Statistical Analysis of Popular Albums

Alisha Foster
Andrew Diaz
Xiaoyu Gui

1. INTRODUCTION

Music taste is something so personal to an individual, yet music bridges differences and serves as a foundational pillar of culture, uniting people of all different backgrounds. Now– let's zoom out and look at how music is perceived and evaluated by society on a wide-scale. The patterns and trends not only inform us about society, but about aspects of our fundamental humanity. In our project, we are motivated by the question: ***What are the relationships between musical genres, emotional sentiment, the sex of vocalists, and societal reception of popular albums?***

Since writing our initial proposal, we have streamlined our research question to be more tailored to the specific relationships we are interested in uncovering, and we have directed our analyses accordingly. Specifically, we investigate similarities and discrepancies in album ratings by vocalists' sex, differences in sentiment values for albums by vocalists' sex and by genre, differences in mean ratings by sentiment value and by genre, and the trend of album ratings over the years. Furthermore, we construct a model with the goal of predicting albums' ratings from the aforementioned features.

2. DATA

2.1. Data Description

To investigate the questions outlined in the introduction, we sourced our data from the Rate Your Music: The Top 5,000 Most Popular Albums dataset available on Kaggle. This dataset, compiled from the popular online music database and community RateYourMusic.com, includes information on the 5,000 most rated albums as of March 11, 2022. The dataset comprises 11 variables and 5,000 observations before any filtering was applied.

This dataset contains information on the album's name, release date, primary and secondary genres, descriptors, release type, the artist's name, its position in the chart, its average rating, number of ratings and number of reviews. However, we removed the 'release_type' column as it uniformly contained the value 'album', offering no additional useful information. This cleaned dataset provides a comprehensive overview of highly-rated albums, including their ratings, genres, and other relevant descriptors, making it a valuable resource for analyzing popular music trends and characteristics.

2.2. Data Background

There have been a couple previous projects which have utilized this dataset before along with similar datasets within the realm of music analytics. We conduct exploratory data analysis in our project but in this specific project, they primarily focus on descriptive statistics, examining trends in music genres, artist popularity, and listener preferences. Another project places an emphasis on predictive analytics by answering the question of what album types will be successful in this day and age given the data from 2022. Additionally, there are a multitude of other datasets which highlight various popular music streaming platforms as their main sources of data, like Spotify and Apple Music, to conduct similar types of analysis as well.

In our project, we build upon these previous analyses by taking a unique approach to exploring the dataset. While some prior studies may have focused solely on descriptive or predictive analytics, our research combines elements of both, with a particular emphasis on the role of descriptors and genres in shaping music consumption. Our process of data collection and analysis allow us to execute various types of tests given the vast diversity of the information collected from this dataset.

2.3. Exploratory Data Analysis

Among all the variables, ‘position’, ‘rating_count’, and ‘review_count’ are discrete quantitative; ‘avg_rating’ is continuous quantitative, and all the others are categorical. Approximately 10% of data is missing in the ‘secondary_genres’ column, and only one value is missing in ‘primary_genres’. We decided to drop the column ‘secondary_genres’ because it is extraneous to our research, as ‘primary_genres’ contains plenty of information regarding the genres of the albums. Then, we dropped the single row containing a null value in the ‘primary_genres’ column. We’ve also extracted the year from the ‘release_date’ column because we are only interested in tracking trends over years. After these preprocessing steps, we dived deeper to each of the variables.

2.3.1. Artist

We found that Bob Dylan is the artist with the highest number of albums in this dataset. This highlights his prolific nature and enduring popularity, as he appears more frequently than any other artist.

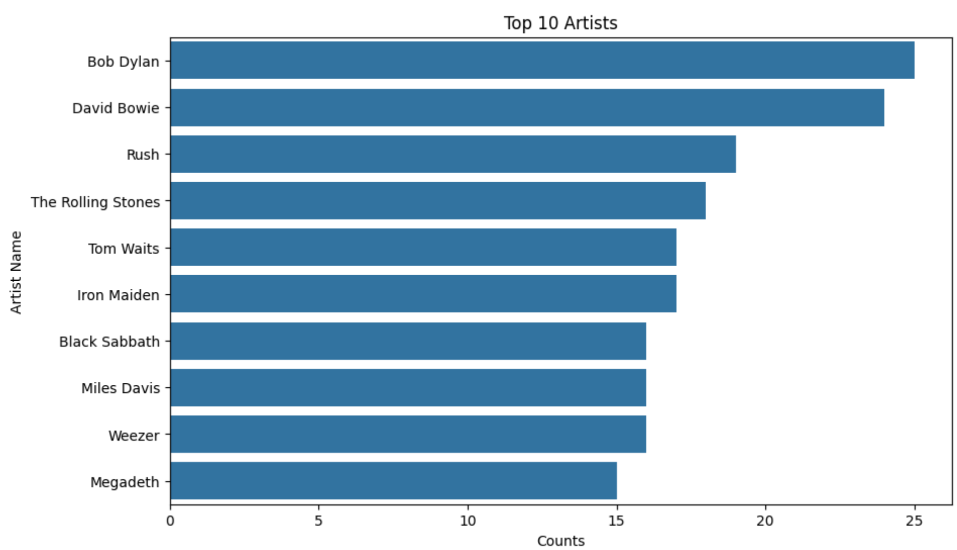

Figure 1: bar chart of the counts of the top 10 most frequently occurring artists

2.3.2. Genre

We found that Singer-Songwriter is the most common genre in this dataset, with Alternative Rock close behind in second place. This indicates a preference among users for these introspective and versatile musical styles.

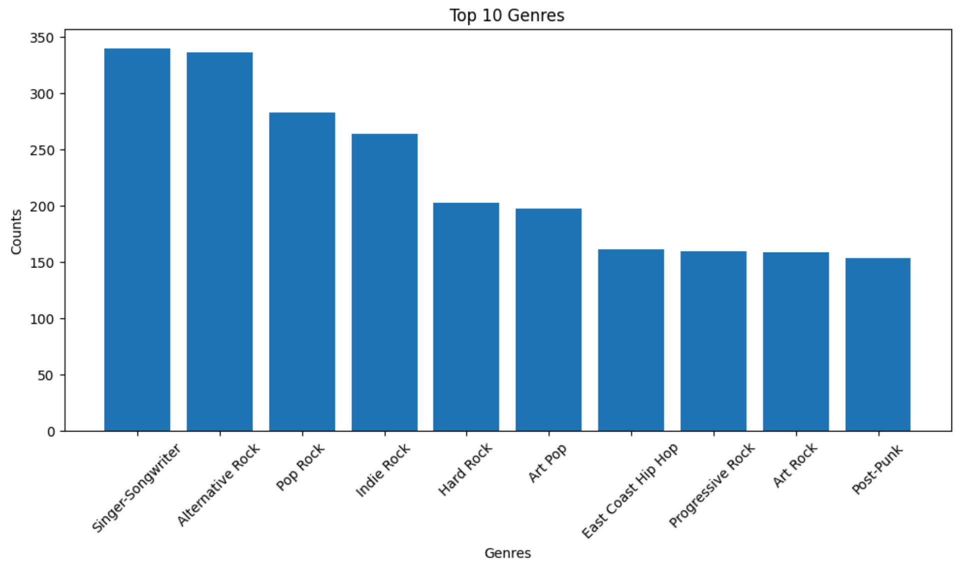

Figure 2: bar chart of the counts of the top 10 most frequently occurring genres

2.3.3. Descriptors

In this dataset, 'descriptors' refers to words assigned to albums that describe the musical body with respect to lyrical themes, melodic tendencies, tone, and overall emotional output. We found that the descriptor 'malevocals' has by far the most occurrences in this dataset, reflecting the dominance of male vocalists in popular music.

