Mariam Abdullah (ma3259)
Will Calandra (wcc9105)
Musonda Sinkala (mks9887)

Title Page/Summary

Group Name: DS 85
Group Members: Mariam Abdullah, Musonda Sinkala, Will Calandra
Member Contributions: Mariam completed questions 1-3, 5-6, and did the write-up for these questions. Musonda completed questions 4, 7, 8, and did the write-up for these questions. Will completed questions 9-10, the extra credit, and did the write-up for these questions.
Preprocessing:

- On Spotify song data: For questions involving parametric statistical significance testing against popularity, we removed songs with a popularity of 0 to achieve normality. For questions involving fitting a neural network, we standardized our features and label encoded our targets to prevent scaling issues.
- On Star rating data: For questions 9 and 10, preprocessing of the star ratings dataset required an aggregation of user ratings for duplicate songs. We defined duplicate songs as duplicates of track name and song features. While there may be differences in context for each song (ex: live performances) that would lead to different user ratings and distinct songs in our data, there were ~90% of ratings missing across both each user and each song. In this way, taking the mean (aggregation) of user ratings for duplicate songs would help reduce the sparsity of our matrix.

Citations: ChatGPT helped direct us to the Surprise package for building our recommendation system. We asked for a simple framework to start, read through the documentation, and paired it with our prior experience in scikit-learn and GridSearchCV. We chose the number 7 as our seed for reproducibility.

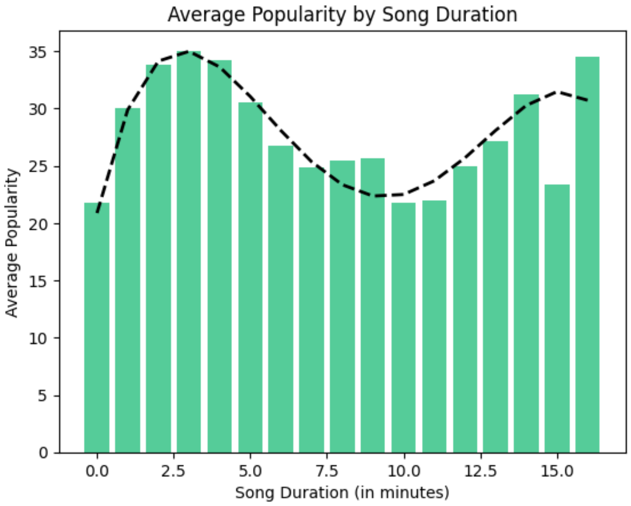
DS-GA 1001 Capstone Project - Group 85

Question 1

Is there a relationship between song length and popularity of a song? If so, is it positive or negative?

Length and popularity are correlated at **-0.057** in the provided dataset, so these variables have a ***slightly*** **negative** linear relationship.

However, correlation only tells us if we can fit a linear function, not any other function or even if piecewise these might be related. To investigate further, we bucketed song lengths by minutes and looked at the average popularity per bucket, given that at least 10 songs fell into that bucket (to reduce variance).



The bar plot went up, down, up, and then back down, so we fit a degree 4 polynomial to it. This shows the relationship is positive until a song is about 3 minutes long, then wobbles, then is slightly negative as the song gets longer.

Question 2

Are explicitly rated songs more popular than songs that are not explicit?

While we want to compare the means of two distribution, they do not appear to be normally distributed due to the high number of songs with popularity 0. Ablating these records, an assumption of normality seems more reasonable.10.93% of explicit songs and 12.42% of others have popularity 0, so this conditioning will slightly improve the mean of explicit songs.

The variances of these distributions are 396.63 and 359.42, which are close but different, so we found p-values for both Welch and independent-T tests. Both are << α=0.05 ∴ stat sig.

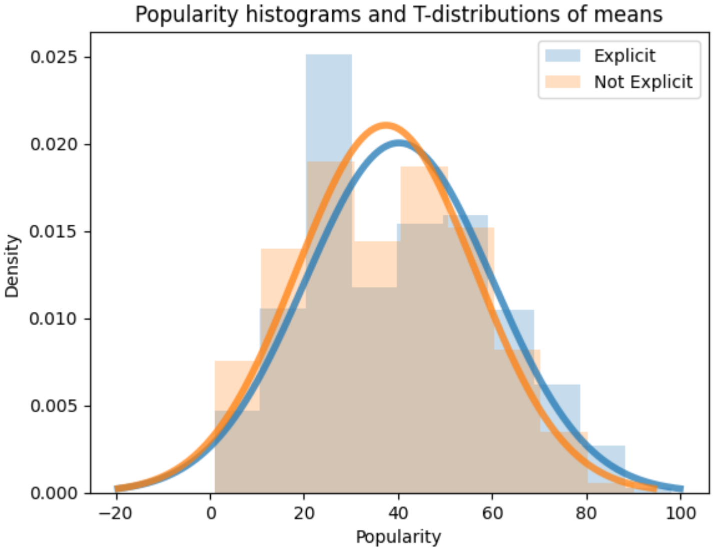
- Independent-T: p = **3.76×10^{-22}**
- Welch: p = **1.63×10^{-20}**

We fitted T-distributions to the mean of each dataset for further inspection.

T-Distribution for mean of	**Mean**	**Stdev**

Explicit	40.21	19.91
Non-Explicit	37.44	18.96

We conclude that explicit songs are **more popular** among songs with popularity > 0. Plotting the PDFs of these T-distributions with (binned) popularity densities shows this by eyeball.



