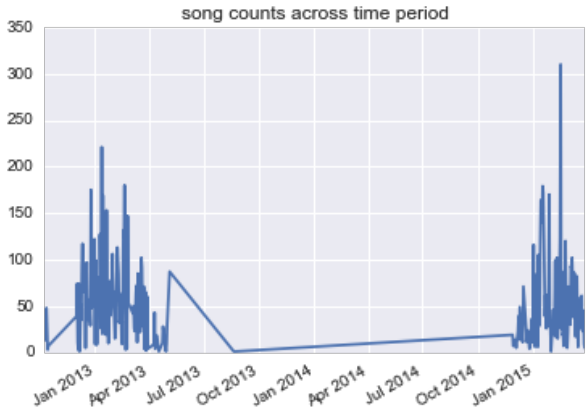
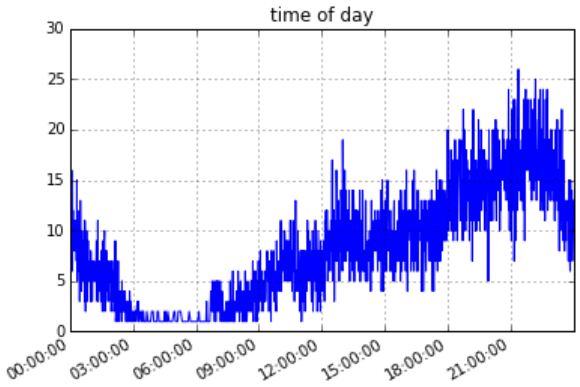
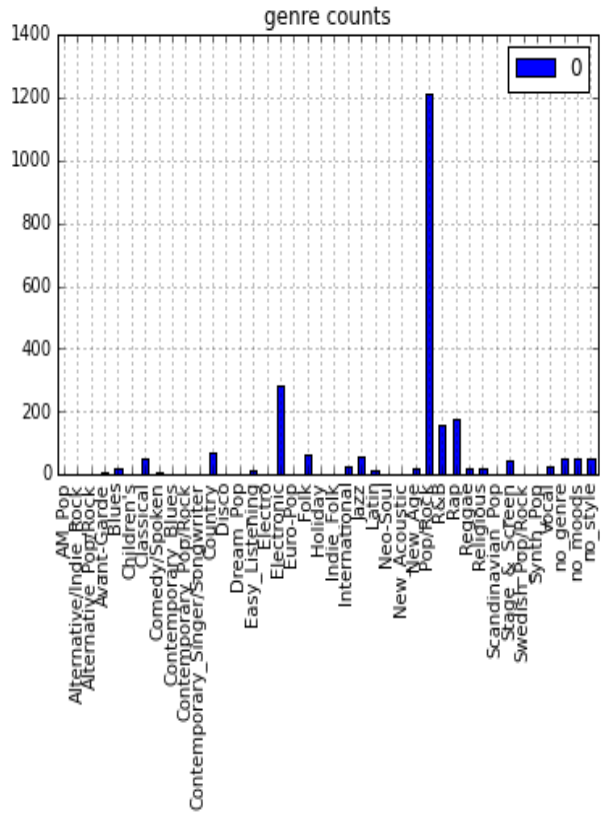
My project incorporates three datasets. The first data set is collected from my personal Facebook data. From that I was able to scrap the history of songs I listened to on Spotify for the time periods of January 2013 to April 2013 and January 2015 to April January. I must have turned off Spotify or stopped listening to it during that time period. The magnitude is **4460 observations** during 3-month period starting 01/01/2013 and **4830 observations** during 3-month period starting 01/01/2015. The headers of the datasets are **3 features** across, timestamp, name of song, name of artist. This was parsed from a HTML document provided by Facebook. These datasets will be collectively referenced the **Facebook dataset.** The charts below show the distribution of times I listened to music, by counting the number of songs per time period. The first shows distinguishes the two time periods I will be studying. The second shows what time of day I listened to music. This is very revealing of my sleep habits. I maintain that I will go to sleep before 3AM regardless of reason and the small spikes is when I fall asleep before I turn off the music. The precision of this chart minutes.

