Capstone Project
Introduction to Data Science, Section 001

Fall Semester, 2023
Group: DS 34
Members: Tanvi Bansal, Natasha Recoder, Tyler Perez

Author Contributions

Each of the 10 questions were owned by one team member and then independently reviewed by another (different) team member. The owner/reviewer designations are shown in Table 0 below. Owners were responsible for making key modeling decisions, developing code implementation, style choices for tables/figures, interpreting results to generate conclusions, and documenting in this report according to the spec sheet. Reviewers were responsible for independently researching topics, challenging assumptions and decisions, reviewing code, and providing feedback to the owner. Data pre-processing was handled according to the guidelines described in the next section and implemented by each owner. chatGPT was used for code syntax reference and debugging assistance throughout the assignment. Additionally, chatGPT was used for strategic guidance on Q8 (neural network); documentation for the recommended packages were reviewed closely for adherence with content reviewed in lecture/lab.

Data Processing

Duplicate removal

The raw data set contained many songs with multiple permutations where one or more features were different; for instance a song present on two different albums by the same artist can have two different entries. We considered such song entries to be duplicates if the artist, track name, and *all* features of interest pertaining to the question were identical. Features of interest are explicitly stated in the body of each question's answer. We made this decision because song permutations caused by differences in features that are not relevant to the question may over-represent the audio factors used in the modeling for these songs and artificially skew the results. The duplicates as defined here were dropped from the dataframe for each question. Additional data processing steps taken for individual questions are stated in their answers.

Correlation

When citing any correlation figures in regards to star ratings and popularity, we used Spearman's Rank Correlation due the lack of a reasonable assumption regarding ordinal data such as Star Rating. Popularity's underlying distribution is unknown, and given the distribution of the data offered it does not look normal.

Random seeding

Random seeding was implemented for all questions regardless of owner/reviewer using Natasha's random seed number (N = 13839901). The random seed was implemented by each owner in their respective code.

Cross-validation

Cross validation was chosen as the method of model evaluation for all questions utilizing an 80/20 train/test split and random state = N (defined in the random seeding section above). We chose cross-validation instead of k-fold validation because of the large computational power required to fold the large data set. In the future we would utilize a big data processor to help achieve k-fold validation with large data sets.

Significance Testing

For questions relating to some kind of effect in popularity, the Mann Whitney test was used. The popularity value is an integer from 0-100 assigned to each song depending on the number of plays, but we do not have evidence that the play count is evenly spaced between intervals of popularity. We determined that it was not reasonable to reduce the popularity data to sample means and chose a Mann Whitney U test to detect effects between groups. Loss of power is generally a concern with opting for the non-parametric test, however we believe the sample size and degrees of freedom are sufficiently large to retain enough power to detect effects.

Question	Owner	Reviewer
1	Tasha	Tyler
2	Tyler	Tasha
3	Tasha	Tanvi
4	Tasha	Tanvi
5	Tanvi	Tyler
6	Tasha	Tyler
7	Tanvi	Tasha
8	Tyler	Tasha
9	Tyler	Tasha
10	Tanvi	Tyler
EC	Tyler	Tanvi
Report Compilation	Tanvi	

Table 0 - Division of labor for author contributions

Questions 1-10

Q1

Duplicates were removed according to the procedure listed above using additional features of duration and popularity.

We calculated the spearman's correlation coefficient between duration and popularity given we do not know how popularity is calculated and what it means. We got a coefficient of -.0881, showing a negative, although small, relationship between song length and popularity with a p-value of 9.49e^-77. Despite the small effect size we have a large amount of data (44,082) and thus due to the small p-value we feel confident in this relationship. The longer, the less popular in general, although only slightly so. In order to confirm this we ran a linear regression which yielded a COD value of 0.00832 (Figure 1). We did not train test split our regression as we are not using it in a predictive manner. We note that this best fit line has negative popularity values beyond a duration of around 2,600,000ms, which is not possible, however we are solely fitting this line to visualize a general positive or negative relationship, not to predict.

Overall we establish that there is a slight negative relationship between duration and popularity with high confidence.

Q2

In this question we sought to figure out if explicit songs tended to be more popular than non explicit songs. We used the Mann Whitney U test in accordance with the assumptions outlined in the introduction. The figures of the two distributions do not look too different, however since we are dealing with large samples the effect is amplified. Therefore we shied away from a T test in favor of a Mann Whitney U Test. Through exploratory analysis we saw the variances between the two dataframes were somewhat similar (447.7 for explicit, and 396.9 for non explicit. As the Mann Whitney U test compares medians we saw the median popularity of an explicit song was 35 and non explicit was 34.

Although the distributions did not seem to be quite different from one another, the large sample size (n=4925 for explicit and n=39127) yielded a quite significant result with a U statistic of 103783931.0 and p value of 9.53e-19. Given our alpha of 0.05, we reject the null and conclude that the difference in median popularities of explicit vs non explicit songs is too significant to happen by chance.

We also calculated a 95% confidence interval for the mean popularity. The CI for explicit songs was [36.72, 37.91] while the interval ranged from [34.17, 34.56] for non explicit songs. To interpret these ranges one can say we are 95% confident that the true mean popularity for explicit and non-explicit songs will fall in these ranges. Neither of these ranges overlap and neither of them are very wide alluding to our parameter being stable given the sampling error. The means were 37.32 and 34.37 for explicit and non explicit songs respectively.

Q3

We dropped duplicates using mode (major vs minor key) and popularity. As explained in the introduction, the Mann Whitney U test was used to compare the median popularity of minor key and major key songs. The median popularity of major songs was 34 while the median popularity of minor songs was 35. To compare if the median of major songs was greater than the median of minor songs with statistically significance we ran the one-tailed Mann Whitney U test which resulted in a U statistic of 224859984, with a p-value of 0.99, thus we are highly confident. Inspecting the two histograms visually we do not see much of a difference, which backs our conclusion and our medians, where the major song median is lower than that of minor songs. (Figure 3a,3b)

Q4

To determine which of the ten specified features predicted the tempo the best we ran simple linear regression models for each feature. We cross validated using an 80/20 train-test split to help account for overfitting. We wanted to ensure each model used the same samples, thus we dropped duplicates that had identical values in each of the 10 feature columns, the popularity column, the artist column and the track name column. The model comparing instrumentalness to popularity had the lowest RMSE of 1845 and highest R^2 value of 0.0404, thus predicted popularity the best, while the other features' models all had RMSE's below 1884. This model had a beta coefficient of -12.13. The feature that predicted popularity the worst was valence with an R^2 of -5.8e-.05, showing an RMSE of 1884 (Table 4). We visualized the data by plotting instrumentalness vs. popularity, along with the regression line to confirm the appearance of a slight negative correlation where increased instrumentalness decreases popularity (Figure 4). Although none of the features predicted popularity well, instrumentalness predicted it the best of the 10 features analyzed. Intuitively this makes sense, as many songs on either end of the spectrum for each feature can be popular and using one feature cannot fully predict how popular a song is, however it is rare for an instrumental song to hit the top charts.

