35-315 Final Project Report

Kyra Balenzano, Evan Feder, David Yuan

5/2/2020

Introduction

The "Spotify: All Time Top 2000 Mega Dataset" is a dataset from Kaggle that contains various audio statistics and ratings of the top 1,994 songs on Spotify. For each song, it includes information such as the Title, Artist, Top Genre, Year of release, BPM (beats per minute), and Duration. The first three are nominal categorical variables, Year is an ordinal categorical variable, and the last two are quantitative variables. In addition, for each song, this dataset includes various quantitative ratings, such as those measuring its level of Energy, Danceability, Loudness, Liveness, Valence, Acousticness, Speechiness, and Popularity.

We added additional Genre, Decade, and Decade Range columns so that we could cluster songs into fewer groups, which in turn make visualizations more clear. The Genre column was created by manually sorting the 149 unique Top Genres into six main categories (Rock, Pop, Country, Hip Hop, Indie, and Other). *[ADDRESS BIAS?]* Meanwhile, the Decade and Decade Range columns were able to be assembled programmatically.

Using this dataset, we will answer three main questions in our report:

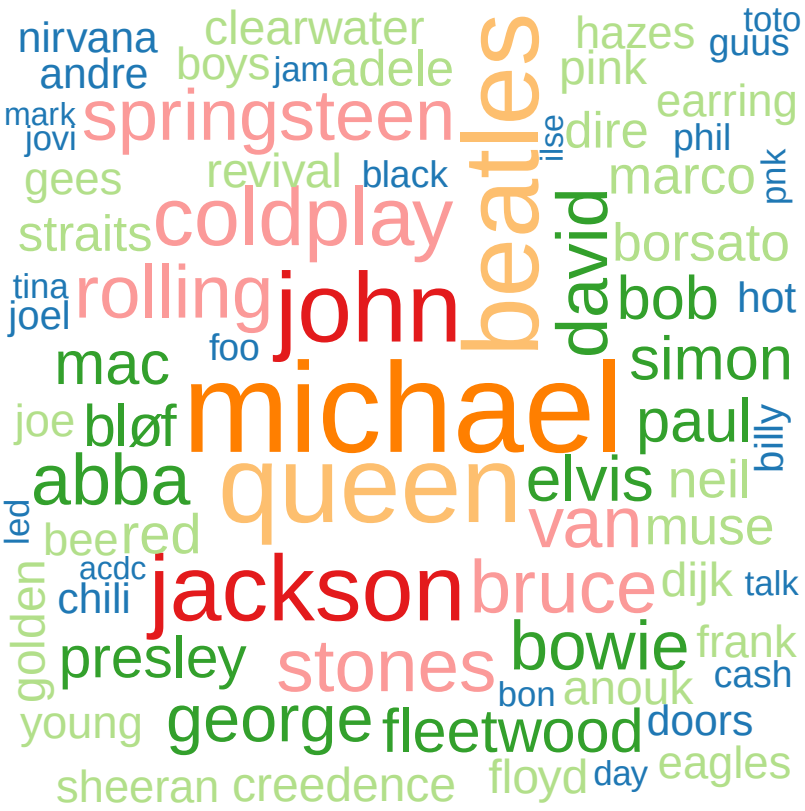
- Which attributes are associated with which genres?

- Which attributes do popular songs tend to have?
- How have the attributes of songs that appear on the top 2000 list changed over time?

