678 Final Project Report

true

Part1 Abstract

In the database recommended by Japanese animation, I found data about a large number of animation ratings, including numerical data such as animation scores, number of participants, and variables of more than 10 groups such as producer and source. This report will use multilevel modeling to analyze the how popularity and number of community fans and other variables affect the rating of anime (and some interesting conclusions/guess about anime-rating). The 10 page pdf is not enough for all graphs and analysis. There is a larger one in the github.

Part2 Introduction

The name of the data set from Kaggle is called "anime recommendation", which contains information on user preference data from 73,516 users on 12,294 anime. In the excel from the data, some useful variables can be used to find which variables can affect the rating of animes and how they exactly affect the rating. The following contents are the explanation of each variable.

Anime_id: myanimelist.net's unique id identifying an anime. Title: full name of anime. Type: Movie, TV, Special, etc. Producer: Different producer companies that produce the anime. Studio: The creator company of anime. Rating: Average rating out of 10 for this anime. ScoreBy: Number of people who rate the anime. Popularity: The popularity of the anime(the lower number means more famous). Members: Number of community fans that are in this anime's "group". Episodes: How many episodes in this show. (1 if movie). Source: The source of the anime, including Manga, original, etc. Aired: The date that the anime start to show.

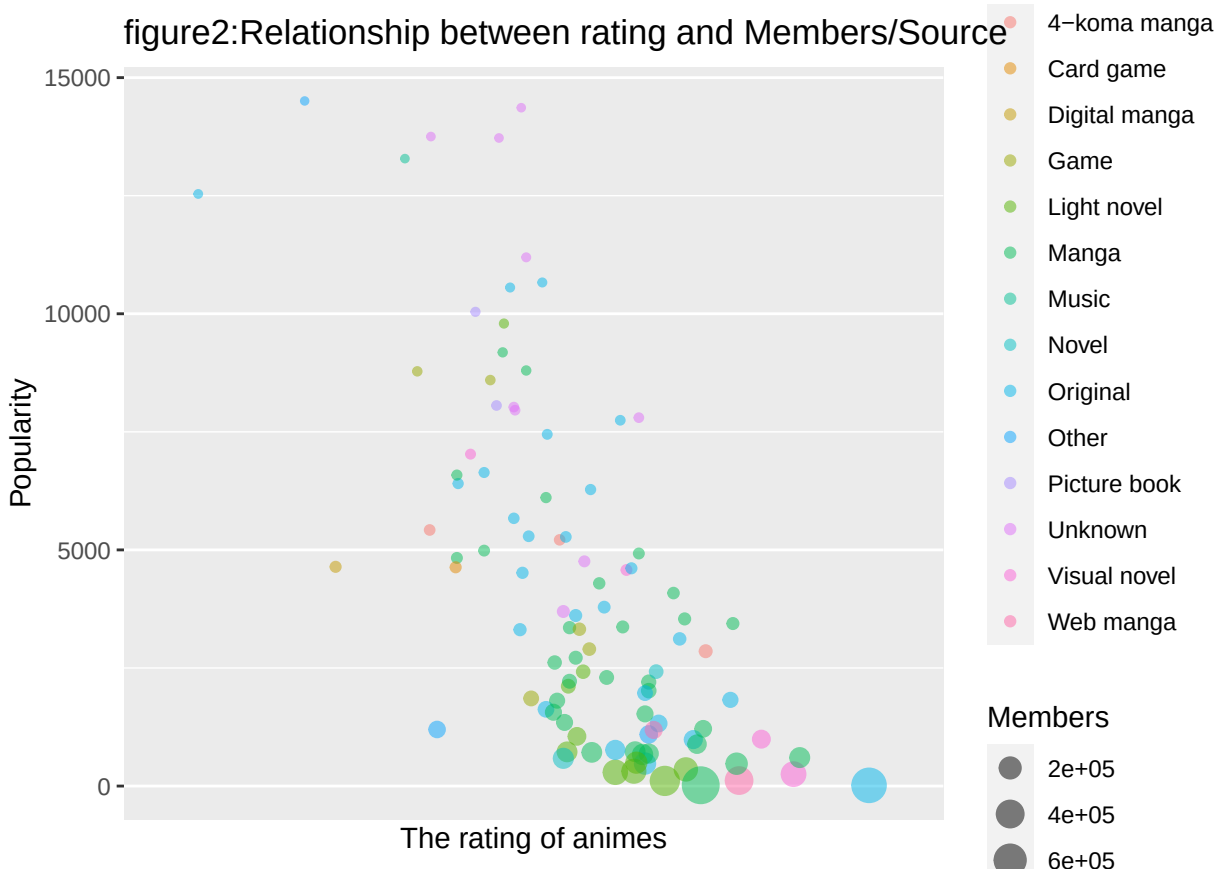
Part3 Methods

Data Cleaning

In the uploading part, I delete the missing data and NAs.

