Reproducible Research: Project 1

Juan Diego Solorzano Gomez

23/12/2020

Course Project 1

Loading and preprocessing the data and required libraries

```
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 4.0.3
```

```
library(ggplot2)
library(lattice)
activity<-read.csv("activity.csv")
activity$day <- weekdays(as.Date(activity$date))
activity$DateTime<- as.POSIXct(activity$date, format="%Y-%m-%d")
clean <- activity[!is.na(activity$steps),]
```

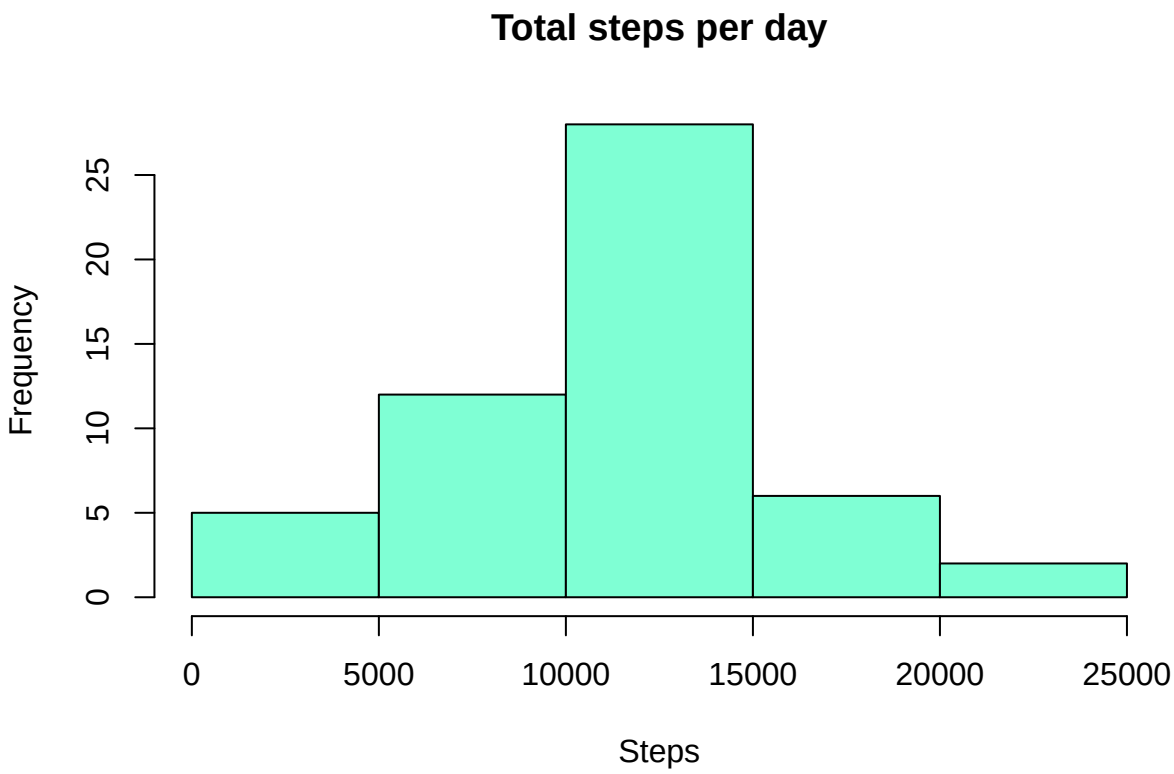
What is mean total number of steps taken per day?

Calculate the total number of steps taken per day is

```
stepsPerDay <- aggregate(activity$steps ~ activity$date, FUN = sum)
colnames(stepsPerDay) <- c("Date","Steps")
```

Histogram of the total number of steps taken each day

```
hist(stepsPerDay$Steps, breaks=5, xlab="Steps", main = "Total steps per day", col="Aquamarine")
```