Q5

First we dropped duplicates of the subset of data containing the 10 predictors. We then fit a multiple regression model and cross-validated resulting in $R^2 = 0.0858$ and RMSE = 1824.818. This R^2 value is weak indicating poor predictive ability of our model, however it is an improvement from the R^2 obtained in Q4; this may be a result of overfitting.

To account for this, we fit and cross-validated a LASSO regression model to the same training/test sets over a range of λ values; results shown in Figure 5.1 (appendix). We chose a LASSO regression to enforce sparsity as we want to avoid potential overfitting and remove predictors with little contribution to the explained variance. The optimal regularization strength $\lambda_{optimal} \approx$ 1e-6; $R^2_{optimal} = 0.0858$, $RMSE_{optimal} = 1824.817$. We see that these results are nearly identical to the simple multiple regression (OLS regression) therefore we can conclude that our simple regression model has low risk of overfitting. We conclude that some combination of the 10 song features of interest is an improved predictor of popularity than any of these features alone.

Q6

We dropped duplicates that shared identical values in the 10 columns corresponding to the features, and the columns containing song title and artists. We visualized the correlation matrix between each of the 10 features and observed strong correlations between features 2 and 3, energy and loudness, and strong negative correlations between features 2 and 5, energy and acousticness, as well as 3 and 5, loudness and acousticness (Figure 6.1).

When running the PCA on the correlation matrix we found the biggest drop between the eigenvalues associated with the 3rd and 4th principal component and thus selected the first three principal components. This decision was supported by the Kaiser criterion. These three components accounted for only 57.5% of the variance, which is suboptimal, but is the natural cutoff (Figure 6.2). Using only the data corresponding to these three principal components we ran K-means clustering. First we calculated the silhouette score for around 50 clusters given there are 52 genre's. We did a broad search of silhouette score with cluster numbers increasing in increments of 5 between 30 and 75 (Figure 6.3a). Given we saw

the best silhouette scores on the lower range we searched between 5 and 25 clusters in increments of 5. Again, we saw the lower range had the best silhouette scores, thus ran a more exhaustive search between 2 and 10 clusters, resulting in the best silhouette score at 2 clusters of (Figure 6.3b). This does not match our expect number of clusters of 52, from the genre labels from column 20. This is potentially due to the fact that our three principal components only account for roughly half of the variance.

We see cluster 0 contains 77.77% of the songs and thus over 50% of most genres. Interestingly, cluster 1 contains over 75% of the songs of the classical, ambient, disney and guitar genres as well as over half of the songs of the chill, acoustic and cantopop genre's (Table 6.1a). These are all genres we would consider calmer and quieter. Cluster 1 meanwhile, has 100% of the hardstyle songs, as well as over 99% of drum and bass, happy, hardcore, edm and death-metal songs (Table 6.1b). These songs are likely to be very different from those in cluster 0 as, by intuition about those genres, they are louder and more energetic. Comparing the median of each feature between clusters we see cluster 0 has higher energy and loudness as expected, although we are not making conclusions about the statistical significance of this difference. Further exploration could be done to statistically compare the differences in each feature between the clusters. Although we did not get 52 clusters as expected the 2 clusters did roughly align with overall moods, which allowed certain genres to fit mostly into one cluster or another.

Q7

First we removed duplicates in the subset of data containing the 13 numeric features (duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo) and plotted the predictor space to visualize the inputs for a classification model (Figure 7.1); the plot indicates weak separability of major/minor key of a song from valence. We fit a logistic regression model cross-validated with resulting AUC = 0.500 (Figure 7.2). This weak AUC score aligns with our intuition that valence is a poor predictor of mode as the distributions of valence for each class are nearly identical.

We attempted three different methods to find a better predictor of mode: univariate logistic regression with each of the 13 numeric features, multivariate linear classification (SVM), and multivariate non-linear classification (randomForest). These methods were chosen to explore the effects of using different single predictors/combinations of predictors and linear/non-linear classification strategies, and ultimately find a better predictor of mode. To remove any inconsistencies between models, each model was trained and tested on the same split data set as the univariate regression with valence. The results of each of these methods are shown in Table 7.1 below; the univariate result only contains the metrics from the best fit which came from the predictor 'key' with AUC = 0.592. The non-linear randomForest classifier performed the best with AUC = 0.653; this aligns with our intuition of the classification objective as we expect non-linear multivariate classifiers to perform best on complex, high-dimensional data sets. We conclude that a better predictor of mode is all 13 numeric song features with a non-linear randomForest classifier.

Q8

To create a neural network predicting track genre, we decided to follow our 10 selected features from question 4; duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, tempo, and valence.

We also standardized each of our 10 features in a gaussian manner to ensure best practice. As many of our features have different ranges, and we do not want our model to be dominated by ones that have large values. Such an example would be in the scale for track duration. In addition to standardization

we employed a train/test split. As is standard we used the Relu activation algorithm to begin our neural network. We also went through 1,000 iterations to try to get the best accuracy while not hinging too much on computational intensity.

The accuracy results yielded by our MLPC Artificial Neural Network were not astounding by any means. Our prediction accuracy was just 30%. Some genres performed much better than others, Comedy (88%) and Grindcore (64%) turned out to be the easiest genres to predict. The only other genres that were predicted correctly more often than not were classical at 51% and Detroit Techno at 57% . The hardest genres to predict were alternative (0%), british (11%), country (11%), and groove (10%). The difficulties arising from this task make some sense given our limited features and the imperfections in our dependent variable. Many of these musical genres are sub genres of others listed in this list such as Black Metal and Heavy Metal or grunge and alt rock. Lastly, the human job of classifying musical genre contains much ambiguity itself as what differentiates one genre from another is not always clear and separable. Lastly the method of determining genre in the spotify dataset was not revealed to us.

