Coursera Data Science Courses Projects Reproducible Research Week 2 Project Author: James C

Date: AUG, 2017

Loading and processing the data Load the data Process/transform the data (if necessary) into a format suitable for your analysis

setwd("C://Users//u182335//Documents//DataScience//Course 5 Week 2")

Have a look at the summary of data and first five rows of data

acti\_data <- read.csv("activity.csv")  
summary(acti\_data)

## steps date interval   
## Min. : 0.00 2012-10-01: 288 Min. : 0.0   
## 1st Qu.: 0.00 2012-10-02: 288 1st Qu.: 588.8   
## Median : 0.00 2012-10-03: 288 Median :1177.5   
## Mean : 37.38 2012-10-04: 288 Mean :1177.5   
## 3rd Qu.: 12.00 2012-10-05: 288 3rd Qu.:1766.2   
## Max. :806.00 2012-10-06: 288 Max. :2355.0   
## NA's :2304 (Other) :15840

head(acti\_data)

## steps date interval  
## 1 NA 2012-10-01 0  
## 2 NA 2012-10-01 5  
## 3 NA 2012-10-01 10  
## 4 NA 2012-10-01 15  
## 5 NA 2012-10-01 20  
## 6 NA 2012-10-01 25

What is mean total number of steps taken per day? The missing values in the dataset are ignored.

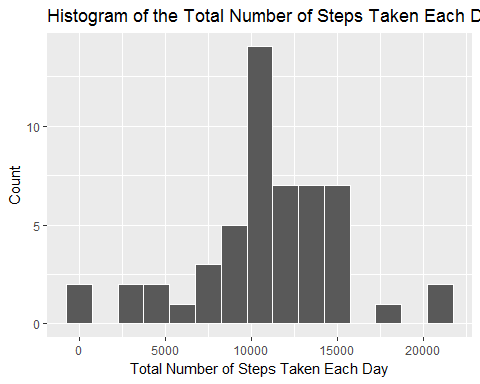
Calculate the total number of steps taken per day

steps <- aggregate(acti\_data$steps, by = list(Date = acti\_data$date), FUN = sum)  
library(ggplot2)  
names(steps)[names(steps) == "x"] <- "Total"  
temp <- as.Date(steps$Date, "%Y-%m-%d")  
steps$Date <- format(temp, format = "%m-%d")  
head(steps)

## Date Total  
## 1 10-01 NA  
## 2 10-02 126  
## 3 10-03 11352  
## 4 10-04 12116  
## 5 10-05 13294  
## 6 10-06 15420

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

hist1 <- ggplot(data = na.omit(steps), aes(Total)) +   
 geom\_histogram(binwidth = 1500, colour = "white") +  
 xlab("Total Number of Steps Taken Each Day") +  
 ylab("Count") +  
 ggtitle("Histogram of the Total Number of Steps Taken Each Day")  
print(hist1)



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

mean(na.omit(steps$Total))

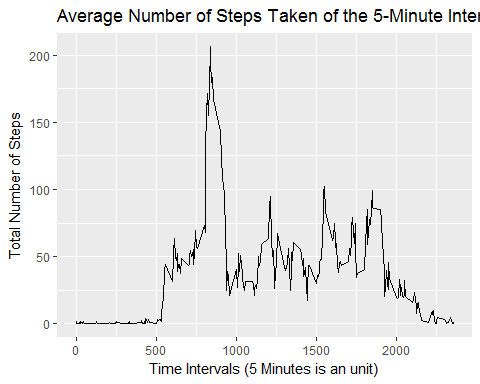
## [1] 10766.19

median(na.omit(steps$Total))

## [1] 10765

What is the average daily activity pattern? Make a time series plot (i.e. ???????????????? = "????") of the 5-minute interval (x-axis) and the aveage number of steps taken, averaged across all days (y-axis) type = "l" means the plot is line graph.

five\_min\_steps <- aggregate(steps ~ interval, data = acti\_data, FUN =mean)  
TimeSeries1 <- ggplot(data = five\_min\_steps, aes(x = interval, y = steps)) +   
 geom\_line() +  
 xlab("Time Intervals (5 Minutes is an unit)") +   
 ylab("Total Number of Steps") +  
 ggtitle("Average Number of Steps Taken of the 5-Minute Interval")  
print(TimeSeries1)

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

head(five\_min\_steps)

## interval steps  
## 1 0 1.7169811  
## 2 5 0.3396226  
## 3 10 0.1320755  
## 4 15 0.1509434  
## 5 20 0.0754717  
## 6 25 2.0943396

five\_min\_steps[which(five\_min\_steps$steps == max(five\_min\_steps$steps)),]

## interval steps  
## 104 835 206.1698

Imputing missing values Note that there are a number of days/intervals where there are missing values (coded as ????????). The presence of missing days may introduce bias into some calculations or summaries of the data.

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

sapply(X = acti\_data, FUN = function(x) sum(is.na(x)))

## steps date interval   
## 2304 0 0

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. I will use the mean for that 5 -minute interval to replace all the missing values in the dataset. At the end, I will check if all the NAs have been replaced.

library(dplyr)

## Warning: Installed Rcpp (0.12.9) different from Rcpp used to build dplyr (0.12.11).  
## Please reinstall dplyr to avoid random crashes or undefined behavior.

