

Faculty of Computing and Informatics (FCI)

Multimedia University

Cyberjaya

**TDS 2101 - INTRODUCTION TO DATA SCIENCE**

**Topic: Billboard Top 100 On Spotify**

**Semester 2, 2019/2020**

Submitted To : Mr. John See Su Yang

Tutorial Section: TT02

| **Student Name** | **Student ID** |
| --- | --- |
| Lou Jia Yu | 1161104266 |
| Perivitta Rajendran | 1171101579 |

TABLE OF CONTENTS

| **TITLE** | **PAGE** |
| --- | --- |
| Introduction | 2 |
| Datasets | 3-5 |
| Question 1 | 6-8 |
| Question 1.2 | 9-11 |
| Question 1.3 | 12 |
| Question 2 | 13-20 |
| Question 2.1 | 21-22 |
| Question 3 | 23-24 |
| Challenges Faced | 25 |
| Conclusion | 26 |
| References | 27 |

**Introduction**

Spotify is currently the world’s [most valuable music company](https://www.npr.org/sections/therecord/2018/04/04/599385111/spotify-is-for-now-the-worlds-most-valuable-music-company), and for good reason. Now worth around $25 billion, the company has had a major impact on both the popularity of music streaming and the way the music industry uses the data these streaming services generate in impactful ways.

An obvious way to measure music's popularity is through the charts. [Billboard](https://www.billboard.com/) has been charting the "Hot 100" songs weekly since 1958. Nowadays, chart rankings are based on a combination of sales both physical and digital, radio play, and online streaming. Billboard states that the approximate weighting is 35-45% for sales, 30-40% for airplay, and 20-30% for streaming. An annual "year-end" chart is also released, ranking that year's songs by calculating a cumulative total of sales, airplay, and streaming. As for our datasets, we have decided to use readings from 1958 to 2019.

But that is not the only way we can see which songs are popular. Over 140 million people worldwide are active users of Spotify, with around half of those using the paid Premium service. The Swedish streaming service has a vast catalogue of 20 million songs. Each year, Spotify releases a playlist of the 100 most streamed songs for that specific year. We can see that there are similarities, but that the lists are far from identical. We can visualize the differences in the rankings using a data visualization method.

One of the most prominent ways Spotify uses the data generated by their customers is to help generate content that each user will consider in-line with their specific tastes. Although Spotify approaches this process from a variety of angles, the overarching goal is to provide a music-listening experience that is unique to each user, and that will inspire them to continue listening and discovering new music that they will be engaged with week after week. This is accomplished through the use of artificial intelligence and machine learning algorithms.

**Data Sets**

There are two datasets named HotStuff.csv and Hot100AudioFeatures.csv included in the zip file. Those datasets are retrieved from Data.World

<https://data.world/kcmillersean/billboard-hot-100-1958-2017>.

Data.World provides data scientists and machine learners a platform to find and publish datasets. Not only that, but it also provides an opportunity to join competitions and solve data science problems by interacting with more talented people.

**HotStuff.csv**

This dataset contains every weekly Hot 100 singles chart from [Billboard.com](http://billboard.com/). Each row of data represents a song and the corresponding position on that week's chart. Included in each row are the following elements:

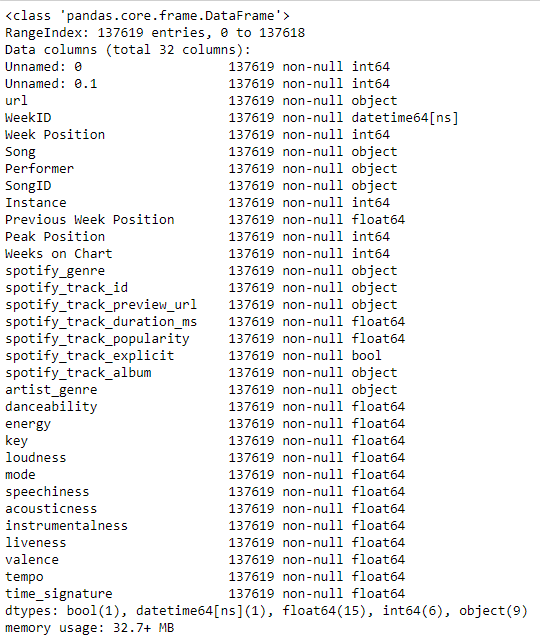
1. Billboard Chart URL
2. WeekID
3. Song name
4. Performer name
5. SongID - Concatenation of song & performer
6. The current week on the chart
7. Instance (this is used to separate breaks on the chart for a given song. Example, an instance of 6 tells you that this is the sixth time this song has appeared on the chart)
8. Previous week position
9. Peak Position (as of the corresponding week)
10. Weeks on Chart (as of the corresponding week)

