VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

DATA VISUALIZATION

J Component Review 3

An Analysis of the Recent Music Trends on Online Streaming Platforms such as Spotify

Submitted by

Sarthak Bajaj- 19BCE0710

Srushti Jagtap - 19BCE0325

Sehaj Kapoor - 19BCE0896

Ranjal Chirag Shenoy - 19BCE0356

Aviral Goyal – 19BCE0883

Faculty: PRAVAT KUMAR JENA

Slot – B1

I. ABSTRACT

Streaming services have become a key player in the cultural industries in sharing media content with audiences. This chapter addresses how on-demand music-streaming services, the world's most popular format for the distribution of recorded music, have driven new professional music industry practices that are affected by, and affect in turn, the ways in which music communicates. It identifies new dynamics in the relationship between listeners and music. It then analyses the ways in which these dynamics afford yet other distribution practices in the music industry, according to two communication patterns. These patterns have particular purposes and methods but share an alignment with the logics of distributed communication, either within or outside of the streaming services, where the struggle for audience attention is paramount. The chapter concludes with a discussion of streaming's impact upon the negotiation of new practices in the music industry derived from the abundance and intangibility of those services, as well as their multiple options for music consumption. The chapter explains how the communication adapted to the streaming paradigm is characterized by content circulation among the layers and fragments of global networks and multiple platforms, linking artists, fans, music, and the industry in new, less predictable ways. The work of communication management hence grows in importance in a streaming-dominated music industry that might also be characterized as a communication industry in its own right.

Music rankings are mainly aimed at marketing purposes but also help users in discovering new music as well as comparing songs, artists, albums, etc. This work will present an interactive way to visualize, find and compare music rankings using different techniques, as well as displaying music attributes. Our visualization makes easier to obtain information about artists and tracks, and also to compare the data gathered from the top major music rankings of Spotify.

Different visualization tools are used to show various aspects of the dataset found. This study is necessary since it helps in the recommendation algorithms that are used in modern day streaming apps for music. Song specific data is also visualized which shows the recent interests of people in music which helps artists to keep up and blend in to produce new music. It also shows which artists have been the most influential and who has given the greatest number of hits in the past decade. The visualizations show the top genres and also the shift of the music taste over the decade.

The goal of this project is to visualize and analyse the recent trends in music over streaming applications such as Spotify for consumers to have a better idea of the popularity of artists as well as the music taste depending on the timelines which can be very useful in recommendation algorithms used in future streaming platforms.

II. INTRODUCTION

The term ‘streaming media’ is ambiguous and, generally speaking, used rather loosely in everyday and industrial contexts. The forces that brought about streaming have been recounted by others (Hesmondhalgh and Meier, 2017) and are beyond the scope of detail than can be recounted here. Historically, ‘streaming’ appears in the 1990s to describe a technical process for delivering media over the internet in ‘real time,’ without the file being downloaded or stored on a local drive. Alternatively, the phrase sometimes refers to forms of ‘on-demand’ services regardless of the technical means of transmission, such as cable video on- demand, and it is possible that viewers might conceive of catch up services (US MVPD video on demand) as streaming as well. More and more, however, ‘streaming’ refers to a particular kind of media service that is increasingly mainstream in music, movies and television. Key features of these services include the availability of subscription payment for on-demand access to a large media catalogue over internet protocols - though we cannot claim to be focusing on ‘subscription streaming services’ given that most music streaming occurs through free versions of these services. The streaming video ecosystem, too, is much broader than the focus here, and the industrial dynamics among various sectors make common claims difficult. In order to make this exploratory paper manageable, we will primarily concentrate on the two largest services in movie, television and music spheres: Netflix and Spotify.

The history of music technology is peppered with breakthroughs that freed people from constraints. Radio meant no longer physically having to go to a live performance if you wanted to listen to music. Vinyl gave people the option of playing records they owned rather than being stuck listening to whatever was on the radio at the time. Then cassettes made a personal music collection portable, so you could listen to it anywhere.

We’re now deep in the digital era, untangling recorded music’s historical link with physical media – and things are changing at speed. Just 15 years ago, people were excited to own a device the size of a deck of cards that stored a thousand songs. Now, countless pieces of internet-connected hardware offer instant access to tens of millions of music tracks.

Modern streaming services – TIDAL, Spotify, Apple Music *et al* – are the culmination of two decades of cutting-edge technological achievements, evolution and iteration. Their architects have unpacked and developed audio file formats, delivery platforms and smart algorithms, and sharp-eyed businesspeople have converted that vision into the subscription models we know today.

Spotify is a digital music, podcast, and video streaming service that gives you access to millions of songs and other content from artists all over the world. It offers streaming of over 30 million songs. The service was founded in 2006 by Daniel Ek and Martin Lorentzon. Spotify has around 140 million active registered users. There are two versions of Spotify: a premium monthly subscription service and a free service which is supported by advertising. It is available in more than 75 countries and is considered one of the best and most famous music streaming service in the world.

In today's millennial world, most people listen to their music on streaming platforms. There is a huge demand and fanbase for music. There are so many different genres and music to choose from and everyone has a different taste for music choices. This project has a great scope as it visualizes all the recent trends in music over the past decade based on the data collected on Spotify which is the leading music streaming service in the world today. It considers all the top ranking songs and artists based on year for it visualization. This analysis is very relevant and needed since all Machine learning based recommendation algorithms are based on the past trends that have followed in music releases.

Africa	Algeria, Egypt, Morocco, South Africa, Tunisia.
Asia	Bahrain, Hong Kong, India, Indonesia, Israel, Japan, Jordan, Kuwait, Lebanon, Malaysia, Oman, Palestine, Philippines, Qatar, Saudi Arabia, Singapore, Taiwan, Thailand, United Arab Emirates, Vietnam.
Europe	Andorra, Austria, Belgium, Bulgaria, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Monaco, Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Spain, Sweden, Switzerland, Turkey, United Kingdom.
North America	Canada, Costa Rica, Dominican Republic, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama, United States.
South America	Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Paraguay, Peru, Uruguay.
Oceania	Australia, New Zealand.

