DATA MINING OF MUSIC INDUSTRY

I. Introduction

The purpose of this project is to explore the available data of music industry and to provide insightful information by analyzing and interpreting trends.

II. Steps of data mining

1. Data Collection
2. Methodologies
3. Data Wrangling
4. Data Exploration & Communication of findings

III. Details

1. Data Collection

The information the management is looking for is:

- Number of streaming
- Top songs
- Top artists
- Relationship between variables includes genre, Beat per minute, valence, acousticness, liveness, speechiness or popularity

Data was collected in this project includes

- Top 200 global and US songs on Spotify by daily counting on 08/04/2020. The purpose of this dataset is to explore current trends regarding streaming, artist and songs. Source: https://spotifycharts.com/regional/global/daily/2020-08-04
- Top 50 global songs on Spotify in 2019. The purpose of this dataset is to find the relationship between features of songs. Source: https://www.kaggle.com/leonardopena/top50spotify2019

2. Methodologies

Tableau, Python, Excel, SQL are applied during the process of data mining of these datasets

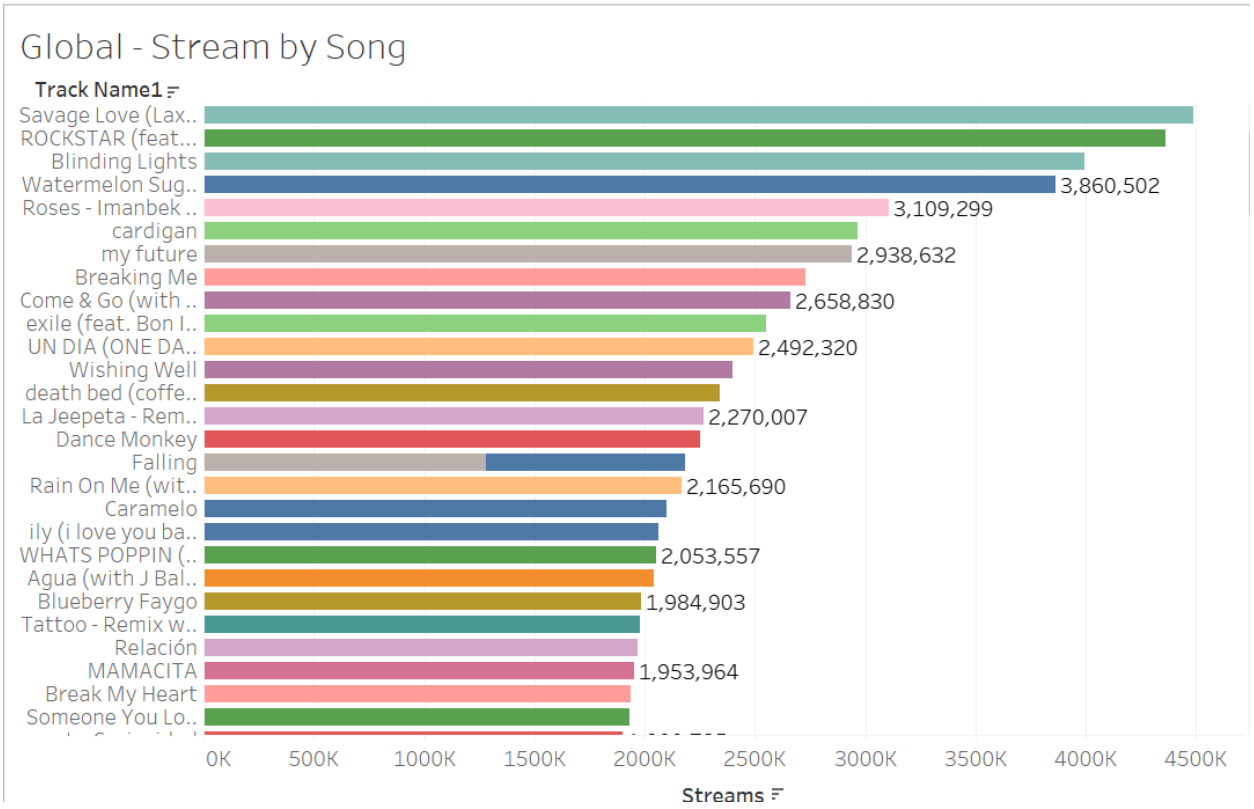
3. Data Wrangling

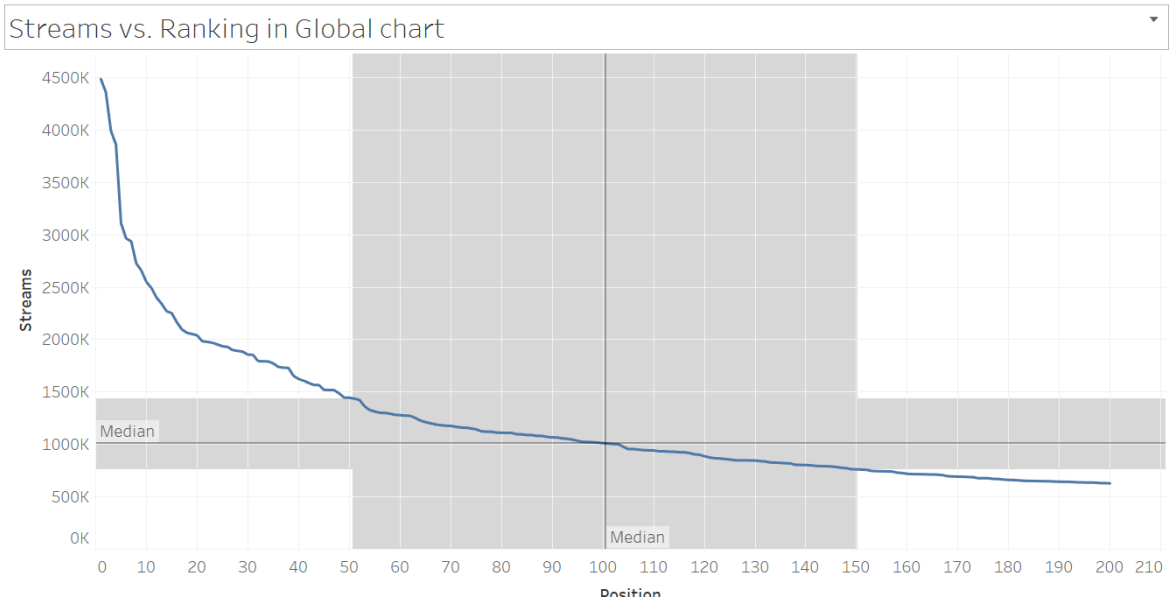
General, the data is pretty clean and well-organized. There is not much to do during the process of data wrangling.

4. Data Exploration

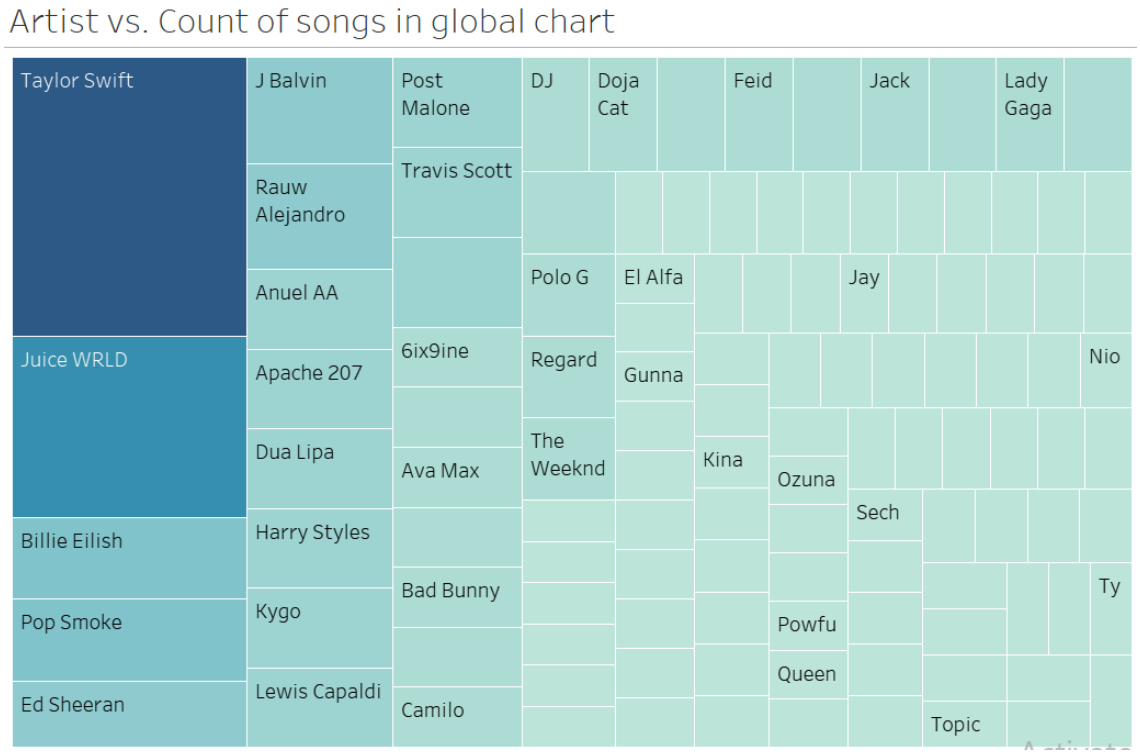
4.1 Global chart

Top 200 songs by Global

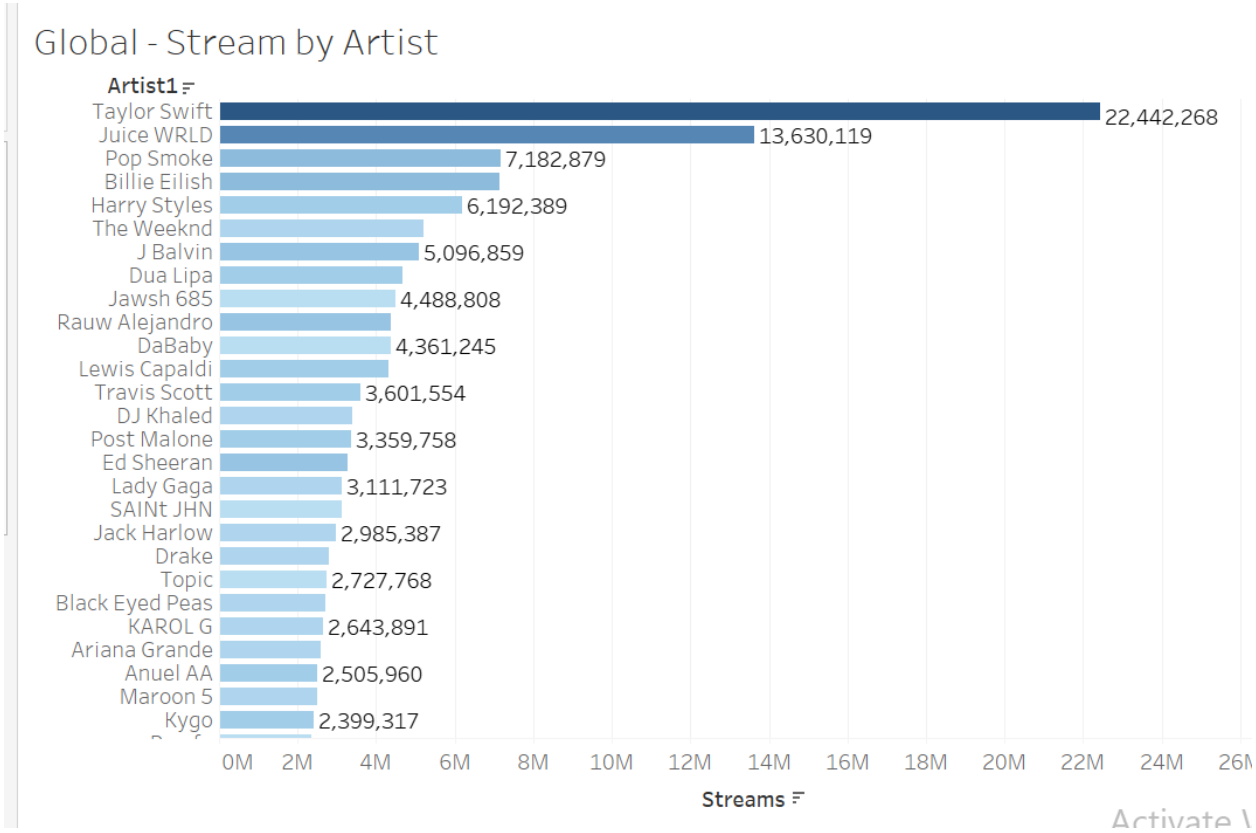




To be in Top 100 of the global chart, any songs should have more than 1 million streaming times.

