Spotify Genre Analysis

DSAN 5100

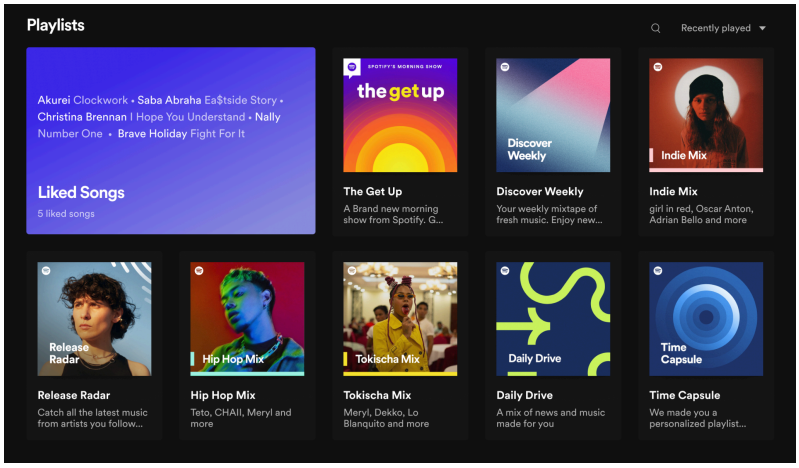
Probabilistic Modeling & Statistical Computing Final Project

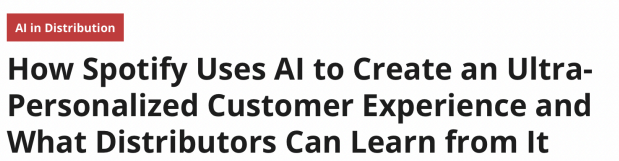
Marion Geary Bauman, Kendall Gilbert, Hannah Kim, Patricia Schenfeld

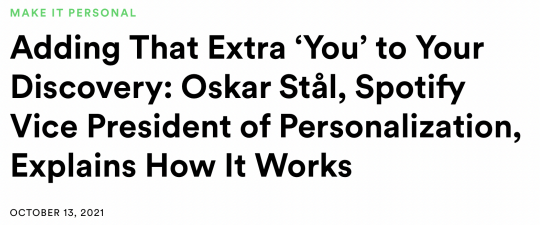
December 6, 2023

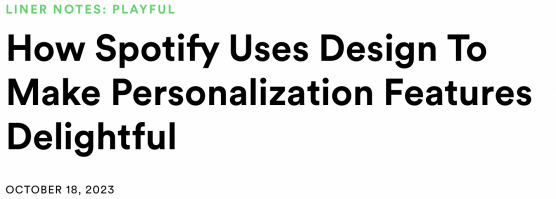
Introduction

Since Spotify's launch in 2008, they have become a globally-recognized name, with over 590 million monthly active users. It is one of the largest streaming services in the world, and has been pushing the boundaries of traditional media consumption, with a focus on personalized music recommendations. Below, you can see an example of many of the personalized playlists created by Spotify for its users, as well as numerous article headlines emphasizing the amazing work Spotify has done in the music personalization industry.



