1. **Motivation**:
   Music is a large part of today's culture and society. A lot of influence and power comes with the popularity of creating a 'hit.' Many musical aspects come together to produce a song. The aim of this study is to identify which musical variable is most influential in making a song popular or not. Once identified, more musicians can place emphasis on this variable to increase their fame and influence. It would change the way a lot of people may produce music by making it more formulaic. Our research question is "What is the single most influential musical variable in making a song popular?"

2. **Data Description**:
   Sources were derived from https://www.kaggle.com/theoverman/the-spotify-hit-predictor-dataset/data#dataset-of-10s.csv. Kaggle.com is a website full of data sets primarily used by universities and students.

   Our data set has an extensive number of predictor variables that are all relevant to the data. The original source had 6398 observations, which was too large for the SAS system to process, so we limited it to the first 251 observations. The dataset was compiled using the Spotify Developer feature, which created the variables and mathematically calculated the values for every song. Descriptions of these 19 variables and how they are calculated, as described by the Spotify website, are listed below. The file was compiled using tracks from the last decade.
   Non numeric identifiers:

- "Track: The name of the track"
- "Artist: The name of the first artist listed for the track"
- "Uri: The resource identifier for the track."

  Predictor variables:

- "Danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable."
- "Energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy."

- “Key: The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C?/D?, 2 = D, and so on. If no key was detected, the value is -1.”
- “Loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.”
- “Mode: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.”
- “Speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.”
- “Acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.”
- “Instrumentalness: Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.”
- “Liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.”
- “Valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).”
- “Tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.”
- “Duration_ms: The duration of the track in milliseconds.”

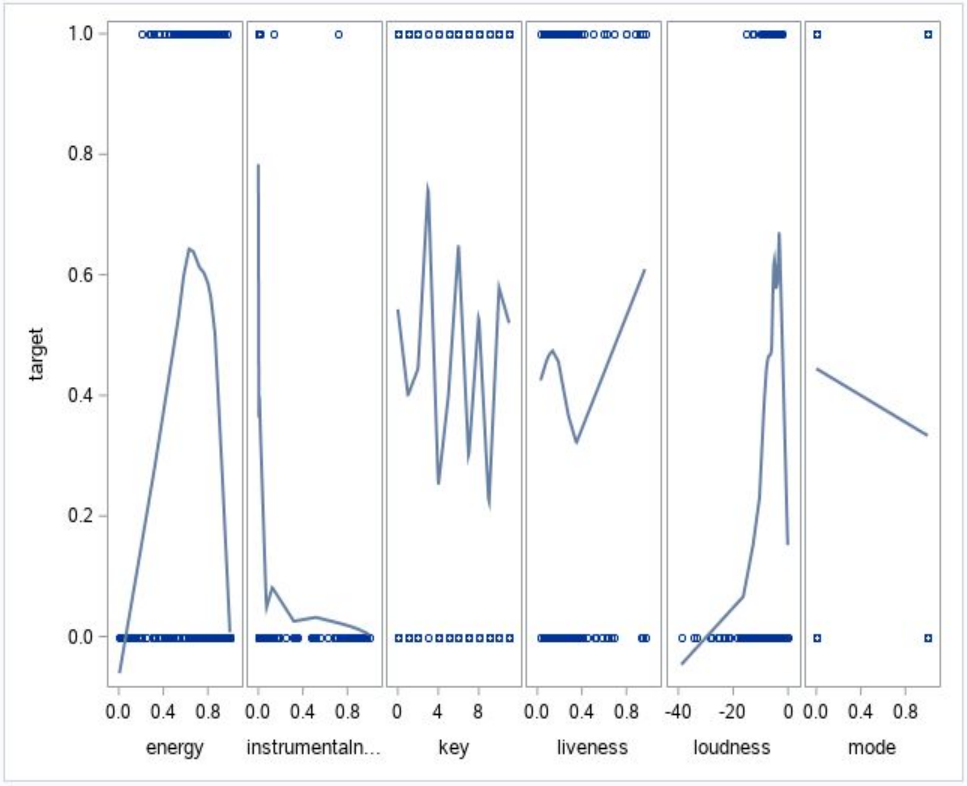
- “Time_signature: An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).”
- “Chorus_hit: This the the author's best estimate of when the chorus would start for the track. Its the timestamp of the start of the third section of the track (in milliseconds). This feature was extracted from the data received by the API call for Audio Analysis of that particular track.”
- “Sections: The number of sections the particular track has. This feature was extracted from the data received by the API call for Audio Analysis of that particular track.”
- “Target: The target variable for the track. It can be either '0' or '1'. '1' implies that this song has featured in the weekly list (Issued by Billboards) of Hot-100 tracks in that decade at least once and is therefore a 'hit'. '0' Implies that the track is a 'flop'.

