# Provisional Database

**Extract**

CSV files
API data

**Transform**

Python
Pandas

**Load**

Postgres

**Store**

AWS

**Dashboard**

Heroku

# Machine Learning Model

| Random Oversampling + SMOTE | Linear Regression | Decision Tree | Neural Network | K-Nearest Neighbors |
|---|---|---|---|---|
| Use oversampling techniques to tackle imbalance of data. Imbalance is caused by Spotify dataset of 600,000 tracks. We do not know which tracks were /are on the top 100 billboard yet so, we will choose over_sampling methods in anticipation of there being an abundance of non-top 100 songs. | 1. Quick to compute and can be updated easily with new data.<br><br>2. Relatively easy to understand | 1. A single decision tree is fast to train<br><br>2. Robust to noise and missing values<br><br>3. RF performs well “out-of-the-box” | 1. Can learn very complex relationships<br><br>2. Extremely powerful for many domains<br><br>3. Hidden layers reduce need for feature engineering | 1. Simple<br><br>2. Powerful<br><br>3. No training involved |

According to https://srnghn.medium.com/machine-learning-trying-to-predict-a-numerical-value-8aafb9ad4d36 these are the most common ML models for predicting a numerical outcome. So for the purpose of education it will be good to employ all four to see what could be the best outcome. We believe Decision Tree and Neural Network are going to work best.

# Dashboard

An interactive chart based on popularity per genre. click on a genre. Potentially see the breakdown of the average audio features in that genre.

Line graph noting the rise of K-pop by year

Are there certain artists or genres that occur more often?

Using the Spotify song ‘part’ breakdown to analyze KPOP songs on the Hot 100 and predict how long a song will stay on the Hot 100.

Comparison of a song’s audio features and how long a song stays on the Billboard’s Top 100. Are songs with the highest danceability more likely to stay on the Billboard Top 100 longer?