EDA

Before the linear regression and modeling, I first use some variables mentioned above to make graphs to predict which methods are more useful in the next parts.

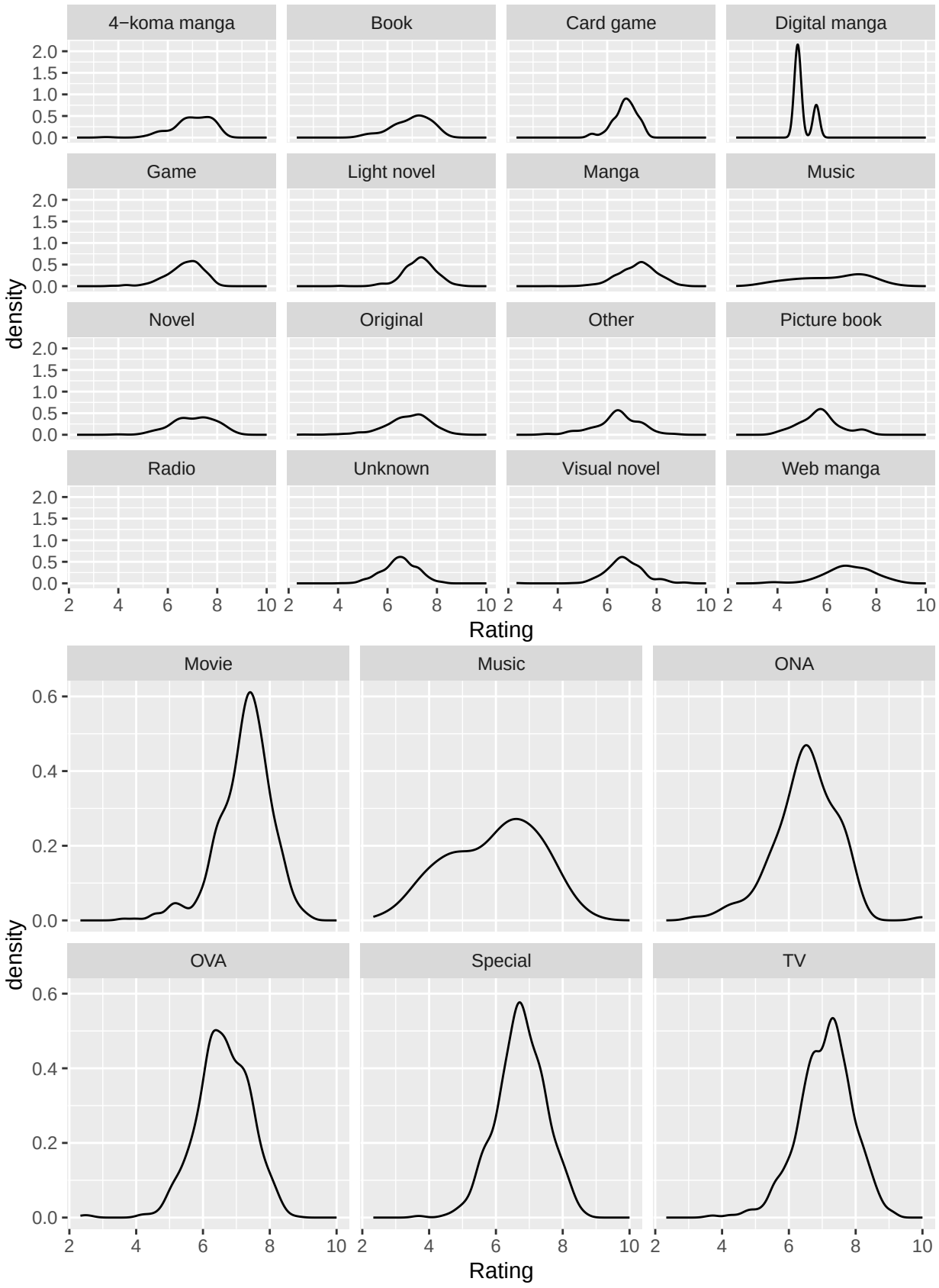


In figure 1, we can find there are key variables that may affect the rating.

