Yui Chi Tiffany Lo
CS591
Professor Terzi
Spring 2015

Homework 6 : Project Update

Based on my interest in music and pop culture, I designed my project around analyzing patterns in popular music in recent years. There are many approaches in terms of datasets to explore this area, so I decided to look at three of them – Billboard Hot 100 charts, lyrics, and The Echo Nest music metadata.

Billboard Hot 100 Charts

Initially, I obtained a spreadsheet of chart data from the 1920s-2013 from the work of the Whitburn Project to complete this project. However, I decided that I would much rather explore current data rather than the past. This meant that I had to scrape the Billboard website for Hot 100 chart data. My dataset contains dates beginning with January 2012 to April 25, 2015. There are 1397 unique songs in the dataset and they range from charting for only a week, up to an impressive 87 weeks by Imagine Dragon's *Radioactive*.

Here are the attributes of the dataset:

```
data columns (total 94 columns):
song        1397 non-null object
artist      1397 non-null object
featuring    379 non-null object
debut       1397 non-null datetime64[ns]
duration    1397 non-null int64
peak        1397 non-null int64
peak_week   1397 non-null datetime64[ns]
1           1397 non-null float64
2           1076 non-null float64
…
```

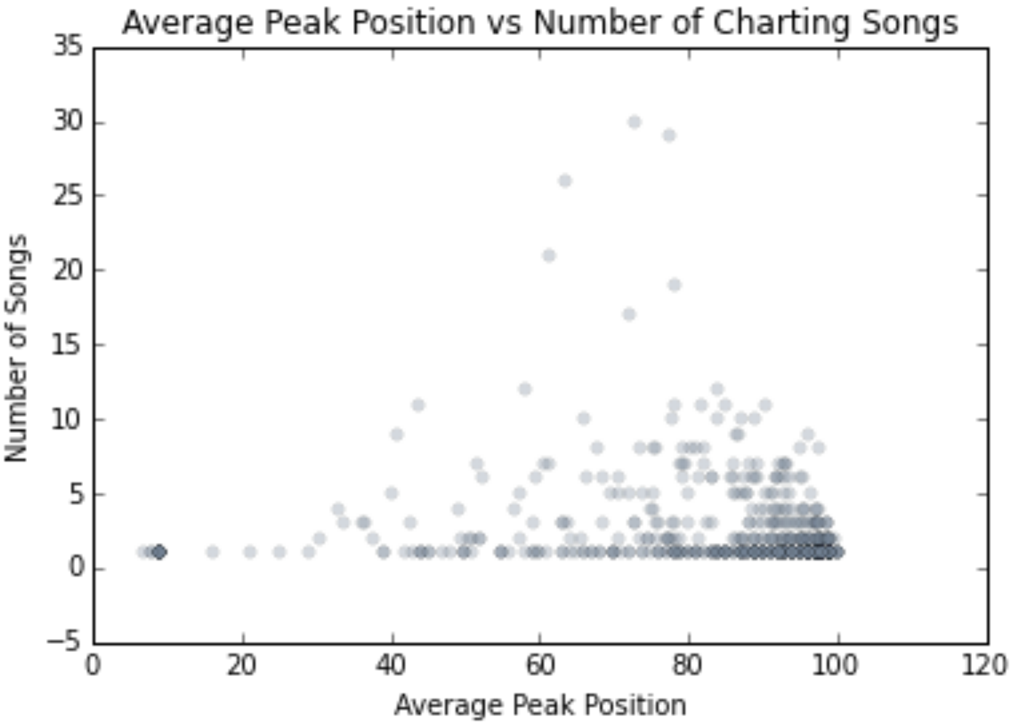
Essentially, this table consists of the song title, artist name, featuring artists, debut week, duration on the chart, peak spot on the chart, week of the peak spot charting, and chart status for each week the song remained on the chart, which ranges up to 87 as that was the maximum number of weeks on the chart.

Based on this data, we can already get an understanding of what artists have had great successes over the past three and a half years of music. Daft Punk has the highest average peak spot of 7, though they only charted for a brief period of time, as compared to artists such as Drake and the Glee Cast who have 30 and 29 different hits on the Billboard Hot 100 charts over these years and therefore a wider spread in terms of their peak spots.