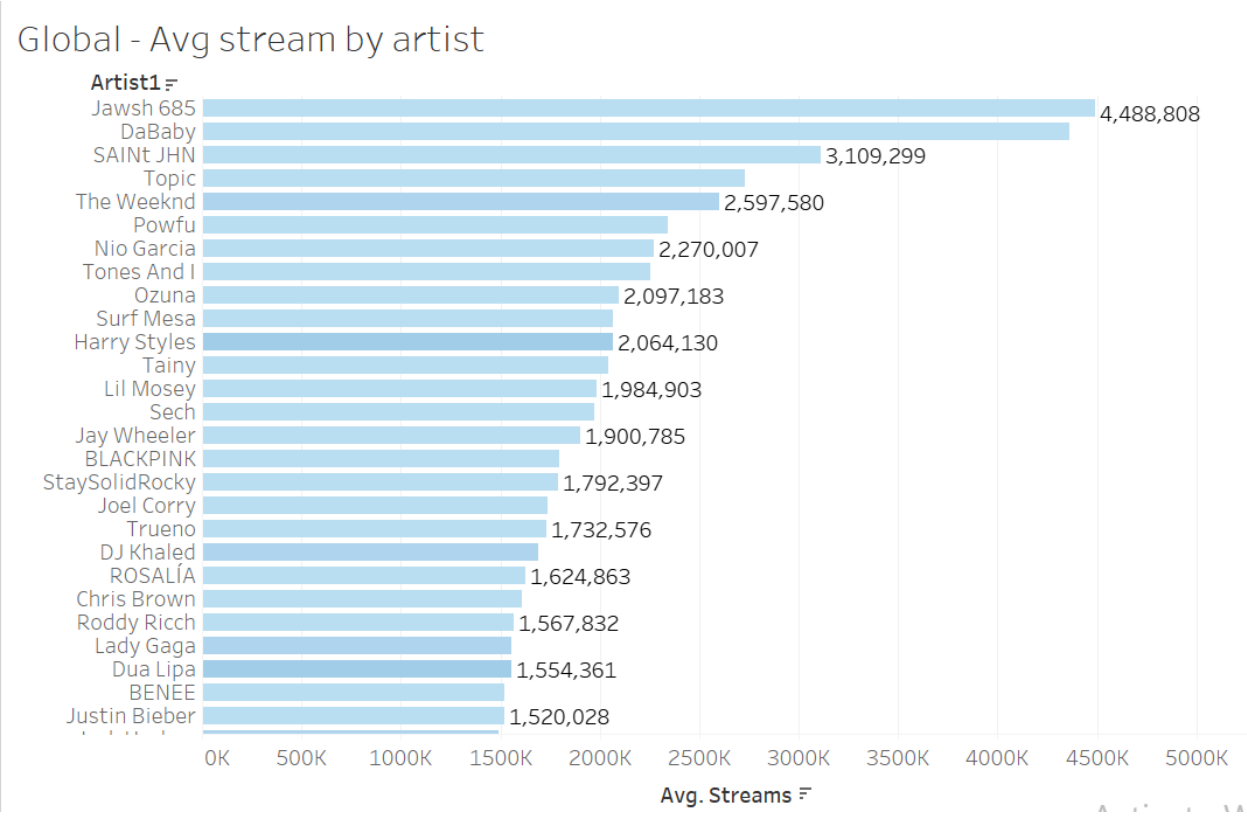


Top artists having highest number of stream



Considering Top 200 songs globally, Taylor Swift is the artist that has the highest times of streaming with more than 22 million views by 17 different songs. Juice WRLD ranks secondly with 13.6 million times by 11 songs. Pop Smoke is getting more than 7 million views from 5 different songs.

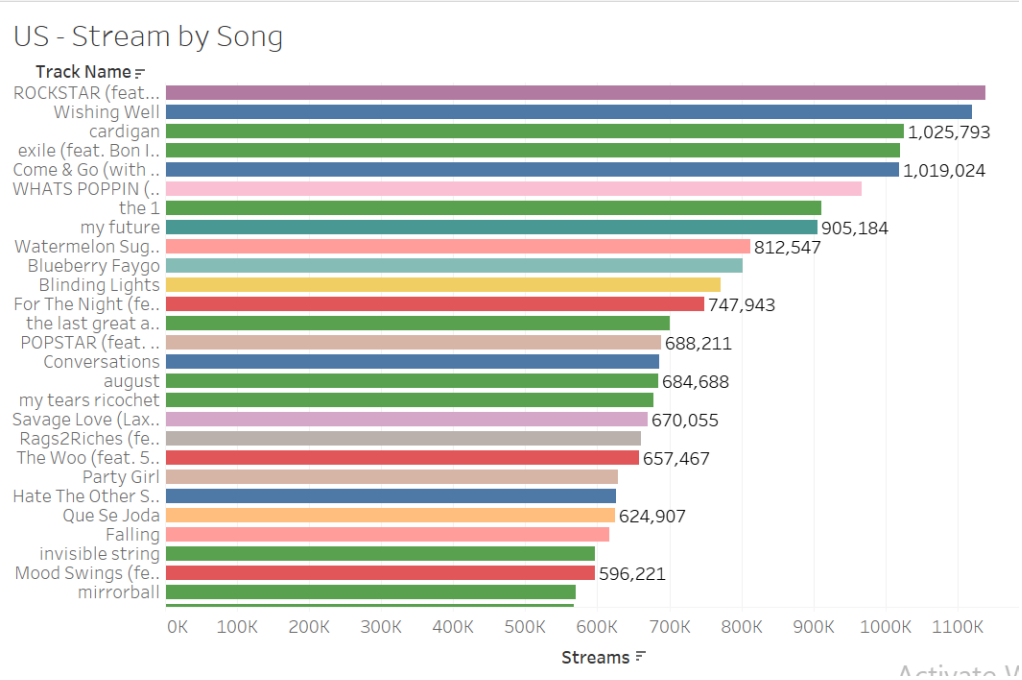
Average number of streams by artists

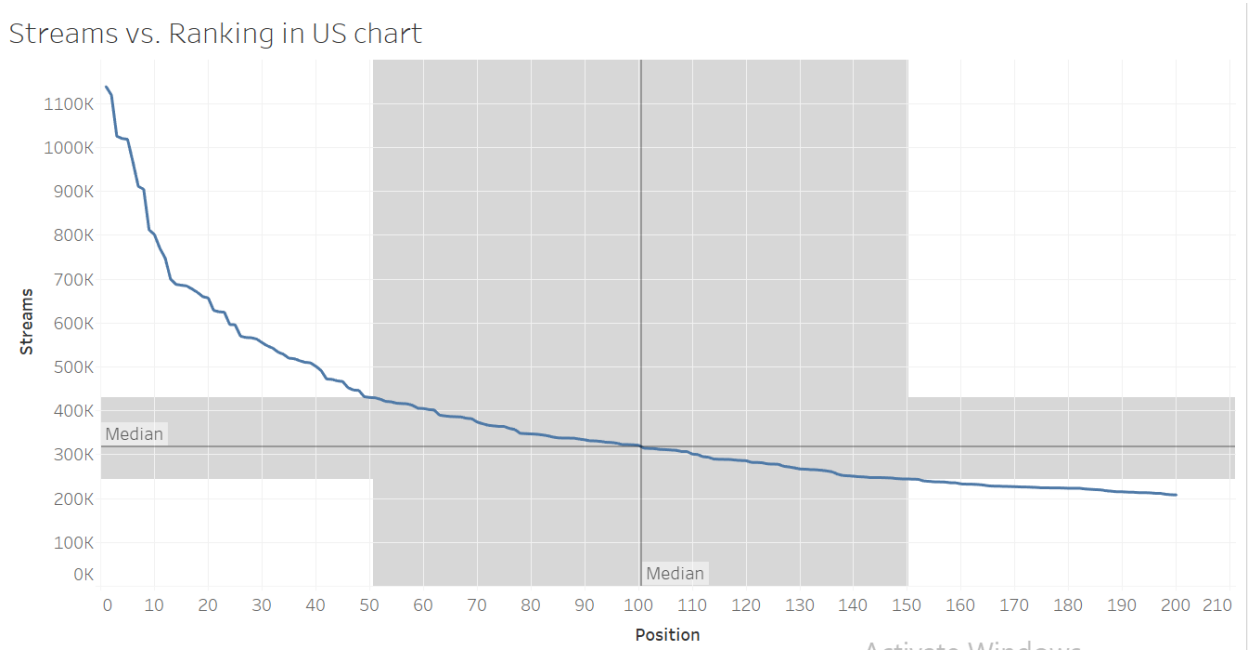


However, regarding average times of streaming by an artist per each song, Taylor Swift, Juice WRLD or Pop Smoke are not in Top 5 or even Top 5. Jawsh 685 has highest view with the song Savage Love. DaBaby comes after that with the song Rockstar. This information amplified that Taylor Swift and Jucies WRLD has a lot of popular songs but those are not the most streamed hits.