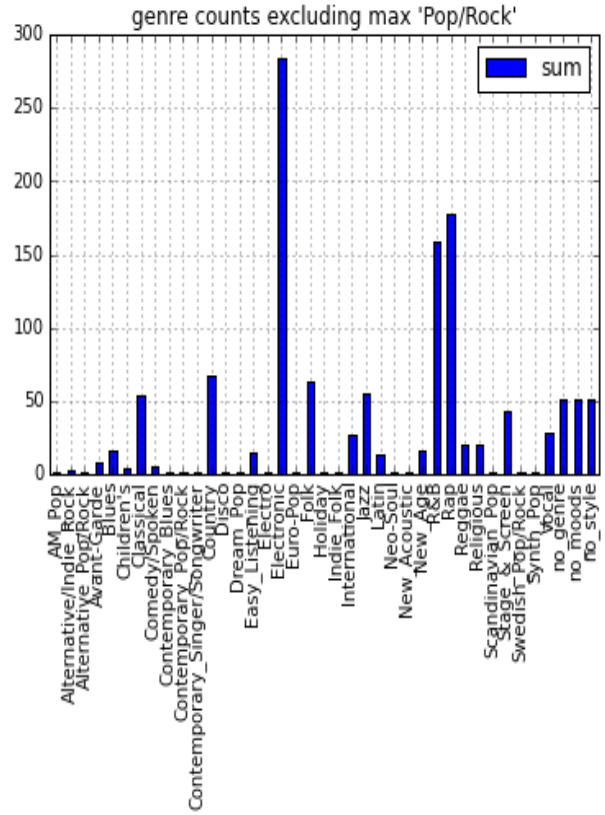


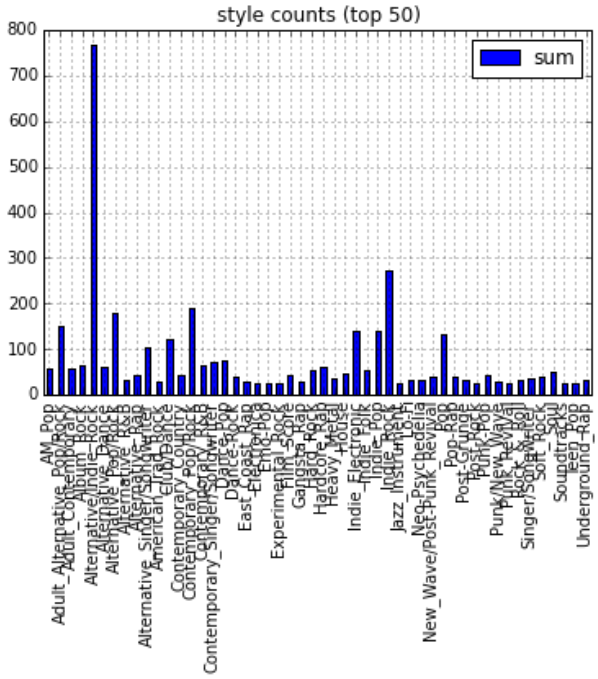


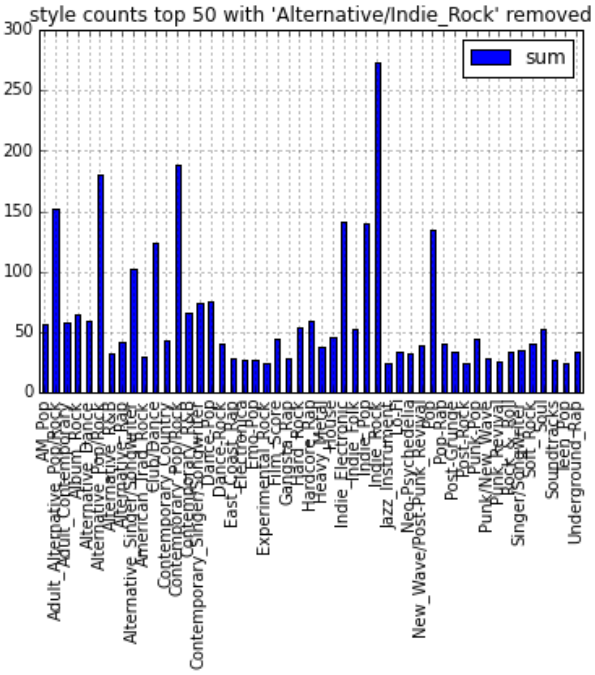
I wrote a scrapper that scrapped allmusic.com for each artist's "genre", "style", and "mood." This will be referenced as the **AllMusic dataset.** There are **41 genres, 525 styles, 274 moods** and **1790 artists**. There is generally 1 to 2 genres, 5-10 styles, and up to 20 moods

