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 (a very specific Spotify-labeled 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 based on trends we saw within the data (Rock, Pop, Country, Hip Hop, Indie, and Other). Because of this manual sorting, there is a possibility of some human bias in results related to genre. Meanwhile, the Decade and Decade Range (two ranges, 1950s-1980s and 1990s-2010s) 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

Question 1: Understanding Genre-Specific Attributes

In our analysis of genres, one of the first things we sought to analyze was how our quantitative variables varied individually by Genre (our manually sorted categories of rock, pop, country, hip-hop, indie, and other), excluding Popularity, since that will be covered extensively in the next section of our report. In our first pass, we created density plots, each colored by Genre, for each of the nine variables. These plots can be seen in Figure 1. Note that we excluded the "other" genre from this analysis, as the data in this category is too diverse to be of much use, and we judged that we probably would not be able to make any meaningful conclusions about it (as it includes such genres originally labelled as "blues", "electro", "reggae", "streektaal", etc.).

As we can see from the above density plots, BPM, Energy, Danceability, and Valence showed some clear differences between the genres, while the other five categories showed mostly similar trends between the five genres. For BPM, we can see two main peaks, with pop, indie, and hip-hop on the lower end, and country and rock having higher BPMs. This seems to be a fairly surprising trend, as many tend to perceive pop and hip-hop songs as having faster tempos, contrary to what this graph shows. For Energy, we can again see two main peaks, with hip-hop being on the higher end and country and indie on the lower end. Again, this is a somewhat surprising trend after seeing the results of the BPM graph, since we assumed higher-paced songs would tend to have higher amounts of energy. The Danceability plot shows that all but hip-hop have a similar distribution in that category. Lastly, for Valence, we can see that three of the genres (country, pop, rock) all have a similar distribution, while hip-hop and indie have different distributions (with slightly higher peaks). Overall, it seems like the five genres are mostly similar, but a few stand out in certain categories; in particular, hip-hop seems to stand out in many categories.

Next, we decided to take a look at a dendrogram using all of the quantitative variables (again excluding Popularity), to get a picture of how similar/different they were when taking all of the

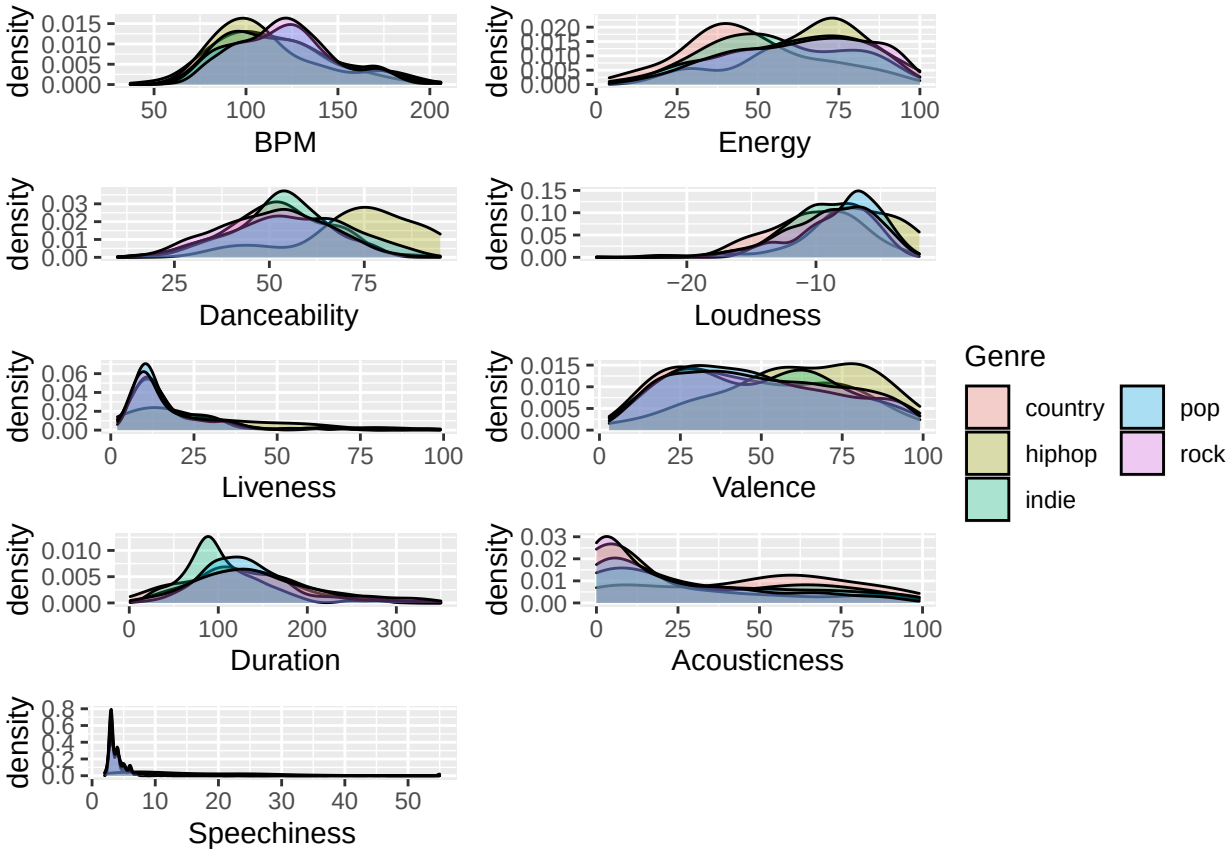
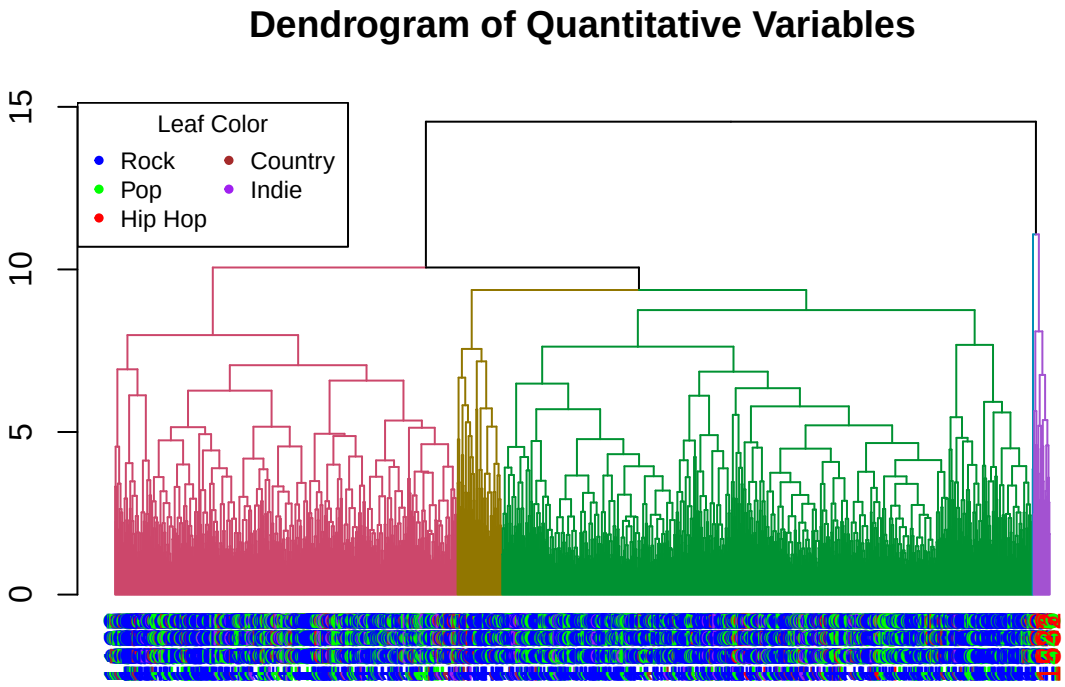


