



Building the Perfect Song

By Asher Palmer

The Dataset

I found this dataset while looking for sets on music. I found a subset of the Million Song Dataset. This subset is 10000 songs or 1% of the original dataset. Originally, I did not want to use this dataset as it had many missing values and needed quite a bit of cleaning.

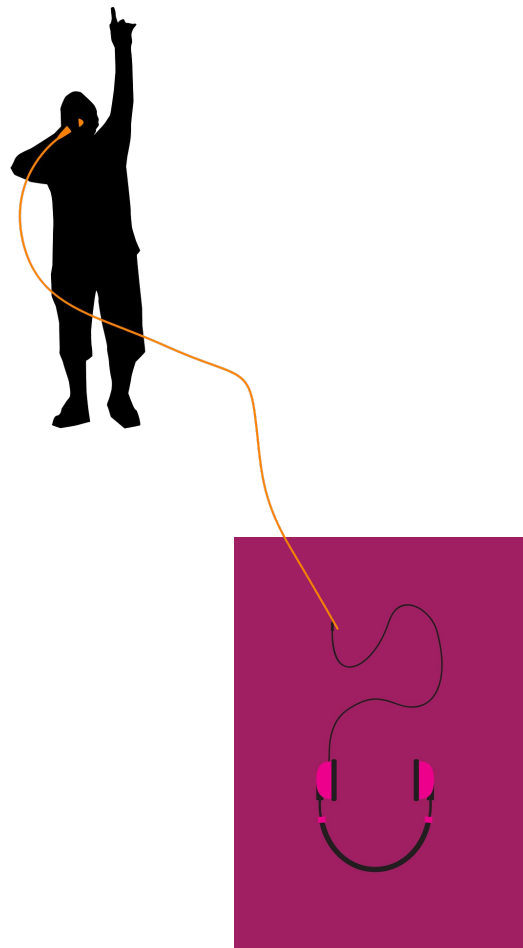
The Music dataset has a number of great attributes such as artist hotness, song hotness, location, latitude & longitude, key the song was made in, the mode, the familiarity, the duration as well as the artist name and song title.

The data ranges from year 0 (these songs did not have input for year) through 2010 and is 10000 rows x 35 columns.

Why Did I Choose This Dataset?

I choose this dataset because I have a very personal connection with music. I've played in band in school and have seen and experienced the way it moves people and can completely change their outlook and provide a safe space.

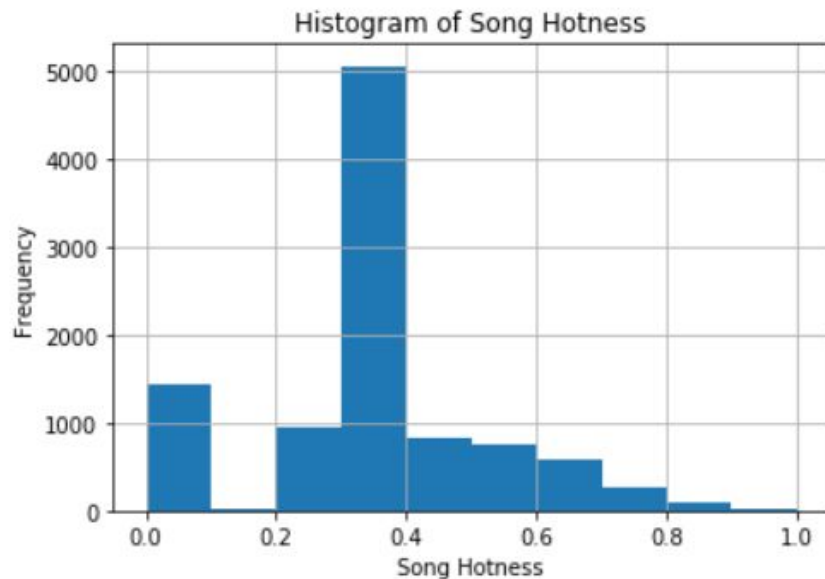
Because music is so influential, it has become a multibillion dollar business and being able to produce the perfect song (particularly multiple songs on a consistent basis) can be very lucrative.



The Parts and Pieces

Many of the features are continuous. One of which, we will focus on is 'song.hottnesss'. This variable is how we define a great song.

