Intro

MangaUpdates is an English database/wiki that tracks translations of manga, manhua, manhwa, and other international comics. The site allows users to save and review works of all types–including queer comics. In recent years, queer media has become more widespread and accepted. How do those queer media measure against their non-queer counterparts in the manga community?

This project aims to study comics of genres Yuri/Shoujo Ai (Girls Love, gay female stories), Yaoi/Shounen Ai (Boys Love, gay male stories), and compare them to non-queer comics in an exploratory data analysis. Furthermore, this project also explores characteristics within queer genres. How different are Boys Love and Girls Love from each other?

Data Description

Data was collected from around 26,000 series by calling the MangaUpdates API and filtering out series with missing data. All of MangaUpdates' Yuri and Shoujo Ai series were sampled, as were two calls worth of Boys Love and non-queer genres. Data attributes are listed in appendix **Figure 0.**

Genre

Of the works sampled, 62% (n = 17344) were queer in any way.
42.1% of all works were Boys Love (BL) (22% Shounen Ai, 25% Yaoi)
18.9% of all works were Girls Love (GL) (9% Shoujo AI, 11% Yuri).
3.9% of GL works had both GL tags, 1.3% of BL works had both BL tags, 0.3% of queer works had both BL and GL tags. **(Figures 1 and 2)** show the composition of these tags.

Series had ~3.5 genres on average, GL had slightly more and BL slightly less. Genre count had a mild positive correlation with category count (r = 0.39) and description length (0.34)

There was notable slight correlation between Shounen Ai and Adult (-0.27), Shounen Ai and Doujinshi (0.30), and Yuri and Adult (0.21) **(Figure 3).** Genres with positive correlations were more likely to appear together, negative correlations one or the other.

To enable smooth data analysis with independent groupings, a new column was derived from the genre variables. Mutual_exclu_queer was a column for all rows, with one of four values: **GL** (Tagged with Yuri and/Or Shoujo Ai but neither BL tag), **BL** (Tagged with Yaoi and/Or Shounen Ai but neither GL tag), **Both** (Tagged with at least one BL tag and at least one GL tag), and **Not Queer** (Tagged with no GL tags and no BL tags).

Series Info

80% of the series had a description. A higher proportion of BL series had a description. The average description length was 315 characters, with GL series averaging a slightly shorter description length. 7 series (mostly anthologies with multiple summaries) had descriptions > 5000 characters. Description length was heavily right-skewed.

The most common series types overall were Manga (53%), Doujinshi (32%), Manhwa (9%), Manhua (5%) and Novels (1%). All other series types were lumped into an "Other" category. Doujinshi made up a larger percentage of both queer groups (GL: 38%, BL: 49%). A slightly lower % of both GL and BL series were manga.

The most common release year was 2013 (6%), followed by 2014 (5%), and 2020 (5%). The oldest series sampled was released in 1947, and the newest in 2025. Release year was a left skewed variable. Queer series were more clustered around recent release years, visible in **(Figure 7).**

Only 65% of series had category tags, but the mean number of tags was 8.8. Queer series averaged slightly fewer category tags.
The most popular non-queer category tags were Full Color (11%), Webtoon (9%), 21st Century (7%), Web Comic (5%), and Older Female Younger Male (5%).
Among GL series, the most popular category tags were Full Color (6%), 21st Century (6%), LGBT Scenes (5%), High School Student (5%), and Web Comic (4%).
For BL series, the most popular category tags were High School Student (6%), Full Color (5%), Webtoon (5%), Web Comic (3%) and Kisses (2%).
Category tags used for less than 1% of all data were not considered in further analysis.

Only 33% of series had any related series, and only 5% had more than one relation. 24 works had outlier relation counts > 15. 55% of BL works had a related series, potentially linked to a high proportion of Doujinshi (often derived from a source material), as 37% of all BL were "Adapted From" another source. Among all works, 23% were "Adapted From", and 4% Sequel.

User stats
Any series without rating/votes were filtered out in the initial data cleaning. The mean bayesian rating was 6.4, with a median of 6.3 and a tiny right skew. Rating behaviour seemed similar across queer and non-queer series.

The average number of votes received was 31, the median number was 5. *Votes* had a strong right skew, most series that had an extraordinary (>500) number of votes were non-queer.

The mean Rank This Year was 6025/9999. Most of the series to break Top 500 were BL or not queer. The distribution had a strong left skew across all groups.

The mean total number of lists a series was on was 260 (vs. median 81), a heavily right skewed variable. Most of the series with an extraordinary total # (>5000) of lists were non-queer. Similar trends (heavy right skew, extraordinary counts mostly non-queer) followed for all list types. Respectively, they had means/medians of: Reading (70/11), Wish (46/24), Complete (81/14), Unfinished (5/1), Custom (46/24). All list counts were strongly positively correlated with each other ($r > \sim 0.8$) and with vote counts ($r > \sim 0.75$). All list counts also had moderate positive correlation with category and genre count ($r > \sim 0.3$)

Translation info

All series were last updated sometime between Dec 2006 and April 2025. Around 7% of series had never been updated, a greater proportion of which were queer. Most updates were recent, in a left-skewed distribution.

Only 36% of works had a latest chapter update. The mean latest update (chapter number that was translated) was chapter 7. The distribution had strong right skew across all groups, most series that had an extraordinary (>100) latest updated chapter and thus series length were non queer.

Most sampled series were complete (78%), followed by unknown (15%) and ongoing (6%). A higher proportion of GL series had an unknown status (28%).

16% of series were licensed, a lower proportion of GL series were licensed (9%). License status was moderately correlated with description length (0.47) and category count (0.33)

57% of series had a full scanlation. BL and GL series had a higher proportion of complete scanlations: BL (64%), GL (70%).

