Can Mathematics Predict Popular Music?

Garrett Davidson
Bahrain School

Table of Contents

Introduction

The research was an examination of the mathematics behind popular Music Information Retrieval techniques and how effectively they predict popularity of songs, namely: do the numbers exhibit an identifiable pattern?

ScoreAHit[1], a company founded upon a similar goal, used mathematics and machine learning to compute an algorithm for predicting whether a song would become popular (i.e. make it to the top 10) following its release. Their research paper outlined their methods and was a large influence on this work. This, then, takes a step farther.

Their algorithm achieved approximately a 60% accuracy, likely because their data were limited to purely acoustical attributes. However, songs' lyrics, social media presence, and the previous popularity of their artists are also influential factors. This paper quantifies these factors to be used in tandem with ScoreAHit's algorithm in an attempt to produce even more accurate results.

By taking into account factors such as the repetitiveness of the lyrics, the uniqueness of the language used in them, the media buzz surrounding the artist at the time of the release and their relative popularities, as well as accounting for ScoreAHit's predictions, it was found that even more accurate predictions are possible. However, this was only achievable for recent songs. This result could likely be refined through learning machines similar to that used by ScoreAHit in order to account for shifts in the popularity of various attributes over time, similar to the changing weights seen in their algorithm.

The Examined Attributes

Three samples sets were used: Top 25 iTunes singles, bottom 25 iTunes singles (those ranked 176-200 on the iTunes charts) and a set of 25 randomly selected songs never reaching the iTunes charts. These three sets were chosen to indicate highly popular, mildly popular, and unpopular songs respectively. For each song, a total of ten attributes were calculated and recorded:

ScoreAHit score – ScoreAHit was put together by researchers studying Hit Song Science from a purely acoustical viewpoint. They produced an algorithm which could be applied to any song that would predict how likely the song is to become popular. The score from their algorithm was used as a line of comparison for the results of this research.

Words per Minute – This is the total number of words in a song divided by its duration in minutes. No attempt was made to account for changes in this value over time throughout the song. While songs may increase or decrease in pace throughout their duration,

it was believed that the average of these (i.e. the total WPM over the whole song) would be the most influential.

Repetitiveness – This is the total number of repeated words for the song. To calculate this value, the lyrics of each song were looped through and for each occurrence of a word which had been previously used in that song, this number increased by one.

Average Repetitiveness – This value was calculated by dividing the repetitiveness value by the total number of words in the song.

Uniqueness – Uniqueness was used as a counterpoint to repetitiveness. It is the total number of unique words in a song. It was calculated in a similar fashion to repetitiveness, except increasing for unique words rather than repeated ones.

Average Uniqueness – This is the uniqueness divided by the total number of words.

Average Commonness – To calculate average commonness, each word in the song was looked up in Wordcount's[2] database. Wordcount ranks words based on how often they are used in the English language. In their system, the most common word "the" is given a rank of 1, the next most common word "of" is ranked 2 and so on through nearly 100,000 different words. Average Commonness is the sum of the rank of each word (in the database) in a song divided by the total number of words (in the database). (See Normalizing the Lyrics section).

Average Unique Commonness – This is the same as the Average Commonness score, except only unique words are taken into account.

Average Artist Peak – The artist of each song was looked up on the Billboard Hot 100[3] list. Billboard Hot 100 has been the definitive US song ranking nearly since its inception in 1958. For each artist, a list was made of each of their previously charting songs. The peak position for each song on the Hot 100 list was then pulled and averaged, producing this value.

Artist's Previous Hits – This is the total number of songs from an artist which previously charted on the Hot 100 list.

The Program

A program was written to calculate all of these values. The program is fed a text file containing a list of songs and artists. It then proceeds to gather the rest of the required information. This includes querying AZLyrics[4] for the song lyrics, Wordcount for the word commonness, ScoreAHit for their score, and Last.fm[5] for the song durations. It then applies the necessary modifications to the lyrics (see the Normalizing the Lyrics section) and begins the calculations. When finished, it outputs the values in a CSV file at the specified location. Also, it outputs a list of all words, if any, which were unable to found in the Wordcount database. It was through this program that all data were collected for this paper.

Normalizing the Lyrics

Words such as names, brands, and non-English words were not present in the Wordcount database and thus ignored for the calculation of the commonness values. However, a few of the more commonly and universally understood foreign words (such as “cuatro” and “Inglés”) were simply translated into English in an attempt to produce a result more reflective of the entirety of the lyrics. All punctuation and formatting marks also had to be removed. In addition, all contractions and slang were expanded and corrected for the sake of these values. Across all of the songs tested, this totaled nearly 300 possible values to be searched for and replaced.

All attributes for songs were calculated after this transformation. This helped to ensure proper quantification of the lyrics, as it made certain that all of the words were in the correct form, equally devoid of punctuation, and readily able to be compared.

