MDA 720

Capstone 1

Spotify Top 50 Analysis

Background

Music is listened to everywhere by everyone every day. Millions of people have a Spotify subscription, including me, and a lot of people listen to songs that are in the top 50 list. The list includes the most popular and trending songs at the moment, and some songs can stay in the list for months if it is streamed enough. I have a huge interest in music, and I would like to find out what kind of songs are trending and why they are in the top 50. By figuring this out, musicians can make music that will be trending and therefore be more popular. They can then create more music that others and myself will more likely listen to. My business would help independent artists and bands to understand what makes a song popular and increase their exposure on Spotify. The data set I will be using includes 50 songs with 13 variables. The variables I will be using are: Track name, Artist name, Genre, Beats per minute, Energy, Danceability, Loudness(dB), Liveness, Valence, Length, Acousticness, Speechiness and Popularity.

Objective/Goals of Project

- Filter and understand the data set
- Find the most popular genres
- Find the most popular songs
- Find out if there is any correlation between the songs that are most popular

Column Explanation

Track name	Name of the specific song
Artist name	Name of the artist singing the song
Genre	Which genre the song is under
Beats per minute	The tempo of the song
Energy	The energy of a song - the higher the value, the more energtic song
Danceability	The higher the value, the easier it is to dance to this song.
Loudness(dB)	The higher the value, the louder the song.
Liveness	The higher the value, the more likely the song is a live recording.
Valence	The higher the value, the more positive mood for the song.
Length	The duration of the song.
Acousticnes	The higher the value the more acoustic the song is.
Speechiness	The higher the value the more spoken words

	the song contains.
Popularity	The higher the value the more popular the song is.

Data Extraction/ Collection/ Scraping

I used data from Kaggle that had extracted the top 50 songs from spotify in 2019 with the 13 track properties from http://organizeyourmusic.playlistmachinery.com/. I imported the relevant libraries and imported the CSV file. I had to encode the file before loading it into a data frame.

```
import csv

with open('top50.csv', encoding='latin1') as file:
    reader = csv.reader(file)
    for row in reader:
        print(row)

df = pd.read_csv('top50.csv', encoding='latin1')
df.head()
```

We have 50 rows and 14 columns, and before I started I needed to find out if there's any NA values. To avoid this, I decided that in case I have too many NAs in a column I would consider the elimination of the column.

df.isna().sum()

Once I ran the code I saw that fortunately the dataset doesn't have any NAs, so I could go ahead with the exploration of the data into more details. I also thought it could be useful for me to have a description and a summary of all the columns.

I checked the info of the data set to see what type the values I'm working with are:

df.info()

In the visualization below we can see the count, mean, standard deviation, minimum and maximum, 25%, 50% and 75% percentile.

df.describe()