60% of series had a recorded original (OG) publisher name. Publisher names outside the top 20 most common for original language (og) were lumped into “Other”. The most common og publishers for non-queer series included Kodansha (8%), Shueisha (8%), and Shogakukan (8%). For BL series, common og publishers included Bomtoon (5%), Gentoosha (3%), and Houbunsha (3%). For GL series, common og publishers featured Ichijinsha (19%), Houbunsha (4%), and Kodansha (2%)

17% of series had a recorded English (EN) publisher name (only licensed translations have corresponding english publishers). Publisher names outside the top 20 most common for English (en) were lumped into “Other”. The most common EN publishers for non-queer series included Tapas (6%), Harlequin K.K. (5%), and Coolmic (5%). The most common EN publishers for BL series included Lezhin (15%), Renta! (10%), and Bilibili (8%). The most common EN publishers for GL series included Seven Seas (17%), Yen Press (8%), and Lezhin (7%).

Analysis 1 - Methodology

After examining all these variables, I became curious to see how they related to each other in further depth and how they might draw lines between the highlighted genres. *Principal Component Analysis* (PCA) is a useful tool to visualize the complexity of these relationships and find initial patterns as something more simple.

I included 15/16 numeric variables in the analysis (*Desc_length, year, bayesian_rating, votes, rank_this_year, lists_reading, lists_wish, lists_complete, lists_unfinished, lists_custom, last_updated_year, latest_ch, relation_n, cat_n, genre_n)* except *total_lists*, which is derived from the sum of all the other lists and not independent. PCA is intended for continuous variables, so the categorical and binary variables were not included.

Analysis 1 - Results

The first principal component explains around 39% of variability between series, the second principal component an additional 11%, diminishing with each further component. With only 2 components, series can be recognized with ~53% accuracy. With 8, this rises to ~87% accuracy. **(Figure 17)**

Most of the variability is explained by the first component traits with a high absolute value **(Figure 18)** – *votes, lists_reading, lists_wish, lists_complete, lists_unfinished, lists_custom.* It appears one of the most distinguishing features for a work is its sheer popularity in numbers. All of these metrics relate to the number of users who engage with a series, moving in the same (negative) direction.

Following those traits, the second component traits show that *description_length*, *year,* and *last_updated_year* also account for a significant amount of variability. These three metrics also move closely together. Year of release and year of last update moving together makes sense; Newer series will probably be edited more recently while old series stop receiving updates. *Description_length* has no immediately visible tie to the other two, but may suggest a trend upwards in summary length through the years.

Connections are also seen in **Figure 19**. All list variables tend to move together (eg. a series with high reading list count will likely also have high wish list count) and with votes. *Rank_this_year* moves nearly opposite from *latest_ch, last_updated_year, and year*, if a series has received no new updates (earlier year/latest update), it makes sense that it would see less activity this year and have its rank "higher" in number. *Cat_n* and *genre_n* move together as well, these two are user tagged and can be added in abundance by eager fans. *Bayesian_rating* also moves against *rank_this_year*, with a high rating contributing to a "low" ranking. *Relation_n* has a curious placement, somewhat related to *lists* but opposite of *year*. Series with many relations may be saved by readers of their related works, while newer series may not have the opportunity to receive spin-offs and related work.

When overlaying a plot of the first two components with the mutually exclusive queer variable, it was visible that there were some differences between the four groups. **(Figure 19)** The three queer groups clustered closer to the average on most scales, the non-queer group had more series with high lists/votes/bayesian_rating, and slightly more series along newness/tag counts. The BL group had a few series high on those two axes, but not as significantly as the non-queer group.

Analysis 2 - Methodology

From here, I wanted to explore which variables were most likely to be related to which queer status in a classification problem.

Out of the possible options: Linear Discriminant Analysis assumes continuous data, making it impossible to estimate probabilities with categorical predictors. Similarly, logistic regression only works with two groups.

Tree-based methods then provided the best tools for this goal. Random forests and decision trees allowed me to consider the huge swath of possible factors, narrowing down the key variables in classifying a work's queer status.

Due to concerns about imbalanced class sizes, series that were both GL and BL were filtered out.

To avoid overloading a tree with every variable which would risk outlier skew and variance, a preliminary random forest was run on all numerical and categorical factors to create a more stable model. The sample was bootstrapped 500 times with 500 trees based on a random subset of predictors to create a forest collection.

The top 50 columns that affected a mean decrease in accuracy and top 50 that affected a mean decrease in GINI were then selected to use in the decision tree model. Including variables with a high *MeanDecreaseAccuracy* increases classification accuracy, and including variables with high *MeanDecreaseGINI* decreases misclassification, both playing their parts as critical classification variables.

Decision trees, while less robust, allow human interpretation of each factor and I wanted to gain a better understanding of the factors that play into queer genres. To build an optimal tree with a balance of fit and complexity, I started with a low complexity parameter (cp) (0.0001) threshold and then found the cp the tree where the out-of-sample error was lowest, pruning the tree of excess for a clean analysis,

Analysis 2 - Results

The random forest achieved an OOB (out of bag) error rate of 0.1215. This shows the model has a strong ability to be applied beyond the test data. When comparing predictions against the real data, the model predicted Not Queer and BL series with only around ~6% error. This error was significantly higher for GL works at 38.8% error.

Some of the most important variables found in the random forest included: list variables (*lists_custom, lists_complete, lists_wish, lists_reading, lists_unfinished, total_lists),* some genre variables *(Josei, Shoujo, H, Shounen, Seinen, etc.),* series stats *(year, genre_n, desc_length, fully_scanlated, etc.),* some user stats (*votes, rank_this_year, bayesian_rating, etc.)* and some miscellaneous categories (*Ichijinsha: og publisher, Other: og publisher, Doujinshi: series type, Manga: series type, Adapted From: relation type)*. **(Figure 21)** For example, *lists_custom* had a mean decrease in accuracy of 60.7 and a mean decrease in Gini of 538.2. This predictor is important to keep, otherwise accuracy and node purity will both decrease significantly.

