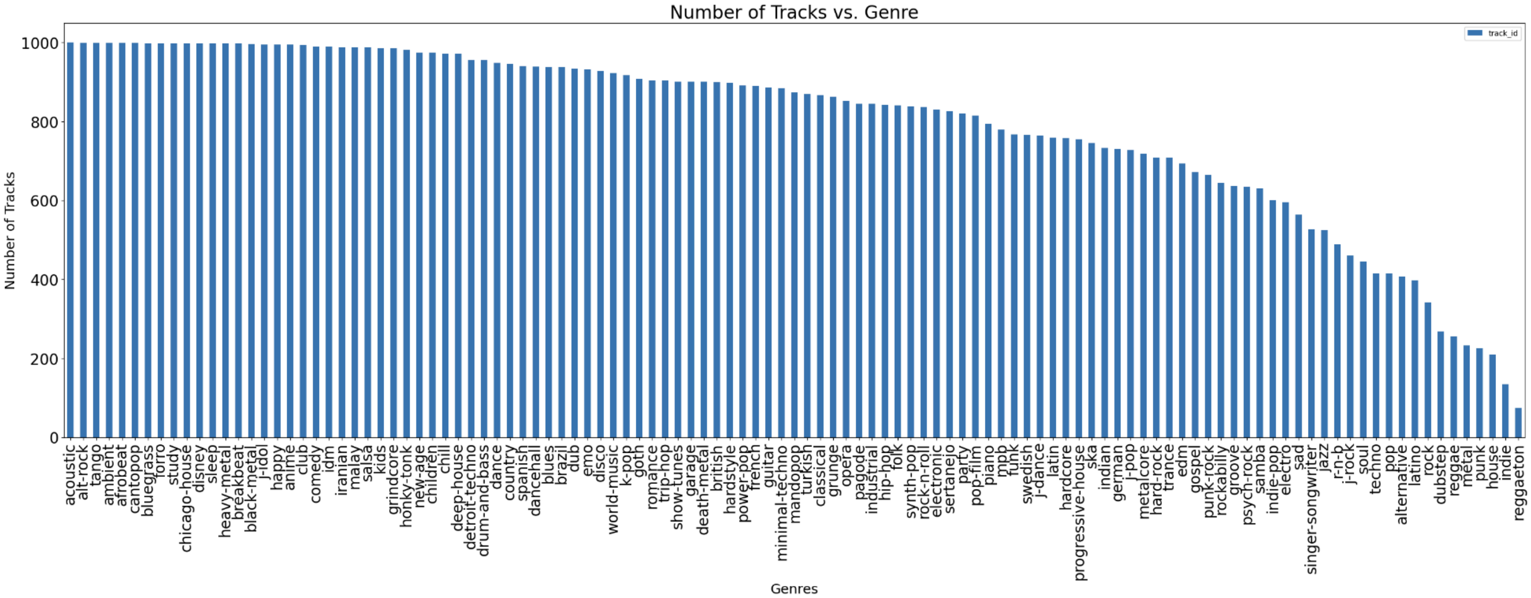
Capstone 2
Abhi Wagh

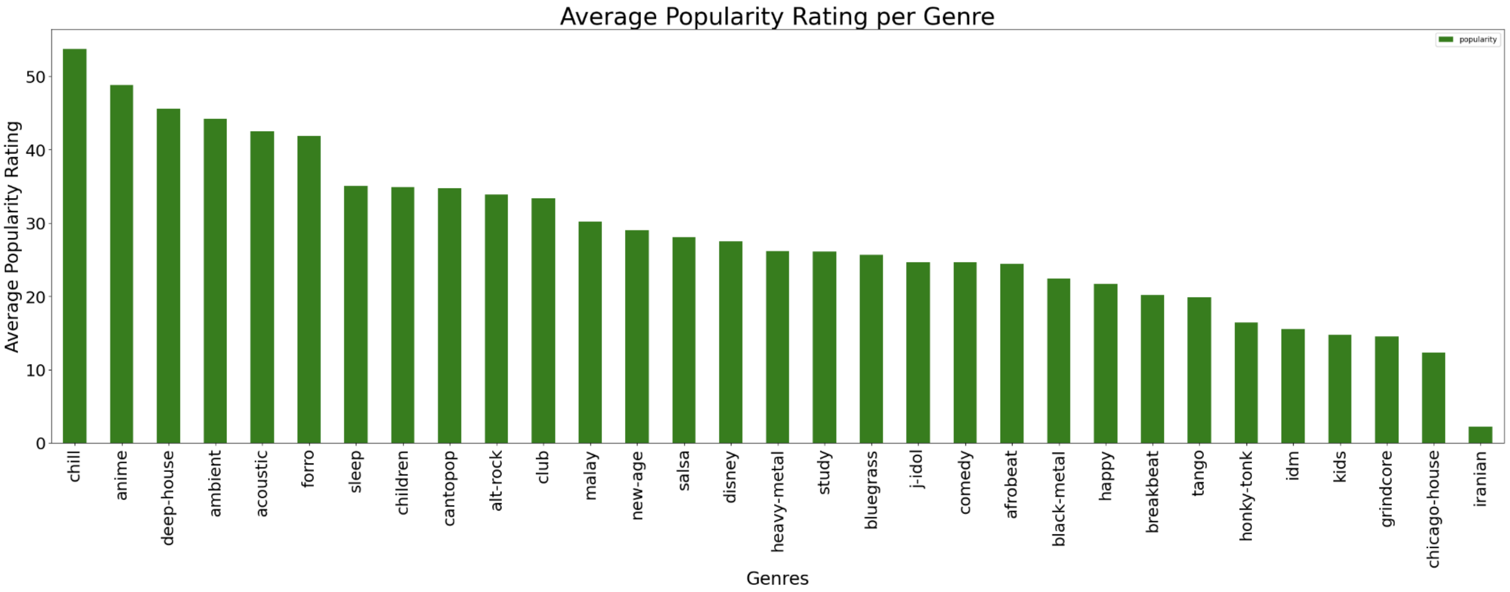
Background

The dataset I chose to examine for this project was from Kaggle. It was collected from Spotify Web API in October 2022 and measures various aspects of tracks, such as tempo, explicit content, and genre. The dataset itself has 114,000 data points. However, upon further analysis, I found that some songs were listed under 2 genres and thus repeated (i.e. 'Sweater Weather' by The Neighborhood was categorized as alt-rock and alternative). After removing duplicates, the dataset had 89,741 results. Before cleaning the dataset, there were 1000 songs listed per genre. This would've made for a consistent dataset where trends could be found with relative accuracy. After cleaning the dataset, however, I found the number of tracks per genre varied a