A Comprehensive Analysis of the Relationship Between Spotify and Billboard Rankings Using Linear Regression, Correlation Analysis, KNN Clustering, and Time-Series Modelling

Kaily Mejia

MGMT 6609: Data Analytics

Professor Thilanka Munasinghe

December 8th, 2023

Abstract

This paper attempts to build different data models to accurately predict a song's ranking on the Billboard Global 200 chart based on it's ranking on the Spotify Global 200 chart. For the Billboard charts second to buying, streaming accounts for the most points a song can get through Billboard's point system. With the world's largest service being Spotify and the Billboard chart's point system it's reasonable to hypothesize that the relationship between Spotify chart and Billboard chart would be clear and measurable. The data used to measure this connection is from twenty-seven weeks of chart information starting from April 13th, 2023, to October 14th, 2023. The data was synthesized from both Spotify and Billboard data that is publicly available. However, the songs considered were only those at appeared on both charts during this period. Using clustering, linear regression, correlation, and time-series analysis a moderately positive correlation between the charts has emerged.

Introduction

Over the past twenty-five years, the music industry has seen incredible changes to their business model. The change in the business model employed by the music industry can be reflected in the billboard music charts (Wikipedia Contributors, 2023). Since its inception in 1940, Billboard has tracked the most popular songs in the United States (Wikipedia Contributors, 2023). However, the charts have expanded to reflect the ever-changing landscape of the music industry. The charts have changed to reflect not only radio play but physical album sales of records, cassettes, and CDs not only in the United States but globally as well. The dawn of digital sales and streaming has completely changed the way people listen to music. It has made music more accessible globally and is now the biggest income generator for the music industry (Wikipedia Contributors, 2023). Reflecting how much streaming accounts for music popularity Billboard changed their rules yet again to account for this change. Billboard calculates a song's rank through a point system with different point

amount allocated toward each interaction with a song. As an example of rules changing to reflect the impact of streaming, streaming a song now counts for more points than radio play (Wikipedia Contributors, 2023).

In the streaming business, Spotify has the largest market share with the largest number of streams when compared to any other streaming service (Wikipedia Contributors, 2019). Given that streams contribute significantly to a song's chart performance and considering Spotify's status as the world's largest streaming service, one would anticipate a discernible and quantifiable relationship between these two charts. When looking at the available data from both sources a hypothesis began to form by looking at other features in the data could influence a song's ranking on the Billboard chart. There were features such as the number of weeks on the chart, number of steams, peak position on the chart, and the previous week's position on the chart. All of these were potential features of influence.

Data Description and Exploratory Data Analytics

To conduct this analysis, compiling the datasets from the two sources, Billboard and Spotify, was essential. Both charts are published weekly and are readily available online. Acquiring the data from Spotify was straightforward, the charts could be easily accessed and downloaded as data files for each week. However, obtaining data from the Billboard charts presented a challenge as it was not available for download. A web-scraping approach was employed, each week of Billboard chart data was taken from the website and formatted as a datafile. The web-scraping included features equivalent to those on the Spotify dataset for direct comparison. This process yielded 27 separate data files for the Billboard weeks that complemented the existing 27 data files from Spotify. To create a unified dataset, each corresponding week's data from both sources was joined based on song title, resulting in a comprehensive data file for each of the 27 weeks. The final dataset produced was each weekly paired Spotify and Billboard file combined.

