Reproducible Research: Course Project 1

A Suarez-Pierre

June 16, 2016

1. Load necessary packages and set working directory

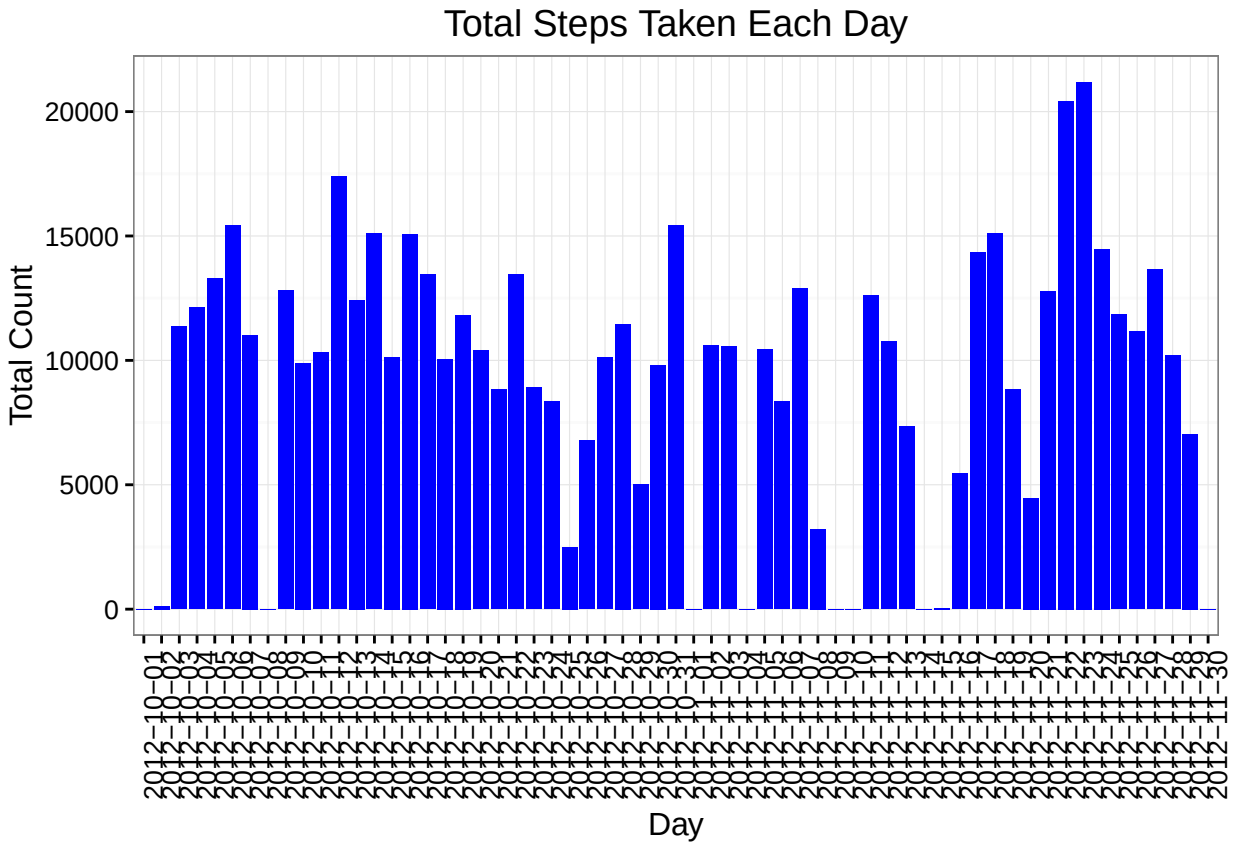
```
library(ggplot2)
library(dplyr)
library(lubridate)
library(gridExtra)


setwd("/Users/asuarez/Desktop/ReproducibleResearch")
```

2. Read the dataset and process the data

```
df = read.csv("activity.csv", stringsAsFactors=FALSE)
```

3. Histogram of the total number of steps taken each day

```
df = group_by(df, date)
statsByDay = summarise(df, steps=sum(steps, na.rm=TRUE))

ggplot(data=statsByDay, aes(x=date, y=steps)) + geom_bar(stat="identity", fill="blue") +
      theme_bw() + theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
      labs(title="Total Steps Taken Each Day", x="Day", y="Total Count")
```



