**Analyzing Top Spotify Tracks of 2017**

Capstone Project – Milestone Project

**The Project**

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. They currently have over 30 million songs and over 140 million active users in 61 countries. At the end of each year, Spotify compiles a playlist of the top 100 songs streamed most often over the course of that year. My goal is to do an intensive analysis of the audio features of these songs to determine if there are any similarities in them. Once we have this information, would it be possible to predict if a song might make it in the top 10 or top 100 songs list?

**The Client**

With physical music sales declining year after year, the entire music industry is transitioning to a more digital music experience. Music streaming services are becoming more popular than ever and only showing signs of growth. The data in this analysis would be valuable for any music steaming service (Spotify, Apple Music, YouTube, Amazon Music, Pandora, Soundcloud, etc.) to help them better determine what songs to include in their more popular suggested playlists and email blasts. This could also help record labels and artists choose the right single to release off an album.

**The Data**

I will be using the ‘Top Tracks of 2017’ Spotify playlist as a reference for this analysis. I will also be using the Spotify Web API along with a Python specific API called Spotipy to extract all the available audio features Spotify uses to categorize these songs, such as danceability, tempo, key, etc.

**Link to the ‘Top Tracks of 2017’ playlist:**

https://open.spotify.com/user/spotify/playlist/37i9dQZF1DX5nwnRMcdReF?si=HXOSfrHJQGyky0-DP9xQoA

**The Approach**

After extracting and cleaning the necessary data, I will do a full analysis of the Spotify 2017 Top 100 Tracks list and try to find any similarities in audio features among these songs. Once we have some indication of the characteristics of these songs, I’d like to test some recently released songs and see if we can predict if they might be in the top 100 tracks list for this year based on a comparison of their audio features and my findings.

**Libraries**

numpy for:

array data structure, the primary input for classifiers

model comparison, matrix manipulation

pandas for:

data loading, wrangling, cleaning, and manipulation

feature selection and engineering

descriptive statistics

matplotlib for:

data visualization

seaborn for:

drawing attractive and informative statistical graphics

examining [relationships](https://seaborn.pydata.org/examples/scatter_bubbles.html#scatter-bubbles) between [multiple variables](https://seaborn.pydata.org/examples/faceted_lineplot.html#faceted-lineplot)

Automatic estimation and plotting of [linear regression](https://seaborn.pydata.org/examples/anscombes_quartet.html#anscombes-quartet) models

scipy for:

Numerical algorithms and domain-specific toolboxes

signal processing, optimization, statistics

statsmodels for:

Statistical models, Linear regression

### Data Wrangling

For the Data Wrangling process, I used Spotify to find the playlist I will be using for this analysis. I used the Spotify API and their documentation to understand all the data features available for download from each song, artist, & playlist. Once I determined the specific data and information I wanted to use for my analysis, I used a few YouTube videos, Spotipy, Sublime Text, & the Mac Terminal to prepare the necessary code to download the data I wanted to download for this analysis.

The way that the Spotify API is setup to download data, it was necessary to download this data in the following steps:

1) Using the Sublime Text editor & help from a few YouTube tutorials on how to use Spotipy, I wrote code & downloaded the artist(s) name, track name, track id #, & album name information as a JSON file. I then transferred this JSON data into a CSV file. Included with this data were links to the specific webpages for each song, artist(s), & album.

2) In order to download the audio features of each song, it was necessary to get the track id #’s for each song from the playlist data I downloaded in the 1st step above. Then I used similar code as in step 1 to download the audio features of each song as a JSON file. I then transferred this JSON data into a CSV file.

3) Since the data had to be downloaded as 2 different CSV files, my next step was to open these 2 files and combine them into 1 CSV file. Once I was able to combine these, it was time to inspect our data and decide which columns & rows may need to be rearranged and/or removed.

**Cleaning**

After close inspection of this data, I was able to determine that we had no missing data in any of the columns or rows. I then arranged the columns in the data in an order that made more sense and would make it easier to analyze. I also deleted any columns that had webpage links since this data is not necessary for this analysis.

A few of the songs from the playlist feature guest artists on certain songs. When a song has a featured artist, Spotify lists a separate row for each featured artist along with their artist profile links and other data specific to that artist. To clean up our data, I added featured artists after the song title in parenthesis and deleted these artist specific rows.

There were no missing values in this data. However, as I mentioned in the cleaning steps above, if a song features a guest artist there were separate rows for each featured artist along with their artist profile links and other data specific to each artist. I did remove this extra data and rows.

There were no outliers in this data, so I was finally able to begin my analysis.

