# Reproducible Research Course Project 1

*Kyle Ward*

*July 9, 2017*

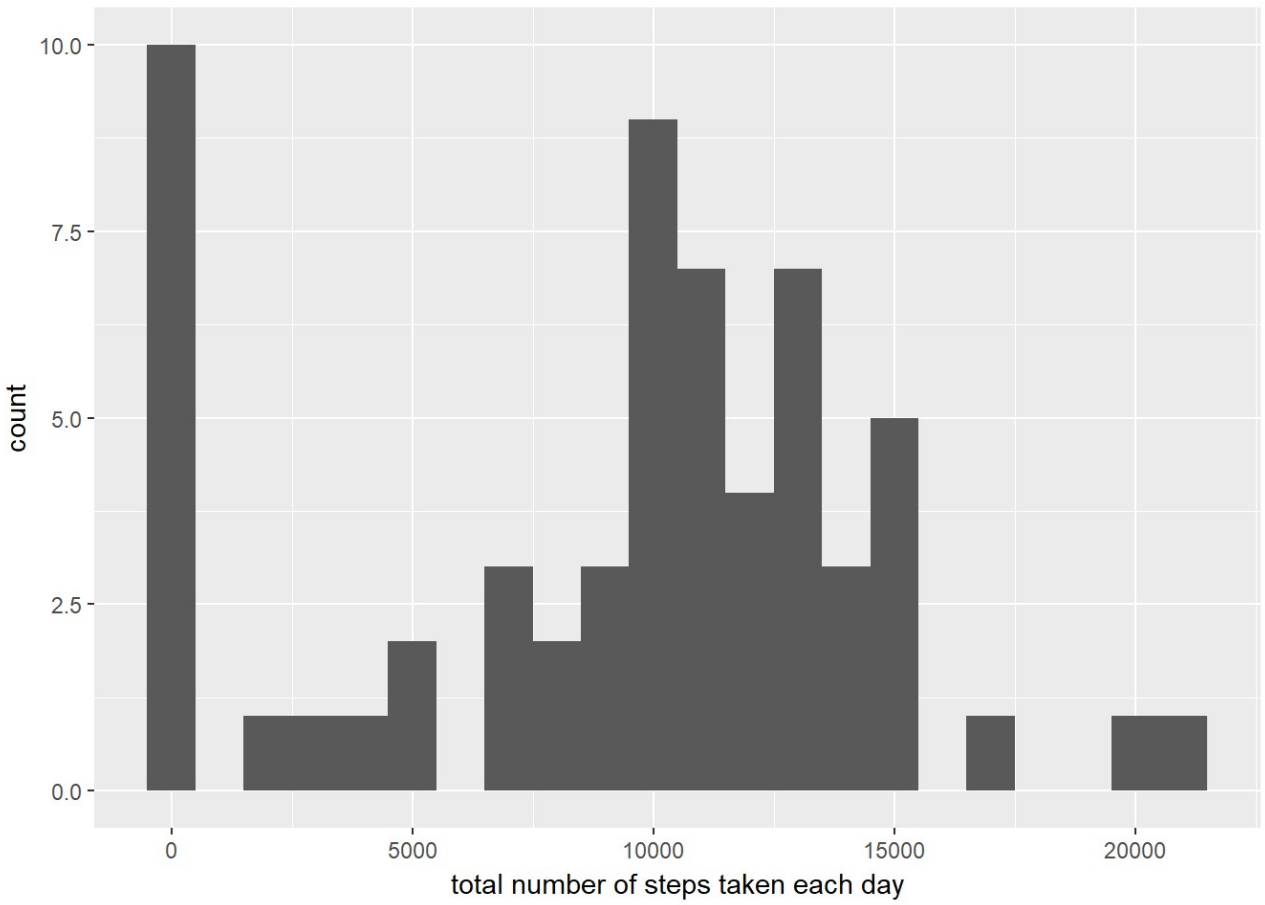
## R Markdown

This an R markdown file generated for completion of the Coursera Reproducible Research Course Project 1. The following computations and visualizations are performed using the ggplot2, dplyr, and scales packages. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

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

```
setwd("~/Documents/Documents/Important Files/Hopkins/Data Science/Reproducable Researc
h/Week 2/projectdata")
activity <- read.csv("activity.csv")
activity$date <- as.Date(activity$date, "%Y-%m-%d")
activity <- as.data.frame(activity)
```

```
library(ggplot2)
total.steps <- tapply(activity$steps, activity$date, FUN=sum, na.rm=TRUE)
qplot(total.steps, binwidth=1000, xlab="total number of steps taken each day")
```




```
mean(total.steps, na.rm=TRUE)
```