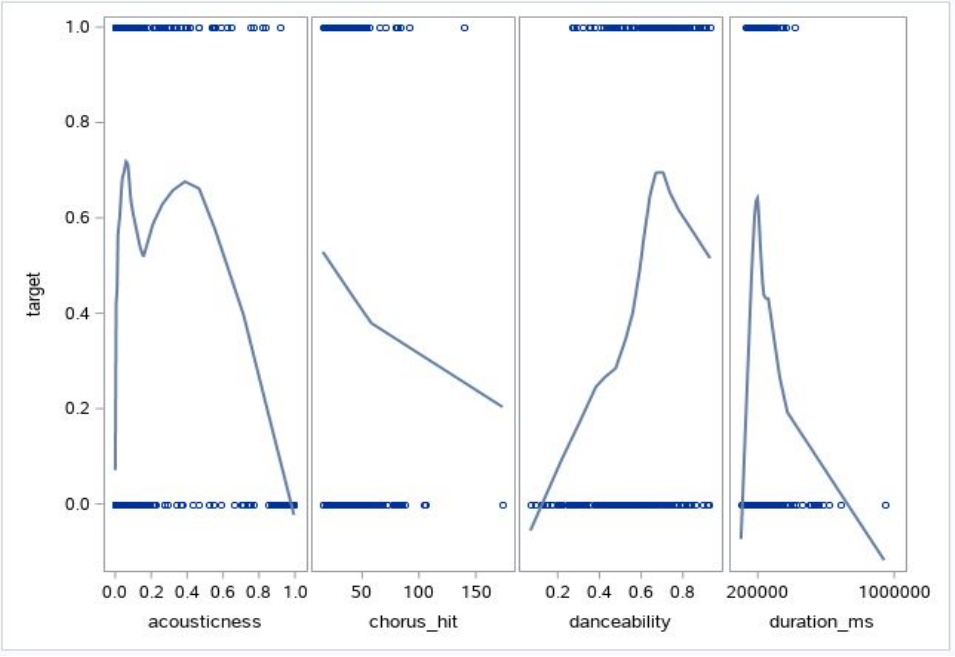
  The author's condition of a track being 'flop' is as follows:
  - The track must not appear in the 'hit' list of that decade.
  - The track's artist must not appear in the 'hit' list of that decade.
  - The track must belong to a genre that could be considered non-mainstream and / or avant-garde.
  - The track's genre must not have a song in the 'hit' list.
  - The track must have 'US' as one of its markets.”