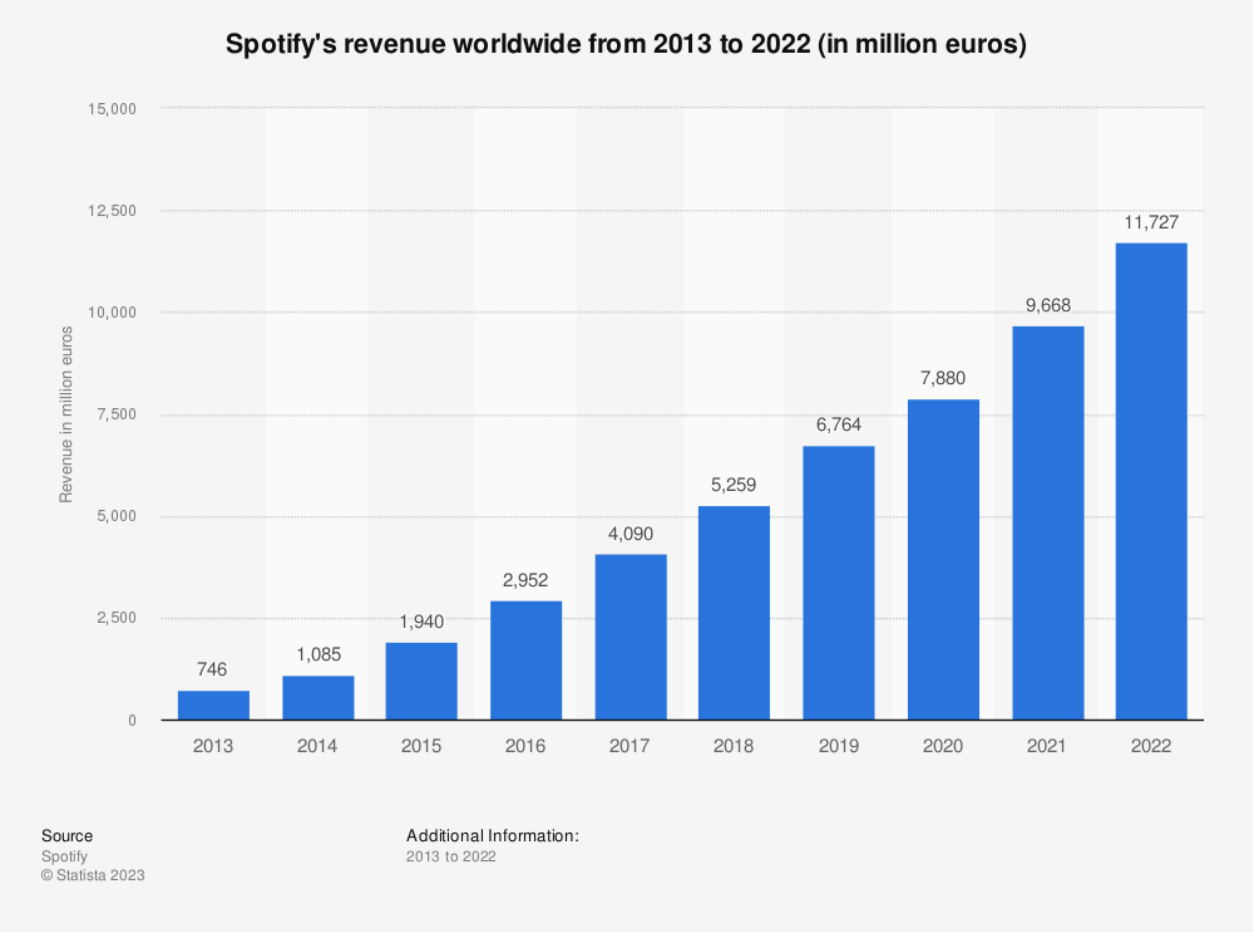






For our project, we wanted to explore the topics of audio features and how these may or may not contribute to the definition of certain genres. In the contemporary digital era, investigating this research topic holds significant relevance. Spotify serves as an exemplary model for innovation within the media sector, and comprehending its dynamics can offer insights into

broader implications for technology, entertainment, and user behavior. This pertains to the evolving landscape of content consumption in the future.



Above is a graph showing the worldwide revenue of Spotify from 2013 to 2022 in million euros. It is evident that Spotify's revenue has rapidly increased over the years, with no sign of decline. This reinforces the fact that Spotify has not only established itself as a dominant force in the global music streaming industry but also underscores its sustained growth trajectory. This continued revenue increase highlights the platform's ability to adapt to evolving market demands, maintain user engagement, and capitalize on emerging opportunities, affirming its position as a formidable player in the digital entertainment landscape.

This consistent and robust growth in Spotify's revenue reiterates the need for a comprehensive exploration of audio features across various music genres within a data science context, as understanding the nuanced patterns and preferences can provide valuable insights for optimizing user experiences, personalized recommendations, and further advancements in the field of music analytics. In our project, more specifically, we are narrowing in on 3 specific data science questions:

1. **Does tempo vary across genres?**
2. **Does 'popularity' have a dependence on 'speechiness' of the genre?**
3. **Is the average 'energy' across genres different?**

Dataset and Preprocessing

Our dataset consists of songs associated with different genres along with their audio features. The genres included are rock, latin, hiphop, R&B, pop, rap, and EDM. Some examples of audio features include tempo, loudness, energy, and danceability. This dataset is quite versatile as there are over 10 different audio features we can explore and perform statistical tests on. To begin, we needed to preprocess our data to ensure it was in a format usable for analysis. After importing the data in R, we took a first glance at the dataframe to see if there were any obvious issues we needed to fix. This gave us a good understanding of what the data looked like and what variables we were working with. One of the most crucial steps was to check for any missing values in the dataset. This ensures the integrity of our data, as missing values could lead to biased analyses and incorrect conclusions. Luckily, our data had no missing values, allowing us to continue with our data cleaning process. We then removed any unnecessary columns, which became apparent to us after getting a first look at the data. We want to simplify our dataset to include only variables that would be usable for analysis. From here, we removed 5 columns. The

