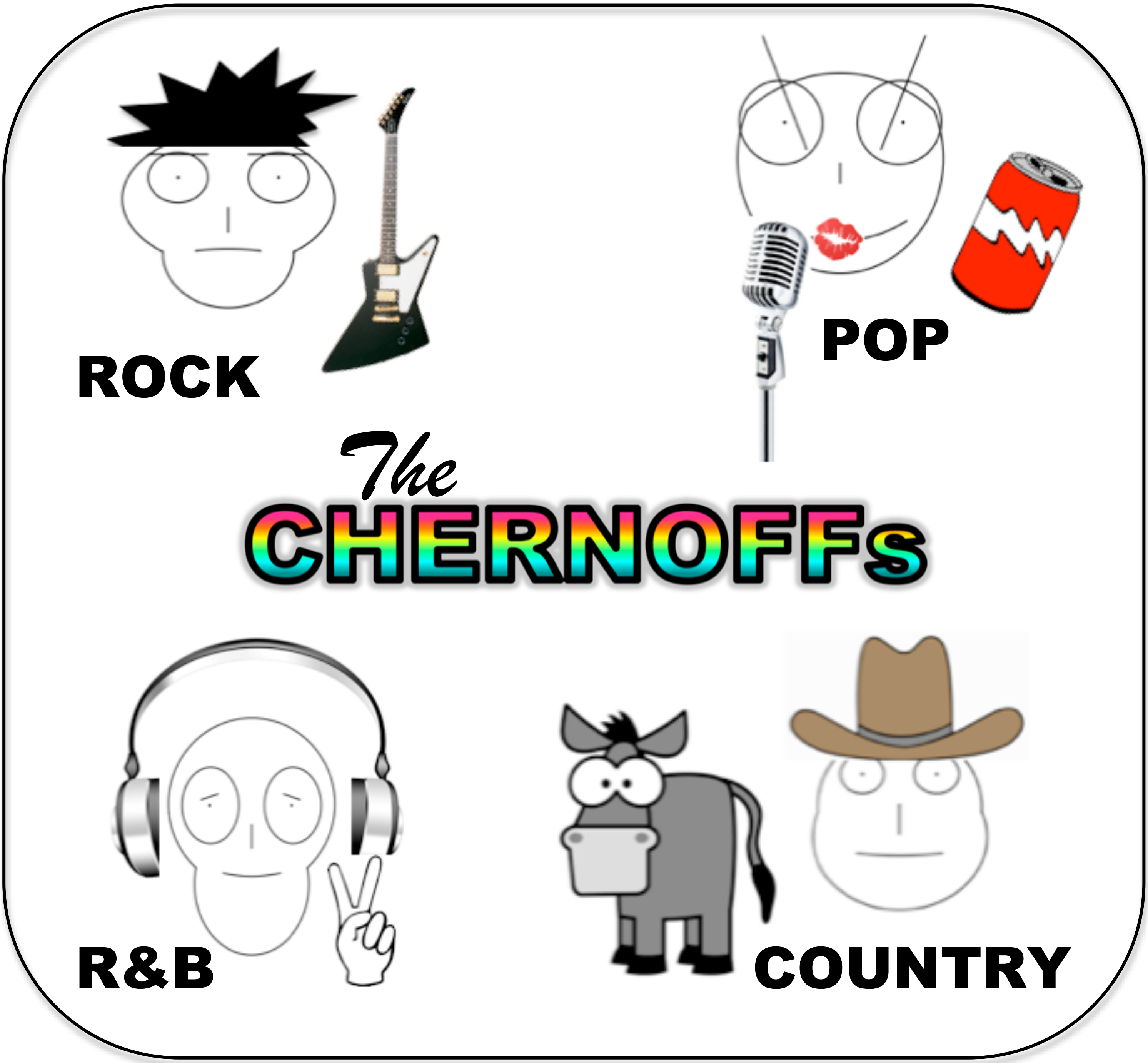
STATS 415

DATA MINING PROJECT

Insights into the associations between song genres, and the rating behavior of music fans

Annalyn Ng & Ben Charoenwong



STATS 415 Charoenwong & Ng

Data Mining Project 2011

Data

The dataset is courtesy of Yahoo! music, containing rating scores on songs. 800,000 unique users rated 10 songs each, resulting in a total of 8 million observations. For each observation, there were 3 factors: user ID, rating score, and song genre. Ratings scores were from “1” to “5”.

