Reproducible Research Project 1

Step Zero- Make sure you're in the right directory with 'activity.csv' First, load in the data and give a quick summary of the dataset.

```
data<-read.csv('activity.csv')
summary(data)
```

```
##      steps                date          interval
##  Min.   :  0.0   2012-10-01:  288   Min.   :   0
##  1st Qu.:  0.0   2012-10-02:  288   1st Qu.: 589
##  Median :  0.0   2012-10-03:  288   Median :1178
##  Mean   : 37.4   2012-10-04:  288   Mean   :1178
##  3rd Qu.: 12.0   2012-10-05:  288   3rd Qu.:1766
##  Max.   :806.0   2012-10-06:  288   Max.   :2355
##  NA's   :2304    (Other)   :15840
```

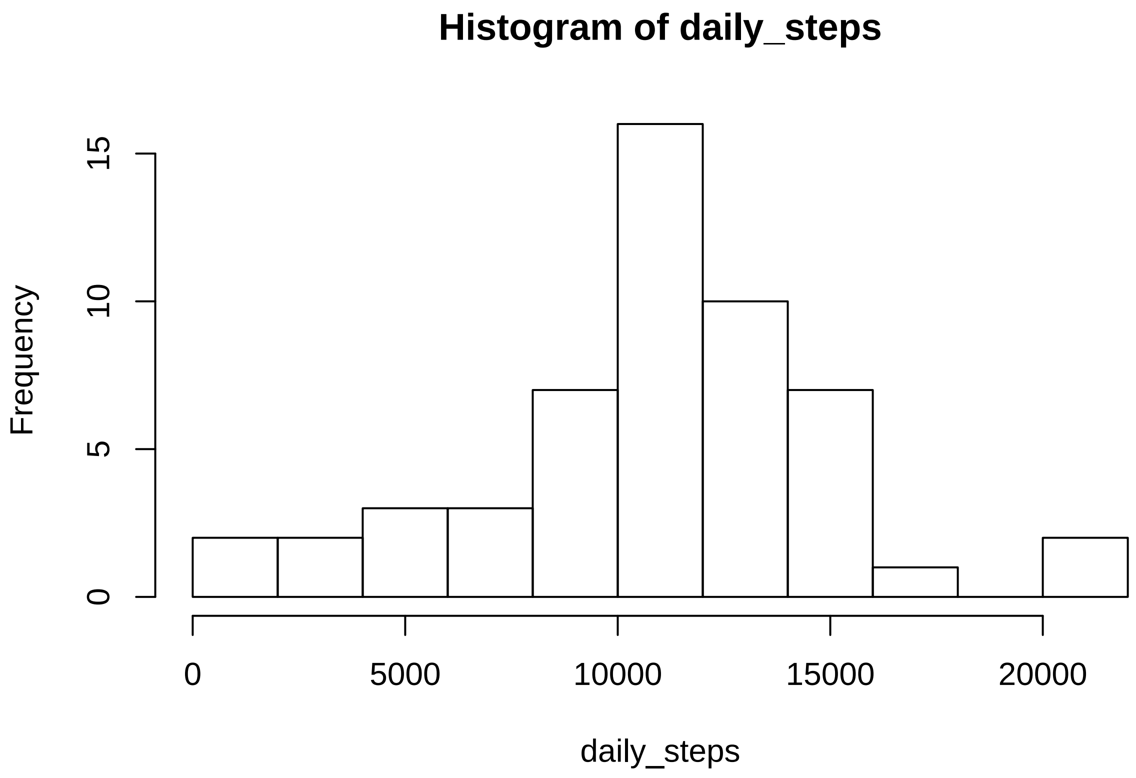
Steps per day

Now find the total number, mean, and median of steps per day. Note, NAs are removed at this point.

```
daily_steps<-tapply(data$steps,as.factor(data$date),sum)
daily_mean=mean(daily_steps, na.rm=TRUE)
daily_median=median(daily_steps, na.rm=TRUE)
```

And produce histogram of daily steps.

```
hist(daily_steps,breaks=15)
```



The mean number of steps per day is 1.0766 × 104 and the median number of steps per day is 10765.

