Soundwave – Lead Data Analyst mini-Project

Pre-Requisite – Gathering the data

- Initial step involves reading the JSON data into whatever tool you wish to carry out your analysis.
- Soundwave_JSON_Import.ipynb demonstrates multiple methods for reading the JSON files into a python environment and also displays how to loop through multiple JSONs and append them.
- Since the JSON files are roughly 120Gb the data will need be imported into a Hadoop Hive. A JSON serde can be used so the Hive may map the JSON data to the table columns.
- I have not explored the setup of a JSON hive in this project.

Platform Trends

Question 1 – Analysis of Streaming Service position from a Soundwave perspective

The iPython notebook Soundwave_DataAnalysis.ipynb displays a method of analyzing Spotify and Tidal's music listenership position from a Soundwave position. The process uses two simplified data sets that contain the key features from the User and Play database. The sample data spans three years from 2013 to 2015. My methodology is as follows:

1. **Import Data**
   Import the data into the respective environment you wish to carry out the analysis and clean the column names. For initial analysis I will firstly just import the Play database.
2. **Add Year Column**
   Format the time column and create a year column based of the time column.
3. **Eliminate iOS Data**
   As instructed remove all iOS data for the purpose of the analysis using the platform field, similarly the device field could be used.
4. **Evaluate Respective market Shares**
   Use the source_name column to compute the respective market shares for 2013-2015

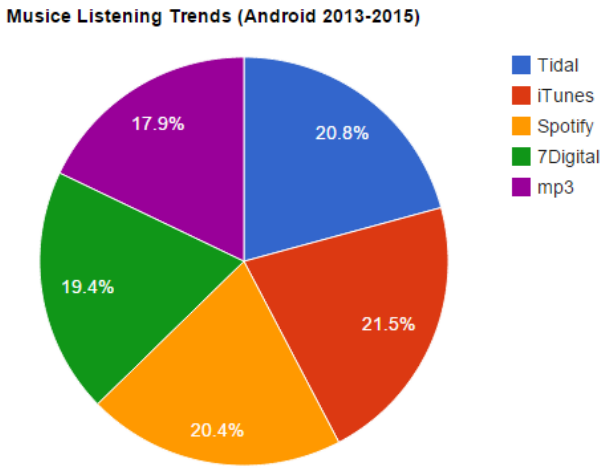

Figure 1: HTML visualization of market share (Chart1.html)

5. **Segment the data by year**

In order to see if the market share of Tidal/Spotify is trending upwards it would be interesting to segment by year and visualize the trend.

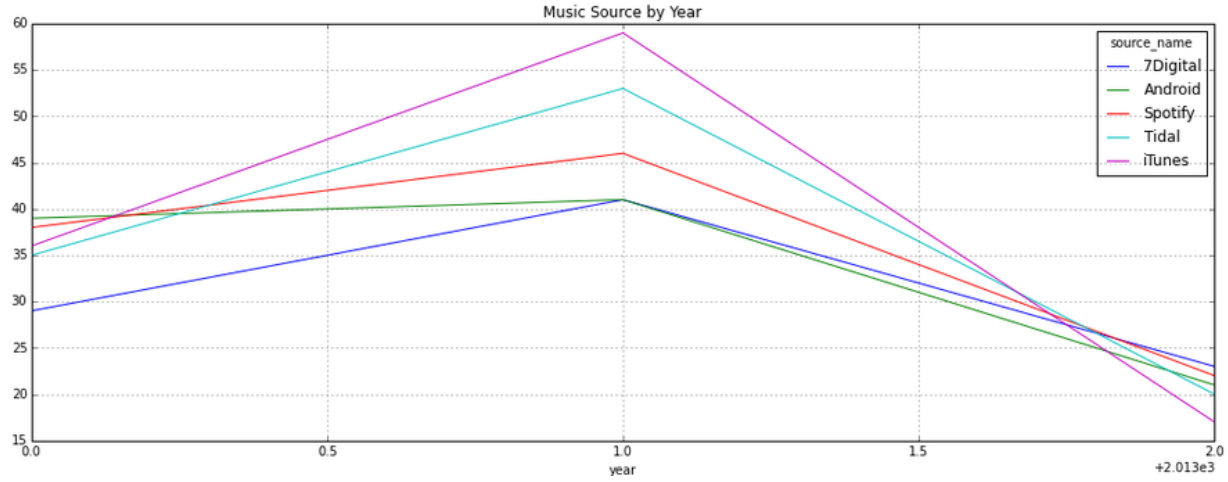

Figure 2: Market Share trend (taken from Sondwave_DataAnalysis.ipynb)

6. **Group the various providers**

For visualization purposes I group Spotify and Tidal into NextGen and iTunes, 7Digital and Android into OldGen.

