Andrew Ferruzza   **111616974**
CSE 332 - Project #1

Top Spotify Songs of the 2010s

Introduction

The dataset I will be using for this project is titled "Top Spotify songs from 2010-2019 BY YEAR." You can find this data on Kaggle at this address.

The data contains the most popular songs for each year of the 2010s decade, based on the number of streams they received on the music streaming service, Spotify. Each year contains about 50-100 songs. As an avid musician and fan of music in general, I thought this data would be interesting to dive into, as it can answer a lot of questions as to why these particular songs were the most listened to by users. The sole dataset contains a total of 604 data points (songs), each with a total of 10 attributes I have selected which are explained in the following section.

Attributes

Each data point contains 10 attributes. They are:

- **Title:** Categorical data which lists the title of the song. All of the songs here are unique, so there are a total of 604 categories. This attribute is self-explanatory.
- **Artist:** Categorical data that gives the artist of the song. Unlike the title, some artists have multiple songs in this dataset, so the number of categories is way lower and not unique for each point.
- **Genre:** Categorical data that describes the style, or genre, of the song. The genres in this set are pretty specific, so they are more aligned with subgenres. Genres would be considered things like rock, pop, blues, etc., while this dataset lists electropop, pop, dance-pop, etc. Due to this, there is a wider range of categories in this column.
- **Year:** First numerical attribute that states the year the song was in the Billboard and received the most plays. This is important because the year doesn't necessarily represent the release date in this dataset -- it represents the year the song charted the highest and received the most streams. Years are in the range of the decade (2010-2019, inclusive).
- **BPM:** Numerical data that stands for "beats per minute." In other words, this is essentially the speed of the song (the higher the BPM, the faster the song). Numbers range all around, most songs sit around 100-140 BPM, although there are plenty of slower ones that land lower than 100.
- **Energy:** Numerical data that represents the energy of the song. Another way to describe this is the "vibe" of the song. Points with higher energy values will most likely want to make you get up and move, or are more "sing-along" type songs, while lower values usually represent ballads or more sad, slower songs. Numbers range all around < 100.
- **dB:** Numerical data--abbreviation of "decibels" which measures how loud the song is. The greater the value, the louder the song is. Most values here range from -2 to -10 dB, but there are some outliers at -60 dB, for example.
- **Duration:** Numerical data which gives the total time of the song, in seconds. Most songs are around 3-4 minutes (180-240 seconds).
- **Speech:** Numerical data -- counts the number of words that are spoken in the song. Spoken word is essentially just using a talking voice, which isn't as common in a lot of mainstream music, but some of these songs surprisingly contain pretty high numbers for this attribute. No values really make it past 50 on this section though.
- **Popularity:** Numerical data which attributes a number to describe how popular the song is. All numbers are below 100. If a number is closer to 100, this means that most people will be familiar

with it. For example, "Hey, Soul Sister" by Train has popularity 83, meaning most people know the tune when they hear it.

Hypotheses

-Now I will list my hypotheses for this data, and what I hope to answer in my upcoming analysis.

-Hypothesis #1 - Higher energy will be level with higher BPM. I think that the energy and BPM will increase together, as people usually feel more energetic with faster songs. Most slower songs are usually ballads, or moody songs, which won't give people high energy.

-Hypothesis #2 - Higher popularity will have lower speech values. Essentially this means the most popular songs won't contain a lot of spoken word or talking bits. As stated above, spoken word in mainstream music isn't a very big thing. So I think the most well-known songs will have mostly, if not all, singing. A lot of very popular songs are structured around creating a simple yet catchy and effective melody that people can sing along to and remember for years. I predict the data will show a negative correlation in this regard.

-Hypothesis #3 - The most popular songs will have an average duration of 3.30 minutes (or 210 seconds). From observation, most popular songs are anywhere from 3 to 4 minutes long, so I think the most popular ones will average out somewhere in this range. A simple scatter plot could possibly show this nicely, with most of the points clustered together in this range. Of course, there would be some outliers. Some songs in the upper 2 minute range have peaked very high on the Billboard and there's even been instances of 5 minute or longer songs reaching high numbers.

-Hypothesis #4 - The most popular genre here will be "dance-pop." The word "pop music" is just a shortened version of "popular music," so it's easy to say that this will be the most popular genre. But I predict, more specifically, that dance-pop will take home the crown. There were many artists, especially earlier last decade, that tried to revive the dance-pop sound that originated in the 80s. This includes artists like Lady Gaga and Katy Perry, who both had numerous Billboard toppers and were all over the mainstream radio during those years. The data visualization for this will ideally be able to show popular genres throughout the years of this decade as well.

-Hypothesis #5 - Energy and loudness (dB) will positively correlate with each other. Going back to a similar idea to Hypothesis #2, I think louder music also has a great effect on how energetic songs make people feel. The slower songs also tend to be lower in dB, which explains this as well. Of course, this is not always the case. There are some electronic dance music (EDM) songs that are played at slower BPMs but have a great amount of energetic elements to them.

-I am positive that the data will show me trends or observations that I didn't even think of or notice just from looking at the dataset.

Conclusion and Closing Thoughts…

Overall, I think this is a neat dataset to hopefully explain *why* these songs in particular became the most streamed songs on Spotify for the decade. Do people just enjoy music that boosts their energy? Do people just like faster and louder songs? What is so good about dance-pop music? These are all questions that I believe can be answered by effective visualizations which I am looking forward to creating this semester.