Problem Set 4

Karl Hickel

a.

install.packages('ggplot2movies')

b.

```r
library(ggplot2movies)
library(ggplot2)
```

c.

```r
data(movies)
summary(movies)
```

```
##     title                year          length            budget
##  Length:58788       Min.   :1893   Min.   :   1.00   Min.   :        0
##  Class :character   1st Qu.:1958   1st Qu.:  74.00   1st Qu.:   250000
##  Mode  :character   Median :1983   Median :  90.00   Median :  3000000
##                     Mean   :1976   Mean   :  82.34   Mean   : 13412513
##                     3rd Qu.:1997   3rd Qu.: 100.00   3rd Qu.: 15000000
##                     Max.   :2005   Max.   :5220.00   Max.   :200000000
##                                                      NA's   :53573
##      rating           votes                r1               r2
##  Min.   : 1.000   Min.   :     5.0   Min.   :  0.000   Min.   : 0.000
##  1st Qu.: 5.000   1st Qu.:    11.0   1st Qu.:  0.000   1st Qu.: 0.000
##  Median : 6.100   Median :    30.0   Median :  4.500   Median : 4.500
##  Mean   : 5.933   Mean   :   632.1   Mean   :  7.014   Mean   : 4.022
##  3rd Qu.: 7.000   3rd Qu.:   112.0   3rd Qu.:  4.500   3rd Qu.: 4.500
##  Max.   :10.000   Max.   :157608.0   Max.   :100.000   Max.   :84.500
##
##        r3               r4                r5                r6
##  Min.   : 0.000   Min.   :  0.000   Min.   :  0.000   Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.:  0.000   1st Qu.:  4.500   1st Qu.: 4.50
##  Median : 4.500   Median :  4.500   Median :  4.500   Median :14.50
##  Mean   : 4.721   Mean   :  6.375   Mean   :  9.797   Mean   :13.04
##  3rd Qu.: 4.500   3rd Qu.:  4.500   3rd Qu.: 14.500   3rd Qu.:14.50
##  Max.   :84.500   Max.   :100.000   Max.   :100.000   Max.   :84.50
##
##        r7               r8               r9                r10
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Min.   :  0.00
##  1st Qu.:  4.50   1st Qu.:  4.50   1st Qu.:  4.500   1st Qu.:  4.50
##  Median : 14.50   Median : 14.50   Median :  4.500   Median : 14.50
##  Mean   : 15.55   Mean   : 13.88   Mean   :  8.954   Mean   : 16.85
##  3rd Qu.: 24.50   3rd Qu.: 24.50   3rd Qu.: 14.500   3rd Qu.: 24.50
##  Max.   :100.00   Max.   :100.00   Max.   :100.000   Max.   :100.00
##
##      mpaa              Action         Animation           Comedy
##  Length:58788       Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  Class :character   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
##  Mode  :character   Median :0.00000   Median :0.00000   Median :0.0000
##                     Mean   :0.07974   Mean   :0.06277   Mean   :0.2938
##                     3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000
##                     Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##
##      Drama        Documentary         Romance           Short
##  Min.   :0.000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.00000   Median :0.0000   Median :0.0000
##  Mean   :0.371   Mean   :0.05906   Mean   :0.0807   Mean   :0.1609
##  3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##
```

```r
dim(movies)
```

```
## [1] 58788    24
```

Rows 58788 and 24 columns

d.

```r
help(movies)
```

possibly rating and budget

e.

is.na(movies)

```r
sum(is.na(movies$budget))
```

```
## [1] 53573
```

Total missing value ˆ and we have 5215 values that are not missing

f.

