Questions you found interesting and what motivated you to answer them

What makes a song on Spotify (and really the greater context is in general) popular? We see things like How many streams or How long on the charts or Chart placement and those things tell you a song is popular, but what are the things that contribute to making the song popular? These were things of interest. The fuel for motivation to find answers was simple curiosity coupled with wanting validation of assumptions. Positive songs, with great beats – that make you want to get up and move, that aren’t too long or too short – and that you can easily move to would probably be popular songs, and songs that are the most popular would have the “just right” mix of these things.

Where and how you found the data you used to answer these questions

We found a dataset online on Kaggle, that consisted of Spotify provided data for 2020/21 regarding the weekly top 200 charted songs. Below is the link to get the dataset from.

<https://www.kaggle.com/sashankpillai/spotify-top-200-charts-20202021/version/2>

We found it doing a simple google search for “free datasets music”

The data exploration and cleanup process (accompanied by your Jupyter Notebook)

* The exploration process started by just reading through the field definitions of what the dataset was comprised of and picking some metrics we thought would be most meaningful to look at.
* Field definitions …
* Then we started to make assumptions through brainstorming about how we could make relations/correlations between some of the fields, and we decided upon our main question of “What makes a song popular?” In retrospect, it would make more sense to rephrase it has “What are the things/traits popular songs share amongst each other?”
* Once we did the initial exploring and decided upon our questions, we started to look at the data and how we would try to answer them using what we’ve learned so far. We concluded using scatter plots were a potentially a good fit for this analysis, so we went down that path at first and then decided to try and take it further by utilizing other aspects of things we’ve learned throughout the course like sublots with multiple charts and different statistical tools/methods like calculating the mean/median/mode etc. for reporting
* The cleanup process began with identifying aspects of the data that were either problematic or that didn’t jive with how we wanted to ensure the integrity of our results.
  + We decided first we didn’t’ want to include any records that were incomplete with missing data, so we removed any records with blank fields. There weren’t many – only 11 records out of 1556. So we are at 1545 records.
  + Then we removed any songs that were duplicated by their unique identifier (Song ID). 29 of those – so we were down to 1516 records to work with.
* Once we did that, we then created a copy of the dataset to be used for all our analysis based on popularity. We updated the data types we needed to work with for plotting and calculations.
* We then created another copy of that we could manipulate further to fit the first things we were looking at charting data for analysis and that was genre data. What were the most popular genres? In the data set, there is only a single record per song with all the meta data attached. The genre data was stored in a single field as a comma delimited list, so if a given song was identified in more than one genre, that’s how the data was stored. To chart this easily, we need to separate this out so we could use plotting.

<walk through notebook>

The analysis process (accompanied by your Jupyter Notebook)

1. First we charted out the top 10 genres based on how many times they appeared in the set by value count
2. Then we created a list to correlate the top genres by popularity/stream count
3. We observed that popularity seems to be directly derived from stream count as the set returns identically when popularity is filtered with or without stream count
4. We discerned that there was no direct correlation on genres based on the comparison of the sample set for the popularity vs the value count for the entire set. A larger sample test would be in order but time constraints didn’t allow for it.
5. We then looked at correlating the variables – first up Danceability (rhythm, beat strength) vs Popularity. Although the r-value doesn’t indicate a conclusive correlation, it was still positive and based on that coupled with the clustering tightening up as the danceability rose coupled with the slight rise in popularity, we still inferred that the trend is the more danceable the song, the more popular it can be. We also generated a histogram to show the distribution of the Danceability (approx. 0.7 being the mode) and a boxplot to better identify the outliers for popularity as well as to reinforce via the quartiles where the values lie on the scatterplot.
6. Then Energy (intensity) vs. Popularity – We again inferred that the more energy a song has, the more popular it can be.
7. Now we have Tempo (speed or pace of the song) vs. Popularity – this one is interesting and merits further scrutiny to find the sweet spot (say the mean potentially 120 bpm) as being the best predictor for what tempo leads to a more popular song. The trend shows (ever so slightly) that as a song gets faster, its potential for popularity actually decreases
8. Valence is interesting. The online documentation defines it as: ”…describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry). “ Valence followed the trend with Energy and Danceability in that the more valence (musical positivity) the more popular a song could be
9. Duration (length of the song in ml) merits investigation and should have its outliers removed. It shows that as a duration increases so would potential popularity, but like tempo its more about the sweet spot and looking at the 68-95-99 rule
10. Loudness vs. Popularity – turn it up baby! Although we can infer that the louder the better, this one is more about the consistency of the levels over the entire track so more analysis to show this would be in order.

Your conclusions including a numerical summary and visualizations of the summary

In conclusion, we feel its safe to say that the more danceable, energetic, and “positive vibe inducing” a song is, the more popular it will probably be. The tempo and duration have to hit the sweet spot, and the loudness needs to more than likely be consistent across the duration (we should plot that) to see a positive trend for popularity. As far as genres go,

The implications of your findings: what do your findings mean?

More analysis required to prove it out. Linear regression can allow for estimation and assumption, but something called an F-test can be done to better identify overall statistical significance.

Post mortem – time – I went through a job interview process but am happy to report I got an offer and am deciding what to do now!

Would have loved to spend a lot more time on drilling in and really determining what the numbers mean, looking into how to get a more accurate picture potentially using this F-test if it made sense???