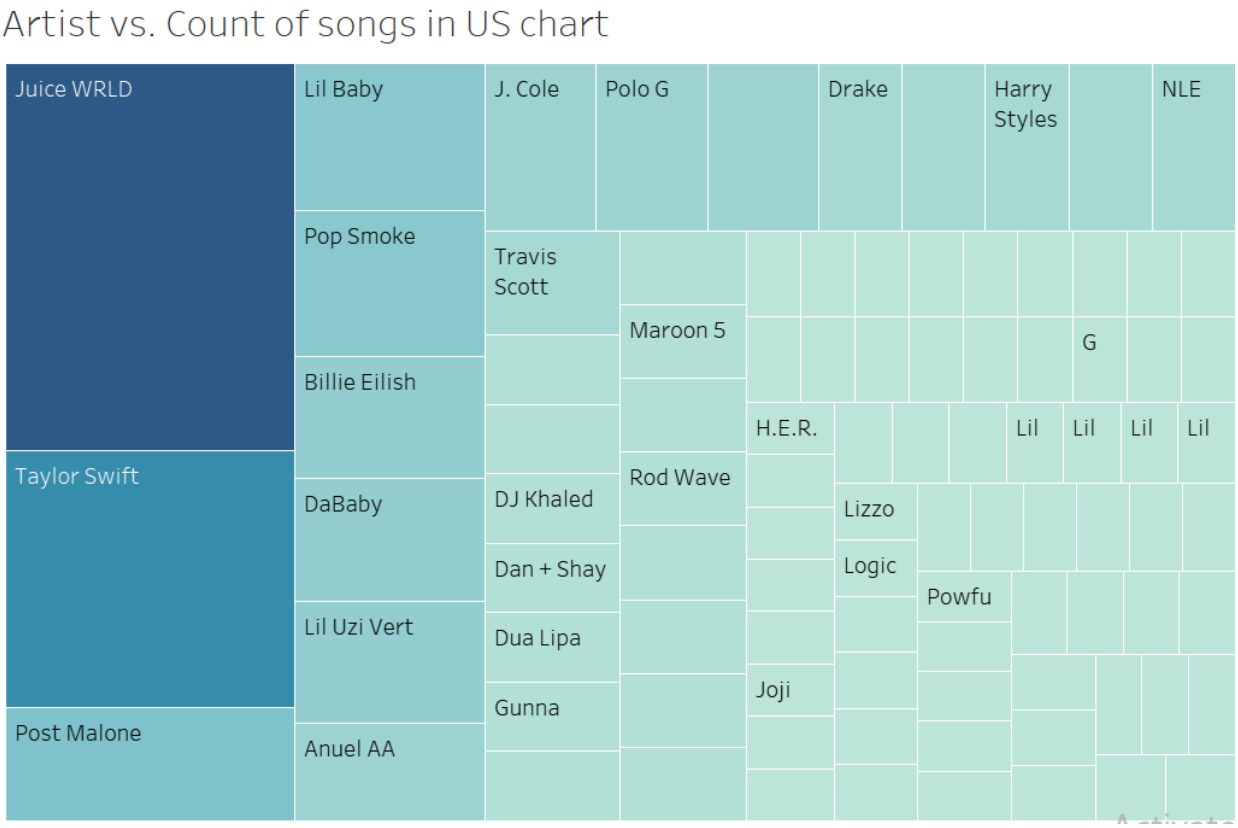
4.2 US Chart

Top songs in the US

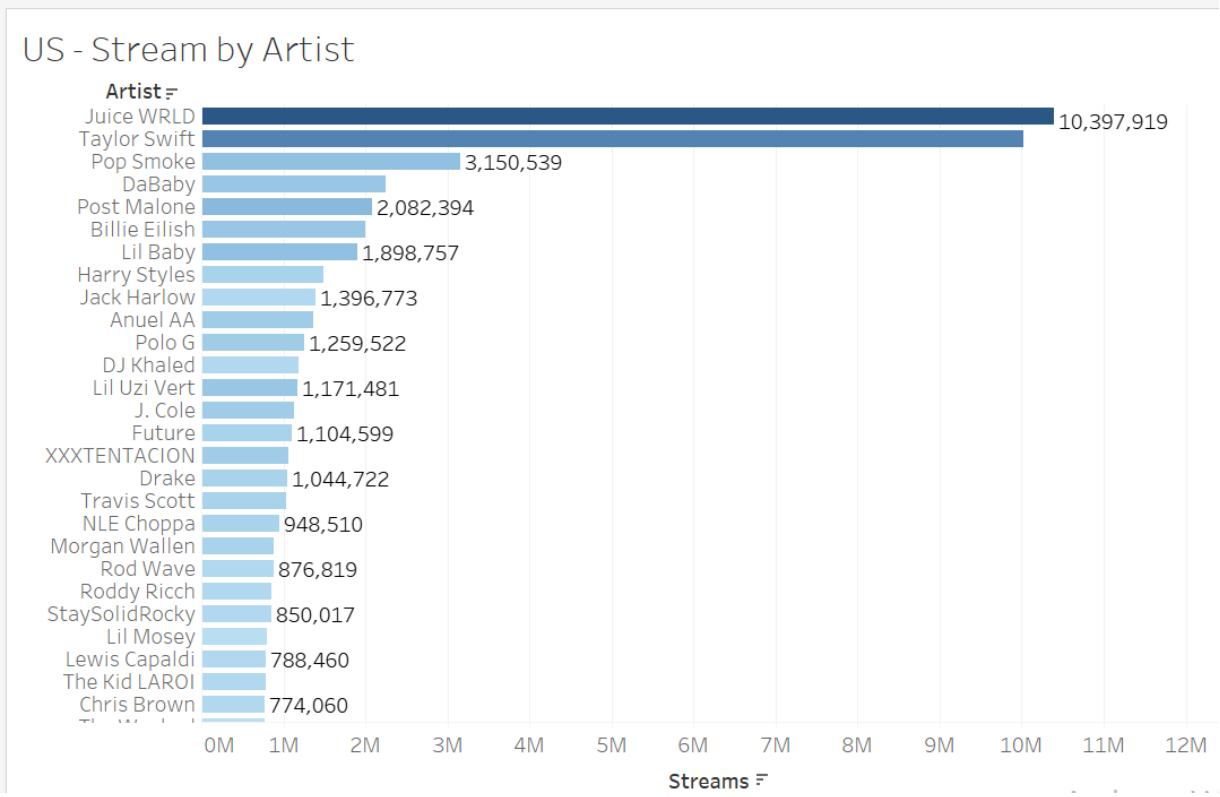




To be in Top 100 of the US chart, any songs should have more than 300,000 streaming times.

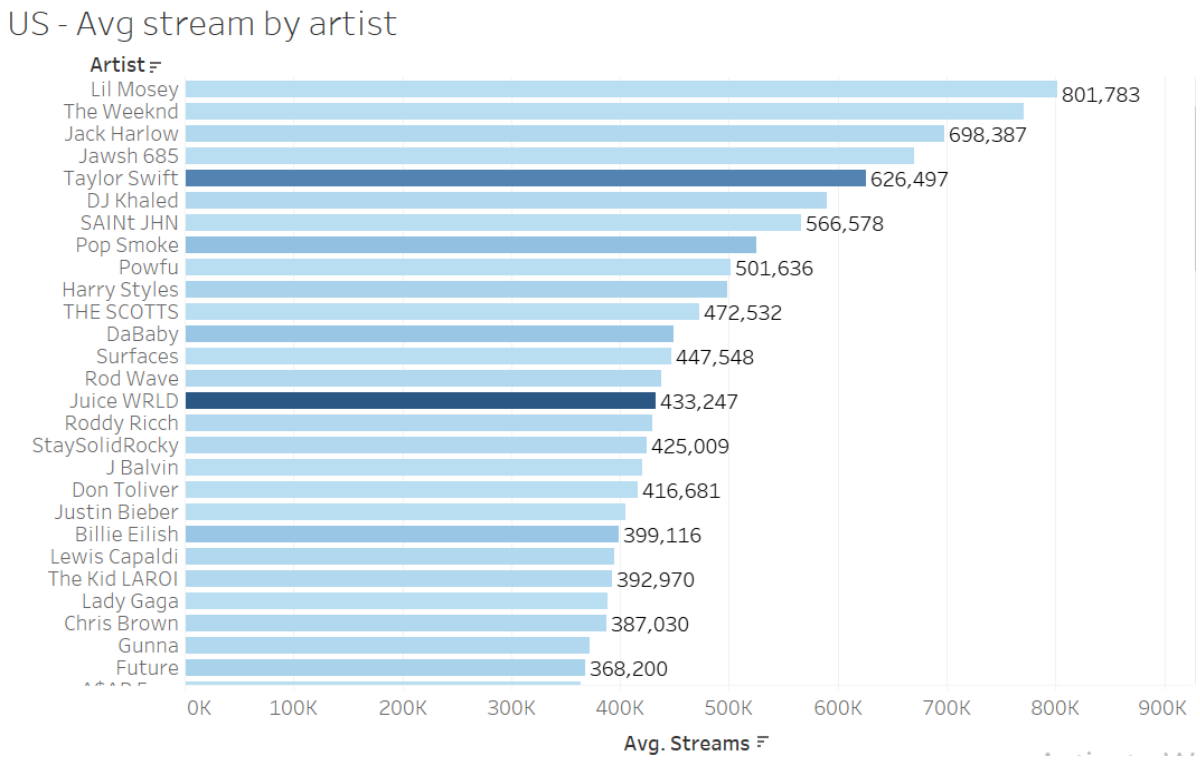


Top artists having highest times of streaming



Considering only in the US, Jucie WRLD with 24 songs has the total views higher than Taylor Swift with 17 songs. Pop Smoke ranks thirdly with more than 3 million views from 6 songs. This information show us that Taylor Swift is more popular than Jucie WRLD all over the world. However, considering only US market, people love Jucie WRLD's music more than Taylor Swift's.

Average stream by artists



Lil Mosey, The Weekend and Jack Harlow are top 3 artists that earn highest views per song. Taylor Swift is also in Top 5 with 626,497 views. Although DaBaby has the hit ROCKSTAR with highest times of streaming (more than 1.1 million views), on average, this artist only gets around 450,000 views average with 5 songs in Top 200. This means that other songs of DaBaby is not as popular as ROCKSTAR although both of them are in Top 200.

4.3 Comparison between Global chart and US chart

The analyst would like to see how many times that songs in global chart are listened by people in the United of States.

Position	Track Name	Artist	Global Stream	US Streams	Percentage Stream by US
1	Savage Love (Laxed - Sir	Jawsh 685	4488808	670055	15%
2	ROCKSTAR (feat. Roddy	DaBaby	4361245	1139206	26%
3	Blinding Lights	The Weeknd	3992791	770648	19%
4	Watermelon Sugar	Harry Styles	3860502	812547	21%
5	Roses - Imanbek Remix	SAINt JHN	3109299	566578	18%
6	cardigan	Taylor Swift	2965720	1025793	35%
7	my future	Billie Eilish	2938632	905184	31%
8	Breaking Me	Topic	2727768	330266	12%
9	Come & Go (with Marsh	Juice WRLD	2658830	1019024	38%
10	exile (feat. Bon Iver)	Taylor Swift	2550073	1020578	40%
	TOTAL STREAM		**33653668**	**8259879**	**25%**

Regarding Top 10 songs that are listened globally, on average, there is 25% of streams that are from people in the US. There are some songs that have higher percent of streams than average number. They are Come & Go, exile, my future, ROCKSTAR and cardigan. This means that these are songs seems to be on the trending in the US.

What are the positions of those Top 10 global songs in US chart?

```
Select g.Position as Global_Position, g.Track_Name, g.Artist, g.Streams as Global_Stream,
u.Streams as US_Stream, u.Position as US_Position
from [regional-global-daily-latest] g
inner join [regional-us-daily-latest] u
on g.url = u.url;
```

Global_Position	Track_Name	Artist	Global_Stream	US_Stream	US_Position
1	Savage Love (Laxed - Siren Beat)	Jawsh 685	4488808	670055	18
2	ROCKSTAR (feat. Roddy Ricch)	DaBaby	4361245	1139206	1
3	Blinding Lights	The Weeknd	3992791	770648	11
4	Watermelon Sugar	Harry Styles	3860502	812547	9
5	Roses - Imanbek Remix	SAINt JHN	3109299	566578	28
6	cardigan	Taylor Swift	2965720	1025793	3
7	my future	Billie Eilish	2938632	905184	8
8	Breaking Me	Topic	2727768	330266	93
9	Come & Go (with Marshmello)	Juice WRLD	2658830	1019024	5
10	exile (feat. Bon Iver)	Taylor Swift	2550073	1020578	4

Although Savage Love has the 1st ranking of streaming, there seem to be a decrease in its popularity in the US with the ranking of 18th. On the other hand, ROCKSTAR is doing well in both global and US billboards. Taylor Swift's recent songs are in both Top 10 global and US chart.

```
Select count(*) as 'Number of songs in both US chart & global chart'
from (Select g.Position, g.Track_Name, g.Artist, g.Streams as Global_Stream, u.Streams as US_Stream
from [regional-global-daily-latest] g
inner join [regional-us-daily-latest] u
on g.url = u.url) sub
```

Number of songs in both US chart & global chart
119

US chart and global chart has total 119 songs in common. This means that nearly 60% of global music in Top 200 are influenced by US music.