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

replace\_with\_mean <- function(num) replace(num, is.na(num), mean(num, na.rm = TRUE))  
meanday <- (acti\_data %>% group\_by(interval) %>% mutate(steps = replace\_with\_mean(steps)))  
head(meanday)

## # A tibble: 6 x 3  
## # Groups: interval [6]  
## steps date interval  
## <dbl> <fctr> <int>  
## 1 1.7169811 2012-10-01 0  
## 2 0.3396226 2012-10-01 5  
## 3 0.1320755 2012-10-01 10  
## 4 0.1509434 2012-10-01 15  
## 5 0.0754717 2012-10-01 20  
## 6 2.0943396 2012-10-01 25

sum(is.na(meanday))

## [1] 0

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

new\_dataset <- as.data.frame(meanday)  
head(new\_dataset)

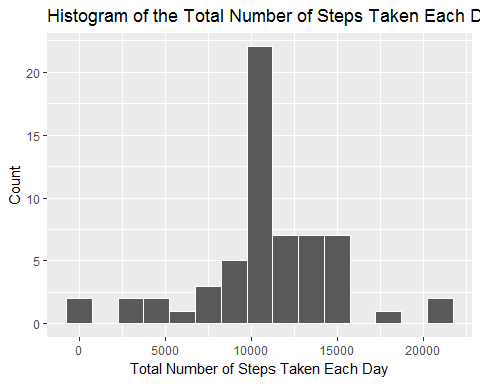
## steps date interval  
## 1 1.7169811 2012-10-01 0  
## 2 0.3396226 2012-10-01 5  
## 3 0.1320755 2012-10-01 10  
## 4 0.1509434 2012-10-01 15  
## 5 0.0754717 2012-10-01 20  
## 6 2.0943396 2012-10-01 25

summary(new\_dataset)

## steps date interval   
## Min. : 0.00 2012-10-01: 288 Min. : 0.0   
## 1st Qu.: 0.00 2012-10-02: 288 1st Qu.: 588.8   
## Median : 0.00 2012-10-03: 288 Median :1177.5   
## Mean : 37.38 2012-10-04: 288 Mean :1177.5   
## 3rd Qu.: 27.00 2012-10-05: 288 3rd Qu.:1766.2   
## Max. :806.00 2012-10-06: 288 Max. :2355.0   
## (Other) :15840

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? Make a histogram of the total number of steps taken each day first by using the new version dataset

new\_steps <- aggregate(new\_dataset$steps, by = list(new\_dataset$date), FUN = sum)  
names(new\_steps)[names(new\_steps) == "x"] <- "Total"  
names(new\_steps)[names(new\_steps) == "Group.1"] <- "Date"  
hist2 <- ggplot(data = new\_steps, aes(Total)) +   
 geom\_histogram(binwidth = 1500, colour = "white") +  
 xlab("Total Number of Steps Taken Each Day") +  
 ylab("Count") +  
 ggtitle("Histogram of the Total Number of Steps Taken Each Day with New Version Dataset")  
print(hist2)



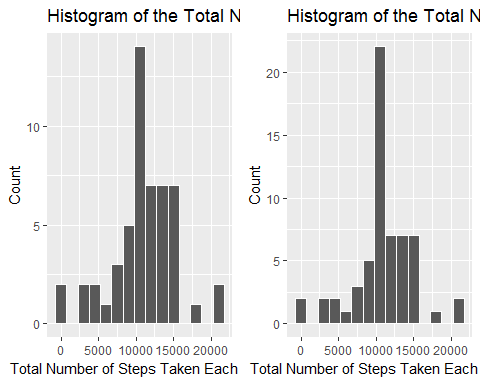
Compare the two plots.

library(grid)  
library(gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

grid.arrange(hist1, hist2, ncol = 2)



Compare the mean and median:

mean(na.omit(steps$Total))

## [1] 10766.19

median(na.omit(steps$Total))

## [1] 10765

mean(new\_steps$Total)

## [1] 10766.19

median(new\_steps$Total)

## [1] 10766.19

We find that the highest count of the new version data is larger than the one we have with NAs. The means of each dataset are the same. The medians of each dataset are slightly different.

Are there differences in activity patterns between weekdays and weekends? Use the dataset with the filled-in missing values which is called new\_steps dataset.

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.

new\_dataset$WeekendOrWeekday <- ifelse(weekdays(as.Date(new\_dataset$date)) %in%   
 c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday"), "Weekday", "Weekend")  
head(new\_dataset)

## steps date interval WeekendOrWeekday  
## 1 1.7169811 2012-10-01 0 Weekday  
## 2 0.3396226 2012-10-01 5 Weekday  
## 3 0.1320755 2012-10-01 10 Weekday  
## 4 0.1509434 2012-10-01 15 Weekday  
## 5 0.0754717 2012-10-01 20 Weekday  
## 6 2.0943396 2012-10-01 25 Weekday

Make a panel plot containing a time series plot (i.e. ???????????????? = "????") 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.