Multilevel Modeling

Then, based on the requirement, i use the multilevel model to analysis and predict the data. First, I check the distributions by type/source.

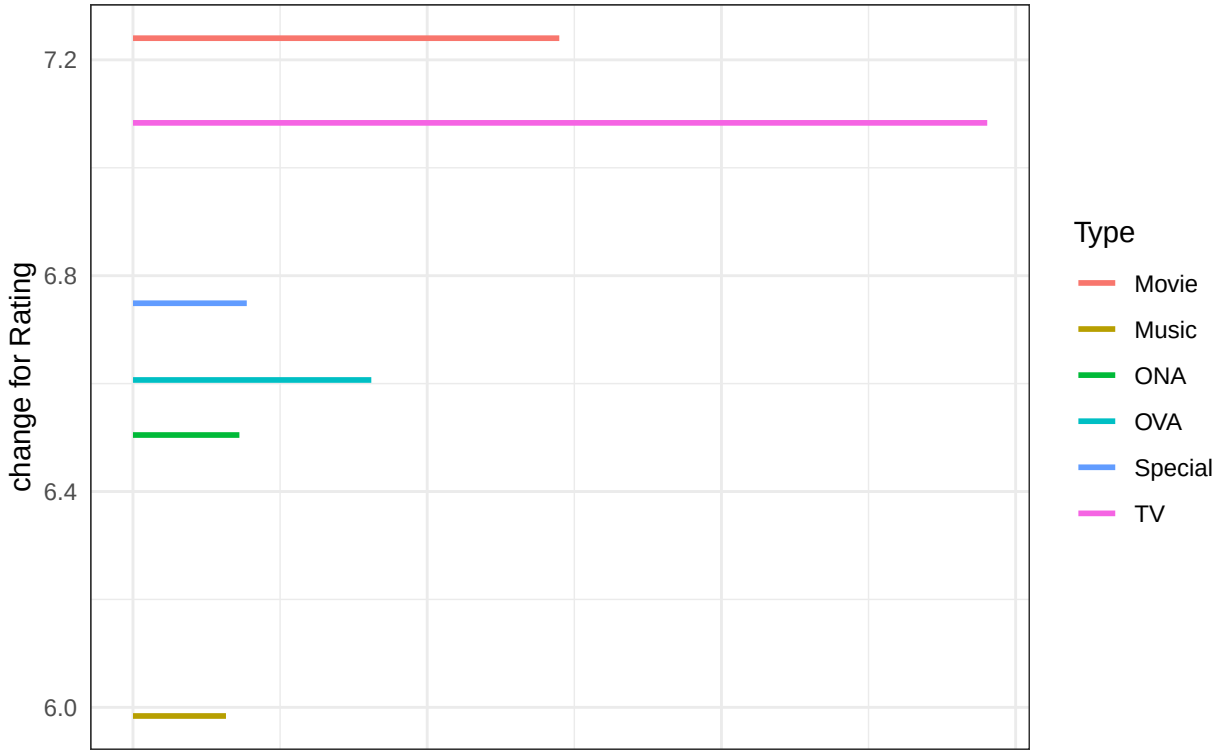
```
## # A tibble: 16 x 4
##    Source         mean    SD  miss
##    <chr>         <dbl> <dbl> <dbl>
##  1 4-koma manga   7     7        0
##  2 Book           6.97  6.97     0
##  3 Card game      6.72  6.72     0
##  4 Digital manga  5.01  5.01     0
##  5 Game           6.68  6.68     0
##  6 Light novel    7.29  7.29     0
##  7 Manga          7.23  7.23     0
##  8 Music          6.19  6.19     0
##  9 Novel          7.05  7.05     0
## 10 Original       6.83  6.83     0
## 11 Other          6.41  6.41     0
## 12 Picture book   5.72  5.72     0
## 13 Radio          6.71  6.71     0
## 14 Unknown        6.5   6.5      0
## 15 Visual novel   6.72  6.72     0
## 16 Web manga      6.88  6.88     0
```



