# Course Project 1 Reproducible Research

jguzmant

5/21/2020

## Activity Data set

It is now possible to collect a large amount of data about personal movement using activity monitoring devices such as a Fitbit, Nike Fuelband, or Jawbone Up. These type of devices are part of the "quantified self" movement – a group of enthusiasts who take measurements about themselves regularly to improve their health, to find patterns in their behavior, or because they are tech geeks. But these data remain under-utilized both because the raw data are hard to obtain and there is a lack of statistical methods and software for processing and interpreting the data.

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.

### Reading file

```
Activity_DF <- read.csv("activity.csv")
Activity_CP <- Activity_DF
Activity_CP$date <- as.Date(Activity_CP$date)
```

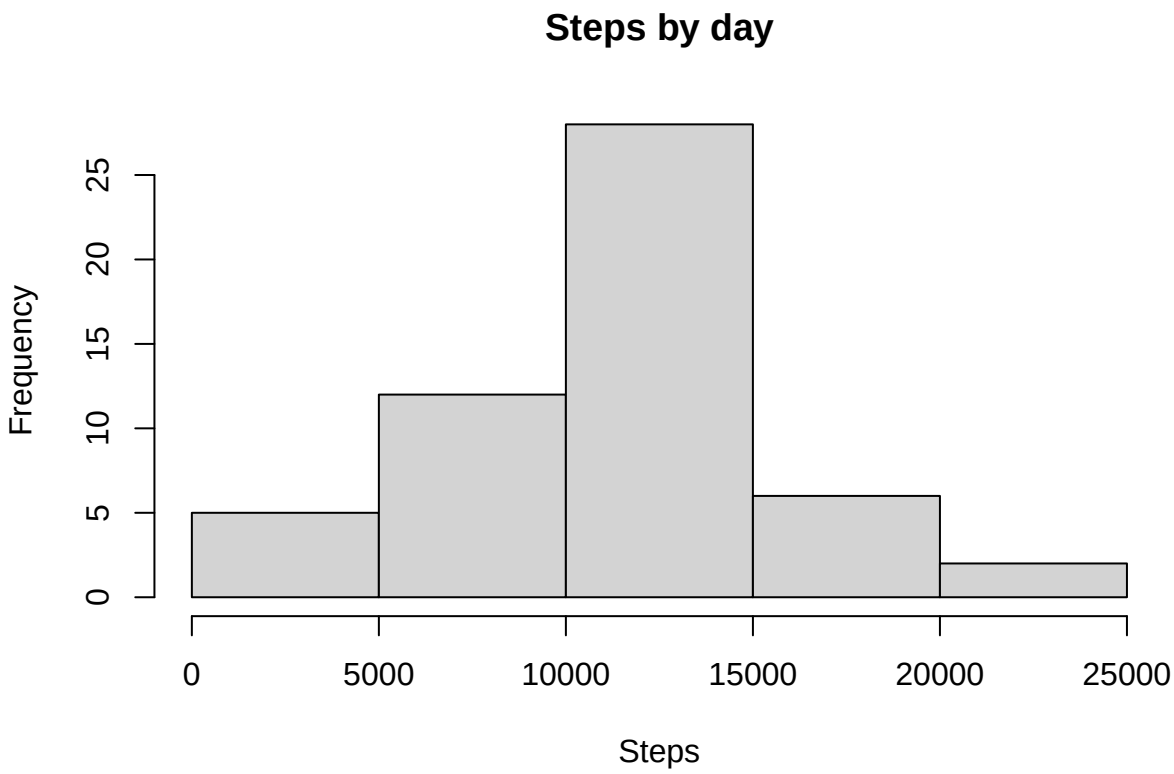
In order to maintain the original Activity table, Activity_CP was created. ## Question 1 What is mean total number of steps taken per day?

### 1.1 Code

```
Agg_Steps_date <- aggregate(steps~date,Activity_DF,sum,na.action = na.omit)
```

### 1.2 Histogram

```
hist(Agg_Steps_date$steps,main="Steps by day",xlab="Steps")
```




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

We already have the data in the form we need to obtain both the mean and median steps per day.