Question 3

Are songs in major key more popular than songs in minor key?

We again need to compare means of two distributions with unequal variances; again, we have reason to believe these distributions are only normally distributed after removing songs with popularity 0. 12.94% of songs in major key and 11.12% of songs in minor key have popularity 0, so this conditioning will slightly improve the mean of songs in major key.

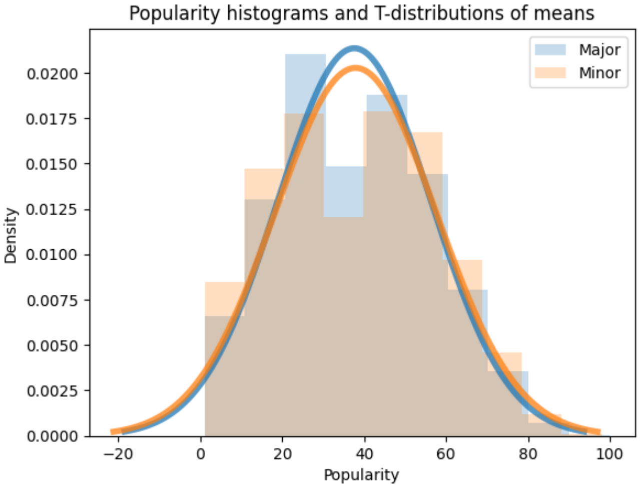
The variances of these distributions are 349.57 and 387.91, which are again close but different, so we found p-values for both Welch and independent-T tests. Both are > α=0.05, so not stat sig.

- Independent-T: p = **0.1185**
- Welch: p = **0.1129**

We fitted T-distributions to the mean of each dataset for further inspection.

T-Distribution for mean of	**Mean**	**Stdev**
Major key	37.63	18.70
Minor key	37.92	19.69

We conclude that **we lack evidence** to say songs in major keys are more popular. Plotting the PDFs of these T-distributions with (binned) popularity densities shows this by eyeball.



Question 4

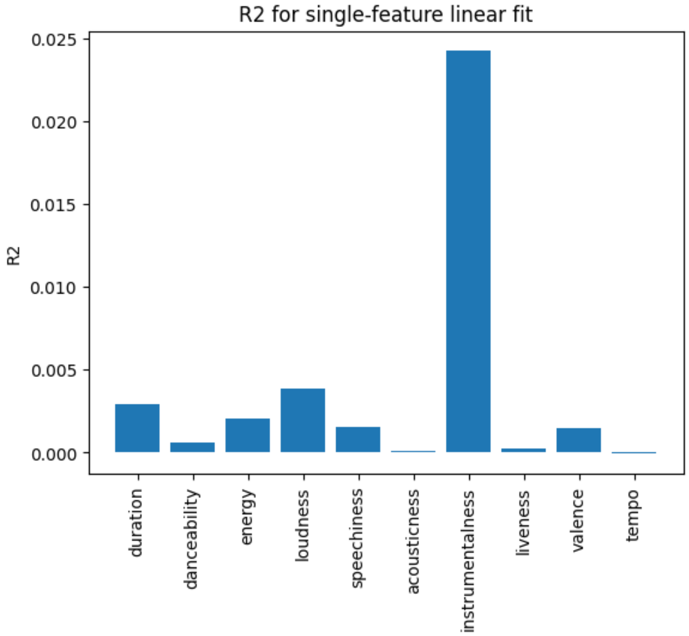
Which of the following 10 song features: [...] predicts popularity best? How good is this model?

To determine which of the selected 10 song features best predicts popularity, we performed **simple linear regression** for each of the 10 song features independently, assuming a linear relationship between the individual features and the target variable (popularity). Simple linear regression was used as it provides a quick and interpretable measure (R-squared) of how well each individual feature predicts the target variable - thus identifying which feature has the strongest linear association with popularity. The table below summaries the performance of each of the models that was built. From the table below it is clear to see that **Instrumentalness best predicts popularity** (relative to the other features) as this model boasts an **R-squared** of **2.4%**. Thus, we conclude that instrumentalness has the strongest linear association with popularity among the 10 song features. However, the overall predictive power of these individual features is relatively low, with R-squared values generally close to zero, suggesting that other factors not considered in this analysis may contribute to popularity

Mariam Abdullah (ma3259)
Will Calandra (wcc9105)
Musonda Sinkala (mks9887)

Feature	R-squared
Duration	0.289%
Danceability	0.06%
Energy	0.20%
Loudness	0.39%
Speechiness	0.15%
Acousticness	0.01%
Instrumentalness	2.42%
Liveness	0.03%
Valence	0.15%
Tempo	0.01%

In question 1 we found that a 4th degree polynomial fits the relationship between duration and popularity better than correlation, so we might be able to improve this. Without applying convoluted/overly complex functions, we found that we can achieve an **R^2** of **0.0286** by regressing instrumentalness against popularity squared instead.