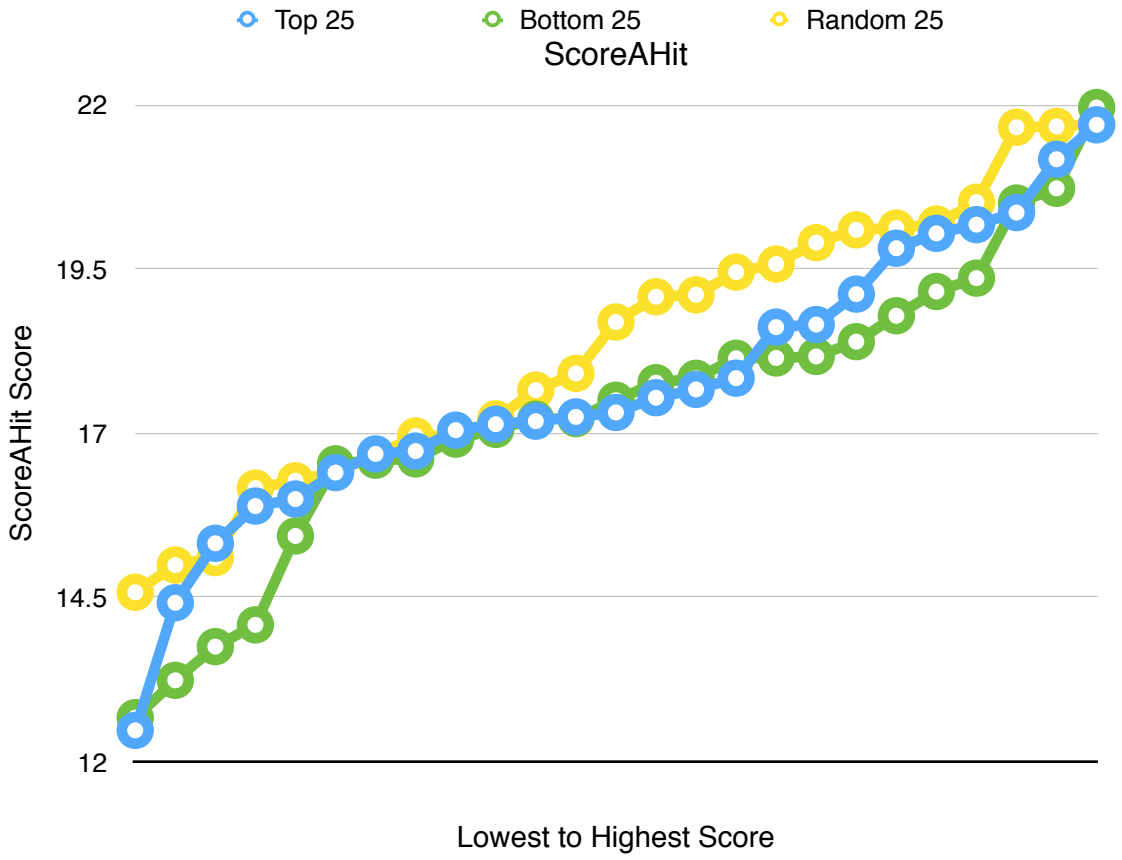
The Songs

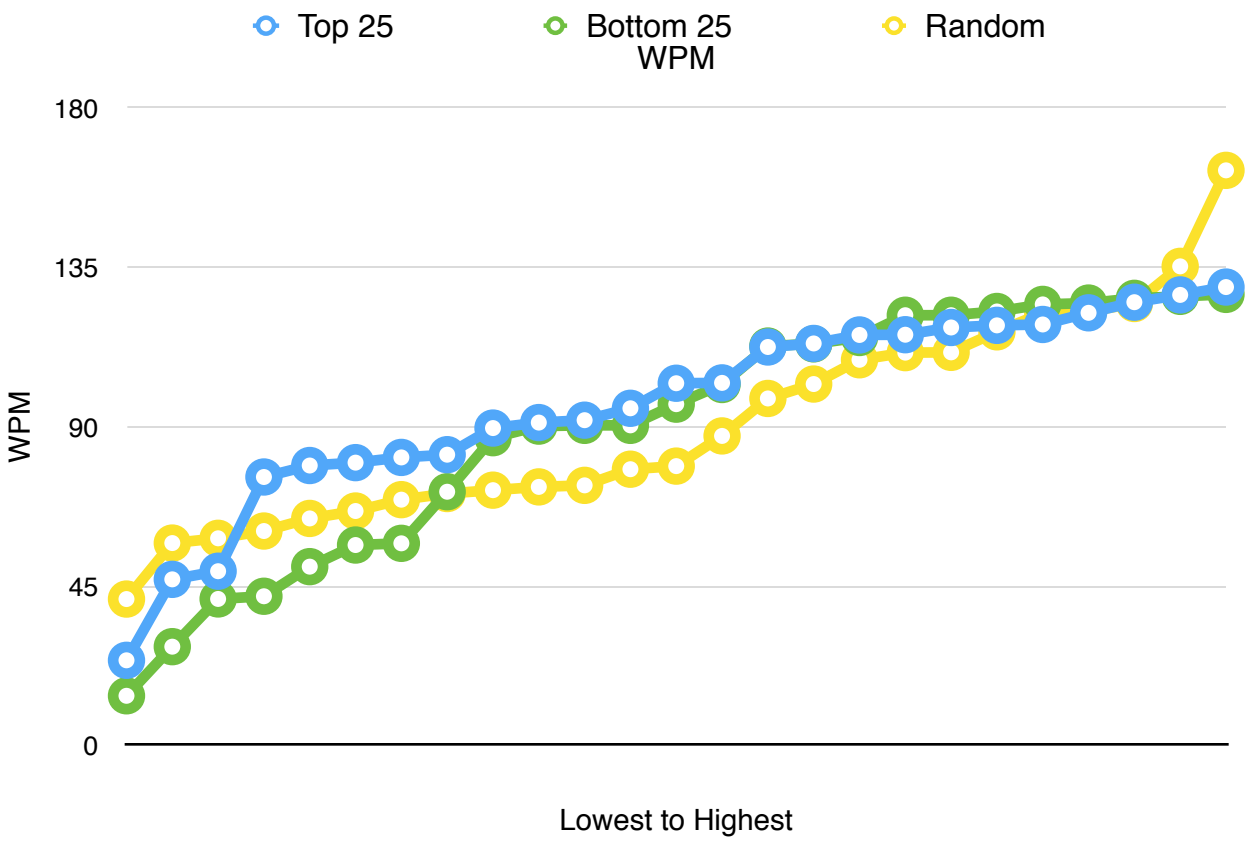
Top 25		Bottom 25		Random 25	
Title	**Artist**	**Title**	**Artist**	**Title**	**Artist**
Adore You	Miley Cyrus	Come and Get It	Selena Gomez	Play It Again	Luke Bryan
All Of Me	John Legend	Thinking About You	Calvin Harris	Break Your Plans	The Fray
Best Day of My Life	American Authors	Coming Home	Diddy	Tennis Court	Lorde
Bottoms Up	Brantley Gilbert	Move That Dope	Future	This Moment	Katy Perry
Burn	Ellie Goulding	Radioactive	Imagine Dragons	The Other Side	Jason Derulo

