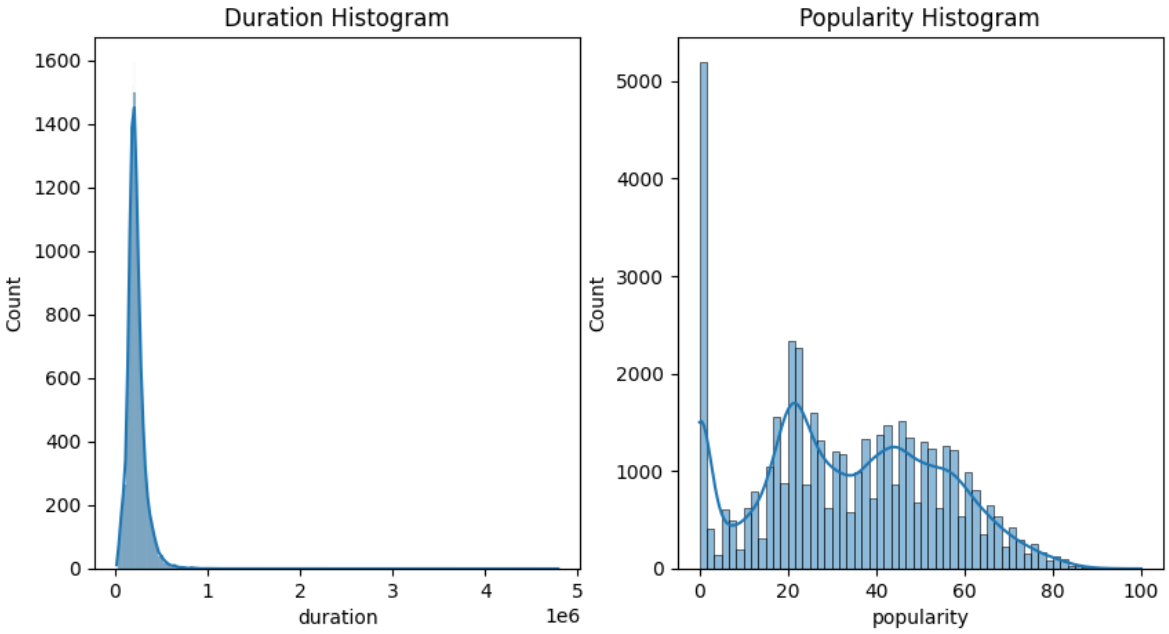
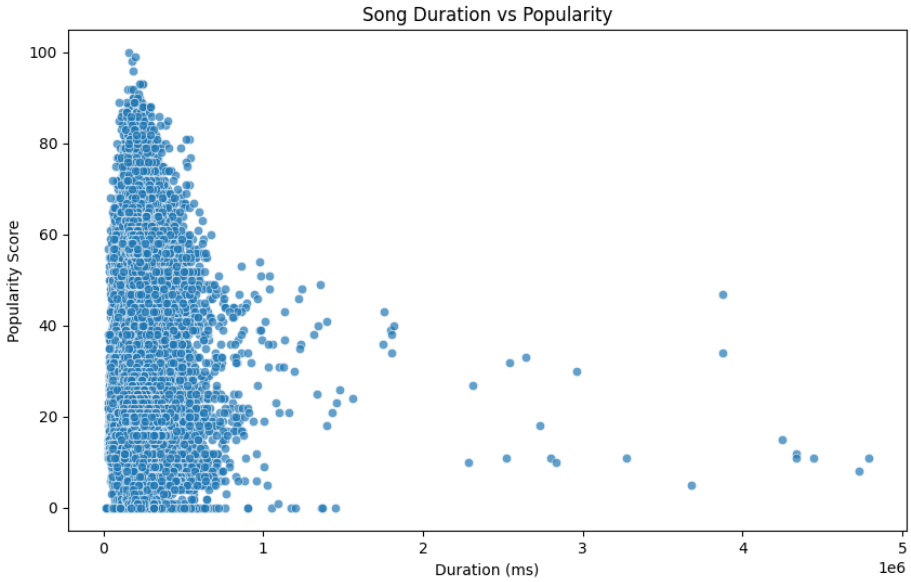
Kristo Papadhimitri
Group 42
Capstone Project

For the first 8 questions I took the Spotify data and removed the duplicate rows. There were rows that had the same song, same duration, same rating, and everything else was the same except song number and genre. I removed these from the first 8 and EC so certain sounds wouldn't be double or triple counted for the analysis. The last two have the original data so they can lineup with the ratings. For seed I used 10206068.

