Netflix and IMDB
USER RECOMMENDATIONS

BY SARAH KIM, PATRICK LIU,
MOE SUNAMI, JULIA DOU

PROBLEM STATEMENT

- ***Which movie should I show an individual user?***
  - *Based on what genres a user watched*
  - *Based on user's ratings of movies*
- ***What kind of movies should Netflix commission next?***

table of CONTENTS

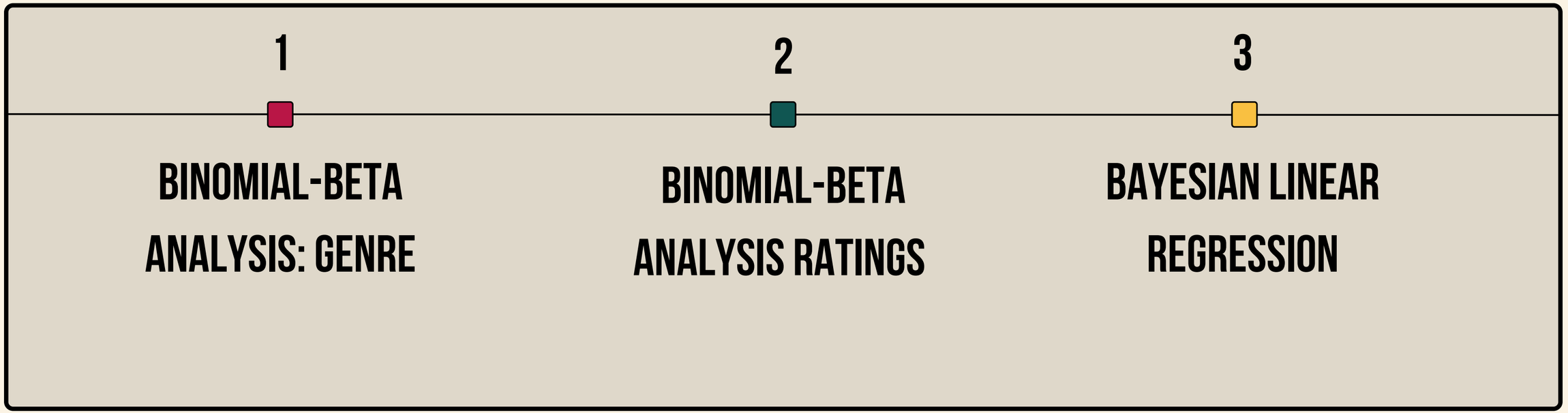
THE DATA:

NETFLIX DATASET

IMDB DATASET

CONSOLIDATED DATASET

ANALYSES


1
2
3
BINOMIAL-BETA ANALYSIS: GENRE
BINOMIAL-BETA ANALYSIS RATINGS
BAYESIAN LINEAR REGRESSION

1. BINOMIAL-BETA ANALYSIS: GENRE

1. WHAT SHOULD NETFLIX SHOW AS ONE OF THE FIRST RECOMMENDED MOVIES TO OUR USER JANE GIVEN THE GENRERAL TREND OF USER PREFERENCE FOR FILM GENRE?
2. WHAT % OF THE TIME DOES A USER WATCH A CERTAIN GENRE?

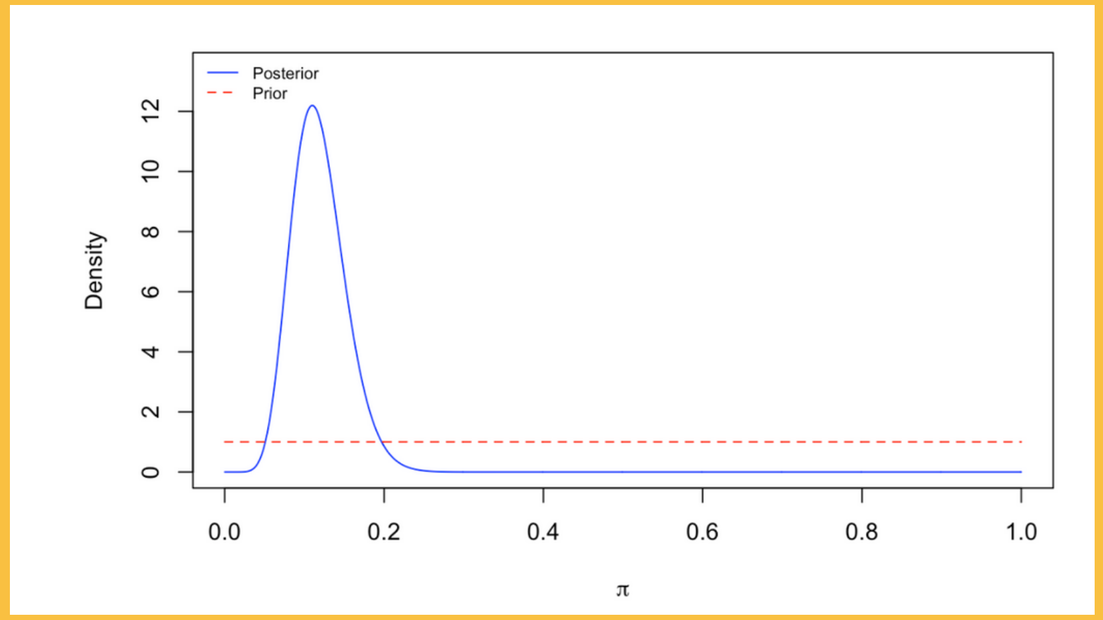
Jane!

