Noelis [redacted]
Trisha [redacted]
Keira [redacted]
Rose [redacted]

Github Repo Link: https://github.com/noelisaponte/ds3000-project/tree/main

Exploring the Relationship Between K-pop Lyrics and Song Popularity

Abstract

Our project, *"Exploring the Relationship Between K-pop Lyrics and Song Popularity"* analyzes how lyrical sentiment and content impact a song's success using machine learning methods, linear regression and polynomial regression. The linear regression model used NumPy and demonstrated that a weighted sum of views, likes, and comments perfectly explained song popularity within the dataset, with a near-zero MSE and R^2 of 1.0. The polynomial regression model (degree 2) captured patterns in the data, yielding an R^2 of 0.999 and low MSE. The results showed that a combination of English and a variety of subject material could affect song appeal, even though lyrical sentiment does not directly correlate with views. Finally, the clustering model showed the distribution of engagement using various attributes that correlate with a song's popularity. For musicians and producers looking to maximize worldwide exposure, our findings can provide useful insights.

Introduction and Data Description

The K-pop industry has witnessed exponential growth over the past decade, crossing cultural and linguistic barriers to achieve widespread global recognition. As the genre evolves, understanding the factors contributing to a song's popularity becomes increasingly valuable for artists and producers. The goal for any artist or producer in the end is to gain popularity so it's imperative they understand what helps to achieve this. This project, titled *"Exploring the Relationship Between K-pop Lyrics and Song Popularity,"* aims to analyze how the sentiment and content of K-pop lyrics influence a song's success. More specifically, it aims to determine whether songs with positive lyrics, or the inclusion of English phrases, contribute to wider fame, and how the language within K-pop lyrics have changed over time. As the K-pop genre grows each year and reaches a wider audience, the addition of some English lyrics could be crucial to making the songs more relatable and easily understandable. By dissecting these elements, the models we create aim to unveil patterns that could predict or explain song popularity in this music genre.

The project includes two main types of data, lyrics data and popularity metrics. The lyrics were scraped from websites such as Genius and AZ Lyrics, and then put into a dataframe. These lyrics were then analyzed for sentiment (positive, negative, neutral) using sentiment analysis libraries like TextBlob. Using the YouTube API, we then collect popularity metrics such as: number of plays, popularity score, and the genre of the song. This provides two numeric features (number of plays, popularity score) and one categorical feature (genre) for analysis. To clean our data we removed any duplicates or irrelevant entries, standardized the sentiment analysis process using Python libraries, and handled multi-genre labeling by selecting a primary genre for each song.

When we created some initial visualizations using our data, we found that there were approximately 50 instances of a positive sentiment, approximately 17 instances of a negative sentiment, and approximately 5 instances of a neutral sentiment (Figure 1.1). The distribution was clearly skewed toward positive sentiments.

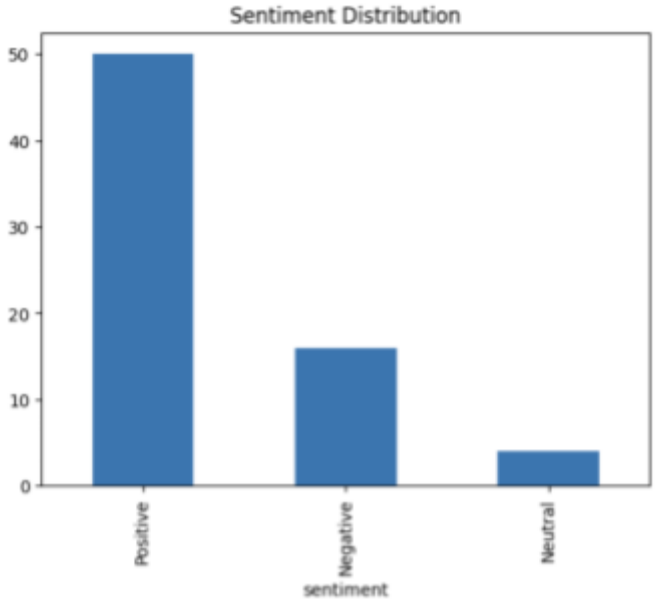

Figure 1.1

When we plotted the distribution of sentiment polarity scores we found that most songs tend to avoid extremes (Figure 1.2). However, overall, there are more songs that have a positive sentiment.

