Data Exploration and Preprocessing

```
##   user_id movie_id rating timestamp                    movie_title
## 1       1        1      5 874965758               Toy Story (1995)
## 2       1      101      2 878542845             Heavy Metal (1981)
## 3       1      188      3 875073128       Full Metal Jacket (1987)
## 4       1       32      5 888732909                   Crumb (1994)
## 5       1       66      4 878543030 While You Were Sleeping (1995)
## 6       1      250      4 874965706      Fifth Element, The (1997)
##   release_date video_release_date
## 1  01-Jan-1995                 NA
## 2  08-Mar-1981                 NA
## 3  01-Jan-1987                 NA
## 4  01-Jan-1994                 NA
## 5  01-Jan-1995                 NA
## 6  09-May-1997                 NA
##                                                                   imdb_url
## 1                   http://us.imdb.com/M/title-exact?Toy%20Story%20(1995)
## 2                 http://us.imdb.com/M/title-exact?Heavy%20Metal%20(1981)
## 3         http://us.imdb.com/M/title-exact?Full%20Metal%20Jacket%20(1987)
## 4                         http://us.imdb.com/M/title-exact?Crumb%20(1994)
## 5 http://us.imdb.com/M/title-exact?While%20You%20Were%20Sleeping%20(1995)
## 6  http://us.imdb.com/M/title-exact?Fifth%20Element%2C%20The%20%281997%29
##   unknown action adventure animation childrens comedy crime documentary
## 1       0      0         0         1         1      1     0           0
## 2       0      1         1         1         0      0     0           0
## 3       0      1         0         0         0      0     0           0
## 4       0      0         0         0         0      0     0           1
## 5       0      0         0         0         0      1     0           0
## 6       0      1         0         0         0      0     0           0
##   drama fantasy filmnoir horror musical mystery romance scifi thriller war
## 1     0       0        0      0       0       0       0     0        0   0
## 2     0       0        0      1       0       0       0     1        0   0
## 3     1       0        0      0       0       0       0     0        0   1
## 4     0       0        0      0       0       0       0     0        0   0
## 5     0       0        0      0       0       0       1     0        0   0
## 6     0       0        0      0       0       0       0     1        0   0
##   western age gender occupation zip_code
## 1       0  24      M technician    85711
## 2       0  24      M technician    85711
## 3       0  24      M technician    85711
## 4       0  24      M technician    85711
## 5       0  24      M technician    85711
## 6       0  24      M technician    85711
```

Preprocessing

Missing Values

```
NA_ratings <- sapply(ratings, function(x) switch( class(x), factor = sum(x==""), sum( is.na(x) ) ))
NA_ratings_df <- as.data.frame( t(NA_ratings) )
NA_ratings_df[which(NA_ratings_df != 0)]
```

```
##   release_date video_release_date imdb_url
## 1            9             100000       13
```

From above, we can see that all values of video_release_date are NA. This suggests that we can remove the feature from our dataset without losing any information.

We can also see that there are 9 observations with missing values. We can take a closer look to at the overvations that have NA value for release date.

```r
# Convert movie release date to seconds from UNIX epoch
ratings$release_date <- apply(ratings[c('release_date')], 1, date_to_sec)
ratings %>% filter(release_date %>% is.na())
```

```
##   user_id movie_id rating timestamp movie_title release_date
## 1       1      267      4 875692955     unknown           NA
## 2       5      267      4 875635064     unknown           NA
## 3     130      267      5 875801239     unknown           NA
## 4     268      267      3 875742077     unknown           NA
## 5     297      267      3 875409139     unknown           NA
## 6     319      267      4 875707690     unknown           NA
## 7     422      267      4 875655986     unknown           NA
## 8     532      267      3 875441348     unknown           NA
## 9     833      267      1 875655669     unknown           NA
##   video_release_date imdb_url unknown action adventure animation childrens
## 1                 NA                1      0         0         0         0
## 2                 NA                1      0         0         0         0
## 3                 NA                1      0         0         0         0
## 4                 NA                1      0         0         0         0
## 5                 NA                1      0         0         0         0
## 6                 NA                1      0         0         0         0
## 7                 NA                1      0         0         0         0
## 8                 NA                1      0         0         0         0
## 9                 NA                1      0         0         0         0
##   comedy crime documentary drama fantasy filmnoir horror musical mystery
## 1      0     0           0     0       0        0      0       0       0
## 2      0     0           0     0       0        0      0       0       0
## 3      0     0           0     0       0        0      0       0       0
## 4      0     0           0     0       0        0      0       0       0
## 5      0     0           0     0       0        0      0       0       0
## 6      0     0           0     0       0        0      0       0       0
## 7      0     0           0     0       0        0      0       0       0
## 8      0     0           0     0       0        0      0       0       0
## 9      0     0           0     0       0        0      0       0       0
##   romance scifi thriller war western age gender    occupation zip_code
## 1       0     0        0   0       0  24      M    technician    85711
## 2       0     0        0   0       0  33      F         other    15213
## 3       0     0        0   0       0  20      M          none    60115
## 4       0     0        0   0       0  24      M      engineer    19422
## 5       0     0        0   0       0  29      F      educator    98103
## 6       0     0        0   0       0  38      M    programmer    22030
## 7       0     0        0   0       0  26      M entertainment    94533
## 8       0     0        0   0       0  20      M       student    92705
## 9       0     0        0   0       0  34      M        writer    90019
```

From above, we can see that the observations with missing values for release_date are also missing information about movie_title and genre. This missing information will make these observations not very useful for rating predictions so we will drop them.

```r
# remove overvations with missing values for release_date
missing_dates <- ratings$release_date %>% is.na() %>% which()
ratings <- ratings[-missing_dates,]
```

The final variable with missing values is imdb_url. We will be removing this feature because it does not give useful information related to movie rating.

Feature Removal

We decided to remove the three following features: movie_title, video_release_date, and imdb_url. As discussed above, we will remove video_release_date because it have all missing values and imdb_url because it has missing values and is noninformative in regards to movie rating. Lastly, we will remove movie_title because it is redundant information since we already have movie_id.

```r
# Remove noninfomative predictors
drops <- c('movie_title','video_release_date','imdb_url')
ratings <- ratings[ , !names(ratings) %in% drops]
```

Feature Preprocessing

Preprocess timestamp and release_date to be consistent

```r
# convert timestamp and release date to class Date for fprocessing later
ratings$timestamp <- ratings$timestamp %>% as_datetime
ratings$release_date <- ratings$release_date %>% as_datetime
```

zip_code

```r
# replace old zipcode column with two digits
ratings$zip_code <- substr(as.character(ratings$zip_code),1,2)
```

Convert categorical variables to factors

```r
# convert all categorical variables to factors
factor_cols <- c('user_id', 'movie_id', 'unknown', 'action',
                       'adventure', 'animation', 'childrens', 'comedy', 'crime',
                       'documentary', 'drama', 'fantasy', 'filmnoir', 'horror',
                       'musical', 'mystery', 'romance', 'scifi', 'thriller',
                       'war', 'western','zip_code')
ratings[,factor_cols] <- data.frame(apply(ratings[factor_cols], 2, as.factor))
```