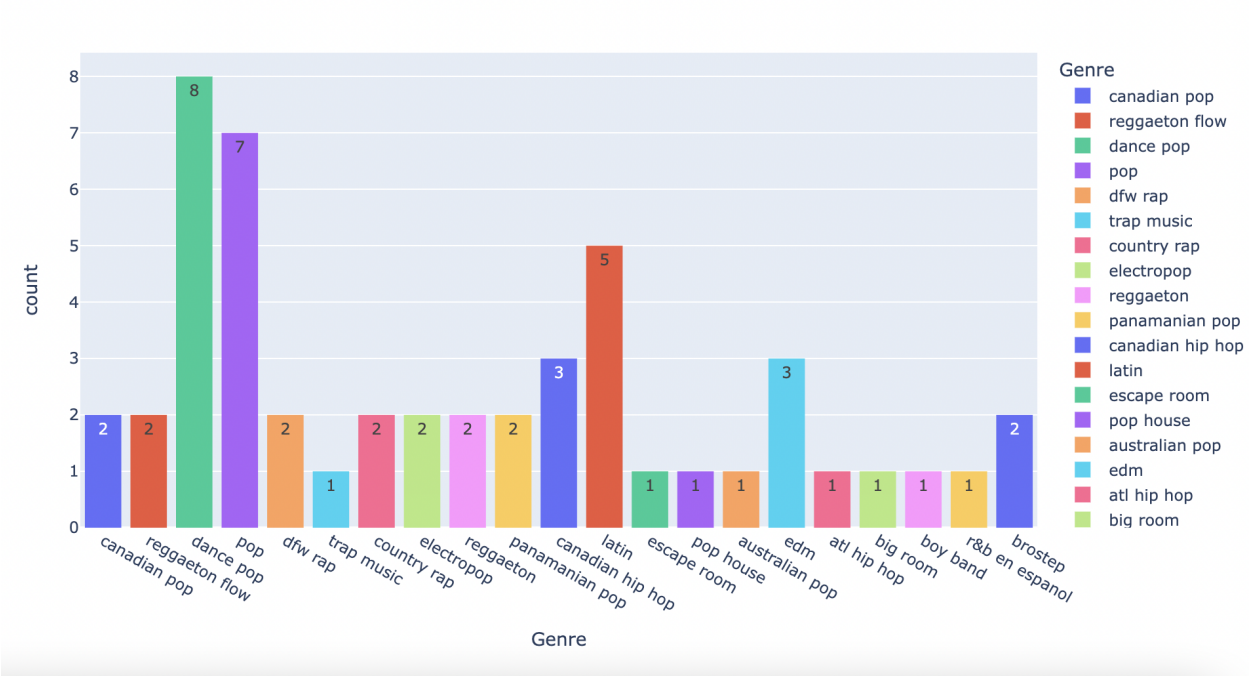
	Unnamed: 0	Beats.Per.Minute	Energy	Danceability	Loudness..dB..	Liveness	Valence.	Length.	Acousticness..	Speechiness.	Popularity
count	50.00000	50.000000	50.000000	50.00000	50.000000	50.000000	50.000000	50.000000	50.000000	50.000000	50.000000
mean	25.50000	120.060000	64.060000	71.38000	-5.660000	14.660000	54.600000	200.960000	22.160000	12.480000	87.500000
std	14.57738	30.898392	14.231913	11.92988	2.056448	11.118306	22.336024	39.143879	18.995553	11.161596	4.491489
min	1.00000	85.000000	32.000000	29.00000	-11.000000	5.000000	10.000000	115.000000	1.000000	3.000000	70.000000
25%	13.25000	96.000000	55.250000	67.00000	-6.750000	8.000000	38.250000	176.750000	8.250000	5.000000	86.000000
50%	25.50000	104.500000	66.500000	73.50000	-6.000000	11.000000	55.500000	198.000000	15.000000	7.000000	88.000000
75%	37.75000	137.500000	74.750000	79.75000	-4.000000	15.750000	69.500000	217.500000	33.750000	15.000000	90.750000
max	50.00000	190.000000	88.000000	90.00000	-2.000000	58.000000	95.000000	309.000000	75.000000	46.000000	95.000000

Data Exploration/Data Visualization

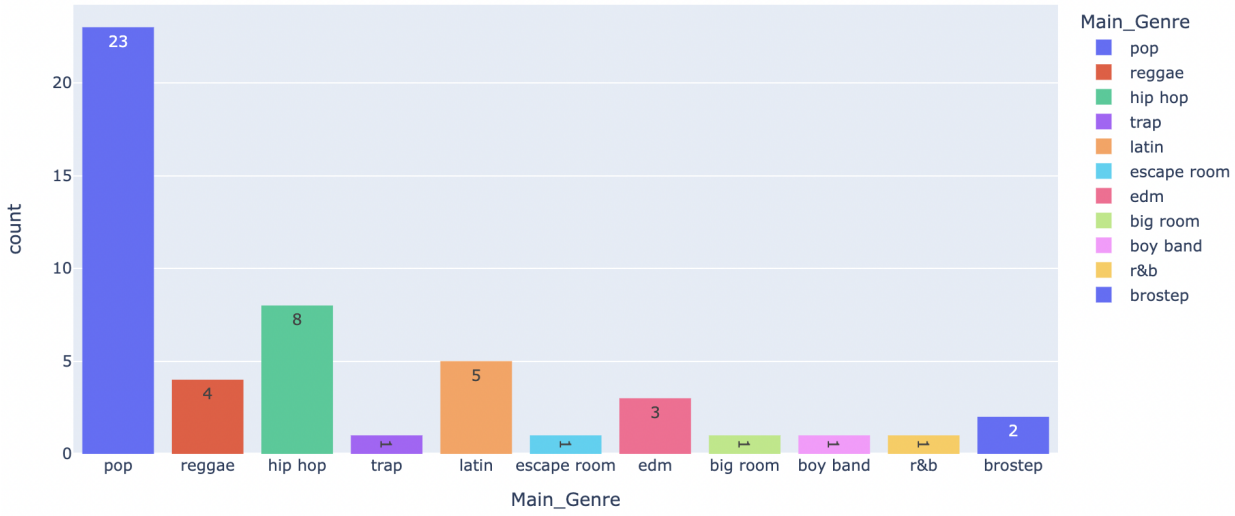
In this section I am going to create different data visualizations and in that way I can better understand the data.

