Evaluating and Predicting Trends

Khawar Murad Ahmed, Shreeman Gautam

University of Utah

:

ABSTRACT

There is a lot of data available with regards to art. As such, there is a quantifiable way to predict current trends with regards to songs to see what kind of sound is popular and what isn't. We decided to do clustering, first with the original number of dimensions, and then applied PCA and did clustering. Subsequently, both returned different clusters with regards to years, which meant that we had two different hypothesis. Based on the clusters returned to us, we applied regression in order to predict the popularity of the song. We found that our hypothesis in 13 dimensions was not quite validated, but at the same time the performance was alright, but not the best. We also realized that using 2 dimensions to make such predictions is ineffective, as there are a lot of components that go into making a song a success.

Data

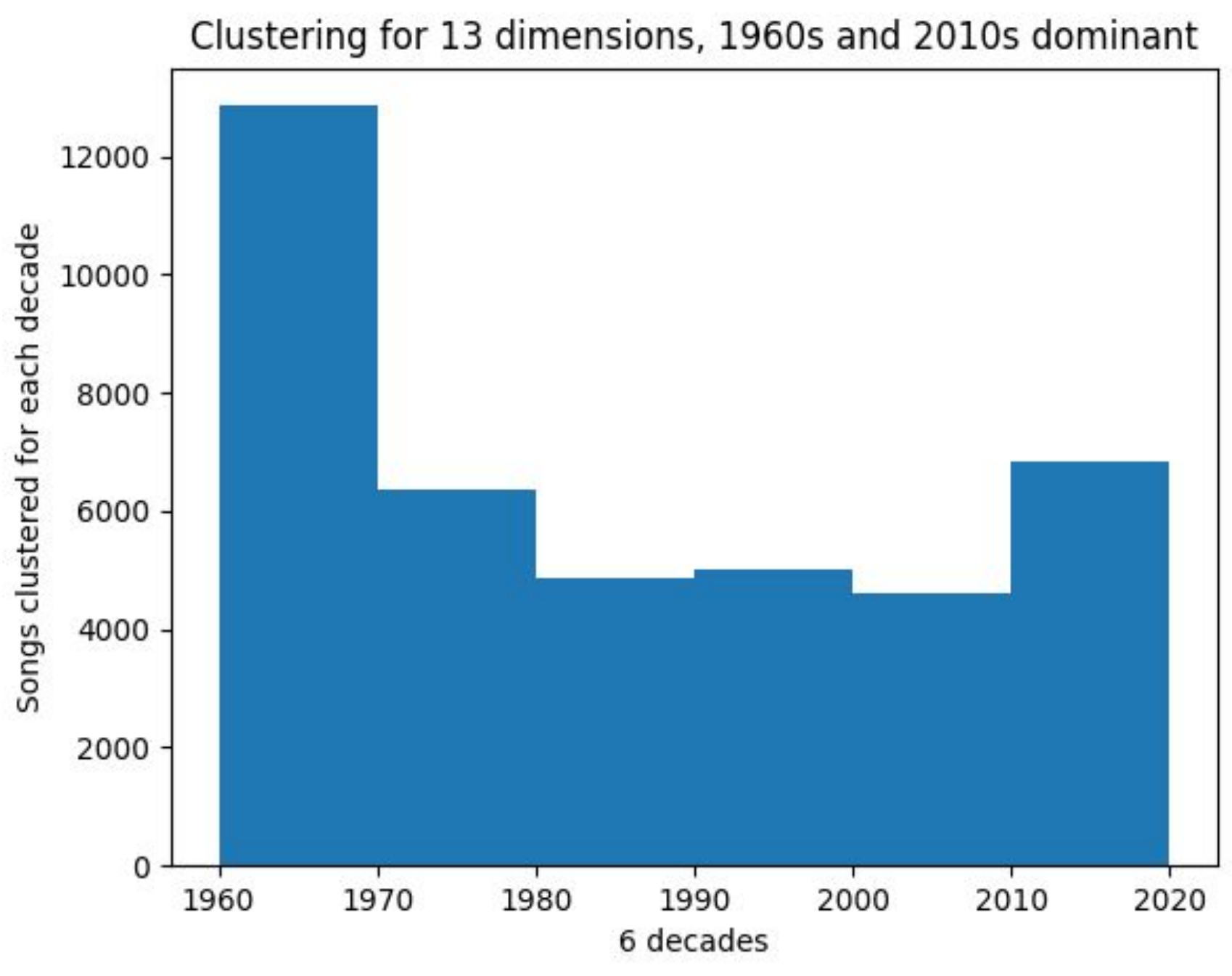
The dataset that was used for this project was found on Kaggle and the dataset was about songs from Spotify. In total, we used 13 dimensions to cluster, and predicted popularity using regression(popularity ranging from 0 to 100). Out of these 13 dimensions, explicit and mode were two categorical features (0 or 1), and the rest were numerical, which are acousticness (range 0 to 1), danceability (range 0 to 1), duration in milliseconds (typically ranging 200k to 300k), energy(range 0 to 1), instrumentalness(range 0 to 1), key(from 0 to 11 starting C as 0), liveness (range 0 to 1), loudness (float ranging from -60 to 0), speechiness (range 0 to 1), tempo (float typically ranging from 50-150), and valence (range 0 to 1). We got rid of release date, artists, song id, and name. In terms of time period, we had songs ranging from 1921-2020. One thing to note is that the popularity variable was taken from Spotify, and was thus by today's standard (older songs would naturally be less popular). For reasons that are somewhat arbitrary, we decided to get rid of all the songs prior to 1960 as we were not familiar with songs or artists from that time period. There were a total of 160,000 songs prior to our preprocessing. After preprocessing, we were left with a total of 121656.