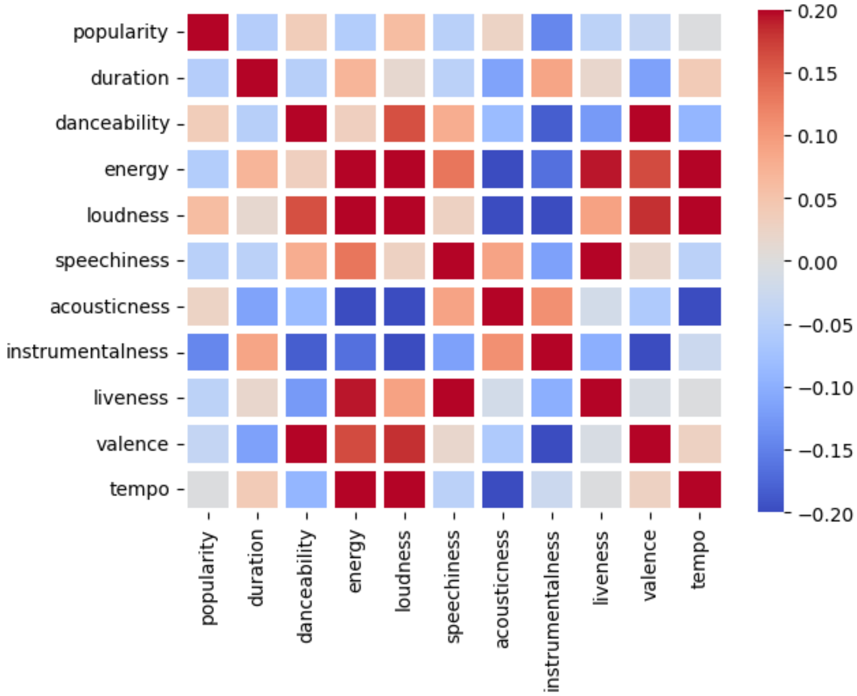


Mariam Abdullah (ma3259)
Will Calandra (wcc9105)
Musonda Sinkala (mks9887)

Question 5

*Building a model that uses *all* of the song features mentioned in question 4, how well can you predict popularity? How much (if at all) is this model improved compared to the model in question 4? How do you account for this? What happens if you regularize your model?*

We improved **R^2** from **0.0242** to **0.0470** by using all 10 features, again employing an 80-20 train-test split. This is because features beyond just instrumentalness have some correlation with popularity, but do not suffer from multicollinearity, so using more helps. However, this model is still lacking since it only explains < 5% of the variance.



Regularizing our data did not really help. We tried both lasso and ridge regression with 15 values of α from 0.0001 to 5, after normalizing our data, and resulting the **R^2** were:

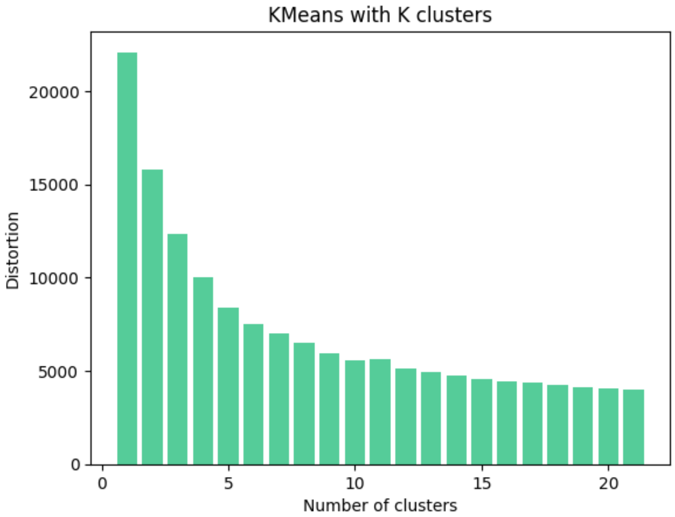
Method	**R^2**
OLS	0.046972
Lasso	0.046972
Ridge	0.046974

In the end, ridge regression helped ever so slightly. However, regressing against the square of popularity like in question 4 yielded an **R^2** of **0.060176**.

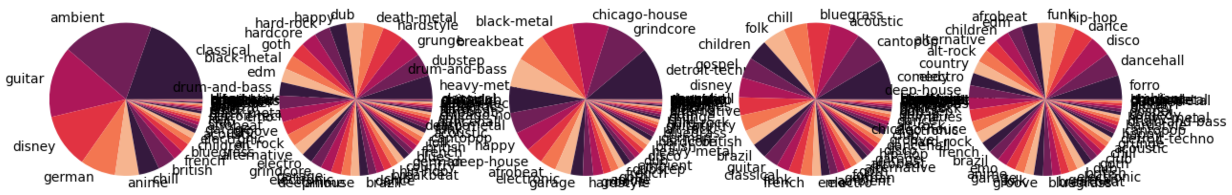
Question 6

When considering the 10 song features in the previous question, how many meaningful principal components can you extract? What proportion of the variance do these principal components account for? Using these principal components, how many clusters can you identify? Do these clusters reasonably correspond to the genre labels in column 20 of the data?

