Anime Popularity Analysis

Zening Ye

11/22/2021

Abstract

In this project, I will use the data from 2014 to 2017 to analyze the following questions: first, what source(s) of Anime affected the audience more in 2014 to 2017? Second, why did the Anime come up a lot in 2014 to 2017? Last but not least, what type(s) of Anime will be the mainstream in the future? Therefore, I conducted a multilevel analysis for the data I have, and tried to provide a top Anime list under the result I got after the analysis. Based on the result, it is hard to identify these questions, since the model fitted was not good enough. In addition, I did not have users' data to combine with Anime data, users' data might help to identify several factors, such as the period of episode, favorite, etc. I also concluded some main aspects that might affect the popularity of Anime based on the current analysis result. Last but not least, as a Anime fan, I created a list of Anime for recommendation after this analysis.

Introduction

Anime has become more and more popular for the past 5 to 6 years. The original word "Anime" means a hand-drawn and computer animation originating from Japan. In Japan, the Anime includes all animated works, regardless of style or origin. There are plenty of Anime works in the world, such as Movie, Music, ONA, TV, etc. However, some Anime did not have more popularity than others, and they did not have more shows on TV or other.

There are plenty of types, sources and genres in Anime, therefore there will be a lot of combinations whether in TV, Movie, OVA, etc. There are some variables I would like to use to fit the model, such as start day, types, genres, sources, rating, popularity and score. These variables might help me to get some direction for the questions I mentioned above. For instance, how the genres might affect the popularity of the source(manga, original, novel), how rating and score will affect the popularity as well. Furthermore, how will these variables influence the future mainstream of the Anime?

Based on these factors, I would like to use multilevel models to illustrate how these factors affect the mainstream and popularity in single or multiple genres in the future.

Method

Data Processiong and Cleaning

The data was came from Kaggle: Anime dataset:. The dataset I have chosen was cleared, however, it was not good enough for me to move to the next steps. Therefore, I first tidy up the data. I removed the variables that I will not use for the analysis in this report such as title, title_english, synopsis, etc. Second, I renamed some columns for easy access in the future, such as arid_from to star_year. Since I'm focusing on the data from 2008 to 2020, I filtered the data from the original dataset and created a new dataset for exploratory

data analysis, which might help to figure out the difference from two different timelines. After cleaning the datasets I will use for the further analysis is listed below:

Column	Description
title	Name of the Anime
year_sta	Date on Aired
duration	Duration of Each Episode in min and hour
episodes	Number of episodes in Anime
genres	Theme of the Anime
popularity	Famous Level
rank	Rank of the Anime
rating	Level of the Anime, eg. PG-13, R, PG
score	Numerical Level for Anime
scored_by	Population of the rating
source	Category of the Anime
type	Type of the Anime

Exploratory Data Analysis

As I mentioned above I will use the data from 2008 to 2020, since during this period, more and more Anime came up, the popularity, rating, and sources are dynamic. The plot below indicates the frequency of the Anime came up in 12 years.

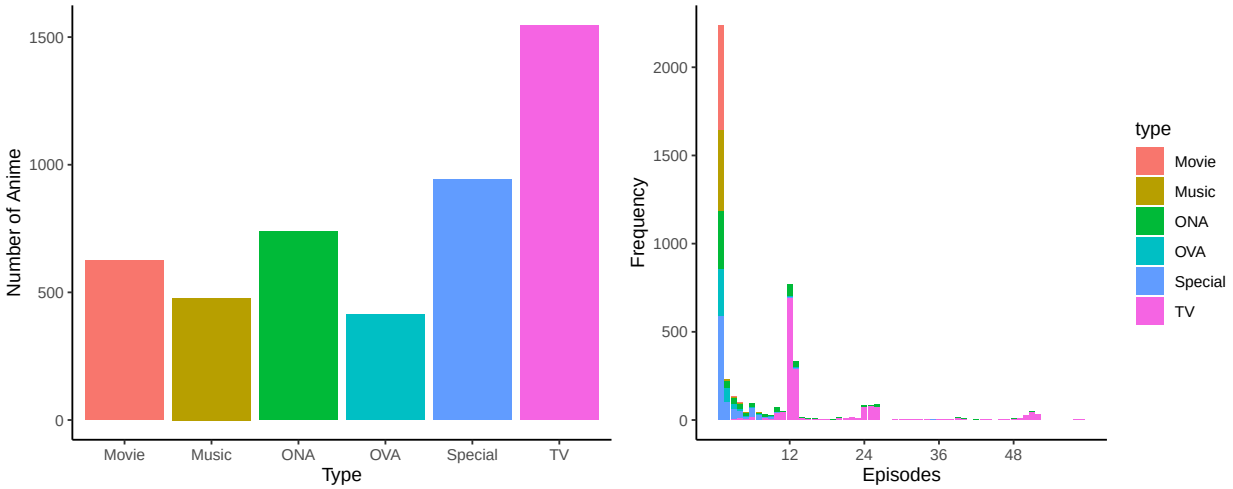

Figure 1: Frequncy of Anime in 2008-2020

For Figure 1, it is obvious to see the type of Anime in TV with episodes under 24 has more popularity and market share in 2008 to 2020. In addition, the second plot in Figure 1 basically represents the entire then entire stream type of Anime in these years.