4. Mean and median number of steps taken each day

Pardon the long list

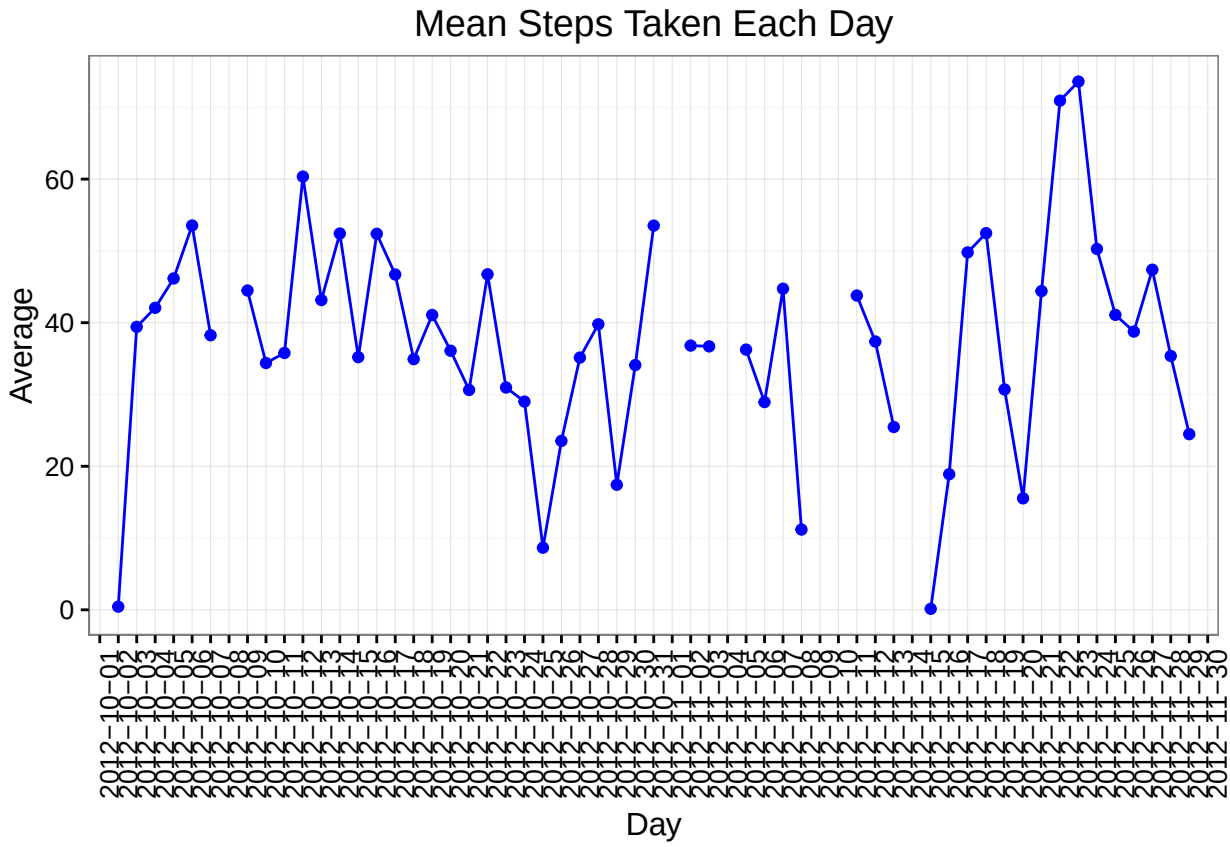
```
statsByDay = summarise(df,
                       mean=round(mean(steps, na.rm=TRUE),2),
                       median=median(steps, na.rm=TRUE))

print(statsByDay, n=62)
```

```
## Source: local data frame [61 x 3]
##
##          date  mean median
##         (chr) (dbl)  (dbl)
## 1  2012-10-01    NA     NA
## 2  2012-10-02  0.44      0
## 3  2012-10-03 39.42      0
## 4  2012-10-04 42.07      0
## 5  2012-10-05 46.16      0
## 6  2012-10-06 53.54      0
## 7  2012-10-07 38.25      0
## 8  2012-10-08   NaN     NA
## 9  2012-10-09 44.48      0
## 10 2012-10-10 34.38      0
## 11 2012-10-11 35.78      0
```

```
## 12 2012-10-12 60.35      0
## 13 2012-10-13 43.15      0
## 14 2012-10-14 52.42      0
## 15 2012-10-15 35.20      0
## 16 2012-10-16 52.38      0
## 17 2012-10-17 46.71      0
## 18 2012-10-18 34.92      0
## 19 2012-10-19 41.07      0
## 20 2012-10-20 36.09      0
## 21 2012-10-21 30.63      0
## 22 2012-10-22 46.74      0
## 23 2012-10-23 30.97      0
## 24 2012-10-24 29.01      