# Analysis of Pitchfork Album Reviews

J.T. Liso, Caleb Fabian, and Richard Riedel

***Abstract*— Pitchfork is the premier website for album reviews of all genres of music. Critics commonly claim that the reviewers are biased toward specific artists or genres. Is this an accurate sentiment? This project seeks to analyze the patterns within album reviews on Pitchfork. We collected all album reviews from Pitchfork's conception in 1999 and performed various numerical analyses, such as normal distributions, linear regression, and frequency analysis, as well as time-series analyses on the data. Afterwards, the data was also compared to other information metrics such as Metacritic review scores and YouTube views. From this, we discovered that individual reviewers tend to have biases, but as a whole, Pitchfork is only slightly biased.**

## I. MOTIVATION

As online music review sites become more prevalent, critics claim that certain sites have certain biases toward various genres and artists. People rely on these websites for recommendations, so the accuracy of the content is vital for rising artists' success. Since Pitchfork.com is the premier mainstream media source for music reviews and news, we have decided to explore their reviews for inaccuracies and biases.

To do this, we collected all reviews saved on Pitchfork's album reviews endpoint[1], which dates back to 1999. We compared the ratings and other metadata from the reviews with online information such as Metacritic scores for the albums and YouTube streams to compare the review score to the popularity of the music. Additionally, we sorted the data based on artist, reviewer, genre, month of review, day of the week of the review, and time between album release and the review. From these sortings, we exploited any of the biases that Pitchfork may have within its organization, whether that is biases from Pitchfork as a whole or individual biases between reviewers.

## II. DATA COLLECTION

Data collection was divided into two separate parts. First, a web crawler was designed and created to scrape the Pitchfork album reviews endpoint, only collecting and sorting the relevant data. Second, we created scrapers for access and organization of YouTube streaming metrics.

### A. Pitchfork Reviews

Pitchfork contains an archive of every album the site has ever reviewed since January 1999, totaling to over 18,000 different reviews. Each review contains a header with the album name, album cover, the artist, and the score given to the album by the reviewer (on a scale from 0 - 10 in .1 increments). Following the header, information about the reviewer, the date of the review, and the textual review of the album in about 1000 words are listed. The review also declares other information such as if the album is a "Reissue" or considered "Best New Music" (given to new albums with 8.0+ scores that are especially good).

We created a DynamoDB database with the following fields of the review:

1) Artist Name - Album Name (key of the table)
2) Artist Name
3) Album Name
4) Album Genre
5) Album Release Year
6) Review Author
7) Review Publish Date
8) Review Title
9) Review Score
10) Review Content

Through Pitchfork's API we gathered every review on the site by navigating through the artists in each genre and then finally getting every review for those artists. This returned a JSON object that had the information we required, plus plenty of extra information.

### B. YouTube Streams

Ideally, we would compare the Pitchfork reviews to album sales, but since the music industry has turned to a stream-based industry, we decided to collect information from a widely used streaming service, YouTube.

Our YouTube webscraper collected the total number of streams on a playlist for each album that is reviewed on Pitchfork. Because this step is dependent on the names of the albums and artists that appear on Pitchfork, this step could not proceed until the data collection of the Pitchfork Reviews is completed. YouTube was searched for the albums by appending the URL https://www.youtube.com with /results?search_query=, the key of the database (Artist Name - Album Name) and then &sp=EgIQA1AU, which signified to YouTube that we only wanted playlists. The first playlist from the query was chosen and the view count was taken from the whole playlist (as opposed to individual songs).

One perceived problem was that YouTube allows many users to post videos or playlists of albums that may have already been uploaded by someone else. Due to this, there may be multiple playlists for a single album. To account for this we chose to use the first playlist that YouTube returned for the search. Another problem that we encountered was that there were playlists with the same or similar title that were not the actual album. These had to be filtered out of the final data.

[1]https://pitchfork.com/reviews/albums/

This stream count combined with the Pitchfork metadata described previously was stored in one cohesive DynamoDB database. This made the various analyses much easier as we could just poll the database for specific artists, reviewers, and albums to get the corresponding data.