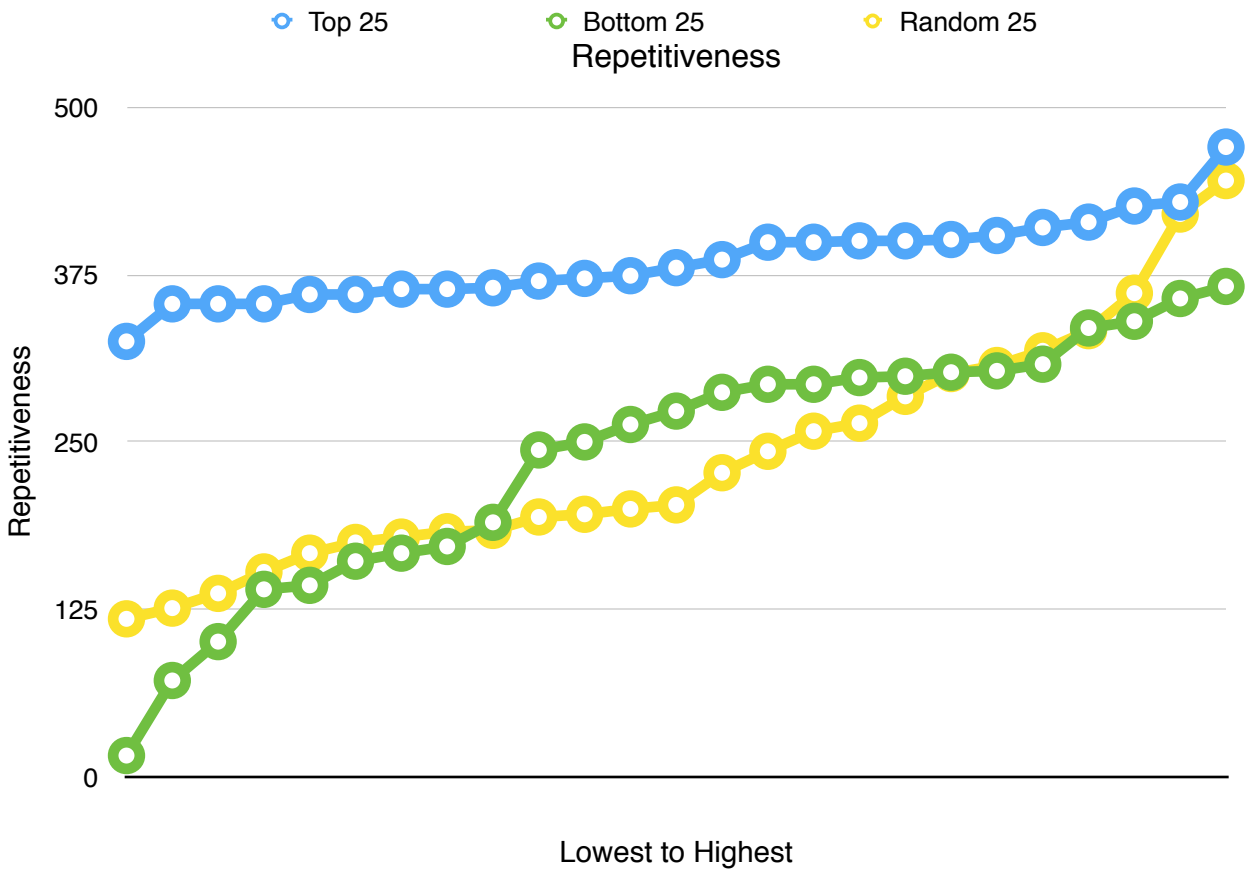
Top 25		Bottom 25		Random 25	
Title	**Artist**	**Title**	**Artist**	**Title**	**Artist**
Chillin' It	Cole Swindell	We Dem Boyz	Wiz Khalifa	Tip It Back	Florida Georgia Line
Counting Stars	OneRepublic	Wave	Beck	Joy	Ellie Goulding
Dark Horse	Katy Perry	Party Girls	Ludacris	Believer	American Authors
Demons	Imagine Dragons	Feel This Moment	Pitbull	Things We Lost in the Fire	Bastille
Everything Is Awesome	Tegan and Sara	Heart Attack	Demi Lovato	Two Pieces	Demi Lovato
Hey Brother	Avicii	Cruise	Florida Georgia Line	Roar	Katy Perry
Let Her Go	Passenger	Waking Light	Beck	Drive	Miley Cyrus
Let It Go	Idina Menzel	Cruise (remix)	Florida Georgia Line	Liar Liar	Avicii
Love Me Again	John Newman	We Were Us	Keith Urban and Miranda Lambert	Underdog	Imagine Dragons
Neon Lights	Demi Lovato	Lose Yourself	Eminem	Rockstar	A Great Big World
Pompeii	Bastille	Let Me See Ya Girl	Cole Swindell	Cheating	John Newman
Royals	Lorde	Santeria	Sublime	The Beginning	John Legend
Say Something	A Great Big World	In The Air Tonight	Phil Collins	Don't Stop The Party	Pitbull
Story Of My Life	One Direction	We Own it	2 Chainz and Wiz Khalifa	Strong	One Direction
Talk Dirty	Jason Derulo	Same Love	Macklemore and Ryan Lewis	Red Camaro	Keith Urban
Team	Lorde	Confident	Justin Bieber	If I Lose Myself	OneRepublic
The Man	Aloe Blacc	Berzerk	Eminem	Circles	Passenger
Timber	Pitbull	Low	Flo Rida	I Just Want you	Cole Swindell
Turn Down For What	DJ Snake	Without You	David Guetta	Love Is The Answer	Aloe Blacc
U	Austin Mahone	Explosions	Ellie Goulding	Still Sane	Lorde