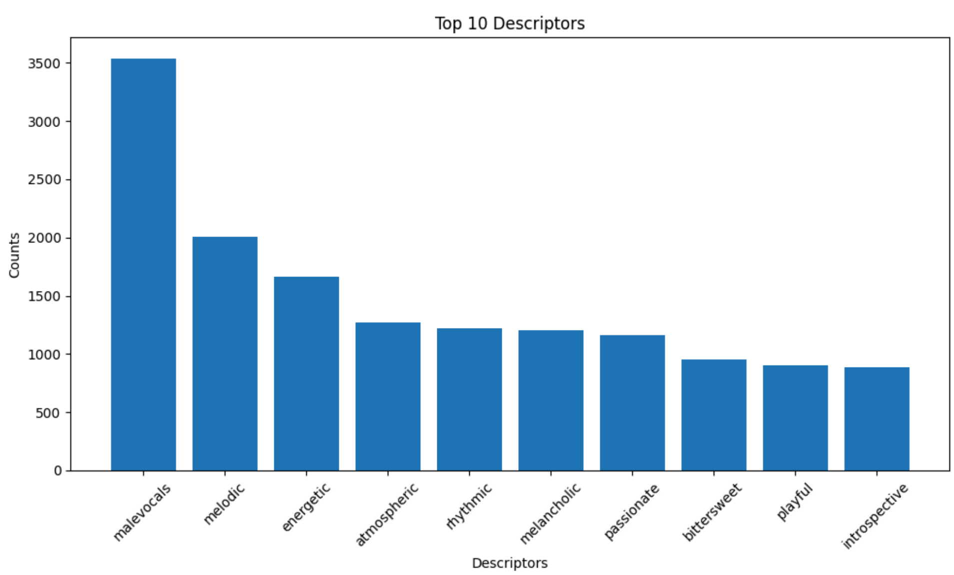

Figure 3: bar chart of the counts of the top 10 most frequently occurring descriptors

2.3.4. Gender of Vocalists

Of the descriptors in this dataset, two are of particular interest to us: ‘malevocals’ and ‘femalevocals’. We use these descriptors to analyze the differences between albums with male vocals versus female vocals in terms of musicality as well as societal reception.

To make our analysis easier, we decided to create a new column in the dataframe called ‘vocals’ and encode 0 to represent neither male nor female vocals, 1 to represent only male vocals, 2 to represent only female vocals, and 3 to represent both male and female vocals.

Before analyzing the data, we didn’t have expectations about the musical differences between albums with male vocals and albums with female vocals. However, we did have expectations about the differences in terms of the number of occurrences in the dataset and societal reception. We expected that ‘malevocals’ would be overrepresented and that ‘malevocals’ would correlate with higher average ratings, as a result of societal prejudice.

We found that the dataset indeed has much more albums with ‘malevocals’ than ‘femalevocals’. In fact, there are more albums with no vocals at all than albums with ‘femalevocals’ (Table 1).

Vocal type	**Number of Albums**
Neither Male or Female Vocals (0)	852
Male Vocals (1)	3346
Female Vocals (2)	607
Both Male or Female Vocals (3)	194

Table 1: distribution of albums across different vocal types. The numbers behind each vocal type are the encodings in our dataset.

According to figure 4, the most prevalent genre in albums with male vocals is Alternative Rock, followed by Pop Rock and Singer-Songwriter. This distribution suggests a strong presence of rock and related genres among albums with male vocals. On the other hand, the most prevalent genres in albums with female vocals are Singer-Songwriter and Art Pop, both with similar frequencies. This distribution indicates a preference for more diverse and modern genres among albums with female vocals, with a noticeable presence of pop and alternative styles.

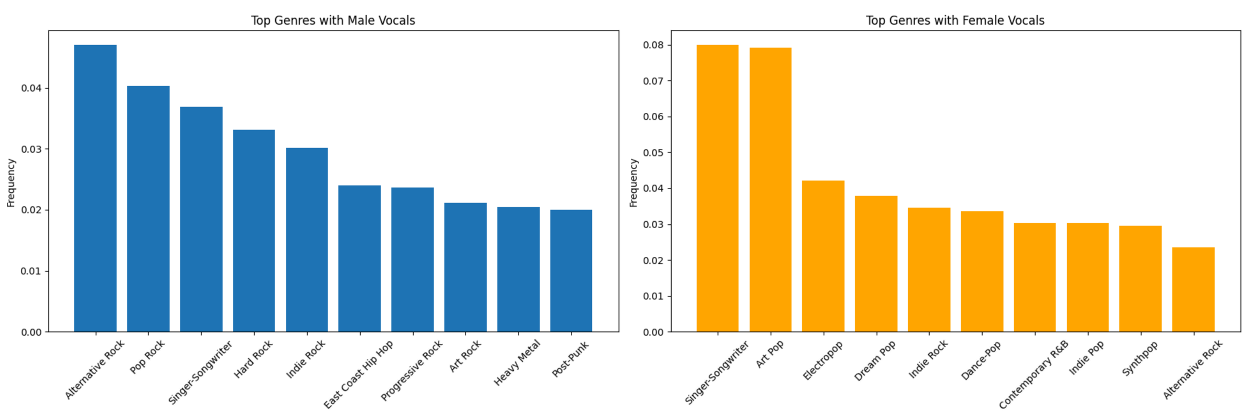

Figure 4: Top 10 Genres in albums labeled with Male and Female Vocals

According to figure 5, the most common descriptors in albums with male vocals are ‘melodic’ and ‘energetic’, indicating a focus on tuneful and lively music. This suggests that albums with male vocals often emphasize energetic and rhythmic elements, with a significant number also being described as introspective or heavy. The most common descriptors in albums with female vocals are ‘melodic’ and ‘love’, indicating a focus on tuneful music and romantic themes. This suggests that albums with female vocals often emphasize melodic and romantic elements, with a notable presence of emotive and lush descriptors.