Feature Engineering

Time intervals

```r
ratings$release_year <- year(ratings$release_date)
ratings$release_month <- month(ratings$release_date)
ratings$timestamp_year <- year(ratings$timestamp)
ratings$timestamp_month <- month(ratings$timestamp)
ratings$time_difference <- as.period(ratings$timestamp - ratings$release_date) %>% day
```

Age intervals

```r
age <- ratings %>% pull(age)
ratings$age_group <- rep(0,nrow(ratings))
ratings$age_group <- findInterval(age,c(10,20,30,40,50,60,70,80))
ratings$age_group <- as.factor(ratings$age_group)
```

```r
# 0, 1, 2, 3, 4, 5, 6, 7
levels(ratings$age_group) <- c("<10","10-20","20-30","30-40","40-50","50-60","60-70","70+")
```

Summary Statistics

Distibution of Feature Values

```r
summary(ratings)
```

```
##     user_id         movie_id         rating    
##  405    :  737     50    :  583   Min.   :1.00  
##  655    :  685    258    :  509   1st Qu.:3.00  
##   13    :  636    100    :  508   Median :4.00  
##  450    :  540    181    :  507   Mean   :3.53  
##  276    :  518    294    :  485   3rd Qu.:4.00  
##  416    :  493    286    :  481   Max.   :5.00  
##  (Other):96382   (Other):96918                 
##    timestamp                     release_date                  unknown  
##  Min.   :1997-09-20 03:05:10   Min.   :1922-01-01 00:00:00   0:99990  
##  1st Qu.:1997-11-13 19:19:19   1st Qu.:1986-01-01 00:00:00   1:    1  
##  Median :1997-12-22 21:43:03   Median :1994-01-01 00:00:00            
##  Mean   :1997-12-31 00:52:41   Mean   :1988-02-09 00:43:11            
##  3rd Qu.:1998-02-23 18:53:04   3rd Qu.:1996-09-28 00:00:00            
##  Max.   :1998-04-22 23:10:38   Max.   :1998-10-23 00:00:00            
##                                                                       
##  action    adventure animation childrens comedy    crime     documentary
##  0:74402   0:86238   0:96386   0:92809   0:70159   0:91936   0:99233    
##  1:25589   1:13753   1: 3605   1: 7182   1:29832   1: 8055   1:  758    
##                                                                         
##                                                                         
##                                                                         
##                                                                         
##                                                                         
##  drama     fantasy   filmnoir  horror    musical   mystery   romance  
##  0:60096   0:98639   0:98258   0:94674   0:95037   0:94746   0:80530  
##  1:39895   1: 1352   1: 1733   1: 5317   1: 4954   1: 5245   1:19461  
##                                                                       
##                                                                       
##                                                                       
##                                                                       
##                                                                       
##  scifi     thriller  war       western        age        gender   
##  0:87261   0:78119   0:90593   0:98137   Min.   : 7.00   F:25738  
##  1:12730   1:21872   1: 9398   1: 1854   1st Qu.:24.00   M:74253  
##                                          Median :30.00            
##                                          Mean   :32.97            
##                                          3rd Qu.:40.00            
##                                          Max.   :73.00            
##                                                                   
##          occupation       zip_code      release_year  release_month  
##  student      :21956   55      : 7581   Min.   :1922   Min.   : 1.000  
##  other        :10662   60      : 4184   1st Qu.:1986   1st Qu.: 1.000  
##  educator     : 9441   02      : 2921   Median :1994   Median : 1.000  
##  engineer     : 8174   10      : 2815   Mean   :1988   Mean   : 2.643  
##  programmer   : 7800   95     : 2783   3rd Qu.:1996   3rd Qu.: 3.000
##  administrator: 7479   20     : 2773   Max.   :1998   Max.   :12.000
##  (Other)      :34479   (Other):76934
##  timestamp_year timestamp_month  time_difference   age_group
##  Min.   :1997   Min.   : 1.000   Min.   : -292   20-30  :39529
##  1st Qu.:1997   1st Qu.: 2.000   1st Qu.:  467   30-40  :25693
##  Median :1997   Median : 9.000   Median : 1389   40-50  :15021
##  Mean   :1997   Mean   : 6.815   Mean   : 3612   50-60  : 8704
##  3rd Qu.:1998   3rd Qu.:11.000   3rd Qu.: 4290   10-20  : 8181
##  Max.   :1998   Max.   :12.000   Max.   :27866   60-70  : 2623
##                                                  (Other):  240
```

Features Types and Values

```r
str(ratings)
```

```
## 'data.frame':    99991 obs. of  34 variables:
##  $ user_id        : Factor w/ 943 levels "  1","  2","  3",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ movie_id       : Factor w/ 1681 levels "   1","   2",..: 1 101 188 32 66 250 258 240 25 85 ...
##  $ rating         : int  5 2 3 5 4 4 5 3 4 3 ...
##  $ timestamp      : POSIXct, format: "1997-09-22 22:02:38" "1997-11-03 07:40:45" ...
##  $ release_date   : POSIXct, format: "1995-01-01" "1981-03-08" ...
##  $ unknown        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ action         : Factor w/ 2 levels "0","1": 1 2 2 1 1 2 1 1 1 1 ...
##  $ adventure      : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
##  $ animation      : Factor w/ 2 levels "0","1": 2 2 1 1 1 1 2 1 1 ...
##  $ childrens      : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 1 ...
##  $ comedy         : Factor w/ 2 levels "0","1": 2 1 1 1 2 1 1 2 2 2 ...
##  $ crime          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ documentary    : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...
##  $ drama          : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 2 1 1 1 ...
##  $ fantasy        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ filmnoir       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ horror         : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
##  $ musical        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ mystery        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ romance        : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
##  $ scifi          : Factor w/ 2 levels "0","1": 1 2 1 1 1 2 2 1 1 1 ...
##  $ thriller       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ war            : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 1 1 1 ...
##  $ western        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ age            : int  24 24 24 24 24 24 24 24 24 24 ...
##  $ gender         : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ occupation     : Factor w/ 21 levels "administrator",..: 20 20 20 20 20 20 20 20 20 20 ...
##  $ zip_code       : Factor w/ 111 levels "00","01","02",..: 84 84 84 84 84 84 84 84 84 84 ...
##  $ release_year   : num  1995 1981 1987 1994 1995 ...
##  $ release_month  : num  1 3 1 1 1 5 7 12 3 1 ...
##  $ timestamp_year : num  1997 1997 1997 1998 1997 ...
##  $ timestamp_month: num  9 11 9 3 11 9 11 9 9 9 ...
##  $ time_difference: num  995 6084 3919 1520 1037 ...
##  $ age_group      : Factor w/ 8 levels "<10","10-20",..: 3 3 3 3 3 3 3 3 3 3 ...
```

Gender

Number of ratings by gender

```
##     F     M
## 25738 74253
```

Rating Proportions by Gender

```
##
##              1          2          3          4          5
##   F 0.01894170 0.02784251 0.06783611 0.08302747 0.05975538
##   M 0.04215379 0.08586773 0.20360832 0.25870328 0.15226370
```

Row wise rating Proportions

```
##
##              1          2          3          4          5
##   F 0.07358769 0.10816691 0.26354029 0.32255809 0.23214702
##   M 0.05676538 0.11563169 0.27418421 0.34837650 0.20504222
```

