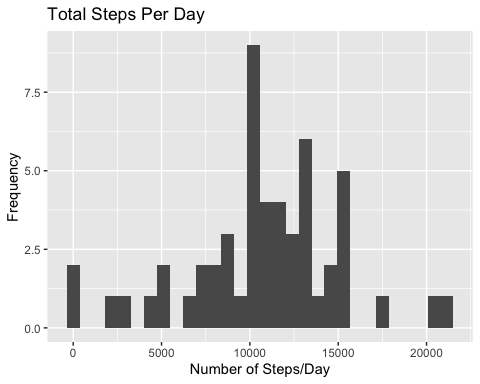
activity <- read.csv("~/Desktop/activity.csv")  
library(ggplot2)

s\_d <- aggregate(activity$steps, by = list(date=activity$date), FUN=sum)  
colnames(s\_d) <- c("Date", "Steps")

qplot(s\_d$Steps, geom = "histogram", xlab = "Number of Steps/Day", ylab = "Frequency", main="Total Steps Per Day")

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

## Warning: Removed 8 rows containing non-finite values (stat\_bin).



mean(s\_d$Steps)

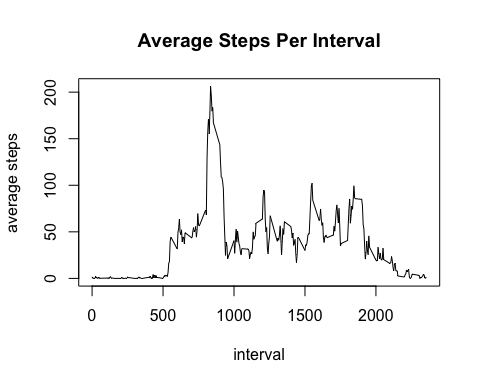
## [1] NA

median(s\_d$Steps)

## [1] NA

**The mean steps taken each day is 10766 and the median steps taken each day is 10765.**

s\_i <- aggregate(activity$steps, by=list(activity$interval), FUN=mean, na.rm=TRUE)  
colnames(s\_i) <- c("interval", "average steps")  
s\_i\_plot <- plot(s\_i, type="l", main="Average Steps Per Interval")

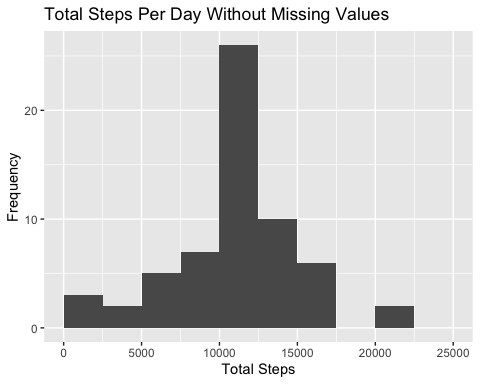
 The maximum number of steps occurs during the interval 835.

summary(activity)

## steps date interval   
## Min. : 0.00 Length:17568 Min. : 0.0   
## 1st Qu.: 0.00 Class :character 1st Qu.: 588.8   
## Median : 0.00 Mode :character Median :1177.5   
## Mean : 37.38 Mean :1177.5   
## 3rd Qu.: 12.00 3rd Qu.:1766.2   
## Max. :806.00 Max. :2355.0   
## NA's :2304

replace\_steps <- s\_i$`average steps` [match(activity$interval, s\_i$interval)]  
inputed <- transform(activity, steps=ifelse(is.na(activity$steps), yes=replace\_steps, no=activity$steps))

input\_sum <- aggregate(steps~date, inputed, sum)  
qplot(input\_sum$steps, geom = "histogram", xlab = "Total Steps", ylab = "Frequency", main="Total Steps Per Day Without Missing Values", breaks=seq(0,25000,by=2500))



noNA\_mean <- mean(input\_sum$steps)  
noNA\_median <- median(input\_sum$steps)  
NA\_mean <- mean(s\_d$Steps)  
NA\_median <- median(s\_d$Steps)  
compare <- data.frame(mean=c(noNA\_mean, NA\_mean), median=c(noNA\_median, NA\_median))  
rownames(compare) <- c("NA\_included", "NA\_substitute")  
print(compare)

## mean median  
## NA\_included 10766.19 10766.19  
## NA\_substitute NA NA

The mean remains the same when the NAs are replaced with the mean total steps/day however, the median is slightly lower when the NA values are replaced.By inputting the missing values there is a slight change in the estimates of total daily/steps.

inputed$date <- as.Date(inputed$date)  
inputed$day <- ifelse(weekdays(inputed$date) %in% c("Saturday", "Sunday"), "weekend", "weekday")  
inputed$day <- as.factor(inputed$day)  
day\_end <- aggregate(steps~interval+day, data=inputed, FUN=mean, na.action = na.omit)  
plot1 <- ggplot(day\_end, aes(interval, steps))  
plot\_day <- plot1+geom\_line(col="blue")+ggtitle("Steps: Weekend vs. Weekday")+xlab("Time Interval")+ylab("Average Steps")+facet\_grid(day ~ .)  
print(plot\_day)