Billboard Chart Feature Name	Billboard Chart Feature Data Type
chart_week	Character
current_week	Integer
title	Character
performer	Character
last_week	Character
peak_pos	Integer
wks_on_charts	Integer

Figure 1: Data Dictionary for Weekly Billboard Chart Data

Spotify Chart Feature Name	Spotify Chart Feature Data Type
chart_week	Character
rank	Integer
uri	Character
artist_name	Character
track_name	Character
source	Character
peak_rank	Integer
previous_rank	Integer
weeks_on_charts	Integer
streams	Integer

Figure 2: Data Dictionary for Weekly Spotify Chart Data

Billboard-Spotify Feature Name	Billboard-Spotify Feature Data Type
chart_week.x	Date (changed before analysis)
current_week	Integer
title	Character
performer	Character
last_week	Integer (changed before analysis)
peak_pos	Integer
wks_on_charts	Integer
chart_week.y	Date (changed before analysis)
rank	Integer
uri	Character
artist_name	Character
track_name	Character
source	Character
peak_rank	Integer
previous_rank	Integer
weeks_on_charts	Integer

streams	Integer

Figure 3: Data Dictionary for Final Dataset

Now with a thorough and comprehensive data set we can start to prepare the data for analysis. The only missing value within the whole dataset came from one Billboard chart feature, “last_week”, resulting from a song’s new appearance on the chart. In the Spotify feature equivalent, “previous_rank”, the value -1 is used instead of a null value. For cohesion the same value, -1, was applied to null values in the “last_week” feature. Additionally, the features, “chart_week.x” and “chart_week.y”, were changed from characters to dates.

During the creation of the final dataset there are a couple of sources of error that may have been introduced. Firstly, in merging the Billboard and Spotify data on son song title the only songs on the final dataset were the intersection of the two charts. If a song only appeared on either chart for a given week but not it was eliminated from the dataset during the merging process. This could introduce error in the models because it doesn’t account for songs that may appear on the Billboard chart regardless of its streaming numbers on Spotify. Spotify is the largest but not the only streaming service used around the world. It’s completely plausible for a song to be streamed enough times to land itself on the Billboard charts but those streams may have occurred outside of Spotify’s platform. Secondly, the songs were merged on song title assuming that the song titles are the same on both Spotify and Billboard’s. There could be some songs that were purged from the dataset because of slight differences in how a song’s title is displayed. This could also introduce error because it is not truly considering all the songs on both charts. These are specific uncertainties to keep in mind during the conclusion of the data analysis.

Analysis

In this comprehensive data analysis, various models were employed to unravel patterns and relationships within the dataset. A linear regression model was used to explore

the linear association between the current week's ranking on the Billboard chart and the Spotify rank. The correlation model, employing a multiple linear regression approach based on correlation analysis, to predict the current week's ranking. KNN clustering was applied to group data points with similar characteristics, revealing inherent patterns within the dataset. Finally, a time-series model, through cross-correlation analysis, unveiled the temporal relationship between Spotify and Billboard rankings over 27 weeks.

Linear Regression Model

The linear regression model serves as the foundational exploration into the linear association between the current week's ranking on the Billboard chart and the corresponding Spotify rank. The leading hypothesis was that the strongest predictor of a song's Billboard ranking would be it's ranking on Spotify's chart. Thus, the predictor variable is the Spotify rank, while the response variable is the Billboard rank. The model is constructed using the "**lm**" function in R, where the linear relationship is explored through a regression line fitted to the data. The function yielded the resulting intercept and coefficient seen in Figure 4. The "**plot**" function was used on the linear regression which plotted Figures 5, 6, 7, and 8. A train-test split was applied so that Mean Absolute Error (MAE) and Root Squared Mean Squared Error (RSME) could be employed as metrics to assess the accuracy and precision of the linear regression model. The results of both Mean Absolute Error and Root Squared Mean Squared Error are pictures in Figure 9. The model's performance was limited, with a mean absolute error (MAE) of 54.53782 and a root mean squared error (RMSE) of 66.57448.

