**Midterm Project proposal**

Aoyi Li

**Personal Statement**

There are two reasons that I chose the music related data. Firstly, I love music and considered to get a triple major in music performance during my undergraduate years. However, there was a lot more requirements when you decide to obtain another degree in other schools, forcing me to give up on this idea. The only thing I can do to compensate for my regret is to do more things related to music and to get close to music fields. Secondly, I had an internship in music department of the world top artists agency company called Creative artists agency (CAA). I really like the working environment in this company, and my career goal is to be a data scientist in new media or art and music related fields. Thus, I chose to analyze the Top Spotify songs from 2010-2019, which stores top songs by billboard and by each year.

**Data Description**

**Title:** Song’s title

**Artists:** Artist’s name

**Top genre:** The genre of the track

**Year:** Song’s year in the Billboard

**Bpm:** Beats.Per.Minute - The tempo of the song

**Nrgy:** Energy - The energy of the a song - the higher the value, the more energetic the song is.

**Dnce:** Danceability - The higher the value, the easier it is to dance to this song.

**dB:** Loudness..dB.. - The higher the value, the louder the song is.

**Live:** Liveness - The higher the value, the more likely the song is a live recording

**Val:** Valence - The higher the value, the more positive mood for the song.

**Dur:** Length - The duration of the song.

**Acous:** Acousticness - The higher the value, the more acoustic the song is.

**Spch:** Speechiness - The higher the value, the more spoken word the song contains.

**Pop:** Popularity - The higher the value, the more popular the song is.

**Question**

I plan to use the data to analyze the popularity for different types of songs, finding what factors will significantly related to the popularity scores. Since there are several ways to group the data, for example, it can be grouped by genre and year, I can use multilevel model to do the regression analysis. I think my focus will be the effect of genre, and I might do some trend analysis based on year. After the analysis, I plan to come up with a suggestion to guide the market or songwriting directions.

**The Data Source**

I found “Top Spotify songs from 2010-2019 - BY YEAR” from Kaggle Open Datasets website.

<https://www.kaggle.com/leonardopena/top-spotify-songs-from-20102019-by-year>

**Proposed Timeline of work**

I plan to use the first week starting at Nov. 9th to do the EDA and Data Processing part, cleaning the data, getting some simple relationships between variables and visualizing the data. If I found a way to combine 2 data files together, I Then, I will take the next week to do the modeling and validation part. In order to get a final model, I will check the original model, doing transformation and variable selection. Lastly, I will use the last week or at least last 5 days to make some changes and updates, then write a report.