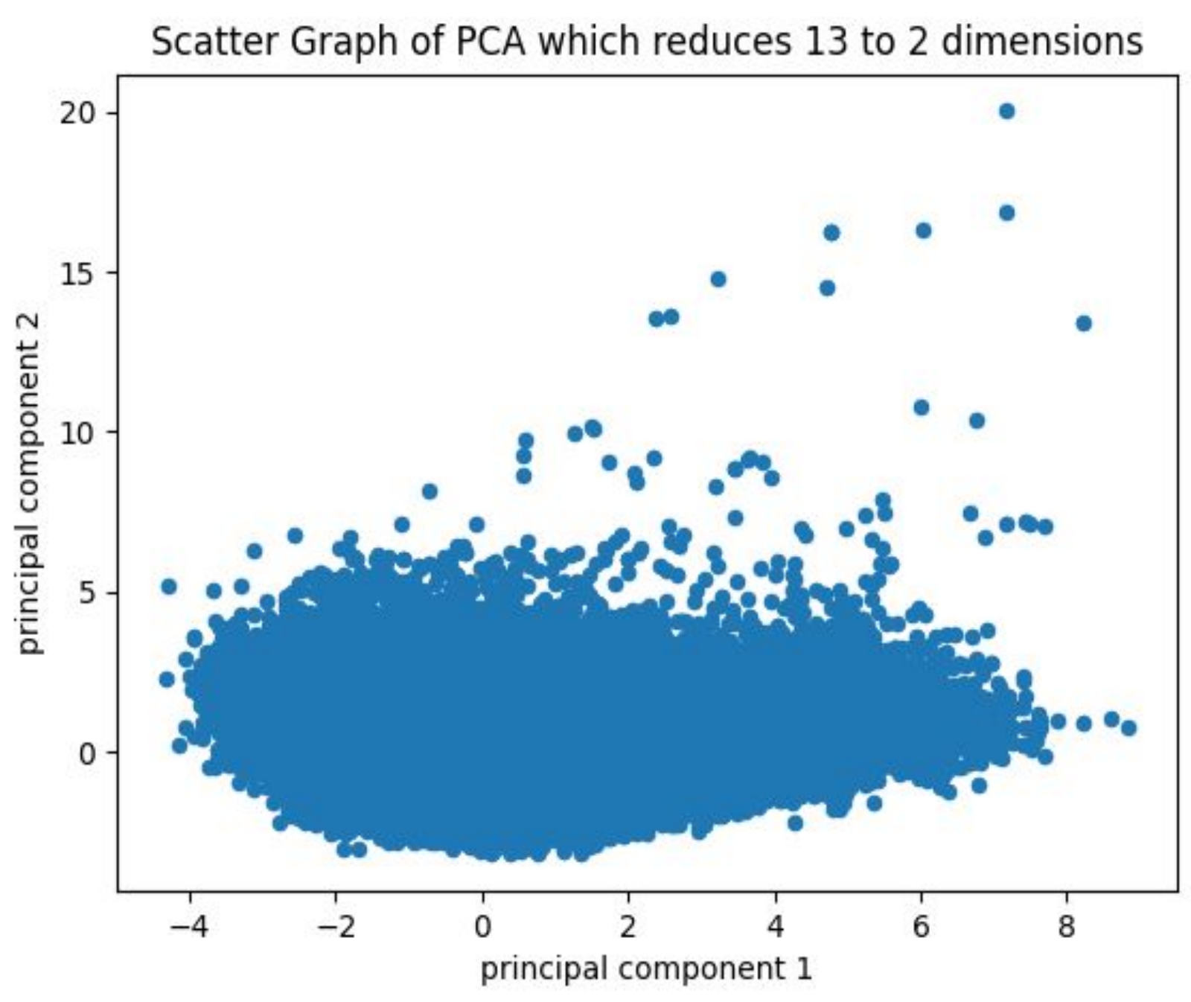
Methodology

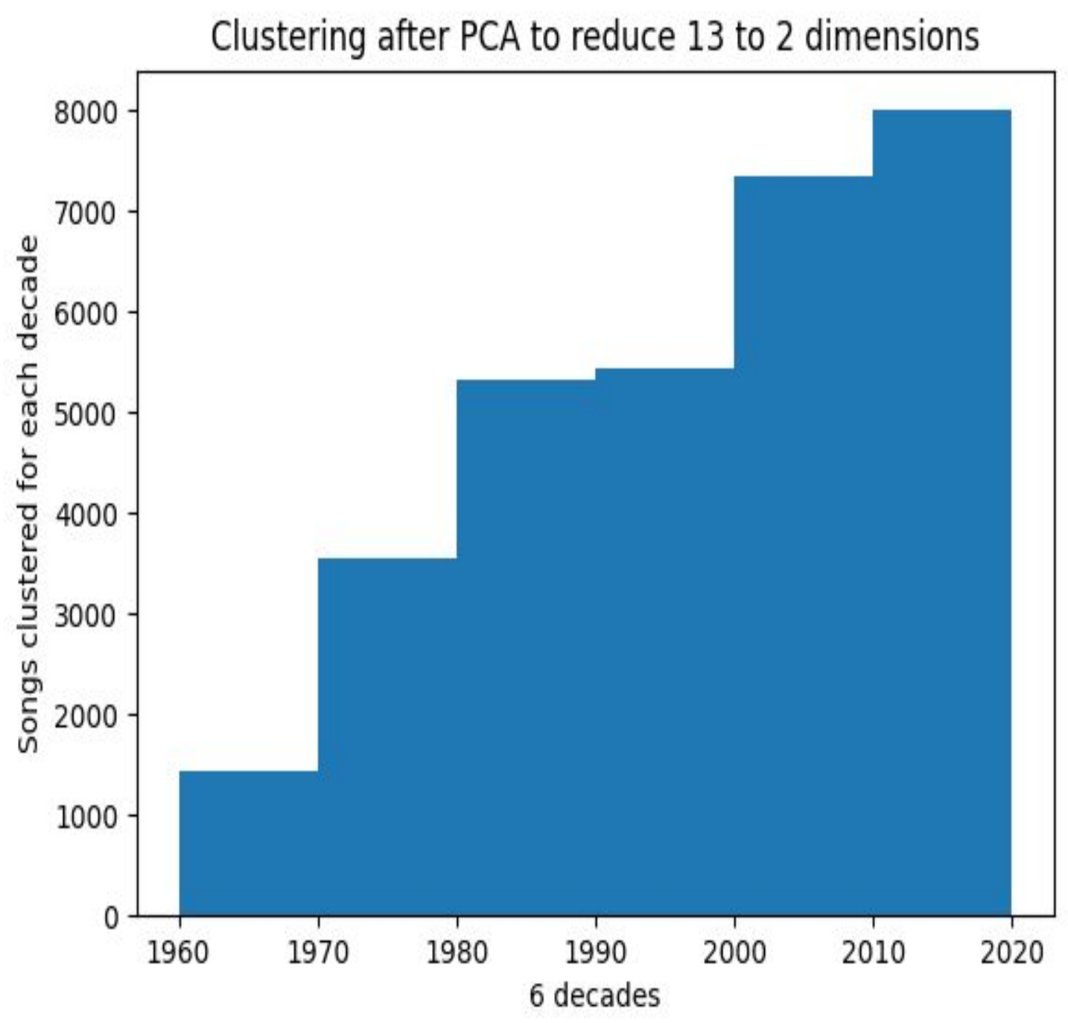
After preprocessing, we used a two-pronged approach in order to best understand our data and use appropriate methods to achieve our goals. First thing we did was cluster our data. This was not our main objective, instead we were just using clustering to see how music has progressed over the years, and whether there were any trends that were noteworthy that could potentially change the way we use prediction. After that, our plan was to apply what we had seen in the dataset, and process data accordingly. Then we would use those trends to filter our dataset which would eventually be used to train a regression model that would predict popularity of a song given the features. We applied our method(clustering and regression) to 13 dimensions (mentioned above), and also used PCA to reduce to 2 dimensions and apply clustering and regression again.

Experiment Results: Clustering

For d=13, we saw an interesting cluster which had grouped together songs from 1960s and 2010s, implying that the sound of those two eras might have been similar. For d = 2 after PCA, we found a standard staircase graph that gravitated towards whatever time period dominated that cluster, and there was no binodal distribution for those clusters. Based on that, we decided to test out regression with two datasets. For the first dataset, we used songs from 2001-19(based on one clustering where 2001-19 are the most dominant) to predict popularity of songs from 2020. For the second dataset, we added songs from 1960-69 and 2010-19 together to predict the popularity of songs from 2020.The training set for 2001-19 had 37900 songs, and the training set for the union of 1960-69 and 2010-19 had 39900 songs. The test set, with songs from 2020, has 1756 songs.





