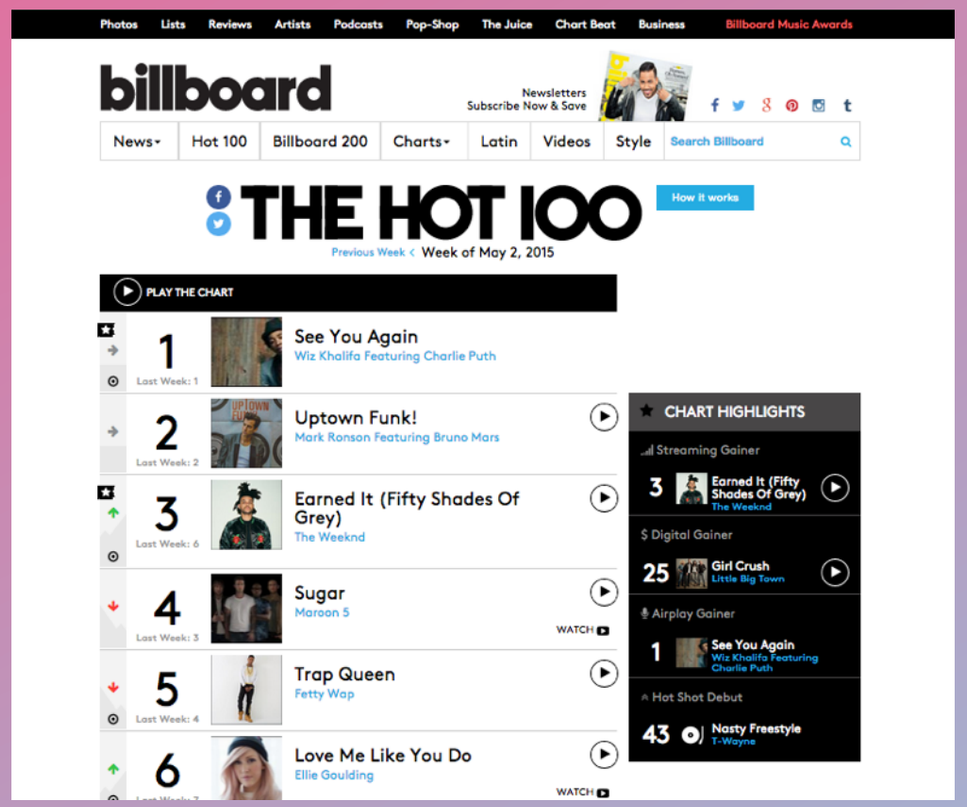
What makes a top-charting single?

Yui Chi Tiffany Lo
CS591 - Professor Terzi
Spring 2015

Obtaining Data



- Scraped data from three websites to get chart data (Billboard), track attributes (Echo Nest), and lyrics (Genius)

Challenges

- Songs with the same name and cover songs complicated data collection process
- Censored titles vs uncensored titles and other naming convention differences between sites
- API limits = Hours of scraping
- Contractions/slang

About the Data

Chart data:

- January 2012 - April 2015
  (Up to consecutive 87 weeks)
- 1397 Songs
  (song, artist, featuring, duration, peak...)

Track attributes:

- time_signature, energy, liveness, tempo, speechiness, acousticness, danceability, instrumentalness, key, loudness, valence, location, longitude, latitude
- 1100 Songs

Lyrics:

- 1379 Songs

About the Data

tempo: the BPM of the song

danceability: A number that ranges from 0 to 1, representing how danceable The Echo Nest thinks this song is.

energy: A number that ranges from 0 to 1, representing how energetic The Echo Nest thinks this song is

key: The key that The Echo Nest believes the song is in

Key signatures start at 0 (C) and ascend the chromatic scale. In this case, a key of 1 represents a song in D-flat

loudness: The overall loudness of a track in decibels (dB)

acousticness: A measure of how acoustic vs. electric a song is; close to 1 indicates that the song is mostly recorded with acoustic instruments and non-modified vocals, close to zero indicates that the song has many electric instruments such as as electric guitars and synths. Vocals may be processed, filtered or distorted.

valence: A measure of the emotional content of a song; close to 1 indicates a positive emotion, close to 0 is a negative emotion. Valence is often combined with energy to yield a four quadrant mood: high energy/high valence, high energy/low valence, low energy/high valence, and low energy/low valence. Low energy/low valance songs are typically sad, whereas high energy/low valence songs are angry

time_signature: Time signature of the key; how many beats per measure

About the Data

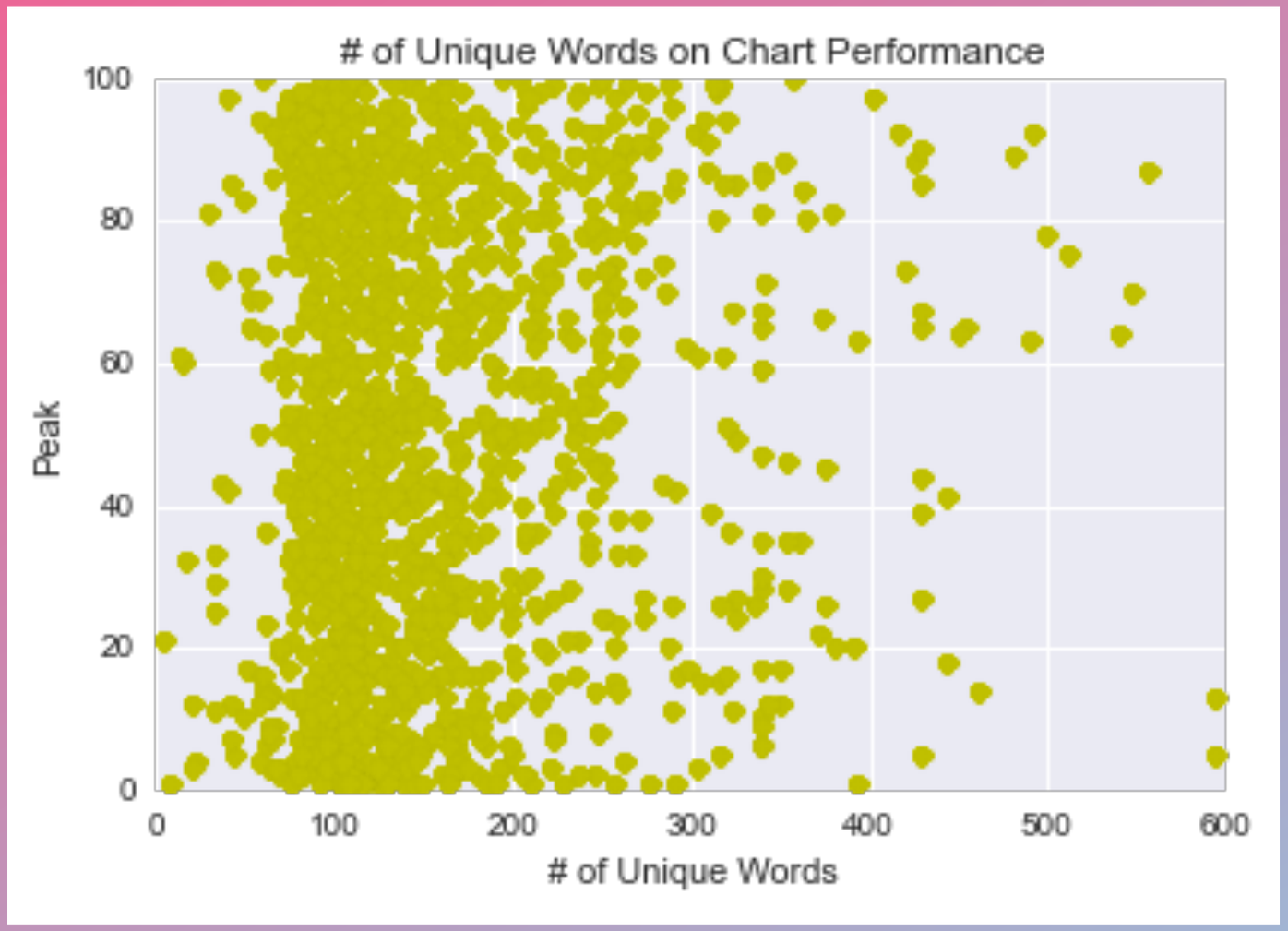
Chart data:

- 542 artists
- Average peak: 50.8138869005
- Average duration: 12.3836793128
- Peak Min: 1 Max: 100
- Duration Min: 1 Max: 87