lot, anywhere from 74 to 1000. The lowest number of tracks was in the reggaeton genre. I got around the variation by subsetting the dataset for genres that had more than 970 tracks and analyzing that data. I also examined overall aspects of the dataset, such as average duration of songs.



Analysis of Subsetted Data (Genres with >970 tracks)

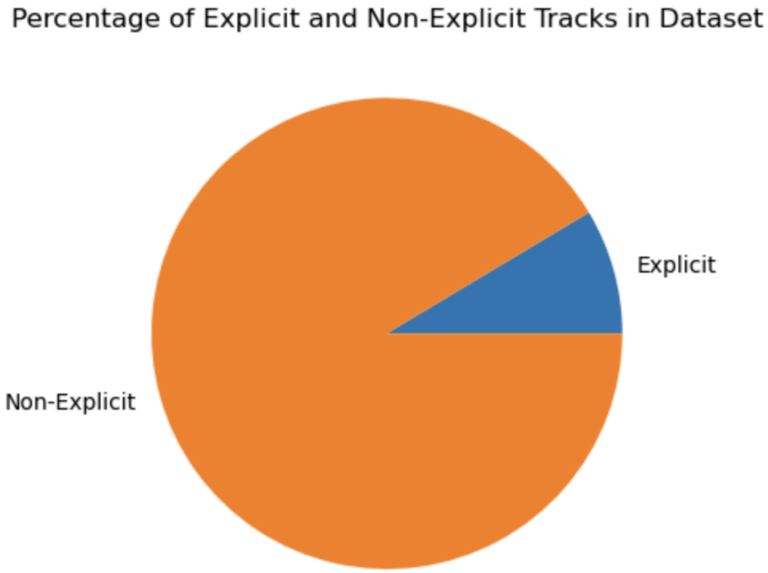
One of the parameters in the Spotify dataset was popularity ratings on a scale from 0 to 100. These ratings were determined by an algorithm that was largely based on how many total streams a song had as well as if it was streamed recently. I analyzed the subsetted data to uncover possible trends. I found that the most popular genre was ‘chill’ with a popularity rating of 53.74. The least popular was ‘Iranian’ with a popular rating below 5. The graph below shows this data.