Age

Number of ratings by age

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   24.00   30.00   32.97   40.00   73.00
```

Rating proportions by age group

```
##
## age_group            1            2            3            4            5
##       <10 0.0000100009 0.0000400036 0.0000900081 0.0001900171 0.0001000090
##     10-20 0.0062905662 0.0094208479 0.0218719685 0.0266924023 0.0175415787
##     20-30 0.0288625976 0.0483343501 0.1065595904 0.1323419108 0.0792271304
##     30-40 0.0147313258 0.0283125481 0.0702063186 0.0871478433 0.0565550900
##     40-50 0.0074506706 0.0156114050 0.0411837065 0.0525447290 0.0334330090
##     50-60 0.0029002610 0.0091408227 0.0240821674 0.0315628407 0.0193617426
##     60-70 0.0007600684 0.0026502385 0.0069906292 0.0104709424 0.0053604824
##     70+   0.0000900081 0.0002000180 0.0004600414 0.0007800702 0.0004400396
```

Row wise rating proportions by age group

```
##
## age_group            1            2            3            4            5
##       <10 0.0000100009 0.0000400036 0.0000900081 0.0001900171 0.0001000090
##     10-20 0.0062905662 0.0094208479 0.0218719685 0.0266924023 0.0175415787
##     20-30 0.0288625976 0.0483343501 0.1065595904 0.1323419108 0.0792271304
##     30-40 0.0147313258 0.0283125481 0.0702063186 0.0871478433 0.0565550900
##     40-50 0.0074506706 0.0156114050 0.0411837065 0.0525447290 0.0334330090
##     50-60 0.0029002610 0.0091408227 0.0240821674 0.0315628407 0.0193617426
##     60-70 0.0007600684 0.0026502385 0.0069906292 0.0104709424 0.0053604824
##     70+   0.0000900081 0.0002000180 0.0004600414 0.0007800702 0.0004400396
```

Table with both gender and age

```r
aggregate(rating ~ age_group + gender, data=ratings, FUN=sum)
```

```
##    age_group gender rating
## 1      10-20      F   9094
## 2      20-30      F  32979
## 3      30-40      F  25096
## 4      40-50      F  14909
## 5      50-60      F   8511
## 6      60-70      F     75
## 7        70+      F    230
## 8        <10      M    162
## 9      10-20      M  19426
## 10     20-30      M 104080
## 11     30-40      M  66230
## 12     40-50      M  39043
## 13     50-60      M  23135
## 14     60-70      M   9496
## 15       70+      M    489
```

```
aggregate(rating ~ age_group + gender, data=ratings, FUN=length)
```

```
##    age_group gender rating
## 1      10-20      F   2560
## 2      20-30      F   9642
## 3      30-40      F   6834
## 4      40-50      F   4201
## 5      50-60      F   2407
## 6      60-70      F     23
## 7        70+      F     71
## 8        <10      M     43
## 9      10-20      M   5621
## 10     20-30      M  29887
## 11     30-40      M  18859
## 12     40-50      M  10820
## 13     50-60      M   6297
## 14     60-70      M   2600
## 15       70+      M    126
```

```
aggregate(rating ~ age_group + gender, data=ratings, FUN=function(x) {length(x)/sum(x)})
```

```
##    age_group gender    rating
## 1      10-20      F 0.2815043
## 2      20-30      F 0.2923679
## 3      30-40      F 0.2723143
## 4      40-50      F 0.2817761
## 5      50-60      F 0.2828105
## 6      60-70      F 0.3066667
## 7        70+      F 0.3086957
## 8        <10      M 0.2654321
## 9      10-20      M 0.2893545
## 10     20-30      M 0.2871541
## 11     30-40      M 0.2847501
## 12     40-50      M 0.2771303
## 13     50-60      M 0.2721850
## 14     60-70      M 0.2737995
## 15       70+      M 0.2576687
```

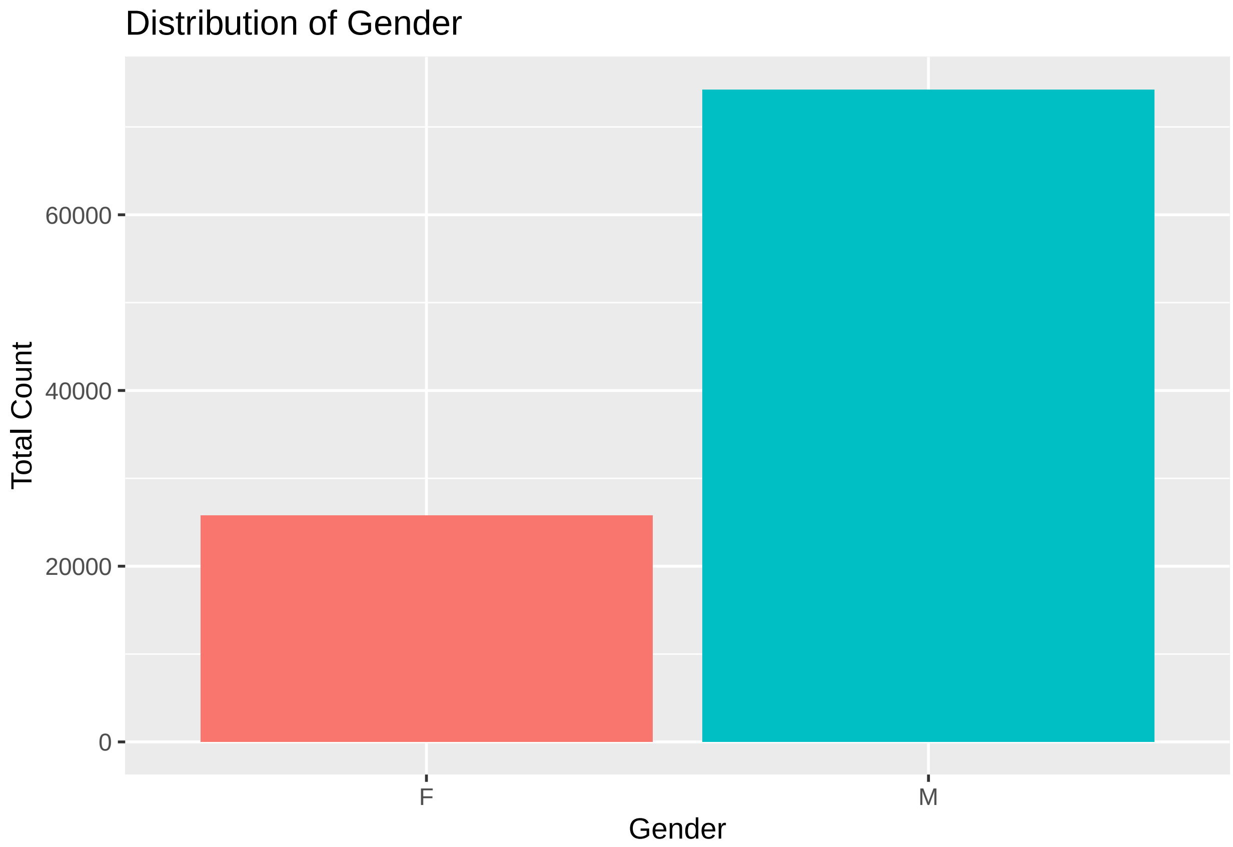
```
aggregate(rating ~ zip_code + age_group + gender, data=ratings, FUN=function(x) {sum(x)/length(x)})
```