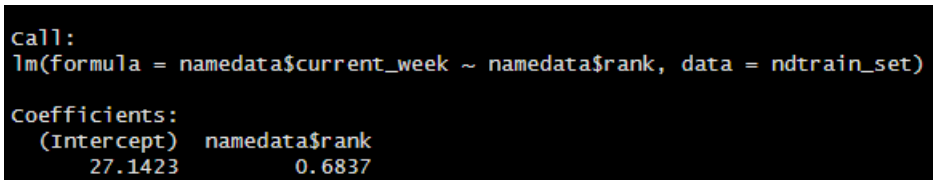

Figure 4: Linear Regression Model Formulation

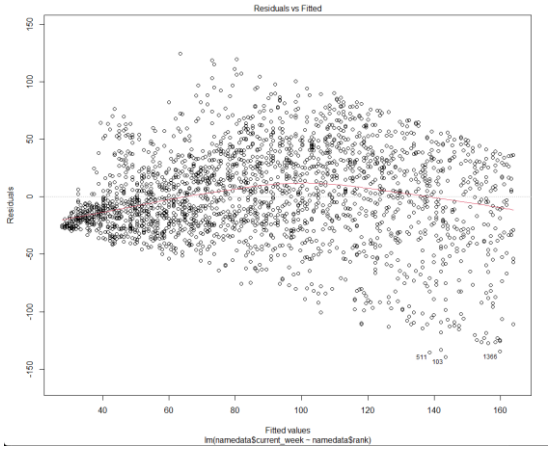

Figure 5: Linear Regression Residuals vs. Fitted Plot

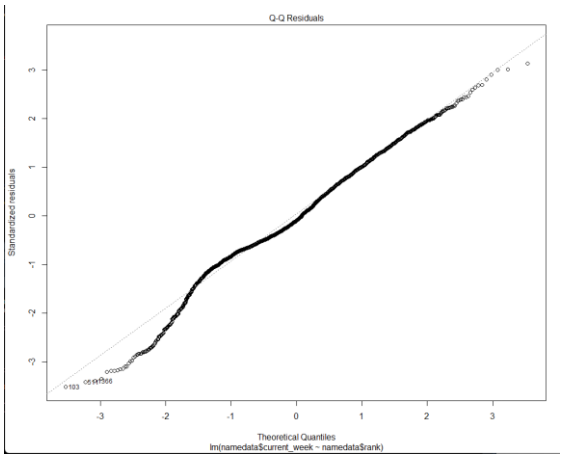

Figure 6: Linear Regression Q-Q Residuals Plot

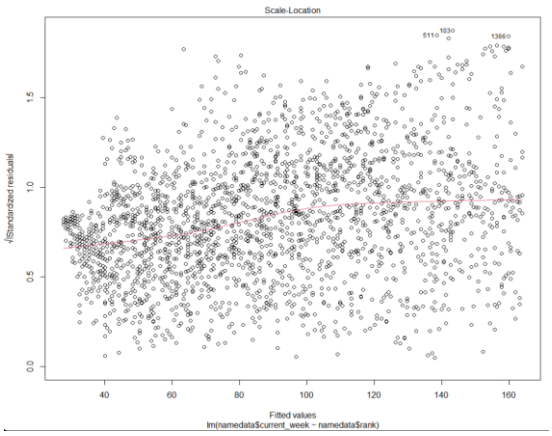

Figure 7: Linear Regression Scale-Location Plot

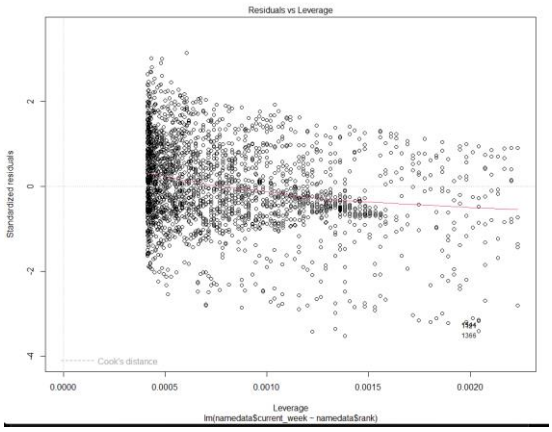

Figure 8: Linear Regression Residuals vs. Leverage Plot

```
> ndlinregMAE
[1] 54.53782
> ndlinregRSMSE
[1] 66.57448
```