### 1.3 Code

```
meanSteps <-as.integer(round(mean(Agg_Steps_date$steps)))
medianSteps <- median(Agg_Steps_date$steps)
```

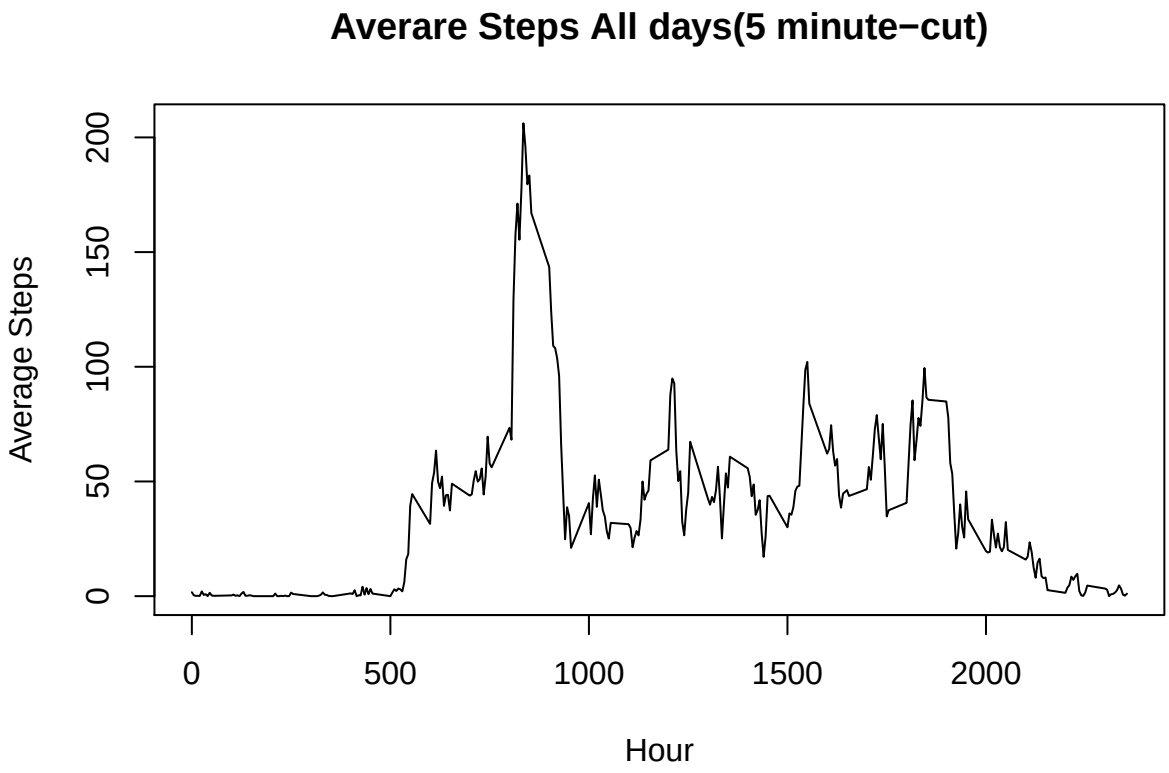
- The **mean** of steps per day is **10,766**
- The **median** steps per day is **10,765**

## Question 2.- What is the average daily activity pattern?

2.1 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)

### 2.1 Code

```
Agg_AVG_Steps_Interval <- aggregate(steps~interval,Activity_CP,mean)
with(Agg_AVG_Steps_Interval,plot(interval,steps,type="l",main="Averare Steps All days(5 minute-cut)",yla
```




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

**2.2 toTime Function** I created a function to convert the format of interval to a constant 5 character string, it takes a character vector and returns a character vector of 5 character each entry, e.g if receives "5" it returns "00:05"

```
toTime <- function(x){
    out <-NULL
    for(i in 1:length(x)){
        xlen <- nchar(x[i])
        newTime<-""
        if(xlen==1)
            newTime<-paste("000",x[i],sep="")
        else if(xlen==2)
            newTime<-paste("00",x[i],sep="")
        else if(xlen==3)
            newTime<-paste("0",x[i],sep="")
```

```
        else {
            newTime<-x[i]
            }
        out[i] <- paste(substr(newTime,1,2),substr(newTime,3,4),sep=":")
    }
out
}
```

**2.2 Code toTime function**

**2.2 Code**

```
MaxSteps <- max(Agg_AVG_Steps_Interval$steps)
TimeMaxSteps <- Agg_AVG_Steps_Interval$interval[grep(MaxSteps,Agg_AVG_Steps_Interval$steps)]
TimeMaxSteps <- toTime(as.character(TimeMaxSteps))
```

####The interval with the max steps is at **08:35**

## Question 3.- Imputing missing values

**3.1 Calculate and Report Total Number of Missing values**

**3.1 Code**

```
NASteps <- sum(is.na(Activity_CP$steps))
```

**The number of NA records in the data set is 2304**

**3.2 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.**

**3.2 Function to fill missing values Code**

```
fillMissing <- function(x,y){

    for (i in 1:length(y[,1])){
        x[x[,3]==y[i,1] & is.na(x[,1]),1] <- y[i,2]
    }
    x
}
```

**3.2 & 3.3 call to function to replace missing values**

```
Activity_CP <- fillMissing(Activity_CP,Agg_AVG_Steps_Interval)
summary(Activity_CP$steps)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   37.38   27.00  806.00
```

It can be noted that there are not missing values in the data set, the original data set is intact since i created a copy of it in the beginning.

