Music Data

*Anne, Stian, Marwa, Michael*

**Project Description**

**What is the dataset you'll be working with?  Please include background on who collected the data, where you accessed it, and any additional information we should know about how this data came to be.**

The dataset we will be working with the billboard dataset as well as Spotify and Apple Music. They allow us to search through albums, artists and tracks. Since we have access to this data by country, it will be possible for us to organize a visual that can show most popular song/artist/album etc. Spotify and Apple Music collect this data for customer visualization - to see who’s listening to what, and we will be able to access it through .csv files from multiple years which contain popular song data.

**Who is your target audience?  Depending on the domain of your data, there may be a variety of audiences interested in using the dataset.  You should home in on one of these audiences.**

We are targeting post-millennials since the majority of streaming service users are under 34 and the majority of trending artists are very young as well.

**What does your audience want to learn from your data?  Please list out at least 3 specific questions that your project will answer for your audience.**

1. What is the most popular song in \_\_\_\_\_? (geographically)
2. What is the most popular artist in \_\_\_\_\_? (geographically
3. What song is most popular across all regions? (geographically)

**Technical Description**

**How will you be reading in your data (i.e., are you using an API, or is it a static .csv/.json file)?**

API with .json files

**What types of data-wrangling (reshaping, reformatting, etc.) will you need to do to your data?**

We will need to combine data of both the popular songs along with the corresponding countries they are most popular in. In addition, with the .csv/.json files we are using, we may need to parse through them in order to attain the information that we want.

**What (major/new) libraries will be using in this project (no need to list common libraries that are used in many projects such as dplyr)**

We will probably be using the decrypt library since the databases and files we are intaking will be very large. In addition, we will be utilizing ggplot2 because we need to create visualizations with maps to represent our data set.

**What questions, if any, will you be answering with statistical analysis/machine learning?**

Our project doesn’t really utilize machine learning since it is a lot of hard-coded set data. The statistical analysis we will be doing is the popularity of different songs and seeing the cultural differences within different regions all over the world. We will be able to see the most popular genres across countries. For example, it’ll be interesting to see if the United States and China have similar song interests even when they are in opposite parts of the world.

**What major challenges do you anticipate?**

Creating the visualization of the map with the correct songs. We need to display all of the popular songs and make sure that it highlights the correct country that it corresponds to. I also think another challenge would be the time crunch and hopefully we are able to at least finish our MVP. We also need to be able to find open-source databases that we can utilize to attain the data of the most popular songs.

**Project Set-up**

<https://github.com/annepham25/music-data>