```
##     zip_code age_group gender   rating
## 1         02     10-20      F 3.435484
## 2         06     10-20      F 3.584615
## 3         25     10-20      F 4.869565
## 4         27     10-20      F 3.923977
## 5         28     10-20      F 4.724138
## 6         38     10-20      F 3.100000
## 7         40     10-20      F 3.423077
## 8         44     10-20      F 3.333333
## 9         47     10-20      F 3.028571
## 10        49     10-20      F 3.460526
## 11        51     10-20      F 2.947368
## 12        53     10-20      F 2.171875
## 13        55     10-20      F 3.648000
## 14        61     10-20      F 3.744681
## 15        63     10-20      F 4.333333
## 16        74     10-20      F 3.380282
## 17        77     10-20      F 3.739130
## 18        78     10-20      F 3.168750
## 19        81     10-20      F 3.827273
## 20        84     10-20      F 3.624060
## 21        93     10-20      F 3.746939
## 22        95     10-20      F 3.628492
## 23        98     10-20      F 3.140351
## 24        99     10-20      F 3.081081
## 25        02     20-30      F 4.255639
## 26        06     20-30      F 3.286517
## 27        07     20-30      F 2.968254
## 28        08     20-30      F 3.677551
## 29        10     20-30      F 2.000000
## 30        11     20-30      F 3.405941
## 31        14     20-30      F 3.849624
## 32        15     20-30      F 3.859729
## 33        19     20-30      F 3.296296
## 34        20     20-30      F 3.545830
## 35        21     20-30      F 3.454545
## 36        22     20-30      F 3.583333
## 37        23     20-30      F 4.019231
## 38        32     20-30      F 3.418605
## 39        33     20-30      F 3.046263
## 40        35     20-30      F 4.191011
## 41        42     20-30      F 3.000000
## 42        45     20-30      F 3.896552
## 43        46     20-30      F 4.111111
## 44        48     20-30      F 3.439799
## 45        50     20-30      F 4.393939
## 46        53     20-30      F 3.600000
## 47        54     20-30      F 3.725490
## 48        55     20-30      F 3.408108
## 49        60     20-30      F 3.936508
## 50        62     20-30      F 3.688073
## 51        63     20-30      F 3.609375
## 52        66     20-30      F 4.303030
## 53        68     20-30      F 4.062500
## 54        71     20-30      F 3.384615
## 55        75     20-30      F 4.267857
## 56        76     20-30      F 2.683721
## 57        78     20-30      F 3.557604
## 58        80     20-30      F 4.195980
## 59         85     20-30      F 3.320755
## 60         90     20-30      F 3.541667
## 61         92     20-30      F 3.845842
## 62         94     20-30      F 3.800000
## 63         96     20-30      F 3.200000
## 64         97     20-30      F 3.301724
## 65         98     20-30      F 3.413793
## 66         R3     20-30      F 3.893617
## 67         V5     20-30      F 2.789474
## 68         00     30-40      F 3.413043
## 69         01     30-40      F 3.871795
## 70         03     30-40      F 4.000000
## 71         07     30-40      F 2.600000
## 72         11     30-40      F 3.811382
## 73         14     30-40      F 3.630952
## 74         15     30-40      F 2.867816
## 75         17     30-40      F 3.580110
## 76         22     30-40      F 3.655172
## 77         27     30-40      F 3.476190
## 78         29     30-40      F 4.321429
## 79         30     30-40      F 3.857143
## 80         32     30-40      F 3.513761
## 81         33     30-40      F 3.861386
## 82         37     30-40      F 3.880866
## 83         39     30-40      F 3.990338
## 84         42     30-40      F 4.050000
## 85         43     30-40      F 3.004310
## 86         44     30-40      F 3.888476
## 87         48     30-40      F 3.877729
## 88         49     30-40      F 3.872340
## 89         52     30-40      F 3.606796
## 90         53     30-40      F 3.806122
## 91         55     30-40      F 3.392593
## 92         59     30-40      F 4.000000
## 93         60     30-40      F 4.833333
## 94         68     30-40      F 3.896947
## 95         77     30-40      F 3.803191
## 96         78     30-40      F 3.590909
## 97         85     30-40      F 3.258621
## 98         90     30-40      F 3.589744
## 99         92     30-40      F 3.583333
## 100        94     30-40      F 3.818713
## 101        95     30-40      F 3.313725
## 102        97     30-40      F 3.233333
## 103        V0     30-40      F 3.104478
## 104        V1     30-40      F 3.272727
## 105        02     40-50      F 3.964912
## 106        06     40-50      F 3.576923
## 107        07     40-50      F 2.966667
## 108        08     40-50      F 3.086957
## 109        11     40-50      F 4.113821
## 110        12     40-50      F 2.959459
## 111        16     40-50      F 4.141243
## 112        19     40-50      F 4.000000
## 113       20     40-50      F 3.851852
## 114       29     40-50      F 2.806452
## 115       30     40-50      F 3.240000
## 116       33     40-50      F 2.900000
## 117       34     40-50      F 3.793103
## 118       43     40-50      F 3.680000
## 119       44     40-50      F 3.854922
## 120       53     40-50      F 3.554140
## 121       55     40-50      F 3.736000
## 122       60     40-50      F 3.138462
## 123       61     40-50      F 3.920000
## 124       62     40-50      F 2.653846
## 125       64     40-50      F 3.326360
## 126       68     40-50      F 3.800000
## 127       70     40-50      F 3.385542
## 128       73     40-50      F 3.860759
## 129       75     40-50      F 2.864662
## 130       77     40-50      F 3.360000
## 131       78     40-50      F 4.056075
## 132       80     40-50      F 3.935780
## 133       83     40-50      F 3.704797
## 134       84     40-50      F 3.117647
## 135       85     40-50      F 3.254054
## 136       89     40-50      F 4.041667
## 137       90     40-50      F 3.592593
## 138       92     40-50      F 3.621514
## 139       93     40-50      F 3.892157
## 140       94     40-50      F 3.000000
## 141       95     40-50      F 3.081081
## 142       97     40-50      F 3.121339
## 143       99     40-50      F 3.643836
## 144       03     50-60      F 4.127660
## 145       04     50-60      F 4.360000
## 146       10     50-60      F 3.658824
## 147       15     50-60      F 4.133333
## 148       17     50-60      F 3.619048
## 149       19     50-60      F 3.495763
## 150       20     50-60      F 3.541667
## 151       21     50-60      F 2.900000
## 152       27     50-60      F 4.018750
## 153       30     50-60      F 3.619048
## 154       43     50-60      F 4.060606
## 155       48     50-60      F 3.937500
## 156       53     50-60      F 4.034483
## 157       56     50-60      F 4.212121
## 158       58     50-60      F 4.518519
## 159       60     50-60      F 2.958904
## 160       62     50-60      F 3.382353
## 161       63     50-60      F 4.200000
## 162       80     50-60      F 4.121951
## 163       90     50-60      F 3.241379
## 164       91     50-60      F 3.777778
## 165       92     50-60      F 3.654762
## 166       94     50-60      F 3.709677
## 167       97     50-60      F 4.041667
## 168       98     50-60      F 3.268657
## 169       78     60-70      F 3.260870
## 170       48       70+      F 3.239437
## 171       55       <10      M 3.767442
## 172       02     10-20      M 3.140000
## 173       05     10-20      M 2.895522
## 174       06     10-20      M 3.288462
## 175       14     10-20      M 3.365702
## 176       17     10-20      M 3.553191
## 177       20     10-20      M 3.489362
## 178       22     10-20      M 3.029412
## 179       24     10-20      M 3.720779
## 180       27     10-20      M 3.705882
## 181       28     10-20      M 2.928205
## 182       29     10-20      M 3.459330
## 183       30     10-20      M 3.531915
## 184       37     10-20      M 4.000000
## 185       44     10-20      M 3.644156
## 186       48     10-20      M 2.961832
## 187       55     10-20      M 3.678363
## 188       56     10-20      M 2.857143
## 189       58     10-20      M 2.955882
## 190       60     10-20      M 3.415698
## 191       76     10-20      M 3.666667
## 192       77     10-20      M 3.348416
## 193       83     10-20      M 3.377451
## 194       84     10-20      M 3.500000
## 195       90     10-20      M 3.572864
## 196       92     10-20      M 3.557252
## 197       93     10-20      M 3.918310
## 198       94     10-20      M 3.814815
## 199       97     10-20      M 3.664319
## 200       98     10-20      M 3.000000
## 201       01     20-30      M 4.147826
## 202       02     20-30      M 3.812371
## 203       03     20-30      M 3.703872
## 204       05     20-30      M 4.075000
## 205       07     20-30      M 3.041096
## 206       08     20-30      M 3.430642
## 207       10     20-30      M 3.754950
## 208       11     20-30      M 3.307692
## 209       12     20-30      M 3.265625
## 210       13     20-30      M 4.267380
## 211       14     20-30      M 3.424165
## 212       15     20-30      M 3.727273
## 213       16     20-30      M 3.921212
## 214       18     20-30      M 3.909091
## 215       19     20-30      M 3.325714
## 216       20     20-30      M 3.728242
## 217       21     20-30      M 2.414501
## 218       23     20-30      M 3.183099
## 219       27     20-30      M 3.129386
## 220       28     20-30      M 3.360000
## 221       29     20-30      M 3.368243
## 222       30     20-30      M 2.780822
## 223       31     20-30      M 3.498667
## 224       32     20-30      M 3.427948
## 225       33     20-30      M 3.520548
## 226       37     20-30      M 3.415584
## 227       38     20-30      M 3.457627
## 228       39     20-30      M 3.791667
## 229       40     20-30      M 3.823151
## 