Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(statsr)
library(BAS)

library(MASS)

library(GGally)
library(broom)
```

Load data

```
load("movies.Rdata")
```

Part 1: Data

This data is a random sample of movies released before 2016 and collected from the IMDB and Rotten Tomatoes movie database. Therefore, the results of this analysis can only be generalized to the sampled population from said databases; considering the breadth of these databases, though technically not true, the results are most likely generalizable to all Western (American and European) films released before 2016.

Since the data is collected from past information and no random assignment of treatment is performed (basically impossible to implement in an experiment) only assocations rather than causations can be found.

Part 2: Data manipulation

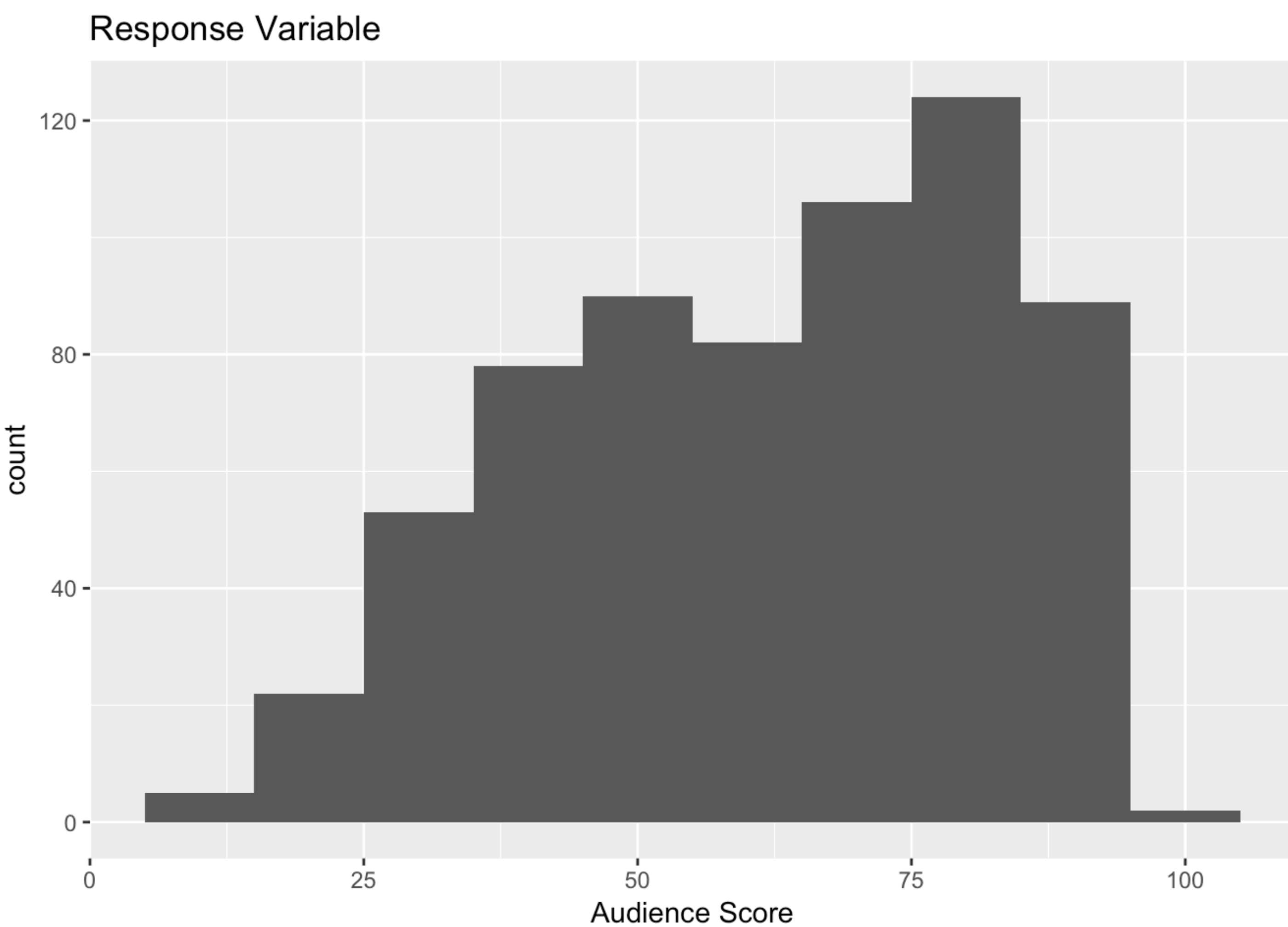
```
 movies_manipulate <- movies

as.character(as.numeric(movies_manipulate$thtr_rel_month)) %>%
  str(thtr_rel_month)
```

```
##  chr [1:651] "4" "3" "8" "10" "9" "1" "1" "11" "9" "3" "6" "12" "1" ...
```

```
movies_manipulate <- mutate (movies_manipulate,
    feature_film = ifelse(title_type == 'Feature Film', 'Yes', 'No'),
    drama = ifelse(genre == 'Drama', 'Yes', 'No'),
    mpaa_rating_R = ifelse(mpaa_rating == 'R', 'Yes', 'No'),
    oscar_season = ifelse(thtr_rel_month >= 10, 'Yes', 'No'),
    summer_season = ifelse(thtr_rel_month >= 5 & thtr_rel_month <= 8, 'Yes', 'No'))
```

Part 3: Exploratory data analysis

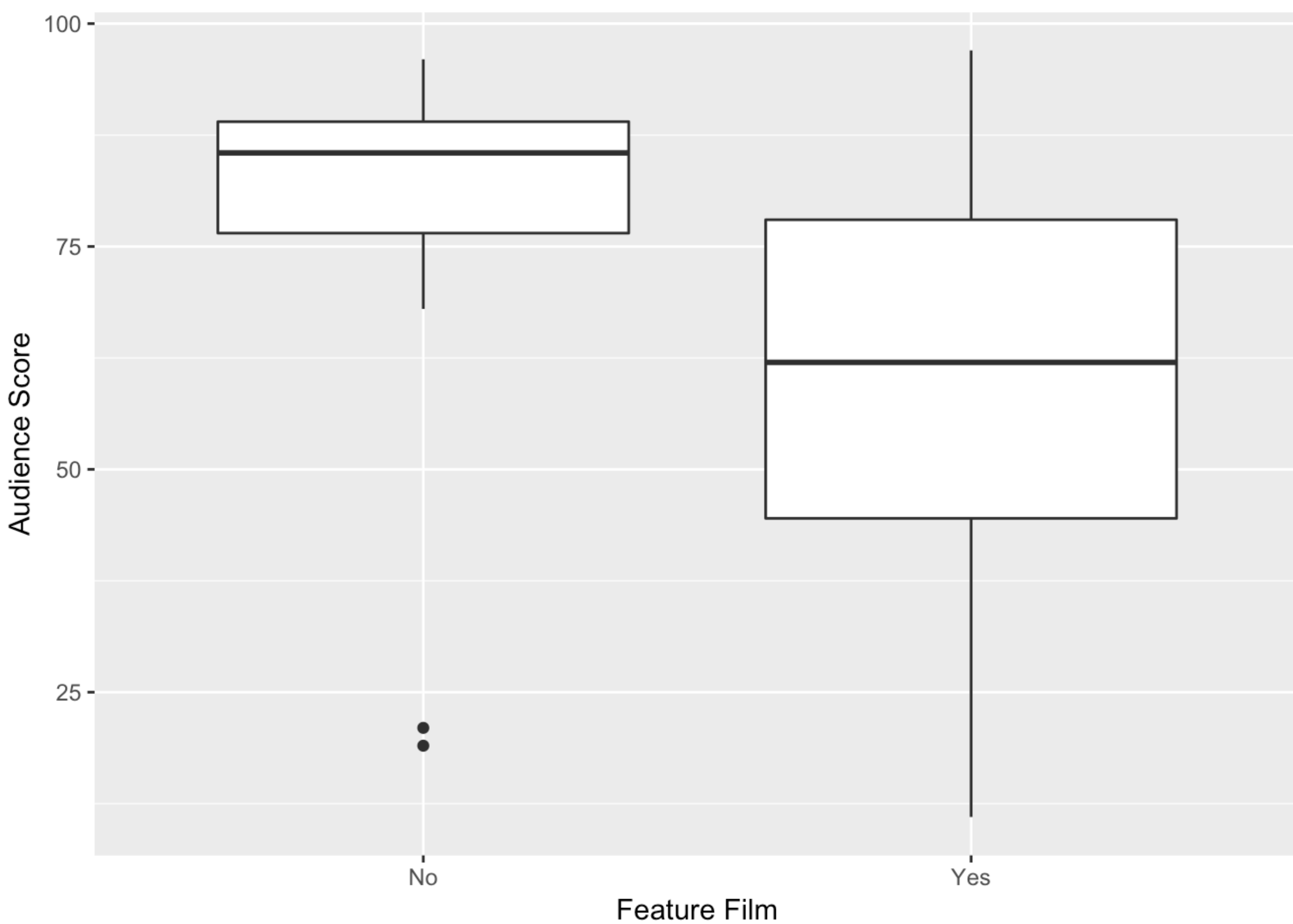
```
# Response Variable: Audience Score
ggplot(data = movies_manipulate, aes(x = audience_score)) + geom_histogram(binwidth =
10) + labs(x = "Audience Score", title = "Response Variable")
```




```
summary(movies_manipulate$audience_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.00   46.00   65.00   62.36   80.00   97.00
```

```
# The graph shows the data as skewed left. That being said, the mean (65) and the med
ian (62.36) are similar to one another, which normally only occurs when the data is r
oughly normal.

# Feature Film
ggplot(data = movies_manipulate, aes(x = feature_film, y = audience_score)) + geom_bo
xplot() + labs(x = "Feature Film", y = "Audience Score")
```

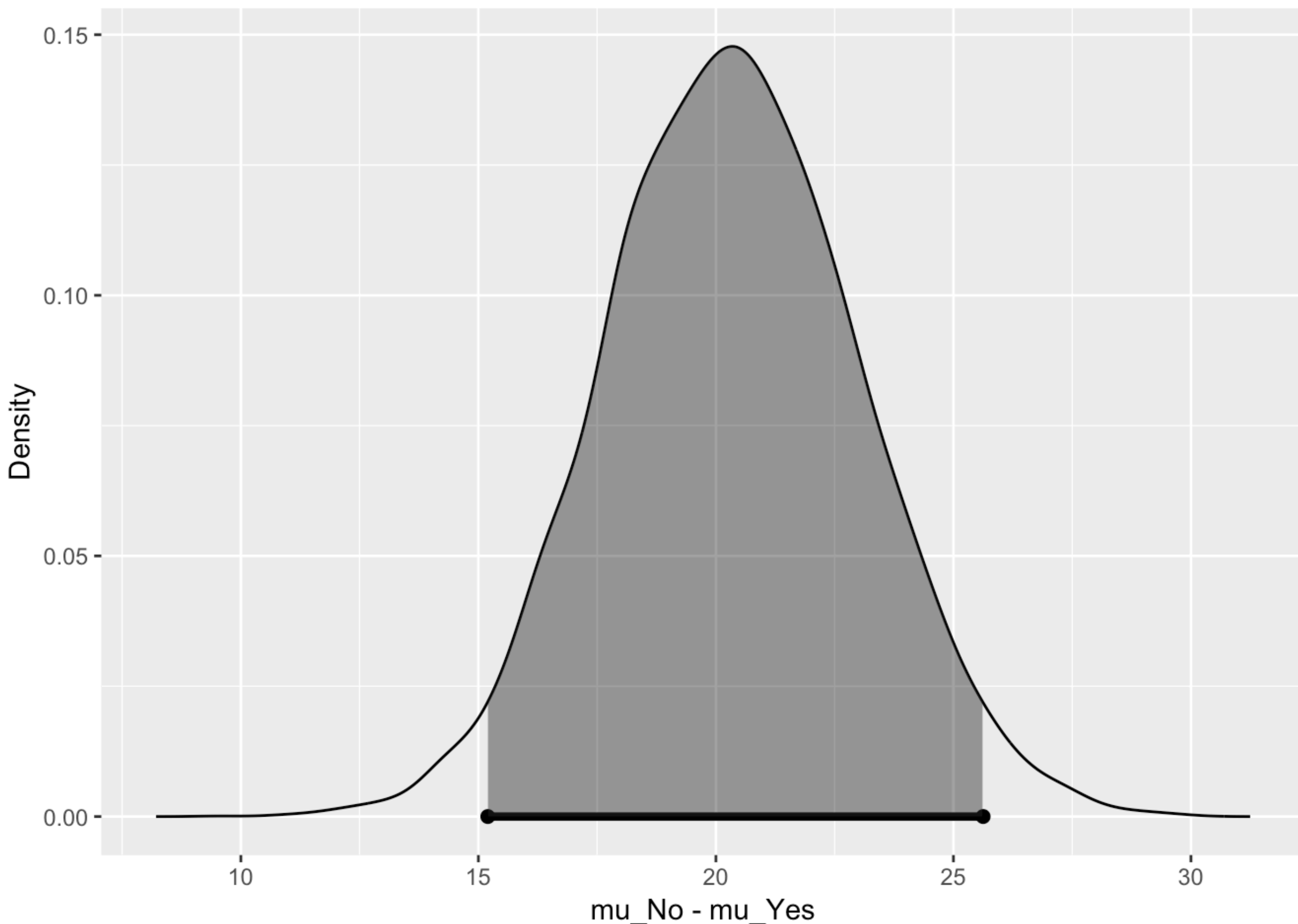



```
# Feature Films are roughly normally distributed in regards to audience score whereas
non-feature films appear to be right-skewed; non-feature films tend to have a higher
audience score based upon this plot.
movies_manipulate %>%
  group_by(feature_film) %>%
  summarise(mean_FF = mean(audience_score), sd_FF = sd(audience_score),
            median_FF = median(audience_score), IQR_FF = IQR(audience_score),
            n = n())
```

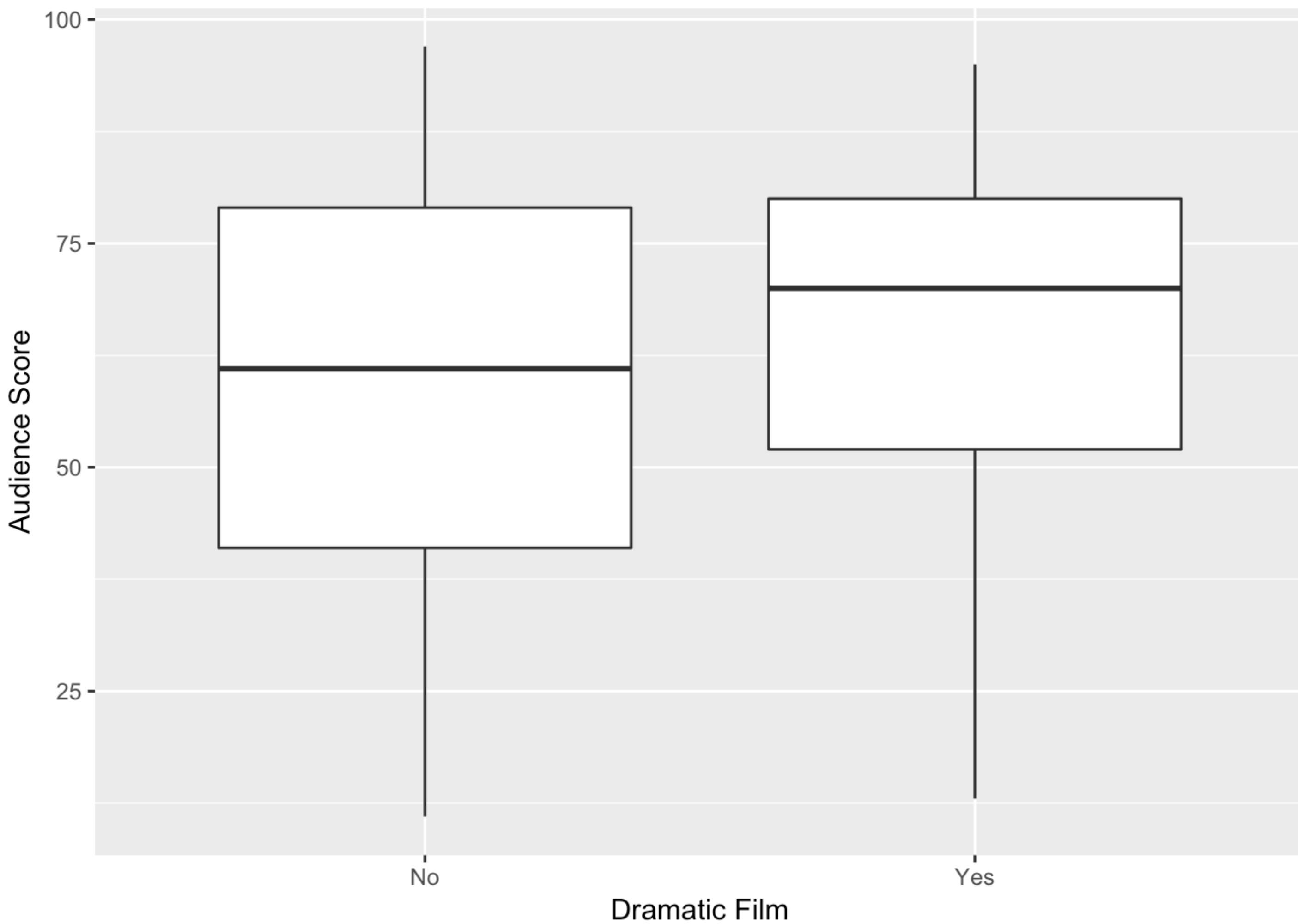
```
## # A tibble: 2 x 6
##   feature_film mean_FF sd_FF median_FF IQR_FF     n
##   <chr>          <dbl> <dbl>     <dbl>  <dbl> <int>
## 1 No              81.0  13.6      85.5   12.5    60
## 2 Yes             60.5  19.8      62     33.5   591
```

