DATA621-FinalProject-SmoothOperators

Rob Hodde, Matt Farris, Jeffrey Burmood, Bin Lin

5/11/2017

Problem Description

Our final project will explore, analyze and model a data set containing information on approximately 5,000 movies. The dataset contains movie data extracted from the IMDB website and is available on Kaggle.com.

The project will develop predictive models for two questions:

1) Will the movie make money, lose money, or break even (approximately)?
2) What is the anticipated gross margin (profit) for the movie?

Data Exploration

Data Exploration

VAR	TYPE
duration	integer
director_facebook_likes	integer
actor_3_facebook_likes	integer
actor_1_facebook_likes	integer
gross	integer
movie_title	character
num_voted_users	integer
cast_total_facebook_likes	integer
facenumber_in_poster	integer
content_rating	character
budget	double
actor_2_facebook_likes	integer
imdb_score	double

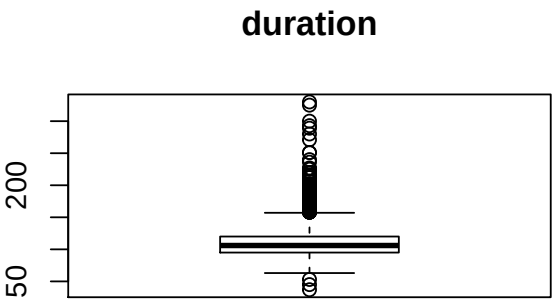
```
##     duration   director_facebook_likes actor_3_facebook_likes
##  Min.   : 37   Min.   :    0.0         Min.   :    0.0
##  1st Qu.: 95   1st Qu.:   10.0         1st Qu.:  188.8
##  Median :106   Median :   60.0         Median :  433.0
##  Mean   :110   Mean   :  792.9         Mean   :  761.8
##  3rd Qu.:120   3rd Qu.:  232.5         3rd Qu.:  690.0
##  Max.   :330   Max.   :23000.0         Max.   :23000.0
##
```

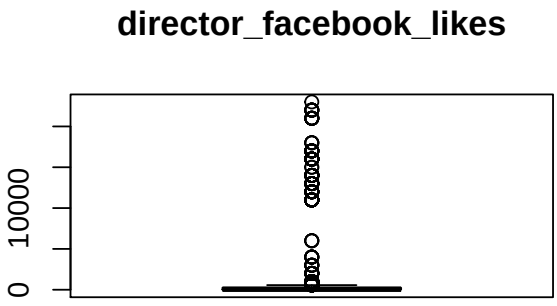
```
##  actor_1_facebook_likes     gross           movie_title
##  Min.   :     0.0       Min.   :      162   Length:3828
##  1st Qu.:   737.5       1st Qu.:  7452337   Class :character
##  Median :  1000.0       Median : 28854152   Mode  :character
##  Mean   :  7664.1       Mean   : 51694432
##  3rd Qu.: 12250.0       3rd Qu.: 66004138
##  Max.   :640000.0       Max.   :760505847
##
##  num_voted_users   cast_total_facebook_likes facenumber_in_poster
##  Min.   :     22   Min.   :     0            Min.   : 0.000
##  1st Qu.:  18267   1st Qu.:  1880            1st Qu.: 0.000
##  Median :  52380   Median :  3962            Median : 1.000
##  Mean   : 103908   Mean   : 11396            Mean   : 1.379
##  3rd Qu.: 125642   3rd Qu.: 16128            3rd Qu.: 2.000
##  Max.   :1689764   Max.   :656730            Max.   :43.000
##
##    content_rating     budget          actor_2_facebook_likes
##  R        :1736   Min.   :2.180e+02   Min.   :     0.0
##  PG-13    :1326   1st Qu.:1.000e+07   1st Qu.:   373.8
##  PG       : 574   Median :2.500e+07   Median :   677.0
##  G        :  89   Mean   :4.548e+07   Mean   :  1994.6
##  Not Rated:  40   3rd Qu.:5.000e+07   3rd Qu.:   975.0
##  Unrated  :  24   Max.   :1.222e+10   Max.   :137000.0
##  (Other)  :  39
##    imdb_score
##  Min.   :1.600
##  1st Qu.:5.900
##  Median :6.600
##  Mean   :6.459
##  3rd Qu.:7.200
##  Max.   :9.300
##
```

