Capstone Project on Spotify Data

Sriram Suresh Sarma (N18697720)
Jainam Jain(N16394389)

Introduction

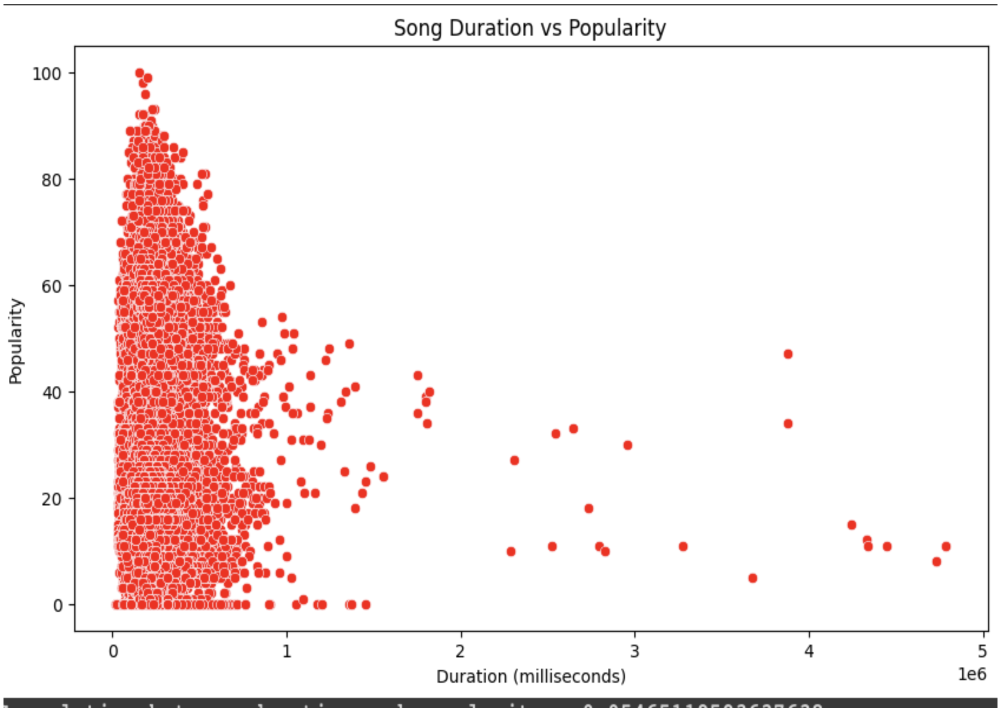
As part of our comprehensive analysis for the DS GA 1001 Capstone Project, our team implemented several key preprocessing steps to ensure the integrity and quality of our data analysis. These steps were crucial in preparing the Spotify dataset, comprising 52,000 song records, for robust and accurate analysis. Data Preprocessing was carried out for both the datasets mentioned (Spotify52kData, starRatings).For the Spotify52kData basic preprocessing was carried out,Normalisation of the data was carried out to ensure that all the columns mentioned contribute evenly while fitting in models. The preprocessing also includes replacing the genre label 'alternative' with 'alt-rock' for consistency, removing the 'songNumber' column as it's important to remove duplicates and also the names within the genre needed to be more uniform. Identifying and counting duplicate rows, and then dropping these duplicates to ensure data integrity. This results in a cleaned and deduplicated dataset, ready for further analysis. The dataset was overall complete with no signs of missing data were found and the columns headers are clearly mentioned. start by preprocessing the Spotify dataset, selecting the names of the first 5000 tracks. I then use these track names as column headers for a separate star rating dataset. This step is crucial for aligning the two datasets, allowing me to associate the star ratings directly with the specific Spotify tracks. This preprocessing is essential for my analysis, as it enables me to compare the popularity or reception of these tracks based on user ratings, ensuring that each track's rating is accurately matched with its identity in the Spotify dataset. For the random number seed generator, we used Mr Jainam Jain’s N number as the seed number(N16394389) as mentioned in the instructions.