```
# Summary statistics confirm that films that are not feature films are right skewed (
mean = 81; median = 85.5 [IQR = 12.5]) in their audience score. Feature films were no
rmally distributed (mean = 60.5; sd = 19.8). This data supports the results of the pl
ot in that non-feature films had higher audience scores. It should be noted that ther
e was a smaller sample of non-feature films, which may imply that there is not enough
data to draw a strong conclusion.
bayes_inference(y = audience_score, x = feature_film, data = movies_manipulate,
                statistic = "mean", type = "ht",
                null = 0, alternative = "twosided",
                prior = "JZS", rscale = 1,
                method = "theoretical")
```

```
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_No = 60, y_bar_No = 81.05, s_No = 13.5764
## n_Yes = 591, y_bar_Yes = 60.4653, s_Yes = 19.824
## (Assuming Zellner-Siow Cauchy prior on the difference of means. )
## (Assuming independent Jeffreys prior on the overall mean and variance. )
## Hypotheses:
## H1: mu_No  = mu_Yes
## H2: mu_No != mu_Yes
##
## Priors: P(H1) = 0.5  P(H2) = 0.5
##
## Results:
## BF[H2:H1] = 338337769673
## P(H1|data) = 0
## P(H2|data) = 1
##
## Posterior summaries for under H2:
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_No = 60, y_bar_No = 81.05, s_No = 13.5764
## n_Yes = 591, y_bar_Yes = 60.4653, s_Yes = 19.824
## (Assuming Zellner-Siow Cauchy prior for difference in means)
## (Assuming independent Jeffrey's priors for overall mean and variance)
##
##
## Posterior Summaries
##                      2.5%         25%        50%        75%       97.5%
## overall mean    68.1555171  69.7737209  70.657337  71.543891   73.332208
## mu_No - mu_Yes  15.1986086  18.5036997  20.308571  22.121333   25.625392
## sigma^2        335.8981696 361.1724847 374.739157 388.808661  418.660797
## effect size      0.7783184   0.9551485   1.049415   1.144329    1.327966
## n_0             15.8499761 175.2344256 426.630287 862.365611 2381.008934
## 95% Cred. Int.: (15.1986 , 25.6254)
```




```
# Based upon the Bayes Factor there is very strong evidence that the distinction betw
een feature films and non-feature films has an effect upon audience score. The probab
ility that a non-feature film has on average 15 to 25 higher audience score is 0.95.

# Drama
ggplot(data = movies_manipulate, aes(x = drama, y = audience_score)) + geom_boxplot()
+ labs(x = "Dramatic Film", y = "Audience Score")
```

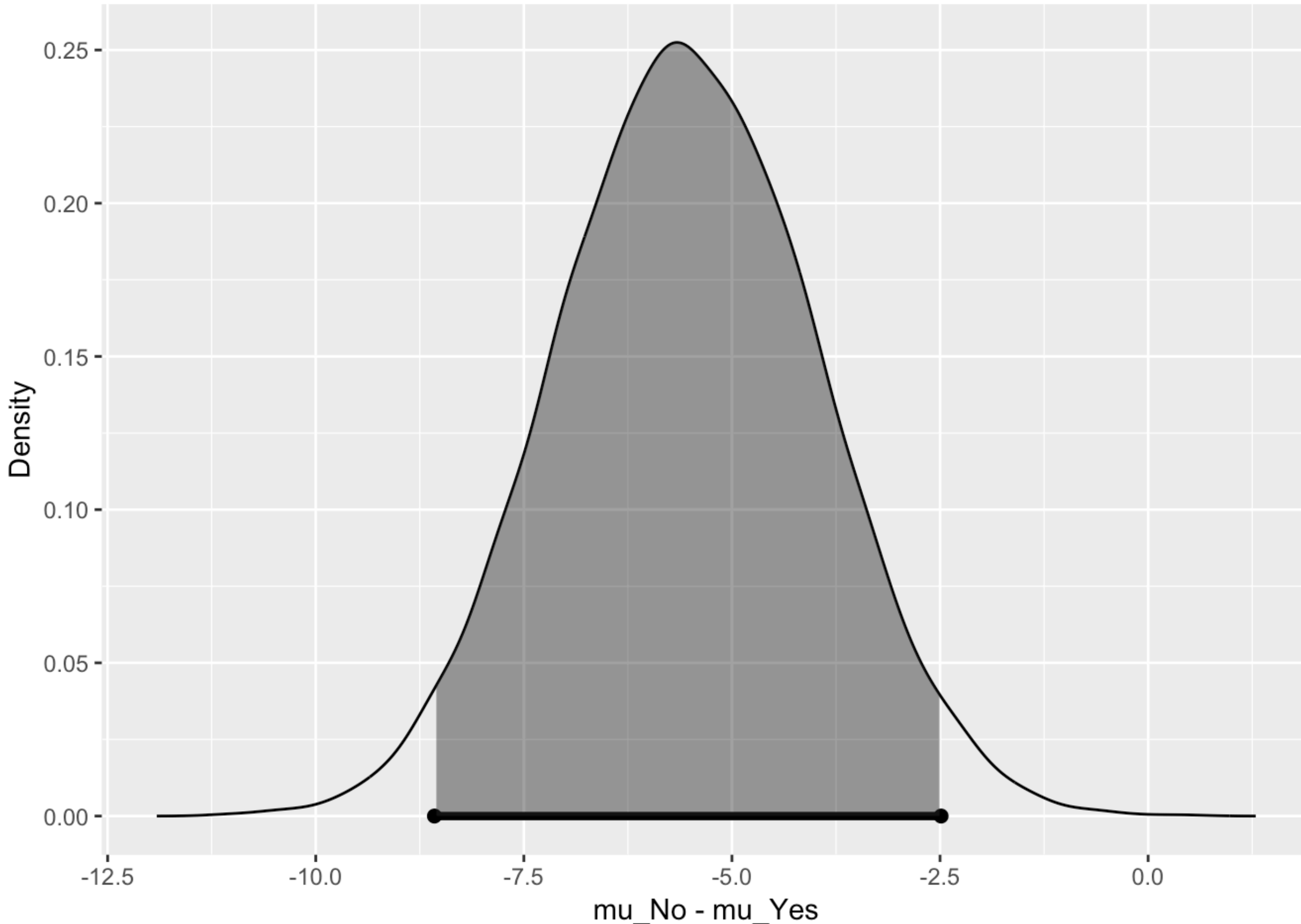



```
# Dramatic films appear to be left skewed though the median value of audience score i
s greater than that of non-dramatic films. Non-dramatic films appear to be normally d
istributed.
movies_manipulate %>%
  group_by(drama) %>%
  summarise(mean_dr = mean(audience_score), sd_dr = sd(audience_score),
            median_dr = median(audience_score), IQR_dr = IQR(audience_score),
            n = n())
```

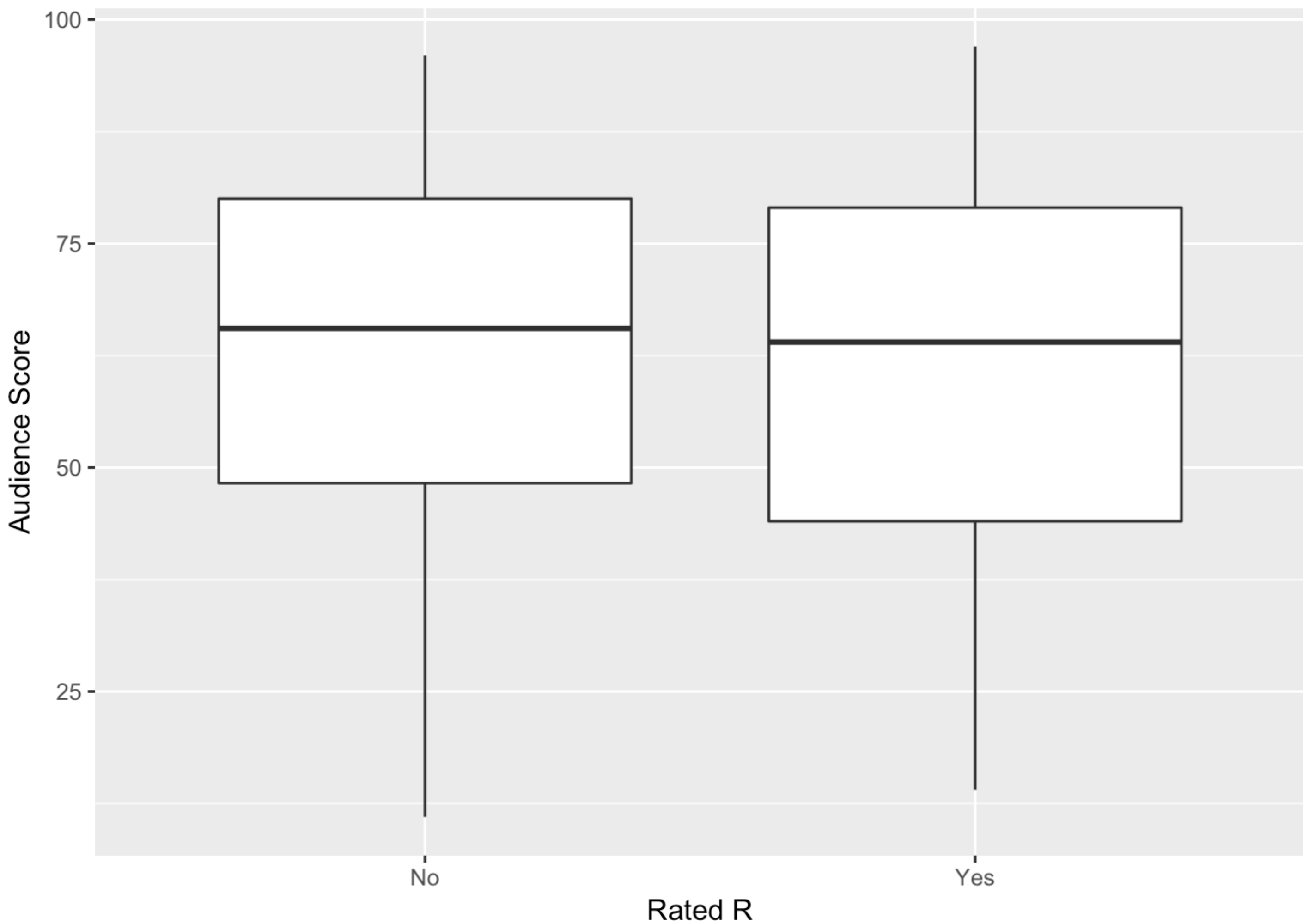
```
## # A tibble: 2 x 6
##   drama mean_dr sd_dr median_dr IQR_dr     n
##   <chr>   <dbl> <dbl>     <dbl>  <dbl> <int>
## 1 No       59.7  21.3        61     38   346
## 2 Yes      65.3  18.5        70     28   305
```

```r
# Dramatic films have a median* audience score of 70 with an IQR of 28. (*Due to the
skewed distribution of data, evidenced by vastly different median and mean values the
median and IQR data is provided.) Non-dramatic films had a mean audience score of 59.
7 with a standard deviation of 21 points.
bayes_inference(y = audience_score, x = drama, data = movies_manipulate,
                statistic = "mean", type = "ht",
                null = 0, alternative = "twosided",
                prior = "JZS", rscale = 1,
                method = "theoretical")
```

```
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_No = 346, y_bar_No = 59.7312, s_No = 21.2775
## n_Yes = 305, y_bar_Yes = 65.3475, s_Yes = 18.5418
## (Assuming Zellner-Siow Cauchy prior on the difference of means. )
## (Assuming independent Jeffreys prior on the overall mean and variance. )
## Hypotheses:
## H1: mu_No  = mu_Yes
## H2: mu_No != mu_Yes
##
## Priors: P(H1) = 0.5  P(H2) = 0.5
##
## Results:
## BF[H2:H1] = 31.9101
## P(H1|data) = 0.0304
## P(H2|data) = 0.9696
##
## Posterior summaries for under H2:
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_No = 346, y_bar_No = 59.7312, s_No = 21.2775
## n_Yes = 305, y_bar_Yes = 65.3475, s_Yes = 18.5418
## (Assuming Zellner-Siow Cauchy prior for difference in means)
## (Assuming independent Jeffrey's priors for overall mean and variance)
##
##
## Posterior Summaries
##                       2.5%         25%         50%          75%
## overall mean    60.9759384  61.9985457  62.5446289   63.0543697
## mu_No - mu_Yes  -8.5740119  -6.6073995  -5.5556489   -4.4907638
## sigma^2        360.3449253 387.1630038 401.5538468  416.8317642
## effect size     -0.4276108  -0.3303862  -0.2770954   -0.2234412
## n_0             33.4898862 353.3138667 834.5173375 1687.7365093
##                      97.5%
## overall mean     64.0748098
## mu_No - mu_Yes   -2.4870239
## sigma^2         448.9100662
## effect size      -0.1233749
## n_0            4419.3541422
## 95% Cred. Int.: (-8.574 , -2.487)
```




```
# There is strong evidence (BF = 30) that dramatic movies on average have a lower aud
ience score. There is a 95% probability that non-dramatic films score 8.6 to 2.4 poin
ts lower than dramatic films on average.

# MPAA Rating R
ggplot(data = movies_manipulate, aes(x = mpaa_rating_R, y = audience_score)) + geom_b
oxplot() + labs(x = "Rated R", y = "Audience Score")
```

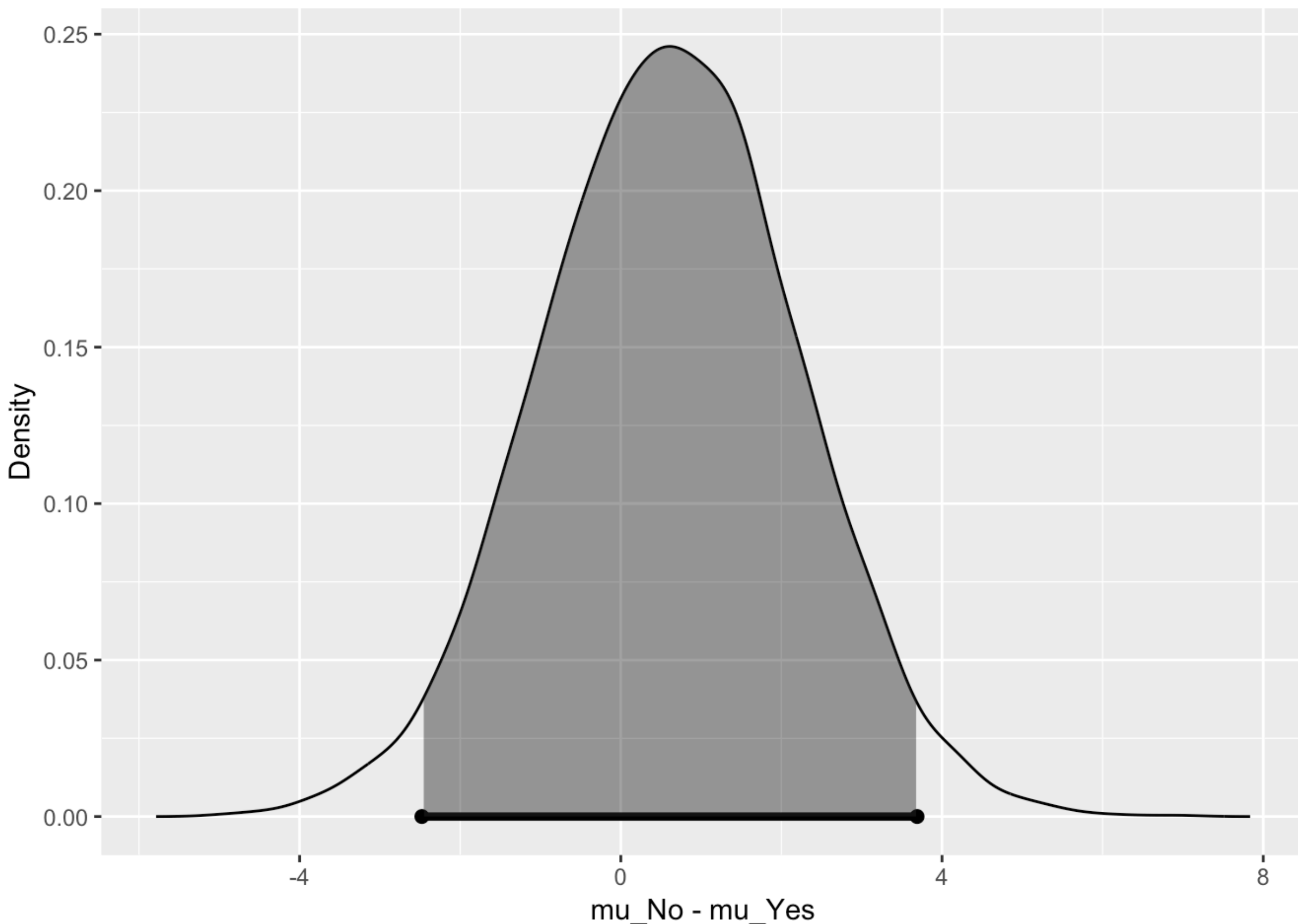