PRELIMINARY ANALYSIS

- Action, Adventure
- Action, Adventure, Fantasy
- Action, Adventure, Horror
- Action, Adventure, Sci–Fi
- Action, Adventure, Thriller
- Action, Adventure, War
- Action, Biography, Crime
- Action, Crime, Drama
- Action, Crime, Thriller
- Action, Drama
- Action, Drama, History
- Action, Drama, Mystery
- Action, Mystery, Thriller
- Adventure, Comedy, Drama
- Adventure, Comedy, Film–Noir
- Adventure, Mystery, Thriller
- Adventure, Sci–Fi
- Adventure, Thriller
- Animation, Action, Crime
- Biography, Comedy, Drama
- Biography, Drama
- Biography, Drama, Family
- Biography, Drama, History
- Biography, Drama, Music
- Biography, Drama, Sport
- Comedy
- Comedy, Crime
- Comedy, Crime, Drama
- Comedy, Drama
- Comedy, Drama, Family
- Comedy, Drama, Fantasy
- Comedy, Drama, Romance
- Comedy, Family, Romance
- Comedy, Music
- Comedy, Music, Musical
- Comedy, Musical, Romance
- Comedy, Mystery, Romance
- Crime, Drama
- Crime, Drama, History
- Crime, Drama, Thriller
- Crime, Thriller
- Drama
- Drama, Family, Fantasy
- Drama, Film–Noir, Romance
- Drama, Horror
- Drama, Horror, Sci–Fi
- Drama, Mystery, Sci–Fi
- Drama, Mystery, Thriller
- Drama, Romance
- Drama, Romance, Sci–Fi
- Drama, Romance, War
- Drama, Sport
- Drama, War
- Drama, Western
- Film–Noir, Mystery
- Film–Noir, Mystery, Thriller
- Horror, Mystery, Thriller

METHODS:

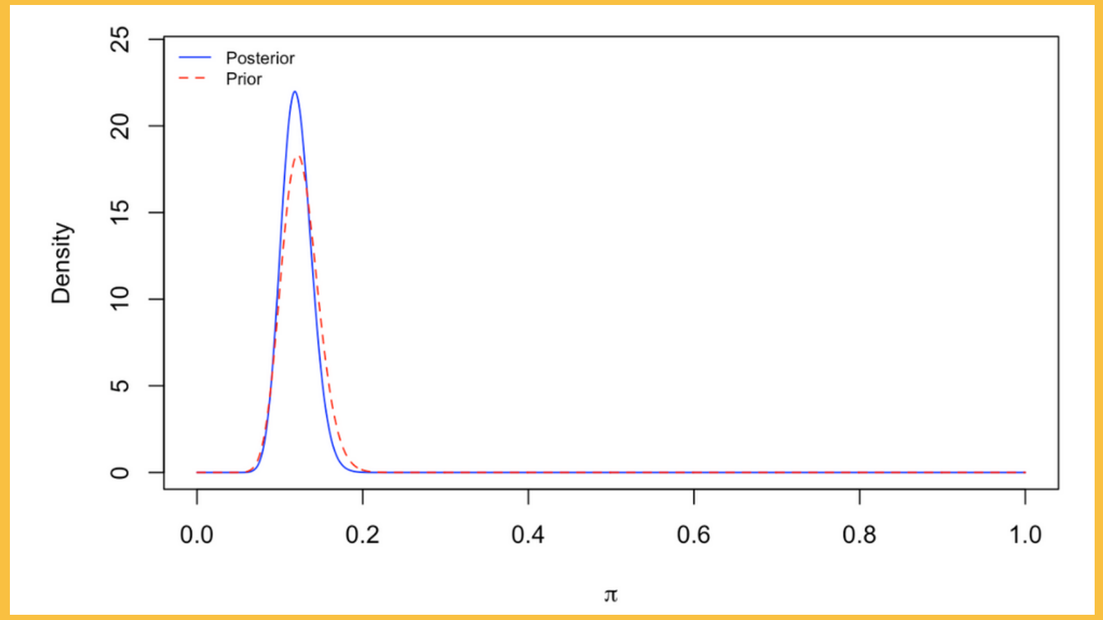
1. uniform prior
2. calculate beta parameters
3. choose a beta distribution
4. posterior results



beta(1, 1)

METHODS:

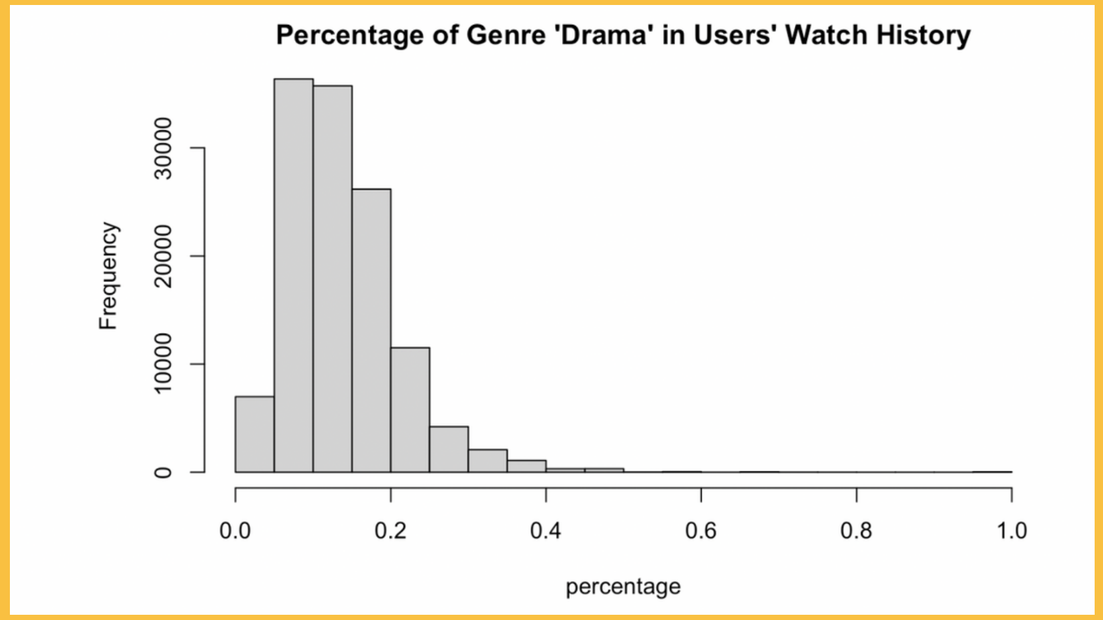
1. uniform prior
2. calculate beta parameters
3. choose a beta distribution
4. posterior results



beta(28.3, 198.53)