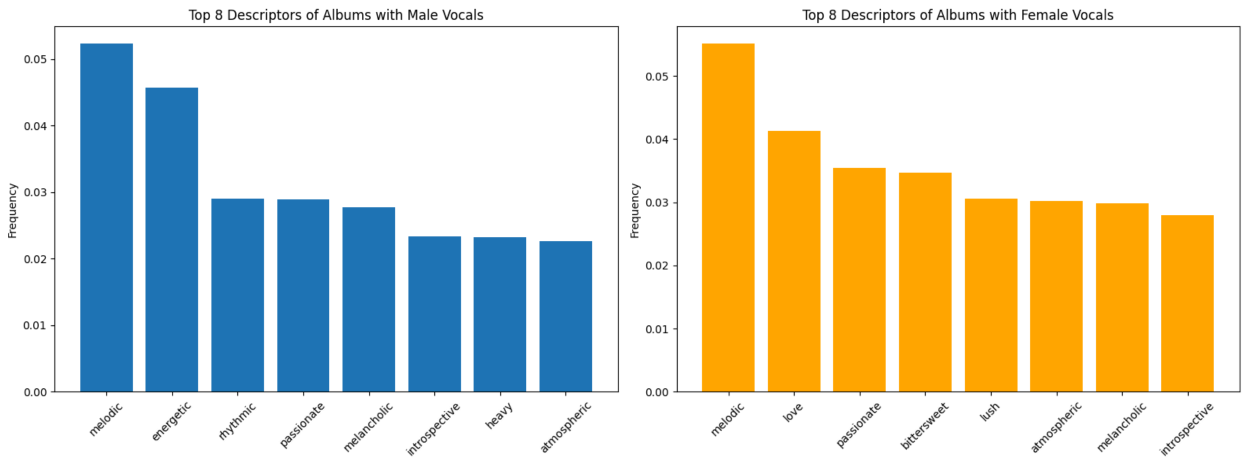

Figure 5: Top 8 Descriptors in albums labeled with Male and Female Vocals

2.3.5. Release Year

Before analyzing the data, we expected that release dates will range from the 1960’s to the 2020’s, with a skew towards the more recent decades. This is somewhat similar to our findings. We found that the recent decades indeed have more entries in this dataset, and there are many peaks and valleys in our plot of release years vs. number of observances. Figure 6 below demonstrates the number of albums released each year in our dataset.

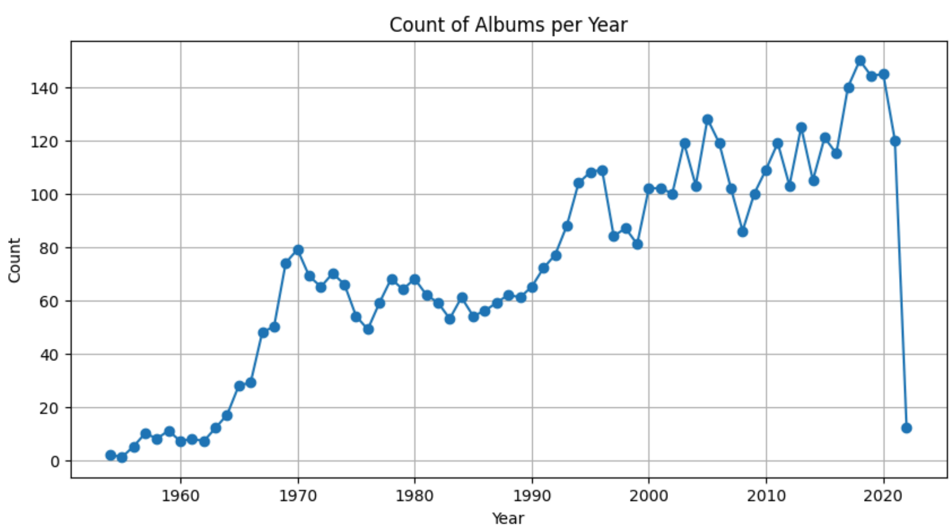

Figure 6: Line chart of Count of Albums per Year

We've also investigated the annual count of albums categorized by vocal type -- male and females -- from the 1960s through 2020. It highlights notable trends in the music industry over several decades. Albums featuring male vocals have historically outnumbered those with female vocals, although the gap has narrowed significantly since the 2000s.

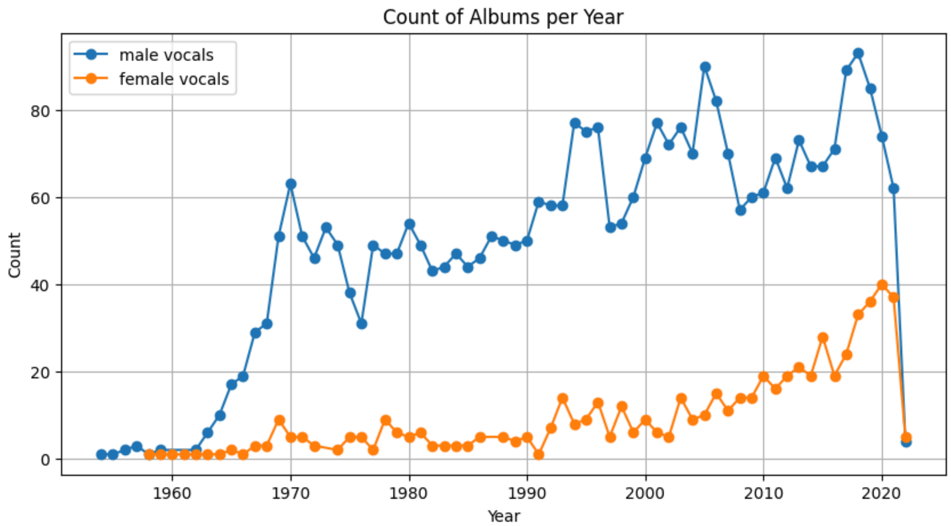

Figure 7: Line chart of count of albums per year categorized by male and female vocals

2.3.6. Average Rating

The average ratings in our dataset span from 0.62 to 4.34 with a mean of 3.54 and a median of 3.62. The distribution of the album rating is skewed to the left.

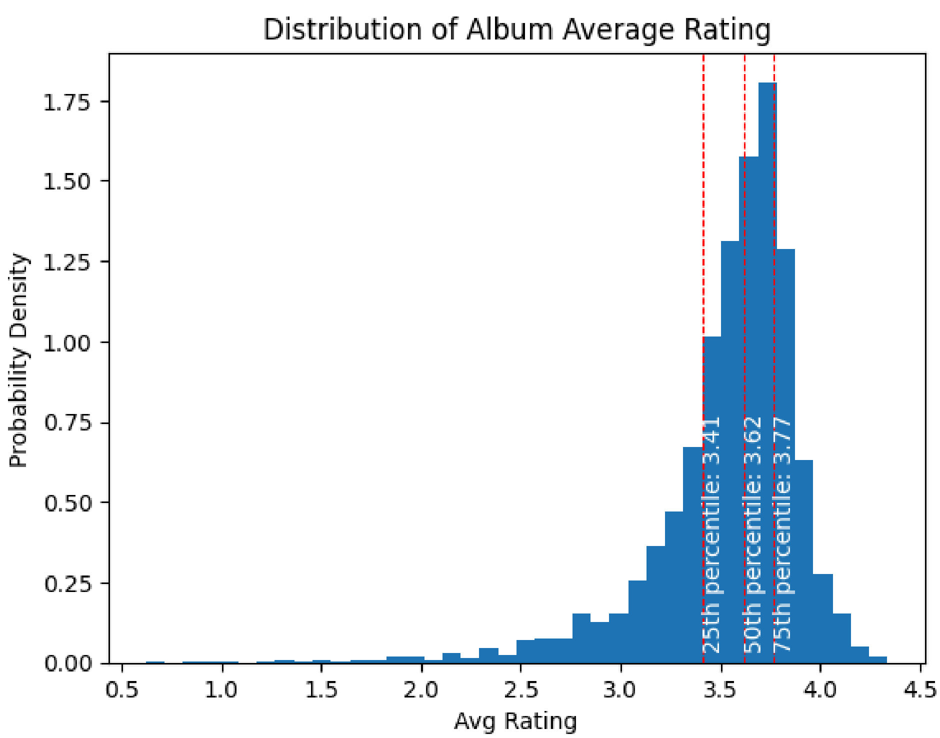

Figure 8: Distribution of Album Ratings with 25th, 50th and 75th percentile

Then, we analyze the trend of the average ratings of all albums for each release year. Interestingly, we found that earlier release years tend to have higher ratings. We believe that people may feel nostalgic towards older albums, contributing to the trend that we see.