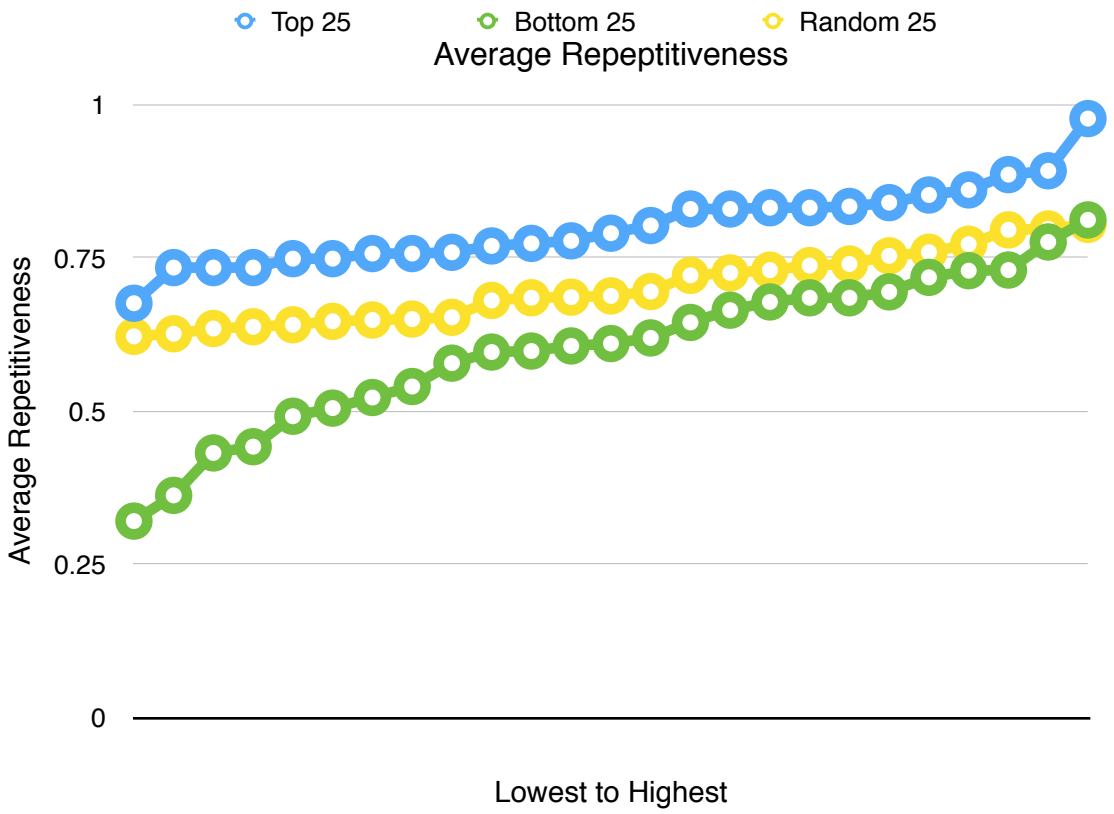
The Results

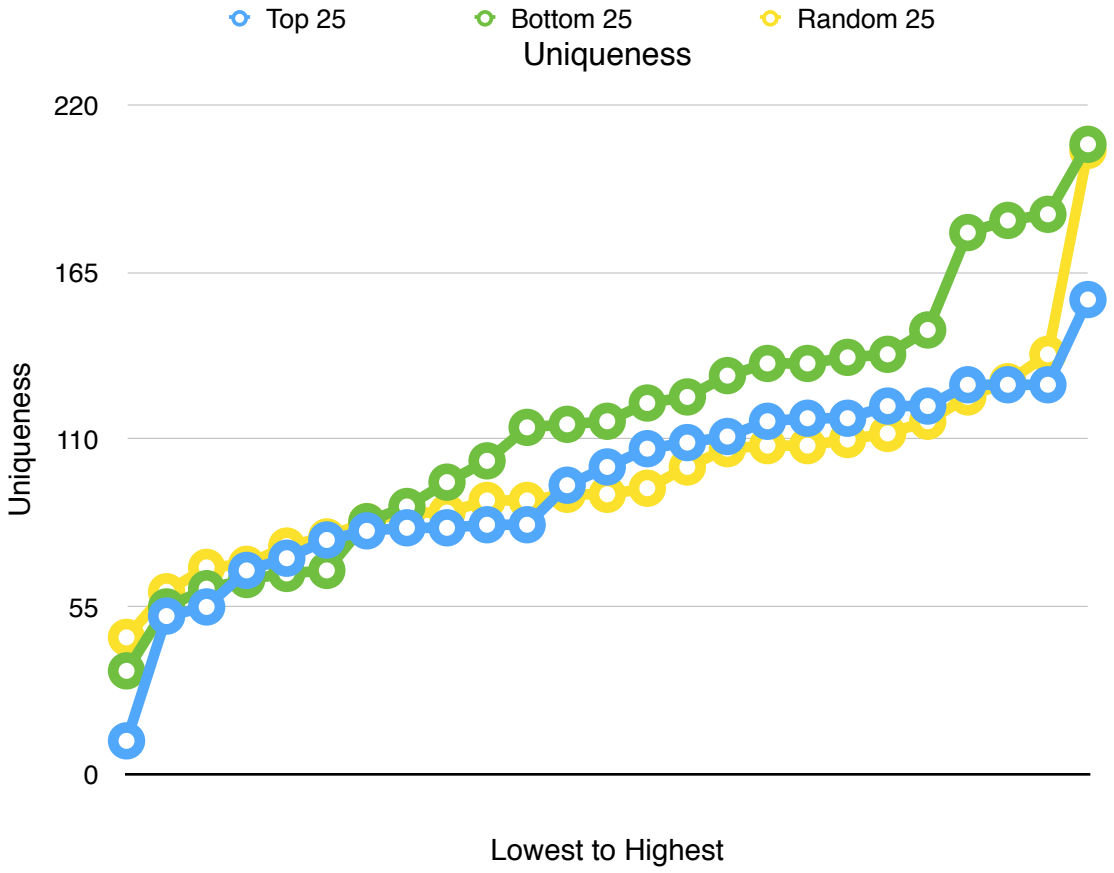
As can be seen, ScoreAHit's algorithm, which claims a 60% accuracy, shows little difference between popular and unpopular songs.

