Analysis of Popular Songs on Spotify and Youtube

Renee Singh and Nila Ragu

Summary of Research Questions

We split our research into two parts: analysis depending on platform, and analysis depending on genres and auditory qualities. For our genre and auditory qualities analysis, we focused on Spotify streams.

For our analysis depending on platform, we explored the following questions:

1. What platform are songs more popular on; Youtube or Spotify?
    a. Most genres have significantly more streams on Spotify than views on Youtube.
2. What is the difference in popularity between genres, depending on the platform?
    a. The most popular genres on Youtube are dance-pop, pop, album rock, modern rock, alternative rock, alternative metal, glam metal, permanent wave, Latin, and British soul.
    b. The most popular genres on Spotify are album rock, dance-pop, modern rock, alternative rock, pop, alternative metal, adult standards, permanent wave, glam metal, and neo-mellow.
3. How do the total streams and views vary depending on the release date?
    a. The more recent a song is released, the higher the number of streams and views it probably has.

For our genre and auditory qualities analysis, we focused on the questions:

1. What are the least popular to most popular genres of all time?
    a. The 5 least popular genres are candy pop, chanson, finnish metal, blues, and downtempo.
    b. The 5 most popular genres are dance-pop, alternative rock, alternative metal, permanent wave, and neo-mellow.
2. What are the auditory quality averages by genre?
    a. Chanson, baroque pop, belgian pop, and acid jazz have some of the least variance in auditory quality averages.
    b. Electro house and latin have some of the most variance in auditory quality averages.
3. How do the auditory qualities of a song impact its popularity?

    a. There seems to be no clear correlation between auditory quality values and the number of streams a song gets.
4. How did the popularity of the top 6 genres of all time rise and fall over the years?
    a. The popularity of album rock and alternative rock mostly decrease over time (with a few peaks), while the popularity of pop, dance pop, and modern rock increase over time. Alternative metal has a slight increase for a few years, but then begins decreasing.
5. How did the total average auditory qualities change over time?
    a. Acousticness seems to decrease slightly over time, while danceability and energy seem to increase slightly over time.
6. What are the average auditory qualities of the top 10 popular artists?
    a. Most average auditory qualities seem to be similar between artists, except for Lewis Capaldi, John Legend, and Justin Bieber, who have the highest amounts of acousticness. As well as Billie Eilish and Justin Bieber, who have the highest speechiness.
7. How many songs from the top 10 genres were released every year?
    a. Before 1990, the majority of songs released were either album rock or adult standards. After 1990, those two genres drastically decreased, and there is an emergence of a variety of different genres.
8. Finally, for our machine learning analysis, we aimed to answer the question: Based on a KNeighborsRegressor learning model, can we predict the number of streams a song has given its auditory features?
    a. The model had a training r-squared value of 0.45, and a testing r-squared value of 0.15. There seems to be no or a weak correlation between the auditory features of a song and the number of streams it gets, making it something we weren't able to predict.

Motivation

These questions were explored to understand the overarching trends in popular songs. By finding how different factors can affect the success of a song, we can understand what listeners enjoy. Genre trends can let us know what new genres are emerging and how existing genres are performing on the charts, and auditory quality trends show how the production of music changes over time. Platform trends can show which platform is the best for reaching an audience. This information could be leveraged by current and future artists, and music producers to tailor their music to reach a wider audience, stay in trend, and contribute to their overall success.

Data Setting

We used three datasets, all downloaded from Kaggle:

- "Spotify and Youtube"
- "Spotify - All Time Top 2000s Mega Dataset"
- "Spotify Tracks DB"

These datasets were named Spotify_Youtube, Spotify_2000s, and Spotify_Features accordingly.

Our initial plan was just to use one dataset, but we realized that the context provided would not be sufficient for our analysis. For example, the "Spotify and Youtube" dataset did not include the genres of each song, which would be helpful to know so we can categorize auditory qualities by genre. On the other hand, the "Tops 2000s Mega Dataset" did not include the factors necessary to determine the popularity of each song. By combining the two datasets, we were able to get more context behind the streams, likes, and views of the songs, as well as how the genres of each song relate to the auditory qualities and popularity of the songs.

Furthermore, we decided to use the "Spotify Tracks DB" for our machine learning model, since it included many more songs that could be used to train and test our model.

Method

For this project, our goal was to find some meaningful connection between the auditory qualities of a song and its popularity, and see if it was possible to predict the popularity of a song. To do so, we decided to conduct a more detailed analysis on 2 aspects of our dataset: a streaming platform analysis, genre and auditory quality analysis. In addition, we explored a third aspect through the research question, "Can we predict the number of streams a song has given its auditory features? The next few paragraphs will discuss the methods we used when processing the data, analyzing the dataset, and creating our learning model.

Step 0: Preprocessing the data

Firstly, we had to process our datasets. We had 3 datasets total:

- Spotify_Youtube.csv
- Spotify_2000s.csv
- Spotify_Features.csv

For our dataset analysis, we merged the Spotify_Youtube dataset with the Spotify_2000s dataset. This dataset was named 'spotify'. The datasets were joined through an inner-join, merged by their song title and artist name columns. This was to ensure that each song had a value in the corresponding genre column.

For our learning model, we merged the Spotify_Youtube dataset with the Spotify_Features dataset. While we initially wanted to use the 'spotify' dataset for this section as well, we found that it was too small to fit the model accurately. Instead, we opted for the new dataset, which had an adequate amount of observations. This dataset was named 'spotify_ml'.