III. LITERATURE SURVEY/ RELATED WORK

Analyzing the Spotify Top 200 Through a Point Process Lens Michelangelo Harris,
Brian Liu, Cean Park, Ravi Ramireddy, Gloria Ren, Max Ren, Shangdi Yu, Andrew
Daw, and Jamol Pender∗ School of Operations Research & Information Engineering Cornell University, Ithaca, NY, October 4, 2019

Learning about Spotify the world's largest music streaming service, streaming audio is a great way of delivering sound without having to download files from the internet. This paper takes a closer and more specific venture into Spotify's popularity. It begins ab expedition through an analysis of the Spotify Top 200, a chart of the 200 most frequently streamed songs each day on Spotify. This investigation is built on a data set scraped from over 20 months of the United States based edition of these rankings, spanning from 2017 to 2018. As defined by Spotify, one day spans from 3:00 PM UTC through 2:59 PM UTC on the next. We have scraped this data set from the publicly available Spotify Top 200 charts, specifically from the U.S. based rankings. The code to do so is derived from open source code used to form a Kaggle data set containing the worldwide Top 200 rankings for all of 2017 [2]. We also learnt from this paper hoe quantify the duration of a pop hit.

Approaching Media Industries Comparatively: A Case Study of Streaming, Daniel Herbert, Amanda Lotz and Lee Marshall

This paper got us to understand the concept of streaming data. We understood what is the importance for streaming data in different multimedia platforms. Key features of these services include the availability of subscription payment for on-demand access to a large media catalogue over internet protocols - though we cannot claim to be focusing on
'subscription streaming services' given that most music streaming occurs through free versions of these services. This paper gave us the analysis of business practices related to streaming of audio. We learnt that suggesting that implications for each industry derive from pre-streaming norms at least as much as the common implications of streaming. There is notable consistency in the consumer experience across streaming media. The experience of these three industries in conversation revealed that the timeline of adjustment of different components such as consumer experience, business practices and content is occurring at different rates in each medium.

Spotify Teardown: Inside the Black Box of Streaming Music by Maria Eriksson,
Rasmus Fleischer, Anna Johansson, Pelle Snickars & Patrick Vonderau, The MIT Press, ISBN 978-0262038904 (Paperback)

In try-ing to explore the way in which Spotify is commonly perceived, the re-searchers analyse Spotify from every possible angle. Chapter 1 "Where Is Spotify" starts by questioning Spotify's corporate history, analysing it through the lens of its funding rounds. This approach proves helpful to understand the way the company was built from a service that relied on pirated material. The second chapter "When Do Files Become Music" explores the system's backstage describing its network infrastructure and the data gathering and sharing processes. One of the chapters also talk about how recommendations are fed to users based on changing circumstances. The final chapter talked about the financial tactics of the company.

Infinite content and interrupted listening The impact of smartphones, streaming and music 'superabundance' on everyday personal music listening behavior, Ellen Moore

Music is something that has been around for almost as long as we can remember. This paper however has a very unique way of understanding and interpretation done by the author. The intention of this research was to produce new insights into everyday listening and our relationship to music in the age of smartphones and the infinite content provided by streaming services. A study was conducted to see if the youth today can survive without the wonderful integration of music and technology. Many of the teens couldn't. And that automatically brings us to the importance of streaming audio.

Metrics and decision making in music streaming, Arnt Maaso & Anju Nyland Hagen

As we have seen, over the last few years, streaming services, and Spotify in particular, have made an increasing amount of data available to different stakeholders on a daily basis, and those stakeholders have enjoyed increasing flexibility to monitor and act based upon these Metrics. No single stakeholder has access to all relevant data points about music, but it is clear that the MSSs and major labels are dominating the metrics race.

These actors also have the resources and skillsets to interpret this data in ways that others do not. When metrics are presented, interpreted and analyzed upon by numerous stakeholders making decisions that is the fed back to algorithms that create consumer software.

Music in Streams: Communicating Music in the Streaming Paradigm Anja Nylund Hagen, Postdoctoral fellow, Dept. of Musicology, University of Oslo

This paper got us into the analysis of how streaming shapes music communication. Music can now be run on the smartphones on the go as well as randomly. How songs are marketed on these sites are also discussed artists are like brands for these companies.

Music Radio as a Format Remediated for Stream-based Music, Andreas Lenander Egidius

This paper drove into the concept of radio. Talked about how spotify and apple music technologies have taken it to a whole different level. Both providing users with endless new features signifying a huge evolution from the radio.

Musical trends and predictability of success in contemporary songs in and out of the top charts, (Royal society of Science)

This is paper is similar to what we will be doing for our project. We will analyze the given dataset that is the songs releases in a particular country. The influence of the musical characteristics were observed in how it managed to change the trends on the charts.

Changing Their Tune: How Consumers' Adoption of Online Streaming Affects Music Consumption and Discovery, Hannes Datta, George Knox, Bart J. Bronnenberg*

Talks about a different perspective of live streaming. In recent years, copyright-related industries have suffered as new digital technologies disrupted their revenue models. One such disruptive technology that is currently taking over the music industry is streaming. Streaming allows consumers unlimited access to a vast library of content at a fixed monthly payment.

The rise of streaming music and implications for music production, R. Scott Hiller∗
Jason Walter†‡, July 11, 2016

In this paper we observed that depending on the mounting popularity of the non-durable option of streaming. With a sufficiently low price and sufficiently high base of listening, the paid streaming model becomes an attractive option for consumers. No longer forced to pay for individual durable bundles, a subscription allows access to a much larger, diverse bundle.

IV. PROPOSED METHODOLOGY/ PROJECT DESCRIPTION

Firstly, a literature survey is being compiled where we have read several research papers on streaming platforms, Spotify as an application, recent music trends, recommendation algorithms, etc. We have built a problem statement based on the content that we found in these articles. Spotify has its own website where many research articles and blogs have been posted.