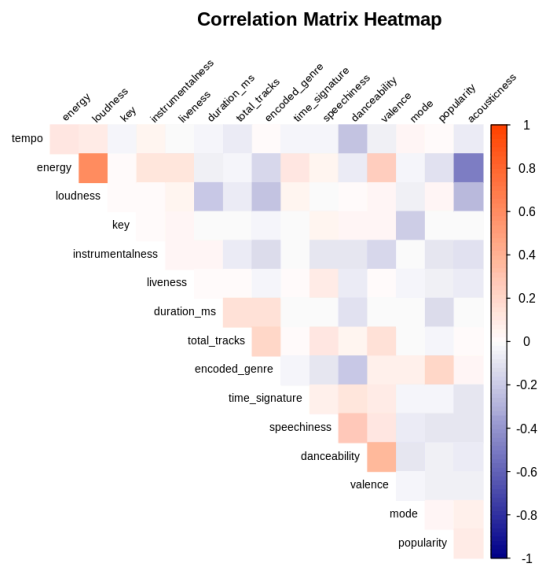
columns we removed were those that were either non-numerical or irrelevant to our analysis. For example, this included ‘Album_cover_link’, which was a column containing a web address to the image of the album cover. Columns as such were easy to spot visually when looking at the data. Lastly, we converted all column names to lowercase letters for consistency and ease of use during analysis. Below you can see a snapshot of what the data looks like after cleaning (all columns not shown). Although track names and artist names are included as well, we decided to focus our analysis on specific genres.

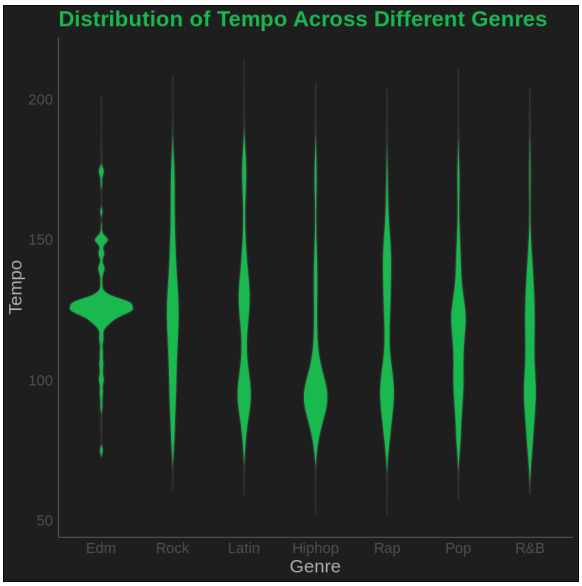
genre	title	artist	duration_ms	popularity	total_tracks	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	v
<chr>	<chr>	<chr>	<int>	<int>	<int>	<dbl>	<dbl>	<int>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	
rock	Baba O'Riley	The Who	300400	75	29	0.489	0.724	5	-8.367	1	0.0352	0.31300	0.185000	0.2870	
rock	More Than a Feeling	Boston	285133	78	8	0.377	0.681	7	-8.039	1	0.0298	0.00088	0.002300	0.0504	
rock	(Don't Fear) The Reaper	Blue Öyster Cult	308120	76	10	0.333	0.927	9	-8.550	0	0.0733	0.00290	0.000208	0.2970	
rock	Jump - 2015 Remaster	Van Halen	241599	78	9	0.572	0.835	0	-6.219	1	0.0317	0.17100	0.000377	0.0702	
rock	Stairway to Heaven - Remaster	Led Zeppelin	482830	79	16	0.338	0.340	9	-12.049	0	0.0339	0.58000	0.003200	0.1160	
rock	American Girl	Tom Petty and the Heartbreakers	214733	73	10	0.550	0.824	2	-5.988	1	0.0334	0.44800	0.000127	0.3660	

Exploratory Data Analysis

Before beginning our statistical analysis, we created numerous visualizations and statistical summaries in order to understand the dataset. Specifically, we were interested in understanding the audio profile of each different genre, as defined by each of the measured audio features. We aimed to understand the distribution of tempo, popularity, speechiness, and energy across each of the genres in our analysis. Data visualization allowed us to view the distributions and understand how they might differ before performing statistical analyses to support our hypotheses.

