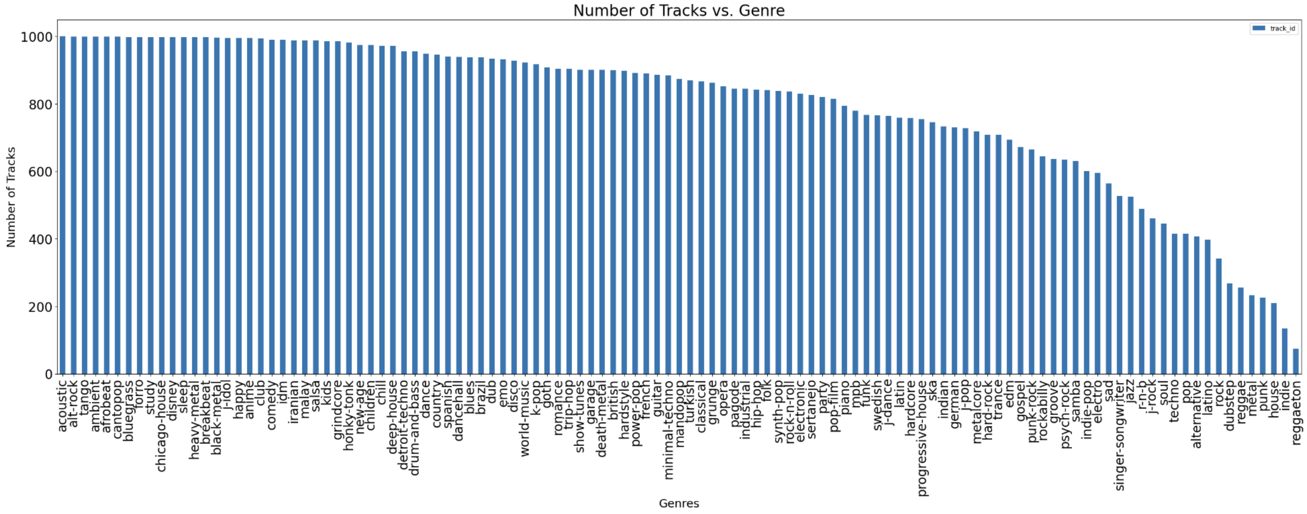
Capstone 2 Report
Abhi Wagh

Dataset Background

The dataset I chose to examine for this project was sourced from Kaggle. It was collected from Spotify Web API in October 2022 and measures approximately sixteen parameters 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. As seen in Figure 1, the lowest number of tracks was in the reggaeton genre. One method of getting around the variation would be subsetting the dataset for genres that had more than 970 tracks and analyzing that data. Instead, I primarily analyzed tracks with a popularity rating above 90 out of a 100.

Figure 1: Number of Tracks per Genre (of cleaned dataset) Bar Chart



Objective

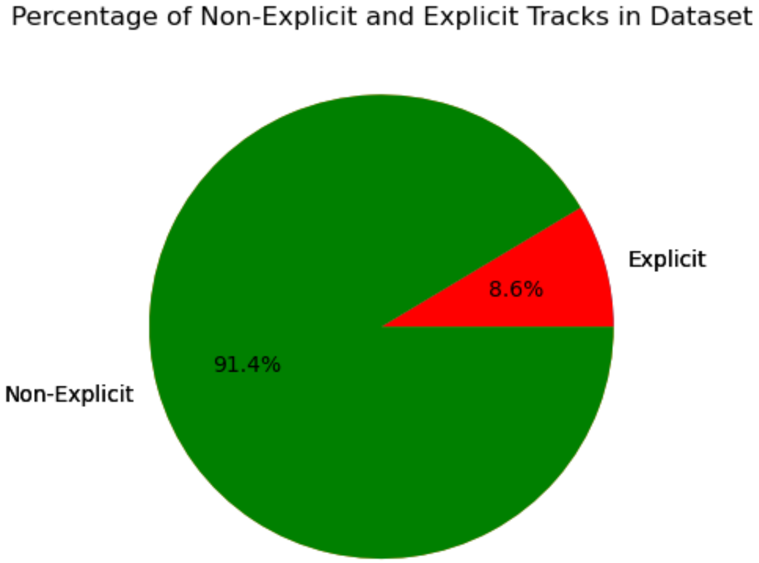
The main goals of exploring this dataset were to analyze the broad parameters of the full cleaned dataset as well as specific parameters of popular tracks. The outputs of these analyses could suggest new ways for Spotify to categorize their track data or inform them about what types of tracks their users enjoy the most. These could translate to actionable insights that help them process their data more efficiently and increase their profit.