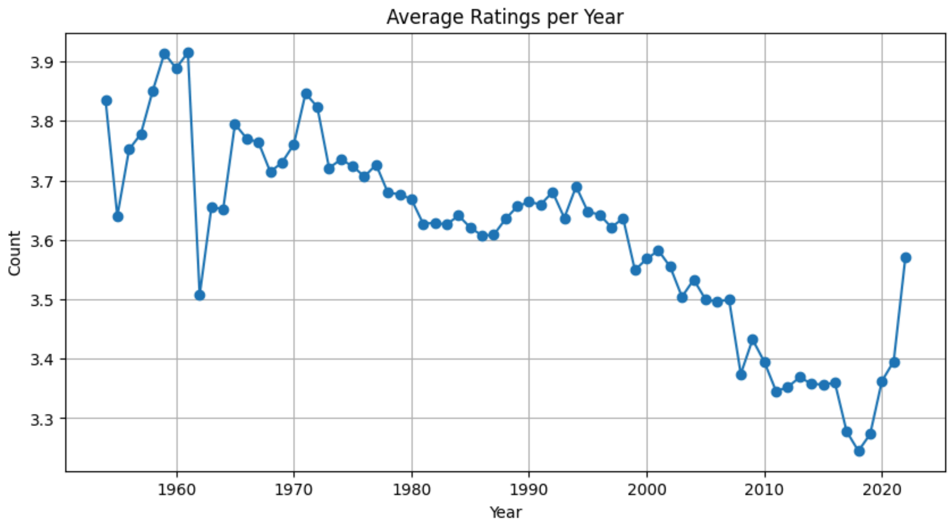

Figure 9: Line charts of Album Average Ratings by Release Year

In addition, we explored the average rating of albums by genre. Picopop is the highest-rated genre with an average rating of 4.02.

Genres	Average Rating
Picopop	4.02
Romanticism	4.02
Orchestral	4.01
Samba Soul	4.01
Video Game Music	4.0
Television Music	4.0
Dance	4.0
Samba-rock	3.99
Lounge	3.98
Afrobeat	3.97

Table 2: Top 10 rated genres and their average ratings

2.3.7. Review count & Rating count

In figure 10, we look at the variables `review_count` and `rating_count` which represent the number of reviews and ratings each album has, respectively. Both distributions are skewed to the right.

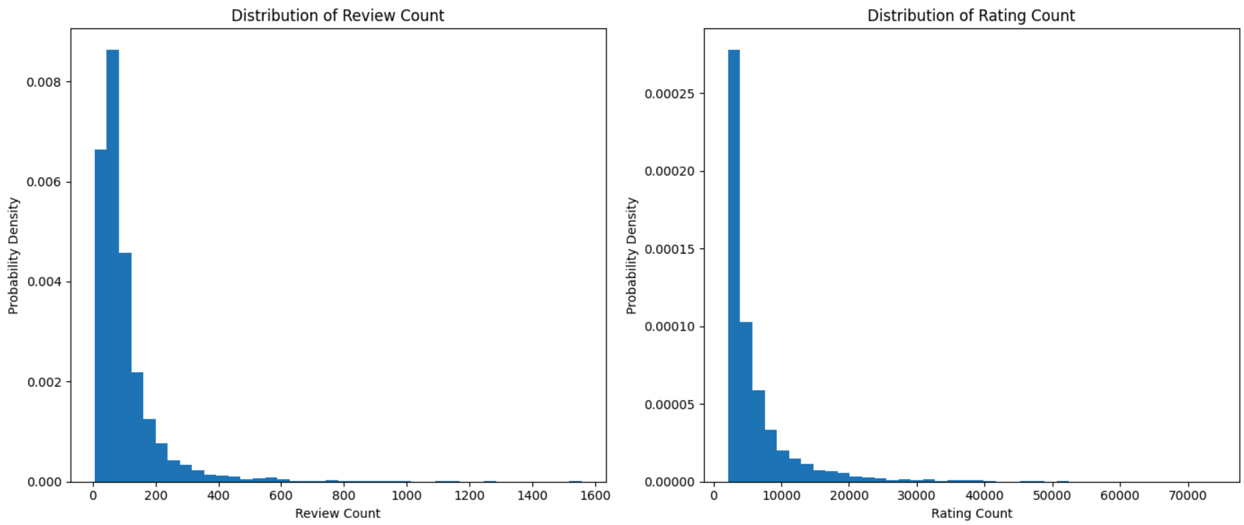

Figure 10: Distribution of Review Count and Rating Count

We noticed that there appears to be a positive correlation between review count and rating count. Albums with more ratings tend to have more reviews.

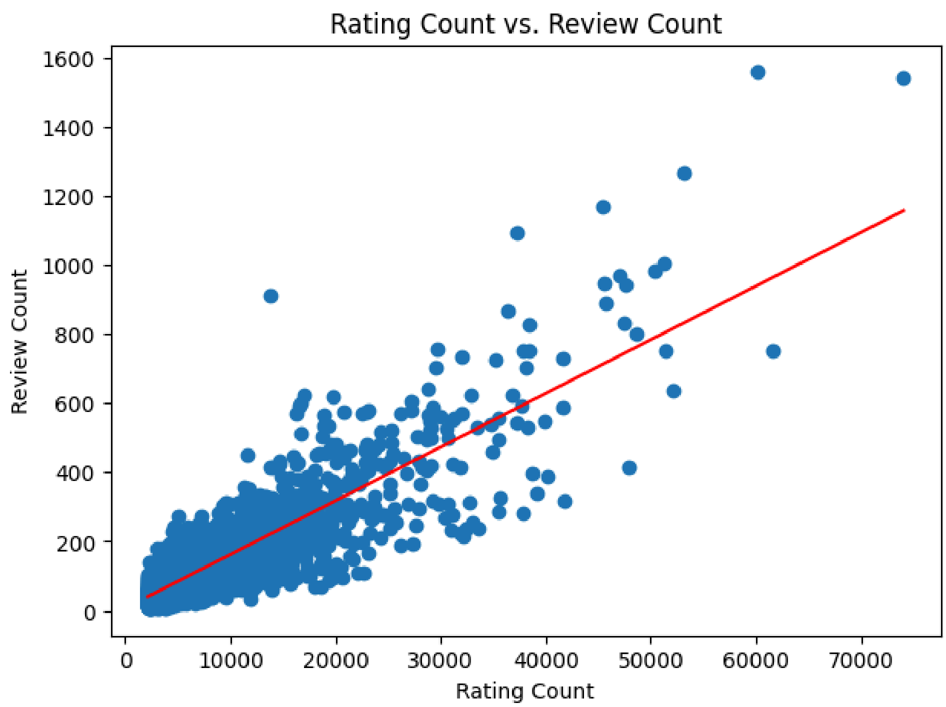

Figure 11: Scatterplot of Rating Count vs. Review Count

3. ANALYSIS

3.1. Sentiment Analysis

In today's digital age, understanding the sentiment behind textual data is of great importance across various industries. Sentiment analysis offers insights into the emotional tone and attitudes expressed within textual content and not only are we able to learn about the overall sentiment trends surrounding albums but also identify specific genres or descriptors that evoke strong emotional responses. With this knowledge, artists and labels can make decisions, based on the data, to enhance the appeal of their albums and elicit emotions from specific audiences. Using this technique, we can grasp a better understanding of what genres or types of albums certain audiences are more attracted to.