As you can see with the histogram to the right, the majority of songs have a rating of 0.3. However, we want a song that scores 1.0 consistently.



Attributes Of A Perfect Song

Looking at the data, there are only two songs in my dataset that have a rating for song hotness of 1.

Nothin' On You by B.o.B. made in 2010. The artist hotness is 0.71253495 and plays for 269.6355 seconds. The familiarity is 0.768224 and played in the key of 10 (B flat). The tempo is 104.038 and is -5.388 decibels loud.

Immigrant Song by Led Zeppelin made in 1970. The artist hotness is 0.63441239 and plays for 145.0575 seconds. The familiarity is 0.787098 and played in the key of 11(B). The tempo is 150.569 and is -10.544 decibels loud.

The Eye Ball Comparison

Just looking at these comparisons, you would think that there isn't much between these two songs other than a similar familiarity and key that they are played in. They are played in different countries and have so many more dissimilarities that I'm not sure data science will find a way to find the perfect song.

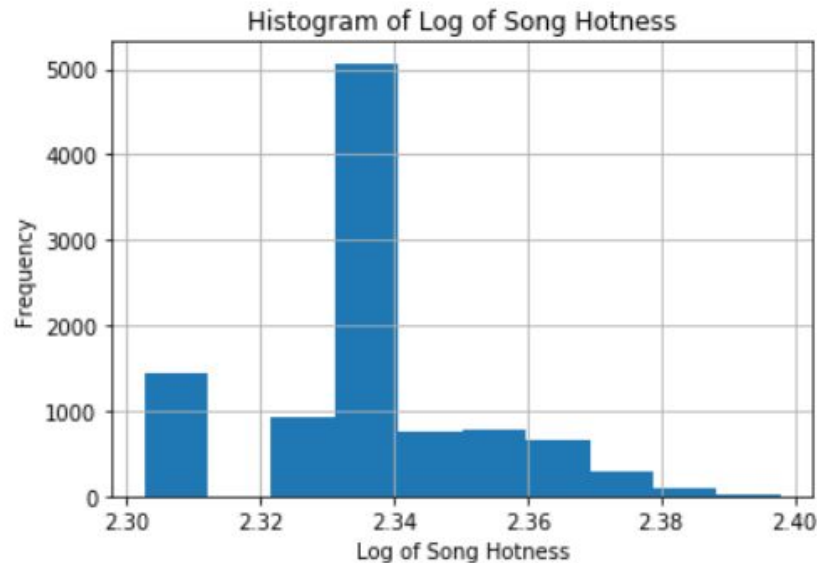
Lucky for us, we have technology to analyze data and be able to make predictions about future songs.



Normalizing the Data

I thought about using a log version of my target variable as a means of making the distribution more normal but all I got was essentially the same distribution, which is close to normal. That's why I decided to use the normal version of my variable.

I did decide to use a variable in my classifications called 'song_hottness' that split my data into songs with a hotness of 0.5 or greater.



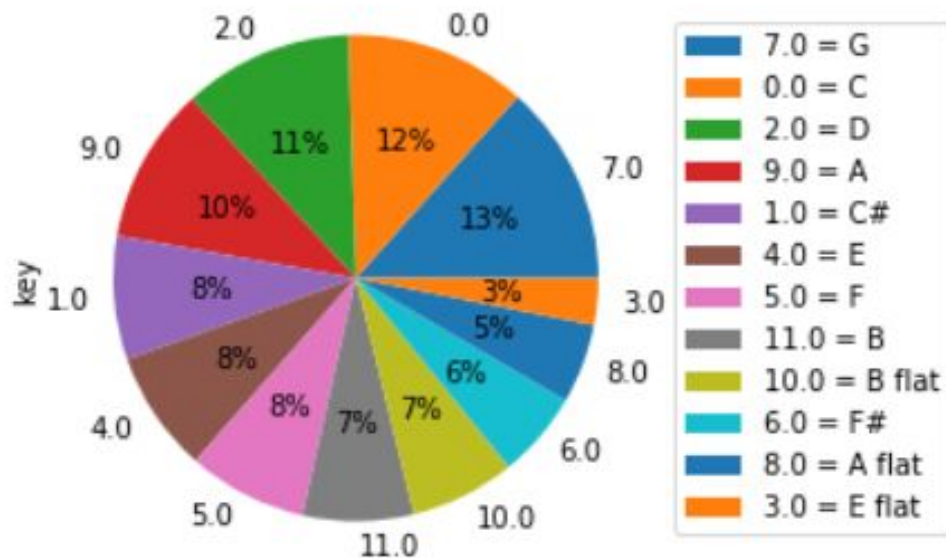
The Frequently Used Key

The key that is used the most is:



