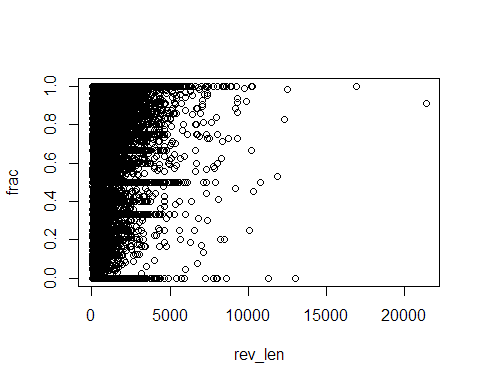
# DS 710 Assignment 11

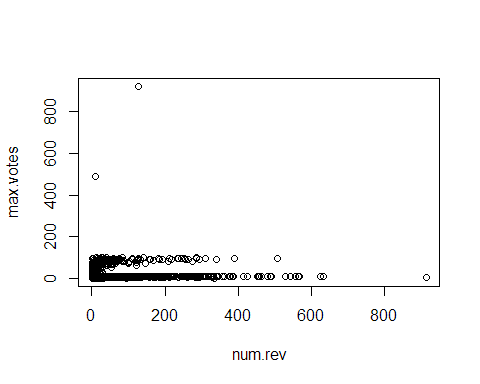
#1a.  
d <- scan('df.csv',   
 what=character(),  
 sep=',',   
 skip=1,  
 nlines=600000)  
  
#1b.  
d.matrix <- t(matrix(d, nr=7))  
d.header <- scan('df.csv', what=character(), sep=',', nlines=1)  
colnames(d.matrix) <- d.header  
  
#1c.  
total <- as.numeric(d.matrix[,3])  
rev\_len <- as.numeric(d.matrix[,5])  
excl <- as.numeric(d.matrix[,6])  
frac <- as.numeric(d.matrix[,7])  
  
#1d.  
total[which(frac > 1)] <- NA  
frac[which(frac > 1)] <- NA  
  
#1f.  
helpful\_rev <- rep(FALSE, length(frac))  
helpful\_rev[which(frac > 0.5)] <- TRUE  
  
#1g.  
#plot(helpful\_rev ~ rev\_len)  
#barplot(by(rev <- len, helpful <- rev, mean))  
#I choose to zoom in to see mean and IQR but it does cut off the range and outliers.  
#boxplot(rev\_len~helpful\_rev,   
# ylim=c(0,1000),   
# xlab=c("Helpful Review"),   
# ylab=c("Review Length"))  
plot(frac ~ rev\_len)



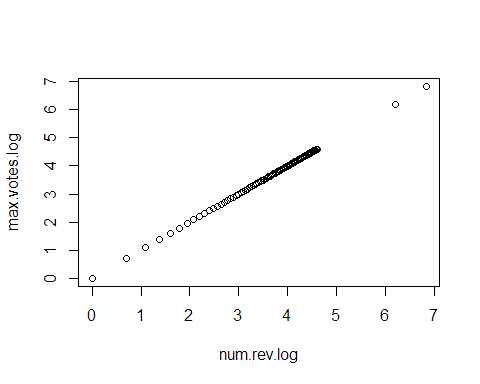
t.test(rev\_len~helpful\_rev)

##   
## Welch Two Sample t-test  
##   
## data: rev\_len by helpful\_rev  
## t = -87.905, df = 382834.6, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -114.4836 -109.4898  
## sample estimates:  
## mean in group FALSE mean in group TRUE   
## 391.5688 503.5555

#1h.  
max.votes <- tapply(d.matrix[,'total'], d.matrix[, 'prod'], FUN=max)  
num.rev <- tapply(d.matrix[,'rev'], d.matrix[, 'prod'], FUN=length)  
  
#1i.  
plot(max.votes~num.rev)



#1j.  
max.votes2 <- max.votes[which(max.votes >= 1)]  
max.votes.log <- log(as.numeric(max.votes2))  
num.rev.log <- log(as.numeric(max.votes2))  
  
#1k.  
plot(max.votes.log ~ num.rev.log)



### 1e.

I found to values (1.5 and 3) in the fraction column. It is obviously not possible to have more helpful votes than the total votes.

### 1g.

Using a 2-sample t-test, we are able to reject the null hypothesis (p-value < 2.2x10-16). That is to say the difference in the mean review length of helpful versus unhelpful is significant.

### 1k.

There is a strong positive linear trend. As the log of num of revs increase, so do the max number of votes for the product.