There are 81 songs in Top 200 US chart that is not in Top 200 Global chart. This means that these songs below are popular only in the US not globally.

```
Select u.Position, u.Track_Name, u.Artist, u.Streams
from [regional-us-daily-latest] u
where URL not in (select URL from [regional-global-daily-latest])
Order by u.Position ASC
```

Position	Track_Name	Artist	Streams
57	The Bigger Picture	Lil Baby	415897
72	Narrow Road (feat. Lil Baby)	NLE Choppa	367260
76	Stay High	Juice WRLD	359927
86	PEEP HOLE	DaBaby	337958
87	Bad Energy	Juice WRLD	337880
88	Tell Me U Luv Me (with Trippie Redd)	Juice WRLD	337722
91	Titanic	Juice WRLD	331554
92	Patek	Future	331454
96	Man Of The Year	Juice WRLD	326195
97	Lion King On Ice	J. Cole	323170
100	Up Up And Away	Juice WRLD	321602
101	Prospect (ft. Lil Baby)	iann dior	314938
103	Tap In	Saweetie	313889
104	High Fashion (feat. Mustard)	Roddy Ricch	312003

5. Relationship between variables

The information is analyzed by the data Top 50 Spotify songs 2019

Data source: https://www.kaggle.com/leonardopena/top50spotify2019

The dataset is about the top 50 most listened songs in the world by Spotify. This dataset has several variables about the songs. The data includes 50 songs and 13 variables

```
df.head()
```

	Unnamed: 0	Track.Name	Artist.Name	Genre	Beats.Per.Minute	Energy	Danceability	Loudness..dB..	Liveness	Valence.	Length.	Acousticness..	Speechi
0	1	Señorita	Shawn Mendes	canadian pop	117	55	76	-6	8	75	191	4	
1	2	China	Anuel AA	reggaeton flow	105	81	79	-4	8	61	302	8	
2	3	boyfriend (with Social House)	Ariana Grande	dance pop	190	80	40	-4	16	70	186	12	
3	4	Beautiful People (feat. Khalid)	Ed Sheeran	pop	93	65	64	-8	8	55	198	12	
4	5	Goodbyes (Feat. Young Thug)	Post Malone	dfw rap	150	65	58	-4	11	18	175	45	

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 14 columns):
Unnamed: 0          50 non-null int64
Track.Name          50 non-null object
Artist.Name         50 non-null object
Genre               50 non-null object
Beats.Per.Minute    50 non-null int64
Energy              50 non-null int64
Danceability        50 non-null int64
Loudness..dB..      50 non-null int64
Liveness            50 non-null int64
Valence.            50 non-null int64
Length.             50 non-null int64
Acousticness..      50 non-null int64
Speechiness.        50 non-null int64
Popularity          50 non-null int64
dtypes: int64(11), object(3)
memory usage: 5.6+ KB
```

Varaibale definition:

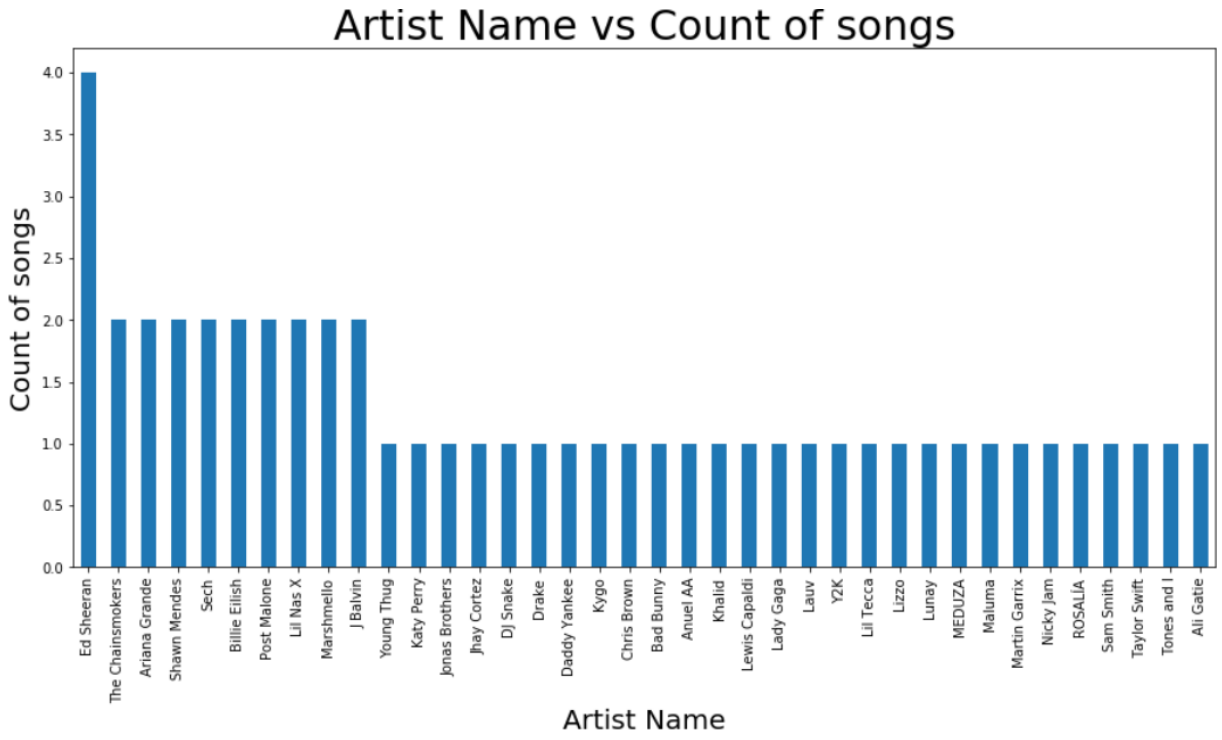
- Danceability: Describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity.
- Valence: Describes the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
- Energy: Represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale.
- Beats per minute (BPM): the speed or pace of a given piece, and derives directly from the average beat duration.
- Loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks.
- Speechiness: This detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value.

- Liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.
- Acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic.

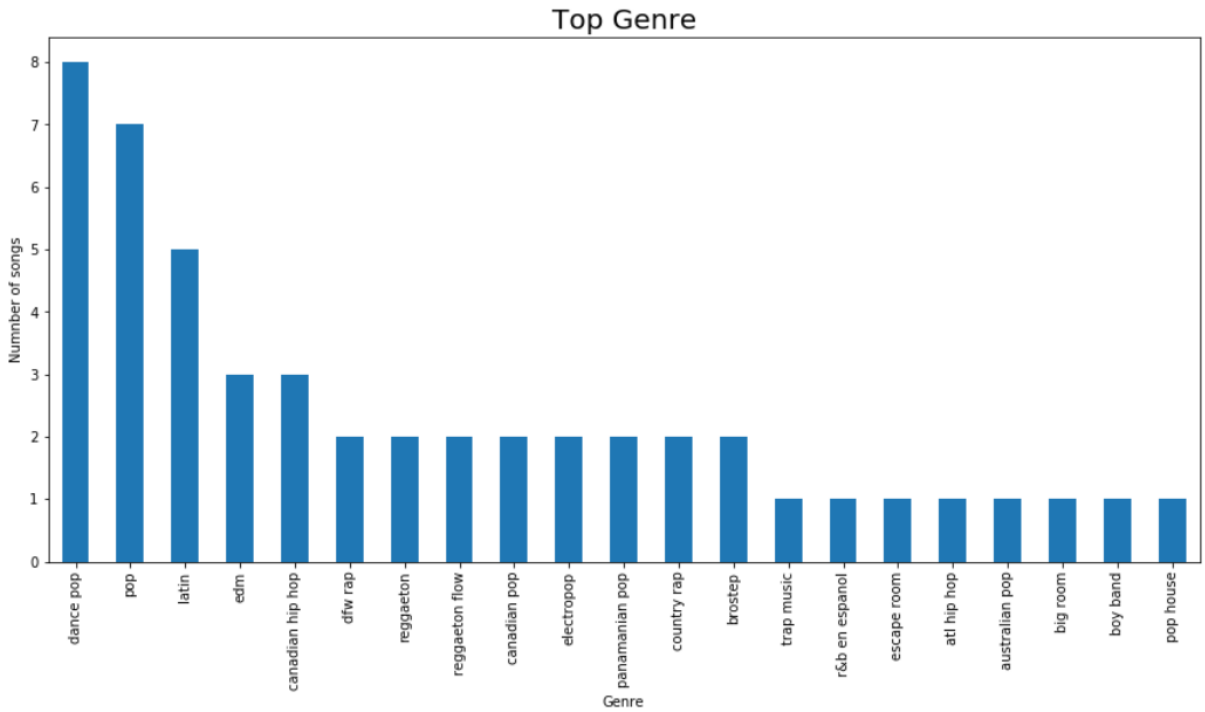
An overview of the quantitative variables

```
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



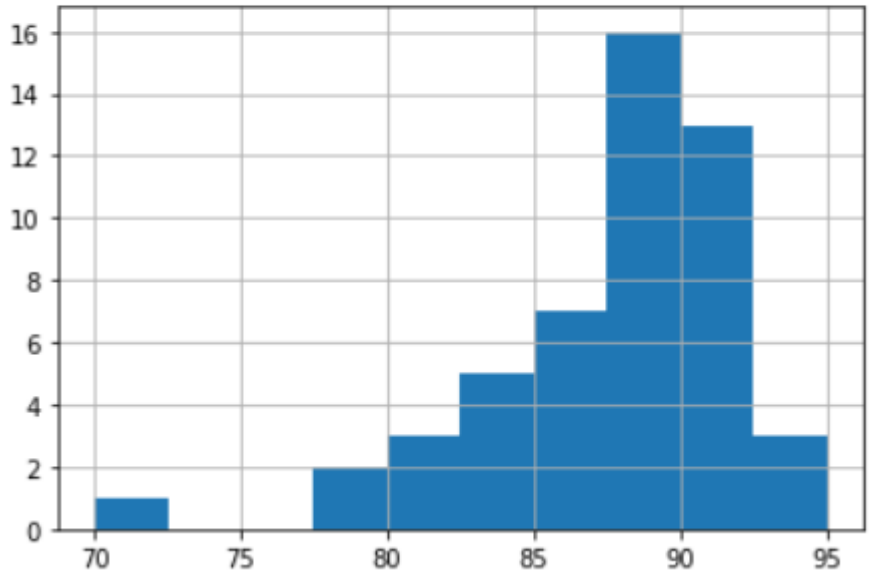
In 2019, Ed Sheeren has the most number of song in top 50 songs of Spotify with 4 songs. The Chainsmokers, Ariana Grande, Shawn Mendes, Sech, Billie Ellish and Post Malone share the 2nd position with 2 songs for each.