### C. The Database

The database that was used was a DynamoDB instance on AWS (Amazon Web Services). This is a noSQL database that requires a key value to query. The column headers of the database looked as follows:

- Artist Name - Album Name (key of the table)
- Artist Name
- Album Name
- Album Genre
- Album Release Year
- Review Author
- Review Publish Date
- Review Title
- Review Score
- Review Content
- YouTube View Count
- YouTube URL

These values could then be pulled in bulk or queried for analysis.

## III. Quantitative Methods

Various quantitative methods were used to analyze the data numerically. Within each of these methods, data was grouped and organized by the various fields outlined in the data section. These methods include creating normal distributions and linear regressions as well as conducting frequency and time-series analyses. Each method is discussed below.

### A. Normal Distributions

With any statistical analysis, one of the first methods always tried is fitting the data to a normal distribution. A normal distribution *N* for dataset *X* is defined as

$$X \sim \mathcal{N}(\mu, \sigma^2) \tag{1}$$

where $\sigma^2$ is the variance (or the standard deviation squared) and $\mu$ is the mean [1]. Normal distributions were used on scores as they were the only numerical data. However, we calculated several different normal distributions by grouping by genre, artist, reviewer, day of week, year, and month of year. Within each grouping the different number of normal distributions were compared to find patterns. These patterns were visualized through probability density curves or box plots (as shown in Section IV).

Pandas[2] and NumPy[3] were used as the tools for calculating the normal distribution. Matplotlib[4] and Seaborn[5] were used for the plotting.

[2]https://pandas.pydata.org/
[3]http://www.numpy.org/
[4]https://matplotlib.org/
[5]https://seaborn.pydata.org/

### B. Linear Regression

Linear regression is another common technique in data analysis. It attempts to find a linear equation to model two variables that best represents a trend in the data. One variable, *x*, is considered to be the independent variable and the second, *y*, is dependent on the independent variable by an estimated slope *m* and estimated *y*-intercept, *b* [2]. This is described by the following equation:

$$y = mx + b \tag{2}$$

In every linear regression analysis we conducted, the dependent variable was the album review score. However, the independent variable differed. We explored the effects of streams, day of week, year, month, and album release number as the dependent variables. A positive slope indicates a direct relationship between the review score and the dependent variable. A negative slope indicates an inverse relationship between the review score and the dependent variable.

Scikit-Learn[6] was used for the calculation of linear regression and Matplotlib was used for the visualization.

### C. Frequency Analysis

Frequency analysis is simply analyzing the number of times a specific instance occurred. These are often visualized through histograms and bar charts. For our project, frequency analysis was used to explore the popularity of artists and genres, as well as activity by reviewers, days of week, months, and years. These were summed by both occurrence (each review counted as one) and by summing the streams for artists and genres. By plotting the frequency analysis of these fields, we could easily figure out if certain fields were disproportionately more common than others.

Pandas was used as the data aggregation tool for the frequency analysis and visualized through Matplotlib.

### D. Time-Series

The final quantitative method used for our project was simple time-series analyses. This answers the questions of how the review scores have evolved over time and how specific times affect the results. To do this, we used a combination of the previously mentioned methods as well as simple plots with time as the x-axis to explore how the scores changed through time.

Pandas was used as the data aggregation tool for the time-series analyses and visualized again through Matplotlib and Seaborn.

## IV. Results

Through the methods outlined in Section III, we generated various graphs to explore potential biases within Pitchfork. These results as well as some preliminary raw results are outlined in this section.

[6]http://scikit-learn.org/stable/

### A. Raw Data Analysis

Before delving into the exploration of biases, we will go over some general features of the data. Overall, the average review score was a 7.02 with a standard deviation of 1.27. As was expected, Rock was the most often reviewed genre with 6546 total reviews, and Global music the least reviewed with only 175 reviews. We noticed that as a genre received more reviews, the score tended to slowly decrease. The two least reviewed genres had the highest scores while the most reviewed genres had scores closer to the mean.