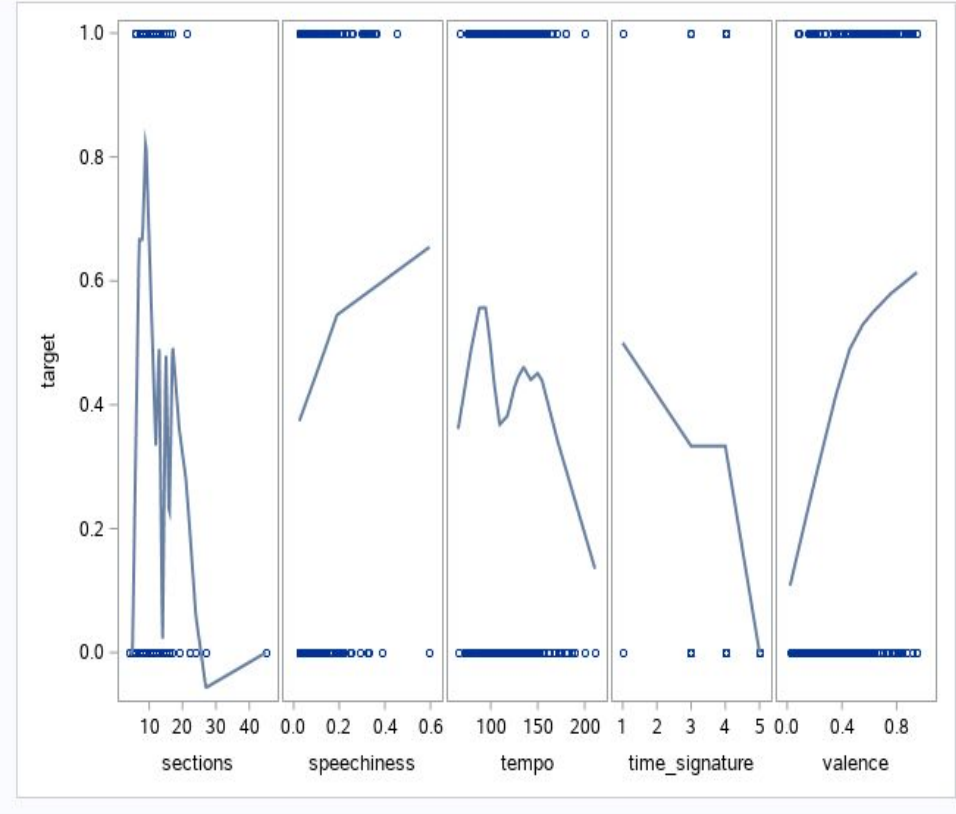
3. **Data Exploration**:
   Our data was stored in a csv file. We manipulated the data by reducing the data set to the first 251 observations. The original CSV files contained over 6000 observations. We also dropped the variable cod_centro, as there was no explanation for it on the Spotify website.

```
FILENAME REFFILE '/home/u42230827/assignments/dataset-of-10s.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=songs10;
    GETNAMES=YES;
RUN;

/*limit observations to first 250. Original dataset contained over 6000*/
data songs11;
set songs10 NOBS=COUNT;
if count - _n_ > 250 then delete;
drop cod_centro;
run;

proc contents data=songs10;

/** Print the results. **/
PROC PRINT DATA=songs11; RUN;
```

SGSCATTER Plots w/ LOESS Smooth:







From the panel displays above, a logistic regression does appear to be an effective way to predict song popularity across certain variables. While the LOESS curve is jagged among the variables key, liveness, mode, sections, sand tempo, the other variables are more promising. Due to the nature of music taste, it is expected that some variable's influence on popularity exists on a bell curve. This is extremely evident among the variables energy, loudness, duration_ms, and acousticness to a slightly lesser extent. The variable valence has a promising LOESS curve

4. **Model Fitting & Analysis**:

A. Logistic Regression

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.3574	0.2931	21.4524	<.0001
valence	1	2.4576	0.5621	19.1153	<.0001

The parameter estimate is 2.4576. Therefore, having a high valence increases the odds of popularity ('hit').
The prediction equation: f(x)=-1.3754+2.4576x

B. Statistical significance

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	20.8089	1	<.0001
Score	20.2604	1	<.0001
Wald	19.1153	1	<.0001

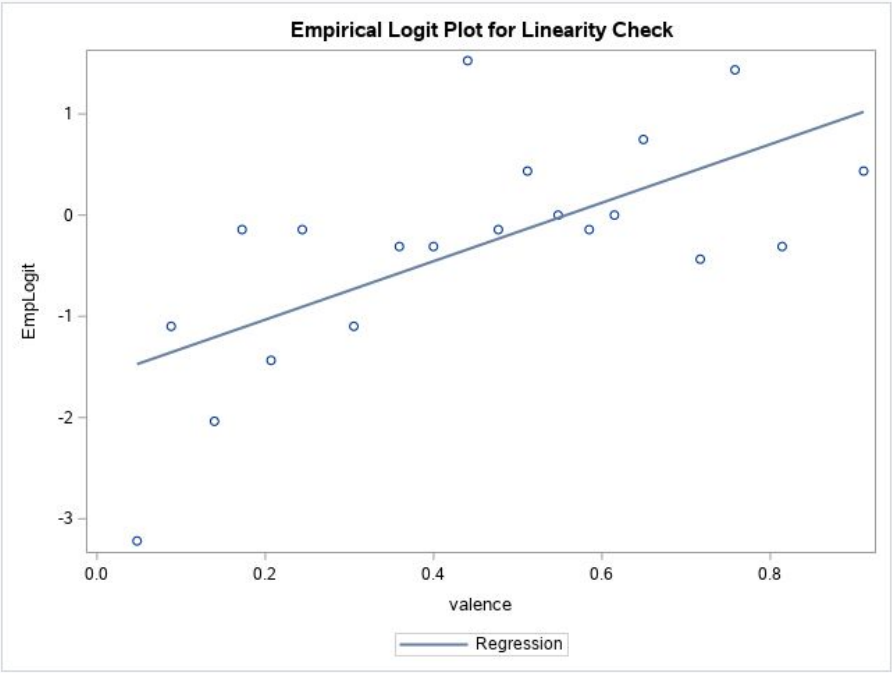
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.3574	0.2931	21.4524	<.0001
valence	1	2.4576	0.5621	19.1153	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
valence	11.677	3.880	35.140

Valence has a significantly low p value and has a wald chi-square of 19.1153. 1 is not contained in the confidence interval which makes the variable significant.