Dance pop (8 songs), pop (7 songs) and latin (5 songs) are the most popular genre in top 50 Spotify 2019.

```
df['Popularity'].hist()
```

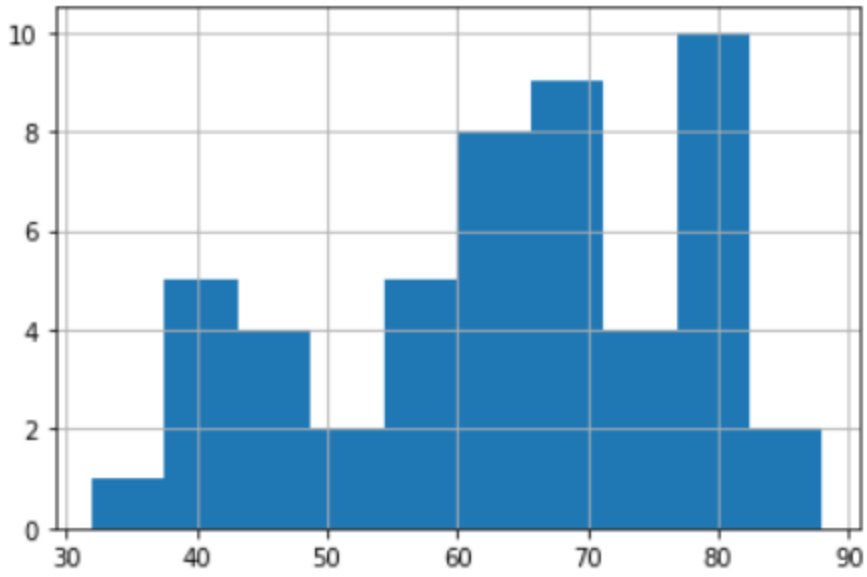
```
<matplotlib.axes._subplots.AxesSubplot at 0x25aef386f48>
```



In general, most of the songs in top 50 has the popularity from 87 to 90.

```
df['Energy'].hist()
```

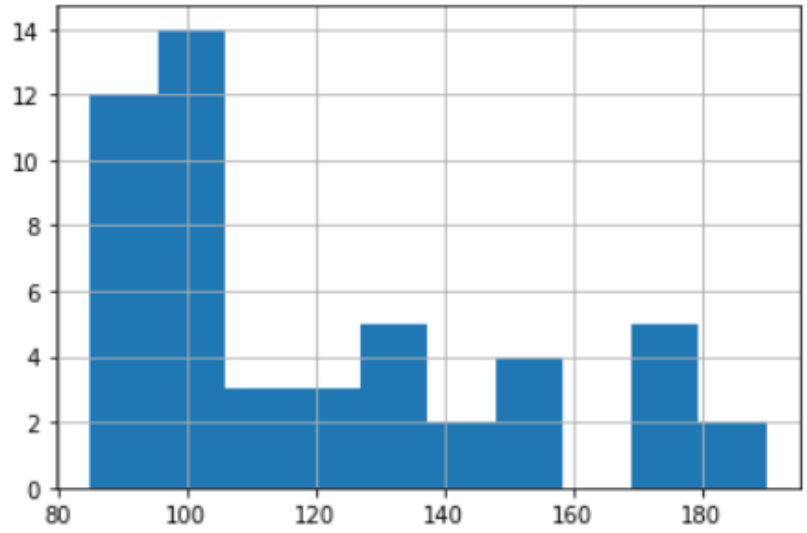
```
<matplotlib.axes._subplots.AxesSubplot at 0x25af56562c8>
```



The distribution of energy focus on three branches: around 40, from 60 to 70 and around 80.

```
df['Beats.Per.Minute'].hist()
```

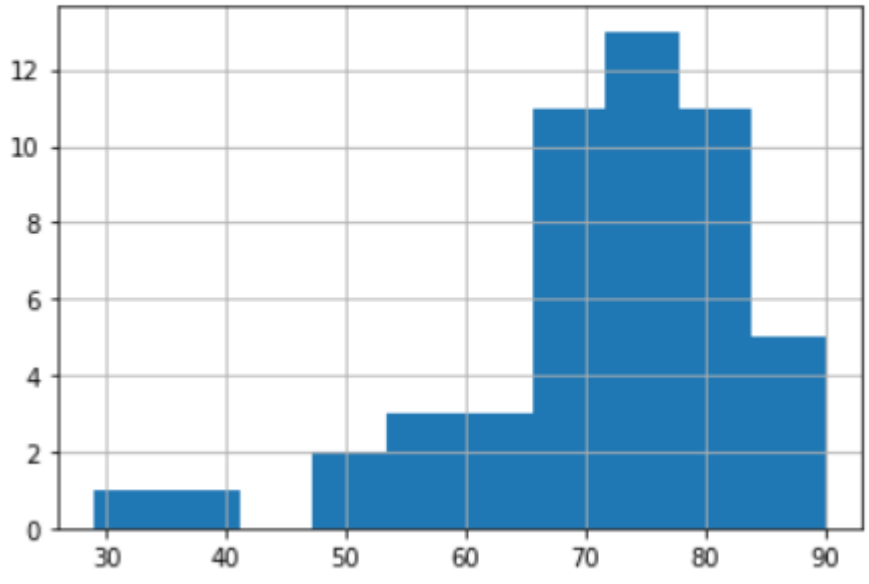
```
<matplotlib.axes._subplots.AxesSubplot at 0x25af557d208>
```



In general, most of the songs in top 50 has the Beats per Minutes from range 85 to 115.

```
df['Danceability'].hist()
```

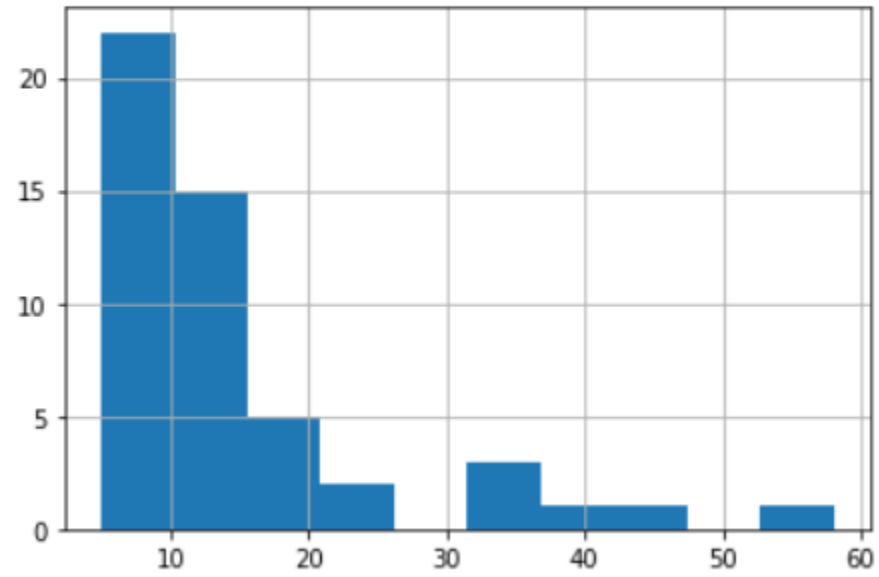
```
<matplotlib.axes._subplots.AxesSubplot at 0x25af7e88c
```



In general, most of the songs in top 50 has the Danceability at around 85. This means most of the songs have the ability make people dance.

```
df['Liveness'].hist()
```

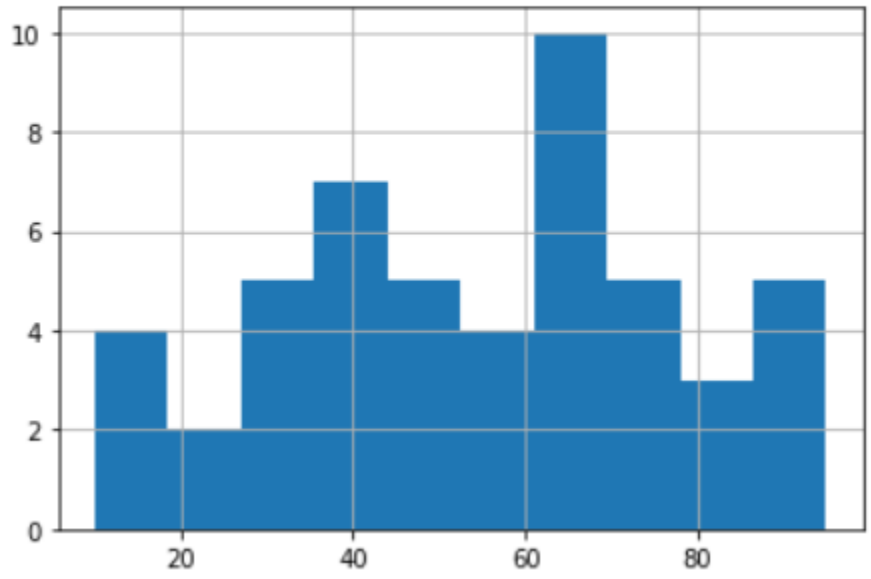
```
<matplotlib.axes._subplots.AxesSubplot at 0x25af811e408>
```



In general, most of the songs in top 50 has the Liveness at around 5.

```
df['Valence.'].hist()
```

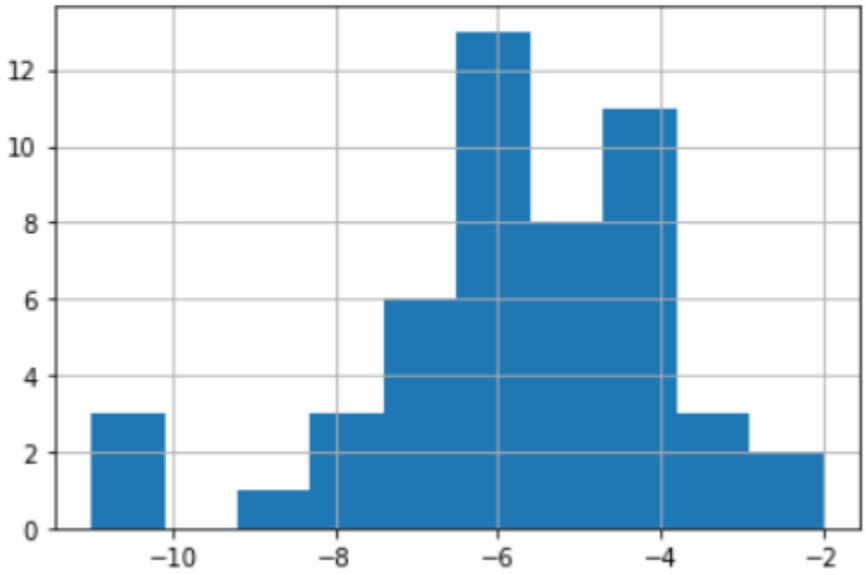
```
<matplotlib.axes._subplots.AxesSubplot at 0x25af81a7908>
```



In general, most of the songs in top 50 has the Liveness from 60 to 70. However, there are a number of songs in top 50 has the Valence at around 40. This means that those song with medium level of cheerfulness has the highest distribution.

```
df['Loudness..dB..'].hist()
```

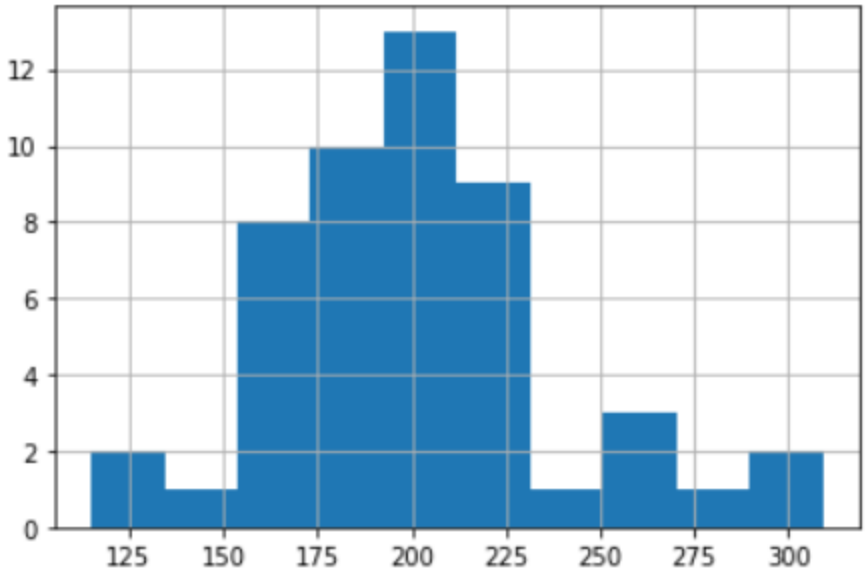
```
<matplotlib.axes._subplots.AxesSubplot at 0x25af81caf88>
```