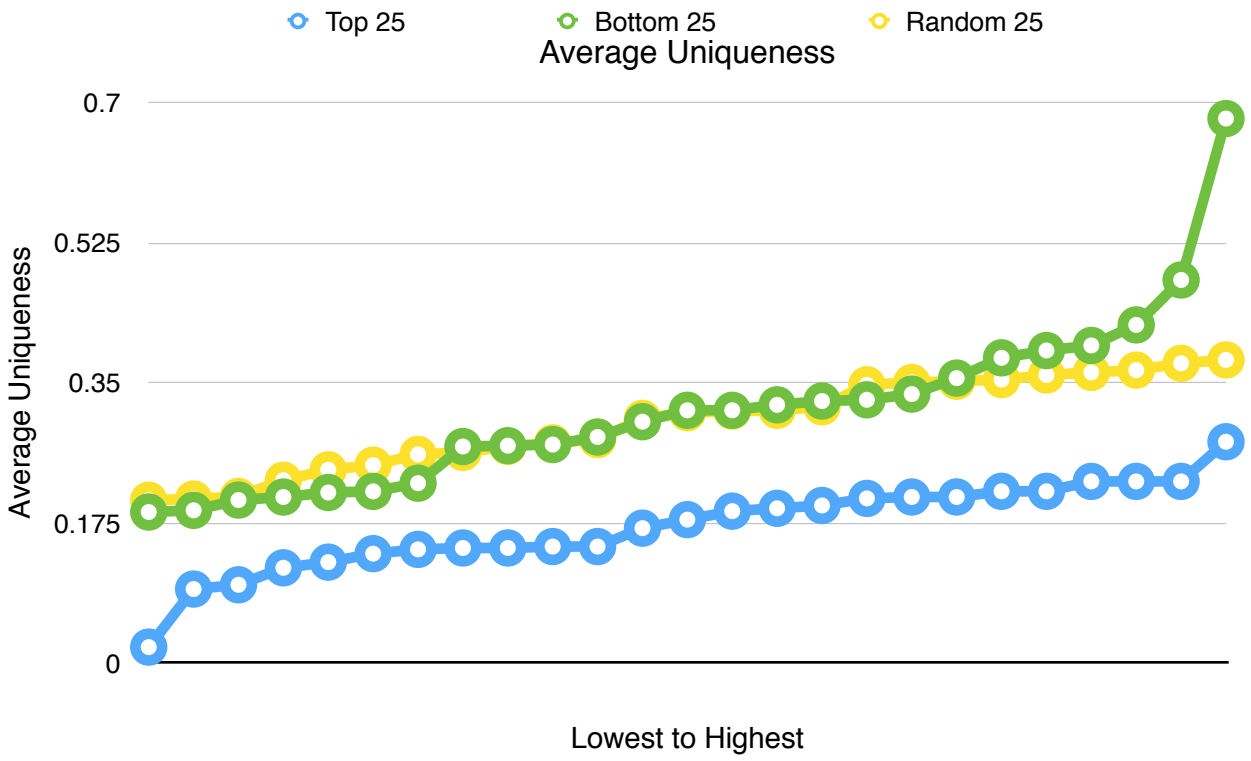


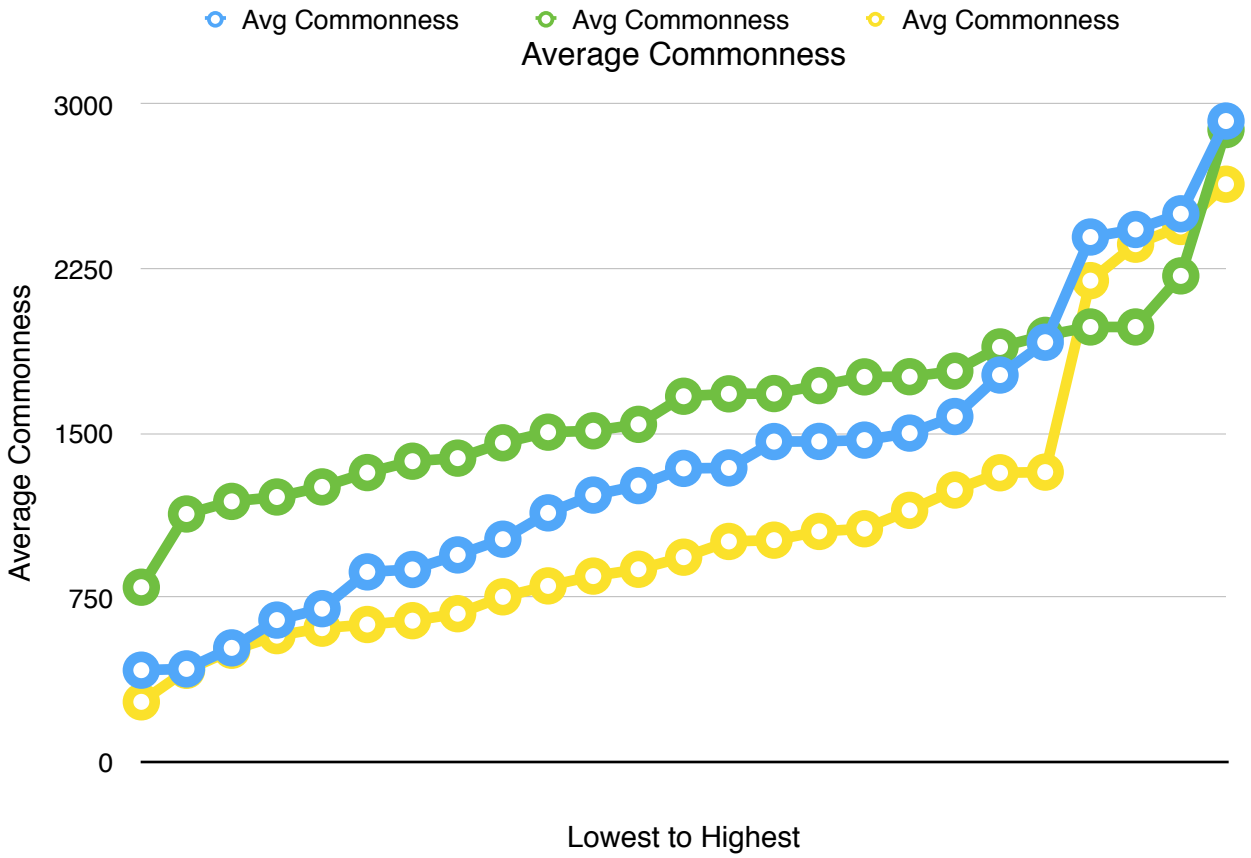

Top 25
Bottom 25
Random
WPM
180
135
90
45
0
WPM
Lowest to Highest