Genres Analysis with Histogram

To get a deeper understanding, I wanted to figure out what the most popular genre was, looking at what the most played genre was. I created a histogram of all the different genres:



To narrow it down even more I divided all genres into main genres: hip hop, jazz, reggae, techno, trap, regga, rap, r&b, rock, pop and blues.



We can clearly see that the most popular genre is pop, with hip hop in second place and latin in third place.

Most popular songs by artist

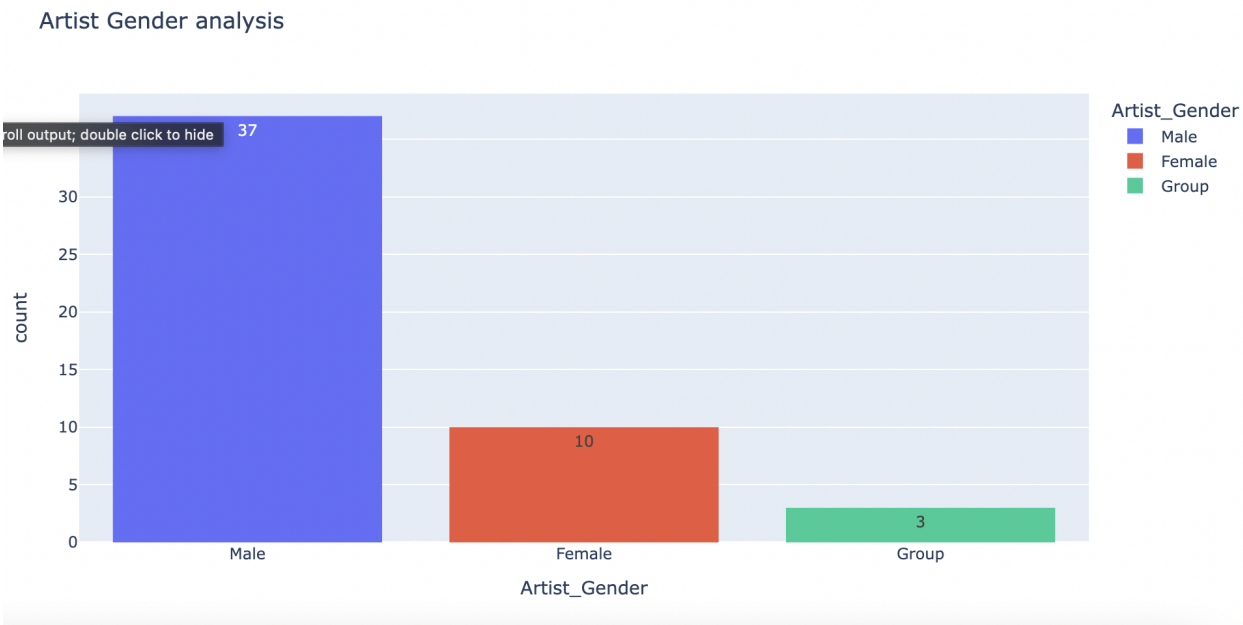
I wanted to find the the top 10 artist with the most popular songs:

	Unnamed: 0	Track.Name	Artist.Name	Genre	Beats.Per.Minute	Energy	Danceability	Loudness..dB..	Liveness	Valence.	Length.	Acousticness..	Speechi
9	10	bad guy	Billie Eilish	electropop	135	43	70	-11	10	56	194	33	
4	5	Goodbyes (Feat. Young Thug)	Post Malone	dfw rap	150	65	58	-4	11	18	175	45	
10	11	Callaita	Bad Bunny	reggaeton	176	62	61	-5	24	24	251	60	
1	2	China	Anuel AA	reggaeton flow	105	81	79	-4	8	61	302	8	
6	7	Ransom	Lil Tecca	trap music	180	64	75	-6	7	23	131	2	
14	15	Money In The Grave (Drake ft. Rick Ross)	Drake	canadian hip hop	101	50	83	-4	12	10	205	10	
17	18	Sunflower - Spider-Man: Into the Spider-Verse	Post Malone	dfw rap	90	48	76	-6	7	91	158	56	
19	20	Truth Hurts	Lizzo	escape room	158	62	72	-3	12	41	173	11	
20	21	Piece Of Your Heart	MEDUZA	pop house	124	74	68	-7	7	63	153	4	
21	22	Panini	Lil Nas X	country rap	154	59	70	-6	12	48	115	34	

Billie Eilish is the most popular artist, and we can see that she produces songs in the pop genre.

Gender analysis with histogram

Next, I wanted to do a gender analysis to see how many songs in the top 50 are produced by women or men.



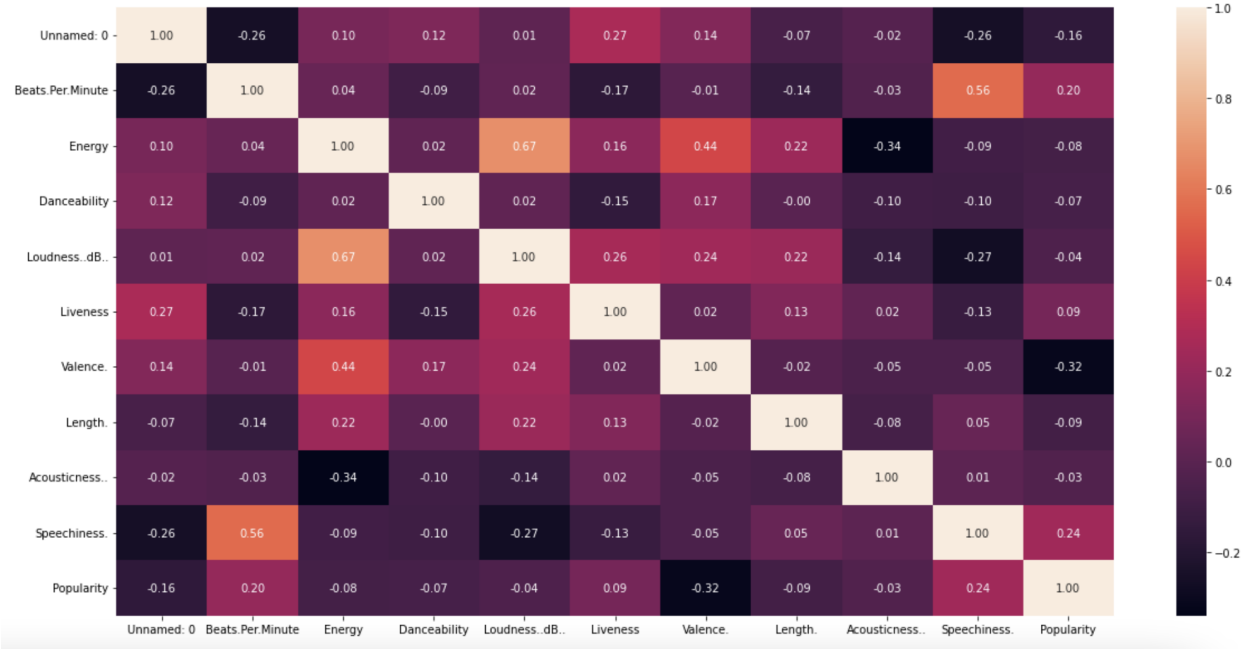
Most of the songs are made by men, 10 of the songs are made by females and three from a group/band.