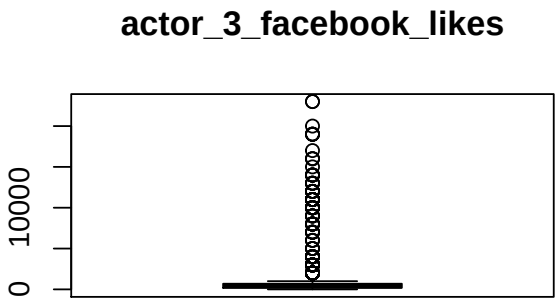
	duration	director_facebook_likes	actor_3_facebook_likes	actor_1_facebook_likes
duration	1.0000000	0.1822411	0.1279962	0.0863409
director_facebook_likes	0.1822411	1.0000000	0.1184843	0.0905543
actor_3_facebook_likes	0.1279962	0.1184843	1.0000000	0.2526590
actor_1_facebook_likes	0.0863409	0.0905543	0.2526590	1.0000000
num_voted_users	0.3434487	0.3013255	0.2697667	0.1817812
cast_total_facebook_likes	0.1232351	0.1197195	0.4895509	0.9450371
facenumber_in_poster	0.0263907	-0.0478417	0.1055483	0.0614101
budget	0.0696018	0.0189881	0.0408678	0.0173849
actor_2_facebook_likes	0.1311685	0.1172937	0.5540722	0.3910139
imdb_score	0.3655775	0.1915761	0.0661996	0.0939598

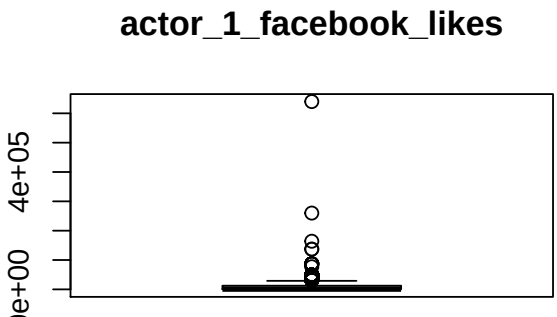
	num_voted_users	cast_total_facebook_likes	facenumber_in_poster	budget
duration	0.3434487	0.1232351	0.0263907	0.0696018
director_facebook_likes	0.3013255	0.1197195	-0.0478417	0.0189881
actor_3_facebook_likes	0.2697667	0.4895509	0.1055483	0.0408678
actor_1_facebook_likes	0.1817812	0.9450371	0.0614101	0.0173849
num_voted_users	1.0000000	0.2516946	-0.0324633	0.0678793
cast_total_facebook_likes	0.2516946	1.0000000	0.0837393	0.0298442
facenumber_in_poster	-0.0324633	0.0837393	1.0000000	-0.0215767
budget	0.0678793	0.0298442	-0.0215767	1.0000000
actor_2_facebook_likes	0.2473172	0.6424574	0.0720087	0.0367048
imdb_score	0.4792715	0.1073363	-0.0671658	0.0298854