230       41     20-30      M 3.727273
## 231       42     20-30      M 2.961538
## 232       43     20-30      M 3.421687
## 233       44     20-30      M 3.776860
## 234       45     20-30      M 3.833333
## 235       46     20-30      M 3.637119
## 236       47     20-30      M 3.745575
## 237       48     20-30      M 3.029412
## 238       49     20-30      M 3.523810
## 239       50     20-30      M 3.590400
## 240       52     20-30      M 3.797753
## 241       53     20-30      M 3.905694
## 242       55     20-30      M 3.445563
## 243       60     20-30      M 3.399707
## 244       61     20-30      M 3.636564
## 245       63     20-30      M 3.688571
## 246       64     20-30      M 3.925926
## 247       65     20-30      M 3.453532
## 248       66     20-30      M 3.837838
## 249       71     20-30      M 3.657500
## 250       75     20-30      M 3.604167
## 251       76     20-30      M 3.519737
## 252       77     20-30      M 3.238965
## 253       78     20-30      M 3.237654
## 254       79     20-30      M 3.325301
## 255       80     20-30      M 3.349112
## 256       83     20-30      M 3.935135
## 257       84     20-30      M 3.841991
## 258       85     20-30      M 3.463811
## 259       87     20-30      M 3.592334
## 260       90     20-30      M 3.855814
## 261       91     20-30      M 3.304833
## 262       92     20-30      M 3.429565
## 263       94     20-30      M 3.493734
## 264       95     20-30      M 3.516667
## 265       96     20-30      M 3.859060
## 266       97     20-30      M 4.045455
## 267       98     20-30      M 4.000000
## 268       99     20-30      M 3.591837
## 269       E2     20-30      M 3.210870
## 270       N2     20-30      M 3.365931
## 271       N4     20-30      M 3.572464
## 272       01     30-40      M 3.371795
## 273       02     30-40      M 3.284247
## 274       03     30-40      M 4.060606
## 275    05        30-40      M 3.589928
## 276    06        30-40      M 3.431250
## 277    08        30-40      M 3.132075
## 278    10        30-40      M 3.357625
## 279    11        30-40      M 3.296296
## 280    12        30-40      M 3.656977
## 281    15        30-40      M 2.600000
## 282    17        30-40      M 3.688889
## 283    18        30-40      M 3.779528
## 284    20        30-40      M 3.526667
## 285    21        30-40      M 3.700787
## 286    22        30-40      M 3.197929
## 287    23        30-40      M 3.833333
## 288    26        30-40      M 4.563380
## 289    27        30-40      M 3.497382
## 290    28        30-40      M 3.579196
## 291    29        30-40      M 3.571429
## 292    30        30-40      M 3.315143
## 293    31        30-40      M 4.450000
## 294    32        30-40      M 3.584416
## 295    33        30-40      M 3.868571
## 296    34        30-40      M 3.493333
## 297    36        30-40      M 4.173077
## 298    37        30-40      M 3.663851
## 299    40        30-40      M 3.328829
## 300    43        30-40      M 3.505535
## 301    44        30-40      M 3.127321
## 302    45        30-40      M 4.000000
## 303    46        30-40      M 3.153846
## 304    47        30-40      M 3.851852
## 305    48        30-40      M 3.760000
## 306    50        30-40      M 4.239796
## 307    51        30-40      M 3.466667
## 308    53        30-40      M 4.333333
## 309    54        30-40      M 3.790576
## 310    55        30-40      M 3.550199
## 311    57        30-40      M 4.307692
## 312    60        30-40      M 3.658537
## 313    61        30-40      M 3.665663
## 314    62        30-40      M 3.570175
## 315    63        30-40      M 3.427885
## 316    67        30-40      M 3.950276
## 317    73        30-40      M 3.798295
## 318    74        30-40      M 2.490385
## 319    75        30-40      M 3.203540
## 320    76        30-40      M 3.500000
## 321    77        30-40      M 4.017341
## 322    78        30-40      M 3.518732
## 323    79        30-40      M 3.370370
## 324    80        30-40      M 3.130785
## 325    85        30-40      M 3.314410
## 326    89        30-40      M 2.451613
## 327    90        30-40      M 3.162534
## 328    91        30-40      M 3.602484
## 329       92     30-40      M 3.361446
## 330       93     30-40      M 3.222222
## 331       94     30-40      M 3.797508
## 332       95     30-40      M 3.784065
## 333       97     30-40      M 3.824503
## 334       98     30-40      M 3.865942
## 335       99     30-40      M 4.098039
## 336       K7     30-40      M 3.918919
## 337       L1     30-40      M 3.854839
## 338       L9     30-40      M 3.376812
## 339       M7     30-40      M 2.666667
## 340       T8     30-40      M 3.431034
## 341       V3     30-40      M 3.335443
## 342       01     40-50      M 3.688725
## 343       02     40-50      M 3.842767
## 344       03     40-50      M 3.673554
## 345       05     40-50      M 3.613793
## 346       06     40-50      M 2.863636
## 347       07     40-50      M 4.428571
## 348       08     40-50      M 3.932836
## 349       10     40-50      M 3.424658
## 350       12     40-50      M 3.513274
## 351       15     40-50      M 3.326203
## 352       17     40-50      M 4.300000
## 353       20     40-50      M 4.190476
## 354       21     40-50      M 3.187739
## 355       23     40-50      M 3.252632
## 356       26     40-50      M 4.363636
## 357       29     40-50      M 3.241290
## 358       30     40-50      M 3.987342
## 359       33     40-50      M 3.545455
## 360       36     40-50      M 3.429907
## 361       40     40-50      M 2.975610
## 362       42     40-50      M 3.050314
## 363       44     40-50      M 3.547170
## 364       45     40-50      M 3.795181
## 365       47     40-50      M 3.903226
## 366       48     40-50      M 4.450000
## 367       50     40-50      M 3.925373
## 368       53     40-50      M 3.864078
## 369       55     40-50      M 3.866029
## 370       60     40-50      M 3.582031
## 371       61     40-50      M 3.665306
## 372       63     40-50      M 3.913580
## 373       64     40-50      M 3.692308
## 374       66     40-50      M 3.923077
## 375       68     40-50      M 3.095238
## 376       70     40-50      M 3.481707
## 377       73     40-50      M 3.781250
## 378       74     40-50      M 3.636735
## 379       75     40-50      M 4.075145
## 380       77     40-50      M 3.365079
## 381       80     40-50      M 3.750000
## 382       83     40-50      M 3.304348
## 383       89     40-50      M 3.791469
## 384       90     40-50      M 3.767442
## 385       91     40-50      M 3.327381
## 386       92     40-50      M 3.553191
## 387       93     40-50      M 3.662577
## 388       94     40-50      M 3.673913
## 389       95     40-50      M 3.725490
## 390       96     40-50      M 3.147651
## 391       97     40-50      M 3.048193
## 392       98     40-50      M 3.740988
## 393       99     40-50      M 3.903226
## 394       M4     40-50      M 3.480000
## 395       V0     40-50      M 3.846154
## 396       Y1     40-50      M 3.688525
## 397       01     50-60      M 3.859296
## 398       02     50-60      M 3.564327
## 399       04     50-60      M 3.347826
## 400       05     50-60      M 3.341365
## 401       06     50-60      M 3.794904
## 402       07     50-60      M 3.640000
## 403       08     50-60      M 3.693333
## 404       14     50-60      M 3.652174
## 405       19     50-60      M 3.869565
## 406       20     50-60      M 3.581712
## 407       22     50-60      M 3.475499
## 408       27     50-60      M 3.834951
## 409       40     50-60      M 4.071429
## 410       45     50-60      M 3.793893
## 411       49     50-60      M 3.000000
## 412       50     50-60      M 3.627907
## 413       53     50-60      M 3.521739
## 414       55     50-60      M 2.794872
## 415       59     50-60      M 4.045045
## 416       60     50-60      M 2.903226
## 417       61     50-60      M 3.428571
## 418       62     50-60      M 4.170000
## 419       63     50-60      M 4.000000
## 420       70     50-60      M 3.784314
## 421       75     50-60      M 3.338983
## 422       78     50-60      M 3.500000
## 423       80     50-60      M 3.128205
## 424       82     50-60      M 3.500000
## 425       84     50-60      M 4.319149
## 426       85     50-60      M 4.548387
## 427       90     50-60      M 4.129808
## 428       91     50-60      M 3.919565
## 429       93     50-60      M 3.801196
## 430       94     50-60      M 3.567073
## 431       95     50-60      M 3.670968
## 432       97     50-60      M 3.154982
## 433       98     50-60      M 3.876543
## 434       99     50-60      M 3.608108
## 435       01     60-70      M 3.839286
## 436       02     60-70      M 3.285714
## 437        06     60-70      M 3.918429
## 438        09     60-70      M 3.428571
## 439        10     60-70      M 3.578947
## 440        12     60-70      M 3.434783
## 441        18     60-70      M 2.555556
## 442        21     60-70      M 3.205479
## 443        22     60-70      M 3.701149
## 444        32     60-70      M 3.189189
## 445        33     60-70      M 3.721311
## 446        48     60-70      M 3.745098
## 447        49     60-70      M 4.159091
## 448        55     60-70      M 3.765625
## 449        61     60-70      M 3.195122
## 450        78     60-70      M 4.262391
## 451        91     60-70      M 3.578125
## 452        94     60-70      M 3.225750
## 453        95     60-70      M 3.758186
## 454        97     60-70      M 3.129032
## 455        98     60-70      M 3.737500
## 456        00       70+      M 4.432432
## 457        37       70+      M 3.982143
## 458        78       70+      M 3.090909
```

Data Visualization

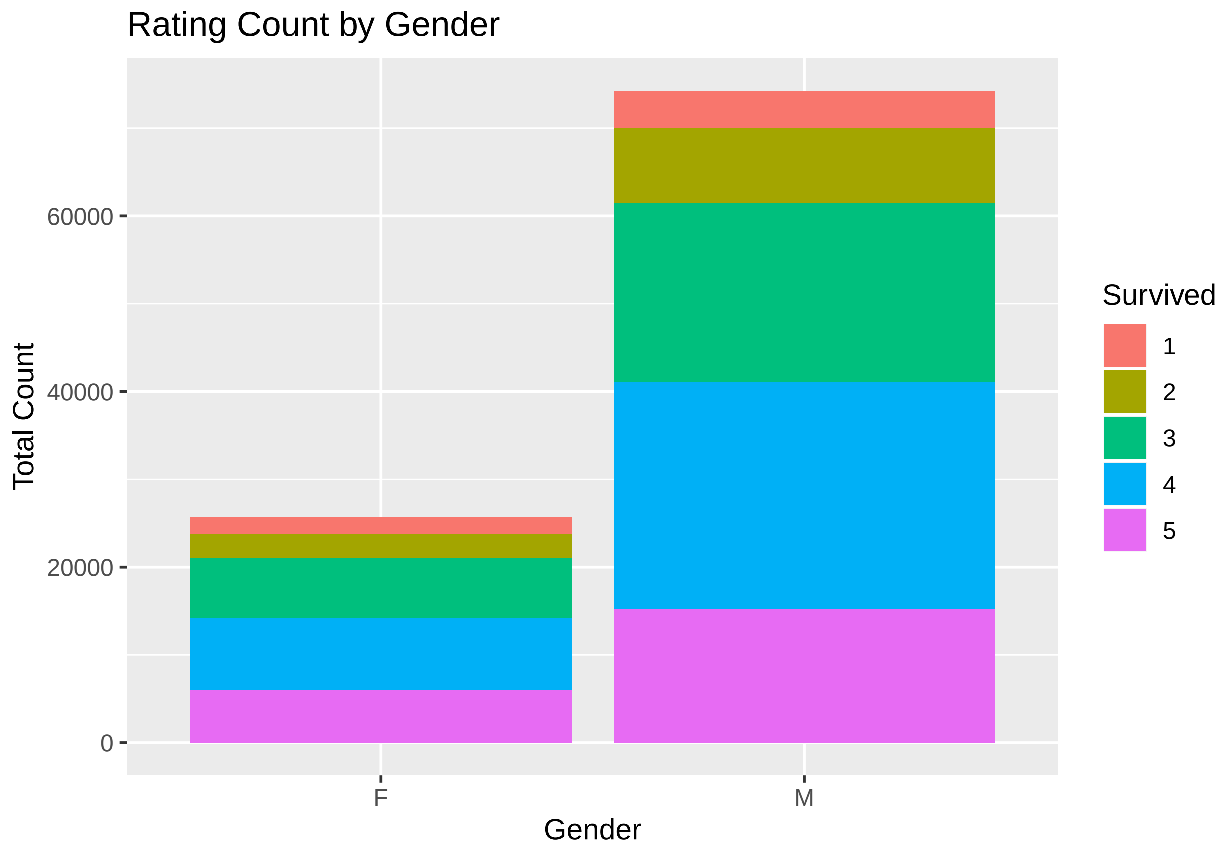
Distribution of Gender