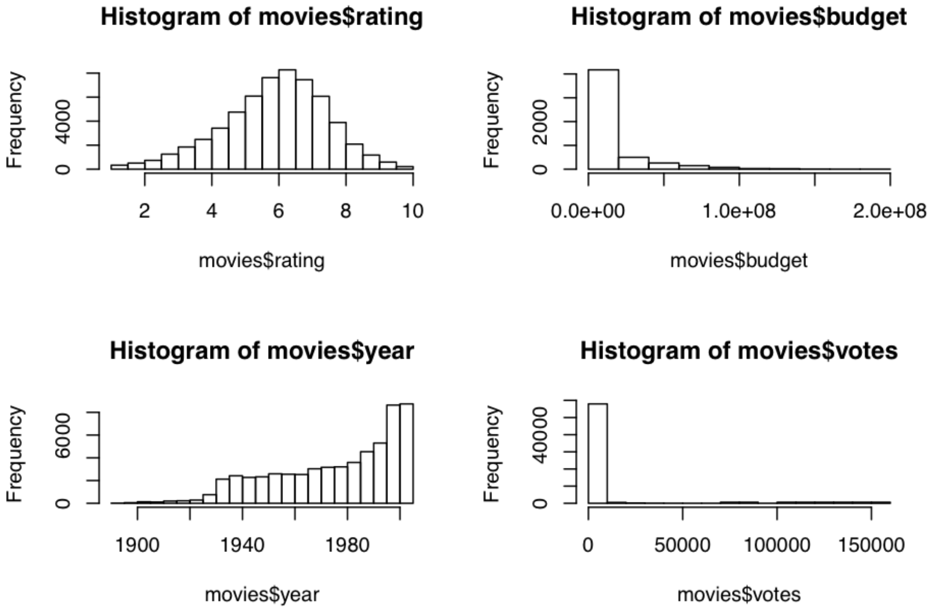
```r
moviesSub <- movies[!is.na(movies$budget),]
colSums(is.na(moviesSub))
```

```
##       title        year      length      budget      rating       votes
##           0           0           0           0           0           0
##          r1          r2          r3          r4          r5          r6
##           0           0           0           0           0           0
##          r7          r8          r9         r10        mpaa      Action 
##           0           0           0           0           0           0 
##   Animation      Comedy       Drama Documentary     Romance       Short 
##           0           0           0           0           0           0 
```

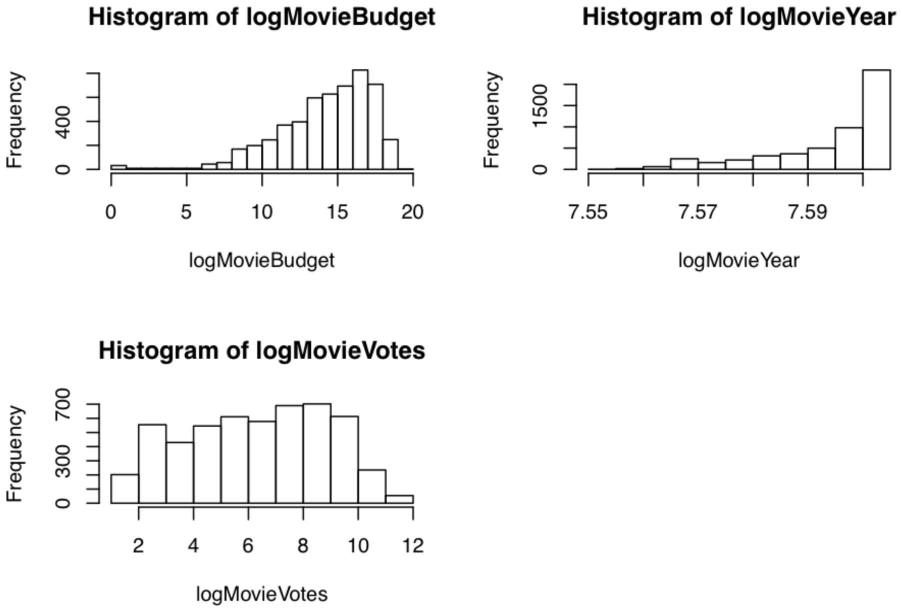
View(moviesSub)

g.

```
par(mfrow=c(2,2))
hist(movies$rating)
hist(movies$budget) #Skewed to the right
hist(movies$year) #Skewed to the left.
hist(movies$votes) #Skewed to the right.
```

