Kodi Obika
Eryk Pecyna

Math 23C Final Project Handout

The dataset that we explored is one provided by RStudio via the "billboard" package. It contains both technical and sonic information concerning songs that charted on the Billboard Hot 100 from 1960 to 2015. From the complete dataset, we extracted the pieces of information concerning each song that we were particularly curious about, such as the song's title, artist, year that it landed on the chart, key, mode, time signature, duration, tempo, explicitness, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, and valence, as measured by Spotify. We also added a column to indicate chart position, resulting in a dataframe with five categorical columns, one logical column, and nine numeric columns.

To start off, we displayed some simple graphics, using both base R and ggplot. In doing so, we found that Billboard charting songs are more than twice as likely to be written in the major mode than the minor mode, and C and G are the most popular keys (irrespective of mode). When taking mode into account, we saw that the nine most popular keys are all major keys, and C and G still reign supreme, with charting songs primarily being written in C major and G major (with close contenders being D major, C#/Db major, and A major), while the most popular minor keys are B minor and F minor. Additionally, it seems that D#/Eb minor, D minor, and G#/Ab minor tend to be avoided. In addition, we used a histogram to find that most songs tend to fall in the 0.6 to 0.7 range for danceability, and we saw that it could most accurately be modeled by a normal distribution centered at the mean danceability (approximately 0.619). We also found that the 95% confidence interval for mean danceability is (0.532, 0.706). We also saw that the average tempo of Billboard charting songs, despite repeatedly rising and falling, has generally gone up over time. Lastly, we created a contingency table showing the association between explicitness and mode.

In regards to analysis, for the contingency table, we used a chi-square test to find statistically significant evidence (p-value of 0) that explicitness and mode are not independent. Additionally, we used a permutation test to find statistically significant evidence (p-value of 0) that there is a difference in valence between songs written in the major mode and those written in the minor mode. The permutation test also showed us that there is not statistically significant evidence at the 5% significance level (p-value of 0.06) of a difference in valence between songs that are explicit and songs that are not.

We further explored the data by plotting a logistic regression for loudness and explicitness. Afterwards, we used the projection matrix approach to find linear regression and used it to see how energy, loudness, and speechiness correlate against chart position, and it turns out that there is essentially 0 correlation for all three. Then we attempted to reconstruct the data via principal components analysis, and found that reconstructing from only the largest eigenvalue yields a good reconstruction of tempo, but not much else. We also found the average correlation between each factor and found that energy has the highest average correlation with the other factors. With that information in mind, we decided to use our linear regression function to find the correlation between energy and a variety of other factors (danceability, loudness, speechiness, acousticness, instrumentalness, livenses, valence, and tempo), and we found that loudness seems to have the highest correlation with energy, while instrumentalness has the lowest.