For Figure 2, in order to show the distribution of Anime in different types, I filtered the original dataset into six different subsets by different types. For instance, I used TV and Movie as examples to illustrate the frequency of source. Apparently, Original and Light Novels are the most famous topics in TV and Movie.

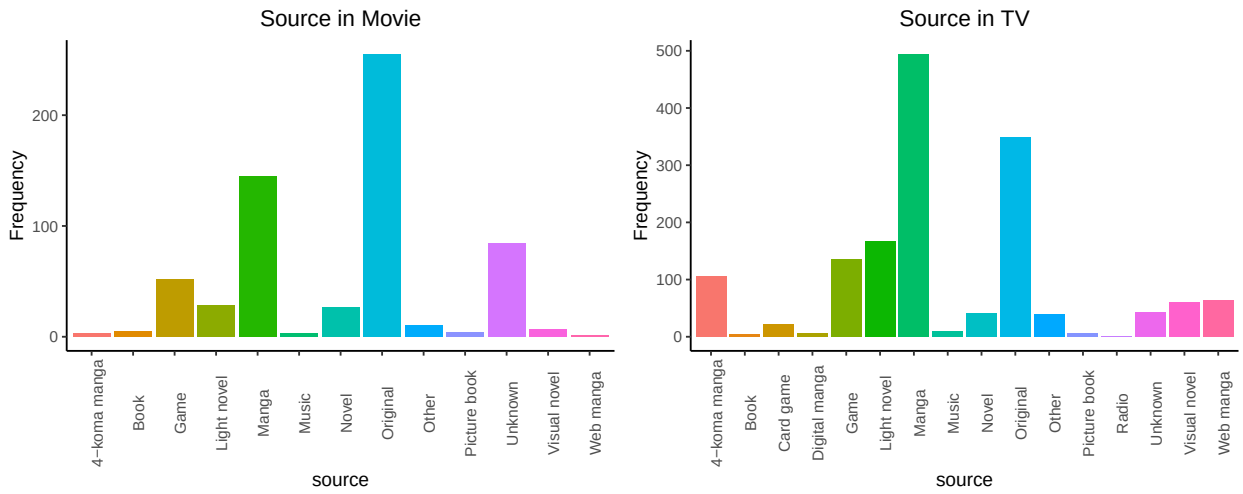

Figure 2: Frequency of Anime in Different Sources

Model Fitting

Before I fit the multilevel model, I would like to use the subset I made above to fit a glm model to check their relationship. Since TV and Movies are the most effective sources on Anime, it might help to visualize the relationship. The variables will include score, rating, source, and rank. Since the popularity level might be too large for fitting the model, I use log(popularity) to fit the linear model.

For two two different subsets I fitted two different multilevel models, the variables I use for multilevel are rating, score, rank, and $rating : score$, for the random effect is $(1|rating)$ and $(1|source : type)$. I also used two individual plots to check the model, "Residuals vs. Fitted" and "QQ Plot". The two multilevel models can be written as follow:

$$\begin{aligned} log(popularity) =& 16.74 + 1.336 \cdot ratingNone + 1.302 \cdot ratingPG + 2.401 \cdot ratingPG13 \\ & - 1.061 \cdot score - 0.0002 \cdot rank - 0.219 \cdot ratingNone : score \\ & - 0.200 \cdot ratingPG : score - 0.415 \cdot ratingPG - 13 : score \\ & + n_j + i_j + \epsilon \end{aligned}$$

$$n_j \sim N(0, \sigma_a^2), i_j \sim N(0, \sigma_b^2)$$

where n_j and i_j are random effects: $(1|rating)$, $(1|source : type)$

$$\begin{aligned} log(popularity) =& 12.26 - 2.495 \cdot ratingNone - 1.699 \cdot ratingPG + 1.094 \cdot ratingPG13 \\ & - 5.631 \cdot score - 0.0001 \cdot rank + 0.332 \cdot ratingNone : score \\ & + 0.265 \cdot ratingPG : score - 0.283 \cdot ratingPG - 13 : score \\ & + n_j + i_j + \epsilon \end{aligned}$$

$$n_j \sim N(0, \sigma_a^2), i_j \sim N(0, \sigma_b^2)$$

where n_j and i_j are random effects: $(1|rating)$, $(1|source : type)$

Therefore, we can interpret these two models by using, for instance, rating levels. For every one unit change in ratingPG13, when other variables are constant, the log(popularity) will increase 2.401 in Movie. Same interpretation with TV, for every one unit change in ratingPG13, when other variables are constant, the log(popularity) will increase in 1.094.

According to the residual analysis (under Appendix), even though I used two different subsets to fit the multilevel model, the "Residuals vs. Fitted" plots indicated that the residual is not good to fit the model, so

it might conclude there is not a lot of correlation between the predictors. On the QQ plots, there are plenty of residual points not on the lines, so it might not follow the normal distribution.

Result

Focusing on the model fitting and EDA I have done above, unfortunately, it is hard to identify the variables that might affect the popularity of Anime. In other words, these variables do not have a significant effect on the final results. Although I derived the Anime genre share from 2008 to 2020 from different subsets in figure 2, I still could not reach a valid conclusion. Moreover, I have analyzed rating and even though it has a significant effect in all variables, it has little effect on the overall model. The current result I have for two individual model is: 1) Movie: for every one unit change in ratingPG13, when other variables are constant, the log(popularity) will increase 2.401 in Movie; 2) TV: for every one unit change in ratingPG13, when other variables are constant, the log(popularity) will increase in 1.094. Furthermore, I made a random effect plot to show how it will affect the model (Appendix).