**Hot100AudioFeatures.csv**

This dataset contains the songs on Spotify. Each column of data contains each song’s features itself and each song’s features in Spotify. Included in each row are the following elements:

1. Performer
2. Song
3. Spotify Genre
4. Spotify Track ID
5. Spotify Track Preview URL
6. Spotify Track Duration ms
7. Spotify Track Popularity
8. Spotify Track Explicit
9. Spotify Track Album
10. Artist Genre
11. Danceability
12. Energy
13. Key
14. Loudness
15. Mode
16. Speechiness
17. Acousticness
18. Instrumentalness
19. Liveness
20. Valence
21. Tempo
22. Time Signature

**Data Types**



*Figure 1.0 - Data types of cleaneddata.csv*

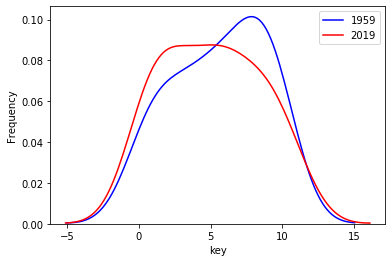
**Question 1**

How music trends influence a song from the 60's to 90’s ?

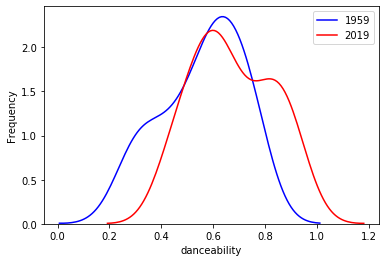
A feature might pertain to a particular musical passage, a movement, work, composer, period, genre, style or culture. Features may be salient because of intratextual or intertextual treatments including prevalence, primacy, recency, evocation, quotation, allusion, parody, and model. Not only that, the sound of music has evolved with society over the years, thus the innovation in sound is really a reflection of our cultural and technological progression. Music that is relevant to the mainstream population now would have been impossible a few years back due to the technological means. Since the musical industry is in an advanced state, the streaming services have made it easier than ever to consume music across various platforms. Like, for example, spotify has also lowered the barriers to entry into the music industry.Below is the music features by Spotify for the audio features settings.

* **Acousticness:** A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
* **Danceability:** 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. A value of 0.0 is least danceable and 1.0 is most danceable.
* **Energy:** Energy is a measure from 0.0 to 1.0 and 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. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
* **Instrumentalness:** Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
* **Liveness**: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides a strong likelihood that the track is live.
* **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. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.
* **Speechiness:** Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audiobook, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
* **Valence:** A measure from 0.0 to 1.0 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).
* **Tempo:** The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
* **Key:** The key the track is in. Integers map to pitches using standard Pitch Class notation .

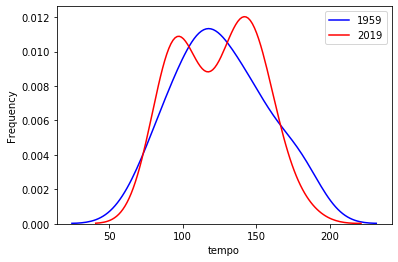
Below are the patterns that we have discovered using the audio datasets.There’s a huge difference between music patterns from 60s to 90s onwards.