Out of the 8 million ratings, 87% are of songs with an unknown genre. In other words, these songs did not fall into any of the recognizable categories within the Yahoo! database. Within the remaining 13% of songs, distribution over the genre groups is grossly uneven. Specifically, the four genre groups of Country, Pop, R&B and Rock comprise approximately 90% of the songs with a known genre. As such, we streamlined our analysis by focusing on these four largest genre groups.

Overview

We had aimed to predict whether a user would rate a song of an unknown genre favorably (“5” or “4”) or unfavorably (“1” or “2”). Average ratings across the four genre groups were calculated per user and used as predictors. This approach is based on the assumption that songs of unknown genre have a distinctive musical style that can be associated with styles of known genres. For instance, we might find that a user with high ratings for rock songs rated songs of the unknown genre favorably, implying that the unknown genre might be closely related to the rock genre.

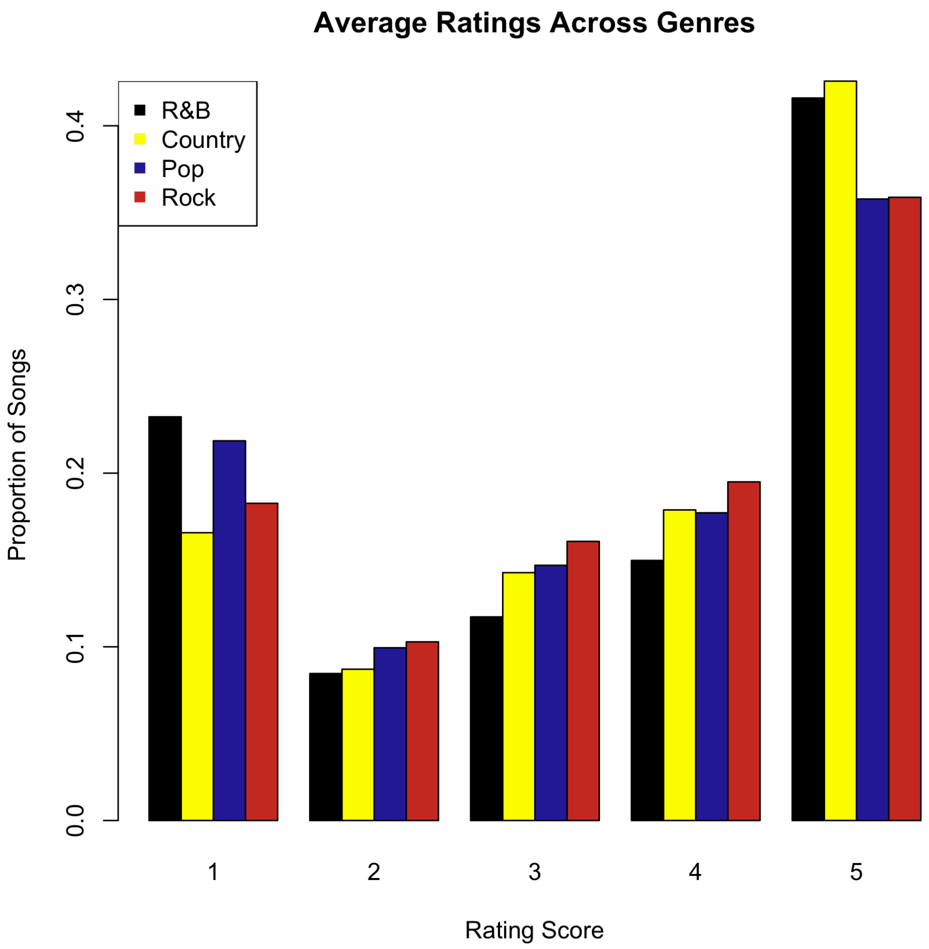
Using classification techniques, we were able to predict whether a user would like a particular song with relatively high accuracy (about 70% success rate). However, further investigation revealed that this might be due to a considerable proportion of users who only rated “5”s. In other words, these users only rated songs they really liked and neglected to rate otherwise, yielding a perfect correlation between their average and individual ratings. This prompted us to perform a clustering analysis to test possible differences in user rating behaviors.