The mean of steps taken per day was

```
mean(stepsPerDay$Steps)
```

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

The median of steps taken per day was

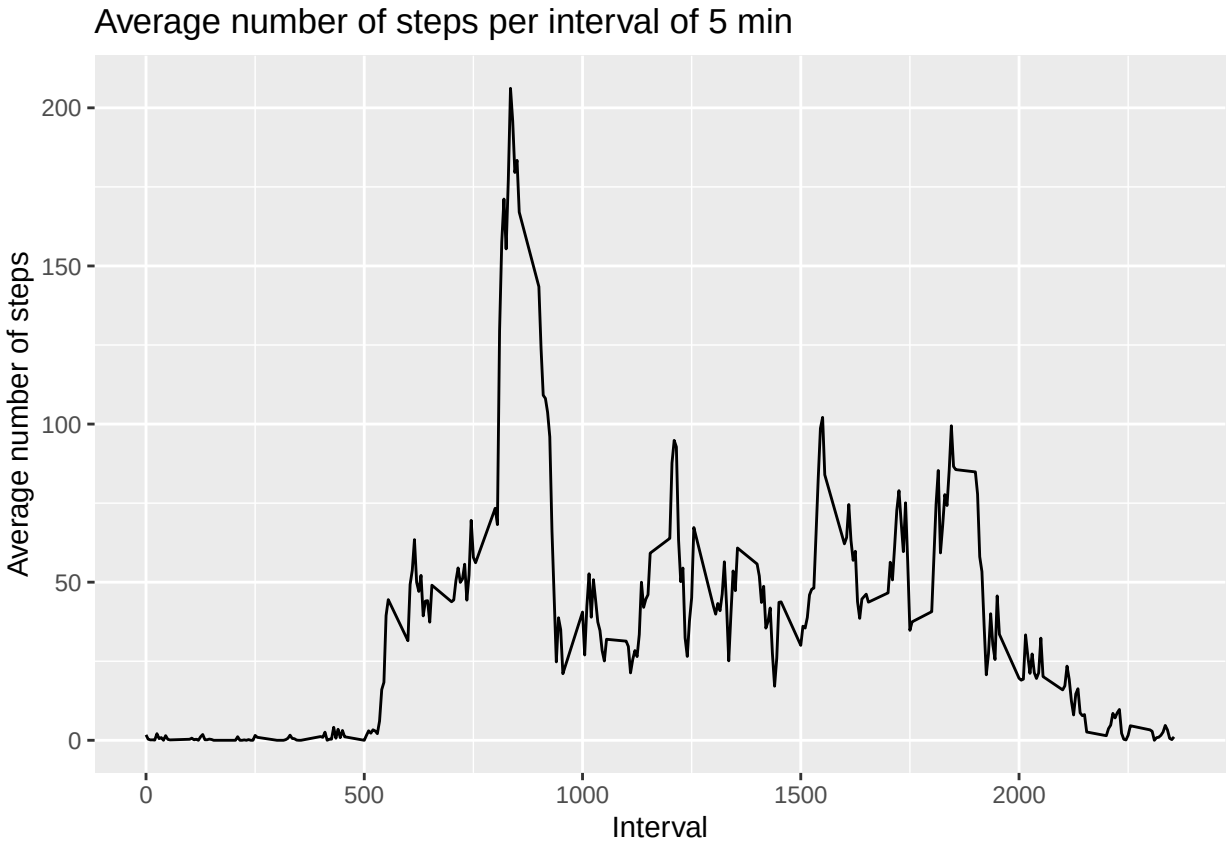
```
median(stepsPerDay$Steps)
```

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

What is the average daily activity pattern?

Make a time series plot of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis)

```
intervalTable <- ddply(clean, .(interval), summarize, Avg = mean(steps))
graphic <- ggplot(intervalTable, aes(x=interval,y=Avg), xlab="Interval",ylab="Average number of steps")
graphic + geom_line()+xlab("Interval")+ylab("Average number of steps")+ggtitle("Average number of steps
```



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

```
maxSteps <- max(intervalTable$Avg)
```

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

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

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

Imputing missing values

The total number of missing values in the dataset is

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

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

Devise a strategy for filling in all of the missing values in the dataset. My strategy will be to substitute the missing steps with the average 5-minute interval based on the day of the week

```
avgTable <- ddply(clean, .(interval, day), summarize, Avg = mean(steps))
nadata<- activity[is.na(activity$steps),]
newdata<-merge(nadata, avgTable, by=c("interval", "day"))
```