Step 1: Platform analysis

In order to decide which streaming platform to focus on, we conducted a platform analysis. We had two options, Spotify streams or YouTube views. This section mainly consisted of 3 bar graphs and one line plot. The first bar graph was formatted to be a segmented graph, and depicted the views and streams of all the songs on the dataset, grouped by genre and sorted in descending order. The other two bar charts were similar to the original, only having views or streams grouped by genre. The final graph was a line plot depicting the views and streams of songs, grouped by their release year. The graph had two lines, one for views and one for streams. In the end, we found that songs generally got more streams than views, and so we decided to use only streams for the rest of the project.

Step 2: Genre and auditory quality analysis

The aim of this section was to look for more insight into the auditory qualities of different genres and look for any patterns or correlations. To do so, we created the following graphs:

- Graph 1: Least Popular to Most Popular Genres of All Time
    - A bar chart ranking all the genres based on their total views. This graph allowed us to make our analysis more concise by picking the most streamed genres for our subsequent graphs.
- Graph 2 and 3: Auditory Quality Averages by Genre & Auditory Qualities of Most Popular to Least Popular Songs (2005-2010)
    - Both of these graphs were both scatter plots. The first graph depicted the average auditory qualities of songs in each genre, and the second graph depicted the auditory qualities of all the songs in the dataset released between 2005 and 2010, sorted so the most popular song was at the top, and least popular was at the bottom.
- Graph 4 and 5: Popularity of Genres Over Time & Average Auditory Qualities Over Time
    - Both of these graphs were line plots. The first graph depicted the popularity of the top 6 genres over time, made by taking the sum of the streams of all the songs in

each genre and plotting it against the release date. While we initially intended to include the top 10 genres, we found that the graph was unreadable. The second graph was a line plot, depicting the average auditory qualities of songs according to release date. This allowed us to look for patterns in how the auditory qualities of songs change over time.

- Graph 6: Average Auditory Qualities Per Artist
    - This graph is a scatter distribution plot, depicting the average auditory qualities of the top 10 most streamed artists in the dataset. This graph allowed us to look for similarities between the auditory qualities of songs by artists with a similar level of popularity.
- Graph 7: Number of Popular Songs Released Per Year in Top 10 Most Streamed Genres
    - For the last graph of this section, we created a segmented bar chart. The graph depicts the number of songs released every year, grouped by genre. This graph allowed us to account for changing genre paradigms.

Step 5: Machine learning algorithm

For the last step of our project, we wanted to fit a learning model and see if it's possible to predict a song's number of streams based on its auditory qualities. Before beginning to fit our model, we created 8 scatterplots for each song's auditory qualities, plotting it against the song's total streams. This allowed us to pick which qualities showed a higher correlation with the streams a song gets. We ended up picking 4 out of 8 of the qualities, as they all had a relatively steeper slope than the others in their scatter distributions. These qualities were Speechiness, Loudness, Liveness, and Instrumentalness. The qualities were normalized to account for differing metrics, ensuring that all of the auditory features were weighted equally.

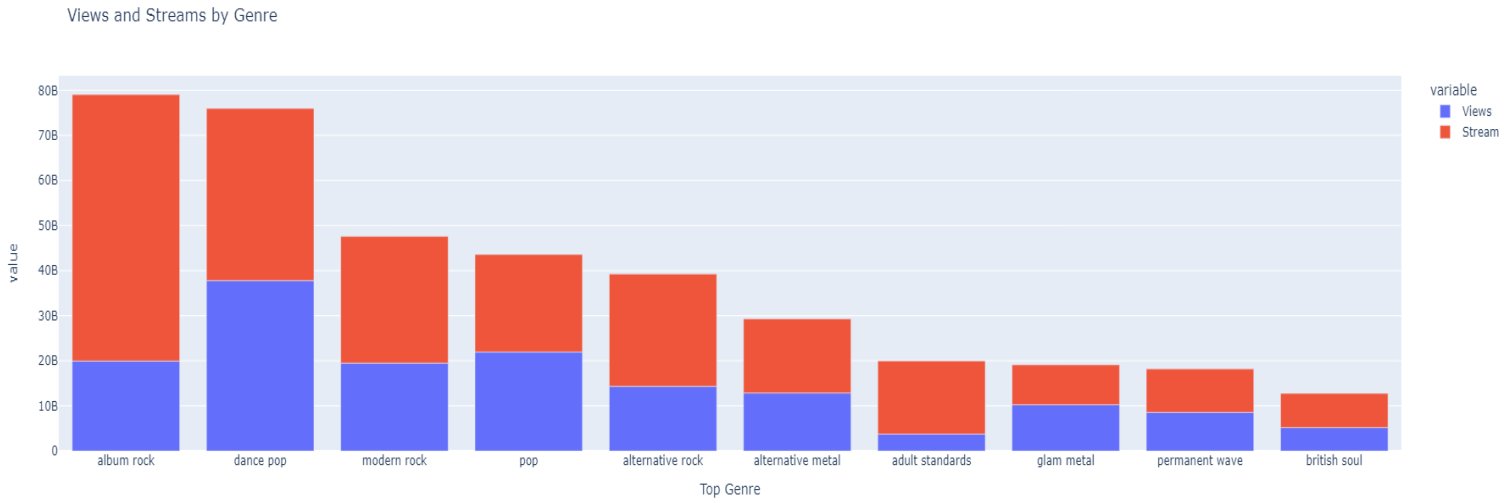
For our algorithm, we used the KNeighborsRegressor learning model. This was chosen as all of our data is all numerical, and KNeighbors is a well-known model with a large amount of documentation available if needed.

