STAT355 Project

Colleen Li

December 2, 2025

Abstract

This report analyzes the Spotify Music dataset (from Kaggle), which contains data for over 8,500 tracks spanning from 2009 to 2025, with the goal of answering three research questions using various statistical methods. The first section explores the key variables of the data set through exploratory data analysis (EDA) to provide summary statistics and visualizations. In the second section, we develop and address three research questions using correlation, t-tests, and ANOVA.

1 Preliminary Analysis

1.1 Data Overview

The data set used in this analysis is the *Spotify Global Music Dataset*, namely `spotify_data clean.csv`, which contains information on Spotify tracks. The data set spans from 2009 to 2025 with each row corresponding to a unique track. Each row contains track features, artist features, and album features.

- Track Features: popularity, explicit content, track number, etc.
- Artist Features: popularity, genre, followers, etc.
- Album Features: album id, track type (single, compilation, or album)

To prepare the data for analysis, the data was checked for missing values (NaNs) using `colSums(is.na())`. Then, entries with missing data were

removed accordingly with `na.omit()` to prepare the data set for analysis. Given the size of the data, around 8,500 entries, it was determined that taking a random sample was not necessary. Analyzing the full dataset would likely be the best approach to get comprehensive and powerful results.
For the analysis, the subset of key variables explored are primarily track popularity, explicitness, artist popularity, artist genre, and album type.

1.2 Exploratory Data Analysis (EDA)

Using data visualizations and summary statistics, track popularity, explicitness, artist popularity, and artist genres were explored in detail.

1.2.1 Track Popularity

Track popularity, denoted as a numeric variable that ranges from 0 to 99, represents how well a track performs on Spotify. After viewing a summary of the track popularity data, the following information is revealed:

Min = 0.00

1st Quartile = 39.00

Median = 58.00

Mean = 52.35

3rd Quartile = 71.00

Max = 99.00

The summary statistics suggests that the track popularity is skewed left because the mean is less than the median. This conclusion is synonymous with the visualization produced in Figure 1.
Figure 1 shows a histogram that was created to visualize the distribution of track popularity. The histogram reveals left skewness and a disproportionate frequency of zeros. Figure 2, a boxplot of the distribution, was also produced to analyze for any potential outliers. When viewing the boxplot, no apparent outliers appear to exist.
To further verify the skewness, `qqnorm()` and `qqline()` were utilized to determine the normality of the distribution. As seen in Figure 3, it is clear

that there is major skewness, and the distribution does not follow a Normal distribution. In hopes of reducing the skewness of the data without compromising meaningful analysis, the next step taken was to remove the large frequency of zeros in the data (which can be seen in Figure 1). There may have been a lack of data in the track's popularity which were defaulted to zero, resulting in a disproportionate amount of zeros. While removing the zeros did improve the distribution towards a Normal distribution slightly, the distribution is still slightly skewed as seen in Figure 4. After the removal of the large frequency of zeros, the data better fits a Normal distribution which can be seen the updated histogram in Figure 5. While there are still outliers and skewness in the distribution, most of the data- except for the ends- follows the qqline in Figure 4, and the data appears to be somewhat Normal in Figure 5. This indicates a moderately Normal distribution, allowing for future tests that involve the assumption of Normality in Part Two.

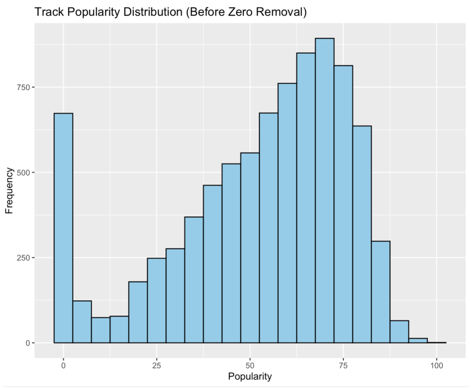


Figure 1: Track Popularity Distribution Before Zero Removal

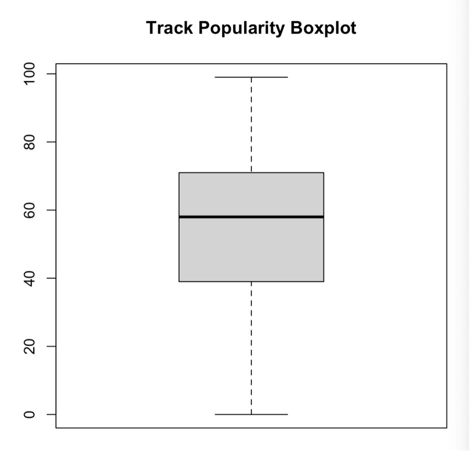


Figure 2: Track Popularity Boxplot (Before Zero Removal)

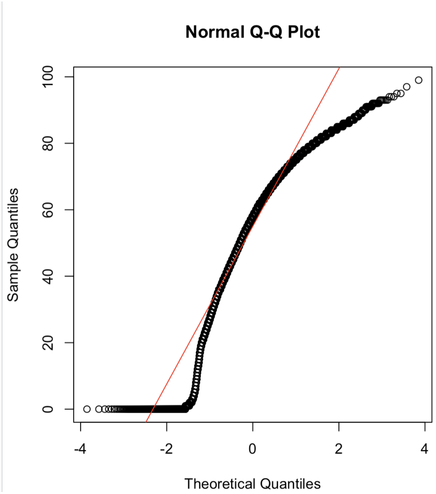


Figure 3: Track Popularity: Before removing large frequency of zeros.

