Reproducible Research: Course Project 1

hcam

11/12/2016

Loading/Prcessing Data

Loading packages to process the dataset.

```
library(dplyr)
library(ggplot2)
library(chron)
```

```
setwd("/Users/hcam/Desktop/Data Sets/")
activity <- read.csv("activity.csv", stringsAsFactors = FALSE)
```

Mean Total Number of Steps Per Day

```
steps <- activity %>%
                group_by(date) %>%
                filter(!is.na(steps)) %>%
                summarise(total = sum(steps))
head(steps)
```

```
## Source: local data frame [6 x 2]
##
##          date total
##         (chr) (int)
## 1 2012-10-02   126
## 2 2012-10-03 11352
## 3 2012-10-04 12116
## 4 2012-10-05 13294
## 5 2012-10-06 15420
## 6 2012-10-07 11015
```

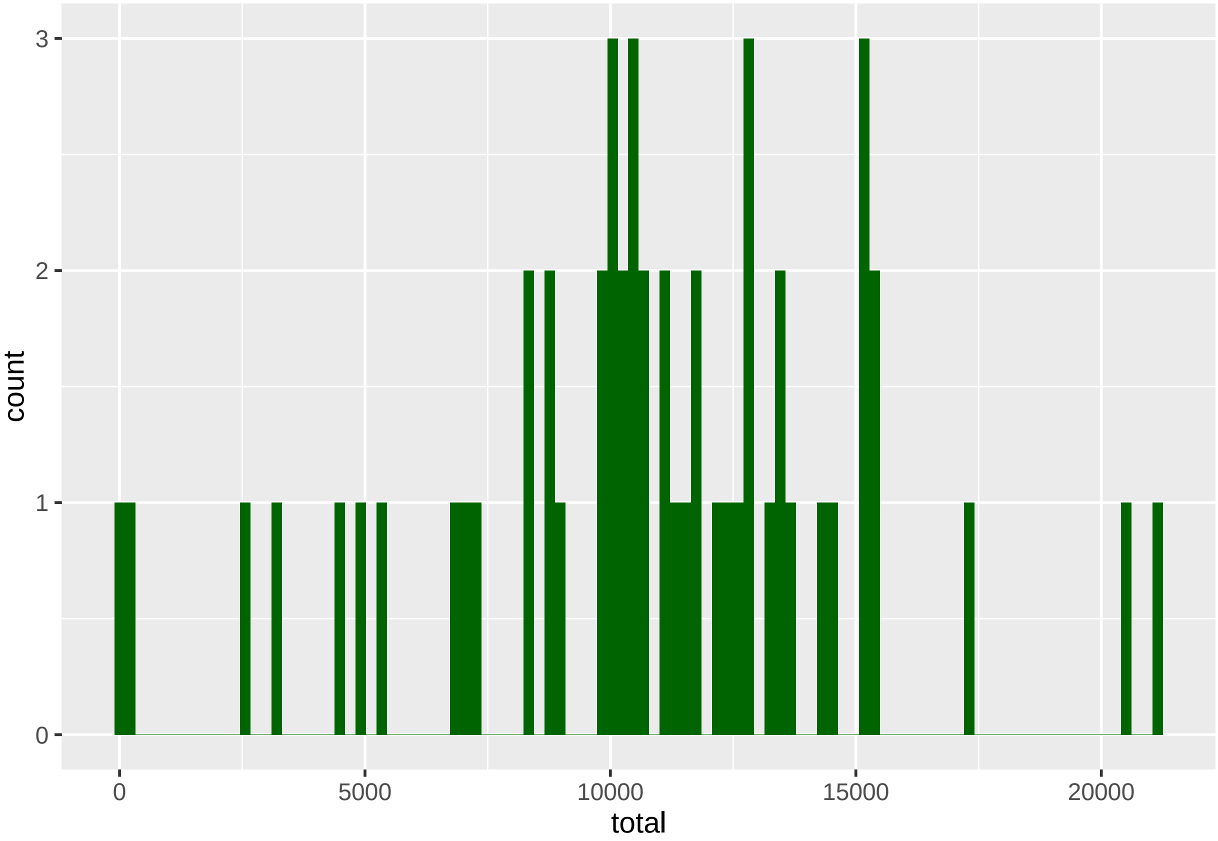
```
summary(steps)
```

```
##      date               total
##  Length:53          Min.   :   41
##  Class :character   1st Qu.: 8841
##  Mode  :character   Median :10765
##                     Mean   :10766
##                     3rd Qu.:13294
##                     Max.   :21194
```

We see the mean total number of steps per day is 10766 and median is 10765.