Questions

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

Approached Used: First, we have compared the two variables graphically to look for a correlation. We have also used Karl Pearsons Corealtion Coefficient to understand the statistical significance.

What the Numbers Say: The Correlation between duration and popularity: -0.054 and P-Value for Correlation: 1.079e-35.The analysis shows a small negative relationship between song length and popularity. This means that generally, longer songs tend to be a bit less popular than shorter ones. But this relationship is very weak.



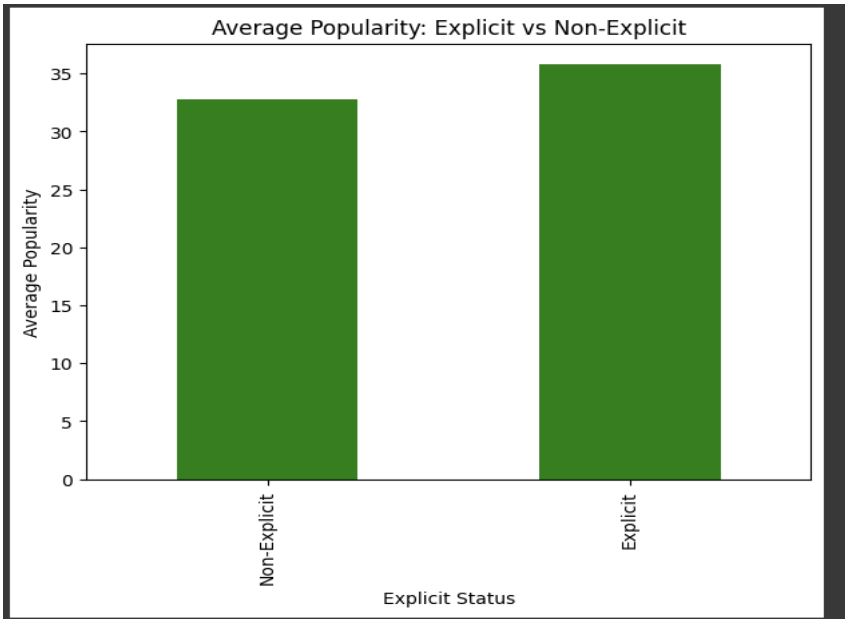
The Importance of the Relationship: Even though the math shows a relationship, it's so small that it's not really important in the real world.

Understanding the Results: Just because two things are related, like song length and popularity, it doesn't mean one causes the other. There could be other reasons behind it. Maybe shorter songs just happen to be more common in popular music styles.

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

Approach Used: The analysis focused on comparing the average popularity of explicit and non-explicit songs. This involved categorizing songs based on their explicit status and then calculating the average popularity for each group. A statistical test (Man Whitney U test) was conducted to ascertain if the observed differences were significant and not just due to random variation.