Analysis

[Remove before handing in] David

[Remove before handing in] Evan

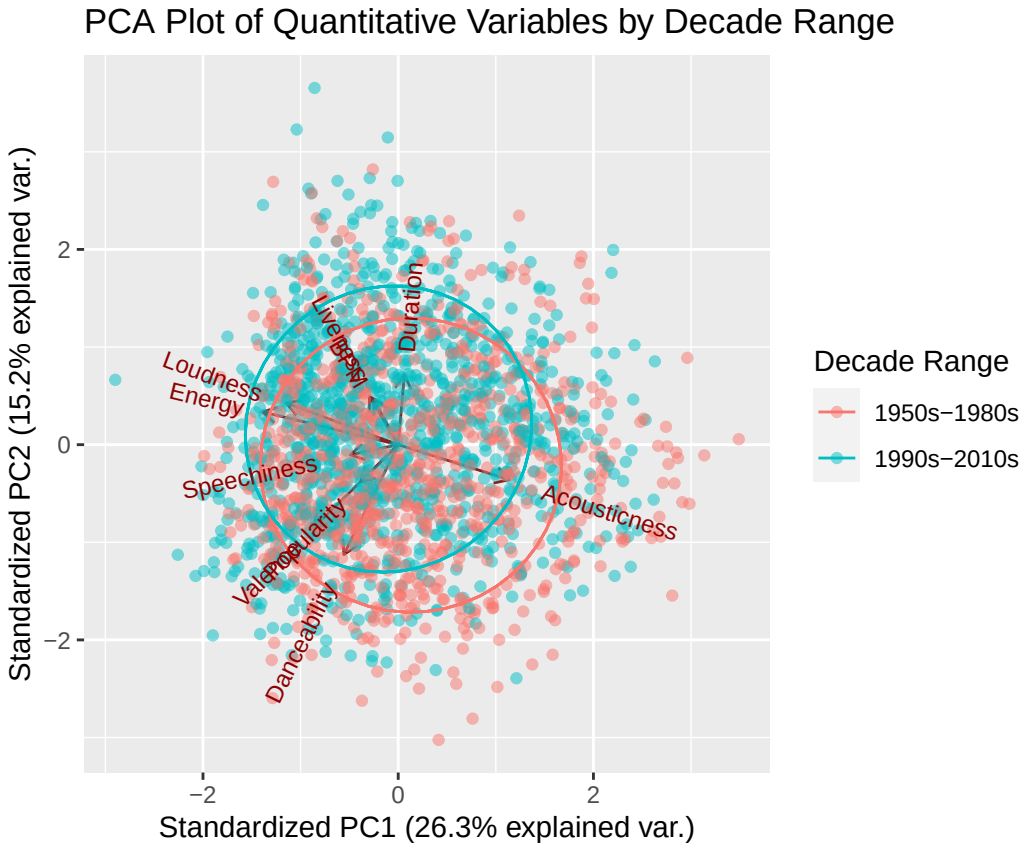


Above word cloud answers the question: which artists are the most popular? / Do particular artists make songs that are more popular?

[Remove before handing in] Kyra

Since we have so many quantitative variables, we first tried to condense them into a couple of dimensions and see if there were any changes over those. We performed this dimension reduction using principal component analysis (PCA), plotted the first two dimensions of this

result, and then colored the datapoints by the Decade Range variable so that we could make some comparisons regarding time without clouding the graph with too many overlapping colors.

