DAS Group Project 2

Group 7

1 Introduction

Introduction paragraph

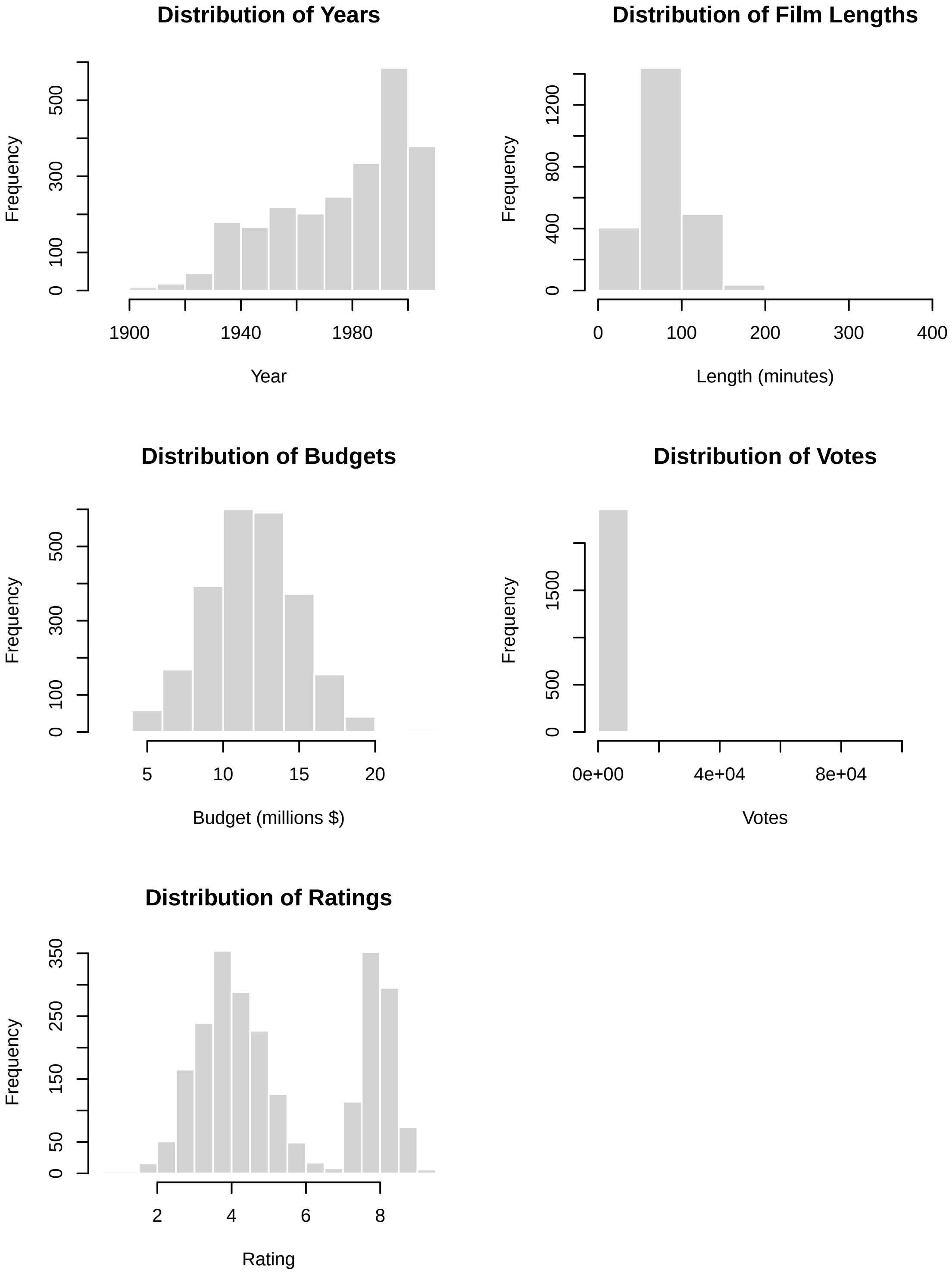
2 Exploratory Data Analysis

```
film_id    year  length  budget   votes   genre  rating 
      0       0      92       0       0       0       0 
```

```
Rows: 2,387
Columns: 8
$ film_id <int> 39891, 33810, 20282, 33131, 50633, 37020, 55337, 28037, 13291,~
$ year    <int> 2003, 2004, 1941, 1959, 1917, 1934, 2003, 1988, 1981, 1935, 19~
$ length  <int> 75, 120, 78, 106, 70, 64, 91, 101, 78, 7, 21, 90, 99, 101, 110~
$ budget  <dbl> 10.9, 19.6, 11.7, 12.0, 14.8, 11.6, 12.6, 10.1, 14.2, 6.6, 10.~
$ votes   <int> 17, 21, 14, 14, 9, 8, 182, 274, 61, 10, 5, 8, 349, 24, 20168, ~
$ genre   <chr> "Action", "Documentary", "Action", "Drama", "Drama", "Drama", ~
$ rating  <dbl> 4.4, 7.3, 2.7, 4.9, 5.6, 4.7, 4.4, 4.3, 4.3, 8.8, 7.3, 8.3, 7.~
$ above_7 <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1,~
```

```
    film_id           year          length           budget     
 Min.   :   33   Min.   :1894   Min.   :  1.00   Min.   : 2.10  
 1st Qu.:14799   1st Qu.:1958   1st Qu.: 74.00   1st Qu.:10.00  
 Median :30259   Median :1984   Median : 90.00   Median :12.00  
 Mean   :29942   Mean   :1977   Mean   : 81.75   Mean   :11.95  
 3rd Qu.:44670   3rd Qu.:1998   3rd Qu.:100.00   3rd Qu.:13.90  
 Max.   :58780   Max.   :2005   Max.   :399.00   Max.   :23.70  
     votes            genre               rating          above_7      
 Min.   :     5   Length:2387        Min.   :0.700   Min.   :0.0000  
```

```
1st Qu.:    12   Class :character   1st Qu.:3.700   1st Qu.:0.0000
Median :    32   Mode  :character   Median :4.700   Median :0.0000
Mean   :   659                      Mean   :5.414   Mean   :0.3523
3rd Qu.:   118                      3rd Qu.:7.800   3rd Qu.:1.0000
Max.   :103854                      Max.   :9.200   Max.   :1.0000
```

