Contents

Reproducible Research Project 1

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.

Loading and preprocessing the data

The data 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.

```
library(plyr)
library(ggplot2)
activity <-read.csv("C:/Users/omard/Desktop/Coursera/repdata-data-activity/activity.csv")
activity$day <- weekdays(as.Date(activity$date))
activity$DateTime<- as.POSIXct(activity$date, format="%Y-%m-%d")

##removing the NA's

clean <- activity[!is.na(activity$steps),]
```

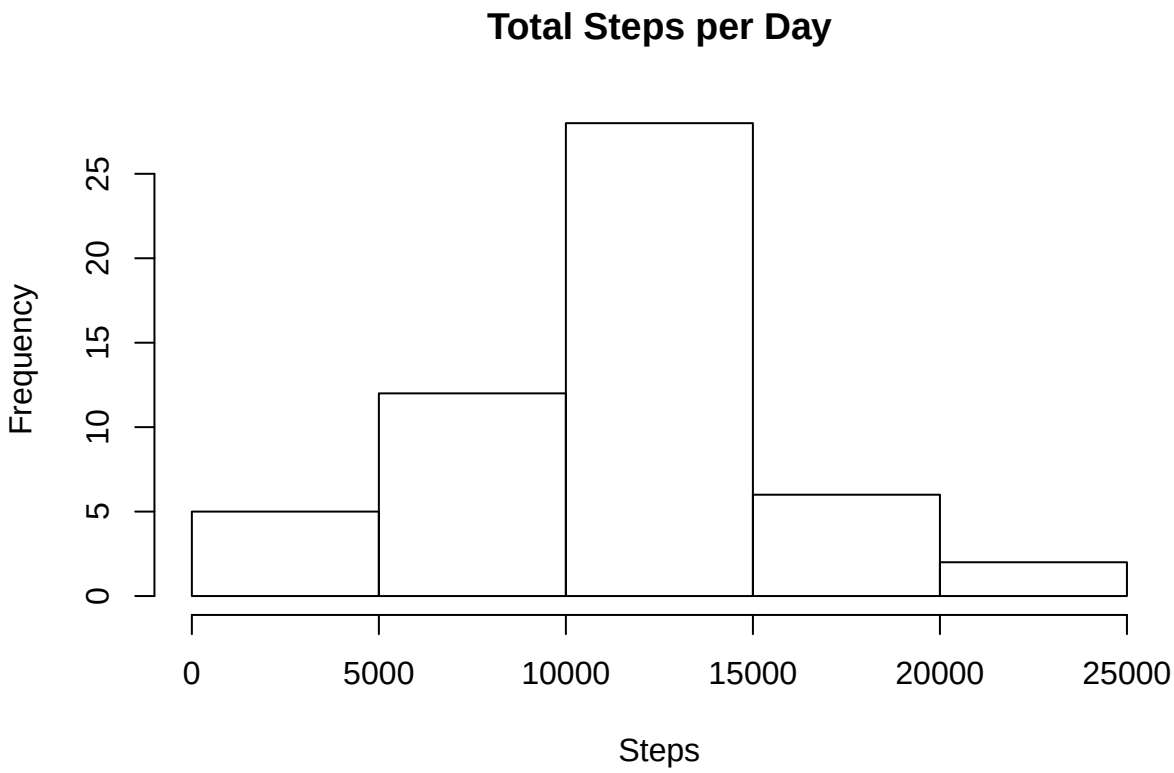
What is mean total number of steps taken per day?

1. Total number of steps taken per day

```
## summarizing total steps per date
Total_Table <- aggregate(activity$steps ~ activity$date, FUN=sum, )
colnames(Total_Table)<- c("Date", "Steps")
```

2. Make a histogram of the total number of steps taken each day

```
## Historgram of total steps per day
hist(Total_Table$Steps, breaks=5, xlab="Steps", main = "Total Steps per Day")
```



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

```
## Mean of Steps
as.integer(mean(Total_Table$Steps))
```

```
## [1] 10766
```

```
## Median of Steps
as.integer(median(Total_Table$Steps))
```

```
## [1] 10765
```