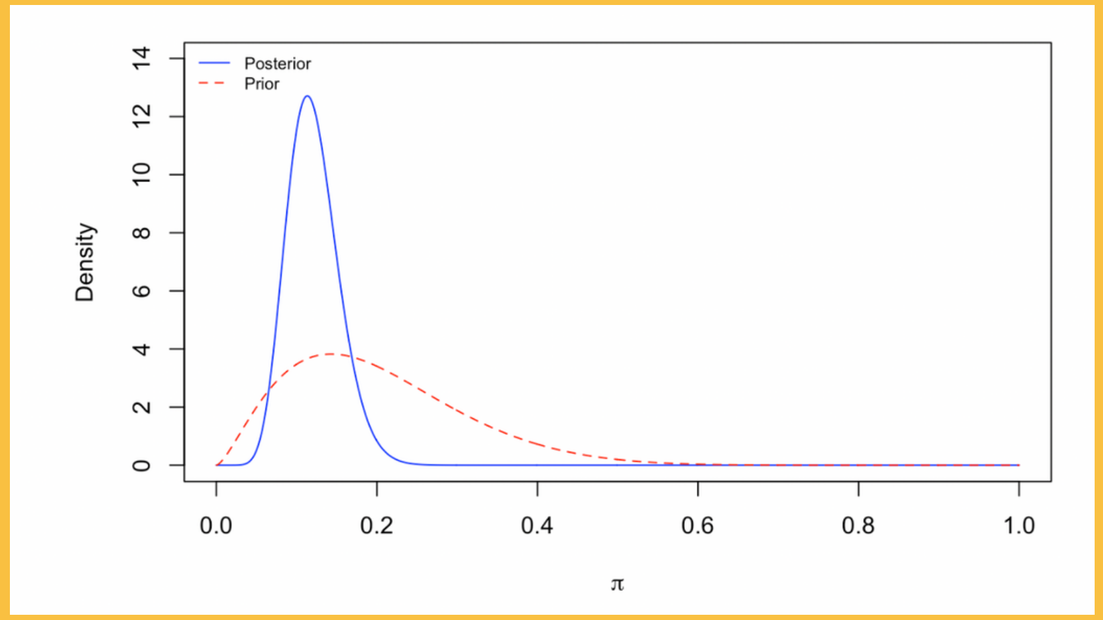
RESULTS:

1. uniform prior
2. calculate beta parameters
3. choose a beta distribution
4. posterior results



RESULTS:

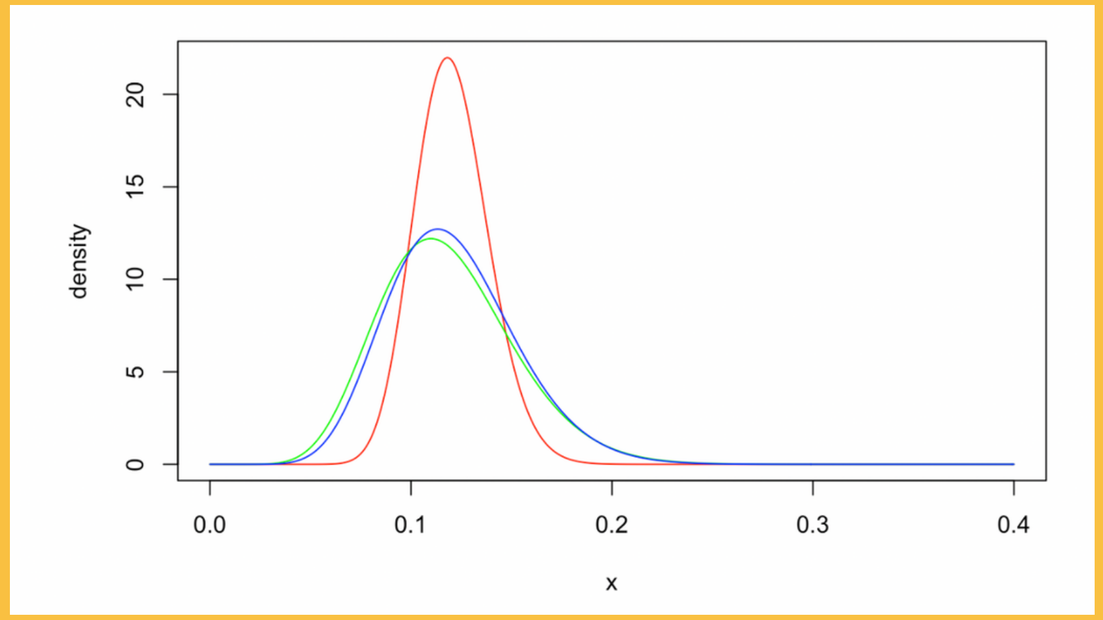
1. uniform prior
2. calculate beta parameters
3. choose a beta distribution
4. posterior results