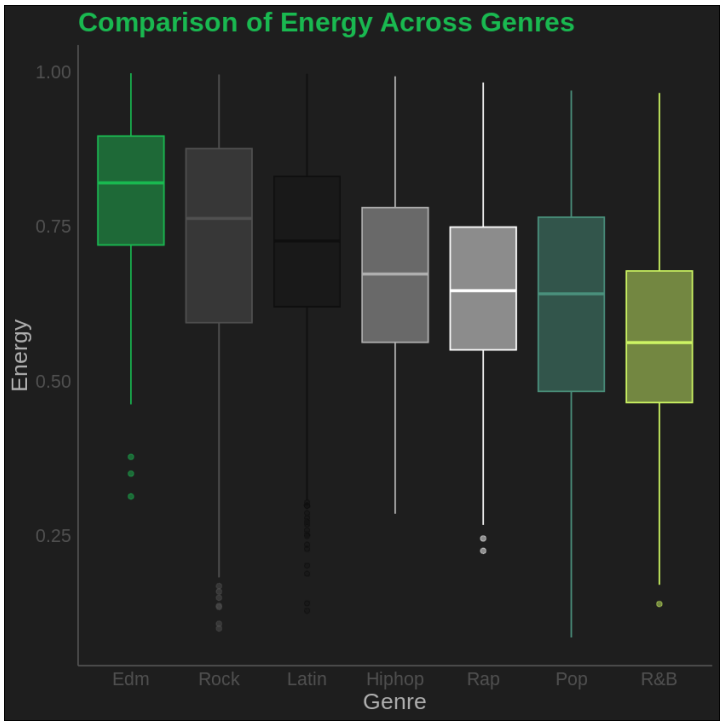
Our exploratory data analysis began with a correlational analysis to assess the relationship between features in our data. The heatmap below shows that, generally, our features are not highly correlated, making them suitable for our analysis. There do appear to be two points of higher correlation: energy and loudness, and energy and acousticness. However, we will not use loudness or acousticness in our analysis, so we do not need to account for these correlations.





After analyzing the correlations between the features in our data, we explored the variables that will be tested in our hypotheses. We started by analyzing the distribution of tempo across songs in each of the seven genres. A violin plot of the distribution of tempo for each genre shows that EDM has a distribution with low variance, with most songs having tempos around 125 beats per minute. All other genres have higher variance in tempo, spanning tempos between 75 and 250 beats per minute. Hiphop has the lowest tempo on average, with most songs having tempos under 100 beats per minute. We observe that rock music appears to have a slightly higher average tempo than Latin music, which has a concentration of lower tempos and a concentration of moderate tempos.

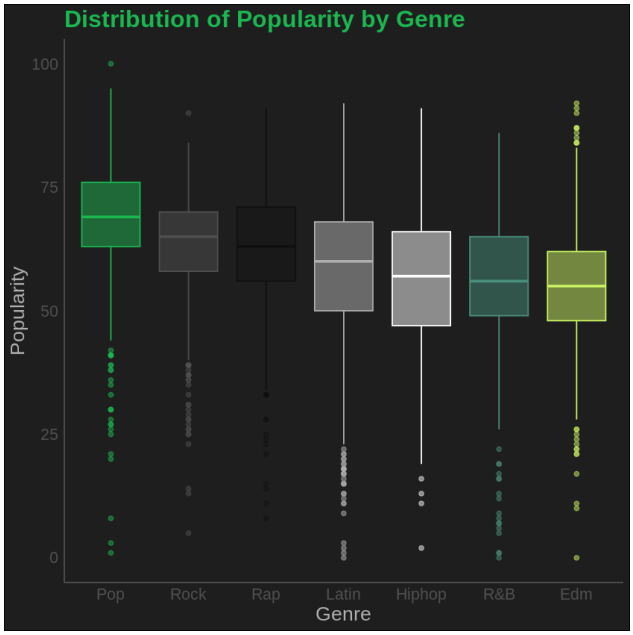
Next, we explore the distribution of energy across each genre in our dataset. According to Spotify, "[e]energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity"[1]. Unsurprisingly, we see that EDM has the highest median energy, where most EDM songs have an energy level of 0.75 or higher. Rock and pop have the widest range of energy values across their songs, with wide interquartile ranges and long tails on both ends of the boxplots. Latin has a slightly lower median energy than rock and a narrower interquartile range, meaning the energy level is more consistent for rock songs. Hiphop and rap have very similar energy distributions, with hip hop having slightly higher energy than rap on average. Pop has a similar median energy to rap, but pop has a much wider range of energy levels, showing the large variance of energy of pop songs. R&B songs have the lowest median energy with a smaller variance in energy level across the distribution.

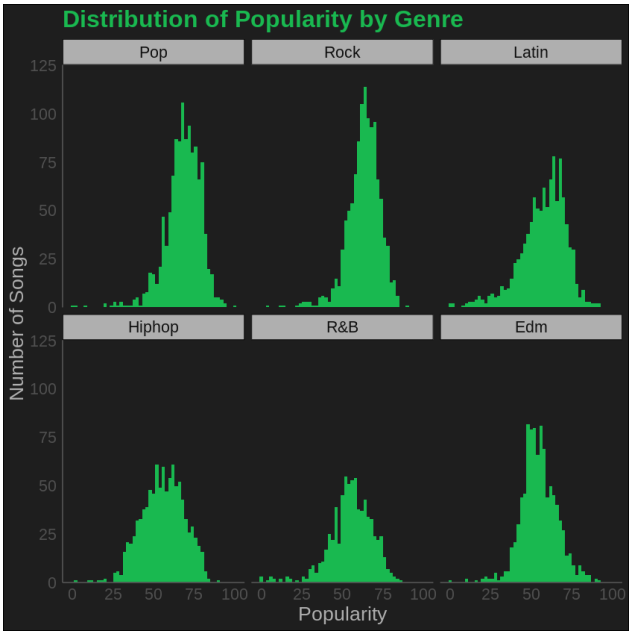


Our analysis now focuses on popularity of songs compared between the different genres. Box plots of the distributions of popularity by genre show that genres with higher median