From the PCA analysis, we already know to keep an eye out for the *lists* and *votes* which seemed to distinguish non-queer series from the rest. As well, our data exploration showed that some categorical variables set apart groupings from one another. *(Ichijinsha showing up as original publisher for many GL works, Doujinshi highly prominent among BL)*

After pruning the tree at its optimal cp, the decision tree model was created **(Figure 23)**. While still complex, it is justified by its solid performance on out of sample data at a relatively low error rate of 0.269 **(Figure 22)**. The number of branches is likely also influenced by the wealth of categorical (encoded as binary) predictors.

From this decision tree, immediately visible are rough clusters for each of the three queer groups: The left cluster shows non-queer works are most likely, and a middle and right cluster tend to lead to BL or GL. BL and non-queer works appear fairly distinct, rarely on the same "branches", while GL is mixed with both. This could explain GL's higher identification error.

The most prominent division is whether a series is of type "Doujinshi" or not, on average 70% of Doujinshi are BL and only 35% of non-Doujinshi are BL. Doujinshi, indie works not published by mainstream companies, may have more freedom to be openly queer.

Within Doujinshi, on the rightmost branch, H-genre series are more likely to be GL or non queer. H Doujinshi have only a 4% probability of being BL, compared to 80% for non-H Doujinshi. Specifically unpopular non-H Doujinshi with few lists have a higher chance of being GL.

On the left branch, most works have the highest likelihood of being not queer. GL series have a small likelihood of appearing when a work is H-genre with few other tagged genres. If a series is *Shounen, Shoujo, Seinen* or *Josei*, it is also more likely to be a non-queer work. These tags are primary manga demographics: Young boys, young girls, older boys, older girls. This suggests that queer works tend to have none of these tags and fall outside these traditional age and gender categories. Three large non-queer terminal nodes appear in this branch, bucketing 18% of the total data at high 88+% probabilities. Two of these fall under branches not tagged LGBT scenes, and the third under both Shoujo and Romance *(well known for heterosexual princes and princesses)*. All three are determined from <6 internal nodes, suggesting this kind of non-queer series may be more easily identifiable.

In the middle cluster with none of those primary demographics, most works are likely to be BL. Two smatterings of high GL probability appear, both have short descriptions and fewer lists. Also, notably, series from the publisher *Ichijinsha* (publisher of Yuri magazine *Yuri Hime*) split off to create a branch high in GL probability.

Lists and *votes* prominently appeared during PCA and still have some influence throughout the tree, with different thresholds leaning the data towards one group or another. In other models, multicollinearity may weaken their effects, but decision trees are able to naturally handle collinearity. They tend to appear below categorical variables in the decision tree. Unaccounted for in PCA, categorical variables may interact with numeric ones in interesting ways.

Additionally, certain specific elements appear to sway groups in certain manners. For example, *All_girls_school* seems to increase the probability a series is GL, *LGBT_scenes* towards BL or GL (although one interesting node on the left with *LGBT_scenes* still ends up most likely non queer), *female_protagonist* towards GL. Having female characters makes logical sense for a GL story. Older *Year*

is less likely to be BL, agreeing with initial data comparison. Also, *Romance* tends towards non-queerness. This distinction between (traditional) romance and queer romance is interesting.

Conclusions

Queerness is complicated and difficult to nail down. This brief study has shown that queer media is also hard to define, with PCA demonstrating that a range of factors pull genres in different directions, and tree methods that many variables are needed to accurately determine a series' queer status. Still, the error rates are low and suggest that the models make a solid sketch.

GL and BL series appear more similar to each other than non-queer series. In the PCA plot, GL and BL points are close to each other while non-queer series have higher *list* and *vote* stats. In the decision tree, BL and GL remain closely grouped while non-queer series branch off based on *Doujinshi* and other genre statuses. Points of GL-BL differentiation rise up usually from narrower margins.

Broadly, queer series still struggle for the same widespread recognition as non-queer series. The main points of differentiation came from user popularity (*lists, votes*), establishment (*year, latest_ch*), mainstream publishing (*Doujinshi*), and a lack of integration with traditional young reader demographics (*Shoujo, Shounen, Josei, Seinen)*.

While queer media has come a long way in acceptance, efforts must continue to be made to encourage their development and promote their success.

This project faced sample size limitations for GL series, both alone and in tandem with BL. While it would have been interesting to study how BL and GL interacted with each other, the sample size was far too unbalanced to analyze in depth. Also, GL had the highest error rate in the random forest, which could hopefully be improved with greater input. It is telling about the state of queer media that a magnitude fewer GL works exists in comparison to even BL.

It could be interesting to group GL and BL together in another study and compare queer vs non-queer series overall, or further break GL and BL down to their two components. Regardless, this exploratory analysis provides a launching point for further research into queer media and English audiences.

