PA1_template.Rmd

Jhon Sanabria

1/19/2021

```
knitr::opts_chunk$set(echo = TRUE)
```

Reproducible Research - Peer Graded Assignment # 1

This is a Mark Down document to answer the questions presented inthe Peer-Graded Assignment 1 of the Rep|roducible Research Curse.

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.

Teh data was downloaded and saved in a file in the computer. In order for R to read it we need to first set the working directory, load libraries to be used and read the data.

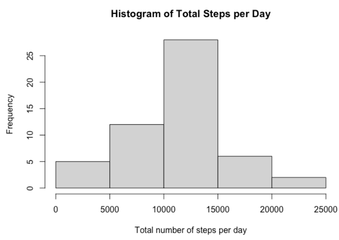
```
setwd("~/Desktop/datasciencecoursera/Reproducible research/week 2 project")
stepsdata <- read.csv("activity.csv", header=TRUE, na.strings = NA)
stepsdata$date <- as.Date(as.character(stepsdata$date))
```

What is mean total number of steps taken per day?

1. Calculate the total number of steps taken per day
2. Make a histogram of the total number of steps taken each day
3. Calculate and report the mean and median of the total number of steps taken per day

```
dailytotal <- aggregate(steps ~ date, data=stepsdata, FUN = "sum", na.rm = TRUE)
colnames(dailytotal) <- c("Date", "Total")

## create histogram of total steps per day (removing NAs)
dailyhist <- hist(dailytotal$Total, xlab = "Total number of steps per day",
                  main = "Histogram of Total Steps per Day")
```




```
## calculate mean and media of total number of steps
meanDT <- mean(dailytotal$Total)
medianDT <- median(dailytotal$Total)
```

The mean of the total number of steps (ignoring missing values) is

```
meanDT
```

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

The Median of the total number pf steps per day (ignoring the missing values) is

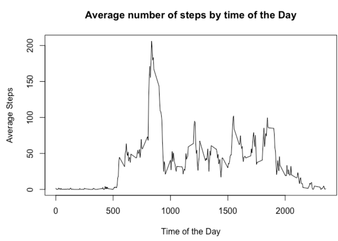
```
medianDT
```

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

What is the average daily activity pattern?

1. 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)
2. Which 5-minute interval, on average across all the days in the dataset, contains the maximum number of steps?

```
intervals <- aggregate(steps ~ interval, data=stepsdata, FUN = "mean", na.rm = TRUE)
Iplot <- plot(x = intervals$interval, y = intervals$steps, type = "l", main =
                    "Average number of steps by time of the Day", xlab =
                    "Time of the Day", ylab = "Average Steps")
```



The interval of the day with the maximun number of steps is

```
maxSIRow <- (which.max((intervals$steps)))
maxI <- intervals[c(maxSIRow), 1]
maxI
```

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

Imputing Missing Data

The presence of missing days may introduce bias into some calculations or summaries of the data. 1. Calculate and report the total number of missing values in the dataset (i.e. the total number of rows with NANAs) 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. Create a new dataset that is equal to the original dataset but with the missing data filled in. 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?

The percentage f missing data is

```
missing <- mean(is.na(stepsdata$steps))
missing
```

```
## [1] 0.1311475
```

Missing data will be replaced using the average number of steps taken per time interval

```
dailymean <- aggregate(steps ~ date, data = stepsdata, FUN = "mean", na.rm = TRUE)
impute_val <- function(interval) {
        intervals[intervals$interval==interval,]$steps
}
imputed <- stepsdata
for(i in 1:nrow(imputed)) {
        if(is.na(imputed[i,]$steps)) {
                imputed[i,]$steps <- impute_val(imputed[i,]$interval)
        }
}
```

After imputing the missing values, the mean and median number steps per day are

```
dailytotalWONAs <- aggregate(steps ~ date, data=imputed, FUN = "sum", na.rm = TRUE)
meanWONAs <- mean(dailytotalWONAs$steps)
```

```
medianWONAs <- median(dailytotalWONAs$steps)
```

Mean (with NAs replaced)

```
meanWONAs
```

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

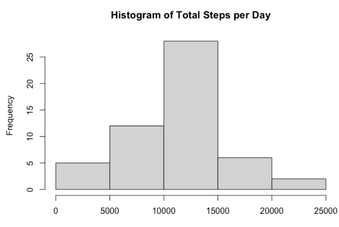
Median (with NAs replaced)

```
medianWONAs
```

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

Here is a histogram of the data with the missing values replaced

```
dailyhistWONAs <- hist(dailytotal$Total, xlab =
        "Total number of steps per day (After replacing missing data)",
                  main = "Histogram of Total Steps per Day")
```



The data seems to be equal to the data before the data imputation

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.
2. 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).

```
day <- weekdays(imputed$date)
wknd <- ifelse(day == "Saturday" | day=="Sunday",
               "Weekend", "Weekday")
imputed <- cbind(imputed, day, wknd)
library(lattice)
xyplot(steps ~ interval | wknd, data=imputed, type="l", layout=c(1,2))
```