Project1

Utkarsha Patil

February 13, 2019

Introduction

This assignment makes use of data from a personal activity monitoring device. This device collects data at 5 minute intervals through out the day. The data consists of two months of data from an anonymous individual collected during the months of October and November, 2012 and include the number of steps taken in 5 minute intervals each day.

Data

The data for this assignment can be downloaded from the course web site:

Dataset: Activity monitoring data [52K] The variables included in this dataset are:

steps: Number of steps taking in a 5-minute interval (missing values are coded as NA) date: The date on which the measurement was taken in YYYY-MM-DD format interval: Identifier for the 5-minute interval in which measurement was taken

The dataset is stored in a comma-separated-value (CSV) file and there are a total of 17,568 observations in this dataset.

Loading and preprocessing the data

I have already downloaded and unzipped activity.csv in my local directory.

```r
data <- read.csv("activity.csv")
```

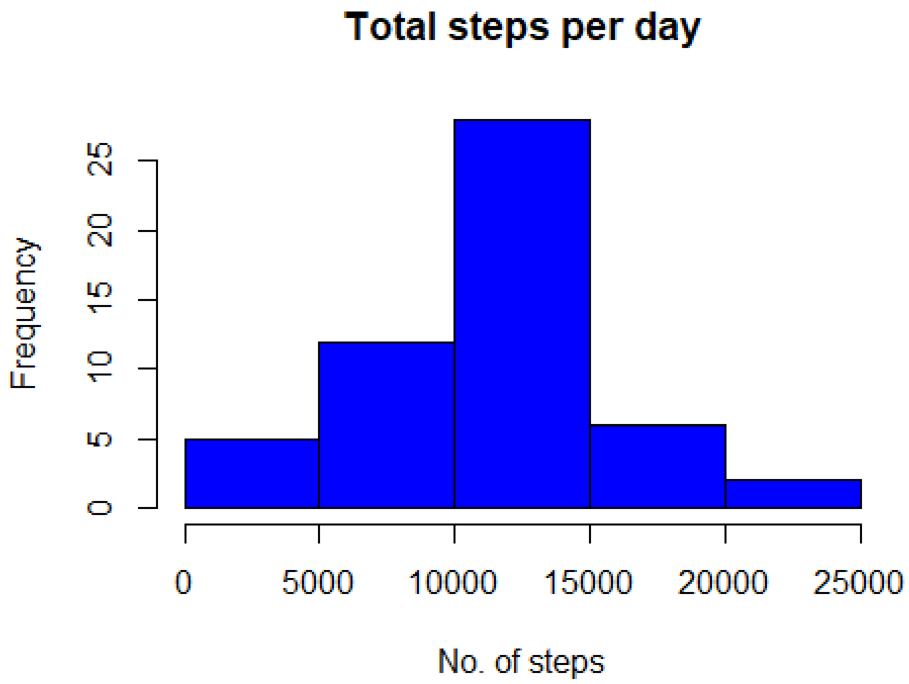
What is mean total number of steps taken per day?

1.Calculate the total number of steps taken per day

```r
#aggregating no. of steps per day
daily_steps <- aggregate(steps ~ date, data, sum)
```

2.If you do not understand the difference between a histogram and a bar plot, research the difference between them. Make a histogram of the total number of steps taken each day

```r
hist(daily_steps$steps, main = paste("Total steps per day"),
col="blue", xlab="No. of steps")
```



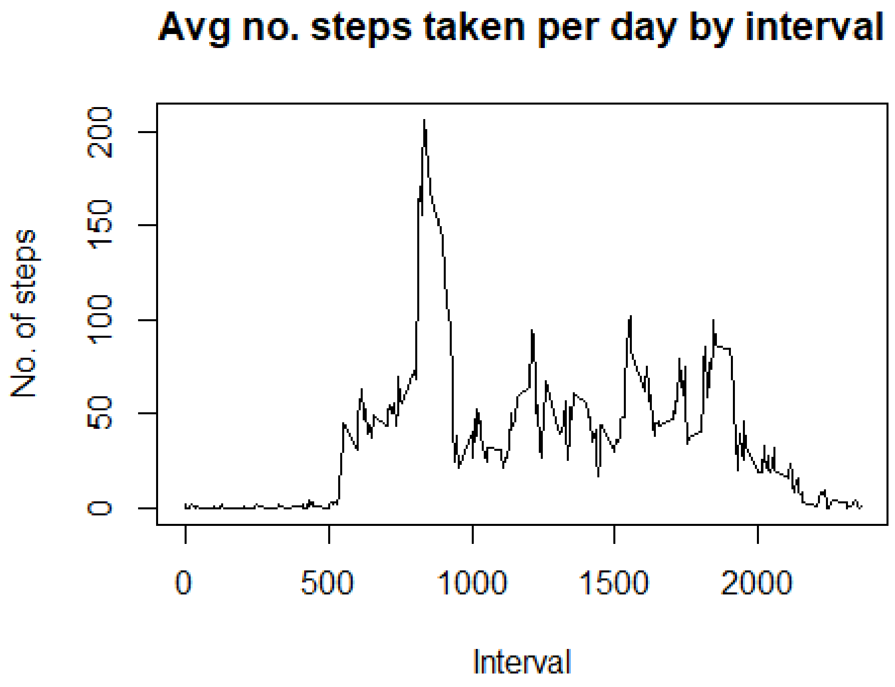
3.Calculate and report the mean and median of the total number of steps taken per day

```
step_mean <- mean(daily_steps$steps)
paste("Mean of total steps taken each day:", step_mean)

## [1] "Mean of total steps taken each day: 10766.1886792453"

step_mean

## [1] 10766.19

step_median <- median(daily_steps$steps)
paste("Median of total steps taken each day:", step_median)

## [1] "Median of total steps taken each day: 10765"
```

What is the average daily activity pattern?