```
ggplot(ratings, aes(x = factor(gender), fill = factor(gender))) +
  geom_bar( show.legend=FALSE) +
  xlab("Gender") +
  ylab("Total Count") +
  ggtitle("Distribution of Gender")
```



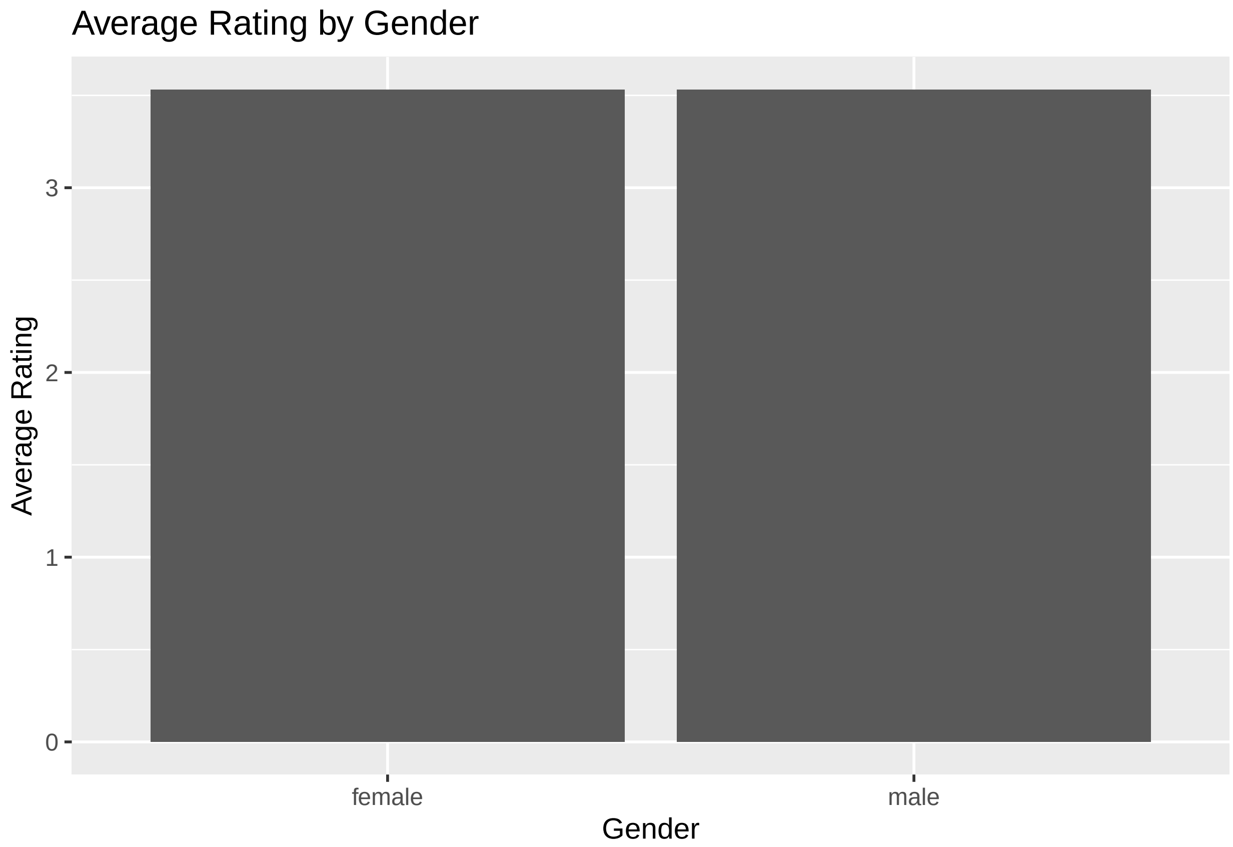
Ratings by Gender