In Table 8 you can see each genre with their precision, recall, fl-score, and support metrics. Additionally, in Figure 9 you can see our Confusion Matrix, which depicts the troubles our model had predicting correct genre classifications.

Additionally to further explore if we could predict track_genre we used the PCA from question 6 and then ran another Neural Network in the same exact aforementioned fashion, but now with our three principal components. This yielded even more disappointing results of 14% accuracy. Many genres possessed 0% accuracy, and the highest was comedy at 84%. Once again corroborating our conclusion that predicting track_genre is quite a challenging expedition.

Q9

In order to tackle both the popularity based model and whether there exists a relationship between popularity and average star rating we first had to deal with missing data. The vast majority of our cells in the star rating data were NaN. We chose to handle these by imputing the missing cells with an average of both the row and column mean. This seemed to be apt and appropriate because in this we take into account both the user's rating habits as well as the average rating for the song itself.

We then calculated the correlation between the popularity column with the average star rating. This yielded a Spearman correlation of .494. This signals that there is a clear yet not super strong positive relationship between popularity and average star rating. Additionally we calculated an Ordinary Least Square Regression model predicting the average star rating based on the popularity of the song. We did a test train split in order to cross-validate and assist in scalability. Through a graphical analysis one can see that this does a pretty poor job at predicting, especially given the fact that popularity is not normally distributed, but does seem linearly separable. Our RMSE was .334 and the p value related to our t-statistic for popularity was indeed significant below 0.05. The COD that was outputted by our model was .262. The plot as well as the model are given in the figures section.

This in all qualifies as a popularity based model as we then sorted the 5,000 songs with the highest average rating. On the basis of our popularity model our 10 greatest hits were: Red Hot Chili Peppers- Can't Stop (2.96), The Offspring - You're Going to Go Far Kid (2.96), The Neighbourhood - Sweater Weather (2.95), Gorillaz, Tame Impala, Bottie Brown - New Gold (2.95), Walk The Moon - Shut Up and Dance (2.95), Linkin Park - Numb (2.92), Nirvana - Smells Like Teen Spirit (2.92), The Offspring

- The Kids Aren't Alright (2.91), System of a Down - Chop Suey (2.91), System of a Down- Toxicity (2.91).

Therefore in summation we say there is a relationship between popularity and average star rating, however we cannot predict either of them from one another in a sound way without taking into account confounders. The 10 "Greatest Hits" differed from the ones that possessed the highest popularity, as the average popularity for the Top 10 was 83.2. As mentioned previously our model is not perfect due to our imputation of missing data, yet it seemed superior to the alternative of leaving data out.

Q10

For this recommender system we chose a user-based collaborative filtering approach for two reasons: 1) the user-item rating matrix is sparse making latent factor modeling and matrix factorization difficult, and 2) we have a large enough sample of users that similarity between users can be estimated. First we removed any songs that had fewer than 5 ratings to limit noise, then centered each user's ratings to remove user-specific bias. Then for each user of interest we predicted the ratings of all songs by computing a weighted average of ratings from a set of other users that are most similar to the user of interest. The similar group was determined by selecting the top 1% of users with the smallest pairwise euclidean distance to the user of interest; we used row-wise missing data removal for each user pair, and euclidean distance was chosen as the similarity metric because we accounted for "tough rater" bias by centering the data set. The predicted song ratings were determined by taking a weighted average of the similar group's ratings weighted by their similarity to the user of interest; similarity is defined as the inverse of the euclidean distance. The 10 songs with the highest predicted rating were selected for the mixtape for the user of interest.

The 10 most common songs in the 10k mixtapes had 7 songs in common with the greatest hits identified in question 9. To compute precision and recall for each user's mixtape we set the threshold to the 75th percentile of the user's actual ratings and considered all songs with actual rating above the threshold to be relevant; results are shown in Table 10.1. However, these results are expected to have a large margin of error because of the sparsity of the user rating dataset; many of the recommended songs had no actual rating recorded and were therefore dropped from the precision/recall estimates. We found that on average each user had rated 1.56 songs in their mixtape and 2,201 users did not rate any songs in their mixtape (therefore no precision/recall estimates). To account for this, we adjusted the recommender system to select the top 10 songs with the highest predicted rating *and* an explicit feedback response; the similarity group was expanded to the top 10% of users to counter lost signal from the removal of new songs from the target list; results are shown in Table 10.1.

The system performs measurably well when recommending songs that a user has already listened to, however we have lowered confidence in the performance metrics when the system includes recommendations for new songs. In the future, using implicit feedback methods (i.e. listen time, clicks, abandonment) or experimental methods (i.e. A/B testing) may be helpful to evaluate the system with new songs. Relying exclusively on the system using explicit feedback methods introduces the cold-start problem and new items will never be represented. While implicit feedback is generally weaker, complementing the explicit data with the plentiful implicit data could help to alleviate our sparsity problem while preserving signal.

EC Does Song Title Length relate to Popularity?

For the extra credit, we explored whether short titled songs are more popular than long titled songs. We firstly did a median split on the song title length by calculating the amount of characters in each song. With doing a median split we lose some confidence in our findings, for instance a song with 14 letters is close to a song with 16 letters but would be bucketed in two categories. This does present issues for edge cases. We were left with one dataframe of songs with more than 15 characters, and another with one 15 or less. The variances within popularity of the two dataframes were quite similar (411.6 and 388.6). Therefore we follow the homogeneity of variance assumption which is needed for such a statistical test. Due to the lack of knowledge of underlying normality in our data and the inadequacy in reducing our data to sample means we came to a concussion that the non parametric test was the most adequate route.

The means and medians of the data set point to short titled songs being more popular than long songs. The median and mean popularity for short were 36 and 36.2 while longer titled songs yielded values of 31.0 and 33.2. We then performed the nonparametric test which yielded a p-value of 1.76-e-60, and U statistic of 21690281.5. Our test presented a Degrees of freedom of 43814. This is a high amount, presenting some stability in our calculations. The large amount of data however may have amplified the statistical significance of our results. We then concluded that the data in favor of shorter titled songs being more popular was too strong to happen by chance and in turn reject the null hypothesis (a=0.05). We plotted both distributions as well as their box and whisker plots. (Figure EC). As seen in our box plots there were quite a few outliers in the long titled data frame, to corroborate our results we did another MU test disregarding outliers and in turn yielded the same significant results.