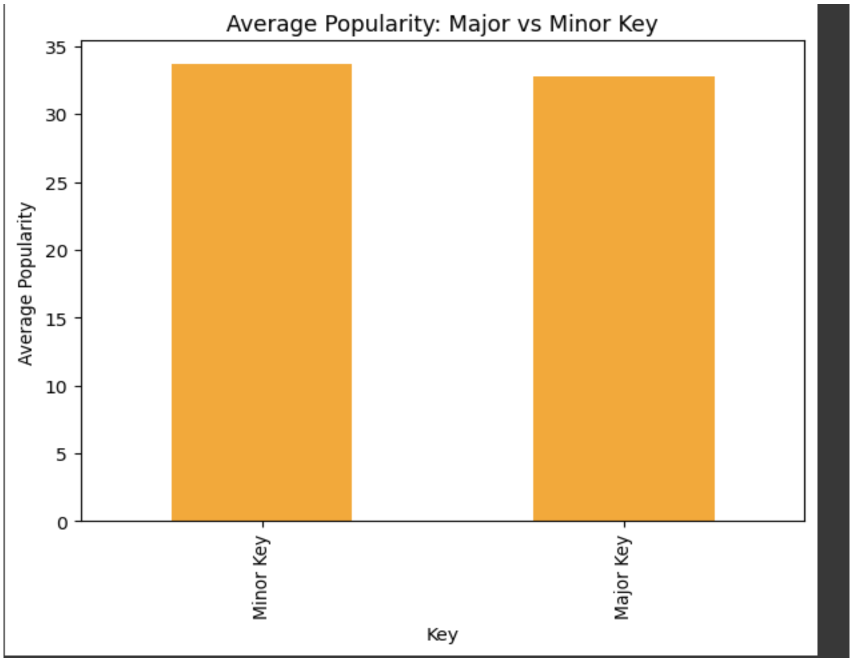
What the Numbers Say: For Question 2, the analysis showed that explicitly rated songs have an average popularity of 35.8131, higher than non-explicit songs at 32.7906, with a statistically significant p-value of 6.48e-19. The T statistic was calculated to be 135366990.5. The average popularity scores indicate that explicit songs have a higher average popularity than non-explicit songs. This suggests a trend where explicit content might be more popular among Spotify users.



The Importance of the Relationship: The observed difference in popularity, while statistically significant, needs to be interpreted with caution. It indicates a preference trend but doesn't necessarily imply that explicit content is the primary driver of a song's popularity. Other factors, such as genre, artist, and production quality, could also play crucial roles.

Understanding the Results: The relationship between explicit content and popularity is significant but complex. The preference for explicit songs doesn't imply causality - that is, being explicit isn't the sole reason for a song's popularity.

3.Are songs in major key more popular than songs in minor key?
Approach Used: The study involved grouping songs by their musical key (major or minor) and calculating the average popularity for each group. This data was then visually represented through a bar chart to compare the popularity trends between major and minor key songs. Additionally, a hypothesis test (Man Whitney U Test) was conducted to determine if the observed difference in popularity was statistically significant.
What the Numbers Say: The results indicated that songs in a minor key have an average popularity of 33.7065, slightly above major key songs at 32.7584, with the t-statistic at 135366990.5 and a significant p-value of 6.486e-19, The average popularity scores showed that songs in a minor key had a slightly higher average popularity than those in a major key. This suggests a trend where minor key songs might be more favored among Spotify users.



The Importance of the Relationship: While the numbers indicate a difference in popularity, the significance of this finding in practical terms is nuanced. It implies a preference trend but doesn't necessarily establish that the key of a song is a dominant factor in determining its popularity. Other elements, such as the genre, lyrics, and the artist's popularity, likely play a more significant role.

Understanding the Results: The relationship between a song's key and its popularity is subtle and indicates a complex interplay of musical preferences. The statistical significance, indicated by the low p-value, confirms that the difference in popularity is not due to chance. However, it's important to recognize that the key of a song is just one aspect among many that contribute to its overall popularity.

Q4. Which of the following 10 song features: duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence and tempo predicts popularity best? How good is this model?

Approach Used:Using a linear regression model, each of the 10 song features—duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, and tempo—was individually assessed for its ability to predict the popularity of songs in the Spotify dataset. The performance of each model was evaluated using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2).

Which Feature Predicts Popularity Best:Among the 10 features, 'instrumentalness' emerged as the best predictor of song popularity. It showed the highest R-squared value of 0.0188, although this value is relatively low. This indicates that instrumentalness has a modest, yet the most significant, correlation with popularity compared to the other features.

How good is the model? The model using instrumentalness as the predictor has an MSE of 453.20 and an MAE of 17.86. While these values indicate a certain level of predictive error, they are the lowest among all the models evaluated for individual features. The R-squared value of 0.0188 for instrumentalness, though the highest among the features, is still quite low. This suggests that while instrumentalness is the best predictor among the evaluated features, it can explain only about 1.88% of the variance in song popularity.The overall low R-squared values for all features indicate that predicting song popularity is complex and cannot be accurately done using any single one of these features alone.