```
# Ratings by Gender
ggplot(subset(ratings, !is.na(gender)), aes(x = gender, fill = as.factor(rating))) +
  geom_bar() +
  ggtitle("Rating Count by Gender") +
  xlab("Gender") +
  ylab("Total Count") +
  labs(fill = "Survived")
```



Average Rating by Gender

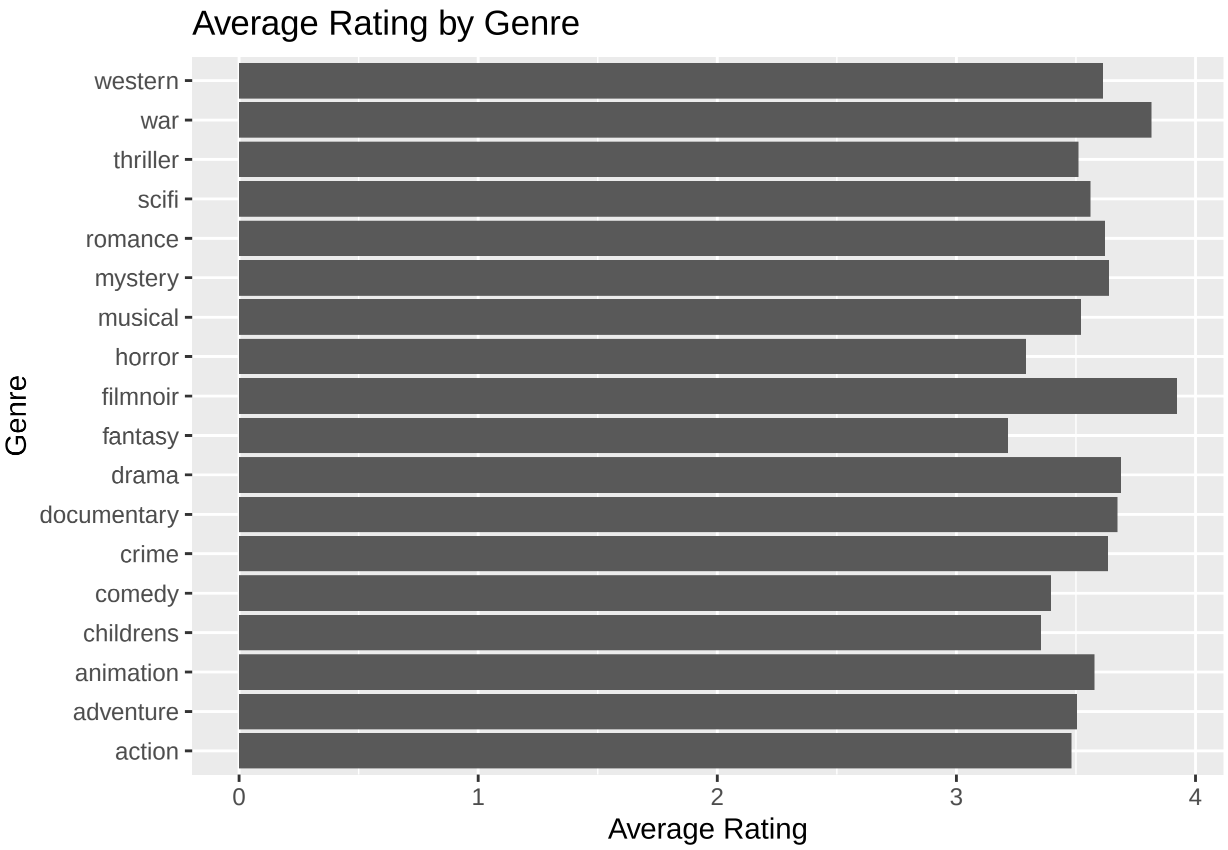
```r
male_mean <- ratings %>% filter(gender=='M') %>% pull(rating) %>% mean
female_mean <- ratings %>% filter(gender=='F') %>% pull(rating) %>% mean
mean_gender <- c(male_mean, female_mean)
gender <- c("male","female")
mean_gender_df <- data.frame(gender, mean_gender)
ggplot(mean_gender_df, aes(x=gender, y=mean_gender)) +
  geom_bar(stat="identity") +
  ggtitle("Average Rating by Gender") +
  xlab("Gender") +
  ylab("Average Rating")
```



Average Rating by Genre

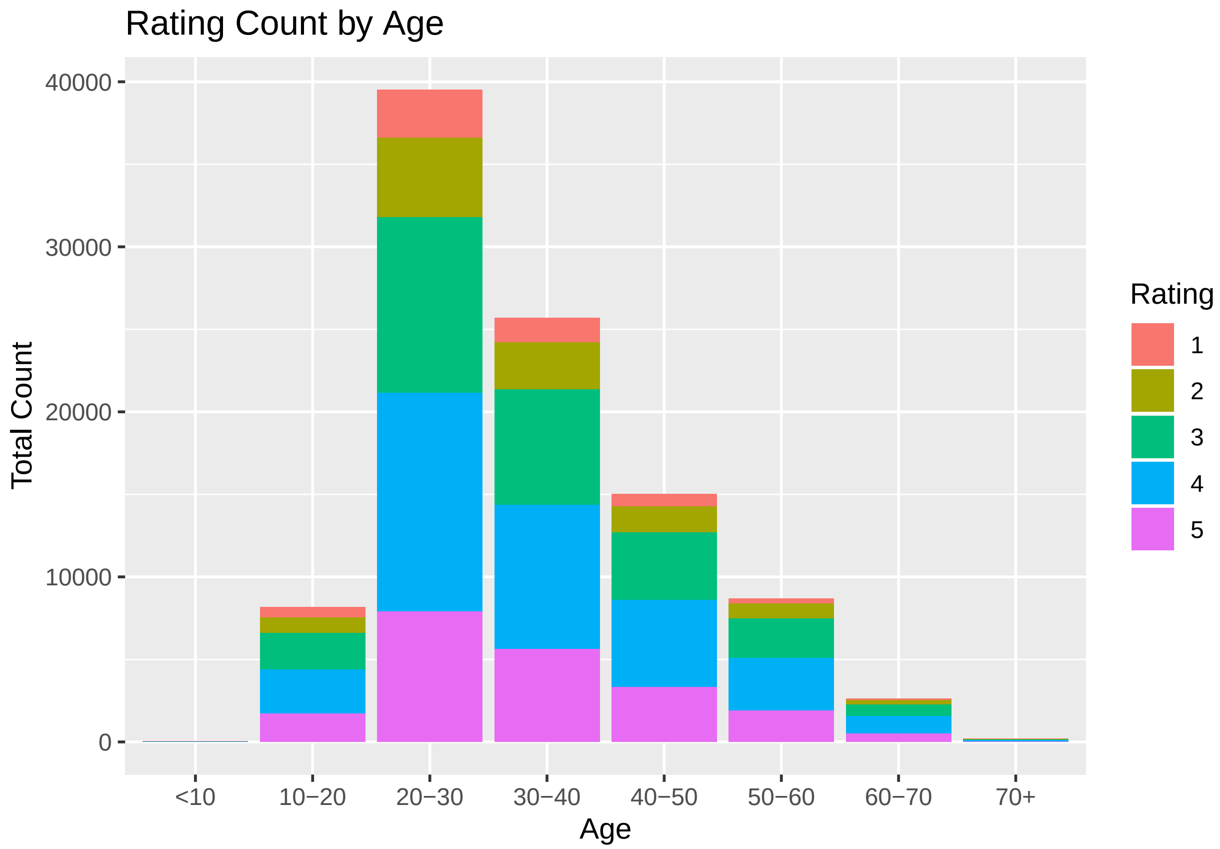
```
# convert genres to factor
genres <- ratings[,7:24]
for(i in 1:ncol(genres)) {
  genres[,i] <- as.factor(genres[,i])
}
ratings[,7:24] <- genres
genres_rating <- cbind(genres, ratings[,3])
colnames(genres_rating)[19] <- "rating"

