RR\_Week2 Project

Shrikant

25/04/2020

## Set Working Directory

I set the working directory to appropriate folder

setwd("C:/Users/shrikant.shendage/Documents/Transfer/RR\_week2 Project")  
library(dplyr)

##   
## 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

library(ggplot2)

## Loading and preprocessing the data

activity<- read.csv("activity.csv")  
str(activity)

## 'data.frame': 17568 obs. of 3 variables:  
## $ steps : int NA NA NA NA NA NA NA NA NA NA ...  
## $ date : Factor w/ 61 levels "2012-10-01","2012-10-02",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ interval: int 0 5 10 15 20 25 30 35 40 45 ...

summary(activity)

## 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(activity)

## 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

## Processing the data

Remove the missing values

act.complete <- na.omit(activity)

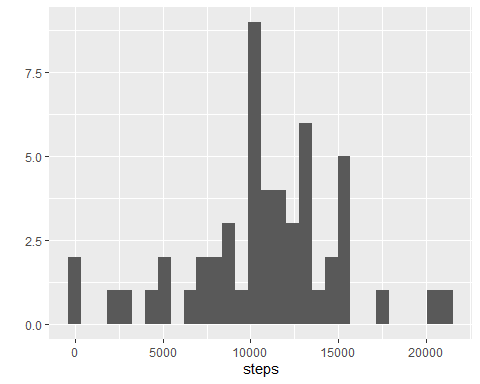
Calculate the total number of steps taken per day

act.day <- group\_by(act.complete, date)  
act.day <- summarize(act.day, steps=sum(steps))  
summary(act.day)

## date steps   
## 2012-10-02: 1 Min. : 41   
## 2012-10-03: 1 1st Qu.: 8841   
## 2012-10-04: 1 Median :10765   
## 2012-10-05: 1 Mean :10766   
## 2012-10-06: 1 3rd Qu.:13294   
## 2012-10-07: 1 Max. :21194   
## (Other) :47

qplot(steps, data=act.day)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



Calculate the mean and median

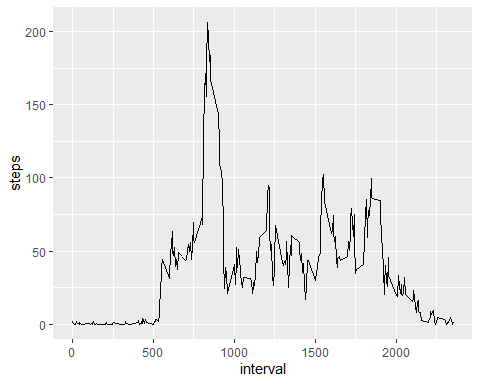
mean(act.day$steps)

## [1] 10766.19

median(act.day$steps)

## [1] 10765

act.int <- group\_by(act.complete, interval)  
act.int <- summarize(act.int, steps=mean(steps))  
ggplot(act.int, aes(interval, steps)) + geom\_line()



5 - minute interval identifiation

act.int[act.int$steps==max(act.int$steps),]

## # A tibble: 1 x 2  
## interval steps  
## <int> <dbl>  
## 1 835 206.

nrow(activity)-nrow(act.complete)

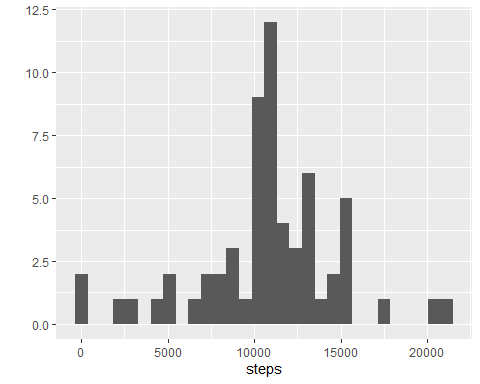
## [1] 2304

names(act.int)[2] <- "mean.steps"  
act.impute <- merge(activity, act.int)  
act.impute$steps[is.na(act.impute$steps)] <- act.impute$mean.steps[is.na(act.impute$steps)]  
act.day.imp <- group\_by(act.impute, date)  
act.day.imp <- summarize(act.day.imp, steps=sum(steps))

Histogram & summary

qplot(steps, data=act.day.imp)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



mean(act.day.imp$steps)

## [1] 10766.19

median(act.day.imp$steps)

## [1] 10766.19

Difference in activity pattern between weekdays and weekends

act.impute$dayofweek <- weekdays(as.Date(act.impute$date))  
act.impute$weekend <-as.factor(act.impute$dayofweek=="Saturday"|act.impute$dayofweek=="Sunday")  
levels(act.impute$weekend) <- c("Weekday", "Weekend")  
act.weekday <- act.impute[act.impute$weekend=="Weekday",]  
act.weekend <- act.impute[act.impute$weekend=="Weekend",]  
act.int.weekday <- group\_by(act.weekday, interval)  
act.int.weekday <- summarize(act.int.weekday, steps=mean(steps))  
act.int.weekday$weekend <- "Weekday"  
act.int.weekend <- group\_by(act.weekend, interval)  
act.int.weekend <- summarize(act.int.weekend, steps=mean(steps))  
act.int.weekend$weekend <- "Weekend"  
act.int <- rbind(act.int.weekday, act.int.weekend)  
act.int$weekend <- as.factor(act.int$weekend)  
ggplot(act.int, aes(interval, steps)) + geom\_line() + facet\_grid(weekend ~ .)