```
# Movies rated R appears normally distributed whereas films not rated R appear to be
left skewed.
movies_manipulate %>%
  group_by(mpaa_rating_R) %>%
  summarise(mean_R = mean(audience_score), sd_R = sd(audience_score),
            median_R = median(audience_score), IQR_R = IQR(audience_score),
            n = n())
```

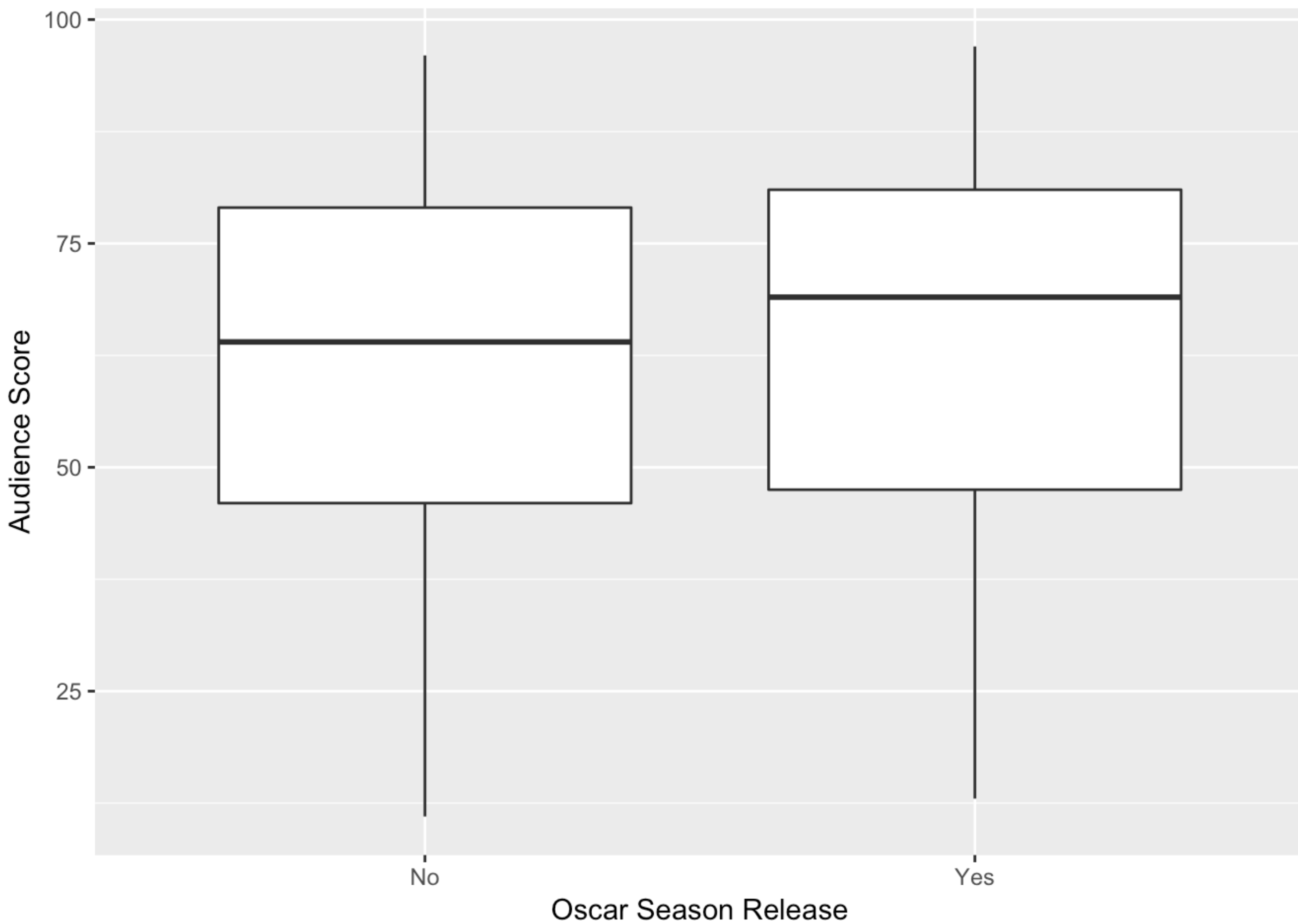
```
## # A tibble: 2 x 6
##   mpaa_rating_R mean_R  sd_R median_R IQR_R     n
##   <chr>          <dbl> <dbl>    <dbl> <dbl> <int>
## 1 No              62.7  20.3     65.5  31.8   322
## 2 Yes             62.0  20.2     64    35     329
```

```r
# Films rated R have a median audience rating of 64 and an IQR of 35 (*statistics for skewed data reported as the mean is different from the median). Films not rated R have a median score of 65.5 and and IQR of 32.
bayes_inference(y = audience_score, x = mpaa_rating_R, data = movies_manipulate,
                statistic = "mean", type = "ht",
                null = 0, alternative = "twosided",
                prior = "JZS", rscale = 1,
                method = "theoretical")
```

```
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_No = 322, y_bar_No = 62.6894, s_No = 20.3167
## n_Yes = 329, y_bar_Yes = 62.0426, s_Yes = 20.1559
## (Assuming Zellner-Siow Cauchy prior on the difference of means. )
## (Assuming independent Jeffreys prior on the overall mean and variance. )
## Hypotheses:
## H1: mu_No  = mu_Yes
## H2: mu_No != mu_Yes
##
## Priors: P(H1) = 0.5  P(H2) = 0.5
##
## Results:
## BF[H1:H2] = 14.8147
## P(H1|data) = 0.9368
## P(H2|data) = 0.0632
##
## Posterior summaries for under H2:
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_No = 322, y_bar_No = 62.6894, s_No = 20.3167
## n_Yes = 329, y_bar_Yes = 62.0426, s_Yes = 20.1559
## (Assuming Zellner-Siow Cauchy prior for difference in means)
## (Assuming independent Jeffrey's priors for overall mean and variance)
##
##
## Posterior Summaries
##                      2.5%          25%          50%          75%
## overall mean    60.8116797  61.81924671  62.36110365 6.289304e+01
## mu_No - mu_Yes  -2.4772052  -0.46468660   0.61345815 1.660548e+00
## sigma^2        368.1590809 394.24302103 409.40801289 4.248300e+02
## effect size     -0.1223484  -0.02293481   0.03041908 8.241127e-02
## n_0             31.8340778 389.22770189 905.94172383 1.821753e+03
##                      97.5%
## overall mean     63.9094670
## mu_No - mu_Yes    3.6895443
## sigma^2         458.7342050
## effect size       0.1815429
## n_0            4727.8516693
## 95% Cred. Int.: (-2.4772 , 3.6895)
```




```
# BF = 14.81; 95 = -2.45, 3.77
# Though the Bayes factor provides positive evidence that R-rated films and not-R-rat
ed films have on average different audience scores, the credible intervals includes t
he value 0 therefore suggesting that there is no significant difference in audience s
cores between the groups.

# Oscar Season
ggplot(data = movies_manipulate, aes(x = oscar_season, y = audience_score)) + geom_bo
xplot() + labs(x = "Oscar Season Release", y = "Audience Score")
```

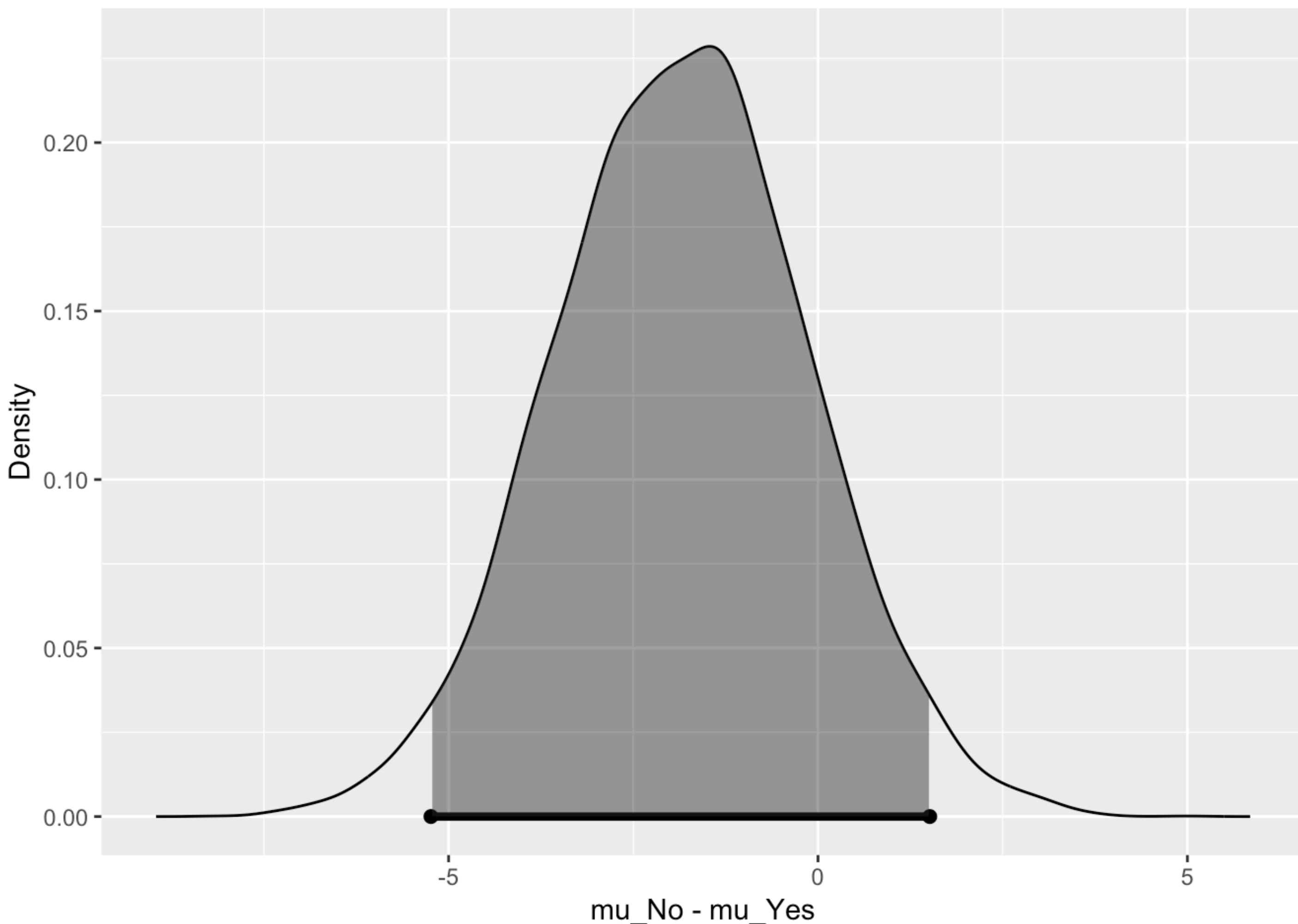



```
# Data for movies released during Oscar season (or not released during Oscar season)
appear similar for both categories and slightly left skewed.
movies_manipulate %>%
  group_by(oscar_season) %>%
  summarise(mean_oscar = mean(audience_score), sd_oscar = sd(audience_score),
            median_oscar = median(audience_score), IQR_oscar = IQR(audience_score),
            n = n())
```

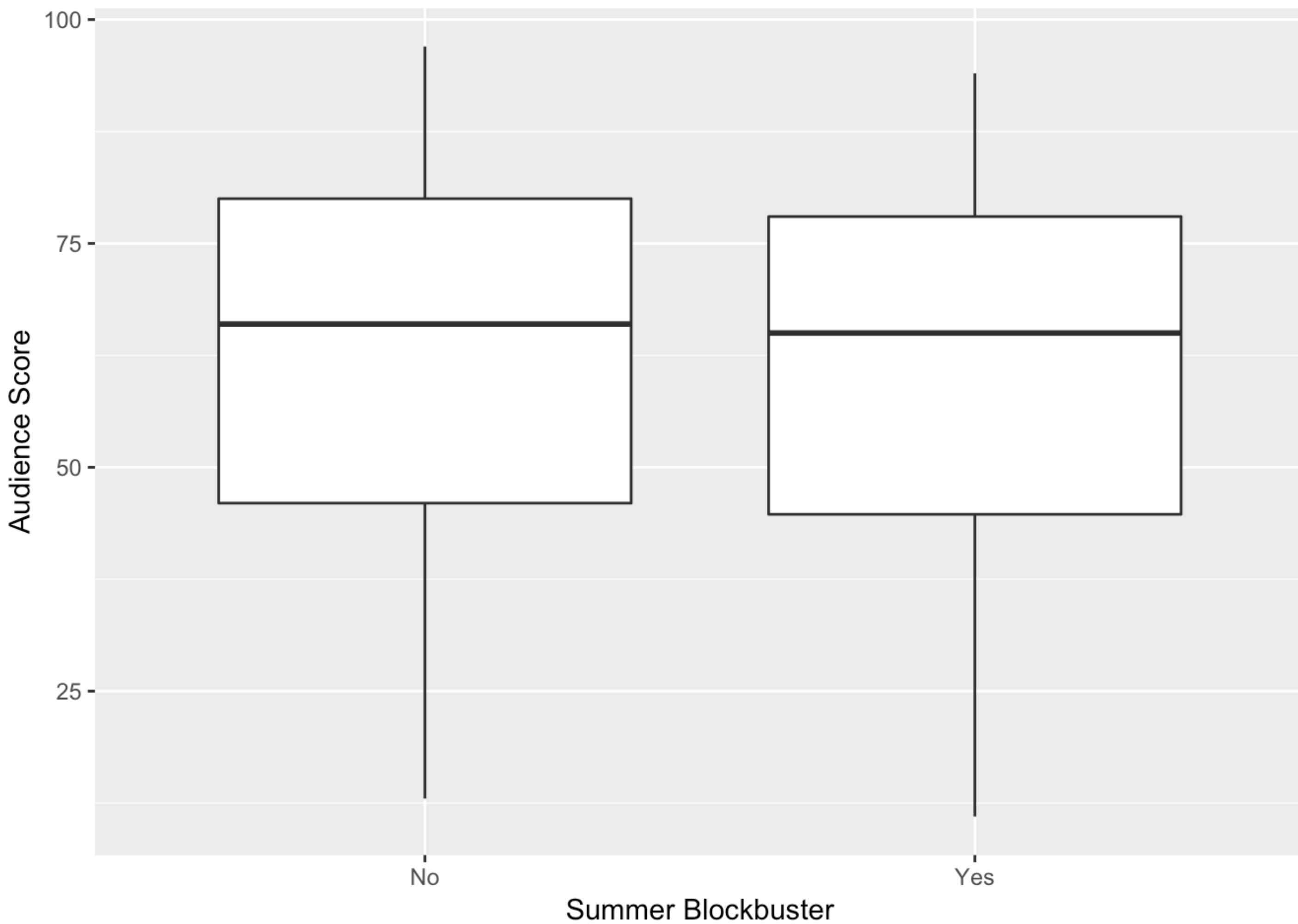
```
## # A tibble: 2 x 6
##   oscar_season mean_oscar sd_oscar median_oscar IQR_oscar     n
##   <chr>             <dbl>    <dbl>        <dbl>     <dbl> <int>
## 1 No                 61.8     20.1           64      33     460
## 2 Yes                63.7     20.5           69      33.5   191
```

```r
# Movies released during Oscar season have a median audience score of 69 with an IQR of 33. Movies not released during Oscar season have a median audience score of 64 wit
h an IQR of 33.
bayes_inference(y = audience_score, x = oscar_season, data = movies_manipulate,
                statistic = "mean", type = "ht",
                null = 0, alternative = "twosided",
                prior = "JZS", rscale = 1,
                method = "theoretical")
```

```
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_No = 460, y_bar_No = 61.813, s_No = 20.1196
## n_Yes = 191, y_bar_Yes = 63.6859, s_Yes = 20.4612
## (Assuming Zellner-Siow Cauchy prior on the difference of means. )
## (Assuming independent Jeffreys prior on the overall mean and variance. )
## Hypotheses:
## H1: mu_No  = mu_Yes
## H2: mu_No != mu_Yes
##
## Priors: P(H1) = 0.5  P(H2) = 0.5
##
## Results:
## BF[H1:H2] = 8.2858
## P(H1|data) = 0.8923
## P(H2|data) = 0.1077
##
## Posterior summaries for under H2:
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_No = 460, y_bar_No = 61.813, s_No = 20.1196
## n_Yes = 191, y_bar_Yes = 63.6859, s_Yes = 20.4612
## (Assuming Zellner-Siow Cauchy prior for difference in means)
## (Assuming independent Jeffrey's priors for overall mean and variance)
##
##
## Posterior Summaries
##                      2.5%         25%          50%           75%
## overall mean    61.0369416  62.1619740  62.74285510   63.32904822
## mu_No - mu_Yes  -5.2391091  -2.9975694  -1.83595206   -0.70603033
## sigma^2        367.0655532 393.5273529 408.58674743  424.40316507
## effect size     -0.2587563  -0.1482319  -0.09088762   -0.03459618
## n_0             35.6509347 367.8029755 905.42544547 1771.71970917
##                      97.5%
## overall mean   6.444667e+01
## mu_No - mu_Yes 1.512721e+00
## sigma^2        4.567457e+02
## effect size    7.409905e-02
## n_0            4.699248e+03
## 95% Cred. Int.: (-5.2391 , 1.5127)
```




```
# BF = 8.23; 95 = -5.35, 1.53
# Though the Bayes factor provides positive evidence that films released during Oscar
season and films not released during Oscar season have on average different audience
scores, the credible intervals includes the value 0 therefore suggesting that there i
s no significant difference in audience scores between the groups.