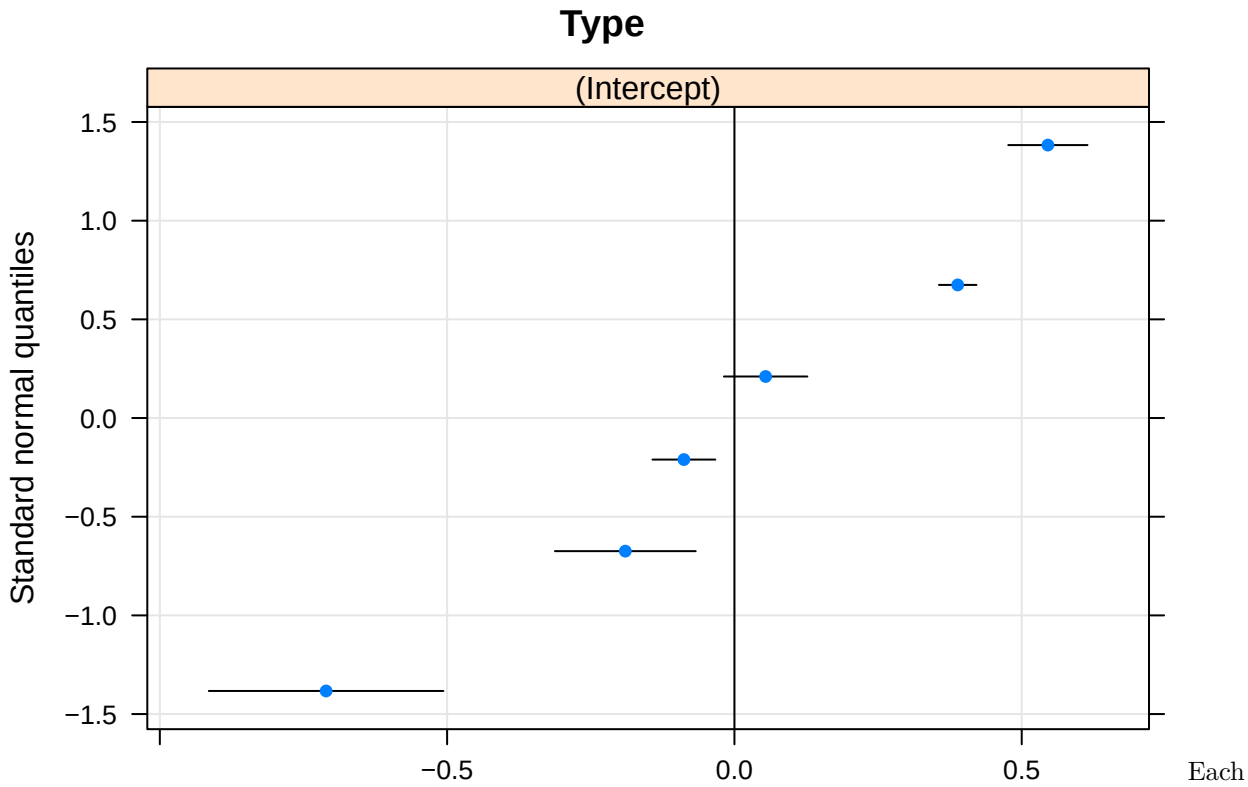

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
```

```
## Formula: Rating ~ 1 + (1 | Type)
##    Data: data2
##
##      AIC      BIC   logLik deviance df.resid
##  11074.3  11093.6  -5534.2  11068.3     4521
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.2171 -0.5895  0.0571  0.6723  4.2638
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Type     (Intercept) 0.1698   0.4121
##  Residual             0.6720   0.8198
## Number of obs: 4524, groups:  Type, 6
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   6.6946     0.1698   39.44
```

figure4:Prediction for Rating based on Type and Members

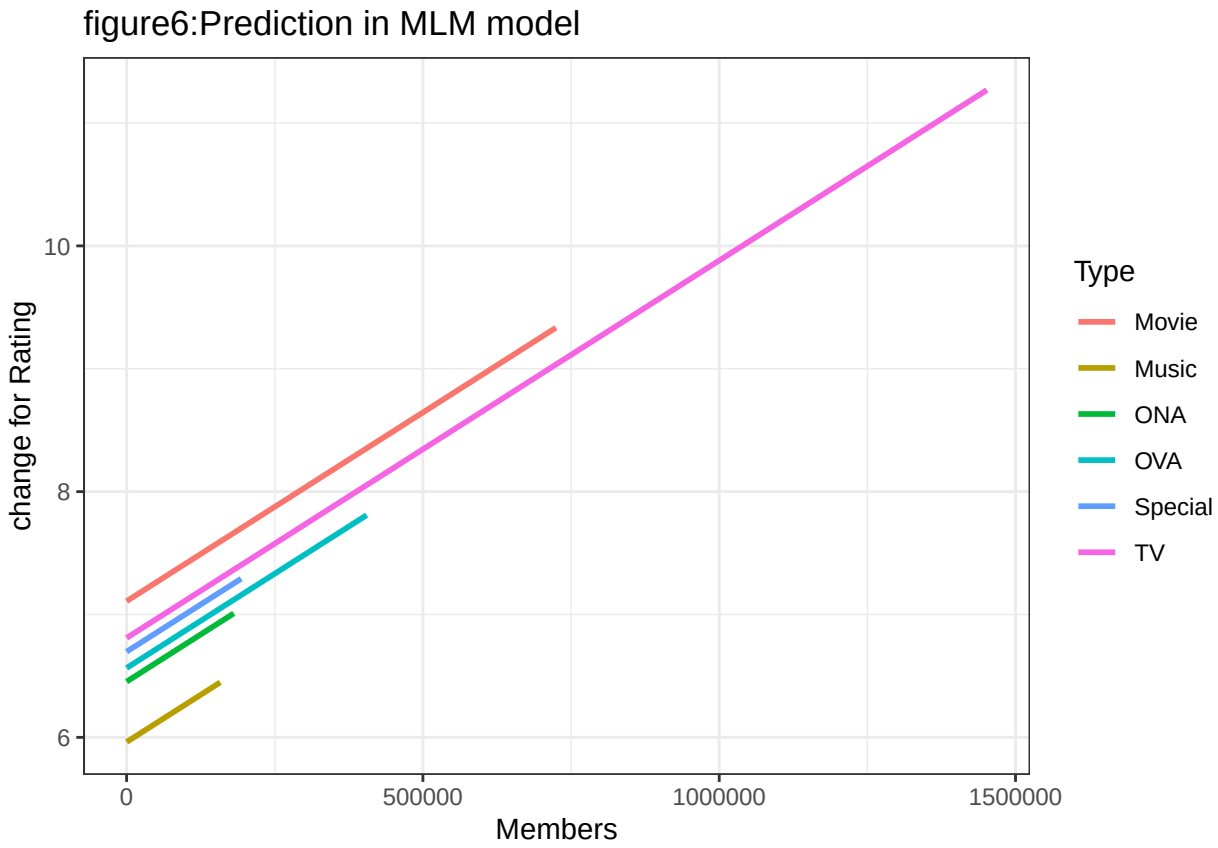