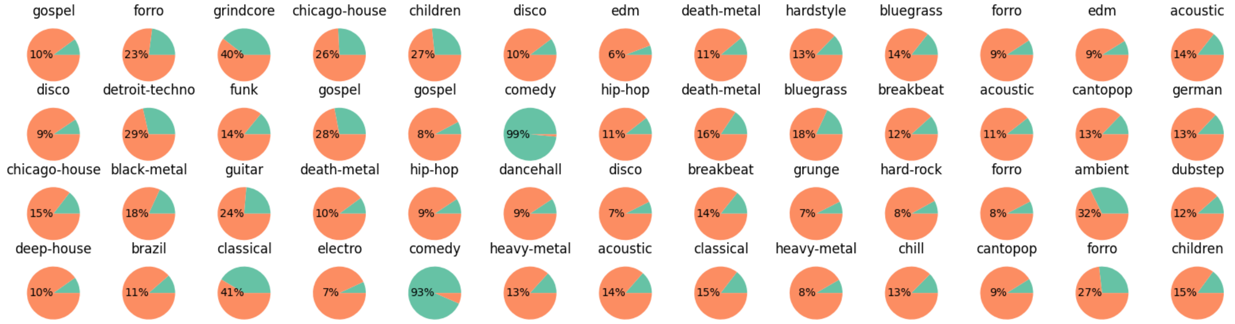
We found that **6** principal components explain **93.76%** of the variance (7 explain 96.59%). Next, we examined the distortion on up to 20 clusters.



The elbow appears around **5 clusters** via eyeball method (see fig above). Accordingly, we labeled the data with which cluster it fell into, then looked at the distribution of genres per cluster. It is resoundingly clear that **clusters do not categorize genres well**, since many genres are split across many clusters.



However, given the scale, 52 genres corresponding well to only 5 clusters seems unlikely. Attempting a better fit, we grouped into 52 clusters and looked at the maximum genre percent per cluster. This works well for comedy and perhaps grindcore, but most clusters only had their top genre at around a 10-15% share.

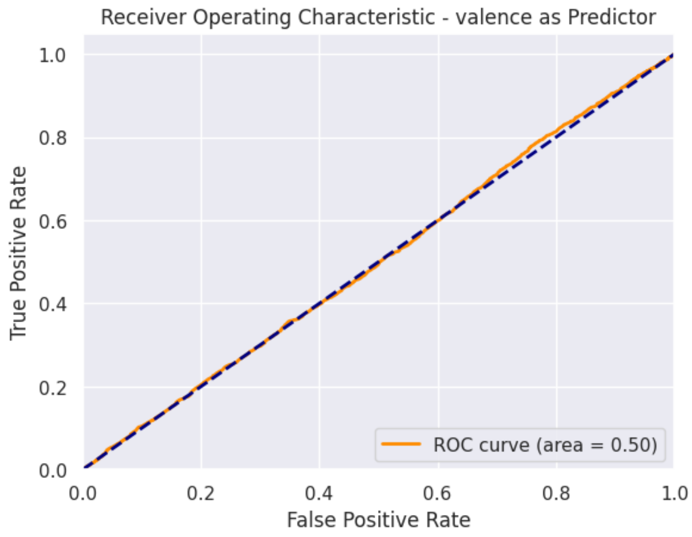


Question 7

Can you predict whether a song is in major or minor key from valence using logistic regression or a support vector machine? If so, how good is this prediction? If not, is there a better one?

To answer this question, we employed **logistic regression** and receiver operating characteristic (ROC) curve analysis for valence to predict the target variable 'mode,' which implies binary classification. We then **did the same for 9 other song features** for the purpose of comparing the results against that of valence. We assumed independence of observations and a linear relationship between features and the log-odds of the target. Given that mode is a binary categorical variable, logistic regression was used as it is suitable for binary classification problems, and ROC analysis provides a means of evaluating classification performance by way of assessing the area under the ROC curve (AUROC) for each feature.' The graph below shows the ROC curve for Valence as a predictor. The ROC being the unity line indicates poor classification performance - which is confirmed by an AUC value of 0.5 for the model. The table that follows indicates the AUC values for the other features that were used to predict ‘mode’.

Mariam Abdullah (ma3259)
Will Calandra (wcc9105)
Musonda Sinkala (mks9887)



Feature	AUC
Duration	0.50
Danceability	0.55
Energy	0.55
Loudness	0.53
Speechiness	0.56
Acousticness	0.56
Instrumentalness	0.54
Liveness	0.51
Valence	**0.50**
Tempo	0.51

Given the AUC values, we can conclude that **valence is not a good predictor of mode** and speechiness, acousticness, and energy exhibit relatively higher discriminatory power in predicting the 'mode'. However, the overall predictive performance is modest, and individual features might have limited standalone predictive value.

Mariam Abdullah (ma3259)
Will Calandra (wcc9105)
Musonda Sinkala (mks9887)

Question 8

Can you predict genre by using the 10 song features from question 4 directly or the principal components you extracted in question 6 with a neural network? How well does this work?

To answer the question, we built a neural network using the MLPClassifier from scikit-learn to predict music genre based on the 10 song features provided. We began by standardizing the 10 features using scikit-learn's StandardScaler after which scikit-learn's MLPClassifier was used to fit the Neural Network.

The architecture of the Neural Network consists of an input layer with 10 neurons corresponding to the features of the songs. We fit two hidden layers with 128 and 64 neurons, respectively, each followed by a Rectified Linear Unit (ReLU) activation function. The output layer has multiple neurons equal to the number of unique music genres, with a softmax activation function, enabling the network to perform multi-class classification. The model was trained using the backpropagation algorithm, with a maximum of 50 iterations, a learning rate of 0.001, and a random seed for reproducibility.

