INST 447 Data Analytics Team Project
Aneesa Munir, Shamyr Douglas, Will Heffernan, Rebekah Warner, Mohamed Ndaloma

Section 1:
Problem definition

As a top music consulting firm our company aims to get a greater understanding of the top ranked songs. By analyzing data consisting of the songs on spotify on the top 200 list since 2016, we will have a greater understanding of the attributes of these to songs and their relationships with each other. Our focus is specifically on variables such as danceability, energy, tempo, and popularity (streams). We aim to analyze relationships between popularity and danceability, tempo and popularity, as well as acousticness and danceability. The conclusions from these analyses will equip our company with the information necessary to help our clients continue to produce singles and albums that stay on the charts for weeks at a time!

Key questions include:
1. **Is a song more likely to be popular if it has higher 'danceability'**
2. **Is there a positive relationship between popularity and the tempo of a song?**
3. **Is there a relationship between the acousticness and danceability?**

Section 2. Description of the dataset

Key variables described in this data set are the number of streams which essentially represent popularity. Danceability refers to how well the song is suited for dancing. Energy refers to the intensity and activity of the song. Down below are more specific and detailed descriptive statistics for our dataset.

Feature	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Danceability	:0.1500	:0.5910	:0.6980	:0.6817	:0.7850	:0.9850
Energy	:0.0218	:0.5340	:0.6510	:0.6365	:0.7590	:0.9890
Speechiness	:0.0232	:0.0440	:0.0722	:0.1219	:0.1630	:0.9660
Acousticness	:0.0000	:0.0444	:0.1450	:0.2368	:0.3560	:0.9940
Instrumentalness	:0.0000	:0.0000	:0.0000	:0.0125	:0.0000	:0.9530
Liveness	:0.0197	:0.0974	:0.1240	:0.1802	:0.2190	:0.9770
Valence	:0.0320	:0.3160	:0.4890	:0.4924	:0.6690	:0.9820
Loudness	:-34.475	:-7.564	:-5.983	:-6.351	:-4.673	:1.509
Tempo	:46.72	:98.01	:120.03	:122.12	:142.03	:212.12
Duration (ms)	:30133	:173038	:198367	:202567	:226003	:690732

| Weeks on Chart | :1.00 | :1.00 | :4.00 | :13.61 | :16.00 | :367.00 |
| Streams | | :2.525e+06 | :8.695e+06 :2.713e+07 | :1.090e+08 | :9.951e+07 | :3.528e+09 |

The mean danceability was .6817 which indicates that on average, each song was moderately danceable. The mean energy score was .63 which means most songs had moderate energy. The mean acousticness score was .2368 indicating that most songs had low acousticness. There was a wide range for acousticness with a minimum score of 0.0 and a maximum of .9940.
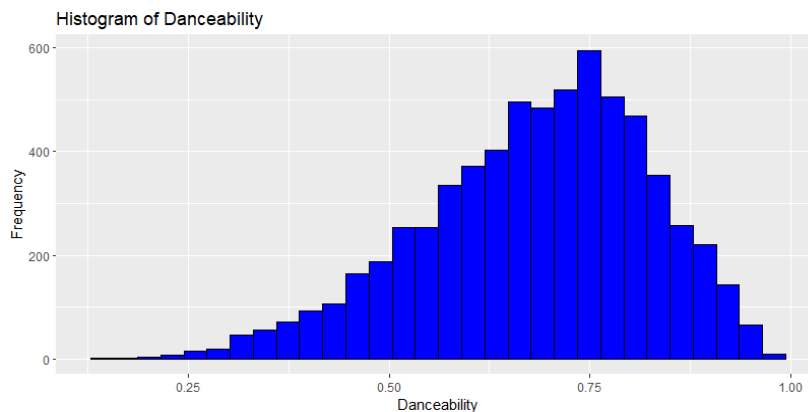
This histogram graph shows the distribution of the danceability score across the tracks. As a high frequency score indicates the tracks that are danceable. As we see most of the songs that are danceable are around 0.60 to 0.75.
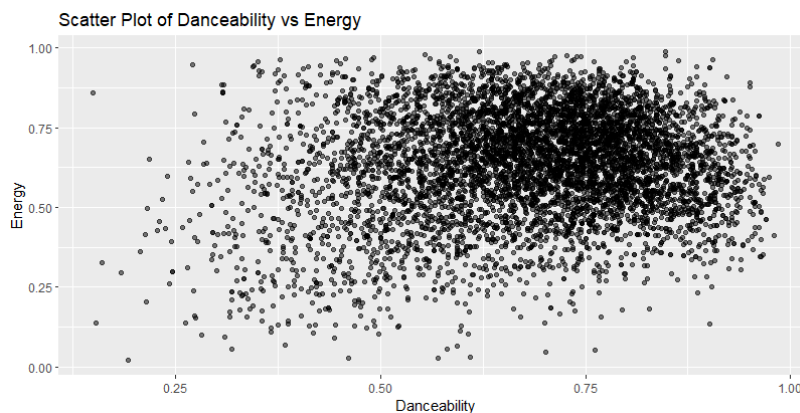
The scatterplot to the left depicts the relationship between danceability and the energy level of the song. The cluster might indicate that more danceable tracks have more energy. The visualization may also indicate that the energy level and danceability of a track tend to coincide.