Other interesting preliminary statistics include:
Average peak: 80.7473156764
Average duration: 12.3836793128

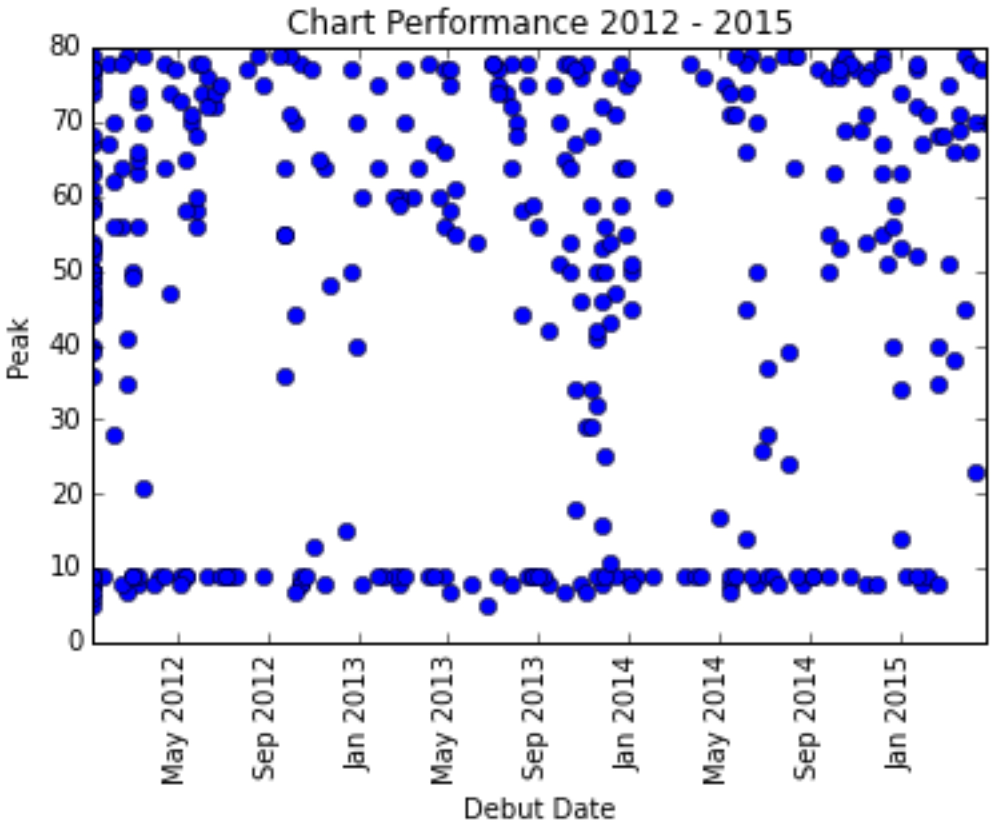
Peak	Min: 5	Max: 100
Duration	Min: 1	Max: 87

Number of artists overall: 546
Number of artists above average peak: 383
One visualization I looked at so far is the relationship between Average Peak Position and Number of Charting Songs per artist. This only includes primary artist recognition and no features that the artist may have contributed to.



We can see that most artists do not have more than three hits on the Billboard Hot 100 during the past 3.5 years. More interestingly, we see that there is a larger concentration of points at the bottom right of the graph, indicating that most artists only had 1-3 hits and they peaked at 80-100. A very small set of artists that appear on the chart dominate the upper positions and they have peak positions that are in the middle range of the rankings between 40-60.

Another perspective I tried to investigate included how the debut date of a song may affect its peak chart position. The graph data is not very conclusive, even as I reduced the points down to a subset of songs that have charted at least a peak of 80. The spread is distributed throughout the x and y axis as there are always songs ranging in the higher and lower ends and songs that leave the charts with each week. I would have to rethink this a bit to get a better image of what is happening.



Lyrics
The lyrics for all the charted songs were obtained by scraping Genius.com, a website that provides lyrics and annotations powered by an online community. Not all the songs were available on the website, though there are very few well-formed, free lyrics repositories to use, so I found this to be the best solution. Eventually I still managed to get lyrics for 1357 songs that are all saved into csv files for further analysis.