In order to test the accuracy of our model, we used a 70/30 train test split, and used r-squared as our accuracy metric. We decided on r-squared due to its 0 to 1 range being easily convertible to a percentage accuracy score.

Results

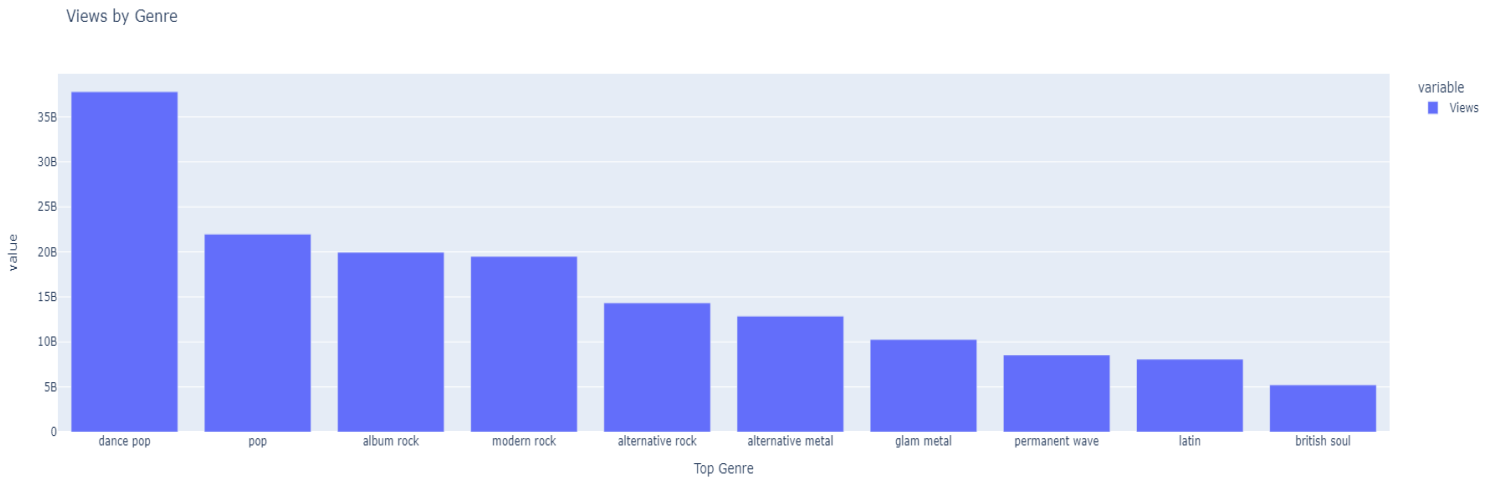
Analysis of Spotify vs. Youtube

1. What platform are songs more popular on; Youtube or Spotify?

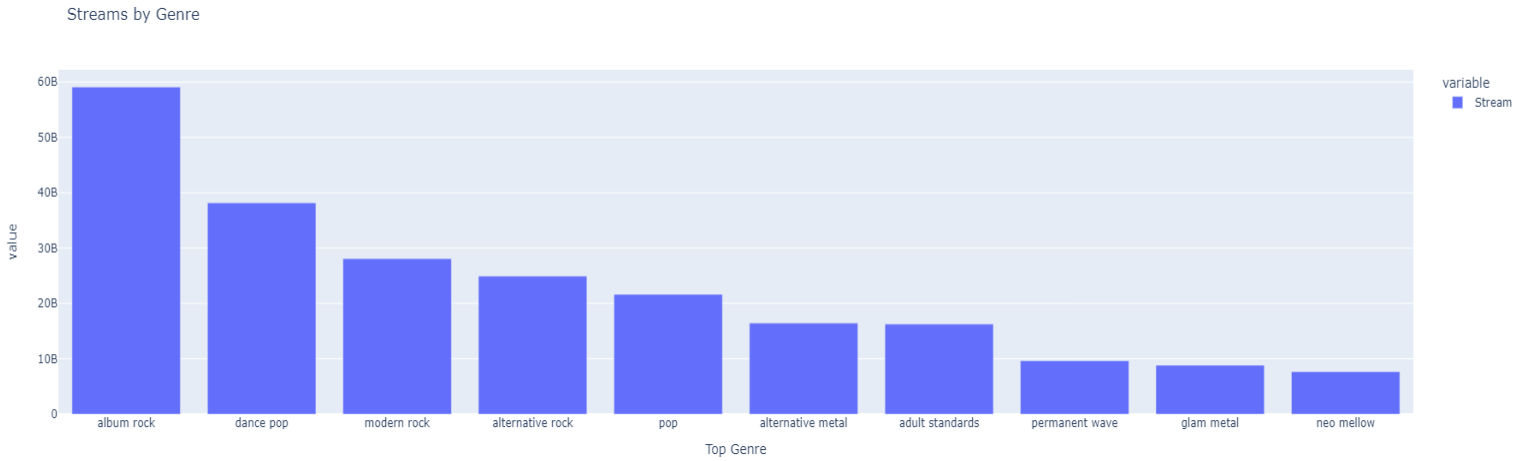


We can see that there are significantly more streams than views for most genres. The two exceptions are pop and glam metal. These two genres are more visual performance oriented than the others, and generally have high budget music videos, which could explain the higher number of views.