Appendix

Data Attribute	Brief Description
Genres	A list of genre tags for a series, out of 34 possible, multiple select possible. Broken out into dummy columns.
genre_n	The number of genres a series has
queer	If a series is tagged with any of “Shoujo Ai”, “Shounen Ai”, “Yaoi”, “Yuri”, boolean. Derived from genre.
girlslove	If a series is tagged with any of “Shoujo Ai”, “Yuri”, boolean. Derived from genre.
boyslove	If a series is tagged with any of “Shounen Ai”, “Yaoi”, boolean. Derived from genre.
mutual_exclu_queer	“Not queer”, “Boys Love” (only), “Girls Love” (only), “Both”
mutual_exclu_gl	“Not GL”, “Shoujo Ai” (only), “Yuri” (only), “Both”
mutual_exclu_bl	“Not BL”, “Shounen Ai” (only), “Yaoi” (only), “Both”
Description	The series description. “” if none.
Desc_length	The number of characters in the description
Type	The type of a series, based on content type (eg. Doujinshi (self-publishedwork /derivative work) vs published), and country of origin (eg. Manga (JP) vs Manhua (CN))
Year	The release year of a series
Bayesian_rating	The weighted rating of a series from user votes, ranging 1-10.
Votes	The total number of user votes on a category
Rank_this_year	The rank of a series this year based on its bayesian rating
Lists_reading	The number of users who have saved a series under status “Reading”
Lists_wish	The number of users who have saved a series under status “Wishlist”
Lists_complete	The number of users who have saved a series under status “Complete”
Lists_unfinished	The number of users who have saved a series under status “Unfinished”
Lists_custom	The number of users who have saved a series under status “Custom”
Total_lists	Derived from the previous five attributes, a total number of saved series
Last_updated	When a series was last updated by users/admins after title creation, UNIX time. 0 if never updated. Last_updated_year a numeric derived attribute.
Latest_ch	The chapter number of the most recent series translation. 0 if never translated.
Status	The series status in its original language. Broken out into 5 categories (Complete, Ongoing, Cancelled, Hiatus, Unknown)
Licensed	Whether a series received a licensed translation, boolean.
Fully_scanlated	Whether a series received a full fan translation, boolean

Categories	A list of series tags more specific than genres, multiple select possible. Broken out into dummy columns for those with >=250 counts.
cat_n	The number of categories a series has
relation_type	A list of the types of any related works, out of 11 possible, multiple select possible. Broken out into dummy columns.
relation_n	The number of relations a series has
pub_name_og	The name of the primary series publisher in its original language. Lumped outside top 20 publishers
pub_name_en	The name of the primary English publisher of a series. Lumped outside top 20 publishers

Figure 0, explanation of variables

```
> contingency_table
          GL non-GL
BL        58  12292
non-BL  4994   9333
> contingency_table2
          non-Shounen Ai Shounen Ai
non-Yaoi           14327       5675
Yaoi                6517        158
> contingency_table3
          non-Shoujo Ai Shoujo Ai
non-Yuri          21625      2114
Yuri               2743       195
```

```
> round(contingency_table / length(MU_contingent$queer) * 100, 2)
          GL non-GL
BL      0.22  46.08
non-BL 18.72  34.99
> round(contingency_table2 / length(MU_contingent$queer) * 100, 2)
          non-Shounen Ai Shounen Ai
non-Yaoi          53.71      21.27
Yaoi              24.43       0.59
> round(contingency_table3 / length(MU_contingent$queer) * 100, 2)
          non-Shoujo Ai Shoujo Ai
non-Yuri          81.06      7.92
Yuri              10.28      0.73
```

Figures 1 and 2, contingency tables of queer genres by count (1) and proportion (2)

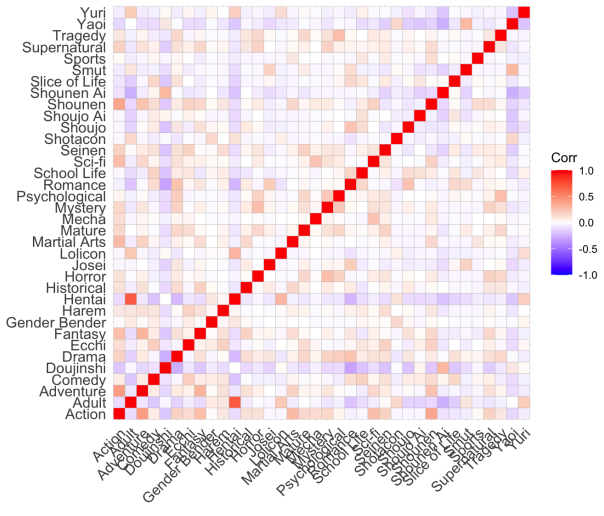


Figure 3, genre correlation plot

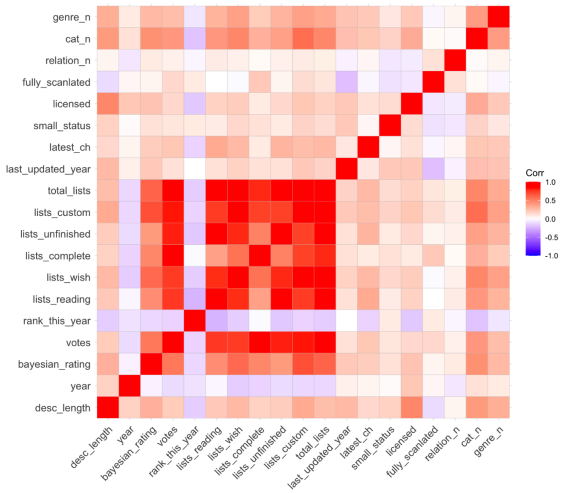
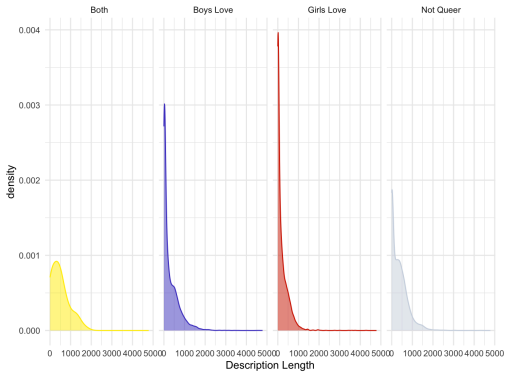