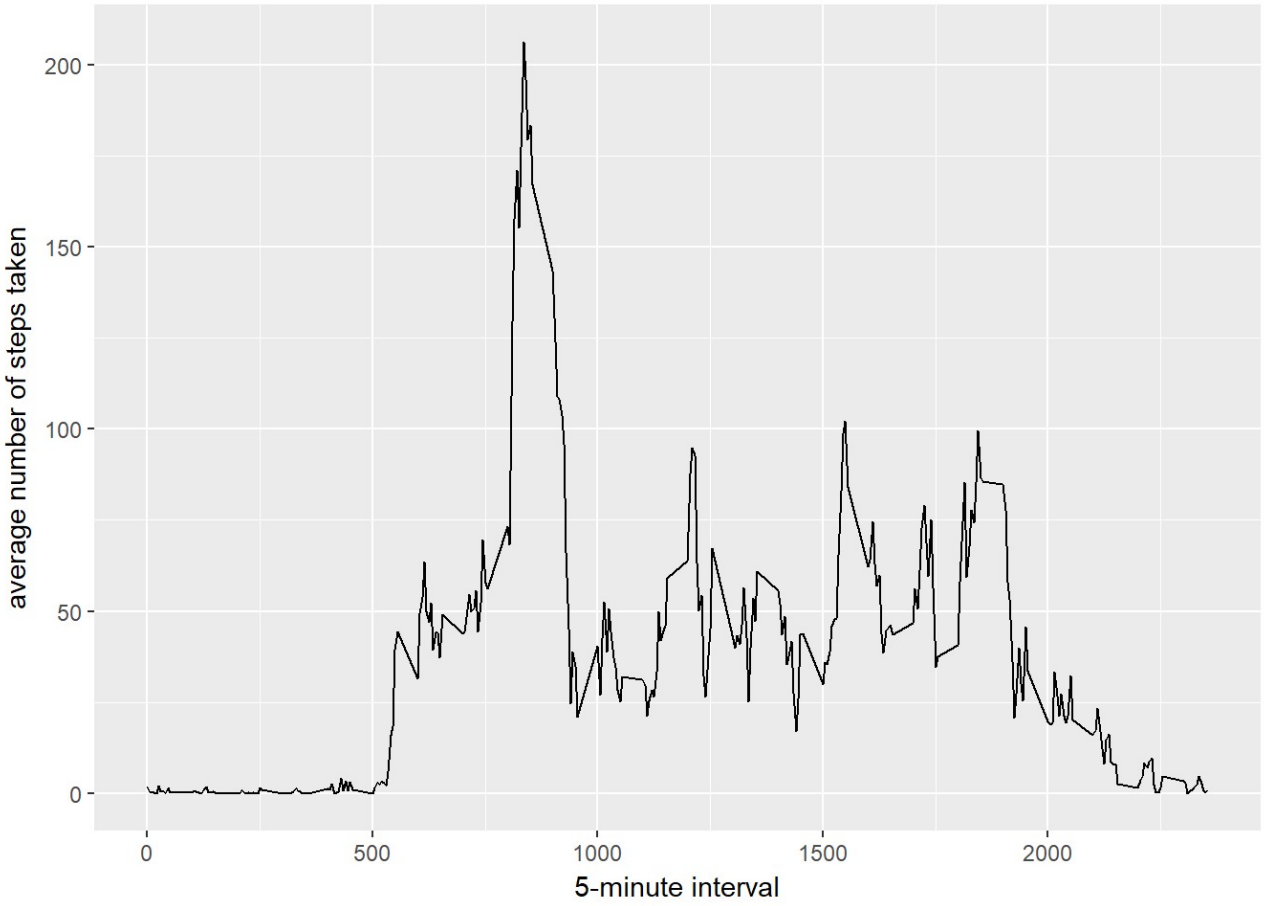
```
## [1] 9354.23
```

```
median(total.steps, na.rm=TRUE)
```

```
## [1] 10395
```

# What is the average daily activity pattern?

```
library(ggplot2)
averages <- aggregate(x=list(steps=activity$steps), by=list(interval=activity$interva
l),
                      FUN=mean, na.rm=TRUE)
ggplot(data=averages, aes(x=interval, y=steps)) +
        geom_line() +
        xlab("5-minute interval") +
        ylab("average number of steps taken")
```




```
averages[which.max(averages$steps),]
```

```
##     interval    steps
## 104      835 206.1698
```

```
library(scales)
sum(is.na(activity))
```

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

```
percent(sum(is.na(activity))/nrow(activity))
```

```
## [1] "13.1%"
```

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

```
averages[which.max(averages$steps),]
```

```
##     interval    steps
## 104      835 206.1698
```