2. What is the difference in popularity between songs, depending on the platform?

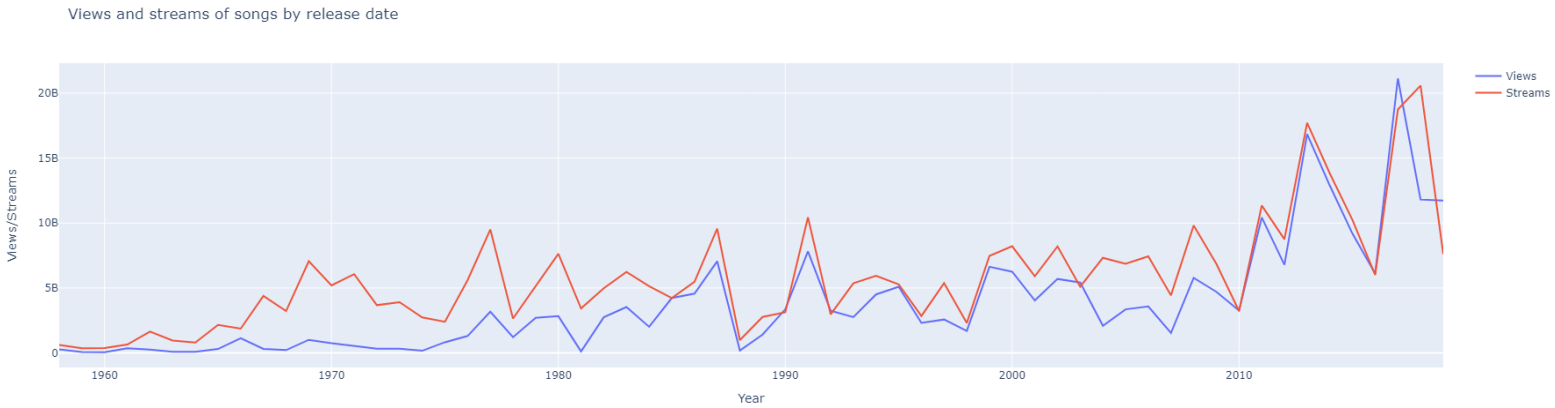


The top 5 genres on Youtube seem to be dance pop, pop, album rock, modern rock, and alternative rock. This could be due to the fact that these genres are more mainstream and as mentioned above, have more music video production and marketing, leading to higher views on Youtube videos.



On Spotify, the top 5 genres are album rock, dance pop, modern rock, alternative rock, and pop, so the same as the top 5 genres on Youtube, but in a different order. The reason the popularity of the genres may be in different order could be due to the amount of music videos produced for each genre. Dance pop and pop may have a plethora of high-budget music videos, which causes them to be the most popular genres for views, while album, modern, and alternative rock may not have music videos, but a higher amount of people listening on Spotify.

3. How do the total amount of streams and views vary depending on release date?

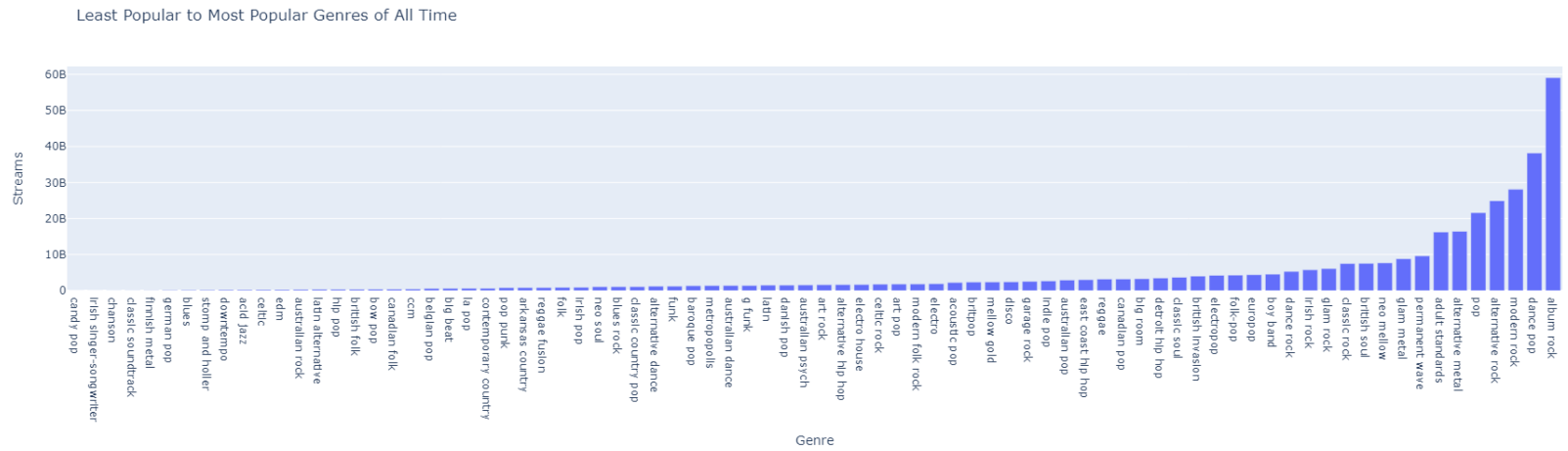


There is a 10.148 billion difference in streams for songs released in 1991 (10.436 billion streams) versus 2018 (20.584 billion streams). On the other hand, there is a 3.989 billion difference in views for songs released in 1991 (7.82 billion views) versus 2018 (11.809 billion views).