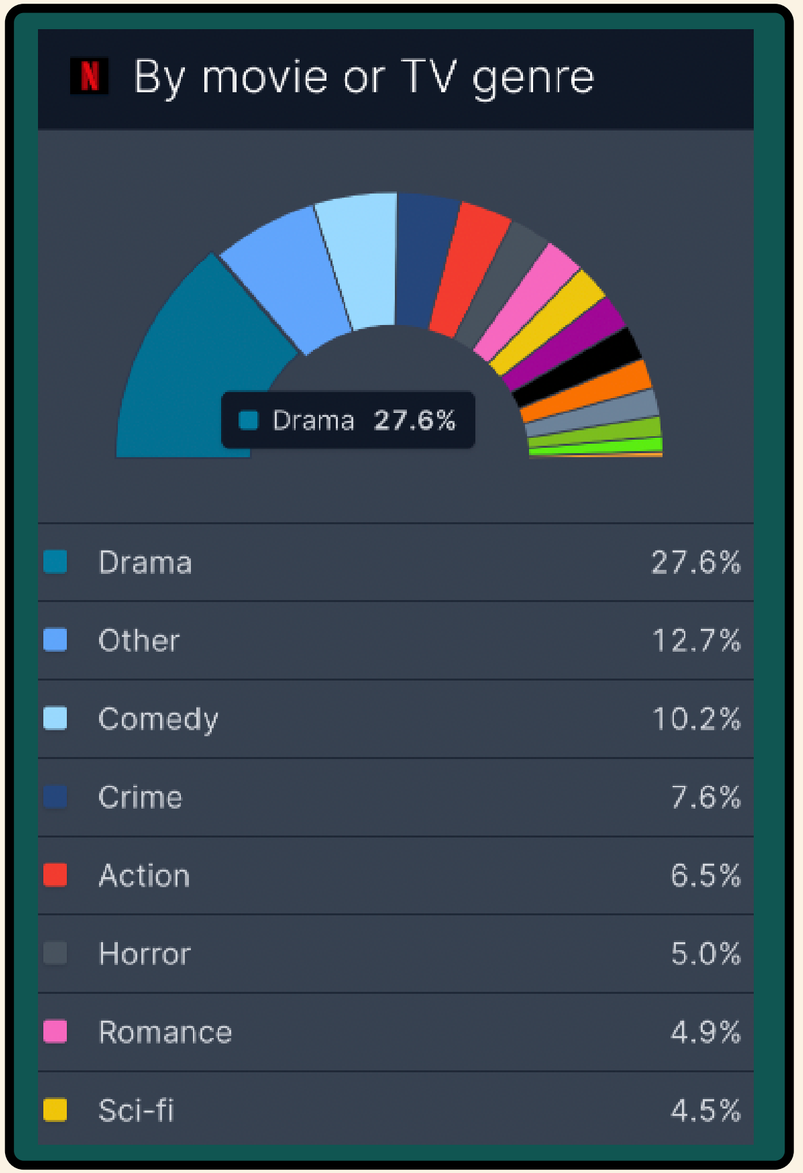
beta(2.5, 10)

RESULTS:

1. uniform prior
2. calculate beta parameters
3. choose a beta distribution
4. posterior results



DISCUSSION


By movie or TV genre
Drama 27.6%
Drama 27.6%
Other 12.7%
Comedy 10.2%
Crime 7.6%
Action 6.5%
Horror 5.0%
Romance 4.9%
Sci-fi 4.5%

2. BINOMIAL-BETA ANALYSIS: RATINGS

1. HOW CAN WE PREDICT A USER BEHAVIOR ON NETFLIX GIVEN THEIR RATING HISTORY?
2. WHAT % OF THE TOTAL RATING IS GIVEN TO 'DRAMA' FOR EACH USER?
3. WHAT % OF 'DRAMA' MOVIES SHOULD BE RECOMMENDED IN THEIR FEED?

METHODS:

1. data wrangling
2. histogram
3. choose a beta distribution
4. posterior results

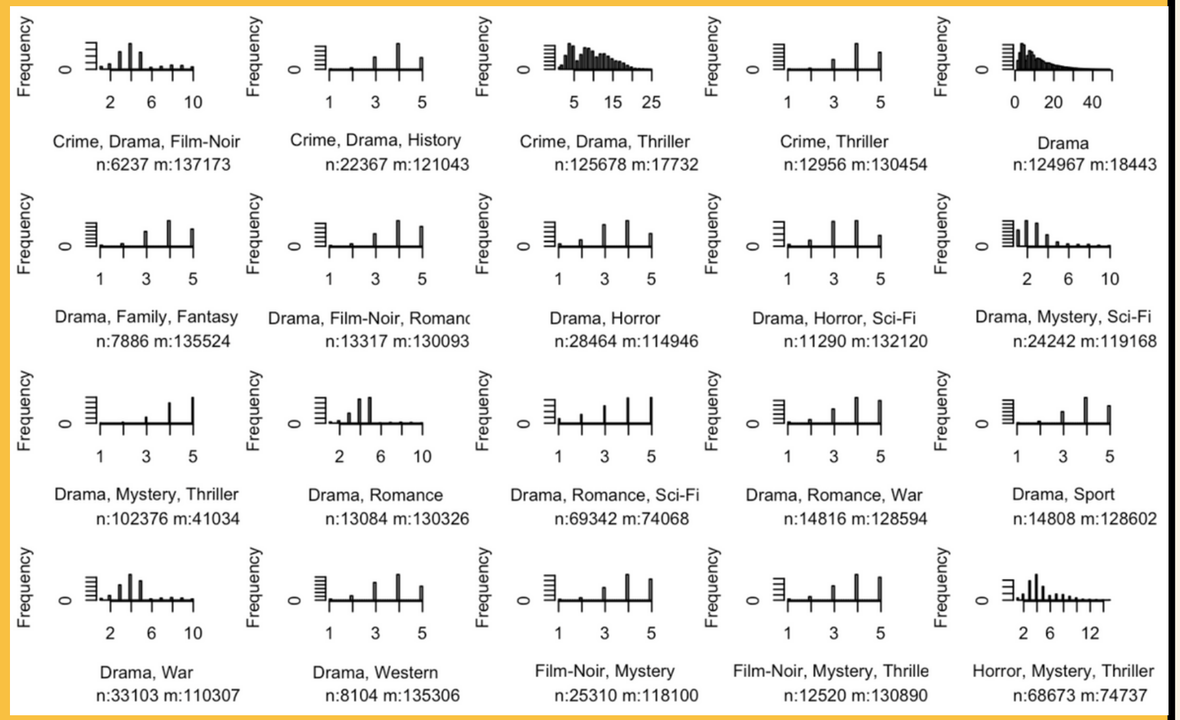
A tibble: 143,410 × 6

User_ID <dbl>	Crime, Drama, Thriller <dbl>	Drama <dbl>	n_total <dbl>
6	13	2	116
7	21	29	156
79	11	10	90
97	10	12	57
134	10	4	97
169	9	8	73
183	5	4	63
188	11	3	105
195	14	5	96
199	16	5	115