Overall Analysis of Whole Cleaned Dataset

I also chose to analyze the cleaned Spotify dataset as a whole. Analyses included percentage of explicit content, overall popular tracks, average duration of tracks, and percentage of live music.

For the percentage of explicit versus non-explicit songs of this dataset, I hypothesized that at least half of all the songs would be explicit. The 'explicit' parameter was based on whether there were any explicit lyrics in a song and yielded a boolean value. 91.42% of all tracks were non-explicit, while 8.58% of all tracks were explicit. It makes sense that the majority of tracks were not explicit, as the dataset included a large variety of genres. Genres such as 'piano', 'chill', 'disney', and 'kids' are likely to have zero explicit lyrics. The pie chart below shows explicit and non-explicit track data.

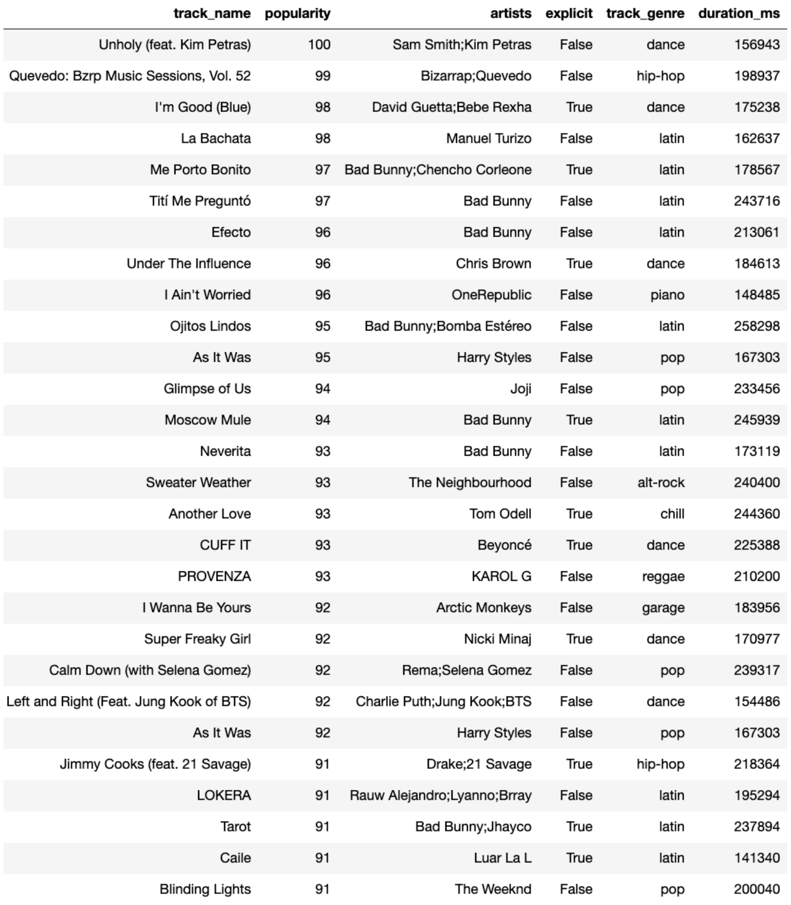


I also looked at the average duration of all the songs in the dataset, approximately 89,741. I found that the average duration of all songs was 3.82 minutes. *This mean value could be possibly skewed, as not all genres had an equal amount of songs. So, some genres could contribute more to the average than others.