0
## 25 2012-10-25  8.65      0
## 26 2012-10-26 23.53      0
## 27 2012-10-27 35.14      0
## 28 2012-10-28 39.78      0
## 29 2012-10-29 17.42      0
## 30 2012-10-30 34.09      0
## 31 2012-10-31 53.52      0
## 32 2012-11-01   NaN     NA
## 33 2012-11-02 36.81      0
## 34 2012-11-03 36.70      0
## 35 2012-11-04   NaN     NA
## 36 2012-11-05 36.25      0
## 37 2012-11-06 28.94      0
## 38 2012-11-07 44.73      0
## 39 2012-11-08 11.18      0
## 40 2012-11-09   NaN     NA
## 41 2012-11-10   NaN     NA
## 42 2012-11-11 43.78      0
## 43 2012-11-12 37.38      0
## 44 2012-11-13 25.47      0
## 45 2012-11-14   NaN     NA
## 46 2012-11-15  0.14      0
## 47 2012-11-16 18.89      0
## 48 2012-11-17 49.79      0
## 49 2012-11-18 52.47      0
## 50 2012-11-19 30.70      0
## 51 2012-11-20 15.53      0
## 52 2012-11-21 44.40      0
## 53 2012-11-22 70.93      0
## 54 2012-11-23 73.59      0
## 55 2012-11-24 50.27      0
## 56 2012-11-25 41.09      0
## 57 2012-11-26 38.76      0
## 58 2012-11-27 47.38      0
## 59 2012-11-28 35.36      0
## 60 2012-11-29 24.47      0
## 61 2012-11-30   NaN     NA
```