# Summer Season
ggplot(data = movies_manipulate, aes(x = summer_season, y = audience_score)) + geom_b
oxplot() + labs(x = "Summer Blockbuster", y = "Audience Score")
```

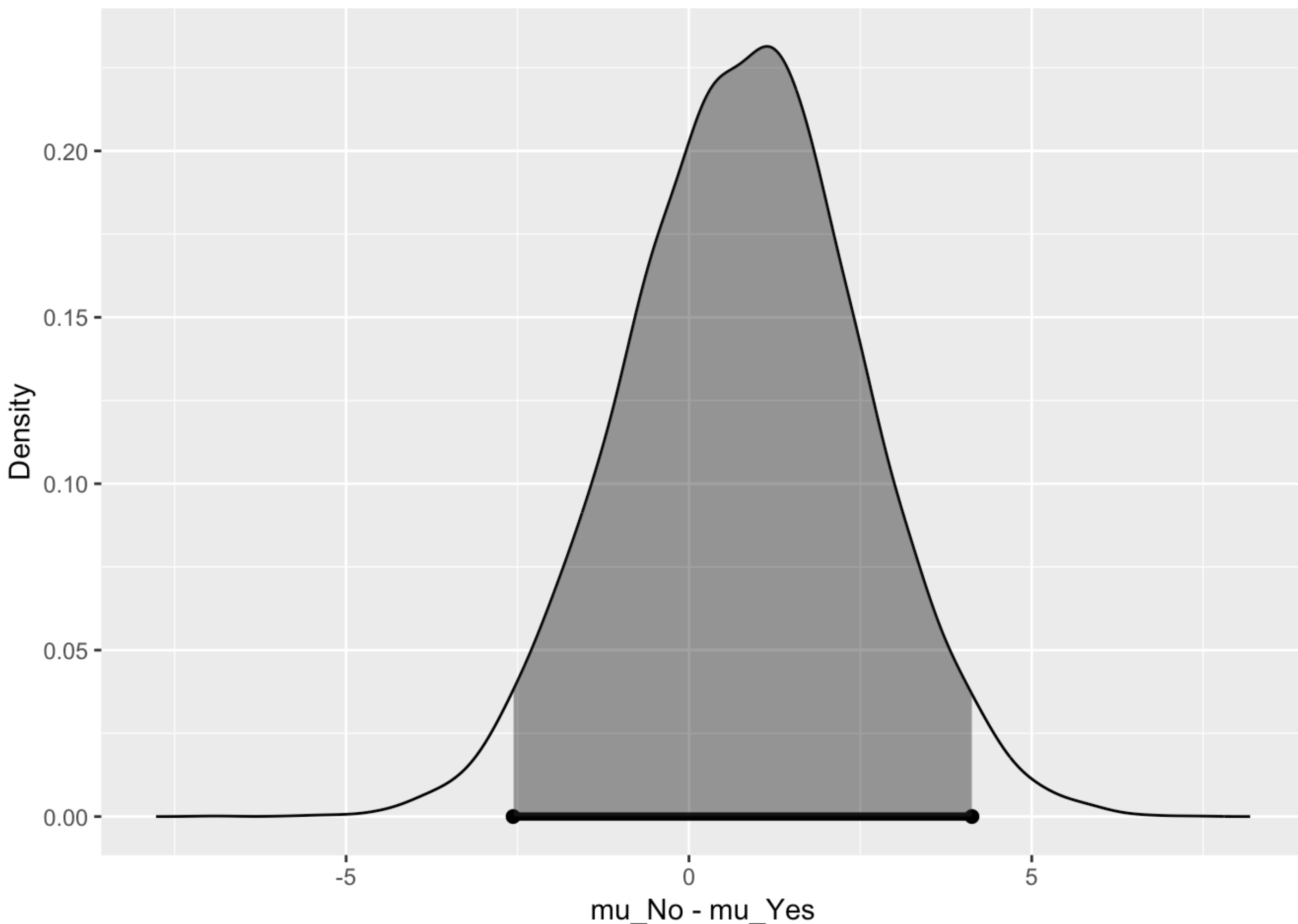



```
# Data for movies released during summer season appear similar for both categories an
d slightly left skewed.
movies_manipulate %>%
  group_by(summer_season) %>%
  summarise(mean_summer = mean(audience_score), sd_summer = sd(audience_score),
            median_summer = median(audience_score), IQR_summer = IQR(audience_score),
            n = n())
```

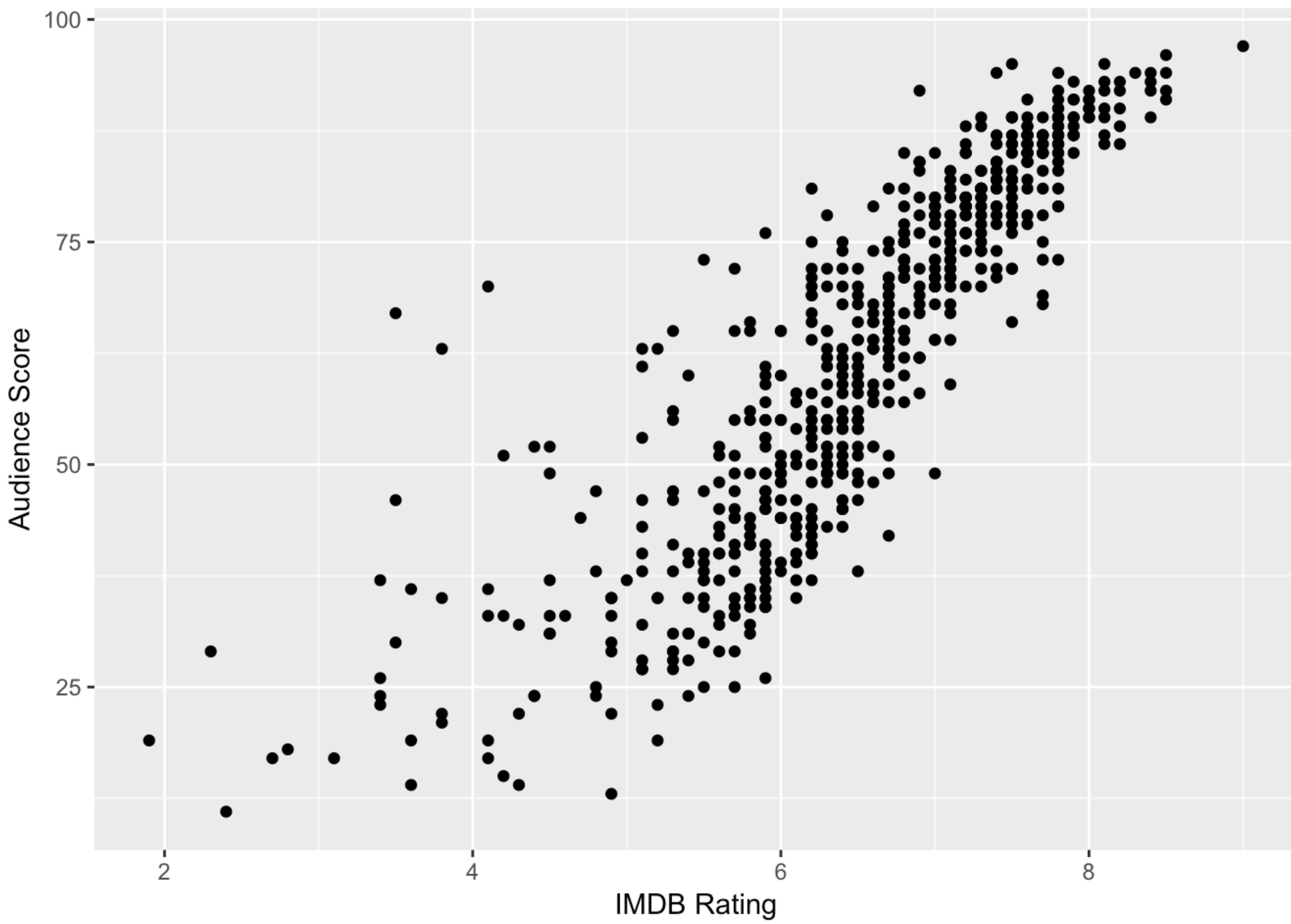
```
## # A tibble: 2 x 6
##   summer_season mean_summer sd_summer median_summer IQR_summer     n
##   <chr>               <dbl>     <dbl>         <dbl>      <dbl> <int>
## 1 No                   62.6      20.4            66       34     443
## 2 Yes                  61.8      19.9            65       33.2   208
```

```
# Movies released during summer season have a median audience score of 66 with an IQR
of 34. Movies not released during summer season have a median audience score of 65 wi
th an IQR of 33.
bayes_inference(y = audience_score, x = summer_season, data = movies_manipulate,
                statistic = "mean", type = "ht",
                null = 0, alternative = "twosided",
                prior = "JZS", rscale = 1,
                method = "theoretical")
```

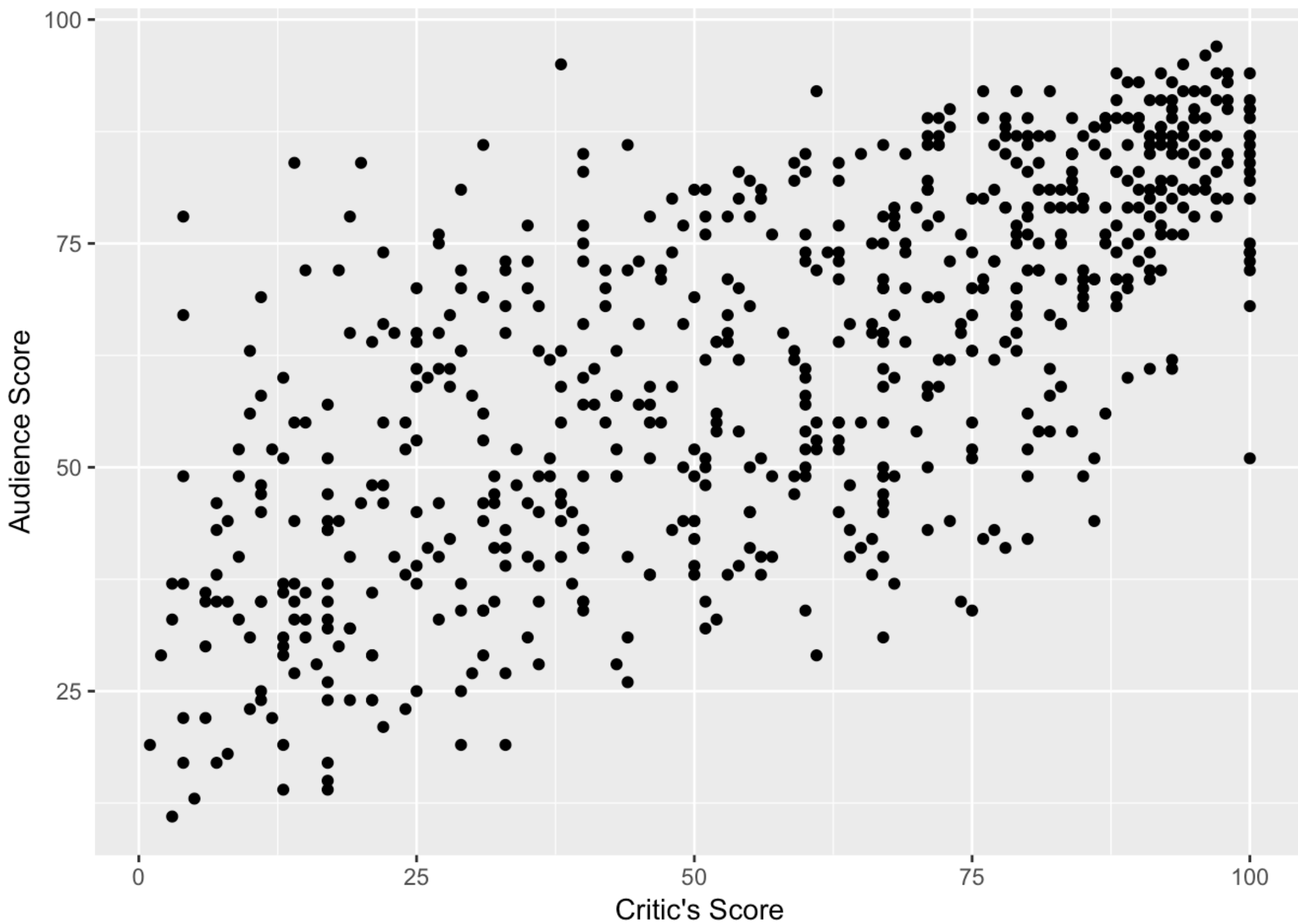
```
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_No = 443, y_bar_No = 62.623, s_No = 20.3857
## n_Yes = 208, y_bar_Yes = 61.8077, s_Yes = 19.9083
## (Assuming Zellner-Siow Cauchy prior on the difference of means. )
## (Assuming independent Jeffreys prior on the overall mean and variance. )
## Hypotheses:
## H1: mu_No  = mu_Yes
## H2: mu_No != mu_Yes
##
## Priors: P(H1) = 0.5  P(H2) = 0.5
##
## Results:
## BF[H1:H2] = 13.4039
## P(H1|data) = 0.9306
## P(H2|data) = 0.0694
##
## Posterior summaries for under H2:
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_No = 443, y_bar_No = 62.623, s_No = 20.3857
## n_Yes = 208, y_bar_Yes = 61.8077, s_Yes = 19.9083
## (Assuming Zellner-Siow Cauchy prior for difference in means)
## (Assuming independent Jeffrey's priors for overall mean and variance)
##
##
## Posterior Summaries
##                        2.5%          25%          50%          75%
## overall mean    60.5420961  61.64057090  62.20322073 6.279452e+01
## mu_No - mu_Yes  -2.5643373  -0.36310847   0.81764689 1.928377e+00
## sigma^2        368.2520323 394.33023170 409.28235192 4.254782e+02
## effect size     -0.1267592  -0.01784972   0.04030192 9.526388e-02
## n_0             33.3172332 377.44936723 913.84753962 1.822056e+03
##                       97.5%
## overall mean     63.9088331
## mu_No - mu_Yes    4.1275295
## sigma^2         458.2827053
## effect size       0.2042352
## n_0            4753.7880722
## 95% Cred. Int.: (-2.5643 , 4.1275)
```




```
# BF = 13.4; 95 = -2.59, 4.10
# Though the Bayes factor provides positive evidence that films released during summe
r season and films not released during summer season have on average different audien
ce scores, the credible intervals includes the value 0 therefore suggesting that ther
e is no significant difference in audience scores between the groups.