C. Empirical Logit Plot

We set the number of bins equal to 20 to ensure accuracy of the logit plot.

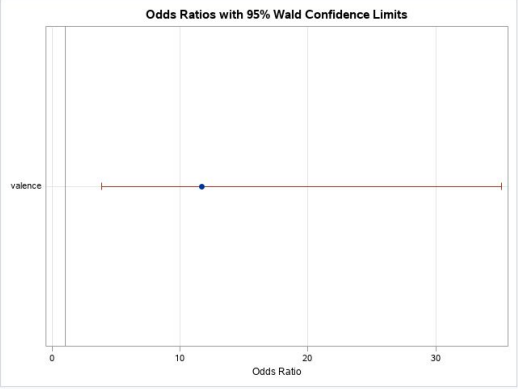


The relationship between valence and popularity appears to be roughly linear. The points are scattered around the line.

D. 95% Confidence Interval for resulting Odds Ratio

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	66.6	Somers' D	0.333
Percent Discordant	33.3	Gamma	0.333
Percent Tied	0.1	Tau-a	0.165
Pairs	15540	c	0.666

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
valence	1.0000	11.677	3.880	35.140



In the odds ratio graph, the reference line of the null hypothesis is not crossed, so the variable is significant. Also, the confidence limits do not contain 1. For each 1 unit increase in valence, popularity increases by 11.6%.

E. Prediction model of probability of success with valence at .9

ət	_LEVEL_	phat
.	1	0.70150

The phat value indicates the probability of a song being popular ('hit') with a valence value of .9, is 70.150%

Obs	B0	B1	p	logit	X
1	-1.3574	2.4576	0.5	0	0.55233

In order to obtain a probability of .5, the value of valence should be .55233

I. Multicollinearity

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.37053	0.37297	3.67	0.0003	0
acousticness	1	-0.32386	0.14783	-2.19	0.0295	3.33386
chorus_hit	1	-0.00122	0.00147	-0.83	0.4067	1.28079
danceability	1	0.04523	0.19855	0.23	0.8200	2.16272
duration_ms	1	-4.29814E-7	5.112707E-7	-0.84	0.4014	3.62905
energy	1	-1.00760	0.24469	-4.12	<.0001	5.56559
instrumentalness	1	-0.48130	0.10168	-4.73	<.0001	1.64347
key	1	0.00123	0.00725	0.17	0.8654	1.08695
liveness	1	0.03159	0.15124	0.21	0.8347	1.16811
loudness	1	0.03483	0.00954	3.65	0.0003	4.10257
mode	1	-0.03116	0.05685	-0.55	0.5841	1.11218
sections	1	-0.00578	0.01228	-0.47	0.6384	3.39014
speechiness	1	0.41985	0.32424	1.29	0.1966	1.20676
tempo	1	-0.00176	0.00096367	-1.83	0.0688	1.15962
time_signature	1	0.11517	0.06908	1.67	0.0968	1.21656
valence	1	0.21930	0.13527	1.62	0.1063	1.62612

The predictors have relatively average multicollinearity. However, energy has a VIF value over 5, so we should eliminate it.

J. variable selection model

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	sections	1	14	0.0195	0.8890
2	key	1	13	0.1082	0.7422
3	danceability	1	12	0.1152	0.7343
4	liveness	1	11	0.1573	0.6917
5	mode	1	10	0.4072	0.5234
6	chorus_hit	1	9	0.6211	0.4306
7	speechiness	1	8	1.6467	0.1994
8	duration_ms	1	7	1.7667	0.1838
9	time_signature	1	6	1.8924	0.1689
10	tempo	1	5	3.0551	0.0805
11	acousticness	1	4	3.6168	0.0572

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	5.8351	1.3911	17.5945	<.0001
energy	1	-5.3899	1.2780	17.7880	<.0001
instrumentalness	1	-6.6179	2.3575	7.8803	0.0050
loudness	1	0.4009	0.0884	20.5463	<.0001
valence	1	1.8014	0.7082	6.4707	0.0110

Using a backward variable selection method, the variables of sections, key, danceability, liveness, mode, chorus_hit, speechiness, duration_ms, time_signature, tempo, and acoustic ness were removed. Our best model includes energy, instrumentalness, loudness, and valence.