Exploring the factors influencing genre popularity across various time periods sheds light on the ever-evolving music space. For instance, genres like jazz and funk thrived during the upbeat 70s, while the 90s witnessed a surge in alternative rock and rap, often characterized by more intense and assertive tones. Depicted in figure 12, we have the average sentiment values of albums from each specific year where we can see relations between these changes as well. These values were calculated from the polarity scores of our Sentiment Intensity Analyzer with the nltk library and used to visualize the changes within the music industry during specific years.

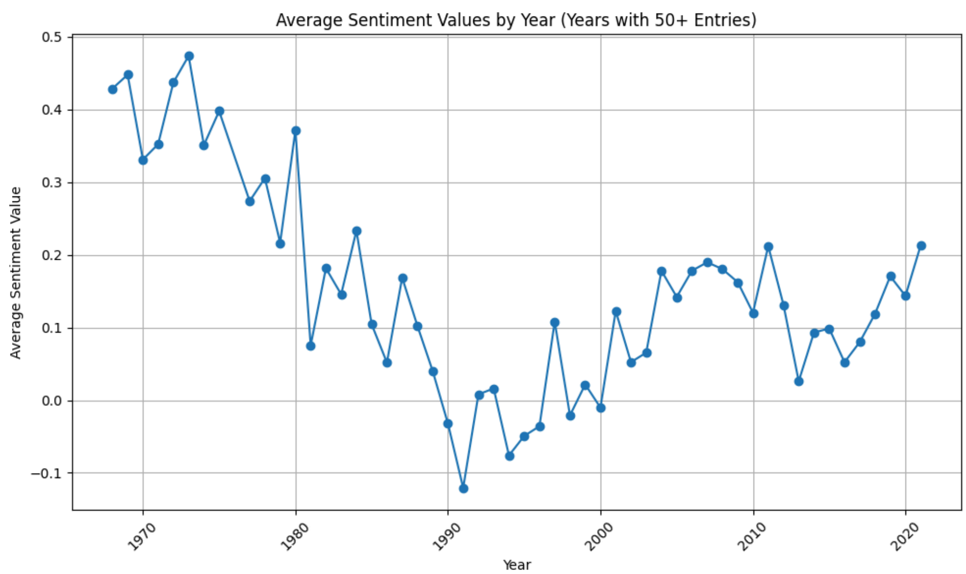

Figure 12: Line Graph of Average Sentiment by Years (Years with 50+ Entries)

The sentiment values found within our new sentiment_df data frame will be used for calculation in our hypothesis testing where we answer what genres have similar mean values and also utilized in our predictive models for further analysis.

3.2. Hypothesis testing

In this section, we used statistical analysis to uncover quantitative associations between sociological factors, musical components, and emotional characteristics. We tested several hypotheses regarding similarities and discrepancies in album ratings by vocalists' sex, differences in sentiment values for albums by vocalists' sex and by genre, differences in mean ratings by sentiment value and by genre, and the trend of album ratings over the years.

3.2.1 Is there a difference in the means of the ratings for albums with male vocals and albums with female vocals?

When we take the mean of 'avg_rating' for the albums with male vocals, we get 3.5155, and for the albums with female vocals, we get 3.5099. Is this difference statistically significant, at the 5% significance level? To answer this question, we conducted a two-sample t-test for difference of means, testing the following hypotheses H_0: the means of ratings for albums with male vocals and albums with female vocals are the same versus H_a: the means are not the same.

This procedure assumes that the data from each sample is normally distributed and independent between and within samples. We have independence between samples, since we organized the data in such a way that the categories 'male vocals' and 'female vocals' are mutually exclusive. We have independence within samples, since the ratings of albums in this dataset are not dependent on the ratings of other albums. As for the normality assumption, our sample sizes, 3346 for male vocals and 607 for female vocals, are sufficiently large to apply the Central Limit Theorem to, and assume that these samples are approximately normally distributed.

The hypothesis test resulted in a probability value of approximately 0.696, indicating that the probability of the data producing the results that it did or more extreme under the null hypothesis is about 69.6%. Thus, we fail to reject the null hypothesis. *Therefore, we conclude that there is not a statistically significant difference between the means of ratings of albums with male vocals and albums with female vocals.*

3.2.2 Are the distributions of the ratings of albums with male vocals and the ratings of albums with female vocals the same?

We have concluded that the means are the same, but we are interested in what other information we can learn about how vocalists' sex and albums ratings are related. Now, we will see whether the ratings are distributed equally for albums with male vocals and albums with female vocals. Below, we visualize the ratings of albums for vocalists of each sex:

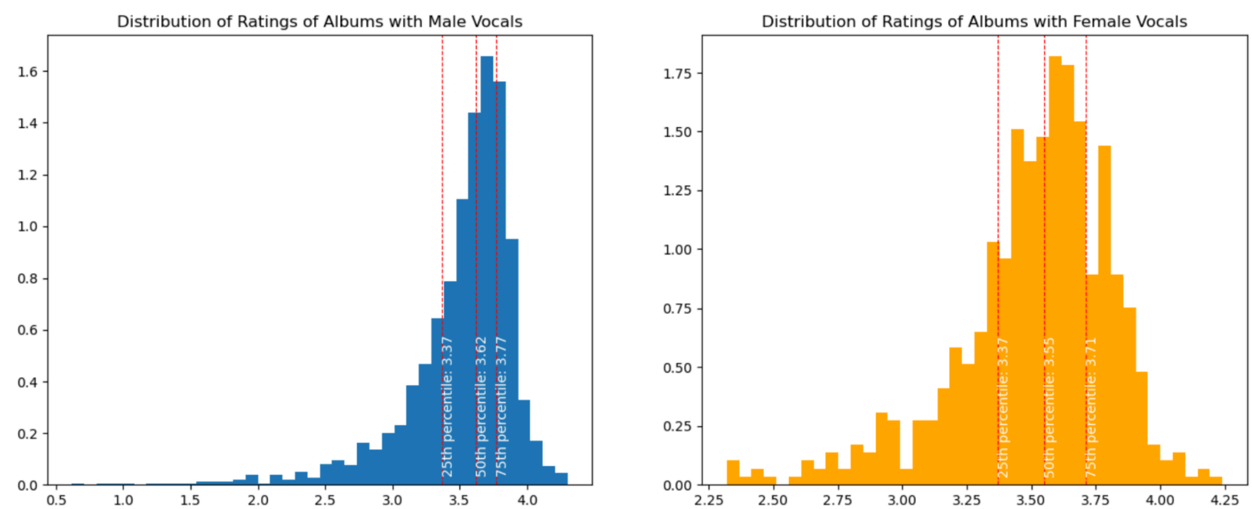

Figure 13: Distributions of the ratings of albums with male vocals (left) and albums with female vocals (right)

In order to answer our question, we conduct a two-sample Kolmogorov-Smirnov Test. This procedure assumes that the ratings of albums with male vocals come from a distribution, and the ratings of albums with female vocals come from a distribution. This is an easy assumption to satisfy. We use the Kolmogorov-Smirnov Test to test our hypotheses H_0: the distributions of ratings of albums with male vocals and albums with female vocals are the same versus H_a: the distributions are not the same.