The **model accuracy on the test set was significantly low**, with an **accuracy of approximately 1.69%**. Thus the neural network, with the specified architecture and parameters, did not perform well in predicting music genre based on the provided features. Given the low model accuracy, predicting music genre directly from the 10 song features using the chosen neural network architecture is not effective. There might be other factors or features that are more informative for predicting music genre, or the model parameters and architecture need further optimization.

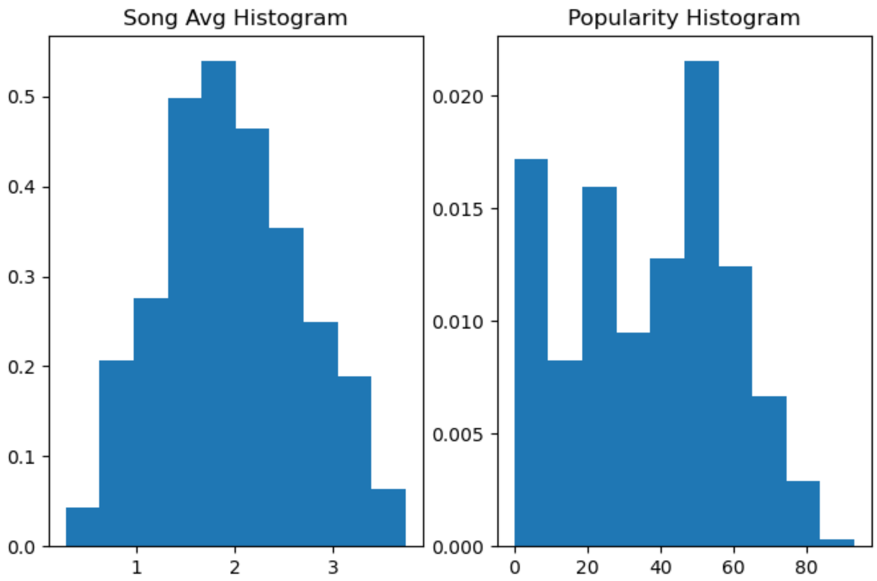
Question 9

In recommender systems, the popularity based model is an important baseline. We have a two part question in this regard: a) Is there a relationship between popularity and average star rating for the 5k songs we have explicit feedback for? b) Which 10 songs are in the "greatest hits" (out of the 5k songs), on the basis of the popularity based model?

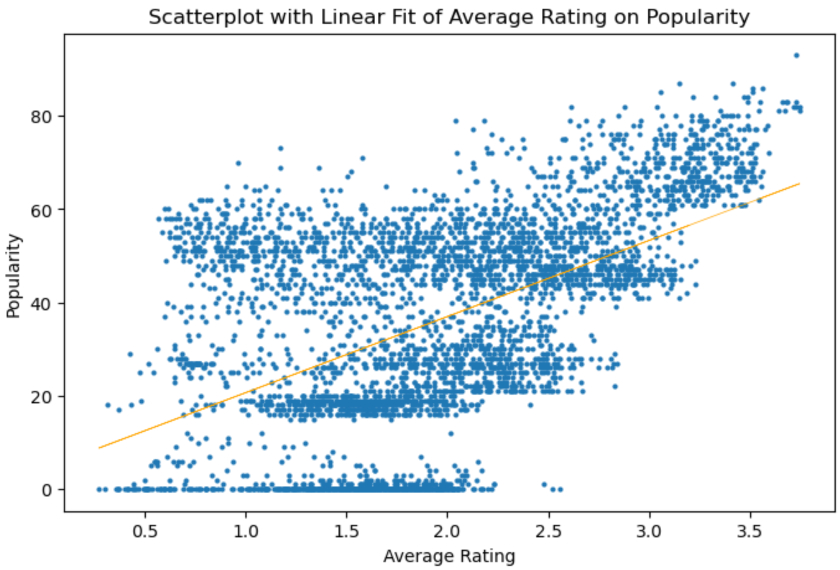
a)

We noticed from the histograms below that the average song rating appears to be normally distributed (which we assume for our test), while popularity appears to have a skewed distribution.

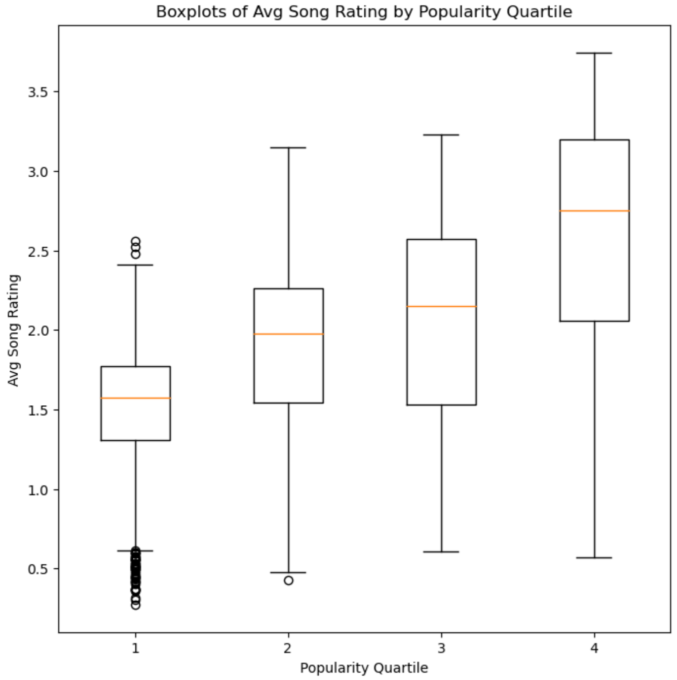
Mariam Abdullah (ma3259)
Will Calandra (wcc9105)
Musonda Sinkala (mks9887)



By visual inspection of this scatterplot, the relationship between popularity and average song rating is not linear, and there appear to be “bands” of data as popularity increases.