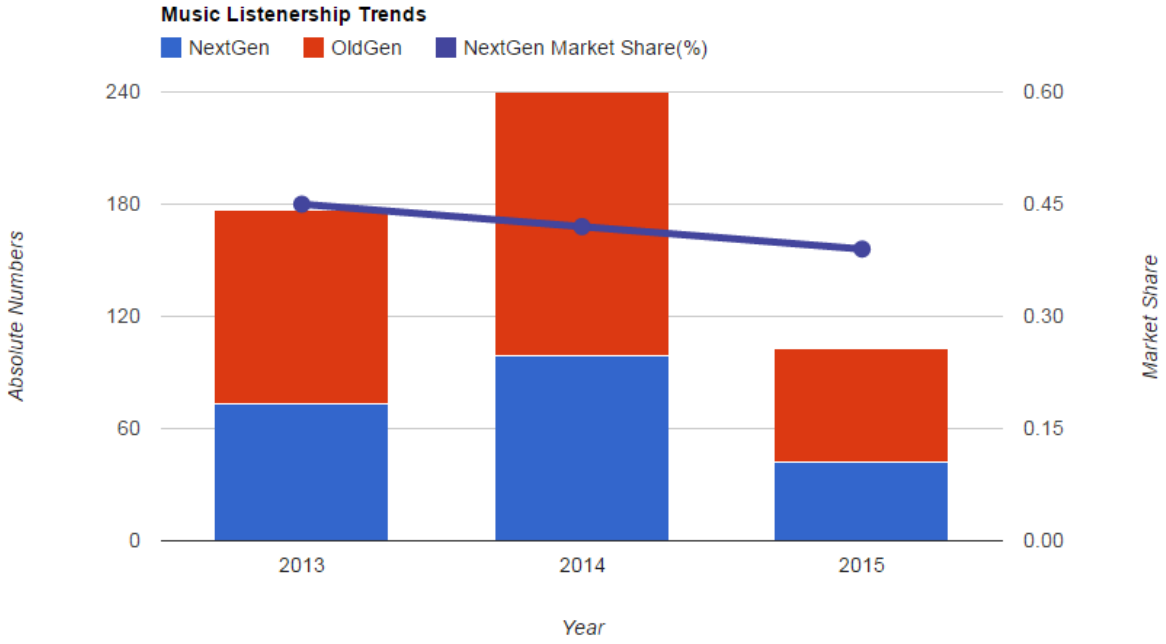

Figure 3: HTML visualization of market share by year (Chart3.html)

7. **Merge User data and analyze based on DOB**

I have a sense that the market share for the NextGen providers would be higher for younger users as they will consume new digital technology more readily. To test this premise merge the Play database with the User database and analyse based on DOB.

