Reproducible Research - Course Project 1

Andrew Golus

March 30, 2017

Load the dataset

```
knitr::opts_chunk$set(echo = TRUE)
setwd("C:/Users/ag827/Desktop/R")
activity <- read.csv("activity.csv")
```

Create histogram of the total number of steps taken each day

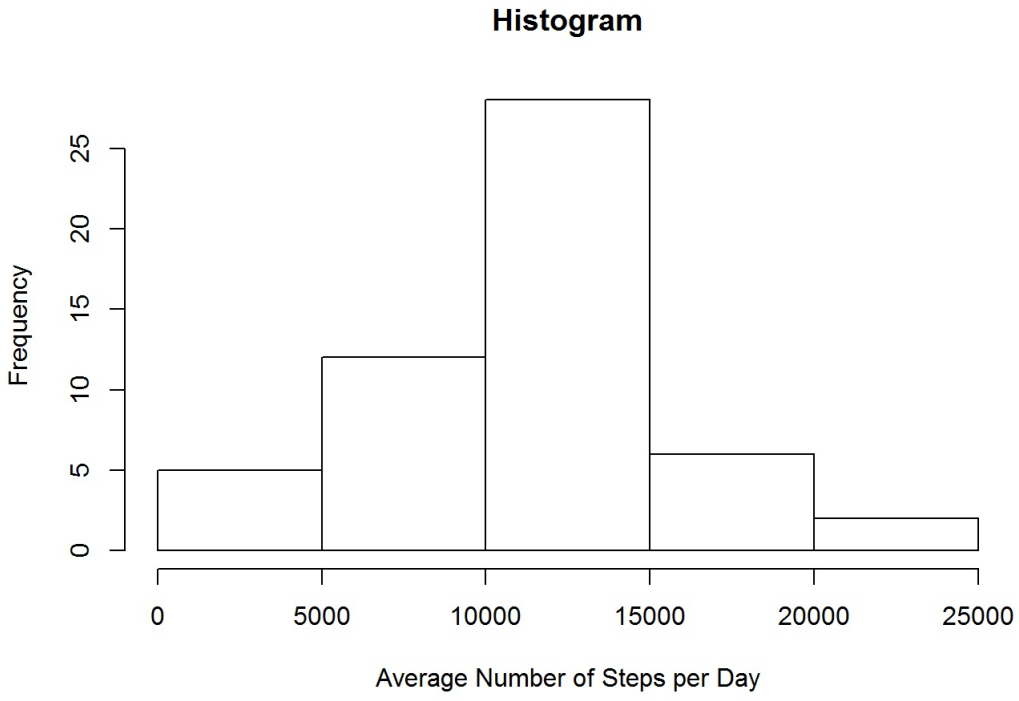
```
library(dplyr)
```

```
##
## 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
```

```
steps.by.day <- activity %>%
        filter(is.na(steps)==FALSE) %>%
        group_by(date) %>%
        summarize(mean.steps = sum(steps))
hist(steps.by.day$mean.steps, main = "Histogram", xlab = "Average Number of Steps per Day")
```



Calculate mean number of steps taken each day

```
mean(steps.by.day$mean.steps)
```

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

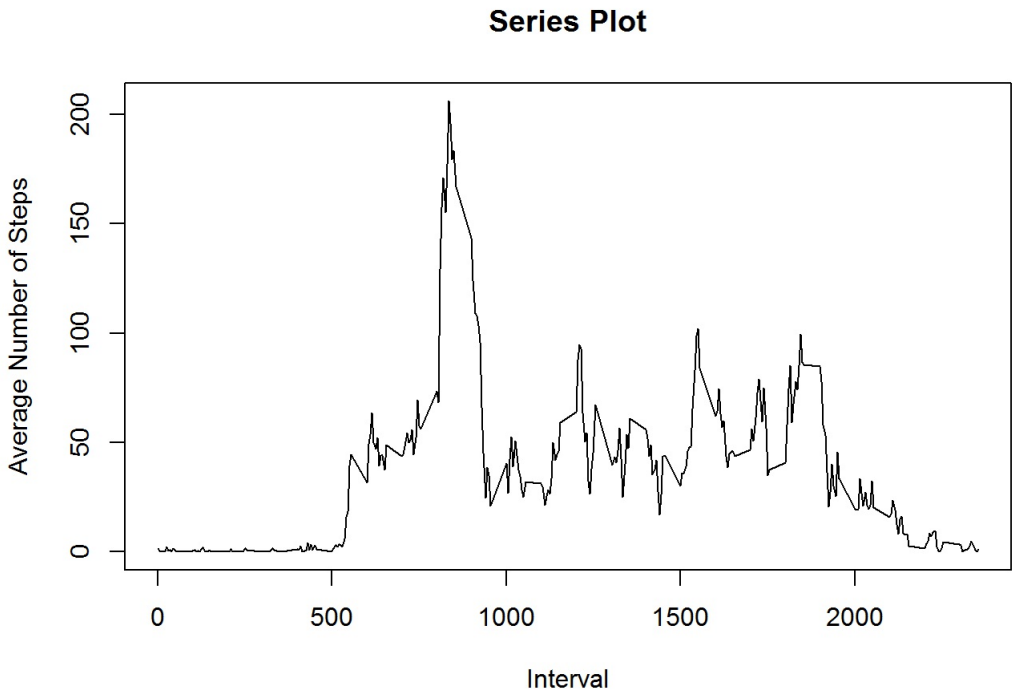
Calculate median number of steps taken each day

```
quantile(steps.by.day$mean.steps, probs = 0.5)
```

```
##   50%
## 10765
```