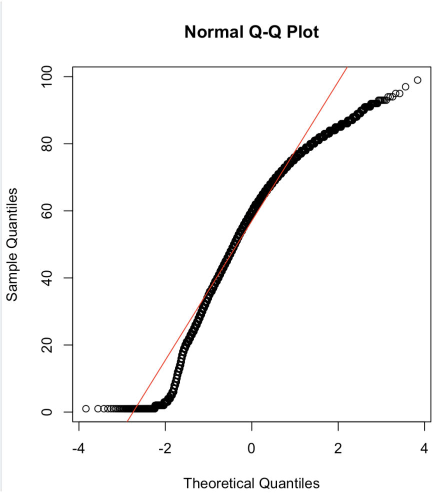


Figure 4: Track Popularity: After removing large frequency of zeros.

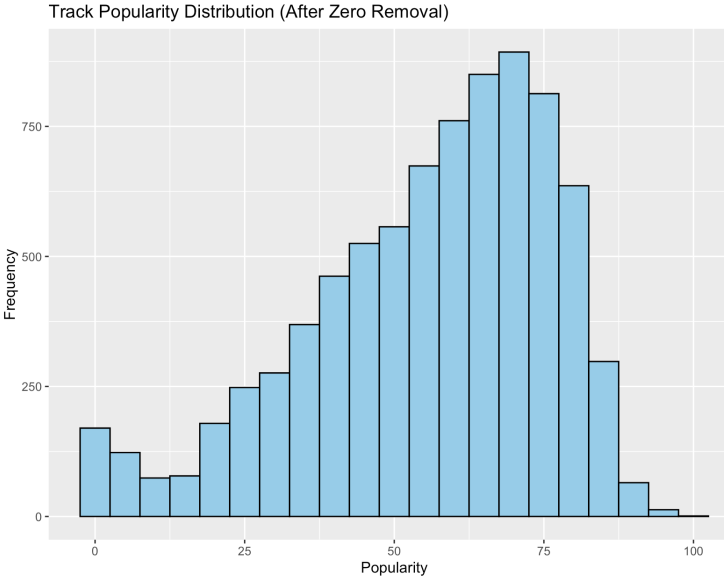


Figure 5: Track Popularity Distribution After Zero Removal

1.2.2 Explicit Tracks

Explicitness is a boolean variable that indicates whether the track is explicit or not. To better understand explicitness in the dataset, a bar plot (Figure 6) was generated to show the count of explicit versus non-explicit tracks. In Figure 6, there is a disproportionate amount of non-explicit tracks in comparison to explicit tracks in the dataset. There is approximately 6250 non-explicit tracks, 2000 explicit tracks, and a negligible amount of uncategorized data. This information provides important considerations to make when conducting testing.

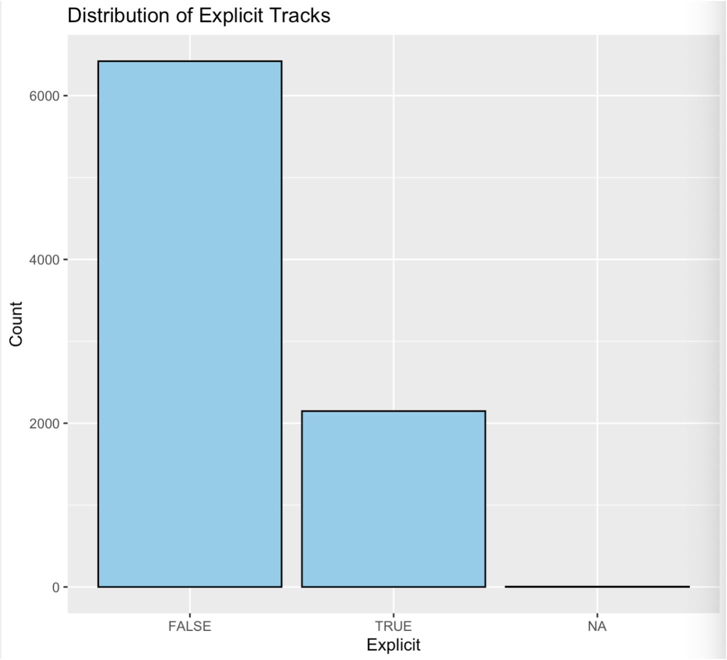


Figure 6: Distribution of Explicit Tracks

1.2.3 Artist Popularity

Artist popularity is a numeric variable representing an artist's popularity on Spotify. Higher values indicate more popularity while smaller values indicate less popularity. After viewing a summary of the artist popularity data, the following information is revealed:

Min = 0.00

1st Quartile = 61.00

Median = 75.00

Mean = 70.69

3rd Quartile = 84.00

Max = 100.00

The summary statistics suggests that the track popularity is skewed left because the mean is less than the median. This conclusion is synonymous with the visualization produced in Figure 7, a histogram that was plotted to better understand the distribution of artist popularity. A boxplot was constructed to inspect for any potential outliers. As seen in Figure 8, there are many outliers in the distribution resulting in the left skewness of the distribution.