Section 3. Hypotheses
Danceability and Streams
H_0: A song's danceability has no relationship to the number of times the song has been streamed.

H$_a$: There is a positive relationship between the danceability of a song and the number of times the song has been streamed.

Tempo and Streams
H$_0$: There is no positive relationship between the tempo and the number of times a song has been streamed.
H$_a$: There is a positive relationship between the tempo and the number of times a song has been streamed.

Acousticness and Danceability
H$_0$: There is no relationship between the acousticness of a song and the danceability of a song.
H$_a$: There is a relationship between the acousticness and dancability of a song.

Section 4. Description of the data analysis procedure and methods
For our data analysis procedures and methods we identified key topics/ areas of interest that we wanted to explore. From there we cleaned the data set which consisted of dropping missing values, irrelevant columns, and duplicate values. We also identified descriptive statistics for variables of interest as well as identified relationships through significance and correlation testing. From there we were then able to visualize and interpret relationships between variables that we intended to focus on

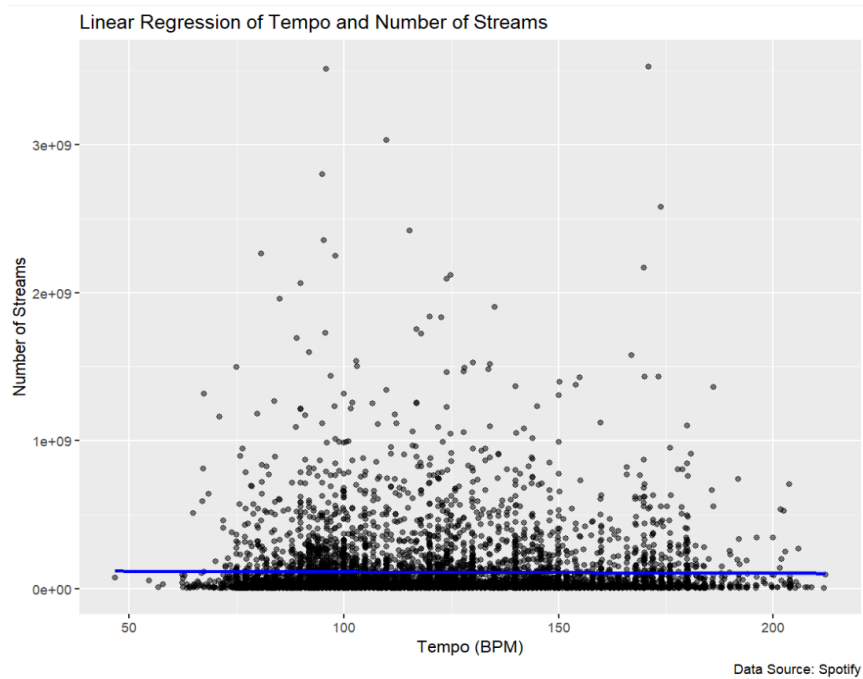**Section 5. Test results**
Danceability and Streams
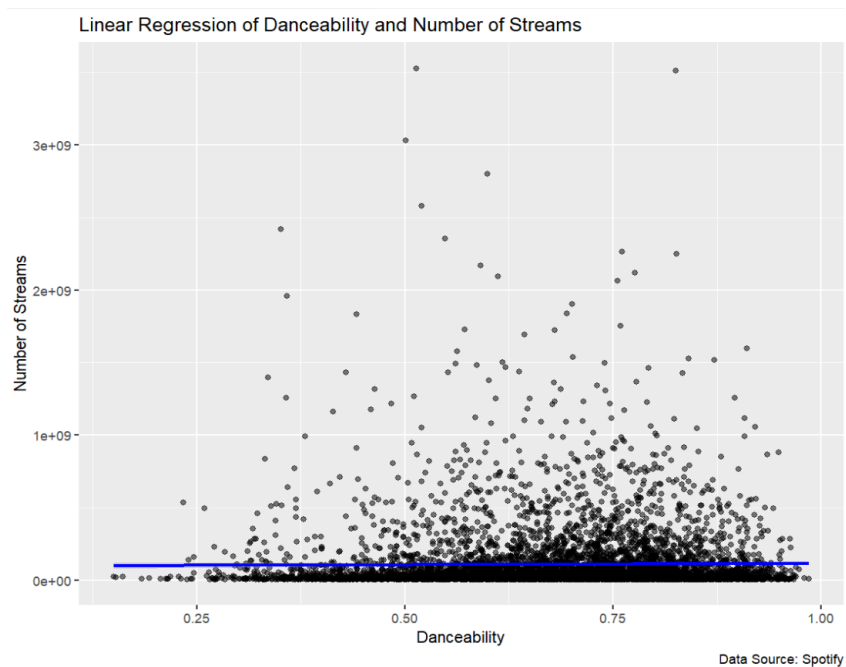
Scatter plot of observations of Danceability levels and Number of Streams a song has. Linear regression is overlaid on the plot in blue.