1–10 of 143,410 rows

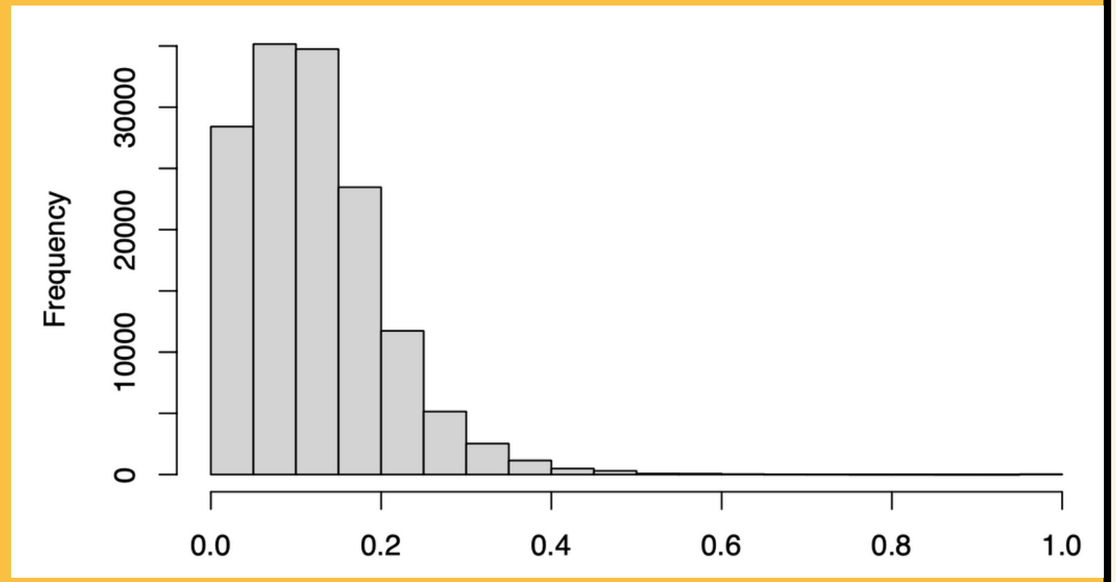
METHODS:

1. data wrangling
2. histogram
3. choose a beta distribution
4. posterior results



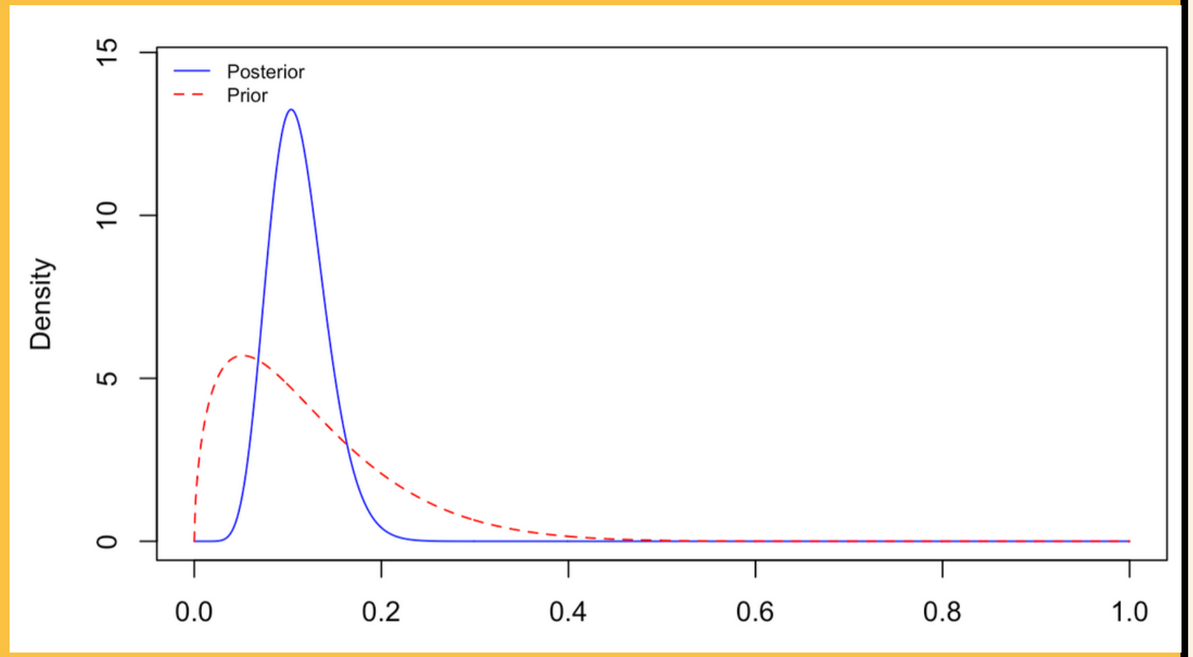
METHODS:

1. data wrangling
2. histogram
3. choose a beta distribution
4. posterior results



RESULTS:

1. data wrangling
2. histogram
3. choose a beta distribution
4. posterior results



RESULTS:

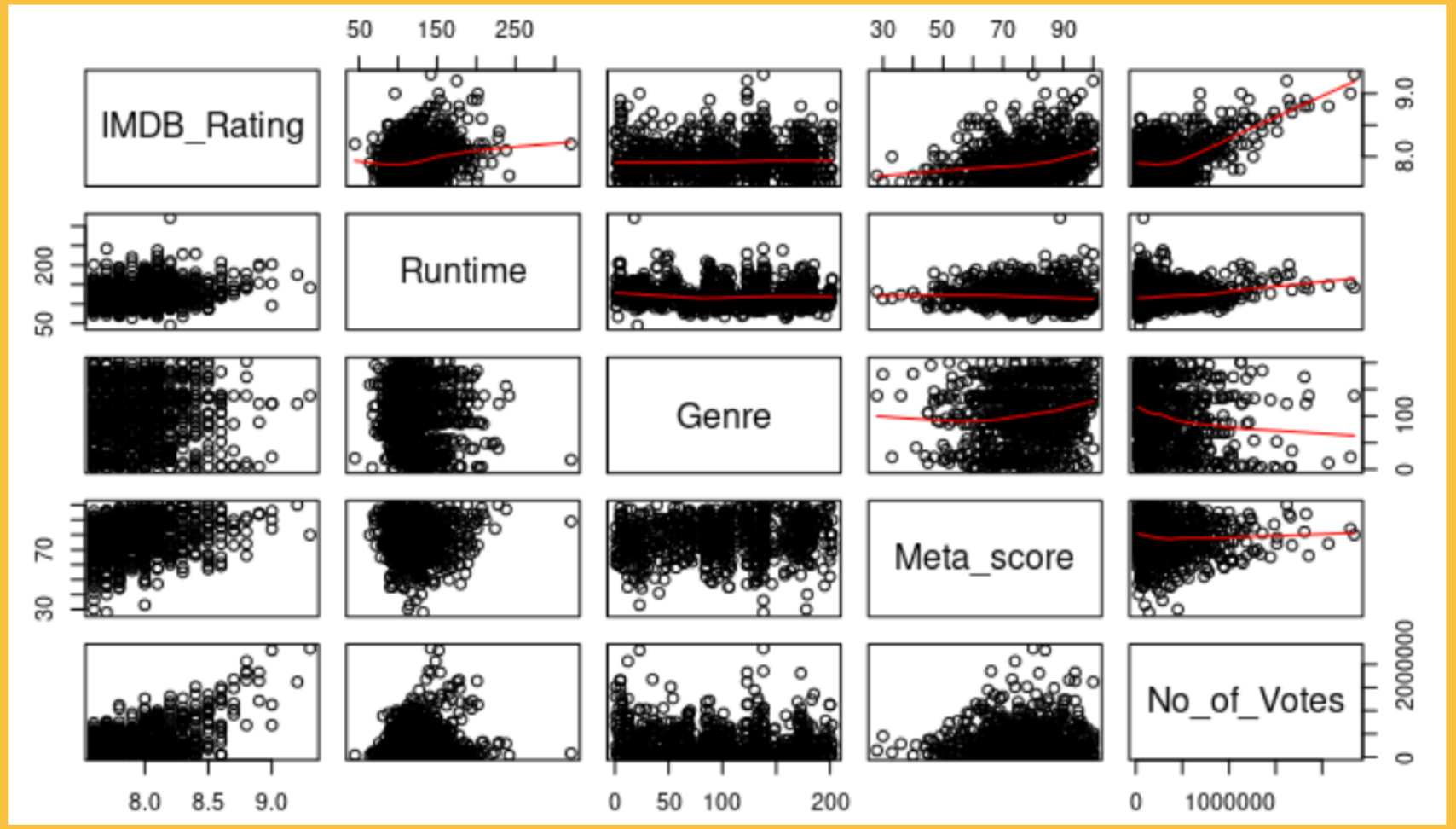
1. uniform prior
2. calculate beta parameters
3. choose a beta distribution
4. posterior results

```
## Posterior Mean            :  0.1112104
## Posterior Variance        :  0.0009407
## Posterior Std. Deviation  :  0.0306708
##
## Prob.    Quantile
## ------   ---------
## 0.005    0.0467325
## 0.010    0.0513104
## 0.025    0.0585696
## 0.050    0.0653342
## 0.500    0.1087196
## 0.950    0.1655968
## 0.975    0.1779483
## 0.990    0.1928566
## 0.995    0.2033318
```

```
quantile(result5, probs = c(0.025, 0.975))
```

```
## [1] 0.05856955 0.17794830
```