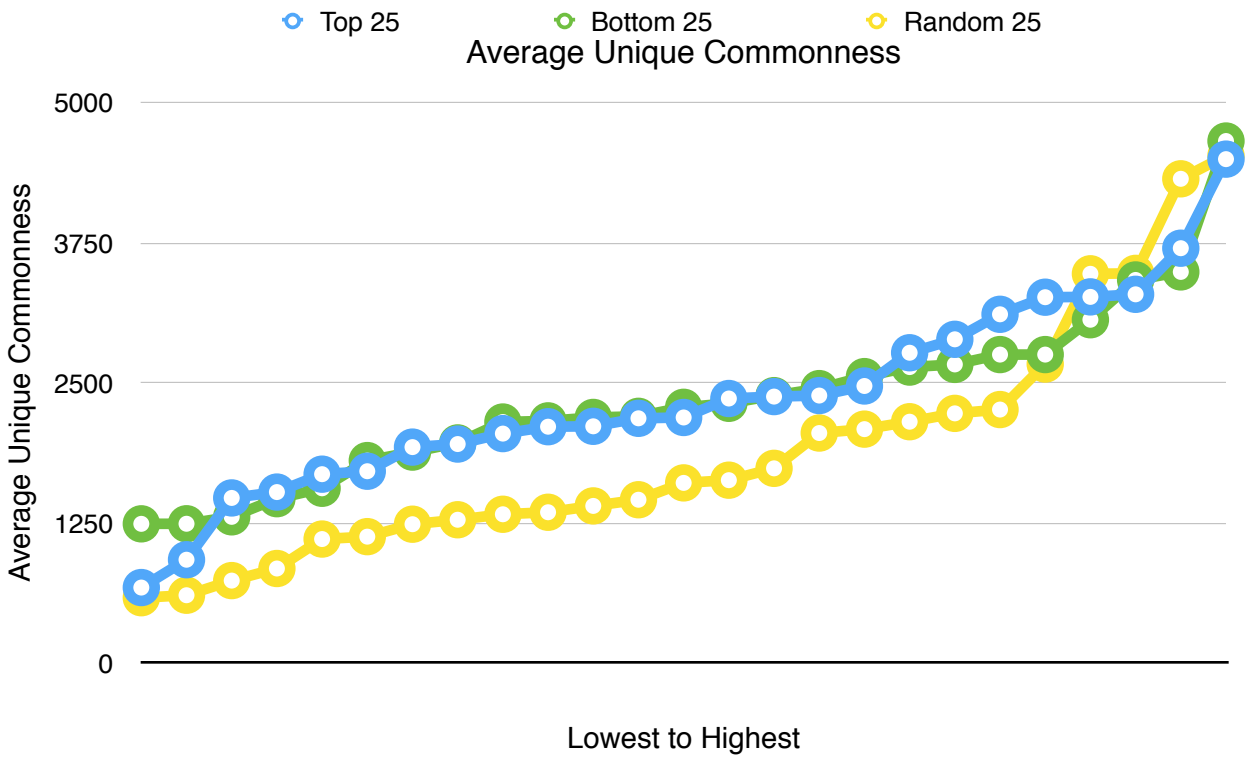

Top 25
Bottom 25
Random 25
Repetitiveness
500
375
250
125
0
Repetitiveness
Lowest to Highest