Appendix I - Figures

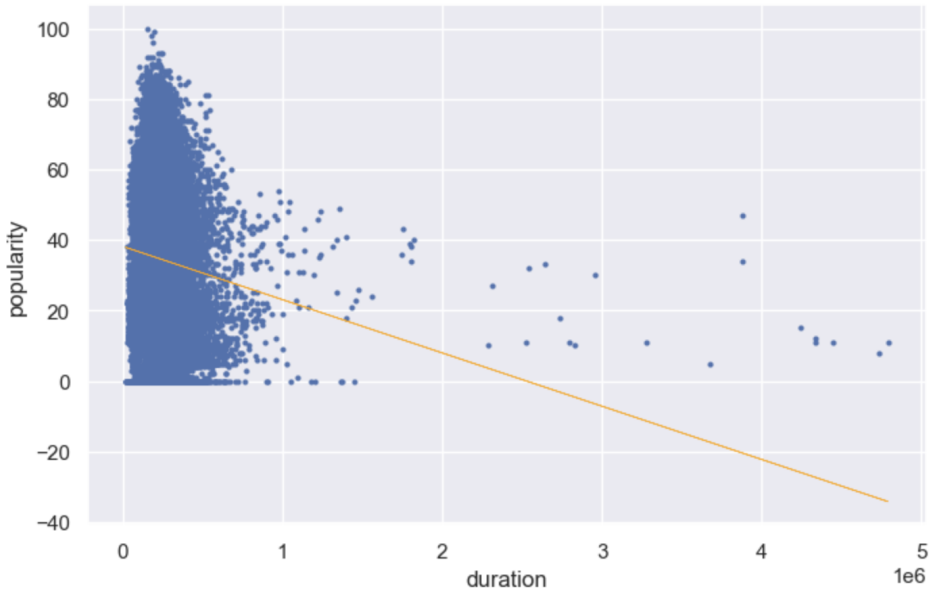

Figure 1- Linear regression and polynomial regression of duration vs popularity. Regression line used to visualize relationship between duration and popularity, not to predict. Slope of regression line = -.0912.

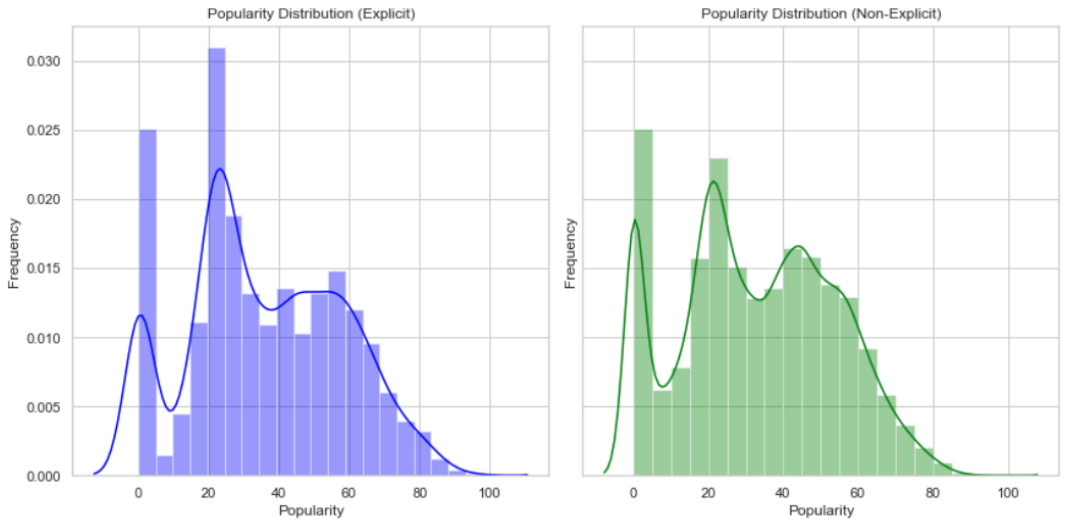

Figure 2- Popularity Distributions of Explicit and Non Explicit Songs

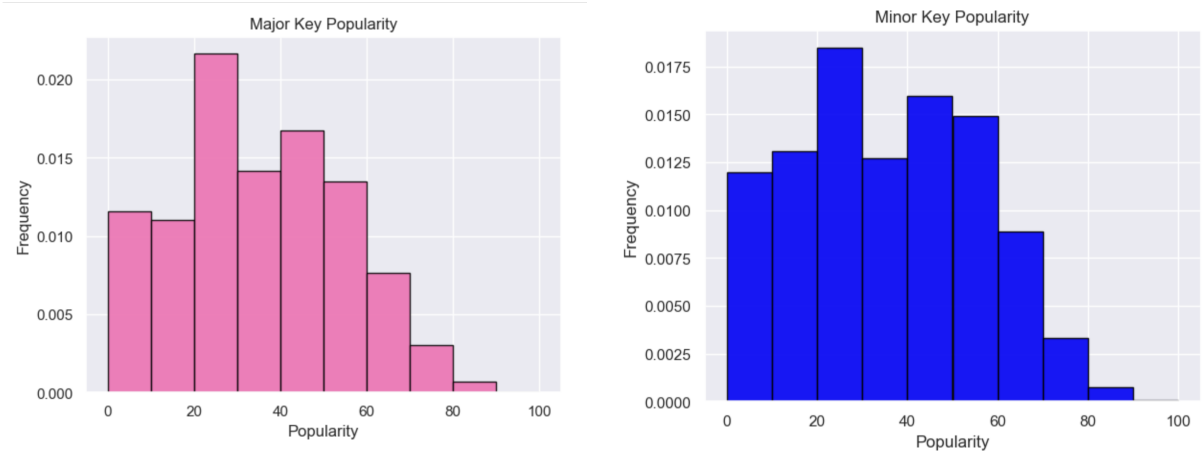

Figure 3- Histograms of popularity distributions of songs in Major key (left) and songs in minor key (right)

	duration	danceability	energy	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo
r2	0.008194	0.005232	0.007888	0.004085	0.004991	0.001388	0.040420	0.002726	-0.000058	0.001265
RMSE	1875.912690	1878.711610	1876.201670	1879.793984	1878.938891	1882.338423	1845.184989	1881.076554	1883.699907	1882.453942

Table 4 - R squared, RMSE, Beta coefficients and alpha values used for 10 ridge regressions regression run between each listed feature and popularity.