1) For this problem we are looking if there is a relationship between duration and popularity. To do this I found the correlation coefficient between duration and popularity. Given (chart 1) we see there are many songs that are low in popularity so the distribution is not normal. Given this

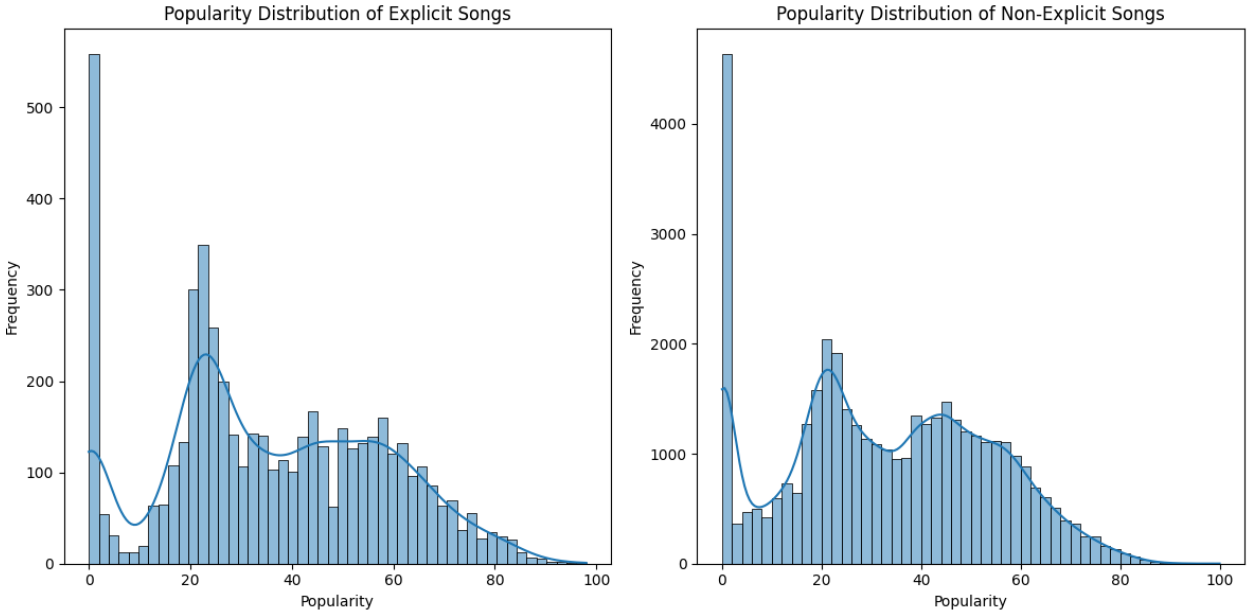
I used Spearman rank correlation coefficient test with duration and popularity values. The result is a correlation coefficient of -0.0543 with a p-value of 2.15 e-31. The low p-value is expected since there is a lot of data for this but the p-value is low which indicates there isn't much of a relationship between duration and popularity. Since it is negative we can say there is a slightly negative correlation between song duration and popularity. A lot of songs tend to fall in a similar duration range so these may not be affected as much by the negative correlation but songs that are greatly longer will likely not be as popular.