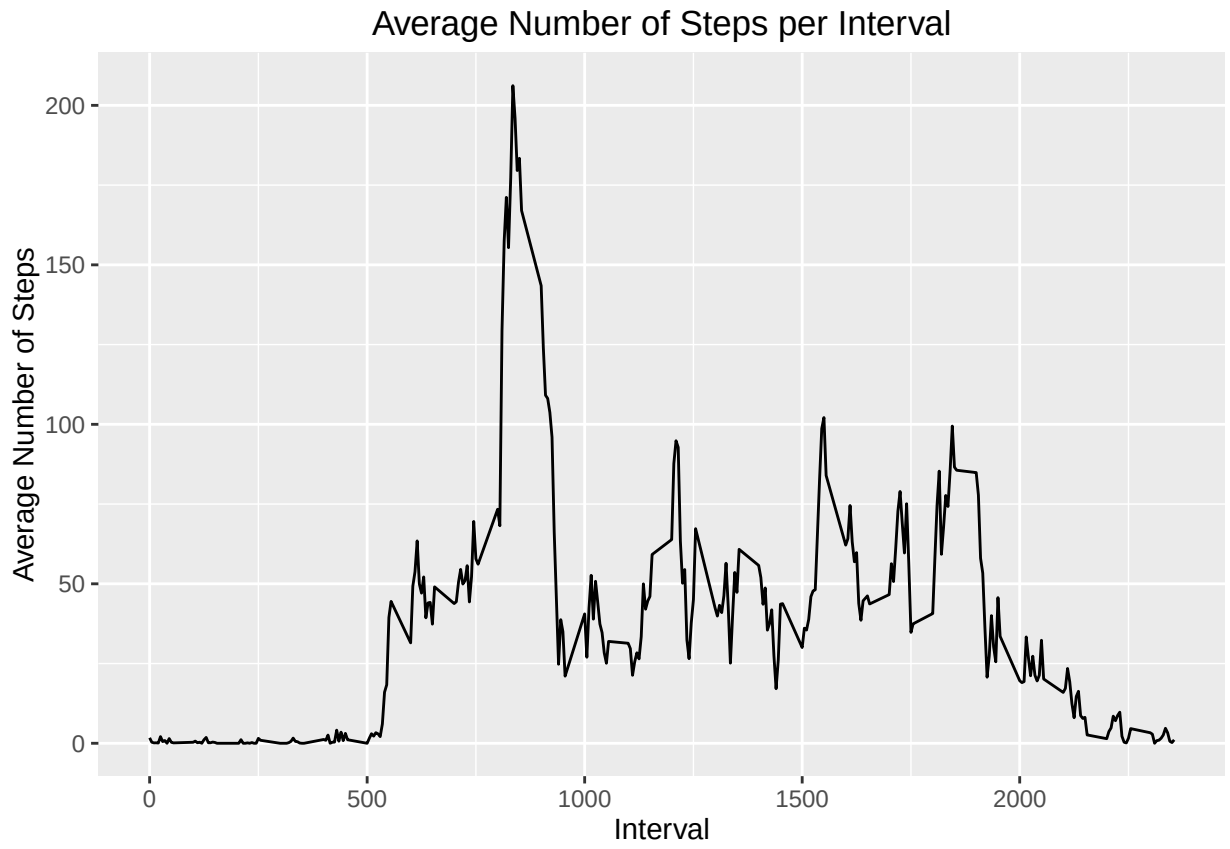
The average number of steps taken each day was 10766 steps.

The median number of steps taken each day was 10765 steps.

What is the average daily activity pattern?

1. Create 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)

```
library(plyr)
library(ggplot2)
## Data without NA's
clean <- activity[!is.na(activity$steps),]

##Average number of steps per interval (5 minutes)

intervalTable <- ddply(clean, .(interval), summarize, Avg = mean(steps))

##Line plot of average number of steps per interval

p <- ggplot(intervalTable, aes(x=interval, y=Avg), xlab = "Interval", ylab="Average Number of Steps")
p + geom_line()+xlab("Interval")+ylab("Average Number of Steps")+ggtitle("Average Number of Steps per In
```



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

```
##Maximum steps by interval

maxSteps <- max(intervalTable$Avg)

##Which interval contains the maximum average number of steps

intervalTable[intervalTable$Avg==maxSteps,1]
```

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

The maximum number of steps for a 5-minute interval was 206 steps.

The 5-minute interval which had the maximum number of steps was the 835 interval.

Imputing missing values

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

```
##Number of NAs in original data set

nrow(activity[is.na(activity$steps),])
```

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

The total number of rows with steps = 'NA' is 2304.