Surprisingly, 82 different albums received a perfect 10 out of 10 score, and 3 albums received a sad 0 out of 10. Furthermore, some artists are reviewed more often than others. There were 5 artists that had the most amount of reviews of 13 - Animal Collective, Prince, Bonnie "Prince" Billy, Brian Eno, and Neil Young. These artists have large discographies that cause the high amount of Pitchfork reviews. However, we did find that these artists had significantly higher mean scores than the average for their respective genres.

Pitchfork has had over 400 unique reviewers, most of whom have only reviewed less than 50 albums. We found no trend between the amount of reviews an author has completed and his/her average review score. However, we found that great disparities exist between reviewers with a lower amount of reviews than reviewers with a higher amount of reviews who more closely represent the mean scores discussed previously.

### B. Author Biases

As is human nature, each individual author contained his/her own biases toward certain genres and artists. We looked at the top 5 authors with the most reviews and noticed some large disparities between average scores of various genres. For example, the author with the most reviews, Ian Cohen (who is also one of the harshest reviewers), strongly favored Folk music, with an average score of 6.78, and is harsh on Electronic music with an average score of 5.96. This is a significant difference of .82 between 2 genres that Cohen frequently reviews. Cohen and the 4 other top authors' favorite and least favorite genres can be seen below.

| Author | Favorite Genre | Least Fav. Genre |
|---|---|---|
| Ian Cohen | Folk (6.78) | Electronic (5.96) |
| Joe Tangari | Global (7.71) | Rock (7.22) |
| Stephen M. Deusner | Folk (7.28) | Electronic (6.75) |
| Brian Howe | Rap (7.18) | Pop (6.67) |
| Stuart Berman | Experimental (7.36) | Pop (6.74) |

On average, there is a .59 difference between the author's favorite and least favorite genre. This is a pretty significant difference. However, within these genres, we did not find any biases toward a particular artist for each reviewer. Their reviews within each genre were pretty diverse over the artists in the genre.

### C. Artist Biases

Artists tend to receive better scores as they release more albums. Consequently, artists with large discographies tended to have higher average review scores. The linear regression for the number of albums had a slope of .028 which is quite a strong positive correlation. The visualization for this linear regression and the means for the numbered album can be seen in Figure 1.

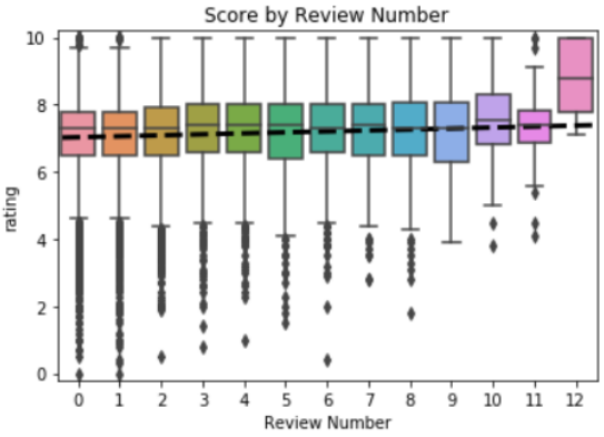


Fig. 1: The number of albums released positively correlates with review score.

Additionally, music fans have claimed certain bands to be "Pitchfork favorites" such as Grizzly Bear, Fleet Foxes, and Bon Iver [3]. Thus, we decided to compare the Pitchfork scores of the albums reviewed by these artists with their respective Metacritic scores. Metacritic is a website that aggregates reviews of music, video games, and other forms of entertainment. From this comparison, we found that, on average, Pitchfork rated albums by the "Pitchfork favorites" .17 points higher than Metacritic. This shows that Pitchfork does have a slight bias toward these artists.

### D. Effects of Time

Many people would agree that certain times of the year call for specific genres of music. For example, Rock, Jazz, and Global music receive the highest scores in September. Electronic music receives the highest score in October. Rap does the best in May. Pop excels in April. Folk music does best in January. Experimental music receives high scores in June, and Metal does the best in March. However, there is no evidence that a certain day of the week causes any review boosts.