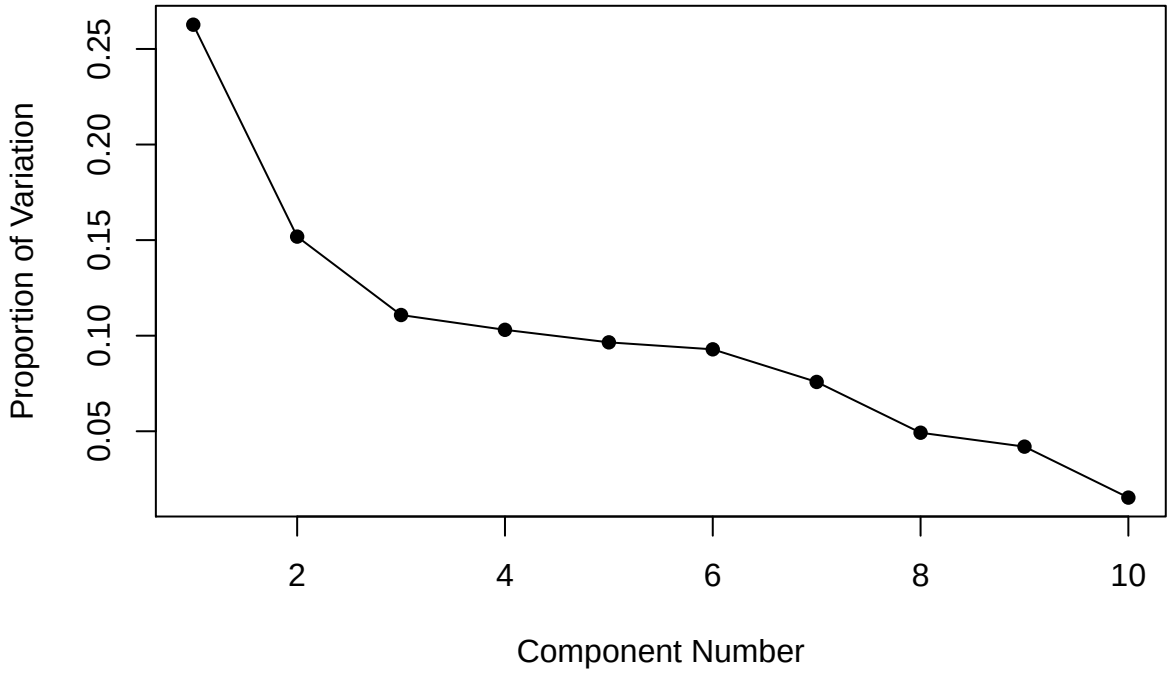


We can see that Decade Range very vaguely clusters by the first two components, since most of the blue datapoints are above `PC2 = 0` and most of the red datapoints are below it. The direction of the arrow for each attribute indicates how the principal components change as that attribute increases. For example, as BPM increases, PC1 decreases and PC2 increases. Given that the blue cluster is more in that direction, this allows us to conclude that songs from the 1990s-2010s tend to have a greater number of beats per minute than songs from the 1950s-1980s. Similarly, as duration increases, both PC1 and PC2 increase. This arrow again faces the blue cluster; therefore, we can conclude that songs from the 1990s-2010s tend to be longer than those from the 1950s-1980s. Finally, as Danceability increases, both PC1 and PC2 decrease. Since it is pointing towards the red cluster, this graph suggests that songs

from the 1950s-1980s are more danceable than those from the 1990s-2010s.

Normal distribution ellipses were drawn on top of the datapoints to visualize the degree by which these clusters differ. Since they overlap quite a bit, especially with respect to PC1, it is questionable whether the principal components, and consequently the quantiative attributes they represent, are significantly different across time/between the two Decade Range groups. In order to more concretely answer this, we can perform a two sample, two-sided t-test to compare the group means of each of the principal components. The null hypotheses are that the mean PC1 of songs from the 1950s-1980s is equal to the mean PC1 of songs from the 1990s-2010s, and the mean PC2 of songs from the 1950s-1980s is equal to the mean PC1 of songs from the 1990s-2010s. The p-value of the test comparing PC1 group means is $4.623e-07$, while that of PC2 is $< 2.2e-16$. So, contrary to what the ellipses seem to suggest, we are able to conclude that at level $\alpha = 0.05$, there is enough evidence to reject the null hypotheses that the PC1 group means are equal and the PC2 group means are equal.

[ADDRESS SCREE PLOT/THIRD DIMENSION?]



To address some of our qualitative variables, we made a comparison word cloud between

