Reproducible Research - Course Project 1

Nick Rumbaugh

March 19, 2018

Course Project 1

Loading and Pre-processing the Data

The first thing to do is load in the data. I do this assuming the data is in the current directory:

```
df <- read.csv('./activity.csv')
```

We want the dates to be in a date format (at least for the final part of the project), so we need to process that column.

```
df$date = as.POSIXct(df$date)
```

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

First, let's filter out the missing values.

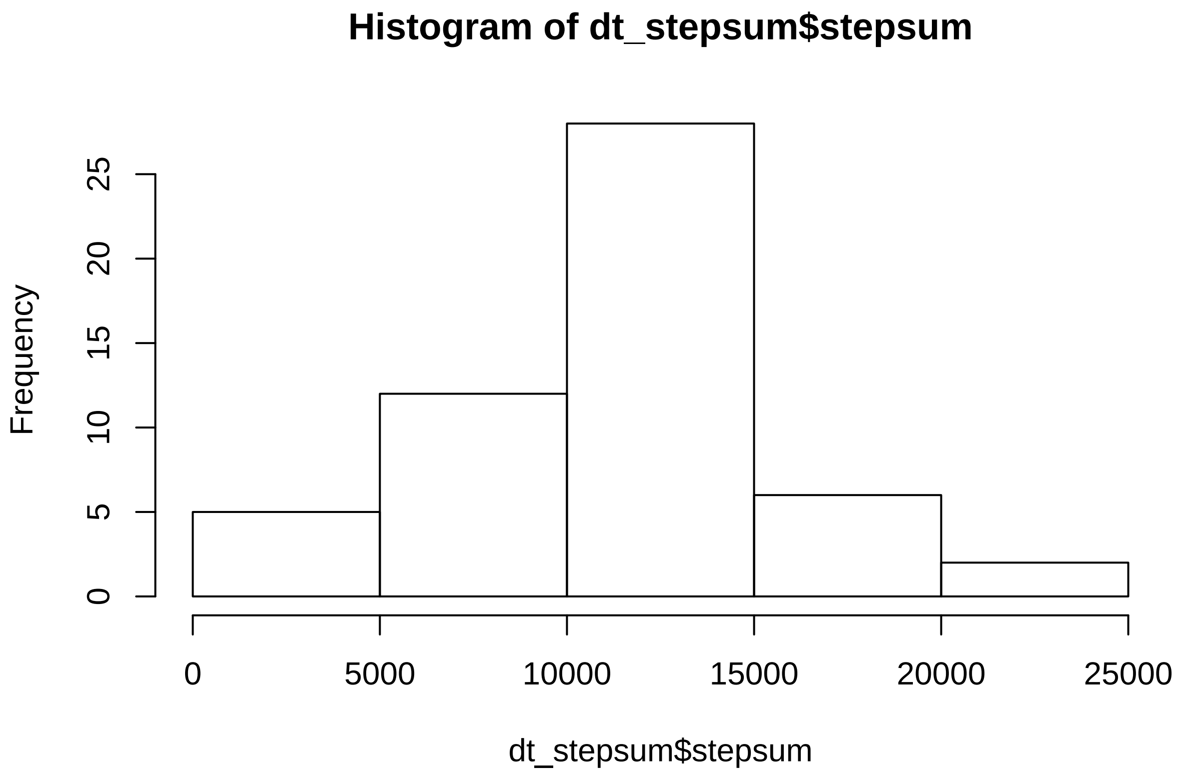
```
missingvalues <- is.na(df$steps)
df_no_mv <- df[!missingvalues,]
```

Now, we can calculate the total number of steps for each day. To do this, I'll use the data.table library.

```
library(data.table)
dt <- data.table(df_no_mv)
dt_stepsum = dt[,list(stepsum=sum(steps)),by=date]
```