Figure 9: Linear Regression Error Calculations

Correlation Model (Multiple Linear Regression)

As an optimization of the linear regression model, the correlation model delves into multiple linear regression based on features with a strong correlation to the current week's ranking. The model utilizes the "**cor**" function to compute the correlation matrix and subsequently selects features with a correlation coefficient above the defined threshold. The correlation matrix was used to make the heat map in Figure 10. Features were then selected based on a correlation coefficient above 0.5, encompassing last_week, peak_pos, wks_on_chart, rank, peak_rank, previous_rank, weeks_on_chart, and streams. The features that were above the threshold were "current_week", "last_week", "peak_pos", "rank", and "previous_rank". Similarly, to the linear regression, the multiple linear regression employed the "**lm**" function. The use of the function gave the intercept and coefficients seen in Figure 11. The resulting regression was also plotted producing Figures 12, 13, 14, and 15. Again, train-split was applied for Mean Squared Error (MSE) and Root Squared Mean Squared Error (RSME) to serve as benchmarks for evaluating the precision and effectiveness of this model. The results of the are shown in Figure 16. The model's performance improved significantly

compared to the previous linear regression model, with an MAE of 17.53078 and an RMSE of 25.05107.

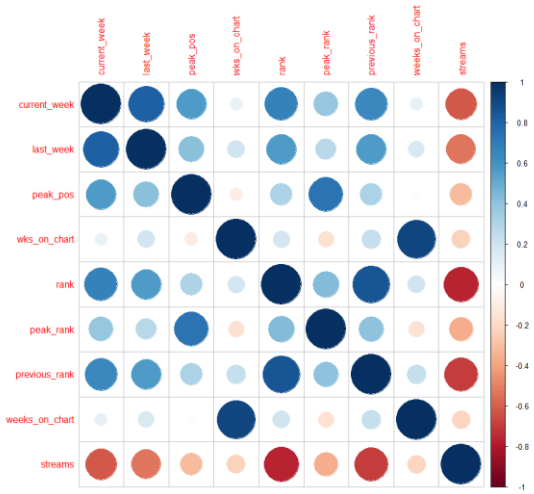

Figure 10: Correlation Matrix Heat Map

```
Call:
lm(formula = ndtrain_set$current_week ~ ., data = ndtrain_set[,
    features])

Coefficients:
  (Intercept)      last_week       peak_pos           rank  previous_rank
      5.41694        0.53962        0.32715        0.26809        0.03641
```