The hypothesis test resulted in a probability value of approximately 0.00001, indicating that the probability of the data producing the results that it did or more extreme under the null hypothesis is about 0.001%. Thus, we reject the null hypothesis. *Therefore, we conclude that the distributions of the ratings of albums with male vocals and the ratings of albums with female vocals are not the same.*

3.2.3 Is the distribution of sentiment value of the descriptors of albums with male vocals and albums with female vocals the same?

We are interested to find out whether there is a difference in sentiment values between albums with male vocals and albums with female vocals. So, we analyze the distributions of the sentiment values of the descriptors of the albums. Below, we visualize the sentiment values of albums for vocalists of each sex:

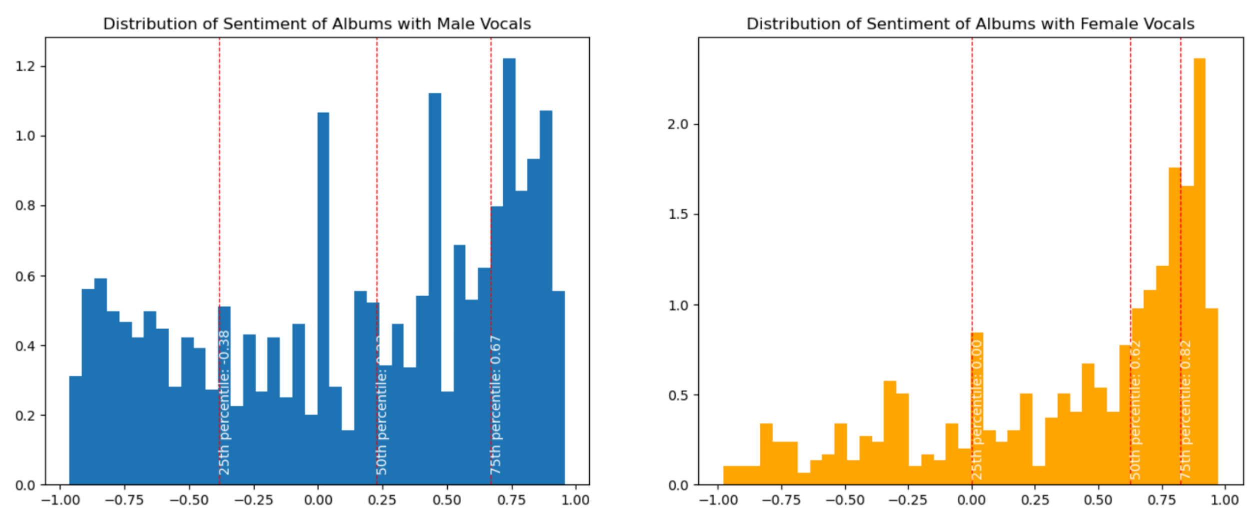

Figure 14: Distributions of the sentiment values of albums with male vocals (left) and albums with female vocals (right)

Just from looking at the figure, we can see that albums with female vocals tend to have higher sentiment value. Is the difference in distribution statistically significant, at the 5% significance level? To answer this question, we run another two-sample Kolmogorov-Smirnov Test. Again, the assumptions are that the two samples each come from a distribution are satisfied.

The hypothesis test resulted in a probability value of approximately 8×10^{-22}, indicating that the probability of the data producing the results that it did or more extreme under the null hypothesis is extremely small. Thus, we reject the null hypothesis. *Therefore, we conclude that the distributions of the sentiment values of albums with male vocals and the sentiment values of albums with female vocals are not the same.*

3.2.4 Is there a statistically significant downward trend in the average ratings over the years?

While exploring our data, we noticed that the average ratings tend to go down over time. Is this downward trend statistically significant, at the 5% significance level? To answer this question, we fit a linear regression model that predicts ratings from years, and we test the hypotheses H_0: the slope of the regression line is equal to zero versus H_a: the slope is less than zero. Below is a plot depicting the regression line through the data:

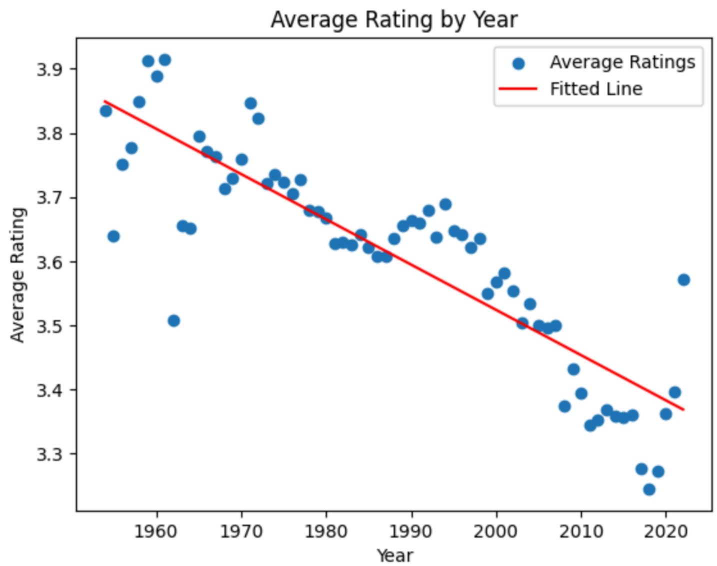

Figure 15: Scatter plot of average rating versus year, with a best fit line

Our regression model resulted in a probability value of approximately 7×10^{-22}, indicating that the probability of the data producing the results that it did or more extreme under the null hypothesis is extremely small. Thus, we reject the null hypothesis. *Therefore, we conclude that there is a significant downward trend in average ratings over the years.*

3.2.5 Are the sex of the vocalist and the rating of albums independent?

To answer this question, we conduct a Test for Independence. This procedure is used for two categorical variables, so we created the vocalist categories (‘no vocals’, ‘male vocals’, ‘female vocals’, and ‘both vocals’) and organized the ratings into bins of quartiles, indicating for each album which quartile its rating falls into. Therefore, our procedure reasonably satisfies the assumption that vocalist sex and album rating quartiles come from a multinomial distribution, the assumption required by the Test for Independence. So, we tested the hypotheses H_0: vocalist sex and rating quartile are independent versus H_a: they are not independent, at the 5% significance level.

Our test resulted in a probability value of approximately 6×10^{-24}, indicating that the probability of the data producing the results that it did or more extreme under the null hypothesis is extremely small. Thus, we reject the null hypothesis. *Therefore, we conclude that the sex of the vocalist and the rating quartile of albums are not independent.*

3.2.6 Which genres have the same mean ratings?

In the dataset, each album has a list of specific genres that describe it. In order to compare across broader genres that everyone is familiar with, we group albums by whether their `genres` variable contains 'pop', 'rock', 'metal', 'jazz', 'hip hop', 'country', or 'ambient'.

To answer this question, we do multiple hypothesis testing where we test for difference of means in ratings between each pair of genres in our list using a two-sample t-test. In order for our type I error to stay equal to 0.1, we apply the Bonferroni Correction to our p-values.