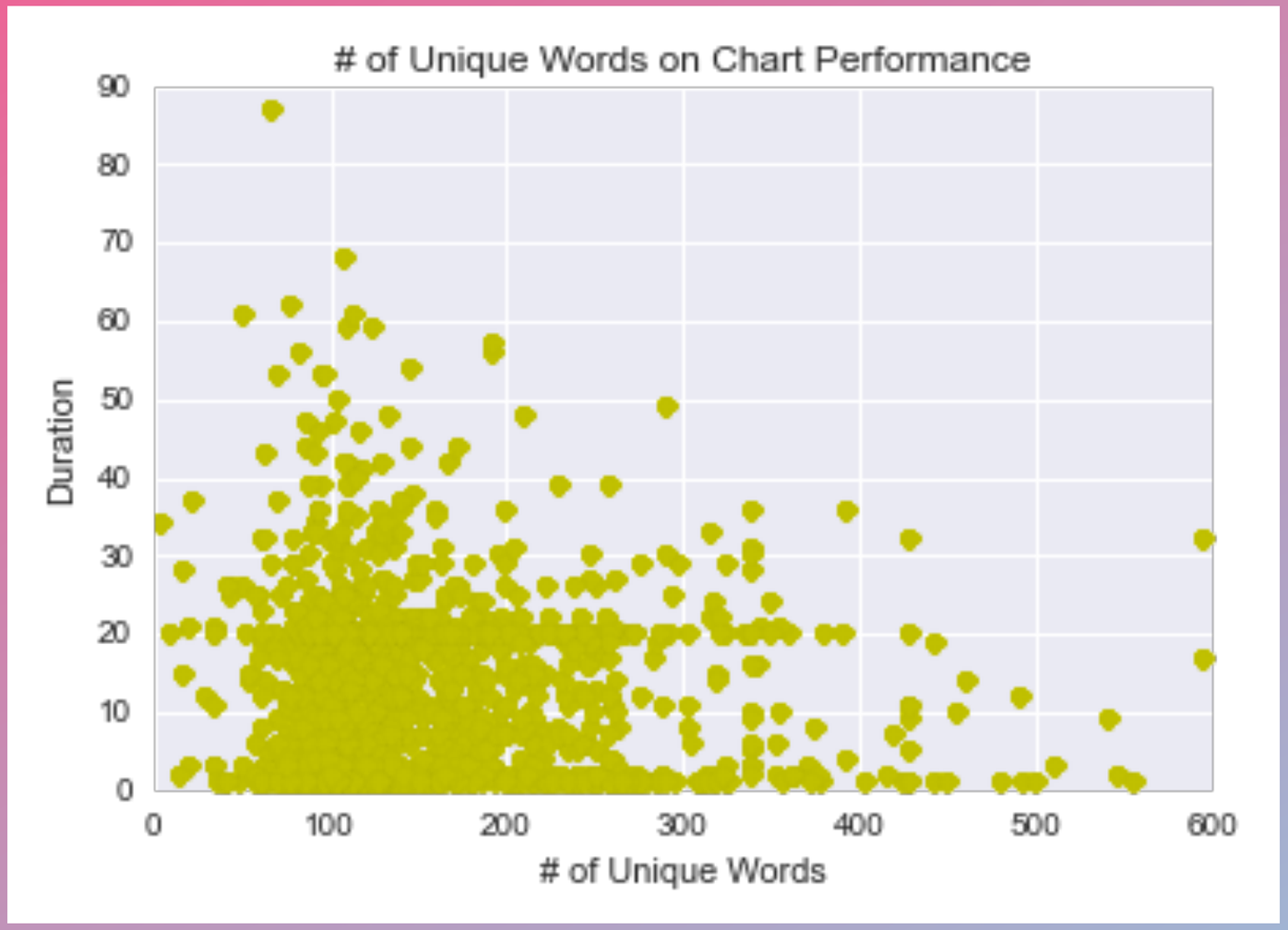
Tools/Techniques

- Textblob (text analysis, sentiment analysis)
- Scikit Learn linear regression models
- Kmeans clustering
- Plotting

Songs that contain fewer unique words will chart more successfully than songs with greater amounts of unique words.



- A significant amount of the charting songs have 100-150 unique words
- Longer songs are less prevalent on the charts
- Charting songs with over 400 words are more likely to peak below Top 20 mark



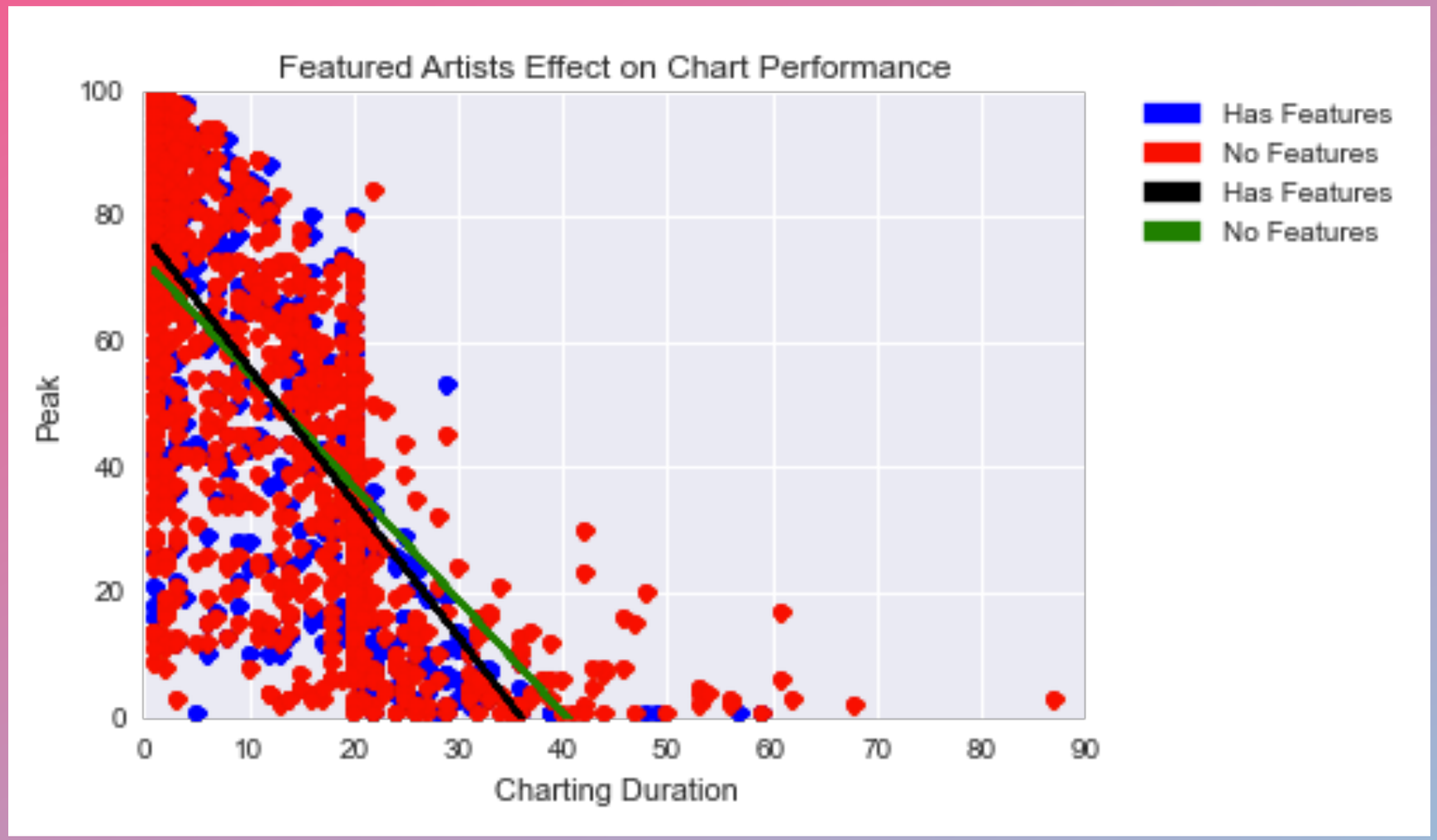
- Most songs last for up to 20-25 weeks
- Songs with over 400 words have little staying power on the charts; under 20 weeks