In general, the distribution of Loudness is mainly around -6 and -4.

```
df['Length.'].hist()
```

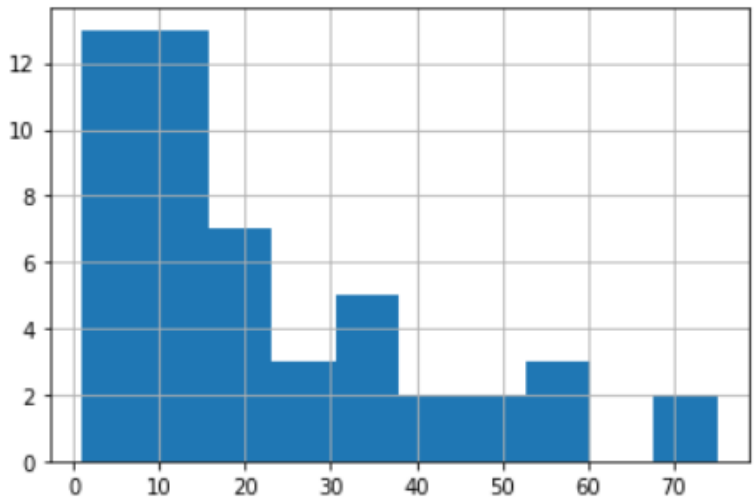
```
<matplotlib.axes._subplots.AxesSubplot at 0x25af7eed408>
```



Most of the songs has the length at around 200 seconds. The distribution of the length is mainly from 175 to 225 seconds.

```
df['Acousticness..'].hist()
```

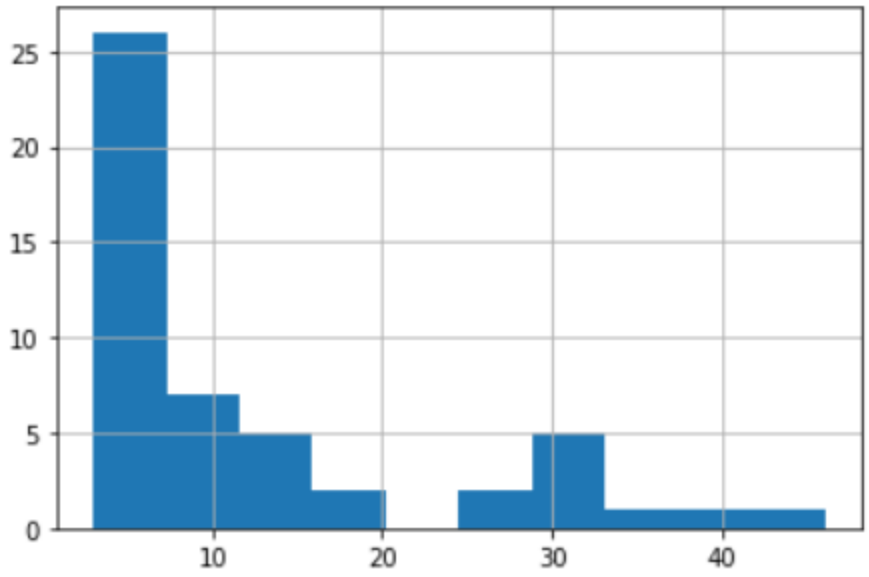
```
<matplotlib.axes._subplots.AxesSubplot at 0x25af8322bc8>
```



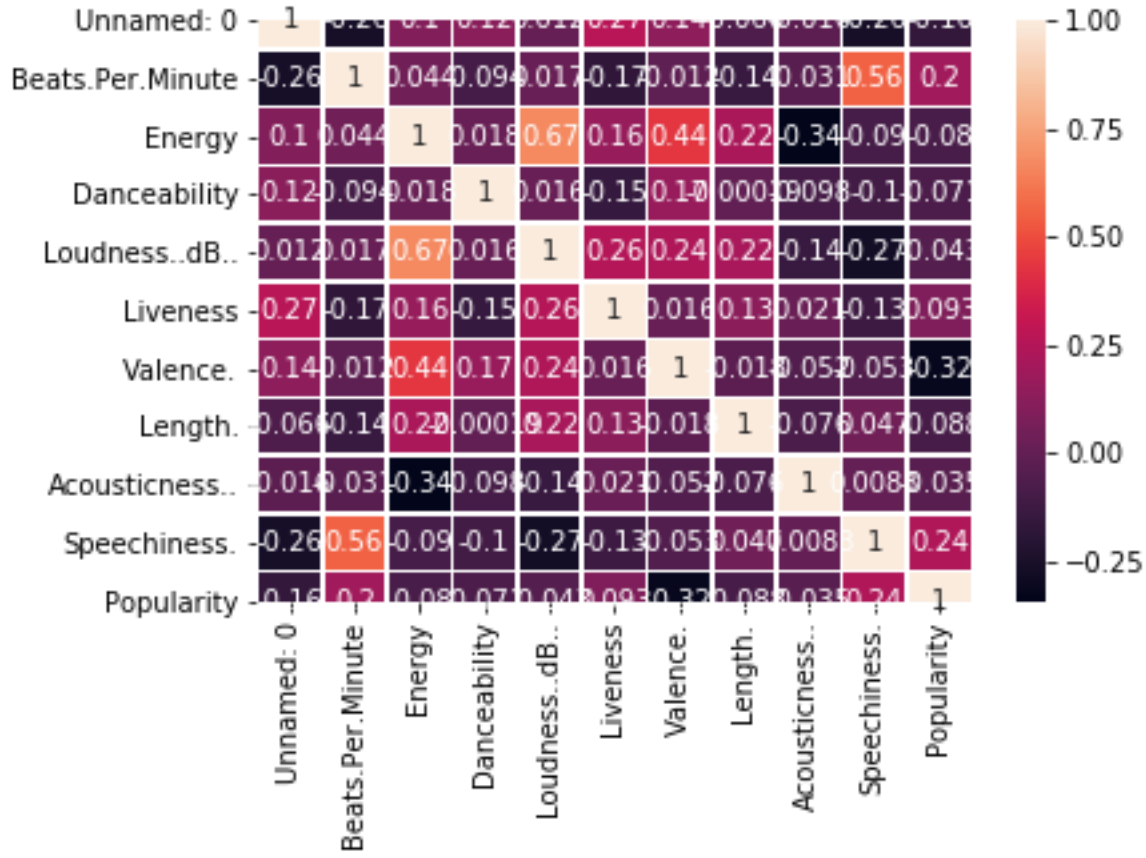
Most of the songs have the acousticness from 0 10 15.

```
df['Speechiness.'].hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x25af834edc
```



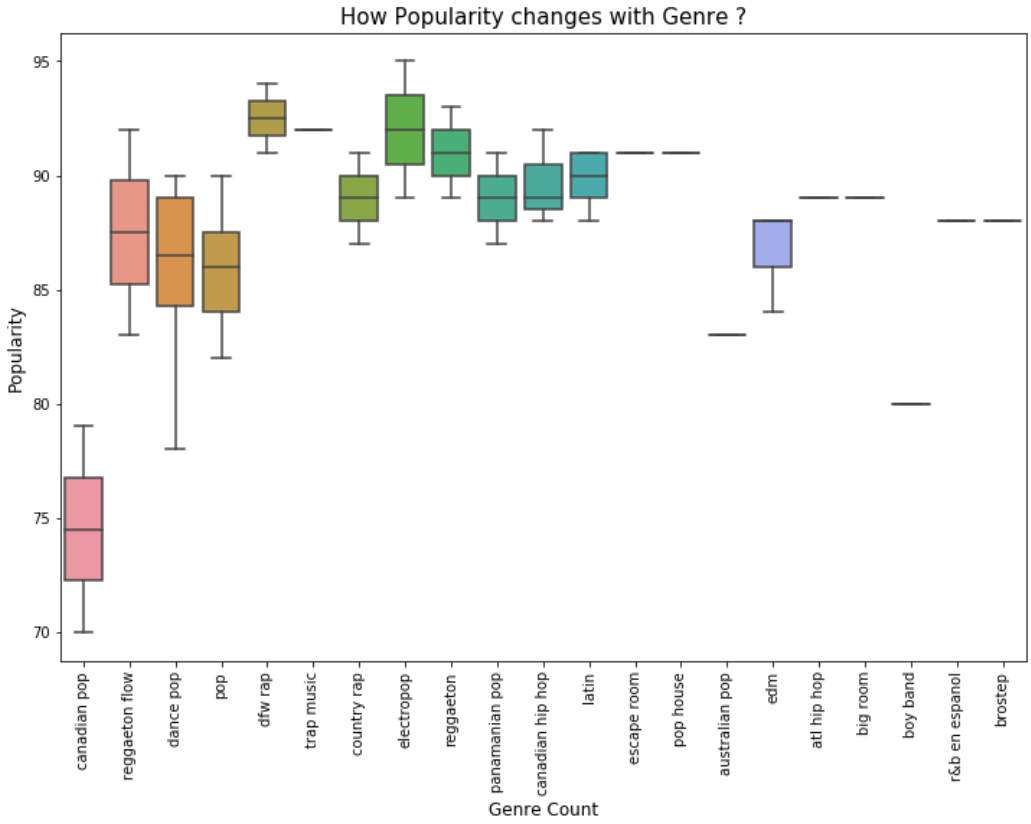
Most of the songs have the speechiness figure less than 8.