Pearson's product-moment correlation

t	df	p-value	95% CI	cor
0.93211	6511	0.3513	-0.012, 0.035	0.0115

Tempo and Streams

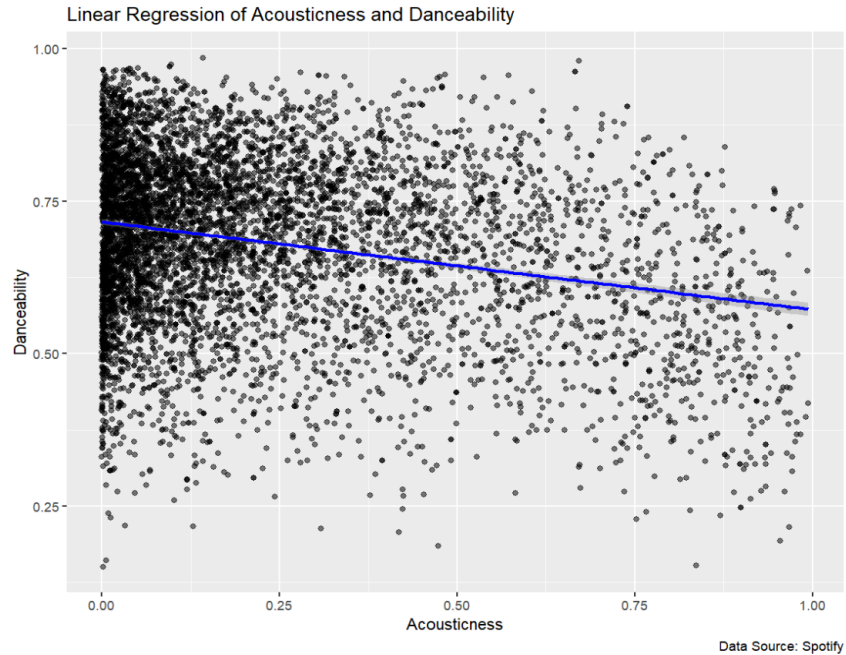

Linear Regression of Tempo and Number of Streams

Scatter plot of observations of Tempo and Number of Streams a song has. Linear regression is overlaid on the plot in blue.

Pearson's product-moment correlation

t	df	p-value	95% CI	cor
-0.97254	6511	0.3308	-0.036, 0.012	0.012

Acousticness and Danceability

After plotting the relationship between Acousticness and Danceability of a song.

Scatter plot of observations of Acoustic levels and Danceability level of songs. Linear regression is overlaid on the plot in blue.

Pearson's product-moment correlation

t	df	p-value	95% CI	cor
-20.696	6511	< 2.2e-16	-0.271, -0.225	-0.248

Section 6. Interpretation of the results and suggestion for decision making

Danceability and Streams
By performing a Pearson's correlation test on danceability and the number of streams a song has, we can observe that there is little to no correlation between the two variables, with an r value of 0.0115 .With a p-value of 0.3513, we fail to reject the null hypothesis and conclude that there is not a significant relationship between the danceability of a song and the number of streams the song has.

Tempo and Streams
From the Pearson's correlation test, we can observe that there is little to no correlation between tempo and the number of streams with an r value of 0.012. Additionally, the p-value of the test is 0.3308, indicating there is not a significant relationship between the two variables. Based on the correlation test, we cannot reject the null hypothesis that there is no relationship between tempo and the number of streams a song has. After analyzing the two variables, we cannot recommend that the tempo of a song can impact how many streams it accrues.

Acousticness and Danceability

From the Pearson's correlation test, we can observe a strong negative relationship between the acousticness and danceability of a song. The p-value of < 2.2e-16, indicates that we reject the null hypothesis and can conclude with 95% certainty that there is a relationship between the acousticness and danceability of a song. The negative t value indicates that this relationship is negative. In other words as a song's acousticness decreases the danceability of the song increases. Given this information, we are able to help consult artists who want to make dancy pop songs and advise them to keep minimal acoustics in their new tracks.