Another parameter I examined was 'liveness'. In the dataset, 'liveness' corresponded to the likelihood of a song being a live performance. 'Liveness' was measured by determining if a live audience could be heard on a track. The values ranged from 0 to 1, with 1 being the highest probability of a live performance. Background information on the dataset suggested that "a value above 0.8 provides strong likelihood that the track is live." I used this as a threshold to filter the data for highly likely live performances. I found that 2865 tracks or 3.19% of all tracks are likely live performances. The rest are likely produced tracks without the presence of a live audience.

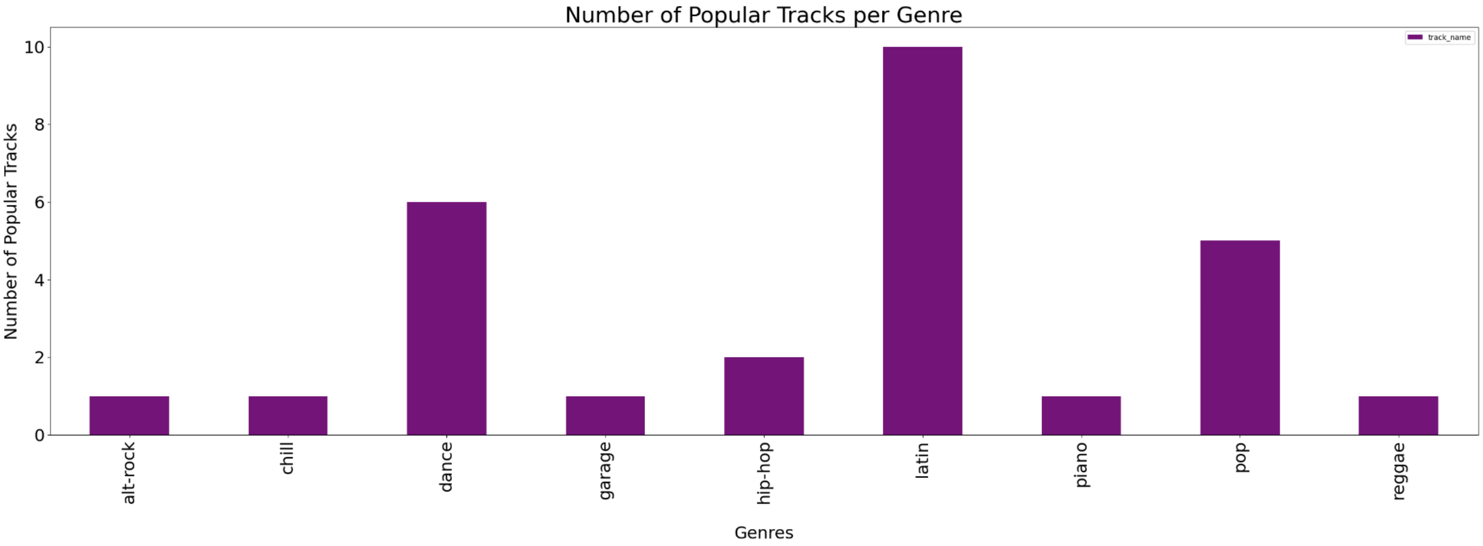
Analysis of Popular Tracks

I was interested in what tracks were most popular in October 2022 on Spotify. Since Spotify has a large variety of genres in their library, I didn't assume all popular tracks would be pop music. However, I hypothesized that most of the popular tracks would be pop music. To conduct this analysis, I filtered the cleaned dataset to output only tracks with a popularity rating above 90. This filter would ensure that I was looking at highly popular tracks. ('Popular' even by numbers is subjective, and I chose to filter for popularity rating >90 per my judgment.) According to the data, 'Unholy' by Sam Smith and Kim Petras was the most popular track in October 2022 on Spotify. The track list of popular songs is below.