Figure 1: Exploratory density plots of quantitative variables by Genre

variables into account at the same time. In this dendrogram, five colors were used to color the branches because we have five genres; this was done so that if the genres tended to cluster together, it would make it more clear. Additionally, the "other" genre was again removed for the same reason as for the above graph.



In the dendrogram above, we colored the leaves by genre in order to visualize how the data ended up clustering. We can see that each cluster tends to be somewhat uniformly divided between the genres, except for the last cluster, which appears to consist of mostly hip-hop songs. From this graph, we can conclude that when taking into account the quantitative variables, there does not appear to be much difference between the five genres aside from hip-hop being different from the rest. This plot also confirms our suspicions that most of this dataset is composed of rock songs.

Lastly for this question, we looked at contour plots using the variables of interest that we highlighted from our density plots from above. The four variables that we looked at are BPM, Energy, Danceability, and Valence. We can look at the six contour plots in Figure 2, which accounts for all six pairings of these variables of interest.

These graphs provide a somewhat surprising conclusion: that despite our observations from the density plots, the five genres do not seem to differ much, if at all, when looking at two variables

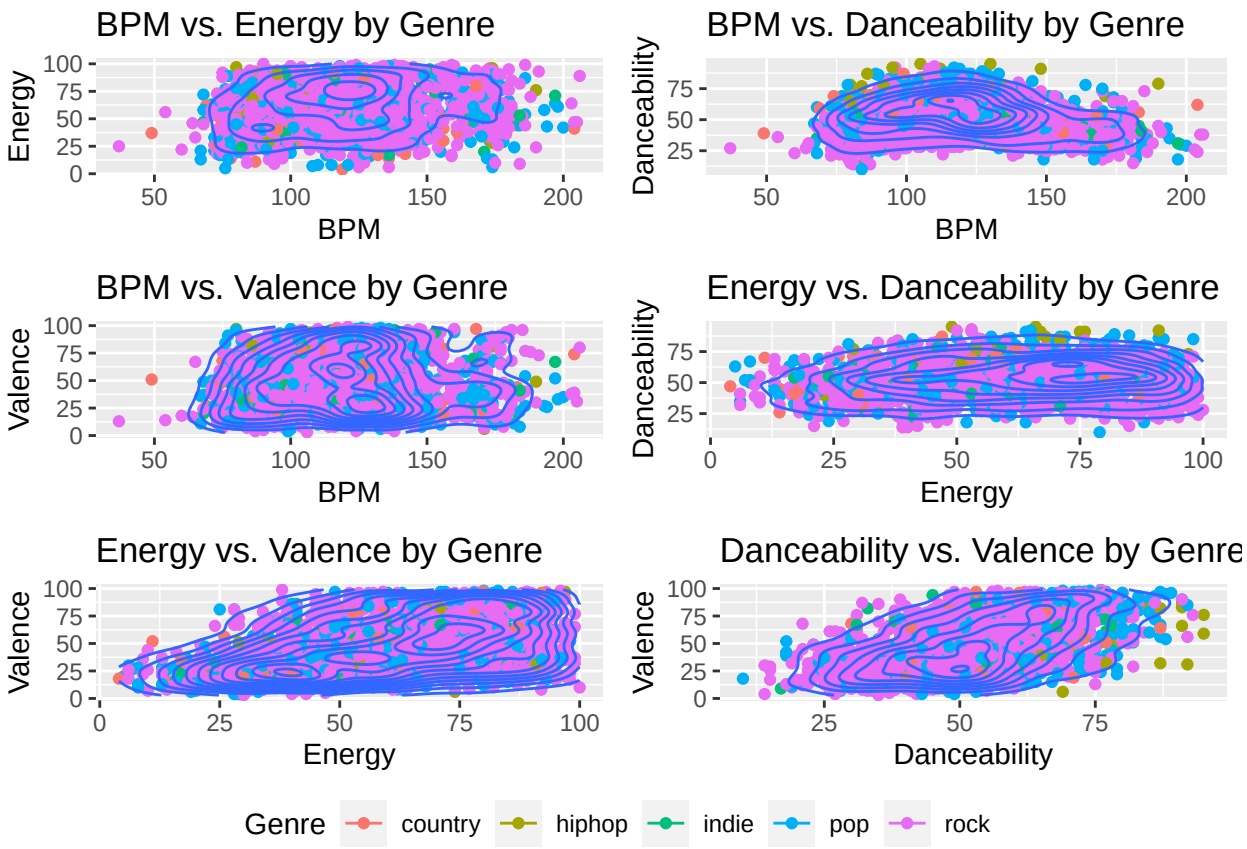


Figure 2: Contour plots of quantitative variables of interest

instead of one, even if just limited to the variables that seem to show differentiation. In the BPM vs. Energy, BPM vs. Danceability, Energy vs. Danceability, and Danceability vs. Valence graphs, there appears to be only one peak, centered around the middle, which means that there is little to no differentiation between the genres for these variables. The BPM vs. Valence graph, while showing two peaks, again does not really show too much differentiation, as those peaks are located fairly close to each other and on similary contour levels; furthermore, the genres seem very similarly distributed between them. The Energy vs. Valence graph is similar: three peaks, but close in distance and similarly distributed in genres.

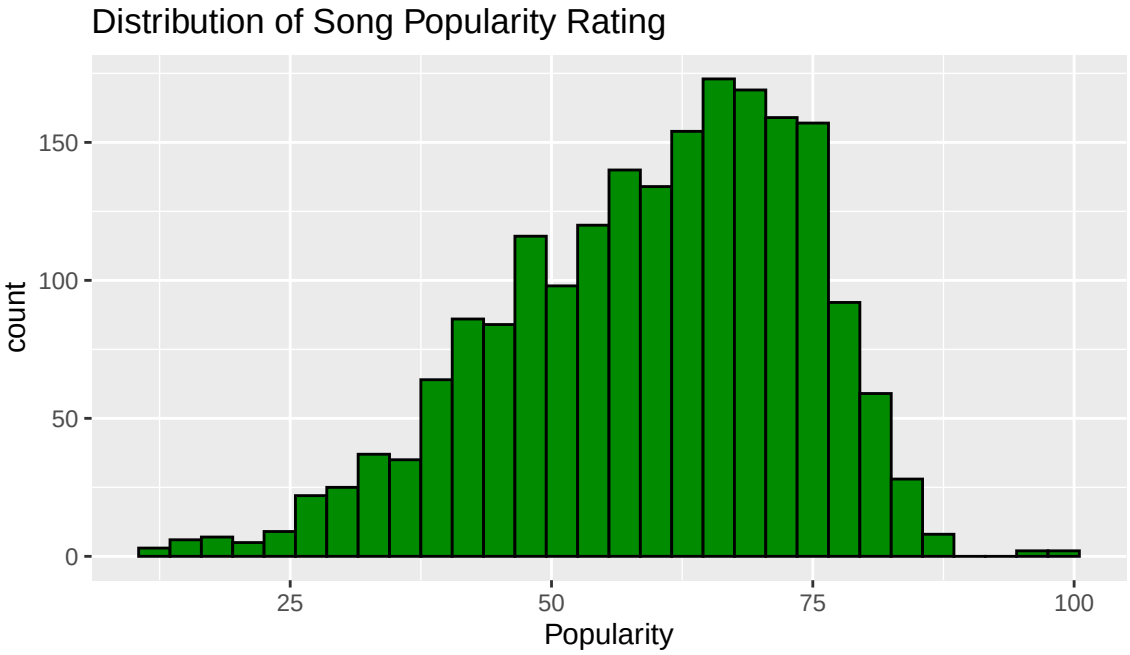
Overall, we have found that though looking at one variable at a time seems to reveal some differences between the genres, it seems like when considering all them holistically, the genres in this dataset seem to be fairly similar, in terms of the ten quantitative variables.