Secondly, we identify the most appropriate dataset for our requirements from dataset repository sites such as www.kaggle.com/datasets.

Thirdly, we use visualization tools such as Python to visualize various aspects of the dataset in the most suitable forms. Spotify Labs and various other online articles have given demo visualization techniques to portray such data. We will do a research on what people want to know about this field and try to visualize that.

Next, we analyse the data and devise conclusions from them about the trends of the past decade. In the end we summarize all the results that we found.

What do people want to know about the music trends over the last decade?

The artists, songs as well as the specifications of each song.

Why do people want to know about it?

To see which artist performed well and which songs were ranked highest in each year and also the type and genre of the songs which changed over the decade depending upon the music preferences of the audience.

How will we visualize the data?

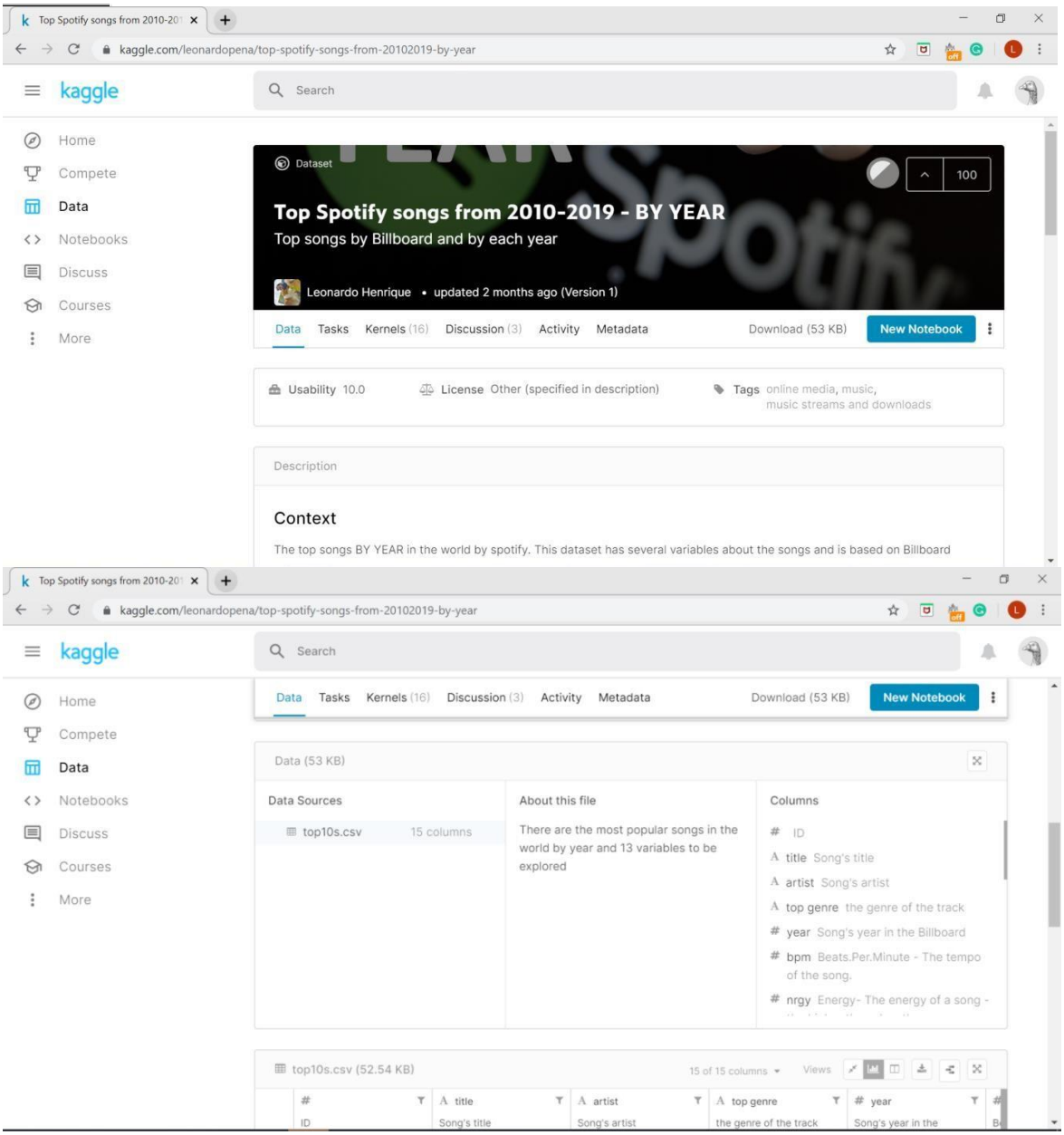
Using datasets available from Spotify and Kaggle.com and by using tools such as seaborn, matplotlib.pyplot and Scikit-Learn in Python to visualize the trends in various forms like scatter plots, area plots, bar charts, pie charts, line charts etc.