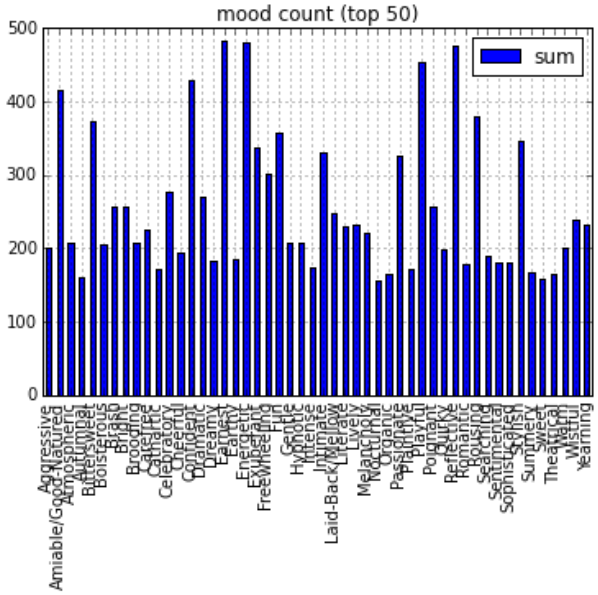
assigned to each artist. From these metrics, it appears that **style** will be the best descriptor, because of the high number of variability, but lower frequency. Genres there is less variety in genres and greater frequency with moods. For example, Lil Wayne is described as having a "genre" of 'Rap' and 'Pop/Rock' and having a "style" of 'Southern Rap', 'Dirty South', 'Hardcore Rap', 'Pop' and finally as having a "mood" of "Ambitious, Angst-Ridden, Boisterous, Brash, Bravado, Confident, Exciting, Trashy, Celebratory, Exuberant, Freewheeling, Harsh, Hedonistic, Humorous, Provocative, Quirky, Rambunctious, Rollicking, Slick, Snide, Street-Smart ,Uncompromising, Whimsical, Confrontational, Sleazy, Stylish, Aggressive, Energetic, Malevolent, Outrageous, Raucous, Rebellious, Reckless, Rousing, Rowdy, Thuggish, Urgent, Visceral, Volatile" These graphs are counts of artists per each category (total artists = 1790)