Normality was then checked using `qqnorm()` and `qqline()` as seen in Figure 9. There are clear outliers and the distribution does not follow a Normal distribution.

Given the outliers from the boxplot in Figure 8, the 1.5 * IQR rule is used to remove the outliers that are heavily skewing the data. The outliers may be due to a lack of data or an innate concentration towards more popular artists in the dataset. However, if the outliers is not due to data error, a drawback of removing the outliers would be losing data from artists with low popularity. Final conclusions and analysis may not fairly take into account artists with lower popularity scores.

After removing the outliers, the new qqplot is shown in Figure 10. Compared to the qqplot before removing outliers in Figure 9, there is significant change in the distribution. The data now more closely follows a Normal distribution. While there is still outliers and skewness in the data, the distribution is moderately Normal. A similar conclusion may be made when inspecting Figure 11, the new distribution of artist popularity. While there is still a left skew, the data is somewhat Normal.

Figure 12 displays an important note to make. The formula for boxplots in R is different from the 1.5 * IQR rule. Therefore, when plotting the new boxplot, there still appears to be a potential outlier in the dataset. This represents a slight variation and potential drawback of differing methods. Using another method besides the 1.5 * IQR method may have created a better removal of outlier; however, in this case, the 1.5 * IQR method is satisfactory for outlier removal.

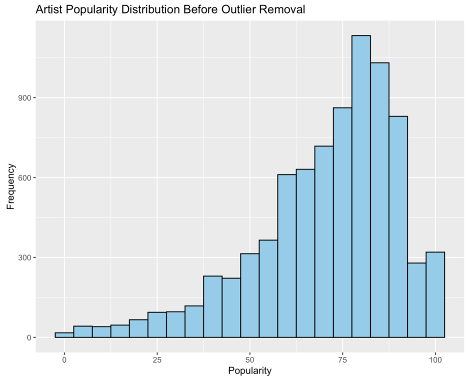


Figure 7: Artist Popularity Distribution Before Outlier Removal

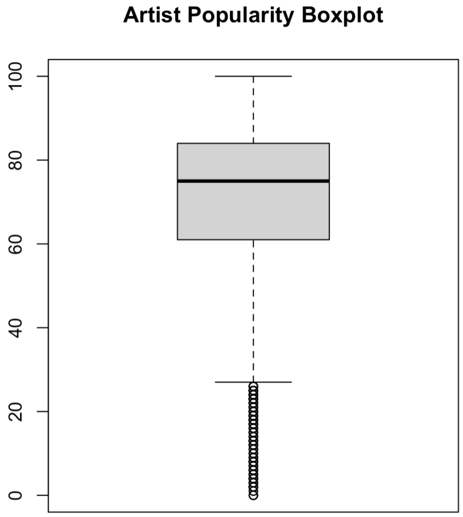


Figure 8: Artist Popularity Boxplot Before Outlier Removal

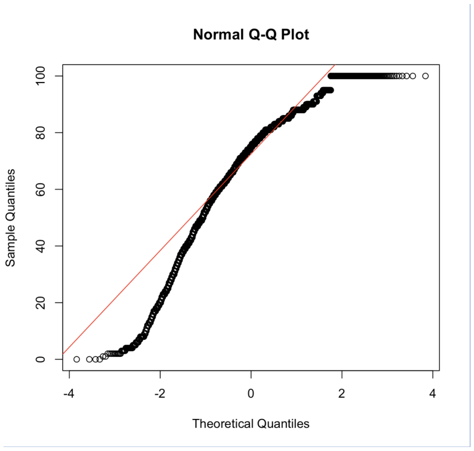


Figure 9: Artist Popularity: Before Removing Outliers.

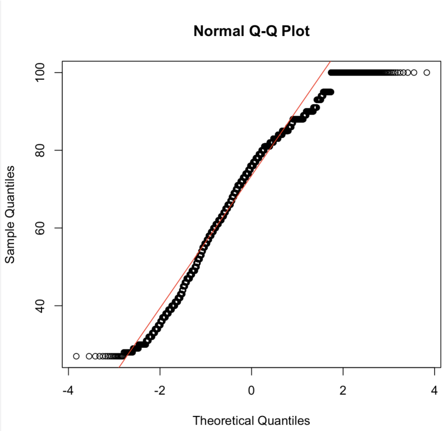


Figure 10: Artist Popularity: After Removing Outliers.