Create time series plot of the average number of steps taken

```
steps.by.interval <- activity %>%
        filter(is.na(steps)==FALSE) %>%
        group_by(interval) %>%
        summarize(mean.steps = mean(steps))
plot(steps.by.interval$interval, steps.by.interval$mean.steps, type = "l", xlab = "Interval", ylab = "Average N
umber of Steps", main = "Series Plot")
```



Determine the 5-minute interval that, on average, contains the maximum number of steps

```
subset(steps.by.interval, steps.by.interval$mean.steps == max(steps.by.interval$mean.steps))[1,1]
```

```
## # A tibble: 1 × 1
##   interval
##      <int>
## 1      835
```

Calculate the total number of missing values in the dataset

```
sum(is.na(activity$steps) == TRUE)
```

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

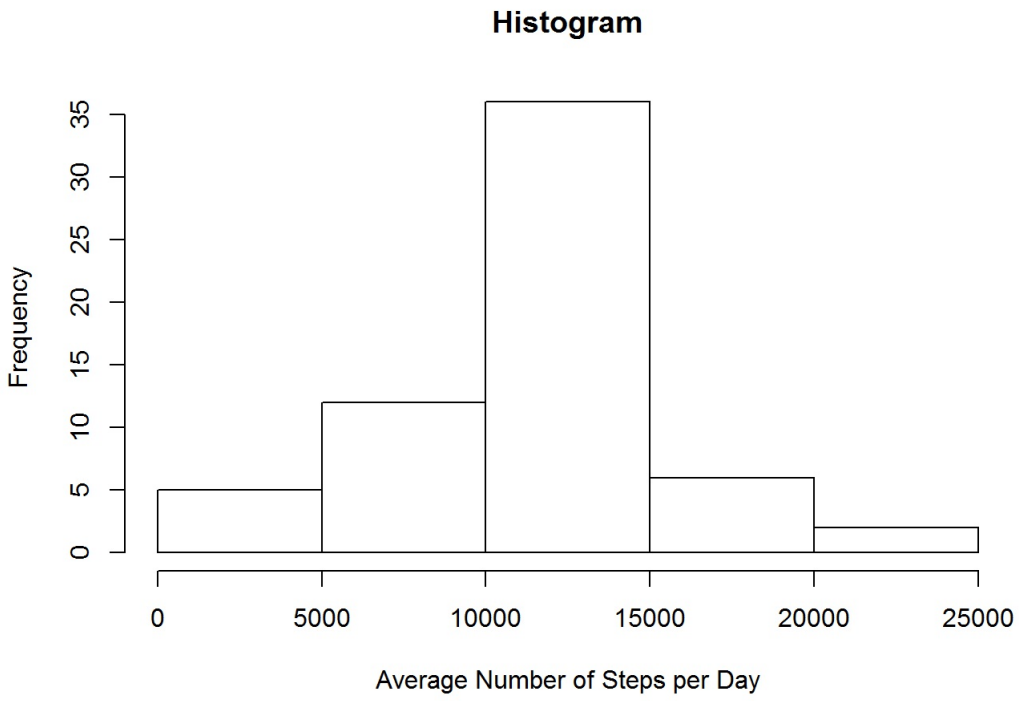
Create a new dataset with imputed missing data