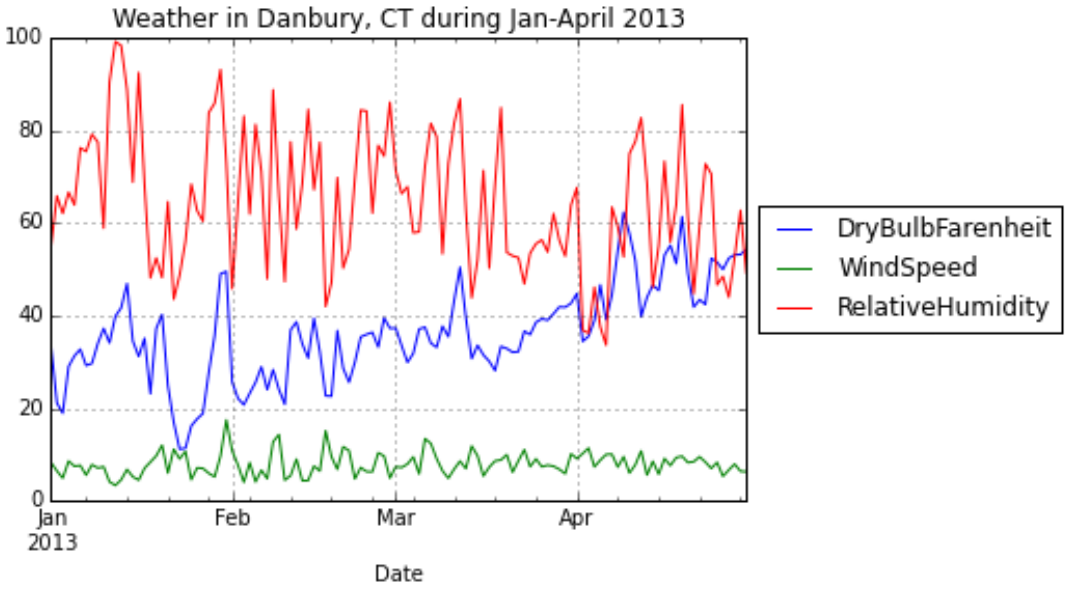


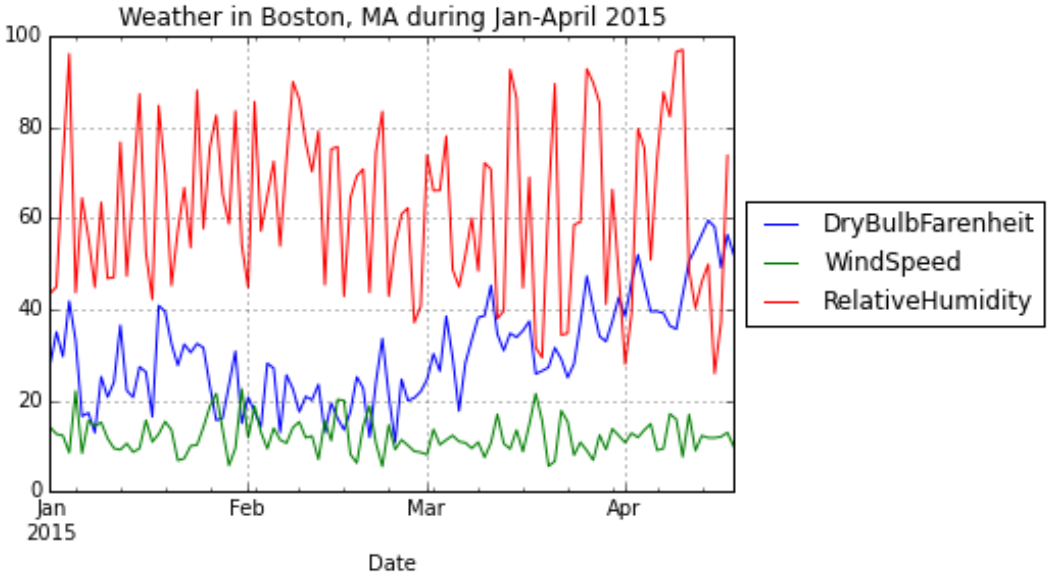


Some interesting features from these charts is the very common occurrence of 'Pop/Rock' accounting for more than half of the entire dataset (~1200/1790). I would be curious to see how effective a binary classifier (Pop/Rock and not Pop/Rock) would be, based on the

styles and moods of the artists. In addition, I have a dataset the size of the Facebook dataset that matches each song observation to the corresponding artist and keywords associated with that artist.

The third dataset, referenced as the **Weather dataset** , is data collected from NOAA.gov. Since I was living in New Milford, CT during the first time period I collected the weather data from the nearest weather station in Danbury (~15 miles away from my general location). For the second time period, I collected data from the Boston Logan Airport weather station(~3 miles away from my general location). In this dataset, information is indexed by day, and there are almost 30 features describing the weather. This is a very sparse matrix. This data is indexed by hour so I can attempt to match hourly trends in weather with hourly trends with my music selection.