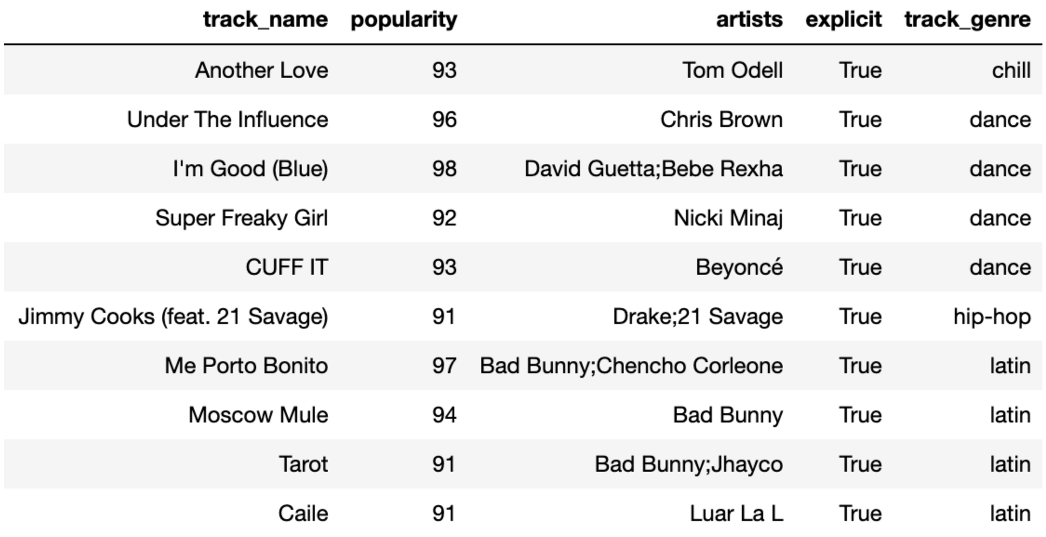


track_name	popularity	artists	explicit	track_genre	duration_ms
Unholy (feat. Kim Petras)	100	Sam Smith;Kim Petras	False	dance	156943
Quevedo: Bzrp Music Sessions, Vol. 52	99	Bizarrap;Quevedo	False	hip-hop	198937
I'm Good (Blue)	98	David Guetta;Bebe Rexha	True	dance	175238
La Bachata	98	Manuel Turizo	False	latin	162637
Me Porto Bonito	97	Bad Bunny;Chencho Corleone	True	latin	178567
Tití Me Preguntó	97	Bad Bunny	False	latin	243716
Efecto	96	Bad Bunny	False	latin	213061
Under The Influence	96	Chris Brown	True	dance	184613
I Ain't Worried	96	OneRepublic	False	piano	148485
Ojitos Lindos	95	Bad Bunny;Bomba Estéreo	False	latin	258298
As It Was	95	Harry Styles	False	pop	167303
Glimpse of Us	94	Joji	False	pop	233456
Moscow Mule	94	Bad Bunny	True	latin	245939
Neverita	93	Bad Bunny	False	latin	173119
Sweater Weather	93	The Neighbourhood	False	alt-rock	240400
Another Love	93	Tom Odell	True	chill	244360
CUFF IT	93	Beyoncé	True	dance	225388
PROVENZA	93	KAROL G	False	reggae	210200
I Wanna Be Yours	92	Arctic Monkeys	False	garage	183956
Super Freaky Girl	92	Nicki Minaj	True	dance	170977
Calm Down (with Selena Gomez)	92	Rema;Selena Gomez	False	pop	239317
Left and Right (Feat. Jung Kook of BTS)	92	Charlie Puth;Jung Kook;BTS	False	dance	154486
As It Was	92	Harry Styles	False	pop	167303
Jimmy Cooks (feat. 21 Savage)	91	Drake;21 Savage	True	hip-hop	218364
LOKERA	91	Rauw Alejandro;Lyanno;Brray	False	latin	195294
Tarot	91	Bad Bunny;Jhayco	True	latin	237894
Caile	91	Luar La L	True	latin	141340
Blinding Lights	91	The Weeknd	False	pop	200040