5. Times series plot of the average number of steps taken

```
ggplot(data=statsByDay, aes(x=date, y=mean, group=1)) + geom_point(col="blue") +
      geom_line(col="blue") + theme_bw() +
      theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
      labs(title="Mean Steps Taken Each Day", x="Day", y="Average")
```



6. The 5-min interval that, on average, contains the maximum number of steps

```
df = group_by(df, interval)

statsByInt = summarise(df, mean=round(mean(steps, na.rm=TRUE), 2))
statsByInt = arrange(statsByInt, desc(mean))

max = (statsByInt$interval[1])
print(max)
```

```
## [1] 835
```

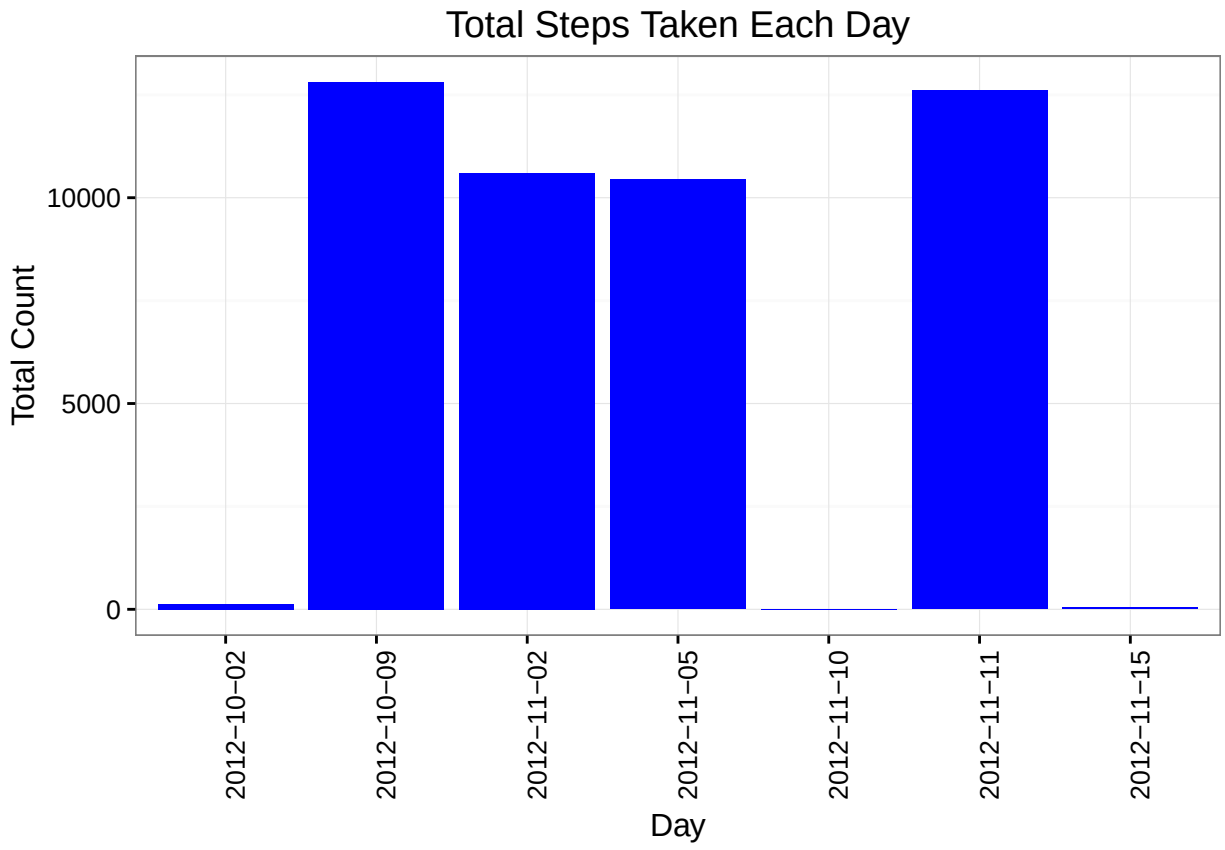
The interval with the max number of steps on average is 835.

7. Histogram of the number of steps taken each day after missing values are imputed

```
df = group_by(df, date)
statsByDay = summarise(df, steps=sum(steps, na.rm=TRUE))

missing = statsByDay$steps==0
missing = c(FALSE, missing) # array containing days after missing data

dayAfter = statsByDay[missing,]
dayAfter = dayAfter[1:7,] #removes empty row at the bottom
```

Now we print the histogram

```
ggplot(data=dayAfter, aes(x=date, y=steps)) + geom_bar(stat="identity", fill="blue") +
      theme_bw() + theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
      labs(title="Total Steps Taken Each Day", x="Day", y="Total Count")
```



8. Panel plot comparing the average number of steps taken per 5-min interval across weekdays and weekends

```
df$date = ymd(df$date)

# Id which are weekdays
```

```
df = mutate(df, weekday=ifelse(wday(date)==1 | wday(date)==7, 0, 1))


# Averages by interval
df = group_by(df, weekday, interval)
statsByInt = summarise(df, mean=round(mean(steps, na.rm=TRUE), 2))


# Separating into 2 dataframes depending on weekday Vs weekend
wkdays = filter(statsByInt, weekday==1)[,2:3]
wkends = filter(statsByInt, weekday==0)[,2:3]


# Generating plots
plot1 = ggplot(data=wkdays, aes(x=interval, y=mean)) + geom_line(col="blue") + theme_bw() +
        labs(title="Weekdays: Avg Steps By Time of Day", x="Time", y="Average")
plot2 = ggplot(data=wkends, aes(x=interval, y=mean)) + geom_line(col="blue") + theme_bw() +
        labs(title="Weekends: Avg Steps By Time of Day", x="Time", y="Average")

grid.arrange(plot1, plot2, nrow=2)
```