Followed by; C, D, & A

Pie Chart of Keys of Songs



Why Is G The Most Used Key?

According to Spotify, “More mysterious, at first glance anyway, is the order of the notes. Why is G Major the top key on all of Spotify? And why is C Major number two?”

Much like electricity going through a circuit, songwriters often take the path of least resistance. On a keyboard or a guitar — both incredibly popular instruments for composing western music — that path is through G Major.”

Put simply:

E is convenient for guitar, but not piano.

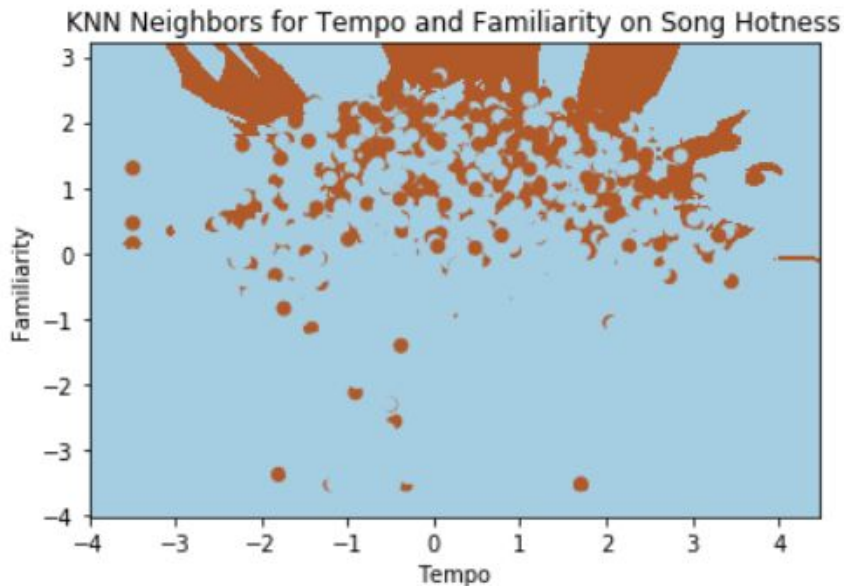
C is convenient for piano, but not guitar.

G is convenient for both guitar and piano.



Let the Testing Begin

Let's take a look at KNN. Looking at Tempo and Familiarity, I wanted to see if there were any clear and decisive lines of what to call a “good song”. As you can see, there's no real correlation between tempo and song hotness but the higher the familiarity, the greater the occurrence of song hotness.



Model Selection

Here are my cross-validation scores.
My model has an average of 83.40%
and features of importance showing
as artist hotness, familiarity and the
year.

```
ANOVA Cross-Validation Scores: [0.83316683 0.83483483]
```

```
ANOVA Cross-Validation Score Averaged Across Folds: 83.40%.
```

```
Selected Features: ['artist.hotttnesss', 'familiarity', 'year']
```

Random Forest

Using Random Forest, my model performs at around 82% and took 183 seconds.

However, my classifier claims around 93% accuracy.

```
Decision Tree Depth: 1  
Model Performance: 82.87%.
```

```
Decision Tree Depth: 2  
Model Performance: 82.87%.
```

```
Decision Tree Depth: 3  
Model Performance: 82.34%.
```

```
Decision Tree Depth: 4  
Model Performance: 83.00%.
```

```
Decision Tree Depth: 5  
Model Performance: 82.65%.
```

```
Decision Tree Depth: 6  
Model Performance: 82.67%.
```

```
Decision Tree Depth: 7  
Model Performance: 82.79%.
```

```
Decision Tree Depth: 8  
Model Performance: 82.30%.
```

```
Decision Tree Depth: 9  
Model Performance: 82.24%.
```

```
Decision Tree Depth: 10  
Model Performance: 81.32%.
```

```
--- Runtime: 183.56333139999992 seconds. ---
```

```
array([0.95704296, 0.92807193, 0.94505495, 0.95      , 0.914      ,  
       0.93      , 0.917      , 0.92092092, 0.93093093, 0.97497497])
```

The Ridge Regression was only 17% accurate.