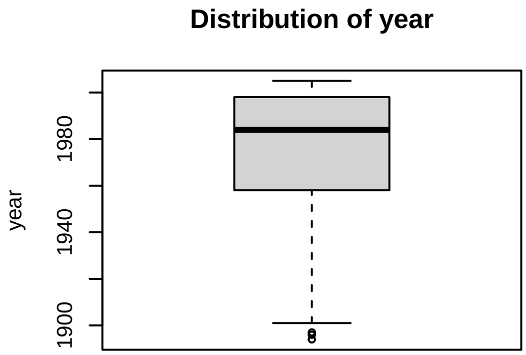

Distribution of Years

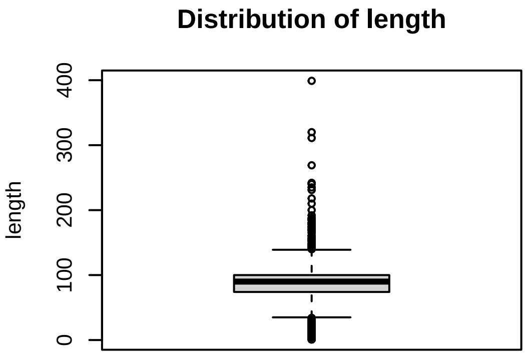
Distribution of Film Lengths

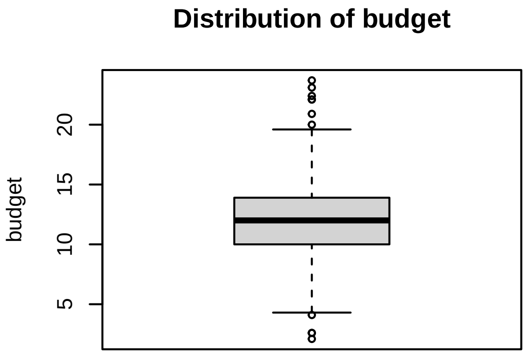
Distribution of Budgets

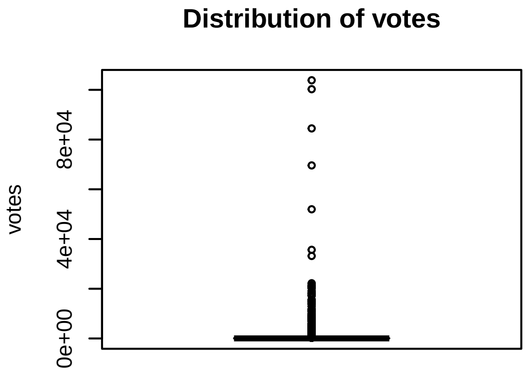
Distribution of Votes

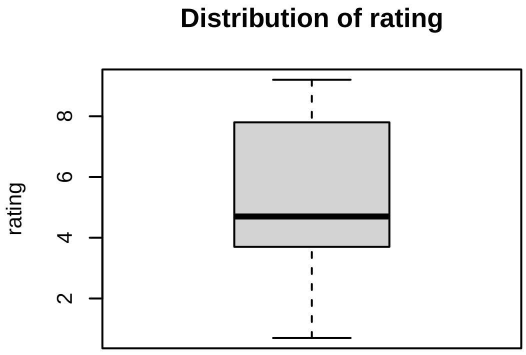
Distribution of Ratings

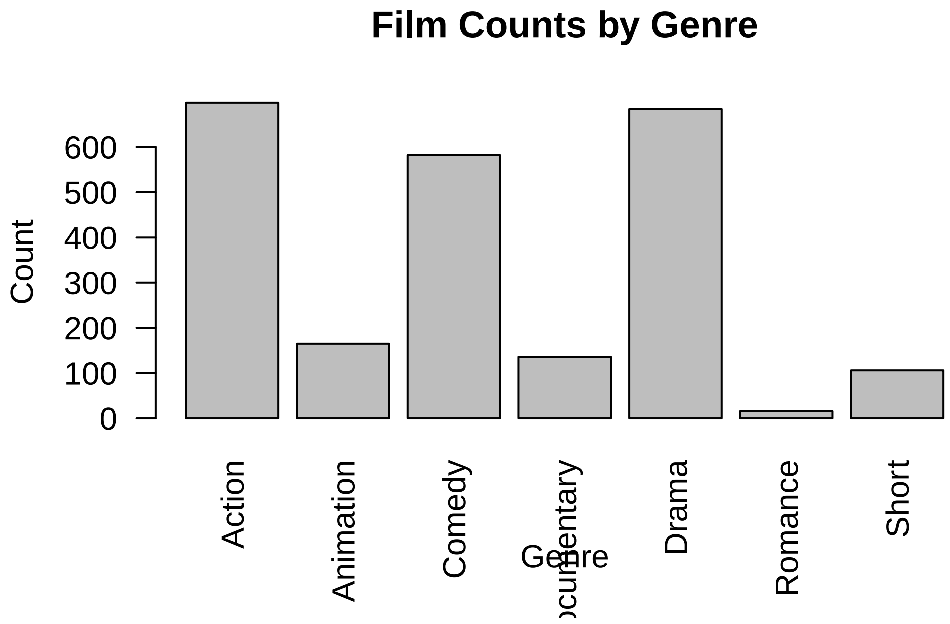