Average Daily Activity Pattern

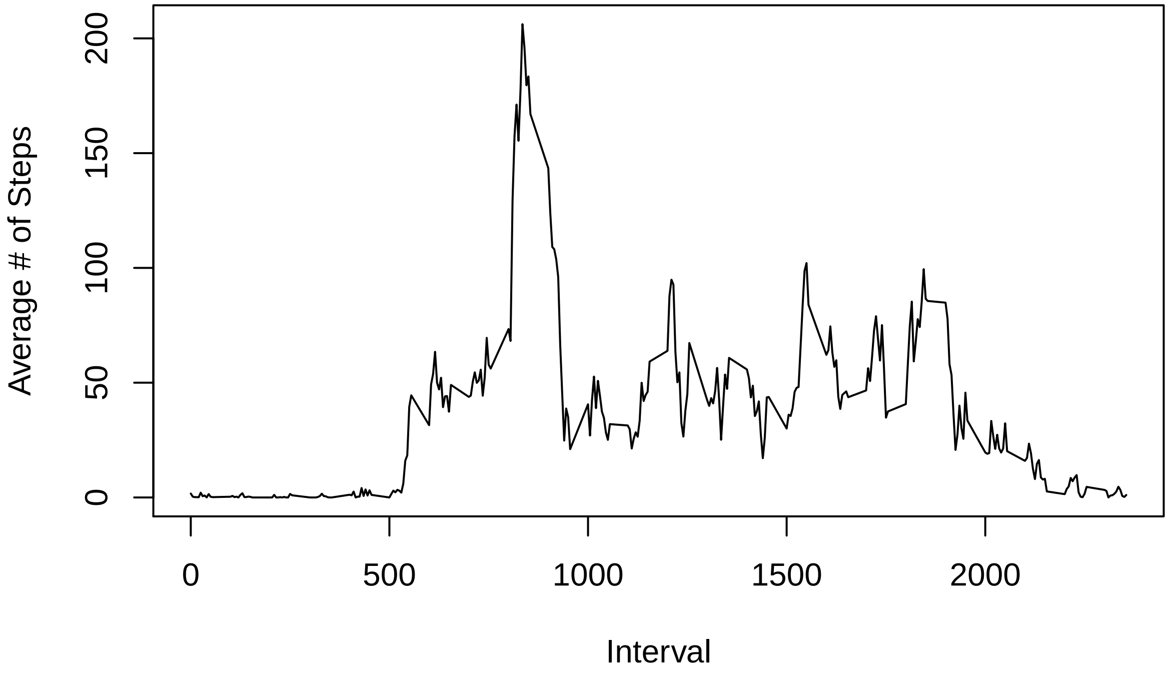
Goal- Make a time series plot of the 5-minute interval and the average number of steps taken, averaged across all days (y-axis).

First, find the average number of steps taken for each interval.

```r
interval_mean<-tapply(data$steps,as.factor(data$interval),mean,na.rm=TRUE)
unique_interval=unique(data$interval)
```

Then plot.

```r
plot(unique(data$interval),interval_mean,type='l',xlab='Interval', ylab='Average # of Steps')
```



On average, the 835th 5-minute interval has the most number of steps.

Impact of missing values

How many of the reported intervals have 'NA' steps?

```r
good=table(data$steps != 'NA')
```

There are 17568 data points in the dataset, 15264 of which are real reported values, leaving 2304 missing values.

Now, if we replace the missing values for each interval with the average of that interval how does the mean and median of the entire dataset change?

First do the replacing-

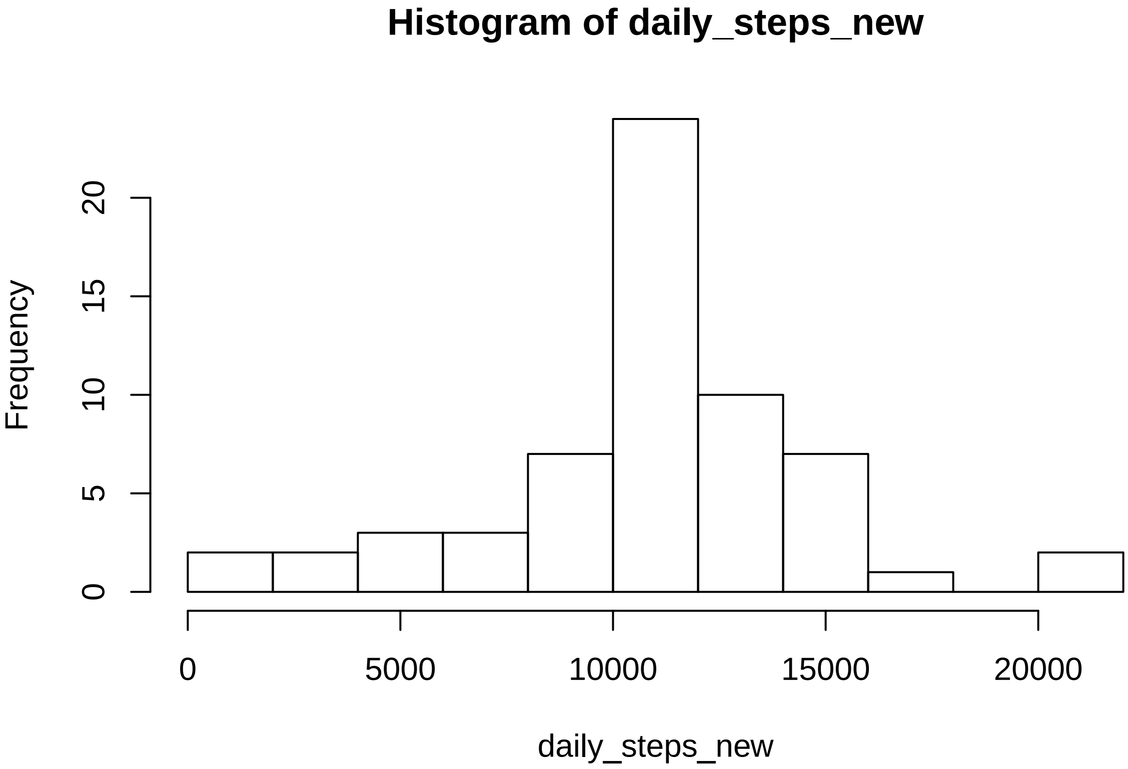
```
data2<-data
for (i in seq_along(data$steps)) {if (is.na(data2$steps[i]) == 'TRUE') {data2$steps[i]<-mean(data2$steps
```