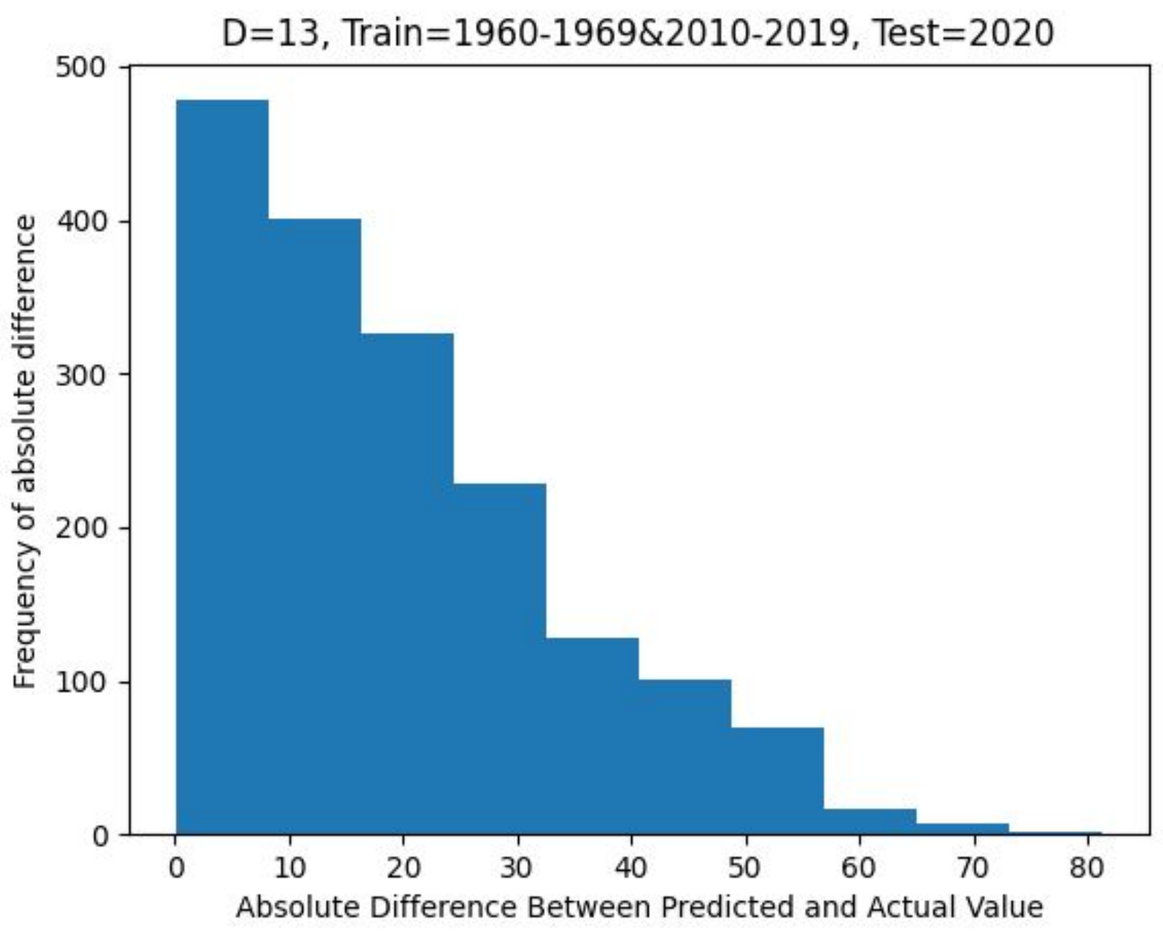


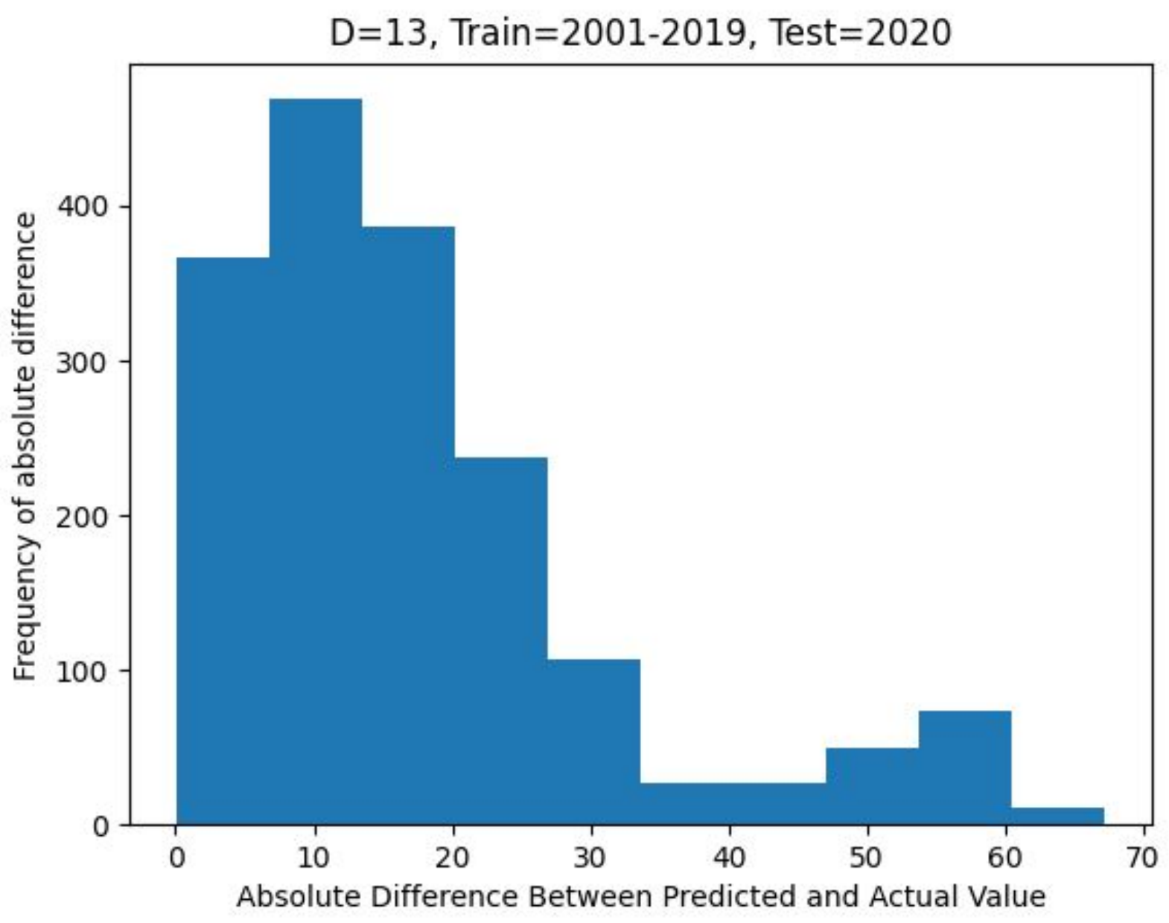
Experiment Results: Regression

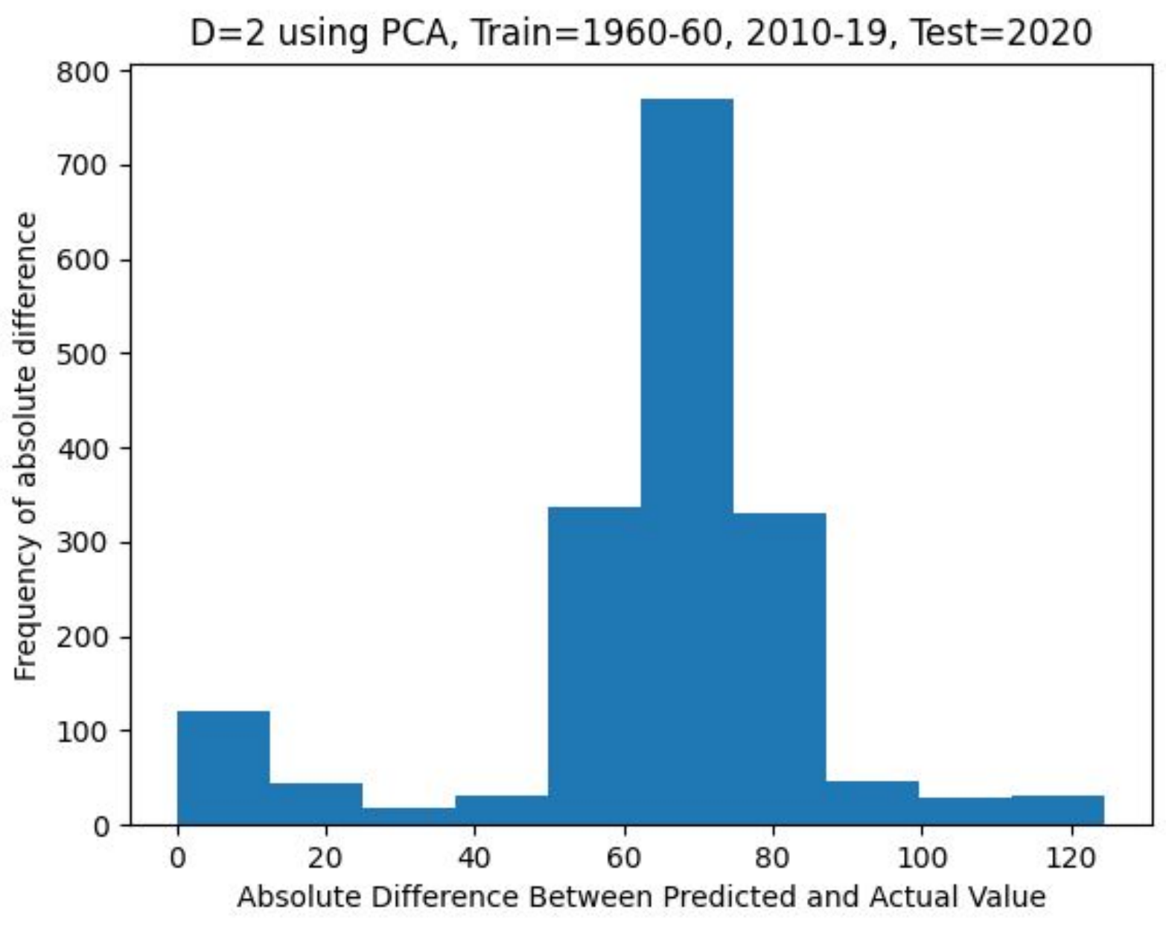
After seeing the clusters, we hypothesized that maybe adding the songs from 1960s to the 2010s and doing regression would give a good popularity score for the 2020 songs. While the performance was not completely bad when compared to just the general dataset, at the same time, it also made the performance of the linear regression worse. We define error as the average of absolute value of observed popularity score minus predicted popularity score given by regression for the relevant dataset. As shown by the table, the error for linear regression was 19.40 when 1960s and 2010s were used for training and 17.56 when 2001-19 was used for training. On the other hand, when PCA was applied, we saw that the performance was really bad, with the error being 64.21 for the dataset when 1960s and 2010s were used, and 63.12 when 2001-19 was used. When we consider that popularity can only be between 1 and 100, we see that this kind of error is really bad. So for that, we think that two dimensions is very less to do regression. For the sake of experimentation, we are going to try out different dimensions to see if for any k, we can see improved performances over the current best model that we have.