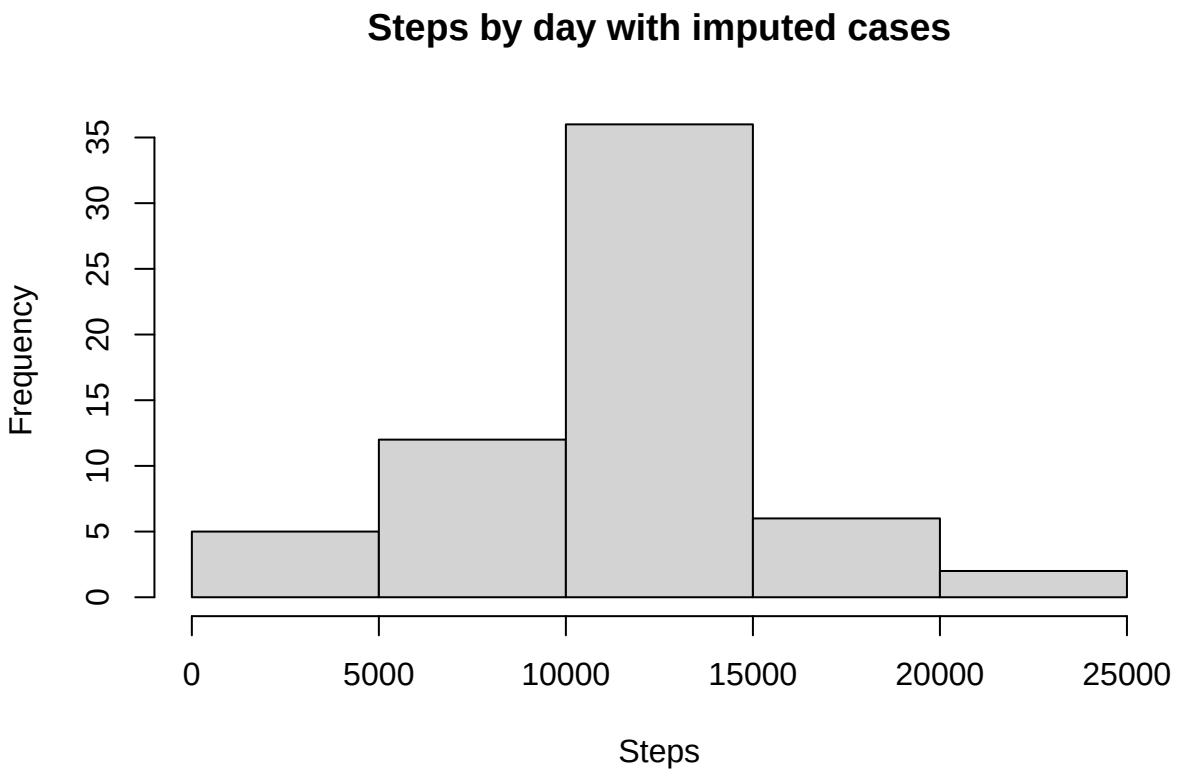
## 3.4 Make a histogram of the total number of steps taken each day and Calculate and report the mean and median total number of steps taken per day.

## 3.4 Code

```
Agg_Steps_date_Complete <- aggregate(steps~date,Activity_CP,sum)
meanSteps_Complete <-as.integer(round(mean(Agg_Steps_date_Complete $steps)))
medianSteps_Complete <- as.integer(median(Agg_Steps_date_Complete $steps))
```

### 3.4 a) Histogram with complete Cases

```
hist(Agg_Steps_date_Complete$steps,main="Steps by day with imputed cases",xlab="Steps")
```



### 3.4 b) Mean and Median for Data frame with imputed values

- The **mean** of steps per day is **10,766**
- The **median** steps per day is **10,766**

**Do these values differ from the estimates from the first part of the assignment?**

The mean and median were not really affected, they change too little to consider it had any impact.