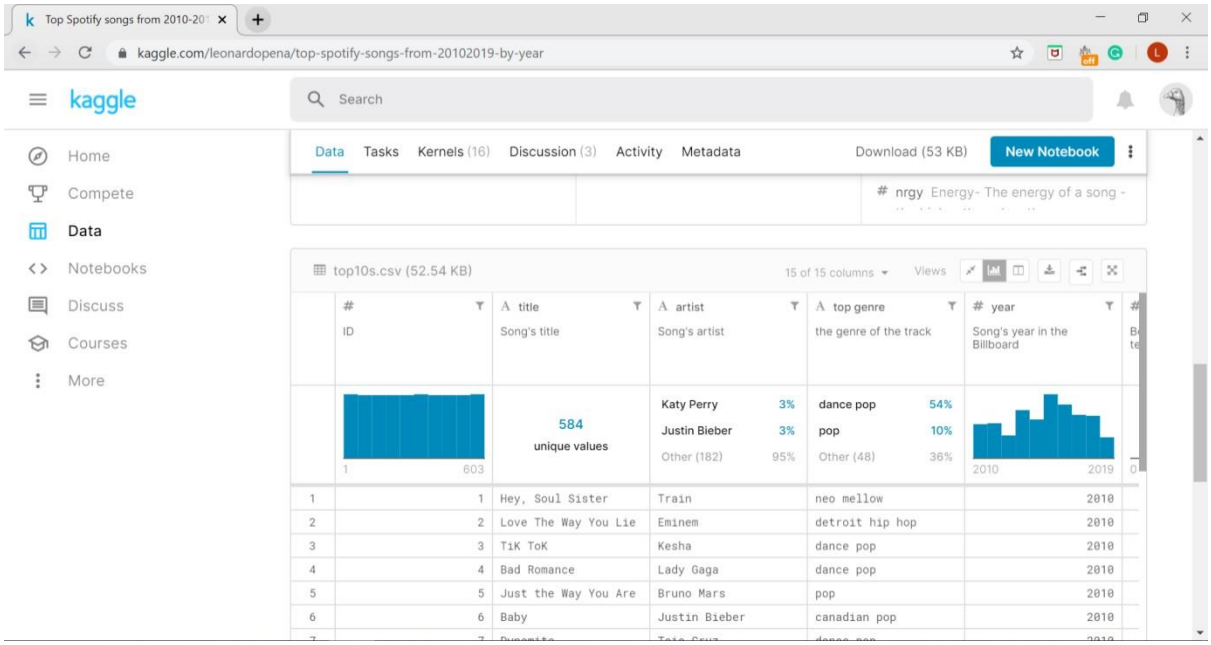
At Last, we use Machine Learning Algorithms such as K-Means Clustering using Scikit Learning in Python to make playlists of similar songs using our analysis. The main use of these analysis is in recommendation systems in music streaming apps.

V. DATASET

Our first task was to obtain a large set of songs. After a bit of research, we decided that the **Top Spotify songs from 2010-2019 – By Year (Top songs by the Billboard and by each year)** dataset would best suit our needs for this project since we could download an excel file of track names and artists which we could then use in Tableau. Our main challenge here was to select the right amount and range of songs. Data could become heavily skewed at this initial step. We wanted to sample as many genres as possible in order to create an accurate high-level view of musical changes over time. We also needed to ensure our data had a sufficient number of songs from all time periods for accurate analysis.

Getting the Dataset from kaggle.com/datasets?search=spotify





There is a total of 603 songs in the dataset.
It has the following columns which represent the different attributes of each song.

1. **# ID**
2. **. title: Song's** title [Categorical]
3. **artist: Song's artist [Categorical]**
4. **top genre: the genre of the track [Categorical]**
5. **# year: Song's year in the Billboard [Quantitative]**
6. **# bpm: Beats. Per. Minute - The tempo of the song. [Quantitative]**
7. **# nrgy: Energy- The energy of a song - the higher the value, the more energetic. Song [Quantitative]**
8. **# dnce: Danceability - The higher the value, the easier it is to dance to this song. [Quantitative]**
9. **# dB: Loudness..dB.. - The higher the value, the louder the song [Quantitative]**
10. **# live: Liveness - The higher the value, the more likely the song is a live recording [Quantitative]**
11. **# val: Valence - The higher the value, the more positive mood for the song. [Quantitative]**
12. **# dur: Length - The duration of the song. [Quantitative]**
13. **# acous: Acousticness. - The higher the value the more acoustic the song is. [Quantitative]**
14. **# spch: Speechiness** - The higher the value the more spoken word the song contains. [Quantitative]
15. **# pop: Popularity**- The higher the value the more popular the song is. [Quantitative]

VI. SCIKIT-LEARN IN PYTHON

Scikit-learn (formerly scikits. learn and also known as sklearn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms - 13 -

including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Scikit-learn is largely written in Python, and uses NumPy extensively for high-performance linear algebra and array operations. Furthermore, some core algorithms are written in Cython to improve performance. Support vector machines are implemented by a Cython wrapper around LIBSVM; logistic regression and linear support vector machines by a similar wrapper around LIBLINEAR. In such cases, extending these methods with Python may not be possible.

Scikit-learn integrates well with many other Python libraries, such as matplotlib and plotly for plotting, NumPy for array vectorization, Pandas data frames, SciPy, and many more.

VII. K-Means Clustering Algorithm

The K-Means algorithm clusters data by trying to separate samples in n groups of equal variances, minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified. It scales well to large number of samples and has been used across a large range of application areas in many different fields.

The k-means algorithm divides a set of N samples X into K disjoint clusters, each described by the mean of the samples in the cluster. The means are commonly called the cluster “centroids”; note that they are not, in general, points from, although they live in the same space.The K-means algorithm aims to choose centroids that minimise the inertia, or within cluster sum-of-squares criterion:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

K-means is often referred to as Lloyd’s algorithm. In basic terms, the algorithm has three steps. The first step chooses the initial centroids, with the most basic method being to choose samples from the dataset. After initialization, K-means consists of looping between the two other steps. The first step assigns each sample to its nearest centroid. The second step creates new centroids by taking the mean value of all of the samples assigned to each previous centroid. The difference between the old and the new centroids are computed and the algorithm repeats these last two steps until this value is less than a threshold. In other words, it repeats until the centroids do not move significantly.

VIII. IMPLEMENTATION

In [26]:
```
import seaborn as sns
import matplotlib.pyplot as plt
sns.set()
%matplotlib inline
```

In [27]:
```
spotify = pd.read_csv('top10s.csv', encoding='cp1252')
spotify.drop(spotify.columns[0], axis=1, inplace=True)
```

First, we import the necessary libraries for visualisation, and we read the dataset and save it as a data frame.