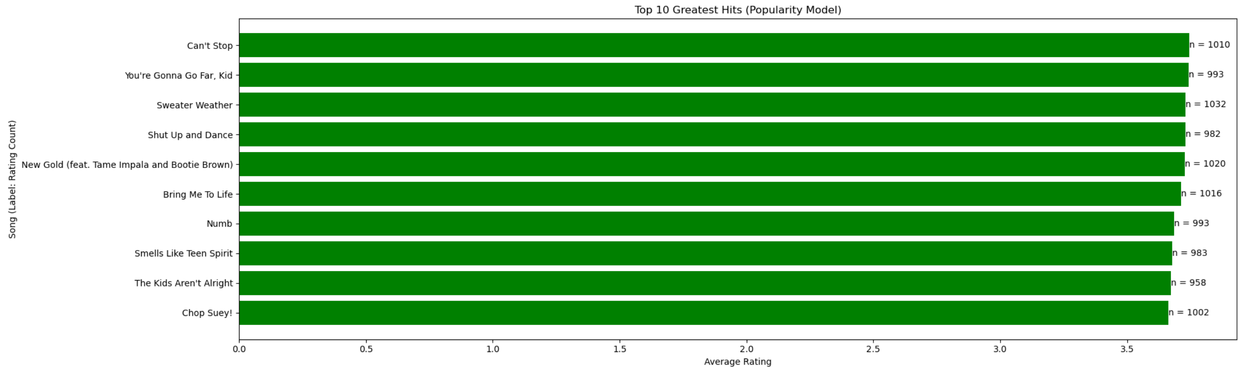
Therefore, we discretized popularity into its quartiles, and since the variances were not equal across groups, we compared the medians of average song rating across groups using a Kruskal-Wallis test. With an **H-statistic of 1062.995 and 3 degrees of freedom, we obtain a p-value of 3.88e-230**, which is a significant result. Visual inspection of the boxplots below show the differences in medians between groups.

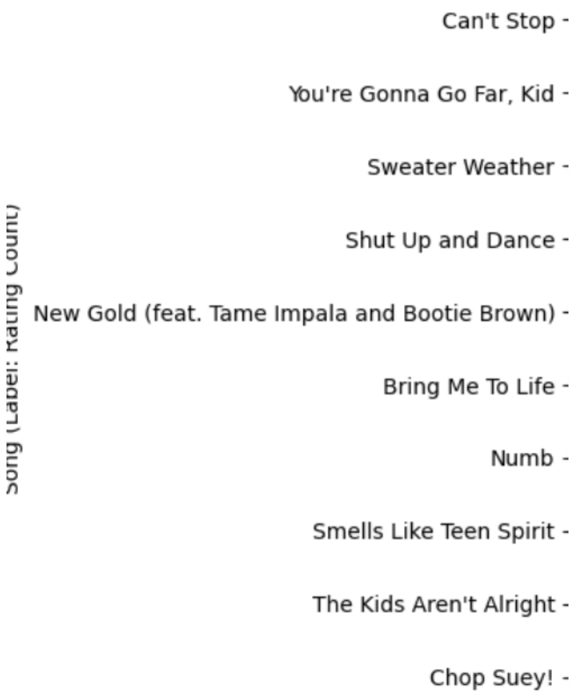


Consequently, we have strong evidence to reject the null hypothesis that on average, median song ratings are equal for varying degrees of popularity. **We conclude that there is a relationship between popularity and average star rating for the 5k songs in which we have explicit feedback.** By visual inspection in addition to a **spearman rank correlation of 0.499**, we deduce that more popular songs typically have higher ratings.

b)

We **calculated the average utility for each song in our ratings matrix (5k songs), and returned the top 10 songs with the highest average rating** that we refer to as our “greatest hits” in this report. Per the bar chart below (with song titles magnified for ease of viewing), they all average ratings above 3.6 stars on our 0-4 star scale, and they each contain ~1000 explicit user ratings.