Distribution of year
year
1900
1940
1980

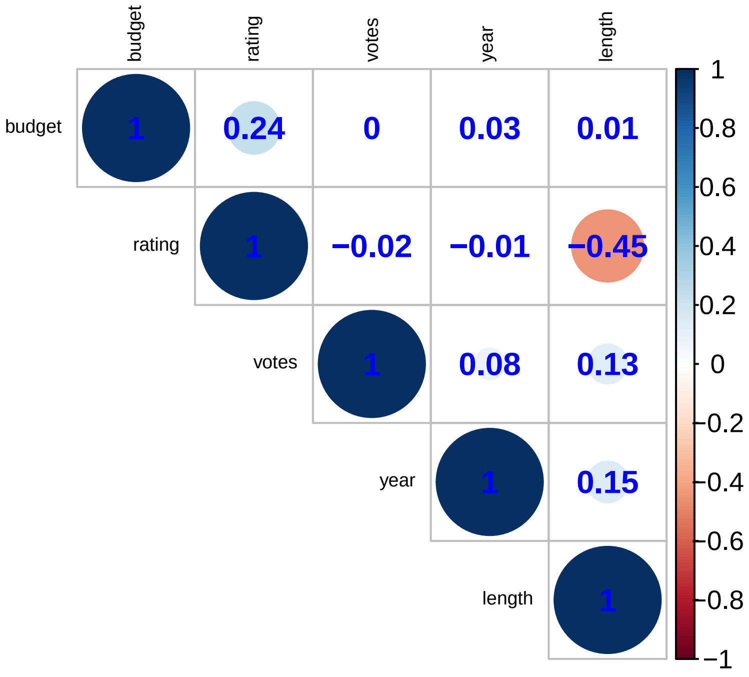

Distribution of length
length
0
100
200
300
400

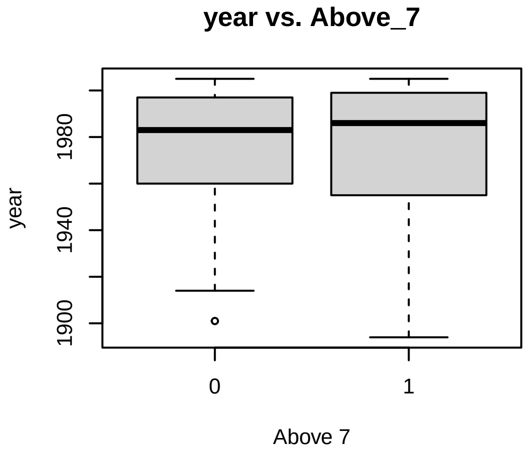

Distribution of budget
budget
5
10
15
20

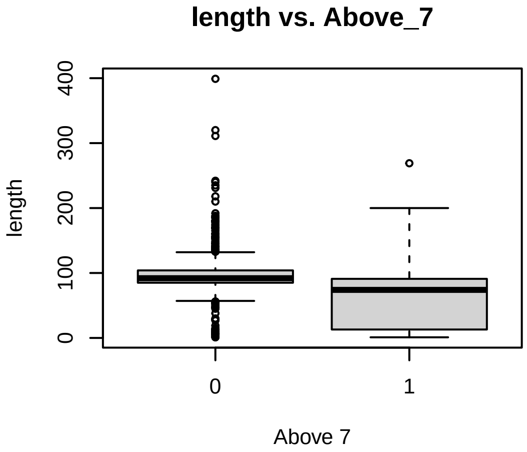

Distribution of votes
votes
0e+00
4e+04
8e+04

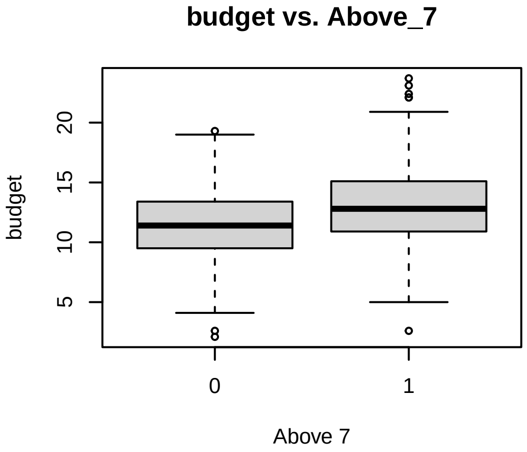

Distribution of rating
rating
2
4
6