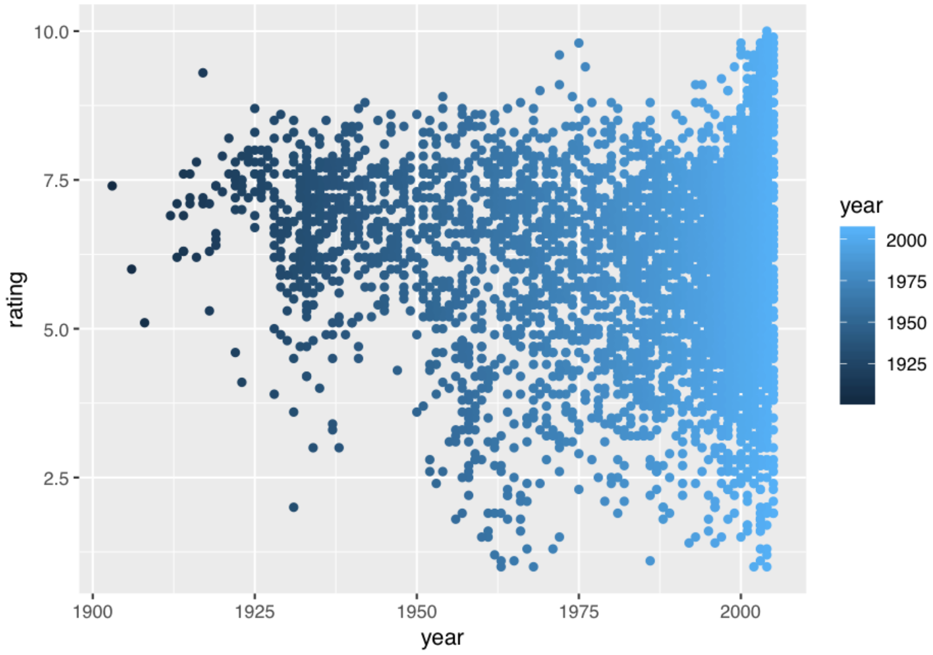


```
logMovieBudget <- log(moviesSub$budget + 1)
hist(logMovieBudget)
logMovieYear <- log(moviesSub$year + 1)
hist(logMovieYear)
logMovieVotes <- log(moviesSub$votes + 1)
hist(logMovieVotes)
```