Conclusion

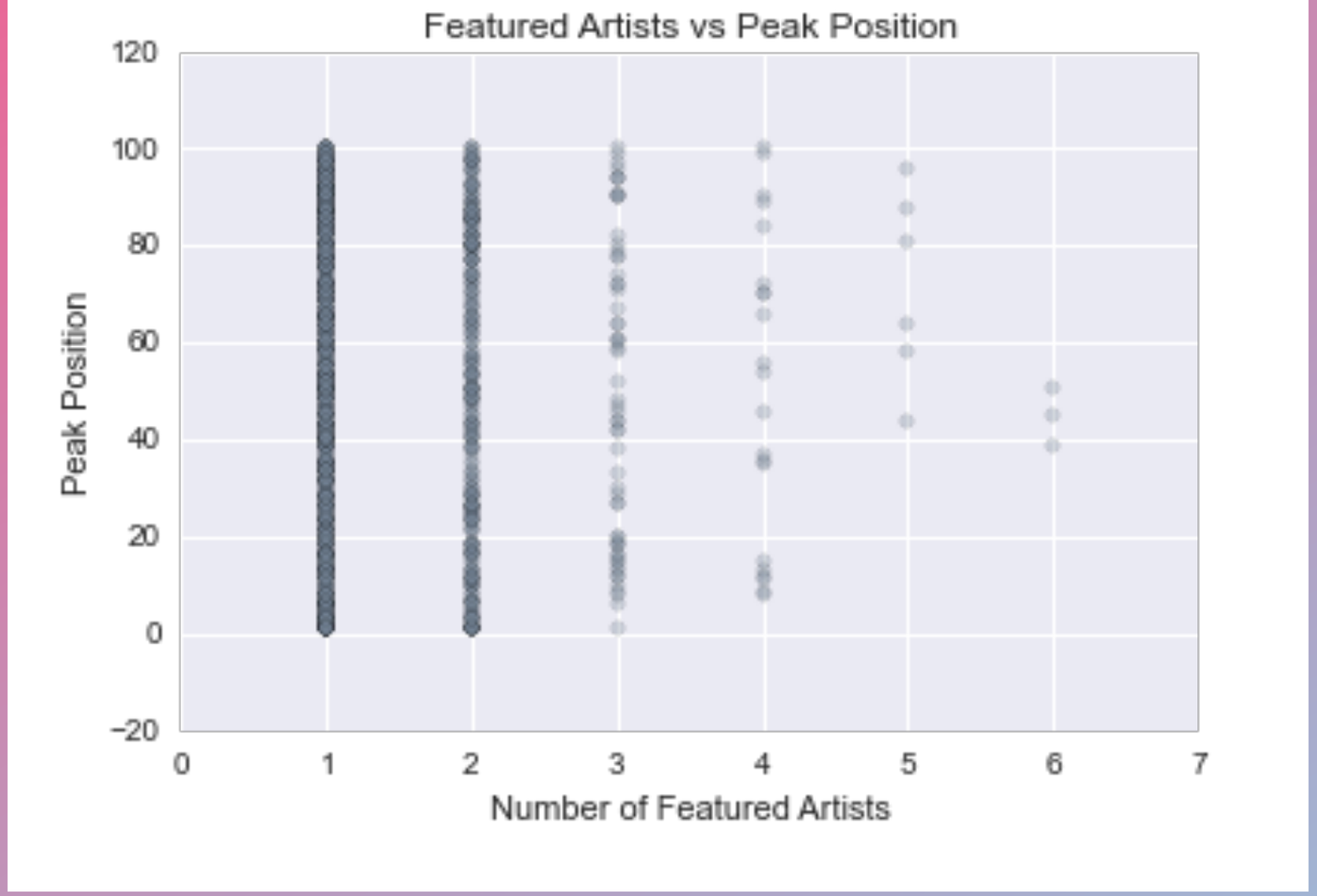
This hypothesis can be inferred as correct based on the proof that the inverse, songs that have more unique words would be less successful than songs with fewer, holds true.

Further analysis and noise reduction in the “fewer unique words” section ranging 100-150 words would be better prove this hypothesis.

Songs that contain features will chart more successfully than songs with no features.



- No song with features pasts the 60 week mark; 5 songs without features do
- Songs with features that pass the 20 week mark peak higher but last shorter than no features
- The trend line for Has Features is much sharper than No Features



- Most songs do not feature additional artists
- Songs that have more than 4 artists do not peak in the Top 40
- Even songs with 4 artists cannot break past the Top 10

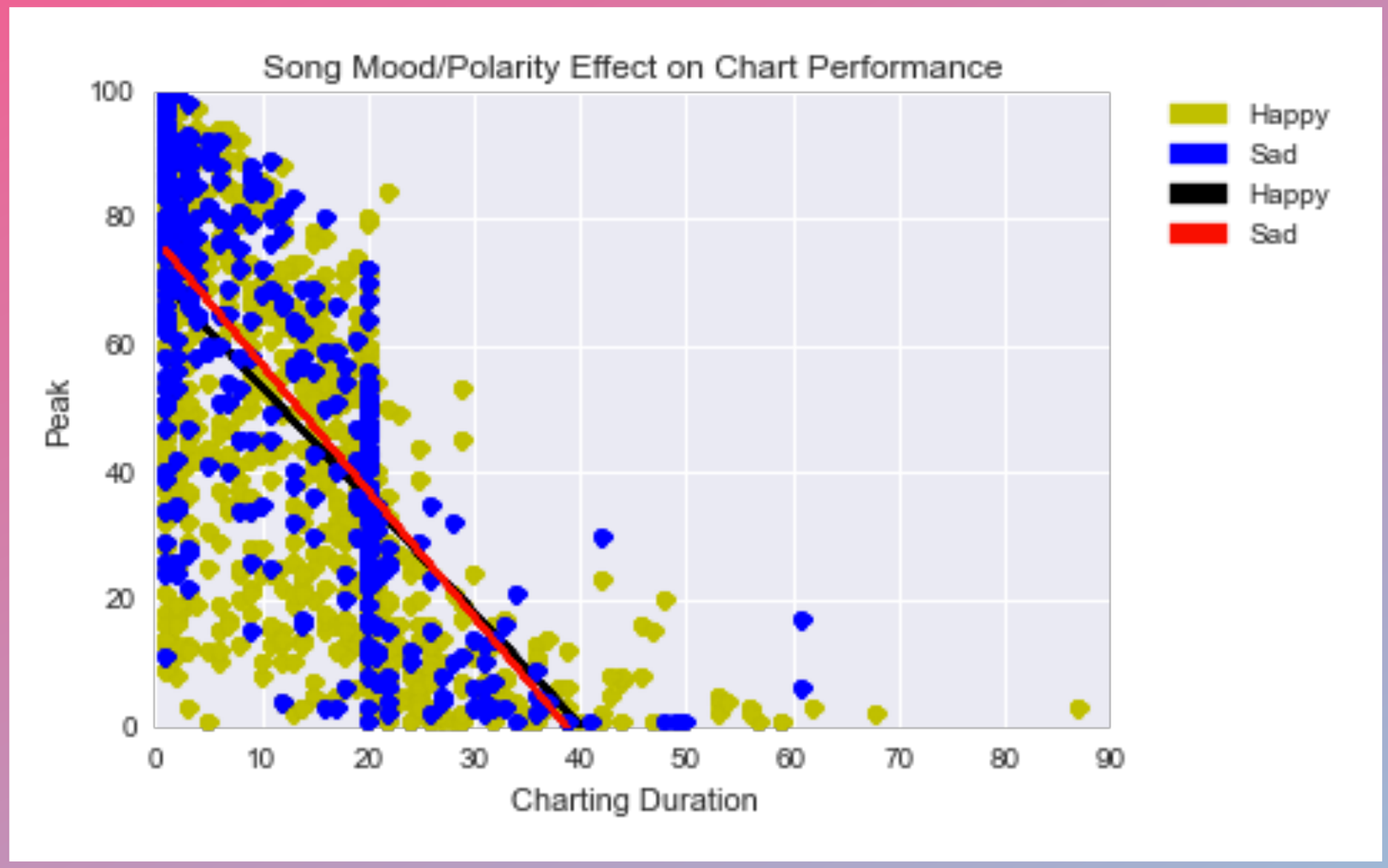
Conclusion

This hypothesis is proven incorrect based on the findings shown, where songs with features chart less successfully than songs performed by a single artist in terms of peak position and duration.