Data Analysis/ Data Mining/Text Mining

Correlation with Heatmap

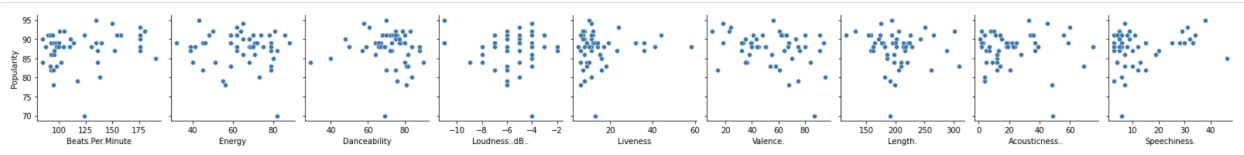
I wanted to find the correlation between the variables. It is a measure of monotonic correlation between two variables, and is therefore better in catching nonlinear monotonic correlations than Pearson's r. Its value lies between -0.4 and +1, -0.4 indicating total negative monotonic

correlation, 0 indicating no monotonic correlation and 1 indicating total positive monotonic correlation. From the graph we can see that in the zones where the color is darker it means we have a high correlation between the two values, for example Acousticnes and energy, or valence and popularity.



Relationship between popularity and other numerical features using scatter plots

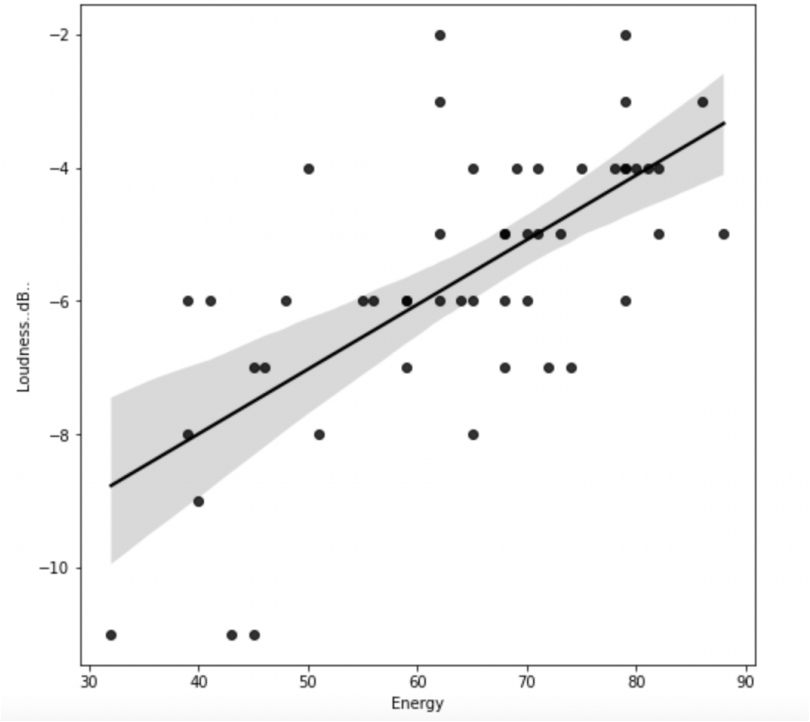
I wanted to see how the different features played a role in the song being popular.



In the different scatter plots you can see the numerical values measured up against popularity. We can for example see that if the Danceability is higher the song is more popular. If we look at the liveliness, most of the songs are in the lower part of the scale, but they are still being listened to.