My strategy is to replace the missing step values with the average number of steps for the interval.

```
x0 <- subset(activity, is.na(activity$steps) == FALSE)
x1 <- subset(activity, is.na(activity$steps) == TRUE)
x1 <- merge(x1, steps.by.interval, all.x = TRUE)
x1 <- x1[,-2]
names(x1) <- c("interval", "date", "steps")
activity.fix <- rbind.data.frame(x0, x1)
steps.by.day <- activity.fix %>%
        filter(is.na(steps)==FALSE) %>%
        group_by(date) %>%
        summarize(mean.steps = sum(steps))
```

Create histogram of the total number of steps taken each day after missing values are imputed

```
hist(steps.by.day$mean.steps, main = "Histogram", xlab = "Average Number of Steps per Day")
```



Calculate mean number of steps taken each day after missing values are imputed

```
mean(steps.by.day$mean.steps)
```

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

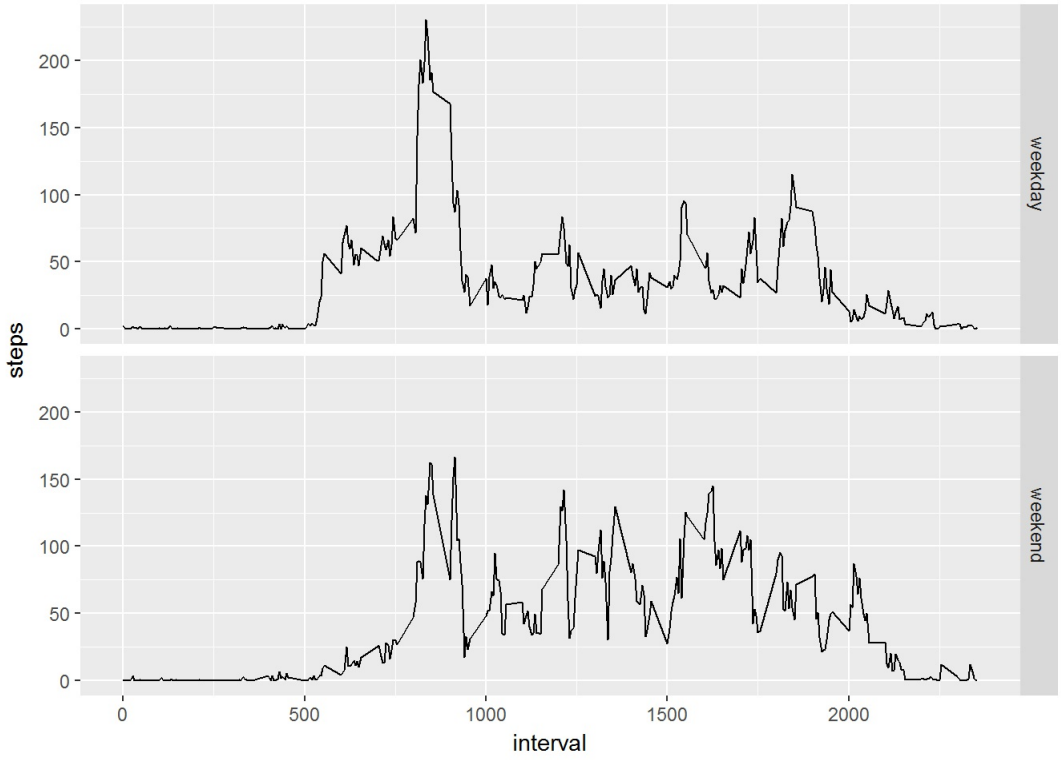
Calculate median number of steps taken each day after missing values are imputed

```
quantile(steps.by.day$mean.steps, probs = 0.5)
```

```
##      50%
## 10766.19
```

Compare the average number of steps taken per 5-minute interval across weekdays and weekends

```
activity.weekday <- activity.fix %>%
mutate(weekday = weekdays(as.Date(date))) %>%
mutate(isweekend = ifelse(weekday == "Saturday" | weekday == "Sunday","weekend","weekday")) %>%
group_by(interval, isweekend) %>%
summarize(steps = mean(steps))
library(ggplot2)
g <- ggplot(activity.weekday, aes(interval, steps))
g + geom_line() + facet_grid(isweekend ~ .)
```


steps
200
150
100
50
0
weekday
weekend
0
500
1000
1500
2000
interval