[1] https://developer.spotify.com/documentation/web-api/reference/get-audio-features

popularities, such as pop and rock, have a smaller interquartile range and shorter tails, indicating lower variance in the popularity levels. Pop songs have the highest median popularity, which is expected based on our experience listening to the radio. Rock and rap have similar popularity distributions, with rock having a slightly higher median popularity. Latin and hip hop have the widest interquartile ranges, indicating more variation in popularity of songs from these genres. Histograms of the popularity for each genre[2] show that more popular genres like pop, rock, and latin have distribution that are skewed left, indicating that more songs in these genres tend to be popular. Less popular genres like hip hop, R&B, and EDM have more normal distributions of popularity, showing that the popularity of songs in these genres is closer to average.

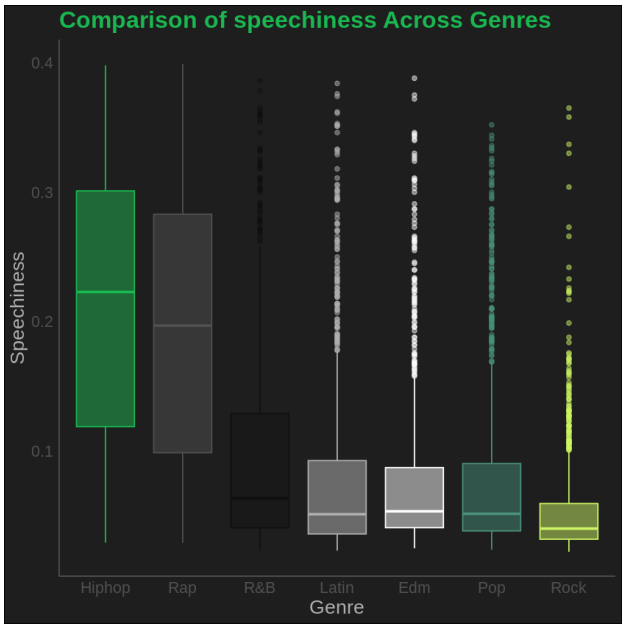


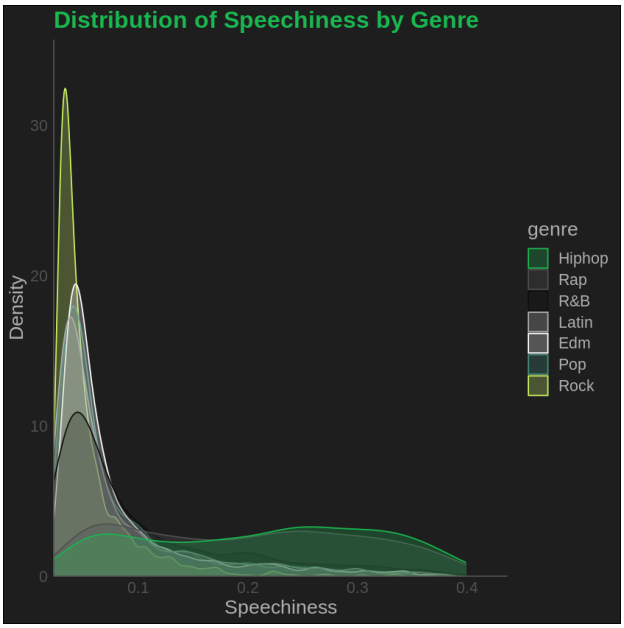


Finally, we analyzed the distribution of speechiness across genres. Speechiness is a measure between 0 and 1 based on the "presence of spoken words" in a song, according to Spotify's developer site[3]. Boxplots of speechiness between different genres show that hiphop and rap have very large variance in speechiness while all other genres have much lower variance in speechiness. Hiphop and rap have the highest median speechiness, including much more spoken

[2] Rap is excluded from the histogram of genre popularity for better readability of the visualization.
[3] https://developer.spotify.com/documentation/web-api/reference/get-audio-features

words than other genres. R&B has a moderate amount of speechiness, with the third highest median speechiness level. Latin, EDM, and pop have similar speechiness distributions, with small interquartile ranges, a large amount of outliers, and low median speechiness. Rock has the lowest median speechiness and the small interquartile range, signaling a low amount of speechiness across all songs in the genre. A density plot of speechiness levels, colored by genre, further demonstrates the significant difference in variance of speechiness across genres. Hiphop and rap have low, flat density plots, showing the wide variation in speechiness of songs in these genres. On the other hand, rock has a high peak of density near the axis, making a highly right skewed distribution, indicating that most rock songs have an extremely small speechiness level. Latin, EDM, and pop have similar distributions, all with right skews, though not as extreme as rock. R&B has a moderate right skew, showing that while R&B songs tend to have a lower speechiness, the genre still has substantial variation in speechiness.