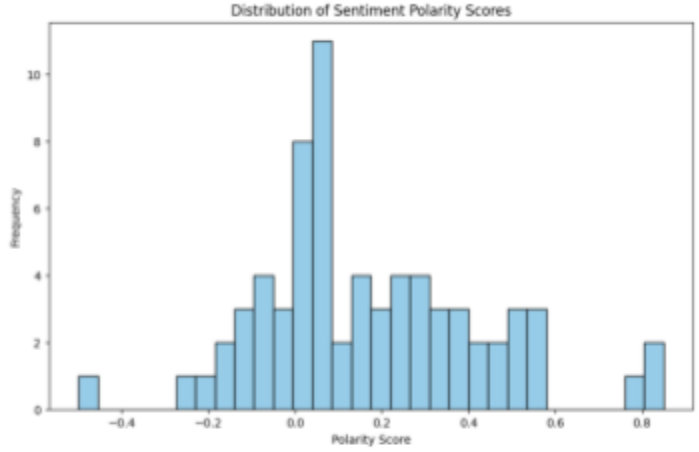

Figure 1.2

When we created a scatterplot of polarity vs views there was no clear correlation (Figure 1.3). Some content with neutral polarity had very high views, while some with higher sentiment had fewer views. Most of the data points with lower views are spread across different sentiment scores, while the highest-viewed content clusters around a neutral sentiment score.

Figure 1.3

Lastly, when we plotted comments vs likes we found that there was a positive correlation between amount of likes and amount of comments. However, the higher the variables, the more spread out.

Method

For a technical expert, the first model we chose to do linear regression using NumPy. This model is easy to read and interpret, making it a good choice to start out with. The target variable we chose is popularity as that showcases a song's success. The variable is a weighted sum of views, likes, and comments. The decision to implement linear regression manually using NumPy instead of a pre-built library like scikit-learn allows us to better understand and control the underlying mathematics, such as computing the line of best fit using the Normal Equation. However, there are inherent pitfalls with this method. Linear regression assumes no multicollinearity among features, but views, likes, and comments tend to be highly correlated. Nonetheless, this is addressed indirectly by the matrix inversion in the computation. Also, outliers in the data, such as a viral song, can skew the results easily.

For an application expert, linear regression is a straightforward machine learning method used to understand and predict relationships between variables. In our project we are exploring how views, likes, and comments relate to the popularity of k-pop songs. Using a weighted sum of these features, we defined a popularity score. Using a linear regression model is well-suited for our problem because the relationship is likely to be linear. Also, by implementing it manually using NumPy, and calculating the line of best fit using matrix operations. The model is evaluated with two metrics: MSE and R^2.

For a technical expert, the second model we chose to implement a polynomial regression using scikit-learn. This method is well-suited for our context as it allows us to capture non-linear relationships in the data that might be present due to the complexity of song lyrics. In the context of our project, polynomial regression offers a more adaptable method than linear regression in this situation, taking into account complex patterns and the subtle differences in lyrics. We think this approach will allow us to capture more nuanced associations that a linear model could overlook, given the inherent complexity of language present in song lyrics. However, no model is perfect. If the degree of the polynomial is too high, polynomial regression may result in overfitting since it thinks that our relationship, despite being non-linear, can be characterized by polynomial terms. In

order to combat this problem the degree for our model will be two. Lastly, while the model is good with complex data, which is usually the case when dealing with songs, it can struggle with generalizations.

For an application expert, polynomial regression is the appropriate model as it can help to find patterns between variables, which can help to estimate how likely a song is to be successful based on its lyrics. Polynomial regression gives us a greater chance of identifying the true associations in our K-pop lyric dataset by capturing variations in the data. This is particularly important in fields where interactions can be intricate and multi-layered, such as language and music. For artists and producers who want to improve lyrics for a higher audience appeal, this method provides easily comprehensible results that highlight the elements of a song's lyrics that can have the greatest impact on its success.

For a technical expert, the third model we chose to implement was a K-means clustering analysis to identify patterns in the data. We specifically focused on clustering based on song lyrics, and metrics such as views, likes, comments, and popularity. We will use the Elbow Method to determine the optimal amount of clusters and Principal Component Analysis for a better visualization which would help us lead to better interpretations in a contextual context.

K-means works well for this project as it clusters data based on similarity and we are looking for similarities between song features, such as lyrics, views, and likes. It's also good at handling multi-dimensional data which is often the case with song lyrics. It's also good for scalability which can be helpful in the future for expanding on our current results. The Elbow Method is good because it helped us identify the optimal number of clusters by examining the inertia. The goal was to find a point where adding more clusters results in diminishing returns. In the context of K-pop songs, we were interested in clustering songs into distinct groups that represent different popularity and engagement levels. The elbow method allowed us to determine the best number of clusters that revealed meaningful segments of songs. PCA is useful because it can reduce the dimensionality of the dataset while retaining as much variance as possible. We were able to visualize the results clearly and help us to identify which features are the most influential. However, no model is perfect. A pitfall of this method is the model's sensitivity to outliers. This can hurt the model if there is a viral song or something that strays from the norm.