**What is the impact of imputing missing data on the estimates of the total daily number of steps?**

The histogram frequency changed, the value is above 35 and before was above 25

##Question 4.-Are there differences in activity patterns between weekdays and weekends?

1.-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.

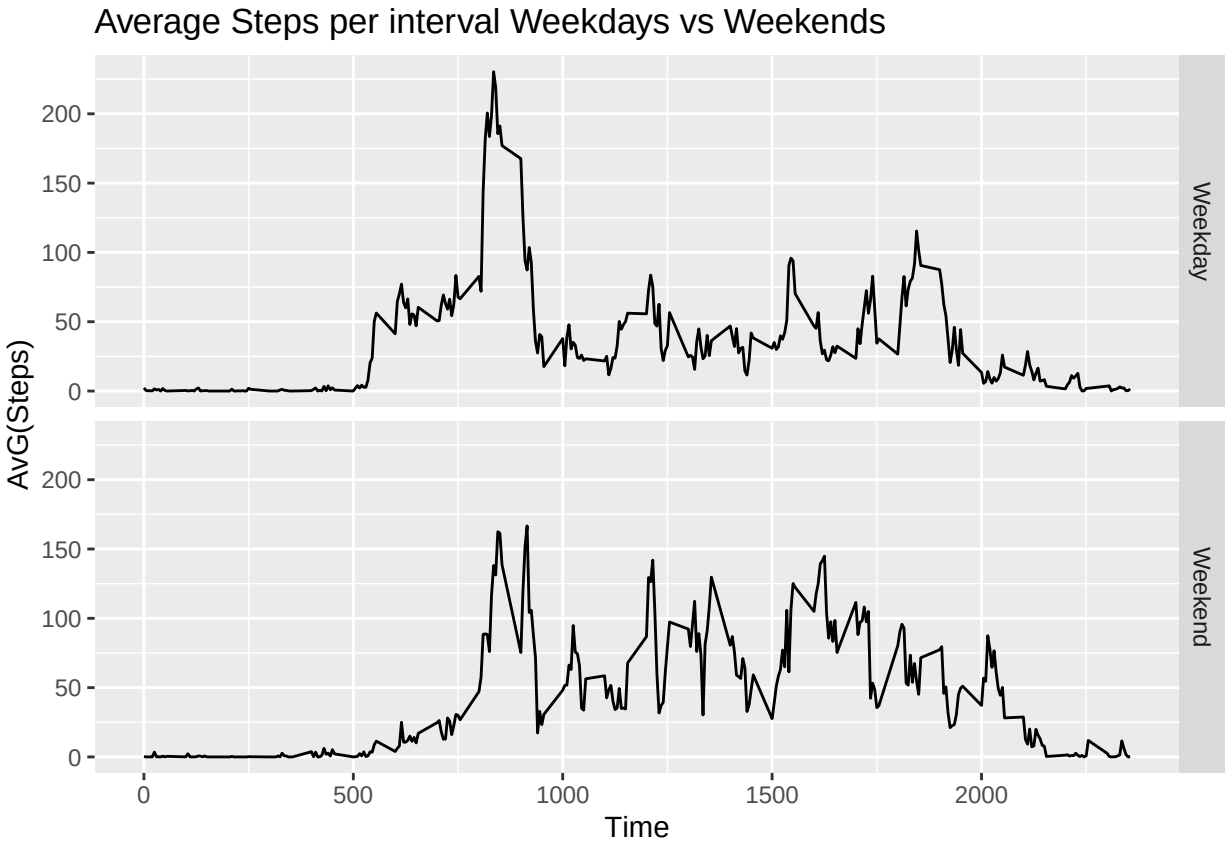
### 4.1 Code

```
Activity_CP$DayType <- weekdays(Activity_CP$date)
Activity_CP$WeekD <- factor(Activity_CP$DayType, levels = c("Monday", "Tuesday" , "Wednesday","Thursday
```

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). See the README file in the GitHub repository to see an example of what this plot should look like using simulated data.

## 4.2 Code

```
library(ggplot2)
Agg_Steps_date_Complete <-aggregate(steps~WeekD*interval,Activity_CP,mean)
outG<- ggplot(Agg_Steps_date_Complete)
outG + aes(x=interval,y=steps)+geom_line()+facet_grid(WeekD~.)+ggtitle("Average Steps per interval Week
```




### There are differences:

- In weekdays the subject starts moving little after 5 am, but on weekends he starts moving 2.5 hrs later
- During weekends he moves more in the 11-17 hrs span, in the same span, during weekdays he barely moves.
- Usually keeps moving ater eight at night on weekends. On weekdays he decreases his movement at 8.