With this table, we can plot up a histogram:

```
hist(dt_stepsum$stepsum)
```



And find the mean and median:

```
mean(dt_stepsum$stepsum)
```

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

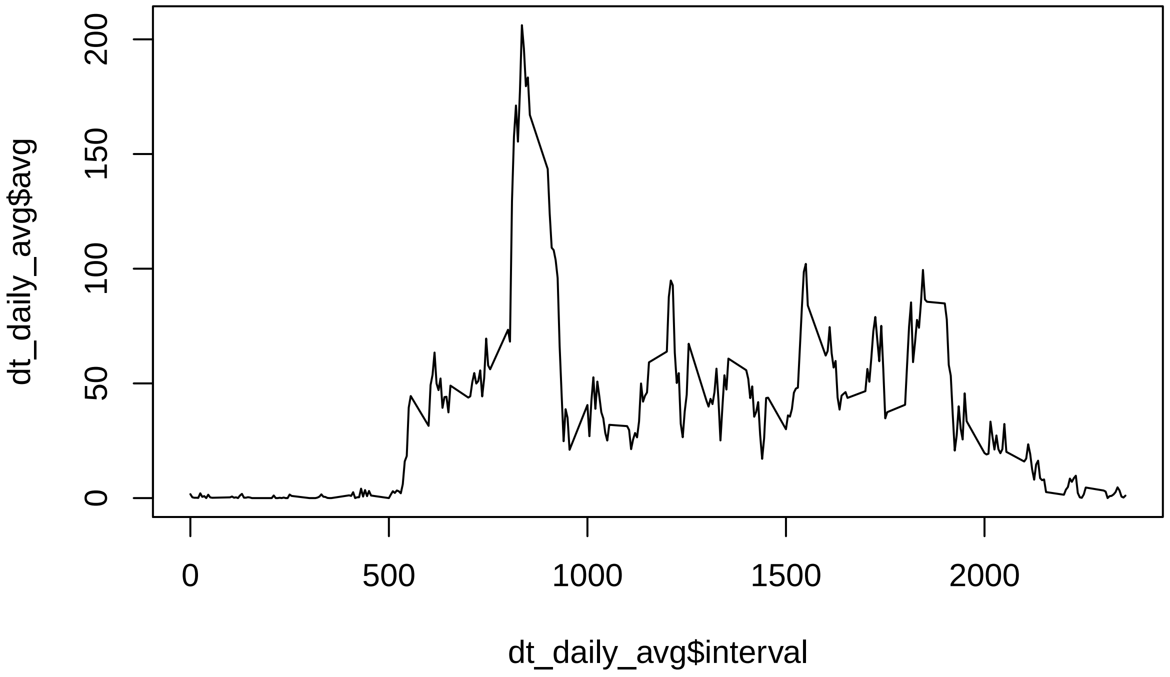
```
median(dt_stepsum$stepsum)
```

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

What is the average daily activity pattern?

This time, we want to group based on interval, not date.

```
dt_daily_avg = dt[,list(avg=mean(steps)),by=interval]
plot(dt_daily_avg$interval, dt_daily_avg$avg, type='l')
```



Now, let's find the maximum

```
dt_daily_avg[order(dt_daily_avg$avg, decreasing = TRUE)][1,]
```

```
##    interval      avg
## 1:      835 206.1698
```

It occurs at interval 835, with an average step count of ~206.

Imputing missing values

We can find the total number of missing steps by summing up a call to is.na on steps.