Additionally, we found that over the 18 years of Pitchfork's existence, the review scores have slowly increased. Using a linear regression, we found the slope of the regression to be .018. Although this is a small positive slope, it is still significant. More importantly, we found that as the time increased between an album's release and the date of the review, the higher the review score is. We called this the "classic effect" as albums that are rated long after their

review are more likely to be considered classic examples of a specific type of music. This can be seen in Figure 2.

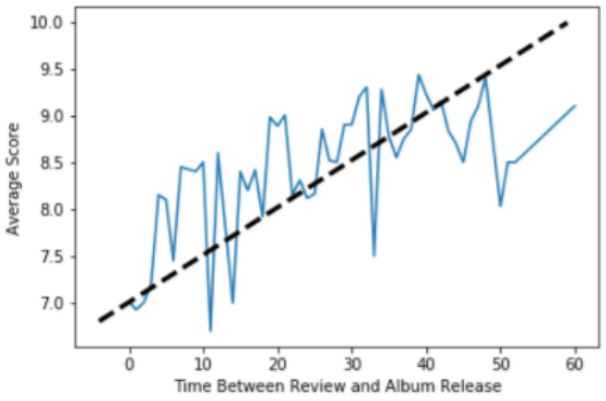


Fig. 2: There is a strong positive correlation between time between album release and review with its score.

### *E. YouTube Streams*

As we hypothesized, there is little correlation between album review score and streams. We found that albums that there was no trend whatsoever between album score and number of YouTube streams. Both low and highly rated albums had high and low streams. In fact, the highest streams came from albums with a rating of 2.8. This group includes popular artists such as Eminem, Bassnectar, and Ed Sheeran.

However, we did find that genres such as Global and Pop had unproportionally higher stream counts than the number of reviews Pitchfork has done on those genres. Global may seem like an odd genre to have such high popularity, but it contains popular artists such as M.I.A. and Bob Marley that account for high amounts of streams.

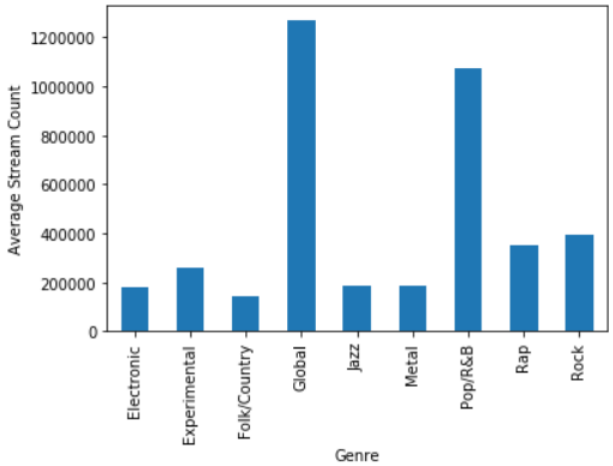


Fig. 3: Global and Pop music have much higher stream counts than other genres. These two genres account for a very small fraction of Pitchfork reviews

## V. Limitations and Issues Encountered

Originally, as discussed in the proposal for this project, we wanted to gather streaming information from both YouTube and Spotify. However, Spotify has deprecated their API making it near impossible to obtain streaming information. The only solutions we could come up with to collect the streaming counts from Spotify were too difficult, were prone to error and could potentially not even work after we finished. Thus, we decided to abandon the Spotify stream counts.

Additionally, when collecting the YouTube stream counts, sometimes streams for unrelated videos would be returned in the total stream count. This issue was noticed when some albums we had never heard of had extremely high stream counts. When this issue was noted, we had to go through and manually check the suspicious stream counts and remove them from the database in order to not skew our results.

Finally, we wanted to include textual analyses of the actual review content in our report, but the two textual analysis methods we had time to explore, Latent Dirilecht Allocation (LDA) and Sentiment Analysis, returned inconclusive results. We attempted to use LDA to find common words for each of the eight genres, but LDA returned the same groups of words for each genre. Sentiment Analysis only returned that an unreasonably small amount of the reviews (in the 10s) contained ”negative” language which disagreed with the score distribution. With more time, perhaps we would have been able to explore the parameters of these analysis methods to get more conclusive results.