# Based upon later model selection; EDA on imdb rating and critics score
ggplot(data = movies_manipulate, aes(x = imdb_rating, y = audience_score)) + geom_poi
nt() + labs(x = "IMDB Rating", y = "Audience Score")
```

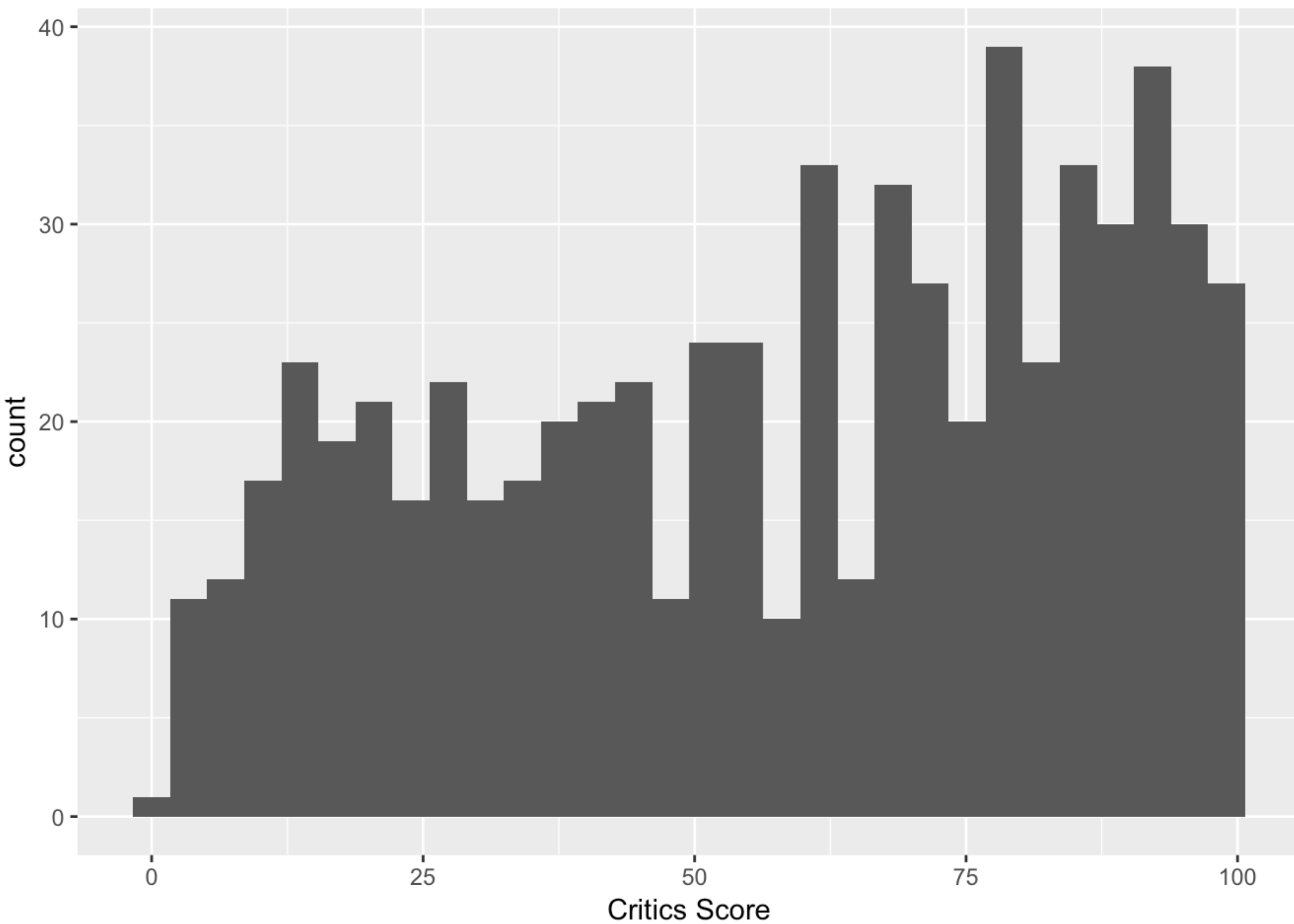



```
ggplot(data = movies_manipulate, aes(x = critics_score, y = audience_score)) + geom_p
oint() + labs(x = "Critic's Score", y = "Audience Score")
```



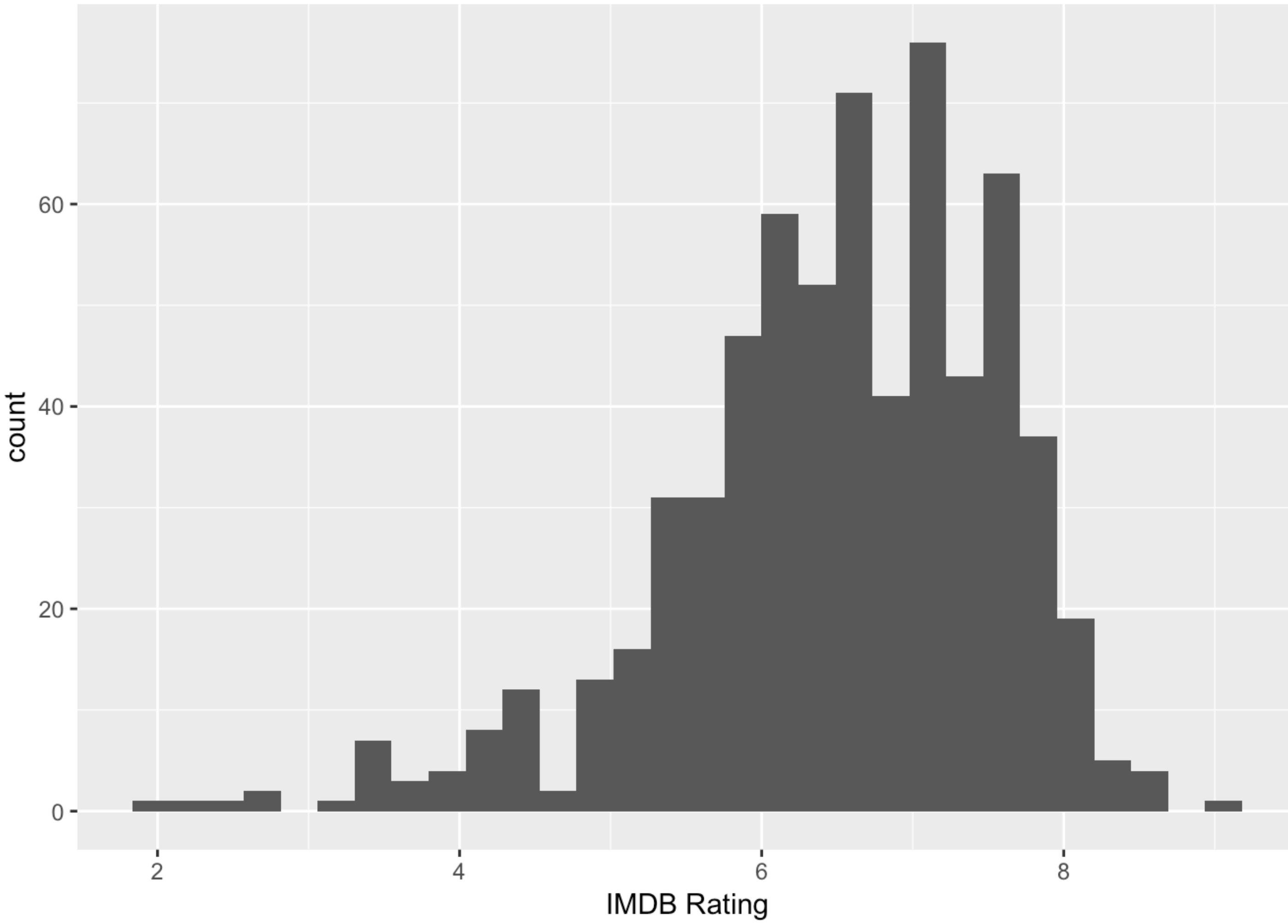

```
# Simply based upon this visualization it looks as if there may be a linear relations
hip between both the variables IMDB rating and Critics score and audience score respe
ctively.

ggplot(data = movies_manipulate, aes(x = critics_score)) + geom_histogram() + labs(x
= "Critics Score")
```

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




```
# There appears to be a uniform distribution of critics scores.
ggplot(data = movies_manipulate, aes(x = imdb_rating)) + geom_histogram() + labs(x =
"IMDB Rating")
```

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




```
# There appeats to be a slightly left-skewed distribution of IMDB ratings though the
data appears slightly normal.
```

Part 4: Modeling

```
# Variables of Interest
audience_score_var <- movies_manipulate %>%
  dplyr::select(audience_score, feature_film, drama, runtime, mpaa_rating_R, thtr_rel
_year, oscar_season, summer_season, imdb_rating, imdb_num_votes, critics_score, best_
pic_nom, best_pic_win, best_actor_win, best_actress_win, best_dir_win, top200_box)
audience_score_var <- na.omit(audience_score_var)
```

```
# Model
audience_score_var_freq <- lm(audience_score ~ ., data = audience_score_var)
tidy(audience_score_var_freq)
```

```
## # A tibble: 17 x 5
##    term                   estimate   std.error statistic  p.value
##    <chr>                     <dbl>       <dbl>     <dbl>    <dbl>
##  1 (Intercept)         124.         77.5            1.61  1.09e- 1
##  2 feature_filmYes      -2.25        1.69          -1.33  1.83e- 1
##  3 dramaYes              1.29        0.877          1.47  1.41e- 1
##  4 runtime              -0.0561      0.0242        -2.32  2.04e- 2
##  5 mpaa_rating_RYes     -1.44        0.813         -1.78  7.60e- 2
##  6 thtr_rel_year        -0.0766      0.0383        -2.00  4.63e- 2
##  7 oscar_seasonYes      -0.533       0.997         -0.535 5.93e- 1
##  8 summer_seasonYes      0.911       0.949          0.959 3.38e- 1
##  9 imdb_rating          14.7         0.607         24.3   2.03e-92
## 10 imdb_num_votes        0.00000723  0.00000452     1.60  1.10e- 1
## 11 critics_score         0.0575      0.0222         2.59  9.73e- 3
## 12 best_pic_nomyes       5.32        2.63           2.02  4.33e- 2
## 13 best_pic_winyes      -3.21        4.61          -0.697 4.86e- 1
## 14 best_actor_winyes    -1.54        1.18          -1.31  1.91e- 1
## 15 best_actress_winyes  -2.20        1.30          -1.69  9.23e- 2
## 16 best_dir_winyes      -1.23        1.73          -0.713 4.76e- 1
## 17 top200_boxyes         0.848       2.78           0.305 7.61e- 1
```

```
summary(audience_score_var_freq)
```

```
##
## Call:
## lm(formula = audience_score ~ ., data = audience_score_var)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.594  -6.156   0.157   5.909  53.125
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.244e+02  7.749e+01   1.606  0.10886
## feature_filmYes     -2.248e+00  1.687e+00  -1.332  0.18323
## dramaYes             1.292e+00  8.766e-01   1.474  0.14087
## runtime             -5.614e-02  2.415e-02  -2.324  0.02042 *
## mpaa_rating_RYes    -1.444e+00  8.127e-01  -1.777  0.07598 .
## thtr_rel_year       -7.657e-02  3.835e-02  -1.997  0.04628 *
## oscar_seasonYes     -5.333e-01  9.967e-01  -0.535  0.59280
## summer_seasonYes     9.106e-01  9.493e-01   0.959  0.33778
## imdb_rating          1.472e+01  6.067e-01  24.258  < 2e-16 ***
## imdb_num_votes       7.234e-06  4.523e-06   1.600  0.11019
## critics_score        5.748e-02  2.217e-02   2.593  0.00973 **
## best_pic_nomyes      5.321e+00  2.628e+00   2.025  0.04330 *
## best_pic_winyes     -3.212e+00  4.610e+00  -0.697  0.48624
## best_actor_winyes   -1.544e+00  1.179e+00  -1.310  0.19068
## best_actress_winyes -2.198e+00  1.304e+00  -1.686  0.09229 .
## best_dir_winyes     -1.231e+00  1.728e+00  -0.713  0.47630
## top200_boxyes        8.478e-01  2.782e+00   0.305  0.76067
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.975 on 633 degrees of freedom
## Multiple R-squared:  0.763,  Adjusted R-squared:  0.757
## F-statistic: 127.3 on 16 and 633 DF,  p-value: < 2.2e-16
```

```
# Statistically significant predictive variables: runtime, year released, imdb rating
, critics score and whether the picture was nominted for an acedemy award
# NB Imdb rating had the lowest p-value followed by critics score
# Simple frequentist linear regression results this just gives a basis to compare Bay
esian Results with

# Bayesian Regression using ZS-null prior as done in week 5 lab.
bma_ZS <- bas.lm(audience_score ~ ., data = audience_score_var,
                            prior = "ZS-null",
                            modelprior = uniform())
summary(bma_ZS)
```

```
##                     P(B != 0 | Y)  model 1     model 2     model 3
## Intercept              1.00000000   1.0000   1.0000000   1.0000000
## feature_filmYes        0.06796946   0.0000   0.0000000   0.0000000
## dramaYes               0.04591717   0.0000   0.0000000   0.0000000
## runtime                0.46420058   0.0000   1.0000000   0.0000000
## mpaa_rating_RYes       0.20274450   0.0000   0.0000000   0.0000000
## thtr_rel_year          0.09499813   0.0000   0.0000000   0.0000000
## oscar_seasonYes        0.07749797   0.0000   0.0000000   0.0000000
## summer_seasonYes       0.08335823   0.0000   0.0000000   0.0000000
## imdb_rating            1.00000000   1.0000   1.0000000   1.0000000
## imdb_num_votes         0.06115184   0.0000   0.0000000   0.0000000
## critics_score          0.88078574   1.0000   1.0000000   1.0000000
## best_pic_nomyes        0.13684669   0.0000   0.0000000   0.0000000
## best_pic_winyes        0.04215714   0.0000   0.0000000   0.0000000
## best_actor_winyes      0.14642057   0.0000   0.0000000   1.0000000
## best_actress_winyes    0.14444247   0.0000   0.0000000   0.0000000
## best_dir_winyes        0.06936269   0.0000   0.0000000   0.0000000
## top200_boxyes          0.04998566   0.0000   0.0000000   0.0000000
## BF                             NA   1.0000   0.8702806   0.2236679
## PostProbs                      NA   0.1388   0.1208000   0.0311000
## R2                             NA   0.7525   0.7549000   0.7539000
## dim                            NA   3.0000   4.0000000   4.0000000
## logmarg                        NA 443.9495 443.8105657 442.4519125
##                         model 4     model 5
## Intercept             1.0000000   1.0000000
## feature_filmYes       0.0000000   0.0000000
## dramaYes              0.0000000   0.0000000
## runtime               0.0000000   1.0000000
## mpaa_rating_RYes      1.0000000   1.0000000
## thtr_rel_year         0.0000000   0.0000000
## oscar_seasonYes       0.0000000   0.0000000
## summer_seasonYes      0.0000000   0.0000000
## imdb_rating           1.0000000   1.0000000
## imdb_num_votes        0.0000000   0.0000000
## critics_score         1.0000000   1.0000000
## best_pic_nomyes       0.0000000   0.0000000
## best_pic_winyes       0.0000000   0.0000000
## best_actor_winyes     0.0000000   0.0000000
## best_actress_winyes   0.0000000   0.0000000
## best_dir_winyes       0.0000000   0.0000000
## top200_boxyes         0.0000000   0.0000000
## BF                    0.2217602   0.2055844
## PostProbs             0.0308000   0.0285000
## R2                    0.7539000   0.7563000
## dim                   4.0000000   5.0000000
## logmarg             442.4433468 442.3676066
```

```
coef_bma <- coefficients(bma_ZS)
confint(coef_bma)
```

```
##                             2.5%        97.5%          beta
## Intercept            6.158491e+01 6.312665e+01  6.234769e+01
## feature_filmYes     -1.287607e+00 0.000000e+00 -1.081424e-01
## dramaYes             0.000000e+00 0.000000e+00  1.791132e-02
## runtime             -8.332129e-02 0.000000e+00 -2.534532e-02
## mpaa_rating_RYes    -2.126881e+00 1.803519e-04 -3.073124e-01
## thtr_rel_year       -5.603212e-02 0.000000e+00 -4.771859e-03
## oscar_seasonYes     -1.033925e+00 0.000000e+00 -8.221176e-02
## summer_seasonYes    -2.994753e-03 1.130680e+00  8.984037e-02
## imdb_rating          1.369330e+01 1.659671e+01  1.496477e+01
## imdb_num_votes      -1.198173e-09 1.673640e-06  2.242282e-07
## critics_score        0.000000e+00 1.057024e-01  6.227229e-02
## best_pic_nomyes     -2.549020e-03 5.011251e+00  5.323609e-01
## best_pic_winyes      0.000000e+00 0.000000e+00 -1.090880e-02
## best_actor_winyes   -2.568386e+00 0.000000e+00 -2.897874e-01
## best_actress_winyes -2.884651e+00 6.602460e-03 -3.146149e-01
## best_dir_winyes     -1.558993e+00 0.000000e+00 -1.227043e-01
## top200_boxyes        0.000000e+00 0.000000e+00  9.039557e-02
## attr(,"Probability")
## [1] 0.95
## attr(,"class")
## [1] "confint.bas"
```

```
# Most likely model (P = 0.1388) includes: Intercept, IMDB Rating, Critics Score

# Which Model to Use
BMA <- predict(bma_ZS, estimator = "BMA", se.fit = TRUE)
BPM <- predict(bma_ZS, estimator = "BPM", se.fit = TRUE)
variable.names(BPM) # intercept, runtime, imdb rating, critics score
```

```
## [1] "Intercept"     "runtime"       "imdb_rating"   "critics_score"
```

```
HPM <- predict(bma_ZS, estimator = "HPM", se.fit = TRUE)
variable.names(HPM) # intercept, imdb rating, critics score
```