Question 2: Exploring Popularity

Interested in making some popular music? The best way to learn how may be to study the professionals. What qualities do popular songs embody? By taking a look at the correlations between Popularity and other attributes, we can see what these popular songs have in common. We should start by investigating just popularity itself across all songs with univariate exploratory data analysis in the form of a histogram. This allows us to get a sense of what sort of popularity ratings are common and uncommon amongst songs.

BPM	Energy	Danceability	Loudness	Liveness
-0.00318	0.103	0.144	0.166	-0.122
Valence	Duration	Acousticness	Speechiness	
0.0959	-0.0367	-0.0876	0.112	

Table 1: Correlations between Popularity and quantitative variables.



Most songs seem to be concentrated in the middle half of the possible popularity range (between 25 and 75). There is definitely a left skew, with a unimodal peak around 65.

We want to look into associations, so let's see how popularity is correlated with other quantitative variables through a pairs plot.

This pairs plot is incredibly messy. What we are really looking for is correlation between just Population and other individual variables.

After isolating the correlation coefficient portion of the pairs plot, we can see how correlated Popularity is with the quantitative variables. As the magnitude of all of these correlation coefficients are under 0.20, we can state that Popularity is not associated with any of these quantitative variables.

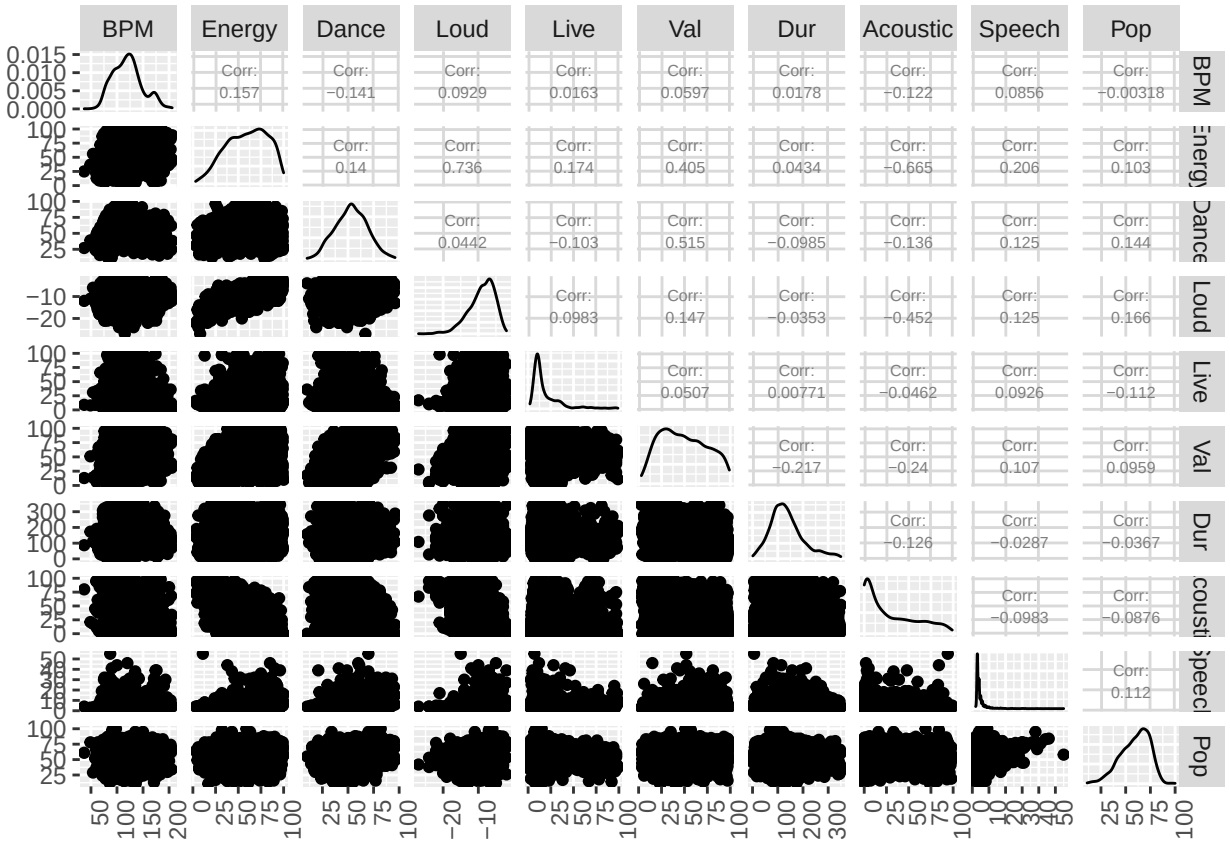
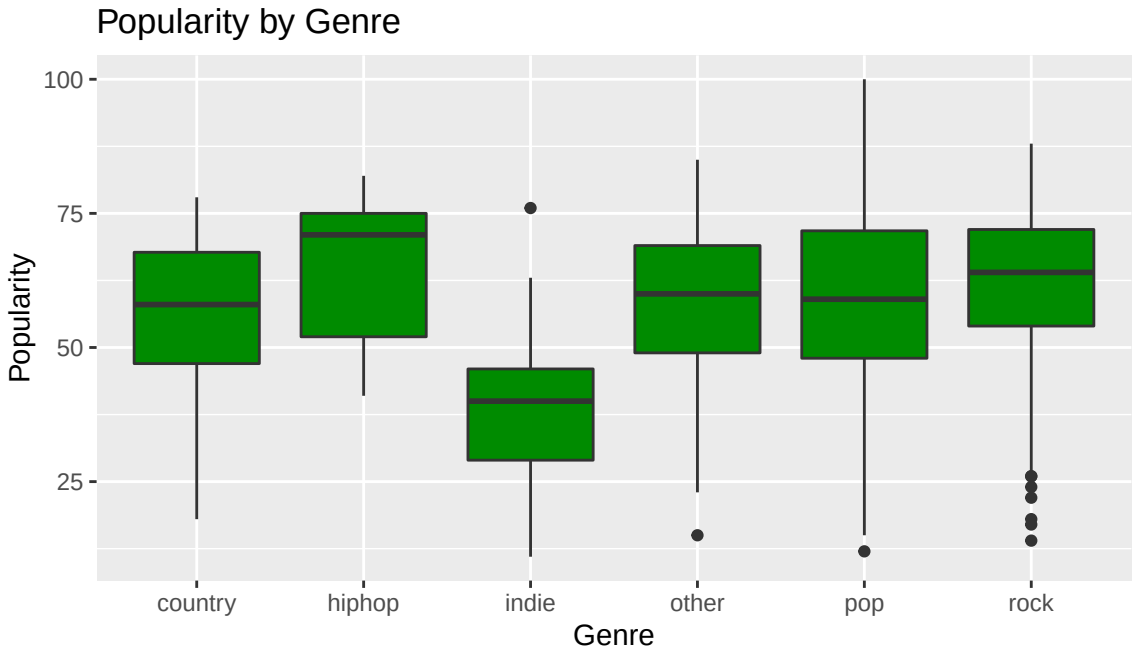