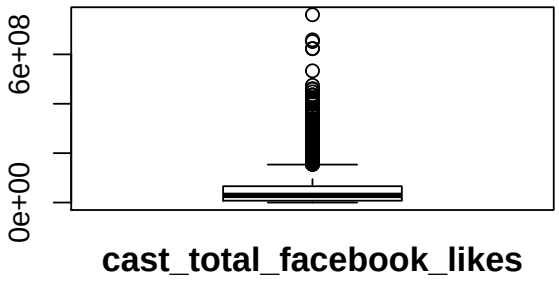
	actor_2_facebook_likes	imdb_score
duration	0.1311685	0.3655775
director_facebook_likes	0.1172937	0.1915761
actor_3_facebook_likes	0.5540722	0.0661996
actor_1_facebook_likes	0.3910139	0.0939598
num_voted_users	0.2473172	0.4792715
cast_total_facebook_likes	0.6424574	0.1073363
facenumber_in_poster	0.0720087	-0.0671658
budget	0.0367048	0.0298854
actor_2_facebook_likes	1.0000000	0.1031776
imdb_score	0.1031776	1.0000000

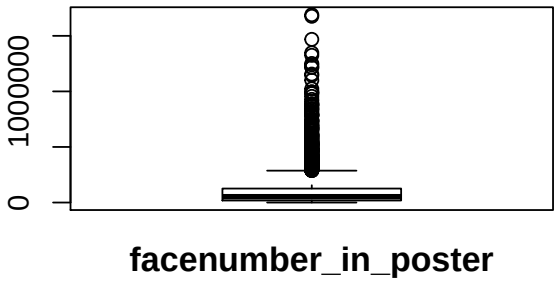


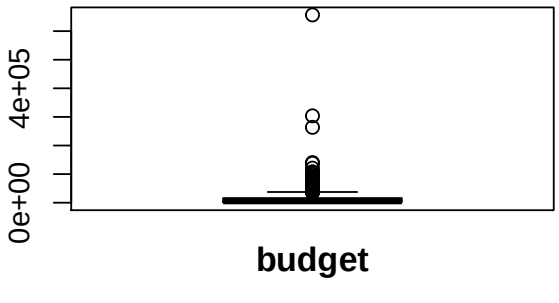


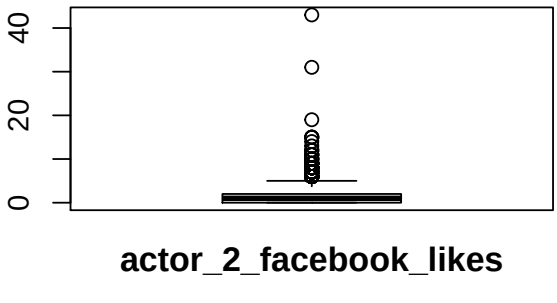

actor_3_facebook_likes
0
10000

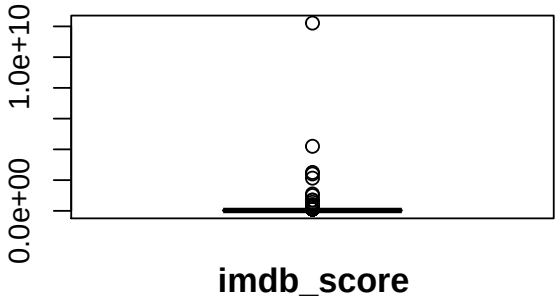

actor_1_facebook_likes
0e+00
4e+05

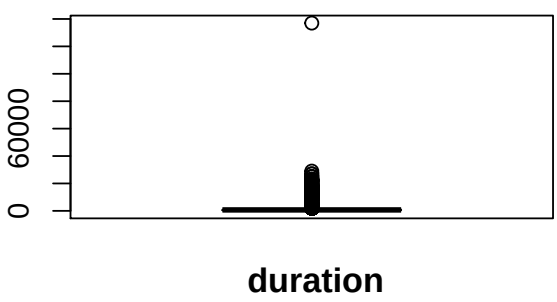

gross
0e+00
6e+08

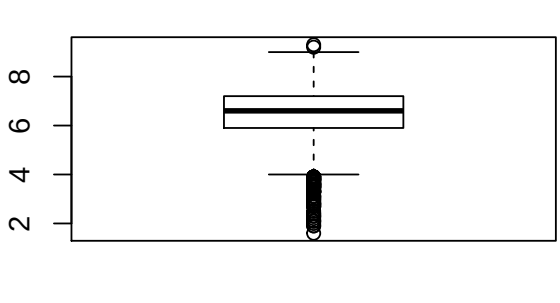

num_voted_users
0
1000000

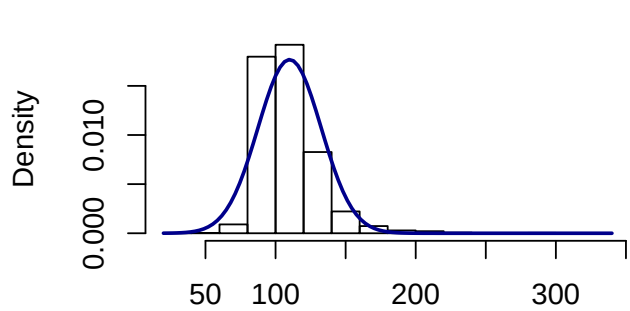

cast_total_facebook_likes
0e+00
4e+05

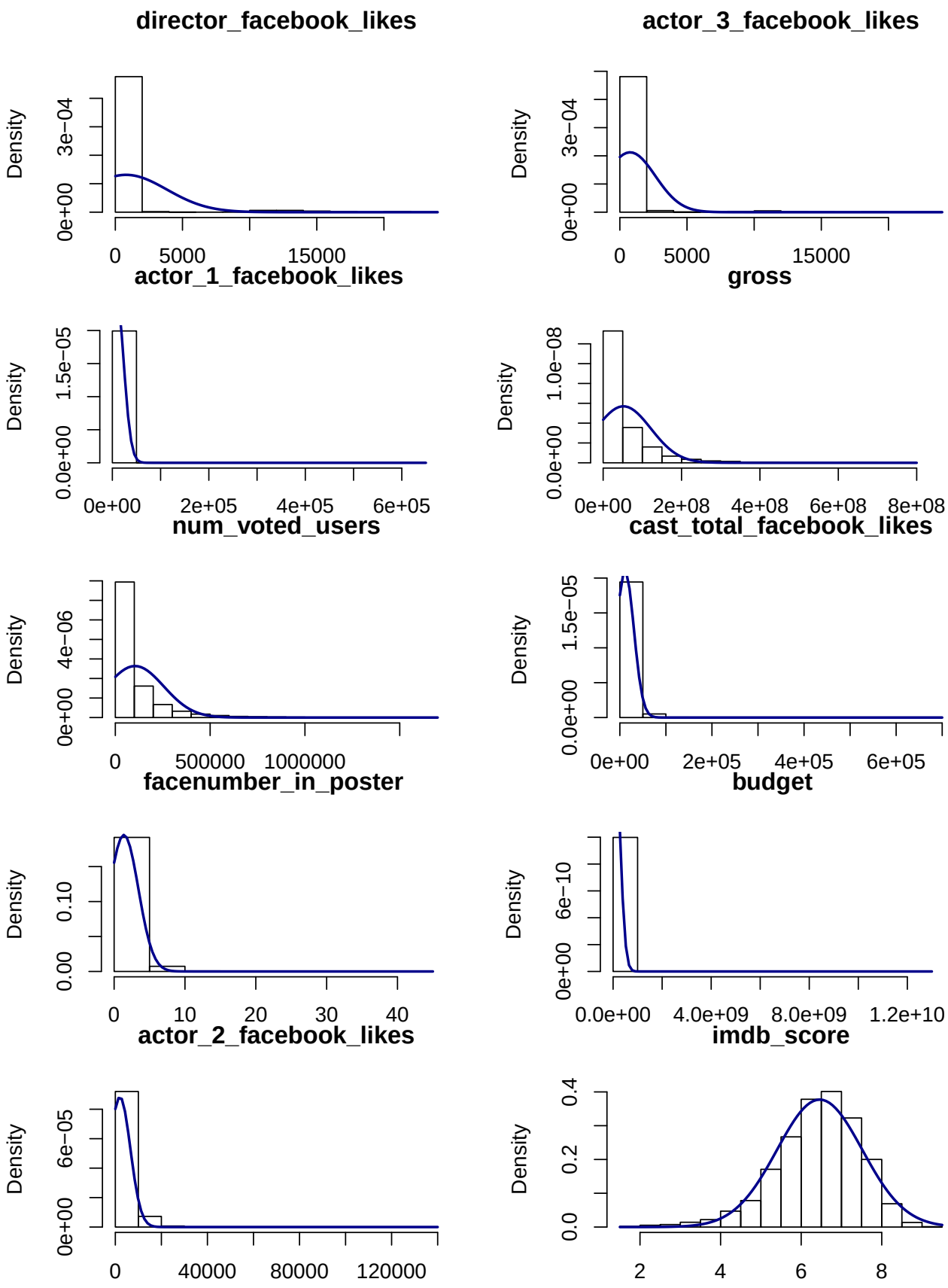