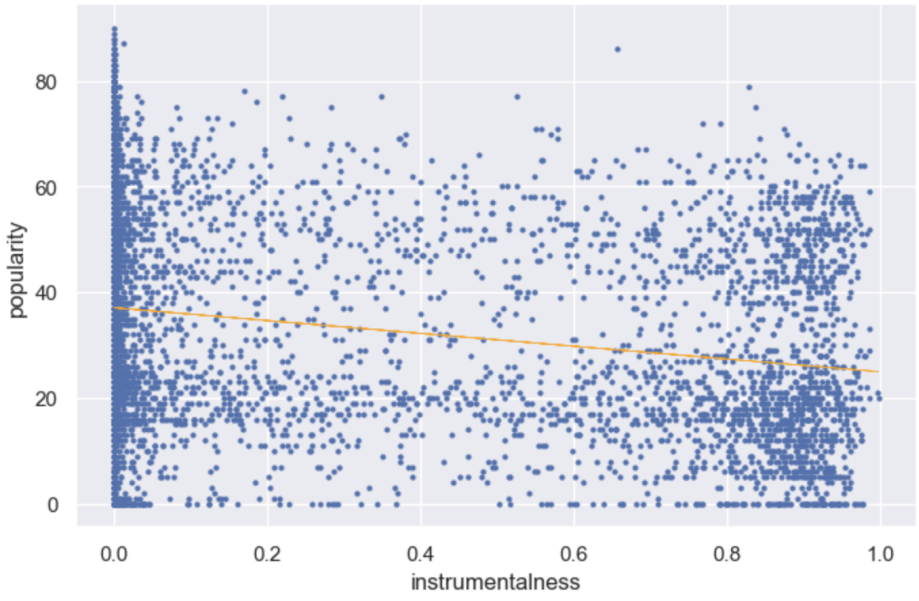

Figure 4- Linear regression of instrumentalness vs popularity. RMSE = 1845, COD = 0.0403

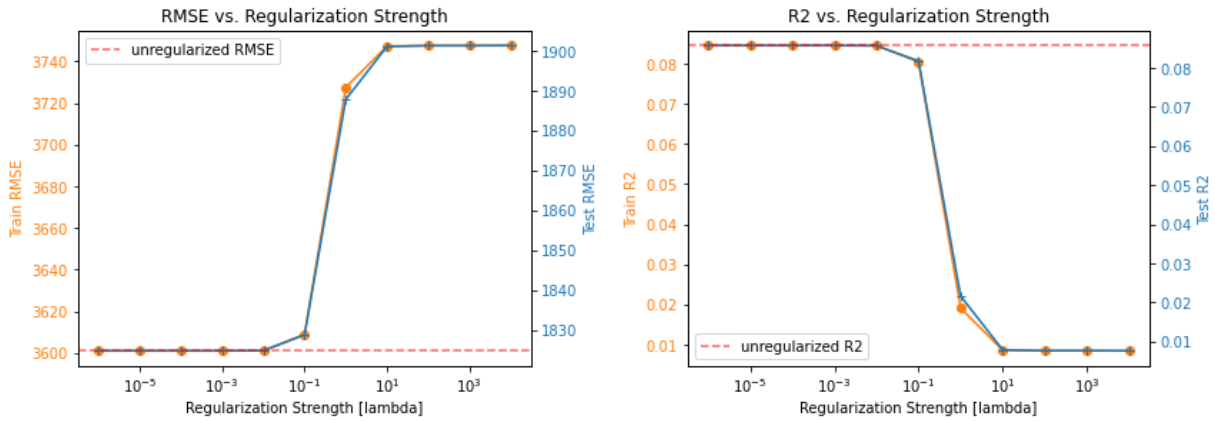

Figure 5.1 - *LASSO multiple regression RMSE as a function of* $\boldsymbol{\lambda}$ *(left) and* R^2 *as a function of* $\boldsymbol{\lambda}$ *(right)*