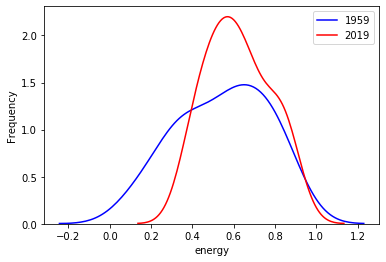
*Figure 3.1 - 2019 has a lower rate of frequency for the key pitch compared to 1959.*



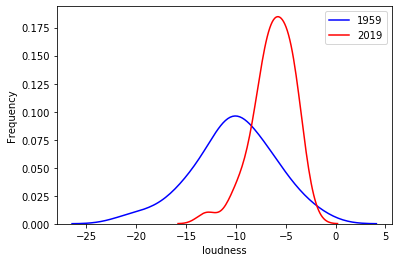
*Figure 3.2 - 2019 has a lower rate of frequency for danceability compared to 1959.*



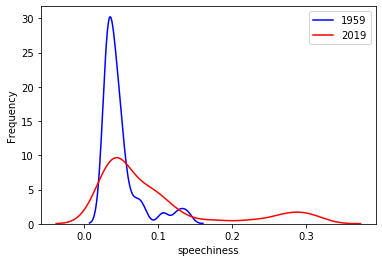
*Figure 3.3 - 2019 has a higher rate of frequency for tempo compared to 1959.*

**

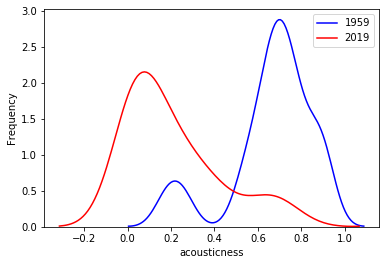
*Figure 3.4 - 2019 has a higher rate of frequency for energy compared to 1959.*

**

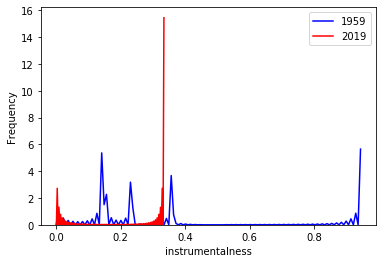
*Figure 3.5 - 2019 has a higher rate of frequency for loudness compared to 1959.*

**

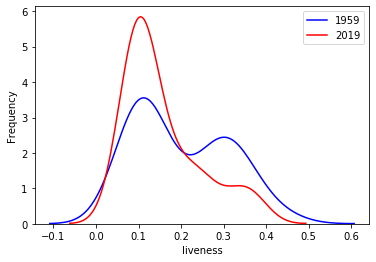
*Figure 3.6 - 2019 has a lower rate of frequency for speechiness compared to 1959.*

**

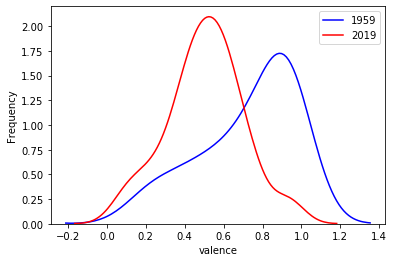
*Figure 3.7 - 2019 has a lower rate of frequency for acousticness compared to 1959.*

**

*Figure 3.8 - 2019 has a higher rate of frequency for instrumentalness compared to 1959.*

**

*Figure 3.9 - 2019 has a higher rate of frequency for liveness compared to 1959.*

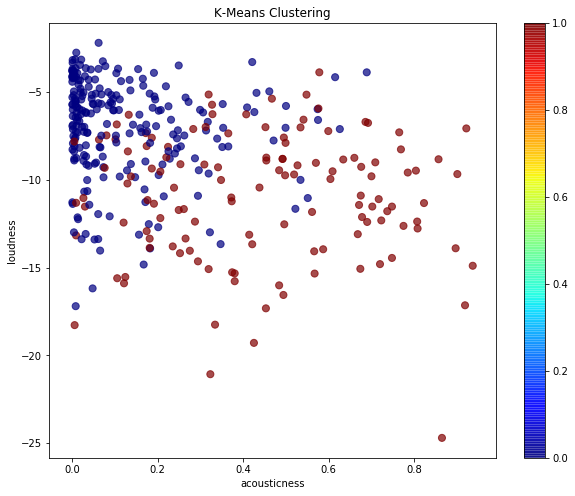
**

*Figure 3.9.1 - 2019 has a higher rate of frequency for valence compared to 1959.*