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

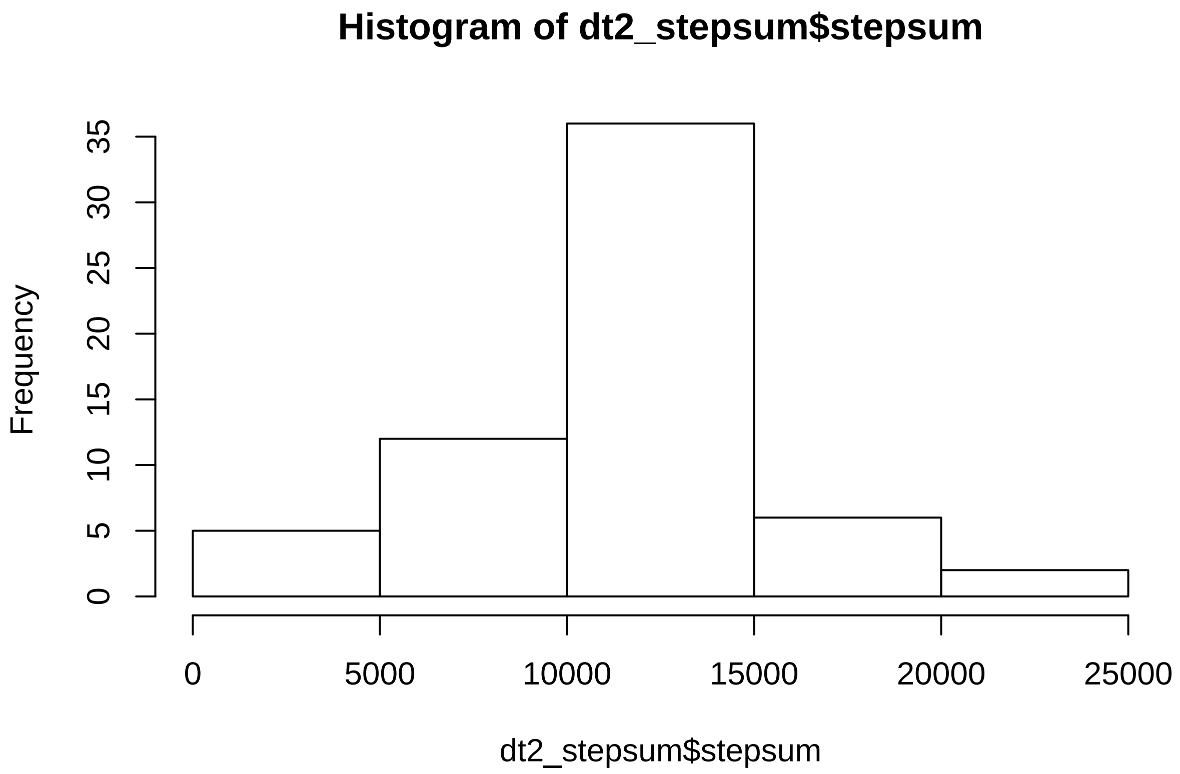
```
## [1] 2304
```

So, 2304 entries that need to be filled in. We will fill those in with the average step count for that interval, averaged over all days.

```
df_filled_in = read.csv('./activity.csv')
df_filled_in$date = as.POSIXct(df_filled_in$date)
for(interval in unique(df_filled_in[missingvalues,]$interval)) {
    dt_daily_avg_at_interval = dt_daily_avg$interval == interval
    df_filled_in[missingvalues & df_filled_in$interval == interval,]$steps <-
                      dt_daily_avg[dt_daily_avg_at_interval,]$avg
}
```

Let's make a new histogram after filling in the values:

```
dt2 <- data.table(df_filled_in)
dt2_stepsum = dt2[,list(stepsum=sum(steps)),by=date]
hist(dt2_stepsum$stepsum)
```



Now, let's look at the mean and median for this dataframe with imputed values.

```
mean(dt2_stepsum$stepsum)
```

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

```
median(dt2_stepsum$stepsum)
```

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

The mean does not differ from before, but the median does. It is now the same as the mean. The mean isn't changed because we used mean values to replace the missing values, but this did affect the median. Exactly how imputing values will change these estimates will depend on the method used.

Are there differences in activity patterns between weekdays and weekends?

First, we need to set up a weekday/weekend variable.

