STATS 767 - Project Writeup

Alice Hankin

2022-05-22

Introduction

The Archive of Our Own (or AO3) is an open-source web archive which describes itself as being a "noncommercial and nonprofit central hosting place for fanworks". As of April 2022, the Archive contains over 9 million works composed mostly of written fanfiction, although fanart, comics and podcasts are also included in this total.

For this project, I will be looking at a small subset of fanworks that can be found on the Archive. I am interested in analysing the trends in the various statistics that AO3 records about each fanwork. In particular, I am interested using principal component analysis descriptively, before using linear discriminant analysis to create a model to categorize the works. I wish to know if it is possible to classify English-language works on Archive Of Our Own that were completed in the first week of 2022 by their rating, using only the numeric information in the metadata.

The Data

In order to download all the necessary metadata, I made use of a web scraper which runs in Python. The CC-BY-NC licenced code can be found on GitHub. I restricted the works to being complete (this is so that I could analyze the number of chapters in the finished work), and English-speaking (so that I could understand the tags used, and so I wouldn't have to deal with the same work translated to multiple languages). It is worth noting here that I have not included 'restricted' works – these are accessible only to those with accounts on the website – due to the limitations of the web scraper.

I ended up with a 14.2 megabyte .csv file containing 28,152 observations of 19 variables. The following table shows the column names in the raw data as well as two examples of observations (fanworks).

work_id	title	author	rating	category	fandom
36133360	Coming Home	['alexcat']	General Audiences	M/M	Sherlock Holmes
696420	Gift	['elstaplador']	General Audiences	F/M	Le Fantôme de l'Opéra

relationship	character	additional tags
Sherlock Holmes/John Watson	Sherlock Holmes, John Watson, Mrs. Hudson	Christmas, Fluff
N/A	Erik (Phantom of the Opera), Christine Daaé	Unrequited Love, Drabble

language	published	status	status date	words	chapters	comments	kudos	bookmarks	hits
English	2022-01-01	Completed	2022-01-01	500	1/1	3	19	2	208
English	2022-01-01	Completed	2022-01-01	100	1/1	N/A	14	1	228

Since several of these columns are unneeded for the analysis, I will remove them. For example, since I have restricted my dataset to only include works in English, the "language" column is not needed. I will also tidy the data. This process involves making the columns numeric or factor variables, dealing with missing data (i.e. all N/A values in the kudos column should really be zeros), and mutating some of the non-numeric columns to be additional numeric variables. The R code for this cleaning step can be found in Appendix 1.

I have printed a random sample from the cleaned dataset. You can see what each column means in Appendix 2.

```
##                          rating      category words chapters comments kudos
## 36173203 Teen And Up Audiences F/M, Gen, M/M  2532        1        0     7
## 36147160     General Audiences           Gen  7699        4       14    47
## 36126127 Teen And Up Audiences           F/M  7485        1       29   160
## 36213145     General Audiences           Gen  1399        1        0    24
```

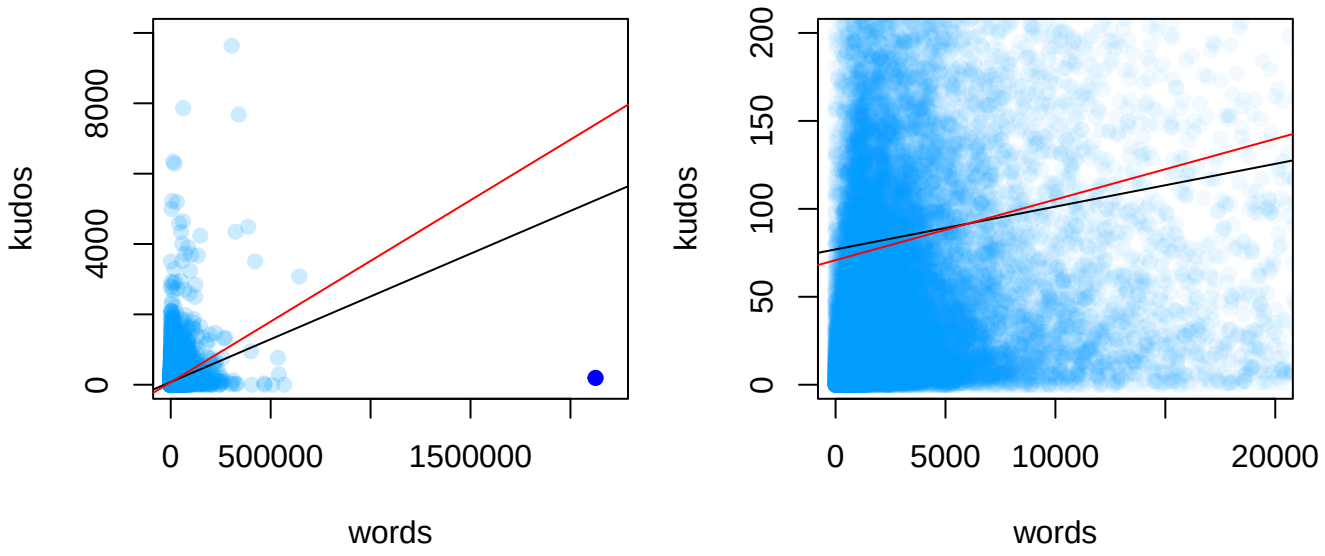
```
## 36265846              Explicit           F/M  2320        1        1    58
##          bookmarks hits num_tags daystocomplete fluff_angst
## 36173203         1   71        0              0     neither
## 36147160         4  410       17              0       fluff
## 36126127        20  794        7              0     neither
## 36213145         0  173        0              0     neither
## 36265846         6 1521        5              0     neither
```

Code for producing this table found in Appendix 3

Exploring the data (1)

Before doing any analysis, it's important to explore the data to get an understanding of its patterns and characteristics. Primarily, I am interested in seeing if we need to transform the data, and if there are any outliers.

Outliers



Code for this plot found in Appendix 4

Here, I have plotted words against kudos. These plots are typical of any one of the panels in the pairs plot for the dataset; any of the numerical variables against any of the others produces a very similar looking result. Note that both of the panels here show the same data, zoomed in by different amounts.

The work with the most words (coloured dark blue) is a possible outlier - this work has 2 million words (which is three times as many as the work with the next smallest length) and 195 kudos. Here, I have applied two linear regression models to the data; one including this data point (in black) and one removing it (in red). They do look very different! I will discuss this point further in the logged data section.

The important thing to notice from these plots is that can see that there is an extremely high density of points in the lower left corner. This is works with fewer than 40 kudos and fewer than 5,000 words. However, there are also many works that have more kudos/words by several orders of magnitude. In fact, there are several works with over half a million words. This is exactly the type of data that is best suited to a log transformation.

Due to this extreme skew that we are seeing, with data values ranging over multiple orders of magnitude, it makes it very hard to produce box-and-whisker plots or histograms that are understandable. So instead, in order to confirm that this log transformation is suitable, I will have a quick look at the quantiles of the variables.

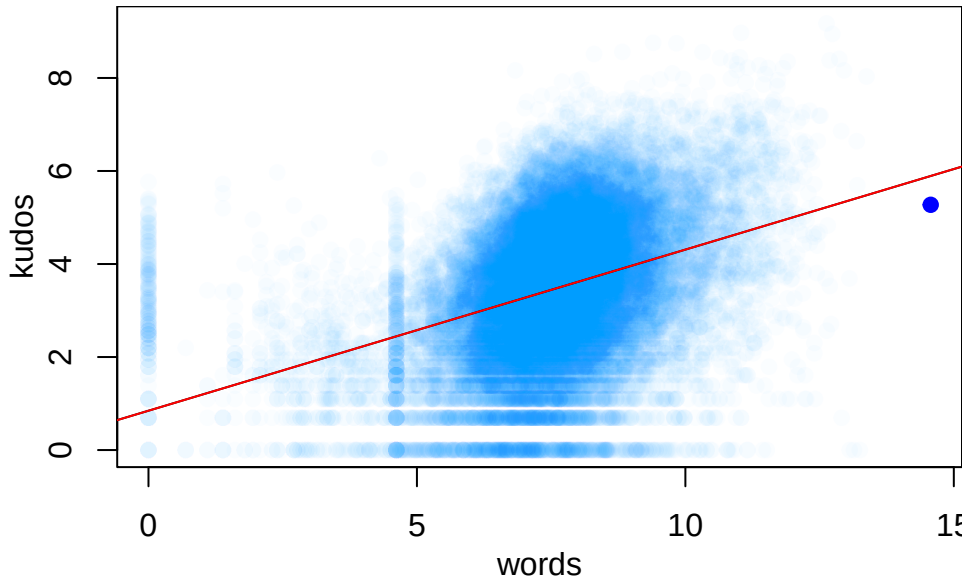
Scaling

```
##           words chapters comments kudos bookmarks      hits num_tags
## 0%         0.00        1        0     0         0      0.00        0
## 25%      969.00        1        0    11         0    131.75        4
## 50%     2026.50        1        3    33         2    386.00        7
## 75%     4497.25        1        9    91         8   1044.00       12
## 100% 2125479.00      366    10265  9639      2290 291493.00       74
##      daystocomplete
## 0%                0
## 25%               0
## 50%               0
## 75%               0
## 100%           3313
```

Code for this table found in Appendix 5

There seems to be a very long right tail for every single one of these variables! We can see from the `comments` and `daystocomplete` variables just how extreme this is in some cases! This table gives good reason for logging the data.

