Summary of Findings and Website Review

# **Task**

Building off of project 3, in which we analyzed song attributes by decade for the most popular songs from 1950 to 2009, we wanted to use machine learning to see if we could predict what decade a song came from based on its attributes.

# **Data**

We initially ran supervised learning on the ~500 songs we pulled in our first project, but the sample size was too small to have any accuracy in predicting songs. We therefore had to execute many more cUrl commands to pull additional songs to try to improve the precision of the model.

Spotify had a limitation of 100 songs per pull, so as a time saving measure, we each did several pulls, which brough us to just over 4000 (4,113) songs to analyze.

Unfortunately, the playlists we pulled from had a lot of songs in common, so we arrived at just under 2000 songs. Then after dropping songs that did not have scores for all attributes, we arrived at our final count of 1810 songs from 6 decades.

In the future, we would recommend pulling out any remastered songs or even covered songs, to ensure songs (and their attributes) are correctly listed by its original release year and decade.

# **Supervised Learning**

We then re-ran the models. Accuracy was still very low, with precision being <.34 for any given training decade. Then on the test data, the 1970 decade was the only decade with decent precision of .67, but overall, the model was not very predictive of decade, having an overall average accuracy of .27.

From this, our conclusion was that an unsupervised machine learning would be best suited for this data.

# **Unsupervised Leaning**

We decided to use Principal Component Analysis so the model could determine the most predictive attributes from songs to cluster and hopefully determine (correctly) what decade they likely came from.

After removing the irrelevant columns from the data, we scaled the data. We utilized the elbow method and arrived at 4 clusters. With 4 PCA components, we can explain about 64% of the variability of the data, using data with Eigen values greater than 1.

In reviewing the by decade attributes from our website, we saw danceability and energy as two factors that seemed to be distinct over the decades in a linear fashion, and we hypothesized that using these as two factors would have high correlation by decade and result in 6 distinct clusters (on for each decade)—however, that was not the case. Instead, by using two components and 4 clusters, we found a good fit for the songs. By hovering over each data point, we found the data was mostly clustered by songs in similar genres rather than close in release decade.

# **Website Usage**

## Data Updates

Using our 1,810 songs, we updated the “by Decades” attribute graphs, which really cemented the trends we previously saw by decade.

Additionally, we updated the song slider, which now plays new songs every time you refresh the page and samples the song. This gives the user a sense of what these attributes and their values translate to.

## New Tableau Page

On this new page, we uploaded two tableau items. The first if a gif file which shows how important energy, loudness, and acousticness. The dt when sorted by release year has some visible clusters, but is overall difficult to see true trends. But when you slim it down to the average of these attributes by release year (the second frame of the gif), the trends start to become clearer, and then which simplified to a line graph with average by decade (the third frame of the gif), it is the most clear. The last frame of the gif is when we added in linear trend lines, and added their trend line equations with r-squared values.

In short, what this tells us is that there is a ton of variability in a given attribute and that only once you take the average of averages (average of the attribute by year and then by decade) do you have a good trend line. This helps explain why our supervised learning had such little predictability or accuracy.

We also included a dashboard from tableau in which we tried different trend line equations on these attributes to see how it would impact the r-squared value. Overall, it seems like a 3rd degree or 4th degree polynomial are the best fits for these attributes, but given that some of the trend lines would then go in very different paths in the future (like for energy), you can see that using any 1 factor could have little to no predictive accuracy for where songs are going for just one attribute.

## New PCA Page

We also added a 3D visualization of the PCA analysis along 3 components PC1, PC2, and PC3. Users can change the clusters to look at. By zooming in or shifting the view, users can see how the clusters group well (at k=3, 4, or 5) or unwell (6+). They also can hover over individual points to see what cluster they fall in, what the song is, and the decade it was released in.

In a quick scan of k=4 clusters, you can see that most of these songs are clustered by a general genre, but when you increase the clusters, it becomes increasingly nuanced for how the model clusters the songs, but it is still not by decade.

# **Key Takeaways**

1. If you were able to pull in a lot of data (1000+ songs per decade) and analyze within one or two specific genres that have had significant changes in attributes over time (example: country, which historically was slower and acoustic, to today which is heavily pop influenced), you may be able to create a model that can predict what decade a song came from. But our model as is has little accuracy in estimating a given song’s decade.
2. It is interesting to look at what songs are similar based on their cluster of the Principal Component Analysis model. For example:
   1. same cluster: Tiny Dancer, Elton John -- 1970 and Mr. Brightside, The Killers -- 2000
   2. different cluster: Fire & Rain, James Taylor, both in 1970
3. Tableau is a great tool in terms of comparing attributes of songs over time, and how much variability there is even in the same year for popular songs.