Figure 11: Multiple Linear Regression Model Formulation

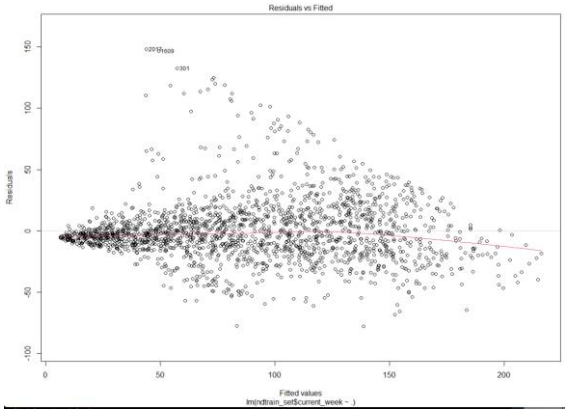

Figure 12: Multiple Linear Regression Residuals vs. Fitted Plot

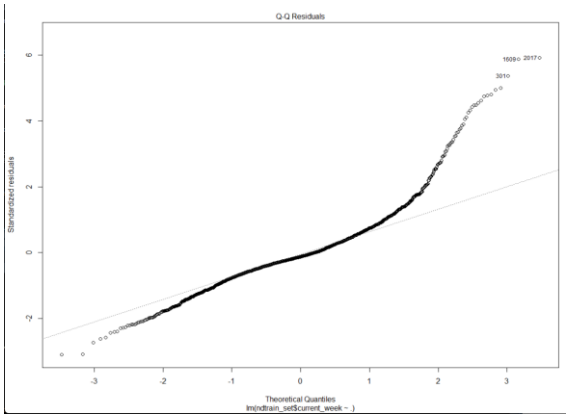

Figure 13: Multiple Linear Regression Q-Q Residuals Plot

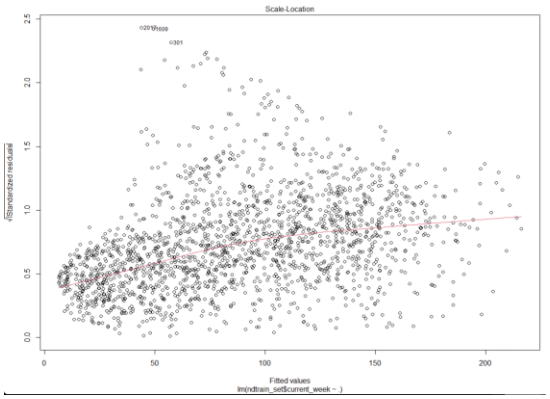

Figure 14: Linear Regression Scale-Location Plot

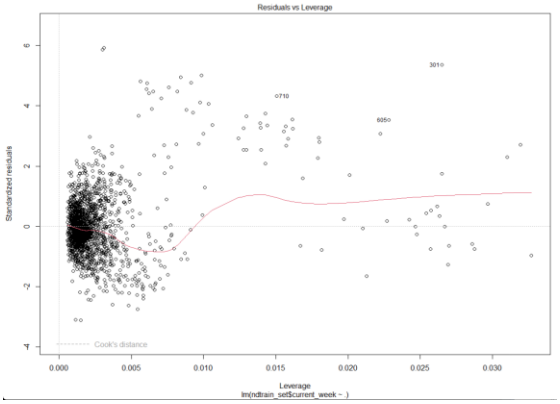

Figure 15: Linear Regression Residuals vs. Leverage Plot

```
> ndcorMSE
[1] 17.53078
> ndcorRSME
[1] 25.05107
```

Figure 16: Multiple Linear Regression Error Calculations

KNN Cluster Model

Using KNN Clustering, both Billboard and Spotify ranks were considered for clustering to visualize clusters formed by K-means, providing insights into the inherent groupings of the song rankings. The hypothesis was that there would be groups the followed around the axis at which Spotify and Billboard ranking were equal. The choice of k, the number of clusters, is set arbitrarily at 20, as each week had 200 ranking slots it seems that 20 clusters that could show the trends in the data without being overfitted. The cluster plot produced from using k equal to 20 is shown in Figure 17. Even though K was chosen arbitrary, the KNN clustering model underwent k-fold cross-validation, exploring different k values to identify the optimal parameter for maximizing accuracy. The results of k-fold cross-validation for k values between 1 and 20 can be seen in Figure 18. The grouping in the plot around the x=y axis is where most of the group would be if Spotify ranking correlated significantly Billboard rankings. Of course, there are many groups along that axis but there are also two very interesting groups outside of those. plot shows two distinct groups of songs based on their Billboard and Spotify rankings. One group of songs has high Billboard rankings and relatively low Spotify rankings. The other group of songs has low Billboard rankings and relatively high Spotify rankings. The KNN model evaluation shows, based on the number of clusters, the accuracy range from 0.01237113 to 0.03092784.