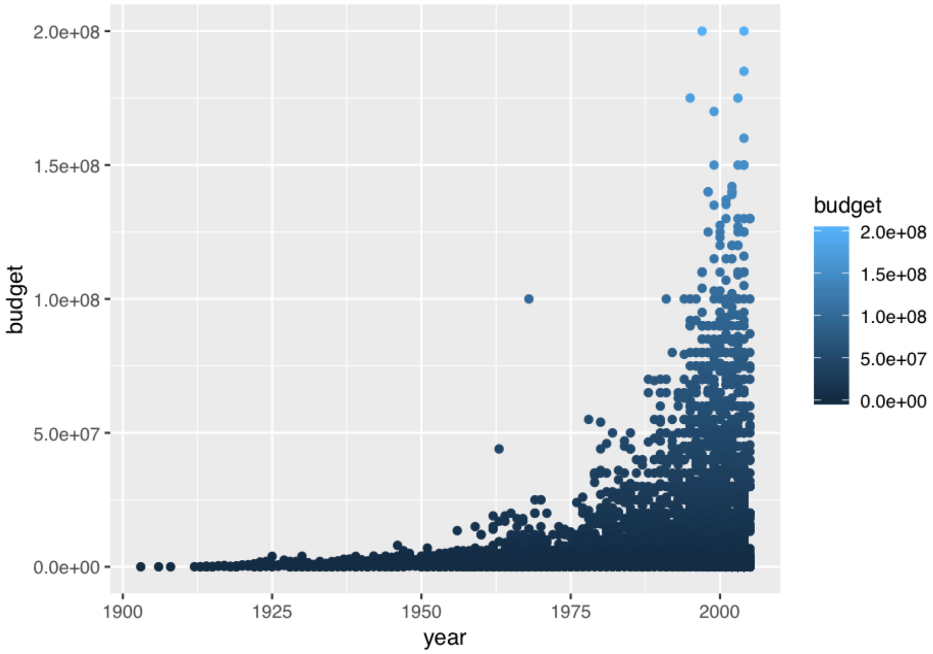


h

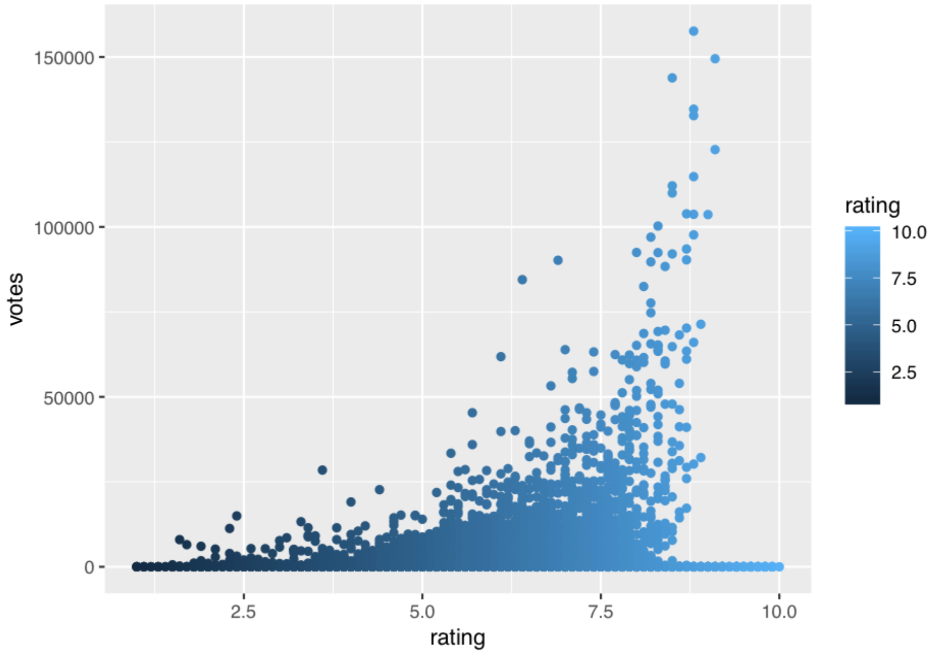
```
par(mfrow=c(2,2))
help(movies)
ggplot(moviesSub, aes(year,rating)) + geom_point(aes(color = year))
```