K. Evaluate best model for significance

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	5.8351	1.3911	17.5945	<.0001
energy	1	-5.3899	1.2780	17.7880	<.0001
instrumentalness	1	-6.6179	2.3575	7.8803	0.0050
loudness	1	0.4009	0.0884	20.5463	<.0001
valence	1	1.8014	0.7082	6.4707	0.0110

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
energy	0.005	<0.001	0.056
instrumentalness	0.001	<0.001	0.136
loudness	1.493	1.256	1.776
valence	6.058	1.512	24.275

All of the variables have p-values less than .05. In addition, the confidence limits all do not contain 1. This indicates significance.

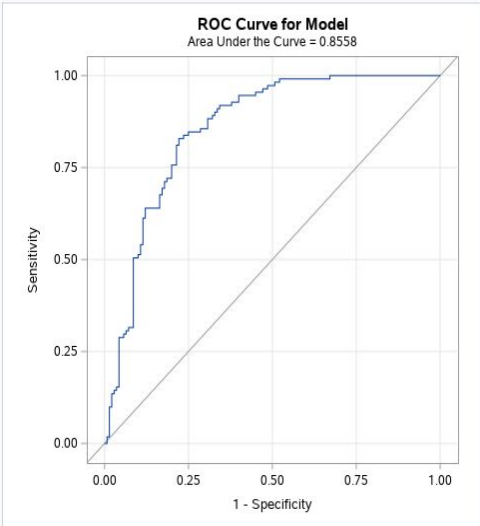
L. Predicted probability of success for variable values in best model

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	5.4573	1.5084	13.0890	0.0003
energy	1	-4.6278	1.3404	11.9201	0.0006
instrumentalness	1	-5.6135	2.0161	7.7527	0.0054
loudness	1	0.3904	0.0983	15.7611	<.0001
valence	1	1.3560	0.7826	3.0025	0.0831

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
energy	0.010	<0.001	0.135
instrumentalness	0.004	<0.001	0.190
loudness	1.478	1.219	1.792
valence	3.881	0.837	17.990

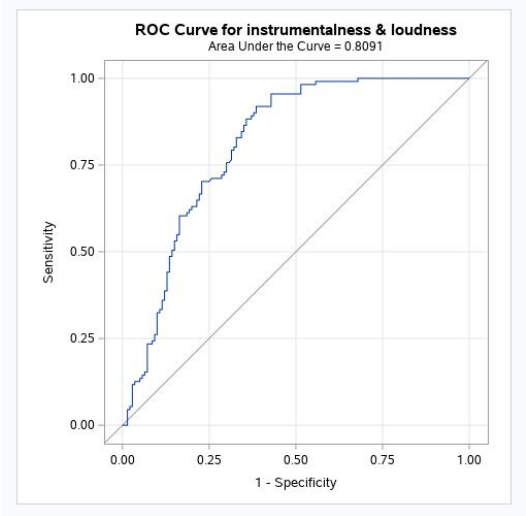
After validating and scoring the data, all of the variables remained significant besides valence. Valence now has a p value greater than .05 and 1 is contained in its confidence interval.

M.ROC Curves



The curve for the best model:

Model we chose to test.



Between the two ROC curves, we can tell that both models are close to 1 but our best model which includes; energy instrumentalness loudness and valence has slightly higher area under the curve. Removing energy and valence was not beneficial overall.

5. **Conclusion**:

Our results indicate that the full model, which includes the variables energy, instrumentalness, loudness, and valence, are more predictive of whether or not a song is a hit than instrumentalness and loudness alone. This makes sense because music is dynamic and involves the interaction of all its components. If someone were trying to create a popular song ('hit'), they should focus their attention on the four variables of energy, instrumentalness, loudness, and valence specifically. In addition, our predicted probabilities of success would help indicate the level each variable should be calculated to in order to further ensure the production of a hit.

We, the project team members, certify that the percentage of the effort listed by each of our names below is an accurate account of the original effort contributed by each team member in the producing of this project and report.

Claire Gambacorta 25%
Tyler Rose 25%
Nicholas Blanton 25%
Dylan Smith 25%