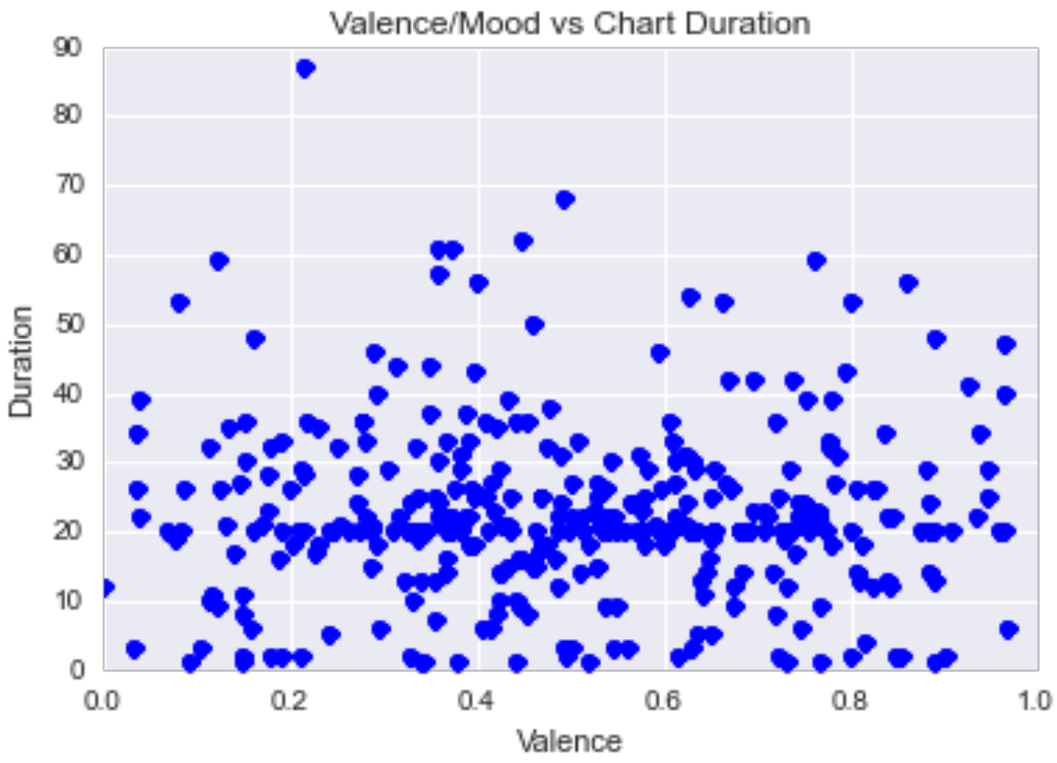
Looking at it in a binary fashion and by number of total artists on a track, songs with features have a smaller chance of high chart peaks or longevity.

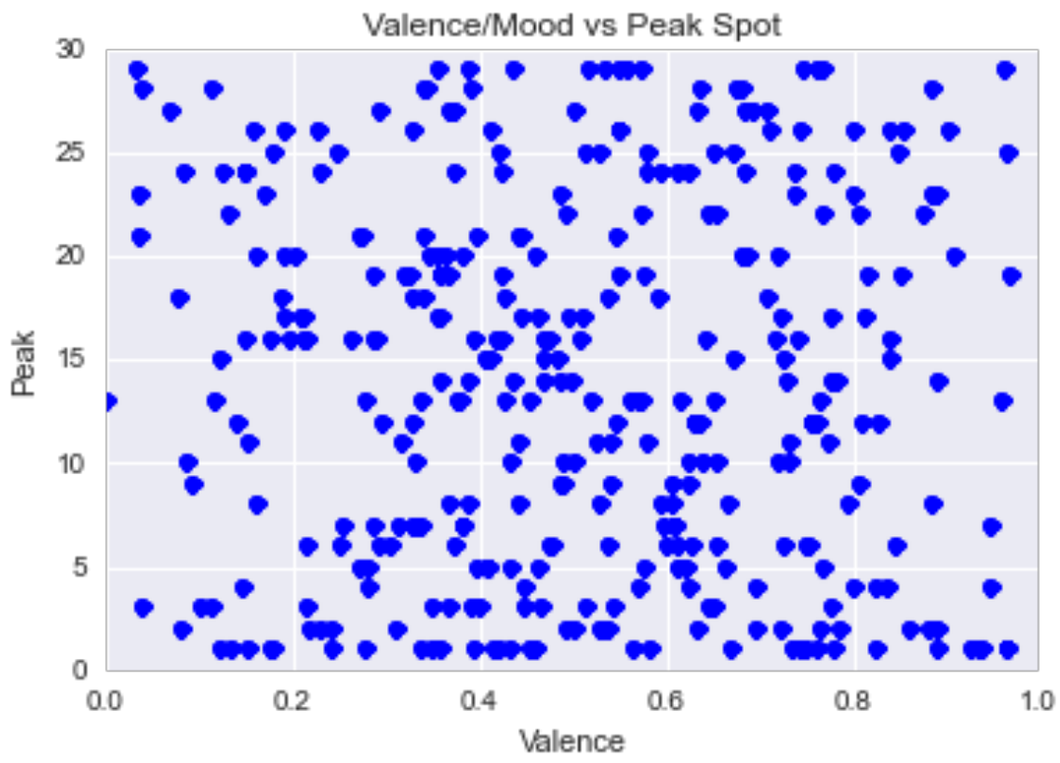
Perhaps, this is because many collaborations do not work very well and that many tracks with features are rap/hip hop which is a less mainstream genre.

Songs that sing about sad subjects will tend to chart for a longer duration than songs that are happy.



- Fewer sad songs last past the 40 week mark than happy songs
- Sad songs' peak in the lower portion of the chart as compared to their happy counterparts
- Sad songs have a sharper downward-sloped best fit line indicating faster peak decline





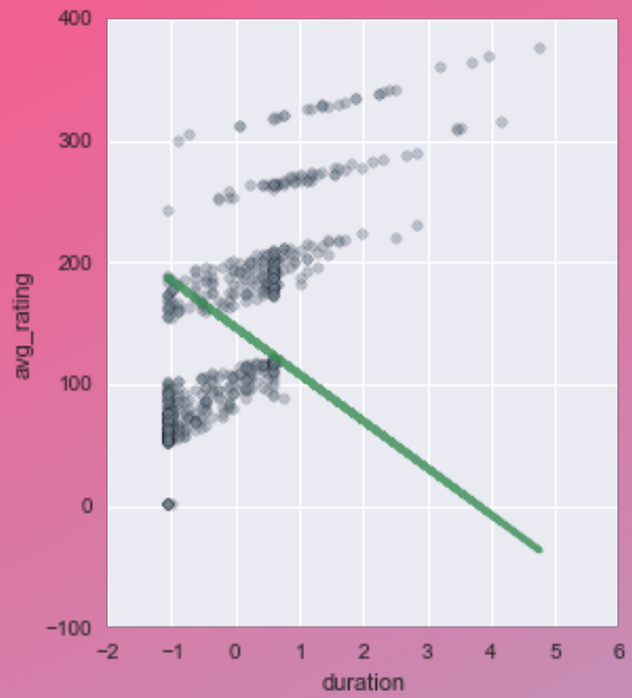
- Valence/Mood did not provide any interesting correlations with neither duration nor peak

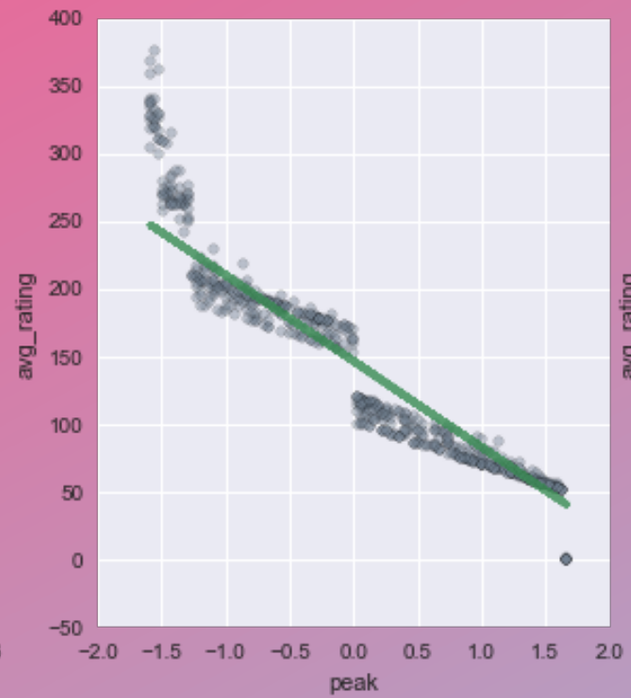
Conclusion

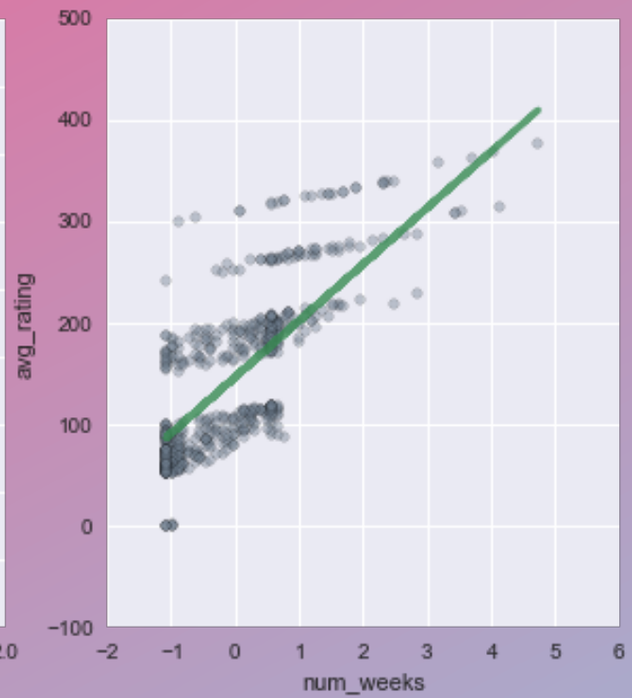
This hypothesis has been neither proven nor disproven by my data. The best fit lines for both happy and sad song indicate that most songs have peaked and left the chart by Week 40. Otherwise, few conclusions could be drawn from the data.