Relationship between Loudness and Energy

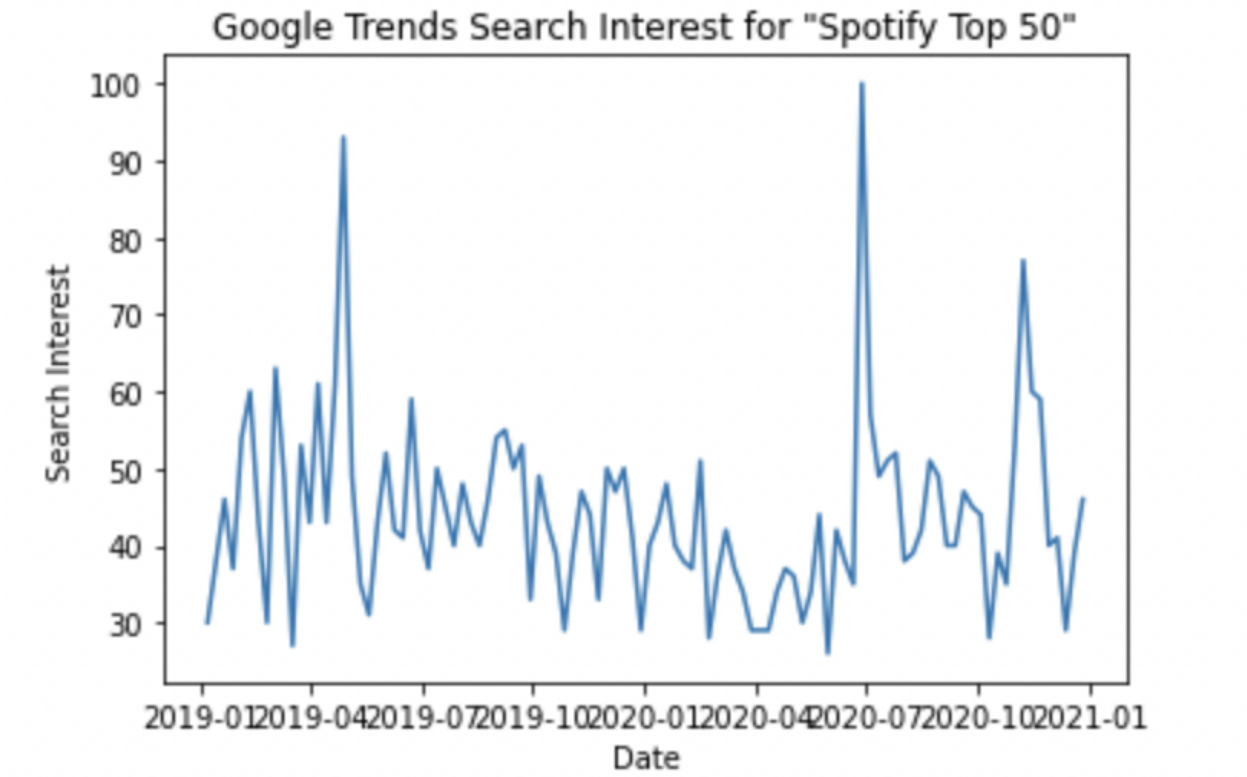
I wanted to find the relationship between loudness in desibel and energy of the song. We can see on the scatter plot that there is a linear relationship between the variables as when the sing is louder in desibel the energy is higher in a lot of the songs.