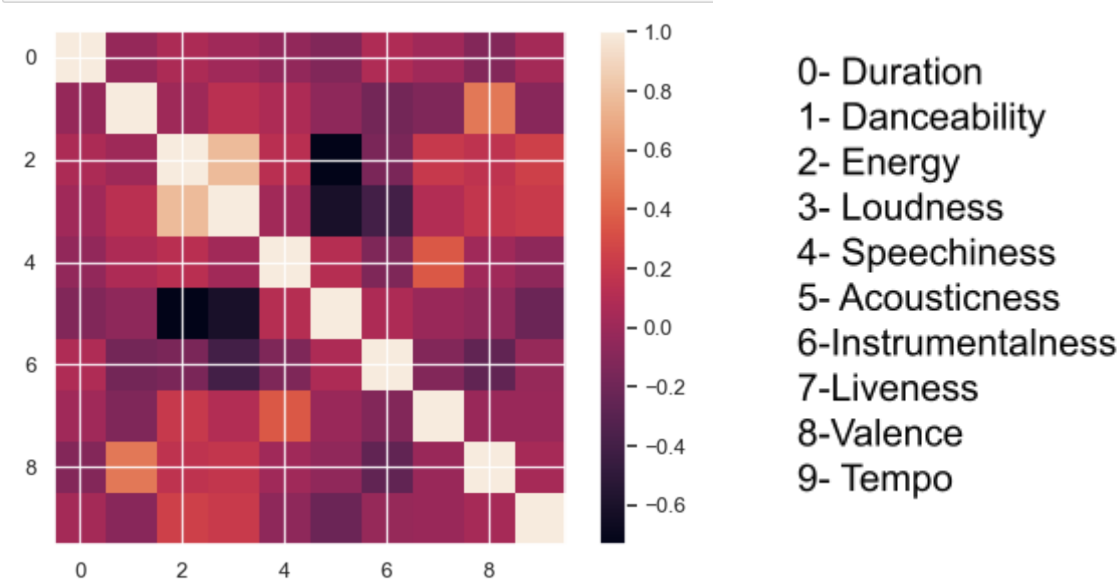

Figure 6.1- Correlation Matrix between features

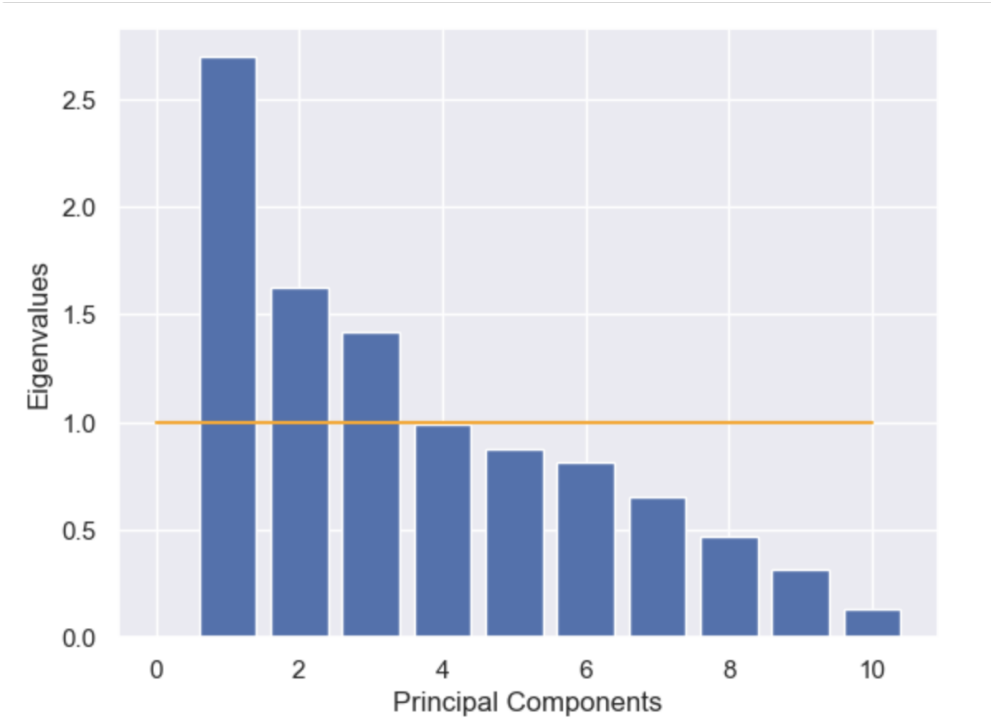

Figure 6.2 - Eigenvalue magnitude corresponding to each principal component. Orange bar signifies Kaiser criterion for selecting principal components.

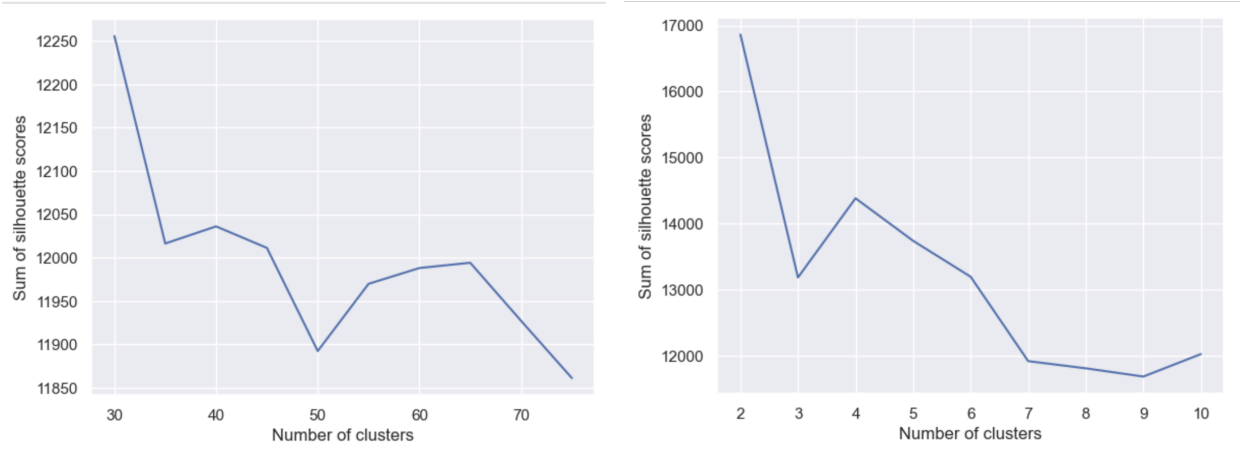

Figure 6.3 - Sum of silhouette score for initial cluster number search (a) and final cluster number search (b)