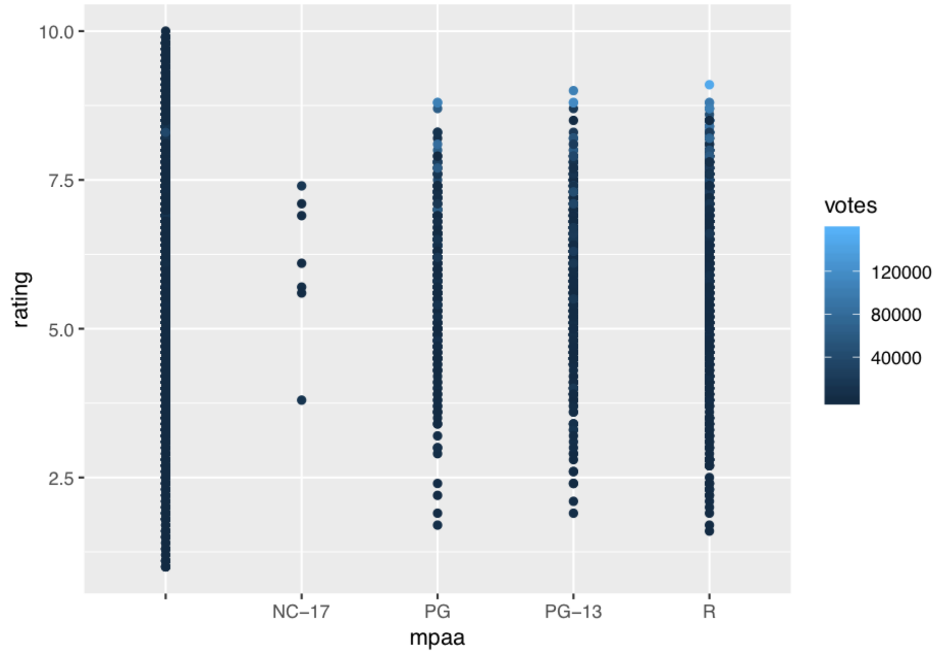
```r
ggplot(moviesSub, aes(year,budget)) + geom_point(aes(color = budget))
```



```
ggplot(moviesSub, aes(rating,votes)) + geom_point(aes(color = rating))
```

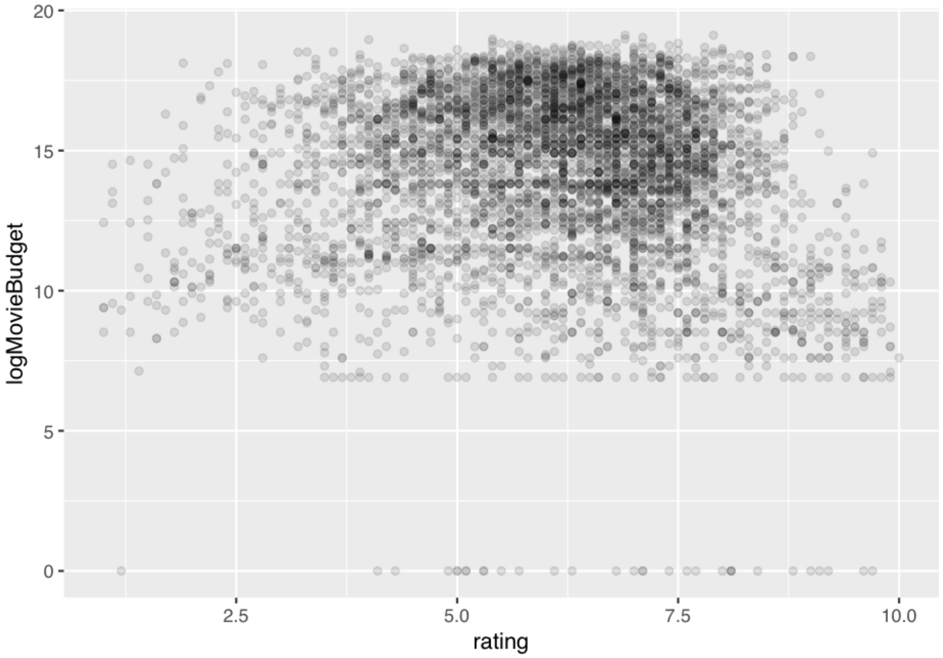


```
ggplot(moviesSub, aes(mpaa,rating)) + geom_point(aes(color = votes))
```



