**Final Report**

**Mairead Lavery and Swaribha Garg**

**Extraction:**

We used a dataset from Kaggle on the Billboard Top 100 songs every year between the years of 1964 and 2015. This dataset was in an excel file (csv) and contained the title, artist, fun song lyrics, and year for each top 100 song. We pulled this data into a data frame and used the song title and artist name to make API calls to [last.fm](http://last.fm) to get additional information about each song genre. This was pulled in in JSON format. We also attempted to pull additional information from the music-story API, however, their API has a 2000 pull limit and our original data set has 5100 songs so we were unable to pull all of the information.

**Transformation:**

From our original Kaggle data set, we transformed the csv file into a pandas dataframe. We then removed the song lyric and source column as it was irrelevant in our analysis. Then we stored the song titles and artists in a dictionary in order to loop through the API calls to [last.fm](http://last.fm). We converted the JSON responses into columns that we then added to our dataframe. If the API call did not return values we were looking for, we stored a “0” in its place. We then used this in our final transformation; we dropped all rows containing a “0” in the “playcount” columns since we were unable to pull the necessary data for these songs. This left us with our final cleaned up pandas dataframe, that we then converted into a dictionary to load into mongoDB.

**Loading:**

We chose to make a mongoDB because our data was already stored in a pandas dataframe, which feeds easily into a mongoDB table. We named the collection “songs\_info” in the “songs\_db” that we created.