Figure 4, correlation matrix between numerical variables



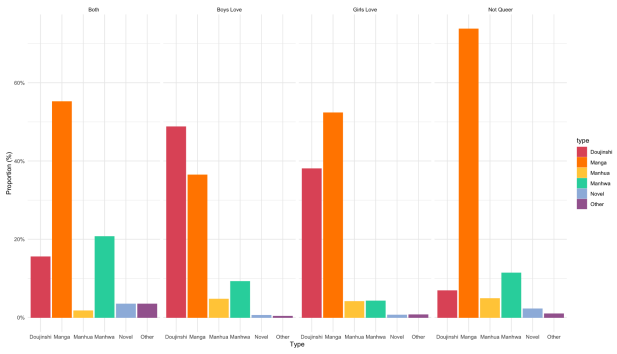
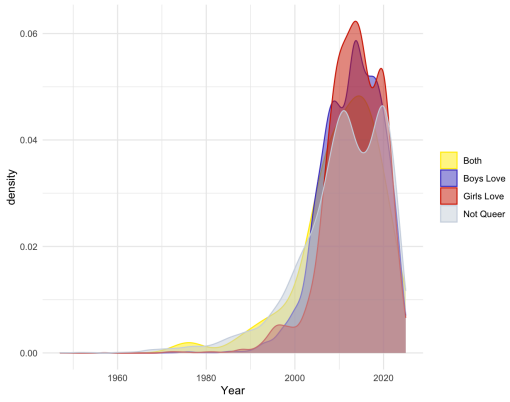


Figure 5, density plot of description length by queer group after filtering out 7 extreme outliers for readability

Figure 6, bar plot of series type by queer group



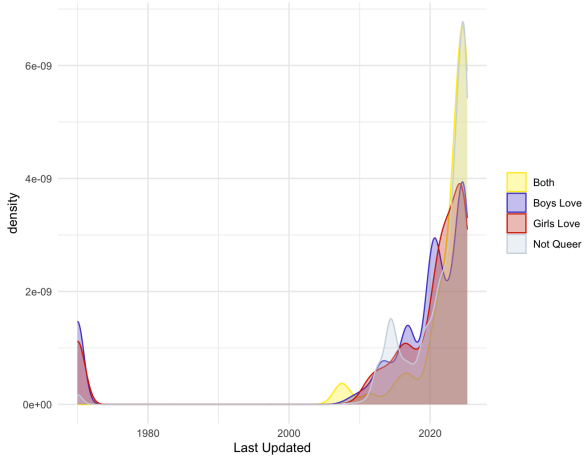
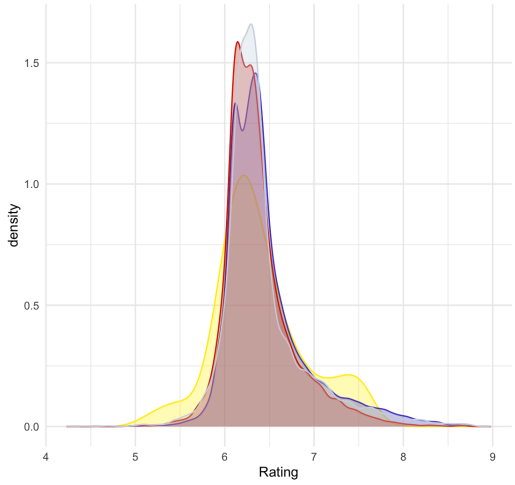


Figure 7, density plot of release year by queer group

Figure 8, density plot of last series update by queer group



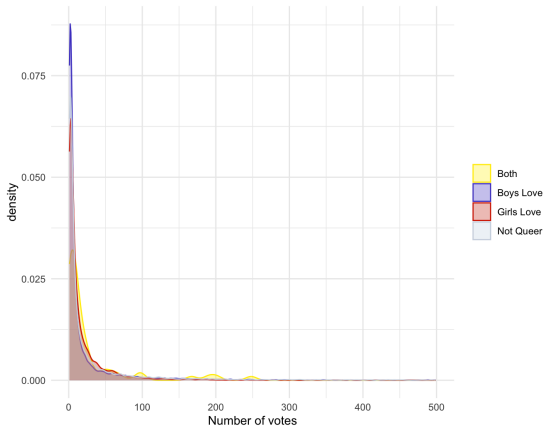


Figure 9, density plot of rating by queer group

Figure 10, density plot of number by queer group after filtering out extreme outliers >500 votes for readability

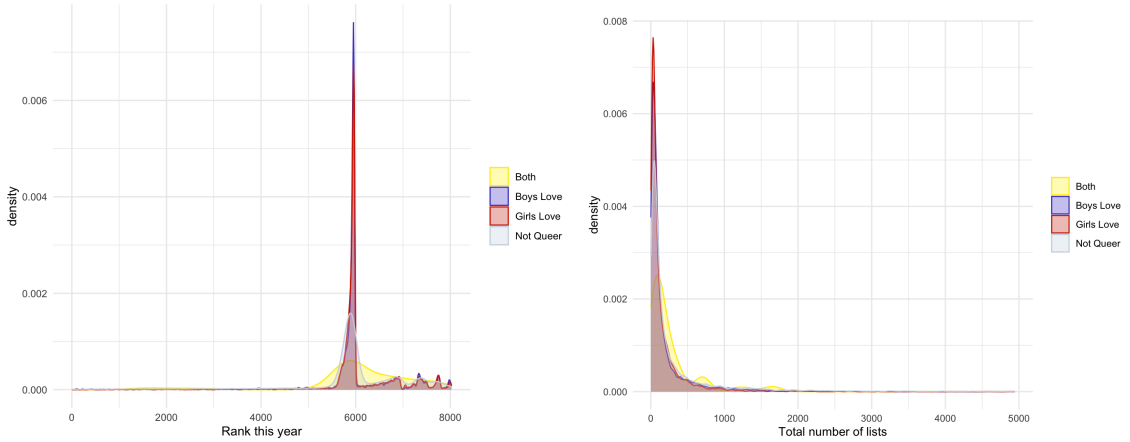


Figure 11, density plot of rank this year by queer group

Figure 12, density plot of total list count by queer group after filtering out extreme outliers >5000 lists for readability