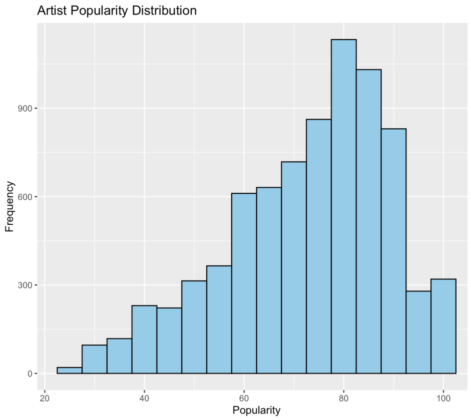


Figure 11: Artist Popularity Distribution After Outlier Removal

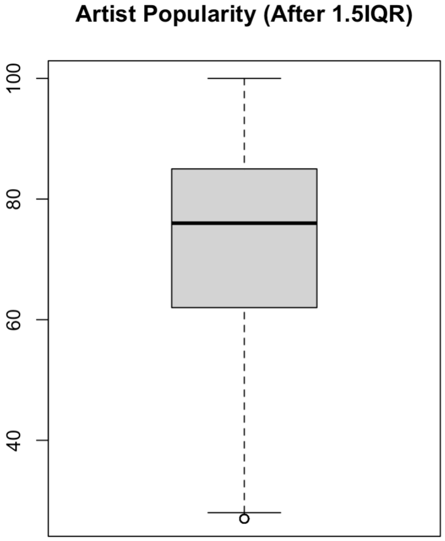


Figure 12: Artist Popularity Boxplot After Outlier Removal

1.2.4 Artist Genres

The genre(s) of each artist is a categorical variable, with some artists having multiple genres. For each artist, each genre is separated by commas. After extracting the genres from each artist, the genres were counted to better understand the dataset. To synthesize and visualize the data, the bar plot in Figure 15 was constructed to show the top 10 most common/popular genres in the dataset.

The first constructed bar plot shows a high concentration of N/A, denoting undefined artist genres, which convoluted the data. Removal of N/As from the data visualization as seen in Figure 14 provided a better understanding of the top ten artist genres. The top ten genres were country, pop, hip hop, pop, indie, folk, rock, soundtrack, rap, soft pop.

A problem in the list of the top ten genres is the repetition of ”pop”. There are two pop categories in the top ten artist genres, likely due to whitespace which results in processing as two different genres. This affects Figures 13 and 14. Aggregating the two ”pop”s into one ”pop” category would place pop as the most common artist genre. Another important consideration to make is the overlap between different genres. For example, a track may be both country and pop. Thus, running statistical tests on specific tracks in relation to genres may involve many complexities as individual tracks may correspond to multiple genres.

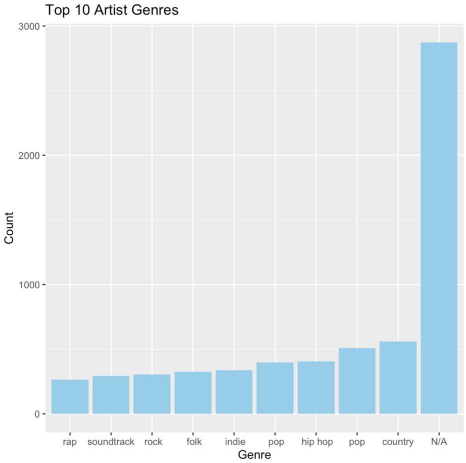


Figure 13: Top Ten Artist Genres: Before N/A Removal

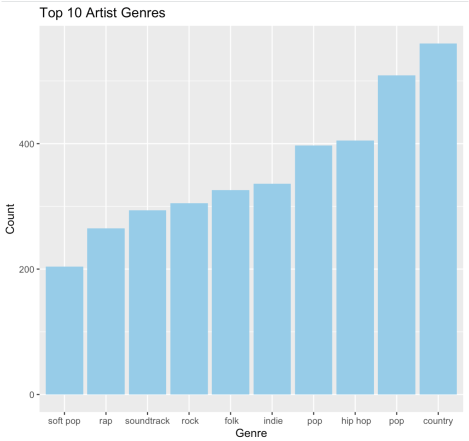


Figure 14: Top Ten Artist Genres: After N/A Removal

1.2.5 Album Type

The album type is a categorical variable denoting whether each track is a single, compilation or an album. A bar plot was created (Figure 15) to show the distribution of the different album types in the data set. There are three different album types: album, compilation, and single. Tracks in albums make up the majority of the data set at around 5,500 tracks. Singles, or tracks not in an album and released individually, make up almost 2,000 tracks in the data set. Lastly, compilations make up a minimal amount of approximately 400 tracks.

The difference in counts of different album types is an important consideration to make in future tests involving this variable. In other words, the availability and frequency of the different types may impact findings involving album type.

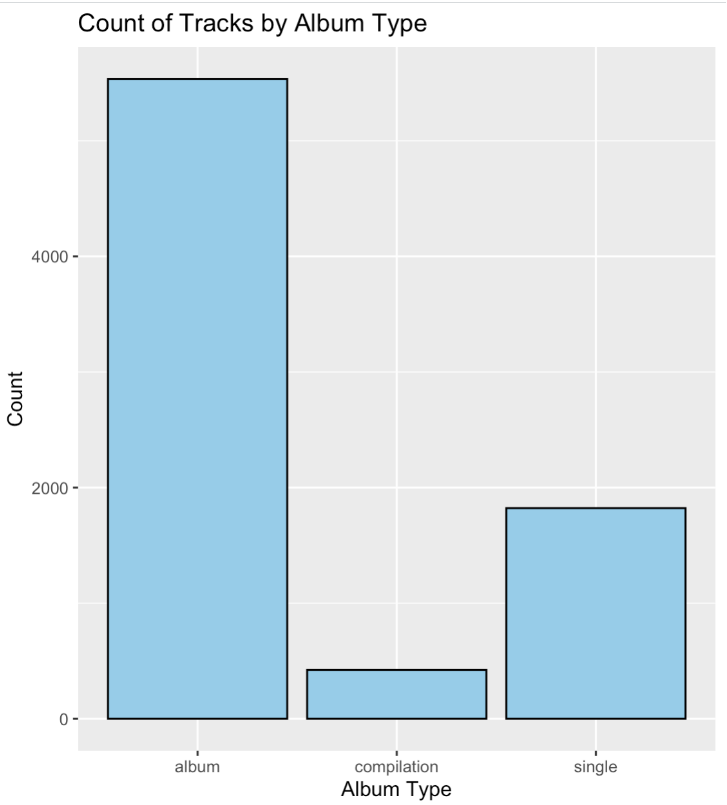


Figure 15: Distribution of Album Type

2 Research Questions

In this section, three research questions were developed based on EDA and the characteristics of the data set. Then, each question is addressed using appropriate statistical methods.