```
## [1] "Intercept"     "imdb_rating"   "critics_score"
```

```
MPM <- predict(bma_ZS, estimator = "MPM", se.fit = TRUE)
variable.names(MPM) # intercept, imdb rating, critics score
```

```
## [1] "Intercept"     "imdb_rating"   "critics_score"
```

```
coef_bma$conditionalmeans[BPM$best,]
```

```
##          Intercept     feature_filmYes            dramaYes 
##       6.234769e+01       -3.006730e+01        8.840792e+00 
##            runtime    mpaa_rating_RYes       thtr_rel_year 
##       4.728376e-02        4.521207e-01       -2.833207e-01 
##     oscar_seasonYes    summer_seasonYes         imdb_rating 
##       3.505303e-01        1.001027e+00        0.000000e+00 
##      imdb_num_votes       critics_score     best_pic_nomyes 
##       6.337589e-05        0.000000e+00        0.000000e+00 
##     best_pic_winyes   best_actor_winyes best_actress_winyes 
##      -3.367846e+00       -5.371803e-01       -1.769083e+00 
##     best_dir_winyes       top200_boxyes 
##       2.316835e+00        0.000000e+00
```

```
coef_bma$conditionalsd[BPM$best,]
```

```
##          Intercept     feature_filmYes            dramaYes 
##       6.804999e-01        2.538711e+00        1.459679e+00 
##            runtime    mpaa_rating_RYes       thtr_rel_year 
##       4.125407e-02        1.390988e+00        6.470565e-02 
##     oscar_seasonYes    summer_seasonYes         imdb_rating 
##       1.707232e+00        1.629747e+00        0.000000e+00 
##      imdb_num_votes       critics_score     best_pic_nomyes 
##       6.967107e-06        0.000000e+00        0.000000e+00 
##     best_pic_winyes   best_actor_winyes best_actress_winyes 
##       7.292645e+00        2.023775e+00        2.238431e+00 
##     best_dir_winyes       top200_boxyes 
##       2.964864e+00        0.000000e+00
```

```
coef_bma$conditionalmeans[HPM$best,]
```

```
##          Intercept     feature_filmYes            dramaYes 
##        62.34769231         -2.07202300          0.89520178 
##            runtime    mpaa_rating_RYes       thtr_rel_year 
##        -0.05311811          0.00000000         -0.06067619 
##     oscar_seasonYes    summer_seasonYes         imdb_rating 
##         0.00000000          1.27584824         14.85788275 
##      imdb_num_votes       critics_score     best_pic_nomyes 
##         0.00000000          0.05741522          5.17974190 
##     best_pic_winyes   best_actor_winyes best_actress_winyes 
##         0.00000000          0.00000000         -2.20909892 
##     best_dir_winyes       top200_boxyes 
##        -1.55509439          0.00000000
```

```
coef_bma$conditionalsd[HPM$best,]
```

```
##           Intercept     feature_filmYes            dramaYes 
##          0.39202101          1.59028105          0.85845797 
##             runtime    mpaa_rating_RYes       thtr_rel_year 
##          0.02280168          0.00000000          0.03656349 
##     oscar_seasonYes    summer_seasonYes         imdb_rating 
##          0.00000000          0.85055972          0.58455686 
##      imdb_num_votes       critics_score     best_pic_nomyes 
##          0.00000000          0.02212119          2.32607865 
##     best_pic_winyes   best_actor_winyes best_actress_winyes 
##          0.00000000          0.00000000          1.29684984 
##     best_dir_winyes       top200_boxyes 
##          1.65045573          0.00000000
```

```
MPM_coef = bas.lm(audience_score ~ ., data = audience_score_var,
           prior="ZS-null",
           modelprior = uniform(),
           bestmodel = bma_ZS$probne0 > .5,
           n.models=1)
coef(MPM_coef)
```

```
## 
##  Marginal Posterior Summaries of Coefficients: 
## 
##  Using  BMA 
## 
##  Based on the top  1 models 
##                      post mean  post SD   post p(B != 0)
## Intercept            62.34769    0.39549   1.00000      
## feature_filmYes       0.00000    0.00000   0.00000      
## dramaYes              0.00000    0.00000   0.00000      
## runtime               0.00000    0.00000   0.00000      
## mpaa_rating_RYes      0.00000    0.00000   0.00000      
## thtr_rel_year         0.00000    0.00000   0.00000      
## oscar_seasonYes       0.00000    0.00000   0.00000      
## summer_seasonYes      0.00000    0.00000   0.00000      
## imdb_rating          14.64833    0.56593   1.00000      
## imdb_num_votes        0.00000    0.00000   0.00000      
## critics_score         0.07316    0.02161   1.00000      
## best_pic_nomyes       0.00000    0.00000   0.00000      
## best_pic_winyes       0.00000    0.00000   0.00000      
## best_actor_winyes     0.00000    0.00000   0.00000      
## best_actress_winyes   0.00000    0.00000   0.00000      
## best_dir_winyes       0.00000    0.00000   0.00000      
## top200_boxyes         0.00000    0.00000   0.00000
```

```
ci_BMA <- confint(BMA, parm = "pred")
opt_BMA <- which.max(BMA$fit)
ci_BMA[opt_BMA,]
```

```
##      2.5%     97.5%      pred 
##  78.89563 119.33803  99.76999
```

```
ci_HPM <- confint(HPM, parm = "pred")
opt_HPM <- which.max(HPM$fit)
ci_HPM[opt_HPM,]
```

```
##      2.5%     97.5%      pred 
##  82.06589 121.87606 101.97097
```

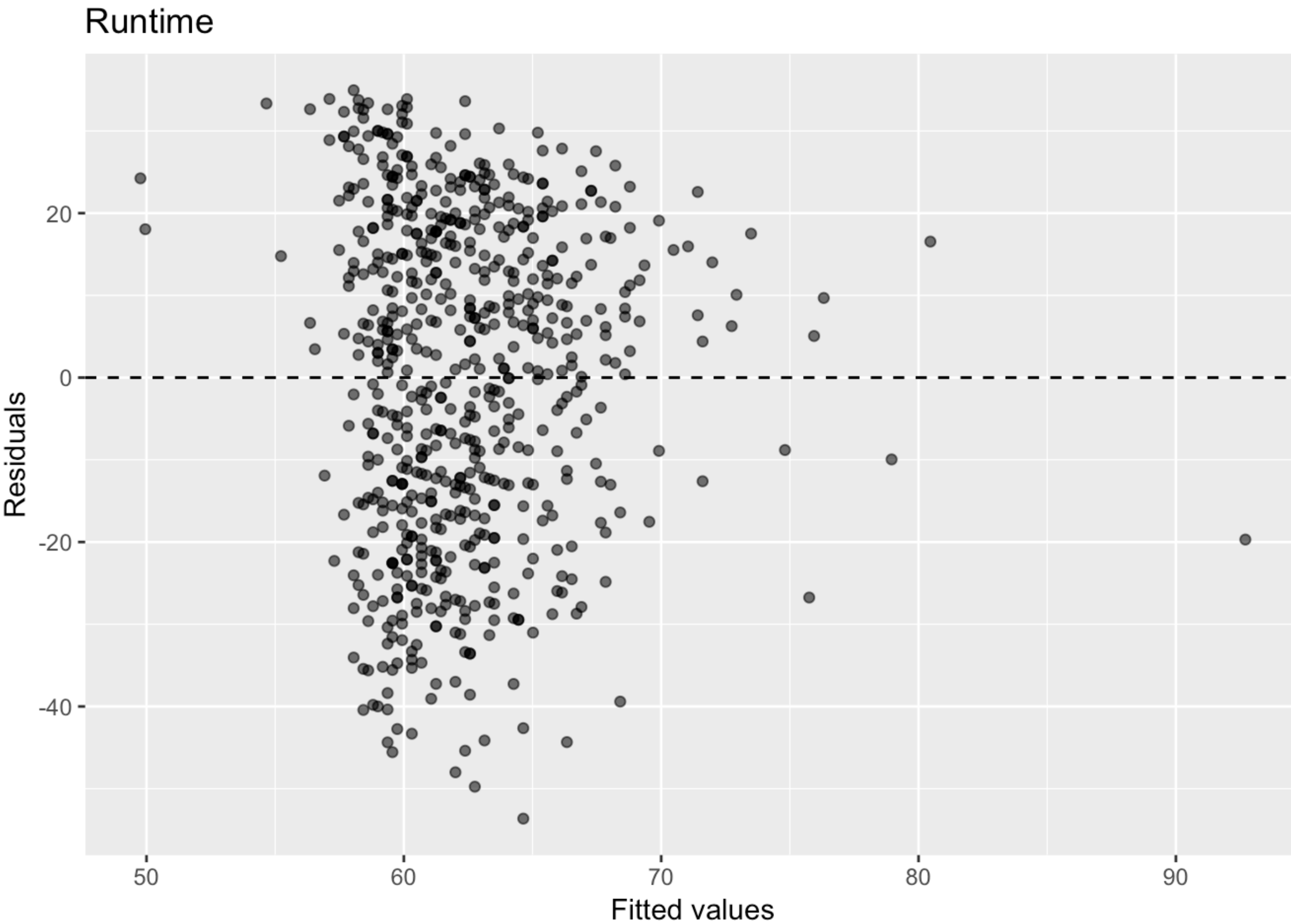
```
ci_MPM <- confint(MPM, parm = "pred")
opt_MPM <- which.max(MPM$fit)
ci_MPM[opt_MPM,]
```

```
##      2.5%     97.5%      pred 
##  82.06589 121.87606 101.97097
```

```
ci_BPM <- confint(BPM, parm = "pred")
opt_BPM <- which.max(BPM$fit)
ci_BPM[opt_BPM,]
```

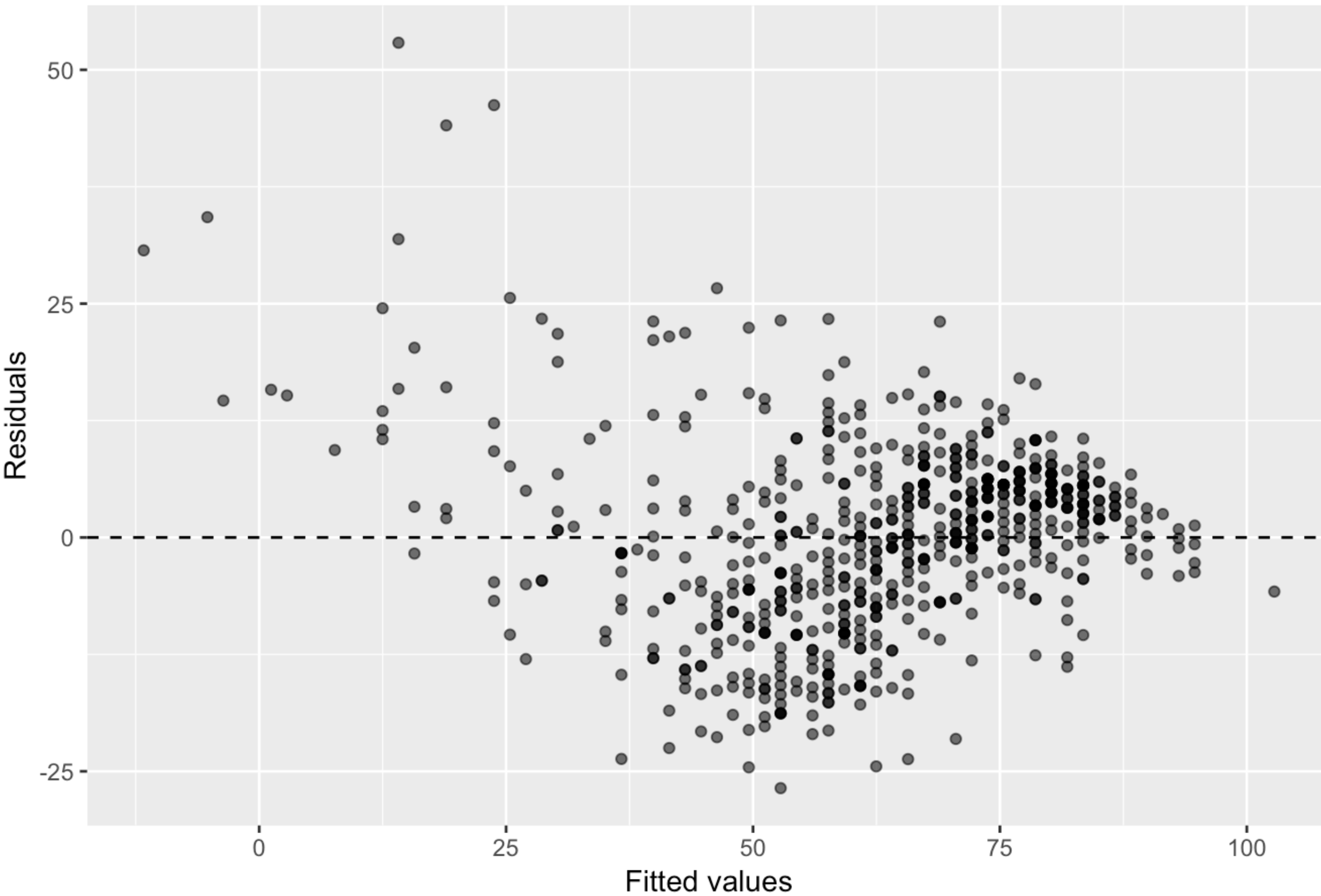
```
##      2.5%     97.5%      pred 
##  77.36827 117.60091  97.48459
```

```
# Variables of interest for each model (BMA, MPM, HPM, MPM) listed above
# CI refers to maximum audience score of a picture based on variables from respsctive
models (there is a 95% probability that the top rated movie is scored between L and U
)
# BMA:  Intercept: 62.3 (95% CI: 61.6-63.1); IMDB Rating: 15.0 (95% CI: 13.7-16.5); c
ritics score: 0.062 (95% CI: 0.0-0.11)
# BMA Model: 99.8 (95% CI: 79.6-120.7)

# BPM: Intercept: 62.3 (sd: 0.55); Runtime: 0.068 (sd: 0.032); IMDB Rating: 0.0 (sd:
0.0); critics score: 0.45 (sd: 0.22)
# BPM Model: 97.5 (95% CI: 77.4-117.6)

# HPM Intercept: 62.3 (sd: 0.54); IMDB Rating: 0.0 (sd: 0.0); critics score: 0.42 (sd
: 0.22)
# HPM Model: 102.0 (95% CI: 82.0-121.9)

# MPM Intercept: 62.3 (sd: 0.40); IMDB Rating: 14.65 (sd: 0.57); critics score: 0.073
(sd: 0.20)
# MPM Model: 102.0 (95% CI: 82.0-121.9)

# Prior to model selection diagnostics will be performed on continuous variables; che
cks assumptions for Bayesian Regression are true
# Simple Linear Model (Non-Bayesian)
lm_audience_runtime <- lm(audience_score ~ runtime, data = movies_manipulate)
lm_audience_imdb_rating <- lm(audience_score ~ imdb_rating, data = movies_manipulate)
lm_audience_critics_score <- lm(audience_score ~ critics_score, data = movies_manipul
ate)
# Augment; obtaint residual and fitted values
lm_audience_runtime_aug <- augment(lm_audience_runtime)
lm_audience_imdb_rating_aug <- augment(lm_audience_imdb_rating)
lm_audience_critics_score_aug <- augment(lm_audience_critics_score)
# Linearity and Constant Variance
ggplot(data = lm_audience_runtime_aug, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Fitted values", y = "Residuals", title = "Runtime")
```

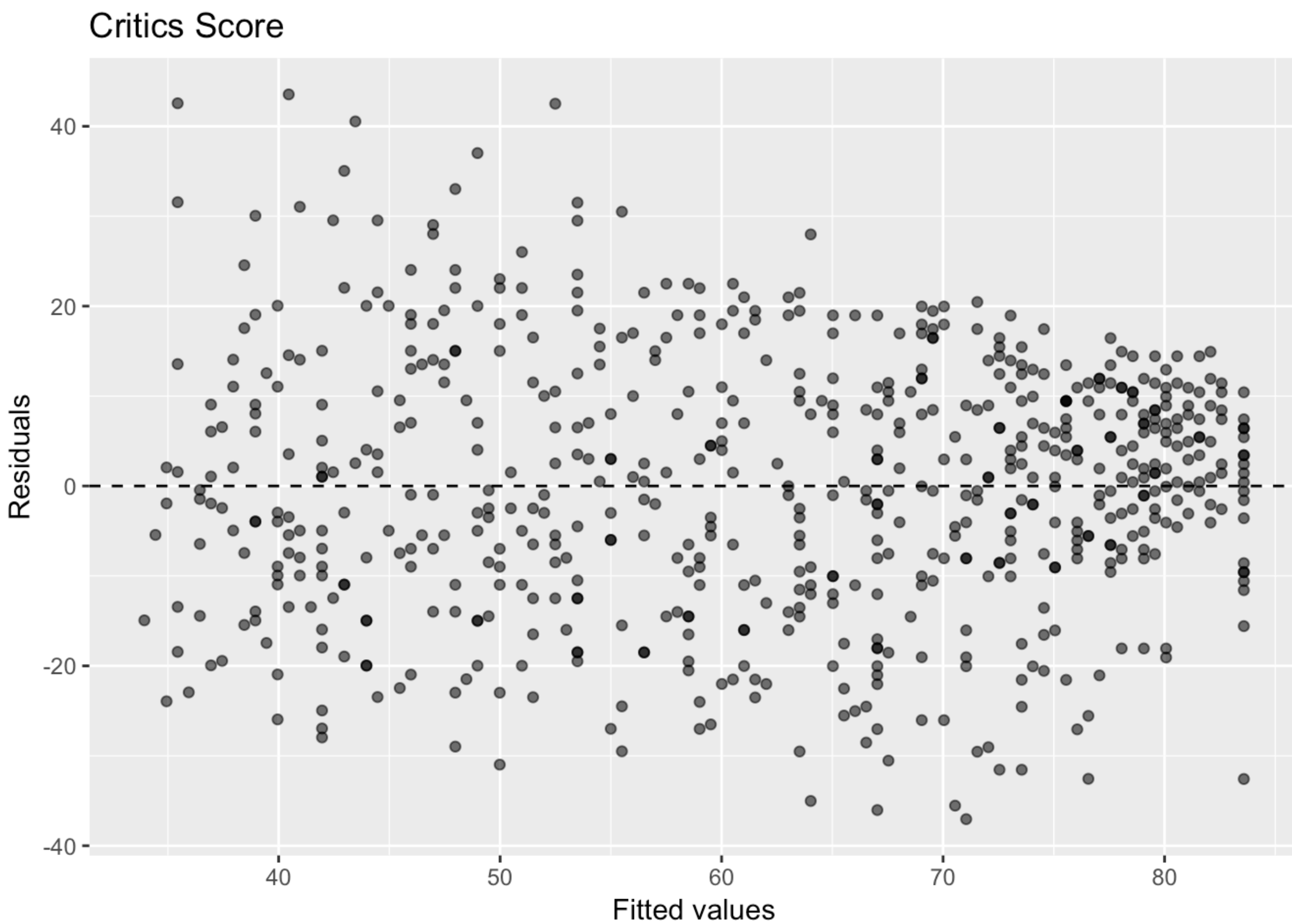



