Final Project Proposal - Kelly Zeng

For my STAT410 Final Project, I would like to look into data on songs and why they are more popular. More specifically, what attributes of songs lead them to be more popular. This topic interests me because listening to music is something I love to do, and I have always wondered what makes a song good (or popular). I also think it would be interesting if the most popular songs today have commonalities, which may lead song producers to utilize similar attributes.

In order to answer this question, I will be using a dataset from Kaggle posted by user Hanna Yukhymenko, who found the data from Spotify. This dataset includes the top 50 songs of the year 2021 on the music streaming platform Spotify. It includes 14 variables (or attributes) that characterize each song. The 14 attributes are popularity (the higher the value, the more popular the song is), danceability (the higher the value, the easier it is to dance to this song), energy (the energy of a song - the higher the value, the more energetic the song), key (the key the track is in), loudness (the higher the value, the louder the song), mode (major is represented by 1 and minor is 0), speechiness (the higher the value the more spoken word the song contains), acousticness (the higher the value, the more acoustic the song is), instrumentalness (the closer the value is to 1.0, the more instrumental the song is), liveness (the higher the value, the more likely the song is a live recording), valence (the higher the value, the more positive mood for the song), tempo (the overall estimated tempo in beats per minute), duration (how long the song is in ms), and time signature (an estimated time signature).

I will use multiple linear regression for this dataset by looking at how all other variables affect the popularity variable. This will allow me to see what attributes correlate with more popular songs and how they correlate, which is my research question. I will also use prediction,

so I can predict how popular a song will be given the values of the other attributes. This will help me see what values of the attributes can create, theoretically, the most popular song. Since I do not know if this data holds under the Gaussian assumption, I could also use bootstrapping to further calculate confidence intervals and hypothesis tests. I might also use ANOVA to see if the addition of certain variables affect the popularity of the song.