**Question 1.2**

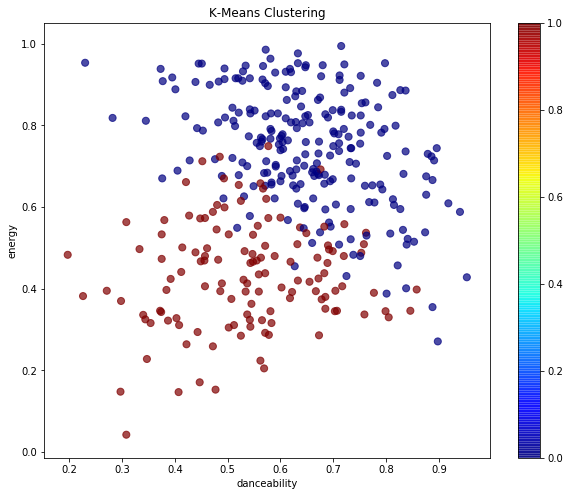
K-means clustering for Audio Features

In short, [K-Means Clustering](https://www.quora.com/What-is-the-k-Means-algorithm-and-how-does-it-work) is a technique that categorizes data based on the mean characteristics of each data point. I first absorbed the more obscure genres into the larger ones. We have decided to analyze the audio features of acousticness and loudness, below is the output from the K-means clustering:

******

*Figure 3.2.1 - Acousticness vs loudness.*

Loudness, in [acoustics](https://www.britannica.com/science/acoustics), attribute of [sound](https://www.britannica.com/science/sound-physics) that determines the intensity of auditory sensation produced. The loudness of sound as perceived by human ears is roughly proportional to the logarithm of sound intensity, when the intensity is very small, the sound is not audible and when it is too great, it becomes painful and dangerous to the [ear](https://www.britannica.com/science/ear). The [sound intensity](https://www.britannica.com/science/sound-intensity) that the ear can tolerate is approximately 1012 times greater than the amount that is just perceptible. This range varies from person to person and with the [frequency](https://www.britannica.com/science/frequency-physics) of the sound. Figure above shows the result from selected months within the year 2019. We can see that the stability is high at the top most axis within the two variables.



*Figure 3.2.2 - Danceability vs Energy*

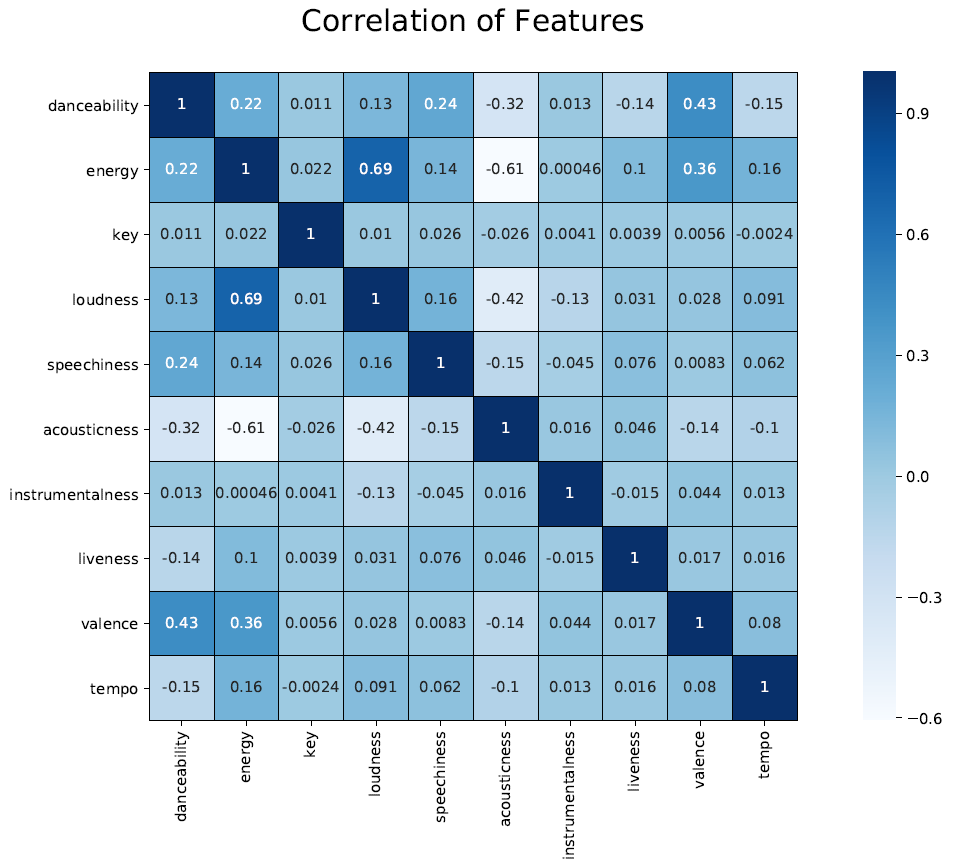
As for the figure above, we all know that energy is a measure of a track's “intensity.” This might seem similar to danceability, but a song can be energetic (fast, loud, noisy), but undanceable due to the audio features. The K-mean clustering shows that there is less correlation among the variables.There’s only correlation among the variables from the range of 0.4 - 0.6 energy, we can assume that, the danceability depends on the energy level of the audio.