For an application expert, the point of this project is to uncover how the features of K-pop lyrics correlate with a song's popularity. Because we have data that is structured and unstructured, it's important to explore patterns that may not be visible to the naked eye. K-mean clustering is the main method, and by doing this we could compare how songs with similar lyrical themes performed in terms of popularity. To determine the amount of clusters we used the elbow method, where we looked for a bend in the graph that tells us the amount of clusters to use in order to avoid choosing too many categories. Lastly, we used PCA which helped us to visualize the data while still being able to preserve the most important and meaningful information to gain the most valuable insights. As a result of the way the model is structured it's appropriate to use it in order to identify trends in K-pop lyrics and understand the factors that contribute to a song's success.

Results

The first model we chose is linear regression which can be measured with two metrics: MSE and R^2, as stated earlier. The MSE value of 6280763466496461suggests the regression line does not fit the data points well. The R^2 value indicates that only 2.07% of the variance in popularity is explained by the polarity. This low value shows that polarity is not a strong predictor of popularity.

The residual plot below shows the difference between the actual and predicted popularity values for a linear regression model (Figure 2.1). The residuals vary significantly which suggests that the model struggles to predict popularity for certain data points accurately. Also, because the data points are not randomly scattered around zero, the relationship between polarity and popularity may not be linear.

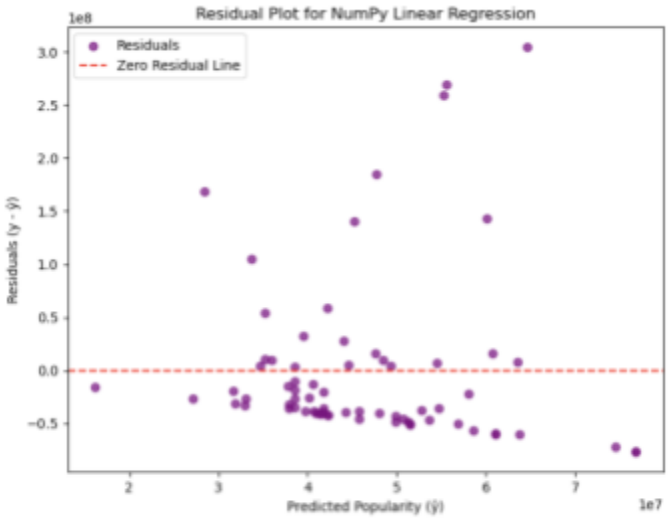

Figure 2.1

For this model we also created residual plots comparing the residual vs polarity and vs order. Both of these plots look similar to the main residual plot indicating that a linear regression model may not be the best way to model the data.

This image displays the results of a Polynomial Regression (Degree 3) model (Figure 2.2). The model achieved an R-squared value of 0.06013, indicating polarity explains 6.1 percent of all the variance in popularity scores. The Mean Squared Error (MSE) is relatively high at 6027869951187205.0, suggesting the model does not predict well. The Model Coefficients and Intercept represent the polynomial terms used in the regression equation. However this model does do a better job at explaining the variation between polarity and popularity than the linear regression model.

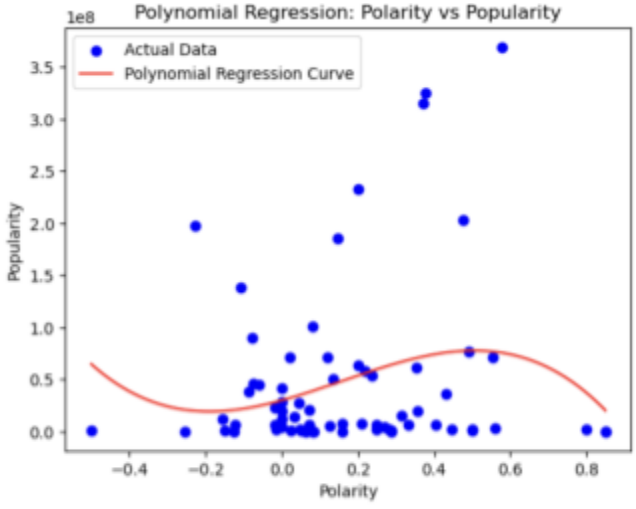

Figure 2.2

We also plotted the Residual Plot for the Polynomial Regression model with Degree 2 (Figure 2.3). The residuals (differences between actual and predicted values) are plotted against the predicted popularity scores. The residual plot shows the MSE in comparison to a linear regression model of 77463. The plot shows most of the residuals around the zero with no clear pattern. Therefore this model fits the data relatively well.