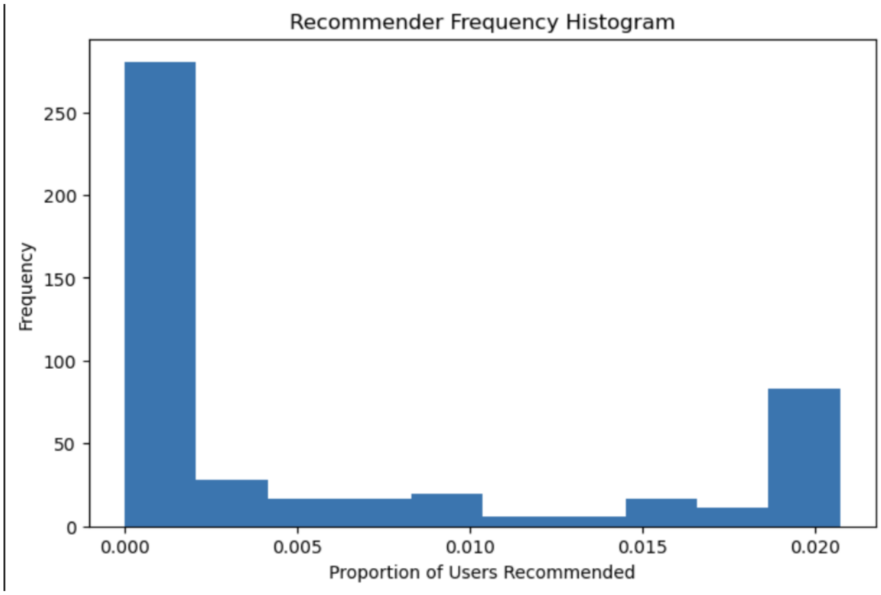


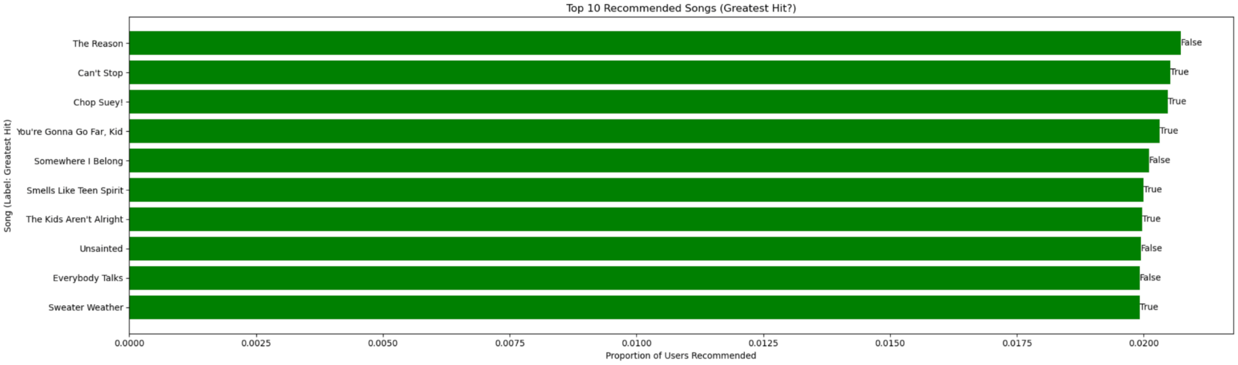


Question 10

You want to create a "personal mixtape" for all 10k users we have explicit feedback for. This mixtape contains individualized recommendations as to which 10 songs (out of the 5k) a given user will enjoy most. How do these recommendations compare to the "greatest hits" from the previous question and how good is your recommender system in making recommendations?

Given the premise that users have different preferences, building a recommendation system may provide insights into missing ratings for which we don't have explicit feedback. We decompose the user ratings into its latent factors using the SVD algorithm, regularizing our model and cross-validating over 5 folds to minimize the RMSE of our predicted ratings. However, since delivery matters in recommendation systems, we sort the top 10 estimated ratings for each user, return the list of songs as recommendations, and evaluate our recommender using mean average precision. We see that in our test set, **we achieve a 0.602 mean average precision on existing ratings, which is similar to the 0.620 mean average precision on the popularity model (recommending the 10 greatest hits to every user)**. We assume relevance by the precision of song recommendations that are above the median rating of 2. Since we do not have ground truth on songs for which users have not yet listened, it is difficult to beat the popularity model offline with existing ratings. We still recommend popular songs to users, but **only 19.97% of the time.** 6 out of the top 10 recommended songs are the popular songs, but at a lower frequency. Therefore, we conclude that **the recommender system gives more preference to learned user tastes over the popularity model, but still performs to the same satisfaction level of users.**