The rise of music streaming and viewing platforms could account for the difference in streams and views. Youtube was launched in 2005, while Spotify was launched 2008. People who listen to music released before those years may prefer other methods of accessing music, while newer generations may rely on streaming. Therefore, newer songs that cater to newer generations may have more streams and views.

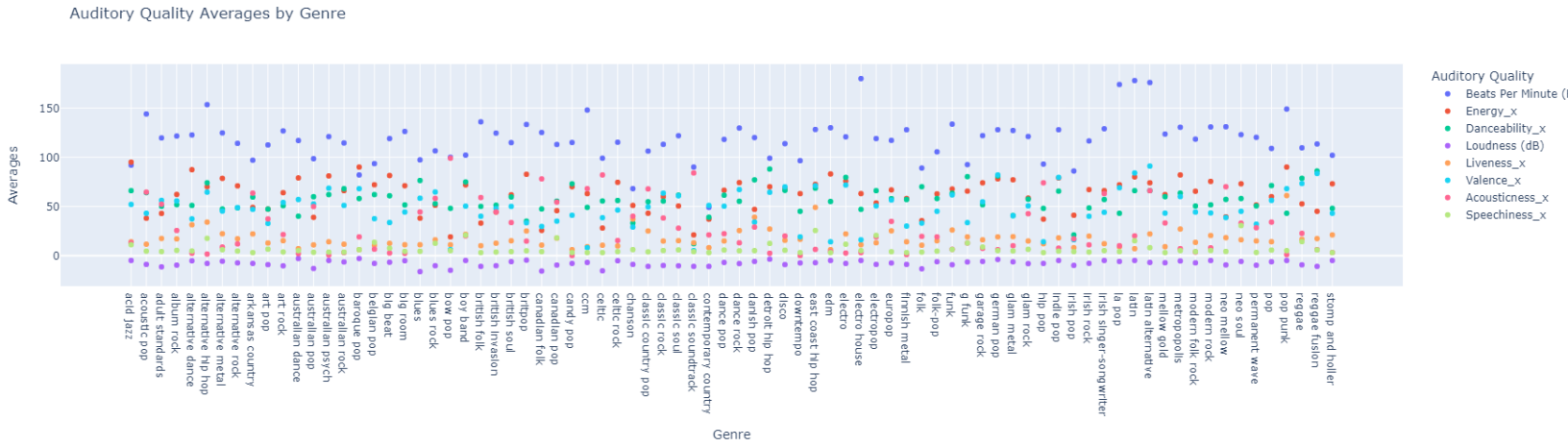
Analysis of Spotify Songs Popularity, Genres, and Auditory Qualities

4. What are the least popular to most popular genres of all time?



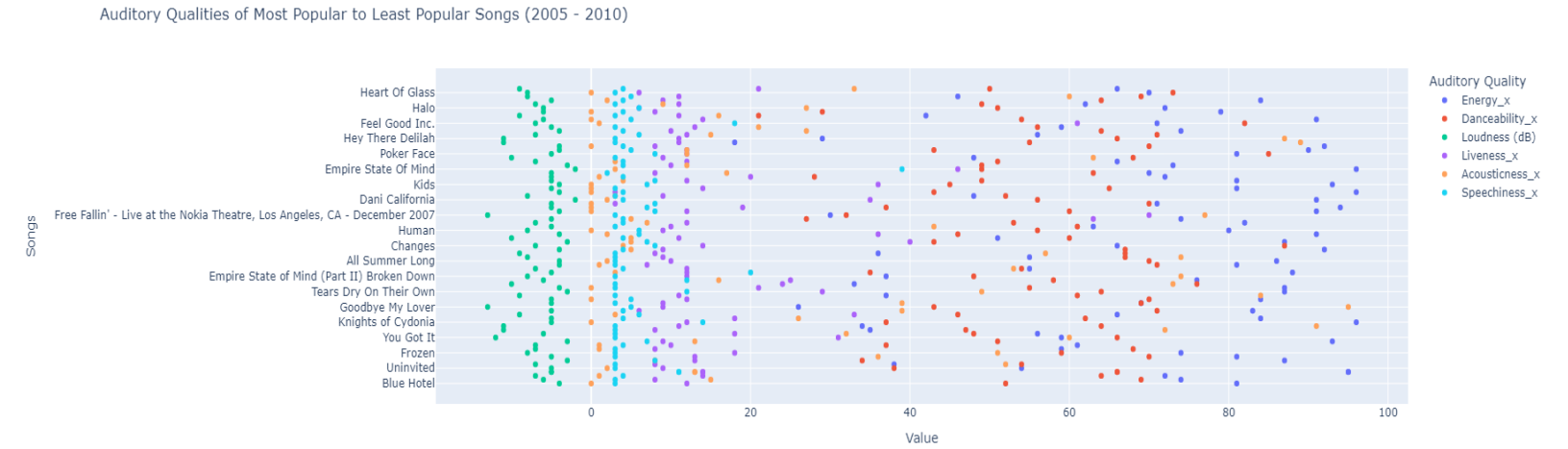
Candy pop seems to be the least popular genre on Spotify, while album rock seems to be the most popular genre. This could be due to more independent or small artists not releasing their music on large streaming platforms. Many up-and-coming artists use various, more accessible and cheaper methods for releasing their music, such as using the platform SoundCloud.