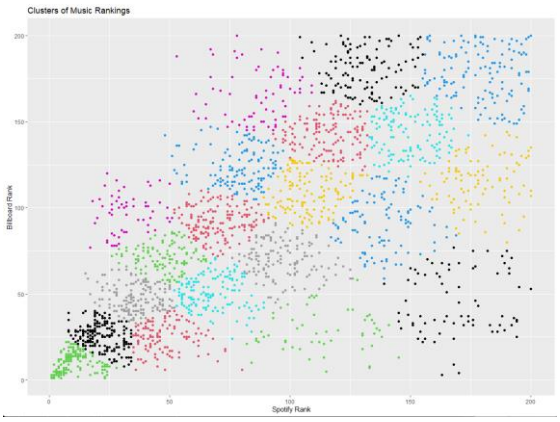

Figure 17: KNN Cluster Plot (K = 20)

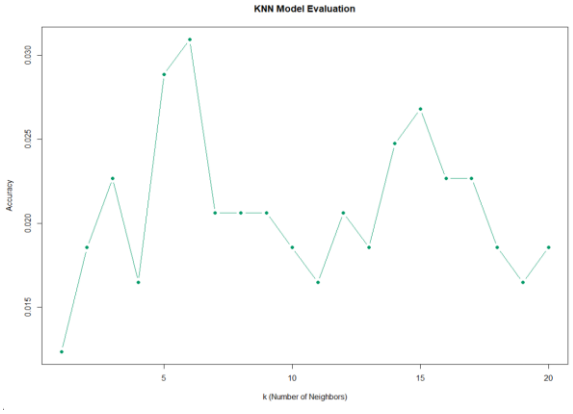

Figure 18: KNN Model Evaluation (K = 1 – 20)

Time – Series Model (Cross – Correlation)

Finally, the cross-correlation between Spotify and Billboard rankings over 27 weeks was analyzed through a time-series model. Using the function "**ccf**" to visualize the cross-correlation of the temporal relationship between Spotify and Billboard rankings. The chosen lag was based on the number of weeks' worth of data at 27. The results of the Cross-Correlation function can be seen in Figure 19. The statistical significance of the model was assessed through a t-test applied to the cross-correlation coefficients. The results of the t-test are shown in Figure 20. The cross-correlation plot shows a strong positive correlation between Billboard and Spotify rankings at a lag of 0. The peak of the correlation coefficient is 0.87, which is a very high correlation. The plot also shows a significant correlation at a lag of 1. This means that changes in Spotify rankings may precede changes in Billboard rankings by one week. The t-test results show that the cross-correlation coefficient between Billboard and Spotify rankings at a lag of 0 is statistically significant (p-value = 1.766e-06). The t-test statistic is 5.36, which is greater than the critical value of 2.571 at a significance level of 0.05. The cross-correlation plot and t-test results provide strong evidence that there is a close and statistically significant relationship between Billboard and Spotify rankings.

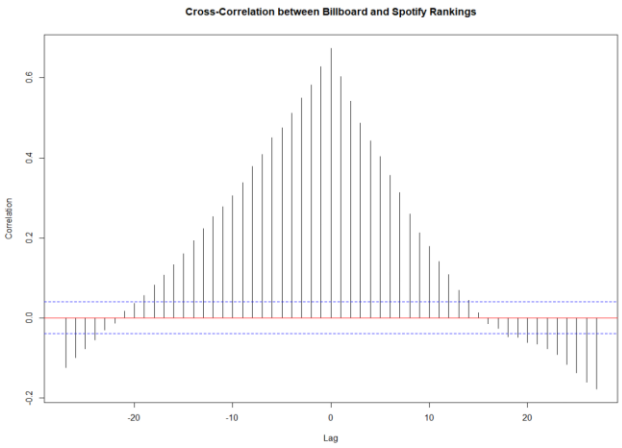

Figure 19: Cross – Correlation between Billboard & Spotify Rankings

```
        One Sample t-test

data:  timecor$acf
t = 5.3604, df = 54, p-value = 1.766e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1090407 0.2393412
sample estimates:
mean of x
0.1741909
```

Figure 20: Time – Series T-Test Results

Discussion

Linear Regression Model

The initial hypothesis that Spotify ranking would be the strongest predictor of Billboard ranking led to the construction of a linear regression model. However, the model's performance, as indicated by the mean absolute error (MAE) and root mean squared error (RMSE), suggested that the linear association alone did not capture the complexity of the relationship. While the linear regression model provided a fundamental understanding, it highlighted the need to explore additional factors that might influence Billboard rankings beyond a straightforward linear relationship.