3. BAYESIAN LINEAR REGRESSION



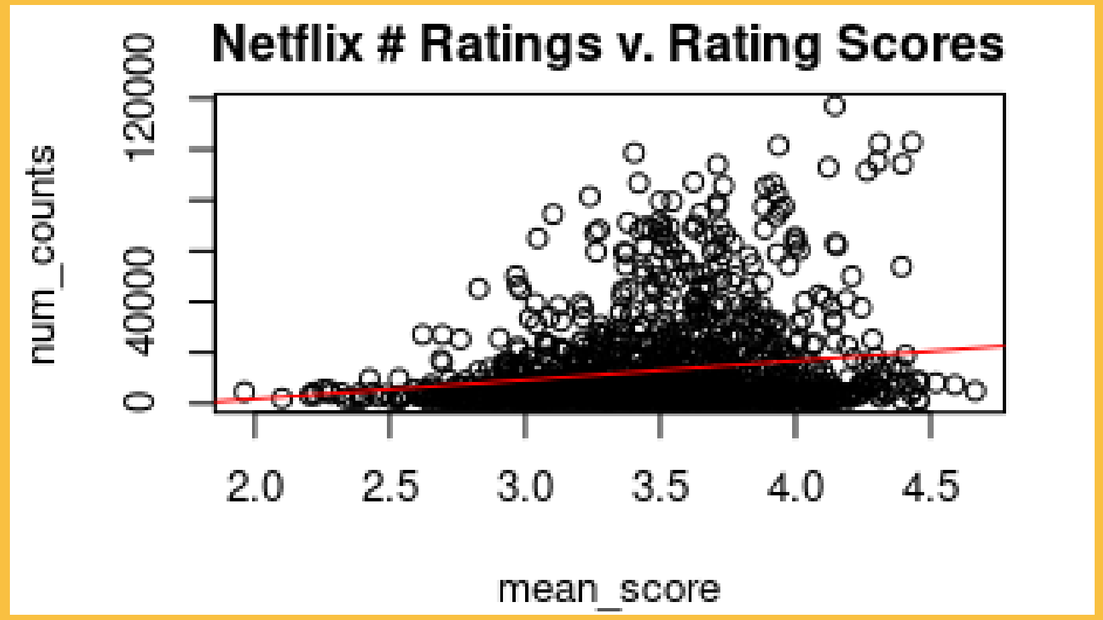
RESULTS: BAYESIAN LINEAR REGRESSION

1. variables
2. first update (Netflix data)
3. posterior result
4. second update (IMDB data)
5. posterior result

- Explanatory: average film ratings by users on IMDB and Netflix
- Response: number of ratings given by users for each film

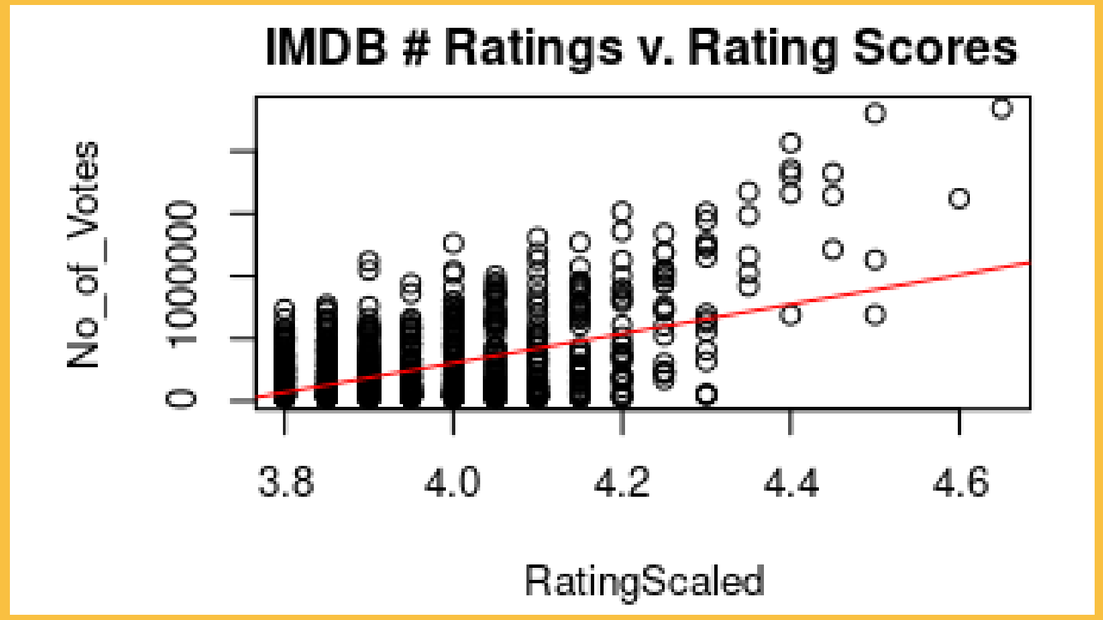
RESULTS: BAYESIAN LINEAR REGRESSION

1. variables
2. first update (Netflix data)
3. posterior result
4. second update (IMDB data)
5. posterior result



RESULTS: BAYESIAN LINEAR REGRESSION

1. variables
2. first update (Netflix data)
3. posterior result
4. second update (IMDB data)
5. posterior result



RESULTS: BAYESIAN LINEAR REGRESSION

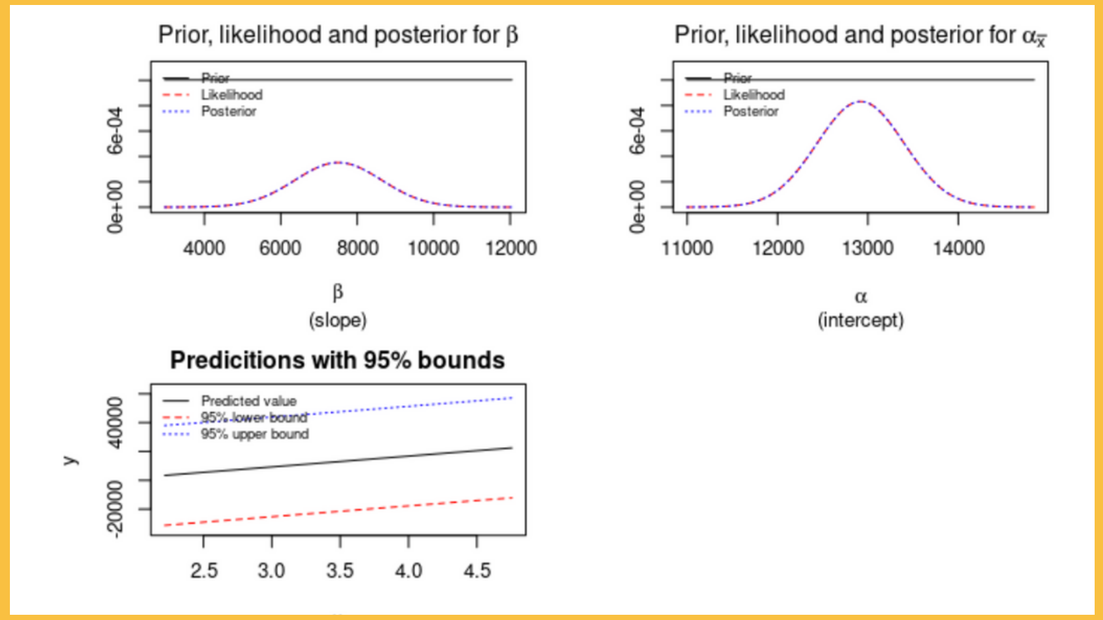
1. variables
2. first update (Netflix data)
3. posterior result
4. second update (IMDB data)
5. posterior result