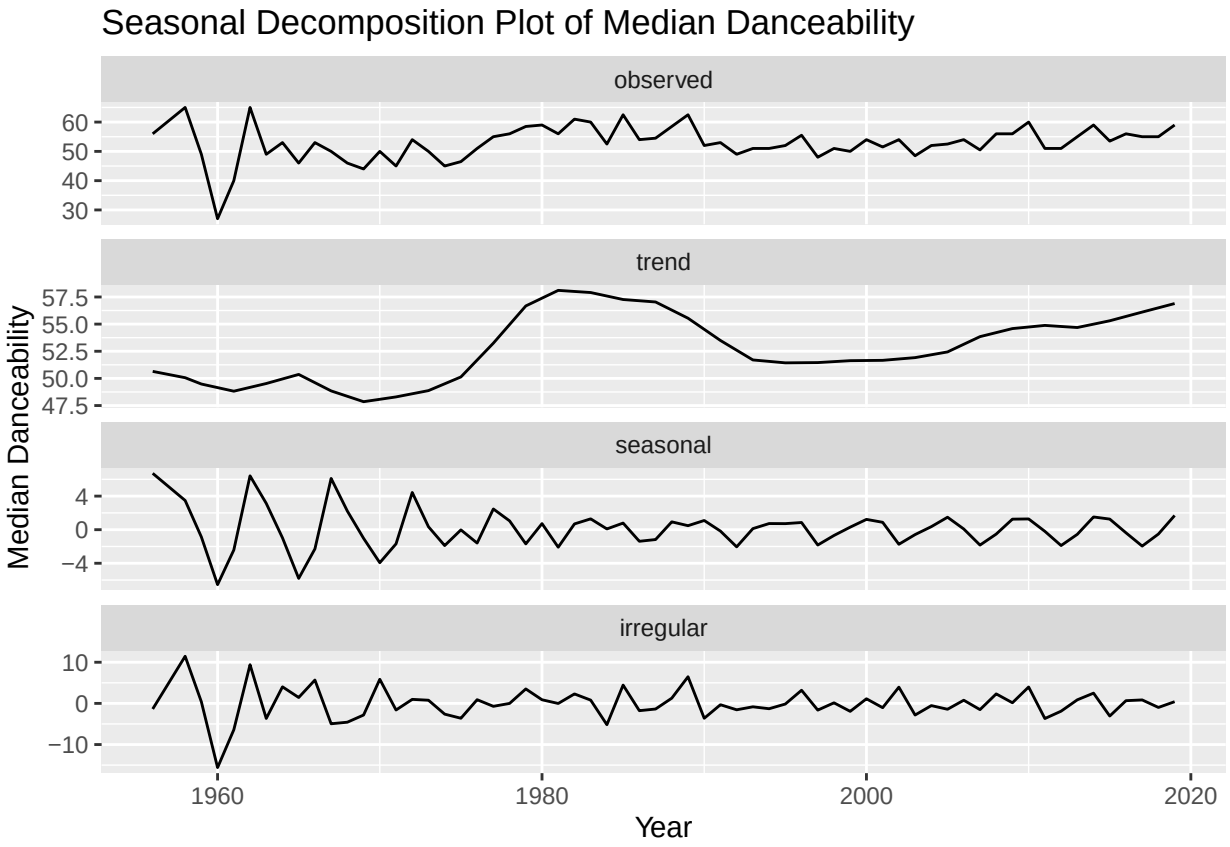
the top genres of songs from the 1950s to the 1980s and the top genres of songs from the 1990s to the 2010s to provide insight on how the top genres have changed, if at all, between these two eras.

In the word cloud, word that are large in size mean that they appear much more on the Decade Range side that they are on (denoted by the labels and colors), while those that are small in size indicate that they appeared roughly an equal number of times betwen the two Decade Ranges. The latter can mean that they only appeared a few times total, or multiple times in both.

Using the above rules to draw conclusions from our graphic, we can observe that are a few song genres that almost exclusively appeared in the 1950s-1980s, such as "adult standards," "classic rock," "album," and "europop." Note that the word cloud broke a few of these phrases a part. Meanwhile, "alternative," "modern," and "pop" music seem to be more popular genres in the 1990s-2010s. So, even though we grouped multiple decades together, making