R-squared simple Ridge Regression model:
0.17392810393574742

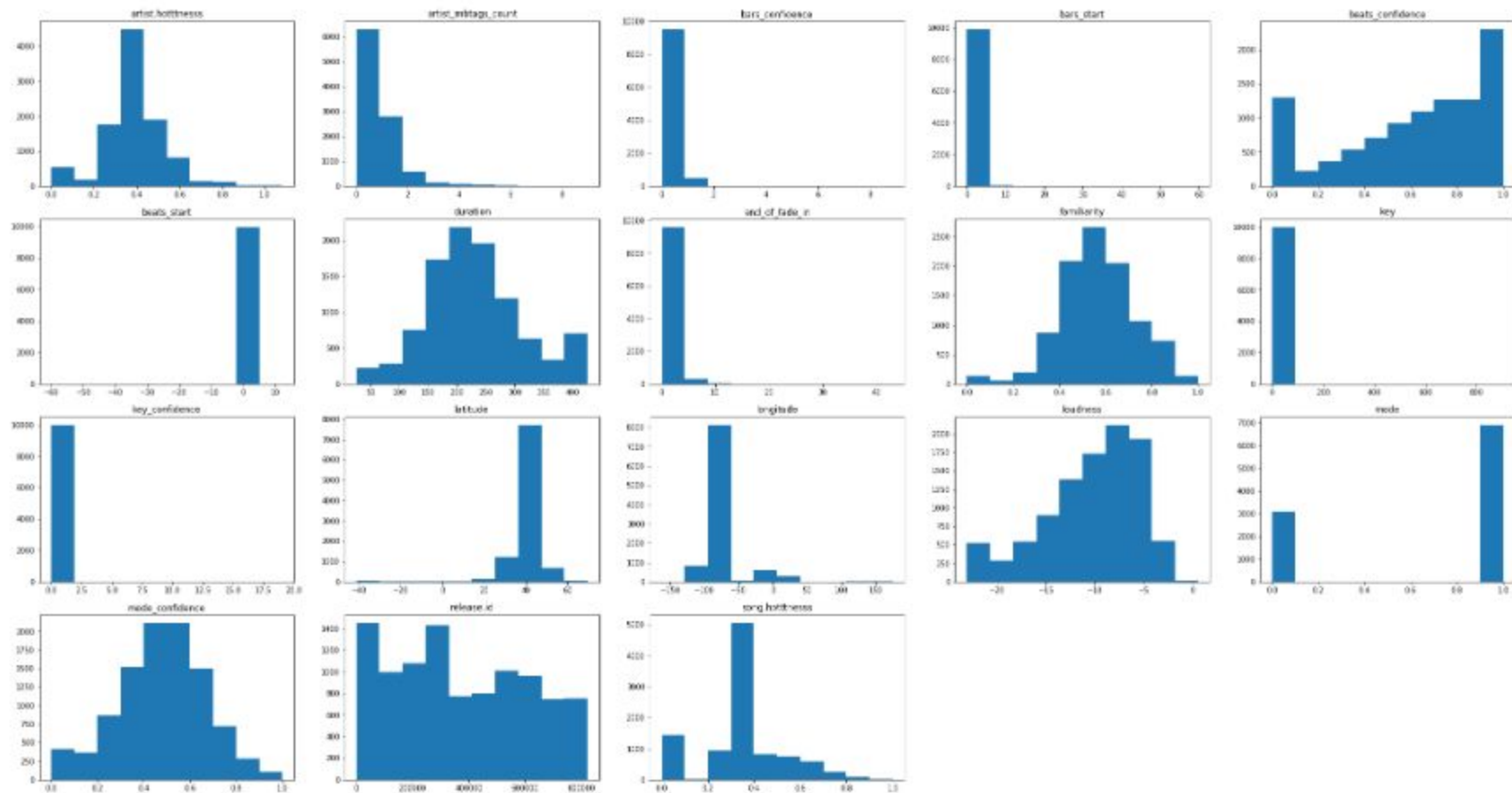
The Lasso Regression was only 9% accurate.