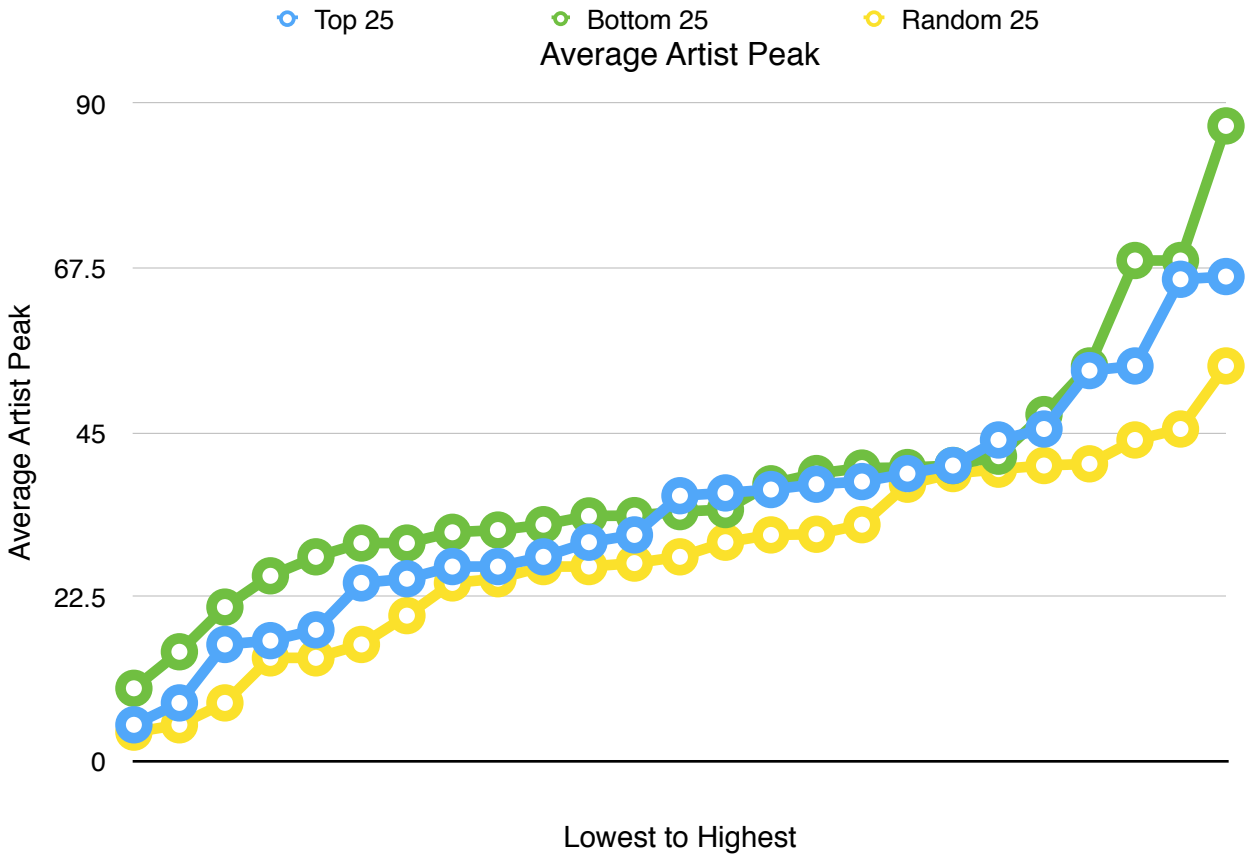

Top 25
Bottom 25
Random 25
Average Repeptitiveness
1
0.75
0.5
0.25
0
Average Repetitiveness
Lowest to Highest