8

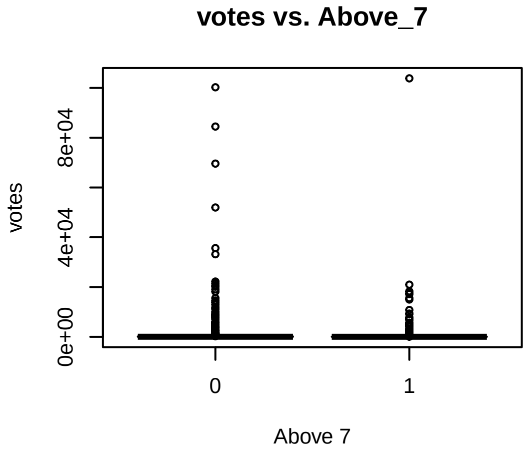

Film Counts by Genre
Count
600
500
400
300
200
100
0
Action
Animation
Comedy
ocumentary
Drama
Romance
Short
Genre

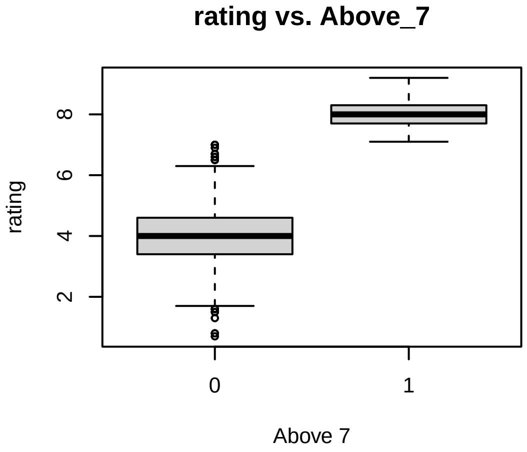

budget
rating
votes
year
length
budget
1
0.24
0
0.03
0.01
rating
1
−0.02
−0.01
−0.45
votes
1
0.08
0.13
year
1
0.15
length
1
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