Then find the sum of the steps per day, along with the mean and median.

```
daily_steps_new<-tapply(data2$steps,as.factor(data2$date),sum)
daily_mean_new=mean(daily_steps_new)
daily_median_new=median(daily_steps_new)
```

Finally, plot the new histogram.

```
hist(daily_steps_new,breaks=15)
```



The mean(median) number of steps taken for the dataset with replaced missing values is 1.0766 × 104(1.0766 × 104). These values are extremely close to those computed from the dataset with missing values- replacing the missing values with the mean of the interval does not appear to significantly alter the distribution of the data.

Difference in activity between weekdays and weekends

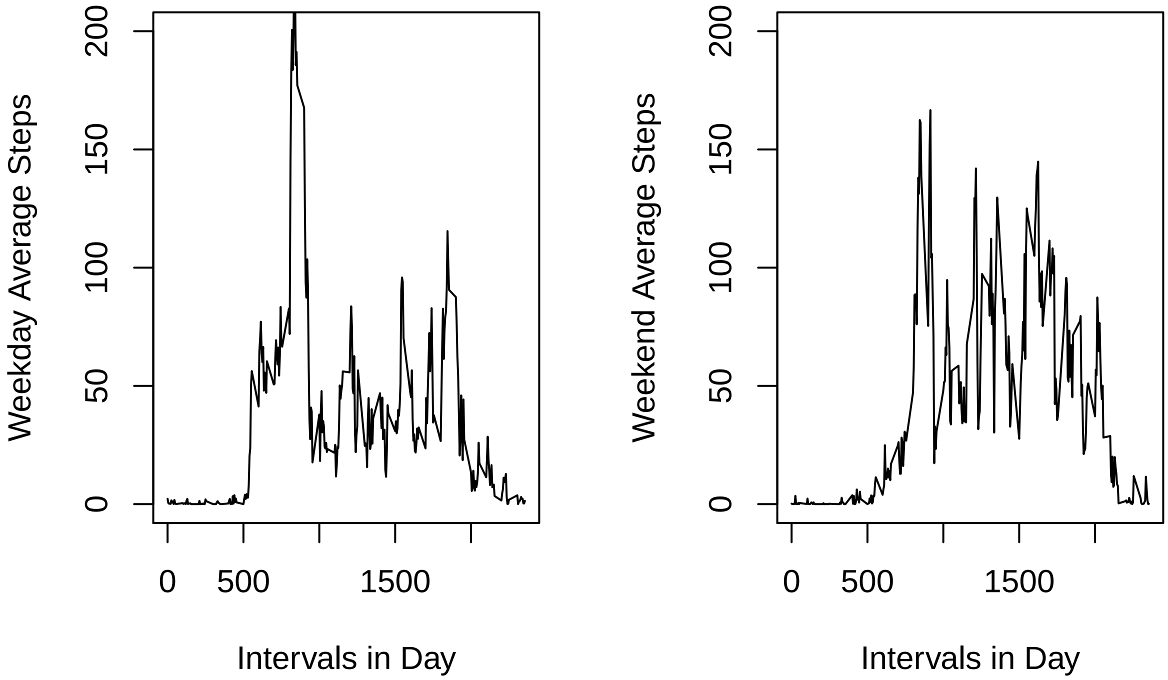
Next I'll look at how the level of activity changes for weekdays vs. weekends. I'll first convert each date to its week day and then make a factor for weekday and weekends.

```
days_oweek <- as.factor(ifelse(weekdays(as.Date(data2$date)) %in% c("Saturday","Sunday"), "Weekend", "We
```

And now plot the average number of steps taken of each 5-minute interval for averaged across all days, seperated by weekday and weekend days.

```
library(lattice)
intervals_inday=data2$interval[data2$date==unique(data2$date)[1]]
weekday_steps<-tapply(data2$steps[days_oweek=='Weekday'],as.factor(data2$interval[days_oweek=='Weekday']
weekend_steps<-tapply(data2$steps[days_oweek=='Weekend'],as.factor(data2$interval[days_oweek=='Weekend']


par(mfrow = c(1, 2))
mar = c(1, 1, 2, 1)
plot(intervals_inday,weekday_steps,type='l', ylim=c(0,200), xlab='Intervals in Day', ylab='Weekday Avera
plot(intervals_inday,weekend_steps,type='l', ylim=c(0,200), xlab='Intervals in Day', ylab='Weekend Avera
```



In general, weekend days show a slightly higher level of activity throughout the day. However, weekdays have a much higher level of activity early in the day.