In [29]:
```
spotify.describe()
```

Out[29]:

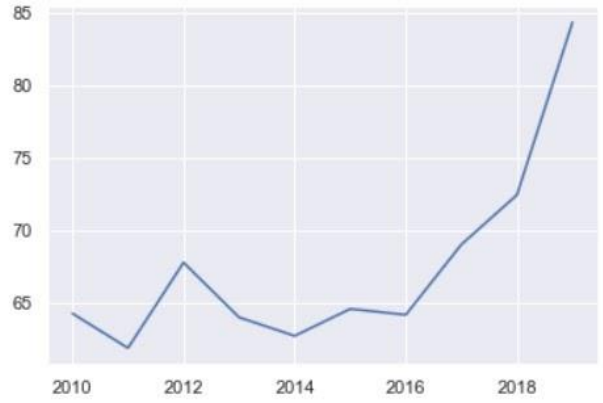
	year	bpm	nrgy	dnce	dB	live	val	
count	603.000000	603.000000	603.000000	603.000000	603.000000	603.000000	603.000000	60
mean	2014.592040	118.545605	70.504146	64.379768	-5.578773	17.774461	52.225539	22
std	2.607057	24.795358	16.310664	13.378718	2.798020	13.102543	22.513020	3
min	2010.000000	0.000000	0.000000	0.000000	-60.000000	0.000000	0.000000	13
25%	2013.000000	100.000000	61.000000	57.000000	-6.000000	9.000000	35.000000	20
50%	2015.000000	120.000000	74.000000	66.000000	-5.000000	12.000000	52.000000	22
75%	2017.000000	129.000000	82.000000	73.000000	-4.000000	24.000000	69.000000	23
max	2019.000000	206.000000	98.000000	97.000000	-2.000000	74.000000	98.000000	42

In [30]:
```
spotify.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 603 entries, 0 to 602
Data columns (total 14 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   title      603 non-null    object
 1   artist     603 non-null    object
 2   top genre  603 non-null    object
 3   year       603 non-null    int64
 4   bpm        603 non-null    int64
 5   nrgy       603 non-null    int64
 6   dnce       603 non-null    int64
 7   dB         603 non-null    int64
 8   live       603 non-null    int64
 9   val        603 non-null    int64
 10  dur        603 non-null    int64
 11  acous      603 non-null    int64
 12  spch       603 non-null    int64
 13  pop        603 non-null    int64
dtypes: int64(11), object(3)
memory usage: 66.1+ KB
```

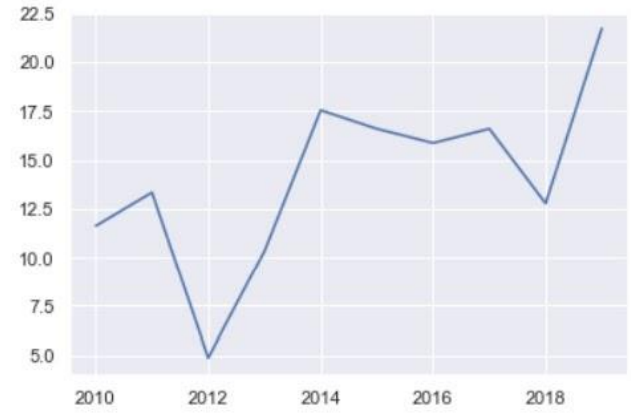
Before starting the project, we do some elementary data exploration to check the range, mean, standard deviation and count. We also check if the attributes are categorical or numerical and if there are any NULL values.

```
In [33]: sns.lineplot(data=by_year["pop"])
Out[33]: <matplotlib.axes._subplots.AxesSubplot at 0x1f2053288b0>
```



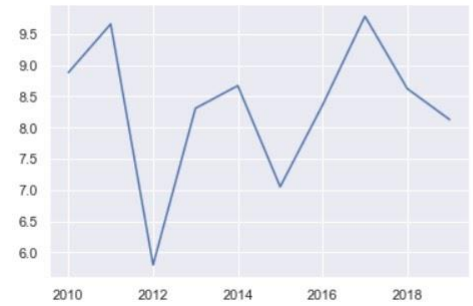
Here, we can see the trend of various popness of music that are present on Spotify. We can see from graph that from 2010 to 2011 there is decrease in their popularity but it increases from 2011 to 2012. Finally, at present the popularity of this online streaming platforms is getting more and more popular according to the graph.

```
In [34]: sns.lineplot(data=by_year["acous"])
Out[34]: <matplotlib.axes._subplots.AxesSubplot at 0x1f2054485e0>
```



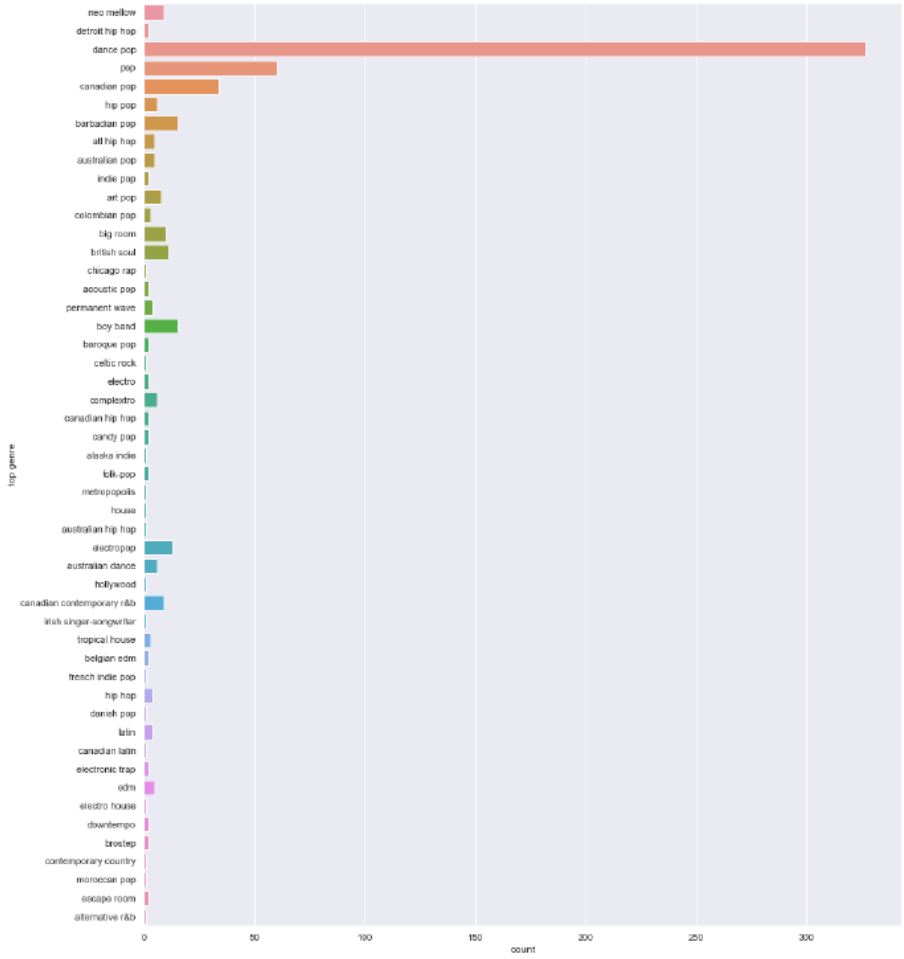
The given graph shows the acoustics of the songs that are present on Spotify. It shows the songs have become more or less acoustic according to their popularity. Like during 2010-2011 the acoustics in songs were increasing but as we move on to the period of 2011-2012 the acoustics of the songs drops drastically indicating that acoustic songs were less popular during that time. Currently the trend is leading towards more and more acoustics in the songs.

```
In [35]: sns.lineplot(data=by_year["spch"])
Out[35]: <matplotlib.axes._subplots.AxesSubplot at 0x1f20549cee0>
```