My hypothesis was incorrect, as most of the popular tracks were not in the pop genre. 35.71% of popular tracks are from the latin genre. This seems to be the most popular genre in the top tracks. Latin music even beat pop music in the top tracks, as pop music made up only 17.86% of the top tracks. *Additionally, it's possible these track rankings can change day-to-day, especially when new albums from well-known artists drop. It can also depend on what songs blow up on TikTok. Currently, we can see that the Billboard Hot 100 hits are sometimes driven by TikTok success. The graph for the genre breakdown of popular tracks is below.

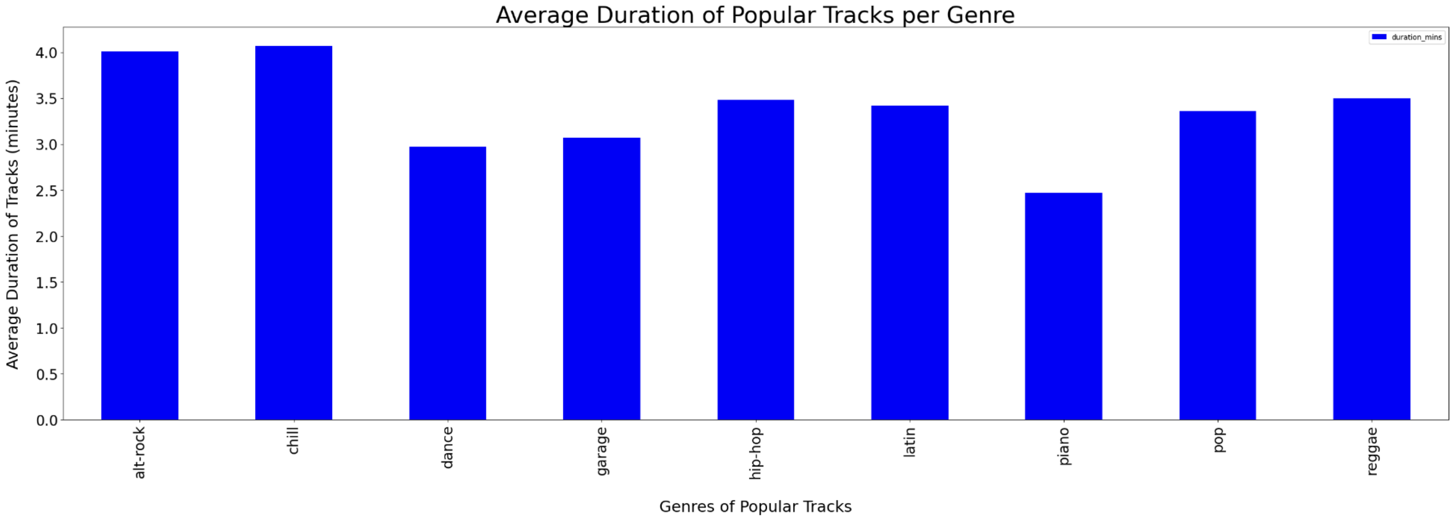


I also analyzed the percentage of explicit songs in the popular tracks list. I hypothesized that a large amount of the top tracks would be explicit. Explicit lyrics may cause more engagement with a track or album itself. Often, tracks have the explicit label on pop music. Further data would be needed to verify this though. 10 out of 28 popular tracks were explicit, while 18 were not explicit. 35.71% of popular tracks were explicit, and 64.29% of popular tracks were not explicit. Compared to the overall dataset analysis, a proportionally higher percentage of explicit tracks are in the top tracks list. The list of explicit tracks is below.