The popularity of Anime was not affected by the current variables I have on my data frame, but also has other aspects that might impact the entire Anime. For instance, the sources of Anime have increased a lot in light novels and manga. After 2008, The resources for light novels are growing exponentially, more and more light novels are being animeized or a series of anime peripherals are appearing. Typical examples include Sword Art Online, The Pet Girl of Sakurasou, The Familiar of Zero, Hyouka, etc. The emergence of these light novels has brought about a huge shift in the entire Anime market, which was previously dominated by manga and originals. The light novels have impacted the original balance, making the Anime market more and more competitive and enriching the daily life of Anime fans.

Not only for the light novel source growing exponentially, but also for the original and manga source. Kyoto Animate, a great Anime company in the world, created plenty of great Anime for us, such as “Miss Kobayashi’s Dragon Maid”, “Love, Chunibyo & Other Delusions”, and Nichijou, etc. I believe more and more light novels and manga will become mainstream in the future, even though it is hard to show under my analysis. Therefore, I created a list for recommendations to others if they want to get into Anime more.

Anime Name	Type	Year
The Pet Girl of Sakurasou	Light Novel	2008
Hyouka	Light Novel	2011
Sword Art Online	Light Novel	2013
Seishun Buta Yarō	Light Novel	2019
My Youth Romantic Comedy Is Wrong, As I Expected	Light Novel	2011
Overload	Light Novel	2012
Eromanga Sensei	Light Novel	2013
Miss Kobayashi’s Dragon Maid	Manga	2013

As you can see from the table, many of them are light novels which means it will become more and more famous in the future, and that is the reason why I enter the area of Anime.

Discussion

Overall, the entire analysis was not good enough, even though the data looks good. The data I selected from Kaggle, I used rating, scour, rank as variables and source:genres as random effect for my multilevel model. Unfortunately, as I mentioned above it is unlikely to answer the questions I ask at the beginning of the report. After checking the model and data, I think there are some limitations for the data. First, the data provide the source level, which is helpful, I thought, for the model fitting. However, it did not provide the feedback of the user to indicate why it is so popular. Second, the variables I used might not be enough for model fitting, but the rest of variables were hard to fill in the lmer function. Last but not least, even though we have the count for type of Anime, we do not have actual market sale data for the Anime, including the type, rating, genres. This data might help to fit a better model in the future.

For the next step, I would like to collect the actual market data that I just mentioned, also I will try to include the market share data since it is more reasonable to indicate the popularity of Anime. In addition, I would like to learn more about modeling fitting and model design, because I realized I have lacked experience with designing a good model for data analysis.

Reference:

Kaggle - Anime dataset: https://www.kaggle.com/thunderz/anime-dataset

Appendix

Liner Model fitting

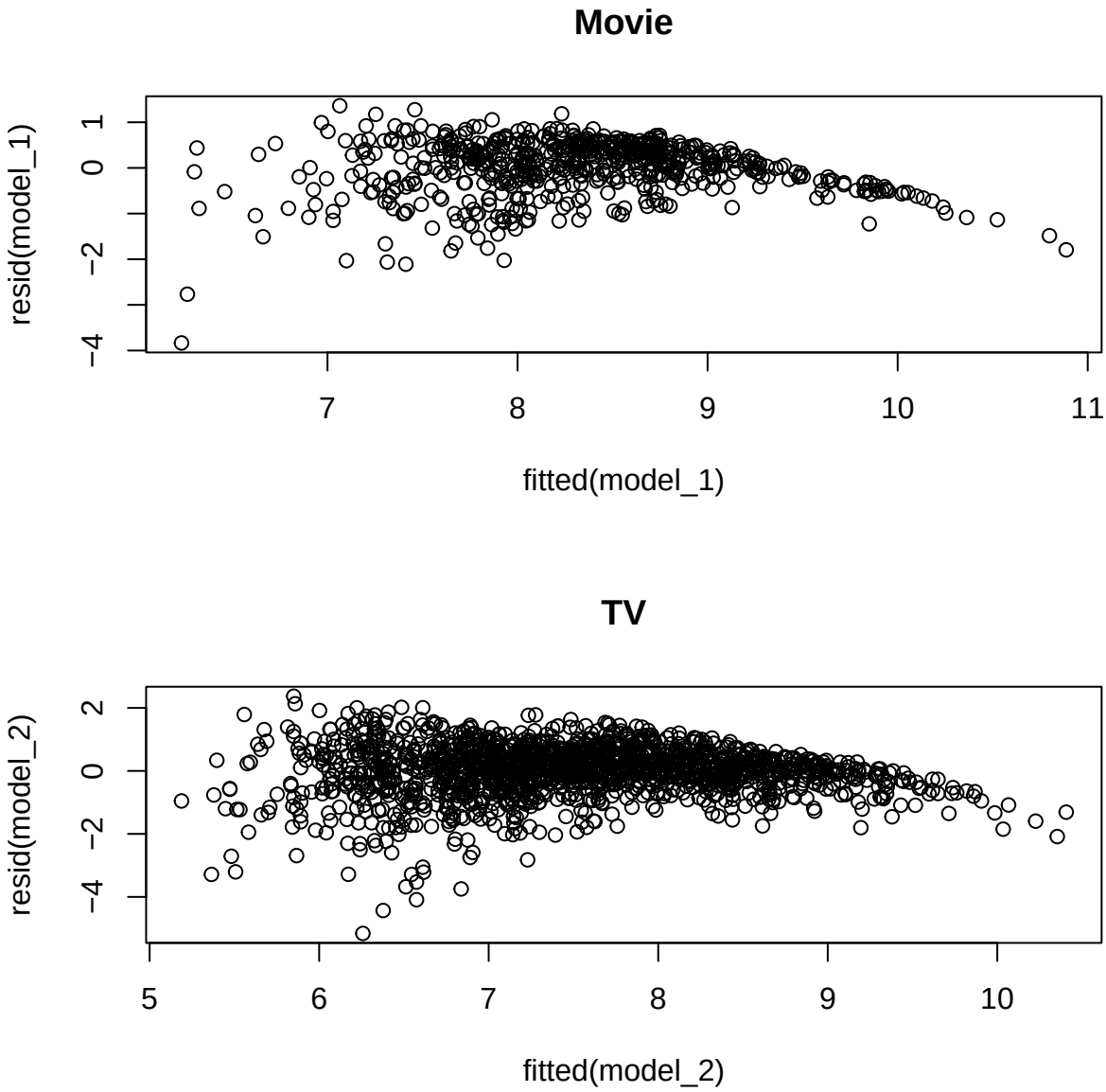

Figure 3: Residual Plot