Our multiple hypothesis test resulted in failing to reject the null hypothesis for 6 out of the 21 genre pairings, which are listed below:

- pop and hip hop
- rock and metal
- rock and hip hop
- metal and country
- metal and ambient
- country and ambient

Therefore, we conclude that for the above genre pairings, the mean ratings are equal. For all other combinations of genres in our list, we reject the null hypothesis, concluding that their mean ratings are not equal.

3.2.7 Which genres have the same mean sentiment values?

To answer this question, we conducted a procedure similar to 3.2.6, using multiple hypothesis testing with two-sample t-tests and Bonferroni Correction. However, we tested the difference between the means of each genre's sentiment values, rather than ratings.

Our multiple hypothesis test resulted in failing to reject the null hypothesis for 6 out of the 21 genre pairings, which are listed below:

- pop and jazz
- pop and country
- rock and country
- metal and hip hop
- jazz and country
- country and ambient

Therefore, we conclude that for the above genre pairings, the mean sentiment values are equal. For all other combinations of genres in our list, we reject the null hypothesis, concluding that their mean sentiment values are not equal.

3.2.8 Which sentiment value ranges have different mean ratings?

To answer this question, we conducted a procedure similar to 3.2.6, using multiple hypothesis testing with two-sample t-tests and Bonferroni Correction. However, we tested between different sentiment value ranges instead of between genres. In order to do so, we labeled each album in our dataset as either 'strongly negative', 'somewhat negative', 'slightly negative', 'neutral', 'slightly positive', 'somewhat positive', or 'strongly positive' based on the sentiment value of the album's descriptors.

Our multiple hypothesis test resulted in rejecting the null hypothesis for 2 out of the 21 sentiment range pairings, which are listed below:

- moderately negative and strongly positive
- neutral and strongly positive

Therefore, we conclude that for the above sentiment range pairings, the mean ratings are different. For all other combinations of sentiment ranges in our list, we fail to reject the null hypothesis, concluding that their mean ratings are equal.

3.3. Predictive Modeling

The music industry increasingly relies on data-driven insights to navigate its dynamic and competitive landscape. To this end, our analysis included the predictive modeling of ‘avg_rating’ for albums, aiming to uncover the factors that significantly influence these ratings. Understanding these factors helps music producers, artists, and marketers make informed decisions, tailoring their strategies to better meet audience preferences and enhance the overall appeal of their music.

3.3.1. Data Preprocessing

Our analysis began with a thorough data preprocessing phase, where categorical variables such as genre and vocalist type were transformed using one-hot encoding. This step was crucial for fitting machine learning models, allowing them to interpret categorical data effectively. We also addressed potential multicollinearity – where independent variables are highly correlated – by dropping specific redundant features like ‘review_count’, which closely paralleled ‘rating_count’.

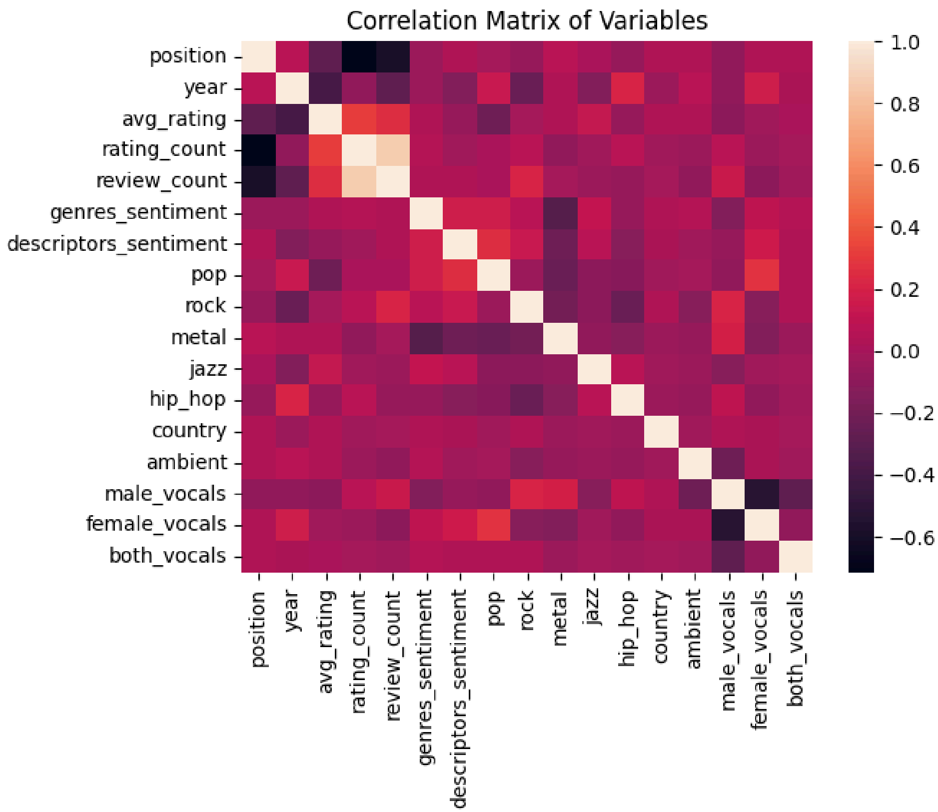

Figure 16: Correlation Matrix of Covariates available in the dataset. The lighter color is associated with higher correlation.

We then split the data into training and testing sets with a proportion of 8:2, maintaining a balance that would allow our models to learn comprehensively from the training data while also being validated against an unbiased testing set.

3.3.2. Baseline Model and F-Test for Significance

The statistical foundation of our project was set by constructing a baseline linear regression model, focusing initially on a select group of predictors: year, rating count, and sentiments derived from genres and descriptors. This model helped establish a preliminary understanding of how these factors might influence album ratings.

Further, we tested the assumptions necessary for a robust linear regression analysis, ensuring that conditions such as linearity, independence, and homoscedasticity were met. In addition, we tested for the significance of coefficients in this model, using the following hypotheses:

- Null Hypothesis (H_0): All coefficients in the model, excluding the intercept, are equal to zero. This suggests that none of the predictors have a statistically significant effect on the response variable.
- Alternative Hypothesis (H_a): At least one of the coefficients in the model, excluding the intercept, is not equal to zero. This implies that at least one predictor has a statistically significant effect on the response variable.

Statistics	Value
R-squared	0.247
Adjusted R-squared	0.246
F-statistic	326.8
Prob (F-statistic)	1.36e-243
Durbin-Watson	1.999
No. Observations	3999
DF Residuals	3994

Table 3: OLS Regression Results of Baseline Model - Model Summary Statistics

Predictor	Coef	Std Err	t-Value	P>|t|	95% CI
Intercept	20.7276	0.627	33.049	0.000	[19.498, 21.957]
year	-0.0087	0.000	-27.583	0.000	[-0.009, -0.008]
rating_count	0.00001803	0.000000871	20.693	0.000	[0.0000163, 0.0000197]
genres_sentiment	0.0362	0.039	0.929	0.353	[-0.040, 0.112]
descriptors_sentiment	0.0712	0.009	-7.912	0.000	[-0.089, -0.054]

Table 4: OLS Regression Results of Baseline Model - Coefficients

The insights we obtained from the baseline model summary:

- R-squared (0.247): This value indicates that approximately 24.7% of the variability in average ratings is explained by the model. This isn't particularly high, suggesting the model captures some, but not all, factors influencing ratings.
- Coefficients:
    - 'year' (-0.0087): This coefficient is negative, indicating that as the year increases, the average rating slightly decreases, holding all else constant. Its p-value (< 0.0001) suggests this is statistically significant. This matches our previous findings of decreasing ratings as time goes by.
    - 'rating_count' (0.00001803): Positive coefficient, statistically significant, indicating a slight increase in average rating with an increase in rating count.
    - 'genres_sentiment' (0.0362): Not statistically significant (p-value = 0.352), meaning changes in genre sentiment don't reliably predict changes in average rating in this dataset.
    - 'descriptors_sentiment' (-0.0713): This is significant and negative, suggesting that more negative sentiment in descriptors correlates with lower ratings.
- F-statistic: The model is statistically significant as a whole, with a very low probability (near zero) that the observed results are due to chance.