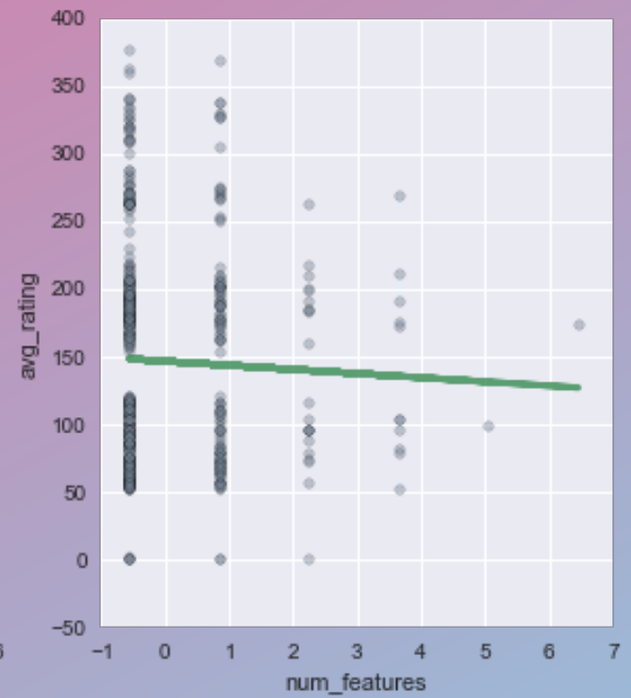
The subjective manner of evaluating happy/sad songs makes it difficult to judge by sentiment analysis or other algorithms. This is a difficult measure to analyze the dataset on.

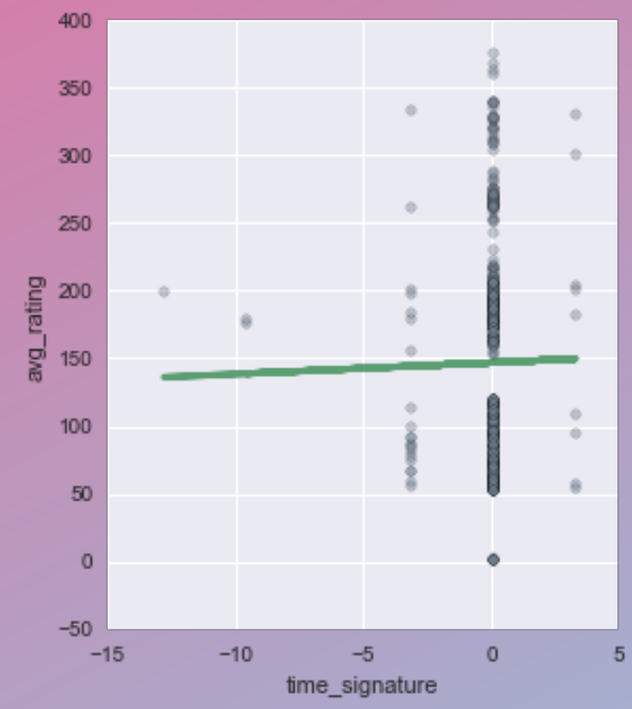
Songs with higher danceability or energy chart higher than their counterparts(in the Top 50).