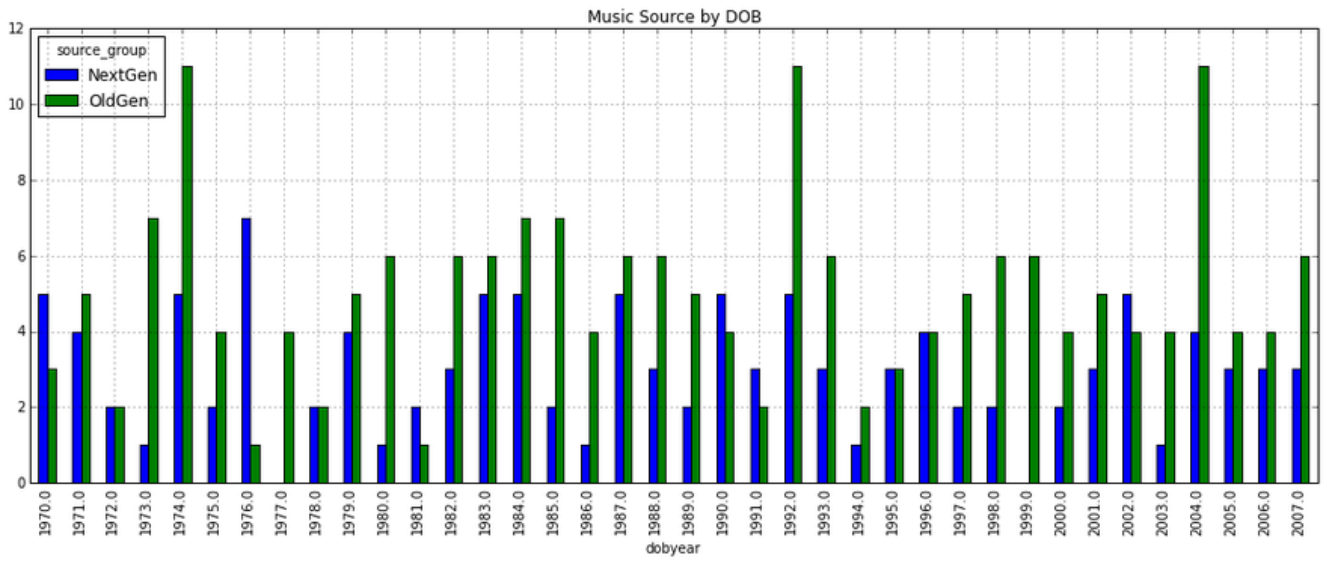

Figure 4: Market Share by DOB

8. **Further Analysis**

   The ways I could segment this data is pretty endless but some of the key ways I would are as follows:

   - Bin DOB into three groups and visualize market share, also split by year (2013-2015) and see if there is an increasing trend in a particular age group.
   - Split by sex and see if NextGen providers are capturing one gender more than another.
   - Split by month and see if there are any seasonal trends.
   - Identify devices where market share is stronger among NextGen and seek to find explanations.
   - Use long and lat co-ordinates to identify regions with high NextGen listenership.
   - Identify genres that are more popular on NextGen.

Question 2 – Can you extend any conclusions to the music listening population?

I feel that the type of person who uses Soundwave is exactly the type of person who uses the NextGen music providers such as Spotify and Tidal. Thus any NextGen market share we analyze should be assumed to be skewed towards a Spotify/Tidal market share representation and does not provide a clear picture of the mobile music listening population.

Question 3 – Using the Play database how you could segment any findings by region.

The Play database contains Longitude and Latitude co-ordinates associated with each play. There are numerous ways we could use these co-ordinates to break down our data by region. The simplest method would be to use a Google maps API call but since these are limited and we would require thousands of API calls this method will have to be dis-regarded. A more prudent method is outline below:

1. Build a longitude/latitude database based on your region requirements. If we want to segment by country our first step should be to build a database with the longitude/latitude co-ordinates of every countries most central position. This should be available online in a usable format.
2. Then we can compute the nearest region for each data point by using the nearest-neighbor algorithm.
3. This method doesn't depend on any external sources and should be relatively efficient, however this method (not the algorithm) does have a degree of error associated with it.

User Ages

Question 1 – Can you support this theory from any body of industrial or academic research?

Wu, Jang and Lu present a methodology for estimating age based on music metadata in their paper "Gender Identification and Age Estimation of Users Based on Music Metadata".

The author predicts a listeners age by using music metadata to construct a Term Frequency – Inverse Document Frequency (TF*IDF) and Gaussian Super Vector (Hotness Factor) feature method. The TF*IDF feature method measures the frequency of each artist and their discriminant power to predict an age with a mean square error of 4.25 years. The Gaussian Support Vector/Hotness Factor works on the basis that each artist exudes a 'hotness' to different ages and results in a predicted age with a mean square error of 3.69 years. Both of these methods use a similar premise to the one outlined by Soundwave in that they believe certain artists are related to certain age group (teenagers popular chart etc.).

http://www.terasoft.com.tw/conf/ismir2014/proceedings/T100_278_Paper.pdf

Question 2 – Outline a method of computing a user's age based on the music they listen to.

Using the methodology outlined in the paper above the initial step in the process would be to collect a sample data set (training data) of music listeners with associated music metadata and an age. From this it will be possible to compute a hotness factor associated with each artist ie. Evaluate what artists are related to what ages.

Secondly using the training data we can construct a document for each user containing the artist's name of the top n users listened to songs. From this list a TF*IDF feature can be engineered. Latent Semantic Indexing (LSI) should then be used to reduce the dimensionality of the TF*IDF feature.

Then using our document created in the creation of the TF*IDF and the hotness factor we can then transform the artist related document into a hotness document. Gaussian Support Vector can then be applied to this bag-of-features to create a usable feature for each user in our model.

These features can then be used in a model to classify the age of a user, also it is important to note the paper identified the correlation between gender and age so it would be useful to use a gender feature in our model. A regression model can then be trained using these features to predict the age of each user in the Sondwave User database. The paper recommends a support vector regression model as the features we engineer will be vectorised.

Now we have a trained SVR model we can use for classifying the age of Users in our User database. However we will also have to carry out some of feature engineering steps carried out on the training data. We will build a term frequency document and use it to build a TF*IDF and GSV feature. Finally we can now feed these features into our model and predict the age of our User.

Question 3 – How else might one classify the ages of the Soundwave user base?