Overall Analysis of Whole Cleaned Dataset

First, to develop an understanding of overall trends, I looked at the dataset as a whole. Analyses included percentage of explicit content, average duration of tracks, and percentage of live music.

I hypothesized that at least half of all the songs would be explicit. The ‘explicit’ parameter was measured 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. *Figure 3* below shows explicit and non-explicit track data.

Figure 3: Pie Chart of Explicit vs. Non-Explicit Tracks



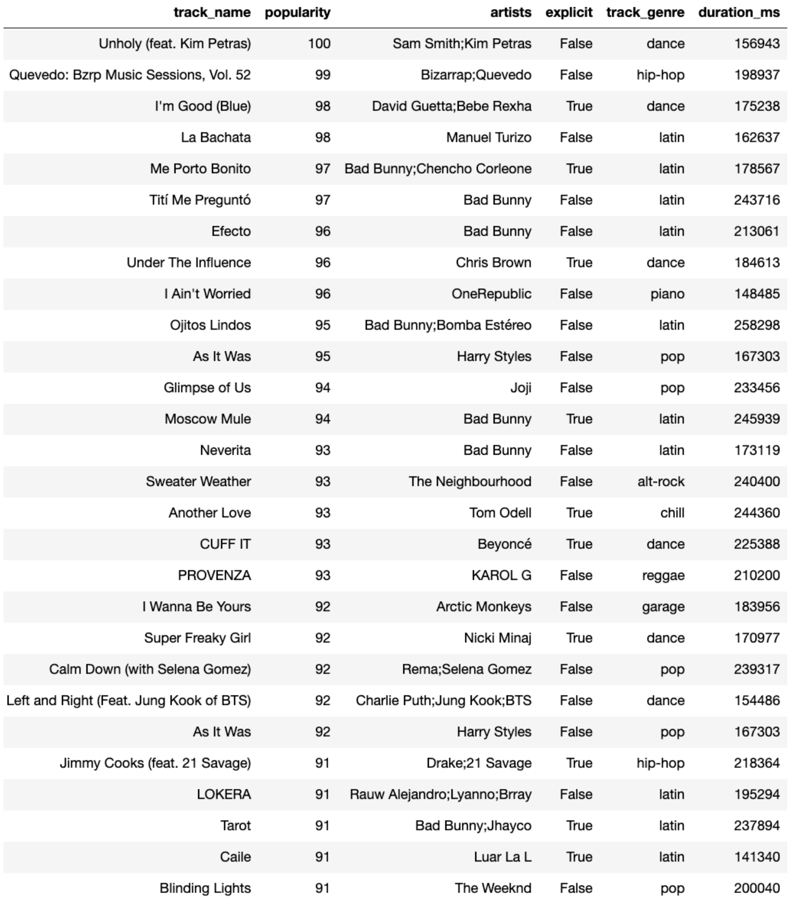
I also looked at the average duration of all the songs in the dataset, approximately 89,741. I found that the median duration of all songs was 3.55 minutes. *This median value may not be an accurate representation of the whole dataset, as not all genres had an equal number of songs. So, some genres could contribute more to the median 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

My next goal was to find the most popular tracks from the dataset. Since Spotify has a large variety of genres in their library, I didn't assume *all* popular tracks would be pop music. However, I did hypothesize that most of the popular tracks would be pop music like past trends on charts like the *Billboard Hot 100* suggest. 'Popularity' was determined by an algorithm on a scale of 0 to 100. The algorithm examined how recent streams for a particular track were as well as its total number of streams. 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.

Figure 4: Popular Tracks Table (Filtered by Popularity Rating > 90)