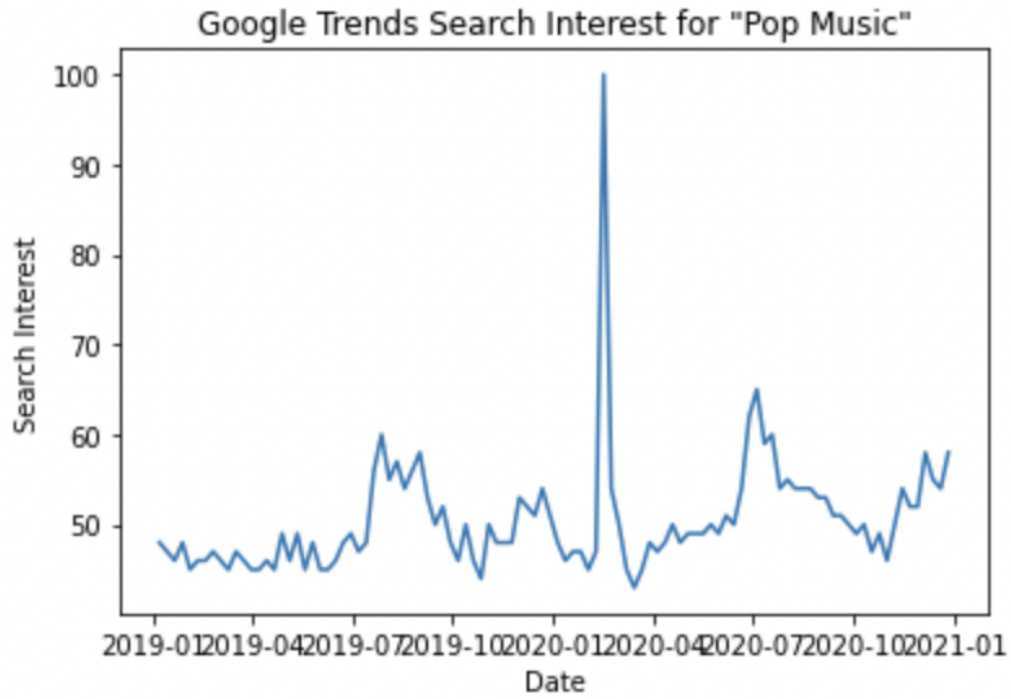
Google Trends Analysis

I wanted to do a google trends analysis to find some of the most searched words in google relating to spotify. I wanted to find the trend for the words “Spotify Top 50”, “pop” and “Billie

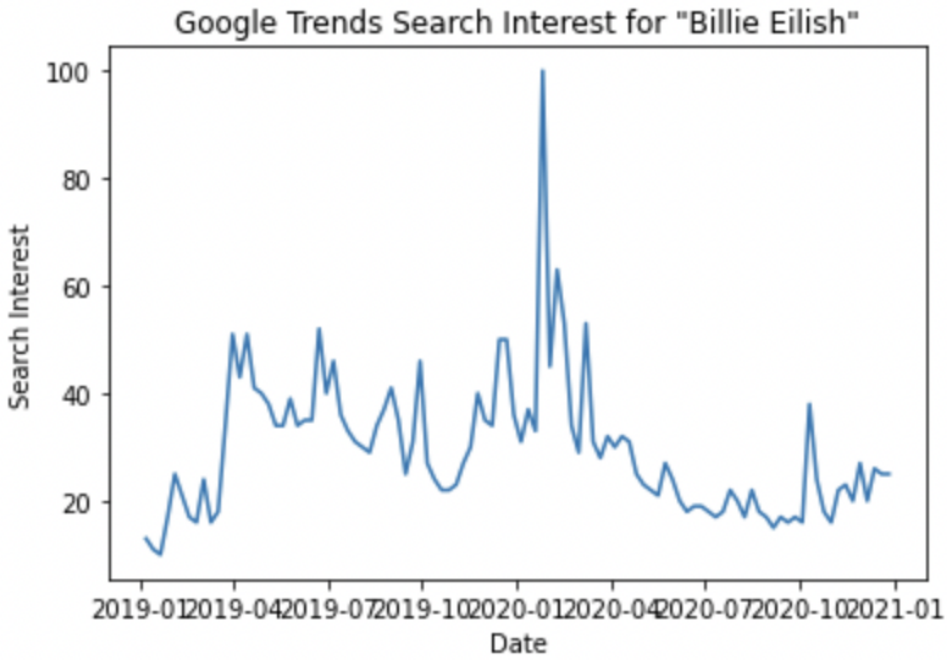
Eilish". This is because we have found out that pop is the most popular music genre and Billie Eilish was the most popular artist.



For the search words "Spotify Top 50" we can see that there are three significant spikes in where it's searched for the most: in April 2019, July 2020 and at the end of 2020.



The search for “pop” music had a significant increase in the beginning of 2020.



People searching for “Billie Eilish”had a peak in the beginning of 2020 but was also a popular search throughout 2019.

Conclusions/Recommendations

After analyzing and visualizing the main aspects of the data there is a lot of interesting information I have found about the most popular songs. For example, one of the most notable things is the genre. We can clearly see that the most frequent genre in the Top 50 is pop. We can also see that there are mainly male artists in the top 50. There are some interesting correlations as well. We can see that there is a high correlation between acousticness and energy. Which means that the most popular songs have high energy while also scoring high in acousticness. Same goes for valence and popularity. The most popular songs have a happier mood. When comparing popularity to all the different features it is high with songs that are more easy to dance to. The songs that are live recordings are also more popular. There is a linear relationship between loudness and energy, which concludes that the higher the song is in volume, the more energetic it is. Spotify Top 50 is a popular google search as well, and if you are one of the most popular artists on that list you will get a lot of google searches. To sum up, the most popular songs are in the pop genre, and songs that have a happier mood tend to be more popular. Most of the songs are created by male artists, however, the most popular artist is a woman: Billie Eilish.

Works Cited

Kaggle - Spotify

Organize your music: (http://organizeyourmusic.playlistmachinery.com/)