```
## $Type
```



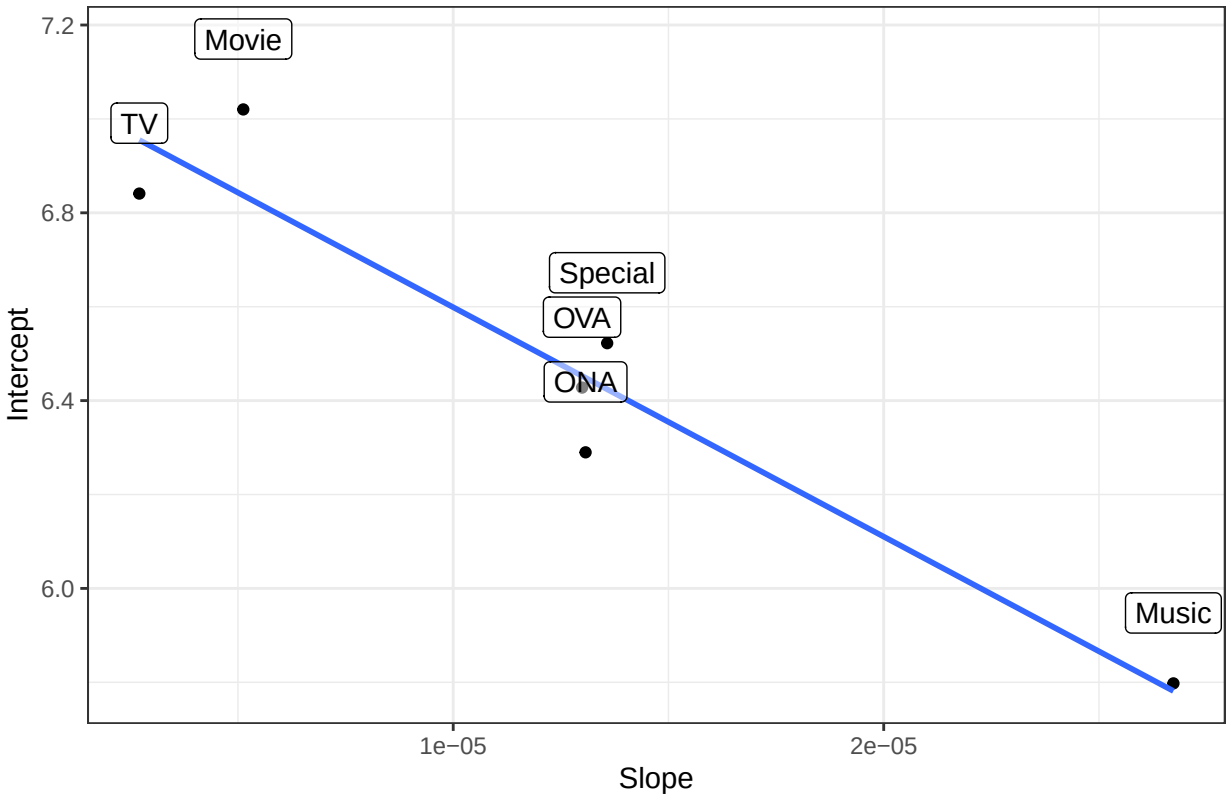
Each Type/Source is able to have a different increasing/decreasing of rating. Different Type and Source will affect the changing of rating.

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Rating ~ 1 + Members + (1 | Type)
##    Data: data2
##
##      AIC      BIC   logLik deviance df.resid
##  10171.7  10197.4  -5081.9  10163.7     4520
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.8508 -0.5443  0.0744  0.6550  4.7800
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Type     (Intercept) 0.1260   0.3549
##  Residual             0.5503   0.7418
## Number of obs: 4524, groups:  Type, 6
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 6.599e+00  1.464e-01   45.08
## Members     3.072e-06  9.708e-08   31.64
##
## Correlation of Fixed Effects:
##         (Intr)
## Members -0.021
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```



