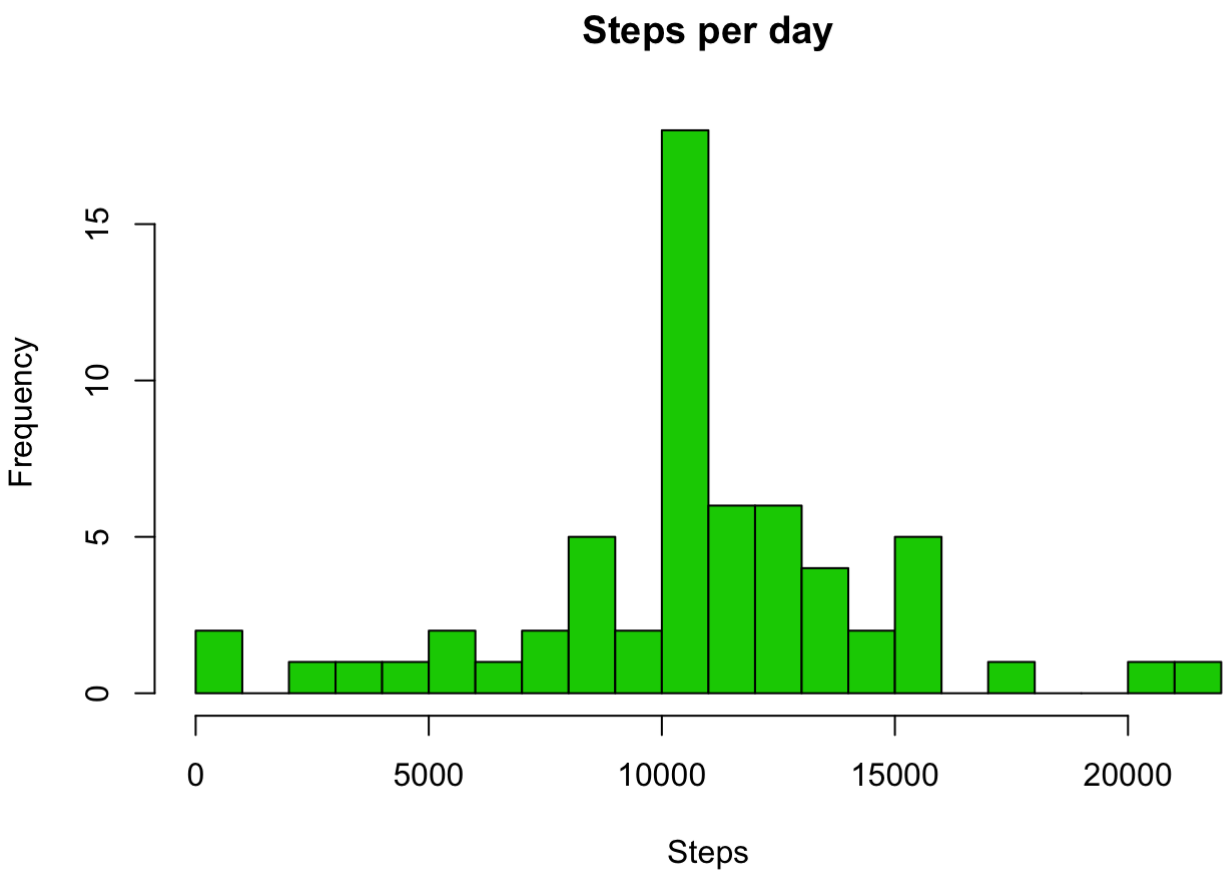
```
library(zoo)
Activity_dat_NA<-Activity_dat
Activity_dat_NA$steps<-na.aggregate(Activity_dat_NA$steps,by=Activity_dat_NA$interval,FU
N=mean)
```

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

```
Activity_dat_new<-Activity_dat_NA
```

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?

```
Agg_Activity_new<- data.frame(aggregate(x=Activity_dat_new$steps,by=list(Date=Activity_d
at_new$date), FUN=sum))
Agg_Activity_new$x<-as.numeric(Agg_Activity_new$x)
hist(Agg_Activity_new$x,xlab="Steps",ylab="Frequency",main="Steps per day",col=3,breaks
= 25)
```




```
mean(Agg_Activity_new$x)
```

```
## [1] 10766.19
```

```
median(Agg_Activity_new$x)
```

```
## [1] 10766.19
```

The mean seems same as the previous values, the median seems to have changed.

Are there differences in activity patterns between weekdays and weekends?

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.

```
Activity_dat_new$Day<-weekdays(Activity_dat_new$date)
Activity_dat_new$Weekend <- as.factor(ifelse(Activity_dat_new$Day =="Sunday"|Activity_da
t_new$Day =="Saturday", "Weekend", "Weekday"))
```

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).

```
library(ggplot2)
Agg_Activity_final<- data.frame(aggregate(x=Activity_dat_new$steps,by=list(interval=Acti
vity_dat_new$interval,weekend=Activity_dat_new$Weekend), FUN=mean))
g<-qplot(interval,x,data=Agg_Activity_final,facets = weekend~.,geom="line",main = "Weeke
nd vs Weekday",ylab = "Steps",xlab = "Interval")
print(g)
```