More EDA

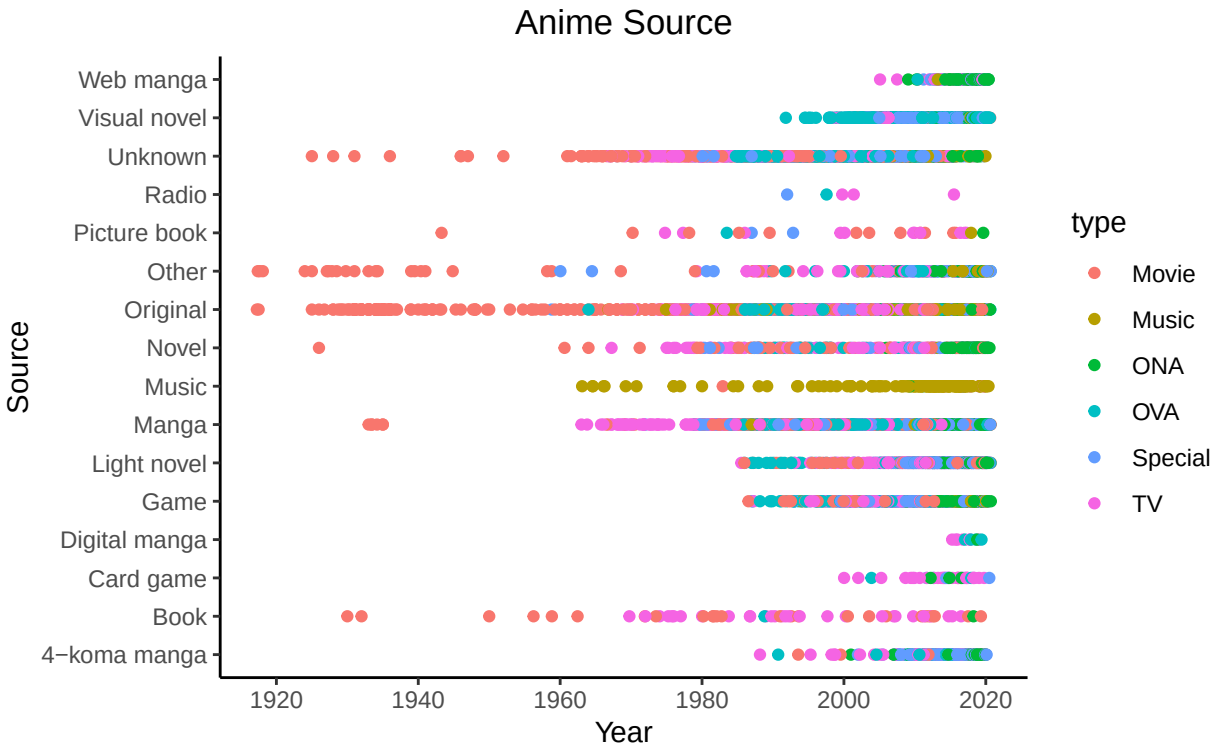

Figure 4: Original Dataset

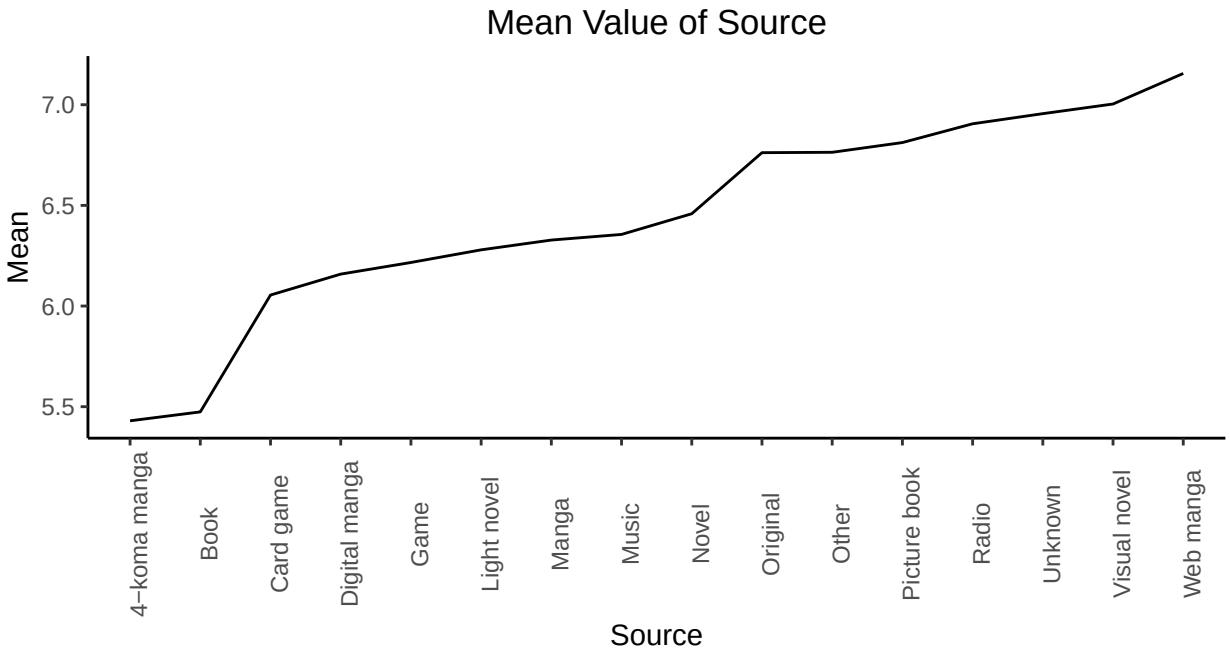

Figure 5: Mean Value of Each Source