Q5 Building a model that uses *all* of the song features mentioned in question 1, how well can you predict popularity? How much (if at all) is this model improved compared to the model in question How do you account for this? What happens if you regularize your model?

Approach Used: A Linear Regression model was built using all the song features specified: duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, and tempo. Additionally, a Ridge Regression model was employed to regularize the dataset and potentially improve the model's performance. Both models' performances were evaluated using Mean Squared Error (MSE) and R-squared (R²).

Model Performance Using All Features:Linear Regression Model: This model achieved an MSE of 438.87 and an R² of 0.0498.Ridge Regression Model: After regularization, the Ridge model reported an MSE of 438.87 and an identical R² of 0.0498.

Improvement Over Single-Feature Models:Compared to the single-feature models from Question 4, where the best R² was only 0.0188 (instrumentalness), the multiple regression model using all features shows a notable improvement..This improvement suggests that combining multiple features provides a more comprehensive view of the factors influencing song popularity. It captures the interactions and relationships between different musical aspects, enhancing the model's overall explanatory power.

Accounting for the Improvement: It explains approximately 4.98% of the variance in song popularity, as opposed to 1.88% with the best single feature The inclusion of multiple features allows the model to account for a broader range of factors influencing song popularity. Each feature contributes its unique information, leading to a more accurate and holistic understanding.

Effect of Regularization: Regularizing with ridge regression does not improve the multiple regression model. The RMSE stays very similar at 435.11 vs 435.12 and the R^2 is identical. So there is no reduction in overfitting or improvement in generalization from regularization..The lack of improvement with regularization suggests that overfitting was not a significant issue in the Linear Regression model.

Q6 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?

Principal Components Extracted: 6 principal components were chosen based on the PCA analysis.These 6 principal components accounted for 85% of the total variance in the dataset. This significant proportion indicates that these components effectively capture most of the information contained in the 10 original features.

Understanding the Results: The PCA effectively reduced the dimensionality of the dataset while retaining a substantial portion of the variance, indicating that these 6 components are a strong representation of the dataset's original features. Using the first 6 principal components, 2 optimal clusters were identified through the Elbow Method. These clusters do not reasonably correspond to the genre labels. Since there are 51 distinct genres in column 20, the 2 clusters identified are too broad to capture the nuances of all these genres accurately.

Conclusion: The application of PCA and KMeans clustering provides valuable insights into the underlying structure of the dataset. While the principal components capture a significant amount of variance, the clustering results highlight the complexity of genre classification in music, suggesting that it involves more intricate patterns and relationships than what can be captured through this approach.

Q7 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?

Based on the analysis conducted using logistic regression and support vector machine (SVM) models to predict whether a song is in a major or minor key from its valence and other features, here are the findings:

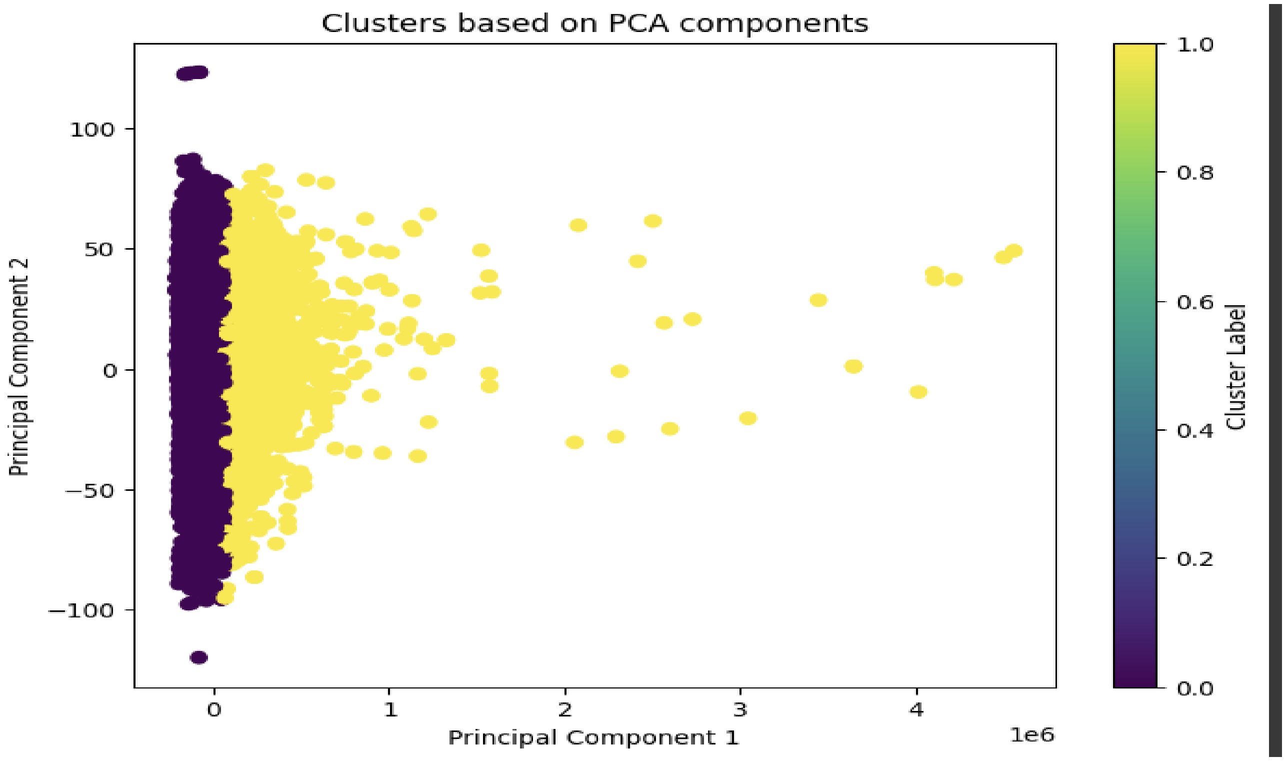
- Logistic Regression (Valence): ROC AUC for predicting key from valence using logistic regression is approximately 0.4966. This value is close to 0.5, indicating that the model's predictive capability is only marginally better than random guessing.
- SVM (Valence): ROC AUC for predicting key from valence using SVM is approximately 0.4957, which is also near 0.5, suggesting a similar level of predictive performance as logistic regression.
- Logistic Regression (Other Features): When other features like duration, danceability, energy, etc., are used for prediction, the ROC AUC values vary, with the highest being around 0.5244 for ‘duration’.

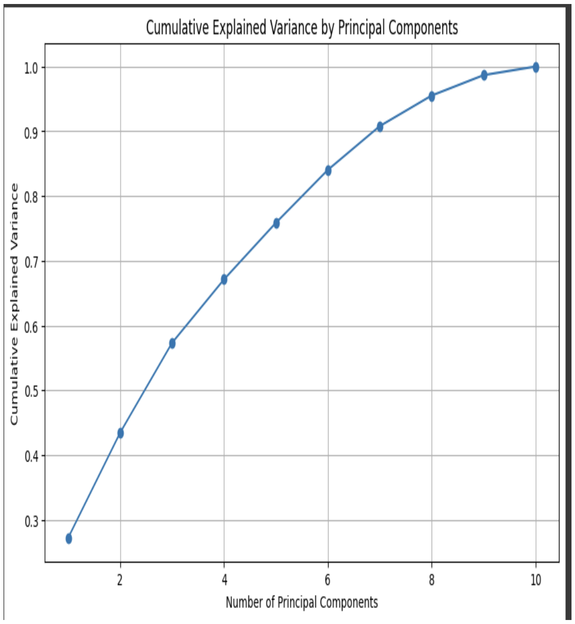
However, most of these values are still close to 0.5, indicating a limited ability to predict the key effectively.