**Question 1.3**

What kind of Genre classification is preferred by music listeners ?

Over the past few years, streaming services with huge catalogs have become the primary means through which most people listen to their favorite music. But at the same time, the sheer amount of music on offer means users might be a bit overwhelmed when trying to look for newer music that suits their tastes.

For this reason, streaming services have looked into means of categorizing music to allow for personalized recommendations. One method involves direct analysis of the raw audio information in a given song, scoring the raw data on a variety of metrics. Our goal is to look through this dataset and classify songs and we will learn how to clean our data, do some exploratory data visualization, and use feature reduction towards the goals.A song is about more than its title, artist, and number of listens. We have another dataset that has musical features of each track such as danceability,energy and moreover. Below is the result of heatmap from our prediction using Hot100AudioFeatures.csv dataset.

******

*Figure 1.1 - Prediction based on Genres*

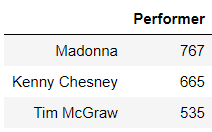
As from what we have examined,there are some interesting insights to point out in this plot. Specifically, there are associations between valence (the happiness of a track) and the energy of a track. Therefore, although there is not necessarily causality between the two variables, a track that is happier is predicted to have more energy. Moreover, there is also a strong association between dancebality and valence seen, which indicates that within the dataset songs that are happier usually tend to be more danceable. These associations are not too large, however, as both values are only around 0.43 so predictions are made with extreme caution. One association that is high, but negatively correlated, is the relation between energy and acoustic (how strong the acoustic is in the song). That means in the data set the more energy there is in the song tends to predict a lower acoustic rating and vice versa.

Not only that, it also can be seen that, there is a strong association between loudness and energy, which indicates that within the dataset songs that are loud and energetic is preferred by music listeners.This heatmap accentuates a correlation between how long someone is on the charts and what their peak rank was. In essence, this data point demonstrates that peak rank sometimes predicts how long a song is on the charts and could be a point of further analysis.Creating a heatmap from a joined table that includes the peak rank does not uncover any new findings, but does open up consideration of what characteristics help create a song the top song.

**Question 2**

Who are the most successful performers with the highest amount of composed songs?

After collecting the data and cleaning it, we then moved with data exploration by looking into feature importance, trends in our dataset, and identifying the differences between top performers and the amount of songs released by each artist. Before making changes,we decided to check which artist has composed most songs from the overall datasets. Below is the result of the top 3 artists who have composed the most songs from the year 1958 to 2019.



*Figure 2.0 - The top 3 Performers who have composed the most songs.*

As for our datasets, we decided to separate the bar plots into two categories, meaning we have split the bar graphs among top 90’s and 2000’s. By using this method, we would be able to find how long the artist has been famous for.

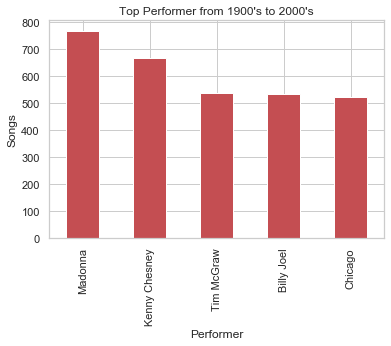
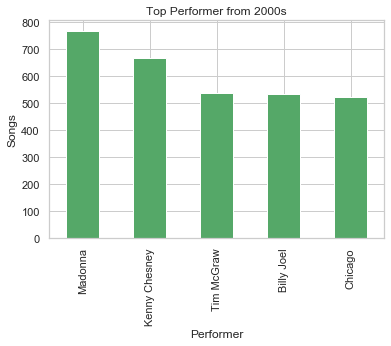


Figure 2.1 - Total *songs composed from the year 1900s - 2000s*

Figure 2.1 shows that the highest ranking goes to Madonna, because she has composed the most songs compared to any other performer in the dataset. But it also can be seen that, there is a sudden dropout of Madonna’s performance after 2000s, and Kenny Chesney took over the first position as shown in Figure 2.2.