Logged data



Code for this plot found in Appendix 6

This is the same scatterplot we saw previously in the outliers sections, however here, I have logged the data. Note that I have had to add one to each data value before taking the logarithm, since many entries are allowed to be zero. The possible outlier we saw before is marked in dark blue; on a log scale it doesn't seem too far from the trend! In fact, I have again plotted the linear regression lines again on this plot, this time using the logged data. The lines essentially overlap. This gives good reason to keep this point in, if I am to use this scaled data - it seems our scaling, as expected, has increased the stability of a linear model (and hopefully our other models too).

One interesting feature of this plot is the increased density of works at `words`$= \log(100+1) \approx 4.62$ and `words`$= \log(0+1) = 0$. These reveal some insight into the way that AO3 is used.

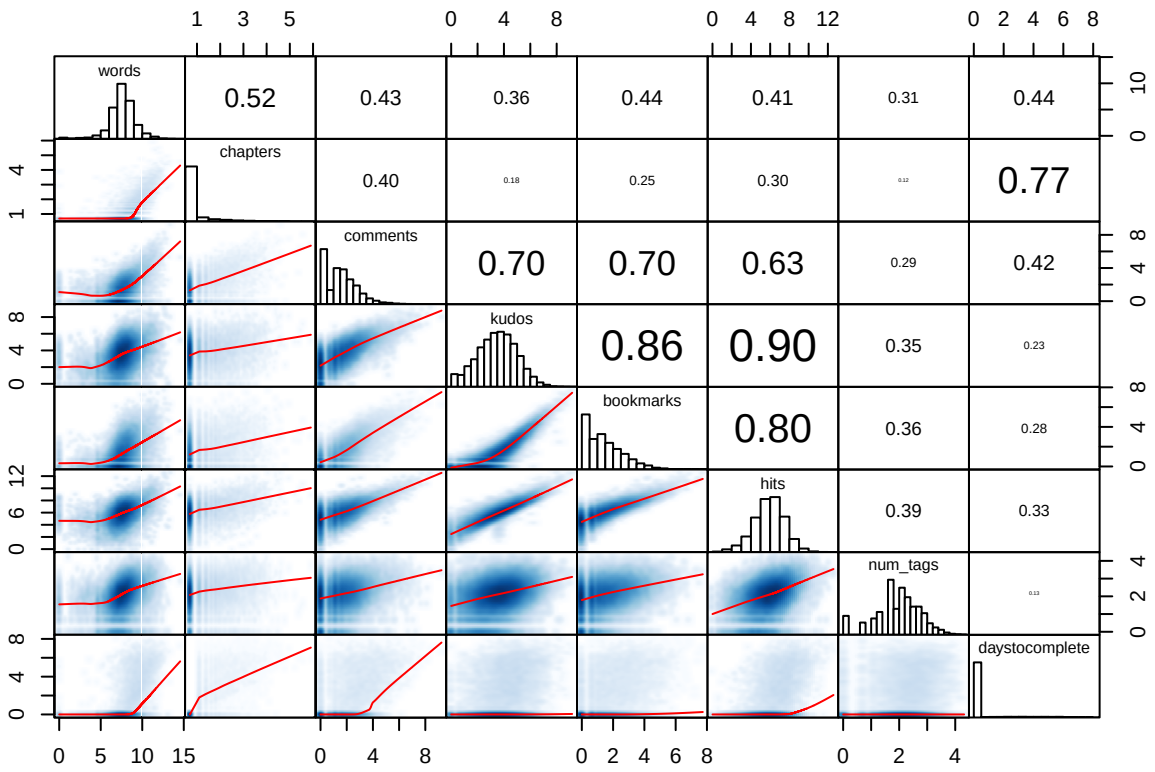
- All those fan works which are not written are classified as having zero words. These might be podfics (i.e. audio-based fanworks), or fanart. These are not too uncommon to see on AO3, but the vast majority of works remain written. Some of these non-written works will have a few words/paragraphs description which might explain this "smudge" of points with low `words` values.

- There is a particular type of fiction called a "drabble", which Wikipedia defines to be "a short work of fiction of precisely one hundred words in length". These are particularly popular on AO3 - it's really interesting that this causes a visible increase in density.

I would suggest that it is these two unexpectedly high-density regions that mean that the trend line looks far from going through the major axis of the ellipse shape that we can see. This means that our data isn't quite multivariate normal, but it doesn't look too far off.

Another reason for this not-quite-normal characterization is that, due to the fact that we have discrete data, some banding is visible on the plot. This is more visible on the `kudos` axis since there is more likely to be works with few (say, 1-20) kudos compared to few words. Of course, our normal assumption requires that the data is continuous, so the data misses the mark slightly here too.

Pairs plot



Code for this plot found in Appendix 7

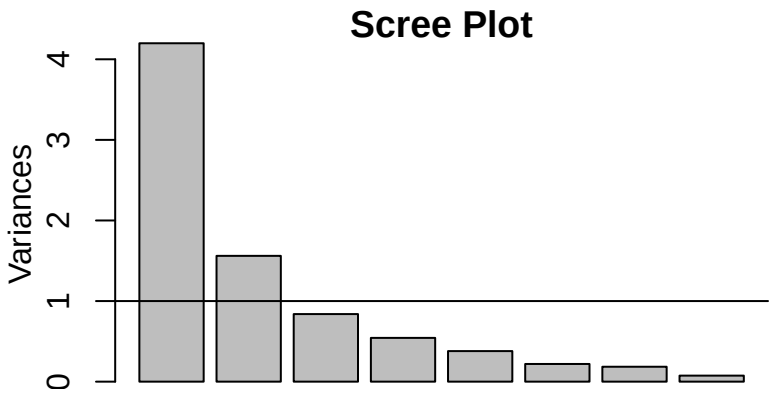
In the bottom/left, we have the pairs plot for our logged data (note - I also centred and scaled the data by subtracting the mean and dividing by the standard deviation). In order to be viewable (and render in a short amount of time), I have used the `smoothScatter` function, which makes the dense areas darker, and the less dense areas lighter coloured. In the top/right panels, we can see the correlations between each pair of variables. On the diagonal, we have the histograms of each individual variable.

- Correlations:
  - We can see that there are very high correlations between `hits`, `kudos`, `bookmarks`, and `comments`. This makes sense since they are all measures of how "enjoyed" or perhaps "good" a work is.
  - `chapters` and `daystocomplete` have a very high correlation - this is likely because the many works that have one chapter also must have a one in the `daystocomplete` column. These also have a relatively high correlation with `words` which makes sense intuitively - these are all measures of "length".

  - Each variable is positively correlated with every other variable, but `num_tags` has the lowest correlations. This is expected since it seems to measure something different than "enjoyable-ness" or "length" (which, themselves, are positively correlated, just not strongly).
- Scatterplots:
  - The scatterplots between the `kudos`/`bookmarks`/`hits`/`comments` variables are all mostly linear. It seems that an increase in one of these values (on the log scale) corresponds to a proportional increase in all the others.
  - There seems to be a small number of points below the $x = y$ line in the `kudos`/`hits` plot (see Appendix 8 for a zoomed in version of this panel). This indicates that these works have more kudos than hits, which shouldn't happen! Checking these works on AO3 tells us that the data has been scraped correctly - perhaps this is indicative of some error with the servers? I can't think of a reason why this would happen.
  - The scatterplots involving `words` flatten on the left-hand side; this is due to those works which are not mainly written, as we've seen before
  - The `daystocomplete` plots are pale coloured; this is because of the sheer density of works with 0 for this variable. This also influences these red trendlines, making them look flat. A similar thing is occurring for `chapters`.
- Histograms: I look at the normality of each variable in depth later on in this report.
  - While `words, kudos, hits, num_tags` look very normal, `comments` and `bookmarks` are still slightly skewed even after the log transformation.
  - `daystocomplete` and `chapters` are still extremely skewed.