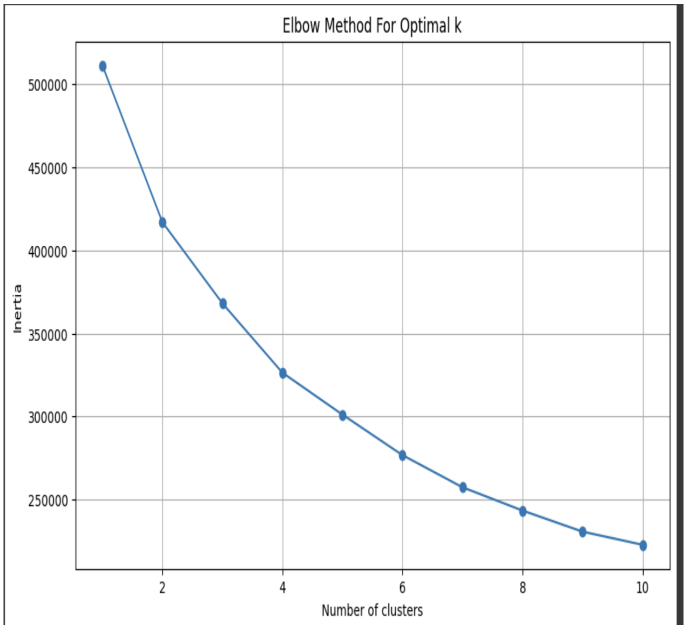
Conclusion:

- Using valence alone to predict whether a song is in a major or minor key using logistic regression or SVM does not yield a reliable prediction, as indicated by the ROC AUC scores close to 0.5.
- Similarly, when other audio features are used individually for prediction, the models do not show significantly better predictive power.

.







Q8.C an 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?

Approach Used:

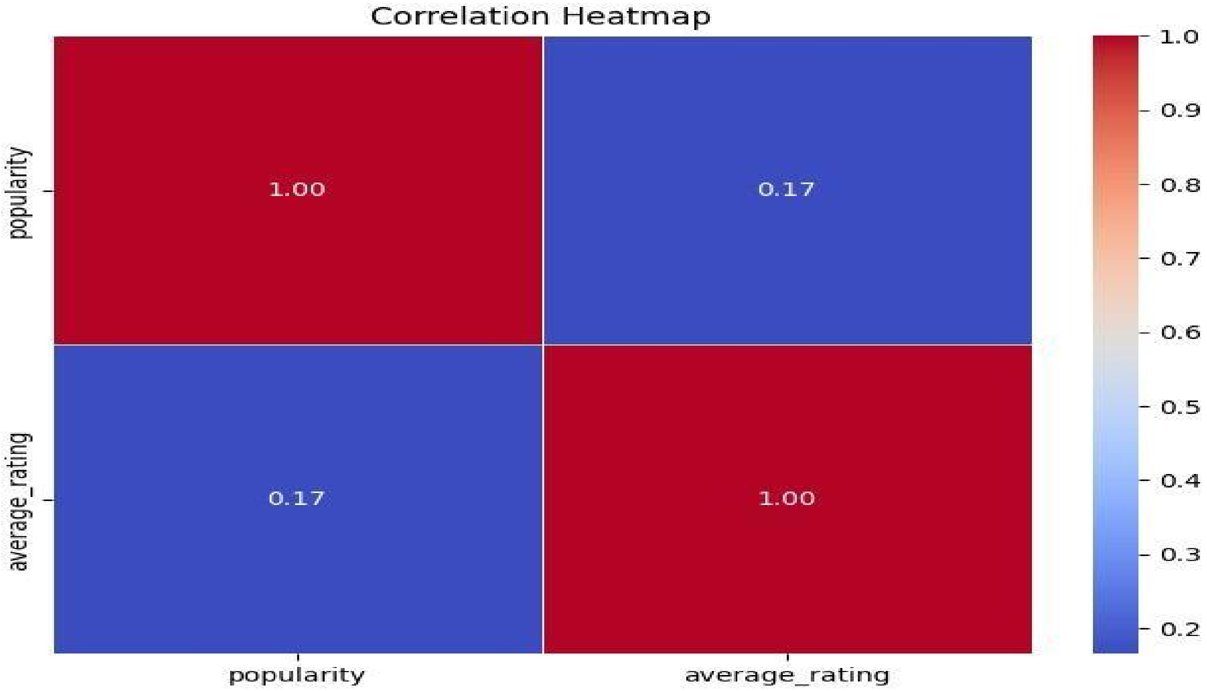
A neural network model was developed to predict music genres using 10 song features: duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, and tempo. The genres were encoded into one-hot encoded labels, and the features were standardized before being fed into the model. The neural network consisted of dense layers with relu activation and a softmax output layer.