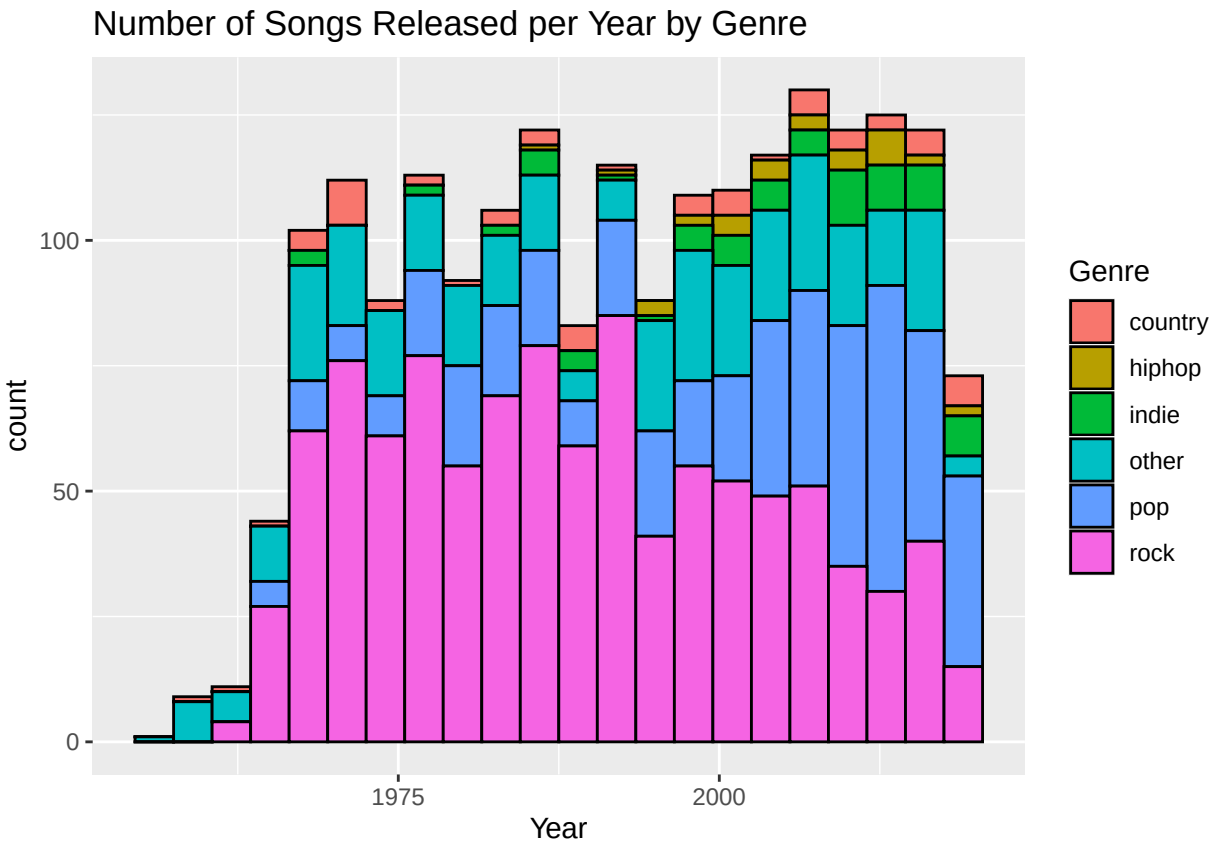
us unable to analyze how top genres may or may not have changed decade-to-decade, it is clear that there are some genres that were/are more prominent in one time period or another, indicating that there was indeed a genre shift over time.

While grouping years into Decade Range categories allows for some insightful graphs, it also makes sense to treat time quantitatively. To that end, to more closely monitor how a single quantitative attribute has changed over time, we constructed a time series plot with decomposition measuring median Danceability. We chose median over mean so that any particular year would not be too adversely affected by outliers, since some years had less songs in this dataset than others.