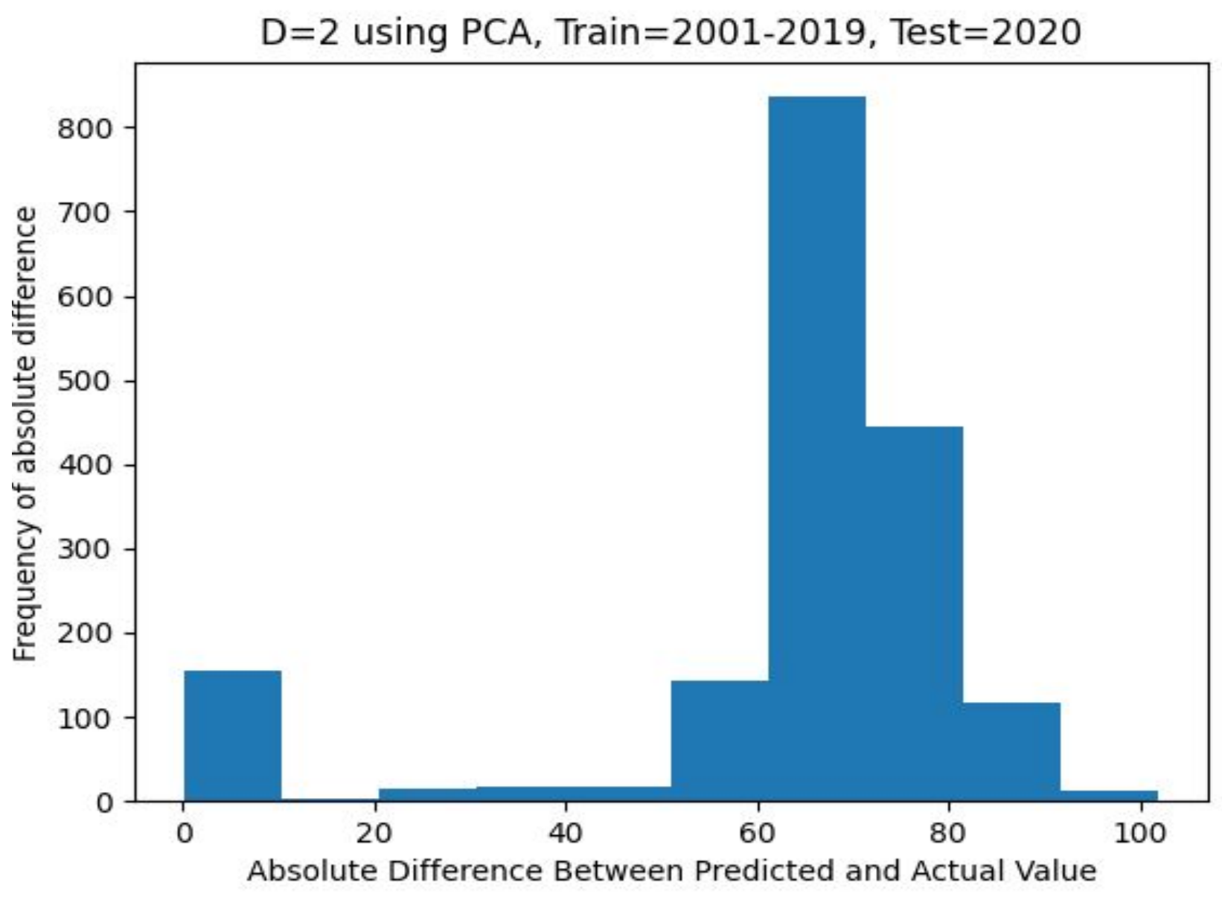
Mean Squared Error for our Regression Models.

Training	1960-69 U 2010-19	2001- 19
D = 2	64.21	63.12
D = 13	19.40	17.56









Discussion

There is a lot that goes into making a song a success, and oftentimes, it is not entirely how the songs sound. As can be seen in two dimensions, there are a lot of songs that lie very close to each other. Essentially, there is no magic formula that makes a song go popular, even though certain trends might be found. That is why, even though our model does not perform exceptionally well, it doesn't perform particularly terrible either. The one factor that we did not/could not include, and is arguably very important, is the artist. Artists are now essentially brands, and the name attached to the song has a major influence on whether people will choose to stream that song, and whether it goes popular. So attaching some sort of coefficient to the artist and incorporating into the model might see the model make more accurate predictions than the ones that we already have.

The other question that pops up is about the data itself. In the data, the definition of popularity was not really mentioned. For example, what makes a song get a popularity coefficient of 95? Currently, short term virality does not always translate to long term popularity. So distinguishing between the two would also be important.