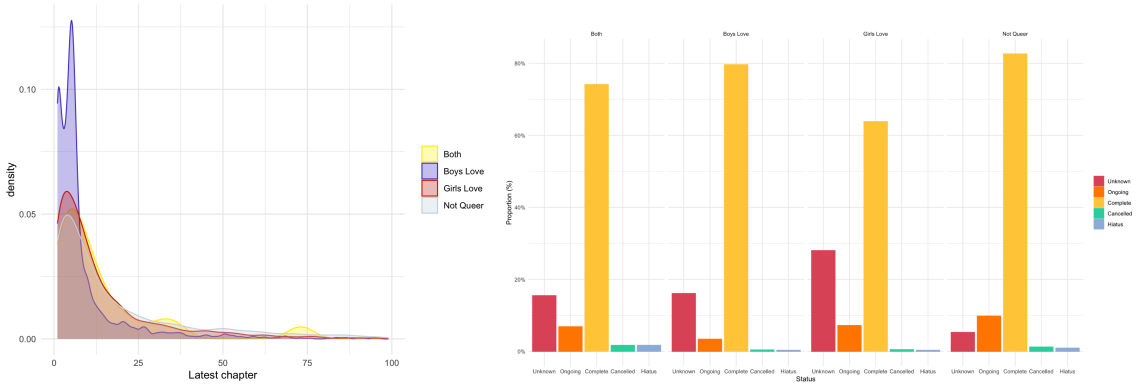


Figure 13, density plot of latest chapter by queer group after filtering out extreme outliers for readibility

Figure 14, bar plot of series status by queer group

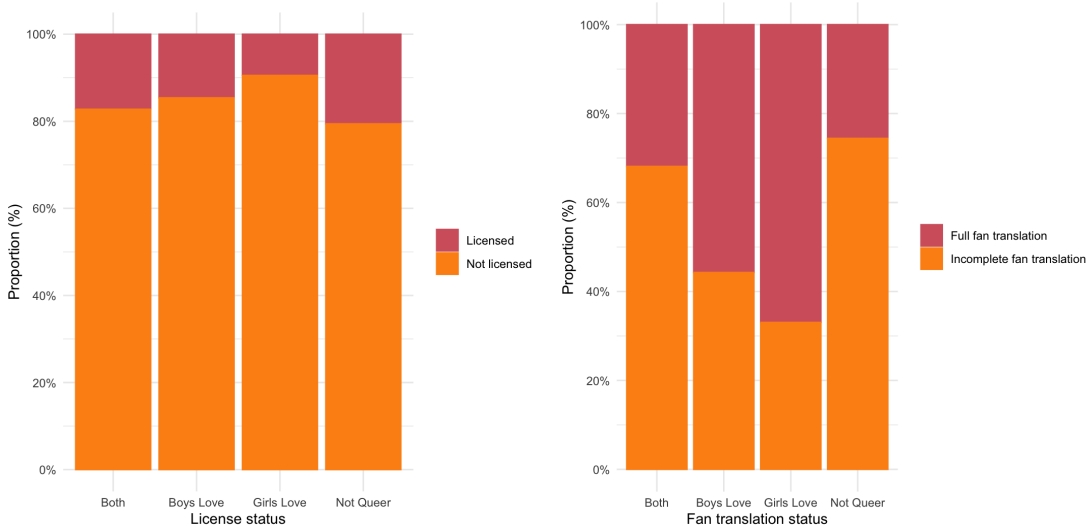
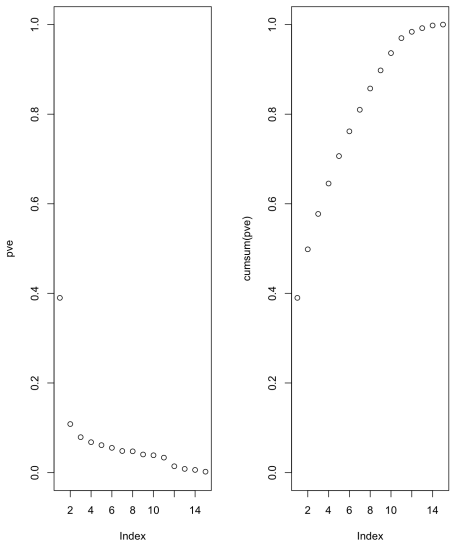


Figure 15, bar plot of licensed status by queer group

Figure 16, bar plot of fan translation status by queer group




```
> pca
Standard deviations (1, .., p=15):
 [1] 2.4187204 1.2747074 1.0887921 1.0088827 0.9584107 0.9100646 0.8500844
 [8] 0.8437490 0.7784847 0.7614461 0.7078088 0.4597023 0.3510062 0.2984351
[15] 0.1727681

Rotation (n x k) = (15 x 15):
                         PC1         PC2         PC3          PC4         PC5
desc_length       -0.16607829 -0.46380103  0.14353651  0.153141601  0.17878644
year               0.03263783 -0.46287180 -0.16262893 -0.242958240  0.36992822
bayesian_rating   -0.28588544 -0.06285673  0.19024539  0.125334559  0.13601678
votes             -0.36557971  0.22417681  0.03913038 -0.092598035  0.13961306
rank_this_year     0.09307122  0.18367913  0.56383563  0.018318155 -0.40114422
lists_reading     -0.34610016  0.08840753 -0.28548284 -0.090157881 -0.10258164
lists_wish        -0.36822275  0.10229427 -0.02812284 -0.021006886 -0.07850989
lists_complete    -0.29094685  0.23624727  0.31212178 -0.039537764  0.25449438
lists_unfinished  -0.34619165  0.17338305 -0.20223906 -0.107646616 -0.09575694
lists_custom      -0.38641644  0.06549025  0.10130318  0.011676334  0.05767640
last_updated_year -0.10740469 -0.36137452  0.39084939  0.018890291 -0.33785687
latest_ch         -0.14712914 -0.16326880 -0.41973391 -0.032913133 -0.59121415
relation_n        -0.03208538  0.06538915 -0.17039929  0.926640401  0.07066807
cat_n             -0.25817969 -0.31777320  0.06167315  0.009847467  0.09606606
genre_n           -0.19300991 -0.33428555  0.06826693  0.104132708 -0.24581544
```