avg_rating
duration

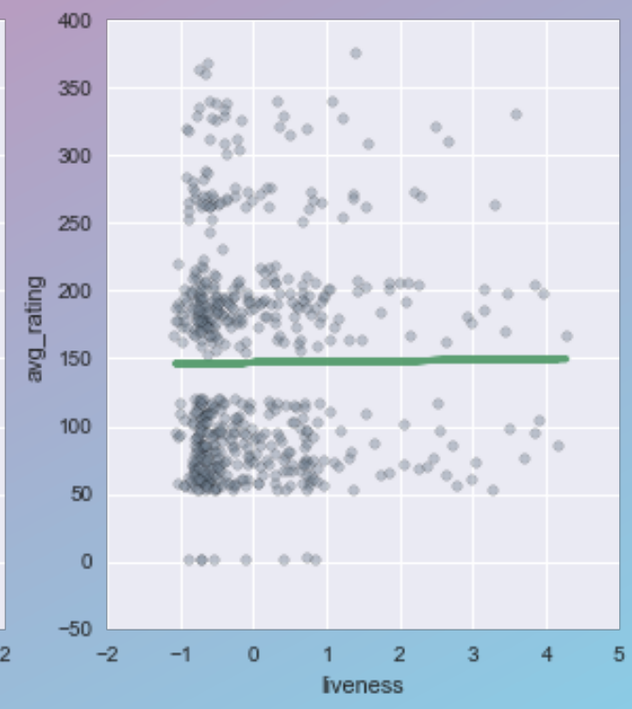

avg_rating
peak

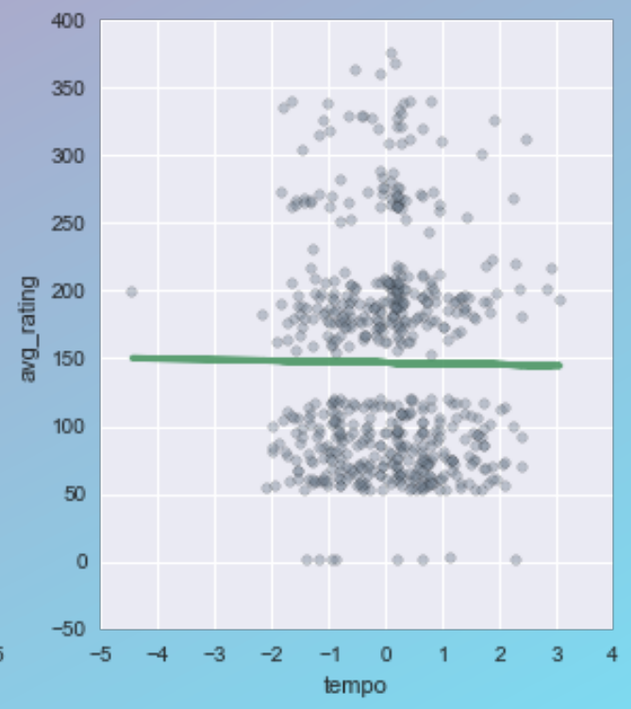

avg_rating
num_weeks

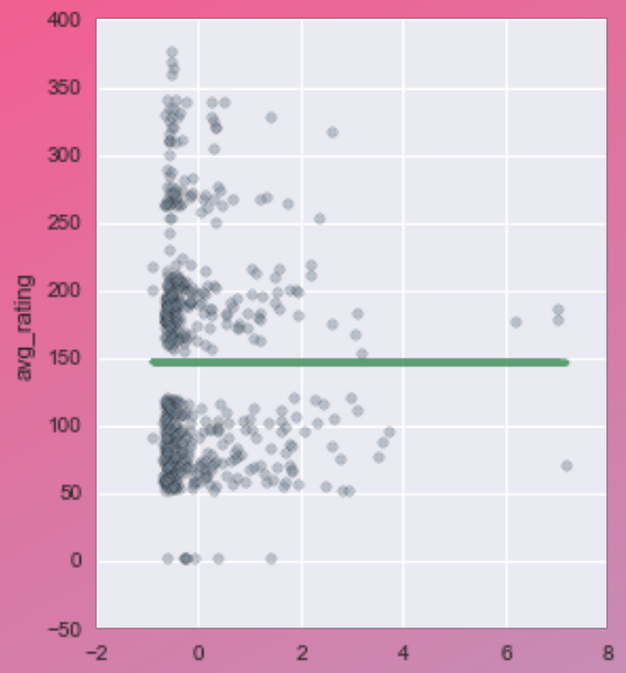

avg_rating
num_features

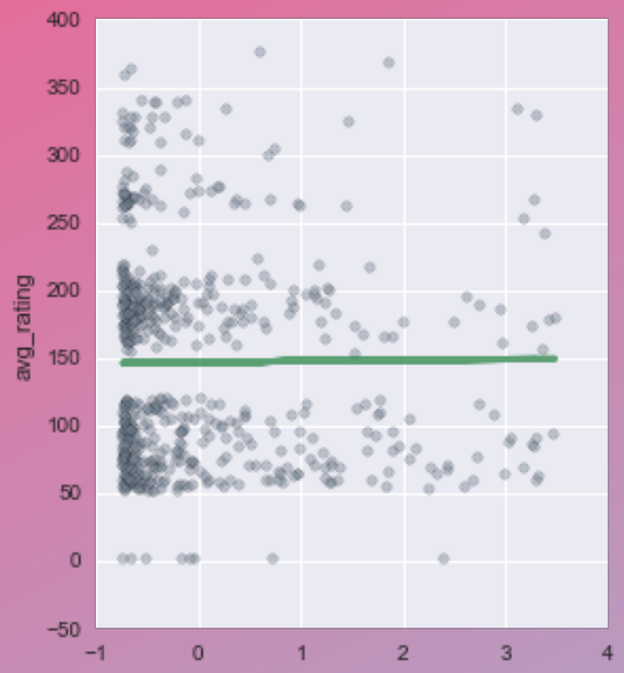

avg_rating
time_signature

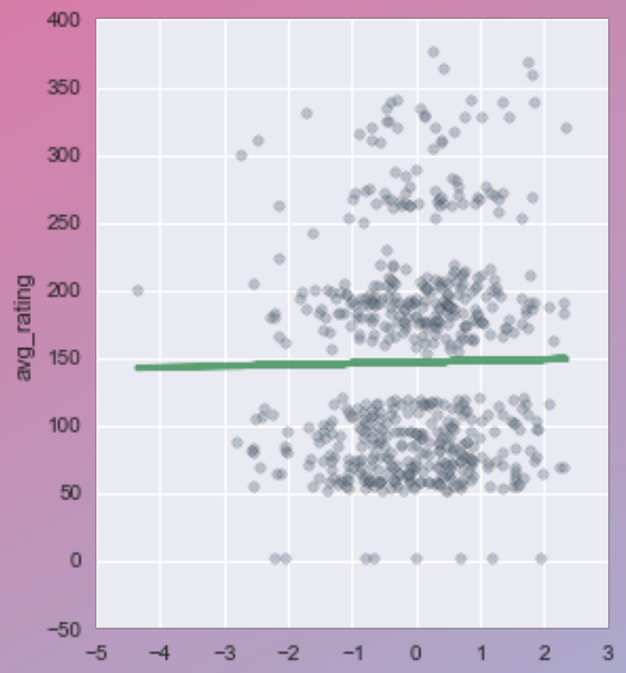

avg_rating
energy

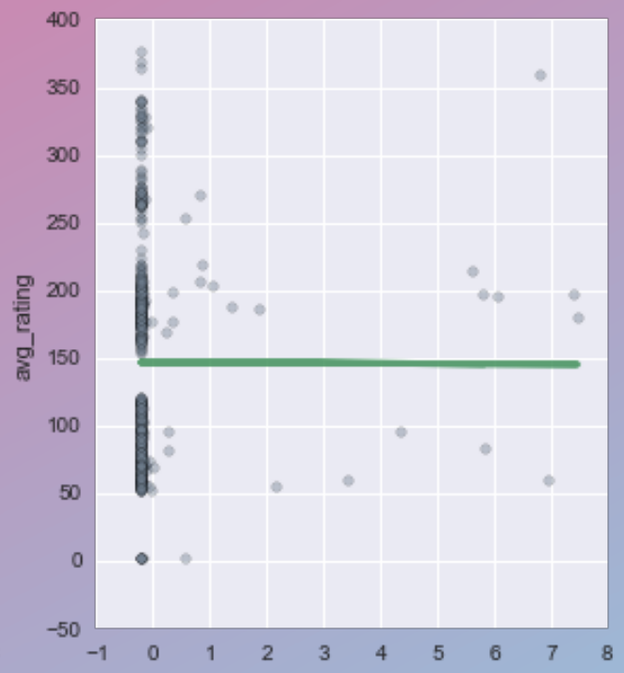

avg_rating
liveness

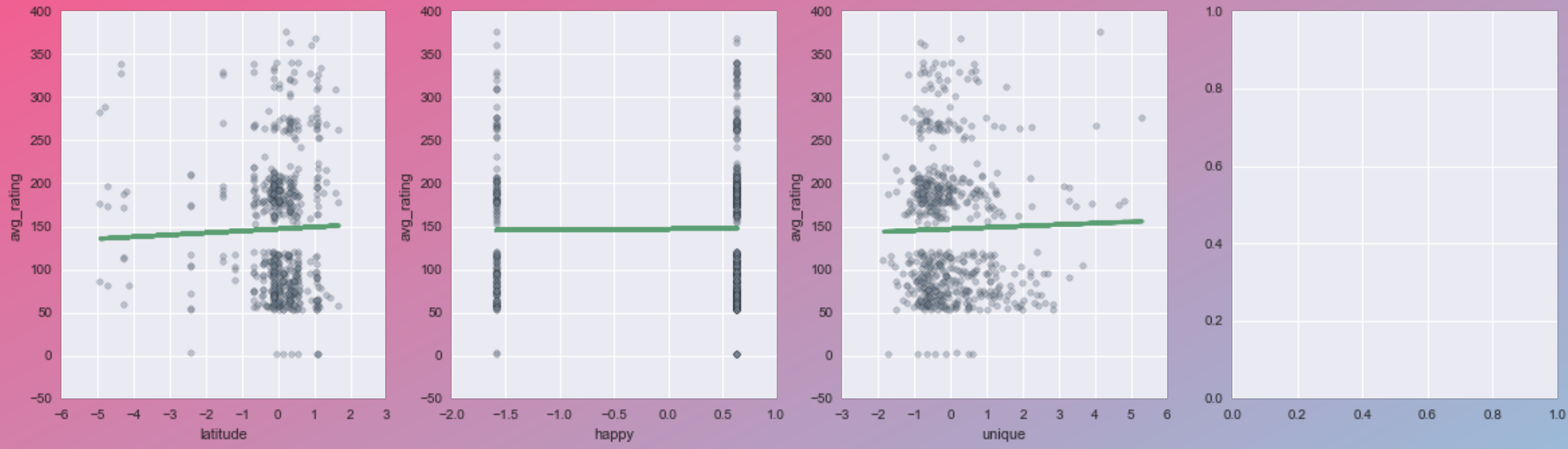

avg_rating
tempo


avg_rating
speechiness
key

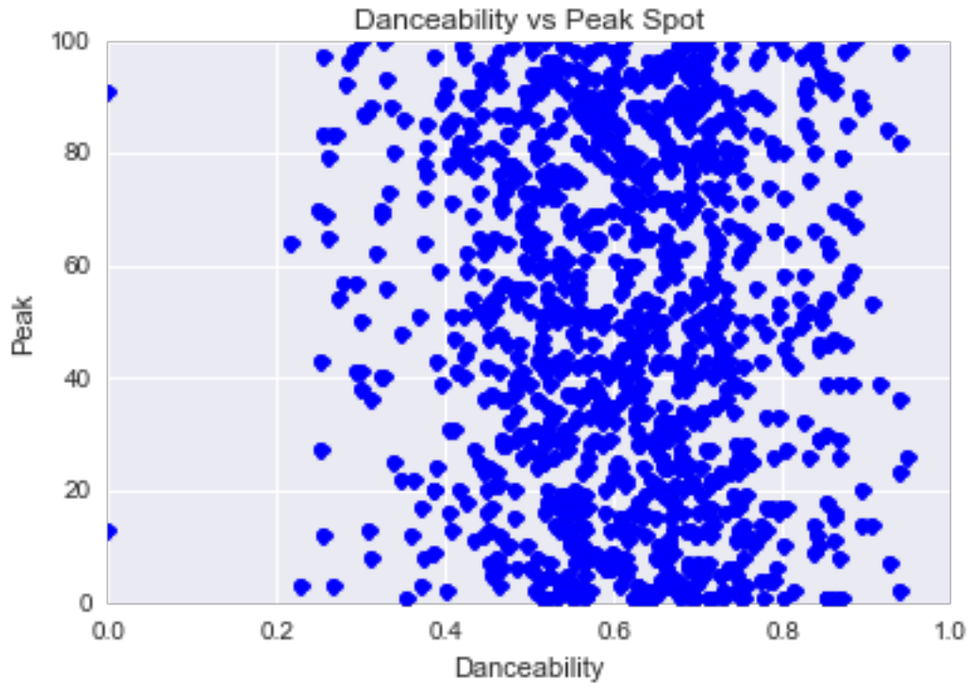

avg_rating
acousticness
loudness

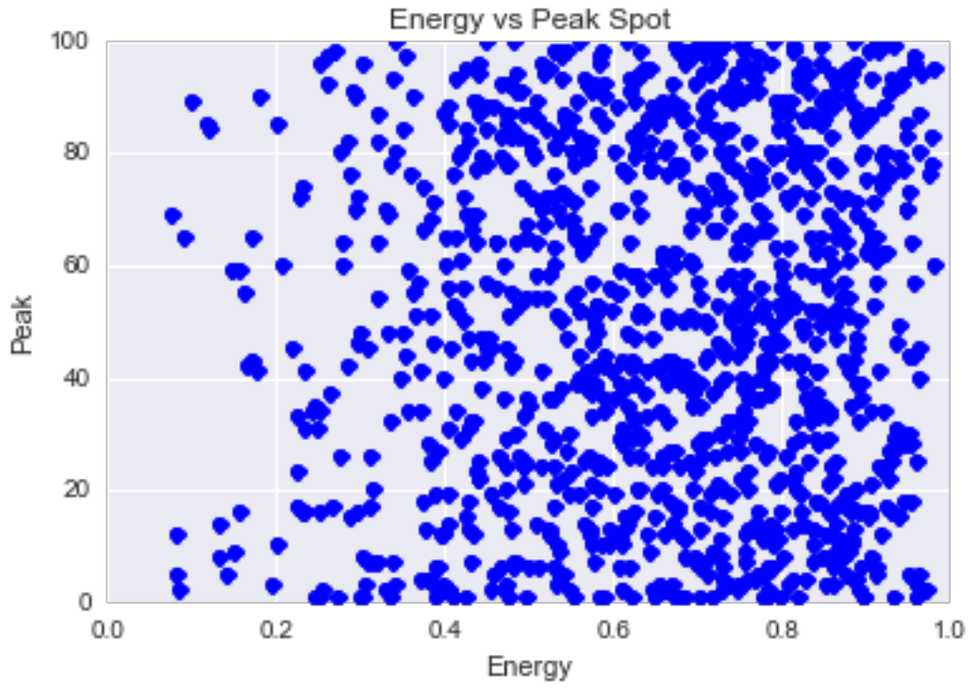

avg_rating
danceability
valence


avg_rating
instrumentalness
longitude



- Many attributes do not have a strong correlation to average rating
- The aforementioned observation that number of features is negatively correlated with chart success is confirmed
- Songs that have are happier have a slight positive correlation to average rating





- The features in isolation did not offer observable trends