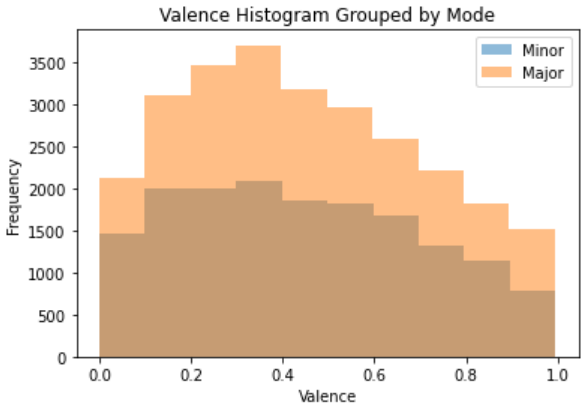

Figure 7.1 - Distribution of valence grouped by mode

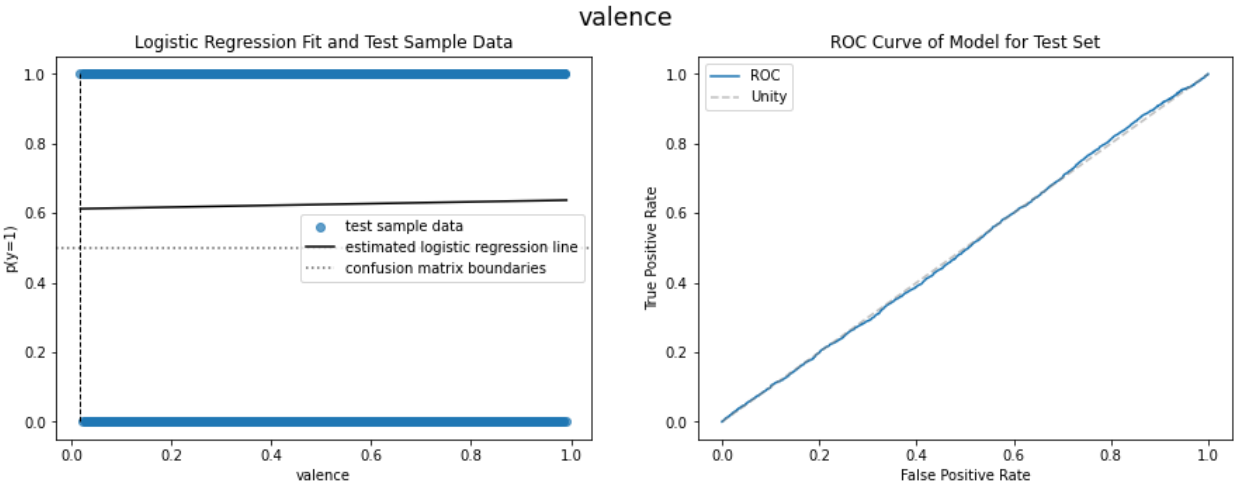

Figure 7.2 - Logistic regression of mode from valence

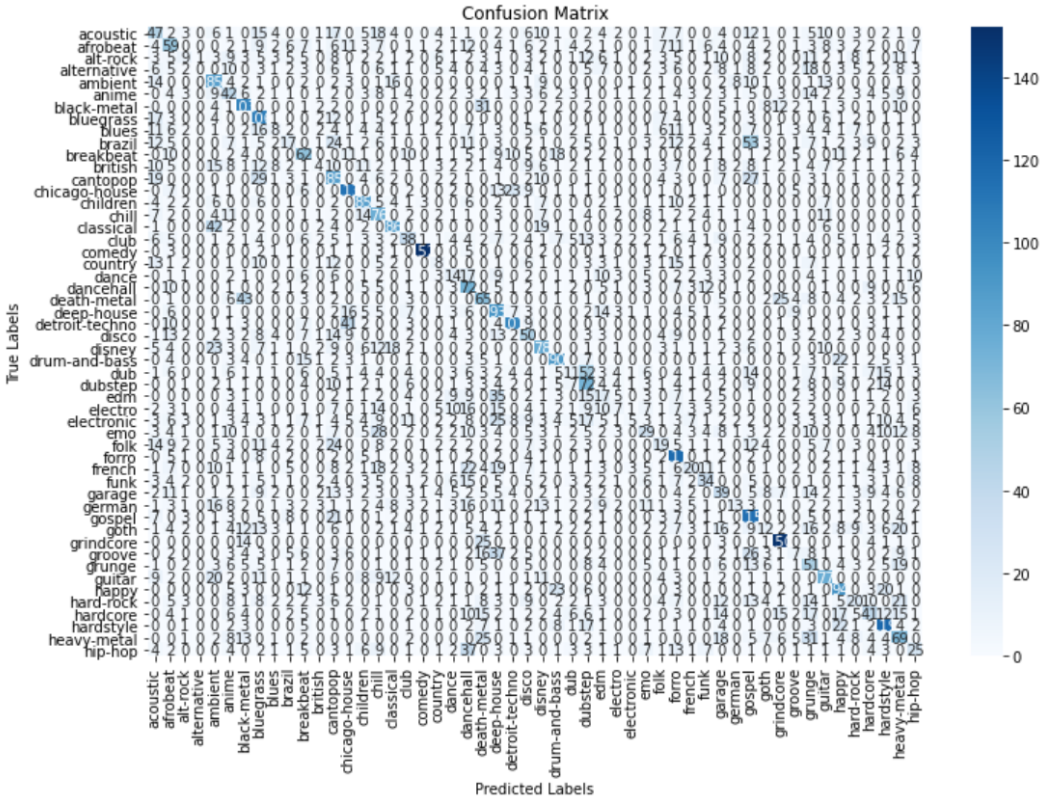

Figure 8- Confusion Matrix of Neural Network to Predict Track Genre

Figure 9- Least Squares Fitting Popularity onto Average Star Rating

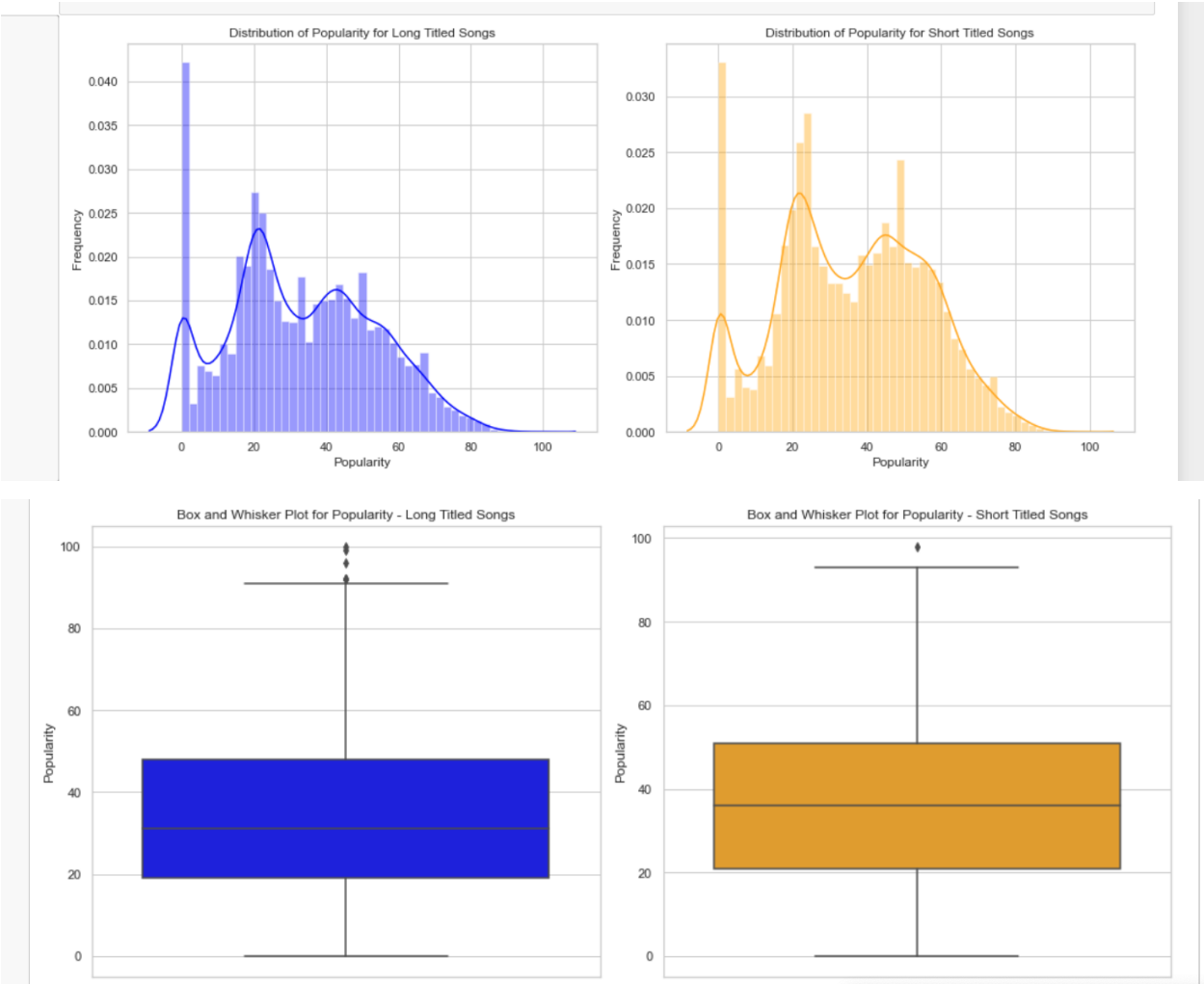

Figure EC- Distributions of Popularity for Long Titled (Left)/Short Titled Songs (Right) and Box Plots for Both

Appendix II - Tables

cluster	0	1
track_genre		
classical	7.894737	92.105263
ambient	10.395010	89.604990
disney	20.102041	79.897959
guitar	21.729238	78.270762
chill	42.361863	57.638137
acoustic	42.752868	57.247132
cantopop	47.076613	52.923387

cluster	0	1
track_genre		
hardstyle	100.000000	0.000000
drum-and-bass	99.572193	0.427807
happy	99.383350	0.616650
hardcore	99.332443	0.667557
edm	99.270073	0.729927
death-metal	99.103139	0.896861
breakbeat	98.977505	1.022495
forro	98.478702	1.521298
heavy-metal	98.286290	1.713710
grindcore	98.170732	1.829268

Table 6.1a,b - Percentage of how much each genre is represented in clusters 0 and 1. 7 genres with the highest percentage in cluster 1 (a) and 10 genres with the highest percentage in cluster 0 (b)

cluster	duration	danceability	energy	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo
0	218997.0	0.588	0.7950	-5.958	0.0579	0.0416	0.000127	0.144	0.473	125.966
1	197682.5	0.519	0.3105	-13.076	0.0401	0.8090	0.009360	0.114	0.291	109.439

Table 6.2 - median value of each feature in each genre

Classification Model Type	**Predictor(s)**	**Model Evaluation**
Univariate Logistic Regression	key	AUC = 0.594
Multivariate Linear Classification (SVM)	Popularity, key, time_signature,duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo	AUC = 0.586
Multivariate Non-linear Classification	Popularity, key, time_signature,duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo	AUC = 0.668

Table 7.1 - Comparison of classification models predicting mode (major/minor key) with selected predictors

	precision	recall	f1-score	support
acoustic	0.20	0.22	0.21	210
afrobeat	0.24	0.27	0.25	218
alt-rock	0.17	0.05	0.08	170
alternative	0.00	0.00	0.00	155
ambient	0.31	0.47	0.38	179
anime	0.19	0.23	0.21	180
black-metal	0.41	0.52	0.46	197
bluegrass	0.32	0.57	0.41	187
blues	0.14	0.05	0.08	149