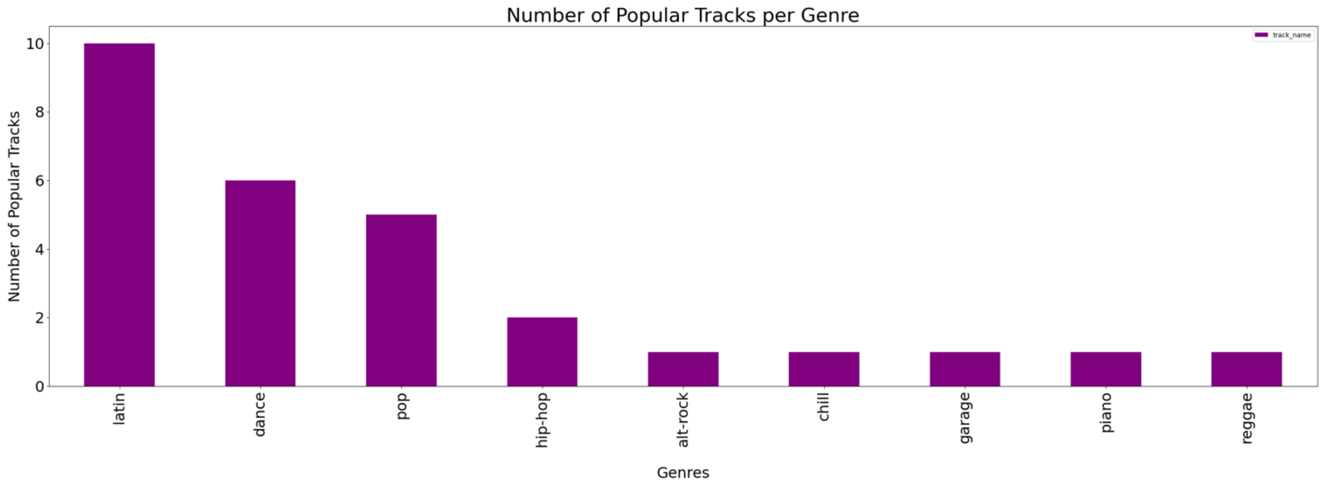


track_name	popularity	artists	explicit	track_genre	duration_ms
Unholy (feat. Kim Petras)	100	Sam Smith;Kim Petras	False	dance	156943
Quevedo: Bzrp Music Sessions, Vol. 52	99	Bizarrap;Quevedo	False	hip-hop	198937
I'm Good (Blue)	98	David Guetta;Bebe Rexha	True	dance	175238
La Bachata	98	Manuel Turizo	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 actually 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. The graph for the genre breakdown of popular tracks is below in Figure 5.

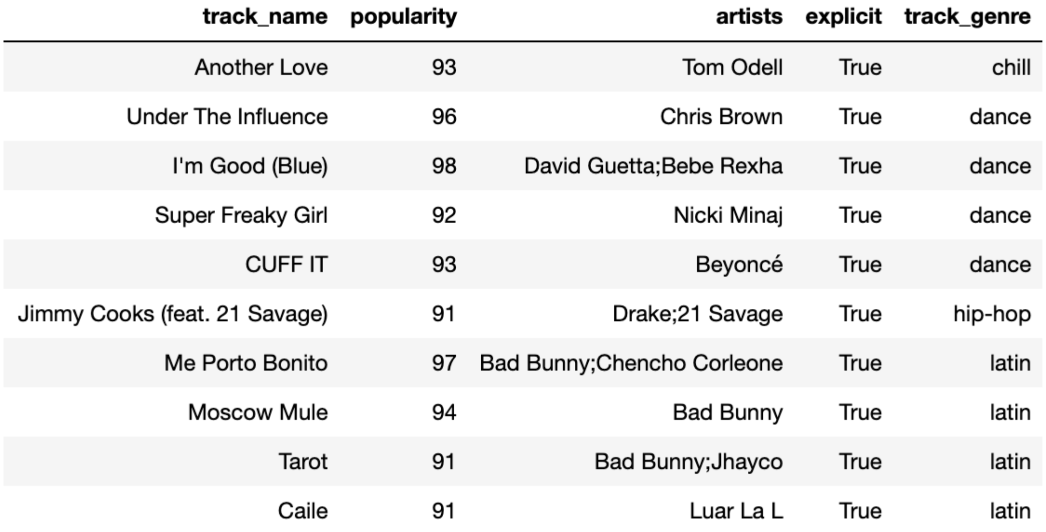
Additionally, it's possible these track rankings can change day-to-day, especially when new albums from well-known artists drop. It can also depend on what songs blow up on TikTok. Currently, we can see that the Billboard Hot 100 hits are sometimes driven by TikTok success.

Figure 5: Number of Popular Tracks per Genre Bar Chart



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 in 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.