PCA (1)

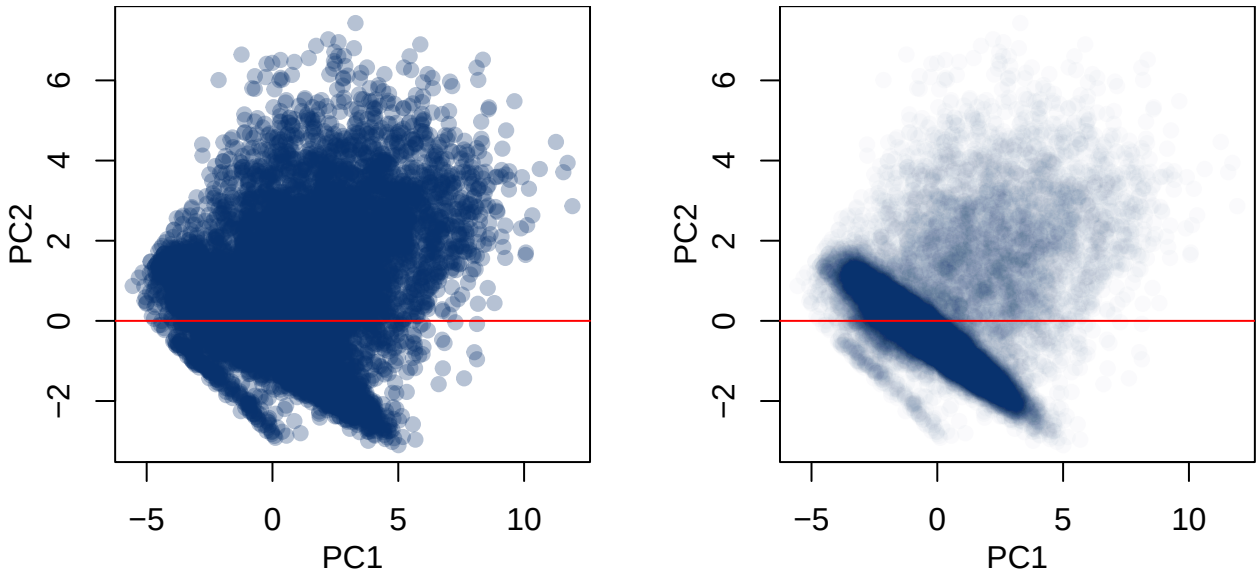
Principal components analysis (PCA) is a dimension-reduction technique. We are interested in seeing (essentially) how many dimensions the data takes up, and hopefully getting an interpretation for what each dimension means. PCA also allows us to explore associations between variables. Initially, we want to draw a scree plot which helps us decide how many components to use based on the amount of variability in the data they explain.



Code for this plot found in Appendix 9

Going by the "eigenvalue greater than 1" rule, it seems reasonable to select two principal components. I can see that there is an "elbow" between the second and third components, so this seems like a good idea. These two components make up 72% of the variability in the data (see Appendix 10).

Let's have a look at plotting the PC1 and PC2 scores for each observation against each other.



Code for this plot found in Appendix 11

Both these images show the same plot, but with transparency turned up to different values. This is very strange-looking! Recall that principal components are defined in such a way that they must be orthogonal to one another. From the left-hand plot, we would think that there is no relationship between the principal components, especially when we also plot the trend line in red. However, the second plot tells a different story. It seems like the otherwise negative trendline is dragged flat by all these values with high PC1 and high PC2 scores. I expect that this is because of all these points in the tails of our distributions. We saw earlier that although our transformations have normalised a lot of the components, the `daystocomplete` and the `chapters` variables were still hugely skewed.

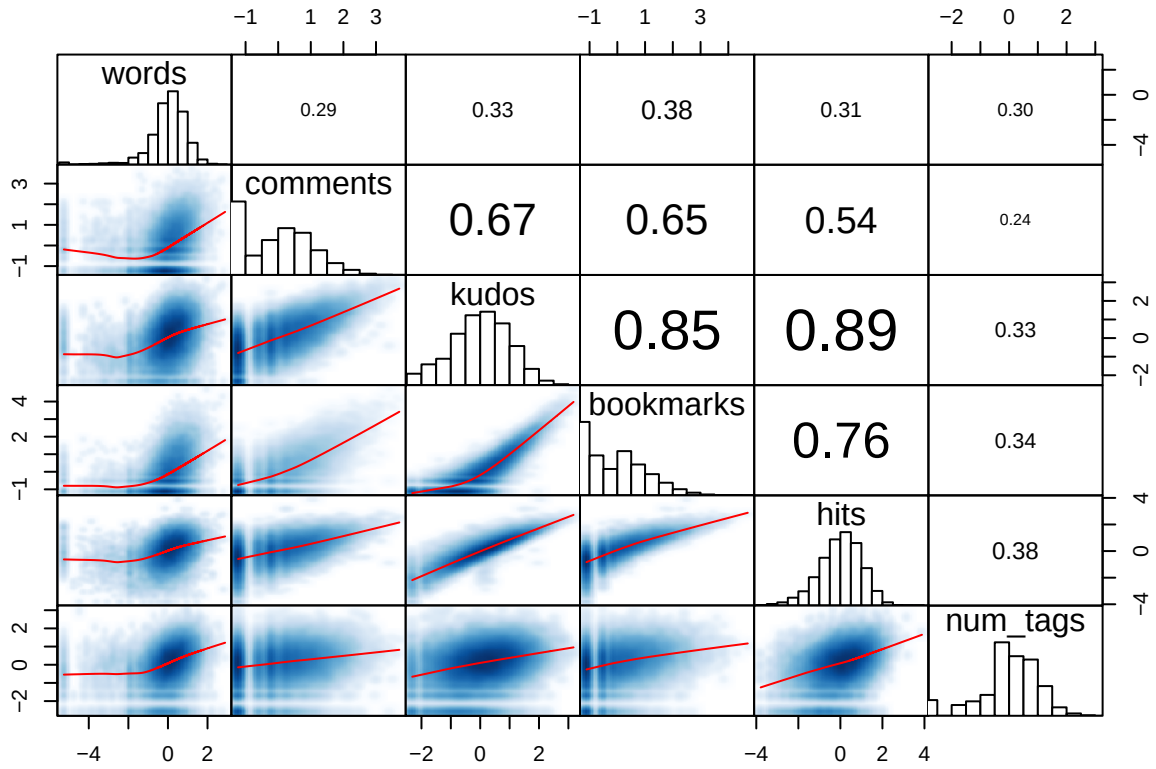
Now, in order to get around this, one might decide to just remove these two variables. However, it is important to think about where these issues are coming from in the first place.

When I downloaded the dataset, I had to decide which period I wanted the works to be completed in - I chose the first week of 2022. What I failed to consider was that there is an inherent difference between multi-chapter works, and works with one chapter only. Whilst a multi-chapter and a single-chapter work may both have been *completed* on the same day, the multi-chapter work has had days (possibly even years) of time in between the upload of the first chapter and the last chapter, giving it a lot more time to accrue kudos, comments, hits, and bookmarks.

So instead of simply removing these two variables, I will instead restrict my dataset to only include works with one chapter. Once I have done this, instead of the 28,152 observations, I have 23,590. Of course, this means my research question has changed slightly (to only include single-chatpter works), but there remains plenty of data to complete the analysis.

Exploring the Data (2)

Let's briefly have a look at the pairs plot for the smaller dataset.

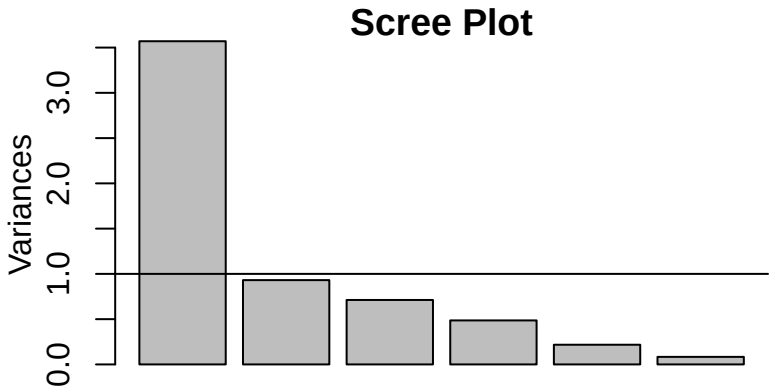


Code for this plot found in Appendix 12

This is what the pairs plot looks like (with the scaled data) after all the multi-chapter works have been removed. This is pretty similar to the plot we saw before. The only thing that we might expect to have changed (since this is the most highly correlated with the removed variables) is `words`, but there isn't anything that stands out to me.

PCA (2)

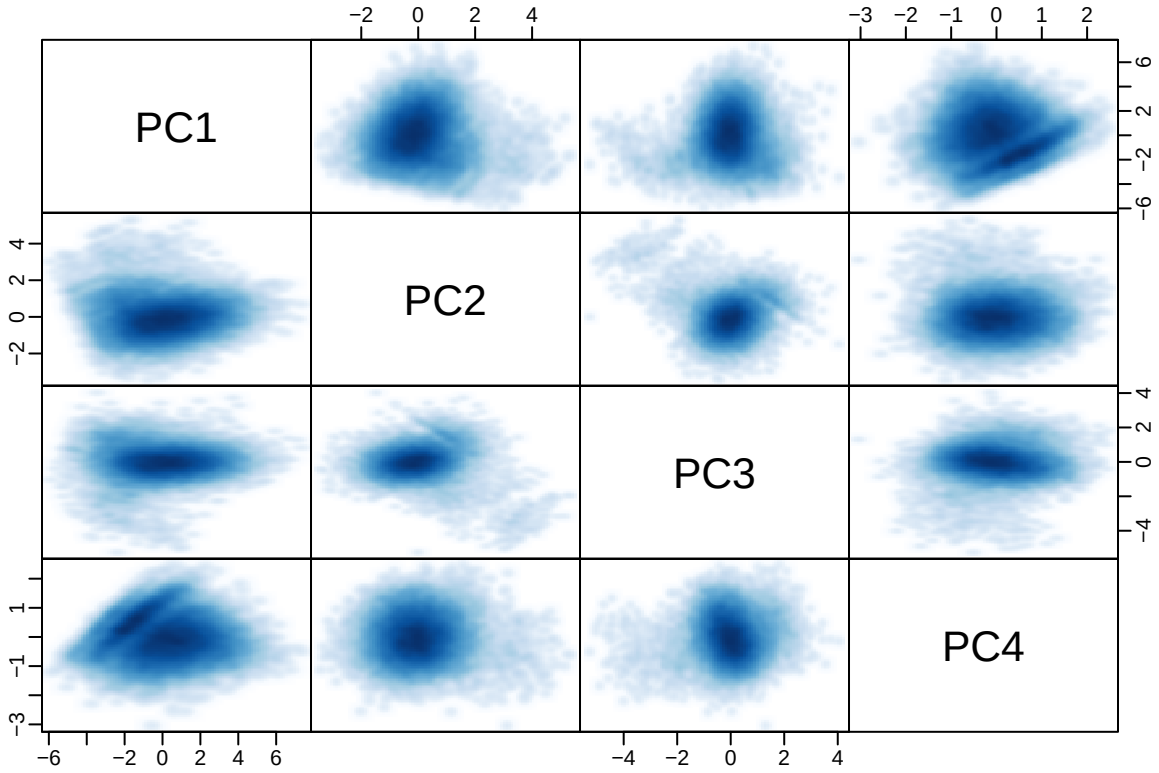
Let's try our PCA again.