Figure 3: Pairs plot of all quantitative variables



Taking a look across genre, one clear trend seems to be that indie songs are less popular than those of the other genres. Indie's second and third quartile range from around 30 to 45, where all other genre's 25th percentile are above 45. Hip-hop appears to have the highest median popularity across the genres, but with some clear skew left so that its inner two quartiles overlap heavily with all non-indie genres. Overall, it is hard to judge any clear differences between the non-indie genres, but indie music is definitely less popular. This difference can be confirmed using a few statistical tests. We first used a One-Way ANOVA Test with the null hypothesis that the group means are equal. The p-value of the test is $< 2.2e-16$, meaning that there is sufficient evidence to reject the null hypothesis. We then performed multiple T-Tests between the popularity of indie songs and each of the remaining genres. In every case, the p-value is significant at a level $\alpha = 0.05$, meaning that we are consistently able to reject the null hypothesis that the group means are equal, in turn suggesting that the group mean between indie music and each of the other genres is significantly different.

If we want to look into what artists seem to have different ratings a popularity, one option is create two word clouds split on rating popularity. To make the two word clouds around equal in terms of amount of data, the less popular one has artists of songs with popularity less than or equal to 60 while the popular one has popularity greater than 60.

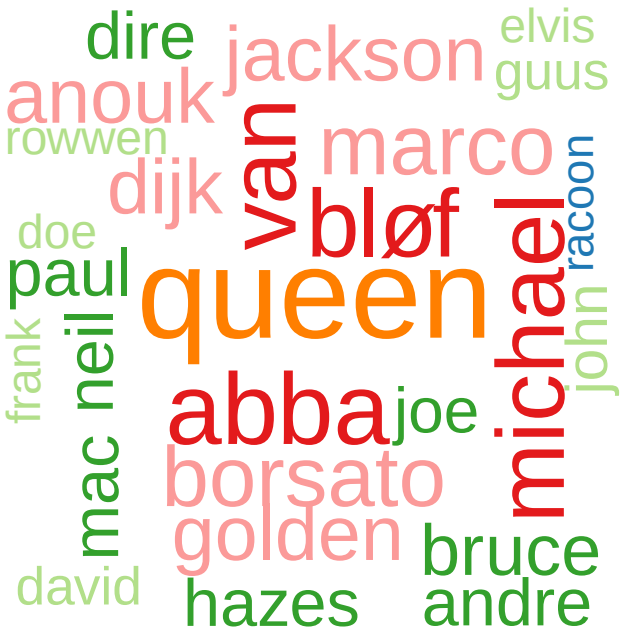


Figure 4: Artists of less popular songs

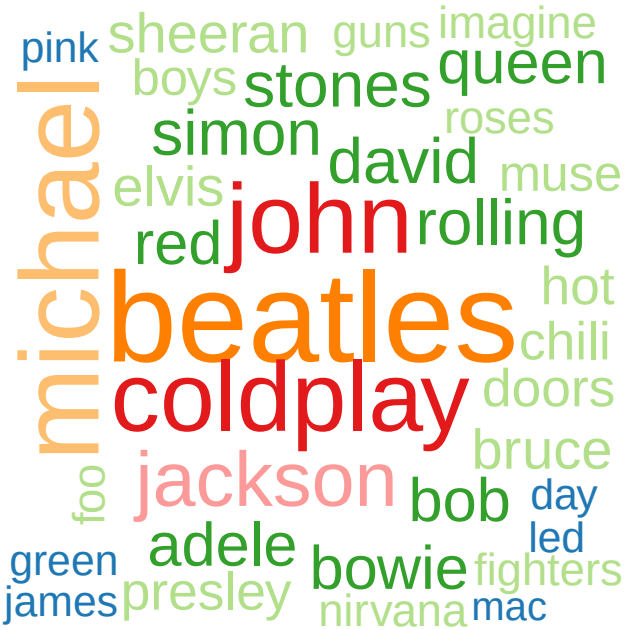
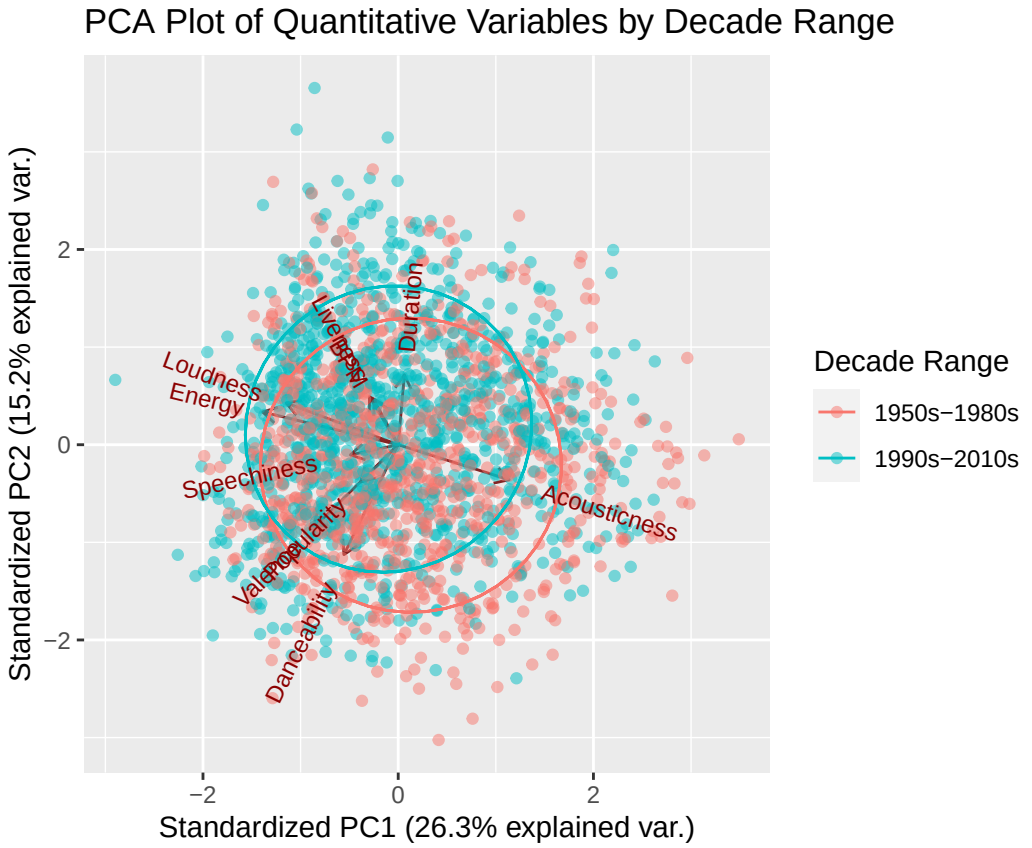


Figure 5: Artists of more popular songs

The larger the name on the word cloud, the more often they appeared in that subset of the data. Perhaps to our surprise, in Figure 4, we see that ABBA is most associated with releasing less popular music. Meanwhile, Coldplay, and The Beatles are associated with releasing more popular music, as seen in Figure 5. There are quite a few artists that appeared to relase music that had a mix of popularities, such as Queen, the Rolling Stones, Bruce Springsteen, and Michael Jackson. Note that we can only interpret these results in terms of associations and not causations; emulating the style of one of the artists on the second word cloud is not guaranteed to make you a superstar, unfortunately.

Question 3: Examining Time Trends

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. (Note: the word `BPM` overlaps with `Liveness`.) 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. 68% of the group's datapoints are contained within each ellipse. 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 for 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 might lead us to believe, we are able to conclude that at a 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. In conclusion, there are statistically-significant differences in these components when we group datapoints by time categories, but this does not necessarily translate to the quantitative attributes themselves.

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.