Some ideas on how to infer age:

- Time of plays – younger users will use the device late at night, middle-aged professional will us it around commuting times etc.
- Frequency of Device Change – How many different mobile devices have been used to play music, younger people will change their device more frequently.
- Users name – Certain names could be used to infer age as certain names have been popular at specific points in time.
- Location Data – a sophisticated geo-location model could help identify students (colleges) and other locations which correlate with age.
- Followers/Following – Using follower and following information we could infer that users age are closely related to those in their cluster. Also I would have a feeling that younger people follow more and have more followers than older people.

Idea

Assist with the Chart Gap

With the demise of music purchases and rise of digital music consumption gauging the popularity of an artist/album/song is becoming increasingly difficult. Sondwave data could be leveraged to identify rising stars in the music industry and labels can then invest money in artist in a more analytical and informed manner. This information can also be used to identify popular genres or artists in particular geographies from which advertisers could select these songs for advertisements.

Gauge Public Reception

An interesting use of Soundwave play data would be to gauge what effect publicity, advertising and news has on listening trends. For example did Chris Brown's assault on Rihanna negatively impact his listenership and increase Rihanna's. Does the repeated rumours of an Oasis reunion result in increased listenership for Oasis, The High Flying Birds or Beady Eye. These analyses may not prove economically beneficial. But if we learn that these events lead to increased demand advertisers could then react accordingly to news as it comes in and push artists and songs that are likely to see an increase in demand and from this we then can build a revenue stream.

Transportation

An analysis of listening trends in the lead up to large concerts and festivals may show that people travelling to these venues begin listening to the artists playing at these venues more frequently. Thus we could then identify likely transportation routes and requirements and any possible shortages.

Booking Artists

This again is related to my first point round the music chart gap. Analysis of Soundwave data could help, bars, clubs and festivals book artists who are being listened to more highly in their respective region. In particular for festivals the aim is to book a number of artists who appeal to as many people as possible, with Soundwave this requirement could be fulfilled in a much more analytical and accurate manner.

Platform Monitoring

Monitoring of time-series platform usage should help identify any bugs associated with a particular operating system.

Identify Music Flows

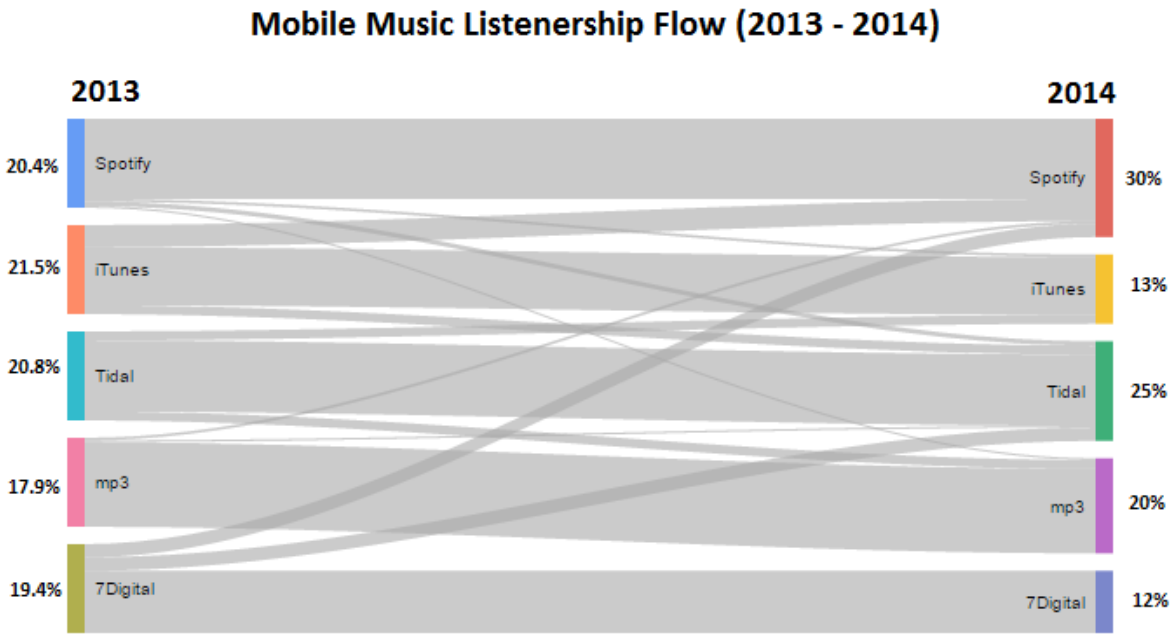

Figure 5: Music Listenership flow (Sankey.html)