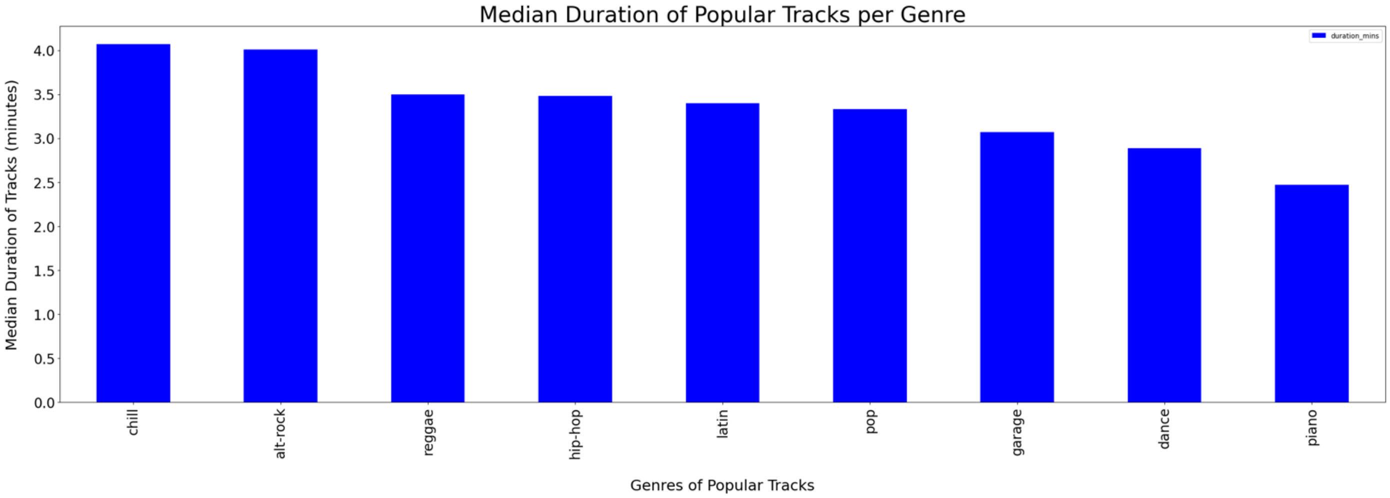
Figure 6: Table of Explicit Tracks from Popular Tracks List



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

Another analysis I did on the popular tracks data was the median duration of songs by genre. I hypothesized that the longest popular songs would be from the piano genre, as classical music tracks tend to be longer than pop tracks. My hypothesis was incorrect. The longest popular songs were in the chill genre at 4.07 minutes, and the shortest were in the piano genre at 2.47 minutes. Interestingly, pop songs in the top tracks list were in the middle with a median value of about 3.33 minutes. The graph for this data is listed below in Figure 7.

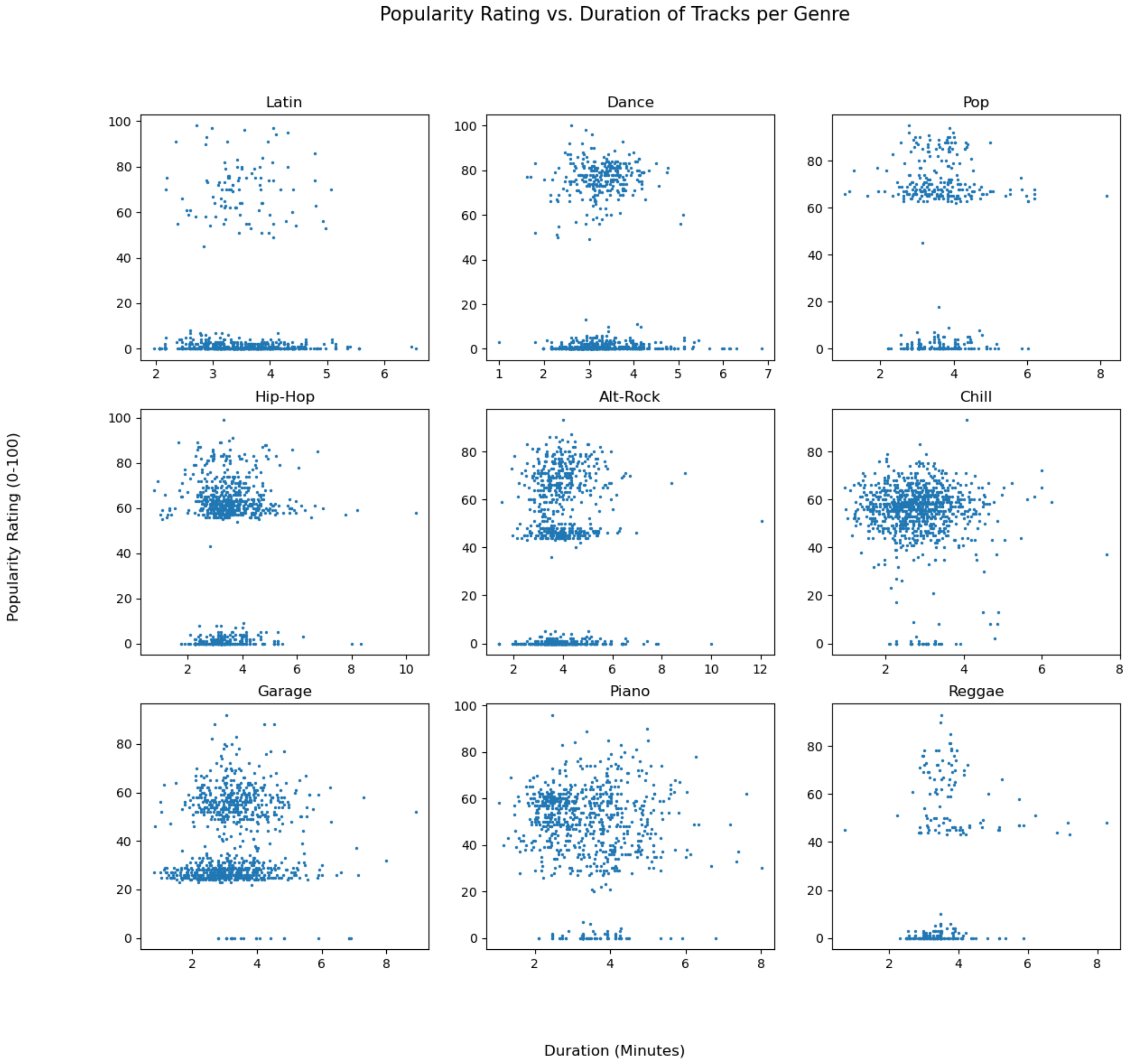
Figure 7: Median Duration of Popular Tracks per Genre Bar Chart