Setting up the histogram

```
steps_hist <- ggplot(steps, aes(x=total))
steps_hist + geom_histogram(fill = 'darkgreen', bins = 100)
```

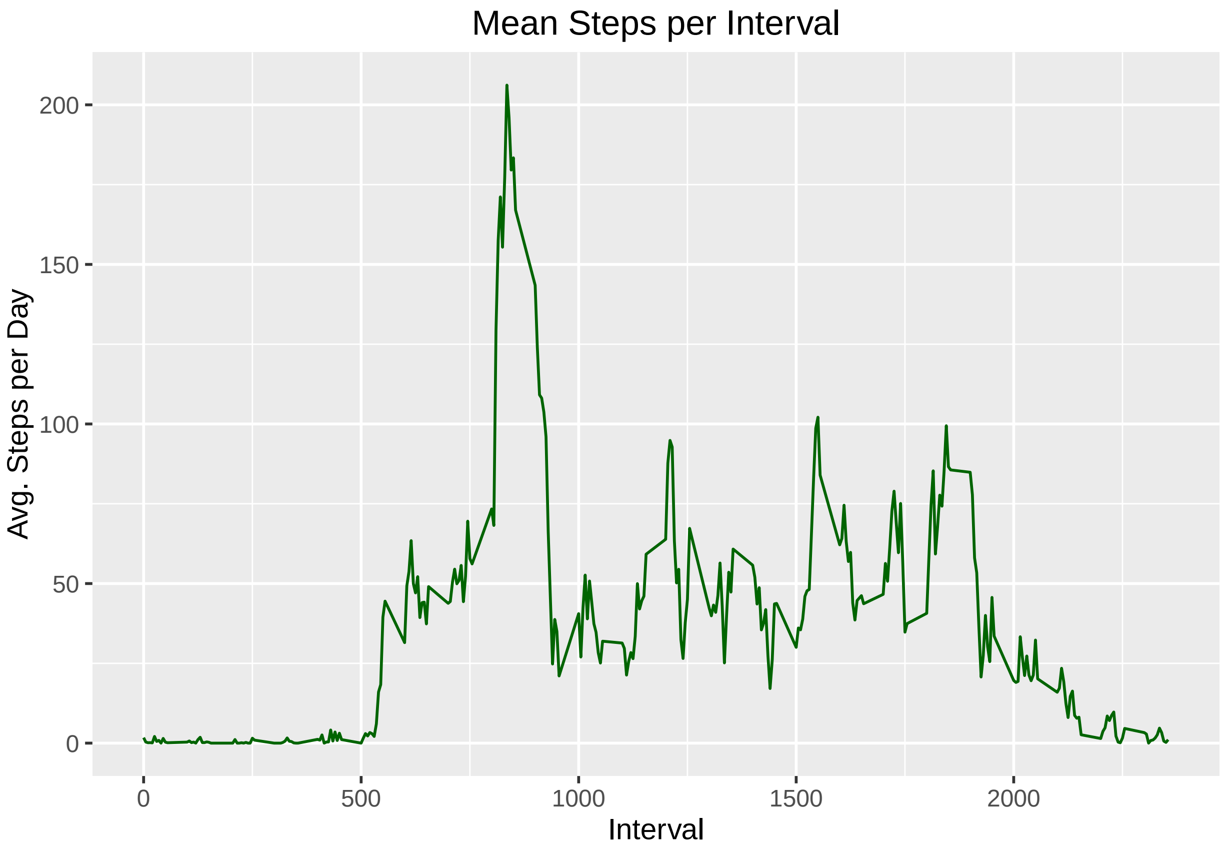


Average Daily Activity Pattern

We will remove the NA values from the dataset, group them by interval, and create a plot.

```
interval <- activity %>%
    group_by(interval) %>%
    filter(!is.na(steps)) %>%
    summarise(mean = mean(steps))

ggplot(interval, aes(x = interval, y = mean)) +
        geom_line(col = 'darkgreen') +
        labs(title = "Mean Steps per Interval", x = "Interval", y = "Avg. Steps per Day")
```



```
interval[which.max(interval$mean),]
```

```
## Source: local data frame [1 x 2]
##
##   interval     mean
##      (int)    (dbl)
## 1      835 206.1698
```

We see that interval 835 has the max average number of steps.