**

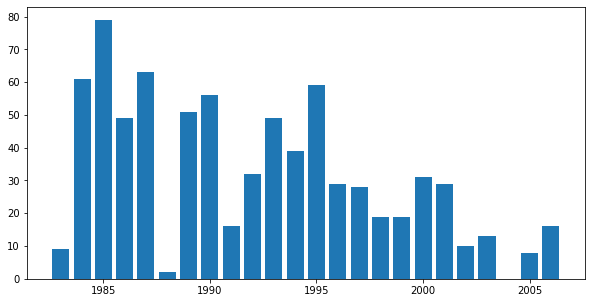
*Figure 2.2 - Total songs composed from the year 2000s - 2019*

Based on the figure 2.2, Kenny Chesney had composed most songs, which is about 583 songs from the year 2000-2019. Secondly, Kelly Clarkson and Jason Aldean have a closer range of values with the difference of 11 additional songs by Kelly Clarkson. That makes Kelly Clarkson to be in second place and the remaining performers are slightly close to each other in terms of the rankings.

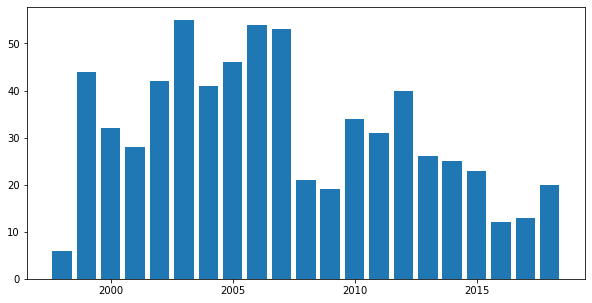
**Question 2.1**

Investigation on Madonna’s drop out

After further data analysis, we have tried to investigate why Madonna started dropping out and here’s the bar graph of Madonna’s composed songs.

******

*Figure 2.3 - Count of overall Madonna’s songs*

**

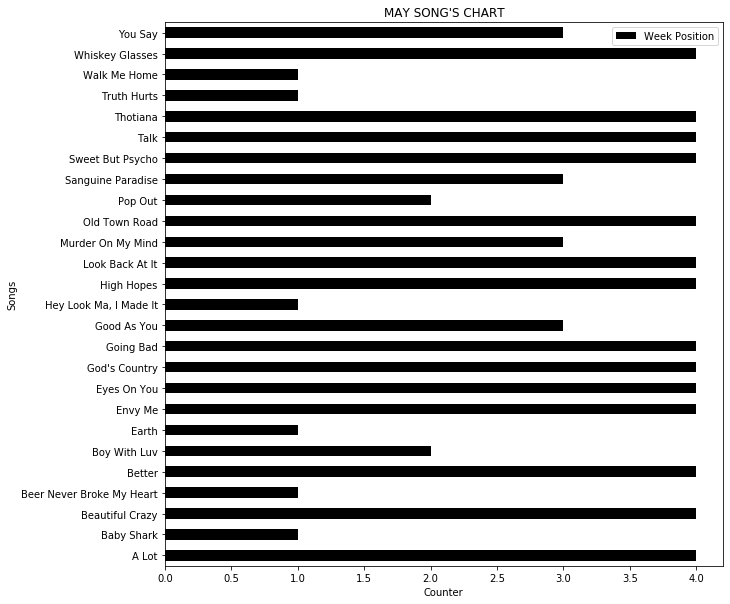
*Figure 2.4 - Count of overall Kenny Chesney’s songs*

From figure 2.3, it is shown that Madonna’s songs have reached the billboard up to year 2005.The highest peak where Madonna’s song made it to the billboard hit list was in the year 1985. It is also shown that Madonna’s songs did not make it to the list in 2004. Since Madonna is famous for pop songs, we can assume that pop songs are less preferred by listeners as the year increases. In addition, we also see that Kenny Chesney’s song started becoming famous throughout the billboard in the year 2003.After comparing between the first two rankers Kenny Chesney’s and Madonna, we see the changes between song genres and how diversion from pop to rock songs.