R^2 for the Lasso Regression model:
0.09329011267733589



Subplots

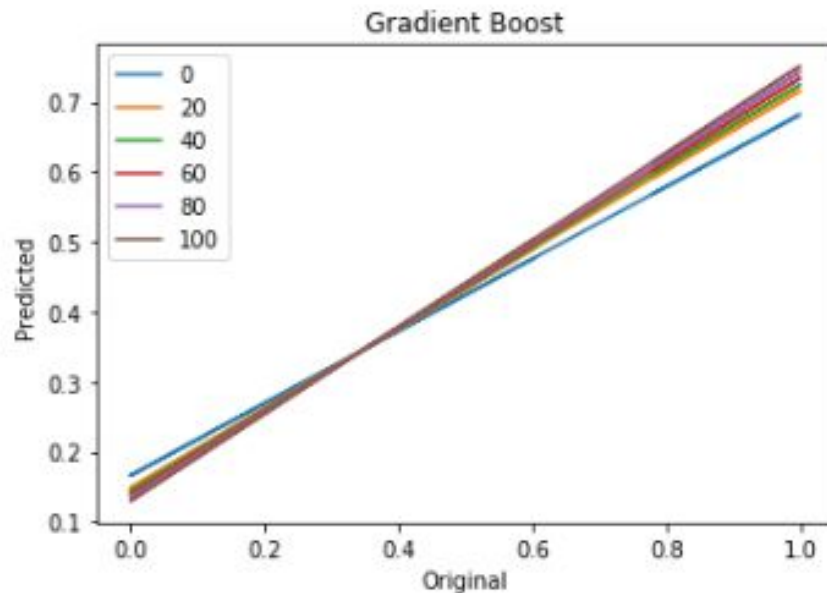
On the following slide are the subplots of my numeric variables. The features have mostly normal distributions but there are some variables that are skewed. I did consider winsorizing these skewed variables in an attempt to normalize them. However, I decided not to as a means of “preserving” the original data.



Model Selection

From the Gradient Boosting Graph, the accuracy is 79.22%. That is the poorest of my models.

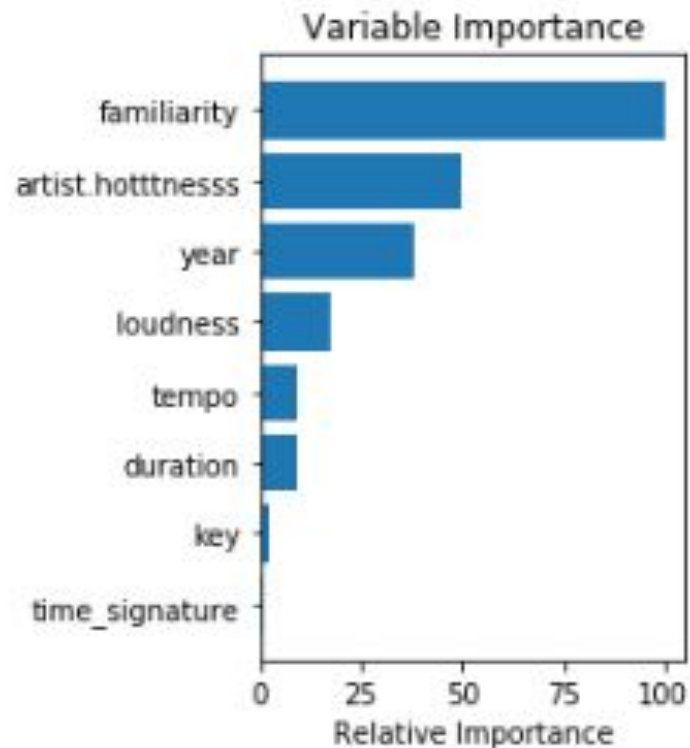
```
Weak learner 0 R^2: -8.109344347702944  
Weak learner 20 R^2: -0.0019854999555357633  
Weak learner 40 R^2: -0.002275624819774391  
Weak learner 60 R^2: -0.0016407235871092052  
Weak learner 80 R^2: -0.0019533867661012394  
Weak learner 100 R^2: -0.0032707399806952697
```