Imputing Missing Values

```
summary(activity)
```

```
##      steps            date              interval
##  Min.   :  0.00   Length:17568       Min.   :   0.0
##  1st Qu.:  0.00   Class :character   1st Qu.: 588.8
##  Median :  0.00   Mode  :character   Median :1177.5
##  Mean   : 37.38                      Mean   :1177.5
##  3rd Qu.: 12.00                      3rd Qu.:1766.2
##  Max.   :806.00                      Max.   :2355.0
##  NA's   :2304
```

We see that there are 2304 NA values. I chose replace the NA values with the mean of the remaining values.

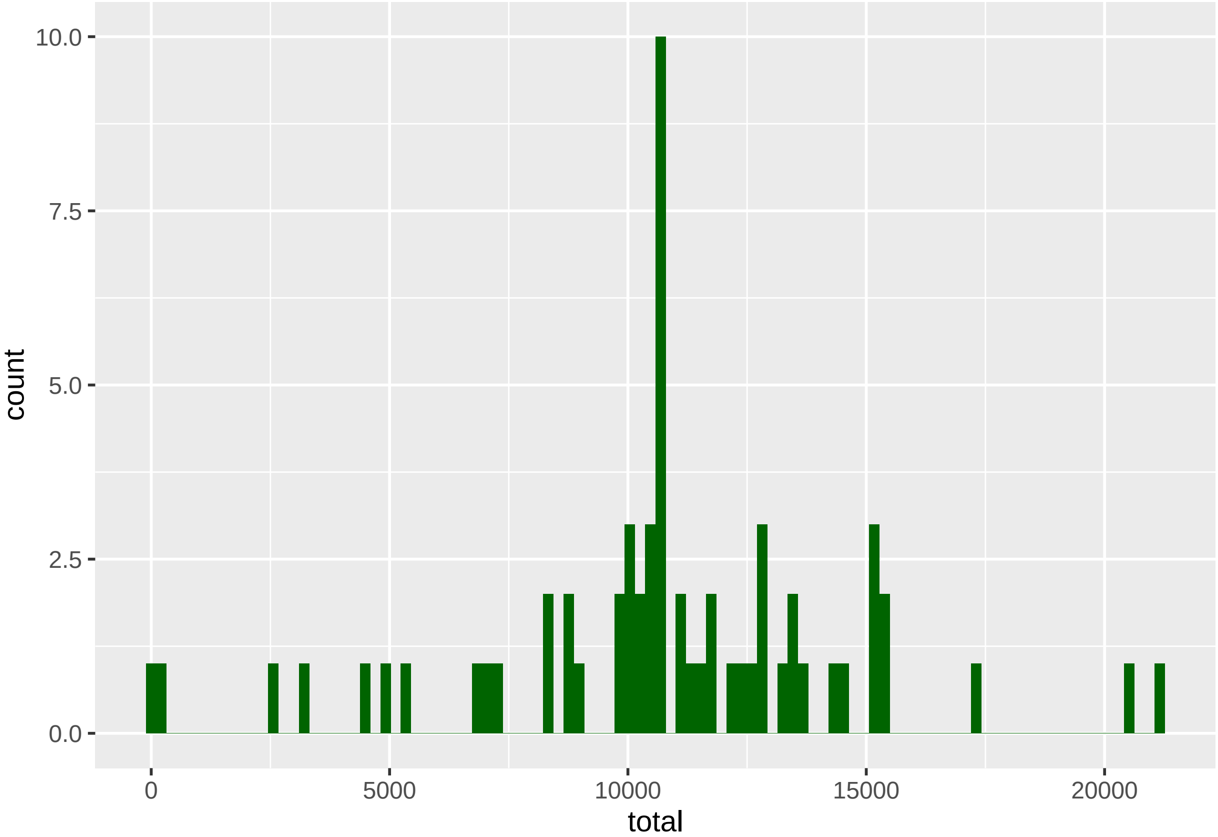
```
mean <- mean(activity$steps, na.rm = TRUE)

activityNA <- activity
activityNA$steps[which(is.na(activityNA$steps))] <- mean
summary(activityNA)
```

```
##      steps            date              interval     
##  Min.   :  0.00   Length:17568       Min.   :   0.0  
##  1st Qu.:  0.00   Class :character   1st Qu.: 588.8  
##  Median :  0.00   Mode  :character   Median :1177.5  
##  Mean   : 37.38                      Mean   :1177.5  
##  3rd Qu.: 37.38                      3rd Qu.:1766.2  
##  Max.   :806.00                      Max.   :2355.0  
```

```
stepsNA <- activityNA %>%
    group_by(date) %>%
    summarise(total = sum(steps))

stepsNA_hist <- ggplot(stepsNA, aes(x=total))
stepsNA_hist + geom_histogram(fill = 'darkgreen', bins = 100)
```



```
summary(stepsNA)
```

```
##      date               total      
##  Length:61          Min.   :   41  
##  Class :character   1st Qu.: 9819
##  Mode  :character   Median :10766
##                     Mean   :10766
##                     3rd Qu.:12811
##                     Max.   :21194
```