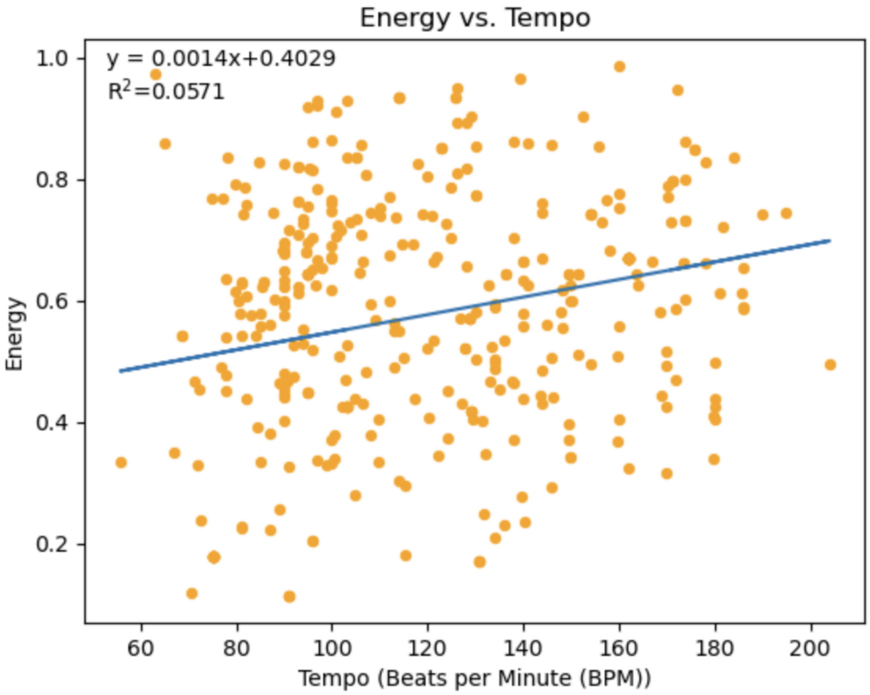
track_name	popularity	artists	explicit	track_genre
Another Love	93	Tom Odell	True	chill
Under The Influence	96	Chris Brown	True	dance
I'm Good (Blue)	98	David Guetta;Bebe Rexha	True	dance
Super Freaky Girl	92	Nicki Minaj	True	dance
CUFF IT	93	Beyoncé	True	dance
Jimmy Cooks (feat. 21 Savage)	91	Drake;21 Savage	True	hip-hop
Me Porto Bonito	97	Bad Bunny;Chencho Corleone	True	latin
Moscow Mule	94	Bad Bunny	True	latin
Tarot	91	Bad Bunny;Jhayco	True	latin
Caile	91	Luar La L	True	latin

The last analysis I did on the popular tracks data was the average duration of songs by genre. The longest popular songs were in the chill genre at 4.07 minutes. The range in average duration for top tracks is from 2.47 minutes to 4.07 minutes. The shortest popular songs on average are from the piano genre. Interestingly, pop songs in the top tracks list average to about 3.36 minutes. The graph for this data is listed below.



Bonus Analysis: How are tempo and energy related for pop music only?

After my main analyses, I wanted to further study relationships between different variables from the dataset. I examined how the ‘tempo’ and ‘energy’ parameters related to each other only for pop music, as the dataset was too large and varied to find any possible correlation. The tempo parameter is measured in beats per minute or BPM, while the energy parameter examines activity of a track on a scale from 0 to 1. The scatterplot showed a potential positive correlation between energy and tempo. However, since the correlation coefficient was calculated to be 0.24, there is likely very little correlation between the two variables.



Conclusion and Future Directions

This dataset was very large with tens of thousands of data points, so it was exciting to analyze trends in. I believe the insights I was able to glean from this dataset are likely accurate, as I made sure to use specific filters and analyze data with categories of close to an equal number of data points. Additionally, there are so many parameters in the data that my current analysis barely scratches the surface of what can be done with this data. There are several future directions I have in mind, such as the following questions: How do danceability and tempo relate to each other? How are energy and loudness related? How many of the tracks are more speech-dominated versus less speech-dominated? Further analyses can also be done per genre as well.