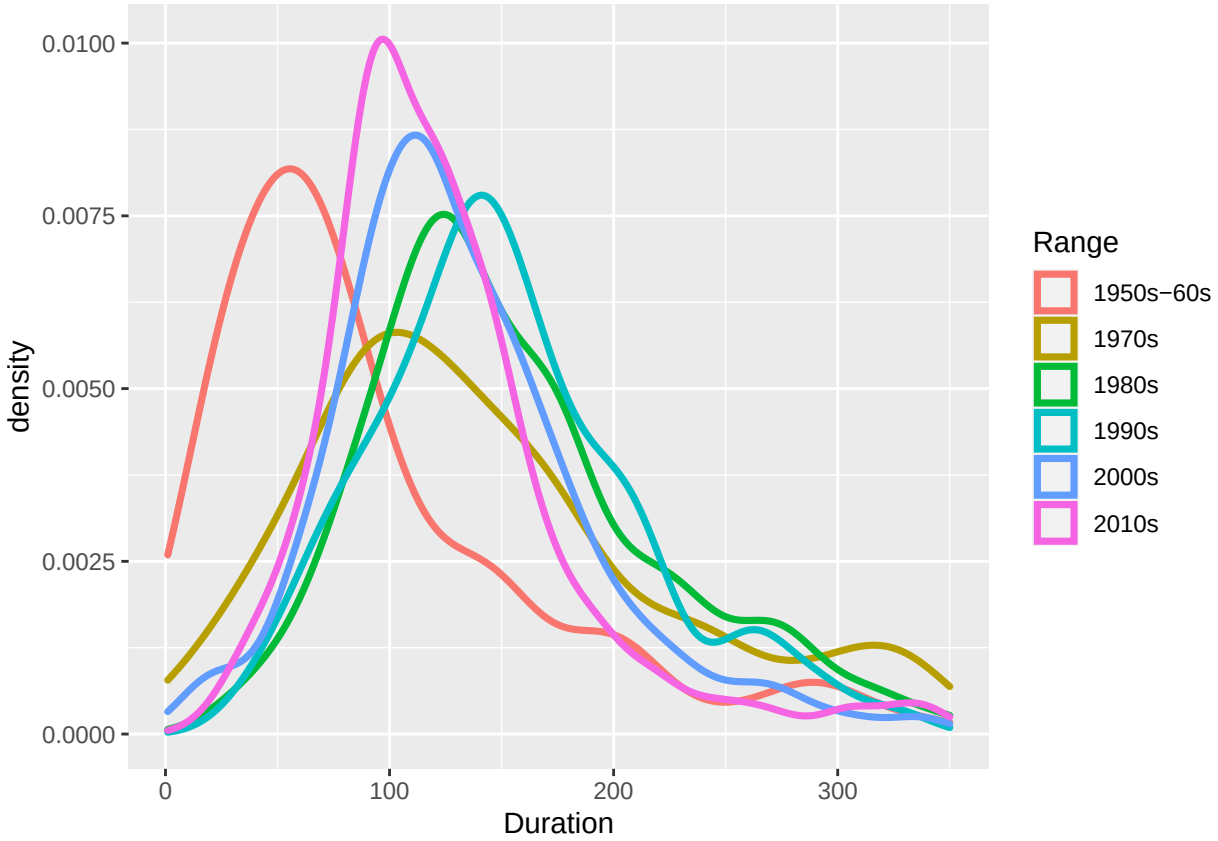


The global trend can be seen in the second facet, which shows that the median Danceability rating for songs started off rather low, climbed throughout the 1970s, reached a peak around 1980, decreased from the late 1980s and 1990s, and has been steadily increasing since 2000. The seasonal trend can be seen in the third facet: the consistent up-and-down nature