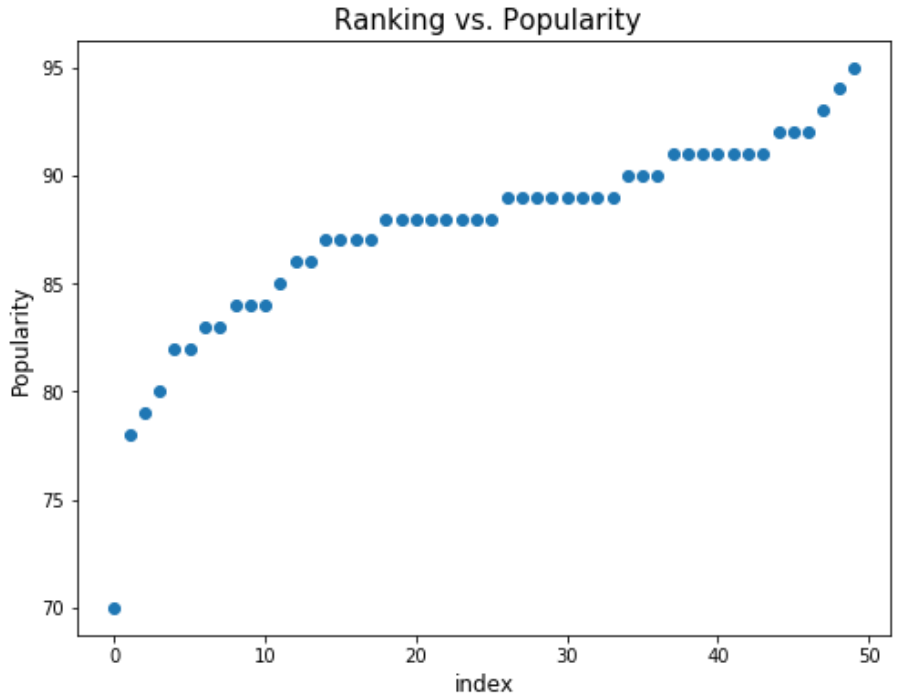
	Unnamed: 0	Beats.Per.Minute	Energy	Danceability	Loudness..dB..	Liveness	Valence.	Length.	Acousticness..	Speechiness.	Popularity
Unnamed: 0	1.000000	-0.259193	0.102649	0.122691	0.011914	0.270659	0.137329	-0.065844	-0.015993	-0.257506	-0.160680
Beats.Per.Minute	-0.259193	1.000000	0.043756	-0.094183	0.017016	-0.167286	-0.011586	-0.139288	-0.031450	0.557052	0.196097
Energy	0.102649	0.043756	1.000000	0.018254	0.670794	0.162768	0.438820	0.224677	-0.339892	-0.089860	-0.080295
Danceability	0.122691	-0.094183	0.018254	1.000000	0.016255	-0.149636	0.172829	-0.000185	-0.098165	-0.103472	-0.071413
Loudness..dB..	0.011914	0.017016	0.670794	0.016255	1.000000	0.258652	0.237614	0.219219	-0.138300	-0.272213	-0.043085
Liveness	0.270659	-0.167286	0.162768	-0.149636	0.258652	1.000000	0.016123	0.131782	0.021328	-0.125286	0.092564
Valence.	0.137329	-0.011586	0.438820	0.172829	0.237614	0.016123	1.000000	-0.017782	-0.052323	-0.053242	-0.317752
Length.	-0.065844	-0.139288	0.224677	-0.000185	0.219219	0.131782	-0.017782	1.000000	-0.076293	0.046755	-0.087639
Acousticness..	-0.015993	-0.031450	-0.339892	-0.098165	-0.138300	0.021328	-0.052323	-0.076293	1.000000	0.008293	-0.034684
Speechiness.	-0.257506	0.557052	-0.089860	-0.103472	-0.272213	-0.125286	-0.053242	0.046755	0.008293	1.000000	0.238553
Popularity	-0.160680	0.196097	-0.080295	-0.071413	-0.043085	0.092564	-0.317752	-0.087639	-0.034684	0.238553	1.000000

Bivariate Exploration

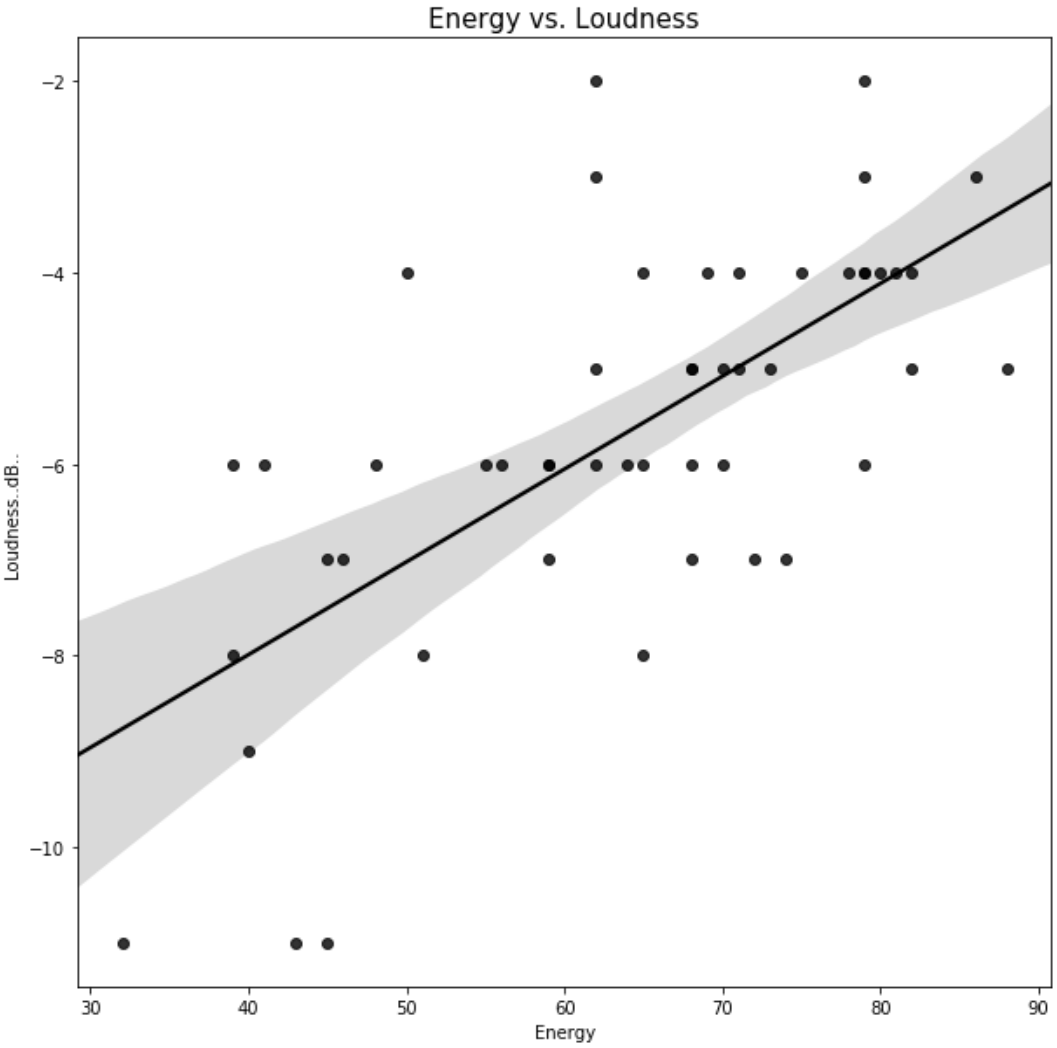
- Beats Per Minute had a positive relationship with speechiness with the correlation of 0.56.
- Energy had a strong relationship with Loudness and Valence but it has a negative relationship with acousticness.
- Danceability didn't really have any strong relationships with other attributes. Together with the distribution of danceability, it can be inferred that any songs in top 50 had an ability to make people dance regardless of other attributes.
- Loudness had a negative relationship with speechiness but strong positive relationship with Energy.
- Liveness didn't really have any extremely strong relationships with other attributes. However, it did have a negative relationship with Beats Per Minutes but positive relationship with Loudness and Ranking of the Songs.
- Valence had a positive relationship with Energy but negative relationship with Popularity.
- Length didn't have any strong relationship with other attributes except slightly positive relationship with Loudness and Energy. This means that longer song tends to have higher level of Loudness and Energy.
- Position (Unamed column) has a slightly positive relationship with Valence and Liveness but a negative one with Speechiness, Beats Per Minute and Popularity.



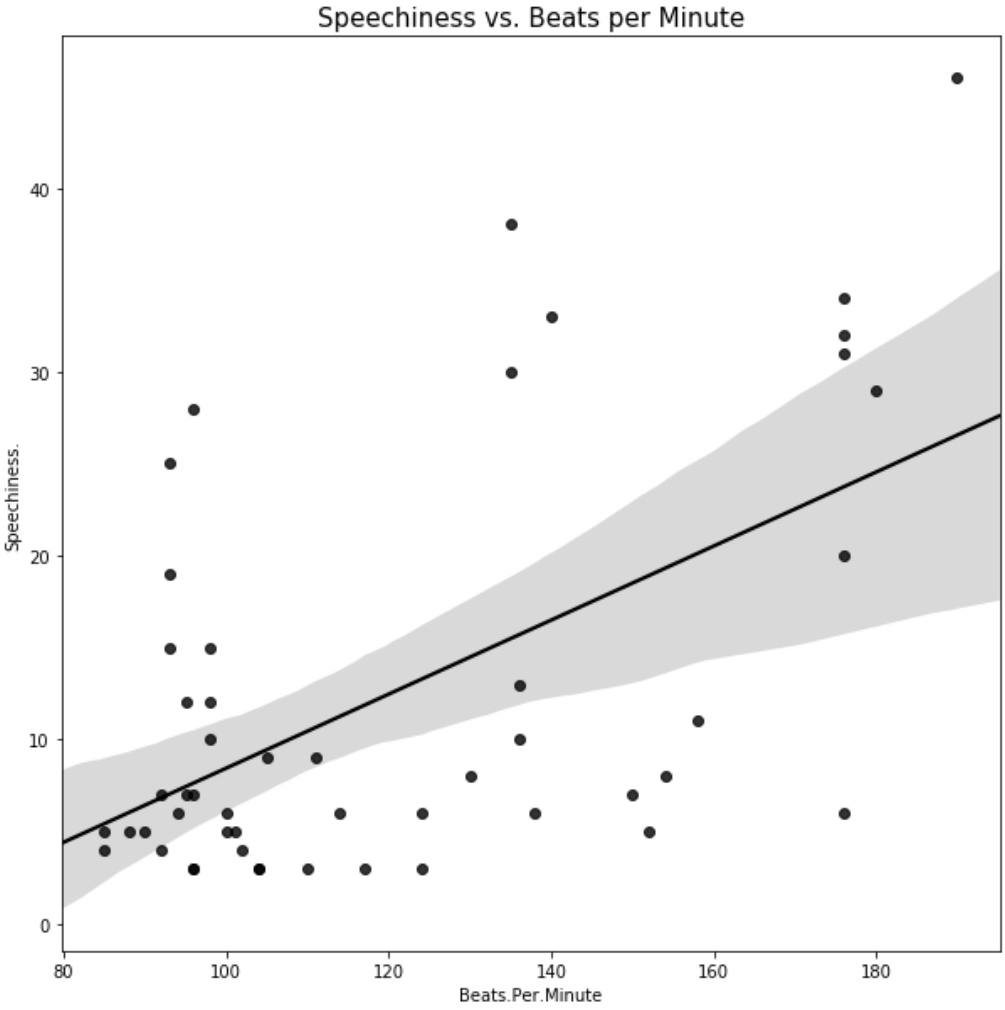
Electric Pop had the highest popularity while Canadian pop's popularity is the lowest.



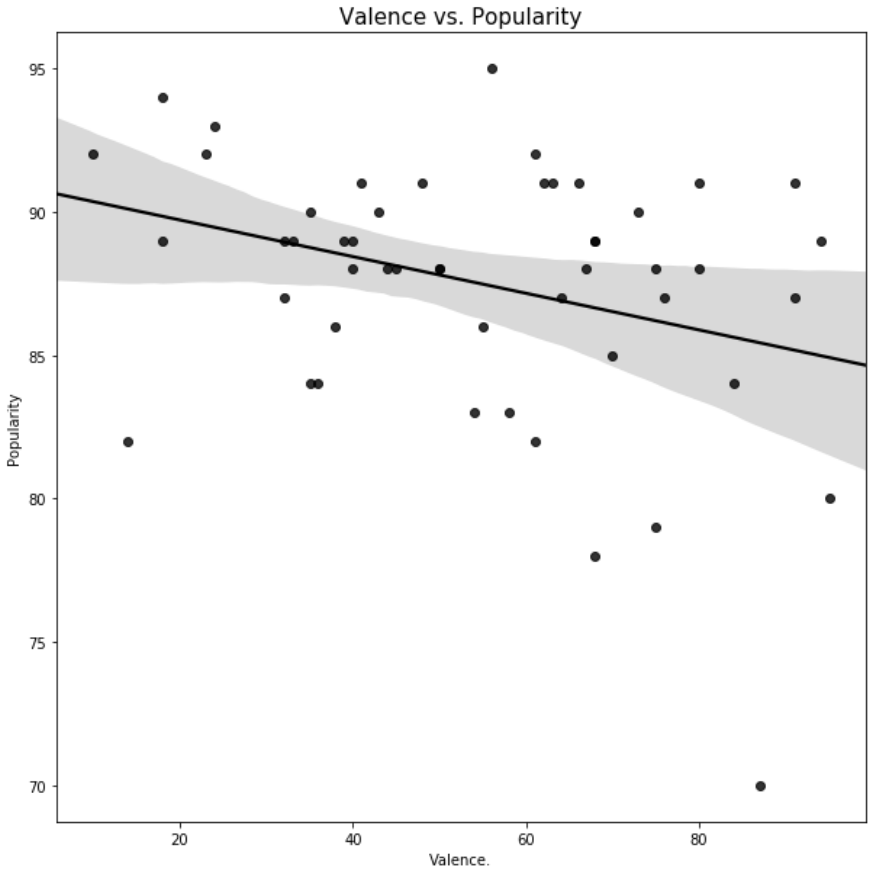
Lower ranking songs tends to have higher level of popularity.