Before elaborating on the above-mentioned analyses, we will first present you with the results of our data exploration and visualization techniques that compare song genres.

Table of Contents

1. Visualizations ---------------- General comparison of genres
2. Dimension Reduction ---------------- Components that predict a song’s rating
3. Classification ---------------- Using derived components for classification/prediction
4. Clustering ---------------- Users grouped via rating behavior
5. Limitations ---------------- Discussion of possible pitfalls in analysis

1. Visualizations



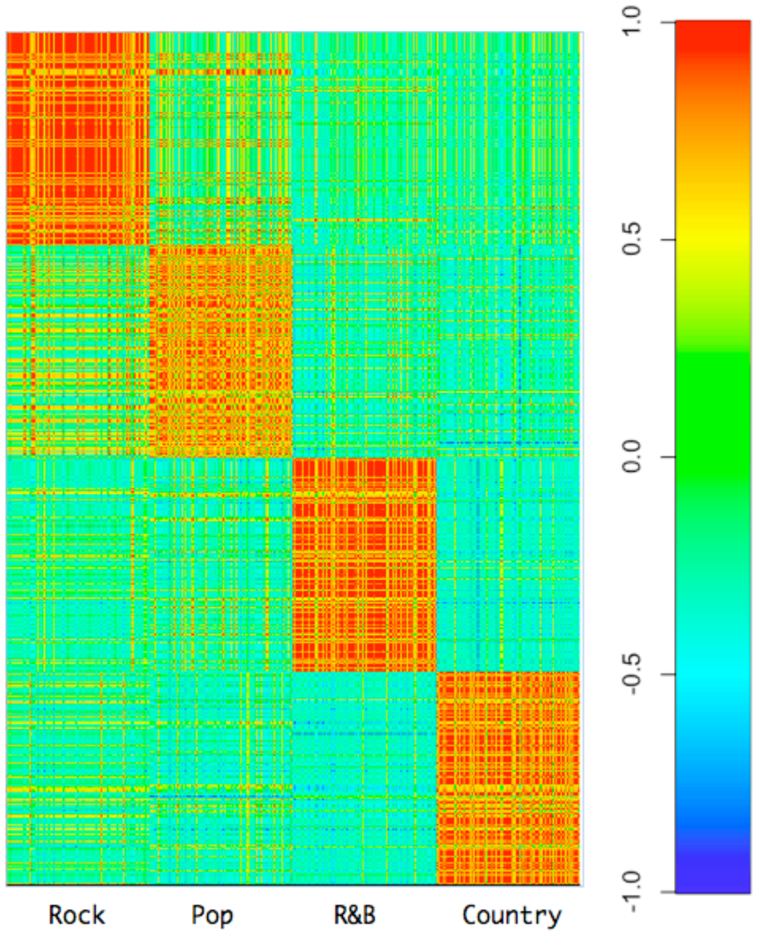
We obtained the frequency counts for the rating scores of songs in each genre, using the normalized counts to construct the barplot.

1.1 Users are more likely to rate when they strongly like/dislike a song

There are relatively more "5"s and "1"s as compared to the other rating scores. This suggests that users only rated when they felt strongly for or against a song, and that they might not rate every song that they listen to.

1.2 Country songs seem the most favored

Country songs scored the most "5"s and the least "1"s on average. This could be attributed to the common themes of hope and optimism contained in the lyrics of Country songs, which users might find hard to dislike. However, it could also imply that users who listen to Country songs are more lenient with their ratings.



This heatmap compares preferences of music fans from the four genre groups. A user was defined to be a fan of, for instance, Rock music if he had rated all his Rock songs as "5"s. Though a user could be a Rock fan and a Pop fan at the same time, he would only appear once in this heatmap in either the Rock or Pop group. 250 music fans were randomly selected for each genre group, and the factors used in the correlation analysis were the user's average ratings for songs across the four genres.