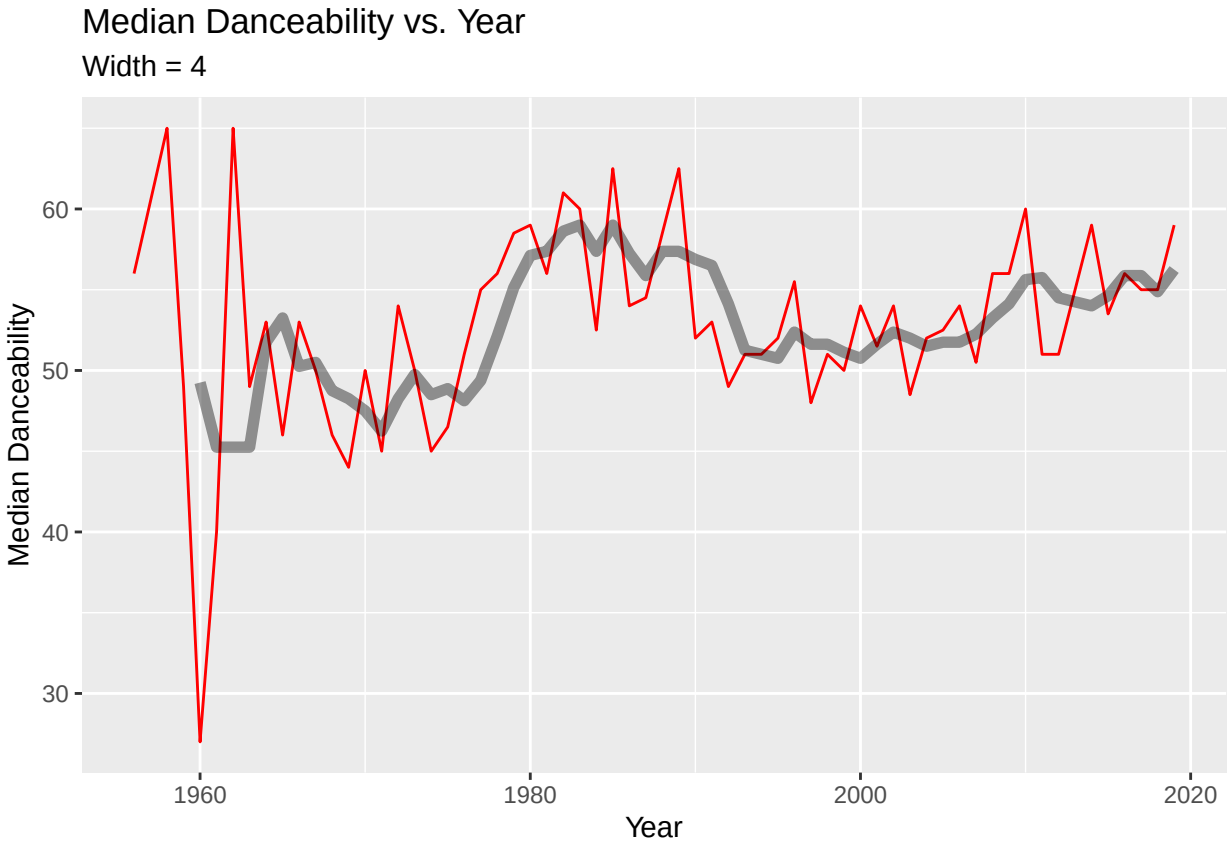
of this plot, especially since 1990, suggests that the median Danceability rating follows a cyclical pattern that lasts about five years per cycle. Therefore, the main takeaway from this graph is that not only does Danceability come in big waves over the span of decades, but it also comes in small waves over the span of a few years.



Graphs that probably will not be included in final version



Interpretation: song duration has changed throughout the decades. All are right skewed but 1950s-1960s seemed to have the smallest duration mode, followed by 1970s, 2010s, 2000s, 1980s, and then 1990s. Seemed to have cycled around.



Conclusions

Actual:

Overall, there is not much of a difference between the five genres, at least in terms of the quantitative variables in this dataset. While we did observe some minor trends in Energy and Danceability where one or two genres somewhat differentiated themselves from the rest, we could not observe any clear difference between them.

We were able to conclude that Popularity is not associated with any of the other qualitative variables in the dataset. Indie music is less popular than other genres, but there seem to be no significant differences among the rest.

Finally, we can conclude that a variety of qualitative and quantitative attributes have

appeared to change over time, such as Top Genre and Danceability, but not necessarily to a statistically-significant degree.

In our future work, we look forward to exploring these relationships with greater granularity and would be interested in experimenting with various subsets of the data to perform subgroup analyses.

Notes: Answering each research question!

Limitations: PCA plot (Adding and interpreting a third dimension to the PCA plot. Adding the third dimension has been learned, but not yet interpreting.)

Future work: potential other questions? Explore every quantitative variable in time analysis individually. Subgroup analysis.