Overall R^2:
0.7922629924941085

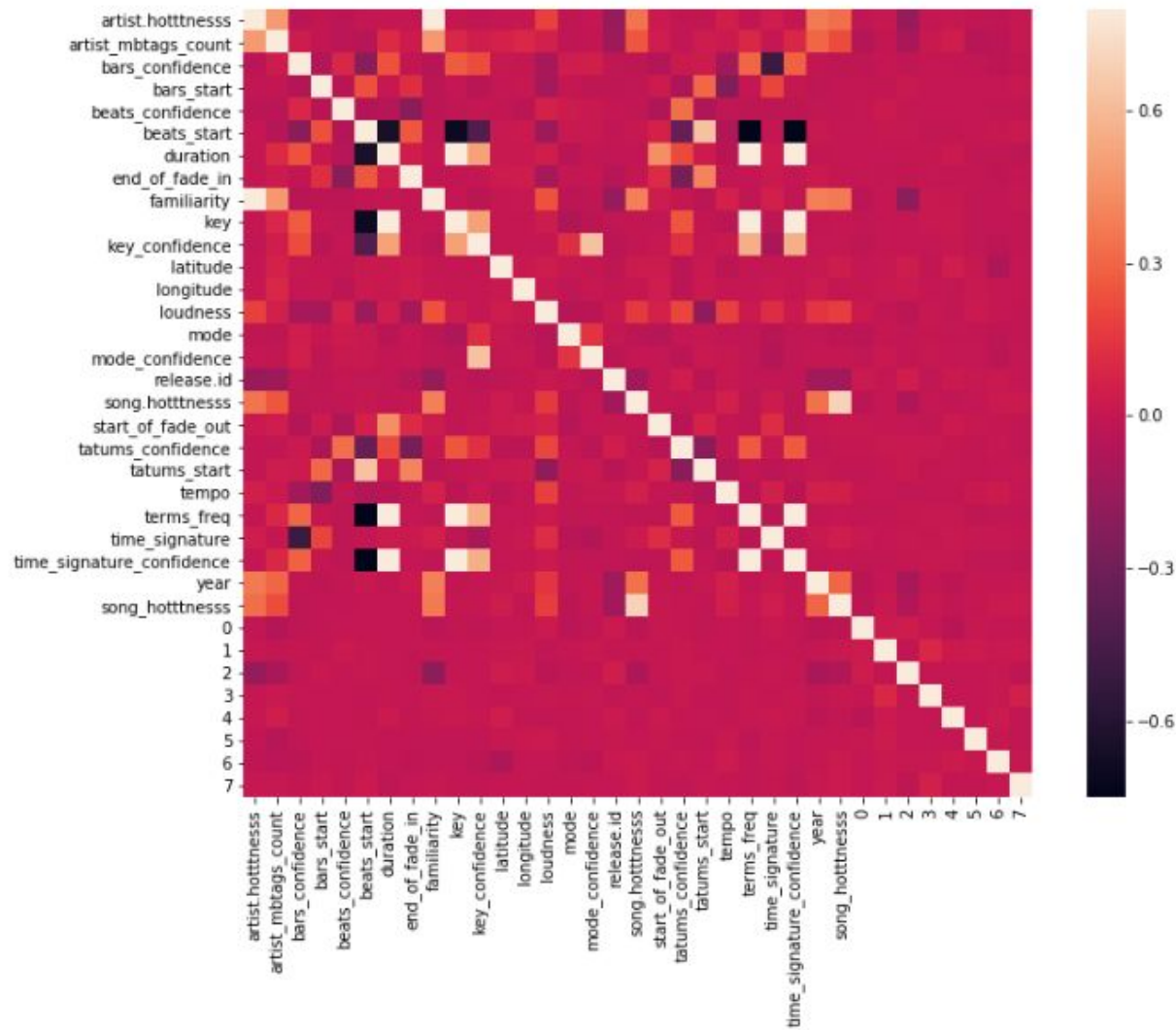
Variable Importance

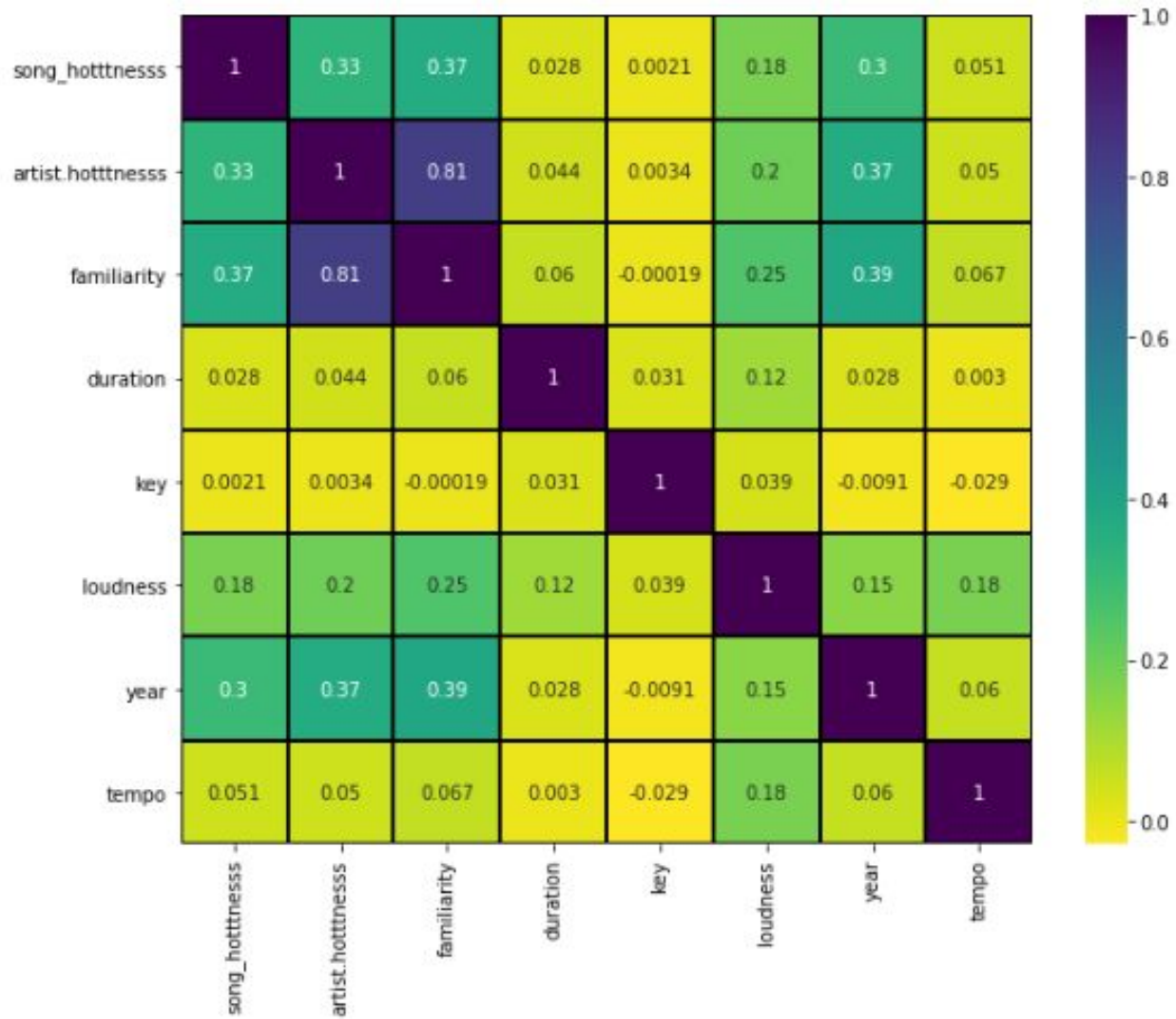
As you can see, familiarity, year and artist.hottnesss are the top three important features. I thought the duration and key of a song would be more influential on song hotness.



Heatmap

On the next slide, you will find the heatmap. It's filled with over 40 features so it's hard to find `song.hottnesss` and follow along the axis. However, `song.hottnesss` is highly correlated with `artist.hottnesss`, familiarity, and year with some slight correlation to loudness. This is all shown with the variable importance image from the last slide but with more pizzazz.