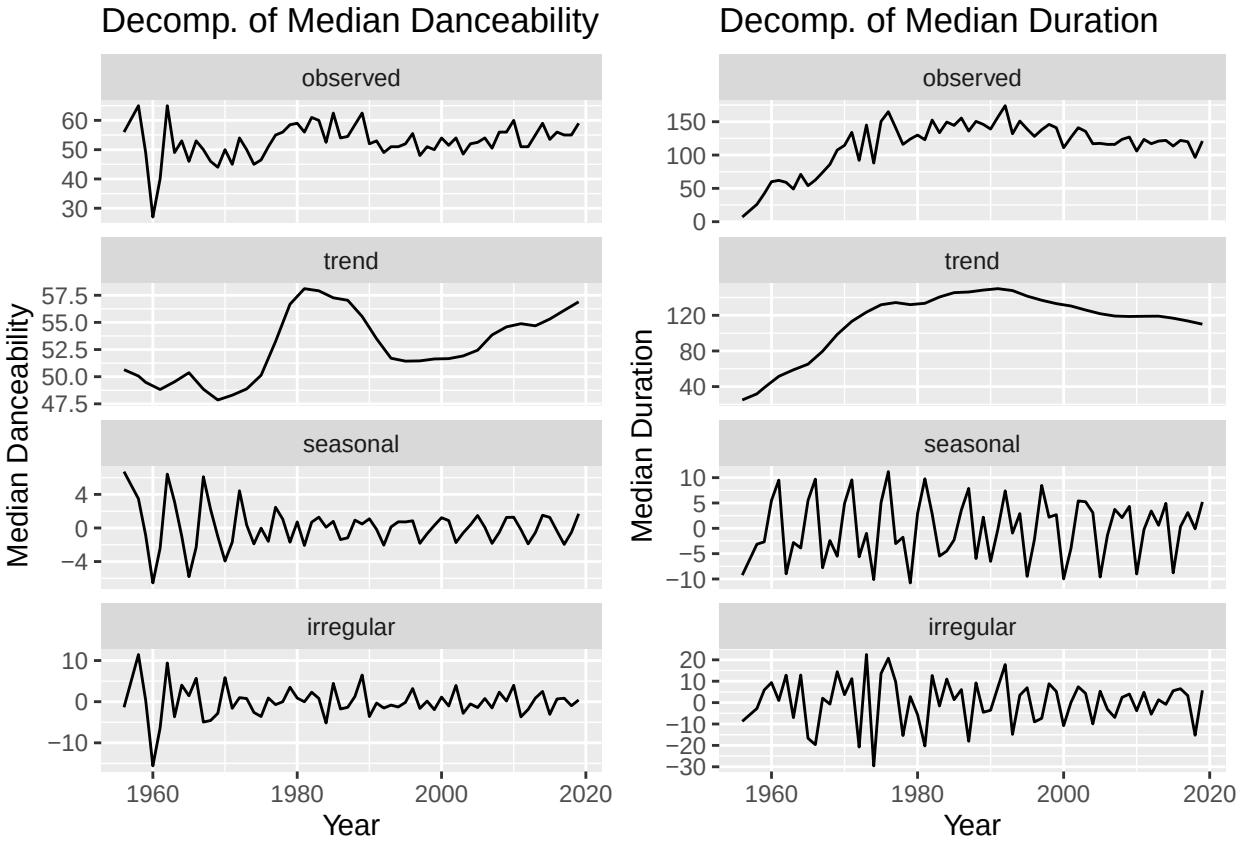
In the word cloud, word that are large in size mean that they appear much more on the Decade Range side that they are on (denoted by the labels and colors), while those that are small in size indicate that they appeared roughly an equal number of times betwen the two Decade Ranges. The latter can mean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

Using the above indicators to draw conclusions from our graphic, we can observe in Figure 6 that there are a few song genres that almost exclusively appeared in the 1950s-1980s, such as "adult standards," "classic rock," "album rock," and "europop." (Note: the word cloud broke a few of these phrases apart.) Meanwhile, "alternative," "modern," and "pop" music seem to be more popular genres in the 1990s-2010s. We can confirm this in one instance using a Chi-Square Test for equal proportions. Suppose we take "album rock"; the null hypothesis is that the proportions of "album rock" songs between the two Decade Range groups are equal. The p-value of this test is $< 2.2e-16$. This is significant at a level $\alpha = 0.05$, meaning that there is sufficient evidence to reject the null hypothesis. This test could be done for every genre listed in the word cloud. So, even

Figure 6: Comparison word cloud of Top Genre by Decade Range

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, including some to a statistically-significant degree, 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 time series plots with decomposition measuring median Danceability and median Duration. 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.