- Durbin-Watson (1.999): This value, close to 2, suggests that there is no serious autocorrelation in the residuals.

Based on the F-statistic and its p-value, we conclude that there is compelling evidence to ***reject*** *the null hypothesis and that the predictors in the model collectively have a significant effect on the dependent variable, `avg_rating`.*

3.3.3. Advanced Full Model and Model Comparison

The baseline model's results revealed that while it captured some influential factors (such as a negative trend in ratings over time and a positive influence of rating counts), it only explained about 24.7% of the variability in album ratings.

To deepen our insights, we extended our analysis to a full model that incorporated additional predictors, excluding only the 'review_count' to avoid multicollinearity. We've also conducted an anova test to determine whether adding these predictors would significantly improve the model fit with the following hypotheses:

- Null Hypothesis (H_0): The reduced model is adequate, and the additional predictors in the full model do not provide a significantly better fit to the data.
- Alternative Hypothesis (H_a): The full model, with the additional predictors, provides a significantly better fit than the reduced model.

df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
3994.0	414.0662	0.0	N/A	N/A	N/A
3983.0	375.3964	11.0	38.6698	37.2992	5.1133e-77

Table 5: ANOVA results comparing the baseline model with the full model. The significant decrease in SSR and the high F-statistic indicate a substantially better fit in the full model.

The insights we obtained from the this comparison summary:

- Degrees of Freedom (df_resid and df_diff): The full model has 11 more parameters (predictors) than the baseline model, as indicated by df_diff = 11.
- Sum of Squares (ssr and ss_diff): The sum of squared residuals decreases from 414.07 in the baseline model to 375.40 in the full model. The difference in sum of squares (ss_diff) is 38.67, indicating an improvement in fit by the full model.
- F-statistic (F): The F-statistic is 37.30. This statistic is used to compare the two models. Higher values generally indicate a more significant difference in model fits.
- p-value (Pr(>F)): The p-value is extremely small (5.11e-77), which is much less than the typical significance level of 0.05.

Since the p-value is very small, we ***reject*** *the null hypothesis that the additional predictors do not improve the model. This indicates that adding the extra predictors significantly improves the model fit.*

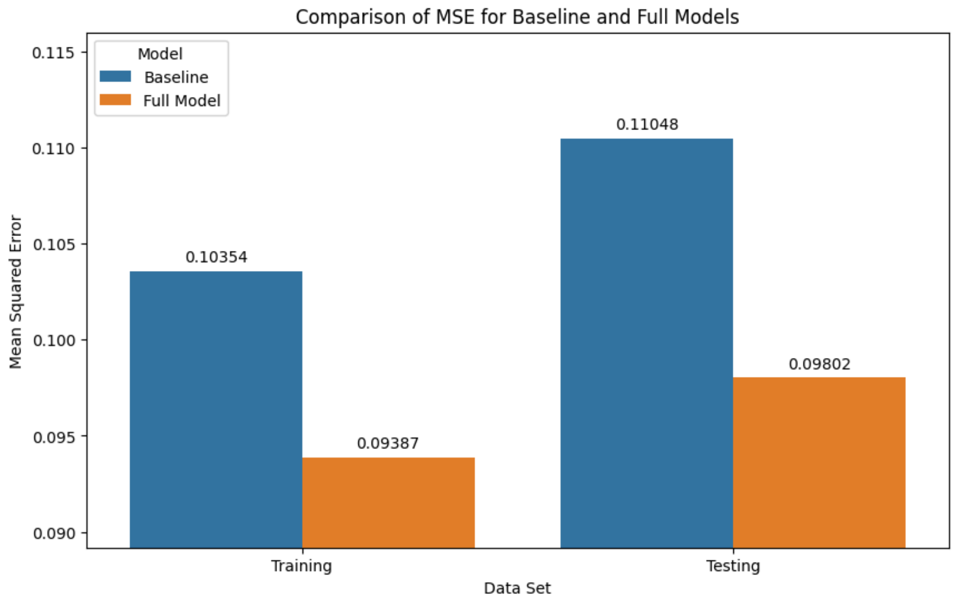

Figure 17: Comparison of Model Performance on Training and Testing Sets

The full model not only lowered the MSE on both the training and testing sets but also provided a comprehensive view of the impact of various predictors on album ratings. This model comparison step showed a significant improvement in fit when we add the additional predictors, as evidenced by a decrease in mean squared error (MSE) and an increase in the explanatory power of the model.

3.3.4. Conclusion

Through data preparation, model selection, and rigorous statistical validation, our analysis successfully identified key factors affecting album ratings. This predictive modeling serves as a valuable tool for stakeholders in the music industry, enabling them to enhance user engagement and satisfaction by aligning their offerings more closely with user preferences and industry trends.

4. Further Discussion

A few limitations to our project arise as a result of the dataset we used. The creation of the Spotify app was in the year 2006 and data collection of specific stats during that time wasn't as accurate/extensive as it is today. In other words, we aren't able to gather the diverse set of reviews and sentiments on specific albums, due to the fact that some may have been created prior to the app's creation, as there is a lack of resources available for calculation during that period. As a result, our analysis of sentiment and rating calculation may not be as extensive and thus can hinder our overall results.

Another limitation which can arise is through the use of the Sentiment Intensity Analyzer from the VADER library. Sentiment analysis is defined as "the process of ‘computationally’ determining whether a piece of writing is positive, negative or neutral." Because we determine our compound scores for the genre and descriptors through the analyzer, there is a possibility that the analyzer itself doesn't have much information to base it's scores on as some pieces of text may be too short or the order in which the descriptors are placed may have a confounding effect on the overall compound score.

Additionally, our dataset is based on the data collected by the site RateYourMusic.com meaning that although we do have a multitude of reviews from a certain population, the results may not be representative for a much larger group of people. We are able to conduct our analysis but there is a chance that the albums themselves are rated differently according to other sources as well.

In the future, we would like to be able to build upon our study and conduct our analysis with better datasets that cover a larger range of albums and music review platforms. Although we were able to come to significant conclusions through our multiple hypothesis tests in the end, the data itself may not be very representative of the sentiments of a larger population outside scope. In conclusion, there are many aspects of this study that can be improved/worked on from the datasets utilized to the choices of what we used to measure the sentiments of various albums as well.

Overall, there is a lot that can be studied from the music industry and while we have shown significant findings in our predictive modeling and testing, we can take a step back and appreciate the diversity of albums in this day and age. By doing so, we can gain a better understanding of the dynamics music plays in the world around us and learn about the diversity of opinion within our communities.