Here we see that as Members of group increases by 1 unit the rating increases by 3.072e-06 and 2.932e-06. We also notice that the random effects are smaller compared to the previous model. This indicates that number of members for group is explaining some variation.

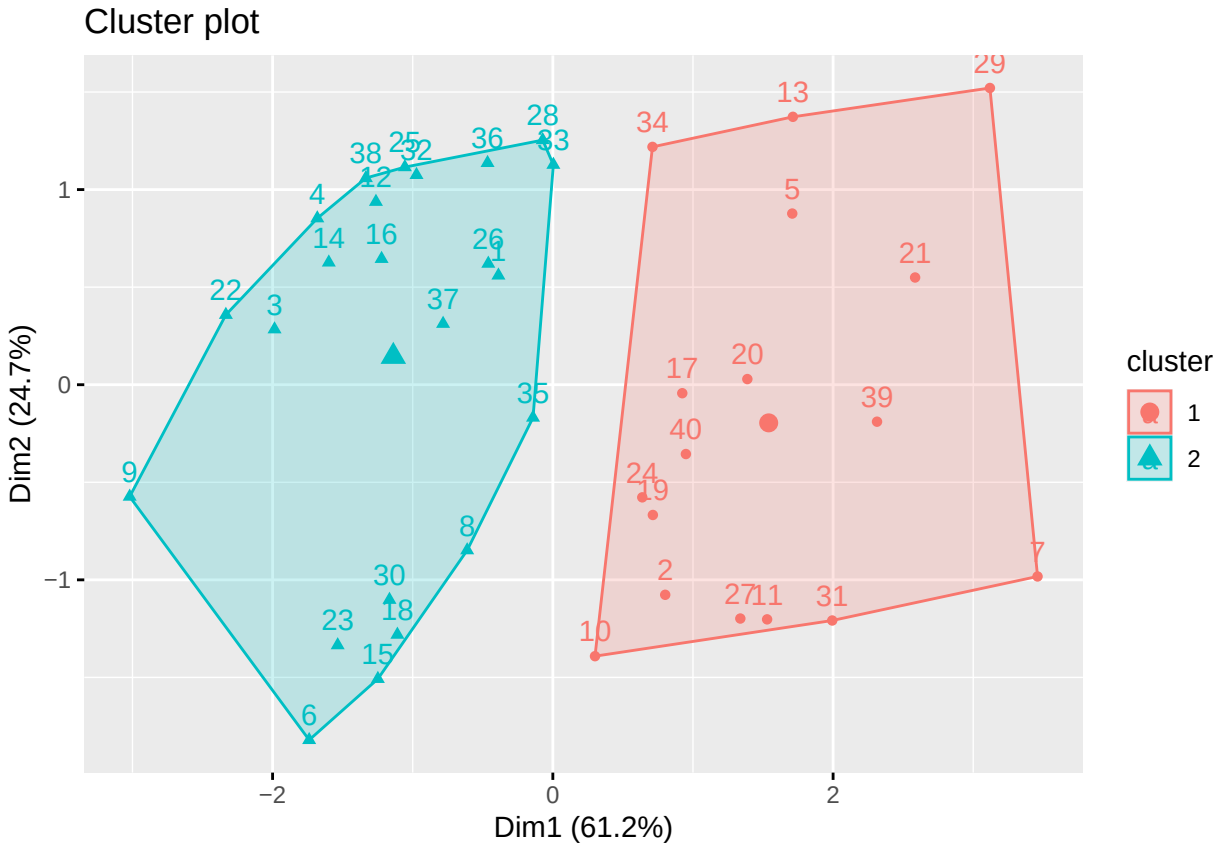
figure10 :effect of members and rating



In figure The graphs shows that we can divide the types into 3 parts. TV and Movie have most members and high increasing in rating but low effect of members to outcome. The Music has low support in rating but members for music has strong effect. The interpret ion will be written in the result/discussion part.

K-means algorithm

For this part, I want to use k-means to cluster observations and want observations in the same group to be similar and observations in different groups to be dissimilar. Due to the page limit, only the graph will be shown.



Word Could and Tpoic Modeling

Finally, I tried what I learn in 615 to analysis the non-numerical data. Due to the limit, only 1 page is shown.