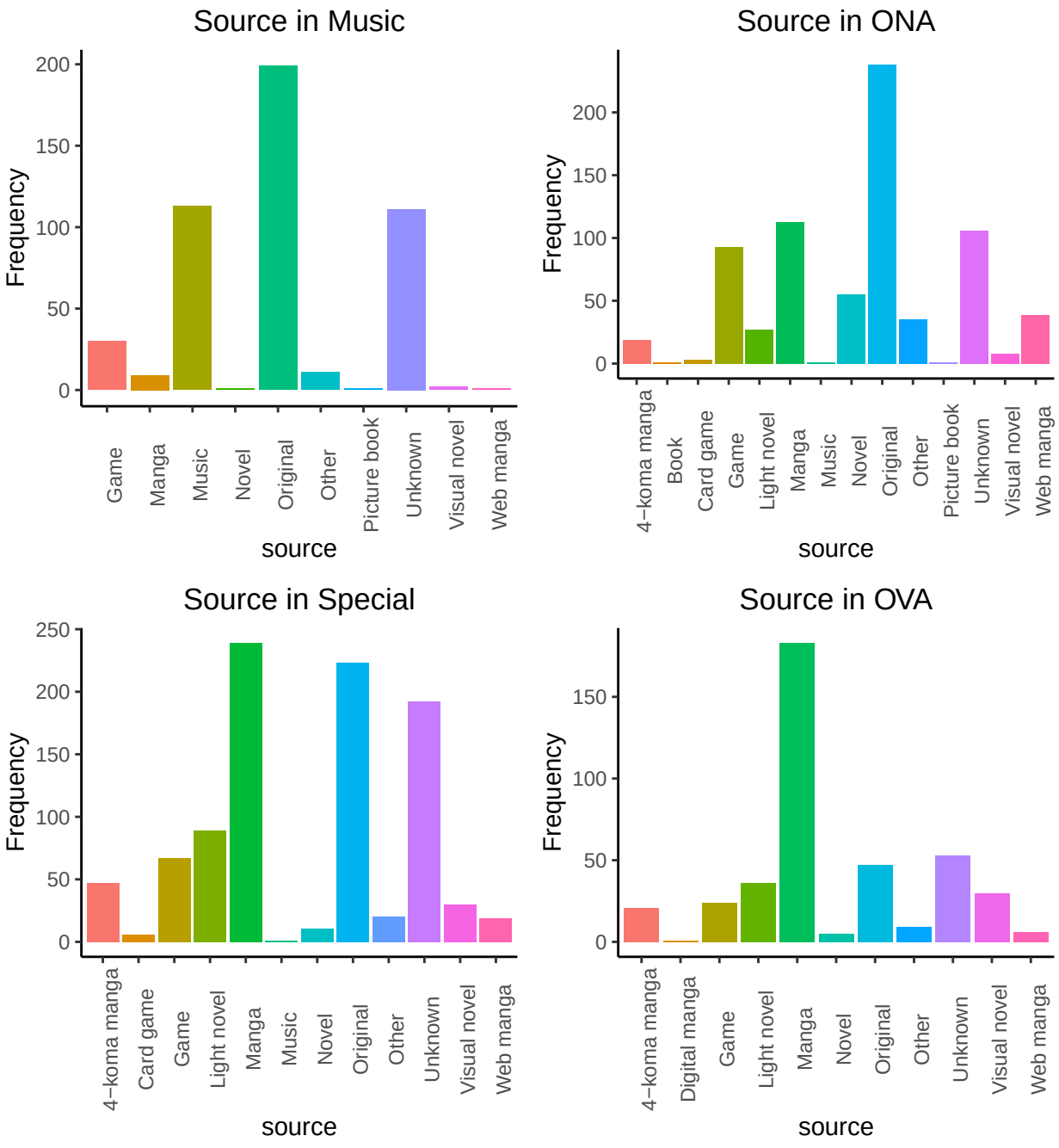

Figure 6: Different Type of Anime

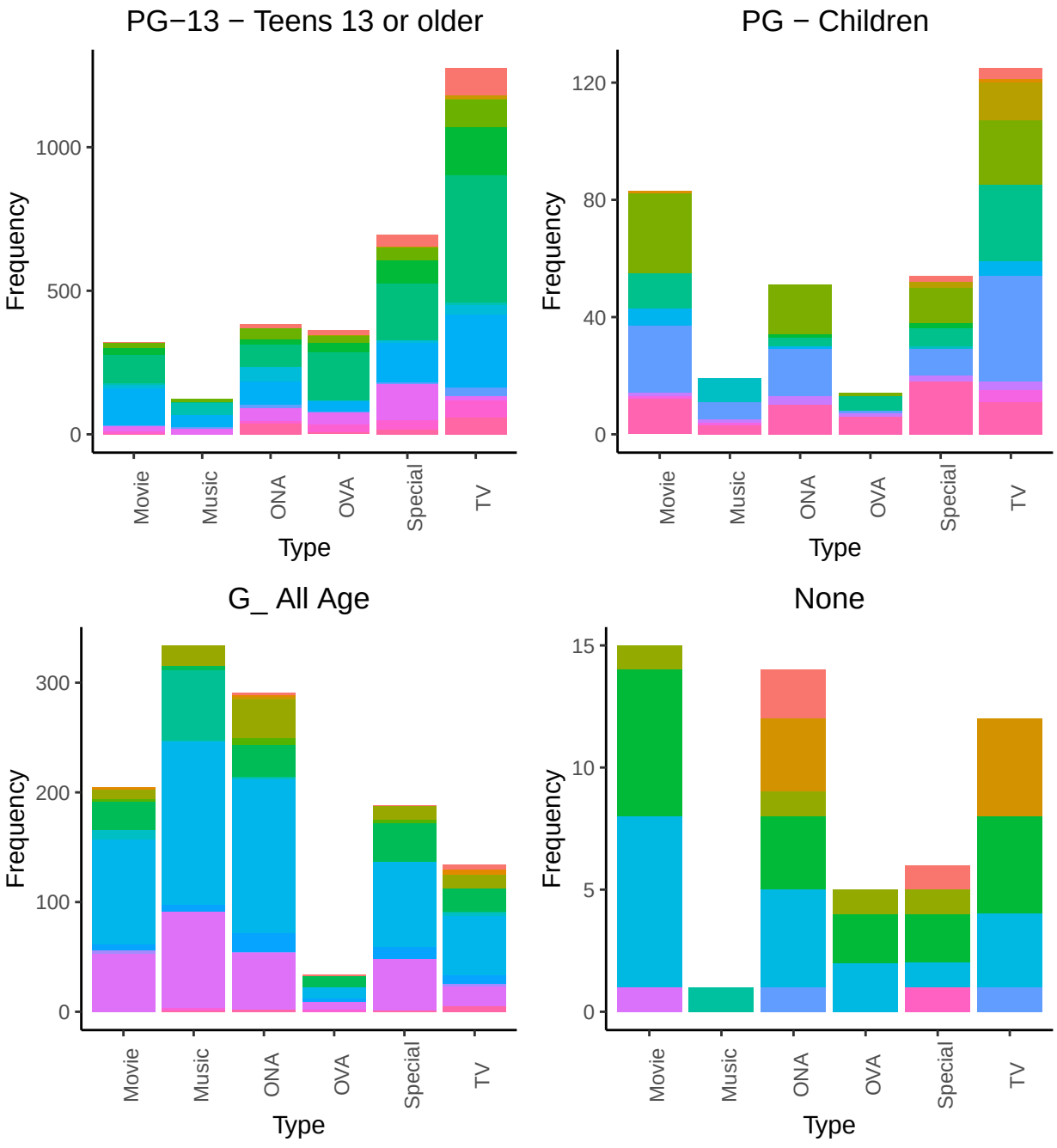

Figure 7: Rating Level