Code for this plot found in Appendix 13

It looks as if only one component is needed. Since the variances of PC2, PC3, and PC4 are very similar, I am hesitant to only include a subset of these (for reproduceability). For this analysis, I will keep four dimensions, but I

don't expect most of the components to be interpretable. These four components make up a huge 95% of the variability in the data; the first component is 59.5% (see Appendix 14).



Code for this plot found in Appendix 15

The pairs plot of the first four principal component scores looks a lot better than the first one did. There do seem to be some visible lines and clusters but this is nowhere near as severe as it was in our previous analysis.

```
##
## Loadings:
##            [,1]   [,2]   [,3]   [,4]
## words       0.506 -0.614  0.601
## comments    0.764                -0.584
## kudos       0.934
## bookmarks   0.906
## hits        0.888
## num_tags    0.499 -0.645 -0.564
##
##                  [,1]  [,2]  [,3]  [,4]
## SS loadings     3.570 0.931 0.712 0.486
## Proportion Var  0.595 0.155 0.119 0.081
## Cumulative Var  0.595 0.750 0.869 0.950
```

Code for this table found in Appendix 16

It looks as if our first principal component refers to generally "how good" any given work is, or how users feel about a work after they've viewed it. It is very strongly correlated with kudos, bookmarks, hits, and comments. It looks like people tend to prefer long works, and works with more tags, too. We have lots of big values here due to the fact that the correlations between our original variables are high.

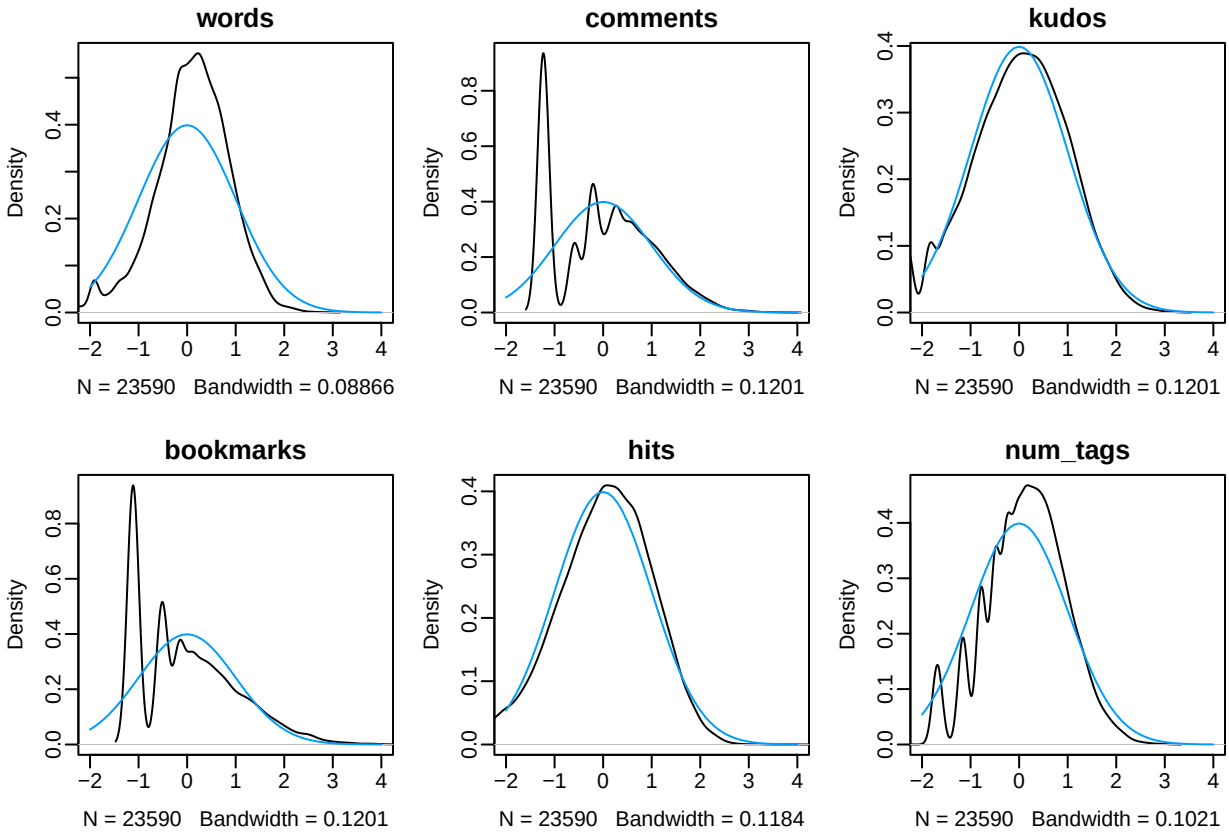
Our second and third components appear to be most highly correlated with `words` and `num_tags`. It looks like, combined, these two components give us similar information to just knowing where a work is in the two-dimensional `words`/`num_tags` space.

The fourth component appears to be a measure of the number of comments. Recall PC4 is orthogonal to PC1, so this must be unrelated to how "good" or "enjoyed" a work is. I suggest that this is a measure of how active a fandom is, or perhaps how close-knit they are.

You could also try rotating the principal component axes to increase interpretability, but I think that these axes are fairly easily interpretable as is, so I won't be including that here.

Next, in preparation for linear discriminant analysis, I want to compare each of the univariate distributions of our original variables to the standard normal density since the histograms aren't all that easy to see in the pairs plot.

Exploring the data (3)



Code for this plot found in Appendix 17

Here, I have used the `density()` function to plot the kernel density estimate (using the default bandwidth) in black. The light blue lines show the standard normal probability density function.

We can see that `hits` and `kudos` match the normal distribution almost perfectly. `words` and `num_tags` also do not seem to be a problem. The 100-word bump is likely what makes the words variable look slightly further from normal than we'd expect. The bumps that we can see in the `num_tags` plot is due to the fact that we have discrete data. The appearance of this graph, as well as some of the other ones with more than one peak is simply due to the bandwidth that the kernel density estimates use.

`bookmarks` and `comments` seem to have a huge spike, before approximating the normal distribution for larger values. I would suggest that perhaps this is because each of these has a much higher number of works with zero in this column. These still remain pretty skewed even after the log transformation.

From this, I would say that the transformed data is not exactly normal, but some variables are pretty close. I would think a normality assumption would be fine. We can see from the pairs plot that all of the scatterplots could be

fairly well approximated with ellipses.

However, in Appendix 18, I do a more rigorous test of normality. I won't go into detail about this here, but it actually shows that the data is very much *not* in fact multivariate normal. Despite this, I think it's okay to use LDA as a classifier. In the worst case scenario, it is just a really bad one!