facenumber_in_poster
0
20
40

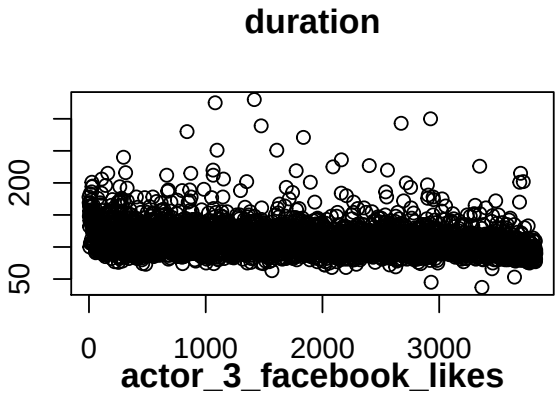

budget
0.0e+00
1.0e+10

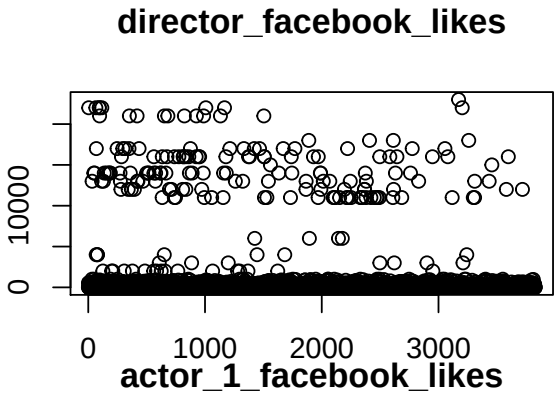

actor_2_facebook_likes
0
60000

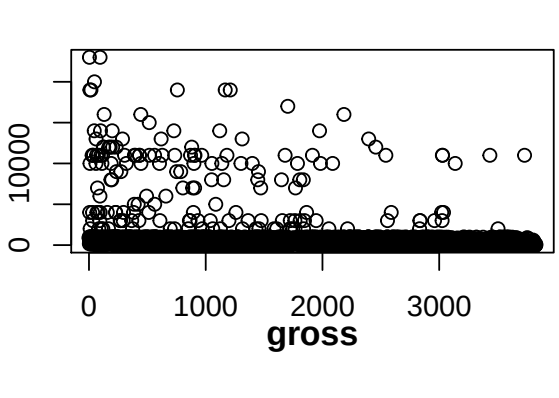

imdb_score
2
4
6
8

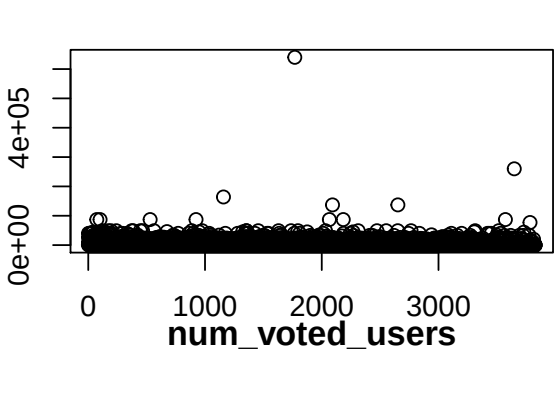

duration
Density
0.000
0.010
50
100
200
300

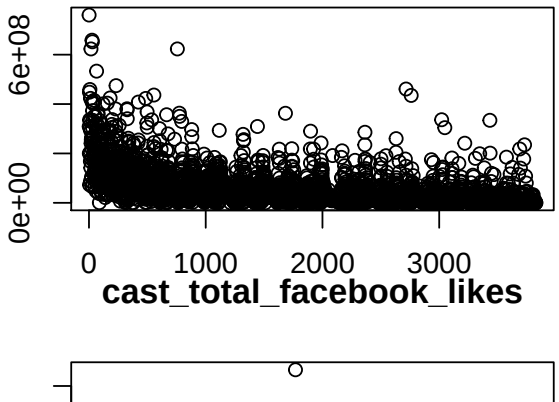

director_facebook_likes
Density
0e+00
3e−04
0
5000
15000
actor_1_facebook_likes
actor_3_facebook_likes
Density
0e+00
3e−04
0
5000
15000
gross
Density
0.0e+00
1.5e−05
0e+00
2e+05
4e+05
6e+05
num_voted_users
Density
0.0e+00
1.0e−08
0e+00
2e+08
4e+08
6e+08
8e+08
cast_total_facebook_likes
Density
0e+00
4e−06
0
500000
1000000
facenumber_in_poster
Density
0.0e+00
1.5e−05
0e+00
2e+05
4e+05
6e+05
budget
Density
0.00
0.10
0
10
20
30
40
actor_2_facebook_likes
Density
0e+00
6e−10
0.0e+00
4.0e+09
8.0e+09
1.2e+10
imdb_score
Density
0e+00
6e−05
0
40000
80000
120000
Density
0.0
0.2
0.4
2
4
6
8