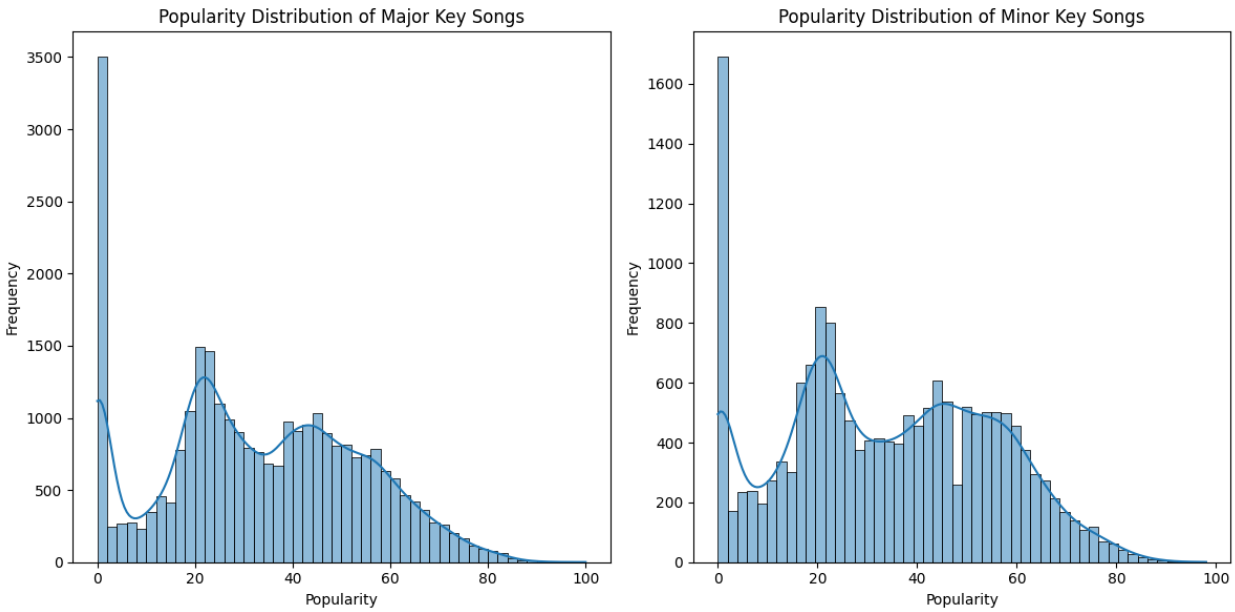


2) For this problem we broke explicit and non-explicit songs into two groups and calculated the mean and compared the popularity of explicit songs to non-explicit songs. I used a Mann-Whitney U test to compare the popularity since the popularity is not normally distributed. This gave a result of explicit songs having an average popularity of 35.863 and non-explicit songs had

an average popularity of 32.958. The p-value for this is 1.107 e-17 which is very small because of the large sample. I then calculated Cohen's d to get a value of 0.136 which means it has a small effect size which makes sense when we compare the averages. Using bootstrapping I got the 95% confidence interval of the mean differences to be [2.274, 3.537] this further confirms that explicit songs are more popular than non-explicit but only a little bit more. If all else is equal and you only care about your song being as popular as possible then have it be explicit.



3) In this problem I broke the dataset into two groups minor key group and major using the mode variable. I calculated the means of the popularity of these two groups and did a Mann-Whitney U test to see the significance and Cohen's d to get an understanding of the effect size. Major key songs had an average popularity of 32.884 and minor songs had an average popularity 33.95 with a p-value of 1.43e-6 which means the results are significant even though it is such a small difference, Cohen's d is -.051 which is very small and means minor is larger than major. Given such a small effect and small difference of means I would not treat this as being a big difference at least not enough to do anything actionable with.



4) For this problem I created 10 models with the 'y' variable being popularity and the x being one of the 10 features. I did this until popularity was compared with each feature, I used 15% of the data for testing and the rest for training. Then I calculated the r^2 and RMSE for each of these models. The result was instrumentalness was the best predictor with R^2 = .0248 and a RMSE of 20.622. Each predictor had an RSME over 20 and a .0248 is not helpful for predicting popularity and RSME of over 20 is significant since popularity can go from 0 – 100. Predicting popularity is more complex than just one feature as a result.

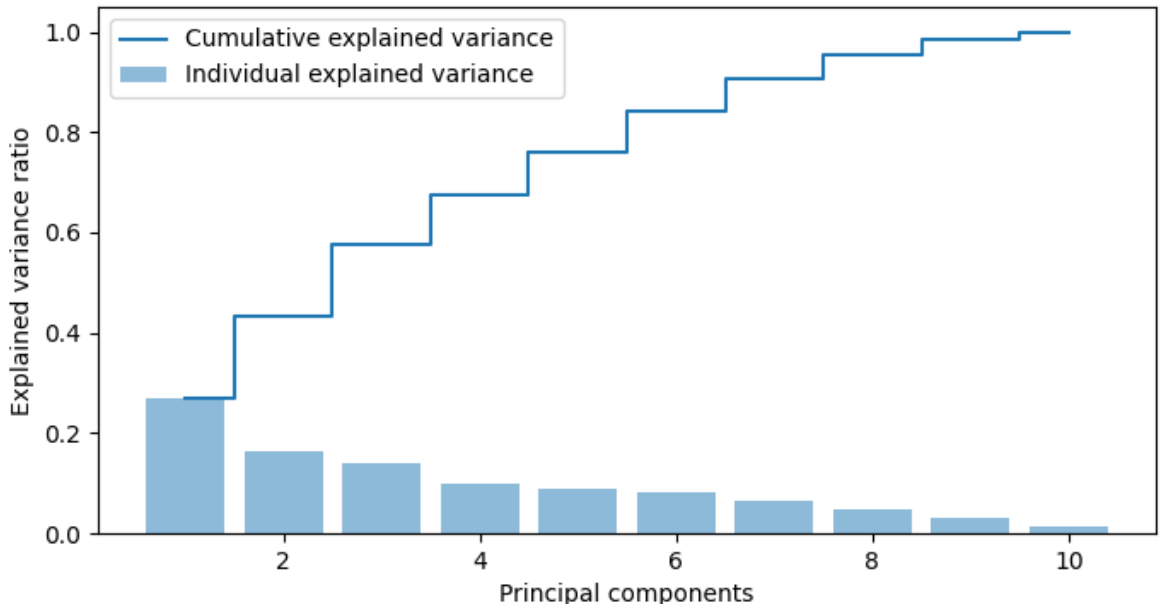
Table 1:

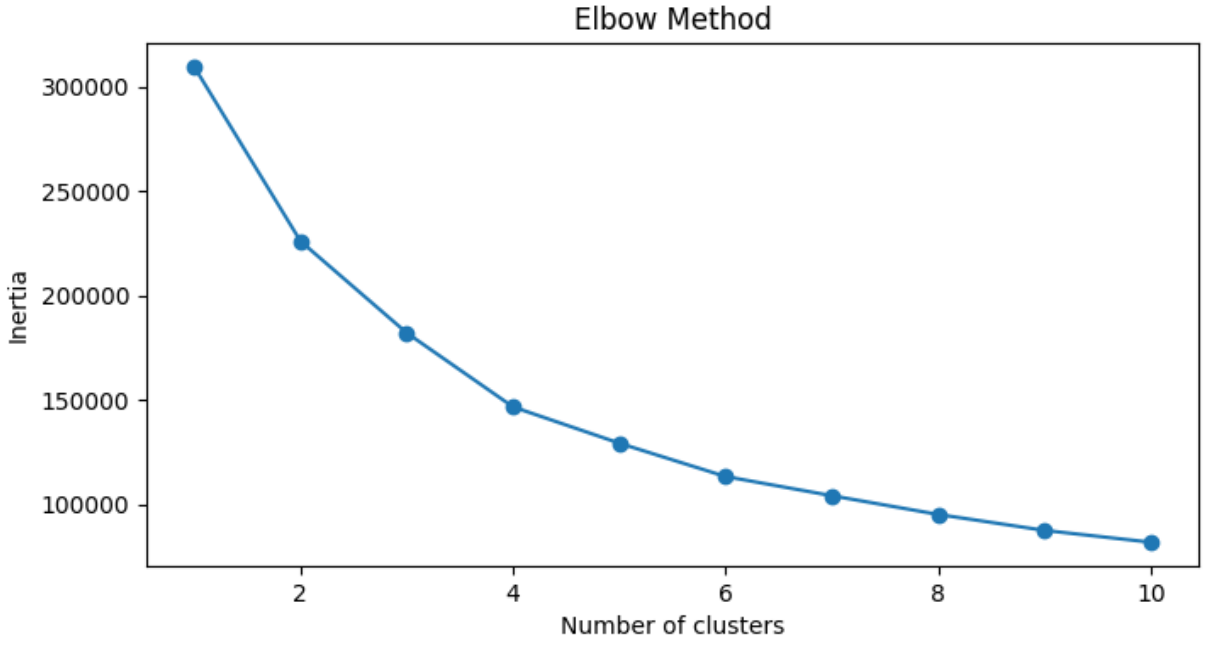
Feature: duration	R^2: 0.004017072484190587	RMSE: 20.8412039352474
Feature: danceability	R^2: 0.000723082607665515	RMSE: 20.875639287925758
Feature: energy	R^2: 0.0052260850540668535	RMSE: 20.82855064162613
Feature: loudness	R^2: 0.002818040326941995	RMSE: 20.85374519303395
Feature: speechiness	R^2: 0.0034675988340293618	RMSE: 20.84695208266725
Feature: acousticness	R^2: 0.0012762518954030355	RMSE: 20.869860428792766
Feature: instrumentalness	R^2: 0.02481794949902205	RMSE: 20.622423711116447
Feature: liveness	R^2: 0.0034977855841267225	RMSE: 20.846636334530103
Feature: valence	R^2: 0.0007868192170498167	RMSE: 20.874973524681923
Feature: tempo	R^2: -0.00038399780455034005	RMSE: 20.88719995415545