1.3 Rock fans have highly homogenous tastes for other music genres

The largest region of red is within the rock fan group. Based on this finding, one might harness information on the general music preferences of Rock fans in selecting a desired song for a single Rock fan. The opposite seems true for Pop fans, which has the most yellow lines intersecting its region of red.

1.4 Considerable overlap in music preferences of Rock fans and Pop fans

There are many regions of yellow in the shared space between Rock and Pop. The two genres originated from about the same time period (1950s to 1960s), with Pop music having its roots in Rock. Depending on how Yahoo! defines its genres, one might even view Rock music as a subset of Pop (Popular) music.

1.5 Negative correlation in music preferences between R&B and Country fans

There are regions of blue in the shared space between R&B and Country. This suggests that fans from the two genre groups have opposing music tastes. Taken to the extreme, this would imply that songs desired by R&B music fans are disliked by Country music fans.

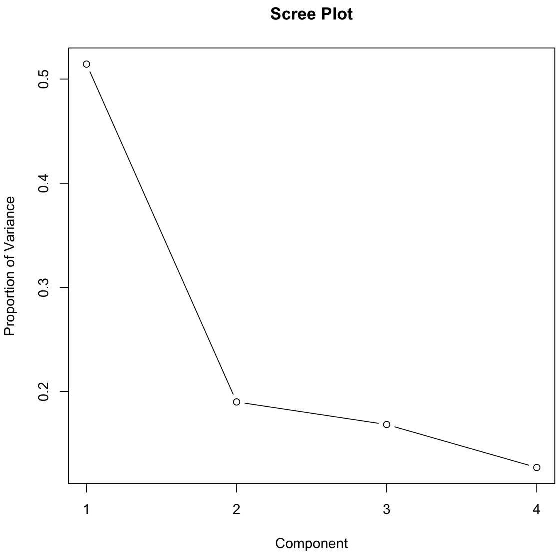
To further assess the associations between genres with respect to music preferences, we present the results of a principal component analysis next.

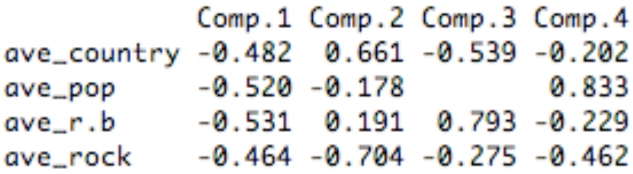
2. Dimension Reduction

We chose principal components analysis (PCA) over multi-dimensional scaling (MDS) because there are extensively more observations than variables, therefore PCA will be computationally more efficient than MDS. Furthermore, we have access to each data point and values for each variable, as opposed to only having the distance/Gram matrix.

The factors employed in PCA are identical to that used previously in the heatmap analysis, which are users' average ratings for songs across all four genres. We ensured that users included in the analysis had rated all four genres of music at least once.

The scree plot and loadings are displayed below:





```
            Comp.1 Comp.2 Comp.3 Comp.4
ave_country -0.482  0.661 -0.539 -0.202
ave_pop     -0.520 -0.178         0.833
ave_r.b     -0.531  0.191  0.793 -0.229
ave_rock    -0.464 -0.704 -0.275 -0.462
```

The scree plot suggests the use of the first two components, accounting for 70% of the total variance. The 1st Component is a weighted average of all the mean ratings of songs by genre, while the 2nd Component weighs average rating for Country songs against that of Rock songs.

It is worthwhile to note that the 3rd Component, which explains a sizeable 17% of the total variance, weighs average rating for Country songs against that of R&B songs. This in support of our earlier conclusion from the heatmap analysis; music preferences of Country fans diverge from that of R&B fans.

Interpreting the 2nd Component offers another insight on how we might classify music preferences. There could be a general dichotomy of preferences for **traditional vs. modern** music. Country music has an element of traditional folk style, whereas Rock music might be seen as a prototype for modern music as both Pop and R&B styles have their roots in Rock genre. Further evidence for overlap between Rock and Pop genres has been established in the heatmap analysis.