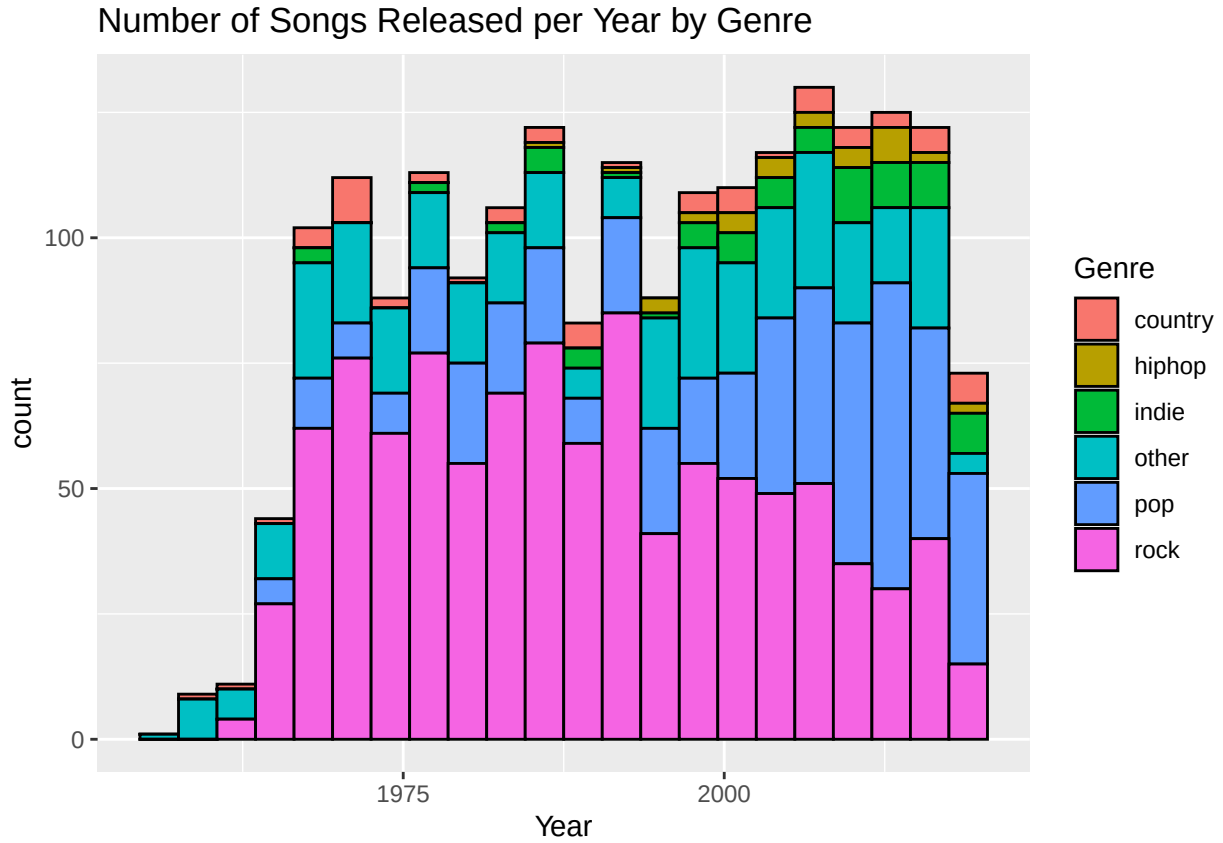


With respect to the lefthand graph, the global trend can be seen in the second facet, which shows that the median Danceability rating for songs started off rather low, climbed throughout the 1970s, reached a peak around 1980, decreased from the late 1980s and 1990s, and has been steadily increasing since 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.

Meanwhile, in the righthand graph, in terms of the global trend, Duration has steadily increased since the 1950s, reached a peak in the early 1990s, and has since slightly declined an plateaued. With respect to the seasonal trend, it is rather jagged; it goes up and down, but not as smoothly as the seasonal decomposition for median Danceability. The main takeaway is therefore that Duration changed a lot between the 1950s and 1990s, but that change has since become less volatile.

Finally, considering the patterns in the previous graphs and the genre distributions explored earlier in our report, we were interested in seeing how the number of top songs produced by each genre changed over time. In fact, perhaps these changes over time align with the rise and fall of certain genres.



In general, we can observe that the number of songs released is more or less uniformly distributed, so that has not changed over time. With respect to genre, however, the number of rock songs has decreased since the turn of the century, while the number of pop songs released/on the Top 2000 list has been increasing since about 1975. We can also see that hip-hop music started appearing in the 1990s. These results further validate what we saw in our comparison word cloud graphic, since the word "rock" appeared on the top half and the words "pop" and "hip-hop" appeared on the bottom half. That being said, we are more concerned with its relation to our previous graphic. Comparing the two, we see that the rise of pop music occurs at the same time as the first peak in Danceability we saw in the previous graph, while the rise of hip-hop music occurs at the same time as the second cycle/wave of Danceability, as well as the peak of Duration. Recall that in our earlier genre analysis graphs, the distribution of Danceability of both pop and hip-hop songs have a left skew. However, it is important to note that these conclusions cannot be interpreted causally.

Conclusions

Overall, there is not much of a difference between the five genres in terms of the quantiative variables in this dataset, aside from when they tended to be released. 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 also able to conclude that Popularity is not associated with any of the other quantitative 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, Danceability, and Duration. Some of these changes are statistically-significant, but others are not.

One limitation to our analysis that we acknowledge is that for the PCA plot, its corresponding scree plot suggested that we should plot the first three dimensions, as the elbow occurred at `k = 3`. However, we did not include it because we were not able to figure out how to both add and interpret the addition of a third dimension. So, our PCA plot is most likely missing some information. In the future, we can research this further so that we can better represent these

principal components. Another limitation comes from the data itself; there were many Dutch songs in the dataset even though after a thorough, manual research of their stats, they would not normally appear in a canonical Top 2000 songs listing. So, perhaps these are not truly the Top 2000 songs of all-time, but rather the Top 2000 songs of a Dutch-biased country or person. In the future, we could eliminate those songs entirely, or perhaps even address them more directly. Finally, as mentioned in our introduction, since we manually sorted the top genres into genres, there may have been some biased introduced in terms of where we ended up putting top genres that fit into multiple generic genre categories. Perhaps getting a music expert's opinion could ensure the accuracy of these categorizations.

Additionally, in our future work, we can analyze these relationships with greater granularity and would be interested in experimenting with various subsets of the data to perform subgroup analyses. We did not do those things in our report because we aimed to answer more overarching questions. Potential questions we can attempt to answer in the future include: "Is Dutch music similar to non-Dutch music?" and "Are there differences between different types of rock music?" We could also explore more prediction questions, such as "What attributes will songs in 2021 have?", since this report mostly concerned itself with exploration and inference. Overall, we see a lot of potential with this dataset and are excited by the possibility of working with it in even more depth moving forward.