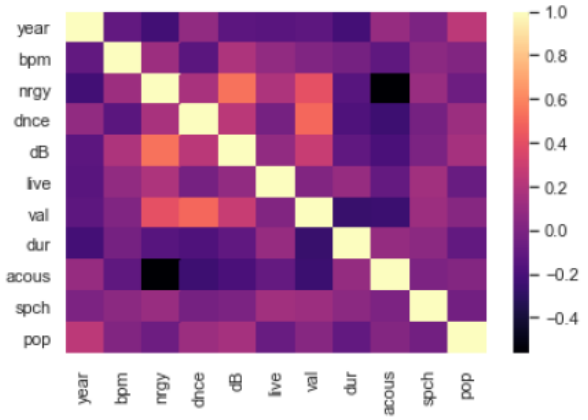
The given graph shows the speechiness of the songs present on Spotify. The songs had more speechiness during 2010-2011 and this reduces drastically during 2012 as this was the year of the popularity of EDM songs with no lyrics. This also reflects the popularity of the songs with respect to speechiness. At present the trend is going towards reducing the speechiness in the songs as songs with more music is getting popular.

```
In [55]: plt.figure(figsize=(16, 20))
         sns.countplot(y='top genre', data=spotify)
Out[55]: <matplotlib.axes._subplots.AxesSubplot at 0x1f20820f790>
```



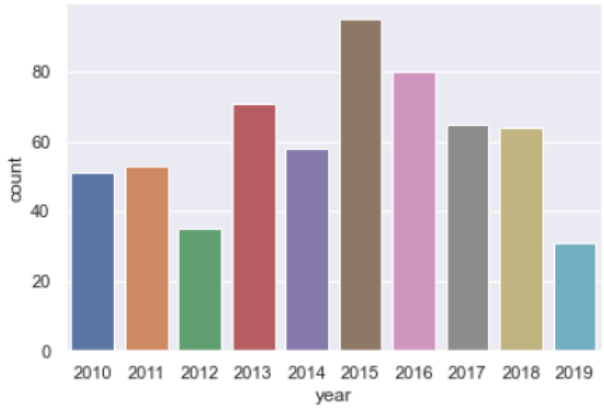
This graph plots various genres of songs present on Spotify and shows their relationship with the number of people who listen to that genre of songs. From the given graph we can clearly conclude that the most famous genre in recent times is the dance pop genre followed by pop genre songs.

```
In [52]: sns.heatmap(spotify.corr(), cmap='magma')
Out[52]: <matplotlib.axes._subplots.AxesSubplot at 0x1f207f7bd60>
```



The given heatmap shows the correlation between various aspects of the songs that we have considered in our dataset. We can see that the energy that a song has is related to the volume of the song in decibels, i.e., louder the song more energy the song has. We can also observe that the energy of a song has a high negative correlation with the its acoustics, i.e., the more acoustic the song the lesser energy it has. We also observe relation between the valence of a song and it's danceability i.e., the more danceable a song is, the more like it is to raise up your mood. There is also a slight relation between valence and energy, as there is a possibility that high energy songs boost up mood.

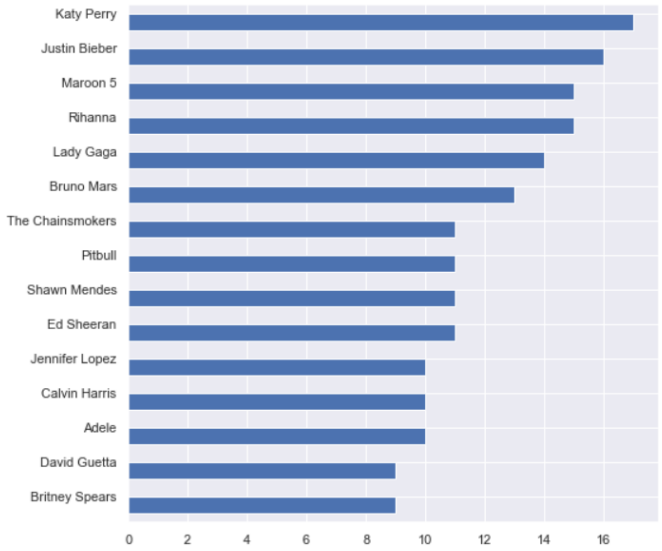
```
In [66]: sns.countplot(x='year', data=spotify)
Out[66]: <matplotlib.axes._subplots.AxesSubplot at 0x1f208247d30>
```



The given graph shows the number of new records that are being released on Spotify. From the graph we can see that in 2015 maximum records were released and in recent years this number has gone down drastically.

```
In [29]: df=spotify['artist'].value_counts()
         df=(df.head(15)).sort_values(ascending=True)
         plt.figure(figsize=(8, 8))
         df.plot(kind='barh',position=1)
```

```
Out[29]: <matplotlib.axes._subplots.AxesSubplot at 0x1dff875ddf0>
```



The given graph shows the 15 most Popular Artist in the last decade. From this we can see the popularity of different artists and can determine in the future that whose song would be most popular among them.