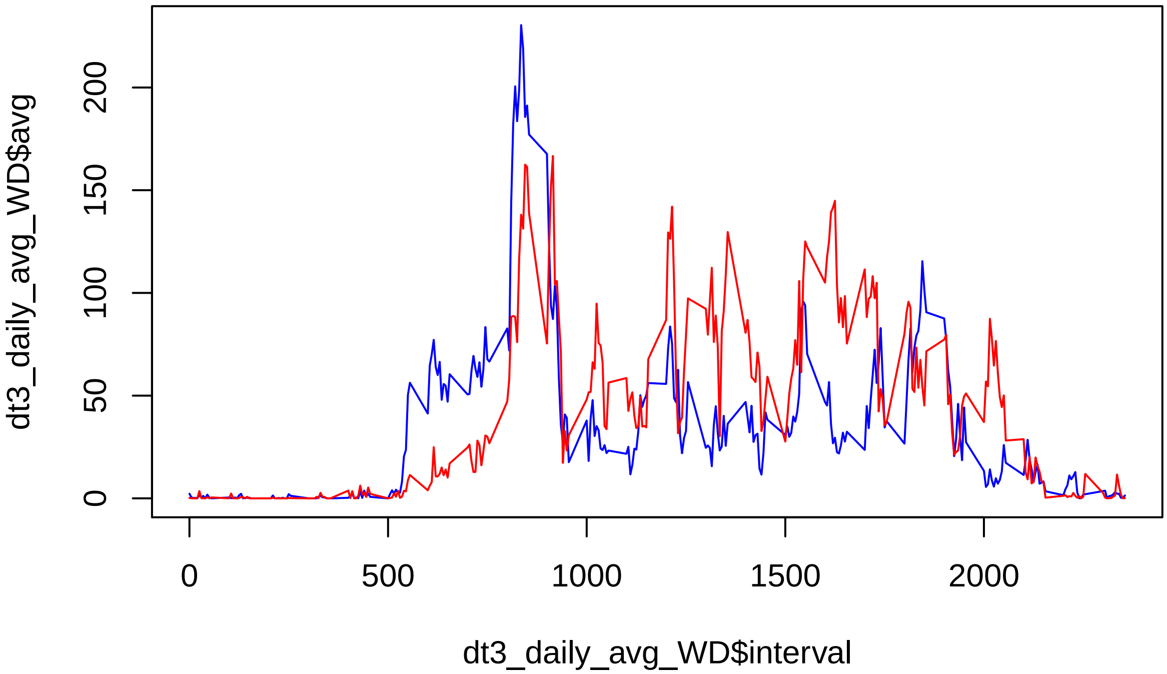
```
df_filled_in$day = 'weekday'
onweekend <- weekdays(df_filled_in$date) %in% c("Saturday", "Sunday")
df_filled_in[onweekend,]$day <- 'weekend'
df_filled_in$day = as.factor(df_filled_in$day)
```

Let's make tables with separate averages for the weekdays and weekends.

```
weekend <- df_filled_in$day == 'weekend'
dt3_WD <- data.table(df_filled_in[!weekend,])
dt3_WE <- data.table(df_filled_in[weekend,])
dt3_daily_avg_WD = dt3_WD[,list(avg=mean(steps)),by=list(interval)]
dt3_daily_avg_WE = dt3_WE[,list(avg=mean(steps)),by=list(interval)]
```

Now, we can plot daily activity levels averaged separately for weekdays and weekends.

```
plot(dt3_daily_avg_WD$interval, dt3_daily_avg_WD$avg, type='l', col='blue')
lines(dt3_daily_avg_WE$interval, dt3_daily_avg_WE$avg, type='l', col='red')
```



The weekend is plotted in red and the weekdays in blue. The plots are roguhly similar overall, with peaks in some of the same places. The weekdays have higher step counts in the morning, while the weekends have higher step counts in the middle of the day and at night, before bed.