5. What are the auditory quality averages by genre?



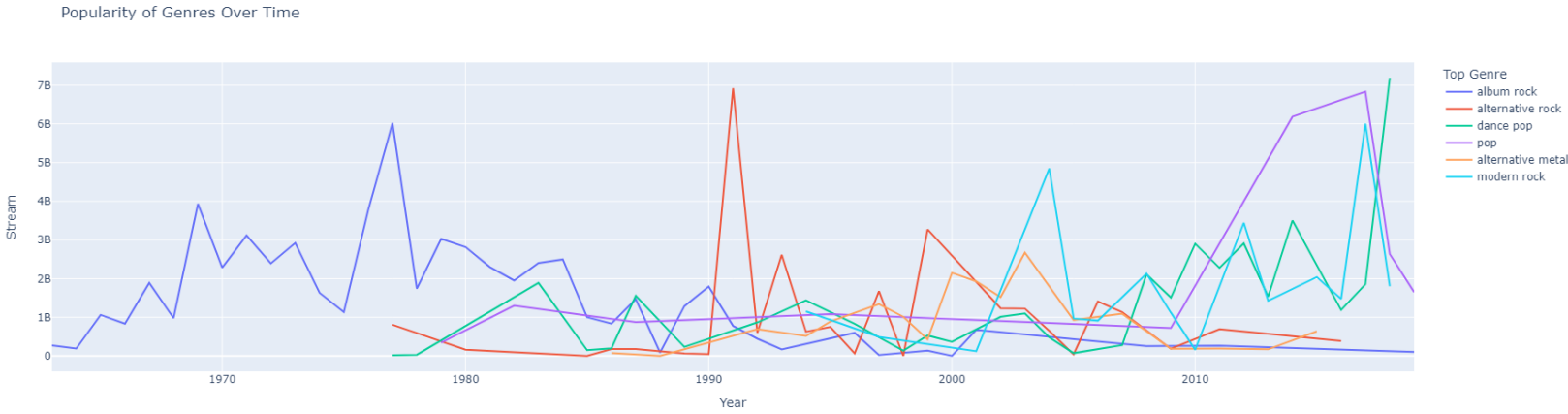
The auditory quality averages vary greatly by genre. Out of all the auditory quality averages, speechiness, loudness, and liveness variables seem to be the most consistent. Acousticness and energy seem to have the most variance in their measurements. Overall, no two genres seem to have the same measurements for each auditory quality. From this, we can see that each genre has a different balance of auditory qualities. This is because genres are differentiated by the style of each song, and the style is influenced by the auditory qualities.

6. How do the auditory qualities of a song impact its popularity?



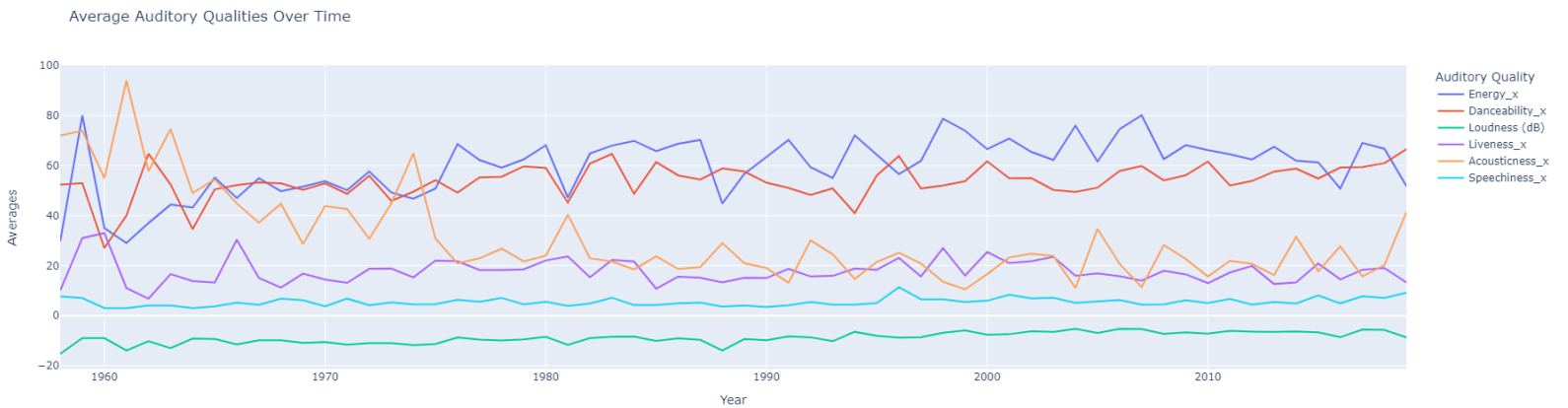
There does not seem to be much of a correlation between the auditory qualities of a song and its popularity. This can be because popularity can depend on a variety of outside factors, such as marketing, lyrics, different streaming platforms, and production quality. Auditory qualities alone may not be enough to determine how well a song will perform.