IX. Using K-Means Clustering for making recommendation playlists

The above analysis of the dataset can be used in making playlists of similar songs using machine learning techniques such as K-Means Clustering. Here we have used scikit-learn library of python for our implementation. Here we have created 3 playlists using K-Means clustering based on the 'popness' and 'acoustic' values of songs. These clusters automatically give us songs of similar types which can be placed in a single playlist. These recommended playlists ensure that people are more likely to listen to songs which are similar in vibe.

Data Pre-processing:

First, we imported our dataset as a csv file. Then we made our target dataset smaller by extracting all the songs from the year 2016. We have scaled the 'pop' and 'acous' values between 0 and 1 using MinMaxScaler to get a more accurate result in our algorithm.

```
#The necessary packages are imported.

from sklearn.cluster import KMeans
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
from matplotlib import pyplot as plt
%matplotlib inline
```

```
#Here we import our dataset as a csv file.
#We have extracted all the songs from the year 2016.

df = pd.read_csv("Downloads/top10s.csv")
df = df[df.year==2016]
df.head()
```

	Unnamed: 0	title	artist	top genre	year	bpm	nrgy	dnce	dB	live	val	dur	acous	spch	pop
363	364	The Hills	The Weeknd	canadian contemporary r&b	2016	113	56	58	-7	14	14	242	7	5	84
364	365	Love Yourself	Justin Bieber	canadian pop	2016	100	38	61	-10	28	52	234	84	44	83
365	366	Cake By The Ocean	DNCE	dance pop	2016	119	75	77	-5	4	90	219	15	5	81
366	367	Don't Let Me Down	The Chainsmokers	electropop	2016	160	87	53	-5	14	42	208	16	17	81
367	368	In the Name of Love	Martin Garrix	big room	2016	134	52	50	-6	45	17	196	11	4	81

```
#Here we scale our values for an efficient system.

sc = MinMaxScaler()
df['pop'] = sc.fit_transform(df[['pop']])
df['acous'] = sc.fit_transform(df[['acous']])
df[['pop','acous']]
```

	pop	acous
363	1.000000	0.070707
364	0.988095	0.848485
365	0.964286	0.151515
366	0.964286	0.161616
367	0.964286	0.111111
368	0.952381	0.020202
369	0.952381	0.202020

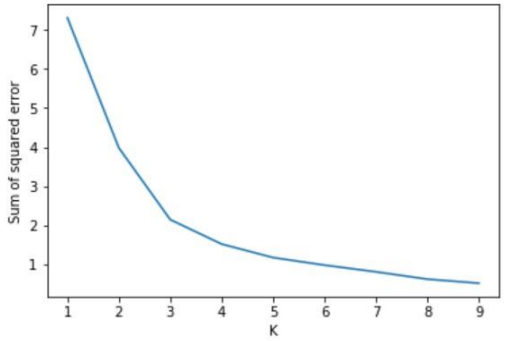
Evaluating the clusters using the Elbow Method:

```
#Here we find the Sum of Squared Error value for each value of k

k_rng = range(1,10)
sse = []
for k in k_rng:
    km = KMeans(n_clusters=k)
    km.fit(df[['pop','acous']])
    sse.append(km.inertia_)
sse
```

```
[7.308102474901041,
 3.9803750657789556,
 2.1442709346572597,
 1.5186851291378707,
 1.1731579774509067,
 0.9799718469440778,
 0.80813799018053811,
 0.62097598513049801,
 0.5180501938569434]
```

```
#We have plotted the sse values to see that at k=3 clusters, we get the optimal solution.

plt.xlabel('K')
plt.ylabel('Sum of squared error')
plt.plot(k_rng,sse)
```

```
[<matplotlib.lines.Line2D at 0x1ef5f8d6080>]
```



Applying the K-Means Algorithm:

```
#Here we use K-Means clustering algorithm on the dataset.

km = KMeans(n_clusters=3)
km
```

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
    n_clusters=3, n_init=10, n_jobs=1, precompute_distances='auto',
    random_state=None, tol=0.0001, verbose=0)
```

```
y_predicted = km.fit_predict(df[['pop','acous']])
y_predicted
```

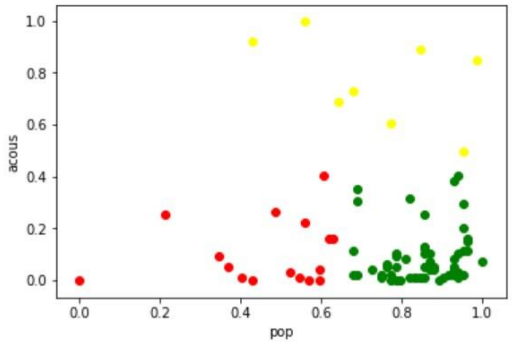
```
array([0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 2, 2, 2, 2, 2, 1,
       2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2])
```

```
df['cluster'] = y_predicted
df.head()
```

	Unnamed: 0	title	artist	top genre	year	bpm	nrgy	dnce	dB	live	val	dur	acous	spch	pop	cluster
363	364	The Hills	The Weeknd	canadian contemporary r&b	2016	113	56	58	-7	14	14	242	0.070707	5	1.000000	0
364	365	Love Yourself	Justin Bieber	canadian pop	2016	100	38	61	-10	28	52	234	0.848485	44	0.988095	1
365	366	Cake By The Ocean	DNCE	dance pop	2016	119	75	77	-5	4	90	219	0.151515	5	0.964286	0
366	367	Don't Let Me Down	The Chainsmokers	electropop	2016	160	87	53	-5	14	42	208	0.161616	17	0.964286	0
367	368	In the Name of Love	Martin Garrix	big room	2016	134	52	50	-6	45	17	196	0.111111	4	0.964286	0