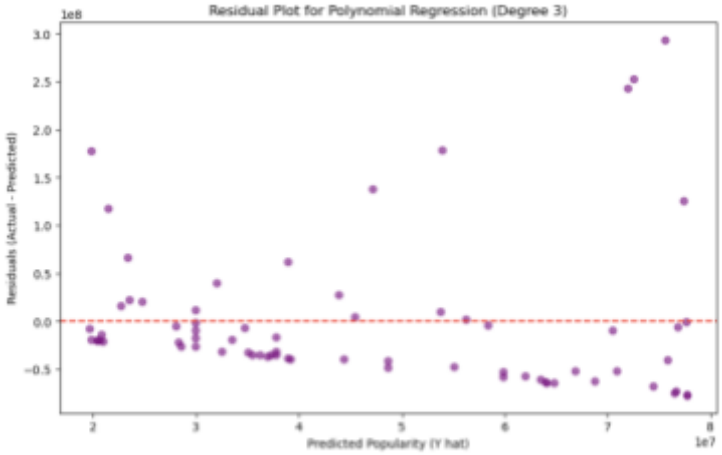

Figure 2.3

We used clustering to group songs by similar lyrical themes or emotional tones, utilizing this unsupervised learning approach to uncover hidden patterns in the data that might not be immediately apparent. The elbow plot models the inertia, or sum of squared distances from each point to its assigned cluster center, across a range of clusters from 1 to 10. The optimal number of clusters is determined by the elbow point, where inertia begins to level off. Based on the plot, three clusters are optimal, as the inertia drops sharply initially and levels out after the 3rd cluster. These clusters represent categories like "High Engagement," "Moderate Engagement," and "Low Engagement," with High Engagement songs having the highest values across all metrics, as shown in the summary table. Additionally, the PCA plot visualizes these clusters in two dimensions, with black "X" markers denoting the centroids, clearly distinguishing the profiles. These profiles can be leveraged to predict engagement patterns for new songs and develop marketing strategies tailored to each engagement level. This is all shown in Figure 2.4.

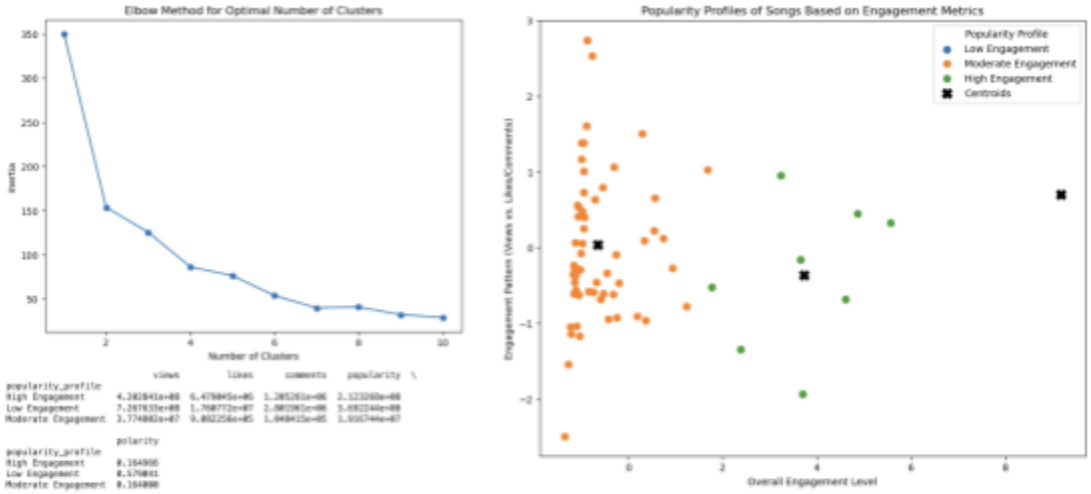

Figure 2.4

To analyze engagement ratios, we used clustering to group songs based on similar audience interactions. The elbow plot shows the inertia, or the sum of squared distances from points to their cluster centers, across a range of 1 to 10 clusters. The optimal number of clusters is determined by the elbow point, where the rate of inertia reduction slows significantly. Based on this, three clusters were chosen, representing "High Engagement," "Moderate Engagement," and "Low Engagement." The summary table highlights the engagement metrics for each cluster, with "High Engagement" songs achieving the highest likes per view and popularity, while "Low Engagement" songs have higher comments per view relative to “High Engagement.” The PCA plot visualizes these clusters in two dimensions, with distinct separation between profiles. Black "X" markers represent the centroids of each cluster, showing the average engagement characteristics for each group. This clustering approach helps identify patterns in audience interactions and can guide targeted marketing strategies for songs based on their engagement profiles. This is all shown in Figure 2.5.