Our exploratory data analysis helped us understand the differences in audio features between all seven genres in our data. We focused specifically on the features that we will be testing in our statistical analysis: tempo, energy, popularity, and speechiness. Our analysis

suggested that there are differences in these audio features between genres, so we will utilize statistical testing techniques in order to verify these differences and challenge our hypotheses.

Statistical Testing/Results

After successfully cleaning our data as well as exploring the features and relationships this dataset obtained in the exploratory data analysis we moved onto statistical analysis. In order to answer our data science questions we complete three different statistical tests within RStudio. These three tests included three Welch's t-tests, two chi-squared tests as well as bootstrapping. These three forms of statistical analysis were completed in an attempt to answer our data science questions:

- **Does tempo vary across genres?**
- **Does 'popularity' have a dependence on 'speechiness' of the genre?**
- **Is the average 'energy' across genres different?**

Welch's t-test is a statistical test used to compare the means of two different groups. This test assumes both groups of data are sampled from populations that follow a normal distribution, but it does not assume that those two populations have the same variance. We can use the central limit theorem to meet the assumption of normality since we have more than 30 songs for each genre. We used Welch's t-test to analyze the average mean of tempo between pop and edm, the average mean of tempo between Latin and rock, and finally the average mean of tempo between R&B and hiphop.

First, we compared the genres pop and EDM in regards to their tempo. Within this statistical test the null hypothesis was that there is no difference between the average tempo within the pop genre and EDM genres. Additionally, the alternative hypothesis was that there is a difference between the average tempo in the pop genre and ED genre. After completing this test,

we obtained a p-value less than 0.05 meaning we had sufficient evidence to reject the null hypothesis and conclude that there is a statistically significant difference between the average tempo in pop music and the average tempo in EDM music. Furthermore, from this test we are 95% confident that the true difference in means falls between -10.12 and -6.19 and the average pop tempo is slower than the EDM genre.

Next, we completed Welch's t-test on the genres latin and rock to get another perspective on the tempo of different genres within this data set. Similar to the first test completed, the null hypothesis was that there is no difference between the average tempo within the latin genre and rock genre. Additionally, the alternative hypothesis was that there is a difference between the average tempo in the latin genre and rock genre. The results we obtained from this test were similar to the first t-test. With a p-value less than 0.05 we had sufficient evidence to reject the null hypothesis and conclude that there is a statistically significant difference between the average tempo in latin music and the average tempo in rock music. We are also 95% confident that the true difference in means falls between -5.93 and -1.02. This analysis allowed us to conclude that the tempo in latin and rock music is very different. The tempo within latin rock is slower than the tempo within rock music.

The last t-test we completed was to compare the means between R&B and Hip Hop music. The null hypothesis was that there is no difference between the average tempo within the R&B genre and Hip Hop genre. Additionally, the alternative hypothesis was that there is a difference between the average tempo in the R&B genre and Hip Hop genre. For the third time we received very similar results. With a p-value less than 0.05% we have sufficient evidence to reject the null hypothesis that tempo within the R&B genre and Hip Hop Genre are the same meaning that we support the alternative hypothesis. Additionally, we are 95% confident that the true difference in

means falls between 1.27 and 6.84. The tempo for R&B music is faster than the tempo within Hip Hop music.

Overall, these three Welch's t-tests provided us with results to conclude that the average tempos with the genres, pop, edm, latin, rock, R&B and Hip Hop differ. The tempo of specific genres depends on the genre. Although this may seem obvious to the traditional music listener, it is interesting to see the numbers and statistical analysis behind this data science question.

We also completed two Pearson's Chi Squared Tests in order to answer our data science question - **Does popularity have an association with speechiness?** Pearson's chi squared test is a statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance. We looked at two different genres when completing these chi squared tests, pop and R&B. The null hypothesis for the chi squares tests carried out was that there is no significant association between the popularity and speechiness within the pop and R&B genres. The alternative hypothesis was that there is a significant association between the popularity and speechiness within the pop and R&B genres.