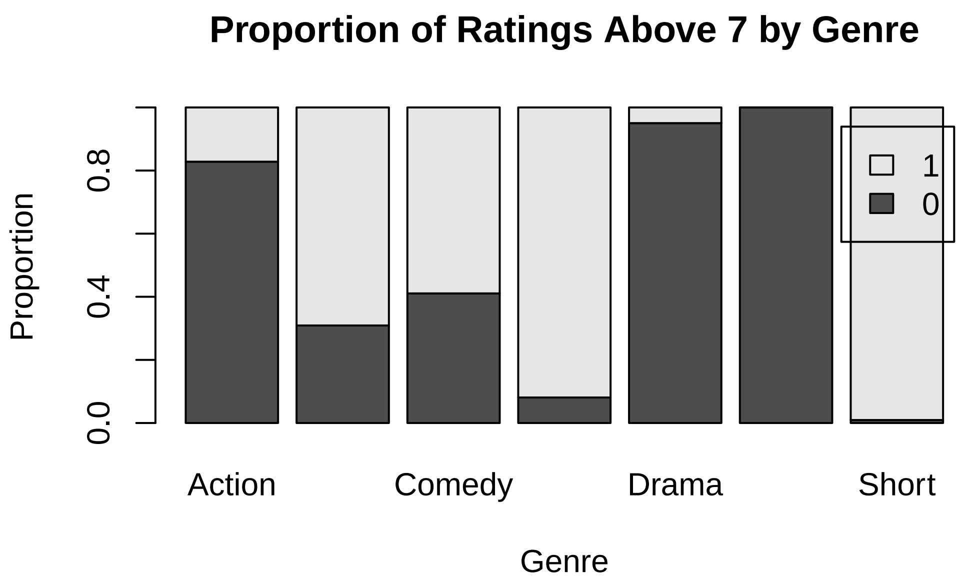

year vs. Above_7
year
1900
1940
1980
0
1
Above 7


length vs. Above_7
length
0
100
200
300
400
0
1
Above 7


budget vs. Above_7
budget
5
10
15
20
0
1
Above 7


votes vs. Above_7
votes
0e+00
4e+04
8e+04
0
1
Above 7

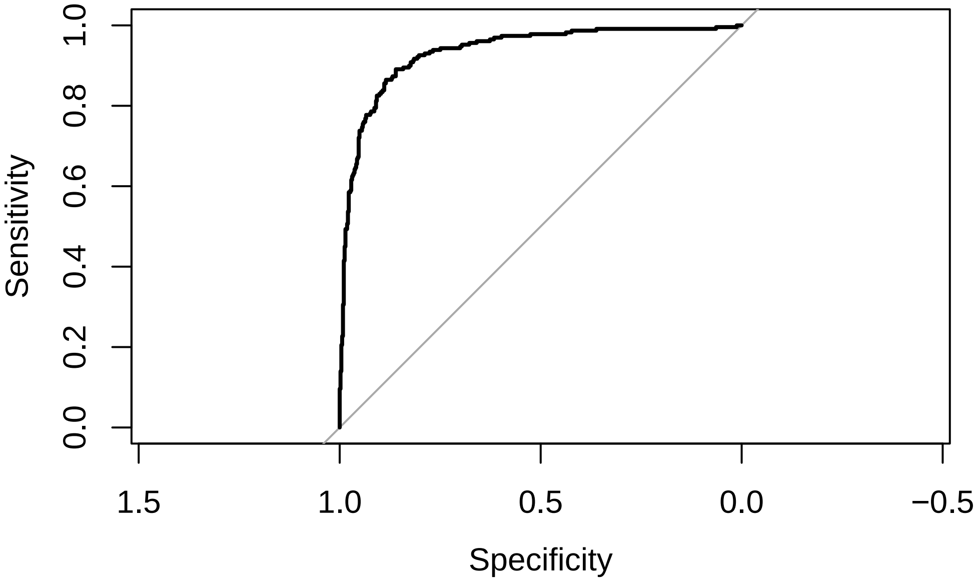

rating vs. Above_7
rating
2
4
6
8
0
1
Above 7



3 Formal Data Analysis

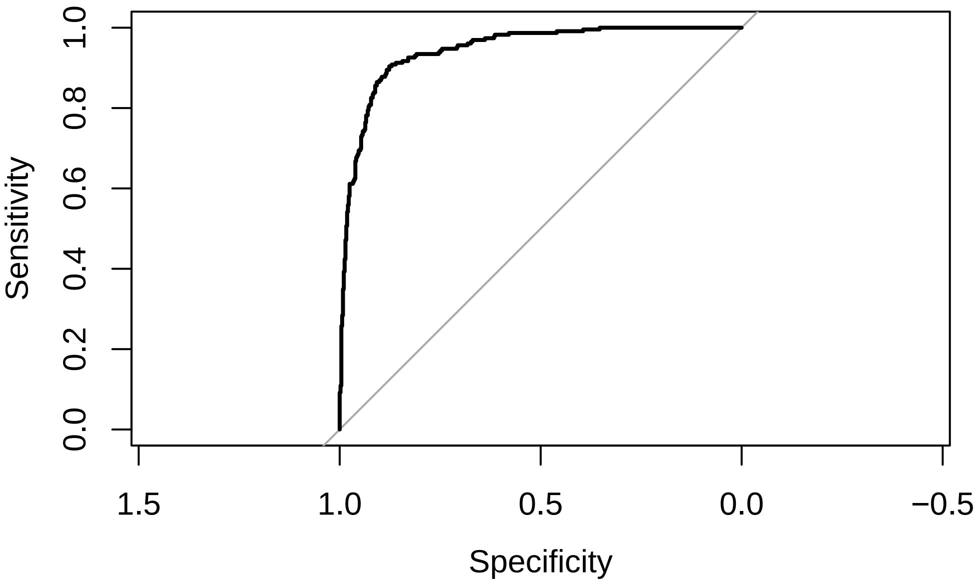
```
Call:
glm(formula = above_7 ~ year + length + budget + votes + genre,
    family = binomial, data = train_data)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -5.361e+00  7.280e+00  -0.736   0.4615
year              9.916e-04  3.716e-03   0.267   0.7896
length           -5.523e-02  4.400e-03 -12.554  < 2e-16 ***
budget            5.022e-01  3.799e-02  13.219  < 2e-16 ***
votes             4.861e-05  1.914e-05   2.540   0.0111 *
genreAnimation   -2.633e-01  4.018e-01  -0.655   0.5124
genreComedy       2.758e+00  2.178e-01  12.664  < 2e-16 ***
genreDocumentary  4.819e+00  4.713e-01  10.224  < 2e-16 ***
genreDrama       -1.906e+00  3.080e-01  -6.187 6.12e-10 ***
genreRomance     -1.664e+01  1.450e+03  -0.011   0.9908
genreShort        1.777e+01  6.322e+02   0.028   0.9776
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2195.45  on 1670  degrees of freedom
Residual deviance:  911.25  on 1660  degrees of freedom
AIC: 933.25

Number of Fisher Scoring iterations: 17


 Accuracy
0.8659218


Sensitivity
   0.862423


Specificity
  0.8733624
```



```
Area under the curve: 0.935
```