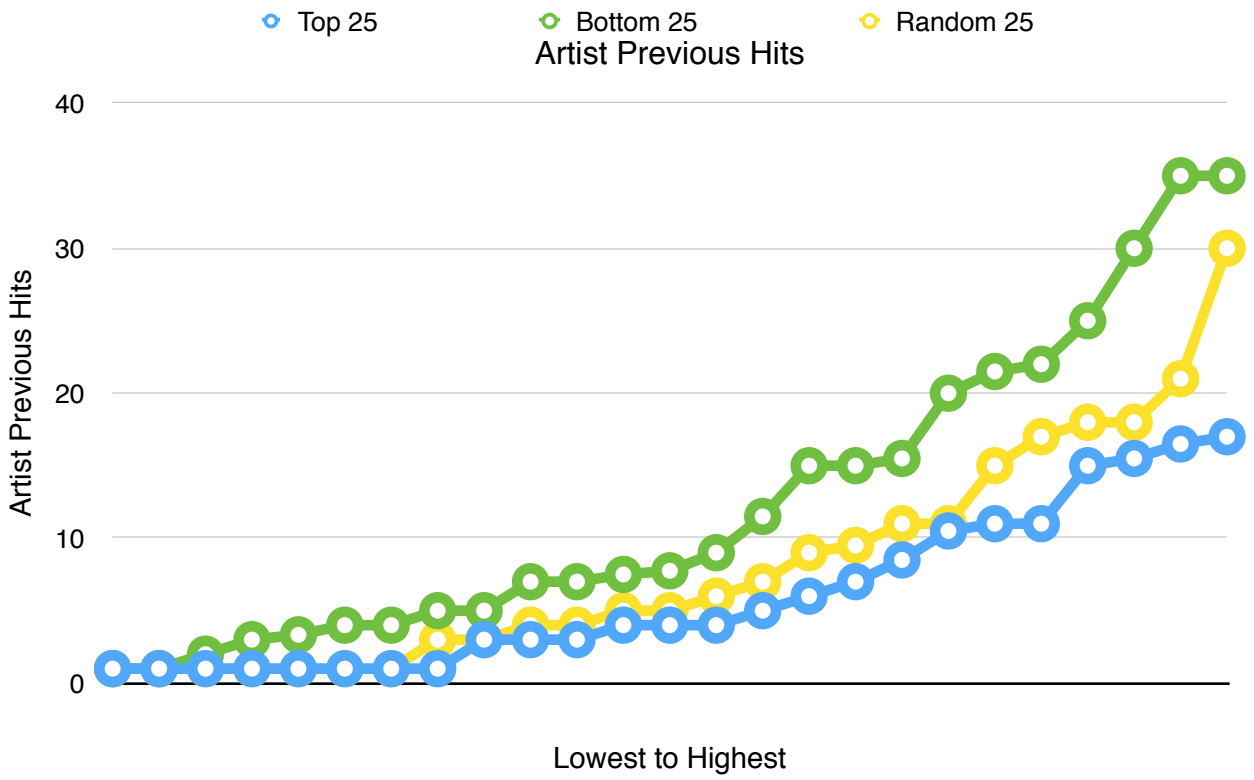

Top 25
Bottom 25
Random 25
Uniqueness
220
165
110
55
0
Uniqueness
Lowest to Highest


Top 25
Bottom 25
Random 25
Average Uniqueness
0.7
0.525
0.35
0.175
0
Average Uniqueness
Lowest to Highest


Avg Commonness
Avg Commonness
Avg Commonness
Average Commonness
3000
2250
1500
750
0
Average Commonness
Lowest to Highest


Top 25
Bottom 25
Random 25
Average Unique Commonness
5000
3750
2500
1250
0
Average Unique Commonness
Lowest to Highest


Top 25
Bottom 25
Random 25
Average Artist Peak
90
67.5
45
22.5
0
Average Artist Peak
Lowest to Highest



These graphs show very obvious trends in the data. However, t-tests were used to verify their validity. The following p-values were produced:

Top 25 – Bottom 25

	p-value
ScoreAHit Score	0.288119773302137
WPM	0.246300187408548
Repetitiveness	0.000000019455770019293 4
Avg Repetitiveness	0.000000013854570325576 5
Uniqueness	0.0401751735009899
Avg Uniqueness	0.000000365347339965686
Avg Commonness	0.0532377369783821
Avg Unique Commonness	0.479650904425369
Average Artist Peak	0.183373971296959
Artist Previous Hits	0.00522323716519614

Top 25 – Random

	p-value
ScoreAHit Score	0.146951614662972
WPM	0.242001340572233
Repetitiveness	0.000000005491643347204 04
Avg Repetitiveness	0.000000327104721902913
Uniqueness	0.372874649856552
Avg Uniqueness	0.000000000060185231959962
Avg Commonness	0.0768889169189782
Avg Unique Commonness	0.049351570759296
Average Artist Peak	0.0811345171216479
Artist Previous Hits	0.142456188490836

Bottom 25 – Random

	p-value
ScoreAHit Score	0.0591571204482308
WPM	0.479673364970876
Repetitiveness	0.498157785433046
Avg Repetitiveness	0.000501575832112954
Uniqueness	0.0668560904869185
Avg Uniqueness	0.260833360672345
Avg Commonness	0.000660124714066096
Avg Unique Commonness	0.0477944913906124
Average Artist Peak	0.0116254490777568
Artist Previous Hits	0.0519045200724735

In the above tables, values highlighted in red (those with a p-value above 0.05) are considered insignificant, those in yellow ($0.01 < p < 0.05$) considered somewhat significant, and those in green ($p < 0.01$) considered very significant in determining the difference between songs in the two categories. All t-tests were single-tailed, with unpaired data of unequal variance.

Conclusion

By far the most significant attribute in differentiating Top 25 songs from others is how repetitive the lyrics are. The p-values calculated when comparing the Top 25 to other are incredibly small, verifying this relationship. This, though, was not a surprising result for those familiar with contemporary popular music.

A more unexpected result was the relationship between word commonness and popularity. The Top 25 and and Bottom 25 song sets both had significantly less common words than those of the random set. This means that, in popular songs, the words used tend to be less likely to appear in everyday conversations.

The most surprising result was that of the Artist Previous Hits attribute. It clearly showed that Top 25 songs were more likely to be achieved from less popular artists, and that Bottom 25 songs were more often achieved by artists with a large number of previously charting songs. At first, this seems like counter-intuitive, but upon further reflection it seems more appropriate. The most likely reason for this result is that any new song of an already popular will tend to be more popular, regardless of its mediocrity, and thus make in onto the list, but it won't necessarily be chart-topping material. However, for a lesser-known artist to chart, his charting song will likely need to be exceptional, and thus make it to the Top.

Sources

1. "Documentation." *ScoreAHit*. ScoreAHit, n.d. Web. 25 Feb. 2014. <http://www.scoreahit.com/Documentation>.
2. Harris, Jonathan. "Wordcount · Tracking the Way We Use Language." *Wordcount · Tracking the Way We Use Language*. N.p., 2003. Web. 25 Feb. 2014. <http://www.wordcount.org/main.php>.
3. "The Hot 100." Billboard.com. N.p., n.d. Web. 25 Feb. 2014. <http://www.billboard.com/charts/hot-100>.
4. "A-Z Lyrics Universe." A-Z Lyrics Universe. AZLyrics.com, n.d. Web. 25 Feb. 2014. <http://www.azlyrics.com/>.
5. "Last.fm." Last.fm. Last.fm Limited, n.d. Web. 25 Feb. 2014. <http://www.last.fm/home>.