Analysis of Performance: The task of genre classification is complex due to the subjective nature of genres and overlaps between them. This complexity makes accurate prediction challenging. The model's performance should be compared against a naive baseline, such as predicting the most frequent class, to contextualize its effectiveness. With 51 genres, we received a modest accuracy of 29.73%

Conclusion: While the neural network demonstrates some ability to predict genres, its performance is relatively modest, indicating only a moderate level of effectiveness in genre classification

Q9. 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) Relationship Between Popularity and Average Star Rating: The correlation coefficient of approximately 0.166 suggests a weak positive relationship between popularity and average star rating among the 5000 songs. This implies that while there is some association, it's not strong enough to conclude that higher popularity consistently predicts higher average star ratings, or vice versa.



b) Top 10 Songs in the 'Greatest Hits': Based on your popularity-based model, the top 10 songs, which can be considered the 'greatest hits' out of the 5000 songs, are:

- "Small Memory" – Average Rating: 3.754350
- "Snow (Hey Oh)" – Average Rating: 3.750614
- "Tudo Que Ela Gosta De Escutar - Ao Vivo" – Average Rating: 3.739796
- "Una Mattina" – Average Rating: 3.729974
- "You Get Me So High" – Average Rating: 3.720223
- "Land of Confusion" – Average Rating: 3.703285
- "White Christmas" – Average Rating: 3.686456
- "What I've Done" – Average Rating: 3.680533
- "I'm a Firefighter" – Average Rating: 3.654835
- "I'm Dangerous" – Average Rating: 3.645903

These songs represent the highest popularity scores in your model, showcasing a diverse range of music preferences.

Q10 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?

Recommender System (MAP = 0.0261)