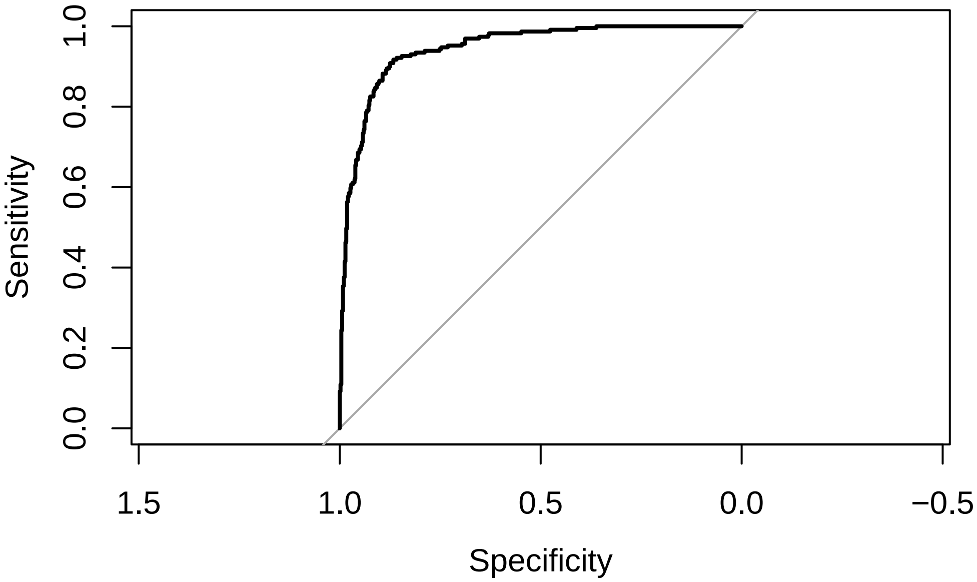
```
Call:
glm(formula = above_7 ~ year + length_log + budget + votes_log +
    genre, family = binomial, data = train_data)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.272e+01  7.643e+00   1.664  0.09604 .
year             -3.962e-03  3.915e-03  -1.012  0.31159
length_log       -3.122e+00  2.804e-01 -11.135  < 2e-16 ***
budget            5.226e-01  3.932e-02  13.289  < 2e-16 ***
votes_log         1.463e-01  5.001e-02   2.925  0.00345 **
genreAnimation   -2.798e+00  6.652e-01  -4.207 2.59e-05 ***
genreComedy       2.608e+00  2.173e-01  12.001  < 2e-16 ***
genreDocumentary  4.790e+00  4.551e-01  10.524  < 2e-16 ***
genreDrama       -2.290e+00  3.459e-01  -6.621 3.57e-11 ***
genreRomance     -1.697e+01  1.479e+03  -0.011  0.99085
genreShort        1.742e+01  5.672e+02   0.031  0.97549
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2195.45  on 1670  degrees of freedom
Residual deviance:  875.22  on 1660  degrees of freedom
AIC: 897.22

Number of Fisher Scoring iterations: 17

 Accuracy
0.8868715

Sensitivity
  0.8911704

Specificity
  0.8777293
```



```
Area under the curve: 0.9451
```