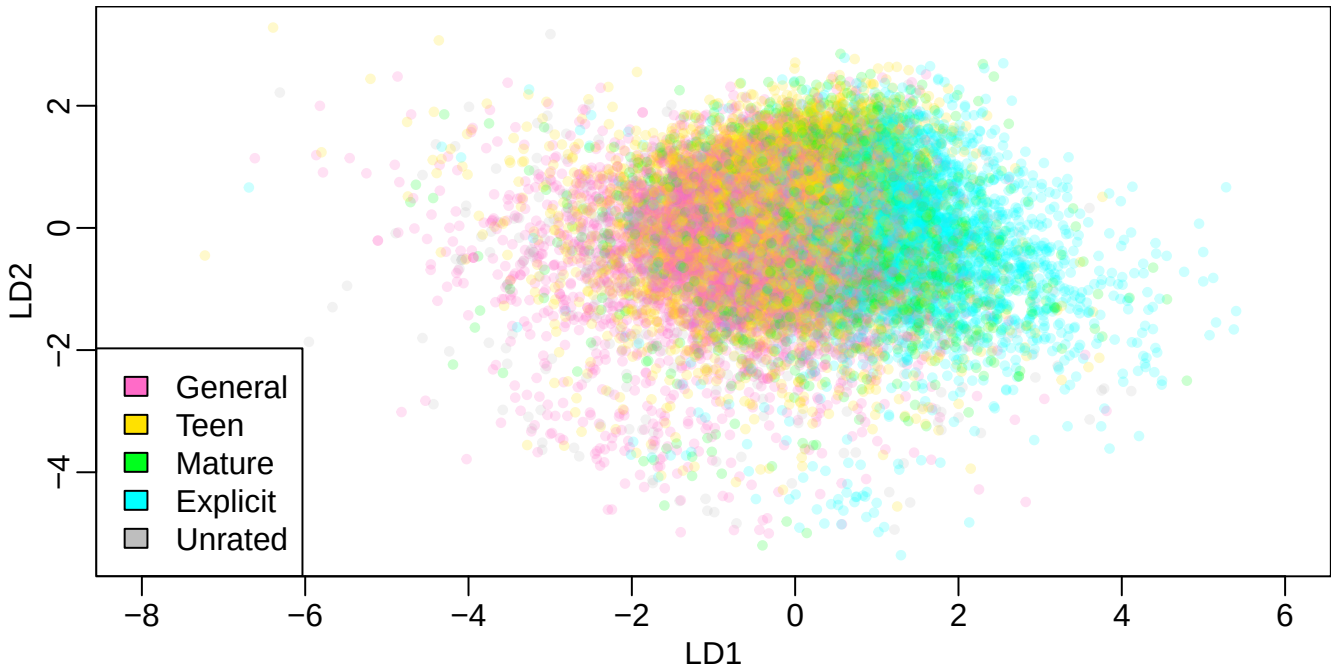
It is worth mentioning that I wanted to use a proper statistical test to see if the other assumption of LDA (equality of covariance) was true. However, the `anova()` function crashed R on my laptop every time I ran it. This is likely due to the sheer number of observations I have. And so, I could not include the results here.

Linear Discriminant Analysis

Linear discriminant analysis (LDA) is similar to PCA, however instead of choosing the axes to maximize variability, they are instead chosen to preserve differences between groups. We can use LDA to classify data, which is my goal of this analysis.

I am interested in classifying the `rating` of the works. Recall from our variable definitions in Appendix 2 that this categorical variable can take on 5 different levels; General Audiences, Teen and Up Audiences, Mature, Explicit, and Not Rated. Note that, although I will be treating these categories as unordered (see full discussion of this in the limitations chapter), it is worth noting that there is some inherent ordering in the categories. By this I mean that General Audiences is a lower rating that Teen and Up audiences which is a lower rating that Mature, and so on.

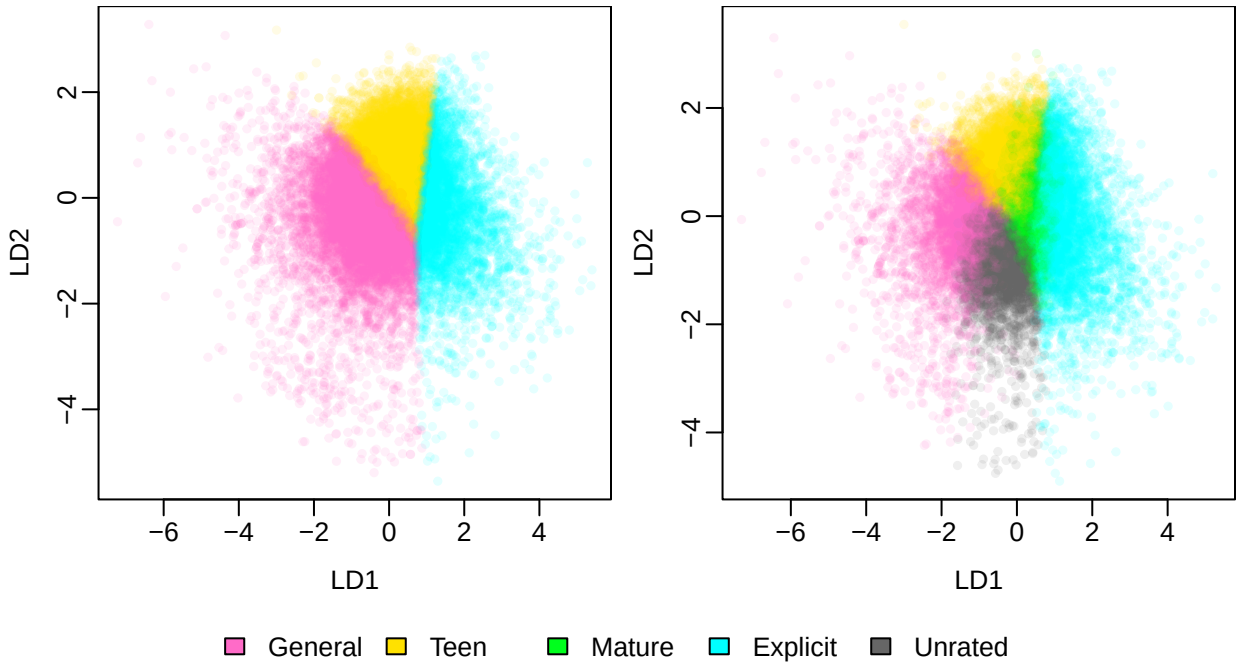
I will use the `MASS` package to perform LDA on our datapoints. Below is a plot of the LD1 and LD2 scores for each point; the colours show the *true* classifications. I have attempted to use alpha-blending to show the different ratings.



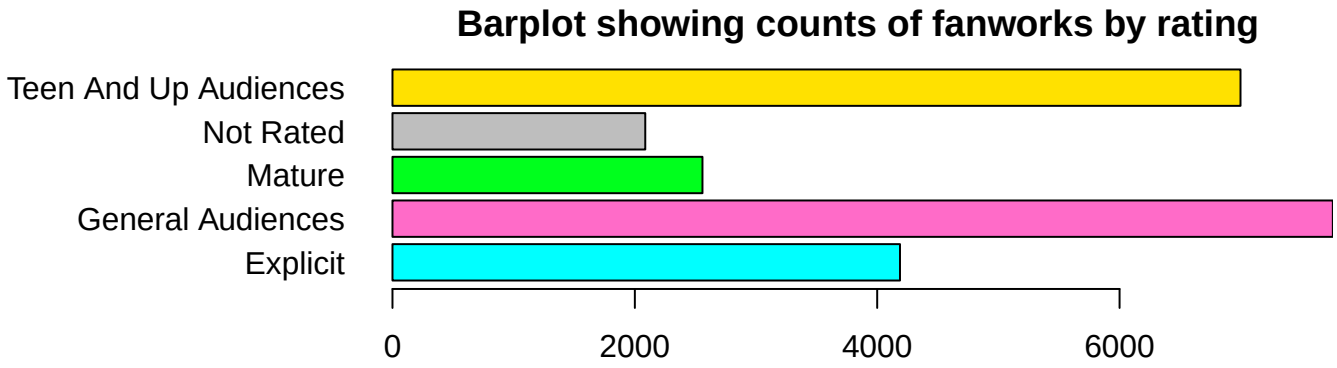
Code for this plot found in Appendix 19

We have so many points that it is hard to see, but I think that LD1 seems to do a decent job of separating the colours. Works with lower values of LD1 seem to generally be rated for younger audiences (i.e. be warmer coloured, and be rated for either general or teen audiences), and works with higher values of LD1 seem to be rated for older audiences (i.e. be cooler coloured, and be in the mature or explicit categories). It's hard to see if there is any difference between the ratings in the LD2 direction.

Now, I am interested in plotting the *predicted* ratings of our works:



Code for this plot found in Appendix 20



Code for this plot found in Appendix 21

The prior I used for the left-hand plot is the default (i.e. it uses the proportions in the training set which are visible in the bar chart above), whereas the prior used in the right-hand plot is one with equal priors for each rating.

- On the left plot, we can see that the model is only classifying the works into three categories; nothing is classified as having no rating or a mature rating. On the bar chart above, it is apparent that these are the smallest categories.
- In the right plot, each of the ratings has several works classified in the category. We can see that these "not rated" works are in the center, which is perhaps what we'd expect if they were a combination of works in other categories. However, it looks like the left-hand plot categorizes all of the grey points as being in the "general" category. I imagine that "general" works are in fact more likely to be labelled as "not rated", since it might be easy to mistake these ratings for each other. It is also more of an issue if explicit works are incorrectly (or ambiguously) labelled then if general works are. In the right-hand plot, the "mature" category seems to be right in between the "explicit" and "teen" categories. But this time the left-hand plot categorizes almost all the green points as "teen". I suppose since these "mature" and "teen" don't have as distinct of a meaning as the other categories; I wouldn't be surprised if these two tags are often used interchangeably.

I would argue that we want our model to be best suited to future data. Any new data is likely to have the same rating ratio as the data in this dataset. And so I will use the default prior for my model.