With this perspective on how music preferences might be classified, we set out to employ classification techniques, using components of PCA as predictors. Our aim is to predict whether a song of an unknown genre would be rated favourably or not.

3. Classification

The goal of this section is to examine if the first two PCA components could be generalized to predict the ratings of songs with an unknown genre. Specifically, classification analysis was employed to predict whether a song with an unknown genre would be rated favorably ("5" or "4") or unfavorably ("1" or "2"). Songs with rating scores of "3"s were excluded as these represent neutrality towards a song.

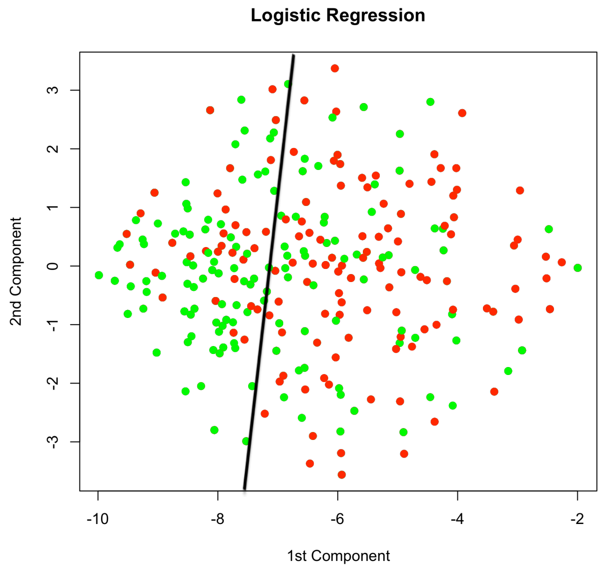
We will compare results from five techniques: logistic regression, linear discriminant analysis (LDA), support vector machine (SVM), quadratic discriminant analysis (QDA) and K-nearest neighbor (KNN). The original dataset was split into two to create a training and test dataset. Supervised learning was conducted on the training dataset, before the classifers were assessed using the test dataset. The values for "K" in KNN, and those of "degree", "sigma" and "cost in SVM were tuned via cross-validation using a leave-one-out process on the training dataset. The parameters that generated the lowest error rate were then used on the test dataset.

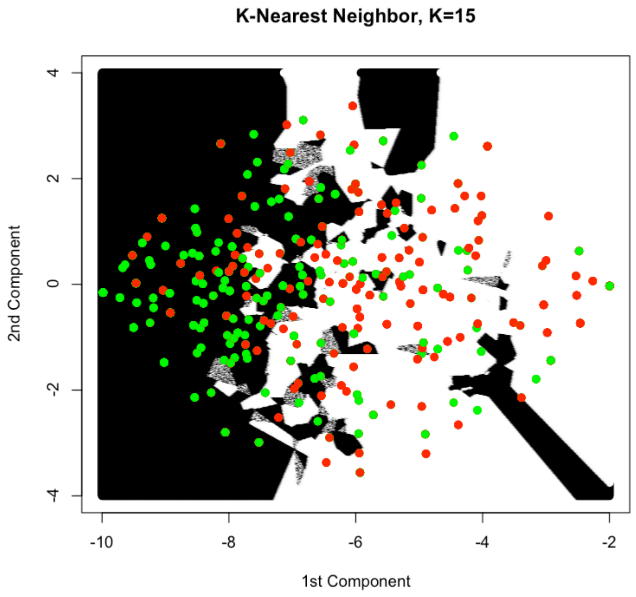
The error rates are as follows:

	Logit Reg	LDA	SVM (deg=1)	QDA	KNN (K=15)
Test Error (%)	29.8	29.8	29.8	33.6	30.0

The baseline error rate seems to be about 30%, implying that the probability of accurately predicting whether a user would like a song of an unknown genre using the first two PCA components is well above chance. Interestingly, the techniques that employed linear boundaries worked best. In addition, having non-linear or multiple boundaries does not seem to improve error rates. Below are the classification plots for songs of unknown genre:

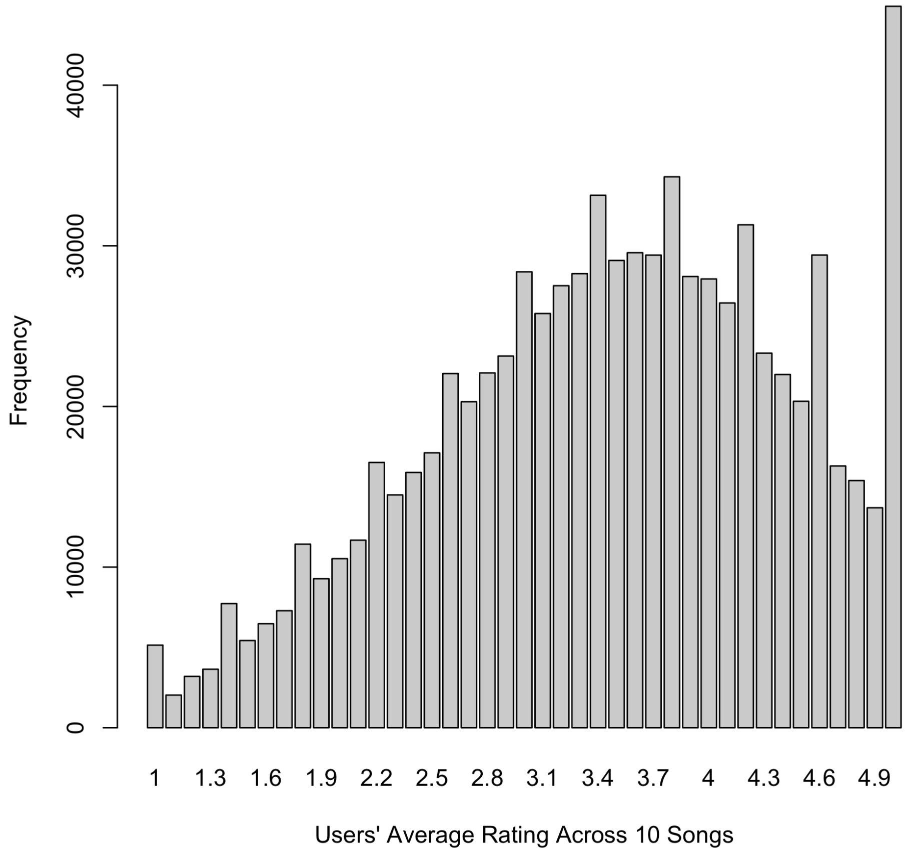
Green represents songs rated **favorably**
Red represents songs rated **unfavorably**.


Logistic Regression


K-Nearest Neighbor, K=15

From the classification plots, notice a general pattern in which the **green** data points are clustered more towards the left relative to the **red** data points. The x-axis represents the 1st PCA component, which is just an average of the mean ratings of songs by genre.

This leads one to suspect the presence of users who only vote when they really like a song and do not vote otherwise, resulting in the strong association between individual ratings and overall average ratings. To verify this hypothesis, we use a histogram to examine users' average ratings across all 10 songs they rated:

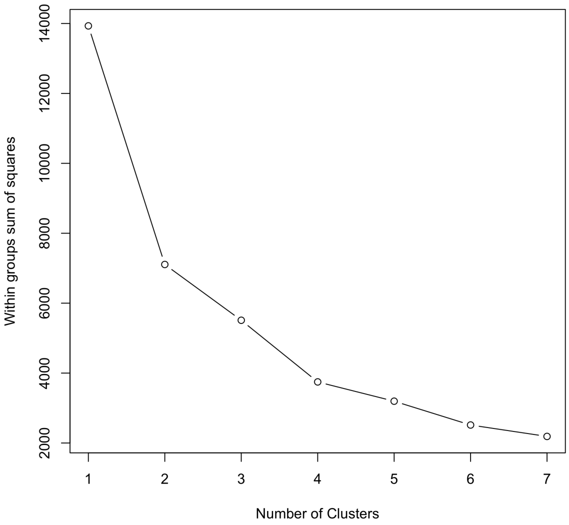


The distribution seems almost normal, except for several spikes due to the response being a discrete variable. More importantly, we observe a huge proportion of users who had a total average rating of "5", implying that they gave all their songs perfect scores, regardless of genre.