i

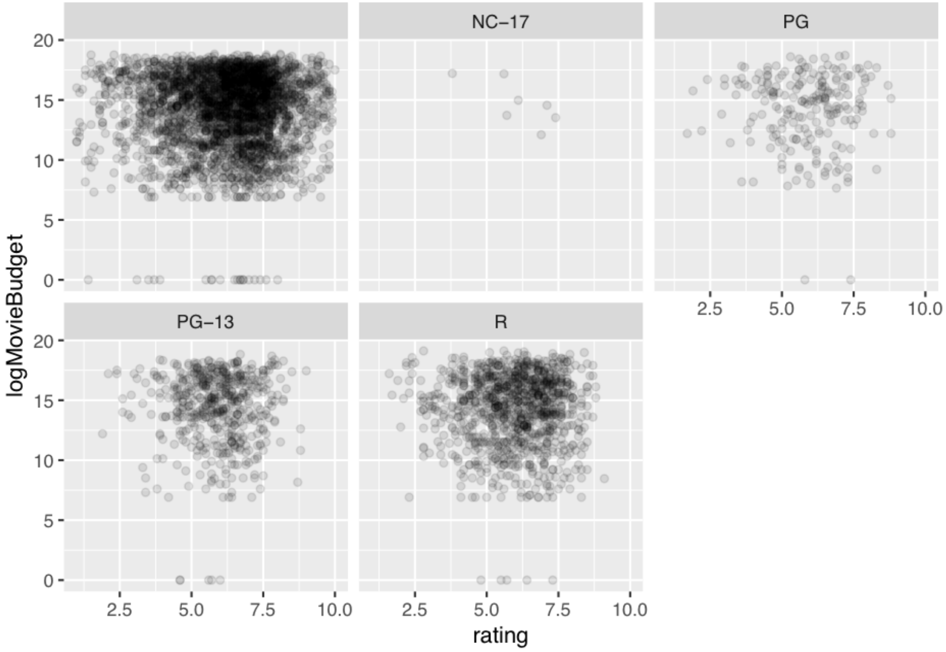
```
ggplot(moviesSub, aes(rating, logMovieBudget)) + geom_point(alpha = 0.1)
```



There appears to be some indication that some movies with very high budgets do get higher ratings but there are still exceptions to this. There are movies whos budgets are high but ratings are not. There are also a number of highly rated movies that have little to no budget at all. There isn't a clear correlation between the two.

j

```
ggplot(moviesSub, aes(rating, logMovieBudget)) + geom_point(alpha = 0.1) + facet_wrap(~mpaa )
```



#Facet wrap displays all of the movies by their MPAA ratings. Each graph looks at their budget and rating.

k

```
xtabs(~mpaa, data = moviesSub)
```

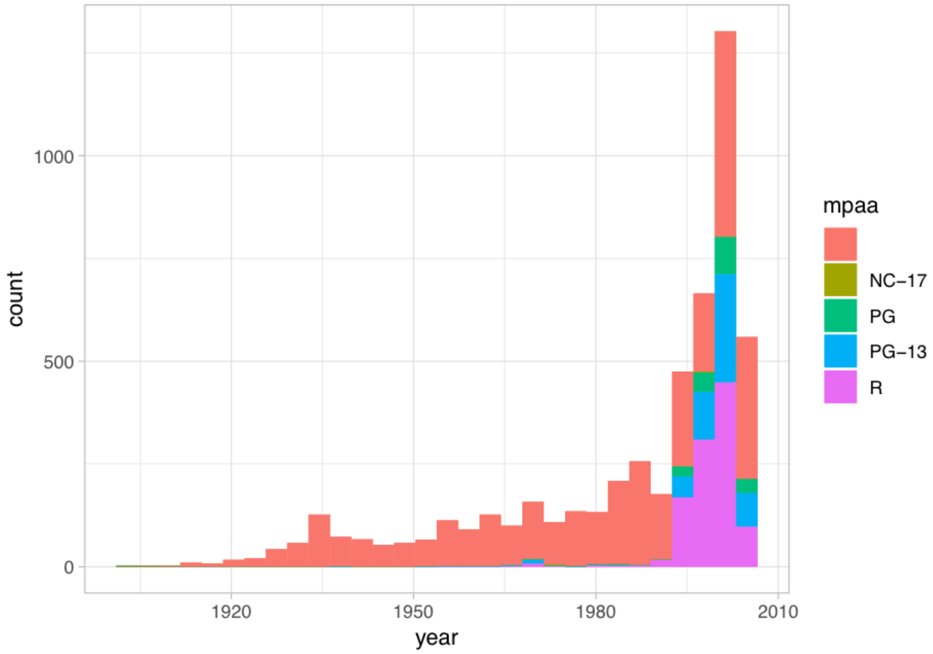
```
## mpaa
##         NC-17     PG PG-13      R
##   3402      7    212   530   1064
```