Name <fctr>	mean_score <dbl>	num_counts <int>
Lost: Season 1	4.665432	4860
The Simpsons: Season 6	4.589824	7331
Family Guy: Freakin' Sweet Collection	4.520766	8090
Six Feet Under: Season 4	4.461601	8581
Inu-Yasha	4.457774	1042
Stargate SG-1: Season 8	4.456026	1535
The Best of Friends: Vol. 4	4.449168	8774
The West Wing: Season 3	4.436258	5185
Lord of the Rings: The Fellowship of the Ring	4.431489	102721
Gilmore Girls: Season 3	4.428943	3849

1-10 of 1,342 rows Previous

RESULTS: BAYESIAN LINEAR REGRESSION

1. variables
2. first update (Netflix data)
3. posterior result
4. second update (IMDB data)
5. posterior result



RESULTS: BAYESIAN LINEAR REGRESSION

1. variables
2. first update (Netflix data)
3. posterior result
4. second update (IMDB data)
5. posterior result

```
Standard deviation of residuals:  17600
             Posterior Mean Posterior Std. Deviation
             -------------- ------------------------
Intercept:   12920          480.19
Slope:       7505           1135
```

- Beta estimate: 7505 additional raters per point added (1-5 scale)
- 95% C.I.: 5280.4 <> 9729.6

RESULTS: BAYESIAN LINEAR REGRESSION

1. variables
2. first update (Netflix data)
3. posterior result
4. second update (IMDB data)
5. posterior result

~/MATH153 Bayesian/153-hw8.Rmd

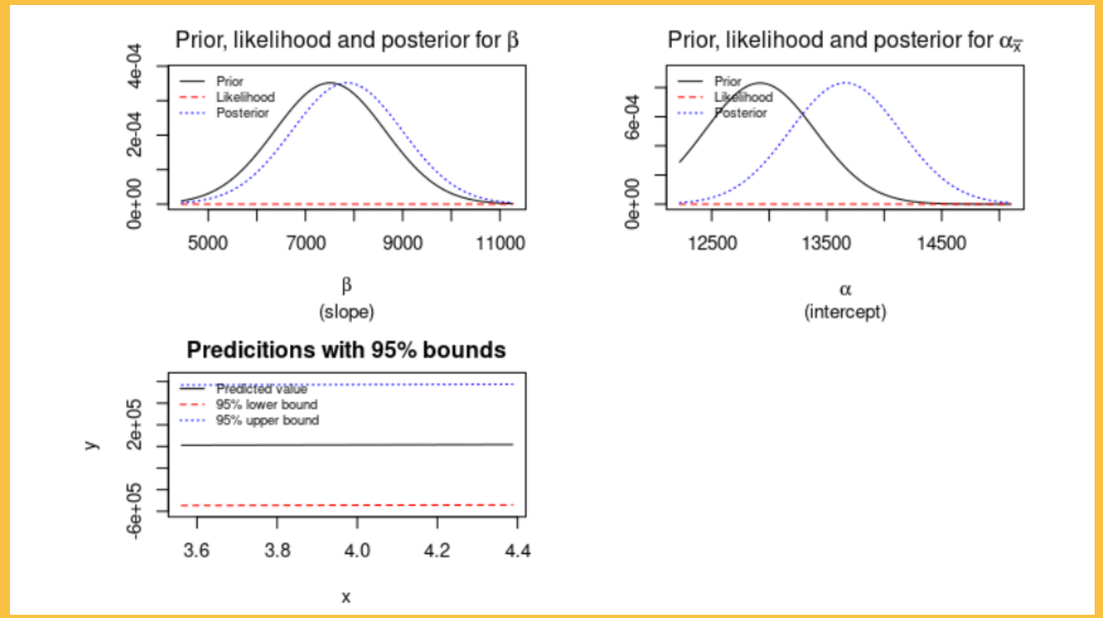
Name	RatingScaled <dbl>	No_of_Votes <dbl>
The Shawshank Redemption	4.65	2343110
The Godfather	4.60	1620367
The Dark Knight	4.50	2303232
The Godfather: Part II	4.50	1129952
12 Angry Men	4.50	689845
The Lord of the Rings: The Return of the King	4.45	1642758
Pulp Fiction	4.45	1826188
Schindler's List	4.45	1213505
Inception	4.40	2067042
Fight Club	4.40	1854740

1-10 of 1,000 rows

Previous 1 2 3 4 5

RESULTS: BAYESIAN LINEAR REGRESSION

1. variables
2. first update (Netflix data)
3. posterior result
4. second update (IMDB data)
5. posterior result



RESULTS: BAYESIAN LINEAR REGRESSION

1. variables
2. first update (Netflix data)
3. posterior result
4. second update (IMDB data)
5. posterior result

```
Standard deviation of residuals:  285000
             Posterior Mean Posterior Std. Deviation
             -------------- ------------------------
Intercept:   13660          479.51
Slope:       7857           1134.8
```

- Beta estimate: 7857 additional raters per point added (1-5 scale)
- 95% C.I.: 5,632.79 <> 10,081.21

THANK YOU !!!