## VI. Related Works

We are not the first group of people to explore Pitchfork review data. One such instance is Neal Grantham exploring the correlations of Pitchfork’s ”Best New Music” grouping in 2015. ”Best New Music” is music Pitchfork rates highly and receives special promotion on the website. Grantham did not find any notable correlations between albums and ”Best New Music” designation [4].

Additionally, Nolan Conaway completed a similar analysis to us in recent years. He explored statistical heaping methods on the data as well as also exploring the ”Best New Music” category as Grantham did. Conaway’s conclusions were similar to Grantham in that little to no correlation exists between the Pitchfork scores and categories such as ”Best New Music” as well as little evidence for biases. He did, however, find that scores are most likely to be given on an exact number (*.0) rather than any other decimal point. Some inspiration for our methods came from his repository as some of his analyses were similar to ones we desired [5].

## VII. Future Work

Although we have some results, there is still more work that can be done in regards to the analyses. If Spotify ever re-instates its API, it would be much more meaningful to use the streaming data gathered from Spotify as opposed to the streaming data from YouTube since Spotify is arguably the most used music streaming service. Additionally, we could look into album sales as a comparison metric as well, but this may not be useful as most people do not purchase albums anymore. Finally, different methods of textual analyses can be explored and used since the ones we tried did not work

well. Do we just need to fine-tune the parameters of LDA and sentiment analysis or do we need to try a completely different textual analysis method to get useful results?

Most of the work conducted in this project was exploratory as fans of music. However, this project is preliminary work for something J.T. has been interested in doing for a while. Is it possible to predict the score an album will receive based on various parameters? In order to do this, we will need to find a proper textual analysis model as this will be a useful metric to learn the prediction model on. The goal of this additional project would be to use the data we collected from this project, such as reviewer and month of review, to predict what the score would be for the album review. Additionally, we may consider scraping data from music discussion forums to hear what others in the music community are saying about a specific record in order to further build our prediction model.

## VIII. Conclusions

The most important effects on an album's review score comes from the author reviewing the album and the time of the album review. Review authors rate a variety of genres and tend to have large disparities between his/her "favorite" genre and his/her "least favorite" genre. Thus, if an author reviews an album in a genre he/she is not too fond of, the score will be affected (and vice versa). Considering the effects of time, we conclude that there is a "classic effect" in that an album reviewed long after its release receives a significantly higher score than albums reviewed shortly after its release. In addition, each genre has a specific month it receives higher review scores in than the other months, supporting the theory that each genre of music has a specific time of year that fits it best.

Since there was no trend between album score and number of streams or album score and number of reviews, we can conclude that Pitchfork has little influence on the popularity of an album. In fact, genres that traditionally do well in the market (Global and Pop music) still have high number of streams compared to the few reviews Pitchfork does on these genres.

Pitchfork does seem to slightly favor some artists over others, especially those classified as "indie rock", but these biases are not too blatant. Pitchfork generally rates the "Pitchfork favorites" slightly higher than Metacritic, but there is not enough evidence to say that Pitchfork strongly favors these artists. Consequently, we conclude that most of the bias of Pitchfork comes from individual reviewers and unintentional effects of time.

## References


[1] Weisstein, Eric W. "Normal Distribution." From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/NormalDistribution.html [Online; accessed 2017-11-28].

[2] "Linear Regression" http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm [Online; accessed 2017-11-28].

[3] "How Pitchfork Ruined the Taste of Alternative Music Listeners. [Part 1]" https://rateyourmusic.com/list/AfterTheRain/how-pitchfork-ruined-the-taste-of-alternative-music-listeners-part-1/ [Online; accessed 2017-11-10].

[4] Grantham, Neal S. "Who Reviews the Pitchfork Reviewers?" 14 January 2015. http://nsgrantham.com/pitchfork-reviews/ [Online; accessed 2017-10-02].

[5] Conaway, Nolan. "Pitchfork-Data" 24 May 2017. Github repository, https://github.com/nolanbconaway/pitchfork-data [Online; accessed 2017-10-02].