Let's do some cross validation to see how good the model is.

```
##
##                                Predicted Explicit Predicted General Audiences
##   Actual Explicit                            2543                         518
##   Actual General Audiences                    451                        5052
##   Actual Mature                               656                         904
##   Actual Not Rated                            285                        1237
##   Actual Teen And Up Audiences                629                        3209
##
##                                Predicted Teen And Up Audiences
##   Actual Explicit                                         1127
##   Actual General Audiences                                2257
##   Actual Mature                                            998
##   Actual Not Rated                                         564
##   Actual Teen And Up Audiences                            3160
```

Code for this table found in Appendix 22

This gives us a correct classification rate of 10755 out of 23590, or about 45.6%. Considering the data we are working with, this is a lot better than I would have expected. If we allow prediction to within one rating either side, we get a correct classification rate of 75.8%.

Let's see if we can analyse what the linear components mean:

```
##
## Loadings:
##            [,1]  [,2]
## words      0.445 0.754
## comments        0.629
## kudos      0.303 0.410
## bookmarks 0.357 0.441
## hits       0.668
## num_tags   0.471 0.384
##
##                  [,1]  [,2]
## SS loadings    1.119 1.499
## Proportion Var 0.186 0.250
## Cumulative Var 0.186 0.436
```

Code for this table found in Appendix 23

I would suggest that the first component measures click-through rate, or perhaps how appealing a work looks from its description. This is strongly correlated with `hits`.

The second component is most strongly correlated with `words` and `comments`. It is interesting that the `words` and `comments` original variables are not strongly correlated with each other; they have a correlation of 0.29. I'm not sure I can think of some other factor that would influence both `words` and `comments` more strongly than any of the other variables!

We saw that LD1 was the best as classifying our ratings (especially explicit works from works for general/teen audiences). Since we said that LD1 represents click-through rate, I would suggest that it is the case that in general, works with higher LD1 scores are more appealing than works with lower LD1 scores. We can see from our scatterplot that it is the explicit fanworks that have higher scores, and works for general/teen audiences with lower LD1 scores. So it looks like, in general, people are more inclined to click on explicit fanworks compared to those written for general/teen audiences. Also since LD1 is positively correlated with `kudos` and `bookmarks`, people tend to rate higher-rated works better after they have read them, too.

From our scatterplot, it looks like Teen and Up rated works have high LD2 scores compared to General Audiences.

We saw that this was highly correlated with `words` and `comments`. I suppose that the longer the work is, the more likely it is to be bumped up from being in the General Audiences category to being in the Teen category. Short works don't really have a chance to get a higher rating. I don't have an explanation for why this category has a higher comment rate than general audience works.

It is worth noting that our first LDA component explains 69.8% of the variability in the data and the second explains a further 18.6% (see Appendix 24).

Conclusion

After restricting my dataset to only include single-chapter works, transforming the data, and doing principal component analysis, I found that 60% of the variation in the data is explained in one dimension. This means that we can get 60% of the information that the 6 original variables tell us in just one new variable - PC1. We saw that, since PC1 is highly correlated with our `bookmarks` and `kudos` variables, it represents how much a particular work is liked. PC1 has a medium to strong positive correlation with every original variable. This indicates that the more comments, kudos, bookmarks, hits, tags, or words a fanwork has, the more people like it. This is unsurprising! It does, however, suggest that if someone is interested in creating a popular work of fanfiction, they should write a lot of words, and tag it properly.

I also did a linear discriminant analysis. Recall that the linear discriminant axes are chosen to preserve group differences. We found that LD1, which represents the appeal a work has, distinguishes Explicit works (with high LD1 scores) from all other ratings (with low LD1 scores) the best. LD2, which represents how many words and comments and work has, discriminates between General Audiences (with low LD2 scores) and Teen Audiences (with high LD2 scores) the best. I suppose this suggest that if you want to become a successful fanfiction author, it is better to write adult content!

I personally found this study very interesting; I've wanted to look at data regarding AO3 for a long time. I'm not sure that the conclusions I've found are particularly surprising or useful to many people, but perhaps it gives a small insight into the inner workings of fandom communities on the internet. There is still so much more you could do with this data - I'm sure that I have only just scraped the surface of all the fascinating analysis that could be done here.

Optional Chapter

I have included a limitations section and a future work section here because I have quite a lot to say under these topics. I am aware that they are an additional page and a half, so feel free to skip this chapter!

Limitations

- Recall that the dataset that I am using is all English-language works completed in the first week of 2022. Now, it is worth noting that any conclusions I have made in this report cannot necessarily be generalized to *all* works on AO3. I used 7 days worth of data, which hopefully removes any weekly effects (e.g. a difference in the works posted on weekends vs weekdays), but yearly or monthly effects are not accounted for. It is very likely that there are shifts in usage of AO3 over time; perhaps the fact that January is a holiday makes the data taken in this month different to other times of the year.
- Recall also that in our second PCA and our LDA that we are only looking at single-chapter works. As we have seen , when using the techniques we've learned so far in this course, and this particular dataset, multi-chapter works have to be studied separately from single-chapter works. Mostly, this issue was due to the fact that I had restricted the dataset to be the works completed in the first week of 2022 which gave the multi-chapter works an "advantage" when it comes to various numeric variables. Now, if we were to instead take a random subset of all the works uploaded at any time on AO3, perhaps we could have gotten around this issue. The problem then would be that we'd then have four types of work; incomplete with an unknown number of chapters (e.g. 2/?), incomplete with a known number of chapters (e.g. 2/3), complete with multiple chapters (e.g. 2/2), and complete with one chapter (i.e. 1/1). At this point, using numeric variables to describe the works (especially with how they relate to the number of words) would be hard. I think my way of restricting the works to be single-chaptered was the simplest way to go, even if it meant that my findings could not be generalised to the multi-chapter case.

- The works I was looking at are from a wide range of fandoms; it is not clear that the trends in the variables I have been looking at are the same for each one. Since some fandoms are bigger than others; these will weight my analysis accordingly. The top three fandoms in my dataset are My Hero Academia, Harry Potter, and Genshin Impact; around a tenth of works are tagged with one of these fandoms. This is pretty significant, especially as there are 5,445 total fandoms in the sample. If there exists a fandom that is very large but not representative of the general AO3 trends, this could have disproportionately influenced my analysis.
- I could not download all the necessary data instantaneously. Scraping AO3 took several days (the data was downloaded between the 4th and the 11th of March). This means that the data that was scraped first had slightly less time to gather kudos/hits/comments. I doubt this will effect the analysis in any significant way, but it is worth mentioning.
- Since the Archive does not record trends in metadata over time, it will be near-impossible to reproduce my dataset exactly; it represents a snapshot of the what the Archive looked like at a past date. This is very unhelpful if we are to consider reproduceability. Although my methods should be generalizable to new data, I am likely the only person with access to the exact data that I used in this project.
- In my LDA, I included the "Not Rated" category as a category of its own. However, it is possible to think of the unrated works as being missing data of the "missing not at random" variety. Because of this, I was hesitant to leave it out, since being labelled as "Not Rated" does tell us something about that work (for example, I believe that "General Audiences" works are more likely to be labelled "Not Rated" than "Explicit" works are). If we instead remove all the unrated works, and perform LDA on the 4 remaining categories, our model does better. It has a 50.1% correct cross-vaildated classification rate (compared to 45.6%) and gets 83.2% to within one level (compared to 75.8%). This is mostly because the model we did use doesn't classify anything in the unrated category anyway! So removing these works that the model is guaranteed to get wrong increases the hit rate.
- I discussed earlier that I couldn't run the test for equality of covariance on my laptop. I could, however, run QDA as well as LDA, and compare the two. It turns out that QDA has a correct (cross-validated) guess rate of 44.27% compared to 45.6% for LDA, so is slightly worse at prediction. This comparison of LDA and QDA makes me think that maybe our equal covariance assumption is fine, since relaxing the assumption doesn't seem to help the classification error.

Further Study

- It would be really interesting to do sentiment analysis on the tags. Perhaps we could determine the number of kudos from the text of the tags themselves? It would be possible to create binary variables, one for each tag used in the entire dataset. Each observation would be given a 0 if the tag was used and 1 if it was not. Although, perhaps it isn't that simple. On AO3, there exists the concept of a canonical tag. Essentially, multiple tags can be used to mean the same thing (for example, Tenth Doctor/Rose Tyler and Rose Tyler/Tenth Doctor) - only one tag for each meaning is the "canonical" one. The tag situation is so messy that there is a whole team of people who volunteer for AO3 called tag wranglers, whose job it is to organize the tags on the Archive. There does exist a downloadable dataset which has information about all the tags used on the Archive, each synonymous tag linked to its canonical counterpart, which would have to be used. Essentially, an analysis of the tags on AO3 would be fascinating, but would likely be a lot of work in practice.
- A comparison of the different fandoms in the dataset would be interesting. LDA would not necessarily be suitable, due to the sheer number of fandoms, but comparing two different popular fandoms could be interesting. Also, a look at how a fandom changes over time could be interesting!
- I have two more categorical variables that it *would* be suitable to do LDA on; `category` and `fluff_angst`. If it hadn't been for the length of this report already, I would have included them here!
- Finally, a test to see if there actually is a statistically significant difference between our different rating groups would be useful!