NA is the most popular movie rating.

l

```
ggplot(aes(year, fill = mpaa), data = moviesSub) + geom_histogram() + theme_light() + stat_bin(bins = 3(
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The reason why there are so many movies that are not rated is because many of these movies are not checked by the mpaa and go straight to market and are not viewed by the general public.

2

a.

```
moviesSub$mpaa <- as.factor(moviesSub$mpaa)
```

b.

```
contrasts(moviesSub$mpaa, contrasts = TRUE)
```

```
##       NC-17 PG PG-13 R
##           0  0     0 0
## NC-17     1  0     0 0
## PG        0  1     0 0
## PG-13     0  0     1 0
## R         0  0     0 1
```

```
contrasts(moviesSub$mpaa, contrasts = FALSE)
```

```
##         NC-17 PG PG-13 R
##       1     0  0     0 0
## NC-17 0     1  0     0 0
## PG    0     0  1     0 0
## PG-13 0     0  0     1 0
## R     0     0  0     0 1
```

NA

c.

```
linearFit <- lm(rating~ I(mpaa == "NC-17") + I(mpaa == "R")+ logMovieBudget + year + length + logMovieV
summary(linearFit)
```

```
##
## Call:
## lm(formula = rating ~ I(mpaa == "NC-17") + I(mpaa == "R") + logMovieBudget +
##     year + length + logMovieVotes, data = moviesSub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7003 -0.8219  0.1646  0.9193  4.5221
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             7.320e+00  2.027e+00   3.612 0.000307 ***
## I(mpaa == "NC-17")TRUE -9.627e-01  5.593e-01  -1.721 0.085266 .
## I(mpaa == "R")TRUE     -2.972e-01  5.558e-02  -5.347 9.34e-08 ***
## logMovieBudget         -2.198e-01  1.151e-02 -19.093  < 2e-16 ***
## year                    3.596e-05  1.027e-03   0.035 0.972074
## length                  3.723e-03  8.246e-04   4.515 6.48e-06 ***
## logMovieVotes           2.523e-01  1.233e-02  20.471  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.474 on 5208 degrees of freedom
## Multiple R-squared:  0.09279,    Adjusted R-squared:  0.09175
## F-statistic: 88.78 on 6 and 5208 DF,  p-value: < 2.2e-16
```

d.

```
linearFit2 <- lm(rating~ logMovieBudget + I(mpaa == "R") + I(mpaa == "NC-17") + Action + Documentary + (
summary(linearFit2)
```