```
# The distribution of residuals about the value 0 is not random
ggplot(data = lm_audience_imdb_rating_aug, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Fitted values", y = "Residuals", title = "IMDB Rating")
```

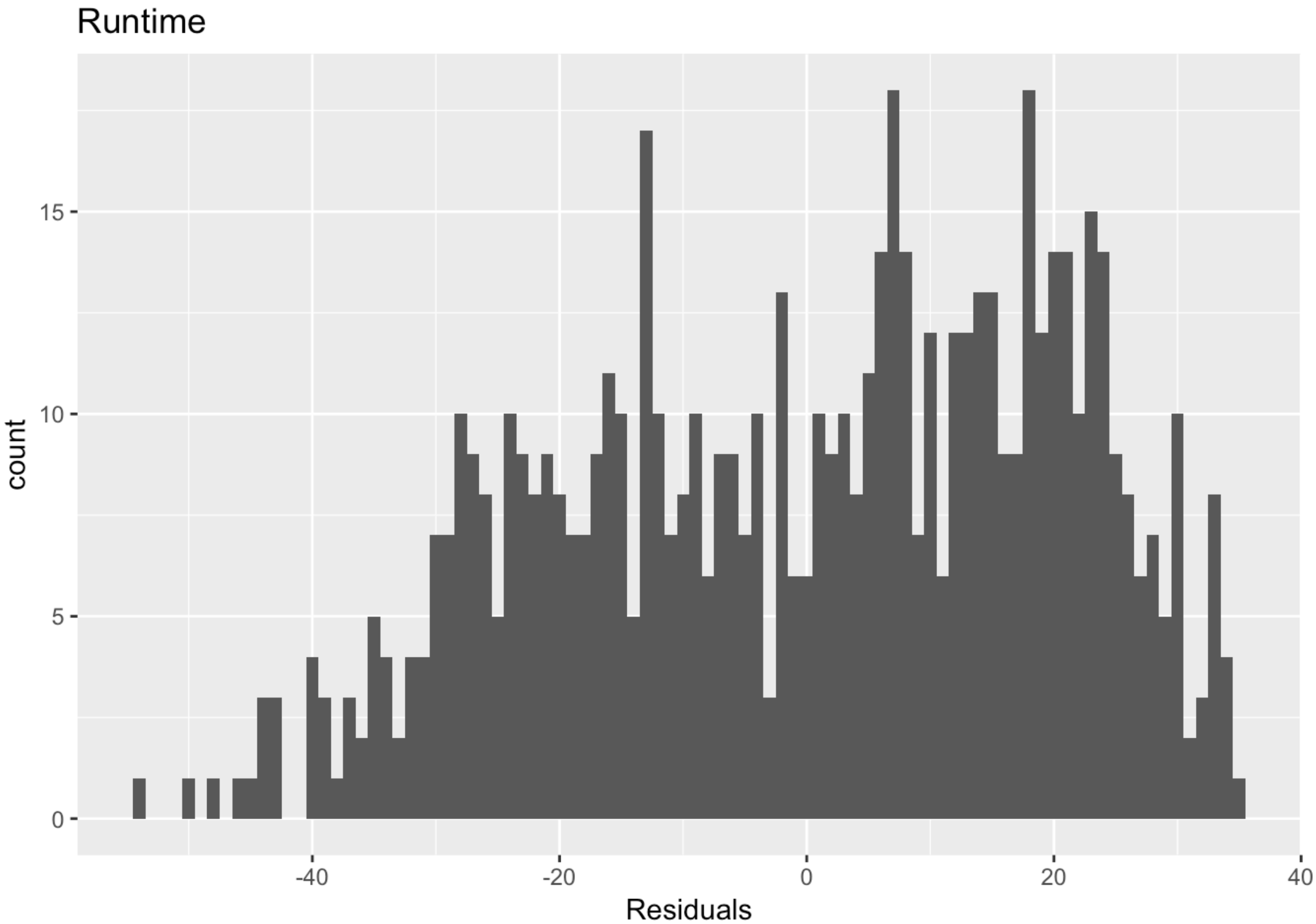
IMDB Rating



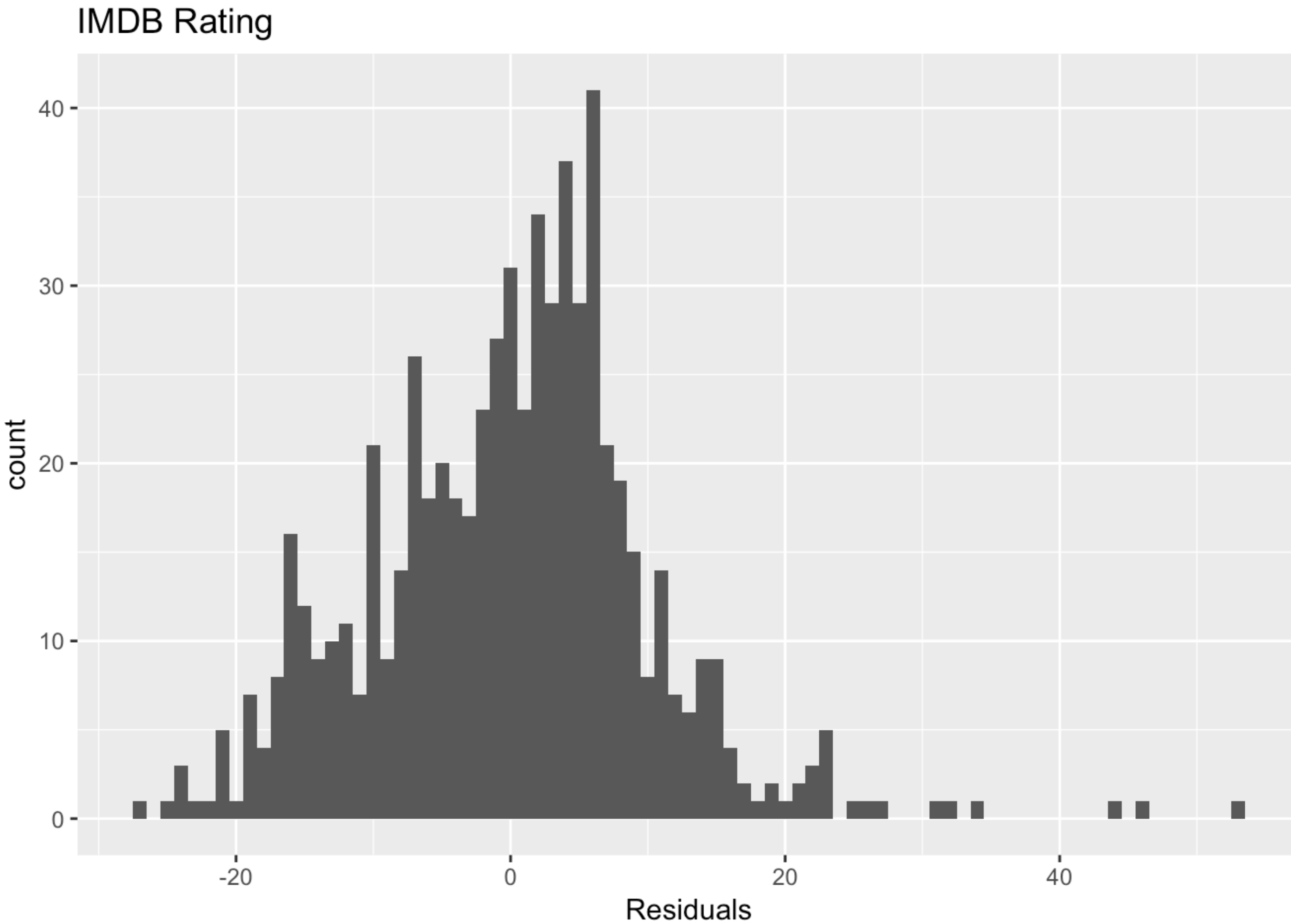

```
# The distribution of residuals about the value 0 is not entirely random, but has les
s obvious structure than the residual values from runtime
ggplot(data = lm_audience_critics_score_aug, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(x = "Fitted values", y = "Residuals", title = "Critics Score")
```



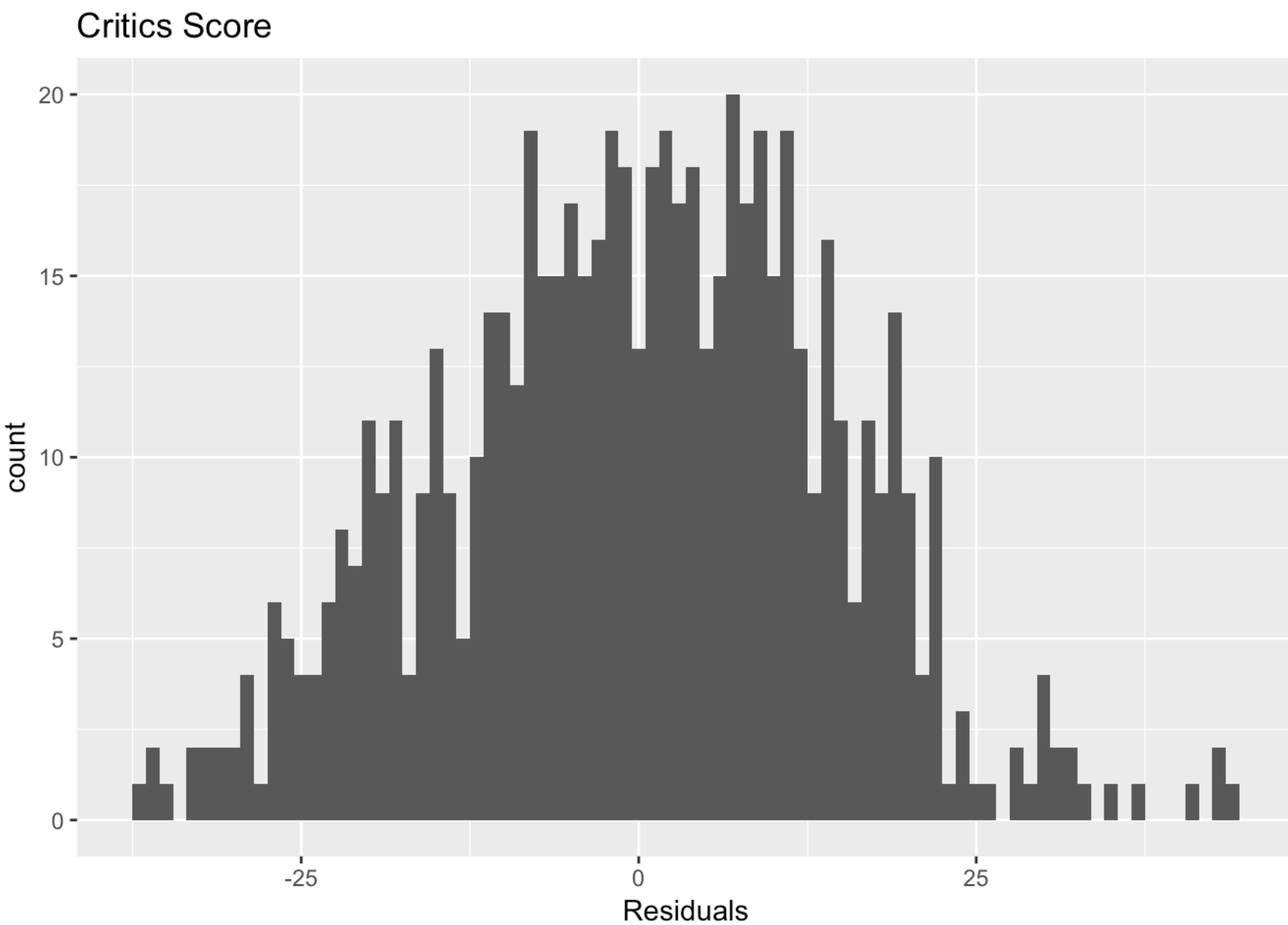

```
# The distribution of residuals about the value 0 random suggesting linearity
# Normality
ggplot(data = lm_audience_runtime_aug, aes(x = .resid)) +
  geom_histogram(binwidth = 1) +
  labs(x = "Residuals", title = "Runtime")
```




```
# The residuals do not appear to be normally distributed
ggplot(data = lm_audience_imdb_rating_aug, aes(x = .resid)) +
  geom_histogram(binwidth = 1) +
  labs(x = "Residuals", title = "IMDB Rating")
```




```r
# The residuals appears to be roughly normal
ggplot(data = lm_audience_critics_score_aug, aes(x = .resid)) +
  geom_histogram(binwidth = 1) +
  labs(x = "Residuals", title = "Critics Score")
```




```
# The residuals show roughly a normal shape though the distribution is broad


# Final Model
# Variables included in all four models (BMA, BPM, HPM, MPM) are imdb rating and crit
ics score. Based on the parsimonious theory (Occam's razor) it makes sense to choose
a model that only has those variables, which removes BPM from final model selection.
Runtime, the variable only found in the BPM model, also does not appear to follow the
requirements for linearity and normality of residuals, which supports its removal fro
m final model.
# Using BIC to confirm model variable selection above is correctly done:
# BIC full model
BIC(audience_score_var_freq) # BIC = 4934.145
```

```
## [1] 4934.145
```

```
lm_post_models <- lm(audience_score ~ imdb_rating + critics_score, data = audience_sc
ore_var)
BIC(lm_post_models) # BIC = 4871.63
```

```
## [1] 4871.629
```

```
# AIC
AIC_Step_AS <- stepAIC(audience_score_var_freq, direction = "backward")
```

```
## Start:  AIC=3006.94
## audience_score ~ feature_film + drama + runtime + mpaa_rating_R +
##     thtr_rel_year + oscar_season + summer_season + imdb_rating +
##     imdb_num_votes + critics_score + best_pic_nom + best_pic_win +
##     best_actor_win + best_actress_win + best_dir_win + top200_box
##
##                    Df Sum of Sq    RSS    AIC
## - top200_box        1         9  62999 3005.0
## - oscar_season      1        28  63018 3005.2
## - best_pic_win      1        48  63038 3005.4
## - best_dir_win      1        51  63040 3005.5
## - summer_season     1        92  63081 3005.9
## - best_actor_win    1       171  63160 3006.7
## - feature_film      1       177  63166 3006.8
## <none>                           62990 3006.9
## - drama             1       216  63206 3007.2
## - imdb_num_votes    1       255  63244 3007.6
## - best_actress_win  1       283  63273 3007.9
## - mpaa_rating_R     1       314  63304 3008.2
## - thtr_rel_year     1       397  63386 3009.0
## - best_pic_nom      1       408  63398 3009.1
## - runtime           1       538  63527 3010.5
## - critics_score     1       669  63659 3011.8
## - imdb_rating       1     58556 121545 3432.2
##
## Step:  AIC=3005.04
## audience_score ~ feature_film + drama + runtime + mpaa_rating_R +
##     thtr_rel_year + oscar_season + summer_season + imdb_rating +
##     imdb_num_votes + critics_score + best_pic_nom + best_pic_win +
##     best_actor_win + best_actress_win + best_dir_win
##
##                    Df Sum of Sq    RSS    AIC
## - oscar_season      1        26  63025 3003.3
## - best_pic_win      1        49  63047 3003.5
## - best_dir_win      1        52  63051 3003.6
## - summer_season     1        94  63093 3004.0
## - best_actor_win    1       169  63168 3004.8
## - feature_film      1       176  63175 3004.8
## <none>                           62999 3005.0
## - drama             1       214  63213 3005.2
## - best_actress_win  1       279  63278 3005.9
## - imdb_num_votes    1       302  63301 3006.1
## - mpaa_rating_R     1       330  63329 3006.4
```