5) Last problem was a simple regression while this one is a multiple regression with all the features in the last problem combined. I used 15% of the data to test with the rest to train and calculated the r^2 and RSME, then I also did Ridge and Lasso regularization and got the R^2 and RSME values of those. For multiple regression I got R^2 = .056 and 20.294, means the model is not great for predicting the popularity. Lasso and Ridge helps minimize the impact of some variables or even 0 in the case of Lasso. This didn't help, Ridge gave R^2 = .0557 and RSME =

20.294, Lasso gave R^2 = .0065 and RSME = 20.815. Ridge gave about the same with Lasso giving worse results but the features aren't good predictors for popularity regardless.

6)For this problem I took the data and tried to extract meaningful principle components from them. Using PCA I got 4 meaningful components that accounted for .665 of the variance. The I used the 4 components to cluster and using the elbow method of clusters I got the optimal clusters to be 3 clusters while this does significantly cluster some song genres, there are a lot more than 3 song genres so it does they don't reasonably correspond in that regard.





7) I took valence and mode and made valence the ‘x’ and mode the ‘y’ for the models. I then did a logistic regression but balanced the weights because the model was highly favoring one over the other I also balanced the weights for SVM. Then I used a classification report to see the results of both models and assess their performance. The accuracy for the logisitic regression is .489 and the accuracy for SVM is .510. These are not great especially for two options where a guess is 50/50. A neural network may be better for this but it is possible that valence is not a good predictor for major or minor key.

Logisitic Regression			
	Precision	Recall	F1-score
0	0.37	0.52	0.43
1	0.64	0.48	0.53

Support Vector Machine			
	Precision	Recall	F1-score
0	0.38	0.48	0.43
1	0.63	0.53	0.57

8) Using the PCA components from question 6, I used the PCA for the ‘x’ and track genre for the ‘y’ and used .85 for training and .15 for test. Then I used tensorflow and keras for the neural network model, the accuracy was not great and I believe it may have been from the model learning too fast so I tried lowering the rate. The test accuracy is 16.017% which is not great and not something I would use to predict track genre.

9) In the first part of the question I took the popularity of the first 5000 songs and compared it to the star ratings by calculating the correlation. Then from the 5000 songs and derived the 10 most popular songs to be used in a popularity based model. This is a good go to list to recommend, especially good for a cold start if you have a new user that you don’t know much about you can recommend the most popular songs they are likely to like some of them and based on their ratings you can continue form there.

songNumber	Artists	Track Name	Popularity	Track Genre
2003	The Neighbourhood	Sweater Weather	93	alt-rock
3003	The Neighbourhood	Sweater Weather	93	alternative
3300	Oliver Tree;Robin Schulz	Miss You	87	alternative
2000	The Neighbourhood	Daddy Issues	87	alt-rock
3000	The Neighbourhood	Daddy Issues	87	alternative
2106	The Killers	Mr. Brightside	86	alt-rock

3004	GAYLE	abcdefu	86	alternative
2002	The Neighbourhood	Softcore	86	alt-rock
3257	The Killer	Mr. Brightside	86	alternative
3002	The Neighbourhood	Softcore	86	alternative

10) For the mixtape we need to put together a mixtape of 10 songs that the user would like based on our recommender model. I went with a simple approach for the mixtape for the user and used the user's ratings and extracted songs that they rated a 4. If the user didn't have 10 songs with a 4 rating I included in songs from the greatest hits as that is a safe choice for recommending. Overall this did not lead to much overlap between the greatest hits and the mixtapes, popularity does not always equal best ratings and many people have different tastes. The overlap was 0.867% all together which is very low. This recommender system can not be rated as is but only with user feedback I believe it will do okay since we are recommending songs that users like but tastes can change and people could dislike the greatest hits.

EC) For this problem I am looking to see if we can predict the 'energy' of a song based on 'tempo'. I am curious to see if the tempo plays a major part in the energy, to do this I used a linear regression of the two variables. Afterwards I took the r^2 of these. In the model I used 85% of the data for training and 15%. The r^2 for this was .0595 which is better than what we saw for predicting popularity in the previous problem but not very good on its own, there may be multiple things that go into the energy of a song but tempo is not a great predictor for energy on its own.