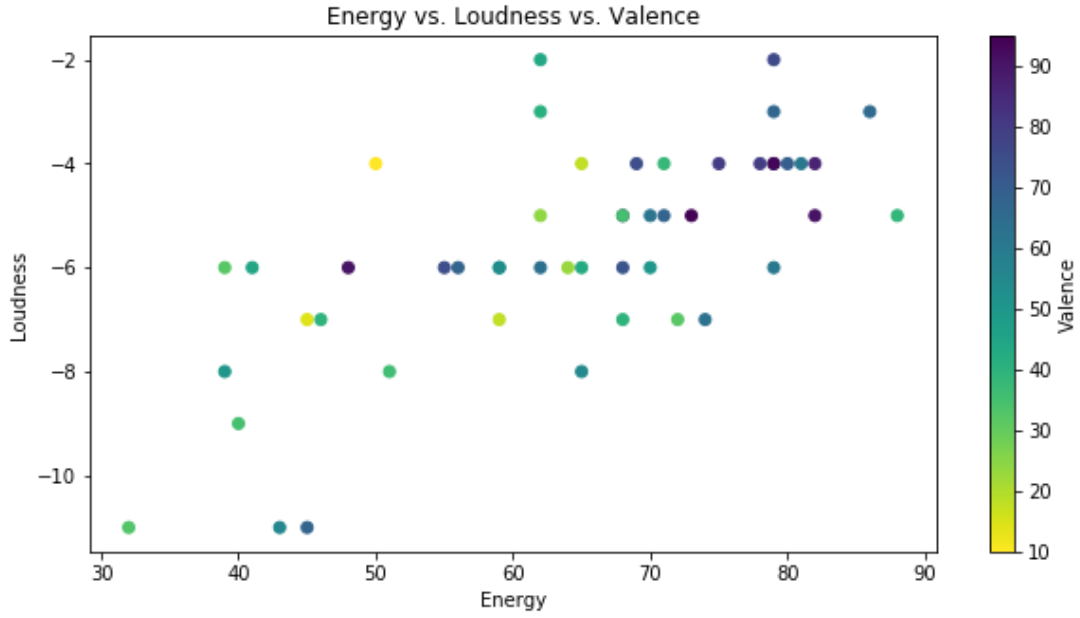
Energy and Loudness had a positive relationship together. The higher the level of loudness of a song is, its energy tend to be higher.



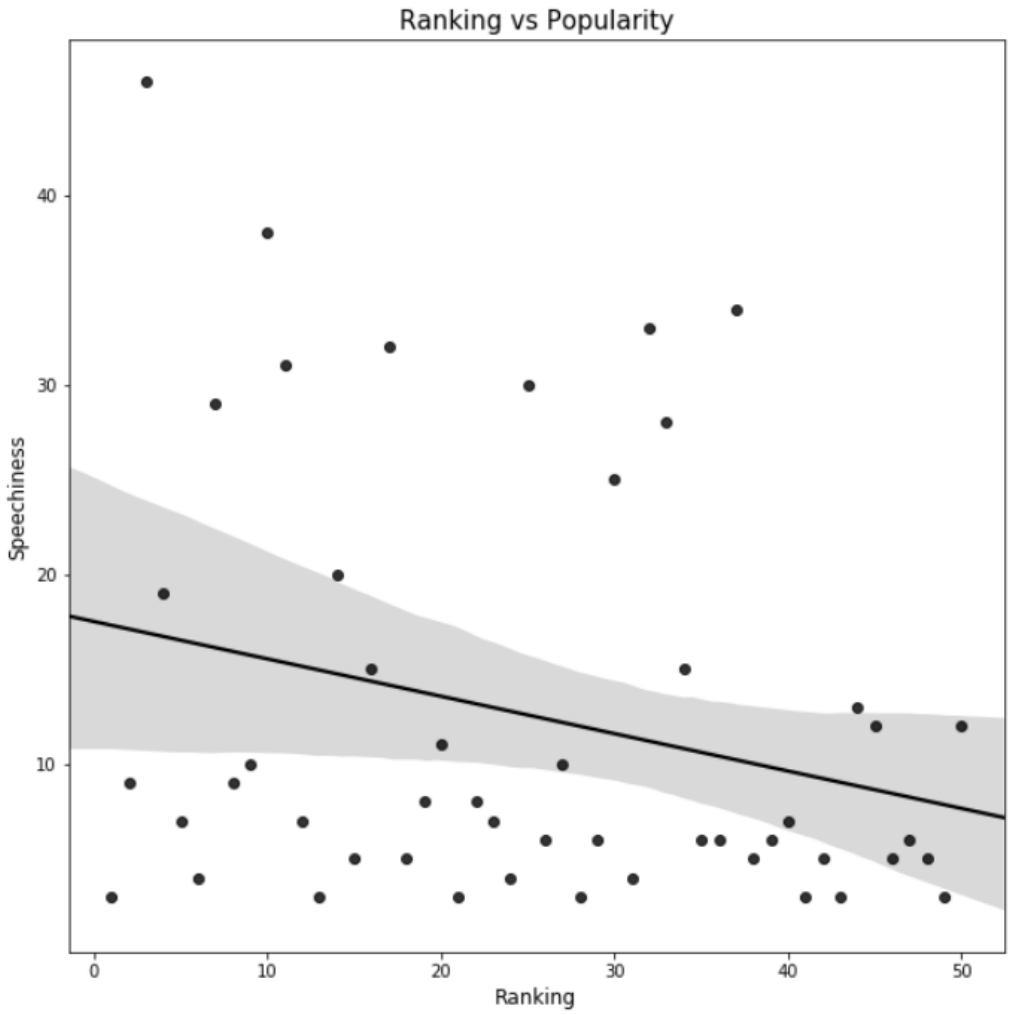
Speechiness and Beats per Minute had a positive relationship with each other.



Popularity and Valence had a negative relationship with each other.

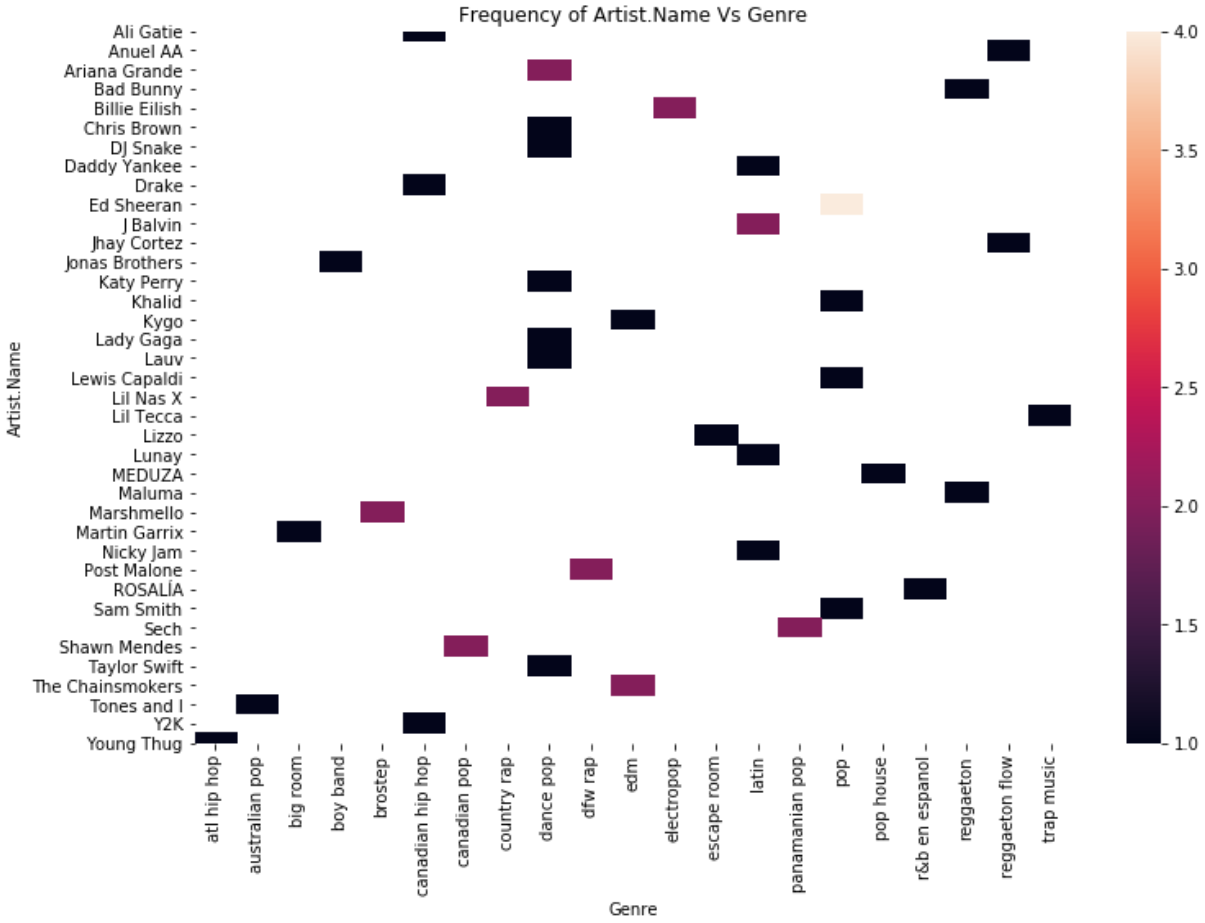


Energy, Valence and Loudness had a positive relationship together. This means that the more cheerful a song is, the higher level of loudness and energy it has.



Higher ranking songs tend to have slightly higher level of speechiness.

As analyzed earlier above, dance pop (8 songs), pop (7 songs) and latin are the most favorite genre in top 50 Spotify 2019. When considering both Artist and Genre together, we can see that Ed Sheeran is the king of pop music with 4 out of 8 pop songs in Top 50. Both of songs of Ariana Grande in Top 50 have the genre of dance pop.



IV. Summary and conclusions

Regarding Top 200 Spotify in global chart and US chart:

- Taylor Swift and Juice WRLD are top artist that have highest number of songs in Top 200 both globally and in US. Taylor Swift is the most popular artist with highest number of songs and highest number of total view.
- US chart and global chart has total 119 songs in common. This means that nearly 60% of global music are influenced by US music considering Top 200 songs by Spotify globally and in the US.

Regarding the relationship between variables in Top 50 Spotify songs 2019:

- Energy, Valence and Loudness had a positive relationship together. This means that the more cheerful a song is, the higher level of loudness and energy it has.
- Speechiness and Beats per Minute had a positive relationship with each other.
- Popularity and Valence had a negative relationship with each other.
- Electric Pop had the highest popularity while Canadian pop's popularity is the lowest
- Higher ranking songs tend to have slightly higher level of speechiness and Beats Per Minutes.
- Most of the song in Top 50 Spotify songs 2019 has an ability to make people dance.