```
## # A tibble: 46 x 2
##    word          n
##    <chr>     <int>
##  1 Comedy     5272
##  2 Action     3404
##  3 Fantasy    2874
##  4 Adventure  2558
##  5 Drama      2350
##  6 Fi         2259
##  7 Sci        2259
##  8 Kids       2249
##  9 Shounen    1784
## 10 Music      1717
## # ... with 36 more rows
```

Part4 Results

Based on the modeling above, there results are: 1. The estimated on numerical variables is 0.009475,0.267762, 0.040018 per unit increase; The linear model works for the three variables, but since the smaller popularity value means more popularity rank, its effect to rating is much smaller than members and scored people. 2. Multilevel modeling distribution works well on type/source, for the members value as the x-label, some specific type/source are more effective in rating, like OVA, Special/ Manga and game.The random effects are changing for different model. This indicates that number of members for group is explaining some variation. 3. There are some words that show high frequency in the genre part, which means high rating animes are often related to some of the words. 4. There are limit of space for this part. The full figures can be seen in another pdf.

Part5 Discussion

Based on the results, especially for the prediction part, the members and popularity are the most 2 important numerical variables that affect the rating. More popularized and discussed anime are easier to get rating higher than average. However, the groups of type and source may be different in different situations. For example, TV type of animes have more members to discuss, but the effect to increase the rating is not that important. In the future, the type of animes will still be TV-mained, but if more people start to become members of other types, there rating will increase faster than TV. Also, it will be harder to get higher rating when evaluating the date of anime(no space to show in the pdf) and about the genre, animes about comedy, action and advanture will still be the main topics for anime.