This confirms our revised hypothesis, and it presents a new possibility of clustering users based on their rating behaviors.

4. Clustering

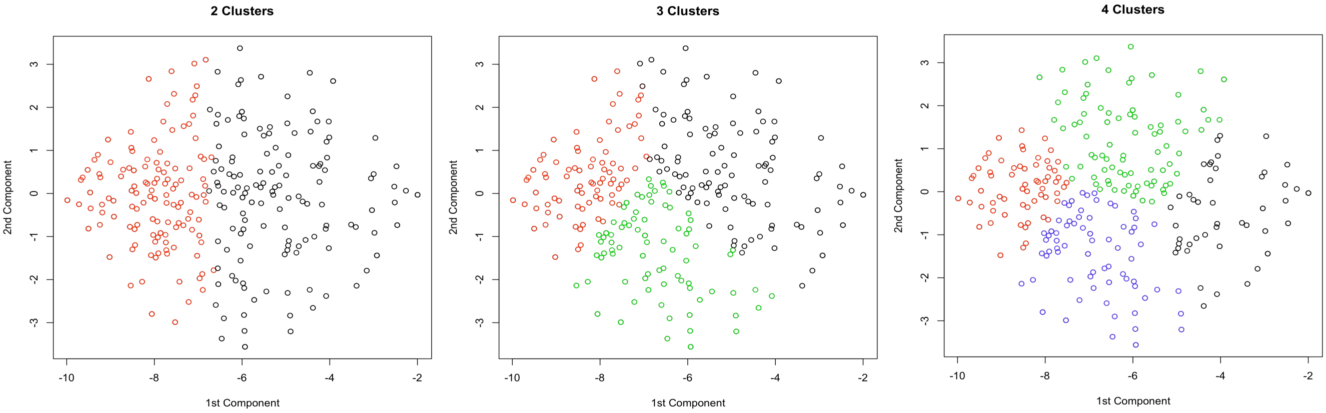
We employed the K-means technique to cluster users based on their rating behavior. As done previously, the first two PCA components were used as predictors. To decide the value of "K", we used a scree plot.



As the number of clusters increase, the within-group sum of squares decreases. We seek to select the right number of clusters that gives a reasonably low intra-cluster scatter. K=4 was chosen, as that is where the decrease in within-group sum of squares starts to flatten out.

In addition to plotting K=4, we also plotted K=3 and K=2. Notice that the left side of the graph, as denoted by the red points, is always grouped as a single cluster, and this phenomenon is invariant to the value of K. On the other hand, the clustering configuration for the right side of the graph varies as K changes.

This suggests that the cluster of users on the left is distinct from the others, leading one to believe that these users possess a unique rating behavior. The cluster is to the left side along the 1st PCA component axis, implying that these users have unusually high average ratings across genres. We might safely conclude that this cluster represents the users found earlier whom only rated songs they extremely liked as "5"s, and did not rate otherwise.



5. Limitations

Most of the existing classification techniques require continuous variables, which led to the use of average ratings in our analyses. The use of average ratings has several shortcomings:

- **Preference for a particular artist** might yield a high average rating for that artist's music genre. For instance, a user might have repeatedly rated a favorite Country artist, skewing the average rating for Country songs upwards, though the high average score might not reflect a general liking for Country songs.

- **A user might have only rated one song per genre**, which does not constitute a representative sample of the user's genre preferences.

- **Users do not rate every song they listen to**, and average rating scores might reflect a user's rating behavior instead of a user's desirability of a song genre. As found in the clustering analysis, a high average rating score might imply that the user only rated a song when it was desirable, and neglected to rate a song that was undesirable.

Concluding Remarks

In the attempt to classify songs as favorable or unfavorable for each user based on genre preferences, we discovered that user ratings might not be a reliable predictor due to differing rating behaviors. Specifically, we found that some users only rated when they liked a song, and hence we would not have information about songs that they dislike or are neutral towards.

In light of this, one needs to obtain more information about each song (e.g. Billboard chart rankings, number of times a user listened to the song's artist) in addition to using average ratings to predict how favorable a song would be for each user.