```
df0=df[df.cluster==0]
df1=df[df.cluster==1]
df2=df[df.cluster==2]

plt.scatter(df0['pop'],df0['acous'],color='green')
plt.scatter(df1['pop'],df1['acous'],color='yellow')
plt.scatter(df2['pop'],df2['acous'],color='red')

plt.xlabel('pop')
plt.ylabel('acous')
```

```
Text(0,0.5,'acous')
```



Displaying the clusters as playlists:

```
#This is one of the clusters which can be a playlist.

df2[['title','pop','acous']]
```

	title	pop	acous
425	Make Me... (feat. G-Eazy)	0.630952	0.161616
426	Keeping Your Head Up	0.619048	0.161616
427	True Colors	0.607143	0.404040
428	Make Me Like You	0.595238	0.040404
429	Champagne Problems	0.595238	0.000000
430	Blown	0.571429	0.000000
432	Pep Rally	0.559524	0.222222
433	Higher	0.547619	0.010101
434	Invitation	0.523810	0.030303
435	One Call Away (feat. Tyga) - Remix	0.488095	0.262626
437	Little Lies	0.428571	0.000000
438	Do You Wanna Come Over?	0.404762	0.010101
439	BURNITUP!	0.369048	0.050505
440	Picky - Remix	0.345238	0.090909
441	Behind Your Back	0.214286	0.252525
442	Million Years Ago	0.000000	0.000000

X. CONCLUSION

Spotify is the largest music streaming service available. The company started in 2006 in a time when piracy caused considerable losses to the music industry. In January 2015 they had 60 million users in total of which 15 million premium users (1) and these numbers seem to be increasing. Spotify offers free streaming of music to its users, though one can purchase a premium membership for added benefits, such as no advertisements and being able to listen to music offline. The large number of users and content of Spotify create a large database of users and songs that users listened to that could hold interesting patterns and information for related companies, such as Spotify themselves, record companies or radio stations. The dataset in question has been provided by, so we first look for general applications of the data and then focus on possibilities that will also be useful to , but will also be interesting from a scientific perspective. By performing some statistics on the entire dataset, we try to determine the worth of the Spotify data for both and for scientific purposes. We will answer a few relatively simple questions regarding interesting patterns found in the data and try to formulate a good model that can be used with this data.

How to use these visualizations?

- Marketing strategies can be created based on the most popular artist and genres that are at the top. The company may ask themselves “which artist sells?”. Online advertisements can be created accordingly.
- Spotify can organize its target user-base as the time period and type of music that people like to listen to at that particular moment.
- Such a visualization and interpretation as well as data collection is of importance. Data is the most valuable asset in today’s day and age such data can be sold to music companies for large sum of money.

We also discovered that there was a transition of the kind of music over the years which included songs to be more upbeat and more EDM songs were getting popular over the later half of the decade. This shows that the music taste of the people is a constant changing process. Artists like Katy Perry, Maroon 5 and Justin Beiber were the major players in the game. The produced the most number of songs with a high success rate.

More and more DJ’s like Martin Garrix, David Guetta and Alan Walker were also becoming main stream and more popular. One of the major takeaways was also that the music industry was at its peak during 2015- 2016 in terms of quantity as well as the quality of songs.

The major application of all this analysis can be put in use for the machine learning algorithms of the Spotify API to create recommendation playlists for the users by mixing similar valued songs to each playlist to make it even more attractive and enjoyable for the users.

REFERENCES

[1] AEgidius, A. L. (2017). Music Radio as a Format Remediated for the Stream-Based Music Use.

Bonneville-Roussy, A., Rentfrow, P. J., & Xu, M. K. (2013). Music Through the Ages: Trends in Musical Engagement and Preferences From Adolescence Through Middle Adulthood.

[2] Datta, H., Knox, G., & Bronnenberg, B. J. (2017). Changing Their Tune: How Consumers' Adoption of Online Streaming Affects Music Consumption and Discovery.

[3] Eriksson, M., Fleischer, R., Johansson, A., Snickars, P., & Vonderau, P. (2019). *Spotify Teardown: Inside the Black Box of Streaming Music*.

[4] Hagen, A. N. (2020). Music in Streams: Communicating Music in the Streaming Paradigm.

[5] Harris, M., Liu, B., Park, C., Ramireddy, R., Ren, G., Ren, M., Yu, S., Daw, A., & Pender, J. (2019). Analyzing the Spotify Top 200 Through a Point Process Lens.

[6] Herbert, D., Lotz, A., & Marshall, L. (2018). Approaching media industries comparatively: A case study of streaming.

[7] Hiller, R. S., & Walter, J. (2015). The Rise of Streaming Music and Implications for Music Production.

[8] Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z., & Komarova, N. L. (2018). Musical trends and predictability of success in contemporary songs in and out of the top charts.

[9] Maasa, A., & Hagen, A. N. (2020). Metrics and decisions-making in music streaming.

[10] Menten, M., Ng, K., O'Rourke, T., & Holmes, R. B. (2018). Temporal Trends in Music Popularity - A Quantitative analysis of Spotify API data.

[11] Moore, E. (2019). Infinite content and interrupted listening: The impact of smartphones, streaming and music 'superabundance' on everyday personal music listening behaviour.

[12] Razlogova, E. (2013). The past and the future of music listening between freedom DJs and recommendation algorithms.

[13] Sital, R. (2019). The Daily Use of Music Streaming Platforms by the Dutch Millennial Age Group.

[14] www.spotify.com

[15] https://www.kaggle.com/datasets?search=spotify

[16] https://en.wikipedia.org/wiki/Spotify

[17] https://scikit-learn.org/