Figure 17, Percentage-of-variance-explained plots

Figure 18, Principal components

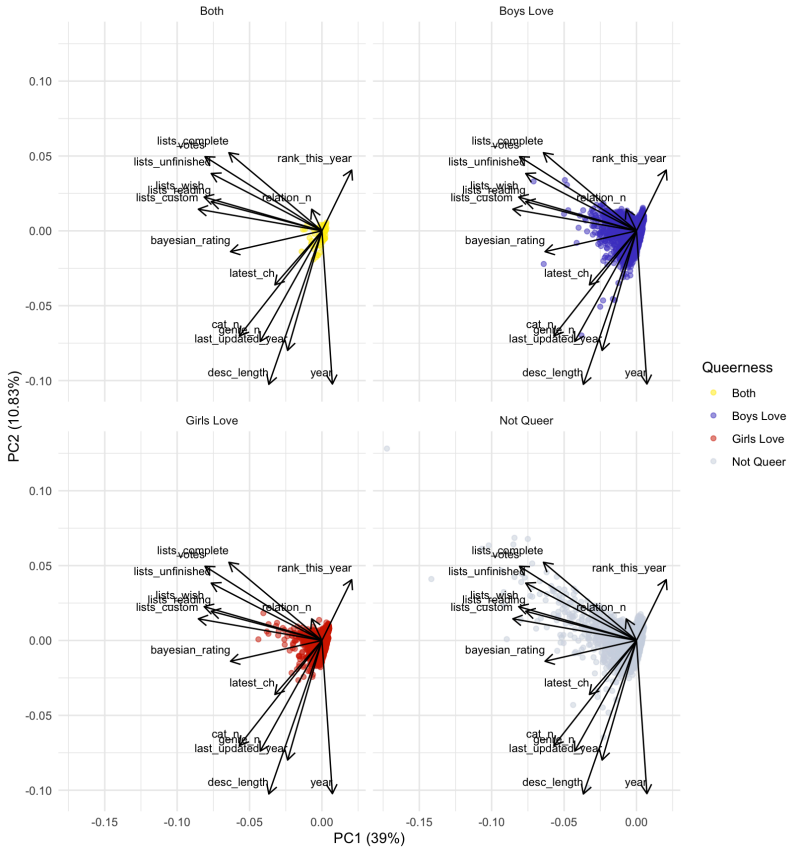


Figure 19, Principal Component analysis plots by queer group

```
Call:
 randomForest(formula = fourthtree_formula, data = MU_df_biggened_5,      ntree =
500, importance = TRUE, do.trace = TRUE, na.action = na.omit)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 15

        OOB estimate of  error rate: 12.15%
Confusion matrix:
           Not Queer Boys Love Girls Love class.error
Not Queer       8764       303        266  0.06096646
Boys Love        451     11570        271  0.05873739
Girls Love       879      1063       3052  0.38886664
```