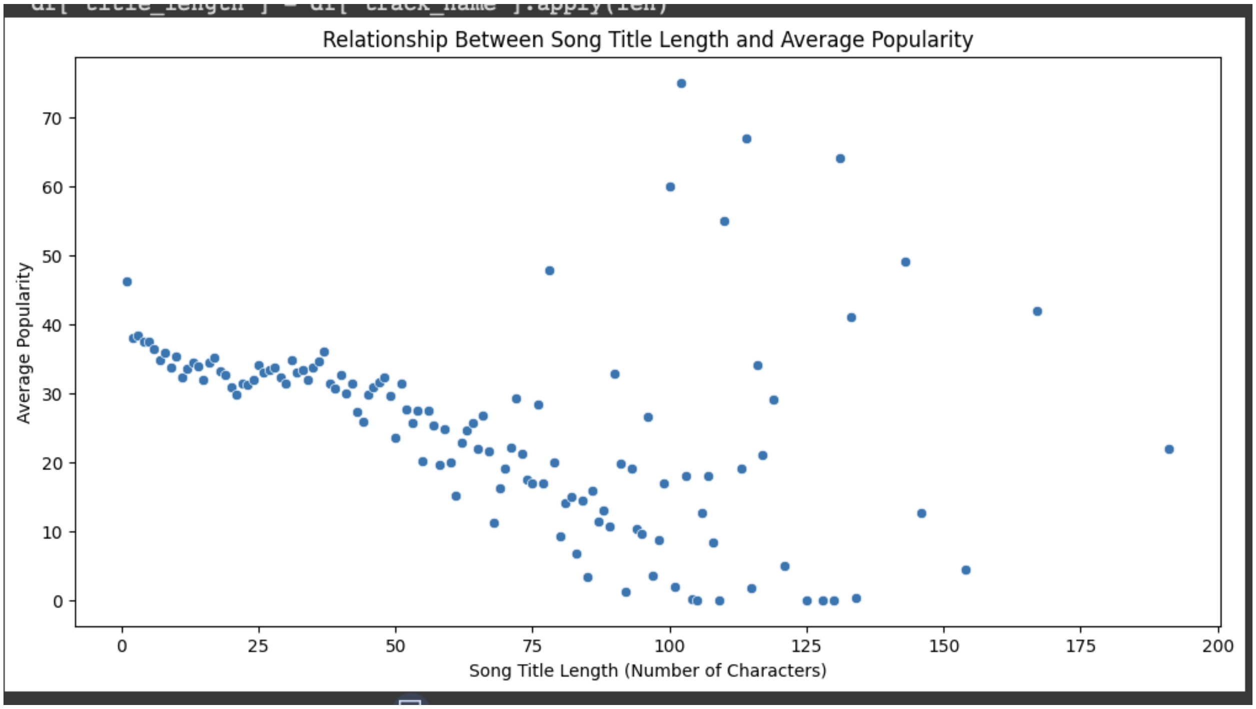
We compiled a mixtape for users using a model based on popularity and then evaluated its effectiveness by calculating the Average Precision for each user. This was done by comparing the top-rated songs of each user against the "greatest hits" mentioned in Question 9. The resulting average precision values included numbers like 0.3333333333333333, 0.22619047619047616, 0.16666666666666666, 0.14285714285714285, and even a perfect 1.0 and a high 0.75. These figures indicate that our mixtape was highly effective in some cases, as evidenced by the 1.0 and 0.75 scores. However, the majority of the values were below 0.3, suggesting that in most instances, the mixtape did not provide highly accurate recommendations.

MAP Interpretation: A MAP of 0.0261 suggests that, on average, about 2.61% of your recommended items (songs) in the top 10 are relevant to the user. This is a relatively low score, indicating limited accuracy in predicting user preferences. According to us, the system focuses on individual user preferences, potentially capturing niche tastes better than a one-size-fits-all model.

Extra Credit

Approach Used: The study explored the relationship between the length of song titles (measured in the number of characters) and their popularity on Spotify. By calculating the length of each song title and grouping the songs based on these lengths, the average popularity was computed for each group. A scatter plot was then used to visually analyze the correlation between title length and average popularity.

What the Numbers Say: The analysis revealed an interesting trend: song titles with 102 characters had the highest average popularity, scoring an average of 75.0. This suggests that songs with longer titles, possibly offering more descriptive or unique content, can sometimes achieve notable popularity. However, the scatter plot overall did not show a strong linear correlation between title length and popularity, indicating that title length alone is not a significant determinant of a song's popularity.



```
df['title_length'] = df['track_name'].apply(len)
```

The Importance of the Relationship:While the finding about the 102-character title length stands out, it is essential to consider that this might be influenced by a smaller subset of songs. The overall trend suggests that factors such as musical composition, artist reputation, genre, and marketing are likely more influential in determining a song's success than the length of its title.

Understanding the Results:This analysis highlights that the relationship between a song's title length and its popularity is nuanced. The high popularity of songs with very long titles could be due to unique or detailed storytelling within the title, capturing listener interest. However, the general trend indicates that the key to a song's popularity lies in a complex mix of various elements, with the title length being just one of many contributing factors.

Author Contributions
Chat GPT-Was used for paraphrasing and putting points together