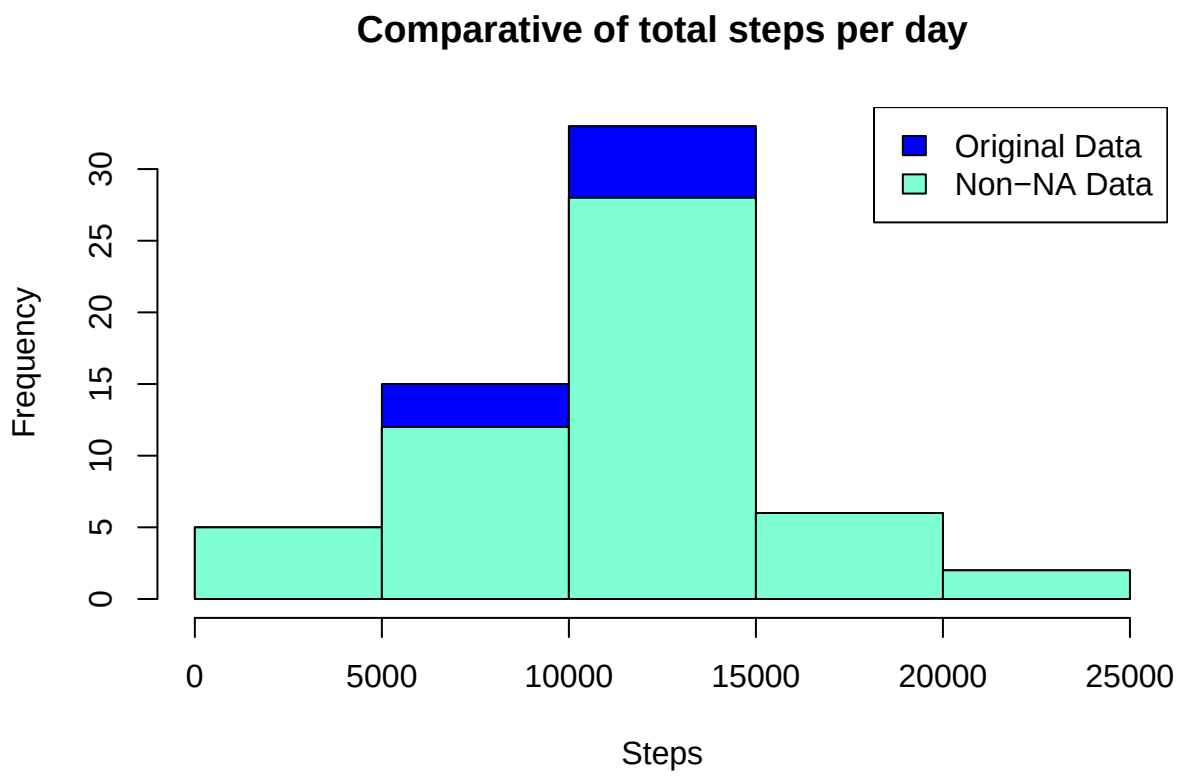
Create a new dataset that is equal to the original dataset but with the missing data filled in.

```
newdata2<- newdata[,c(6,4,1,2,5)]
colnames(newdata2)<- c("steps", "date", "interval", "day", "DateTime")
```

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

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?

```
stepsPerDay2 <- aggregate(mergeData$steps ~ mergeData$date, FUN = sum)
colnames(stepsPerDay2) <- c("Date","Steps")
hist(stepsPerDay2$Steps, breaks=5, xlab="Steps", main = "Comparative of total steps per day", col="Blue
hist(stepsPerDay$Steps, breaks=5, xlab="Steps", main = "Comparative of total steps per day", col="Aquama
legend("topright", c("Original Data", "Non-NA Data"), fill=c("Blue", "Aquamarine"))
```



The fixed mean is

```
mean(stepsPerDay2$Steps)
```

```
## [1] 10821.21
```

The fixed median is

```
median(stepsPerDay2$Steps)
```

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

Despite the differences in the statistics, little difference is observed between the distribution of the steps

Are there differences in activity patterns between weekdays and weekends?

Create a new factor variable in the dataset with two levels - "weekday" and "weekend" indicating whether a given date is a weekday or weekend day.

```
mergeData$DayCategory <- ifelse(mergeData$day %in% c("sábado", "domingo"), "Weekend", "Weekday")
```

Make a panel plot containing a time series plot of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all weekday days or weekend days (y-axis).

```
intervalTable2 <- ddply(mergeData, .(interval, DayCategory), summarize, Avg = mean(steps))
xyplot(Avg~interval|DayCategory, data=intervalTable2, type="l",  layout = c(1,2),
       main="Average steps per interval based on type of day",
       xlab="Interval", ylab="Average Number of Steps")
```