- Low energy songs that manage to chart peak in the Top 20
- High Danceability songs peak across the Top 30 range

Conclusion

This hypothesis has been neither proven nor disproven by my data. The best fit line for danceability indicates a slight positive correlation, while energy indicates a almost ignorable negative correlation.

Most songs have a similar level of danceability or energy, so the middle ranges of these two scales are very densely populated. Perhaps creating some scale that creates more diversity in the songs' scores would help better define this area.

Add. Findings

Drake tops the list with 30 charting singles; followed by Glee Cast with 29

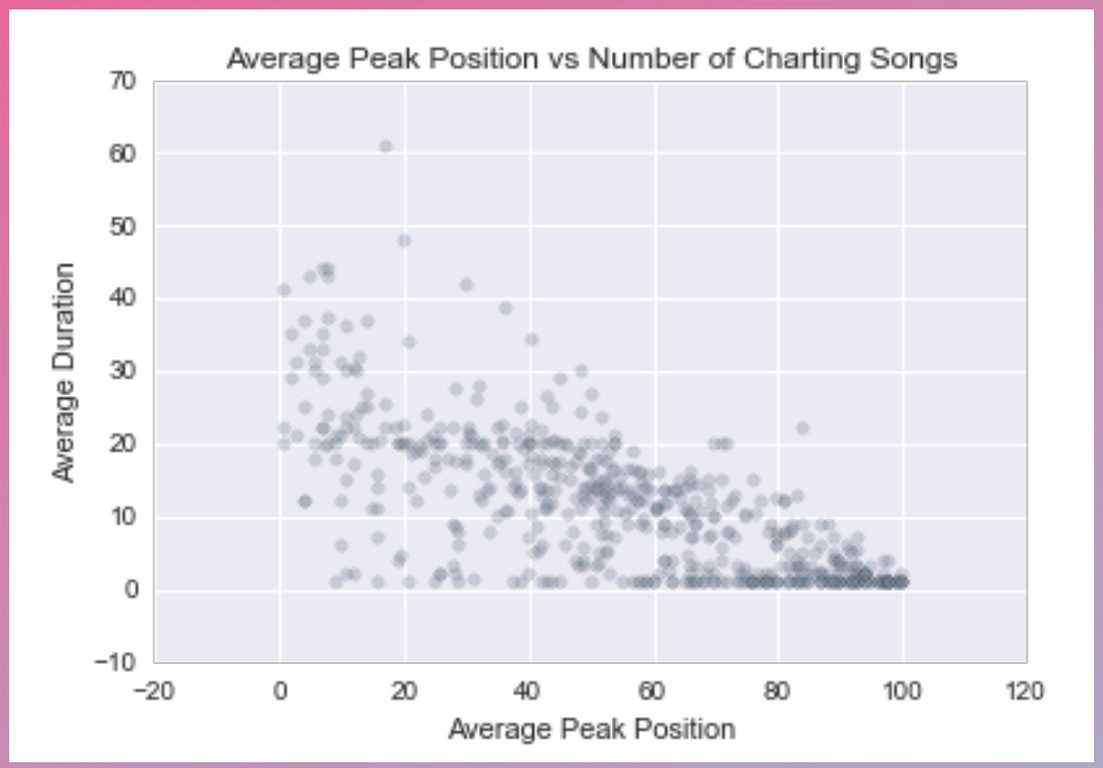
Top 5: Drake, Glee Cast, Taylor Swift, Justin Bieber, One Direction

Many songs that peak in the Top 50 fall off the chart by the 20th week

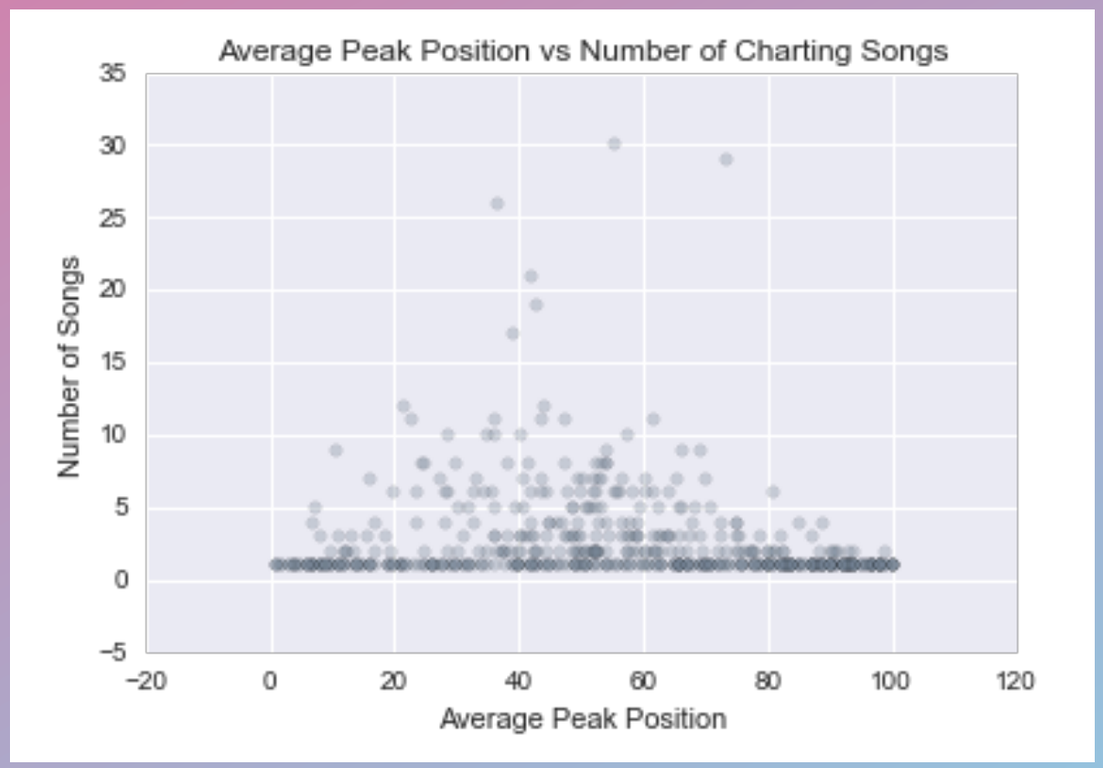
Within 20 weeks, most songs that remain on the chart will peak in the Top 20