StatsModel Regression

I have included a statsmodel regression for those that would like a closer look at the numbers and how I am forming my hypothesis.

Optimization terminated successfully.
Current function value: 0.357650
Iterations 19

Logit Regression Results

```
=====
Dep. Variable:      song_hotttnesss      No. Observations:      8000
Model:              Logit                Df Residuals:           7966
Method:              MLE                  Df Model:                33
Date:               Tue, 03 Sep 2019      Pseudo R-squ.:          0.2222
Time:               22:47:04              Log-Likelihood:         -2861.2
converged:          True                  LL-Null:                -3678.6
                                      LLR p-value:                0.000
=====
```

```
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
artist.hotttnesss      1.1166      0.465      2.402      0.016      0.206      2.028
artist_mbtags_count    -0.0900      0.038     -2.377      0.017     -0.164     -0.016
bars_confidence         0.0122      0.148      0.083      0.934     -0.278      0.302
bars_start              0.0227      0.027      0.851      0.395     -0.030      0.075
beats_confidence       -0.0656      0.112     -0.585      0.559     -0.285      0.154
beats_start            -0.1873      0.236     -0.792      0.428     -0.651      0.276
duration                0.0203      0.005      3.881      0.000      0.010      0.030
end_of_fade_in         -0.0159      0.021     -0.757      0.449     -0.057      0.025
familiarity             5.4545      0.448     12.177      0.000      4.577      6.332
key                    -0.0088      0.010     -0.920      0.357     -0.028      0.010
key_confidence          0.0465      0.193      0.240      0.810     -0.333      0.425
latitude                0.0087      0.004      2.266      0.023      0.001      0.016
longitude              -0.0008      0.001     -0.730      0.466     -0.003      0.001
loudness                0.0536      0.008      6.796      0.000      0.038      0.069
mode                   -0.1169      0.073     -1.591      0.112     -0.261      0.027
mode_confidence         0.2164      0.281      0.770      0.441     -0.334      0.767
release.id             -5.273e-07      1.47e-07     -3.591      0.000     -8.15e-07     -2.4e-07
start_of_fade_out      -0.0209      0.005     -3.925      0.000     -0.031     -0.010
tatums_confidence       0.0537      0.120      0.448      0.654     -0.181      0.288
tatums_start           0.2321      0.243      0.956      0.339     -0.244      0.708
tempo                   0.0017      0.001      1.607      0.108     -0.000      0.004
terms_freq             -2.1828      0.663     -3.294      0.001     -3.482     -0.884
time_signature          0.0345      0.035      0.996      0.319     -0.033      0.102
time_signature_confidence 0.0077      0.097      0.079      0.937     -0.183      0.198
year                   0.0006      3.94e-05     15.564      0.000      0.001      0.001
0                      -0.0755      0.034     -2.250      0.024     -0.141     -0.010
1                      0.0454      0.033      1.366      0.172     -0.020      0.111
2                      0.0154      0.032      0.482      0.629     -0.047      0.078
3                      0.0053      0.034      0.158      0.875     -0.061      0.072
4                      -0.0385      0.033     -1.155      0.248     -0.104      0.027
5                      -0.0080      0.034     -0.238      0.812     -0.074      0.058
6                      0.0610      0.034      1.810      0.070     -0.005      0.127
7                      0.0736      0.033      2.202      0.028      0.008      0.139
intercept              -4.0590      0.738     -5.496      0.000     -5.506     -2.612
=====
```

Areas of Improvement

1. I could add more features to test different theories on song hotness
2. I could do more normalizing techniques for all of my variables
3. I could find more ways to reduce my bias towards the data

Conclusion

The perfect song::

Artist Hotness of: 0.3855

Familiarity: 0.5654557054013589

Made in the year: 2006

However, my research doesn't matter because music is something that is intangible and has unexplained qualities. The same song can produce multiple emotional reactions to different people. However, that does not mean hit songs cannot have a formula.

References

The Most Popular Keys of All Music on Spotify

<https://insights.spotify.com/us/2015/05/06/most-popular-keys-on-spotify/>