**Vivian Sabga**

**Alison Sgambati**

**MA 346 Final Report**

**Report:**

The goal of our data project was to investigate whether or not there is any correlation between the number of followers an artist has and if their top song is explicit or not. Essentially, we wanted to discover if an artist has more followers when they create explicit music. Our goal was to determine if a song being explicit or not had any effect on the number of followers an artist had. After thorough research and analysis, we found that there is no direct correlation between an artist producing explicit music and the number of followers they have. We explored our datasets, removed unnecessary columns, filled in missing values and merged the two datasets based on the Artist ID and ID columns. We then grouped the data frame by Artist ID and selected the first song out of each artists’ top ten songs. We created a new data frame with this manipulated data and conducted another merge between this dataset and the original dataset of the artists information, so we were able to evaluate the data based on the top song of each artist and successfully answer the question, “Is there any correlation between the number of followers an artist has and whether their top song is explicit or not?”

To determine this, we first conducted a chi-squared test to calculate the correlation between the binomial variable (Is Explicit) and an integer variable (Followers). From this statistical method we were able to conclude that there is no significant correlation between the number of followers and the explicitness of songs because the p-value was relatively large, indicating that there is not enough evidence to reject the null hypothesis. Hence, we had to accept the null hypothesis that, ‘there is no significant correlation between followers and explicit songs.’

We then conducted a logistic regression to determine if there was a correlation between the binomial variable (Is Explicit) and an integer variable (Followers), we did this just to ensure the results from our chi-squared test were accurate. From this test we were able to confirm that the number of followers may not be a statistically significant predictor of whether a song is explicit or not, as the p-value is relatively high.

We then conducted one final statistical method to be sure of our results. We used a Point Biserial Correlation because we had a Boolean column and an integer column. The point-biserial correlation coefficient is a useful tool for assessing the strength and direction of the linear relationship between a binary variable and a continuous variable. The columns we are focusing on are the "Is Explicit" column and the "Followers" column. The correlation coefficient ranges from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no correlation. The p-value helps assess the statistical significance of the correlation. Our results yielded a p-value of 0.24625795105762002 suggesting that there is not enough evidence to reject the null hypothesis that there is no correlation between the two variables.

After confirming our conclusion, we decided to create visualizations to better understand and represent our results. We wanted to show a graphic that our hypothesis testing was true and that there is no correlation between the mean of followers with explicit songs and non-explicit songs.

A screenshot of a computer

Description automatically generated

From this graph we were able to see that the mean values are very close in number, and it is easy to tell that the difference is extremely small.

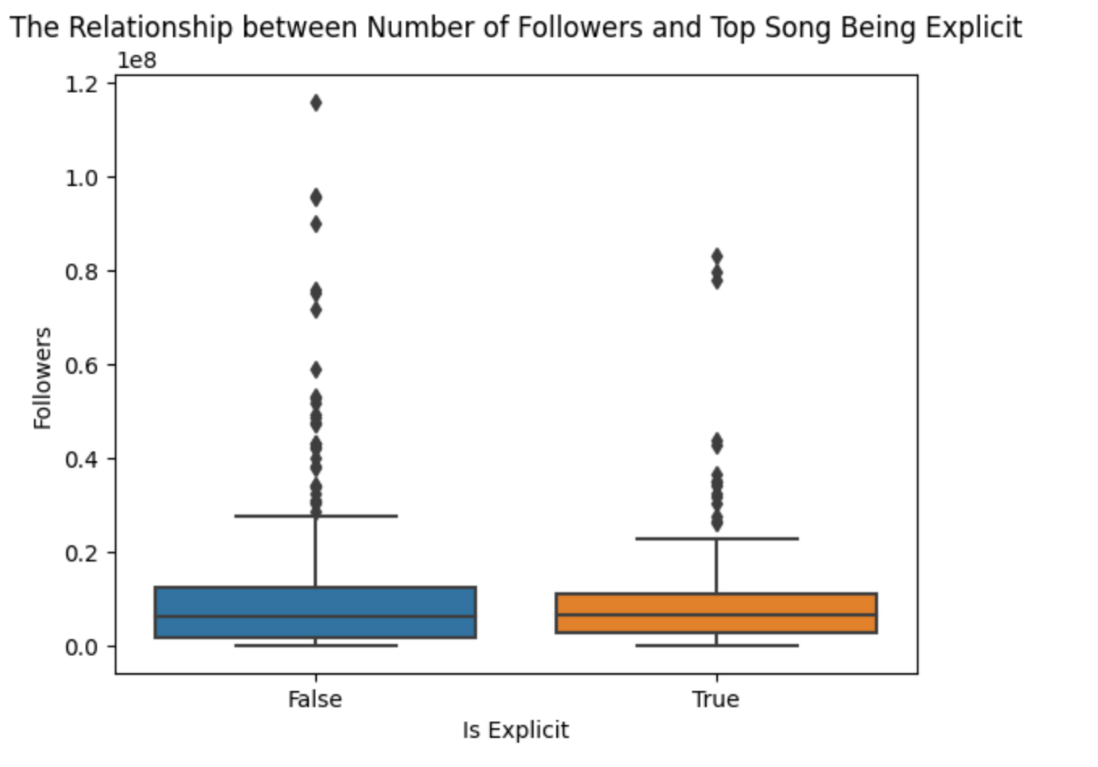
Then we wanted to break the followers into categories and decipher if the follower count by category played a role in if the songs were deemed explicit or not, so we created another graph showing this.