Weekdays vs Weekends

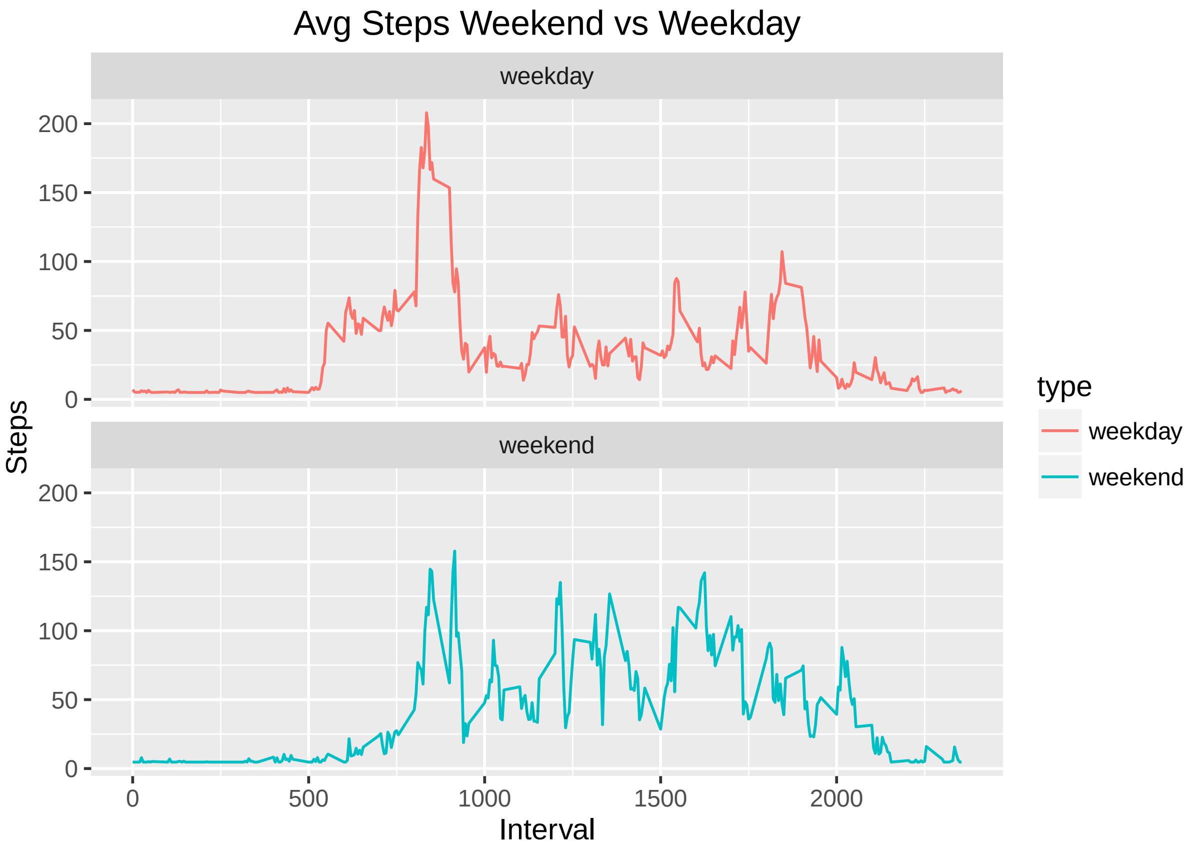
We will make use of the chron package and the is.weekend function to differentiate weekday or weekend.

```
type <- is.weekend(activityNA$date)
activityNA$type <- ifelse (type == "TRUE", "weekend", "weekday")

activityNA_type<- activityNA %>%
   group_by(interval, type) %>%
   summarise(mean = mean(steps))

wkdy_plot <- ggplot(activityNA_type, aes(x =interval , y=mean, color=type)) +
    geom_line() +
    labs(title = "Avg Steps Weekend vs Weekday", x = "Interval", y = "Steps") +
    facet_wrap(~type, ncol = 1, nrow=2)

wkdy_plot
```



We can see from the two plots that there are differences in activity patterns for weekdays and weekends. The pattern for weekdays seem to peak earlier in the day and on weekends the steps remain constant throughout the day.