2.1 Do Tracks with Higher Artist Popularity Tend to Have Higher Track Popularity?

This question is significant in the music industry as it is synonymous with determining if artists' fan bases, brand recognition, and previous success may influence the performance of the artists' tracks. Understanding the relationship between artist popularity and track popularity may offer valuable insights for record labels, marketing teams, and artists. Depending on the correlation, resources might be allocated differently to attract the widest audience and gain the most traction.

- Null Hypothesis (H_0): There is no correlation between artist popularity and track popularity.

- Alternative Hypothesis (H_A): There is a correlation between artist popularity and track popularity.

  Method: A correlation test and a linear regression model were performed to examine the relationship between artist popularity and track popularity. Correlation testing and linear regression allows for an understanding of the potential linear relationship between artist popularity and track popularity. The correlation coefficient will also provide insights into the strength and direction of the association. The linear model will quantify the expected change in track popularity based on changes in artist popularity.

Assuming that the data is approximately Normal, which is moderately satisfied after transformations (namely removal of zeros and outlier removal) made in Preliminary Analysis, the correlation coefficient was determined to be 0.3823531 (Figure 16), indicating a weak positive correlation.

There is a weak positive relationship between artist popularity and track popularity. Therefore, tracks from more popular artists are sometimes more popular; however, this relationship is likely not linear. While popular artists may produce more popular tracks, other factors contribute to the variation in track popularity.

This becomes more apparent when plotting the two variables in Figure 17. The scatterplot is widely distributed, suggesting large ranges and outliers. This makes sense has some tracks are "one hit wonder" while others may be "flops". Additionally, musicians often release albums with many tracks; regardless of the artists, it is typical that some tracks perform well while others do not. There is also a significant amount of clustering at the higher end of the popularity values. Because many tracks and artists have high popularity scores, the dataset itself may have very popular artists and tracks that dominates the distribution. This concurs with the distributions from Preliminary Analysis. Track popularity and artist popularity, Figures 5 and 11, respectively, after transformation both show left skewness.

When analyzing the linear regression model between the artist popularity and track popularity, a summary of the analysis is shown in Figure 18.

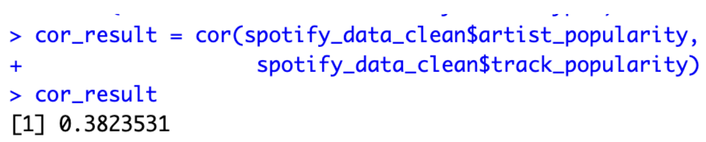
- The coefficient for artist popularity is 0.47476, meaning that for every 1 unit increase in artist popularity, the track popularity is expected to increase by approximately 0.47476 units when holding all other factors constant.

- The low p-value of $< 2.2e - 16$ indicates that the relationship is statistically significant. There is statistically strong evidence that artist popularity has a significant relationship with track popularity. However, from the scatterplot in Figure 17 and the correlation r in Figure 16, the relationship between the two variables is likely not linear. While there is likely a relationship between artist popularity and track popularity, a non-linear model would likely better explain the relationship.

- R-squared of 0.1461 denotes that 14.61% of the variation in track popularity can be explained by artist popularity. This suggests that artist popularity only explains a portion of the variance in track popularity; many other factors also influence track popularity.

From the correlation test and the linear regression model, artist popularity may correspond to more popular tracks. However, this relationship is weak and likely not linear; the relationship results in a wide distribution with many outliers. While a linear model may not ideal, there is still a statistically significant relationship between the two variables.

This finding indicates that artist popularity, from fan bases, brand, and/or previous success, has a relationship with track popularity. Exploring more factors would provide strategic insights for record labels, marketing teams, and artist. The finding that there is a statistically significant non-linear relationship reveals that further research into this topic may be meaningful and impactful in producing successful tracks.



```
> cor_result = cor(spotify_data_clean$artist_popularity,
+                  spotify_data_clean$track_popularity)
> cor_result
[1] 0.3823531
```