1.Make a time series plot (i.e. type="l") of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis)

```
interval_steps <- aggregate(steps ~ interval, data, mean)
plot(interval_steps$interval, interval_steps$steps, type="l",
xlab="Interval", ylab="No. of steps", main="Avg no. steps taken
per day by interval")
```



2.Which 5-minute interval, on average across all the days in the dataset, contains the maximum number of steps?

```
max_interval <- interval_steps[which.max(interval_steps$steps),1]
paste("At 5-minute interval, on average across all the days in
the dataset, contains the maximum number of steps:",
max_interval)

## [1] "At 5-minute interval, on average across all the days in
the dataset, contains the maximum number of steps: 835"
```

Imputing missing values

1.Calculate and report the total number of missing values in the dataset (i.e. the total number of rows with NAs)

```
NA_data <- sum(!complete.cases(data))
NA_data

## [1] 2304
```

2.Devise a strategy for filling in all of the missing values in the dataset. The strategy does not need to be sophisticated. For example, you could use the mean/median for that day, or the mean for that 5-minute interval, etc.

```
#Aggregating mean of original data for 5 minute interval
mean_i <- aggregate(steps ~ interval, data, mean)
```

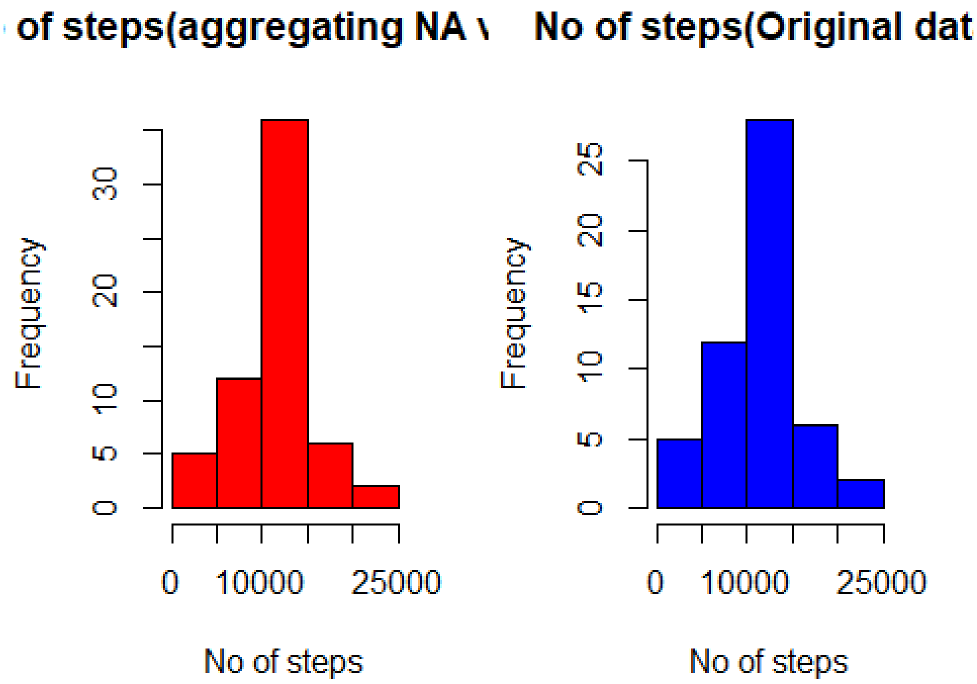
```
#Merging the mean of total steps for that day with original data
new_data <- merge(x=data, y=mean_i, by="interval")
```

3.Create a new dataset that is equal to the original dataset but with the missing data filled in.

```
new_data$steps <- ifelse(is.na(new_data$steps.x),
new_data$steps.y, new_data$steps.x)
head(new_data)

##   interval steps.x       date  steps.y    steps
## 1        0      NA 2012-10-01 1.716981 1.716981
## 2        0       0 2012-11-23 1.716981 0.000000
## 3        0       0 2012-10-28 1.716981 0.000000
## 4        0       0 2012-11-06 1.716981 0.000000
## 5        0       0 2012-11-24 1.716981 0.000000
## 6        0       0 2012-11-15 1.716981 0.000000
```

4.Make a histogram of the total number of steps taken each day and Calculate and report the mean and median total number of steps taken per day. Do these values differ from the estimates from the first part of the assignment? What is the impact of imputing missing data on the estimates of the total daily number of steps?

```
new_data_agg <- aggregate(steps ~ date, new_data, sum)
par(mfrow=c(1,2))
hist(new_data_agg$steps, col="red", xlab="No of steps",
ylab="Frequency", main="No of steps(aggregating NA values)")
hist(daily_steps$steps, col="blue", xlab="No of steps",
ylab="Frequency", main="No of steps(Original data)")
```


of steps(aggregating NA
Frequency
0
5
10
20
30
0
10000
25000
No of steps
No of steps(Original dat
Frequency
0
5
10
15
20
25
0
10000
25000
No of steps