# show average rating by genre
mean <- rep(0,ncol(genres_rating)-1)
for(i in 1:(ncol(genres_rating)-1)) {
  mean[i] <- genres_rating %>% filter(genres_rating[[i]] == 1) %>% pull(rating) %>% mean
}
genres <- names(genres)
df <- data.frame(genres, mean)
ggplot(df, aes(x=genres, y=mean)) +
  geom_bar(stat="identity") +
  coord_flip() +
  xlab("Genre") +
  ylab("Average Rating") +
  ggtitle("Average Rating by Genre")
```



Rating by Age Group

```
ggplot(ratings, aes(x = age_group, fill = as.factor(rating))) +
  geom_bar() +
  ggtitle("Rating Count by Age") +
  xlab("Age") +
  ylab("Total Count") +
  labs(fill = "Rating")
```



Split train and test

```r
# Split ratings back into train and test
ratings_train <- merge(ratings_train, ratings, by=names(ratings_train))
ratings_test <- merge(ratings_test, ratings, by=names(ratings_test))
```

Model Training

```r
lm(rating ~., data=ratings[,-c(4,5)])
base_model <- lm(rating ~., data=ratings[,-c(1,2,4,5,29:32)])
summary(base_model)
model1 <- lm(rating ~., data=ratings[,-c(4,5,27:32)]) # .34
model2 <- lm(rating ~., data=ratings[,-c(1,4,5,27:32)]) # .2112
model3 <- lm(rating ~., data=ratings[,-c(2,4,5,27:32)]) # .28
```