There are a total of 34 #1 singles in the last 3.5 years

Add. Findings

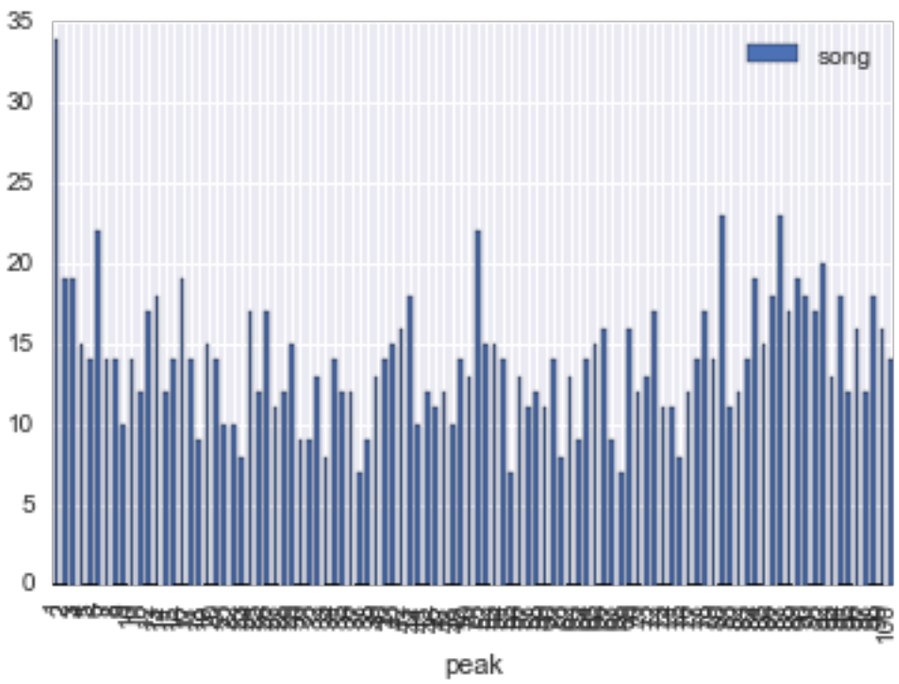


Most artists only chart for a couple weeks at the end of the chart

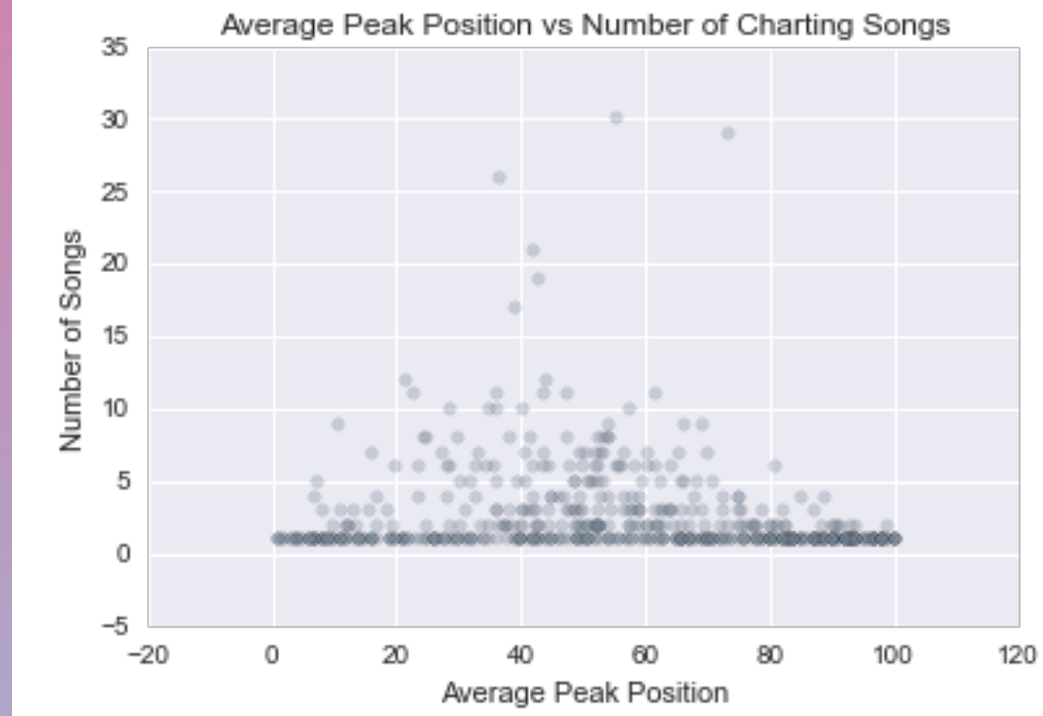


The charts are dominated by a very small group of artists; most other artists have only 2 singles

Add. Findings

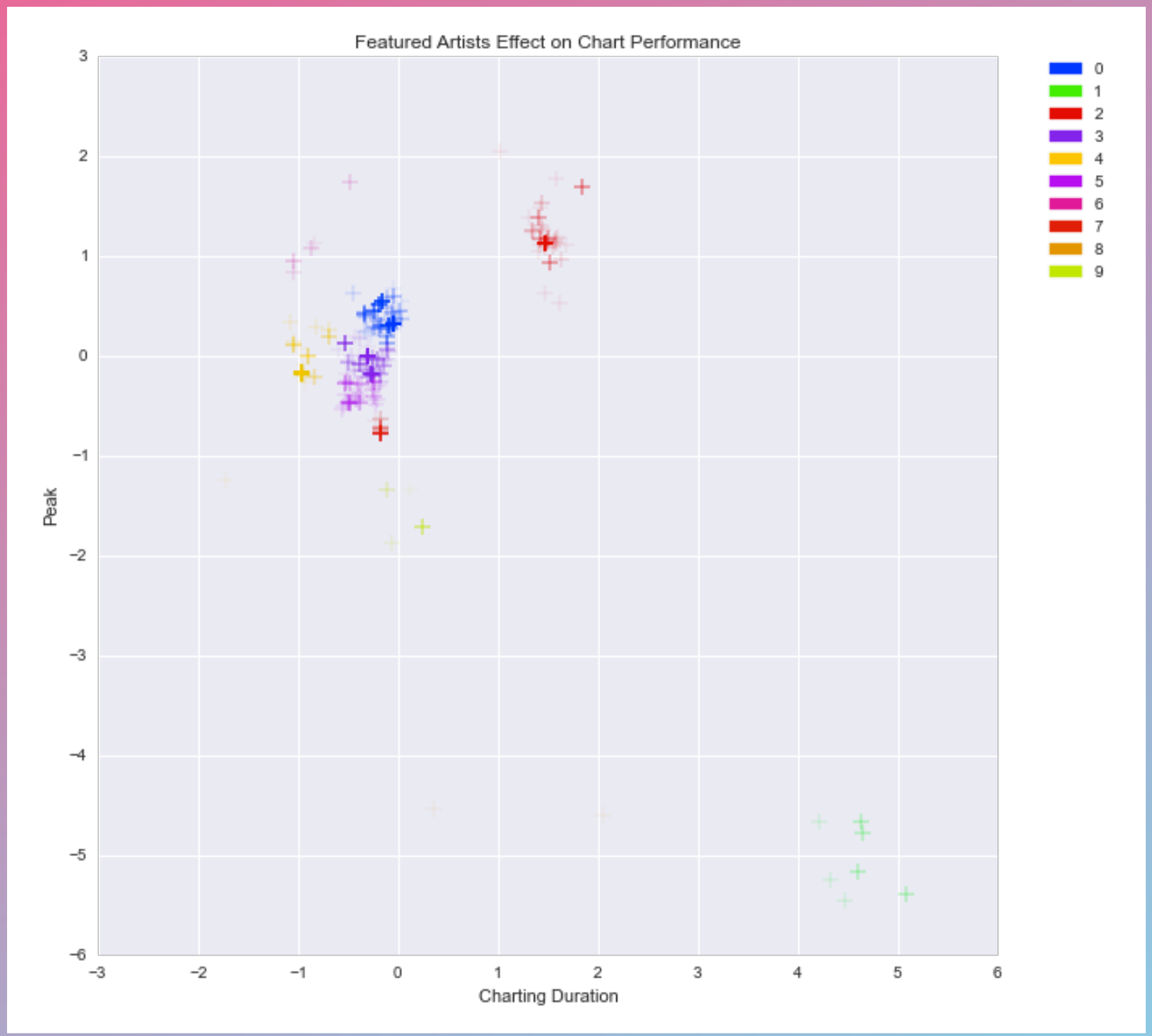


Songs' peak positions are distributed fairly evenly across the 1-100 range



The charts are dominated by a very small group of artists; most have 2 singles

Add. Findings



- Few charting songs are from Canada, the Caribbean/Latin America, and Australia/NZ
- New York, London, Florida, and Nashville are the most dense areas producing a bulk of our hits today

Add. Findings

Lyrics Clustering

Cluster 0: let, know, like, just, heart, feelings, iwill, time, bridge, cause

Cluster 1: loving, like, know, heart, let, just, ca, cause, feelings, iwill

Cluster 2: baby, like, just, know, loving, night, yeah, got, let, oh

Cluster 3: oh, yeah, like, know, got, just, loving, cause, ca, time

Cluster 4: got, like, yeah, girl, know, just, got, ta, want, cause

Cluster 5: na, wan, wan, gon, gon, like, just, know, make, girl

Cluster 6: sd, did, say, like, know, things, day, only, time, man

Cluster 7: n****s, sh*t, f**k, got, like, b*tch, know, just, money, man

Cluster 8: b*tch, f**k, like, got, n****s, sh*t, make, bad, know, just