Appendices

Appendix 1 - tidying the data

```
library(tidyverse)
library(lubridate)
library(dplyr)
fanworks = read.csv("all.csv")
fanworks[,17:20][fanworks[17:20]=="null"]<-0
fanworks[,6][is.na(fanworks[6])]<-"None"
rownames(fanworks) = fanworks$work_id
fanworks = fanworks[,c(5:6,10,12,14:20)] %>%
  mutate(num_tags = {a = str_count(additional.tags, ',')+1; ifelse(is.na(a), 0, a)}) %>%
  mutate(daystocomplete = abs(as.integer(ymd(status.date)-ymd(published)))) %>%
  mutate(fluff_angst = {fluff = grepl("Fluff|fluff", additional.tags);
  angst = grepl("Angst|angst", additional.tags);
  fluff_angst = ifelse(fluff>0, ifelse(angst>0, "both", "fluff"),
                       ifelse(angst>0, "angst", "neither"))}) %>%
  dplyr::select(-c("additional.tags", "status.date", "published")) %>%
  mutate(chapters = str_extract(chapters, "^[^\\./:]+"))
fanworks[, 3:10] <- sapply(fanworks[, 3:10], as.numeric)
fanworks[, c(1:2,11)] <- sapply(fanworks[, c(1:2,11)], as.factor)
```

Appendix 2 - definitions of the variables

- The row name, previously `work_id`, is the unique identifier of each work.
- `rating` is a categorical variable. Each work has one of five ratings - General Audiences, Teen and Up Audiences, Mature, Explicit, or Not Rated. This describes the suitability of a work for different audiences.
- `category` is a categorical variable, which has seven possible values - M/M, F/M, F/F, Gen, Multi, Other, and None. This describes which relationships, pairings, or orientations are present in the work. For example, M/M indicates that male/male relationships are present, Gen means that no relationships are present, and Multi means that more than one kind of relationship, or a relationship with multiple partners is present.
- In the uncleaned data, there was an `additional tags` column. Tags are used in order to help users search and filter works on the Archive website; the original column associated each work with a string of every tag on the work. I have mutated this to be a numeric variable, `num_tags`, which instead gives the number of tags.
- In the uncleaned data, we had the `published` variable which gave the date the work was initially published, and `status date` which gave the date the work was most recently updated. I mutated these columns to instead have the `daystocomplete` variable, which takes the difference of the date published and the date completed.
- `words` and `chapters` are self-explanatory - giving the number of words and number of chapters in each work
- Both users and non-users of AO3 can leave comments on works; `comments` is a numeric variable which gives the number of comments on the work
- On AO3, kudos are given in a similar way that "likes" are given on Twitter or YouTube. `kudos` is a numeric variable referring to the number of kudos left on the work.
- Users on AO3 can add a work to their "bookmarks" page on their profile. `bookmarks` is a numeric variable which counts how many times this has been done.
- `hits` is a numeric variable which gives the number of times the work has been viewed.
- `fluff_angst` gives an indication as to whether the words "fluff" and/or "angst" were used in the tags. This is a categorical variable.

Appendix 3 - example of five observations at random

```
fanworks[sample(28152,5),]
```

Appendix 4 - scatterplot of words against kudos

```
library(scales)
par(mfrow=c(1,2), mar=c(4,4,1,1), oma=c(0,0,0,0))
limits = cbind(c(2200000, 10000), c(20000,200))
trans = c(0.2, 0.05)

for (i in 1:2){
  cols = rep(scales::alpha("#009dff", trans[i]), nrow(fanworks))
  cols[4832] = scales::alpha("blue",1)

  plot(fanworks[,c(3,6)], col=cols, pch=19,
       xlim=c(0,limits[1,i]), ylim=c(0,limits[2,i]))
  abline(lm(fanworks[,6]~fanworks[,3]), col="black")
  abline(lm(fanworks[-4832,6]~fanworks[-4832,3]), col="red")
}
```

Appendix 5 - quantiles of each variable

```
numeric_vars = fanworks[,3:10]
sapply(numeric_vars, function(x) quantile(x))
```

Appendix 6 - scatterplot of words against kudos (log scale)

```
par(mar=c(3,3,0,0))
cols = scales::alpha(rep("#009dff", 28152), 0.02)
cols[4832] = "blue"
plot(log(fanworks[,c(3,6)]+1), col=cols, pch=19, mgp=c(2,1,0))
abline(lm(log(fanworks[,6]+1)~log(fanworks[,3]+1)), col="black")
abline(lm(log(fanworks[-4832,6]+1)~log(fanworks[-4832,3]+1)), col="red")
```

Appendix 7 - pairs plot

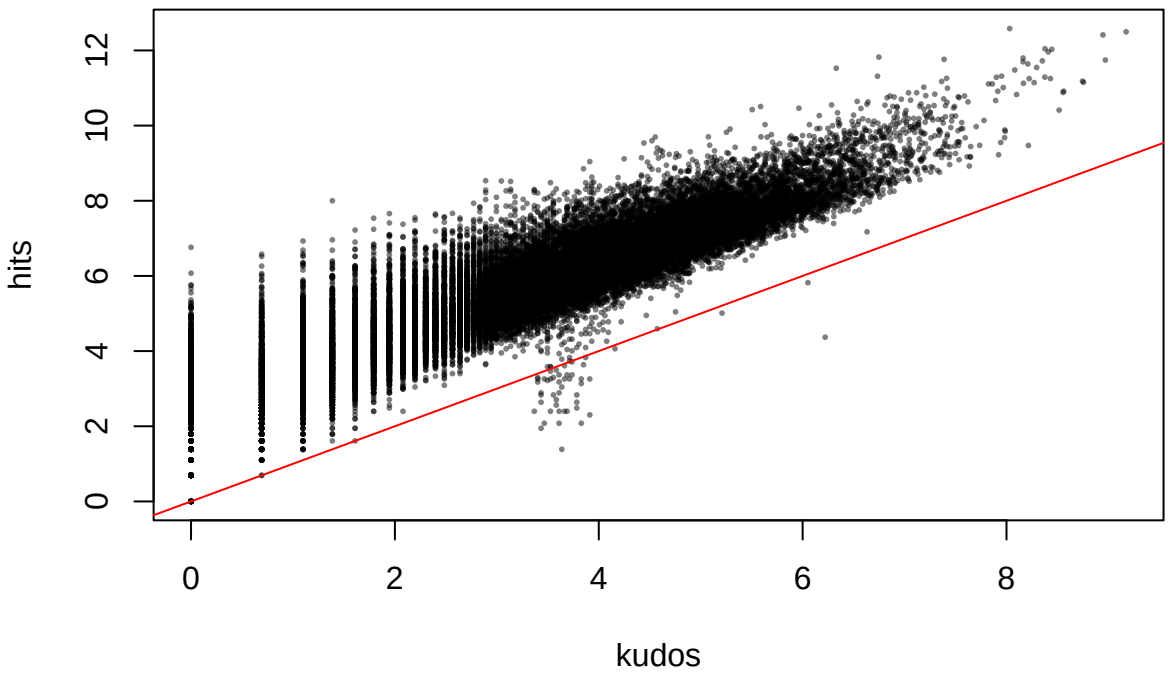
```
library(psych)

pairs.panels(log(numeric_vars+1), density = F,cor = TRUE, stars=F, ellipses = F,
             smoother=T, rug=F, gap=0, hist.col="white", scale=T, cex=1.5)
```

Appendix 8 - hits against kudos

Here, the red line marks the line $x = y$.

```
plot(log(numeric_vars+1)[,c(4,6)], pch=16,cex=0.4, col=scales::alpha("black",0.5))
abline(0,1, col="red")
```




```
hits_kudos = log(numeric_vars+1)[,c(4,6)]
sum(hits_kudos[,1]>hits_kudos[,2])
```

```
## [1] 48
```

There are 48 points below the red line, almost all with a similar number of kudos (between 20 and 50).

Appendix 9 - scree plot

```
par(mar=c(0.1,3,1,0))
s_fanworks = sapply(log(numeric_vars+1), function(x) (x-mean(x))/sd(x))
results = prcomp(s_fanworks)
plot(results, main="Scree Plot", mgp=c(2,1,0))
abline(h=1)
```

Appendix 10 - Variability of first two PCA components

```
cumsum(results$sdev^2)/8
```

```
## [1] 0.5248762 0.7200079 0.8248152 0.8926775 0.9400929 0.9675191 0.9906425
## [8] 1.0000000
```

Appendix 11 - first two principal component scores of each observation

```
par(mfrow=c(1,2), mar=c(3,3,0,2))
plot(results$x[,1:2], col=scales::alpha("#08326E", 0.3), pch=19, mgp=c(2,1,0))
abline(lm(results$x[,1] ~ results$x[,2]), col="red")
plot(results$x[,1:2], col=scales::alpha("#08326E", 0.02), pch=19, mgp=c(2,1,0))
abline(lm(results$x[,1] ~ results$x[,2]), col="red")
```

Appendix 12 - a second pairs plot

```
fanworks_new = dplyr::filter(fanworks, chapters==1)
numeric_new = fanworks_new[,c(3,5:9)]
scaled_new = sapply(log(numeric_new+1), function(x) (x-mean(x))/sd(x))

pairs.panels(scaled_new, density = F,cor = TRUE, stars=F, ellipses = F,
             smoother=T, rug=F, hist.col = "white", gap=0, scale=T, cex=1.5)
```