2. Strategy for filling in all of the missing values in the dataset: NA's are replaced by the day of the week 5-minute interval Average

```
## First calculate & summarize the average number of steps per weekday and interval

avgTable <- ddply(clean, .(interval, day), summarize, Avg = mean(steps))

## Second Create dataset with all NAs for substitution

nadata<- activity[is.na(activity$steps),]

## Merge NA data with average weekday interval for substitution

newdata<-merge(nadata, avgTable, by=c("interval", "day"))
```

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

```
## Format the new substituded data in the same format as clean data set

newdata2<- newdata[,c(6,4,1,2,5)]

colnames(newdata2)<- c("steps", "date", "interval", "day", "DateTime")

##Merge the clean and substitued data

mergeData <- rbind(clean, newdata2)
```

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 (NAs substituted with Daily 5-minutes interval Averages).

```
##Create sum of steps per date to compare with  clean data

Total_Table2 <- aggregate(mergeData$steps ~ mergeData$date, FUN=sum, )

colnames(Total_Table2)<- c("Date", "Steps")
```

```
## Mean of Steps with NA data substituted by 5-minutes interval averages

as.integer(mean(Total_Table2$Steps))
```

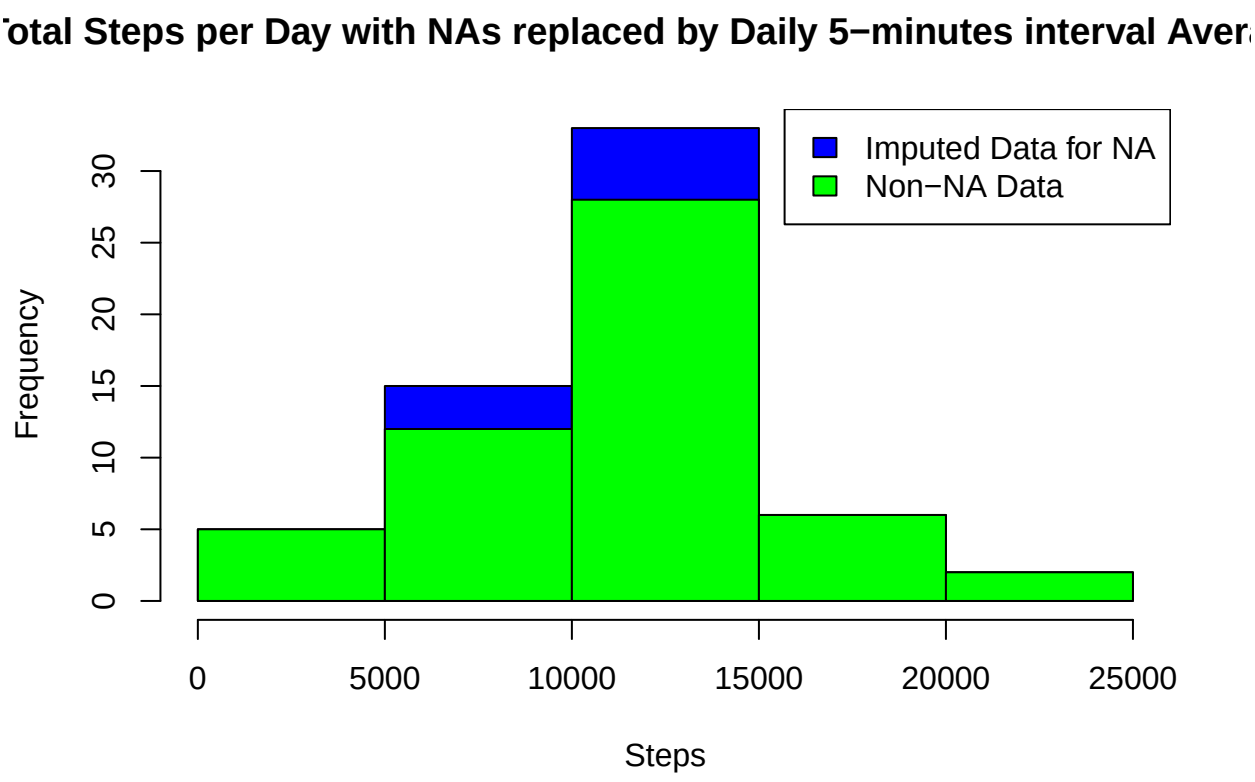
```
## [1] 10821
```

```
## Median of Steps with NA substituted

as.integer(median(Total_Table2$Steps))
```

```
## [1] 11015
```

hISTOGRAM

```
## Creating the histogram of total steps per day,differentiating clean (non-NA) vs I mputted data for NA

hist(Total_Table2$Steps, breaks=5, xlab="Steps", main = "Total Steps per Day with NAs replaced by Daily
hist(Total_Table$Steps, breaks=5, xlab="Steps", main = "Total Steps per Day with NAs replaced by Daily 5
legend("topright", c("Imputed Data for NA", "Non-NA Data"), fill=c("blue", "green") )
```



The new mean of the imputed data is 10821 steps compared to 10766 steps for the clean data

The new median of the imputed data is 11015 steps compared to 10765 steps for the clean data.