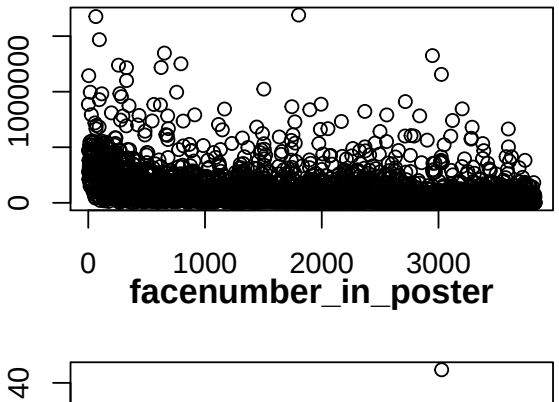

duration
50
200
0
1000
2000
3000

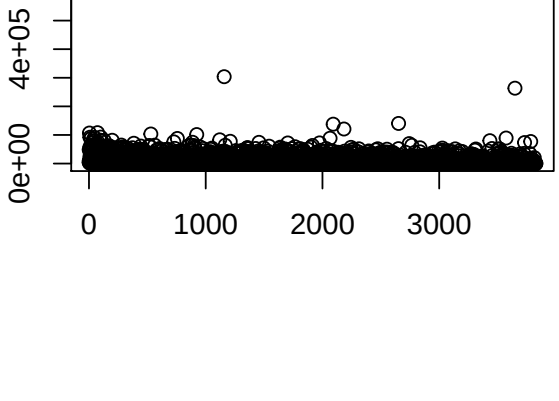

director_facebook_likes
0
10000
0
1000
2000
3000

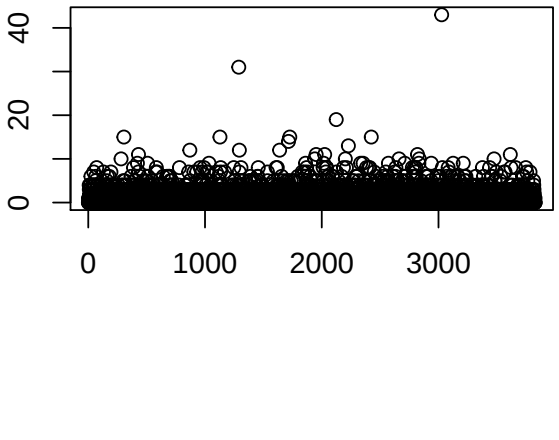

actor_3_facebook_likes
0
10000
0
1000
2000
3000

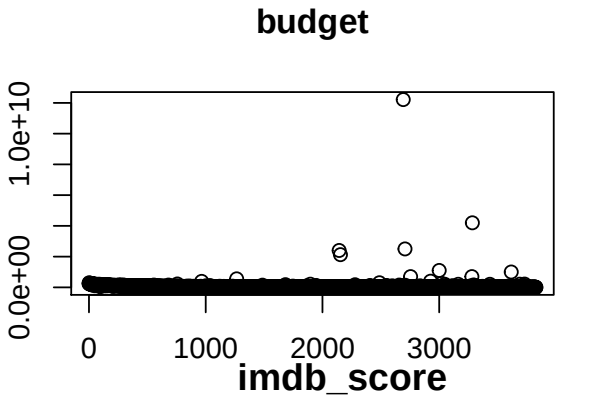

actor_1_facebook_likes
0e+00
4e+05
0
1000
2000
3000

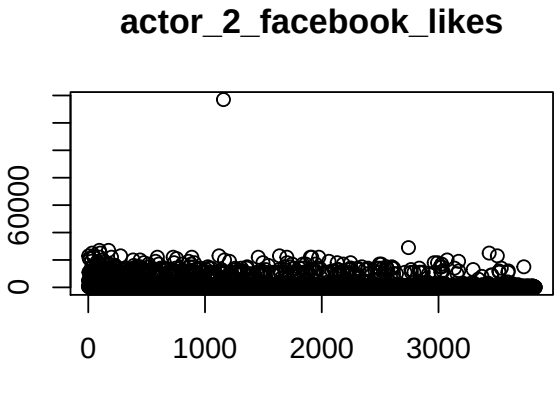

gross
0e+00
6e+08
0
1000
2000
3000

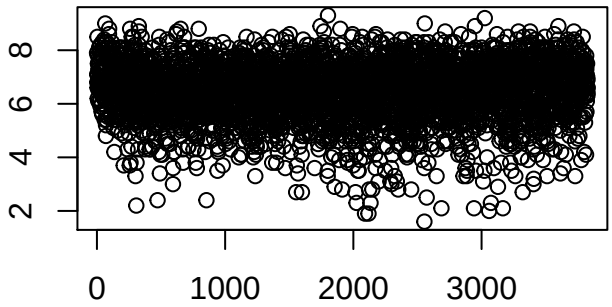

num_voted_users
0
1000000
0
1000
2000
3000


cast_total_facebook_likes
0e+00
4e+05
0
1000
2000
3000


facenumber_in_poster
0
20
40
0
1000
2000
3000







After exploring the data, we noticed there is a scattering of NAs across the variables. Due to the relatively low number of total NAs, we choose to remove all rows with NAs, leaving 3,828 rows of data.

Next, the content_rating variable is converted to a factor so the rating categories can be used with the regression models.

Data Preparation

Data Preparation

Build Models

Build Models

Smooth Operators - All Done!