Correlation Model (Multiple Linear Regression)

To address the limitations of the linear regression model, the analysis expanded to a correlation model. This model delved into multiple linear regression, considering features with strong correlations with the current week's ranking. The correlation model, with the

significant features, significantly improved predictive accuracy compared to the linear regression model. The mean squared error (MSE) and root squared mean squared error (RSME) demonstrated a more precise prediction of Billboard rankings. However, the error was still too large to determine that the model is mathematically accurate. Even with the optimization to the original linear regression model with the multiple linear regression the model wasn't accurate enough. While there is a measurable relationship between the two it is not strong enough to make accurate linear model from it.

KNN Cluster Model

KNN clustering was introduced to identify inherent patterns and groupings within the dataset. The cluster plot revealed intriguing groupings that deviated from the x=y axis, suggesting distinctive trends in the data. The exploration of different k values in k-fold cross-validation indicated an optimal cluster count. While the choice of k was arbitrary, the results showcased a range of accuracies, and the grouping plot hinted at two distinct song groups based on Billboard and Spotify rankings.

Time – Series Model (Cross – Correlation)

The time-series analysis, employing cross-correlation, brought a temporal dimension to the exploration. The plot revealed a strong positive correlation at a lag of 0, suggesting an immediate relationship between Spotify and Billboard rankings. The t-test results provided statistical significance to the cross-correlation coefficients, reinforcing the conclusion that changes in Spotify rankings significantly influence changes in Billboard rankings.

Conclusion

The comprehensive analysis of the dataset utilized multiple models, each shedding light on different aspects of the relationship between Spotify and Billboard rankings. The data analysis revealed a positive correlation between Spotify rankings and Billboard rankings, indicating a relationship between rankings of the two charts. However, the relationship is not

perfect, suggesting that there are other factors that aren't accounted for in Spotify data. The project initially operated under the assumption of a straightforward linear relationship. However, as the analysis progressed, the need to consider additional features and explore the temporal dimension became evident. This evolution was crucial to gaining a more nuanced understanding of the factors influencing rankings. The progression from linear regression to correlation analysis, KNN clustering, and time-series modelling reflects an increasing sophistication in the modelling approach. Each model added a layer of complexity, uncovering distinct facets of the Spotify-Billboard relationship.

In subsequent exploration, it would be interesting to combine streaming data from sources outside of Spotify. While Spotify is the largest streaming service by market share it is only 30% of the global market share (Wikipedia Contributors, 2019). There is a large part of the global market that is not accounted for in this data that should be because the Billboard Global 200 chart reflects global song popularity. The gaps that need to be filled are probably from other streaming sources. There are many other streaming services globally available that count towards Billboard ranking that could all more accuracy to the models. It is also important to note that most of Spotify's streams come from North America and Europe (Wikipedia Contributors, 2019) which could mean that steaming patterns in those geographic locations is overrepresented in these models.

In conclusion, the synthesis of linear regression, correlation analysis, clustering, and time-series modelling provided a multifaceted understanding of the Spotify-Billboard relationship. This project serves as a foundation for future explorations, encouraging a continuous refinement of models and approaches to unravel the ever-changing landscape of music chart dynamics.

References

Billboard. (2023, December 2). *Billboard Global 200*. Billboard. https://www.billboard.com/charts/billboard-global-200/

Spotify. (2023, November 30). *Spotify Charts - Spotify Charts are made by fans*. Charts.spotify.com. https://charts.spotify.com/charts/view/regional-global-weekly/latest

Wikipedia Contributors. (2019, June 13). *Spotify*. Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/Spotify

Wikipedia Contributors. (2023, November 27). *Billboard Hot 100*. Wikipedia. https://en.wikipedia.org/wiki/Billboard_Hot_100#:~:text=The%20Hot%20100%20is%20ranked