7. How did the popularity of the top 6 genres of all time rise and fall over the years?



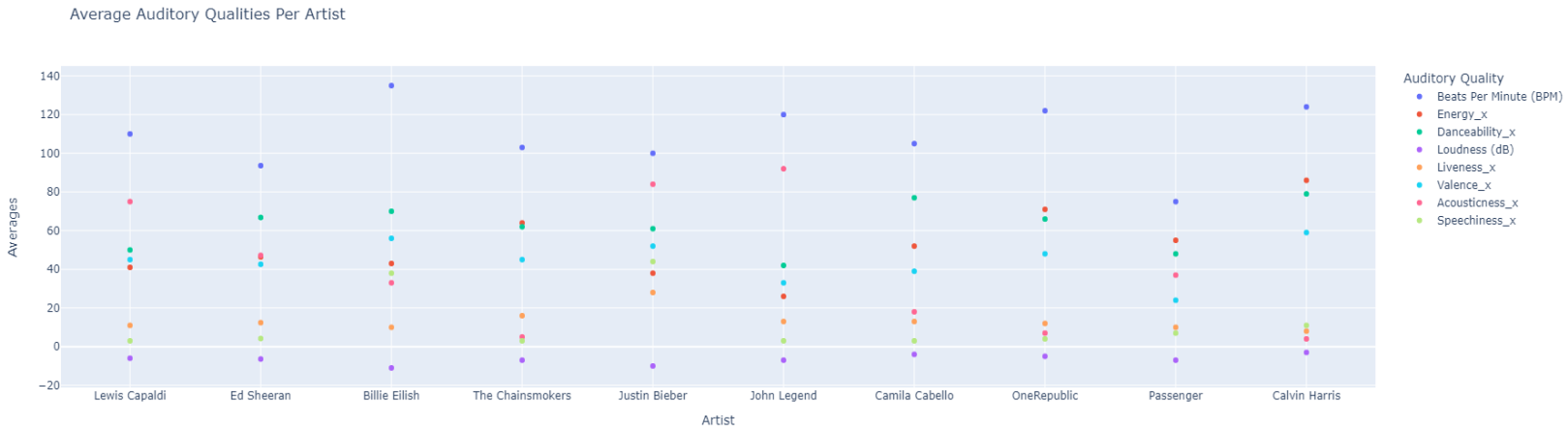
This graph shows that album rock streams had a steady decline over the years, while dance-pop, pop, alternative metal, and modern rock streams all mostly increased over the years (with a few exceptions). Alternative rock seems to quickly peak around 1977 and then decline with a few other lower peaks. The popularity of alternative rock peaks for songs released in 1991. Dance-pop peaks in 2018, pop and modern rock peaks in 2017, and alternative metal peaks in 2003. Before 1977, there wasn't much competition from other popular genres; the popular songs released from those years all seem to be album rock. The decrease in album rock and alternative rock over time could be due to the different subcategories of rock that emerged, such as modern rock. This could lead to people releasing more music from various genres throughout the years, therefore also increasing the streams for other genres.

9. How did the total average auditory qualities change over time?



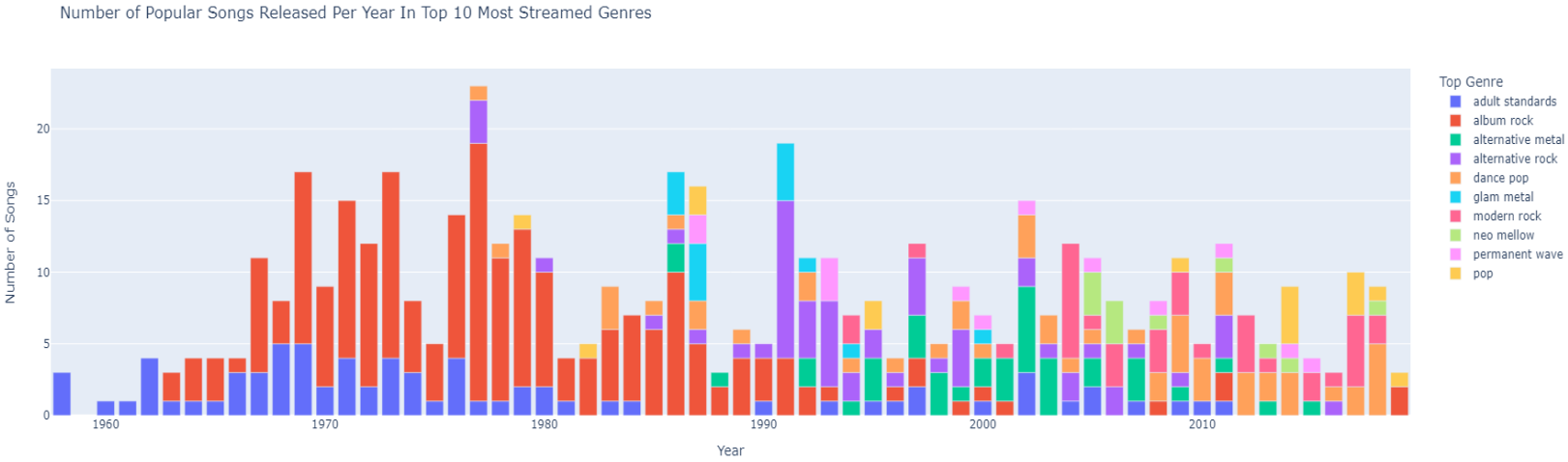
Speechiness, loudness, and liveness seem to remain stable throughout the years. On the other hand, danceability and energy increase over time, while acousticness decreases over time. The decrease in acousticness could be because of the increased popularity of dance-pop and pop, as we've seen above. That would also explain the increase in danceability and energy, which are both defining characteristics of dance-pop and pop music.