brazil	0.21	0.08	0.11	218
breakbeat	0.32	0.32	0.32	192
british	0.11	0.02	0.04	175
cantopop	0.22	0.39	0.28	218
chicago-house	0.48	0.59	0.53	191
children	0.41	0.52	0.46	164
chill	0.25	0.44	0.32	173
classical	0.51	0.48	0.49	178
club	0.28	0.20	0.23	186
comedy	0.88	0.83	0.86	183
country	0.11	0.07	0.09	110
dance	0.14	0.12	0.13	117
dancehall	0.22	0.44	0.29	164
death-metal	0.27	0.33	0.30	195
deep-house	0.26	0.47	0.33	199
detroit-techno	0.57	0.52	0.55	193
disco	0.23	0.28	0.26	176
disney	0.36	0.38	0.37	204
drum-and-bass	0.47	0.49	0.48	182
dub	0.15	0.05	0.08	214
dubstep	0.24	0.38	0.29	191
edm	0.12	0.12	0.12	143
electro	0.15	0.05	0.07	153
electronic	0.29	0.02	0.04	212
emo	0.23	0.14	0.17	207
folk	0.18	0.11	0.14	171
forro	0.34	0.69	0.45	170
french	0.24	0.10	0.14	194
funk	0.28	0.23	0.25	146
garage	0.17	0.20	0.18	194
german	0.34	0.07	0.12	174
gospel	0.28	0.57	0.38	200
goth	0.16	0.06	0.09	203
grindcore	0.64	0.75	0.69	200
groove	0.10	0.04	0.06	169
grunge	0.17	0.29	0.21	177
guitar	0.40	0.41	0.41	186
happy	0.39	0.51	0.44	185
hard-rock	0.18	0.11	0.13	188
hardcore	0.28	0.18	0.22	223
hardstyle	0.45	0.56	0.49	214
heavy-metal	0.26	0.31	0.28	222
hip-hop	0.20	0.16	0.18	160
accuracy			0.30	9559
macro avg	0.28	0.30	0.27	9559
weighted avg	0.28	0.30	0.28	9559

Table 9.1- Classification Metrics for Each Musical Genre

```
t-statistic for 'popularity': 16.44547005168054
P-value for 'popularity': 0.0
Mean Squared Error: 0.11808475457995235
Regression Coefficients:
Intercept: 1.6321768898614974
Coefficient for 'popularity': 0.009438527351773666
Coefficient of Determination (COD): 0.2624643906230286
```

Table 9.2- Least Squares Model Fitting Popularity onto Average Star Rating

	Recommender System with New Songs	**Recommender System without New Songs**

Mean Average Precision	0.952	0.972
Mean Recall	0.011	0.056
Top 10 most recommended songs (in descending order)	Sweater Weather Californication Shut Up and Dance You're Gonna Go Far, Kid Bring Me To Life New Gold (feat. Tame Impala and Bootie Brown) Can't Stop The Kids Aren't Alright Awake and Alive Unsainted	Sweater Weather New Gold (feat. Tame Impala and Bootie Brown) Bring Me To Life Shut Up and Dance Californication You're Gonna Go Far, Kid Can't Stop Numb Losing My Religion Toxicity
# Songs in Intersection with Greatest Hits	6	7

Table 10.1 - Performance metrics for the personalized mixtape recommender system