The Echo Nest Music Metadata
A few options for online music databases exist, yet again finding a free and robust solution was challenging. The Echo Nest provides a sufficient free-tier of their expansive music API with some interesting information about a very wide range of songs. They have a Python library API that was very easy to use and with several hours of patiently scraping and sleeping due to their stringent request limits, I was able to download data for 1101 songs. There were errors in my coding system of songs of the same title from different artists that may have caused discrepancies here.

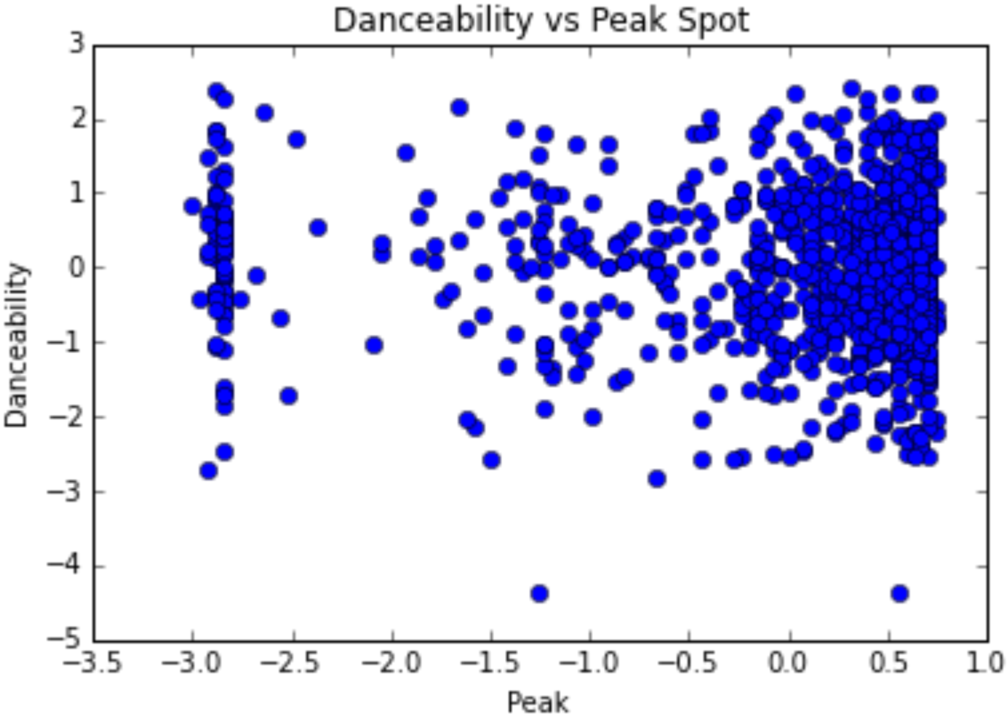
From this database, I obtained the following fields:

```
data columns (total 17 columns):
song              1100 non-null object
artist            1100 non-null object
featuring          311 non-null object
time_signature     1100 non-null int64
energy            1100 non-null float64
liveness          1100 non-null float64
```

```
tempo              1100 non-null float64
speechiness          1099 non-null float64
acousticness         1099 non-null float64
danceability        1099 non-null float64
instrumentalness    1099 non-null float64
key                1099 non-null float64
loudness            1099 non-null float64
valence            1099 non-null float64
location           1077 non-null object
longitude           1099 non-null object
latitude           1099 non-null float64
```

Many of these fields are less technical, but interesting features of the music as determined by The Echo Nest's algorithms. They characterize the type of the music and also provide geographical information which may be helpful to know.

Here's an example of combining the datasets:



Hypotheses I intend on proving:

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

   I intend on using the Natural Language Processing Toolkit (NLTK) to analyze the corpus of lyric files I collected, calculating the number of unique words in each song and other statistics that may cluster songs and define other patterns of the charts.

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

   I intend on using my collected Billboard data to generate data based on features, multiple artists, and bands versus individual artists to get a better understanding of whether solo acts or ensembles fair better on the charts.