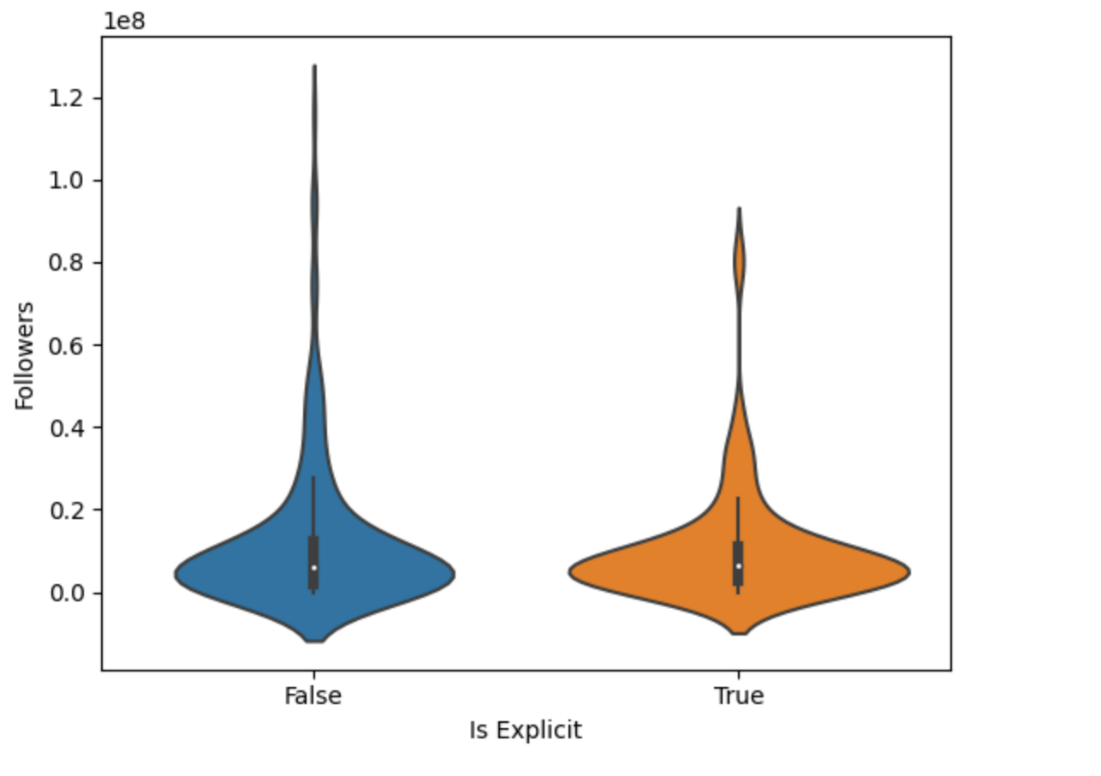
A graph of followers

Description automatically generated with medium confidence

The graph shows that the largest number of followers for both True and False is over 500,000. Most of the artists on the list have a large number of followers and there is no correlation seen between the number of categories a follower falls in and their top song being explicit or not.

Finally, we created two graphs, a Box-Whisker Plot and a Violin Plot to show the distribution between the mean numbers of artists with a top explicit song and artists without a top explicit song.





From these two graphs we were able to see that the distribution is generally equal between the two categories of artists with explicit songs and artists without. However, the range of average number of followers for those artists whose top song is not explicit is wider as they have a higher number of outliers.

The datasets we used were published on Kaggle. The first dataset includes information about the 10,000 most listened to artists in the United States. The second dataset includes information about each of these artists' top ten songs. The datasets cover a diverse range of musical genres and spans from 1964 to 2023. These datasets were suitable for our analysis because they each contained data that is relevant to conducting our hypotheses. Additionally, the data used can be used to compare several different statistics and relate to each other.