In order to complete these chi squared tests we had to prepare and set up our data in a specific way. First, we classified each song as low popularity, moderate popularity or high popularity. Low popularity was a popularity value less than 30, moderate popularity was a popularity value greater than or equal to 30 and less than 70. Finally, high popularity was considered for values greater than 70. Next, we divided speechiness into three different bins. These bins were all of equal sizes. This was done in order to make the contingency table easier to read and analyze in our future testing. Contingency tables are a type of table in a matrix format that displays the multivariate frequency distribution of the variables. These tables were crucial to complete these chi squared tests. Finally, we filtered out the specific genre that we were looking at

for the specific chi squared test. We completed chi squared tests on this modified dataset for the pop genre as well as the R&B genre. It was interesting to see how these test results differed for the specific genre.

The first chi squared test completed was specifically on songs classified within the pop genre. After analyzing how the binning was working and setting up the contingency table, we were able to complete the chi squared test. Unfortunately, the first time running this statistical test we ran into a warning message. This warning message was a result of low values within the contingency table. If we disregard this warning we could pull conclusions from our data - this report will analyze the chi squared test with the warning and the Fisher's Exact test and permutation test we completed in an attempt to get rid of this error. With a p-value higher than 0.05 at 0.2892 there is insufficient evidence to reject the null hypothesis meaning that there is no association between popularity and speechiness within the pop genre.

As stated earlier, we also completed the Fisher's Exact test. Fisher's exact test is a statistical significance test used in the analysis of contingency tables used when values in the contingency table are small. In the process above, we categorized the popularity into three different groups. In order to compute the Fisher's test, not only do we need the classified popularity values, but we also have to classify the speechiness column into two groups as following: speechiness lower than 0.05 was labeled as 'low speechiness', while the ones with higher than 0.05 was categorized as 'high speechiness'. Once a contingency table was created using both classified popularity and speechiness data, we computed the Fisher's test. This test did not give any warning, and with a p-value of 0.0008872, which is below the significance level of 0.05, we can conclude that there is significant evidence to reject the null hypothesis. This implies a significant association between popularity and speechiness within the pop genre.

Additionally, we also completed a permutation test in order to analyze the association between pop music popularity and speechiness. Permutation tests are non-parametric statistical tests that rely on the assumption of exchangeability. In order to obtain a p-value, we randomly sample (without replacement) possible permutations of our variable of interest. The p-value is the proportion of samples that have a test statistic larger than that of our observed data. With the permutation test, we were able to compare the association into details. The test was conducted among pop songs categorized by high popularity versus moderate popularity and songs categorized between moderate and low popularity. The p-value between the high popularity group and moderate popularity group was 0.99682, indicating insufficient evidence to reject the null hypothesis. Hence, there is no significant association between popularity and speechiness within the high popularity and moderate popularity groups of the pop genre. The p-value between the moderate popularity group and low popularity group was 0.1869, which was lower than that between high popularity and moderate popularity group. Yet, since the value is still greater than 0.05, we do not have sufficient evidence to reject the null hypothesis. Therefore, it is concluded that there is no significant association between popularity and speechiness within the moderate and low popularity groups of this genre. Combining these two outcomes, it leads to a conclusion that there is no significant association between popularity and speechiness within the pop genre.

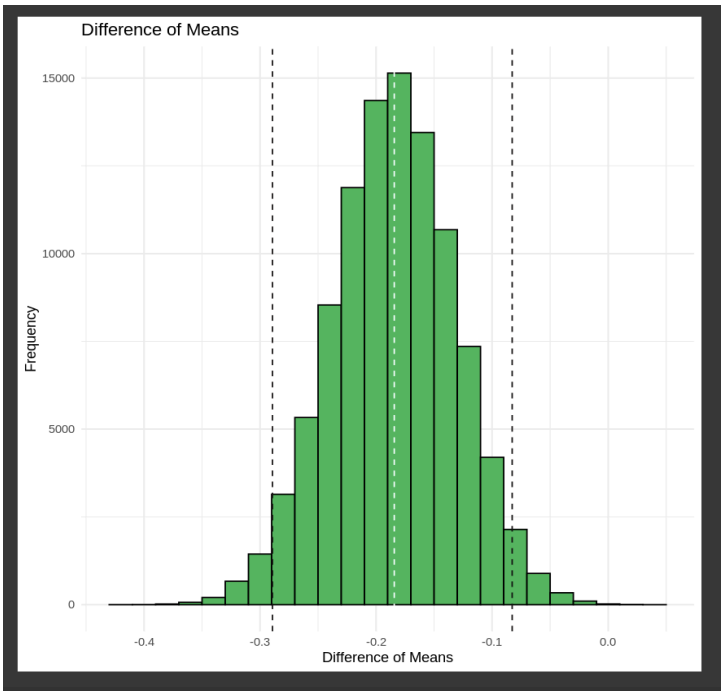
Overall, the results from these three tests — chi-squared test, fisher's test and permutation test — have provided insufficient evidence to reject the null hypothesis, indicating no significant association between popularity and speechiness within the pop genre. Although the results were different in Fisher's test, the collective outcome from the other two tests supports that there is a significant association.