Figure 16: Correlation Test between Artist Popularity and Track Popularity

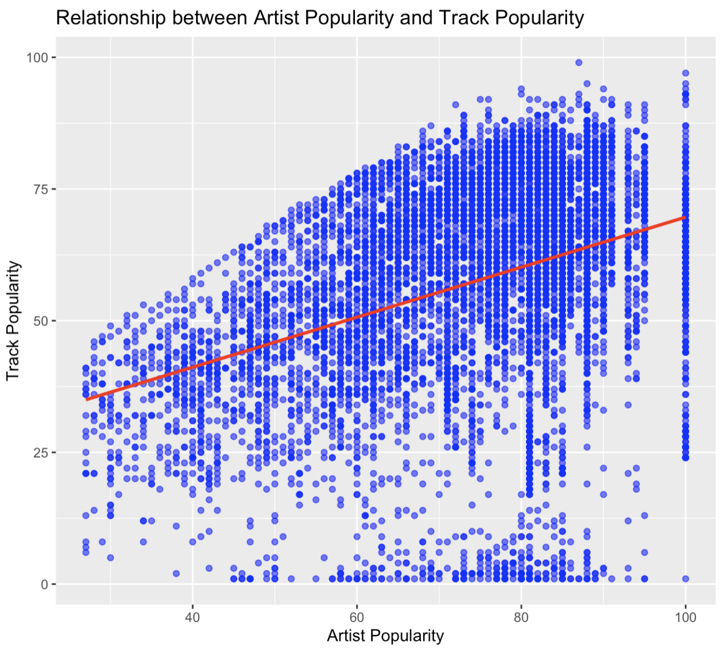
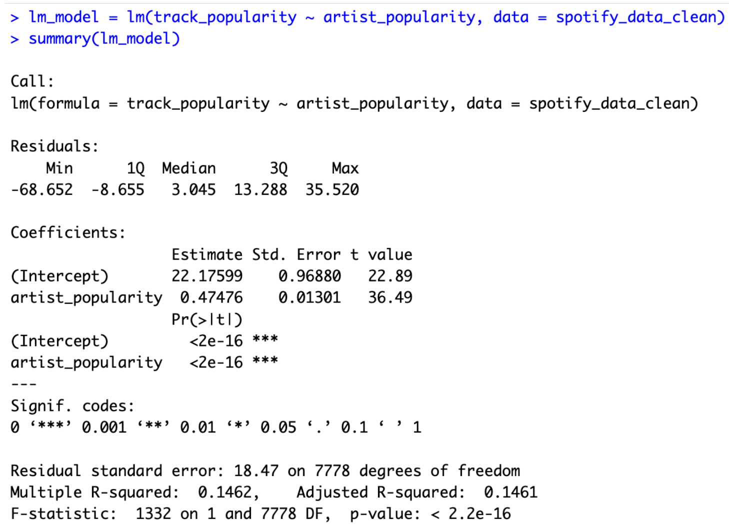


Figure 17: Relationship between Artist and Track Popularity Scatterplot



```
> lm_model = lm(track_popularity ~ artist_popularity, data = spotify_data_clean)
> summary(lm_model)

Call:
lm(formula = track_popularity ~ artist_popularity, data = spotify_data_clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-68.652  -8.655   3.045  13.288  35.520 

Coefficients:
                  Estimate Std. Error t value
(Intercept)       22.17599    0.96880   22.89
artist_popularity  0.47476    0.01301   36.49
                  Pr(>|t|)    
(Intercept)         <2e-16 ***
artist_popularity   <2e-16 ***
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.47 on 7778 degrees of freedom
Multiple R-squared:  0.1462,	Adjusted R-squared:  0.1461 
F-statistic:  1332 on 1 and 7778 DF,  p-value: < 2.2e-16
```

Figure 18: Linear Regression Model between Artist and Track Popularity

2.2 Is There a Significant Difference in Track Popularity Between Explicit and Non-Explicit Tracks?

This question compares the performance of explicit and non-explicit tracks. This is of interest as it could help artists better understand which type of musical content might have higher success. Analyzing this question may provide answers for artists when they determine the content of their tracks. Do explicit tracks appeal more to specific demographics or do they face censorship and reduce the track's visibility? Do non-explicit tracks attract more mainstream radio play resulting in more popularity or do the lyrics come off as lacking?

- Null Hypothesis (H_0): There is no significant difference in track popularity between explicit and non-explicit tracks.

- Alternative Hypothesis (H_A): There is a significant difference in track popularity between explicit and non-explicit tracks.

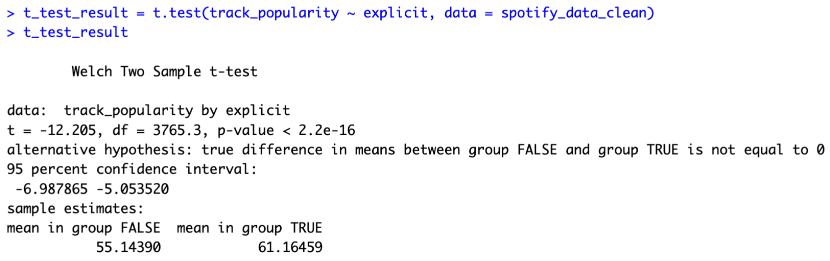
  Method: A two-sample t-test to compare the mean track popularity between explicit and non-explicit tracks. The test allows for a comparison of means between two independent groups. In each track, the track may be explicit or non-explicit; it cannot be both. Independence is satisfied. The assumptions of normality are reasonably met after data transformation from Preliminary Analysis. The p-value from this test will denote whether or not there is a statistically significant difference in track popularity between the two groups. The test focuses on the means of both groups, so the two sample independent t-test provides an answer to whether one group results, on average, in popular tracks than the other.