```
##
## Call:
## lm(formula = rating ~ logMovieBudget + I(mpaa == "R") + I(mpaa ==
##     "NC-17") + Action + Documentary + Comedy + logMovieBudget +
##     year + length + logMovieVotes + I(mpaa == "NC-17"), data = moviesSub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5502 -0.7842  0.1782  0.8908  4.5774
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             7.219e+00  2.034e+00   3.549 0.000391 ***
## logMovieBudget         -2.053e-01  1.154e-02 -17.793  < 2e-16 ***
## I(mpaa == "R")TRUE     -3.131e-01  5.537e-02  -5.655 1.64e-08 ***
## I(mpaa == "NC-17")TRUE -9.845e-01  5.546e-01  -1.775 0.075933 .
## Action                 -3.994e-01  5.778e-02  -6.912 5.35e-12 ***
## Documentary             8.612e-01  1.342e-01   6.418 1.50e-10 ***
## Comedy                 -8.719e-02  4.484e-02  -1.944 0.051930 .
## year                    3.757e-06  1.032e-03   0.004 0.997095
## length                  3.231e-03  8.308e-04   3.889 0.000102 ***
## logMovieVotes           2.652e-01  1.233e-02  21.513  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.461 on 5205 degrees of freedom
## Multiple R-squared:  0.1092, Adjusted R-squared:  0.1077
## F-statistic: 70.89 on 9 and 5205 DF,  p-value: < 2.2e-16
```

e.

According to our model, no, having a higher budget does not result in a positive movie rating. In fact it hinders it. For every dollar increase in budget we have a -.2 decrease in our rating.

f.

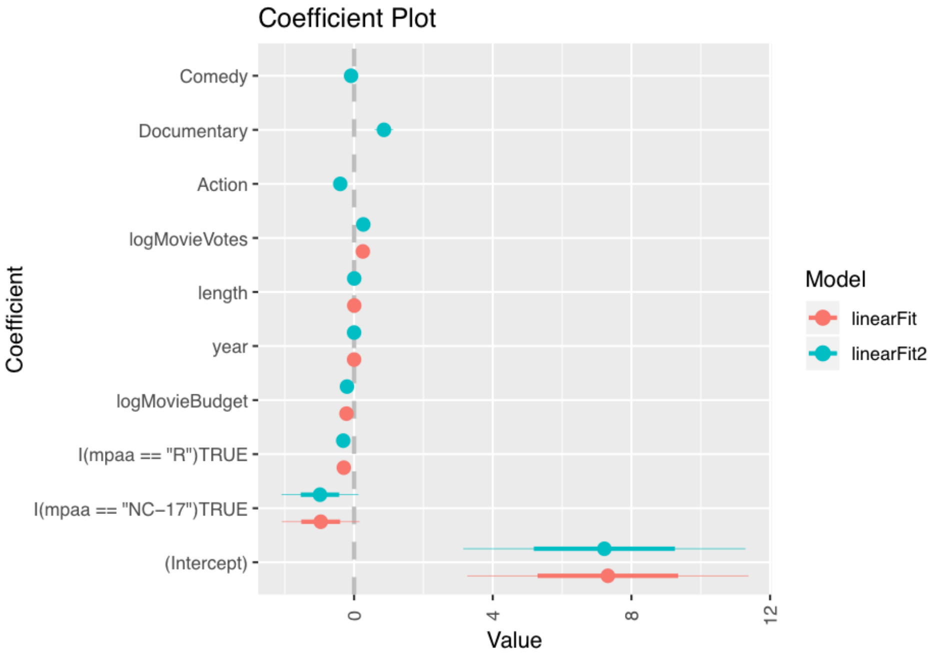
No, our negative coefficient would indicate that there is a negative impact on the rating. NC-17 has a -.9 effect on ratings and R has a -.3 effect on ratings.

g.

Documentaries receive higher ratings than non documentaries, that includes action and comedy. It get .86 higher than non documentaries.

h

```
library(coefplot)
multiplot(linearFit, linearFit2)
```



3.

a.

P(x) = the chances of landing on heads which is .5 1 - p(x) is the chance that it does not land on heads. .5/.5 = 1.

b.

p(x) is the chance of rolling a 1 in a six sided dice 1/6 = 0.166 1 - p(x) = 1 - .166 = .834 (0.166/1-0.166) (1/6)/(5/6) = 1/5 It is 5 time not likely that it will not land on 1.

c.

.9/.1 = 9. It is 9 time more like that it will not rain