We also completed a chi squared test to test if there is an association between speechiness and popularity within the R&B genre. We chose to look at this genre as it is very different from pop and we wanted to achieve different results. Luckily, we received no warning messages when completing this test due to the larger values within the contingency table. The chi squared provided us with key insights into the R&B genre. With a p-value less than 0.05 there is sufficient evidence to reject the null hypothesis meaning that there is a statistically significant association between speechiness and popularity within the R&B genre. It can be concluded that population level may be higher or lower depending on the speechiness levels in this music. For example, a song with low speechiness may be very popular or the opposite.

The chi-squared test is a statistical test used to determine if there is a significant association between two categorical variables. The chi squared tests within our project helped us analyze the association between two different features within this dataset - speechiness and popularity.

The third statistical test conducted to address our third question — **Is the average 'energy' across genres different?** — was the bootstrapping method. This method aimed to compare the difference between the energy feature values of pop and EDM genres. Initially, we sampled from the actual data to obtain subsets for each group. Then we achieved the 95% confidence interval which was [-0.2833, -0.0874] by computing the difference in average values. Since the interval does not include zero, we concluded that a difference exists between the average energy features. This conclusion is further supported by the minimal bias value of -5.910,

meaning that this method provided an accurate estimate of the difference of the average values. Additionally, the histogram plot gave us more detailed insights into the outcomes of this approach. Confirming that the actual difference in average values lies within the confidence interval, it is concluded that the difference of energy between the pop songs and edm songs are statistically significant.



Overall, these three statistical methods allowed us to answer the data science questions:

- **Does tempo vary across genres?**
- **Does 'popularity' have a dependence on 'speechiness' of the genre?**
- **Is the average 'energy' across genres different?**

After rejecting the null hypothesis for all three t-tests conducted it can be concluded that the tempo of songs differs greatly depending on the genre. We are now confident in the findings that the average tempo of EDM songs is faster than Pop songs, the average tempo of R&B songs is faster than Latin songs and the average tempo of Rock songs is faster than Hip Hop songs. Furthermore, using chi squared tests, fisher's test and permutation tests it can be concluded that there is no significant association between the popularity and speechiness of the pop songs. However, the chi squared test showed that there is a significant association between the popularity

and speechiness of the R&B songs. Finally, bootstrapping illustrates that the difference of energy between pop and EDM songs are statistically significant.

Conclusions

In conclusion, this project provided us with the opportunity to analyze real data and make informed decisions using statistical analysis. Within this project, we cleaned our data set, completed exploratory data analysis and successfully completed a variety of statistical techniques in order to answer our data science questions that we established at the beginning of this project. Completing these data science steps helped us develop crucial insights into our data that would have been unknown otherwise. The results obtained within the statistical testing allow us to conclude that the features of music typically depend on the genre. If given more time, it would be interesting to dive deeper into more of the audio features like instrumentalness or see if specific discography from a certain artist has any impact on certain features.

Works Cited

Berk, Michael. “How to Use Permutation Tests.” *Medium*, Towards Data Science, 21 Sept. 2021, www.towardsdatascience.com/how-to-use-permutation-tests-bacc79f45749.

Cohen, Benj. “How Spotify Uses AI to Create an Ultra-Personalized Customer Experience and What Distributors Can Learn from It .” *Distribution Strategy Group*, 27 Sept. 2023, www.distributionstrategy.com/how-spotify-uses-ai-to-create-an-ultra-personalized-customer-experience-and-what-distributors-can-learn-from-it/.

“Fisher’s Exact Test.” *From Wolfram MathWorld*, www.mathworld.wolfram.com/FishersExactTest.html. Accessed 6 Dec. 2023.

“Get Track’s Audio Features.” *Web API Reference | Spotify for Developers*, www.developer.spotify.com/documentation/web-api/reference/get-audio-features. Accessed 6 Dec. 2023.

Götting, Marie Charlotte. “Spotify Revenue 2013-2022.” *Statista*, 26 Sept. 2023, www.statista.com/statistics/813713/spotify-revenue/.

Goldrick, Stacy. “Adding That Extra ‘you’ to Your Discovery: Oskar Stål, Spotify Vice President of Personalization, Explains How It Works.” *Spotify*, 28 Oct. 2021, www.newsroom.spotify.com/2021-10-13/adding-that-extra-you-to-your-discovery-oskar-stal-spotify-vice-president-of-personalization-explains-how-it-works/.

Goldrick, Stacy. “How Spotify Uses Design to Make Personalization Features Delightful.” *Spotify*, 18 Oct. 2023, www.newsroom.spotify.com/2023-10-18/how-spotify-uses-design-to-make-personalization-features-delightful/.