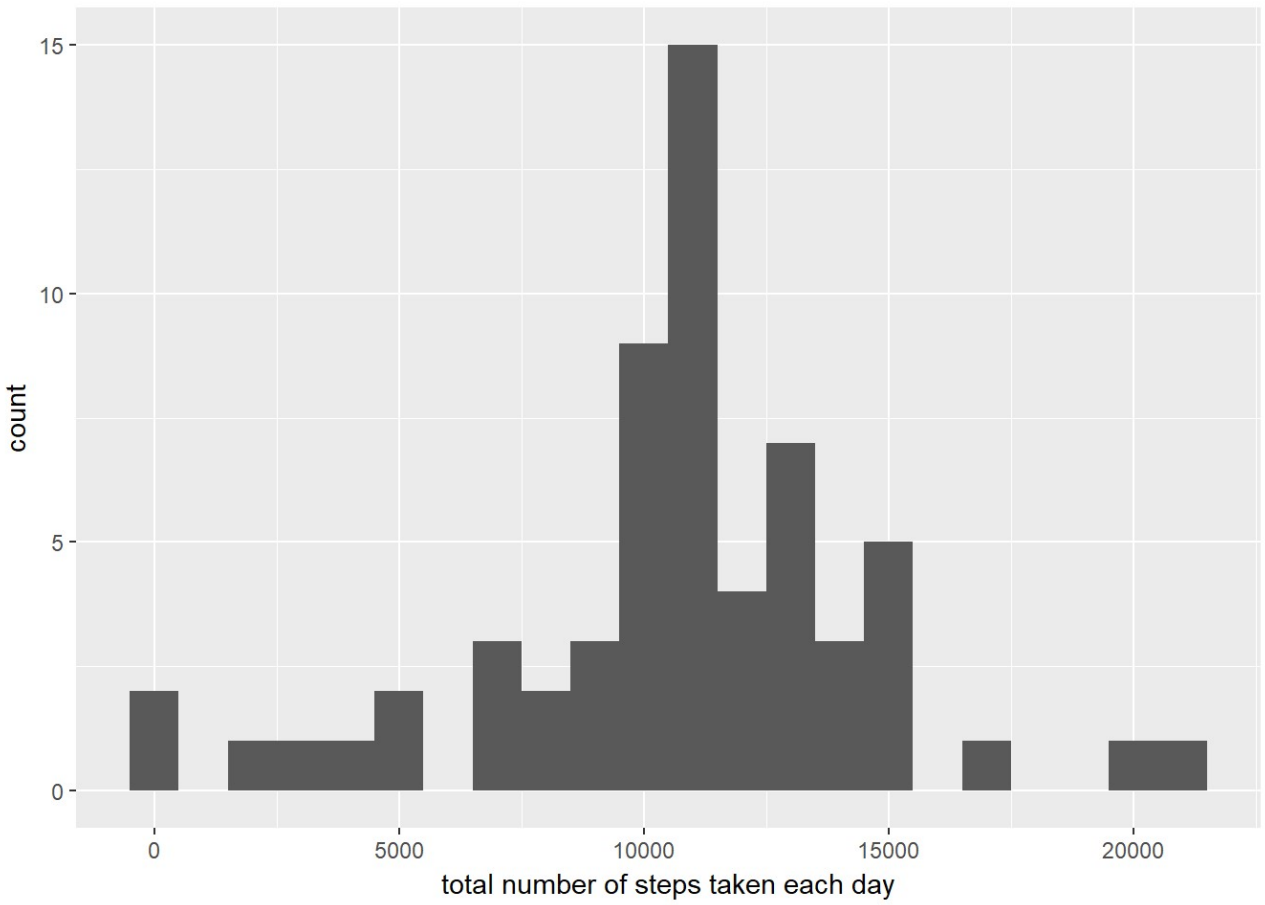
# Imputing Missing Values

```
missing <- is.na(activity$steps)
table(missing)
```

```
## missing
## FALSE  TRUE
## 15264  2304
```

```
fill.value <- function(steps, interval) {
        filled <- NA
        if (!is.na(steps))
                filled <- c(steps)
        else
                filled <- (averages[averages$interval==interval, "steps"])
        return(filled)
}
filled.activity <- activity
filled.activity$steps <- mapply(fill.value, filled.activity$steps, filled.activity$int
erval)

total.steps <- tapply(filled.activity$steps, filled.activity$date, FUN=sum)
qplot(total.steps, binwidth=1000, xlab="total number of steps taken each day")
```




```
mean(total.steps)
```

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

```
median(total.steps)
```

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

# Are there differences in activity patterns between weekdays and weekends?

```
weekday.or.weekend <- function(date) {
        day <- weekdays(date)
        if (day %in% c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday"))
                return("weekday")
        else if (day %in% c("Saturday", "Sunday"))
                return("weekend")
        else
                stop("invalid date")
}
filled.activity$date <- as.Date(filled.activity$date)
filled.activity$day <- sapply(filled.activity$date, FUN=weekday.or.weekend)

averages <- aggregate(steps ~ interval + day, data=filled.activity, mean)
ggplot(averages, aes(interval, steps)) + geom_line() + facet_grid(day ~ .) +
        xlab("5-minute interval") + ylab("Number of steps")
```