10. What are the average auditory qualities of the top 10 popular artists?



For the top 10 popular artists of all time, most of the auditory qualities seem to be similar. One exception would be the acousticness measurement for Lewis Capaldi, John Legend, and Justin Bieber who all have a higher level than others. Another exception is the speechiness measurement, which Justin Bieber and Billie Eilish have the most of. Finally, John Legend has the least energy out of all the artists. Other than those exceptions, the rest seems to be fairly equal. This could be the result of what is popular among listeners. For example, if the majority of listeners prefer fast songs, the artists that get popular may all tend to have higher bpm in their songs. However, the slight differences in the measurements also shows that an artist's popularity can depend on many factors, not just the auditory qualities of their music. The artist's personality, social media presence, marketing and public relations team, and storytelling capabilities are all other factors that could influence their popularity.

11. How many songs from the top 10 genres were released every year?



From 1960 - 1984, the primary genres that popular songs were released were adult standards and album rock. After 1984, there was a gradual increase in the variety of genres, and in 1993, there was an explosion of popular songs being released in different genres. The increase of popular songs in many different genres could be due to the demographic shifting between generations. While an older generation enjoyed listening to and producing album rock and adult standards, a younger generation may be more interested in experimenting and producing music that bends the rules of genres that had been set before.

12. Based on this model, can the number of streams be predicted given speechiness, loudness, liveness, and instrumentalness features?

For our machine learning analysis, we decided to use the KNeighborsRegressor learning model and r-squared accuracy metric. After merging our datasets, we had 12,601 observations to use as input for the learning model. In order to potentially increase the accuracy of our model, we plotted all 8 of the auditory features against the streams to look for a correlation. The graphs for speechiness, liveness, loudness, and instrumentalness had the steepest slope in the scatter distribution, indicating a higher degree of correlation. For this reason, we only included these 4 features in our learning model. We also normalized the auditory features using MinMaxScaler from sklearn, to ensure each feature was weighted equally, between the values of 0 to 2.

For our accuracy metric, we chose to use r-squared due to its straightforward 0-1 scale. Our training accuracy was around 45%, and our testing accuracy was around 15%. This denotes either no correlation, or a weak correlation at most. Hence, we weren't able to predict the streams a song would get based on its auditory features.

Impact and Limitations.

There are some potential limitations of our work.

Both of the datasets used focus on music on Spotify and YouTube. Both of these platforms, as well as most mass media, are heavily centralized in North America. This removes any local or cultural context from other parts of the world, so representing our findings as universal would be inaccurate.

Furthermore, these datasets do not include any information about the demographics of the people who streamed/listened to the music. Many artists tend to make music that reflects their personal experiences, often in relation to their identity. This can be hopefully mitigated to some degree through genre analysis, but without accurate demographics, we will not be able to make a statement on what type of music is popular to what smaller sub-group of people.

Lastly, many smaller or new artists may need more traction than others. A song's popularity isn't solely based on its auditory qualities– artists usually have marketing teams and social influence, giving them a much more extensive outreach than smaller or newer artists. In addition, many streaming platforms like Spotify require artists to publish their music through a label or music distributor, which usually creates costs that may be less affordable for smaller artists. This causes two complications: our analysis of a song's popularity may be affected by the confounding variable of outreach, and some songs with the potential to be popular would not appear on any datasets altogether.

Overall, artists can use our analysis to understand the general trends across Youtube and Spotify and the trends of genres and auditory qualities over the years. However, it's important to note that many outside factors can contribute to the trends seen in our visualizations, and the results cannot be generalized for all global artists and songs.

Challenge Goals

Multiple datasets

- We merged two different datasets to create a more detailed analysis of the top songs. For example, the top 2000s songs dataset does not include factors for judging a song's popularity, so we combined it with a Spotify Youtube dataset that has the number of likes, views, and streams for each song.
- We expanded on this challenge goal by creating another merged dataset for our machine learning model. We merged the Spotify Youtube dataset with a Spotify features dataset, which includes many more songs and features we can analyze, therefore creating a more accurate model.

External library (plotly)

- We used plotly to visualize our findings. Using plotly allowed us to create interactive visualizations, making our analysis a bit more fun and easy to comprehend.

Machine learning

- We also used machine learning to predict a song's popularity given different measurements of auditory qualities. To do this, we used a KNeighborsClassifier.

Plan Evaluation

The initial setup of our data took more time than expected. We ran into some problems using Google Colab, such as being unable to upload the data, getting disconnected from the notebook, and having errors with saving. However, we were able to mitigate these obstacles and got back on track. The machine learning part also took much longer than expected. We struggled with getting a high accuracy rate for our model, so we had to spend a few extra days researching different models and techniques that we could use to increase our accuracy.

Fortunately, our visualizations took just the amount of time we expected, and we didn't have any major problems with them. Creating our visualizations on time gave us some more flexibility to figure out the machine-learning aspect of our project.

Testing

Since most of our code for this project was plots generated through plotly using our merged dataset, we tested our work through smaller datasets with which we could easily hand draw graphs. We created 3 test datasets: a smaller version that only had 3 observations, a slightly larger version with 5 observations, and the final one containing a column for song titles that the other two datasets did not include. We then used the same code we used for our main graphs to create test plots, and then compared them to a hand-plotted version of the graph. To test our

machine learning analysis, we used the same test datasets, subsetting the data the same way we did in our analysis. We then used boolean statements and indexing to check if the column values were what we expected them to be.

Collaboration

Other than our mentor, we consulted various posts on Stack Overflow.