Assuming that the two groups, explicit and non-explicit, are independent, and track popularity is approximately Normal (which is moderately satisfied after transformations in Part One: Preliminary Analysis), a two-sample t-test is conducted. The test results are displayed in Figure 19.
The test resulted in a small p-value of $< 2.2e - 16$. Therefore, we reject the null hypothesis. There is a statistically significant difference in track popularity between explicit and non-explicit tracks.
The mean popularity of non-explicit tracks is 55.14390, while the mean popularity of explicit tracks is 61.16459. When viewing the 95% confidence

interval for the difference in means, the difference is between -6.987865 and -5.053520. Therefore, on average, explicit tracks tend to be more popular than non-explicit tracks by between 5.053520 to 6.987865 points.

Some important considerations to make here is that there is a significant imbalance in the number of explicit and non-explicit tracks in the dataset, as seen in Figure 6. A large difference in sample sizes between groups may lead to situations where one group (explicit or non-explicit) disproportionately influences the results, even if the true difference between the group is minimal. To improve the analysis, bootstrapping or resampling may mitigate this concern.

The results imply that it may be easier to stand out in explicit tracks as there tend to be less of them. Additionally, it appears that censorship is not a major barrier for explicit tracks as they can perform just as well- if not better- than non-explicit tracks. This test ultimately provides a baseline for decision making regarding lyrics and content of upcoming tracks. However, it is difficult to provide a concrete analysis given the limitations from the structure of the data set. Replicating this test on a data set with equal amounts of explicit and non-explicit tracks would likely provide a more fair analysis.



```
> t_test_result = t.test(track_popularity ~ explicit, data = spotify_data_clean)
> t_test_result

        Welch Two Sample t-test

data:  track_popularity by explicit
t = -12.205, df = 3765.3, p-value < 2.2e-16
alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
95 percent confidence interval:
 -6.987865 -5.053520
sample estimates:
mean in group FALSE  mean in group TRUE 
           55.14390            61.16459 
```

Figure 19: t-test for Track Popularity by Explicitness

2.3 Do Tracks from Albums Tend to Be More Popular Than Singles?

This question explores the importance of the format of a track and whether it has a significant impact on its popularity. Understanding the association between album type and track popularity, strategic decisions can be made about releasing music and discographies. Additionally, results of this question may provide additional insight into how artists should plan their releases.

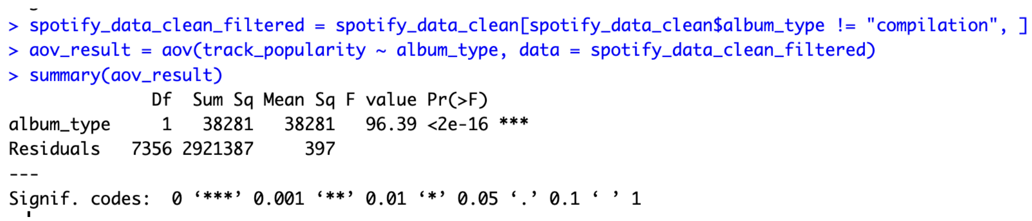
- Null Hypothesis (H_0): There is no difference in track popularity between album tracks and singles.

- Alternative Hypothesis (H_A): Tracks from albums are more popular than singles.

  Method: ANOVA will be used to compare the mean track popularity between album tracks and singles. An ANOVA test is appropriate for this question because it allows for comparison of mean track popularity across at least two groups. The assumptions for the ANOVA test are reasonably satisfied. The data is independent as tracks are either in an album or released as a single; tracks cannot be both an album and a single. The assumption of normality is also reasonably satisfied after data transformations during Preliminary Analysis. The test will ultimately determine whether there is a statistically significant difference between the means of the two types of tracks: tracks in an album and tracks released as singles. Given that there is a significant p-value from the ANOVA, a post-hoc analysis will be performed with Tukey's HSD. Tukey's HSD provides additional information by identifying which group differs in terms of mean popularity.
  Note: Throughout the testing, compilation albums are excluded to ensure that only relevant categories (albums and singles) are considered in the analysis.

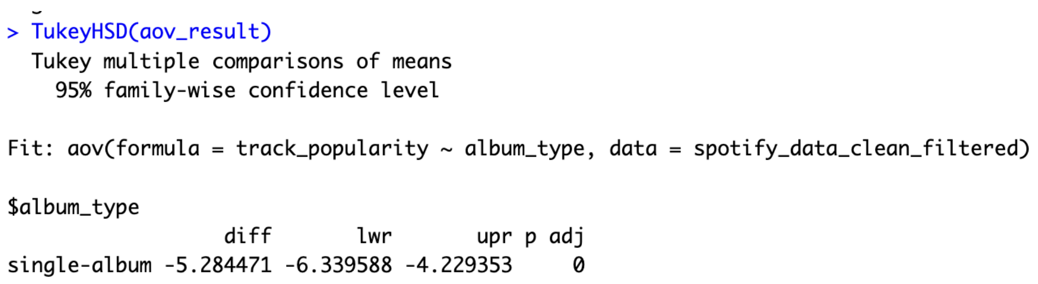
After excluding "compilation" types from the analysis and assuming (1) the data is independent (either album track or single) and (2) track popularity is approximately normally distributed in each group, an ANOVA test is conducted. A summary of the ANOVA test is shown in Figure 20. The test produced an extremely small p-value of $< 2e - 16$, so the null hypothesis is rejected. There is statistically significant evidence that means of tracks from albums are significantly different from means of singles.
Post-hoc analysis is performed using `TukeyHSD()` because the p-value was significant. The Tukey HSD test shows that tracks from albums tend to have higher popularity than tracks from singles, with a difference of -5.284471.
To better visualize this difference, a box plot was produces as seen in Figure 22. The analysis question focused only on album and single types, so