Same distribution shape

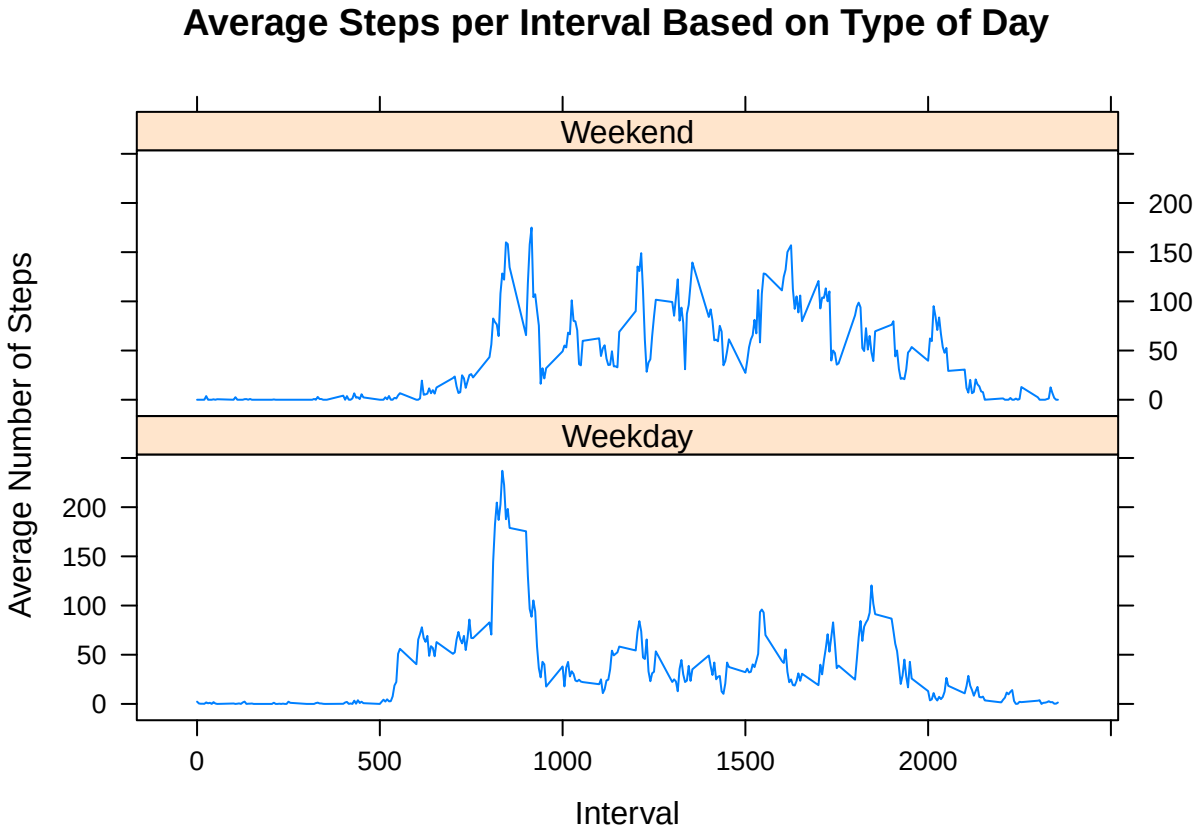
Are there differences in activity patterns between weekdays and weekends?

1- Create a dataset with two levels - "weekday" and "weekend" indicating whether a given date is a weekday or weekend day.

```
##Create a new factor variable in the dataset with two levels - "weekday" and "weekend"
mergeData$DayCategory <- ifelse(mergeData$day %in% c("Saturday", "Sunday"), "Weekend", "Weekday")
```

2- Make a panel plot containing 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 weekday days or weekend days (y-axis).

```
library(lattice)
## Summarize data by interval and type of day
intervalTable2 <- ddply(mergeData, .(interval, DayCategory), summarize, Avg = mean(steps))

##Plot data in a panel plot
xyplot(Avg~interval|DayCategory, data=intervalTable2, type="l",  layout = c(1,2),
       main="Average Steps per Interval Based on Type of Day",
       ylab="Average Number of Steps", xlab="Interval")
```



Yes, the step activity trends are different based on whether the day occurs on a weekend or not. Much higher activity during the week-end