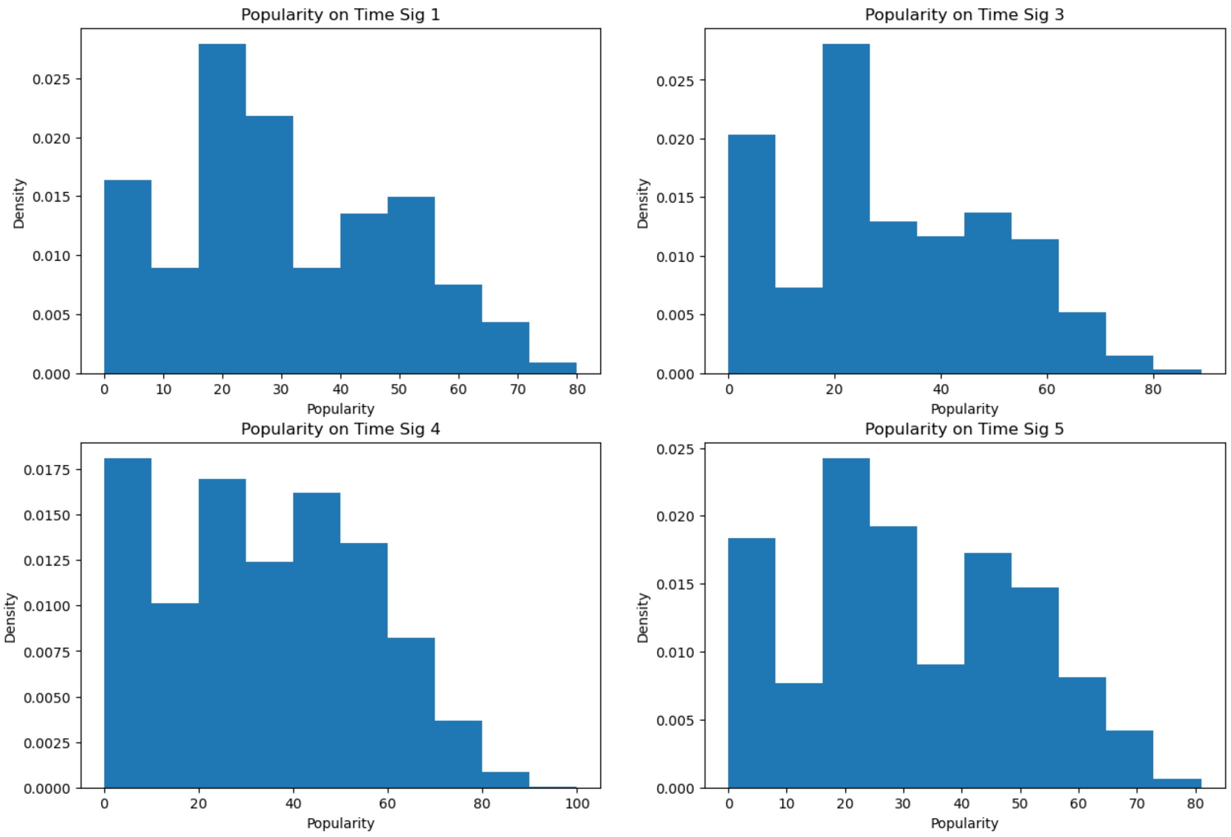


Mariam Abdullah (ma3259)
Will Calandra (wcc9105)
Musonda Sinkala (mks9887)

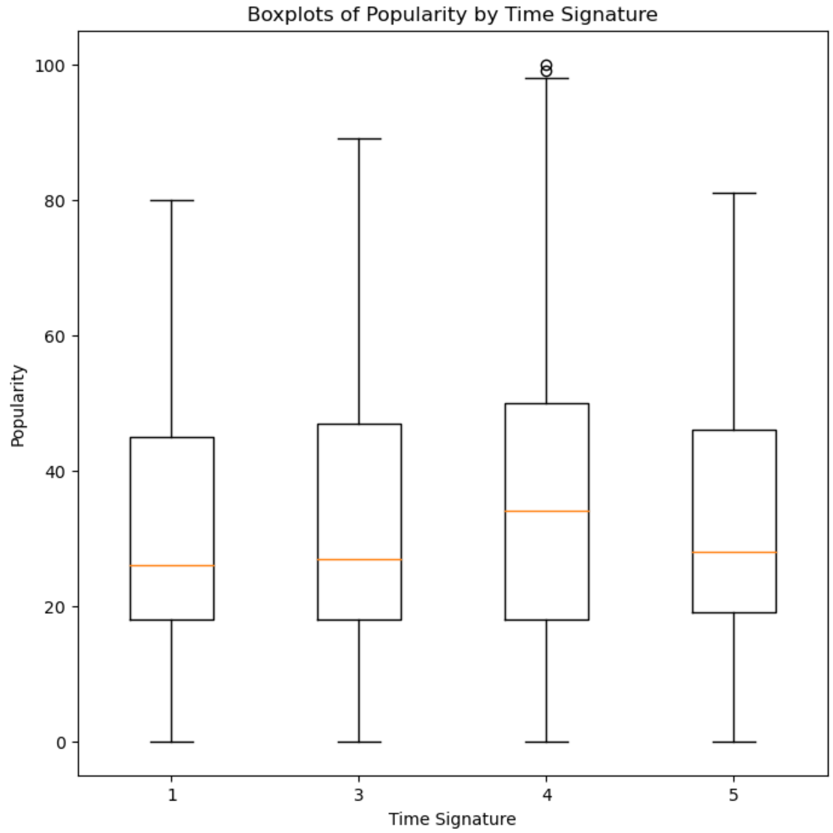
Extra Credit

Tell us something interesting about this dataset that is not trivial and not already part of an answer (implied or explicitly) to these enumerated questions [Suggestion: Do something with the number of beats per measure, something with the key, or something with the song or album titles]

For this question, we were curious if there is a relationship between time signature and the popularity of our songs. We see from the histograms below that popularity does not distribute normally for each time signature. Therefore, a Kruskal-Wallis test would be appropriate for this data.



With an **H-statistic of 81.09 and a p-value of 1.78e-17 at 3 degrees of freedom, we conclude that popularity medians do not distribute equally for songs across time signatures**. However, in a follow-up test involving songs of time signatures 1, 3, and 5, an H statistic of 2.516 and a p-value of 0.284 at 2 degrees of freedom indicate that these songs may have equal popularity medians. Additionally, in a one-sided Mann-Whitney U-test between songs involving time signatures of 4 and the rest of the data, we observe a **U-statistic of 110282139.5 and a p-value of 2.832e-19**. Therefore, **we can also conclude that songs with a time signature of 4 have a higher median popularity on average than the rest of the music.** This conclusion can be visually supported by the boxplot below:


Boxplots of Popularity by Time Signature
Popularity
Time Signature