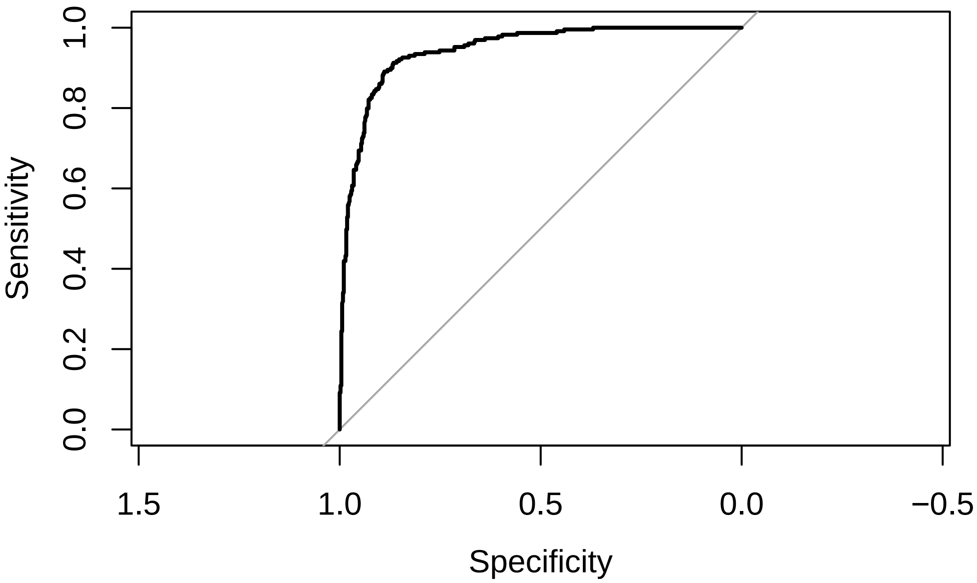
```
Call:
glm(formula = above_7 ~ length_log + budget + votes_log + genre,
    family = binomial, data = train_data)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         5.08479    1.16889   4.350 1.36e-05 ***
length_log         -3.14990    0.28168 -11.183  < 2e-16 ***
budget              0.52035    0.03921  13.272  < 2e-16 ***
votes_log           0.13636    0.04896   2.785  0.00535 **
genreAnimation     -2.82857    0.66752  -4.237 2.26e-05 ***
genreComedy         2.60473    0.21686  12.011  < 2e-16 ***
genreDocumentary    4.72592    0.44982  10.506  < 2e-16 ***
genreDrama         -2.29767    0.34875  -6.588 4.45e-11 ***
genreRomance      -16.87365 1494.07345  -0.011  0.99099
genreShort         17.31928  566.97080   0.031  0.97563
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2195.45  on 1670  degrees of freedom
Residual deviance:  876.25  on 1661  degrees of freedom
AIC: 896.25

Number of Fisher Scoring iterations: 17

 Accuracy
0.8854749

Sensitivity
  0.8870637

Specificity
  0.8820961
```



```
Area under the curve: 0.9457


Call:
```

```
glm(formula = above_7 ~ length_log + budget + genre, family = binomial,
    data = train_data)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         4.8250     1.1248   4.290 1.79e-05 ***
length_log         -2.9618     0.2621 -11.298  < 2e-16 ***
budget              0.5166     0.0389  13.281  < 2e-16 ***
genreAnimation     -2.5261     0.6343  -3.982 6.82e-05 ***
genreComedy         2.6824     0.2156  12.441  < 2e-16 ***
genreDocumentary    4.6662     0.4487  10.400  < 2e-16 ***
genreDrama         -2.2495     0.3442  -6.536 6.33e-11 ***
genreRomance      -16.7858  1507.9078  -0.011    0.991
genreShort         17.2922   574.9628   0.030    0.976
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2195.45  on 1670  degrees of freedom
Residual deviance:  884.03  on 1662  degrees of freedom
AIC: 902.03

Number of Fisher Scoring iterations: 17

 Accuracy
0.8882682

Sensitivity
  0.8870637

Specificity
  0.8908297
```



```
Area under the curve: 0.945
```

4 Conclusions

5 Reference