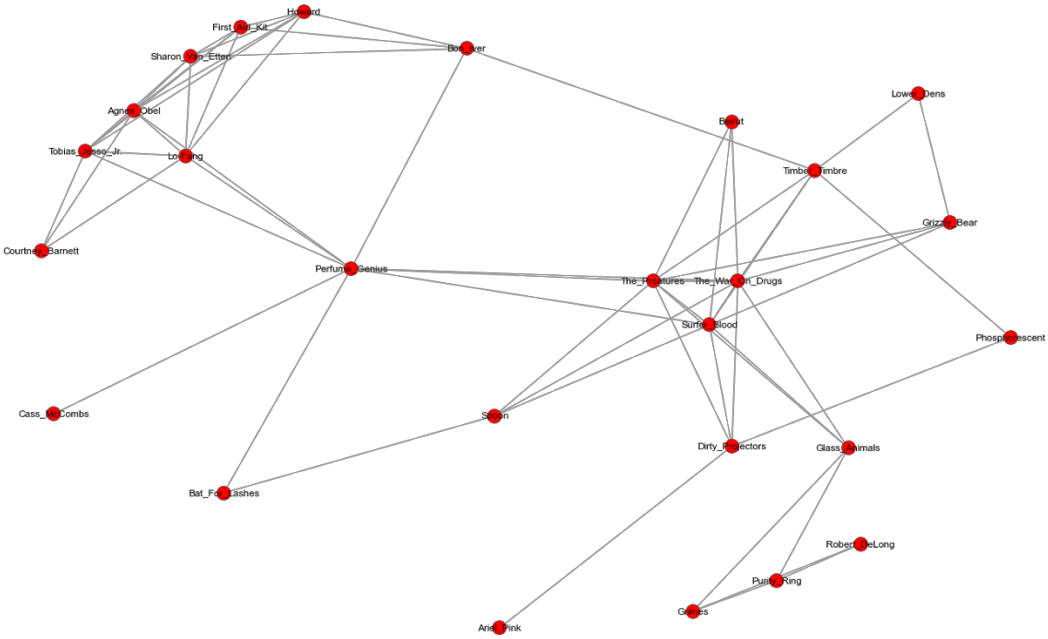
I have a lot of data so it will be important to focus my tasks. First, I want to explore the **AllMusic dataset** and the artists a bit more with a graph. Each artist will be represented as a node and edges will be determined by a Jaccard distance. The top *k* most similar artists will have edges drawn (I still need to determine *k,* but it will be on the power of 10^1)

The next direction I want to explore is connecting the **Facebook dataset** and the **AllMusic dataset**. I will cluster the data and investigate any interesting clusters. I do not see dimensionality reduction being useful in this context because each feature and observation plays a distinct role holistically. I will then introduce the **Weather dataset** and recluster the data and investigate interesting clusters. I will group the data based on different features, specifically, time of day (morning/afternoon/evening) and strong weather to investigate a basis for my music selection.

I also want to build a classifier to for Rock/Pop and not Rock/Pop songs. Since the distribution of Rock/Pop and not Rock/Pop songs (and artists) is about 50-50 this will be a good feature to try to classify. The features I will are styles and moods to classify the genre (pop/rock or not pop/rock).

Time permitting, I will look into building a graph where each song is a node. Edges will be determined based on the temporal distance (provided by the **Facebook dataset**).

This is a preliminary graph drawn from the **AllMusic** data set. Each node represents an artists and the edges are determined by a Jaccard similarity of more than 0.995 in represent to only the **moods** of the artists. The nodes represent a subset of the first 50 artists I listened to, as unconnected nodes were eliminated.


Howard
First_Aid_Kit
Sharon_Van_Etten
Bon_Iver
Agnes_Obel
Tobias_Jesso_Jr.
Low_Cut
Courtney_Barnett
Perfume_Genius
Beirut
Lower_Dens
Timber_Timbre
Grizzly_Bear
The_Pixies
The_War_On_Drugs
Surfer_Blood
Phosphorescent
Cass_McCombs
Sohn
Dirty_Projectors
Glass_Animals
Bat_For_Lashes
Robert_DeLong
Purity_Ring
Geotics
Ariel_Pink