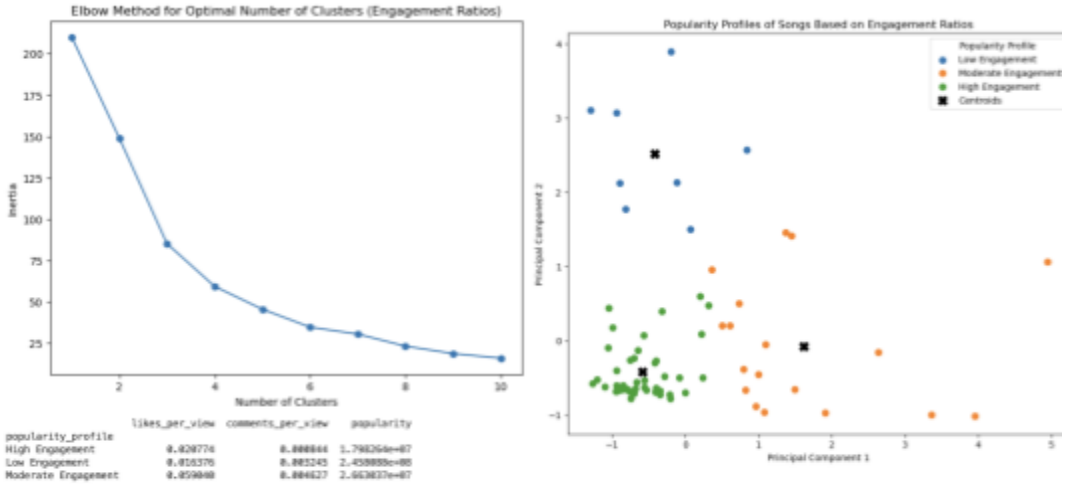

Figure 2.5

Discussion

The goal of our project was to analyze the relationship between K-pop lyrics and song popularity.
We initially went into this project with two main questions in mind:

- Are K-pop songs with positive lyrics more likely to achieve widespread fame? Has the inclusion of English lyrics contributed to their popularity?
- How have the themes and emotional tone of K-pop lyrics evolved over time?

After creating the models we found that, while highly viewed songs tended to cluster around neutral sentiment scores, there was no clear correlation between sentiment and views. We also found that including English lyrics tended to influence song appeal, but we couldn’t explicitly state a causational relationship. Our analysis didn’t really focus on themes over time, so if we were to expand on it at a later time, we could look at when the songs themselves were released.

While our models seemed reasonable, no model is perfect. The linear regression model may look over some of the complexities and intricacies in song lyrics. The same can be said about our polynomial regression model. It could potentially be improved by looking at more features. The clustering analysis at the end was helpful in providing actionable insights but can lack the ability to be generalized due to the metrics used. While the

statistics were good, the results should be accepted at face value with caution. To improve this model in the future we could look at more songs in order to confirm a generalization, or add more data. To build off of this, some limitations of our study is the dataset scope and the use of TextBlob, as it may not have been able to capture nuanced multilingual content.

Artists and producers could take action and include more English lyrics in their songs, but should avoid focusing solely on sentiment. For us in the future, an action we could take is looking at more variables such as where the song is streamed, and how long it is. While we are confident in these actions, again, we recommend proceeding with caution because there is a chance of overfitting.

Some unanticipated question that arose from the analysis were:

- What role do factors such as marketing and streaming platforms play in a song's popularity?
- How have the cultural shifts over time affected the inclusion of English lyrics?
- Is there a certain threshold of how many English lyrics a song can have before it negatively affects a song's success?

For future work, we would want to look at ways to answer the questions above, and dive deeper into one of our original questions about how the themes and emotional tone of K-pop has changed over time. We also think that it would be interesting to explore some more advanced machine learning models to see some of the non-linear relationships and to see if we would get the same results as some of the simpler models.