compilation is excluded in the analysis. When comparing the album and singles box plots, the first quartile, median, and third quartile of album are all greater than that of single. This agrees with the ANOVA and TukeyHSD analysis.
There appears to be many outliers in the album box plot. This signifies that albums may include tracks that are not very popular. While the overall finding suggests that album tends to provide additional exposure and value to tracks, often making them more popular than standalone singles, the outliers show that potential conclusion is not definitive. Additionally, the distribution of the different album types in the data set (Figure 15) shows that there are significantly more tracks in albums compared to singles. A drawback to the finding is the difference in counts of albums vs singles as it may have altered the final conclusion.
This test implies that albums may provide more visibility and increase track popularity. In the data set, there are more tracks in albums than tracks as standalone singles. Given these preliminary results, artists might make strategic decisions about how to stand out or appealing to fan bases when releasing music and discographies. Considering additional factors may result in a more complete understanding of the relationship between album type and singles. Additionally, further studies with similar amounts of data in albums and singles would provide a more fair analysis.



```
> spotify_data_clean_filtered = spotify_data_clean[spotify_data_clean$album_type != "compilation", ]
> aov_result = aov(track_popularity ~ album_type, data = spotify_data_clean_filtered)
> summary(aov_result)
              Df  Sum Sq Mean Sq F value Pr(>F)    
album_type     1   38281   38281   96.39 <2e-16 ***
Residuals   7356 2921387     397                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 20: ANOVA for Track Popularity by Album Type



```
> TukeyHSD(aov_result)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = track_popularity ~ album_type, data = spotify_data_clean_filtered)

$album_type
                  diff       lwr       upr p adj
single-album -5.284471 -6.339588 -4.229353     0
```

Figure 21: Post-Hoc TukeyHSD Analysis

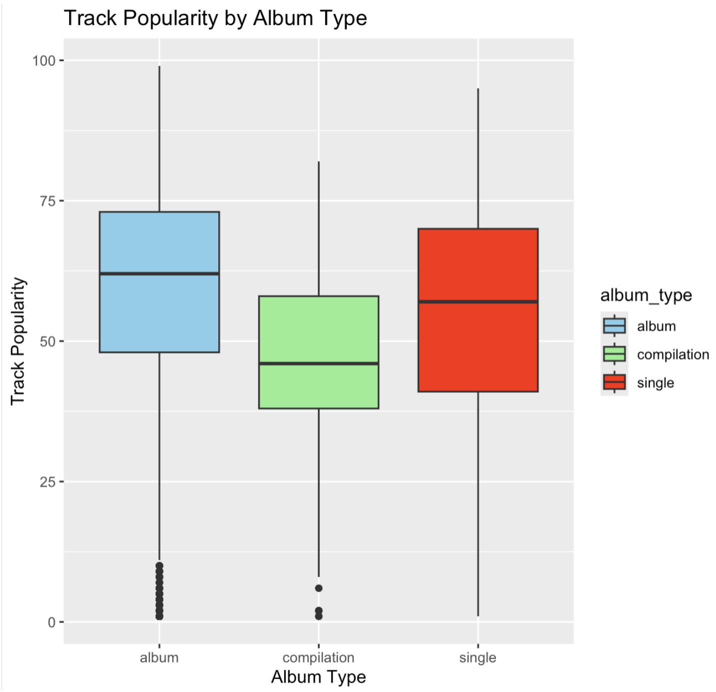


Figure 22: Track Popularity Boxplot by Album Type

3 Conclusion

After analyzing the Spotify Global Music Dataset to explore various factors, three major questions were answered: (1) do tracks with higher artist popularity have higher track popularity?, (2) is there a significant difference in

track popularity between explicit and non-explicit tracks?, and (3) do tracks from albums tend to be more popular than singles?
The analysis revealed a weak positive correlation between artist popularity and track popularity. While there is likely a relationship between artist and track popularity, a linear model results in a weak portrayal of the relationship as other factors likely contribute to track popularity. Additionally, it was discovered that explicit tracks are significantly more popular than non-explicit tracks, with explicit tracks averaging a higher popularity score. However, it is important to note that the imbalance in the number of explicit and non-explicit tracks in the dataset may have influenced these results. Lastly, tracks from albums are statistically significantly more popular than singles, suggesting that being a part of an album provides additional exposure, allowing album tracks to be more likely to succeed. Similar to the previous question, the imbalance of album types may have influences the final results. While additional, unaccounted factors may also play a role in the findings, the findings highlight the importance of artist popularity, track explicitness, and album context in determining track popularity. Further analysis, such using additional variables and methods, would provide an improved understanding of the variables and their relationships.

4 References

Wardabilal. (2025). *Spotify global music dataset (2009–2025)* [Data set]. Kaggle. https://www.kaggle.com/datasets/wardabilal/spotify-global-music-dataset-20092025/data