**Question 3**

Prediction of monthly chart based on Billboard chart rankings

According to the dataset, we had analyzed a few songs from the monthly chart. As for our prediction, we have selected songs from the May and June chart. After getting the results of the monthly charts, we decided to transfer it into a graph form so that the differences can be spotted. Below is the bar graph for both May and June song charts.



*Figure 3.1 - Song count based on the May Chart.*



*Figure 3.2 - Song count based on the June Chart.*

Based on the charts on figure 3.1 and 3.2, we have made some predictions for the following month (July). Since, there are some songs reappearing twice under both May and June’s chart, so those songs have high possibilities to get into the billboard chart for July.

For example, songs like “Whisky Glasses”, “Talk”, “Sweet but Psycho” with the maximum amount of count have high chances to be included in July’s chart. Moreover, there are also some songs that have been dropped out from the list and new songs have been added to the list, as day counts.

Therefore, we predicted songs like “Going Bad”, “God’s Country”, and “Envy Me” will be in the top ranking in July.

**Challenges Faced**

1. We had confusion on choosing the most suitable method to answer the questions.
2. At the beginning we had difficulties identifying where maximum data quality errors occur so that you can assess the root cause and design the plan according to that.
3. Identify any duplicates and validate the accuracy of the data as this will save a lot of time during analysis. Tracking all the cleaning operations performed on the data is very important so that you repeat or remove any operations as necessary. We had to keep checking our datasets for errors.
4. Handling huge datasets was quite difficult, due to large memory, the anaconda software had crashed multiple times and there was some missing data when we tried using the Google collaboration Jupyter.
5. At some point, we managed to figure out solutions and we could think of multiple ideas but some of it was way too detailed and complicated and we did not continue with it, instead we went with methods that we could truly understand and code using the basic programming knowledge.
6. As for the K-means, we had to go through some troubles for plotting the graphs due to the genres. Based on our datasets, the genres were listed as string data type, but for K-means it does not accept any string values, so we had to choose a different option, where we had to audio features to visualize K-means clustering.

**Conclusion**

Going into this endeavour, we were uncertain if it is even possible to predict, better than random, if a song will be popular or not. After testing our model on new songs pulling from Spotify billboard chart datasets, we observed that it is significantly simpler to correctly predict a bad song rather than a hit. It may have been easier to predict known hit songs if our datasets were skewed, with only a few hit songs, unfortunately, we did not use smaller datasets because we thought it would make a huge change to the results. Moving forward, we would like to explore how additional features such as artist location or release date can influence a song’s popularity. In addition, using full datasets will make a better prediction than using skewed data.

Alongside improving customer experiences, Spotify is able to use the massive amount of data generated by its users to inform its own ad campaigns and better target consumers. At the most basic level, this is done by reviewing what they’ve learned about their listeners and using those insights to develop ads that strategically target their ideal audience. In general, now that streaming far outranks music purchases, the industry has had to shift its focus from record sales to the collection of this kind of data in order to decipher how the public is responding to an artist, album, or song. Since this data also provides a deeper insight into listening trends, audience markets, and more, it is hopefully a sustainable change for those within the field. Either way, trends predict that Spotify will continue to be one of the largest music data sources for some time, and that data will continue to make for better business decisions across industries.

**References**

* Spotify: Big Data Shows Big Results. (2019, October 7). Retrieved from

https://www.northeastern.edu/graduate/blog/spotify-big-data/

* Spotify Audio Analysis. (n.d.). Retrieved from :

https://spotify-audio-analysis.glitch.me/analysis.html

* Apple Music Unveils Free Data Analytics Tool for Artists, Managers. (2019, August 8). Retrieved from:

https://www.billboard.com/articles/business/streaming/8526884/apple-music-data-analytics-tool-artists

* Ashrith. (2019, March 22). What Makes A Song Likeable? Retrieved from https://towardsdatascience.com/what-makes-a-song-likeable-dbfdb7abe404
* Displayr. (2020). Visualized: Can we Quantify the Most Popular Music? | Displayr. [online] Available at: https://www.displayr.com/most-popular-music/