Figure 20, Random Forest summary results

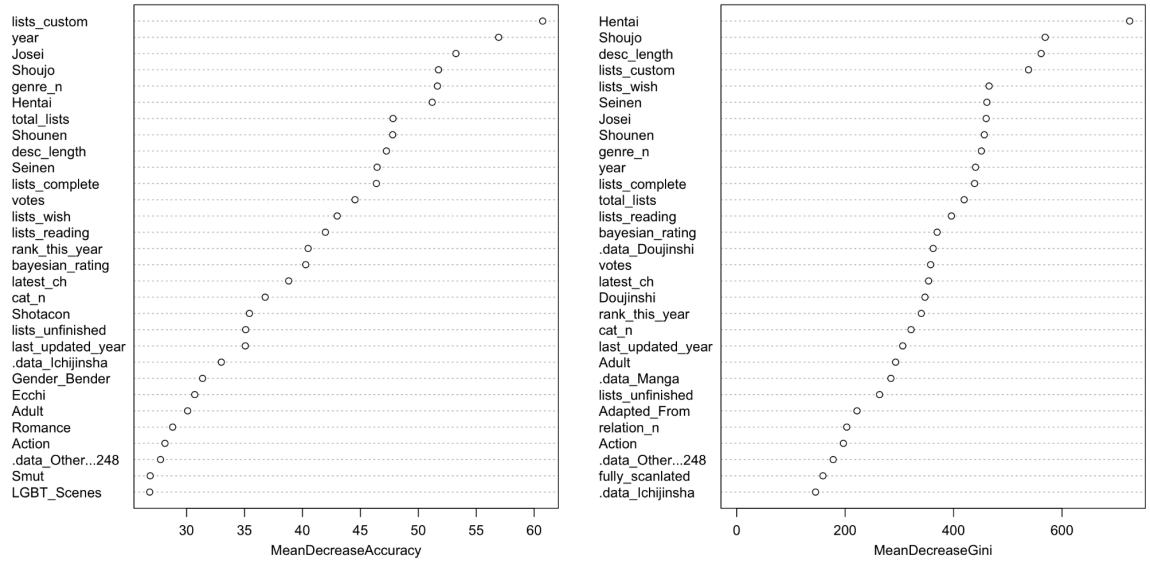


Figure 21, Random Forest Variable Importance plot

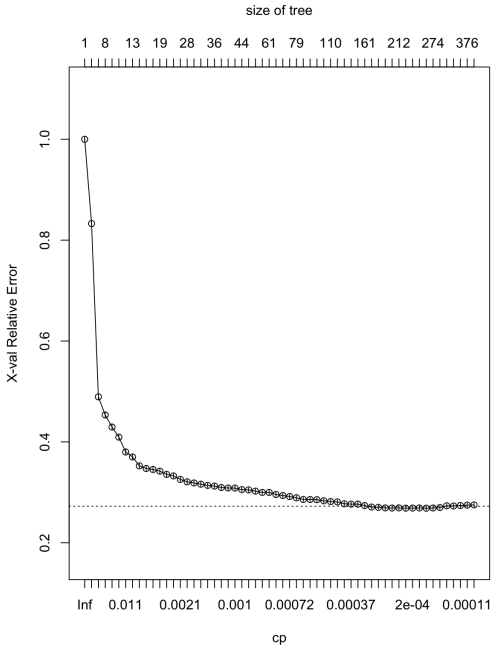


Figure 22, Graph of complexity parameter vs out of sample error

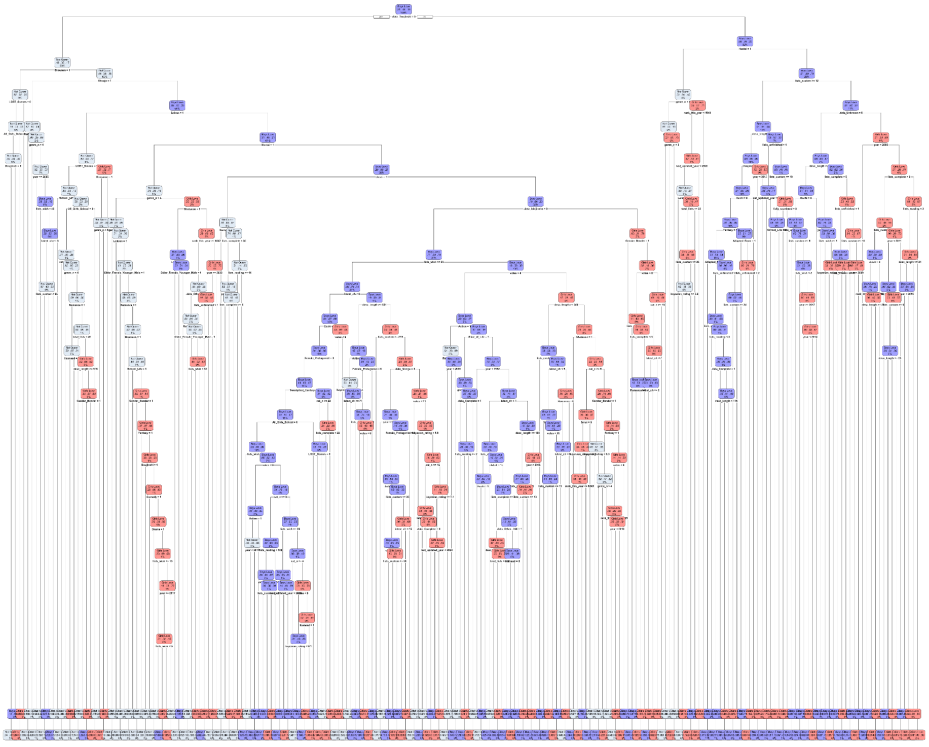


Figure 23, Decision tree. Higher resolution

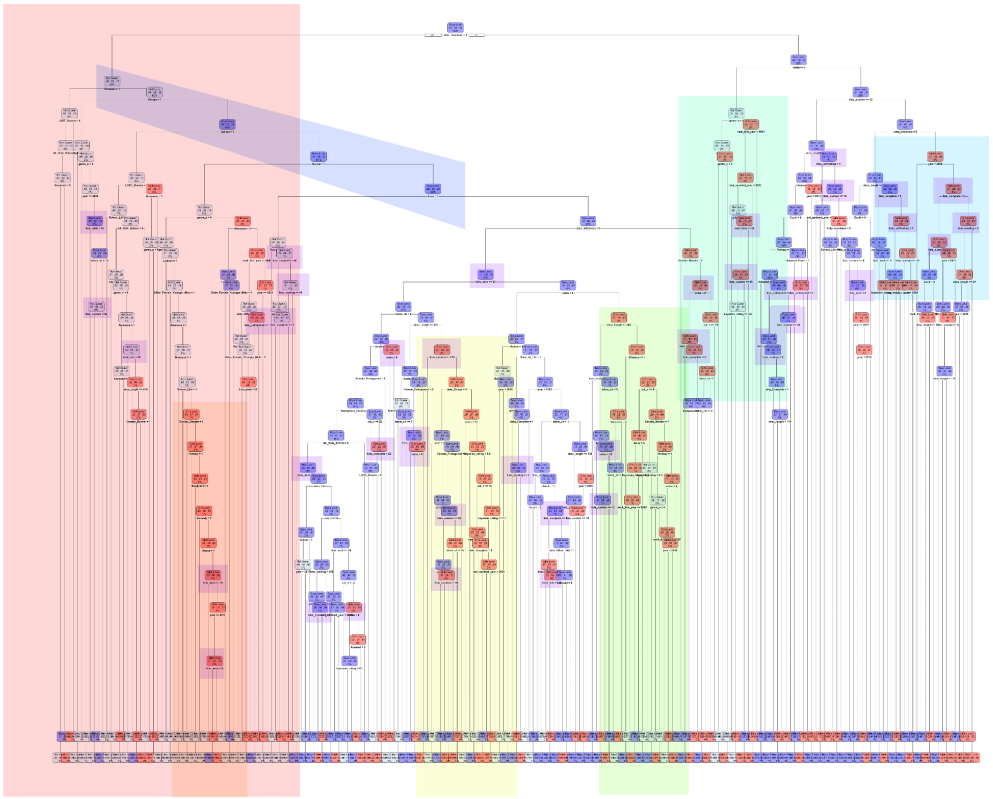


Figure 24, highlighted decision tree. Higher resolution