**Adding Inferential Statistics**

The first bit of Inferential Statistics I used in my data were 3 separate heatmaps. These Heatmaps gave me a better understanding of a few of my data points which were closely correlated. ‘Energy’ & Loudness’ were the closest correlated variables in my data. The correlation was also not very high, just a bit above average.

The tests I used to measure the level of correlation are the Pearsons, Spearman, and Kendall tests. For these tests, I used the ‘Energy’ & ‘Loudness’ columns as they had shown the closest correlation in some Heatmap plots I did. After analyzing these tests, the Spearman Test had the closest correlation significance with an R-Value of 0.778 & a P-Value of 1.828. Since our P-Value is so large, it signifies that there is a non-significant result.

I also applied some List-wise deletion & Pair-wise deletion tests but found no significant discoveries in my data using these tests. The results were actually very comparable to the Heatmaps I had previously tested.

The final statistical testing I tried was the T-Test. Our T-Score was pretty low, coming in at 4.1224. A small t-score tells you that the groups are similar. After comparing the t statistic with the critical t value (computed internally) we get a low p value of 0.0013. A low P-Value is good as they indicate your data did not occur by chance. Thus, it proves that the mean of the two distributions are different and statistically significant.

After close examination of my data I was not able to find a Dependent Variable. More specifically, there is not one variable that determines if a song will make it in the Top 100 songs of the year chart. There are multiple variables that determine this. Since my data contains more than 2 Independent Variables, this is a multiple logistic regression model.

# **Apply Data Storytelling**

Something interesting I was able to count from this data was how many artists on our list have multiple songs in the top 100 songs list.

### Top 10 Artists With Most Songs on the Top 100 songs of 2017 list

Drake 5

Post Malone 5

Kendrick Lamar 5

Ed Sheeran 3

Lil Uzi Vert 3

The Chainsmokers 3

The Weeknd 3

21 Savage 3

Migos 3

Khalid 2

After a quick analysis of the top 100 tracks of 2017, there are clearly a few musical attributes that help define a top song. Low Acousticness, Liveliness, & Speechiness along with High Danceability & Energy are a few of the obvious results. The clearest indicators are Time Signature of 4, Instrumentallness of 0 (which means songs have lyrics), & a high modality, which means the song is in a Major Key.

Comparing Danceability & Energy, there are a few similarities in their histogram charts. Also, their mean values of this top 100 songs list are rather close to each other as can be seen below.

("Mean value for 'Danceability':", 0.7335499999999999)

("Mean value for 'Energy':", 0.59642)

We are not able to make a time series plot of this data as there are no dates to reference in any of this data. Song duration would be the closest thing to a time-series plot we may be able to get here.

According to the heatmap I plotted of the top 100 songs, the closest correlations in this data are Energy & Loudness. I also did a heatmap of the top 10 & top 5 songs on this list and was surprised to see that in the top 5 songs only, Valence & the songs duration in miliseconds has the most significant correlation.

Having found some clear indicators and the mean averages of all the audio features, I wonder if there is a way to predict if a song has the potential to be in the top 10 or top 100 songs list?

After the above exploration of the data, a few of the mean values for the top 100, top 10, & top 5 songs show some clear indicators for the following:   
  
- Time Signature of 4.0   
- Tempo of about 120   
- Mode of 1 (Which means the song is in a Major Modality)   
- Most Popular key / pitch value of 1 ('C♯,D♭')   
- Mean song duration of 221,063.55 miliseconds (which is just over 3 & a half minutes)

- Mean value of remaining audio features:

Danceability 0.733550   
Energy 0.596420   
Loudness -6.378080   
Speechiness 0.141210   
Acousticness 0.165205   
Liveness 0.149470   
Valence 0.473075

I think the most interesting thing I’ve discovered from this analysis is the method Spotify uses to put a numeric value on the audio features they use to define each song. I'm most surprised that the song's genre is not a part of their audio features.

In this analysis, there are a few audio features that immediately became apparent as the clear leaders with all the data leaning towards one direction or in the case of the songs time signature, which was almost unanimous what the most used time signature was. But there are a lot of other audio features that are much more varied in their results per song. Artists and song titles might also be a factor which I had not immediately considered to analyze. I think I may need to do some type of analysis on words used in song titles to see if there are any words that are most common in these songs.

My initial quest with this analysis was to find some trends and correlations between these audio features to try and build a model to predict if a song could stand out as a possible contender to be included in next year’s top 100 songs list, but at this current point I'm thinking it may not be possible with the available data provided.