```
## - best_pic_nom      1       404  63403 3007.2
## - thtr_rel_year     1       415  63414 3007.3
## - runtime           1       535  63534 3008.5
## - critics_score     1       681  63680 3010.0
## - imdb_rating       1     58606 121604 3430.5
##
## Step:  AIC=3003.31
## audience_score ~ feature_film + drama + runtime + mpaa_rating_R +
##     thtr_rel_year + summer_season + imdb_rating + imdb_num_votes +
##     critics_score + best_pic_nom + best_pic_win + best_actor_win +
##     best_actress_win + best_dir_win
##
##                    Df Sum of Sq    RSS    AIC
## - best_pic_win      1        46  63071 3001.8
## - best_dir_win      1        56  63081 3001.9
## - best_actor_win    1       174  63200 3003.1
## - summer_season     1       177  63202 3003.1
## - feature_film      1       182  63207 3003.2
## <none>                           63025 3003.3
## - drama             1       222  63247 3003.6
## - best_actress_win  1       281  63307 3004.2
## - imdb_num_votes    1       302  63328 3004.4
## - mpaa_rating_R     1       329  63354 3004.7
## - best_pic_nom      1       387  63412 3005.3
## - thtr_rel_year     1       410  63436 3005.5
## - runtime           1       587  63613 3007.3
## - critics_score     1       679  63704 3008.3
## - imdb_rating       1     58603 121628 3428.6
##
## Step:  AIC=3001.78
## audience_score ~ feature_film + drama + runtime + mpaa_rating_R +
##     thtr_rel_year + summer_season + imdb_rating + imdb_num_votes +
##     critics_score + best_pic_nom + best_actor_win + best_actress_win +
##     best_dir_win
##
##                    Df Sum of Sq    RSS    AIC
## - best_dir_win      1        94  63165 3000.7
## - best_actor_win    1       163  63234 3001.5
## - feature_film      1       171  63242 3001.5
## - summer_season     1       174  63245 3001.6
## <none>                           63071 3001.8
## - drama             1       220  63291 3002.0
## - imdb_num_votes    1       271  63342 3002.6
## - best_actress_win  1       294  63365 3002.8
## - mpaa_rating_R     1       330  63401 3003.2
## - best_pic_nom      1       342  63414 3003.3
## - thtr_rel_year     1       397  63468 3003.9
## - runtime           1       586  63657 3005.8
## - critics_score     1       680  63751 3006.8
## - imdb_rating       1     58858 121929 3428.2
```

```
##
## Step:  AIC=3000.75
## audience_score ~ feature_film + drama + runtime + mpaa_rating_R +
##     thtr_rel_year + summer_season + imdb_rating + imdb_num_votes +
##     critics_score + best_pic_nom + best_actor_win + best_actress_win
##
##                    Df Sum of Sq    RSS    AIC
## - summer_season     1       167  63332 3000.5
## - best_actor_win    1       171  63336 3000.5
## - feature_film      1       183  63348 3000.6
## <none>                           63165 3000.7
## - drama             1       228  63394 3001.1
## - imdb_num_votes    1       247  63412 3001.3
## - best_actress_win  1       299  63464 3001.8
## - best_pic_nom      1       326  63491 3002.1
## - mpaa_rating_R     1       345  63510 3002.3
## - thtr_rel_year     1       368  63533 3002.5
## - critics_score     1       651  63816 3005.4
## - runtime           1       673  63839 3005.6
## - imdb_rating       1     58895 122061 3426.9
##
## Step:  AIC=3000.46
## audience_score ~ feature_film + drama + runtime + mpaa_rating_R +
##     thtr_rel_year + imdb_rating + imdb_num_votes + critics_score +
##     best_pic_nom + best_actor_win + best_actress_win
##
##                    Df Sum of Sq    RSS    AIC
## - feature_film      1       156  63488 3000.1
## <none>                           63332 3000.5
## - best_actor_win    1       195  63527 3000.5
## - drama             1       204  63536 3000.6
## - imdb_num_votes    1       260  63592 3001.1
## - best_pic_nom      1       297  63629 3001.5
## - best_actress_win  1       297  63629 3001.5
## - mpaa_rating_R     1       356  63688 3002.1
## - thtr_rel_year     1       361  63693 3002.2
## - runtime           1       690  64022 3005.5
## - critics_score     1       732  64064 3005.9
## - imdb_rating       1     58763 122095 3425.1
##
## Step:  AIC=3000.06
## audience_score ~ drama + runtime + mpaa_rating_R + thtr_rel_year +
##     imdb_rating + imdb_num_votes + critics_score + best_pic_nom +
##     best_actor_win + best_actress_win
##
##                    Df Sum of Sq    RSS    AIC
## - drama             1       121  63609 2999.3
## - imdb_num_votes    1       173  63661 2999.8
## <none>                           63488 3000.1
## - best_actor_win    1       219  63706 3000.3
```

```
## - thtr_rel_year     1       277  63765 3000.9
## - best_pic_nom      1       291  63778 3001.0
## - best_actress_win  1       306  63794 3001.2
## - mpaa_rating_R     1       453  63941 3002.7
## - runtime           1       715  64203 3005.3
## - critics_score     1       875  64363 3007.0
## - imdb_rating       1     63189 126677 3447.1
##
## Step:  AIC=2999.3
## audience_score ~ runtime + mpaa_rating_R + thtr_rel_year + imdb_rating +
##     imdb_num_votes + critics_score + best_pic_nom + best_actor_win +
##     best_actress_win
##
##                    Df Sum of Sq    RSS    AIC
## - imdb_num_votes    1       148  63757 2998.8
## <none>                           63609 2999.3
## - best_actor_win    1       209  63818 2999.4
## - thtr_rel_year     1       272  63881 3000.1
## - best_actress_win  1       274  63883 3000.1
## - best_pic_nom      1       307  63916 3000.4
## - mpaa_rating_R     1       391  64000 3001.3
## - runtime           1       631  64240 3003.7
## - critics_score     1       916  64525 3006.6
## - imdb_rating       1     63434 127043 3447.0
##
## Step:  AIC=2998.81
## audience_score ~ runtime + mpaa_rating_R + thtr_rel_year + imdb_rating +
##     critics_score + best_pic_nom + best_actor_win + best_actress_win
##
##                    Df Sum of Sq    RSS    AIC
## <none>                           63757 2998.8
## - thtr_rel_year     1       201  63958 2998.9
## - best_actor_win    1       219  63976 2999.0
## - best_actress_win  1       266  64023 2999.5
## - mpaa_rating_R     1       367  64124 3000.5
## - best_pic_nom      1       442  64199 3001.3
## - runtime           1       519  64276 3002.1
## - critics_score     1       879  64635 3005.7
## - imdb_rating       1     67356 131113 3465.4
```

```
# Using the variables from the model with the lowest AIC score
lm_post_AIC <- lm(audience_score ~ runtime + mpaa_rating_R + thtr_rel_year + imdb_rat
ing + critics_score + best_pic_nom + best_actor_win + best_actress_win, data = audien
ce_score_var)
BIC(lm_post_AIC) # BIC = 4890.2
```

```
## [1] 4890.199
```

```
# The BIC using the variables selected from the models aboves lowers the BIC score. S
tepwise model that yields the lowest AIC value, which is used in some model selection
processes, produced a model or group of variables that did not have a loweor BIC than
the model created from the above information.
# Penn State College of Human Health: "So what’s the bottom line? In general, it migh
t be best to use AIC and BIC together in model selection."

# Ultimately the BMA model is chosen as the best model as it has the least number of
variables, which themselves produce lowest BIC. It is chosen above the HPM and MPM mo
dels because those models have a predicted maximum value greater than 100, which is i
mpossible under the Rotten Tomato scoring rubric.

# For every 1 increase in IMDB rating audience score increases on average by 15 point
s; there is a 0.95 probability that for every increase in IMDB rating the audience sc
ore increases on average between 13.7 and 16.5 points; for every 1 increase in critic
s score the audience score increases on average by 0.062 points; there is a 0.95 prob
ability that for every 1 point increase in critics score the audience score increases
on average between 0 to 0.11 points.
```

Part 5: Prediction

```
# Manchester by the Sea
MS_b <- data.frame(77, 'Yes', 'Yes', 137, 'Yes', 2016, 'Yes', 'No', 7.8, 198685, 95,
'yes', 'no', 'yes', 'no', 'no', 'no')
colnames(MS_b) = colnames(audience_score_var)

MS_b
```

```
##   audience_score feature_film drama runtime mpaa_rating_R thtr_rel_year
## 1             77          Yes   Yes     137           Yes          2016
##   oscar_season summer_season imdb_rating imdb_num_votes critics_score
## 1          Yes            No         7.8         198685            95
##   best_pic_nom best_pic_win best_actor_win best_actress_win best_dir_win
## 1          yes           no            yes               no           no
##   top200_box
## 1         no
```

```
summary(MS_b)
```

```
##  audience_score feature_film drama      runtime    mpaa_rating_R
##  Min.   :77     Yes:1        Yes:1   Min.   :137   Yes:1
##  1st Qu.:77                          1st Qu.:137
##  Median :77                          Median :137
##  Mean   :77                          Mean   :137
##  3rd Qu.:77                          3rd Qu.:137
##  Max.   :77                          Max.   :137
##  thtr_rel_year  oscar_season summer_season  imdb_rating  imdb_num_votes
##  Min.   :2016   Yes:1        No:1          Min.   :7.8   Min.   :198685
##  1st Qu.:2016                              1st Qu.:7.8   1st Qu.:198685
##  Median :2016                              Median :7.8   Median :198685
##  Mean   :2016                              Mean   :7.8   Mean   :198685
##  3rd Qu.:2016                              3rd Qu.:7.8   3rd Qu.:198685
##  Max.   :2016                              Max.   :7.8   Max.   :198685
##  critics_score best_pic_nom best_pic_win best_actor_win best_actress_win
##  Min.   :95    yes:1        no:1         yes:1          no:1
##  1st Qu.:95
##  Median :95
##  Mean   :95
##  3rd Qu.:95
##  Max.   :95
##  best_dir_win top200_box
##  no:1         no:1
##
##
##
##
##
```

```
# The BIC prior is chosen based on week 5 lab and the discussion board responses.
MS_baslm_b <- bas.lm(audience_score ~ ., data = audience_score_var,
                 prior = 'BIC',
                 modelprior = uniform())
colnames(MS_b) = colnames(audience_score_var)

MS_baslm_b
```

```
##
## Call:
## bas.lm(formula = audience_score ~ ., data = audience_score_var,
##     prior = "BIC", modelprior = uniform())
##
##
##  Marginal Posterior Inclusion Probabilities:
##           Intercept      feature_filmYes             dramaYes
##             1.00000              0.06537              0.04320
##             runtime     mpaa_rating_RYes        thtr_rel_year
##             0.46971              0.19984              0.09069
##     oscar_seasonYes     summer_seasonYes          imdb_rating
##             0.07506              0.08042              1.00000
##       imdb_num_votes        critics_score      best_pic_nomyes
##             0.05774              0.88855              0.13119
##     best_pic_winyes    best_actor_winyes  best_actress_winyes
##             0.03985              0.14435              0.14128
##     best_dir_winyes        top200_boxyes
##             0.06694              0.04762
```

```
summary(MS_baslm_b)
```

```
##                       P(B != 0 | Y)    model 1       model 2       model 3
## Intercept                1.00000000     1.0000     1.0000000     1.0000000
## feature_filmYes          0.06536947     0.0000     0.0000000     0.0000000
## dramaYes                 0.04319833     0.0000     0.0000000     0.0000000
## runtime                  0.46971477     1.0000     0.0000000     0.0000000
## mpaa_rating_RYes         0.19984016     0.0000     0.0000000     0.0000000
## thtr_rel_year            0.09068970     0.0000     0.0000000     0.0000000
## oscar_seasonYes          0.07505684     0.0000     0.0000000     0.0000000
## summer_seasonYes         0.08042023     0.0000     0.0000000     0.0000000
## imdb_rating              1.00000000     1.0000     1.0000000     1.0000000
## imdb_num_votes           0.05773502     0.0000     0.0000000     0.0000000
## critics_score            0.88855056     1.0000     1.0000000     1.0000000
## best_pic_nomyes          0.13119140     0.0000     0.0000000     0.0000000
## best_pic_winyes          0.03984766     0.0000     0.0000000     0.0000000
## best_actor_winyes        0.14434896     0.0000     0.0000000     1.0000000
## best_actress_winyes      0.14128087     0.0000     0.0000000     0.0000000
## best_dir_winyes          0.06693898     0.0000     0.0000000     0.0000000
## top200_boxyes            0.04762234     0.0000     0.0000000     0.0000000
## BF                               NA     1.0000     0.9968489     0.2543185
## PostProbs                        NA     0.1297     0.1293000     0.0330000
## R2                               NA     0.7549     0.7525000     0.7539000
## dim                              NA     4.0000     3.0000000     4.0000000
## logmarg                          NA -3615.2791 -3615.2822108 -3616.6482224
##                             model 4       model 5
## Intercept                 1.0000000     1.0000000
## feature_filmYes           0.0000000     0.0000000
## dramaYes                  0.0000000     0.0000000
## runtime                   0.0000000     1.0000000
## mpaa_rating_RYes          1.0000000     1.0000000
## thtr_rel_year             0.0000000     0.0000000
## oscar_seasonYes           0.0000000     0.0000000
## summer_seasonYes          0.0000000     0.0000000
## imdb_rating               1.0000000     1.0000000
## imdb_num_votes            0.0000000     0.0000000
## critics_score             1.0000000     1.0000000
## best_pic_nomyes           0.0000000     0.0000000
## best_pic_winyes           0.0000000     0.0000000
## best_actor_winyes         0.0000000     0.0000000
## best_actress_winyes       0.0000000     0.0000000
## best_dir_winyes           0.0000000     0.0000000
## top200_boxyes             0.0000000     0.0000000
## BF                        0.2521327     0.2391994
## PostProbs                 0.0327000     0.0310000
## R2                        0.7539000     0.7563000
## dim                       4.0000000     5.0000000
## logmarg               -3616.6568544 -3616.7095127
```

```
MS_pred_bd = predict(MS_baslm_b, newdata = MS_b, estimator = "BMA", se.fit = TRUE)
MS_pred_bd$fit # 83.50
```

```
## [1] 83.49721
```

```
# CI
MS_pred_bd_ci <- confint(MS_pred_bd, estimator = "BMA")
MS_pred_bd_ci # CI: 63.3.-103.4
```

```
##          2.5%    97.5%     pred
## [1,] 63.98796 104.4521 83.49721
## attr(,"Probability")
## [1] 0.95
## attr(,"class")
## [1] "confint.bas"
```

```
# Sensitivity Analysis to see how other prediction methods compare with the results a
bove.
MS_pred_ba = predict(MS_baslm_b, newdata = MS_b, estimator = "HPM", se.fit = TRUE)
MS_pred_ba$fit # 82.91
```

```
## [1] 82.90536
## attr(,"model")
## [1]  0  3  8 10
## attr(,"best")
## [1] 8776
## attr(,"estimator")
## [1] "HPM"
```

```
MS_pred_ba_ci <- confint(MS_pred_ba, estimator = "HPM")
MS_pred_ba_ci # CI: 63.1-102.7
```

```
##          2.5%    97.5%     pred
## [1,] 63.11821 102.6925 82.90536
## attr(,"Probability")
## [1] 0.95
## attr(,"class")
## [1] "confint.bas"
```

Part 6: Conclusion

Though many variables were collected during the sampling process only several of them are statistically significant in the simple linear regression model for explaining audience score. After applying Bayesian linear regression even fewer variables were found to be significantly associated with audience score. BIC was primarily used to determine which variables should be included in the analysis as well as the BMA prediction method.

The research question is which variables are associated with audience score and is it possible to predict the audience score of a movie using Bayesian linear regression. Having found the predictive model a regression was run and the variables for the movie Manchester by the Sea were entered into the model. The predicted audience score is 83.5. There is a 0.95 probability that Manchester by the Sea receives an audience score on average between 63.3 and 103.4.

Manchester by the Sea received an audience rating of 77, which sits within the credible interval found in the prediction stage. The biggest concern with this prediction is the large credible interval, which suggests a lack of accuracy. That being said, the predicted value was close to the actual value therefore there is some validity in using this model to predict the audience score of movies released in 2016.