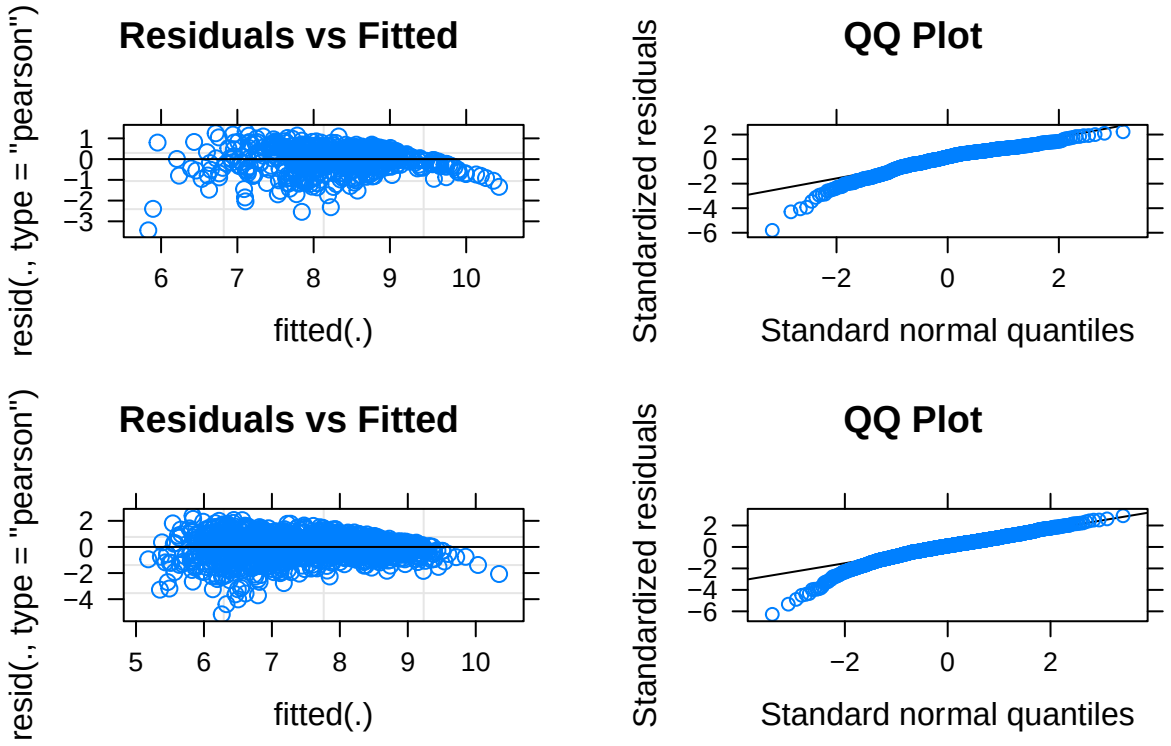

Figure 8: Type of Anime: Movie and TV

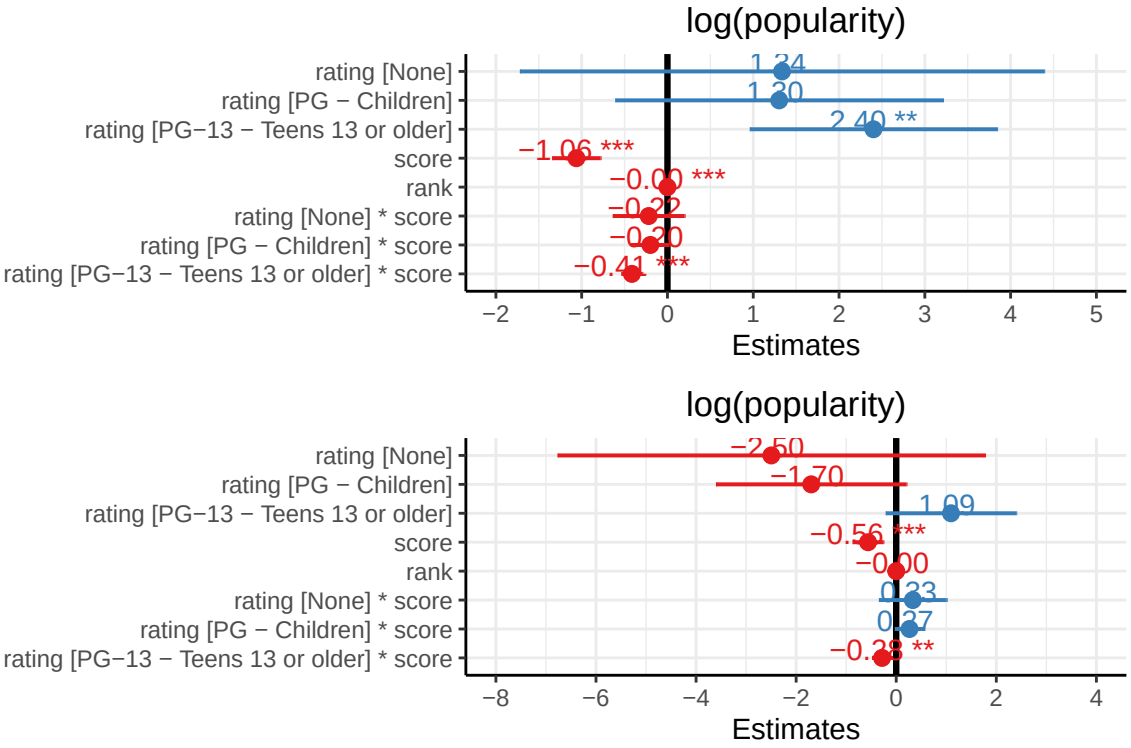

Figure 9: Random Effect Plots

Result for model fitting:

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      log(popularity) ~ rating + score + source + rank
##  observations: 624
##  predictors:   18
## ------
##                                Median MAD_SD
## (Intercept)                    18.2    1.2
## ratingNone                      0.0    0.2
## ratingPG - Children             0.0    0.1
## ratingPG-13 - Teens 13 or older -0.4    0.1
## score                          -1.3    0.1
## sourceBook                      0.1    0.4
## sourceGame                      0.0    0.4
## sourceLight novel              -0.4    0.4
## sourceManga                     0.0    0.4
## sourceMusic                    -0.6    0.5
## sourceNovel                     0.0    0.4
## sourceOriginal                  0.0    0.4
## sourceOther                    -0.2    0.4
## sourcePicture book              0.5    0.5
## sourceUnknown                   0.1    0.4
## sourceVisual novel             -0.3    0.4
## sourceWeb manga                -0.4    0.7
## rank                            0.0    0.0
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 0.6    0.0
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      log(popularity) ~ rating + score + source + rank
##  observations: 1547
##  predictors:   21
## ------
##                                Median MAD_SD
## (Intercept)                    14.5    1.0
## ratingNone                     -0.5    0.3
## ratingPG - Children             0.0    0.1
## ratingPG-13 - Teens 13 or older -0.8    0.1
## score                          -0.9    0.1
## sourceBook                      0.3    0.4
## sourceCard game                 0.4    0.2
## sourceDigital manga            -0.5    0.3
## sourceGame                      0.0    0.1
## sourceLight novel              -1.0    0.1
## sourceManga                    -0.1    0.1
```

```
## sourceMusic                      0.2    0.3
## sourceNovel                      0.4    0.2
## sourceOriginal                   0.1    0.1
## sourceOther                      0.1    0.2
## sourcePicture book              -0.1    0.3
## sourceRadio                      0.2    0.8
## sourceUnknown                    0.4    0.2
## sourceVisual novel              -0.3    0.1
## sourceWeb manga                 -0.1    0.1
## rank                             0.0    0.0
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 0.8    0.0
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg

## Linear mixed model fit by REML ['lmerMod']
## Formula: log(popularity) ~ rating + score + rank + rating:score + (1 |
##     rating) + (1 | source:type)
##    Data: movie
## REML criterion at convergence: 1159.385
## Random effects:
##  Groups      Name        Std.Dev.
##  source:type (Intercept) 0.09877
##  rating      (Intercept) 0.42116
##  Residual                0.59168
## Number of obs: 624, groups:  source:type, 13; rating, 4
## Fixed Effects:
##                            (Intercept)                              ratingNone
##                              16.7388517                               1.3357979
##                    ratingPG - Children        ratingPG-13 - Teens 13 or older
##                               1.3023276                               2.4010011
##                                   score                                    rank
##                              -1.0613187                              -0.0001888
##                         ratingNone:score             ratingPG - Children:score
##                              -0.2186107                              -0.1996323
## ratingPG-13 - Teens 13 or older:score
##                              -0.4148522
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
## optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer warnings; 2 lme4 warnings

## Linear mixed model fit by REML ['lmerMod']
## Formula: log(popularity) ~ rating + score + rank + rating:score + (1 |
##     rating) + (1 | source:type)
##    Data: tv
## REML criterion at convergence: 3850.07
## Random effects:
##  Groups      Name        Std.Dev.
##  source:type (Intercept) 0.34724
##  rating      (Intercept) 0.01401
```

```
##  Residual                0.82225
## Number of obs: 1547, groups:  source:type, 16; rating, 4
## Fixed Effects:
##                            (Intercept)                                ratingNone
##                              1.226e+01                                -2.495e+00
##                    ratingPG - Children        ratingPG-13 - Teens 13 or older
##                              -1.699e+00                                 1.094e+00
##                                   score                                      rank
##                              -5.631e-01                                -1.215e-05
##                          ratingNone:score              ratingPG - Children:score
##                              3.319e-01                                 2.651e-01
## ratingPG-13 - Teens 13 or older:score
##                              -2.831e-01
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
## optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer warnings; 2 lme4 warnings

## $`source:type`
##                       (Intercept)
## 4-koma manga:Movie  0.016239410
## Book:Movie          0.003603308
## Game:Movie         -0.008057235
## Light novel:Movie  -0.096927143
## Manga:Movie         0.057601527
## Music:Movie        -0.045446424
## Novel:Movie         0.014777903
## Original:Movie     -0.005522407
## Other:Movie        -0.065628061
## Picture book:Movie  0.040352002
## Unknown:Movie       0.126531994
## Visual novel:Movie -0.028284875
## Web manga:Movie    -0.009239998
##
## $rating
##                              (Intercept)
## G - All Ages               2.614445e-12
## None                      -6.797908e-15
## PG - Children             -4.854468e-13
## PG-13 - Teens 13 or older  1.476291e-11
##
## with conditional variances for "source:type" "rating"

## $`source:type`
##                     (Intercept)
## 4-koma manga:TV   0.024680008
## Book:TV           0.110088667
## Card game:TV      0.257559202
## Digital manga:TV -0.293402682
## Game:TV          -0.009969698
## Light novel:TV   -0.883920043
## Manga:TV         -0.093774532
## Music:TV          0.170916491
## Novel:TV          0.336531849
## Original:TV       0.068902901
```

```
## Other:TV          0.124125561
## Picture book:TV   0.066273079
## Radio:TV          0.021719408
## Unknown:TV        0.428747454
## Visual novel:TV  -0.281072310
## Web manga:TV     -0.047405355
##
## $rating
##                             (Intercept)
## G - All Ages               7.822154e-15
## None                      -1.280418e-18
## PG - Children              2.103932e-16
## PG-13 - Teens 13 or older  1.175100e-13
##
## with conditional variances for "source:type" "rating"
```