Appendix 13 - a second scree plot

```
par(mar=c(0.5,3,1,0))
results2 = prcomp(scaled_new)
plot(results2, main="Scree Plot", mgp=c(2,1,0))
abline(h=1)
#sum(results2$sdev[1:4]^2)/6
```

Appendix 14 -

```
cumsum(results2$sdev^2)/6
```

```
## [1] 0.5950233 0.7502159 0.8688799 0.9498860 0.9861100 1.0000000
```

Appendix 15 - pairs plot of first four principal component scores

```
pairs(results2$x[,1:4], panel = function(...) smoothScatter(...,nrpoints = 0,
                                                            add=TRUE, nbin=100),
      oma=rep(1.5,4),gap=0, mgp=c(0,0.5,0))
```

Appendix 16 - correlations between original variables and principal components

```
cor_mat = matrix(NA, ncol=4, nrow=ncol(scaled_new))
for (i in 1:6){
  for (j in 1:4){
    cor_mat[i,j] = cor(scaled_new[,i],results2$x[,j])
  }
}
rownames(cor_mat) = colnames(scaled_new)
class(cor_mat) = "loadings"
print(cor_mat, cutoff=0.4)
```

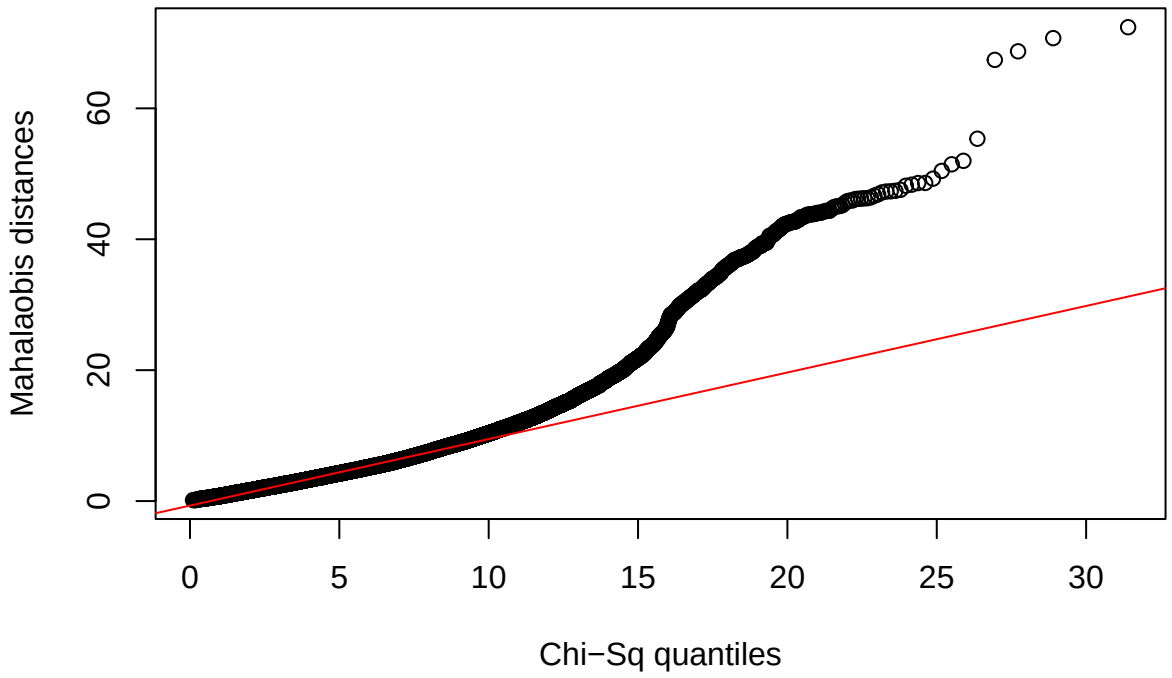
Appendix 17 - a normality check

```
par(mfrow=c(2,3), oma=rep(0,4), mar=c(4,4,2,0), mgp=c(2,0.5,0))
labels = colnames(scaled_new)
for(i in 1:6){
  plot(density(scaled_new[,i]), xlim=c(-2,4), main=labels[i])
  lines(seq(-2,4,0.1), dnorm(seq(-2,4,0.1)), col="#009dff")
}
```

Appendix 18 - rigorous test of multivariate normality

```
library(mvabund)
mod<-manova(scaled_new ~ as.factor(fanworks_new[,1]))
```

```
Resid<-mod$resid
Resid<-abs(Resid)
modperm<-manylm(Resid~factor(fanworks_new[,1]), cor.type = 'R', test="F")
Sigma<-summary(mod)$SS$Resid/(nrow(scaled_new)-5)
mahalanobis(mod$residuals,0,Sigma)->mahal
qqplot(qchisq(ppoints(nrow(scaled_new)),ncol(mod$resid)),y=mahal, xlab="Chi-Sq quantiles",
       ylab="Mahalaobis distances")
qqline(mahal, distribution=function(p) qchisq(p, df=ncol(mod$resid)), col="red")
```



Our Malalanobis distances look close to the line initially, but start straying very far in the higher quantiles, indicating that multivariate normality isn't the best fit.

```
ks.test(mahal, "pchisq", ncol(mod$resid))
```

```
## Warning in ks.test.default(mahal, "pchisq", ncol(mod$resid)): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  mahal
## D = 0.10266, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Our Kolmogorov-Smirnov test results give a highly significant result, meaning we can reject the null hypothesis that are data are multivariate normal.

Appendix 19 - first two LDA scores coloured by true rating

```
library(MASS)

lda = lda(scaled_new, fanworks_new[,1])
scores = predict(lda)

p = hcl(h=seq(0,360,length.out=6),c=200,l=90, fixup=T)

cols = fanworks_new$rating
cols[cols == "General Audiences"] = p[1]; cols[cols == "Teen And Up Audiences"] = p[2]
cols[cols == "Explicit"] = p[4]; cols[cols == "Mature"] = p[3]
cols[cols == "Not Rated"] = "grey"

par(mar=c(2.5,2.5,0,0))
plot(scores$x, col=scales::alpha(cols,0.2), pch=19, cex=0.6, xlim=c(-8,6),
     xlab="LD1", ylab="LD2", mgp=c(1.5,0.5,0))
legend(x = "bottomleft", fill= c(p[1:4], "grey"),
       legend=c("General", "Teen", "Mature", "Explicit", "Unrated"))
```

Appendix 20 - first two LDA scores coloured by predicted rating

```
par(oma=c(0,0,0,0))
m <- matrix(c(1,2,3,3),nrow = 2,ncol = 2,byrow = TRUE)
layout(mat = m,heights = c(0.9,0.1))

cols = as.character(scores$class)
cols[cols == "General Audiences"] = p[1]; cols[cols == "Teen And Up Audiences"] = p[2]
cols[cols == "Explicit"] = p[4]; cols[cols == "Mature"] = p[3]
cols[cols == "Not Rated"] = "grey40"

par(mar=c(3.5,3.5,0,0))
plot(scores$x, col=scales::alpha(cols,0.1), pch=19, cex=0.6, xlab="LD1",
     ylab="LD2", mgp=c(2,0.5,0))

lda2 = lda(scaled_new, fanworks_new[,1], prior = rep(1/5,5))
scores2 = predict(lda2)

cols = as.character(scores2$class)
cols[cols == "General Audiences"] = p[1]; cols[cols == "Teen And Up Audiences"] = p[2]
cols[cols == "Explicit"] = p[4]; cols[cols == "Mature"] = p[3]
cols[cols == "Not Rated"] = "grey40"

plot(scores2$x, col=scales::alpha(cols,0.1), pch=19, cex=0.6, xlab="LD1",
     ylab="LD2", mgp=c(2,0.5,0))

par(mar=c(0,0,0,0))
plot(1, type = "n", axes=F, xlab="", ylab="")
legend(x = "bottom", fill= c(p[1:4], "grey40"), horiz = TRUE, bty='n', inset=0,
       legend=c("General", "Teen", "Mature", "Explicit", "Unrated"))
```

Appendix 21 - plot of counts for each rating

```
par(mar=c(2,10,2,0))
barplot(table(as.factor(fanworks_new[,1])), hori=T, las=1, col=c(p[c(4,1,3)], "grey", p[2]),
```

```
        main="Barplot showing counts of fanworks by rating")
```

Appendix 22 - confusion matrix for LDA cross validation

```
lda_cv = lda(scaled_new, fanworks_new[,1], CV=TRUE)
(tab = table(paste("Actual", fanworks_new$rating), paste("Predicted", lda_cv$class)))
```

Appendix 23 - correlations of original variables with LDA components

```
cor_mat = matrix(NA, ncol=2, nrow=6)
for (i in 1:6){
  for (j in 1:2){
    cor_mat[i,j] = cor(scaled_new[,i],scores$x[,j])
  }
}
rownames(cor_mat) = colnames(scaled_new)
class(cor_mat) = "loadings"
print(cor_mat, cutoff=0.2)
```

Appendix 24 - singular values for LDA

```
lda$svd/sum(lda$svd)
```

```
## [1] 0.69836745 0.18618255 0.08517119 0.03027882
```