The last major analysis I conducted on this dataset was to create scatterplots of popularity rating versus duration of songs by genre. I examined the genres that were part of the popular songs list, as shown above. A variety of patterns appeared in the subplots. All the genres except ‘Chill’ and ‘Piano’ seemed to have data with a sharp gap between mid-to-high popularity ratings. ‘Pop’ and ‘Hip-Hop’ showed this contrast the greatest, with many ratings close to 0 and many ratings between 60-100. Another interesting pattern was that compared to the amount of low ratings, there was a proportionally smaller number of high popularity ratings for the ‘Latin’ genre. This is surprising, as the majority of popular tracks in October 2022 were from the Latin genre. ‘Garage’ also had a split in popularity ratings, but the lower ratings were higher than a lot of the other genres. The lower ratings were around the 20-30 mark, whereas the higher ratings were between 45-80. Overall, these scatterplots didn’t show much, if any, correlation between ‘popularity’ and ‘duration’ of tracks per genre. They did, however, show where the data for each genre was mostly grouped. For example, tracks in the ‘Chill’ genre fell between under a minute to 5 minutes long.

Several different patterns were evident per genre’s scatterplot for a few reasons. One was because each genre had a different number of tracks listed in the dataset, so some plots like the one for ‘Reggae’ look sparse compared to others. Another reason could be because the ‘popularity’ parameter is algorithm-based and not data directly collected from the Spotify user base. The algorithm could certainly be accurate, as it used existing data of the number and timeline of streams for songs. However, it may be what caused the gap in popularity ratings for a lot of the scatterplots. In reality, it could be the case that Spotify's user base is so diverse that the spread for popularity ratings could look completely different if they were the sources of the data. Further analysis could be done to compare algorithmically-derived data versus direct survey data.

Figure 8: Scatterplots of Popularity Rating vs. Duration of Tracks per Genre



Actionable Insights

This analysis could provide several actionable insights for Spotify:

- 3.19% of all tracks in the given dataset are categorized as 'live.' This could be an untapped market, as there are droves of engagement for live performances from artists on other platforms like YouTube. An example is the NPR 'Tiny Desk' show that features artists of small and large followings singing live. Increasing the number of live performances on Spotify could increase both engagement and/or new users.

- Popular tracks from this dataset were largely between around 2.5-4 minutes. This could be a useful insight to make more playlists with songs of this duration, again for more engagement.

- Mid-September to mid-October is Hispanic Heritage Month. The Latin music genre had a lot of popular tracks in October 2022. (More data would be needed to verify if this pattern was due to Hispanic Heritage Month or other factors like the recent popularity in international music genres, among others.) Spotify does this to some degree, but perhaps they could have a more robust program all year-round that spotlights artists of all sizes that are of different ethnic backgrounds or LGBTQ+.

- In my analysis, I found some genres like 'Iranian' that ranked very low in popularity. ('Iranian' as a genre had a median rating of near zero in October 2022.) This means songs of those genres likely have very few plays. Spotlighting these genres could also bring new users and engagement.

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 use median values instead of means to avoid influence from outliers. Additionally, there are so many parameters in the dataset 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 many of the tracks are more speech-dominated versus less speech-dominated? Further analyses can also be done per artist as well.